Lev Manovich

Trending: The Promises and the Challenges of Big Social Data

[Uploaded: 4/28/2011]

Today the term “big data” is often used in popular media, business, computer science and computer industry. For instance, in June 2008 *Wired* magazine opened its special section on “The Petabyte Age” by stating: “Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions” (“The Petabyte Age”). In February 2010, *Economist* started its special report “Data, data everywhere” with the phrase “the industrial revolution of data” (coined by computer scientist Joe Hellerstein) and then went to note that “The effect is being felt everywhere, from business to science, from government to the arts” (“Data, data everywhere”).

Discussions in popular media usually do not define “big data” in qualitative terms. However, in computer industry the term has a more precise meaning: “Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set” (“Big data”)..

Since its formation in 2008, NEH Office of Digital Humanities has been systematically creating grant opportunities to help humanists work with large data sets. The following statement from 2011 grant competition organized by NEH together with a number of other research agencies in USA, Canada, UK, and Netherlands provides an excellent description of what is at stake:

“The idea behind the Digging into Data Challenge is to address how "big data" changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences -- ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records -- what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials.” (“Digging into Data Challenge”).

The projects funded by 2009 Digging Into Data Challenge and earlier NEH 2008 Humanities High Performance Computing grant program begin to map the landscape of data-intensive humanities. They include analysis of 18th century European thinkers; maps, texts, and images associated with 19th century railroads in the U.S., criminal trial accounts (data size: 127 million words); ancient texts, detailed 3D map of ancient

Rome; and the project by my lab to develop tools for the analysis and visualization of large image and video data sets.

At the moment of this writing, the largest data sets being used in digital humanities projects are much smaller than big dataused by scientists; in fact, if we use industry's definition, almost none of them qualify as big data (i.e. the work can be done on desktop computers using standard software, as opposed to supercomputers.) But this gap will eventually disappear when humanists start working with born-digital user-generated content (such as billions of photos on Flickr), users online communication (comments about photos), user created metadata (tags) and transaction data (when and from there the photos were uploaded). This web content and data are infinitely larger than all already digitized cultural heritage, and, in contrast to the fixed number of historical artifacts, is grows constantly. (I expect that the number of photos uploaded to Facebook daily is larger than all artifacts stored n all world's musems.)

In this text I want to address some of the theoretical and practical issues raised by the possibility of using massive amounts of such social and cultural data in humanities and social sciences. My observations are based on my own experience working with large cultural data sets carried out in our Software Studies Initiative (softwarestudies.com) at UCSD since 2007. The issues which I will discuss include the differences between "deep data" about a few people and "surface data" about lots of people; getting access to transactional data; and the new "data analysis divide" between data experts and researchers without computer science training.

The emergence of social media in the middle of 2000s created opportunities to study social and cultural processes and dynamics in new ways. For the first time, we can follow imaginations, opinions, ideas, and feelings of hundreds of millions of people. We can see the images and the videos they create and comment on, monitor the conversations they are engaged in, read their blog posts and tweets, navigate their maps, listen to their track lists, and follow their trajectories in physical space. And we don't need to ask their permission to do this, since they themselves encourage us to do by making all this data public.

In the 20th century, the study of the social and the cultural relied on two types of data: "surface data" about lots of people and "deep data" about the few individuals or small groups. The first approach was used in all disciplines that adapted quantitative methods (i.e., statistical, mathematical or computational techniques for analyzing data. The relevant fields include quantitative schools of sociology, economics, political science, communication studies, and marketing research.

The second approach was typical of humanities: literary studies, art history, film studies, history. It was also used in non-quantitative schools in psychology (for instance, psychoanalysis and Gestalt psychology), sociology (Wilhelm Dilthey, Max Weber,

Georg Simmel), anthropology, and ethnography. The examples of relevant methods are hermeneutics, participant observation, thick description, semiotics, and close reading.

For example, a quantitative sociologist worked with census data that covered most of the country's citizens. However, this data was collected only every 10 year and it represented each individual only on a "macro" level, living out her/his opinions, feelings, tastes, moods, and motivations ("US Census Bureau"). In contrast, a psychologist would be engaged with a single patient for years, tracking and interpreting exactly the kind of data which census did not capture.

In the middle between these two methodologies of "surface data" and "deep data" were statistics and the concept of sampling. By carefully choosing her sample, a researcher could expand certain types of data about the few into the knowledge about the many. For example, starting in 1950s, Nielsen Company ("Nielsen Company") collected TV viewing data in a sample of American homes (via diaries and special devices connected to TV sets in 25,000 homes), and then used this sample data to predict TV ratings for the whole country (i.e. percentages of the population which watched particular shows). But the use of samples to learn about larger populations had many limitations.

