Lev Manovich

Cultural Data

Possibilities and limitations of the digital data universe

Digitization of cultural heritage over last 20 years has opened up very interesting possibilities for the study of our cultural past using computational "big data" methods. Today, as over two billion people create global "digital culture" by sharing their photos, video, links, writing posts, comments, ratings, etc., we can also use the same methods to study this universe of contemporary digital culture.

In this chapter I will discuss a number of issues regarding the "shape" of the digital visual collections we have, from the point of view of researchers who use computational methods. They are working today in many fields including computer science, computational sociology, digital art history, digital humanities, digital heritage and Cultural Analytics – which is the term I introduced in 2007 to refer to all of this research, and also to a particular research program of our own lab that has focused on exploring large visual collections.

Regardless of what analytical methods are used in this research, the analysis has to start with some concrete existing data. The "shapes" of existing digital collections may enable some research directions and make others more difficult. So what is the data universe created by digitization, what does it make possible, and also impossible?

The Islands and The Ocean

Before born-digital content, media creators first used physical and later electronic media (video and audio). Starting in the middle 1990s, gradually more and more of this content has being digitized. We can call such content *born-analog*.

The very first project to digitize cultural texts and make them freely available was Project Gutenberg that started in 1970. Today the largest sites for digitized content include Europeana (over 53 milion "artworks, artefacts, books, videos, and sounds from across

Europe" as of 2016[1]), Digital Public Library of America (over 13 million items as of 2016), HathiTrust (13 million volumes as of 2015), Digital Collections at the Library of Congress and Internet Archive. The latter contains digital collections of various types of media ranging from largest collection of historical software to 10.7 billion historical texts (as of 12/2016).[2]

The sites typically offer a number of useful ways to navigate these massive collections. For example, the Digital Public Library of America (DPLA) supports direct search, view by Timeline, Map view, and Thematic Exhibitions. Both DPLA and Europeana also encourage and help developers create experimental interfaces and apps that expand how their artifacts can be viewed and used. But in terms of using them for Cultural Analytics research, they do have one limitation. While the works in these and other collections can always be viewed online, not all works can be downloaded (or downloaded in mass using an API), because of the restrictions imposed by owners of the works.

The site which in my view is most interesting in this genre is Google Arts & Culture[3] It has fewer works but the most fluid interface. This site grew from the earlier Google Art Project that worked with many museums to scan artworks and then presented them online in a "virtual museum" interface. Today it offers virtual tours of many museums, millions of digitized artworks and photographs from the past, contemporary art. Media projects and photo stories are also created. The interfaces include zoom, timeline, search by color, thematic exhibitions, and also categories (artists, mediums, art movements, partners, names of objects, and places). When I was exploring the website (July 2016), it was offering 3,000 thematic exhibitions on all kinds of cultural topics.[4] When we started our own Cultural Analytics Lab (culturalanalytics.info) in 2007, it was a bet. While contemporary culture was already well represented on the web, the kinds of large-scale online digital collections with multiple navigation functions and API like Europeana or DPLA did not yet exist. But I assumed that within the next few years, millions of digital images of historical art, photography and other media would become available. However, it was not clear at that time how inclusive they would become.

In the article I wrote about Cultural Analytics in March 2009, I described my experience of trying to use the existing digital image collections available at that time.[5] I was interested in the following question: What did people paint around the world in 1930 – aside from a number of modernist "isms" that encompassed at best 150 artists (working in Paris, Amsterdam, Berlin and a few other cities) who are now included in the Western art historical canon? I was not thinking of "paintings in tens of thousands of small museums in small

1 Europeana collections, http://www.europeana.eu/portal/en (accessed July 25, 2016).
2 Internet Archive Search (mediatype: texts), https://archive.org/search.php?query=mediatype%3Atexts (accessed December 3, 2016).
3 Google Arts and Culture, Main Page, https://www.google.com/culturalinstitute (accessed July 26, 2016).
4 Google Arts and Culture, Exhibition Overview, https://www.google.com/culturalinstitute/beta/search/exhibit (accessed July 26, 2016).
5 Manovich 2009.

cities," rather of paintings of nationally "important" artists that have entered in art history canons in their countries.

I did a search on artstor.org – a leading commercial service for digital images of art used in most art history classes in U.S. and also in other countries. In 2009 it already contained close to one million digital images of art, architecture and design. These images came from many important USA museums, art collections, and university libraries.[6] To collect the images of artworks that are outside of the usual Western art historical canon on Artstor, we excluded Western Europe and North America from the search. This left the rest of the world: Eastern Europe, South-East Asia, East Asia, West Asia, Oceania, Central America, South America, and Africa. Not a small area! But when we searched Artstor for paintings done in these parts of the world in 1930, we only found a few dozen images. So, while there were very large numbers of images of paintings of canonical artists from Europe and USA painted in the same year, there were only a few images for a whole continent like East Asia.

