


Statistical Data Analysis in the Era of Big Data

Thomas Lengauer*

DOI: 10.1002/cite.202000024[‡]

 This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Big data is on everyone's lips and often raises emotions. On the one hand, the notion is a basis for much technological optimism, mostly directed towards new business models, or simplifications and optimizations in professional and private life. On the other hand, it is a basis for dystopic perspectives, which are targeted, e.g., at profiling of the individual and their privacy space, overarching optimization in daily life and intransparency of decision making. In this article, after a short historical prolog, it is discussed what distinguishes big data from traditional data analysis. The underlying mathematical methods are introduced and scientific successes are reported. Additionally, the risks and limits – especially regarding the derivation of causal relationships – of data analysis are discussed.

Keywords: Causality, Classification, Correlation, Modeling, Regression

Received: February 26, 2020; *accepted:* April 23, 2020

1 History of Data Analysis

We have been learning from data for a long time. The general process is always the same: The first step is to collect data as systematically as possible. This can happen just by observing the system to be investigated, or also in a controlled experiment, which systematically creates boundary conditions for the observations and carries out targeted interventions. In a second step, we then search for meaningful patterns in the data collected. This process is called data analysis. The uncovered patterns are then interpreted; this is how regularities are derived that are often called rules or laws. Often the knowledge generation process ends at this point. In the best of all cases, however, once the regularities have been discovered, a third step searches for causal relationships that explain them by reducing them to more fundamental principles.

References to systematic collection and analysis already date back to the Babylonians [1]. At this point let us consider an example that is often referred to as the first systematic data analysis of modern times: gaining insight into the laws governing the orbits of the planets in the solar system and their scientific basis. The process began with a meticulous measurement and collection of the coordinates of planetary trajectories over several decades by the Danish astronomer Tycho Brahe (1546–1601). Johannes Kepler (1571–1630), who assisted Tycho Brahe during the last year of his life, has since then continued to work on his data collection, and in 1627 he published it as the *Tabulae Rudolphinae* [2, 3]. He analyzed the data, resulting in Kepler's three laws, which have been posed in the early 17th century (see [4]) and which formulate mathematically the geometry of a planet's orbit and the relationships between its orbital velocity and the distance from its central star. It is important to stress that Kepler's three laws do not provide an explanation for

the formulas found. Therefore, in the context of this article it is preferred to speak of “rules” rather than laws. The planet moves in an elliptical orbit around the central star, which is located at one of the two foci of the ellipse. The orbital velocity is faster when the planet is closer to the central star. Kepler's laws are precise mathematical formulas for the shape of the ellipse and the orbital velocity of the planet. Kepler's reasoning was based on Renaissance ideas about forces and soul but presented an attempt to physicalize the planetary theory (see [5], p. 216). As we see in the following the discovery of such patterns in data is extremely useful, per se, and in many cases can form the basis for subsequent decisions and strategies. In the case of planetary orbits, the process of gaining knowledge did not end there – fortunately and characteristically of how science proceeds. Rather, on the basis of Kepler's laws and other observations, Isaac Newton (1642–1727) derived his law of gravitation and published it in 1687 in his work *Philosophiae Naturalis Principia Mathematica* (see [6], p. 8, 381, 383, 400–510), or *Principia* for short. The law of gravitation provides a fundamentally new explanation of Kepler's laws. More specifically, Kepler's laws can be derived mathematically from Newton's law. From then on Newton's law of gravitation provided the axiomatic basis for celestial mechanics, i.e., it was assumed to be a given from Newton's perspective and

Prof. Dr. Thomas Lengauer
lengauer@mpi-inf.mpg.de
Max Planck Institute for Informatics, Saarland Informatics Campus, Campus E1 4, 66123 Saarbrücken, Germany.

[‡]English version of: T. Lengauer, Statistische Datenanalyse in der Zeit von Big Data, *Nova Acta Leopoldina* 2019, NF 424, 187–206. With kind permission of the Deutsche Akademie der Naturforscher Leopoldina – Nationale Akademie der Wissenschaften.

not in need of further explanation. This situation has only changed with the development of the more comprehensive general theory of relativity in 1915 put forth by Albert Einstein (1879–1955) [7, 8], which reproduces Newton's law of gravitation (in the limit for slow speed compared with the speed of light and weak gravity).

A characteristic of the data analysis just discussed is that the examined system, here the solar system, is only observed. There is no intervention into the system – this is also not possible with the solar system. This kind of data analysis will be discussed first. If one intervenes in the system, e.g., in a controlled experiment, data analysis takes on a different character, which will be discussed later.

2 What is Different with Big Data?

The process described in Sect. 1 is a prime example of meticulously performed data analysis, but as a table with a total of a little over 250 pages certainly not for big data. More significant than the scope of the data collection is the fact that the relevant patterns, mathematically formulated in Kepler's laws, were extracted manually by Johannes Kepler by reviewing the data collection. The data collections that form the basis of today's big data analyses are much more voluminous. Furthermore, the relevant patterns are often very complex. For these two reasons, finding relevant patterns in such data sets without the aid of sophisticated algorithms and the use of computers is no longer possible.

In partial deviation from the criteria commonly used for big data¹⁾, I would like to state the following preconditions for big data collections and analyses.

- 1) The data volume must be very large. I am not asking for the size of petabytes or more, which is put forth often for the term big data. For the purpose of this paper, it is sufficient to assume a quantity that requires data analysis using complex computer algorithms.
- 2) Data analysis, not data generation, must be the bottleneck in knowledge generation. It is a characteristic of big data analysis that we have access to large amounts of data with comparatively little effort. This is a reversal of the traditional situation where, as a rule, data generation is much more complex and expensive than data analysis. This was definitely the case with Tycho Brahe. In contrast, today we generate data with high-throughput experiments in science and via the internet in everyday life to an extent that is no less than avalanche-like. Thus, data analysis becomes the bottleneck in the process of knowledge generation.

1) Big data is often characterized by the five V: Volume (large amount of data, corresponding to our criterion (1)), Velocity (speed of data generation, according to our criterion (2)), Variety (heterogeneity of data, i.e., different origins; this criterion is in the foreground in our discourse), Veracity (reliability of data; see Sect. 5), and Value (an economic criterion, which we will not consider here).

