Digital Forensics and the Big Data Deluge - Some Concerns Based on Ramsey Theory

Martin Olivier

Chapter 1

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

Martin 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. A potential paradigm shift based on big data is looming – many researchers believe that massive volumes of data have enough substance to capture knowledge without the theories needed in earlier epochs. Patterns in big data are deemed to be sufficient to make predictions about the future, as well as about the past as a form of understanding. This chapter uses an argument developed by Calude and Longo [6] to critically examine the belief system of the proponents of data-driven knowledge, especially as it applies to digital forensic science.

From Ramsey theory it follows that, if data is large enough, knowledge is imbued in the domain represented by the data purely based on the size of the data. The chapter concludes that it is generally impossible to distinguish between true domain knowledge and knowledge inferred from spurious patterns that must exist purely as a function of data size. In addition, what is deemed a significant pattern may be refuted by a pattern that has yet to be found. Hence, evidence based on patterns found in big data is tenuous at best. Digital forensics should therefore proceed with caution if it wants to embrace big data and the paradigms that evolve from and around big data.

Keywords: Digital forensic science, 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, 2008 A.M. Turing Award Laureate [9]

> *"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, Vice President of Artificial Intelligence, Facebook [13]

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

In 1782, John Smeaton, a British engineer, offered his scientific knowledge of sea currents as evidence in a case involving the silting of the harbor at Wells-next-the-Sea in Norfolk [22]. At that time, evidence relying on, say Newton's work, would have been classified as hearsay evidence unless Newton was called to confirm it – a challenge because Newton passed away in 1727. Since 1782, science and expert witnesses have become entrenched in legal proceedings.

We are currently at another watershed moment in history. With the advent of big data, data science and deep learning, patterns are being uncovered at an increasing rate and are 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.

Numerous calls have been made to engage intelligent techniques:

> *"[Artificial Intelligence] 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"* [16].
>
> *"[M]achine learning could play an important role in advancing these [code attribution and automated reverse engineering] research areas"* [16].
>
> *"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"* [17].
>
> *"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"* [23].

However, from Ramsey theory, it is known that any dataset that is large enough will contain a multitude of regular patterns. The patterns stem from the size of the dataset, 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 could lead to a spurious system of (in)justice.

This chapter follows the logic of a generic argument by Calude and Longo [6] – based on Ramsey theory 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.

This chapter also discusses some aspects of patterns and repetitions with specific reference to inferences based on the patterns. This is illustrated using court cases where short patterns played a significant role. The chapter explores the guaranteed presence of (often spurious) patterns in large datasets. Finally, it illustrates the inherent dangers that arise if digital forensic findings are based on inferences from patterns in big data.

2. 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 the 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 long before one begins to doubt the integrity of the dice. Conversely, one does not expect that the same person will win a lottery on a fairly regular basis – if this were to happen, one would 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 to be anomalous.

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

In the examples above, the probabilities of the anomalies can be calculated rather accurately using basic probability theory and encountering them (on average) once in given periods of time or volumes are expected. More regular occurrences would, with very high probabilities, be indicative of anomalies.

However, as the chapter will explain, in a large dataset, data clusters that exhibit certain traits have to occur with mathematical certainty. The sizes and prevalence of the clusters are functions of data size and may be totally unrelated to what the data are purported to represent. It seems natural to denote the more prevalent clusters as normal and the less prevalent clusters as anomalies.

Such differentiation between normality and anomaly is often the basis of intrusion detection in computer networks and it is increasingly being applied in digital forensics. This claim will be substantiated below. However, if the occurrences of normal data and anomalies are due to the size of the data, rather than some justifiable theory, then the distinction between normality and anomaly is very tenuous at best (and would be wrong in many cases). If this is the case, such differences should not serve as the basis of scientific findings in forensic science.

To make matters more concrete, consider a web server request that contains 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 with anomalous requests. In addition, if lengthy requests can be linked to known vulnerabilities in servers, then the odds would increase that they are indeed malicious requests.

