PAPER NO. 09/2023

December 2023

[Data science, data crime and data law]

[Maria Grazia Porcedda, Trinity College Dublin]

[David S. Wall, University of Leeds]

Data science, data crime and data law

Maria Grazia Porcedda[1] & David S. Wall[2]

[This is a draft of chapter 17 'Data science, data crime and data law' forthcoming in V. Mak, A. Berlee and E. Tjong Tjin Tai, *Research Handbook in data Science and Law* (Edward Elgar 2024). Please only cite the version of record (published version).]

1. Introduction

This chapter explores the relationship between data science, data crimes and the law. At first edition we discussed how big data is driving data crimes. We showed how the law and data science could mutually help each other by shedding light on the ethical and legal devices necessary to enable big data analytic techniques to identify the key stages at which data crimes[3] take place and prevent them.

That work gave impetus to research on big data and cybercrime[4] that led to data crime modelling. In this chapter, we retrace our steps and supplement the analysis with considerations on how ransomware and Artificial Intelligence are adding a whole new significance to (big) data crime.

So, in section 2 we explore the strengths and weakness of big data analytics, including the implications of the shift from information ethics to data ethics. We observe the difference between personal and non-personal data which has implications for our understanding of data crime and the use of analytics to fight it.

[1] Assistant Professor in IT Law, School of Law, Trinity College Dublin, Ireland.

[2] Professor of Criminology, Centre for Criminal Justice Studies, School of Law, University of Leeds, UK.

[3] Data Crimes are conceptualised here as cybercrimes in which data is the primary focus; examples include data theft (including Big Data), data destruction, data extortion and the use of stolen data to commit further crimes. For further discussion of data crimes, see sources in footnote 4 below and later in this chapter. Please note that this is an evolving field.

[4] Maria Grazia Porcedda and David S. Wall, 'The Chain and Cascade Effects in Cybercrime: Lessons from the TalkTalk Case Study' (*IEEE Euro S&P 2019*); Maria Grazia Porcedda and David S. Wall, 'Modelling the Cybercrime Cascade Effect in Data Crime' (*IEEE Euro S&P 2021*).

In section 3, we look at the data crimes created by Big Data and dissect the relationship between cybercrime and data crimes to show that in order to understand risks, threats, and harms, it is necessary to look into the technology, the information and also the data. By so doing, it becomes possible to uncover the nature of the impact of the internet technologies in terms of the 'cyber lift' they create which, importantly, includes the 'cascade effect' of cybercrime when 'upstream' cybercrimes such as data theft subsequently cascade downstream to cause further crime. We discuss ransomware attacks, which represent the sinister face of data (even Big Data) crime today and then consider the impact of applying Artificial intelligence to resolve the data crime model.

In the last part, section 4, we explore how data science and law could mutually challenge or help each other. We examine the considerable evolution of the applicable law on data science and data crime in the EU, with a special focus on data economy instruments in force in summer 2023 (free flow of non-personal data, PSI Directive, DGA, DSA and DMA) and open matters that need to be resolved for data science to support the fight against data crime.

2. Data Science: power and limits of big data analytics

2.1. Theoretical considerations on data analytics and science

Data Science, as outlined in the introductory chapter of this collection, concerns the application of data analytics (statistical techniques) to obtain useful information from existing computer datasets. Those data sets, often referred to as Big Data, are characterized by the three 'Vs'[5] - volume, velocity, and variety. The term 'Big Data' can also be used in a looser sense, to refer to the products of datafication,[6] for example,

[5] Doug Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety* (Meta Group Inc. Application Delivery Strategies 2001).
[6] Kenneth Neil Cukier and Viktor Mayer-Schoenberger, 'The Rise of Big Data. How It's Changing the Way We Think About the World' *Foreign Affairs* (4 April 2013).

the data sets that result from various devices which underpins the emerging data economy.[7] Indeed, data science stems from the combination of big data sets that result from datafication[8] with the algorithmic techniques and tools used to process them, particularly data analytics software that combine statistical analysis with machine learning.

Big Data is marketed as a descriptive and predictive tool capable of identifying new truths about social and physical phenomena that were previously un-researchable on such a large scale. As such, it holds great promises for boosting the economy, and addressing some societal problems.[9] In the words of the General Data Protection Regulation (GDPR): [10]

> "[b]y coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression... Within social science, research …results obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people and improve the efficiency of social services" (Recital 157).

The potential of Big Data analytics to "collect and analyse large amounts of data to identify attitude patterns and predict behaviours of groups and communities",[11] makes

[7] European Commission, *Building a European Data Economy* ((Communication) COM (2017) 9 final, 2017).

[8] That is, the transformation of information collected digitally into data that can be analysed and monetized. See Cukier and Mayer-Schoenberger, 'The Rise of Big Data. How It's Changing the Way We Think About the World'.

[9] Council of Europe (2017). European Commission, Communication 'A European strategy for data', COM (2020) 66 final.

[10] Regulation 2016/679/EU of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such data, and Repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1.

[11] Council of Europe (2017), p. 1.

it an interesting tool to assist in the fight of offences, including data crimes, but there are limitations on the use of this technology, as we discuss later in section 4.

2.1.1. The strength of data science is the variety of its data

What makes Big Data so powerful is the fact that it encompasses such a wide variety of personal and non-personal data, which are often regulated by different bodies of laws.[12]

The category of personal data contains data revealing information of varying degrees of sensitivity and pseudonymized data (including public sector information). Personal data are defined differently in different jurisdictions, but it is understood that they include information capable of identifying a natural person either directly, or indirectly (see definitions as § 2 in ISO/IEC29100[13], art. 2a of Convention 108,[14] and Art. 4(1) in the GDPR[15]). Pseudonymized data consist in replacing "identifying information with an alias" (§ 2.2.4 ISO/IEC29100); as for the GDPR, a data subject can only be identified with the use of additional information, which should be kept separately and protected by means of technical and organizational measures (Art. 4(5) GDPR).

The category of non-personal data includes raw/technical data, trade secrets, state secrets, IP, public sector information,[16] and anonymized data. Raw and technical data are produced by systems in an intelligible form. Trade and state secrets are confidential information, whose access is restricted according to some form of authorization. Intellectual property is a creative product of intellectual labour, the access and fruition of which rests under the control of the holder of copyright and intellectual property rights. Anonymized data either do not relate to a natural person,

[12] This is the case of the European Union, e.g. with the General Data Protection Regulation (2016/679/EU). See section 4 for further instruments.

[13] International Organization for Standarization (ISO), *International Standard ISO/IEC 29100:2011(E) Information technology — Security techniques — Privacy framework* (2011).

[14] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Council of Europe, CETS n. 108, 28 January 1981, as modified by additional Protocols no. 181 and no 223 (not yet in force).

[15] General Data Protection Regulation (2016/679/EU).

[16] See in this regard also Chapter 8 on Property Law and 9 on Intellectual Property Law of this volume.

or they no longer relate to an individual, for example, when "identity information is either erased or substituted by aliases for which the assignment function or table is destroyed" (e.g. § 4.4.4 ISO/IEC29100).[17] GDPR addresses anonymous data in passing[18] stating that 'identifiability' depends on all the means reasonably likely to be used based upon objective factors such as cost, time and technology (recital 26), including technological developments. The latter puts a dent into the possibility of truly anonymous data. In fact, because of the advances in data science and its enabler, cloud computing (see *infra*), anonymity may be more of an aspirational goal than a given.[19] The distinction between anonymized and pseudonymized data is important and becomes fundamental when discussing the use of data science in the fight against crime, as we address later in section 4.

Figure 1, below, summarizes the different types of data that can be analysed simultaneously with big data analytics (and hence co-exist in a given dataset).

[17] International Organization for Standarization (ISO) (2011), *International Standard ISO/IEC 29100:2011(E) Information technology — Security techniques — Privacy framework*.

[18] Since anonymous data is no longer personal, it is not addressed in the legislation (Recital 26 of the GDPR).

[19] See among others, Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland, 'Unique in the shopping mall: On the reidentifiability of credit card metadata' (2015) *Science* 347 (6221), 536-539; Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye Estimating the success of re-identifications in incomplete datasets using generative models (2019) *Nature Communications* 3069.

