M S Olivier. "Digital forensics and the big data deluge — some concerns based on Ramsey Theory". *Advances in Digital Forensics XVI*, G Peterson and S Shenoi (eds), Springer 2020.
Preprint manuscript.
The final version is available at Springer via http://dx.doi.org/[DOI not available yet]

Chapter 1

DIGITAL FORENSICS AND THE BIG DATA DELUGE — SOME CONCERNS BASED ON RAMSEY THEORY

M Olivier

Abstract Constructions of science (that slowly change over time) are deemed to be the basis of the reliability with which scientific knowledge is regarded. One of the more recent potential paradigm shifts is based on the increase of available data; many researchers deem such *big data* to have enough 'substance' to capture knowledge without the theories needed in earlier epochs. The patterns in big data are deemed to be sufficient to make predictions about the future (and about the past as a form of understanding). The current chapter uses an argument developed by Calude and Longo in 2017 to critically examine the belief system of the proponents of such data-driven knowledge — particularly as it applies to digital forensic science.

From Ramsey theory it follows that, if data is large enough, 'knowledge' is 'imbued' on whatever domain is represented by the data purely based on the size of the data.

The chapter concludes that it is (a) generally impossible to distinguish between true knowledge of the domain and (b) knowledge inferred from spurious patterns that must exist purely as a function of the size of the data. In addition, what is deemed a significant pattern may be refuted by a pattern simply not yet found. Hence 'evidence' based on patterns found in big data is tenuous at best. The field of digital forensics should therefore proceed with caution if it wants to embrace big data and the paradigms that evolved from and around big data.

Keywords: Digital forensic science, artificial intelligence, big data, Ramsey Theory, epistemology

1. Introduction

> *"Today, machine learning programs do a pretty good job most of the time, but they don't always work. People don't understand why they work or don't work. If I'm working on a problem and need to understand exactly why an algorithm works, I'm not going to apply machine learning."*
> Barbara Liskov, Turing laureate [15]

> *"Deep learning and current AI, if you are really honest, has a lot of limitations. We are very very far from human intelligence, and there are some criticisms that are valid: It can propagate human biases, it's not easy to explain, it doesn't have common sense, it's more on the level of pattern matching than robust semantic understanding."*
> Jerome Pesenti, VP of artificial intelligence at Facebook [17]

From ancient times, science has operated on the basis of observation of interesting patterns. Patterns observed in the movement of celestial bodies, interaction between physical objects, and even human behaviour simplified prediction and, eventually, culminated in scientific understanding.

In 1782 John Smeaton, a British engineer, was allowed to offer his scientific knowledge of sea currents as evidence in a case involving silting-up of the harbour at Wells-next-the-Sea in Norfolk [27]. Before then evidence that relied on, say, Newton's work would have been classified as hearsay evidence unless Newton was called to confirm it — a challenge since Newton passed away in 1727. Science and expert witnesses have since 1782 become entrenched in legal proceedings.

Currently we are at another watershed moment in history. With the advent of big data, data science and deep learning, patterns are uncovered at an ever increasing rate and used to predict future events. In forensic science pressure is increasing to use these technologies to 'predict' the past to provide a scientific basis for finding facts that may be useful in legal proceedings.[12]

However, from Ramsey theory, it is known that any data set that is big enough will contain a multitude of regular patterns. The patterns stem from the size of the data set, rather than anything represented by the data; the patterns are guaranteed to exist even in random data. A finding derived from big data may therefore have more to do with the size of the data than with the case being litigated. Such spurious patterns may lead to a spurious system of (in)justice.

The current chapter follows the logic of a more generic 2017 argument of Calude and Longo [10] — based on Ramsey (and Ergodic) theory —

to reflect on the role that big data (and related) technologies ought to play in forensic science, with a specific focus on digital forensic science.

The chapter proceeds as follows. Firstly, some aspects of patterns and repetitions are discussed, with specific reference to inferences based on such patterns. This is illustrated using court cases where short patterns played a significant role. The chapter continues by exploring the guaranteed presence of patterns (that often are spurious) in large data sets. Finally it illustrates the inherent dangers if digital forensic findings were to be based on inferences from patterns in big data.

2. On patterns and repetition

It is all too human to expect chaos in nature, and then to interpret a pattern in the chaos as something of special significance. Conversely, many aspects of nature (such as the coming and going of seasons) produce expectations of a regular pattern, and any deviation from that pattern is often deemed significant. In games of chance some events, such as throwing a pair of dice and getting a double is deemed lucky, and a series of such doubles may be deemed a lucky streak. However, the streak cannot continue for very long before one begins to doubt the integrity of the dice. Conversely, one does not expect that the same person will win the lottery on a fairly regular basis — if this were to happen one would soon doubt the integrity of the lottery system. In such sequences of events there are often sequences that would seem 'normal' and sequences that would seem like an anomaly.

On purely statistical grounds, if the probability of encountering some phenomenon is $p = 10^{-6}$ then one would expect to encountering the phenomenon, on average, once in a million cases inspected. If it is the probability of being born with a specific unusual medical condition, then the usual absence of the condition will in all likelihood be labelled as *normal*, and when a child is born with this condition, it may be deemed to be *abnormal* or, in the language used below, an *anomaly.*

In the examples above, the probability of these 'anomalies' can be calculated rather accurately using basic probability theory and encountering them (on average) once in a given period of time or volume is expected. A more regular occurrence would, with a very high probability, be indicative of some anomaly.

