FT Magazine

How computers routed the experts

AUGUST 31, 2007 by: Ian Ayres

Six years ago, Ted Ruger, a law professor at the University of Pennsylvania, attended a seminar at which two political scientists, Andrew Martin and Kevin Quinn, made a bold claim. They said that by using just a few variables concerning the politics of the case, they could predict how the US Supreme Court justices would vote.

Analysing historical data from 628 cases previously decided by the nine Supreme Court justices at the time, and taking into account six factors, including the circuit court of origin and the ideological direction of that lower court's ruling, Martin and Quinn developed simple flowcharts that best predicted the votes of the individual justices. For example, they predicted that if a lower court decision was considered "liberal", Justice Sandra Day O'Connor would vote to reverse it. If the decision was deemed "conservative", on the other hand, and came from the 2nd, 3rd or Washington DC circuit courts or the Federal circuit, she would vote to affirm.

Ruger wasn't buying it. As he sat in that seminar room, he didn't like the way these political scientists were describing their results. "They actually used the nomenclature of prediction," he told me. "[But] like a lot of legal or political science research, it was retrospective in nature."

After the seminar he went up to them with a suggestion: why didn't they run the test forward? As the men talked, they decided to run a horse race, to create "a friendly interdisciplinary competition" to compare the accuracy of two different ways to predict the outcome of Supreme Court cases. In one corner stood the predictions of the political scientists and their flow charts, and in the other, the opinions of 83 legal experts – esteemed law professors, practitioners and pundits who would be called upon to predict the justices' votes for cases in their areas of expertise. The assignment was to predict in advance the votes of the individual justices for every case that was argued in the Supreme Court's 2002 term.

The test would implicate some of the most basic questions of what law is. In 1881, Justice Oliver Wendell Holmes created the idea of legal positivism by announcing: "The life of the law has not been logic; it has been experience." For him, the law was nothing more than "a prediction of what judges in fact will do". He rejected the view of Harvard's dean at the time, Christopher Columbus Langdell, who said that "law is a science, and ... all the available materials of that science are contained in printed books".

Many insiders watched with interest as the contest played out during the course of the Court's term; both the computer's and the experts' predictions were posted publicly on a website before the decision was announced, so people could see the results as opinion after opinion was handed down.

The experts lost. For every argued case during the 2002 term, the model predicted 75 per cent of the court's affirm/reverse results correctly, while the legal experts collectively got only 59.1 per cent right. The computer was particularly effective at predicting the crucial swing votes of Justices O'Connor and Anthony Kennedy. The model predicted O'Connor's vote correctly 70 per cent of the time while the experts' success rate was only 61 per cent.

How can it be that an incredibly stripped-down statistical model outpredicted legal experts with access to detailed information about the cases? Is this result just some statistical anomaly? Does it have to do with idiosyncrasies or the arrogance of the legal profession? The short answer is that Ruger's test is representative of a much wider phenomenon. Since the 1950s, social scientists have been comparing the predictive accuracies of number crunchers and traditional experts – and finding that statistical models consistently outpredict experts. But now that revelation has become a revolution in which companies, investors and policymakers use analysis of huge datasets to discover empirical correlations between seemingly unrelated things. Want to hedge a large purchase of euros? Turns out you should sell a carefully balanced portfolio of 26 other stocks and commodities that might include some shares in Wal-Mart.

In *Freakonomics*, Steven D. Levitt and Stephen J. Dubner showed dozens of examples of how statistical analysis of databases can reveal the secret levers of causation. Yet *Freakonomics* didn't talk much about the extent to which quick quantitative analysis of massive datasets – call it "super crunching" – is affecting real-world decisions. In fact, decision-makers in business and government are using statistical analysis to drive a wide variety of choices – and shunning the advice of traditional experts along the way.

Instead of simply throwing away the know-how of experts, wouldn't it be better to combine super crunching and experiential knowledge? Can't the two types of knowledge peacefully coexist? There is some evidence to support this possibility. Indeed, traditional experts are shown to make better decisions when they are provided with the results of statistical prediction.

But evidence is mounting in favour of a different and much more dehumanising mechanism for combining human and super-crunching expertise. Several studies have shown that the most accurate way to exploit traditional expertise is merely to add the expert evaluation as an additional factor in the statistical algorithm. Ruger's Supreme Court study, for example, suggested that a computer that had access to human predictions would rely on the experts to determine the votes of the more liberal members of the court (Stephen Breyer, Ruth Bader Ginsburg, David Souter and John Paul Stevens, in this case) – because the unaided experts outperformed the super-crunching algorithm in predicting the votes of these justices.

