# 3. Cultural Analytics, Social Computing and Digital Humanities

*Lev Manovich*

## Social Computing vs. Digital Humanities

I define Cultural Analytics as the analysis of massive cultural data sets and flows using computational and visualization techniques. I developed this concept in 2005, and in 2007, the Software Studies Initiative[1] research lab was established to start working on Cultural Analytics projects.

Our work is driven by a number of theoretical and practical questions: What does it mean to represent 'culture' by 'data'? What are the unique possibilities offered by the computational analysis of large cultural data in contrast to qualitative methods used in humanities and social sciences? How can quantitative techniques be used to study the key cultural form of our era – interactive media? How can we combine computational analysis and visualization of large cultural data with qualitative methods like 'close reading'? Put another way, how can we combine the analysis of larger patterns with the analysis of individual artefacts and their details? How can computational analysis do justice to variability and diversity of cultural artefacts and processes, rather than focusing on the 'typical' and 'most popular'?

Eight years later, the work of our lab has become a tiny portion of a very large body of research. Thousands of researchers have published tens of thousands of papers analysing patterns in massive cultural data sets. This is data describing activity on most popular social networks (Flickr, Instagram, YouTube, Twitter, etc.), user-created content shared on these networks (tweets, images, video, etc.), and users' interactions with this content (likes, favourites, reshares, comments). Researchers have also started to analyse particular professional cultural areas and historical periods, such as website design, fashion photography, 20th-century popular music, and 19th-century literature. This work is being carried out in two newly developed fields: Social Computing and Digital Humanities.

Given the scale of that research, I am not interested in proposing Cultural Analytics as some alternative 'third way'. However, I think that the ideas this

1 Software Studies Initiative: www.softwarestudies.com.

term stands for remain relevant. As we will see, Digital Humanities and Social Computing have carved out their own domains in relation to the types of data they study, while 'Cultural Analytics' continues to be free of such limitations. It also attempts not to take sides vis-à-vis humanities vs. scientific goals and methods. In this article I don't take sides vis-à-vis humanities vs. scientific goals and methods. In this chapter I reflect on both paradigms, pointing out opportunities and ideas that have not yet been explored.

Digital Humanities scholars use computers to analyse mostly historical artefacts created by professionals, such as writers, artists and musicians. To take an example, one area of study could be novels written by professional writers in the 19th and 20th century. Yet for reasons of access, they stop at the historical boundaries defined by copyright laws in their countries. According to the United States copyright law, for example,[2] the works published in the last 95 years are automatically copyrighted. (So, for example, as of 2015, everything created after 1920 is copyrighted, unless it is recent digital content that uses Creative Commons licenses.) I have no qualms about respecting copyright laws, but in this case that means that digital humanists are shut out from studying the present.

The field of Social Computing is thousands of times larger. Here, researchers with advanced degrees in computer science study online user-created content and user interactions with this content. Note that this research is carried out not only by computer and information scientists who professionally identify themselves with the 'Social Computing' field, but also researchers in a number of other computer science fields such as Computer Multimedia, Computer Vision, Music Information Retrieval, Natural Language Processing, and Web Science. Therefore, social computing can also be used as an umbrella term for all computer science research that analyses content and activity on social networks. These researchers work with data from after 2004, when social networks and media sharing services started to become popular.[3] The data sets are usually much larger than the ones used in digital humanities. It is not uncommon to find tens or hundreds of millions of posts, photos or other items. Since the great majority of user-generated content is created by regular people rather than professionals, Social Computing studies the non-professional, vernacular culture by default.

2 A branch of computer science focused on the intersection of computational systems and social behaviour. See www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/social-computing.

3 Since it takes 1-2 years to do research and publish a paper, typically a paper published in 2015 will use the data collected in 2012-2014.

The scale of this research may be surprising to humanities and arts practitioners who may not realize how many people are working in computer science and related fields. For example, an October 2015 search on Google Scholar for 'Twitter dataset algorithm' returned 102,000 papers, a search for 'YouTube video dataset' returned 27,800 papers, and a search for 'Flickr images algorithm' returned 17,400 papers. Searching for 'computational aesthetics dataset', I got 14,100 results. Even if the actual numbers are much smaller, this is still impressive. Obviously not all these publications directly ask cultural questions, but many do.

