THE NEW ATLANTIS
A JOURNAL OF TECHNOLOGY & SOCIETY

Why Data Is Never Raw

Nick Barrowman

A curious fact about our data-obsessed era is that we're often not entirely sure what we even mean by "data": Elementary particles of knowledge? Digital records? Pure information? Sometimes when we refer to "the data," we mean the results of an analysis or the evidence concerning a certain question. On other occasions we intend "data" to signify something like "reliable evidence," as in the saying "The plural of anecdote is not data."

In everyday usage, the term "data" is associated with a jumble of notions about information, science, and knowledge. Countless reports marvel at the astonishing volumes of data being produced and manipulated, the efficiencies and new opportunities this has made possible, and the myriad ways in which society is changing as a result. We speak of "raw" data and laud it for its independence from human judgment. On this basis, "data-driven" (or "evidence-based") decision-making is widely endorsed. Yet data's purported freedom from human subjectivity also seems to allow us to invest it with agency: "Let the data speak for itself," for "The data doesn't lie."

Out of this quizzical mix, it is perhaps unsurprising that near-magical thinking about data has emerged. In the 2015 book *Digital Destiny: How the New Age of Data Will Transform the Way We Work, Live, and Communicate*, Shawn DuBravac describes a collection of "properties of data" and expresses them in anthropomorphic terms. DuBravac, former chief economist at the Consumer Electronics Association and a self-styled futurist and "trendcaster," claims that data "seeks permanence," "wants to replicate," "seeks instantaneity," "wants to be understood," and "seeks movement."

> Data is immediate.... When data comes into being, when it is first tracked, captured, or copied, it wants to immediately be utilized—to exert force and influence.... Data constantly moves toward efficiency. It removes barriers; it closes distances; it destroys the moments between recognition and understanding. Because data wants to be understood, it abhors friction.

Nick Barrowman *is Senior Statistician at the Children's Hospital of Eastern Ontario Research Institute in Ottawa, Canada. This essay reflects his own views.*

This projection of human-like qualities onto data is ostensibly metaphorical, but it can muddle our thinking. It seems aimed at obscuring how intertwined is the production of data with human judgment, and the use of data with human agency. And once our agency has been obscured, it is not hard to imagine that data has a mind of its own, that to solve our great problems we have only to collect the data and set the computers running.

Provenance

The word *data* is derived from the Latin meaning "given." Rob Kitchin, a social scientist in Ireland and the author of *The Data Revolution* (2014), has argued that instead of considering data as *given* it would be more appropriate to think of it as *taken*, for which the Latin would be *capta.* Except in divine revelation, data is never simply given, nor should it be accepted on faith. How data are construed, recorded, and collected is the result of human decisions—decisions about what exactly to measure, when and where to do so, and by what methods. Inevitably, what gets measured and recorded has an impact on the conclusions that are drawn.

For example, rates of domestic violence were historically underestimated because these crimes were rarely documented. Polling data may miss people who are homeless or institutionalized, and if marginalized people are incompletely represented by opinion polls, the results may be skewed. Data sets often preferentially include people who are more easily reached or more likely to respond.

In scientific research, the choice of what to measure and how is fundamental. But in many cases, especially in the social sciences, what we want to capture doesn't already have a clear measurement. It must therefore be "operationalized" somehow—meaning we must create a technique for measuring it. This necessarily requires emphasizing some aspects over others. Just as thought involves focusing, data collection involves narrowing attention; something is always left out.

Suppose that a database contains the annual income of every household in an underdeveloped country, along with data on a variety of other variables, like the locations of the households, the age and sex of the members, their occupations, and so forth, all collected in a door-to-door survey. Analysis of the data could begin right away, without any additional information.

But we might ask several questions. For what purposes was the data collected? For example, will taxes be based on the income reported by a household? Would the households that were surveyed have any interest in misrepresenting their income? Would households whose members participated in the informal economy, through bartering or working for under-the-table wages, be able to report the exact amounts of their income? Who carried out the door-to-door survey, what were their incentives, and how might their characteristics, questioning techniques, and the political situation in the country have affected responses? Was the data collected on paper and later entered into a computer? If so, who performed the data entry, and under what conditions? Was the data reviewed, cleaned, filtered, or otherwise altered after being entered?

The way data is collected places limits on the inferences we can obtain. If we had reason to believe that the households surveyed were likely to misreport their incomes because they understood the purpose of the survey to be related to taxation, then the data's accuracy would be disputable. If poorer households were likelier to participate in the barter economy, then the data on their incomes might not represent their economic situation as accurately as the data on incomes reported by middle-class households.

