Big Data and Serious Crime Investigations

by

Khalid Al-Ali

A thesis submitted in partial fulfilment of the requirements of
Liverpool John Moores University for the award of the degree of
Doctor of Philosophy in Policing Studies
December 2025

Declaration

That no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Acknowledgments

I am profoundly grateful and would like to thank God for all his blessings. Words cannot express how thankful I am to my lovely parents, my incredible mother and father that I owe a debt of gratitude which I can never repay. I would like to thank them for all their support, blessings and prayers that enabled and guided me to what I have achieved. My heartfelt thanks to my dear sisters, my brother, and my whole family who have supported through this journey.

Also, I am beyond grateful and thankful to my incredible country, the land of dreams and possibilities the United Arab Emirates, and to our great rulers and the leadership that have always supported us with endless opportunities to develop ourselves and our country. I am extremely thankful to Dubai Police Head Quarters for sponsoring me to achieve my MSc and PhD degrees, starting from His Excellency Lieutenant General Dhahi Khalfan – Deputy Chief of Police and Public Security in Dubai for believing in me and for his personal support, for which I will forever be grateful. I would also like to thank His Excellency Lieutenant General – Abdullah Al Marri Commander in Chief of Dubai Police for his continuous support to the whole force in providing us with opportunities to develop ourselves academically and professionally. Which placed us among the top-ranking police forces worldwide and enabled us to serve our country, our communities, and our people. The journey was made possible by Dr. Khamis Al Muhairi and Ali Al Shehhi, thank you for being my steadfast companions and for the support that carried me over rough terrain. I would also like to thank Dubai Police Scholarship Department for all their support throughout the years, starting from Ms. Rula, Dr. Hamid Murad, Fahad Al Shaer, Khalid Ibrahim, Mohammed Al Obiadly, and Majid Al Balooshi.

I cannot thank enough my wonderful supervisory team Dr Jude Towers, Dr Helen Selby-Fell, and Dr Adrian James who have supported and guided me all the way through my master and PhD degrees. I was very proud and happy greeting them on stage during my master's degree graduation, and I cannot wait to repeat that after my PhD graduation. Their contribution in my research and academic life has been invaluable, and a special thank you to Dr Jude for going above and beyond in every step. I was lucky enough to work with such a great inspirational supportive team that have always been there for me.

I would also like to extend my heartfelt thanks to my dear friends who were instrumental throughout this journey. For all our happy memories, the encouragement, joyful moments, and the ups and downs, thank you all for being there for me: Dr Mohammad Khalifa, Dr Saeed Al Bedwawi, Dr Alhamza Mohammed, Zain Malik, Bilal Malik, Mohammad Al Rahma, Hassan Al Mulla, Fahad Al Saleh, Faisal Al Saleh, Mohammad Bin Karam, Mohammad Al Hadidi, Shuaib Al Raeesi, Mansoor Al Farsi, Hazza Al Farsi, Abdulrahman Al Beshri, and Ali Al Ali.

My beloved grandparents, and my cousin, childhood friend, and brother, Salman, it saddens me that you are not here to celebrate this achievement, I wish you were here today, but I know you would be celebrating with us in spirit.

Lastly, I would like to express my gratitude to all those whose names I may have missed but who played a role in helping me reach this milestone.

Abstract

The growing complexity of criminal activity has encouraged policing to explore and adopt advanced methods to support their operations in crime detection, investigation and prevention. Given the increased risks and harms associated with serious crimes for victims, individuals, communities and countries, these types of crime have become a central focus for policing. At the same time, big data has expanded and there have been developments in the artificial intelligence tools that support its analysis and applications across multiple disciplines where its advantages are perceived. This area has become an interesting focus for research within policing, aiming to advance operational practices. Therefore, this thesis aims to explore what is known about big data and if it can be useful in serious crime investigations for policing. A scoping review was conducted to explore and map the existing literature, clarify key concepts and gaps, and guide the focus of the empirical research. This study adopted a qualitative research approach, collecting data through semi-structured interviews with two groups of participants, one comprising policing professionals and the other of big data experts. The findings reveal the strong potential for big data to be useful in serious crime investigations by providing strategic and operational advantages. However, effective implementation was viewed as complex, with challenges in terms of conceptual, cultural, financial and technical aspects and in employing skilled human resources, along with key privacy and bias concerns. This thesis contributes to the field by clarifying the concept of big data and proposes a definition of its application in the context of serious crime investigations, which can be used by policing internationally. Collectively, the findings offer insights into the core elements that need to be addressed by policing to be able to achieve the strategic and operational advantages.

Keywords: Artificial intelligence, Big data, Crime detection, Crime prevention, Criminal investigations, Policing and big data, Serious crimes

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial intelligence
ANPR	Automatic Number Plate Recognition
APCC	Association of Police and Crime Commissioners
API	Application programming interfaces
CAS	Crime Anticipation System
CCPA	California Consumer Privacy Act
CDCI	Crime Detection and Criminal Identification
DNA	Deoxyribonucleic Acid
EU	European Union
GDPR	General Data Protection Regulation
HDFS	Hadoop Distributed File System
HMM	Hidden Markov Modelling
IT	Information Technology
LJMU	Liverpool John Moores University
NASA	National Aeronautics and Space Administration
NCA	National Crime Agency
OLAP	On-line Analytical Processing
ONS	Office of National Statistics
PRECOBS	Pre Crime Observation System
ROCU	Regional Organised Crime Units
ROI	Return on investment
SAP HANA	SAP High-performance Analytic Appliance
SC	Serious crimes
SOC	Serious and organised crimes
UAE	United Arab Emirates
UAEU	United Arab Emirates University
UK	United Kingdom
UNCHC	UNC Health Care
UREC	University Research Ethics Committee

Glossary of Key Terms

An **algorithm** is a technical set of instructions that a computer performs, for instance to search, analyse, detect and train data modules, performing commands based on its programmed task (Janssen and Kuk, 2016).

Artificial intelligence (AI) represents a wide range of technological tools and systems that are able to assist in procedures such as data collection and analysis, rapidly analysing large volumes of data, identifying patterns and performing complex calculations with the aim of supporting the end user in decision making (Gkikas and Theodoridis, 2022).

Big data refers to high volumes of various types of data that are generated at high velocity and creates value by giving the user the capacity to search, cross-reference and analyse large data sets by incorporating a wide range of analytical techniques (Babuta, 2017).

Big data analytics is a systematic approach to analysing large complex volumes of data and associated attributes to extract useful information by using AI tools and algorithms, such as data mining and machine learning (Feng et al., 2019).

Data mining is among the AI tools that enable the process of discovering hidden information from big data through different computational methods (Hassani et al., 2016).

Datafication is the process of transforming a diverse range of information, such as subjects, objects and practices, into digital data (Southerton, 2020).

Machine learning is an AI tool that is programmed to learn from historical datasets to identify patterns, detect irregularities and make future predictions to inform decision making (Udeh et al., 2024).

Serious crimes are criminal activities that are deemed dangerous by nature. They can be committed by one or more offenders and may be life threatening to the victim or lead to severe physical, emotional and/or financial harm. They are classified differently based on a country's legislation (Paoli et al., 2016).

Chapter 1. Introduction

1.1 Background and context

The big data revolution has reshaped human activities and guided decision making across multiple areas, such as cybersecurity, counterterrorism and policing (Richards and Kings, 2014). The magnitude, speed and diversity of data available to governments, the private sector and individuals, together with advancements in computing powers, are transforming societies (Brady, 2019). The concept of big data can be understood in different ways, with various definitions evolving across fields (Chan and Moses, 2016; Kitchin and McArdle, 2016; Babuta, 2017). There is a lack of agreement on the definition of big data given that is an emerging discipline undergoing rapid evolution (Mauro, Greco and Grimaldi, 2016). However, it is broadly agreed that big data involves high volumes of information collected rapidly through a range of technologies (Jurkiewicz, 2018). Given the continued debate and variability in how big data is defined, this thesis does not adopt a single universal definition to avoid the risk of narrowing the concept. Rather than adopting a debatable definition, drawing on the literature, this thesis views big data in terms of being represented by certain characteristics of relevance to the scope of this research, namely: volume, velocity, variety, value and variability (Zikopoulos et al., 2012; Dijcks, 2013; Dumbill, 2013; Richards and King, 2014; Kitchin and McArdle, 2016; Broeders et al., 2017; Brady, 2019; van der Voort et al., 2019; Bell et al., 2021).

Two main definitions of big data in policing were found in the literature. Schuilenburg and Soudijn (2023, p.1) described it as *"...the use of large volumes of data made accessible by means of algorithms and gathered with the objective of making society safer"* which appeared to have a broad objective of securing the society. In contrast, Neiva, Granja and Machado (2022, p.1167) described it as *"...the processing and analysis of large amounts of information, aimed at supporting policing activities, defining security governance policies, and advancing with criminal investigations"*. This thesis adopts the definition proposed by Neiva, Granja and Machado (2022) as it clarifies the perceived aims of using big data in policing and focuses on advancing criminal investigations.

Within the field of policing, serious crime has a daily and direct impact on citizens, public services and businesses (Winchester, 2020). As an indication of the scale of this phenomenon, in the United Kingdom (UK), there are more than 100 government and law enforcement agencies aimed at tackling serious and organised crimes (National Audit Office [NOA], 2019). Moreover, the big data revolution has encouraged police organisations to enhance and expand their crime fighting and detection methods (Sandhu and Fussey, 2021), particularly as big data technologies have the potential to generate predictions about crime patterns and support investigative outcomes (Neiva, Machado and Silva, 2023).

The following sub-sections provide more in-depth background on current understandings of the uses of big data in policing in relation to serious crime investigations and examine real-life applications of big data in policing, together with serious and organised crime (SOC) strategies in the UK and Europe.

1.1.1 Policing and big data

Growing complexity in the methods of committing crimes in smart societies provides a motivation for police forces to advance their strategies and capabilities to secure society (Ezzeddine, Bayerl and Gibson, 2023). A crime can be conceptualised as a phenomenon that negatively affects the safety and stability of societies (Xu, Cheng and Sugumaran, 2020). Crime destroys lives and devastates communities, and SOC elevates harm and can endanger national security (Home Office, 2023). The role of investigating and solving crimes has been entrusted to policing officers and specialists globally (Nath, 2006). When a crime is reported, the main purpose of a criminal investigation is to investigate if a crime occurred and then to identify the suspect(s) (Almansoori, 2019) and bring offenders to justice (Newburn, 2008). It has been established that the methods and patterns of criminal activities are constantly developing, with criminals exploring and using the latest available technologies to commit crime and avoid being captured (Hassani et al., 2016; Yadav et al., 2017). In the modern age, crimes are not limited to the streets; increasingly, sophisticated criminals are using the internet, giving them access globally to commit crime (Hassani et al., 2016). Therefore, societies' efforts to control crime effectively continue to be a significant area for research (Xu, Cheng and Sugumaran, 2020).

It is clear that big data will have an impact on policing practices as it is doing on all other aspects of the social world (James, 2016). There is an increasing tendency to introduce big data technologies in police departments with the aim of preventing crime by predicting where crimes are likely to occur and supporting criminal investigations after a crime has been committed (Neiva, Machado and Silva, 2023). However, there is a lack of research on the different ways that police utilise big data applications and whether and to what extent big data is useful in policing (Schuilenburg and Soudijn, 2023). As previously noted in Chapter 1 (see 1.1), the use of big data in policing has been described as *"…the use of large volumes of data made accessible by means of algorithms and gathered with the objective of making society safer"* (Schuilenburg and Soudijn, 2023, p.1) and as *"…the processing and analysis of large amounts of information, aimed at supporting policing activities, defining security governance policies, and advancing with criminal investigations"* (Neiva, Granja and Machado, 2022, p.1167). As differences were found between the definitions of big data, this thesis adopts Neiva, Granja and Machado's (2022) definition as a central reference because it offers additional context relevant to big data in policing, particularly its application in criminal investigations, thus broadly aligning with the scope of this thesis.

1.1.2 Serious crime

An initial search of the literature showed that the definition of "serious crime" varies, both in the academic literature and in practice, across policing in different jurisdictions. Therefore, this section explores these differing definitions to develop an operational definition for the purpose of this thesis in the context of policing.

The term "serious crime" was first introduced in an EU policy document published in 1995 but really started to take shape and circulate in the early 2000s (Paoli et al., 2016). Serious crimes are a cause for concern to both the police and the public and are challenging to solve, usually requiring considerable resources to investigate (Almansoori, 2018). The term was increasingly used in shaping security policy in the UK, which resulted in the establishment of the Serious and Organised Crime Agency in 2005 and the UK Serious Crime Act 2007 (Paoli et al., 2016). The National Crime Agency (NCA), which replaced the Serious and Organised Crime Agency in 2013 in the UK states that serious and organised crimes affect UK citizens more often than any other national security threat (Winchester, 2020); SOC have also been described as *"...the most deadly"* national security threat faced by the UK (Home Office, 2018, p.3).

Serious crimes have a daily impact on citizens, public services, businesses and the national reputation of the country (Winchester, 2020) and have been defined as *"...criminal activity that is planned, coordinated and committed by people working individually, in groups, or as part of transnational networks"* (National Audit Office [NAO], 2019, p.5). These organised crime groups operate in changing and unpredictable ways and often use violence to intimidate communities and vulnerable people (National Audit Office, 2019). The driving force behind SOC is frequently, though not exclusively, financial gain (Home Office, 2013, 2023). Advancements in technology offer criminal networks new methods of identifying their victims and committing crimes (Home Office, 2018). For a long time, many such criminals have been one step ahead of policing in harnessing technology and therefore a new strategy and framework were developed as a response to combat serious crimes (Home Office, 2013).

In the UK alone, SOC is estimated to have cost the country £24 billion per year since 2013, increasing to £37 billion in 2020 and reaching an estimated £47 billion per year by 2023 (Home Office, 2013, 2023; National Crime Agency, 2021). These figures can be extrapolated to the wider context, representing a significant drain on the global economy, as well as the human victims. Thus, there is an urgent need to address SOC, which is one of the aims of this thesis. Cyber-crimes are typically conceptualised as serious (Winchester, 2020; Europol, 2023); they pose an increasing threat and challenge to individuals and businesses (APCC, 2020). For example, it is estimated that the cost of cyber-crime in the UK is around £27 billion a year (APCC, 2020). Cyber-crimes are a threat that is expected to evolve rapidly over the next 5–10 years with emerging methods such as deepfake and infrastructure hacking (APCC, 2020) and

the current investment in technology in UK policing is dwarfed by the cost of cyber-crime to the UK. Hence, strengthening the UK police force's ability to predict, investigate and be more responsive to such crimes is a crucial consideration for the National Policing Digital Strategy in the UK (APCC, 2020).

A content analysis conducted by Paoli et al. (2017) evaluated the term "serious crime" in 93 EU policy documents published in the period 1995–2013 and 104 academic articles published from 2004 to 2013. From the 93 EU policy documents analysed, only four defined serious crime, with Directive 2005/60/EC (Art. 5) providing the most structured and comprehensive definition:

> *…(a) terrorism offences, including membership in a terrorist group, criminal activities for the purpose of financing terrorism and inciting; (b) drug trafficking offences; (c) "the activities of criminal organisations;" (d) "fraud, at least serious;" (e) "corruption;" and (f) "all offences which are punishable by deprivation of liberty or a detention order for a maximum of more than 1 year... (Paoli et al., 2017, p.275, citing the European Parliament and Council, 2005, p. 21)*

A United Nations Conference in 2012 set out the concept and definition of serious crime as follows:

> *[Serious crime is] defined in article 2, subparagraph (b), of the Organized Crime Convention as meaning "conduct constituting an offence punishable by a maximum deprivation of liberty of at least four years or a more serious penalty". (United Nations Office on Drugs and Crime [UNODC], 2012, p.2)*

These definitions of serious crime do not require motivation, or name specific offences; rather, they use the sentencing penalty as the means of categorising serious crimes. In contrast, other sources, including some police forces, categorise serious crime by offence type, as shown in Table 1.1.

Table 1.1 presents 23 types of serious crime identified by three police forces (Dubai Police, 2023; The Metropolitan Police, 2023; New York Police, 2023), two sources in the literature (Paoli et al., 2016; Winchester, 2020), and one policy source (Europol, 2023).

Table 1.1. Types of serious crime in the literature.

	Arson	Assault/ Aggravated assault	Bribery/ Corruption	Border vulnerabilities	Burglary	Criminal damage	Cybercrime	Currency counterfeiting	Drug offences	Fraud and Forgery	Homicide/ Murder/ Manslaughter	Human trafficking
Paoli et al., (2016)	✓	✓			✓				✓		✓	
Winchester (2020)			✓	✓			✓		✓	✓		✓
Dubai Police (2023)		✓			✓				✓		✓	✓
Metropolitan Police (2023)		✓			✓	✓			✓	✓		
New York Police (2023)		✓			✓						✓	
Europol (2023)							✓	✓	✓	✓		✓

Table 1.1. Types of serious crime in the literature (cont'd).

	Illegal firearms	Illicit waste trafficking	Intellectual property crime	Kidnap/ Abduction	Money laundering	Motor vehicle theft	Organised immigration crime	Organised property crime	Rape and sexual offences	Robbery	Trafficking of endangered species
Paoli et al. (2016)				✓		✓		✓	✓	✓	
Winchester (2020)	✓			✓	✓		✓		✓		
Dubai Police (2023)						✓			✓	✓	
Metropolitan Police (2023)									✓	✓	
New York Police (2023)						✓			✓	✓	
Europol (2023)		✓	✓				✓	✓			✓

Serious crimes can also be categorised by what Winchester (2020, p.2) terms *"most harmful"*, progressing from cybercrimes to assault, burglaries and murder. Regardless of the type of crime, it is argued these crimes can be considered serious because of their impact on both the victims and on societies, affecting citizens and public services more than any other national security threat (Home Office, 2018; Winchester, 2020). However, *"The harm to the UK from serious and organised crime is difficult to measure as there is no single measure or separate criminal offence; it is a range of serious offences committed by organised offenders"* (Home Office, 2023, p.12). Referrals and the economic cost are often cited. For example, referrals to the NCA reporting child sexual abuse and exploitation increased 700% from 2012 to 2020; the number of county lines drug supply networks increased from 720 to approximately 2,000 between 2018 and 2019; 3.6 million fraud incidents were reported in England and Wales in 2018, with an overall cost of £190 billion; 6,993 potential modern slavery and human trafficking victims were identified in 2018, a 36% increase from 5,142 in 2017; 61,646 sexual crimes were committed against under 16s in 2018, a 9% increase from 56,346 in 2017 (National Audit Office, 2019; Winchester, 2020). Regarding fraud, data from the Crime Survey for England and Wales indicated that there was a 31% increase from 3,200,000 fraud incidents in the year ending March 2024 to 4,159,000 incidents in March 2025 (Office of National Statistics [ONS], 2025). However, as Winchester (2020, p.3) notes, *"Despite such figures, a large amount of SOC remains hidden or underreported"*.

Furthermore, the policing of SOC is complex, with responsibility often located within several different bodies. For example, in the UK there is no single entity or body that has overall charge in responding to SOC (Winchester, 2020). A National Audit Office report that examined the government's strategic response to SOC indicated that there are more than 100 government and law enforcement agencies (including policing) and organisations that tackle SOC in the UK (National Audit Office, 2019). Table 1.2 illustrates the range of organisations with responsibility for various aspects of security and law enforcement.

Table 1.2. Organisations with responsibility for addressing aspects of serious and organised crime in the UK.

Law enforcement	Policy and legislation	Justice	Intelligence and security	Other
National Crime Agency	Home Office	Ministry of Justice	Ministry of Defence	Department for Business. Energy and Industrial Strategy
43 territorial police forces	Cabinet Office	Her Majesty's Prison and Probation Service	UK Armed Forces	Financial Conduct Authority
9 regional organised crime units (ROCUs)	HM Treasury	The Attorney General's Office	National Ballistics Intelligence Service	Local authorities
British Transport Police	Northern Ireland Executive	Crown Prosecution Service	National Prisons Intelligence Coordination Centre	Ministry of Housing, Communities and Local Government
Civil Nuclear Constabulary	Scottish Government	Youth Justice Board	Government Agency Intelligence Networks	Citizens Advice
Border Force	Welsh Government	Victim Support	National Fraud Intelligence Bureau	Department for Environment, Food and Rural Affairs
Immigration Enforcement			Action Fraud	Environment Agency
HM Revenue and Customs			National Cyber Security Centre	Department of Health and Social Care
Police Scotland			Security and intelligence agencies	Public Health England
Police Service of Northern Ireland			Centre of the Protection of National Infrastructure	Department for Digital, Culture, Media and Sport
Serious Fraud Office			Stabilisation Unit	Department for Education
College of Policing			Maritime and Coastguard Agency	Intellectual Property Office
			Marine Management Organisation	Department for Transport
				Early Intervention Foundation
				Department for Work and Pensions
				Local Enterprise Partnerships
				Foreign and Commonwealth Office
				Department for International Development

(National Audit Office, 2019, p.53)

1.1.3 Current applications of big data in policing

Big data is viewed as a tool that is useful in supporting policing activities, including assisting in criminal investigations, crime prediction, mass surveillance and DNA databases (Neiva, Granja and Machado, 2022). Examples of big data applications in policing at a national and international level are presented in the following paragraphs.

At the national level among countries, for example, French police forces set up a strategy after the terrorist attacks in 2015 that aimed to collate online data in a centralised database in an easily sharable format, supported by big data technologies to assist in policing activities such as crime prediction and mass surveillance (Neiva, Granja and Machado, 2022). The Canadian police have also moved towards data and intelligence-driven approaches that include storing and using big data (O'Connor et al., 2022), although no examples of these uses and applications are provided. In an interview with Lieutenant General Dhahi Khalfan Tamim, Deputy Chief of Police and Public Security in Dubai, it was confirmed that the Dubai Police is using AI with facial recognition to identify wanted suspects (Tamim, 2024).

The Netherlands Police are using big data in frontline policing, criminal investigations and intelligence (Schuilenburg and Soudijn, 2023). In frontline policing, big data is being employed for predictive policing, with police surveillance teams deployed based on a system known as CAS that determines the risk of crime occurrence (Schuilenburg and Soudijn, 2023). Data analysts and social network specialists are also working on financial investigations and economic crimes (Schuilenburg and Soudijn, 2023). The use of big data in criminal investigations to date represents different levels of complexity, from simple tools to the implementation and development of a new digital infrastructure aimed at investigating complex digital crimes (Schuilenburg and Soudijn, 2023). In intelligence, the activities undertaken as part of frontline policing and criminal investigations generate new data that are recorded in different systems (Schuilenburg and Soudijn, 2023). This created the need for a Business Intelligence system to collect and analyse relevant data and convert it into intelligence products that can contribute to criminal investigations and policing activities (Schuilenburg and Soudijn, 2023). To use big data in policing intelligence, the Netherlands Police have recruited Business Intelligence Operations staff with experience in data warehousing, reporting, analytics, text and data mining to address different types of crime, such as cybercrime, organised crime on the dark web and terrorism (Schuilenburg and Soudijn, 2023).

Europol and Eurojust collaborated in setting up the Europol Information System, which contains data from various police forces within the EU (Schuilenburg and Soudijn, 2023). Europol recognises big data as a tool that is useful to advance the effectiveness of policing activities (Neiva, Granja and Machado, 2022). Another example can be seen in the Prum System, which is a European network that enables the automated exchange

of fingerprints, DNA profiles and motor vehicle information (Schuilenburg and Soudijn, 2023). DNA databases are considered one of three big data applications in the field of policing, in addition to crime prediction and mass surveillance (Schuilenburg and Soudijn, 2023). Finally, the Egmont Group of Financial Intelligence Units (FIUs) is an international initiative that assists in intelligence sharing among various national units (Schuilenburg and Soudijn, 2023).

1.1.4 Serious and organised crime strategies in the United Kingdom

It is important to note that this thesis did not purposely concentrate on the UK context, but the volume and relevance of UK-based studies, strategies and reports made it a logical and important focal point to explore areas relevant to the scope of this thesis.

SOC strategies in the UK have been updated every five years since 2013 by the Home Office, the most recent being published in 2023. The UK also has a National Policing Digital Strategy 2020–2030. The overarching aim of these strategies is to combat, investigate and prevent SOC due to the growing threat it presents (Home Office, 2013, 2018, 2023): *"Our mission is to reduce serious and organised crime in the UK, using the full reach and power of our intelligence and law enforcement agencies in partnership with the private sector and communities"* (Home Office, 2023). The purpose of exploring these strategies in this chapter is to understand the current position in the UK regarding the usefulness and/or use of big data in tackling SOC.

Serious and Organised Crime Strategy 2013

The 2013 SOC strategy aimed to substantially reduce the level of SOC affecting the UK through four steps (Home Office, 2013): pursue, prevent, protect and prepare. The first step, "pursue", entailed the NCA developing a reliable intelligence picture of organised crimes and collaborating with police forces and all concerned entities to coordinate and support their response (Home Office, 2013). The second step, "prevent", comprised preventative programmes, including improved education and communications about organised crimes, local coordination with troubled families and gangs and a wider use of interventions through Serious Crime Prevention Orders (Home Office, 2013). These first two steps, pursue and prevent, aimed to reduce the threat, whereas the following two steps aimed to reduce vulnerability (Home Office, 2013).

The third step was "protect", which began by changing the roles and responsibilities at the border and enhancing immigration enforcement by building new border capabilities (Home Office, 2013). Also, this step aimed to increase efforts to reduce fraud perpetrated on government procurement and share threat reports of fraud cybercrimes impacting the private sector (Home Office, 2013). Finally, the fourth step,

"prepare", sought to continue to develop emergency service interoperability, enabling a better response to terrorism and different crimes (Home Office, 2013). In addition, a new unit would be established to coordinate a national response to major cyber-attacks and cybercrimes (Home Office, 2013).

In terms of data and data analysis, the strategy aimed to develop the capabilities to detect and investigate SOC (Home Office, 2013). For example, Chief Constables and Police and Crime Commissioners led programmes to increase capabilities in collecting and analysing intelligence (Home Office, 2013). Moreover, in relation to data analysis, it was stated that *"The collection and analysis of bulk data are essential to the investigation and disruption of serious and organised crimes"* (Home Office, 2013, p.32). Hence, it was a priority for the NCA to establish *"…bulk data processing"* capabilities that could analyse lawfully obtained data from various sources in different areas, such as economic crimes and cybercrimes (Home Office, 2013, p.32). Most data were derived from government agencies, as well as some from the private sector (Home Office, 2013). Notably, the strategy referred to large data sets, or what is now referred to as big data, using the term "bulk data", clearly stating the essential role of data in investigating and disrupting serious crimes.

Serious and Organised Crime Strategy 2018

A review conducted by the NAO in 2017 found that the 2013 SOC strategy did not effectively address the increasing complexity and scale of SOC (National Audit Office, 2019). It was estimated that the government and law enforcement bodies spent around £2.9 billion tackling SOC but there were significant failings with the 2013 strategy (National Audit Office, 2019). It was established that the government's understanding of SOC was inconsistent and underdeveloped in some parts. For instance, in terms of the four "P" work standards, *"…work under the Pursue strand of the strategy dominated the efforts of government and law enforcement bodies, and work under the Prevent, Protect, and Prepare strands of the strategy needed improvement"* (National Audit Office, 2019, p.9). In addition, organisations' efforts to tackle SOC were disjointed and uncoordinated, law enforcement efforts were duplicated and the method used to prioritise threats needed improvement (National Audit Office, 2019). Also, the review indicated that there were gaps in the capabilities of law enforcement to respond to the complexity and evolving nature of SOC (National Audit Office, 2019), although these were not explicitly explained or demonstrated. Therefore, the 2018 SOC strategy aimed to address these shortcomings.

As has been noted, SOC is increasing in volume and complexity (Home Office, 2018). Despite some progress, the scale of the challenge posed by SOC is significant and the new strategy's approach was revised (Home Office, 2018). The new strategy aimed to protect citizens and the prosperity of the UK by leaving

no safe space for criminals to operate within the UK, overseas, online or offline (Home Office, 2018). The 2018 SOC strategy had four objectives to achieve its intended aims:

1. *Relentless disruption and targeted action against the highest harm serious and organised criminals and networks.*
2. *Building the highest levels of defence and resilience in vulnerable people, communities, businesses and systems.*
3. *Stopping the problem at source, identifying and supporting those at risk of engaging in criminality.*
4. *Establishing a single, whole-system approach.*
   (Home Office, 2018, pp.3–4)

This strategy aimed to provide a framework that outlined how to mobilise the capabilities of police forces, security and intelligence agencies to disrupt SOC (Home Office, 2018). It also aimed to equip the government, the private sector, communities and individual citizens to align their efforts to eliminate the harms of SOC (Home Office, 2018). In addition, from 2013, the government introduced robust legislation that aimed to ensure policing agencies would have the required powers to disrupt SOC. These included the Serious Crime Act 2015, the Modern Slavery Act 2015 and the Criminal Finances Act 2017 (Home Office, 2018). In relation to data, the UK government introduced the Data Protection Act 2018 and the GDPR, which provided additional insights into the laws governing data (Home Office, 2018). These regulations aimed to provide transparency to the public, stating that their data would be protected and used lawfully by the government and in a proportionate way based on the threat posed (Home Office, 2018).

Regarding technology, the 2018 SOC strategy aimed to harness the latest leading technologies in data analytics, biometrics, and behavioural and social sciences to stay ahead of threats and keep pace with the fast rate of change in crime (Home Office, 2018). The strategy outlined its endorsement of the potential role of big data, stating *"We will support initiatives that seek to explore the ethics of using artificial intelligence in the exploitation and interpretation of big data"* (Home Office, 2018, p.32). This included using innovative detection technologies and algorithms to detect concealed weapons and employed analytical tools that could alert police agencies to patterns in communications that might indicate sexual exploitation, considered one of the most severe categories of SOC (Home Office, 2018). The quote given above was the only direct reference to big data in this strategy. However, the role of data and data analysis in combating serious crimes was frequently referred to in the 2018 SOC strategy, which clearly stated that it would put data and intelligence at the heart of its approach because of the potential in enabling government agencies, especially the NCA, to effectively penetrate criminal networks (Home Office,2018). As the strategy stated,

"The ability to harness data is vital both to understand and disrupt serious and organised crime effectively" (Home Office, 2018, p.25).

In addition, the strategy aimed to sharpen and deepen the capabilities of specialists to combat cybercrimes, along with other online crimes (Home Office, 2018). Finally, similar to the 2013 SOC strategy, this version used the term "bulk data", perhaps referencing the concept of big data. The Investigatory Powers Act 2016 gave agencies the ability to intercept communications in specific circumstances, enabling an authority framework to examine *"…bulk data"* (Home Office, 2018, p.25).

Serious and Organised Crime Strategy 2023

The latest SOC strategy confirmed that crimes and criminal networks are increasing in their sophistication and depend greatly upon modern technology, such as the internet, to operate across borders (Home Office, 2023). Most SOC directly affecting the UK has an international element; for example, the criminals could be foreign nationals, or the criminal activity could be conducted overseas by British nationals. This makes the investigation of SOC challenging (Home Office, 2023). The 2023 SOC strategy is founded on the growing threats of SOC, with a mission to reduce incidence in the UK through use of the full scope and power of the intelligence and policing forces in partnership with the private sector and with communities (Home Office, 2023).

The 2023 SOC strategy covers five sectors: in-country, UK border, international, technology and capabilities, and multi-agency response (Home Office, 2023). The first, in-country line of action, aims to keep the public safe by disrupting and dismantling organised crime groups operating in and against the UK through building resilience in local communities, designing out crime and raising online barriers (Home Office, 2023). The second, the UK border line of action, aims to strengthen the border by identifying and intercepting known and unknown individuals and goods entering or leaving the country illegally (Home Office, 2023). The third, the international line of action, is overseas and aims to disrupt international organised crime groups through intelligence-led operations and raise barriers to reduce the harm reaching the UK (Home Office, 2023). The fourth line of action is to ensure the availability of the latest technology, best intelligence, data collection, analysis and investigative capabilities to identify and disrupt SOC (Home Office, 2023). The fourth objective aims to ensure better outcomes from the first three lines of action together, seeking to conduct effective intelligence, data collection and analysis which can lead to better disruption of and protection against SOC (Home Office, 2023). The fifth and final line of action is a multi-agency response to ensure that all public and private sector partners work together as effectively as possible and have the required capacity and skills to do so (Home Office, 2023). The fifth objective aims to enhance

the delivery of the outcomes of the first four lines of action, making the multi-agency response more coordinated and effective and providing maximum effect and value for money (Home Office, 2023).

Regarding big data, unlike the 2013 and 2018 SOC strategies in which bulk/big data was mentioned, there is no direct mention of big data in the 2023 SOC strategy. However, data and data analysis are among the core elements of this strategy. The fourth objective states, "*Our fourth line of action is ensuring the best intelligence and data collection, analysis and investigative capabilities are in place to identify and disrupt organised criminals*" (Home Office, 2023, p.40). Border Force, Home Office Intelligence, NCA and other agencies are expected to work together, using intelligence and data to identify, target and disrupt organised crime groups that seek to exploit the UK border (Home Office, 2023). This requires securing and making the best possible use of data and linking it with advanced intelligence and detection capabilities to strengthen and protect the UK border against, for example, illicit commodities and class A drugs (Home Office, 2023). As part of this, investment has been made in an advanced next-generation data analytics system (CERBERUS) that helps to identify high threat movements of goods and people (Home Office, 2023). The strategy outlines several projects that speak to the potential usefulness of big data for the investigation of serious crime in the UK. This includes the following: as part of the UK's exit from the EU, the UK is supposed to start receiving data for the first time (from 2025) on the movement of EU goods, which the government argues will "*...be a key element in powering data analytics*" (Home Office, 2023, p.34); data sharing agreements with the EU, including the Prum data system, which is an international biometric database through which DNA and fingerprint data can be exchanged (Home Office, 2023); the I-LEAP project, which will deliver connectivity to Interpol's databases, providing real-time access for UK law enforcement agencies to subjects and objects of interest (Home Office, 2023). The UK government's focus is on responding to the challenges created by technology by ensuring the best intelligence and data collection, analysis and sharing capabilities are available to address them (Home Office, 2023).

Linked to the national UK strategy is the National Policing Digital Strategy 2020–2030, put forward by the APCC. This has a stronger focus on the data and technology aspect (APCC, 2020). It aims to create a digital transformation in policing and improvements in data and technology, developing the skills of the people who lead, manage and use them (APCC, 2020). The strategy recognises the rapid growth of data, together with an increasing and complex need to analyse large datasets to discover trends and patterns, as well as to use AI to support decision making (APCC, 2020). The APCC (2020, p.3) argues that "*The potential benefits are immense*" from analysing these large datasets and data-driven insights have the potential to be a *"...force multiplier*". The APCC (2020) emphasises the positive potential of big data and how it is considered a strategic asset for UK policing. However, it also highlights that it is challenging to foresee how the next decade of policing will unfold (APCC, 2020). Despite this, the APCC recommended an assessment of digital

trends and behaviours around data and new technologies that will change the nature and volume of demand, affecting the ability of UK policing to respond to future demand and need (APCC, 2020). In addition, the National Policing Digital Strategy identifies key digital trends that raise some pressing questions about the direction of policing in the UK, with big data being one of these trends (APCC, 2020). It questions the bots, algorithms, automation and big data will be managed and harnessed and their potential influence in daily life and across UK police forces (APCC, 2020).

1.1.5 European Union security strategy and big data

Big data and AI are important in emerging strategies that aim to tackle SOC, such as the EU's Security Strategy 2020–2025 (Schuilenburg and Soudijn, 2023). The strategy argues that AI and big data should integrated into security policy, as both will be effective in fighting crimes (Schuilenburg and Soudijn, 2023). Since the EU's Security Strategy recognises the potential of big data in policing, especially serious crimes (Schuilenburg and Soudijn, 2023), the strategy was further examined to explore its links to big data. In the section on tackling evolving threats, the strategy suggests that common approaches be taken to ensure that AI, big data and high-performance computing are integrated into security policies designed to be effective in both fighting crime and ensuring fundamental rights (European Commission, 2020). In this regard, the strategy contends that "*Artificial intelligence could act as a powerful tool to fight crime, creating enormous investigative capabilities by analysing large amounts of information and identifying patterns and anomalies*" (European Commission, 2020, p.12). Moreover, it offers examples of big data and AI applications in policing to assist in identifying online terrorist content, uncover suspicious transactions in selling dangerous products and help citizens in emergencies (European Commission, 2020). In terms of big data, electronic information and evidence are needed in about 85% of investigations of serious crimes (European Commission, 2020). The strategy also emphasises the importance of high levels of compliance with fundamental rights in effectively protecting citizens (European Commission, 2020).

Taken together, the prior sections have provided the background and context for this thesis by examining big data, policing and serious crimes. They also help set the focus of the thesis within the wider literature. The following section now turns to the research gap addressed by this thesis.

1.2 Research gap

Despite the growing interest in data-driven policing, there is still uncertainty about what big data means in the policing context and how far can it deliver practical value in serious crime investigations. A recent study highlighted that there is limited research on how police use big data applications and whether and to what extent the use of big data can be effective in policing (Schuilenburg and Soudijn, 2023). In addition, as police forces enter an era of big data policing, they must consider the quality of the data they collect and

use, as little is currently known about the quality of the data available (O'Connor et al., 2022). Moreover, the big data revolution raises complex ethical challenges for policing, resulting in legal and ethical constraints (Babuta, 2017). Taken together, this creates a problem for research and practice as big data is inconsistently conceptualised, with many claims made about its benefits at the same time as users face technical and ethical limitations. Thus, it becomes difficult to evaluate its usefulness. There is a need for a clearer understanding of what big data is and what is known about it and a careful assessment of the potentials, challenges and concerns regarding its use in policing. Therefore, this thesis explores (a) how big data can be understood and defined in the context of serious crime investigations in policing, (b) the potential advantages and disadvantages of its use and (c) any challenges and concerns that can affect implementation in practice. Also, the literature tends not to examine big data and serious crime together. Therefore, this thesis includes several studies that focus on each area independently and then synthesises them to help establish a basis for linking the two domains in this thesis.

Moreover, researching the two central domains aligns with the "Safe and Resilient City" strategic goal established by the Dubai Police, which aims to protect individuals, reduce crime rates and secure society (Dubai Police, 2025). In consideration of this, and the researcher's professional background, the researcher was awarded a scholarship to undertake this research based on the recognised need within the profession to better understand big data as an emerging field with a view to advancing serious crime investigations. The scholarship from Dubai Police supported the research but did not interfere with or influence the findings, as its interest was in supporting the generation of knowledge that could be useful to professionals in the field.

1.3 Rationale and significance

The importance of researching these domains lies in the serious consequences for individuals, communities and public trust of the application of big data in policing. If big data tools are used to inform policing decisions, the quality of the inputs and outputs is critical as the decisions made based on them will affect individuals and communities. In addition, police forces may face various types of constraints and competing demands, which could increase the appeal of approaches that promise higher efficiency. Big data may support and advance policing operations, but it may also introduce unanticipated risks if there is a failure to evaluate its use with care. Understanding both the potential value and possible challenges of big data is therefore essential for realistic decision making about its adoption.

1.4 Research aims and questions

The overarching aim of this thesis is to explore what is known about big data and examine its potential usefulness for policing in serious crime investigations.

The research began with a scoping review conducted in two stages. First, scoping review question 1 (SRQ1), "What is known about the usefulness of big data across multi-disciplinary fields?", explored the literature on the usefulness of big data across multiple disciplines to establish a broad understanding of its potential value and possible limitations. Second, a more focused scoping review was conducted to examine the fields of big data and policing through scoping review question 2 (SRQ2), "What is known about big data and can it be useful in serious crime investigations?", which formed the central question of this thesis. The two-stage scoping review supported clear progression from broad evidence mapping to a focused synthesis on the application of big data in policing.

In line with the above, the thesis addresses the following aims:

1. To develop a comprehensive understanding of the concept of big data, including its definitions and characteristics, to form a theoretical basis as a foundation for further research within the context of serious crimes.
2. To explore the potential usefulness of big data in serious crime investigations by identifying the advantages, disadvantages challenges and availability of artificial intelligence (AI) tools in relation to big data.
3. To develop a definition of big data suitable for use across policing internationally.

1.5 Scope and boundaries of the thesis

The scope of this thesis encompasses two main fields: big data in relation to policing. Rather than developing or testing a specific technological system in this context, it examines the usefulness of big data in terms of how it can support serious crime investigations, for example in terms of decision making, and crime prevention and detection, while also considering potential technical and practical challenges, along with concerns such as bias and privacy. This thesis does not aim to produce a statistical estimation of the effectiveness of crime reduction or detection tools but to develop a structured understanding of what is known about the topic under study and explore informed perspectives, thus helping develop a conceptual understanding and identify the usefulness, challenges and conditions related to the applications of big data in policing.

1.6 Methodological approach

The study adopted a qualitative research design, first undertaking a scoping review to map the breadth of existing research in the literature. Following this, in an empirical phase, semi-structured interviews were conducted with 17 participants divided into two groups: nine police officers and eight specialists with big data expertise. The participants were selected and recruited through a purposive sampling strategy based on

their knowledge and experience. In addition, the study drew on the support of one gatekeeper to assist in gaining access to six of these interviewees. The interview schedule was carefully designed to support open-ended responses, avoid leading questions and allow the participants to express their views freely. The data were then analysed through reflexive thematic analysis, following the six-phase approach developed by Braun and Clarke (2006, 2020), to identify patterns in meanings across the participants' accounts.

1.7 Theoretical and conceptual positioning

The analysis was informed by a critical realist framework, supported by a relativist perspective as an epistemological viewpoint. This approach was deemed appropriate because it recognises that certain structures and conditions shape what is possible in practice, while also acknowledging that knowledge about these conditions is partial and shaped by context and role (Sayer, 2000; Bhaskar, 2008; Fletcher, 2016).

In this thesis, the participants' accounts are treated as valuable evidence of how the usefulness of big data can be understood (perceived advantages), as well as identifying where certain conditions or challenges need to be mitigated to attain the potential advantages, thereby enhancing usefulness.

1.8 Contributions to knowledge

In this section, I briefly outline the contributions made by this thesis to knowledge. These are addressed in greater detail in the discussion and conclusion (Chapters 5 and 6). This thesis makes three contributions to the field in three areas: theoretical/conceptual, empirical and practical.

Theoretically, the thesis advances our understanding of big data by setting out the debates in the field and clarifying the definitions and characteristics to be operationalised in the context of policing. By establishing and synthesising the foundations drawn from the literature on big data and the findings in relation to serious crimes, this thesis proposes a stable definition of big data in serious crime investigations that can be used by policing internationally.

Empirically, this thesis contributes by providing an evidentiary basis for arguing that big data has the potential to be useful in serious crime investigations, specifically in terms of identifying strategic and operational advantages. In addition, the study identifies central challenges (cultural/conceptual, financial and technical, and related to human resources) and concerns (privacy and bias) that need to be addressed and overcome to ensure big data is utilised successfully in policing. These empirical insights are based on data drawn from two groups of participants – policing specialists and big data specialists – who are often required to work together in practice, thereby grounding the discussion in the perspectives of professionals in the field.

Practically, this thesis proposes a hierarchical framework (see Figure 5.2) that translates the findings into a practical guide for policing by presenting the core elements required to attain the strategic and operational advantages. This figure supports a responsible and effective analysis of big data, particularly in relation to protecting the privacy of individuals and avoiding biased outcomes that could affect decision making.

1.9 Thesis structure

Following this introductory chapter, the thesis is structured in five further chapters, moving from the theoretical groundwork drawing on the literature, through the empirical study, to a focused analysis of the empirical findings and then interpretation.

Chapter 2 presents the literature reviewed using a structured scoping review approach, first mapping big data across disciplines in response to SRQ1 and then developing a comprehensive understanding of big data and examining its usefulness in serious crime investigations in the context of policing in initially addressing SRQ2.

Chapter 3 sets out the methodological approach to the empirical study, including the development of the central research question, the theoretical and conceptual frameworks, sampling, data collection and analysis, the researcher's positionality and ethical considerations.

Chapter 4 presents the findings from the reflexive thematic analysis of the data collected through semi-structured interviews, which resulted in the development of six key themes.

Chapter 5 critically discusses the empirical findings in relation to the literature and interprets them according to a critical realist framework, and outlines the empirical, theoretical, and practical contributions to knowledge.

Finally, *Chapter 6* concludes by summarising the main conclusions as well as making theoretical and practical recommendations.

In summary, this chapter has introduced and positioned the thesis within the wider context and provided a rationale for examining what is currently known about big data and its potential value for policing. Also, it has clarified the central research problem and main gaps in existing knowledge, as well as presenting the central research question (SRQ2) and the aims guiding the thesis. The chapter has also briefly outlined the research design and set out structure of the thesis. The next chapter provides a review of the relevant literature, building the foundation for the empirical analysis that follows.

Chapter 2. Scoping Review

2.1 Introduction

This chapter reviews the existing literature and establishes the evidence base for the thesis by mapping what is known about big data and its usefulness. In doing so, the focus narrows from addressing broad applications to the specific context of policing and serious crime investigations. The chapter is structured in three sections. Section 1 presents the scoping review methodology and addresses SRQ1, which examines the usefulness of big data across multiple disciplines. This helps to identify the potentials and limitations of the application of big data beyond a single field.

Section 2 then focuses on existing knowledge and understandings of big data in line with SRQ2 and Research Aim 1, developing a comprehensive account of the key definitions, characteristics and developments that shape how big data can be understood and used in practice. Building on this foundation, Section 3 addresses SRQ2 and Research Aim 2 by reviewing the usefulness of big data in serious crime investigations, paying particular attention to the reported advantages, challenges and concerns regarding its use.

Dividing the chapter in this way strengthens the review, as it first establishes that big data has recognised potential across different domains, then clarifies what big data means and how it has expanded and developed and finally applies this understanding to the policing context. This staged structure also supports a critical review, as it enables comparison between broader claims about big data and the specific constraints and expectations that shape its use in policing.

2.2 Scoping review methodology

This section presents the scoping review methodology used to explore the knowledge base concerning big data, policing and big data, and the potential usefulness of big data in serious crime investigations. A scoping review was used as it can address topics where there is a paucity of published research (Arksey and O'Malley, 2005; Peterson et al., 2017) related to the specific research area. It aimed to identify existing evidence and gaps, thus highlighting areas requiring further research (Arksey and O'Malley, 2005; Mak, 2022; Rodger, Admani and Thomas, 2024). Arksey and O'Malley (2005, p.21) define a scoping review as an exercise that *"aim[s] to map rapidly the key concepts underpinning a research area and the main sources and types of evidence available, [which] can be undertaken as stand-alone projects in their own right, especially where an area is complex or has not been reviewed comprehensively before"*. This approach enabled the identification and mapping of existing and emerging literature on a given topic, as well determining if there were any gaps in the literature (Mak, 2022).

In this thesis, the scoping review aimed to identify the current knowledge base and knowledge gaps concerning the usefulness of big data in serious crime investigations. It was also designed to include any relevant knowledge from other fields in which big data may be used in similar circumstances, for example in finance and healthcare. An advantage of scoping reviews is that they provide greater flexibility than traditional systematic reviews by including diverse relevant literature (Peterson et al., 2017), which is important in a rapidly developing and under-researched field, such as the use of big data in serious crime investigations.

The framework for conducting the scoping review in this thesis was developed by Arksey and O'Malley (2005), which is the most cited protocol in the literature (Cacchione, 2016) and was employed by Mak (2022) and Rodger, Admani and Thomas (2024) in their studies. Arksey and O'Malley (2005, p.22) suggest that researchers follow five stages, as follows:

Stage 1: Identifying the research question

The first stage identified the research question(s) to guide the building of the search strategy (Mak, 2022; Rodger, Admani and Thomas, 2024; Arksey and O'Malley, 2005). In this study, the scoping research questions were as follows:

SRQ1: What is known about the usefulness of big data across multi-disciplinary fields?

SRQ2: What is known about big data and can it be useful in serious crime investigations?

Stage 2: Identifying relevant studies

The second stage entailed identifying relevant studies through a search strategy (Arksey and O'Malley, 2005; Mak, 2022; Rodger, Admani and Thomas, 2024). The purpose of scoping the field was to be as comprehensive as possible in identifying studies to answer the scoping research questions (Arksey and O'Malley, 2005). In this study, the parameters of the search were as follows:

(i) Databases and websites searched

- Liverpool John Moores University online library (Discover)
- Google Scholar
- Research Gate, Wiley and Taylor and Francis digital libraries (incorporating a range of papers covering both the fields of policing and big data)
- Policing journals: *Policing: A Journal of Policy and Practice*, *The Police Journal: Theory, Practice, and Principles*, *Policing and Society: An International Journal of Research and Policy*, *International Journal of Police Science*

- European Union (EU) commission for security strategies
- UK government website for Serious and Organized Crime Strategies and Data Ethics frameworks
- Dubai Police, London Metropolitan Police, New York City Police websites (for serious crimes categories)

(ii) Keywords (sorted in alphabetical order)

- "big data" and "advantages, disadvantages, concerns"
- "big data" and "ethics" "legislations"
- "big data and policing"
- "big data" and "healthcare" "finance" "business"
- "big data investigations"
- "big data" and "characteristics, definitions, developments"
- "healthcare data breach annual report"
- "big data ethical challenges in healthcare"
- "EU security strategies police and big data"
- "potential of big data"
- "serious crimes strategies"
- "serious crimes/investigations and police/policing"
- "volume of worldwide data in 2024"
- "what is big data"

(iii) Eligibility criteria for the scoping review

Inclusion:

- Books, academic articles, studies published in peer-reviewed journals, policies, strategies, and reports focusing on big data and/or policing and across multidisciplinary fields
- Publications from January 1997 to September 2025. The start date of 1997 was chosen because it is when the concept of big data was first introduced by National Aeronautics and Space Administration (NASA) researchers (Ylijoki and Porras, 2016; Jurkiewicz, 2018); September 2025 was selected as the end date to ensure that the scoping review would be as up to date as possible.
- No geographical criteria were applied and therefore no studies were excluded based on location.

Exclusion:

- Studies not published in English.
- Any publications before 1997.
- Non-scholarly sources, such as news articles, letters and opinion pieces.
- Conference abstracts without full published papers.

Stage 3: Study selection

This stage established a screening protocol and eligibility criteria for the literature identified by the search strategy. Developing a protocol assisted in clarifying the types of studies that will be eligible for inclusion (Rodger, Admani and Thomas, 2024), ensuring relevant literature was included and removing irrelevant literature to ensure an effective and efficient use of time in Stages 3 and 4 and provide a transparent and consistent approach in addressing the significant body of literature identified via the search strategy. In the screening process for this study, the title of each publication identified in the literature search was first reviewed to determine those relevant to either or both of the scoping review questions (SRQ1: What is known about the usefulness of big data across multi-disciplinary fields? SRQ2: What is known about big data and can it be useful in serious crime investigations?).

For scoping reviews, authors need to report clearly the process undertaken, as well as any limitations of their approach, to ensure transparency (Pham et al., 2014; Rodger, Admani and Thomas, 2024). To maintain the integrity of the scoping review process, it is necessary to set out the sifting process. As Peters et al. (2021, p.3, citing Peters et al. [2020] and Levac, Colquhoun and O'Brien [2010]) argue, *"…scoping reviews may be iterative and flexible and whilst any deviations from protocol should be transparently reported, adjustments to the questions, inclusion/exclusion criteria and search may be made during the conduct of the review"*. For the sake of clarity, this thesis evolved from an implicit literature review previously conducted, in which studies were screened and excluded, but this was not a systematic process recorded prior to undertaking the more focused scoping review. Since Arksey and O'Malley's (2005) framework is iterative (Peters et al., 2021) and was developed before the PRISMA system of numerical reporting (Moher et al., 2009), the search and screening method in this thesis is transparently reported. This chapter does not present the overall number of studies that were read, evaluated and included/excluded across the research as a whole. However, the identification process in Stage 2 clearly presents the databases and websites, keywords and eligibility criteria applied, enabling the search to be reproduced by other researchers.

In support of flexibility in conducting and reporting the sifting process, scoping review reporting has varied across studies (Tricco et al., 2016; Peters et al., 2021; Viitanen et al., 2022). In Viitanen et al. (2022, p.143), the authors reported, "*Our scoping review did not strictly follow the PRISMA protocol [and] we did not*

produce a PRISMA flow diagram, but instead explained the process…". Similarly, the search strategy in the scoping review for this thesis is presented in Stage 2. Moreover, Tricco et al. (2016) examined 494 scoping review studies and found that less than half (47%) used study flow figures but remained methodologically robust by being transparent about the process and acknowledging this as a possible limitation. Thus, scoping reviews allow flexibility and may not always require an extensive sifting process (Tricco et al., 2016; Viitanen et al., 2022). This thesis acknowledges the potential limitation of not recording screening counts and thus not providing a detailed study selection record; however, the search strategy and eligibility criteria are reported transparently.

Stage 4: Charting the data

In the fourth stage, the data derived from the literature were categorised based on the scoping review questions and the themes beginning to emerge from the literature base under review (Arksey and O'Malley, 2005; Mak, 2022). Table A1 (Appendix A) summarises the 84 sources that shaped the empirical study, presenting the author(s), year of publication, country and context, aim and focus, methods and data sources, and key points relevant to both SRQ1 and SRQ2. These studies shaped the main themes that informed the development of the interview protocol and provided a foundation for the analysis, represented in the recurring themes in the findings and revisited in the discussion.

Stage 5: Collating, summarising and reporting the results

In line with Arksey and O'Malley's (2005) framework, the fifth stage involved collating, summarising and reporting the results of the literature reviewed, moving beyond a simple description of the studies included to an interpretive synthesis of the evidence base. In this study, 84 sources from the databases searched in stage 2 met the inclusion criteria for the scoping review. The study designs and document types were qualitative (n = 9), quantitative (n = 8), mixed methods (n = 8), technical and computer science (n = 8), legal-doctrinal (n = 8), government policy, strategy, or official documents (n = 17), reviews, scoping, or mapping studies (n = 14), and conceptual or theoretical works (including handbooks and industry reports; n = 12). The included sources form a methodologically varied set of studies which reflects on the multi-disciplinary nature of big data and serious crime investigations. These variations indicated that much of the existing work in these fields has developed through conceptual, technical, policy, and legal contributions alongside a notable body of empirical research.

Across the sources included (N = 84), the geographical extent of the evidence base included studies adopting a global or international perspective, presenting cross-jurisdictional discussions rather than focusing on a single country (n = 23), studies focused on the UK (England and Wales; n = 15), and contributions examining the EU (n = 10), the United States (US; n = 10), India (n = 6), and Nigerian/African contexts (n

= *3*). Smaller clusters included contributions from the Netherlands (*n* = 2), the United Arab Emirates (UAE; *n* = 2), and Italy (*n* = 2). Other individual studies focused on Australia (*n* = 1), Brazil (*n* = 1), Canada (*n* = 1), Hong Kong (*n* = 1), Iraq (*n* = 1), Indonesia (*n* = 1), Norway (*n* = 1), Malaysia (*n* = 1), Finland (*n* = 1), Spain (*n* = 1). There was also a multi-country study (*n* = 1) that examined perspectives across the UK, Netherlands and Germany. Taken together, this distribution shows both concentration and diversity. While a considerable proportion of the literature concern a small number of jurisdictions, particularly the UK, US and EU, there is also a growing body of work from a range of other national contexts.

Of the 84 sources included, most focused directly on policing, crime, serious and organised crime and security analytics (*n* = 43), big data and information systems (*n* = 10), governance, security policy and public-sector data (*n* = 7), data protection, privacy and civil rights (*n* = 6), broader conceptual discussions of big data and security (*n* = 6), finance and fraud (*n* = 4), healthcare and health informatics (*n* = 3), consumer and managerial applications (*n* = 3), and humanitarian aid or disaster management (*n* = 2).

The final corpus comprised 84 sources that met the predefined inclusion criteria. Scoping reviews intend to map the breadth of the field by including a wide range of sources and study designs (Arksey and O'Malley, 2005), without necessarily restricting inclusion to a narrowed set of study types for example peer-reviewed literature (Mak, 2022). Consequently, the resulting set of sources across sectors, policing and serious crime being the core, encompassing different methodological approaches and study designs, provided sufficient conceptual and empirical richness to identify recurring themes, areas of agreement and differences. On this basis, the corpus was considered adequate to address the scoping review questions.