The following table summarizes the differences between the two fields:

**Table 3.1. Comparing Social Computing and Digital Humanities.**

| | **Social Computing and various fields of computer science where researchers study social networks and shared media** | **Digital Humanities (research quantitative analysis using computer science techniques)** |
|---|---|---|
| **Number of publications** | Tens of thousands | Few hundred |
| **Period and material studied** | Websites and social media content and activity after 2004 | Historical artefacts up to the early 20th century |
| **Authors of artefacts studied** | Regular people who share content on social networks | Professional writers, artists, composers, etc. |
| **Typical size of data sets** | Thousands to hundreds of millions of items, billions of relations | Hundreds to thousands of items |

Why do computer scientists rarely work with large historical data sets of any kind? Typically, they justify their research by referencing already existing industrial applications – for example, search or recommendation systems for online content. The general assumption is that computer science will create better algorithms and other computer technologies useful to industry and government organizations. The analysis of historical artefacts falls outside this goal, and, consequently, only a few computer scientists work with historical data (the field of Digital Heritage being one exception).

However, looking at many examples of computer science papers, it becomes clear that they are actually doing Humanities or Communication Studies (in relation to contemporary media) but at a much larger scale. Consider these recent publications: 'Quantifying Visual Preferences Around the

World' (Reinecke & Gajos 2014), and 'What We Instagram: A First Analysis of Instagram Photo Content and User Types' (Hu et al. 2014). The first study analyses worldwide preferences for website design using 2.4 million ratings from 40,000 people from 179 countries. Studies like this of aesthetics and design traditionally belong to the humanities. The second study analysed the most frequent subjects of Instagram photos – a method comparable to art history studies of the genres in the 17th-century Dutch art which would be more appropriately categorized as humanities.

Another example is a paper called 'What is Twitter, a Social Network or a News Media?' (Kwak et al. 2014). First published in 2010, it has since been cited 3,284 times in other computer science publications.[4] It was the first large-scale analysis of Twitter as a social network, using 106 million tweets by 41.7 million users. The study looked in particular at trending topics, showing 'what categories trending topics are classified into, how long they last, and how many users participate'. This is a classic question of Communication Studies, going back to the pioneering work of Paul F. Lazarsfeld and his colleagues in the 1940s who manually counted the topics of radio broadcasts. But I would argue that given that Twitter and other micro-blogging services represent a new form of media, like oil painting, printed books and photography before them, understanding the specificity of Twitter as a medium is also a topic for humanities.

A small number of publications lie at the intersection of Digital Humanities and Social Computing. They take computational methods and algorithms developed by computer scientists for studying contemporary user-generated content and apply them to historical media artefacts created by professionals. The most prominent examples are 'Toward Automated Discovery of Artistic Influence' (Saleh et al. 2014), 'Infectious Texts: modelling Text Reuse in Nineteenth-Century Newspapers' (Smith et al. 2013), 'Measuring the Evolution of Contemporary Western Popular Music' (Serrà et al. 2012) and 'Quicker, faster, darker: Changes in Hollywood film over 75 years' (Cutting et al. 2011).

Until a few years ago, the only project that analysed cultural history on the scale of millions of texts was carried out by scientists rather than by humanists. I refer here to N-Gram Viewer created in 2010 by Google scientists Jon Orwant and Will Brockman following the prototype by two Harvard PhD students in Biology and Applied Mathematics. More recently, however, we see people in Digital Humanities scaling up the data they study. For example, in 'Mapping Mutable Genres in Structurally Complex

4 https://scholar.google.com/citations?user=M6i3Be0AAAAJ&hl=en.

Volumes' literary scholar Ted Underwood (2013) and his collaborators analysed 469,200 volumes from Trust Digital Library. Art historian Maximilian Schich and his colleagues (2014) have analysed the life trajectories of 120,000 notable historical individuals. And even larger historical data sets are becoming available in the areas of literature, photography, film and TV, although they have yet to be analysed. In 2012, The New York City Municipal Archives released 870,000 digitized historic photos of NYC (Taylor 2012). In 2015, HathiTrust made data extracted from 4,801,237 volumes (containing 1.8 billion pages) available for research (2016). In the same year Associated Press and British Movietone uploaded 550,000 digitized news stories covering the period from 1895 to today to YouTube (Associated Press 2015).

