University of New South Wales Law Research Series

IS BIG DATA CHALLENGING CRIMINOLOGY?

JANET CHAN AND LYRIA BENNETT MOSES

[2020] *UNSWLRS* 81

UNSW Law
UNSW Sydney NSW 2052 Australia

E: unswlrs@unsw.edu.au
W: http://www.law.unsw.edu.au/research/faculty-publications
AustLII: http://www.austlii.edu.au/au/journals/UNSWLRS/
SSRN: http://www.ssrn.com/link/UNSW-LEG.html

Is Big Data Challenging Criminology?

Janet Chan and Lyria Bennett Moses, UNSW Law

This is the original version of a paper (23 January 2015) that was amended in response to referee comments and published as Chan J; Bennett Moses L, 2016, 'Is Big Data challenging criminology?', Theoretical Criminology, vol. 20, pp. 21 - 39, http://dx.doi.org/10.1177/1362480615586614

Abstract

The advent of 'Big Data' and machine learning algorithms is predicted to transform how we work and think. Specifically, it is said that the capacity of Big Data analytics to move from sampling to census, its ability to deal with messy data and the demonstrated utility of moving from causality to correlation have fundamentally changed the practice of social sciences. Some have even predicted the end of theory—where the question why is replaced by what—and an enduring challenge to disciplinary expertise. This article critically reviews the available literature against such claims and draws on the example of predictive policing to discuss the likely impact of Big Data analytics on criminological research and policy.

Introduction

In a provocative and much quoted magazine article, Chris Anderson (2008) has concluded that with the advent of the ‘Petabyte Age’, this ‘data deluge makes the scientific method obsolete’. In particular, he contends that this massive volume of data ‘forces us to view data mathematically first and establish a context for it later’. He uses the example of Google which, he claims, has succeeded by simply using ‘applied mathematics’ and assuming that ‘better data, with better analytical tools, would win the day’. In short, in Anderson’s view, Google succeeded without having a theory or a model, because, as Google’s research director Peter Norvig was quoted as saying, ‘All models are wrong, and increasingly you can succeed without them’. And this logic doesn’t just apply to advertising or the generation of a translation app—significantly, Anderson uses examples from physics, biology and genetics to argue that this kind of approach is transforming the scientific method:

> Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete...There is now a better way. Petabytes allow us to say: ‘Correlation is enough.’ We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. (Anderson 2008)

Anderson’s claims were elaborated and confirmed in a book by Mayer-Schönberger and Cukier (2013) where the authors characterize Big Data as ‘*a revolution that will transform how we live, work and think’*. Many examples were used to show that the ‘ascendency’ of Big Data has led to ‘three shifts in the way we analyze information that transform how we understand and organize society’: because of the availability of large volumes of data, (i) sampling is a thing of the past, Big Data is increasingly giving us access to *all* the data (*N = all*), (ii) data accuracy is no longer a problem we need to be concerned about, and (iii) there is no need to search for causality, correlation provides ‘novel and invaluable insights’ (2013:12, 14). Their conclusion is: ‘Big data is about *what*, not *why*’, we can ‘let data speak for itself’ and in ‘many situations this is good enough’ (2013:14).

The authors do not support Anderson’s (2008) end-of-theory claim, since ‘Big data itself is founded on theory’, but insist that Big Data ‘does fundamentally transform the way we make sense of the world’ and ‘Causality …is being knocked off its pedestal as the primary fountain of meaning’ (Mayer-Schönberger and Cukier, 2013: 71, 67). In particular, the authors suggest that the social sciences is ‘one of the areas that is being most dramatically shaken up by N= all’— social scientists will lose their monopoly as experts in social research:

> … the pioneers in big data often come from fields outside the domain where they make their mark. They are specialists in data analysis, artificial intelligence, mathematics, or statistics, and they apply those skills to specific industries. … To be sure, subject-area experts won’t die out. But their supremacy will ebb. From now on, they must share the podium with the big-data geeks, just as princely causation must share the limelight with humble correlation. This transforms the way we value knowledge, because we tend to think that people with deep

> specialization are worth more than generalists—that fortune favors depth. Yet expertise is like exactitude: appropriate for a small-data world where one never has enough information, or the right information, and thus has to rely on intuition and experience to guide one's way. (2013:141-142)

Reactions to these claims about Big Data vary among academics. Uprichard's (2013) warning that 'the big data hype is generating… a methodological genocide… even has a flavor of being a disciplinary genocide' is, by her own admission, 'a bit melodramatic', but it does reflect a sense of panic and urgency about this imminent threat. Her call for social scientists (especially sociologists) to 'fight back' by improving 'our quantitative skills' and finding a way of 'voicing our capacity to deal with big data' is suggestive of an emerging 'jurisdictional conflict' between social scientists and Big Data analysts (cf Abbott, 1998). Other academics feel less threatened by the warning that Big Data will bring about 'the death of the theorist' because 'there will always be things that are left unsaid, things that haven't been measured or codified' and 'even big data patterns need someone to understand them' (Steadman, 2013). Nevertheless, anxieties remain because 'Social science appears to be escaping the academy' and being housed in the offices of 'data scientists' working for Google, Facebook, Amazon and other commercial firms (Williamson, 2014).

In this paper we will critically review the available literature against claims about the gaming-changing influence of Big Data on criminology. The paper is organized as follows. The next sections explain the concept of Big Data in greater detail and provide a brief overview of the current use of Big Data in criminological research. It then goes on to examine more closely the claims made in relation to Big Data, particularly the side-lining of theory and causation in favour of pattern-identification and correlation. The paper concludes with a discussion of the likely impact of Big Data analytics on the theory and practice of criminology.

What is Big Data?

Defining ‘Big Data’ is not a straightforward exercise. It is possible to define Big Data by reference to the size and type of data sets being employed, the capabilities of a data storage, processing and/or analytic system, as a set of marketing claims about what is enabled by particular technologies, or as a social and cultural phenomenon. The reason for the diversity of definitions is the variety of technologies employed, platforms and systems potentially involved and purposes to be achieved. The single label exists to capture a broad trend in how data is captured, stored and used rather than to identify a particular product or process.

