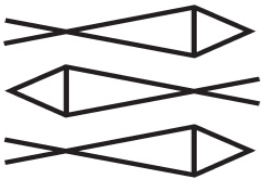
LIVING IN DATA

A CITIZEN'S GUIDE TO A BETTER INFORMATION FUTURE

JER THORP



2. I Data You, You Data Me (We All Data Together)

Open the window and let the words in. Let them flow into the room in a stream, all of the words, hundreds of thousands of them, let them fill the space, let them hang in the air, tiny sparkling motes of language. Let them drift, and organize. Let them be carried by eddies of usage and syntax until they find a place to rest. When they've settled, when they've coated every surface like ash, get up and walk the room, see each word and its neighbors, read the room like a map.

Over by the door you see a cluster of fruit words: "pear" and "apple" and "orange" and "kumquat." "Strawberry" and "yuzu" and "lychee" and "banana." Above your head, on the ceiling, are "hope" and "dream" and "imagination," "whimsy" and "fancy" and "caprice." Right under your chair you find all of the synonyms for "love." Proper nouns are here, too, in dense clusters. You find the presidents over near the window, and just beside them all the countries and states and provinces and cities.

Each word has found its particular place in the room by seeking out those with which it has the most affinity: the words with which it has been most often written. Near "business," you spot "money," of course, and "bank" and "Warren Buffett" and "the Federal Reserve." Further out, there's "power" and "politics" and "economics," "number" and then "math" and then "Newton" and "Poincaré." Further still and you're reading words that are almost never written in the same

sentence as “business”: “Cretaceous” and “alabaster” and “millipede” and “orgasm.”

Data’s closest neighbors in this room are “statistics,” “information,” and “database.” “Indicators,” “analysis,” “metrics,” “graphs,” “analytics,” “measurement,” “surveys.” “Terabytes,” “files,” “metadata,” “end points.” It’s in a clean neighborhood, with well-groomed streets and good schools. A suburb of science, just a short train ride from “truth,” “fact,” and “evidence.” Data lives on this map of words comfortably apart from “human,” and even further still from “art” or “dirt” or “laughter.”

Sit now, and watch the words.

It will seem, if you wait for only a minute, that nothing is happening. Even if you sit for an hour, a week, a month, you might not notice anything. But if you were to find a more patient eye, to find a way to sit and watch for years and decades, you would see that there is much activity. Each time it is written, used in a sentence somewhere in a newspaper, in an essay or a novel or a textbook, a word shifts in space. Fast-forward over centuries and our room is alive. Words bustle and budge with each other within their syntactic clusters; they dart from one side of the room to another. Whole new groups of words appear. Systems of meaning collapse like sandpiles.

“Data” has always been a restless word. It first appeared in the English language on loan from Latin, where it meant “a thing given, a gift delivered or sent.” It spent its early years in the shared custody of theology and mathematics. The clergyman Thomas Tuke wrote this in 1614 about the difference between mystery and sacrament: “Every Sacrament is a Mysterie, but every Mysterie is not a Sacrament. Sacraments are not Nata, but Data, Not Naturall but by Divine appointment.” Here “data” holds its Latin meaning as something given, but because its giver is the almighty God, it carries with it a particular strength of truth. In 1645, the Scottish polymath Thomas Urquhart wrote *The Trissotetras; or, A Most Exquisite Table for Resolving All Manner of Triangles*. In it, he defined “data” as “the parts of the triangle which are given to us.” By 1704, data had found a hold in mathematics beyond geometry. Another clergyman, John Harris, defined “data” in his *Lexicon Technicum* as follows: “such

things or quantities as are supposed to be given or known, in order to find out thereby other things or quantities which are unknown." Data as givens, things we already know, truths like gravity and pi and the Holy Ghost.

