The Great "Data" Kerfuffle, Resolved At Last!

Geoff Nunberg
School of Information, UC Berkeley
Draft, October 12, 2015

A version of this paper was presented as a topical lecture at the Annual Meeting of the American Association for the Advancement of Science in San Jose, February 14, 2015

When we speak of data we invariably speak in what the geologist and historian of science Scott Montgomery calls "the scientific voice." No other word—not *hypothesis*, not *theory*, not *empirical*—evokes the scientific mindset so definitively. You can't write the history of the word without recapitulating the history of modern science, writ small, and as we enter the era of Big Data, it is becoming a proxy for the future of science, as well.

When it comes to usage, though, it's fair to say that no other word in the scientific lexicon is the subject of such unscientific squabbling. Should *data* take a singular or plural verb? Researchers, editors, journalists— it's hard to find anyone in the world of science who doesn't have a decided opinion about that question, invariably shaped by pedantry, ideology or folklore, and unruffled by scientific curiosity. People feel no need to cite empirical evidence about the use of the word itself or to consider the more general linguistic phenomenon that it exemplifies. We feel free to opine about *data* without having to appeal to the very stuff it names.

This is not a new debate. People often talk about a "growing tendency" to use *data* as a singular noun. But perceptions of changing usage are subject to a well-documented recency illusion. "I literally blazed with wit" may strike us as a typically contemporary solecism, but the speaker was Thackeray. And it turns out that people have been saying things like "the data is" and "this data" since the nineteenth century, and critics have been grumbling about those constructions for almost as long. In a note in a 1907 number of the *American Machinist*, the editor of a scientific review complained that use of *data* as a singular had become so common in technical papers that "it is apparent that proof-readers as well as engineers need some instruction in the proper use of the language." By 1927 the dispute had reached the letters section of *Science*. The surgeon W. W. Keen, the editor of Gray's Anatomy, wrote that:

> I have more than once publicly protested against that abomination "data is." We say "phenomenon is" and "phenomena are," and I do not recall in Latin any singular verb used in English with a plural noun, excepting poor "data is."

In response, a correspondent named Charles Blake argued that "We speak and hence write English by ear and not by rules of grammar. If 'this data' sounds better than "these data" it will be used."

Making allowances, either of those remarks could have been pulled off the web yesterday. This seems to be just another one of those eternal squabbles over usage, framed as a clash between incommensurable grammatical dogmas (or dogmata, if you wish). On the one side are the *data* pluralists, who insist on paying obeisance to the origin of the word. On the other are the *data* singularians, who insist that English is not Latin and that ordinary usage trumps etymology. And so it goes. The only point that's clear is that nobody has had anything new or surprising to say about this subject since the time when Calvin Coolidge was walking the earth.

Once a principle becomes a dogma, it is impervious to the force of reason. Inflexible pluralism leads to inconsistencies and clashes that range from embarrassing to comical. The American Psychological Association style guide insists that *data* can only be plural. But one doesn't have to leave the pages of that very manual to find examples like: "Tables are efficient, enabling the researcher to present a large amount of data in a small amount of space." (Note that "a large amount of" can be used only with a singular mass noun: we might say "they sell a large amount of footwear" but not, "They sell a large amount of shoes.") Sometimes, in fact, one runs into sentences that seem to treat *data* as both a singular and a plural at the same time. A notice in the 1895 report of the New Jersey Weather Service read, "During the year much data have been furnished to the various departments of the State Experiment Station." That "much data have" could only have been produced by somebody who had been instructed that *data* had to take a plural verb but didn't think to change *much* to *many* in the process—an indication that the singular usage was both common and controversial even then.

At the limit, one finds people treating *data* as a plural where logic cries out for the singular. The *Economist* stylebook insists on the plural in all instances, which is no doubt what led to the appearance of the sentence, "Yet even as big data are helping banks, they are also throwing up new competitors from outside the industry." That sentence is the sort of thing that happens to a writer would rather risk appearing foolish than being considered incorrect. If one really thought of "Big Data" as a plural then it would have to refer to a collection of large things. (When we talk about "big elephants," after all, we're referring to a group of elephants that are individually large, not a large group of elephants.) But it makes no sense to speak of individual data (or data points) in this way. The datum $n > 3$ is neither bigger nor smaller than the datum $n = 3.141592653589793\ldots$

Those examples underscore the perils of undeviating grammatical dogmatism. They're the linguistic equivalents of the figure sometimes called the devil's fork, which seems to have three prongs on one end and four on the other: you can't get from the beginning to the end without doing a little mental stutter-step on the way. And even if you want to defend the use of the word in the plural in many contexts, as I certainly do, the plural fetishists don't make that job any easier. Some of these people are animated by a strain of pure pedantry—what Samuel Johnson once defined as "the unseasonable ostentation of learning." These are people who pride themselves as knowing that *data* comes from a Latin plural—or anyway, who wouldn't want anybody else to think they *don't* know that *data* comes from a Latin plural. There's some irony here in that those who actually read the language would know that Latin *data* didn't mean remotely what the word does for us. In fact when the word was first adopted into English in the early seventeenth century, it was chiefly as a term for the incontestable truths of mathematics and theology, which is the exact opposite of what we take data to be.

