1 *Data before the Fact*

Daniel Rosenberg

Is data modern? The answer depends on what one means by "data" and what one means by "modern." The concept of data specific to electronic computing is evidently an artifact of the twentieth century, but the ideas underlying it and the use of the term are much older. In English, "data" was first used in the seventeenth century. Yet it is not wrong to associate the emergence of the concept and that of modernity. The rise of the concept in the seventeenth and eighteenth centuries is tightly linked to the development of modern concepts of knowledge and argumentation. And, though these concepts long predate twentieth-century innovations in information technology, they played a crucial role in opening the conceptual space for that technology. The aim of this chapter is to sketch the early history of the concept of "data" in order to understand the way in which that space was formed.

My point of departure for this project is a happenstance textual encounter that eventually became a kind of irritation: in reading the 1788 *Lectures on History and General Policy* by the polymath natural philosopher and theologian Joseph Priestley, I stumbled on a passage in which Priestley refers to the facts of history as "data."[1] In the text, his meaning is clear enough, but the usage surprised me. I had previously associated the notion of data with the bureaucratic and statistical revolutions of the nineteenth century and the technological revolutions of the twentieth. And while I don't begrudge Priestley his use of the term, it seemed very early.

Of course, if one were to pick an eighteenth-century figure likely to be interested in data, Priestley is about as good a choice as one might make. After all, Priestley was an early innovator in the field we now call data graphics. His 1765 *Chart of Biography* is a great achievement in this field, an engraved double-folio diagram displaying the lives of about two thousand famous historical figures on a measured grid.[2] It was one of the earliest works to employ the conventions of linearity and regularity now common in

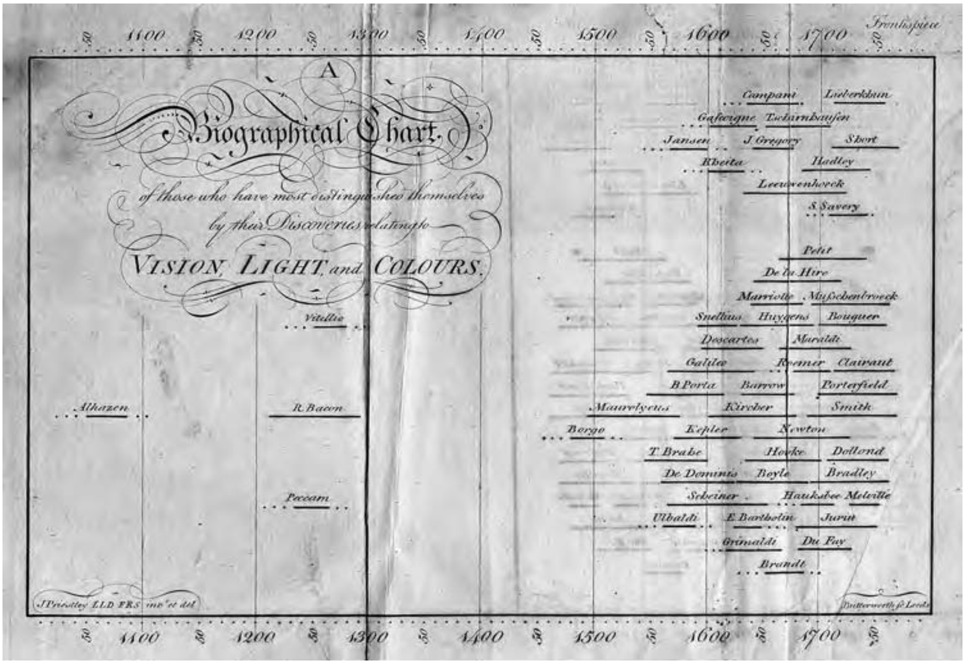


Figure 1.1 Joseph Priestley, Biographical Chart from *History and Present State of Discoveries Relating to Vision, Light, and Colours*, 1772. Biographical information extracted from *Chart of Biography* showing lives of key figures in the history of optics. Courtesy of Rare Book Division. Department of Rare Books and Special Collections, Princeton University Library.

historical timelines and the most important work of its kind published in the eighteenth century (figure 1.1).

Furthermore, Priestley was an empiricist and an experimentalist—among his many achievements, the isolation of oxygen from air in 1774 is the best remembered—and he brought an interest in aggregate phenomena to the many domains in which he researched and wrote. In his historical works, both diagrammatic and textual, Priestley was not only interested in individual facts—when was Newton born, when did he die?—but in large constellations of information. He examined fields of scientific endeavor quantitatively, grouping historical figures by their domains of achievement and plotting their lives on a measured timeline in order to observe patterns of occurrence and variations in density.

Framing historical data in a graphic such as Priestley's is second nature today, and this is in part due to Priestley himself. Today, we look at timelines and intuit historical patterns with no trouble. But all of this was new when Priestley published his charts, and the aggregate views they offered were regarded as an important and novel contribution to both social and natural science. Indeed, it is the *Chart of Biography*, not an achievement in experimental science that is named on Priestley's document of induction to the Royal Society. Later writers such as the political economist William Playfair, who debuted early versions of the line graph and bar chart in his 1786 *Commercial and Political Atlas*, credited Priestley for his innovative work in this area, too.[3]

In fact, the term "data" appears in Priestley's works many times. In his *Experiments and Observations on Different Kinds of Air*, Priestley uses "data" to refer to experimental measurements of volume. In the *Evidences of Revealed Religion*, Priestley notes that scripture offers us "no sufficient data" on the physical nature of Christ's resurrected body. In his *Essay on a Course of Liberal Education for Civil and Active Life*, Priestley writes, "Education is as much an art (founded, as all arts are, upon science) as husbandry, as architecture, or as ship-building. In all these cases we have a practical problem proposed to us, which must be performed by the help of data with which experience and observation furnish us."[4]

Nor is Priestley unique in this. The term "data" appears in a wide variety of contexts in eighteenth-century English writing. But what were these early usages? What was their importance in the language and culture of the eighteenth century? And what was their connection to the usages familiar today? What was data apart from modern concepts and systems of information? What notion of data preceded and prepared the way for our own?

All of these questions are that much more pressing since, in recent histories of science and epistemology, including foundational works by Lorraine Daston, Mary Poovey, Theodore Porter, and Ann Blair, the term "data" does heavy lifting yet is barely remarked upon.[5] Consider, for example, the first lines of Mary Poovey's landmark book, *A History of the Modern Fact*. "What are facts?" Poovey asks. "Are they incontrovertible data that simply demonstrate what is true? Or are they bits of evidence marshaled to persuade others of the theory one sets out with?" Facts may be conceived either as theory-laden or as simple and incontrovertible, Poovey says. In the latter case, we call them "data."[6]

Of course, it would not be difficult to engage in some one-upmanship. If facts can be deconstructed—if they can be shown to be theory-laden—surely data can be too. But it is not clear that such a move would be useful from either a conceptual or a practical point of view. The existing historiography of the fact is strong in its own terms, and no special harm is done by an unmarked, undeconstructed deployment of the term "data." What is more, there is a practical consideration: one has to have some language left to work with, and after thrilling conceptual histories of truth, facts, evidence, and other such terms, it is helpful to retain one or two irreducibles. Above all, it is crucial to observe that the term "data" serves a different rhetorical and conceptual function than do sister terms such as "facts" and "evidence." To put it more precisely, in contrast to these other terms, the semantic function of data is *specifically* rhetorical.

The question then is: what makes the concept of data a good candidate for something we would *not* want to deconstruct? Understanding this requires understanding what makes data different from other, closely related conceptual entities, where data came from, and how it carved out a distinctive domain within a larger conceptual and discursive sphere.

So, what was data prior to the twentieth century? And how did it acquire its pre-analytical, pre-factual status? In this, etymology is a good starting point. The word "data" comes to English from Latin. It is the plural of the Latin word *datum*, which itself is the neuter past participle of the verb *dare*, to give. A "datum" in English, then, is something given in an argument, something taken for granted. This is in contrast to "fact," which derives from the neuter past participle of the Latin verb *facere*, to do, whence we have the English word "fact," for that which was done, occurred, or exists. The etymology of "data" also contrasts with that of "evidence," from the Latin verb *vidēre*, to see. There are important distinctions here: facts are ontological, evidence is epistemological, data is rhetorical. A datum may also be a fact, just as a fact may be evidence. But, from its first vernacular formulation, the existence of a datum has been independent of any consideration of corresponding ontological truth. When a fact is proven false, it ceases to be a fact. False data is data nonetheless.

