(Big) Data and algorithms: Looking for meaningful patterns

Taina Bucher

In the 1670s the Dutch businessman and scientist Antonie van Leeuwenhoek discovered the hitherto unknown microscopic world. Using his handcrafted microscope, he was the first to observe and describe bacteria and other microorganisms. Equipped with a lifetime of microscopic experimentation and technical refinement, making over 500 optical lenses and creating 25 single-lens microscopes, van Leeuwenhoek is not just considered the father of microbiology but a pioneer in revealing the unseen world using his self-made microscopes (Lane, 2015). Fast forward to our own day and age and there is an entirely different unseen world that engages scientists and business people alike. The future, or more precisely the prediction of what is to come based on what is and has been, is the unseen that people want to discover today. If the men and women of the Golden Age of Dutch science and technology discovered the unseen world through microscopes, the world today is increasingly 'discovered' through large datasets and predictive analytics[1].

Let's stay in the world of biology for just a little while longer. This time, not bacteria but data. Not miniscule worlds but large ones. Not microscopes but Google. When launching Google Flu Trends in 2008 the public health tracking system was hailed as a potential new innovation in epidemiology. By mining the millions of search queries of web users, the idea was that the flu tracker would be able to estimate flu activity even before the official Centres for Disease Control and Prevention (CDC) had the chance to register any outbreaks. Traditional flu monitoring, overseen by the CDA in the case of the US, depends on national networks of physicians reporting patient cases with Influenza-like symptoms. However, to classify as sick, patients first have to be diagnosed by a doctor and by then it is often too late, they are already sick and have already had plenty of opportunity to infect others. At first, Google Flu Trends seemed remarkably successful. The predictive models created based on CDC data from between 2003 and 2007, proved to be "consistently one to two weeks ahead of the CDC surveillance reports" (Butler, 2008). Yet, when Google Flu Trends provided estimates double that of the CDC during the 2012 flu season, the wonders of big data started to fade (Butler, 2013). As Cheney-Lippold writes, it wasn't that Google had "missed the forest for the trees. It missed the sick tree for all the other trees who'd been

[1] I borrow the opening story from an article published in The Atlantic: see Brynjolfsson, E. & Mcafee, A. (2011).

frantically Googling ways to help the sick tree" (2017: 123. No technique of observation and measurement, whether microscopic nor macroscopic, is ever immune to bias or failure.

While new technical terrains often provide new ways of understanding and measuring the world, as in the above cases, revealing the worlds of the previously unseen or unmeasurable, these techniques are only half the story. In this chapter on data and algorithms, it is therefore important to highlight a technical as well as historical, cultural, political and economic understanding of the 'datafied' and algorithmically constructed present. One of the biggest truisms about big data is the "end of theory"-thesis famously articulated by former Wired editor Chris Anderson in 2008. According to Anderson, we now live in a 'Petabyte Age', or as Barnes puts it: "an age of ten to the power of 15, binary 2 to the power of 50, bytes" (2013: 298). The point is not just that the petabyte age is big, but different "because more is different" (Anderson, 2008). More is supposedly different because we no longer need to hypothesize what things mean, just follow the data. Despite this lingering techno-optimism of the 'big data revolution', which is particularly evident in the business world and tech industry, much critical scholarship on big data and algorithms have consistently scrutinized such overly simplified notions of 'data as the new oil' (see for example, Amoore & Piotukh, 2015; boyd & Crawford, 2012; Crawford et al., 2014; Kitchin, 2014; Neff et al., 2017). Indeed, as Hargittai suggests in relation to the methodological challenges of using big data, "bigger is not always better; size is not all that matters when it comes to datasets" (2015: 74). There is more to big data than its petabyte. As Crawford et al. argue, big data does not constitute the end of theory; it *is* theory (2014: 1664). To understand how big data has become a *Weltanschauung* as Crawford et al. (2014) put it, some necessary grounds will have to be covered first.

This chapter begins by defining some of the key terms, including data and big data, before moving on to historical and technical background in terms of databases, the rise of statistical society and what Hacking (1990) has termed an "avalanche of numbers". Next, the chapter considers the broader 'datafication' of society, understood as the "process of rendering into data aspects of the world not previously quantified" (Kennedy et al. 2015: 1). The second part of the chapter moves more specifically into the terrain of 'algorithms', providing some definitional clarity to key terms and contextual understanding in terms of exemplifying how algorithms (and data, big data and all the related terms) need to be seen as part and parcel of larger sociotechnical systems and assemblages.

