The Evidentiary Value of Big Data Analysis

Marco Pollanen & Bruce Cater
Trent University, Canada

Abstract

Big data is transforming the way governments provide security to, and justice for, their citizens. But it also has the potential to increase surveillance and government power. Indeed, information gathered from license plate recognition, mobile phone usage, biometric matches of DNA, facial recognition, financial transactions, and internet search history is increasingly allowing government agencies to search and cross-reference. The opportunity for big data searches then raises the question: what is the probative value of the information that results?

The scientific method begins with the development of a hypothesis that is then tested against data that will either support or refute the hypothesis. That method is essentially followed in a conventional criminal investigation in which, after a suspect is first identified, evidence is gathered to either build a case against, or rule out, that suspect.

The analysis of big data, by contrast, may at times be more akin to trawling for data first, only to subsequently define a hypothesis. In this paper, we investigate the conditions in which this approach may lead to problematic outcomes, including higher rates of false positives. We then sketch a big data analysis legal/policy framework that may address these problems.

Keywords: database searches, forensic science, big data analysis, criminal databases

Introduction

With the advent of smartphones, we may now leave in our wake a near-complete digital trail of our activities, from everything we search for and read online, to continuous location information, to a record of our interactions through social media and texting apps. Wearable technology is evolving in the direction of recording complete health and biometric information. Contactless transit cards, and arrays of video cameras coupled with improvements in facial recognition and license plate reading algorithms, allow our movements to be more accurately tracked. And cashless transactions allow every detail of purchases to be recorded.

At the same time, advances in computing and statistical analysis are increasingly allowing for the contents of these vast databases to be analyzed and for inferences to be drawn.

In the United States, dragnets – where large numbers of people are indiscriminately questioned or detained – are a violation of civil liberties and considered unconstitutional. Yet, in recent years, we have seen the growth of digital dragnets that provide law enforcement with access to ever-larger DNA databases, financial and communications records, and the results of widespread electronic surveillance. These dragnets are often justified by threats to security or rationalized as resulting in metadata only. What this paper will show, however, is that, regardless of how these searches are framed, the data they collect may lead to erroneous conclusions.

One feature of the scientific method is that a hypothesis is formulated *before* data are drawn to test that hypothesis. This feature is important because, without establishing a hypothesis first, it is easy to fall into the trap of data dredging, in which inadvertent patterns are uncovered and misleading conclusions are drawn. Indeed, it is for this reason that modern forensic investigations ideally follow a path that is characteristically much like the scientific method – a suspect is first identified (i.e., a hypothesis is first formed), then evidence to test that hypothesis is gathered.

Of course, the scientific method does vary, depending on the field to which it is applied. In laboratory experiments in physics, for example, repeated real-time observations of a phenomenon can be made and direct evidence can be obtained in abundance, whereas in a field such as paleontology, only scant indirect evidence of an event a long time ago is available. It follows that these fields must employ different mixtures of *inductive* and *deductive* reasoning.

Inductive reasoning involves a bottom-up approach, in which broad generalizations are formed from specific observations. Typically, observations and measurements are made until a pattern that is sufficiently clear is found, from which a tentative hypothesis is formed. Thereafter, the hypothesis is further tested and explored until a general conclusion or theory can be drawn. *Deductive* reasoning, by contrast, is a top-down approach, in which the analysis of observations serves to test a theory by testing hypotheses that arise from that theory.

Both approaches involve experiments that are designed to rule out hypotheses in a process known as falsification – a hypothesis or theory is always open to challenge, and it only gains credibility as the number of attempts to falsify it increases. And both approaches involve the formulation of a scientific hypothesis and the subsequent testing of that hypothesis.

In this paper, we will discuss the implications of deviating from that path. To introduce one type of problem that arises when hypotheses are not established first, we will begin by outlining a well-known result from probability theory: the birthday paradox (Bloom, 1973). We will then illustrate something akin to the birthday paradox with real examples from DNA database matches. Finally, we will examine the problems that arise in the context of Big Data searches, and we will discuss potential solutions to those problems.

The Birthday Paradox

Suppose that, in a group of N people, each of the 365 possible birthdays is equally probable, ignoring leap years.

If $N = 2$, the probability that they share a birthday is 1/365 = 0.0027 – that is, we could assign the first person any birthday, and the second person would have a 1/365 chance of having the same birthday.

A question that naturally arises is: how large does N need to be for it to be likely that (at least) two of the people share a birthday? As it turns out, $N = 23$ is sufficient to give us a 50.7% probability that (at least) two of the people will share the same birthday. Many would see $N = 23$ as being a surprisingly small number.

