Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "Making sense of data." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 41–62. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-3>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 09:27 UTC.

Copyright © Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy 2022. Released under a CC BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/3.0/>). You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

Making sense of data

‘What do the data say?’ is one of those questions we might hear more and more often in the offices of executives and data analysts, who aim to make their decisions and develop strategies based on the digital data collected by corporations and through online media. After a second thought, we might answer: ‘Well, data don’t say anything; they don’t speak; they don’t have a mouth, and they don’t deliver the interpretations on which decisions and strategies are based.’ It seems to be an odd feature of our times that data have been bestowed with the status of an oracle who tells us what to do and how to proceed. In this popular view, and in the discursive practices discussed in Chapter 2, data are seen as ‘facts’, prepared in numerical or textual form, which seem to bear the quality of being authentic and original, objective, pre-interpretive and impartial, guaranteeing value-free description. In contrast to them, the interpretations which are based on them, the sense and meaning we attribute to them, and the narratives we create out of data analysis seem to result from a secondary, additive process which pollutes the authenticity of the original data, making them somehow ‘false’ and thus ‘distorting’ an accurate representation of reality. As one person interviewed by the KPLEX project put it:

I would say that narrative is much more like, that’s completely fake. Like, any sort of model or meaning or you know representation or any of these kind of stories that we come up with, they’re not an accurate representation of reality, for sure. And as much as possible, people try to use data to back it up, to show that their narrative, their representation of the world is correct.

(WP2 INT6)

The interpretation and narratives produced by humans on the basis of these data imply, in this view, the danger of deformation and misrepresentation, the addition of a potentially false layer, the misdirection of perception and the pollution by the human stain of meaning-making: ‘I think the narrative

comes after an interpretation of data. So, it's like you look into the data, you give yourself some kind of idea about what's going on. So, it's like yeah you're adding basically, an interpretation, and that of course can be false' (WP2 INT2). But how have we arrived at such a staunch opposition of data and narratives, of 'facts' and 'interpretation'? What kind of facticity is being seen in data, especially in big data, in contrast to the meaning transported by narratives? Which traits and functions differentiate the use of data and narratives? What is the surplus value of interpretation and narrativization, produced by humans, thus laying bare qualities which show that data don't render storytelling obsolete?

Interpretation in the humanities: Two examples

How we use data (and big data) to come to understand the world is not universal or natural, it is a social construction and a cultural process, engaging deeply with values and beliefs as well as language practices. And so, in order to answer these questions, it is necessary to turn away from data science and look at scientific approaches where the fabrication of meaning and the creation of interpretive narratives are part of the daily business and central to the social functions of these disciplines. In the humanities, the methodological approaches of history and literature researchers developed and handed down along generations of scientific practice start from the close inspection and comprehensive scrutiny of data. More often than not, the term 'data' is not used here; rather, historians speak of 'sources' and literature scholars of 'primary material' or 'primary literature'. The use of these terms reflects on the one hand the long-standing traditions of both disciplines and point back to times when the term 'data' was uncommon (as the discussion of this issue in the last chapter has shown). On the other hand, 'data' designates something already processed, while 'sources' or 'primary literature' point to the original, authentic and unaltered material, such as can be found in archives or libraries. Furthermore, such disciplines have developed comprehensive methodologies and prescriptions on how to deal with the materials under scrutiny. In historical research, so-called auxiliary sciences have been established which frame the examination of a whole range of different classes of sources, such as handwritten material in general ('palaeography'), more specifically documents, records, and proceedings or coats of arms, seals and coins. The inspection of these sources always comprises a thorough discussion of source provenance, a reconstruction of the context in which such a source was

produced, a critical assessment of the reliability of the source and its explanatory power for a given research question.

Humanistic research processes, in contrast to data-driven ones, are often perceived to be subjective, lacking rigour, or even emotion-driven, but they are nothing of the sort. Historians traditionally visit archives, consult finding aids in order to determine which sources would be valuable for their research project and which are not, and discuss with archivists whether or not there are further sources which might be relevant for their research. They gather research literature in order to determine which sources have already been investigated and thus be part of their scientific endeavour or not. In a next step, historians perform a critical assessment of these sources and take notes on them, a process which often implies the inspection of hundreds or even thousands of different sources. Gaining an overview of the multitude of available sources allows the development of a research hypothesis which is then fostered by at times multiple re-readings of sources, selecting some sources while dismissing others. This comprehensive process of knowledge creation is often reflected in the final product – the research article or monograph – where it may be documented why certain sources have been used while others were not deemed valuable for the analysis. Since historical sources are mostly unique copies and cannot be created *ex post* (in contrast to the process of data collection), the scarcity of these sources often leads to the phenomenon that the same sources are re-read by different generations of historians. Think of historical evidence from antiquity, where the availability of material is limited and the discovery of new sources, for example, through archaeological excavations, forms the exception rather than the norm. These multiple re-readings of sources reveal a peculiarity common to the humanities, namely the phenomenon that there is no exhaustive interpretation of any source; rather, they can be consulted again and again by different generations of historians re-visiting these sources working on different research questions with a different focus, be it on constructions of gender in earlier times, on manifestations of everyday life, on culturally different forms of handling disputes, on conceptions of space and time, the networks between people or the like. This stands in contrast to the data practices of an organization like CERN, where data volumes can be so large that throwing data away is a normal practice. In history, however, the process of knowledge creation is regarded as never-ending: any particular time will create new questions to be answered by the scrutiny of historical sources.

