scripted

Volume 11, Issue 2, September 2014

"THEIRS' NOT TO MAKE REPLY, THEIRS' NOT TO REASON WHY"- A WORKSHOP REPORT ON BIG DATA, FORENSIC REASONING AND THE TRIAL

Burkhard Schafer [*]

DOI: 10.2966/scrip.110214.145

[*] Burkhard Schafer is Professor for Computational Legal Theory at the University of Edinburgh and Director of the SCRIPT Centre for IT and IP law. With Colin Aitken, he is also a co-founder and co-director of the Bell Centre for Forensic Statistics and legal Reasoning. The workshop was jointly organised by both centres, with financial contributions from the School of Law; the School of Mathematics and Statistics; and a personal donor gratefully acknowledged.

Over the last few years, "Big Data", has emerged as a major topic in the discussion on the future of the internet and an internet driven economy. As Bollier and Firestone in "The Promise and Perils of Big Data" put it, "[…] a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging, and the implications for business, government, democracy and culture are enormous." [1] By analysing more efficiently the ever increasing amounts of data that companies hold about their customers, products and processes, companies understand their own business better and better. They are able to quantify more and more of the crucial parameters of the business and thus, become better and better at predicting and managing its future.[2] Or as Eric Siegel writes in the introduction to his influential *Predictive analytics: the power to predict who will click, buy, lie, or die*:

> "You have been predicted — by companies, governments, law enforcement, hospitals, and universities. Their computers say, "I knew you were going to do that!" These institutions are seizing upon the power to predict whether you're going to click, buy, lie, or die. Why? For good reason: predicting human behavior combats financial risk, fortifies healthcare, conquers spam, toughens crime fighting, and boosts sales."[3]

Predicting how the market will react to a new music video, which customers to target with the latest advert, if a piece of news will result in a run on the bank or if a pattern of changes in Facebook statuses is indicating an emerging flu epidemic are all examples of the predictive power of Big Data analytics. Harnessing this is of course also of potential interest to law and law enforcement – the Minority Report may have edged just a little bit closer to reality, as Big Data may enable police to predict unrest or civil disorder from mining Twitter discussion, or cybersecurity experts to identify an upcoming denial of service attack through analysing internet traffic patterns. However, the focus on predicting future behaviour has meant that the impact of Big Data on forensic reasoning and the trial has been largely neglected. Fact finding in the context of a trial is typically concerned with one specific individual event that took place in the past – did the accused commit the crime he is charged with, did the defendant cause the harm for which damages are sought? This focus on reconstructing a unique past event aligns legal reasoning about facts more closely with history, archaeology or geology than the laboratory sciences and their aim to develop reliable predictions of the future through the discovery of universally applicable patterns and relations.[4]

And yet, it cannot be doubted that trial and legal process have been profoundly influenced by modern science. The forensic science process that began in the 17th century and culminated in the emergence and proliferation of dedicated forensic disciplines in the 20th century revolutionised the way in which facts are established in a legal setting. This difference in basic epistemic assumptions and aims between legal trial and scientific discovery caused lasting tensions, which the law of evidence tries to mitigate and manage. Imprinting its own normative logic upon scientific practice as a social phenomenon, the law of evidence tries to determine the nature of scientific expertise. Decisions such as the *Daubert* decision in the

---

[1] Bollier, D. and Charles M. F. The promise and peril of Big Data. Washington, DC, USA: Aspen Institute, Communications and Society Program, 2010 p.1

[2] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data. The management revolution. Harvard Bus Rev, 90(10), 61-67.

[3] Siegel, E. (2013). Predictive analytics: the power to predict who will click, buy, lie, or die. John Wiley & Sons.

[4] Farber, D. A. (1997). Adjudication of Things Past: Reflections on History as Evidence. Hastings LJ, 49, 1009.

US,[5] or legislative initiatives such as the recently proposed reform of the law on scientific evidence in England[6] try to guide lawyers in distinguishing reliable from unreliable science, trustworthy from untrustworthy experts.