What is the importance of having such large cultural data sets? Can't we simply use smaller samples? I believe that there are a number of reasons. First of all, to have a representative sample, we first need to have a much larger set of actual items to draw from or at least a good understanding of what this larger set includes. So, for example, if we want to create a representative sample of 20th-century films, we can use IMDb (2015), which contains information on 3.4 million films and TV shows (including separate episodes). Similarly, we can create a good sample of historical US newspaper pages using the Historical American Newspaper collection of millions of digitized pages from the Library of Congress (2016). However, in many other cultural fields such larger data sets do not exist and without them, it may be impossible to construct representative samples.

The second reason is the following: without a large enough sample, we can only find general trends and patterns, but not local patterns. For example, in the already mentioned paper 'What We Instagram', three computer scientists analysed 1,000 Instagram photos and came up with the eight most frequent categories (selfie, friends, fashion, food, gadget, activity, pet, captioned photos). The sample of 1,000 photos was randomly selected from a larger set of photos shared by 95,343 unique users. It is possible that these eight categories were also most popular among all Instagram photos shared worldwide at the time when the scientists did their study. However, as we at the Software Studies Initiative saw from projects analysing Instagram photos in different cities and their parts (for example, the centre of Kyiv during the 2014 Ukrainian Revolution in *The Exceptional and the Everyday* (Manovich et al. 2014)), people also share many other types of images beyond Hu et al.'s eight categories. Depending on the geographic area and time period, some of these types may replace the top eight in popularity. In other words, while a small sample allows finding the 'typical' or 'most

popular,' it does not reveal what I call 'content islands' – types of coherent content with particular semantic and/or aesthetic characteristics shared in moderate numbers.

## Cultural Analytics

When I first started thinking about Cultural Analytics in 2005, both Digital Humanities and Social Computing were just getting started as research fields. I felt the need to introduce this new term to signal that our lab's work would not simply be a part of digital humanities or social computing but would cover subject matter studied in both fields. Like digital humanists, we are interested in analysing historical artefacts, but we are also equally interested in contemporary digital visual culture: Instagram as well as professional photography, artefacts created by dedicated non-professionals and artists outside of the art world like those found on deviantart.com,[5] and accidental creators, such as those who occasionally upload their photos to social media networks.

Like computational social scientists and computer scientists, we are also attracted to the study of society using social media and social phenomena specific to social networks. An example of the former would be using social media activity to identify similarities between different city neighbourhoods (Cranshaw et al. 2012). An example of the latter would be analysing patterns of information diffusion online (Cha et al. 2012). However, if Social Computing focuses on the *social* in social networks, Cultural Analytics focuses on the *cultural*. Therefore, the most relevant part of social sciences for Cultural Analytics is sociology of culture, and only after that sociology and economics.

We believe that content and user activities on the Web (on social networks and elsewhere) give us the unprecedented opportunity to describe, model and simulate the global cultural universe while questioning and rethinking basic humanities concepts and tools that were developed to analyse 'small cultural data' (i.e. highly selective and non-representative cultural samples). In the very influential 1869 definition by British cultural critic Matthew Arnold (1869), culture is 'the best that has been thought and said in the world'. The academic institution of humanities has largely followed this definition. And when they started to revolt against their canons and to include the works of previously excluded people (women, non-whites, non-Western authors, queer, etc.), they often included only 'the best' created by those who were previously excluded.

5 'The largest online social network for artists and art enthusiasts', http://about.deviantart.com/, retrieved 22 August 2015.

Cultural Analytics is interested in *everything created by everybody*. In this, we are approaching culture the way linguists study languages or biologists study life on earth. Ideally, we want to look at every cultural manifestation, rather than selective samples, in a systematic perspective not dissimilar to that of cultural anthropology. This larger inclusive scope combining the professional and the vernacular, the historical and the contemporary, is exemplified by the range of projects we have worked on in our lab since 2008. We have analysed historical, professionally created cultural content in all *Time* magazine covers (1923-2009); paintings by Vincent van Gogh, Piet Mondrian and Mark Rothko; 20,000 photographs from the collection of the Museum of Modern Art in New York (MoMA); and one million manga pages from 883 manga series published in the last 30 years. Our analysis of contemporary vernacular content includes *Phototrails* (the comparison of visual signatures of 13 global cities using 2.3 million Instagram photos) (Hochman et al. 2013), *The Exceptional and the Everyday: 144 Hours in Kyiv* (the analysis of Instagram images shared in Kyiv during the 2014 Ukrainian Revolution) (Manovich 2014) and *On Broadway* (the interactive installation exploring Broadway in NYC using 40 million user-generated images and data points) (Goddemeyer et al. 2014). We have also looked at contemporary amateur or semi-professional content using one million artworks shared by 30,000 semi-professional artists on deviantart.com. Currently, we are exploring a data set of 265 million images tweeted worldwide between 2011 and 2014. To summarize, our work doesn't draw a boundary between (smaller) historical professional artefacts and (bigger) online digital content created by non-professionals. Instead, it draws freely from both.

