Cite as: Dresp-Langley, B.; Ekseth, O.K.; Fesl, J.; Gohshi, S.; Kurz, M.; Sehring, H.-W. Occam's Razor for Big Data? On Detecting Quality in Large Unstructured Datasets. *Appl. Sci.* **2019**, *9*, 3065. https://doi.org/10.3390/app9153065

Review

Occam's Razor for *Big Data*? On Detecting Quality in Large Unstructured Datasets

Birgitta Dresp-Langley [1], Ole Kristian Ekseth [2,*], Jan Fesl [3], Seiichi Gohshi [4], Marc Kurz [5] and Hans-Werner Sehring [6,*]

[1] Centre National de la Recherche Scientifique, UMR 7357 ICube Lab, CNRS-Strasbourg University, 67200 Strasbourg, France
[2] NTNU Trondheim, 7491 Trondheim, Norway
[3] Institute of Applied Informatics, Faculty of Science, University of South Bohemia Czech Republic, 370 05 České Budějovice, České
[4] Informatics Department, Kogakkan University, Ise Mie 516-0016, Japan
[5] Department of Mobility & Energy, University of Applied Sciences Upper Austria, 4232 Hagenberg Austria
[6] Namics—A Merkle Company, 20357 Hamburg, Germany
* Correspondence: oekseth@gmail.com (O.K.E.); hans-werner.sehring@namics.com (H.-W.S.)

Received: 23 June 2019; Accepted: 24 July 2019; Published: date

Abstract: Detecting quality in large unstructured datasets requires capacities far beyond the limits of human perception and communicability and, as a result, there is an emerging trend towards increasingly complex analytic solutions in data science to cope with this problem. This new trend towards analytic complexity represents a severe challenge for the principle of parsimony (*Occam's razor*) in science. This review article combines insight from various domains such as physics, computational science, data engineering, and cognitive science to review the specific properties of *big data*. Problems for detecting data quality without losing the principle of parsimony are then highlighted on the basis of specific examples. Computational building block approaches for data clustering can help to deal with large unstructured datasets in minimized computation time, and meaning can be extracted rapidly from large sets of unstructured image or video data parsimoniously through relatively simple unsupervised machine learning algorithms. Why we still massively lack in expertise for exploiting *big data* wisely to extract relevant information for specific tasks, recognize patterns and generate new information, or simply store and further process large amounts of sensor data is then reviewed, and examples illustrating why we need subjective views and pragmatic methods to analyze *big data* contents are brought forward. The review concludes on how cultural differences between East and West are likely to affect the course of *big data* analytics, and the development of increasingly autonomous artificial intelligence (AI) aimed at coping with the *big data* deluge in the near future.

Keywords: *big data*; non-dimensionality; applied data science; paradigm shift; artificial intelligence; principle of parsimony (*Occam's razor*)

1. Introduction

The Cisco Global Cloud Index 2016–2021 Forecast [1] estimates that nearly 850 zeta bytes (ZB) of data will be generated by all people, machines, and things by 2021, up from the 220 ZB generated in 2016. Most of the more than 850 ZB generated by 2021 will be "ephemeral in nature" and will be "neither saved nor stored". The Cisco forecast states further that "most of this ephemeral data is deemed not useful to save", and that "approximately 10 percent of it is useful, which means that there will be 10 times more useful data being created (85 ZB, 10 percent of the 850 total) than will be stored or used (7.2 ZB) in 2021". Apart from the fact that grasping the significance of this statement in itself represents a challenge, the problem of identifying what is and what is not useful in the growing jungle of information overflow represents, indeed, one of the most pressing challenges for science, business, and society. As a consequence, the *big data* issue, coupled with that of finding new data analytics, radically challenge established theory and practice across the sciences, engendering a new form of scientific uncertainty and paradigm shifts [2] in all major fields of science, from physics to the humanities. Already more than 10 years ago, Anderson [3], among other visionaries, predicted that the *big data* deluge will, ultimately, lead to the end of science and its quest for causality. Whole cohorts of freshly recruited data scientists will occupy their time looking for correlations in non-dimensional input, and any scientist trained the "old-fashioned way" knows only too well that correlation does not imply causality. Instead, universities and research labs now train the young to cope with the volume–variety–velocity–veracity–value (the 5-'v' problem) of a data-driven science and society (Figure 1), which the *big data* deluge has brought upon us without asking for our opinion.

Getting a grip on the many problems represented by *big data* involves being able to detect what is and what is not meaningful in a more and more complex jungle of facts and figures. The effective processing of an increasingly large amount of data, accumulating in the cloud and, in principle, designated to stay there, requires capacities far beyond the limits of human perception and communicability. *Big data* represent a number of problems of steadily increasing magnitude for society, which call for new analytic approaches and conceptual solutions, and science urgently needs to develop the necessary expertise, methods, and procedures to ensure that the data that will ultimately be retained for a useful exploitation will be of the best possible quality, convey genuine meaning, and produce beneficial effects on both science and society.

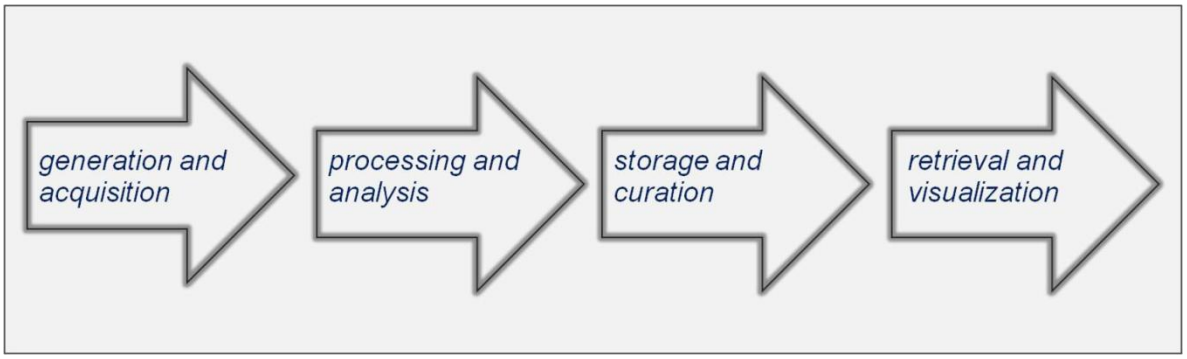

Figure 1. The *big data* value chain [1]. All steps in the chain, from data generation to retrieval and visualization for exploitation to the benefit of science, business, and society, are subject to problems relative to the *volume*, *variety*, *velocity*, and *veracity* of the data, affecting their ultimate *value* to science and society (*5-'v' problem*).

The principle of parsimony, or *Occam's razor*, stems from writings on the logic of explanation by the cleric William of Occam [4]. In science, *Occam's razor* reflects both a general rule of prudence and a conservative guideline for investigation that consists of aiming for the simplest among possible explanations for, or model approaches to, the phenomenon under study. While the principle of parsimony has been implicitly adhered to across and within all major scientific disciplines (mathematics, biology,

physics), its fundamentals and application are severely challenged by the emerging trend towards increasingly complex analytic solutions in data science within the context of *big data*. This new problem space requires new strategies of data analysis and requires maintaining an open mind to alternative possibilities, a new philosophy as to how data analysis is be carried out, rather than any fixed set of method [5,6]. The complexity of the *big data* problem space is such, that there can be no single, novel or established, solution for data quality control [5], and there is an urgent need for wider critical reflection within the scientific community on the epistemological implications of this unfolding data revolution along with the rapid changes in research practice presently taking place [3].

This collaborative review article results from an international panel consortium effort at the 'Cognitive 2019' conference in Venice [7] and represents a "think-tank" approach towards critical conceptual knowledge construction on some of the problems represented by *big data*. The paper benefits from insight from various domains such as physics, computational science, data engineering, and cognitive neuroscience. Section 1 summarizes the specific properties of *big data* and the problems these pose for science and, in particular, applied data science. Section 2 provides a synthetic overview of big data analytic solutions from machine learning to artificial intelligence, followed by a more detailed discussion of two examples of parsimonious analytics in the *big data* context: Clustering for determining data structure of any kind, and parsimony-driven analysis of single-pixel change in large sets of image data with minimalistic artificial intelligence. Why we still massively lack in expertise for exploiting *big data* wisely in order to extract relevant information for specific tasks, recognize patterns and generate new information, or simply store and further process large amounts of sensor data is summarized in Section 3 on the example of the *smart city* concept. Issues relative to model building and reasons why we need subjective views and pragmatic methods to analyze *big data* contents are brought forward in Section 4. Section 7 concludes the review by evoking a few thoughts on how cultural differences between East and West are likely to affect the course of *big data* analytics, and the development of increasingly autonomous artificial intelligence (AI) in the near future. *Big data* represents a problem our cultural development has brought upon us, and without this problem, there would be no need for highly developed artificial intelligence.

2. Testing *Big Data* Quality: Potential and Limitations

The *big data* concept is a new trend that occurs in many scientific directions. It may seem that such a concept has been around for many years but, in reality, the *big data* of today have created radically new challenges and also pose a number of new problems. These are directly related to their properties, which are the following:

(a) *Uniqueness*—the volumes of *big data*sets are mainly presented in the scale of EXA, PETA, or ZETA bytes [8]. This means that such datasets are retrieved from many specific unique sources, which cannot be easily exchanged by other optional e. q. values from sensor networks located in a specific area. Another reason for is related to the time necessary for the data production and the collection time, and it is not rare that these move into the range of tens of months or more [9].

(b) *A-dimensionality—big data* mostly have no concrete structure, are unsorted, and their value distribution functions are typically unknown [10]. For many *big data* types, it is impossible to sort the data according to the value of a specific parameter (e.g., speech samples, pictures), because they are not straightforwardly comparable.

(c) *Specificity*—this feature has much in common with data uniqueness. In general, the datasets are retrieved from many sources, and their content is quite specific. This means that what is valid for one dataset need not be for another. This specificity can have many different reasons—like the data type, resource type, geographical context [11], or other.

(d) *Cost*—their storing and processing requires large and expensive, high-capacity data stores and powerful distributed computing systems.
(e) *Unpredictability*—it is mostly unknown which are the "correct" or "expected" data values. This implies that the data optimization cannot be performed at the same time as the data collection. To predict the data quality, it is mostly necessary to analyze the entire dataset.

2.1. Big Data Reading and Storing

For an unknown big dataset, it is suitable to first consider the way of storing. That means to analyze data storing structures—the data representation format (textual or binary form), the data encryption type, etc.—clearly; one of the most important conditions prior to efficient data processing is the proper location for storing. The most frequent solutions today are XML files, distributed file systems [12], content delivery networks [13], clouds, or special databases [14]. The concrete efficient solution depends on the data type. Completely different solutions may be adopted, for example, for image datasets versus textual information from social networks.

2.2. Ecosystem and Tools for Big Data Processing

The preprocessing phase concerns data cleaning, outlier detection, normalization, interpolation of missing values, or noise filtering (noise reduction). For an unknown dataset being considered a high-quality dataset, it is practically impossible to perform these operations. The reason lies in the lack of knowledge of the data distribution before preprocessing, which makes it impossible to sort the data on the basis of criteria. The next steps of the processing phase concern dimensionality reduction, feature selection, and discretization. The general approach of selecting a representative subset of the data is not working very well here, because big datasets are mostly unstructured. A similar conclusion can be drawn for data normalization. There are several algorithmic approaches, designed especially for big datasets. Still, these approaches require prior knowledge relative to which the correct data values should be [15]. For the preprocessing of *big data*, distributed frameworks are used. The major platforms used are listed here below. Many of them are developed by the Apache Foundation in the framework of their Apache *big data* processing ecosystem.