Do these rules, guidelines and heuristics need revisiting as a result of the Big Data revolution? It seems that a strong *prima facie* case can be made that just as modern science has both changed and challenged the logic of the trial, so can and will Big Data. Potential or actual examples of Big Data analytics in forensic contexts are already emerging. Can for instance forensic linguistics use the abundance of samples of written English on the Internet to determine the frequency with which an unusual slang expression is used or a spelling mistake made, for a stylometric identification of the author of a blackmail note? Can data on the pollution of a river, collected by "citizen scientists" on their smartphones and uploaded on the internet be used in prosecutions for environmental crimes? Can courts in their "gatekeeper function" use Big Data from social networking sites and online publishers to determine more accurately if a scientific idea or method is "generally accepted" by the scientific community, for instance by the pattern of retweets that indicate that a publication announcement is well received by the peers of the author?

The last two examples indicate an important change that Big Data science might bring about for the legal process. Traditionally, reliability of scientific expertise was (in parts) achieved by a system of quality control and accreditation. DNA laboratories, just like government owned national DNA databases, are subject to more or less stringent regulation that can give us a degree of confidence in the quality of the underlying data and the processes by which it is collected, curated and interpreted. By contrast, "variability", including variability in quality, is one of the hallmarks of Big Data. The hope is that the sheer volume means that statistically, low quality information will not result in wrong predictions, but be "filtered out" through the statistical methods that are employed. What do the lack of centrally controlled data quality and the heterogeneous and unsystematic nature of Big Data mean for the administration of justice? To accept expert witness statements on the basis of data that is *a priori* known to be of low quality in parts will require a major adjustment in the way in which judges have traditionally exercised their gatekeeping function. Indeed, recent decisions such as *R v. T*[7] point if anything towards a greater insistence by the courts on high levels of transparency and demonstrable data quality to make probabilistic judgements by experts permissible than had been in the past. If this trend continues, then it could create barriers for the use of Big Data analytics in forensic settings, leaving for the moment as an open question if this would result in the welcomed exclusion of unreliable methods or the unnecessary rejection of trustworthy ones.

A final question in relation to the use of Big Data in criminal proceedings highlights another set of challenges and dangers, namely: will the availability of larger and larger amounts of health care data prevent the next Harold Shipman or cause the next miscarriage of justice, as happened in the case of Lucia de Berk? Her case in particular brings several of the issues surrounding a forensic use of Big Data into sharp relief. Data from the Dutch health care system was used to "establish" statistically that the chance of a nurse being present at the scene of the unexplained deaths that she was charged with was one in 342 million. However, this statistical analysis was in several respects seriously flawed. In a forensic setting, in particular in adversarial systems of adjudication, we rely on defence counsels to "make reply" when the prosecution introduces expert evidence and challenge it vigorously. This however

[5] Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993)
[6] http://lawcommission.justice.gov.uk/areas/expert-evidence-in-criminal-trials.htm
[7] R.vT [2010] EWCA Crim2439;[2011] 1 Cr. App. R. 9

requires a solid understanding of the underlying scientific and mathematical principles, together with a high degree of transparency of the analytical methods that were used to derive the result. However, the tools that are used for Big Data analytics are often proprietary and protected by trade secrets, making independent scrutiny difficult. Even where such an independent analysis is possible, the complexity of the statistical analysis will regular go beyond the capabilities even of comparatively well-trained judges or counsel for the a parties.

The case of Lucia de Berk also illustrates that potentially an even deeper sea-change will be heralded by the use of Big Data in forensic settings. The pattern in the data was so obvious and the correlation between her working shifts with the unexplained deaths so strong, that for the purposes of prosecution it was not necessary to build a conventional story that led from a compelling motive together with proving that she had the means at her disposal and the opportunity to use them. One advantage of such conventional narratives" that explain the why and how of a suspect's actions through the everyday ontology of causal relations was that they led to testable predictions. Assuming the prosecuting narrative is true, we should expect to find additional evidence, which should be absent if the defence narrative, is correct. This approach underpins the concept of falsification in science just as much as the practice of critical scrutiny through cross-examination and with it the adversarial legal process.[8] In the past, both legal and scientific thinking converged in their emphasis on causal accounts of this type - accounts that allow the finder of facts "to reason why" (and indeed how). Big Data by contrast may leave us with a Humean world where correlations are all there is.