For instance, in the example of Nelson's TV ratings, the small sample did not tell us anything about the actual hour by hour, day to day patterns of TV viewing of every individual or every family outside of this sample. Maybe certain people watched only news the whole day; others only tuned in to concerts; others had TV on not never paid attention to it; still others happen to prefer the shows which got very low ratings by the sample group; and so on. The sample stats could not tell us anything about this. It was also possible that a particular TV program would get zero shares ("Nielsen ratings") because nobody in the sample audience happened to watch it – and in fact, this occurred more than once ("Nielsen ratings").

Imagine that we want to scale up a low-resolution image using a digital image editor like Photoshop. For example, we start with 10x10 pixel image (100 pixels in total), and resize it to 1000x1000 (one million pixels in total). We don not get any new details – only larger pixels. This is exactly what happens when you use a small sample to predict the behavior of a much larger population. A "pixel" that originally represented one person comes to represent 1000 people who all assumed to behave in exactly the same way.

The rise of social media along with the progress in computational tools that can process massive amounts of data makes possible a fundamentally new approach for the study of human beings and society. We no longer have to choose between data size and data depth. We can study exact trajectories formed by billions of cultural expressions, experiences, texts, and links. The detailed knowledge and insights that before can only be reached about a few people can now be reached about many more people. In 2007, Bruno Latour summarized these developments as follows: "The precise forces that mould our subjectivities and the precise characters that furnish our imaginations are all open to inquiries by the social sciences. It is as if the inner workings of private worlds

have been pried open because their inputs and outputs have become thoroughly traceable." (Latour, "Beware, your imagination leaves digital traces").

Two years earlier, in 2005, PhD student Nathan Eagle at MIT Media Lab was already thinking along the similar lines. He and his advisor Alex Pentland put up a web site called "reality mining" ("MIT Media Lab: Reality Mining") and wrote how the new possibilities of capturing details of peoples' daily behavior and communication via mobile phones can create Sociology in the 21st century ("Sociology in the 21st Century"). To put this idea into practice, they distributed Nokia phones with special software to 100 MIT students who then used these phones for 9 months – which generated approximately 60 years of "continuous data on daily human behavior." Eagle and Pentland published a number of articles based on the analysis of data they collected. Today many more computer scientists are working with large social data sets; they call their new field "social computing." According to the definition provided by the web site of The Third IEEE International Conference on Social Computing (2011), social computing refers to "computational facilitation of social studies and human social dynamics as well as design and use of information and communication technologies that consider social context." ("Social Computing.")

Now, let us consider Google search. Google's algorithms analyze billions of web pages, plus PDF, Word documents, Excel spreadsheets, Flash files, plain text files, and, since 2009, Facebook and Twitter content. (More details: en.wikipedia.org/wiki/Google_Search). Currently Google does not offer any service that would allow a user to analyze patterns directly n all of this text data the way Google Insights for Search does with search queries and Google's Ngram Viewer does with digitized books – but it is certainly technologically conceivable. Imagine being able to study the collective intellectual space of the whole planet, seeing how ideas emerge and diffuse, burst and die, how they get linked together, and so on – across the data set estimated to contain at least 14.55 billion pages ("The size of the World Wide Web").

To quote again *Wired* "Petabyte Age" issue: "Because in the era of big data, more isn't just more. More is different" ("The Petabyte Age").

Does all this sounds exiting? It certainly does. So what may be wrong with these arguments? Do we indeed witness the collapse of deep data / surface data divide? Does this collapse open a new era for social and cultural research?

I am going to discuss four objections to the optimistic vision I just presented. These objections do not imply that we should not use new data sources about human culture and human social life, or not analyze them with computational tools. I strongly believe that we should do this – but we need to carefully understand what is possible in practice, as opposed to in principle. We also need to be clear about what skills digital humanists need to take advantage of the new scale of human data.

1. Only social media companies have access to really large social data - especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not.

A researcher can obtain some of this data through APIs provided by most social media services and largest media online retailers (YouTube, Flickr, Amazon, etc.). API (Application Programming Interface) is a set of commands that can be used by a user program to retrieve the data stored in a company's databases. For example, Flickr API can be used to download all photos in a particular group, and also retrieve information about each photo size, available comments, geo location, list of people who favored this photo, and so on ("Flickr API Methods").

The public APIs provided by social media and social network companies do not give all data that these companies themselves are capturing about the users. Still, you can certainly do very interesting new cultural and social research by collecting data via APIs and then analyzing it – if you are good at programming, statistics, and other data analysis methods. (In my lab we have recently used Flickr API to download 167,000 images from "Art Now" Flickr group, and currently working to analyze these images to create a "map" of what can be called "user-generated art.")

