Ian Ayres's *Super Crunchers* is Not about Super Crunching

Bruce D. McCullough
Drexel University

bdmccullough@drexel.edu

I regularly teach Ph.D. and MBA level statistics courses in data mining. Consequently, when I heard of the book *Super Crunchers* I ordered it immediately. On the first page of the introduction I was disappointed. Ayres writes of Princeton economist Orley Ashenfelter as an example of a super cruncher, *i.e.*, someone who analyzes large datasets: "In his day job, Orley crunches numbers. He uses statistics to extract hidden information from *large datasets*." [emphasis added] My Ph.D. is in economics, I am familiar with some of Ashenfelter's work, and I was surprised to learn that Ashenfelter has analyzed large datasets. Ayres goes on to give examples of these analyses: "As an economist at Princeton, he's looked at the wages of identical twins to estimate the impact of an extra year of school. He's estimated how much states value a statistical life by looking at differences in speed limits." I consulted Ashenfelter's *c.v.*, found the titles of the relevant journal articles, read the journal articles, and discovered that neither performed a statistical analysis on more than one thousand observations. Is one thousand a large dataset? No.

Data mining can roughly be described as the statistical analysis of large datasets. What separates traditional applied statistics from data mining is that the methods used for traditional statistics do not work on large datasets. Ayres clearly indicates that he is aware, at least nominally, that super crunching is data mining. On page 10 he writes, "What is Super Crunching? It is statistical analysis that impacts real-world decisions....The sizes of the datasets are really big...And when I say that Super Crunchers are using large datasets, I mean really large. Increasingly business and government datasets are measured not in mega- or giga- but in tera- and even petabytes[.]" (Henceforth, I shall use "super crunching" and data mining interchangeably.)

Why on earth, then, does Ayres use Ashenfelter as an example of a super cruncher, *i.e.* a data miner, *i.e.*, someone who uses "really big" datasets? The "Why Me?" section of his introduction yields great insight into this question. Ayres writes,

> *I'm a number cruncher myself. Even though I teach law at Yale, I learned econometrics while I was studying at MIT for a Ph.D. I've crunched numbers on everything from bail bonds and kidney transplantation to concealed handguns and reckless sex. You might think that your basic Ivory-tower egghead is completely disconnected from real-world decision making (and yes, I am the kind of absent-minded professor who was once so engaged in writing an article on a train that I went to Poughkeepsie instead of New Haven). Still, even datamining by eggheads can sometimes have an impact on the world.*

Ayres conveys the impression that he has actually done super crunching, and that his cited examples (bail bonds, etc.) are examples of data mining. Is either of these impressions correct?

The average reader would not know that econometrics at MIT did not include courses on the statistical analysis of large datasets (at least back when Ayres took his degree there), and most readers would not bother to check Ayre's *c.v.* to determine whether the indicated research actually analyzed large datasets. But I did. Letting *n* denote the "sample size" (*i.e.*,, number of observations) used in the cited statistical analysis, I found the sample sizes for bail bonds (n = 1366), concealed handguns (n = 62,350), and reckless sex (n = 5898). These sample sizes are far from the millions and millions that Ayres admits are the hallmark of true super crunching. He then spends a two full pages discussing his work with Steve Levitt (of *Freakonomics* fame) on Lojack (the electronic device for deterring car theft), as if it is an example of super crunching. It isn't. Ayres and Levitt analyzed car theft data on 57 cities over the period 1981-1994: n = 751.

Ayres continues in the next paragraph: "I have been a cook myself in the data-mining cafe." If this is not a claim to have actually done some data mining, I don't know what is. Is there any evidence, beyond his own say-so, that Ayres has actually performed the statistical analysis of a large dataset? His *c.v.* is rather difficult to read, in part because under "Scholarly Articles and Chapters" he actually includes letters to the editor (*e.g.*, "*HLA Matching in Renal Transplantation*, 332 *The New England Journal of Medicine* 752 (1995) with Robert Gaston and Mark Deierhoi" gives the impression of being an article but is actually a letter to the editor), so it is difficult to know what is an article and what is not. Nonetheless, a rough approximation is that he has published about 80 law review articles and seven articles in economics journals. None of the articles in economics has anything to do with data mining, and I strongly doubt (with one possible exception to be discussed anon) that any of the law review articles does, either.