Another common pattern in intrusion detection involves a port scan. Methods for hiding port scans often interfere with some of the regular features in typical port scans. A port scan is often an indication of nefarious intention, unless the port scan was performed as part of an official security assessment.

Correlating anomalous events such as unusual web requests and port scans with reported computing incidents may be useful. However, it is important to remember that causality may also work in the other direction, where the incident causes the anomaly. A computer system that has lost connectivity typically makes an unusually large number of attempts to re-establish connectivity. More importantly for the purposes of this chapter, anomalous patterns may be entirely unrelated to incidents to which they apparently correlate and deriving any significance from the patterns would be incorrect. Making this case convincingly has to be postponed. Understanding the belief in patterns begins at a much simpler point – where a small correlation is just too significant to ignore.

2.1 Small Correlations

Unexpected patterns are often deemed significant even in small datasets. To the best of this author's knowledge, the interpretation of patterns in a cyber-related court case has not led to significant scrutiny of the presented evidence. Therefore, a well-known and widely discussed matter is used to reflect on the use of patterns as evidence in court.

Consider the infamous, now discredited, Meadow's law, which is based on patterns: "One sudden infant death is a tragedy, two is suspicious and three is murder, until proved otherwise" [15].

Meadow's law formed the basis of expert evidence in a number of cases. Arguably, the most prominent case was Regina v. Sally Clark [10]. Sally 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 Harry's death to be unnatural and revised his finding about Charles, whose death he also deemed to be unnatural.

Sir Samuel Roy Meadow (of Meadow's law fame) was an expert witness in the ensuing murder trial. His 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. However, she was released from jail in 2003 after a successful second appeal [11].

The pattern played a major role in Sally Clark's conviction and in the failure of her first appeal [10]. The judgment in the second appeal provides interesting insights into how the pattern was construed by the prosecution and jurors. This is discussed in more detail below.

2.2 Patterns and/or Knowledge

The previous paragraph illustrates that a potentially strong belief may be formed even when a very short pattern is considered. 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 in both appeals.

In contrast, machine learning, especially in the context of big data, has tended to ignore underlying knowledge and focus on patterns. Langley [14] 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, underlying knowledge about problem domains was significant. Knowledge representation was at the core of expert systems and domain-specific heuristics improved the speed of machine 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" [14].

In a similar vein, Anderson [1] published an article in *Wired* 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 this article, Anderson 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 has been a concern in the context of digital forensics ever since it emerged as an academic discipline [2]. Some of the earliest concerns were about finding the proverbial needles in haystacks as the sizes of the haystacks increased [18]. Dramatic increases in the amount of storage associated with computers have made comprehensive forensic imaging very difficult. The emergence of the cloud has only exacerbated the problem.

However, in parallel with these concerns, a new field of study developed under the big data rubric. The principle underlying this field is that the universe and aspects of it behave according to some patterns. If enough data is available, the analysis of the data can reveal the patterns. Once the patterns are known, behavior becomes predictable. This knowledge can be monetized or other benefits may be derived from it. Meanwhile, the name of the field has changed over time – data mining, data analytics, data science. Machine learning and deep learning are closely associated with the field. This chapter uses the term big data 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 digital forensics.

3. What Constitutes Correlation?

The Sally Clark case illustrates pattern recognition and correlation in a small dataset.

In the second appeal [11], the court pointed out that the previous courts (erroneously) accepted that the deaths of her 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 considered each of these points systematically and rejected 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 arguments. However, the incorrect reasoning in the original trial and the first appeal was only rectified by the second appeal [11].

In contrast, when an expert witness uses such methods, the court is ill equipped to deal with them, unless they are rebutted by other experts. 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 that the prevalence of Sudden Infant Death Syndrome (SIDS) was one in 8,543 cases. Some claim that Meadow obtained this incidence from a 1995 article in *Lancet* [3]. Hence, with the probability p of a SIDS case estimated to be $p = \frac{1}{8543}$, Meadow determined the probability of repeated cases by multiplying the estimated probability p by the number of cases, assuming that the occurrences of SIDS were independent.