(a) *MapReduce* [16] (Hadoop v2)—allows the storing of files via HDFS or the processing of stored data values. This solution uses an efficiently scalable distributed architecture.
(b) *Computing Engines* [17] (Spark, Storm)—approximately 100x faster than traditional MapReduce, used in many solutions.
(c) *Processing Pipelines* [18] (Kafka, Samza)—basically targeted on efficient caching of procedures that allow further processing.
(d) *Databases* [19] (Cassandra, Hive)—allow multiple-times faster data search in comparison to the traditional SQL-based approach.
(e) *AI-based frameworks* [19] (Mahout, ML over Spark)—help to use the traditional machine learning methods on *big data*sets.

2.3. The Problem of Big Data Quality Evaluation

The evaluation of the quality of an unknown big dataset is a highly complex and difficult task. For specific data, it is practically impossible to find missing values, check or normalize the values, and to detect the often-substantial amount of additional noise. Such datasets have in common that it is, in principle, unclear how to compare their values, or consider which values should be correct. Commonly known datasets can be evaluated by using the traditional approaches, which have been adapted to the scale of big datasets. The current analytic ecosystem solutions summarized above are able to process these

datasets quickly and efficiently. However, due to the unstructured character of big data, it is not easy to reduce the data volumes to representative data sub-sets, which are much smaller, but preserve the original properties of the whole dataset. The identification of outliers represents a similar problem case, because it requires prior knowledge of predicted/predictable values. Data analysis and data mining cannot be considered separately from *big data* quality assessment [20], and it is necessary to use a large variety of these methods, or analytics, to discover whether valuable information or knowledge exists or not in *big data,* and whether this knowledge is useful or not. Poor data quality will lead to low data efficiency and produce decision errors. Given the data diversity, hierarchical structures for a data quality framework with assessment procedures [20] will not be readily adaptable to any kind of data but remain very specific, applicable only to a particular domain. Thus, the era of *big data* is inspiring the search for new approaches and algorithms. This search for new solutions will require exploring radically new possibilities in directions well off the already beaten tracks of computer science and incite data scientists to venture in territories nobody has ever ventured before.

3. Big Data Analytics: From Machine Learning to Artificial Intelligence

A major challenge for *big data* analytics consists of the accurate and fast segmentation of large datasets consisting of parameter values, generally expressed in numbers, into quality clusters, i.e., those which optimally define the input, which is less than a straightforward matter. For example, for a dataset with n = 1000 feature rows, there would be 10^{301} possible combinations to evaluate in cases where prior knowledge relative to how the data are organized is unavailable. While the current trend is to believe that artificial intelligence and deep learning in increasingly larger neural network architectures will provide the all-encompassing answer to questions on how to address this problem, the scientific principle of parsimony (Occam's razor) implies starting by trying to improve the accuracy of data mining without "reinventing the wheel" altogether. This should be possible by building on machine learning algorithmic approaches which have proven their worth. When combined adequately in a building block approach, simpler approaches may prove far more powerful than other, more complex algorithms and improve the quality of data mining and software analysis, as will be illustrated later herein on the basis of specific examples.

Machine-learning (ML) systems [21] identify objects in images and select relevant results of search. The classic supervised machine learning approaches were initially limited in their ability to process data in its raw form and the analytic systems required specific domain expertise to program feature extractors that transform raw data into a valid internal representation vector for learning and input classification. This "old-fashioned" way of programming had the considerable advantage that the result obtained could be linked to well-identified and controlled steps in the machine-learning procedure. In the current context of *big data* analytics, unsupervised ML algorithms prevail. We propose that the most prevailing analytic approaches to *big data* may be arbitrarily ranked into three categories: (1) '*Educated*', (2) '*Wild*', and (3) '*Artificial Intellligence*' (AI); an overview is given in Figure 2.

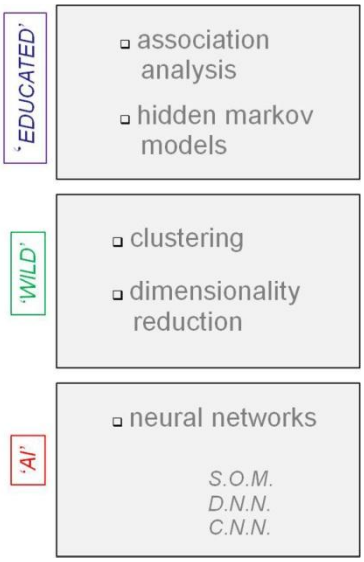

Figure 2. Analytic approaches to *big data*: From unsupervised machine learning ('educated' or 'wild') to artificial intelligence (AI).

Unsupervised ML methods of the 'educated' type (Figure 2, top) imply educated guesses (hypotheses) relative to the data structure and are aimed at generating predictions. The most common of these would be hidden Markov models, which have been applied in the *big data* context to generate spatio-temporal predictions based on clusters of interest in large video datasets [22], and association analysis or rule-mining, aimed at finding frequent items in a database with the least complexities to

generate predictions. Association analysis for *big data* is limited by the problem of finding large enough item clusters in datasets [23]. Making assumptions about trends in large data to generate predictions makes no sense if nothing about the quality of the data is known a priori.

The unsupervised ML methods of the 'wild' type (Figure 2, middle) make no particular assumptions about the data, and are therefore a first choice for detecting quality and structure in unknown *big data*sets. These different methods can easily be adapted and combined for the identification of meaningful groups in large and unknown datasets. They employ algorithms for data clustering [24, 25] and dimensionality reduction, often by principal component analysis [24], to reduce the size of very large data in high dimensional (generally Euclidean) space to smaller sets of weighted data points, or core datasets. Essential properties of clustering algorithms as a first-choice analytic solution for detecting quality in *big data* will be pinpointed further in a subsection here below.

Neural network structures, which are inspired by functional properties of neurons in the primate brain, are synonymous with AI (Figure 2, bottom). The so-called deep neural networks (DNNs) [26] are the most recent and also the most complex breed of representation-learning algorithms, with multiple levels of representation obtained from groups of non-linear modules that transform representations at one level (starting with the raw input) into representations at higher, increasingly abstract levels. With many such transformations, increasingly complex functions can be learned. For classification tasks, higher layers of representation amplify certain input data (important for discrimination, and suppress irrelevant variations. The key aspect of deep learning is that it is, by nature, unsupervised, i.e., feature processing is no longer controlled by a human expert, and the models learn to process features from the initial input data fed into a given machine learning procedure, of which there are many [21**].** Deep learning may outperform other methods in *big data* classification with respect to accuracy of prediction [27], but not necessarily in other specific tasks, such as image-based cell-type annotation, for example [28]. It is not always clear which processing step in a deep learning approach would account for better results obtained [29].

Convolution neural networks (CNN) are a type of deep learning models, often termed deep CNN [27]. They have an input layer, and output layer, and hidden layers. The hidden layers usually consist of convolution or pooling layers, and all layers are fully connected. Convolution layers apply a convolution operation to the input, then the information is passed to the next layer. Pooling combines the outputs of clusters of neurons into a single neuron in the next layer. Fully connected layers connect every neuron in one layer to every neuron in the next layer. In a convolution layer, neurons only receive input from a subarea of the previous layer. In a fully connected layer, each neuron receives input from *every* element of the previous layer. The network works by extracting features from images, which are learned while the network trains on a set of images. CNNs learn to detect features through tens or hundreds of hidden layers. Each layer increases the complexity of the learned features. The problem of any of these deep learning algorithms with respect to data quality detection is that they generally do not apply parsimony rules for eliminating excessively large amounts of random or meaningless data in a set [30]. This constitutes a drawback for any kind of data domain where precision of data description matters.

The simplest form of an AI neural network is the self-organized map (SOM), sometimes also called Kohonen map [31–33]. To provide a snapshot view of the difference in complexity between the SOM and a DNN, a graphic illustration of their functional properties is shown in Figure 3 here below. In its most elementary form, the SOM has a functional architecture with a single input layer, a single processing layer (neurons), and a single output layer. One of the advantages of the SOM analytic is that it reduces the input dimensionality in order to represent the input distribution as a map. In image analysis [34], such dimensionality reduction allows performing analyses of large image series with millions of pixels per image with to-the-single pixel precision. In its most elementary form, the SOM uses a winner-take-all learning strategy, and has hitherto unsuspected potential as a parsimonious and effective image data analytic, as will be explained in further detail in the subsection here below.

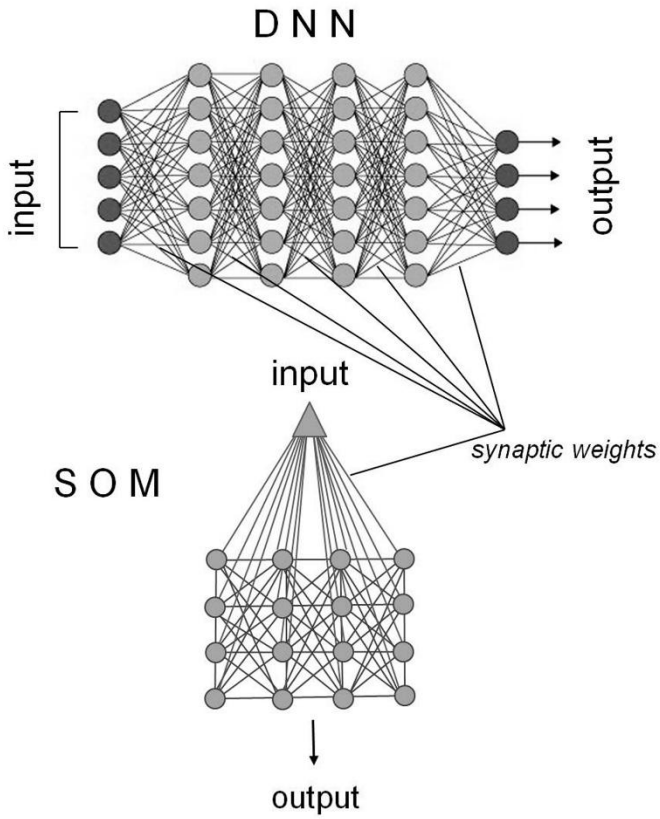

Figure 3. Functional architecture of a deep neural network (DNN) (top) with four fully connected hidden layers of neurons between input and output layers, and that of a self-organized map (SOM) with a single layer of fully connected neurons between input and output. The SOM reduces the input dimensionality in order to represent the input distribution as a map. For learning, it generally uses winner-take-all, while a DNN may use a multitude of learning algorithms, at multiple steps of processing.

3.1. Clustering Algorithms: Old Dogs for New Tricks

Clustering has been around for a rather long time [25] and, as has become clear from the overview given in the previous section, before diving into the deep blue sea of deep learning, clustering methods are among the most essential parsimonious analytics to consider first. They prove ideal candidates for building on simplicity in order to (1) improve the quality of large data analysis and (2) minimize the necessary computation times. Cluster analysis is well-suited for identifying meaningful groups in large and unknown datasets [35–47]. Algorithms such as "k-means" [37] use pair-wise similarity metrics such as the Euclidean to identify the cluster–vertex memberships, where the Euclidean is by far the most popular choice for "k-means" functions. Considerable reduction in computation times can be obtained through optimization of similarity metrics [47] and local heuristics [48].

For cluster algorithms to provide accurate results, characteristics in the training data need to be captured by the algorithm. Cluster algorithms are dedicated to specific data topologies, hence the need for data normalization. Yet, the smoothing of data results in a loss of information [49,53]. For a similarity metric to improve prediction quality, the similarity metric needs to reflect the core attributes of each dataset [48], a requirement that is satisfied the "hpLysis" software solution for example [53]. The use of data normalization strategies may hide key features of datasets [49], an observation that may explain, just to give an example, why genes with a strong differentiation are not always detected in microarray analysis [49]. The default strategy here consists of applying averaged normalization [53–58]. While [59]

argues for tuning algorithms towards data topologies, this strategy is rarely applied. There is no "best clustering algorithm", and any algorithm imposes a structure on the data given [59].