In the study of literature, the focus is not on unique sources, but on printed books, that is, on technically reproduced material. In traditional literary research,

studies most often concentrate on a single work or on the work of a single author. Similar to historical research, the methodology is hermeneutical: a given work is read and re-read multiple times, gathering evidence from the text(s) according to the research question, contextualizing it by cultural and biographical sources or secondary literature, alternative readings of the text and complementary interpretations found in studies published earlier. Since the focus is more on contemporary culture or cultural heritage than on the reconstruction of history, the interdependencies between the text or passages from the text and the surrounding cultural horizons are explored. The methodology, hermeneutics, can be understood as a circumference through subjective comprehension. It is an iterative, self-reflexive process, going back and forth between preconception and understanding, ultimately aiming at interpretation. As such this practice seeks to create not so much a 'black box' as a prismatic one, through which light refracts in different waves. Humanists focus on the exemplary and are not necessarily interested in generalization and exact quantities. A humanist can develop her comprehension of individual authors; her comprehension of the relationship between these single authors and the intellectual contexts they live in; and her comprehension of these intellectual contexts, that is, the larger environment of individuals, the discursive structures or cultural patterns prevailing in the period which is being researched. This methodological process, summarized in the conception of the hermeneutic circle, includes reflections on the conditions determining the structures and patterns which can be found in the text. Moreover, it entails considerations of the limitations imposed by the text – what it can tell and what is not contained – and on the limited range of insights offered by the method. Literary scholars are aware of the social constructedness of meaning in literary products and of the processes of sensemaking peculiar to literature, of the true nature of their data, as it were, as *capta*, which are brought into dialogue with the cultural context. As such, hermeneutics is opposed to positivism, which is seen to be narrowed to the statements which can be found in a text and which disregards context or regards it as *quantité négligeable*.

In both of these disciplines the hallmarks of the humanities are clearly visible: hermeneutic, critical and speculative thinking, thorough evaluation of the sources, thick description capturing the context, profound understanding of the particular, attentiveness towards meaningful differences.¹ The studies presented by historians and literary scholars ultimately aim at spinning a narrative web that conveys the interpretation elaborated by the researcher. These narratives contain what has been acquired along the way up to the writing of the study: precise observation of details, distinction between relevant and irrelevant information

(or 'signal' and 'noise'), complex syntheses, long chains of argumentation, the handling of alterity and particularity. It is not contested that these narratives are the ultimate result of the subjective choices and decisions by the individual researcher, and it is a commonplace that the studies presented by history or literature scholars are subjective: they are themselves placed in the context of their creation, and they entail the use of rhetorical devices and persuasive strategies. However, there is a paradox here in that such studies are nonetheless objective. In order to better understand the subjectivity-objectivity paradox of the humanities, one of the most famous sentences ever written in modern historiography can be cited. In 1885, the German historian Leopold von Ranke defined 'blos zeigen, wie es eigentlich gewesen'² as the task of historiography. For a long time, this has been translated as 'merely to show how it really was'. Such an understanding of the occupation of a historian implies that the historian employs a positivistic approach, discovers objective facts and uncovers the truth about the past. But a more appropriate translation would be 'merely to show how it essentially was'; according to this small shift, the task of the historian would therefore be to deliver the essence of the past, an essence that has to be 'distilled' from the sources and might not be directly contained in them. It is obvious that Ranke understood very well that history is a subjective discipline,³ and he was fully aware of the function of a historian within society. To determine where we come from, who we are, which norms are formative for our behaviour, how history shapes the objectives of our actions and those of institutions, and what the historical record means for the present and the future, all this entails the creation of identities. Or, to present it in another way: if people's understanding of themselves requires history, then identity formation and foundation no longer work without history.

This function of historiography is not specific to the nineteenth century or the historicism of Ranke's time. Another famous German historian, Thomas Nipperdey, began his impressive *Deutsche Geschichte* (published in 1983 for the first time) with the sentence: 'Am Anfang war Napoleon.'⁴ Invoking Genesis, German history and identity are described with regard to Napoleon's massive influence on the Germans and in contrast to the major opponent of the Germans at the beginning of the nineteenth century. Knowledge created by historians is therefore ultimately relational, and the task of the historian is to deliver a comprehensive interpretation to her contemporary society and thus to provide meaning and orientation, a function which is characteristic of the humanities as a whole. And yet, the subjective dimension inadvertently contained in historical writing does not impair objectivity. The guarantee of this objectivity – for

historians as well as for literature researchers – is method. Alongside the development of the discipline of history, several methods have been established: the historical-critical method of the German tradition following Schleiermacher or books like *The Historian's Craft* (Marc Bloch)⁵ and *What Is History?* (E.H. Carr)⁶ describe the methodological rules, sources and models against which the resulting narrative can be tested – and if it is in discord with them, such a narrative can be refuted as being a distorted image of the past. The chosen method thus restrains the multitude of interpretations by what can be found in the sources, and the method itself, its prescriptions and the models established by it remain non-arbitrary. As such, it guarantees verifiability and the intersubjective validity of the historical narrative. If this was not the case, erroneous and misleading interpretations of the past – take the works of holocaust deniers as an example – could not be refuted. By applying a rigorous method, historians can check which models stand the scrutiny of the sources and in this way ensure transparency about the processing of historical narratives. The objectivity of the narrative – be it in history or literature studies – is thus warranted by its embedding in science as a social system. It's also about the nature of knowledge – tolerance for ambiguity and uncertainty can be higher in a system based on 'preponderance of evidence' (careful interpretation) than on 'beyond a reasonable doubt' (proof of fact). Hence refuting conspiracy theories isn't necessarily effective for those determined to believe those elements of them that could be true.