In English, "data" is a fairly recent word, though not as recent as one might guess. The earliest use of the term discovered by the *Oxford English Dictionary* occurs in a 1646 theological tract that refers to "a heap of *data*." It is notable that this first *OED* citation is to the plural, "data," rather than the singular, "datum." While "datum," too, appeared in seventeenth-century English, its usage then, as now, was limited—so limited, that in contrast to the well-accepted usage of the plural form, some critics have doubted whether the Latin *datum* was ever naturalized to English at all.[7]

"Data" did not move from Latin to English without comment. Already in the eighteenth century, stylists argued over whether the word was singular or plural, and whether a foreign word of its ilk belonged in English at all. In Latin, *data*, is always plural, but in English, even in the eighteenth century, common usage has allowed "data" to function either as a plural or as a collective singular. Guides differ, but usage authorizes both, and analogy to parallel Latin loan words gives no unambiguous guide.[8] Indeed, it seems preferable in modern English to allow context to determine whether the term should be treated as a plural or as a collective singular, since the connotations are different. When referring to individual bits or varieties of data and contrasting them among one another, it may be sensible to favor the plural as in "these data are not all equally reliable"; whereas, when referring to data as one mass, it may be better to use the singular as in "this data is reliable." According to Steven Pinker, in English today, the latter usage has become usual.[9] The fact that a standard English dictionary defines a "datum" as a "piece of information," a fragment of another linguistically complex mass noun, further strengthens this intuition.[10]

As Pinker argues, however much priggish pleasure professors may take in pointing out that the term *data* in Latin is plural, foreign plurals may be deployed in English as singulars. Were they not, we would be incorrect in referring to *an* agenda, *an* insignia, or *a* candelabra. Each of these words is a plural in its source language. Moreover, Pinker writes, "whenever pedants correct, ordinary speakers hypercorrect, so the attempt to foist 'proper' Greek and Latin plurals has bred pseudo-erudite horrors such as *axia* (more than one *axiom*), *peni*, *rhinoceri*, and . . . *octopi*." None of these exist in the source language. In the case of the last: "It should be . . . 'octopuses.' The *-us* in *octopus* is not the Latin noun ending that switches to *-i* in the plural, but the Greek *pous* (foot). The etymologically defensible *octopodes* is not an improvement."[11]

However controversial they may have been, in seventeenth-century English, neither "data" nor "datum" was particularly common. In these early years, the term "data" was still employed, especially in the realm of mathematics, where it retained the technical sense that it has in Euclid, as quantities *given* in mathematical problems, as opposed to the *quaesita*, or quantities *sought*, and in the realm of theology, where it referred to scriptural truths—whether principles or facts—that were given by God and therefore not susceptible to questioning. In the seventeenth century, in theology, one could already speak of "historical data," but "historical data" referred to precisely the sorts of information that were outside of the realm of the empirical. These were the God-given facts and principles that grounded the historian's ability to determine the *quaesita* of history.

This formulation is not marginal: technical historical practice during the early modern period involved accommodation of historical facts to scriptural data in order to make the unknown known. Some of the most heroic efforts of this sort took place in the realm of chronology, especially in efforts to correlate European and non-European historiographical traditions. Ancient records of comets and other astronomical phenomena that posed interpretive problems for histories based on scripture provide other examples. And it is notable that chronology is one of the fields in which the English word "data" flourished earliest.

In seventeenth-century philosophy and natural philosophy, just as in mathematics and theology, the term "data" functioned to identify that category of facts and principles that were, by agreement, beyond argument. In different contexts, such agreement might be based on a concept of self-evident truth, as in the case of biblical data, or on simple argumentative convenience as in the case of algebra, given $X = 3$, and so forth. The term "data" itself implied no ontological claim. In mathematics, theology, and every other realm in which the term was used, "data" was something given by the conventions of argument. Whether these conventions were factual, counter-factual, or arbitrary had no bearing on the status of givens as data.

When used in English, "data" had a much narrower meaning than did either *data* in Latin or "given" in English. Whether in mathematics, theology, or another field, use of the term "data" emphasized the argumentative context as well as the idea of problem-solving by bringing into relationship things known and things unknown. The "heap of data" that the *OED* unearthed in Henry Hammond's 1646 theological tract, *A Brief Vindication of Three Passages in the Practical Catechisme*, is not a pile of numbers but a list of theological propositions accepted as true for the sake of argument—that priests should be called to prayer, that liturgy should be rigorously followed, and so forth.[12]

It is also the case that the Latin word *data*, as a conjugation of the verb *dare*, was in constant use during the seventeenth and eighteenth centuries. In early modern Latin, as in classical Latin, *data* is everywhere. But *data* in Latin rarely translates to "data" in English. A 1733 translation of Bacon's *Novum Organum* gives a good example of the dynamic. Aphorism 105 of Book 1 of the *Novum Organum* reads as follows:

> *Inductioenim quae procedit per enumerationem simplicem res puerilis est, et precario concludit, et periculo exponitur ab instantia contradictoria, et plerumque secundum pauciora quam par est, et ex his tantummodo quae praesto sunt, pronunciat.*

> For that Induction which proceeds by simple Enumeration, is a childish thing; concludes with Uncertainty; stands exposed to Danger from contradictory Instances; and generally pronounces upon scanty Data; and such only as are ready at hand.[13]

Here we have the word "data" in the English translation, but no *data* at all in the Latin original. In fact, in the Latin, we have not even got a substantive, only the neuter substantival adjective *pauciora*, which means a small number of something—a something that Bacon's eighteenth-century translator took to be "data." All of this is made even more complicated by the fact that Bacon himself did not use the term "data" when writing in English. "Data" arrives in Bacon's corpus belatedly, posthumously, and just exactly when we would expect it, in the early 1730s.

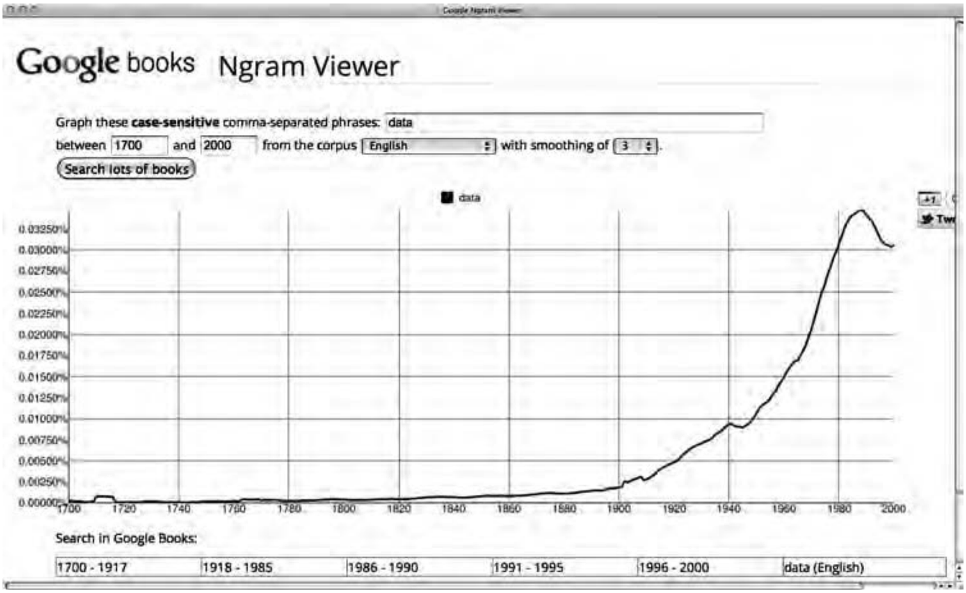
Nor is the phenomenon of posthumous data-fication limited to Bacon. The same effect took hold in the works of Newton at virtually the same time. Bacon's translator, the physician Peter Shaw, interpolated the term "data" into the *Novum Organum* in 1733; Newton's translator, John Colson, got "data" into Newton's works three years later in 1736. In contrast to what happened in the case of Bacon, Colson did not actually put the English word "data" in Newton's mouth. But he used the term extensively in his analytic notes on Newton's works. Usually, he employed "data" in the restrictive Euclidean context in the contrast of mathematical *data* and *quaesita*. But not always. Colson's most notable usage occurs in his hagiographic introduction to his translation of Newton's *The Method of Fluxions and Infinite Series*.

