Defining Data: Questions of "Numerical Facts"

Malinda Dietrich
malinda.dietrich@colorado.edu
University of Colorado Boulder
Boulder, Colorado

Morgan Scheuerman
morgan.klaus@colorado.edu
University of Colorado Boulder
Boulder, Colorado

Katy Weathington
katy.weathington@colorado.edu
University of Colorado Boulder
Boulder, Colorado

ABSTRACT

The terms "data" and "information" have different meanings depending on who is asked. The first author, in an effort to unpack these different meanings in relation to ontological and epistemological understandings of the world and how knowledge is produced, interviewed individuals from both professions that interact with "data" and individuals from the general population. We unpack some of the preliminary findings prior to proposing an intervention that suggests providing more clarity and background in the process of using and displaying "data".

INTRODUCTION

On July 28th, Jonathan Swan of AXIOS interviewed U.S. President Donald Trump [2]. During the interview, Swan asked the President questions about COVID-19 data. Some of the data were comparisons between the United States and other countries, including how many people were tested for COVID-19, how many people tested positive for COVID-19, and how much "control" the country has over the pandemic. At one point, the President showed Swan some paper "charts" that resulted in the two disagreeing on the statistics. As the interview continues, President Trump expressed multiple times that Swan was misinterpreting the information. This vignette demonstrates how present "data" is in our daily vocabulary, as well as society's slippage with the word "data".

The researchers have begun a qualitative study that seeks to better understand how different groups of people conceptualize and understand the term "data". Many people make the claim that data are all around us–companies are motivated by "big data" and use it to drive future decisions, research

puts emphasis on what the data shows or how the data can be modeled to make assumptions about the world, and our technologies increasingly show us ads tailored to our wants and needs because this data is accurately predicting our behavior. Despite “data” being commonplace in everyone’s everyday lives, what does the word mean?

THE PROBLEM WITH “DATA”

The word "data" was used as early as the 1640s to mean “a fact”. The first author, in her own words, would define data as something that falls underneath “information”; data is the recording of something (whether that be numerical or not), and once that record is organized, it becomes information. While data and information are seemingly interchangeable, the researcher sees information as an organization of the data into a narrative or further analysis. However, if one looks in the Merriam Webster dictionary, the word “data” is defined three ways: first, as “factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation”; second, as “information in digital form that can be transmitted or processed”; and third, “information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful” [1]. The first definition highlights a positivist understanding of recording measurements as objective and factual. The second definition, which is rather vague and unclear, exemplifies digitized information into a database; truthfully, this definition might be gesturing towards electric pulses or binary that we consider to be “digital” as well. Finally, the third definition begins to ascertain how data points are collected and sifted through. While this word was used as early as the 1640s to mean “a fact,” “data” can mean something different to people within different disciplines. Clearly, there is no monolithic concept of data; rather, there is a wide range of understanding.

Prior work in social science and humanities disciplines has explored areas of how data is used and transited. In communication and media studies, scholars have identified and examined the ways in which technologies allowed for the inscription and storage of “data” [7]. Others have created books that focus on “keywords” (following the work of Cultural Studies scholar, Raymond Williams) of communication technology and digital media research; these "keywords" draw upon the word “data” but never include "data" as an entry with its own definition [5, 9]. Others have studied the culture surrounding the commodification of “data” [3, 6]. Some scholarship has focused on how we “became” our data through a historical approach towards understanding the institutional practices that normalized the collection of “data” [8] and how we are defined by “what our data is made to say about us” [4]. However, throughout all these works, the term “data” remains undefined; it is implied that the reader already understands exactly what “data” is.

Computer scientists Ward and Barker surveyed different companies and reports to determine how different companies were conceptualizing the term “Big Data” [10]. The authors found that one of

the most cited definitions is one from a Gartner report in 2001 that highlights Volume, Velocity, and Variety as primary characteristics of data; they state that “the evidence presented in the Gartner definition is entirely anecdotal. No numerical quantification of big data is afforded.”

The lack of a singular, colloquial definition of “data” often leads to issues of misinterpretation of data–we can look to the example above, or some of the many Universities that have implemented “COVID-19 Dashboards” to be more forthcoming with some of the data they have collected that ultimately get critiqued due to causing confusion and frustration with many of the faculty and students.