For the practice of science, it has been claimed that the thinking in causal, explanatory categories will be swept aside by the Big Data revolution, resulting in a new, data driven practice of scientific research. Some proponents of Big Data are going as far as suggesting that it heralds the "end of theory" altogether[9]: Intelligent search algorithms that mine huge amounts of data for patterns will replace the "academic hunches" that lead to the formulation of tentative causal hypothesis on the basis of limited data. Mayer-Schönberger and Cukier put it like this:

> "Since Aristotle, we have fought to understand the causes behind everything. But this ideology is fading. In the age of Big Data, we can crunch an incomprehensible amount of information, providing us with invaluable insights about the what rather than the why."[10]

The Aristotelian thinking in terms of causal relations is however deeply ingrained in the practice of reasoning about facts in law, in terms of both physical causes ("what caused his death") and mental states ("why did she kill him").[11] How will the courts react to expertise that denies them in principle answers to this type of question? Who in this new world is the expert, who "owns" the numbers? Permitting experts to quantify the strength of the evidence

---

[8] see e.g.Schafer, B., & Keppens, J. (2005). " And then there was none"-Indirect proof and hypothetical reasoning in Law. Archiv fuer Rechts-und Sozialphilosophie, 177-187.

[9] Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, (Science: Discoveries). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[10] Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

[11] See e.g. Bex, F., Bench-Capon, T., & Atkinson, K. (2009). Did he jump or was he pushed? Artificial Intelligence and Law, 17(2), 79-99; Walton, D., & Schafer, B. (2006). Arthur, George and the mystery of the missing motive: towards a theory of evidentiary reasoning about motives. International Commentary on Evidence, 4(2).

is a relatively recent phenomenon, and itself the result of a long and often acrimonious struggle between scientific experts and lawyers over control in the courtroom. In one traditional model of legal fact finding, experts provide the bare facts, it is the role of judge or jury to weigh the evidence and assess its credibility. Expressing evidence in probabilistic terms, central for many modern forms of forensic evidence such as DNA, was often seen as an intrusion into the territory of the finder of facts. It was only with the growing importance of DNA evidence that statistical assessments of evidential weight by the expert witness became acceptable. A compromise of sorts was reached that laid out clear preconditions under which a causal account of the evidence could be couched in probabilistic terms. However, the expert always remained the person trained in the natural sciences e.g. biology or chemistry, not the statisticians. Big Data calls this historical accommodation and its underlying epistemology into question just as much as it challenges the role of the traditional forensic scientists as expert witness. In the world of Big Data, if some of its more aggressive proponents are to be believed, it would have to be the data analyst as a generalist in all forms of data analysis, independent of domain, and not the forensic biologist, chemist or anthropologist with their domain specific knowledge, who would take centre stage in the proceedings.

So far there has been little discussion in the forensic science and evidence law communities on these opportunities and challenges. If some of the claims of Big Data evangelists are to be believed, then the "Big Data paradigm" will bring considerable disruption to the practice of forensic statistics, forensic science and legal reasoning, and with that the administration of justice. But how serious and credible are these challenges? How prepared is the legal system? Is there a need for new forms of training for lawyers or juries, are there new ways needed to communicate data driven expert evidence in the courtroom? Are there needs for reform in the law of evidence, the regulation of scientific expertise in the courtroom and the way in which the complementary roles and duties are assigned to judges, party lawyers, jurors and witnesses?

To address this gap and to begin a dialogue between lawyers, statisticians, scientists and educators on this topic, the SCRIPT Centre for IT and IP Law organised a round table workshop jointly with the Bell Centre for Forensic Statistics and Legal Reasoning on the 5th of September in Edinburgh.

