WHENCE "DATA"?

The surprising origins of a ubiquitous term

by Daniel Rosenberg

OPEN THE NEWSPAPER on any given day and you are likely to find one or more stories about the importance of data in our everyday lives. These stories are no longer clustered mainly in the business or science section, as they were just a few years ago, but also in the sports, entertainment, and fashion pages, and very often in the headlines themselves.

If the press is to be believed, Germany won the last football World Cup because of data, and Barack Obama the last two US presidential elections. The losers in these contests were data aficionados, too, of course. Sport is now governed by the statistical rules of "moneyball," and politicians are "data guys"—to use the phrase favored by Obama's last electoral opponent. "Data" has acquired a kind of aura, as if it unlocked a realm beyond opinion, beyond partisanship, beyond theory.

Claims about the ubiquity of data in our environment may be more or less accurate, but even as claims they represent something powerful: the *idea* of data—"data-ism" even—has become central to contemporary culture, to our understanding of the world, and ourselves.

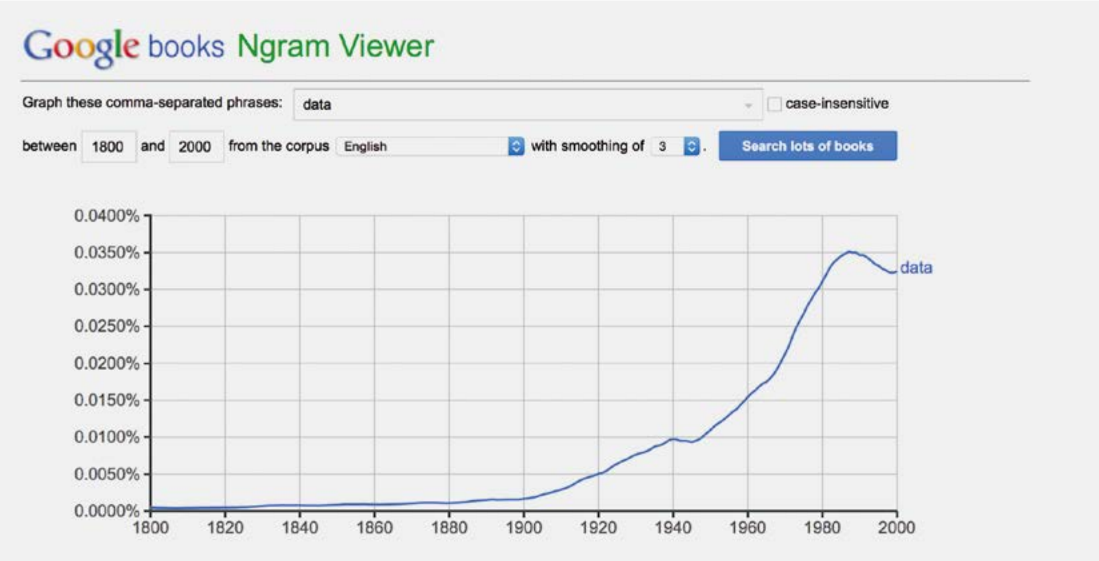
Neither the idea of data nor the technical practices that support it are altogether new. In one way or another, we have inhabited data cultures since the first tax rolls were inscribed and populations counted. And even as a subject of explicit discussion, the term "data" has been around for some time. In English, we've been talking about "data" for more than three centuries now. And, in important ways, the history of the term is a history of modernity itself.

TRAVEL BACK IN TIME to the 1640s, and people are already talking about "data," not in the arts and letters section of the local shipping news, granted, but in a number of specific and important contexts. In some ways, this is not surprising: the seventeenth-century world was steeped in many kinds of data immediately recognizable as such today, from demographer John Graunt's mortality tables to the gold-clasped accounts book that Louis XIV kept in his pocket to the "weather clock" designed by Christopher Wren—the architect who rebuilt London after the Great Fire of 1666—for recording temperature and barometric pressure in real time. Yet the "data" being discussed at the time was distinct from any of these things, and, in general, on the subject of "data," it wasn't a Graunt or a Wren who was doing the talking.

