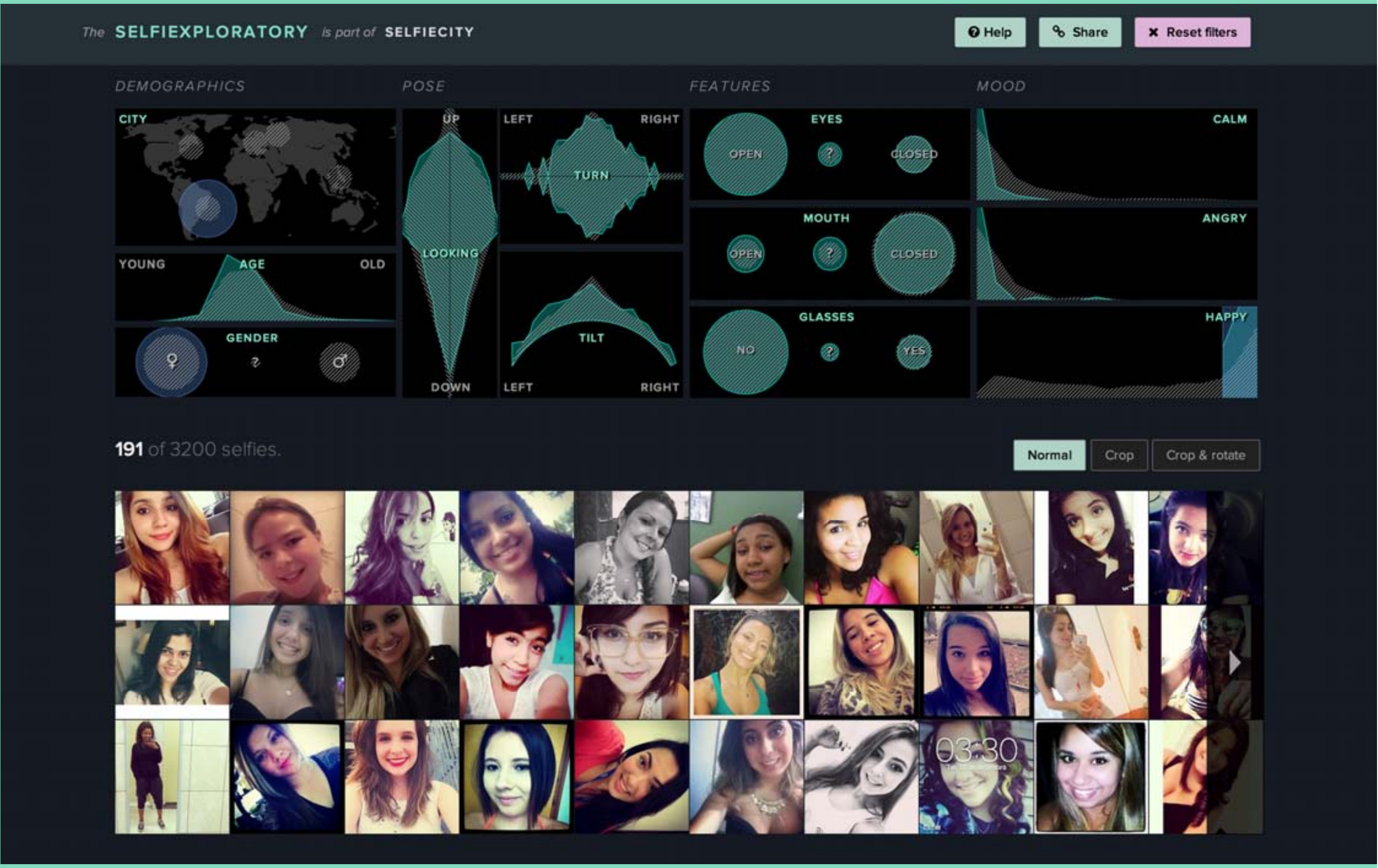
Featured Article



191 of 3200 selfies.

Figure 1: An example of the web interface for interactive exploration of image collections from our project *Selfiecity* (http://selfiecity.net, 2014).
A visitor can filter the collection of 3200 Instagram self-portraits by using graphs in the upper part of the screen. The left column contains graphs and controls for filtering images using cities, ages, and gender information. Age and gender estimates were obtained by using Amazon Mechanical Turk service. Other columns contain graphs that show features extracted by face analysis software from https://rekognition.com/.
They include face orientation (up/down, left/right, and degree of tilt), presence of smile and glasses, open/close eyes and mouth and seven emotions detected in faces (only three emotion graphs are included).

Invited Article

Data Science and Digital Art History

Lev Manovich

Abstract: I present a number of core concepts from data science that are relevant to digital art history and the use of quantitative methods to study any cultural artifacts or processes in general. These concepts are objects, features, data, feature space, and dimension reduction. These concepts enable computational exploration of both large and small visual cultural data. We can analyze relations between works on a single artist, many artists, all digitized production from a whole historical period, holdings in museum collections, collection metadata, or writings about art. The same concepts allow us to study contemporary vernacular visual media using massive social media content. (In our lab, we analyzed works by van Gogh, Mondrian, and Rothko, 6000 paintings by French Impressionists, 20,000 photographs from MoMA photography collection, one million manga pages from manga books, one million artworks of contemporary non-professional artists, and over 13 million Instagram images from 16 global cities.) While data science techniques do not replace other art historical methods, they allow us to see familiar art historical material in new ways, and also to study contemporary digital visual culture.

In addition to their relevance to art history and digital humanities, the concepts are also important by themselves. Anybody who wants to understand how our society "thinks with data" needs to understand these concepts. They are used in tens of thousands of quantitative studies of cultural patterns in social media carried out by computer scientists in the last few years. More generally, these concepts are behind data mining, predictive analytics and machine learning, and their numerous industry applications. In fact, they are as central to our "big data society" as other older cultural techniques we use to represent and reason about the world and each other – natural languages, material technologies for preserving and accessing information (paper, printing, digital media, etc.), counting, calculus, or lens-based photo and video imaging. In short, these concepts form the data society's "mind" – the particular ways of encountering, understanding, and acting on the world and the humans specific to our era.

Keywords: data science, data mining, visualization, data analysis, features, metadata, social media, algorithm, dataset

Introduction[1]

Will art history fully adapt quantitative and computational techniques as part of its methodology? While the use of computational analysis in literary studies and history has been growing slowly but systematically during 2000s and first part of 2010s, this has not yet happened in the fields that deal with the visual (art history, visual culture, film, and media studies).

However, looking at the history of adoption of quantitative methods in the academy suggests that these fields sooner or later will also go through their own "quantitative turns." Writing in 2001, Adrian Raftery points out that psychology was the first to adopt quantitative statistical methods in 1920s-1930s, followed by economics in 1930s-1940s, sociology in 1960s, and political science in 1990s.[2] Now, in 2015, we also know that humanities fields dealing with texts and spatial information (i.e., already mentioned literary studies and history) are going through this process in 2000s-2010s. So I expect that "humanities of the visual" will be the next to befriend numbers.

This adaption will not, however, simply mean figuring out what be counted, and then using classical statistical methods (developed by the 1930s and still taught today to countless undergraduate and graduate students pretty much in the same way) to analyze these numbers. Instead, it will take place in the context of a fundamental social and cultural development of the early 21st century – the rise of "big data," and a new set of methods, conventions, and skills that came to be called "data science." Data science includes classical statistical techniques from the 19th and early 20th century, additional techniques and concepts for data analysis that were developed starting in 1960s with the help of computers, and concepts from a number fields that also develop in the second part of the 20th century around computers: pattern recognition, information retrieval, artificial intelligence, computer science, machine learning, information visualization, data mining. Although the term "data science" is quite recent, it is quite useful as it acts as an umbrella for currently most frequently used methods of computational data analysis. (Alternatively, I could have chosen machine learning or data mining as the key term for this article, but since data science includes their methods, I decided that if I am to refer to all computational data analysis using a single term, data science is best right now.)

