Aleš Završnik

12 Big data

Abstract: The chapter presents the changing meaning of Big data over time and its relation to digitalization in contemporary 'datafied' society, in which Big data has become viewed as the 'new oil.' The chapter then delves into the question of relevance for criminology. It presents different views and framings of its benefits and risks in the crime control domain and for the production of criminological knowledge. It presents typical uses of Big data in crime control practice and some of the risks to fundamental liberties theorof. Connections with the notion of Artificial Intelligence, narratives of risk, prediction, and pre-emption are also presented.

Keywords: Big data, datafication, human rights, production of knowledge, prediction, pre-emption

Big data gained social science relevance due to increased data processing practices by public institutions and private companies that involve large amounts of data about citizens and (potential) customers. 'Big' data, as opposed to 'small' data, was perceived as a game changer for analyzing and understanding the social realm and for the design of policies, such as health policy, social policy, macroeconomic policy, and also crime policy. Ultimately, the trend was dubbed as a Big data "revolution" (Lavorgna and Ugwudike, 2021). The term 'Big data' is used to describe large new data sources, as well as the associated processes, i. e., collection, analysis, insight, and Big data 'mining'—a process of uncovering patterns in large datasets. Critical authors define Big data not only as the capacity to search, aggregate, and cross-reference large datasets but also as a "specific socio-technical phenomenon" (boyd and Crawford, 2012: 663).

Big data plays an increasingly large role in fields relevant to criminologists. Data gathered from several sources is claimed to offer new perspectives and insights into reasons, factors, and circumstances of past and potential future crimes. The expectation towards Big data is high: not only would big datasets offer new understandings of crime but implicitly also new approaches on how to act upon these, whether it is to prevent crime (e. g., with predictive policing software) or investigate already committed crime (e. g., with a prediction of a risk score of parolee). Its biggest promise seems to be to predict future crime (e. g., place, actors—perpetrators and victims) in order to pre-emptively 'strike' and prevent crime from happening, i. e., to colonize the future (see Prediction by Ķīlis, Gundhus and Galis). The fact that Big data has its origins in the business world has ramifications for the crime control domain, especially in pur-

Acknowledgments: Aleš Završnik acknowledges receiving financial support from the Slovenian Research and Innovation Agency (research core funding no. P5 – 0221 and the research project "Proteverb – Legal, ethical and technological aspects of processing textual and speech data for scientific, research and development purposes", No. V5 – 2265).

Open Access. © 2025 the author(s), published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. https://doi.org/10.1515/9783111062037-013

suing values of effectiveness in a manner of ‘doing more with less.’ It is linked to the neoliberal turn when shrinking police budgets meant that the police must ensure the same level of protection with limited resources (Beck and McCue, 2009).

The concept of Big data is relatively new, but the origins can be traced back to the 1960s and ’70s with the rise of data centers and relational databases. Around 2005, with the advent of social media platforms, it was clear that users were generating data that could be ‘monetized’ and made ‘actionable’ (Zuboff, 2015). In the 2010s, Big data had become a topic of discussion in various domains, and authors were showing enthusiasm that parallels some of the most significant movements in the history of computing, such as the development of personal computing in the 1970s, the World Wide Web in the 1990s, and social media in the 2000s.

The ‘big’ in Big data has acquired different meanings over time, and ‘data’ in Big data generated its own field of research in the social sciences, including software studies and critical data studies. What seemed ‘big’ a decade ago became ‘small’ with the exponential growth of technological capacities. A telling example comes from increasingly powerful Large Language Models (LLM), which are built on supposedly the ‘whole internet.’ For instance, the ChatGPT model required a data center with 10,000 GPUs, i. e., Graphics Processing Units, which are electronic circuits that can perform high-calibre mathematical calculations. ChatGPT training lasted 9 months and cost $100 million in electricity. In 2023, a new data centre with 50,000 GPUs was built, which meant that the cost of the processors alone approached $2 billion (Zgonik, 2023). The volume of ‘big’ has skyrocketed due to more and more data generated by digital appliances—from phones, wristwatches, to fridges, heating systems, and other objects of mundane life connected to the Internet of Things (IoT) or the ‘Internet of everything” (see Internet of Things by Milivojevic). Some have hence also claimed the opposite—that ‘small’ quality data rather than ‘big’ ‘anything’ data is something to aspire to in order to acquire actionable insights to be acted upon.