This highly uneven distribution of digitized cultural artifacts is not due to Artstor's choices. Artstor does not digitize images itself. Instead, it makes images available that have been submitted by museums and other cultural institutions. The results of our search reflects what participating museums collect and what they think should be digitized first. In other words, a number of major US collections and a slide library of a major research university (where by 2007 the proportion of Asian students was 45%) together contained only a few dozen paintings created outside of the West in 1930 which were digitized. In contrast, searching for Picasso returned around 700 images. Describing this example, I wrote in this 2009 article:

> If this example is any indication, digital art repositories may be amplifying the already existed biases and filters of modern cultural canons. Instead of transforming the "top forty" into "the long tail," digitization can be producing the opposite effect.
> What remains outside of the digitized collections is all the rest: provincial nineteen century newspapers sitting in some library somewhere; millions of paintings in tens of thousands of small museums in small cities around the world; millions of thousands of specialized magazines in all kinds of fields and areas which no longer even exist; millions of home movies and photographs… This creates a problem for Cultural Analytics, which has a potential to map everything that remains outside the canon – and to begin writing a more inclusive cultural history without "great names." We want to understand not only the exceptional but also the typical; not only the few "cultural sentences spoken by a few "great men" but the patterns in all cultural sentences spoken by everybody else; what is outside a few great museums rather than what is inside and what has already been discussed extensively and too many times.

6 The very first large institutional collection that formed the core of Artstor was the slide library of the University of California, San Diego (UCSD) – the same university where I had been teaching since 1996. The library had over 200,000 slides, and they were all digitized and included in Artstor. In 2009, this was the largest single collection in Artstor. The slides were either directly created by art history faculty teaching in Visual Art Department, or by art library staff following lists of images faculty provided. This collection is very interesting because it reflects the biases of art history as it was taught over a few decades when color slides were the main media for teaching and studying art.

I worried that what has been digitized, is only an "island," and that a massive cultural "ocean" remains inaccessible for quantitative analysis. Luckily, such amplification of biases and focus on only "what is important" did not happen. Exploring the online libraries of digitized cultural artifacts seven years later, I am amazed by their richness and variety. The reason is that Europeana, DPLA, Library of Congress, NYPL, Internet Archive or Google Arts & Cultures do not just offer us images of high art like art *museums*. Instead, they are extensions of traditional *libraries*. And the libraries in modern times have an important function besides offering readers books and periodicals – they are places to which numerous people and organizations donate their archives. As these archives started to be digitized, an amazingly rich and varied historical cultural landscape started to emerge online.

For example, here are three examples among hundreds of digital image collections from the New York Public Library (NYPL):

> "Photographs of The Catskill Water Supply System in Process of Construction." 55 albumen print photographs created between 1906 and 1915.[7]
> "Buttolph Collection of Menus" – A collection of Miss Frank E. Buttolph (1850–1924), a somewhat mysterious and passionate figure, whose mission in life was to collect menus donated to NYPL in 1899, 18,964 digitized items.[8]
> "Catalog of the Chiroptera, by G. E. Dobson" – 31 digitized prints from a 1878 book.[9]

And here are examples listed in the blog post from europeana.eu referred to as "highlights of the new datasets ingested in the last months":

> Almost 100 objects (drawings, paintings, photographs) from Telegraph Museum in UK.
> Over 3 000 photographs, XIX and XX century, mostly buildings from Culture Centre in Helsingborg.
> Collection of 620 botanical drawings by Georg Schweinfurth from Botanic Garden and Botanical Museum Berlin-Dahlem.[10]

Comparing these collections with those of the digital image offerings of the largest *art museums*, we find that they are complete opposites of each other. Although modern art museums' collections like that of libraries also developed through both their purchasing programs and donations, what was donated to them – or what museums chose to accept – was quite different. Libraries ended up housing millions of all kinds of heterogeneous items, few of them financially valuable. In contrast, modern art museums have traditionally focused on what has been recognized as very valuable. Indeed, original European "museums" included estates of very rich people, parts of royal palaces, or treasures of cathedrals and churches. For example, Vatican Museums originated in 1506 when Pope Julius II purchased

7 http://digitalcollections.nypl.org/collections/photographs-of-the-catskill-water-supply-system-in-process-of-construction (accessed July 26, 2016).
8 http://digitalcollections.nypl.org/collections/buttolph-collection-of-menu (accessed July 26, 2016).
9 http://digitalcollections.nypl.org/collections/catalogue-of-the-chiroptera-by-ge-dobson (accessed July 26, 2016).
10 http://www.europeana.eu/portal/en/search?q=europeana_collectionName%3A11630%2A&view=grid (accessed March 8, 2017). See also Strzelichowska 2016.

the ancient sculpture of Laocoön and his Sons and placed it on public display. (I should note that digitized collections of *design and crafts* museums such as Victoria and Albert in London or Cooper-Hewitt in New York are closer to that of libraries – their holdings are more varied and also organized in more categories than those of art museums.)

Libraries vs. Museums

However, there is also another aspect in museum's history. Some of the original European museums contained not art but "curiosities." One such famous museum is The Kunstkamera that was established in St. Petersburg in 1716 by Peter the Great to present "natural and human curiosities and rarities." Another is the British Museum that opened in London in 1759, that initially showed a private collection of the physician and scientist Sir Hans Sloane.

Art history since the 20th century has created a highly controlled system that divides our visual heritage into "art" and everything else, and organizes the former by artists (their national origin, time period, and medium and style). The digital collections of art museums today also look ordered and systematic.

We are used to their ordered classifications. In comparison, the meta-collections of digitized visual artifacts by Europeana, DPLA and others may remind us of the cabinets of curiosities. Instead of a military-like "parades" of art history played in physical museums or on their sites, we find "trivia" and "ephemera." (The latter word comes from Greek and New Latin where it referred to insects or flowers that were alive sometimes for less than a day.)

Browsing through page after page describing endless collections that often contain a few dozen or even only a few items – like the ones in the examples above – I often feel uncanny. In this view, the past looks un-periodic and un-systematized. Endless "deposits" of human material cultures have remained inside libraries, have then been digitized and are now connected by common metadata standards, web protocols, Javascript code, APIs and other computer machinery.