How do we know? These days, there are plenty of new data tools for doing the research. Google, for example, offers an online device called the Google Books Ngram Viewer to chart the frequency of words and phrases by year in the books included in its database. With only a few keystrokes, an everyday user can perform quantitative analysis on a corpus of over five million books, a feat impossible for a scholar with the best resources in the world only a few years ago.

For the term "data," the Google Ngram Viewer (see *Image 1*) produces a very intuitive graph, a curve that creeps along close to zero, begins to pick up in the nineteenth century, and rockets skyward in the middle of the twentieth. At first blush, this seems right, even obvious. In the increasingly mechanized and bureaucratized world of the nineteenth century, data gathering and analysis mattered more and more. In the networked electronic world of the twentieth and twenty-first centuries, data went nova.

But we ought to be careful about how we use these new big data tools in the arena of culture where they are mostly unfamiliar, particularly when they provide results that reinforce what we are inclined to expect. There are a lot of easy mistakes to be made. Consider, for example, the diagram on the following page produced with the same



Graph these comma-separated phrases: data

□ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Image 1

Google tool, depicting the frequency of the term “atomic bombs” from 1800 to 2000 (see *Image 2*). The chart shows a massive usage spike around the end of World War II. This is followed by a substantial fall and then a kind of steady persistence up to the present. The result is so intuitive, it seems virtually unarguable: the first atomic bomb set off a panic, which soon settled into a generalized cultural anxiety. It’s a great story. If only it were true. Factor in the additional term, “nuclear weapons,” and the anxiety no longer levels off (see *Image 3*).

The term “data,” too, turns out to be a good example for how tricky it can be to interpret big data such as that behind the Google Ngram Viewer in the cultural sphere. In the case of “data,” the Ngram Viewer correctly identifies the moment when the concept “data” takes off as a subject of discussion in the general culture, yet it obscures the crucial early moments in the story of “data” in the seventeenth and eighteenth centuries, when “data” so-called first emerged as a term of intellectual importance.

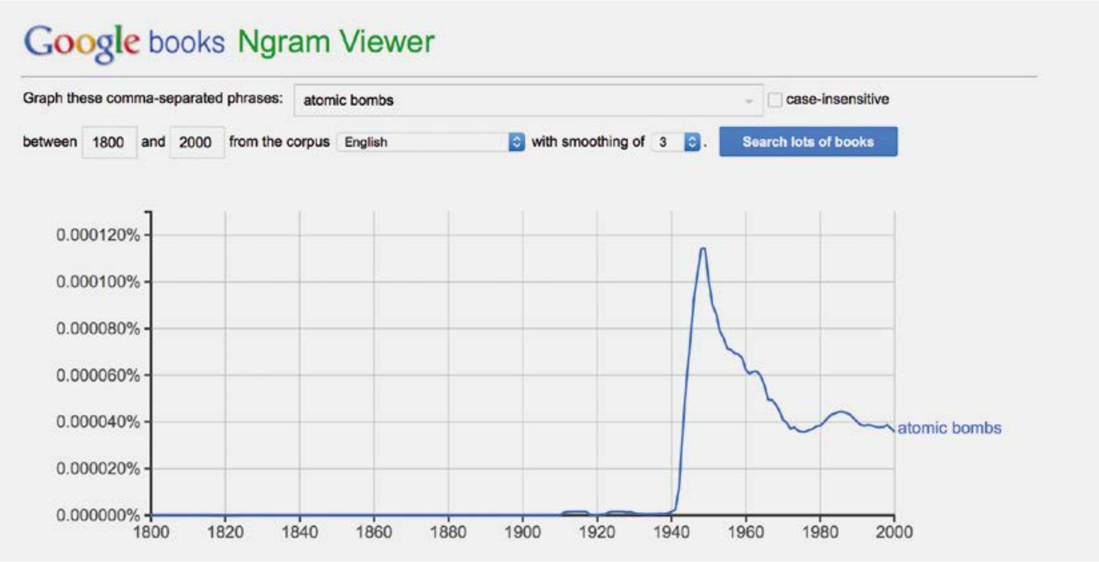
TO BE FAIR, it is hard to blame Google for stumbling over the seventeenth- and eighteenth-century data on “data.” “Data” is a funny term and very hard to search. One reason is that many digital resources, Google Books included, are not yet very good for the period before the nineteenth century. But there are others, too: not least of all, the presence of the word “data” in Latin, a language still used extensively in the early modern period. Careful examination of the sources clarifies why the real quantitative rise of the English word “data” in the seventeenth and eighteenth centuries does not show up in the Google Ngram: it is offset by the decline of Latin at the same time, resulting in a flat curve in the Ngram Viewer.

