Lev Manovich

Data

Representing Phenomena as Data

In the early 21st century, collection and analysis of data using computers has become central to the functioning of societies. The new field called *data science* that includes classical statistic and newer methods to handle "big data" become very popular. Dozens of professional fields have started to employ data scientists to extract value and generate predictions from their data. In the academic world, new disciplines *digital humanities* and *computational social science* that focus on computational analysis of large social or cultural data emerged and grew quickly.

If we want to use computers to analyze some phenomenon or process, how do we start? First, we need to represent this phenomenon or process in such a way that computers can act on this representation. It includes numbers, categories, digitized texts, images, audio, and other media types, records of human activities, spatial locations, and connections between elements (i.e., network relations). Only after that such a representation is constructed, we can use computational methods to analyze it.

Creating such a representation involves making three crucial decisions:

1. What are the boundaries of this phenomenon?

For example, if we are interested in studying "contemporary societies," how can we make this manageable? Or, if we want to study the subject of "modern art," how we will choose what to include—time periods, countries, artists, artworks, publications, exhibitions, or other information? In another example, let's say that we are interested in contemporary "amateur photography." Shall we focus on studying particular photo enthusiast groups on Flickr, or shall we collect large sample of images shared worldwide from Instagram, Facebook, Weibo, VK, and other social networks and media sharing services—since everybody today with a mobile phone with a built-in camera is automatically a photographer?

2. What are the objects we will represent?

For example, if we want to represent the phenomenon of modern art, we may include the following data objects: artists, artworks, correspondence between artists, lists of their exhibitions, reviews in art journals, passages in art books, auction prices, numbers of their followers on social media. The entrance to 2012 *Inventing Abstraction* exhibition at MoMA (New York) featured a large network visualization showing connections between 85 artists in this exhibition based on the number of letters they exchanged.[1] In this representation, modern abstract art was presented by a set of connections between artists, as opposed to other kind of objects I listed above.

In "contemporary society" example, we can, for instance, construct a sample of people chosen at random, their demographic and economic characteristics, their connections to each other, biological daily patterns as recorded by sensors they wear, and their social media (if they give us permission). If we want to understand patterns of work in a hospital, we may use as our data objects both people (doctors, nurses, patients) and also medical procedures performed, tests, forms, doctors' notes, medical images produced, and so on. Data science uses a number of equivalent terms to refer to data objects. These terms come from other fields that were using data much earlier and which data science draws on. They are *data points*, *records*, *items*, *samples*, *measurements*, *independent variables*, *target variables.* This is useful to know if you want to read data analysis publications, learn data skills using online tutorials, or try software.

3. What characteristics of each object we will include?

Characteristics of objects may be also referred as *properties*, *attributes*, *metadata*, or *features*. humanities fields, cultural heritage and library science, people refer to objects' characteristics that are already available in the data (because somebody already recorded them) and additional characteristics we have added via, for example, manual tagging as *metadata*. In social sciences, the process of manually creating descriptions of objects is called *coding*. In data science, researchers use algorithms to automatically *extract* various statistics (i.e., summarized compact descriptions) from the objects. These statistics are referred as *features* and this process is called *feature extraction*.

Although it is logical to think of the three questions above as three stages in the process of creating a structured representation that a computer can analyze—limiting the scope, choosing objects, and choosing their characteristics—it is not necessary to proceed in this 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 original plan and limit analysis to characteristics we do have. In short, the processes of creating a data representation and analyzing this data often proceed in parallel and drive each other.

Depending on our perspective, we could assume that a phenomenon such as, for example, "contemporary society," objectively exists regardless of how we study it—i.e., what we decide to use objects and their characteristics. Alternatively, we can assume that a phenomenon is equal to a set of objects and their properties used in all different qualitative and quantitative studies, publications and communication about it until now (books, articles, popular media, academic papers etc.). That is, a *phenomenon is constituted by its representations and the conversations about it*. This includes the created datasets, the research questions used in studies, and the results of the analysis of these datasets. Given that in the

[1] *Museum of Modern Art*, "Network Diagram of the Artists in *Inventing Abstraction, 1910-1925*" (2012), accessed August 8, 2016, http://www.moma.org/interactives/exhibitions/2012/inventingabstraction/?page=connections.

academy research people typically start with already existing research and either refine it or add new methods and questions, this perspective makes a good sense. So, Facebook phenomenon as it is “defined” in computer science and computational social science is all published research on it until now. 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.” The first perspective maybe called empiricist, while the second is related to Foucault’s concept of *discourse* where statements constitutes the objects of knowledge.