A popular definition of Big Data is that it involves (at least) three Vs– Volume (the amount of data), Velocity (the speed at which data is being added and processed) and Variety (the fact that data may come from multiple sources using different formats and structures). This definition focuses on the data, generally for discussing and solving technical problems in analyzing such data, for example, how data can be securely or inexpensively stored and communicated, how data can be efficiently processed, or how data sets with different structures can be combined. Some definitions go beyond the traditional three Vs by requiring additional features. For example, Kitchin (2014b) has suggested that to be ‘Big Data’, data sets should generally be exhaustive in scope (capturing entire populations or systems), fine-grained in resolution, indexical in identification (attaching to specific people or things rather than groups or types), relational in nature (so that data sets can be merged) and flexible (with the ability to add new fields or expand in size).

In all of these feature-based definitions, what counts as large, efficient, flexible or diverse is a moving target. The McKinsey Global Institute recognizes this by defining Big Data as 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze' (Manyika et al., 2011). Big Data in this sense sits perpetually on the technological frontier – 'older' approaches fall outside the definition once they come to be viewed as 'typical database software tools'.

These definitions of Big Data are focused exclusively on its *technological* features, but as boyd and Crawford (2012) remind us, Big Data is more appropriately seen as 'a cultural, technological, and scholarly phenomenon' that is based on the 'interplay' of three elements:

> (a) *Technology*: maximizing computation power and algorithmic accuracy to gather, analyze, link and compare large data sets.
>
> (b) *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
>
> (c) *Mythology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity and accuracy. (2012:663; see also Jurgenson, 2014)

What is often neglected in discussions about the influence of Big Data is an appreciation of its mythic qualities. Boyd and Crawford have tried to debunk Big Data myths by exposing its misleading claims to objectivity and accuracy, the confusion of size with representativeness, the absence of contextual information to guide interpretation, and unexamined issues relating to ethics and differential access, but one aspect of Big Data may be becoming irreversible: it ‘creates a radical shift in how we think about research … Big Data reframes key questions about the constitution of knowledge, the process of research, how we should engage with information, and the nature and the categorization of reality’ (2012:665). This leads us to ask: to what extent has this happened to criminology and what are the consequences?

The use of Big Data in criminology

To discuss whether Big Data may present a challenge to criminology, we will need first of all to define what criminology is and how Big Data is being used in criminology. There are, of course, many questions, theories, methods and interventions that come under the rubric of criminology; the list is endless: the study of crime prevalence and trends; theories of causation of crime, crime prevention and corrections; research methods for studying crime; questions about the definition, social construction and control of crime; features of the criminal justice system; strategies for the prevention of crime and treatment of offenders; and so forth. One persistent issue often discussed by criminologists is whether criminology is a discipline or an intersection of numerous disciplines including sociology, psychology, law, and political science. In a recent volume, *What is Criminology?*, the editors admit that ‘Borders, in criminology as elsewhere, are policed, yet they are also frequently transgressed’ (Bosworth and Hoyle, 2011, front cover jacket). Even if we agree that no academic borders are impervious and acknowledge that, given the diversity of intellectual influences and investigative methods being used for criminological research, it may be a waste of time to define what *really* is criminology, many of us would still be comfortable with defining criminology broadly as a field of social scientific practice devoted to advancing knowledge about crime and deviance (see Chan, 2000). Note, however, that this does not imply that researchers in these areas would necessarily call themselves ‘criminologists’, nor do they necessarily see their work as ‘doing’ criminology.

Despite the breadth of this definition, the current overlap between criminology and Big Data is relatively small. There are two main areas where Big Data has been used for researching crime and deviance. First of all, the use of Big Data such as social media as *data* in criminological research is becoming more prevalent. This is consistent with a growing trend in social research that takes advantage of the availability of ‘naturally occurring or “user-generated” data at the level of populations in real or near-real-time’ (Edwards et al. 2013:245). Secondly, there is a trend towards using computer *modeling/algorithms* as a predictive tool to guide policing strategies and other criminal justice decisions (Uchida, 2013; Berk and Bleich, 2013). While this second trend does not necessarily involve Big Data, it is a development that is increasingly tied to the availability of Big Data.

The use of Big Data as one type of research data

Criminological researchers have started to use Big Data such as social media as one type of research data. As user-generated data, social media gives researchers access to self-report data about social media users’ activities, perceptions and opinions (Edwards et al., 2013). Such data can complement or act as surrogates for more traditional social science data such as those derived from experiments, surveys and interviews. Social media data is also unique in that it is generated in real time and at the level of populations (2013:249). Criminology has had a long tradition of dealing with self-report data, from victim and public attitudes surveys to studies of delinquents and adult offenders; the availability of

Big Data in the form of social media communications can be extremely attractive. Nevertheless, there are not many published examples of such research.

One prominent example is Procter et al.'s (2013) project that analyzed a corpus of 2.6 million public messages (tweets) on Twitter around the time of the August 2011 riots in England to examine the role of social media during the riots. This type of analysis is not dissimilar to traditional media content analysis, except for the size of the dataset. Without access to adequate natural language processing (NLP) tools, the researchers resorted to 'less sophisticated computational tools to expose underlying structures in the corpus… to identify potentially significant fragments' and then used 'established qualitative methods' to analyze these fragments (2013:199).

Another example involves using social media for criminal intelligence. Watters and Phair (2012) tested a 'new methodology' for using social media data to detect the marketing and distribution of illicit drugs. While the researchers argue that there is scope for automation, the so-called Automated Social Media Intelligence Analysis (ASMIA) is in fact a manual process using Google's search engine. They conducted two experiments that show that ASMIA can assist with intelligence gathering but 'longstanding problems with word sense disambiguation and deeper-level sematic processing will remain' (2012:74). This way of using social media data is similar to the work of intelligence or investigative officers in policing or regulatory agencies.