Data: given, taken and made

When asked to write about data and algorithms in 2018, there is an implicit understanding of relating the discussion to the ways in which data have assumed such a significant role in society today with the advent of very large datasets – more commonly described under the banner of "big data" – and its supporting and enabling technologies. A decade ago, writing about data may have meant something quite different. That said, a decade ago there would probably not have been a chapter on data and algorithms in a media and communications handbook. This does not mean, however, that data or algorithms are anything new. Data are basic forms of enumeration and encodings of the phenomena they represent or describe (Barocas et al. 2014). What is new is the scale and proliferation of these enumerated and encoded phenomena that are increasingly used to drive and support all kinds of decision-making processes in society. But let's not get ahead of ourselves. Why were data and algorithms not taken-for-granted concepts in media and communications a decade ago when these concepts have been around for longer than the discipline has existed? After all, the term 'data' has been part of the English language at least since Antonie van Leeuwenhoek's time. While relatively new in its rhetorical and discursive significance, data understood more narrowly as the classification and quantification of observations, need to be understood in a much longer historical context of quantification, documentation and archiving. The question of why data, now, may be the same as for other disciplines. Before the advent of what is now commonly referred to as "big data", the word data did not assert itself in the same way as it does today. Today, we have become accustomed to data driving something, as in data-driven businesses and data-driven organizations. Data have taken on a more active role as cheaper and easier to use technologies support both the collection and scalability of data in new ways. Add to this the fact that most dominant media outlets, including social media platform, the entertainment industry and news organizations, are increasingly relying on data-driven and algorithmically processed insights, and the (renewed) relevance of these terms for media and communication scholars should be quite evident.

Not only has the past decade seen the explosive rise of terms previously tucked down in a computer science textbook or statistical bureau, but the related terms have multiplied as well (at least in regards to the scholarly discussions on them). We have data and big data, but we also have social media data, open data, personal data, structured data, unstructured data, small data, thick data, primary data, secondary data, metadata, mundane data, log data and so forth. All of these terms connote different kinds of data, with their own histories, questions and problems, only some of which will be discussed as part of this chapter. Etymologically the word data is derived from the Latin dare, meaning 'to give'. In general use, however, "data refer to those elements that *are* taken; extracted through observations, computations, experiments, and record keeping" (Kitchin, 2014: 2). Moreover, data are always *made*. That is, as Barocas et al. point out, "they are artifacts of human intervention, not facts imparted by nature itself" (2014:2). As Helles (2013) exemplifies, "the Web server log file that we find online does not become data before we begin to conceptualize it within the context of a research project", or the context of a business model for that matter. As such, data are representative in nature as they provide information on certain aspects of the phenomena we are interested in studying or knowing. As Kitchin points out, data need not be explicit in its representative nature, but can also be "implied (e.g., through an absence rather than presence) or derived (e.g., data that are produced from other data, such as percentage change over time calculated by comparing data from two time periods)" (2014: 1).

Ultimately, data needs to be processed, analysed and made sense of. Whether it is made sense of by humans or machines, or most likely, a combination of both, the process of making data always involves multiple agents (Helles & Jensen, 2013). Data are never simply raw nor do they exist in vacuum, but are stored, recorded, collected, processed, analysed and employed by a complex ecosystem of users, digital infrastructure, databases, businesses, public and private institutions, algorithms, policy makers, and governments alike. Researchers, for instance, use data to advance the state of knowledge. They may rely on primary data that they have themselves collected, secondary data others have made available to them, or tertiary data, which are derived forms of data that also include someone else's interpretations, such as statistical results. Authorities too, not only use, but depend on data, including hospitals, which patient records are essential to their service, schools that keep track of their students' performance, and government agencies that meticulously record information about their populations, most notably through the Census Bureau responsible for producing data about a nation's people and economy. While the data points these authorities use and collect may vary, the idea is the same: store, process and assemble, it will be useful for making decisions.

