FROM TRAGEDY TO STATISTIC: HOW BIG DATA HAS CHANGED THE PRACTICE OF LAW

BRYONT CHIN HAN MING

Data is information, and information never stops growing. In 2012, 2.5 exabytes[1] of data were generated every single day.[2] 90% of the data in the world today has been created in the past two years.[3] In Singapore, there is currently scant regulation pertaining to cybersecurity and data use. The *Personal Data Protection Act*[4] and the *Computer Misuse and Cybersecurity Act*[5] are the Legislature's responses to these issues. These are mainly focused on preventing and penalizing cybercrime, especially intrusions into important government and private networks. However, while hacking is undoubtedly a perennial concern, in recent years many have asked another important question: how should the law respond to the increasing importance of big data and data analytics?

The term "big data" has been variously defined, but it centrally refers to "extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions".[6] Multinational companies and governments can collect these data sets, analysing them to generate insights which cannot typically be gleaned through traditional data analysis. Data collection, storage, and analysis is a field that is growing

---

[1] Equivalent to 2.5 billion gigabytes, or 2.5 x 10^{18} bytes.
[2] Ralph Jacobson, "2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?" (24 April 2013), IBM Consumer Products Industry Blog, online: <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>.
[3] *Ibid*.
[4] (No 26 of 2012).
[5] (Cap 50A, 2007 Rev Ed Sing).
[6] From *The Oxford English Dictionary*, 2nd Ed, *sub verbo* "big data".

very important very quickly: according to International Data Corporation, a market intelligence and analytics firm, worldwide revenues for big data and business analytics will grow from $130.1 billion in 2016 to more than $203 billion in 2020.[7]

A. ACCELERATING LEGAL PRACTICE

The meteoric rise in importance of this field has smashed existing practices in law firms. Legal practice is one of the most information-heavy industries, with the huge volume of precedent cases, regulations, case analyses, and legal commentaries. It used to be common to see armies of lawyers and paralegals sifting through information as part of due diligence. For the discovery process in *United States v. CBS Inc.*,[8] the studios examined six million documents for more than $2.2 million—much of it going to lawyers and paralegals who worked for months at high hourly rates.[9] But that was in 1978. Today, thanks to advances in artificial intelligence, "e-discovery" software can analyse documents in a fraction of the time, for a fraction of the cost. The Californian data analytics firm Blackstone Discovery has software that can analyse 1.5 million documents for less than $100,000, less than a fifth of what it would have cost thirty years ago.

The possibilities of data analytics for lawyers go beyond just search engines; even something as basic as the pricing of legal services has been completely changed. Clients and lawyers now have access to huge databases about prices that firms charge across the jurisdiction. The TyMetrix LegalView service continuously aggregates tens of billions of dollars' worth of legal invoices, allowing law firms to choose the best position for themselves: an accessible low-cost provider or a high-end premium law firm.[10] Another firm, Sky Analytics, offers companies a "Right Rate

---

[7] This is a compound growth rate of 11.7% per annum. See Gil Press, "6 Predictions For The $203 Billion Big Data Analytics Market", *Forbes* (20 January 2017), online: <https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#20ece4b72083>

[8] 459 F Supp 832 (CD Cal 1978).

[9] John Markoff, "Armies of Expensive Lawyers, Replaced by Cheaper Software", *The New York Times* (4 March 2011), online: <http://www.nytimes.com/2011/03/05/science/05legal.html>.

[10] Joe Dysart, "How lawyers are mining the information mother lode for pricing, practice tips and predictions", online: (May 2013) *ABA Journal* <http://www.abajournal.com/magazine/article/the_dawn_of_big_data/?utm_source=feedburner&utm_medium=feed&utm_campaign=ABA+Journal+Magazine+Stories>.

Advisor" tool, which assesses an external lawyer from many aspects to advise companies whether to accept or reject his fees.[11] Over the years, a common complaint levelled at lawyers is the opacity of their fee structures and the resultant uncertainty. With a huge amount of data now available about legal fees, data analysis systems like TyMetrix LegalView and Sky Analytics' "Right Rate Advisor" can make legal fees more transparent, making legal services more accessible for all.

B. PREDICTIVE ANALYSIS AND ABUSE IN THE CRIMINAL LAW

However, arguably the most important use of data analytics is in the criminal law. Algorithms can now piece together information from a myriad of sources – police reports, arrest statistics, surveillance camera footage, and other information generated by the police – to better understand crime activity in a particular area and, more importantly, where criminals are likely to strike next. Based on this "predictive policing" system, the police can better deploy their resources to areas that are more prone to crime, and can respond faster and more effectively when someone breaks the law.

