<< Back to Article

WIRED MAGAZINE: 16.03

This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize

By Jordan Ellenberg 02.25.08

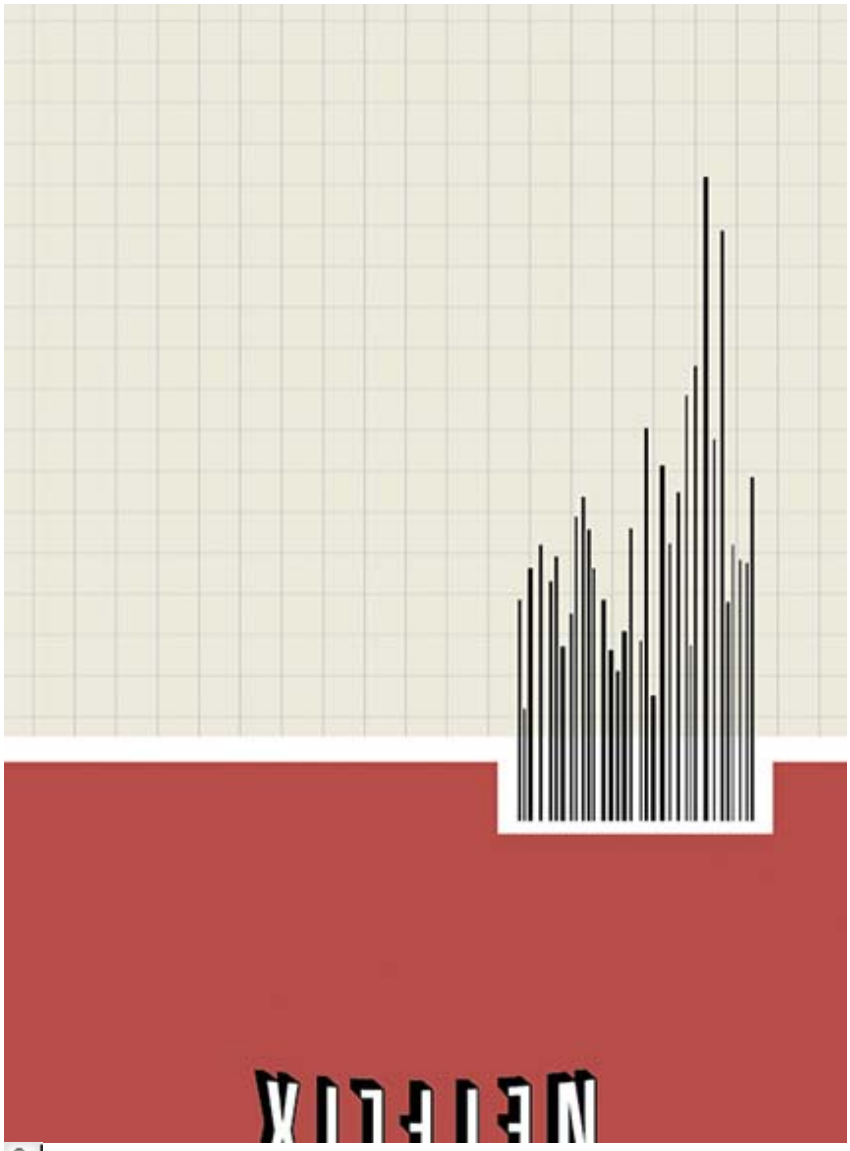

Illustration: Jason Munn

Check out who's ahead on the Netflix Prize leaderboard.

Forum for discussion about the Netflix Prize and dataset.

Read a detailed description of the Netflix Prize from James Bennett and Stan Lanning. (PDF)

At first, it seemed some geeked-out supercoder was going to make an easy million.

In October 2006, Netflix announced it would give a cool seven figures to whoever created a movie-recommending algorithm 10 percent better than its own. Within two weeks, the DVD rental company had received 169 submissions, including three that were slightly superior to Cinematch, Netflix's recommendation software. After a month, more than a thousand programs had been entered, and the top scorers were almost halfway to the goal.

But what started out looking simple suddenly got hard. The rate of improvement began to slow. The same three or four teams clogged the top of the leaderboard, inching forward decimal by agonizing decimal. There was BellKor, a research group from AT&T. There was Dinosaur Planet, a team of Princeton alums. And there were others from the usual math

powerhouses — like the University of Toronto. After a year, AT&T's team was in first place, but its engine was only 8.43 percent better than Cinematch. Progress was almost imperceptible, and people began to say a 10 percent improvement might not be possible.

