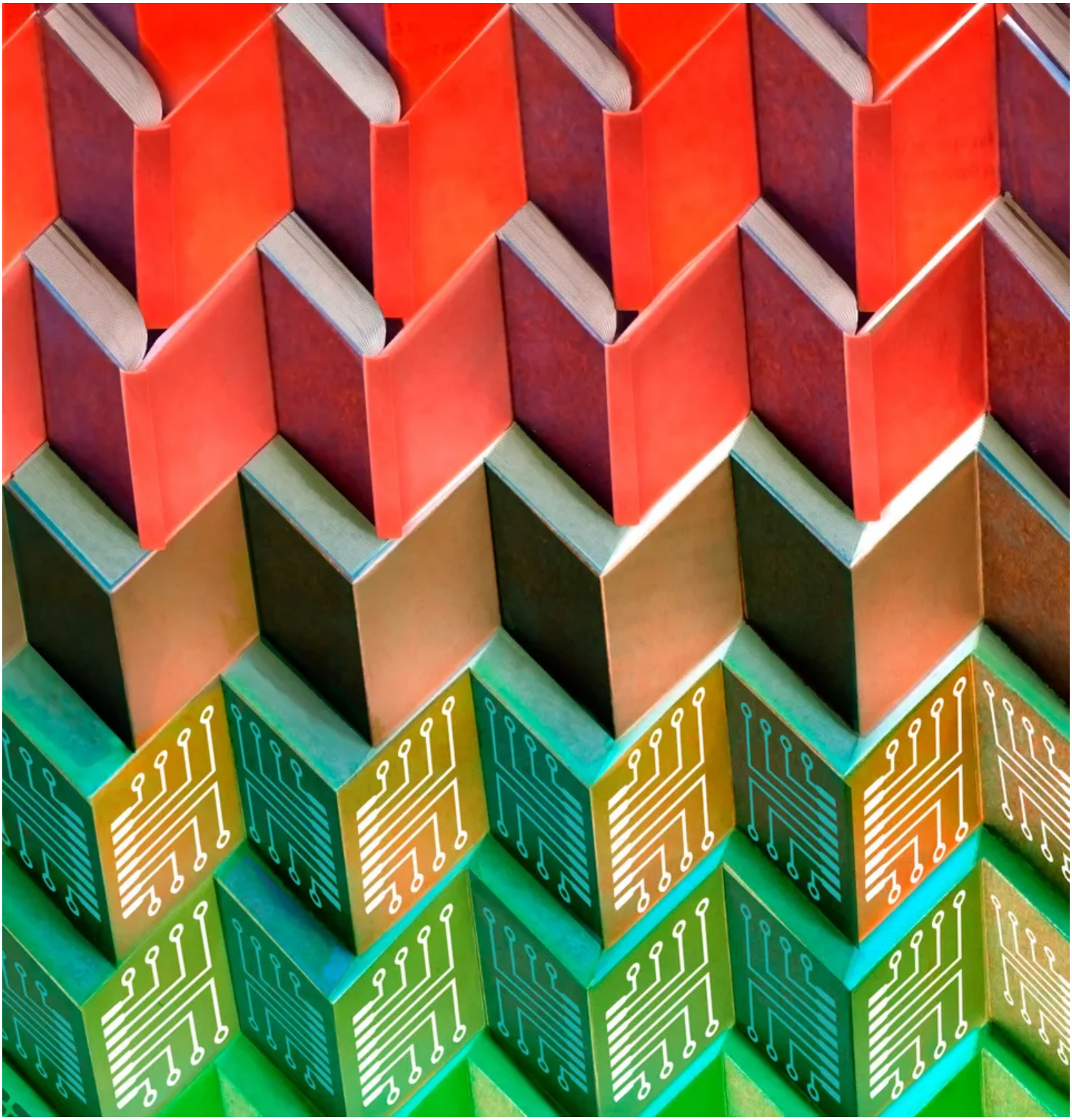
AMERICAN CHRONICLES APRIL 3, 2023 ISSUE

THE DATA DELUSION

We've uploaded everything anyone has ever known onto a worldwide network of machines. What if it doesn't have all the answers?

By Jill Lepore

March 27, 2023



The age of data is variously associated with late capitalism, authoritarianism, techno-utopianism, and the dazzle of "data science." **Illustration by Kelli Anderson**

One unlikely day during the empty-belly years of the Great Depression, an advertisement appeared in the smeared, smashed-ant font of the New York *Times'* classifieds:

WANTED. Five hundred college graduates, male, to perform secretarial work of a pleasing nature. Salary adequate to their position. Five-year contract.