Database technologies

In order for data to be useful, they need to be organized and kept in databases. While data can be lost, forgotten or simply overlooked, most data thought valuable are usually collected and classified so that it later can be retrieved for analytical or informational purposes. Understanding database technologies, both analogue and digital, is therefore essential if we want to understand how data is made available for analysis. In the most basic understanding of the term, a phone book could be a database, because it provides a structured list of a few data points such as name and associated phone numbers. A database, broadly conceived, is a record-keeping and information-retrieval system. Its origins predate the computer, going back to libraries, archives, and other government, business, and medical record-keeping. In this broad definition, books, libraries, and archives can be conceived of as databases: they provide a way to store and maintain data. Through a description of its own structure, a database also provides the means for finding and retrieving the data it contains. Books contain a table of contents and, in many cases, an index at the back; libraries include a catalog that is organized according to a specific classification system; and archives likewise depend on indexical and other systems of organization. More specifically, the term database is most commonly used to describe how computers store, manage, and organize data. A database is a collection of data that is encoded and arranged according to a common format. The term is also used interchangeably about systems that manage collections of data and about the tools and techniques that support the manipulation and operation of these data.

In the context of computing, the term database is used more narrowly to describe how computers store, describe, and organize data. Here we might distinguish between three levels at which the term database is used (Dourish, 2014). At the most general level, database is used to merely denote a collection of data. More specifically, database refers to a collection of data that is encoded and arranged according to a common format. This common format, importantly, makes data amenable to a common set of operations, including sorting, comparing, and processing the data in consistent and reliable ways. At a third level, database refers to software management systems that implement the relationship between data formats and data. At this level, database is often used interchangeably with the systems that manage collections of data (e.g., Oracle) and the tools and techniques that support the manipulation and operation of these data (e.g., SQL).

Importantly, databases are constructed artifacts that are designed to “hold certain kinds of data and enable certain kinds of analysis, and how they are structured has profound consequences as to what queries and analysis can be performed” (Kitchin, 2014: 21-22; Ruppert, 2012). In the case of relational databases, which is the still the most common way of digitally storing and structuring data, data are organized into one or more tables of columns and rows, each with a uniquely identifiable key. Relations between tables are established on the basis of their interactions. Pioneered by E. F. Codd in the early 1970s, relational databases became a de facto standard for digital storage and retrieval when the development of the Structured Query Language (SQL), an English-like syntax for interacting with a relational database, enabled easier management of the data contained in the relational database. Although relational databases have been around for more than 40 years, their position has changed as new database models have done away with the tabular schema. With the steady increase in available data and web services with greater workloads, there have been new demands for data storage and processing. As a result, new kinds of data models and database management systems have evolved, collectively known as post-relational or NoSQL databases. It should be pointed out, however, that these terms do not refer to “a single implementation or conceptual model, but rather to a range of designs that in different ways responds to problems with the relational model” (Dourish, 2017: 123). These databases are typically used to store and retrieve data from Web server logs and social media platforms. Unlike relational databases, which can mainly cope with **structured data** (i.e. data that is easily organized and stored in a defined data model), NoSQL databases are useful for operating on **unstructured data** (i.e. data that does not have a pre-defined data model or is not organized in a pre-defined model) too, as they do not require that fields be specified in advance. In NoSQL databases, Kitchin points out, “data are typically distributed and replicated across many machines rather than centralised into one location”.

This move into vast data territories and the development of new storage and processing technologies in parallel is precisely what some scholars have identified as the key characteristic of the big data age, understood as the “transformation in what can be collected or sampled as data, and how it can be rendered analysable” (Amoore & Piotukh 2015:345). On the one hand, Amoore & Piotukh suggest that the big in big data refers to the notion that big data pushes at the limits of traditional relational databases, and on the other hand, which is also the more common understanding of big data, the data is considered big because “it exceeds and changes human capacities to read and make sense of it” (2015: 343). The shift, from relational databases designed for structured data, to post-relational databases built for the capacity to hold unstructured data implies an expansion of the kinds of data forms that can be parsed and detaches analysis from a specific index to allow for analysis to be deterritorialized and conducted across jurisdictions (Amoore & Piotukh, 2015). The advent, then, of post-relational databases that hinge on distributed processing, but also on new important hardware changes in processor designs and improved memory as Dourish points out (2017), have contributed to the fact that "we can now collect information that we couldn't before, be it relationships revealed by phone calls or sentiments unveiled through tweets" (Mayer-Schoenberger & Cukier 2013: 30). In other words, understanding the technical changes in both software and hardware is essential for an understanding of the ideological, political and social grounds of datafication, the idea of harnessing (big) data and algorithms to analyse social behaviour.

