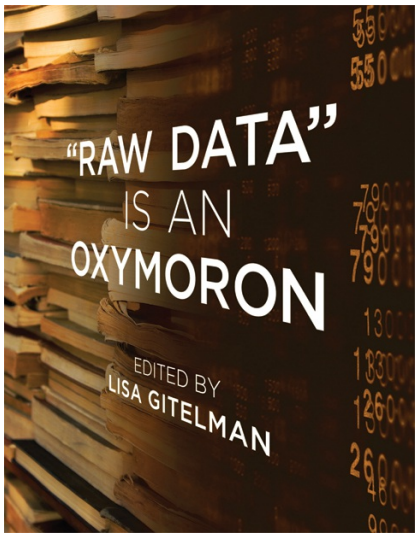
~~Book~~ Review: "Raw Data" is an Oxymoron

Blog Admin

We live in the era of Big Data, with storage and transmission capacity measured not just in terabytes but in petabytes. Data collection is constant and even insidious, with every click and every "like" stored somewhere for something. This edited collection seeks to remind us that data is anything but "raw", that we shouldn't think of data as a natural resource but as a cultural one that needs to be generated, protected, and interpreted. ***Niccolò Tempini*** *finds that all of the matters discussed in this book are as inherently political as they are urgent.*

"Raw Data" is an Oxymoron. Lisa Gitelman (ed.). MIT Press. March 2013.



Find this book:

amazon

Edited by NYU media historian Lisa Gitelman, this book is a stimulating and useful collection of essays which characterize practices at the heart of the increasingly data-centric society in which we live. It provides us with conceptual tools with which to critically engage with emerging topics like, for instance, Big Data.

Newspapers and the business world are abuzz about the data revolution. In these times, when writers sometimes put too little commitment into weighing competing claims, there is need for academic voices to step into the debate. Academic work reminds us that much research about data and its organizational and social implications has been produced. While this book is not heavy on 'back-linking', it provides some well-read and original empirical research; a breath of oxygen through the hot air we have been getting from preachers of all sorts.

The elegant title of this collection states its main argument well: there is no such thing as 'raw data'. In her introduction to the volume, Gitelman points to the interpretive essence of data: "data are imagined and enunciated against the seamlessness of phenomena". Since data are always imagined, they embed assumptions. Through the book, we learn how the material context and format of producing data sets anticipates and affects their interpretive possibilities.

So what are the lessons that the volume teaches us?

The first lesson is that data have an essentially social, and not epistemological, function. Daniel Rosenberg tracks the etymological history of the term 'data' through the semantic oscillations that characterize the concept's history. He shows us that the original essence of data is to be rhetorical. Data, plural of the Latin *datum,* means *given*. Data is both what a speaker feeds into a conversation, and what should be taken-for-granted, in order to persuade: "Facts are ontological, evidence is epistemological, data is rhetorical". Over time, the connotation of the term changed, "from being reflexively associated with those things that are outside of any possible process of discovery to being the very paradigm of what one seeks through experiment and observation".

As a rhetorical tool, we would expect data to be instrumental in shifting the focus from one to another aspect of a phenomenon. Indeed, Travis D. Williams reminds us that "a data set is already interpreted by the fact that it is a set: some elements are privileged by inclusion, while others are denied relevance through exclusion". Therefore, when trying to see the world by looking though data, it is essential to approach them systemically and immerse ourselves in the original context of their practice and consumption.

The second lesson offered is that data aggregation, the fundamental operation that renders data's value, has implications. Numbers and figures appearing in social analyses are the result of historical processes that shaped according to the specific objectives of their original use. Kevin R. Brine and Mary Poovey, illuminate this issue by reconstructing a process of research in the history economics. They show that assumptions are embedded in a data model upon its creation. Data sources are shaped through 'washing', integration, and algorithmic calculations in order to be commensurate to an acceptable level that allows a data set to be created. Only after these operations have occurred can a theory can fit the data. By the time the data are ready to be used, they are already 'at several degrees of remove from the world.'

Maintaining the stability of a data set over time is a painstaking effort. David Ribes and Steven J. Jackson study scientists involved in a project about ecological change. In this case, scientists study change by taking samples from a river (think Heraclitus). They engage with the ontological problem of defining change in data collection: in the tools, in the data models, in the object of study. What does it mean to measure the same thing, and until what point do scientists regard the changing data as describing degrees of difference rather than something qualitatively different? It is difficult to imagine perhaps, but maintaining backwards-compatible datasets involves precisely a 'complex ontological choreography as scientists and technicians work to make data "the same" in a changing ecology of technologies, organizations, field sites, and institutional arrangements.'

Matthew Stanley analyzes a current problem in astronomy to show that when context cannot be disambiguated evaluating data is particularly difficult. Through the compelling case of the construction of anti-slavery arguments in the USA of 19th century, Ellen Gruber Garvey adds that when data can be aggregated, it is also as a result of a long process of de-contextualization and re-contextualization. This involves the imaginative selection of the data, of 'alienable bits' from a bigger amount of material, and their elaboration in an aggregated format that will make something evident that was not being recognized before – such as the torture and inhuman nature of enslavement.

All of the matters discussed in this book are as inherently political as they are urgent. As data are "reworked, processed through an online algorithm or spat out to somewhere and somewhen to the computer screen of a vigilant operator", warns Geoffrey Bowker in an illuminating afterward, "my possibilities for action are being shaped". More importantly, the social shift over data-centric operations and techniques of government implies the disappearance of those aspects of the world that are less measurable. "Computers may have the data, but not everything in the world is given".

One of the merits of this book is to show us, through well-seasoned empirical data, just how much we can learn from the past. Of all the revolutionary promises coming from Big Data, much of its potential could be lost in the wind if businesses and organizations fail to notice that only a relatively small subset of problems can be resolved by flooding social settings with billions of sensors and a bunch of stats geeks.

Many of the biggest problems facing data-centric problem solving are philosophical in nature. Back in the 1950s Hanson argued that observation was always theory-laden. If we arrange technologies to do much more observation on our behalf, we will need to be especially aware of which theory informing what the data will be rhetorically supporting.

Niccolò Tempini is a PhD Candidate in Information Systems at the London School of Economics and Political Science. You can follow Niccolò on Twitter @tmpncl. Read more reviews by Niccolò.