But there are also singularian absolutists who categorically condemn the plural usage. One sees the use of the plural described as "pretentious and outmoded," or as "a deliberate archaism." The writer John August compares it to over-pronouncing Italian at the Olive Garden. "No one is impressed," he says, "and frankly, we're just a little embarrassed for you." The columnist Kevin Drum has been waging a campaign against plural *data* for some time:

> I won't rest until [everyone] accepts the plain fact that data should be treated as a singular noun in all circumstances.…I'm not sure I've ever heard anyone say "data are," but lots of diehards with PhDs still use it in print.

This is dogma, too, though it's populist rather than pedantic. Where is it written that one shouldn't use any construction in formal writing that would sound unnatural in informal conversation? (If that were the case, one could never use "one" as a personal pronoun.) Using *data* in the plural in the pages of *Nature* or *Science* isn't like speaking Italian at Olive Garden, it's like speaking Italian at Piperno's in Rome.

As with other ideological schisms, each of the parties has its territorial strongholds. The singular has an edge in most of the informal precincts of English, like fiction, broadcast, and popular newspapers and magazines, though there's a lot of variation—not surprisingly, the plural is more common on PBS and in the *Economist* than on CBS and in *Rolling Stone*. The plural is dominant in academic and scientific publications:

Genre	per mil words	ratio pl/sg
fiction	37.9	0.46
broadcast	42.93	0.44
newspaper	158.84	1.05
magazine	99.6	0.38
academic	673.32	4.55

Frequency of *data* and ratio of "these data" to "this data" by genre:
BYU Corpus of Contemporary English

And within the academic domain, the plural tends to be most frequent in the hard sciences and least frequent in the discursive social sciences and the humanities:

	ratio pl/sg
Jrnl. Cell Biology	14.48
Am. Educ. Research	6.21
Language	3.41
Am. Hist. Review	1.38
ACM	0.39
Representations	0.03

Ratio of "these data" to "this data" in selected journals

In most fields, the frequency of the use of *data* in the plural correlates with the frequency of the word itself. In the pages of *Science*, for example, where *data* is naturally quite frequent, the plural wins out by better than twenty to one, depending on how you do the search. (The ratio of "the data are" to "the data is" is much higher than the ratio of "many data" to "much data," and the plural is more frequently used when the verb is in the past than in the present.)

There is one important exception to this generalization. In the *Communications of the ACM*, the standard organ for computer science, the word *data* is both very frequent and overwhelmingly in the singular, as it generally is when it denotes digital signals or content. That digital use of the word is increasingly common, which is one reason why searches on Google and in news databases seem to show that the singular has been gaining on the plural over the last 30 years or so. In fact the digital uses of the noun are probably the only ones that belong to everybody's active vocabulary. A 2015 Super Bowl ad for Verizon that featured Kim Kardashian lamenting the waste of unused cell data ended with "It's your data. Keep it." What else could they have said? "They're your data. Keep them" would have left 120 million other Americans scratching their heads.

Why do scientists use the plural more often than others do? The style guides have a lot to do with it, and so does a strain of pedantry that many scientists and engineers are naturally susceptible to. After all, "sticklerism" is just the name we give to a variety of OCD that serves scientists well when they're preparing slides or debugging programs. But most scientists aren't really pedants; for them, using the plural is more like a club handshake than a rosette in the lapel. Nobody grows up using *data* as a plural—indeed, few of us grew up using *data* at all. Most science-minded people first picked up the plural rule about the same time they began actively using the word itself in college STEM classes, at about the same time humanities majors were learning to say "interrogate" instead of "consider." When I ask Berkeley undergraduate STEM majors who first told them that *data* should be plural, they almost invariably point to their graduate student TA's: it's a nugget of linguistic lore that's passed along to each new class of novitiates as a badge of membership in the great confraternity of science.