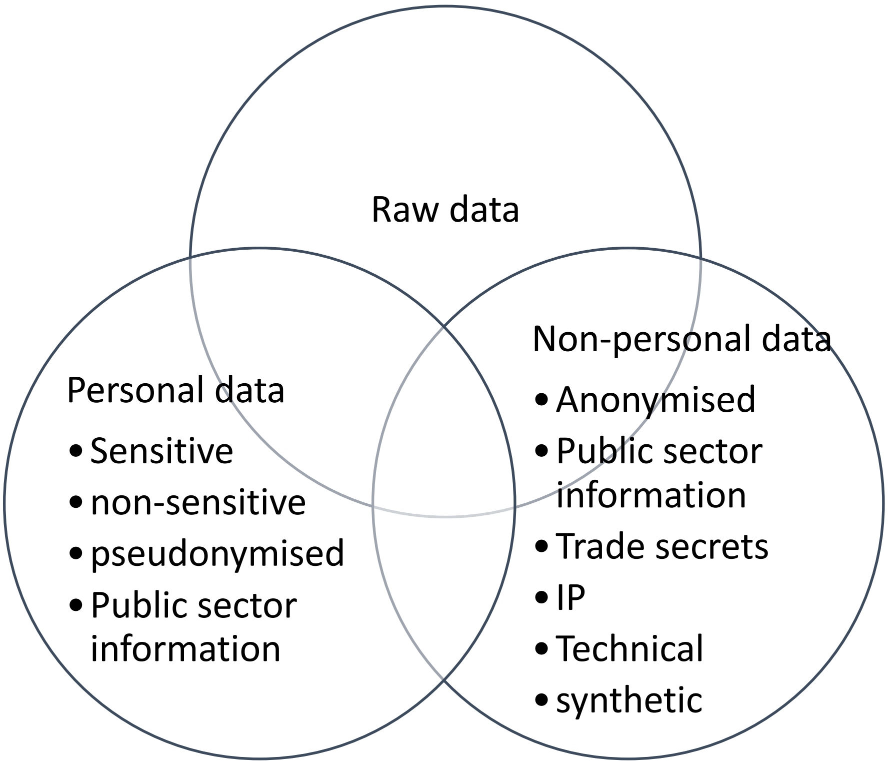


Figure 1: Types of data in large data sets

2.1.2. Endogenous and exogenous limitations of data science

Floridi and Taddeo[20] have commented on the pre-eminence of data as an object by noting that data science caused a shift from information ethics to data ethics with regard to data-related algorithms and practices. This shift casts a shadow on the information contained within the data and also the computers and technology making the data flows possible. Although in practice, the two are closely related; indeed, the two are often conflated in legislation.[21]

We agree with Floridi and Taddeo that focusing upon data can add new insight into understanding and fighting crime, particularly where cybercrime is understood to be a

[20] Floridi and Taddeo (2017), 'What is data ethics?'.
[21] For a discussion of this debate, see Maria Grazia Porcedda, *Cybersecurity, Privacy and Data Protection in EU Law. A Law, Policy and Technology Analysis* (Hart Publishing 2023), ch 4.

data crime (as we discuss *infra*). Nevertheless, we also believe that when focussing upon the use of analytics to support the fight against data crime it is necessary to look at the data, the information it contains and the underlying technology as part of the same ecosystem. One should not overlook the fact that data represent information and that data flows are made possible by technological applications. These elements are crucial if we want to appraise the exogenous and endogenous limitations of data science.[22]

The consideration of the types of information contained within data is a necessary task to appreciate the exogenous limitations of data science. Particularly, the risks against the person and the community because the GDPR and Charter state that the benefits of data science cannot be (in the European Union) lawfully achieved at the expense of the protection of a natural persons' data. The Council of Europe released their Guidelines on the use of Big Data which account for the collective dimension of the risks of big data.[23] In turn, the process of protecting data shields information concerning the private life of individuals and their private communications and personal data from unwarranted use. This respects the right to private and family life that is enshrined in several regional and international sources of human rights law and is protected as separate rights in some jurisdictions (chiefly the EU).[24] Because of this, safeguarding privacy and data protection represents an important limitation to the indiscriminate use of data analytics.[25]

[22] For critics, see ibid.

[23] The data economy is powered by the making available of data sets collected from sectors such as government, research and physical devices (e.g., Internet of Things). Council of Europe, Consultative Committee of the Convention for The Protection of Individuals with Regard to Automatic Processing of Personal Data Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data (T-PD (2017)01).

[24] Note that, in the European Union, the protection of personal data is a fundamental right enshrined in article 8 of the Charter of Fundamental Rights of the European Union. In the Council of Europe, it is safeguarded as part of the right to respect for private and family life (article 8 ECHR), as specified by Convention 108 (fn 13).

[25] For further discussion on this point, see Omer Tene and Jules Polonetsky, 'Privacy in the Age of Big Data: A Time for Big Decisions' (2012) 64 Stanford Law Review Online; Christopher Kuner and others, 'The Challenge of 'Big Data' for Data Protection' (2012) 2 International Data Privacy Law; Ann Cavoukian and Jeff Jonas, *Privacy by Design in the Age of Big Data* (Information and Privacy Commissioner, Ontario, Canada, 2012).

The endogenous limitations ironically stem from the very technological enablers powering the data analytics, namely networked and cloud computing, machine learning and artificial intelligence. On the one hand, the predictive capabilities of big data and machine learning are far from perfect and therefore their use to support criminal investigations have been oversold,[26] as we discuss in section 4. On the other hand, data science harbours new cybercriminal opportunity to victimise at personal, business and nation state levels. At first edition, we argued that the demand for Big Data is stimulating criminal data markets, making data both a target of criminal action and also used to facilitate offences against data in cyberspace, plus criminals also appear to be using Big Data analytics to maximise their victimisation.[27] Over time this has taken on the sinister shape of ransomware attacks and we are now at the threshold of further game changes effected by Artificial Intelligence.

3. Data Crime: The Downside of Data Science and the AI-turn

Big Data has become the target of cybercrime because of its inherent market and strategic value when stolen, combined with the fact that its digital and networked qualities make it possible to steal in bulk. So, we need to understand Data Crimes in terms of Cybercrime. Before that, however, we need to untangle the term cybercrime which often confuses readers because of its general application. Cybercrime is understood here in terms of the level of transformation by networked and digital technologies and also its *modus operandi.*[28] By simply imagining what the crime would look like if the cyber-element (impact of internet technologies) were to be removed, a more accurate understanding of the cybercrime can be achieved. This 'transformation test'[29] indicates how the crime has been transformed by technologies.

[26] Chan and Bennett Moses (2014), 'Using big data for legal and law enforcement decisions: testing the new tools', p. 643, 678.

[27] As discussed in David Wall 'How Big Data Feeds Big Crime, *Current History: A journal of contemporary world affairs* 117(795): 29-34 (2018)

[28] David S Wall, *Cybercrime: The transformation of Crime in the Information Age, 2nd Edition* (Polity 2024); Wall (2017), 'Crime, security and information communication technologies: The changing cybersecurity threat landscape and implications for regulation and policing', pp. 1075-1096.

[29] Wall (2024), *Cybercrime: The transformation of Crime in the Information Age, 2nd Edition*

So, for the purpose of this discussion hacking into a computer system and stealing the data contained within are cyber-dependent crimes in so far as the offences are dependent upon digital and networked technologies. If the technologies are removed from the crime, then it simply will not happen. This is in contrast, for example, to 'cyber-assisted crimes' which use computers and information systems to organise them, e.g., using internet communications to organise drugs deals. If the internet is removed, then the crime will still take place: the offenders will simply use other forms of communication to commit them. Or cyber-enabled crimes such as fraud and extortion which exploit the globalised 'cyber' element of networked and digital technologies. If the internet is removed from these cyber-enabled crimes, they will still take place at a more limited (probably localised) level.

While understanding the degree to which a crime has been transformed by technology is useful to identify issues relating to the scalability and globalisation of the offending, and the subsequent weaknesses in the procedural laws, it says little about the offence itself which can only be really understood in terms of its modus operandi. Here we must differentiate between crimes against the machine, crimes using the machine and crimes in the machine. This slightly different terminology reflects the legal approach towards the fight against cybercrime[30] as demonstrated by the 2001 Budapest Convention (Council of Europe Cybercrime Convention).[31] 'Crimes against the machine' attack the confidentiality, integrity and availability of computer data and systems. They include: illegal access by hacking; illegal interception (man in the middle attacks); data interference by infecting a machine with malware; system interference via DDoS; and Misuse of devices, which typically refers to creating and making available tools to commit the offences listed above.[32] Crimes that use the machine in the Convention,

[30] See Chs. 4,5,6 Wall (2024), *Cybercrime: The transformation of Crime in the Information Age, 2nd Edition* and also Wall (2017), 'Crime, security and information communication technologies: The changing cybersecurity threat landscape and implications for regulation and policing', pp. 1075-1096.
[31] Convention on Cybercrime, Council of Europe, CETS n. 105 23 November 2001.
[32] Respectively articles 2 to 6. Council of Europe, *Explanatory Memorandum to the Cybercrime Convention* (2001).

are forgery (falsifying electronic documents) and fraud[33] (scams) committed by using computer data and systems. Finally, crimes in the machine are offences where the computer contents are illegal[34]. In the Convention these are content-related offences such as child abuse imagery[35] (Article 9) and the racist and hate categories contained in the Convention's Additional Protocol.[36]

The Budapest Convention is a good starting point to demonstrate the link between Data Crime and Cybercrimes not only because it is the longest existing international legal instrument on the subject matter,[37] but also because it has acted as the model law for cybercrime legislation, at least for what concerns 'crimes against the machine', in several jurisdictions.[38] Pursuant to the Convention, cybercrimes are offences that concern 'computer data and systems' (Article 1). Following the Explanations to the Convention, a computer system is "a device consisting of hardware and software developed for automatic processing of digital data" (§ 23). In turn, the understanding

[33] Respectively articles 7 and 8 of the Convention.

[34] For a discussion of this crime, see Yaman Akdeniz, *Report. Freedom of Expression on the Internet. A study of legal provisions and practices related to freedom of expression, the free flow of information and media pluralism on the Internet in OSCE participating States* (2011).