> To improve Inventions already made, to carry them on, when begun, to farther perfection, is certainly a very useful and excellent Talent; but however is far inferior to the Art of Discovery, as having *pou sto* (foundations), or certain data to proceed upon and where just method, close reasoning, strict attention, and the Rules of Analogy, may do very much. But to strike out new lights, to adventure where no footsteps had been set before . . . this is the noblest Endowment that a human Mind is capable of, is reserved for the chosen few . . . and was the peculiar and distinguishing Character of our great Mathematical Philosopher.[14]

The quotation is interesting both because of the forcefulness of the distinction that it makes between the arts of invention and discovery and because of the high value that it places on the latter. Discovery, according to Colson is "the noblest Endowment" of the human mind; invention, on the other hand, is merely "useful."

From the point of view of this lexicographic history, what is most interesting is the presence of the English word "data," here used in a mode that is entirely characteristic of Colson's period. *Pou sto* is ground to stand upon, as in the famous phrase of Archimedes—"give me ground to stand upon, and I will move the world"—and undoubtedly Colson intended the phrase to be heard in this context. By the 1730s, there would have been nothing odd about using the term "data" to refer to facts discerned through experimentation, but here Colson uses "data" in the usual competing sense of principles or axioms given on the basis of which methods may be devised and facts discovered.

This is what one learns from reading. But what about the data on "data"? Might a quantitative approach be possible too? Might it be possible to study the corpus of printed English books in order to discover when "data" became a common term in English, how it was naturalized from Latin, and when it achieved its various meanings? Fortunately, today we are swimming in data for lexicographic research provided by both specialized and general databases along a spectrum from stand-alone electronic books to massive archiving and scanning endeavors such as Project Gutenberg and Google Books. Some of these resources are set up in ways that generally mimic print formats. They may offer various search features, hyperlinks, reformatting options, accessibility on multiple platforms, and so forth, but, in essence, their purpose is to deliver a readable product similar to that provided by pulp and ink. Others—still relatively few—foreground the aggregate and statistical features of the textual corpora that they access, and in a few cases



Google books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases: data

between 1700 and 2000 from the corpus English with smoothing of 3.

Search in Google Books:

1700 - 1917	1918 - 1985	1986 - 1990	1991 - 1995	1996 - 2000	data (English)

Figure 1.2 Image: Relative frequency of "data" in Google Books, by year, 1700–2000, generated by Google Ngram Viewer.

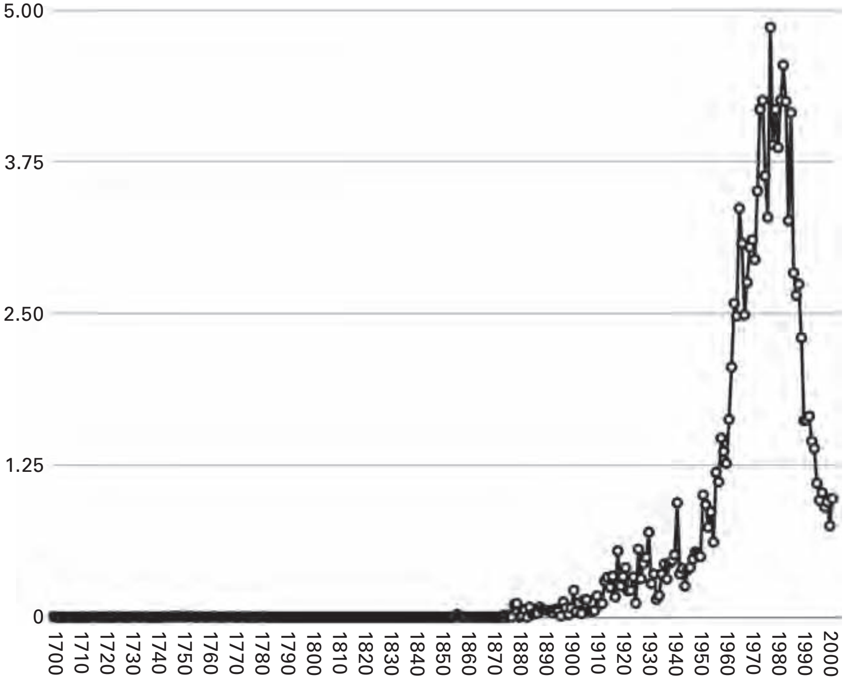


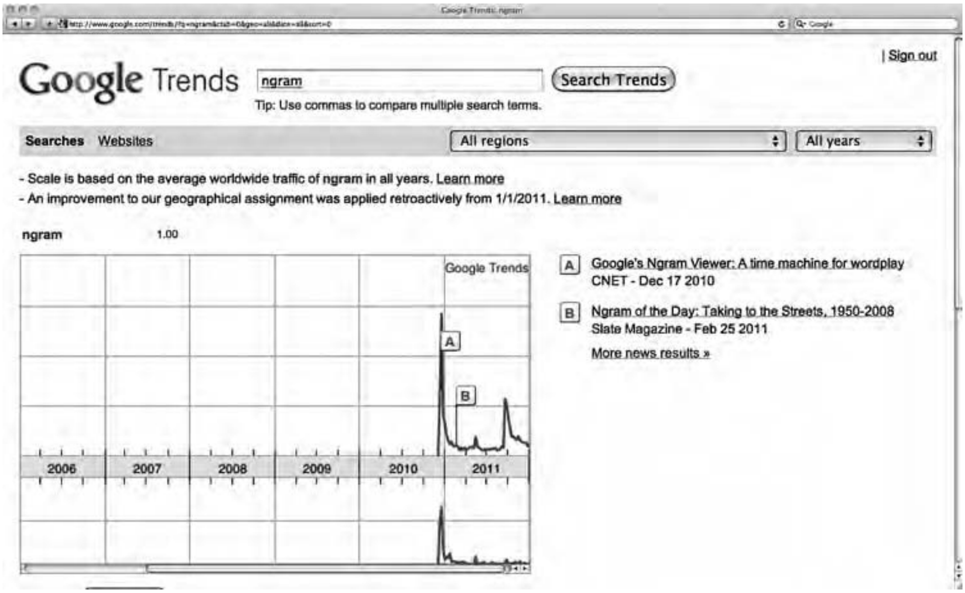
Figure 1.3 Relative frequency of "data" in Google Books corpus, 1700–2000, generated manually. *Note:* Data generated by repeated date-limited Google searches.

they do so even to the exclusion of the possibility of conventional reading, from beginning to end.

Much has been written about Google Books, but a large part of this scholarly literature has focused on the ways in which Google interacts with and places stress upon authors, publishers, libraries, and competing databases—stress that largely has to do with the fate of books in the electronic age.[15] Since the beginning of 2011, however, new attention has been focused on the research potential of Google Books as a linguistic corpus rather than as an electronic library. To facilitate research, Google has been making its book corpus accessible in two new ways: the raw data, abstracted from individual works, can be downloaded for analysis according to the interests of individual researchers, or it can be searched through a simple online interface called the Google Books Ngram Viewer. An "ngram" is a phrase consisting of a defined number of words (n): the Ngram Viewer allows corpus searches on these phrases and returns statistical results. While the Ngram Viewer is limited in the kinds of searches it can perform, its basic trick is already impressive: presented with one or more search phrases of up to five words and a historical timeframe, the Ngram Viewer can instantly produce a graph of relative usage frequency over time.