Topics addressed included a discussion of the current practice of statistical and probabilistic analysis in court, so to speak the legacy that "small data" has created for the legal system and on which any future developments will have to build. Colin Aitken from the University of Edinburgh, representing the forensic statistics community, introduced an ambitious project of the Royal Statistical Society to develop a multi-volume Practitioner Guide that aims to give an overview of all the relevant statistical knowledge that the stakeholders in the criminal justice system need for assessing the probative value of evidence. This guidance for judges, lawyers, forensic scientists and expert witnesses will give a comprehensive and standard setting account of the way in which probabilistic methods for data analysis should be used in courts.[12] The ensuing discussion tried to gauge if this project needs to be expanded to cover Big Data analytics, or if those aspects of Big Data that have validity are already adequately covered by it. This followed a line of reasoning indicated by Big Data sceptics such as danah

[12] Vol 2 on DNA evidence is available here : http://www.maths.ed.ac.uk/~cgga/Guide-2-WEB.pdf

boyd and Kate Crawford who warn against the danger of side-lining more appropriate analytical tools as a result of the marketing hype surrounding Big Data.[13]

Edinburgh's Burkhard Schafer tried to place the forensic potential of Big Data into the wider context of a sometimes paradoxical search for certainty in the legal fact finding process. Traditional, pre-scientific methods of fact finding such as confession and trial by ordeal held the (deceptive) promise of absolute certainty by relying on epistemically privileged observers: the accused himself, and an omniscient and interventionist God are the only possible candidates for an account of past events that does not involve inferences under uncertainty. As the belief in the latter waned, and the problem of false confession, especially when extracted under torture, became too obvious to be ignored, modern science offered a radically different alternative. The very possibility of certainty was abandoned under the onslaught of radical, Humean scepticism, but as a replacement emerged the possibility to give a precise expression to the degree of our ignorance. The scientific revolution, and ultimately the revolution of forensic science in court, thus not only increased our knowledge, it also increased our knowledge about its limitations. At the moment of radical and potentially destructive scepticism, an alternative thus emerged through a major historical compromise: a belief in a clockwork world that follows strict laws underpins our trust in its intelligibility and our ability to reason reliably about it, even if we cannot have certain knowledge of these laws. But the emergence of probability theory, often intimately linked historically to questions of legal reasoning, created a new type of knowledge, precise and quantifiable knowledge of the limits of what we can know. This in turn allowed the formulation of central legal concepts such as "proof beyond reasonable doubt" or Blackstone's ratio. If the claims about a radical change in the nature of science necessitated by Big Data come to fruition, this historical compromise is in peril and a return to the radical scepticism of Hume a possibility, with as yet unclear consequences of our understanding of legal reasoning about facts.

Marco Gomes (IBM) represented the industry perspective, with a fascinating insight on the role of Big Data analytics in forensic science and fraud detection. While his focus was on the more common use of Big Data to predict criminal behaviour and help the prevention of crime, it also gave an account of the advances that have been made in the field of data analytics. For the lawyers in particular, his talk opened up a discussion on issues of privacy and data protection. At present, exclusionary rules can prevent the use of evidence that was unlawfully obtained. To make this determination though, the process of gathering evidence has to be fully explicit and transparent. Obvious issue arise if the complexity of the data collection and analysis process, another defining feature of Big Data, make this type of legal scrutiny problematic or impossible. If only one or two pieces of data in a very large data set are of legally problematic, does this "contaminate" the entire analysis and make it legally inadmissible?

Christopher Laing from Northumbria University talked about digital forensics and the role of Big Data to guide the investigative process. This involves prioritising the right devices ("triage") and case auditing requirements. Digital and computer forensics is the forensic discipline most obviously affected by Big Data. Better analytical tools and methods are not just an opportunity in this context; they are a necessity, if we do not want a backlog of cases that could bring the justice system to a standstill. His talk focussed on the potential of Big Data to develop more rational and transparent methods on device triage (what devices analyse first, where should our priorities lie given constraints on resources) and case auditing.