[35] Colloquially but often wrongly referred to as 'child pornography'.

[36] Additional Protocol to the Convention on Cybercrime Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed through Computer Systems, Council of Europe, ETS n. 189. Title 4 of the convention addresses 'Offences related to infringements of copyright and related rights', which we do not address here, also due to the disagreement of the authors with copyright related offences as cybercrime.

[37] The Convention is no longer the only adopted international cybercrime instrument thanks to the African Union convention on Cyber Security and Personal Data Protection, which has recently entered into force https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection.

[38]For a list of the countries which have signed the Convention, see <https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185/signatures?p_auth=hTmKR7WR> accessed 15 November 2017). Among countries whose law has been influenced by the Convention there is the EU. The Directive on Attacks against information systems, which harmonizes substantive law across Member States to enable cross-border cooperation in the fight against 'computer-related crime' (Art. 87 TFEU, Consolidated versions of the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU), OJ C 83/01 (Lisbon Treaty).), contains clear connection clauses to the Convention. Directive 2013/40/EU of the European Parliament and the Council of 12 August 2013 on Attacks against Information Systems and Replacing Council Framework Decision 2005/222/JHA, OJL 218. In any case, the term 'cybercrime' does not have an autonomous legal significance, see Porcedda (2023), ch 7, building on Marise Cremona's 'A Triple Braid: Interactions between International Law, EU Law and Private Law' in Marise Cremona and Hans-W Micklitz (eds), *Private Law in the External Relation of the EU* (Oxford University Press 2016).

of data is taken from the ISO definition, as data put in such a form that it can be directly processed by the computer system (§ 25). Hence, cybercrimes tend to be offences which involve computer data or the systems where such data are processed. In other words, Data Crimes, especially data theft, are true (cyber-dependent) cybercrimes,[39] however, the stolen data can subsequently be used to commit cyber-enabled frauds, hence the cascade effect mentioned earlier.

Having established that Data Crimes are primarily cyber-dependent cybercrimes, we propose data crime be explored as Cybercrime, to differentiate between 'crime against the machine' (e.g., data breach), 'crime using the machine' (e.g., DDoS), and 'crime in the machine' (e.g., data that can be used against the owner). As for crime against confidentiality, integrity and availability, big data repositories can be hacked and the information contained therein copied, altered, further distributed etc., or repositories can be made unavailable through DDoS. Big data analytics can also be the target of forgery or be used to commit fraud: in this respect, they are the "IP" of a criminal activity. Finally, big data analytics can be used to support crimes of other kinds, for example, offenders joining databases together to build sophisticated profiles of individuals (credential stuffing) that increase both the chances of victimisation and also repeat victimisation.

While this differentiation holds the promise of reaching greater analytical depth, we further argue that it should not come at the expense of the analytical importance of information and the technology powering data science. It is only by accounting for the data, information, and technology that we can begin to untangle the epistemological difference between the threats, risks and harms entailed by data crimes as large cybercrimes. This is important because these qualities represent different epistemologies, yet they are regularly conflated, especially when it comes to

[39] In that they would disappear without internet technologies, Wall (2018) 'How Big Data Feeds Big Crime.

differentiating between what could and what does happen, causing some confusion.[40] In a nutshell, from a data perspective, risks and threats are anticipatory, in that they 'could happen' (not will happen), whereas harms and crimes represent different legal states of 'what has happened'.[41] Only when these differences are resolved, which requires looking at the data, the information, and the technology, can we achieve the granularity necessary to respond to such cybercrimes as data crimes effectively.[42] In fact, it is the technology which bears most inherent risks, because the likelihood of loss, the imminent threat, increases when the exclusivity of the data held in datasets increases, thus, further motivating the attacker. In turn, the harm becomes a function of the information contained in the data. These links are exemplified in the diagram below.

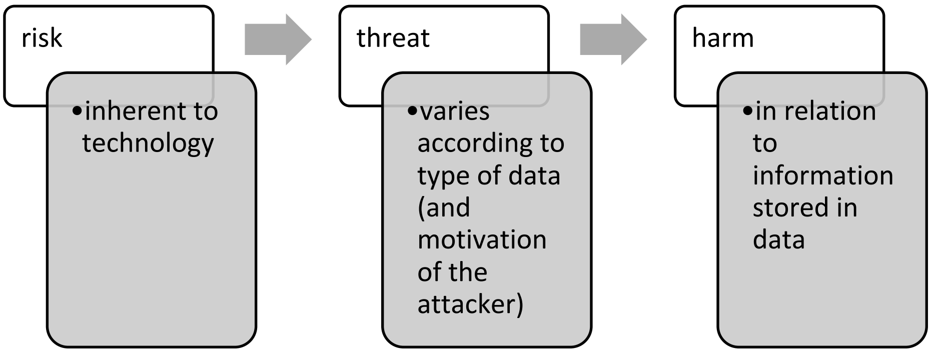

risk
threat
harm
•inherent to technology
•varies according to type of data (and motivation of the attacker)
•in relation to information stored in data

Figure 2: Risk, threat and harm in relation to technology, data and information

[40] David Wall, 'Crime, security and information communication technologies: The changing cybersecurity threat landscape and implications for regulation and policing' in Roger Brownsword, Elaine Scotford and Karen Yeung (eds), *The Oxford Handbook of the Law and Regulation of Technology* (Oxford Univeristy Press 2017), p. 1083. Risks, threats and harms are connected by the risk-based approach (risk management and assessment) which underpins information security and the protections of personal data. Nevertheless, these are conceptually different, as we will discuss later o On the point, see Raphael Gellert, 'Data Protection: a risk regulation? Between the risk management of everything and the precautionary alternative' 5 *International Data Privacy Law* 3-19.

[41] David Wall, 'Crime, security and information communication technologies: The changing cybersecurity threat landscape and implications for regulation and policing' in Roger Brownsword, Elaine Scotford and Karen Yeung (eds), *The Oxford Handbook of the Law and Regulation of Technology* (Oxford Univeristy Press 2017), p. 1083.

[42] Cybercrimes are crimes which are either enabled by or are wholly dependent upon the internet, see later discussion.

Accordingly, we will next explore the combination between the risks inherent to the technology and large data sets. Such an approach enables us to identify two mechanisms, the 'cyber lift' and the 'cascade effect', the upstream and downstream aspects which specifically define data crimes, as we discuss next.

3.1. Data science and the risks inherent in technology: The 'cyber lift' and the 'cascade effect'

Data science and data crime are made possible by cloud computing applications. This is due to two combined effects. Firstly, cloud technologies in all its three configurations of IaaS, PaaS and SaaS[43] are driving the development of the Internet as we know it today, as well as many of the services that enable the collection of big data. Secondly, cloud technologies are also providing storage and increasing processing capacity, necessary to analyse big datasets. However, due to the combination between 'networked' and 'digitised'[44] (the 'cyber lift'), cloud computing is also causing a further escalation in the scope of cybercrime which is explained below.[45]

Earlier we introduced Wall's 'transformation test', which shows how cybercrime is mediated by technology, and whereby the best way to ascertain what is and what is not a cybercrime is to (mentally) take away the internet from the crime being observed and think about what is left.[46] The test helps explain what is the 'cyber-difference', namely that cyber-assisted crimes still take place, cyber-enabled crimes lose the global, informational and distributed lift that is characteristic of 'cyber' and cyber-dependent crimes simply disappear.[47] Similarly, if you apply this principle to the cloud,

[43] Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). On cloud, see Michael Armbrust and others, *Above the Clouds: A Berkeley View of Cloud Computing.* (Technical Report No UCB/EECS-2009-28, 2009).

[44] Wall (2024), *Cybercrime: The transformation of Crime in the Information Age, 2nd Edition*

[45] David S Wall, 'Towards a Conceptualisation of Cloud (Cyber) Crime' (5th International Conference on Human Aspects of Information Security, Privacy and Trust, Vancouver, 9-14 July 2017).

[46] Wall (2024), *Cybercrime: The transformation of Crime in the Information Age, 2nd Edition*

[47] See discussion in Wall (2017), 'Towards a Conceptualisation of Cloud (Cyber) Crime'; Wall (2007), *Cybercrime: The transformation of Crime in the Information Age*; Michael Levi and others, *Technical Annex of The Implications of Economic Cybercrime for Policing* (2015).