Datafication

The exponential growth in available data generated from user interactions in online systems has not only led to more data being collected and stored, but also to what scholars have termed *datafication*, the "widespread *belief* in the objective quantification and potential tracking of all kinds of human behaviour and sociality through online media technologies" (van Dijck, 2014: 198). As Kennedy et al. suggest, "datafication refers to the process of rendering into data aspects of the world not previously quantified" (2015: 1). Central to this (re)newed belief in the power of quantification is a type of data, which, as van Dijck suggests, "not too long ago was considered worthless by-products of platform-mediated services" (2014: 199). **Metadata,** or data about data, is essential to the utilization of big data. These are the kind of data that provide additional and associative information to whatever data point one is interested in. In the case of a single email, for example, the metadata provides information about who the receiver and sender is, the time and date of the message, the length and amount of words contained in the message and so on. Or, when scrolling down your Facebook feed, pretty much every additional information is being logged, from when you log in, which device you are using, what you click on, location, through to the duration of your activities (Facebook Data Policy, 2016). While these kinds of data may not be as interesting in and of themselves, in the aggregate, however, the patterns generated with the help of metadata may be invaluable.

The widespread belief in datafication has been embraced by a number of institutions, private and public, businesses and researchers alike, where data and metadata are commonly treated as traces of human behaviour, or so-called digital footprints. The availability and ease of collecting huge datasets has led to a rush of collecting data for its own sake, oftentimes without a clear purpose in sight. The amount of business and trade press literature on big data is simply overwhelming. The general advice seems to be somewhere along the lines of 'collect as much as possible, even the things you don't think are useful, and worry about analysis later', thinking that more information is always better. A common critique levelled at social media research, for example, is that it has privileged the study of Twitter, simply because of its publicly available data. As Lomborg writes, "research too often gets seduced by the sheer availability and abundance of data" (2017: 7), while overlooking or turning a blind eye to the messiness and unrepresentativeness of the data collected. Researcher, however, are not alone in being seduced by the abundance in data. As boyd and Crawford argue, there is a "deep government and industrial drive toward gathering and extracting maximal value from data, be it information that will lead to more targeted advertising, product design, traffic planning, or criminal policing" (2012: 675).

In one of the first critical assessments of the term big data in media and communication studies, boyd and Crawford define **big data** as "a cultural, technological, and scholarly phenomenon that rests on the interplay of: Technology, Analysis and Mythology" (2012). Big data, the authors suggest, is no more a technical phenomenon than it is a social and epistemic one. It changes not just how we might collect and analyse data but how we think about objects of knowledge in and of themselves. According to Anderson's 'Petabyte vision' we no longer start from theories or prior knowledge, data will generate it for us. But do numbers speak for themselves, boyd and Crawford rhetorically ask, warning that data will lose its meaning and value if we lose sight of its context (2012: 670). Context means knowing more about the kinds of data that are being generated, who gets to access data and to what end data is deployed. Context means understanding not just the possibilities but, more importantly, the limitations of big data analytics. To evoke the Google Flu Trends experiment, context means understanding that the device was "better at using browser data to trace the spread of worries about the symptoms of flu than it was at predicting the spread of the virus itself" (Halford and Savage, 2017: 3). It would, however, be a mistake to assume that big data does not *have* context. As Seaver nicely puts it, "the nice thing about context is that everyone has it" (2015). Drawing on fieldwork amongst data scientists and developers working on music recommender systems, Seaver points out, how practitioners are very much geared towards the question of context, as knowing more about individual users is key to providing personalized content and recommendations. Neff et al. further point out, how data, whether big or small, is "always already context-rich because of how people imagine data and construct, produce, or define the dataset " (2017: 89).

If boyd and Crawford paved the way for a critical discourse on big data with their six provocative questions about the meaning and governance of data, many more questions and concerns have since been added to the list by scholars and practitioners alike. Some of the recurring issues have to do with the significance of big data for governments (Rieder & Simon, 2016), the health sector (Rückenstein & Dow Schüll, 2017), surveillance (Lyon, 2014), privacy and personal integrity (Crawford & Schultz, 2014) to name but a few. Discussions around ethics and methods have been prevailing too, with scholars advocating for ethical data sharing practices (Zook et al., 2017; Zwitter, 2014), attending to the specificities of digital devices themselves (Ruppert et al., 2013) and for more *data-activist research practices* (Milan & van der Velden, 2016). As so-called big data has come of age, there is also a growing need to account for the concept in historical and sociological terms (Beer, 2016). As mechanisms of quantification, classification, measurement and prediction, data and algorithms are as much imbued in the history of computation and software engineering as they are in the history of statistics, accounting, and bureaucratization. As such, the historical and cultural contexts of the big data era intersect with the social history of calculation and ordering of various types, including the history and politics of statistical reasoning and large numbers (Desrosieres & Naish, 2002; Foucault, 2007; Hacking, 2006; Power, 2004), practices of quantification, numbering and valuation (Callon & Law, 2005; Espeland & Stevens, 2008; Verran, 2001), the cultural logic of rankings and ratings (Espeland & Sauder, 2007; Sauder & Espeland, 2009), and ideas of critical accounting and auditing (Power, 1999; Strathern, 2000).