Digitization and the change of interpretive practices in the humanities

With the beginning of the twenty-first century and the advent of digitization, the practice of historians and literary scholars began to change profoundly, as long-established epistemic cultures began to come into contact with the very different requirements of data-centric research. In the twentieth century, access to sources and primary literature was limited by the necessity to physically access cultural heritage institutions like archives and libraries. Travel bursaries were needed to visit these institutions, to consult finding aids and catalogues and to consult with archivists and librarians in order to unlock the knowledge embodied by them. Digitization now implies what could be described in Foucauldian terms as an 'epistemic rupture':⁷ a profound reorganization of the knowledge base and the ways in which relevant sources are collected, organized and narrativized. With more and more metadata such as archival descriptions and information

in catalogues becoming available online, the access to pertinent sources is potentially broadened; at the same time, other sources become obfuscated where metadata are not yet available or never will be for reasons of privacy or lack of description. On the side of the researchers, rich information environments require them to develop new skills to reduce noise and enhance signal, thus necessitating knowledge of new techniques which are not part of the classical formation provided by these disciplines.

Digitization can be regarded as a disruptive force here, since it does not entail simply bringing the long-established finding aids and cataloguing procedures into digital form. Rather, this information has to be brought into new formats according to metadata standards ensuring data exchange between institutions, interoperability, aggregation and scaling. The creation of metadata, a kind of description that seeks to organize knowledge rather than merely annotate and let the algorithm find patterns, is therefore crucial for the representation of complex knowledge, since it represents a shift in the scope of re-contextualization and at the same time redefines the terrain of researchers' exploration. In order to avoid levelling differences and to bypass reductive and limiting metadata structures, the experts establishing those data have to anticipate which possible research roads the users will take, a human determination of signal versus noise not always applied by data-centric approaches. Archivists and librarians need to provide context within given metadata standards, thus deprioritizing some knowledge by dint of elevating other knowledge over it. With online accessible metadata, archival descriptions and catalogue formats become the new interface between researchers and cultural heritage institutions. This profound shift in their relationship produces a range of yet uncertain consequences: firstly, the aim of having all descriptions of available material online signals to users that the dialogic exchange with the archivists and librarians may be superfluous. For historians, this may result in bypassing the knowledge embodied in cultural heritage practitioners, as will be discussed in the next chapters. With users becoming enabled to explore the entirety of the cultural heritage collections online, their terrain of discovery becomes enlarged, and the archivists' scope of re-contextualization gets reduced. Cultural heritage practitioners themselves notice their own detachment from knowledge they presided over. Thus the power structures regarding access to knowledge and knowledge creation change. Furthermore, researchers struggle to acquire the competences necessary to perform complex searches. This presents a huge challenge since the most fruitful knowledge enquiries require an intimate knowledge about the specific data structures in which relevant information is presented, not just familiarity

with relevant keywords. For historians, their formative training provided the skills necessary to navigate through finding aids and the traditional hierarchical structure of archival collections through which contextual connections could be traced. Keyword searches may act as an approximation, with important consequences for the complexity of the search results, especially with regard to archival sources that may not be captured by this kind of approach.

For literary scholars, digitization of primary material free of copyright provided access to broad corpora; hundreds of thousands of books were now available in digital format, endowed with machine and human readable transcriptions via optical character recognition (OCR). While Computational Literary Criticism had been confined to a manageable number of texts and had focused on authorship attribution until the 1980s, the availability of large corpora required the application of statistical methods. This shift in the paradigm of knowledge production was not welcomed by every part of the community of philologists.⁸ The need to establish filtering procedures to pick a manageable corpus of texts relevant for the research question brought the interplay between individual researchers, their research environments and the libraries and infrastructures providing metadata to the fore, and decisions on the tools and methods appropriate for the available research objects had to be reflected carefully upon. It is not by chance that text-oriented digital humanists took up classical knowledge discovery instruments like indices and registers known from book printing. This observation can be substantiated with regard to the enthusiasm of the digital humanities for subject, topic and person indexes and directories of places, which correspond to techniques such as topic modelling or named entity recognition (NER), or their preference for visualizations such as maps and chronologies or timelines. All these examples refer to classification and categorization systems as have always been contained in historical books, and which preform insight and knowledge. Moreover, the context-dependency of data introduces the suspicion that they might not render it possible to determine the meaning. As one of the KPLEX interviewees from a computer science background stated:

If I read this and then I read this one, I cannot say which one is correct because it's like it depends on the context and I am not in this context. So it's really hard to it's like once you get the data and you remove the context, you don't know any more what – maybe there are hidden parameters and you don't know them, there are hidden variables that, you know, condition the way data have been collected, and once you don't know any more what these variable are it's very hard to describe.

(WP2 INT2)

New techniques like data cleaning had to be acquired by philologists who were accustomed to a strong reverence for the printed word. Data cleaning does not only imply the correction of OCR errors, but also the processing of texts by removing stopwords, calculating term frequencies, identifying the most frequent bi- and trigrams and so on. These techniques are apt for substituting distant reading where close reading is the norm, but at the same time those readers accustomed to interpret the polysemy and ambiguity of words and phrases immediately identified data cleaning as a procedure which obfuscates as much as it elaborates. The provision of metadata, for example, the indications of the literary genre recorded in a library catalogue, was questioned as a valid source of information, since they immediately become recognizable as time-bound social constructions.

The establishment of metadata as the new interfaces between researchers and their research material also has had consequences for the creation of narratives. Certainly, scholars are still required to directly access sources and primary material in the very same way as they did in the twentieth century. But with the shift to digital discoverability and the establishment of newly created metadata as the primary point of entry, metadata occupy a new role. As the organizing scheme that facilitates accessibility and discoverability of data, they inevitably support and preform the transformation of data into narratives. In this manner, metadata become a sort of actor, shaping and reshaping the range of possible narratives that can be created.⁹ It is hard to predict the consequences this has for scholars and the narratives they produce, and even more difficult to estimate the outcomes of this Foucauldian power shift in knowledge creation for the insights and narratives gained through the analysis of big data aggregated on the basis of these metadata.