However, as the chapter will explain in more detail below, in a large dataset data 'clusters' (for lack of a better term at this stage) exhibiting certain traits have to occur, where *have to* indicates mathematical certainty. The size and prevalence of such clusters is a function of the size of the data, and may be totally unrelated to what the data is purported to represent. It seems natural to represent the more prevalent clusters as 'normal' and the less prevalent clusters as an anomaly.

Such differentiation between normality and anomalies is often the basis of intrusion detection in data networks and it is increasingly being applied in digital forensics. This claim will be substantiated below. However, note that if the occurrence of 'normal' data flow and 'anomalies' is a result of the size of the data, rather than some justifiable theory, the distinction between normal and anomaly is very tenuous, at best (and will be wrong in many cases). If this is the case, such differences could not be the basis of a scientific finding in forensic science.

To make matters more concrete, consider a request to a Web server containing an extremely long URL. Often this is indicative of an attempt to exploit a buffer overflow vulnerability in the server. 'Normal' requests are typically relatively short, compared to these 'anomalous' requests. In addition, if the lengthy requests can be linked to some known vulnerability in (some versions) of such servers, the odds increase that it is indeed a malicious request. Another well-known pattern from the intrusion detection literature is a port scan. Various methods exist that attempt to hide such port scans that are based on interfering with some of the 'regular' features of a typical port scan. A port scan is often an indication of nefarious intentions (unless the port scan was performed as part of an official security overview). Correlating such anomalous events (such as unusual Web requests or port scans) with reported computing incidents may be useful. However, one should remember that causality may also work in the other direction, where the incident causes the anomalous events. A computer system that lost connectivity will typically send an unusually high number of attempts to re-establish connectivity. More importantly for the purposes of the current chapter: 'anomalous' patterns may be entirely unrelated to incidents they apparently correlate with, and deriving any significance from the pattern would be incorrect. However, making this case convincingly has to be postponed. The starting point for understanding our belief in patterns start at a much simpler point: Where a 'small' correlation seems just too significant to ignore.

2.1 Small correlations

Even in small datasets an unexpected pattern is often deemed significant. To the best of the author's knowledge, the interpretation of patterns in a cyber-related court case has not led to significant scrutiny of evidence presented. We therefore use a well-known and widely discussed matter as a starting point to reflect on the use of patterns as evidence in court.

An infamous ‘law’ using patterns is Meadow’s (now discredited) law: “one sudden infant death is a tragedy, two is suspicious and three is murder, until proved otherwise” [19, p.29]. This ‘law’ formed the basis of expert evidence in a number of cases. Arguably, the most prominent of those cases was R v Sally Clark [3]. Ms Clark’s first son, Charles, died in December 1996, aged 11 weeks. The pathologist found that the death was due to natural causes. Sally Clark’s second son, Harry, died in January 1998, aged 8 weeks. The pathologist ruled the death to be unnatural and revised his finding about Harry, whose death he now also considered to be unnatural. Sir Samuel Roy Meadow was one of the expert witnesses in the ensuing murder trial, and evidence was based on the ‘law’ carrying his name, although the ‘law’ was not mentioned explicitly during his testimony. Sally Clark was found guilty and sentenced to life imprisonment. Sally Clark was released from jail in 2003, after a second appeal [4] was successful.

The ‘pattern’ played a major role in her first conviction and in the failure of her first appeal [3]. The judgement in the second appeal provides interesting insights into how the ‘pattern’ was mentally constructed by the prosecution and jurors. This will be discussed in more detail below.

2.2 Patterns and/or knowledge

The previous paragraph illustrates the potentially strong belief that may be formed — even when considering very short patterns. Court arguments turned on many facets of the Sally Clark case, and the notion of probability was deemed of minor importance; rather, medical knowledge was deemed paramount in the original trial and both appeals.

In contrast, machine learning — especially in the context of big data — in recent years tended to ignore underlying knowledge and rather focussed on patterns. Langley [18, p.278] describes the development as follows: “During the 1990s, a number of factors led to decreased interest in the role of knowledge. One was the growing use of statistical and pattern-recognition approaches, which improved performance but which did not produce knowledge in any generally recognized form.”

During earlier periods of artificial intelligence, some underlying knowledge about the problem domain somehow impacted machine; in expert systems, knowledge representation was at the core of the work; in machine learning, domain specific heuristics improved the speed of learning. However, as machine learning developed, the focus shifted to an “increasing reliance on experimental evaluation that revolved around performance metrics [which] meant there was no evolutionary pressure to study knowledge-generating mechanisms” [18, p.278].

In a similar vein, Anderson [2] published an earlier article in Wired Magazine with the provocative title (borrowed from an earlier claim by a George Box): *The end of theory: The data deluge makes the scientific method obsolete.* In it he, for example, declares

> *“Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”*

2.3 Big data

Big data was a concern in the context of digital forensics since digital forensics emerged as a recognised academic discipline [5, p.24]. Some of the earliest concerns included that finding the needle in the haystack became more challenging as the size of the haystack increased [25]. The fact is that the typical amount of storage associated with a computer increased dramatically and this made imaging of such storage harder or impossible. Emergence of the cloud exacerbated these issues and paper after paper was published that lamented the growth in data volumes.

However, in parallel with these concerns a new field of study developed under the *big data* rubric. The principle that underlies this field is that the universe (or many aspects of it) behave according to some pattern. If enough data is available it can be analysed and the patterns discovered. Once the pattern is known behaviour becomes predictable. This knowledge of the future can be monetised or other benefits may be derived from it. The name of the field has changed over time with *data mining*, *data analytics* and *data science* being some of the prominent examples. The notion of *machine learning* or *deep [machine] learning* is closely associated with the field. In this chapter the term *big data* will be used, unless specific differentiation is required.