- 3) Data analysis is no longer possible by hand. This does not apply to the above example of the derivation of celestial mechanics. The difficulty of manual data analysis is usually not only due to the high volume of data, but also to another characteristic of today's data collections, namely the high dimensionality of the data. What this means is explained in the next section.
- 4) Powerful statistical methods, often referred to as data mining or machine learning, enable the computer to detect even complex patterns in the data. This makes the application of statistical methods the essential step in generating knowledge with big data.

Today we encounter big data in all areas of life and science. In daily life, since the emergence of the internet and due to the increasing interconnectedness of technologies, massive data collection is happening in a wide variety of areas. Data are collected whenever we browse the internet, whenever we watch television, when we drive, when we use our mobile phones, when we do banking or shopping. In public life our traces are captured by webcams. Domestic appliances increasingly exhibit networked intelligence, and networking and data collection is also gaining ground in energy supply.

Similarly, big data is increasingly becoming an essential part of practically all scientific disciplines. Elementary particle physics is a prime example. The Higgs particle was found only by collecting and analyzing vast amounts of data [9, 10]. In astronomy, detailed three-dimensional (sometimes four-dimensional) models of the entire universe and the stars and galaxies it contains are being developed [11]. The earth sciences and environmental sciences [12–14] collect a wide range of data on various aspects of the state of our planet and project this data into the future and into the past. In chemistry, comprehensive data sets are collected on both chemical compounds and their properties, and comprehensive data sets resulting from quantum mechanical calculations are made available worldwide [15–17]. The economic and social sciences are basing their research on extensive and diverse data collections [18, 19]. And access to complete genomic information was an essential prerequisite for the transformation of biology and medicine into a quantitative science highly driven by molecular data [20, 21].

A lot has been said and written about big data. Here, the focus is on a special scientifically relevant aspect of big data analyses, namely the difference between the discovery of regularities (associations), which describe patterns, and of laws (causal relations) that explain patterns.

3 Statistical Data Analysis

This section provides a brief introduction into the basics of statistical data analysis. A particular variant of data analysis, called supervised learning, will be presented. This form of data analysis aims at predicting unknown values from a set

of known values about an object or process. The encoding of the input usually results in a point in a Euclidean space which is high-dimensional, in general. For illustration an example data set²⁾ is used providing information on the fuel consumption of motor vehicles as well as their year of construction and their weight. All three values together form a data point that provides information on a vehicle. The latter two values are regarded as inputs and the first value as output. These roles are assigned to the values for the purposes of data analysis and are not anchored in the data set. This assignment is done to analyze the dependence of fuel consumption on the weight and year of manufacture of the vehicle. Since the data are from the United States, fuel consumption is measured in miles per gallon (mpg) and the weight in British pounds (lbs). Thus, we have two inputs (weight and year) – the data space on which the analysis is based is, thus, two-dimensional – and one output (fuel consumption). In real cases, the dimensionality of the data space can be much higher. This will be commented on later.

There are basically two forms of supervised learning (see Fig. 1):

- 1) Regression: Here a number (label) is assigned to each data point that functions as output. In the example data set, this number is an estimate of the fuel consumption in miles per gallon. The two input variables (weight, year) span a plane over which a vertical axis extends, marked as *fuel consumption*. The fuel consumption values estimated by a mathematical model are represented by surfaces in Fig. 1 (a, b). The actual (measured) values engulf this surface as a point cloud. Fig. 1a shows a linear model, represented by a planar surface. Fig. 1b shows a more complex nonlinear model, represented by a (slightly) curved plane. The vertical lines that extend from the data points onto the surfaces represent the errors that the model makes in estimating the fuel consumption of individual vehicles. In both cases the surfaces are positioned such that the sum of the square errors over all data points is minimized. The error of the linear model is greater than that of the nonlinear model, since the linear model must fulfil the additional boundary condition that the plane must not be curved.
- 2) Classification: Instead of estimating a continuous output value, here one of finitely many classes is assigned to each data point. The number of classes is usually small. In the example of Fig. 1 (c–e) we have two classes: *low fuel consumption* (mpg ≥ 20 , blue) and *high fuel consumption* (mpg < 20 , orange). Fig. 1c shows this situation in the same fashion as Fig. 1a and 1b does for regression. Figs. 1d and 1e show views onto the input plane from the top. The direction of view is given by the arrow in Fig. 1c. The goal of the data analysis is to

subdivide the input data space – here the plane – into areas that are assigned to the two classes. The classification of a new data point is then performed by assigning the class to the point that is represented by the color of the location of the point. Fig. 1d again shows a linear model that is characterized by a straight dividing line between the areas representing the two classes. This line is called the decision boundary. The model makes a number of errors (30 out of 392) on our data set, which are given by points whose color – that is, their actual class – is different from that of their background – i.e., their class estimate. The decision boundary is chosen such that the number of such false classifications is as small as possible. Fig. 1e shows a nonlinear model with a complex curved decision boundary. Similar to regression, its predictions are better on our data set than those of the linear model. In regression they are closer to the true values, on average, in classification this model makes much fewer errors (misclassifications) on the data set (2 out of 392).

In general, the art of adequate data analysis comprises 1.) the appropriate coding of the input, 2.) the correct selection of the error function (in our case the square error in the regression and the number of incorrect classifications (misclassification error) in classification), and 3.) the appropriate selection of the model, e.g., the correct choice between a linear (simple) and a nonlinear (complex) model. The deficit of simple models is that the accuracy of their estimates may be limited by constraints such as the linearity constraint, which may not be fully reflected in the data. If a model is too complex, however, there is a risk that, although it can reproduce the given data very accurately, it may not be able to generalize accurately to unseen data. Such a model is called overtrained. Fig. 1e is an example of an overtrained model of the example data set. The fact that this model is overtrained is easily recognized by the complex decision boundary, which is obviously adapted to the specific data set and does not reflect the real relationship between high or low fuel consumption and year of manufacture and weight. Therefore, the model makes only few errors on the given data. However, it is not to be expected that the model makes accurate estimates on future data. After all, the fuel consumption of a motor vehicle also depends on variables other than its weight and year of manufacture. However, we do not have such data at our disposal. Therefore, it is safe to assume that the nonlinear classification model in Fig. 1e makes significant errors on new data. However, a linear model for the relationship between weight and year of construction on the one hand and fuel consumption on the other hand appears to be too simple. As can be shown, the nonlinear regression model in Fig. 1b predicts fuel consumption on future data more accurately (with smaller error and higher predictive power) than the linear model in Fig. 1a. And compared to the nonlinear classification model, the nonlinear regression model is considerably smoother and offers a plausible functional dependence of fuel consumption on weight and year