The historical sciences and big data

It is remarkable to note the distance of historical science from the terms ‘data’ and ‘big data’, especially with regard to the observation that ‘big data’ – or at least their forerunner – have always been at the hands of historians. One may think of the nineteenth century, for example, where the economy was still predominantly agrarian and strong bureaucracies like the one in Prussia collected large amounts of historical records on agriculture and demography for statistical and fiscal purposes. The disinterest on the side of historians for these sources which can still be found in the archives, and the accompanying

indifference to numbers, is therefore in need of an explanation. This can only be understood by a deep dive into the history of historiography, and here the story begins in Germany, which was, according to British Historian Blackbourn, 'not only the home of modern, archival, "scientific" history, but an international leader in philosophy, philology and law',¹⁰ so while the trends we are looking at have general applicability, we can use the German historical frame as a point of reference for their origins.

One of the important pioneers of German historicism was Johann Gottfried Herder, who reaccentuated the interpretation of the past and brought the conception of 'Volk' as an agent to the connection between period and culture.¹¹ The German 'Volk' as a subject of history, from which cultural and societal conditions silently emerge: it was this master narrative which forms Herder's lasting contribution to historiography. Furthermore, before the establishment of the discipline, Wilhelm von Humboldt's reflections on how scientific knowledge of past realities can be created by contemporaries and how understanding is possible at all paved the way for a theory of historical hermeneutics as well as the development of the historical-critical method. Finally, as a third important source, enlightenment's philosophy of history has to be named, especially with respect to romantic conceptions of the 'Volk' as the source of language and customs, of law and poetry, and Hegel's concept of history as having a direction and a telos, as progress towards freedom.

Nourished by these sources, classical philology yielded classical and ancient studies, and developed in Germany throughout the nineteenth century into historical science. German historical scholarship gained an extraordinary prestige, especially with regard to historical encyclopaedias. Moreover, also jurisprudence developed into a historical science, with leading figures such as Karl von Savigny and Karl Friedrich Eichhorn. With historicism taking hold, Niebuhr and especially Leopold von Ranke transformed historiography into a strict science. Ranke performed a recalibration of the focus of historiography on politics, government, constitution, law, churches and institutions. With this shift, the emphasis of historical science went over from culture and 'Volk' to the nation state. In this way, historical science and historiography became between 1840 and 1870 a politically oriented, engaged science dealing with the identity of the citizens of a nation, with freedom and emancipation, progress and political unity. German unification in 1871 then provided a boost for nationalist historians like Sybel, Droysen and Treitschke; at the same time, the prospering economy and the advantages of the German Reich – such as the harmonization of currency and patents, the rule of law and

a liberal commercial code – were welcomed by the liberals.¹² A common notion of progress undergirded the establishment of the paradigm of modernization. It is not surprising that contemporaries were obsessed with statistics to show how everything was becoming larger, better or quicker. The German railway became a symbol of progress, as became everything connected to steam-power – steam engines, steam ploughs and steamboats. Beyond this fixation on technical modernization and material improvement, social and cultural advance became a testimony to German progress, and the national historiography of the German late nineteenth century concentrated these elements into a narrative creating a sense of identification. With its focus on the nation-state and the triumphant representation of the German Reich as the telos of history, historical science took a central position within the sciences and the humanities remained among them as a ‘Wissenschaft’ then and to this day. While this process took place in Germany during the phase of the differentiation of the scientific disciplines in the nineteenth century, at the same time classical national economy and the theory of the liberal-capitalistic economic system were developed mainly in Great Britain (Adam Smith, David Ricardo). In contrast to this upsurge, hardly little more than the reception of these advances can be noted in Germany. It is therefore not surprising that almost no integration into the historical sciences took place there, which can best be seen from the disregard German historians had (and still have) for Karl Marx and his theory. German national economy developed a focus on the social question and on a rather sociologically oriented empirical-statistical social research. If data on economy, on social classes and social stratum as well as on behaviour make up what we would now call ‘big data’, it is obvious that the establishment and consolidation of disciplines like macroeconomics, statistics, sociology and psychology – and the collection of the respective data – have taken place alongside with the formation of the specific functions of historical science and its neglect of numerical data in the late nineteenth century. Furthermore, in contrast to the occupation of the historians working in archives and being concerned with primary sources, these disciplines are marked by a strong division of labour, with longer chains from the data collector up to the scientist compared to how historians work. While macroeconomics, statistics, sociology and psychology may therefore be much more acquainted to the term ‘data’, it has to be noted that their process of knowledge production is marked by an alienation from the sources in a more profound way than in the historical sciences.

Numbers and description, narrative and interpretation

In order to explain the distance of the historical and literary sciences from the term ‘data’ and to characterize the way in which these disciplines make sense out of the primary material, one has to look at the yet under-researched question of the epistemological relationship between numbers and interpretation. Here it is noteworthy that etymology points to the common roots of the terms ‘to count’ and ‘to re-count’,¹³ of ‘to tally’ and ‘to tell’¹⁴ or, in German, of ‘zählen’ and ‘erzählen’.¹⁵ The historical meaning of these words therefore refers to the relationship between description and interpretation, between enumeration and storytelling. ‘Counting’ may have been used here, in a pre-modern and pre-scientific context, in a self-referential, tautological way to count the most obvious (such as fingers, people, oxen). Understood in this way, it points to objectively independent entities which cannot be modified at will. This brings us back to the concerns of the last chapter, as what is counted therefore meets the definition of ‘datum’, a ‘given’ which cannot be changed. At the same time this very fundamental understanding meets the definition of ‘fact’, insofar as it can be proven to be true with evidence and is independent of opinion and interpretation. ‘Re-counting’, in its pre-modern meaning, simply designates telling a story or constructing an account of that which has been counted and which has been ordered (first this, then that and so on), in the sense of narratively going along a sequence of numbers. It is here where Mary Poovey has mounted her *History of the Modern Fact*, by which she narrates how two functions – describing and interpreting – came to seem separate from each other in the early modern period, and how numbers, condensed in figures, came to seem distinct from interpretation. In analysing historical debates about induction and by studying authors such as Adam Smith, Thomas Malthus and William Petty, she comes to the following conclusion: ‘Separating numbers from interpretive narrative, that is, reinforced the assumption that numbers were different in kind from the analytic accounts that accompanied them.’¹⁶ Poovey is able to show that in the pre-history of ‘the modern fact’ two dimensions became intertwined: first of all, the assumption about epistemology that systematic knowledge should be derived from (and as such is superior to) non-interpretive descriptions. Numbers and quantities were seen as a disinterested representation of objectively verifiable units. Figures, therefore, simultaneously seem to describe discrete particulars, represent quantified items and serve analytic accounts, and thus seemed to be different in representational kinds. According to Poovey, the second dimension