To address issues relative to data topologies, [59–63], strategies for dimensionality reduction and feature extraction as a pre-processing step to clustering are proposed [61]. However, these strategies are known to provide a relatively poor performance [64] when compared with the fine-tuning of algorithmic building blocks [45]. Clustering algorithms are generally used for image segmentation, tracking of individuals in CCTVs, classification of digital signals, and so forth. Comparisons of measurements from 20+ cluster algorithms and 300+ correlation metrics have shown that established cluster algorithms have a speed and accuracy considerably higher than that of other recently proposed algorithms. Some work [49,50] identified 100+ new cluster algorithms that may be successfully applied to large datasets in fields such as digital signal processing and/or image analysis. Measurements across a large number of cluster algorithms and more than 100 real-life datasets reveal how the right parameter choice for clustering can outperform a large number of other recent algorithmic approaches to the problem of large datasets [65].

In the case of big data, the motivation is to increase the performance of clustering for data mining, and to identify new data mining algorithms without "re-inventing the wheel" altogether. This can be achieved, as argued here, by the integration of "trivial" algorithms to outperform more "complex" ones [48,49]. In this context, exploiting machine learning strategically to identify new clustering algorithms considerably improves the quality of image analysis and digital signal processing, for example. It represents a performance boosting strategy based on:

(1) Using building blocks from the parameterization of a large number of cluster algorithms to optimize the clustering accuracy,
(2) A combination of local and global heuristics that outperforms the speedups enabled by any single clustering algorithm, and
(3) An unsupervised automated (machine learning) procedure to construct many new clustering algorithms from the training data.

A major challenge when estimating a set of new algorithms consists of finding a precise definition that states, in clear terms, what exactly a clustering algorithm is supposed to do, at which level, and under which boundary conditions. In practice, there is no well-defined separation between algorithm, software, and software–algorithm configuration. Although mostly 'wild' by nature as clarified in Section 2 here above, clustering algorithms sometimes carry implicit (tacit) assumptions about input data and "ideal" cluster outcomes, i.e., heuristics, as is the case for column-based normalization, Kendall's Tau correlation metric [50,51,66], or the "Silhouette" convergence metric [67], for example.

In the context of *big data*, clustering algorithms have similar potential and face similar challenges as genetic algorithms (GA) and fuzzy clustering. This is due to the combinatorial space between "training data" versus "user hypothesis" versus "real-life data". For k-means, GA, and fuzzy clustering to provide accurate predictions, they may be adjusted to topological traits [68–77]. A genetic algorithm (GA) is understood as a "generic term subsuming all machine learning and optimization methods inspired by neo-Darwinian evolution theory" [74,75], and is often used in a combination with other methods for accurate inference of knowledge, like fuzzy clustering, which is a permutation of k-means [76]. There is no clear separation between GA versus k-means, and GAs may be merged directly with k-means [70–77], or k-means may be optimized through a GA seed selection [78–83]. In the context of big data, the use of classical clustering algorithms combined with accurate training is inseparable from the practical application of GA, or evolutionary algorithms in general. Dimensionality or model order reduction in very large data can be achieved by Fuzzy C-means clustering [84], or by astutely combining k-means with principal component analysis [85].

3.2. SOM for Single-Pixel Change Detection in Large Sets of Image Data

In the image and vision sciences, at the intersection between machine learning algorithms for computer vision and computational neuroscience, low-level artificial neural networks have been proposed well before the area of big data and deep learning neural networks. The earliest AI approach in this context is, as explained in Section 3 herein, the self-organizing map or SOM [31–33,86,87], is largely inspired by the functional properties of visual neurons in the non-human and human primate [88–91], and by psychophysical data on human detection [92–101]. How elementary output parameters of the SOM may be exploited for detecting certain qualities in large, arbitrarily archived image datasets may be illustrated here on the example of rapid, unsupervised detection of strictly local changes invisible to humans in image time series. Such local changes in images may reveal different meaningful states of a physiological structure, tissue, or cell, and reflect progression or recession of a pathology, or the progressive response of a cell structure to treatment, not detectable by any of the currently available medical image analysis tools.

A simple functional architecture of the self-organizing map may be applied to the unsupervised classification of massive amounts of patient data from different disease entities ranging from inflammation to cancer, as shown recently [35]. Other recent work [102–105] has shown the quantization error (*QE*) in the output of a basic self-organized neural network map (*SOM*); in short, the SOM-QE is a parsimonious and highly reliable measure of the smallest local change in contrast or color data in random-dot, medical, satellite, and other potentially large image data. The SOM is easily implemented, learns the pixel structure of any target image in about two seconds by unsupervised "winner-take-all" learning, and detects local changes in contrast or color in a series of subsequent input images with a to-the-single-pixel precision, in less than two seconds for a set of 20 images [106,107]. The QE in the SOM output permits to scale the spatial magnitude and the direction of change (+ or −) in local pixel contrast or color with a reliability that exceeds that of any human expert [102].

Applied to the automatic classification of arbitrarily archived scanning electron microscopy (SEM) images of $CD4^+$ T-lymphocytes (so-called *helper cells*) with varying extent of HIV virion infection [101], a four-by-four SOM (for an illustration, see Figure 3) may be trained on any image of a given series, whatever its size, using unsupervised winner-take-all learning. The functional architecture of the SOM is minimal, with a constant number of neurons and a constant neighborhood radius [31–33,86,87]. Winner-take-all learning and subsequent SOM processing to generate QE output for the automatic classification of images from a series of 20 takes less than two seconds. Each of the images from the series is associated with a SOM-QE output value. These values are then automatically listed in ascending order of magnitude on the y-axis, the image corresponding to each QE value is ranked on the x-axis as function of its SOM-QE label. This analysis achieved a 100% correct classification of the $CD4^+$ cell images in the order of the magnitude of HIV virion infection displayed in an image. A human expert electron microscopist with more than 25 years of experience took 52 min to sort the color-enhanced micrographs in the correct order, and 69 min to roughly sort the original grayscale images, but not with 100% correct. Moreover, for the human expert to be able to perform such a task at all, it is necessary to visualize all the images of a given series simultaneously on the computer screen, using software that allows zooming in and out of single images for close-up two-by-two comparisons. Thus, since quality and clinical meaning in larger sets of imaging data cannot easily be perceived by expert visual inspection, we need to resort to parsimonious and effective computational blocks to assist expert analysis as the image datasets grow larger. The SOM-QE outperforms the RGB mean in the detection of single-pixel changes in images with up to five million pixels [107], which could have important implications in the context of unsupervised image learning and computational building block approaches to large sets of image data, including deep learning blocks for automatic detection of contrast change at the nanoscale in transmission or scanning electron micrographs (TEM, SEM), or at the sub-pixel level in multispectral and hyper-spectral imaging data [108]. Whatever the method of image analysis, at the beginning, any input image is broken down into pixels, and the analytic process starts from there; from utmost simplicity towards increasing complexity. With respect to

the principle of parsimony, the simplest network architecture that does the job should be the one that is chosen. However, this is not systematically adhered to in contemporary data science, where complexity often outvotes simplicity [109], and where data scientists can "learn" to build a DCNN "in just ten minutes" through *Google*. This is a hazardous trend [5,9], and we may lose the track of what works best, at which level of explanation, and within which limits. This may then result in seemingly successful data analysis by sheer "fluke", where nobody asks why exactly it works, and which processing step(s) in the model explain(s) the results obtained.

4. Can Cities Become Really "Smart" or Will the *Big Data Jungle* Continue to Proliferate?

Mark Weiser stated in his article "The computer for the 21st century" [110] back in 1991 that "the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it". Following this vision, Weiser is often seen as the forefather of ubiquitous computing [111], whereas he defined a smart environment as "the physical world that is richly and invisibly interwoven with sensors, actuators, displays, and computational elements, embedded seamlessly in the everyday objects of our lives, and connected through a continuous network". Thus, the evolution of connected devices of everyday life—commonly referred to as Internet of Things (IoT)—allows for an ever-increasing amount of data characterizing different facets of our everyday lives. When referring to "smart cities", a lot of different aspects have to be considered. As shown in Figure 4, inspired by [112], the landscape of building blocks of smart cities is very widespread. In addition to the illustrated areas where large amounts of unstructured data are generated (e.g., smart grid, smart health, smart people, smart and interconnected cars, etc.) a lot of other data-producing entities in different areas play a vital role in this ever-increasing amount of unstructured data. With the advent of inter-connected sensing devices delivering constant streams of data, the *big data* jungle is constantly growing and proliferating. Different applications exchange or produce information using embedded, virtual, or other sensor devices [113]. Be it a smart-home combined with smart-grid technology, or interconnected vehicles (i.e., "smart cars"), data-delivering smart phones, or even data extracted from online user behavior in the area of e-commerce [114–116]. Data are being generated everywhere at any time. Thus, with the technological capabilities and the sensing devices integrated into our daily lives, we have the capability to sense the world, and gather massive amounts of—rather unstructured—data. The problem is that we still massively lack in using this data wisely in order to (i) extract relevant information for specific tasks, (ii) recognize patterns and generate new information, or (iii) simply store and further process data. Possible reasons for this fact could be:

- The variety of different types of data and the common unstructured nature of data arising from (i) different application domains, (ii) different devices delivering the data, and (iii) proprietary solutions;
- The fact that there is not "the one big data" algorithm—artificial intelligence as a research field provides a lot of different algorithmic and technical approaches; nevertheless, all of them are application-specific;
- That data could be error-prone;
- Data could simply be misused (i.e., using data in a "wrong" way or within the wrong context or domain);
- Computing resources and tools are not braced for coping with massive amounts of data.

Taking into account all the aforementioned aspects, the research challenge in applied science has now shifted from data acquisition to data analysis. The data are there—or at least we believe that we know how to access them—and it is now time to extract relevance, to decide what matters, and what does not. The simple observation that the amount of data is steadily increasing has been subject to research in recent

related work. For example, Marx [114] presents challenges for *big data* from the perspectives of medicine and biology and argues that a "data explosion" has happened (and is currently gaining in volume). Additionally, with such an increasing amount of data, the fact that in an entity on a city scale, where different areas (e.g., smart homes, smart grids, smart cars, connected people, etc.)—each for itself delivering massive amounts of unstructured data—deliver data, it is almost impossible to extract relevant information or detect quality and meaning [117]. Pruning the big data jungle is subject to research, whereas the recent advances in artificial intelligence seem to be promising approaches. Unsupervised learning [118] that intends to discover patterns in unlabeled and unstructured data or detect anomalies could be an algorithmic solution. As already shown and discussed in the article sections here above, there is not "the one algorithm for *big data* analysis". The complex problem space represented by *big data* cannot be addressed by any single, novel or established, analytic solution. The potential value of any computational solution will depend on the application domain, the characteristics of the dataset given, and the specific use case in its context. Thus, for each specific problem, the best algorithmic solution is bound to be different from that for other problems in the complex universe of *big data*.