Data science includes many ideas developed over many decades, and hundreds of algorithms. This sounds like a lot, and it is. It is much more than can be learned in one or two graduate methods classes, or summarized in a single article, or presented in a single textbook. But rather than simply picking particular algorithms and techniques from a large arsenal of data science, or borrowing whatever technique happens to be the newest and therefore is currently in fashion (for example, "topic modeling" or "deep learning") and trying to apply them to art history, it is more essential to fist understand the most fundamental assumption of the field as a whole. That is, we in art history (or any other humanities field) need to learn the core concepts that underlie the use of data science in contemporary societies. These concepts do not require formulas to explain, and they can be presented in one article, which is what I will attempt here. (Once we define these core concepts, a variety of terms employed in data science today can also become less confusing for the novice.)

Surprisingly, after reading thousands of articles and various textbooks over last eight years, I have not found any short text that presents these core concepts together in one place. While many data science textbooks, of course, do talk about them, their presentation often takes place in the context of mathematically sophisticated techniques or particular applications which can make it hard to understand the generality of these ideas.[3] These textbooks in general can be challenging to read without computer science background.

Since my article is written for humanities audience, it is on purpose biased–my examples of the application of the core concepts of data science come from humanities as opposed to economics or sociology. And along with an exposition, I also have an argument. I will suggest that some parts of data science are more relevant to humanities research than others, and therefore beginning "quantitative humanists" should focus on learning and practicing these techniques first.

From World to Data

If we want to use data science to "understand" some phenomenon (i.e., something outside of a computer), how do we start? Like other approaches that work on data such as classical statistics and data visualization, data science starts with representing some phenomenon or a process in a particular way. This representation may include numbers, categories, digitized texts, images, audio, spatial locations, or connections between elements (i.e., network relations). Only after such a representation is constructed, we can use computers to work on it.

In most general terms, creating such a representation involves making three crucial decisions:

What are the boundaries of this phenomenon? For example, if we are interested to study "contemporary societies," how can we make this manageable? Or, if we want to study "modern art," how we will choose what time period(s), countries, artist(s), and artworks, or other information to include? In another example, let's say that we are interested in contemporary "amateur photography." Shall we focus on studying particular groups on Flickr that contain contributions of people who identify themselves as amateur or semi-pro photographers, or shall we sample widely from all of Flickr, Instagram, or other media sharing service

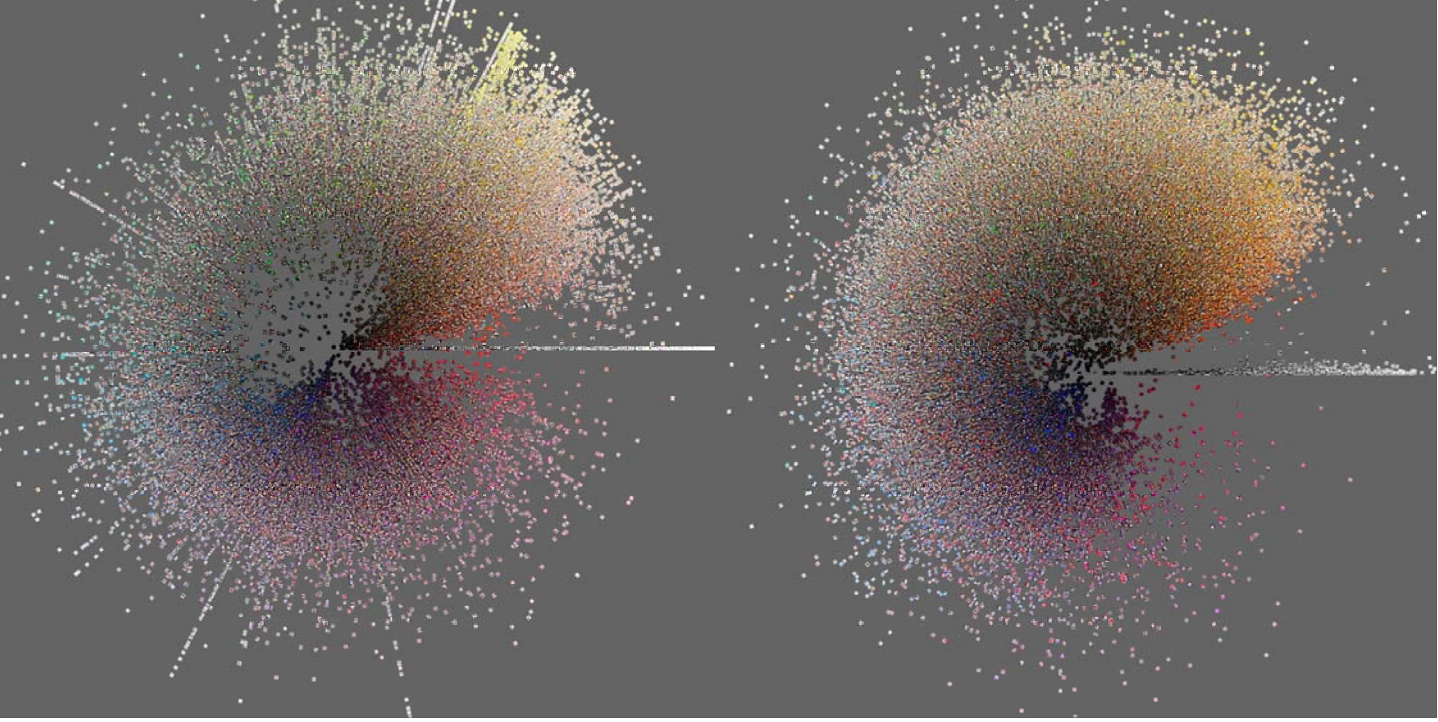


Figure 2: An example of visualizations of image collections that uses image features automatically extracted by a computer. (Source: our project Phototrails, http://phototrails.net/, 2013). Left: a random sample of 50,000 Instagram images from Bangkok. Right: a random sample of 50,000 Instagram images from Tokyo. In each visualization, images are organized by mean hue (angle) and brightness mean (distance to the center).

- since everybody today with a mobile phone with a built-in camera automatically becomes a photographer.

What are the objects we will represent? For example, in modern art example, we may include the following "objects" (in data science they can be also called data points, records, samples, measurements, etc.): individual artists, individual artworks, correspondence between artists, reviews in art journals, passages in art book, auction prices. (For example, 2012 *Inventing Abstraction* exhibition in MoMA in NYC featured a network visualization showing connections between artists based on the number of letters they exchanged.[4] In this representation, modernist abstract art was represented by a set of connections between artists, rather than any other kind of object I listed above.) In a "society" example, we can for instance choose a large set of randomly chosen people, and study social media they share, their demographic and economic characteristics, their connections to each other, and biological daily patterns as recorded by sensors they wear. If we want to understand patterns of work in a hospital, we may use as elements people (doctors, nurses, patients, and any others), also medical procedures to be performed, tests to be made, written documentation and medical images produced, etc.

What characteristics of each object we will include? (These are also referred to as metadata, features, properties, or attributes.). In humanities, we usually refer to characteristics that are already available as part of the data (because somebody already recorded them) and characteristics we have added (for example, by tagging) as metadata. In social science, the process of manually adding descriptions of data is called coding. In data science, people typically use algorithms to automatically extract additional characteristics from the objects, and they are referred as features (this process is called "feature extraction"). For example, artists' names is an example of metadata; average brightness and saturation of their paintings, or the length of words used in all titles of their works are examples of features that can be extracted by a computer. Typically features are numerical descriptions (whole or fractional numbers) but they can also take other form. For example, a computer can analyze an image and generate a few words describing content of the image. In general, both metadata and features can use various data types: numbers, categories, free text, network relations, spatial coordinates, dates, times, and so on.

Fig. 1 shows the examples of metadata and features used in one of the projects of my lab. We assembled a collection of 3200 Instagram self-portaits and created an interactive web interface for exploration of this collection. The examples of metadata are the same of the cities where Instagram images were shared. The features include estimate of the people age and gender, and results of computer analysis (emotions, face position and orientation, presence and amount of smile, etc.)