Telecommunication data is another source of Big Data that has been used in criminological research. Traunmueller et al. (2014) used such data to evaluate urban crime theories such as Jacob's (1961) environmental and Felson and Clark's (1998) opportunity theory. The researchers obtained access to anonymized and aggregated data from a mobile phone provider in the UK for three weeks between December 2012 and January 2013 for the Metropolitan Area of London. This data, which consisted of 12 million 'footfall count entries', was broken down into 23,164 grid cells; footfall counts were provided per hour as well as broken down by gender and age group. The researchers created metrics that measure the diversity of people and ratios of visitors, residents, workers, female population and young people for each spatio-temporal unit and correlate these with crime data provided by the Metropolitan Police and the City of London Police. This kind of analysis is similar to traditional criminological research except for the size of the dataset.

It is clear from these examples that the use of Big Data as one type of research data poses minimal challenge to the way criminological research is conceptualized and conducted, except in relation to the size of the datasets which may call for different analytical techniques. The questions posed by the researchers are similar to conventional criminological research. These studies consciously or unconsciously make use of established concepts (such as coding categories used by Proctor et al. (2013)) or theories (such as urban crime theories in Traunmueller et al. (2014)) drawn from criminology or other social sciences. Contrary to Mayer-Schönberger and Cukier's (2013) claim that messy data is not a problem with Big Data, researchers are very concerned about data quality. Procter et al. (2013) note that their Twitter data is biased because of their reliance on hashtags to select tweets, their exclusion of DMs (Direct Messages) which are not public, and the unrepresentativeness of Twitter users. Traunmueller et al. (2014) similarly discuss the limitations of their research in terms of being able to access data from only one mobile phone provider (albeit a major one with almost 25% of the market share in 2013). They also had to 'cleanse' the telecommunication data to remove inconsistent data. In general, Big Data such as telecommunication data is not readily available to researchers, hence the reliance on social media data. Yet social media data suffers from numerous flaws, including its 'low fidelity', unrepresentativeness, the absence of key demographic variables in substantial proportions of the data, and the danger of identification of users (Edwards et al., 2013:247-256).

The use of computer modeling/algorithms as predictive tools

The second major area involving Big Data research in criminology relates to the use of computer modeling/algorithms as a predictive tool for risk analysis or crime prevention. Criminology is no stranger to the use of data mining and predictive analysis. Even before the advent of Big Data, the use of new technology and statistical analysis has been a growing trend in policing research (see Uchida, 2009; Perry et al., 2013). The rise of 'predictive policing' goes beyond hotspot analysis, problem-oriented policing, and crime mapping to use data and analytics to 'forecast where and when the next crime or series of crimes will take place' (Uchida, 2013: 3871). These predictions can be about 'places and times with an increased risk of crime', 'individuals at risk of offending in the future', creating 'profiles that accurately match likely offenders with specific past crimes', or identifying groups or individuals at risk of becoming victims of crime (Perry et al., 2013:8-9).

While Big Data can be used to enhance predictive approaches, it is not a necessary element. Prediction can be based on a model (Groff and La Vigne, 2002; Johnson and Bowers, 2004) in ways that engage with traditional empirical approaches. For example, Johnson and Bowers (2004) draw on interviews with offenders to propose a model for 'prospective hot spotting' based on past burglary events, verifying this model through the analysis of data. The types of analytics used in predictive policing can range from computer-assisted queries, analysis of databases, to regression and classification models, risk terrain analysis, and advanced data mining techniques (Perry et al., 2013:10-12).

The introduction of social media data has brought Big Data to predictive policing. For example, Gerber (2014) has shown that the incorporation of linguistic analysis of spatiotemporally tagged tweets has improved the performance of crime prediction models. Social media data has also complicated the calculation of crime rates. Malleson and Andresen (2014:6-8) have shown that ‘different spatial patterns of crime rates emerge when using two different population-at-risk measures: the residential population (measured by the 2011 UK census) and the ambient population (measured by counting the number of messages posted to the Twitter social media service)’. The massive volume and unstructured nature of Big Data such as social media data have spawned research in developing software algorithms to analyze such data. Williams et al. (2013), for example, built and tested a ‘social media tension-monitoring engine’ using a ‘Collaborative Algorithm Design’ involving experts in criminology, sociology of language and computer science. A corpus of 1022 tweets were collected one month before and one month after a well known incident of racial abuse between two professional football players in 2011. The results indicated that the tension-monitoring engine and the MCA (Membership Categorisation Analysis) methodology used for classifying the level of tension ‘function as a sound alternative to human police coders’ (2013:476).

The predictive analytics presented thus far does not deviate substantially from normal research methods used in quantitative criminology. The use of predictive policing, for example, is very much informed by established concepts and theories such as situational crime prevention, problem-oriented policing and repeat victimization theory (Uchida 2013). However, one area of predictive analysis that can potentially challenge established research practice is the use of machine-learning procedures such as ‘random forests’ to assess offenders’ risk of reoffending (Berk and Bleich, 2013). For Berk and Bleich criminal justice forecasting should not be confused or conflated with explanation:

> As a formal matter, one does not have to understand the future to forecast it with useful accuracy. Accurate forecasting requires that the future be substantially like the past. If this holds, and one has an accurate description of the past, then one has an accurate forecast of the future. That description does not have to explain why the future takes a particular form and certainly does not require a causal interpretation. (2013:516)

The authors have argued for an approach that is focused solely on maximizing forecasting accuracy. They are not concerned with ‘why certain predictors improve forecasting accuracy and no attempt is made to interpret them as explanations for the forecasted behavior’ (2013:517). Indeed, the authors suggest that if shoe size happens to be a good predictor of recidivism, it ‘can be included as a predictor’, because ‘[w]hy shoe size matters is immaterial’ (2013:517). This seems to echo Mayer-Schönberger and Cukier’s (2013) suggestion that Big Data will lead to the end, or at least the decline in importance, of theory in social research. If true, such a development would radically alter practices within criminology, as being satisfied with knowing *what* without understanding *why* ‘overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality’ (Mayer-Schönberger and Cukier, 2013:7). It is time we examine this claim more closely.