2) The data set is available at <http://www-bcf.usc.edu/~gareth/ISL/data.html>

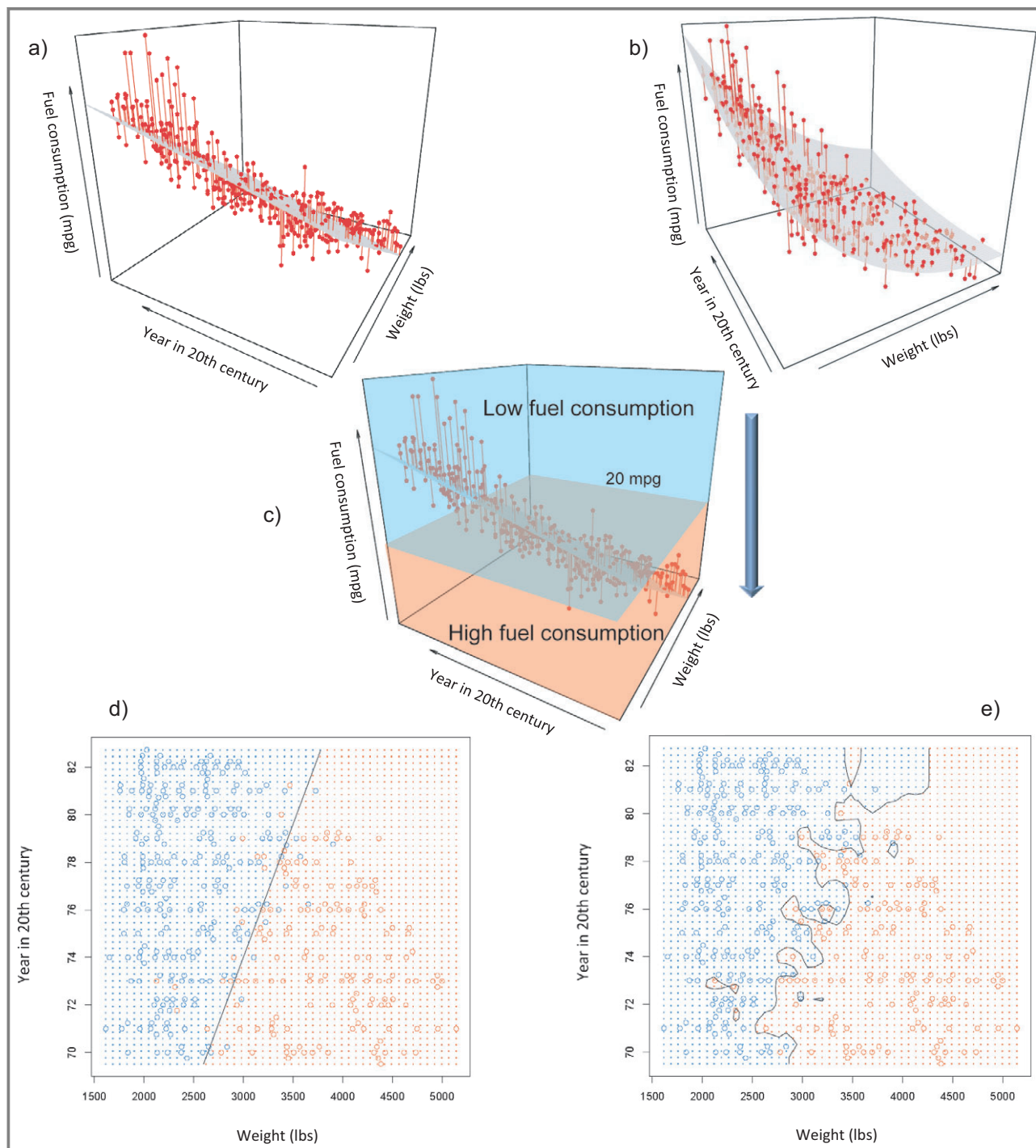


Figure 1. Statistical methods for data analysis. a) Regression, linear model, b) regression, nonlinear model. The data points are colored red. The gray surfaces represent the estimated fuel consumption values. The vertical red lines indicate the deviations of the actual from the estimated values. c) Class definition for binary classification into high and low fuel consumption. d) Classification, linear model, e) classification, nonlinear model. The data points are colored blue (low fuel consumption) and orange (high fuel consumption). The black lines show the decision boundaries obtained with a linear and a nonlinear model, respectively. The direction of view of the top onto the input plane in parts d) and e) is indicated by the vertical arrow in part c).

of manufacture. Thus, models can be too simple, but also too complex. The appropriate choice of model complexity is a central problem in data analysis.

Another major problem in data analysis is the frequently occurring high dimensionality of the data. The dimensionality of the data is the number of features that exist for each

data point. In our example we have two features: weight and year of manufacture; thus, the data set is two-dimensional. We also have 392 data points. The number of data points, thus, clearly exceeds the dimensionality of the data space. Consequently, the data space is also well filled with data points, as can be seen in Fig. 1d and 1e. Only in the upper right corner data points are missing. This is often quite different with today's data sets. For example, if we want to analyze the influence of genes on diseases, we often have to deal with comparatively few data points. Each data point represents one patient, and the cohort sizes usually range from a few dozen to hundreds. (Some very large studies, however, already include more than 100 000 patients [22–24].) In contrast, the human has about 20 000 genes, all of which can be measured using modern molecular biological methods. The dimensionality of the data set is, therefore, also in this range. This means that the dimensionality of the data space far exceeds the number of observations. This happens in many cases from all areas of data collection.

For a high-dimensional data set there are two problems, which are illustrated in Fig. 2. With increasing dimensionality, the number of data points required to fill the data space to a certain density increases exponentially. For data spaces with hundreds or thousands of dimensions, there are usually not nearly enough data available to adequately sample the complete data space. Furthermore, with increasing dimensionality, an ever greater fraction of data points are located near the margin of the space, i.e., in an unfavorable position for all procedures that infer the output of the examined data point from the outputs of points in its vicinity. High-dimensional data spaces consist practically only of margin. If the number of dimensions exceeds the number of data points, we usually have a serious problem, and a significant portion of today's research aims at being able to still make useful predictions in such cases.