Instead of having the statistics as a servant to expert choice, the expert becomes a servant of the statistical machine. Mark E. Nissen, professor at the Naval Postgraduate School in Monterey, California, who has tested computer-versus-human procurement, sees a fundamental shift toward systems where the traditional expert is stripped of his or her power to make the final decision. "The newest space – and the one that's most exciting – is where machines are actually in charge," he says, "but they have enough awareness to seek out people to help them when they get stuck." It's best to have the man and machine in dialogue with each other, but, when the two disagree, it's usually better to give the ultimate decision to the statistical prediction.

The decline of expert discretion is particularly pronounced in the case of parole. In the past 25 years, 18 states have replaced their parole systems with sentencing guidelines. And those states that retain parole have shifted their systems to rely increasingly on super-crunching risk assessments of recidivism. Just as your credit score powerfully predicts the likelihood that you will repay a loan, parole boards now have externally validated predictions framed as numerical scores in formula. Still, even reduced discretion can give rise to serious risk when humans deviate from the statistically prescribed course of action.

Consider the case of Paul Herman Clouston. For more than 50 years, Clouston has been in and out of prison in several states for everything from car theft and burglary to escape. In 1972, he was convicted of murdering a police officer in California. In 1994, he was convicted in Virginia of aggravated sexual battery, abduction and sodomy, and of assaulting juveniles. He had been serving time in a Virginia penitentiary until April 15 2005, when he was released on mandatory parole six months before the end of his nominal sentence.

As soon as Clouston hit the streets, he fled. He failed to report for parole and failed to register as a violent sex offender. He is now one of the most-wanted men in Virginia. But why did this 72-year-old, who had served his time, flee? The answer is the Sexually Violent Predator Act (SVPA). In April 2003, Virginia became the 16th US state to enact such a statute, under which an offender, after serving his full sentence, can be found to be a "sexually violent predator" and subject to civil commitment in a state mental hospital until a judge is satisfied he no longer presents an undue risk to public safety. Clouston probably fled because he was worried that he would be deemed a sexual predator (defined in the statute as someone "who suffers from a mental abnormality or personality disorder which makes the person likely to engage in the predatory acts of sexual violence").

Virginia made Clouston "most wanted" for the same reason – and because it was embarrassed that Clouston had been released. You see, Virginia's version of the SVPA contained a super-crunching innovation. The statute included a "tripwire" that automatically sets the commitment process in motion if a super-crunching algorithm predicts that the inmate has a high risk of sexual offence recidivism. Under the statute, commissioners of the Virginia Department of Corrections were directed to review for possible commitment all prisoners about to be released who "receive a score of four or more on the Rapid Risk Assessment for Sexual Offender Recidivism". The Rapid Risk Assessment for Sexual Offender Recidivism (RRASOR, and pronounced "razor") is a points system based on a regression analysis of male offenders in Canada. A score of four or more on the RRASOR translates into a prediction that the inmate, if released, would in the next 10 years have a 55 per cent chance of committing another sex offence.

John Monahan, a leading expert in the use of risk-assessment instruments, notes: "Virginia's sexually violent predator statute is the first law ever to specify, in black letter, the use of a named actuarial prediction instrument and an exact cut-off score on that instrument." Clouston probably never should have been released: he had a RRASOR score of four. The state has refused to comment on whether they failed to assess Clouston's RRASOR score as directed by the statute or whether the committee reviewing his case chose to release him notwithstanding the statistical prediction of recidivism. Either way, the Clouston story seems to be one where human discretion led to the error of his release.

It was a mistake, that is, if we trust the RRASOR prediction. Before rushing to this conclusion, however, it's worth looking at what exactly qualified Clouston as a four on the RRASOR scale. The RRASOR system is based on just the four factors shown in the figure below. Clouston would receive one point for victimising a male, one for victimising a non-relative, and two more because he had three previous sex-offence charges. It's hard to feel any pity for Clouston, but this man could be funnelled towards lifetime commitment based in part upon crimes for which he had never been convicted.

RRASOR FIGURES

PRIOR SEXUAL OFFENCES	
None	0
1 conviction or 1-2 charges	1
2-3 convictions or 3-5 charges	2
4+ convictions or 6+ charges	3
AGE OF RELEASE	
More than 25	0
Less than 25	1
VICTIM GENDER	
Only females	0
Any males	1
RELATIONSHIP TO VICTIM	
Only related	0
Any non-related	1

SOURCE: John Monahan and Laurens Walker ”Social Science in Law: Cases and Materials” (2006).