The linguistic neighbors of "data" remained, for a century or two, consistent. "Math," "numbers," "quantities," "evidence," "unknowns." Some new words arrived as mathematicians and philosophers worked to order their universe: "qualitative," "quantitative," "ordinal," "cardinal," "ratio." At the turn of the twentieth century, with Galton and Pearson and the birth of modern statistics, came a new way for data to be thought of, and a new way for it to live: as the contents of a table. Fifty years after that, data became bound to one of its stalwart allies, a word that would change the way in which data is commonly understood: "computer." Between 1970 and the end of the millennium, data changed quickly: from a thing of God and mathematics to a collection of bits and bytes. The word still adhered firmly to concepts of truth, but it was a different kind of veracity, one stamped into thin layers of silicon.

In the twentieth century, the tech industry and its marketing departments drove a lot of changes in our linguistic landscape. In 1964, Doug Engelbart put two orthogonal wheels in a wooden box. Two decades later "mouse" had crawled away from "rat" and "shrew" and "hamster," toward "keyboard" and "joystick" and "terminal." In 1989, Tim Berners-Lee wrote a set of programs on his NeXT computer in Geneva, Switzerland, and set a whole cadre of words on the move: "browse," "web," "page," "surf," "highway." "Link," "refresh," "bookmark." Starting in the 1990s, Silicon Valley went about its own program of linguistic disruption, with "Google" and "timeline" and "tweet" and "snap" and "swipe."

In the last decade, the way in which we collectively define "data" is undergoing perhaps its most dramatic change. "Data," born of God and raised by computers, has found its way to the mess of human lives. It's there now with "social" and "genetic" and "sentiment," with "migrant" and "gender" and "identity." As "data" settles in with its new neighbors, it is changing the way we think of each of them.

★ ★ ★

The room of words we're in is imaginary, but the map of meaning that defines it is real. I made it by taking a corpus of text gathered from Google News and processing it with a program that calculates word vectors. What this program does is look at the position of every word in every sentence and keep a running tally of the relationships between them. Each word gets a position—a vector—in relation to each other word. This means that every word ever used in a Google News story gets a position in relation to every other word, "cat" or "zeitgeist" or "religion." For words that often appear close to "religion"—"God" or "church" or "pew"—this position will be close to zero. For words that almost never sit in the same sentence with "religion"—"squid" or "pappardelle"—this number will be close to one. The number of vectors in the map that I'm using is huge—remember that every word gets a position in relation to every other word. Think of any strange combination of words you can—"rhizome" and "whiskey," "harness" and "Bob Dole," "zeitgeist" and "origami"—and there's probably a vector that's been calculated. The corpus from Google News contained roughly a hundred billion words, with a total vocabulary of three hundred million words. Out of this come nearly a billion vectors.

So far we've been content to use this map of words to investigate proximity: how close one word is to another, and how this changes over time. A word map this vast and multidimensional allows us to do some other things, though, that speak to us about the ways that language is interconnected, and also about how the data set was created and assembled. In particular, we can map the unique relation between a pair of words and a different pair of words. This is done using a neural network. The network is trained on the relations that exist between all of the words in the vocabulary (in this case three hundred million) and then can be asked to conjure new links. These "fill in the blanks" connections are fun to explore:

"Air" is to "bird" as "water" is to "fish."
"Joy" is to "child" as "thrill" is to "adult."

The example that was offered by the Google engineers who produced one popular word-vectorization software package called word2vec is this: What word is connected to “woman” in the same way that “king” is to “man”? The trained neural network dutifully offers up an answer: “queen.”

In 2016, Tolga Bolukbasi, then a machine learning student, exposed troubling gender bias in word2vec’s output. When queried, for example, as to what word is connected to “woman” in the same way that “doctor” is to “man,” the system answers “nurse.” When asked about “computer programmer” in the same context, word2vec offers up “homemaker.” There are other word pairs that show extreme female-male gender differences in word2vec: “sewing” and “carpentry,” “hairdresser” and “barber,” “interior designer” and “architect,” “diva” and “superstar,” “giggle” and “chuckle.” Gendered relations are evident even indirectly; “receptionist” is closer to “softball” than it is to “football.”