THE STUDY

The first author and her collaborators are conducting a study focusing on how different data professionals and individuals from the general public defined both “data” and “information”. The study seeks to aid scholarship in the understanding of how data is conceptualized and used to impact people daily. As questions around data become more central to aspects of our lived experience (e.g., politics, public health, weather), the results of this research will hold important implications for designing technologies and platforms, and for how the collection of data can be communicated more transparently.

We have interviewed 12 participants from different data professions: four from finance (consultants, corporate strategists, and financial analysts); two climate scientists (researchers of hurricanes and the polar ice caps); two from digital analytics; one in data visualization; one from geographic information systems (or GIS); one from broadcast television services; and one from higher education grant writing. We continue to interview more participants, but none of the participants have been from the general population.

During the interviews, the first author asked the participants a two different sets of questions. Beginning with demographic questions, the researcher then moved to asking general definition questions. She classified these as questions trying to investigate how individuals conceptualize “data” and “information”. Then, the researcher posed some data professional specific questions. This includes questions such as: what does working with data (and/or information) usually look like? What counts as working with data? Who is equipped to work with “data”? And who does this work benefit?

The first author is deciding to use interview methods for this project in order to unpack how different individuals are using the term data. By employing the use of semi-structured interviews, the researcher is and has been able to "go deeper" into particular topics of interest or themes that arise such as gathering more information on the "data sets" utilized by these different professions. "Going deeper" in this sense, allows for more contextual material and information that may lead to better understanding of the participant’s ontological and epistemological orientations, which in turn, help us understand where distinctions or discrepancies in shared knowledge arise.

PRELIMINARY FINDINGS

For the preliminary understanding of the study, we will speak through our findings in relation to some of the definitions used above. While above the researchers frame the discussion around one dictionary definition of data, Merriam Webster is only a single source. In future work, it is important that the researchers moved beyond this definition as a single frame of reference.

Some of the participants defined data and information similarly to the researcher: one (data) as underneath the other (information), and information being the “story of the data”. The participants who held these definitions were in Geographic Information Systems, Hurricane research, and Finance. Some participants made a different distinction than the researcher between data and information: they felt that data was purely quantitative, while information was purely qualitative. Other participants spoke of data and information as interchangeable, objective truths. For instance, one of the climate scientists spoke of measuring “factual information” such as the temperature or the depth of polar ice caps and called this measurements both data and information.

A few other participants spoke on how their interactions with data required that it be “processed to be meaningful.” Participant P04 noted that “data is information that we collect from our surroundings, that we use to inform decision making... just things that we pick up from our surroundings that can aid us in decision making.” Many other participants felt that context matters. In other words, to understand the data, one must understand the context around that set of data. Finally, some participants discussed human elements of “data” and data science more generally: some spoke on who collects the data (if not them); who cleans and organizes data; and how it makes them feel (oftentimes, lonesome) to be the primary person interacting with the “raw” data or information.

POTENTIAL INTERVENTIONS

To move forward with this research data, the researcher proposes an intervention. One participant suggested that those working with data provide more context around their data points. This participant suggested to “show why you did what you did” through the documentation of how the person found the data, collected it, interested it, and why they chose those methods (similar to how researchers document our progress). Acknowledging that this work is done in some formats (like End User License Agreements) these legal documents and processes are abstracted. In other words, the researcher poses design questions around clarity: how can these processes be documented and made clear?

Ultimately, this research and work is ongoing. Further questions and interventions are being explored. To conclude, the researchers propose the questions in which they are exploring: what are data? How do different people define or conceptualize this term? Ethically, how can these meanings be better communicated to the public?

REFERENCES

[1] [n.d.]. Data. https://www.merriam-webster.com/dictionary/data

[2] 2020. President Trump Exclusive Interview. https://www.youtube.com/watch?v=zaaTZkqsaxY&feature=emb_title

[3] Louise Amoore. 2013. *The politics of possibility: Risk and security beyond probability.* Duke University Press.

[4] John Cheney-Lippold. 2017. *We Are Data: Algorithms and the Making of Our Digital Selves.* NYU Press.

[5] Matthew Fuller. 2008. *Software Studies A Lexicon.* MIT Press.

[6] Rob Kitchin. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage Publishing.

[7] Friedrich Kittler. 1990. *Discourse Networks 1800/1900.* Stanford University Press.

[8] Colin Koopman. 2019. *How we became our data: A genealogy of the informational person.* University of Chicago Press.

[9] Benjamin Peters. 2016. *Digital Keywords: A Vocabulary of Information Society and Culture.* Princeton University Press.

[10] Jonathan Ward and Adam Barker. 2013. Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821* (Sept. 2013).