A team of Harvard researchers led by the physicist Erez Lieberman Aiden and the biologist Jean-Baptiste Michel designed the Ngram Viewer. They introduced it with a clever publicity strategy: they aimed both low and high, promoting the Ngram Viewer as both an amusing geegaw and a tool for serious scholarly research. In their January 2011 *Science* article, "Quantitative Analysis of Culture Using Millions of Digitized Books," Michel and Aiden present the Ngram Viewer as a tool for what they call *culturomics*, quantitative cultural analysis modeled on *genomics* and the other *-omic* fields booming in the natural sciences.[16]

Michel and Aiden's publicity strategy proved successful, stirring up notice in key media venues such as the *New York Times* and in the blogosphere, where the ease of use and linking prompted a lot of kitchen culturomics. Briefly, it seemed that everyone was ngramming.[17]



ngram

Tip: Use commas to compare multiple search terms.

All regions

All years

- Scale is based on the average worldwide traffic of ngram in all years. Learn more
- An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

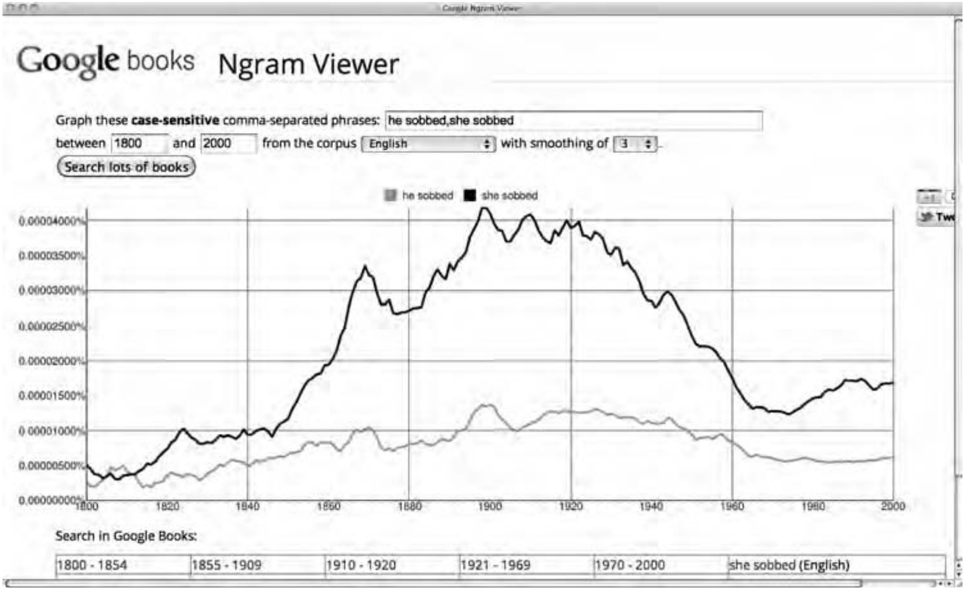
ngram 1.00

A Google's Ngram Viewer: A time machine for wordplay
CNET - Dec 17 2010

B Ngram of the Day: Taking to the Streets, 1950-2008
Slate Magazine - Feb 25 2011

More news results »

Figure 1.4 Search volume for "ngram," May 2010–December 2011, generated by Google Trends.



Google books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases: he sobbed,she sobbed

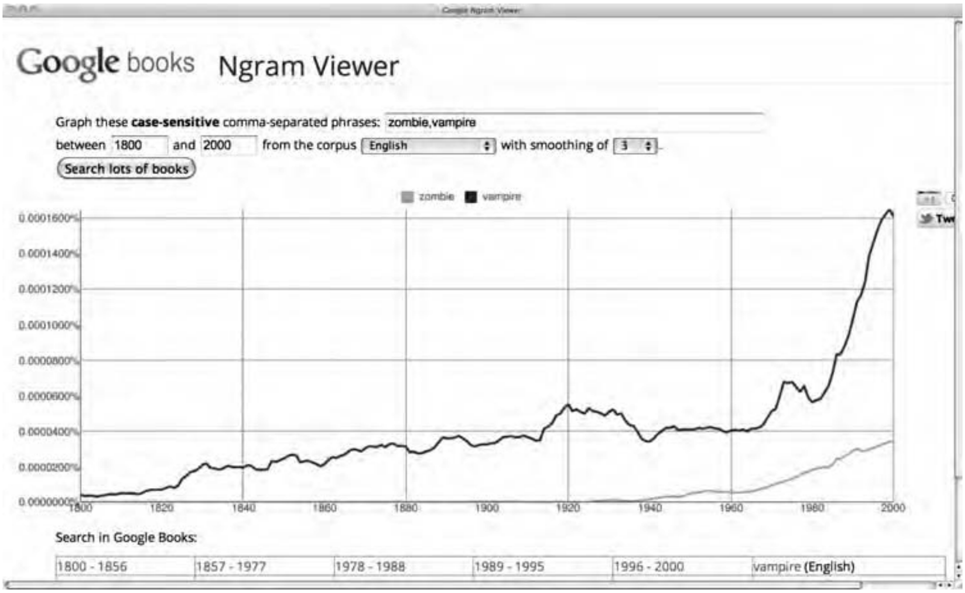
between 1800 and 2000 from the corpus English with smoothing of 3.

Search lots of books

Search in Google Books:

1800 - 1854 | 1855 - 1909 | 1910 - 1920 | 1921 - 1969 | 1970 - 2000 | she sobbed (English)

Figure 1.5 Relative frequency of "he sobbed" vs. "she sobbed" in Google Books, 1800–2000, as conceived by jezebel.com, generated by Google Ngram Viewer.



Google books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases: zombie,vampire

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

Search lots of books

Search in Google Books:

1800 - 1856 | 1857 - 1977 | 1978 - 1988 | 1989 - 1995 | 1996 - 2000 | vampire (English)

Figure 1.6 Relative frequency of "zombie" vs. "vampire" in Google Books, 1800–2000, as conceived by theatlantic.com, generated by Google Ngram Viewer.

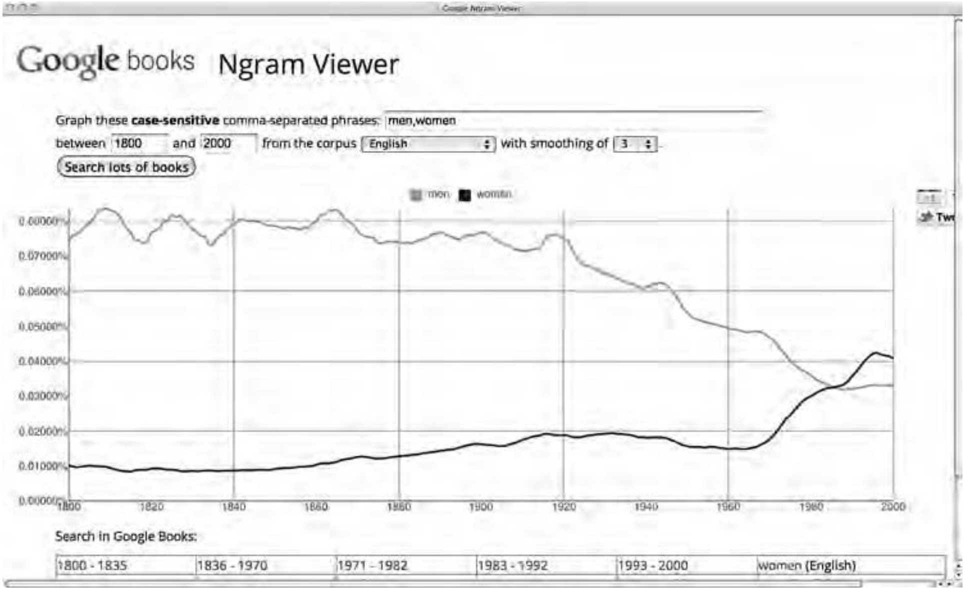
The Harvard team got the ball rolling with some provocative diagrams of their own, plotting the changing importance in the linguistic corpus of a variety of people, events, and things. "'Galileo,' 'Darwin,' and 'Einstein' may be well-known scientists," write Michel and Aiden, "but 'Freud' is more deeply ingrained in our collective subconscious." "In the battle of the sexes, 'women' are gaining ground on the 'men.'"[18] Even *years* themselves could be tracked through the corpus, and these produced interesting regularities.