As with all scoping reviews, the findings reported here were shaped by the search strategy and inclusion criteria. While multiple databases and supplementary sources were searched, relevant work published in other or less accessible sources may not have been captured. In addition, the charting and classification of studies were conducted systematically, which could include interpretive judgments. Hence, the review should be perceived as a rigorous mapping of accessible literature rather than an exhaustive census of all possible sources. Nonetheless, the breadth and diversity of the sources included provide a robust foundation for thematic synthesis and helping shape the empirical study. It is also important to note that these sources do not represent all the references cited in this thesis. Additional methodological and theoretical studies were used to inform the research design and all sources, whether or not included in the charted dataset, are included in the list of references.

Following Arksey and O'Malley's (2005) framework, the categories and themes were derived through the processes in stages 4 and 5. In stage 4, the data were charted through the systematic extraction and

organisation of relevant information from the included studies. This stage enabled the identification of recurring concepts and patterns, which were grouped into categories. In stage 5, the charted data were collated and interpreted, with the related findings grouped into categories. These shaped the broad themes representing patterns within the literature. For clarity, within this thesis, the categories refer to the groupings of similar findings that were identified from the data extracted from the included studies, whereas the themes represent the broader patterns that were developed through the interpretation of the dataset and shaped by the categories. The results are presented by the overarching themes which are divided across three sections. Section 1 addresses SRQ1, exploring the expansion of big data and its usefulness across multi-disciplinary such as healthcare and finance. Section 2 responds to the first part of SRQ2, recounting the foundations of big data and definitions, characteristics and key developments. Finally, Section 3 examines specifically what is known about policing/serious crimes and big data to explore its potential usefulness, applications, challenges, and frameworks, addressing the second part of SRQ2.

2.3 Section 1: The usefulness of big data across multi-disciplinary fields

This section explores the literature in relation to SRQ1: "What is known about the usefulness of big data across multi-disciplinary fields?" Using the search and inclusion criteria set out above, studies were identified that speak to the potential usefulness of big data across fields such as finance and health. These fields were not chosen specifically but emerged through the literature search. The potential of big data and the challenges of privacy, security, bias, technicalities and legislation in relation to its use in different fields are also applicable to policing. Similarities were found across these fields, which shaped the formation of the following themes: the expansion of big data; the use of big data in healthcare; the use of big data in finance; challenges in utilising big data in both fields. To support transparency and coherence, an analytical summary is provided after each subsection to interpret the key points and highlight their relevance to the research.

2.3.1 The expansion of big data

The spread of the use of smartphones, social media and different technologies has led to an increase in the amount of data and created new data sources (Horita et al., 2017). The rapid development of the Internet, Internet of Things and cloud computing have led to a substantial growth in data in almost every industry, making big data an interesting area for academia, governments, public sector, policymakers and private industries around the world (Jin et al., 2015; Ylijoki and Porras, 2016).

The concept of big data can be interpreted in various ways, with different definitions having evolved based on its features and characteristics in particular fields (Chan and Moses, 2016; Babuta, 2017). Moreover, big data is considered an emerging discipline in academia and thus has encountered the lack of agreement

concerning its conceptualisation and definition common among researchers for a phenomenon in the early stages of its evolution (Mauro, Greco and Grimaldi, 2016). The volume, velocity, variety and veracity of data available to governments, businesses and people, together with developments in computing power, can and are changing societies in substantive ways (Brady, 2019). Big data is perceived to have great value and is already transforming the way people think, live and work (Jin et al., 2015; Abouelmehdi, Beni-Hessane and Khaloufi, 2018). As Babuta (2017, p.1) notes, *"We're sitting on absolutely monumental amounts of information collected from different sources"*. The increasing volume of information and datasets are being mined to form predictions in new ways and to find new insights into challenges facing societies (Richards and Kings, 2014).

Although the expansion of data in the last two centuries was significant, it cannot be compared with the growth of data in terms of size and variety being experienced at this current moment (Broeders et al., 2017). The unprecedented growth of data and the subsequent creation of large databases for use by businesses and governments is spawning a field of research studies aimed at understanding and discovering how to utilise these data efficiently (Brady, 2019). Indeed, research centres have been established in various universities around the world, such as the University of California, Columbia University, New York University, Tsinghua University, the Eindhoven University of Technology and the Chinese University of Hong Kong (Jin et al., 2015). In addition, several universities have launched undergraduate and postgraduate courses in data analytics to train data scientists and data engineers to work with big data (Jin et al., 2015).

Similarly, governments are developing ways to use big data. For example, the US administration launched a Big Data Research and Development Initiative in 2012 with an investment of over $200 million, involving six federal government agencies: the Department of Defense, the Defense Advanced Research Projects Agency, the Department of Energy, the National Institutes of Health, the National Science Foundation and the US Geological Survey (Jin et al., 2015). This initiative aims to develop infrastructure, tools and techniques to gain knowledge and value from big data in various areas: health, the environment, emergency response and disaster resiliency, manufacturing, robotics and smart systems, secure cyberspace, transportation and energy, education, and workforce development (Jin et al., 2015). An initiative like this indicates the diverse fields in which big data has potential. This is related to the first scoping review question, concerning the application of big data across multi-disciplinary fields; as Jin et al. (2015, p.63) suggest *"Big data has made a strong impact in almost every sector and industry today".*

According to Babuta (2017), big data can play a major role in transforming and developing different fields, such as business, healthcare and finance. Richards and King (2014, p.393) describe this phenomenon as a *"big data revolution"* affecting many human activities and influencing decisions, not least by making

predictions about human behaviour that has not yet happened in areas as diverse as shopping, education, policing, cybersecurity and terrorism prevention. It is argued that a leading organisation is one that can use big data effectively and benefit from it, as it is considered a competitive advantage (Surbakti, 2020). Big data is also considered a significant enabler that can generate value for public organisations and private companies, such as creating new business opportunities, advancing research and development and supporting decision making (Ylijoki and Porras, 2016).

Despite the failure of some big data projects and the slow progress of others, Surbakti (2020) contends that the potential and opportunities offered are promising and widespread. Big data is now considered one of the most promising – indeed, increasingly essential – technologies in the medical and financial services fields (Awrahman, Fatah and Hamaamin, 2022; Aderemi et al., 2024), among others. Considering its application in these fields offers the potential to gain a better understanding of the potential of big data and its usefulness for serious crime investigations.

2.3.2 The use of big data in healthcare

The medical field is one in which where big data is considered fundamental in terms of promising to make a positive change in healthcare (Abouelmehdi, Beni-Hessane and Khaloufi, 2018; Awrahman, Fatah and Hamaamin, 2022). The objective of generating big data is to attain information that can be analysed to generate insights useful in healthcare and for forecasting the future (Awrahman, Fatah and Hamaamin, 2022). The integration of diverse big healthcare datasets, along with big data technologies, is viewed as having the potential to improve patient outcomes, improve policymaking, predict outbreaks, avoid preventable diseases, decrease hospital costs and improve quality of life more generally (Abouelmehdi, Beni-Hessane and Khaloufi, 2018; Awrahman, Fatah and Hamaamin, 2022). The digitisation of healthcare data is leading to a "big data revolution" in this field, impacting biomedical signals, genomic data, sensing data, biomedical images and real time signals (e.g. electrocardiograms [ECGs], and pulse and blood pressure monitoring) (Awrahman, Fatah and Hamaamin, 2022).

There are also now numerous corporations in the big data healthcare space that provide tools for various purposes, such as Kafka, Sqoop, Apache Spark, Hadoop HDFS, the early warning score (EWS), Hive, and Spark SQL, enabling big data analytic services in the healthcare sector (Awrahman, Fatah and Hamaamin, 2022). This sector is particularly advanced in North America. For example, Flatiron Health in New York analyses data points from cancer patients to develop research and enhance patient care, supporting the work of oncologists, academics and researchers in advancing treatments for cancer patients (Awrahman, Fatah and Hamaamin, 2022). SCIO Health in Connecticut uses algorithms and integrated data to provide insights and detect gaps in healthcare that can worsen health outcomes and lead to increased healthcare costs.

Identifying these gaps can assist medical professionals in detecting at-risk patient groups and helping avoid complications (Awrahman, Fatah and Hamaamin, 2022). Hortonworks in California collates billions of records of pharmaceutical data that are used by pharmaceutical companies and researchers to advance more effective research for clinical trials and marketing (Awrahman, Fatah and Hamaamin, 2022). UNC Health Care in North California (UNCHC) has implemented a new system that allows clinicians to access and analyse unstructured patient data rapidly through natural language processing and provide insights enabling timely intervention, reducing re-admissions and providing safer care for high-risk patients (Abouelmehdi, Beni-Hessane and Khaloufi, 2018). Toronto Infant Hospital in Canada uses big data analytics to improve outcomes for newborns at risk of serious hospital infections (Abouelmehdi, Beni-Hessane and Khaloufi, 2018).

However, although big data analytics in healthcare is delivering benefits and has the potential to further improve healthcare outcomes, there are still barriers and challenges (Abouelmehdi, Beni-Hessane and Khaloufi, 2018) to the development and implementation of big data analytics in this sector. These are discussed in greater depth in the challenges section of this chapter (see 2.3.4).

The widespread adoption of big data by medical companies and hospitals suggests that there is ~~strong~~ potential for its application in other sectors, such as policing, as they share several valuable parallels. Both fields rely on sensitive and private data concerning individuals in the community, highlighting the importance of data quality, validity and privacy. Also, just as the medical field aims to forecast the future in terms of predicting outbreaks and detecting diseases, policing aims to predict and forecast crimes to prevent them, as well as detecting crimes, calling for further research into the use of big data in policing and specifically in serious crime investigations to determine its usefulness.

2.3.3 The use of big data in finance

The exceptional growth in the volume, velocity and variety of data and developments in technologies aiding data analytics have provided financial institutions with opportunities to harness information offering various advantages (Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024). Aderemi et al. (2024) argue that big data analytics is considered a cornerstone for innovation and achieving operational excellence in the financial services industry. The more traditional data analytics tools are becoming increasingly ineffective for deriving competitive insights from the big data now available, making the move to develop and utilise big data analytics *"inevitable"* in this sector (Ravikumar, Murugan and Sriram, 2022).

Big data analytics can provide financial institutions with actionable insights through real-time analysis of the complex big datasets they are developing. They can also help reduce costs, facilitate informed decision

making, enhance customer experience, understand customer behaviour, predict market trends, optimise risk management processes, predict stock prices, facilitate fraud detection and anti-money laundering practices, and identify new growth opportunities (Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024). Aderemi et al. (2024) note that real-time analysis is needed in the context of investment strategies as the speed and accuracy of decision making can significantly affect financial outcomes. Indeed, they argue that *"The significance of Big Data Analytics in financial services cannot be overstated"* (p.148) as it enables the financial industry to address some of its most pressing challenges, such as fraud detection and prevention, regulatory compliance and customer service.

Similarly, the large volume of data generated daily is leading banking services in the financial sector to use big data analytics for customer management, developing marketing campaigns and strategies, customising financial products and services, and enhancing their competitiveness (Ravikumar, Murugan and Sriram, 2022). Customer credit scoring is another application, as the quality of credit scoring can be improved by applying big data analytics to reduce the level of risk associated with credit provision (Ravikumar, Murugan and Sriram, 2022). Digital finance companies in India, such as Axio and EarlySalar, use real time applications of big data analytics for credit scoring and customer analytics to better understand customers' profiles and credit risk (Ravikumar, Murugan and Sriram, 2022).

Risk management is becoming an area in which big data analytics can play a pivotal role in the financial industry. Identifying potential risks is an important proactive approach for risk management, helping minimise losses and ensuring compliance with regulatory requirements, which have become more stringent since the financial crisis of (Aderemi et al., 2024).

Fraud detection has emerged as another key area within the financial sector in which big data offers potential. Through big data analytics, patterns and inconsistencies in large datasets can be identified, enabling proactive fraud prevention strategies and compliance with regulatory requirements (Aderemi et al., 2024). In the current era of digital financial transactions (which can be legal, illicit or illegal), increasing in financial fraud poses a significant challenge to the security and integrity of global financial systems (Udeh et al., 2024). Within this field, big data has emerged as an essential tool for detecting and preventing fraud in digital transactions, which can cause substantial financial losses, reduce trust and confidence among consumers and businesses, disrupt business operations and threaten the stability of markets (Udeh et al., 2024). Financial fraud in digital transactions can occur in different forms, such as identity theft, payment fraud, account takeover and malware attacks. Hence, *"The importance of detecting and preventing financial fraud cannot be overstated"* (Udeh et al., 2024, p.1747).

Unlike traditional approaches that rely on predefined rules and thresholds for analysis, big data analytics has the advantage of enabling the analysis of large volumes of data in real time (Udeh et al., 2024). In fraud detection, big data analytics is the process of analysing high volume and complex datasets to extract valuable actionable insights (Udeh et al., 2024). Analysing big data in real time can encompass a range of techniques, including machine learning algorithms, data mining and predictive analytics. These processes can enable users to uncover hidden patterns and correlations, detect anomalies and identify suspicious activity that does not match expected patterns, for example through the analysis of various data points and variables, such as transaction history and logs, geographic location, device fingerprinting, and biometric data (Udeh et al., 2024).

There are numerous examples of financial and e-commerce institutions that use big data analytics in their operations. For example, PayPal and Capital One use big data analytics to improve their risk assessment models and prevent fraudulent transactions, analysing multidimensional data points, transaction data, user behaviour patterns, and external threat intelligence feeds. This enables them to detect and block suspicious activities in real time and take proactive measures, reducing fraud losses and improving customer trust (Udeh et al., 2024). Alibaba, a global e-commerce organisation, also uses big data analytics to combat fraud in its online marketplace through the analysis of the same multidimensional points used by PayPal and Capital One to detect and block fraudulent activities, protect its customers, and maintain the integrity of its platform (Udeh et al., 2024).

Insights drawn from big data analytics in the financial field illustrate the potential to support predictive and investigative policing efforts. Both fields generate high volumes of data daily. In finance, big data is used to support real-time analysis, decision making, understanding customer behaviour and predicting trends, which are important areas that could also advance policing and criminal investigations. In addition, fraud detection stands out as a key application of big data in finance, offering proactive fraud prevention strategies, detecting fraudulent transactions and uncovering hidden patterns by non-police investigators. These are aspects that warrant further exploration as they could be useful for serious and organised financial crimes investigations.

2.3.4 Challenges in utilising big data in healthcare and finance

The use and analysis of big data to enhance decision making present both opportunities and challenges (Nnaji et al., 2024). The challenges come in two types, the first of which is implementation. For example, as Surbakti (2020) argues, organisations often struggle to use big data effectively. Surbakti's (2020) study found that while many executives understood that the use of big data and big data analytics could provide additional value, less than a quarter of those surveyed believed their organisations were currently able to

create that additional value through the utilisation of big data. Second, there is a trickier challenge in that the way big data works means fundamental changes to decision making and this has considerable societal implications (Surbakti, 2020). As Udeh et al. (2024, p.1753) and others (see, e.g., Awrahman, Fatah and Hamaamin, 2022; Nnaji et al., 2024) point out, these challenges include issues *"... [ranging from] data privacy and security concerns to ensuring fairness and transparency in algorithms, and from addressing biases and ethical considerations to navigating regulatory compliance and legal implications"*, meaning that *"stakeholders must grapple with various complexities to harness the full potential of big data"*.

Despite the considerable potential offered by big data analytics in the financial industry, the development and implementation of big data in this sector also comes with challenges, especially related to data privacy, security and governance, requiring a robust framework to protect sensitive information and ensure the ethical use of data (Aderemi, et al., 2024). In addition, the lack of skilled professionals able to analyse and interpret complex datasets is another challenge that affects the implementation of big data analytics in the financial industry (Aderemi et al., 2024). As the volume of data grows, so do the analytical tools and technologies that give big data analytics an important role and potential in shaping the future of financial services (Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024). Thus, we are seeing financial institutions continue to invest in analytics technologies and skilled professionals to support the potential for innovation and development in this field (Aderemi et al., 2024). The issues concerning privacy and security, ethics and bias, technical challenges and legislation are addressed further in the following paragraphs.

Privacy and security

Privacy and security are recognised as challenges to the utilisation of big data across the sectors investigated above and as essential areas to address for the use of big data to progress (Abouelmehdi, Beni-Hessane and Khaloufi, 2018; Awrahman, Fatah and Hamaamin, 2022; Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024; Nnaji et al., 2024; Udeh et al., 2024). As Abouelmehdi, Beni-Hessane and Khaloufi (2018, p.1) argue, *"Big data, no matter how useful for the advancement of medical science and vital to the success of all healthcare organizations, can only be used if security and privacy issues are addressed"*. In the financial sector, there is concern around the privacy of sensitive data when employing big data analytics because financial transactions include large volumes of personal and financial data, such as account numbers, transaction details and user identities, which must be protected from unauthorised access and misuse (Udeh et al., 2024).

Similarly, in healthcare, a key risk is data breaches and unauthorised access to personal and sensitive information, either accidentally or maliciously (Nnaji et al., 2024). Thus, while a degree of automation and the use of big data analytics may lead to better patient care and reduced costs, it also increases the probability

of data security and privacy breaches (Abouelmehdi, Beni-Hessane and Khaloufi, 2018). Redspin's 7th annual breach report, released by CynergisTek in 2016, reported that hacking attacks on healthcare organisations increased by 320% in 2016 and that 81% of the records breached resulted from hacking attacks (Abouelmehdi, Beni-Hessane and Khaloufi, 2018). In 2023, over 700 data breaches were reported in the US health sector, with the highest breach compromising over 133 million sensitive records (Balogun et al., 2025). Ransomware is another cyber threat to healthcare organisations and nearly doubled globally from 2022 to 2023, with a 128% increase in the US alone (Balogun et al., 2025). While there is recognition that providers need to take serious proactive comprehensive approaches to protect data from such cyber-attacks, breaches still frequently occur (Abouelmehdi, Beni-Hessane and Khaloufi, 2018).

The challenge in ensuring privacy concerns *"...the ability to protect sensitive information about personally identifiable health care information"* and in security regards *"...protection against unauthorized access..."* (Abouelmehdi, Beni-Hessane and Khaloufi, 2018, p.4). Security focuses on protecting big data from hacking attacks and theft for profit. While this is vital to protect data, it is insufficient to address privacy (Abouelmehdi, Beni-Hessane and Khaloufi, 2018); both are needed.

As a foundational insight, identifying the challenges concerning the use of big data related to privacy and security in the financial and healthcare fields lays the groundwork for further research should policing face the same challenges. That is, it will be of value to understand if big data can be useful in serious crime investigations without compromising individuals' rights or the public trust by addressing possible challenges such as unauthorised access, hacking, or ransomware, as found in the healthcare and finance sectors.

Ethics and bias

Although integrating big data and analytics into strategic decision making can be transformative, it is not without ethical challenges (Aderemi et al., 2024; Nnaji et al., 2024). The ethical use of big data analytics goes beyond complying with legal standards; it includes challenges of consent, transparency and potential bias (Nnaji et al., 2024). It is important to ensure that those whose data are collected, stored and used are aware of and consent to how this is done, although this can be complex and difficult to predict in the future given the speed at which the field of big data analytics is evolving (Nnaji et al., 2024). Another challenge is ensuring the fairness and transparency of algorithms that are developed and deployed in predictive analytical models (Udeh et al., 2024). Machine learning algorithms and predictive analytic techniques rely on historical data to identify patterns and then predict future events, meaning any bias in the historical data is replicated and reinforced by algorithms, which may lead to discriminatory results (Udeh et al., 2024).

It will also be essential to address ethical considerations, such as consent, transparency and bias, to ensure big data can be properly utilised in serious crime investigations, together with ensuring the fairness and transparency of algorithms as a means of providing a useful foundation for deploying big data in law enforcement applications.

Technical challenges

As well as the challenges outlined above, the implementation of big data presents some technical and human resources challenges for organisations aiming to benefit from its use (Ylijoki and Porras, 2016; Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024). A key issue identified in the financial sector is the lack of skilled professionals able to analyse and interpret complex datasets (Aderemi et al., 2024). The shortage of technical skills among human resources is hindering the adoption of big data in the banking sector and increasing reluctance to adapt to the change (Ravikumar, Murugan and Sriram, 2022). Thus, financial institutions are seeking to invest in technologies and skilled professionals to support the potential for innovation and development of analytics in this field (Aderemi et al., 2024).

Big data is huge in volume and has a complex structure and thus processing it requires high computational power, long duty cycles and significant real-time requirements (Jin et al., 2015). These requirements not only pose a challenge in designing system architectures, computing frameworks and processing systems but also present demands in terms of the energy consumption needed for such computational power (Jin et al., 2015). The evaluation of the energy efficiency of big data processing systems remains both a technical challenge and a cost/benefit question (Jin et al., 2015). The continuous evolution and production of big data can also be technically challenging in terms of analysing, storing and recovering the high volumes of data since standard database systems cannot be used to process and store the information due to the massive volume involved (Awrahman, Fatah and Hamaamin, 2022). Recently, cloud computing has offered new possibilities for the data mining of big data in the medical field, although several challenges remain to be overcome (Awrahman, Fatah and Hamaamin, 2022). First, cloud computing offers a simple and flexible way to mine data, but it increases the risk of privacy disclosure (Awrahman, Fatah and Hamaamin, 2022). Second, importing and exporting high volumes of big data to the cloud may increase financial costs and constrain transfer speeds (Awrahman, Fatah and Hamaamin, 2022).

Accuracy in big data analytics is essential. However, big data can contain typographical mistakes, abbreviations and notes giving rise to errors which affect the usefulness of the data collected and the outputs of analysis in addition to the potential bias described above (Awrahman, Fatah and Hamaamin, 2022). For example, in the fast-growing area of noise data, heterogeneous results can be caused by differences in degrees of quality and completeness, which may lead to false discoveries (Awrahman, Fatah and Hamaamin,

2022). The use of unreliable or inaccurate data in big data analytics can generate misleading information and affect organisations' decisions negatively (Ravikumar, Murugan and Sriram, 2022).

The findings concerning challenges in terms of technical aspects and the availability of skilled professionals suggest they are more likely to be related to big data than the fields in which it is applied. Further exploration of these challenges could provide insights into the readiness of policing infrastructure to adopt and utilise big data; if similar challenges are found, recommendations can be made to address them properly.

Legislation

Data privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union (EU) and the California Consumer Privacy Act (CCPA) in the US, impose strict requirements on organisations regarding personal data collection, processing and storage (Udeh et al., 2024). Complying with these regulations requires strong data protection measures, such as encryption, access controls and data anonymisation, to protect sensitive data and prevent unauthorised access (Udeh et al., 2024). The increase in data breaches and cyberattacks creates significant risks for the security of data; therefore, organisations must implement robust cybersecurity measures, such as intrusion detection systems, network segmentation and threat intelligence feeds, to detect and respond to security threats successfully (Udeh et al., 2024). Following such data privacy regulations prioritises data privacy and security, enabling organisations to build trust with customers and regulators and mitigate the risks associated with data breaches and cyber threats (Udeh et al., 2024). However, differences and/or conflicts in data protection laws and regulations across different jurisdictions may complicate compliance efforts for organisations globally (Nnaji et al., 2024). Therefore, it is necessary to follow regulatory requirements and best practices in the relevant field to achieve the highest data protection possible, as well as implementing data governance frameworks that enable organisations to manage data more responsibly and ensure protection against breaches (Nnaji et al., 2024).

Understanding how data are governed by regulations such as the GDPR and CCPA to increase data privacy and security measures can help guide the responsible adoption of big data in policing. Also, further research could be undertaken to explore if there are existing frameworks or laws that govern the application of big data in serious crime investigations to address challenges such as privacy and bias.

Section 1: Conclusion

The purpose of SRQ1 was to provide an overview of big data's potential across multidisciplinary fields. The findings indicated that with the expansion of technologies used by individuals, private companies and governments around the world, there has been a big data revolution. As yet, the concept of big data lacks clarity among scholars and users in the field and the study of big data is considered an emerging discipline.

Nonetheless, the continuous growth and expansion of data have made big data an interesting field for the private sector and governments, which have invested in it. The big data revolution has highlighted its potential and influenced its involvement in and transformation of various fields, such as defence, energy, business, health and education. These specific areas were not initially targeted or predefined for study but rather emerged through the search conducted for SRQ1, which led to the identification of healthcare and finance as potential fields for further scrutiny.

The widespread adoption of big data by medical companies and hospitals suggested strong potential for its application in other sectors, such as policing, as they share several valuable parallels. Both fields rely on sensitive and private data concerning individuals in the community, highlighting the importance of data quality, validity and privacy. Also, just as the medical field seeks to forecast the future in predicting outbreaks and detecting diseases, policing aims to predict and forecast crimes to prevent them, along with detecting crimes, calling for further research into the application of big data in policing and specifically in serious crime investigations to determine its usefulness.

Insights drawn from big data in the financial field illustrate the potential to support predictive and investigative policing efforts. Both fields generate high volumes of data daily. In finance, big data is used to support real-time analysis, decision making, understanding customer behaviour and predicting trends, which are important areas that could well advance policing and criminal investigations too. In addition, fraud detection stood out as a key application of big data in finance, offering proactive fraud prevention strategies, detecting fraudulent transactions and enabling non-policing investigators to uncover hidden patterns. These are potentially fruitful areas for further exploration in relation to SOC investigations.

As a foundational insight, discovering challenges to the use of big data, such as issues of privacy and security, in the financial and healthcare fields lays the groundwork for further research examining these challenges in the context of policing. Crucially, it will be necessary to determine whether big data can be employed in serious crime investigations without compromising individuals' rights or the public trust by addressing possible challenges, such as unauthorised access, hacking, or ransomware, which were found in healthcare and finance. The review showed that it is essential to address ethical considerations, such as issues of consent, transparency and bias, for big data to be properly utilised, highlighting the need for caution and further exploration in relation to serious crime investigations. It is also essential to ensure fairness, and the transparency of the algorithms employed to provide a useful foundation for the application of big data in law enforcement.

The findings concerning the challenges related to technical and skilled professionals seemed to be related more to big data itself rather than the fields in which it is applied. However, efforts from the financial and banking institutions can be seen in their investments in advanced technologies and skilled professionals to support their operations. Further exploration of these challenges could provide more insights into the readiness of policing infrastructures to adopt and utilise big data and if similar challenges were found, recommendations could be made to address them properly. Moreover, as the fields of healthcare and finance possess high volumes of personal data, understanding how the use of data is governed by regulations such as the GDPR and CCPA to assure data privacy and security measures could guide the responsible adoption of big data in policing. Also, further research might explore if there are any frameworks or laws that govern the use of big data in serious crime investigations to address challenges such as privacy and bias.

In summary, the findings in Section 1 of Chapter 2 suggested that the big data revolution is expanding and as an emerging discipline, the concept is not yet clearly understood in the field. This lack of conceptual clarity justifies the focus of Section 2, which develops a more structured understanding of the concept of big data as a foundation for further research within the context of policing. It has shown potential in healthcare and finance in supporting decision making, providing real-time analysis, understanding customer behaviour and forecasting. However, there are also various challenges, ranging from technical issues and a shortage of skilled professionals to aspects such as privacy, data security and bias.

2.4 Section 2: Current knowledge and understanding of big data

The preliminary findings from SRQ1 suggest that there is the potential for big data to be used across broad multidisciplinary fields, including academia, government and the private sector. The review of the knowledge base here raises several relevant areas for further interrogation relevant to the thesis topic, namely the usefulness of big data for serious crimes investigation, including: debates around the concept of big data and its definition; the ways in which big data can be deployed; the challenges faced across the sectors of healthcare, finance and marketing, which are arguably at the forefront of the utilisation of big data. This section explores the debates concerning the concept and definition of big data and its evolution in more depth to provide context for the development of second main scoping review question of the thesis: "What is known about big data and can it be useful in serious crime investigations?". This is followed in Section 3 by a discussion of what is currently known explicitly about the use of big data in policing and more specifically in serious crime investigations.

2.4.1 What is big data?

The term "big data" was first coined by researchers at the National Aeronautics and Space Administration in the US (NASA, 2025) in 1997 and its definition is still evolving (Ylijoki and Porras, 2016; Jurkiewicz,

2018). The NASA researchers, Cox and Ellswort (1997), observing that large data sets that take up or exceed the capacity of the main memory or local disk are a challenge for computer systems, stated, *"We call this the problem of big data"* (p.1). As they noted, the most common solution is to acquire additional resources (Cox and Ellswort, 1997).

It is generally agreed that the main elements of big data are that it comprises huge volumes of information that are gathered through different technologies rapidly and require continuous developments in processing to keep up with it (Jurkiewicz, 2018). In 1998, the notion of "big data" was used in data-mining context and in a hardware-related presentation and in 2003, it was used in relation to statistics (Ylijoki and Porras, 2016). Shortly after, in 2001, big data was described in relation to three aspects, volume, velocity and variety, which was considered a significant milestone in defining it (Ylijoki and Porras, 2016). In the following decade, companies like Google and Amazon developed big data solutions which were found to add value to their business (Ylijoki and Porras, 2016). In 2008, an article about big data was published in *Wired*, which increased public interest in the use of big data and its effects in science (Ylijoki and Porras, 2016). The next significant milestone was in 2011, when the McKinsey Global Institute published reports about big data that drew attention to its potential value and this has since been followed by various newspaper articles and the publication of scientific papers and books (Ylijoki and Porras, 2016).

According to Zainab and Dhanda (2018, p.39), *"Big Data are massive datasets that cannot be processed by means of traditional techniques"*. In greater detail, Mauro, Greco and Grimaldi (2016, p.131) define big data as *"the information asset characterised by such a high volume, velocity, and variety to require specific technology and analytical methods for its transformation into value".* Overall, Mauro, Greco and Grimaldi (2016) argue that the definition of big data should clearly refer to an identifiable entity, rather being related to or dependent on the field in which it is applied. In a rather different vein, Neiva, Granja and Machado (2022, p.1167) state that *"Big Data refers to a set of techniques that aggregate and analyses massive datasets to enhance the effectiveness and efficiency of decision-making in the areas where it is applied"*.

While there are various definitions of big data and experts have not yet agreed on one specific conceptualisation (Babuta, 2017), most refer to the "three Vs" as characteristics: high volume, high variety and high velocity (Richards and King, 2014; Broeders et al., 2017; van der Voort et al., 2019). High volume refers to the large amounts of data, high variety refers to the various sources of data that are stored and saved in different formats and high velocity refers to the speed of data analysis (Broeders et al., 2017). Over time, the definition of big data has expanded, focusing on different features and elements (van der Voort et al., 2019). As noted by Broeders et al. (2017), various authors have added "Vs": veracity (IBM, 2015), variability (Hopkins and Evelson, 2011; TechAmerica, 2012), value (Dijcks, 2013; Dumbill, 2013) and

virtual (Zikopoulos et al., 2012). This results in a potential seven "Vs" that can be considered features of big data.

Moreover, continuous technological developments and applications feed the debate concerning what defines big data and what distinguishes it from other forms of data. The concept of big data is flexible and can be interpreted in different ways among those who seek to employ big data-related technologies for various purposes – scholarly, commercial, or governmental (Chan and Moses, 2016). Perry (2017) argues that big data is much more than a lot of data, taking the view that although high volumes of data drive big data, solely associating the concept with the volume of data is a mistake; rather, what distinguishes it is the value that can be extracted from it. James (2016) agrees with this and considers that although the volumes of data are growing considerably, the term "big data" does not just describe its vastness but also implies the unrealisable benefits that may be gained if the correct tools are found and used to make value of the data. In the context of this study, the ability to gain value from previously unusable or combined data from various sources has the potential to provide new insights that may advance policing (James, 2016).

Emerging disciplines, such as the study of big data, often encounter a lack of agreement among scholars regarding the definition of the core concept (Mauro, Greco and Grimaldi, 2016; Ylijoki and Porras, 2016). The extent of agreement in the scientific community when defining a concept can be an indicator of a discipline's level of development and in the case of big data, the lack of agreement indicates that the concept is still developing (Mauro, Greco and Grimaldi, 2016; Ylijoki and Porras, 2016). As the number of stakeholders increases, there is increasing significance in arriving at a common understanding of what constitutes big data and the terminology associated to improve the quality of communication among those from different backgrounds as it currently a volatile term (Ylijoki and Porras, 2016). Although various scholars proposed their own definitions of big data, none have prevented subsequent scholars and studies from proposing new definitions (Mauro, Greco and Grimaldi, 2016).

Mauro, Greco and Grimaldi (2016) suggest that the quick and chaotic evolution of big data has hindered the development of a universally accepted definition of big data, while Surbakti (2020) further contends that this has prevented agreement on a universally accepted definition of the effective use of big data. The effective use of big data is a complex phenomenon that is rarely discussed in the literature, but Surbakti (2020) suggests that it is the ability to realise data into actionable insights and contends that having a consensual definition is an important step for organisations to realise the worth of big data:

> *Effective use of Big Data refers to harnessing value from Big Data by defining the problem, showing the value to the organization, and making solution by using insights from data and build Big Data infrastructure and data analytics by considering scalability and sustainability. (Surbakti, 2020, p.4)*

Hence, it can be proposed that the effective use of big data is related to value, highlighted by several scholars and studies as one of the characteristic Vs of big data. This is addressed further in the discussion in Chapter 5.

2.4.2 Data forms

The growth of data has led to the creation of large databases for governments and businesses and has initiated scientific research to understand and utilise data in the most beneficial way (Brady, 2019). Approximately every 20 years throughout the 19th, 20th and 21st centuries, new forms of information technologies have emerged:

> *…telephones (1870–1890s), phonographs (1870–1890s), cinema (1890–1920s), radio (1900–1920s), television (1940–1950s), mainframe computers (1940–1950s), personal computers (1970–1980s), the internet and World Wide Web (1980–2000s), cell phones (1980–2000s), and smart phones (2000s–present). (Brady, 2019, p.299)*

Brady (2019, p.300) has described the volume of digital communication data as a *"tsunami"*, expressing the huge size of data available that is generated by different technologies. In addition, widespread digital datafication creates data in different formats that can be stored and analysed by computers (Brady, 2019). Datafication is defined as:

> *…the process by which subjects, objects, and practices are transformed into digital data. Associated with the rise of digital technologies, digitization, and big data, many scholars argue datafication is intensifying as more dimensions of social life play out in digital spaces. (Southerton, 2020, p.1).*

Rai (2020) defines three forms of data – structured, semi-structured and unstructured – and contends that both structured and unstructured data are significant forms of big data. In the same vein, Kitchin and McArdle (2016) argue that variety is one of the characteristics of big data, which includes structured, semi-structured and unstructured data. From an alternative perspective, van der Voort et al. (2019) describe the structured form of data as traditional and less structured or unstructured forms as nontraditional.

Structured data are data that can be processed, stored and analysed in their original format and comprise one type of big data (Rai, 2020). This type of data constitutes highly organised information that can be accessed from a database through simple search algorithms (Rai, 2020). An example of structured data would be an Excel table containing employees' details, job positions and salaries in a company database, presented in an organised form (Rai, 2020). It is estimated that firms use 15–20% of their structured data to examine fraud detection patterns and undertake risk modelling (Zikopoulos et al., 2012). In contrast, unstructured data do not come in a specific form, making them very difficult to process and analyse and posing a challenge to organisations that handle big data (Zikopoulos et al., 2012; Rai, 2020). Technological limitations may also

create challenges for institutions in gaining value from unstructured data as a whole (James, 2016). Around 80% of the world's data are unstructured and the volume is increasing 15 times faster than that of structured data (Zikopoulos et al., 2012). Examples of unstructured data include emails, free text, images and videos (Babuta, 2017; Rai, 2020). Given the mass of data in unstructured form, data scientists are exploring and researching different ways to analyse them (Brady, 2019). Finally, semi-structured data contain both structured and unstructured data forms (Rai, 2020). Such data are not classified under a specific database but contain significant information that separates individual elements within the data (Rai, 2020). According to Zikopoulos et al. (2012) frameworks such as the open-source platform Hadoop can deal effectively with semi-structured data; Awrahman, Fatah and Hamaamin (2022) note it is one of the tools used to provide big data analytics in the healthcare sector.

Data generation

It has been estimated that around 90% of data worldwide was produced from 2018 to 2020 (Association of Police and Crime Commissioners [APCC], 2020; Rai, 2020). Furthermore, it is estimated that 2.5 exabytes of data are created daily, which is equivalent to 200 million 5 GB DVDs (Loenen, Kulk and Ploeger, 2016). The World Wide Web alone generates around 1,500,00 terabytes every day, providing an increasing amount of data that are available to scientists to study and analyse in different fields (Brady, 2019). In addition, Dijcks (2013, p.3) notes that *"The McKinsey Global Institute estimates that data volume is growing grew 40% per year, and will grow 44x between 2009 and 2020"*. In terms of the source of data, Brady (2019, p.209) points out that:

> *Human interactions through phone calls, email, texts, tweets, social media posts, and other technological methods are now digitally recorded, time- and location-stamped, and attributable to nodes in networks in ways that go far beyond the much more ephemeral media of the past.*

Nonetheless, in the digital world the "right" data might not always be available, whether because the data are not there, or are bad quality, outdated, corrupted, biased, or manipulated (Broeders et al., 2017). Notably, individuals may have little idea of the data that are being collected about them or even shared with third parties (Richards and Kings, 2014).

2.4.3 Big data characteristics

This sub-section explores the characteristics of big data to provide a better understanding of what shapes the concept and highlight the various perspectives of experts in the field, recognising that the concept of big data is open to interpretation and there is no consensus as yet on its definition and characteristics (Babuta, 2017; Jurkiewicz, 2018; Clissa, Lassing and Rinaldi, 2023).

Researchers in the field of data, such as Brady (2019), Broeders et al. (2017), Ferguson (2015), Horita et al. (2017). Loenen, Kulk and Ploeger (2016), Rai (2020) and Richards and Kings (2014), agree that the volume of data is expanding significantly. However, when it comes to big data, *"…the volume of data is not the only characteristic that matters"* (Dijcks, 2013, p.3). As noted in 2.4.4, big data has been defined as having characteristics described as the "three Vs": volume, velocity and variety (Kitchin and McArdle, 2016), the terms adopted by IBM (Zikopoulos et al., 2012). To these Dijcks (2013) adds "value". Bell et al. (2021) further contend that the understanding of the features of big data has expanded to encompass volume, velocity, variety, value and veracity, visibility, viability and variability.

Kitchin and McArdle (2016) argue that big data has more features, such as exhaustivity, resolution, indexicality, relationality, extensionality and scalability. From a rather broader perspective, but continuing with the "V" theme, Uprichard (2013, p.1) describes the features of big data as including the following: versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy, virility, valueless, vampire-like, venomous, vulgar, violating and very violent. Lupton (2015) contends that the "V" words most commonly used to characterise big data have been drawn from the domains of data science and data analytics but fail to elucidate the attributes of big data. As an alternative, Lupton (2015, p.1) proposes the "13 Ps": portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, privacy, polyvalent, polymorphous and playful. Offering a different viewpoint, Lyon (2014, p.5) proposes the key characteristics of huge volume, high velocity, extensive variety, exhaustive in scope, fine-grained resolution, indexical in identification, flexible, and offering scalability.

Extrapolating from the above, most studies (Zikopoulos et al., 2012; Dijcks, 2013; Dumbill, 2013; Richards and King, 2014; Kitchin and McArdle, 2016; Broeders et al., 2017; Brady, 2019; van der Voort et al., 2019; Bell et al., 2021) agree on four Vs to represent the characteristics of big data, addressed in turn in the following paragraphs: volume, velocity, variety and value.

Volume

Data that are generated by machines are produced in greater quantities and higher volumes than non-traditional data or data that are not generated by machines (Dijcks, 2013). For example, one airplane jet engine generates around 10 TB of data in 30 minutes and there are over 25,000 airline flights per day (Dijcks, 2013). This single data source can generate large volumes of data per day that amount to petabytes (Dijcks, 2013). Massive volumes of different types of data – environmental, financial, medical, surveillance and so on – are being stored (Zikopoulos et al., 2012) but a significant amount of the data generated is not analysed at all (Zikopoulos et al., 2012).

While Broeders et al. (2017, p.310) consider that volume pertains to *"...the use of large amounts of data"*, Kitchin and McArdle (2016, p.6) argue that *"In the context of Big Data, volume generally refers to the storage space required to record and store data"*. Data volumes have developed from terabytes to petabytes to zettabytes and this volume of data cannot be stored in traditional computer systems (Zikopoulos et al., 2012). To give an idea of the scale of existing data, according to Ferguson (2015), if printed in books, they would cover the surface of the US in 52 layers, and if on CDs and stacked up, they would reach to the moon in five separate piles. Since that comparison was made, the available data have been expanding daily. However, Brady (2019) argues that the real impact of the big data revolution is not the amount of data itself, but the change in our environment that demands new perspectives to handle datafication, networking, connectedness and computer authoring. Big data as a term does not refer to a specific size in terms of volume but implies volumes that exceed the normal or simple scales of data (Clissa, Lassing and Rinaldi, 2023).

In terms of handling big data, a well-known software system in the field is Apache Hadoop (Thabet and Soomro, 2015; Wadhwani and Wang, 2017). This is open-source software that can handle high volumes of data in real time by distributing the datasets for processing over a cluster of machines to manage structured, semi-structured and unstructured data (Thabet and Soomro, 2015; Wadhwani and Wang, 2017). However, Wadhwani and Wang (2017) contends that Hadoop is still a new technology and many professionals are not familiar with it, meaning it will likely require training resources. As an alternative, Wadhwani and Wang (2017) suggests another tool, Grid Computing, which offers the potential to face the challenges posed by the volume of big data. Grid Computing operates on several servers that are connected by a high-speed network and has two main advantages: first, it has high storage capability; second, it has high processing power (Wadhwani and Wang, 2017). Spark is another tool that offers high computing performance and can be used to manage high volumes of diversified data (Wadhwani and Wang, 2017). There are two additional ways for organisations to deal with the high volume of big data: (i) it can shrink the data; (ii) it can invest in an appropriate IT infrastructure to properly manage high volumes of data (Wadhwani and Wang, 2017). The approaches suggested aim to assist organisations in exploring, managing and gaining insights from big data (Wadhwani and Wang, 2017).

Velocity

As the volume and variety of the stored and collected data has developed, the velocity of data has also expanded (Zikopoulos et al., 2012). Kitchin and McArdle (2016) identify velocity as a key characteristic of big data that differentiates from small data. It refers to data that are created in real time and comprises two types: frequency of generation and frequency of handling, recording, publishing and exhaustivity (Kitchin and McArdle, 2016, p.7). Somewhat in agreement, Zikopoulos et al. (2012) describe data velocity as the

speed at which data is flowing, arriving and being stored and the rates of its retrieval, whereas Broeders et al. (2017) consider velocity to be the speed of data that is being processed, mostly analysed in real time.

An example of data velocity in social media streams can be found in tweets on Twitter (Dijcks, 2013), now known as X. Although historically a single tweet comprised a maximum of 140 characters only, Twitter generated a high velocity of data that could exceed 8 terabytes per day (Dijcks, 2013). The number of characters possible in a single tweet then expanded to 280 in 2017 (BBC, 2017), potentially leading to the generation of higher volumes of data. Indeed, with the increase in data and information sensors, organisations nowadays handle high flows of data measured in petabytes rather than terabytes (Zikopoulos et al., 2012). The flow of data is at such a high pace that it is increasingly difficult – if at all possible – for conventional systems to handle it (Zikopoulos et al., 2012).

One of the procedures used to manage the velocity of big data is sampling, defined by Wadhwani and Wang (2017, p.7) defines the sampling process as among *"…various statistics approaches used to select, manipulate and examine an elective subset of data points in order to recognize patterns and inclinations in the massive data set being inspected"*. In addition, systems such as Hybrid SAAS, PAAS, LAAS, and cloud computing have high potential to manage the challenges posed by big data's velocity (Wadhwani and Wang, 2017).

Variety

According to Broeders et al. (2017) data variety concerns the various data sources that are stored in different forms. The rapid growth in the number of sensors and smart technological devices used by organisations across the world has led to data complexity (Zikopoulos et al., 2012; Dijcks, 2013). This is because the data gather does not only include traditional data but also other formats: raw, structured, semi-structured and unstructured (Zikopoulos et al., 2012; Kitchin and McArdle, 2016). Hence, data variety describes all the types of data that contribute to the provision of insights and decision-making processes and as pointed out by Zikopoulos et al. (2012, p.8), *"To capitalize on the Big Data opportunity, enterprises must be able to analyse all types of data, both relational and nonrelational: text, sensor data, audio, video, transactional, and more"*.

One of the approaches proposed for managing big data variety is On-line Analytical Processing (OLAP) Tools (Wadhwani and Wang, 2017). OLAP Tools finds connections between pieces of information and arranges data in a logical way to make them easy to access, having the advantage of being able to process high volumes of data rapidly with low lagging (Wadhwani and Wang, 2017). Wadhwani and Wang (2017) also suggests an additional tool, the SAP HANA tool, which is an in-memory data platform that performs

real-time analytics, develops real-time applications, and can be deployed on-premises or in the cloud (Wadhwani and Wang, 2017).

Value

The definitions of big data characteristics have revolved around "Vs", including volume, velocity, variety, veracity, value, visibility, viability, and variability (Bell et al., 2021). Among these, Bell et al. (2021, p.2) contend that *"While all characteristics are arguably important, value has been highlighted as a key concern"*. Kitchin and McArdle (2016) view value as the ability to extract insights from data and reprocess them. In the business sector, for example, the value of data can be manifested in delivery accuracy, inventory accuracy, advanced product development and enhanced customer management (Bell et al., 2021). Dijcks (2013) considers that there is always good value information that might or might not be visible in a large volume of data. Hence, a potential challenge is finding the valuable information and extracting it for analysis so that benefits can be gained from it (Dijcks, 2013). James (2016) notes that attention should be paid to the sources of big data to make sure it is of high quality, which can be challenging in policing. This is because differentiating between good and bad data, whether system-generated or user-generated, can be difficult, especially as in the case of the latter, there might be few quality control measures (James, 2016). In summary, the term "big data" refers not only to its high volume but also the value that can be obtained from it (James, 2016; Perry 2017).

2.4.4 Big data ethics

According to the Data Ethics Framework published by the UK, data ethics is an emerging branch of applied ethics that demonstrates the approaches when data is generated, analysed, and disseminated (Government Digital Service, 2018). Data ethics includes data protection laws and related legislation that regulate the proper use and good practice of new technologies and information assurance (Government Digital Service, 2018). The size of data and quantity of information now available and able to be used for surveillance, investigation and prosecution are rapidly increasing (Broeders et al., 2017), along with different forms of data storage and computers that execute complicated data-processing analysis, which has resulted in different government entities interfering in the lives of citizens (Broeders et al., 2017). Using big data analytics in security policies affects freedom and security – both fundamental rights – at an individual and societal level (Broeders et al., 2017). As Babuta (2017, p.34) points out, *"The big data revolution brings with it complex ethical questions and the ethical implications of big data are as yet poorly understood"*. Ethical concerns regarding big data usually focus on three phases: collection, analysis and dissemination (Babuta, 2017). Crucially, according to Jurkiewicz (2018), there are some cases of big data ethical issues that organisations and politicians do not address. Although there are many benefits that can be gained from

big data, these ethical issues affect society and governments need to exploit their resources to protect their citizens and their rights (Jurkiewicz, 2018).

Jurkiewicz (2018) provides several instances of where big data ethical concerns can occur. First, data collected under the guise of social betterment, related to the individual's behaviour and preferences, are instead collected for profit, while publicly being advertised as for the individual's own benefit (Jurkiewicz, 2018). An example is software such as Blackboard and Turnitin, used by universities to assess levels of potential plagiarism, which may access information on the user's computer, such as course content, communications between students and professors and grade reports (Jurkiewicz, 2018).

Second, there is the issue of facilitating unequal wealth distribution, as it is estimated that 25–35% of jobs could be replaced by AI by 2025 (Jurkiewicz, 2018). Such changes will negatively affect the lower economic groups in society, since unskilled jobs will be replaced by AI first, potentially leading to an unfair divide between economic classes in society (Jurkiewicz, 2018). Furthermore, individuals and organisations that can afford advanced computer systems, collect big data and have the human resources to analyse it will likely be able to conduct deals ahead of those who cannot afford these systems (Jurkiewicz, 2018). This could lead to higher profits for those who can afford these systems, generating unequal market opportunities (Jurkiewicz, 2018).

Third, concealed data collection for profit may occur with the widespread use of camera systems worldwide; ostensibly installed to provide safety and security, they can generate data that can be sold for high profit (Jurkiewicz, 2018). There are also products offered by different companies, such as Apple's Siri, Google Home and Amazon Echo, which are supposed to assist humans in their daily activities but collect data and record sounds even when they are not in use, which constitutes an invasion of the user's privacy (Jurkiewicz, 2018).

Fourth and finally, big data technologies and tools can induce human dependency and lead to a decrease in logical thinking through increasing reliance on advanced technologies (Jurkiewicz, 2018). Advanced big data applications and algorithms can make multiple decisions faster than humans on our behalf, which will lead to an increase in our reliance on them (Jurkiewicz, 2018).

2.4.5 Legislation and big data

The first data protection legislation was introduced in the 1960s, when computers emerged and threatened individual privacy (De Hert and Papakonstantinou, 2009). This led to the introduction of Data Protection Acts in Europe during the 1970s and 1980s (De Hert and Papakonstantinou, 2009). The main source of concern and threat to privacy at that time was from state and government administrations (Costanzo,

D'Onofrio and Friedl, 2015; De Hert and Papakonstantinou, 2009). Government administrations gathered large volumes of data from their citizens to provide healthcare and education services (Costanzo, D'Onofrio and Friedl, 2015). Intelligence and police forces were untouched by these data protection measures since intelligence officers who were in charge of security-related surveillance did not act under court standards or regulations but were subject instead to government officials and parliament committees (Costanzo, D'Onofrio and Friedl, 2015). Since the 1970s, technology was a significant cause of a change in the nature of the relationship between policing and data protection legislation (Costanzo, D'Onofrio and Friedl, 2015). In the 1980s, the focus shifted towards the private sector, which realised commercial benefits from processing personal data (De Hert and Papakonstantinou, 2009). Therefore, individuals and their personal data had to be protected from both public administrations and private organisations in the face of activities such as data mining, data matching and profiling, which were introduced at that time (De Hert and Papakonstantinou, 2009).

As Bignami (2007, p.233) notes, *"Data privacy is one of the oldest human rights policies in the European Union"*. In the 1990s, the EU instituted data protection laws that aimed to prevent rights abuses in the market as it faced a challenge in protecting privacy rights with the developing systems of criminal justice (Bignami, 2007). Protection was enshrined as follows: *"Proposed in 1990 and adopted in 1995, the Data Protection Directive (95/46/EC) guarantees the right of individuals' data protection as well as the flow of data in the European Union (EU)"* (Costanzo, D'Onofrio and Friedl, 2015, p.239). Furthermore, according to Cardock, Stalla-Bourdillon and Millard (2017, p.142), *"Transparency is a key principle of EU data protection law and the obligation to inform is key to ensuring transparency".* Transparency aims to give data subjects information that enables them to evaluate the trustworthiness of the data controller (Cardock, Stalla-Bourdillon and Millard, 2017).

However, Loenen, Kulk and Ploeger (2016) contend that enhanced computing power, advanced data-mining techniques and public big data require an extension of the scope of the EU Data Protection Framework, particularly given that *"The open government data policies may conflict with the individual's right to information privacy as protected by the EU Data Protection Directive…"* (p.338). Furthermore, Ferguson (2015, p.373) argues that *"Big data remains largely under-regulated"*. With the widespread use of advanced technologies, legislative frameworks must develop to include new measures that regulate the use of these technologies, addressing the legal and ethical constraints that regulate the use of big data (Babuta, 2017). Indeed, Jurkiewicz (2018, p.52) notes that *"Calls for regulation of big data and protection of individual data are growing".* The EU is seeking to implement strict laws that regulate the collection and use of big data, requiring that the algorithms used are clear and understandable for citizens (Jurkiewicz, 2018). What is more, the EU can impose strong penalties, which can reach billions of dollars for those that do not comply

with its regulations (Jurkiewicz, 2018). In the UK, the data ethics framework is designed to guide anyone working directly or indirectly with data, analysing and creating policy or operational decisions through a set of principles (Government Digital Service, 2018). In contrast, the US has no such laws regulating big data (Jurkiewicz, 2018).

Section 2: Conclusion

Section 2 has established a baseline understanding of big data by reviewing how the concept has developed and how it is commonly defined in the literature. The review also identified that the most frequently cited characteristics are the "three Vs" – volume, velocity and variety – while highlighting that there is still no agreed definition or fixed set of characteristics across experts. This ongoing lack of consensus appears to be reinforced by continuous developments in technology, which repeatedly reshape what counts as "big" and what analytical capabilities are considered standard.

The section also examined the forms of data and how big data is generated. While a wide range of characteristics were reported in the literature, the main focus remained on the three Vs, with value frequently emphasised as an additional key feature. Finally, the section reviewed the ethical and legal approaches relevant to data practices. It found that most ethical frameworks and legislative instruments identified were designed to address data in general, rather than being explicitly tailored to big data. Although these sources are still relevant, the review did not identify frameworks or legislation that directly and specifically address big data as a distinct category.

A key implication is that although big data is widely discussed, it remains conceptually unclear and continues to evolve in practice. This lack of stability reduces consistency across studies, as different authors might apply the term in different ways. Taken together, these issues justify the need for this thesis to develop a clear and more stable working definition of big data for policing so that the concept can be applied consistently when assessing its usefulness within that context.

2.5 Section 3: The uses of big data in policing and its potential usefulness in serious crime investigations

This thesis has earlier explored the broader background and context of big data and policing (see 1.1), including the concept of serious crime and its types (1.1.2), current applications of big data in policing (1.1.3) and SOC strategies in the UK and Europe (1.1.4 and 1.1.5, respectively). Building on the earlier discussion, the review in this section focuses more specifically on the potential usefulness of big data in serious crime investigations. This allows the discussion to move from the wider background to a more focused body of literature relevant to the aims of the thesis.

Moreover, building on the findings presented in Section 2 (2.4), which explored discussions around the concept of big data, this section explores the uses of big data in police resource management, crime detection, prediction and prevention. Moreover, there is a critical analysis of challenges posed by big data in relation to policing and serious crime investigations, followed finally by an exploration of the expectations of policing professionals around big data. This comprehensive approach allows critical evaluation to answer SRQ2 and determine if big data can be useful in serious crime investigations following the empirical analysis and discussion with reference to the literature.

2.5.1 Big data analytics and police resource management

AI and big data analytics are being adopted by police forces to increase their operational efficiency in managing police resources and personnel (Ezzeddine, Bayerl and Gibson, 2023). In the policing context, analysing data and information is a fundamental aspect of supporting and delivering police services (Newburn, 2008). It is driven by the need to focus on reducing harm, manage risks and forecast demand (Newburn, 2008). In some forces, advanced computerised systems are used by data analysts to track crimes and assist police officers and detectives in solving them (Nath, 2006). Big data analytics has gained the attention of the security community due to its ability to analyse and link security-related data on an unprecedented scale (Cardenas, Manadhata and Rajan, 2013).

Big data analytics is defined by Feng et al. (2019, p.1) as "*...a systematic approach for analysing and identifying different patterns, relations, and trends within a large volume of data*". Data analysis in the policing context aims to translate raw information into operationally viable and valuable intelligence (James, 2016). Feng et al. (2019) and Zainab and Dhanda (2018) contend that big data analytics offers promising outcomes for policing, such as better understanding crimes, providing insights to track criminal activities, predicting incidents, identifying patterns, effectively deploying resources and enhancing decision making.

Feng et al. (2019) argue that big data analytics can effectively contribute to resolving the challenges that face big data sets, such as being too vast, unstructured and fast moving to be managed by traditional methods. Also, Pramanik et al. (2017) point out that big data analytics offer advanced innovative technologies that can reform security intelligence by enabling users to discover vital security knowledge from large databases. Cardenas, Manadhata and Rajan (2013) argue that one of the fundamental impacts of big data technologies (in policing) is that they are able to provide affordable infrastructures for security purposes, with new big data technologies such as Hadoop, Hive, Pig, RHadoop, stream mining and NoSQL enabling their users to quickly analyse large heterogeneous datasets. The rapid growth and increase in big

data policing software is partly due private companies developing increasing numbers of programs (Schuilenburg and Soudijn, 2023).

In the current information age, the volume of data is growing rapidly, and vast amounts are available to organisations to be managed and used in their decision-making tasks (Zainab and Dhanda, 2018), including those related to policing (Nath, 2006; Newburn, 2008; Feng et al., 2019; Ezzeddine, Bayerl and Gibson, 2023). As police forces move into an era of big data policing, they need to consider the quality of data they collect, as little is currently known about the quality of data that the police acquire and utilise (O'Connor et al., 2022). Police analysts and scholars have acknowledged that the quality of police data is one of the challenges they face, such as incomplete and inconsistent data, important details missing, and a lack of sufficient details to conduct proper analysis (O'Connor et al., 2022). Nonetheless, these data are used to allocate resources and guide operations (O'Connor et al., 2022), reported earlier as an advantage of the application of big data analytics in policing. O'Connor et al. (2022) suggested that the quality of police data could be improved by assigning data tasks to trained analysts, increasing training for frontline officers in data entry, acquiring more user-friendly software, ensuring accountability for data entry and quality, establishing data quality checks and audits, and having police leaders take data quality seriously.

