Essays

DAVID AUERBACH

The Stupidity of Computers

Published in: Issue 13: Machine Politics

Publication date: Winter 2012

Dumb

COMPUTERS ARE NEAR-OMNIPOTENT cauldrons of processing power, but they're also stupid. They are the undisputed chess champions of the world, but they can't understand a simple English conversation. IBM's Watson supercomputer defeated two top *Jeopardy!* players last year, but for the clue "What grasshoppers eat," Watson answered: "Kosher." For all the data he could access within a fraction of a second—one of the greatest corpuses ever assembled—Watson looked awfully dumb.

Hard Labor

CONSIDER HOW DIFFICULT IT IS to get a computer to do anything. To take a simple example, let's say we would like to ask a computer to find the most commonly occurring word on a web page, perhaps as a hint to what the page might be about. Here is an algorithm in pseudocode (more or less plain English describing the basic outline of a program), in which we have a table, WordCount, that contains the number of occurrences of each word on the page:

```
FOREACH CurrentWord IN WordsOnPage WordCount[CurrentWord] = WordCount[CurrentWord] + 1
MostFrequentlyOccurringWord = "????" FOREACH CurrentWord IN WordCount.keys() IF
(WordCount[CurrentWord] > WordCount [MostFrequentlyOccurringWord])
MostFrequentlyOccurringWord = CurrentWord PRINT MostFrequentlyOccurringWord,
WordCount[MostFrequentlyOccurringWord]
```

The first FOREACH loop counts the number of times each word occurs on the page. The second FOREACH loop goes through that list of unique words, looking for the one that has the largest count. After determining the most commonly occurring word, it prints the word and the number of occurrences.

The result, however, will probably be a word like *the*. In fact, the least commonly occurring words on a page are frequently more interesting: words like *myxomatosis* or *hermeneutics*. To be more precise, what you really want to know is what *uncommon* words appear on this page *more commonly* than they do on other pages. The uncommon words are more likely to tell you what the page is about. So it becomes necessary to track and accumulate this sort of data on vast scales.

Any instructions to a computer must be as laboriously precise as these. Some of the towering achievements in computer science have been in the creation of brilliantly clever, efficient, and useful algorithms such as Quicksort, Huffman Compression, the Fast Fourier Transform, and the Monte Carlo method, all reasonably simple (but not obvious) methods of accomplishing precisely specified tasks on potentially huge amounts of precisely specified data. Alongside such computational challenges there has been the dream of artificial intelligence: to get computers to think.

Search

ALMOST TWO DECADES ON, it's easy to forget the mess that resulted when you tried to use the early search engines—Lycos, AltaVista, Northern Light. In addition to spotty coverage of the nascent web, none of them had any particularly skillful way of ordering their results. Yahoo, which presented users with a search box, was in fact not a search engine at all, but a manually updated, far-from-universal directory of web pages. Either way, users frequently had to go through pages and pages of results, often with completely incoherent descriptions or summaries, in order to find something even resembling the desired information. As spam pages increased on the web, this problem grew worse. But how was a computer to know what a user was looking for, just on the basis of a list of search words?

One early search engine attempted to mask its inadequacy by achieving a semantic understanding of the queries being entered in its search boxes. Ask Jeeves (now known simply as Ask) encouraged users to type actual questions rather than keywords: "Where can I buy shoes?" rather than "shoe shops." Its front page featured a tuxedoed English butler, the eponymous Jeeves, who was unfortunately more like Bertie Wooster.

Jeeves did not handle the queries very well. Where other search engines searched for particular words on a page, and the more of them the better, Jeeves interpreted questions in a heuristic fashion, looking for interrogative keywords such as "where," "who," and "how," then transformed these into a more standard query, possibly with some special handling. "How old is President Clinton?" could become "President Clinton's age" as a query, after isolating "How old" and using a hard-coded rule that equated the answer to that question with a plausible number, say between 0 and 100. Jeeves then searched its archive of the internet for appearances of "President Clinton's age" or instances of "President Clinton" near a two-digit number (it certainly didn't know that "President Clinton" was equivalent to "Bill Clinton"). If it found a few websites with the phrase "President Clinton's age is 52" or "President Clinton is 52 years old" or even "President Clinton is 52," it could be reasonably confident that this was correct. But it only knew to look for those patterns because humans had hard-coded them. Other cases required similar handcrafted code. It was a cumbersome process and it didn't "scale." The particular case of asking for a person's age was covered, but anything more complicated—"How old is the New York subway system?" or "How old is primogeniture?" much less "How old is 12 in dog years?"—would confuse poor Jeeves.