One of the most ambitious predictive policing systems in the world is China's "Police Cloud" system. Government databases scoop up everything from addresses, to medical history, supermarket membership, and delivery records. This information is linked to each citizen's unique identification number, and is used by security bureau authorities to look for patterns in an individual's behaviour. These databases are massive: for instance, police in Jiangsu, China have amassed 780 million data points on its citizens,[12] collecting records of citizen's incomes, navigation data, and their purchases from major e-commerce companies, among other things.[13] In Shandong,

---

[11] *Ibid*. Sky Analytics has revealed that the Right Rate Advisor refers to, among other factors, the external lawyer's years of experience, his or her position in the firm, the size of the firm, and the cost of living where the lawyer is based.

[12] *What do China's police collect on citizens in order to predict crime? Everything*, online: Quartz <https://qz.com/1133504/to-predict-crime-chinas-tracking-medical-histories-cafe-visits-supermarket-membership-human-rights-watch-warns/>.

[13] Echo Huang, "China: Police 'Big Data' Systems Violate Privacy, Target Dissent", *Human Rights Watch* (20 November 2017), online: <https://www.hrw.org/news/2017/11/19/china-police-big-data-systems-violate-privacy-target-dissent>.

China, the police can access patient records,[14] names and causes of petitioners and political troublemakers,[15] and social media usernames. With the Police Cloud, even the most intimate parts of a citizen's life are open to the government.

The potential for abuse is obvious. In 2015, the Office of the Central Committee of the CCP announced their intention to embrace technology like the Police Cloud to achieve "social stability". China's Ministry of Public Security designed the Police Cloud system to surveil seven categories of "focus personnel", including petitioners, those who "undermine stability", and people "involved with terrorism".[16] Such vague definitions mean that essentially anyone could be designated a threat and placed under surveillance.

Chinese citizens do not have the right to be notified when placed under surveillance and have no legal avenues for contesting it.[17] At present, China has no privacy or data protection law protecting personal data from misuse. The police are under no obligation to obtain a court order to conduct surveillance, or provide any evidence that the people whose data they are collecting from are associated with or involved in criminal activity. There are essentially no effective privacy protections against government surveillance, giving the Chinese police nearly unchecked power.

This is especially alarming if one remembers that predictive policing algorithms are not, and never will be, perfect. Predictive policing systems can only make predictions based on past data, which may not reflect actual risk patterns. Erroneous data will also result in erroneous predictions. A California woman recently won a civil rights lawsuit[18] against the San Francisco Police Department after a number-plate reader misidentified hers as a stolen car and she was held at gunpoint by officers, forced to her knees, and detained for 20 minutes.

[14] Including names and illnesses.

[15] *Ibid.* A tender document from Tianjin boasted that its Police Cloud system could monitor "petitioners who are extremely [persistent]" and "Uyghurs from South Xinjiang". It could even pinpoint their residences and track their movements on maps.

[16] *Ibid.*

[17] *Ibid.*

[18] *Green v City and County of San Francisco*, 751 F.3d 1039 (2014)

Especially dangerous is the very real possibility of bias in the input data or the man-made algorithms, causing bias in the predictions. The LAPD has seen a "feedback loop" in its PredPol system sparked by skewed input data.[19] A racial bias in the existing crime statistics made the algorithm direct officers to certain neighbourhoods – typically those with many racial minorities – regardless of the true crime rate in that area. Errors and bias in data will inevitably survive in the resulting predictions. Authorities would do well to be cautious in relying too much on predictive policing systems.

That said, predictive policing's short track record seems promising. Chicago's 7th District Police reported that shootings in that district dropped 39% from January to July 2017 compared to the same period last year.[20] The murder rate also dipped 33% during that period, while the murder rate in the city as a whole rose.[21] Predictive policing systems have also beaten human analysts in other real-world trials.[22] While the many inherent shortfalls of these systems must be acknowledged, and applications of these systems must take these failures into account, predictive policing is clearly a technology with great potential.

C. PRIVACY IS NOT UNASSAILABLE

Apart from criticisms of the data analysis systems, data collection on such a scale also raises many legitimate privacy concerns. If GPS data shows Alice was at a hotel at 5 o'clock on Tuesday, and Bob was at the same hotel at the same time, it can be inferred that they might have been together. Further inferences can be drawn from conversation logs between them, if any. Therein lies the core of big data analysis: to reach conclusions that can be drawn through the correlation of many data

[19] Danielle Ensign et al, "Runaway Feedback Loops in Predictive Policing" (2017) arXiv:1706.09847v2 [cs.CY]

[20] Juliet van Wageren, "Cities Give Predictive Policing a Second Look", *Slate Tech Magazine* (12 December 2017), online: <https://statetechmagazine.com/article/2017/12/cities-give-predictive-policing-second-look>.