It was established that that big data analytics have the potential to predict incidents, identify patterns and detect criminal activities (Feng et al., 2019; Zainab and Dhanda, 2018), and the following sections provides greater depth of these areas in relation to big data and serious crime investigations.

2.5.2 Crime detection, prediction and prevention

This section explores crime detection, prediction and prevention in relation to big data, as they are a part of crime investigations in which big data can play a potential role given its capabilities in various fields, such as healthcare and finance, as suggested earlier. Police forces are increasing their use of AI for security purposes as a result of the growing complexity of the crime landscape, seeking to advance their ability to predict, identify and combat new crime trends (Ezzeddine, Bayerl and Gibson, 2023).

Crime detection and data mining

Crime detection is defined as "*The process of uncovering criminal activity (or verifying reported crime) and acquiring evidence in order to identify and prosecute its perpetrators*" (Lexico, 2021, p.1). According to Assouli, Benahmed and Gasbaoui (2021), link prediction is a tool used in social complex network research to identify links between node pairs. Link prediction has promising potential to provide an effective way of discovering links between members in criminal groups and could contribute to preventing crimes and terrorist activities (Assouli, Benahmed and Gasbaoui, 2021). Big data can potentially be used in mass

surveillance to detect criminal activity in real time by combining digital data from various sources, such as video surveillance, facial recognition and electronic communications (Neiva, Machado and Silva, 2023). Applications of this can be seen in the activities of the Metropolitan Police and South Wales police, as they use facial recognition technologies to identify people through CCTV, mainly at large events, for crime detection and prevention purposes (Ezzeddine, Bayerl and Gibson, 2023). One of the primary techniques used in big data analysis is data mining, which is applied in various fields to gain useful information and new knowledge from data (Feng et al., 2019). Feng et al. (2019, p.2) argue that "*With the support of such techniques, [big data analytics] can help easily identify crime patterns which occur in a particular area and how they are related with time*".

It has been established that data mining demonstrates potential in crime detection as part of SOC investigations and as such it will be explored further in the next section.

Data mining

Data mining tools may be another means by which big data can be useful in serious crime investigations. Sharma (2014) suggests that as greater volumes of complex criminal data become available, data mining potentially becomes an option for use. AI is the subfield that establishes the foundations for data mining (Pramanik et al., 2017). Here, various AI algorithms are being developed to enable automated learning from data, to build crime prediction and detection models, criminal behaviour profiling models and models that can cluster criminal data (Pramanik et al., 2017). Some police departments and security and intelligence agencies already rely on diverse data-mining techniques to prevent and detect crimes and terrorism (Pramanik et al., 2017).

Some scholars (e.g., Sharma, 2014; Hassani et al., 2016) argue that criminology is one of the most significant fields in which data mining and its techniques have the potential to achieve remarkable results. Data mining is defined by Sharma (2014, p.1) as *"…the extraction of knowledge from large databases*". Tomar and Manjhvar (2016) assess the main principle of data mining as being to select information from huge datasets and transform it into an understandable form to be used in the future.

The review of the literature identified 33 data-mining techniques that can be used to investigate crimes (Nath, 2006; Hajian, Domingo, and Martinez-Balleste, 2011; Sharma, 2014; Hassani et al., 2016; Tomar and Manjhvar, 2016; Pramanik et al., 2017; Prabakaran and Mitra, 2018; Jha, Sivasankari and Krishnappa, 2020). These include artificial neural networks, association rule mining, classification, clustering, cumulative logistics modelling, decision trees, dimensionality reduction, discrimination analysis, entity extraction, the fuzzy c-means algorithm, genetic algorithms, hidden Markov modelling (HMM), the

influenced association rule, intelligent agents, the J48 algorithm, K-means clustering, K-mode clustering, kernel density estimation, link analysis, logistic regression, machine learning, the naïve Bayes rule, naïve Bayesian classifiers, the nearest neighbour method, neural networking, outliner analysis, prediction, the random forest algorithm, regression, rule induction, social network analysis, support vector machines, and text mining.

Hassani et al. (2016) state that the data-mining techniques most often used for crime analysis are the following: entity extraction, which relies on huge amounts of clean data to identify patterns in texts, images, or audio materials; clustering, which sorts data into related groups; association rule mining, decision trees, support vector machines, naïve Bayes rules and neural networks, which are classification techniques; social network analysis, which develops a series of interconnected nodes from relational data. Notably, Hajian, Domingo, and Martinez-Balleste (2011) found that data-mining techniques can detect discriminatory decisions, defined as "*...the act of unfairly treating people on the basis of their belonging to a specific group*" (p.1). Given that bias and discrimination are significant challenges within policing that are likely to be exacerbated by the use of big data, data mining may be a technique of significant use in resolving such issues.

Research conducted by Tayal et al. (2015) proposed a Crime Detection and Criminal Identification (CDCI) method employing data-mining techniques. The CDCI method is divided into six modules. The first is data extraction and starts by extracting unstructured crime data from various sources (Tayal et al., 2015). The second module is data preprocessing, which integrates and reduces the extracted crime data into structured crime events (Tayal et al., 2015). The third module starts by clustering crime events, using the k-means technique, into two clusters sharing similar attributes for crime detection (Tayal et al., 2015). This is followed by the fourth module, which uses a Google Map Application Programming Interface that embeds Google maps through Java to provide user-friendly and improved visuals (Tayal et al., 2015). The fifth module is classification using K-nearest neighbour clustering, which is a data-mining tool that discovers similarities among different crimes and arranges them into predefined classifications for crime identification and prediction (Tayal et al., 2015). The sixth and final module is termed WEKA and it uses a Java-based graphical user interface to verify the K-means crime results (Tayal et al., 2015). Prior research by Tayal et al. (2015) suggests that data-mining tools can successfully be used in criminal investigations. The findings of their research indicate that the CDCI method can assist police in narrowing down the identification of criminals, thus helping reduce the cost and duration of investigations (Tayal et al., 2015).

Prediction and prevention

Police forces around the world are increasingly applying advanced computer science technologies and statistics to predict events and automate their tasks (Vestby, 2019; O'Connor et al., 2022). As pointed out by Egber and Krasmann (2019, p.907):

> *The rise of big data and the fabrication of mathematical algorithms have made predictive policing possible, and these factors might constitute a qualitative difference to previous forms of forecasting and prevention in policing.*

Predictive analytics is one of the advanced techniques that enables users to benefit from real-time and stored data (Zainab and Dhanda, 2018). Neiva, Machado and Silva (2023) argue that big data technologies have the potential to make probabilistic predictions about the place and time where a crime is most likely to happen. The big data revolution has motivated police organisations to improve and develop their "hot spot policing", which is also known as crime mapping (Sandhu and Fussey, 2021). According to Sandhu and Fussey (2021, p.66), predictive policing refers to *"...police work that utilises strategies, algorithmic technologies, and big data to generate near-future predictions about the people and places deemed likely to be involved in or experience crime*". Predictive policing is the best-known example of big data policing and aims to predict the chances of crime occurrences in a specific area during a specific time (Schuilenburg and Soudijn, 2023).

Predictive policing is a technique derived from intelligence-led policing (Vestby, 2019; Sandhu and Fussey, 2021). The predictions are used to coordinate the deployment of police forces to prevent crimes in the future (Sandhu and Fussey, 2021; Schuilenburg and Soudijn, 2023). By applying predictive analytics to big data, valuable information can be produced to help predict future events and behaviours (Zainab and Dhanda, 2018). When it comes to predicting crimes, the analysis goes through multiple steps to finally reach the predictive results: data collection, data classification, pattern identification, and finally, crime prediction and visualisation (Zainab and Dhanda, 2018).

There are several software systems that claim to provide police forces with crime prediction software, such as the following: PredPol and HunchLab in the US, the Pre Crime Observation System (PRECOBS) in Germany, KeyCrime in Italy, Maprevelation in France, and the Crime Anticipation System (CAS) in the Netherlands (Schuilenburg and Soudijn, 2023). Among these, PRECOBS is one of the leading software options for predictive policing, used in German-speaking countries, which provides crime prediction based on past experiences and historical crime data to identify high-risk areas in which to employ policing actions, such as intensified patrolling (Egber and Krasmann, 2019). In the UK, the Durham Constabulary adopted the Harm Assessment Risk Tool (HART) to predict the probability of recidivism (criminals re-committing

a crime) within two years of being released from prison, aimed at determining if certain individuals can benefit from rehabilitation programmes (Ezzeddine, Bayerl and Gibson, 2023). The predictions provided by these software packages are based on comparative mathematics driven by computer algorithms which are capable of "*...high speed analysis of big data about crime*" (Sandhu and Fussey, 2021, p.66).

It is important to note that not all predictive technologies are complex. PredPol, for example, only uses three data points to produce its crime forecasts (Sandhu and Fussey, 2021). A key technology that underlines many of the predictive policing software options is machine learning (Vestby, 2019). Machine learning models for crime detection have been used by the UK Serious Fraud Office to identify legally privileged materials among millions of documents during an investigation and for crime prevention by the Norwegian Labor Inspection Authority to predict high-risk workplaces that are inspected by the agency (Vestby, 2019; Jha, Sivasankari and Krishnappa, 2020). Deploying predictive policing software has been shown to be effective in several cities in the US although no clear positive effects have been found in other countries (Schuilenburg and Soudijn, 2023). According to Babuta (2017), PredPol has proved useful in reducing property and burglary crimes in California. However, it is necessary to assess the relationship between the quality of predictions and the quality of the input data used to reach these predictions (Sandhu and Fussey, 2021). It is important to point out that when discussing machine learning in police decision making, it is not compared to ideal decision making but to normal human decision making (Vestby, 2019).

Crime prediction is fundamental for criminal justice and policing decision makers who aim to prevent crimes (Vomfell, Hardle and Lessmann, 2018). The overall aim of predictive policing technologies is to enable automated police decision making and to reduce the overall harms of crime by assisting in early intervention strategies (Sandhu and Fussey, 2021). Crime prevention refers to strategies that involve individuals, communities, businesses, government and non-government organisations seeking to address different social and environmental aspects that might contribute to crime and disorder (Abdul Jalil, Mohd and Noor, 2017; Sandhu and Fussey, 2021). In the UK, preventing serious crimes and violence is considered a priority for the government and police forces nationally (Brennan, 2022).

Predictive policing is likely to raise questions regarding accountability and may result in organisational and regulatory changes (Egber and Krasmann, 2019). Questions around the need for new forms of accountability can be raised if there are concerns that external software programs provided by private companies may generate false results or lead falsely to suspecting individuals (Egber and Krasmann, 2019). The issue is who should be held accountable for false-positive results and who within the police might be responsible for controlling the criteria applied for prediction and suspicion (Egber and Krasmann, 2019). Thus, while predictive policing may offer a more effective and objective approach to the prevention and prosecution of

crime, *"...much will hinge on how the police – and society – deal with the technology*" (Egber and Krasmann, 2019, p.916). Sandhu and Fussey (2021) found that little is known about the experiences of the police officers who use these predictive technologies and that while they recognise the potential of predictive policing, they also acknowledged the risk of biased predictions.

Pattern recognition has emerged as one of outcomes of crime prediction and will be explored further in this section. The ability to process and analyse data has considerably improved with developments in information technology capabilities (Loenen, Kulk and Ploeger, 2016). Due to the large size and variety of databases used in big data analytics, it is hoped that the outcomes will provide a highly precise risk analysis (Broeders et al., 2017). Some big data analytic techniques can reveal unexpected connections that may assist in creating risk profiles, for instance leading to the efficient use of police resources (Broeders et al., 2017). AI researchers have developed certain methods that combine traditional statistical methods, machine learning and statistical learning to produce methods of pattern recognition (Brady, 2019). Broeders et al. (2017, p.314) consider that *"Pattern recognition lies at the heart of big data"*. Ferguson (2015, p.371) takes the view that *"...with enough data, police will be able to predict criminal networks from patterns or connections"*.

As this research aimed to explore if big data could be useful in serious crime investigation, aspects of pattern recognition could be a key area, establishing if policing could identify criminal networks by applying pattern recognition technologies (Ferguson, 2015). However, as Broeders et al. (2017, p.314) warn *"...not all threats, security issues and types of crime show patterns that can be analysed in a meaningful way*", although they acknowledge that there are *"Increasingly sophisticated algorithms [that] can extract patterns from…data, enabling important advances in science, medicine, and commerce"* (Barocas et al., 2017, p.1).

2.5.3 Challenges of big data in policing

As well as discussing the advantages of big data in this thesis, it is important to address the challenges to attain a balanced assessment, in addition to comparing if there are any similarities/differences related to the challenges identified previously in response to SRQ1. Therefore, the possible challenges of big data in policing will be explored to discover if they might have any implications for the use of big data in SOC investigations.

As Babuta (2017, p.4) argues, *"Despite its widely transformative capabilities, big data is not without limitations"*. Although big data analytics applications have significant potential in the security field, there are several challenges that should be addressed to achieve their true potential (Cardenas, Manadhata and Rajan, 2013). Utilising big data in criminal investigations potentially entails various levels of complexity

that range from simple tools to the implementation of new digital infrastructures required for complex digital criminal investigations (Schuilenburg and Soudijn, 2023). As pointed out by Munoz, Smith and Patil (2016, p.21), one of the big data challenges for the policing community is to:

> *use new technologies to enhance trust and public safety in the community, especially through measures that promote transparency and accountability and mitigate risks of disparities in treatment and outcomes based on individual characteristics.*

The growing size, variety and frequent changes in big data require new big data analytics techniques to manage and benefit from these datasets (Zainab and Dhanda 2018; Feng et al., 2019). If new technologies are designed and implemented carefully, they could assist in policing decision making based on analysis employing factors and variables with a lower risk than human instincts and biases (Munoz, Smith and Patil, 2016).

The findings indicate that several challenges associated with big data have emerged and these will be critically analysed in relation to policing and SOC investigations.

The Vs as challenges

As mentioned previously (see 2.4.3), volume, variety and velocity are some of the distinctive characteristics of big data. According to Wadhwani and Wang (2017) and Brady (2019), these characteristics can present some technical challenges. With the increase in the number of sensors and smart devices, data gathered by organisations have become more complicated, encompassing different forms: traditional, non-traditional, raw, semi-structured and unstructured (Zikopoulos et al., 2012). Large volumes of data can lead to challenges in storage and management and developments in data variety can present challenges in changing data from one form to another (Wadhwani and Wang 2017; Brady, 2019). Furthermore, the growth in data velocity may force data analysts to prioritise and amend data on the run and choose only what is considered important (Wadhwani and Wang 2017; Brady, 2019). Finally, the veracity of data can potentially add a layer of complication over the volume, variety and velocity of data (Wadhwani and Wang, 2017; Brady, 2019). Therefore, new technologies are needed to overcome the challenges of big data arising from its characteristics (Thabet and Soomro, 2015).

To facilitate a more comprehensive understanding of the challenges presented by volume, variety and velocity, each is addressed individually in the following paragraphs. The discussion of the Vs in this section is not revisited as describing them as characteristics (see 2.4.3) but examine them in the context of challenges they pose for practice.

- Volume

The volume of stored data is increasing every minute. It was estimated that there were around 800,000 petabytes stored worldwide in 2000 and this was expected to increase to 35 zettabytes in 2020 (Thabet and Soomro, 2015). Updated figures indicate that worldwide data actually amounted to 59 zettabytes in 2020 and it was estimated that it reached 149 zettabytes in 2024 (Gkikas and Theodoridis, 2022). Similarly, in policing Hassani et al. (2016) argue that the volumes of crime data are constantly growing, leading to a big data evolution that demands new methods to analyse sets of data effectively and accurately. This constant growth in data has presented a considerable challenge to policing (Hassani et al., 2016). For example, the Metropolitan Police Service in London receives approximately 38 million records per day only from the Automatic Number Plate Recognition (ANPR) network (Babuta, 2017). The ANPR system is used when a vehicle passes an ANPR camera and its registration number is detected and checked against database records of vehicles of interest (Metropolitan Police, 2025). This technology is used to provide lines of enquiry and evidence and help detect and disrupt criminal activity at the local, regional and national levels. Currently, ANPR cameras submit an average of 60 million records per day for the Metropolitan Police in London (Metropolitan Police, 2025), an increase from around 38 million in 2017 (Babuta, 2017).

In terms of big data more generally, Wadhwani and Wang (2017) notes that social media is generating high volumes of data and mobile phones are one of the main sources of data generation. Wadhwani and Wang (2017, p.4) compares the challenge of coping with the volumes of big data to dealing with mountains and oceans of data, expressing how hard this is. Today, the availability of automated machines enables users to constantly track and record various types of data: environmental, business, medical and surveillance (Thabet and Soomro, 2015). The challenge becomes clear when the flow of data in an organisation increases, while the ability to process these high volumes of data decreases, therefore creating what is known as a "*blind zone*" (Thabet and Soomro, 2015, p.3). A blind zone situation occurs when it becomes hard for an organisation to decide and differentiate between the important, valuable and invaluable data it owns (Thabet and Soomro, 2015). Also, in terms of volume as a challenge, using large data warehouses to manage big data sets is expensive (Cardenas, Manadhata and Rajan, 2013).

However, Neiva, Machado and Silva (2023) argue that big data applications are expected to be more cost-effective and productive, enabling more efficient criminal justice decisions and reducing crime rates, particularly as advanced technologies are no longer exclusive to big corporations, with falling costs accelerating the pace of development (APCC, 2020). An example of this is that in 1967, storing one gigabyte of data cost £800,000 whereas today it would cost less than £0.016 (APCC, 2020). In a different vein, Jin et al. (2015) suggest that the real big data challenges lie in the variety of diversified data, the velocity of

timely response requirements and the veracity of data given uncertainties. These challenges are addressed further below.

- Variety

Zikopoulos et al. (2012, p.7) contend that "*The volume associated with the Big Data phenomena brings along new challenges for data centers trying to deal with its variety*". Generally, with the widespread use of sensors and smart devices that collect data, it is unusual to get structured data; organisations typically obtain raw, semi-structured and unstructured data from various sources, such as web pages, search indexes, images, videos, social media, and e-mails (Thabet and Soomro, 2015; Wadhwani and Wang, 2017). Thabet and Soomro (2015, p.3) consider that "*The massive volume of data caused by Big Data phenomenon [has] presented new challenge which is the variety of data types and format*". This leads to complexities arising from the different types and formats of data in a dataset (Wadhwani and Wang, 2017). In Thabet and Soomro's (2015) study, they found that only 20% of data could be processed by the systems available at the time, leaving 80% of data unutilised, not analysed or benefited from.

Addressing the variety of types of data is an essential requirement for the analytic and decision-making process (Zikopoulos et al., 2012). For an organisation to succeed, it will need to be able to gain insights from the different types of traditional and non-traditional data it possesses (Zikopoulos et al., 2012). In the context of policing, forces in the UK have access to various databases that contain a variety of forms of data. For instance, the Police National Computer is a nationwide database that contained over 12.2 million personal records, 62.6 million vehicle records, and 58.5 million driver records in 2017 (Babuta, 2017). There is also the Police National Database, which is a national intelligence handling system that contains data comprising local police records and enables officers to search across 220 different databases operated by individual police forces in the UK (Babuta, 2017). Moreover, the Ident1 database contains over 7 million fingerprint records (Babuta, 2017). As of 2024, recent figures show an increase in the number of records reaching over 28 million fingerprint forms relating to 8.7 million individuals (Home Office, 2024). This reflects not only an increase in data variety but also volume, an issue taken up in the discussion in Chapter 5.

- Velocity

Velocity, as already noted in discussing the characteristics of big data, entails the ability of current software packages to handle and process continuously generated data (Thabet and Soomro, 2015). The high velocity of data generation requires it to be analysed in motion and real time to find valuable insights, which in some cases also have a short shelf-life that may add an additional challenge (Thabet and Soomro, 2015). Wadhwani and Wang (2017), in agreement with Thabet and Soomro (2015), consider that one of the

significant challenges of big data is its velocity and dealing with the continuous high flow of data in the absence of the appropriate technology and tools. Although big data can offer organisations substantial insights from the terabytes or petabytes of data that flow to them daily, some lack the infrastructure to cope with the high volume of data (Thabet and Soomro, 2015). Due to the rapid growth in such datasets, tools, techniques and strategies are needed to successfully extract information and knowledge (Zainab and Dhanda, 2018).

By critically analysing the overall challenges of volume, variety and velocity, it is evident that they are related to another big data challenge, namely, the readiness of the technological infrastructure, which is considered in what follows.

Technological infrastructure

The rapid growth of data acquisition, cloud computing and storage technologies in different areas of businesses, research organisations and governments has led to challenges in handling the data collected (Feng et al., 2019). Wadhwani and Wang (2017) suggests that to be able to manage the challenges of big data, we need to understand its computational complexities, computational techniques and security threats. If real value is to be gained from big data, the appropriate tools must be available and employed to capture and organise the different data types collected from various sources (Dijcks, 2013). Thus, organisations must develop their technological infrastructure to handle the high volume, high velocity, and high variety of data from different sources and integrate them with existing data (Dijcks, 2013).

A major challenge of big data is its unlimited scalability and the need to process high volumes rapidly (Wadhwani and Wang, 2017). It is essential to consider the pace at which a project can grow and evolve in order to manage big data, otherwise big data projects may have to be put on hold and allocated additional resources to reach the expected outcomes because of the unexpected growth and evolution of the project (Wadhwani and Wang, 2017). Similarly, Srinuvasu and Santhosh (2017) contend that the main challenge of big data lies in storing and processing it in a specified time frame, which is dependent on the readiness of the organisation's infrastructure.

Munoz, Smith and Patil (2016) point out that the public and private sectors must collaborate to derive the greatest benefits from big data technologies. This requires developing hardware and software infrastructure, operational and management software and application programming interfaces (APIs) to create functional big data managing models (Thabet and Soomro, 2015, p.7). To attain the most appropriate information technology (IT) infrastructure for big data implementation, it is essential to address four requirements: performance, availability, scalability, and flexibility (Thabet and Soomro, 2015). First, performance

necessitates measuring the system's degree of response; as the system's performance increases, so will the cost of developing the IT infrastructure (Thabet and Soomro, 2015). Second, availability entails determining, for example, whether the system is required to be running 24/7 without any interruption; the longer the system needs to be available, the higher the cost of development (Thabet and Soomro, 2015). Third, scalability requires users to define the size of the big data infrastructure, the storage capacity and computing power; it is also essential to consider additional parameters to meet possible future challenges (Thabet and Soomro, 2015). Fourth, flexibility in the IT infrastructure system is measured by the speed and ability to add more resources and recover from failures; due to the nature of big data projects and continuous flow of data, the physical infrastructure must be resilient (Thabet and Soomro, 2015).

As noted by Babuta (2017), police forces possess enormous amounts of information gathered from different sources. However, there is a lack of the technological capability to analyse this information in the most beneficial way. Currently, analysing these data is a difficult task since police forces tend not to have access to advanced tools such as data mining. In this regard, Babuta (2017, p.1) argues that "*If the police were able to effectively apply such technology to the data they collected, they would greatly enhance their operational efficiency and crime-fighting capabilities*". Big data analytics requires high-performing technical tools in terms of computing power to process and store data, which organisations typically do not have access to (Babuta, 2017). Brady (2019) points out that while there are advances in computer power when it comes to managing big data, the volume of such data is increasing faster than our capability to process it. Acknowledging the challenges and the potential of big data in policing, Babuta (2017, p.32) states that "*It is hoped that in the near future, forces in the UK will have access to the infrastructure necessary to implement advanced analytical tools and methods*".

Privacy

Wadhwani and Wang (2017) argues that when extracting data and identifying trends, extra care should be given to assuring individuals' privacy and securing their data. Thabet and Soomro (2015, p.6) note that privacy is considered "*...a major concern*" when it comes to the context of big data. Private information on online platforms such as Facebook, X (formerly known as Twitter), and so on, can be shared and some users may not understand the consequences of sharing data or how information can be linked to identify the personal details of users (Thabet and Soomro, 2015). Another example of privacy concerns is related to online services which ask users to share their location; users may assume that their identity is hidden, not realising that the location-based service provider can discover the user's identity by tracing the location information back to an office location or home residence (Thabet and Soomro, 2015).

In addition, big data analytics and data mining requires access to big data sets, which may raise privacy concerns (Hassani et al., 2016). Data privacy in the context of crime prevention has also surfaced as a challenge as there have been calls to share data between the business sector and policing, which could go against the data collection principles of data reuse (Cardenas, Manadhata and Rajan, 2013). Babuta (2017) stresses the urgency of addressing this matter to ensure the police can effectively use big data technologies without violating citizens' right to privacy. Nonetheless, the SOC strategy 2023 encourages cooperation between the private sector and the government in combating SOC (Home Office, 2023). In this regard, Europol also faces constraints in the direct exchange of personal data with private parties, which it claims obstructs it from effectively supporting its members in combating serious crimes (European Commission, 2020).

Ethics and bias

The big data revolution brings with it complex ethical questions for policing and a fundamental challenge to the use of big data can be found in legal and ethical restrictions (Babuta, 2017). Richard and Kings (2014) emphasise that law is an important element in big data ethics, in addition to establishing ethical principles and identifying best practices. In relation to policing, Babuta (2017, p.36) contends that "*At present, while the police's use of data is legally governed by data protection legislation, there is no clear decision-making framework for the ethical use of big data technology in law enforcement*".

A potential risk is that some big data methods can result into data determinism, which arises when *"…individuals are judged based on probabilistic knowledge (correlations and inferences) of what they might do, rather than what they actually have done"* (Broeders et al., 2017, p.314). Developments in big data support predictive policing, a practice described as follows:

> *A police department engaged in predictive policing uses data mining methods to find correlations between criminal outcomes and various input data they have collected-crime locations, social networks, or commercial data. (Selbst, 2017, p.113)*

However, in 2016, 17 civil rights organisations released statements expressing their concerns regarding some of the outcomes of predictive policing (Selbst, 2017). The civil rights organisations argued that by using data mining and predictive policing, there is the possibility of racist outcomes, in addition to the absence of transparency among police forces (Selbst, 2017). They observed that if big data and data mining tools were not used carefully, there would be the possibility of producing discriminatory patterns and biased decisions (Selbst, 2017). Therefore, analysing anonymised crime data rather than personal data is less challenging from an ethical and legal perspective (Babuta, 2017).

Moreover, data provenance is a challenge for big data analytics (Cardenas, Manadhata and Rajan, 2013). Since big data is collected from various sources, it is challenging to be certain that every data source is reliable and suitable to ensure the algorithms produce accurate and unbiased results (Cardenas, Manadhata and Rajan, 2013). Therefore, when using data analytics to develop predictive tools, it is necessary to ensure that the algorithms do not depend on factors and variables that might classify a particular community based on characteristics such as race, religion, income level or education (Munoz, Smith and Patil, 2016). For instance, a machine learning algorithm that considers past arrests as a factor in its analysis could indicate that a certain community requires more policing, irrespective of the face that the community may be positively changing over time (Munoz, Smith and Patil, 2016).

Newburn (2008) stresses that it is vital to consider the quality of data and information used in analytical tools such as crime maps or offender network associations. Knowing the quality of data entails measuring how reliable the dataset is for decision making (Thabet and Soomro, 2015). However, differentiating between correct/incorrect and reliable/unreliable data is challenging even if the best data cleaning tools are used (Jin et al., 2015). Thus, it is vital to determine the quality of the dataset and its relevance to the case under analysis (Wadhwani and Wang, 2017), since no matter how elaborate the analytical tool, it will not be able to replace or make up for poor quality data (Newburn, 2008).

Computer scientists have started to investigate different ways of eliminating or at least reducing discriminatory data outcomes that could become a part of the model building process (Barocas et al., 2017). Their approaches have traced the sources of unfairness or discriminatory data in the machine learning process. It has been assumed that a key problem lies in the training data, which is a fundamental source of potential bias. For example, some of the tools used to analyse big data can create decision-making patterns that are implicitly discriminatory, which is an unintentional outcome of extracting data. Importantly, Barocas et al. (2017) argue that *"Implicit discrimination by algorithms requires our attention because such data-driven methods are deployed in many of our most crucial social institutions".* They stress that biased outcomes affect civil rights and can be traced back to two possible origins: (i) computer scientists might not have enough knowledge of civil rights and fairness matters; (ii) civil rights scholars might not have enough knowledge and understanding of big data and its complexities (Barocas et al., 2017). Therefore, it is suggested that investments should be made for computer scientists and civil rights scholars to work together, advancing their knowledge and providing the intellectual resources to achieve fairness and avoid biased outcomes (Barocas et al., 2017).

2.5.4 Frameworks and legislation

To attain the benefits of using big data analytics, a framework that governs its use in the field of security should be developed (Broeders et al., 2017). This framework will help add layers of protection to assure essential rights and ensure that there is no improper use of the data collected and analysed. Although there are potential benefits of using big data in the field of security, individuals' freedoms and fundamental rights might be at risk of violation (Broeders et al., 2017). For example, during the 1990s the EU created data protection laws to prevent any rights violations (Bignami, 2007). Years later, the EU faced a challenge in protecting privacy rights in the face of governments exercising their powers to protect their national security (Bignami, 2007). Hence, *"The challenge for the European Union is to protect privacy in its emerging system of criminal justice"* (Bignami, 2007, p.233). While governments and states are responsible for the safety and security of their citizens, they are also responsible for protecting their fundamental freedom and rights (Broeders et al., 2017).

As previously mentioned in relation to ethical challenges, predictive policing techniques may generate discriminatory results, and the degree to which these results are biased are unclear to the police and the society (Selbst, 2017). That is because there is no incentive for police forces that are already focusing on crime control to spend additional resources to investigate possibly biased results (Selbst, 2017). In this case, Selbst (2017, p.110) argues that *"…neither the typical constitutional modes of police regulation nor a hypothetical anti-discrimination law would provide a solution…".* The solution Selbst (2017) proposes is to provide *"algorithmic impact statements"* to cover the legislative gap. Algorithmic impact statements would require police forces to assess the efficiency of predictive policing methods and whether they generate any discriminatory results (Selbst, 2017). In addition, this regulation would allow the public to be involved, viewing and commenting on the process. Thus, *"Such a regulation would fill the knowledge gap that makes future policy discussions about the costs and benefits of predictive policing all but impossible"* (Selbst, 2017, p.110).

Furthermore, the new technological methods adopted by the police could lead to discriminatory results in terms of increases in arrests, jailing and physical harm especially to people of colour (Selbst, 2017). If these issues cannot be resolved by existing legal restrictions, new legal constraints must be created (Selbst, 2017). However, Munoz, Smith and Patil (2016) propose that police forces can work to avoid these issues through implementing key steps, such as ensuring transparency and accountability in the data input process and eliminating data referring to certain races to ascertain that the tools used are not biased.

Finally, deploying AI and big data applications in policing may trigger uncertainty and scepticism concerning the potential ethical and moral consequences (Ezzeddine, Bayerl and Gibson, 2023). As a result

of the ongoing debates in this area, regulations and legislation are needed, as well as taking into consideration public opinion to allow for informed decision making about adopting AI applications for policing purposes (Ezzeddine, Bayerl and Gibson, 2023).

2.5.5 Expectations of big data in criminal investigations

This section discusses two studies, Neiva, Granja and Machado (2022) and Neiva, Machado and Silva (2023), which report on research exploring big data and policing, specifically criminal investigations, and hence are of relevance to the aims of this thesis. The reason for including these papers in a sub-section is that they address one of the latest studies presenting an overview of policing professionals' perspectives, yet to be found in other studies. The former paper discussed the views of 22 policing professionals from 16 countries and the latter included an analysis of 14 articles published between 2015–2022 in 22 countries.

Neiva, Granja and Machado (2022) aimed to examine the expectations of professionals involved in police cooperation within the EU regarding the use of big data in criminal investigations. This study involved 22 policing professionals from 16 countries in the EU and found that the participants perceived big data to be potentially beneficial for criminal investigations, especially in cases where forensic evidence such as fingerprints and DNA could not offer productive lines of inquiry. In addition, they considered that big data can assist in solving cold cases by generating new information useful in a criminal investigation (Neiva, Granja and Machado, 2022). Another potential benefit is the ability to strengthen the connections between different datasets, which can enable professionals involved in EU police cooperation to connect information from different databases, ultimately advancing a criminal investigation (Neiva, Granja and Machado, 2022). The study described big data analytics as an *"analytical weapon"* that has the potential to connect and analyse data from various sources to construct valuable intelligence for a criminal investigation (Neiva, Granja and Machado, 2022, p.1171).

Although the findings indicated various potential advantages of big data in criminal investigations, the participants also outlined some potential concerns about privacy, bias and the high volumes of data (Neiva, Granja and Machado, 2022). Large volumes of data require the allocation of resources to distinguish between valuable and non-valuable datasets and handling large amounts of data can be challenging for a criminal investigation. In addition, big data analytics involving genetic and DNA data can lead to issues with profiling and biased outcomes, giving rise to ethical problems due to the potential for inaccurate predictions. Thus, Neiva, Granja and Machado (2022) argue that there is a need to develop regulations that provide the required data protection and support the use of big data in policing activities and criminal investigations specifically.

Although there is a framework for EU policing agencies (Directive EU 2016/680), which contains data protection legislation, it does not explicitly refer to big data (Neiva, Granja and Machado, 2022). In addition, while the current use of data by police is governed by data protection legislation, there is no clear framework that governs the ethical use of big data technologies in policing (Neiva, Granja and Machado, 2022). As has previously been argued, using large datasets has the potential to raise ethical issues (James, 2016). Nonetheless, in Neiva, Granja and Machado (2022, p.9) study, this was mostly seen by the participants as *"…a potentially useful technique".* In addition, Neiva, Granja and Machado (2022) found overall agreement that big data is of public interest for use in criminal investigations and arresting criminal suspects. The authors referred to using big data in criminal investigations as having *"…underexplored investigative potential"* (2022, p.10), especially after a crime has been committed. This indicates a research gap and aligns with the contribution that this thesis aims to make to the body of knowledge.

In the review conducted in their later paper, Neiva, Machado and Silva (2023) suggested that police forces have begun to develop their capability to utilise big data in various ways. However, there is limited knowledge about the perspectives of professionals in police forces regarding big data and the application of its technologies in the field of policing and criminal investigations. Neiva, Machado and Silva (2023) aimed to fill this gap by analysing 14 articles published in 2015–2022 to better understand the views of policing professionals in 22 countries. They identified both optimistic and oppositional views about using big data in crime investigations. The optimistic views were that big data can improve the objectivity and efficiency of policing, lead to better management of police resources, and help develop strategies to predict crime, facilitate data exchange, undertake analysis, and advance criminal investigations (Neiva, Machado and Silva, 2023).

Also, Neiva, Machado and Silva (2023) argue that big data can be used at different times, both before a crime has been committed for prediction and following for investigation. It is argued that big data derived from social media can advance policing by developing strategies that are proactive and more focused on prediction and prevention rather than reactive practices after the occurrence of a crime (Neiva, Machado and Silva, 2023). In addition, big data technologies are perceived as part of the "scientification" of police work to provide new tools to analyse large volumes of data and predict social problems (Neiva, Machado and Silva, 2023).

However, despite the potential in policing and criminal investigations, there are some potential risks, such as errors, biases, lack of regulation and privacy threats (Neiva, Machado and Silva, 2023). Some of the data used to make future predictions are based on police officers' past categorisations, which could potentially produce biased predictions (Neiva, Machado and Silva, 2023). Policing professionals emphasised the

significance of human involvement in decisions, not just relying on the data emerging from the analysis but rather inspecting to identify any errors that might have been produced by the software (Neiva, Machado and Silva, 2023). Also, some concerns were expressed about the lack of regulations addressing and regulating big data, which could lead to unintended social consequences (Neiva, Machado and Silva, 2023). However, these social consequences were not explained. Moreover, it was argued that the usefulness of big data technologies in policing was not clear and empirical evidence would be required to present the added value (Neiva, Machado and Silva, 2023). Additional practical barriers could exist when utilising big data in policing, for instance having the economic resources to implement the technologies, lack of professional training for police officers, administrative tension between analysts and police managers, and challenges in accessing data and resources (Neiva, Machado and Silva 2023).

Section 3: Conclusion

Overall, Section 3 has shown that the growing complexity of crime is increasing the need for more advanced policing capabilities. The literature suggests that the big data revolution is influencing policing in terms of introduction new approaches to crime detection, prevention and prediction. However, a key limitation is the lack of detailed research on how police can use big data applications in practice and the extent to which they may achieve measurable outcomes, demonstrating their usefulness. Although recent studies have started to define big data in policing and outline its potential uses, the evidence is still nascent.

The section also reviewed definitions of "serious crime" and highlighted its financial and social impacts, which emphasised on addressing them. In addition, a comparison of serious crime types across the six sources (Table 1.1) showed some similarities and differences in classifications, the latter due to legislative distinctions. Despite these differences, the harm associated with serious crimes was consistently recognised. The section further noted the complexity of tackling serious and organised crimes in the UK, given that 100 agencies are involved, increasing the need for coordinated intelligence and highlighting the importance of addressing this matter.

The literature also indicated growing international interest in using big data applications in policing, for instance DNA and biometric databases, facial recognition and financial intelligence tools. The UK SOC strategies (2013, 2018, 2023) were examined to assess the role of big data, if any. The earlier strategies referred to the "bulk" of data, possibly indicating "big" data, whereas the 2023 strategy did not explicitly mention it. Nevertheless, the 2023 strategy placed data and data analysis at the centre of efforts to combat SOC, suggesting a data-led direction even without the big data label. Wider strategies reinforce this movement. The National Policing Digital Strategy (2020-2025) explicitly supports the application of AI and big data in advancing policing and criminal investigations, while acknowledging the uncertainty of future

policing directions, which could be linked to rapid technological change and uneven readiness across jurisdictions. In addition, the EU Security Strategy (2020–2025) recognises the importance of employing big data and AI to address evolving security threats.

Taken together, the background and context in Section 1.1 established the broader context in which big data is being considered in policing in view of the evolution of crime; strategies are increasingly becoming data led and international cooperation is expanding around data sharing. This creates a clear rationale to examine specific analytical capabilities that are expected to deliver operational value. Therefore, in what follows, I focus more directly on big data analytics applications and tools, alongside the practical and ethical challenges that shape whether these approaches can move from strategic aspirations to effective use in serious crime investigations.

Examining big data analytics in policing identified commonly reported advantages, including improved operational efficiency in resource management and support for crime detection, prediction and prevention. The literature notably highlighted the role of big data in crime detection, focusing especially on data mining tools compared to other forms of AI. This suggests that the evidence on data mining tools may be more developed than for newer or less established AI tools. The review further indicated that predictive policing if often presented as the most recognised example of big data policing, with several systems cited in the literature, including PredPol, HunchLab, PRECOBS, KeyCrime, Maprevelation, CAS and HART. However, the literature does not consistently provide clear or comparable evidence of effectiveness across these systems, and it is not always explicitly stated whether their operations meet the big data threshold in terms of volume, variety and real-time processing.

Moreover, predictive policing was repeatedly linked to questions of accountability and legitimacy, implying that its adoption may require organisational change and will require strong oversight and clear regulation. Section 3 then synthesised challenges identified across the literature, showing a strong emphasis on technical and resource constraints. The challenges were commonly described as related to the growing volume, variety and velocity of data and the cost of management. In policing, these challenges translate into practical difficulties in building and maintaining the required infrastructure. This is especially the case where police forces hold large volumes of data but lack the tools, integration and capacity required to undertake big data analytics projects. Together with the technical barriers, privacy concerns were raised, with calls to secure personal data. Ethical and bias concerns were also prominent, including fears that predictive systems may produce biased outcomes, particularly if they rely on certain identifying variables or historical data. As a result, the studies reviewed stressed the importance of data quality and caution in training algorithmic models.

Finally, the literature highlighted the need for governance frameworks and legislation addressing accountability, transparency and fairness in big data use, especially to reduce the risk of biased decision making. Lastly, two recent studies were included because of their focus on policing professionals' expectations of using big data in criminal investigations, which closely aligns with the scope of this thesis. Their inclusion strengthens the review by providing up-to-date and practice-relevant insights.

Section 3 supports the need for this study by showing a gap between the growing strategic push for data-led policing and the limited empirical evidence on how big data is used in practice and how useful it might be. Although the literature highlights predictive and detection-focused applications, these are often presented without consistent evaluation of their effectiveness. At the same time, persistent technical, financial and infrastructure challenges, along with privacy and bias concerns, suggest that usefulness is conditional rather than guaranteed. Together, these points justify the focus of this thesis on developing a clear operational understanding of big data in policing and evaluating its usefulness in serious crime investigations.

Chapter 3. Methodology

3.1 Introduction

This chapter outlines the methodological approach adopted for the interview component of the thesis, which builds on the scoping review and contributes to answering the overarching research question. The chapter first outlines the development of the research question and the rationale for the approach chosen. It then presents the conceptual and theoretical frameworks guiding the thesis, followed by the sampling strategy and justification for the participant selection. The chapter next describes the qualitative research methodology, including the data collection and analysis methods. Finally, it presents the researcher's positionality and ethical considerations. Overall, these sections provide a clear explanation of the methodological choices and approaches taken to ensure this study was robust, aligned with the research aims and conducted ethically.

3.2 Development of the main research question

This thesis began exploring big data and policing/criminal investigations in general in its early stages and then narrowed the scope to serious crime investigations. Identifying the two main areas of research – big data and policing – emerged from my professional experience and field of work. This is consistent with Johnson et al.'s (2020, p.139) view that *"Formulating a research question is often stimulated by real-life observations, experiences, or events in the researcher's local setting that reflect a perplexing problem begging for systematic inquiry"*. Day-to-day observations and practice pointed to the need to explore in depth the role of big data in advancing serious crime investigations. This is discussed further in relation to the conceptual framework (see 3.3.1).

A research question starts as a set of propositions that describe the relationship among certain concepts or experiences (Johnson et al., 2020). In this thesis, it is the relationship between serious crimes and big data. In the first stages of the research, the development of the research question was informed by the evolving findings from the literature. It started as "How can big data be useful in crime prevention and detection?" but was refined as it assumed that big data would be useful without sufficient academic evidence. In a further stage, the research question was amended to "Can big data effectively contribute to the detection of serious crimes and suspects?" and later to "Can big data be useful in serious crime investigations?". Initial research questions are usually broad in focus to be researchable, but as Johnson et al. (2020, p.139) point out, these *"…initial qualitative research questions guide the inquiry but often change as the author's understanding of the issue develops through the study"*.

Developing and refining the primary research question to focus on the phenomenon of interest and the context in which it is positioned is necessary to achieve research rigour and quality (Johnson et al., 2020). As the understanding of the researched areas developed, the scoping review question was refined to align more closely with the research aims and evidence emerging from the literature. As a result, the final research question was formulated as: "What is known about big data and can it be useful in serious crime investigations?". The first part of the research question was formulated as "What is known about big data…", which was addressed in Section 2.4, prior to exploring the relationship with policing/serious crime investigations more specifically. The rationale for this approach – exploring big data first – was that preliminary findings pointed to debates and ambiguity regarding the concept, its definitions and characteristics in the field. Therefore, this approach ensured that the thesis could proceed based on a better understanding of the concept of big data before exploring it in the context of policing. Building on this foundation, the second part of the research question, "…and can it be useful in serious crime investigations?" explored serious crime investigations in relation to big data.

The research aims were outlined in the introduction (see 1.4); however, they are restated here to clearly show their relation to the chosen methodology. The research aims of this thesis are:

1. To develop a comprehensive understanding of the concept of big data, including its definitions and characteristics, to form a theoretical basis as a foundation for further research within the context of serious crimes.
2. To explore the potential usefulness of big data in serious crime investigations by identifying the advantages, disadvantages, challenges and the availability of artificial intelligence (AI) tools in relation to big data.
3. To develop a definition of big data suitable for use across policing internationally.

After developing the initial research question, a conceptual framework is constructed to provide a logical argument for the research (Johnson et al., 2020), and this will be presented in the following section.

3.3 Theoretical and conceptual framework

3.3.1 Conceptual framework

Developing a conceptual framework defines and justifies the research questions and the selected methodology to enable them to be answered (Johnson et al., 2020). In addition, the conceptual framework is essential to establish a research topic that is based upon a thorough and integrated review of relevant literature (Johnson et al., 2020). In the context of this thesis, the relevant literature was explored in response to the central research question. Also, the conceptual framework identifies and integrates key concepts,

assumptions and best practices in a way that demonstrates the problem in relation to the research question(s) (Johnson et al., 2020). As noted previously (see 3.2), the research problem first emerged from my professional practice and experience and was further developed through the initial stages of exploring the literature. Consequently, for this thesis, Chapter 2 presented key concepts of the researched areas ("The expansion of big data" in 2.3.1 and "What is big data?" in 2.4.1), and assumptions ("Expectations of big data in criminal investigations" in 2.5.5). These, together with the other sections in the literature review, support one another in offering a comprehensive understanding of the key concepts and their interconnections. Together, they provide the required foundation to address the main research question in an integrated and coherent manner.

According to Johnson et al. (2020), an effective conceptual framework comprises three essential parts: (i) theories and/or concepts and principles relevant to the phenomenon of interest; (ii) what is known and unknown from prior work, observations and examples; (iii) the researcher's observations, ideas and suppositions regarding the research problem and question. These elements are presented in the following paragraphs to define the conceptual framework.

First, to ground this thesis based on existing knowledge, an initial scoping review was conducted through SRQ1, which highlighted the expansion of big data across various sectors: healthcare (Abouelmehdi, Beni-Hessane and Khaloufi, 2018; Awrahman, Fatah and Hamaamin, 2022), finance and fraud detection (Ravikumar, Murugan and Sriram, 2022; Aderemi et al., 2024). The review provided insights from big data applications in these sectors, such as its ability to advance operations, as well as several challenges. Building on these insights, the review extended to address the role of big data in policing, with several studies suggesting the expansion of big data into criminal investigations (Jin et al., 2015; Babuta, 2017; Broeders et al., 2017; Surbakti, 2020; Neiva, Granja and Machado, 2022; Neiva, Machado and Silva, 2023; Schuilenburg and Soudijn, 2023). This was further researched to explore its potential in serious crime investigations through the research question.

Second, the studies reviewed established the relation between policing and big data, although the extent to which big data could be useful in serious crime investigations remained unclear. To the best of my knowledge, no studies have exclusively focused on big data and serious crime investigations and consequently, this thesis researched the foundations of serious crimes (Home Office, 2013, 2023; Paoli et al., 2016; Almansoori, 2018; National Crime Agency, 2021; Winchester, 2020) and made the link to big data. Within the existing literature, the three most recent publications closest to the scope of this thesis concerned the studies conducted by Neiva, Granja and Machado (2022), Neiva, Machado and Silva (2023),

and Schuilenburg and Soudijn (2023). Collectively, these provide valuable insights into the role of big data in policing/criminal investigations broadly and frontline policing and intelligence more specifically.

This thesis focuses on serious crime investigations, examining their foundational elements, definitions, types and strategies. Neiva, Granja and Machado (2022) and Neiva, Machado and Silva (2023) examined the expectations and perspectives of policing professionals concerning the role of big data in criminal investigations in general, which produced findings that partially align with some of those in this thesis. These include the advantages of advanced policing operations and crime analysis and challenges such as privacy, data security and bias. Schuilenburg and Soudijn (2023) provided an overview of the current use of big data applications by the Netherlands Police in frontline policing, criminal investigations and intelligence. Partial similar findings were also found in this study, such as the potential of advancing criminal investigations, along with complexities in implementation and the need for qualified human resources. However, these studies present certain limitations, such as being broad in scope, lacking depth in the conceptualisation of criminal investigations and big data and not including the perspectives of big data professionals.

It should be noted that the research for this thesis commenced in 2020, and these studies were found in 2024, when exploring the latest literature to ensure the discussion would include the most recent developments. While these studies contribute significantly due to the scarcity of studies found that examine these areas, their scope is limited. Such limitations are understandable given that they are academic articles rather than doctoral theses, which can provide more comprehensive and in-depth research. Consequently, this thesis explored areas that have not been addressed by the existing studies reviewed in Chapter 2, thereby contributing to greater depth in the field.

Three, based on my professional observations, it was apparent that policing organisations possess high volumes of data but struggle to manage and transform these into actionable insights. This led to the idea of exploring whether big data might play a transformative role in policing operations, such as crime prevention and detection. Also, it was observed that some policing organisations attempt to use and adopt the latest technologies and tools promptly to advance their operations. This aligns with Neiva, Machado and Silva's (2023, p.217) view that their scoping review was timely, *"Considering the speed with which Big Data is advancing in the field of policing, and with police professionals increasingly being asked to use such technologies…"*. Beyond the potential advantages of applying big data and AI in policing, less attention has perhaps been paid to the broader context surrounding big data, such as the challenges it presents, its disadvantages and possible concerns. Therefore, they were explored as one of the aims of this thesis. I had

anticipated that a greater number of studies exploring serious crimes and big data would be found in the literature, but the scoping review highlighted the lack of such studies.

Finally, developing a strong conceptual framework facilitates the selection of an appropriate study method in qualitative research to help readers trust the research and the researcher (Johnson et al., 2020). This leads to the presentation of the methodology in the following sections.

3.3.2 Theoretical framework

Critical realism provides a broad explanation of ontology and epistemology, offering a comprehensive philosophy of science that provides a practical way of researching complex real-world problems (Sayer, 2000; Fletcher, 2016), in the context of this thesis, big data and policing. In this thesis the critical realist framework helps examine what is known about big data and assessing its potential usefulness in serious crime investigations in policing. It has the capacity to address the dual nature of big data initiatives as a technical artefact and policing as a social and organisational phenomenon.

The critical realist perspective takes the view that the nature of reality is not determined by our perceptions or theories; rather, all human knowledge about reality is provisional and contextually shaped (Sayer, 2000; Bhaskar, 2008). The central principle of critical realism is the separation between ontology, in terms of what exists, and epistemology, namely our knowledge of what exists (Sayer, 2000; Bhaskar, 2008; Fletcher, 2016). This thesis adopts the view that socio-technical structures, such as human resources and skills, culture, laws, ethics, datasets, analytical tools and software, exist and can produce real effects. Hence, outcomes are formed by real conditions that exist beyond individuals' opinions or descriptions (Bhaskar, 2008).

Ontologically, critical realism contends that reality is stratified into three layers or levels: the empirical, the actual and the real (Fletcher, 2016; Lawani, 2021). First, the empirical level concerns the domain of events that we experience and observe through perception or measurement (Fletcher, 2016; Lawani, 2021). Second, the actual level consists of events that take place whether we experience or interpret them or not; these events are often different from what is observed at the empirical level (Fletcher, 2016; Lawani, 2021). Third, the real level is the deepest domain and consists of underlying causal structures or mechanisms of objects or entities which are physical, social and internally related (Fletcher, 2016; Lawani, 2021).

Epistemology is a branch of philosophy that engages with the theory of the nature of knowledge and lays the foundations for how we as individuals understand the world we live in and the determinations we make (Sol and Heng, 2022). It concerns the process whereby the researcher asks how knowledge or what is assumed to exist is known and why we jointly decide certain things are true and others are not (Coe at al.,

2021; Sol and Heng, 2022). Epistemologically, critical realism accepts epistemic relativism which means that our knowledge of the world is incomplete and can be differently mediated through the descriptions and discourses that are available to us (Sayer, 2000). This thesis adopts an epistemic relativism approach and recognises that knowledge about big data and using it in serious crime investigations will vary depending on the professionals' roles, and that understanding is shaped by context and experience.

Epistemic relativism supports an approach that takes participants' accounts seriously while not assuming that any single point of view offers a complete or final description of reality (Sayer, 2000). Guided by critical realism, this thesis values the participants' accounts as meaningful yet incomplete insights, evaluated based on their explanatory strengths and coherence with observed reality.

Through this lens, our understanding of the world is shaped by existing theories and perspectives, but the world itself is not created or fully defined by them and an independent reality exists to be discovered (Fletcher, 2016). Methodologically, this method supports the use of the participants' accounts and existing literature as important sources of insights, while recognising that they provide a perspective based on interpretations rather than direct access to reality. Also, it justifies the comparison of different perspectives and sources of evidence to build a credible explanation.

A main aim of critical realism is to explain social events by identifying the underlying causal mechanisms and presenting how their effects appear across the different layers of reality (Fletcher, 2016). Therefore, rather than describing the potential advantages of utilising big data in serious crime investigations, this framework offers a means of moving beyond the surface level to undertake an in-depth exploration of the challenges and concerns that need to be mitigated to enable successful implementation of big data usage. Within a big data policing context, the factors shaping its usefulness are the key conditions that determine whether big data can produce reliable and useful outputs, such as the availability of datasets, data quality, and technical, technological and organisational factors. These cannot directly be observed as isolated causes but must be inferred from how the applications of big data operate and are able to achieve the desired outcomes. For example, identifying their impact in practice can be undertaken by exploring various aspects, such as: the advantages they offer in advancing police tasks and operations during investigations and how they can be achieved; what human and technical resources are needed; how these tools fit within daily work. Therefore, exploring the participants' views and relating them to the literature is necessary to make clear claims about the potential usefulness of big data in serious crime investigations and the conditions that support or limit the usefulness of applications.

Critical realism argues that qualitative methods can produce rich explanations of the underlying mechanisms that shape the phenomenon under study (Lawani, 2021). In this research, semi-structured interviews were conducted to attain rich explanations from the participants' accounts concerning the areas explored. The inclusion of two participant groups, one with policing expertise and one with big data expertise, provided different perspectives that were important from a methodological standpoint. Based on their different experiences, the analysis was able to compare the empirical evidence provided by their accounts and develop explanations about the underlying social and technical mechanisms influencing the adoption of big data in the research context (see 3.4).

Consistent with the ontology and epistemology of critical realism, thematic analysis was used in this study to identify and develop key themes and sub-themes related to the phenomenon under study (Lawani, 2021), as set out in Section 3.6. As already noted, critical realism is a philosophical framework that is used in social scientific research to move beyond describing perceptions towards explaining conditions (Fletcher, 2016). Hence, usefulness was not treated as a fixed feature of big data and its related technologies as it can change depending on conceptual understanding, technical and operational constraints, and how benefits are balanced against challenges in policing. Accordingly, different participants and sources may provide different interpretations that reflect their position rather than a single point of view. Variations in outcomes are thus interpreted as context-dependent rather than inconsistencies. Overall, the aim of adopting critical realism as a theoretical framework is that it moves the analysis beyond description towards explaining the potential usefulness of big data, for whom and under what circumstances.

3.4 Sampling strategy and overview

During the first stages of designing a research project, the researcher will consider the interview participants (Barrick, 2020). Different sampling strategies can be adopted before collecting qualitative data and each has its own aims and purposes which are appropriate to answer different research questions (Creswell, 2012). These include the following: maximal variation sampling; extreme case sampling; typical sampling; theory or concept sampling; homogeneous sampling; critical sampling; opportunistic sampling; snowball sampling; purposeful sampling; confirming and disconfirming sampling (Creswell, 2012, pp.207–209).

In qualitative studies, it is common to select participants based on *"purposeful sampling"* (Creswell, 2012, p.206), also known as *"purposive sampling"* (Johnson et al., 2020, p.141). This entails the researcher purposefully selecting individuals to explore a certain phenomenon (Creswell, 2012) or the *"...intentional selection of research participants to optimize data sources for answering the research question"* (Johnson et al., 2020, p.141). The aim of choosing this strategy is that the research question(s) may best be answered by individuals who have certain experience in the area of focus (Johnson et al., 2020). Thus, in this study,

the purposeful sampling strategy was used to identify and invite professionals to participate based on their experience and knowledge, enabling exploration of the concepts of serious crime investigations and big data.

3.4.1 Sample size

One-to-one interviews are among the most common data collection approaches in qualitative research. These are conducted in a form of conversation that involves the researcher, a participant and the themes of focus (Barbour, 2008; Moen and Middelthon, 2015). There are often fewer participants in qualitative studies than in quantitative research, since qualitative research typically focuses on gaining a better understanding of a phenomenon, particular issue or theme (Dworkin, 2012). Nonetheless, the sample size in qualitative research varies from one study to another (Creswell, 2012). In qualitative research, the guiding principle to assess the adequacy of purposive sampling is to reach saturation, the point at which no new information is generated (Hennink and Kaiser, 2022). While Dworkin (2012, p.1319) points out that *"An extremely large number of articles, book chapters, and books recommend guidance and suggest anywhere from 5 to 50 participants as adequate"*, Hennink and Kaiser (2022) suggest that saturation can commonly be achieved through 9–17 interviews.