Whence comes Ayres's alleged expertise in the analysis of large datasets? On page 132 Ayres does write, "With the yeoman help of Mark Cohen (who really bore the laboring oar), I have

crunched data on more than three million car sales." The notes in the back of the book for page 132 cite an article in the *California Law Review* that is not yet published so I cannot examine it. If Cohen was pulling the oar, does one instance playing coxswain in the data mining boat make Ayres an expert in super crunching? It is obviously due to his lack of experience in data mining (super crunching) that Ayres confuses traditional applied statistics with data mining (super crunching); hence his exemplars of super crunching given above have anything but large datasets. It's a mistake he makes repeatedly throughout the book. Regrettably, most readers of this book will be unaware that Ayres has no credible experience in data mining, at least none that this reviewer could discover from reading the book or from reading his *curriculum vitae*. It would be interesting to engage Ayres in a discussion of his favorite data mining software and favorite data mining methods.

Chapter Two, "Creating Your Own Data with the Flip of a Coin" gives a good layman's discussion of experimental design. Experimental design is used in tandem with data mining, as his examples make clear; but experimental design is not data mining. He discusses how the credit card company Cap One uses experimental design to refine credit card offers, noting that in 2006, Cap One "conducted 28,000 experiments -- 28,000 tests of new products, new advertising approaches, and new contract terms." But this is not data mining, it's 28,000 applications of traditional applied statistics. How the Cap One data miners came up with 28,000 hypotheses that needed testing -- *that's* probably the result of data mining. But Ayres nowhere discusses this. In the same chapter, Ayres describes how the director of customer relationship management at an airline randomly assigned delayed passengers to three groups who received different treatments: a letter of apology, a letter of apology and compensation, and a control group that received nothing. There's no point in discussing the result because this is not data mining, it's traditional applied statistics.

Chapter Three, "Government by Chance" spends pages discussing unemployment insurance experiments. While these are good examples of randomization, their sample sizes (usually about 15,000) makes them non-examples of data mining. For the same reason, the pages spent discussing the use of randomization in the criminology literature are not about super crunching. Chapter Four is about "evidence based medicine" and most of the examples have little to do with super crunching and much to do with traditional applied statistics; this is true of the remaining chapters in the book, so there is no point in discussing them.

There are irritating excursions into left-wing moralizing that don't belong in a book about numbers. As one example, when Ayres discusses online match-making services, he writes,

> *eHarmony has gotten into even more trouble for its refusal to match same-sex couples...Out of the top ten matching sites, eHarmony is the only one that doesn't offer same-sex matching. Why is eHarmony so out-of-step?...Its refusal to match gay and lesbian clients, even in Massachusetts and where same-sex marriage is legal, seems counter to the company's professed goal of helping people find lasting and satistfying marriage partners.*

Perhaps it is too much to expect that a man who writes books with titles like *Straightforward: How to Mobilize Heterosexual Support for Gay Rights* would be able to stop himself from editorializing on sexual mores in a book about data mining. One might at least hope that a Ph.D. economist could recognize that a firm in an industry might desire the benefits of market segmentation and not chase the same customers as other firms, but but this elementary economics seems to have escaped Ayres. Indeed, eHarmony has released a statement to this effect:

> *The research that eHarmony has developed, through years of research, to match couples has been based on traits and personality patterns of successful heterosexual marriages. Nothing precludes us from providing same-sex matching in the future, it's just not a service we offer now based upon the research we have conducted.*

For someone who refers to himself as "well placed to explore the rise of data based decision making," Ayres seems singularly unable to appreciate niche-marketing in the age of the internet.