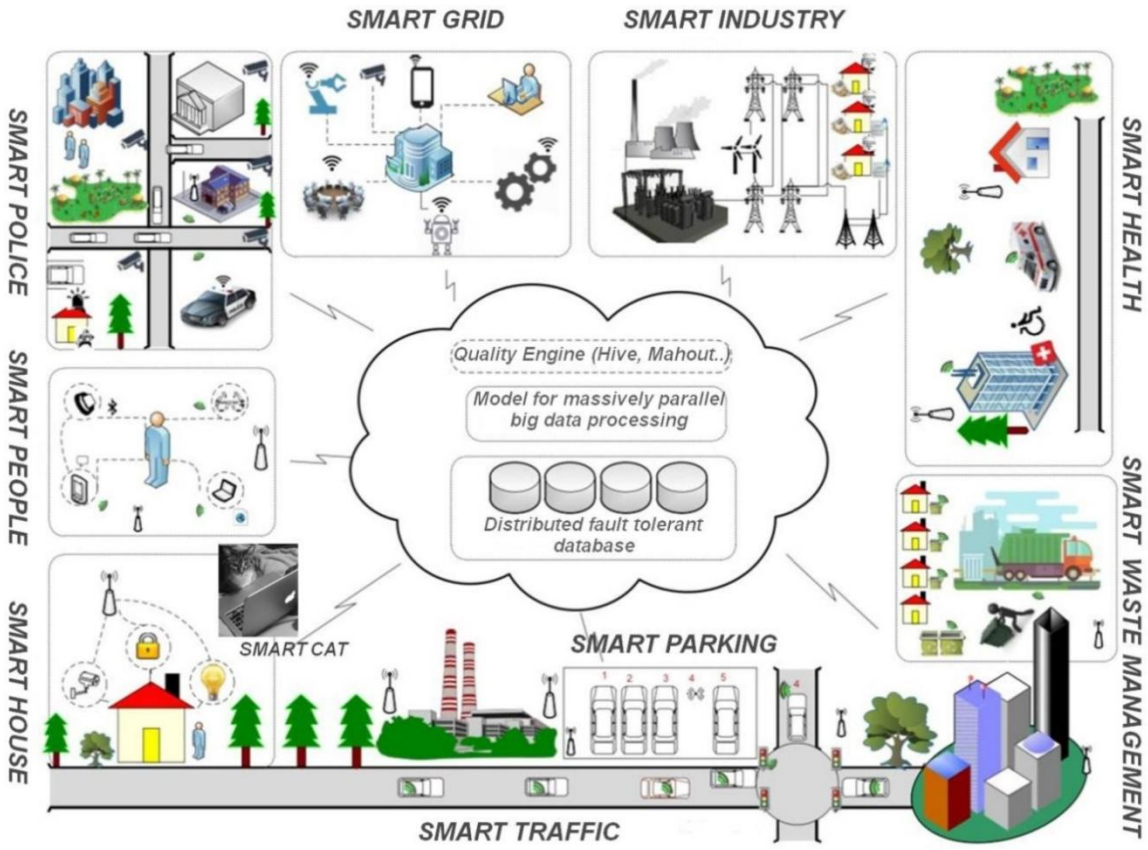

Figure 4. Idealistic view of a landscape of building blocks of *smart city* components through *big data* technologies, adapted from [112].

The sheer amount of data in itself poses a major problem, whatever the kind of data or application domain. As shown in Helbing et al.'s [117] model for digital growth (for an illustration see Figure 5), computational resources double about every 18 months (*Moore's Law* [119]) and data resources double about every 12 months. These two resources follow an exponential growth, while the processes to analyze data commonly follow a factorial growth resulting in large amounts of data that cannot be processed at

all. Frangi et al. [120] refer to this phenomenon as "dark data"—i.e., the data that cannot be processed due to the high systemic complexity, information content, and the meaning they carry. The big data jungle is inexorably growing, since the world is polluted with sensors and data-delivering entities.

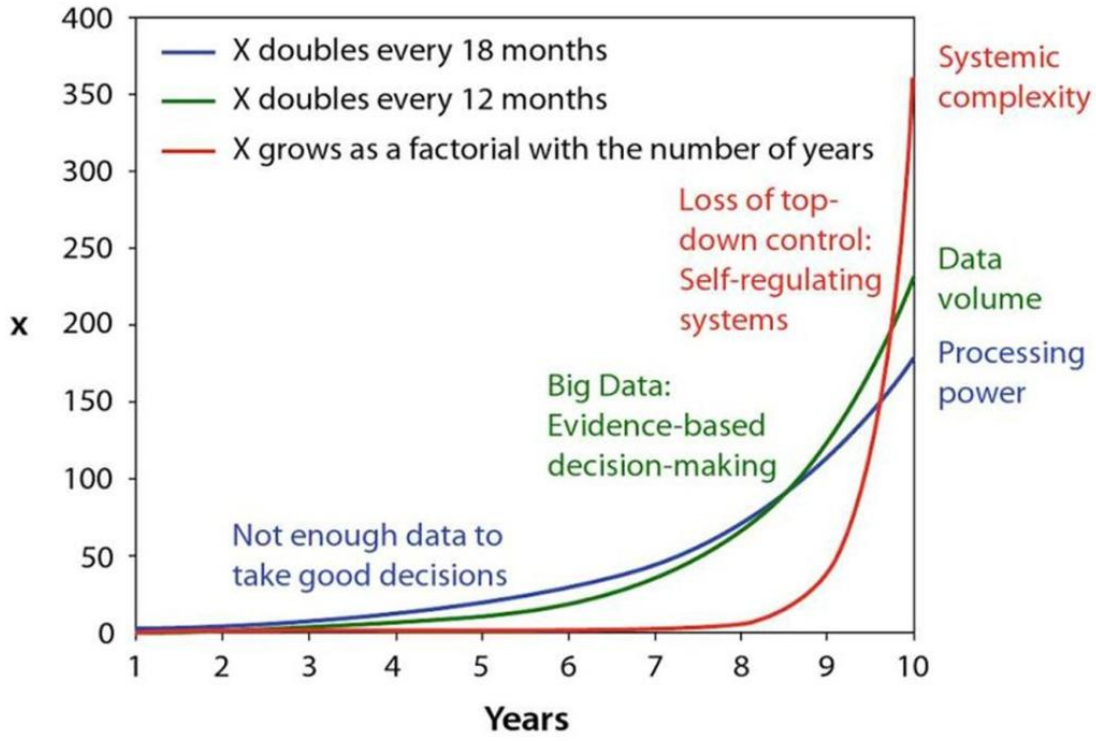

Figure 5. Helbing et al.'s [117] model for digital growth comparing computational and data resources with systemic complexity to analyze data.

5. The Chicken-or-the-Egg *Paradoxon* of Model Building or Why We Need Subjective Views and Pragmatic Methods to Analyze *Big Data* Contents

Data management for *big data* analysis requires particular attention, not only as a result of the quantity of data involved, but also because of the requirements of the data analysis process itself. Traditional database approaches concentrate on the management of data and of the schema behind it. More generally, we can, say, speak of *instances* and of a *model* for their description. A database model is predefined. It is used as a blueprint for newly created instances, it formulates rules and constraints by which to manage instances, and it defines how data can be queried.

One class of big data research studies the querying of large amounts of data, where queries require some model for the data. The kind of big data analysis addressed in this article works quite the other way around: Given data with "partially known" structures, find a model that generalizes properties of interest and that derives rules and constraints for datasets.

The ANSI-SPARC Architecture [121] that received particular attention for relational databases defines three levels of models: The *external* level that defines user-specific views on the data, the *conceptual* level that describes the whole data of a database logically, and the *internal* level that describes how the data are maintained on the physical level. Applying the architecture (in a kind of reverse manner) to big data analysis, we aim to find external views to describe particular aspects of a large set of data that has been created according to an (unbiased) conceptual model. Most information systems incorporating a

database work with data that describe real-world entities. The data are a representation of things that exist outside of that system. This leads to the observation that a conceptual model serves two purposes at once: On the one hand, it is a technical description of the computational properties of data. For example, it determines that some data represent a non-negative integer number. On the other hand, it reflects an application domain. For example, it formulates that some data represent the weight of a person and thus that the unit of measure of grams is attached to it and that values typically lie in the range of 2000 to 200,000. While the former aspect of technical representation is mostly covered by type systems/database schemas, the latter aspect of domain-specific interpretation is buried in constraints and in application code that make use of the data.

Typically, a model is considered adequate if all data under consideration conform to it, if it allows capturing the meaning of a suitably large extent of the modeled application domain, and if it is minimal in the sense that no model element can be omitted without violating the first two requirements (*Occam's razor*). For all of these properties, it is required to know the bounds within which a model may be applied with respect to the application domain at hand. A first step towards deriving such a model is to find *content* in the set of data—that significant data that can be interpreted as a meaningful description of a domain entity [122]. There are two basic ways content can be represented by data: Firstly, the data can represent the entity directly, e.g., if those data are a (unique) product number that identifies some sales item, or if the data provide a photo of the entity. Secondly, the data can describe one property of the entity, and a sufficient number of such properties characterizes the entity.

Data representing entities directly serve as a sign as studied in semiotics. The triadic semiotics of Charles Sanders Peirce proposes a three-layered hierarchy of signifiers (*Firstness*, *Secondness*, and *Thirdness*) [123]. There are three kinds of representations [124]: On the first level, *icons* refer to entities directly. On the second level, an *index* has a relation to the signified entity. On the third level, a *symbol* denotes a rule that describes entities. This allows distinguishing data that directly signify an entity, data that establish an entity relationship, and data signifying a rule or constraint on entities. An instance on one layer is built upon ones on the next lower layer. Therefore, Peirce's semiotics indicates that data that directly signify entities are required as a basis for derived model elements like relationships and rules.

These different levels of signification are, e.g., similar to normal forms of relational database schema. Third normal form provides a kind of iconic signification by a primary key. Second normal form typically achieves attribute independence from partial key candidates by the creation of referenced tables, thus creating "Secondness" indexes. Tables in first normal form typically require recognizing structures in order to identify rows that together describe an entity. Therefore, such row sets are interpreted as symbols. The creation of hierarchies of entity descriptions through content has been studied in the *concept-oriented content management* (*CCM*) approach [125]. The most important insight is the subjectivity of the signs found. This is discussed below. Furthermore, it turned out that the ability to recognize signs depends on a conceptualization of the application domain. Dual to the signifying part of data, these findings directly relate to the descriptive parts of data that provide a conceptualization of entities. The simplest kind of such descriptive data is a reference to a concept (which in turn is represented by data, as a "Thirdness"). This is, e.g., employed on the Semantic Web [126].

In a model, a *concept* names a phenomenon in the domain, and rules and constraints may come with it. This way, concepts describe classes of domain entities. A technical concept is a data type ("integer") and might be given for a particular database. A domain concept captures domain knowledge ("typical human weight") and can, therefore, not be predefined in a generic way. It carries some domain semantics that may be given as *intensional semantics*, i.e., by rules and constraints that data of a certain class have to fulfill, or by *extensional semantics* through the set of data that it is assigned to [127].

Figure 6 illustrates this problem on the example of interpretations of the painting *Napoleon Crossing the Alps* by Jacques-Louis David (example taken from a CCM application [128]). The photo of the painting is an iconic representation of that painting and of Napoleon since he is depicted. To art historians, the

image is also an index for Hannibal crossing the Alps since the scenery visually cites this historical event. The image is a symbol for equestrian statue, an object of art that is characterized by the ingredients to be seen in the picture (equestrian, horse, pose, …). The photo is intentionally defined by the concepts attached. The concept *Strength*, in turn, is extensionally defined by some further images. The two ways of defining concepts have been derived from the epistemology of Ernst Cassirer [129–131]. From our perspective, Cassirer's work contains a possible answer to the question whether to start with a model (as databases do) or to start with data (as data analytics do): Instances and models co-evolve.

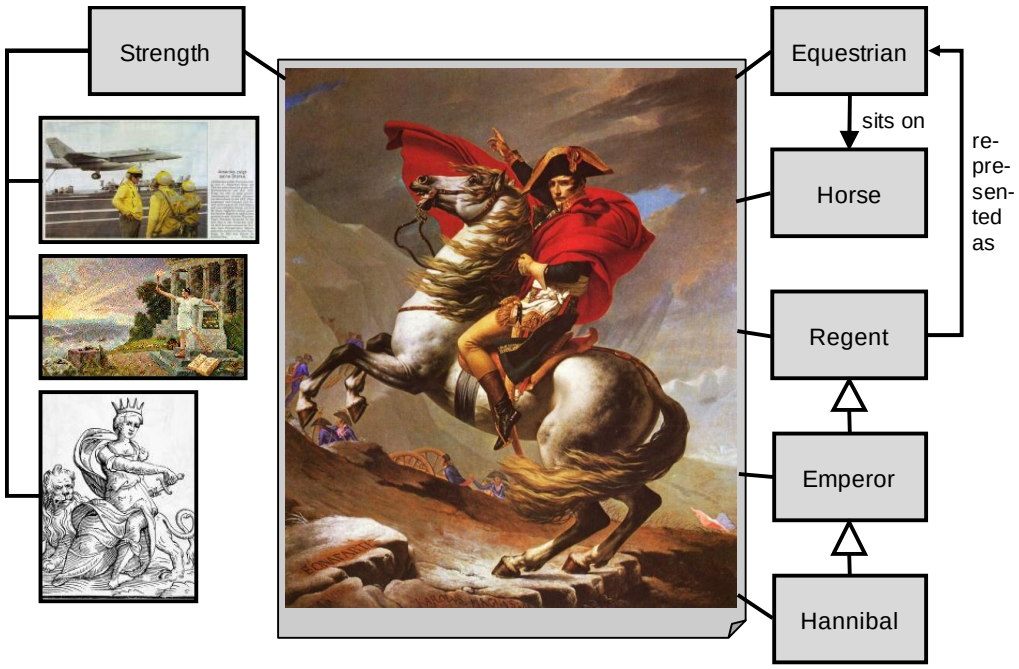

Figure 6. Intentional and extensional concept semantics in concept-oriented content management (CCM). Associations between images (Napoleon Crossing the Alps) and related concepts can be read in two ways. Sets of images define a concept, and concepts give meaning to an image.

