BIG DATA AND LAW ENFORCEMENT - UK AS A CASE STUDY

[1]DAN BOUHNIK, [2]STEVEN YAKOBASHVILI

[1,2]Bar Ilan-University, Ramat Gan, 5290002 Israel
[1]Jerusalem College of Technology, Jerusalem, 9116001 Israel
E-mail: [1]Dan.biu.rg@gmail.com, [2]stiven54@gmail.com

Abstract - Big Data is referred to as an analytical process in which a large number of basic units are processed to produce a finished project.It can incorporate a wide range of analytical techniques, & the data used can take many different forms originating from virtually a limitless number of potential sources. The majority of information that organizations hold is in an unstructured format, such as free text, images and video. The system is able to extract meaningful information from both structured and unstructured data. This allows the data to be interrogated in a much deeper way than with traditional data science methods, yielding significantly richer products. The purpose of this paper is to demonstrate the different uses of this process in the law enforcement field, while focusing on different methods of crime analysis. While the gathering of information gets more and more efficient using this process, as demonstrated hereinafter the concerns of using the methods affiliated with it are raising concerns on the objectivity of the data collect. This paper demonstrates the main issues of the process when applied in law enforcement agencies worldwide, while focusing on the UK and the US as case studies.

Keywords - Big Data, Law Enforcement, Crime Analysis.

I. INTRODUCTION

Big data refers to a process that is used when traditional data mining and handling techniques can't uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach known as big data, which incorporates massive parallelism on readily-available hardware.

The purpose of this approach is to answer questions, solve problems and provide a broad view out of the raw analyzed data. The term 'Big' data shouldn't be only referred to as a defined size or scale, but rather the capacity to search, process, and cross-reference large data sets.
In other words, big data analytics (advanced analytics) becomes necessary when data is collected on such a large scale that it cannot be analyzed with traditional data-management tools and methods. A key feature of advanced analytics is the use of algorithms, which increasingly incorporate Artificial Intelligence (AI) methods underpinned by machine learning. As AI entails that the machine processing the data learns new rules through experience, the processing methods and calculations involved are often opaque to a human observer.[1]Big data can incorporate a wide range of analytical techniques[2], and the data used can take many different forms, originating from a virtually limitless number of potential sources. The majority of information that organizations hold is in an unstructured format, such as free text, images and video. A key distinction between advanced analytics and traditional data analytics is that the latter is unable to make sense of unstructured data. The former, however, is able to extract meaningful information from both structured and unstructured data. This allows the data to be interrogated in a much deeper way than with traditional data science methods, yielding significantly richer products.

Major retail corporations use advanced analytics to inform pricing strategies and inventory control, as well as other uses as demonstrated in this paper.

Despite its wide and adaptive capabilities, big data is not without limits and restrictions. Advancedanalytics is by no means objective, but rather relies on a considerable amount of interpretation. This interpretation, as demonstrated hereinafter is a major risk factor as per the law enforcement field requires.
For instance, the data-cleaning process – which involves the manual elimination of variablesand data points that are deemed either irrelevant or unreliable – is inevitably subjective.

Additionally, datasets used for advanced analytics are often unreliable, containing datafrom disparate sources, which has been collected using varied and sometimes questionablemethods. These methods are very often not controlled by the legal system, and the methods may very well be considered illegal, if detected and investigated.
In many cases, the data used originates from the internet, and large data sets fromInternet sources are often unreliable, prone to outages and losses, and these errors and gapsare magnified when multiple data sets are used together[3], to sometimes what's considered to be a pre-defined conclusion.

As the amount of available data and the number of potential sources increase, so too doesthe potential for error – and particularly the possibility of detecting relationships where noneexist (false positives).

When analyzing these false positives, the resultscan serve as an important reminder that a failure to understand the inherentlimitations present in a dataset can easily lead to misinterpretation, and that data analysis ismost effective when researchers take account of the complex methodological processes thatunderlie the analysis of that data.[3]

This is especially true in the context of big data, where evena small oversight can have a dramatic impact on the findings.Another limitation is the high technical requirement of big data analytics, both in terms ofcomputing power and storage. Data complexity has evolved drastically in both volume andstructure, and manyorganizations do not have the infrastructure required to store and processdata at the rate at which it is now being collected.

Traditional data warehouses are not sufficientto meet the requirements of advanced analytics and so organizations must instead invest increating data pools – centralized data stores that can accommodate all forms of data, bothstructured and unstructured.