Fig. 2 shows examples of visualizations that present large image collections using features. 50,000 Instagram images shared in Bangkok are compared with 50,000 Instagram images shared in Tokyo using two features extracted by computer analysis – average color saturation, and average hue.

I suggest that in digital art history we adapt the term "features" to refer to both information that can be extracted from objects through computer analysis and the already available metadata. In natural and social sciences, the most common term is "variable," and it is used in the context of experiments. But since in humanities we do not do systematic experiments like in the sciences, for us the term "features" is better. It only implies that we represent objects by their various characteristics - but it does not imply any particular methods of analysis. (However, in the section "Classical Statistics and Statistical Graphs" below I will use "variable" because this was the term used during the period described in this section.)

Although it is logical to think of the three questions above as three stages in the process of creating a data representation– limiting the scope, choosing objects, and choosing their characteristics – it is not necessary to proceed in such linear order. At any point in the research, we can add new objects, new types of objects and new characteristics. Or we can find that characteristics we wanted to use are not practical to obtain, so we have to abandon our plans and try to work with other characteristics. In short, the processes of generating a representation and using computer techniques to work on it can proceed in parallel and drive each other.

Depending on our perspective, we could assume that a phenomenon (such as "contemporary society," for example) objectively exists regardless of how we study it (i.e., what we use as objects and their properties). Or we can also assume that a phenomenon is equal to a set of objects and their properties used in different qualitative and quantitative studies, publications and communication about it (books, articles, popular media, academic papers, etc.) That is, a phenomenon is constituted by its representations and the conversations about it. My description of the three questions above assumes the first position, but this is done only for the convenience of explaining the steps in moving "from world to data."

Objects + Features = Data

Together, a set of objects and their features constitutes the "data" (or "dataset").

People in digital humanities always like to remind us that data is something that is "constructed" - it does not just exist out there. But what does this mean exactly? Any data project, publication, or data visualization includes some aspects of the phenomena and excludes others. So it is always "biased." But this this is something that in most cases can be corrected. For example, in the case of a survey of social media use that only samples people in the U.S. and asks them particular questions about their social media use (such as popular Pew Internet surveys), we can add people from different countries and we can also ask them additional questions. But the concept of "data" also contains more basic and fundamental assumptions that cannot be changed, and this is equally important. Before we can use computers to analyze a phenomena or activity, it has to be represented as a finite set of individual objects and also a finite set of their features. For example, consider music. The computational analysis of music typically divides a music track into very small intervals such as 100 ms and measures some properties of each sample. In this way, analog media is turned into discrete data.

How is a "data representation" of some phenomenon today different from other kinds of cultural representations humans used until now, be they representational paintings, literary narratives, historical accounts, or hand drawn maps? Firstly, a data representation is modular, i.e. it consists from separate elements: objects and their features. Secondly, the features are encoded in such a way that we calculate on them. This means that the features can take a number of forms – integers, floating point numbers, categories represented as integers or text labels, etc. – but not just any form. And only one format can be used for each feature.

But the most crucial and interesting difference, in my view, is that a data representation has two types of "things" which are clearly separated: objects and their features. What is chosen as objects, what features are chosen, and how these features are encoded – these three decisions are equally important for representing phenomena as data – and consequently, making their computable, manageable and knowable though data science techniques.

Practically, objects and features can be organized in various ways, but the single most common one is a familiar table. An Excel spreadsheet containing one worksheet is an example of a table. A table can be also stored as a standard text file if we separate the cells by some characters, such as tabs or commas (these are stored as .txt or .csv files, respectively). A relational database is a number of tables connected together though shared elements.

A table has rows and columns. Most frequently, each row is reserved to represent one object; the columns are used to represent the features of the objects. This is the most frequent representation of data today, used in every professional field, all natural and social science, and in government services. It is the way data society understands phenomena and individual, and acts on them.

Classical Statistics and Modern Data Science: From One to Many Variables

Classical Statistics and Statistical Graphs: Dealing with One or Two Variables

Statistics comes from the word "state," and its rise in the 18th and 19th century is inseparable from the formation of modern bureaucratic, "panopticon" societies concerned with counting, knowing and controlling its human subjects, and also its economic resources. Only in the middle of the 19th century, the meaning of "statistics" changes – it becomes a name for an independent discipline concerned with producing summaries and reasoning about any collections of numbers, as opposed to only numbers important for the states and industry.

For our purposes – understanding core principles of contemporary data science and how they are different from classical statistics - we can divide the history of statistics in three stages. The first stage encompasses 18th and first part of the 19th century. During this stage, statistics means collecting and tabulating various social and economic data. During this stage, William Playfair and others develop a number of graphing techniques to represent such collections visually. Playfair is credited with introducing four fundamental techniques: bar chart and line graph (1786), and pie chart and circle graph (1801). The titles of the books where Playfair first used these techniques exemplify the kinds of number gathering that motivated the invention of these techniques: *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England during the Whole of the Eighteenth Century* (1786); *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe* (1801).

These graphing techniques invented by Playfair are still most popular today, despite the invention of other data visualization techniques in later periods. Note that they all visualize only a single characteristic of objects under study. Built into all statistical and graphing software and web services, they continue to shape how people use and think about data today – even though computers can do so much more!

(Note: When you make a graph in a program such as Excel, you often also select an extra column that contains labels. So even though these techniques show only patterns in a single characteristic – i.e., some numbers stored in a single column - in order to include the labels for the rows, a second column is also used. But it is not counted as a data variable.)

In the 19th century topical maps also became popular. An example is a map of country where the brightness of each part represents some statistics, such as literacy rate, crime rate, etc.[5] Although such maps are two-dimensional graphical representation, they still only use a single variable (i.e. a quantity is used to determine the brightness or graphic style for each part of the territory shown on a map).

In the second stage of statistics history (1830s-1890s), the analytical and graphical techniques are developed to study the relations between two characteristics of objects (i.e., two variables). In 1880s Francis Galton introduces concepts of correlation and regression. Galton was also probably the first to use a technique that we now know as a scatterplot. Today scatterplot remain the most popular techniques for graphing two variables together.[6]

One of the most famous uses of statistics in the 19th century exemplifies “data imagination” of that period. In 1830s Belgian Adolphe Quetelet measured height and weight in a large number of children and adults in different ages and published his results in a book that was to become famous: *A Treatise on Man and the development of his aptitudes* (1835). Quetelet concluded that these characteristics measured in large numbers of people follow a bell-like curve (called now Gaussian or normal distribution). Along with analyzing height and weight as single separate variables, Quetelet also analyzed their relations in many people, creating in 1832 the modern “body mass index.” He found that, on the average, “the weight increases as the square of the height.”[7]

More canonical examples can be found in the book considered to be the founding text of sociology – *Suicide* by Émile Durkheim (1897).[8] The book has dozens of data tables. Durkheim used such summary statistics to compare suicide rates in different population groups (Protestants vs. Catholics, single vs. married, soldiers vs. civilians, etc.). He then proposed theoretical explanations for these differences. (Note that the book does not have a single statistical graph, not any statistical tests of the significance of the differences.)

In the third stage (1900-1930) the statistical concepts and methods for the analysis of one or two variables were further refined, extended, systematized, and given rigorous mathematical foundation. These include summarizing a collection of numbers (measures of central tendency such as mean and median, and measures of dispersion, such as variance and standard deviation), analyzing relations between two variables (correlation and regression), doing statistical tests, and designing experiments that gather data to be analyzed with statistics. The key work in this period was done by Karl Pearson, Charles Spearman, Ronald Fisher working in England and the American Charles Pierce.[9]

The content of contemporary introductory textbooks on statistics for college students is very similar to the content of Fisher's book *Statistical Methods for Research Workers* published in 1925 – and we may wonder why we keep using the concepts and tools developed *before* computers to analyze "big data" today. The practicality of manual computation was an important consideration for the people who were consolidating statistics in the beginning of the 20th century. This consideration played key role in shaping the discipline, and consequently still forms the "imaginary" of our data society.