In the Sally Clark case, Meadow concluded that the probability of two SIDS deaths would be p^2 – or about one in 73 million. He proceeded to illustrate the rarity of two SIDS deaths using a sports betting analogy. Although the judge downplayed the importance of this number in his instructions to the jury, its effect arguably stuck. Of course, two deaths

in a family may not be independent – they may have been due to the genetic makeup of the children – and hence, squaring the probability (without showing independence) was incorrect. This was one of the issues raised in a press release by the Royal Statistical Society [21] after the denial of the first appeal [10].

The second aspect raised by the Royal Statistical Society [21] was the emphasis on the 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 a sequence of events S and proceed to conclude that another unlikely sequence of events B is the logical inference.

As a second example, consider the case of Australian, Kathleen Folbigg. Four of her children died very young: the first in 1989 at age 19 days, the second in 1991 at eight months, the third in 1993 at ten months and the fourth in 1999 at 19 months. While experts used the same calculations as Meadow during pretrial hearings, by the time Folbigg's trial started in March 2003, the British Court of Appeals had already discredited Meadow's law and calculations.

Meadow's law was excluded by the court, but his ideas nevertheless featured during the trial. A Professor Berry 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." On the other hand, a Professor Herdson deemed the events to be too different to correspond to a pattern in which SIDS deaths would occur, and used the absence of a specific pattern (amongst others) to be indicative of unnatural causes of death.

In the Sally Clark and Kathleen Folbigg cases other evidence was influential in the eventual findings of the various courts. In fact, this other evidence was eventually more important than the presence or absence of patterns.

In the Sally Clark case, microbiological test results for Harry were not available to the defense and were only discovered by them after the first appeal. The second appellate court found that the 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 prosecution did not apply for a re-trial and the convictions were set aside.

In the Kathleen Folbigg case, diaries that she maintained played a significant role in the proceedings and the outcome of the trial. Public

interest eventually led to a judicial inquiry by Reginald Blanch, former Chief Judge of the New South Wales District Court, who reviewed the case and heard new evidence. In his July 2019 report, Reginald Blanch concluded that "the Inquiry does not cause me 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 behavior 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" [4].

4. Correlation in Big Data

Many papers express concern about or reject the notion that data can speak for itself without the need for a theory. One only has to look through the many papers that cite Anderson's claim [1] to find such critiques.

Calude and Longo [6] make a critique that should be taken seriously in digital forensics. They "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 by Calude and Longo].

Calude and Longo use a number of theorems from Ramsey theory and ergodic theory that are relevant in the current context. This chapter only focuses on the final claim made by Calude and Longo that is based on Ramsey theory, but a different exposition is provided.

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 involves a scenario where people attend a party. Any two people at the party will either have met previously or be mutual strangers. If colors are used to represent the relationships between pairs of people, the case where they have previously met may be represented by the color green while the case where they are mutual strangers may be represented by the color 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 color (or, stated differently, to have c monochromatic cases).

If, for example, c is chosen to be one, it is easy to show that $n = 2$. Specifically, the relationship between two attendees a and b can be

represented graphically as an edge between vertices a and b; the edge is green if they know each other and red if they are mutual strangers.

Furthermore, if $c = 2$ then $n = 3$. Specifically, attendees a, b and c can be depicted graphically as a triangle with vertices a, b and c, and edges (a, b), (a, c) and (b, c) whose colors represent the relationships. Since there are two colors (red and green) and three edges, at least two edges must have the same color.

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

As illustrated by the party problem, it is natural to think about Ramsey theory in terms of graphs. In graph theory, a 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 – where all the vertices are connected. In this context, the task is to color a complete graph using two colors. One color (say green) is used to color an edge if the relationship holds between the vertices connected by the edge; the other color (say red) is used to color an edge if the relationship does not hold between the two connected vertices. Then, the Ramsey number $R(s, t)$ is the smallest n such that graph K_n must either contain a clique of s (or larger) with green edges or a clique of size t (or larger) with red edges. Note that, instead of saying that a subgraph consists of, say, green edges, it is more appropriate to say that a subgraph is induced by red edges. The former term is used here for reasons of simplicity.