For this thesis, the initial aim was to conduct 36 semi-structured interviews in total, 18 with professionals from the policing field and 18 participants in the field of big data. Out of the 74 individuals approached, the overall response rate was 23%, resulting in a total of 17 participants, 9 of whom were police officers and 8 big data experts; this was within the range considered adequate to achieve saturation by Dworkin (2012) and Hennink and Kaiser (2022). In addition, seven institutes were approached as gatekeepers (see 3.4.3).

3.4.2 Participants

At the outset, it was intended that the roles of the policing participants would vary, including policing commanders, detectives, analysts and academics, in addition to participants with backgrounds and experience in big data and AI, such as technology company directors, big data and AI specialists, analysts and academics. The rationale for this diversification was to ensure that the data collected would represent the perspectives of those at different organisational levels to generate in-depth understanding. The actual roles and areas of expertise of the professionals who participated in this thesis are given in Table 3.1, in which the policing participants are referred to as P1, P2, P3, up to P9 and the big data experts are referred to as A1, A2, A3, up to A8.

Table 3.1. Participant information.

Participant code	Role/Rank	Area of expertise
P1	CID officer	Artificial intelligence and facial recognition projects
P2	Police officer	PhD in computer engineering, formulating digital transformation
P3	Police officer	Engineering innovation department, safe city solutions
P4	Police officer	Drone applications in policing
P5	CID officer	Expert in cryptocurrency and virtual assets investigations
P6	Police officer	PhD in traffic engineering, artificial intelligence and smart cities
P7	CID officer	Senior lecturer in policing, organised crimes
P8	Chief constable	Former senior investigating officer of serious crimes
P9	Head of crime	Former senior investigating officer of serious crimes
A1	Professor	Computing, data science and cybersecurity
A2	Professor	Expert in digital surveillance technologies and public policy
A3	Professor	Electronic governance and policy implementation
A4	Assistant professor	Computer science, big data and machine learning
A5	Senior principal research scientist	Security and privacy preservation, machine learning and the internet of things
A6	Professor	Criminology, technology and law
A7	Chief technology officer	Data science and the ethical adoption of artificial intelligence
A8	Regional digital and engineering director	Big data and digital security solutions

The participants were assigned codes as identifiers and their roles and areas of expertise are presented in brief in Table 3.1 to protect their anonymity and prevent potential identification. A detailed description or provision of additional information about their positions, areas of expertise, professional qualifications and names of police forces/universities could unintentionally lead to deductive disclosure, revealing their identities, particularly with regard to their qualifications and scope of work. The participants were based in four different countries at the time of data collection, representing a geographically diverse sample across Australia, Singapore, the UAE and the UK.

To minimise the risk of potential bias in response patterns, the first step was to ensure diversity in sampling by approaching a range of participants (see Table 3.1 and Appendix B, which presents the roles of the invited professionals who chose not to participate, without disclosing any names or personal details) from various police forces, universities, private technology companies and other entities. This approach aimed to include perspectives from various contexts and backgrounds to increase the breadth and credibility of the findings. It resulted in interviewing participants from three police forces, six universities and two private technology companies. It is important to stress again that the names of the organisations will not be disclosed as it could compromise the participants' anonymity indirectly given the specialised nature of their roles and fields of

expertise. The purpose of assuring the anonymity of the participants was to reduce any pressure and encourage honest responses to avoid any favourable answers. In addition, to avoid the risk of potential bias in response patterns, the interview questions were carefully designed with neutral wording to avoid leading the participants towards taking a specific position. Further details regarding the development of the interview questions are provided in the following section.

3.4.3 Gatekeepers

In qualitative research, the researcher needs to gain permission from the individuals participating in the research and often any sites visited to collect data (Creswell, 2012). The process of interviewing participants can be extensive; therefore, it is useful to identify and use the assistance of a gatekeeper (Creswell, 2012). Creswell (2012, p.211) defines a gatekeeper as *"...an individual who has an official or unofficial role at the site, provides entrance to a site, helps researchers locate people, and assists in the identification of places to study"*. This study identified and approached seven gatekeepers with the aim of gaining access to both policing and big data professionals with a local and broad geographical reach. Four of the gatekeepers were policing institutes: the Cambridge Centre for Evidence-Based Policing, the College of Policing, Dubai Police and the Metropolitan Police in London. The other three were the Ministry of Artificial Intelligence in the UAE, the big data faculty at University College London and the UAEU Big Data Analytics Centre.

As I had sponsorship from the Dubai Police, I had support to reach out to the officers in the police force to identify participants and approach them for interview. In addition, the Dubai Police force has strategic partners and contacts with technology corporations in the public and private sector, both locally in the UAE and globally, which I was able to approach to interview participants and collect data. The Dubai Police yielded six successful interviews, but the other six gatekeepers/institutions that were approached did not yield any. However, the email correspondence with the Metropolitan Police did result in discovering useful information regarding their use of big data, reported in the findings in Chapter 4.

3.5 Qualitative research methodology

This study aimed to answer the research question and achieve its aims by collecting and analysing data using qualitative research methods. Data collection was undertaken by conducting semi-structured interviews and purposefully selecting the participants through theory or concept sampling. The semi-structured interview, its advantages and disadvantages, the interview guides, and possible ethical issues are also discussed (see 3.6 and 3.9).

Sahin and Ozturk (2019) argue that a methodology is required to produce scientific knowledge. This guides the researcher through all the steps, from creating a research question to the final findings. Qualitative

research methods are usually applied when the researcher is interested in attaining a better understanding of a specific topic from the participants' perspectives (Rosenthal, 2016). In this thesis, the aim was to have a better understanding of the role and potential that big data could offer in serious crime investigations from the perspectives of the intended interview participants.

Scientific research can draw on primary or secondary data (Kothari, 2004, p.95):

> *The primary data are those which are collected afresh and for the first time, and thus happen to be original in character. The secondary data, on the other hand, are those which have already been collected by someone else and which have already been passed through the statistical process.*

This thesis adopted a qualitative research methodology to collect primary data, considered the most appropriate approach based on the scoping review questions and research aims.

The researcher should decide the type of data that will be collected for the research and select one or more data collection method (Kothari, 2004). Among these, Harrell and Bradley (2009, p.2) note that *"Interviews can be used as a primary data gathering method to collect information from individuals about their own practices, beliefs, or opinions"*. Collecting primary data is an essential element in many studies and using the right technique is vital to ensure that data are collected in a scientific manner (Harrell and Bradley, 2009). Choosing the most appropriate data collection technique enhances the validity and reliability of the study (Harrell and Bradley, 2009). In qualitative research, the aim is to develop an in-depth exploration of the researched domains (Creswell, 2012). Qualitative research methods offer different approaches for exploring participants' experiences and different practices (Moen and Middelthon, 2015).

Accordingly, due to the novel domain researched in this thesis, the scarcity of prior studies and the need for in-depth exploration, a qualitative research approach was be applied to collect primary data, enabling the thesis to address the central research question and achieve its aims.

3.6 Data collection

3.6.1 Semi-structured interviews

There are different forms of interview methods that are applied to collect data to explore different phenomena: structured, semi-structured and unstructured (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). A structured interview is when the interviewer asks the interviewee a set of standard and pre-established questions in a specific order about a certain topic (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). In a structured interview, the interviewees are allowed to answer from a list of a limited number of responses and therefore it is seen as a rigid form of data collection in which all

the interviewees are asked the same questions in the same order to provide a brief answer or to select an answer from a list of responses (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011).

An unstructured interview is informal and does not have specific guidelines or pre-established questions (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). In an unstructured interview, the interviewer asks the participants several broad questions to engage in an open informal discussion. This type of interview is useful to get different narratives about the participants' experiences when the interviewer does not have specific questions (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011).

Semi-structured interviews fall between the structured and unstructured types and are the most common approach in qualitative research (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). A semi-structured interview involves a prepared set of questions asked in a way that allows the participants to answer in their own words and share their own experiences (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). To ensure that all the participants provide responses relevant to the topic under study, interviewers use a set of identified themes and questions in a systematic manner. This type of interview is considered more flexible and open to clarification than other formats (Easwaramoorthy and Zarinpush, 2006; Sandy and Dumay, 2011). This study employed semi-structured interviews as it they offered advantages in terms of contributing valuable data within the scope of the research.

A semi-structured interview is an effective method for collecting data as it is grounded in a human conversation, which enables the researcher to modify the form, pace and ordering of the questions to collect the most useful responses from the participant (Sandy and Dumay, 2011). One of its most important features is flexibility, allowing the participants to respond using their own terms, expressions and language (Sandy and Dumay, 2011; Wilson, 2014). Moreover, semi-structured interviews can provide support when addressing complex themes (Wilson, 2014). As this thesis progressed, complexities in the researched areas were found and the interviews were able to help understand the areas of serious crime and big data from the participants' perspectives and their own experiences. Indeed, Barrick (2020, p.403) defines a semi-structured interview as *"…a guided conversation in which a researcher enquires about how research participants understand their social worlds"*.

Sandy and Dumay (2011) note that semi-structured interviews require proper planning before, during and after the interviews, especially in the way that the questions are asked and interpreted. A semi-structured interview combines structured and unstructured aspects by having prepared and predefined questions together with open-ended exploration (Wilson, 2014). As Wilson (2014, p.24) states, *"The general goal of*

the semi-structured interview is to gather systematic information about a set of central topics, while also allowing some exploration when new issues or topics emerge". Thus, in qualitative interviews, researchers ask open-ended questions that allow the participants to best voice their opinions and experiences without any constraints (Creswell, 2012). These attributes led to the choice of semi-structured interviews in this study.

Despite the strengths of semi-structured interviews, they have some limitations. For instance, the presence of the researcher could affect the participants' responses so that they provide answers they believe the researcher wants to hear (Creswell, 2012). The researcher needs to be careful not to give hints or suggestions during the interview that could guide the participant to provide certain answers (Wilson, 2014). Even so, the researcher's posture and body language can affect participants' responses. For example, if the researcher yawns, it might indicate to the participant that the researcher is not interested in what he/she is saying (Wilson, 2014). Thus, it is recommended that the researcher has some training beforehand to avoid putting words in the participant's mouth (Wilson, 2014). For this research, I undertook the LJMU Research Ethics Training before conducting the interviews. In addition, the researcher's background, sex, age and other demographic characteristics can have an impact on the information the participants are willing to reveal during the interview (Wilson, 2014).

Adams (2015) points out that the process of preparing, setting, conducting and analysing semi-structured interviews is time-consuming and requires considerable effort, especially when it comes to analysing the huge volume of data which often comprises hours of transcripts. Moreover, if the interview is recorded, it is possible to have issues with the equipment, which can affect the recording and transcription (Creswell, 2012).

3.6.2 Interviewing guidelines

To assist in ensuring that semi-structured interviews are conducted in the proper manner and to avoid any procedural or ethical flaws before, during and after, Creswell (2012, p.222) proposes an interviewing procedure checklist. Some of the elements address who will participate in the research, the most appropriate type of interview and obtaining consent from the participants (Creswell, 2012). In relation to this thesis, the participants' fields of expertise were identified, the type of interview was selected and after completing all the ethical requirements and before conducting the interviews, informed consent was obtained from the participants. In addition, Creswell's (2012) guide highlights the importance of avoiding leading questions, asking open-ended questions and withholding judgments about the participants' views. Moreover, the researcher should make sure that the interview setting is comfortable and quiet and thank the participants after concluding the interviews (Creswell, 2012).

Wilson (2014, pp.30–34) also proposes several steps to avoid any mistakes in developing and undertaking semi-structured interviews. He emphasises the importance of the first few minutes of the interview, when the researcher should make a good first impression by being calm, confident, knowledgeable, flexible and professional; this can lead to the success of the interview (Wilson, 2014). Also, it is important to determine the goals of the research and the topics that will be discussed. The researcher should develop a list of questions that will be asked (Wilson, 2014). Wilson (2014) proposes that researchers create an interview guide consisting of several steps to aid the interview process, such as introducing oneself, explaining the goals of the interview, assuring confidentiality and setting out the way the data will be used.

Creswell (2012, p.229) notes various issues that can possibly occur in the data collection and interview phases, such as facing difficulties in scheduling an interview, avoiding any interruptions during the interview, and making sure that the recording equipment is working properly if the interview is going to be recorded. It is important to be confident during the interview, learn to listen carefully rather than talk, and be prepared to handle emotional outbursts (Creswell, 2012). The purpose of considering these guiding steps is to assure the data collection process and avoid any procedural or ethical flaws.

3.6.3 Developing the interview questions

The research aims and key concepts were identified from the literature reviewed in Chapter 2, which guided the formulation of the interview questions, developed after a reflective process to ensure relevance to the scope of this thesis. The interview questions were designed to explore the research question (What is known about big data and can it be useful in serious crime investigations?).

In qualitative research, the ability to compare responses is one of the criteria informing the selection of diverse research participants, seeking to enhance the solidarity of the research findings by incorporating varying perspectives (Lindsay, 2019). Since this study explored two areas of focus, serious crimes and big data, separate sets of interview questions were developed for each participant group, one tailored for policing (13 questions, see Appendix C) and one for big data big data expert participants (14 questions, see Appendix D), although built on a similar thematic basis. The slight difference in the number of questions resulted from the need to explore additional technical themes with the big data participants. The intention was not to conduct a comprehensive and detailed comparison between the two groups of participants but rather to explore the various aspects in which the participants had particular expertise, broadly maintaining common questions shaped by the themes and findings from the literature reviewed in Chapter 2.

The following section presents the development of the interview questions designed to answer the research question and achieve the research aims, as well as outlining the slight difference between the two sets of question.

Similar interview questions

As the literature indicated that definitions of serious crime and big data are debated (see 1.1.2 and 2.4.1), the interviewees were asked to define these two terms to explore their perspectives. Also, given the numerous characteristics of big data proposed (see 2.4.3) and the ambiguity in terms of what distinguishes data from big data (see 2.4.1), the interview questions sought to gain insights concerning these issues. Moreover, as the study aimed to discover if big data could be useful in serious crime investigations, the participants were asked about their opinions on the application of big data in policing, particularly detecting serious crimes and suspects. The literature identified various key factors in terms of advantages, challenges and concerns (see 2.5.2 and 2.5.3) around using big data in policing in general and serious crime investigations in particular and thus the participants were asked about these areas to see how these theoretical insights aligned with their practical and academic experiences. Analysing the literature (see 2.5.5), it was evident that the use of big data in serious crime investigations would likely necessitate identification of the types of datasets, variables and technological tools for employment. Consequently, the interview questions addressed these issues. The description of the development of the interview questions did not include numbering them, that is to prevent duplication and potential confusion arising from the overlapping questions across the two sets; for further details, see Appendices C and D.

The wording of the interview questions was carefully reviewed before they were submitted for ethics approval from the LJMU to avoid leading and minimise the risk of response bias. For example, one of the open-ended questions for both participant groups was as follows: “In your opinion, what do you think of using big data in the policing field?”. Starting the question with “In your opinion” aimed to encourage the participants to state their personal views, avoiding any bias, followed by an open-ended question that sought to collect rich data. Another form of question that aimed to encourage the participants to state their opinion started with “Do you think…?”, as well as seeking further elaboration by following up with “If yes, why and how? If not, why?” (e.g. “Do you think big data can be used by police forces specifically in the criminal field? If yes, how? If not, why?” and “Do you think that big data can be effective in detecting serious crimes and/or suspects? If yes, why and how? If not, why?”). This allowed the participants to express both agreement and disagreement without leading them. The remaining questions were designed to be more general and open-ended, starting with “What…?”, “Do you think…?”, and “Are there any…?”, giving the participants the freedom to share their perspectives and expand on areas that might not have been fully addressed in the literature to enrich the depth of the data.

Focused interview questions

This section addresses the differentiated questions in the interviews. "P-Q" refers to the interview questions addressed to the policing participants and "BD-Q" to the big data expert participants. While most questions were similar for both groups, built on the same thematic basis that merged serious crimes and big data, as discussed above, some questions were differentiated slightly. This applied to four policing questions (P-Q2, P-Q3, P-Q9, P-Q12) and four big data questions (BD-Q9, BD-Q10, BD-Q11), which focused on areas that were perceived to be more relevant to the expertise of the particular participant group.

The policing participants were asked where their police force was in terms of using big data (P-Q2), which was not considered relevant for big data participants as they were presumed to be working in universities and private companies. Also, the policing participants were asked to define serious crimes and big data P-Q3, while the big data expert participants were asked to define big data only, as it was assumed that they might lack policing knowledge due to their data science academic backgrounds. P-Q9 aimed to gather insights from the participants regarding their views about the future of big data in the criminal field, seeking to address their expectations and predictions. Since several strategies related to SOC were identified in section 1.1.4, P-Q12 aimed to explore if there were successful applications or strategies related to the use of big data in serious crime detection/policing.

For the big data expert participants, the questions concerned if they proposed any solutions to the challenges of using big data (BD-Q9) due to their technical background. Moreover, as the big data participants were expected to come from different backgrounds, BD-Q10 aimed to explore if they perceived big data analytics to be effective in subject/discipline fields other than policing and if so, how policing might learn from these applications. Also, due to their expected technical backgrounds, BD-Q11 asked about any AI tools that could be useful in big data analytics and if they presented any potential or challenges for policing.

The interview schedules were developed through an evolving process, revising the initial drafts to form two sets of interview questions that were both comprehensive and considered the participants' backgrounds in relation to the scope of this thesis (see Appendices C and D).

3.6.4 Conducting the interviews

After completing the LJMU Research Ethics Training, a UREC Ethics Application Form for studies associated with low risks was submitted to and approved by the Research Governance Assessment (Reference 22/LCP/005). Identifying and inviting participants and conducting the interviews took place from March to September 2022.

The participants were mainly approached through an email, which contained an introduction to the research, a participant information sheet, the interview questions and a consent form (see Appendix E). For those who replied saying they were interested in participating and signed the consent forms, an interview was scheduled and conducted. First, several big data participants were interviewed to gain insights into their perceptions of its definitions and characteristics, as well as advantages, disadvantages, concerns, challenges and tools, aligned with the themes found in the literature. Having developed a better understanding of big data, interviews were undertaken with the policing participants to explore the definitions and types of serious crime, the potential of big data, and challenges and concerns. These areas were identified from reviewing the literature and the themes that emerged and were explored further to allow analysis in light of the perceptions of the participants. The findings from the review of the literature and the empirical data were then critically analysed and are interpreted in the discussion in Chapter 5 to answer the research question, achieve the research aims and provide practical recommendations and suggestions for future academic study.

Semi-structured interviews are not only conducted physically face to face; they can be conducted, for example, over the phone or through an online meeting (Barrick, 2020). For this study, four interviews were conducted in person (P1, P3, P8, and A1), while nine were conducted online (P2, P4, P5, P7, P9, A2, A3, A4, and A7), based on the participants' preference. Also, due to the low response rate from the 74 professionals overall who were invited to participate, the option was given to be interviewed by answering the interview questions and sending them back by email. This alternative approach is called e-mail interviewing and is used to collect qualitative data (Opdenakker, 2006; James, 2007). It resulted in four e-mail interviews, one with a policing officer (P6) and three with big data professionals (A5, A6, and A8). Using email interviewing as an alternative approach can still give participants a voice and generate narratives that represent their thinking and experiences (James, 2007). In addition, they have the advantage of providing extended access to participants without geographical constraints, unlike face-to-face interviews (Opdenakker, 2006). Moreover, the interviewee can answer the questions at his or her own convenience and their choice of place and time (Opdenakker, 2006). However, while the four email interviews conducted in this study did yield useful findings, they provided fewer elaborative details and less context compared to the face-to-face and online interviews.

Finally, regarding the policing and big data professionals who were invited to participate but chose not to, they had vital roles and valuable experience in their fields and it was anticipated that their participation would have led to additional new findings. Like those who did participate in the interviews, those who did not were invited by email but either did not respond (the majority) or agreed to take part, then rescheduled and did not respond after that. Several follow-up emails were sent but did not get any response. In addition,

several professionals were identified and contacted through the LinkedIn platform based on the relevance of their experience for the scope of this thesis, but this approach did not result in in any interviews or lead to the collection of any data. Nonetheless, it would be important to invite them to participate in future research as they have substantial knowledge and experience in their fields. Therefore, their professional roles and areas of expertise are presented in Appendix B.

3.7 Data analysis

3.7.1 Thematic analysis

This thesis adopted thematic analysis to analyse the findings from the interviews, conducted to better understand the concept of big data and explore its potential use in serious crime investigations. This section presents this approach in detail.

For a study to be perceived as trustworthy, the qualitative researcher must demonstrate and disclose how the data analysis process has been conducted with enough detail to enable the reader to determine the credibility of the process (Nowell et al., 2017). One qualitative data analysis method is thematic analysis, introduced by Braun and Clarke (2006; see, also, Nowell et al., 2017). Braun and Clarke (2006, p.6) define thematic analysis as *"... a method for identifying, analysing, and reporting patterns (themes) within data"*. A theme identifies important areas within the data collected in relation to the research question(s), representing a recurring pattern or meaning within a dataset (Braun and Clarke, 2006).

Thematic analysis is commonly used in qualitative research as it offers a flexible approach to analysing qualitative data and the interpretation of various aspects of the area researched (Braun and Clarke, 2006; Nowell et al., 2017). In addition, thematic analysis is perceived as a useful research tool that has the potential to provide a rich and detailed analysis, by identifying, organising, describing and reporting the themes found in a data set (Braun and Clarke, 2006; Nowell et al., 2017; Kiger and Varpio, 2020).

As noted by Bryne (2021), Braun and Clarke have pointed out that researchers who adopt their approach should include the latest relevant publications beyond their 2006 article. Consequently, this thesis drew on Braun and Clarke (2020, 2021), Bryne (2021) and McLeod (2024).

3.7.2 Reflexive thematic analysis

There are three approaches to thematic analysis: coding reliability approaches, reflexive approaches and codebook approaches. It is recommended that researchers specify the type of thematic analysis used (Braun and Clarke, 2020). This thesis employed reflexive thematic analysis as it was deemed the most suitable methodological approach based on the following considerations. This approach involves themes developed from codes and conceptualised as patterns of common meanings underpinned by a central concept (Braun

and Clarke, 2020). Reflexive thematic analysis is useful to address exploratory research questions (McLeod, 2024), as in the case of this thesis, namely:

"What is known about big data and can it be useful in serious crime investigations?"

The reflexive approach, developed by Braun and Clarke (2006, 2020; see, also, McLeod 2024), consists of six phases: familiarisation with the data; generating initial codes; generating initial themes; reviewing themes; refining, defining and naming themes; writing up. These are presented in turn below.

1. Familiarisation with the data: The first phase seeks to gain an overall, thorough understanding of the data and includes transcribing audio recording if necessary and engaging in repeated readings of the transcripts (McLeod, 2024).

2. Generating initial codes: The second phase involves systematically identifying sections of data that are relevant to the research question as initial codes to help organise and categorise the data (McLeod, 2024). The coding process is a subjective, unstructured process that requires the researcher to engage with the data to develop an in-depth understanding (Braun and Clarke, 2020). Coding can be done by two methods: assigning semantic codes, which are identified from the explicit or surface meaning of the data, or developing latent codes, which go beyond the descriptive surface level of the data in attempt to identify deeper meanings and ideas (Bryne, 2021; McLeod, 2024). The decision to use semantic or latent codes, or a mix of the two, depends on the available data, as both types are valuable in thematic analysis and can contribute to an insightful analysis of qualitative data (McLeod, 2024). This coding can be done manually, by highlighting a paper transcription, or using software (McLeod, 2024).

3. Generating initial themes: The third phase transforms and clusters the initial codes in broader patterns to generate potential themes (McLeod, 2024). Theme development requires analytical and explanatory work from the researcher because they do not exist independently but are generated by the researcher from the previously identified codes (Braun and Clarke, 2020; McLeod, 2024). Here, researchers are encouraged to engage in reflexive thematic analysis, which acknowledges that theme generation is a creative and active process rather than claiming that themes emerge (Braun and Clarke, 2021; McLeod, 2024).

4. Reviewing themes: The fourth phase involves evaluating and reviewing the initial themes against the coded data as a quality check to ensure that the themes are representative. In this phase, the

researcher may also need to refine, discard, or generate new themes based on the evaluation process (McLeod, 2024).

5. Refining, defining and naming themes: The fifth phase consists of naming and developing a clear definition for each theme to clarify its scope, boundaries, and its contribution to the overall analysis (McLeod, 2024).

6. Writing up: Finally, the sixth phase encompasses writing up the findings and integrating the themes to present a coherent narrative of the data (McLeod, 2024). Writing up should not only describe the data but should also include insightful interpretations that relate the findings back to the research questions and connect them to the existing literature (McLeod, 2024).

In reflexive thematic analysis, where the researcher plays an interpretive role in analysing the data, the analysis can be inductively led or deductively led, depending on how the researcher initially approaches the dataset (DeJockheere et al., 2024). A researcher who adopts a deductive approach produces codes and generates themes by critically and flexibly applying a theoretical framework to the data (Bryne, 2021; Boyd, 2024). In a deductively led approach, the researcher starts with a structured codebook based on existing literature and theory, while remaining open to new and unanticipated concepts that do not fit the initial codes derived from the theoretical framework (DeJockheere et al., 2024). The deductive approach interprets the data collected to evaluate if the coded segments align with the themes suggested by the theoretical framework (Boyd, 2024), in the case of this thesis, comparing the findings from the data to the findings from the literature. Consequently, the researcher may need to add new codes to the coding scheme to ensure that the analysis accurately represents the data (DeJockheere et al., 2024).

This study conducted a comprehensive reflexive thematic analysis following the six stages in the framework outlined above. The themes and sub-themes were derived through a deductively led approach guided by a pre-established research question and concepts derived from existing literature. The familiarisation phase involved multiple readings of the transcripts to ensure a sound understanding of the data and this was followed by manual coding rather than using qualitative data analysis software to ensure closer engagement with the data. Both semantic and latent coding were applied where appropriate, depending on the nature and richness of the available data and the extent to which they could be interpreted. While the interview questions focused on areas of relevance identified in the literature in terms of the perceived advantages, disadvantages, challenges and concerns related to using big data, the participants were not presented with a list of these aspects. This allowed the participants to raise additional and/or unexpected points that were not captured by the initial framework and the analysis thus remained open to concepts that went beyond the

themes drawn from the literature. This approach reduced the risk of leading the participants to articulate predefined advantages, challenges or concerns, enabling them rather to present their own accounts in their own terms.

As the analysis progressed, the initial codes were established and examined to identify patterns and relationships, which were then transformed into broader themes and sub-themes that resulted in the identification of six main themes presented in detail in Chapter 4. The patterns and recurring themes concerned debates around the definition and concept of big data, financial, technical and human resource challenges, and privacy concerns, which were identified both in the existing literature and within the collected data. These themes were derived through an in-depth deductively led analysis guided by prior theoretical understanding which allowed a meaningful connection to be made between the established knowledge from the literature and the participants' perspectives.

Finally, regarding the quality of the study in terms of reliability, McLeod (2024) argued that coding reliability is not appropriate in reflexive thematic analysis as it attempts to quantify and control subjectivity in a research approach that clearly values the researcher's contribution. Braun and Clarke (cited by McLeod, 2024) suggested that attempts to impose positivist constructs, such as reliability and coder agreement, on reflexive thematic analysis can weaken its core strength as it relies on a rich reflective analysis. However, this does not mean that reliability was overlooked. Instead, McLeod (2024, p.25) proposes that *"If the analysis is carried out by a single researcher, it is recommended to seek feedback from an external expert to confirm that the themes are well-developed, clear, distinct, and capture all the relevant data".* As the data collection and analysis for this thesis were carried out by a single researcher, regular meetings were held with the research supervisors as experts to ensure the reliability and rigour of the data collection, analysis and write up. During these meetings, I shared the drafts of the findings chapter with the themes generated for feedback and critical discussion. The aim here was not to conduct inter-coder reliability (McLeod, 2024) but rather to enhance the transparency, rigour and credibility of the data collection and analysis process through the supervisors' involvement.

3.8 Researcher positionality

In this section, I discuss my viewpoint and positionality as a police officer in relation to constructing the research questions, the interpretation of the literature, sampling and analysis.

In line with the critical realist epistemology, my professional background and experience in policing influenced developing explanatory interpretations that went beyond surface level descriptions. My identity as an insider (police officer) and outsider (academic researcher) shaped different stages of the research, such

as framing the focus of the thesis, developing the research questions, interpreting the literature, sampling and data analysis.

This exploration of the usefulness of big data in serious crime investigations was informed by a practical need that I encountered in the field during my professional and operational practice, which informed the development of the question around usefulness. I was awarded a scholarship to undertake research into this phenomenon. My positionality and experience helped me identify the areas important to explore (big data and serious crim investigations) and how to frame the research question based on the evolving findings while remaining within the scope of the research (for the development of the research questions, see 3.2).

My background and experience played a role in how I read and understood the existing literature in relation to the policing field. It allowed me to relate academic topics and debates to practicalities in the field, which provided a foundation for the thesis based on an understanding of the concepts of big data and serious crimes. This was followed by exploring the advantages, challenges and concerns in relation to the use of big data in serious crime investigations as these areas emerged and were identified in the review of the literature and in light of my own experience, recognising that these aspects are important to decision makers in assessing the potential of big data in this context. Also, attention was paid to exploring ethics and legislation due to my professional awareness of their impact and their important roles were confirmed by the results of reviewing the literature and analysis of the empirical data.

In relation to sampling, my insider perspective and positionality allowed me to identify relevant participants based on their specific roles within police forces, in addition to accessing gatekeepers, which helped me reach several participants. However, I remained aware of potential biases and the need to avoid any negative influence when recruiting big data and policing professionals. It was important to remain reflexive when analysing the data so as not to allow my professional experience and assumptions to influence how my interpretation; the aim was to understand the participants' perspectives based on their own accounts and not let my background shape or skew their intended meaning. Also, my linguistic background, as a native speaker of Arabic and fluent in English, enabled me to conduct interviews in either language, facilitating my ability to establish rapport and communicate with a diverse range of participants.

3.9 Ethical considerations

The interviews were conducted after completing all the ethical requirements and obtaining the required informed consent forms to ensure that no ethical dilemmas would arise in recruiting and interviewing the participants or collecting and analysing the data. This section discusses potential ethical challenges and addresses how they were handled in this research.

According to Patton (2002a, p.405), *"Interviews are interventions"* and thus can affect the participants. Importantly, the researchers are required to protect the anonymity of participants, for instance by assigning numbers or aliases in processing and analysing the data and reporting the findings (Creswell, 2012). Bloom and Crabtree (2006) concur that protecting participants' anonymity is vital, since they might well share information in an interview that could jeopardise their position. Patton (2002b) provides a checklist that can act as a guide for conducting qualitative research and a means of evaluating the methods used. This addresses possible ethical issues when collecting qualitative data, applicable to the type of data collected through interviews in this study.

Patton (2002a) argues that qualitative research presents some unique ethical challenges because of its open-ended nature of inquiry and the direct contact between the researcher and participant. The seventh point in the checklist, *"Design the evaluation with careful attention to ethical issues"*, covers the ethical aspect of qualitative research (Patton, 2002b, p.11), considered at the different stages of data collection in this study. This includes the following: explaining purpose, promises and reciprocity, risk assessment, confidentiality, informed consent, data access and ownership, interviewee mental health, advice, data collection boundaries, and ethical versus legal (Patton, 2002b, pp.11–12). In addition, if interviews are going to be recorded, many ethical committees require an informed consent form that must be signed by the participant prior to the interview taking place (Bloom and Crabtree, 2006).

Another ethical concern when conducting interviews is to ensure that the participant has the intent to continue with the interview itself (Bloom and Crabtree, 2006). It is recommended that the researcher verbally ask the participant for his/her consent to undertake the interview and ensure the participant understands that he/she has the right to stop the interview and disengage at any time (Bloom and Crabtree, 2006).

The ethical issues presented here were considered and guidelines were followed while collecting data for this thesis, in addition to adhering to the interview guidelines from LJMU research ethics training. This sought to ensure ethical issues would not arise in the different stages of this research.

3.10 Summary

This chapter has presented the approach used to address the research question: "What is known about big data and can it be useful in serious crime investigations?" It has established the conceptual framing of the thesis by stating the need for the research, initiated based on my professional observation and examined with reference to the existing literature. The thesis adopts critical realism as the theoretical framework, supported by an epistemic relativist framing to reflect the view that knowledge is shaped by context, but

that different claims can still be compared and evaluated based on evidence and logical reasoning. This positioning provides a coherent basis to examine what is known about big data from the perspective of experts in the field and how it is conceptualised, in addition to its perceived usefulness in serious crime investigations.

The chapter has justified the adoption of a qualitative approach and set out the use of semi-structured interview with 17 participants sampled across two groups: nine police officers and eight with big data experts. Including these two groups strengthened the research design by enabling the comparison of perspectives, hence supporting credible interpretation by examining areas of agreement and differences across their accounts. In addition, a gatekeeper was acknowledged as playing a role in supporting the implementation of six interviews. In addition, the chapter provided rigour by articulating a clear account of how the interview schedules were developed, with separate sets of questions for the two participant groups. Careful attention was paid to the wording of items to ensure open-ended and non-leading questions that would allow the participants to express both agreement and disagreement.

Data analysis was conducted using reflexive thematic analysis, following the six-phase approach outlined by Braun and Clarke (2006, 2020). This provided a transparent structure to move from initial coding to the development of themes. This approach was made it possible to interpret patterns across the participants' accounts while remaining aware of the context, variation and meaning. In addition, the chapter acknowledged the researcher's positionality as part of methodological transparency. This is significant, as my professional background and experience informed my ability to engage with the research topics, develop relevant questions and interpret accounts with a degree contextual awareness, at the same time as maintaining reflexivity to exclude unexamined assumptions in the analysis.

Finally, ethical considerations were embedded through the research process, including training in adopting an ethical approach to research and obtaining formal approval from LJMU's ethics committee to proceed with data collection. The participants were then approached and for those who chose to participate, their identities were anonymised; however, their roles and fields of expertise have been presented to demonstrate the relevance of their contributions.

In the next chapter, the findings from the analysis of the empirical data are presented as themes and sub-themes developed through the reflexive thematic analysis.

Chapter 4. Findings

4.1 Introduction

This chapter sets out the findings of the data collected through semi-structured interviews to answer the research question:

"What is known about big data and can it be useful in serious crime investigations?"

The existing literature provided an initial theoretical foundation concerning big data, policing and criminal investigations in general. However, there was a lack of research on the potential of the application of big data in serious crime investigations and the extent to which the theoretical findings aligned with the day-to-day practices of professionals in the field remained unclear. Hence, the position of this study was that exploring the views of professionals operating within the fields of big data and serious crime investigations could offer contributing insights into how they perceive and interpret big data and serious crimes, as well as validating and/or challenging the findings from the literature.

Data were collected between March and September 2022. Of the 74 professionals and 7 institutes invited to take part, 17 participants agreed to be interviewed: nine specialised in policing and eight experts in big data and its related fields. In reporting the findings, the policing participants are referred to as P1, P2, P3 and so on (up to P9) and the big data expert participants are referred to as A1, A2, A3 and so on (up to A8). The themes and sub-themes were derived through a reflexive thematic analysis, following the five-step approach developed by Braun and Clarke (2020). These steps comprising the following: becoming familiarised with the data; identifying and reviewing sections relevant to the research question; generating the initial themes and reviewing them; refining, naming and writing the themes; writing up (McLeod, 2024).

The data were manually coded by hand following a deductively led approach. This was adopted as the review of the literature in Chapter 2 had highlighted key themes that provided a foundation for developing the interview questions. This enabled the data analysis to remain focused on exploring the predefined concepts, ensuring the findings would respond to the research question and aims. Full details of the methodology are provided in Chapter 3.

In this chapter, the findings from the analysis of the data are presented under six themes, quoting from the participants' interviews (for a full list of quotations, see Appendix F). This will be followed by the study limitations, particularly in terms of the data, before a summary of the findings. Table 4.1 provides a summary of the main themes and sub-themes derived from the analysis of the interview data. The

participants who contributed to each theme (and sub-theme where relevant) will be presented in the introduction to each one in turn in the following sections.

Table 4.1. Themes and sub-themes derived from the interview data.

Themes	Sub-themes
1. Defining big data	• Ongoing debate • Perspectives on definitions • Big data characteristics • Big data criteria
2. Big data and policing	• Big data and serious crime investigations • Types of serious crimes
3. Advantages and disadvantages	• Advantages of using big data in serious crime investigations • Disadvantages of using big data in serious crime investigations
4. Challenges in using big data in serious crime investigations	• Big data concept challenges • Financial challenges • Human resources and training challenges • Technical challenges • Operational challenges
5. Concerns regarding the use of big data in serious crime investigations	• Bias • Privacy
6. Tools and datasets	• Technological tools • Types of datasets

4.2 Theme 1: Defining big data

With regard to the definition of big data and understanding of the concept, there was debate and disagreement in the literature. Therefore, the participants were asked about their definitions of big data and what they considered to be its characteristics and the criteria for classifying it. This aimed to gain a deeper understanding and clear insights to address the first part of the research question, "What is known about big data?", and to achieve research aims 1 and 3. All the participants (100%) contributed and engaged in responding to Theme 1 and their perspectives are presented below.

4.2.1 Ongoing debate

The data showed that P2, P4, A1, A2, A3, A4, A6 and A7 considered that there was debate in the field concerning the definition of big data, representing 47% of the total participants. The fact that only two of the nine policing participants mentioned this may suggest a lack of awareness, whereas six out of the eight big data expert participants mentioned it, indicating greater awareness of this issue. This finding suggests that while the debate is highlighted in the literature, it is not particularly prevalent in policing practice.

Among the policing participants, P2 suggested that disagreement in the field and differences in defining big data could make it difficult for big data to be discussed, saying:

"Big data depends on its definition and characteristics, what is considered and not considered big data? There is a huge debate in the field, you cannot discuss something when we don't agree on the definition, and everyone has their own perspective."

P4 also argued that the definition of big data would depend on the angle from which you were examining it, as the definition would vary based on the perspective of the individual providing or using big data: *"To put a set definition, it depends on which side of the table are you, are you a person providing the data or using the data?"*

Among the big data expert participants, there was more awareness that this was a matter of debate and their definitions differed, as illustrated in the following quotes:

"There is no rigid definition of big data." (A1)

"I think I follow a traditional three V one." (A3)

"The generic definition of this is since the term big data was introduced, was referred to the increase in the volume of data that is difficult to store, process and analyse. So that definition was basically set in the early days when big data was introduced." (A4)

"This isn't the kind of term to be defined in standards." (A6)

"So we do not really have a clear definition of big data." (A7)

Also acknowledging that there are different definitions of the concept, A2 suggested:

"Okay, so you can find a textbook definition which will talk about the three Vs… I think I would unpack the definition of big data slightly differently."

4.2.2 Perspectives on definitions

The participants were then asked to provide their definitions of big data, and it was observed that there were two approaches to defining it. The first group/approach followed the traditional definition found in the literature, known as the three Vs: volume, velocity, and variety. The second group/approach comprised the participants suggesting their own definitions or expressing their understanding of big data.

In terms of the policing participants, two of the nine referred to the traditional three Vs definition:

"Big data is usually known by its three Vs: volume, velocity, and variety." (P5)

"… initially, what was considered the first three Vs of big data, volume, velocity, variety, and then we have got, you know, more recently you've got value and veracity." (P9)

P2 suggested that definitions based on the three Vs are restrictive and it is undetermined if a dataset would still be considered to comprise big data if it lacked one or more of the Vs. As P2 put it:

> *"For example, if a data set has some Vs but not the other, does it lose its big data characteristics? Some big data sets have volume and variety but not the velocity, is it considered big data or not?"*

Five of the policing participants defined big data using general terms close in meaning to the three Vs. Following are examples of the policing participants' definitions:

> *"Multiple input of information that can benefit us in understanding how, when, and what is the reason for things to occur and analysing data to understand the past to predict the future." (P1)*

> *"Big data from a technical standpoint is to use hardware that accepts data input from sensors and multiple sources which is then analysed by artificial intelligence to create relationships to support decision making." (P3)*

> *"Larger, more complex data sets, especially from new data sources." (P6)*

> *"High volume with high intensity data that has unknown value until explored." (P7)*

> *"Big data is the diffusion of lots of data from disparate sources which creates links that may identify people engaged in serious and organised crime or detecting a serious crime that is taking place which would not be otherwise known if the user did not have the ability to bring data together." (P8)*

Although the policing participants' own definitions used general terms, they appear to reflect the three Vs. For example, *"Larger…"* (P6) and *"…lots of data"* (P8) can be related to volume, *"data input from sensors and multiple sources"* (P3) can be related to velocity and/or variety, *"Multiple input of information"* (P1) and *"new data sources"* (P6) can be related to variety, and *"unknown value until explored"* (P7) concerns value.

Among the big data expert participants, two of the eight interviewees referred to the traditional three Vs, as follows:

> *"So you can find a textbook definition, which will talk about the three Vs, which is very much driven by a kind of data scientists' definition, I think the three Vs are something like the velocity variety and volume." (A2)*

> *"I think I follow a traditional three V one, I do believe that visualisation is also an important factor in my definition." (A3)*

Five big data expert participants suggested their own definitions and one (A6) did not contribute. Overall, the big data expert participants' definitions consisted of more technical and scientific terms, differing from the policing participants' definitions, which were slightly generic and broad:

> *"Big data is data that comes from many sources without any limit, such as the web or internet which is an open-source big data." (A1)*

> *"I think I would unpack the definition of big data slightly differently. So, I would say that what we should seek to do is to seek to understand how big data offers something significantly different to data processes than that has happened in the past." (A2)*

> *"Big data are relevant, connected data, and size is a term that is too abstract." (A5)*

> *"Big data is having fast moving and slow-moving data signals, fast moving signals are constantly updated like a phone location for example, whilst an example of slow-moving signals is a permanent address or a bank account number." (A7)*

In addition, A7 elaborated:

> *"Big data normally for me means using the data that an organisation holds and combining that with additional data in and building what we call data universes [which] are connected ecosystems."*

A4 contended that there is a generic definition of big data from when the term was introduced in its early days: it is the increase in the volume of data that is difficult to store, process and analyse. However, A4's own definition of big data was:

> *"Big data is a set of techniques and technologies that need a new form of integration to uncover large hidden values from large data sets that are diverse, complex, and in massive scales."*

Moreover, the big data expert participants used technical terms in their definitions, which can also be related to some of the Vs. For example, *"...data that comes from many sources"* (A1) can be related to variety, *"without any limit"* (A1) and *"...fast moving signals"* (A7) can be related to velocity, *"large data sets that are diverse, complex, and in massive scales"* (A4) can be related to volume, velocity and variety, and *"seek to understand how big data offers something significantly different"* (A2) and *"to uncover large hidden values"* (A4) can be related to value.

Both the policing and big data expert participants provided big data definitions that revolved around several Vs, such as volume, velocity, variety and value, either directly or indirectly. Despite the differences in their definitions, there was no indication that the participants were wedded to the view that a particular definition was correct or incorrect; rather, each participant defined big data from their own point of view. This

illustrates the challenge in reaching a definition that everyone agrees on, as views can differ, for example, in terms of what Vs to include or exclude.

4.2.3 Big data characteristics

Variations were also evident in the participants' views of the characteristics of big data. Several characteristics were suggested by both the policing and big data expert participants, which included and went beyond the three Vs of volume, velocity and variety. The findings again suggest that there is no conclusion when it comes to an agreed standard/set of characteristics that shape the concept of big data.

According to P3, *"With big data the quality of data is important, and this is what characterises big data"*. This participant considered that collecting data from different sources would make it big enough to create relationships between data to allow optimal decision making. In contrast, if big data were of low quality, it would not be useful in decision making.

Other participants took a very different view. For instance, P4 considered the current variables used to describe big data unstable, meaning that the definition of big data and its characteristics would be liable to constant change:

> *"If we restrict characterising big data with a changing variable that is constantly changing like the power of computers that we have, then the definition of big data will constantly change."*

The use of the word "we" seemed to refer generally to scholars and professionals with an interest in big data rather than limited to a particular profession or specific group.

P6 went beyond the traditional three Vs of volume, velocity and variety, and added: *"Veracity, validity, variability, volatility, visualisation, and value"*. As P6 was one of the participants interviewed by email, this response did not include any additional explanation. P9 also suggested that volume, velocity, variety, value and veracity characterise big data, broadly aligned with P6's contribution.

A4 contended that the main characteristic of big data is its volume, followed by secondary characteristics: variety and velocity. This was based on the view that big data comes from various sources in different forms – structured and un-structured – in high volume and needs to add value and be useful to the organisation. A4 highlighted value as one of the significant characteristics of big data, saying:

> *"So, if we have all these huge amounts of data, but we can't get insights from this data, then that might not be something useful to us. So value, it's also one of the main characteristics of data."*

In contrast, A7 argued that velocity is the main characteristic of big data, stating: *"It is normally velocity, which is how fast the volume is produced"*.

As can be seen, the policing and big data expert participants' views varied and there was no agreement on a definitive set of characteristics that would describe big data. Rather, they could comprise any of the following: quality, volume, velocity, variety, veracity, validity, variability, volatility, visualisation, and value. These characteristics were proposed by the participants as examples and they did not provide additional explanations for how they arrived at them. When discussing the increasing number of big data characteristics found in the literature, A3 suggested that this was common – indeed typical – in academia:

> *"That is academia for you isn't it, because we all like to become wordsmiths and invent new terms and expand existing terms to progress our own academic careers."*

A7 made the same point, stating:

> *"That is the beauty and pain of working in AI. Everybody makes something up to sound like an expert."*

It was established that the participants would consider one or more Vs as characteristics of big data, such as volume, velocity and variety, but not quality or value. This characterisation is linked to P2's argument that it is not clear if a dataset would still be considered to comprise big data if it lacked one or more of the Vs. While volume was mentioned as one of the features of big data, but the required amount was not quantified; rather, they used terms such as "high" or "huge" amounts of data. Moreover, P2 pointed out that what is considered a high volume of data today might not be in the future as data storage technologies develop. Therefore, if a dataset is considered big data because of its high volume and that volume becomes average due to developments in storage technologies, it is unclear if the same dataset would still be considered big data.

Overall, it appeared that the participants held different views of what they considered the main characteristics of big data. None dismissed other perspectives in the field but simply suggested their own point of view.

4.2.4 Big data criteria

Given that the debate around defining big data and determining its essential characteristics remains unresolved, the participants were asked if they could propose specific criteria that might be used to classify a dataset as big data. By inviting the participants to suggest their perceived criteria, it was possible to move

beyond the theoretical definition and conceptualisation of big data to draw on their real-life practical experiences.

A4 argued that complexity is a criterion of big data, regardless of data size: if the data were complex, the data set could be considered big data. With the developments in storage technologies, A4 considered that volume was becoming less challenging than complexity:

> *"The volume is getting actually less attention compared to the complexity which is the variety of data… So, if we have data that is of various sources, and with different forms, then we can say that this data is basically big data."*

A8 argued that if a data set ranged between 1 and 10 terabytes, depending on its complexity, it could be considered big data. This was the only instance in which a participant specified a certain size for a dataset to be considered big data:

> *"It's a little bit subjective, but we can say that starting from 1–10 TB depending on the complexity of the data (video, text, pictures, etc.) we can consider that this is big data."*

Hence, complexity, suggested as a criterion by A4, surfaced again in A8's proposition in terms of diversity of data format: *"video, text, pictures, etc."*. Although this is a partial explanation and complexity may extend beyond this, it reflects an important starting point.

A3, when asked whether there are criteria for classifying big data, said:

> *"No, I don't think, I think because the term is so much a slogan or a catchphrase, it can be used for any type of data set which is larger than normal."*

There was not clarification or quantification of what would be considered "larger than normal", making this ambiguous. It seems that A3 perhaps considered the term "big data" a surface level, catch all term, seeming to reflect the point made by A6 that *"This isn't the kind of term to be defined in standards so not sure that there is a criteria-based definition"*, which again was not explained in greater depth.

The views of P3 and A5 were aligned in considering relevance as a criterion. As P3 argued, *"No criteria at the moment, the main criteria is to be relevant"*, whereas A5 only suggested *"Relevancy"* as an answer to the question in the emailed interview. Overall, no further explanations about relevance were introduced, making it unclear as a criterion to be considered.

In terms of the perspective/criteria that considers big data to be data that cannot be handled by normal computers and may require super computers, A7 said:

> *"I wouldn't agree with that definition because most of the models are run on the cloud anyway. So if you say it cannot be handled by a computer, and then you kind of be like, 'Why are you not working on the cloud?'"*

Also, A7 also highlighted not restricting the definition of data based on the technology or infrastructure used, saying, *"I would say it depends on your architecture and would not define the data on the technology that you are using"*. As an example, A7 said:

> *"So for example, your phone is actually processing a lot of big data. When it does image recognition and other things, when Siri works on your voice recognition. It is actually big, big data AI models."*

Finally, partial agreements were noted in this aspect, with participants A4 and A8 agreeing on complexity, A3 and A6 agreeing that big data is not a term to be defined in standards, P3 and A5 agreeing on relevance, and A7 not restricting the criteria based on the technology being used. The differences observed regarding the definition, characteristics and criteria determining what constitutes big data illustrate the lack of agreement and stability in the understanding of big data.

4.3 Theme 2: Big data and policing

The participants were asked about their views regarding the use of big data in policing to understand their point of view broadly before narrowing the scope to address big data in relation to serious crime investigations. All the participants (100%) responded and there was overall agreement that using big data in policing would be beneficial for police forces, as it could enable them to better utilise their financial and human resources, become more data driven, improve police operations, and assist in decision making. P1 viewed the use of big data as highly important, saying, *"It is essential, significant, and a must, and to develop your police force you need to understand and use big data"*. Moreover, as P2 noted, *"With the global development, using big data in policing is considered a part of our job, and it is an important part of the digital world"*. These appeared to be common positions, linking the adoption of big data to the development of police forces. Similarly, P3 stated, *"Using big data in policing is very relevant and it can be done as any other field"*.

P4 gave an example of the use of big data and its advantages in their police force's use of drones to search for wanted individuals, which directly contributes to the safety of society. P4 added, *"I am a huge advocate of using big data in policing in general as it can improve police operations and lead to better results"*. P6 also provided an example of the police force's use of big data in relation to traffic accidents:

> *"We are using big data to investigate the traffic accidents to cluster accidents and analyse the most reasons of causing accidents, locations, time, nationalities, gender and other factors and present the*

trend of these factors. Based on that our department build their studies to draw the strategies to avoid traffic accidents."

While P7 saw both positives and negatives in the use of big data in policing, saying, *"Although it brings some challenges, it has great opportunities"*. Several participants highlighted various challenges, which are discussed further under Theme 4.

With regard to the big data expert participants, A1 considered that *"The use of big data in the policing field is definitely effective"*. A2 had some reservations, saying:

"I think the thing about new technologies or emerging technologies is that they are new by definition, and therefore we do not really know the impacts and consequences that they will have."

A2 did not refer to big data specifically, but rather to *"new or emerging technologies"*. A possible explanation could be that new technologies will be employed to utilise or analyse datasets and due to their novelty, it is difficult to predict all their consequences. A4 was not very experienced in the field of policing, but viewed big data as potentially useful, saying:

"So, object detection could be one of the applications of big data policing because it basically can help police to track and monitor individuals who are suspected."

A5 stated, "*Good, just like any other social and economic problems*".

4.3.1 Big data and serious crime investigations

The participants were first asked about big data in policing generally to establish a baseline understanding. This was then followed by asking if it could be effective and useful in serious crime investigations, thereby narrowing the scope to provide more depth and focus. In this regard, there was robust agreement among the policing and big data expert participants concerning the effectiveness of big data.

P1 expressed the view that *"Yes, it can be very useful"* and suggested several examples, given under Theme 3 (see 4.4). P2 also said that *"...big data can for sure be effective"* in detecting serious crimes. P3 viewed big data as potentially useful, stating *"It can be effective but it depends on how clean your data is as it can create bias in the outcomes"*.

P4 also said, *"I am certain that big data can be effective in detecting serious crimes and there are very positive results"*. Similarly, P6 stated, *"Yes, to know how the crimes happened and the ways to prevent these crimes"*. P8 suggested that *"...if it was applied within the legal constraints, its potential is enormous"*. P9 agreed that it could be useful but highlighted several challenges to effective implementation: *"I think [it]*

can and can be very effective, but it is having the resources, the equipment, and the staff in order to deal with and manage it" (see also Theme 4, 4.5).

Among the big data experts, A3 stated, *"I think it can probably be become very very effective. I also think there are some risks here"*. A6 responded positively, noting, *"Sure, facial recognition is an example of that"*. As pointed out by A7:

> *"It can lead to tremendous insights, we have used it successfully from fraud detection to demand forecasting, from optimisation of staffing to counter terrorism in all these fields, it can really help you getting insights".*

Here, the reference to "we" referred to the private company A7 worked for, which developed technological solutions employing big data for police forces. A8 similarly commented on the experience of the private company he worked for, which provided security services, stating:

> *"From our experience in …. we think big data can indeed support crimes and/or suspects detection but its need to gather data from the field (e.g. CCTV, gate entry, database coming from authorities, etc.). In order to support the data processing with crossing data from various sources and to support the algorithm learning, even if we also think we have to follow white box paradigm."*

4.3.2 Types of serious crime

The review of the literature in Chapter 2 established that there are differences in classifying the types of serious crime among police forces. Therefore, the participants were asked about the types of crimes that they considered big data to be useful for addressing in their investigations. The purpose was to bridge the findings from the literature and the perspectives and knowledge of practitioners in the field. The types of serious crime suggested as examples by the participants were as follows: all kind of offenses in public places, arson, burglary, cyber-crimes, cyber-bullying on social media, drug trafficking, financial crimes, fraud, murder, online sexual offending, rape, robbery, terrorism, and treason.

P2 considered that big data could be useful in all types of crime, saying, *"Big data is involved in every crime"*. In contrast, P8 suggested that big data could be helpful but would be more effective in certain areas:

> *"There are all sorts of types of fraud going on online, it could be really helpful in that field in particular but I think in all serious crimes. I think it has value and it has the potential to be exploitable particularly in fraud."*

From the big data analyst perspective, A8 viewed its applications as quite comprehensive, suggesting:

> *"AI that analyse big data can detect cyber-bullying on social media, all kind of offenses in public places thanks to CCTV, predict crimes (e.g. robbery in gas station) before it happen in some districts during a specific time slot, etc."*

Several participants named the types of crime without further elaboration, as follows: *"Cybercrimes, financial crimes, fraud, drug trafficking"* (P5); *"Terrorism, treason, arson, murder, rape, and robbery"* (P6); *"Burglary"* (A6); *"Fraud, drugs, online sexual offending"* (A7).

Finally, the participants agreed on the usefulness of big data in policing and serious crime investigations and this was consistent across both groups. However, they also pointed out that police forces should be cautious when adopting new technologies and set out several challenges, which will be presented in Theme 4 (see 4.5). However, the overall agreement on the effectiveness of big data analytics was robust.

4.4 Theme 3: Advantages and disadvantages

Building on the themes previously identified in the literature, the participants were asked their views on any advantages/disadvantages (Theme 3), challenges (Theme 4), and concerns (Theme 5) regarding the use of big data. As these themes are closely related to the second research aim and the research question, namely, exploring if big data can be useful in serious crime investigations, as proposed in the literature, through the participants' perspectives.

Participant A6's contribution to theme 3 was excluded as the email response simply provided references to five studies rather than articulating a position. P7's responses, in a relatively brief interview, did not contribute to this theme. The other participants' (88%) views are expressed in the following sub-sections.

4.4.1 Advantages of using big data in serious crime investigations

The policing and big data expert participants were asked if they perceived any advantages of using big data in serious crime investigations. P1, P5, P6 and A8 shared a common standpoint, suggesting that big data could be used by police forces to be proactive in predicting crimes and forming strategies to prevent them and being reactive by detecting criminals, patterns and high-risk crime areas. P1 argued that *"big data can be used in crime prediction, planning and forming strategies"*. In addition, P1's police force was establishing a safe city platform through video analytics with the use of AI, aiming to enable them to achieve these advantages. As P5 suggested, *"It can be used for the police force to be proactive and reactive"*, while P6 proposed that it could *"Detect, predict, and suggest best solutions to deal with crimes before happening"*.

According to A8, big data could be of use, saying, *"That is what we can retrieve from big data, learn from the past to better predict and anticipate the future"*. A8 went on to state:

> *"It also help us to recognise criminals and criminal behaviours by using deep learning algorithm and various pattern to target 0 crime and anticipate when it comes to predict offenses in a specific spatio-temporal window based on historic data. We can take advantage of big data for many use cases, from the need to store huge number of statistical data in order to identify the district with the*

higher risk to estimate automatically a criminal path on hours of CCTV video coming from various sources."