Modern Data Science: Analyzing Many Features Together

In the 20th century, statistics gradually develop methods for the analysis of many variables together (i.e., "multi-variable analysis"). The use of digital computers for data analysis after WWII facilitates this development. As computers get faster, analyzing more and more features together becomes more practical. By the early 21st century, a representation of phenomena that has hundreds or thousands of features has become commonplace. The assumption that objects are described using a large number of features is standard in data science, and this is one of its differences from classical early statistics.

While basic statistical classes today still focus on the techniques for the analysis of one or two variables, data science always deals with many features. Why? In social sciences, the goal is explanation, and its ideal method is systematic experiments. The goal of experiments is studying how some conditions may be affecting some characteristics of a phenomenon or activity. For example, how does a person's background (place of birth, ethnicity, education, etc.) affect her current position and salary? How does an athlete's preparation and diet affect her performance in multiple sports competition? If there are many factors and effects, it is not easy to understand what is affecting what. Therefore, in an ideal 20th century experiment, a researcher wanted to only measure one condition and one effect. All other factors ideally are hold constant. In an experiment, one condition (called independent variable) is systematically changed, and the values of a single characteristics thought to be affected by this condition (called dependent variable) are recorded. After the experiment, statistical techniques (graphing, correlation, regression and others) are used to study the possible relationship between the two variables.

In modern data science the key goal is automation. Data science (like Artificial Intelligence field earlier) aims to automate decision-making, prediction, and production of knowledge. Based on the available information about the customer, shall a bank make a loan to this customer? Does a photograph contain a face? Does this face match an existing face in a database? Based on the phrase a search engine user typed, what web pages are most relevant to this phrase? In principle, each of these questions would be best answered if a human or a team spent sufficient time studying all relevant information and coming up with the answer. But this would require lots and lots of time for a single answer. Given the scale of information available in many situations (for example, the web contains approximately 14-15 billion web page), this time will approach infinity. Also, how many different conditions (variables) the data may contain, even infinite time will not help humans fully understand their effects.

Therefore, credit ranking systems, face recognition systems, search engines and countless other technological systems in our societies use data science algorithms and technologies to automate such tasks. In summary, the goal of data science is automation of human cognitive functions – trying to get computers to do cognitive tasks of humans, but much faster.

Achieving this goal is not easy because of what computer sciences call "semantic gap." This is the gap between knowledge that a human being can extract from some data, and how computer sees the same data. For example, looking at a photograph of a person, we can immediately detect that the photo shows a human figure, separate the figure from the background, understand what a person is wearing, face expression, and so on. But for a computer, a photograph is only a matrix of color pixels, each pixel defined by three numbers (contributions of red, green and blue making its color). A computer has to use this “low-level” information to try to guess what the image represents and how it represents it.. Understanding a meaning of a text is another example of the semantic gap. A human reader understands what the text is about, but a computer can only “see” a set of letters separated by spaces.

Trying to “close the semantic gap” (this is the standard phrase in computer science publications) is one of the motivations for using multiple features. For example, the case of image analysis, a computer algorithm may extract various features from images, in addition to just the row RGB values of the pixels. Computer can identify regions that have similar color value and measure orientations of lines and properties of texture in many parts of an image. The hope is that together all these features will contain enough information for an algorithm to identify what an image represents.

In summary, 20th century statistical analysis and contemporary data science use variables in exactly the opposite way. Statistics and quantitative social science that uses it ideally wants to isolate one independent and one dependent variable, because the goal is understanding the phenomenon. Data science wants to use many features in the hope that together they contain the right information for automating recognition, classification, or another cognitive task.

Feature Space

Before we move on, a quick summary of what we learned so far about representing phenomena as data. We represent a phenomenon as a set of objects (also called data points, measurements, samples, or records) that have features (also called attributes, characteristics, variables, or metadata). Together, the objects and their features is what we mean by "data" (or "datasets"). Features can be represented in a variety of ways: whole and fractional numbers, categories, spatial coordinates, shapes and trajectories, dates, times, etc.

These are the basic requirements/ conventions of modern data analysis and also data visualization. Now, let's start our next "lesson." To the concepts above (objects and features) we are going to add the third core concept: feature space.

We assume that our data is stored in a table. But now we will conceptualize our data table as a geometric space of many dimensions. Each feature becomes one of the dimensions. Each object becomes a point in this space. This is a "feature space," and it is the single most important and also most relevant for us in humanities the concept from contemporary data science, in my opinion.

The easiest way to understand this is by considering a familiar 2D scatter plot. Such a plot represents data in two dimensions. One dimension (X) corresponds to one feature (i.e., one column in a data table); the second dimension (Y) corresponds to a second feature (another column in the table). (Fig. 3 uses a space of two features to compare paintings Vincent van Gogh created in Paris and in Arles).

If we want to also add a third feature, we can make a three-dimensional scatterplot, if our software allows this. And if we have 10 features, our plot now conceptually exists in a 10-dimensional space. And so on. However, while mathematics and computer science have no problems working with spaces that may have arbitrary numbers of dimensions, we humans cannot see or plot them directly, because we exist physically and can only see in three dimensions. But we can still use computational techniques to think about objects in multi-dimensional spaces, and study their relations.

Use of Feature Space in Data Science

Once we represent some phenomenon or a process as a set of objects defined by many features, and conceptualize this representation as a multi-dimensional space, many analytical operations become possible. Many fundamental applications of data science correspond to such different operations, explicitly or implicitly.

For example, we can use a set of techniques called exploratory data analysis (described below) to "look" at the structure of the space and visualize it. To perform cluster analysis, we divide the space into parts, each containing points that are more similar to each other than to points outside this part. In classification, we identify the points belonging to two or more categories. ("Binary classification" deals with two categories; "multiclass classification" deals with more than two classes. If cluster analysis and classification sound similar, it is because they are, but while the first is completely automatic technique, classification needs some data that already has category information.) In many search algorithms, a computer finds the points in the space that are most similar to the input terms (these are points that that are closest to the input in feature space – see the section on measuring distance in feature space below). Some of the recommendation algorithms work similarly – starting from the points that a user have previously favored, they find and display other points the closest to them (of course they do not show the points directly but the media objects represented by them such as movies, songs or people to follow on social media).

These operations rely on more basic ones such as computation of similarity/difference between points in a feature space. The degree or similarity/difference can be equated with the simple geometric distance between the points in the space).

I would like to mention a few more terms because they are so common in data science that you will inevitably encounter them. "Exploratory data analysis" is also called "unsupervised learning." In contrast, "supervised learning" needs part of the data already labeled as belonging to this or that category. Algorithms then use this labeled data along with its features to "learn" how to classify new data. The practical application of unsupervised learning is part of the field of "predictive analytics.")

Among the contemporary applications of data science, probably the most common is automatic classification. However, in my view it is the least interesting one for humanities. Why should we use computers to classify cultural artifacts, phenomena or activities into a small number of categories? Why not instead use computational methods to question the categories we already have, generate new ones, or create new cultural maps that relate cultural artifacts in original ways?

This is why this article does not go into any detail about the widely used data science methods you will find extensively covered in standard data sciences textbooks and courses – i.e., classification methods. But while these textbooks typically only devote a small part to exploratory data exploration, I think that for the humanities we need to reverse this ratio.

Figure 3: Comparing paintings created by van Gogh in Paris (left) and Arles (right) on brightness and saturation dimension. X-axis – average brightness; y-axis – average saturation. The visualization shows that on these dimensions, van Gogh's Paris paintings have more variability than his Arles paintings. We can also see that most paintings created in Arles occupy the same part of the brightness/saturation space as Paris paintings; only a small proportion of Arles's paintings explore the new part of this space (upper right corner). (Visualization by Lev Manovich / Software Studies Initiative).