Although APIs themselves are not complicated, all truly large-scale research projects, which use the data with these APIs so far have been done by researchers in computer science. A good way to follow the work in this area is to look at papers presented at yearly WWW conferences ("WWW2009" and "WWW2010"). Recent papers investigated how information spreads on Twitter (data: 100 million tweets) (Kwak, Lee, Park, Moon), what qualities are shared by most favored photos on Flickr (data: 2.2 million photos), how geo-tagged Flickr photos are distributed spatially (data: 35 million photos) (Crandall, Backstrom, Huttenlocher, Kleinberg), and how user-generated videos on YouTube compare with similar videos on Daum, the most popular UGC (user-generated content) service in Korea (data: 2.1 million videos) (Cha, Kwak, Rodriguez, Ahn, Moon).

It is worth pointing out that even researchers working inside the largest social media companies can't simply access all the data collected by different services in a company. Some time ago I went to a talk by a researcher from Sprint (one of the largest US phone companies) who was analyzing the relations between geographic addresses of phone users and how frequently they called other people. He did have access to this data for all Sprint customers (around 50 million.) However, when he was asked why he did not used other data Sprint collects such as instant messages and apps use, he explained that these services are operated by a different part of the company, and that the laws prohibit employees to have access to all of this data together. He pointed out that like any other company, Spring does not want to get into lawsuits for breach of privacy, pay huge fines and damage their brand image, and therefore they are being very careful in terms of who gets to look at what data. You don't have to believe this, but I do. For example, do you think Google enjoys all the lawsuits about Street View? If you were running a business, would you risk losing hundreds of millions of dollars and badly

damaging your company image?

2. We need to be careful of reading communications over social networks and digital footprints as “authentic.” Peoples’ posts, tweets, uploaded photographs, comments, and other types of online participation are not transparent windows into their selves; instead, they are often carefully curated and systematically managed (Ellison, Heino, Gibbs, “Managing impressions online”).

Imagine that you wanted to study cultural imagination of people in Russia in the second part of 1930s and you only looked at newspapers, books, films, and other cultural texts - which of course all went through government censors before being approved for publication. You would conclude that indeed everybody in Russia loved Lenin and Stalin, was very happy, and was ready to sacrifice his/her life to build communism. You may say that this is unfair comparison, and it would be more appropriate to look instead at people’s diaries. Yes, indeed it would be better – however if you were living in Russia in that period, and you knew that any night a black car may stop in front of your house and you would be taken away and probably shot soon thereafter, would you really commit all your true thoughts about Stalin and government to paper? In 1993 famous Russian poet Osip Mandelstam wrote a short poem that criticized Stalin only indirectly without even naming him – and he paid for this with his life (“Stalin Epigram”).

Today, if you live in a pretty large part of the world, you know that the government is likely to scan your electronic communications systematically (“Internet Censorship by country”). In some of the countries, it also may arrest you simply for visiting a wrong web site. In these countries, you will be careful in what you are saying online. Some of us live in other countries where a statement against the government does not automatically put you in prison, and therefore people feel they can be more open. In other words, it does not matter if the government is tracking us not; what is important is what it can do with this information. (I grew up in Soviet Union in 1970s and then moved to the US; based on my experience living in both societies, in this respect the difference between them is very big. In USSR, we never made any political jokes on the phone, and only discussed politics with close friends at home.)

Now, let us assume that we are living in a country where we are highly unlikely to be prosecuted for occasional anti-government remarks. But still, how authentic are all the rest of our online expressions? As Ervin Goffman and other sociologists pointed out a long time ago, people always construct their public presence, carefully shaping how they present themselves to others – and social media is certainly not an exception to this (“Ervin Goffman”), The degree of this public self-construction varies. For instance, most of us tend to do less self-censorship and editing on Facebook than in the profiles on dating sites, or in a job interview. Others carefully curate their profile pictures to construct an image they want to project. (If you scan your friends Facebook profile pictures, you are likely to find a big range). But just as we do in all other areas of our everyday life, we exercise some control all the time when we are online – what we say, what we upload, what we show as our interests, etc.

Again, this does not mean that we can't do interesting research by analyzing larger numbers of tweets, Facebook photos, YouTube videos, etc. – we just have to keep in mind that what all this data is not a transparent window into peoples' imaginations, intentions, motifs, opinions, and ideas. Its more appropriate to think of it as an interface people present to the world – i.e., a particular view which shows only some of the data of their actual lives and imaginations and which may also include other fictional data designed to project a particular image (Ellison, Heino, Gibbs).