Today, Big data should also be conceptualized together with other technologies, forming the *Big data pipeline* or *Artificial Intelligence (AI) supply chain.* ‘Ingredients’ (Big data) do not lead to meaningful outputs without a proper ‘recipe’ (AI). Since the meaning of large amounts of data must be extracted in order to furnish ‘actionable’ insights, AI is an essential part of the Big data pipeline. The two—Big data and AI—are related as part of the same logic and values.

Criminological relevance

The first explicit recognition of the relevance of Big data for criminology can be traced to the 2010s when Berk (2012) analyzed changes in computational criminology brought about by machine learning. Forecasting in criminal justice, together with predictive policing, were two specific applications of Big data for crime control. Procedures from computer science and applied mathematics have been used before the ‘advent’ of Big data to animate theories about crime and law enforcement, but a culture of causal modeling thoroughly dominated such methods. Machine learning, instead, comes from a different culture characterized by an ‘algorithmic perspective’ (Berk, 2013).

Big data entered into criminology through several paths: through the increasing use of new types of data (social media or user-generated data), through using computer modeling/algorithms as a predictive tool to guide policing strategies and other criminal justice decisions (Chan and Bennett Moses, 2016) and also through the theorization of the increased focus on risk (Zedner, 2007). The latter is part of wider skepticism of Big data in criminology and shows how criminologists have scrutinized the meaning, circulation, and power relations associated with Big data. Authors reflected on whether the claim that the ‘data deluge’ would make scientific methods obsolete has merits for criminology (Chan and Bennett Moses, 2016) and analyzed predictive policing programmes (Egbert and Leese, 2021). Critical criminology challenged the very novelty of Big data and claimed that the digital turn has failed to fulfill the old dreams of more just and equal societies (Završnik, 2018). What may be new with the Big data ‘revolution’ is the exponential acceleration of neoliberalism, for example reflected in reduced state power and a reinforced private sector, an increase in social and wealth inequalities facilitated by the powerful elite to gain insight into different populations more than ever before. The pressure to institute a process of ‘datafication’—turning everything into data or numbers—in order to ‘monetize’ data and create ‘actionable’ insights is at the core of Big data logic (see Datafication by Chan). Big data, hence, serves specific political ends. Similar to Desrosières’ (2002) analysis of statistics, for which he lucidly noted that it offered new justifications for modern state interventions back in the 19th century, today, Big data offers new justifications for policy interventions. Big data is, hence, in this sense, not an ‘objective’ knowledge but has always been a political endeavor.

Other social sciences were also relevant for criminological attempts to grasp the Big data ‘revolution.’ As a multidisciplinary field of research, criminology has drawn knowledge from critical data studies and critical security studies, critical media studies, critical legal studies, and the work of many authors studying at the intersection of political science, sociology, Science and Technology Studies (STS), ethics, and surveillance studies.

Uses of Big data in crime control

Examples of uses of Big data in the crime control domain include:

1) predictive policing software, traditionally using Big data to predict where and when crimes are likely to occur. The focus is typically on locations (‘hot spots’) or individuals (‘heat lists,’ ‘persons of interest’) (e. g., Kaufmann et al., 2019), which is the case in the software PRECOBS (*Pre Crime Observation System*) which creates graphically supported insights into areas where increased risk for follow-up incidence is estimated within the next 72 hours. Another example is *HunchLab*, now Azavea, a predictive policing program that inputs crime data, cen-

sus data, population density and adds variables like the location of schools, churches bars, clubs, and transportation centers (Ferguson, 2017);

2) social network analysis to map the relationships between criminals and their associates, to identify potential suspects, and analyze criminal networks;
3) DNA databases and other biometric databases for profiling (Kaufmann, 2022);
4) analytic tools for criminal investigations, e. g., to extract knowledge on criminal networks from multiple data sources in real-time, gunshot detection, or video analysis of child abuse images;
5) analytic tools to assist bail courts, e. g., with generating risk scores in bail procedures and assisting parole bodies (Bennett Moses and Chan, 2014).
6) sentencing tools to predict an offender's likelihood of future recidivism, e. g., COMPAS (*the Correctional Offender Management Profiling for Alternative Sanctions*, used in the USA) or *the Risk of Reconviction* (in the UK) and the LSI (*Level of Service Inventory*, used internationally) (Ryberg and Roberts, 2022);
7) tools for legal reasoning and legal research (Legal-tech).