The question is not just, as Beer puts is, "how we should do the history of big data" (2016), but also to recognize, as Barnes suggests, there is no single history but rather a "conjuncture of different elements, each with their own history, coming together at this our present moment" (2013: 298). While many scholars have rightfully focused on the lineage of calculation, statistics and numbers when accounting for the history of big data, the specifics matter. The "avalanche of numbers" (Hacking, 1991, 2015), which occurred as nation-states started to classify and count their populations in the 19th century certainly forms a general backdrop for an understanding of what is at stake today. But in order to arrive at the present moment, we must also acknowledge the complex and disjunctive route it takes to get there, via, but not limited to, the social history of census, punch cards, bureaucratization, wartime machinery, the rise of computers, automated management of populations, biopolitics, machine learning techniques, and so much more.

Data subjects

Consider this recent advertisement for Spotify's Premium subscription model distributed through various social media platforms during Christmas 2017. Above a bright coloured background, the message reads quite simply: "Data has feelings too. Hold it for longer with offline listening Premium". Against the background, the ad mimics a New Year's resolution, encouraging consumers to "Spend more time with your data". Not only does this ad anthropomorphize data, showing how music is not necessarily the most important part of the streaming service's business model. It also shows how data has become part of the social imaginary. The datafication of society has made data mundane in the sense that people not only image data in particular ways but also that data has become part of how people image their social existence through software-mediated practices of consumption. The fact that Spotify can run an ad campaign framing data as a friend that needs attention and care is only possible because data has become part of people's everyday practices and contexts.

Data are not part of people's lives; people also actively make data on a daily basis. Online services and social media platforms do no longer produce content through the educated guesses of expert individuals trained in trend forecasting, gut feelings based on decades of experience from the industry or academic educations in film theory or musicology. Not only, at least. Today content production is supported and driven by the explicit and implicit emission of user-generated data. Companies like Spotify, Netflix, Facebook, and Google provide information and recommendations based on what they think we want, predictions derived by aggregated user data. As **data subjects** (Ruppert, 2011) humans have in a sense themselves *become* data. People's actions and interactions with online services serve as inputs for the construction of personal profiles, or what Cheney-Lippold (2017) calls 'measurable types'. Whether we are speaking of "soccer moms in Florida that are really passionate about action films" or "female college educated Scandinavian who listens to hip hop and jazz”, measurable types are used to classify and filter what we get to see online. As Cheney-Lippold argues, traditional categories like gender are never absolute, you are never just ‘male’ or ‘female.’ Rather, based on statistical confidence and probability, you might be 92 percent confidently ‘male’ and 32 percent confidently ‘female’ (2017: 34). Based on further inputs, such as clicks, purchase behaviour and other actions, these measures may subsequently either rise or fall. In other words, while you may be 92 percent confidently male today, tomorrow the confidence score may have dropped to 70 percent. Thus, the data subject is generated through a malleable and changing ‘algorithmic identity’ (Cheney-Lippold, 2011), emerging in and through data.

Such is the work of ‘profiling machines’ (Elmer, 2004) that seek to produce a sense of identity through detailed consumer profiles, which are geared towards anticipating future needs. Based on statistical inferences and inductive reasoning, profiling algorithms do “not necessarily have any rational grounds and can lead to irrational stereotyping and discrimination” (de Vries, 2010: 80). Still, the question of whether ‘correct’ identifications are being constructed may be beside the point. As de Vries (2010) argues, misidentification is not simply a mismatch or something that should be considered inappropriate. Misidentifications may also give some leeway for thinking about how identity construction is experienced. Experiencing algorithmic landscapes is as much about what the algorithm does in terms of making certain connections as it is about people’s personal engagements. A particular landscape, the anthropologist Ingold (1993) suggests, owes its character to the experiences it affords to the ones that spend time there - to their observations and expectations. In my research on how people encounter algorithms online, several participants reported the limits to algorithmic identity construction (Bucher, 2017). One of the interviewees, who identified as transgender, described how she felt there was no obvious space for her in Amazon’s purchasing recommendations. Either there were suggestions for makeup or power tools, but never anything in between. To Amazon you are still either male or female although the degree to which you may be one or the other may differ. As a person in transition, she felt her queer subject position became too much. Amazon seemed willing to try and categorize people according to fluid demographic buckets, just not the ones that might endanger their profits and risk offending someone. More than simply describing strange feelings, experiences like these describe some of the many, mundane moments in which people variously encounter the algorithmic realities and principles underlying contemporary media platforms.