3. Is it really true that "We no longer have to choose between data size and data depth" as I stated? Yes and no. Imagine this hypothetical scenario. On the one side, we have ethnographers who are spending years inside a particular community. On another side, we have computer scientists who never meet people in this community but have access to their social media and digital footprints - daily spatial trajectories captured with GPS, all video recorded by surveillance cameras, online and offline conversations, uploaded media, comments, likes, etc. According to my earlier argument, both parties have "deep data" – but the advantage of computer science team is that they can capture this data about hundreds of millions of people as opposed to only small community.

How plausible is this argument? For thousands of years, we would learn about other people exlusiely through personal interactions. Later, letter writing became an important new mechanism for building personal (especially romantic) relationships. In the 21st century, we can have access to a whole new set of machine captured traces and records of individual's activities. Given that this situation is very new, it is to be expected that some people will find the concept that such machine records can be as meaningful in helping us to understand communities and individuals as face-to-face interaction hard to accept. They will argue that no matter how good are computer scientists' data sources, data analysis ideas and algorithms, they will never arrive at the same insights and understanding of people and dynamics in the community as ethnographers. They will say that even the most comprehensive social data about people which can be automatically captured via cameras, sensors, computer devices (phones, game consoles) and web servers can't be used to arrive at the same "deep" knowledge.

It is possible to defend both positions – but what if both are incorrect? I think so. I believe that in our hypothetical scenario, ethnographers and computer scientists have access to *different* kinds of data. Therefore they are likely to ask different questions, notice different patterns, and arrive at different insights.

This does not mean that the new computer-captured "deep surface" of data is less "deep" than the data obtained through long-term personal contact. In terms of the sheer number of "data points," it is likely to be much deeper. However, many of these data points are quite different than the data points available to ethnographers.

For instance, if you are physically present in some situation, you may notice some things which you would not notice if you watching a high-res video of the same situation.

But at the same time, if you do computer analysis of this video you may find patterns you would not notice if you were on the scene physically only. Of course, people keep coming up with new techniques that combine on the scene physical presence and computer and network-assisted techniques. For a good example of such innovation, see valleyofthekhans.org project at UCSD. In this project, photos captured by small unmanned aerial vehicles out by archoelogical team moving around a large area in Mongolia are immediately upoaded to a special *National Geographic* site exploration.nationalgeographic.com. Thousands of people immediately start tagging these photos for interesting details – which tells archeologists what to look for on the ground.) ("Help Find Genghis Khan's Tomb.")

The questions of what can be discovered and understood with computer analysis of social and cultural data versus traditional qualitative methods are particularly important for digital humanities. My hypothetical example above used data about social behavior, but the "data" can also be 18th century letters of European thinkers, 19th century maps and texts about railroads, hundreds of thousands of images uploaded by users to a Flickr group, or any other set of cultural artifacts. When we start reading these artifacts with computers, this often makes many humanists really nervous.

I often experience this reaction when I lecture about digital humanities research done in my lab Software Studies Initiative at UCSD (softwarestudies.com). The lab focuses on development of methods and tools for exploration and research of massive cultural visual data - both digitized visual artifacts and contemporary visual and interactive media' ("Software Studies: Cultural Analytics"). We use digital image analysis and new visualization techniques to explore cultural patterns in large sets of images and video – user-generated video, visual art, magazine covers and pages, graphic design, photographs, feature films, cartoons, motion graphics. Examples of visual data sets we analyzed include 20,000 pages of *Science* and *Popular Science* magazines issues published between 1872-1922, 780 paintings by van Gogh, 4535 covers of Time magazine (1923-2009) and one million manga pages ("One million manga pages").

In our experience, practically every time we analyze and then visualize a new image video collection, or even a single time-based media artifact (a music video, a feature film, a video recording of a game play), we find some surprising new patterns. This equally applies to collections of visual artifacts about which we had few a priori assumptions (for instance, 167,000 images uploaded by users to "Art Now" Flickr) and artifacts that already were studied in details by multiple authors.

As an example of the latter, I will discuss a visualization of the film *The Eleventh Year* by a famous 20th century Russian director Dziga Vertov (Manovich, "Visualizing Large Image Collections for Humanities Research". The visualization itself can be downloaded from our Flickr account (Manovich, "Motion Studies: Vertov's The Eleventh Year").

My sources were the digitized copy of the film provided by Austrian Film Museum, and the information about all shot boundaries created manually by a museum researcher. (With other moving image sources, we use open source software shotdetect that

automatically detects most shot boundaries in a typical film.) The visualization uses only first and last frame of every shot in the film, disregarding all other frames. Each shot is represented as a column: first frame is on the top, and last frame is right below.