In general, the binary relationship used above – that some relationship holds or does not hold – is too restrictive. It is useful 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 events that occurred in a computer system under investigation, then the time between the events may for some reason be deemed to be a possibly relevant relationship. As an arbitrary example, events that occurred hours apart, minutes apart and seconds (or less) apart form such a partition – assuming a definition of time exists 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 [6]) is based on undirected graphs, where the relationships between objects or events are symmetric. An appropriate example is the time between events. However, the question of whether an event preceded another event, coin-

cided with it or followed it is asymmetric and is, therefore, not covered by the current discussion. In any case, the exclusion of asymmetric relationships is not material in this chapter.

5.1 Finite Ramsey Theorem

In 1930, Ramsey [20] proved the following theorem that is the foundation of the theory carrying his name:

> *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 Δ_n belong to the same C_i."*

An r-combination is a set of r elements that occur in a dataset. If the dataset contains the values $\{a, b, c, d\}$, then the 3-combinations 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 colors, as used previously).

An analogy with the training phase of supervised machine learning can provide insights into the theorem. In supervised learning, a number of inputs are provided to a classifier along with the class associated with the inputs. Let r inputs be used for each instance to be classified and let every instance be assigned to one of the μ classes. Let n be some number that is chosen. Then, 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 learning that may occur. It merely 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. Upper bounds can be determined for cases where it cannot (yet) be calculated precisely.

The fact that a certain relationship between members of some set holds relatively often in a dataset may be of interest in unravelling an incident. Ramsey's theorem warns us to proceed with care. However, it seems much more likely that an activity of interest in a digital forensic investigation would consist of several actions that together constitute

an anomalous (or otherwise useful) indication of what transpired (or is otherwise useful).

For example, in a case involving network communications, a message may be deemed to be 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.

As a more concrete example, consider the problem of authorship attribution, which often uses contiguous sequences of linguistic elements called n-grams. These elements may be letters, words, word pairs, phonemes or other entities that experimentally turn out to be useful. In a 2018 authorship attribution competition [12], "n-grams were the most popular type of features to represent texts in" one of the primary tasks in the competition. "More specifically, character and word n-grams [were] used by the majority of the participants."

Although the Finite Ramsey theorem does not play a significant role in the remainder of this chapter, it sets the stage for the Van der Waerden theorem of 1927, which is part of Ramsey theory. Once again, the logic of Calude and Longo [6] is employed.

5.2 Van der Waerden's Theorem

The Finite Ramsey theorem provides a threshold beyond which a certain number of relationships among the members of a set is guaranteed. In contrast, Van der Waerden's theorem 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 or longer than the computed threshold, 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 states 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. This pattern is referred to as a periodic pattern, in the sense that, once the pattern starts, every p^{th} value in the sequence is the same for at least k occurrences. The threshold (or minimum sequence length) from which point the repetitions are guaranteed is known as the Van der Waerden number. The Van der Waerden number depends only on two values: (i) number of distinct values that occur in the sequence; and (ii) number of repetitions k that are desired. The sequence may correspond to a series of process states, where a process is in the ready queue (`R`), executing

(E), blocked (B), suspended (S) or terminating (T). Its execution history may correspond to the process sequence:

R E B R E S E T

where the process states are listed using the first letters of their names.

In this example, the alphabet has five values. To have a guaranteed periodic repetition that repeats, say $k = 100$ times, it is only needed to determine the Van der Waerden value for an alphabet of size five and a pattern of length 100.

Again using concepts from graph theory, the alphabet can be a set of colors and, rather than talking about the size of the alphabet, it is more convenient to simply refer to the number of colors in the sequence. Of course, the colors may represent relationships between elements of some set (as it did in the 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 in an examination.

