The Deluge of Spurious Correlations in Big Data*

Cristian S. Calude
Department of Computer Science, University of Auckland
Auckland, New Zealand
www.cs.auckland.ac.nz/~cristian

Giuseppe Longo
Centre Cavaillès (République des Savoirs), CNRS,
Collège de France & École Normale Supérieure Paris, France
Department of Integrative Physiology and Pathobiology
Tufts University School of Medicine Boston, USA
http://www.di.ens.fr/users/longo

January 22, 2016

...fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza.
Dante Alighieri, Inferno, Canto XXVI.[1]

Abstract

Very large databases are a major opportunity for science and data analytics is a remarkable new field of investigation in computer science. The effectiveness of these tools is used to support a "philosophy" against the scientific method as developed throughout history. According to this view, computer-discovered correlations should replace understanding and guide prediction and action. Consequently, there will be no need to give scientific meaning to phenomena, by proposing, say, causal relations, since regularities in very large databases are enough: "with enough data, the numbers speak for themselves". The "end of science" is proclaimed. Using classical results from ergodic theory, Ramsey theory and algorithmic information theory, we show that this "philosophy" is wrong. For example, we 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 "randomly" generated, large enough databases, which – as we will prove – implies that *most correlations are spurious*. Too much information tends to behave like very little information. The scientific method can be enriched by computer mining in immense databases, but not replaced by it.

*To appear in *Foundations of Science*.

[1] Italian: ... you were not born to live like brutes, but to pursue virtue and knowledge.

1 Data deluge

Speaking at the Techonomy conference in Lake Tahoe, CA in August 2010, E. Schmidt, then Google CEO, estimated that every two days humanity creates a quantity of data equivalent to the entire amount produced from the dawn of time up to 2003 [46].

According to Grossman [23], every day humanity generates 500 million tweets, 70 million photos on Instagram, and 4 billion videos on Facebook.

Under the title "What is big data?" IBM researchers estimate in [25] that

> Every day, we create 2.5 quintillion bytes of data – so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

From scarcity and difficulty to find data (and information) we now have a deluge of data.

2 Data science is the end of science

In June 2008, C. Anderson, former editor-in-chief of *Wired Magazine*, wrote a provocative essay titled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" in which he states that "with enough data, the numbers speak for themselves". "Correlation supersedes causation, and science can advance even without coherent models, unified theories", he continues. Anderson is not unique in stressing the role of petabytes, stored in the cloud, in replacing the scientific method. Succinctly, George Box's maxim that *all models are wrong, but some are useful* is replaced by *all models are wrong, and increasingly you can succeed without them.*[2]

The idea behind this new "philosophy" is that sufficiently powerful algorithms can now explore huge databases and can find therein correlations and regularities. Independently of any analysis of "meaning" or "content" of such correlations – which are notions very difficult to define – rules for prediction and, possibly for action, are then provided by the machine. The strength and generality of this method relies on the immense size of the database: the larger the data, the more powerful and effective is the method grounded on computationally-discovered correlations. Consequently, there is no need to theorise, understand, criticise ... the sense of the discovered correlations: "No semantic or causal analysis is required". According to Anderson, petabytes allow us to say: "Correlation is enough .... We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot".

A moderate, yet criticised[3], position was taken by the NSF [37]:

> Computational and Data-Enabled Science and Engineering (CDS&E) is a new program. CDS & E is now clearly recognizable as a distinct intellectual and technological discipline lying at the intersection of applied mathematics, statistics, computer

[2] Anderson attributed the last sentence to Google's research director Peter Norvig, who denied it [36]: "That's a silly statement, I didn't say it, and I disagree with it."

[3] See more in [4].

> science, core science and engineering disciplines ... We regard CDS&E as explicitly recognizing the importance of data-enabled, data-intensive, and data centric science. CDS&E broadly interpreted now affects virtually every area of science and technology, revolutionizing the way science and engineering are done. Theory and experimentation have for centuries been regarded as two fundamental pillars of science. It is now widely recognized that computational and data-enabled science forms a critical third pillar.

The big target of the "data-enabled, data-intensive, and data centric science" new "philosophy" is science itself. The scientific method is built around hypotheses derived from observations and "insight". It actually started when humans looking at the stars and the moon tried to understand the sky by myth and theory, an endeavour which gave us not only knowledge, but also human sense to our biological life. Theories are then constructed while new objects of knowledge, such as electromagnetic waves or quanta, are proposed and studied. Experiments help to falsify hypotheses or their theoretical consequences and, when this is possible, suggest mathematical models.

The position pioneered by Anderson and his followers explicitly announces the "End of Science". Why? Because, according to them, science founded on models is relatively arbitrary and often wrong as it is based on adductive and deductive praxis, as well as excessive critical thinking that leaves room even for aesthetic judgements (for example, "appreciations of symmetries" in the broadest sense, from Euclid to H. Weyl and R. Feynman, to modern biology, see [30]). The conclusion is that

> science as we know it will be replaced by robust correlations in immense databases.

3 The power and limits of correlations

In general, "co-relation" denotes phenomena that relate covariantly, that is, they vary while preserving proximity of values according to a pre-given measure. "Co-relation" is essentially "co-incidence", that is, things that occur together.

For example, a correlation can be a relationship between two variables taking (numeric) values. If one variable value increases, then the other one also increases (or decreases) "in the same way". For example, "as the temperature goes up, ice cream sales also go up" or "as the temperature decreases, the speed at which molecules move decreases" or "as attendance at school drops, so does achievement" or "as weather gets colder, air conditioning costs decrease".

Correlations can be useful because of their potential predictive power: *use or act on the value of one variable to predict or modify the value of the other.*

From about the 10th century BC onwards, in the region from Korea and China to Morocco, early astronomers detected various regularities, from which they were able to compute future observations of astronomical bodies. The observations and computational methods were gradually improved over the millennia, so that by about 1200 some astronomers could make reliable predictions of solar eclipses [34]. In the 10th century AC, Chinese doctors remarked that mild smallpox infection could prevent more severe illnesses. They were later able to confirm this correlation by successfully experimenting with smallpox inoculation on large numbers of individuals [35, pp. 114–174].