”Vertov” is a neologism invented by director who adapted it as his last name early in his career. It comes from the Russian verb *vertet*, which means “to rotate something.” “Vertov” may refer to the basic motion involved in filming in the 1920s - rotating the handle of a camera – and also the dynamism of film language developed by Vertov who, along with a number of other Russian and European artists, designers and photographs working in that decade wanted to defamiliarize familiar reality by using dynamic diagonal compositions and shoting from unusual points of view. However, my visualization suggests a very different picture of Vertov. Almost every shot of *The Eleventh Year* starts and ends with practically the same composition and subject. In other words, the shots are largely static. Going back to the actual film and studying these shots further, we find that some of them are indeed completely static – such as the close-ups of a people faces looking in various directions without moving. Other shots employ a static camera that frames some movement – such as working machines, or workers at work – but the movement is localized completely inside the frame (in other words, the objects and human figures do not cross the view framed by the camera.) Of course, we may recall that a number of shots in Vertov’s most famous film *Man with A Movie Camera* (1929) were specifically designed as opposites: shooting from a moving car meant that the subjects were constantly crossing the camera view. But even in this most experimental of Vertov’s film, such shots constitutes a very small part of a film.

One of the typical responses to my lectures is that computers can’t lead to the same nuanced interpretation as traditional humanities methods and that they can’t help understand deep meanings of artworks. My response is that we don’t want to replace human experts with computers. As I will describe in the hypothetical scenario of working with one million YouTube documentary-style videos below, we can use computers to quickly explore massive visual data sets and then select the objects for closer manual analysis. While computer-assisted examination of massive cultural data sets typically reveals new patterns in this data which even best manual “close reading” would miss – and of course, even an army of humanists will not be able to carefully “close read” massive data sets in the first place – a human is still needed to make sense of these patterns.

Ultimately, completely automatic analysis of social and cultural data will not produce meaningful results today because computers’ ability to understand the content texts, images, video and other media is still limited. (Recall the mistakes made by IBM Watson artificial intelligence computer when it competed on the TV quiz show Jeopardy! in early 2011) (“Watson (computer)”).

Ideally, we want to combine human ability to understand and interpret - which computers can’t completely match yet - and computers’ ability to analyze massive data

sets using algorithms we create. Lets us imagine the following research scenario. You want to study documentary-type YouTube videos created by users in country X during the period Y, and you were able to determine that the relevant data set contains 1 million videos. So what do you do next? Computational analysis would be perfect as the next step to map the overall “data landscape”: identify most typical and most unique videos, automatically cluster all videos into a number of categories; find all videos that follow the same strategies, etc. At the end of this analytical stage, you may be able to reduce the set of one million videos to 100 videos which represent it in a more comprehensive way than if you simply used a standard sampling procedure. For instance, you reduced set may contain both most typical and most unique videos in various categories. Now that you have a manageable number of videos, you can actually start watching them. If you find some video to be particularly interesting, you can then ask computer to fetch more videos which have similar characteristics, so you can look at all of them. At any point in the analysis, you can go back and forth between particular videos, groups of videos and the whole collection of one million videos, experimenting with new categories and groupings. And just as Google analytics allows you to select any subset of data and look at its patterns over time (number of viewed pages) and space (where do visitors come from), you will be able to select any subset of the videos and see various patterns across this subsets.

This is my vision of how we can study large cultural data sets – whether these are billions of videos on YouTube or billions of photos on Flickr, or smaller samples of semi-professional or professional creative productions such as 100 million images on deviantart.com, or 250,000 design portfolios on coroflot.com. Since 2007, our lab has gradually working on visualization techniques that would enable such research exploration.

4. Imagine that you have software that combines large-scale automatic data analysis and interactive visualization. (We are gradually working to integrate various tools which we designed in our lab to create such a system. See “Cultural Analytics Research Environment.”) If you also have skills to examine individual artifacts and the openness to ask new questions, the software will help you to take research in many new exiting directions. However, there are also many kinds of interesting questions that require expertise in computer science, statistics, and data mining – something which social and humanities researchers typically don’t have. This is another serious objection to the optimistic view of new “big data”-driven humanities and social research I presented above.

The explosion of data and the emergence of computational data analysis as the key scientific and economic approach in contemporary societies create new kinds of divisions. Specifically, people and organizations are divided into three categories: those who create data (both consciously and by leaving digital footprints), those who have the means to collect it, and those who have expertise to analyze it. The first group includes pretty much everybody in the world who is using the web and/or mobile phones; the

second group is smaller; and the third group is much smaller still. We can refer to these three groups as new “data-classes” of our “big data society” (my neologisms).