Given the popularity of *big data* it was only natural that researchers would posit the use of big data methods in the digital forensic realm.

3. What constitutes correlation?

The Sally Clark case is a good illustration of both ‘pattern recognition’ and correlation in a small data set.

In the second appeal [4], the court pointed out that the previous courts (erroneously) accepted that the deaths of the two children were related (or correlated) on the following grounds (quoted verbatim):

i) Christopher and Harry were about the same age at death namely 11 weeks and 8 weeks.

ii) They were both discovered unconscious by Mrs Clark in the bedroom, allegedly both in a bouncy chair.

iii) Both were found at about 9.30 in the evening, shortly after having taken a successful feed.

iv) Mrs Clark had been alone with each child when he was discovered lifeless.

v) In each case Mr Clark was either away or about to go away from home in connection with his work.

vi) In each case there was evidence consistent with previous abuse.

vii) In each case there was evidence consistent with recently inflicted deliberate injury.

The appeal ruling considers each of these points systematically and rejects every point. It should be noted that these points were raised by the prosecution, rather than the expert witnesses and the court was, in principle, equipped to deal with such an argument. However, the incorrect reasoning of both the court of first instance and of the first appeal case was only rectified by the second appeal case [4].

In contrast, where an expert witness uses such methods, the court is ill equipped to deal with it, unless it is rebutted by another expert. The closest that any expert witness came to including anything similar in expert testimony was Meadow's testimony on the rarity of two infant deaths in one family. Meadow cited from a work where the prevalence of Sudden Infant Death Syndrome (SIDS) was one in 8 543 cases.[3]. Hence, with the probability p of a SIDS case estimated to be $p = \frac{1}{8543}$, he multiplied these probabilities to determine the probability of repeated cases by the number of cases, as if occurrences of SIDS were independent. In the Sally Clark case, he concluded that the probability of two such deaths would be p^2 — or about one in 73 million. He continued by illustrating the unlikelihood of such an event using a comparison from sports book betting. Though the judge downplayed the importance of this number in his instructions to the jury, the effect of it arguably stuck. Of course, two deaths in a family may very well not be independent — it may be caused by the genetic makeup of children in the family, and hence squaring the probability (without showing independence) was incorrect. This was one of the issues raised by the Royal Statistical Society in its press release [13] after the first appeal [3] failed.

The second aspect raised by the Royal Statistical Society [13] was the emphasis on a small probability of a specific outcome. The probability of SIDS is indeed small, but so is the probability (or relative prevalence) of parents murdering multiple children: One cannot focus on the small probability of some sequence of events S and therefore conclude that another unlikely sequence of events B as the logical inference to be made.

As a second example, consider the case of Australian, Kathleen Folbigg. Four of her children died at a young age in 1989 (age: 19 days), 1991 (age: 8 months), 1993 (age: 10 months) and 1999 (age: 19 months). While experts used the same calculation during pretrial hearings, when Folbigg's trial started in March 2003, the British Court of Appeals had already discredited Meadow's law and calculations.

Meadow's Law was therefore excluded, but his ideas still featured during the trial. A Professor Berry, for example, testified that "[t]he sudden and unexpected death of three children in the same family without evidence of a natural cause is extraordinary. I am unable to rule out that Caleb, Patrick, Sarah and possibly Laura Folbigg were suffocated by the person who found them lifeless, and I believe that it is probable that this was the case." A Professor Herdson, on the other hand, deemed the events too different to be a pattern in which SIDS death would occur, and used the absence of a specific pattern (amongst others) to be indicative of unnatural causes of death.

In both cases mentioned here other evidence was influential in the eventual findings of the various courts (that eventually were more important than the presence or absence of a pattern). In the case of Sally Clark microbiological test results of Harry were not available to the defence and was only discovered by them after the first appeal. The appeal court found that availability of these results, along with expert testimony, could have impacted the jury's decision and concluded that the guilty verdict was unsafe. On its own, the guilty verdict regarding Christopher's death was unsafe. The Crown did not apply for a re-trial and the convictions were set aside.

In the Kathleen Folbigg case diaries that she kept played a significant role in proceedings and the outcome of the trial. Public interest eventually led to a judicial inquiry by the former Chief Judge of the New South Wales District Court, Reginald Blanch, to review the case (and hear new evidence). In July 2019 the report he concluded that "the Inquiry does not cause me [Reginald Blanch] to have any reasonable doubt as to the guilt of Kathleen Megan Folbigg for the offences of which she was convicted. Indeed, as indicated, the evidence which has emerged at the Inquiry, particularly her own explanations and behaviour in respect of her diaries, makes her guilt of these offences even more certain." In addition, “there is no reasonable doubt as to any matter that may have affected the nature or severity of Ms Folbigg’s sentence” [7, p.496].

4. On correlation in big data

Many papers have been written that express concern over, or reject the notion that data can speak for itself without the need for any theory. One only has to look through the many papers that cite Anderson’s claim [2] to find such critiques.

One paper that criticises this claim in a manner that should be taken seriously in digital forensics is the paper by Calude and Longo [10] who “prove that very large databases have to contain arbitrary correlations. These correlations appear only due to the size, not the nature, of data. They can be found in ‘random’ generated, large enough databases, which [. . .] implies that *most correlations are spurious*” [emphasis in original].

Calude and Longo use a number of theorems from Ergodic and Ramsey theory that are relevant in the current chapter. However, we will only focus on the final claim by Calude and Longo (based on Ramsey theory), and provide a different exposition.