Then, in November 2007, a new entrant suddenly appeared in the top 10: a mystery competitor who went by the name "Just a guy in a garage." His first entry was 7.15 percent better than Cinematch; BellKor had taken seven months to achieve the same score. On December 20, he passed the team from the University of Toronto. On January 9, with a score 8.00 percent higher than Cinematch, he passed Dinosaur Planet.

The Netflix challenge is just one example of a kind of problem called *data mining* — trying to make useful sense out of a gigantic dataset, typically rather noisy, completely unintelligible to the naked eye, and, despite its size, often painfully incomplete. Data mining is what Google does when it transforms the vast and ever-changing array of links on the Web into one number, PageRank, which it uses to figure out which page comes up first in your search. It's what intelligence agencies do — or at least what we surmise they do — when they search for red-flag patterns in a heterogeneous stew of visa applications, phone calls, and flight and hotel reservations. And it's what computer-aided detection software does for doctors when it boils down millions of observations of electrons passing through tissue into a single binary variable — tumor or no tumor.

Secrecy hasn't been a big part of the Netflix competition. The prize hunters, even the leaders, are startlingly open about the methods they're using, acting more like academics huddled over a knotty problem than entrepreneurs jostling for a $1 million payday. In December 2006, a competitor called "simonfunk" posted a complete description of his algorithm — which at the time was tied for third place — giving everyone else the opportunity to piggyback on his progress. "We had no idea the extent to which people would collaborate with each other," says Jim Bennett, vice president for recommendation systems at Netflix. When I ask Yehuda Koren, BellKor's leader, whether the prize money would go to him and his teammates or to AT&T, he pauses. He seems honestly to have never considered the question. "We got a big prize by learning and interacting with other teams," he says. "This is the real prize for us."

"Just a guy in a garage" was the exception to all this openness. He didn't even have a link attached to his screen name, which kept creeping higher and higher on the leaderboard. By mid-January, there were just five teams, out of 25,000 entrants, ahead of him. And still, no one knew who he was or by what statistical magic he kept improving. "He's very mysterious," says Koren with unconcealed interest. "I hope you will at least be able to find out his name."

His name is Gavin Potter. He's a 48-year-old Englishman, a retired management consultant with an undergraduate degree in psychology and a master's in operations research. He has worked for Shell, PricewaterhouseCoopers, and IBM. In 2006, he left his job at IBM to explore the idea of starting a PhD in machine learning, a field in which he has no formal training. When he read about the Netflix Prize, he decided to give it a shot — what better way to find out just how serious about the topic he really was?

In 2001, Potter cowrote a book called *Business in a Virtual World* that described how companies could best take advantage of new technology. So he's well aware of the commercial value of improving recommender systems, which tend to perform poorly, sometimes comically so. (You liked *The Squid and the Whale*? Try this Jacques Cousteau documentary.) "The 20th century was about sorting out supply," Potter says. "The 21st is going to be about sorting out demand." The Internet makes everything available, but mere availability is meaningless if the products remain unknown to potential buyers.

Potter says his anonymity is mostly accidental. He started that way and didn't come out into the open until after *Wired* found him. "I guess I didn't think it was worth putting up a link until I had got somewhere," he says, adding that he'd been seriously posting under the name of his venture capital and consulting firm, Mathematical Capital, for two months before launching "Just a guy." When he started competing, he posted to his blog: "Decided to take the Netflix Prize seriously. Looks kind of fun. Not sure where I will get to as I am not an academic or a mathematician. However, being an unemployed psychologist I do have a bit of time."

Oh, and he's not really in a garage: He works in a back bedroom on the second floor of his home in a quiet Central London neighborhood. The room is painted a cheery bright green and his children's toy boxes line the walls. His hardware rack is what he calls an "elderly" Dell desktop, recently refitted with 6 gigs of RAM to speed things up a bit. He doesn't run any experiments overnight; the rattling of the fan keeps his family awake.

Netflix Prize seeker Gavin Potter in his London home with his math consultant (and daughter) Emily.
Photo: Ed Hepburne-Scott

Next to Potter's computer there's a sheet of notebook paper. On it is an intricate computation in a neat, squarish hand. Not his — the calculation was done by his oldest daughter, Emily, a high school senior who plans to start a degree at Oxford next fall. She is, for the moment, serving as her father's higher-math consultant. "He gives me bits of calculus to do," she says, in a manner that suggests she feels ready to assume a position of greater responsibility on the project. (Emily has received no authoritative word as to what portion of any prize money would accrue to her personal accounts.)

Potter has had to work hard to understand and implement the complex mathematics that most contestants use. But he's no stranger to computers — as a young man he built an Ohio Scientific Superboard home computer from a kit and wrote software to predict the outcome of Premier League football matches. Anyway, his strategy isn't to out-math the mathematicians. He wants to exploit something they're leaving untapped: human psychology.