An excellent indicator of what Google’s Ngram is missing may be found using a much older reference, the *Oxford English Dictionary.* But it would be a mistake to think that this simply reflects the virtues of old humanistic techniques in comparison with the data-driven approaches of today. A monument of pen-and-paper scholarship, the *OED* was nonetheless a highly novel project, remarkable even now. Today, we would call its approach “crowd sourcing”: evidential quotations were contributed by ordinary readers, mailed to the *OED*’s editorial offices on paper slips, and filed in a purpose-built data collection center known as the *scriptorium,* where they were sorted and stored. From the *OED*, we learn that the term “data” emerged in English not in the 1940s but in the 1640s, and the origins of “data” as traced by the *OED* turn out to be surprising.

When it first entered English, “data” was less the province of the scientist than the priest. Consider the very first use of “data” cited by the *OED*, from a series of published letters between the prominent Anglican theologian, Henry Hammond, one-time chaplain to Charles I, and the Presbyterian controversialist, Francis Cheynell.

In the letters, Hammond defends the “set forms” of the Anglican liturgy against Cheynell’s critique. In refuting Cheynell, Hammond paraphrases the tangle of theological propositions posited by his rival (“that there were an ordinary gift of Prayer and that to be stirred up and exercised, that Ministers should study to pray seasonably, . . . that he that hath not ordinary wisdom to pray as he ought, is not called by Christ to be a Minister of the Gospel” . . .) in order to dismiss them with a single stroke. Hammond writes, “Were, I say, all this granted to you, yet sure from all this heap of *data* (if they were *concessa* too) it would not follow that it was necessary . . . to abolish all set forms in the publique service of God.”

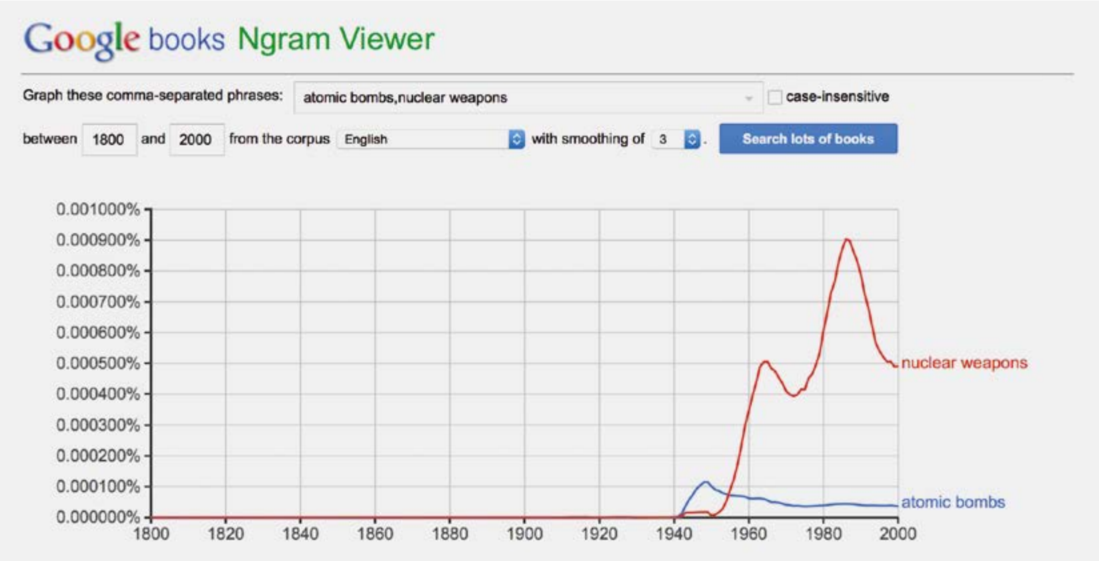
In this first *OED* citation, “data” are stipulations, things taken for granted in an argument. Though he does not agree