After reading Bolukbasi’s paper, the artist and researcher Matthew Kenney investigated similar biases around race. “If we think about historically where algorithms have had the most dire impact,” Kenney told me, “it’s really around marginalized groups.” Kenney started digging into “word embedding” models like word2vec, with a particular focus on how biases in the model might affect what he calls “downstream classification tasks.” How might racial biases in these models trickle down to processes in which Black people are being considered for jobs? Or for home loans? Kenney found that word2vec, trained on a corpus of news articles, positioned the word “black” more closely to “criminal” than it did “white.” Digging deeper, he tested a variety of given names against the word “criminal.” The word2vec placed stereotypically Black names, such as Darnell and DeShawn, closer to “criminal” than it did Mike, Conner, Jake, or Brad.

Where does word2vec’s racial and gender bias lie exactly? In the words of the news stories written by journalists? In the selection of a subset of those stories from particular publications by Google News? In the vectorization of language by word2vec, or in its neural-net learning of linguistic relationships? There’s a layer cake of decisions here, and each one likely carries at least a part of the responsibility

for binding “nurse” so closely to “woman,” and “black” so closely to “criminal.”

There’s an instinct perhaps to forgive word2vec in this case: Isn’t it after all simply mirroring biases that exist in the English language and in culture as a whole? To understand where the real danger lies, we need to consider why word2vec exists. It is not a tool that is intended to be used only to create a map of language, a playground for the linguistically curious. It’s a tool that is built to be able to make decisions—specifically classifications based on language. In a talk in 2018 in Minneapolis, Meredith Whittaker posited an example very close to what Kenney had been imagining: where a developer at a large corporation uses word2vec inside a piece of software she’s written to process HR applications. The software goes on to be used company-wide, and thus word2vec’s peculiar and problematic biases are now part of hiring practices, playing a role in who gets hired and who doesn’t, or perhaps more important who is even offered an interview.

As it turns out, Whittaker’s HR analogy was remarkably prescient. A software system developed internally for Amazon over the course of four years was scrapped in October 2018 when it was shown to be dramatically biased against women. The system rated résumés lower if they contained the word “women’s” and if they listed all-female colleges and higher if they used words that have been shown to be more common in male résumés, such as “executed” and “captured.” As if to demonstrate the absurdity of these systems’ flaws, researchers studying another algorithmic HR system uncovered the two things that were most likely to flag a candidate for an interview: if their first name was Jared and if their résumé contained the word “lacrosse.”

This issue with tools like word2vec seems a modern problem, but the roots of it are set into the soil of the seventeenth century, when “data” drifted into English from Latin. We are still stuck with the idea that data is given to us, if not from God, from somewhere similarly divine. In the case of word2vec, there seems to be some common belief that we can use it to investigate language itself, rather than a

very particular model of language, gathered by particular humans working for a particular company, in a particular culture and time.

In the research paper that Bolukbasi and his colleagues published, they suggested there might be two methods for "debiasing" a model like word2vec. They called the first method "hard debiasing": a scorched-earth tactic where words that are known to contain bias would be removed from the model. They also proposed a second, very engineery solution to word2vec's problems: if they could mathematically measure the system's bias, they could then mathematically remove it. If we know that "nurse" is closer to "woman" than it is to "man," we can scale the relations in the vector space and, voilà, no more bias. It is perhaps conceivable that this approach, which the authors termed "soft debiasing," could work to remove gender bias, but it would rely on the (mostly male, mostly white) engineers knowing exactly what that bias looks like. Engineers would have to work closely with linguists and gender theorists to find not only the blunt examples of gender bias but the subtle ones. Once they were done with that, they would presumably move on to other forms of bias. Kenney's work shows us that racism would need a cleanup, and after that there would still be ageism, ableism, and countless other isms to be addressed.

