Doting on Data

Book Review by Rebecca Goldin

Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart
Ian Ayres
Hardcover, Bantam, 2007
US$25.00, 272 pages
ISBN-13: 978-0-553-80540-6
Paperback, Bantam, 2008
US$14.00, 320 pages
ISBN-13: 978-0-553-38473-4

I remember driving halfway across the country for summer vacation with my parents—a trip that invariably included hours of "Twenty Questions" with my sister. I would typically pick a noun that was suggested by the environment—"tree", "license plate", or "sign"—and she had to figure out what it was I had in mind by asking yes/no questions. Eventually, I lost interest in the game—by picking a sufficiently obscure object, it was too easy to avoid discovery.

Until I found the website `http://www.20q.net/`, which can read my mind. While the computer suggested that I use only nouns that most people would recognize, it "guessed" I had "igloo" in my mind after nineteen questions, and I wasn't able to fool it for more than thirty questions until I tried "differential equation" which, admittedly, isn't a commonly known noun. The most impressive feature of the game is how it can make a good guess when the questions are so vague. How did the program get igloo by knowing it was hard and bigger than a breadbox?

Rebecca Goldin is professor of mathematics at George Mason University. She is also the director of research at Statistical Assessment Service (STATS), a nonprofit media education and watchdog group affiliated with George Mason. Her email address is `rgoldin@math.gmu.edu`.

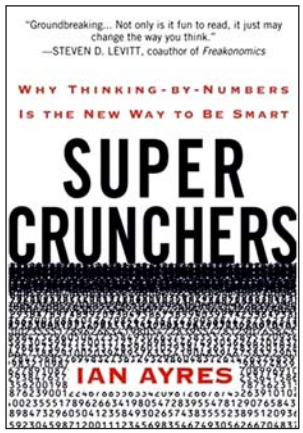
The game is an example of a super cruncher, a computer equipped with huge amounts of data that it can analyze and use to make predictions. The theme of the book *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart* by Ian Ayres is that super crunchers are better than humans at predicting outcomes—from whether a computer user will click on an ad to how good the wine will be this year.

A Good Read with Some Statistics Thrown In

For Ayres, "super crunching" is a catch-all term used for any computer activity involving data—sometimes super crunching means doing regression analysis, and other times it involves factors that are correlated with outcome. *Super Crunchers* is filled with amusing and illustrative stories of how society interacts with crunching computers. Some of these are banal—like how Amazon suggests you might like a book based on what you have purchased (and what you have browsed) and what others with similar Amazon-behavior have done. It's no surprise that sellers and advertisers want your information—and pay money to collect and organize it—in order to target you better for advertisements and future sales. What may be surprising is just how far this goes. Every time you use your discount card at the grocery store, you are participating in their data collection efforts on your eating habits in exchange for discounts on your purchases. Every time you swipe your credit card, you are sharing your shopping life with huge data-collection companies that will likely sell this information, perhaps even to a company that seems irrelevant to what you buy. Insurance companies, for example, would be happy to know your age and income, and the ages of your children, for

targeted advertising. If you buy diapers, they know you have a baby. If you also shop at Gymboree (an upscale children's clothing store), they know your income is above a certain level. Ayres points out that, with just a little number crunching, they can predict how likely you are to get a divorce in the next three years.

But *Super Crunchers* also details more serious consequences of data combing. Some of these include making decisions on behalf of humans, for the better or for the worse. An example comes from the dubiously named "Sexually Violent Predators Act", passed by my state of Virginia in 2003. The law allows the state to consider committing a sexual offender to a mental institution, even after he has served his full prison term, if he is considered a sexually violent predator. The law is meant to keep imprisoned anyone "who suffers from a mental abnormality or personality disorder which makes the person likely to engage in the predatory acts of sexual violence". Since not all cases can be reviewed in a timely fashion, a computer makes an initial judgment, based on a short algorithm—a simple "number crunching" formula that is considered a good predictor for future acts of sexual violence. The formula includes variables for assault *charges* (not convictions) and whether the sex offense was committed against a male or female. The computer gives each offender a score—and a high enough score triggers a human review for commitment (which then goes to a judge). While humans are involved after the trigger, the stakes are high: computers are (ostensibly) predicting whether ex-sex offenders will offend again.