Due to the prohibitive costs involved, it is unfeasible for public sector organizations to invest in dedicated in-house infrastructure to perform big data analytics, leading to an increase inpartnerships with third-party organizations that provide the service within a secure privatecloud environment or private data center, a model known as 'Infrastructure as a Service' (IAAS).
Another key limitation lies in the legal and ethical constraints on the use of big data.As demonstrated further on, the use of outsourcing services to preform sensitive large scale collection of data raises many unsolved ethical issues. For example, are we sure that the data won't be used by the third party company to further benefitfrom the collected data for its own use? After all, we all must note that big data is big business, moneywise. Forrester predicts the global Big Data software market will be worth $31B this year, growing 14% from the previous year. The entire global software market is forecast to be worth $628B in revenue, with $302B from applications..[4] These figures has are usually business oriented, and not always legally and constitutionally restricted.

Many public sector organizations are understandably apprehensive of storing sensitive data inthe cloud or in off-premise facilities, especially given the restrictive nature of data-protectionlegislation. This is especially true for the police, who handle large volumes of personal data.However, the security of private cloud networks has improved considerably over time, and theuse of cloud-based data services is increasingly becoming the norm in both the private andpublic sectors.

Thelegislation governing digital surveillance, interception of communications and other means ofdata collection – such as the Regulation of Investigatory Powers Act 2000 and the InvestigatoryPowers Act 2016 are currently under development as the countries are trying to balance the legal aspects of mass information gathering and the goal of analyzing huge amounts of raw data.Data protection legislation restricts large-scale exploratory analysis of personal data, such as the predictive risk assessment methods discussed earlier. For this reason, pseudo nominationis likely to be the only way to perform big data analytics on personal datasets while complying with data protection laws.

II. CRIME ANALYSIS

While big data has been employed in a diverse range of domains, its use in policing has beenlimited. The three main ways in which the police use big data are through the collection andstorage of DNA information, mass surveillance, and predictive policing.

Police forces have thecapability to cross-reference DNA samples against a database of millions of other records with ahigh degree of accuracy. However, the police's ability to trawl and analyze large volumes of datacollected through mass surveillance is far more limited. This is because DNA information is storedas numbers, which is far more straightforward to analyze than large volumes of unstructureddata, such as images and video.

Predictive policing is defined as 'taking data from disparate sources, analyzing them and thenusing results to anticipate, prevent and respond more effectively to future crime'[5]. Predictivepolicing is based on the notion that analytic techniques used by retailers to predict consumerbehavior can be adapted and applied to policing to predict criminal behavior.
The use of quantitative analysis to make predictions about crime levels is by no means a newapproach. In the 1990s, the New York Police Department (NYPD) was at the forefront ofthe intelligence-led policing revolution.[6]

2.1 CompStat: Ahead of time invention

The CompStat (Compare Statistics) system, was used to track spatial crimepatterns and identify hotspots by sticking pins in maps. Weekly CompStat meetings forcommanding officers became compulsory. The

system allowed the performance of eachofficer's precinct to be quantitatively measured, providing a level of accountability never beforeseen in law enforcement.[7]

The dramatic reductions in crime rates that followed within precincts that implemented CompStatled almost every law enforcement agency in the US to adopt the practice of automated mappingand statistical analysis of crime data.

The practice of hotspot policing – where crime hotspotsare identified using spatial analysis and police activity is targeted to these areas accordingly – has been shown to be a highly effective crime prevention strategy in a range of differentenvironments and for many different crime types.

The technological capabilities of the police have advanced immeasurably since the days ofCompStat, and law enforcement agencies worldwide now have access to both a much greatervolume of useable data and far more sophisticated and efficient methods of analysis.

The innovation of predictive policing is the application of artificialintelligence to such large data sets. CompStat relied heavily on thecollection and display of past crime data; predictive policing appliescomputer analysis to similar information. The identification of futuregeographic places likely to be targeted by criminals has attracted themost attention.
These predictive models all rely on well-documentedobservations about the spatial distribution of criminal behavior. Crime isnot found randomly across a city, but rather tends to fall within limited, and often very small, areas. (Crime tends to be "lumpy.") For example,researchers found that over a fourteen-year period, about fifty percent ofthe crime in Seattle was limited to 4.5 percent of the city's streetsegments. [15] Based upon this connection between crime and place,computer models adopt different approaches towards the prediction ofcrime.

III. UK AS A CASE STUDY

All police forces in the UK have access to the Police National Computer (PNC), a nationwide database, which as of May 2017 contained more than 12.2 million personal records, 62.6 million vehicle records and 58.5 million driver records.

This is distinct from the Police National Database (PND), a national intelligence-handling system containing copies of locally held police records. PND was created in 2011 to allow police officers to search across the 220 different databases operated by individual police forces in the UK. Some of these local databases hold huge amounts of data: the Metropolitan Police Service (MPS) in London operates an Automatic Number Plate Recognition (ANPR) network which receives approximately 38 million reads a day.[8]

As well as PNC and PND, all UK police forces also have access to Ident1, the central national database for biometric information, which contains more than 7 million fingerprint records, and the National DNA Database, which holds the DNA information of more than 5 million individuals.