Eliza and Shrdlu

JEEVES WASN'T EXACTLY FAKING IT—he really did try to "understand" the question, but his ability to do so was shockingly narrow. To computer programmers, he was just an updated version of Joseph Weizenbaum's ELIZA, a 1966 program that parodically simulated a Rogerian therapist, mostly by rephrasing its interlocutor's statements as questions. For example:

> How are you today? What would you like to discuss? >> `I'm having a nervous breakdown.` Is it because you are having a nervous breakdown that you came to me? >> `Yes.` You seem to be quite positive. >> `My father hates me.` Tell me more about your family. >> `I'm a mess.` Did you come to me because you are a mess? >> `No.` You are being a bit negative.

ELIZA could recognize keywords like *yes, no,* and *father* to deliver particular responses. But that was about as far as it went.

A few years later came SHRDLU, developed at MIT by the artificial intelligence researcher Terry Winograd. SHRDLU was a genuine computing achievement. It allowed for natural language interaction with a computer under narrowly circumscribed conditions. Differently shaped blocks were displayed on a screen, and the computer could manipulate these in response to typed instructions from a person:

PERSON
Pick up a big red block.

COMPUTER
OK. [does it]

PERSON
Grasp the pyramid.

COMPUTER
I don't understand which pyramid you mean.

PERSON
Find a block which is taller than the one you are holding and put it into the box.

COMPUTER
By "it," I assume you mean the block which is taller than the one I am holding.

COMPUTER
OK. [does it]

PERSON
What does the box contain?

COMPUTER
The blue pyramid and the blue box.

PERSON
What is the pyramid supported by?

COMPUTER
The box.

PERSON
How many blocks are not in the box?

COMPUTER
Four of them.

PERSON
Is at least one of them narrower than the one which I told you to pick up?

COMPUTER
Yes, the red cube.

PERSON
Is it supported?

COMPUTER
Yes, by the table.

PERSON
Can the table pick up blocks?

COMPUTER
No.

PERSON
Can a pyramid be supported by a block?

COMPUTER
Yes.

[...]

PERSON
A "steeple" is a stack which contains two green cubes and a pyramid.

COMPUTER

I understand.

PERSON
Are there are any steeples now?

COMPUTER
No.

PERSON
Build one.

COMPUTER
OK. [does it]

PERSON
Call the biggest block "superblock."

COMPUTER
OK.

PERSON
Have you picked superblock up since we began?

COMPUTER
Yes.

PERSON
Why did you drop it?

COMPUTER
By "it," I assume you mean the superblock. I can't explain a nonexistent event.

Every human sentence had to be transformed into a logical syntax that represented its meaning. The sentence "pick up a red block" was represented as the following:

```
(PICK UP A BIG RED BLOCK)
(CLAUSE MAJOR TOPLEVEL IMPER ACTV PRT TRANS)
(RSS1)
(PICK) (VG IMPER) NIL (PICK (VPRT VB INF TRANS MVB)
(UP (PRT))
(A BIG RED BLOCK)
(NG OBJ OBJ1 DET NS INDEF)
(OSS4)
(A (DET NS INDEF)) (BIG (ADJ)) (RED (ADJ)) (BLOCK (NOUN NS))))))
NIL)
```

Such programs seemed to hold promise for more generalized communication, but the secret of SHRDLU's success lay precisely in the very narrow domain it worked within. It could understand English, but only as long as you were talking about colored blocks and the movements that could be performed on them.

Alan Turing, the analytical genius who broke the German ENIGMA code during World War II and formulated some of the fundamental principles of computer science, famously proposed a "test" for whether a computer was intelligent: could it, in text-only conversation, convince a person that it was human? Turing predicted in 1950 that a computer would have at least 128 megabytes of memory and be able to pass his test with reasonable frequency by the year 2000. He died in 1954, and although personal computers hit that memory mark ten years ago, no computer has come close to passing his test.

The Gonz

STILL, COMPUTERS COULD SCORE an occasional victory by cheating. In 1989, the computer scientist Mark Humphrys wrote an absurdly primitive chatbot, MGonz, which not only convinced another user (at Drake University, reached in an early online chat room) that it was human, but browbeat this confession out of him:

> MGONZ
> when was the last time you had sex
>
> SOMEONE AT DRAKE
> yesterday.
>
> MGONZ
> ok honestly when was the last time you got laid
>
> SOMEONE AT DRAKE
> I told you it was yesterday.
>
> MGONZ
> you are obviously an asshole
>
> MGONZ
> ok honestly when was the last time you got laid
>
> SOMEONE AT DRAKE
> ok ok it was over 24 hours ago. for you it must have been 20 years

MGonz had no understanding of what the other user was saying, but succeeded, as Humphrys explained, through “profanity, relentless aggression, prurient queries about the user, and implying he was a liar when he made responses to these.”