When $N = 60$, there are fewer people than required to cover 1/6th of the birthdays. Yet, the probability that two of those people share the same birthday rises to 99.4%.

And when $N = 200$, the probability that at least two people share a birthday is an astounding 99.9999999999999999999999999998% – roughly equivalent to the probability of winning a multi-million-dollar jackpot in a lottery four times in a row – despite the fact there are only enough people to cover slightly over half (54%) of the birthdays.

The fact that these probabilities are so unimaginably high is referred to as the birthday problem or birthday paradox (Bloom 1973), not because it is a real paradox – after all, the reasons are well understood – but because it is an apparent paradox in the sense that it defies human intuition.

The birthday problem is relevant to our understanding of the effect that hypothesis formulation has on the validity of forensic discovery. To see this, consider the following cases:

Case I: Suppose a suspect has been first identified. Only thereafter is it determined whether she/he meets a key fact in the investigation. For example, his/her footprint must be of a certain size or he/she is excluded as a suspect. This is akin to asking the question: what is the probability that the birthday of a given individual matches the birthday of another particular individual in the group? The answer to that question, importantly, is independent of the number of individuals in the group; the hypothesis would be validated by random chance alone with a probability of only 1/365 or 0.27%.

Case II: Suppose a suspect has been identified first, but he/she only needs to match *some* fact in the investigation. This is akin to asking: what is the probability that the birthday of a given individual matches the birthday of any other individual in the group. The probability of this match happening by random chance alone grows as the number of people in the group does. For example, when $N = 2$ it is 0.27%, but when $N = 200$ it grows to 42%.

Case III: Suppose that no suspect has been identified and that there no specific exclusionary criteria. Instead, the data are dredged through to find a hypothesis. In the case of birthdays, if N=200, a match will be found by random chance alone with a probability of 99.999999999999999999999999998%, the implication being that the discovery of a match is of no evidentiary value whatsoever.

DNA Database Matches

In recent years, considerable controversy has arisen over the question of how to calculate the probability that a random DNA sample will match the profile of one found in a DNA database. Everyone does have a unique DNA, of course, but DNA databases typically only store a profile consisting of measurements from a fixed set of locations (or loci) on the chromosome. Typically, 9 to 13 independent loci are selected for the database, with two unrelated samples matching at a particular locus, with a probability of about 7.5%, making the odds that two random unrelated profiles will match at a fixed set of 9 (13) loci about 1 in 13 billion (1 in 420 trillion).

Statistical results from *The Arizona DNA Offender Database* (Kaye, 2009) have been particularly controversial. Of the 65,493 profiles in the *Database* at the time, 122 pairs were found to match at 9 loci, 20 at 10 loci, and 1 pair at each 11 and 12 loci. Many found those results to be astounding, for, as noted above, the probability of two random samples matching at 9 loci is about 1 in 13 billion, and at 12 loci about 1 in 32 trillion.

There are, however, several reasons why we would expect to see a large number of matches. The first is due to a sort of birthday paradox, as described above. The second reason is that, in the case of 9 loci, for example, the loci for which the matches occur could be different for different pairs of matches. From a set of 13 loci there are 715 different ways to choose 9 of them, so allowing partial matches increases the odds of a match by an additional factor of 715.

While these considerations do not fully explain the high number of matches, they do come close – for example, in the case of 9 loci, the expected number of matches would be 68, not 122 as were found. But, given the scale of the numbers being dealt with, that *is* fairly close, particularly given the crudeness of the genetic model in which it is assumed that all individuals are unrelated, and all loci are independent with equal probabilities of random matches. A more sophisticated analysis has been done by Mueller (2008).

One issue that arises immediately from DNA matches is that the science is relatively sophisticated and the odds of a random match can seem so overwhelmingly long that it seems possible to identify and convict a suspect by means of a DNA match only. But this is problematic if the match originated from a database search alone. A few cases are instructive.

In what was the first widely reported false match (Fowler, 2003) from a DNA database, a man in the United Kingdom was arrested for a burglary that occurred some 200 miles away and that involved the burglar climbing through a window. In that case, the only evidence was a match from a database search, which would occur with a probability of 1 in 37 million, corresponding to 6 loci. The only problem was that the man was severely disabled and incapable of committing the crime for which he was arrested – a fact that did not clear him.