Similarly, computer algorithms can predict our likelihood to repay a loan, and hence what interest rate we should pay. Unfortunately these algorithms can also be used to exploit us, though Ayres does not elaborate on this problem. If the computer is designed to predict what interest rate we'd be *willing* to pay rather than what we should pay, those on the receiving end of the loan might get short shrift, while those initiating the loans—and, presumably, designing or contracting the software—will benefit.

Ayres makes the important point that super crunching is not only about how profiteers can better exploit information about us to their benefit. First of all, the benefits of super crunching can go both ways in a financial transaction. If an advertiser finds its audience more easily, then those who sell the products profit, but so do those who receive the ads (by receiving information that they're interested in). More importantly, super crunching can actually be a weapon against excess-profit seekers. For example, when I buy clothes for my kids, I do a quick search for a "free shipping" promotion coupon. Truth is, had the coupon not existed, I would have paid the shipping costs and the company would have profited. But in under five minutes, I can often find a discount that it was presumably offering to those who would have shopped at that company only if there was a discount. I have just profited from a super cruncher. Similarly, when buying plane tickets, many people bypass specific airline websites in favor of sites like Expedia, Sidestep, and Yahoo!, which can search for prices and itineraries on all the major carriers at once.

But the level at which crunching data can improve our lives is not limited to our Internet or shopping experience. Doctors are increasingly turning to evidence-based medicine, which helps them determine the impact of thousands of drugs and lifestyle choices on disease. Hospital error can be reduced by analyzing the circumstances under which it occurs. Crunching can even be used to prove discrimination. Computers are used to catch criminals (via face recognition), thwart ideologues (by exposing their lack of sound statistical evidence for ideas on such topics as gun control), and promote social progress (through clinical trials, or recognition of discrimination, among many examples). The book is rife with wonderful stories about how reasoning with numbers and crunching data "solved" a mystery or changed the world for the better.

Mathematical Content

As far as turning us into super crunchers is concerned, the book falls short, leaving the reader with very limited explanations about how super crunchers actually work, aside from noting that they use large data bases. The main exception to this is a long and eloquent discussion on the meaning of "standard deviation" and normal curves, which should equip people who have never understood the topic with powerful machinery. I can only agree with the author that this concept is one of the most fundamental and ubiquitous statistical tools in existence. Readers—even savvy ones—will find the discussion refreshing and filled with real-world incidents in which understanding the mathematics can give one insight. For example, Ayres details how a colleague found corrupt sports betting based on data that didn't fit into a normal curve. He illustrates the notion of standard deviation with a discussion of IQ scores.

At one point, Ayres' language suggests a possible conflation of distinct concepts. Ayres points to a "head-to-head'' political survey, in which

Candidate A is preferred by 48 percent of the respondents, Candidate B is preferred by 52 percent, and there is a 2 percent margin of sampling error. What this means is that we can be 95 percent confident that the "true value" of Candidate B's support in the overall population from which the sample was drawn is somewhere between 50 and 54 percent. This does *not* mean, as Ayres states, that the true value has a 95 percent *chance* of being somewhere between 50 and 54 percent.

The distinction may be subtle, but it's an important one. Confidence intervals are based on a fixed true value, so that probability statements about the true value are inappropriate. Ayres seems to assume that the true value of a population's preferences is a random variable that is described by a Gaussian curve centered at 52 percent, and then asserts that there is a 97.5 percent chance that Candidate B is ahead of Candidate A based on this curve. But talking about the *chance* that the true value is greater than 50 percent does not make sense, since it either is or is not. It may be true, though Ayres does not explore the topic this way, that a probability could be defined making reference to many election polls in many elections—but additional assumptions would be needed.