We must acknowledge the extraordinary insights of these very early observations of non-obvious correlations and their practical relevance. More scientific knowledge has been added

to them: a) Clairaut's mathematical derivation of the orbit of Halley's comet, and its delays, on the grounds of Newton's laws and equations, in the 18th century, b) the understanding of microbial infections and vaccines in the 19th century. These works could frame the original correlations into relatively robust theories that gave them "scientific meaning". This scientific method, which allows unification and generalisation, can also detect correlations that may be considered spurious – like the relation between the orientation of a comet's tail and the Emperor's chances of a military victory.

In a 1992 study at the University of Illinois [1] 56,000 students were asked about their drinking habits and grades. The results indicate a correlation between these two variables: the more a student drinks, the more her/his GPA goes down. Based on the established correlation one could predict that a student who drinks a bottle of wine every day would be likely to flunk out of school.

Pitfalls of exaggerating the value of prediction based on correlated observables have been discussed in the literature for many years. For example, the conclusion of Ferber's 1956 analysis [14] is:

> Clearly the coefficient of correlation is not a reliable measure for [the practical problem of selecting functions (hypotheses) for predictive purposes], nor does there appear to be at the present time any alternative single statistic adequate in this respect relating to the period of observation. Unsettling as it may seem, there does not appear to be any statistical substitute for a priori consideration of the adequacy of the basic hypothesis underlying a particular function.

Even more importantly, it is well-known that *correlation does not imply causation*. In the analysis of the Illinois survey referred above, one notes [1]:

> However, a correlation does not tell us about the underlying cause of a relationship. We do not know from the Illinois data whether drinking was correlated with lower grades because (1) alcohol makes people stupid, or (2) the students who tend to drink tend to be poorer students to begin with, or (3) people who are hung-over from a drinking binge tend to skip class, or (4) students in academic trouble drink in order to drown their sorrows, or some other reason. There can be hundreds of possible explanations for a correlation: the number is limited only by your imagination and ingenuity in thinking up possible reasons for a relationship between two variables.

Then the purely syntactic/relational perspective, with no "understanding", is lucidly spelled out in [1]:

> For purposes of making a prediction, the underlying reason for a correlation may not matter. As long as the correlation is stable – lasting into the future – one can use it to make predictions. One does not need an accurate cause-effect explanation to make a prediction. As long as a "pattern" (correlation) continues into the future, we can use it to make a prediction, whether or not we understand it.
>
> What a correlation does not tell you is why two things tend to go together. Maybe alcohol consumption is not the root cause of bad grades, in the Illinois study. Perhaps the students who drank never opened their books and never studied, and that is why they got bad grades. We do not know. The study did not seek to explore underlying causes of the correlation.

The following example quoted from [38] is worrying:[4]

> A 2010 study [44] conducted by Harvard economists Carmen Reinhart and Kenneth Rogoff reported a correlation between a country's economic growth and its debt-to-GDP ratio. In countries where public debt is over 90 percent of GDP, economic growth is slower, they found. The study gathered a lot of attention. Google Scholar listed 1218 citations at the time of writing, and the statistic was used in US economic policy debates in 2011 and 2012, [48][5].
>
> The study isn't conclusive, though – in fact, it's far from it. As noted by John Irons and Josh Bivens of the Economic Policy Institute, it's possible that the effect runs the other way round, with slow growth leading to high debt. Even more worryingly, the research didn't hold up on replication [53]. But by the time that became clear, the original study had already attracted widespread attention and affected policy decisions.

As a last example in Figure 1 we present a hilarious high correlation.

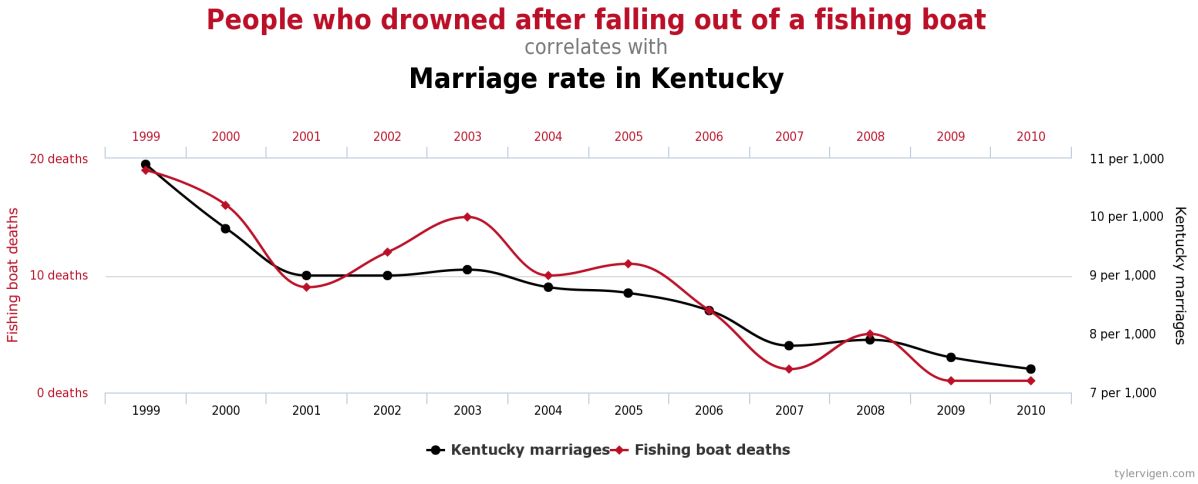

Figure 1: A correlation with $r = 0.952407$ [2].

In spite of long-known facts pointing to the contrary, the core of the recent argumentation is that because of the possibility of mining huge databases, *correlation supersedes causation*. Shortcomings of this position have been discussed in the literature. For example, Poppelaar's blog from 5 April 2015 titled "Do numbers really speak for themselves with big data?" [41] concludes that

[4]This example points to another important issue: no data collecting is strictly objective – see the analysis in [22] of Reinhart and Rogoff's bias in their collection of data in several countries for 218 years.

[5]European policy makers largely referred to that paper till 2013. For example, O. Rehn, EU Commissioner for Economic Affairs (2009–13) referred to the Reinhart-Rogoff correlation as a key guideline for his past and present economic views ([51], address to ILO, April 9, 2013) and G. Osborne, British Chancellor of the Exchequer (since 2010), claimed in April 2013: "As Rogoff and Reinhart demonstrate convincingly, all financial crises ultimately have their origins in one thing [the public debt]." [32].