After a description of the methods of data analysis a case study is presented, which shows that the analysis of large data sets can supply information that would not be available otherwise.

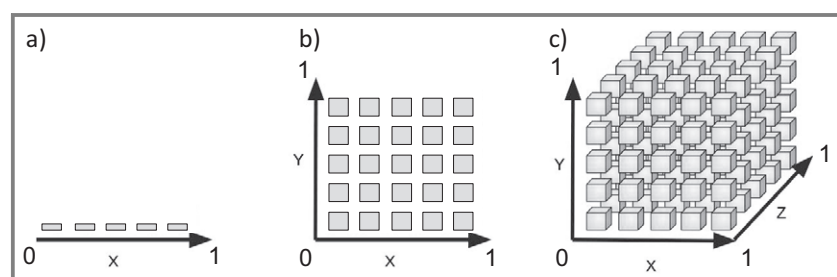


Figure 2. a) One-dimensional data space. Five evenly distributed data points fill the space to a certain density. Two of these points, i.e., 40 %, lie at the margin of the data space. b) Two-dimensional data space. To fill it with data points to the same density, you need 25 points. Of these points 16 points, or 64 %, lie at the margin. c) Three-dimensional data space. Now, already 125 data points are needed. Of these, 98 points (or 78.4 %) lie at the margin. This figure is a modified version of a figure at www.freecodecamp.org/news/the-curse-of-dimensionality-how-we-can-save-big-data-from-itself-d9fa0f872335.

4 Medicine: Big Data for the Benefit of the Patient

Medicine as a science is concerned with the differences between healthy and sick people. Diseases usually have a specific molecular basis. They manifest themselves in dysregulations of highly complex molecular interaction networks inside or between cells of our body. All of these interactions are subject to the natural laws of physics and chemistry, which are well known.

Thus, we know the mathematical foundations of the functions and malfunctions of the human body, in principle. However, a physicochemical analysis of the molecular processes based on the application of the known laws of nature is not possible, in general. This is because the systems involved (molecules, cells, organs, organ systems) are far too complex for exact calculation. Therefore, we have to proceed in a data-driven fashion. This in itself is nothing new. Medicine has always been data driven. The doctor has based the diagnosis and therapy on information about the patient's medical history, laboratory values and medical experience. In a sense, based on the information available, a model of the patient was derived, on which the diagnostic and therapeutic decisions are based. However, neither the model nor the process of its derivation are generally of a consistently systematic nature. In the age of big data, this procedure is submitted to a high degree of mathematization and systematization. In its course, the measured laboratory values, especially those of genomic character, and their mathematical interpretation with the help of computers is becoming increasingly important.

This is illustrated using the example of HIV therapy. The human immunodeficiency virus (HIV) evolves extremely rapidly in the body of the infected patient. Here the same type of process can be observed that is known from the development of antibiotic resistance in bacteria. Only in the case of HIV, it is not only observed in the infected population as a whole, but also in the individual patient and over very short periods of time – months, weeks or even days.

An HIV patient can simultaneously harbor a wide variety of similar but different HI viruses, and these viruses are constantly changing in order to escape the attacks by the patient's immune system and administered drug therapy. For this reason, today there are over two dozen different drugs against HIV, which are administered to patients in combinations of about three drugs, affording over a thousand therapy options.

Whether a virus is resistant to a certain drug is encoded in the viral genome, but in a way that is not readily apparent to humans and which is also difficult to measure in the laboratory. So, this is a typical data analysis problem [25]. The

input for the problem consists of an appropriate coding of the genome sequence of the virus that is found dominantly in the patient. (Recently it has also become possible to determine with high accuracy the variety of different viruses present in the patient [26, 27].) The corresponding data space has several thousand dimensions. The output is a list of the estimated resistances of the virus to the range of HIV drugs available (see Fig. 3). In this case, the database (the big-data aspect of the problem) comprises more than 230 000 therapy change episodes of HIV patients collected by an international consortium [28]³⁾. On these data, statistical learning was applied, as described in Sect. 3, to derive mathematical models that estimate the level of resistance of the virus to the available drugs based on the viral genome sequence. At www.geno2pheno.org such a decision-support system for HIV therapy is provided free of charge via the internet [29, 30]. The system is used in Germany and beyond to treat HIV patients.

Fig. 4 shows an analysis report returned by the system. This report is from 2003 for an HIV patient who has a virus with so many resistance mutations that no promising therapy could be found using conventional methods. The table has one row per drug. The first column (from the left) gives the drug name. The second column contains the model's estimate of the resistance factor to the corresponding drug, a numerical laboratory value that represents the strength of resistance between different drugs, the third column contains a normalized form of the resistance factor (z-score). If this value is greater than 4, there is a considerable degree of resistance of the virus to the drug, which renders the use of the drug problematic. For the patient in question, no effective medical treatment was available according to this rule. Thus, it is plausible that conventional methods that only provide a classification of a drug into *effective* and *not effective* were not able to suggest a therapy.

However, our data analysis provides numerical values, on whose basis a therapy can be suggested consisting of the drugs with the lowest z-scores, e.g., even if they are a little above 4. However, there is still room for improvement.

The data analysis was extended to include an interpretation of the prediction. This goes beyond the methods presented in Sect. 3. In the right part of the table the mutations in the viral genome are listed that confer resistance to the virus (resistance mutations, red) and those which increase the efficacy of the drug (resensitizing mutations, green). Each mutation is coded by a number (the position of the mutation in the corresponding protein sequence of the virus), followed by a letter (the amino acid, into which the viral protein is mutated at this position). This information puts the resistance value of

