social sciences

MDPI

Article

Researching Culture through Big Data: Computational Engineering and the Human and Social Sciences

Teresa Duarte Martinho

Universidade de Lisboa, Instituto de Ciências Sociais, Av. Professor Aníbal de Bettencourt 9, 1600-189 Lisboa, Portugal; teresa.duartemartinho@ics.ulisboa.pt

Received: 31 October 2018; Accepted: 8 December 2018; Published: 11 December 2018

Abstract: The emergence of big data and data science has caused the human and social sciences to reconsider their aims, theories, and methods. New forms of inquiry into culture have arisen, reshaping quantitative methodologies, the ties between theory and empirical work. The starting point for this article is two influential approaches which have gained a strong following, using computational engineering for the study of cultural phenomena on a large scale: 'distant reading' and 'cultural analytics'. The aim is to show the possibilities and limitations of these approaches in the pursuit of scientific knowledge. The article also focuses on statistics of culture, where integration of big data is challenging procedures. The article concludes that analyses of extensive corpora based on computing may offer significant clues and reveal trends in research on culture. It argues that the human and social sciences, in joining up with computational engineering, need to continue to exercise their ability to perceive societal issues, contextualize objects of study, and discuss the symbolic meanings of extensive worlds of artefacts and discourses. In this way, they may help to overcome the perceived restrictions of large-scale analysis such as the limited attention given to individual actors and the meanings of their actions.

Keywords: human and social sciences; data science; big data; distant reading; cultural analytics; statistics of culture; positivism/interpretivism/interactionism

1. Introduction

The emergence and dissemination of big data, together with the consolidation of information as a major category of thought, has produced for the human and social sciences a period of questioning and reconsideration of many of their objectives, theories, and methods (Savage and Burrows 2007) One of the consequences of the accessibility of the expanding amount of data generated by the new information technologies was that digital systems became the driving force for a new way of producing scientific knowledge. Its name varies from 'data science' to 'computer science' to 'data-enabled science' to 'web-enabled science', endorsing the creation of data-driven rather than knowledge-driven science (Martins 2011; Kitchin 2014). 'Scientific knowledge' is here used to designate an intellectual construct which seeks greater understanding of the world, generating ideas and theories and attempting to render them compatible with facts derived from observation. It is therefore knowledge which is not definitive, but may be questioned and adjusted as new facts emerge and as the scientist or researcher adopts a critical reflexive attitude to his work.

The dissemination of 'data science' is at the heart of a debate on the relationship of the human and social sciences and the 'big data' phenomenon, on account of the methodological and epistemological consequences of the data science model and its defense of the primacy of observational-inductive procedures. One of the sides in this cross-cutting debate is enthusiastic about big data, emphasizing how

the range and variety of the data available, in much larger quantities in relation to behaviors and attitudes, enables more comprehensive and as yet unrealized portraits of society (Lazer et al. 2009; Reagan et al. 2016; Manovich 2018). The unconditional fans of big data add to these virtues the savings in time and costs of research work. Others adopt a more detached, skeptical approach, drawing attention to the weaknesses and risks inherent in data mining-based research work and stressing that big data embody elements of society and culture and have their own politics (Lupton 2015). The main weaknesses attributed to research based on data mining are the following: predominance of correlations of facts and identifying of patterns, to the detriment of interpretation; tendency to reduce human action to behaviors, avoiding enquiry into the reflexive aspect (Beer 2016; Lupton 2016; Carrigan 2018)[1]. Others maintain that the human and social sciences gain by welcoming the convergence with computational engineering. They argue this cross-fertilization embodies within it the ability to translate information into knowledge and to apply the skills of synthesizing, contextualizing, and reflecting theoretically in order to articulate and understand the large quantity and variety of results (McFarland et al. 2015; Halford and Savage 2017).

The above-mentioned debate is part of a broader process of reorganization and transformation of knowledge and learning, guided by the search for and accumulation of information. This practice is often justified on the grounds of openness and transparency, principles which governments, universities, and cultural institutions have adopted. These trends have contributed to the creation of very large sets of cultural artefacts (e.g., books, images, newspapers and magazines, music, films, TV series) on digital media. The unprecedented amount of large cultural corpora offers unheard of possibilities for the study of culture by the human and social sciences, giving rise to lines of research in collaboration with computational engineering with the aim of facilitating the production of macroscopic social analyses and studies. Culture, in this article, designates artistic works and other forms of symbolic creation, and those of its aspects which are studied by the human and social sciences: (i) the internal structure of works: genres, styles, forms, topics, artistic movements; (ii) the contexts in which they are produced, disseminated and consumed; (iii) the interconnectedness of the cultural sphere with other social fields.

This article focuses on the earliest kinds of research in the human and social sciences which were based on synergies with computational engineering for the study of very extensive sets of cultural artefacts: 'distant reading' and 'cultural analytics'. The choice is justified, for the purposes of this article, not so much by the specific type of culture studied, but rather because they have characteristics which make analysis of these two types of research potentially relevant to a discussion of the implications of the use of big data and computation in broader culture research. The first characteristic is the common interest in very large corpora, in particular digitalized literature archives or image collections shared in virtual communities of artists. The large quantity of cultural objects studied becomes the condition for a predominantly quantitative approach to culture. The second characteristic is that they generally favor macro and longitudinal perspectives, with a view to detecting patterns and trends in the mass of data which are then reproduced in abundant visual forms (charts, tables, diagrams).

While "cultural sociologists have until recently made very few ventures into the universe of big data" (Bail 2014), it was in the field of human sciences that the first form of inquiry strongly grounded in computational methods emerged (Moretti 2013). In the year 2000 the researcher behind it, Franco Moretti, dubbed it 'distant reading'. It is used to designate the analysis of large groups of works, basically extensive literary texts stored in digital archives, written and published at various times and in different countries, mainly in English. 'Distant reading' seeks to capture trends in matters of genre, style, movements and topics, its practitioners arguing that the proliferation of digitalized literature

[1] Some writers suggest that the debate should begin by distinguishing between the 'material phenomenon' and the 'ideational phenomenon' of big data, the former referring to its social, technical and historical elements, the latter inquiring as to how the notion of big data developed in technical, philosophical, and cultural terms. Although they are connected, the heuristic value of the distinction lies in helping to identify, and not subsume, the way sponsors of big data infrastructure and technologies find material sponsorship within organizational contexts with specific characteristics (Beer 2016; Carrigan 2018).

collections provides opportunities for a better understanding of the past (Moretti 2013). Another form of macro analysis of the culture field, mainly based on analyzing images, is the tendency known as 'cultural analytics', an expression coined in 2005 by Lev Manovich. This approach, which seeks to combine computational engineering, the history of art, and sociology, favors using massive content generated and stored through digital interfaces, paying special attention to images published and disseminated through social networks (Manovich 2018).

We complement our analysis of 'distant reading' and 'cultural analytics' with an incursion into the official statistics of the European Union (EU) in the field of culture, for three reasons which, in our view, justify their inclusion in this mosaic. First, because official statisticians have demonstrated an increasing interest in incorporating big data: in 2016, *Culture Statistics* started to publish indicators drawn from Wikipedia, with a view to assessing how interest in Europe's cultural heritage has evolved (Culture Statistics 2016). Secondly, culture statistics share with the above-mentioned forms of inquiry the preference for macro and longitudinal approaches, with a strong visual component (charts, tables, and diagrams), based on handling very large sets of quantitative data. Thirdly, the incorporation of big data is challenging and reshaping procedures for handling quantitative data and producing scientific knowledge, revealing a shift in the epistemic values related to data modeling objectives (Pietsch 2013). In addition to these commonalities, statistics are significant because they are a source used by social scientists, above all those who study cultural practices and consumption in a broader spatial and temporal perspective, with a view to determining trends (Christin and Donnat 2014).