While these participants emphasised the potential to be both proactive and reactive, P3 said, *"Policing is very responsive and not proactive, we do not have systems to predict crimes and it all depends on how clean the data is"*.

P3, P4, A3 and A7 shared a common perspective in considering that using big data could make an organisation more data driven, leading to better decision making, making it possible to discover patterns and crime hotspots, and enabling the proper utilisation of resources. As P4 noted:

"Data in the policing field is very important, it allows us to react quickly in a more efficient manner, overall refines the process and allows you to get better results."

Similarly, P3 stated, *"It makes the organisation more data driven to better utilise resources…it can automate the process of having a system that takes decisions by itself in the future"* and went to say, *"It can help discovering crime hotspot heat maps, find patterns, and study them to build better prevention programmes"*. In addition, P3 suggested that humans have capability limitations in rapidly analysing high volumes of data sets to find patterns, *"…humans have limitations, we cannot look at different data sources and make quick conclusion or try to find a pattern and build relationships, whereas big data with AI can do that in a fast way"*. A3 agreed with P3 regarding the capability of identifying patterns of criminal behaviour through analysing big data and contended that *"you can identify patterns for example criminal behaviour"*. Whereas P3 highlighted the high speed that these analyses can be obtained.

A7 gave an example of a big data project their technology company developed for a police force in the UK to identify hotspots and direct their patrols, recounting:

"I remember 10 years ago when the whole big data started, we built a demand forecasting for UK police force to basically understand exactly where the hotspots and where should they put their police cars."

Regarding patterns, A7 argued that there are several algorithms that are effective:

"A lot of prediction optimization and grouping algorithms are extremely effective, when we look at for example fraud detection and similar kind of cases, you are trying to find the needle in the haystack."

A7 added:

"So one of the things that we see is that every human being has a pattern, and just like normal people have a pattern, so do fraudsters have a pattern, so do criminals have a pattern, and there is actually a lot of interesting use of policing information and policing data for retail. So yeah, it is a very good field for the police to invest in."

P2 suggested that searching a big data system to discover leads in criminal investigations could offer benefits in terms of time, saying, *"If I suspected someone and I do not have a big data system, I will be searching in every system individually such as the ANPR, car registrations, etc."*. P2 visualised a big data system connecting all the databases that detectives need during their investigations, enabling a comprehensive search and more efficient use of time "*...in a click of a button"*.

P8 and A2 shared this perspective, contending:

"So, the advantages are obvious that you can search masses of data technically and get answers to questions very quickly. They might take you weeks, months or even, never, in most of this investigation." (P8)

"So I think the advantages of big data are that advances in computing being that you can process such large quantities of data that were not previously possible for mankind. So, you can take into account you know, vast datasets that are beyond the sight of human comprehension, you know, they their data is massive." (A2)

Similarly, A5 stated, *"Yes, at least potentially. It can help understand better about events"*.

A7 went into considerable detail concerning the implementation of big data projects:

"In the long run, it always makes sense to have people in house because the policing sector is a specialist sector, and you could then basically build up your own teams. That is at the beginning, very expensive, so you invest in the right people, you invest in the right technology in the right training, and you need to make sure they have the access to the data, and they actually need to have a certain level of confidentiality."

A8 proposed two use cases in which big data and AI could be advantageous to the police in their investigations:

"Use case 1: We can have what we started to recall in the last question: operators lose a lot of time to watch about ten hours and more of video to try identify and catch a suspect when he was seen making an offense. Using an AI trained thanks to a huge volume of data, we can learn how to identify a same individual on multiple images and calculate his path based on CCTV camera's locations automatically without mobilizing an officer to do it.

Use case 2: Other example, thanks to geolocation data collected from a smartphone's suspect, we can map and identify many aspects of its life like his occupation, where he lives, his gym location, etc., in order to predict some behaviours."

Finally, both P9 and A1 agreed that despite the challenges and concerns (addressed in Themes 4 and 5), the advantages of big data outweigh its disadvantages for police forces in serious crime investigations. These views reflected the overall agreement among the participants, recognising the potential it could deliver:

> *"To the police the advantages are more than the disadvantages." (A1)*

> *"I think that the advantages far outweigh the disadvantages." (A9)*

4.4.2 Disadvantages of using big data in serious crime investigations

The participants were asked if there were any disadvantages to using big data in serious crime investigations. The analysis showed that most framed them as challenges (Theme 4) rather than disadvantages. As A2 suggested, *"I think, maybe disadvantages is not quite the right word, maybe challenges is better"*. However, a few disadvantages were suggested by the participants.

P8 argued that it was a mistake to think that big data was the perfect solution in policing:

> *"I think it will be a mistake to think it's the answer to all of our problems. You know, there is always going to be a role for traditional policing, talking to people doing other types of investigative tactics such as physical surveillance."*

However, P8 also said, *"[I] can't see many disadvantages, other than the logistical limitations that we've already discussed in terms of people and systems"*. In contrast, A4 suggested several disadvantages that were also outlined as technical challenges, related to the high volumes of data and the tools used to analyse them, which will be presented further in theme 4.

According to P4, *"A drawback is if you have untrained data scientists or untrained algorithms you will provide wrong data"*. This is related to the perceived harms proposed by A2, i.e. *"bad big data science"*, which could lead to discrimination or biased outcomes, although A2 also stated, *"I am not saying that would happen, it is the risk of that happening"*. This was also articulated as a potential for bias and profiling concern by several other participants, presented further under Theme 5 (see 4.6).

4.5 Theme 4: Challenges in using big data in serious crime investigations

The process of developing this theme was interpretive and generated through my active engagement in analysing the data collected from the participants, 82% of whom contributed to it. The identified sub-themes indicated difficulties in understanding the concept of big data, along with issues with financial and human resources, as well as technical and operational challenges. Addressing the challenges as well as the advantages was necessary to convey a balanced and comprehensive understanding, as focusing on the

advantages only would risk overlooking the hidden challenges that could obstruct the effective and useful implementation of big data in serious crime investigations.

4.5.1 Conceptual challenges concerning big data

Several participants pointed to difficulties in understanding the concept of big data in some organisations. As noted by P3, *"The science of big data and artificial intelligence is new, and the culture of big data is still not there"*. In terms of policing, P3 suggested that some senior commanders might not be familiar with big data and *"…so they must be shown the capabilities of big data and how it can actually improve decision making"*. P3 recommended that exposure was needed to spread the understanding of big data. P4 considered that the challenge went beyond big data and encompassed the computing field as a whole, stating:

> *"The issue is not about big data…but the lack of understanding of how computing works and what are its subfields, people see one person on a computer and expect him to do all kinds of computing jobs."*

P4 suggested that more awareness about computing and its subfields was needed and emphasised that *"…training should not only be for the employees but will be beneficial for the upper management as well"*. This relates back to the point made by P3 that senior commanders also need to be exposed to the capabilities of big data, leading to a better understanding.

A4 argued that fear of change could also be a challenge for some organisations, saying *"There are some organisations or even maybe the police, they feel more comfortable in the way they operate than changing to new technology"*. A4 suggested that introducing new technologies could change the way they operate, leading to some resistance. In line with P3 and P4, A4 considered exposure and training valuable and added:

> *"The people in charge should provide all the type of training that helps these officers to not only know about big data but also learn some of the technologies and some of the techniques that can help them do their work better."*

However, P3 also noted, *"There is a resistance to change from seniors or employees that think the automation can replace their jobs"*.

4.5.2 Financial challenges

A number of policing and big data expert participants contended that building a big data project to be used in serious crime investigations could be expensive and required substantial funding. This could be a challenge for police forces lacking the necessary funding to develop the required infrastructure and utilise advanced technologies. Thus, they would continue to rely on traditional methods of policing. As P1 noted, *"The technologies cost is very high"*.

P2 outlined the concept of return on investment (ROI), an idea that to the best of my knowledge did not appear in the literature reviewed in relation to policing/serious crime investigations. P2 stressed the value of analysing data for policing as not all data can be stored and analysed. P2's perception was that there was an issue in determining if a system could be useful for criminal investigations with the rising costs of technologies:

> *"For example, if there was a terrorist attack and it costed the country around 1 billion in business losses, whereas the police had a chance to purchase a system for millions that would enable to prevent this attack, then it will be a good investment."*

In P2's view, it would be difficult for police forces to make this assessment, as they might purchase advanced technologies for millions, but would not know the ROI unless serious crimes occurred and in some cases they might not. Hence, P2 argued, *"An equation is needed to evaluate the need of big data, if the system is for criminal investigations, is it a good system even if it solves one case? Is this correct based on its high costs?"*. Also concerning the issue of value, P4 queried, *"If it saves people lives, can we put a cost on it?"*.

P4's expressed the view that many would hope for possessing these costly systems in an ideal world, but with the financial challenges and possibly other barriers, it could be difficult to achieve fully. P8 gave an example illustrating the difficulties, noting that funding for police forces in the UK was facing budget cuts and this created limitations that *"... are preventing the exploitation of big data"*. In addition, P7 pointed out that some police forces do not always have access to public funding for the big data technology they need, which might lead them to work on projects with the private sector.

Among the big data experts, A7 said that big data and AI are indeed expensive, but that in this field, cost is not the sole consideration:

> *"It is something that is not quick and easy, it is something that is actually expensive and requires a lot of amount of skills. That is why it is actually a field where if you go with startups, or if you go with cheap vendors, you run a massive risk because you are dealing with sensitive information."*

A7 added, *"...just building AI is really cheap, building trusted AI really building good AI is really expensive and needs a lot of expertise"*. A4 also addressed this, stating, *"Financial matters are a core of big data challenge, because investment in these kind of projects require a huge amount of money, which some organisations might not really like"*. This was reflected in a comment made by P8, who said, *"I express frustration [about] being able to see the potential of big data but because of financial and other restraints not being able to exploit it to its full capacity"*.

4.5.3 Human resources and training challenges

The participants highlighted the need for qualified human resources to develop and operate a big data project and presented this as a challenge:

> *"It is not easy to find the qualified specialists in the field of data." (P1)*

P4 also recounted that having employees with a *"lack of training can be a challenge"* and highlighted the need to *"...put the correct resources in the correct operation"*. P7 similarly said, *"Finding the resources to manage and search for big data is challenging"*.

From the big data experts' perspective, A4 concurred, saying, *"There are many projects that failed due to lack of having the right people to deal with data"*, linking back to P4's point that it is necessary to allocate the appropriate resources for operations. A1 argued that the future of big data projects will require *"...skill sets and specialists in big data for transformation"* and stressed that *"...it is a challenge not a disadvantage"*.

Making a somewhat different argument, A2 stated:

> *"I think there are challenges for public services to acquire the skills to use big data effectively. Now that cannot be underestimated because data scientists are in high demand and they are all getting all the good ones are getting swallowed up by commercial companies."*

While having qualified human resources was mostly framed as a "challenge" by the other participants, A2 viewed it as a "genuine concern", highlighting its significance and the urgency of addressing it:

> *"But I think the concern that I would bring forward, which may be different to other people would be around you know, the skills required around data science or how policing will get that skill set to use data science and big data properly. I think that is a genuine concern, and I know that that is happening across the public sector."*

In relation to this, P8 suggested that getting into the field of big data was one thing, but having the human resources to properly use it was another:

> *"But often you have got really good intelligence and you have got no resources to do anything with it. So, exploiting big data is one thing and then been able to do something about it is another thing".*

The human resource challenge can also be considered an operational challenge, as the police already possess the data but lack the resources to turn them into valuable insights.

4.5.4 Technical challenges

The analysis indicated possible technical challenges related to big data in terms of its forms, the required infrastructure, data reliability, quality and storage time. The following observations were highlighted by the participants as technical challenges that could face police forces when utilising big data.

P2 argued that one technical challenge concerned data retention; there are solutions, such as saving it on the cloud or in data warehouses, but the main point is until when?

> *"One of the technical challenges is how to save the big data and until when, I cannot save all the data forever, and what software will analyse all this data?" (P2)*

P2 implied that this could lead a police force to select the datasets needed from big data and not save all the data. P2 gave Twitter's big data as an example, saying *"For example, can we save all Twitter's data since it was created? And will all of it be useful? If the answer is yes, we need huge infrastructure investments"*.

P2, P3, P4 and A4 agreed that there were technical challenges related to the variety in data formats, the need for data cleaning, and finding the most practical and useful datasets. P2 and P3 suggested that having different data formats from various sources could be challenging. In this regard, P2 suggested that *"With every new system we need new infrastructure and so on".* P2 added that thousands of videos are uploaded daily on the TikTok platform, for example, and asked what video analytics software could analyse all of them to detect a crime, identifying this is also as a challenge.

Also, P4 argued that usually with big data the data can be unfiltered and require cleaning, which is the process of extracting useful data from big data and is a significant process:

> *"…data cleaning is checking for missing values, and it is the process of using raw data to refine it to data that can be used afterwards."*

P4 added that in some policing projects, finding the most practical and appropriate data can be challenging. Moreover, A4 noted that volume is one of the challenging aspects of big data, in addition to variety, as big data comes in different formats (structured and unstructured) and in different forms:

> *"So volume is basically one of the challenges in big data, but now with a revolution of internet we have also data that comes in various forms of video audio and text… we are no longer dealing with only structured data we have to deal with unstructured data and velocity is also becoming a challenge because we are getting data more than one can imagine."*

P5 argued that *"the biggest challenge is ensuring the data is correct and of high quality"*. P5 suggested that there should be quality assurance at different stages of data collection and entry points. In the same vein, P3

highlighted the need for a data governance framework that entities (including policing) could use to facilitate data collection and determine what is considered of high quality. This is related to A2's point that one of the challenges for public services is the reliability of data. While P3, P5 and A2 framed data quality as a *"challenge"*, P4 viewed it as a *"concern"*, stating, *"My concern with data is the source"*, as problematic sources and non-clean data could be misleading and harm investigations.

P8, P9 and A4 consider that high volumes of data created a technical challenge for policing. As P8 noted from experience, *"So the forces held a huge amount of data but it was not joined up as they did not have the technical capability to export all of that data"*. P8 was referring to police forces in the UK facing this technical challenge, which was also related to the financial challenges highlighted earlier. P9 suggested that if a police force decided to collect big data, they should have the tools and infrastructure needed to handle high volumes:

> *"So if you are going to start collecting big data as a police force or as the national agency, to effectively deal with the data that you do collect, then you are going to have to invest in some kind of equipment to deal with it. If you are not going to deal with it, then what is the point in having it?"*

However, A4 suggested that with enhancements and developments in storage technology, such as data warehouses and cloud technologies, data volume is becoming less and less challenging:

> *"So in terms of the size, because now we have machines with huge space that can accommodate huge amount of data and also we have cloud technologies."*

Rather, the challenge now concerned the ability to reduce the size of data without compromising its quality:

> *"The other disadvantage of having huge amount of data is also the reduction of the data, how do we reduce the size of data without compromising the quality of the data that we have?"*

A7 highlighted a technical challenge causing a security issues not only for police forces but also countries around the world related to linguistic issues. A7 argued that this especially occurred in the Middle East, which involves systems that are built in English and face difficulties in translating to Arabic. A7 gave the example of identifying individuals entering a country, since the same name in Arabic can be spelled in multiple ways in English:

> *"That is always a big challenge for all countries around the world to know who is coming in and is it really the person who is saying that it is? The other big problem that we have especially in the Middle East is the translation between Arabic and English. A lot of the systems are built in English and cannot deal with Arabic language, and also you have 50 different ways of spelling Mohammed, with two m's with one m, with a's with one a, because it depends on who translated it when the data was entered."*

Furthermore, A8 stressed that there is a technical concern in securing the high volumes of big data, especially private data, due to the increasing threats of cyber-attacks:

> *"There is big concerns on how to secure huge volume of critical private data in a connected world with is more and more threat with cyber-criminality and cyber war."*

4.5.5 Operational challenges

The analysis of the challenges identified by the participants also revealed insights in terms of operational challenges. One such was facial recognition, which A1 argued could be fundamental in some investigations, saying, *"For example, facial recognition, in some cases if you cannot have it you cannot investigate"*. Similarly, P2 described data as a *"double-edged weapon"*:

> *"…not having data is a challenge, having excess data is a challenge, it becomes like it is not there because you will be lost and will not benefit from it."*

The issue of excess data was also raised by P8:

> *"But one of the huge challenges the police face is they have got more intelligence and information and data than they can cope with…lack of intelligence was never our problem, it was having sufficient operational resources to act on the intelligence."*

In terms of managing large volumes of data, P9 contended that one of the challenges was ensuring that critical information that could lead to progress in an investigation was not missed, especially in time-critical investigations. P9 described such critical information as *"…that golden nugget",* referring to its significance. Regarding time-critical investigations, P9 also suggested that a possible challenge for the police in the UK was collecting certain types of data legally, such as phone calls, given the lengthy process:

> *"Of course the other issue is for police and law enforcement in the UK, particularly is to get the lawful access. You have got to have the authorisation in place and that authorisation process is quite cumbersome, require quite burdensome with regard to the administration."*

P7 contended that the various challenges, such as lack of human resources, technology and huge amount of work, investigators might avoid using big data.

A2 pointed to additional challenges in terms of putting data-sharing protocols in place between different entities, including the police, stating, *"I think there are challenges in terms of building data-sharing protocols, so there has to be some sort of protocol governance structure in place whenever the police share data with somebody. So that is a challenge"*.

Furthermore, P2 and A2 suggested potential operational challenges for policing in the future, particularly highlighting the role of digital policing. P2 questioned if the metaverse could be considered a source of data and whether verbal assault and sexual harassment occurring between characters in the metaverse, for example, would still be considered crimes, noting the absence of legislation in this regard:

> *"In future, is the metaverse a source of data? Is verbal assault and sexual harassment in a metaverse considered a crime? With no known identities and no legislation, the characters assaulted each other, not a known person. What if they were known to each other and took their revenge in the real world?"*

In a somewhat different vein, A2 pointed out that *"The world of criminality massively moved online during the pandemic, and obviously policing has to go online as it is going to be a massive area of future activity"*, adding *"...online is the new frontline policing"*.

4.6 Theme 5: Concerns regarding the use of big data in serious crime investigations

Considering the participants' concerns in parallel with the advantages of using big data allowed the study to reflect not only on the opportunities but also the possible risks that could result from using advanced technologies in policing. The two main concerns that were highlighted by the participants and shaped the following sub-themes were bias and privacy.

4.6.1 Bias

Bias and profiling concerns were highlighted by P3, A2, A3, A7 and A8 (29% of the participants), who emphasised the importance of addressing these, especially when using big data and AI in policing. The participants' concerns were founded in the potential risks of incorrectly identifying or arresting individuals during criminal investigations due to flaws in the data being fed to the algorithms, negligence in building AI models and insufficient auditing.

According to P3, *"Bias is the main concern and the main reason is how clean is the data"*. Similarly, A7 contended, *"...our world is extremely biased, and our data is biased, that is the main concern that I have"*. A3 weighed the positives and negative, stating:

> *"I think it can probably become very effective but I also think there are some risks here. One obvious risk which has been addressed several times is profiling."*

Moreover, A7 raised another concern:

> *"Also, the second concern is that people are lazy and build models under pressure…and some AI models are often not audited, often built by somebody maybe a junior who does not really understand the impact."*

In referring to "*does not really understand the impact"*, it seemed that A7 was alluding to the consequences for a police force of taking decisions based on a biased analysis provided by the algorithm, which could affect an individual or a community. An example of such a situation was provided by A2:

> *"I think this is where police forces have got into difficulties in recent years because they have not thought through some of these ethical consequences, and face recognition is an example of that. So the Metropolitan Police and South Wales Police have been trialling face recognition in public spaces. There has been a lot of criticism about them being ineffective, about bias. And it was legally deemed at the High Court that what they were doing was not legal at the end of the day."*

Another example was provided by P3 concerning the risks of prediction using biased data:

> *"In some police forces projects the AI software or algorithm is biased, for example in the US where they did a trial project which had outcomes that most of the criminals are from a certain race, and that is because the data that is fed to the algorithm is not clean so it creates bias."*

While P3 nonetheless viewed big data and AI as having considerable potential, this was conditional: *"It is effective if the data sources are clean"*. The issue of clean data was also established as one of the technical challenges identified in Theme 4, indicating the interrelatedness of challenges and concerns.

A3 gave several suggestions for dealing with big data projects:

> *"I think it is important that you have first of all legal provisions, but also to actually try to roll out some kind of impact assessments, but you have some frameworks, and you have some guidelines, so people are actually enlightened about the risks, to me it is very much an educational challenge."*

In this regard, A7 suggested that there should be certain requirements, but did not expand on what they were, saying, *"So there are a lot of factors that we say these are the key minimum for ethical, trusted AI"*. Notably, A8 argued that algorithms may not recognise all ethical concerns, making it challenging to apply the outcomes of analysis:

> *"Moreover, any algorithm doesn't understand all ethic concerns, so the output of this data processing could be no ethical at all for human using it."*

A8 also suggested that it will take time to refine the use of big data:

> *"The big data needs also some time to be enough accurate which could not be relevant in some cases or need to have an organisation proactive to invest on such topic."*

A7 expressed a similar view and proposed the following:

> *"Yes, so AI could be audited. AI could be tested for biases could and transparency, explainability and so on. It is still open research and people are working on it, and the problem at the moment is*

we do not have enough people to build the AI, so how are we expecting to find enough people to actually audit that on top of it."

A7 did also note the existence of tools to detect bias, stating, *"So we have toolkits to actually analyse models to try to open up the black box"*, but went on to say, *"Sometimes you can fix the bias, but sometimes you cannot. Sometimes you just don't have the information and no real way of fixing it"*.

Accordingly, A2's advice to police forces is relevant: *"So I think new technology should be used where appropriate by policing agencies, but there's always a degree of caution involved"*. This was in line with A7's view of the importance of caution in policing:

"These in policing as important as anywhere else, but I think especially important in policing…We now say entrusted AI has to be transparent, explainable, robust, safe, identity protection, and unbiased."

4.6.2 Privacy

Privacy surfaced as a common concern amongst both the policing and big data expert participants (P2, P4, P5, P7, P8, P9, A1, A2, A3, A4, A5, A7, and A8; representing 72% of the total), who stressed the importance of addressing this when developing big data projects to be used in serious crime investigations. From what they said, it was apparent that they were referring to the privacy of individuals in the community, whether innocent or suspected of a crime, in terms of the personal data held, used, or requested by the police during their investigations. An example would be a CCTV video in which members of the public appeared during a crime or searching through a big data set that involved people's personal data.

This was such an important issue that P2 stated when it came to the use of big data in serious crime investigations *"I have no other concerns other than the privacy"*. Similarly, A5's response to being asked about any concerns was simply *"Privacy"* and A1 also said that when it came to big data, *"A disadvantage to the people is their privacy"*. A4 elaborated on this issue, stating:

"One of the main disadvantages is basically the privacy because sometimes data that we collect could be a personal data especially in crimes and all those areas. We have to look at, for example, individuals' profiles and all this information so privacy is one of the big issues in big data."

According to P2, attitudes to the privacy of individuals variable, depending on the culture of the society. Thus, for example, CCTV and facial recognition applications are more acceptable and normal in some societies but are rejected and considered a violation of privacy in others. Hence, as P2 noted:

"Privacy of individuals can be considered as a floating concern and it depends on the culture of the society… In our culture having a camera is not considered a violation of privacy, whilst in other cultures it could be."

Here, P2 contrasted the culture of the UAE, where it is culturally acceptable to secure the community using cameras and facial recognition technologies, and practice in Western countries, where privacy controls tend to be high, with individuals having the right to delete their data from digital records. This corresponds to A2's view of prevalent attitudes in the UK and other parts of the world:

"In the UK having some face recognition is a brilliant example, there is a massive resistance across the world to having face recognition in public spaces."

Similarly, A8 stated:

"Moreover, in some countries, people are concerned on the use of their data less eager to accept that any companies or organisation like the police administration can have access to private information and track them in everyday life".

P2 and A1 justified the need to employ technology, with P2 remarking, *"You have a crime, you have to protect the innocent people privacy and discover the criminal, if there was total privacy and no cameras installed that might lead to not discovering the criminal"*. However, P4 questioned this, saying:

"Privacy is something we always take into account, would you rather have more privacy but risk safety and security? Or would you rather have more security but less privacy? It is always a balance."

P4 noted sometimes, privacy would take priority, remarking, *"In some projects like the beach project, we would let go of them just to let people have more privacy"*. P4, P8, P9 and A2 also argued for the importance of achieving balance, as illustrated in the following quotes:

"It is all about or it tries to be about balance, about protecting the public and protecting intrusion into people's private lives, it is all about trying to get that balance." (P8)

"I think you have got to be realistic, so there is got to be the balance." (P9)

P9 gave an illustrative example of this in the UK:

"In the UK for instance, right to privacy is a qualified right, and if you are involved in serious and organized crime depending on your level of criminality, then you forgo the right to privacy, but you have got to put a lot of consideration in to justify that from a law enforcement perspective, and I think in this country, I think the balance is just about right."

P9 highlighted that in certain instances, intrusion might be necessary but should be justified from a law enforcement perspective. P7 and P9 also noted the need for care in the case of intrusion into the privacy of individuals as collateral in a criminal investigation:

> *"Searching through big data in the digital environment may lead into privacy intrusion of innocent people and it should be justified." (P7)*

> *"It is got to be justified, it is got to be proportional, and it is got to be necessary." (P9)*

However, A2 took a different stance:

> *"Yeah, so I do not necessarily believe in arguments about balance. I think that I think we should strive to achieve security, safety and privacy. I do not buy into you have you cannot have both at the same time. I think that is a very old argument, going back 20 years, and misspelled over the years. And when people repeat that argument, you kind of think, okay, they have not got they have not developed a mature thinking around the technology."*

P5 and P9 viewed privacy as an ongoing concern:

> *"Privacy is a wide argument, and there are legislations such as the GDPR in Europe to ensure there is no data abuse." (P5)*

> *"There are always going to be the points if there is a concern and possibly that up against the arguments about privacy versus security." (P9)*

A3 discussed the potential measures to protect privacy, from current legislation to the potential for further measures:

> *"I mean, how can you actually protect privacy? I mean, part of it has obviously been all these data protection acts around the world. Establishment of data protection agencies or we call them privacy commissioners here, and I think we need something similar in artificial intelligence or big data."*

A3 also described an algorithm impact assessment introduced in Canada:

> *"I mean, there are also a growing interest in you know about privacy impact assessments. You must have heard about that. So, the Canadians have invented something called the algorithmic impact assessment. So, when you introduce a new system, prior to that you should actually both tick some boxes go through and check for any risks here, and how do we mitigate the risks? So, it is sort of what I would call a regulation by design really."*

Finally, A7 stressed the importance of protecting privacy in policing given the impact this can have on society, saying, *"Privacy protection for policing is more important than any other industry"*. This links back to A2's argument for caution on the part of the police when using new technologies:

"So I think you know when the police use a new technology, they have to be super cautious in terms of thinking about the impact it will have on a whole set of different relationships…so it is changing relationships."

4.7 Theme 6: Tools and datasets

Having established that big data has the potential to advance serious crime investigations, the participants were asked if there were any tools or specific types of datasets that could be beneficial in this regard. The exploration of these aspects is related to the second research aim.

4.7.1 Technical tools

The participants noted that technical tools would be necessary in resolving the challenges and concerns identified in Sections 4.5 and 4.6 and employing big data thereafter. P2, P4, A1, A4, A5, A6, and A8 (41%) contributed their perceptions and view concerning this sub-theme.

The participants suggested several examples of AI tools and techniques they considered useful in analysing big data for serious crime investigations, as follows: anomaly detection, clustering-based techniques, deep learning, document analysis, facial recognition, link analysis, natural language processing, pattern recognition, predictive modelling, social network analytics, text and image analytics.

P4 noted, *"The most useful AI tools depend on the programming language, and there a lot of tools and algorithms that are available, it also depends on the type of data you have"*. A1 also highlighted the availability of various AI tools and their usefulness in dealing with large volumes of data, for instance in financial crimes:

"A successful application of big data analysis in financial crimes for example if 20 computers were seized, a lot of time and resources are needed to analyse all of them. Therefor the use of AI and its tools will be more effective, such as pattern recognition, link analysis, document analysis, and natural language processing".

A4 articulated the benefits of deep learning, particularly for unstructured data:

"One of them basically and the most popular adopted technique is deep learning, and the reason for that is because it really works very well with unstructured data. Also, one of the advantages of deep learning in big data analytics is self-learning. As a part of reinforcement learning approach, it basically helps also in tackling some problems of huge amount of data."

In addition, asked about the techniques that could be useful in policing, A4 and A8 proposed the following:

"So clustering-based techniques is very useful in policing, it helps in clustering segments of group of individuals or to identify crime hot spot areas that can basically be covered by police patrols…

Predictive model is also one of the interesting techniques that can be used in the policing field, and it was recently used because of COVID-19." (A4)

"Clustering and anomaly detection." (A8)

A5 considered it challenging to list the most useful AI tools, stating, *"It is hard for me to list out the so called most useful AI tools, there are many AI tools used in cybersecurity and most of them have their own unique features"*. A6 noted, *"Depends on your definition of AI, but facial recognition, text and image analytics, social network analytics"*. P2 and P4 also viewed facial recognition as valuable, with P2 giving the example *"FR can be helpful in searching for individuals"*.

4.7.2 Types of datasets

The participants (P1, P2, P4, P6, P7, P9, A1, A2, A4, A7, and A8; representing 64% of the total) pointed out that knowing what type of data would be necessary for the application of big data in serious crime investigations. They proposed the following: CCTV videos, communications data, crime historical data, criminal and traffic system, emails, emergency service calls, government records, IP addresses, live camera feeds, mobile geolocation, open-source intelligence, phone numbers, social media, tax records, and telephone behaviour.

P1 and P2 both noted that having data is fundamental for investigations. P1 pointed out that *"Data and information is essential to investigate; without data you cannot do anything"* and proposed useful databases, such as the traffic and criminal systems. P2 remarked, *"Text is highly important if it came from a trusted source, then audio and video…"*.

P1, P9, A4 and A8 suggested that social media could be of value:

"Social media." (P1)

"There are going to be huge data on social media." (P9)

"We have social media." (A4)

"Social network data (Tweets, Facebook, Instagram)." (A8)

However, P2 and A2 argued that social media might not be a trusted source of data:

"Facebook might not be trusted because of people impersonating another people." (P2)

"So if you look around social media and Facebook and Twitter there are millions of fake accounts." (A2)

According to P4, *"Firsthand data works better but it is very expensive to collect and takes a lot of time"*, presumably referring to primary data collected directly from the user. P6 proposed the use of *"Live camera feeds, emergency service calls, and complaints registered with the police"* as datasets and sources that could be useful in serious crime investigations. In addition, P7 suggested *"Phone numbers and IP addresses"*, adding that *"...numerical data has the greatest value"*. Similarly, P9 noted, *"The first one is obviously communications data…the contents of emails are potentially huge, but again, it is looking for that tiny piece of critical information or that piece of evidence in a massive volume of data"*.

Like P9, A7 referenced communication data, specifying, *"Telephone signals and segments and I think the number one I would go for is telephone behaviour"*. A7 suggested that the users' telephone behaviour could reveal information useful during investigations for geospatial location and network analysis. Likewise, A8 suggested *"Mobile geolocation data, CCTV videos, Car plate and all police data legacy"*.

In addition, from the perspective of the big data experts, A1 suggested the value of *"open-source intelligence, crime historical data and past reports"*. Moreover, A2 noted that, *"Public services are used to working with their administrative datasets that are very reliable you know, so if we look at the tax records, they are very accurate"*. Similarly, A7 stated, *"I think their own records the government holds on like the employees' information"* as an example of datasets.

4.8 Limitations

This section outlines the limitations that should be considered when interpreting the findings of this thesis. Clarifying these limitations aims to provide transparency and ensure the robustness of the results. The limitations mainly concern the rapid pace of developments in big data and its related technology, the timing of the empirical data collection and the professionals who were invited but did not participate, all of which may have shaped the content and breadth of the insights presented.

4.8.1 Rapid technological developments and publication gap

The first limitation is related to the fast-paced developments in big data and its related technological fields. There is a considerable lag in the publication of academic papers and by the time this thesis is completed and made publicly available, some of the findings, for instance the challenges identified, may partially have been addressed in practice. Nonetheless, the most recent literature still highlights these key challenges, which suggests that the findings remain relevant.

4.8.2 Timing of the empirical data (2022)

The second limitation relates to the timing of data collection. The semi-structured interviews were conducted in 2022, which is a considerable time ago given the rapidity of developments in the field. It is

therefore possible that some practices and organisational measures have changed since the data were gathered. As a result, the findings of this thesis should be interpreted as a reflection of the participants' views and practices at the time, rather than final accounts of the existing situation.

4.8.3 Nonparticipation of key professionals

The third limitation is associated with the professionals who were approached but eventually did not participate in data collection. Several of these professionals held senior and specialised roles (see Appendix B) and appeared to have substantial experience in the fields of serious crime investigations and big data. Their perspectives might have introduced supplementary or even contrasting insights that are not captured in the data analysed. Although these professionals did not participate, this does not undermine the value and robustness of the findings from the data collected. The two groups of participants had relevant roles and substantial experience in the fields under study and offered valuable insights that address the research question.

4.9 Summary of findings

This chapter has presented the findings from the reflexive thematic analysis of the 17 semi-structured interviews with two participant groups, comprising nine police officers and eight big data specialists. Overall, the findings show strong support for the potential usefulness of big data in policing and especially in serious crime investigations, but they also demonstrate that usefulness cannot be assumed; indeed, the participants repeatedly argued that the usefulness of big data depends on the availability of suitable data and the capacity to manage and analyse datasets.

A key finding is that the foundational concept of big data was not consistently defined across participants. In Theme 1, 47% of the participants explicitly recognised an ongoing debate in terms of defining big data, with big data specialists (6/8) being more aware of this than policing participants (2/9). At the same time, all the participants demonstrated a baseline understanding of what big data is, either defining it in relation to the three Vs or using their own terms, thus indicating a shared recognition at a practical level even when the language used differed. However, the varied views on certain characteristics of big data, such as value and data quality, and on whether it is possible to define a fixed set of criteria for big data reinforce the importance of reaching a clear operational definition to support consistent interpretation for policing.

Themes 2 and 3 showed broad and consistent agreement that big data can support policing, for instance by improving decision making, enabling more efficient use of resources and supporting serious crime detection, prediction and prevention. These perceived advantages were linked to proactive approaches, faster access to relevant information and an improved ability to identify patterns and hotspots. The reported advantages

centred on faster investigations and improved ability to search large volumes of information, suggesting that the participants viewed big data as a capability that could support operational efficiency, for example in identifying leads during investigations. The suggested potentials position big data as supporting decision making to advance investigative capacity rather than being a replacement for professional investigative judgement. While the advantages were strongly articulated, the disadvantages highlighted were framed as challenges in implementation, suggesting that the participants viewed them as barriers that could be potentially addressed.

Themes 4 and 5 provided greater depth to this picture, illustrating that the challenges and concerns are substantial and may affect whether the benefits of big data usage can be realised in practice. Cultural resistance, financial pressures, a shortage of skilled human resources and technical difficulties with infrastructure in relation to data storage, data cleaning, data quality, security and cyber risks were described as challenging factors. Operational challenges were also highlighted, especially in terms of the risk of missing critical information among the high volumes of data and managing excessive data with a lack of human resources. In parallel, bias and privacy emerged as central concerns. Bias was recognised as a serious risk, mainly linked to the poor quality of training data, which could lead to individuals being wrongly identified, negligence in building models, or models being built by junior employees and underestimating the impact of biased outcomes. Privacy was highlighted as more prominent concern, with the participants focusing on the protection of personal data, justifications for intrusions into privacy and ensuring their legitimacy in terms of needfulness and proportionality and aiming to achieve a balance between security and privacy.

Finally, Theme 6 identified a wide range of AI tools and types of datasets that could support policing in undertaking serious crime investigations. This supports the position that the implementation and adoption of big data applications should be guided by their purpose, data availability, capability and risk management.

Overall, the findings provide a strong empirical foundation for the next chapter, in which the six themes will be examined critically in relation to the literature and interpreted through the critical realist framework to identify the conditions that confer usefulness.

Chapter 5. Discussion

5.1 Introduction

This chapter critically interprets the findings from the analysis of the data in relation to the research question, aims, the literature reviewed in Chapter 2 and the implications for the field of policing, focusing on the usefulness of big data in serious crime investigations. This will be followed by a comprehensive interpretation and analysis of the interview data against the findings from the literature to answer the research question of this thesis. The overall synthesis of the discussion that integrates the findings will be presented in the conclusion of this chapter to offer a clear account of how the empirical findings can advance existing research. Finally, the empirical, theoretical, and practical contributions of this thesis will be presented. To orient the reader, the research question and aims were as follows:

Research question:

What is known about big data and can it be useful in serious crime investigations?

Research aims:

1. To develop a comprehensive understanding of the concept of big data, including its definitions and characteristics, to form a theoretical basis as a foundation for further research within the context of serious crimes.
2. To explore the potential usefulness of big data in serious crime investigations by identifying the advantages, disadvantages, challenges and availability of artificial intelligence (AI) tools in relation to big data.
3. To develop a definition of big data suitable for use across policing internationally.

Big data in serious crime investigations: Insights from the research question

The outcomes of the scoping review indicated that big data can indeed be useful in serious crime investigations. However, achieving usefulness and ensuring effective implementation can be complex. The findings will be presented across six themes, introduced below.

As the scoping review and empirical analysis established that there are debates and varying conceptualisations of big data in the literature and among the participants, exploring fundamental aspects such as definitions, characteristics and criteria was perceived to be essential to form a clear understanding, aligned with Research Aims 1 and 3 and presented in Theme 1 (see 5.2). This is followed by an overview of the potential role of big data in serious crime investigations, contributing to answering the research question and presented in Theme 2 (see 5.3). The potential advantages are presented in Theme 3 (see 5.4),

challenges in Theme 4 (see 5.5), concerns in Theme 5 (see 5.6), and technical tools and datasets in Theme 6 (5.7). The findings discussed directly address Research Aims 1, 2, and 3 and contribute to answering the research question.

5.2 Theme 1: Concept of big data (Research Aim 1)

5.2.1 Definitions

Before exploring the potential usefulness of big data in serious crime investigations, it was deemed necessary to clarify the conceptualisation as the scoping review showed a lack of consensus among scholars in terms of defining big data. The field of big data is an emerging discipline and thus the concept is open to different interpretations and definitions (Mauro, Greco and Grimaldi, 2016; Babuta, 2017).

In this thesis P2, P4, A1, A2, A3, A4, A6, and A7 acknowledged the debate and the lack of agreement concerning the definition and conceptualisation of big data, leading to challenges in reaching a common understanding. As A7 stated, *"So we do not really have a clear definition of big data"* and P2 suggested that this could make it challenging to discuss big data. Only two of the nine policing participants referenced the debate in the field, while six of the eight big data expert participants did so, perhaps suggesting a greater awareness within academia. This also supports the methodological approach of including two groups of participants based on their experience in each field, which provided detailed insights into the variations in understanding and conceptualisation.

The review of the literature in the scoping review showed that several definitions of big data share common characteristics, known as the three Vs: volume, velocity and variety (Richards and King, 2014; Broeders et al., 2017; van der Voort et al., 2019). As work in the field has evolved, characteristics have been added, such as value, veracity, virtual and variability (Hopkins and Evelson, 2011; TechAmerica, 2012; Zikopoulos et al., 2012; Dijcks, 2013; Dumbill, 2013; IBM, 2015; van der Voort et al., 2019); however, the original three remain the most prevalent.

The critical analysis of the interview data indicated two approaches to defining big data among the policing and big data expert participants. P5 and P9 defined big data directly, using the traditional three Vs, while P1, P3, P6, P7 and P8 defined big data using general terms, although these were also related conceptually to the notions of volume, velocity and variety. In addition, the critical analysis of the policing participants' views showed that the definitions provided by P1, P3, P7 and P8 indicated a tendency towards expecting value from big data. They suggested that big data *"can benefit us"* (P1), *"support decision making"* (P2), *"has unknown value until explored"* (P7), and offers benefits *"which would not be otherwise known if the user did not have the ability to bring data together"* (P8).

Among the big data expert participants, A2 and A3 also described big data in terms of the three Vs (volume, velocity and variety), whereas A1, A2, A4, A5 and A7 used their own expressions, which drew on scientific and technical terms. The link to the notion of value made by P1, P3, P7 and P8 was also observed in the characterisations of A2 and A4, with A4 noting the potential of big data *"to uncover large hidden values"*. A2 first gave what was described as *"a textbook definition"*, namely the three Vs, but then focused on the value that big data can offer, stating that *"we should…seek to understand how big data offers something significantly different"*.

Moreover, value was one of the big data characteristics highlighted by several scholars in the literature (Dijcks, 2013; Dumbill, 2013; Kitchin and McArdle, 2016; Bell et al., 2021) due to its vital role in big data projects. According to those who raised the importance of value, there must be value in collecting, managing, storing and analysing big data (i.e. purpose). Also, it is important to assess and determine the value of any big data project during the planning phases due to the high financial costs and challenges in utilising big data (Zikopoulos et al., 2012), as also noted by P2.

The purpose of asking the participants to define big data was to observe their understanding, how they visualised it and their possible awareness of the debate concerning what constitutes big data. The main difference observed was that the big data expert participants appeared to be more aware of the debate in the field compared to the policing participants, perhaps indicating varying degrees of knowledge among the participants. Nevertheless, despite the non-technical terms used, the policing participants indicated that they were aware of and familiar with the notion of big data.

The differences in defining the concept of big data do not preclude the possibility of reaching mutual understanding but they do need to be acknowledged. As shown in this thesis, the characteristics and elements considered to be related to big data continue to evolve over time, potentially differently within different fields and/or between scholars and practitioners. Thus, in this study, it was important to gain a better understanding of the characteristics of big data relevant to the (evolving) definitions, discussed further in the next section.

5.2.2 Big data characteristics

This thesis aimed to identify the core elements of what constitutes big data. This thesis identified that the three Vs (volume, velocity and variety) are the characteristics of big data most commonly cited. In addition to these, the findings suggest additional characteristics potentially helpful in defining big data, such as value, as noted by P1, P2, P3, P7, P8, A2 and A4 and the literature (Zikopoulos et al., 2012; Dijcks, 2013; Dumbill, 2013; Kitchin and McArdle 2016; Bell et al., 2021).

Moreover, scholars went beyond the three Vs, proposing the following: exhaustivity, resolution, indexicality, relationality, extensionality, and scalability (Kitchin and McArdle, 2016); versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy, virility, valueless, vampire-like, venomous, vulgar, violating, and very violent (Uprichard, 2013); veracity, value, visibility, viability and variability (Bell et al., 2021); portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, privacy, polyvalent, polymorphous, and playful (Lupton, 2015).

The diverse range of characteristics suggested by prior studies reflect the ongoing debate in the field, which resulted in the identification of 41 characteristics in the literature. Table 5.1 presents the characteristics identified in the literature (column 1), those suggested by the participants (column 3) and those that are shared or similar (column 2). As can be seen from Table 5.1, of the 10 characteristics mentioned by the participants, 5 were also identified in the literature: value, variability, variety, velocity, and volume. A possible explanation for the range of characteristics found in the literature is that they were identified in studies in different fields, leading them to focus on different characteristics. Furthermore, as noted by A3, this is what commonly happens in academia, as researchers like to be wordsmiths, inventing new terms or expanding on existing terms to progress their academic career. Relating this to the field of AI, A7 remarked, *"That is the beauty and pain of working in AI. Everybody makes something up to sound like an expert"*.

As well as addressing the central research question in terms of what is known about big data, the five characteristics identified in column 2 can perhaps be considered a useful baseline for establishing a consistent definition of big data for policing, discussed further in Theme 3.

Nonetheless, one of the five characteristics was notable and may help explain the continuing difference in defining big data. Variability was highlighted by several studies (Hopkins and Evelson, 2011; TechAmerica, 2012; Bell et al., 2021) and P6. Neither the studies cited nor the participant gave any additional context, such as a definition, simply mentioning it as a characteristic of big data. Also, although this characteristic was highlighted by one participant only, which may indicate a gap in awareness among the other participants or within the field, it links directly to the ongoing debates surrounding the definition of big data. However, the Oxford English Dictionary gives a definition of variability that is perhaps useful, stating that it is *"The fact or quality of being variable in some respect; tendency towards, capacity for, variation or change"* (Oxford English Dictionary, 2025). This offers a possible insight into the ongoing debates and differences concerning how big data is defined and visualised, since it tends towards variation and change, making it challenging to set clear fixed boundaries.

Table 5.1. Overview of big data characteristics.

Literature review	Similar/shared characteristics	Participants
1. Exhaustivity	1. Value	1. Quality
2. Extensionality	2. Variability	2. Validity
3. Indexicality	3. Variety	3. Value
4. Partial	4. Velocity	4. Variability
5. Personal	5. Volume	5. Variety
6. Perverse		6. Velocity
7. Playful		7. Veracity
8. Political		8. Visualisation
9. Polymorphous		9. Volatility
10. Polyvalent		10. Volume
11. Portentous		
12. Practices		
13. Predictive		
14. Privacy		
15. Productive		
16. Provocative		
17. Relationality		
18. Resolution		
19. Scalability		
20. Value		
21. Valueless		
22. Vampire-like		
23. Variability		
24. Variety		
25. Velocity		
26. Venomous		
27. Veracity		
28. Versatility		
29. Very violent		
30. Viability		
31. Vibrancy		
32. Vigour		
33. Violating		
34. Virility		
35. Virtuosity		
36. Visibility		
37. Visionary		
38. Vitality		
39. Volatility		
40. Volume		
41. Vulgar		

The findings discussed in Theme 1 will be addressed further in relation to serious crime investigations in the following sections, seeking to develop a definition that reflects the recent developments in the field drawing on the participants' perspectives and the literature.

5.3 Theme 2: Big data and serious crime investigations (Research Aim 2)

This theme concerns the broad agreement on the usefulness and potential of the application of big data in serious crime investigations evidenced in the findings. This will be followed by an analysis of the types of serious crimes in the field and finally the drive to combat serious crimes.

5.3.1 Broad agreement on the value of big data in serious crime investigations

The analysis of the empirical interview data established that there was comprehensive support among the participants for the use of big data in policing generally and in serious crime investigations particularly (cf. 4.2). For instance, they commented:

> *"It is essential, significant, and a must, and to develop your police force you need to understand and use big data." (P1)*

> *"It can lead to tremendous insights, we have used it successfully from fraud detection to demand forecasting, from optimisation of staffing to counter terrorism in all these fields, it can really help you getting insights." (A7)*

P1's observation linking the development of the police force to the need to understand big data also supports the rationale for examining the meaning of big data first in this thesis. A7's point supporting its use by presenting the potential advantages it can provide to the police will be critically discussed further in Theme 3.

The overall agreement among the participants aligned with the findings from studies in the literature which supported the positive role that big data and AI can play in assisting law enforcement in their operations and criminal investigations (Richards and King, 2014; Ferguson, 2015; Nath, 2006; Broeders et al., 2017; Pramanik et al., 2017; Zainab and Dhanda, 2018; Feng et al., 2019; Vestby, 2019; APCC, 2020; Jha, Sivasankari and Krishnappa, 2020; Xu, Cheng and Sugumaran, 2020; Neiva, Granja and Machado, 2022; O'Connor et al., 2022; Neiva, Machado and Silva 2023; Schuilenburg and Soudijn, 2023). However, to the best of my knowledge, none of the studies in the literature focused particularly on serious crimes. What is more, the existing literature in the field of policing has not engaged in a comprehensive exploration of big data, leaving a gap in the fundamental understanding of it as a concept, one this thesis has sought to fill.

The emergence of recent studies exploring big data in criminal investigations (Neiva, Granja and Machado, 2022; Neiva, Machado and Silva, 2023; Schuilenburg and Soudijn, 2023), identified following the empirical investigation in this study, reflects growing interest in the field. Moreover, in addressing Research Aim 3, seeking to explore if there were any recent applications and projects to strengthen the contribution of this

thesis by linking the conceptual findings to real world projects, several projects were identified as examples from the literature:

- The French Police centralised big database (Neiva, Granja and Machado, 2022)
- The Canadian Police big data and intelligence approach (O'Connor et al., 2022)
- The Dubai Police facial recognition artificial intelligence project (Tamim, 2024)
- The Netherlands Police utilisation of big data in front line policing, criminal investigations and intelligence (Schuilenburg and Soudijn, 2023)
- The Europol Information System, supported by data from police forces within the EU (Schuilenburg and Soudijn, 2023)
- The Prüm system, a connected European network that encompasses fingerprints, DNA profiles and motor vehicle information (Schuilenburg and Soudijn, 2023)
- The Egmont Group of Financial Intelligence Units (Schuilenburg and Soudijn, 2023)

These projects were cited as examples without delving into the details of their internal operations, implementation processes and outcomes. Hence, this thesis goes further and explores the potential advantages (Theme 3), challenges (Theme 4) and concerns (Theme 5) of the application of big data in policing and serious crime investigation to respond to the research question and fulfil Research Aim 2.

5.3.2 Serious crimes: Types and definitions

As this thesis focused on serious crime investigations rather than crimes in general, definitions and types of serious crimes were explored to provide conceptual clarity and establish the policing boundaries for the application of big data. Multiple definitions were found in the literature, with variations observed between them. Three are provided here. First, the European Commission's definition of serious crimes proposed in directive 2005/60/EC (Art. 5) is as follows:

> *…(a) terrorism offences, including membership in a terrorist group, criminal activities for the purpose of financing terrorism and inciting; (b) drug trafficking offences; (c) "the activities of criminal organisations;" (d) "fraud, at least serious;" (e) "corruption;" and (f) "all offences which are punishable by deprivation of liberty or a detention order for a maximum of more than 1 year..." (European Parliament and Council, 2005, p.21, cited by Paoli et al., 2016, p.275)*

Second, according to the definition provided at the United Nations Conference in 2012:

> *[Serious crime is] defined in article 2, subparagraph (b), of the Organized Crime Convention as meaning "conduct constituting an offence punishable by a maximum deprivation of liberty of at least four years or a more serious penalty. (CTOC, 2012, p.2)*

Third, the NAO (2019, p.5) defined serious crime as *"…criminal activity that is planned, coordinated and committed by people working individually, in groups, or as part of transnational networks"*. The first and second definitions set out sentences (a maximum of more than 1 year, up to 4 years or a more serious penalty) while the third did not. Another difference is that the first definition mentioned the types of crime, while the second and third did not.

According to Paoli et al. (2016), the term "serious crime" was first introduced in an EU policy in 1995 and was later circulated broadly in the early 2000s. The term further expanded as it was used in security policies in the UK (Paoli et al., 2016). Examining the various SOC strategies in the UK, it was observed that term "serious crime" is frequently associated with "organised crime" in UK legislation and policies (Home Office, 2013, 2018, 2023; National Crime Agency, 2021). The link between serious and organised crime is likely because they are often linked in practice, since organised crime usually involves serious offences.

Just as there were differences in the definitions of serious crime, there were variations in what type of crime police forces categorised and classified as serious. Table 5.2 presents the types of serious crimes proposed by the international policing body Europol, law enforcement agencies (the Dubai Police, the Metropolitan Police, the New York Police), and Paoli et al. (2016). In addition, sets out the types of serious crime the participants considered big data would be useful for handling in their investigations. That is, the study sought to explore where big data could be applied in practice and connect theoretical insights to potential applications.

Table 5.2. Overview of types of serious crime.

Europol	Paoli et al. (2016)	Winchester (2020)	Dubai Police (2023)	Metropolitan Police (2023)	New York Police (2023)	Participants' suggestions
• Currency counterfeiting • Cybercrime • Drug production • Fraud • Illicit waste trafficking • Intellectual property crime • Migrant smuggling • Organised property crime • Sports corruption • Trafficking in human beings • Trafficking of endangered species • Trafficking of firearms	• Assault • Homicide and murder • Robbery • Burglary • Rape • Property crime • Motor vehicle theft • Drug trafficking and possession • Arson • Sexual crimes • Kidnapping	• Bribery/ Corruption • Border vulnerabilities • Cybercrime • Drug offences • Fraud and Forgery • Human trafficking • Illegal firearms • Kidnap/ Abduction • Money laundering • Organised immigration crime • Rape and sexual offences	• Abduction • Aggravated assault • Burglary • Drugs • Grand theft auto • Human trafficking • Rape • Robbery • Theft • Wilful murder	• Burglary • Criminal damage • Drugs • Fraud and forgery • Robbery • Sexual offences • Theft and handling • Violence against the person	• Burglary • Felony assault • Grand larceny • Grand larceny of motor vehicle • Murder and non-negligent manslaughter • Rape • Robbery	• Offences in public places • Arson • Burglary • Cyber-crimes and cyber-bullying on social media • Drug trafficking • Financial crimes • Fraud • Murder • Online sexual offending • Rape • Robbery • Terrorism • Treason

Sources: Europol (2023), Paoli at al. (2016), Winchester (2020), Dubai Police (2023), Metropolitan Police (2023), New York Police (2023), P5, P6, P8, A6, A7, A8.

Exploring the types of serious crime provides a better understanding of those that could potentially be prevented or investigated by analysing big data and using AI tools. As noted by A8, *"Big data is involved in every crime"*. Despite the difference in classifying the types of crime, this thesis found that most serious crimes suggested by the participants corresponded to the types of crimes classified as serious in the literature and by police forces.

In addition, no set definitions of serious crimes were provided by the police forces explored in the scoping review – the Dubai Police, the New York Police, or the Metropolitan Police – besides listing the types of crime considered to be serious or major. However, the Metropolitan Police do provide a clarification for each type of major crime individually, for example *"violence against the person"*, which includes serious offences such as *"murder"*, *"actual bodily harm"* and *"grievous bodily harm"*, as well as an individual definition of a major crime, such as robbery, which is defined as *"Theft with the use of force or a threat of force. Both personal and commercial robbery are included. Snatch theft is not included"* (Metropolitan Police, 2023). Finally, the findings clarify what is meant by serious crimes, identify their types and define the scope in terms of exploring the usefulness of big data.

5.3.3 The drive to combat serious crime

The study observed a drive to address serious crimes, in addition to acknowledging their consequences in terms of harm to societies and countries. There are more than 100 government organisations and law enforcement agencies (see Table 1.2) tasked with combatting serious crimes, rather than a single entity with the overall charge of tackling and responding to such crimes (National Audit Office, 2019; Winchester, 2020). This highlights the scale of the problem and the complexity of addressing it. On the positive side, the wide distribution of responsibilities indicates that it is being taken seriously at different policing and governmental levels. However, this disjointed approach can create challenges, as many organisations are involved, risking the duplication of work in tackling serious crimes, the ineffective use of resources and possibly divided funding. Hence, there is a need for clearer roles or a joint strategy to ensure efforts support each other.

Moreover, efforts to combat SOC were noted in the UK through the development of SOC strategies over the past 10 years (Home Office, 2013, 2018, 2023). These strategies have evolved over the years as crimes have become more complex, crossing borders and taking advantage of modern technology (Home Office, 2023). The drive to combat serious crimes derives from the harm and consequences for societies and countries (Xu, Cheng and Sugumaran, 2020). Indeed, they are considered a critical national security threat (Home Office, 2013). An example of the financial harm resulting from SOC in the UK can be seen in its cost, estimated in 2013 to exceed £24 billion per year, increasing to at least £47 billion in 2023 (Home Office, 2013, 2023).

As the cost of SOC continues to rise, the effectiveness of the strategies aimed at combatting them are open to question, although it is also not known whether the harm would have been worse without having the strategies in place. However, the government and police forces are to be praised for their efforts in developing these strategies as they are trying to combat SOC. Besides other possible consequences, the substantial increase in the cost of SOC in the past 10 years suggests that the financial damage is increasing, illustrating the urgent need to address serious crimes. In addition to the economic costs, child sexual abuse and exploitation, drug supply lines, fraud, human trafficking and sexual crimes have significantly increased in the past years, and a high percentage of these remain hidden or unreported (National Audit Office, 2019; Winchester, 2020).

The continuing development of SOC strategies and amendments made by the UK government is an acknowledgment that SOC are evolving, as should the methods used to combat them (Home Office, 2013, 2018, 2023). Examining the three strategies showed that they draw on technology, data and intensive analysis in striving to overcome SOC. While they do not refer to the term "big data", the 2013 and 2018 strategies do mention "bulk data" (Home Office, 2013, 2018). The absence of references to big data in the SOC strategies underscores the significance of this thesis in seeking to evaluate whether big data could be useful and add value in this field. This gap not only highlights a missed opportunity in earlier approaches but also raise questions of why big data has not been utilised sooner, as several findings have highlighted its potential.