At Google, computer scientists are working on the algorithms that scan a web page a user is on currently and select which ads to display. At YouTube, computer scientist’s work on algorithms that automatically show a list of other videos deemed to be relevant to one you are currently watching. At BlogPulse, computer scientists work on algorithms that allow companies to use sentiment analysis to study the feelings that millions of people express about their products in blog posts. At certain Hollywood movie studios, computer scientists work on algorithms that predict popularity of forthcoming movies by analyzing tweets about them (it works). In each case, the data and the algorithms can also reveal really interesting things about human cultural behavior in general – but this is not what the companies who are employing these computer scientists are interested in. Instead, the analytics are used for specific business ends. (For more examples, see “What People Want (and How to Predict it)”).

So what about the rest of us? Today we are given a variety of sophisticated and free software tools to select the content of interest to us from this massive and constantly expanding universe of professional media offerings and user-generated media. These tools include search engines, RSS feeds, and recommendation systems. But while they can help you find what to read, view, listen to, play, remix, share, comment on, and contribute to, in general they are not designed for carrying systematic social and cultural research along the lines of “cultural analytics” scenario I described earlier.
While a number of free data analysis and visualization tools have become available on the web during last few years (Many Eyes, Tableau, Google docs, etc.), they are not useful unless you have access to large social datasets. Some commercial web tools allow anybody to analyze certain kinds of trends in certain data sets they are coupled with in some limited ways (or at least, they wet our appetites by showing what is possible). I am thinking of already mentioned Google Ngram Viewer, Trends, Insights for Search, Blogpulse, and also YouTube Trends Dashboard, Social Radar, Klout. (Searching for “social media analytics” or “twitter analytics” brings up lists of dozens of other tools.)

For example, Google Ngram Viewer plots relative frequencies of words or phrases you input across a few million English language books published over last 400 years and digitized by Google (data sets in other languages are also available). You can use it to reveal all kinds of interesting cultural patterns. Here are some of my favorite combinations of words and phrases to use as input: “data, knowledge”; “engineer, designer”; “industrial design, graphic design.” In another example, YouTube Trends Dashboard allows you to compare most viewed videos across different geographic locations and age groups.

Still, what you can with these tools today is quite limited. One of the reasons for this is that companies make money by analyzing patterns in the data they collect about our online and physical behavior, and target their offerings, ads, sales events, and promotions accordingly; in other cases, they sell this data to other companies.

Therefore they don't' want to give consumers direct access to all this data. (According to an estimate by ComScore, in the end of 2007 five large web companies were recording "at least 336 billion transmission events in a month.") ("To Aim Ads, Web Is Keeping Closer Eye on You").

If a consumer wants to analyze patterns in the data which constitutes/reflects her/his economic relations with a company, here the situation is different. The companies often provide the consumers with professional level analysis of this data - financial activities (for example, my bank web site shows a detailed breakdown of my spending categories), their web sites and blogs (Google Analytics), or their online ad campaigns (Google AdWords).

Another relevant trend is to let a user compare her/his data against the statistical summaries of data about others. For instance, Google Analytics shows the performance of my web site against all web sites of similar type, while many fitness devices and sites allow you to compare your performance against the summarized performance of other users. However, in each case, the companies do not open the actual data, but only provide the summaries.

Outside of the commercial sphere, we do see a gradual opening up of the data collected by government agencies. For USA examples, check Data.gov ("Data.gov"), HealthData.gov ("Health.Data.gov"), and Radar.Oreilly.com ("GOV 2.0 Coverage and Insight"). As Alex Howard notes in Making Open Government Data Visualizations That Matter, "Every month, more open government data is available online. Local governments are becoming data suppliers." Note, however, that this data is typically statistical summaries, as opposed to transactional data (the traces of people online behavior) or their media collected by social media companies.

The limited access to massive amounts of transactional social data that is being collected by companies is one of the reasons why today large contemporary data-driven social science and large contemporary data-driven humanities are not easy to do in practice. (For examples of digitized cultural archives available at the moment, see the list of repositories ("List of Data Repositories") that agreed to make their data available to Digging Into Data competitors.) Another key reason is the large gap between what can be done with the right software tools, right data, and no knowledge of computer science and advanced statistics - and what can only be done if you do have this knowledge.

For example, imagine that you were given full access to the digitized books used in Ngram Viewer (or maybe you created your own large data set by assembling texts from Project Guttenberg, or another source) and you want software to construct graphs which show changing frequencies of topics over time, as opposed to individual words. If you want to do this, you better have knowledge of computational linguistics text mining (A search for "topic analysis" on Google Scholar returned 38,and 000 articles for the first field, and 38,000 articles for the second newer field.)