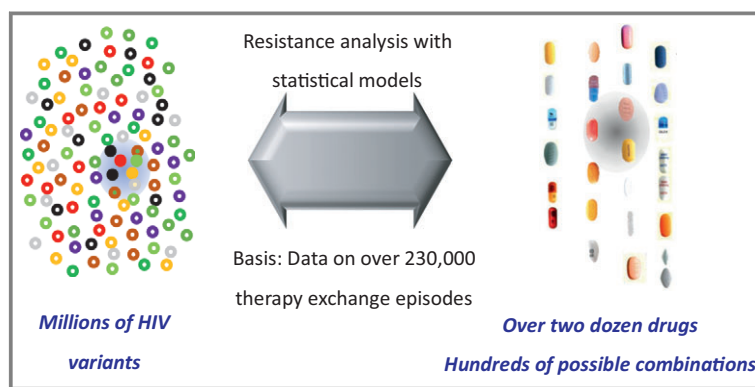


Figure 3. Using data mining in large databases on HIV drug resistance, from a set of over two dozen HIV drugs (right), drug combinations (right, background shaded gray to black) that are effective against the predominant population of HI viruses in the patient (left, blue background) are proposed.

| Drug | RF(*) | z-score | Scored Mutations(**) |
|------------|---------|---------|---------------------------------------|
| ZDV | 257.276 | 9.945 | 215Y 210W 41L |
| ddl | 4.057 | 6.087 | 184V 121H 178L 215Y 177E |
| d4T | 2.477 | 4.594 | 215Y 178L 118I 184V 210W 121H 41L |
| 3TC | 149.029 | 18.504 | 184V 41L 215Y |
| ABC | 7.501 | 11.968 | 184V 215Y 210W 41L |
| TDF | 4.613 | 6.684 | 215Y 41L 98G 184V 178L 118I 177E 135T |
| NVP | 149.235 | 5.243 | 103N 135T 210W 98G 211K |
| EFV | 60.034 | 7.450 | 103N 98G 135T 210W 214F 177E |
| SQV | 3.908 | 4.658 | 63P 46I 37N 71V 72T 60E 76V 57K |
| IDV | 52.999 | 11.171 | 46I 76V 63P 82A 61E 62V 60E 71V |
| NFV | 33.156 | 8.246 | 46I 63P 10F 3I 76V 60E 62V |
| APV | 59.452 | 13.085 | 76V 54M 10F 46I 63P |
| LPV | 100.108 | 14.521 | 46I 76V 10F 63P 82A |
| ATV | 19.252 | 7.838 | 82A 46L 54M 76V 93L 71V 62V |

Figure 4. Analysis report for an HIV patient.

the virus into context with characteristics of the viral genome. And this increases the expressiveness of the prediction significantly. On the basis of this report, the physicians treating the patient selected the two encircled drugs. The drug SQV was administered due to its low resistance level, with the hope that it will still be effective. LPV cannot be expected to be effective due to its high resistance level. But this high value can be partially attributed to the resistance mutation 76V (colored red for LPV) which, at the same time, is indicated to reduce the resistance level of the virus to SQV, since there it is colored green. The doctors hoped that the virus would be striving to maintain its resistance to LPV, thus fixing the 76V mutation, which hopefully maintains the effectiveness of SQV. It turned out that this therapy has been effective for years.

We can see from this example that the task of data analysis goes beyond the mere estimation of the output value. Rather, one also wants to weigh the input values according to how informative they are for the output value.

In the development of our HIV models, all problems described in Sect. 3 were addressed. Due to the high

3) see www.euresist.org

dimensionality of the data space there is a danger of over-training. We usually use linear models because they limit the danger of overtraining, on the one hand, and because they facilitate the kind of interpretation of the estimate just described, on the other hand.

It is important to point out that the kind of explanation given here is associative and not causal. This is analogous to the way we as humans interpret facial expressions of our fellow humans. Whether they are happy or sad, aggressive or friendly – we do not infer this from a causal analysis of their psychological and neurological state. The relevant information for this purpose is usually missing. Rather, inferences are drawn on the basis of associations that we have learned through our extensive experience with interpreting facial expressions in the course of our lives.

Of course, our HIV resistance models make mistakes. That is why for some analyses we offer we also provide reliability estimates. Specifically, in addition to the prediction we also return a value that informs the treating physician about how much the prediction can be trusted. We have assessed the predictive accuracy of the models and a number of models have proven to be clinically useful. Especially for highly therapy-experienced patients, who are already highly resistant to HIV, statistical data analysis is much better at identifying promising treatment options than manual approaches to therapy selection.

5 Risks and Limitations of Statistical Data Analysis

In the previous section, the possibilities afforded by the analysis of large data sets were exemplified. In this section we want to take a critical turn in the discussion of data

analysis. After all, data analysis is also fraught with risks and has its limits.

To start with the risks: For this purpose, an example from modern genome-based medicine is inspected, which was already touched upon briefly in Sect. 3. Here one aims at identifying associations between disease patterns and genomic variants of the patient. In short, we are looking for disease genes. The term *disease gene* is quite misleading, because usually everyone has the gene in question. However, the variants of the gene can differ between different people. Only some variants are associated with diseases. What we are looking for are gene variants that occur frequently together with the disease pattern under investigation. Such statistical correlations are investigated with so-called genome-wide association studies (GWAS) (Fig. 5). Here, a large cohort of usually thousands or tens of thousands of subjects is studied, some of whom are healthy and some are sick. We read their genomes or those parts of the genome that are considered to be relevant to the disease. And then, using suitable statistical methods, it is investigated whether genomic variants at certain locations in the genome (gene loci) are found predominantly among patients with a certain disease pattern.

Using GWAS, one has already found numerous associations of gene loci with disease patterns [32]⁴⁾. The study results are shown in diagrams such as Fig. 5. Here the 22 chromosomes of the human genome are arranged along the horizontal axis (the sex chromosomes are not considered). Along the vertical axis a measure of the conspicuousness of the gene locus is plotted. Conspicuousness here means the increased occurrence of a variant of the gene in people with the disease in question [33]. Larger numbers mean higher conspicuousness. It can be seen from the point cloud that only a few dozen gene loci lie above the cloud.