> Just as individuals forget the past, so do societies. To quantify this effect, we reasoned that the frequency of 1-grams such as "1951" could be used to measure interest in the events of the corresponding year, and we created plots for each year between 1875 and 1975. The plots had a characteristic shape. For example, "1951" was rarely discussed until the years immediately preceding 1951. Its frequency soared in 1951, remained high for 3 years, and then underwent a rapid decay, dropping by half over the next 15 years. Finally, the plots enter a regime marked by slower forgetting: Collective memory has both a short-term and a long-term component. But there have been changes. The amplitude of the plots is rising every year: Precise dates are increasingly common. There is also a greater focus on the present. For instance, "1880" declined to half its peak value in 1912, a lag of 32 years. In contrast, "1973" declined to half its peak by 1983, a lag of only 10 years. We are forgetting our past faster with each passing year.[19]

Precisely what one makes of these word-frequency trends is, of course, open to question. "Women" are not women, nor are "men" men, and there are good bureaucratic reasons unrelated to "collective memory" why 1951 would appear in documents from 1950, but the researchers argue that within the terms of the linguistic corpus the data speaks for itself.

The value of these diagrams immediately became a subject of scholarly debate. Some humanities scholars were highly skeptical; others, such as Anthony Grafton and Geoffrey Nunberg received them more favorably. Grafton invited Michel and Aiden to address the American Historical Association in two special sessions in 2011 and 2012, the second of which was substantially devoted to rebutting misconceptions including the notion that culturomics sets out to replace historians with computer programmers.[20]

More significant than the Ngram Viewer was Google's decision to make its raw data—if the term can be applied at all—available for download so that scholars could run the numbers themselves without going through the ngram interface.[21] This resource



Graph these **case-sensitive** comma-separated phrases: men,women

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

Figure 1.7 Relative frequency of "men" vs. "women" in Google Books, 1900–2000, as conceived by Michel and Aiden, generated by Google Ngram Viewer.

is likely to produce significant new research; at the same time, it should also elicit new critique.

At the time that I began research for this study, the Google Ngram Viewer was not yet available, and although it was possible to produce similar results by hand, at that time, Google Books offered neither the most obvious nor the most promising corpus with which to conduct a study such as this. As figure 1.3 demonstrates, repeating a search for the term "data" year by year and dividing the results by the results of searches for a control word in each of the same years in order to offset the effect of changing corpus size produces a curve consistent with that produced by the Ngram Viewer. This gives some indication of the promise of the corpus but only creases its surface.

In any event, I did not begin with Google Books, but rather with the subscription database Eighteenth-Century Collections Online (ECCO) from the educational publisher, Gale. ECCO is in many ways a primitive tool, and it suffers from several of the key faults for which Google Books has been criticized including inconsistent scanning quality. But ECCO has some notable advantages too. The corpus of ECCO I, based on the English Short Title Catalogue, is large, comprising more than 136,000 unique titles, 155,000 volumes, and 26 million pages of text, backed up by an accessible analog microfilm collection from which it was generated and by well-catalogued books. A later supplement, ECCO II, raises the totals to 182,000, 205,000, and 32 million, respectively. Additionally, ECCO is well defined and much more stable than Google Books, which is changing all the time. ECCO's sources are well chosen, well known, and accessible. Its out-of-the-box search functions are more flexible. And at this point in time, the metadata is much better.

In fact, there is so much that is good about ECCO that a decade ago one might have thought ECCO would have had the kind of revolutionary effect on scholarship that Google and the culturomics advocates claim Google Books will have today. ECCO has opened new research avenues, but it hasn't made that kind of impact. In 2002, ECCO's publisher promoted it as a "research revolution." A breathless review called it a "resource that scholars will die for."[22] My graduate school friends called it "the dissertation machine."

The first thing that limited ECCO's effect, of course, is that it was not made openly available like Google Books. Additionally, though ECCO is a full-text database, it does not allow users to cut and paste text. And while users can search for words under the page images, they cannot reveal what the computer sees; they cannot see the characters that the computer recognizes in the page image. Ironically, over time ECCO's publisher has loosened its rules on downloading page images. So, for database subscribers, it has become easy and quick to download page images of full books from ECCO. Yet regular users cannot even download a single page of text as interpreted by ECCO's optical character recognition (OCR) software, which suggests that over time Gale determined there is no percentage in books, not even in digitized images of books, unless the books are already packaged as data.[23]

The future is in data.

Using ECCO, I began trying to understand the sense of "data" in Priestley. Happily, my first searches turned out to be promising. On the one hand, the ECCO results are consistent with those of Google. Speaking from a strictly quantitative point of view, the big "data" takeoff is unquestionably a post-Enlightenment phenomenon. On the other hand, ECCO shows clear trends in usage in the eighteenth century that laid the foundations for all later developments, which are difficult to perceive in Google's projections. The eighteenth century produced important new ways of thinking data, and Priestley was situated, felicitously, just exactly where those new ways of thinking happened.[24]

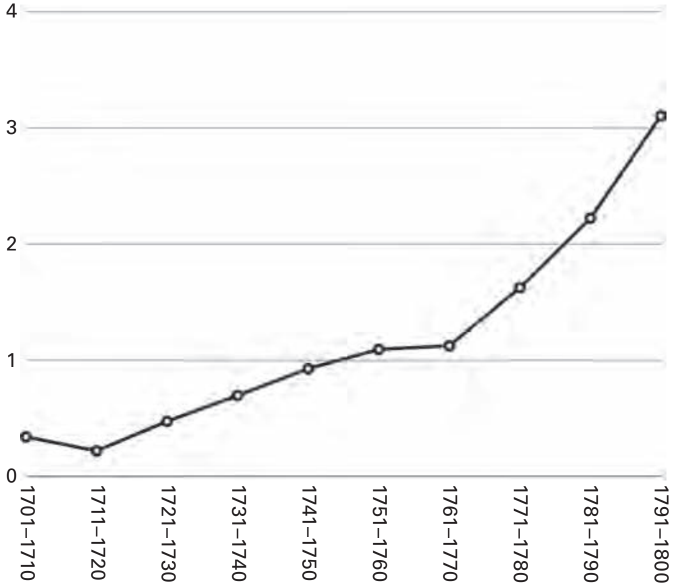


Figure 1.8 Percentage of works including the English common noun "data" in the corpus of ECCO I, 1701–1800.

The ECCO numbers are interesting, and they are also surprising in their clarity given the Google Books results, which suggest that the strong trends in the history of the term "data" begin in the nineteenth century and only accelerate definitively in the twentieth. First, from a statistical point of view, "data" was neither a rare nor an especially common term in eighteenth-century English. For comparison, a simple full-text ECCO search for the word "truth" produces hits in about 112,000 books or about 82 percent of the 136,000 total included in ECCO I. "Evidence" shows up in 66,000 books or 49 percent of total. "Fact" appears in about 35,000 or 28 percent. Even if we were to take the most generous count for "data," uncorrected for Latin usages, scanning errors, and so forth, we would find no more than 10,545 works in which "data" appears, or about 8 percent of total. And a stricter analysis of those occurrences produces a significantly smaller number, closer to 2 percent. In the eighteenth century, "data" was still a term of art.[25]

The further one goes into the data on "data," the more complicated it becomes. In my larger project, I aim to examine every usage of the term "data" in the ECCO corpus,

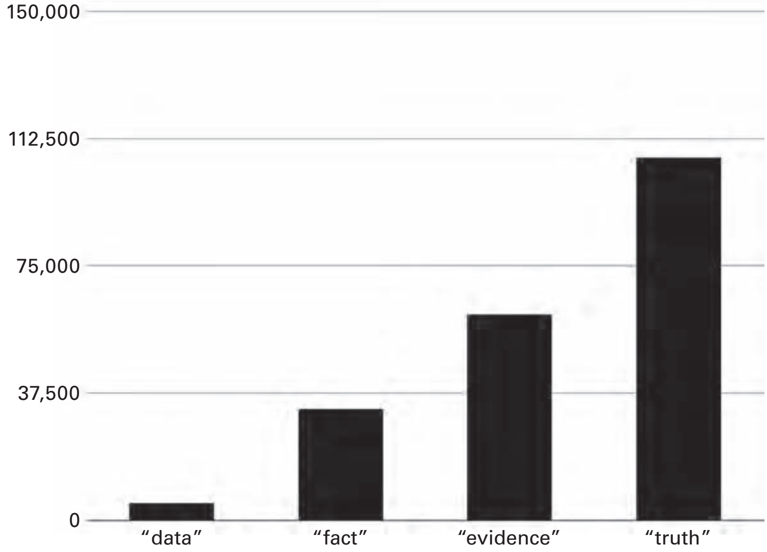


Figure 1.9 Works in the ECCO I corpus containing "data," "fact," "truth," and "evidence," 1701–1800.