Or imagine that you interested in how social media facilitates information diffusion, and you want to use Twitter data for your study. In this case, you can obtain the data using Twitter API, or third party services that collect this data and make it available for free or for a fee. But again, you must have the right background to make use of this data. The software itself is free and readily available – R, Weka, Gate, Mallet, etc. - but you need the right training (at least some classes in computer science and statistics) and pior practical experience which uses this training to get meaningful results.

Here is an example of what can be done by people with the right training. In 2010 four researchers from Computer Science department at KAIST (South Korea's leading university for technology) published a paper entitled "What is Twitter, a social network or a news media?" Using Twitter API, they were able to study the entire Twittersphere as of 2009: 41.7 million user profiles, 1.47 billion social relations, 106 million tweets. Among their discoveries: over 85% of trending topics are "headline news or persistent news in nature." (Note that the lead author on the paper was a PhD student. It is also relevant to note that the authors make their complete collected data sets freely available for download, so it can be used by other researchers.) (For more examples of the analysis of "social flows", see papers presented at IEEE International Conference on Social Computing 2010.)

In this article I have sketcjed an optimmistic vision of a new existing paradigm opened to humanities and social sciences. I will then discussed four –objections to this optimistic vision. There are other equally important objections that I did not discussed because they are already debated in popular media and in academia by lots and lots of people. For example, a very big issue is privacy (would you trust academic researchers to have all your communication and behavior data automatically captured?)

So what conclusions should we draw from this analysis? Is it true that "surface is the new depth" – in a sense that the quantities of "deep" data that in the past was obtainable about a few can now be automatically obtained about many? Theoretically, the answer is yes, as long as we keep in mind that the two kinds of deep data have different content.

Practically, there are a number of obstacles before this can become a reality. I tried to describe a few of these obstacles, but there are also others I did not analyze. However, with what we already can use today (social media companies APIs, Infochimps.com data marketplace and data commons, free archives such Project Guttenberg, Internet Archive, etc.), the possibilities are endless – if you know some programming and data analytics, and also are open to asking new types f questions about human beings, their social life and their cultural expressions and experiences.

I have no doubt that eventually we will see many more humanities and social science researchers who are equally good at most abstract theoretical arguments as well the

latest data analysis algorithms which they can implement themselves, as opposed to relying on computer scientists. However, this requires a big change in how students particularly in humanities are being educated.

The model of big data humanities research that exists now is that of collaboration between humanists and computer scientists. It is the right way to start “digging into data.” However, if each data-intensive project done in humanities would have to be supported by a research grant which would allow such collaboration, our progress will be very slow. We want humanists to be able to use data analysis and visualization software in their daily work, so they can combine quantitative and qualitative approaches in all their work. How to make this happen is one of the key questions for “digital humanities.”

I am grateful to UCSD faculty member James Fowler for an inspiring conversation a few years ago about the collapse of depth/surface distinction. See his work at jhfowler.ucsd.edu.

References

"Big data." Wikipedia.org. n.p. Web. 17 July 2011. <http://en.wikipedia.org/wiki/Big_data>.

Cha, Meeyoung, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. 2007 ACM Internet Measurement Conference. Web. 17 July, 2011. < http://an.kaist.ac.kr/traces/papers/imc131-cha.pdf>.

Crandall, David J., Lars Backstrom, Daniel Huttenlocher, Jon Kleinberg. Mapping the world's photos. 18th international conference on World wide web, 2009. Web. July 17, 2011. <www.cs.cornell.edu/~dph/papers/photomap-www09.pdf>.

"Cultural Analytics Research Environment." 2007. Web. 2 May 2011. http://lab.softwarestudies.com/2008/12/cultural-analytics-hiperspace-and.html>.

"Data.gov." Data.gov. Data.gov. n.d. Web. 2 May 2011. <http://www.data.gov/>.

"Digital footprint." Wikipedia.org. n.p. Web. 31 March 2011. <http://en.wikipedia.org/wiki/Digital_footprint>.

"Flickr API Methods." Flickr.com. n.d. Web. 17 July 2011. < http://www.flickr.com/services/api/>.

"Digging into Data Challenge." DiggingIntodata.org. DiggingIntoData, n.d. Web. 31 March 2011. <http://diggingintodata.org/>.

Ellison, Nicole,, Rebecca Heino, Jennifer Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication, 11*(2). 2006. Web. 17 July 2011. <http://jcmc.indiana.edu/vol11/issue2/ellison.html>.

"Ervin Goffman." Wikipedia.org. n.p. Web. 27 March 2011. <http://en.wikipedia.org/wiki/Erving_Goffman>.