[21] *Ibid*.

[22] During a four-month trial in Kent, London, 8.5% of all street crime occurred within and next to the areas the PredPol system designated as high crime areas, predictions from police analysts scored only 5%. An earlier trial in Los Angeles saw the machine score 6% compared with human analysts' 3%. See "Don't even think about it", *The Economist* (20 July 2013), online: <https://www.economist.com/news/briefing/21582042-it-getting-easier-foresee-wrongdoing-and-spot-likely-wrongdoers-dont-even-think-about-it>.

points, which could not have been drawn from the data points themselves. If these data-sets are from public sources, does the individual have a right to privacy in the new information revealed through analysis? In *United States v Maynard*,[23] the DC Circuit held that although the appellant's individual journeys on public roads were public information, since the compilation of data on these journeys would not have been reasonably expected, the appellant's behaviour patterns revealed through analysis of all of these trips together remain private.

How, then, are the inferences drawn from big data to be kept private? Data-sets are usually anonymized before use to protect the privacy of the individuals that the data was collected from. Anonymization is done through deletion of personally identifiable information ("PII"), or obfuscation thereof (for example, by changing a postcode from 123456 to 123***). Top government agencies and leading technology companies have embraced anonymization to protect privacy rights: the US Department of Defence has recommended anonymization "whenever practicable",[24] and Google has said that its anonymization techniques can make identification "very unlikely".[25] Prominent legal scholars also share this faith in anonymization,[26] and claim that anonymization can make data reidentification "impossible".[27]

This faith is misguided. Unfortunately, data reidentification is easier than most would think. If all PII is removed from a data-set, there will be nothing left, since every piece of data is potentially useful in identifying an individual. Therefore, for data, complete privacy means zero utility. For a data-set to have any use at all, some PII must be retained. The widespread faith in anonymization is based on the belief that it is possible to remove enough PII to prevent identification and still

[23] 615 F.3d 544.

[24] Technology and Privacy Advisory Committee, *Safeguarding Privacy in the Fight Against Terrorism* (United States of America: Technology and Privacy Advisory Committee, 2004) at 50 (Recommendation 2.2). The document is available at: http://www.cdt.org/security/usapatriot/20040300tapac.pdf.

[25] Chris Soghoian, "Debunking Google's log anonymization propaganda", *CNET* (11 September 2008), online: <http://news.cnet.com8301-13 739_3-10038963-46.html>

[26] Ira S. Rubinstein et al, "Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches" (2008) 75 U Chi L Rev 261, at 266 and 268

[27] Barbara J. Evans, "Congress' New Infrastructural Model of Medical Privacy" (2009) 84 Notre Dame L Rev 585, at 619-20

retain enough to keep the data-set useful. However, very little PII is actually necessary for deanonymization and thus identification. It has been shown that 87.1% of people in the US can be uniquely identified by their combined five-digit ZIP code, birth date, and sex.[28] 53% of American citizens are uniquely identified by their city of residence, birth date, and sex.[29]

Using just three easily obtainable pieces of information, the vast majority of people can be identified using "anonymized" data. Privacy, in the age of big data, is a much more elusive ideal than previously thought. While it would be asking too much for legislators to come up with rules that could prevent any subsequent deanonymization, legislators must at the very least abandon the idea that removing PII is sufficient protection of privacy in today's security climate. Since this assumption is the foundation to nearly all privacy laws in use today, a paradigm shift is necessary for legislators and industry leaders in this field.

D. THE CLARION CALL TO REGULATION

The law is often slow to respond to societal change. Legislatures must mire themselves in debate before promulgating laws far overdue by the time they are passed. The courts, while more flexible, are still reactive rather than proactive. While both bodies stagnate, the total amount of information grows, and private companies and states worldwide are responding to this wealth of information with increasing urgency. In particular, legal practice has been affected: the availability of these huge amounts of data and the analysis has changed case analysis, pricing of legal services, and even the generation of evidence. Police departments worldwide now routinely draw on predictions based on huge data-sets collected from surveillance networks and private companies. Proponents of such policing systems posit this helps the authorities respond faster and more effectively to changing crime trends; critics argue that this breaches the citizen's right to privacy and gives the authorities untrammelled power to a dystopian degree. Underlying these are the individual's privacy rights and the unfortunate reality that it is harder to protect these rights than previously thought.

---

[28] L. Sweeney, "Simple Demographics Often Identify People Uniquely" (2000) Carnegie Mellon University, Data Privacy Working Paper 3.

[29] *Ibid*.

Whatever the stance taken, we cannot afford to ignore these issues. A single death is a tragedy but a million deaths a statistic; the countless records of crimes and punishments that form the law have many lessons to teach us, but only if we decide to listen.