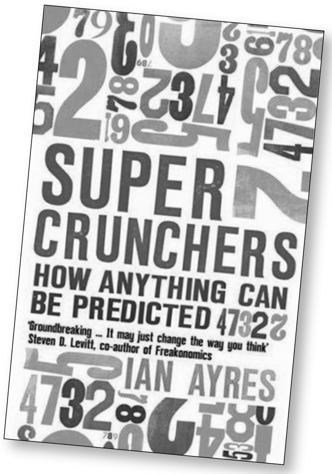
# THE PREDICTIVE POWER OF STATISTICS

Data crunching may soon be more important than intuitive decision-making, writes **Ross Farrelly**



**Super Crunchers: How Anything Can Be Predicted**

by Ian Ayres

John Murray, 2007
$35.00, 272pp
ISBN 9780719564642

On a recent visit to my local medical centre, I came face to face with the fact that Dr Google now seems to know more than my family GP. I was ushered into the consultation room, where I explained my symptoms to a slightly ruffled and obviously overworked doctor. He listened attentively, nodded thoughtfully, and then asked me to wait for a moment while he entered a few keywords into Google.

'Right, then,' he continued brightly, 'what you have is …' Then he proceeded to diagnose my ailment and prescribe the necessary treatment to clear it up. I was taken aback. What was he doing looking it up on Google? I expected that he'd have enough experience to know what my problem was. And what if his internet connection was down? Would he just guess?

I was reminded of this incident and my reaction to it when reading *Super Crunchers*. Ian Ayres' latest book addresses this very question the episode raised: what constitutes a correct use of data, and how can conclusions derived from it be used alongside an expert's intuition and experience?

Ayres, who is the William K. Townsend Professor at Yale Law School, and also a professor at Yale's School of Management, has written a very readable and highly topical book that sits easily alongside such titles as *Blink*, *Freakonomics*, *The Tipping Point*, and *The Wisdom of Crowds.* His main theme is that that we stand at a significant moment in history, where intuitive decision-making is now being complemented by data crunching, and may well soon be overtaken by it. Ayres argues that those who refuse to acknowledge the influence of high-powered data analysis will lose their competitive edge and will soon fall behind those who embrace the possibilities of this emerging field.

Ayres reasons that a number of factors have combined to make the collection, analysis, and utilisation of data essential activities for government, business, and consumers. As the cost of data storage decreases, it becomes feasible for companies to collect and store vast quantities of data. Furthermore, an increasing number of transactions are 'born digital,' so data-collection has become cheaper and quicker. These two factors, combined with the increasing power of personal computers, make data-based decision-making one of the fastest-growing and most important developments in modern society.

**Ross Farrelly** is a Sydney-based writer.

Throughout the book, Ayres uses the term 'super crunching' as a synonym for the less sexy but more commonly used term 'data mining.' Over the last fifteen years, data mining has emerged from the related fields of classical statistics, artificial intelligence, online analytical processing (OLAP), and machine learning. Data mining distinguishes itself from these older disciplines because it deals with huge data warehouses whose capacity is measured in terabytes ($10^{12}$ bytes) or even petabytes ($10^{15}$ bytes). Data mining is concerned with extracting predictions, patterns, and other useful information from these large datasets in such a way that ordinary mortals (not just statisticians) can understand them. Since data is constantly being added to many of these data warehouses, data mining is also concerned with real-time extraction of meaningful information and predictions.

Ayres opens his account of super crunching by referring to websites such as Digg, the iTunes Music Store, Amazon.com, and del.icio.us, which compile lists of most-read articles and books, most-downloaded songs, and most popular bookmarks. He uses these sites to illustrate how technology is taking the place of experts. In the past, we gained referrals by asking well-informed friends. Now we follow the advice of a website. He follows a similar vein by explaining how web-based matchmaking services and recruitment firms use real-time data collection and powerful regression models to predict how successful a prospective match will be. Once we listened to the advice of friends and colleagues when making such partnerships. Now we are guided, knowingly or unknowingly, by the coefficients in an unseen regression equation.

Ayres points out that while businesses exploit the power of super crunching to maximise profit, customers can also harness this new technology to get a better deal. He gives the example of Farecast, a website that not only reports and compares current prices of airline tickets, but also predicts whether or not the price will rise or fall in the following day or two.

One of the most interesting aspects of *Super Crunchers* is its discussion of the use of randomised trials to objectively test whether or not planned changes to strategies will actually produce the desired results. Ayres tells how Capital One, one of the largest issuers of credit cards in the US, used randomised trials to test alternative wording of marketing campaigns. The strength of this approach was that because of randomisation, the two groups exposed to the alternative wordings were identical in all but one factor: the different wording they had been exposed to. Without randomisation, differences in the performance of the two alternatives would have been confounded by hidden factors, and the results would have been much more difficult to interpret. Online, real-time data collection allows randomised trials to be conducted continuously. Ayres tells how he used Google Adwords to test two alternative titles for his book, making *Super Crunchers* itself something of a child of the super crunching phenomenon.