Thousands of desperate, out-of-work bachelors of arts applied; five hundred were hired ("they were mainly plodders, good men, but not brilliant"). They went to work for a mysterious Elon Musk-like millionaire who was devising "a new plan of universal knowledge." In a remote manor in Pennsylvania, each man read three hundred books a year, after which the books were burned to heat the manor. At the end of five years, the men, having collectively read three-quarters of a million books, were each to receive fifty thousand dollars. But when, one by one, they went to an office in New York City to pick up their paychecks, they would encounter a surgeon ready to remove their brains, stick them in glass jars, and ship them to that spooky manor in Pennsylvania. There, in what had once been the library, the millionaire mad scientist had worked out a plan to wire the jars together and connect the jumble of wires to an electrical apparatus, a radio, and a typewriter. This contraption was called the Cerebral Library.

"Now, suppose I want to know all there is to know about toadstools?" he said, demonstrating his invention. "I spell out the word on this little typewriter in the middle of the table," and then, abracadabra, the radio croaks out "a

thousand word synopsis of the knowledge of the world on toadstools."

Happily, if I want to learn about mushrooms I don't have to decapitate five hundred recent college graduates, although, to be fair, neither did that mad millionaire, whose experiment exists only in the pages of the May, 1931, issue of the science-fiction magazine *Amazing Stories*. Instead, all I've got to do is command OpenAI's ChatGPT, "Write a thousand word synopsis of the knowledge of the world on toadstools." Abracadabra. *Toadstools, also known as mushrooms, are a diverse group of fungi that are found in many different environments around the world*, the machine begins, spitting out a brisk little essay in a tidy, pixelated computer-screen font, although I like to imagine that synopsis being rasped out of a big wooden-boxed nineteen-thirties radio in the staticky baritone of a young Orson the-Shadow-knows Welles. *While some species are edible and have been used by humans for various purposes, it is important to be cautious and properly identify any toadstools before consuming them due to the risk of poisoning*, he'd finish up. Then you'd hear a woman shrieking, the sound of someone choking and falling to the ground, and an orchestral stab. *Dah-dee-dum-dum-DUM!*

If, nearly a century ago, the cost of pouring the sum total of human knowledge into glass jars was cutting off in their prime hundreds of quite unfortunate if exceptionally well-read young men, what's the price to humanity of uploading everything anyone has ever known onto a worldwide network of tens of millions or billions of machines and training them to learn from it to produce new knowledge? This cost is much harder to calculate, as are the staggering benefits. Even measuring the size of the stored data is chancy. No one really knows how big the Internet is, but some people say it's more than a "zettabyte," which, in case this means anything to you, is a trillion gigabytes or one sextillion bytes. That is a lot of brains in jars.

Forget the zettabyten Internet for a minute. Set aside the glowering glass jars.

Instead, imagine that all the world's knowledge is stored, and organized, in a single vertical Steelcase filing cabinet. Maybe it's lima-bean green. It's got four drawers. Each drawer has one of those little paper-card labels, snug in a metal frame, just above the drawer pull. The drawers are labelled, from top to bottom, "Mysteries," "Facts," "Numbers," and "Data." Mysteries are things only God knows, like what happens when you're dead. That's why they're in the top drawer, closest to Heaven. A long time ago, this drawer used to be crammed full of folders with names like "Why Stars Exist" and "When Life Begins," but a few centuries ago, during the scientific revolution, a lot of those folders were moved into the next drawer down, "Facts," which contains files about things humans can prove by way of observation, detection, and experiment. "Numbers," second from the bottom, holds censuses, polls, tallies, national averages—the measurement of anything that can be counted, ever since the rise of statistics, around the end of the eighteenth century. Near the floor, the drawer marked "Data" holds knowledge that humans can't know directly but must be extracted by a computer, or even by an artificial intelligence. It used to be empty, but it started filling up about a century ago, and now it's so jammed full it's hard to open.