All Data Is Cooked

We tend to think of data as the raw material of evidence. Just as many substances, like sugar or oil, are transformed from a raw state to a processed state, data is subjected to a series of transformations before it can be put to use. Thus a distinction is sometimes made between "raw" data and processed data, with "raw data" often seen as a kind of ground truth, a "just the facts, ma'am" empirical starting point. For example, Lea Ypi, in *Global Justice and Avant-Garde Political Agency* (2012), critiques a common view of how argument proceeds in political theory, the idea that there is "a first, pre-interpretive, stage" at which "we are concerned with the identification of the raw data of interpretation." According to this naive view, rival political theorists all begin with the same "basic, uncontroversial, sociological facts in need of critical scrutiny and interpretation."

Although the idea of raw data is useful, it can also be misleading, because, as we've seen, even the initial collection of data already involves intentions, assumptions, and choices that amount to a kind of pre-processing. A related issue is that of how to handle seemingly bad data points. Digital instruments may identify certain measurements as faulty and automatically discard them (using error detection circuits and algorithms), or they may attempt to counter measurement error by combining multiple measurements (using averaging or other methods).

When people use the term "raw data," they usually mean that *for their purposes* the data provides a starting point for drawing conclusions. Suppose a sound level meter is used to measure how loud it is in a particular part of a factory. At the simplest level, the device registers changes in air pressure caused by sound waves. Successive sound pressure readings are then combined with measurements of the sound pressure waves' frequencies, waveforms, and durations, producing an estimate of *loudness*—a technical term in the sound-measurement field that describes sound intensity, adjusting for the differing hearing sensitivity of the human ear to different frequencies. Readings judged invalid—whether because they contain sharp background noise, such as a crash or a honking car horn, or because they don't register any sound at all—can be flagged as invalid or filtered out. In addition, sound meters generally perform a kind of averaging over some time interval to obtain an "equivalent continuous sound level" for that time period.

Suppose further that these values are then downloaded to a computer, where the operator treats them as "raw data." Examination of the sound levels may reveal to the operator that it is particularly loud at certain times of the day. The factory manager, upon learning this, may request that future testing be performed, in order to gather more "raw data." In this example, at least three different levels of "raw data" could be identified: the sound pressure readings, the instantaneous measures of loudness, and the time-averaged sound levels. Additionally, the manager might see the summaries provided by the operator as "raw data" for making decisions about sound-dampening measures.

In the memorable words of Geoffrey Bowker, informatics professor at the University of California, Irvine, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care." "Raw" carries a sense of natural or untouched, while "cooked" suggests the result of cognitive processes. But data is always the product of cognitive, cultural, and institutional processes that determine what to collect and how to collect it. In this sense, "raw data" is indeed a contradiction in terms. In the ordinary use of the term "raw data," "raw" signifies that no processing was performed *following* data collection, but the term obscures the various forms of processing that necessarily occur before data collection.

Scientific measurement is often seen as the epitome of rigorous data collection, and data inherits some of the prestige of science. But scientific measurements are only interpretable in terms of pre-existing theory, and that theory is baked into the scientific instruments used to make those measurements. The history of science shows how theory and appropriate instruments had to be developed before quantities such as force, pressure, and voltage—which we could today easily regard as capturable in "raw" data—could be measured. The French physicist and philosopher Pierre Duhem provided a concrete illustration in *The Aim and Structure of Physical Theory* (1914):

> Go into the laboratory; draw near this table crowded with so much apparatus: an electric battery, copper wire wrapped in silk, vessels filled with mercury, coils, a small iron bar carrying a mirror. An observer plunges the metallic stem of a rod, mounted with rubber, into small holes; the iron oscillates and, by means of the mirror tied to it, sends a beam of light over to a celluloid ruler, and the observer follows the movement of the light beam on it. There, no doubt, you have an experiment; by means of the vibration of this spot of light, this physicist minutely observes the oscillations of the piece of iron. Ask him now what he is doing. Is he going to answer: "I am studying the oscillations of the piece of iron carrying this mirror?" No, he will tell you that he is measuring the electrical resistance of a coil. If you are astonished and ask him what meaning these words have, and what relation they have to the phenomena he has perceived and which you have at the same time perceived, he will reply that your question would require some very long explanations, and he will recommend that you take a course in electricity.