Obviously, online social networks today do not include every human being, and the content shared is sometimes specific to these networks (e.g. Instagram selfies), as opposed to something which existed before. This content is also shaped by the tools and interfaces of technologies used for its creation, capturing, editing and sharing (e.g. Instagram filters, its Layout app, etc.). The kind of cultural actions available are also defined by these technologies. For example, in social networks you can 'like', share or comment on a piece of content. In other words, just as in quantum physics, the instrument can influence the phenomena we want to study. All this needs to be carefully considered when we study user-generated content and user activities. While social network APIs make it easy to access massive amounts of contents, it is not 'everything' by 'everybody'.

## The General and the Particular

When the humanities were focused on 'small data' (content created by single authors or small groups), the sociological perspective was only one of many options for interpretation – unless you were a Marxist. But once we started studying online content and the activities of millions of people, this perspective became almost inevitable. In the case of 'big cultural data', the cultural and the social closely overlap. Large groups of people from different countries and socio-economic backgrounds (sociological perspective) share images, video, texts, and make particular aesthetic choices in doing this (humanities perspective). Because of this overlap, the kinds of questions investigated in *sociology of culture* of the 20th century (exemplified by its most influential researcher, Pierre Bourdieu (2010)) are directly relevant for Cultural Analytics.

Given that certain demographic categories have been taken for granted in our thinking about society, it appears natural today to group people into these categories and compare them in relation to social, economic or cultural indicators. For example, Pew Research Center regularly reports the statistics of popular social platform use, breaking up their user sample by demographics such as gender, ethnicity, age, education, income and residence (urban, suburban and rural) (Duggan et al. 2015). So if we are interested in various details of social media activities (such as types of images shared and liked, filters used or selfie poses) it is logical to study the differences between people from different countries, ethnicities, socio-economic backgrounds or levels of technical expertise. The earlier research in social computing did not, and most of the current work still does not consider such differences, treating all users as one undifferentiated pool of 'humanity'. More recently, however, we have started to see publications separating users into demographic groups. While we support this development, we also want to be careful in how far we want to go. Humanistic analysis of cultural phenomena and processes using quantitative methods should not be simply reduced to sociology and only consider common characteristics and behaviours of human groups.

The sociological tradition is concerned with finding and describing the *general* patterns in human behaviour, rather than with analysing or predicting the behaviours of particular individuals. Cultural Analytics, too, is interested in patterns that can be derived from the analysis of large cultural data sets. However, ideally *the analysis of the larger patterns will also lead us to individual cases,* such as individual creators, their particular creations or cultural behaviours. For instance, the computational analysis

of all photos made by a photographer during her long career may lead us to the outliers – the photos that are most different from all the rest. Similarly, we may analyse millions of Instagram images shared in multiple cities to discover the types of images unique to each city.

In other words, we may combine the concern of social science, and sciences in general, with the *general* and the *regular*, and the concern of humanities with the *individual* and the *particular*. The just described examples of analysing massive data sets to zoom in on the unique items illustrate one way of doing this, but it is not the only way.

## The Science of Culture?

The goal of science is to explain phenomena and develop compact mathematical models for describing how these phenomena work. Newton's three laws of physics are a perfect example of how classical science approached this goal. Since the middle of the 19th century, a number of new scientific fields have adopted a new probabilistic approach. The first example is the statistical distribution describing likely speeds of gas particles presented by Maxwell in 1860, now called the Maxwell-Boltzmann distribution. Throughout the 18th and 19th centuries, many thinkers were expecting that, similar to physics, the quantitative laws governing societies would also be eventually found (Ball 2004), yet this never happened. The closest 19th-century social thought came to postulating objective laws was in the works of Karl Marx. Instead, when positivist social science started to develop in the late 19th and early 20th century, it adopted the probabilistic approach. So instead of looking for deterministic laws of society, social scientists study correlations between measurable characteristics and model the relations between 'dependent' and 'independent' variables using various statistical techniques.