dates back to the seventeenth century, where members of the British Royal Society discussed arguments about the possibility to collect data immune from theory or interpretation and thus opened a gap between theory-free particulars and systematic knowledge. Such a separation of observation from systematic account, inherited from the early modern period, severs the connection between description and interpretation and implies the non-interpretive, value-free and ontologically different status of numbers. Even though it can be read from nineteenth-century texts that even numbers were understood to be interpretive because every quantification embodies theoretical assumptions of what should be counted, by that time two modes of representation had become customarily graphically separated: numbers (presented in figures) and text (forming the narrative commentary).

The relationship between description and interpretation within the larger frame of the epistemological process changed with the reconfiguration of the conception of objectivity in the late nineteenth and early twentieth century and the emergence of mechanical objectivity.¹⁷ As Daston and Galison explain in their book, objectivity and subjectivity are inextricably bound together, with objectivity meaning the suppression of some aspects of the subject as well as the artificial division of the researcher-subject into an active experimenter and a passive observer. In the nineteenth century, scientists developed an ethos obliging researchers to restrain themselves in order to repress their wilful intervention and acquire virtues like patience, tirelessness and industriousness; furthermore, procedures were developed in order to reliably bring observations from nature onto the page documenting the scientist's activities through a strict protocol. The development of photography is emblematic in this emergence of mechanical objectivity, and two intertwined processes are at work here: On the one hand, the separation of the development and activities of machines from the human beings who conceived of them, with the result that machines were attributed freedom from the wilful interventions that had come to be seen as the most dangerous aspects of subjectivity. Thus machines – be they cameras, sensors or electronic devices, or indeed the data they produce – have become emblematic for the elimination of human agency and embody objectivity without subjectivity. On the other hand, the ideal of mechanical objectivity also implied alterations on the side of the subject of the researcher, favouring non-intervention and evolving around practices including 'training the senses in scientific observation, keeping lab notebooks, drawing specimens, habitually monitoring one's own beliefs and hypotheses, quieting the will, and channelling

the attention.’¹⁸ Thus an ethics of objectivity was developed, which called for a morality of self-restraint in order to hold researchers back from intervention and interferences like overinterpretation, aestheticization and theoretical overreaching, precisely the kinds of hesitations around narrative we see in the KPLEX interview cited earlier in this chapter. The emergence of mechanical objectivity thus also reconfigured the relationship between description and interpretation: while photography became emblematic for objective truth, self-elimination in the interpretation of automatically transferred records became a precept for the scientists. Even though contemporary researchers may have been fully aware of an epistemology comprising the design and construction of machines for the purpose of data collection according to given research questions, the persistence of the belief that machines deliver impartial and objective records of a reality is remarkable and can only be explained with the auratization of mechanically fabricated products like photographs. At the same time it is obvious that such a conception of a division between a mechanical device responsible for collecting and recording data and the researcher responsible for impartially making sense of what had been collected is disjunct from the activities of historians and literature researchers who are occupied with the interpretation of sources and research material produced and collected prior to the beginning of their research. This is not to say that historians do not use mechanically produced materials like photographs for their studies or that they did not adhere to virtues like self-discipline, self-restraint and self-control. Methodologies like the critical assessment of the sources have been developed as technologies to control and curtail the dangers contained in the judgements of the researchers. But historians have simply not invented mechanical devices to collect the information that they need for answering their research questions. In this way, the term ‘data’ has become alien and unusual to historians – and to humanists in general – and data do not form the primary source out of which interpretation and narratives are created.

Science as a social system: The social construction of meaning

While the development of mechanical objectivity can be seen as decisive in setting the course for the epistemic bifurcation between the natural sciences and the humanities, there is still the need to explain why common sense still attributes objectivity and impartiality to data and identifies them with facts,

whilst interpretation and narrative are seen as being ontologically distinct from them. This has to do with the different social systems and the peculiarities and customs that have emerged in the various scientific disciplines. As has been noted above, for historians and literature scholars, the establishment of models, the strict application of methods, the reference to sources to enable transparency, and the questioning and discussion of the research results by the audience of peers are basic elements by which the objectivity of the scientific narrative is guaranteed. These agonistic disputes seem to function well as an internal corrective, even though the case of the German 'Historikerstreit' shows that such a debate may remain unresolved. Generally speaking, scientific knowledge is socially constructed because it is partly derived from collective beliefs held by scientific communities.¹⁹ Beyond the activity of the individual researcher, the embedding of research results within a broader scientific discourse shows that scientific research is a collective activity. Each scientific collective develops characteristic habits and customs of directed perception which shapes the ways in which the members of the collective perceive and think about the world, and which is responsible for a shared understanding of what is accepted as evidence.²⁰ Daston and Galison call this 'trained judgement'.²¹ Acquired habits are also responsible for the community of researchers to 'forget' that an object has been declared a scientific fact by virtue of convention and learning. Although researchers are aware of the acquisition of these abilities, they accept it as tacit knowledge. Learning to see like a scientist is a question of accumulated experience. Collectives of scientists therefore develop and learn as part of their professional formation, and the acceptance and recognition within the scientific collective form an important precondition for the approval or dismissal of the hypotheses brought forward in scientific studies within the larger scientific discourse in which they are debated. Much of what is termed 'science' therefore relies on social practices and procedures of adjudication. As the historian of science Naomi Oreskes has recently commented, the heterogeneity of the scientific community supports the strength of the achieved consensus: 'Objectivity is likely to be maximized when [...] the community is sufficiently diverse that a broad range of views can be developed, heard, and appropriately considered.'²² Ultimately, the processes performed within such scientific communities result in what is designated as 'truth'.