5. Ramsey theory

Ramsey theory studies the number of objects that should be present in a collection for order to emerge. Perhaps the best-known example is based on a scenario where people attend a party. Any two people at the party will either have met previously or be mutual strangers. If one uses colours to represent the relationship between any two people, the case where they have previously met may be represented by the colour green, while the case where they are mutual strangers may be represented by the colour red. The fundamental question in Ramsey theory is what is the minimum number of people who need to be at the party to have at least c cases of the same colour (or, stated differently, to have c monochromatic cases). If, for example, c is chosen to be 1, it is easy to show that $n = 2$: If we use letters to represent the attendees, a and b will know one another (green) or be mutual strangers (red). If $c = 2$ then $n = 3$: With three guests, a, b and c, the situation may be depicted graphically as a triangle with a, b and c as the vertices and (a, b), (a, c) and (b, c) as the edges, representing the relationships. With two colours (red and green), colouring the three edges requires two edges to be coloured using the same colour.

The notation $R(s, t)$ is used to depict so-called Ramsey numbers. $R(s, t)$ is the minimum number of objects in a set such that some relationship holds amongst at least s members of the set, *or* does *not* hold amongst at least t members of the set.

As illustrated by the party problem mentioned above, it is natural to think in terms of graph theory about Ramsey theory. In the language of graph theory, complete graph is one where every vertex is connected to every other vertex. For n vertices the corresponding complete graph is denoted by K_n. A clique is a subgraph that is 'complete' — in other words, the vertices in the subgraph are all connected. In this context, colour a complete graph using two colours. One colour (say green) is used to colour the edge if the relationship holds between the vertices connected by the edge; the other (say red) is used to colour the edge if the relationship does not hold between the two connected edges. Then the Ramsey number $R(s, t)$ is the smallest n such that K_n *has* to either contain a clique consisting of green edges of size s (or larger), or a clique of size t (or larger) consisting of red edges.[4]

In general the binary relationship used above (that some relationship holds, or does not hold) is too restrictive. It is useful (and possible) to talk about any set of relationships that form a partition of the possible relationships that may hold between the vertices. If the vertices represent, for example, events that occurred in a computer system being investigated, one may distinguish the time between the events may for some reason be deemed a possibly relevant relationship. As an arbitrary example, events that occurred hours apart, minutes apart and (seconds or less) apart forms such a partition — assuming some definition of time for events that occurred multiple times. (Obviously a more precise notion of the informal concepts of *hours*, *minutes* and *seconds* would also be required.)

A cautionary note is required at this stage: The Ramsey theory introduced here (following the exposition by Calude and Longo [10]) is based on undirected graphs, where the relationship between objects or events is symmetric; the time between events is an appropriate example; however, the question whether an event preceded another event, coincided with it or followed it is asymmetric and hence not covered by the current discussion. The exclusion of such asymmetric relationships is not material in this chapter.)

5.1 The Finite Ramsey theorem

In 1930 Ramsey proved the theorem that forms the foundation of the theory carrying his name [26]:

> *"Given any r, n, and μ, we can find an m_0 such that, if $m \geq m_0$ and the r-combinations of any Γ_m are divided in any manner into μ mutually exclusive classes C_i $(i = l, 2, \ldots, \mu)$, then Γ_m must contain a sub-class*

> Δ_n *such that all the r-combinations of members of $Delta_n$ belong to the same C_i."*

An *r-combination* is, as the name suggests, a set of r elements that occur is the dataset. If the dataset consists of the values $\{a, b, c, d\}$ then the 3-combinations that are present are $\{a, b, c\}$, $\{a, b, d\}$, $\{a, c, d\}$ and $\{b, c, d\}$. Every 3-combination is assigned to one of μ classes (or colours, as used previously).

An analogy with the training phase of supervised machine learning may help readers who are better versed in computer science than mathematics to get a picture of what the theorem says. In supervised learning a number of inputs are provided to the classifier, as well as the class associated with those inputs. Say r inputs are used for each instance to be classified and every instance is assigned to one of μ classes. One may then pick any number n. Using only μ and n, a number m_0 can be determined such that any selection of m_0 instances in the training data will have at least n instances that belong to the same class. Note that this analogy says nothing about the resulting learning that may occur; it simply says that having at least n instances of the same class in the training data is unavoidable.

More formally, what the Finite Ramsey theorem does predict (and guarantee) is that there is some (finite) number m_0 such that after classifying m_0 of the r-combinations, n of the r-combinations *will* have been assigned to one of the classes. The theorem says nothing about the first class that will reach this n threshold. It just says that the threshold will have been reached. The point (m_0) at which a class is guaranteed to reach the n threshold can sometimes be calculated precisely. For those cases where it cannot (yet) be calculated precisely, upper bounds can be determined.

The fact that a certain relationship between members of some set held relatively often in a dataset may be of interest in unravelling some incident. Ramsey's theorem warns us proceed with care. However it seems much more likely that an activity of interest in a digital investigation will consist of several actions that together constitute an anomalous (or otherwise useful) indication of what transpired (or is otherwise useful).

As a simplistic example, in a case involving network communications a message may be deemed as significant in terms of the hosts involved in sending the message and the ports used. Hence tuples consisting of these four values may be deemed useful and classified in some manner. Whether these values would be sufficient (or even relevant) cannot be answered without more context, and this example is not explored in more detail.