After deterministic and probabilistic paradigms in science, the next paradigm was computational simulation – running models on computers to simulate the behavior of systems. The first large-scale computer simulation was created in the 1940s by the Manhattan Project to model a nuclear explosion. Subsequently, simulation was adapted in many hard sciences, and in the 1990s it was also taken up in the social sciences.

In the early 21st century, the volume of digital online content and user interactions allows us to think of a possible 'science of culture'. For example, by the summer of 2015, Facebook users were sharing 400 million photos and sending 45 billion messages daily (Smith 2015). This scale is still much

smaller than that of atoms and molecules.[6] However, it is already bigger than the numbers of neurons in the whole nervous system of an average adult, which is estimated at 86 billion. But since science now includes a few fundamental approaches to studying and understanding the phenomena – deterministic laws, statistical models and simulation – which of them should a 'science of culture' adapt?

Looking at the papers of computer scientists who are studying social media data sets, it is clear that their default approach is statistics.[7] They describe social media data and user behaviour in terms of probabilities. This includes the creation of statistical models – mathematical equations that specify the relations between variables that may be described using probability distributions rather than specific values. A majority of papers today also use supervised machine learning, an automatic creation of models that can classify or predict the values of the new data using already existing examples. In both cases, a model can only account for part of the data, and this is typical of the statistical approach.

Computer scientists studying social media use statistics differently than social scientists. The latter want to *explain* social, economic or political phenomena.[8] Computer scientists are generally not concerned with explaining patterns in social media by referencing some external social, economic or technological factors. Instead, they typically either analyse social media phenomena internally or try to predict the outside phenomena using information extracted from social media data sets. The example of the former is a statistical description of how many favourites a photo on Flickr may receive on average after a certain period of time.[9] The example of the latter is the Google Flu Trends service that predicts flu activity using a combination of Google search data and the US Centers for Disease Control and Prevention's official flu data (Stefansen 2014).

The difference between deterministic laws and non-deterministic models is that the latter describe probabilities, not certainties. The laws of classical mechanics apply to any macroscopic objects. In contrast, a probabilistic model for predicting the number of favorites for a Flickr photo as a function of time since it was uploaded cannot tell us exactly the

6 1 $cm^3$ of water contains 3.33 *1022 molecules.

7 Computer scientists also use many recently developed methods including techniques of data mining and machine learning that were not part of 20th-century statistics. I discuss these differences in 'Data Science and Digital Art History,' *International Journal for Digital Art History*, issue 1 (2015), https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631.

8 For example, the effect of family background on children's educational performance.

9 See 'Delayed information cascades in Flickr.'

numbers of favourites for any particular photo. It only describes the overall trend. This seems to be the appropriate method for a 'science of culture'. If instead we start postulating deterministic laws of human cultural activity, what happens to the idea of free will? Even in the case of seemingly automatic cultural behaviour (people favouring photos on social networks with certain characteristics such as pretty landscapes, cute pets or posing young females), we don't want to reduce humans to mechanical automata for the passing of memes.

The current focus on probabilistic models in studying online activity leaves out the third scientific paradigm – simulation. As far as I know, simulation has not yet been explored in either Social Computing or Digital Humanities as a tool for studying user-generated content, its topics, types of images, etc. If scientists at IBM's Almaden research centre simulated human visual cortex using 1.6 billion virtual neurons with 9 trillion synapses in 2009 (Fox 2009), why can't we think of simulating, for instance, all content produced yearly by users of Instagram? Or all content shared by all users of major social networks? Or the categories of images people share? The point of such simulations will not be to get everything right or to precisely predict what people will be sharing next year. Instead, we can follow the authors of the influential textbook *Simulation for the Social Scientist* (Gilbert & Troitzsch 2005) when they state that one of the purposes of simulation is 'to obtain a better *understanding* of some features of the social world' and that simulation can be used as 'a method of *theory development'* (emphasis added). Since computer simulation requires developing an explicit and precise model of the phenomena, thinking of how cultural processes can be simulated can help us to develop more explicit and detailed theories than we use normally.[10]

And what about 'big data'? Does it not represent a new paradigm in science with its own new research methods? This is a complex question that deserves its own article.[11] However, as a way of conclusion, I do want to mention one concept interesting for humanities that we can borrow from big data analytics and then push in a new direction.