In contrast, the National Policing Digital Strategy 2020–2030 directly mentions big data and recognises its positive role, considering it a strategic asset for policing (APCC, 2020):

> *This game-changing promise of big data and machine learning requires policing to treat data as a strategic asset in how it is captured, managed and analysed. (p.3)*

A notable finding is that the strategy describes big data as *"game-changing promise"* and a *"strategic asset"*, reflecting the strong view of its potential, albeit also reflecting the need to look carefully at the evidence and the requirements for effective implementation. Indeed, the National Policing Digital Strategy argues that analysing large datasets has immense potential for policing and could be a *"force multiplier"* (APCC, 2020, p.3). The participants similarly reflected optimism and caution in discussing the potential of big data in policing and serious crime investigations:

> *"With the global development using big data in policing is considered a part of our job…" (P2)*

> *"…if it was applied within the legal constraints, its potential is enormous." (P8)*

> *"The use of big data in the policing field is definitely effective." (A1)*

> *"I think it can probably be become very very effective. I also think there are some risks here." (A3)*

Notably, P8 voiced frustration that the potential of big data was not being exploited to its full capacity due to several challenges. However, there are emerging studies and there is evidence of a growing interest in researching the application of big data in criminal investigations (APCC, 2020; Neiva, Granja and Machado, 2022; Neiva, Machado and Silva, 2023; Schuilenburg and Soudijn, 2023). This suggests that the field has started to acknowledge the relevance of big data in criminal investigations. Nonetheless, the literature remains at an early stage, with limited empirical evidence to support many of its perceived potentials due to the discovered challenges. Hence, this thesis aims to contribute by addressing this gap to offer a critical examination of the usefulness of big data in serious crime investigations.

5.4 Theme 3: Advantages of using big data in serious crime investigations (Research Aims 2 and 3)

This theme addresses the advantages of the application of big data in serious crime investigations drawing on the participants' perspectives and the findings from the literature, contributing to achieving Research Aim 2 and answering the research question.

Recognising the advantages of big data in policing starts by examining two definitions in the literature. The first, by Neiva, Granja and Machado (2022) highlights the benefits as follows:

> *...the processing and analysis of large amounts of information, aimed at supporting policing activities, defining security governance policies, and advancing with criminal investigations. (p.1)*

The second, proposed by Schuilenburg and Soudijn (2023), focuses solely on securing society and making it safer:

> *...the use of large volumes of data made accessible by means of algorithms and gathered with the objective of making society safer. (p.1)*

To the best of my knowledge, there are no definitions specifically defining the use of big data in the context of serious crime investigations. Therefore, having established a fundamental understanding of the concept of big data in Theme 1 and serious crime in Theme 2, this thesis aims to propose a definition of the value of big data in serious crime investigations (Research Aim 3) at the end of this section, after discussing and analysing its advantages.

Critically analysing the literature and synthesising the participants' arguments led to the proposition of 10 advantages, 5 strategic and 5 operational, that police forces can exploit from analysing big data in serious crime investigations. The strategic advantages include gaining better insights, forming strategies, enabling demand forecasting, conducting faster operations with higher efficiency and the proper

utilisation and deployment of resources. The operational advantages encompass real-time analysis, detecting serious crimes and criminals, facilitating pattern extraction, predicting crime, and controlling crime. The following sub-sections provide a critical analysis of the relevant advantages, with a final reflection at the end of the section integrating them.

5.4.1 Serious crime investigations and prevention

This sub-section concerns the perceived advantages of using big data in terms of investigating, detecting and preventing serious crimes. These include discovering and identifying criminals and criminal networks, real-time analysis, pattern extraction, crime prediction and control.

Regarding the identification of criminals and criminal networks, the findings suggest that using big data and AI tools can assist the police in understanding and detecting crimes, identifying online terrorist content, uncovering suspicious transactions in selling dangerous products, discovering crime hotspots, and identifying criminal networks, identified both in the literature (Pramanik et al., 2017; European Commission, 2020; Assouli, Benahmed and Gasbaoui, 2021) and by P3, P5 and P6. As an example, Assouli, Benahmed and Gasbaoui (2021) noted that link prediction can be used as a tool to perform complex searches of social networks to identify links between node pairs, which can be used to discover links between members of criminal groups.

The findings also suggested that big data could be used for mass surveillance to detect criminal activity in real time, analysing surveillance video, facial recognition programs and electronic communications (Neiva, Machado and Silva, 2023). P1, P2, P4, A1 and A6 agreed on the usefulness of facial recognition in searching for wanted criminals and detecting and arresting them to further investigate their crimes. This corresponds with applications of facial recognition by the Dubai, Metropolitan and South Wales police forces, which exploit this technology for crime detection and prevention purposes (Ezzeddine, Bayerl and Gibson, 2023; Tamim, 2024). Due to the growing complexity of crimes, combining big data with AI tools can advance the ability of policing operations to identify, predict and combat new crime trends (Ezzeddine, Bayerl and Gibson, 2023). Also, the findings suggest that these tools and applications can provide real-time analysis in minutes or hours, in contrast to manual analysis, which could take days, weeks, or months (Broeders et al., 2017). Indeed, P8, A2 and A4 argued that the developments in computing powers can enable police forces to perform real-time analysis, which can support serious crime investigations by providing critical information that would previously have taken significantly longer if done manually by humans.

Furthermore, identifying patterns or pattern extraction were among the capabilities recognised in the literature as being of value in analysing big data to support reactive approaches in serious crime investigations, as well as proactive methods to prevent them (Broeders et al., 2017; Zainab and Dhanda,

2018; Feng et al., 2019; van der Voort et al., 2019; European Commission, 2020) and also endorsed by P1, P3, P5, P6, A3, A7, and A8. As P5 suggested, *"[Big data] can be used for the police force to be proactive and reactive"*.

In investigating crimes as a reactive approach, detectives can use big data analytics to identify patterns of similarities between related/serial crimes, whereas the proactive approach consists of creating crime prevention strategies by predicting future crimes based on historical data and understanding criminal behaviours. These approaches are already commonly used in policing, but the analysis of the empirical data and observations in this study suggests that the potential to conduct such analyses in real time offered by big data is what makes it transformative. This finding reinforces the description of big data as a *"game-changing promise"* (APCC, 2020, p.3) since it can enable much faster responses and more timely decision making, leading to a shift in practice.

In this regard, P2 argued that using big data systems can reduce the time needed to run an investigation, which can also lead to a better utilisation of resources. For instance, P2 noted that an investigating officer would be able to inquire about a suspect from a single big data system rather than having to search multiple individual systems. Having a single big data system connecting various databases can provide information *"...in a click of a button"* (P2). This point was also noted by (Schuilenburg and Soudijn, 2023).

Emerging evidence can be seen in the Prüm system, which includes European DNA, fingerprints, and motor vehicle databases, and has been described as one of the big data applications in policing (Schuilenburg and Soudijn, 2023). This system can enable police detectives to perform a DNA, fingerprint, or vehicle information search and compare it with millions of records across European countries to support their criminal investigations, corresponding with P2's notion of the availability of data and ease of analysis. The correlation between these findings suggests that big data in serious crime investigations is gaining wide recognition in both practice and research.

Similarly, the literature showed that big data analytics can detect patterns and inconsistencies to identify fraud, both in the financial field (Aderemi et al., 2024; Udeh et al., 2024) and policing (Vestby, 2019; Jha, Sivasankari and Krishnappa, 2020), as also reported by A7. Fraud, as a serious crime (see Table 5.2), cost around £190 billion in 2018 due to 3.6 million fraud incidents (Winchester, 2020, p.3). Also, figures from the Crime Survey for England and Wales indicated an increase in fraud incidents from 3,200,000 in the year ending March 2024 to 4,159,000 in March 2025 (ONS, 2025), representing a 31% increase. The cost of SOC in the UK exceeded £24 billion per year in 2013 and was estimated to increase to at least £47 billion in 2023 (Home Office, 2013, 2023). Clearly the cost of SOC is far greater, although fraud is considered a serious crime (Winchester, 2020; Metropolitan Police, 2023), and the lack of

alignment between the numbers and inconsistencies may contribute to underestimating the scale and consequences of the problem of SOC. Whatever the case, the findings emphasise the need to address serious crimes and given the growing evidence of the usefulness of big data, it is essential that police forces explore advanced solutions to combat such crimes.

Moreover, when the advantages of using big data reactively and proactively are combined, they can result in predicting the time and place where crimes are more likely to occur by analysing big data through probabilistic mathematical algorithms, also known as crime mapping and hotspot policing (Pramanik et al., 2017; Zainab and Dhanda, 2018; Egber and Krasmann, 2019; Feng et al., 2019; Sandhu and Fussey, 2021; Neiva, Machado and Silva, 2023; Schuilenburg and Soudijn, 2023), as highlighted by P1, P5, P6, A4, A7, and A8. Taken together, the findings indicate that prediction based on big data can contribute to controlling crime (Xu, Cheng and Sugumaran, 2020). Similarly, P1 considered that analysing big data could be used to understand how, when and what types of crimes may occur, leading to the development of strategies and provision of support for decision making to prevent them.

5.4.2 Decision making and advancing operations

This sub-section discusses the potential of big data in advancing decision making, forming strategies and increasing the operational speed and efficiency of policing. The review of the literature suggested that the use of big data analytics has expanded in the security community due to its ability to analyse and connect security-related data at an unprecedented scale (Cardenas, Manadhata and Rajan, 2013). More recent findings indicate that these analytical capabilities have advanced even further (Loenen, Kulk and Ploeger, 2016; Home Office, 2018; APCC, 2020; Schuilenburg and Soudijn, 2023; Aderemi et al., 2024). Advancements in analytical capabilities can enhance operational speed and efficiency in analysing big data. Feng et al. (2019, p.1) defined big data analytics as *"…a systematic approach for analysing and identifying different patterns, relations, and trends within a large volume of data"*. In the context of policing, the analysis of data translates raw information into operationally practical and valuable intelligence (James, 2016). In addition to the advantages identified previous, big data analytics has been shown to offer promising outcomes for policing in terms of enhancing decision making and improving risk analysis (Broeders et al., 2017; Zainab and Dhanda, 2018; Feng at al., 2019; APCC, 2020). In the context of predictive policing in serious crime investigations, the application of big data analytics aims to enable automated police decision making that can assist in developing early intervention strategies to combat crimes (Sandhu and Fussey, 2021).

In general, the advocates of big data and big data analytics contend that the advantages and potential of big data are based on a simple theory: big (more/greater volumes of) data can lead to better decision making (van der Voort et al., 2019). However, Newburn (2008) and O'Connor et al. (2022) highlight that as policing moves into an era of using big data, it is extremely important that they collect data that

is of high quality, as this will play a vital role in ensuring the effectiveness of analytical tools, such as crime maps or offender network associations. In addition, P3, P4, P5, A2 and A4 emphasised the importance of the quality of data used in decision making for policing purposes, framing this as a challenge. Data quality refers to assessing how reliable the dataset is in terms of its use in decision making (Thabet and Soomro, 2015).

This study found that the quality of data is more critical for decision making than the size of the dataset. While large and varied datasets can offer a range of information, low-quality datasets pose risks for police forces in terms of decision making in serious crime investigations. Poor decision making can lead to developing ineffective or even harmful strategies that can be detrimental for policing approaches in tackling crimes. Also, low-quality data can lead to biased decisions, identified as one of the concerns that will be discussed further in Theme 4 (see 5.5.4).

Furthermore, the findings indicated that it is important not to compare police decisions based on machine learning to ideal decision making, but to normal human decision making (Vestby, 2019). This suggests that decision making based on AI tools in policing should be compared to human decision making that may be subject to error and bias. Hence the value of machine learning and other AI tools that can assist in decision making lies not in achieving perfection, but in showing potentially improving on existing human practices.

5.4.3 Proper utilisation of resources and demand forecasting

The study found that the analysis of big data in policing could lead to demand forecasting, resulting in the better utilisation and effective deployment of resources (Newburn, 2008; Zainab and Dhanda, 2018; Feng et al., 2019; O'Connor et al., 2022; Ezzeddine, Bayerl and Gibson, 2023; Neiva, Machado and Silva, 2023), as reported also by P3, P4, A3, A7, and A8. Recent studies have suggested that police forces are starting to adopt AI tools and big data analytics to increase their operational efficiency (Ezzeddine, Bayerl and Gibson, 2023), which aligns with the optimistic views of policing professionals (Neiva, Machado and Silva, 2023). The findings did not specifically suggest that the proper utilisation of resources is related to serious crime investigations, but rather the potential for policing in general.

The advantages identified above appear to be linked, creating a possible sequence in which each advantage reinforces another. Figure 5.1 was developed based on the findings of this thesis, which presents an overview of this linkage, discussed further below.

Figure 5.1. Synthesised sequence of key advantages.

Figure 5.1 synthesises the key advantages of analysing big data in serious crime investigations. Starting with real-time analysis, it can enable the detection of serious crimes such as fraud and cybercrimes, in addition to the live detection of wanted individuals through facial recognition technologies and vehicles through automated number plate recognition systems. Moreover, police forces may swiftly identify suspects through comparing fingerprints and DNA samples from crime scenes with international DNA databases, affording cross-border reach. A recent example of this application in Europe can be seen in the Prüm system. Furthermore, through analysing the detection of serious crimes and criminals, patterns can be extracted to gain better insights leading to the prediction of where crimes are more likely to occur, offering targeted detection and prevention strategies that aim to control crimes. Also, as better insights are gained from detection, demand forecasting can be performed to increase the efficiency of operations, supporting the proper utilisation and deployment of resources. Hence, it is evident that the value of analysing big data in policing – especially for serious crime investigations – lies in building applications, one upon another, to achieve overall effectiveness. It is important to note that Figure 5.1 presents a suggested sequence of advantages rather than a fixed or mandatory progression. The ordering intends to show a logical flow, in which some potential advantages enable later ones; however, this does not imply that each advantage will arise in strict sequence.

In addition, recent studies (see, e.g., Neiva, Granja and Machado, 2022; Schuilenburg and Soudijn, 2023) have argued that despite the surge in the use of big data in policing, academic studies have focused on predictive applications rather than reactive applications after a crime is committed. The study findings address this gap and present the potential of big data, not only in terms of its predictive capabilities but also in detecting and investigating serious crimes. Moreover, Figure 5.1 represents a novel contribution of this thesis, as it systematically links the advantages that were initially identified separately and reframes them in a structured progression.

5.4.4 Definition of big data in serious crime investigations

To the best of my knowledge the literature has provided no prior definition of the use of big data in serious crime investigations. Also, as the literature and the participants' responses revealed, there is no single agreed upon definition of big data, with perspectives varying across disciplines. Hence this thesis synthesises the insights from Themes 1, 2 and 3 seeking to generate a definition that can establish conceptual clarity and ensure the concept of big data can be understood and applied in the context of policing and especially in serious crime investigations.

Drawing on the five characteristics identified in both the literature and by the participants, the proposed definition includes value, variability, variety, velocity, and volume. This provides a potential baseline from which the debate can be addressed, without precluding the addition of supplementary characteristics in the future, as big data and technology are constantly evolving.

Concerning the differing definitions and classifications of serious crimes, it is challenging to impose a fixed definition or description as they are based on legislation and conceptions that vary from one country to another. For this reason, the proposed definition treats serious crime in a way that allows it to be applied flexibly by police forces based on their laws and classifications. However, serious crime investigations will be linked to achieving certain strategic and operational advantages as they represent the overarching aim of utilising big data. Therefore, big data in the context of serious crime investigations in this thesis can be defined as:

> *Proposed definition*
> High volumes of information that are diverse, variable and generated at high velocity, analysed by artificial intelligence tools to yield insights for policing to achieve strategic and/or operational advantages in serious crime investigations.

The definition synthesises the perspectives capturing both the technical and contextual dimensions of big data in serious crime investigations as revealed in this thesis. Moreover, this definition aligns with Mauro, Greco and Grimaldi's (2016) view that the definition of big data should clearly identify the area of practice (serious crime investigations) rather than depending on the general field that it is applied in.

5.5 Theme 4: Challenges of using big data in serious crime investigations (Research Aim 2)

Theme 4 addresses Research Aim 2, which aimed to explore the possible disadvantages or challenges that police forces may encounter as a result of using big data in serious crime investigations. The findings showed that the barriers were framed and interpreted as challenges rather than as disadvantages. As A2 put it, "I think maybe disadvantages is not quite the right word, maybe challenges is better". This reflected the interpretation among the participants' responses and the findings in the literature, indicating that such challenges are seen as areas for development that can be addressed rather than persistent disadvantages that cannot be fixed. In addition, big data has the potential to be useful to governments and their citizens, albeit with important ethical, technical, policy and implementation implications (Broeders et al., 2017).

A critical examination of the data suggests that although big data has considerable potential (Theme 3, cf. 5.4), its effective implementation in serious crime investigations can be challenging. The challenges identified start from how the concept of big data is adopted and understood in policing, followed by financial, technical, human resource and operational challenges, all of which will be discussed under the following sub-themes: cultural adoption, high financial costs, technical challenges, and data quality.

5.5.1 Cultural adoption

One of the findings that emerged from the participants' responses (P3, P4, A4) was that adapting and adopting big data in organisations can be culturally challenging. This did not appear in the literature reviewed in Chapter 2, suggesting that it may be a more recent practice-based insight drawn from the participants' experience.

P3 associated this challenge with the lack of a culture of using big data as it is a new phenomenon. In policing, P3 remarked that some senior managers do not have sufficient knowledge of big data and its capabilities and therefore they need to be shown its potential for improving decision making and other advantages. Moreover, P4 commented on a related challenge, namely that there is a lack of understanding of how computing and its sub-fields work, as some managers see an employee working on a computer and expect him/her to be able to perform all types of computing tasks or jobs. P4 argued the need for awareness raising about computing and its sub-fields, in addition to continuous training for both employees and higher management. Similarly, A4 recommended that police forces educate their officers not only about big data but also other technologies and techniques that could assist them in doing their work better, considered an essential aspect of training and development. P3, P4 and A4 frequently highlighted the importance of exposure to the capabilities of big data and training to overcome this cultural challenge, particularly as P3 and A4 noted that there can be resistance and/or fear of change among employees, both junior and senior, in organisations, as they think that adopting new technologies and automation could replace them, leading to job losses. In addition, A4 argued that when it organisations consider it appropriate to adopt new technologies, the higher management should develop clear policies for the employees to follow to ensure the organisation adapts and develops.

As the discussion shows, the cultural challenge arising in the adoption of big data was considered to be related to a lack of knowledge and awareness of its capabilities. Hence, the solutions to this challenge proposed by P3, P4 and A4 were to ensure understanding of and exposure to the potential of big data, implemented through continuous training and education, as well as the development of policies to ease its integration in practice. This observation highlights the importance of the data collected as it bridges the gap between theory and practice and potentially suggests areas for future research.

5.5.2 High financial costs

The study identified financial challenges in creating the infrastructure required to properly utilise and analyse big data, as building it can be expensive and it is not easily affordable by all private or government entities. This challenge was highlighted by multiple participants (P1, P2, P4, P7, P8, A4, A7). It was an interesting perception as Neiva, Machado and Silva (2023) contended that big data applications are expected to be cost-effective, although they did not elaborate on what they meant or how it would come about. Moreover, the APCC (2020) found that the cost of storing one gigabyte of

data declined from £800,000 in 1967 to less than £0.016 in 2020. However, the analysis shows that the financial costs go beyond data storage. The substantive expense lies in establishing an end-to-end infrastructure capable of collecting, storing, managing and analysing big data, requiring investments in measures that include data governance and security, assuring data quality, implementing AI analytics tools, undertaking daily operations and funding human resources. This is what policing appears to be struggling with (P1, P2, P4, P7, P8, A4, A7). As noted by A7, it is recommended that police forces initiate and build their big data projects internally, which can be costly in the beginning but will be cost-effective over the long term. A4 termed this an "investment", highlighting the high up-front costs: "Financial matters are a core of big data challenge, because investment in these kinds of projects require a huge amount of money, which some organisations might not really like".

However, although the findings suggest that developing big data projects can be challenging due to the high initial costs, Theme 3 highlights the strong potential of big data to address problems that cost billions annually. A critical interpretation suggests that the balance between cost, capability and potential must be carefully evaluated, as the enthusiasm to adopt advanced technologies might outpace the readiness needed for effective implementation by police forces. In this vein, P2 proposed the need for a return-on-investment formula or an equation that police forces could use to aid decision making in developing and assessing their big data projects. P2 suggested that with variety of big data systems and technologies available, it can be challenging to evaluate and assess what systems to purchase to combat serious crimes, particularly given the high costs. The different perspectives articulated, noting the enthusiasm for and potential of big data, while at the same time emphasising the high cost and the need to balance this in terms of the value added in the face of a lack of a return-on-investment formula, illustrates the complexity of the topic. As P8 pointed out, it is frustrating to see the potential of using big data analytics to support serious crime investigations and policing but not be able to explore it fully due to financial constraints. Notably, A7 warned against going with new startups or cheap vendors to develop big data projects for policing as a way of reducing the cost, as this could pose significant risks, particularly in dealing with sensitive information. While the literature reviewed did not specifically highlight the risks of working with startup companies or cheap vendors for policing, Neiva, Machado and Silva (2023) did point out issues related to decision making, bias and privacy, illustrating the potential repercussions noted by A7.

5.5.3 Technical challenges

The findings highlighted various technical challenges related to the characteristics of big data that could pose issues for police forces seeking to develop big data projects and/or utilise big data analytics for serious crime investigations (P2, P3, P4, P5, P8, P9, A2, A4, A7, A8). The technical challenges identified concerned the forms of big data, the infrastructure required, storage, data quality, reliability and security.

The first challenge identified concerned the high volumes of big data and the difficulties not only in having the required infrastructure for storage but also determining and specifying the duration of the retention of data (P2). High volumes of data as a challenge for policing was also highlighted in the literature; Hassani et al. (2016) point out that crime data are constantly growing, resulting in the need for advanced methods to analyse big data in an effective and accurate way. This was reflected in the reports of the participants concerning the application of big data in practice (P2, P3, P4, A4).

Moreover, in addition to crime data, policing faces the challenge of dealing with supplementary sources of data, such as ANPR records collected for use in investigations (Metropolitan Police, 2025). In 2017, the Metropolitan Police in London received approximately 38 million such records per day (Babuta, 2017), increasing to 60 million records per day at the national level in the UK in 2025 (Metropolitan Police, 2025). This means that approximately 21.9 billion records are generated yearly at the national level, demonstrating the magnitude of the storage requirements for only one platform. P2 argued that this can lead police forces to select specific datasets and avoid saving all the data as huge investments would be needed to create the required infrastructure. Given the financial challenges identified in this study, storing such data constitutes an additional burden. Currently, national ANPR data in the UK are stored for a maximum of one year from the date of collection (Metropolitan Police, 2025). However, in the case of serious crime investigations, such a short duration could affect investigations of historical offences and the identification of long-term patterns.

The second technical challenge identified was that big data contains a variety of data formats drawn from different sources and these require cleaning and refining to find the most useful sets (P2, P3, P4, A4). These forms range from text to audio and video and are not limited to structured data but also comprise unstructured data (A4). Both the empirical findings and the literature point to variety and the associated high volumes of data as an issue (see, e.g., Zikopoulos et al. 2012; Thabet and Soomro, 2015). As noted by P2, "*With every new system we need new infrastructure and so on*". Similarly, Zikopoulos et al. (2012, p.7) argued "The volume associated with the Big Data phenomena brings along new challenges for data centers trying to deal with its variety".

However, the literature review also highlighted that having a variety of data types is fundamental for data analysis and decision making (Zikopoulos et al., 2012). Moreover, an organisation's effective use big data depends on its capability to gain insights from the various types of data it possesses (Zikopoulos et al., 2012). Examples in policing are IDENT1, the UK's central national biometric database (Babuta, 2017; Home Office, 2024), the Police National Computer (PNC) database, which holds millions of personal, vehicle and driver records (Babuta, 2017), and the Police National Database, a national intelligence system that contains local police records built from 220 databases (Babuta, 2017).

These databases are just examples to illustrate the variety of data types that are generated, collected and utilised within policing contexts. As the findings suggested, this diversity/variety results in high volumes of data, identified as a challenge. While the variety of data and databases offers valuable potential for investigating serious crimes, it also introduces complexity and challenges in data management and storage. This creates a tension between potential and practicality, as the potential cannot be achieved without the requisite technical infrastructure to handle such high volumes and variety. Beyond this, P4 contended it is challenging for policing to find the most useful datasets for serious crime investigations.

The third technical challenge, also related to a characteristic of big data, is velocity, i.e. how fast data are produced (A7). Velocity was highlighted as a technical challenge by A4, linked to the high volume and variety of data. Thabet and Soomro (2015) and Wadhwani and Wang (2017) also considered that velocity could be a significant challenge as a result of not having the appropriate technology and tools to manage the continuous high flow of data. Data generated at high velocity requires analysis in real time to gain valuable insights (Thabet and Soomro, 2015).

Hence for policing to benefit from big data analytics, it is essential to develop and source the appropriate technical infrastructure to manage the high volumes, variety, and high velocity of data (Dijcks, 2013; Munoz, Smith and Patil, 2016; Srinuvasu and Santhosh, 2017; Wadhwani and Wang, 2017). As noted by Babuta (2017, p.1), "*If the police were able to effectively apply such technology to the data they collected, they would greatly enhance their operational efficiency and crime-fighting capabilities*". The empirical findings suggest that this is still an issue that poses an obstacle for policing, as reported by P2, P3, P4, P5, P8, P9, A2, A4, A7 and A8.

5.5.4 Data quality

The findings indicated that another challenge encountered in policing is ensuring that the data used in investigations are accurate and of high quality (P3, P4, P5, A2, A4). Using inaccurate or unclean data could be misleading in an investigation (P4), leading to a risk of bias or faulty decision making, a concern discussed further in Theme 5. Part of the challenge in this regard is the capability to reduce the size of data without compromising the quality (A4). Thus, it can be seen that the issue of the high volume of data is not only related to the challenges posed by variety and velocity but also quality. The findings highlighted that policing needs a data governance framework and quality assurance policies that cover all phases, including aspects of data collection and entry points, to ensure the data collected are correct (P3 and P5).

These findings align with prior research, as police analysts and scholars have argued that the quality of police data is a concern, particularly since the rise of big data, as relatively little is known about its collection and quality (O'Connor et al., 2022). Similarly, P4 pointed to concerns regarding the sources

of data used in policing. This finding is consistent with James (2016), underlining that attention must be paid to the sources of data to assure quality and credibility.

Again, to enhance the quality of data on which policing relies to attain the advantages outlined in Theme 3, it would be of value to have a data governance framework and quality assurance policies in place. Several studies in the literature have highlighted the importance of using high-quality data for crime mapping and offender network association, and to avoid misleading predictions (Newburn, 2008; Jin et al., 2015; James, 2016; Wadhwani and Wang, 2017; Sandhu and Fussey, 2021). All of these are considered from the perspective of the potential advantages offered by analysing big data in serious crime investigations.

Moreover, the findings highlighted a technical challenge that not only has implications for policing but also for the use of data more broadly at the national level. As A7 pointed out, many systems are built in English and do not lend themselves to translation into languages such as Arabic. This could lead, for instance, to security risks for countries in authenticating the identity of individuals entering the country (A7), for example because there is often not a straightforward transliteration from a name in Arabic to English. For instance, Mohammed can be spelt Mohamed, Mohammad, etc. From my professional experience, current systems do not translate names from English to Arabic; rather data are entered either through scanning documents, such as passports and identification cards and official documents issued in Arabic-speaking countries typically list the individual's name in both Arabic and English. This links back to the need for a data governance framework and quality assurance policies (P3 and P5) to ensure any variations in the data entered/collected do not lead to confusion or inaccuracies in the analysis, potentially harming serious crime investigations.

5.5.5 Qualified human resources

P1, P4, P7, P8, A1, A2 and A4 argued that it is challenging for policing to source the required personnel in terms of qualified human resources to manage and operate their big data projects. It is important to have data science specialists with the necessary skills and qualifications, as a lack of appropriate personnel could lead to the failure of the project (A4). According to A2, qualified specialists tend to be hired by private commercial companies, likely due to the higher remuneration compared to the public sector. Giving the existing financial constraints in policing, attracting skilled data scientists and specialists away from the private sector presents an additional economic challenge.

What is more, the absence of a big data culture among management in the policing sector, as noted by P3, P4 and A4, could contribute to the difficulty of recruiting and retaining human resources. The lack of a well-established culture within a police force can create difficulties in developing projects, leading to hesitancy among qualified human resources when it comes to joining the organisation and therefore

creating recruitment challenges. As P8 argued, police forces often have very good intelligence but no resources to exploit it. Hence, while big data could be useful in serious crime investigations, a major challenge for implementation is having the people to utilise it and realise its potential (cf. Theme 3), thereby creating an operational challenge.

5.6 Theme 5: Concerns of bias and privacy (Research Aim 2)

This theme concerns the findings related to bias and privacy risks. These two elements were identified as areas of concern because of the significant implications and consequences they may have for policing practice and decision making.

5.6.1 Bias

Biased decision making or results from analysing big data in serious crime investigations were underlined as a concern by the participants (P3, A2, A3, A7, A8). Similarly, the literature review recognised the potential of big data but also acknowledged the risk of bias (Hajian, Domingo, and Martinez-Balleste, 2011; Cardenas, Manadhata and Rajan, 2013; Broeders et al., 2017; Selbst, 2017; Sandhu and Fussey, 2021; Neiva, Granja and Machado 2022; Neiva, Machado and Silva 2023).

P3 argued that bias is a core concern, and the main cause of bias was the poor/low quality and cleanness of data. This aligns with Cardenas, Manadhata and Rajan's (2013) study, which established that the collection of big data from various sources makes it challenging to ensure all sources are reliable and that the analysis can produce unbiased results. Also, given that some of the data used to develop future predictions are based on police officers' categorisations, there is the potential for human bias (Neiva, Machado and Silva, 2023). As noted in relation to Theme 4, ensuring the high quality of data is one of the technical challenges that policing faces, thus making bias a significant concern.

A2 and A7 stressed the need for caution when employing the results obtained from new technologies in policing more than any other field. As A7 suggested "These in policing are as important as anywhere else, but I think especially important in policing…". A2 supported the use of technology in policing but stressed that "…there is always a degree of caution involved". These views reflect the importance of a lack of bias in policing, a field in which decision making and actions have implications for peoples' freedom. Similarly, Barocas et al. (2017) highlighted that biased outcomes and a lack of understanding on the part of computer scientists about the applications of data may affect civil rights. These were the consequences A7 was concerned about when suggesting that some developers build unaudited AI models under pressure but do "…not really understand the impact".

A further issue in terms of bias was raised with regard to profiling, which A3 framed as one of the clear risks that has frequently been addressed. This is in line with the literature, which also suggests that profiling and biased outcomes can result from analysing big data, leading to ethical complications

(Neiva, Granja and Machado 2022). Besides low-quality data being one of the causes of biased outcomes (P3), it has been argued that the data used to train big data models can potentially lead to biased results (Barocas et al., 2017). Examples of policing projects that have faced biased and ethical consequences were also identified. For instance, the Metropolitan Police and South Wales Police have faced legal challenges for using facial recognition technologies (A2). In addition, a policing project in the US that relied on AI software was believed to have generated biased results, suggesting that most criminals were of a particular race (P3). Such issues can arise from feeding the algorithm non-clean data (Barocas et al., 2017; P3). This highlights the importance of overcoming technical challenges, providing and maintaining high-quality data for analysis in serious crime investigations.

Crucially, algorithms do not understand ethical concerns, potentially giving rise to unethical propositions for decision making by humans (A8). The findings indicated that with more training, the accuracy of algorithms can be improved, but this would require proactive investment on the part of the policing organisation (A8). This links back to the need to overcome financial challenges and ensure the necessary technical infrastructure is in place to tackle any potential bias. In addition to algorithm training, Hajian, Domingo, and Martinez-Balleste (2011) propose that data-mining tools can assist in detecting biased decisions. However, Selbst (2017) cautions that failure to employ data-mining tools properly can itself lead to biased decisions or discrimination in patterns. While the findings neither supported nor challenged the use of data mining as an effective tool, the literature highlighted several tools that could potentially assist in overcoming some of the challenges related to big data, taken up in Theme 6 (see 5.7). As noted by Munoz, Smith and Patil (2016), if new technologies are carefully designed and implemented, they can support decision making based on measurable factors and variables, thereby reducing the risks associated with human bias and decision making.

Among the suggestions made by the participants, A3 proposed that there are impact assessment frameworks police forces can use when developing big data projects. The aim of using such frameworks is to assess the risks accompanying decisions made based on analysing big data and weigh the benefits of these applications against implications of possibly biased decisions. In light of this, it would be possible to educate the public about the risks of these applications, which A3 framed as an educational challenge. This is related to another concern among the public – privacy – discussed in 5.6.2.

In addition to impact assessment frameworks, the findings from the empirical data and the review of the literature suggested additional measures to combat bias-related challenges. A7 suggested that AI algorithms could be tested and audited for bias but noted that there is a human resource challenge in recruiting qualified data scientists to build these projects. A7 also referenced toolkits that can analyse big data models, aiming to detect for bias, but their effectiveness can be hampered by a lack of sufficient

information. This would benefit from further research to examine their possible utilisation in big data policing projects, as the literature reviewed did not refer to such toolkits.

The aim of testing algorithms for bias is to ensure transparency and explainability to the public (A7). This aligns with Munoz, Smith and Patil's (2016) view that police forces can address issues of bias through transparency and accountability. Moreover, Munoz, Smith and Patil (2016) recommend inspecting the data input process to ensure there are no data referring to a certain race, thus helping avoid biased outcomes. As argued by A7, there is an obligation to ensure that the AI algorithms used to analysis big data in serious crime investigations are transparent, explainable, robust, safe and unbiased. This obligation is important in policing as its decision making can have serious consequences for people's rights and public confidence.

5.6.2 Privacy

Privacy was highlighted as a core concern by the participants when asked if they had any misgivings about using big data in serious crime investigations. The literature likewise frequently noted privacy matters, such as risks to personal data, challenges in securing data, concerns about data sharing between entities and the need for more robust privacy legislation in the context of big data.

The findings demonstrated that participants' concerns about privacy were related to the risk that personal data could be viewed and searched by the police from the different forms of big datasets available in criminal investigations (A1 and A4). The literature also indicated privacy concerns in the context of big data as it contains various forms of personal data (Bignami, 2007; Cardenas, Manadhata and Rajan, 2013; Soomro, 2015; Babuta, 2017; Wadhwani and Wang, 2017). Data privacy has long been treated as a basic human right, with the drafting of laws and legislation to prevent any violations (Bignami, 2007). Yet, there are continuing challenges in protecting privacy as the criminal justice system has expanded (Bignami, 2007).

The participants acknowledged the importance of protecting individuals' privacy and suggested that policing should aim to achieve a balance between protecting the public from crimes and guarding their privacy (P4, P8, P9, A2). The right to privacy is qualified not absolute and thus collateral intrusion into an individual's privacy is possible as part of a criminal investigation, but any such intrusion should be justified, proportional and necessary based on law enforcement policies (P7 and P9). P9 argued that the balance between security and privacy in the UK is sufficient and appropriate. Policing recognises privacy rights by following certain principles, such as that any intrusion must be "justified, proportional, and necessary"; thus, investigators will not view or use personal data unless they can clearly explain why such data are needed. Given the importance of this issue, strict policing policies are needed to ensure that access to any form of information is governed by privacy-protecting guidelines.

Views on the importance of affording privacy and the balance with security differed. For instance, P9 suggested that those who are involved in serious and organised crimes forgo their right to privacy. In contrast, A2 was not convinced by the argument for balance in terms of offering flexibility and argued that the police should aim to achieve both security and privacy. From A2's point of view, those advocating a balance between security and privacy did not have sufficient knowledge of developments in technology. Taken together, it appears that proponents of the balanced approach argued for maximising both privacy and security, with the priority depending on the project or case, such that either privacy or security might be privileged, guided by the principles of legality, necessity and proportionality. From the alternate perspective, proponents of achieving both argued that systems should be developed to deliver robust security and privacy in parallel, without any compromise concerning the predominance of one or the other.

However, despite the latter argument, put forward by A2, there was no evidence that privacy and security could both fully be achieved without compromise; on the contrary, privacy was the concern most frequently highlighted among the participants. In addition, privacy concerns related to big data are not only found in policing but also in healthcare (Abouelmehdi, Beni-Hessane and Khaloufi, 2018; Balogun et al., 2025) and finance (Aderemi et al., 2024; Udeh et al., 2024), making privacy a common challenge.

Furthermore, in relation to the perception of privacy and security, the analysis showed that privacy can be visualised and interpreted differently between different cultures and societies. For example, CCTV and facial recognition are acceptable in some contexts but are rejected or considered a violation of privacy in others (P2). P2 considered that people in Western countries may view privacy differently than those in the UAE, where having cameras and other forms of technology to enhance security is not viewed as a violation of privacy, unlike in the West. This corresponds with a point made by A2 about cultural resistance to the use of facial recognition in the UK. A8 also contended that in some countries, individuals are concerned about the use of their data by private companies and police forces but did not specify where. Similarly, the literature points to data privacy concerns in the context of crime prevention and the sharing of data between the private sector and policing (Cardenas, Manadhata and Rajan, 2013).

These findings indicate varying perceptions of privacy, from those arguing for balance between privacy and security to those advocating an attempt to achieve both in parallel, as well as noting wavering cultural acceptance. As pointed out by P5 and P9, the debate about privacy is widespread and will be ongoing, but they considered that there are data governance laws to safeguard privacy rights. However, A3 felt that the current data protection legislation is insufficient and measures are required specifically to cover big data, AI applications and technological developments, reflecting calls to extend data protection legislation to cover advancements in computing powers and data-mining tools in the literature (Loenen, Kulk and Ploeger, 2016). A3 proposed the use of an algorithmic impact assessment tool that

could help police forces identify any privacy risks in their big data projects and suggest ways of mitigating them. Taken together, these points illustrate that existing legislation may not provide sufficient protection for privacy rights, particularly given the continuous developments in big data applications, underscoring the need for continuous legal adaptation. In parallel, police forces may well need to develop, tailor and embed privacy impact assessment tools in their big data projects to achieve the highest levels of privacy.

Babuta (2017) highlighted the significance of addressing privacy matters to enable police forces to use big data and its technologies effectively without violating privacy rights. One of the perceived advantages of analysing big data (discussed in Theme 3, see 5.4) was found to be the ability to identify trends and correlations, and Neiva, Granja and Machado (2022) and Wadhwani and Wang (2017) highlighted the importance of securing the privacy of personal data in that process. Despite the challenges and concerns identified, the findings in 5.4 suggest significant benefits of using big data; accordingly, policing should address privacy barriers to attain the full value of analysing big data. However, the findings do not endorse compromising privacy or tolerating biased outcomes, both of which remain primary concerns that need to be overcome.

Drawing the above arguments together, P4 posed a challenging question for policing and the public: Is it better to have more privacy but risk safety and security, or accept less privacy for greater security? For instance, P2 and A1 argued the need for cameras and facial recognition, stating that they play a critical role in investigating crimes and their absence could prevent the identification of suspects or valuable evidence. As policing has already moved into the era of big data, decisions about how far to use various tools will largely rest with governments and the public, based on their priorities and cultural acceptance, as well as their ability to overcome the challenges previously discussed. Moreover, as big data projects and associated technologies expand, privacy will continue to be at risk because of the various types of data they run on, including, most importantly, personal data. Therefore, any adoption should be paired with clear safeguards, strict oversight and transparent accountability.

Another observation is that police have long used different forms of data in serious crime investigations, for instance police databases, DNA records and other online data systems (Neiva, Granja and Machado, 2022; O'Connor et al., 2022; Schuilenburg and Soudijn, 2023; Metropolitan Police, 2025), so data usage itself is not new. The question here is whether current privacy concerns have arisen from labelling these sources "big data" or whether they have always existed regarding the risks of data use. Are such concerns driven by the expanding digitalisation of the high volumes and variety of data, such as ANPR and CCTV recordings, facial recognition, travel histories, phone metadata, social media and other data sources? Lastly, A2 and A7 stressed that protecting the privacy of individuals in serious crime investigations is

vital in policing, more so than in other fields, due to the tremendous impact on society if the wrong decisions are taken.

5.7 Theme 6: Tools and datasets (Research Aim 2)

Having established that the analysis of big data has considerable potential in policing, this study explored whether there were any AI tools and datasets perceived to be useful for serious crime investigations. The rationale was that identifying which data types and tools were truly useful could guide future research and help police make better decision about what to adopt. The analysis of the participants' views suggested various tools and datasets could be useful, but this thesis does not examine how they would be implemented in serious crime investigations. Future research can concentrate on use cases and test these tools and datasets based on investigative needs to assess their effectiveness within the operating police force.

5.7.1 AI tools

A5 noted that it was challenging to identify the most useful AI tools because there are many and each has its own specific features. The findings from other participants support A5's point, showing different tools serving different purposes in policing. However, a critical analysis evidenced that there are several AI tools that can be used by the police to analyse big data and advance serious crime investigations, offering the advantages discussed in Theme 3 (see 5.4): anomaly detection, clustering-based techniques, deep learning, document analysis, facial recognition, link analysis, natural language processing, pattern recognition, predictive modelling, social network analytics, and text and image analytics (P2, P4, A1, A4, A5, A6 and A8). The literature also indicated that the rapidly increasing volume and diversity of big data means analytics tools, techniques and methods are needed to manage datasets and extract information and knowledge from them (Zainab and Dhanda, 2018; Feng et al., 2019).

In relation to financial crimes, classified as among serious crimes, the A1 tools proposed as of potential use in investigations included pattern recognition, link analysis, document analysis and natural language processing. As an example of the usefulness of such tools, AI noted that when 20 computers are seized and require a full examination for investigative purposes, a manual analysis would require substantial time and staff, whereas automated analytics can deliver much faster results. This issue was also represented in the usage of AI tools in the literature. For instance, machine learning models were reported as being used for crime detection purposes by the Serious Fraud Office in the UK, inspecting millions of documents during an investigation to identify legally privileged materials (Vestby, 2019). In addition, machine learning was employed for crime prevention purposes by the Norwegian Labor Inspection Authority to predict high-risk workplaces to be inspected by the agency in Norway (Jha, Sivasankari and Krishnappa, 2020). These studies found that such analytical tools can process digital evidence at a scale which human investigators cannot achieve within the same timeframe. They can

rapidly search and compare vast amounts of material across numerous domains to assist investigators in identifying relevant information. These examples of applications illustrate the value expected from analysing big data, matching some of the potential benefits identified earlier in Theme 3, such as pattern extraction, crime detection and prediction, and the provision of greater insight.

However, despite the capabilities of such tools, P9 identified a significant challenge in terms of the chance of missing a small but critical piece of evidentiary information due to the high volumes of data. Accordingly, the outputs of these tools should not be treated as conclusory and investigators should not rely on these tools alone, especially where there is any doubt about the potential to miss critical information or where the items seized are the sole potential source of evidence in a case. In addition, if a case is time sensitive, a safer approach would be to have investigators adding layers of checks. Thus, to mitigate problems, more than one tool can be used to avoid blind spots and a second review can be conducted manually for high-risk items and sensitive cases.

According to A4, deep learning is another tool frequently used in the field of big data and works well with unstructured and high volumes of data in policing projects. Deep learning has the advantage of handling high volumes of unstructured data, previously identified as a technical challenge in Theme 4 (see 5.5). Hence this tool can be advantageous for policing in training algorithms to analyse big data and thus advance serious crime investigations, aiming to achieve the potential benefits discussed in Theme 3 (see 5.4).

In addition, A4 pointed to predictive models as among the important techniques that can be used in policing, predicting future crimes by analysing historical data. Similarly, in the literature, Pramanik et al.'s (2017) study suggested that AI algorithms be developed to support automated data learning and build models for crime prediction and detection, criminal behaviour profiling and criminal data clustering. Predictive analytics is an advanced method that allows users to make use of both real-time and historical data to assist in predicting crimes, events and behaviour (Zainab and Dhanda, 2018), representing some of the potential advantages found in Theme 3 (see 5.4). In addition, A4 and A8 proposed that anomaly detection and clustering-based techniques can be useful in policing to identify crime hotspots and direct police patrolling of these areas. Moreover, Vestby (2019) noted that various predictive software packages draw on machine learning. Hence, combining the tools and techniques available with machine learning, noted by Vestby (2019) as a key technology in the field, could offer a more data-driven training approach to be used by policing in serious crime investigations.

P2, P4 and A6 also pointed to facial recognition, ANPR, text and image analytics and social network analytics as useful. In serious crime investigations, facial recognition and ANPR can swiftly compare images from CCTV, ANPR records and crime scenes against databases to help investigators identify

suspects and missing persons. Such tools can also reveal patterns of movement and repeated appearances at specific locations to help link individuals and vehicles to certain events, providing avenues to obtain additional evidence. Text analytics can process high volumes of emails, social media posts and messages to uncover hidden connections, while image analytics can help detect objects or identify locations to help investigators. Social network analysis can map relationships between people, organisations and suspected criminal groups and networks. These capabilities are highlighted as perceived advantages of big data analytics in policing, helping build a better understanding of crimes and providing insights in terms of tracking criminal activity, identifying patterns, predicting incidents, deploying resources effectively and improving decision making (Zainab and Dhanda, 2018; Feng et al., 2019).

The literature reviewed also showed a strong focus on data mining and its tools as useful methods to analyse big data in policing (Nath, 2006; Hajian, Domingo, and Martinez-Balleste, 2011, 2016; Sharma, 2014; Tayal et al., 2015; Tomar and Manjhvar, 2016; Pramanik et al., 2017; Selbst, 2017; Prabakaran and Mitra, 2018; Jha, Sivasankari and Krishnappa, 2020). However, the empirical data did not highlight data mining. A possible explanation is that the participants focused more on the tools presented and thus data mining did not play a major role in their accounts. However, criminology is one of the key areas in which data mining and its tools have the potential to achieve positive results (Sharma, 2014; Hassani et al., 2016). Some police forces and security agencies in the US have started to use different data-mining tools and techniques for crime prevention and detection, including in the field of terrorism (Pramanik et al., 2017).

The review of the literature identified 33 data-mining tools and techniques (see section 2.5.2) of potential use in investigating crimes (Nath, 2006; Hajian, Domingo, and Martinez-Balleste, 2011; Sharma, 2014; Hassani et al., 2016; Tomar and Manjhvar, 2016; Pramanik et al., 2017; Prabakaran and Mitra, 2018; Jha, Sivasankari and Krishnappa, 2020). This contributes to the thesis by presenting the wide range of analytical options available for policing. It also provides an overview that can guide future research and assist policing practitioners in considering which tools may be most suitable for their investigative needs and policing projects.

Moreover, the literature showed that data mining can be used in investigating and detecting crimes such as fraud, as well as online, violent and sexual crimes (Prabakaran and Mitra, 2018). For fraud detection, three data-mining techniques are recommended: genetic algorithms, hidden Markov models (HMM) and naïve Bayes. For violent crimes, the Fuzzy c-means algorithm can be used to cluster data and has been shown to give better results than the K-means algorithm (Prabakaran and Mitra, 2018). For sexual crimes, the recommended data-mining techniques are kernel density estimation, logistic regression and the random forest algorithm, whereas for cybercrimes, influence-based association rules and the J48 algorithm are proposed (Prabakaran and Mitra, 2018).

In terms of software, there are various options, among which is Apache Hadoop, a widely known big data open-source platform that can process high volumes of data through distribution across a cluster of machines able to handle structured, semi-structured and unstructured data (Cardenas, Manadhata and Rajan, 2013; Thabet and Soomro, 2015; Wadhwani and Wang, 2017). Moreover, a key impact of big data technologies is their ability to provide low-cost infrastructures for security purposes, with tools such as Hive, Pig, RHadoop, stream mining and NoSQL being examples (Cardenas, Manadhata and Rajan (2013). However, as discussed in Theme 4 (see 5.5), the participants reported that big data technologies are expensive and financially challenging for policing. In this regard, it should be noted that the claims about affordability in Cardenas, Manadhata and Rajan (2013) may no longer hold true given the current costs of technology and market conditions.

In addition to the tools identified earlier, the literature also referred to several systems reported to support crime prediction. Examples of these are PredPol and HunchLab in the US, PRECOBS in Germany, KeyCrime in Italy, Maprevelation in France and the Crime Anticipation System (CAS) in the Netherlands (Schuilenburg and Soudijn, 2023). PRECOBS is claimed to be one of the leading software systems for predictive policing in German-speaking countries (Egber and Krasmann, 2019). In addition, the Durham Constabulary in the UK implemented the Harm Assessment Risk Tools (HART), which estimates the likelihood that an offender will commit another crime within two years of release from prison, thus identifying those who may benefit from rehabilitation programmes.

The predictions provided by such software applications are based on comparative mathematics derived from computer algorithms that can conduct high speed analysis of big datasets (Sandhu and Fussey, 2021). However, it remains unclear how effective such software is in making predictions and whether these platforms should be classified as big data applications in policing. This will depend on the data sources, volume of data and analytical methods employed. Thus, it is questionable whether they should be considered big data applications rather than crime prediction software and examples of emerging technology. Therefore, further research is needed to assess their performance and examine whether these systems face the same challenges identified in Theme 4 (see 5.5) and concerns in Theme 5 (see 5.6).

5.7.2 Types of datasets

Researching the types of datasets used in big data analysis is important because data are the core component. Also, exploring types of datasets can help clarify practical requirements, enabling policing organisations to plan the development and necessary upgrading of their technical infrastructure. As noted by P4, besides the availability of various AI tools presented above, their use by police forces will depend greatly on the type of data they possess. The results did not describe how datasets could be used beyond identifying them as beneficial for serious crime investigations.

The participants referred to the following types of data: CCTV recordings, communications data, historical crime data, criminal and traffic system data, emails, emergency service calls, government records, IP addresses, live camera feeds, mobile geolocation data, open-source intelligence, phone numbers, social media, tax records, and telephone behaviour (P1, P2, P4, P6, P7, P9, A1, A2, A4, A7, and A8). Analysing data in policing entails translating raw information into operationally practical knowledge (James, 2016). Therefore, the purpose of collecting and analysing big data is to convert the information embedded in the types of data mentioned above into actionable insights that can support serious crime investigations.

As P1 pointed out, nothing can be done in such investigations without data. P2 argued the prime importance of text data, provided they were from a trusted source, followed by audio and video. The importance of trusted text data derived from their ability to help investigators build timelines, link related information and search for key details in large volumes of data. Also, text data tend to be widely available from databases such as criminal and traffic systems, as well as emails, social media and records of phone behaviour. Audio data are also important as these files capture content that might not appear in written records, such as emergency calls and social media, potentially providing significant details about events. In addition, video data have value as they can provide visual evidence of people and their actions through CCTV recordings, facial recognition software and social media.

Social media was frequently mentioned as one of the data sources for text, audio and video and P1, P9, A4 and A8 all suggested that information from social media platforms could be useful in serious crime investigations. However, P2 and A2 cautioned that they are non-trusted sources of data, as there are millions of fake accounts and details. These different views suggests that social media offers both benefits and limitations in serious crime investigations. Some participants viewed social media as useful because it could provide information quickly about people, events and place, potentially helping generate leads in an investigation, whereas others viewed such data as unreliable. Hence, social media may best be considered a source of intelligence rather than firm evidence, unless verified by additional sources of information. Through this approach, social media can be exploited as a source while also reducing the risk of acting on inaccurate information.

According to the European Commission (2020), electronic information and digital evidence are required in roughly 85% of serious crime investigations, reflecting the findings of earlier studies that modern investigations rely on data-driven inputs in digital formats, such as text, audio and video (Babuta, 2017; Home Office, 2023). Also, this supports the rationale of identifying relevant types of database types, as it aligns with the discussion of technical challenges (see 5.5.3). Given the vital role of electronic evidence in investigations, policing must be prepared to store, manage and analyse high volumes of data. P4 noted that firsthand data are of particular use during investigations due to their reliability, but such data are

very expensive. This reflects the financial challenges identified in Theme 4 (see 5.5.2), with the pressures of cost extending beyond developing technical infrastructure and recruiting skilled human resources to obtaining reliable firsthand data from trustworthy sources. Regarding the reliability of data, P7 and P9 argued that numerical data, such as phone numbers and IP addresses, tend to be mostly accurate, in addition to data on telephone behaviour. In addition, P9 highlighted the importance of communications data and emails.

The literature offered other types and sources of data, such as DNA databases (Neiva, Granja and Machado, 2022), facial recognition (Neiva, Machado and Silva, 2023; Tamim, 2024), the Europol Information System and the Prum System (Schuilenburg and Soudijn, 2023), and I-LEAP (Home Office, 2023). Besides DNA and facial recognition databases, which often have national applications, the other systems are international, covering European countries. Such databases are large and interconnected, making them useful across borders in policing and serious crime investigations. This reflects the cross-border nature of serious crimes (Europol, 2023; Home Office, 2023), since international data sharing can support identification and the verification and linking of information beyond national boundaries.

Moreover, the literature identified several big data projects in policing but did not specify the types of data or online platforms that were used; examples include the French Police (Neiva, Granja and Machado, 2022) and the Canadian Police (O'Connor et al., 2022). Identifying the types of datasets and platforms used in active big data projects could be beneficial for other police forces, potentially showing what datasets are required, how they are managed and the required infrastructure, as well as providing transparency. Overall, this section shows that policing can draw on a wide range of datasets in relation to big data, with each type offering different investigative advantages and limitations.

5.8 Conclusion: Moving towards ensuring the usefulness of big data in serious crime investigations

The research question and aims of this thesis sought to establish if big data could be useful in serious crime investigations. As stated at the beginning of this chapter, this thesis found enthusiasm and overall agreement concerning the use of big data in policing. However, to move to the point where analysing big data can properly be utilised and achieve its potential in serious crime investigations, a working definition and concept of big data needs to be established; such a definition for policing has been developed and presented in Theme 3 (see 5.4).

This should be followed by mitigating the challenges identified, namely, the adoption of a coherent concept and issues concerning financial constraints, the availability of qualified human resources and technical and operational obstacles. Technical implementation barriers are being overcome as technology advances, being that they are largely, although not fully, about technological developments,

such as storage and processing power. However, without sufficient funding and effective management, even such technical challenges may remain unsolved.

Effective management is not only about handling finances, even when enough funding is available. It also includes encouraging organisations to adopt big data and adapt to technological advances, as well as recruiting qualified human resources to successfully implement projects and attain the potential advantages of big data. Alongside tackling these challenges, it is equally important to address the concerns of privacy and bias, as the findings suggested that these are more salient in policing than in other fields. This highlights the need for legal frameworks suited to address the use of big data in policing, directly addressing privacy- and bias-related risks.

Overall, this thesis suggests that if the challenges and concerns are addressed, ensuring the availability of the appropriate AI tools and high-quality big data sets, robust regulations and policies will be required to ensure the safe and proper use of AI to analyse big data, ensuring it is useful for serious crime investigations, both operationally and strategically. Figure 5.2 suggests how these aspects might theoretically be achieved in a potential virtuous hierarchical framework. The figure presents the core elements that could lead to the operational and strategic advantages of using big data in serious crime investigations identified in this thesis.