As a more concrete example, consider approaches commonly used in authorship attribution. Often so-called n-grams are used. Such n-grams are contiguous sequences of linguistic elements. These elements may be letters, words, word pairs, phonemes or a host of other entities that experimentally turn out to be useful. In a well-known authorship attribution competition such n-grams are often the basis of approaches that perform well. In the 2018 competition "n-grams were the most popular type of features to represent texts in" one of the primary tasks of the competition [16]. "More specifically, character and word n-grams [were] used by the majority of the participants."

The detail of the Finite Ramsey theorem does not play a significant role in the remainder of the current chapter; however, it sets the scene for the Van der Waerden theorem, which forms part of Ramsey theory. The Van der Waerden theorem dates from 1927. Note that we are again following the logic of the paper by Calude and Longo in this regard.

5.2 Van der Waerden's theorem

The Finite Ramsey theorem provides a threshold beyond which a certain number of relationships amongst members of some set is guaranteed. Van der Waerden's theorem, in contrast, considers regular occurrences of some value in a sequence of values. It provides a threshold for the length of the sequence. Once the sequence is as long as the calculated threshold (or longer), it is mathematically guaranteed that some value will occur 'regularly' at least k times in the sequence for any given k. Formally, Van der Waerden's theorem says that the repeated value will appear in an arithmetic progression. More informally, these k (or more) identical values will have the same number of values separating them. Below we will refer to the pattern as a *periodic* pattern, in the sense that, once the pattern starts, every p^{th} value in the sequence will be the same for (at least k occurrences). The threshold (or minimum sequence length) from which point these repetitions are guaranteed is known as the *Van der Waerden number*. The Van der Waerden number depends only on two values: (1) The number of distinct values that occur in the sequence, and (2) the number of repetitions, k that are desired. The sequence may, for example, be the sequence of states a process executes over time, where is may be in the ready queue (R), executing (E), blocked (B), suspended (S) or terminating (T). Its execution history may then, for example, be a sequence such as the following (using the letters behind the names of the states above:

```
R E B R E S E T
```

```
1 2 3 4 5 6 7 8    1 2 3 4 5 6 7 8 9    1 2 3 4 5 6 7 8 9
G R G R R G R G    G R G R R G R G R    G R G R R G R G G
                           R   R   R        G     G     G

      (a)                 (b)                  (c)
```

Figure 1. Van der Waerden example

In this example, the ‘alphabet’ consists of five values. To have a guaranteed periodic repetition that repeats, say $k = 100$ times, one only needs to determine the Van der Waerden value for an ‘alphabet’ of size 5 and a pattern of length 100. Very few Van der Waerden numbers are known, but upper bounds are simple to calculate.

Again using concepts from graph theory, the ‘alphabet’ is often deemed to be a set of colours — and hence, rather than talking about the size of the alphabet, one will simply refer to the number of colours in the sequence. Of course the colours may represent relationships between elements of some set (as it did in the introduction to Ramsey theory above). The sequence to which Van der Waerden’s theorem is used may, in the case of digital forensics, be the sequence of changes in relationships between entities deemed to be of interest for an examination.

To illustrate the concept, the Van der Waerden number for $k = 3$ repetitions based on two colours is 9. Suppose the two colours are red (R) and green (G). Then it is possible to construct a sequence of eight colours that have no periodic repetition of length $k = 3$. Consider, as an example, the string in Figure 1(a) (with the positions of the colours indicated above each colour). This sequence has no periodic repetitions. To extend it, the next item in the sequence has to be R or G. Since the Van der Waerden number is 9, a repeating pattern is guaranteed. If R is added, R occurs at positions 5, 7, and 9, as illustrated in Figure 1(b). In the language used above, from position 5 onwards, every second colour is red, and this is true for $k = 3$. In contrast, if G is added as the ninth colour, G occurs in positions 3, 6 and 9. Every third character (starting at position 3) is green and it repeats $k = 3$ times; this is depicted in Figure 1(c).

An important aspect of Van der Waerden’s theorem is illustrated by the example above: The theorem does not predict which value will recur, and does not predict the distance between the recurring values. However, it guarantees that a periodic pattern of the required length will be present in the sequence.

To present to work using more formal notations, assume that each member of a sequence of integers $\{1, 2, 3, \ldots, N\}$ is mapped to one of a finite number c of colours. Given a number k, a value w exists such that the numbers $\{1, 2, 3, \ldots, w\}$ contains at least k integers of the same colour that are equidistant from one another.

Let Σ be an alphabet consisting of c symbols. Let $s_1 s_2 s_3 \ldots s_n$ be a string over Σ. Then, for any value k, a value w exists such that the same symbol would be repeated at least k times at equidistant positions in the string. Stated differently, for any string of length w, there would be values j and p such that

$$s_j = s_{j+p} = s_{j+2p} = \ldots = s_{j+(k-1)p}$$

The smallest number for which every string produced has at least k periodic repetitions given an alphabet of size c is the Van der Waerden number, denoted as $W(c, k)$. Rather than using an alphabet consisting of symbols, it is useful in the current chapter to think of an alphabet consisting of colours. The value of $W(2, 3)$ is usually used to demonstrate the concept. It is easy to show that $W(2, 3) > 8$ since it is simple to produce a string using two symbols such that the same symbol does not occur at equidistant positions. As an example, consider the alphabet $\Sigma = \{R; G\}$ and the following string using the alphabet. (For ease of reference the position of symbols are indicated above the symbols.)

As noted (similar to the Finite Ramsey theorem), this theorem does not indicate which symbol (or colour) will be repeated. Few Van der Waerden values are known, but upper bounds have been established.