Data analysis is performed for an *observer*, a user, or an algorithm that interprets the data. Such an observer will only perceive those data as meaningful for which there is a concept, and a concept will only be defined and used if there is sufficient evidence (in the form of data) for its existence. Which data actually constitute a sign depends on its observer, on the context the observer is in, and on the task at hand. An observer perceives the signs that are relevant with respect to those parameters. By the alternation of concept formulation and data interpretation, significant data are identified. It is used to build up hierarchies of signifiers to abstract from the concrete dataset and to carve out its content. The cycle of alternating concept creation and search for evidence for these concepts leads to relevance, and to the abovementioned urge to build minimal models. Only those concepts that are needed to describe instances and that cannot be expressed by other model elements will remain in the evolving model (*Occam's razor*). Overly complex concepts may be broken down into smaller sub-concepts if there is in turn enough evidence for these. Due to the dependency on observers' contexts and tasks, models are built in a pragmatic way. There will be no fixed syntax in the form of signifiers, and no fixed semantics given by concepts. Instead, the content found in a dataset will express a subjective view that is valid only for the context of the observer for whom it was defined.

Since both the identification of data that directly refer to a domain entity as well as the identification of data that describe (a set of) domain entities depend on an observer, there are coexisting views on the

same data. These views need to capture a wider range of variants than the external views known for databases, though, since they do not only consist of restrictions of one global conceptual model, but instead offer freedom in the structuring of data and in the formulation of models. On top of that, models change over time. This is due to the co-evolution of instances and models, so that all newly identified data and every new model element can potentially lead to model additions and changes. Furthermore, users may want to adapt existing models to different, but related contexts in order to reuse results. Therefore, whichever method is used to analyze big data, the results in the form of significant data and models need to be stored in a personalized way for one particular observer. Personalization has become a common means in content management, but on a different level. The modeling requirements formulated here call for both model and content personalization capabilities.

In contrast to the need to formulate personalized models, users need to be able to communicate using data. Such communication takes place at least over time, because data are created and analyzed at different points in time. Furthermore, producers and consumers of data may differ. Communication requires shared models, though. Pragmatic communication in the face of personalization can be achieved by personalized models that are related to each other. Extensive personalization on the one hand and related models on the other have extensively been studied in the CCM approach. In CCM, model relationships are established by models derived from each other, and by reusing models for base domains in different personalized models.

The two dimensions of model relationships are illustrated by Figure 7. It shows four model constellations. In the lower left there is a singular model $Model_1$. To the right, a constellation in which $Model_1$ incorporates (parts of) two other models: $Model_2$ and $Model_3$. The graph in the upper left shows $Model_1$ being refined by two personalized variants $Model_{11}$ and $Model_{12}$. The edges depict the personalization relationships. The upper right of Figure 7 shows combinations of the two model relationships. Current research investigates the *minimalistic meta modeling language* (*M3L*) [132,133] that exhibits properties that make it suitable for the management of *big data* analytics results as proposed in this section. The M3L supports signification on the various levels. It regards context as a major modeling principle, and it allows contextual concept definitions. This enables different forms of personalization, variants, and model relationships between them.

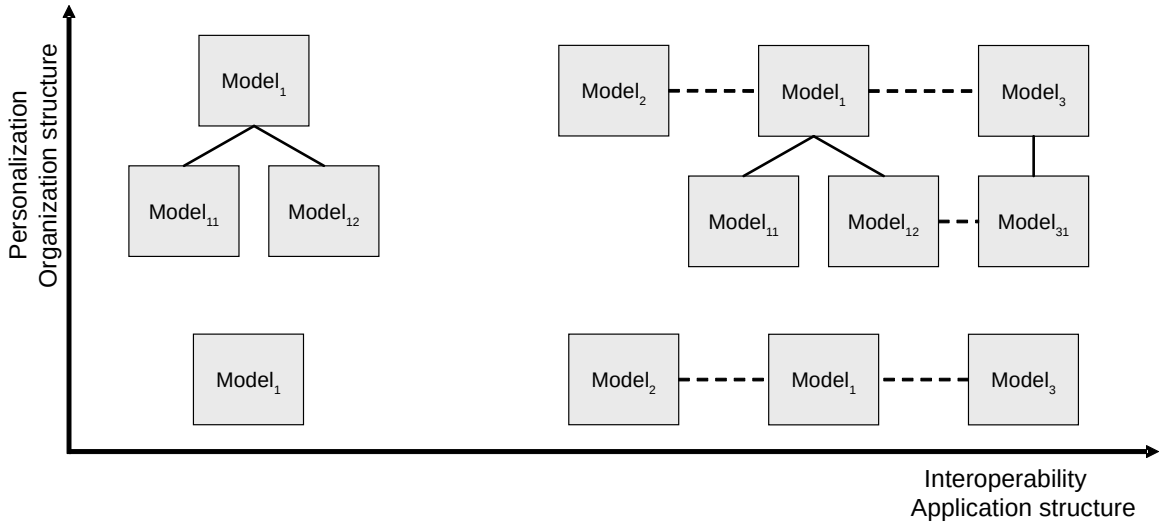

Figure 7. Dimensions of model relationships in CCM. Four model constellations that exploit model reuse and model personalization.

By its deductive properties based on contextual concept definitions and concept refinements, in conjunction with rules that allow both synthesizing concepts as well as matching instances, the M3L allows higher concepts that abstract from individual instances. Without introducing the M3L, we provide some examples to illustrate that. A concept like *Peter is a Person* is an icon for a person named *Peter* if the concept *Person* is given. Through a semantic rule that produces one concept from another, we can define indexical signification. For example, the definitions

Person {Sex; Status}

MarriedFemalePerson is a Person {Female is the Sex; Married is the Status} |= *Wife*

MarriedMalePerson is a Person {Male is the Sex; Married is the Status} |= *Husband*

establish an indexical relationship between *Married Female Person* and *Wife* (the former is directly resolved to the latter). The evaluation of M3L definitions allows deductions. For example, with the definitions

MalePeter is a Peter {Male is the Sex}

MarriedPeter is a Peter {Married is the Status},

the additional fact that

MarriedMalePeter is a MalePeter, a MarriedPeter.

results in the deduction that *Peter* is also a *Husband*.

Such M3L evaluations can be viewed as symbolic references. Each concept defines a context. All concepts in a context can be regarded as a model, so that contextual definitions allow establishing model relationships as, for example, in:

A: Napoleon Crossing The Alps

A is a Model {NapoleonCrossingTheAlps is a HannibalPicture}

B is a Model {NapoleonCrossingTheAlps is an EquestrianStatue}

A2 is an A {NapoleonCrossingTheAlps is an ImageOfStrength}

where the concept in *A2* is derived from the concept in *A* because the two models have a relationship (refinement, in this case, can be used for personalization). For model building, it is generally important to respect the boundaries of models with respect to the part of the application domain for which they are valid. The boundaries are required to judge which deductions can be applied to a dataset. It has to be noted that each deduction may lead to a pragmatic model change. Model boundaries are also relevant when transferring results across models along the relationships between these models.

To conclude, we can say that whatever means we use to analyze *big data*, it needs to be clarified how a given model is derived, which part of a domain it describes, and for what purpose. To this end, the means of management of the models and the information derived call for attention in addition to the investigation of methods for deriving models for *big data*sets. Such data management needs to be able to cope with model evolution as well as with coexisting models that allow different questions to be answered with the help of evidence contained in datasets.

6. Science and Culture from West to East or How We Relate Emotionally to *Big Data* and Artificial Intelligence

There has been an explosion in the amount of digital data since 2010 [134]. The amount of data is expected to reach 35 ZB in 2020. Over 95% of data were accumulated in this decade. Using this big amount of data has become practical owing to the advancement in digital technologies. We call it "*big data*" because it is a field that focuses on ways to analyze data, systematically extract information from them, or otherwise deal with datasets that are too large or complex to be dealt with by traditional data-processing software applications. However, this definition is fluid and can be understood in various ways. Thus, there is no clear definition for *big data.*

At the beginning of the first chapter of the famous book "BIG DATA", Viktor Mayer-Schoenberger and Kenneth Cukier [135] have taken up the Google Flu Trends (GFT). They studied how Google has mined a five-year log on the web, including hundreds of billions of searches on building an algorithm, and they claimed that "it is more effective than government statistics that cause delays in reporting", stating further that "we have built an influenza prediction model using 45 search words that have been proven to be a timely influenza index." Unfortunately, the GFT was not as effective as advertised. Shortly after the GFT was launched, the first problem occurred in 2009. It could not predict the swine flu epidemic at all. A report published in the Nature magazine in February 2013 observed that it predicted the influenza epidemic that occurred at the end of 2012 50% more severely than it actually was. Furthermore, the most inconvenient verification results since the launch of GFT were announced in March 2014. Media always emphasize the examples that work effectively. The apparent magic around "*big data*" is no exception here. However, there are many failures buttressing successful examples. Scientists have been continuously searching for theories of phenomena. Researchers tried to obtain a new theory that could explain the phenomena when no theory could explain it. These theories were based on causality. However, *big data* analysis does not have causality, and we have no golden tool for analyzing *big data*sets. There is no guarantee that a particular method of analysis that works successfully for one *big data*set will work for another. The reasons for this problem cannot be straightforwardly explained. Sometimes it works, and sometimes it does not and, therefore, dealing with *big data* is a big challenge.

Deep learning, very popular among data scientists, faces the same challenge. The problem is clearer in the case of deep learning than in that of *big data*. Although deep learning algorithms have many parameters, the principle of deep learning is always the same. Deep learning creates deep learning systems in minimizing cost functions by using a large amount of learning data. However, we do not know how to adjust such a complex system when it does not work well. The same challenge arises with *big data*, and the result is what matters most. The cause is largely unknown, but *big data* and deep learning may perform quite well, sometimes. We cannot discover the solution to cases where things do not work efficiently. Shall we call these new technologies science, or engineering? They are, in fact, a black box. Can we entrust our future to a black box? It is good when situations are convenient, but convenience often causes us to lose certain abilities. We forget how to read maps when GPS is practically applied everywhere. We have learnt how to calculate the square root a few decades ago, but presently, few people will be asked to give the square root with paper and pen, since anyone can easily find the square root of any number using a smart phone.

Big data and deep learning are tightly linked together [136,137]. As a consequence, this connection will be advanced further in various fields of study. When you fall sick in the future, the doctor might say "let us ask the black-box about your symptoms." How are we to feel about this? So-called technological singularity is an unavoidable issue when we discuss *big data* and deep learning [138]. The abilities of artificial intelligence (Al) may well exceed that of human beings by the year 2045. There is a high risk that about 47% of all U.S. jobs will be automated in the next 10 to 20 years [139]. Bill Gates argued that AI is a threat to humanity shortly after Microsoft's top announces that Al is not a threat to humanity. In addition, Stephen Hawking considered Al dangerous. However, there are different opinions regarding the current Al technology. AI could not pass the university entrance exam, even though it could win in a quiz show [140]. Deep learning, as the most common type of AI technology, is presently not real learning. However, having the name learning is somewhat superficial. The current technology cannot create intelligence beyond human capabilities. If that is the case, why should the singularity that AI surpasses human intelligence be discussed? Furthermore, why should technological singularity, where AI surpasses human intelligence, be discussed at all? We do not hear that technological singularity is considered a problem among experts in Japan. It is only the media and the non-specialist community of people that worry about technological singularity. Beginning with Kurzweil, who advocated it, people in science and technology

who are the keenest on discussing technological singularity as a challenge for humanity are, indeed, mostly Westerners.