Graph these comma-separated phrases: atomic bombs ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Image 2



Graph these comma-separated phrases: atomic bombs,nuclear weapons ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Image 3

with Cheynell's propositions, for the sake of argument, and, because they have no bearing on the larger matter at hand, Hammond *concedes* them. They are, as he says, both *data* and *concessa* . . . and a heaping helping, no less.

The notion that a theological proposition or directive might be called "data" feels strange today, but to Hammond and his contemporaries, it was not surprising in the least. In Latin, after all, the word "data" is nothing more than the plural of the neuter past particle of the verb *dare*, to give. For Hammond, "data," were "givens," facts, propositions, or principles, treated as matters beyond argument because they were true, as in the case of statements in the Bible, or because they were agreed upon for the sake of argument, as here. For an Oxford-educated clergyman steeped in Latin learning, nothing was more natural than to call such givens "data."

For Hammond, "data" does not name one kind of *thing* or another. It simply identifies what is given. A parallel linguistic strategy was employed in mathematics in the same period. In math, one may posit values arbitrarily—let X = 3, and so forth. Such values, too, were known as "data." Here again, it is essential to note that calling something "data" says nothing whatever about its truth. To the contrary, the appellation "data" signals that the question of reference to the world is at least temporarily placed out of bounds. A math problem may well be inspired by facts in the world. The X above might be apples or oranges, but once we decide that X is "data," any question of counting actual fruit is off the table.

In a certain kind of situation, an early modern writer might well have accepted that his or her "data" were "facts," but such an argument would not have meant much one way or another, since the point of calling facts "data" was precisely to moot that question. And in a different kind of case, such as that of Hammond, where the "data" were merely "concessa" for the sake of this particular argument, the author would certainly have rejected the equivalence of "data" and "facts."

A century later, the same principles were still active, but typical uses of the word "data" were changing. This did not happen all at once. Take, for example, the 1761 pamphlet *Experimental Magnetism* by another Oxford-trained scholar, the long-forgotten Temple Henry Croker. In it, Croker makes the following intriguing statement: "Till Experimental Philosophy was introduced, All Science was founded upon Data."

Without some historical context, it is hard to understand what Croker could possibly have meant by this. In fact, from a modern perspective, Croker appears to have his terms exactly backwards. For him, the *abandonment* of "data" was a crucial and definitive step toward modern science. "Data" were not experimental facts; they were axioms given *prior* to experimental investigation. Further scientific advance, Croker writes, "must result, not from Fancy but from Facts, not from artfully devised Systems, but from real Experiments"—from real experiments and facts, not from "data."

Alas, Croker made no great contribution to the history of science. His research into perpetual motion foundered, as did less grandiose plans for a horizontal windmill. Yet his statement about data was no crank gesture. For him, as for many in his day, from John Wesley to Tobias Smollett, "data" still meant "givens," as it did in Latin, and as it did for Hammond and Cheynell in the previous century. But at the time Croker was writing, and as his own argument suggests, assumptions about what constituted givenness were themselves changing. Both the epistemological and the linguistic ground were shifting beneath Croker's feet.