The paper by Calude and Longo expresses the (real) concern that the spurious regular pattern may be discovered and treated as a 'natural' law from which events in the future may be inferred. Recall that the (minimum) length k of the regular pattern can be determined arbitrarily and that any machine learning application that needs k inputs for learning (and testing), will learn the pattern and make highly accurate predictions within the repeated pattern.

Forensics may indeed use such a 'law', but often data analysis in digital forensics is retrospective.

Consider, as a simple example, a case where some incident occurs at some time t. A possible approach for someone investigating the incident is to collect as much data as possible leading up to the incident. Assume data is available from some time t_0. From the Van der Waerden theorem it is known that some regular pattern of at least length k exists in that data, with k only limited by the size of the available data.

One viable approach is to search the data for anomalies by working from time t backwards until some anomaly has been found (or no anomaly is found, and the start of the data is reached). Assume that the search for an anomaly stops at time $t' < t$ (without excluding the possibility that $t' = t_0$). Say the repeating pattern occurs from time t_a to time t_b. (Note that this does not suggest that all available data should be sorted according to time; however in many cases data about events will have an associated time or, at least, be ordered relatively. It may be useful to consider different strategies to 'visualise' the data being considered.[5]

Given the ever increasing size of available data it is possible to assume that in the general case that warrants a thorough investigation, sufficient data will be available to guarantee a pattern of length k, where k exceeds maximum sequences typically used for machine learning; in any case, if a longer k is required, more data is simply required and availability of data is generally not a problem. In days gone by, logs were destroyed because disc space was limited, but the cost of disc space has steadily decreased (reducing the need to delete data) and the growth of big data has disincentived data deletion merely because the data is 'old'.

5.3 The logic of inference

Suppose a spurious pattern is discovered — that is a pattern for which no causal reason exists.

As a temporal example, suppose that evidence is available from a time t_0 up to a time t_1. Say the incident occurred at a time t with $t_0 \leq t \leq t_1$. In order to simplify discussion, let us use two brackets to indicate the recurring pattern. A square bracket indicates that the pattern started at exactly the time written before or after it, while a round bracket indicates that some time elapsed. Hence $t_0[)t$ would indicate that the recurring pattern was present at the time of available evidence, but stopped some time before the incident. Similarly, $t[)t_1$ would indicate that the pattern started exactly when the incident occurred, but did not continue until the end of the period for which evidence exists. The notation remains readable without expressly mentioning t_0 and t_1, so the simplified expression of when the incident occurred will be used. (Of course, if the incident occurred repeatedly, the exposition becomes more complex, but a single occurrence will suffice for the current discussion.)

Any pattern that coincides with the incident will probably be deemed as significant. Hence $(]t$, $t[)$ and $t[]$ are likely to be seen as traces of cause or effect, with $(]$ possibly seen as causal traces and $t[)$ and $t[]$ seen as traces of effect. Note that such cause or effect interpretations would most probably be wrong, but would seem rather convincing. Similarly, a pattern that covers the incident — (t) — may (incorrectly) be seen as traces of some enabling condition.

More generally, the investigator may observe the pattern, and spend time to try and determine why the pattern disappeared (or began in the first place) in the hope that it might shed light on the case. If machine learning is deployed on the dataset, it may learn from the pattern what is deemed to be normal, and then flag subsequent values as anomalous.

The discussion above assumed that a spurious pattern was discovered and used for analysis. However, the starting point of the discussion was that the pattern was spurious. Therefore, it is, by definition, useless in the analysis of the case.

One possible defence for the use of patterns is that they may be useful as a starting point to search for causality. As noted in this chapter, this is indeed true — many laws of nature were first observed as patterns and later understood in causal terms. However, the underlying question in the current scenario is whether the search for patterns is, at least, useful as a mechanism to reduce the search space for causality.

The short answer is that there are too many patterns in a big data set; finding them all and testing them in some way for significance would simply be too time consuming.

For a somewhat more formal discussion of the notions being considered at this stage of the discussion let us assume that the relationship between data points are being classified the relationship expressed as a colour. Neither the arity of the relationship, nor the number of possible categories (or colours) into which such relationships can be classified are important for the current discussion. They merely have an effect on whether there is enough data such that the Van der Waerden theorem can be applied. While a more precise calculation is possible for a specific case, this chapter will assume that its setting in the big data context implies that sufficient data is available.

To be more concrete, assume that a bag of coloured relationships emerge and are arranged in a sequence S. The sequence is the result of pre-processing mentioned earlier; it may, in principle, be a temporal sequence of events, with information deemed to be of little significance removed, but any other mechanism to arrange the relationships would also be acceptable.

Suppose further that a pattern of, say, length n is deemed significant; n may depend on the machine learning techniques to be used, or any other prerequisite for significance. Let s be the number of elements in the sequence s. Let w_n be the Van der Waerden number that guarantees a pattern of length n. As implied earlier, the chapter assumes that $s/geqW_n$ from the context of big data in which the chapter is set.

Before continuing, it is important to reflect for a moment on the classification of a specific collection of data points into a particular class (or, in the language of graph theory, a particular colour that it shares with other collections of data points). Some classifications are straightforward: Using data communications as an example again, in a typical TCP/IP context, the expected port ranges for requests or responses, the direction of the request or response, and many other attributes can be classified as 'normal' or 'anomalous' without much debate. However, the question whether this particular classification scheme would be useful (or lead to the best possible evidence) is far from clear in a non-trivial case. In the bigger scheme of things, it is known that the corpora from which machine learning occurs often encode irrational categories. (See, for example, recent papers that illustrate how racism may be — and has been — learned through artificial intelligence [14, 8, 22]; confusion between patterns in criminal behaviour and patterns of criminal behaviour is just one example that may impact on corpora used to characterise crime).[6] The point is that classification in training sets often do include irrational assumptions that are propagated when machines learn these biases as factually correct, or does not disclose such bias (such as biased accuracy) in its classifications. For the purposes of the current chapter it is sufficient to take note that a somewhat different classification of relationships between data points will yield a different sequence S' of relationships, that may well contain one or more patterns that differ from what was observed in S.