In our modern times, we have accumulated a vast amount of scientific and technical knowledge, but our education has been strongly affected by our different histories, cultures, and religions. Western culture and thinking have been influenced mainly by Christianity, a monotheistic religion. In contrast, the Japanese religion, Shinto, is polytheistic like the religions of ancient Rome and Greece. Of course, we accept and understand religious differences. In 2016, the BBC compiled a list of the 21st Century's 100 greatest films [141], which included "Spirited Away," a Japanese animated film by Hayao Miyazaki. The film was made for young teenage girls. In the film, a girl's parents accidentally crashed a party held by the gods, who changed them into pigs. Consequently, they lost their memories and lived just like pigs. However, their daughter was able to restore them to their human forms. Such a story is very difficult for monotheists to understand, but the BBC chose "Spirited Away" as number four on their list of greatest films ever. This choice is a reflection of current popular understanding of religious differences, but individual thinking is still strongly influenced by our individual religions.

Back in the 19th century, science advanced dramatically and people believed that humans would be able to create other human beings in the future. The novel *Frankenstein* is an excellent example for illustrating this [142]. Although *Frankenstein* has been made famous through films, the original novel is quite different from the films, where Frankenstein's creation, an artificial but biological man, is depicted as very large, powerful and violent while having rather low intelligence. In contrast, the novel depicts him as intelligent enough to master multiple languages. However, he is rather ugly and disliked by people because of his looks. At the end of the novel, Dr. Frankenstein deservedly dies for having created an artificial human from body parts of dead people, implicitly considered a sin in Christianity, where only God is to create humans and where the dead are to be left in peace and not to be exploited for unholy purpose. When Westerners discuss attempts at creating other artificial forms of life, bearing in mind that all life is sacred in Western religions, they may indeed believe, consciously or unconsciously, that any such attempt is "sinful". Yet, such belief can be altered, i.e., effectively manipulated. Consider, for example, the visual appearance given to current products of "popular" artificial intelligence.

Figure 8 here above shows a cartoon character (on the left) named "Astro Boy" [143], who is a super robot that can fly, possesses the strength of one million horsepower, and understands all languages. Although "he" is an android, Westerners tend to think of "him" as *cute*, *cool*. Not many Westerners are likely to think of him as *eerie*, *scary*, or *weird*. In contrast, what do we feel when we see the female humanoid robot [144,145] in Figure 8 (on the right)? "She" too is an android, "her" name is Madoka Mirai, and "she" probably looks uncanny to many Westerners. Although both creations are androids, Astro Boy is mostly considered *cute*, whereas Madoka Mirai may be considered *eerie* or *scary* by most in the Western culture. What makes the difference between these two? An important one is almost certainly Madoka Mirai's close resemblance to a human being. Just as the creation of another human was considered a sin in Mary Shelley's *Frankenstein*, people in Western cultures may feel similarly when they see this humanoid robot. Eastern Asians, such as the Japanese, Chinese, and Koreans, do not feel upset about artificial life and humanoids at all. There is most certainly a close connection between religion, collective belief systems, and how people from different cultures feel about artificial intelligence (AI) and its products. Deep learning technology for *big data* mining inevitably involves AI. Currently, it is impossible to use this technology for creating new forms of AI that would surpass human intelligence. AI still can only respond to questions that already have their answers, but it cannot provide solutions to unanswered questions that we face in data science every day. Furthermore, possessing the ability to solve problems is completely different from possessing willpower, motivation to act, and, ultimately, consciousness. At present, AI cannot understand our culture. In the year 2045, how will AI think and feel about the androids shown in Figure 8?

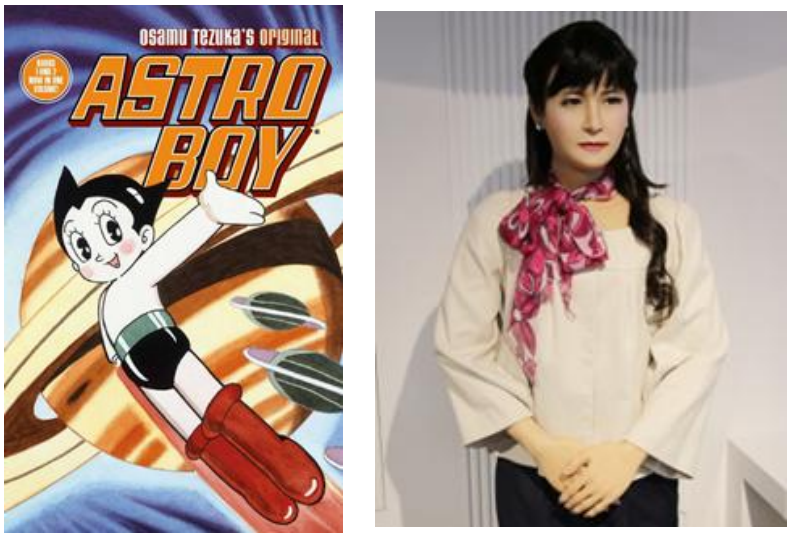

Figure 8. The robot Astro Boy (**left**) and the humanoid Madoka Mirai (**right**).

7. Conclusions

We still have limited understanding of how it will be possible to translate the *big data* potential into actual scientific, social, and economic value. The future trend in data science is announced as one towards fully automated data analytics, which implies artificial intelligence-based analysis and decision making in all fields [146], in other words: AI everywhere https://www.cioreview.com/news/artificial-intelligence-is-everywhere-nid-28010-cid-175.html.

This trend has raised high hopes in different fields of science for a better and faster understanding of structural complexity in living system [147–149]. The key argument defended in this review article here is not one against AI, but against any systematic replacement of the classical hypothesis-driven methods by new, increasingly complex and therefore difficult to control, data analytics that have not yet proved their worth. The different domain examples discussed in this review were aimed at illustrating some of the reasons why the generation of novel, data-driven hypotheses needs to be based on interpretable models. This will always and inevitably require validation and experimental testing. Complexity in analytical approaches does not necessarily outperform simplicity [150]. The currently emerging counter trend against fully automated *big data* analysis is that of exploratory data analysis http://www.creative-wisdom.com/teaching/551/Reading_materials/Yu_EDA_Oxford.pdf largely inspired by the philosophical and analytical work of John W. Tukey [151], who compared valid data (*big* or small) analysis to good detective work. This work should not be fully automated but may and will benefit from wisely and adequately developed artificial intelligence. Good detective work in data science has to

- Start by asking the right question(s);
- Look for the right clues;
- Draw the right conclusions from the clues available.

We are, thus, led to reconsider the somewhat inflexible principle of parsimony (*Occam's razor*) in traditional science within the context of contemporary data science, under a new and different light, where the advantages of simplicity have to be weighed against a seemingly growing need for complexity. This new context ultimately requires responsible decision making by domain experts, on a day-to-day basis.

Big data have not appeared all of a sudden out of nowhere but are a result of our cultural and technological development. They may be the ultimate consequence of the evolution of our species that will lead directly to our own extinction [152], programmed by ourselves, and taking final shape in fully

autonomous AI. Since we may not wish to be kept as pets by fully autonomous monsters (they may not even want to keep us as pets!), we may finally accept that the time for our extinction has come, with the planet's resources getting scarcer. Soon there may not be enough food and water for everyone everywhere on Earth. Fully autonomous AI does not need water, air, or food—we do. However, there is also hope. Fully autonomous AI will consume a lot of energy and we may realize that we should save it for better purpose. Also, modern data science is not evolving behind closed doors, or in a void. Analytic systems aimed at capturing what is deemed to make sense in all these unstructured data are designed by domain experts, and the procedures and algorithms these analytics use are based on scientific reasoning. They can be, and generally are, tested and refined through scientific investigation. They can also be reassessed in the light of progress. Inductive strategies for identifying patterns within data do not occur in a scientific vacuum, they are discursively framed by previous findings, theories, and speculations or intuitions grounded in experience and knowledge. The fear of a data science of the future, where insights and knowledge are automatically generated without asking the right questions, is therefore no more, but also no less, than a collective fantasy. As we have all learnt from history, collective fantasies can be dangerous. Yet, new analytics and algorithms most certainly arise scientifically, not arbitrarily, to be tested in the light of state-of-the-art expert domain knowledge. Ultimately, we may hope that this will be enough to ensure that only the methods and algorithms that have proven their true worth will survive the *big data* revolution. The latter was brought upon us directly by *our* cultural and technological development. Without it, there would be no need for paradigm shifts in science, or highly developed artificial intelligence. At present, science is still struggling to find clear general definitions and guidelines for key concepts directly related to *big data* ("data science", "deep learning", "artificial intelligence"), yet, these concepts are presumed to enable us to cope with the problem represented by *big data*. This somehow seems to require leaving the beaten tracks of science. The logic of scientific explanation according to William of Occam's *summa logicae* [4] requires that the nature of the *explanandum*, or what is to be explained, is adequately derived from the *explanans*, or explanation given. Considering the case of *big data*, these are no more than a particular expression of any current *explanandum*; *big data* neither go along consistently, nor systematically (i.e., predictably) with any current approach exploited to model them. If data science does not tread off the beaten tracks of traditional science very carefully, it may end up not seeing the forest for the trees in the *big data* jungle.

Funding: This research received no external funding.

Acknowledgments: We thank our colleagues from our research teams for their support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Cisco Global Cloud Index Methodology and Forecast 2016–2021*; Cisco Systems: San Jose, CA, USA, 2018**.**
2. Kitchin, R. Big data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*, 1–12.
3. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 23 June 2008. Available online: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed on 25 July 2019)**.**
4. Ockham, W. *Theory of Terms: Part 1 of the Summa Logicae*; University of Notre Dame Press: Notre Dame, IN, USA, 1974.
5. Seni, G.; Elder, J. *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions*; Morgan and Claypool: San Rafael, CA, USA, 2010.
6. Zikopoulos, P.C.; Eaton, C.; DeRoos, D.; Deutsch, T.; Lapis, G. *Understanding Big Data*; McGraw Hill: New York, NY, USA, 2012.

7. Sehring, W.; Kurz, M.; Fesl, J.; Ekseth, O.K.; Dresp-Langley, B.; Gohshi, S. On the perception of meaning in big data. In Proceedings of the 11 International Conference on Advanced Cognitive Technologies and Applications, Venice, Italy, 5–9 May 2019**.**

8. Liu, W.; Park, E.K. Big Data as an e-Health Service. In Proceedings of the International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 3–6 February 2014; IEEE: Piscataway, NJ, USA, 2014.

9. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286.

10. Sabharwal, S.; Gupta, S.; Thirunavukka, K. Insight of big data analytics in healthcare industry. In Proceedings of the International Conference on Computing, Communication and Automation (ICCCA), Noida, India, 29–30 April 2016.

11. Thanh, T.D.; Mohan, S.; Choi, E.; Kim, S.; Kim, P. A Taxonomy and Survey on Distributed File Systems. In Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management, Gyeongju, Korea, 2–4 September 2008.

12. Sarkar, D.; Rakesh, N.; Mishra, K.K. Content delivery networks: Insights and recent advancement. In Proceedings of the Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 22–24 December 2016.

13. Diogo, M.; Cabral, B.; Bernardino, J. Consistency Models of NoSQL Databases. *Future Internet* **2019**, *11*, 43.

14. Nereu, J.; Almeida, A.; Bernardino, J. Big Data Analytics: A Preliminary Study of Open Source Platforms. In Proceedings of the ICSOFT 2017-12th International Conference on Software Technologies, Madrid, Spain, 26–28 July 2017.