This study aims to demonstrate the possibilities and limitations of 'distant reading', 'cultural analytics', and statistics which use big data in producing scientific knowledge. In actual fact, the increasing volume and variety of data has enabled the extension of the geographical and longitudinal scope of analyses on a new scale, generating opportunities for the human social sciences to demonstrate their strong competences in contextualizing objects of analysis and in discussing meanings of large corpora. Nevertheless, strategies for collecting, storing, and analyzing such data are not without difficulties and risks.

The second, third, and fourth sections of this article are devoted to analyzing distant reading, cultural analytics, and the official culture statistics of the EU using big data. Each section contains an assessment of the possibilities and the limitations detected in these approaches, all of which have a common interest in quantifying social facts. In order to illustrate this counterpoint more clearly, several concrete examples are offered: the study, characteristic of the 'distant reading' method, of the representation of how London developed, drawing on about 5000 English novels published between 1700 and 1900 (Heuser et al. 2016); an analysis of a million works of art disseminated between 2001 and 2010 on the digital platform DeviantArt, in the categories 'Traditional Art' and 'Digital Art' (Yazdani et al. 2017); and the first cultural indicators from the EU using big data, based on millions of views of Wikipedia pages on cultural heritage. The conclusion offers an overall view of the different sections of the article, bringing them together for a considered, integrated view of the scope and limitations of the linkages between the study of culture in the human and social sciences by means of big data and its ties to computational engineering.

2. A Change of Scale and Culture under the Serial Logic of 'Distant Reading'

This section addresses the 'distant reading' approach, setting out its aims and assessing its scope and limitations in the study of culture through quantification and the use of big data. According to its mentor, Franco Moretti, a historian of literature intellectually and politically affiliated with Marxism, it was the lack of a cultural atlas, noted by Fernand Braudel in *The Mediterranean* (Braudel), which most inspired him to draw up an atlas of European literature (Moretti 1998). 'Serial history' became the engine of 'distant reading', as applied to the study of literature and other cultural and artistic forms on a large temporal and spatial scale, extending the scope of quantitative analysis. Moretti coined the term in 2000, without it having occurred to him that there existed already the *Bilderatlas Mnemosyne* project, developed by the art historian Aby Warburg, between 1924 and 1929, which contained a thousand

photographic reproductions of images from Classical Antiquity and the Renaissance. This historical laboratory of Western iconography remained little known and appreciated until the 1980s, because its approach to cultural history had become peripheral in mid-20th century, when modernist thought was dominated by the methodology of positivism (Ostrow 2014).

At stake here is the adoption of a heuristic consisting of experimenting with a change in the scale and context in which social phenomena are analyzed (Abbott 2004), an example of which is Fernand Braudel's thesis on the Mediterranean in the 16th century. In addition to establishing a particular connection between geography and history, Braudel's macro approach in part quantifies history, and creates graphs and maps, to illustrate variations in climate, demography, and other areas. But it was the writers of the 'new economic history'—a tendency also labelled 'cliometrics', the literal meaning of which is 'the measurement of history', beginning in the 1950s and developing much faster since the 1990s, who most specialized in the quantitative approach to history, reconstituting the past and studying topics such as demography, transport and international trade through the systematic use of statistics and mathematical methods, compiling vast series of quantitative data to that end. One of the consequences of 'cliometrics', according to some critics, is that the historical narrative tends to emerge as the simple result of applying theories and concepts borrowed from neoclassical and/or new institutional economics and of imposing them on the historical facts (Boldizzoni 2011).

The repercussions of quantitative history can be perceived in those who, like Moretti, shifted their analytical focus from "exceptional events" (which, in a study of culture, correspond, for example, to canonical works) to "the large mass of facts" (all works produced in a given period, far beyond the canon), in order better to capture relationships, repetitions, and patterns (Moretti 2005, pp. 3–4). The large-scale perspective and mass analysis make it possible to achieve 'distant reading', in order to grasp stability and change in literature through quantitative analysis of as many texts as possible, mainly in the form of an ability to understand the evolution of literary genres, and see the gradual emergence of certain topics and variations in stylistic choices, like the titles of works, the lexicon, punctuation, and the grammatical structure of sentences. To that extent, 'distant reading' could be offered as an opportunity "to do something additional that might enrich our understanding of the past", primarily because of its potential for "trac[ing] blurry family resemblances among texts instead of defining fixed categories [whereas before] (. . .) it was hard to trace loose family resemblances among thousands of volumes by hand" (Underwood in Dinsman 2016, pp. 13–14). This would give rise to new ways of understanding literary history: "as we slice [digital] libraries in new ways [through machine learning] we keep stumbling over long, century-spanning trends that have little relationship to the stories of movements and periods we used to tell" (Underwood in Dinsman 2016, p. 17).

The adoption of quantification and macro approaches by culture researchers evokes the grand theoretical debates in the social sciences and the specific nature of the methods associated with them. At the most radical level of theoretical debate we find the opposition between positivism and interpretivism. Positivists argue that life in society can be measured, and that those measurements are comparable between different contexts; the logical-positivist conception, which the writers of the Vienna Circle preferred to call empiricist-logical, stresses the importance of empirical verification and the axiological neutrality of science (Outhwaite 1996). Interpretivism, for its part, favors interaction and interpretation, and maintains that actions and events which seem measurable, quantifiable and serializable only become meaningful through interactive processes. In opposition to positivism, interpretivism and interactionism are parts of a general orientation or style of thought focused on meaning, comprehension, action, interaction, language, and context (Outhwaite 2005). The 'distant reading' approach may therefore be said to embody a predominantly positivist standard, in that it seeks to identify evolutionary trends in the mass of facts and in this way, work around the alleged limitations and inadequacies of approaches centered on events and the individual.

The particular features of Moretti's 'distant reading' lie mainly in a quantitative approach to literature through computational engineering, encompassing various forms of data mining[2], with the common use of algorithms to explore large sets and databases and to identify patterns. The ability to find "quantitative evidence" of change and stability is closely associated with the multiplication of large collections of digitalized books, which has intensified since the beginning of the 21st century, together with an industry which applies computational engineering techniques. Amongst the examples of this are Google Books and the Chadwyck-Healey databases. It is from these and other collections of digitalized literature that the corpora of various 'distant reading' research projects have been formed. Here we highlight "The Emotions of London" (Heuser et al. 2016), an example, among the most recent, of a macroscopic study of literature carried out in the Stanford Literary Lab, co-founded by Moretti in 2010 at Stanford University[3]. The study sample comprises 5000 English novels published between 1700 and 1900. This example is relevant to this section in that it illustrates how 'distant reading' may neglect contextualization. In line with our definition of culture within the scope of this article, artistic works and other forms of symbolic creation may also be approached in terms of the contexts in which they were produced, disseminated, and consumed, and taking into account the interconnectedness of the cultural sphere with other social fields.