From a pessimistic perspective, it is possible that up to s of the classifications made in the sequence S may be incorrect. If r colours are used then it is (obviously) possible to arrive at r^s colourings of a sequence of length s, of which the specific coloured sequence S is just one. Since $s \geq w_n$, each of these r^s will have a periodic pattern of at least length n, which would, in principle, make the pattern significant. While it should (hopefully) be possible to discard the bulk of these r^s colourings as nonsensical, demonstrating that they are all nonsensical will be a mammoth task. It is entirely possible that a single incorrect classification rule leads to a pattern that would not have existed. In addition, the pattern depends on the order of the relationships and other pre-processing tasks that are often based on the intuition of the person mining the big data set. If the pattern discovered in S forms incriminating evidence, how does the examiner show that a somewhat different — and possibly more accurate — classification of some relationships would not have lead to the discovery of an equally convincing pattern that would have served as the basis of exculpatory evidence. And the converse outcome, where incriminating evidence is overlooked and an exculpatory pattern found — based on a tiny misclassification — is equally serious.

In the context of evidence the potential existence of a meaningful patterns in s^r datasets, where s is already a large number, is sufficient to cast doubt on any pattern found. Unlike the small datasets considered earlier, the sheer number of possible patterns precludes exploring each as an alternative and excluding each. Any finding based on such a pattern should be approached with caution — it is far too easy for the opposing counsel to cast doubt on one's conclusions. The obvious exception is where there is a theoretical basis from forensic science that can speak to the significance of specific patterns. However, such patterns may be searched for in cases where they would be of help, rather than be discovered vis a process such as mining.

6. Conclusion

The increasing volume of data that may pertain to a criminal or civil matter of law is a well known challenge facing investigators of such cases. However, techniques associated with the *big data* movement thrive on large volumes of data; learning from such data is touted as a viable solution for many problems — even without fully understanding the problem.

This chapter used the same logic as Calude and Longo to explore the impact of the mere size of data on what may be discovered in big data. Using Ramsey theory and, more specifically, Van der Waerden's theorem, it was shown that spurious patterns are mathematically guaranteed to exist in large enough data sets. This implies that a discovered pattern may be spurious — in other words, it may be a function of the size of the data rather than the content of what the data purportedly represents. The discovery of a pattern does not exclude the discovery of other patterns that may contradict whatever was inferred from a discovered pattern. And it is computationally infeasible to find all patterns in big data.

If forensic conclusions are based on a pattern that has been found, the opposing side has a simple rebuttal for any such conclusion: How does the examiner know that a meaningful pattern has been examined? Without being able to justify the conclusion, there is no way of distinguishing between a meaningless result derived from a spurious pattern, and a correct, but unreliable result derived from a meaningful pattern.

Practitioners (and researchers) are therefore advised to avoid calls to jump on the bandwagon to use technologies of big data until such a time that resulting findings can be shown to yield evidence that is compatible with the requirements of presenting the truth, the whole truth, and nothing but the truth, which, by definition, has to be free from bias.

Notes

1. Typical examples of a call to use various 'intelligent' techniques look as follows: "AI in digital forensics ... does have a lot to offer the digital forensics community. In the short term it is likely that it can be immediately effective by the use of more complex pattern recognition and data mining techniques" [20]; "machine learning could play an important role in advancing these [code attribution and automated reverse engineering] research areas [11, p.S161]; "Artificial Intelligence (AI) is an area of computer science that has concentrated on pattern recognition and ... we highlighted some of the main themes in AI and their appropriateness for use in a security and digital forensics context" [21] and "AI is the perfect tool to aggregate information from the specifications for cyber security ... This use of AI will lift the burden of classification of these data for the cyber analyst and provide a faster and more effective result for determining who is to blame and how to respond" [28].

2. One should also remember the inherent privacy challenges posed by big data [24, 23].

3. Others claim that he obtained this incidence from a paper published in 1995 in The Lancet [6].

4. To use terminology correctly, one would not talk about a subgraph consisting of, say, blue edges, but rather about the subgraph induced by the blue edges. We use the shorter description here for the sake of simplicity.

5. As examples in which data may be visualised, consider any of the following options: the data may indeed be sorted as one long (linear) sequence of events; or the data from various logs may be placed in parallel 'line' so that the times of the various recorded events line up; or the data may be sorted according to event type (whether in one long line or in parallel lines); or the data may be subdivided into more lines with one line per user on whose authority the event occurs; or Where multiprocessors are used (including cloud computing) the data may be stratified per node and/or per instance; or the data may be ordered in any other way. Patterns may occur on a given time line, across time lines at some specific time, or involve various time lines in some systematic manner. None of this matters as far as the conclusion is concerned. However, thinking about such cases may make it simpler for the forensic examiner to intuitively accept that some pattern may indeed be discovered. The Van der Waerden Theorem guarantees that the pattern will be present.