A 1775 letter from Benjamin Franklin to his friend, the scientist and theologian Joseph Priestley, illuminates this point. Here, Franklin employs the term "data," with some irony, to describe an imaginary political calculus on whether or not to go to war. In suggesting that Britain reconsider its opposition to American independence, Franklin writes,

> *Tell our dear good friend [Richard Price], who sometimes has his doubts and despondencies about our firmness, that America is determined and unanimous; a very few tories and placemen excepted, who will probably soon export themselves.—Britain, at the expence of three millions, has killed 150 Yankies this campaign, which is £20,000 a head; and at Bunker's Hill she gained a mile of ground, half of which she lost again by our taking post on Ploughed Hill. During the same time 60,000 children have been born in America. From these data his mathematical head will easily calculate the time and expense necessary to kill us all and conquer our whole territory.*

Franklin employs "data" to refer to quantitative facts gathered through observation and collection and subject to mathematical analysis, much as we do today. That Franklin might use "data" so casually suggests he took his usage to be transparent. What's more, even fifteen years earlier, when Croker was writing, it was already possible to poke fun at the pseudo-scientific way that people talked about "data" as in this social satire modeled on Laurence Sterne:

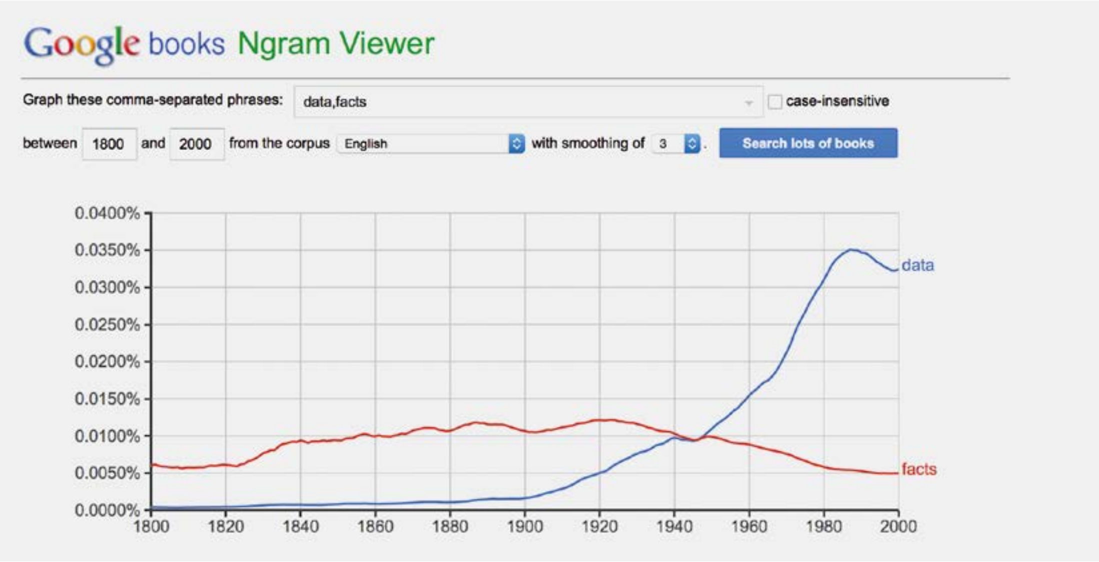
> *Sarah, now advanced to her seventy-sixth year, was, had she been stretched out to her utmost length about five feet three inches, honest measure; and as she was generally seen making an obtuse angle from her middle of about 95°36', it will be easy for mathematicians to compute the length of the line, they will imagine to be extended from the tip of her coif to the toe of her shoe. But as this is a matter of science, out of my reach, I can but shew my good will by assigning these data, little doubting that my second edition of this third volume will contain the calculation at length to one millionth part of an hair's breadth.*

For the author, John Carr, "data" conjures empirical, quantitative science in both its usual practice and its excess—the beachhead of the calculator in the fields of social life. From the middle of the century, Croker's usage was waning, and a modern sense was catching on.