For Bolukbasi, now a researcher at Google Brain, bias was a problem to be debugged. What he seems to have missed—what so many programmers seem to miss—is that the issue is not so much about bias as it is about power, about exclusion and authorship. Whittaker argues that any technical solution to debiasing data will hit a fundamental snag. "Data is reductive," she explains, underlying the fact that any tool like word2vec is necessarily a subset of the real, messy linguistic world. "So whose vision of what is significant and is not significant are we adhering to? Who gets to make the determination about what is or is not muted?" Developers are still looking at the data *and the bias in the data* as things plucked from reality, rather than things that are authored by the developers, the code they write, and the social and political realities in which they were educated and in which they live. As the novelist and

photographer Teju Cole reminds us, “Authorship, after all, is not only what is created but also what is selected.”

★ ★ ★

Johanna Drucker, an author and cultural critic at UCLA, has taken perhaps the most drastic stance to date on the word “data”: that it should be replaced wholesale. In a 2011 essay, she argues that the original root of “data,” as a given, is anchored in a realist take on the universe: that there are real truths out there, given by God or physics, just waiting to be discovered. If I take a thermometer and measure the temperature in the room I’m sitting in right now, I might get a number: 22.3 degrees centigrade. If I take that number as a given, something that existed before my measurement, it’s too easy for me to think of my 22.3 as a real, indisputable truth. The number 22.3 was in the room already; my thermometer simply captured it. Drucker and her humanist compatriots would see my 22.3 differently. The number, they’d argue, is actually an artifact of a system of real-world things: an instrument (a thermometer), an act (my measuring of the temperature at the particular time and place I chose), and a set of cultural constructs (the centigrade scale, Arabic numerals, the concept of temperature).

The very word “data,” Drucker argues, is wrapped tightly in realist rhetoric and needs to be discarded. In exchange, she offers up a new word: “capta,” from the Greek root “to take” rather than “to give.” The key for Drucker is that “take” is active; if we understand that knowledge is something that is constructed, rather than picked up off a metaphysical curb, we accept the truth of our own role in its creation. “Capta,” Drucker explains, “is not an expression of idiosyncrasy, emotion, or individual quirks, but a systematic expression of information understood as constructed, as phenomena perceived according to principles of interpretation.” We need only to look at an example of the creation of data to understand Drucker’s argument.

This sentence that you are reading right now is not data. It doesn’t take long, though, if we consider the sentence for even a few

moments, for data to emerge. The number of words, the composition of the sentence's grammatical parts, verbs and nouns and prepositions. The number of vowels. The height of the sentence, in millimeters. Its length, in inches. The kerning between that first capital *T* and the *h* that follows. The amount of time it takes you to read it. How it makes you feel. It's a human instinct to measure, to describe, to make records.

When I see a bird fly past me, I file that experience into my brain, and I accompany it with a set of observations. The bird was small and brown, I might say. It was fast. Small, brown, fast—these are data that have been born from my brief flittering experience. Both the number of data and their character are constrained by who was doing the measuring (me) and what instrument was being used to measure (my forty-four-year-old eyes). Maans Booysen, a friend of mine and a legendary birder, would have come up with a set of different data from the same experience. He would surely have noted that the bird was a male barn swallow. Catching sight of a few stray white feathers on its undercarriage, he'd note that it was young, less than a year old. Having seen that the bird was carrying a small twig in its beak, Maans would have guessed that it was building a nest. Male, barn swallow, four months old, nesting. Maans has an almost supernatural ability to catch birds in flight with his camera (often held in one hand, a cigarette in the other), and if he'd had it with him, he likely would have snapped an image of the bird on the wing. From that image we could have found out more: that it was seventy-eight millimeters beak to tail, that its beak was the perfect orange of a clementine peel, that its outside right toenail was slightly bent. Taking this analogy to its furthest flight, we might catch the bird, photograph each of its forty thousand feathers, analyze its DNA, assay the bacteria living in its gut.