What’s more, the statutory trigger expressly discriminates based on the sex of his victims. These factors are not chosen to assess the relative blameworthiness of different inmates. They are solely about predicting the likelihood of recidivism. If it turned out that wholly innocent conduct (putting barbecue sauce on ice cream, for example) had a statistically valid, positive correlation with recidivism, the RRASOR system, at least in theory, would condition points on such behaviour.

Since the statute was passed, the attorney-general’s office has sought commitments against only about 70 per cent of the inmates who scored a four or more on the risk assessment, and only about 70 per cent of the time have courts granted the state’s petition to commit these inmates. The Virginia statute thus channels discretion, but it does not obliterate it. To cede complete decision-making power to lock up a human to a statistical algorithm is in many ways unthinkable.

The problem is that discretionary escape hatches have costs too. In 1961, the Mercury astronauts insisted on a literal escape hatch. They balked at the idea of being bolted inside a capsule that could only be opened from the outside. They demanded discretion. However, it was discretion that gave Liberty Bell 7 astronaut Gus Grissom the opportunity to panic upon splashdown. In Tom Wolfe’s memorable account, *The Right Stuff*, Grissom “screwed the pooch” when he prematurely blew the 70 explosive bolts securing the hatch before the Navy SEALs were able to secure floats. The space capsule sank and Grissom nearly drowned.

In context after context, decision makers who wave off the statistical predictions tend to make poorer decisions.

Experts are overconfident in their ability to beat the system. We tend to think that the restraints are useful for the other guy but not for us. So we don’t limit our overrides to the clear cases where the formula is wrong; we override where we think we know better. And that’s when we get in trouble. Parole boards that make exceptions to the statistical algorithm and release inmates who are predicted to have a high probability of violence time and again find that the high probability parolees have higher recidivism rates than those predicted to have a low probability. Indeed, in Virginia only one man out of the dozens civilly committed under the SVPA has ever been subsequently released by a judge who found him – notwithstanding his RRASOR score – to no longer be a risk to society. Once freed, this man abducted and sodomised a child and now is serving a new prison sentence.

What does all this mean for human endeavour? If we care about getting the best decisions overall, there are many contexts where we need to relegate experts to supporting roles in the decision-making process. We, like the Mercury astronauts, probably can't tolerate a system that forgoes any possibility of human override, but at a minimum, we should keep track of how experts fare when they wave off the suggestions of the formulas. And we should try to limit our own discretion to places where we do better than machines.

This is in many ways a depressing story for the role of flesh-and-blood people in making decisions. It looks like a world where human discretion is sharply constrained, where humans and their decisions are controlled by the output of machines. What, if anything, in the process of prediction can we humans do better than the machines?

In a word, hypothesise. The most important thing left to humans is to use our minds and our intuition to guess at what variables should and should not be included in statistical analysis. A statistical regression can tell us the weights to place upon various factors (and simultaneously tell us how, precisely, it was able to estimate these weights). Humans, however, are crucially needed to generate the hypotheses about what causes what. The regressions can test whether there is a causal effect and estimate the size of the causal impact, but somebody (some body, some human) needs to specify the test itself.

So the machines still need us. Humans are crucial not only in deciding what to test, but also in collecting and, at times, creating the data. Radiologists provide important assessments of tissue anomalies that are then plugged into the statistical formulas. The same goes for parole officials who judge subjectively the rehabilitative success of particular inmates. In the new world of database decision-making, these assessments are merely inputs for a formula, and it is statistics – and not experts – that determine how much weight is placed on the assessments.

Still, universities are loath to accept that a computer could select better students. Book publishers would be loath to delegate the final say in acquiring manuscripts to an algorithm. But at some point, we should start admitting that the superiority of super crunching is not just about the other guy. It's not just about baseball scouts and parole officers and legal experts. Super crunching is affecting real-world decisions in many different contexts that touch us as consumers, as patients, as workers and as citizens.

Kenneth Hammond, the former director of Colorado's Center for Research on Judgment and Policy, reflects on the resistance of clinical psychologists to evidence that their predictions cannot match the accuracy of an algorithm's: "One might ask why clinical psychologists are offended by the discovery that their intuitive judgments and predictions are (almost) as good as, but (almost) never better than, a rule. We do not feel offended at learning that our excellent visual perception can often be improved in certain circumstances by the use of a tool (eg, rangefinders, telescopes, microscopes). The answer seems to be that tools are used by clerks (ie, someone without professional qualifications); if psychologists are no different, then that demeans the status of the psychologist." It may be demeaning but it's true: there has been a shift of discretion from clinicians to clerks, from traditional experts to a new breed of super crunchers, the people who control the equations.