Pedantry, deference to authority, collegiality—all of those factors play a role in explaining why scientists use the plural more often than the general public does. But even taken together, they don't fully account for scientists' partiality for the plural or their occasional insistence on the singular. Nor is the choice simply a matter of "personal preference," as some usage writers suggest. In fact, I'll show here that scientists' usage of *data* has a more rational explanation. *Data* has slightly different meanings in the singular and in the plural, which scientists are unconsciously sensitive to: they choose one or the other form on a case-by-case basis according to what the writer is trying to say.

That hypothesis can be tested empirically. If scientific use of the plural were chiefly due to its symbolic importance or to grammatical prescription, then we wouldn't expect to see any systematic semantic difference between the contexts favoring one or the other usage. But if the choice of the singular or plural varies with the linguistic surroundings—if instances of "the data are" is more frequent preceding some types of adjectives than others, say—then we'll conclude that it is motivated at least in part by considerations of meaning. From that it would follow that the reason why scientists treat *data* as a plural more often than others do isn't that they tend to think differently about the grammar of the noun, but that they tend to think differently about the notion of data itself.

The difference between saying "these data" and "this data" is the difference between treating *data* as a plural count noun like *pebbles* and *peas* and treating it as a singular mass noun like *gravel* and *succotash*. (That's why it makes no sense to compare *data* to other Latin plurals, such as *criteria*. "This criteria…" is a mistake for "this criterion," but "this data" isn't a mistaken way of saying "this datum.") It's an important distinction. The singular-plural kerfuffle may seem

to be just a tempest in a teacup, but it's a teacup that's bobbing on the surface of an ocean of research and theory—in linguistics, in philosophy and logic, in psycholinguistics and in developmental psychology. The majority of human languages have a mass-count distinction or something like it, some way of categorizing nouns according how speakers individuate the elements they name, and the distinction provides an ideal point of entry for exploring basic questions about the relations between language, cognition and the world.

In English, the count-mass distinction determines, among other things, whether a noun takes a singular or plural verb ("the gravel is"; "the pebbles are") and how the noun is measured—we say "too many pebbles" but "too much gravel." Count nouns take the indefinite article, but mass nouns don't—we say "She threw a shoe at him," but not "she threw a footwear at him," since *footwear* is a mass noun. (The singular *data* is sometimes loosely described as a "collective noun," but it little in common with true collectives like *herd*—we don't say "The rancher has much herd" and we can speak of several herds, for example.)

To some extent, the distinction between count and mass nouns seems to be arbitrary. Words can be close in meaning but fall into different categories: in English, we say *many shoes* but *much footwear*, *many leaves* but much *foliage*, *many sequins* but *much glitter*. And other languages often classify nouns differently from the way we do. The words for lightning, dandruff and furniture are mass in English but count in Italian, the words for contents and people are count in English but mass in Italian.

But it's also clear that that a large part of the mass-count alternation is motivated. There's a reason why *dog* is a count noun and *water* is mass noun; as the philosopher Jeffrey Pelletier has put it, it goes to "the metaphysical question of the primary existence of gunk vs. things." The difference is obvious in principle: it's easier to discern the individual constituents of a pack of dogs than of the water in a lake. But not all categories are as easy to classify as those. The distinguishability of the individual elements may depend on any of several factors, which in turn determine whether the name of the category will be count or mass. How large, how perceptible or how contiguous are the elements, for example? The linguist Arnold Zwicky points out that we tend to use mass terms for ground cover plants like ice plant, since it's hard to tell where one leaves off and the next begins. The names of plants used in hedges such box and privet are mass nouns, too; they may be easily separable as individuals, but they're usually planted so as to blend together when seen from a distance. But *petunia* and *foxglove* are count nouns because you can distinguish the plants from afar. And *clover* can go either way depending on whether we're planting a field or looking for one with four leaves. Similarly, the small stones that wholly cover

the surface of a path count as gravel, but when they're spread out with spaces between them they turn into pebbles.

Which kind of noun we use for a type of thing can also depend on whether we interact with the elements one-by-one or in quantity. That principle explains how there can be mass and count terms that designate the same things. We can construe the greenery on a tree either as an ensemble of indistinguishable entities, in which case we use the mass noun *foliage*, or as a collection of distinct things, in which case we use the count plural *leaves*. Gazing at our yard in October we can admire either the lovely autumn foliage or the lovely autumn leaves. But when November comes and we have to interact with the individual elements on our lawn it's the leaves we rake, not the foliage. The linguist Anna Wierzbicka suggests that sequins are count because they're sewn on one-by-one whereas glitter is mass because it's sprinkled or brushed on surfaces. And then there's *chad,* which was chiefly used as a mass noun before November of 2000, when people scooped it up by the handful, but was converted to a count noun as the nation watched ballot inspectors in Florida examining those little bits of paper one-by-one to decide whether they were hanging or bulging or dimpled.