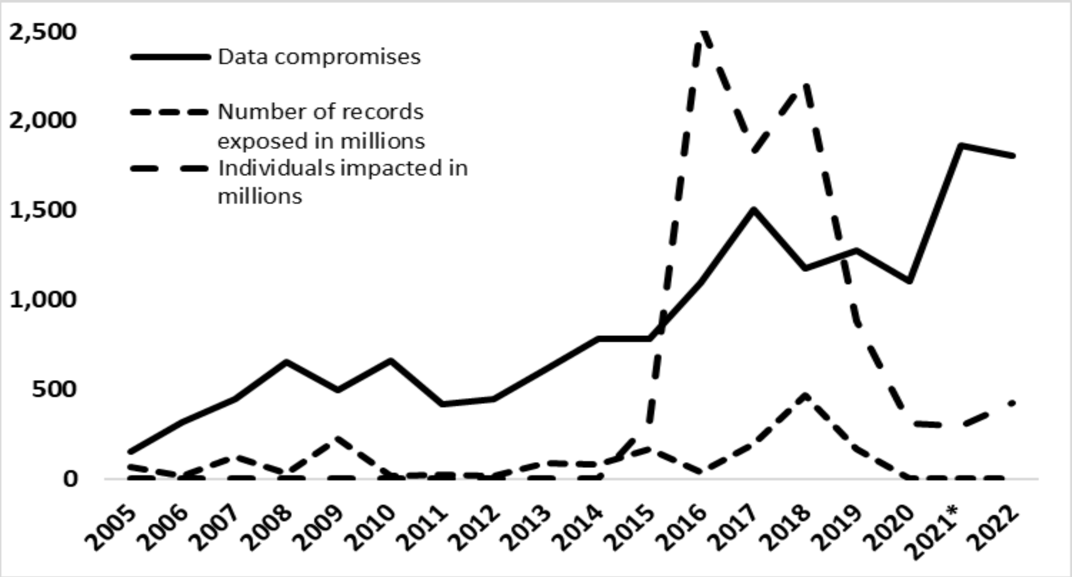
we can understand the 'lift' given by cloud technologies which increase the speed and volume of cybercrime and reduce the relative costs. In principle, because it is hard to demonstrate in practice, you could have cloud-assisted, cloud-enabled and cloud-dependent cybercrime.[48]

So, networked and digital technologies mean that criminals no longer needed to commit a high risk $50 million robbery when they could commit 50 million low risk $1 robberies using a networked computer.[49] The changes of scale that cloud technologies now bring to the table enable the same criminals to commit 50 billion robberies of, say, 0.1 cent, to achieve a greater yield and reduce the risk of prosecution even further.[50] In sum, cloud technologies not only increase the capacity of the internet in terms of volume, computing speed and reduced computing costs, but also increase the sheer volume of data flows when combined with the vast range of new forms of devices from the Internet of Things which cloud technologies also facilitate. Consequentially, attacks upon systems have become much more substantial in size and content than in the past, especially in terms of data breaches. The losses are increasing, and 'lost data' always leaves a question hanging as to whether the data will be reused for another purpose. This problem of 'sleeper fraud' has always been a potential threat,[51] but Big Data techniques employed by offenders are turning potential threats into probable threats by increasing their chances of achieving successful victimizations. The stolen data might, for example as outlined earlier, be joined to other data to increase its value, for example, basic email access data can be joined to membership personal data by a common email address, and then joined, say to other data by email or even social security number, and so on. By combining (Big) Data from a variety of sources, offenders can build up sophisticated and (non) personal sets of data about individuals or businesses, which can then be used at a later stage to

[48] Wall (2017), 'Towards a Conceptualisation of Cloud (Cyber) Crime'.
[49] Ibid.
[50] Ibid.
[51] Lexis-Nexis, *Waking up from the sleeper fraud nightmare, White Paper* (2016). Also see, David S. Wall, 'Micro-Frauds: Virtual Robberies, Stings and Scams in the Information Age' in Thomas J. Holt and Bernadette H. Schell (eds), *Corporate Hacking and Technology-Driven Crime: Social Dynamics and Implications* (IGI Global 2010).

victimise the owner. To give an idea of the problem and potential for data combination, Figure 3 illustrates the increase in the number of data breaches in the US between 2005 and 2022.[52] It shows not just an expansion over the years in the overall number of breaches, but also the increase in the number of records compromised and the numbers of individuals who are impacted by this loss.

Figure 3: Year on year increase in online data breaches 2005-2023



Source: United States; Identity Theft Resource Center; 2005 to 2022; data compromises include data breaches, data exposures, and data leaks; individuals impacted may go beyond the United States, statistics obtained from Statista[53]

Of interest is the relative drop since 2019 in the number of records compromised and individuals impacted. This drop may be due in part to advances in data security, for example, outsourcing data storage, especially with large databases, but this drop may also indicate that data thieves are focusing upon smaller and more obtainable databases. This is certainly the case with Ransomware attacks, where the data is stolen to encourage payment of the ransom. A recent study of Ransomware victims found

[52] Although some of the impact of the data theft will fall outside of the US boundaries.

[53] Permission to reproduce sought from Statista.

that most business and organisational victims tend to be small-medium enterprises of about 25-45 staff with turnovers of about $5m-$10m.[54]

The trend in Figure 3 indicates the breadth of data that is available for misuse. Moreover, although the overall number of affected individuals appears to have fallen, the trends suggest that those who are affected are more likely to be the victim of an actual crime with financial loss, damaged computer or business systems or even emotional damage amongst staff.

By focusing upon the data as a harm indicator, we can explore the way that offenders who prey upon the large datasets often bundle types of offending together. DDoS attacks have, for example, been used to distract system administrators while probing a network for vulnerability to gain entry via an SQL injection[55] to expose databases and give offenders access to the data.[56] Alternatively, spam attacks[57] could be used to place RAT (Remote Access Trojan) malware on a victim's computer to allow a third party (offender) remote access. But attacks upon, or disclosure of, one type of data often leads to several offence types, which we refer to as the cascade effect, where the previously mentioned 'upstream' (cyber-dependent) crimes cascade into (cyber-enabled) crimes 'downstream'. This effect is described below in Figure 4 and the method behind these 'tipping points' was explained in our 2021 paper.[58]

Figure 4: The cybercrime cascade effect

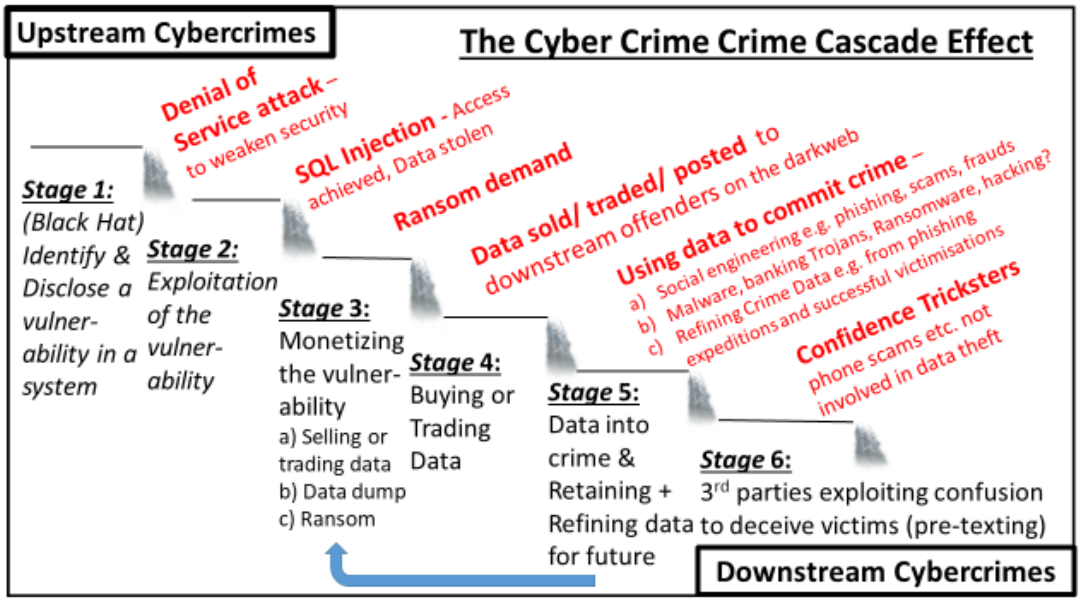
[54] David S. Wall, 'The Transnational Cybercrime Extortion Landscape and The Pandemic: Ransomware and changes in offender tactics, attack scalability and the organisation of offending', European Law Enforcement Research Bulletin, (SCE 5) (Oct 5, 2021) https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/475

[55] Data theft (hack) is the theft of bulk data by hackers who have, to date, tended to perform a DDoS attack as a decoy to confuse the computer security before breaching the system, e.g., via an injection of Structured Query Language statements exploiting a pre-existing security vulnerability to steal the data. See at <https://www.owasp.org/index.php/SQL_Injection> accessed 12 February 2018.

[56] Porcedda and Wall (2019) (n 4).

[57] Spamming is the distribution of unsolicited bulk emails. They choke up bandwidth and present risks to the recipient, should they respond. Mass Spam attacks (fueled by cloud technologies and internet of things botnets) intensify this effect.

[58] Porcedda and Wall, D.S. (2021) (n 4).


Upstream Cybercrimes
The Cyber Crime Crime Cascade Effect
Denial of Service attack – to weaken security
SQL Injection - Access achieved, Data stolen
Ransom demand
Data sold/ traded/ posted to downstream offenders on the darkweb
Using data to commit crime –
a) Social engineering e.g. phishing, scams, frauds
b) Malware, banking Trojans, Ransomware, hacking?
c) Refining Crime Data e.g. from phishing expeditions and successful victimisations
Confidence Tricksters phone scams etc. not involved in data theft
Stage 1: (Black Hat) Identify & Disclose a vulner-ability in a system
Stage 2: Exploitation of the vulner-ability
Stage 3: Monetizing the vulner-ability
a) Selling or trading data
b) Data dump
c) Ransom
Stage 4: Buying or Trading Data
Stage 5: Data into crime & Retaining + Refining data for future
Stage 6: 3rd parties exploiting confusion to deceive victims (pre-texting)
Downstream Cybercrimes

One of the interesting issues regarding these larger offences is that there appears to be some evidence, as mentioned earlier, that offenders are using data analytics and artificial intelligence processes to join various stolen databases together and process them. Datasets of potential victims profiled, for example, by a common characteristic, such as a profession, can enable offenders to focus their attacks and increase access to victims, or even directly profile victim's credentials. These refined data sets effectively become the 'intellectual property' of the offenders to commit further crime[59] and this is a fact that is not lost on offenders.