Netflix headquarters is a faux-Tuscan palazzo on the edge of Silicon Valley. The three-story building overlooks Interstate 280 in Los Gatos and shares a parking lot with an apartment complex from which it is architecturally indistinguishable. The interior is done up in brushed steel and decorated with tastefully arranged orchids. It looks like the entryway of a pan-Asian restaurant.

Founded in 1997, the company has more than 7 million subscribers, who have the option to rate movies on a scale of 1 to 5. In 2000, to encourage users to keep their subscriptions active, Netflix rolled out Cinematch, which used those ratings to help customers find new movies they'd like. When a user logs in, the service suggests "Movies You'll Love" — a list of films that the algorithm guesses will get a high rating from that particular user.

In March 2006, hoping to accelerate progress on Cinematch, the company decided to crowdsource the algorithm. Netflix constructed a data set of 100 million of the ratings customers had previously supplied and made it available to any coder who wanted a crack at it. The programmers use the data to write algorithms that predict how well users will like movies they haven't yet rated. Netflix tests the algorithms on a different ratings data set, which they've kept secret. Top scores are then

posted on a leaderboard.

The benchmark Netflix uses for the contest is called root mean square error, or RMSE. Essentially, this measures the typical amount by which a prediction misses the actual score. When the competition began, Cinematch had an RMSE of 0.9525, which means that its predictions are typically off by about one point from users' actual ratings. That's not very impressive on a five-point scale: Cinematch might think you're likely to rate a movie a 4, but you might rank it a 3 or a 5. To win the million, a team will have to make predictions accurate enough to lower that RMSE to 0.8572.

How much difference could that possibly make? A lot, Bennett says. Netflix offers hundreds of millions of predictions a day, so a tiny reduction in the frequency of insultingly stupid movie suggestions means a lot fewer angry users.

Over the last few years, the RMSE of Cinematch has steadily improved, as has Netflix's success at retaining customers from month to month. Bennett can't prove the two are related, but he's willing to bet on his belief that they are. He refuses to speculate on the dollar value of a 10 percent improvement to Cinematch, but he's certain it's substantially more than $1 million.

Contest participants retain ownership of the code they write, but the winning team must license it (non-exclusively) to Netflix. The company is already incorporating some of BellKor's ideas into its own system and in the future may buy code from other contestants, as well.

The data set, 100 times larger than any of its kind previously made public, is like a new, free library for specialists in data mining. So the contest has already brought Netflix a chorus of goodwill from computer scientists, who have, in turn, been happy to provide Netflix with free labor. "It's up to them to innovate now," Bennett says. "We're just the enablers." The Netflix team didn't publicize the strategies that were on the to-do lists of its own researchers — but one by one they were rediscovered, implemented, and evaluated by contestants. Netflix's programmers watched the leaderboard and read the forum obsessively. Various people had various bets on specific teams, Bennett says. "They all turned out to be wrong! But we didn't mind."

Since the prize has been such a success, might Netflix use the same model to solve other problems? I ask Bennett if there are more contests on the way. He pauses for a moment, thinking about what he wants to tell me. "One at a time," he says finally.

Many of the contestants begin, like Cinematch does, with something called the k-nearest-neighbor algorithm — or, as the pros call it, kNN. This is what Amazon.com uses to tell you that "customers who purchased Y also purchased Z." Suppose Netflix wants to know what you'll think of *Not Another Teen Movie*. It compiles a list of movies that are "neighbors" — films that received a high score from users who also liked *Not Another Teen Movie* and films that received a low score from people who didn't care for that Jaime Pressly yuk-fest. It then predicts your rating based on how you've rated those neighbors. The approach has the advantage of being quite intuitive: If you gave *Scream* five stars, you'll probably enjoy *Not Another Teen Movie*.

BellKor uses kNN, but it also employs more abstruse algorithms that identify dimensions along which movies, and movie watchers, vary. One such scale would be "highbrow" to "lowbrow"; you can rank movies this way, and users too, distinguishing between those who reach for *Children of Men* and those who prefer *Children of the Corn*.

Of course, this system breaks down when applied to people who like both of those movies. You can address this problem by adding more dimensions — rating movies on a "chick flick" to "jock movie" scale or a "horror" to "romantic comedy" scale. You might imagine that if you kept track of enough of these coordinates, you could use them to profile users' likes and dislikes pretty well. The problem is, how do you know the attributes you've selected are the right ones? Maybe you're analyzing a lot of data that's not really helping you make good predictions, and maybe there are variables that do drive people's ratings that you've completely missed.