[13] See e.g. boyd, d. and Crawford K. (2012). "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." Information, Communication, & Society 15:5, p. 662-679.

The discussion took up the vision of the investigative process that this approach entails. "Actuarial justice" and "actuarial policing", terms coined by Malcolm Feeley and Jonathan Simon to describe a justice system based on calculation of risks using the statistical methods of insurance companies,[14] are part of the reality of policing the risk society Its dangers for the legitimacy of police work and resource allocation have been widely discussed. Big Data could add a new dimension to this debate, by reducing the strain on some resources while potentially creating new problems elsewhere.

Finally, Rónán Kennedy from the National University of Ireland, Galway talked about the possible role of Big Data and environmental prosecutions. Environmental regulation presents a particularly appropriate context for the forensic use of 'Big Data', as it is so closely tied to developments in both science and technology. The challenges of properly managing the quality of the environment are complex and difficult, and rely more often than not on complex computational models that are in turn driven by Big Data.[15] Environmental law and science have long been linked in a way that is distinctive, something which can be traced through the development of classification and statistical analysis in the 18th and 19th centuries and into the modern focus on standard setting. In the courts, the two have an uneasy relationship but science is often key to determining legal liability. Rónán's paper explored the resulting questions, highlighting how regulators are using Big Data in practice, the extent to which they are opening their systems to input from citizen science and allowing NGOs and the general public to have access to their datasets.

The workshop, attended by practicing lawyers, computer scientists, statisticians, medical researchers, legal academics and forensic practitioners was a first step to developing a shared vocabulary to discuss the likely impact of Big Data on the trial process. Its aim was also to contribute to "foresighting", and anticipating as far as possible the necessary changes, if any, that the legal system may have to contemplate as Big Data enters the scientific mainstream. A core function of the trial is not just a reliable determination of facts, it also has a symbolic and legitimising role. It is not enough that justice is done; it has to be seen to be done. This requires a degree of transparency and accountability that is potentially inimical to the underlying logic of Big Data analytics, especially when based on proprietary software tools that are intended for competitive markets (of which at least in England, the forensic service market is an example). To discharge its legitimising function, the legal system assigns complementary yet also antagonistic roles to the judge, jury, witness and legal representatives. For juries and defence solicitors alike, the right "to reply, and to reason why" is central. In a Big Data environment, this right may need particular protection. This includes new and better forms of communicating the results of Big Data analytics to laypeople, e.g. through visualisation tools. It includes the need for potentially new forms of training for lawyers. It may require legal intervention e.g. in the regulation of "forensic data analysts" as a discipline. The adversarial process requires a degree of openness about the underlying assumptions, methods and techniques of forensic practitioners which may not be best served through traditional forms of scientific publication, and may be positively hindered through intellectual property and trade secret restrictions on access to data and software specifications. Initiatives such as the "recomputation initiative" that aims to utilise the Internet for new forms of dissemination of scientific knowledge could be an aspect of the

[14] Feeley, M., & Simon, J. (1994). Actuarial justice: Power/knowledge in contemporary criminal justice. The Futures of Criminology. London: Sage.

[15] See e.g. Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present.

solution.[16] Its aim is to make available not just the raw research data, but all the software tools and documentation necessary to replicate the results claimed in the academic papers that they accompany, Where currently, legal approaches such as the Daubert standard rely on traditional peer review, recomputation is considerably closer to the adversarial ethos of the trial and the type of open scrutiny that it demands.

Both the SCRIPT Centre for IT and IP Law, and the Bell Centre for Forensic Statistics and Legal Reasoning which organised this workshop will continue to provide forum for this ongoing discussion, continuing the series of talks and events on this topic in February with a workshop that will focus on fire investigation as domain.

[16] http://www.recomputation.org/blog/2014/08/25/recomputation-dot-org-involved-in-new-data-intensive-research-institute/