Examining the end-of-theory claim

What theory is dead?

In order to examine the claim that Big Data dispenses with theory, we need to look more closely at the definition of *theory*. Theory has been an integral part of criminology, not only for those interested in developing theories, but also for empirical researchers concerned with theory building and theory testing. In fact, theories underpin every research enterprise, even for researchers who claim to be 'atheoretical', because theory helps us make sense of the world. The idea that correlation is good enough, or that the data speaks for itself, is convincing for researchers only if the correlation makes sense in light of some existing or new theory. Thus, in a sense, all data analytics involves theory. There are theories as to which types of algorithms are best suited to solving which kinds of problems, as well as the suggestion that employing large enough data sets allows one to solve all problems.

Decisions as to what data is collected, what data formats are used, how they are ontologically defined, how they are stored (and for how long), and the choice of analytic tools employed are all based on theory, albeit often unexamined (Gitelman, 2013; Kitchin, 2014b; Jurgenson, 2014). In the case of data analysis within law enforcement, software commonly makes assumptions that implicitly or explicitly rely on theories of crime and criminal behaviour. For example, theories have been developed as to the length of time that historic data is relevant for predicting future crimes (eg Johnson and Bowers, 2004). These theories, or at least the 'knowledge' that the predictive value of historic data decreases over time, play an important role in the design and deployment law enforcement analytic software. Further, law enforcement agencies do not have equal access to all data; conclusions drawn will inevitably be shaped by the availability of different datasets, which in practice are a long way from *N = all* (cf Leonelli, 2014).

Choices as to data availability may rest on theory concerning the relevance of different datasets to law enforcement intelligence and investigations, but may equally be ‘the serendipitous result of social, political, economic and technological factors.’ (Leonelli, 2014: 7).

So what is meant by the ‘end of theory’? There are actually two related interpretations here. The first is that researchers need not start with a theory and then test it against data, but can rather gain insights directly from the data itself (Kitchin, 2014b). The second is that questions around theory are less relevant, because so long as correlations are identified, there is no need to understand causation or the mechanisms that explain why particular patterns exist (Mayer-Schönberger and Cukier, 2013). The first reflects an inductive or data-driven epistemology (Kitchin, 2014b), while the second is a judgment on what kind of knowledge is useful, authentic and ‘objective’ (Jurgenson, 2014). The two interpretations are related because the ‘insights’ gained from examining data in the absence of a hypothesis or model will typically be knowledge about correlation, which is then (under the second interpretation) perceived as more authentic and objective than knowledge gained through the traditional scientific method.

However, the two interpretations are not isomorphic. Analyzing data in the absence of a fully developed theory is not truly a Big Data idea. Exploratory research may use relatively raw observational data as a basis for developing new theories rather than testing pre-identified hypotheses, although as noted above this is never completely theory-independent. Further, it is possible to glean knowledge about causation through analysis of observed (as opposed to experimental) data by making assumptions, generally grounded in a developing theory (Morgan and Winship, 2007). Further, observations about patterns and correlations may be interesting for their own sake. For example, the mere fact that certain characteristics of people correlate with particular categories of crime may be an interesting observation that adds to our knowledge *about* those categories of crime or that species of deviance. The more controversial and interesting question is whether the ability of Big Data techniques to generate more accurate and precise knowledge about correlations obviates the need for theory. This is the second interpretation.

Machine-learning and the end of theory

As pointed out earlier, the use of predictive analytics does not necessarily make theory irrelevant, but it has been suggested in the context of using machine learning procedures for prediction that accuracy should be the only consideration and explaining why a predictor is useful is ‘immaterial’ (Berk and Bleich, 2013:517). It is important, therefore, to examine more closely what machine learning is and why it makes theory unnecessary.

The concept of ‘machine learning’ implies that a computer performs an algorithm that allows it to make predictions (e.g. about whether a parolee will reoffend within two years of release) on the basis of historical data (e.g. parole files on individual prisoners with demographic details and offending history). There are many types of machine learning algorithms but a general feature is that the algorithm will use some of the historical data as a ‘training set’ from which it will ‘learn’ actual or probabilistic correlations, saving some of the data for verification purposes. One machine learning technique that is popular and has been found to performs well in predicting recidivism is the random forest algorithm (Berk, 2013; Berk and Bleich, 2013). This will be explained briefly below. While the details are quite technical (and there are better guides, see e.g. Berk (2012) from which our summary is derived), a basic understanding is necessary to see how these methods differ from traditional approaches driven by a theory of the domain being analyzed.

For example, suppose we are interested in predicting the dependent variable Recidivism (which has two values Yes or No) from a range of independent variables (predictors). A random forest is an ensemble of classification trees, which can be created in the following way (see Berk , 2012). A random sample of size N is drawn with replacement from a 'training' data set of N observations (observations not selected are later used as 'test data'). A small sample of predictors (e.g. three) is then drawn at random, and the 'best' one is used as a basis for determining the first partition on the classification tree. Note that 'best' here means the partition that most reduces heterogeneity of outcomes at each node by reference to a chosen index (Berk prefers the Gini index). For each terminal node, the Bayes classifier is used to assign a class to that node, meaning that the class assigned is the most probable class for that node. The observations that were not selected to grow that particular tree (approximately a third of the training set) are dropped down the tree (sorted according to the partition criteria) until they land at a terminal node, where they are assigned the class associated with that terminal node. This entire process is repeated numerous times to produce a large number of classification trees, each assigning approximately a third of the training data to a class. For each observation in the training data, classification is by vote over all the trees for which that observation was 'out of the bag' (i.e. was dropped down the tree, not used in constructing the tree). Because this data is part of the training set, the actual classification is known and it is therefore possible to construct a confusion table that records actual against predicted classifications, so as to work out (for example) the percentage of predicted classifications that correspond with actual classifications. To make a new prediction, one 'votes' over all the trees that were produced in this process, choosing the class with the most 'votes'. Note that the technique may shift based on considerations such as limited budgets, subjective evaluation of the seriousness of different crimes and the differing ‘costs’ of different predictive errors (Berk, 2012).