not only to count for frequencies but also to examine each usage in context and to code each for semantic characteristics. The first and most pervasive problem that has turned up in this work is that a majority of usages of "data," even in the English language books in the database, turn out to be Latin. Often the Latin word *data* appears in quotations, footnotes, or conventional phrases such as *data desuper* (given from above) included in longer English texts. Other hits refer to the title of Euclid's book *Data*. Still others turn out to be scanning errors. In one instance, the search engine pulled up a reference to a certain King Data, a giant who fattened his twenty-five children by feeding them on puddings stuffed with enchanted herbs.[26] As a consequence it has been useful to examine hits individually, to sort the good from the bad and to code them, to engage in the constructive process of data making so well described in recent ethnographies of scientific practice. My own data may once have been raw, but by the time I began any serious interpretation, I had cooked it quite well.

Getting an accurate count for "data" has been a challenge. The process of scrutinizing each hit and eliminating those that were not English-language common nouns shrank the pool of viable instances. In fact, it certainly shrank the total number too far. Many works identified by ECCO as containing the word "data" in fact contain more instances

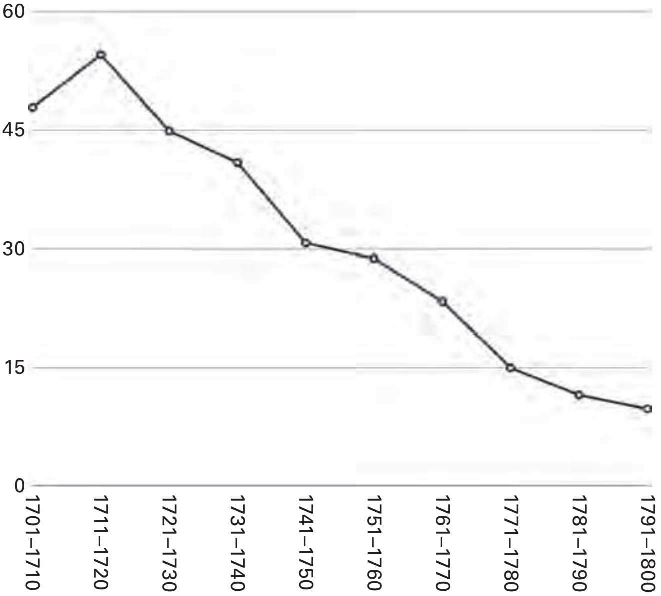


Figure 1.10 "Data" in Latin as Percentage of Total "Data" Hits in ECCO I, 1701–1800.

than ECCO shows; that is, even in works where the OCR algorithms correctly identified "data" once, they often missed it other times. And it is safe to say that there are at least as many instances in which data escaped the ECCO text search as instances in which ECCO thought it saw "data" but was mistaken. Estimating the numbers is challenging: on the one hand, there are more ways for an OCR program to overlook an instance of the word than to produce a false hit; on the other hand, since the term "data" frequently appears in a given work more than once (roughly 38 percent of the time according to my results), a significant number of OCR misses will be compensated for by correct recognitions of occurrences elsewhere in the same work.

Because the number of meaningful search hits for "data" turned out to be only about 2,300, it was possible to read them all well, to code them according to several protocols, and to produce very rich records for each instance. It was also possible to read extensively in the source works to gain a nuanced understanding of context. This has allowed me to pose a fairly wide variety of questions about the term and about key trends in its usage. And while this research is not yet complete, there are already a number of preliminary results, of which I highlight four, as follows.

First: the word “data” entered the English language in the seventeenth century and was naturalized in the eighteenth. There are a number of different sources of evidence for this, and the evidence is unambiguous. The data derived from the ECCO database shows a substantial increase in usage of the term during the eighteenth century. The number of books in which the English word “data” appears rises from 34 in the first decade of the century to 885 in the last decade, and the number of books in which “data” appears rises relative to the total number of books included in ECCO for that decade, from 0.3 percent of the total in the first decade to 3 percent of the total in the last. While this tenfold increase in relative frequency did not make data a common word, it did make it familiar. At the beginning of the century, the term “data” was italicized in the vast majority—88 percent—of cases, an indication that the word was still considered a Latin loan. By the end of the century, “data” was italicized in only 19 percent of cases. These two trends strongly reinforce one another.

Second: the term “data” came into English in the early eighteenth century principally through discussions of mathematics and theology, roughly 70 percent of instances. At

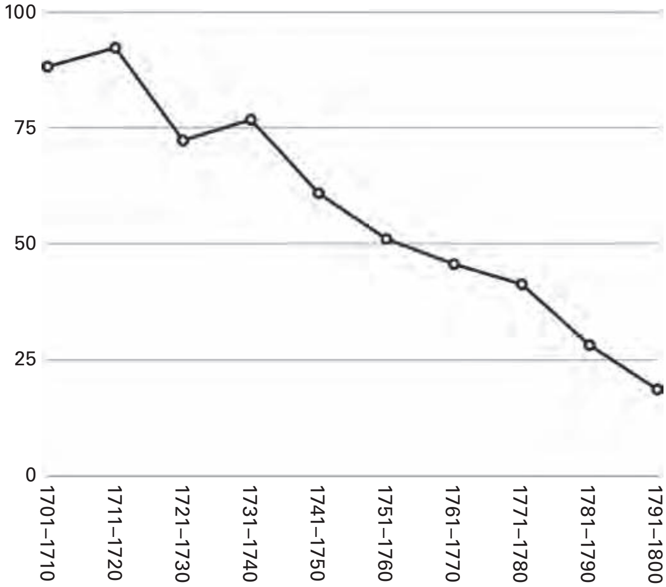


Figure 1.11 Percentage of instances of the English word “data” in ECCO I where term is italicized, 1701–1800.

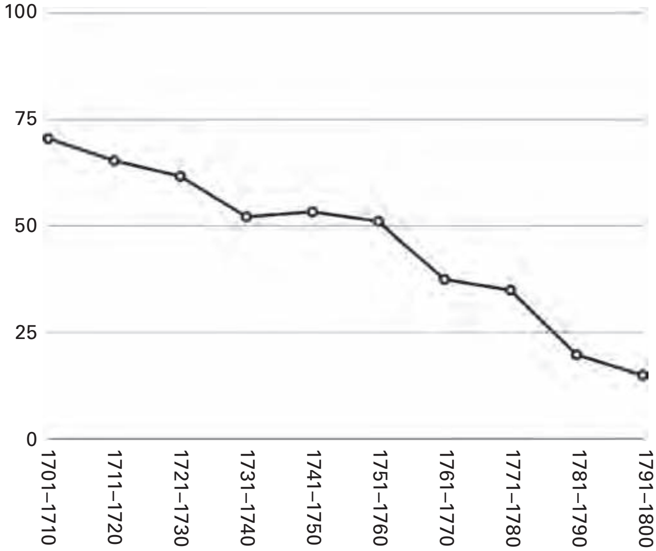


Figure 1.12 Image: Percentage of total "data" hits in English in ECCO I pertaining to Mathematics and Theology, 1701–1800.

century's end, mathematics and religion accounted for only about 20 percent of total instances, which were now dominated by empirical contexts such as those of medicine, finance, natural history, and geography.

Third: over the course of the eighteenth century, the main connotations of the term "data" shifted. At the beginning of the century, "data" was especially used to refer either to principles accepted as a basis of argument or to facts gleaned from scripture that were unavailable to questioning. By the end of the century, the term was most commonly used to refer to facts in evidence determined by experiment, experience, or collection. It had become usual to think of data as the result of an investigation rather than its premise. While this semantic inversion did not produce the twentieth-century meaning of data, it did make it possible. Still today we think of data as a premise for argument; however, our principal notion of data as information in numerical form relies on the late eighteenth-century development.