Language

WITH ANYTHING MORE COMPLICATED than red blocks (or profanity and abuse), you ran into the problem of logically representing language. The problem was twofold. First, a program had to resolve the ambiguity inherent in a sentence’s syntax and semantics. Take the fairly simple sentence “I will go to the store if you do.” For an English speaker, this sentence is unambiguous. It means, “I will go to the store only if you go with me (at the same time).” But to a computer it may be confusing: does it mean that I will go to the store (how many times? and which store do I mean?) if you ever, in general (or habitually), go to the store, or just if you go to the store right now, with me? This is partly a problem with the word *if,* which can be restrictive in different ways in different situations, and possibly with the concept of *the store*, but there are lots of words like *if* and lots of concepts like *the store*, and many situations of far greater ambiguity: uncertain referents, unclear contexts, bizarre idioms. Symbolic logic cannot admit such ambiguities: they must be spelled out explicitly in the translation from language to logic, and computers can’t figure out the complex, ornate, illogical rules of that translation on their own.

Second, a program analyzing natural language must determine what *state of affairs* that sentence represents in the world, or, to put it another way, its meaning. And this reveals the larger problem: what is the relation of language to the world? In everyday life, people finesse this issue. No one is too concerned with exactly how much hair a man has to lose before he is bald. If he looks bald, he is. Even the legal profession can address linguistic confusion on an ad hoc basis, if need be. But when reality must be represented in language with no ambiguity (or with precisely delineated ambiguity, which is even harder), we’re stuck with the messiest parts of the problem.

Philosophers began to tackle this problem toward the end of the 19th century, with Gottlob Frege, Bertrand Russell, and the logical positivists (who believed language could be represented logically) taking an early lead, only to find themselves abandoned in the 1940s by a former adherent (Wittgenstein) and routed in the 1950s by W. V. O. Quine. But the early proponents of what came to be known as “semantic artificial intelligence” were heavily influenced by the work of the logical positivists, and spent many decades (and counting) trying to write programs that would give computers the tools to understand language. Barring a couple of holdouts, artificial intelligence has moved on.

Google

MEANWHILE, END RUNS WERE MADE around the problem of understanding human language. By the mid-'90s a number of researchers—including, most famously, two Stanford grad students named Sergey Brin and Larry Page—were trying to improve the quality of search results by ranking their importance. They found that by analyzing the topology of the web—which pages link to other pages—computers could roughly determine the most "interesting" and "relevant" pages *without any semantic understanding of natural language at all*. The importance of a page about Sergei Prokofiev could be determined, in part, from the number of pages that linked to it with the link text "Sergei Prokofiev." And in part from the importance of those other pages vis-à-vis Prokofiev. And in part from how often Prokofiev is mentioned on the page. And in part from how much *other* stuff was mentioned on the page. These signals of a page's standing are determined from the topological layout of the web and from lexical analysis of the text, but *not* from semantic or ontological understanding of what the page is about. Sergei Prokofiev may as well be selenium mining.

The problem of understanding language had not been solved, but it was immediately apparent that Google's results were vastly better than any other search engine's. It returned more relevant, content-rich pages than any search engine ever had. From the outside, Google did not alter the basic search paradigm. A user still searched for keywords and got results back. But the improvement in ranking was so profound and noticeable that Google quickly surpassed its competitors.

Rather than underscoring the impoverishment of semantic analysis, the great success of Google has indicated how much can be done in the absence of that analysis.

Money

A FOOTNOTE FROM THE HISTORY OF SEARCH. In an essay called "What Happened to Yahoo," the venture capitalist Paul Graham, who was a programmer at Yahoo in the late '90s, recalls the psychological impediments to improving search.

> I remember telling [Yahoo CEO] David Filo in late 1998 or early 1999 that Yahoo should buy Google, because I and most of the other programmers in the company were using it instead of Yahoo for search. He told me that it wasn't worth worrying about. Search was only 6 percent of our traffic, and we were growing at 10 percent a month. It wasn't worth doing better. I didn't say "But search traffic is worth more than other traffic!" I said "Oh, OK." Because I didn't realize either how much search traffic was worth. I'm not sure even Larry and Sergey did then. If they had, Google presumably wouldn't have expended any effort on enterprise search. If circumstances had been different, the people running Yahoo might have realized sooner how important search was. But they had the most opaque obstacle in the world between them and the truth: money. As long as customers were writing big checks for banner ads, it was hard to take search seriously. Google didn't have that to distract them.