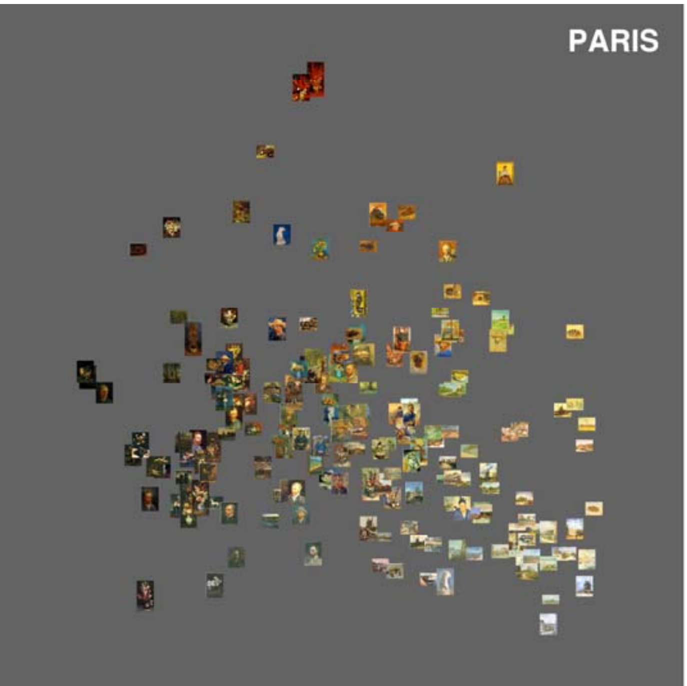
Accordingly, in the rest of this article I will discuss data exploration techniques.

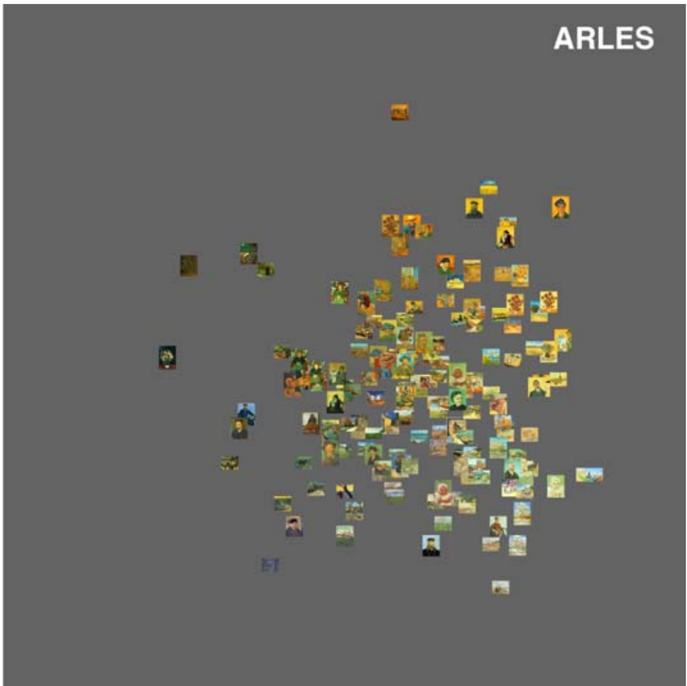
Difference as Distance in Feature Space

We learned that we could conceptualize a set of objects with many features as points in a multi-dimensional space. What are the benefits of such a representation for humanities?

The most basic method of humanities until now has been the same as in everyday human perception and cognition – comparison. (This is different from natural and social sciences that have been using mathematics, statistics, data visualization, computation and simulation to study their phenomena and objects.) In a 20th century art history class, a two-slide projector setup allowed for simultaneous viewing and comparison between two artifacts. Today in an art museum, a label next to one artifact point out the similarities between this artifact and a few other artifacts (or artists) in the same exhibition.

Manual comparison does not scale well for big data. For example, for our lab's project *On Broadway* that visualizes a single street in NYC using many data sources, we collected all publically visible Instagram images from the whole NYC area for five months in 2014. The result was 10.5 million images. Let's say we want to understand some patterns in this nice sample of contemporary vernacular photography – what are the subjects of these images, what are common and uncommon compositions, how this may differ between parts of NYC, how many





images are by people using techniques of professional commercial photography, and so on. Simply looking at all these images together will not allow us to answer such questions. And in fact, no popular commercial or free image management or sharing software or web service can even show that many images together in a single screen.

However, data science techniques can allow us to answer the questions such as the ones I posed above for very large datasets. By representing each image as a point in a space of many features, we can now compare them in quantitative way. In such representation, the visual difference between images is equated with a distance in feature space. This allows us to use computers to compute differences between as many images (or other types of cultural objects) as we want. Such computation then becomes basis for doing other more "high-level" operations: finding clusters of similar images; determining most popular and most unusual types of images; separating photos that use the language of professional photography, and so on.[10]

Using only two features is useful for developing an intuition about measuring distance in a multi-dimensional feature space. Consider a visualization in Fig. 3 showing images of van Gogh paintings that uses average brightness (X axis) and color saturation (Y axis). The geometric distance between any two images corresponds to the difference between them in brightness and saturation. Note that, of course, this example disregards all other types of difference: subject matter, composition, color palette, brushwork, and so. However, this is not only a limitation but also an advantage – by letting us isolate particular features, we can compare artifacts only on dimensions we want.

We can also compute and add as many features as we want. And although we may not be able to visualize and see directly the space of, for example, 50 or 500 features, we can still calculate the distance between points in this space. If the distance between two points is small, it means that the corresponding objects are similar to each other. If the distance between two points is large, it means that the corresponding objects are dissimilar to each other.

There are many ways to define and calculate distance, and data science uses a number of them. One popular way that is easiest to understand is using Euclidian geometry. (Another popular way is "cosine similarity," defined as the cosine of an angle between two vectors in feature space.) Note that in these calculations, we do not need to give equal weight to all features; if we believe that some of them are more important, we can also make them more important in the computation.

The concept of a geometric feature space allows us to take the most basic method of humanities – a comparison – and extend it to big cultural data. In the same time, it allows us (or forces us, if you prefer) to quantify the concept of difference. Rather than simply saying that artifact "A" is similar to artifact "B," and both "A" and "B" are dissimilar to "C," we can now express these relations in numbers. While this quantification may appear to be unnecessary if we are only considering a small number of artifacts, once we start dealing with thousands, tens of thousands, millions, and beyond, it becomes a very useful way of comparing them.

Exploring Feature Space

Let's say we want to understand some cultural field in a particular period – Ming Dynasty Chinese painting, realist art in Europe in late 19th century, graphic design in 1990s, social media photography in early 2010s, etc. What kinds of subject matter (if the field has a subject matter), styles and techniques are present? How they develop over time? Which of them were more popular and which were less popular? Art historians so far relied on human brain's abilities that developed evolutionary to see patterns and understand similarity and difference between sets of artifacts. They seemed to do well without using mathematics, graphic methods, statistics, computation, or contemporary data science. But the price for this "success" was the most extreme exclusion – considering only tiny sample of "important" or "best" works from every period or field. In the words of the pioneer of digital humanities Franko Moretti,

> “What does it mean, studying world literature? How do we do it? I work on West European narrative between 1790 and 1930, and aleady feel like a charlatan outside of Britain or France... 'I work on West European narrative, etc....' Not really, I work on its canonical fraction, which is not even one per cent of published literature. And again, some people have read more, but the point is that there are thirty thousand nineteenth-century British novels out there, forty, fifty, sixty thousand—no one really knows, no one has read them, no one ever will. And then there are French novels, Chinese, Argentinian, American...[11] ”

Moretti's point certainly applies to all other humanities fields; and it applies even more to the analysis of contemporary culture. Who can look at even a tiniest percentage of photos shared on Instagram every hour – or for example hundreds of million Instagram photos with a tag #fashion? Who can visit hundreds of cities around the world in a single month to understand the differences in street fashion between all of them? Who can browse through billions of web pages to understand the landscape of current web design?

Let's apply the concepts we learned – objects, features, feature space, distance in feature space, and various operations this representation allows (exploration, clustering, etc.) to this problem. First we need create an appropriate data set. As we already know, this means represent some cultural field as a large set of objects with various features. Each feature captures some characteristic of the objects. The features can use existing metadata (such as dates or names), extracted automatically by a computer, or added manually (in social sciences, this process is called “coding,” in humanities, we call this “annotation” or “tagging”).