While Spotify wants us to believe that “data have feelings”, making sense of how people have feelings for data is important too (Kennedy & Hill, 2017). If we turn to the phenomenon of self-tracking, understood as an individual’s use of technology to record, monitor and reflect upon features of daily life (Lomborg & Frandsen, 2016), we may see how data “only make sense in the context in which people decide to collect their data and the social relationships and expectations, places and spaces in which they do so” (Lupton, 2017: 10). Whether someone is monitoring his or her heart rate or keeping track of calories burnt during physical exercise, these metrics can only tell limited details about the body. Without interpretation and additional contextual information, the result of these digital tracking devices means very little. As Lupton point out, “when people review their data, they actively relate them to the contexts in which they were generated. People consider such aspects as the time of day, the weather, how their bodies felt, whether they were lacking sleep, were hungry”, etc. (2017: 11). Whether we are talking about large technical systems such as the Google Flu tracker or individual’s fitness trackers, big data or small data, what we cannot lose sight of in our datafied age is the importance of interpretation and the need to contextualize data in everyday practices.

Algorithms: Making sense of data

While the significant power and potential of big data (the quantity of information produced by people, things, and their interactions) cannot be denied, its value derives not so much from the data itself but from the ways in which it has been brought together into new forms of meaningfulness by the associational infrastructure of the respective software systems in which algorithms play a key role.[2] In the standard computer science understanding of the term, an algorithm refers to a set of instructions for solving a problem or completing a task following a carefully planned sequential order. Perhaps, the most common way to define an algorithm is to describe it as a recipe, understood as a step-by-step guide that prescribes how to obtain a certain goal, given specific parameters. Understood as a procedure or method for processing data, the algorithm as recipe would be analogous to the operational logic for making a cake out of flour, water and eggs. Without the specific instructions for *how* to mix the eggs and flour or *when* to add the sugar or water, for instance, these ingredients would remain just that. For someone who has never baked a cake, step-by-step instructions would be pivotal if they wanted to bake one. For any computational process to

[2] Portions of the section on algorithms are adapted from Bucher 2018.

be operational, the algorithm must be rigorously defined, that is, specified in such a way that it applies in all possible circumstances. A program will execute a certain section of code only if certain conditions are met. Otherwise, it takes an alternative route, which implies that particular future circumstances are already anticipated by the conditional construct of the ’if…then statement’ upon which most algorithms depend.

Programmers usually control the flow by specifying certain procedures and parameters through a programming language. In principle, the algorithm is “independent of programming languages and independent of the machines that execute the programs” (Goffey, 2008:15). The same type of instructions can be written in the languages C, C#, or Python, and still be the same algorithm. This makes the concept of the ‘algorithm’ particularly powerful, given that what an algorithm signifies is an inherent assumption in all software design about order, sequence, and sorting. The actual steps are what is important, not the wording *per se*. Designing an algorithm to perform a certain task implies a simplification of the problem at hand. From an engineering perspective, the specific operation of an algorithm depends largely on technical considerations, including efficiency, processing time, and reduction of memory load, but also on the elegance of the code written (Fuller, 2008; Knuth, 1984). The operation of algorithms depends on a variety of other elements - most fundamentally, on **data structures**. To be actually operational, algorithms work in tandem not only with data structures but also with a whole assemblage of elements, including data types, databases, compilers, hardware, CPU and so forth.