Philosopher Kazem Sadegh-Zadeh gives a modern-day example in his *Handbook of Analytic Philosophy of Medicine* (2012):

> The data are not simply collected or inscribed as if they were something pre-existent. Rather, their production is in fact *data engineering*. Consider, for example, the tremendous amount of advanced mathematics, physics, and computational technology contained in a data recording device such as [a] nuclear magnetic resonance spectrometer or [the] Large Hadron Collider....

The very production of data is thus always relevant to its interpretation. As Lisa Gitelman and Virginia Jackson point out in the introduction to the aptly titled *"Raw Data" Is an Oxymoron* (2013), "At a certain level the collection and management of data may be said to presuppose interpretation." "Raw data" is not merely a practical impossibility, owing to the reality of pre-processing; rather, it is a conceptual impossibility, for data collection itself already is a form of processing.

Simply put, the *context* of data—why it was collected, how it was collected, and how it was transformed—is always relevant. There is, then, no such thing as context-free data, and thus data cannot manifest the kind of perfect objectivity that is sometimes imagined.

The End of Politics

The contemporary fascination with data, born of stunning technological developments, has fostered some dubious beliefs. We are tempted to suppose that data is self-contained and context-independent, and that with sufficient data, concerns about causation, bias, selection, and incompleteness can be disregarded. It is a seductive vision: Raw data, uncorrupted by theory or ideology, will lead us to the truth; complex problems will be solved simply by throwing enough data at them. No experts will be required, apart from those needed to produce the data and herald their findings; no theory, values, or preferences will be relevant; nor will it be necessary to scrutinize any assumptions.

When data is thought of as a generic commodity, its most salient characteristic becomes its quantity. In a 2008 article, Chris Anderson, the former editor-in-chief of *Wired* magazine, proclaims the dawn of "The Petabyte Age," and talks about data as if it were grain being dumped into a mill: "We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." Conceiving of data in this way facilitates grand narratives—and seductive sales pitches—but it is possible only in a Neverland of context-free data.

Readers versed in the jargon of organizational theory may be familiar with the acronym DIKW, standing for data, information, knowledge, and wisdom. The origins of this hierarchical model of insight are murky, but since the 1980s it has become ubiquitous in discussions of data processing. (Some versions of the idea include other concepts, like understanding.) At the bottom of the hierarchy is data, seen as raw or unprocessed, and, on its own, effectively useless. Information is taken to be data that has been put to use in answering straightforward questions about who, what, when, where, and how many. Information in turn provides the raw material for knowledge, which, finally, if reflected upon and considered in a broader context may lead to wisdom, which is concerned not only with accuracy and efficiency but also with values.

This framework, although vague, has a certain plausibility. But in recent years data has come to be seen less as the inherently useless raw material for humans to process and refine than as a source of power and insight in its own right. Confidence seems to have shifted away from the products of human reasoning toward the potency of pure, pre-ideological, pre-theoretical, raw data. The DIKW hierarchy has, to an extent, been turned upside down.

But experts and non-experts alike encounter the world not merely through sense experiences but through frameworks of understanding shaped by numerous factors including theories, values, norms, and faith. It may be tempting to suppose that this context can simply be set aside. But assumptions, when not made explicit, do not simply cease to exist. Rather, they influence the direction of inquiry, the choice of measurements, the way the data is collected and transformed, and the conclusions that are drawn.

Assumptions inevitably find their way into the data and color the conclusions drawn from it. Moreover, they reflect the beliefs of those who collect the data. As economist Ronald Coase famously remarked, "If you torture the data enough, nature will always confess." And journalist Lena Groeger, in a 2017 ProPublica story on the biases that visual designers inscribe into their work, soundly noted that "data doesn't speak for itself—it echoes its collectors."

In order for decisions to be critically evaluated, their supporting values and assumptions must also be scrutinized. This is especially vital when it comes to political decision-making. Data or "data-based" policy is increasingly seen as a panacea when it comes to political decision-making. On this view, we now have an opportunity to replace the messiness of politics with the rational order of data. But the attention on data often only obscures the underlying values and assumptions—and the importance of exposing and subjecting them to critical scrutiny only grows. When values, preferences, and interests clash, politics is not only inevitable but essential. No algorithm can determine which decision is best; any such conclusion merely raises the question: By whose values is it the "best"?

According to technocratic fantasies, politics will be made irrelevant by mountains of incontrovertible data. Those who refuse to accept "what the data is telling us" must be either malevolent or stupid. The only apparently rational option for ordinary citizens is to assent to "the data" and the decisions of technocratic elites. Under such an ideology, it should come as no surprise if the citizenry becomes disengaged, suspicious of data-talk to the point that the very idea of data is discredited.