Challenges of Big data in crime control

Since data are the focal point of the Big data paradigm, *privacy and surveillance concerns* form an overarching human rights issue (Kerr and Earle, 2013). Increasing collection and processing capabilities of Big data analytics change *surveillance* of daily life and can have a chilling effect on free speech and other civil liberties. The relation of Big data to surveillance has been theorized as a distinct type of Big data surveillance (Andrejevic and Gates, 2014) and dataveillance. Reflecting on Snowden's revelations of 'dragnet; investigation practices by intelligence agencies, Lyon (2014) observed a transition in surveillance studies from information technology and networks to Big data, which intensified and expanded certain surveillance trends. The future orientation and the quest for pattern discovery of Big data surveillance raised concerns related to privacy, social sorting, and pre-emption (Lyon, 2014).

Data on identifiable individuals collected in legally and/or ethically problematic fashion, such as social media scraping, remain part of Big datasets with unclearly defined data subjects' rights. For instance, it is unclear whether and how data subjects can exercise the rights guaranteed in the personal data protection regime vis-à-vis facial recognition technologies (see Facial Recognition by Fussey).

The *quality of data*, relating to (in)accuracy, completeness of data, representativeness of social groups, etc., varies in the criminal justice domain. Which data is taken in, and which data is left out of the calculus? Data is not a natural resource but a cultural one, it is always 'baked' with (underpinning) human values, interests, and cultural expectations (Gitelman, 2013). Criminal justice data *per definition* does not include unreported crimes (dark figure of crime), which can make it difficult to prevent the so-called 'garbage in—garbage out' effect: poor data leads to poor outcomes. Big datasets, then, tend to indicate social practices (reporting crimes) rather than social reality.

One aspect of data quality relates to *biased data*, which may lead to *discrimination* (see Bias by Oswald and Paul). Research on sentencing prediction instruments has confirmed how criminal history is, in fact, a proxy for race (Harcourt, 2015), meaning that data for such instruments is biased by racialized histories. Police data often reflects biased functioning of police operations, and Big data analytics would perpetuate such biases. If digitized societies are divided along gender, race, wealth, and other lines, and structural inequalities (Ávila et al., 2019: 97), practices based on Big data will *per definition,* reflect such biases (called 'closed-loop'). Examples abound and include the over-policing of minority neighborhoods and the under-policing of white-collar crime as predictive policing tools focus on street and property crimes. As O'Neil (2016) vividly expressed, models are opinions embedded in math. The interpretation of results of Big data analytics is not straightforward as well. It is inherently affected by human knowledge of the analyzed domain and data. Data scientists must work alongside domain-specific scientists in order to ascribe meaning to the calculated results.

The *cost* is another concern, as Big data tools are increasingly sophisticated and expensive to train and maintain, e. g., Microsoft and OpenAI used $100 million worth of energy for training ChatGPT, and they keep spending $700k per day for running it (Zgonik, 2023).

The *'Black box' effect* and a lack of *transparency* are general problems of Big data analytics (Pasquale, 2015). Criminologists have also raised this concern over the latest fifth generation of machine learning-based (ML) risk assessments. Ávila et al. (2021) claim that while the latest ML-based risk assessments are focused on the elimination of biases and self-adjust to new data over time, they also deepen the black box problem. They claim that opacity, proprietary nature, and fluid characteristics of predictive models undermine legal protections.

Blurring regulatory boundaries is another critical aspect of Big data use in crime control. The new mathematical language serves security purposes well (Amoore, 2014). Here, new concepts are being invented in order to understand crime (knowledge production) and act upon it (crime control policy). Such concepts include 'meaning extraction,' 'sentiment analysis,' and 'opinion mining.' However, these concepts are blurring the boundaries in the crime control domain: instead of the relatively well-defined concepts of criminal law, such as suspect, reasonable doubt, etc., which serve as regulators of and thresholds for the intervention of law enforcement agencies, new concepts no longer sufficiently confine agencies nor prevent abuses of power (Završnik, 2021).