Wikipedia

COMPUTERS ARE CONTINUING TO FOLLOW the path of Google, not Jeeves: nonsemantic analysis of huge amounts of data. Intelligence in the form of semantics has not been completely absent from search, but it has been shunted to the side. The genius of using the topology of the web as opposed to its meaning is that you can make use of classifications made by actual intelligence: the humans who have created links and other indicators of what is significant and useful to them.

The largest experiment in a different kind of crowdsourcing—Wikipedia—has shifted the responsibility for organizing information to humans. The results have been impressive. Despite Jimmy Wales's admission that the site's content is not reliable or of academic quality, Wikipedia today is the de facto first stop for information on any reasonably esoteric subject. Wikipedia's pages are the first search result on Google and Bing for everything from "Abraham Lincoln" to "Dante Alighieri" to "triskaidekaphobia." And the quality of the entries has improved significantly over the years and continues to improve.

Of course there are problems. Fights arise among contributors (you can see them duking it out in the edit history of any contentious article, from "Ayn Rand" to "Obama" to "taxes"), and recently an inner elite has come to exercise greater control over edits to keep things organized and nonlibelous. Wikipedia's coverage is heavily slanted toward subjects that its contributors specialize in, which is why the Chilean writer José Donoso's entry consists of only three paragraphs and the Japanese manga *Sailor Moon*'s entry runs over fifty pages. Furthermore, the site still relies on outdated material from the public domain, specifically the 1911 *Encyclopaedia Britannica*. Its most blatantly racist entries have been scrubbed from Wikipedia, and other entries have been marked for revision and replacement, but thousands are still present in whole or in part.

Semantically, Wikipedia has attempted to bring structure to its mass of text by categorizing its articles. Sadly this process has been too haphazard to produce anything comprehensive enough for a computer to use. Progressive rock fans have added a page for Etron Fou Leloublan, but while the band was slotted into "French experimental music groups" and "rock in opposition," they were not slotted into "progressive rock groups," "experimental rock," "art rock," or "avant-garde music," all of which exist as similarly incomplete categories on Wikipedia. W. G. Sebald's *The Rings of Saturn* is listed under "novels," though the Library of Congress disagrees. Wilhelm von Humboldt is listed under "German linguists," "German philosophers," "Christian philosophers," "classical liberals," "people from Potsdam," and several others, but not as a classicist, an anthropologist, a literary critic, or a philologist. Such problems are not fatal to Wikipedia's goals, but they show that even with concerted effort, merging people's written contributions into clear, universal semantic categories is extremely difficult. There is no technical reason why Wikipedia should not someday contain all the world's knowledge in some form. But it will not be a form that computers can understand.

The Rise of Ontologies

WITH WIKIPEDIA AND ITS inadequate categories, one enters the realm of *ontology*. The word originally meant the philosophical study of the nature of being. In the context of information science, it has taken on a different meaning having to do with the modeling of reality. In essence, an ontology is an explicit, formal definition of a conceptual framework for any number of kinds of entities, as well as any number of relationships between them. In contrast to a taxonomy, which is merely a hierarchical ranking of entities using a single relation, an ontology can have any number of hierarchical and nonhierarchical relationships between its entities. The key words, however, are "explicit" and "formal." An ontology is by definition a model of reality that is amenable to logical representation.

Ambiguities and restrictions in an ontology can cause problems. To take a simple example, China has forced citizens to change their names if the name contains a character that government computers do not recognize. In a more complex case from early 2009, Amazon mysteriously deranked a huge number of gay-themed books, vastly decreasing their visibility on the site. Ex-employee Mike Daisey offered the believable explanation that "a guy from Amazon France got confused on how he was editing the site, and mixed up 'adult,' which is the term they use for porn, with stuff like 'erotic' and 'sexuality.'" Amazon deranks any books it labels as porn, so when the "erotic" and "sexuality" categories were treated as porn, the effect was disproportionately on gay-themed books. The ensuing outrage was justified because even though there was no *intent* to single out gay literature, the underlying ontology *already* lumped disproportionate chunks of gay literature into the "erotic" and "sexuality" categories. Edmund White's and John Updike's books have comparable amounts of sexual content, but only the former was de- listed. In this way, a bias built into the underlying ontology emerged, facilitated but not created by computers.

Such ontologies have always existed, from Aristotle's categories to astrology to the *Diagnostic and Statistical Manual of Mental Disorders* (DSM). We rely on these categories; they help us reduce the complexity of the world. Like race and gender, they are cultural categories that evolve organically, with shifting biases and inadequacies. When the Library of Congress finds a book difficult to classify, they either refine their ontological system or they fit it into the existing system as best they can. It is always a matter of approximation.