Labyrinth, kaleidoscope, Kunstcamera, Memex' hypertext, random access memory, relational database – none of these models describe my experience of navigating digital cultural collections. For instance, consider Europeana with its 53 million items. The idea behind this massive multi-year project was to connect digitized artifacts from thousands of European museums and regional archives. So, rather than having to search all their individual sites, you can use the Europeana platform as a single point of access. The platform provides a common interface to all of the objects but it does not store them. They are stored at individual museums and archives. European Film Gateway, one of Europeana's projects, does the same for dozens of European film archives.

Technically and conceptually, this works brilliantly. But experientially, the result has some unintended consequences. Instead of creating a kind of "united Europe" – a single pan-European space for cultural heritage – Europeana may be fragmenting it. As I browse through endless separate collections or individual items from these collections that fit my search terms, countries, geographic relations, and time periods are dissolved. Instead of a

"European" continent, it feels that I am looking at random survived files of many alien civilizations that got all mixed together.

This feeling is created by both very heterogeneous topics, and by equally heterogeneous styles. Photographs created in all kinds of techniques, engravings, etching, newspaper illustrations, covers of cigarette cases, early hand-colored photos, paintings … images are in rectangular formats, round frames, part of a text page, drawn in a corner of a hand written letter… texts typed, types set, hand written, printed on early dot matrix printers, carefully drawn with a brush … every possible subject and form of visual communication is here. (If Instagram platform during 2010–2015 can be thought as the extreme example of visual constraints, with all image being the same size and format and belonging to one medium, a digital historical collection is the other extreme).

But this heterogeneity, richness and variety is actually a good thing. It makes us aware of how rigid and limited our concepts of an "image" are today – a few clearly separated mediums, rectangular formats, and also separation between images and texts. So, while the abundance of communication "species" in digital libraries is on first sight disorientating – and it is certainly a challenge for large scale analysis using Computer Vision systems initially developed for contemporary photos – in the long run it is best for us. It forces us to face the human visual culture as it really exists historically – thousands of variations and their combinations, rather a net set of a small number of categories.

Cultural Sampling

The "islands" of digitized historical contents are constantly growing. But will they ever be big enough to let us understand the "ocean" – i.e., construct a sufficiently detailed map of the human visual history of the last few centuries? Richness and variety do not mean comprehensiveness. In other words: while digitization and organization of digitized items by Europeana, DPLA, and other projects continues, the most basic question for any quantitative study of cultural history remains unaddressed. This question is, how can we compile *representative samples* that systematically cover everything created in a particular period, geographic area and media – or in many such periods and areas together?[11]

Anthropologists do use sampling methods in their research when they excavate sites or study groups of people (such as in urban anthropology that looks at contemporary cities). But there is a basic large question which is more difficult to address: Since the kinds and quantities of artifacts that remained from various ancient civilizations vary significantly, do they together add to a representative sample? (Of course, as excavations of sites and analysis of new artifacts continue, this sample is being continuously refined.)

Since I am a historian of modern visual culture and media of the last 200 years, I am confident that for this period we do not have any comprehensive sample of visual culture in this period before the arrival of social media. So, while the "islands" are increasing

11 For an overview of different sampling methods, see Cook 2011 and Chambers / Skinner 2003.

in size and number, reconstructing the whole ocean maybe may become very difficult. I am using the term "sample" in the sense it is used in statistics: a smaller subset of the larger data. Constructing proper samples and determining the validity of predictions based on these samples is a one of the main areas of statistics. In all social sciences including sociology, demographics, psychology, and political science these questions are particularly crucial, since these disciplines often use small human groups for surveys or observation. Construction of proper samples is also crucial for marketing research, human-computer interaction research and all other applied fields where researchers want to find people's attitude about existing products, interest in new products and new features, their lifestyle aspirations, etc. And while the arrival of big social media data in the second part of the 2000s has changed the situation significantly, because now businesses can follow online millions of individuals tracking what pages they visit, what they click on, which ads they look at, and what they purchase, small groups continued to be widely used. (You can ask people who agreed to participate all kinds questions, or place them in situations and see what they chose – something which is not always possible online.)

We do not have systematic samples of modern visual and media culture. Instead we have numerous separate collections and archives that are being digitized. Therefore, the kind of question I asked in 2009 –What did people painted around the world in 1930? – is still unanswerable. And for many other questions, the situation is even worse. Consider for example the history of photography. While working on a book about Instagram aesthetics in the context of modern design, art and photography, I had a pretty big sample of Instagram: 16 million photos shared in 17 global cities between 2012 and 2016.[12] It is important to note that these are not photos with particular tags. Instead, they are *all* geo-coded photos shared in larger city areas during a particular period. According to a few of computer science publications that analyzed large samples of Instagram posts in 2014, during that time Instagram users shared locations for 20% of their photos.[13] This means that our datasets also represent approximately 20% of all Instagram photos shared in a given area and period. From a sampling point of view, these are very good samples. Not only are they quite substantial but we also know what part of a "population" is represented. ("Population" in statistics is a technical term that refers to the whole data that for practical reasons is not accessible to us. Instead, we can use small samples from which we can probabilistically infer characteristics of the whole data.)

I certainly did not expect to find anything like these samples for vernacular photography in the 20th century. But I assumed that after all digitization work of the last twenty years, I can easily find samples of at least few thousand digitized photographs for particular decades, and maybe even for particular countries. It turns out that nothing like this existed.