Between 1700 and 1900, London went through profound changes. One of the aims of the authors was to detect those changes in fictional representation and to quantify and map emotions based on literature. Like other products of 'distant reading', the resulting text sits alongside many figures, diagrams and visualizations, which are the organizing principle of the document. One of the conclusions drawn by "The Emotions of London" is that there is a "growing *divarication*" and a "striking discrepancy" between fiction and reality, reflected in the over-representation of Westminster and the City and the relative lack of prominence given to Tower Hamlets, Southwark and Hackney (Heuser et al. 2016, pp. 3–9). Even though the authors insist on presenting "synthesized" visualizations of the partial representation of the city in literature, it is equally "striking" that they show little interest in establishing what factors might help to understand the discrepancy between the city in fiction and the city in actuality. This is due to the erosion of the human landscape, that of its authors and their biographies.

The preceding comment exposes one of the problems of 'distant reading', the difficulty of finding the individual actor in enduring organizational frameworks. As Ben Merriman notes, the prevailing conception of literary creation and culture is one which subsumes human action, when in fact books and cultural artefacts, unlike self-reproducing organisms, are the outcome of intentional human activity—and this is particularly noticeable when 'distant reading' makes analogies between changes in literary style and the evolution of biological species (Merriman 2015). Given the predominantly empiricist nature of the approach, identifying the patterns takes precedence over understanding them. An understanding of changes in literary activity demands consideration of authorial practices, of the institutions involved in publishing, and of how readers reacted to it. This study on the representation of London in 18th and 19th century fiction notably lacks biographies of the writers of the several thousand works included in the *corpus*. Examining this aspect in greater depth would contribute to explaining the greater or lesser diversity in the number of areas of London covered in their books, with research into, for example, the artistic and political movements to which the authors belonged, the social classes in which they moved, and the professions they practiced.

---

[2] Given the variety of terms available to describe the computational analysis of data (e.g., pattern recognition, machine learning), the term 'data mining' has been chosen as the most wide-ranging term to describe methods aimed at automating human cognitive functions like recognition and classification.

[3] Stanford University is one of the bodies behind the Stanford Digital Library Technologies Project, which took place between 1999 and 2004 and helped to create the Google search engine. The project was undertaken by computer science researchers with the aim of designing and implementing the infrastructure and services needed for collaboratively creating, disseminating, sharing, and managing information in a digital library context. See http://diglib.stanford.edu:8091/.

These limitations cannot be overcome, according to Marcel Lepper, by adopting the alternative of 'close reading', if only because the two approaches tend to see each other in stereotypical fashion: 'close reading', dedicated to the study of small *corpora*, is considered to be elitist and sectarian; 'distant reading' of archives and much larger *corpora*, is seen as levelling, technophilic and superficial (Lepper 2016, p. 155). He therefore suggests a third heuristic approach, which may bring them together and, according to Lepper, without losing sight of the changes in global text production and storage and the material-political-financial conditions that determine the accessibility of texts. Through the co-existence of quantitative approaches and fundamental qualitative principles it may be possible to build a bridge between the two approaches (Lepper 2016). For Murray G. Philips, Gary Osmond, and Stephen Townsend too, 'distant reading' becomes more productive when linked to the traditional aptitudes involved in 'close reading', namely the ability to contextualize entities and events and raise new issues to achieve more in-depth knowledge of the social phenomena thrown up by the patterns and hypotheses produced by the quantitative analysis of large data sets (Philips et al. 2015).

Another problem which has been detected in 'distant reading' is its "pseudo-scientific" nature, deriving from some of its procedures. Maurizio Ascari notes that the abundant use of abstract models imported from other disciplines—graphs (from Economic History), maps (from Geography) and trees (from Evolutionary Theory), to back up arguments, may lead to failures of verification (Ascari 2014). The author attributes the "theoretical fault" to "Moretti's choice to start from abstract models, the validity of which he subsequently tries to prove, rather than evolving theoretical models from the field of inquiry of popular literature itself" (Ascari 2014, p. 4). There is also a perceived scientific deficit, according to Katherine Bode, in the tendency not to explain the procedures for decanting data sets and to omit the size of the *corpora* thus produced (Bode 2018). In the above-mentioned study on London in literature, the number of works studied overwhelms, through omission, the number of authors who produced them.

One risk associated with the specialized work of 'distant reading' and digital humanities in the study of digitized cultural collections is to be found in the consequences, for producing scientific knowledge, of the connections between the academy and research centers on the one hand, and information technology industries involved in supplying databases and software on the other. Even though 'distant reading' and the digital humanities are not the first in the domain of the human and social sciences to embody the convergence of the university, research centers and the private sector, the ways in which these bodies interact calls into question the apparent lack of barriers to the openness and transparency of information. In connection with a debate on access to digitalized literary collections and the lack of opportunities to share the *corpus* between universities, Jonathan Reeve notes that it is not reasonable that a text which has long been out of copyright should cost researchers a certain amount, even with discounts [referring to a specific situation regarding access to works in collections held by Proquest] (Reeve 2015). The premise that works in the public domain should be the property of for-profit entities threatens "the ideals of science and academia" by hindering the "repeatability of experiments in text analysis" and making the cost of entry into the field "unnecessarily high" (Reeve 2015). Beneath the enthusiasm of the founder of 'distant reading' for the research potential of digital databases and computational tools (Moretti 2013, p. 212) lie inequalities in access to and use of information for academic purposes, rather than its equitable dissemination free of marketing imperatives.

A survey of the main potentialities and limitations of 'distant reading' reveals, first of all, that while on the one hand the ability to plot broad perspectives on large temporal and spatial scales, based on fairly extensive series of digitalized cultural artefacts, makes it possible to detect patterns and tendencies of continuity and change, on the other hand it misses the individual actors and an understanding of their actions and motives. Secondly, importing procedures from other branches of academic research, like Economic History and Biology, in order to link facts and theory, may not only enhance the scientific quality of that method of producing knowledge, it may also diminish it. Thirdly, in requiring not only computational tools but also objects of study partially owned by the

information technology industries (databases and data sets containing large lots of cultural works, many of them produced in the past), it contributes to yet another form of specialization in this market, and to rendering the process of producing scientific knowledge a competitive endeavor. To counter this tendency, arguments have been made in favor of making sharing agreements between researchers and other collaborative and non-profit initiatives (Reeve 2015; Hall 2017).

3. Cultural Analytics and Social Media as an Observatory for Digital Culture

The object of study in this section is the research program dubbed 'cultural analytics' by Lev Manovich in 2005, which has taken up an intermediate position between the humanities and social computation (Manovich 2015). This section ties in with the previous one in that the ideas examined here share with 'distant reading' the aim of analyzing massive sets of cultural objects. However, the macro nature of 'cultural analytics' amplifies them, in that its purpose is to study "everything created by everybody" and it aims to provide "larger inclusive scope combining professional and vernacular, historical and contemporary" (Manovich 2015, p. 7). 'Cultural analytics' focuses in particular on the observatory it sees as being unique for portraying the social world: the most popular social networks, their content, and the way users interact within them. It thus constitutes another approach relevant to the ongoing exploration of the possibilities and limitations of the study of culture using big data and computational engineering.

Research undertaken using this approach, based in the Software Studies Initiative[4] research center, takes as its starting point a set of issues which seek to use data generated by big data sources to interpret the significant cultural changes driven by the process of digitalization, the central locus of the digital in the mediation of culture and the predominance of images to assert identity and communicate with others. Its mentor's biography sheds light on the outlines of the project, his interest in images and visual culture, and the propensity to aestheticize the data. Before moving from Russia to the United States, in the 1980s, he studied mathematics, programming, visual arts, and architecture.