> With data, enough computing power and statistical algorithms patterns will be found. But are these patterns of any interest? Not many of them will be, as spurious patterns vastly outnumber the meaningful ones.[6] Anderson's recipe for analysis lacks the scientific rigour required to find meaningful insights that can change our decision making for the better. Data will never speak for itself, we give numbers their meaning, the Volume, Variety or Velocity of data cannot change that.

The nomothetic studies that seek relationships between demographic variables and cultural traits by mining very large, cross-cultural datasets [45] is another example among many discussed in the literature, see [15, 16, 28, 5].

Our analysis of the big data-driven "philosophy" will be presented in mathematical terms. Specifically, we will use a) ergodic theory to show that one cannot reliably predict by "analogy" with the past, even in deterministic systems, chaotic or not, and b) Ramsey theory to prove that, given any arbitrary correlation on sets of data, there exists a large enough number (size) such that *any* data set larger than that size realises that type of correlation. Since this large-enough data set is arbitrary, it could have been obtained by a random number generator (series of coin tosses or measurements of quantum fluctuations of the vacuum),[7] which, by design, excludes any underlying structure or rule [6, 7] that could justify action through possible future extensions. Note that it is exactly *the size of the data* that allows our result: *the more data, the more arbitrary, meaningless and useless (for future action) correlations will be found in them.*

Thus, paradoxically, the more information we have, the more difficult is to extract meaning from it. Too much information tends to behave like very little information.

4 The goal of the paper

Our goal is not to criticise the fields of data science, data mining or data analytics *per se*. As discussed in the previous section, the correlational method is a scientific praxis with roots going far back in human history. It is also clear that learning how to analyse extremely large data sets correctly and efficiently will play a critical role in the science of the future, and even today. Rather, *our aim is to document the danger of allowing the search of correlations in big data to subsume and replace the scientific approach.*

Our analysis will first use an argument from dynamical systems in [8], then employ results from Ramsey theory [20] and algorithmic information theory [6, 12], theories[8] which provide the foundations of the behaviour – power and limits – of the algorithms, in particular of those dealing with large sets of numbers.

5 Ergodic theory

One of the main ideas supporting data analytics is that a series of correlations will continue or iterate similarly along the chosen parameter (recurrence). If, for example, time is the main parameter (as in Figure 1), then the correlation will extend into the future by iterating a similar

[6] See also the huge collection of spurious correlations [2] and the book [55] based on it, in which the old rule that "correlation does not equal causation" is illustrated through hilarious graphs.

[7] This was informally observed also in [50, p. 20]: "With fast computers and plentiful data, finding statistical relevance is trivial. If you look hard enough, it can even be found in tables of random numbers".

[8] They are branches of finite combinatorics and the theory of algorithms, respectively.

"distance", typically, between the chosen observables. The recurrence of correlation is a rather natural phenomenon which applies to many familiar events, like diurnal and seasonal cycles and their observables consequences. This idea and the belief in determinism – from the same antecedents follow the same consequents – more or less implicitly justify the prediction methods based on the recurrence of regularities.

The debate on recurring (time) series has a long history, going back to the end of the 19th century studies in the geometry of dynamical systems (Poincaré) and statistical physics (Boltzmann). Poincaré proved the unpredictability of deterministic non-linear systems: physical dynamics (orbits) where minor fluctuations (i.e. below measurement) could yield measurable but unpredictable evolutions (the origin of chaos theory, see [11, 7]). Moreover, he also proved the fundamental "recurrence theorem" in ergodic theory.[9] This theorem roughly states that in any deterministic system, *including chaotic systems*, the future, soon or late, will be analogous to the past (will somehow iterate). More precisely, for every subset A of non-null measure and for every probability-preserving transformation (dynamics) T, almost every point of A will return again to A, after a sufficiently long but finite number of iterations of T – called *Poincaré recurrence time*. Probabilities are special types of measures, so we can formally state the result using the framework of probability theory:

> **Poincaré recurrence theorem**. Let $T : X \to X$ be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$ and let A be a set in $\mathcal{B}$ with measure $\mu(A) > 0$.[10] Then with μ-probability equal to one, the orbit $\{T^n(x)\}_{n=0}^{\infty}$ of an element $x \in A$ returns to A infinitely often.

This is a "stability" or predictability result, as it guaranties that almost any orbit will eventually iterate in a similar way (and this for ever, again and again). In other words, every orbit will come close to a previous value, with probability one (almost certainty), even in the presence of chaos. This mathematical result seems to confirm the data science approach: if the system is believed to be deterministic, one can avoid the mathematical analysis and just "follow the (implicit) rule or regularity": by recurrence, it will dictate the future. As a consequence, in large databases, whose data come from an underlying (though unknown) deterministic dynamics, one expects that a regularity will show up again and again, hence allowing prediction and action.

A more subtle detail comes from the following classical result proved in [26]:

> **Kac's lemma.** The average recurrence time to a subset A in Poincaré recurrence theorem is the inverse of the probability of A.

This means that the "smaller" A is, the lower are its probabilities to be hit by the orbit or the longer it takes to get back to it.

Consider now a "relational orbit" in a database, that is, a series of values $v_1, v_2, \ldots, v_p$ relating p observables. The analyst does not know or does not want to know the possible law that determines the underlying dynamics, yet she cannot exclude the possibility that there is a law (to be given, for example, as a system of equations or an evolution function). In this case, the probabilities to find again values close to the v_i's in a small set A containing v_i's

[9] A branch of mathematics which studies dynamical systems with an invariant measure and related problems.
[10] The measure of A, $\mu(A)$, is the probability of A.

are very small and, thus, the recurrence time is very long.[11] Moreover, recent developments of Kac's lemma in [8] show that in any deterministic system as above, these probabilities *decrease exponentially* also with the size (dimension) of the phase space (observables and parameters) and the recurrence time *increases exponentially* with that size. Similarly, if the (underlying) equations or evolution function are non-linear, in particular chaotic and presenting attractors, then the probability referred above decreases exponentially also with the dimension of the attractors. [8][12] Actually, the paper [8] proves that the dimension of the phase space is at least as important as chaos for the unpredictability of deterministic dynamics and concludes:

> When the number of effective degrees of freedom underlying a dynamical process is even moderately large, predictions based solely on observational data soon become problematic, as in the case of weather forecasting.[13]

In real life problems, both dimensions – of the phase space and attractors – are large. Thus it is very unlikely or it takes a long recurrence time for an orbit (thus a regularity, a pattern) to iterate again. Furthermore, in databases one may decide to correlate only a few observables (for example, $n = 2$ in Figure 1), yet one cannot exclude that these observables depend on many more parameters — indeed this happens in cases where (causal) dependence is far from unique.