The objects can be photographs, songs, novels, paintings, websites, user generated content on social networks, or any other large set of cultural artifacts selected using some criteria. They can be all works of a single creator, if we want to understand how her/his works are related to each other. Instead of the cultural artifacts, the objects in our representation can be also individual cultural consumers and features can represent some characteristics of their cultural activities: for example, web sites visited by a person, a trajectory though a museum and time spent looking at particular artworks, or the positions of faces in selfie photos (see our project http://www.selfiecity.net for the analysis of such data.)

Once we represent some cultural field or cultural activity field as data (objects and their features), we can conceptualize each object as a point in a multi-dimensional feature space. This allows us to use “exploratory data analysis” techniques from data science and also techniques from data visualization field to investigate the “shape” of this feature space.

The space may have different structures: all points may cluster together, or form a few clusters, or lie at approximately equal distances from each other, etc. Any of these patterns will have an appropriate cultural interpretation. If most points form a single cluster, this means that in a particular cultural field most works/activities have similar characteristics, and only a small number are significantly different. Or we can find a few large clusters that lie at sufficient distances from each other (this can be quantified by measuring distances between the centers of the clusters.). And if we find that there are no clusters, this means that a given cultural space has a high degree of variability, and every work is significantly different from the rest.[12]

Note that just as it was the case with van Gogh example, even if we use many different features, we cannot be sure that we have captured the right information to quantify difference as we humans see it. But producing a single "correct" map should not our main goal.. Every selection of features and choice of parameters of the algorithm will create a different map of cultural artifacts we are interested in. And every map can show us something new.

Using modern data analysis and visualization software, we can generate multiple views of the same data quickly and compare them. This helps us to expand our understanding of a cultural phenomenon, and also notice the relations and patterns we did not see before. In other words, data science allows us not only just to see the data that is too big for our unaided perception and cognition; it also allows us to see data of any size (including very familiar canonical cultural datasets) *differently*.

Dimension Reduction

We want to explore the structure of a feature space: the presence, positions and the shapes of clusters, the distances between them, and the average distances between individual points. How to do this? It will be great if we can visualize this space. If we only have two features, we can directly map each of them into one dimension and create a conventional 2-D scatterplot. If we have many features, a space of many dimensions can be represented as a series of separate scatter plots, each plot showing a pair of features. This visualization technique is called a scatterplot matrix.

Scatterplot matrixes become less useful if we have lots of dimensions. Each plot only shows a particular projection of the space onto two dimensions, i.e., a single flat surface. If the shapes of point clusters are truly multi-dimensional,

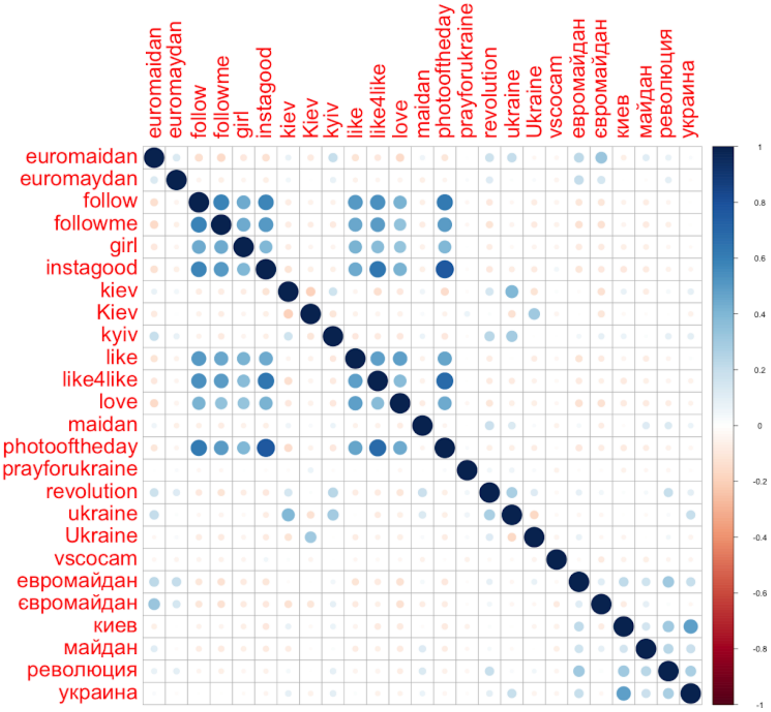


Figure 4: Heat map visualization of top tags assigned by Instagram users to images shared in the center of Kyiv during February 2014 Maidan revolution in Ukraine. Using Instagram API, we collected images for February 17-22. During this period 6,165 Instagram users shared 13,208 images which they tagged with the total of 21,465 tags (5845 unique tags). Visualization shows 25 most frequently used tags. The intensity of color/size indicates how frequently the two tags were used together. (Visualization by Lev Manovich / Software Studies Initiative.)

studying a large number of separate 2-D plots may not help us to see these shapes.

Another method for visualizing points in a space of many dimensions (i.e., many features) is to use a distance matrix. Distance matrix is computed directly from a data table. In a distance matrix, each cell represents a numerical distance between two objects from the original table. By converting the values of the cells into gray tones, colors, or shapes, we can turn the distance matrix into a visualization. Such visualization is called a heat map. Like scatterplot matrixes, heat maps can also quickly become very dense as we add features, and they also have the same limitation of making it hard to see the shapes of multi-dimensional clusters. Fig. 4 is an example of a heatmap visualizations used to explore tags assigned by Instagram users to images they share.

Data science developed another approach for seeing and interpreting the structure of a space of many dimensions. It is called dimension reduction. Along with objects, features, feature space and distances, dimension reduction is another fundamental concept of data science important for humanities.

Dimension reduction is the most widely used approach today for exploring data that has arbitrary larger number of features. It refers to various algorithms that create a low dimension representation of a multi-dimensional space. If this new representation only has two or three dimensions, we can visualize it using one or two standard 2D scatterplots.

Note that typically each axis in such scatterplot(s) no longer corresponds to a single feature. Instead, it represents a combination of various features. This is the serious challenge of dimension reduction algorithms – while they allow us to represent data using scatterplots where we can see the structure of a space easily, it can be quite challenging to interpret the meaning of each dimension. But even if we cannot say exactly what each axis represents, we can still study the shape of the space, the presence or absence of clusters, and the relative distances between points.

Dimension reduction is a projection of a space of many dimensions into a fewer dimensions – in the same way as a shadow of a person is a projection of a body in three dimensions into two dimensions. Depending on the position of the sun, some shadows will be more informative than others. (For example, if the sun is directly above my head, my shadow becomes very short, and my body shape is represented in a very distorted way. But if the sun is at 30 or 45 degree angle, my shadow will contain more information.) Similarly, the idea of dimension reduction is to preserve as much of the original information as possible. But it is crucial to keep in mind that some information will be always lost.

Different dimension reduction techniques use different criteria as to what kind of information should be preserved and how this is to be achieved. The following are among three very widely used data exploration methods that use dimension reduction:

Multi-dimensional scaling (MDS): We want to preserve the relative distances between points in a multi-dimensional space while projecting it into a lower dimension space.

Principal Component Analysis (PCA): We want to preserve most variability (spread of the data) when we go from all to fewer dimensions.

Factor analysis: Similar to MDS and PCA, but its original motivation was different. The idea of factor analysis is to extract “factors” - a smaller number of “hidden” variables that are responsible for the larger set of observed (recorded, measured) variables.[13]

Fig. 5 is an example of MDS visualization. We explore top 25 Instagram’s tags for 13,208 images Kyiv during February 2014 Maidan revolution in Ukraine, and find distinct semantic clusters.

Figure 5: Visualization of the data from Fig. 4 using multi-dimensional Scaling (MDS). The tags that are often used together appear close to each other in the plot. On the right, we see a tight cluster of the tags that represent the “universal” Instagram language: #like, #follow, #instagood, etc. (these same tags are popular in lots of locations around the world). On the left, we see another cluster of tags associated with Maidan revolution. The visualization suggests that there is little interaction between these two types of tags: one group of Instagram users was using generic tags while another group was primarily tagging the local and specific events.