While the quality of such consensual scientific knowledge may depend on factors like the level of abstraction and complexity, the agreement on the appropriate methods to answer the debated questions, or the ideological orientation of the participants, the reliance on sources which have not been

created and collected according to certain research questions remains as a feature distinguishing these humanistic disciplines from natural sciences like biology. In the latter disciplines, the design and development of devices and laboratories constructed to assume the task of data collection assumed an enormous extent in the twentieth century and involved significant material investments. Data were now created in abundance, and an important feature of scientific discovery consisted in the use of procedures to create an account out of the chaos of available perceptions, that is, creating order out of disorder by making sense of the observations at hand. In their pioneering study entitled *Laboratory Life*, Latour and Woolgar analysed what they described as the social construction of scientific knowledge and as the process of establishment and enforcement of statements and data as ‘facts’.²³ Social constructivism, the sociological theory which forms the background of the study, focuses on the social situatedness of knowledge construction through human interaction. It is precisely this situatedness which allows the process of fact construction to be described as one in which social factors disappear and the traces of their constructedness become difficult to detect. This does not imply that facts are entirely constructed, because then they would be artefacts; in what has now become a famous paradoxical formula, Latour and Woolgar describe the way ‘in which the term fact can simultaneously mean what is fabricated and what is not fabricated’.²⁴ On the one hand, they point to the etymology of the word ‘fact’ as derived from the Latin terms ‘facere’ and ‘factum’ (to make or to do) and underline the process of its construction; on the other hand, they explain that scientific statements ‘are referred to as a thing “out there” (objectivity and fact)’.²⁵ In other words, what is termed ‘fact’ can certainly be found in and verified by the collected data, but at the same time it is subject to a process of purification from social factors. Science appears here as a negotiation process in an agonistic arena, in which some forces bring forward statements which prove the artefact-like status by proving social involvement (subjectivity) in the process, while other forces push statements forward which speak for the fact-like status or the ‘out-there-ness’ (objectivity). The whole process ensures that once the debate settles facts are taken for granted. The paradoxical nature of ‘facts’ therefore reflects the process of data production by assertedly neutral and impartial machines which have been designed by humans according to specific research interests and which therefore rather deliver ‘capta’ than ‘data’, and the scientific – and therefore inevitably social – activities around the construction of facts and their recognition as such by a scientific collective.

The larger part of Latour and Woolgar's study is devoted to the analysis of the process of social construction and describes the long line from the construction of facts, the result of which is that facts appear not to be constructed by anyone, up to the presentation of a credible account which aims at affirming and defending the existence of those facts in the arena formed by the collective of scientists. Other than simplistic conceptions of a binary relationship between data/facts (or description/observation) and accounts (or interpretation/narrative), Latour and Woolgar describe the construction of facts as a complex process of negotiations which includes moving back and forth between the two poles. This process includes the elimination of alternative interpretations of scientific data and the rendering of these alternatives as less plausible (e.g. by delegitimising interpretations brought forward by rivalling scientists), the vanishing of elements which are described as circumstances from the final accounts, the introduction of inequality into a set of equally probable statements in such a way that a statement is taken to be more probable than all the alternatives (a technique to decrease noise in the data), the increase of cost for others to raise equally probable alternatives which results out of the material provisions with which the laboratory is equipped (the operationalization of economic capital), the credibility (or accumulated symbolic capital) of the scientist which supports her accreditation and creates the credit needed for holding her account accountable and, as a final activity, the writing and publication of scientific articles, which form a central activity of creating order and a socially accepted reality. While the study by Latour and Woolgar is restricted to the specific research setting of a laboratory, it is fascinating to see how rare studies of this kind are, and to perceive the lack of profound observation of scientific practices in general. Certainly, this is one of the reasons why the collection, analysis and interpretation of big data are shrouded in mystery, especially when it comes to the analysis of the complex chains from data to interpretation in the inaccessible towers of knowledge production erected by large tech companies.

Making sense of big data

Latour and Woolgar analysed a setting in which data were produced in abundance. The social construction of facts includes various procedures enabling the discrimination of signal and noise; the researchers in the laboratory inspect large datasets conflated in tables and graphs. This situation has intensified considerably with the rise of big data. These are *per se* data which

have been collected by machines and are too big to be inspected by humans. Their magnitude has consequences: they are so large that typical applications to store, compute and analyse them are inappropriate. Often processing them is a challenge for a single computer; thus, a cluster of computers has to be used in parallel. Or the amount of data has to be reduced by mapping an unstructured dataset to a dataset where individual elements are key-value pairs; on a reduced selection of these key-value pairs mathematical analyses can be performed ('MapReduce'). Even though big data are often not collected in response to a specific research question, their sheer size (millions of observations $\times$ of variables y) promises to provide answers relevant for a large part of a society's population, as will be discussed more broadly in Chapter 7. From a statistical point of view, what happens is that large sample sizes boost significance; the effect size is more important. However, on the other hand, large does not mean all; one has to be aware of the universe covered by the data. Statistical inference – conclusions drawn from data about the population as a whole – cannot easily be applied, because the datasets are not established in a way that ensures that they are representative.²⁶ Beyond parallel computing and MapReduce, the use of machine learning seems to provide viable solutions to handle big data. Machine learning designates algorithms that can learn from and make predictions on data by building a model from sample inputs. It is a type of artificial intelligence in which the system learns from lots of examples; results – such as patterns or clusters – become stronger with more evidence. It is for this reason why big data and machine learning seem to go hand in hand. Machine learning can roughly be divided into two approaches: on the one hand, analytic techniques which use stochastic data models, most often classification and regression in supervised learning and, on the other hand, predictive approaches, where the data mechanism is unknown, as it is the case with neural nets and deep learning.