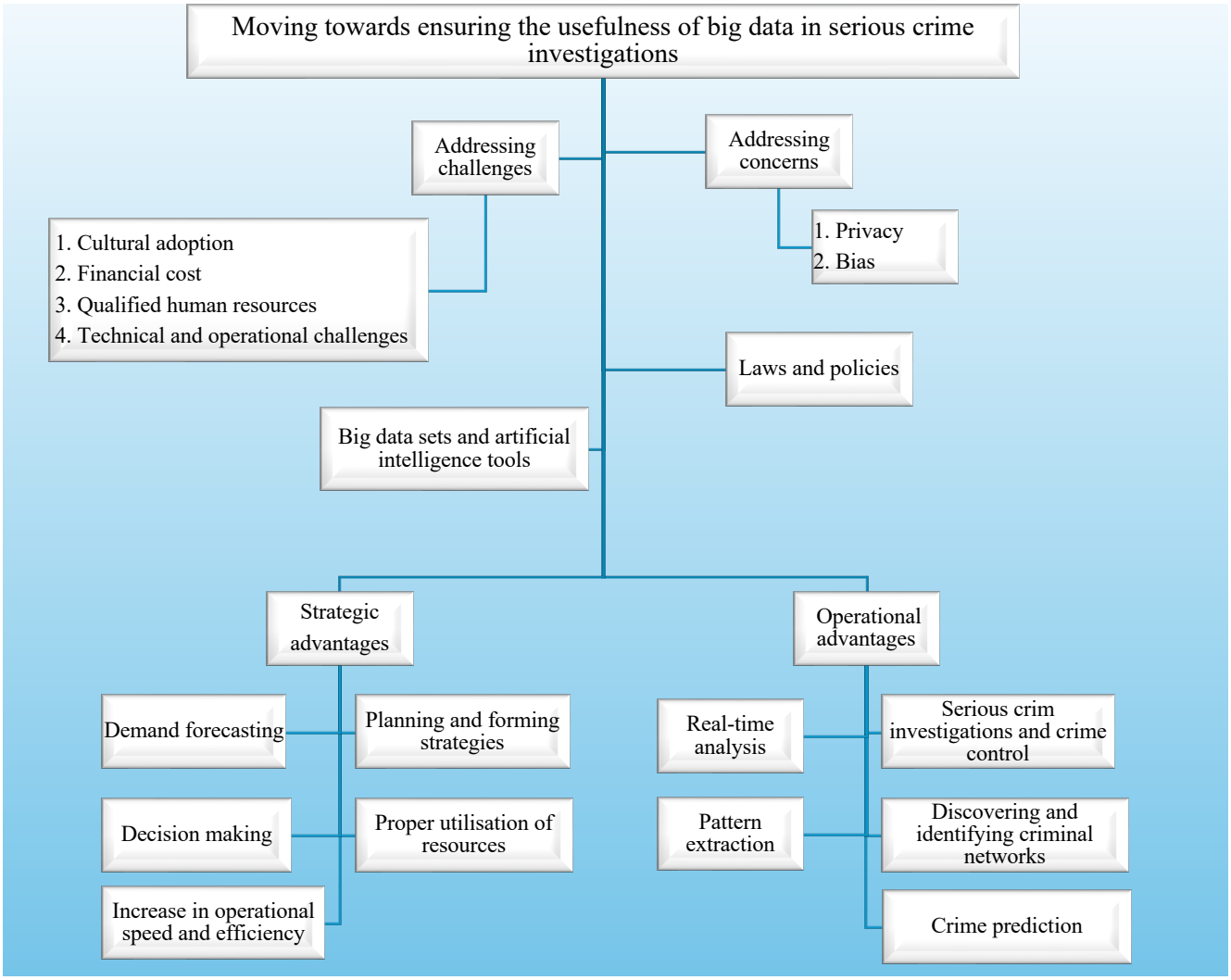

Figure 5.2. Moving towards ensuring the usefulness of big data in serious crime investigations.

Figure 5.2 provides an analytical overview of the discussion by grouping the findings into the main components. The structure of the discussion demonstrates how each element contributes to answering the research question and achieving the research aims, providing visual mapping to help the reader by synthesising the findings into a coherent interpretation.

A critical interpretation of the preceding challenges suggests that their combined effect manifests as an operational barrier within policing. To be able to overcome the challenges confronting the use of big data, it is first necessary to understand its complexities (Wadhwani and Wang, 2017). The findings of this study highlight the main challenges, representing a first step towards understanding the conceptual, technical, financial and human resource complexities associated with big data, which were highlighted by 82% of the participants.

It can be inferred that there are two main challenges that need to be overcome and are central to developing a big data project for policing and specifically to advance serious crime investigations, namely financial and human resource challenges. Having sufficient and sustained funding is

fundamental to recruit skilled data specialists who can design, develop and manage the infrastructure of big data systems. These professionals can not only bring the technical expertise required to make sense of the data but can also promote awareness and understanding of big data across different levels of the organisation. Their involvement can help encourage a more data-driven culture and reduce the resistance that often faces the adoption of new technologies or working methods. Overcoming this conceptual barrier is an important step towards building a vision of how big data can support serious crime investigations and achieve the perceived advantages.

Having established the cultural and conceptual foundations, the police force could be in a better position to move forward strategically. This includes developing plans to strengthen the technical infrastructure and address the main challenges related to data quality, managing and storing high volumes of data, addressing velocity and variety, and establishing clear governance frameworks. These elements are essential: as suggested by the findings, without a stable infrastructure, even the most skilled human resources will be constrained and without knowledgeable staff, the technical systems are unlikely to be utilised to their full potential. Hence tackling the challenges posed by financial and human resource issues will form the foundation for developing both conceptual understanding and technical capability. This is an operational and strategic necessity to achieving the potential advantages of big data, as when financial investments and skilled data specialists come together, policing organisations will be better equipped to utilise big data effectively.

Finally, despite the challenges and concerns identified in relation to big data usage, the study found that big data projects are being implemented in policing (1.1.3) and (2.5.5). However, the existence of these projects does not necessarily mean that the issues have been resolved, as the findings indicated that they are ongoing. It is not clear whether these policing organisations managed to overcome these issues or whether certain challenges or concerns were not fully addressed in practice. Overall, this indicates a lack of clarity about how risks are handled and suggests a possible gap between deployment and effective oversight.

5.9. Contributions to knowledge

This section outlines the primary contributions of this thesis, presented as empirical, theoretical, and practical contributions to knowledge.

5.9.1 Empirical contributions

This thesis makes an empirical contribution by providing evidence that utilising big data has potential for serious crime investigations. Research specifically in this context is scarce. The research question identified the forms of usefulness of big data and the challenges and concerns regarding its application in serious crime investigations. Taken together, this thesis clarifies both the promise and limitations of

current knowledge, indicating where conclusions about usefulness in policing are supported by evidence and where they remain preliminary.

The evidence provided in this thesis shows central challenges and concerns that must be addressed to ensure successful implementation of big data analytics. The findings provide evidence from both research and practice to an area that may be influenced by technical promises and expectations than by actual empirical insights, especially in policing.

In addition, this thesis brought together two groups of participants, interviewed separately but often required to work together in practice, especially in developing and operating big data projects to advance serious crime investigations. By purposefully including both groups, the thesis captured different perspectives, illustrating areas of agreement and difference in interpretation. This bridging approach produced empirical insights into how the two perspectives might be aligned, what could obstruct implementation and what organisational conditions are required for big data applications in policing.

5.9.2 Theoretical/conceptual contributions

This thesis contributes to the ongoing conceptual debates about how big data is defined and understood. The findings established that variety is not a temporary feature but a shaping component of big data. Variety is always present, as big data continues to evolve rather than being a fixed construct. Also, this thesis makes a theoretical contribution by proposing a new definition of big data in serious crime investigations, namely that it is: *High volumes of information that are diverse, variable and generated at high velocity, analysed by artificial intelligence tools to yield insights for policing to achieve strategic and/or operational advantages in serious crime investigations* (cf. 5.4.4). Conceptual debate and multiple definitions are common in academic work. However, in policing practice, a more stable definition may be necessary to support effective communication and shared understanding. The definition is grounded by the core foundations of big data and connected to the potential advantages of utilising it in serious crime investigations.

5.9.3 Practical contributions

Figure 5.2 presents a hierarchical framework that presents the core elements required to achieve the strategic and operational advantages of using big data in serious crime investigations. By organising the elements in a clear categorisation, the framework shows how the foundational requirements support the attainment of the desired outcomes. In practical terms, Figure 5.2 can also be used as a checklist to support planning and decision making, enabling policing organisations to assess whether key components are in place, such as financial funding, qualified human resources, technical tools, and privacy and ethical protocols, before proceeding to developing or deploying a big data project. This can

help reduce the risk of implementing systems that are technically possible but operationally weak, difficult to manage, or vulnerable to privacy and ethical concerns.

Chapter 6. Conclusion

6.1 Overview of the thesis

This thesis has examined what is known about big data and its potential usefulness for policing and specifically for serious crime investigations. It has aimed to develop a clear understanding of big data and a conceptual foundation to be applied in the context of serious crimes, thus addressing Research Aim 1, and to assess its usefulness while recognising technical, operational, organisational and ethical challenges that shape implementation, addressing Research Aim 2. In addition, a practical definition that can be used by policing was developed by addressing Research Aim 3. Collectively, these achievements provide clear insights to address the overarching research question.

To address these aims, the thesis combined a scoping review with a qualitative empirical research approach, such that the review of the literature provided a structured account of how big data has been conceptualised, how it has developed and expanded, and its potential advantages across fields and within policing. The empirical study then explored the professional perspectives of two participant groups, comprising police officers and big data experts, through semi-structured interviews. A critical realist framework, supported by a relativist epistemological stance, guided the thesis in considering the accounts as context-dependent evidence while allowing reasonable judgement in evaluating the claims. Reflexive thematic analysis was applied to develop themes, which represented common patterns across the data collected. To maintain consistency, the same thematic structure of six themes, articulated in the findings (Chapter 4), was followed in the discussion (Chapter 5) to allow a systematic analysis of each theme in relation to the existing literature.

In terms of answering the research question, this thesis found that big data is widely discussed but not consistently defined. The literature showed that the term emerged through the practical challenges of managing high volume datasets and it is commonly described in terms of the three Vs – volume, velocity and variety – together with various additional characteristics. However, both the literature and empirical findings from the interviews indicated that there is no universal definition, which lead to the development of Research Aim 3. At the same time, all the participants demonstrated a functional understanding of big data, confirming that it is operationally understood even when there is no conceptually standardised definition. Hence, the objective of Research Aim 3 was to establish a stable and actionable definition of big data to be used internationally by synthesising the commonalities in the findings to create a foundational baseline for future practical and academic development (see 5.4.4).

The central conclusion of this thesis is that big data can be useful for policing in serious crime investigations, but its usefulness is conditional not guaranteed. Strong and consistent support for its usefulness was observed across the interviews, with the participants linking its potential to improved

decision making, effective deployment of resources, greater operational efficiency and advanced capabilities for crime detection, prediction and prevention (see Figure 5.1).

However, the findings also highlighted that usefulness was widely viewed as conditional. Notably, the participants frequently emphasised that achieving the potential advantages would depend on having clean data, the appropriate analytical and technological capabilities, and organisational capacity in terms of cultural acceptance and qualified human resources to manage and analyse the high volumes of data. In addition, it would be necessary to address concerns of biased outcomes and privacy risks. The review of the literature strengthened this conclusion, showing that recent policing and security strategies are increasingly emphasising the role of data and advanced analytical tools, such as AI applications, even when not specifically referencing "big data". At the same time, the evidence base in policing is still developing, especially in terms of the effectiveness of evaluating some big data applications. This gap between ambition and operational evidence is one of the key justifications for conducting this research.

6.2 Recommendations

Building on the findings presented in Chapters 2 and 5, this section outlines the recommendations for practice and policy. These recommendations are intended for police forces and academia and are most applicable in the context of advancing serious crime investigations.

6.2.1 Recommendations for academia

The recommendations here are intended to support future research on the application of big data in policing and specifically in serious crime investigations, aiming to producing robust research-based evidence that can inform policing policy and practice.

Recommendation 1: Conduct empirical research on existing big data projects in policing

Future research should examine ongoing big data projects to identify if they are achieving any strategic, operational, or other types of advantage and what challenges they face during implementation and use. In particular, studies should assess how risks and concerns related to privacy and bias are addressed in practice, including whether they are mitigated, managed, or possibly left unresolved. In addition, the technical and operational approaches that are used to develop such projects should be explored. Such research can document the full project pathway, including how systems are designed, how data are collected and processed, how the outputs are analysed, and how policing performance is evaluated over time in relation to criminal investigations. This would strengthen existing evidence by moving beyond scoping the potential advantages to assessing verified operational outcomes, identifying effective safeguards and clarifying financial costs.

Recommendation 2: Expand comparative and cross-national research on serious crimes and policing strategies

Given that much of the available literature on strategies aiming to tackle serious crimes is UK-based, further research and publications should include evidence from a wider range of countries and police forces. While the existing literature is valuable, there are limitations in terms of generalisation and reciprocity because policing is shaped by national and international differences in operational practice, governance structures and public trust and relationships with the police. Expanding cross-national research would strengthen the evidence by enabling meaningful comparison of what approaches work for whom and under what conditions. Comparative studies could identify which elements of big data practice are broadly transferable and which elements depend on local context. Overall, this would reduce reliance on narrow geographic evidence base and support more informed research, evidence and practice-based decision making across policing.

6.2.2 Recommendations for practice

This section outlines the recommendations for practice, focused on improving the readiness of police forces not only to keep pace with rapid technological developments but also to shape how these tools are adopted and governed in practice.

Recommendation 1: Develop adaptive legal and regulatory frameworks for big data and artificial intelligence usage in policing

Law makers and/or policing regulators should develop legal frameworks that are fit to regulate and cope with the rapid pace of technological advancements in big data and related analytical tools. These frameworks can be treated as a first step towards addressing the key concerns of privacy and bias by setting clear standards for lawful and proportionate use of big data. In practice, this should include explicit requirements for privacy protection, bias assessments, transparency and clear accountability for data-driven decisions. Also, such legislation could include periodic reviews aimed at safeguarding privacy and keeping pace with technological developments and their potential risks.

Recommendation 2: Build balanced technological awareness across policing roles

Police forces should provide training in technological awareness for staff at all levels, from senior commanders to frontline officers, so they can understand both the potential benefits of technology and its risks. Cultural resistance can arise from limited clarity about what can realistically be achieved through employing advanced AI tools, alongside concerns about privacy and bias. Being informed about the advantages and concerns can shape how staff perform in their day-to-day tasks to support a responsible data-driven approach. Greater awareness can encourage more accurate data entry, consistent recording standards, careful case classification and timely updates of incident records. It can also improve decisions about what data it is appropriate to collect and utilise and when to escalate potential

privacy or bias concerns. Over time, these actions will contribute to the quality and reliability of police databases, which can be analysed using big data tools to attain the desired results. Such an approach would support both operational effectiveness and stronger governance by linking daily practices to the reliability of big data-driven policing.

Recommendation 3: Invest in police skills and capabilities to lead technological change

Police forces should invest in officers and staff, providing education, training and professional development so they are not only familiar with new technologies but are able to lead effective and responsible usage. Developing in-house capabilities can enable staff to make informed routine decisions regarding the adoption, use and supervision of technological tools, rather than relying on external expertise from private companies. By having strong knowledgeable and experienced staff internally, police forces will be in a better position to question private vendors' claims about technological capabilities, set clear requirements and maintain control over sensitive data. This could lead to higher standards of privacy and data security, limiting unnecessary data sharing, improving oversight and ensuring that risks remain within the organisation.

6.3 Conclusion

This concluding section provides an overall synthesis of the thesis, drawing together its main arguments, key findings, and overall contribution.

Chapter 1 established the background and context of the thesis by examining big data and policing, its current applications, the different forms of serious crime, and the strategies used to address such crimes in the UK and Europe. It highlighted the expansion and uncertainty of big data in the policing context and the extent to which it can offer practical value in serious crime investigations. The chapter also underlined the ethical complexity of the phenomenon under study and established the rationale and significance of researching the main two fields. Finally, it outlined the research aims and questions, the scope of the thesis, its methodological and theoretical positioning, and its overall structure.

Chapter 2 introduced the scoping review methodology adopted in the thesis, which was guided by Arksey and O'Malley's (2005) five stage framework. It outlined the stages of identifying the research questions, identifying the relevant studies through database and website searches, applying eligibility criteria and study selection procedures, and charting the data extracted from the literature. The scoping review section also explained how the 84 sources informed the empirical study and supported the development of key themes from the literature. Finally, it outlined how the reviewed literature was collated, summarised and reported to provide an interpretive understanding of the findings. Chapter 2 comprised three sections. Section 1 identified the expansion of big data, its use in the healthcare and financial sections, the potential and challenges associated with its applications across both sectors. The

findings suggested that big data continues to expand as an emerging discipline, although its meaning has remained insufficiently defined within the field. The lack of conceptual clarity provided the basis for Section 2, which developed a more structured understanding. Section 2 established a baseline understanding of big data by examining how the concept has developed, and how it has been understood and defined in the literature. It was found that there is no single agreed definition or set of characteristics, and this lack of consensus appeared to be shaped by the continuing technological developments which have repeatedly changed what is understood as big data. Taken together, these findings highlighted the need for this thesis to develop a clear and stable working definition of big data within the policing context. Building on the foundation established in Chapter 1 on big data and policing, Section 3 continued the review by focusing more specifically on the potential usefulness of big data in serious crime investigations. In doing so, it moved from the broader background discussion to a more focused body of literature that is aligned with the aims of the thesis. More specifically, the section examined the uses of big data in police resource management, and crime detection prediction, and prevention. It also explored the challenges associated with the use of big data in policing and serious crime investigations. Finally, it explored the expectations of policing professionals regarding big data. The findings suggested that the increasing complexity of serious crime has strengthened the drive for more advanced policing capabilities, with big data showing potential to support crime detection, prediction and prevention. At the same time, the findings indicated that the evidence on how big data is applied in practice is still emerging, and the extent to which it achieves useful outcomes is not yet fully understood or clear. Taken together, these findings supported the need for this thesis to develop an operational understanding of big data in serious crime investigations.

Chapter 3 outlined the development of the research question and the rationale for the methodological approach adopted in this thesis. It then presented the conceptual and theoretical frameworks that informed the thesis, in addition to presenting the sampling strategy and the rationale for including two groups of participants, which together resulted in a total sample of 17 participants. The chapter also described the qualitative methodology applied, namely, the use of semi-structured interviews for data collection, reflexive thematic analysis for analysing the data and researcher positionality. Collectively, the sections outlined set out a clear foundation for the methodological approach as ethically appropriate, rigorous and consistent with the study aims.

Chapter 4 built on the literature review by addressing the gap identified in relation to the potential use of big data in serious crime investigations. The extent to which the theoretical findings aligned with day-to-day practices of professionals in the field remained unclear. Therefore, this thesis aimed to explore the perspectives of the two groups of professionals, aiming to provide valuable insights into how big data and serious crime are understood and potentially applied in practice, while also helping to validate or challenge the findings from the literature. Using Braun and Clarke's (2020) reflexive thematic

analysis, six themes were developed: defining big data, big data and policing, advantages and disadvantages, challenges in using big data in serious crime investigations, concerns regarding the use of big data in serious crime investigations, and tools and datasets. Overall, the six themes showed that the participants did not hold a consistent definition of big data, although there was broad agreement that it could support policing by improving decision making, enabling better deployment of resources, and advancing approaches in serious crime detection, prediction and prevention. Big data was viewed as a tool to enhance operational efficiency and investigative capabilities rather than replacing professional policing judgment. At the same time, the participants identified organisational, technical, financial and operational challenges, alongside key concerns related to bias and privacy. Taken together, these six themes provided the empirical foundation for Chapter 5.

Chapter 5 explored the usefulness of big data in serious crime investigations through an interpretive discussion of the empirical data in relation to the literature. Given the lack of consensus regarding the concept of big data, the chapter first clarified the concept by discussing its definitions and characteristics. As the findings indicated, there is no single agreed definition of big data, with understandings varying across disciplines. Accordingly, this thesis synthesised the insights derived from Themes 1, 2 and 3 to develop a definition that provides conceptual clarity and supports a more consistent understanding and application of big data in policing and particularly in the context of serious crime investigations. The findings highlighted a drive to address serious crime by recognising the substantial harm it causes to societies and countries. In addition, the findings showed close alignment between the participants' perspectives and the literature in supporting the positive role of big data in serious crime investigations. By critically analysing the findings, this chapter identified both strategic and operational advantages for policing.

At the same time, the findings emphasised important cultural, financial, technical and human resources challenges affecting the implementation of big data in policing. The insights included concerns related to bias and privacy, both of which may have significant implications for policing practice. Privacy emerged as a central concern in relation to personal data protection, data security and legal safeguards, whereas bias in the interpretation of big data is related to possible biased outcomes and results. The findings further suggested a range of AI tools and types of datasets that may support the use of big data in serious crime investigations, although the empirical data placed less emphasis on data mining than the literature. Overall, these insights demonstrate the breadth of the analytical options available to policing, which may guide and inform future research and practice. Finally, Chapter 5 concluded by discussing and presenting the empirical, theoretical and practical contributions of the thesis.

Chapter 6 offered overview of the thesis and concluded by presenting recommendations for both academia and practice. The recommendations for academic emphasise the need for further empirical

research on existing big data projects in policing, as well as comparative and cross-national research on serious crime and policing strategies. The recommendations for practice focus on the development of adaptive legal and regulatory frameworks for the use of big data and AI in policing, building balanced technological awareness across policing roles, and investment in police skills and capabilities to support technological advancements. This thesis concludes that there is clear and credible support for the application of big data in serious crime investigations based on its potential to deliver strategic and operational advantages for policing. However, the evidence suggests that the usefulness of big data is constrained by certain challenges and concerns that need to be addressed. This thesis lays the foundations for future research and provides support for more informed decision making about adopting big data in policing within the context outlined above.

References

Abdul Jalil, M., Mohd, F. and Noor, N.M. (2017) A comparative study to evaluate filtering methods for crime data feature selection. *Procedia Computer Science*, v. 116, pp.113-120.

Abouelmehdi, K., Beni-Hessane, A. and Khaloufi, H. (2018) Big healthcare data: preserving security and privacy. *Journal of Big Data*, v. 5(1), pp.1-18.

Adams, W. (2015) Conducting semi-structured interviews. In: Newcomer, K.E., Hatry, H.P. and Wholey, J.S. (eds.) *Handbook of practical program evaluation.* San Francisco, CA: Jossey-Bass, pp.492-505.

Aderemi, S., Olutimehin, D.O., Nnaomah, U.I., Orieno, O.H., Edunjobi, T.E. and Babatunde, S.O. (2024) Big data analytics in the financial services industry: trends, challenges, and future prospects: a review. *International Journal of Science and Technology Research Archive*, v. 6(1), pp.147-166.

Almansoori, R. (2019) *Major crimes police interviewing in Dubai: an examination of the investigative interviewing triangle* [online] Available at: https://pure.port.ac.uk/ws/portalfiles/portal/20473876/Thesis_final_with_amendments_.pdf [Accessed 3 October 2023].

APCC (2020) *National digital policing strategy 2020–2030.* The Association of Police and Crime Commissioners [online] Available at: https://www.apccs.police.uk/latest-news/national-digital-policing-strategy-2020-2030/ [Accessed 7 October 2024].

Arksey, H. and O'Malley, L. (2005) Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, v. 8(1), pp.19-32.

Assouli, N., Benahmed, K. and Gasbaoui, B. (2021) How to predict crime – informatics-inspired approach from link prediction. *Physica A: Statistical Mechanics and its Applications*, v. 570, pp.1-14.

Awrahman, B.J., Fatah, C.A. and Hamaamin, M.Y. (2022) A review of the role and challenges of big data in healthcare informatics and analytics. *Computational Intelligence and Neuroscience*, v. 1, pp.1-10.

Babuta, A. (2017) *Big data and policing,* London: Royal United Services Institute for Defence and Security Studies.

Balogun, A.Y., Olaniyi, O.O., Olisa, A.O. and Gbadebo, M.O. (2025) Enhancing incident response strategies in U.S. healthcare cybersecurity. *Journal of Engineering Research and Reports*, v. 27(2), pp.114-135.

Barbour, R.S. (2008) *Introducing qualitative research.* 1st ed. London: Sage Publications.

Barocas, S., Bradley, E., Honavar, V. and Provost, F. (2017) *Big data, data science, and civil rights.* Washington DC: Computing Community Consortium.

Barrick, L. (2020) Interviews: in-depth, semistructured. *International Encyclopedia of Human Geography*, v. 7(2), pp.403-408.

Bell, D., Lycett, M., Marshan, A. and Monaghan, A. (2021) Exploring future challenges for big data in the humanitarian domain. *Journal of Business Research*, v. 131, pp.453-468.

Bhaskar, R. (2008) *A realist theory of science.* 1st ed. London: Routledge.

Bignami, F. (2007) Privacy and law enforcement in the European Union: The Data Retention Directive. *Chicago Journal of International Law*, v. 8(1), pp.233-255.

Bloom, B.D. and Crabtree, B.F. (2006) The qualitative research interview. *Medical Education*, v. 40(4), pp.314-321.

Boyd, P. (2024) Reasoning within hybrid thematic analysis. *Link Journal*, v. 8(2), pp.1-14.

Brady, H.E. (2019) The challenge of big data and data science. *The Annual Review of Political Science*, v. 22(1), pp.297-323.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology*, v. 3(2), pp.77-101.

Braun, V. and Clarke, V. (2020) Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy in Research*, v. 21(1), pp.37-47.

Braun, V. and Clarke, V. (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis?. *Qualitative Research in Psychology*, v. 18(3), pp.328-352.

Brennan, I. (2022) *Victims of serious violence in England and Wales, 2011-2017* [online] Available at: https://osf.io/preprints/osf/uqkem [Accessed 16 March 2023].

Broeders, D., Schrijvers E., van der Sloot, B., van Brakel, R., de Hoog, J. and Hirsch Ballin, E. (2017) Big data and security policies: Towards a framework for regulating the phases of analytics and use of big data. *Computer Law & Security Review*, v. 33, pp.309-323.

Byrne, D. (2021) A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, v. 56(3), pp.1391-1412.

Cacchione, P.Z. (2016) The evolving methodology of scoping reviews. *Clinical Nursing Research*, v. 25(2), pp.115-119.

Cárdenas, A.A., Manadhata, P.K. and Rajan, S.P. (2013) Big data analytics for security. *IEEE Security and Privacy*, v. 11(6), pp.74-76.

Chan, J. and Moses, L.B. (2016) Making sense of big data for security. *The British Journal of Criminology*, v. 57(2), pp.299-319.

Clissa, L., Lassnig, M. and Rinaldi, L. (2023) How big is big data? A comprehensive survey of data production, storage, and streaming in science and industry. *Frontiers in Big Data*, v. 6, pp.1-5.

Coe, R., Waring, M., Hedges, L.V. & Ashley, L.D. (2021) *Research methods and methodologies in education.* 3rd ed. London: SAGE Publications.

Costanzo, P., D'Onofrio, F. and Friedl, J. (2015) Big data and the Italian legal dramework: opportunities for police forces. In: Babak, A., Saathoff, G.B., Arabnia, H.R., Hill, R., Staniforth, A. and Bayerl, P.S. (eds.) *Application of big data for national security. A practitioner's guide to emerging technologies.* Oxford: Elsevier, pp.238-249.

Cox, M. and Ellswort, D. (1997) Application-controlled demand paging for out-of-core visualization. *Proceedings. Visualization '97 (Cat. No. 97CB36155)*. IEEE Xplore, pp.1-13.

Cradock, E., Stalla-Bourdillon, S. and Millard, D. (2017) Nobody puts data in a corner? Why a new approach to categorising personal data is required for the obligation to inform. *Computer Law & Security Review*, v. 33, pp.142-158.

Creswell, J. W. (2012) *Educational research: planning, conducting, and evaluating* q*uantitative and qualitative research.* 4th ed. Boston: Pearson Education, Inc.

de Hert, P. and Papakonstantinou, V. (2009) The data protection framework decision of 27 November 2008 regarding police and judicial cooperation in criminal matters – A modest achievement however not the improvement some have hoped for. *Computer Law & Security Review*, v. 25(5), pp.403-414.

DeJonckheere, M., Vaughn, L.M., James, T.G. and Schondelmeyer, A.C. (2024) Qualitative thematic analysis in a mixed methods study: guidelines and considerations for integration. *Journal of Mixed Methods Research*, v. 3(18), pp.258-269.

Dijcks, J.-P. (2013) *Oracle: big data for the enterprise* [online] Available at: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf [Accessed 9 October 2022].

Dubai Police (2023) *Major crime statistics* [online] Available at: https://www.dubaipolice.gov.ae/wps/portal/home/opendata/majorcrimestatistics [Accessed 17 December 2023].

Dubai Police (2025) *Strategic goals* [online] Available at: https://www.dubaipolice.gov.ae/wps/portal/home/aboutus/ourstrategy/value?lang=en [Accessed 30 January 2025].

Dumbill, E. (2013) *Making sense of big data* [online] Available at: https://www.liebertpub.com/doi/abs/10.1089/big.2012.1503 [Accessed 9 October 2022].

Dworkin, S.L. (2012) Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior*, v. 41(6), pp.1319-1320.

Easwaramoorthy, M. and Zarinpoush, F. (2006) *Interviewing for research.* Toronto: Imagine Canada.

Egbert, S. and Krasmann, S. (2019) Predictive policing: not yet, but soon preemptive? *Policing and Society*, v. 30(8), pp.905-919.

European Commission (2020) *EU Security Union Strategy: connecting the dots in a new security ecosystem* [online] Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1379 [Accessed 10 August 2024].

Europol (2023) *Defining serious and organised crime* [online] Available at: https://www.europol.europa.eu/socta/2017/defining-serious-and-organised-crime.html [Accessed 12 July 2024].

Ezzeddine, Y., Bayerl, P.S. and Gibson, H. (2023) Safety, privacy, or both: evaluating citizens' perspectives around artificial intelligence use by police forces. *Policing and Society*, v. 33(7), pp.861-876.

Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y. and Liu, Q. (2019) Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access*, v. 7, pp.106111-106123.

Ferguson, A.G. (2015) Big data and predictive reasonable suspicion. *University of Pennsylvania Law Review*, v. 163(2), pp.327-410.

Fletcher, A.J. (2016) Applying critical realism in qualitative research: methodology meets method. *International Journal of Social Research Methodology*, v. 20(2), pp.181-194.

Gkikas, D.C. and Theodoridis, P.K. (2022) AI in consumer behavior. In: Virvou, M., Tsihrintzis, G.A., Tsoukalas, L.H. and Jain, L.C. (eds.) *Advances in artificial intelligence-based technologies.* Cham: Springer International Publishing, pp.147-176.

Golafshani, N. (2003) Understanding reliability and validity in qualitative research. *The Qualitative Report*, v. 8(4), pp.597-606.

Government Digital Service (2018) *Data ethics framework.* [Online] Available at: https://assets.publishing.service.gov.uk/media/5f74a4958fa8f5188dad0e99/Data_Ethics_Framework_2020.pdf [Accessed 23 August 2020].

Hajian, S., Domingo-Ferrer, J. and Martínez-Ballesté, A. (2011) Discrimination prevention in data mining for intrusion and crime detection. *2011 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, Paris, France, 2011, pp.47-54.

Harrell, M.C. and Bradley, M.A. (2009) *Data collection methods: semi-structured interviews and focus groups.* Santa Monica: National Defense Research Institute.

Hassani, H., Huang, X., Silva, E.S. and Ghodsi, M. (2016) A review of data mining applications in crime. *Statistical Analysis and Data Mining*, v. 9(3), pp.139-154.

Hennink, M. and Kaiser, B.N. (2022) Sample sizes for saturation in qualitative research: a systematic review of empirical tests. *Social Science & Medicine*, v. 292, pp.1-10.

Home Office (2024) *Forensic information databases annual report 2023 to 2024* [online] Available at: https://www.gov.uk/government/publications/forensic-information-databases-annual-report-

2023-to-2024/forensic-information-databases-annual-report-2023-to-2024-accessible?utm_source=chatgpt.com#national-fingerprint-database [Accessed 26 August 2025].

Home Office, 2013. *Serious and organised crime strategy.* [Online] Available at: https://assets.publishing.service.gov.uk/media/5a7c448b40f0b62dffde0f40/Serious_and_Organised_Crime_Strategy.pdf [Accessed 12 September 2021].

Home Office, 2018. *Serious and organised crime strategy 2018.* [Online] Available at: https://assets.publishing.service.gov.uk/media/5bd99ee8e5274a6e39bf2c2e/SOC-2018-web.pdf [Accessed 12 September 2021].

Home Office, 2023. *No Place to hide: Serious and organised crime strategy 2023 to 2028.* [Online] Available at: https://assets.publishing.service.gov.uk/media/65798633254aaa0010050bdc/SOC_Strategy_23-28_V9_Web_Accessible.pdf [Accessed 15 May 2024].

Hopkins, B. and Evelson, B. (2011) *Expand your digital horizon with big data* [online] Available at: https://1library.net/document/qmkmlw5z-expand-digital-horizon-data-brian-hopkins-boris-evelson.html [Accessed 24 June 2021].

Horita, F.E., Albuquerque, J.P.d., Marchezini, V. and Mendiondo, E.M. (2017) Bridging the gap between decision-making and emerging big data sources: an application of a model-based framework to disaster management in Brazil. *Decision Support Systems*, v. 97, pp.12-22.

IBM (2015) *What is big data* [online] Available at: www.01.ibm.com/ software/data/bigdata/ what-is-big-data.html [Accessed 26 March 2020].

James, A. (2016) Understanding police intelligence work. *Policing: A Journal of Policy and Practice*, v. 14(2), pp.387-388.

James, N. (2007) The use of email interviewing as a qualitative method of inquiry in educational research. *British Educational Research Journal*, v. 33(6), pp.963-976.

Janssen, M. and Kuk, G. (2016) The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, v. 33(3), pp.371-377.

Jha, B., Sivasankari, G.G. and Krishnappa, V. (2020) Fraud detection and prevention by using big data analytics. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, India. IEEE Xplore.

Jin, X., Wah, B.W., Cheng, X. and Wang, Y. (2015) Significance and challenges of big data research. *Big Data Research*, v. 2(2), pp.59-64.

Johnson, J.L., Adkins, D. and Chauvin, S. (2020) Qualitative research in pharmacy education: A review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education*, v. 84(1), pp.138-146.

Jurkiewicz, C.L. (2018) Big data, big concerns: ethics in the digital age. *Public Integrity*, v. 20(1), pp.46-59.

Kiger, M.E. and Varpio, L. (2020) Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, v. 42(8), pp.846-854.

Kitchin, R. and McArdle, G. (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, v. 3(1), pp.1-10.

Kothari, C.R. (2004) *Research methodology: methods and techniques.* 2nd ed. New Delhi: New Age International .

Lawani, A. (2021) Critical realism: what you should know and how to apply it. *Qualitative Research Journal*, v. 21(3), pp.320-333.

Lexico (2021) *Crime detection* [online] Available at: https://www.lexico.com/definition/crime_detection [Accessed 22 April 2020].

Lindsay, S. (2019) Five approaches to qualitative comparison groups in health research: a scoping review. *Qualitative Health Research*, v. 29(3), pp.455-468.

Loenen, B.v., Kulk, S. and Ploeger, H. (2016) Data protection legislation: a very hungry caterpillar. The case of mapping data in the European Union. *Government Information Quarterly*, v. 33(2), pp.338-345.

Lyon, D. (2014) Surveillance, Snowden, and big data: capacities, consequences, critique. *Big Data & Society*, v. 1(2), pp.1-13.

Mauro, A. D., Greco, M. and Grimaldi, M. (2016) A formal definition of big data based on its essential features. *Library Review*, v. 65(3), pp.122-135.

McLeod, S. (2024) *Thematic analysis: a step by step guide.* Simply Psychology [online] Available at: https://www.simplypsychology.org/wp-content/uploads/simplypsychology.org-Thematic-Analysis-A-Step-by-Step-Guide.pdf.

Metropolitan Police (2023) *Crime type definitions* [online] Available at: https://www.met.police.uk/police-forces/metropolitan-police/areas/stats-and-data/stats-and-data/met/crime-type-definitions/ [Accessed 23 December 2023].

Metropolitan Police (2025) *Automatic number plate recognition (ANPR)* [online] Available at: https://www.met.police.uk/advice/advice-and-information/rs/road-safety/automatic-number-plate-recognition-anpr/ [Accessed 15 August 2025].

Moen, K. and Middelthon, A.-L. (2015) Qualitative research methods. In: Laake, P., Benestad, H.B. and Olsen, B.R. (eds.) *Research in medical and biological sciences: from planning and preparation to grant application and publication.* Amsterdam: Elsevier Ltd, pp.321-378.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ,* v.339, pp.1-8.

Munoz, C., Smith, M. and Patil, D. (2016) *Big data: a report on algorithmic systems, opportunity, and civil rights.* Washington DC: Executive Office of the President.

NAO (2019) *Tackling serious and organised crime* [online] Available at: https://www.nao.org.uk/wp-content/uploads/2019/03/Tackling-serious-and-organised-crime.pdf [Accessed 7 May 2024].

NASA (2025) *National Aeronautics and Space Administration Homepage* [online] Available at: https://www.nasa.gov [Accessed 26 February 2025].

Nath, S.V. (2006) Crime pattern detection using data mining. *Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006*, Hong Kong, China. IEEE Xplore, pp.1-4.

NCA (2021) *National Crime Agency Annual Plan 2020-2021* [online] Available at: https://www.nationalcrimeagency.gov.uk/who-we-are/publications/439-national-crime-agency-annual-plan-2020-2021-1/file#:~:text=The%20latest%20estimate%20of%20the,likely%20to%20be%20an%20underestimate.&text=There%20are%204%2C772%20known%20Organised,involved [Accessed 14 November 2022].

Neiva, L., Granja, R. and Machado, H. (2022) Big data applied to criminal investigations: expectations of professionals of police cooperation in the European Union. *Policing and Society*, v. 32(10), pp.1167–1179.

Neiva, L., Machado, H. and Silva, S. (2023) The views about big data among professionals of police forces: a scoping review of empirical studies. *International Journal of Police Science & Management*, v. 25(2), pp.208–220.

New York Police Department (2023) *Seven major felony offenses by precinct 2000–2022* [online] Available at: https://www.nyc.gov/assets/nypd/downloads/pdf/analysis_and_planning/historical-crime-data/seven-major-felony-offenses-by-precinct-2000-2022.pdf [Accessed 8 December 2023].

Newburn, T. (2008) *Handbook of policing.* 2nd ed. Devon: Willan Publishing.

Nnaji, U.O., Benjamin, L.B., Eyo-Udo, N.L. and Etukudoh, E.A. (2024) A review of strategic decision-making in marketing through big data and analytics. *Magna Scientia Advanced Research and Reviews*, v. 11(01), pp.84-91.

Nowell, L.S., Norris, J.M., White, D.E. and Moules, N.J. (2017) Thematic analysis: striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, v. 16(1), pp.1-13.

O'Connor, C.D., Ng, J., Hill, D. and Frederick, T. (2022) Thinking about police data: analysts' perceptions of data quality in Canadian policing. *The Police Journal: Theory, Practice and Principles*, v. 95(4), pp.637-656.

ONS (2025) *Census 2021. Crime in England and Wales: year ending March 2025*. Office for National Statistics [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmarch2025 [Accessed 23 May 2025].

Opdenakker, R. (2006) Advantages and disadvantages of four interview techniques in qualitative research. *Qualitative Social Research*, v. 7(4), pp.1-13.

Oxford English Dictionary (2025) Variability. *Oxford English Dictionary* [Online] Available at: https://www.oed.com/dictionary/variability_n?tab=meaning_and_use#15867819 [Accessed 18 September 2025].

Paoli, L., Adriaenssen, A., Greenfield, V.A. and Conickx, M. (2017) Exploring definitions of serious crime in EU policy documents and academic publications: a content analysis and policy implications. *European Journal on Criminal Policy and Research*, v. 3(23), pp.270-285.

Patton, M.Q. (2002a) *Qualitative research & evaluation methods.* 3rd ed. London: Sage Publications.

Patton, M.Q. (2002b) *Evaluation checklists* [online] Available at: https://wmich.edu/sites/default/files/attachments/u350/2018/qual-eval-patton.pdf [Accessed 21 April 2021].

Perry, J.S. (2017) *What is big data? More than volume, velocity and variety* [Online] Available at: https://developer.ibm.com/blogs/what-is-big-data-more-than-volume-velocity-and-variety/?mhsrc=ibmsearch_a&mhq=big%20data [Accessed 12 March 2020].

Prabakaran, S. and Mitra, S. (2018) Survey of analysis of crime detection techniques using data mining and machine learning. *Journal of Physics: Conference Series*, v. 1000(1), pp.1-10.

Pramanik, M.I. Lau, R.Y.K., Yue, W.T., Ye, Y. and Li, C. (2017) Big data analytics for security and criminal investigations. *WIREs: Data Mining and Knowledge Discovery*, v. 7, pp.1-19.

Ravikumar, T., Sriram, M. and Murugan, N. (2022) Applications and risks of big data in financial services. In: *10th International Conference on Emerging Trends in Corporate Finance and Financial Markets*, October 13-14, 2022, Bangalore, India, Mysuru: SDM Institute for Management Development (SDMIMD), pp.1-7.

Rai, A., 2020. *What is big data - characteristics, types, benefits & examples* [Online] Available at: https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/ [Accessed 6 May 2020].

Richards, N.M. and King, J.H. (2014) Big data ethics. *HeinOnline*, v. 49(2), pp.393-432.

Roopa, S. and Rani, M. (2012) Questionnaire designing for a survey. *The Journal of Indian Orthodontic Society*, v. 46(4), pp.273-277.

Rosenthal, M. (2016) Qualitative research methods: why, when, and how to conduct interviews and focus groups in pharmacy research. *Currents in Pharmacy Teaching and Learning*, v. 8(4), pp.509-516.

Şahin, M.D. and Öztürk, G. (2019) Mixed method research: theoretical foundations, designs and its use in educational research. *International Journal of Contemporary Educational Research*, v. 6(2), pp.301-310.

Sandhu, A. and Fussey, P. (2021) The “uberization of policing”? How police negotiate and operationalise predictive policing technology. *Policing and Society*, v. 31(1), pp.66-81.

Sandy , Q.Q. and Dumay, J. (2011) The qualitative research interview. *Qualitative Research in Accounting & Management*, v. 8(3), pp.238-264.

Sayer, A. (2000) *Realism and social science.* London: Sage Publications Ltd..

Schuilenburg, M. and Soudijn, M. (2023) Big data policing: the use of big data and algorithms by the Netherlands Police. *Policing: A Journal of Policy and Practice*, v. 17(1), pp.1-9.

Selbst, A.D. (2017) Disparate impact in big data policing. *Georgia Law Review*, v. 52(1), pp.109-196.

Sharma, M. (2014) Z - CRIME: a data mining tool for the detection of suspicious criminal activities based on decision tree. *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, Delhi. IEEE, pp.1-6.

Sol, K. and Heng, K. (2022) Understanding epistemology and its key approaches in research. *Cambodian Journal of Educational Research*, v. 2(2), pp.80-99.

Southerton, C. (2020) Datafication. In: Schintler, L. and McNeely, C. (eds.) *Encyclopedia of big data* [online] Available at: https://link.springer.com/referenceworkentry/10.1007/978-3-319-32001-4_332-1 [Accessed 20 January 2022].

Srinuvasu, M.A. and Santhosh, E.B. (2017) Big data: challenges and solutions. *International Journal of Computer Sciences and Engineering*, v. 5(10), pp.250-255.

Surbakti, F.P.S. (2020) Understanding effective use of big data: challenges and capabilities (A management perspective). *Jurnal Metris*, v. 23, pp.1-14.

Tamim, D.K. (2024) *How does Dubai protect 200 nationalities?* [Interview] (23 July 2024).

Tayal, D.K., Jain, A., Arora, S., Agarwal, S., Gupta, T. and Tyagi, N. (2015) Crime detection and criminal identification in India using data mining techniques. *AI & Society*, v. 30(1), pp.117-127.

TechAmerica (2012) *Demystifying big data. A practical guide to transforming the business government.* Washington, DC: TechAmerica Foundation.

Tembrioti, L. and Trangaridou, N. (2013) Reflective practice in dance: a review of the literature. *Research in Dance Education*, v. 15 (1), pp.4-22.

Thabet, N. and Soomro, T.R. (2015) Big data challenges. *Journal of Computer Engineering & Information Technology*, v. 4(3), pp.1-10.

Tomar, N. and Manjhvar, A.K. (2016) An improved optimized clustering technique for crime detection. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, Indore, India, IEEE, pp.1-5.

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., Levac, D., Ng, C., Pearson Sharpe, P., Wilson, K., Kenny, M., Warren, R., Wilson, C., Stelfox, H.T. and Straus,

S.E. (2016) A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology*, v. 16(15), pp.1-10.

Udeh, E.O., Amajuoyi, P., Adeusi, K.B. and Scott, A.O. (2024) The role of big data in detecting and preventing financial fraud in digital transactions. *World Journal of Advanced Research and Reviews*, v. 22(2), pp.1746-1760.

UNODC (2012) *The notion of serious crime in the United Nations Convention against Transnational Organized CrimeNot.* United Nations Office on Drugs and Crime. Vienna, United Nations, pp.1-8.

Uprichard, E. (2013) *Focus: big data, little questions?* [online] Available at: https://discoversociety.org/2013/10/01/focus-big-data-little-questions/ [Accessed 29 April 2020].

van der Voort, H.G., Klievink, A.J., Arnaboldi, M. and Meijer, A.J. (2019) Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, v. 36(1), pp.27-38.

Vestby, A. (2019) Machine learning and the police: asking the right questions. *Policing Journal Oxford Publishing*, v. 15(1), pp.44-58.

Viitanen, J. et al., 2022. Patient experience from an eHealth perspective: A scoping review of approaches and recent trends. *Yearbook of Medical Informatics,* v. 31(1), pp.136-145.

Vomfell, L., Härdle, W.K. and Lessmann, S. (2018) Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems*, v. 113, pp.73-85.

Wadhwani, K. and Wang, Y. (2017) *Big data challenges and solutions* [online] Available at: *https://www.researchgate.net/publication/313819009_Big_Data_Challenges_and_Solutions*.

Wilson, C. (ed.) (2014) Semi-structured interviews. In: *Interview Techniques for UX Practitioners.* Amsterdam: Morgan Kaufmann, pp.23-41.

Winchester, N. (2020) *Policing in the UK: serious and organised crime* [online] Available at: https://researchbriefings.files.parliament.uk/documents/LLN-2020-0016/LLN-2020-0016.pdf [Accessed 1 April 2024].

Xu, Z., Cheng, C. and Sugumaran, V. (2020) Big data analytics of crime prevention and control based on image processing upon cloud computing. *Journal of Surveillance, Security and Safety*, v. 1, pp.16-33.

Yadav, S., Timbadia, M., Yadav, A. and Vishwakarma, R. (2017) Crime pattern detection, analysis & prediction. *International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, pp.225-230.

Ylijoki, O. and Porras, J. (2016) Perspectives to definition of big data: a mapping study and discussion. *Journal of Innovation Management*, v. 4(1), pp.69-91.

Zainab, K. and Dhanda, D.N. (2018) Big data and predictive analytics in various sectors. *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India. IEEE.

Zikopoulos, P. and Eaton, C. (2012) *Understanding big data, analytics for Enterprise Class Hadoop and streaming data.* 1st ed. New York, NY: McGraw-Hill.

Appendix A. Scoping Review Studies

Table A.1. Relevant studies identified in the scoping review.

Author(s) (Year)	Country/context	Aim and focus of the study	Methods/data sources	Key points
Abdul Jalil, Mohd and Noor (2017)	Malaysia (crime data analytics)	To compare different filtering selection methods for crime data to identify effective techniques for selecting relevant features in crime datasets.	Quantitative data mining technical study that applies several filtering methods to crime data and evaluates their performance.	- Insights on crime prevention strategies. - The effectiveness of WEKA data mining software in analysing crime data.
Abouelmehdi, Beni-Hessane and Khaloufi (2018)	General (healthcare and data context)	To examine how security and privacy can be protected when working with high volumes of healthcare data, review techniques, and challenges.	Conceptual review article that draws on existing literature to discuss security and privacy risks and possible technical solutions for big healthcare data.	- Highlights the potential value of big data in healthcare. - Describes security and privacy risks related to sensitive health data. - Recommends technical approaches such as encryption, anonymisation, and access control. - Highlights that technical measures need to be combined with policy and governance practices.
Aderemi et al. (2024)	Financial services industry with examples focused on Nigeria	To review how big data analytics is used in the financial services industry and identify key challenges and future prospects.	Narrative review of academic and industry sources, synthesising prior work rather than collecting new empirical data.	- Presents current applications and the potential of big data analytics in the financial field, such as risk management fraud detection. - Identifies technical challenges, such as data integration, infrastructure and quality. - Highlights ethical and privacy concerns, along with skills and organisational capabilities that may limit effective adoption of big data analytics.

Almansoori (2019)	United Arab Emirates (major crime investigations)	To examine major crimes by interviewing police officers using the Investigative Interviewing Triangle.	Doctoral thesis with qualitative and quantitative data (interviews and surveys) collected from police officers, victims, and convicted offenders in Dubai.	- Introduction to police tasks in investigating crimes. - Highlights the importance of addressing serious crimes.
Association of Police and Crime Commissioners (APCC, 2020)	United Kingdom (national policing)	To set out the National Digital Policing Strategy 2020–2023, which defines and discusses digital capabilities, data and technology priorities for UK policing.	Strategic policy document produced by the APCC and NPCC with input from policing and technology stakeholders, synthesises assessments of current digital capabilities and future needs.	- Highlights the importance of addressing serious and organised crime and using new advanced technologies in policing. - Acknowledges and supports the potential benefits of analysing high volumes of data to support decision making. - Highlights the rapid growth of data and practical complexities of policing capabilities.
Assouli, Benahmed and Gasbaoui (2021)	Global (crime prediction and informatics)	To propose and evaluate an informatics link prediction approach for crime prediction.	Quantitative technical modelling study that represents crime data as a network structure and applies link prediction algorithms.	- Highlights the role of link prediction as a tool that can be used to link criminal groups. - Suggests that link prediction can be used by policing to prevent serious crimes.
Awrahman, Fatah and Hamaamin (2022)	Iraq (healthcare informatics)	To review the role of big data in healthcare informatics and the challenges associated with it.	Review article that summarises recent literature on big data in healthcare, its applications, and challenges.	- Presents how big data analytics can support better patient outcomes, cost reduction and improved healthcare management. - Discusses technological enablers to handle big data in healthcare. - Identifies key challenges, such as data quality, security and privacy. - Concludes that to realise the potential of big data in healthcare, it is necessary to address the challenges.

Babuta (2017)	United Kingdom (policing and law enforcement)	To assess the requirements, expectations and priorities of UK law enforcement agencies regarding big data and explore the potential advantages and challenges.	Qualitative research with UK police with practice-based analysis of police data systems and technologies, and interviews with officers and staff.	- Finds that the UK police hold vast quantities of data across fragmented, incompatible systems. - Identifies organisational, technical and cultural barriers to effective big data analysis. - Discusses the emerging potential of using big data in policing, such as predictive hotspot mapping and open-source analytics. - Highlights legal, ethical and governance challenges around privacy and accountability.
Balogun et al. (2025)	United States (healthcare cybersecurity)	To examine and propose ways to enhance incident response strategies for cybersecurity threats.	Quantitative study that uses case examples, frameworks, and scenario-based analysis of incident response in healthcare cybersecurity, with conceptual modelling, case study, and review of practices.	- Highlights the risks of data breaches of sensitive records. - Emphasizes on the Ransomware as a in increasing threat. - Reflects on privacy concerns that are related to big data.
Barocas et al. (2017)	United States (civil rights and data science)	To examine how big data and data science interact with civil rights in the United States.	Policy-oriented report that synthesises legal, technical, and social science literature.	- Highlights the role of pattern recognition algorithms to extract patterns to advance diverse fields. - Suggests that computer scientists have started to explore and investigate ways to reduce or eliminate bias. - Emphasises on the importance of focusing on implicit discrimination of data-driven methods that are deployed in several crucial social institutions.

Bell et al. (2021)	Global (big data in international humanitarian work)	To explore how big data is currently used in the humanitarian domain and identify future challenges of deploying big data analytics.	Mixed empirical and conceptual paper that draws on existing literature and uses examples from humanitarian practices through data integration and conceptual analysis.	- Clarifies the concept of big data. - Expands on the characteristics that represent big data and suggests 8 Vs. - Highlights the importance of all characteristics but argues that value is central.
Bignami (2007)	European Union (privacy and data retention in law enforcement)	To analyse the EU Data Retention Directive and its implications for privacy and data protection.	Doctrinal legal and policy analysis that examines the background of Data Retention Directive and interprets it in light of EU fundamental rights.	- Highlights that data privacy is one of the oldest human rights legally. - Data protections laws faced challenges to prevent rights abuse with developments in the criminal justice system and powers to protect national security.
Brady (2019)	US (political science and social science)	To examine the opportunities and challenges that big data and data science pose for political science and social research.	Conceptual and review article that synthesises existing studies from political science and related fields.	- Argues that big data can offer new forms of data analytical tools that may enhance social research. - Highlights ethical concerns, including privacy and bias. - Identifies challenges related to validity and generalisability when using big data sources.
Brennan (2022)	England and Wales (serious violence victimisation)	To analyse patterns of serious violence victimisations in England and Wales between 2011 and 2017.	Quantitative analysis of existing data on serious violence in England and Wales.	- Highlights the importance of addressing serious crimes by the government and police forces nationally.
Broeders et al. (2017)	Netherlands (security policy and regulation of big data analytics)	To develop a framework to regulate the use of big data in security policy.	Conceptual and policy-oriented analysis that draws on legal, ethical and governance literature, and examples of big data use in security context.	- Discusses the concept of big data and its characteristics. - Highlights the rapid increase in volumes of data than can be used in criminal investigations. - Argues that big data analytics can lead to positive unexpected connections related to the efficient use of police resources.

				- Suggests that pattern recognition is central to big data applications. - Despite the potential, emphasises the need for safeguarding measures when big data is used to support law enforcement decisions.
Cardenas, Manadhata and Rajan (2013)	Global (security analysis)	To discuss how big data analytics is transforming security tools and practices.	Conceptual practice-focused article that synthesises trends in security analytics and uses examples from large-scale log analysis and security operations.	- Highlights the attention that big data analytics gained in the security community and its potential in analysing data at an unprecedented scale. - Argues that one of the impacts of big data technologies is it is providing affordable infrastructures and suggests several tools. - Despite the recognised potential, storage, reliability, and privacy challenges are needed to be addressed to achieve true potential
Cardock, Stalla-Bourdillon and Millard (2017)	European Union (data protection law and information obligations)	To demonstrate that existing ways of categorising personal data are inadequate to meet legal obligation to inform under EU data protection law.	Conceptual doctrinal legal analysis that analyses EU data protection legislation and engagement with legal and technical literature with no primary empirical data.	- Highlights that transparency is central in data protection laws and the obligation to inform is key to ensure transparency. - Suggests that transparency aims to give data subjects enough information to evaluate the trustworthiness of the data controller.

Chan and Moses (2016)	Australia (security and policing)	To examine how big data technologies may transform the production of security and implications of big data.	Qualitative empirical with conceptual framing that draws on empirical data from an Australian study of security agencies (interviews/qualitative data).	- Argues that big data offers significant promise in enhancing security but is shaped by existing institutional logics and practices. - Identifies tensions between the expectations of big data and practical constraints. - Highlights risks around privacy and the expansion of data collection for security purposes.
Clissa, Lassing and Rinaldi (2023)	Global (science and data production)	To provide a comprehensive survey of how much data is produced, stored, and stream across scientific and industrial domains.	Quantitative analytic survey of existing datasets and reports that complies and analyses statistics on data production, storage, and capabilities.	- Contributes to understanding and shaping the concept of big data and its characteristics as foundations. - Argues that big data as a term does not necessarily refer to a specific size in terms of volume but implies to volumes that exceed normal datasets.
Costanzo, D'Onofrio and Friedl (2015)	Italy (policing and big data legal framework)	To examine how big data is regulated under the Italian legal framework.	Legal and policy analysis chapter in an edited volume that reviews Italian and EU legal provisions in relation to big data and policing.	- Reflects on the high volumes of data that governments have. - Highlights that data protections laws might not be fully applied on intelligence and police forces as they act under government officials and parliament committees. - Emphasises that the technology developments caused a significant change between policing and data protection legislation.
Cox and Ellswort (1997)	Global (high-performance computing)	To present and evaluate an application that allows interactive visualisation of large datasets.	Quantitative technical experimental computer science paper that proposes a framework for demand paging in visualisation.	- Illustrates the foundation and how the term 'big data' emerged as a technical challenge during NASA projects.