The ideas in Michel Foucault’s *The Archeology of Knowledge* published in 1969[2] are also very relevant for computational analysis of cultural phenomena in general. If statistics and quantitative social science calls for us to seek unity and continuity in the data, Foucault’s discourse concept allows for a different perspective where our collected data, i.e., *statements* in Foucault’s terms, may contain contradictions, multiple positions, and represent not a coherent system but system in transition. Thus, if we find correlations or patterns that describe only part of the data, this does not mean that our method is weak. Instead, it is normal that an institution, or social or cultural process generates a large body of statements that may follow different logics and not correspond to each other. Also relevant is another of Foucault’s ideas: that we should analyze discourse on the level of “things said,” as an *archive* of statements that are related to each other rather than to something outside. For me, large samples of user-generated content are such archives. So rather than always asking how user-generated content (e.g., Instagram images shared by a group of people in a given area, their tags and descriptions) does or does not reflect the urban, social, economic, and demographic dimensions outside, thus treating it as signs, it is equally productive to instead consider this content as its own universe of visual subjects, styles, texts and network relations.

Data = Objects + Features

Together, *a set of objects and their features constitutes the “data” (or “dataset”)* that we can work with using computers.

Most data representations include some aspects of the phenomena and exclude others. So, they are “biased.” And this is not a new development. Any two-dimensional map, for example, represents some characteristics of a physical territory but does not show others. But a map does not need to show everything. A map is not a painting, a photograph or a 3D model—it is diagram that presents only the information we need to have and omits the rest. (While the information shown was fixed in printed maps, in interactive contemporary maps we can select what layers and details to show, to search for places, see accidents, get navigation instructions—so their utility as instruments is greatly increased, although visually they may use the same conventions as older paper maps.)

In the case of quantitative studies that use data, their limitations can often be easily corrected. For example, let’s say that we did a survey of social media usage in a particular area

[2] Michel Foucault, *The Archaeology of Knowledge*, trans. A. M. Sheridan Smith (London and New York: Routledge, 2002). Original work published in 1969.

by asking a random sample of people a series of questions. (Pew Research center regularly conducts such surveys in the U.S.) We can enlarge our data by carrying out more surveys in other areas. We can also do a new survey and ask additional questions, and so on.

But the concept of data also depends on a basic and fundamental condition that cannot be changed. Understanding this condition is really important. Before we can use a computer to analyze a phenomenon, behavior, or activity, they have to be represented as *a finite set of individual data objects that have finite number of features*. For example, computational analysis of music typically starts with dividing music track into very small intervals such as 100ms and measures sound characteristics of each sample. To use our previous examples of cultural data, names of artists and their works, passages in art books, or people in a survey are all examples of individual data objects.

How is a *data representation* of some phenomenon or process different from other kinds of cultural representations humans used until now, be they representational paintings, literary narratives, historical accounts, or hand-drawn maps? First, a data representation is *modular*, i.e., it consists of 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, spatial coordinates, time unites, etc.—but not just any form. And only one format can be used for each feature.

In other words, today "data" is not just any arbitrary collection of items existing in some medium such as paper. In a computational environment, *"data" is something a computer can read, transform, and analyze*. This imposes fundamental constraints on how we represent anything.

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 them *computable, manageable, knowable* and *shareable* though data science techniques.

Practically, objects and features can be organized in various ways, but the single most common format is a *table*. An Excel or Google 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). Typically, each row represents one object, and each column represents one feature. A set of objects with their features stored in a table-like format is *the most frequently used representation of data today, used in every professional field, all natural and social sciences (and now entering humanities), NGOs, and governments*. It is the way *data society understands phenomena and individuals, and acts on them*.

In summary, while human societies have used data-like representations for thousands of years, the adoption of digital computers have imposed a number of constraints on what counts as *data* (or *datasets*) today. Datasets are not just any collections of some information, they are objects structured in ways that allows them to exist within a computational medium.