"GOV 2.0 Coverage and Insight." Radar.Oreilly.com/gov2. O'Reilly Media. n.d. Web. 2 May 2011. <http://radar.oreilly.com/gov2/>.

"Health.Data.gov." Data.gov/health. Data.gov. n.d. Web. 2 May 2011. <http://www.data.gov/health>.

"Help Find Genghis Khan's Tomb From the Comfort of Your Home." Wired.com. Wired.com, n.d. Web. 2 May 2011. <http://www.wired.com/geekdad/2010/07/mongolia-valley-of-the-khans/>.

"Internet Censorship by country." Wikipedia.org. n.p. Web. 30 April 2011. <http://en.wikipedia.org/wiki/Internet_censorship_by_country>.

Latour, Bruno. "Beware, your imagination leaves digital traces." Times Higher Education Literary Supplement (April 6, 2007). Print. July 15, 2011. Web. < http://www.bruno-latour.fr/poparticles/index.html,

"List of Data Repositories." DiggingIntodata.org. DiggingIntoData. 14 April 2011. Web. <http://diggingintodata.org/Home/Repositories/tabid/167/Default.aspx>.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a Social Network or a News Media?" The 19th international conference on World Wide Web. July 17, 2011. Web. <an.kaist.ac.kr/~haewoon/papers/2010-www-twitter.pdf>.

"Making open government data visualizations that matter." Gov20.govfresh.com. Alex Howard. 13 March 2011. Web. <http://gov20.govfresh.com/making-open-government-data-visualizations-that-matter/>.

Manovich, Lev. "Visualizing Large Image Collections for Humanities Research". *Media Studies Futures*. Ed. Kelly Gates. Blackwell, forthcoming 2012. Web. July 18, 2011. <http://manovich.net/DOCS/media_visualization.2011.pdf>

Manovich, Lev. "Motion Studies: Vertov's The Eleventh Year". Web. July 18, 2011. < http://www.flickr.com/photos/culturevis/5952098107/in/set-72157623326872241>.

"MIT Media Lab: Reality Mining." Reality.media.mit.edu. MIT, n.d. Web. 2 May 2011. <http://reality.media.mit.edu/>.

"Nielsen Company." Wikipedia.org., n.p. Web. 5 April 2011. <http://en.wikipedia.org/wiki/Nielsen_Company>.

"Nielsen ratings." Wikipedia.org., n.p. Web. 17 July 2011. <http://en.wikipedia.org/wiki/Nielsen_ratings>.

"One million manga pages." Softwarestudies.com. Software Studies Intitative. n.d. Web. 2 May 2011. <http://lab.softwarestudies.com/2010/11/one-million-manga-pages.html>.

"Social Computing: Introduction." The Third IEEE International Conference on Social Computing, n.p. Web. 17 July 2011. http://www.iisocialcom.org/conference/socialcom2011/.

"Software Studies: Cultural Analytics." Softwarestudies.com. Software Studies Initiative. n.d. Web. 2 May 2011. <http://lab.softwarestudies.com/2008/09/cultural-analytics.html>.

"Sociology in the 21st Century." Reality.media.mit.edu. MIT, n.d. Web. 2 May 2011. <http://reality.media.mit.edu/soc.php>.

"Stalin Epigram." Wikipedia.org., n.p. Web. 17 July 2011. < http://en.wikipedia.org/wiki/Stalin_Epigram>.

"The Petabyte Age: Because More Isn't Just More — More Is Different." Wired. 7 June, 2008. Print. July 15, 2011. Web. < http://www.wired.com/science/discoveries/magazine/16-07/pb_intro>.

"The size of the World Wide Web." WorldWideWebSize.com. Maurice de Kunder, n.d. Web. 31 March 2011. <http://www.worldwidewebsize.com/>.

"To Aim Ads, Web Is Keeping Closer Eye on You." NYTimes.com. Louise Story. 10 March 2008. Web.

<http://www.nytimes.com/2008/03/10/technology/10privacy.html?pagewanted=1>.
"US Census Bureau." Census.gov. USCensusBureau, n.d. Web. 2 May 2011. <http://www.census.gov/>.

"Watson (computer)." Wikipedia.org. n.p. Web. 2 May 2011. <http://en.wikipedia.org/wiki/Watson_%28artificial_intelligence_software%29>.

"What People Want (and How to Predict It)." SloanReview.MIT.edu. MIT. n.d. 2 May 2011. <http://sloanreview.mit.edu/the-magazine/2009-winter/50207/what-people-want-and-how-to-predict-it/>.

"WWW2009." www2009.org. n.d. Web. 2 May 2011. <http://www2009.org/>.

"WWW2010." www2010.org. n.d. Web. 2 May 2011. <http://www2010.org/www/>.