With the population of the United Kingdom being 64 million, on average we would expect any 6-loci DNA profile to be shared by two people. But, by conducting a DNA database search, we essentially trawled through millions of hypotheses to fit the evidence, violating the first tenet of the scientific method – that we must first have a hypothesis. This illustrates what is known as the prosecutor's fallacy (Thompson & Shumann, 1987), in which investigations and prosecutions revolve around a probability of match. The correct interpretation is that if the suspect is innocent, there is a 1 in 37 million chance that there is a match. However, with the prosecutor's fallacy, the clauses are reversed and the logically incorrect interpretation is adopted – if the DNA matches, there is a 1 in 37 million chance that the suspect is innocent.

It is not just investigators and prosecutors who incorrectly weigh DNA evidence. A 30-year old cold-case (Murphy, 2015) facilitated the analysis of partial matches in the Arizona DNA database. The defendant in that case was identified and convicted largely due to the partial match of the badly degraded DNA sample to a profile found in a California database. The judge allowed only the prosecution's statistic that the chance that an individual picked at random would match the crime-scene DNA was 1 in 1.1 million. Jurors were not informed that the match was a result of a database trawl, whereby 9-loci partial matches are not uncommon, nor were they informed that about 40 people in California would be expected to have a profile that matches the crime-scene sample. The fact that a partial match was used is not uncommon, as crime scene evidence can be degraded and mixtures of DNA samples can result. Furthermore, different databases often use different loci for profiles, and searches can be done using the profiles of close relatives.

A further problem with assigning astronomical odds to a single piece of evidence, such as a DNA database match, is that those probabilities would be dwarfed by real-life considerations, such as laboratory errors and contamination. For example, a man in Australia was convicted of raping a woman found unconscious at a nightclub based solely on a random match in the Australian DNA database (Roberts & Hunter, 2012), despite other evidence suggesting that the individual could not possibly be a suspect. Only through post-conviction serendipity was it discovered that the original rape-kit was likely contaminated at the laboratory, leaving no clear evidence that a crime even took place.

Even when evidence is found at the crime-scene and it is correctly attributed to an individual, the relevance of the sample to the crime must be established. Typically, DNA establishes, at most, the presence of or contact with an individual, not that they committed a crime. In another case, a man in the United Kingdom (Barnes, 2012) was jailed for eight months when a partial match was found between his DNA profile in a database and a crime-scene sample from a murder scene. It has been suggested that because the suspect was a taxi driver, he likely came into contact with the victim individual and some of his shed skin cells clung to that person.

While there are potential pitfalls in interpreting DNA evidence, especially when it comes from random matches found by trawling through databases, it is important to note that DNA evidence is still some of the most reliable types of evidence there is, and that it has likely lead to the exoneration of more people far more often than it has resulted in false convictions. By comparison, while identification by eyewitnesses carries a lot of weight in courts, studies have shown how utterly unreliable eyewitness testimony can be (National Research Council Report, 2014). We introduced the issues with the *Arizona DNA Offender Database* to demonstrate how the birthday problem arises in criminal investigations.

In the next section we will discuss how these problems might be amplified as the number and type of databases used in forensic investigations increases.

Big Data Searches

In recent years, aided by technological advances, and often rationalized as necessary to fight terrorism, mass surveillance has been increasing. In the US, for example, metadata for hundreds of billions of telephone calls has been collected (Cauley, 2006); the exterior of all letter mail is photographed (Miga, 2013); databases containing information on financial transactions, e-mails, and internet surfing habits are maintained; and social media are monitored (Kawamoto, 2006). The FBI has a face-recognition system with a database of over 400 million photos (Kelley, 2016). Combined with the ever-increasing array of CCTV cameras, it may be possible to recognize individuals in any public location. In the United Kingdom, there are up to 6 million CCTV cameras (Barrett, 2013) – about one for every 11 individuals. Location information can also be obtained from license plate recognition or from databases of transit card usage.

In addition to government databases, private companies such as Google and Facebook have access to vast amounts of information about individuals, except where one exerts considerable effort to maintain their privacy. This is especially true due to the near ubiquitous use of smartphones, which potentially provide details of the contents of every digital communication one partakes in, and to one's location history; they can map one's photographs, social connections, and browsing and search histories; and they can potentially track health and biometric information through a phone's sensors. This information is also available to governments seeking to increase surveillance.

The average individual leaves a vast digital trail throughout their day, from which it may be possible to surmise when he/she woke up and how long they slept, their location throughout the day, including where they work, where they shop, and what they bought, read, or wrote. By combining the available information with information from biometric sensors in smartphones or wearable devices, it may be possible to develop algorithms to give an idea of what an individual thought and felt throughout the day or to predict behavior.