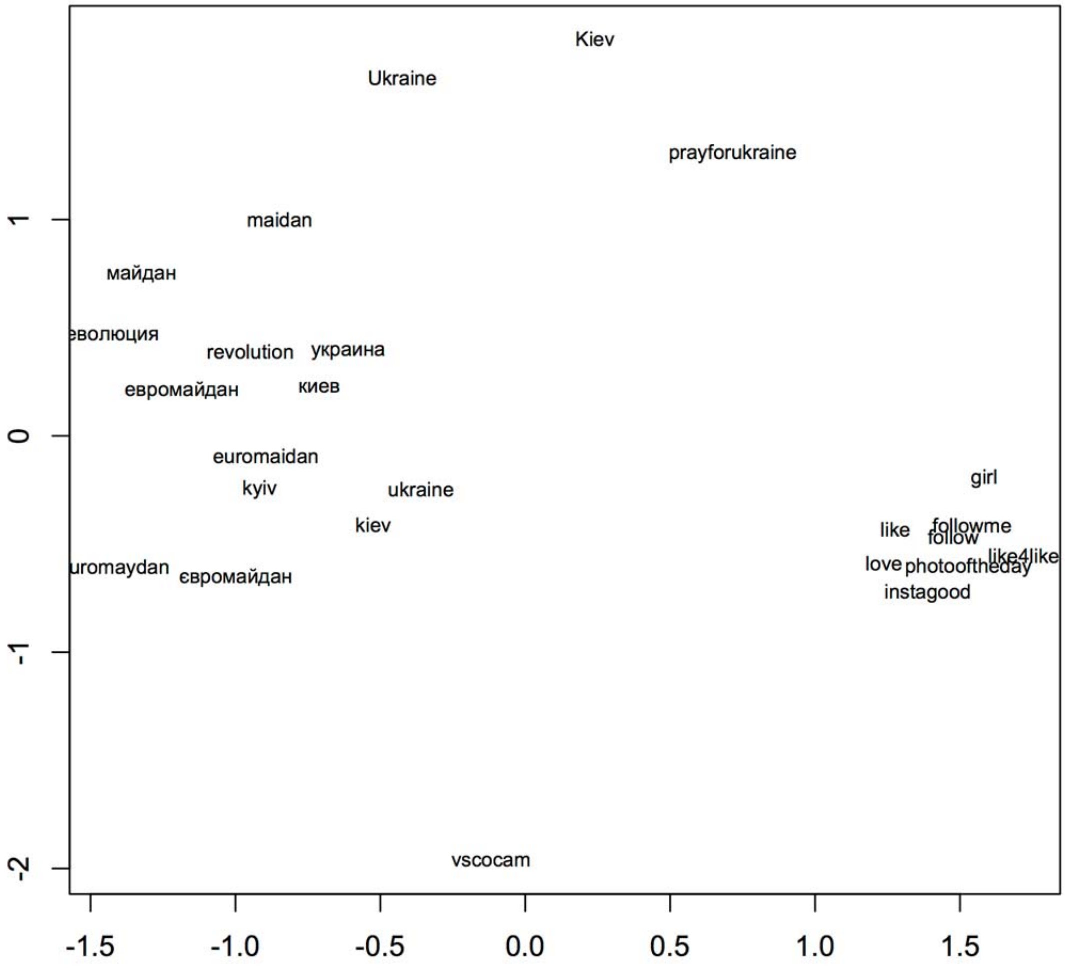
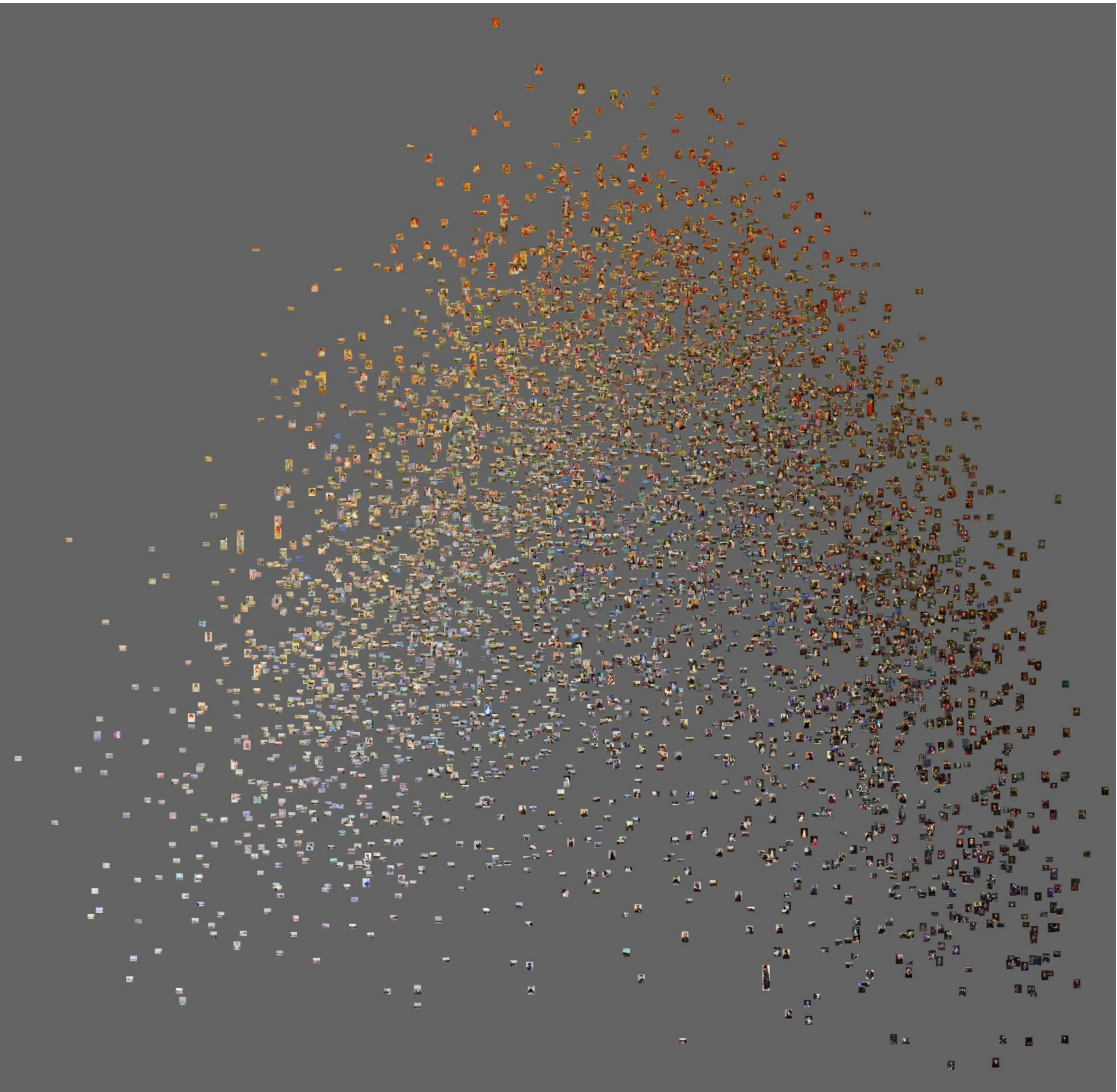


Fig. 6 shows an example of a visualization of approximately 6000 paintings of French Impressionists that uses PCA. In such visualization, images that are similar to each in terms of particular visual features are grouped together. Such visualizations allow us to compare many images to each other, and understand patterns of similarity and difference in large visual datasets.



Conclusion

To explore is to compare. And to compare, we need first to see. To see big cultural data, we need to turn to data science.

Until the 21st century, we typically compared small numbers of artifacts, and the use of our human cognitive capacities unaided by machines was considered to be sufficient. But today, if we want to compare tens of thousands or millions of cultural artifacts (born digital user generated content is a prime example of such scales, but some digitized collections of historical artifacts can be also quite large) we have no choice but to use computational methods. In other words: To "see" contemporary culture requires use of computers and data science.