3.2. Ransomware as data crime and the AI-turn[60]

A major step change in the cybercrime threat landscape was the shift towards data as the focus of cybercrime. This was especially the case after the introduction of crypto-

[59] Please note, this use of the term intellectual property is nominal as it would be hard for offenders to exercise their intellectual property rights in most (if not all) courts under existing laws. See for more on the IP rights protection of data also Chapter 9 of this volume.

[60] The phrase AI-turn is not intuitive, but indicates a debate how AI is beginning to enter the governance agenda. See further, Maria Sapignoli, Anthropology and the AI-Turn in Global Governance, *AJIL Unbound* 115: 294-298 (2021).

ransomware[61], which encrypted all the victims' data rather than locking access to the computer.[62] Crypto-ransomware deprived victims of their data until a ransom was paid for a decryptor to release it. The next step change was the added threat of data extortion whereby the attackers exfiltrated key data before beginning the encryption process. This data theft not only increased the leverage on the victims to pay the ransom, sometimes asking two ransoms, one for the decryptor and a further ransom to return the stolen data and delete copies.[63] The next step was for ransomware groups to move towards data extortion and away from encryption. The argument being that decryption was never successful in returning all the files and that data extortion was a better criminal business model. These changes in the threat landscape made offenders not only realise how much data could be taken, but also its overall value when processed and sold onwards.

To give an idea of the scale of data 'theft', it was estimated that the 2022 Medibank hack resulted in 9.7m medical records of 36% of Australian citizens being stolen and subsequently made available for sale on darkmarkets.[64] Data theft has also become a weapon in the recent Russia/ Ukraine conflict.[65] Ukrainian hackers have claimed to have stolen the personal details of 1600 Russian troops who served in the City of Bucha, which was devastated during the conflict and also the scene of war crimes. They also claim to have the details of 620 Russian spies registered with the FSB.[66]

[61] See further Lena Connolly and David S. Wall, 'The Rise of Crypto-Ransomware in a Changing Cybercrime Landscape: Taxonomising Countermeasures, *Computers and Security*, 87(10 July 2019) https://doi.org/10.1016/j.cose.2019.101568

[62] Ibid.

[63] Wall, (2021) (n 56)

[64] Michael Slezak and Marty Smiley, 'Medibank, Optus, Woolworths data hacks show how a 'decade of anti-security policy' is putting Australia at risk, experts say', *ABC News*, (20 October 2022) https://www.abc.net.au/news/2022-10-21/medibank-optus-data-hack/101558932.
Briana Morris-Grant, 'Hackers have released stolen Medibank data on the dark web. What does this mean for customers?', *ABC News*, (9 November 2022), https://www.abc.net.au/news/2022-11-10/medibank-data-breach-latest-dark-web-leak/101632746

[65] Matt Burgess, 'Russia Is Leaking Data Like a Sieve', *WIRED*, (13 April, 2022), https://www.wired.com/story/russia-ukraine-data/

[66] Ibid

Not only has the scale of data theft and extortion become massive, but the attackers are using Artificial Intelligence (AI) based tools to help improve their ability to victimize.[67] Attackers seemingly use AI-based tools to test and improve their own malware, to infect their victim's AI systems with inaccurate data in their favour and to map out their victim's existing AI models, including security, with a view to countering their actions.[68] They also link AI processes to identify potential victims in terms of their business size, sector and vulnerability. Until recently, AI was the preserve of skilled computer scientists, but publicly available programmes such as ChatGPT (Chat Generative Pre-trained Transformer) are making AI much more accessible. ChatGPT is a large language model (LLM) of AI. Released in November 2022, ChatGPT is a 'dialogue-based AI chatbot capable of understanding natural human language and generating impressively detailed human-like written text'.[69]

Such is the level of concern about ChatGPT that it has already been made the subject of a EUROPOL warning about its use for criminal purposes.[70] The main concern is that its ability to process, manipulate, and generate text can lead to its use in 'Fraud and social engineering, especially for phishing purposes'.[71] More particularly, its ability to re-produce language patterns can be used to impersonate the style of speech of specific individuals or groups and 'mislead potential victims into placing their trust in the hands of criminal actors'. EUROPOL have warned offenders that its ability to create and spread messages reflecting a specific narrative with relatively little effort can also be used to circulate disinformation. Finally, they warn that ChatGPT's ability to

[67] Dave Shackleford, 'How hackers use AI and machine learning to target enterprises', *TechTarget*, (October, 2019), https://www.techtarget.com/searchsecurity/tip/How-hackers-use-AI-and-machine-learning-to-target-enterprises

[68] Ibid.

[69] Samantha Lock, 'What is AI chatbot phenomenon ChatGPT and could it replace humans?', *The Guardian*, (5 December, 2022), https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans

[70] EUROPOL The criminal use of ChatGPT – a cautionary tale about large language models, *EUROPOL Press Release*, (27 March 2023), https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models

[71] Ibid

produce code in different programming languages can provide potential offenders who have little technical knowledge with a resource to produce malicious code.[72]

These AI processes not only introduce unpredictability into crime data but are potentially leading to an avalanche of 'downstream' cyber-enabled cybercrimes. Moreover, Geoffrey Hinton, the so-called godfather of artificial intelligence, said upon his retirement from Google that AI processes are currently 'not more intelligent than us, as far as I can tell. But I think they soon may be'.[73] Hinton also said that 'international competition would mean that a pause would be difficult. "Even if everybody in the US stopped developing it, China would just get a big lead," he said.

Whether or not the discussion is at the level of world power or individual cybercrime, Hinton alludes that to not use AI is to lose pace on the competitors. On this, of course, one point that must not be lost is that AI programmes like ChatGPT are marketed as a powerful tool to enable law enforcers to identify attackers and counter their activities.[74] In this respect, data science may actually be part of the solution as long as its application not only keeps responsible agencies ahead of the offenders but also meets stringent safeguards.

4. Data Law: the big data economy and the use of analytics in the fight against data crime

One of the underlining questions of this edited book is what can data science do for the law, and vice versa? The fact that data science and law could be mutually beneficial, including helping law enforcement fighting data crimes and crimes at large, depends on the law being capable of channelling data science, and doing so in the right

[72] Sead Fadilpašić, 'Hackers are using ChatGPT to write malware', *techradar*, (9 January 2023), https://www.techradar.com/news/hackers-are-using-chatgpt-to-write-malware

[73] Zoe Kleinman and Chris Vallance, 'AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google', *BBC News Online*, (2 May, 2023), https://www.bbc.com/news/world-us-canada-65452940

[74] More on the topic of the use of data science techniques by law enforcement in general, not specified to data crimes, see Chapter 12.

direction. If this means hindering developments that may harm society at large, then it also means to foster the ones that are more welcome.

We begin by reflecting upon the use of personal and non-personal data science to support the fight against crime and then we look specifically into data crimes. We reflect on the predictive value of big data analytics in the wider context of criminal investigations and how this can be lawfully applied to (big) Data Crime. We then discuss developments in data law and reflect on the nexus between the creation of the data economy and the surge of data crime.

4.1. **Policing crime using (big) data science**

The potential for data science to support the fight against crime through predictive policing is widely discussed. Most of the hopes around big data analytics for security purposes seem to revolve around personal data. Chan and Moses drew upon an empirical study of Australian police to note that law enforcement seem to be more interested in the opportunity of using large personal data sets that would help identify the suspects of a criminal investigation.[75] Along similar lines, Ferguson listed the various experimental programs that combine large data sets of personal data with data analytics capabilities, e.g. to obtain real-time identification of individuals in the street or to exculpate suspects, thus reducing false positives.[76] Ferguson proposes that exculpatory facts derived from large data sets could be included in the reasonable suspicion analysis, for example, as a self-contained check on the regular discretionary powers granted by the police.[77] Failure to do so may seriously invalidate the prosecution process.

[75] Chan and Bennett Moses (2017), 'Making Sense of Big Data for Security'.
[76] Ferguson (2015), 'Big Data and Predictive Reasonable Suspicion'.
[77] Ferguson (2015), 'Big Data and Predictive Reasonable Suspicion', p. 309 and 345.

The creation of large data sets containing personal information for policing purposes, including those stemming from public-private partnerships[78] precedes the data science hype and is the object of a well-developed body of research across disciplines. These bodies of scholarship point to the various assumptions, misassumptions and claims made about the predictive value of analytic tools,[79] and highlight the dangers hiding behind the seducing proposition of finding the needle in the haystack, and especially so for the rights to the protection of personal data and the protection of private and family life (anticipated in section 2).[80] Big data analytics have spurred debates around explanability, the black box problem and 'FAT', which are discussed elsewhere in this volume. In the EU, for example, several publicly funded projects[81] have been tackling the issue of how to use big data science in a legitimate manner, particularly in the light of the adoption of the EU Directive 2016/680,[82] which regulates the protection of

[78] This includes data retention by Telcos, but also databases created and hosted by private parties on demand of the government.