The Van der Waerden number for $k = 3$ repetitions based on two colors is 9. Assume that the two colors are red (R) and green (G). Then, it is possible to construct a sequence of eight colors that have no periodic repetition of length $k = 3$.

```
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      G     G     G

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

Figure 1. Van der Waerden example.

Consider the string in Figure 1(a) where the positions of the colors R and G are indicated above each color. The sequence has no periodic repetitions.

To extend the sequence, 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 shown in Figure 1(b). In the language used above, from position 5 onwards, every second color is R and this is true for $k = 3$. In contrast, if G is added as the ninth color, the G occurs in positions 3, 6 and 9. Every third character (starting at position 3) is G and it repeats $k = 3$ times as shown in Figure 1(c).

An important aspect of Van der Waerden's theorem is illustrated by the example above. Specifically, the theorem does not predict which

value will recur and it 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 the work using more formal notation, assume that each member of a sequence of integers $\{1, 2, 3, \ldots, N\}$ is mapped to one of a finite number of colors c. Given a number k, a value w exists such that the numbers $\{1, 2, 3, \ldots, w\}$ contain at least k integers of the same color that are equidistant from each other.

Let Σ be an alphabet with c symbols. Let $s_1 s_2 s_3 \ldots s_n$ be a string on 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, which is denoted as $W(c, k)$. The value of $W(2, 3)$ is used to demonstrate the concept. It is easy to show that $W(2, 3) > 8$ because it is simple to produce a string using two symbols such that the same symbol does not occur at equidistant positions.

As with the Finite Ramsey theorem, Van der Waerden's theorem does not indicate which symbol (or color) will be repeated. Few Van der Waerden values are known, but upper bounds have been established.

Calude and Longo [6] express 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 a case where an incident occurs at time t. An investigator would collect as much data as possible leading up to the incident. Assume that data is available from time t_0. From the Van der Waerden theorem it is known that some regular pattern of at least length k exists in the data, with the value k limited only by the size of the available data.

A viable approach is to search the data for anomalies by working from time t backwards until an anomaly has been found or no anomaly is found if the start of the data has been reached. Assume that the search for an anomaly stops at time $t' < t$ without excluding the possibility

that $t' = t_0$. Also, assume that the repeating pattern occurs from time t_a to time t_b. Note that this does not suggest that all the available data should be sorted according to time; however, in many cases, data about events would have an associated time or, at least, be ordered relatively.

At this point, it is instructive to consider strategies for visualizing the data. The options include: (i) data may be sorted as one long (linear) sequence of events; (ii) data from various logs may be placed in parallel lines so that the times of the various recorded events line up; (iii) data may be sorted according to event type (whether in one long line or in parallel lines); (iv) data may be subdivided into more lines with one line per user on whose authority the event occurs; (v) data may be stratified per node and/or per instance when multiprocessors or cloud computing are used; or (vi) data may be ordered in some 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 these matters as far as the conclusion is concerned. However, thinking about such cases may make it simpler for a digital forensic practitioner to intuitively accept that a pattern may indeed be discovered. Van der Waerden' theorem guarantees that a pattern will be present.

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 the maximum sequences typically used in machine learning. In any case, if a longer k is required, more data would be needed and the availability of this data would not be a problem. In days gone by, logs were destroyed because storage space was limited, but storage capacities have increased significantly while storage costs have decreased, eliminating the need to delete logged data. Moreover, the growth of big data has disincentivized data deletion merely because the data is old.

5.3 Logic of Inference

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

As a temporal example, assume that evidence is available from time t_0 up to time t_1. Assume that the incident occurred at time t with $t_0 \leq t \leq t_1$.

In order to simplify the discussion, two brackets are used to indicate a recurring pattern. A square bracket indicates that the pattern started at exactly the time written before or after it whereas a round bracket indicates that some time has elapsed. Thus, $t_0[)t$ would indicate that the recurring pattern was present at the time of the available evidence