Because computers are incapable of creating meaningful ontological categories, and because the invention of new ontologies requires collective acceptance before they can be used, the most workable approach is to have the expected consumers of an ontology create and maintain it themselves. We could call this process the crowdsourcing of ontology.

Amazon

WIKIPEDIA'S CATEGORIES are so haphazard that no one would use them authoritatively. But what about existing, explicit ontologies? In contrast to the flattened, semantically ignorant world of search, hierarchical, semantic ontologies are ubiquitous in a different internet application: shopping. Library of Congress data for books, feature specifications for electronics, age ranges for children's toys: these classifications are attached to products, and are effectively their metadata, providing a ready-made ontology for an object. They meet computers more than halfway.

Computers handle ontologies of this sort with ease, since they contain as little ambiguity as possible. Shopping provides a ready-made set of categories that don't require computers to have any understanding of the underlying relationships. It is just a matter of importing existing classificatory ontologies into the web. If the specification for an ontology already exists, computers do not have the intractable task of filling in the blanks. Any remaining ambiguities must be clarified by hand.

Of all the online companies, Amazon has made the most of these ready-made ontologies. They didn't have to explain their categories to people or to computers, because both sides already agreed what the categories were. In the early days of the web, Yahoo's categories felt arbitrary and disputable: "humor" could contain everything from a Sam Kinison pinup page to a collection of Henny Youngman one-liners to the full text of *Tristram Shandy*, with no further distinction between the links, and far more pages were missing than included. So they lost out to search. But Amazon's categories—"electronics," "furniture," "jewelry"—were concrete, ubiquitous, and universal. They could tell customers which were the bestselling toasters, which toasters had which features, and which microwaves were bought by people who had bought your toaster. The ontology was already there.

Amazon also had your entire purchasing history, and its servers could instantly compute recommendations you would be likely to accept. When you browse or purchase a book, for example, Amazon can recommend other books on the basis of a huge number of factors: the book's author, genre, publication date, or press; your wish list; your past purchasing history; the purchasing history of people who bought or browsed this book; your geographical location; the purchasing history of other people near your geographical location; and your personal data (age, gender, marital status, parental status) as inferred from purchases.

A person who regularly buys diapers and buys a book by Jean Piaget may be likely to buy other books about child rearing, while a person who buys books by Jean Piaget and L. S. Vygotsky may be likely to buy other psychology books. The challenges Amazon faces in making its recommendations are different from those search engines face. Search engines have to work with a minimum of ontological information, because it has proved extremely difficult to work within an ontology on the web. Amazon has a regulated, regimented ontology, so their task is to figure out how individual categories within that ontology relate to each other, building up further data about the extant ontology. It's all symbol manipulation to a computer, but the underlying nature of the data is completely different.

Amazon still has troubles. Aside from the haziness of categories like "erotic," there can simply be problems in importing the metadata. Amazon claimed that several reports on meat imports and exports were authored by the trio of "Chilled the Fresh," "Frozen Horse," and "Ass Meat Research Group," as a consequence of using the report title as a list of authors. These mistakes must be corrected by hand, as there is no easy way for a computer to detect them. (How could it? Plenty of human names incorporate actual words; likewise plenty of research groups or products have esoteric names attached to them.) Having explicit metadata just reduces the number of such errors for humans to find and correct.

Amazon would love to know more. If they had my medical records, they would be able to recommend pills for my headaches. If they had the contents of my bookshelves, they could stop recommending books I've already bought elsewhere. If they knew I was writing this article, they would recommend books about the future of computers.

Facebook, Twitter, Metadata

AMAZON HAS NOT, to my knowledge, tried to gather much information about its customers beyond their activity on Amazon. Rather, they have tried to get customers to do as much of their shopping on Amazon as possible. But another site has been much more aggressive in gathering information about its users: Facebook.

People do not normally add metadata to things they write or purchase. It's tedious. People do, however, enjoy classifying themselves and their interests. In retrospect, the commercial implications of this are obvious, but when social networks first arose they did not seem like great repositories of metadata. Sites like Friendster and Myspace allowed users to classify themselves by a few broad demographic categories. Friendster suffered from heavy downtime and technical stagnation, while Myspace quickly descended into anarchy. Neither knew how to capitalize on the popularity of its site. Facebook, though, tried to turn the unstructured profile fields of Friendster into data. No other social network has been so clever about understanding and marketing its users' natural impulse toward self-definition.