[79]Janet Chan and Lyria Bennett Moses, 'Using big data for legal and law enforcement decisions: testing the new tools' 37 *UNSW Law Journal*; Janet Chan and Lyria Bennett Moses, 'Is Big Data challenging criminology?' 20 *Theoretical Criminology* 21-39; Adam Edwards, 'Big Data, predictive machines and security. The minority report' in M. R. McGuire and Thomas J Holt (eds), *The Routledge handbook of Technology, Crime and Justice* (Routledge 2017); Guthrie Ferguson, 'Big Data and Predictive Reasonable Suspicion' 163 *University of Pennsylvania Law Review* 327-410; Elizabeth Groff and Dan Birks, 'Simulating Crime Prevention Strategies: A Look at the Possibilities' 2 *Policing* 175-184; Matthew L. Williams, Pete Burnap and Luke Sloan, 'Crime Sensing with Big Data: the affordances and limitations of using open-source communications to estimate crime patterns' 57 *British Journal of Criminology* 320-340; Carrie B. Sanders and James Sheptycki, 'Policing, crime and 'big data'; towards a critique of the moral economy of stochastic governance' 68 *Crime, Law and Social Change* 1-15; Richard Berk and Justin Bleich, 'Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment' 12 *Criminology & Public Policy* 513-544; Janet Chan and Lyria Bennett Moses, 'Making Sense of Big Data for Security' 57 *British Journal of Criminology* 299-319.

[80] This is a longstanding debate, see among many Stefano Rodotà, *Elaboratori Elettronici e Controllo Sociale* (Mulino 1973);Frank Dumortier and others, 'La Protection des Données dans l'Espace Européen de Liberté, de Sécurité et de Justice' 166 Journal de Droit Européen 23; Franziska Bohem, *Information sharing and data protection in the Area of Freedom, Security and Justice – Towards harmonised data protection principles for EU-internal information exchange* (Springer 2012). For an attempt to reconcile big data and security objectives, see Cavoukian and Jonas (2012), *Privacy by Design in the Age of Big Data*.

[81] With different degrees of acceptance, e.g., TRESSPASS (https://cordis.europa.eu/project/id/787120); more recently, see VIGILANT (https://cordis.europa.eu/project/id/101073921).

[82] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of the Prevention, Investigation, Detection or Prosecution of Criminal Offences or the Execution of Criminal Penalties, and on the Free Movement of such Data, and Repealing Council Framework Decision 2008/977/JHA [2016] OJ L 119/89. For guidance (not endorsed by the EDPB), see Article 29 Working Party, Opinion on some key issues of the Law Enforcement Directive (EU 2016/680), wp258 (2017);Eleni Kosta and Franziska Bohem (eds), *The EU Law Enforcement Directive (LED*) (Oxford University Press, forthcoming).

personal data in the course of law enforcement activities. Art. 11 on 'automated individual decision-making' (Art. 11) can be read as authorising law enforcement agencies to use (big) data science for making decisions, provided such use is authorised by national law, which foresees appropriate safeguards, at a minimum the right to obtain human intervention on the part of the controller (as further specified by Recital 38). Algorithms, after Articles 11(2) and (3), can also process special categories of data, such as ethnic origins and political beliefs, subject to suitable safeguards, and can perform 'profiling', so long as it does not result in discrimination against natural persons. Personal data processed by analytics for policing purposes are exposed to the same risks as it would be the case for, say, marketing purposes under civil law, with the obvious difference that the consequences for the data subjects can be devastating and life changing, and range from being imprisoned to committing suicide.[83]

Whilst a discussion of the pros and cons of the Directive is beyond the scope of this article, its adoption has nonetheless spurred a debate about the viability of algorithmic decision-making in policing (and further afield) which applies beyond the EU rules. What is interesting is that the downsides of algorithmic policing that rely on personal data largely overlap with the downsides of analytics based on non-personal data, which we discuss next.

In general, non-personal (big) data science can be used to improve analytical capacity and understand broader trends that support (predictive) policing. One such example is by providing the information to build simulations for analysis *in silico*, as discussed by Groff and Birks.[84] Also, big data could lead to the development of a quantitative standard (in terms of statistical likelihood) to validate reasonable suspicion, which could be adapted to the seriousness of crime.[85] Moreover, non-personal big data could more easily pass the legal tests of permissibility, and be ethically acceptable. This does not, however, mean that predictive policing products based on non-personal data,

[83] Jess Bidgood, 'Body of Missing Student at Brown Is Discovered', New York Times (25 April 2013).
[84] Groff and Birks (2008), 'Simulating Crime Prevention Strategies: A Look at the Possibilities'.
[85] As suggested for instance by Ferguson (2015), 'Big Data and Predictive Reasonable Suspicion', p. 406.

such as PREDPOL, and its variations PROMAP and PRISM[86], are free from criticism[87]. Yes, they can help police plan their police coverage of an area and even highlight the potential for conducting operations. They can also assist as an anticipatory tool, but only a tool and nothing more. All they can do is 'anticipate' as they cannot really 'predict' in the strict sense of the word. They cannot foretell the future because it has not happened yet and neither can they identify predatory individuals 'beyond reasonable doubt', or even on the 'balance of probability'.

The problems, which, as anticipated earlier, concern both analytics processing non personal as well as personal data sets, arise with the mission creep of the predictive claims. Whatever the robustness of predictive data, what they fail to do is enable the predictors to say specifically who will be the offenders or the victims. Williams et al.[88] clearly state that big (social) data should be used in combination with sources that carry greater validity. Yet, some vendors of big data analysis systems are either making specific claims about the ability to predict specific types of offenders, or they are giving the illusion that the products will do so.[89] Not only would such tools fail to prove criminal intent against specific individuals, let alone any notion of conclusive evidence of wrongdoing, but they can also strengthen anti-social stereotypes towards specific groups in society and potentially interfere with the due process of law. It is interesting to note that only one officer interviewed by Chan and Moses raised the fact that big data analytics may lead to discrimination by strengthening stereotypes. Moreover,

[86] Edwards (2017), p. 453.

[87] Ferguson (2015), 'Big Data and Predictive Reasonable Suspicion', p. 394.

[88] Williams, Burnap and Sloan (2017), 'Crime Sensing with Big Data: the affordances and limitations of using open-source communications to estimate crime patterns'.

[89] A search for the words - Identifying offenders through BIG Data analytics - indicates a range of discussion about the claims made of data analytics. The main claims are that techniques applied in other areas (e.g., health and medical science) could be transferable to the study of crime. Another strand of debate relates to the alleged ability of data science to re-identify data that has been actively de-identified. See, for example, arguments made in blogs such as Kwapien, A. (2016) 'How Big Data Helps To Fight Crime?', and especially the discussion in Cavoukian, A. and Castro, D. (2014) Big Data and Innovation, Setting the Record Straight: De-identification Does Work. The key issue in the discussions is one high levels numbers of false positives that give the outputs of the techniques a lack of certainty to act upon.

data may be inaccurate,[90] yet not audited due to lack for provision of oversight, and lead to an intolerable number of false positives.[91] Ferguson for example writes that the FBI files contain, reportedly, hundreds of thousands of errors.[92] Thus, unless the machine learning data input is checked, data science could possibly worsen the very problems it is seeking to resolve. We will come back to this in the last part of this section.

4.2. **Data science supporting the fight against data crime**

At present, the process of identifying data crimes relies mainly upon the data holder reporting a breach of security, which does not always happen, and often only when the holder has to do so[93], or when the data has been identified as causing downstream crimes, say, through attempts to monetize it via frauds, accounts takeover or even extortion arising from the stolen data.

As mentioned in the previous section, personal and non-personal data analytics can assist in the fight against any crimes. But this time, by using machine learning and artificial intelligence-based systems, one can observe key characteristics and combinations of these forms of offending to identify the crimes and their knock-on effects when they take place.[94] As stated earlier, at present they are only detected when the data holder is aware of the breach, or when the stolen data is identified downstream, say, through frauds, accounts taken over or even extortion arising from the stolen data. The intention of computer science input is to identify algorithms applicable to the anonymized data via Machine Learning and Artificial Intelligence to identify security breaches when they occur, assist practitioners to both identify

[90] Ferguson (2015), 'Big Data and Predictive Reasonable Suspicion', pp. 389 and 398. This is a well-known problem in data protection circles. Data quality/accuracy is one of the transversal principles of data protection.
[91] See for more on these issues also Chapter 15 regarding methods in this volume.
[92] Ibid, p. 399.
[93] See further Porcedda (2023), ch 6.
[94] This was one of the ambitions of the CRITiCaL project and RAMSES project <https://ramses2020.eu/>.

patterns of offending, and possibly the path of offending which could lead law enforcement to the offenders.