From the outside, these four drawers look alike, but, inside, they follow different logics. The point of collecting mysteries is salvation; you learn about them by way of revelation; they're associated with mystification and theocracy; and the discipline people use to study them is theology. The point of collecting facts is to find the truth; you learn about them by way of discernment; they're associated with secularization and liberalism; and the disciplines you use to study them are law, the humanities, and the natural sciences. The point of collecting numbers in the form of statistics—etymologically, numbers gathered by the state—is the power of public governance; you learn about them by measurement; historically, they're associated with the rise of the administrative state; and the disciplines you use to study them are the social sciences. The

point of feeding data into computers is prediction, which is accomplished by way of pattern detection. The age of data is associated with late capitalism, authoritarianism, techno-utopianism, and a discipline known as data science, which has lately been the top of the top hat, the spit shine on the buckled shoe, the whir of the whizziest Tesla.

All these ways of knowing are good ways of knowing. If you want to understand something—say, mass shootings in the United States—your best bet is to riffle through all four of these drawers. Praying for the dead is one way of wrestling with something mysterious in the human condition: the capacity for slaughter. Lawyers and historians and doctors collect the facts; public organizations like the F.B.I. and the C.D.C. run the numbers. Data-driven tech analysts propose "smart guns" that won't shoot if pointed at a child and "gun-detection algorithms" able to identify firearms-bearing people on their way to school. There's something useful in every drawer. A problem for humanity, though, is that lately people seem to want to tug open only that bottom drawer, "Data," as if it were the only place you can find any answers, as if only data tells because only data sells.

In "How Data Happened: A History from the Age of Reason to the Age of Algorithms" (Norton), the Columbia professors Chris Wiggins and Matthew L. Jones open two of these four drawers, "Numbers" and "Data." Wiggins is an applied mathematician who is also the chief data scientist at the *Times;* Jones is a historian of science and technology; and the book, which is pretty fascinating if also pretty math-y, is an adaptation of a course they began teaching in 2017, a history of data science. It begins in the late eighteenth century with the entry of the word "statistics" into the English language. The book's initial chapters, drawing on earlier work like Theodore Porter's "Trust in Numbers," Sarah Igo's "The Averaged American," and Khalil Gibran Muhammad's "The Condemnation of Blackness," cover the well-told story of the rise of numbers as an instrument of state power and the place of

quantitative reasoning both in the social sciences and in the state-sponsored study of intelligence, racial difference, criminology, and eugenics.

Numbers, a century ago, wielded the kind of influence that data wields today. (Of course, numbers are data, too, but in modern parlance when people say "data" they generally mean numbers you need a machine to count and to study.) Progressive-era social scientists employed statistics to investigate social problems, especially poverty, as they debated what was causation and what was correlation. In the eighteen-nineties, the Prudential Insurance Company hired a German immigrant named Frederick Hoffman to defend the company against the charge that it had engaged in discrimination by refusing to provide insurance to Black Americans. His "Race Traits and Tendencies of the American Negro," published in 1896 by the American Economic Association, delivered that defense by arguing that the statistical analysis of mortality rates and standards of living demonstrated the inherent inferiority of Black people and the superiority of "the Aryan race." In vain did W. E. B. Du Bois point out that suffering more and dying earlier than everyone else are consequences of discrimination, not a justification for it.

Long before the invention of the general-purpose computer, bureaucrats and researchers had begun gathering and cross-tabulating sets of numbers about populations—heights, weights, ages, sexes, races, political parties, incomes—using punch cards and tabulating machines. By the nineteen-thirties, converting facts into data to be read by machines married the centuries-long quest for universal knowledge to twentieth-century technological utopianism. The Encyclopædia Britannica, first printed in Edinburgh in 1768—a product of the Scottish Enlightenment—had been taken over for much of the nineteen-twenties by Sears, Roebuck, as a product of American mass consumerism. "When in doubt—'Look it up' in the *Encyclopaedia Britannica*," one twentieth-century newspaper ad read. "The Sum of Human Knowledge. *29 volumes, 28,150 pages, 44,000,000 words of text. Printed on thin, but strong opaque India*

paper, each volume but one inch in thickness. THE BOOK TO ASK QUESTIONS OF." When in doubt, look it up! But a twenty-nine-volume encyclopedia was too much trouble for the engineer who invented the Cerebral Library, so instead he turned seven hundred and fifty thousand books into networked data. "All the information in that entire library is mine," he cackled. "All I have to do is to operate this machine. I do not have to read a single book." (His boast brings to mind Sam Bankman-Fried, the alleged crypto con man, who in an interview last year memorably said, "I would never read a book.")