In conclusion, there is no way to reliably predict and act by "analogy" with the past, not even in deterministic systems, chaotic or not, specifically if they model real life phenomena. The data miner may believe that if a regularity shows up, it could have some "effective" role in prediction and action, in particular if the database is "big" as to reach the above recurrence limits.[14] No, this is of no use, as we will show in the following sections: big databases *have* regularities, but they are mostly "spurious", a notion that we will define using algorithmic randomness. Mathematically, we will show that spurious correlations largely prevail, that is, their measure tends to one with the size of the database. The proof of this result uses a different approach which does not need the hypothesis of determinism and is complementary to the one adopted in this section.

6 Ramsey theory

This theory, named after the British mathematician and philosopher Frank P. Ramsey, studies the conditions under which order must appear. Typical problems in Ramsey theory are (see more in [21, 20]):

> Are there binary (finite) strings with *no* patterns/correlations?

[11] Ehrenfest's example [56] is a simple illustration. Let an urn U_1 contain 100 numbered balls and U_2 be an empty urn. Each second, one ball is moved from one urn to the other, according to the measurement of events that produce numbers from 1 to 100. By Kac's lemma, the expected return time to (almost) all balls in U_1 is of (nearly) 2^{100} seconds, which is about 3×10^{12} times the age of the Universe. Boltzmann already had an intuition of this phenomenon, in the study of the recurrence time in ergodic dynamics, of gas particles for example, see [8].

[12] The dimension of an attractor is the number of effective degrees of freedom.

[13] Our footnote: see [31, 9].

[14] For example, with one dimension far exceeding the age of the Universe in seconds, yotta of yottabytes.

> How many elements of some structure must there be to guarantee that a *given regularity* holds?

Ramsey theory answers the first question in the *negative* by providing a precise answer to the second. Significantly, Graham and Spencer subtitled their *Scientific American* presentation of Ramsey theory [21] with the following sentence:

> Complete disorder is an impossibility. Every large set of numbers, points or objects necessarily contains a highly regular pattern.

Our analysis will use two Ramsey-type results. In both cases correlations – let us call them *Ramsey-type of correlations* – appear only because of the size of the data.

Let $s_1 \dots s_n$ be a binary string. A monochromatic arithmetic progression of length k is a substring $s_i s_{i+t} s_{i+2t} \dots s_{i+(k-1)t}$, for $1 \leq i$ and $i + (k-1)t \leq n$ with all characters equal for some $t > 0$. The string 01100110 contains no arithmetic progression of length 3 because the positions 1, 4, 5, 8 (for 0) and 2, 3, 6, 7 (for 1) do not contain an arithmetic progression of length 3. However, both strings 011001100, and 011001101 do: 1, 5, 9 for 0 and 3, 6, 9 for 1 and this change was obtained by adding just one bit. In fact, *all 512 binary strings of length 9 have a monochromatic arithmetic progression of length 3.*

The theorem below states that *all* sufficiently long strings of digits/colours, taken from a finite set, have "long enough" arithmetic progressions of the same digit, i.e. a sequence of equi-distributed positions of the same digit/colour (monochromatic) *of a pre-given length* [20]. The importance of the theorem lies in the fact that *all* strings display one of the simplest types of correlation: a sequence of equi-distributed positions of the same value.

> **Finite Van der Waerden theorem.** *For any positive integers k and c there is a positive integer γ such that every string, made out of c digits or colours, of length more than γ contains an arithmetic progression with k occurrences of the same digit or colour, i.e. a monochromatic arithmetic progression of length k.*

The Van der Waerden number $W(k, c)$ is the smallest γ such that every string, made out of c digits or colours, of length $\gamma + 1$ contains a monochromatic arithmetic progression of length k. How big is $W(k, c)$? For binary strings, one has $W(3, 2)$=9. In [18] it was proved that $W(k, 2) < 2^{2^{2^{2^{k+9}}}}$, but conjectured to be much smaller in [19]: $W(k, 2) < 2^{k^2}$.

In the examples above, for the sake of simplicity, only binary relations ($n = 2$) have been considered, such as smallpox inoculation and resistance to infection in a large set of individuals, or the number of marriages and the number of drowned people, over many years. In what follows we will use the finite Ramsey theorem [40] to study the more general case of n-ary relations. Ramsey theorem deals with arbitrary correlation functions over n-ary relations, as, in principle, in a database one can correlate sets of n elements, for any fixed $n \geq 2$.

First we illustrate the Ramsey theorem with the classic "party problem".[15] Suppose a party has six people. Consider any two of them. They might be meeting for the first time, in which case we call them *mutual strangers*, or they might have met before, in which case we call them

[15] Appeared in Putnam Mathematical Competition in 1953 and in the problem section of the *American Mathematical Monthly* in 1958 (Problem E 1321).

mutual acquaintances.[16] The problem asks whether *in every party of six people either: a) at least three of them are (pairwise) mutual strangers or b) at least three of them are (pairwise) mutual acquaintances?* The problem is equivalent to the statement that in the complete graph in Figure 2 – which has six vertices and every pair of vertices is joined by an edge – every colouring of edges with red and blue necessarily contains either a red triangle or a blue triangle.

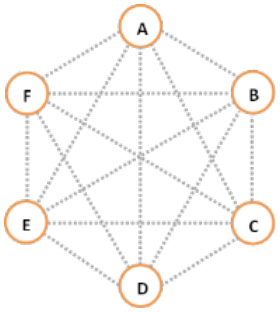

Figure 2: Party graph [3].

The answer is *affirmative*. To prove we just need to choose any vertex in Figure 2, say A, and note that there are five edges leaving A, each coloured red or blue. Then the *pigeonhole principle* says that at least three of them must have the same colour, hence a monochromatic triangle will appear. In Figure 3 the edges are AF, AE, AD, AC, AB, coloured in order with blue, red, red, red, red, blue: the triangle AEC is monochromatic as is coloured in red.