This is an edited extract from 'Super Crunchers: How Anything Can Be Predicted' by Ian Ayres, published this month by John Murray.

Crunch time: when the computer gets it wrong

Once upon a time – say, a whole three decades ago – academics who studied maths and science tended to be divorced from the financial world. Wall Street and the City of London were wary of excessive academic zeal, or an overly numerical approach to allocating cash, **writes Gillian Tett (http://www.ft.com/comment/columnists/gilliantett)**, Financial Times Capital Markets editor.

No longer. If there is one essential factor that has marked the rise of 21st-century finance, it is the relentless ascent of the so-called "quant" – a financier who can employ quantitative (or highly numerical) skills to develop computer-based strategies, often using techniques transplanted from academic research in science or maths.

This quant trend first got underway three decades ago, with the mass production of computers, and just as some economists were starting to apply their rigorous numerical approaches to finance for the first time - and moving over to Wall Street. In the 1980s and 1990s, the use of quantitative approaches in finance became more widespread, helping to create complex products such as derivatives. Indeed, this field could not exist without "quants", since it is impossible to trade or value these products without numerical skills.

But this decade, the role of the "quant" has been turbo-charged. These days, "quants" do not simply create models to value markets: they also invent computer programs that can search the markets for trading opportunities – and then execute these trades too. When times are good in the financial markets – as they have been for much of this decade – these techniques have often been extremely lucrative, enabling the "quants" to make vast fortunes for themselves and their banks. But this summer the markets turned sour, and some of the quants' models went haywire, creating equally vast losses. Indeed, some observers blame much of the recent market storm on quant problems. And while this finger-pointing might be unfair, one thing is clear: these days the geeks are truly inheriting the earth – or at least, grabbing the limelight in 21st-century finance.

Grape expectations: the price of wine

Orley Ashenfelter is an economist at Princeton university, a former editor of the prestigious American Economic Review and a wine enthusiast. About 30 years ago, that led to some trouble. He decided that instead of using the "swishing and spitting" approach of wine gurus such as Robert Parker to predict auction prices, he would use statistics. Bordeaux are best when the grapes are ripe and their juice is concentrated. In years when the summer is hot, grapes get ripe. And, in years of below-average rainfall, the fruit gets concentrated. So it's in the hot and dry years that you tend to get the legendary vintages, **says Ian Ayres**.

Ashenfelter put these facts into a formula: Wine quality = 12.145 + 0.00117 winter rainfall + 0.0614 average growing season temperature - 0.00386 harvest rainfall.

By plugging the weather statistics for any year into this equation, he was able to predict the general quality of any vintage.

Ashenfelter started publishing his predictions in a newsletter called Liquid Assets. But his ideas reached a much larger audience in 1990, when The New York Times published a front-page article about his prediction machine. Where Parker had rated the 1986 Bordeaux as "very good and sometimes exceptional", Ashenfelter disagreed. Moreover, he predicted the 1989 Bordeaux, barely three months in the cask and yet to be tasted by critics, would be "the wine of the century". And, he said, 1990 was going to be even better.

Traditional wine critics were not pleased. Britain's Wine magazine said "the formula's self-evident silliness invite[s] disrespect". When Ashenfelter gave a wine presentation at Christie's wine department, dealers in the back hissed.

And Parker said Ashenfelter was "rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director".

He had a point. Just as it's more accurate to see the movie, shouldn't it be more accurate to actually taste the wine? But because Bordeaux spend 18-24 months in oak casks before they are set aside for ageing in bottles, experts have to wait four months just to have a first taste, after the wine is placed in barrels. And even then it's a rather foul, fermenting mixture. It's far from clear that tasting this undrinkable early wine offers accurate information about the wine's future quality.

What was clear was this: Ashenfelter's predictions were astonishingly accurate. The 1989s turned out to be a truly excellent vintage and the 1990s were even better. And while few wine experts have publicly acknowledged the power of Ashenfelter's predictions, their own forecasts now correspond much more closely to his simple equation results. Take that, Robert Parker.

Follow the topics mentioned in this article

The University of Pennsylvania

Supreme Court of the United States

Andrew Martin

Kevin Quinn

North America

Take a tour of myFT

Comments have not been enabled for this article.