An important distinction needs to be made between algorithms that are pre-programmed and behave more or less deterministically and algorithms that have the ability to “learn” or improve in performance over time. Given a particular input, a deterministic algorithm will always produce the same output by passing through the same sequence of steps. The learning kind, however, will learn to predict outputs based on previous examples of relationships between input data and outputs. Unlike a deterministic algorithm that correctly sorts an alphabetized list, many of the algorithms that run the Internet today do not necessarily have one easily definable, correct result. The kinds of algorithms and techniques to which I am referring here are called **machine learning**, which is essentially the notion that we can now program a computer to learn by itself (Domingos, 2015). In contrast to the strict logical rules of traditional programming, machine learning is about writing programs that learn to solve the problem from examples. Whereas a programmer previously had to write all the ‘if…then’ statements in anticipation of an outcome herself, machine learning algorithms lets the computer learn the rules from a large number of training examples without being explicitly programmed to do so. In order to help reach a target goal, algorithms are ‘trained’ on a **corpus** of data from which they may ‘learn’ to make certain kinds of decisions without human oversight.

Just like rule-based algorithms, machine learning algorithms come in many different flavors. Similar to humans, the machine itself learns in different ways. One of the most common ways in which algorithms learn is called **supervised learning**. Essentially an inductive approach to learning, algorithms are given a training set comprising the characteristics that engineers want the algorithm to detect and compare with new data (Flach, 2012). Importantly, the training set includes data about the desired output. When the training data does *not* include data about desired outputs, the approach is called **unsupervised learning**. Often, machine learning algorithms may fall somewhere in between: the data only contains a few desired outputs, which is also called semi-supervised learning (Domingos, 2015). Before an algorithm can be applied to learn from data, **models** have to be constructed that formalize the task and goals, so that it can be processed by a computer. For instance, before an algorithm can perform the task of finding the most important news feed stories, models have to be created to represent the relationship between news and relevance.

In data-intensive environments such as social media, machine learning algorithms have become a standard way of learning to recognize patterns in the data, to discover knowledge, and to predict the likelihood of user actions and tastes. Another way to put this is to say that machine learning is largely enabled by proliferating data from which models may learn. In the age of so-called ‘big data’, having the biggest pool of data available from which to detect patterns is often seen as a competitive necessity. The bigger the database, the better the conditions for algorithms to detect relevant patterns. Commercial application of machine learning is commonly called **data mining**, which basically refers to the routinized and automated processes of discovering patterns from models (Barocas et al., 2014: 6). Though data mining has become somewhat of a contemporary buzzword, the concept has been around for over 25 years, pioneered by IBM research fellow Rakesh Agrawal in a paper demonstrating the utility of consumer algorithms coupled with retail data (Agrawal et al., 1993). However, the world has moved way beyond analyzing patterns in

Marks & Spencer's retail data as was the case with the aforementioned research paper. Today, data mining has become a standardized way of collecting and saving traces of human activity, whether these data subjects are consumers, citizens, criminals or 'users'. What is important is that we do not lose sight of what data mining is for, or whom and in what situations. Beyond simply collecting, storing and analyzing data, the critical task is to interrogate the purposes and processes of data mining in a broader perspective. As van Dijck asks: "why do we look for certain patterns in piles of metadata, in whose interests, and for what purposes?" (2014: 202). It is when data mining becomes an argument for even more data mining - more data to make borders secure or more data to make more effective business decisions – that we need to be particularly alert (Vaidhyanathan, 2017). As we confront the world of increased enumeration, quantification and prediction, there is also a need to ask what the possibilities are to remain invisible, silent, indeed, undiscoverable? Notwithstanding the men and women of the Golden Age of Dutch science and technology or the epidemiologists of Google, sometimes the unseen should remain just that.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases.* Paper presented at the Acm sigmod record.

Amoore, L., & Piotukh, V. (2015). Life beyond big data: Governing with little analytics. *Economy and Society, 44*(3), 341-366.

Barocas, S., Rosenblat, A., boyd, d., Gangadharan, S. P., & Yu, C. (2014). *Data & Civil Rights: Technology Primer*. Retrieved from Data & Society: http://www.datacivilrights.org/

Beer, D. (2016). How should we do the history of Big Data? *Big Data & Society, 3*(1), 1-10. Doi: 2053951716646135

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Brynjolfsson, E. & Mcafee, A. (2011) "The big data boom is the innovation story of our time". *The Atlantic*, retrieved from https://www.theatlantic.com/business/archive/2011/11/the-big-data-boom-is-the-innovation-story-of-our-time/248215

Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, communication & society, 20*(1), 30-44.

Bucher, T. (2018). *IF…THEN: Algorithmic power and politics*. New York: Oxford University Press.

Butler, D. (2008). Web data predict flu. *Nature, 456*, 287-288.

Butler, D. (2013). When Google got flu wrong. *Nature, 494*(7436), 155.