In 2015, the UK connected into the second-generation Schengen Information System(SIS II), a European-wide governmental database used for national security, law enforcement and border control.

Researchersfound that UK police forces have access to a vast amount of digital data, butcurrently lack the technological capability to use it effectively. With mass surveillance and thelarge-scale collection of data now becoming a matter of daily routine, the challenge aheadlies in devising accessible, affordable systems that can be used to process and analyze thesequantities of data efficiently and reliably.

3.1 Databases: A success or a mess?

As in many public service domains, law enforcement data is typically stored across multiple fragmented databases, with no unified system for querying several databases simultaneously.

Officers are required to input the same data multiple times, accessing several individual systems to manually collate information and manage cases.

Data applications operate in isolation, with no facility to copy information between different systems. This problem has created demand for systems that allow users to access and input information across multiple databases simultaneously.

Another issue noted by officers is the lack of a standardized lexicon or naming conventions for police databases. For example, an officer recording a firearms offence may use the words 'shoot', 'fire' or 'discharge' when entering case information into a database. This complicates the search process for investigators and analysts who are later trying to retrieve such data.

IV. PREDICTIVE HOTSPOT MAPPING

The occurrence of advanced analytics has enabled the development of sophisticated predictive crime mapping tools that use statistical models to identify

areas that are at increased risk of experiencing crime, based on past criminal events. The practice of using past crimes to predict future crime is founded on the observation that repeat victimization accounts for a large proportion of all crime. In addition, crime is often contagious, with the risk of victimization for crimes such as burglary increasing for houses near a burgled property in the aftermath of the initial crime.[9]

The system initially processes several years of crime data to lay down a 'background' level of crime, using an Epidemic Type Aftershock Sequence model – a self-learning algorithm based on seismological models that predict earthquake aftershocks.[10] The software uses three data points as the input for forecasting: crime type; crime location; and crime date and time. These are deemed to be the most objective variables, since they do not include any personal data.

Using self-exciting point process modelling (a time series model in which the occurrence of past points makes the occurrence of future points more probable), the data is used to pinpoint small (500 x 500 feet) boxes that indicate the times and places where crime is most likely to occur.

These boxes are automatically generated for each shift of each day, allowing officers to respond to potential crime locations in real time.

One of the main attractions of the software is that it is affordable and requires minimal training.

4.1 Is the mapping "too hot"?

Software such as PredPol generates grid-based predictions, which are indifferent to the geographical landscape. More recent research has shown that a calibrated network-based model that generates street segment predictions delivers a significantly higher level of predictive accuracy than traditional grid-based methods. Street segment predictions are more useful for policing purposes than arbitrary grids, and these promising findings suggest that further research is needed to refine such models and integrate them within existing predictive policing software.

Another shortcoming of tools such as the aforementioned, is that they rely entirely on the analysis of past events to predict future crime. They do not take into account the many complex and dynamic environmental factors that make certain places more vulnerable to crime.
It is important to also bear in mind that the use of predictive mapping results in a prioritization of resources that is based simply on crime volume, rather than harm caused. In any given area, crimes that occur most frequently will receive a greater police response than those that occur less frequently, despite some crimes causing more harm than others.
When considering the issues of the matter, these 3 main concerns should be considered:

- Software and applications can be abused - We don't know how data is inputted or analyzed
- Was the data manipulated?
- Are we developing a reduced reasonable suspicion standard?

V. PREDICTIVE RISK ASSESSMENT OF INDIVIDUALS

The police's current use of data analytics is heavily location-based, but the uses of big data extend far beyond spatial analysis.
Data can also be used to calculate the risks associated with particular individuals.

Forces routinely collect information on known offenders, and this data could prove valuable in identifying potential repeat offenders, especially when analyzed in combination with data from partner agencies.

In 2016, the Supreme Court of Wisconsin used a risk assessment tool – COMPAS – to conclude that the defendant in question posed a great enough risk to society to be ineligible for probation. Another example is Durham Constabulary in Germany that is currently in the process of developing an AI-based system to assess the risk of individuals reoffending. The Harm Assessment Risk Tool (HART) uses as its input variables an individual's past offending history, their age, postcode and other background characteristics. Predictive algorithms are then used to classify each individual as being at low, medium or high risk of re-offending.[11]

Researches reflects the fact that the model favors classifying individuals as medium or high-risk in order to reduce the likelihood of false negatives.

Another issue to address is how predictions should be acted upon when an individual is identified as posing an increased risk. Acting upon such predictions may result in negative social effects, such as perpetuating bias or racial discrimination. [12]

A ProPublica investigation into the COMPAS sentencing algorithm found that only 20% of individuals identified as likely to commit a violent crime actually did so, and that black defendants were twice as likely to be deemed at risk of offending than white defendants.