was collected, 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 during which the evidence was collected. 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 would become more complex, but a single occurrence will suffice for the current discussion.

Any pattern that coincides with the incident would likely be deemed 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 and effect interpretations would most probably be wrong, but would appear to be 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 attempt to 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 flag subsequent values as anomalies.

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, by definition, it is useless in the analysis of the case.

One possible defense for the use of patterns is that they may be useful as starting points 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 dataset. Finding all the patterns and testing them for significance would be too time consuming.

For a more formal discussion, assume that the relationships between data points are expressed as colors. Neither the arity of the relationships nor the number of possible categories (or colors) into which the relationships can be classified are important in the current discussion. They merely have an effect on whether there is enough data to enable the application of Van der Waerden's theorem. While a more precise calculation is possible for a specific case, the assumption is that the big data context implies that sufficient data is available.

To be more concrete, assume that a bag of colored relationships emerges and that the elements of the bag are arranged in a sequence S. The sequence is the result of the pre-processing mentioned earlier. It may be a temporal sequence of events with information of little significance eliminated or some other mechanism would be used to arrange the relationships.

Assume that a pattern of length n is deemed significant, where the value of n may depend on the machine learning technique to be used or any some other prerequisite for significance. Let s be the number of elements in a sequence. Let w_n be the Van der Waerden number that guarantees a pattern of length n. As implied earlier, it is assumed that $s \geq W_n$ in the context of big data.

Before continuing, it is important to reflect on the classification of a specific collection of data points into a particular class (or, in the language of graph theory, a particular color that it shares with other collections of data points). Some classifications are straightforward. For example, in the TCP/IP networking context, the expected port ranges for requests or responses, directions of requests or responses, 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 the big 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 [5, 8, 19]. Indeed, confusion between patterns in criminal behavior and patterns of criminal behavior is just one example that may impact corpora used to characterize crime.

The point is that classifications of training sets often engage irrational assumptions that are propagated when machines learn the biases as factually correct or the machines do not disclose the biases (e.g., biased accuracy) in their classifications. For the purposes of this work, it is sufficient to note that a different classification of relationships between data points would yield a different sequence S' of relationships, which may contain one or more patterns that differ from the patterns 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 colors are used, then it is possible to arrive at r^s colorings of a sequence of length s, of which the specific colored sequence S is just one of the sequences. Since $s \geq w_n$, each r^s would have a periodic pattern of at least length n, which would make the pattern significant. While it should be possible

to discard the bulk of these r^s colorings as nonsensical, demonstrating that they are all nonsensical would be a mammoth task.

It also possible that a single incorrect classification rule could lead to a pattern that would not have existed. In addition, a pattern depends on the order of the relationships and other pre-processing tasks that are often based on the intuition of the individual who mines a large dataset. If the pattern discovered in S is incriminating evidence, how does the investigator show that a somewhat different – and possibly more accurate – classification of relationships would not have led to the discovery of an equally convincing pattern that may be exculpatory evidence? 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 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 the exploration of each pattern as an alternative and keeping or excluding it. 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 the conclusions. The obvious exception is when a theoretical basis from forensic science exists that can speak to the significance of specific patterns. However, such patterns should be searched for in cases where they would be of help, rather than be discovered via a process such as data mining.