"Data" may also be "facts," but by using the term "data" we are putting them in a specific rhetorical light, accepting them as stipulated. When we use "facts," we are placing emphasis elsewhere, as etymology suggests. In contrast to "data," from *dare*, "to give," "fact" is from the Latin verb *facere*, meaning "to make" or "to do." Thus, when we call something a fact, we emphasize that it truly exists. In a certain kind of argument, "facts" are likely also to be treated as "givens" or "data." In another kind of argument, in algebra for example, "givens" may just as well be arbitrary. What unites these cases, what makes data "data," is not existential truth but status as an accepted premise for argument. Moreover, as often as not, in the early modern period, facts and data were framed as contraries. (In our age of "big data," this possibility feels arrestingly prescient.)

Ironically, with the rise of empiricism in the eighteenth century, the terminological waters grew cloudier. The term "data" grew in importance. It was employed in more arenas, and the fields of mathematics and theology accounted for an ever smaller fraction of total uses. At the same time, "data" came more often to be used in the sense of raw, unprocessed information. As "data" came to be regularly employed in empirical fields such as medicine, finance, natural history, and geography, it became usual to think that "data" could be the result of an investigation, not only its premise. Broadly speaking, this association held for the next century and a half.

And then something changed again.



Graph these comma-separated phrases: data,facts □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Image 4

With the emergence of electronic computing, a new terminological need arose. Just as in the seventeenth century, in the second half of the twentieth, it became important to distinguish between facts and givens. This second time around, some term was needed to name the values upon which we calculate, independent of the question of what they represent. Some term was needed to name the stuff that computers work on (see *Image 4*).

Like the first transformations in the term “data” when it came out of Latin, this more recent change is hard to perceive from the simple quantitative data on language alone—what linguists refer to as the “bag of words.” In the word counts produced from Google Books and other corpus-based resources, the history of “data” looks like one big explosion starting in the middle of the twentieth century, cresting around its end. Of course, that’s right from one perspective: we live in an age of data, both big and personal. And it is no accident that the word “data” shows up so frequently in our literature. What this quantitative account misses is the way in which the application of “data” changed during this same period.

Yes, in strictly quantitative terms, “data” mattered more in the nineteenth century than in the eighteenth, more in the twentieth than the nineteenth, and, toward the end of the twentieth century, more than ever before. But, in the realm of usage, the story is a bit more back-to-the-future than onward-and-upward. That is to say, the *ways* we use “data” now hark back all the way to the days of Henry Hammond.

AS HAMMOND’S USAGE SUGGESTS, from the beginning, “data” was a rhetorical concept. “Data” means today, as it always has, that which is given. As a consequence, for three centuries, the term has served as a kind of historical and epistemological mirror, showing us what we take for granted. Without changing meaning, “data” has repeatedly changed referent. It went from being reflexively associated with things outside of any possible process of discovery to the very paradigm of what one seeks through experiment and observation. It changed referent again in our contemporary period when it came to be associated with quantified information structured, stored, and communicated by computer.

This most recent change laid the linguistic groundwork for a wide range of now-ubiquitous uses such as “personal data,” “big data,” and the like. But we should be clear: from the point of view of our everyday language, this recent explosion of “data” is only a revolution in that older, classical sense of the term, as a circling back whence we came. And our understanding of how “data” works in our language and culture may benefit from this perspective.

“Data” matters enormously in our world and the ways we talk about it. It is ubiquitous and powerful. For this reason, it is tempting to imagine that “data” is also new. From the point of view of artifacts—mortality tables, account books, temperature records, and the like—we would do well to take a longer perspective. This is true, too, for language. Here, a little history, and indeed a little data, taken with the correct dose of salt, may clarify matters and put them in a different light.

It is tempting to want to discover the essence of “data,” to determine exactly what kind of fact it is. But this misses the most important reason why the term “data” has proven useful in so many areas of our contemporary culture. “Data” first emerged as a tool for setting aside questions of ontology. It re-emerged at the center of our general culture as it produced ontologies of its own. □