Statistical subtleties aside, the notion that we *can* talk about standard deviations and interpret them is an important one, and except for the issue just mentioned, Ayres' description is excellent for a layman's understanding.

On another point, while Ayres provides an excellent defense of the strength of data, he does not always acknowledge the mathematical truths that can undermine the way numbers are interpreted. For one, Ayres frequently talks of past success as if it obviously engenders future success. If one person flips a coin 15 times in a row, the chance of getting at least 14 heads (or 14 tails) is less than 1 tenth of one percent. However, if 1,000 people flip coins 15 times, the likelihood that *someone* throws a coin and gets at least 14 heads (or 14 tails) is 62 percent. This classic experiment on the difference between a random occurrence and multiple testing is disregarded by Ayres.

The neglect of how multiple testing can affect modeling is reflected in his discussion, for example, of predictions of the quality of Bordeaux wines next year. He tells the very entertaining story of a modeler who was ignored by the skeptical wine-tasting experts but whose regression model ultimately proved a better predictor than the experts the following year. Since the best model is chosen based on *past* success (i.e., modeling how good the wine was in past seasons), without more information we cannot rule out the possibility that the model was successful at least partly by chance—like choosing the "best" coin among the 1,000 coin flippers. Will the regression model predict the next year's wine any better than a coin with an excellent track record? And if the wine turns out to be as expected according to the regression model, does it mean that the model predicted the wine quality, or that the modeler got lucky?

Ayres notes that people are reluctant to admit that computers are better than people at certain predictive skills. A reader has to be self-conscious, then, in challenging this point of view. What computers can do that humans cannot is process very large quantities of data—and if computers have the right data, they can do more than humans. It's a credit to our current computer capacity that crunching can do so much good for the world—think of the world without weather predictions! But if computers have the wrong data, not enough data, or the wrong algorithm, super crunching can get it really wrong.

Unfortunately, Ayres is reluctant to talk about when computer predictions don't get it right due to a faulty model or faulty data. A recent example of such an occurrence reported in the *Washington Post* after Ayres' book appeared involves the banking crisis. AIG Financial Products got involved with credit-default swaps in part thanks to a computer model that had crunched historical data about corporate debt. The model suggested huge profits were to be had, with very little assumption of risk. Ten years later, its parent company AIG is at the heart of the economic crisis, and over 150 billion government dollars have been invested to keep the company afloat. The *Washington Post* quoted executives who felt that the model was too good to be true—but ironically they allowed their faith in the computer model to trump their intuition. These days, much of the criticism of "what went wrong" with the current financial crisis suggests that human oversight and judgment could have played a positive role.

Super Crunching Requires Data

Super crunching has an implicit message: the data is "out there", and it's just a question of how to collect it and how to crunch it. There is certainly some truth in this—evidence-based medicine has made great progress due to huge amounts of data, thereby offering the possibility of decreases in hospital and doctor error and increases in effectiveness (again, Ayres has some great stories in this vein).

But sometimes, data is simply insufficient or unobtainable. Even worse, the existence of "related" data can lead us astray. What if the best predictors of whether a convicted sexual predator will commit another crime include whether he has a loving relationship with his parents, or whether he has experienced a self-awakening through therapy? While there are ways of "measuring" these factors (such as interviews and psychiatric evaluations), they are costly and time-intensive, and the *data*

showing how they are related to subsequent acts of sexual violence may simply not exist.

Ayres tells the story of a successful model for teaching, called Direct Instruction (DI). Teachers using DI for reading must follow a carefully scripted and fairly tedious class plan to teach students to read. Many educators are up in arms about DI, Ayres notes, because advocating this way to teach reading takes away their autonomy. However, according to Ayres, it is the most successful program ever documented through "scientific means". Ayres touches on the complaint that this method of teaching could increase reading skills while killing the love of reading, diminishing creativity, and stymieing the development of verbal comprehension. (I would add that it also may not challenge intellectually those students who have an easy time reading.) He then dismisses these claims largely because they are not supported by the data—he argues that, for those who don't accept DI, the *philosophy* of teaching is trumping the *science* of teaching.