6. Conclusions

The increasing volumes of data that pertain to criminal and civil matters is a well-known challenge facing investigators. However, big data techniques thrive on large volumes of data and learning from such data is touted as a viable solution for many problems, even when the problems are not fully understood.

This chapter has used the same logic as Calude and Longo to explore the impact of data size on what may be discovered in the data. Ramsey theory and, more specifically, Van der Waerden's theorem demonstrate that spurious patterns are mathematically guaranteed to exist in large enough datasets. This implies that a discovered pattern may be spurious – in other words, it may be a function of the size of the data instead of the content that the data purportedly represents. The discovery of a pattern does not exclude the discovery of other patterns that may contradict what was inferred from a discovered pattern. And, of course, it is computationally infeasible to find all the 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 investigator know that a meaningful pattern has been examined? Without being able to justify the conclusion, there is no way to distinguish between a meaningless result derived from a spurious pattern and a correct, but unreliable, result derived from a meaningful pattern.

Digital forensic practitioners and researchers would be well advised to avoid calls to jump on the big data bandwagon and wantonly use its technologies until the 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, must be free from bias.

References

[1] C. Anderson, The end of theory: The data deluge makes the scientific method obsolete, *Wired*, June 23, 2008.

[2] N. Beebe, Digital forensic research: The good, the bad and the unaddressed, in *Advances in Digital Forensics V*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 17–36, 2009.

[3] P. Blair, P. Fleming, D. Bensley, I. Smith, C. Bacon and E. Taylor, Plastic mattresses and sudden infant death syndrome, *Lancet*, vol. 345(8951), p. 720, 1995.

[4] R. Blanch, Report of the Inquiry into the Convictions of Kathleen Megan Folbigg, State of New South Wales, Parramatta, Australia (`www.folbigginquiry.justice.nsw.gov.au/Documents/Report%20of%20the%20Inquiry%20into%20the%20convictions%20of%20Kathleen%20Megan%20Folbigg.pdf`), 2019.

[5] J. Buolamwini and T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, *Proceedings of Machine Learning Research*, vol. 81, pp. 77–91, 2018.

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

[7] J. Clemens, Automatic classification of object code using machine learning, *Digital Investigation*, vol. 14(S1), pp. S156–S162, 2015.

[8] K. Crawford and T. Paglen, Excavating AI: The Politics of Training Sets for Machine Learning, *Excavating AI* (`www.excavating.ai`), September 19, 2019.

[9] S. D'Agostino, The architect of modern algorithms, *Quanta Magazine*, November 20, 2019.

[10] England and Wales Court of Appeal (Criminal Division), Regina v. Sally Clark, EWCA Crim 54, Case No: 1999/07495/Y3, Royal Courts of Justice, London, United Kingdom, October 2, 2000.

[11] England and Wales Court of Appeal (Criminal Division), Regina v. Sally Clark, EWCA Crim 1020, Case No. 2002/03824/Y3, Royal Courts of Justice, London, United Kingdom, April 11, 2003.

[12] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht and B. Potthast, Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection, in *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*, L. Cappellato, N. Ferro, J. Nie and L. Soulier (Eds.), Volume 2125, CEUR-WS.org, RWTH Aachen University, Aachen, Germany, 2018.

[13] W. Knight, Facebook's head of AI says the field will soon "hit the wall," *Wired*, December 4, 2019.

[14] P. Langley, The changing science of machine learning, *Machine Learning*, vol. 82(3), pp. 275–279, 2011.

[15] R. Meadow, Fatal abuse and smothering, in *ABC of Child Abuse*, R. Meadow (Ed.), BMJ Publishing Group, London, United Kingdom, pp. 27–29,1997.

[16] F. Mitchell, The use of artificial intelligence in digital forensics: An introduction, *Digital Evidence and Electronic Signature Law Review*, vol. 7, pp. 35–41, 2010.

[17] F. Mitchell, An overview of artificial intelligence based pattern matching in a security and digital forensic context, in *Cyberpatterns*, C. Blackwell and H. Zhu (Eds.), Springer, Cham, Switzerland, pp. 215–222, 2014.

[18] M. Pollitt and A. Whitledge, Exploring big haystacks, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 67–76, 2006.

[19] I. Raji and J. Buolamwini, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products, *Proceedings of the AAAI/ACM Conference on AI, Ethics and Society*, pp. 429–435, 2019.

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

[21] Royal Statistical Society, Royal Statistical Society concerned by issues raised in Sally Clark case, News Release, London, United Kingdom, October 23, 2001.

[22] 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*, Longman, London, United Kingdom, 1812.

[23] J. Wulff, Artificial intelligence and law enforcement, *Australasian Policing*, vol. 10(1), pp. 16–23, 2018.