15. Dean, J.; Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. *Mag. Commun. ACM* **2008**, *51*, 107–113.

16. Hedjazi, M.A.; Kourbane, I.; Genc, Y.; Ali, B. A comparison of Hadoop, Spark and Storm for the task of large scale image classification. In Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018.

17. Gürcan, F.; Berigel, M. Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges. In Proceedings of the 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 19–21 October 2018.

18. Chacko, A.M.; Basheer, A.; Kumar, S.D. Capturing provenance for big data analytics done using SQL interface. In Proceedings of the IEEE Conference on Electrical Computer and Electronics (UPCON), Allahabad, India, 4–6 December 2015.

19. Srinivasulu, A.; Subbarao, C.D.V.; Kumar, J. High dimensional datasets using hadoop mahout machine learning algorithms. In Proceedings of the International Conference on Computing and Communication Technologies, Hyderabad, India, 11–13 December 2014.

20. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 1–10, doi.org/10.5334/dsj-2015-002.

21. Holder, L.B.; Haque, M.M.; Skinner, M.K. Machine learning for epigenetics and future medical applications. *Epigenetics* **2017**, *12*, 505–514, doi:10.1080/15592294.2017.1329068.

22. Lv, Q.; Qiao, Y.; Ansari, N.; Jun, L.; Yang, J. Big data driven hidden markov model based individual mobility prediction at points of interest. *IEEE Trans. Veh. Technol.* **2017**, *66*, 5204–5216.

23. Abdel-Basset, M.; Mohamed, M.; Smarandache, F.; Chang, V. Neurotrophic association rule mining algorithm for big data analysis. *Symmetry* **2018**, *10*, 106.

24. Feldman, D.; Schmidt, M.; Sohler, C. Turning big data into tiny data: Constant-size core sets for k-means, PCA, and projective clustering. *arXiv* **2018**, arxiv:1807.04518v1**.**

25. Kendall, M.G. *Rank Correlation Methods*; American Psychological Association: Washington, DC, USA, 1948.

26. Le Cun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *215*, 437, doi:10.1038/nature14539

27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates: Red Hook, NY, USA, 2012.

28. Pliner, H.A.; Shendure, J.; Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *bioRxiv* **2019**, doi:10.1101/538652
29. Smith, A.M.; Walsh, J.R.; Long, J.; Davis, C.B.; Henstock, P.; Hodge, M.R.; Maciejewski, M.; Mu, X.J.; Ra, S.; Zhang, S.; et al. Deep learning of representations for transcriptomics-based phenotype prediction. *bioRxiv* **2019**, doi:10.1101/574723
30. Wenliang, L.K.; Seitz, A.R. Deep Neural Networks for Modeling Visual Perceptual Learning. *J. Neurosci.* **2018**, *38*, 1617–1620.
31. Kohonen, T. Analysis of a simple self-organizing process. *Biol. Cybern.* **1982**, *44*, 135e140.
32. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59e69.
33. Kohonen, T. *Self-Organizing Maps*; Springer: Berlin/Heidelberg, Germany, 2001.
34. Binder, H.; Hopp, L.; Lembcke, K.; Löffler-Wirth, H. Personalized Disease Phenotypes from Massive OMICs Data. In *Artificial Intelligence*: *Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2017, doi:10.4018/978-1-5225-1759-7.ch019.
35. Clauset, A.; Moore, C.; Newman, M.E. Hierarchical structure and the prediction of missing links in networks. *Nature* **2008**, *453*, 98–101.
36. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.
37. Pelleg, D.; Moore, A. Accelerating exact k-means algorithms with geometric reasoning. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; pp. 277–281.
38. Pelleg, D.; Moore, A.W. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*; Carnegie Mellon University: Pittsburgh, PA, USA, 2000; Volume 1, pp. 727–734.
39. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA; pp. 1027–1035.
40. Yau, C.; žurauskienė, J. Pcareduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* **2016**, *17*, 140.
41. Bezdek, J.C.; Pal, N.R. Some new indexes of cluster validity. *IEEE Trans. Syst. Man Cybern. Part B* **1998**, *28*, 301–315.
42. Gasch, A.P.; Eisen, M.B. Exploring the conditional co-regulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* **2002**, *3*, 0059-1.
43. Yeung, K.Y.; Ruzzo, W.L. Principal component analysis for clustering gene expression data. *Bioinformatics* **2001**, *17(9)*, 763–774.
44. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.
45. Vendramin, L.; Campello, R.J.; Hruschka, E.R. On the comparison of relative clustering validity criteria. In Proceedings of the SIAM International Conference on Data Mining, Calgary, AL, Canada, 2–4 May 2019; pp. 733–744.
46. Kockara, S.; Mete, M.; Chen, B.; Aydin, K. Analysis of density based and fuzzy c-means clustering methods on lesion border extraction in dermoscopy images. *BMC Bioinform.* **2010**, *11*, 26.
47. Otair, M. Approximate k-nearest neighbor based spatial clustering using kd tree. *Int. J. Database Manag. Syst.* **2013**, *5*, 97.
48. Sibson, R. Slink: An optimally efficient algorithm for the single-link cluster method. *Comput. J.* **1973**, *16*, 30–34.
49. Ekseth, O.K.; Hvasshovd, S.-O. How an optimized DBSCAN implementation reduces execution-time and memory-requirements for large data-sets. In Proceedings of the Patterns 2019, Barcelona, Spain, 18–22 February 2018; Department of Computer Science (IDI), NTNU: Trondheim, Norway, 2018; pp. 6–11.
50. Ekseth, O.K.; Hvasshovd, S.O. An empirical study of strategies boosts performance of mutual information similarity. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=2ahUKEwjFm4Xp4tLjAhWj3OAKHTJ4DpEQFjACegQIBBAB&url=https%3A%2F%2Farxiv.org%2Fabs%2F1201.0490&usg=AOvVaw3sRE_hzE6HK9082UEqmdeIZakopane, Poland, 16–20 June 2019; pp. 321–332.

51. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the KDD-96, Portland, Oregon, 2–4 August 1996; Institute for ComputerS cience, University of Munich: Miinchen, Germany, 1996; Volume 96, pp. 226–231.
52. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall Incorporated: Upper Saddle River, NJ, USA, 1988.
53. Ole Kristian Ekseth. Hplysis: A High-Performance Software Library for Big-Data Machine-Learning. Available online: https://bitbucket.org/oekseth/hplysis-cluster-analysis-software/ (accessed on 6 June 2017).
54. Changyong, F.; Hongyue, W.; Naiji, L.; Tian, C.; Hua, H.; Ying, L. Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **2014**, *26*, 105.
55. Qiu, X.; Wu, H.; Hu, R. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinform.* **2013**, *14*, 124.
56. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications* **2015**, *19(1)*, 29-33.
57. Wu, J.; Xiong, H.; Chen, J. Adapting the right measures for k-means clustering. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 877–886.
58. Guo, M.; Wang, H.; Potter, S.S.; Whitsett, J.A.; Xu, Y. Sincera: A pipeline for single-cell rna-seq profiling analysis. *PLoS Comput. Biol.* **2015**, *11*, 1004575.
59. Mazandu, G.K.; Chimusa, E.R.; Mulder, N.J. Gene ontology semantic similarity tools: Survey on features and challenges for biological knowledge discovery. *Brief. Bioinform.* **2016**, *18*, 886–901.
60. Jain, A.K. Data clustering: 50 years beyond k-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666.
61. Ben-Hur, A.; Guyon, I. Detecting stable clusters using principal component analysis. In *Functional Genomics: Methods and Protocols*; Humana Press: New York, NY, USA, 2003; pp. 159–182.
62. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **2007**, *52*, 258–271.
63. Lawson, D.J.; Falush, D. *Similarity Matrices and Clustering Algorithms for Population Identification Using Genetic Data*; Department of Mathematics, University of Bristol: Bristol, UK, 2012.
64. Valafar, F. Pattern recognition techniques in microarray data analysis. *Ann. N. Y. Acad. Sci.* **2002**, *980*, 41–64.
65. Patra, B.K.; Nandi, S.; Viswanath, P. A distance based clustering method for arbitrary shaped clusters in large datasets. *Pattern Recognit.* **2011**, *44*, 2862–2870.
66. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics.* **2001** *17(4)*, 309-18. PMID: 11301299.
67. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
68. Mallick, P.K.; Mihir, N.M.; Kumar, S. White Patch Detection in Brain MRI Image Using Evolutionary Clustering Algorithm. In *Research Advances in the Integration of Big Data and Smart Computing*; IGI Global: Hershey, PA, USA, 2016; pp. 323–339.
69. Kim, K.; Hyunchul A. Recommender systems using cluster-indexing collaborative filtering and social data analytics, *International Journal of Production Research* **2017**, *55(17)*, 5037-5049. 10.1080/00207543.2017.1287443
70. Marung, U.; Nipon, T.; Auephanwiriyakul, S. Top-N recommender systems using genetic algorithm-based visual-clustering methods. *Symmetry* **2016**, *8*, 54.
71. Kapil, S.; Meenu, C.; Ansari, M.D. On K-means data clustering algorithm with genetic algorithm. In Proceedings of the Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 22–24 December 2016.
72. Kim, K.; Hyunchul, A. A recommender system using GA K-means clustering in an online shopping market. *Expert Syst. Appl.* **2008**, *34*, 1200–1209.
73. Ekseth, O.K.; Gribbestad, M.; Hvasshovd, S.O. Inventing wheels: Why improvements to established cluster algorithms fails to catch the wheel. In *DISP—FCA Handbook*; St. Huge College: Oxford, UK, 2019.
74. Evolutionary Algorithms. In *Encyclopedia of Machine Learning*, Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2011.

75. Hakrabarti, D.; Kumar, R.; Tomkins, A. Evolutionary clustering. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006.

76. Hua, C.; Li, F.; Zhang, C.; Yang, J.; Wu, W. A Genetic XK-Means Algorithm with Empty Cluster Reassignment. *Symmetry* **2019**, *11*, 744.

77. Ahmed, M.N.; Yamany, S.M.; Mohamed, N.; Farag, A.A.; Moriarty, T. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Trans. Med. Imaging* **2002**, *21*, 193–199.

78. Yang, Y. Image segmentation based on fuzzy clustering with neighborhood information. *Opt. Appl.* **2009**, *39*, 136–146**.**

79. Legacy Documentation. Fuzzy Logic. 2004. Available online: https://reference.wolfram.com/legacy/applications/fuzzylogic/TOC.html (accessed on 25 July 2019).

80. Cowgill, M.C.; Harvey, R.J.; Watson, L.T. A genetic algorithm approach to cluster analysis. *Comput. Math. Appl.* **1999**, *37*, 99–108.

81. De Hoon, M.J.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20*, 1453–1454.

82. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A.; Charrad, M.M. Package 'nbclust'. *J. Stat. Softw.* **2014**, *61*, 1–36.

83. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

84. Narain, A.; Chandra, D.; Singh, R.K. Model order reduction using Fuzzy C-Means clustering. *Trans. Inst. Meas. Control.* **2014**, *36*, 992–998, doi:10.1177/0142331214528968.

85. Sakthi, M.; Thanamani, A.S. An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA. *Int. J. Comput. Sci. Inf. Technol.* **2011**, *2*, 955–959.

86. Kohonen, T. Automatic formation of topological maps of patterns in a self-organizing system. In Proceedings of the 2nd Scandinavian Conference on Image Analysis, Espoo, Finland, 15–17 June 1981; pp. 214–220.

87. Kohonen. MATLAB Implementations and Applications of the Self-Organizing Map. 2014. Available online: http://docs.unigrafia.fi/publications/kohonen_teuvo/MATLAB_implementations_and_applications_of_the_self_organizing_map.pdf (accessed on 25 July 2019).