BellKor (along with lots of other teams) deals with this problem by means of a tool called singular value decomposition, or SVD, that determines the best dimensions along which to rate movies. These dimensions aren't human-generated scales like "highbrow" versus "lowbrow"; typically they're baroque mathematical combinations of many ratings that can't be described in words, only in pages-long lists of numbers. At the end, SVD often finds relationships between movies that no film critic could ever have thought of but that do help predict future ratings.

Singular value decomposition is one example of a family of techniques in data mining known as "dimension reduction." A classic example of dimension reduction is the work of Frederick Mosteller and David Wallace on the Federalist Papers. They showed that frequencies of certain words distinguished those papers written by James Madison from those by Alexander Hamilton. Madison used "upon" and "while" much more frequently than Hamilton, while for "although" and

"whilst" the situation was reversed. So for each paper of disputed authorship, one can write down four numbers, corresponding to the frequencies of "upon,""while," "although," and "whilst." If the former two numbers are large and the latter two are small, you can confidently ascribe the paper to Madison. In this way, Mosteller and Wallace settled an argument that historians had been feuding about since the 19th century, with no firm conclusion in sight.

The danger is that it's all too easy to find apparent patterns in what's really random noise. If you use these mathematical hallucinations to predict ratings, you fail. Avoiding that disaster — called overfitting — is a bit of an art; and being very good at it separates masters like BellKor from the rest of the field.

In other words: The computer scientists and statisticians at the top of the leaderboard have developed elaborate and carefully tuned algorithms for representing movie watchers by lists of numbers, from which their tastes in movies can be estimated by a formula. Which is fine, in Gavin Potter's view — except people aren't lists of numbers and don't watch movies as if they were.

Potter likes to use what psychologists know about human behavior. "The fact that these ratings were made by humans seems to me to be an important piece of information that should be and needs to be used," he says. Potter has great respect for the technical prowess of BellKor — he is, after all, still behind the team in the rankings — but he thinks the computer science community studying this problem suffers from a bad case of groupthink. He refers to the psychological model underlying their mathematical approach as "crude." His tone suggests that if I weren't taping, he might use a stronger word.

It's easy to *say* you should take human factors into account — but how, exactly? How can you use psychology to study people about whom you know nothing except what movies they like?

Some things are easy. For example, the Netflix data set now covers eight years of ratings. If you think people's tastes change over time, you might want to weigh recent ratings more heavily than older ones.

A deeper part of Potter's strategy is based on the work of Amos Tversky and Nobel Prize winner Daniel Kahneman, pioneers of the science now called behavioral economics. This new field incorporates into traditional economics those features of human life that are lost when you think of a person as a rational machine, or as a list of numbers representing cinematic taste.

One such phenomenon is the anchoring effect, a problem endemic to any numerical rating scheme. If a customer watches three movies in a row that merit four stars — say, the *Star Wars* trilogy — and then sees one that's a bit better — say, *Blade Runner* — they'll likely give the last movie five stars. But if they started the week with one-star stinkers like the *Star Wars* prequels, *Blade Runner* might get only a 4 or even a 3. Anchoring suggests that rating systems need to take account of inertia — a user who has recently given a lot of above-average ratings is likely to continue to do so. Potter finds precisely this phenomenon in the Netflix data; and by being aware of it, he's able to account for its biasing effects and thus more accurately pin down users' true tastes.

Couldn't a pure statistician have also observed the inertia in the ratings? Of course. But there are infinitely many biases, patterns, and anomalies to fish for. And in almost every case, the number-cruncher wouldn't turn up anything. A psychologist, however, can suggest to the statisticians where to point their high-powered mathematical instruments. "It cuts out dead ends," Potter says.

We've entered the long twilight struggle of the Netflix Prize. "The last 1.5 percent is going to be harder than the first 8.5 percent," Potter tells me. In the past three months, BellKor's score has barely budged and now stands at 8.57 percent. Potter, meanwhile, is at 8.07 percent, and his pace has slowed, too. It's entirely possible that neither will ever make it to 10 percent. After all, there's a certain inherent variability to human choices that even the savviest computer can't predict.

Maybe the psychologist and the computer scientists would make more headway if they joined forces. Indeed, BellKor's leading program is actually a blend of 107 different algorithms, and the team is open to adding new ones. Potter has begun mixing more pure mathematics in with his psychology-inspired programs. But the two teams haven't expressed any interest in merging.

Potter says he's "still got juice left," but perhaps not quite enough to get to 10 percent. He's still hopeful though, and he's still testing new ideas. After all, if he wins, he'll be the guy who pointed the way to a new synthesis between psychology and computer science — and pocketed a million dollars in the process.

*Jordan Ellenberg (*ellenbergwired@gmail.com*) is a math professor at the University of Wisconsin and author of the novel* The Grasshopper King.