And why bother? By the nineteen-thirties, the fantasy of technological supremacy had found its fullest expression in the Technocracy movement, which, during the Depression, vied with socialism and fascism as an alternative to capitalism and liberal democracy. "Technocracy, briefly stated, is the application of science to the social order," a pamphlet called "Technocracy in Plain Terms" explained in 1939. Technocrats proposed the abolition of all existing economic and political arrangements—governments and banks, for instance—and their replacement by engineers, who would rule by numbers. "Money cannot be used, and its function of purchasing must be replaced by a scientific unit of measurement," the pamphlet elaborated, assuring doubters that nearly everyone "would probably come to like living under a Technate." Under the Technate, humans would no longer need names; they would have numbers. (One Technocrat called himself 1x1809x56.) They dressed in gray suits and drove gray cars. If this sounds familiar—tech bros and their gray hoodies and silver Teslas, cryptocurrency and the abolition of currency—it should. As a political movement, Technocracy fell out of favor in the nineteen-forties, but its logic stuck around. Elon Musk's grandfather was a leader of the Technocracy movement in Canada; he was arrested for being a member, and then, soon after South Africa announced its new policy of apartheid, he moved to Pretoria, where Elon Musk was born, in 1971. One of Musk's children is named X Æ A-12. Welcome to the Technate.

The move from a culture of numbers to a culture of data began during the Second World War, when statistics became more mathematical, largely for the sake of becoming more predictive, which was necessary for wartime applications involving everything from calculating missile trajectories to cracking codes. “This was not data in search of latent truths about humanity or nature,” Wiggins and Jones write. “This was not data from small experiments, recorded in small notebooks. This was data motivated by a pressing need—to supply answers in short order that could spur action and save lives.” That work continued during the Cold War, as an instrument of the national-security state. Mathematical modelling, increased data-storage capacity, and computer simulation all contributed to the pattern detection and prediction in classified intelligence work, military research, social science, and, increasingly, commerce.

Despite the benefit that these tools provided, especially to researchers in the physical and natural sciences—in the study of stars, say, or molecules—scholars in other fields lamented the distorting effect on their disciplines. In 1954, Claude Lévi-Strauss argued that social scientists need “to break away from the hopelessness of the ‘great numbers’—the raft to which the social sciences, lost in an ocean of figures, have been helplessly clinging.” By then, national funding agencies had shifted their priorities. The Ford Foundation announced that although it was interested in the human mind, it was no longer keen on non-predictive research in fields like philosophy and political theory, deriding such disciplines as “polemical, speculative, and pre-scientific.” The best research would be, like physics, based on “experiment, the accumulation of data, the framing of general theories, attempts to verify the theories, and prediction.” Economics and political science became predictive sciences; other ways of knowing in those fields atrophied.

The digitization of human knowledge proceeded apace, with libraries turning books first into microfiche and microfilm and then—through optical character

recognition, whose origins date to the nineteen-thirties—into bits and bytes. The field of artificial intelligence, founded in the nineteen-fifties, at first attempted to sift through evidence in order to identify the rules by which humans reason. This approach hit a wall, in a moment known as "the knowledge acquisition bottleneck." The breakthrough came with advances in processing power and the idea of using the vast stores of data that had for decades been compounding in the worlds of both government and industry to teach machines to teach themselves by detecting patterns: machines, learning. "Spies pioneered large-scale data storage," Wiggins and Jones write, but, "starting with the data from airline reservations systems in the 1960s, industry began accumulating data about customers at a rapidly accelerating rate," collecting everything from credit-card transactions and car rentals to library checkout records. In 1962, John Tukey, a mathematician at Bell Labs, called for a new approach that he termed "data analysis," the ancestor of today's "data science." It has its origins in intelligence work and the drive to anticipate the Soviets: what would they do next? That Netflix can predict what you want to watch, that Google knows which sites to serve you—these miracles are the result of tools developed by spies during the Cold War. Commerce in the twenty-first century is espionage for profit.

While all this was going on—the accumulation of data, the emergence of machine learning, and the use of computers not only to calculate but also to communicate—the best thinkers of the age wondered what it might mean for humanity down the line. In 1965, the brilliant and far-seeing engineer J. C. R. Licklider, a chief pioneer of the early Internet, wrote "Libraries of the Future," in which he considered the many disadvantages of books. "If human interaction with the body of knowledge is conceived of as a dynamic process involving repeated examinations and intercomparisons of very many small and scattered parts, then any concept of a library that begins with books on shelves is sure to encounter trouble," Licklider wrote. "Surveying a million books on

ten thousand shelves," he explained, is a nightmare. "When information is stored in books, there is no practical way to transfer the information from the store to the user without physically moving the book or the reader or both." But convert books into data that can be read by a computer, and you can move data from storage to the user, and to any number of users, much more easily. Taking the contents of all the books held in the Library of Congress as a proxy for the sum total of human knowledge, he considered several estimates of its size and figured that it was doubling every couple of decades. On the basis of these numbers, the sum total of human knowledge, as data, would, in the year 2020, be about a dozen petabytes. A zettabyte is a petabyte with six more zeroes after it. So Licklider, who really was a genius, was off by a factor of a hundred thousand.