We might, in other words, end up with a lot of data about the bird that briefly flew past us. Depending on who we were, and what instrument we were using, that information varies in detail and accuracy. It also varies in character. "The bird seemed to be happy," my son might say, an observation that could be placed right

alongside the flight speed in meters per second, measured by comparing successive images from a tripod.

Data about anything—a sentence, a bird, the temperature of a room, the age of the universe, the sentiment of a tweet, the flow of a river—is an artifact of one fleeting moment of measurement and is, as Drucker's concept of capta gets at, as much a record of the human doing the measuring as it is of the thing that is being measured.

This idea that all data are constructed, that they are results of human action, is crucial, and I will return to it in the next chapters. For now, though, I'd like to consider a fate for the word "data" that is in many ways even more drastic than Drucker's revisioning.

* * *

So far we've focused on the meaning of "data," the ways in which the word is understood, and the linguistic neighborhoods it has, over time, come to occupy. We've seen that the definition of "data" has changed—from mathematical givens, to pieces of evidence, to assemblages of electronic bits and bytes. In all of these definitions, "data" is a thing, a noun. What if, along with a change in meaning, "data" could undergo a shift in use? What if "data" was a verb?

I data you; you data me. They data us; we data them.

As your *Concise Oxford* sails toward me from across the room, let's take some time to consider the arguments.

Since it drifted into the English language, "data" has been a plural noun. Specifically, it is the plural of "datum"—one datum, two data. Data purists have long made it a personal cause to nitpick improper usage; indeed there may be some of you who have already been irritated by my fast-and-loose changeups, my tendency to say that data is as well as data are. I could argue that I am trying to be true to data's roots—in Latin, the word "data" is both a neuter plural and a feminine singular—but the truth is that I am mostly standing in line with modern usage. Over the last decade, "data" has turned into a particular kind of singular: it has become, commonly, a mass noun.

Mass nouns are words that are treated as a single thing, no matter how much of that thing there may be. Blood, homework,

software, trash, love, happiness, advice, peace, confidence, flour, bread, and honey—all mass nouns, because they cannot be counted. I promise that you'll only read the phrase "big data" once in this chapter, and it's already over: this particular catchphrase was adopted exactly because we'd together passed a kind of Rubicon, where data could no longer be counted. It had become so vast that it could no longer be operated on by lowly humans, but instead had to be computed by always vaster and more elaborate systems of algorithms and semi-structured databases. As technology reacted to this dramatic shift in scale, so did language, and the word "data" found itself massified.

Because "data" has already endured such a drastic grammatical change, surely we can persuade the gods of common usage to shift the word's accepted part of speech entirely: Can we make "data" into a verb? In case this still seems too outlandish, consider two synonymic neighbors of "data": "record" and "measure." Both of these words exist as nouns (*I made a record*), as verbs (*We measured the temperature of the room*), and indeed as verbal nouns (*They found a list of measurements and recordings*). In comparison, isn't it strange to keep "data" confined to the dull, inactive realm of the noun?

The verbal forms of "record" and "measurement" make communication about the act of making records and taking measurements much easier. Rather than saying, "I am going to be making a record of this conversation," I can simply say, "I am recording this conversation." If we verbified "data," rather than having to say that the National Security Agency (NSA) is collecting data on our every interaction, movement, and metabolic function, we could simply say, "They data us."

Data is not inert, yet its perceived passivity is one of its most dangerous properties. When we are warned that a government is collecting data about its citizens, we may be underwhelmed specifically because this act of collection seems to be so harmless, so indifferent. But of course data is not collected and then left alone: it is used as a substrate for decision-making and as an instrument for differentiation, discrimination, and damage. Drucker's "capta" means to address this by reminding us that data is taken, but the fact that the

word remains a noun still gives it a character of inertness. Putting an active form of the word “data” into common parlance could serve as a reminder that the systems of data collection and use are humming with capacity for influence, action, and violence.

Making “data” a verb also exposes to us the power imbalances that have kept our collective endeavors drastically off-kilter. Grammatically speaking, “data” as verb would present a number of possibilities for subject/object combinations: I data you. You data me. We data you. You data us. They data me. They data us. We data them.