What has been digitized and made available online are various collections of vernacular photography from particular private collections. They added certain photos to their collec-

12 Manovich 2016.
13 Manikonda et al. 2014.

tions because each photo was *interesting* to them for some reason. Museum exhibitions of vernacular photography that I consulted were similarly "non-objective" – they were assembled by curators who had particular curatorial ideas. I also found some user groups on Flickr with "found photographs" contributed by group members. Every collection I consulted was the result of individual or groups' taste and ideas of what should be included. Often people were only interested in more "artistic" and "avant-garde" examples of vernacular photography, rather than the typical.

To my knowledge, nobody has ever thought to create a *representative sample* that would contain characteristics of the field of vernacular photography as a whole in particular historical periods, types of cameras and printing, and so on (for example, photos made with Kodak Brownie cameras of 1900, or first portable 35-mm Leicas in 1925, or prints using Kodacolor after 1942, or Polaroid prints after 1972.) So now that we have learned from computer science studies of massive social media samples that we can look at any culture as a statistical population asking about distributions, averages, variance, clusters, and so on, we want similar historical samples. But they do not exist.

For example, the National Gallery of Art in Washington presented an exhibition in 2010 called *The Art of American Snapshot, 1888–1978: From the collection of Robert E. Jackson*. According to the curators, "Organized chronologically, the exhibition focuses on the changes in culture and technology that enabled and determined the look of snapshots. It examines the influence of popular imagery, as well as the use of recurring poses, viewpoints, framing, camera tricks, and subject matter, noting how they shift over time."[14]

The online exhibition catalog shows that curators did an excellent job of capturing a number of aspects of vernacular photography and its evolution. However, since the exhibition only had 200 photographs for a 90-year period, that meant that the historical map exhibition constructs was very "low resolution" (to use the spatial metaphor) and also not complete. If we want to understand differences in snapshot photography between different countries, or find gradual changes in style or subjects that are not related only to the introduction of new photography technologies, or see if there may be some regional or demographic differences, we cannot accomplish this with 200 photos.

For a comparison, consider the Gallup U.S. Daily poll.[15] For this poll, Gallop interviews (over the phone) 500 people across U.S. every day. For a country of 300 million people, this looks like a tiny sample. But because Gallup selects people at random and conducts these interviews every day, it accumulates 15,000 responses per month, and 175,000 per year.[16] We also learn that "Gallup also weights its final samples to match the U.S. population

14 The Art of the American Snapshot, 1888–1978: From the Collection of Robert E. Jackson, National Gallery of Art, Washington, DC, October 7–December 31, 2007, Online Exhibition Information: https://www.nga.gov/exhibitions/snapshotinfo.shtm (accessed January 10, 2017).

15 Gallup, Online Methodology Center, http://www.gallup.com/178685/methodology-center.aspx (accessed January 10, 2017).

16 Gallup, How Does the Gallup U.S. Daily Work? – http://www.gallup.com/185462/gallup-daily-work.aspx?utm_source=METHODOLOGY (accessed January 10, 2017).

according to gender, age, race, Hispanic ethnicity, education, region, population density, and phone status." This weighting is done using data from a number of other surveys. For example, to weight by population density, Gallop uses U.S. Census reports. This systematic approach to sampling and analysis of the results is typical of all natural and social sciences, public administration, demographics, public polls, marketing research, and countless other areas. In fact, the only area where it is absent is humanities.

The question humanists have been asking is about *canon*, and how to make canons in their field more representative. There is a parallel here with the kind of weighting Gallup and other organizations that collect demographic data do. However, sometimes in the attempts to compensate for a lack of representation of older canons, the new canons are "weighted" more towards groups that were previously not represented. So as a result, we once again get something completely driven by ideologies, rather than a balanced sample.

A "balanced cultural sample" can be defined in multiple ways, all equally informative and complementary to each other. For example, we can include a proportion of all works produced in particular media, period, and place. Or we can focus instead not on what has been produced, but what audiences actually read, watched, or listened to. We may decide to select only works that achieved certain recognition (which would be equivalent of likes and favorites in contemporary social media), or disregard this information. But whatever we do, we need a *systematic procedure*, not simply a taste judgment. Statistics has developed a sophisticated theory of sampling which includes many methods, and since these methods are used today in all sciences, they should be adopted for analysis of historical cultural artifacts as well – if we are interested in understanding them as a kind of ecological or geological system, where all participants and artifacts are important – as opposed to only a set of "masterpieces."

The idea of creating systematic and representative samples of culture is interesting by itself, because it leads to all kinds of follow up questions. And since our textbooks, museums, cultural portals, classes, and documentaries always represent human arts and cultures using only selected examples, the questions about cultural sampling are important in general, even if we are not conducting quantitative analysis. They relate to how we understand, represent and teach human cultural history – and also how we think about our cultural present, with its new scale of numbers of participants, their cultural interactions and experiences.

For example, imagine a hypothetical scenario where we can include any painting created in France in the 19th century in our sample. Now imagine that we want to create a representative sample, so we randomly select X number of paintings. Such a sample will include several academic salon paintings, realistic paintings, portraits and so on. And it would miss the 19th century art which we now recognize as most important – works by Impressionists and Post-impressionists. Why? It has been estimated that 13 key French Impressionists artists together created 13,000 paintings and pastels during their lifetimes.[17] But this is a

17 Cutting 2005.

very small number in comparison to all paintings created by artists living in France during the whole 19th century. So, a random sample would likely miss them all.