In a recent article, the mentor of 'cultural analytics' suggests that a field of research be set up, to be called 'computational media studies' (Manovich 2018). This would be a counterpoint to the phenomenon of 'media analytics', which includes computational analysis, greatly intensified since 2010, of the content of social media and of individual and collective online activity by firms, including those involved in the culture industry in various stages of dissemination and marketing of cultural goods (websites devoted to sales and search engines, amongst others). The purpose of 'media analytics', a kind of polling and marketing center deployed on digital platforms, is systematically to expose users and consumers to new options for consumption, both adapting to and going beyond individual histories of tastes and interactions with other users. The data collected from real-time algorithmic analyses are not published, and have been commodified.

The proposal for 'computational media studies' seeks to appropriate the data analysis methods used by the culture industry to investigate the effects of those operations. In overall terms, it is an attempt to widen the quantitative study of diversity and cultural standards on the basis of massive data sets relating to culture: films, books, and songs, but also individual posts, messages, and images shared on multiple platforms. In sum, it is suggested that culture studies should intensify their use of research methodologies long practiced by other disciplines, analyzing posts, tweets, and other forms of expression to capture opinions on politics and culture as well. Thus, in the same way that some writers observed the effect of cultural homogenization on North American society in the 1940s and 1950s as a result of the cinematographic production of the major film conglomerates, Manovich suggests that today one might ask: do the computational recommendation systems activated by Amazon, YouTube, Netflix, and Spotify help diversify users' choices of films, music, videos, and apps, or do they rather direct them to top lists and other rankings?

---

[4] Founded in 2007, was renamed Cultural Analytics Lab in 2016.

The analogy raises stimulating questions and assumptions. However, and herein lies one of the restrictions of 'cultural analytics', it is difficult to envisage how it will contribute to "develop[ing] more explicit and detailed theories than we use normally" (Manovich 2015, p. 12). Avoiding issues which have the greatest critical potential, in favor of the fixation on the tools and methods imported from computer science, is one of the limitations of this way of approaching cultural phenomena, all the more prominent because its promoter states these as his aims. In this connection it is revealing to note how the creator of 'cultural analytics' admits to the obstacles in his research program. Specifically, he declares that the functions of digital platforms shape and condition cultural production, so the content found on social networks, and the types of participation in them, are formed according to the available range of features and interfaces of the technologies accessible for creating, editing, and sharing content (Srnicek 2017).

An admission of this sort may implicitly point to the relativization of the idea that social networks are powerful tools for expanding unlimited creativity and for knowing and understanding it. However, and this is one of the reasons for the reduced impact of 'cultural analytics', the implications of the limitations of social media for how culture is generated and received—which show how technology exerts control by means of the standards it establishes—are not questioned in the context of theoretical and conceptual frameworks. Theory, as Gary Hall points out, is postponed, this being due to the lack of involvement with the problematic nature of social facts as objects of study. Nevertheless, theory reasserts itself, in so far as any methodology carries a theory within it; and if it is not made explicit and specified, then there is a risk that the theory will be reductionist (Hall 2014).

The empiricist slant expands the disconnect between the research program's intentions—"follow imaginations, opinions, ideas, and feelings of hundreds of millions of people", in "Trending: The Promises and the Challenges of Big Social Data" (Manovich 2012)—and its effective ability to capture the symbolic. This disconnect increases to the extent that 'cultural analytics' sees itself as being different to social computation, because it prioritizes the cultural and takes the sociology of culture, and one of the writers who made the longest-lasting contributions to it, Pierre Bourdieu, as its main points of reference (Manovich 2015). Invoking Bourdieu seems inconsequential, in as much as 'cultural analytics' does not develop the proposed conceptual grammar to capture the relationships between economic and cultural power or between the individual and society. Nor does it incorporate the possibilities for examining the issues surrounding cultural capital, as contained in the empirical research program developed by Bourdieu. Likewise, it is far removed from the view he took of how theory is constructed, stressing the importance for the social sciences of taking into account the material and symbolic properties of the social universe. In addition, the French sociologist argued for a dialectical relationship between theory and practice. In this connection, social theories which distance themselves from empirical work, and empirical research which takes place as if it could be done without theory, are both equally unsatisfactory.

Using 'cultural analytics' and 'computational media studies' for practical applications—for example, forecasting—is another feature of these suggested ways of interpreting culture which overrides more far-reaching and in-depth methods of investigation. This can be seen in an assessment of the results of study on aesthetic reception on the web, based on 9 million Flickr images with Creative Commons Licenses. Having established that the discrepancy between the high quality of "valuable photographers" contributions and the meagre interest in viewing their images was due to the photographers' low levels of participation in online interaction, the authors suggested creating an algorithm to find "unpopular" images, which might help the creators to find audiences for their work. For the promoter of 'cultural analytics', this application is an example of the potential of 'computational media studies' "to go beyond generating descriptions and 'critique' of cultural situations by offering constructive solutions that can change these situations" (Manovich 2018, p. 484).

The problems and contradictions of 'cultural analytics' become particularly noticeable when researchers look at social networks using a qualitative methods-based approach, instead of the quantitative focus and the fixation on detecting patterns by means of computational analysis.

An example of this is DeviantArt, one of the objects studied by 'cultural analytics' (Yazdani et al. 2017). DeviantArt is described as a digital platform "for emerging and established artists to exhibit, promote, and share their works with an enthusiastic, art-centric community"[5]. This initiative is particularly relevant to this section for a number of reasons. Not only does it host millions of images of works of art, it also contains an art world which may be analyzed in terms of the classic issues in the sociology of culture, such as iconographic analysis and the construction of artistic identity. Yazdani et al. see it as a "challenging" object for qualitative approaches. At the same time, they note that a source which they consider to be so diverse in terms of content, styles and techniques, is "good motivation" for the use of computational methods (Ibid.). The reason for this is that it would make it possible to detect and analyze various temporal patterns, between 2001 and 2010, in a sample of 270 thousand images of works of 'Traditional Art' and 'Digital Art' (out of a total sample of one million items), a quantity much higher than that of any museum collection for the same period (Ibid.).

The novelty which they assign to the results—like the much faster growth of 'digital art' compared to that of 'traditional art'—could not, in their view, have been "predicted using existing qualitative work in art history or media theory". Underlying this, there remains a latent question: "In the early 21st century, [when] the volume of digital online content and user interactions allows us to think of a possible 'science of culture' "(Manovich 2015, p. 10), is there still a place for humanists and social scientists? Even though the opposite is insinuated, the answer is in the affirmative, the human and social sciences are necessary, if only to research what 'cultural analytics' neglects in objects of study like DeviantArt. On the one hand, the exploration of the themes and interconnections of all the imaginaries of contemporary visual culture which flow into this giant digital curio cabinet: films, fashion catalogues, Japanese anime films, TV series which combine the fantastic with horror, and cartoons, among others. On the other hand, looking at DeviantArt at the point where web worlds intersect with art worlds shows how the digital atmosphere, apparently ruled by a 'creativity consensus', makes it possible to extend the analysis of traditional issues in the sociology of art and culture, such as recognition, artistic identity, and intellectual property, to a web environment and a social network dedicated to art (Perkel 2011). A site apparently based on the values of community and egalitarian participation is, in the final analysis, sought after by artists to canvas for capital in the form of visibility, at the same time as it generates inequality—as illustrated by the existence of a section with exclusive access for paying members, even though it was created to allow for feedback and mutual help between registered artists in the network (Perkel 2011).