Exposed to this rich possibility of cause and effect, the common usages of data today become strikingly narrow: in our lived data experiences, we are objects rather than subjects. Google reads our every email, placing us ingloriously in marketing buckets based on what we write to our friends, colleagues, and lovers. Uber’s algorithms note our late-night voyages as records of romantic trysts. Images of our faces are captured by cameras on street corners, stored in databases to be cross-referenced by police departments and immigration enforcement officers. They data us; then they data us again.

Both my verbification of “data” and Drucker’s wholesale replacement of it are probably bridges too far for common usage. I don’t suspect we’ll find any textbooks on data-ing science on bookshelves anytime soon, nor will we hear about captabases or capta warehouses. Capta and data-ing are both useful constructs, though, in reminding us of two important things. First, that data is not found; it is constructed. Neither God nor the universe gives us data; we make it ourselves. Data is a human artifact. Second, that data’s construction acts in a real way on the world, that in making data we change the systems from whence it came.

Occasionally in the rest of this book we’ll return to these linguistic hacks as shortcuts for remembering these two things: data’s human provenance and its fundamentally active nature. I recommend that you try the same thing in your life: if you read a headline about data, try replacing it with “capta,” or rewriting the text with “data” as a verb. Try using “capta” or “data” as a verb in conversation, however jarring

it might seem. The environmental scientist Lauret Savoy has written that “names are one measure of how we choose to inhabit the world.” So, too, are definitions.

I’ll admit that we’re probably stuck for now with “data” as a noun. It is, however, a noun adrift. In the last ten years, “data” seems to be undergoing its most drastic change in meaning, one that I think offers a chance for the word to be broadly redefined. To understand this, let us shoo out the words that we had gathered in our room and welcome in another particular set: the words used in *The New York Times*. Specifically, let’s gather 19,057,600 words in 10,325 stories written about data between 1984 and 2018. Whereas our last set of words contained the entirety of the English language, with all of its nuance, this new set is restricted both in its range and in its structure. We are less likely to find “pomegranate” or “cephalopod” or “Wurlitzer” in this new set, nor will we find any of the hundreds of words and abbreviations that are banned by the *Times*’s style guide. What this new map of words does give us, though, is a setting well suited to exploring how the usage and meaning of “data” have changed in very recent times.

Something we can ask with this *New York Times* set is which words are changing their position relative to “data”; which nouns and verbs and adjectives are becoming less associated or more associated with “data” in the language of the news. Here is where we find evidence that something is indeed afoot. The words that are moving away from “data” are the ones that it has lived closely with for much of the last century: “information,” “digital,” “software,” “network.” Among the words that are moving toward “data” are some that seem to summarize recent events: “scandal,” “privacy,” “politicians,” “misinformation,” “Facebook.” There are also words that we might not previously have expected to find in the same sentence with “data”: “lives,” “deserve,” “place,” “ethics,” “friends,” “play.”

“Data,” it seems, is being pulled by strong currents. One eddy seems to be drawing it toward a dystopic future, an inevitable payback for a decade of unsavory practices. The other, seen through a hopeful lens, might bring data to a more utopian place. One in which it is bound tightly to our lives, to the places where we work and

play, to the friends with whom we share the experience of being. A future where “data” is in fact closer to “art” and “dirt” and “laughter.” Also to “community” and “empowerment” and “equality.”

Is it possible, then, that we might give it a push?

To do this, we need to unfold and examine the act of data. We need to understand how data is created, how it is computed upon, and how it is communicated. How it is made, changed, and told. We need to recognize that each of these steps cannot be looked at in isolation, that to know data we need to be able to look at the end-to-end process of it, to view data not as a noun, or a verb, or a thing but as a system and a process.

In doing so, we might begin to imagine a future perfect for data, where not only will they have data-ed us, but we will have data-ed them. A future, perhaps, where we might all data together.