The important point here is not so much to understand the technical details of random forest algorithm, but rather to make some observations about the methodology involved and the kinds of information it generates. The first point to note is that this approach seems to confirm Chris Anderson’s (2008) thesis about ‘the end of theory’. While it is possible to theorize about why this particular technique generates insights (thus form theories about its predictive capacity compared with alternative approaches), there is no need to develop a hypothesis *ex ante* as to what kinds of relationships might exist. There are, for example, no assumptions that a relationship between two variables will be linear or logarithmic, or that a distribution will be normal. Further, because a large number of potential predictors can be used, one can take a ‘kitchen sink’ approach (Berk, 2013) and treat multiple factors as potentially important. The algorithm can spot complex non-linear relationships between outcomes and multiple predictors with no pre-existing theory to suggest the benefits of exploring any particular relationships ex ante.

Secondly, while the algorithm can predict the classification of a new data point, the algorithm gives no causal explanation for such a prediction. In other words, the algorithm itself does not present its ‘reasons’ for making a particular classification. It is not possible to look at the set of classification trees and infer particular comprehensible relationships between predictors and classifications that might enable an intuitively causal understanding. The most that can be extracted from a random forest procedure is the contribution a particular predictor (e.g. age) makes to forecasting accuracy or the degree to which each predictor is related to the response (Berk, 2012: 66-9). As Anderson (2008) suggests, we can have conclusions without theory (other than theories of statistics and machine learning) and, in particular, correlations without a theory of causation.

Finally, these techniques are particularly difficult for non-experts to understand. Answers are produced by the algorithm but there is no logic that can be questioned as such. People have to trust the process. This is what often gives rise to the 'mythology' associated with Big Data, as described by boyd and Crawford (2012). Because these techniques make predictions with seeming clairvoyance based on methods that people who rely on those predictions cannot understand, there is an 'aura' around the predictions. There are no opportunities to challenge claims about future behaviour based on particular alleged correlations (e.g. race and crime) as they remain hidden. It is difficult to respond to an algorithm that one may not understand based on data and employing software and hardware to which one may not have access. Further, some predictive analytic techniques offer a perceived degree of mathematic precision that suggests a 'scientific' oracle.

Criminology without theory?

Our review of the literature has suggested that it is not Big Data *per se*, but a particular way of using Big Data, namely the use of machine-learning algorithms for making predictions, that may have the most serious impact on criminological research practice. This is because machine-learning appears to remove the need for criminological expertise in data analysis and produce results that may be impossible to explain. Our concern about this kind of analysis is not that it is incapable of generating insights, but rather that, without more, those insights are not of a kind that necessarily yields useful predictions or justifies action.

There are several problems with relying on correlations alone in predictive analysis (see generally Harcourt, 2007). These can be discussed more precisely using an example such as predictive policing.

First of all, the ability to make reliable predictive statements based on historical data is dependent on an assumption of continuity, i.e., that factors relevant to predicting behaviour in the past will continue to be relevant in the future. Thus despite the mantra that Big Data offers 'N = all', where attempting to make a prediction, researchers never truly have *all* the data because data about the future has not yet been collected. Predictive policing is thus not the equivalent of seeing future crimes in a crystal ball (Perry et al., 2013). This fact renders Big Data more useful for predicting some kinds of crimes than others. For example, the location of future burglaries may be more likely to relate to historical data than the location of future kidnappings (Lever, 2014). Further, given that few variables in the world are held constant, even predictable crime-types are susceptible to broader social changes (e.g. where communities become more or less affluent over time). Determining the circumstances in which historical correlations can yield accurate future predictions requires *some* theory to inform the likelihood of stability. In the case of predictive policing, the ability to forecast 'stranger crimes' based on historic data can be justified by theories such as routine activity theory, rational choice theory and crime pattern theory (Perry et al., 2013: 3). In the absence of a theory that justifies the assumption of continuity, or evidence of its truth, machine learning that draws on historical data cannot be used reliably to predict future crime.

Secondly, the problem becomes more acute where action is taken in response to predictions based on correlations. For example, where police organizations rely on data analytics to guide the deployment of police resources, an understanding of causation is essential in predicting the impact of such interventions. Indeed, this is implicit in the definition of causation offered by Pearl (2000: 347): ‘Y is a cause of Z if we can change Z by manipulating Y, namely, if after surgically removing the equation for Y, the solution for Z will depend on the new value we substitute for Y.’ Suppose Z represents crime rates in an area and Y represents the factor that police wish to manipulate, then their efforts will only be successful if Y is not merely correlated with Z but is itself a cause of Z. Of course, causation is rarely a simple matter in criminology or social science. There are different types of causes – whether the immediate reason for a particular event (proximal cause: ‘the gunshot caused the injury’) or a factor that operates (with other factors) to increase the chance of a particular event (distal cause: ‘smoking causes cancer’). Nevertheless, *if* a decision based on correlations is made with the goal of altering outcomes, then it is implicitly assumed that a cause (in Pearl’s sense) has been identified.

Thirdly, decisions based on historical correlations will generate feedback loops that may weaken or undermine the goal of the intervention (Harcourt, 2007). In crime prevention, feedback can occur where there are negative consequences for offenders to engage in behaviours that are found to be correlated with crime, this may lead to avoidance of those behaviours (without changing the likelihood of engaging in criminal conduct). For instance, if police focus surveillance on people with tattoos (because of a purely hypothetical finding that tattoos correlate with crime), people (particularly those planning to commit crime) may decide not to get tattoos without changing their propensity to commit crime. Because there is no causal link, the intervention (or knowledge of the potential intervention) can change behaviour in ways that reduce the extent to which the correlation will continue into the future. Other purely correlative conduct (such a membership of a club, wearing particular clothes or driving a particular vehicle) is susceptible to loss of relevance in the future due to this kind of feedback effect.