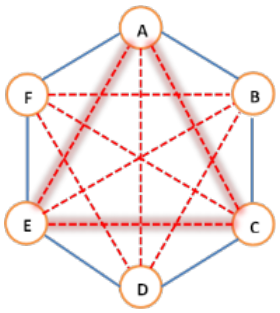

Figure 3: Red triangle in the party graph [3].

We do not know whether in the party there is a group of three people who are pairwise mutual strangers or a group of three people who are pairwise mutual acquaintances, but we know that a group of three people with the same type of pairwise acquaintanceship always exists. Clearly, this result is true for any party with more than six people.

We continue with a piece of notation. If A is a finite set then $[A]^n$ is the set of all subsets of A containing exactly n elements (n-ary relations or n-sets on A). In order to guide intuition, the number c may also be viewed as a number of "colours" partitioning a set. For example, if $A = \{x, y, z, u\}$ and $n = 2$ then $[A]^2 = \{\{x, y\}, \{x, z\}, \{x, u\}, \{y, z\}, \{y, u\}, \{z, u\}\}$;

[16] This is a Ramsey type problem: the aim is to find out how large the party needs to be to guarantee similar pairwise acquaintanceship in (at least) one group of three people.

an example of partition of $[A]^2$ with $c = 3$ colours, red, blue and yellow, is $\{\{\{x,y\},\{x,z\},\{z,u\}\},\{\{y,z\},\{y,u\}\},\{\{x,u\}\}\}$ (if $\{x,y\},\{x,z\},\{z,u\}$ are coloured with red and $\{y,z\},\{y,u\}$ are coloured with blue, then $\{x,u\}$ is coloured with yellow). Any of the six permutations of the colours red, blue and yellow produces a valid colouring.

> **Finite Ramsey theorem.** *For all positive integers b, n, c there is a positive integer γ such that for every finite set A containing more than γ elements and for every partition $P\colon [A]^n \to \{1, 2, \ldots, c\}$ there exists a subset B of A containing b elements whose n-sets are monochromatic, i.e. $P(x)$ has the same value (colour) for every x in $[B]^n$.*

We note the general framework and the strength of both theorems: the first deals with arbitrary, long enough *well-orderings*, the second with n-ary relations on arbitrary, large enough *sets*. In particular, given *any* positive integers b, n, c, the finite Ramsey theorem is valid for *all* sets A containing more than γ elements, and for *every* partition P of the set of subsets of n elements of A into c colours. In all cases we do not know in advance what colour will be given to the elements of interest; however, due to the finiteness of both theorems, in every particular instance we can algorithmically find all the monochromatic elements and their colourings.

7 Correlations everywhere

We will first use the regularity of arithmetic progressions in the total order analysed by Van der Waerden theorem, then the colouring of arbitrary sets of arbitrary objects studied by the finite Ramsey theorem.

The Van der Waerden theorem shows that in *any coding* of an arbitrary database of a large enough size into a string of digits, there will be *correlations of a pre-determined arbitrary length*: the same digit sitting on a long sequence of equi-distributed positions. These equi-distributed positions of the same digit are not due to the specific information coded in the data nor to a particular encoding: *they are determined only by the large size of the data.* Yet, the regular occurrence, for a pre-chosen number of times, of an event may be considered by the data miner as a "law" or as sufficient information to replace a law.

For correlations, assume that the database associates one digit or colour, out of c possible ones, according to a fixed correlation function (for example, 7 or red is produced for fishermen the month when marriages and drownings are correlated, 3 or green for cheese producers as it correlates quarterly cheese consumption with some other statistical data). Then the data miner can consider the correlation relevant if it occurs at a regular pace (an arithmetic progression) for at least k times. Of course, the string of length γ, where the arithmetic progression appears, may be given in time or space or in any countable order.

Thus, by Van der Waerden theorem, for any k and c, the required length k for one of these occurrences or correlations will be surely obtained in *any* string of length greater than γ, whatever (possibly spurious) criteria are chosen for correlating events. Moreover, and this is very important, *as the string of length greater than γ is arbitrary, its values, out of the c possible colours, may be chosen in any "random" way: by throwing a dice or measuring a quantum observable.* Yet, at least one correlation, as a monochromatic arithmetic progression of the desired length, will be found in the string. We will discuss later the chances of stepping on an algorithmic random string.

Next we use the finite Ramsey theorem to look in more general terms into this phenomenon. Instead of dealing with a well-ordering of a pre-given coding by numbers, we search for large enough regularities in sets of n elements taken from arbitrary, but large enough, sets. So, it is no more the regularity of appearance in a well-ordering of an arbitrary string that is searched for, but the structuring, according to an arbitrary partition, of n-ary relations of elements from an arbitrary set.

Let D be a relational database. In full generality, we may consider that *a correlation of variables in D is a set B of size b whose sets of n elements form the correlation (the set of n-ary relations on B, or n values that are considered to be correlated, or monochromatic using the language of colours).* In other words, when a correlation function – defined according to some given criteria (proximity or, conversely, apartness of some observable values or whatever) – selects a set of n-sets, whose elements form a set of cardinality b, then they become correlated. Thus, the process of selection may be viewed as a colouring of the chosen set of b elements with the same colour – out of c possible ones. We insist that *the criterion of selection – the correlation function – has no relevance here, it is arbitrary: it only matters that, for some reason which may be spurious, all n-sets of a set with b elements have the same colour, that is, turn out to be correlated.* Then Ramsey theorem shows that, given any correlation function and any b, n and c, there always exists a large enough number γ such that *any* set A of size greater than γ contains a set B of size b whose subsets of n elements are all correlated – that is, monochromatic. In other words: Do we want a size b set of values that are correlated by sets of n elements out of c possibilities, for whatever b, n and c *you* choose? Ramsey theorem gives us a large enough number γ such that in *any* set with more elements than γ if we choose in *any* way a partition of n-sets into c classes, we are guaranteed to find a correlation of size b and arity n. We do not know *a priori* what will be the colour of the monochromatic set, in the same way as the data miner does not know in advance which correlation will pop out from the data. However, in every particular instance we can algorithmically find all the monochromatic elements and their colourings.

The analysis above, as well as the one in the following section, are independent of any possible (or not) law-like origin (determination) of the processes modelled in the database, as it is only based on cardinality (number of elements in the database). This analysis complements the one in Section 5. Moreover, the arguments also reinforce each other, as the search for regularities require, by the previous results on the low probabilities of recurrences in deterministic systems, very large databases.