De Hert and Papakonstantinou (2009)	European Union (police and judicial cooperation in criminal matters)	To assess the 2008 Data protection Framework Decision for police and judicial cooperation and evaluate if it improves the protection of personal data.	Doctrinal legal and policy analysis that examines framework decision, compares it with existing EU and national data protection guides, relevant legal and policy literature, no primary empirical fieldwork.	- Introduces the emergence of data protection legislation in relation to the computer evolution as the reason for privacy concerns. - Highlights that privacy concerns shifted from the government to the private sector in the 1980s. - Emphasises on the importance of protecting personal data from both governments and private organisations in activities such as data mining and profiling.
Dijcks (2013)	Global (enterprise IT and big data infrastructure)	To describe Oracle's approach of big data for the enterprise by outlining how organisations can acquire, store, and analyse large datasets.	Conceptual technical report that uses descriptive explanations of Oracle's big data architecture and uses illustrative technical scenarios rather than empirical evaluation.	- Explains big data characteristics and demonstrates the increasing volumes of data worldwide. - Highlights that there is always valuable information in large datasets, but it might not always be visible. - Emphasises on the importance of developing proper technical infrastructure to be able to extract the useful information from big data.
Dubai Police (2023)	United Arab Emirates (official crime statistics)	To provide official statistical data on serious crime and their categories in Dubai.	Official government statistics of police crime data.	- Identifies types and categories of serious crime.
Dumbill (2013)	Global (big data, technology, and enterprise)	To introduce the concept of big data and explain why it matters and how organisations can start to make sense of it.	Conceptual article that draws on industry experience and big data practices with no formal empirical data collection.	- Introduces the concept of big data and expands beyond the three Vs and adds value. - Highlights the importance of value as an element in the concept of big data.
Egber and Krasmann (2019)	Europe (crime prediction software in German-speaking countries)	To analyse how predictive policing technologies are used in German-speaking countries.	Qualitative socio-legal analysis that uses case examples of crime prediction software and explores	- Suggests that the evolution of big data supports predictive policing and refers to the usefulness of the PRECOBS software.

			practices in predictive policing.	- Argues that despite the advantages of predictive policing, its success will likely depend on how the police and society deal with technology. - Highlights the importance of accountability measures in relation to software that might produce false outcomes.
European Commission (2020)	European Union (security policy and strategy)	To set out the EU Security Union Strategy 2020-2025, identify the strategic priorities, and actions to strengthen security including tackling serious and organised crime.	Policy document with internal EU assessments and policy documents underpinning the strategy.	- Emphasises on the role of big data and AI as effective solutions to fight crimes. - Suggests different advantages of their use such as identifying patterns, online criminals, and suspicious transactions. - Highlights the important role of electronic information and digital evidence as an important part of serious crime investigations. - Emphasises the importance of the highest levels of compliance towards fundamental rights.
Europol (2023)	European Union (serious and organised crime threat assessment)	To define serious and organised crime as part of Europol's framework.	Policy document with outputs from Europol's SOCTA data collection and analysis.	- Identifies types and categories of serious crime. - Reflects on the cross-border nature of serious crimes.

Ezzeddine, Bayerl and Gibson (2023)	UK, Netherlands and Germany (perspectives of citizens on AI use by police)	To explore how citizens understand and evaluate police use of AI in the face of struggles between safety and privacy.	Q methodology with 43 participants in the UK, Netherlands, and Germany.	- Suggests that complexities in modern crimes motivated policing to explore advanced technologies. - Argues that big data and AI are used by policing to increase operational efficiency and manage police resources. - Suggests that the public view is more complex than a simple privacy vs security argument. - Highlights concerns regarding surveillance, discrimination and lack of safeguards.
Feng et al. (2019)	Global (technical crime data analytics)	To develop and demonstrate a big data analytics and mining framework to visualise crime data and forecast crime trends.	Quantitative study that uses crime datasets with data mining techniques within big data environment.	- Clarifies the concept of big data analytics. - Highlights its promising outcomes for policing to better understand crime, track criminal activity, predict incidents, and effective deployment of resources. - Argues that data mining is one of the primary techniques in big data analytics. - Highlighted technical challenges related to handling big data's volume.
Ferguson (2015)	United States (policing and big data)	To analyse how big data and predictive policing technologies interact with the US Fourth Amendment standards.	Doctrinal legal analysis article that uses US constitutional case law and discusses predictive policing systems and data practices with no primary data collection.	- Highlights the worldwide expansion in scale of data. - Argues that big data remains broadly under-regulated. - Suggests that with enough data, police will be able to predict crimes and identify patters.

Gkikas and Theodoridis (2022)	Global (consumer behaviour and marketing)	To examine how AI is used to understand and influence consumer behaviour.	Literature-based book chapter that reviews literature on AI, ML, and data mining, and examples of their applications.	- Provides updated figures on the estimated worldwide size of data. - Suggests a recent definition of artificial intelligence.
Government Digital Service (2018)	United Kingdom (public sector data and AI projects)	To provide a Data Ethics Framework that guides appropriate and responsible use of data in the UK.	Framework document with policy and ethics work within GDS and across government as a principle-based guide.	- Highlights the emergence of data ethics into practices where data is generated, analysed, and disseminated. - Clarifies what data ethics laws and frameworks encompass to be used as a practical guide for the government and private sectors.
Hajian, Domingo and Martinez- Balleste (2011)	Spain (data mining and crime detection)	To propose methods for preventing discrimination in data mining models used for crime detection.	Quantitative technical paper that uses synthetic or real datasets with attributes to illustrate discrimination awareness in data mining.	- Highlights the ability of data mining techniques to detect discriminatory decisions. - Acknowledges the risks of bias in crime detection models.
Hassani et al. (2016)	Global (crime data mining)	To provide a review of data mining applications in crime by summarising over 100 studies and identifying different techniques.	Review article of mapping literature on data mining applications in crime which is drawn from multiplate databases.	- Highlights the need for advanced technological methods to manage the increasing volumes of crime data. - The constant growth of data is creating challenges for policing. - Highlights privacy concerns in relation to big data analytics and data mining. - Argues that data mining has potential benefits in the criminal field.

Home Office (2013)	United Kingdom (serious and organised crime strategy)	To set out the UK government's 2013 Serious and Organised Crime Strategy and explain how the government and law enforcement will respond to serious and organised crime.	Policy and strategy document of internal government assessments and intelligence on SOC.	- Suggests that criminals are harnessing technology which kept them ahead of policing. - Demonstrates the financial harm of serious and organised crime. - Presented different steps that aim to tackle SOC which included the collection and analysis of 'bulk data' to develop investigative capabilities.
Home Office (2018)	United Kingdom (serious and organised crime strategy)	To update and replace the 2013 strategy and setting out how the UK will mobilise full force to tackle serious and organised crime.	Policy and strategy document of cross-government assessments of SOC threats and harms with inputs from law enforcement agencies.	- Emphasises on the danger that SOC poses to national security and the urgency to address them. - Highlights the developments of technology that criminal networks are using. - Clearly states its support to initiatives that will ethically use AI to exploit and interpret big data to disrupt SOC effectively.
Home Office (2023)	United Kingdom (serious and organised crime strategy)	To present the UK's 2023-2028 Serious and Organised Crime Strategy by outlining its approach to use intelligence, law enforcement powers, and partnership.	Policy and strategy document with updated threat assessments and evidence on SOC in and against the UK.	- Emphasises on the increasing harm caused by SOC with complexities in measuring it. - Clearly states the aim of providing advanced technologies for data collection and analysis to identify and disrupt SOC. - Outlines several projects which are perceived as big data policing projects such as the Prum and I-LEAP. - Provides updated figures of the financial harm of SOC to the UK which is estimated as £47 billion per year.

Home Office (2024)	United Kingdom (forensic databases)	To report on the operation and performance of the UK's forensic information databases for 2023-2024.	Official government annual report with administrative statistics from the National DNA Database and National Fingerprint Database.	- Provides recent figures on the expansion and increase of forensic records which reflects on policing technical infrastructure. - Suggests an examples of a perceived big data forensic database such as IDENT1.
Hopkins and Evelson (2011)	Global (enterprise IT)	To explain what big data means for organisations.	Conceptual strategy paper based on Forrester research, client work, and market analysis, and surveyed 60 clients on their plans towards big data technologies.	- Highlighted that variability is one of the distinguishing features of big data. - Elaborates how the field of big data evolved and its characteristics.
Horita et al. (2017)	Brazil (focusing on disaster management and civil defence context)	To develop and demonstrate a framework to integrate emerging big data sources for decision making.	Conceptual and model-based framework, entailing a case study in disaster management using big data sources combined with disaster management information.	- Highlights challenges related to data quality, integration, interpretation and institutional capacity. - Demonstrates the expansion of big data across different technologies.
IBM (2015)	Global (enterprise technology)	To explain IBM's definition of big data and introduces its Vs.	Corporate conceptual report encompassing IBM's internal expertise, marketing, and prior industry reports.	- Defines big data in terms of its Vs and their extension to veracity. - Presents its opportunities for insight and innovation while highlighted its technical challenges.

James (2016)	United Kingdom (policing and intelligence)	To review and critically comment on the book Understanding Police Intelligence Work and summarise its contributions.	Book review with no primary empirical data, draws on the content of the book to highlight key themes and contributions.	- Reflects on the increasing volumes of big data and suggests its term does not only imply to its size but its unrealised benefits. - Highlights the potentials that policing can gain from big data but also acknowledges technical challenges such as infrastructure and data quality. - Emphasises on the need for strict data quality control measures for policing.
Janssen and Kuk (2016)	Global (technology and governance)	To examine the challenges and limitations of big data algorithms when used in technocratic governance.	Conceptual analytical article with theoretical discussion of big data algorithms and governance, illustrates examples from public sector, with no primary empirical dataset.	- Provides an explanation and definition of an algorithm.
Jha, Sivasankari and Krishnappa (2020)	India (fraud detection and big data)	To present big data analytics approach for fraud detection and prevention.	Quantitative paper that uses transactional or fraud related data with data mining and machine learning methods.	- Highlighted the potential role that data mining can play in policing. - Presents how big data analytics can support automatic detection of fraud patterns through large datasets.
Jin et al. (2015)	General (cross-sector perspectives on big data research)	To review the importance, opportunities and challenges of big data research.	Conceptual overview paper that synthesises existing discussion and presents opportunities and challenges in big data research, no empirical data collection.	- Explains the significance of big data and its potential to generate new insights and drive innovation by providing examples of academic and government projects. - Identifies technical challenges related to storage, analysis, privacy and data security. - Highlights the importance of data quality, value and veracity.
Jurkiewicz (2018)	United States (context of public	To analyse the ethical challenges posed by big data for public organisations and	Conceptual paper that draws on ethical theory and public administration literature and	- Discusses the evolution of the big data concept and its characteristics.

	administration and ethics)	how ethical frameworks can respond.	uses examples of digital use in public and private sectors, with no primary empirical data.	- Argues that big data raises ethical, privacy and transparency concerns that some organisations do not address. - Highlights the risks of profiling, surveillance and loss of control over personal information. - Emphasises the need for updated ethical guidelines for the digital era.
Kitchin and McArdle (2016)	Global (big data characteristics across multiple domains)	To explore what actually make 'big data' big and its characteristic.	Mixed empirical conceptual analysis study that uses 26 datasets from different domains with a systematic comparison.	- Clarifies the concept of big data and its characteristics. - Highlights value as one of the important characteristics of big data. - Expands beyond the traditional Vs and suggests additional 6 characteristics.
Loenen, Kulk and Ploeger (2016)	European Union (data protection and mapping)	To examine how EU data protection legislation affects the use and release of geospatial data.	Doctrinal legal policy analysis that examines EU data protection legislation, policy documents, with no primary empirical data.	- Enhanced computing powers and developments in data mining techniques requires an expansion of the scope of the EU Data Protection Framework.
Mauro, Greco and Grimaldi (2016)	Italy (information systems)	To develop a formal and general definition of big data by identifying and synthesising its essential features from existing literature.	Literature-based conceptual analysis that reviews and compares existing big data definitions.	- Demonstrates that existing big data definitions vary but share a set of features. - Distinguishes between essential features of big data. - Argues that a clear formal definition can support more precise academic discussion and practical application.

Metropolitan Police (2023)	United Kingdom (crime recording and categories)	To provide definitions of crime types used in Metropolitan Police, including distinctions between major and minor crime types.	Official government operational guide with internal Metropolitan Police practices.	- Identifies types and categories of serious crime.
Metropolitan Police (2025)	United Kingdom (ANPR use in policing)	To explain what ANPR is and how it is used to detect and disrupt crime.	Official public information guide with ANPR policy and operational practice.	- Provides a recent figure on number of ANPR records which reflects on the expanding volume of police data that may create technical challenges.
Munoz, Smith and Patil (2016)	United States (civil rights, data, and algorithms)	To assess how big data and algorithmic systems affect civil rights.	Policy analytical report that reviews existing research and case examples, consultation with experts and stakeholders.	- Argues that if new technologies were designed and implemented carefully, they can assist policing in decision making with lower risk than human bias. - Emphasises that algorithm development should not depend on variables that might indicate to a particular community.
Nath (2006)	Hong Kong (crime patten detection using data mining)	To explore how data-mining techniques can be used to detect crime patterns and demonstrate their potential in crime analysis.	Conference paper (IEEE) that applies quantitative data-mining methods in the analysis of crime data and uses examples of police datasets to show pattern discovery.	- Suggests that data mining can identify hidden or non-obvious patterns in crime data, thus supporting crime analysis and decision making. - Highlights the potential advantages of data mining for resource allocation and identifying crime hotspots. - Emphasises that data quality and appropriate method selection are important to ensure reliable results and outcomes.

National Audit Office (2019)	United Kingdom (government response to serious and organised crimes)	To examine whether the Home Office and the National Crime Agency tackle serious and organised crime in an effective way.	Public policy report that reviews strategy and planning documents from Home Office and NCA, and an analysis of governance information.	- Highlights the role of 100 government and law enforcement agencies that are tasked to tackle SOC. - Suggests that 2013 SOC strategy did not effectively address the complexity and scale of SOC.
Neiva, Granja and Machado (2022)	EU (police cooperation and criminal investigations)	To explore the expectations of professionals involved in police cooperation in the EU regarding the use of big data in criminal investigations.	Empirical qualitative study, interviews professionals and analyses their expectations, perceived opportunities and concerns about big data technologies.	- Finds that big data is seen as a promising tool to support policing and criminal investigations. - Identifies opportunities to improve efficiency and information sharing and enhance analytical capabilities. - Highlights concerns about data quality, ethics, privacy and organisational capability. - Concludes that realising the potential of big data to improve criminal investigations in the EU requires that both technical and governance challenges be addressed.
Neiva, Machado and Silva (2023)	International scope (policing and big data)	To conduct a scoping review of empirical research on the views of policing professionals regarding big data.	Scoping review that includes empirical studies on policing professionals' views regarding big data in policing, uses a descriptive-analytical approach to synthesise findings from 14 articles.	- Identifies both optimistic and oppositional views among police professionals concerning big data. - Optimistic views highlighted potential benefits, such as better crime prediction, improved efficiency and enhanced investigative capacity. - Oppositional views emphasised concerns about technological capabilities, data quality, organisational readiness and ethics. - Suggests that policing professionals recognise the benefits but acknowledge the risks.

New York Police (2023)	United States (police recorded crime statistics)	To provide historical crimes statistics for seven major felony offences.	Official statistical report of police crime data held by NYPD.	- Identifies types and categories of serious crime.
Newburn (2008)	United Kingdom (comprehensive reference on policing)	To provide an overview of policing theory, research and practice.	Handbook that synthesises existing research and debates rather than presenting one empirical study.	- Highlights the role of data analysis in policing practices and decision making. - Emphasises the need for high data quality to achieve reliable results.
Nnaji et al. (2024)	Marketing and business decision-making in the Nigerian/African context	To review how big data analytics are used to support strategic decision making in marketing and to identify key benefits and challenges.	Narrative review paper that summarises and discusses existing literature on big data analytics in marketing strategy.	- Suggests that despite the advantages of big data analytics, there are implementation challenges. - Identifies challenges such as data quality, ethics, bias, organisational reform and privacy concerns.
O'Connor et al. (2022)	Canada (crime incident data systems)	To explore how police data analysts in Canada evaluate the quality of police data.	Qualitative empirical study that interviews Canadian police data analysts, with thematic analysis of their accounts.	- Highlights the importance of reflecting on police data as little is known about its quality and policing is moving to an era of big data. - Data quality is one of challenges that police analysts face and this can affect police operations and resource allocations. - Argues that the quality of police data can be improved by establishing quality checks and auditing, providing additional training for front-line officers and ensuring support from police leaders.

Paoli et al. (2017)	European Union (EU criminal policy and literature on serious crime)	To examine how "serious crime" is defined in EU policy documents and academic publications.	Mixed (qualitative/quantitative) study with a collection of EU policy documents, sample pf academic publications, and systematic coding.	- Highlights the evolvement, differences in defining serious crimes, and their classifications.
Prabakaran and Mitra (2018)	Global (crime detection methods)	To provide a survey of crime detection techniques that rely on data mining and machine learning.	Literature review with conference proceedings paper that uses published research on crime detection using data mining and ML, with narrative comparison approach.	- Argues that data mining can be used in investigating and detecting serious crimes and suggest different data mining techniques.
Pramanik et al. (2017)	Global (security and criminal investigations)	To review and conceptualise how big data analytics can be applied to security and criminal investigations.	Conceptual review article with research literature on big data analytics in security, policing, and criminal investigations.	- Argues that law enforcement agencies are already using data mining techniques to prevent and detect serious crimes. - Highlights potential advantages for policing from big data analytics.
Ravikumar, Murugan and Sriram (2022)	India (financial services context)	To examine how big data is applied in the financial sector and identify the main risks of its applications.	Conceptual thematic study based on secondary data that synthesises existing literature and reports on big data in the financial sector.	- Describes key applications of big data, such as risk analysis, fraud detection and regulatory compliance. - Suggests that big data analytics have become essential for data-driven decision making in financial institutions. - Highlights challenges of shortages in skilled human resources and risks, such as poor data quality, which can lead to misleading analytics and harm business decisions.
Richards and Kings (2014)	United States (legal and ethical discussion of data practices)	To explore the ethical issues raised by big data practices.	Conceptual and legal analysis drawing on examples of big data use, with no primary empirical data.	- Demonstrates the big data revolution across multiple disciplines. - Discusses big data characteristics – volume, velocity and variety.

				- Calls for broader ethical and legal frameworks to address data collection, analysis and accountability.
Sandhu and Fussey (2021)	United Kingdom (predictive policing technology)	To explore how police officers engage and operationalise predictive policing technologies.	Qualitative empirical study interviewing UK police officers involved in designing and trialling predictive software in policing.	- Argues that the big data revolution influenced policing to explore advanced crime detection and prevention tools. - Suggests predictive policing analyses big data, which can lead to advantages such as improved resource deployment, high speed crime analysis and improved decision making. - Highlights that not all predictive policing technologies are complex. - Emphasises the quality of data used for predictive policing.
Schuilenburg and Soudijn (2023)	Netherlands (police use of big data and algorithms)	To examine how the Netherlands Police currently use big data and algorithms and map the areas of deployment.	Quantitative document analysis study that uses novel data from IT-related job vacancies from the Netherlands Police to identify areas of big data applications in police work.	- Acknowledges the importance of big data to tackle serious and organised crimes. - Identifies three main areas where big data is used: frontline policing, criminal investigations and intelligence. - Using big data for criminal investigations faces complexities, such as digital infrastructure implementation, which requires skilled staff. - Highlights international European policing collaboration to exchange and access big data, as Europol recognises its valuable role.

Selbst (2017)	United States (policing, discrimination law, and big data)	To examine how big data policing tools can produce disparate impact on protected groups and explore how US anti-discrimination laws applies.	Doctrinal law review article that utilises US discrimination law and example of big data policing systems, with no primary empirical data.	- Highlights incidents of possible biased outcomes from using data mining in predictive policing, and the need for new legislation. - Argues that police forces focus on crime control with no incentives to investigate possible biased results.
Sharma (2014)	India (technical crime analysis)	To present Z-CRIME which is a data mining tools based on decision tree algorithms to detect suspicious criminal activity.	Quantitative technical study that employs crime-related datasets and the implementation and testing of the Z-CRIME tool.	- Argues that with the expansion and increase in volumes of criminal data, data mining can be useful in crime detection and suggests Z-CRIME tool.
Surbakti (2020)	Indonesia (management and organisational context)	To explore what effective use of big data means from a management perspective and to identify organisational capabilities and challenges.	Qualitative multiple case study of eight organisations and draws on existing literature and managerial perspectives to discuss the use of big data.	- Highlights the proper use of big data as a competitive advantage. - Distinguishes between having big data technologies and using them effectively within organisations. - Identifies challenges such as skills gaps, organisational culture and governance.

Tayal et al. (2015)	India (crime detection and criminal identification)	To develop and test data mining techniques for crime detection and criminal identification.	Quantitative technical study that employs crime records from India and applications of data mining.	- Proposed a Crime Detection and Criminal Identification (CDCI) method that applies data mining techniques and can be beneficial to advance criminal investigations.
TechAmerica (2012)	United States (federal government and public sector big data)	To demystify big data for government leaders by explaining what big data is and outlining its potential benefits.	Policy report with contributions from the Federal Big Data Commission and participating industry and government experts.	- Highlighted variability as one of the identifying characteristics of big data and clarified the expansion of big data Vs.
Thabet and Soomro (2015)	General (cross-sector discussion of big data challenges)	To identify and discuss the main challenges of big data and outline a theoretical framework.	Conceptual article that synthesises existing literature on big data challenges and organises them into categories.	- Highlights technical challenges, such as volume, velocity and variety. - Discusses data processing, storage and quality challenges. - Highlights management challenges, such as privacy and security.
Udeh et al. (2024)	Nigeria (financial fraud in digital transactions)	To examine how big data can be used to detect and prevent financial fraud. Discusses its techniques, advantages and challenges.	Review conceptual paper that draws on existing literature and examples of fraud detection systems that use big data analytics.	- Describes how big data analysis can support early detection of suspicious patterns in digital financial transactions. - Presents techniques and tools used in fraud detection. - Highlights challenges of data quality, privacy and infrastructure limitations and the need for robust governance models.
van der Voort et al. (2019)	EU (public sector decision making in the European context)	To investigate how algorithms and big data interact with public decision making.	Qualitative empirical multiple case study of two big data processes in public decision-making using interviews and qualitative case material to examine interactions between data analysts and decision makers.	- Discusses big data characteristics, expansions and the evolution of the concept. - Suggests that greater volumes of data can lead to better decision making. - Argues that algorithms hold promise for objectivity, efficiency and rationality.

Vestby (2019)	Norway (policing and machine learning decision models)	To examine what police and other stakeholders think of machine learning in policing.	Conceptual analytical article that uses examples of machine learning in policing and discusses key issues, not a primary empirical study.	- Highlights the importance of machine learning for crime prediction and detection. - Presents practical examples of using machine learning by the UK Serious Fraud Office and the Norwegian Labor Inspection Authority. - Distinguishes between human and machine learning decision making in policing.
Wadhwani and Wang (2017)	Global (big data in IT organisations)	To outline the main challenges associated with big data and discuss the technical and organisational solutions.	Literature-based review of existing academic and industry literature on big data with descriptive discussion of issues and challenges with no primary empirical data.	- Identifies different softwares that are be able to address big data's technical challenges such as: Apache Hadoop, Spark, Grid Computing, OLAP, and Hybrid SAAS.
Winchester (2020)	United Kingdom (national serious and organised crime)	To provide an overview of serious and organised crime in the UK including its scale, forms and government responses.	Research briefing that synthesises official statistics, government reports and research on serious and organised crime in the UK, no new empirical data.	- Describes and lists the types of serious crime in the UK. - Highlights the challenges in tackling serious and organised crime, such as the complexity of networks and resource constraints. - Provides context on UK government strategies and measures aimed to combat serious and organised crime.

Xu, Cheng and Sugumaran (2020)	Global (crime prevention and control using image and cloud-based analytics)	To design and demonstrate a big data analytics approach for crime prevention and control.	Quantitative technical modelling paper that utilises crime related image data and implementation of image processing with big data techniques.	- Argues that using big data to predict crimes can consequently contribute to crime control.
Yadav et al. (2017)	India (crime pattern detection and prediction)	To develop and test a crime patter detection, analysis and prediction.	Quantitative paper the uses crime datasets and applications of data mining.	- Highlights that criminals are exploring the latest technologies to commit their crimes, suggest that data mining can uncover patterns and hotspots, and effectively allocate resources.
Ylijoki and Porras (2016)	Finland (information systems context)	To map and analyse the different definitions of big data in the literature and discuss how they shape perspectives in understanding them.	Mapping study / structured literature review that collects and compares definitions from academic and other sources and groups them into different categories.	- Shows no single agreed definition of big data – rather coexisting multiple perspectives. - Highlights the value that big data can add for public and private organisations. - Argues that definitional ambiguity can cause confusion but also highlights the multi-dimensional nature of big data. - Recommends that researchers and practitioners be explicit about what definition they are using.

Zainab and Dhanda (2018)	India (presented at the *Smart 2018* IEEE conference)	To review how big data and predictive analysis are being used across various sectors, its potential benefits and challenges.	Conference paper based on secondary sources of conceptual discussion of big data and predictive analytics and sectoral use cases rather than new empirical data collection.	- Outlines the concept of big data and predictive analytics and their relevance for decision making. - Highlights the potential of big data analytics to predict incidents, discover patterns and detect crimes. - Identifies challenges such as data quality and the need for advanced technological infrastructure.
Zikopoulos et al. (2012)	Global (enterprise in information technology)	To introduce the reader to big data analytics in enterprise environments and explain concepts and technologies.	Professional book that is technically oriented, synthesises tools, architectures, and examples from practice.	- Explains the types of data, their rapid increase and characteristics. - Suggests cases where big data analytics can add value, such as fraud detection and operational optimisation. - Highlights that an effective use of big data depends on its technical capabilities.

Appendix B: Roles of Invited Professionals Who Did Not Participate

Police officers' roles:

- Commander specialised in human trafficking, online child abuse, cybercrime, and major crime
- Community police officer
- Criminal investigation detective
- Data analyst and serious crimes investigator
- Deputy Assistant Commissioner
- Detective inspector specialised in risk analysis and management, and major crimes
- Director in an international policing organisation
- Director of AI department
- Expert in cyber security
- Head of Command Specialised in Criminal Justice at a local, regional, and national level
- Head of digital forensics
- Project manager in a Cybercrime Directorate

Professional roles:

- Advisory member of the Scientific Council for Government Policy
- Associate Professor at the Department of Information Security and Communication Technology
- Computer scientist expert in machine learning and pattern recognition
- Current Chief Executive Officer of digital security solutions company and former police officer specialised in counter-terrorism at a commanding level
- Cyber Security and Digital Transformation Advisor
- Data mining consultant, trainer, speaker, and author
- Director of a big data institute
- Director of a Centre for Emerging Technology and Security
- Director of a Centre for Evidence-Based Policing
- Director of Studies and Technology in a university
- Expert in advanced studies in statistics, data mining, and algorithm analysis
- Founder and Chief Executive Officer of an AI and Big Data Centre
- Head of data analytics in semi-government entity
- Higher education consultant in data architecture and modelling

- Information System Department Chair, doctor teaching information management science
- Principal Investigator of a Big Data Surveillance Project
- Principal Investigator of a Data Stories Project
- Professor and director of a Centre for Research on Security Practices
- Professor, Turing Fellow, and member of an AI Centre
- Research Associate with the National Centre of Geo-computation and a Centre for Applied Data Analytics
- Researcher at the Fundamental Rights Centre, and member of Crime and Society Research Group
- Smart City experience advisor

For the following – Professors, Associate Professors and Doctors – each separate designation represents an individual

Professor

- Applied artificial intelligence and technology ethics
- Artificial intelligence
- Data mining, text mining, and machine learning
- Global security and technology
- Information technology, cybersecurity, AI, and computer security
- Law, and researching the interaction between law and digital technology in the use of AI and innovative technology in policing and national security
- Law, evidence and criminal procedure, privacy, and civil rights
- Pervasive computing
- Policing, privacy, and technology
- Political science and public policy
- Social statistics specialised in confidentiality, privacy and anonymisation, synthetic data
- Span network complexity, geo-computation, space-time analytics, and big data mining
- Statistics

Associate Professor

- Data analytics
- Data science, machine learning, and predictive analytics
- Information systems

Doctor

- Actuarial forecasting of criminal justice behaviour
- Big data analytics, data mining, and data warehousing
- Criminal justice
- Cybersecurity, computer network management, and business intelligence
- Data science
- Information management science

Appendix C. Policing Participants' Interview Questions

1. Can you tell me about your current role and previous experience?
2. Can you tell me more about where your police force is with the use of big data?
3. From your experience and understanding, what are the definitions of big data and serious crimes?
4. What are the big data's characteristics, and is there a criteria to consider a dataset to be big data?
5. In your opinion, what do you think about using big data in the policing field in general?
6. Do you think that big data can be used by police forces specifically in the criminal field? If yes, how? If not, why?
7. Do you think that big data can be effective in detecting serious crimes and/or suspects? If yes, why & how? If not, why?
8. Are there any concerns regarding using big data in detecting serious crimes and/or suspects?
9. Is there a future of using big data in the criminal field, if yes, how? If not, why?

The following questions were based on whether they considered the use of big data effective:

10. Are there any advantages and/or disadvantages of using big data in detecting serious crimes and/or suspects?
11. In your opinion and from your experience, what are the types of serious crimes that big data and its tools can be effective in detecting?
12. Are there any successful applications/methods/strategies that big data is utilised in detecting serious crimes and/or suspects?
13. What are the most useful types of big data datasets/variables that can assist in detecting serious crimes and/or suspects?

Appendix D. Big Data Participants' Interview Questions

1. Can you tell me about your current role and previous experience?
2. From your experience and understanding, what is the definition of big data? And what are the big data's characteristics?
3. Is there a criterion to consider a dataset to be big data?
4. In your opinion, what do you think about using big data in the policing field?
5. Do you think that big data can be used by police forces in the criminal field?
6. Do you think big data can be effective in detecting serious crimes and/or suspects? If yes, can you share any examples?
7. Are there any advantages and disadvantages of using big data to detect serious crimes?
8. Are there any concerns regarding using big data analysis to detect serious crimes and/or suspects?
9. Are there any proposed solutions to the big data challenges?
10. Is big data analysis and its tools effective in detecting any other related subjects in different fields? If so, how? And what can policing learn from these applications?
11. What are the most useful artificial intelligence tools that can be effective in analysing big data sets? And what are the opportunities and challenges for these in policing?

The following questions were based on whether they consider the use of big data effective:

12. What are the most effective tools, methods, current applications to detect serious crimes and/or suspects?
13. Are there any specific types of serious crimes that big data and its tools would be effective to detect?
14. What are the most useful types of datasets and variables that can assist in detecting serious crimes and/or suspects?

Appendix E. Ethical Consent Form

EMAIL CONSENT SCRIPT - Obtaining informed consent from research participants via email

Study title: *Big data and serious crimes*
Research Ethics Committee Reference Number: 22/LCP/005

Please read the statements below. If you are happy with all of the statements, please copy and paste them into an email and send it to me at [*K.I.ALALI@2018.ljmu.ac.uk*]. This will be considered to constitute giving your consent to participate in the study.

If you have any questions about the study or the statements below, please do not hesitate to contact me.

1.	I confirm that I have read the information sheet dated 9/03/2022 (version 1) for the above study, or it has been read to me. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2.	I understand what taking part in the study involves	
3.	I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions I can withdraw from the study at any time, without giving a reason and without penalty or my legal rights being affected.	
4.	I have been advised about potential risks associated with taking part in this study and have taken these into consideration before consenting to participate	
5.	I understand that the study involves taking audio recordings of the interview and I am happy to proceed. I understand that I can stop the recording at any time during the interview and/or refuse that the interview to be audio recorded.	
6.	I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	
7.	I understand that personal data will be retained beyond the duration of the study	
8.	I understand that personal data will remain confidential and that all efforts will be made to ensure I cannot be identified in reports or any further outputs	
9.	I agree to take part in this study	YES / NO

Appendix F. Participants' Quotes

Theme 1: Defining big data

Ongoing debate

P2: *"Big data depends on its definition and characteristics, what is considered and not considered big data? There is a huge debate in the field, you cannot discuss something when we don't agree on the definition, and everyone has their own perspective."*

P4: *"Big data depends on its definition and characteristics, what is considered and not considered big data? There is a huge debate in the field, you cannot discuss something when we don't agree on the definition, and everyone has their own perspective."*

A1: *"There is no rigid definition of big data."*

A3: *"I think I follow a traditional three V one."*

A6: *"This isn't the kind of term to be defined in standards."*

A7: *"So we do not really have a clear definition of big data."*

A2: *"Okay, so you can find a textbook definition which will talk about the three Vs… I think I would unpack the definition of big data slightly differently."*

Perspectives on definitions

P1: *"Multiple input of information that can benefit us in understanding how, when, and what is the reason for things to occur and analysing data to understand the past to predict the future."*

P2: *"For example, if a data set has some Vs but not the other, does it lose its big data characteristics? Some big data sets have volume and variety but not the velocity, is it considered big data or not?"*

P3: *"Big data from a technical standpoint is to use hardware that accepts data input from sensors and multiple sources which is then analysed by artificial intelligence to create relationships to support decision making."*

P5: *"Big data is usually known by its three Vs: volume, velocity, and variety."*

P6: *"Larger, more complex data sets, especially from new data sources."*

P7: *"High volume with high intensity data that has unknown value until explored."*

P8: *"Big data is the diffusion of lots of data from disparate sources which creates links that may identify people engaged in serious and organised crime or detecting a serious crime that is taking place which would not be otherwise known if the user did not have the ability to bring data together."*

P9: *"… initially, what was considered the first three Vs of big data, volume, velocity, variety, and then we have got, you know, more recently you've got value and veracity."*

P1: *"Multiple input of information"*

P3: *"data input from sensors and multiple sources"*

P6: *"Larger…"* and *"new data sources"*

P7: *"unknown value until explored"*

P8: *"…lots of data"*

A1: *"Big data is data that comes from many sources without any limit, such as the web or internet which is an open-source big data."*

A2: *"I think I would unpack the definition of big data slightly differently. So, I would say that what we should seek to do is to seek to understand how big data offers something significantly different to data processes than that has happened in the past."*

A2: *"So you can find a textbook definition, which will talk about the three Vs, which is very much driven by a kind of data scientists' definition, I think the three Vs are something like the velocity variety and volume."*

A3: *"I think I follow a traditional three V one, I do believe that visualisation is also an important factor in my definition."*

A4: *"Big data is a set of techniques and technologies that need a new form of integration to uncover large hidden values from large data sets that are diverse, complex, and in massive scales."*

A5: *"Big data are relevant, connected data, and size is a term that is too abstract."*

A7: *"Big data is having fast moving and slow-moving data signals, fast moving signals are constantly updated like a phone location for example, whilst an example of slow-moving signals is a permanent address or a bank account number."*

A7: *"Big data normally for me means using the data that an organisation holds and combining that with additional data in and building what we call data universes [which] are connected ecosystems."*

A1: *"…data that comes from many sources"*

A1: *"without any limit"*

A2: *"seek to understand how big data offers something significantly different"*

A4: *"large data sets that are diverse, complex, and in massive scales"*

A4: *"to uncover large hidden values"*

A7: *"…fast moving signals"*

Theme 2: Big data characteristics

P3: *"With big data the quality of data is important, and this is what characterises big data"*

P4: *"If we restrict characterising big data with a changing variable that is constantly changing like the power of computers that we have, then the definition of big data will constantly change."*

P6: *"Veracity, validity, variability, volatility, visualisation, and value"*

A4: *"So, if we have all these huge amounts of data, but we can't get insights from this data, then that might not be something useful to us. So value, it's also one of the main characteristics of data."*

A3: *"That is academia for you isn't it, because we all like to become wordsmiths and invent new terms and expand existing terms to progress our own academic careers."*

A7: *"It is normally velocity, which is how fast the volume is produced"*

A7: *"That is the beauty and pain of working in AI. Everybody makes something up to sound like an expert."*

Big data criteria

P3: *"No criteria at the moment, the main criteria is to be relevant"*

A3: *"No, I don't think, I think because the term is so much a slogan or a catchphrase, it can be used for any type of data set which is larger than normal."*

A4: *"The volume is getting actually less attention compared to the complexity which is the variety of data… So, if we have data that is of various sources, and with different forms, then we can say that this data is basically big data."*

A6: *"This isn't the kind of term to be defined in standards so not sure that there is a criteria-based definition"*

A7: *"I wouldn't agree with that definition because most of the models are run on the cloud anyway. So, if you say it cannot be handled by a computer, and then you kind of be like, 'Why are you not working on the cloud?'"*

A7: *"I would say it depends on your architecture and would not define the data on the technology that you are using"*

A8: *"It's a little bit subjective, but we can say that starting from 1–10 TB depending on the complexity of the data (video, text, pictures, etc.) we can consider that this is big data."*

Big data and policing

P1: *"It is essential, significant, and a must, and to develop your police force you need to understand and use big data"*

P2: *"With the global development, using big data in policing is considered a part of our job, and it is an important part of the digital world"*

P3: *"Using big data in policing is very relevant and it can be done as any other field"*

P4: *"I am a huge advocate of using big data in policing in general as it can improve police operations and lead to better results"*

P6: *"We are using big data to investigate the traffic accidents to cluster accidents and analyse the most reasons of causing accidents, locations, time, nationalities, gender and other factors and present the trend of these factors. Based on that our department build their studies to draw the strategies to avoid traffic accidents."*

P7: *"Although it brings some challenges, it has great opportunities"*

A1: *"The use of big data in the policing field is definitely effective"*

A2: *"I think the thing about new technologies or emerging technologies is that they are new by definition, and therefore we do not really know the impacts and consequences that they will have."* and *"new or emerging technologies"*

A4: *"So, object detection could be one of the applications of big data policing because it basically can help police to track and monitor individuals who are suspected."*

A5: "*Good, just like any other social and economic problems*"

Big data and serious crime investigations

P1: *"Yes, it can be very useful"*

P2: *"...big data can for sure be effective"*

P3: *"It can be effective but it depends on how clean your data is as it can create bias in the outcomes"*.

P4: *"I am certain that big data can be effective in detecting serious crimes and there are very positive results"*

P6: *"Yes, to know how the crimes happened and the ways to prevent these crimes"*

P8: *"...if it was applied within the legal constraints, its potential is enormous"*

P9: *"I think [it] can and can be very effective, but it is having the resources, the equipment, and the staff in order to deal with and manage it"*

A3: *"I think it can probably be become very very effective. I also think there are some risks here"*

A6: *"Sure, facial recognition is an example of that"*

A7: *"It can lead to tremendous insights, we have used it successfully from fraud detection to demand forecasting, from optimisation of staffing to counter terrorism in all these fields, it can really help you getting insights"*

A8: *"From our experience in …. we think big data can indeed support crimes and/or suspects detection but its need to gather data from the field (e.g. CCTV, gate entry, database coming from authorities, etc.). In order to support the data processing with crossing data from various sources and to support the algorithm learning, even if we also think we have to follow white box paradigm."*

Types of serious crime

P2: *"Big data is involved in every crime"*

P5: *"Cybercrimes, financial crimes, fraud, drug trafficking"*

P6: *"Terrorism, treason, arson, murder, rape, and robbery"*

P8 : *"There are all sorts of types of fraud going on online, it could be really helpful in that field in particular but I think in all serious crimes. I think it has value and it has the potential to be exploitable particularly in fraud."*

A6: *"Burglary"*

A7: *"Fraud, drugs, online sexual offending"*

A8: *"AI that analyse big data can detect cyber-bullying on social media, all kind of offenses in public places thanks to CCTV, predict crimes (e.g. robbery in gas station) before it happen in some districts during a specific time slot, etc."*

Theme 3: Advantages and disadvantages

P1: *"big data can be used in crime prediction, planning and forming strategies"*

P2: *"If I suspected someone and I do not have a big data system, I will be searching in every system individually such as the ANPR, car registrations, etc."*

P2: "*…in a click of a button"*

P3: *"Policing is very responsive and not proactive, we do not have systems to predict crimes and it all depends on how clean the data is"*

P3: *"It makes the organisation more data driven to better utilise resources…it can automate the process of having a system that takes decisions by itself in the future"*

P3:*"It can help discovering crime hotspot heat maps, find patterns, and study them to build better prevention programmes"*

P3: *"…humans have limitations, we cannot look at different data sources and make quick conclusion or try to find a pattern and build relationships, whereas big data with AI can do that in a fast way"*

P4: *"Data in the policing field is very important, it allows us to react quickly in a more efficient manner, overall refines the process and allows you to get better results."*

P5: *"It can be used for the police force to be proactive and reactive"*

P6: *"Detect, predict, and suggest best solutions to deal with crimes before happening"*

P8: *"So, the advantages are obvious that you can search masses of data technically and get answers to questions very quickly. They might take you weeks, months or even, never, in most of this investigation."*

A1: *"To the police the advantages are more than the disadvantages."*

A2: *"So I think the advantages of big data are that advances in computing being that you can process such large quantities of data that were not previously possible for mankind. So, you can take into account you know, vast datasets that are beyond the sight of human comprehension, you know, they their data is massive."*

A3: *"you can identify patterns for example criminal behaviour"*

A5: *"Yes, at least potentially. It can help understand better about events"*

A7:*"I remember 10 years ago when the whole big data started, we built a demand forecasting for UK police force to basically understand exactly where the hotspots and where should they put their police cars."*

A7: *"A lot of prediction optimization and grouping algorithms are extremely effective, when we look at for example fraud detection and similar kind of cases, you are trying to find the needle in the haystack."*

A7: *"So one of the things that we see is that every human being has a pattern, and just like normal people have a pattern, so do fraudsters have a pattern, so do criminals have a pattern, and there is actually a lot of interesting use of policing information and policing data for retail. So yeah, it is a very good field for the police to invest in."*

A7: *"In the long run, it always makes sense to have people in house because the policing sector is a specialist sector, and you could then basically build up your own teams. That is at the beginning, very expensive, so you invest in the right people, you invest in the right technology in the right training, and you need to make sure they have the access to the data, and they actually need to have a certain level of confidentiality."*

A8: *"Use case 1: We can have what we started to recall in the last question: operators lose a lot of time to watch about ten hours and more of video to try identify and catch a suspect when he was seen making an offense. Using an AI trained thanks to a huge volume of data, we can learn how to identify a same*

individual on multiple images and calculate his path based on CCTV camera's locations automatically without mobilizing an officer to do it"

Use case 2: Other example, thanks to geolocation data collected from a smartphone's suspect, we can map and identify many aspects of its life like his occupation, where he lives, his gym location, etc., in order to predict some behaviours."

A8: *"That is what we can retrieve from big data, learn from the past to better predict and anticipate the future"*

A8: *"It also help us to recognise criminals and criminal behaviours by using deep learning algorithm and various pattern to target 0 crime and anticipate when it comes to predict offenses in a specific spatio-temporal window based on historic data. We can take advantage of big data for many use cases, from the need to store huge number of statistical data in order to identify the district with the higher risk to estimate automatically a criminal path on hours of CCTV video coming from various sources."*

A9: *"I think that the advantages far outweigh the disadvantages."*

Disadvantages of using big data in serious crime investigations

P4: *"A drawback is if you have untrained data scientists or untrained algorithms you will provide wrong data"*

P8: *"I think it will be a mistake to think it's the answer to all of our problems. You know, there is always going to be a role for traditional policing, talking to people doing other types of investigative tactics such as physical surveillance."*

"[I] can't see many disadvantages, other than the logistical limitations that we've already discussed in terms of people and systems"

A2: *"I think, maybe disadvantages is not quite the right word, maybe challenges is better"*.

A2: *"bad big data science"* and *"I am not saying that would happen, it is the risk of that happening"*

Theme 4: Challenges in using big data in serious crime investigation

Conceptual challenges concerning big data

P3: *"The science of big data and artificial intelligence is new, and the culture of big data is still not there"*

P3: *"...so they must be shown the capabilities of big data and how it can actually improve decision making"*

P3: *"There is a resistance to change from seniors or employees that think the automation can replace their jobs"*

P4: *"The issue is not about big data…but the lack of understanding of how computing works and what are its subfields, people see one person on a computer and expect him to do all kinds of computing jobs."*

P4: *"…training should not only be for the employees but will be beneficial for the upper management as well"*

A4: *"There are some organisations or even maybe the police, they feel more comfortable in the way they operate than changing to new technology"*

A4: *"The people in charge should provide all the type of training that helps these officers to not only know about big data but also learn some of the technologies and some of the techniques that can help them do their work better."*

Financial challenges

P1: *"The technologies cost is very high"*

P2: *"For example, if there was a terrorist attack and it costed the country around 1 billion in business losses, whereas the police had a chance to purchase a system for millions that would enable to prevent this attack, then it will be a good investment."*

P2: *"An equation is needed to evaluate the need of big data, if the system is for criminal investigations, is it a good system even if it solves one case? Is this correct based on its high costs?"*

P4: *"If it saves people lives, can we put a cost on it?"*

P8: *"… are preventing the exploitation of big data"*

P8: *"I express frustration [about] being able to see the potential of big data but because of financial and other restraints not being able to exploit it to its full capacity"*

A4: *"Financial matters are a core of big data challenge, because investment in these kind of projects require a huge amount of money, which some organisations might not really like"*

A7: *"It is something that is not quick and easy, it is something that is actually expensive and requires a lot of amount of skills. That is why it is actually a field where if you go with startups, or if you go with cheap vendors, you run a massive risk because you are dealing with sensitive information."*

A7: *"...just building AI is really cheap, building trusted AI really building good AI is really expensive and needs a lot of expertise"*

Human resource and training challenges

P1: *"It is not easy to find the qualified specialists in the field of data."*

P4 *"lack of training can be a challenge"*

P7: *"…put the correct resources in the correct operation"*

P8: *"But often you have got really good intelligence and you have got no resources to do anything with it. So, exploiting big data is one thing and then been able to do something about it is another thing"*

A1: *"There are many projects that failed due to lack of having the right people to deal with data"*

A2: *"…skill sets and specialists in big data for transformation"*

A2: *"…it is a challenge not a disadvantage"*

A2: *"I think there are challenges for public services to acquire the skills to use big data effectively. Now that cannot be underestimated because data scientists are in high demand and they are all getting all the good ones are getting swallowed up by commercial companies."*

A2: *"But I think the concern that I would bring forward, which may be different to other people would be around you know, the skills required around data science or how policing will get that skill set to use data science and big data properly. I think that is a genuine concern, and I know that that is happening across the public sector."*

A4: *"Finding the resources to manage and search for big data is challenging"*

Technical challenges

P2: *"One of the technical challenges is how to save the big data and until when, I cannot save all the data forever, and what software will analyse all this data?"*

P2: *"For example, can we save all Twitter's data since it was created? And will all of it be useful? If the answer is yes, we need huge infrastructure investments"*

P2: *"With every new system we need new infrastructure and so on"*

P4: *"My concern with data is the source"*

P4: *"…data cleaning is checking for missing values, and it is the process of using raw data to refine it to data that can be used afterwards."*

P5: *"the biggest challenge is ensuring the data is correct and of high quality"*

P8: *"So the forces held a huge amount of data but it was not joined up as they did not have the technical capability to export all of that data"*

P9: *"So if you are going to start collecting big data as a police force or as the national agency, to effectively deal with the data that you do collect, then you are going to have to invest in some kind of equipment to deal with it. If you are not going to deal with it, then what is the point in having it?"*

A4: *"So volume is basically one of the challenges in big data, but now with a revolution of internet we have also data that comes in various forms of video audio and text… we are no longer dealing with only*

structured data we have to deal with unstructured data and velocity is also becoming a challenge because we are getting data more than one can imagine."

A4: *"So in terms of the size, because now we have machines with huge space that can accommodate huge amount of data and also we have cloud technologies."*

A4: *"The other disadvantage of having huge amount of data is also the reduction of the data, how do we reduce the size of data without compromising the quality of the data that we have?"*

A7: *"That is always a big challenge for all countries around the world to know who is coming in and is it really the person who is saying that it is? The other big problem that we have especially in the Middle East is the translation between Arabic and English. A lot of the systems are built in English and cannot deal with Arabic language, and also you have 50 different ways of spelling Mohammed, with two m's with one m, with a's with one a, because it depends on who translated it when the data was entered."*

A8: *"There is big concerns on how to secure huge volume of critical private data in a connected world with is more and more threat with cyber-criminality and cyber war."*

Operational challenges

P2: *"double-edged weapon"*

P2: *"…not having data is a challenge, having excess data is a challenge, it becomes like it is not there because you will be lost and will not benefit from it."*

P2: *"In future, is the metaverse a source of data? Is verbal assault and sexual harassment in a metaverse considered a crime? With no known identities and no legislation, the characters assaulted each other, not a known person. What if they were known to each other and took their revenge in the real world?"*

P8: *"But one of the huge challenges the police face is they have got more intelligence and information and data than they can cope with…lack of intelligence was never our problem, it was having sufficient operational resources to act on the intelligence."*

P9: *"…that golden nugget"*

P9: *"Of course the other issue is for police and law enforcement in the UK, particularly is to get the lawful access. You have got to have the authorisation in place and that authorisation process is quite cumbersome, require quite burdensome with regard to the administration."*

A1: *"For example, facial recognition, in some cases if you cannot have it you cannot investigate"*

A2: *"I think there are challenges in terms of building data-sharing protocols, so there has to be some sort of protocol governance structure in place whenever the police share data with somebody. So that is a challenge"*

A2: *"The world of criminality massively moved online during the pandemic, and obviously policing has to go online as it is going to be a massive area of future activity"* and *"…online is the new frontline policing"*

Theme 5: Concerns regarding the use of big data in serious crime investigations

Bias

P3: *"Bias is the main concern and the main reason is how clean is the data"*

P3: *"In some police forces projects the AI software or algorithm is biased, for example in the US where they did a trial project which had outcomes that most of the criminals are from a certain race, and that is because the data that is fed to the algorithm is not clean so it creates bias."*

P3: *"It is effective if the data sources are clean"*

A2: *"I think this is where police forces have got into difficulties in recent years because they have not thought through some of these ethical consequences, and face recognition is an example of that. So the Metropolitan Police and South Wales Police have been trialling face recognition in public spaces. There has been a lot of criticism about them being ineffective, about bias. And it was legally deemed at the High Court that what they were doing was not legal at the end of the day."*

A2: *"So I think new technology should be used where appropriate by policing agencies, but there's always a degree of caution involved"*

A3: *"I think it can probably become very effective but I also think there are some risks here. One obvious risk which has been addressed several times is profiling."*

A3: *"I think it is important that you have first of all legal provisions, but also to actually try to roll out some kind of impact assessments, but you have some frameworks, and you have some guidelines, so people are actually enlightened about the risks, to me it is very much an educational challenge."*

A7*"Also, the second concern is that people are lazy and build models under pressure…and some AI models are often not audited, often built by somebody maybe a junior who does not really understand the impact."*

A7: *"…our world is extremely biased, and our data is biased, that is the main concern that I have"*

A7 *"So there are a lot of factors that we say these are the key minimum for ethical, trusted AI"*

"The big data needs also some time to be enough accurate which could not be relevant in some cases or need to have an organisation proactive to invest on such topic."

A7: *"Yes, so AI could be audited. AI could be tested for biases could and transparency, explainability and so on. It is still open research and people are working on it, and the problem at the moment is we*

do not have enough people to build the AI, so how are we expecting to find enough people to actually audit that on top of it."

A7: *"So we have toolkits to actually analyse models to try to open up the black box"*

A7: *"Sometimes you can fix the bias, but sometimes you cannot. Sometimes you just don't have the information and no real way of fixing it"*

A7: *"These in policing as important as anywhere else, but I think especially important in policing…We now say entrusted AI has to be transparent, explainable, robust, safe, identity protection, and unbiased."*

A8: *"Moreover, any algorithm doesn't understand all ethic concerns, so the output of this data processing could be no ethical at all for human using it."*

Privacy

P2: *"I have no other concerns other than the privacy"*

P2: *"Privacy of individuals can be considered as a floating concern and it depends on the culture of the society… In our culture having a camera is not considered a violation of privacy, whilst in other cultures it could be."*

P2: *"You have a crime, you have to protect the innocent people privacy and discover the criminal, if there was total privacy and no cameras installed that might lead to not discovering the criminal"*

P4: *"Privacy is something we always take into account, would you rather have more privacy but risk safety and security? Or would you rather have more security but less privacy? It is always a balance."*

P4: *"In some projects like the beach project, we would let go of them just to let people have more privacy"*

P5: *"Privacy is a wide argument, and there are legislations such as the GDPR in Europe to ensure there is no data abuse."*

P7: *"Searching through big data in the digital environment may lead into privacy intrusion of innocent people and it should be justified."*

P8: *"It is all about or it tries to be about balance, about protecting the public and protecting intrusion into people's private lives, it is all about trying to get that balance."*

P9: *"I think you have got to be realistic, so there is got to be the balance."*

P9: *"In the UK for instance, right to privacy is a qualified right, and if you are involved in serious and organized crime depending on your level of criminality, then you forgo the right to privacy, but you have got to put a lot of consideration in to justify that from a law enforcement perspective, and I think in this country, I think the balance is just about right."*

P9: *"It is got to be justified, it is got to be proportional, and it is got to be necessary."*

P9: *"There are always going to be the points if there is a concern and possibly that up against the arguments about privacy versus security."*

A1: *"A disadvantage to the people is their privacy"*

A2: *"In the UK having some face recognition is a brilliant example, there is a massive resistance across the world to having face recognition in public spaces."*

A2: *"Yeah, so I do not necessarily believe in arguments about balance. I think that I think we should strive to achieve security, safety and privacy. I do not buy into you have you cannot have both at the same time. I think that is a very old argument, going back 20 years, and misspelled over the years. And when people repeat that argument, you kind of think, okay, they have not got they have not developed a mature thinking around the technology."*

A2: *"So I think you know when the police use a new technology, they have to be super cautious in terms of thinking about the impact it will have on a whole set of different relationships…so it is changing relationships."*

A3: *"I mean, how can you actually protect privacy? I mean, part of it has obviously been all these data protection acts around the world. Establishment of data protection agencies or we call them privacy commissioners here, and I think we need something similar in artificial intelligence or big data."*

A3: *"I mean, there are also a growing interest in you know about privacy impact assessments. You must have heard about that. So, the Canadians have invented something called the algorithmic impact assessment. So, when you introduce a new system, prior to that you should actually both tick some boxes go through and check for any risks here, and how do we mitigate the risks? So, it is sort of what I would call a regulation by design really."*

A4: *"One of the main disadvantages is basically the privacy because sometimes data that we collect could be a personal data especially in crimes and all those areas. We have to look at, for example, individuals' profiles and all this information so privacy is one of the big issues in big data."*

A7: *"Privacy protection for policing is more important than any other industry"*

A8: *"Moreover, in some countries, people are concerned on the use of their data less eager to accept that any companies or organisation like the police administration can have access to private information and track them in everyday life"*

Theme 6: Tools and dataset

Technical tools

P2: *"FR can be helpful in searching for individuals"*

P4: *"The most useful AI tools depend on the programming language, and there a lot of tools and algorithms that are available, it also depends on the type of data you have"*

A1: *"A successful application of big data analysis in financial crimes for example if 20 computers were seized, a lot of time and resources are needed to analyse all of them. Therefor the use of AI and its tools will be more effective, such as pattern recognition, link analysis, document analysis, and natural language processing"*

A4: *"One of them basically and the most popular adopted technique is deep learning, and the reason for that is because it really works very well with unstructured data. Also, one of the advantages of deep learning in big data analytics is self-learning. As a part of reinforcement learning approach, it basically helps also in tackling some problems of huge amount of data."*

A4: *"So clustering-based techniques is very useful in policing, it helps in clustering segments of group of individuals or to identify crime hot spot areas that can basically be covered by police patrols… Predictive model is also one of the interesting techniques that can be used in the policing field, and it was recently used because of COVID-19."*

A5: *"It is hard for me to list out the so called most useful AI tools, there are many AI tools used in cybersecurity and most of them have their own unique features"*

A6: *"Depends on your definition of AI, but facial recognition, text and image analytics, social network analytics"*.

A8: *"Clustering and anomaly detection."*

Types of datasets

P1: *"Data and information is essential to investigate; without data you cannot do anything"*

P1: *"Social media."*

P2: *"Text is highly important if it came from a trusted source, then audio and video…"*

P2: *"Facebook might not be trusted because of people impersonating another people."*

P4: *"Firsthand data works better but it is very expensive to collect and takes a lot of time"*

P6: *"Live camera feeds, emergency service calls, and complaints registered with the police"*

P7: *"Phone numbers and IP addresses"*, adding that *"…numerical data has the greatest value"*

P9: *"The first one is obviously communications data…the contents of emails are potentially huge, but again, it is looking for that tiny piece of critical information or that piece of evidence in a massive volume of data"*

P9: *"There are going to be huge data on social media."*

A1: *"open-source intelligence, crime historical data and past reports"*

A2: *"Public services are used to working with their administrative datasets that are very reliable you know, so if we look at the tax records, they are very accurate"*

A2: *"So if you look around social media and Facebook and Twitter there are millions of fake accounts."*

A4: *"We have social media."*

A7: *"Telephone signals and segments and I think the number one I would go for is telephone behaviour"*.

A7: *"I think their own records the government holds on like the employees' information"*

A8: *"Social network data (Tweets, Facebook, Instagram)."*

A8: *"Mobile geolocation data, CCTV videos, Car plate and all police data legacy."*