This kind of problem can be corrected by continuously updating the data from which correlative inferences are drawn, but this is not the case with a second type of feedback problem. Police deployment decisions not only affect the likelihood a crime will be committed in a particular place and the likelihood that a crime committed in a particular place will result in an arrest, but also the likelihood that a crime committed in a particular place will be recorded. Data collection is thus distorted by police deployment decisions which, if they are based on 'hot spot policing', are systematically skewed. This may perpetuate the perceived status of a location as a 'hot spot' even where this is no longer justified by actual crime rates in the area. If there are initial biases in the crimes noticed by police (for example racial bias affecting initial patrolling patterns), then this will be fed back into the algorithm and 'confirm' police biases (see Pasquale 2015). This feedback effect is self-perpetuating, thus unable to be solved by regular updating.

Harcourt (2007) raises another feedback problem. Profiling of geographical areas and communities presents dangers that whole areas or communities are stigmatized. Such stigmatization may lead initial statistical correlations to become a self-fulfilling prophecy, not only perpetuating stereotypes but in fact increasing the rate of crime. Again, this is due to the fact that changing police deployment constitutes an intervention in the system. In a full causal model, the impact of such police deployments would be recognized as an additional potential cause of crime itself, mediated by community attitudes towards police and minority group stigmatization.

Finally, as we noted earlier, nothing is atheoretical and data analytics carries with it its own implicit, generally unexamined, assumptions. For example, in the case of software products sold to police departments, there is an inbuilt assumption that the appropriate response to data is to make particular *kinds* of interventions. Thus, a program may focus on location data in order to suggest police deployments for each day. Or it may include data on the timing of deployments (Townsley and Pease, 2002). Either way, there is an assumption that changing police deployments is the best *causal* factor to toggle in order to reduce crime rates. Even data updates within the system will not eliminate that assumption. The assumption may be justified, but generally only because a theory (for example, deterrence theory) suggests that it is justified. Even there, there is no comparison with alternative policy interventions that *might*, given knowledge of the causes of crime in an area, be more effective. For example, one might be better deploying resources to enhance street lighting, changing alcohol lock-out times or building youth centres. Indeed, as stated above, in a broader causal model, increasing deployments to a 'hot spot' area might *increase* crime rates due to enhanced stigmatization and hostility towards police. These questions need to be addressed by a theoretical understanding developed outside the software itself.

These types of problems also apply to Berk's (2012) forecasting of recidivism among parolees. Berk gives the hypothetical example of shoe size being correlated with re-offending and argues that, if the correlation exists, then there is no reason not to use it as a predictor. We would argue that an understanding of the reasons for such a hypothetical correlation would be important. If, in fact, the correlation with shoe size arose due to an independent cause of both foot size and reoffending potential (such as growth hormone levels), then refusing parole may be less cost-effective than subsidizing medical treatment in balancing the costs of imprisonment against the risks to public safety. If the differences in reoffending rates due to shoe size had an alternative cause (such as a greater historical inclination to release prisoners with large shoe size early due to shoe shortages in prison stocks), then the correlation is completely misleading as a basis for justifying future parole decisions.

We would therefore argue that criminal justice decisions or interventions ought to take account of more than even accurate information about correlation. They require criteria like Pearl's (2000) definition of causation in order to predict the effect of interventions. This often calls for an understanding of *mechanisms.*[1] A mechanism 'translates causes into effects' (Stinchcombe, 2005:238) or explains the 'cogs and wheels' that provide a plausible account of how inputs and outputs are linked (Hedstrom and Swedberg, 1988: 7). An understanding of the mechanism connecting two variables both helps us distinguish correlation from causation and enhances understanding of the reasons for the causal signal (1988: 9). This deeper theory is necessary if we wish to explain, for example, how individual characteristics, experiences and environmental factors interact in leading the person to commit a crime, which in turn helps to *explain* individual differences with respect to place or differences over time in criminal activity (Wikstrom, 2006).

Not only does an understanding of causes and mechanisms provide a more rational basis for decisions to introduce interventions, it is also likely to be important to decision-makers. The disregard of theory means that a lot of Big Data analysis (like the random forest algorithm described above) will be regarded as a 'black box'. As Brennan and Oliver (2013: 558) state, 'their logic can be inscrutable to human users… This failure to offer explanatory tools might be awkward for decision makers who must provide justifications for their decisions (e.g., judges, probation officers, and parole boards)'. For some decision-makers, the need to explain their decisions may lead to a preference for *understanding* even aside from concerns about the limited ability to predict the effect of interventions based on historical correlations.

This suggests that it is important, not only that algorithms are built with reference to theory, but that the design of the algorithms and the assumptions (derived from theory) on which they are built are made transparent (Bennett Moses and Chan, 2014).

A criminology that aspires to usefulness in policy and law enforcement must therefore continue to engage with theory. The kind of knowledge generated by machine-learning techniques can be useful, but it is limited. In particular, the effectiveness of interventions (by policy-makers or police) based on pure correlations cannot be predicted *ex ante*. Thus even well-funded programs that have relied on correlative 'risk factors' rather than causes of antisocial behaviour have had only marginal success (Moffitt and Caspi, 2006). Instead, an *ex post* evaluation of the effectiveness of an intervention is required. Even strong advocates of machine-learning analytics concede that 'if the goal is to use one or more risk factors to design and test interventions, then many would argue that the only sound approach is randomized experiments or very strong quasi-experiments' (Berk and Bleich, 2013:517). These experiments or quasi-experiments would need to be designed with a sufficient understanding of diverse causes and mechanisms to factor out feedback effects and compare alternative policy interventions. If interventions *are* consistently effective, one can then measure improvements against costs, including the social costs of increased data retention and surveillance. Should the balance favour intervention based on predictive analysis, the challenge for criminology will be to articulate an explanation, with reference to causes and mechanisms, explaining why this is the case.