While all of this information can be a boon for law enforcement in their quest to solve crimes, the growing number and size of databases also have the potential to lead to an increase in the number of falsely accused individuals. To see this, consider a DNA database in which an individual profile will typically contain information regarding 13 loci. This would be equivalent to an individual having a record in 13 different databases, each containing the information of a single loci. Thus searching through multiple databases of digital information would also be subject to the Birthday Paradox as we have seen with DNA. Moreover, when an individual matches information in only some databases but not others, this further magnifies the problem of false identifications from partial DNA matches have been shown to have with the Arizona DNA database.

While there are many similarities with searching through digital information databases and DNA databases, there are causes for greater concern. DNA analysis occurs in a laboratory setting, and while the measurements have errors associated with them, they can be estimated. Laboratory errors do occur, of course, but it is still a scientific setting where one would believe every attempt would be made to estimate and minimize these errors. On the other hand, analysis of databases of other digital records might involve information that was not originally intended for forensic examination, such as facial recognition on grainy photos or the inaccuracies of finding the location of a mobile phone user. These errors may be poorly understood and might contribute significantly to birthday paradox collisions. Furthermore, DNA analysis involves trying to match a set number of loci, while a trawl through digital data may involve an unknown number of databases and is problematic because the probability of a match would be incalculable. We have seen that partial matches of only some loci significantly increase the chance of misidentification with DNA databases, but in that case we know which loci cannot be matched and probabilities could be adjusted accordingly. It would be even more problematic if the databases investigated were not known or revealed. For example, suppose while in an investigation, all Google searches for a particular explosive were flagged. If a suspect had searched for that particular compound, that would certainly be used to build the case against him or her. On the other hand, if the suspect was *not* one of the individuals who had made that particular search, that fact might not be factored into the calculation of their probable guilt and it would almost certainly be inadmissible in court.

The concern with searching through a large number of databases for suspects that could fit the evidence of a crime is, of course, that people may be falsely accused. But with such an overwhelming amount of circumstantial evidence pointing to them, it could be difficult to exonerate them. For example, perhaps a murder has been committed, and by pure chance alone an individual is found whose license plate was caught driving nearby at a similar time, traces of whose DNA are found on the murder victim (perhaps because they ate at the same restaurant), and perhaps the day before they bought the same brand of duct tape used in the crime. As technology and pattern recognition algorithms get better, it is likely that even more casual links in vast arrays of data will be found.

Discussion

Databases are important tools for fighting crime and protecting national security. Indeed, as criminals become more sophisticated in their use of technology, there is arguably a need for law enforcement to do the same. A significant problem arises, however, because trawling through a database without a suspect in mind – essentially in violation of the scientific method – may result in erroneous conclusions. A primary goal, then, should be to put these searches on a more scientific footing.

There is not necessarily a singular scientific method due the variety of very different scientific disciplines. Likewise, no singular definition might be easily developed for Big Data analysis, as there are so many different reasons for database searches. For example, the needs of an investigator trying to solve a crime that occurred in the past might be different from those of an agency looking for patterns in data to prevent a future terrorist attack. In any case, however, to be scientific, it is necessary for a hypothesis to first be formed. One obvious way to achieve this would be to use a database for identifying a suspect (i.e., formulating a hypothesis), and then to use only further evidence gathered from other sources for the purposes of prosecuting. A second possible approach would involve (perhaps randomly) separating a list of all available databases into two parts. From one part, a suspect could be identified, while, from the second part, searches could be conducted to build a case against that suspect. Of course, many would object to either approach, for they would be seen as leaving evidence unused.

It is also important to understand the statistical characteristics of many of the key databases used in order to understand their scope and potential inaccuracies. This would be important for assigning a probability of a match for use in the legal system. Many of the examples in this paper were drawn from criminal cases involving DNA databases, not because these cases are reflective of the only context in which the issues discussed will arise, but because they are the only concrete examples that have, to date, been subjected to some legal scrutiny. By contrast, although it has been widely reported (Press, 2012) that counterterrorism efforts involve mining social media to predict terrorist attacks, it would, at this stage, be a formidable task to assign the probability of a random mismatch in those circumstances. Even for a computer to search human language to find an initial match requires cutting-edge computational linguistic models that are in their research infancy. Furthermore, there is no fixed database that exists at a moment in time that can be searched again at a later date – rather, the information content on social media is constantly in flux. Even when a database is government-maintained, it may be fraught with problems – for example, in the US there many anecdotal reports of misidentifications due to coincidental mismatches of identities on the no-fly list (Kreig, 2015) and department of motor vehicle records (Davis, 2017).