Consider even the billions of documents that the U.S. government deems "classified," a number that increases by fifty million every year. Good-faith research suggests that as many as nine out of ten of these documents really shouldn't be classified. Unfortunately, no one is making much headway in declassifying them (thousands of documents relating to J.F.K.'s assassination, in 1963, for instance, remain classified). That is a problem for the proper working of government, and for the writing of history, and, not least, for former Presidents and Vice-Presidents.

In "The Declassification Engine: What History Reveals About America's Top Secrets" (Pantheon), the historian Matthew Connelly uses tools first developed for intelligence and counterintelligence purposes—traffic analysis, anomaly detection, and the like—to build what he calls a "declassification engine," a "technology that could help identify truly sensitive information," speed up the declassification of everything else, and, along the way, produce important historical insights. (Connelly, like Wiggins and Jones, is affiliated with Columbia's Data Science Institute.)

The problem is urgent and the project is promising; the results can be underwhelming. After scanning millions of declassified documents from the State Department's "Foreign Relations of the United States" series, for instance, Connelly and his team identified the words most likely to appear before or after redacted text, and found that "Henry Kissinger's name appears more than twice as often as anyone else's." (Kissinger, who was famously secretive, was the Secretary of State from 1973 to 1977.) This is a little like building a mapping tool, setting it loose on Google Earth, and concluding that there are more driveways in the suburbs than there are in the city.

By the beginning of the twenty-first century, commercial, governmental, and academic analysis of data had come to be defined as "data science." From being just one tool with which to produce knowledge, it has become, in many quarters, the only tool. On college campuses across the country, data-science courses and institutes and entire data-science schools are popping up like dandelions in spring, and data scientist is one of the fastest-growing employment categories in the United States. The emergence of a new discipline is thrilling, and it would be even more thrilling if people were still opening all four drawers of that four-drawer filing cabinet, instead of renouncing all other ways of knowing. Wiggins and Jones are careful to note this hazard. "At its most hubristic, data science is presented as a master discipline, capable of reorienting the sciences, the commercial world, and governance itself," they write.

It's easy to think of the ills produced by the hubristic enthusiasm for numbers a century ago, from the I.Q. to the G.D.P. It's easy, too, to think of the ills produced by the hubristic enthusiasm for data today, and for artificial intelligence (including in a part of the Bay Area now known as Cerebral Valley). The worst of those ills most often have to do with making predictions about human behavior and apportioning resources accordingly: using

algorithms to set bail or sentences for people accused or convicted of crimes, for instance. Connelly proposes that the computational examination of declassified documents could serve as "the functional equivalent of CT scans and magnetic resonance imaging to examine the body politic." He argues that "history as a data science has to prove itself in the most rigorous way possible: by making predictions about what newly available sources will reveal." But history is not a predictive science, and if it were it wouldn't be history. Legal scholars are making this same move. In "The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future" (PublicAffairs), Orly Lobel, a University of San Diego law professor, argues that the solution to biases in algorithms is to write better algorithms. Fair enough, except that the result is still rule by algorithms. What if we stopped clinging to the raft of data, returned to the ocean of mystery, and went fishing for facts?

In 1997, when Sergey Brin was a graduate student at Stanford, he wrote a Listserv message about the possible malign consequences of detecting patterns in data and using them to make predictions about human behavior. He had a vague notion that discrimination was among the likely "results of data mining." He considered the insurance industry. "Auto insurance companies analyse accident data and set insurance rates of individuals according to age, gender, vehicle type," he pointed out. "If they were allowed to by law, they would also use race, religion, handicap, and any other attributes they find are related to accident rate." Insurers have been minimizing risk since before the Code of Hammurabi, nearly four thousand years ago. It's an awfully interesting story, but for Brin this was clearly a fleeting thought, not the beginning of an investigation into history, language, philosophy, and ethics. All he knew was that he didn't want to make the world worse. "Don't be evil" became Google's model. But, if you put people's brains in glass jars and burn all your books, bad things do tend to happen. ♦