Conclusion

Commentators such as Anderson (2008) and Mayer-Schönberger and Cukier (2013) have claimed that Big Data will lead to three potential challenges to scientific methods: that disciplinary expertise will 'lose their monopoly on making sense of empirical social data', that disciplinary experts will be replaced by generalist data analysts, and that scientific practice will be transformed as causality 'is being knocked off its pedestal as the primary foundation on meaning' (2013:30, 67). This paper has examined these claims in relation to the practice of criminology.

The first two claims are linked and they might be dismissed on the grounds that criminological scholarship has always been multidisciplinary. Criminology has always accommodated a diversity of intellectual influences and investigative methods. While Big Data may require the use of programmers and statisticians to handle the technical side of data storage and data analysis, such reliance on technical personnel in the past has not led to the demise of criminology. It is possible, of course, for generalist data analysts to become specialists in the analysis of criminal justice data, but past experience has shown that the growth in number of such specialists does not threaten the authority or status of criminological researchers. In fact, the most serious challenge to criminology has already happened 15 years ago with the birth of 'crime science' which self-consciously and deliberately dissociates itself from the social and sociological aspects of criminology (see Laycock 2005; Clarke 2010), concentrating on the use of science for crime reduction. Tim Hope (2011) has suggested that, at least in Britain, the voices of 'official criminology and crime science' are increasingly being legitimated while those of 'academic criminology' are being controlled or silenced. This may well happen here if governments continue to cut back on funding for academic and basic research. Nevertheless, as Brennan and Oliver (2013:556) have pointed out, criminal justice agencies ‘are often highly resistant to change and to new or novel methods that require substantial new learning, that are unfamiliar, or that require abandoning familiar methods’.

The third claim is perhaps the most serious challenge and it has been the main focus of this paper. Our review of the literature has shown that it may be too early to tell whether Big Data will dramatically transform the practice of criminology, because at the time of writing very few researchers are actually using Big Data. Nevertheless, two emerging trends deserve attention: first, the use of Big Data as one type of research data and secondly, the use of computer modeling/algorithms as predictive tools. Our analysis suggests that the use of Big Data as one type of research data is unlikely to transform the practice of criminology except new analytical techniques are required to process the very large volumes of data. Even the use of computer modeling and analytics does not by and large constitute a significant departure from past criminological practices, but there is one exception: the use of machine-learning procedures in predictive analysis is one area where established ways of doing criminology may well be threatened. This relates to the claim that, with Big Data, causal theory is irrelevant because correlation is sufficient.

Our discussion of machine-learning procedures and the attendant dismissal of causal theories shows that there can be serious problems with such an approach. Not only are predictions based solely on algorithm-derived correlations opaque and difficult to interpret (and hence difficult to justify to stakeholders), when such predictions are used to guide decisions or interventions, problems of instability, feedback loops, unintended consequences or injustices can undermine the credibility or viability of the predictions.

Brennan and Oliver (2013:556-) have predicted that machine-learning techniques might well be 'overlooked, ignored, or undervalued by criminological researchers and practitioners', given that criminal justice agencies have a 'history of weak adoption, poor implementation, and rejection of technical advances in risk assessment and classification', the practical and political problems of a 'black box' technology which decision makers find 'inscrutable', and the lack of competency among researchers and practical users in such technical methods. Only time will tell whether or not this is indeed what will happen, but judging by the enthusiastic uptake of predictive policing tools by police organizations, especially in the US (see Perry et al., 2013), even though the 'statistical techniques used in predictive analytics are largely untested and have not been rigorously evaluated' (Uchida 2013:3878), Brennan and Oliver may have been unduly pessimistic. It is safe to say that, when the time comes for these techniques to be independently and rigorously evaluated, the expertise of 'traditional' criminology will be in great demand.

Kitchin (2014a:10) is probably right to suggest that while Big Data provides researchers with new sources of data, new approaches to the generation, analysis and visualization of such data, and enables new questions to be asked, it is unlikely to lead to new disciplinary paradigms in the humanities and social sciences because 'it is unlikely that suitable Big Data will be produced that can be utilized to answer particular questions, thus necessitating more targeted studies'. Instead of replacing traditional small data studies, Big Data will present opportunities for scholars to work with 'massive quantities of very rich social, cultural, economic, political and historical data' (2014a:10). Such opportunities will also present new challenges about how to analyze and make sense of such empirical data within a reflexive framework. For criminology, this is not necessarily a new challenge, as researchers have regularly encountered new approaches (often coming from other disciplines) and often need to negotiate contradictory demands and make difficult political choices.

Acknowledgement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Abbott A (1998) *The Systems of Professions: An Essay on the Division of Expert Labor.* Chicago: University of Chicago Press.

Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*, 23 June 2008. Available at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 17 July 2014).

Bennett Moses L and Chan J (2014) Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools. *UNSW Law Journal* 37(2): 643-678.

Berk RA and Bleich J (2013) Statistical Procedures for Forecasting Criminal Behavior. *Criminology & Public Policy* 12(3): 513- 544.

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

Bosworth M and Hoyle C (2011) (eds) *What is Criminology?* Oxford: Oxford University Press.

boyd d and Crawford K (2012) Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, Communication and Society* 15(5): 662-679.

Brennan T and Oliver WL (2013) The Emergence of Machine Learning Techniques in Criminology. *Criminology & Public Policy* 12(3): 551-562.

Chan J (2000) Globalisation, Reflexivity and the Practice of Criminology. *Australian and New Zealand Journal of Criminology* Millennium Issue 33(2):118-135.

Clarke RV (2010) Crime science. In: McLaughlin E and Newburn T (eds) *Handbook of Criminological Theory*. London: Sage, pp. 271-283.