Risk, prediction, and pre-emption

A central use of Big data analytics in criminology is to garner predictions and identify risks with the final goal of managing and pre-empting the risks. Anticipation re-focuses crime control actors (Zedner, 2007). They reorient their practices and "focus on the future more than on the present and the past. In the context of neo-liberal governance, this anticipation is likely to place more weight on surveillance for managing consequences rather than research on understanding causes of social problems such as crime and disorder." (Lyon, 2014: 6 – 8). Big data, so the critique, tends to accompany this shift from causation (e. g., aiming to uncover factors leading to crime) to correlation (e. g., aiming to uncover factors of crime that need to be 'managed' here today).

Pre-emptive approaches in the crime and security domain have been growing steadily since the 1990s and have been extensively augmented after 9/11. Such approaches are a bureaucratic incentive to over-collect data (Lyon, 2014). Big data is at the core of the transition toward pre-emptive approaches in tackling crime. It enabled prediction and triggered a new philosophy of *pre-emption.* Predictive analytics that transcends human perception have been one of the most attractive aspects regarding the application of Big data in crime control, for example 'connecting the dots' in terabytes of data in money laundry schemes would be impossible for a human eye, while Big data analytics can help follow the money by finding hidden correlations. However, Kerr and Earle (2013) warn that Big data's promise of increased efficiency, reliability, and utility might be seen as the justification for a fundamental jurisprudential shift from an *ex-post facto* system of penalties and punishments to *ex-ante* preventative measures. The new form of 'pre-emptive prediction,' as Kerr and Earle (2013) define it, is intentionally used to diminish a person's range of future options. Predictions are used to assess the likely consequences of allowing or disallowing a person to act in a certain way (Kerr and Earle, 2013). Predictions, here, are not concerned with an individual's actions but with whether an individual or group should be permitted to act in a certain way. Big data is thus used "not only to understand a past sequence of events, but also to predict and intervene before behaviours, events, and processes are set in train (sic) [i. e. motion]" (Lyon, 2014: 4). Pre-emption means acting to prevent an anticipated event from happening. Taken to its extreme, the philosophy of pre-emption is not merely pro-active—it is aggressive. Kerr and Earle (2013) exemplify that no-fly lists employing predictive algorithms curtail liberties. Before their development, high-risk individuals were generally at liberty to travel unless the government had sufficient reason to believe that such individuals were in the process of committing an offense. But now, a no-fly list obliterates the need for such evidence. Prediction replaces the need for proof. Big data underwrites anticipatory and pre-emptive approaches that move crime policy towards actuarialism and consequentialist concerns with managing crime rather than seeking its causes in an attempt to eliminate them.

Conclusion

The discussions on Big data remain highly pertinent to contemporary criminology, albeit sometimes being framed as a discussion on risk, prediction, prevention and pre-emption, automation and 'algorithmization' of crime and crime control. AI tools should be understood as part of the 'Big data pipeline.' Analysis of automated decision-making in criminal justice generating new forms of ‘automated justice’ (Marks et al., 2015), ‘algorithmic justice’ (Završnik, 2021), ‘simulated justice’ (O’Malley, 2010) should be conceptualized together with findings of Big data studies. Big data is then part of the wider trend of ‘algorithmic governance’ and ‘algorithmic governmentality’ (Hannah-Moffat, 2019). It should also be read together with critiques that were transitioning from ‘the rule of law’ to ‘the rule of algorithms’ (‘algocracy’).

Big data can *enhance* criminology’s scientific method in understanding patterns of crime and analyzing and verifying theories of crime. Big data analytics may also improve the effectiveness, legitimacy of criminal justice actors and increase the investigative powers of law enforcement agencies. However, a *critique* of Big data focusing on Big data’s contribution to reducing democratic freedoms, reconfiguring privacy and redefining the role of information in contemporary societies needs to accompany the implementation of Big data tools. A discussion on the limits and thresholds in the use of Big data analytics in crime and social control is needed, one of which is, for example, bulk biometric surveillance. The change in orientation of traditional criminal justice based on an after-the-fact system of punishments to one based on future-oriented preventative measures—anticipation and pre-emption induced by Big data must be examined and regulated. The need for *cross-disciplinary dialogue* between developers, data scientists, analysts, criminologists, and others about the legal, socio-political, and discriminatory effects of Big data analytics cannot be understated.