There is the additional problem of the investment in the aestheticization of data, another of the procedures of 'cultural analytics' which gives away the limited role of theory within it. Emphasis is placed on the visualization of image features like tonality and brilliance, automatically identified by computer, which leads to the aestheticized arrangement of the data. Visualizations become a research routine, compensating for the lack of a theoretical framework by reinforcing technological and methodological procedures. Images serve basically as the raw material for other images, with the particularity that 'cultural analytics' is not an artistic movement—if it were, the appropriation and assemblage would have a different meaning. The central role of data in 'cultural analytics', a system which binds itself with sociology, produces an effect of naivety: the aestheticization of social facts, according to writers who have studied the prominence of the cultural among those social facts, is a consequence of the growth in consumption and cultural production in modern society (Connor 1996; Morató 2003).

While it may be acknowledged that 'cultural analytics' and 'computational media studies' are fit to identify patterns, relationships, and tendencies, and to suggest hypotheses on digital culture based on vast sets of cultural data, their main limitation is the fact that they are confined to data-driven research. The expression " 'surface data' about lots of people", suggested by the mentor of 'cultural

[5] See https://about.deviantart.com/.

analytics' to explain the object he is interested in, exempts Manovich, in a way, from having to think separately about how he might develop theories and concepts (Manovich 2012, p. 2). Contrary to what he states in some articles, the qualitative approach is hardly encouraged, and is systematically devalued. What prevails is a fixation on extracting nuggets from collections of data according to certain parameters, and on recognizing patterns and trends, as if actions took place only in an instant, and to the detriment of a more comprehensive understanding of interaction.

4. Looking for New Sources: The Statistics of Culture and the Incorporation of Big Data

This section analyses official EU culture statistics. As shown in the Introduction, these statistics have some relevant characteristics in common with the two previous approaches, apart from the fact they refer to the cultural field. First, they have increasingly focused on the use of big data, seeking to exploit the possibilities for handling the very large sets of quantitative data generated by these sources. Secondly, statisticians are also interested in carrying out longitudinal analyses and identifying patterns and trends, making abundant use of visual means of viewing data (charts, tables, and diagrams). In addition, the interest of official statisticians, including those in the cultural field, in using big data sources points to the emergence of new ways of producing statistical information which are symptomatic of changes in the epistemic values which guide this activity (Pietsch 2013). There are similarities here which make it relevant to assess the extent and the limitations of the use of big data in the field of statistics and the most recent developments in this area.

Since 2013 Eurostat, the body responsible for the statistics of the European Union (EU), has shown its interest in using big data sources in a more systematic way[6]. It launched a number of exploratory projects using databases generated, for example, by smartphones, Google searches and Wikipedia views, denoting a desire to expand and renew its sources of statistical information in various areas. The initial results of a pilot project combining big data and cultural topics emerged in the third edition of *Culture Statistics*, published in 2017. The project is based on views of pages of Wikipedia articles related to established places like world heritage sites. According to the editors of *Culture Statistics* and those most directly involved in this initiative, there are two reasons why this project represents an opportunity for EU statistics on culture. First, since all use of computer systems leaves digital traces of activity, these start to be of interest as a source of official statistics because they are more up to date and more detailed and reliable, and they are "direct measurements of phenomena", and thus go beyond the reach of the traditional "indirect reporting by a survey respondent" (Signorelli et al. 2016, p. 2). Secondly, the increasing resort to Wikipedia as a knowledge source makes it a "relevant big data source for producing official statistics" (Culture Statistics 2016, p. 32). The volume of information contained in the digital encyclopedia has been growing (in 2016, it contained 39 million articles in 246 languages), as has its audience: in 2015, 45% of individuals aged between 16 and 74 viewed its pages, particularly in the 16 to 24 age group (Culture Statistics 2016, p. 32).

The published results of the exploratory big data project can be seen in two diagrams containing methodological notes, but lacking an interpretation. This option seems to be in line with (i) Eurostat's reservations regarding the exploratory nature of big data statistics, which have not yet reached full maturity[7]; (ii) the lack of definition regarding topics and phenomena which information obtained through Wikipedia can capture. Consulting pages of the digital encyclopedia is justified on the one hand as "an attempt to assess the population's interest in cultural heritage" (Culture Statistics 2016, p. 9) and on the other as "a measure of popularity of the sites or a measure of 'cultural consumption' of world heritage"[8].

---

6 See ESS Scheveningen Memorandum on 'big data' (September 2013).
7 In http://ec.europa.eu/eurostat/web/experimental-statistics/world-heritage-sites.
8 In http://ec.europa.eu/eurostat/web/experimental-statistics/world-heritage-sites.

While it is limited in terms of location (European Union and non-European Union) and some languages (English, French, Spanish, and Italian), the information on views of world heritage sites is an indicator of the particular attractiveness of ancient and monumental spaces, including those close to, or which are a part of, cosmopolitan modernity. Thus, the top site in the group of 20 most-highlighted places in the corresponding Wikipedia pages is 'Paris, Banks of the Seine' (6,186,339 views), followed by the 'Historic Center of Rome' (5,759,186 views)[9]. Even in the subsection containing the 'Top 5 World Heritage Sites in number of page views of related Wikipedia articles by language' indicator, the two sites mentioned are present in all the most-visited pages in the various languages.

In addition to the lack of a more in-depth interpretation, it is not clear what the motives are for the interest and perambulation through the pages of Wikipedia relating to world heritage sites and the places specifically mentioned above. A search by the name of a place does not always lead to a physical visit or purchase, even though a correlation is sometimes found, as has already been noted (Demunter 2017)[10]. In the specific case of the number of views of Wikipedia related to travel destinations, those involved in promoting the incorporation of big data in official statistics recognize all of the following at the same time: the potential contribution to tourist flows prediction; the impossibility of ascertaining how many among those who search for a place intend to visit it, and within this sub-group how many actually undertake the physical journey; the need to refine the analysis so as to make solid correlations with tourist visits (Demunter 2017). To this overall assessment we should add that the sociological study of how Wikipedia and its inner politics operate has shown that there are multiple biases present in the production of this platform and in the database itself. These call into question the whole notion that big data can impartially and objectively portray social reality (Jemielniak 2014; Adams and Bruckner 2015).

Eurostat's incursion into big data sources, and the exploratory work undertaken on the potential for incorporating them into official statistics, including those on culture, demonstrates a tendency to value the power of big data as reflections of real life. In a way this is an extension of the idea that statistics are 'a mirror of society', a label which cuts across many online presentations of public entities responsible for producing statistics. Work on incorporating big data has forged a vision of a new operating model for official departments and bodies devoted to collecting and handling data. Among the many transformations which underpin the architecture of the "statistical office of the future", and according to those who have worked in Eurostat on incorporating big data, the following stand out. Work on automatically generated data flows replaces surveys and censuses, with production shifted to certifying the data. In line with this approach, 'data collection designers' are upstaged by 'product designers', who are occupied and focused on 'nowcasting', the shortest-range and fastest form of forecasting, and "forecasting" itself. Compared to these tasks, producing descriptive indicators becomes something of an anachronism (Skaliotis 2015; Kotzeva 2015).