Callon, M., & Law, J. (2005). On qualculation, agency, and otherness. *Environment and Planning D: society and space, 23*(5), 717-733.

Cheney-Lippold, J. (2017). *We Are Data: Algorithms and The Making of Our Digital Selves*: NYU Press.

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication (19328036), 8.*

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev., 55*, 93.

De Vries, K. (2010). Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information technology, 12*(1), 71-85.

Desrosières, A., & Naish, C. (2002). *The politics of large numbers: A history of statistical reasoning*: Harvard University Press.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Dourish, P. (2014). NoSQL: The Shifting Materialities of Database Technology. *Computational Culture*(4).

Dourish, P. (2017). *The Stuff of Bits: An Essay on the Materialities of Information*: MIT Press.

Elmer, G. (2004). Profiling Machines: Cambridge, MA: MIT Press.

Espeland, W. N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds1. *American journal of sociology, 113*(1), 1-40.

Facebook. (2016). Data Policy.   Retrieved from https://www.facebook.com/full_data_use_policy

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*: Cambridge University Press.

Foucault, M. (2007). *Security, Territory, Population*. New York: Palgrave Macmillan.

Fuller, M. (2008). *Software studies: A lexicon*. Cambridge, Mass.: MIT Press.

Goffey, A. (2008). Algorithm. In M. Fuller (Ed.), *Software Studies: A Lexicon*. Cambridge, Mass.: MIT Press.

Hacking, I. (1990). *The taming of chance*. Cambridge, UK: Cambridge University Press.

Hacking, I. (1991). How should we do the history of statistics?” i Burchell, Graham; Colin Gordon & Peter Miller: The Foucault Effect–Studies in Governmentality. *London: Harvester/Wheatsheaf.*

Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*: Cambridge University Press.

Halford, S., & Savage, M. (2017). Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research. *Sociology*, 0038038517698639.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 63-76.

Helles, R. (2013). The big head and the long tail: An illustration of explanatory strategies for big data Internet studies. *First Monday, 18*(10).

Helles, R., & Jensen, K. B. (2013). Introduction to the special issue' Making data-Big data and beyond'. *First Monday, 18*(10).

Ingold, T. (1993). The temporality of the landscape. *World archaeology, 25*(2), 152-174.

Kennedy, H., & Hill, R. L. (2017). The Feeling of Numbers: emotions in everyday engagements with data and their visualisation. *Sociology*, 0038038516674675.

Kennedy, H., Poell, T., & van Dijck, J. (2015). Introduction: Special issue on Data and agency. *Data & Society*.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*: Sage.

Knuth, D. E. (1984). Literate programming. *The Computer Journal, 27*(2), 97-111.

Lane, N. (2015). The unseen world: reflections on Leeuwenhoek (1677) ‘Concerning little animals’. *Phil. Trans. R. Soc. B, 370*(1666), 20140344.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203-1205.

Lomborg, S., & Frandsen, K. (2016). Self-tracking as communication. *Information, communication & society, 19*(7), 1015-1027.

Lupton, D. (2017). Data Thing-Power: How Do Personal Digital Data Come to Matter?

Lyon, D. (2014). Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data & Society, 1*(2), 2053951714541861.

Milan, S., & Velden, L. v. d. (2016). The alternative epistemologies of data activism. *Digital Culture & Society, 2*(2), 57-74.

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data, 5*(2), 85-97.

Power, M. (2004). Counting, control and calculation: Reflections on measuring and management. *Human relations, 57*(6), 765-783.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society, 3*(1), 2053951716649398.

Ruckenstein, M., & Schüll, N. D. (2017). The Datafication of Health. *Annual Review of Anthropology, 46*, 261-278.

Ruppert, E. (2011). Population objects: Interpassive subjects. *Sociology, 45*(2), 218-233.

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society, 30*(4), 22-46.

Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American sociological review, 74*(1), 63-82.

Seaver, N. (2015). The nice thing about context is that everyone has it. *Media, Culture & Society, 37*(7), 1101-1109.

Vaidhyanathan, S. (2017). The Incomplete Political Economy of Social Media. *The SAGE Handbook of Social Media*, 213.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197.

Verran, H. (2001). *Science and an African logic*. Chicago: University of Chicago Press.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., . . . Narayanan, A. (2017). Ten simple rules for responsible big data research. *PLoS computational biology, 13*(3), e1005399.

Zwitter, A. (2014). Big data ethics. *Big Data & Society, 1*(2), 2053951714559253.