This, of course, raises an additional question. Seeing that "data" became much more commonly used during the eighteenth century, why did it take until the twentieth century for the term to become truly ubiquitous? It is clear that the fundamental semantic structure of the term "data" essential to the modern usage was settled by about 1750.

It appears, however, that while the newly outfitted term responded to and exemplified the epistemological perspective of the mid-eighteenth century, the term also was not fully required by it. Moreover, for all of the scientific achievements of the nineteenth century, the term "data" was still not of broad cultural importance. In effect, after its invention, the term went through a period of cultural latency. Though its usage expanded constantly within certain domains, throughout this period it played only a small role in the general culture. Ironically, this long period of latency may partly account for the great usefulness of the term in the twentieth century. In the twentieth century, when "data" reached its point of statistical takeoff, it was already a well-established concept, but it remained largely without connotative baggage. The arrival of computer technology and information theory gave new relevance to the base concept of data as established in the eighteenth century. At the same time, because the term was still relatively uncommon, it was adaptable to new associations.

Fourth: the *OED* is right and Google is wrong. Or at the very least, Google is not yet particularly helpful on this question. There are definitive quantifiable trends in both the currency and usage of the term "data" in the eighteenth century. It took some fairly heavy work with the ECCO data to make these trends visible, but having done it, it is clear that the *Oxford English Dictionary* account of the history of the term is mirrored in the quantitative results.

There are a number of reasons why raw Google Books results do not quite do the job for "data." First, Google Books is not yet very good or representative for periods before the nineteenth century. And even as Google Books advances, differences in the source base are still likely to pose thorny problems for quantitative comparison before the modern period. Lack of proximity search, wildcards, and other tools that aid such work as distinguishing Latin from English usages create challenges as well.

The difficulty in recognizing the true lexicographic issues in eighteenth-century English—regardless of the database one uses—is further heightened by the fact that the rise of the English-language usage of "data" during the eighteenth century coincides precisely with the decline in the general use of Latin in the Anglophone world. Without sorting, the raw numbers are highly ambiguous since the rise in the usage of "data" in English is largely offset by the decline in the use of Latin altogether. This effect is not strictly limited to the eighteenth century, but it is most significant in that transitional period.

The problem of investigating the history and semantics of "data" points to another considerable blind spot: unless search engines are full-featured, permitting good techniques of disambiguation such as proximity searching, common terms—arguably those terms we need most to understand well—may fall outside of the realm of practical investigation. Sometimes this happens by rule: for example, it is typical for search engines to exclude grammatical articles and Boolean operators from possible searches. In many cases, these restrictions virtually rule out the possibility of investigating the linguistics of conjunctions using database search functions. In other cases, blind spots of this sort are created by accident. It happens that the term "data" appears very frequently in metadata. To take one telling example: every work included in Project Gutenberg includes a legal disclaimer employing the term "data." For this reason, a simple search of Project Gutenberg to identify works in its corpus including the word "data" will produce results coextensive with the corpus itself. The online library catalog WorldCat produces another problem since it embeds the term "data" in the titles of many archival collections. None of these problems is insuperable. But none is certain to be fixed any time soon either.

It is worth adding that just because the *OED* is right and Google is wrong today doesn't mean that Google will continue to be wrong. If Google had good metadata, and if it allowed proximity searches and wildcards, we would be a long way toward being able to use it for a lot of quantitative humanities applications, whether or not one wishes to refer to these applications as "culturomic" and whether or not one regards such approaches as fundamentally new.

For the moment, it is a win for nineteenth-century reading practices, but it is not a success that is likely to stand for long. Even the venerable *OED* is moving to embrace a data-driven approach, which is as good a signal as any that we should all be ready to engage with quantitative humanities approaches in a strong, critical fashion. Among other things, as humanists, we need to pay much better attention to the epistemological implications of *search*, an entirely new and already dominant form of inquiry, a form with its own rules, and with its own notable blind spots both in design and use. In any event, I do think that my eventual results will be good news for reading even if they are not bad news for data. What is more, as we have seen with Priestley, the techniques made possible by the data-fication of our literature in many ways are consistent with ideas about ideas and writing native to the eighteenth century. In other words, at least in examining this corpus, there is a pleasing echo of the primary material, such as the charts of Priestley and Playfair, in the contemporary analytic techniques.

In the end, what does the history of the term "data" have to tell us about data today? There are a number of possible answers to this question, but one is worth particular

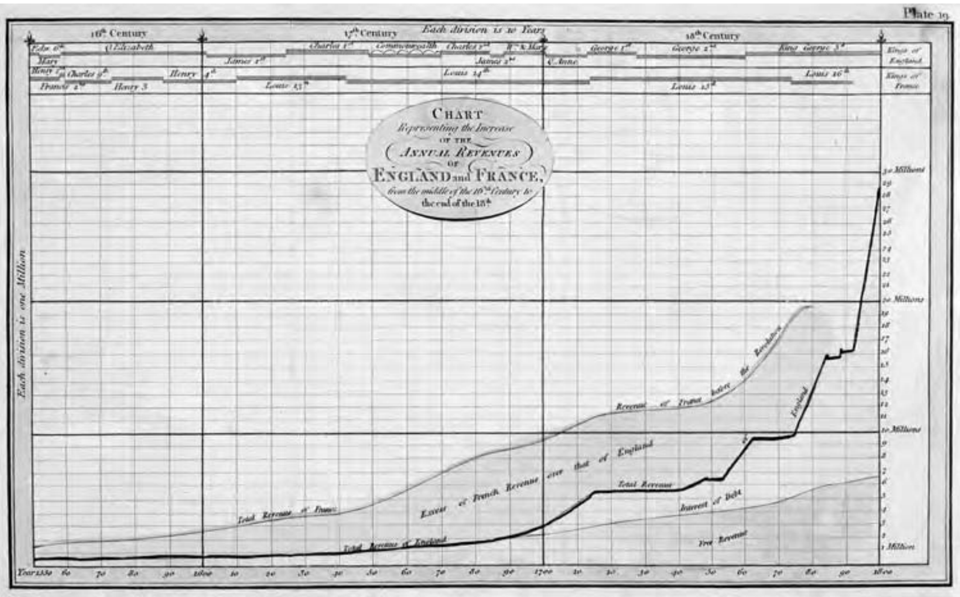


Figure 1.13 Line graph with timeline from William Playfair's *An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations*, 1805. Courtesy of the Library Company of Philadelphia.

attention. This observation is supported by the numbers but not generated by them: from the beginning, data was a rhetorical concept. Data means—and has meant for a very long time—that which is given prior to argument. As a consequence, the meaning of data must always shift with argumentative strategy and context—and with the history of both. The rise of modern economics and empirical natural science created new conditions of argument and new assumptions about facts and evidence. And the histories of those terms and others in the same family nicely illustrate the larger epistemological developments.

The history of data is connected to these other histories in very important ways, but in equally important ways, it remains an outlier. Curiously, the preexisting semantic structure of the term "data" made it especially flexible in these shifting epistemological and semantic contexts. Without changing meaning, during the eighteenth century data changed connotation. It went 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.

It is tempting to want to give data an essence, to define what exact kind of fact data is. But this misses the most important aspect of the term, and it obscures why the term became so useful in the mid-twentieth century. Data has no truth. Even today, when we speak of data, we make no assumptions at all about veracity. Electronic data, like the data of the early modern period, is given. It may be that the data we collect and transmit has no relation to truth or reality whatsoever beyond the reality that data helps us to construct. This fact is essential to our current usage. It was no less so in the early modern period; but in our age of communication, it is this rhetorical aspect of the term "data" that has made it indispensable.

Notes

1. Joseph Priestley, *Lectures on history, and general policy; to which is prefixed, an essay on a course of liberal education for civil and active life* (Dublin: 1788) 104. My thanks to McKenna Marsden and Dennis O'Connell for their invaluable assistance in this project.

2. Joseph Priestley, *A Chart of Biography* (London: J. Johnson, 1765).

3. Daniel Rosenberg, "Joseph Priestley and the Graphic Invention of Modern Time," *Studies in Eighteenth-Century Culture* 36 (Spring 2007): 55–104.