The goal of statistical modelling is to find a model which allows quantitative conclusions to be drawn from data. It has the advantage of the data model being transparent and comprehensible to the analyst. Statisticians in applied research consider data modelling as the template for statistical analysis and focus within their range of multivariate analysis tools on discriminant analysis and logistic regression in classification, and multiple linear regression in regression. This approach has the advantage that it produces a simple and understandable picture of the relationship between the input variables and response. But the assumption that the data model is an emulation of nature is not necessarily valid and can lead to wrong conclusions. What sounds objective since it is 'based on statistics' is not necessarily correct as, if the model is a poor emulation of reality, the conclusions

may be wrong. This is the reason Cathy O'Neill has called an algorithm 'an opinion formalized in code'²⁷ – it does not simply provide objectivity, but works towards the goals for which it was written.

In recent years, powerful new tools for big data analysis have been developed: neural nets and deep learning algorithms. The goal of these tools is predictive accuracy; they are hardware hungry and data hungry, but have their strength in addressing complex prediction problems where it is obvious that stochastic data models are not applicable. Therefore, the approach is designed in another way here: what is observed is a set of *x*'s that go in and a subsequent set of *y*'s that come out. The challenge is to find an algorithm such that for future *x* in a test set, the result of that algorithm being applied to *x* will be a good predictor of *y*. The goal is to have the algorithm produce results with a strong predictive accuracy. The focus does not lie with the model by which the input *x* is transformed into the output *y*; it does not have to be a stochastic data model. Rather, the model is unknown, complex, mysterious and irrelevant. This is the reason why accurate prediction methods are addressed as complex 'black boxes.' As opposed to the case with stochastic models, the goal is not interpretability, but accurate information. And it is here on the basis of an opaque data model, where neural nets and deep learning extract features from big data and identify patterns or clusters which have been invisible to the human analyst. It is fascinating to see that humans don't decide or predetermine what those features are. The predictive analysis of big data can identify and magnify patterns hidden in the data.

The interpretation of the results of big data analyses therefore implies a choice between the devil and the deep blue sea: either the analyst may rely on a data model that is comprehensible to her, in full awareness of the deficiencies of such a model, since the complexity of the world far outstrips the stochastic data models which have been constructed to explain it ('all models are wrong'²⁸), or the analyst resigns herself to the 'black box' and the high predictive power associated with it, being fully aware that correlation is not equivalent to causation, resulting in the separation of understanding and knowledge. With the availability of a growing amount of big data, computational power and the work of capable developers of algorithms, the approach based on stochastic data models has become outdated in favour of predictive accuracy. But the lack of interpretability of the relationship between prediction and response variables accompanying the latter approach widens the gap between understanding and knowledge and renders narratives dysfunctional, since they rely on the power to explain causal relationships. Scientific narratives have always carved out such relationships in multiple ways and are therefore key to human understanding

of the complex world around them. With the advance of big data and AI, central elements of scientific understanding – like principles, laws, models, generalizations and representativeness – lose their worth. Such a development leads to structural blindness: since it is still humans who make decisions on the basis of their interpretations of data, they are deprived of key tools used in the interpretive process, which enable the generation and collateralization of narratives, discussion of uncertainties and doubts, and ensure the adherence to methodological rules and models against which the results can be tested. As a consequence, Law Professor Frank Pasquale has warned that ‘authority is increasingly expressed algorithmically. Decisions that used to be based on human reflection are now made automatically.’²⁹ Finally, the analysis of big data and the application of artificial intelligence in specialized laboratories owned by private corporations entail that the researchers working there are being cut off from the obligation to have their research results discussed and approved by relevant scientific communities, as it is the case with state-funded researchers who have to justify why they should be funded.

While it has become obvious that big data inherit the aura of data and facts as disinterested representations ensuring objectivity, it is often overlooked that big data have their specificities as well: as scalable data, they view individuals as representatives of social groups, identify subjects as ‘users’ or ‘customers’ rather than as citizens or members of a community (yet more revealing linguistic habits of technology firms, such as were discussed in the last chapter), focus on consumption patterns rather than the significance and meaning attributed to them by the individual and take behaviour into perspective instead of the inner life.³⁰ Human beings are not perceived as individuals in a holistic way, but in their partial existence as producers of data or money. Out of this homogenizing force results an indifference towards meaningful particularities of a given world, an insensitivity towards cultural differences and an incapacity for the evaluation of uncertain, hazardous and conflicting information. Finally, the inability to incorporate knowledge which is not scalable – like context – impairs the development of a profound understanding of the particular. Against this backdrop it is necessary to underline that the humanities not only involve a different way of creating meaning entailing the critical assessment of sources, the representation of non linearity in developments, reflections of the method, an openness to double-checking and rivalling interpretations, the discussion of alternative approaches and research results, and the inclusion of thick description and context to ensure narrative richness, but also that the function of the narratives delivered by them is quite distinct from those of interpretations

of big data: the provision of identities, of orientation and societally relevant sensemaking, and a critical reflection of the past for the present and the future.