Other observers of statistics take a more detached and cautious view of this model, stressing the importance of not confusing it with 'computer science' and its associated procedures (e.g., sorting, adding, selecting, matching, concatenating, and aggregating) (Hand 2015, 2018). This position offers a glimpse of a defense of the specific nature of statistics in the face of the assumption that it is less suited to the universe of big data (Borne 2013). The rigor which some computing engineers see in big data sources, because they regard them as being direct measurements of social phenomena, as mentioned above, is questioned by the above-quoted statisticians, who emphasize how important it is to distinguish correlation from causality, and to keep in mind the relevance of statistical inference. The broader framework for this discussion is that of a "shift in epistemic values regarding the aims of modeling", marking the transition from 'parametric' to 'non-parametric modeling', from data to

[9] Data available online: http://ec.europa.eu/eurostat/web/experimental-statistics/world-heritage-sites.

[10] Demunter mentions an article by researchers in Epidemiology, which presents as "the first comparison to evaluate Google, Twitter, and Wikipedia as possible data sources for influenza surveillance against a common gold standard" (Sharpe et al. 2016).

algorithmic models or from model-based to model-free approaches (Breiman 2001; Pietsch 2013). Wolfgang Pietsch observes that, among other differences, 'non-parametric modeling' (the most recent type) "is geared almost exclusively to prediction and manipulation and rarely to understanding in terms of general laws or rules. By contrast, parametric modeling usually emphasizes understanding. (. . .). Presumably, this shift in epistemic values is at the root of the mentioned divide between the different 'cultures' of statistical modeling" (Pietsch 2013)[11].

At the same time, within EU official statistics circles there is a noticeable perception that the transition to a statistical department along the above lines raises certain issues, related, among other things, to methodology and the ties to entities which manage big data sources. The various risks to the quality of information are identified in a document produced by Eurostat (Tourism Statistics: Early Adopters of 'Big Data'? 2017). By way of illustration, consider the use of data generated by mobile phones for the purpose of studying mobility between cities: not only does smartphone ownership vary by country and region, but some functions can be periodically deactivated by users as well; the resulting selection effect is comparable to the problem of high rates of non-responses to sample-based statistical surveys. In addition, it is possible that partial control of the data by the mobile network operators calls into question the rigor and independence of official statistics, which, as we have seen in the new "statistical office of the future" model, tend mainly to arrogate to themselves the role of certifying externally produced data.

5. Conclusions

This article has sought to analyze the implications of using big data in procedures, theories, concepts, and discoveries produced in research in the human and social sciences, in connection with the study of cultural phenomena and artefacts which are increasingly stored in digital form. Taking as its starting point two approaches which are based on synergies with computational engineering, 'distant reading' and 'cultural analytics', the article has sought to demonstrate the possibilities and limitations of these approaches to producing scientific knowledge. The analysis was complemented by an investigation of the statistics of culture, on account of their expanding interest in big data as alternative sources. The mosaic of these research methods demonstrates how the growing wealth of data generated by the new information technologies is agitating and incentivizing the human and social sciences into adopting a new way of producing scientific knowledge. This method is increasingly guided by the 'data science' model, based on data mining and machine learning, where the starting points for new research projects and questions are more and more 'born digital'. With this shift, new epistemic values are being asserted; in the specific case of statistics, it has been observed that 'non-parametric modeling', related to algorithmic models and model-free approaches, is geared almost exclusively to prediction and manipulation.

It is acknowledged that in using massive data sets, 'distant reading' and 'cultural analytics' offer the ability to identify patterns and trends and raise new issues and hypotheses which are not fixed at the outset. At the same time, the conclusion is reached that there are significant limitations and problems with the methods analyzed, which at one moment may cover the cultural production of several centuries and at another moment the far larger torrent of content permanently generated on social networks. A principal limitation is the fact that they are unable to capture individual actors and the meanings and purposes of their actions which produce cultural and artistic creations. Another key problem of these methods is the disconnect between, on the one hand, the investment in a quantitative approach and the aestheticized arrangement of data and, on the other, the undervaluation of theory, as a result of lack of involvement with the problematic nature of society, and even the disparaging of the scope of qualitative methods.

---

[11] In this debate, initiatives such as the MOSAIC Project favor combined and not separate teaching of modeling, statistics, computation and calculus. See http://mosaic-web.org/.

We can thus see how this overall assessment is part of a wider discussion in research circles on topics and themes beyond culture, and on the relationship between the human and social sciences and the big data phenomenon, on account of the methodological and epistemological consequences of the data science model and its defense of the primacy of observational-inductive procedures. As was mentioned in the Introduction, some authors argue that the human and social sciences gain by welcoming the convergence with computational engineering, because this cross-fertilization embodies within it the ability to translate information into knowledge and to apply the skills of synthesizing, contextualizing and reflecting theoretically in order to articulate and understand the large quantity and variety of results (McFarland et al. 2015; Halford and Savage 2017).

Bearing in mind some of the examples mentioned in the 'distant reading' and 'cultural analytics' approaches, it is noticeable that analyses of extensive *corpora* based on computing may point to clues and trends which are significant for research into culture, on condition that room is left for contextual knowledge, the ability to situate objects of study historically and sociologically, and discussion of the symbolic meanings of large amounts of artefacts and discourses. But the scenario is complex and challenging for the dialogue between the human and social sciences and computational engineering, particularly for some of the human sciences, by virtue of their historical lack of proximity with quantitative approaches. Specifically in connection with the complementarity of quantitative methodologies based on computer science and qualitative analyses—e.g., "the performative theoretical interpretations that have long been a prominent feature of the humanities" (Hall 2014, p. 27)—some writers regard this as a non-viable combination, in the sense that one would be faced with incommensurate entities, not having a common measure. For those who equate machine learning with a "theory of learning", the latter, together with statistics, have the potential to be "philosophical interlocutors for the humanities, helping us to think about interpretation on a scale where variation and uncertainty are central problems" (Underwood in Dinsman 2016, p. 13).

This scenario is challenging and perplexing, and points to the need for new developments. It is important to mention, albeit in summary fashion, a significant wave of authors who have argued for mixed methods approaches and the incorporation of computational methods into existing social science disciplines, including the sociology of culture. What is at issue here is the aim of achieving balance between theory and data in cultural sociology (Ghaziani 2009). As this article concludes, the tendency to demote theory is one of the main problems of approaches which prioritize computational methods and the identification of patterns. Christopher Bail mentions a need to take advantage of one of "most promising elements of the big data movement (. . .) [:] so much of the qualitative data that has been collected is longitudinal (. . .) [and] the most pressing questions in cultural sociology concern change over time" (Bail 2014, p. 474). Several others welcome the possibilities opened up by incorporating computational methods into the sociology of culture (DiMaggio et al. 2013; Bail 2014; Muller et al. 2016; Abramson et al. 2017; Baumer et al. 2017; Nelson 2017). They agree in arguing for the importance of incorporating the 'topic modeling' method in content analysis (e.g., newspaper and television transcript archives), for its potential usefulness in organizing, searching and understanding data sets on large-scale cultural artefacts which span an extended period of time. In particular, in studying the contentious debate on government support for the arts, using 8000 articles published in five U.S. newspapers between 1986 and 1997, Paul DiMaggio and others argue that topic models "may be perplexing", but "render operational such concepts as frames, polysemy, heteroglossia, and the relationality of meaning" (DiMaggio et al. 2013, p. 603).