Edwards A, Housley W, Williams M, Sloan L and Williams M (2013) Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodologies.* 16(3):245-260.

Felson M and Clark R (1998) *Opportunity Makes the Thief: Practical theory of crime prevention.* Home Office, UK.

Gerber, MS (2014) Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*. 61(2014):115-125.

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

Groff ER and La Vigne N G (2002), Forecasting the Future of Predictive Crime Mapping. In: Tilley N (ed) *Analysis for Crime Prevention.* Criminal Justice Press, Willan Publishing, pp. 29-57

Harcourt BE (2007) *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. Chicago: University of Chicago Press.

Hedström P and Swedberg R (1988) Social mechanisms: An introductory Essay. In: Hedström P and Swedberg R (eds) *Social Mechanisms: An Analytic Approach to Social Theory*. Cambridge: Cambridge University Press, pp. 1-31

Hope T (2011) Official criminology and the new crime sciences. In: Bosworth M and Hoyle C (eds) *What is Criminology?* Oxford: Oxford University Press, pp. 456-474.

Jacob J (1961) The Death and Life of Great American Cities. Random House.

Johnson SD and Bowers KJ (2004) The Stability of Space-Time Clusters of Burglary. *The British Journal of Criminology* 44(1): 55-65

Jurgenson N (2014), View from Nowhere, *The New Inquiry*. Available at http://thenewinquiry.com/essays/view-from-nowhere/ (accessed 23 January 2015).

Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1: 1-12.

Kitchin R (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London, Thousand Oaks, New Delhi, Singapore: Sage.

Laycock G (2005) Defining Crime Science. In: Smith MJ and Tilley N (eds) *Crime science: new approaches to preventing and detecting crime*. Uffculm: Willan Publishing, pp. 3–24.

Leonelli S (2014) What difference does quality make? On the epistemology of Big Data In biology. *Big Data & Society* 1: 1-11

Lever R (2014) Researchers use Twitter to predict crime, *Agence France Presse* , 20 April 2014.

Malleson N and Andresen MA (2014) The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science,* DOI: 10.1080/15230406.2014.905756

Manyika J et al. (2011) Big data: The Next Frontier for Innovation, Competition and Productivity. McKinsey Global Institute. Available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed 23 January 2015).

Mayer-Schönberger V and Cukier K (2013) *Big Data: A revolution that will transform how we live, work and think*. London: John Murray.

Moffitt T & Caspi A (2006) Evidence from behavioral genetics for environmental contributions to antisocial conduct. In: Wikstrom, P-OH and Sampson RJ (eds) *The Explanation of Crime: Context, Mechanisms and Development* . Cambridge: Cambridge University Press, pp. 108-152.

Morgan S & Winship C (2007) *Counterfactuals and Causal Inference: Analytical Methods for Social Research*. Cambridge University Press.

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

Pearl J (2000) *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Perry WL, McInnis B, Price CC, Smith SC and Hollywood JS (2013) *Predicting Policing: The Role of Crime Forecasting in Law Enforcement Operations.* Rand Corporation, available at www.rand.org (accessed 17 December 2014).

Procter R, Vis F, and Voss A (2013) Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology* 16(3):197-214.

Steadman I (2013) Big data and the death of the theorist, *Wired,* 25 January 2013. Available at http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory (accessed 17 July 2014).

Stinchcombe AL (2005) The logic of social research. Chicago: University of Chicago Press.

Townsley M and Pease K (2002) Hot Spots and Cold Comfort: The Importance of Having a Working Thermometer. In: Tilley N (ed) *Analysis for Crime Prevention* Crime Prevention Studies 13. Criminal Justice Press, Willan Publishing, pp. 59-69

Traunmueller M, Quattrone G, and Capra L (2014) Mining mobile phone data to investigate urban crime theories at scale. In: Aiello LM and McFarland D (eds) *Social Informatics. Lecture Notes in Computer Science*. Springer International Publishing, pp. 396-411.

Uchida, CD (2009) *Predictive Policing in Los Angeles: Planning and Development.* Justice and Security Strategies Inc. Available at http://newweb.jssinc.org/wp-content/uploads/2012/01/Predictive-Policing-in-Los-Angeles.pdf (accessed 23 January 2015).

Uchida CD (2013) Predictive Policing. In: Bruinsma G and Weisburd D (eds) *Encyclopedia of Criminology and Criminal Justice.* Springer 3871-3880.

Uprichard E (2013) Focus: Big Data, Little Questions? *Discover Society*, 1 October 2013. Available at http://www.discoversociety.org/2013/10/01/focus-big-data-little-questions/ (accessed 17 July 2014).

Watters PA and Phair N (2012) Detecting illicit drugs on social media using Automated Social Media Intelligence Analysis. In: Xiang Y et al. (eds) *Cyberspace Safety and Security Lecture Notes in Computer Science*. Berlin Heidelberg: Springer, pp. 66-76*.*

Wikstrom P-OH (2006) Individuals, settings, and acts of crime: situational mechanisms and the explanation of crime. In: Wikstrom, P-OH, Sampson, RJ (eds) *The Explanation of Crime: Context, Mechanisms and Development.* Cambridge: Cambridge University Press, pp.61-107

Williams ML, Edwards A, Housely W, Burnap P, Rana O, Avis N, Morgan J and Sloan L (2013) Policing cyber-neighbourhoods: tension monitoring and social media networks. *Policing & Society* 23(4):461-481.

Williamson B (2014) 'The death of the theorist and the emergence of data and algorithms in digital social research', *The Impact Blog,* 10 February 2014. Available at http://blogs.lse.ac.uk/impactofsocialsciences/2014/02/10/the-death-of-the-theorist-in-digital-social-research/ (accessed 17 July 2014).

Notes

[1] According to Wikstrom (2006), a full scientific explanation for a phenomenon is a theory or hypothesis that is consistent with relevant statistical correlations, is supported by experimentation (enabling inferences of causation) and provides explanation by suggesting plausible mechanisms.