Notes

- 1 Warburg Institute, *Lorraine Daston: Exempla and the Epistemology of the Humanities*, 2016, <https://www.youtube.com/watch?v=8JlXflyqsG4>.
- 2 Von Ranke, Leopold, 'Geschichten der romanischen und germanischen Völker von 1494 bis 1514', in *Leopold von Ranke's Sämmtliche Werke*, 2nd edn (Leipzig: Duncker und Humblot, 1874), 33/34, VII.
- 3 This can be read, for example, from Ranke's statement which reveals – through the use of the irrealis – his consciousness about a remaining rest of subjectivity: 'Ich wünschte mein Selbst gleichsam auszulöschen, und nur die Dinge reden, die mächtigen Kräfte erscheinen zu lassen.' Von Ranke, Leopold, 'Englische Geschichte vornehmlich im siebzehnten Jahrhundert', in *Leopold von Ranke's Sämmtliche Werke*, 3rd edn (Leipzig: Duncker und Humblot, 1870), 15.2, 103.
- 4 Nipperdey, Thomas, *Deutsche Geschichte 1800–1866: Bürgerwelt und starker Staat*, 5th edn (Munich: C.H. Beck, 1991), 12.
- 5 Bloch, Marc, *The Historian's Craft* (New York: Knopf, 1953).
- 6 Carr, Edward Hallett, *What Is History?* (London: Macmillan, 1961).
- 7 Foucault, Michel, *The Order of Things: An Archaeology of the Human Sciences* (New York: Pantheon Books, 1970).
- 8 A valuable and concise history of Computational Literary Criticism is provided by Beausang, Chris, 'A Brief History of the Theory and Practice of Computational Literary Criticism (1963–2020)', *Magazén*, 2, 2020, [JournalArticle_4061](#).
- 9 Edmond, Jennifer and Georgina Nugent Folan, 'Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources', in *Metadata and Semantic Research*, ed. Emmanouel Garoufallou, Sirje Virkus, Rania Siatiri and Damiana Koutsomihia, Communications in Computer and Information Science (Cham: Springer, 2017), 253–60.
- 10 Blackbourn, David, *History of Germany 1780–1918: The Long Nineteenth Century* (Malden, MA: Blackwell, 2005), 208.
- 11 Nipperdey, *Deutsche Geschichte 1800–1866*, 498–533; Herder, Johann Gottfried and Michael N Forster, *Philosophical Writings* (Cambridge, UK: Cambridge University Press, 2002); Barnard, Frederick M, *Herder's Social and Political Thought* (Oxford: Clarendon Press, 1967).
- 12 Blackbourn, *History of Germany 1780–1918: The Long Nineteenth Century*, 204–14.
- 13 'Count [ME]: The verb to count is from Latin *computare* "to calculate", the root also of Computer, account [ME], and recount [LME] "tell" (which can also be used for

- both “narrate” and “count”).’ Cresswell, Julia, *Oxford Dictionary of Word Origins*, 2nd edn (Oxford: Oxford University Press, 2010), 104.
- 14 ‘To Tell, [Taellan or Tylean, Sax. Teler, Du. Zaehlen, Teut.] to count or number. To Tell, [Tellan, Tax. Talen, Du. Tale, Dan.] to relate, to make known.’ Bailey, Nathan, *An Universal Etymological English Dictionary* (Hildesheim/NewYork: Georg Olms Verlag, 1721).
 - 15 ‘erzählen Vsw std. (8.Jh.) mhd. *erzeln*, *erzellen*, ahd. *irzellen*. Bedeutet ursprünglich ‚aufzählen‘; dann, in geordneter Folge hersagen, berichten; woraus durch Verallgemeinerung die heutige Bedeutung entstand. Abstraktum: Erzählung; Nomen Agentis: Erzähler.’ – ‘zählen Vsw std. (8.Jh.) mhd. *zel(e)n*, *zellen*, ahd. *zellen*, as. *tellian*. Ae. *tellan*, anord. *telja* aus g. *tal-ija-* ‚zählen‘. Ableitung von Zahl. ‘Präfigierung: erzählen; Abstraktum: Zählung’. Kluge, Friedrich, *Etymologisches Wörterbuch der deutschen Sprache*, 25th edn (Berlin/Boston: De Gruyter, 2011), 258, 1000.
 - 16 Poovey, Mary, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (Chicago: University of Chicago Press, 1998), xv.
 - 17 Daston, Lorraine J and Peter Louis Galison, *Objectivity* (New York: Zone Books, 2015), 115–90.
 - 18 Ibid., 199.
 - 19 Fleck, Ludwig, *Genesis and Development of a Scientific Fact* (Chicago: Chicago University Press, 1979).
 - 20 And this disciplinarity is then inscribed on the data themselves – data are constituted as belonging to one discipline and not another, in a linguistic sense of their not being understandable by other disciplines but also in that they are playing a role in defining that discipline in a Bernstein(ean) sense, as discussed in Chapter 2.
 - 21 Daston and Galison, *Objectivity*, 370–1.
 - 22 Oreskes, Naomi, *Why Trust Science?* (Princeton: Princeton University Press, 2019), 53.
 - 23 Latour, Bruno and Steve Woolgar, *Laboratory Life: The Social Construction of Scientific Facts*, Beverly Hills (Beverly Hills/London: Sage Publications, 1979).
 - 24 Ibid., 236.
 - 25 Ibid., 180.
 - 26 Kitchin, Rob, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: Sage), 19.
 - 27 O’Neil, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016), 53.
 - 28 So, Richard, Jean, “All Models Are Wrong”, *PMLA*, 132/3 (2017), 668–73.
 - 29 Pasquale, Frank A, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015), 8.
 - 30 Madsbjerg, Christian, *Sensemaking: What Makes Human Intelligence Essential in the Age of the Algorithm* (London: Little, Brown, 2017), 199–211.