Like other social networks, Facebook began by having users list their interests, favorite TV shows, schools, and so on. They then attempted to structure and organize this data in ways that would prove lucrative. Realizing that social networking data was invaluable because it was: 1) a collection of disparate information about a single person; 2) already categorized into *real-life* ontologies; and 3) more or less public (exactly *how* public is an ongoing issue), Mark Zuckerberg and company set out to exploit their data fully.

They employed a two-pronged strategy, one social and one viral. Socially, they encouraged people to share personal data and interests, but they were careful about enforcing structure. Rather than just typing in one's school, users originally had to select from a list of known universities; Facebook also made sure that you had a working email address to confirm that you went there. Thus figuring out that "Harvard," "Crimson," and "Mark Zuckerberg's alma mater" referred to the same place was no longer an issue. As it grew, Facebook continued to impose structure on information, but the kind of information it cared about changed. It cared less about where you went to school and a lot more about your tastes and interests—i.e., what you might be willing to buy. This culminated in a 2010 redesign in which Facebook hyperlinked *all* their users' interests, so that each interest now led to a central page for that artist, writer, singer, or topic, ready to be colonized by that artist's management, publisher, or label. "The Beatles," "Beatles," and "Abbey Road" all connected to the same fan page administered by EMI. Updates about new releases and tours could be pushed down to fans' news feeds.

The other, viral prong has been more contentious. Facebook is very keen to amass data about what its users do on other sites. In 2007, they launched the Beacon feature, which, in its initial form, would track your shopping activity on other sites and publish it to your Facebook wall automatically: "David has just bought tickets for *The Room* on Fandango!" The other sites would get publicity, Facebook would get your data, and users would get no compensation for their disclosure. This was a step too far, and, facing immediate objections and bad PR, Facebook soon removed the "automatic" part of the equation. After a class-action suit, petitions, and headlines like "Facebook's Beacon More Invasive Than Previously Thought" rendered Beacon radioactive, Facebook shut it down completely in late 2009.

Facebook has since retrenched, but their constant efforts to collect and share users' personal information continue to provoke controversies over privacy. The latest is about a feature called Timeline, which makes users' entire history of activity on Facebook (and sites that integrate with Facebook) visible at a glance. The only way to edit this information is by going through every status update and photo during a seven-day "review" period. Or you can just let everything go public, which Facebook would prefer you do.

Facebook's success, in other words, has been a triumph of metadata. Another recent champion in the metadata Olympics is Twitter. It emerged on Twitter, almost spontaneously, that users would add metadata to posts if it would help promote what they'd written. Because Twitter messages are so short (the 140-character limit is an artifact of the limitations of the mobile signaling protocols on which SMS messages traveled), you cannot always find all messages on a given topic by searching for particular words. So users add "hashtags" identified by the "#" marker ("#iranelection," "#Kanye," "#OWS") to enable their messages to be found by searches. The Twitter servers did not need to treat these tags differently from normal words; it was users themselves who recognized "#" as signaling metadata. Anyone looking for an immediate update on a topic could easily search a relevant tag. Having recognized the value in that metadata, Twitter now features the most popular recent subjects in a "trending topics" sidebar. The result is a real-time gold mine of data for marketers.

Big Brother

THE BUREAUCRACY of government intelligence sifts through data using the same techniques as Twitter or Wikipedia, for purposes less innocuous than trying to sell you the new Nike sneaker or confuse you as to the proper classification of a prog rock band. Dana Priest and William Arkin's excellent series on US intelligence in the *Washington Post* in 2010 reported that "Every day, collection systems at the National Security Agency intercept and store 1.7 billion emails, phone calls and other types of communications." All are sorted and analyzed with ontological models no more definite or complete than those of Twitter or Wikipedia. In signals intelligence nowadays, there isn't enough human intelligence to process all the signals. Computers pick up the slack, with all their deficient understanding. We see it in health insurance, where claims are questioned or denied through partly automated processes that determine what treatment is "reasonable" and what premiums people should pay. At a certain scale, the quantitative must take over.

The government may be further behind than we think; FBI director Robert Mueller admitted in the 9/11 hearings that FBI databases only searched one word at a time: "flight" or "school" but not "flight school." More recently, the cables released by WikiLeaks at the end of 2010 each contain a handful of tags, à la Twitter. The CBC observed that one cable discussing Canada was tagged "CN" rather than "CA," designating Comoros instead of Canada, and a few cables were tagged with the nonexistent tag "CAN." It's safe to assume that these tags were also assigned manually.

The haphazardness and errors of government intelligence are not reassuring for national security, but neither are they reassuring for privacy mavens. Who knows what shortcuts are being taken in the service of expediency as surveillance data is processed? Who knows which Canadians may be classified as Comoros dissidents? Under such circumstances, it may seem quaint to complain about "profiling." Everything and everyone is being profiled all the time, often incompetently.