But Ayres doesn't address the problem that the very tests that show the success of DI *become* the meaning of the word "success" to those who believe that all educational success can be read in the data. A deep understanding of character development, the ability to make connections among texts that are read, and the ability to articulate the author's purposes are much more difficult to assess in a simple test format. If you believe that tests can measure progress in reading, you are bound to turn to a surrogate (reading "skills"). Direct Instruction may do the best at increasing scores on skills tests. However, data from skills tests is a partial indicator of the desired outcome and should not be used to the exclusion of other evaluative techniques. This is not to say that the program isn't effective—it's to say that the evidence is far from sufficient to conclude what Ayres does. There is a parallel issue in mathematics, though Ayres does not address DI in mathematics: Some skills are easy to measure, and others simply go unmeasured in any controlled environment. The stakes are high, as we see with the No Child Left Behind legislation, which mandates scientifically based educational techniques and places a high, explicit premium on standardized test performance as an outcome measure.

The Ethics of Super Crunching

Like many people who deal with data crunching, Ayres notes that there is a strange bias in the world: Somehow, politically left-leaning people have deemed number crunching a tool for political conservatives, one that has vaguely demonic (or at least, inhuman) interests at heart. He argues vigorously against this view, maintaining that super crunching is by itself politically neutral and can be used by anyone to benefit. He cites his own laudable work in documenting discrimination in car loans. While many current uses of super crunching seem pro-business, Ayres notes the benefits to the average person. Super crunching is the way that Google targets its ads and can guess what you mean when you type an ambiguous phrase into its search engine. It's how a hospital can recognize dangerous drug interactions, and how governments can test social programs to decrease poverty. However, super crunching is not always used to benefit people—it's how airlines maximize their revenues, and how predatory lending companies got some people's personal information. Ayres predicts that consumer-driven crunching machines will become the norm, and websites will allow consumers to "fight back" against pricing schemes. He cites `http://farecast.com` as a prime example—it's a website that "predicts" whether airline fares are going up or down on a particular route (so you can decide to buy now or to wait before purchasing).

However, Ayres' book is remarkably silent on a host of important ethical issues surrounding super crunchers. While race cannot be explicitly used to determine a car loan rate, is it ethical to use one's mother's educational level, which has a high correlation with race? And if not, then what about other data about an individual, from where he or she lives to what he or she buys? Think of the "Twenty Questions'' game; with a fairly limited amount of data, a computer could predict race with high certainty. Would it be legal or ethical to use the kind of music one buys to determine his or her loan rate (along with a credit score, of course)?

The slope gets very slippery when we return to the Sexually Violent Predators Act from an ethical point of view. Are we looking at a future in which any trouble with the law could trigger a "review" that could lead to commitment to a mental institution for fear of future acts of violence? Perhaps we should just screen all people, rather than convicted offenders. What if we *could* find social factors that, via crunching large amounts, would allow us to determine who is likely to commit a crime and who is not (and should we arrest those who likely will, even if they haven't yet)? When does the individual stand separate from the statistic? The 1997 movie *Gattaca* told of a society in which one's life was determined by computer predictions based on genetics. Already, women can find out if they have a gene that will make them extremely likely to get breast cancer, and a simple blood test tells us if we are at high risk of heart disease. Super crunching has the possibility of constraining us if used incorrectly. Our ability to find correlations should be tempered by respect for our individuality.

Overall, Ayres' book is a good read, in the spirit of the well-known book *Freakonomics*, by Steven Levitt and Stephen J. Dubner. Don't expect you'll be a super cruncher when you're done, but you might turn to regression analysis to choose your wine.