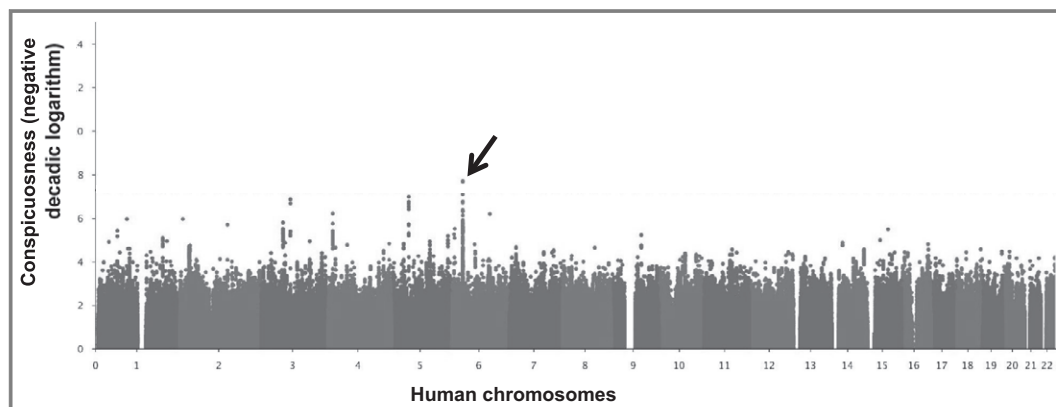


Figure 5. Results of a genome-wide association study for the detection of genome loci that are associated with the intensity of HIV infection. After correct adjustment of the results taking into account the phenomenon of multiple testing, only the gene locus marked with the arrow is significantly associated with the disease pattern. For this gene locus, which codes for a protein of the immune system, the biological basis of this association has been known. Figure adapted from [31], under Creative Commons License CC BY 4.0.

4) see also the GWAS catalog at www.ebi.ac.uk/gwas/

This specific study considered here analyzes the severity of the HIV infection as measured by the virus concentration in the blood of the patient [31]. Fig. 5 would suggest that about two dozen gene loci have a possible association with the disease, namely those that stand out above the ocean of dots in the figure. But this is not the case. According to a more stringent significance analysis, only one gene locus, namely the one marked with the arrow, actually exhibits a statistically significant association with the disease. If we are not stringent enough then, for many genes, we falsely assume an association with the disease. Why is this so? The reason is that we study a great many gene loci. It is possible, but very unlikely, that a single gene locus looks conspicuous purely by chance, i.e., without a biological basis – just as drawing a jackpot in the lottery is possible but very unlikely. However, if thousands of gene loci are examined, such conspicuous outcomes can also occur purely by chance. If many people play the lottery, there is a high probability that there will be a winner. All gene loci that appear conspicuous in Fig. 5, except the one marked with the arrow, are such lottery winners – they are conspicuous without any basis in the data. Such a false call is called a *false positive*. The risk of false positives is particularly high if the analysis includes many similar tests (multiple testing [33]). False positives are a great nuisance in data analysis. They also occur in everyday life, for example in a medical test that suggests a risk of illness that is not confirmed by a follow-up examination, or in a wrongly assumed low creditworthiness that results from an automated analysis of credit standing.

In today's world, in which a great many hypotheses are based on data analyses false positives turn from a risk to a real danger. If the data analysis is performed improperly or conclusions are drawn imprudently, this can lead to erroneous inferences, overinterpretations and eventually to the formation of prejudice, exclusion and stigmatization.

The second problem with data analysis is its high suggestiveness. Today's data analyses in general and the discussed methods in particular are often based only on observations of the system under investigation. There is no controlled intervention into the system. Furthermore, the analyses are associative. They determine patterns of associated occurrence of various features. Two features are associated if they occur simultaneously (correlation), or if the occurrence of one feature is observed more frequently in the absence of the other feature (anti-correlation), in each case in comparison with the individual occurrence of the two features. Associative analyses can afford amazing predictive power. Recently, corresponding case studies received a lot of attention. For example, on the basis of a few dozen likes from Facebook, data analysis can be used to characterize the personality of the user more precisely than can be done by a close friend or partner [34]. Or intimate aspects of the personality structure can be derived from facial images by means of data analysis [35]. Even though the accuracy of such predictions is high, it is not possible to provide a causal justification for the individual predictions. And of course,

there are always false predictions that can have far-reaching negative consequences in certain contexts. To the user, they often suggest causal relationships for which they do not actually provide evidence, and which often are not even true. This is where today's data analysis meets serious limitations.

Let us consider the following examples:

- 1) A genome-wide association study uncovers the association of a gene (variant) with a disease pattern. Does this prove that the gene (variant) is the cause of the disease?
- 2) A statistical study shows that a certain cancer is more common among people who live near a nuclear power plant. Does this prove that the nuclear power plant emits radiation, which is the cause of the increased incidence of the cancer?
- 3) A study shows that Parkinson's disease is less common among smokers. What about causal relationships here?

Let us call the two associated quantities X and Y . In general, associations between X and Y suggest a causal relationship between the two quantities. However, what is cause and what is effect here remains unclear [36]. Fig. 6 shows three possibilities of a causal relationship between X and Y . In Case A is X the cause of Y . In case B it is the opposite. In the most complex case C the association between X and Y is due to a third quantity Z , which is causally related to X and Y . In many cases there are such so-called confounding factors. These are variables that were not measured in the statistical study but affect several measured variables. In a concrete instance of case 2), an actual study has shown that the radioactive radiation emitted by nuclear power plant is so low that it cannot be the cause of the increased frequency of cancer. However, there are other variables, such as the demographic composition of the population in the vicinity of a nuclear power plant, which may well have an influence on the cancer rate. Striking examples such as the fact that the number of pairs of storks in a region is associated with the regional birth rate [37], show how much care must be taken when interpreting the results of data analysis.

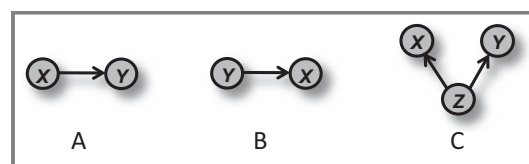


Figure 6. Three possibilities for causal dependencies between associated variables.

However, there are also correlations for which not even such a confounding factor can be found. This is again due to the phenomenon discussed at the beginning of this section, which occurs during multiple testing: If you only have enough data sets available, you will find some that show high correlations, but which are occurring purely by chance. A collection of such *fake spurious correlations* can be found in [38]⁵⁾.