Suggested reading

Amoore, L. (2014). Security and the incalculable. *Security Dialogue*, 45(5), 423 – 439.

Chan, J., & Bennett Moses, L. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, 20(1), 21 – 39.

Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology*, 23(4), 453 – 470.

Završnik, A. (ed.) (2018). *Big Data, Crime and Social Control.* Abingdon and New York: Routledge

References

Amoore, L. (2014). Security and the incalculable. *Security Dialogue*, 45(5), 423 – 439.

Andrejevic, M., & Gates, K. (2014). Big Data surveillance: Introduction. *Surveillance & Society*, 12(2), 185 – 196.

Ávila, F., Hannah-Moffat, K., & Maurutto, P. (2021). The seductiveness of fairness: Is machine learning the answer? – Algorithmic fairness in criminal justice systems. In M. Schuilenburg & R. Peeters (eds.), *The Algorithmic Society. Technology, Power, and Knowledge* (pp. 87 – 103). Abingdon & New York: Routledge.

Beck, C., & McCue, C. (2009). Predictive Policing: What can we learn from Wal-Mart and Amazon about fighting crime in a recession? *The Police Chief Magazine*, 76(11), 18 – 24.

Bennett Moses, L., & Chan, J. (2014). Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal*, 37(2), 643 – 678.

Berk, R. (2012). *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* New York: Springer.

Berk, R. (2013) Algorithmic criminology. *Security Informatics*, 2(1).

boyd, d., & Crawford, K. (2012). Critical questions for big data' *Information, Communication & Society*, 15(5), 662 – 679.

Chan, J., & Bennett Moses, L. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, 20(1), 21 – 39.

Desrosières, A. (2002). *The Politics of Large Numbers: A History of Statistical Reasoning.* Cambridge, MA: Harvard University Press.

Egbert, S., & Leese, M. (2021). *Criminal Futures: Predictive Policing and Everyday Police Work.* Abingdon & New York: Routledge.

Ferguson, A. G. (2017). *The Rise of Big Data Policing. Surveillance, Race, and the Future of Law Enforcement.* New York: New York University Press.

Gitelman, L. (ed.) (2013). *'Raw Data' is an Oxymoron.* Cambridge, MA: MIT Press.

Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology*, 23(4), 453 – 470.

Harcourt, B. E. (2015). Risk as a proxy for race. *Federal Sentencing Reporter*, 27(4), 237 – 243.

Kaufmann, M. (2022). DNA as in-formation. *WIREs Forensic Science.* Online First.

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. *The British Journal of Criminology*, 59(3), 674 – 692.

Kerr, I., & Earle, J. (2013). Prediction, preemption, presumption: How big data threatens big picture privacy. *Stanford Law Review Online*, 66(65), 65 – 72.

Lavorgna, A., & Ugwudike, P. (2021). The datafication revolution in criminal justice. *Big Data & Society*, 8(2).

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, 1(2), 1 – 13.

Marks, A., Bowling, B., & Keenan, C. (2015). Automatic justice? Technology, crime and social control. In R. Brownsword, E. Scotford, & K. Yeung (eds.), *The Oxford Handbook of Law, Regulation and Technology* (pp. 705 – 730). Oxford: OUP.

O'Malley, P. (2010). Simulated justice: Risk, money and telemetric policing. *British Journal of Criminology*, 50(5), 795 – 807.

O'Neil, C. (2016). *Weapons of Math Destruction.* New York: Crown.

Pasquale, F. (2015). *The Black Box Society.* Cambridge, MA: Harvard University Press.

Ryberg, J., & Roberts, J.V. (eds.) (2022). *Sentencing and Artificial Intelligence.* New York: Oxford University Press.

Završnik, A. (ed.) (2018). *Big Data, Crime and Social Control.* Abingdon and New York: Routledge.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18(5), 623 – 642.

Zedner, L. (2007). Pre-crime and post-criminology? *Theoretical Criminology*, 11(2), 261 – 281.

Zgonik, S. (2023, 8 July). *Dr. Jure Leskovec za N1.* N1 News.

Zuboff, S. (2015). Big Other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1), 75 – 89.