Stories such as these, in which companies use their considerable financial resources and technical expertise to wring every last cent out of unsuspecting customers, may not be to everybody's taste. However, the book describes a second application of randomised trials, which I found to be of great interest. Ayres describes the use of randomised trials to test the efficacy of proposed changes in government policy. He illustrates the point with the example of the *Progresa* program for education, health, and nutrition in Mexico. The program, introduced by President Ernesto Zedillo in 1995, is a conditional welfare program in which potential recipients must meet certain criteria before they receive cash payments. To receive welfare, recipients had to keep their children in school, report for nutritional monitoring, and seek prenatal care if they were pregnant. Politics in Mexico is such that no poverty program survives a change of president, so Zedillo knew he would need to prove the efficacy of his reforms if he wanted them to have lasting benefits for the Mexican poor. He decided to introduce randomised trials to test the effect of conditional welfare. The results were overwhelmingly positive, and when he lost the presidency in 2000 his successor 'scrapped' *Progresa* and replaced it with his own *Oportunidades* program, which just happened to have identical personnel, procedures, incentives, and outcomes.

The idea of testing policy reforms before they are introduced on a statewide or federal scale is an excellent one, and yet it is strangely absent from Australian politics. The *Progresa* program has obvious similarities to the reforms suggested by Noel Pearson to end passive welfare dependency in the Cape York Peninsula, and yet to my knowledge there has been no suggestion that the proposed reforms be subjected to testing by randomised trials before being introduced. ANU economist Andrew Leigh has written on this subject, rebutting the main objections to randomised trials in government policy and arguing that it is in the interest of political parties that value substance over rhetoric to submit their polices to such evaluation. He describes limited small-scale randomised trials in the NSW Drug Court and the Department of Family and Community Services, but points out that there is much scope in Australia for further development in this area.

Ayres also points out that some phenomena have been subject to random assignment for years and inadvertently provide a rich source of data that can be used to test many hypotheses. The pairing of college roommates, the assignment of judges to trials, and the sex of the village head in parts of India are the subjects of some of the examples mentioned.

Is high-powered data mining an unmitigated good? Ayres does not think so. He points out cases where statistical algorithms have incorrectly labelled people as criminals and excluded them from benefits to which they are entitled. He also points out the difficulty of checking other people's calculations and conclusions. However, the trend towards making datasets publicly available is mitigating this to some degree.

Throughout *Super Crunchers*, Ayres takes us on a rollicking ride through diverse subject areas including predicting the value of French red, estimating the value of American baseball players, the evidence-based medicine controversy, Kasparov versus Deep Blue, point shaving in college basketball, and direct instruction in education. He concludes the book with a plea for more emphasis on statistical literacy, and predicts that people who are equally at home with intuition and statistical analysis and able to toggle back and forth between the two will be at the forefront of research in the twenty-first century.

To emphasise his point about the need for statistical literacy, Ayres relates the story of Lawrence Summers, who, while president of Harvard University in 2005, made some comments at a conference about the scarcity of female professors in science and maths. Summers lost his job in part because of a lack of statistical literacy in the media and the reading public. Summers' comments regarding the relative intelligence of men and women related to the *spread* or *standard deviation* of IQ scores. However, they were interpreted as referring to *average* intelligence, which is quite another thing. Many believe that the controversy following his comments may have contributed to his resignation from Harvard in 2006.

*Super Crunchers* has an informative companion website (supercrunchers.com), which provides a number of online examples of the types of analyses Ayres describes. For example, one of the links allows you to enter a number of characteristics of your marriage and obtain an estimate of the probability that you will still be married in a certain number of years. (By some quirk, the probability my wife and I will still be together in fifteen years is estimated to be 107.65%!)

*Super Crunchers* is an entertaining, informative and well-written book. It assumes no specialised knowledge of mathematics or statistics. If you're looking for an in-depth discussion of contemporary data mining techniques, this is not the book you need, but as lively survey current of developments it's a valuable resource. It focuses almost exclusively on the US, and rarely mentions research or developments in any other country. Ayres subtitles his book *How Anything Can Be Predicted*. This is obviously an overstatement. Theoretically, this is entirely possible, but ethically and politically not all methods of prediction are acceptable. Moves in the US to open a futures market on the location and type of the next terrorist attack proved to be distasteful and never got off the ground.

So, having read *Super Crunchers*, will I now be any happier when I find myself being diagnosed by a computer or rejected from a job I know I can do because a statistical algorithm says I'm overqualified? No. But I will be more informed about how and why these things are happening, and probably be more prepared should I ever need to circumvent the ever-growing power of the super cruncher.