5) www.tylervigen.com/spurious-correlations

Recently, the emergence of big data in statistics has initiated a dynamic development, among other activities, with the aim of deriving causal relationships from statistical associations. However, such associations cannot be revealed by simply observing the system under investigation. Rather we must intervene in the system in a controlled manner, i.e., by experiment. Prospective clinical studies to investigate the effectiveness of drugs and therapies are an example of such an approach. The resulting data are analyzed statistically in order to reveal causal relationships, for example that the administration of the drug is responsible for the patient's improvement. In recent years, this approach has been increasingly supplied with a theoretical foundation and, today, is also applied outside of medicine. An introduction to this field accessible to the general reader is provided by Pearl and Mackenzie [39]. Introductory textbooks are also available [40, 41]. Applications include data analyses in genomics [42] and astronomy [43]. However, the development of these methods has not yet matured to the point where they can be applied to high-dimensional data. And even such methods usually afford no understanding of the mechanistic basis of the uncovered cause-effect relationship. In a clinical trial, e.g., it is proven that a drug works, but not elucidated how or why it works. This deeper understanding must still be achieved by theory formation and experimental confirmation of the theory.

However, the most common type of data analysis used today, namely associative data analysis based on observations, can help to formulate hypotheses about causal relationships or to select a subset of promising hypotheses from an initially overwhelming variety of possible hypotheses. An example is the reduction of the set of hypotheses for the causal relationship between variants in the genome of a patient and his or her disease pattern using a genome-wide association analysis. Here, from the diversity of all genome variants, those are selected that are highly associated with the disease pattern and can, thus, provide an indication of causal relationships. Such clues can then be investigated further using other methods of molecular biology.

6 Data Analysis and Theory Formation: A Strong Alliance

So, what is the conclusion? Is big data the motor of innovation in our present age that many people perceive it to be? According to this viewpoint, the data and especially the patterns contained in them contain all of the useful information. It is quite sufficient that the data afford the derivation of models with high prediction accuracy. It is not necessary that the models also provide causal understanding. Anyway, causality has entered the universe only via the cognitive abilities of humans (and maybe rudimentarily of some other highly developed animals). Previously, nature did fine without this concept and fared well with only the mecha-

nisms of adaptation and learning, which also underlie associative data analysis. So is big data the beginning of a development in society towards increasingly relying more on associations than on the understanding resulting from careful deduction of causal relationships – with all the accompanying symptoms mentioned, such as suspicions and prejudices resulting from data analysis errors, and with the illusion of suggested causal connections that have not been checked?

I think that in the age of big data, associative data analysis is a powerful tool that needs to be used with great care. An associative data model with high predictive power has a high applicability, as we have seen in the HIV example. For administering a therapy to a patient, it is desirable, but not absolutely necessary, to understand the causal relationships. This would require a separate controlled trial for each possible drug combination, which is not feasible. The same applies to many areas of life, wherever an understanding of the causal relationships is out of reach, e.g., in finance or in the analysis of psychological or social processes. However, it is also important to critically examine the results of data analysis and of the process by which the data were generated. In particular, one must always be aware that such analyses do not reveal any causal links. In this way one can guard against the suggestiveness of the results of data analysis. In the future, this statistical competence will become an increasingly important aspect of citizens' digital literacy.

In science, data analysis can serve as an effective initial filter for investigations that uncover causal relationships. In the past, the mind of the researcher alone served as a tool for establishing a plausible hypothesis, which could then be systematically validated or falsified by experiments. Today we can use the instrument of data analysis to systematically select – from an initially unmanageable variety of hypotheses – a limited number of promising hypotheses, which are then validated using methods that have been used in science for centuries. In such a scenario, data analysis is supplemented with a causal analysis and theory building to explain the mechanistic basis for the observed data whenever possible. Especially in science this has been happening comprehensively already for a long time. Here, the predictions resulting from the data analysis are scrutinized with great care. As an instrument for limiting the variety of hypotheses, data analysis in such a scenario is of enormous benefit and often indispensable, indeed.

I would like to thank Christian Lengauer, Jörg Rahnenführer and Nico Pfeifer for a critical review of the manuscript. The case studies in Fig. 1 and 5 were contributed by Anna Hake and Nico Pfeifer, respectively.



Thomas Lengauer is emeritus member at the Max Planck Institute for Informatics in Saarbrücken, Germany, where he was Director from 2001 to 2018. Previously he was Professor at University of Paderborn and University of Bonn and Institute Director at GMD – German Research Center for Information Technology, Sankt Augustin, Germany.

He holds doctoral degrees from Free University of Berlin (Mathematics) and Stanford University (Computer Science). He is member of the Council of the German National Academy of Sciences Leopoldina and President of the International Society for Computational Biology.