It would also be important to study frequently-used databases to understand the potential for misidentification. In the case of DNA, of course, law enforcement agencies have fought access by researchers to study random match probabilities (Kaye, 2009). Yet, potential violations of privacy could easily be circumvented by, say, scrambling or encrypting the data in some way to avoid identifying individuals, without compromising the ability of researchers to analyze key characteristics of the data. In order to make investigations more scientific, it is important to carefully document all database searches included in the hunt for a suspect – even those searches that lead to negative results. For these purposes, the development of a standardized set of databases and search criteria would be appropriate.

Conclusion

Governments and law enforcement agencies increasingly have access to vast amounts of electronic data that can be searched to identify a suspect or a potential crime. This fact has received a great deal of attention in the literature from those concerned about the privacy implications. The potential for coincidental mismatches, however, has not been explored.

In this paper, we have shown that as the databases grow, the potential for random mismatches may grow exponentially if searches are not done in a way that is consistent with the scientific method. A simple strategy to avoid this problem would be to split datasets in such a way that a part of the data is used to identify a suspect and effectively form a hypothesis, while the remaining parts of the database are used to test that hypothesis. Ultimately, however, more research is required in this area to model potential database searches so as to more fully understand the full extent of the problem.

References

Barrett, D. (2013, July 10). One surveillance camera for every 11 people in Britain, says CCTV survey. *The Telegraph*. Retrieved September 7, 2016, from http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html

Barnes, H. (2012, August 31). DNA test jailed innocent man for murder. BBC News. Retrieved September 6, 2016, from http://www.bbc.com/news/science-environment-19412819

Bloom, D. (1973). A Birthday Problem. *American Mathematical Monthly 80*, 1141–1142.

Cauley, L. (2006, May 11). Advertisement NSA has massive database of Americans' phone calls. *USA Today*. Retrieved September 7, 2016, from http://usatoday30.usatoday.com/news/washington/2006-05-10-nsa_x.htm

Davis, L. S. (2017, April 3). For 18 years, I thought she was stealing my identity. Until I found her. Retrieved April 26, 2017, from https://www.theguardian.com/us-news/2017/apr/03/identity-theft-racial-justice

Fowler, R. (2003, April 27). DNA, the second revolution. The Guardian. Retrieved September 5, 2016, from https://www.theguardian.com/uk/2003/apr/27/ukcrime7

Kaye, D. H. (2009). Trawling DNA Databases for Partial Matches: What Is the FBI Afraid Of? *Cornell Journal of Law and Public Policy*, *19*(1).

Kawamoto, D. (2006, June 9). Is the NSA reading your MySpace profile? CNET. Retrieved September 7, 2016, from http://archive.is/20120720043006/http://news.com.com/2061-10789_3-6082047.html#selection-925.5-929.1

Kelly, H. (2016, June 16). FBI's face-recognition system searches 411 million photos. CNN Money. Retrieved September 7, 2016, from http://money.cnn.com/2016/06/16/technology/fbi-facial-recognition

Krieg, G. (2015, December 7). No-fly list nightmares: The program's most embarrassing mistakes. Retrieved April 26, 2017, from http://www.cnn.com/2015/12/07/politics/no-fly-mistakes-cat-stevens-ted-kennedy-john-lewis

Miga, A. A. (2013, August 2). AP Interview: USPS takes photos of all mail. AP Online. Retrieved September 7, 2016, from http://www.highbeam.com/doc/1A1-10eae68abcc8439fa78610fe561ab6fc.html?refid=easy_hf

Mueller, L. D. (2008). Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? *Journal of Genetics*, *87*(2), 101–108. https://doi.org/10.1007/s12041-008-0016-4

Murphy, E. E. (2015, October 8). The Dark Side of DNA Databases. The Atlantic. Retrieved September 6, 2016, from http://www.theatlantic.com/science/archive/2015/10/the-dark-side-of-dna-databases/408709/

National Research Council of the National Academies. (2014). *Identifying the culprit: Assessing eyewitness identification.* Washington, D.C.: National Academies Press.

Press, T. A. (2012, February 13). FBI seeks social media data mining tool. Retrieved April 26, 2017, from http://www.cbc.ca/news/technology/fbi-seeks-social-media-data-mining-tool-1.1221437

Roberts, P., & Hunter, J. (Eds.). (2012). *Criminal evidence and human rights: Reimagining common law procedural traditions*. Bloomsbury Publishing.

Thompson, E. L., & Shumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, *2*(3). https://doi.org/10.1007/BF01044641

Corresponding author: Marco Pollanen
Contact email: marcopollanen@trentu.ca