10 For the example of how agent-based simulation can be used to study the evolution of human societies, see 'War, space, and the evolution of Old World complex societies', http://peterturchin.com/PDF/Turchin_etal_PNAS2013.pdf.

11 If we are talking about research methods and techniques, the developments in computer hardware in the 2000s, including the increasing CPU speed and RAM size, and the use of GPUs and computing clusters, were probably more important than availability of larger data sets. And while use of machine learning with large training data sets achieved remarkable successes, in most cases it does not provide explanations of the phenomena.

The 20th-century social science was working on what we can call 'long data'.[12] That is, the number of cases was typically many times bigger than the number of variables being analysed. For example, imagine that we surveyed 2,000 people asking them about their income, family educational achievement and their years of education. As a result, we have 2000 cases and three variables. We can then examine correlations between these variables, or look for clusters in the data, or perform other types of statistical analysis.

The beginnings of social sciences are characterized by the most extreme asymmetries of this kind. The first positivist sociologist, Karl Marx, divided all humanity into just two classes: people who own means of production and people who don't, i.e. capitalists and the proletariat. Later sociologists added other divisions. Today these divisions are present in numerous surveys, studies and reports in popular media and academic publications – typically, gender, race, ethnicity, age, educational background, income, place of living, religion, and some others. But regardless of details, the data collected, analysed and interpreted is still very 'long'. The full populations or their samples are described using a much smaller number of variables.

But why should this be the case? In the fields of computer media analysis and computer vision, computer scientists use algorithms to extract thousands of features from every image, video, tweet, email, and so on.[13] So while Vincent van Gogh, for example, only created about 900 paintings, these paintings can be described on thousands of separate dimensions. Similarly, we can describe everybody living in a city on millions of separate dimensions by extracting all kinds of characteristics from their social media activity. For another example, consider our own project *On Broadway* where we represent Broadway in Manhattan with 40 million data points and images using messages, images and check-ins shared along this street on Twitter, Instagram and Foursquare, as well as taxi rides data and the US Census indicators for the surrounding areas.[14]

In other words, instead of *long data* we can have *wide data* – very large and potentially endless number of variables describing a set of cases. Note that if we have more variables than cases, such representation would go against the common sense of both social science and data science. The latter refers to the process of making a large number of variables more manageable

12 I am using this term in a different way than Samuel Arbesman in his 'Stop Hyping Big Data and Start Paying Attention to "Long Data"', wired.com, 29 January 2013, www.wired.com/2013/01/forget-big-data-think-long-data/.

13 I explain the reason for using a large number of features in 'Data Science and Digital Art History.' (Manovich 2015).

14 Described at length in the following chapter.

as *dimension reduction*. But for us, 'wide data' offers an opportunity to rethink fundamental assumptions about what society is and how to study it, and similarly, what is culture, an artistic career, a body of images, a group of people with similar aesthetic taste, and so on. Rather than dividing cultural history using one dimension (time), or two (time and geographic location) or a few more (e.g. media, genre), endless dimensions can be put in play. The goal of such 'wide data analysis' will not be only to find new similarities, affinities and clusters in the universe of cultural artefacts, but to question a taken-for-granted view of things, where certain dimensions are taken for granted. This is one example of the general Cultural Analytics method: estrangement (*ostranenie*)[15], making our basic cultural concepts and ways of organizing and understanding cultural data sets foreign to us so that we can approach them anew. In this way, we use data and data-manipulating techniques to question how we think, see and ultimately act on our knowledge.

## References

Arnold, Matthew. 1960. *Culture and Anarchy*. Cambridge: Cambridge University Press.

Associated Press. 2015. "AP Makes One Million Minutes of Historical Footage Available on YouTube." 22 July. Accessed 12 February 2016. www.ap.org/content/press-release/2015/ap-makes-one-million-minutes-of-history-available-on-youtube.

Ball, Philip. 2004. *Critical Mass: How One Thing Leads to Another*. New York: Farrar, Straus and Giroux: 69-71.

Bourdieu, Pierre. 2010. *Distinction: A Social Critique of the Judgement of Taste*. London: Routledge.

Cha, Meeyoung, Fabrício Benevenuto, Yong-Yeol Ahn & Krishna P. Gummadi. 2012. "Delayed information cascades in Flickr: Measurement, analysis, and modelling." *Computer Networks* 56: 1066–1076.