This computer "vision" can be understood as extension of the most basic act (or method) of humanities - comparing cultural artifacts (or periods, authors, genres, movements, themes, techniques, topics, etc.) So while computer-enabled seeing enabled by data science may be radical in terms of its scale – how much you can see in one "glance," so to speak – it continues the humanities' traditional methodology.

In this article I introduced a number of core concepts of data science: *objects, features, feature space, measuring distance in feature space, dimension reduction.* In my view, they are most basic and fundamental concepts of the field *relevant to humanities.* They enable exploration of large data, but they are also behind other areas of data science and their industry applications. In fact, they are as central to our "big data society" as other main cultural techniques we use to represent and reason about the world and each other – natural languages, lens-based photo and video imaging, material technologies for preserving and accessing information (paper, printing, digital media, etc.), counting, or calculus. They form data society's "mind" – the particular ways of encountering, understanding, and acting on the world and the humans specific to our time.

Figure 6: Example of a visualization of an image collection using Principal Component Analysis. The data set is digital images of approximately 6000 paintings by French Impressionists. We extracted 200 separate features from each image, describing its color characteristics, contrast, shapes, textures and some aspects of composition. We then used Principal Component Analysis to reduce the space of 200 features to a smaller number of dimensions, and visualized the first two dimensions. In a visualization, images that are similar to each in terms of features we extracted are grouped together. One interesting finding is that the types of images popularly associated with Impressionism (lower left part) constitute only a smaller part of the larger set of artworks created by these artists. At least half of the images turn to be rather traditional and more typical of classical 19th century painting (darker tones and warm colors.) Note that our data set contain only approximately ½ of all painting and pastels created by participants in Impressionist exhibitions in 1874-1886. (Visualization by Lev Manovich / Software Studies Initiative).

Notes

[1] Acknowledgments: This article draws on the research by Software Studies Initiative between 2007 and 2015. I would like to thank all graduate and undergraduate students who participated in our projects and our external collaborators. Our work was supported by The Andrew Mellon

Foundation, The National Endowment for the Humanities, The National Science Foundation, National Energy Research Scientific Computing Center (NERSC), The Graduate Center, City University of New York (CUNY), California Institute for Telecommunications and Information Technology (Calit2), University of California Humanities Research Institute, Singapore Ministry of Education, Museum of Modern Art (NYC) and New York Public Library.

[2] Adrian Raftery, "Statistics in Sociology, 1950-2000: A Selective Review." Sociological Methodology 31 (2001): 1-45, https://www.stat.washington.edu/raftery/Research/PDF/socmeth2001.pdf (accessed April 24, 2015).

[3] For example, David Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining* (Cambridge, Mass.: The MIT Press, 2001); Jure Leskovec, Anand Rajaraman, and Jeff Ullman, *Mining of Massive Datasets*. 2n edition (Cambridge: Cambridge University Press, 2014); Nina Zumel and John Mount, *Practical Data Science with R* (Shelter Island: Manning Publications, 2014).

[4] MoMA (Museum of Modern Art), Network diagram of the artists in *Inventing Abstraction, 1910-1925* exhibition (2012), http://www.moma.org/interactives/exhibitions/2012/inventingabstraction/?page=connections (accessed April 24, 2015).

[5] For historical examples, see Michael Friendly and Daniel Denis, "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization" (n.d.), http://datavis.ca/milestones/ (accessed April 24, 2015).

[6] Michael Friendly and Daniel Denis, "The Early Origins and Development of the Scatterplot," *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 103–130, http://www.datavis.ca/papers/friendly-scat.pdf (accessed April 24, 2015).

[7] Quoted in Garabed Eknoyan, "Adolphe Quetelet (1796–1874)—the average man and indices of obesity," *Nephrology Dialysis Transplantation* 23, no. 1 (2008): 47-51, http://ndt.oxfordjournals.org/content/23/1/47.full (accessed April 24, 2015).

[8] Émile Durkheim, *Le Suicide. Étude de Sociologie* (Paris, 1897).

[9] For a highly influential presentation of statistics in this period, see Ronald A. Fisher, *Statistical Methods for Research Workers* (Edinburgh: Oliver and Boyd, 1925), http://psychclassics.yorku.ca/Fisher/Methods/index.htm (accessed April 24, 2015).

[10] For one of the first publications in now what is a big field of computational analysis of large photo datasets, see Ritendra Datta et al., "Studying aesthetics in photographic images using a computational approach," *ECCV'06 Proceedings of the 9th European conference on Computer Vision* Volume Part III (2006): 288-301.

[11] Franko Moretti, "Conjectures on World Literature," *New Left Review* 1, January-February (2000): 55, http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature (accessed April 24, 2015).

[12] See Lev Manovich, "Mondrian vs Rothko: footprints and evolution in style space," 2011, http://lab.softwarestudies.com/2011/06/mondrian-vs-rothko-footprints-and.html (accessed April 24, 2015).

[13] For one of the original formulation of factor analysis in psychology, see Louis Leon Thurstone, "Vectors of Mind," *Psychological Review* 41 (1934): 1-32, http://psychclassics.yorku.ca/Thurstone/ (accessed April 24, 2015).

Bibliography

Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Wang. "Studying aesthetics in photographic images using a computational approach." *ECCV'06 Proceedings of the 9th European conference on Computer Vision* Volume Part III (2006): 288-301.

Durkheim, Émile. *Le Suicide. Étude de Sociologie*. Paris, 1897.

Fisher, Ronald A. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925. http://psychclassics.yorku.ca/Fisher/Methods/index.htm

Friendly, Michael and Daniel Denis. "The Early Origins and Development of the Scatterplot." *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 103–130. http://www.datavis.ca/papers/friendly-scat.pdf

---. “Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization." N.d. http://datavis.ca/milestones/

Eknoyan, Garabed. “Adolphe Quetelet (1796–1874)—the average man and indices of obesity.” *Nephrology Dialysis Transplantation* 23, no. 1 (2008): 47-51. http://ndt.oxfordjournals .org/content/23/1/47.full

Hand, David, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. Cambridge, Mass.: The MIT Press, 2001.

Leskovec, Jure, Anand Rajaraman, and Jeff Ullman. *Mining of Massive Datasets*. 2n edition. Cambridge: Cambridge University Press, 2014. Full book text is available at http://www. mmds.org/.

Manovich, Lev. “Mondrian vs Rothko: footprints and evolution in style space.” 2011. http://lab.softwarestudies.com/2011/06/mondrian-vs-rothko-footprints-and.html

MoMA (Museum of Modern Art). Network diagram of the artists in *Inventing Abstraction, 1910-1925* exhibition. 2012. http://www.moma.org/interactives/exhibitions/2012 /inventingabstraction/?page=connections

Moretti, Franko. “Conjectures on World Literature.” *New Left Review* 1, January-February (2000): 54-68. http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-litera ture

Raftery, Adrian. “Statistics in Sociology, 1950-2000: A Selective Review.” *Sociological Methodology* 31 (2001): 1-45. https://www.stat.washington.edu/raftery/Research/PDF/soc meth2001.pdf

Thurstone, Louis Leon. “Vectors of Mind.” *Psychological Review* 41 (1934): 1-32. (Address of the president before the American Psychological Association, Chicago meeting, September, 1933.) http://psychclassics.yorku.ca/Thurstone/

Zumel, Nina and John Mount. *Practical Data Science with R*. Shelter Island: Manning Publications, 2014.

Lev Manovich is a Professor of Computer Science at The Graduate Center, City University of New York, and founder and director of Software Studies Initiative.

He is the author of seven books including Software Takes Command (Bloomsbury Academic, 2013), Soft Cinema: Navigating the Database (The MIT Press, 2005), and The Language of New Media (The MIT Press, 2001) which was described as "the most suggestive and broad ranging media history since Marshall McLuhan." In 2014 he was included in the list of 50 "most interesting people building the future" (The Verge).

Correspondence e-mail: manovich.lev@gmail.com