8 The deluge of spurious correlations

In the previous section we have showed how Ramsey-type of correlations appear in *all* large enough databases. How "large" is the set of spurious correlations?

First and foremost, the notion of "spurious correlation" was not (yet) defined: how can one answer then the above questions?

According to Oxford Dictionary [39], the adjective *spurious* means

> Not being what it purports to be; false or fake. False, although seeming to be genuine. Based on false ideas or ways of thinking.

This (dictionary) definition is semantic, hence it depends on an assumed theory: one correlation can be spurious according to one theory, but meaningful with respect to another one. To

answer the question posed in the beginning of section in the strongest form, we need a definition of "spurious" which is independent of *any* theory (using formal deductions, mathematical calculations, computer algorithms, etc.).

In order to satisfy the above broad constraint we define a spurious correlation in a very restrictive way:

> a correlation is *spurious* if it appears in a "randomly" generated database.

A spurious correlation in the above sense is also "spurious" according to any possible definition because, by construction, its values are chosen at "random", as all data in the database. As a consequence, such a correlation cannot provide reliable information on future developments of any type of behaviour. Of course, there are other reasons making a correlation spurious, even within a "non-random" database. However, to answer the question posed in this section it will be enough to show that most correlations are "spurious" even according to our restricted definition, when dealing with large enough sets of numbers (big data). Clearly, this is *a fortiori* true if one broadens the definition of spurious correlation.

As any database can be seen as a string in Van der Waerden theorem or a set in Ramsey theorem, our analysis will be based on algorithmic information theory, a mathematical theory studying (algorithmic) randomness for individual objects, in particular, strings and (finite) sets of numbers [6, 12]. This theory is based on classical computability, that is the general theory of algorithms, the foundation of todays computing [10]. As hinted above and discussed at length in [7], randomness is defined as "unpredictability relative to an intended theory". Since theories predict by computing from their formal frames (axioms, equations, evolution functions, ...), algorithmic randomness provides a general notion of randomness which works for any theory.

Using this approach we will answer the question: How "large" is the set of spurious correlations, in the above sense? Surprisingly, *spurious correlations form the majority, in fact a quantifiable majority of correlations.*

As the notion of algorithmically randomness is given in terms of "incompressibility", we start with a familiar example of "compressibility". We use *compressing/decompressing* programs such as zip, gzip, compress, etc., to obtain a compressed version of a given (arbitrary) file. These programs have two tasks: a) to try to obtain – via compression – a shorter variant of the original file, and b) to be lossless, that is, to guarantee that no information is lost during both processes of compression and decompression. How do they work? The answer is important for our discussion: a compression program searches for *regularities* in the input file and uses them to compute a shorter description of the file and the decompression program reconstructs the original file from its description.

The algorithmic complexity evaluates the size of the compressed file. For this it is enough to work only with decompression programs for finite strings. Here is an axiomatic framework to introduce this type of complexity. A function K from binary strings to binary strings satisfying the following three conditions is called an *algorithmic complexity*.

1. Conservation of information: For every computable function f there exists a constant c – depending only on f – such that for every string x we have $K(f(x)) \leq K(x) + c$.

2. Weak computability: There is a program enumerating all pairs strings and positive integers (x, n) such that $K(x) \leq n$.[17]

[17] The function K is incomputable.

3. Localisation: There exist two constants c and C such that the number of strings x with $K(x) < n$ is greater than $c \cdot 2^n$ and smaller than $C \cdot 2^n$.

Fix U a universal Turing machine and denote by $K_U(x) = \min\{y \mid U(y) = x\}$. Then K_U satisfies the above three axioms[18] and, conversely, if K satisfies the above three axioms, then there exists a constant M – depending only on U and K – such that for all x, $|K(x) - K_U(x)| \leq M$.

Let $K = K_U$. A string x is m-compressible if $K(x) \leq |x| - m$. This means that there is an input y of length $y| \leq |x| - m$ such that $U(y) = x$. In other words, U was able to find patterns in x to compress it to a shorter description y. It is not difficult to see that for all non-negative integers $m \leq n$, the number of strings x of length n having $K(x) < n - m$ is less or equal to $2^{n-m} - 1$. Taking $m = 0$ we get the following interesting fact: *for every non-negative integer n there exists a string x of length n such that $K(x) \geq n$.*[19] Such a string is *incompressible* by U: there is no y shorter than x such that $U(y) = x$. One can prove – see [6, 12] – that the incompressible strings have *most* properties associated with randomness, hence they are called *algorithmic random.*[20] The compressibility power of a universal Turing machine U is much higher than that of any practical algorithm - gzip, compress, etc. –, but by Ramsey theory, still limited: *there are correlations U cannot detect.* This is the surprising meaning of Ramsey type theorems, such as those used in this paper: no matter how one *algorithmically* defines randomness for strings or finite sets of numbers, they always contain regularities,, since any long/large enough string/set contains regularities.

Algorithmically random finite sets can be defined naturally [54]. These sets have intriguing regularities. Here is an illustration for Boolean matrices which are relevant for databases. With the exception of a finite set, the rank of *every* algorithmically random Boolean matrix of size $n \times n$ – thought over the field $\mathbf{F}_2 = \{0, 1\}$ – is greater than $n/2$. Otherwise we could select $n/2$ columns of the matrix such that all other columns are linear combinations of the selected ones. Further, based on this observation we could compress the original matrix to a string of $3/4 \cdot n^2 + \mathrm{O}(n \log n)$ bits instead of n^2 required by algorithmic randomness (see [49] for a proof and [27, 33] for other relevant results).

Fix a real number α in the open interval $(0, 1)$. A string x of length n can be compressed by αn bits if its complexity $K(x) \leq n - \alpha n$. The number of strings x of length n which are compressible by αn bits is smaller than $2^{(1-\alpha)n} - 1$, hence the probability that a string x of length n has $K(x) < n - \alpha n$ is smaller than $2^{-\alpha n}$, a quantity which converges exponentially to zero when $n \to \infty$. In other words, for large n, very few strings of length n are compressible or, dually, *most strings are algorithmically random.*

Here is a simple illustration of this phenomenon. According to [24] the best average rate of algorithmic compressibility is about 86.4% (i.e. a string x is compressed into a file $\mathrm{Zip}(x)$[21] which has about 86.4% of the length of x). The probability that a binary string x of (relatively short) length 2048 is reduced by 13.6% is smaller than $2^{-\frac{136}{1000} \cdot 2048} < 10^{-82}$, a very small number.[22]

[18] Traditionally, K_U is called Kolmogorov complexity associated to U.