References

- [1] G. Grasshoff, Globalization of ancient knowledge: From Babylonian observations to scientific regularities, in *The Globalization of Knowledge in History*, (Ed: J. Renn), epubli, Berlin **2012**.
- [2] J. Kepler, *Tabulae Rudolphinae*, Saurius, Ulmae **1627**.
- [3] J. Kepler, *Hamonices Mundi*, Tampachius, Lincii Austriae **1619**.
- [4] J. Kepler, *Astronomia Nova*, Voegelin, Heidelberg **1609**.
- [5] B. Stephenson, *Kepler's physical astronomy*, Springer-Verlag, New York **1987**, p. 216.
- [6] I. Newton, *Philosophiae naturalis principia mathematica*, Jussu Societatis Regiae ac Typis Josephi Streater, Londini **1687**, p. 8, 381, 383, 400–510.
- [7] A. Einstein, Die Grundlage der allgemeinen Relativitätstheorie, *Ann. Phys.* **1916**, *49* (7), 769–822.
- [8] A. Einstein, *Über die spezielle und die allgemeine Relativitätstheorie*, Springer Verlag, Heidelberg **2009**.
- [9] S. Chatrchyan et al., Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, *Phys. Lett. B* **2012**, *716* (1), 30–61.
- [10] D. Horváth, Search for the Higgs boson: a statistical adventure of exclusion and discovery, *J. Phys. Conf. Ser.* **2014**, *510*, 012001.
- [11] Ž. Ivezić, T. C. Beers, M. Jurić, Galactic Stellar Populations in the Era of the Sloan Digital Sky Survey and Other Large Surveys, *Ann. Rev. Astron. Astrophys.* **2012**, *50* (1), 251–304.
- [12] C. Vitolo, Y. Elkhatib, D. Reusser, C. J. A. Macleod, W. Buytaert, Web technologies for environmental Big Data, *Environ. Modell. Software* **2015**, *63*, 185–198.
- [13] S. Sellars et al., Computational Earth Science: Big Data Transformed Into Insight, *EOS, Trans. Am. Geophys. Union* **2013**, *94* (32), 277–278.
- [14] H.-D. Guo, L. Zhang, L.-W. Zhu, Earth observation big data for climate change research, *Adv. Clim. Change Res.* **2015**, *6* (2), 108–117.
- [15] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, *Phys. Rev. Lett.* **2015**, *114* (10), 105503.
- [16] S. J. Lusher, R. McGuire, R. C. van Schaik, C. D. Nicholson, J. de Vlieg, Data-driven medicinal chemistry in the era of big data, *Drug Discov. Today* **2014**, *19* (7), 859–868.
- [17] K. Rajan, Materials Informatics: The Materials “Gene” and Big Data, *Ann. Rev. Mater. Res.* **2015**, *45* (1), 153–169.
- [18] B. W. Hesse, R. P. Moser, W. T. Riley, From Big Data to Knowledge in the Social Sciences, *Ann. Am. Acad. Political Social Sci.* **2015**, *659* (1), 16–32.
- [19] J. D. Levin, L. Einav, *The data revolution and economic analysis*, NBER Innovation Policy and the Economy, Vol. 14, National Bureau of Economic Research, Cambridge, MA **2014**, 1–24.
- [20] E. Pennisi, Genomics. 1000 Genomes Project gives new map of genetic diversity, *Science* **2010**, *330* (6004), 574–575.
- [21] The ENCODE Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature* **2012**, *489* (7414), 57–74.
- [22] K. Michailidou et al., Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer, *Nat. Genet.* **2015**, *47*, 373.
- [23] E. Marouli et al., Rare and low-frequency coding variants alter human adult height, *Nature* **2017**, *542* (7640), 186–190.
- [24] P. M. Visscher et al., 10 Years of GWAS Discovery: Biology, Function, and Translation, *Am. J. Hum. Genet.* **2017**, *101* (1), 5–22.
- [25] T. Lengauer et al., Chasing the AIDS virus, *Commun. ACM* **2010**, *53* (3), 66–74.
- [26] A. Thielen, T. Lengauer, Geno2pheno[454]: A Web Server for the Prediction of HIV-1 Coreceptor Usage from Next-Generation Sequencing Data, *Intervirology* **2012**, *55* (2), 113–117.
- [27] M. Döring et al., geno2pheno[ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data, *Nucleic Acids Res.* **2018**, *46* (W1), W271–W277.
- [28] M. Zazzi et al., Predicting Response to Antiretroviral Treatment by Machine Learning: The EuResist Project, *Intervirology* **2012**, *55* (2), 123–127.
- [29] N. Beerenwinkel et al., Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype, *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (12), 8271–8276.
- [30] T. Lengauer, T. Sing, Bioinformatics-assisted anti-HIV therapy, *Nat. Rev. Microbiol.* **2006**, *4* (10), 790–797.
- [31] I. Bartha et al., A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control, *eLife* **2013**, *2*, e01123.
- [32] Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature* **2007**, *447* (7145), 661–678.
- [33] P. C. Sham, S. M. Purcell, Statistical power and significance testing in large-scale genetic studies, *Nat. Rev. Genet.* **2014**, *15* (5), 335–346.
- [34] W. Youyou, M. Kosinski, D. Stillwell, Computer-based personality judgments are more accurate than those made by humans, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (4), 1036–1040.
- [35] Y. Wang, M. Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, *J. Pers. Soc. Psychol.* **2018**, *114* (2), 246–257.
- [36] O. O. Aalen, A. Frigessi, What can statistics contribute to a causal understanding?, *Scand. J. Stat.* **2007**, *34* (1), 155.
- [37] R. Matthews, Storks Deliver Babies (p=0.008), *Teach. Stat.* **2000**, *22* (2), 36–38.
- [38] T. Vigen, *Spurious Correlations*, Hachette Books, New York **2015**.
- [39] J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause And Effect*, Basic Books, New York **2018**.

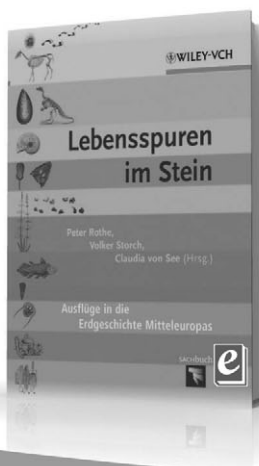
- [40] J. Peters, D. Janzing, B. Schölkopf, *Elements of Causal Inference, Foundations and Learning Algorithms*, MIT Press, Cambridge, MA 2017.
- [41] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, 2009.
- [42] J. B. Pingault et al., Using genetic data to strengthen causal inference in observational research, *Nat. Rev. Genet.* **2018**, 19 (9), 566–580.
- [43] D. Wang, D. W. Hogg, D. Foreman-Mackey, B. Schölkopf, A. Causal, Data-driven Approach to Modeling the Kepler Data, *Publ. Astron. Soc. Pac.* **2016**, 128 (967), 1–13.



Neugierig?

Sachbücher von WILEY-VCH

Jetzt auch als E-Books unter:
www.wiley-vch.de/ebooks



PETER ROTHE, VOLKER STORCH
und CLAUDIA VON SEE (Hrsg.)

Lebensspuren im Stein

Ausflüge in die Erdgeschichte
Mitteleuropas

ISBN: 978-3-527-32766-9
November 2013 300 S. mit
80 Farbabb.
Gebunden ca. € 24,90

Sie heißen Perm, Karbon, Jura, Kreide oder Silur und stehen für geologische Bezeichnungen von Erdzeitaltern. Die faszinierende Wissenschaft der Paläontologie – eine Disziplin zwischen Biologie und Geologie – beschäftigt sich mit den Lebenswelten der Erdzeitalter. Die Autoren stellen die biologische Vielfalt Mitteleuropas während der Erdgeschichte auf einen Blick dar und bieten so eine herausragende und bisher nicht dagewesene Übersicht.

Das Sachbuch basiert auf der höchst erfolgreichen Serie des Magazins *Biologie in unserer Zeit* und ist sowohl die ideale Einführung für Studenten als auch ein fachkundiger Begleiter für alle von der Paläontologie Begeisterten.

Irrtum und Preisänderungen vorbehalten. Stand der Daten: August 2013