4. Joseph Priestley, *Experiments and Observations on Different Kinds of Air*, vol. 3 (London: 1777), 54; idem., *Discourses Relating to the Evidences of Revealed Religion*, vol. 3 (London: 1794–1799), 231; *An Essay on a Course of Liberal Education for Civil and Active Life* (London: 1765), 144.

5. Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven, CT: Yale University Press, 2010), 2; Lorraine Daston, "The Factual Sensibility," *Isis* 79, no. 3 (September 1988): 466; Theodore Porter, *The Rise of Statistical Thinking*, 1820–1900 (Princeton, NJ: Princeton University Press, 1986), 3.

6. Mary Poovey, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (Chicago: University of Chicago Press 1998), 1.

7. Examples from different domains: Walter E. Myers, "A Study of Usage Items," *College Composition and Communication* 23, no. 2 (May 1972): 155–169; Carter A. Daniel and Charles C. Smith, "An Argument for Data as a Collective Singular," *ACBA Bulletin* 45, no. 3 (September 1983): 31–33; Susan E. Bates and Edward J. Benz Jr., "Troublesome Words, Linguistic Precision and Medical Oncology," *The Oncologist* 14, no. 4 (April 2009): 445–447.

8. Nearly any complete style guide will include some discussion of the problem in general or in the particular case of "data." A classic discussion of foreign loan words in English is H. W. Fowler, *The King's English*, 2nd ed. (Oxford: Clarendon Press, 1908), 26–36. On the usage of "data" in contemporary English, see *American Heritage Dictionary of the English Language*, 3rd ed.

Technical literature on the subject includes the following: Chaim Zins, "Conceptual Approaches for Defining Data, Information, and Knowledge," *Journal of the American Society for Information Science and Technology* 58, no. 4 (2007): 479–493; Carter A. Daniel and Charles C. Smith, "An Argument for *Data* as a Collective Singular," *Business Communication Quarterly* 45, no. 3 (September 1982): 31–33; Walter E. Meyers, *College Composition and Communication* 23, no. 2 (May 1972): 155–169.

9. Steven Pinker, *Words and Rules: The Ingredients of Language* (New York: Basic Books, 1999), 178.

10. Oxford Dictionaries Online, "Datum," http://oxforddictionaries.com (accessed February 10, 2012). See also Geoffrey Nunberg, "Farewell to the Information Age," in *The Future of the Book*, ed. Geoffrey Nunberg (Berkeley: University of California Press, 1996), 103–138.

11. Pinker, *Words and Rules*, 55. The eighteenth-century usage question revolved mainly around the propriety of using foreign suffixes to create plurals for naturalized loan words. "Certain it is that *effuviums* and *phenomenons* are as good as *deliriums* and *lexicons*, that therefore *effuvia* and *phenomena* are no better than *deliria* and *lexica*; that *postulata* and *data* are as great *errata* or *arcana* as *peccata* or *viscera*, *regalia* or *paraphernalia*, and (if possible) more so than *postulatum* or *datum*. . . ." James Elphinston, *The Principles of the English Language Digested*, vol. 1 (London: 1765), 6. See also, John Wesley, *Dr. Free's Edition of the Rev. Mr. John Wesley's Second Letter* (London: 1759), 5; Paul Rapin de Thoyras, *The History of England*, vol. 2, trans. N. Tindal, 4th ed. (London: 1757), 5.

12. Henry Hammond, "A Brief Vindication of Three Passages in the Practical Catechisme," in *The Workes of the Reverend and Learned Henry Hammond* (London: 1674), 248.

13. Francis Bacon, *The Philosophical Works of Francis Bacon, Baron of Verulam, Viscount St. Albans, and Lord High-Chancellor of England*, vol. 2 (London: 1733), 397.

14. Isaac Newton, *The Method of Fluxions and Infinite Series* (London: 1736), xx.

15. In this vast literature, see, for example, Nicholson Baker, *The Double Fold: Libraries and the Assault on Paper* (New York: Vintage, 2002); Robert Darnton, *The Case for Books* (New York: PublicAffairs, 2009).

16. Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331 (2011), published online ahead of print: December 16, 2010; Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin Nowak, "Quantifying the Evolutionary Dynamics of Language," *Nature* 449 (2007). See also http://www.culturomics.org.

17. Patricia Cohen, "Analyzing Literature by Words and Numbers," *New York Times*, December 3, 2010; idem., "In 500 Billion Words, New Window on Culture," December 16, 2010; idem., "Five-Million-Book Google Database Gets a Workout, and a Debate, in Its First Days," December

21, 2010; idem.; Ben Zimmer, "The Future Tense," *New York Times*, February 25, 2011. Anna North, "New Google Graphs Reveal Centuries of Dicks, Pimps and Hos," December 17, 2010, http://jezebel.com/5714665/word-graphs-reveal-centuries-of-dicks-pimps-and-hos (accessed May 10, 2012); Alexis Madrigal, "Vampire vs. Zombie: Comparing Word Usage through Time," December 17, 2010, http://www.theatlantic.com/technology/archive/2010/12/vampire-vs-zombie-comparing-word-usage-through-time/68203 (accessed May 10, 2012); Dan Klein, "A Short History of Words: New Google Tool Reveals Relative Popularity of 'Shmuck,' 'Zionist,' Other Terms," December 17, 2010, http://www.tabletmag.com/scroll/53877/a-short-history-of-words/?utm_source=Tablet+Magazine+List&utm_campaign=fcdbed4176-12_20_2010&utm_medium=email (accessed May 10, 2012).

18. Michel et al., "Quantitative Analysis," 181–182.

19. Ibid., 178–179.

20. Geoffrey Nunberg, "Counting on Google Books," *Chronicle of Higher Education*, December 16, 2010, http://chronicle.com/article/Counting-on-Google-Books/125735 (accessed May 10, 2012); Anthony Grafton, "From the President," *AHA Perspective*, March 2010, http://www.historians.org/Perspectives/issues/2011/1103/1103pre1.cfm (accessed May 10, 2012).

21. Beyond our general critique of the notion of "raw data," the specifics in this case show that the data is not raw at all. In order to correct for anomalies in the larger Google Books corpus, the Ngram Viewer, in fact, operates only on a subset of the larger data, roughly five million of the fifteen million books digitized by Google by the end of 2010. This, of course, is not a small number. Though it is only one third of the works in Google books in 2010, Michel and Aiden estimate that it represents about 4 percent of all books ever published. Michel et al., "Quantitative Analysis," 176.

22. Advertisement for Eighteenth Century Collections Online, published in *Choice* 40, no. 9 (May 2003): 1525; *Library and Information Update* 2, no. 9 (September 2003): 10.

23. Happily, users have discovered the same thing and have begun successfully pressing Gale to allow them to have limited access to scanned text and even to correct it. In April 2011, Gale authorized the Text Creation Partnership at the University of Michigan to manually key and release 2,231 texts from ECCO: http://www.lib.umich.edu/tcp/ecco/description.html (accessed May 10, 2012). Related efforts at rescanning and crowdsourced keying and correction are being organized by 18th Connect: see http://www.18thconnect.org.

24. To control for the variation in the size of the ECCO corpus from decade to decade in the eighteenth century, I divided the number of hits for "data" by the number of hits for a common control word. Experiments with several different control words suggested that using "the" as control produced a stable result.

25. The challenge of interpreting numbers such as these is highlighted by the very different results produced by a simple word search in Google compared to a parallel search in Google

Books. On the day that I composed this text, a simple search in Google for "facts" resulted in 163,000,000 hits, while "data" produced 1,160,000,000, seven times as many "data" as "facts." A Google Books search on the same day produced the inverse ratio, 166,000,000 "facts" and 28,500,000 "data," that is, six times as many "facts" as "data."

[26] Antonio de Herrera y Tordesillas, *The General History of the Vast Continent and Islands of America*, vol. 3, 2nd ed. (London: 1740), 119. For more on the difficulties presented by ECCO, see Patrick Spedding, "'The New Machine': Discovering the Limits of ECCO," *Eighteenth-Century Studies* 44, no. 4 (Summer 2011): 437–453.