[19] The number of strings x of length n having $K(x) \geq n - m$ is greater or equal to $2^n - 2^{n-m} + 1$.

[20] In view of results discussed in Section 6, they cannot have *all* properties associated with randomness.

[21] For every x, $\mathrm{Zip}(x)$ is an incompressible string for Zip, but for some x, $\mathrm{Zip}(x)$ is compressible by U.

[22] 10^{82} is approximately the number of hydrogen atoms in the observable Universe.

9 Conclusions

The analysis presented in this paper suggests that the slogan or "philosophy" declaring that *correlation supersedes causation and theorising* is mathematically wrong. Scientists and philosophers have always worried about fallacies one commits by looking at correlations only: *Cum hoc ergo propter hoc*[23] and the related *Post hoc ergo propter hoc*[24] fallacies are illustrative. For example, economists are wary of uncritically using correlations: that's why they perform further statistical tests, e.g. the Granger causality test, knowing well that even this is not enough to yield knowledge of genuine causality.

Our work confirms the intuition that the bigger the database which one mines for correlations, the higher is the chance to find recurrent regularities and the higher is the risk of committing such fallacies. We first showed that if the "world" modelled in a big database is somehow law-like in the multi-dimensional universe of events it describes, then the probability that a series of related observable values (regularity) iterates again is non zero, but extremely low: recurrence may occur, but only for immense values of the intended parameters and, thus, an immense database.

But, then, one steps into the results in the subsequent sections: given k and c (Van der Waerden theorem) or (b, n, c) (Ramsey theorem), there is a (large) γ, such that any data set of size more than γ contains a regularity with the characteristics given by the given parameters *independently of any law-like assumption on the underlying phenomena described by the elements of the database*. Moreover, as proved in Section 8, the larger is γ, the more probable becomes that the sets larger than γ are "randomly" generated. Even using a restrictive definition of spurious correlation – one that appears in an algorithmically random database – we showed that the overwhelming majority of correlations are spurious. In other words, there will be regularities, but, by construction, most of the time (almost always, in the mathematical sense), these regularities cannot be used to reliably predict and act.

[Natural] "[S]ciences are essentially an exercise in observing correlations, then proposing explanations for them", [17]. They operate with three types of correlations: two types of *local* correlations, that can be accounted for in terms of either an influence of one event on another ("if it is rainy, then the ground is wet"), or common local causes ("gum disease, oral cancer, loss of taste, mouth sores and bad breath are likely to affect youths with smoking habits") and *non-local* correlations (like quantum correlations of entangled qubits). Non-local correlations have much more complex physico-mathematical explanations than local correlations. But as we showed, there are also the "Ramsey-type of correlations", which abound, and cannot be accounted for anything else except size. Even worse, the last type of correlations cannot be algorithmically distinguished from the others.

The theory of algorithms, which gave us computers and algorithmic analyses (of databases, in particular), also provides the tools for understanding the limits of a pure algorithmic analysis of big data. Our limiting or "negative" results, as it often happens [29], do not "destroy" data science, but open the way for more reflections. Big data correlations do not make causation obsolete nor do they cause Democritus's type of science to die, but they pose the challenge to integrate the new algorithmic tools with the classical ones in an extended scientific method.

The fundamental Greek practice of scientific observation, thinking and debating on different theoretical interpretations of phenomena was enriched by the experimental method (since

[23] Latin: with this, therefore because of this.
[24] Latin: after this, therefore because of this.

Galileo) and mathematics (since Descartes and Newton). Suggestions to narrow down the scientific methods to just the collection of "empirical evidences" or to the use of mathematics – "in a discipline there is as much science as there is mathematics"[25] – didn't get much support. The new "philosophy" proposed by big data is similar. Big data analytics cannot replace science and, symmetrically, no theory can be so good to supplant the need for data and testing. Implicit or, better, explicit and revisable theorising should accompany meaningful measurements of "evidences" and mathematical modelling, as well as reliable analyses of databases.[26]

Despite the looming threat of spurious correlations, there are many examples of successful use of data analytics in robust scientific studies, see [42, 52, 47, 43]. Thus, we interpret Dante's verse in the epigraph of this paper as "we cannot rely on computational brute-force ("come bruti"), but we must pursue scientific commitment and knowledge ("virtute e canoscenza").

Acknowledgment

The authors have been supported in part by Marie Curie FP7-PEOPLE-2010-IRSES Grant. Longo's work is also part of the project "Lois des dieux, des hommes et de la nature", Institut d'Etudes Avancées, Nantes, France. We thank A. Vulpiani for suggesting the use of Kac's lemma, G. Tee for providing historical data and A. Abbott, F. Kroon, H. Maurer, J. P. Lewis, C. Mamali, R. Nicolescu, G. Smith, G. Tee, A. Vulpiani and the anonymous referees for useful comments and suggestions.

References

[1] Correlation and prediction. http://www.intropsych.com/ch01_psychology_and_science/correlation_and_prediction.html, 1992.

[2] Spurious correlations. http://www.tylervigen.com/spurious-correlations, Nov 2015.

[3] A. Ahn. The party problem (Accessed: 12 december 2015). http://mathforum.org/mathimages/index.php/The_Party_Problem_(Ramsey's_Theorem).

[4] G. E. Andrews. Drowning in the data deluge. *Notices Amer. Math. Soc.*, 59(7):933–941, August 2012.

[5] A. S. Calude. Does big data equal big problems? http://blogs.crikey.com.au/fullysic/2015/11/13/does-big-data-equal-big-problems, Nov 2015.

[6] C. Calude. *Information and Randomness—An Algorithmic Perspective*. Springer, Berlin, second edition, 2002.

[7] C. S. Calude and G. Longo. Classical, quantum and biological randomness as relative. *Natural Computing*, 2015. http://dx.doi.org/10.1007/s11047-015-9533-2.

[25] The danger of purely speculative theories in today's physics is discussed in [13].

[26] The big data can be used for scientific testing of hypotheses as well as for testing scientific theories and results.

[8] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani. Predicting the future from the past: An old problem from a modern perspective. *American Journal of Physics*, 80(11):1001–1008, 2012.