Evidence that Ayres isn't familiar with elementary data mining methods (*e.g.,* how to discriminate between true and spurious predictive relationships via the train-test-validate procedure) comes from the way that he uncritically accepted Malcolm Gladwell's portrayal of Epagogix, the company that claims to be able to predict which movies will be successful and which will not. Ayres presents Epagogix as if it has a factual basis for its remarkable claims, but he cites nothing other than the Gladwell article. Steve Sailer put it best, writing in his blog:

> *I can assure you that it's easy to draw up a list of right predictions you've made in the past, as these guys did for Gladwell, but it's a lot harder to predict the future. One of the [Epagogix] businessmen says: ``With Mel Gibson's `The Passion,' people always say, `Who could have predicted that?' And the answer is, we could have.''*
>
> *Well, swell. I could have, too. Granted, I didn't, technically speaking, predict that. But I could have.*
>
> *What Gladwell should have done is tell them, "I'm Malcolm F****** Gladwell, the highest earning magazine journalist in America, and if you want me to write a huge article in the New Yorker about what geniuses you are, then you are going to have to pass my test. Here it is: Pick any ten movies scheduled for release over the next three months, read their shooting scripts, and then write down how much money they will*

make over their first four weeks of release. If you thoroughly beat the Hollywood Stock Exchange predictions, then I'll write the article. Otherwise, I won't."

Gladwell, not a statistician, perhaps can be excused for not knowing the difference between ex-post and ex-ante prediction. Yet Ayres swallowed the Epagogix marketing materials just as credulously as Gladwell did. In a book that regales the reader with the glories of randomization as the touchstone of statistical veracity, such an oversight is indefensible. How could Ayres, a self-professed "econometrician", have missed this? As he habitually passes off regular-size samples as if they were examples of data mining (super crunching), it is not surprising. However, not knowing very much about data mining does not imply that one does not know much about statistics. A better explanation is needed. Perhaps the jacket blurb can help with this conundrum.

The jacket blurb, for whom presumably the publisher and the author must share the blame, describes Ayres as "an econometrician and a lawyer." Not just an economist, mind you, but an econometrician; and he is an econometrician before he is a lawyer. Is there any truth to Ayres' claim that he is an econometrician? As noted, Ayres has published about 80 law review articles and perhaps one-tenth as many economics articles, so might he not better be described as a lawyer? And of the economics articles, not a single one would qualify as an "econometrics" article. Would Ayres dare call himself an econometrician in front of people who know better, or does he just do this before an unsuspecting public who cannot know any better? Consulting the directory of the American Economic Association, where members are free to identify their fields of specialization, Ayres's primary field is "Basic Areas of Law" and his secondary field is "Market Structure, Firm Strategy, and Market Performance." "Econometrics" is noticeably absent from his list of qualifications, at least when he is self-identifying to other economists, who would recognize the fatuity of his claim to be an econometrician. Of course, passing himself off as an econometrician to unsuspecting readers is in the same vein as passing off sample sizes of one thousand as large datasets. And in several cases, Ayres passes off the words of other writers as his own.

In his review of this book, *New York Times* economics columnist David Leonhardt noted that Ayres reproduced sentences from one Leonhard's articles without quotation marks. Subsequently the Yale student newspaper uncovered several instances where Ayres plagiarized passages in *Super Crunchers* from other writers. One wonders what would be found if all the words in this book were cross-referenced to the words in the articles on which Ayres based the book. Now *that* would be an example of data mining, more specifically, text-mining.

In sum, we have a book about super crunching written by someone who might once have seen super crunching done by someone else, and most of the examples in the book are not about super crunching. If this sounds appealing, buy the book.

About the author:

B. D. McCullough is Professor of Decision Sciences at Drexel University, specializing in statistics and data analysis.