6. Results have lead to a resolution in which Amazon "shareholders request that the Board of Directors prohibit sales of facial recognition technology to government agencies unless the Board concludes, after an evaluation using independent evidence, that the technology does not cause or contribute to actual or potential violations of civil and human rights" [1]. In addition various researchers published an open letter to "call on Amazon to stop selling Rekognition to law enforcement" until the measures are in place that prevent inherent racial bias in the tool to be abused [12]. Subsequently the State of California has enacted the The Body Camera Accountability Act (AB 1215) [9] that prohibits (until 2023) the use of facial recognition technology on cameras worn by law enforcement officers, with one of the primary grounds that *"Facial recognition and other biometric surveillance technology has been repeatedly demonstrated to misidentify women, young people, and people of color and to create an elevated risk of harmful 'false positive' identifications."*

References

[1] Amazon. Risks of sales of facial recognition software Amazon.com, Inc. — 2019. In *Notice of 2019 Annual Meeting of Shareholders To Be Held on Wednesday, May 22, 2019*, pages 18–22. Amazon,

Seattle, Washington, Apr. 2019. Item 6—Shareholder Proposal Requesting a Ban on Government Use of Certain Technologies.

[2] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, June 2008. `https://www.wired.com/2008/06/pb-theory/`.

[3] R. v. Sally Clark, Oct. 2000. Court of Appeal (Criminal Division), EWCA Crim 54 (2nd October, 2000); Case No: 1999/07495/Y3, Royal Courts of Justice, Strand, London.

[4] R. v. Sally Clark, Apr. 2003. Court of Appeal (Criminal Division), EWCA Crim 1020 (11th April 2003), Royal Courts of Justice, Strand, London.

[5] N. Beebe. Digital forensic research: The good, the bad and the unaddressed. In G. Peterson and S. Shenoi, editors, *Advances in Digital Forensics V*, volume 306 of *IFIP Advances in Information and Communication Technology*. Springer, 2009.

[6] P. Blair, P. Fleming, D. Bensley, I. Smith, C. Bacon, and E. Taylor. Plastic mattresses and sudden infant death syndrome. *Lancet*, 345(8951):720, Mar. 1995. Letter to the editor.

[7] R. O. Blanch. *Report of the Inquiry into the convictions of Kathleen Megan Folbigg*. State of New South Wales, Australia, July 2019.

[8] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[9] *An act to add and repeal Section 832.19 of the Penal Code, relating to law enforcement*, chapter 579. State of California, 2019. Approved by Governor October 08, 2019. Filed with Secretary of State October 08, 2019.

[10] C. S. Calude and G. Longo. The deluge of spurious correlations in big data. *Foundations of Science*, 22(3):595–612, Sept. 2017.

[11] J. Clemens. Automatic classification of object code using machine learning. *Digital Investigation*, 14, Supplement 1:S156–S162, Aug. 2015.

[12] “Concerned Researchers”. On recent research auditing commercial facial analysis technology, Mar. 2019. `https://medium.com/@bu64dcjrytwitb8/on-recent-research-auditing-commercial-facial-analysis-technology-19148bda1832`.

[13] Royal Statistical Society, The. Royal Statistical Society concerned by issues raised in Sally Clark case. News release, The Royal Statistical Society, Oct. 2001.

[14] K. Crawford and T. Paglen. Excavating AI: The politics of training sets for machine learning, Sept. 2019.

[15] S. D'Agostino. The architect of modern algorithms. *Quanta Magazine*, Nov. 2019.
`https://www.quantamagazine.org/barbara-liskov-is-the-architect-of-modern-al`

[16] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, and B. S. M. Potthast. Overview of the author identification task at PAN-2018 cross-domain authorship attribution and style change. In L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, editors, *Working Notes of CLEF 2018 — Conference and Labs of the Evaluation Forum*, volume 2125 of *CEUR Workshop Proceedings*, Avignon, France, Sept. 2018.

[17] W. Knight. Facebook's head of ai says the field will soon 'hit the wall'. *WIRED*, Dec. 2019.
`https://www.wired.com/story/facebooks-ai-says-field-hit-wall/`.

[18] P. Langley. The changing science of machine learning. *Machine Learning*, 82(3):275–279, Mar. 2011.

[19] R. Meadow. Fatal abuse and smothering. In R. Meadow, editor, *ABC of Child Abuse*, pages 27–29. BMJ, 3 edition, 1997.

[20] F. Mitchell. The use of artificial intelligence in digital forensics: An introduction. *Digital Evidence and Electronic Signature Law Review*, 7, 01 2014.

[21] F. R. Mitchell. An overview of artificial intelligence based pattern matching in a security and digital forensic context. In C. Blackwell and H. Zhu, editors, *Cyberpatterns*. Springer, 2014.

[22] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 429–435, New York, NY, USA, 2019. ACM.

[23] M. S. Olivier, "A layered architecture for privacy-enhancing technologies," *South African Computer Journal*, **31**, 53–61, 2003.

[24] M. S. Olivier, "Database privacy," *SigKDD Explorations*, **4**, 2, 20–27, 2003.

[25] M. Pollitt and A. Whitledge. Exploring big haystacks. In M. S. Olivier and S. Shenoi, editors, *Advances in Digital Forensics II*,

volume 222 of *FIP Advances in Information and Communication*. Springer, 2006.

[26] F. P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, s2-30(1):264–286, 1930.

[27] J. Smeaton. *Reports of the late John Smeaton, F.R.S., made on various occasions, in the course of his employment as a civil engineer*, volume II. M. Taylor, 2nd edition, 1837.

[28] J. Wulff. Artificial Intelligence and law enforcement. *Australasian Policing*, 10(1):16–23, 2018.