This is exactly the same problem, which accompanies a great deal of quantitative social media research in Computer Science. In many articles, authors explain how they carefully construct a random sample drawn from all users of Pinterest, Instagram or Twitter. Using such samples, they then develop statistical models that account for some characteristics of the behavior and posts of these users. This research is very interesting and important. But using a single global sample of a network with hundreds of millions of people from most countries in the world sharing billions of daily text posts, images and video has serious limitations. We can only see the "typical." So we miss all kinds of regional variations, and presence and activity of endless users who don't have the typical behaviors and posts. In other words, if any of these networks have their own "Impressionists," they are not visible in the analysis that uses single random samples.

Sometimes, the sampling procedures used end up only including particular types of users. For example, in the paper "Analyzing User Activities, Demographics, Social Network Structure and User-Generated Content on Instagram" (2014), the researchers state: "To the best of our knowledge, we believe this is the first paper to conduct an extensive and deep analysis of Instagram's social network, user activities, demographics, and the content posted by users on Instagram."[18] This is how they describe the method they used to create a user sample for their study:

> First, we retrieved the unique IDs of users who had pictures that appeared on Instagram's public timeline by using Instagam API, which displays a subset of Instagram media that was most popular at the moment. This process resulted in a sample of unique users. However, after careful examination of each user in this sample, we found that these users were mostly celebrities (which explains why their posts were so popular). To avoid the sampling bias, for each user in this sample, we crawled the IDs of both their followers and friends, and later merged two lists to form one unified seed user list which contained 1 million unique users.

The final dataset has 5,659,795 images for 369,828 users (the rest had private accounts). Out of these images, 1,064,041 have geo-locations. But how well are these users representing the Instagram universe? Most people follow other people as opposed to celebrities. People who do follow celebrities and their friends are likely only one type of Instagram user. Additionally, given that the number of Instagram users in every country differs, with the biggest countries also often having larger number of users, such a "random" sample likely better represents some countries than others.

These considerations do not invalidate the results in this and all other papers that use a single large sample from massive global social networks. Their findings are valid. They just may not apply to every type of user or type of post on such networks. (Note that we are not talking about individual users but groupings, each with their own characteristics. In other words, these are like 19th century Impressionists who had common characteristics.)

18 Manikonda et al. 2014.

We also need to recall here perhaps the most fundamental "Achilles' heel" of statistics. "The goal of statistics is to represent the facts in the most condensed way" (1833). But we pay a big price for such compression. The measures used in descriptive statistics summarize some population (i.e., a set of items) but they may not correspond to any concrete members of this population. For an example, let's take a series of numbers: 1,1,2,3,2,9,9,10,11,11,11. The average (called "mean" in statistics) of this series is 6.36. But we don't have any actual numbers close to this mean! No. 4, 5, or 6. Instead, we have two "clusters": 1 to 3, and another one from 9 to 11. (This is called a *bimodal distribution*.)

In other words, the standard statistical measures of a large population can easily miss the presence of various groupings in this population. So, if we represent some "cultural population" – be it 19th century paintings or 20th century cinema, Instagram today, or global music videos – with a single random sample, we can miss all kinds of groupings (1960s New Wave or 1920s Soviet Montage school in cinema history; contemporary music videos from India, Korea, Vietnam, Thailand or Kazakhstan which have their own differences despite overall similarity; and so on.) And the characteristics which we will find may describe the "average" which never existed in reality. That is, it may not correspond to any actual group. And rather than capturing the presence of multiple distinct groups, it can hide them from view.

In fact, I would like to claim that in human societies and cultures there are no "averages." Certainly, we can follow Adolphe Quetelet who in the early 1830s was the first to start to measure the physical characteristics of humans such as height and weight and found that their distributions followed "normal" curves.[19] If we perform such measurements today, we will find similar distributions. And, in a sample of a million people, certainly many would have the exact height specified by the mean. In the same way, if we for example measure the length of tens of thousands of modern novels, we will find that some do have exactly the same length as the average novel.

But such results only hold if we limit the study of cultural artifacts, interactions, and experiences to one characteristic at a time. If we look at several selfies sampled from Instagram, we can calculate the average degree of smile, size of a face in a photo, and its positon. And if the sample size is big enough, some actual selfies will have exactly the same numbers as the averages. But just as a face of every person is unique, like their fingerprints, their photos are also unique. So if we multiply the number of characteristics, eventually we will not find any real selfie that matches the sample averages on all of them. The same applies to any other type of cultural expression, past or present.

There is one field that does think about cultural sampling and it is using statistical methods to create and analyze these samples. This field is the *sociology of culture*. The most well-known book in this field remains famous *Distinction: A Social Critique of the Judgement of Taste* by French sociologist Pierre Bourdieu. Published in 1979, it has been recognized as one of the ten most important books of sociology in the 20th century. Bourdieu offered

19 Tyler 1872.

powerful intellectual ideas and theories that connected people's cultural tastes and their socio-economic statuses. These theories were grounded in the statistical analysis of two large surveys of tastes of the French public conducted in the 1960s. Bourdieu collaborated with French "data scientists" (to use contemporary term) who developed a new analytical and visualization methods to represent relations between many elements, and he used this method in all of his later studies including *Distinction.*

Today sociologists of culture continue to use surveys of groups of people, but they also use samples from cultural publications. One example of the former is a study where the researchers "asked 1544 German-speaking research participants to list adjectives that they use to label aesthetic dimensions of literature in general and of individual literary forms and genres in particular (novels, short stories, poems, plays, comedies)."[20] The example of the later is a study called "Institutional Recognition in the Transnational Literary Field, 1955–2005." It uses "a sample of articles from 1955, 1975, 1995 and 2005 in French, German, Dutch and US elite papers (*N*=2,419)."[21] Here is another example: an analysis of fashion discourse during 1949–2010 that uses 1301 fashion reviews from The New York Times and The International Herald Tribune.[22] Although such samples are rather small in comparisons to social media scale, they are sufficient to answer particular questions the researchers asked in these studies.