Some of the contributions mentioned above stand out for their particular combination of grounded theory and topic modeling. In a study of analysis of the same dataset using two separate approaches, grounded theory and statistical topic modelling, Eric P.S. Baumer and others detected several points of convergence and divergence. Among other results they found that, even though statistical topic modeling, unlike grounded theory, ignores contextual information, it is able to "identify patterns that, at some level, align with those found by human researchers" (Baumer et al. 2017, p. 1406). These authors argue that the rigorous combination of these different approaches demands that

researchers who use them keep in mind the different epistemological traditions from which they come: grounded theory draws on the interpretivist tradition, while most computational techniques derive from the positivist tradition, which assumes the possibility of an objective physical and social world. With a view to incorporating computational methods into inductive sociological content analysis, Laura K. Nelson in turn puts forward a 'methodological framework' which she calls "computational grounded theory", using the following argument: given that such a procedure "provides a method to calculate how prevalent or representative each pattern is within the larger corpus", it is more valid than traditional grounded theory, "which asks the reader to simply trust the representativeness of particular examples or quotes" (Nelson 2017, p. 32).

Finally, the overall analysis and assessment made in this article point to a need, in research in the human and social sciences, to encourage reflexive thought on digital media and their complex relationship to culture and other aspects of the societal world, like politics and economics. The concern with better ways of using digital platforms and the mass of big data they contain risks blurring critical thinking, thereby reducing the capacity of the human and social sciences to denaturalize the social world.

Funding: This research received no external funding.

Acknowledgments: This text was translated with the support of FCT, through its strategical project UID/SOC/50013/2013.

Conflicts of Interest: The author declares no conflict of interest.

References

Abbott, Andrew. 2004. *Methods of Discovery: Heuristics for the Social Sciences*. New York: W. W. Norton & Company.

Abramson, Corey M., Jacqueline Joslyn, Katharine A. Rendle, Sarah B. Garrett, and Daniel Dohan. 2017. The Promises of Computational Ethnography: Improving Transparency, Replicability, and Validity for Realist Approaches to Ethnographic Analysis. *Ethnography* 19: 254–84. Available online: https://journals.sagepub.com/doi/full/10.1177/1466138117725340 (accessed on 29 September 2018). [CrossRef]

Adams, Julia, and Hannah Bruckner. 2015. Wikipedia, Sociology, and the Promise and Pitfalls of 'Big Data'. *Big Data & Society* 2. Available online: http://journals.sagepub.com/doi/abs/10.1177/2053951715614332 (accessed on 15 July 2018).

Ascari, Maurizio. 2014. The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres. *Genre* 47: 1–19. [CrossRef]

Bail, Christopher A. 2014. The Cultural Environment: Measuring Culture with Big Data. *Theory and Society* 43: 465–82. Available online: https://www.jstor.org/stable/43694728?seq=1#metadata_info_tab_contents (accessed on 19 November 2018). [CrossRef]

Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing Grounded Theory and Topic Modeling: Extreme Divergence. *Journal of the Association for Information Science and Technology* 68: 1397–410. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23786 (accessed on 17 November 2018). [CrossRef]

Beer, David. 2016. How Should We Do the History of Big Data? *Big Data & Society* 3. Available online: http://journals.sagepub.com/doi/10.1177/2053951716646135 (accessed on 18 June 2018).

Bode, Katherine. 2018. *A World of Fiction. Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press.

Boldizzoni, Francesco. 2011. *The Poverty of Clio: Resurrecting Economic History*. Princeton: Princeton University Press.

Borne, Kirk. 2013. Statistical Truisms in the Age of 'Big Data'. Available online: http://www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html (accessed on 31 August 2018).

Braudel, Fernand. 1972–1973. *The Mediterranean and the Mediterranean World in the Age of Philip II*. New York: Harper and Row. First published 1949.

Breiman, Leo. 2001. Statistical Modeling: The Two Cultures. *Statistical Science* 16: 199–231. Available online: http://www2.math.uu.se/~{}thulin/mm/breiman.pdf (accessed on 1 September 2018). [CrossRef]

Carrigan, Mark. 2018. The Evisceration of the human under digital capitalism. In *Responses to Post-Human Society: Ex Machina*. Edited by Ismael Al-Amoudi and Jamie Morgan. London and New York: Routledge, pp. 165–81.

Christin, Angèle, and Olivier Donnat. 2014. Pratiques Culturelles en France et aux États-Unis: Éléments de Comparaison 1981–2008. *Culture Études* 1: 1–16. Available online: https://www.cairn-int.info/article-E_CULE_141_0001--french-and-american-cultural.htm (accessed on 14 July 2018). [CrossRef]

Connor, Steven. 1996. Cultural Sociology and Cultural Sciences. In *The Blackwell Companion to Social Theory*. Edited by Bryan S. Turner. Oxford: Blackwell Publishers, pp. 340–68.

Culture Statistics. 2016. Available online: http://ec.europa.eu/eurostat/documents/3217494/7551543/KS-04-15-737-EN-N.pdf (accessed on 2 June 2018).

Demunter, Christophe. 2017. Tourism Statistics: Early Adopters of Big Data? Paper presented at the 'Sixth UNWTO International Conference on Tourism Statistics. Measuring Sustainable Tourism', Manila, Philippines, June 21–24; Available online: http://cf.cdn.unwto.org/sites/all/files/pdf/demunter_session5_conf2017manila_central_paper.pdf (accessed on 4 July 2018).

DiMaggio, Paul, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41: 570–606. [CrossRef]

Dinsman, Melissa. 2016. The Digital in the Humanities: An Interview with Ted Underwood. Available online: https://lareviewofbooks.org/article/digital-humanities-interview-ted-underwood/#! (accessed on 10 September 2018).

Ghaziani, Amin. 2009. An "Amorphous Mist"? The Problem of Measurement in the Study of Culture. *Theory and Society* 38: 581–612. Available online: https://www.jstor.org/stable/40345672?seq=1#page_scan_tab_contents (accessed on 23 October 2018). [CrossRef]

Halford, Susan, and Mike Savage. 2017. Speaking Sociologically with Big Data: Symphonic Social Science and the Future for 'Big Data' Research. *Sociology* 51: 1132–48. Available online: http://journals.sagepub.com/doi/abs/10.1177/0038038517698639 (accessed on 8 June 2018). [CrossRef]

Hall, Gary. 2014. Towards a Post-Digital Humanities: Cultural Analytics and the Computational Turn to Data-Driven Scholarship. Author Post-Print (Accepted) Deposited in CURVE January 2014. First published 2013. Available online: https://curve.coventry.ac.uk/open/file/c5331c38-e060-4756-8582-0719f07295f2/1/post-digital%20humanities.pdf (accessed on 16 July 2018).

Hall, Gary. 2017. The Inhumanist Manifesto: Extended play. The Techne Lab, University of Colorado. Available online: http://art.colorado.edu/research/Hall_Inhumanist-Manifesto.pdf (accessed on 2 September 2018).

Hand, David J. 2015. Big Data. Promises and Pitfalls. Paper presented at the Conference 'Policy-Making in the 'Big Data' Era, Opportunities and Challenges', Cambridge, UK, June 15–17; Available online: https://www.youtube.com/watch?v=Yz9_JGezoFk (accessed on 7 July 2018).

Hand, David J. 2018. Statistical Challenges of Administrative and Transaction Data. *Journal of the Royal Statistics Society Series A—Statistics in Society* 181: 555–78. Available online: https://rss.onlinelibrary.wiley.com/doi/10.1111/rssa.12315 (accessed on 7 July 2018). [CrossRef]

Heuser, Ryan, Franco Moretti, and Erik Steiner. 2016. The Emotions of London. The Emotions of London. Stanford Literary Lab. Pamphlet 13. Available online: https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf (accessed on 20 July 2018).