Systems underpinned by machine learning will inevitably reproduce the inherent biases present in the data they are provided with – if particular minorities have been disproportionately targeted by police action in the past, the algorithm will disproportionately assess those individuals as posing an increased risk. Acting on these predictions will then result in those individuals being disproportionately targeted by police action, creating a 'feedback loop' by which the predicted outcome simply becomes a self-fulfilling prophecy. [12,13]

VI. LEGAL USE OF DATA

Data protection legislation restricts large-scale exploratory analysis of personal data, such as the predictive risk assessment methods discussed earlier. For this reason, pseudonominationis likely to be the only way to perform big data analytics on personal datasets while complying with data protection laws.
The use of advanced analytics has the potential to increase anonymity in this regard, as algorithmic machine learning-based systems do not require the input of a human observer during the analytical process. Data can be anonymized prior to analysis, and the identities of individuals of interest only revealed once the analytical process has concluded.

6.1 UK as a case study

The UK is one of the world's foremost surveillance states, with one of the highest numbers of CCTV cameras per capita of any country, and – as of November 2016 – the most comprehensive powers of digital surveillance ever seen in a democracy .

The Investigatory Powers Act 2016 introduced new powers allowing the UK intelligence agencies and police to carry out not just targeted interception of communications, but also indiscriminate 'bulk' collection and analysis of communications data. It also requires CSPs to store all UK citizens' internet connection records and provide these records on request to the police, intelligence agencies and other authorities without requiring a warrant.

In December 2016, the European Court of Justice ruled that the general and indiscriminate collection of electronic communications is illegal under EU law.

However, it is at present unclear whether these provisions of the Act are currently beingimplemented by authorities in the UK.[14,15]

VII. SUMMARY

As noted in this paper, the uses of big data are much wider than it was 10-15 years ago. The development of information gathering, storage facilities and the vast acceleration of cloud based solutions have made the data mining and conclusion process much more effective than in the past.

Yet, we should all be worried about the fact that many aspects of information security are not always considered, and sometimes the use of outsourcing companies for efficiency purposes are lacking the concern of data breach, and privacy issues as well.
More effective and controlled tools must be developed, so that the law enforcement agencies will be able to more wisely use their vast capabilities to prevent crime and terror on one hand, and on the other hand be able to provide a safer working environment for the raw mining of data and for producing useful "workable" information from it.

BIBLIOGRAPHY

[1] UK Information Commissioner's Office Homepage, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection', 9 (2017).
[2] James Manyika et al.: 'Big Data: The Next Frontier for Innovation, Competition, and Productivity', pp. 27–31, McKinsey Global Institute (2011).
[3] Boyd, Danah, and Kate Crawford.: "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society, 15(5), 662-679 (2012).
[4] Forrester:Midyear Global Tech Market Outlook for 2017 To 2018 (2017).
[5] Pearsall, B.: Predictive Policing: The Future of Law Enforcement? National Institute of Justice. 16 (2010).
[6] Bureau of Justice Assistance, US Department of Justice, and Police Executive Research Forum,
[7] Police Executive Research Forum: "Compstat: Its origins, evolution, and future in law enforcement agencies." (2013).
[8] Metropolitan Police, 'One Met: Total Technology (2014).
[9] Bowers, Kate J., Shane D. Johnson, and Ken Pease.: "Prospective hot-spotting: The future of crime mapping?" British Journal of Criminology 44, no. 5, 641-658 (2004).
[10] Johnson, Shane D.: "Repeat burglary victimization: a tale of two theories." Journal of Experimental Criminology 4, no. 3, 215-240 (2008).
[11] Bachner, Jennifer: Predictive policing: preventing crime with data and analytics. IBM Center for the Business of Government, 26 (2013).
[12] Cope, Nina: "'Intelligence Led Policing or Policing Led Intelligence?' Integrating Volume Crime Analysis into Policing." British journal of criminology 44.2 188-203 (2004).
[13] Ferguson, Andrew Guthrie.: "Big data and predictive reasonable suspicion." U. Pa. L. Rev. 163, 327 (2014).
[14] Alexander Babuta, Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations and Priorities, Royal United Services Institutefor Defense and Security Studies(2017).
[15] Anthony A. Braga et al.: The Relevance of Micro Places to Citywide Robbery Trends: A Longitudinal Analysis of Robbery Incidents at Street Corners and Block Faces in Boston, 48 J. RES. CRIME & DELINQ. 7, 10 (2011) (citing David L. Weisburd et al., Trajectories of Crime at Places: A Longitudinal Study of Street Segments in the City ofSeattle, 42 CRIMINOLOGY 283 (2004).

★★★