When I first thought of cultural analytics in 2005, I imagined being able to construct detailed world-wide maps of particular fields – such as painting, cinema, graphic design or music video – for long historical periods. But as I realized that digitization efforts are not creating systematic samples such maps would require, I had to abandon these ideas for the time being. So instead, I focused on a different type of sampling that I could do given what has been digitized – by type of media. Starting in 2008, in our lab, we have worked on over 40 datasets that cover almost every major type of visual media today. We analyzed comics and Manga series, video games, feature films, documentaries, motion graphics, music video, political video ads, print magazines, historical photographs, born-digital photographs and other images, and interactive virtual worlds. We also deliberately included dataset that lie at the extremes of a high – low and professional – non-professional dimensions: from paintings of van Gogh, Mondrian and Rothko to 10 million Instagram photos shared in New York City by 5 million people. And we have also deliberately balanced Western and non-Western cultural sources. The latter include Japanese video games, music videos from across Korea, Instagram photos shared in seventeen global cities that cover four continents. We published analysis using Instagram photos shared in Tel Aviv, Israel during Fallen Soldiers and Victims of Terrorism Remembrance Day, and another analysis of Instagram photos shared during February 2014 Maidan revolution in Kiev, Ukraine.

20 Knoopa et al. 2016.
21 Verboord et al. 2015.
22 Van de Peer 2014.

In fact, the advantage of using social media data is that it is not "canonical" or "national." Popular networks such as Facebook, Instagram, and others are used in every country except the few where they are/were blocked for periods of time (In the case of Facebook, Bangladesh, China, Iran, North Korea, Syria[23]). As of May 2016, the messaging app WhatsApp that started in China was used in 109 countries, with one billion users sending 42 billion messages daily.[24] And by the same time, 80% of Instagram 500M active users were outside U.S.[25]

For example, when we were creating our Instagram samples datasets between 2012 and 2016, Instagram API allowed anyone to download all geo-tagged photos shared within a particular rectangular area defined by its latitude and longitude. Each area could be 5km × 5km in size, and collecting from a number of areas was not more complicated. So it was equally easy to download images from parts of Manhattan, or Moscow, or Bangkok, or Kiev, and so on. (To download all geotagged images shared during five months in Manhattan, we combined a number of areas to enclose the island in a large rectangle, and then filtered out the data outside of Manhattan boundaries).

This means that in practice, comparing many areas from around the world is as easy as comparing nearby areas from the same city – as long people share sufficient amounts of social media in these global areas. The global perspective is "built in" in social media. This of course also applies to the standard formats, constraints and affordances particular networks and apps provide for their users. Everyone who used Twitter between 2007 and 2017 had to fit their messages into the same 140 characters. Everyone who was using Instagram between 2010 and 2015 had to submit to its square image format and the same size: 640 × 640 (or 612 × 612). Everyone has access to exactly the same functions (adding hashtags, optional geo-tagging, etc.) and the same UI. This by itself raises an important question: does social media software lead to less diversity in user-generated content? This was one of the key questions for me during my eight years of research.

Data Representation

However, like every other type of data about society, social media data has its own limitations, and they are not insignificant. I will briefly discuss five issues which are all about *representation* – what gets represented (and available for research) and what is absent. While the use of social networks and the web continues to grow around the word, billions of people do not use them. Here is a concrete example from our own research of how this situation limits what we can "see" using their data. In 2014, Twitter agreed to provide selected researchers with access to any part of their data if they used it in new interesting ways. Thirteen hundred labs from around the world applied, and we were one of six labs

23 Kirkland 2014.
24 Smith 2016.
25 Facebook 2016.

that won. I asked Twitter to give us all tweets with geo-located images shared with them. Twitter added images functionality in 2011, and we were given access to all tweets with geo-located images shared worldwide between 2011 and 2014. When we plotted locations of a random sample of 100 million tweets from this data, approximately half of the populated Earth surface had no coverage.

The second issue has to do with demographics of users who do use social networks. In "developed" countries and global megacities, people from all demographic groups use the networks. In a country like the USA, there is no significant differences in social network use between women and men, or different races, or people with different level of education – but there are still big differences between age groups. This is also true globally – although the differences are getting smaller with time. A report on social media use among people who were online in 34 countries in first quarter in 2016 found that 92% of those who are in 45–54 age group have social media accounts; for people in 55–65 age group the figure is 82 %.[26]

In many developing countries, the proportions of people using social networks among those using the web are higher than in developed countries. At first, this looks like good news because it could mean that we get data on cultural activities of larger proportion of populations in these countries. However, the reality is different. As the report explains, "As many as 98% of Internet users in countries like Malaysia, Brazil, Indonesia and Vietnam are on at least one network. In part, that's a result of their lower Internet penetration levels, which means online adults in these regions are more likely than their counterparts in Europe or North America to come from young, urban and relatively affluent segments. [27]

The third issue is uneven spatial distribution of social networks activity and content even in big urban areas where we see very high use – until we zoom in. The amount of sharing and participation can vary dramatically between city areas, as we show in the *Inequaligram* project. We collected and analyzed 7,442,454 public geo-tagged Instagram images shared in Manhattan over five months. The inequality we found between the more populated and less populated parts of Manhattan was staggering. We found that the ratio between a square km area with most images and the area with least images was 250,000:1. According to our analysis, 50% of all images shared by local residents are within only 21% of Manhattan area. For visitors, this difference is almost twice as big: 50% of their images were shared in only 12% of the Manhattan area. In summary, even for such a densely populated urban area as Manhattan, its Instagram collective image only reflects part of it and not all.