Such an approach, for which there is proof of concept, is itself not devoid of criticism, as discussed earlier in relation to data science and policing. Of course, if successful, then it subsequently raises several questions about admissibility of evidence in court and the direction of the burden of proof. The prime intention is to realistically scope out the possibilities within the limitations of predictive modelling, and at least identify various factors and help to reduce the number of false positives. Such activities understandably raise some instinctive ethical concerns, as well as legal ones, chiefly in the form of the need to identify a suitable legal basis to justify such machine assistance, which, as we argue in the next section, should concern both personal and non-personal data.

In addition to help studying the technical characteristics of cybercrime, data science could be very helpful to assist in the elaboration of improved data crime scripts. Large data sets could first lead to highlight trends, thereby informing the creations of simulations *in silico*.[95] This would help focussing the scarce resources of police officers dedicated to pursuing cybercrime. Big data analytics could also help formulating a model to assess the harms caused by data crime, e.g., in combination with the types of data affected. Data sets held by different bodies could be put together to come up with metrics to assess crime. Recently, it was proposed that the use of Machine learning and artificial intelligence could be put to use to study cybercrime, also for private enforcement, beyond sectors such as banking where these methods are widely used.[96] This was the case of an e-ads company that partnered with a data science company “to identify behavioral patterns among its users, in order to find those

[95] Groff and Birks (2008), ‘Simulating Crime Prevention Strategies: A Look at the Possibilities’

[96] For an overview of the techniques used to detect credit card fraud, see for instance S. Benson Edwin Raj and A. Annie Portia, "Analysis on credit card fraud detection methods," *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*, Tamilnadu, 2011, pp. 152-156. See also in this regard PayPal as mentioned in Chapter 14 section 5.1.

attempting to abuse its systems, and to encourage compliance with abuse policies".[97] In the first quarter following the adoption of big data analytics, the company reported a reduction of frauds of 25%, and an increase in transactions of 20%. The use of machine learning and artificial intelligence to autonomously detect data crimes is still not fully developed,[98] and should be watched closely operationally, ethically, and legally, to avert an avalanche of discriminatory automated individual decisions.

More recent developments in data science have been to secure the enabling environment by creating cloud enclaves to address data crime and facilitate data science. Cloud enclaves have limited input and output abilities which means that transactions cannot be independently observed, especially by offenders[99]. Of course, as identified earlier, the advantages of AI benefit both sides as cloud enclaves can also protect offending activities. Finally, AI can be used to identify and investigate the existence of child sex abuse images, although this practice simultaneously raises privacy and data protection concerns. Also, a 2023 BBC report found that paedophiles were using (AI) technology themselves to generate life-like child sexual abuse material[100] which they sell. Further emphasizing the need for law enforcement to keep ahead of the offenders.

4.3. **Data law and the data economy-data crime nexus**

Here we critically engage with legislative developments that have taken place since the adoption of the first EU Communication addressing the data economy in 2017. The EU has adopted or is in the process of adopting seven instruments that create the

[97] Thomas Claburn, 'Smyte might brighten fraud plight: How machine-learning can be used to thwart crooks' *The Register* (17 August 2017).ra

[98] Katyanna Quach, 'In the red corner: Malware-breeding AI. And in the blue corner: The AI trying to stop it' *The Register* (2 August 2017); Ian Thomson, 'AI quickly cooks malware that AV software can't spot' *The Register* (2 July 2017).

[99] On Secure enclaves or trusted execution environments (TEEs), see Jatinder Singh, Jennifer Cobbe, Do Le Quoc, Zahra Tarkhani, 'Enclaves in the Clouds' (2021) *Communications of the ACM*, 64 (5), 42-51 https://cacm.acm.org/magazines/2021/5/252176-enclaves-in-the-clouds/fulltext.

[100] Crawford, A. and Smith, T. (2023) 'Illegal trade in AI child sex abuse images exposed', *BBC News Online*, 28 June, https://www.bbc.com/news/uk-65932372

conditions for the development of the data economy and address some known shortcomings of the technological environment in which such data economy is supposed to flourish, including the much-awaited AI Act.

The first instrument aims to ensure the free flow of non-personal data, such as ‘aggregate and anonymised datasets used for big data analytics’, ‘data on precision farming’ ‘or data on maintenance needs for industrial machines’[101] It lays down rules that prohibit data localisation and encourages the establishment of codes of conduct to enable the porting of data for professional users.[102]

The recast of the Public Sector Information Directive (hereafter PSI Directive)[103] aims to promote the use of open data and stimulate innovation in products and services through minimum rules governing the re-use and the practical arrangements for facilitating the re-use of existing documents held by public sector bodies of the Member States and specific public undertakings.[104] The framework encourages the creation of ‘high-value datasets’, that is ‘documents the re-use of which is associated with important benefits for society, the environment and the economy’, based on their potential to: ‘(a) generate significant socioeconomic or environmental benefits and innovative services; (b) benefit a high number of users, in particular SMEs; (c) assist in generating revenues; and (d) be combined with other datasets’.[105] High-value sectors are the geospatial, earth observation, environment, meteorological, statistics, companies, company ownership and mobility.[106] The PSI Directive excludes the creation of data sets based on data which is ‘not accessible due to commercial and statistical confidentiality and data that is included in works or other subject matter

[101] Regulation (EU) 2018/1807 Of The European Parliament And Of The Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union [2018] OJ L303/59, Rec 9.
[102] Arts 1, 4 and 6, Reg 2018/1807.
[103] Directive (Eu) 2019/1024 Of The European Parliament And Of The Council of 20 June 2019 on open data and the re-use of public sector information [2019] OJ L 172/56 (hereafter PSI Dir).
[104] Defined in Art 1(b)(i)-(iv), PSI Dir.
[105] Arts 2(10) and 14, PSI Dir.
[106] Annex, PSI Dir.

over which third parties have intellectual property rights', which fall within the remit of the Data Governance Act instead (hereafter DGA).[107]

The DGA 'lays down conditions for the re-use of data held by public sector bodies which are protected on grounds of '(a) commercial confidentiality, including business, professional and company secrets; (b) statistical confidentiality; (c) the protection of intellectual property rights of third parties; or (d) the protection of personal data, insofar as such data fall outside the scope of Directive (EU) 2019/1024.'[108] The DGA further legislates frameworks for the notification and supervision of providers of data intermediation services, for voluntary registration of entities which collect and process data made available for altruistic purposes, and for the establishment of a European Data Innovation Board.'[109] The DGA is accompanied by the Data Act, which creates conditions for the development of the data economy in the private sector and for which a political agreement was reached in June 2023.[110]

Two additional instruments, the Digital Markets Act and Digital Services Act, lay down obligations for those intermediation services that are making the data economy technically possible and that enjoy a dominant position in the digital ecosystem. The Digital Markets Act (hereafter DMA) harmonises rules ensuring contestable and fair markets to the benefit of business users and end users in the digital sector across the Union where gatekeepers are present.[111] A digital service is a gatekeeper when it has a significant impact on the internal market, it provides a core platform service which is

[107] Regulation (Eu) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) [2022] OJ L152/1, Rec 10 (hereafter DGA).

[108] Arts 1(a) and 3(1), DGA. Exclusions are covered by Art 3(2), DGA.

[109] Art 1(b)-(d), DGA.

[110] European Commission, Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), COM (2022) 68, see updates at https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=COM:2022:68:FIN andhttps://ec.europa.eu/commission/presscorner/detail/en/ip_23_3491.

[111] Regulation (Eu) 2022/1925 of the European Parliament And Of The Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L 265/1 (DMA), Art 1.

an important gateway for business users to reach end users and (foreseeably) enjoys an entrenched and durable position, in its operations.[112] A core platform service includes (a) online intermediation services; (b) online search engines; (c) online social networking services; (d) video-sharing platform services; (e) number-independent interpersonal communications services; (f) operating systems; (g) web browsers; (h) virtual assistants; (i) cloud computing services; and (j) online advertising services.[113]

The Digital Services Act, which amends the twenty-three year old e-Commerce Directive, lays down the conditional exemption from liability of providers of intermediary services engaged in mere conduit, caching and hosting, as well as due diligence obligations tailored to certain specific categories of providers of intermediary services.[114] The DSA aims to create a framework to manage risks associated with the dissemination of illegal content, broadly defined as 'any information that is not in compliance with Union law or the law of any Member State' and to police such contents. Examples of illegal contents include the dissemination of child sexual abuse material or illegal hate speech or other types of misuse of their services for criminal offences, and the conduct of illegal activities, such as the sale of products or services prohibited by Union or national law, including dangerous or counterfeit products, or illegally-traded animals.[115] The DSA also addresses the actual or foreseeable impact of services on the exercise of fundamental rights, on democratic processes, civic discourse, electoral processes, public security, as well as negative effects on the protection of public health, minors and serious negative consequences to a person's physical and mental well-being, or gender-based violence.[116]

[112] Art 3(1), DMA.
[113] Art 2(2), DMA.
[114] regulation (eu) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1 (DSA), Arts 1 and 3(g).
[115] Art 3(h), DSA and Rec 80.
[116] Recs 81-83, DSA.