Cranshaw, Justin, Raz Schwartz, Jason I. Hong & Norman Sadeh. 2012. "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." The 6th International AAAI Conference on Weblogs and Social Media (Dublin.)

Cutting, James E., Kaitlin L. Brunick, Jordan DeLong, Catalina Iricinschi, Ayse Candan. 2011. "Quicker, faster, darker: Changes in Hollywood film over 75 years." *i-Perception*, vol. 2: 569-576.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart & Mary Madden. 2015. "Demographics of Key Social Networking Platforms." Pew Research Center Internet Science Tech RSS. 9 January. www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2.

Fox, Douglas. 2009. "IBM Reveals the Biggest Artificial Brain of All Time." *Popular Mechanics*. December 17. www.popularmechanics.com/technology/a4948/4337190/.

Gilbert, Nigel & Klaus G. Troitzsch. 2005. *Simulation for the Social Scientist*, 2nd edition: 3-4.

15 The term 'ostranenie' was introduced by Russian literary theorist Viktor Shklovsky in his essay 'Art as a Technique' in 1917. See www.vahidnab.com/defam.htm.

Goddemeyer, Daniel, Moritz Stefaner, Dominikus Baur & Lev Manovich. 2014. "On Broadway." http://on-broadway.net/.

HathiTrust. 2016. "HTRC Extracted Features Dataset Page-level Features from 4.8 Million Volumes [v.0.2]." HTRC Portal. Accessed 12 February 2016. https://sharc.hathitrust.org/features.

Hochman, Nadav, Lev Manovich & Jay Chow. 2013. "Phototrails." http://phototrails.net/.

Hu, Yuheng, Lydia Manikonda & Subbarao Kambhampati. 2014. "What We Instagram: A First Analysis of Instagram Photo Content and User Types." *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (*ICWSM*).

IMDb. 2015. "IMDb Database Statistics." www.imdb.com/stats. Accessed 10 August 2015.

Kwak, Haewoon, Changhyun Lee, Hosung Park & Sue Moon. 2014. "What is Twitter, a Social Network or a News Media?" *Proceedings of the 19th International World Wide Web* (*WWW*) *Conference* (*ACM*): 591-600.

Library of Congress. 2016. "About Chronicling America." News about Chronicling America RSS. Accessed 12 February 2016. http://chroniclingamerica.loc.gov/about/.

Manovich, Lev, 2015. "Data Science and Digital Art History." *International Journal of Digital Art History*, issue 1. http://www.dah-journal.org/issue_01.html.

Manovich, Lev, Mehrdad Yazdani, Alise Tifentale & Jay Chow. 2014. "The Exceptional and the Everyday: 144 hours in Kyiv." www.the-everyday.net/.

Reinecke, Katharina & Krzysztof Z. Gajos. 2014. "Quantifying Visual Preferences Around the World. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14 (New York: ACM): 11-20.

Saleh, Babak, Kanako Abe, Ravneet Singh & Arora Ahmed Elgammal. 2014. "Toward Automated Discovery of Artistic Influence." *Multimedia Tools and Applications* (Springler, 19 August): 1-27.

Schich, Maximilian, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási & Dirk Helbing. 2014. "A network framework of cultural history." *Science,* 1 August: 345 (6196): 558-562.

Serrà, Joan, Álvaro Corral, Marián Boguñá, Martín Haro & Josep Ll. Arcos. 2012. "Measuring the Evolution of Contemporary Western Popular Music." *Nature Scientific Reports*. 2, article number: 521.

Smith, Craig. 2015. "By the Numbers: 200 Amazing Facebook Statistics (January 2015)." DMR. 24 January. http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/15.

Smith, David A., Ryan Cordell & Elizabeth Maddock Dillon. 2013. "Infectious texts: modelling text reuse in nineteenth-century newspapers." *Proceedings of the 2013 IEEE Conference on Big Data*: 84-94.

Stefansen, Christian. 2014. "Google Flu Trends Gets a Brand New Engine." Research Blog. 31 October. http://googleresearch.blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html.

Taylor, Alan. 2012. "Historic Photos from the NYC Municipal Archives." *The Atlantic*. Accessed 12 February 2016.

Underwood, Ted, Michael L. Black, Loretta Auvil & Boris Capitanu. 2013. "Mapping Mutable Genres in Structurally Complex Volumes." *Proceedings of the 2013 IEEE Conference on Big Data*: n.p.