What of *data*? Unlike pebbles and petunias, data is immaterial (I'll keep the word singular here for the sake of consistency). And data is a cultural construct. Not every systematic collection of facts automatically counts as data. A huge proportion of the discourse of sports consists of discussions of quantitative observations about performance. But even in the post-Moneyball age, people rarely refer to those facts as data. They're stats. To count as data, an observation has to be the product of certain kinds of organized institutions that are subject to standardized procedures, like laboratories, bureaucracies, insurance companies.

In fact it's only relative to those procedures that *data* is reckoned as count or mass. Usage writers sometimes offer broad rules of thumb for making the distinction. The *Random House Learner's Dictionary* advises that one should treat *data* as a mass noun when speaking of "a body or collection of facts" and as a count noun when speaking of "individual facts [or] statistics." The dictionary illustrates that difference by contrasting "do your data support your conclusions?" with "The data is inconclusive." But that counsel isn't very helpful by itself. What makes a set of observations a body of facts? What is it about the verb *support* that militates for treating it as a count noun or about *inconclusive* that militates for treating it as mass? (In actual scientific writing, as it happens, the tendency goes the other way: *inconclusive* is almost twice as likely to favor the plural usage as *support* is.)

As with leaves and foliage, what determines how we think of data is how we interact with it. In scientific practice, there are four kinds of things we do with data. We collect or generate it,

we process it, we analyze it, and we evaluate it or appraise its significance. These categories define a rough continuum. At the collection end of the scale we're interacting with the data individually in a way that's analogous to sewing sequins on a dress or counting chads, and we'd expect that these would be situations that favor the use of the plural. It's true that we don't actually individuate data as such—*data* is one of those aggregate plural nouns like *minutiae* and *particulars*, which have no common singular form. (Other than in surveying contexts, the singular noun *datum* is extremely rare, even in scientific writing; rather we talk about data points and pieces of data, with the idea that we can chunk data into significant bits relative to a particular goal or project.) But we can observe or interact with those chunks of data individually, when we're looking at the observations in terms of their source, their individual reliability, their scarcity or abundance, and so on. And in that case we'd predict that we'll tend to see *data* in the plural when it occurs with verbs and adjectives that are relevant to these sorts of activities.

This prediction is borne out across several corpora. Here are the results of searches in Google Scholar for "the data are" and "the data is" followed by various adjectives and participles. These are drawn from papers in the sciences: the higher the ratio, the more frequent the use of the plural. (The results were filtered to eliminate most strings in which *data* is not the subject of the verb, such as "The scarcity of risk estimates in the data is surprising" and "Listening to the data is important.")

	pl/sg
expressed as	15.70
tabulated	8.54
cross-sectional	7.31
insufficient	5.88
self-reported	4.11
scarce	3.65
precise	3.17
anecdotal	2.86
sketchy	2.85

Ratio of "data are" to "data is" preceding adjectives in articles in Google Scholar

The plural is also dominant when we look at expressions like "consistent with" or "fit," which are notions we invoke when we're trying to pair our specific observations with a predictive model or hypothesis:

	pl/sg
consistent with	11.60
derived	6.24
fit	4.52
compatible with	3.33
analyzed	2.38

Note that these expressions are largely particular to scientific discourse. When we say "x is consistent with y," we're expressing a concept that is only operative within the broad framework of scientific reasoning. Since terms like these overwhelmingly favor the plural, they contribute to the greater frequency of the plural in scientific writing. Scientists, that is, tend to think about data differently.

Now compare these proportions to the ones we see with modifiers like *interesting, key, crucial*, and *vital*, which involve evaluating the data for its wider implications. In these cases, there's a decided preference for the singular:

	pl/sg
overwhelming	1.14
crucial	0.74
vital	0.72
useless	0.68
key	0.47
remarkable	0.41

When you say "the data is key to identifying opportunities " you presume the validity of both the data and the conclusions drawn from the data, and go on to make an extra-scientific judgment about it. There's a similar pattern with the words we use to describe the reliability or implications of conclusions drawn from data, particularly in reference to the scientific literature in general. Here again, the singular is favored:

	pl/sg
says that	1.4
doesn't lie	1
means that	0.64

When you assert that the data says such-and-such or that it means such-and-such, you're really just using *data* as a cover term for scientific results, in something like the way journalists use "studies show":

The data says that with the poor, a little money can buy a lot of happiness.

I'm not saying Obamacare is perfect but the data doesn't lie.