These instruments try to create a favourable environment for big data innovation while also protecting fundamental rights, which is a difficult balancing act. In placing such obligations on economic actors, the legislator was cognisant of the potential damage that may derive from the unwitting disclosure of confidential business data, e.g., trade secrets or information protected by intellectual property. Similarly, the PSI and DMA acknowledge the need to adequately protect personal data, for instance by making its reuse conditional on licenses or requirements such as anonymisation and access through secure processing environments.[117] The DGA appreciates the risks of de-anonymisation and subsequent re-identification of individuals and suggests prohibiting re-identification from anonymised data sets.[118] Among the illegal activity tackled by the DSA is the sale of products or services prohibited by EU or Member State Law, within which could fall the proceeds of data crime.

The architecture created by the interaction of these instruments, however, is silent about data crime and ransomware trends, and begs the question whether the protective measures they identify can work in practice. First, the DGA mentions 'state-of-the-art privacy-preserving methods that could contribute to a more privacy-friendly processing of data', such as anonymisation, differential privacy, generalisation, suppression and randomisation, the use of synthetic data or similar method, and acknowledges that public sector bodies will need help to 'make optimal use of such techniques'.[119] Privacy-preserving techniques are high on the policy agenda, as shown by reports issued by the OECD, the United Nations[120] and other organisations, but there is an outstanding challenge in operationalising such techniques.

[117] See for instance Art Rec 44, Art 8 PSI Dir, and Rec 15, Arts 2(20) and 5 DGA.
[118] Recs 8, 15, DGA.
[119] Rec 7, DGA.
[120] E.g. Organisation for Economic Cooperation and Development, Good Practice Principles for Public Service Design and Delivery in the Digital Age (2022) https://www.oecd.org/publications/oecd-good-practice-principles-for-public-service-design-and-delivery-in-the-digital-age-2ade500b-en.htm; United Nations. *The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics* (2023).

There is no catalogue of 'state of the art measures', which is a market-driven process where the biggest market players, including the gatekeepers and VLOPs addressed by the DMA and DSA can exert great influence.[121] The effacement of technology from information technology law creates a state of technological indeterminacy whereby it is those who handle data that are responsible for identifying the technological solutions that best meet their resources.[122] The creation of the data economy precedes the identification of robust measures of de-identification of personal data.

For data science to power the data economy without creating data crime and ransomware negative externalities, there is an urgent need to provide more than an encouragement to use privacy-preserving techniques but to test them and actively favour those that are going to prevent routine de-anonymisation and re-identification. The same applies to measures for securing any data. Both would require addressing the concrete question of how to help small entities shouldering the costs of taking up such measures – as they are seemingly the entities most targeted by ransomware attacks (section 3 above).

These raise the question of whether data economy instruments may unwittingly play in favour of the expansion of the attack surface for data crime offenders, including ransomware. The data economy instruments reviewed here mention neither the existing cybercrime legal landscape, nor acknowledge that the cybersecurity framework is only partly operational.[123] Such instruments overlook the underlying objective of cybersecurity, for example, resilience. Attacks are presumed, yet offenders are rarely caught and when they are, criminal law instruments are unable to offer relief to affected individuals. This places great emphasis on pre-emptive and preventive instruments adopted under the aegis of the EU.

[121] Maria Grazia Porcedda, *Cybersecurity, Privacy and Data Protection in EU Law. A Law, Policy and Technology Analysis* (Hart Publishing 2023), ch 5.
[122] Maria Grazia Porcedda, *Cybersecurity, Privacy and Data Protection in EU Law. A Law, Policy and Technology Analysis* (Hart Publishing 2023), ch 5.
[123] Ibid, ch 6.

Under the country-of-origin principle, enforcement of EU rules is typically left to national administrative law mechanisms and relief for affected individuals to national private law regimes governing liability, both under the interpretive guidance of the Court of Justice of the European Union (hereafter CJEU). Indeed, the data protection and cybersecurity frameworks currently lack a methodology for identifying harms derived from the abuse of data and information systems. The matrix of potential harms of data crime we proposed in 2018 has yet to be matched by official guidance.[124]

	Raw data	Anonymise	Technical	IP	Trade	Personal	Sensitive
Material damage							
Fraud					x	x	x
Financial loss	x	x	x	x	x	x	x
Other economic disadvantage		x	x	x	x	x	x
Security and continuity of services[125]	x		x		x		
Health							
Physical						x	x
Mental						x	x
Non-material damage							

[124] If an unauthorised reversal of anonymisation/ pseudonymisation, then also natural persons. The table highlights that the misuse of personal data, which contain information relevant for the private and family life of individuals and also their participation in society, is a threat common to most risks. In other words, personal data bear the likelihood of causing the greatest harms, though the law is silent on this point. It also shows that financial loss is likely to occur for all types of data. The table outlined above is silent about the magnitude of the harms entailed by data crimes. Part of the problem here rests in the absence of a suitable methodology. Data Theft exists in a sort of legal netherworld as it is not illegal in all jurisdictions to hold unauthorised data; only the methods by which it is appropriated and subsequently used are always illegal.

Alteration, theft or misuse of identity					x	x	x
Damage to reputation	x			x	x	x	x
Other significant social disadvantage			x	x	x	x	x
Rights							
Loss of control over personal/own data	x					x	x
Limitation of rights;	x			x	x	x	x
Unauthorised reversal of pseudonymisation and anonymisation						x	x
Loss of confidentiality of data protected by professional secrecy					x	x	x
Discrimination	x	x				x	x
Harmed categories							
		Organizations*		Organizations	Organizations	Natural	Natural

The CJEU started interpreting provisions in the context of personal data breaches and cybersecurity incidents only in 2023.[126] More cases are pending before the Court, and it may be sometime before a common approach is found. Identifying a workable matrix of harms is as crucial a step as it is a spinous one, due to the country-of-origin principle and the number of stakeholders that should be involved in such an exercise.

What is clear is that data analytics, the data economy and data crime feed into a cycle that must be looked at together. If resilience is the motto of our digital economies, then data crime is going to happen, meaning that it is shortsighted to leave the

[126] The CJEU started interpreting provisions in the context of personal data breaches and cybersecurity incidents only in 2023, starting with Case C-300/2, *Österreichische Post*, ECLI:EU:C:2023:370.

identification of risks, the minimisation of harms and redress for damages ex post-facto. We need to set up a framework for a holistic and interdisciplinary approach that integrates the prevention-response nexus and addresses any externalities that have been created along the road. To achieve this, we reiterate the importance of standardisation of data collection with regards to data crimes as we discussed at first edition.[127]

5. Conclusion

In this chapter we have explored the use of data science techniques with regard to data crimes. We discussed the strengths and weaknesses of the concept of Data Science, or Big Data analytics and illustrated that that there has been a shift from information ethics to data ethics. Focussing on the data helped us make sense of the fact that big data and the data economy can actually create criminal markets and incentivise offenders to commit data crimes. These are typically upstream cybercrimes with secondary downstream crime that include ransomware attacks and data breaches that can be pan-European[128] or spanning multiple other jurisdictions. At the same time, however, we argued that the focus on data should not lead to overlook the heuristic relevance of information contained in the data, and the technology powering Data Science. It is only by looking at technology, data and information as part of the same ecosystem that we can understand the major implications and harmful consequences of upstream and downstream crimes for victims, and challenges they create for the law, law enforcement and also the judiciary. In this scenario, we need to observe how developments in AI will affect data crime.

In the chapter we discussed the evolution of the EU legal framework to support the data economy, namely the Public Sector Information Directive, the free flow of non-

[127] As we discussed in the first edition of this chapter (2018).

[128] Apostolos Malatras and others, 'Pan-European personal data breaches: Mapping of current practices and recommendations to facilitate cooperation among Data Protection Authorities' 33 *Computer Law and Security Review* 458-469.

personal data Directive, the Data Governance Act, the Digital Services Act and the Digital Markets Act (with the Data Act ad AI act noted in passing). We observed with concern the potential for a data crime-data market nexus, calling for data analytics, the data economy and data crime to feed into a cycle that must be looked at together. We point to shortcomings of the legal framework in accounting for the state of cyber insecurity, including gaps in the conceptualisation of harms that are made difficult to redress due to the technology neutral character of the law and the country-of-origin principle.

We also discussed the potential application of data science (data analytics) to policing, which highlights some interesting misconceptions. Commentators often quote from the *Minority Report* film when extolling the virtues of data analytics, but it is worth remembering that the predictions were actually made by psychic pre-cogs, not computers, computers which actually mistook Tom Cruises' character for Mr Yakamoto (Cruise's character had Mr Yakamoto's eyes transplanted). In this, the predictive technology actually failed.[129] One ray of hope in this, otherwise, gloomy tale is that Data Science, and particularly Big Data analytic techniques which utilise algorithms via Machine Learning and Artificial Intelligence can assist practitioners both identify patterns of offending and also (possibly) the path of offending which could lead law enforcement to data criminals. This subsequently raises a number of questions about admissibility of evidence in court, and also the direction of the burden of proof, both of which require a suitable legal basis. However, resolving this issue, as well as discussing how to address the data crime-data market nexus, is for another paper.

ACKNOWLEDGEMENTS

This chapter was originally based upon research conducted for the EPSRC funded research project "Combatting cRiminals In The Cloud" (CRITiCal - EP/M020576/1)

[129] Wall (2018) 'How Big Data Feeds Big Crime'.

which was an interdisciplinary project with Leeds, Newcastle and Durham Universities looking at the impact of cloud technologies on cybercrime between 2015 and 2023.