The forth issue is what content people share, what comments they make, and what they are willing to say online. Social networks are not a mirror of society. Just as people in other areas of their lives play roles, follow norms, present particular identities and behave in ways expected from them (by "mainstream," or their particular "subcultures," or "tribes), they do this online. And because their posts and comments can be seen by all other network

26 GWI 2016.
27 Ibid.

users (unless they make posts or the whole account private), appear in Google search, and are saved by the networks, shared with marketers, etc., they are likely to be extra-careful. And, just as with professional cultural products, some of user-generated content is driven by conventions, stereotypes and models people see around them. For example, we find endless photos in "table top" genre on Instagram created by regular users, overwhelming proportions of selfies smile (see our selfiecity.net and selfiecity.net/London for more details), and travel photos follow their own conventions. All this means that *the "culture" we can analyze using social media is its own universe*, and not a simple sample of people's cultural activities, taste and opinions outside the networks.

Finally, the fifth issue is access to social media data. In the middle part of 2000, all large social networks created APIs that allow people to freely download large data samples containing user posts and all public information about them visible online – date and time a post was shared, location (if user shared this information), username, tags, comments, and numbers of likes and re-shares. In the case of visual networks such as Instagram and Flickr, image and video along with their user descriptions and all other information was also available for downloads. Flickr launched its API in 2004, and Facebook and Twitter in 2006.[28]

While these APIs were intended for developers building apps that use data from the platforms, and for users to share contents between networks and also their blogs, computer science researchers, data visualization artists, and other creative technologists realized that they can also freely access this data, and numerous studies and projects were created. Hundreds of thousands of computer and social scientists and students used these APIs to download data, analyze it and publish papers.

However, there have always been limits on how much data can be downloaded. For example, during the period we were actively downloading Instagram data (2012–2016), it had a limit of 3000 images per hour, and only images from the last few days were available. Nevertheless, we were able to assemble 16 million Instagram photos shared in 17 global cities in different periods between 2012 and 2016. But given that in 2016 people are sharing 80 millions of images on Instagram per day, what we were able to assemble was a tiny portion.

However, because of the concerns with privacy and unauthorized use of posts, some of the biggest networks gradually limited or closed API access to bulk user data. Facebook limited the use of its API on April 30, 2015, and Instagram stopped allowing bulk downloads on June 1, 2016. At this moment (end of 2016), Twitter is still accessible, along with some networks popular in particular geographic areas such as Russian VK.

In summary, we know that social media and the web are not used by everyone; the proportions and demographics of those who use social media varies from place to place; and what people publish and share constitutes its own cultural reality as opposed to being a transparent window into the realities outside. We should always keep these limitations in mind. At the same time, using the web and social media data and contemporary

28 Lane 2012.

technologies for tracking and analyzing it *questions the very idea of representation*. This concerns the very foundation of modern research methods based on *sampling*.

These methods assume that for practical reasons we cannot have access to the complete "population" (i.e., full data). We can only access and analyze one or more samples of the population. Accordingly, modern statistics is divided into two areas. *Inferential statistics* is a set of methods for estimating characteristics of the population based on its sample(s). *Descriptive statistics* only describes the properties of whatever data we have, and it does not assume that this data came from a larger population.

However, when we analyze web and social media content and interactions, we often can have *full data*. Certainly, the companies that run social networks, media sharing sites or publishing platforms can record all interactions happening on their platforms. This is true for Facebook, YouTube, Twitter, Pinterest, Spotify, Amazon, Scribd, Shutterstock, Behance, academia.edu, and other social media and publication services. This does not mean that a company will be analyzing all their data, or keeping it forever, or even have its own researchers work on it – because companies don't want to sued, have bad publicity or get in trouble with governments. So the data is anonymized, sampled when needed, and only particular parts of the data are made available to internal researchers depending on what lab they work for. However, the largest companies certainly take advantage of having massive data about user interactions on their platforms, using it to train systems that recommend other users to follow or other videos to watch and decide which posts from friends to show, select trending topics etc. Big data is also driving the main source of income for big social media companies – i.e. automatic advertising systems such as Google AdWords and Facebook Ads.

Although academic researchers do not have direct access to complete data from these companies, it is possible to use their APIs to download complete data that satisfies particular criteria, such as all activity on a particular platform within a particular time period. Many papers use such datasets. In our own work, we also followed this approach. We were using Instagram API to download all publically shared geo-coded images shared in a particular geographic area over a period of time. In fact, every Instagram dataset we used was generated in this way. For example, to create a dataset of 7,442,454 public Instagram images shared in Manhattan over five months, we used a single Mac to run our custom download program 24/7 during this whole period. As far as we know, the images we downloaded are *all* images people shared within this area and time with geo-location (which constitutes approximately 20% of everything shared).