Finally I might mention one other set of terms, which apply predominantly to digital data, where the singular again dominates:

	pl/sg
structured	0.59
secure	0.28
encrypted	0.17
safe	0.16

The preference for the mass conception of digital data follows from the same principles that militate for mass terms in other contexts. The elements that make up such data are essentially indistinguishable and can't be broken down into semantic units—you can't watch the data you're downloading to your cell and say, oh, there's Kim's outfit, and there's her vacation, and there's her backhand. In fact since the introduction of packet-switching half a century ago, the whole point of the digital project has been to turn the word *data* into an irreducible mass term—from the phenomenal point of view, it's just undifferentiated gunk.

The figures I've been citing are drawn from scientific publications, of course. If we look at the same terms in news stories, the plural will be less frequently used in all cases, sometimes dramatically. Here are the ratios of plural to singular use in Google News and Google Scholar.

	Google News	**Google Scholar**
says that	0.23	1.4
doesn't lie	0.26	1
surprising	0.36	1.64
disproves	0.36	4.1
overwhelming	0.51	1.14
confirm	0.64	4.87

Ratio of "data are" to "data is" preceding adjectives in articles in Google Scholar and Google News

In part, these discrepancies reflect the grammatical fastidiousness of scientists and the role of editors and style books. But they also reflect differences in the way the words are being used. In

scientific articles, “the data disprove” and “the data confirm” are apt to be followed by something like “…the existence of a separate valise-α-ketoglutarate transaminase within the operon,” where *data* refers to the specific observations made in a study, which disprove or confirm a specific scientific hypothesis. In a news article, we’re more likely to run into “the data disproves the assertions of anti-gun advocates,” where *data* refers to an often unspecified body of research that is held to disconfirm an unspecified set of assertions. The fact is that people engaged in everyday discourse talk and think about data differently from way the scientists do.

These observations tell us several things. First, even within scientific discourse, people’s use of *data* as a singular or plural noun isn’t determined primarily by the rulings of style guides or their beliefs about what’s “correct.” Those considerations certainly play a role, but what matters most is the conception of data that’s relevant in the particular context. The fact that scientists are five times more likely to use the plural when *data* is followed by *tabulated* as when it’s followed by *surprising* shows that they are discerning in the way they treat the noun grammatically. And if scientists use *data* in the plural more than others do, it’s in large part because they interact with data more immediately and because the operations they perform on data impose a granular conception on it. That granularity is usually lost when the results of those operations are handed over to the public—in the wider public discourse, *data* is often just a term for the conclusions drawn from investigations or surveys.

None of this means that these differences in conceptualization are only available to speakers of English. They’re no less important to speakers of languages like Italian or German, where the word for *data* can only be plural. The fact that Language A explicitly marks a distinction that language B lacks doesn’t mean the speakers of B can’t perceive it. Unlike the French, we English-speakers don’t have different words for the rivers that flow into other rivers and the rivers that disgorge into the sea, but that distinction isn’t lost on us, even so. Still, there’s a considerable body of psycholinguistic research showing that people attend more closely to the distinctions that their language makes it easier to communicate, particularly when they’re obligatory. The psycholinguist Dan Slobin calls this phenomenon “thinking for speaking.” If your language requires you to choose a different numeral phrase depending on whether things you’re counting are hard or soft, you’re apt to foreground that feature when you’re asked to group objects according to their similarity.

English-speakers have to make a similar distinction when deciding whether to treat *data* as a singular or plural. As we’ve seen, even scientists tend to make that choice according to how they’re conceiving of data in the context. Not that they stop to think which form to use when they’re taking about the data’s consistency with a model, no more than they stop to consider

whether to describe a collection of small stones as pebbles or gravel. But they make the choice on rational grounds even so. And in the course of acquiring the intuitions that underlie the distinction, young scientists are also learning how to think about data.

True, you can relieve writers of the responsibility for having to make a decision by reducing the choice of verb form to a categorical rule. In that case you privilege mechanical conformity at the expense of discernment and reason—it's to speak, as it were, in the unscientific voice. There are sure to be those who counsel that we stick with the plural only to avoid rousing the ire of the sticklers, however unreasonable they are—that we yield to what I've called the pedant's veto. But apart from being intellectually indefensible, that course inevitably leads to embarrassing inconsistencies and incongruities and deprives the English language of a subtle and useful distinction. And we can't have too many of those.

GEOFFREY NUNBERG is a linguist who teaches at the UC Berkeley School of Information. His commentaries on language have appeared in publications such as the *New York Times*, *The Washington Post*, and *The Atlantic* and are a regular feature on the NPR program Fresh Air. He is also the emeritus chair of the Usage Panel of the American Heritage Dictionary.