The Social Dimension

BETWEEN THE UNSTRUCTURED CHAOS of the web and the ordered categories of shopping products lie varying degrees of specificity. If we want computers to work with a part of the world, it helps to have a fixed, discrete ontology. But people can't always agree on a fixed ontology, or even agree to use one. People make mistakes and, as on Wikipedia, they often disagree. My iTunes library has tagged my music with genres including “Unclassifiable,” “Peace Punk,” “Indie Terror,” “Microphones in Trees,” and “Fuck Genres.” The two-stage process of getting us to agree on categories in an ontology and then to *use* them is far from trivial. When it happens, it is usually because we are forced into it by the government, employers, schools, or especially the market—whether the dating market or the job market or the supposedly new market known as the social network.

With the widely adopted ontologies of social networks, the sorts of analyses done on Amazon products can now be done on people. As with search engines, Facebook keeps analyses of data quite private, even if the data itself is considerably less private. But one social networking site has made some analyses public: the dating site OkCupid. Their OkTrends blog illuminates the correlations they’ve been able to draw from their data, and gives some insight into the possibilities offered by a large database of personal information. Besides offering users a barrage of multiple-choice questions (no semantic understanding necessary!) to help match them with other users, OkCupid allows people to create their own questions to use in matching, compiling in the process an increasingly elaborate ontology of personality, albeit a messy one. Their conclusions about users have included the following: vegetarians are more likely to enjoy giving oral sex; the richer you are, the more likely you are to be looking for casual sex, regardless of what country you live in; the majority of Idaho and Wyoming residents would rather lose the right to vote than the right to bear arms. OkTrends revealed that, on the basis of phrases distinct to particular demographics, white men disproportionately like Tom Clancy and Van Halen, white women like the Red Sox and Jodi Picoult, Latino men and women both like merengue and *bachata*, Asian women like chocolates and surfing the net, Indian men like cricket, Indian women like *bhangra* and *The Namesake*, and Middle Eastern women like “different cultures.”

Now, the OkTrends blog is written by Christian Rudder, a professional prankster and humorist and former math major who is no doubt aware that the sheer number of uncontrolled variables at work makes it dangerous to take any of these conclusions at face value. But the thing to remember is that these sorts of analyses *are* being taken at face value within social networks, national security agencies, insurance companies, and any other organization with access to a large amount of quantitative social data. Add to that the inevitable biases built into the ontologies—recall the Amazon mixup above—and the result is a labyrinth of approximate classifications that are being overlaid onto us.

Anti-Rhizome

IT SEEMS WE’VE RECREATED an ad hoc hierarchy where we were promised the death of hierarchy. What ever happened to the rhizome? Wasn’t it going to abolish all that?

Deleuze and Guattari’s *Capitalism and Schizophrenia* (1972–80) proposed the term *rhizome* as an alternative to the supposedly more hegemonic tree structures of information. The rhizome was to be an antihierarchical structure in which any piece could connect to any other, “an acentered, nonhierarchical nonsignifying system without a General and without an organizing memory or central automaton, defined solely by a circulation of states,” as D&G put it in *A Thousand Plateaus*. The rhizome was to offer liberation from the top-down managerial style associated with corporate capitalism. Many tech enthusiasts, noticing that the web is not hierarchical and has total flexibility in linkages, made the jump to using the term to describe the internet.

Yet the rhizome makes an awkward fit. The adjective frequently dropped from Deleuze and Guattari’s description of the rhizome is *nonsignifying*. This is because, as we’ve seen, the very nature of information on the web is symbolic, that is to say signifying. Hegemony lies not in the shape of the network, but in its semantics.

A glance at some of the “rhizomatic” maps of information in Manuel Lima’s recent *Visual Complexity: Mapping Patterns of Information* reveals the problem. In the political sphere, complicated charts analyzing networks of links from one political blog to another show clusters of linkages tightly within sets of “conservative” and “liberal” blogs. Another chart from a separate analysis shows clusters using a different taxonomy: progressive, independent, and conservative. Who decided on these categories? Humans. And who assigned individual blogs to each category? Again humans. So the humans decided on the categories and assigned the data to the individual categories—then told the computers to confirm their judgments.

Naturally the computers obliged.