Why may we want to use *complete* cultural data? If we are only interested in extracting general patterns, characteristics, and types – for example, the 10 most common types of images on Instagram – we certainly do not need all of the data. But such summarization and aggregation common to the use of statistical methods in 19th and 20th century is only one way to use cultural data. As I explained above, using small samples from diverse cultural "population" (such as trillions of Instagram images) may only reveal the "typical" and "most popular" and miss "regional variations" and "presence and activity of endless users who do not have the typical behaviors and posts." Therefore, ideally Cultural Analytics should try

to obtain and analyze complete data generated by some cultural process (be it career of a single photographer or all photos shared on Instagram).

Rather than only treating culture as "data points" that together create patterns that we want to discover, disregarding the individual points afterwards, Cultural Analytics should pay equal attention to both patterns and individual artifacts, experiences and interactions. As creators and audience members, we engage and enjoy concrete artifacts and experiences, and not "patterns." A particularly successful artifact is often described as "unique" – i.e. it cannot be reduced to already existing patterns. As aesthetic subjects, we search and enjoy such uniqueness. One of the goals of Cultural Analytics is to help us find truly unique artifacts in the infinite universes of media now being created. And even if other artifacts are not unique in most ways, they may still have something unique in other ways, which can get lost if we reduce them to patterns. For instance, every human face is unique, and therefore even the most conventionally-driven photo of this face will be special for us. (In this aspect, Cultural Analytics should combine special perspective of sciences and of humanities – the former's concern with general laws and regularities, and the latter's concern with unique cultural objects.)

To conclude, I would like to note one techno-cultural development of the last 20 years that connects many issues I have discussed – the rise of search as a new dominant mode for interacting with information. This development is just one of many consequences of the dramatic and rapid expansion of information and content being produced which we have experienced since the middle of the 1990s. To serve the search results, Google, Bing, Baidu, Yandex, and other search engines analyze many different types of data – including both metadata of particular web pages (so-called "meta elements") and their content. For example, according to Google, its search engine algorithm uses more than 200 input types.[29]

However, Google, Yandex or Bing do not reveal the measurements of web pages they analyze – they only serve their conclusions, i.e. which sites best fits the search string user entered determined by their propriety algorithms that combine these measures. In contrast, the goal of Cultural Analytics is to enable what we may call "deep cultural search" – give users the open-source tools so they themselves can analyze any type of cultural content in detail and use the results of this analysis in new ways.

References

Cook, Sam (2011): "Sampling methods." In: *Revise Sociology*, May 4, https://revisesociology.wordpress.com/2011/05/04/5-sampling-methods/ (accessed March 3, 2017).

Chambers, R. L. and C. J. Skinner (eds., 2003): *Analysis of Survey Data*, Chichester: Wiley.

Cutting, James E. (2005): *Impressionism and Its Canon,* Lanham: University Press Of America.

Facebook (2016): "Instagram." In: Marketing auf Facebook, https://www.facebook.com/business/products/ads/instagram-ads (accessed July 31, 2016).

Google (2017): "Algorithms." In: *Google Inside Search – How Search Works*, https://www.google.com/insidesearch/howsearchworks/algorithms.html?hl=en (accessed January 17, 2017).

29 Google 2017.

GWI (2016): *GlobalWebIndex's Quarterly Report on the Latest Trends in Social Networking*, Q1, http://insight.globalwebindex.net/social (accessed January 17, 2017).

Kirkland, Alice (2014): "10 countries where Facebook has been banned." In: *Index* (Index on Censorship), February 4, https://www.indexoncensorship.org/2014/02/10-countries-facebook-banned/ (accessed January 17, 2017).

Knoopa, Christine A., Valentin Wagnera, Thomas Jacobsen, Winfried Menninghaus: "Mapping the aesthetic space of literature 'from below'." In: *Poetics* 56 (June), pp. 35–49.

Lane, Kin: "History of APIs." In: *API Evangelist* (Blog), December 12, 2012, https://apievangelist.com/2012/12/20/history-of-apis/ (accessed December 3, 2016).

Manikonda, Lydia, Yuheng Hu, and Subbarao Kambhampati (2014): "Analyzing user activities, demographics, social network structure and user-generated content on instagram." *arXiv* preprint arXiv:1410.8099.

Manovich, Lev (2016): *Instagram and Contemporary Image,* online publication available under Creative Commons license, http://manovich.net/index.php/projects/instagram-and-contemporary-image (accessed January 10, 2017).

– (2009): "How to Follow Global Digital Cultures, or Cultural Analytics for Beginners." In: Felix Stalder and Konrad Becker (eds.), *Deep Search*: The Politics of Search Beyond Google, Piscataway: Transaction Publishers, pp. 1–13, http://manovich.net/index.php/projects/how-to-follow-global-digital-cultures (accessed July 26, 2016).

Van de Peer, Aurélie (2014): "Re-artification in a World of De-artification: Materiality and Intellectualization in Fashion Media Discourse (1949–2010)." In: *Cultural Sociology* 8.4, pp. 443–461.

Smith, Craig (2016): WhatsApp Statistics. In: *DMR Stats | Gadgets* (Blog), http://expandedramblings.com/index.php/whatsapp-statistics/ (accessed July 26, 2016).

Strzelichowska, Aleksandra (2016): "Maggy's picks: new content in Europeana." In: *europeana blog*, July 25, http://blog.europeana.eu/2016/07/maggys-picks-new-content-in-europeana/ (accessed July 27, 2016).

Tylor, Edward Burnett: "Quetelet on the Science of Man." In: *Popular Science Monthly* 1 (May 1872), pp. 45–55.

Verboord, Marc, Giselinde Kuipers, Susanne Janssen (2015): "Institutional Recognition in the Transnational Literary Field, 1955–2005." In: *Cultural Sociology* 9.3, pp. 447–465.