Jemielniak, Dariuz. 2014. *Common Knowledge? An Ethnography of Wikipedia*. Stanford: Stanford University Press.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks: Sage Publications.

Kotzeva, Mariana. 2015. New Frontiers for Official Statistics. Paper presented at the 'European Data Forum', Luxembourg, November 16–17; Available online: http://2015.data-forum.eu/sites/default/files/KOTZEVA_SEC.pdf (accessed on 4 July 2018).

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert Lazlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and et al. 2009. Life in the Network: The Coming Age of Computational Social Science. *Science* 323: 721–23. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/ (accessed on 1 September 2018). [CrossRef] [PubMed]

Lepper, Marcel. 2016. Big Data, Global Villages. *Philological Encounters* 1: 131–62. Available online: http://booksandjournals.brillonline.com/content/journals/10.1163/24519197-00000006 (accessed on 9 July 2018).

Lupton, Deborah. 2015. *Digital Sociology*. London: Routledge.

Lupton, Deborah. 2016. *The Quantified Self*. Cambridge and Malden: Polity Press.

Manovich, Lev. 2012. Trending: The Promises and the Challenges of Big Social Data. Available online: http://dhdebates.gc.cuny.edu/debates/text/15 (accessed on 18 June 2018).

Manovich, Lev. 2015. The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. Available online: http://manovich.net/content/04-projects/088-cultural-analytics-social-computing/cultural_analytics_article_final.pdf (accessed on 15 June 2018).

Manovich, Lev. 2018. 100 Billion Data Rows per Second: Media Analytics in the Early 21st Century. *International Journal of Communication* 12: 473–88. Available online: http://manovich.net/index.php/projects/media-analytics (accessed on 1 July 2018).

Martins, Hermínio. 2011. *Experimentum Humanum. Civilização Tecnológica e Condição Humana*. Lisboa: Relógio D'Água.

McFarland, Daniel, Kevin Lewis, and Amir Goldberg. 2015. Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *The American Sociologist* 47. Available online: https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/amsoc.pdf (accessed on 8 June 2018). [CrossRef]

Merriman, Ben. 2015. A Science of Literature. *Boston Review. A Political and Literary Review*. Available online: http://bostonreview.net/books-ideas/ben-merriman-moretti-jockers-digital-humanities (accessed on 12 July 2018).

Morató, Arturo Rodríguez. 2003. The Culture Society: A New Place for the Arts in the Twenty-First Century. *The Journal of Arts Management, Law, and Society* 32: 245–56. Available online: https://www.tandfonline.com/doi/abs/10.1080/10632920309596978 (accessed on 3 September 2018).

Moretti, Franco. 1998. *Atlas of the European Novel 1800–1900*. London and New York: Verso. First published 1997.

Moretti, Franco. 2005. *Graphs, Maps, Trees. Abstract Models for Literary History*. London and New York: Verso.

Moretti, Franco. 2013. *Distant Reading*. London and New York: Verso.

Muller, Michael, Shion Guha, Eric P. S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. Paper presented at the 19th International Conference on Supporting Group Work, Sanibel Island, FL, USA, November 13–16; Available online: https://dl.acm.org/citation.cfm?doid=2957276.2957280 (accessed on 18 November 2018).

Nelson, Laura K. 2017. Computational Grounded Theory: A Methodological Framewok. *Sociological Methods & Research*, 1–40. Available online: https://journals.sagepub.com/doi/abs/10.1177/0049124117729703 (accessed on 19 November 2018).

Ostrow, Saul. 2014. Introduction. Aby's Warburg: Culture's Image Network. In *Art History as Cultural History. Warburg's Projects*. Edited by Richard Woodfield. London and New York: Routledge, pp. 1–6.

Outhwaite, William. 1996. The Philosophy of Social Science. In *The Blackwell Companion to Social Theory*. Edited by Bryan S. Turner. Oxford: Blackwell Publishers, pp. 83–106.

Outhwaite, William. 2005. Interpretativism and Interactionism. In *Modern Social Theory. An Introduction*. Edited by Austin Harrington. Oxford: Oxford University Press, pp. 110–31.

Perkel, Daniel. 2011. Making Art, Creating Infrastructure: DeviantArt and the Production of the Web. Ph.D. dissertation, University of California, Berkeley, CA, USA. Available online: http://people.ischool.berkeley.edu/~{}dperkel/diss/DanPerkel-dissertation-2011_update.pdf (accessed on 11 July 2018).

Philips, Murray G., Gary Osmond, and Stephen Townsend. 2015. A Bird's-Eye View of the Past: Digital History, Distant Reading and Sport History. *The International Journal of the History of Sport* 32: 1725–40. Available online: https://www.tandfonline.com/doi/full/10.1080/09523367.2015.1090976 (accessed on 14 July 2018). [CrossRef]

Pietsch, Wolfgang. 2013. Big Data—The New Science of Complexity. Available online: http://philsci-archive.pitt.edu/9944/1/pietsch-bigdata_complexity.pdf (accessed on 12 July 2018).

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 31. Available online: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1 (accessed on 26 June 2018). [CrossRef]

Reeve, Jonathan. 2015. A Proposal for Data Sharing Protocol. Available online: http://jonreeve.com/2015/03/proposal-for-a-corpus-protocol/ (accessed on 10 June 2018).

Savage, Mike, and Roger Burrows. 2007. The Coming Crisis of Empirical Sociology. *Sociology* 41: 885–99. Available online: http://journals.sagepub.com/doi/pdf/10.1177/0038038507080443 (accessed on 11 June 2018). [CrossRef]

Sharpe, J. Danielle, Richard S. Hopkins, Robert L. Cook, and Catherine W. Striley. 2016. Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. *JMIR Public Health Surveill* 2. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5095368/ (accessed on 3 July 2018). [CrossRef] [PubMed]

Signorelli, Serena, Fernando Reis, and Silvia Biffignandi. 2016. What Attracts Tourists While Planning for a Journey? An Analysis of Three Cities through Wikipedia Page Views. Paper presented at the '14th Global Forum on Tourism Statistics', Venice, Italy, November 23–25; Available online: https://www.researchgate.net/publication/310605164_What_attracts_tourists_while_planning_for_a_journey_An_analysis_of_three_cities_through_Wikipedia_page_views (accessed on 3 July 2018).

Skaliotis, Michail. 2015. Big data in the European Statistical System. Paper presented at the Conference by STATEC and EUROSTAT 'Savoir pour Agir: La Statistique Publique au Service des Citoyens', Luxembourg, October 20; Available online: https://statistiques.public.lu/fr/agenda/detail-agenda/2015/10/SKALIOTISWorldstatsdaySTATEC.pdf (accessed on 3 July 2018).

Srnicek, Nick. 2017. *Platform Capitalism*. Cambridge and Malden: Polity Press.

Tourism Statistics: Early Adopters of 'Big Data'? 2017. Available online: http://ec.europa.eu/eurostat/documents/3888793/8234206/KS-TC-17-004-EN-N.pdf (accessed on 12 July 2018).

Yazdani, Mehrdad, Jay Chow, and Lev Manovich. 2017. Quantifying the Development of User-Generated Art during 2001–2010. *PLoS ONE*. Available online: http://journals.plos.org/plosone/article/related?id=10.1371/journal.pone.0175350 (accessed on 26 July 2018). [CrossRef] [PubMed]