Could a computer have determined a blog's political stance on its own? Unlikely. One could imagine a program that analyzed instances of known political figures with positive textual markers, but "Obama," say, and "genius" could just as well appear on a Republican blog, where it might be sarcastic, as on a Democratic blog, where it might be sincere. And the classification of a blog into political categories couldn't be done by seeing which other blogs were linked from it, because that is what the charts were supposed to measure. So the classification had to be done manually. Depending on the researchers (who could create two categories, or three, or far more), the results were completely different. The rhizome had not freed anyone from the received list of political stances, nor from the need to judge subjectively where a blog fell on the political spectrum.

The dissemination of information on the web does not liberate information from top-down taxonomies. It reifies those taxonomies. Computers do not invent new categories; they make use of the ones we give them, warts and all. And the increasing amount of information they process can easily fool us into thinking that the underlying categories they use are not just a model of reality, but reality itself.

Future Prospects

GORDON MOORE OBSERVED IN 1965 that the number of transistors that could be cheaply placed on an integrated circuit tended to double every two years, a prediction that has held true since and has been called Moore's law. Roughly speaking, computational processing power has grown at the same rate. While people have repeatedly predicted its end, the exponential growth has remained stunning: computers are literally a *million times* more powerful than they were forty years ago.

This has brought us Google and the iPhone, but it has not brought us HAL 9000. So what does the future hold? There are two pathways going forward.

First, *we will bring ourselves to computers*. The small- and large-scale convenience and efficiency of storing more and more parts of our lives online will increase the hold that formal ontologies have on us. They will be constructed by governments, by corporations, and by us in unequal measure, and there will be both implicit and explicit battles over how these ontologies are managed. The fight over how test scores should be used to measure student and teacher performance is nothing compared to what we will see once every aspect of our lives from health to artistic effort to personal relationships is formalized and quantified.

We will increasingly see ourselves in terms of these ontologies and willingly try to conform to them. This will bring about a *flattening* of the self—a reversal of the expansion of the self that occurred over the last several hundred years. While in the 20th century people came to see themselves as empty existential vessels, without a commitment to any particular internal essence, they will now see themselves as contingently but *definitively* embodying types derived from the overriding ontologies. This is as close to a solution to the modernist problem of the self as we will get.

This will not mean a killing of creativity or of ineffable spirit, but it will change the nature of our creativity. The increasingly self-referential and allusive nature of art has already made "derivative" less of a pejorative, and the ability to mechanically *process* huge amounts of data with computer assistance will play a larger role in the construction of art of all kinds. David Shields in his book *Reality Hunger* was wrong to say that fiction as an art form is dying out; it just awaits new creators to reinvent the form for a more quantitative age.

Second, *we will bring computers to us*, not semantically but physically. Computers will be able to interface more and more directly with the real world. As people have repeatedly learned, with great frustration, the task of manually ordering the world into a semantically meaningful format is too huge and imprecise a task—too imprecise for computers and too huge for humans. But as processing power increases and the size and cost of computers shrink to the point of microscopic disposability, they can be embedded into everything: roads, paper, clothing, skin, organs, medicines, food. From that will come a way forward.

Just as Google was able to create a smarter search engine once they possessed enough data, computers that are able to absorb the world at a sufficient level of detail—sights, sounds, textures, touches—can begin to construct a model of the world from the ground up, one not based on verbal or logical representation of concepts, but on basic sense data. This will not yield intelligence per se, but computers will be able to analyze the basic physical stuff of the world just as they currently analyze the basic informational stuff of the web. This is not so much nanotechnology as it is merely computer *ubiquity*, taking computers out of their present position as relatively discrete devices and making them a functional part of every piece of the real world. In such a situation, semantics will matter less, and the possibility of creating an artificial being that can behave intelligently in an *emergent fashion* becomes remotely plausible.

Reductive Ontologies

BUT WE WILL NOT SEE COMPUTERS acquire minds anytime soon, and in the meantime we will end up accommodating the formalist methodologies of computer algorithms. The problem is one of ambiguity as much as nonneutrality. A reductive ontology of the world emerges, containing aspects both obvious and dubious. Search engines crawl Wikipedia and Amazon, Facebook tries to create their own set of inferred metadata, the categories propagate, and so more of the world is shoehorned into an ontology reflecting ad hoc biases and received ideas, much as the 1911 *Encyclopaedia Britannica* just happens to have become one of the most-read sources on the planet in the past decade. These problems do not arise from malicious intent, but from expediency and happenstance.

There is good news and bad news. The good news is that, because computers cannot and will not "understand" us the way we understand each other, they will not be able to take over the world and enslave us (at least not for a while). The bad news is that, because computers cannot come to us and meet us in our world, we must continue to adjust our world and bring ourselves to them. We will define and regiment our lives, including our social lives and our perceptions of our selves, in ways that are conducive to what a computer can "understand." Their dumbness will become ours. +