88. Hubel, D.H.; Wiesel, T.N. Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.* **1959**, *148*, 574–591.

89. Hubel, D.H. Integrative processes in central visual pathways of the cat. *J. Opt. Soc. Am.* **1963**, *53*, 58–66.

90. Hubel, D.H.; Wiesel, T.N. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **1965**, *28*, 229–289.

91. Hubel, D.H.; Wiesel, T.N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **1968**, *195*, 215–243.

92. Dresp, B. The effect of practice on the visual detection of near-threshold lines. *Spat. Vis.* **1998**, *11*, 1–13.

93. Dresp, B. Do positional thresholds define a critical boundary in long-range detection facilitation with co-linear lines? *Spat. Vis.* **2000**, *13*, 343–357.

94. Fischer, S.; Dresp, B. A neural network for long-range contour diffusion by visual cortex. *Lect. Notes Comput. Sci.* **2000**, *1811*, 336–342.

95. Dresp, B.; Fischer, S. Asymmetrical contrast effects induced by luminance and colour configurations. *Percept. Psychophys.* **2000**, *63*, 1262–1270.

96. Tzvetanov, T.; Dresp, B. Short- and long-range effects in line contrast detection. *Vis. Res.* **2002**, *42*, 2493–2498.

97. Spillmann, L.; Dresp-Langley, B.; Tseng, C.H. Beyond the classic receptive field: The effect of contextual stimuli. *J. Vis.* **2015**, *15*, 7.

98. Dresp, B.; Grossberg, S. Contour integration across polarities and spatial gaps: From local contrast filtering to global grouping. *Vis. Res.* **1997**, *37*, 913–924.

99. Carandini, M.; Demb, J.B.; Mante, V.; Tolhurst, D.J.; Dan, Y.; Olshausen, B.A.; Gallant, J.L.; Rust, N.C. Do we know what the early visual system does? *J. Neurosci.* **2005**, *25*, 10577–10597.

100. Kapadia, M.K.; Westheimer, G.; Gilbert, C.D. Spatial contribution of contextual interactions in primary visual cortex and in visual perception. *J. Neurophysiol.* **2000**, *84*, 2048–2062.

101. David, J.A.S.; Green, M. *Signal Detection Theory and Psychophysics*; John Wiley & Sons Inc.: New York, NY, USA, 1966.
102. Wandeto, J.M.; Dresp-Langley, B. Ultrafast automatic classification of SEM image sets showing CD4 + cells with varying extent of HIV virion infection. *7ièmes Journées de la Fédération de Médecine Translationnelle de Strasbourg*, Strasbourg, France, 25–26 May 2019.
103. Wandeto, J.M.; Nyongesa, H.; Remond, Y.; Dresp-Langley, B. Detection of smallest changes in medical and random-dot images comparing self-organizing map performance to human detection. *Inform. Med. Unlocked* **2017**, *7*, 39–45.
104. Wandeto, J.M.; Nyongesa, H.; Remond, Y.; Dresp-Langley, B. Detection of smallest changes in medical and random-dot images comparing self-organizing map performance and expert performance. In Proceedings of the European Conference on Visual Perception (ECVP), Berlin, Germany, 27–31 August 2017.
105. Dresp-Langley, B.; Wandeto, J.M.; Nyongesa, H.K.O. Using the quantization error from Self-Organizing Map output for fast detection of critical variations in image time series. In *ISTE OpenScience, Collection from Data to Decisions*; Wiley & Sons: London, UK, 2018.
106. Dresp-Langley, B. Principles of perceptual grouping: Implications for image-guided surgery. *Front. Psychol.* **2015**, *6*, 1565.
107. Wandeto, J.M.; Dresp-Langley, B. The quantization error in a Self-Organizing Map as a contrast and color specific indicator of single-pixel change in large random patterns. *Neural Netw.* **2019**, in press**.**
108. Kerekes, J.P.; Baum, J.E. Spectral Imaging System Analytical Model for Subpixel Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1088–1101.
109. Bashivan, P.; Kohitij, K.; DiCarlo, J.J. Neural population control via deep image synthesis. *Science* **2019**, *364*, eaav9436, doi:10.1126/science.aav9436.
110. Weiser, M. The Computer for the 21 st Century. *Sci. Am.* **1991**, *265*, 94–105.
111. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **2013**, *20*, 1645–1660.
112. Hashem, I.A.T.; Chang, V.; Anuar, N.B.; Adewole, K.; Yaqoob, I.; Gani, A.; Ahmed, E.; Chiroma, H. The role of big data in smart city. *Int. J. Inf. Manag.* **2016**, *36*, 748–758.
113. Kurz, M.; Ferscha, A. Sensor abstractions for opportunistic activity and context recognition systems. In *European Conference on Smart Sensing and Context*; Springer: Berlin/Heidelberg, Germany, 2010.
114. Marx, V. Biology: The Big Challenges of Big Data. *Nature* **2013**, *498*, 255–260.
115. Weiser, M.; Gold, R. The origins of ubiquitous computing research at PARC in the late 1980s. *IBM Syst. J.* **1999**, *38*, 693–696.
116. Akter, S.; Wamba, S.F. Big data analytics in E-commerce: A systematic review and agenda for future research. *Electron. Mark.* **2016**, *26*, 173–194.
117. Helbing, D.; Frey, S.; Gigerenzer, G.; Hafen, E.; Hagner, M.; Hofstetter, Y.; van den Hoven, J.; Zicari, R.; Zwitter, A. Will Democracy Survive Big Data and Artificial Intelligence? *Towards Digit. Enlight.* **2019**, 73–98, doi.org/10.1007/978-3-319-90869-4_7.
118. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
119. Schaller, R.R. Moore's law: Past, present and future. *IEEE Spectr.* **1997**, *34*, 52–59.
120. Frangi, A.F.; Tsaftaris, S.A.; Prince, J.L. Simulation and synthesis in medical imaging. *IEEE Trans. Med. Imaging* **2018**, *37*, 673.
121. Brodie, M.L.; Schmidt, J.W. Final Report of the ANSI/X3/SPARC DBS-SG Relational Database Task Group. *ACM Sigmod Rec.* **1982**, *12*, 1–62.
122. Schmitz-Rigal, C. *Die Kunst offenen Wissens, Ernst Cassirers Epistemologie und Deutung der modernen Physik, Cassirer- Forschungen*; Ernst Meiner Verlag: Hamburg, Germany, 2002; Volume 7.
123. Peirce, C.S. *Collected Papers of Charles Sanders Peirce*; Harvard University Press: Cambridge, MA, USA, 1931.
124. Firstness, Secondness, and Thirdness in Peirce|Semiotics and Visual Communication. Available online: https://undcomm504.wordpress.com/2013/02/24/firstness-secondness-and-thirdness-in-peirce/ (accessed on 7 June 2019).

125. Sehring, H.-W.; Schmidt, J.W. Beyond Databases: An Asset Language for Conceptual Content Management. In Proceedings of the 8th East European Conference on Advances in Databases and Information Systems, Budapest, Hungary, 22–25 September 2004; Benczúr, A., Demetrovics, J., Gottlob. G., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 99–112.
126. Bergman, M. A Foundational Mindset: Firstness, Secondness, Thirdness. Available online: http://www.mkbergman.com/1932/a-foundational-mindset-firstness-secondness-thirdness/ (accessed on 7 June 2019).
127. Schmidt, J.W.; Sehring, H.-W. Conceptual Content Modeling and Management. In *Perspectives of Systems Informatics, 5th International Andrei Ershov Memorial Conference, Novosibirsk, Russia, 9–12 July 2003*; Broy, M., Zamulin, A.V.; Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 469–493.
128. Schmidt, J.W.; Sehring, H.-W.; Warnke, M. Der Bildindex zur Politischen Ikonographie in der Warburg Electronic Library-Einsichten eines interdisziplinären Projektes. In *Archivprozesse. Die Kommunikation der Aufbewahrung*; Pompe, H., Scholz, L., Eds.; DuMont Television Network: Cologne, Germany, 2002; pp. 238–268.
129. Cassirer, E. *Die Sprache, Band 11 Philosophie der Symbolischen Formen der Reihe Gesammelte Werke*; Felix Meiner Verlag GmbH: Hamburg, Germany, 2001.
130. Cassirer, E. *Das mythische Denken, Band 12 Philosophie der symbolischen Formen der Reihe Gesammelte Werke*; Felix Meiner Verlag GmbH: Hamburg, Germany, 2002.
131. Cassirer, E. *Phänomenologie der Erkenntnis, Band 13 Philosophie der symbolischen Formen der Reihe Gesammelte Werke*; Felix Meiner Verlag GmbH: Hamburg, Germany, 2002.
132. Sehring, H.-W. Content Modeling Based on Concepts in Contexts. In Proceedings of the Third International Conference on Creative Content Technologies, Rome, Italy, 25–30 September 2011; Sehring, H.-W., Fohl, W., Eds.; ThinkMind(TM) Digital Library: Venice, Italy, 2011; pp. 18–23.
133. Sehring, H.-W. Context-aware Storage and Retrieval of Digital Content: Database Model and Schema Considerations for Content Persistence. *Int. J. Adv. Softw.* **2018**, *11*, 311–322.
134. IDC. The Digital Universe Decade? Are You Ready? Available online: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf (accessed on 25 July 2019).
135. Mayer-Schoenberger, V.; Cukier, K. *BIG DATA.*; Eamon Dolan, Mariner; Houghton Mifflin Harcourt; New York, March 2014, reprint.
136. Sutton, R.S.; Barto, A.G. Introduction to Reinforcement Learning. 1998. Available online: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf (accessed on 25 July 2019).
137. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F. Traffic Flow Prediction with Big Data: A Deep Learning Approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873.
138. Kurzweil claims the singularity will happen. Available online: https://www.kurzweilai.net/futurism-ray-kurzweil-claims-singularity-will-happen-by-2045 (accessed on 25 July 2019).
139. Frey, C.; Osborne, M. The Future of Employment: How Susceptible are Jobs to Computerization? Available online: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (accessed on 25 July 2019).
140. What we gain from the digitalization of medical care. Available online: https://www.nii.ac.jp/en/about/upload/NIIToday_en66.pdf (accessed on 25 July 2019).
141. The 21st century's 100 greatest films. Available online: http://www.bbc.com/culture/story/20160819-the-21st-centurys-100-greatest-films (accessed on 25 July 2019).
142. Shelley, M. *Frankenstein*, Simon and Schuster Digital, available online: https://www.simonandschuster.com/getliterary/ (accessed on 25 July 2019).
143. AI Robot Astroboy. Available online: https://www.youtube.com/watch?v=XPRVj1T1zgs (accessed on 25 July 2019).
144. Madoka Mirai: Creepily human-like robots. Available online: https://www.cbsnews.com/pictures/creepily-human-like-robots-2/12/ (accessed on 25 July 2019).

145. Japan trends: Kyoto temple creates android version of Buddhist Goddess of Mercy. Available online: https://www.japantrends.com/kyoto-temple-android-robot-buddhist-goddess-mercy-kannon/ (accessed on 25 July 2019).
146. Li, D.; Du, Y. *Artificial Intelligence with Uncertainty*; CRC Press: Boca Raton, FL, USA, 2017.
147. Günther, W.A.; Rezazade Mehrizi, M.H.; Huysman, M.; Feldberg, F. Debating big data: A literature review on realizing value from big data. *J. Strateg. Inf. Syst.* **2017**, *26*, 191–209, doi:10.1016/j.jsis.2017.07.00.
148. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med.* **2019**, *6*, 34, doi:10.3389/fmed.2019.00034.
149. Ohno, N.; Katoh, M.; Saitoh, Y.; Saitoh, S. Recent advancement in the challenges to connectomics. *Microscopy* **2016**, *65*, 97–107, doi:10.1093/jmicro/dfv371.
150. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 2nd ed.; Morgan Kaufmann: Waltham, MA, USA, 2011.
151. Tukey, J.W. Exploratory data analysis. *Methods* **1977**, *2*, 131–160, reprinted for sale in 2019.
152. Orr, D.W. Armageddon versus extinction. *Conserv. Biol.* **2005**, *19*, 290–292.
e, reference material.