[9] S. Chibbaro, L. Rondoni, and A. Vulpiani. *Reductionism, Emergence and Levels of Reality*. Springer Berlin / Heidelberg, 2014.

[10] S. B. Cooper. *Computability Theory*. Chapman Hall/CRC, London, UK, 2004.

[11] R. L. Devaney. *An Introduction to Chaotic Dynamical Systems*. Addison-Wesley, Redwood City, CA, second edition, 2003.

[12] R. Downey and D. Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, Berlin, 2010.

[13] G. Ellis and J. Silk. Scientific method: Defend the integrity of physics. *Nature*, 516:321–323, 2014.

[14] R. Ferber. Are correlations any guide to predictive value? *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 5(2):113–121, 1956.

[15] L. Floridi. Big data and their epistemological challenge. *Philosophy and Technology*, 25(4):435–437, 2012.

[16] M. Frické. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651–661, 2015.

[17] N. Gisin. *Quantum Chance: Nonlocality, Teleportation and Other Quantum Marvels*. Springer, London, 2014.

[18] T. Gowers. A new proof of Szemerédi's theorem. *Geometric and Functional Analysis*, 11(3):465–588, 2001.

[19] R. Graham. Some of my favorite problems in Ramsey Theory. *INTEGERS, The Electronic Journal of Combinatorial Number Theory*, 7(2):#A2, 2007.

[20] R. Graham, B. L. Rothschild, and J. H. Spencer. *Ramsey Theory*. John Wiley and Sons, New York, 2nd edition, 1990.

[21] R. Graham and J. H. Spencer. Ramsey theory. *Scientific American*, 262:112–117, Sept. 1990.

[22] A. Grjebine. *La dette publique et comment s'en débarrasser*. Press Universitaire de France, Paris, 2015.

[23] L. Grossman. What's this all about? The massive volume of data that humanity generates is a new kind of problem. The solution is very old: art. *Time Magazine*, 6 July 2015 (double issue).

[24] C. Hoffman. Benchmarked: What's the best file compression format? http://www.howtogeek.com/200698/benchmarked-whats-the-best-file-compression-format/, May 2015.

[25] IBM. What is big data? http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, May 2011.

[26] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.*, 53:1002–1010, 1947.

[27] B. Khoussainov. Algorithmically random universal algebras. In M. Burgin and C. S. Calude, editors, *Information and Complexity*. World Scientific Series in Information Studies, Singapore, 2016 (to appear).

[28] R. Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 2014.

[29] G. Longo. On the relevance of negative results. Influxus, http://www.influxus.eu/article474.html, 2008.

[30] G. Longo and M. Montévil. *Perspectives on Organisms: Biological Time, Symmetries and Singularities*. Springer, Berlin and Heidelberg, 2014.

[31] P. Lynch. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(3431):3431–3444, 2008.

[32] J. Lyons. George Osborne's favourite "godfathers of austerity" economists admit to making error in research. http://www.mirror.co.uk/news/uk-news/george-osbornes-favourite-economists-reinhart-1838219, April 2013.

[33] Y. I. Manin. Cognition and complexity. In M. Burgin and C. S. Calude, editors, *Information and Complexity*. World Scientific Series in Information Studies, Singapore, 2016 (to appear).

[34] C. Montelle. *Chasing Shadows: Mathematics, Astronomy, and the Early History of Eclipse Reckoning*. Johns Hopkins University Press, Baltimore, 2011.

[35] J. Needham. *Science and Civilisation in China: Medicine*, volume 6. Cambridge University Press, 2008.

[36] P. Norvig. All we want are the facts, ma'am. http://norvig.com/fact-check.html, 2008.

[37] NSF. Computational and data-enabled science and engineering. http://www.nsf.gov/mps/cds-e/, 2010.

[38] C. O'Grady. Louder vowels won't get you laid, and other tales of spurious correlation. http://arstechnica.co.uk/science/2015/06/louder-vowels-wont-get-you-laid-and-other-tales-of-spurious-correlation, June 2015.

[39] Oxford Dictionaries. Spurious (Accessed: 30 november 2015). http://www.oxforddictionaries.com/definition/learner/spurious, Nov 2015.

[40] J. Paris and L. Harrington. A mathematical incompleteness in Peano Arithmetic. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 1133–1142, Amsterdam, 1977. North Holland.

[41] J. Poppelars. OR at Work. http://john-poppelaars.blogspot.fr/2015/04/do-numbers-really-speak-for-themselves.html, April 2015.

[42] A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, UK, 2011.

[43] D. A. Reed and J. Dongarra. Exascale computing and big data. *Commun. ACM*, 58(7):56–68, June 2015.

[44] C. Reinhart and K. Rogoff. Growth in a time of debt. *American Economic Review*, 2:573–578, 2010.

[45] S. Roberts and J. Winters. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE*, 8(8):e70902, 08 2013.

[46] E. Schmidt. Every 2 days we create as much information as we did up to 2003. http://techcrunch.com/2010/08/04/schmidt-data, Aug 2010.

[47] R. Schutt and C. O'Neil. *Doing Data Science*. O'Reilly Media, 2014.

[48] J. Sessions. The case for growth: Sessions lists benefits of discretionary cuts. http://www.sessions.senate.gov/public/index.cfm/news-releases?ID=E36C43B4-B428-41A4-A562-475FC16D3793, March 2011.

[49] A. Shen. Around Kolmogorov complexity: basic notions and results. http://dblp.uni-trier.de/rec/bib/journals/corr/Shen15, 2015.

[50] G. Smith. *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie With Statistics*. Overlook/Duckworth, New York – London, 2014.

[51] J. Smith. From Reinhart & Rogoff's own data: UK GDP increased fastest when debt-to-GDP ratio was highest – and the debt ratio came down! http://www.primeeconomics.org/articles/1785, April 2013.

[52] J. M. Stanton. *Introduction to Data Science*. Syracuse University, Syracuse, 2012.

[53] M. A. Thomas Herndon and R. Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38:257–279, 2014.

[54] N. K. Vereshchagin. Kolmogorov complexity of enumerating finite sets. *Information Processing Letters*, 103(1):34 – 39, 2007.

[55] T. Vigen. *Spurious Correlations*. Hachette Books, New York, 2015.

[56] C. Walkden. Magic post-graduate lectures: Magic010 ergodic theory lecture 5. http://www.maths.manchester.ac.uk/~cwalkden/magic/.