The New York Times
nytimes.com

PRINTER-FRIENDLY FORMAT SPONSORED BY

November 23, 2008

THE SCREENS ISSUE

If You Liked This, You're Sure to Love That

By CLIVE THOMPSON

THE "NAPOLEON DYNAMITE" problem is driving Len Bertoni crazy. Bertoni is a 51-year-old "semiretired" computer scientist who lives an hour outside Pittsburgh. In the spring of 2007, his sister-in-law e-mailed him an intriguing bit of news: Netflix, the Web-based DVD-rental company, was holding a contest to try to improve Cinematch, its "recommendation engine." The prize: $1 million.

Cinematch is the bit of software embedded in the Netflix Web site that analyzes each customer's movie-viewing habits and recommends other movies that the customer might enjoy. (Did you like the legal thriller "The Firm"? Well, maybe you'd like "Michael Clayton." Or perhaps "A Few Good Men.") The Netflix Prize goes to anyone who can make Cinematch's predictions 10 percent more accurate. One million dollars might sound like an awfully big prize for such a small improvement. But in fact, Netflix's founders tried for years to improve Cinematch, with only incremental results, and they knew that a 10 percent bump would be a challenge for even the most deft programmer. They also knew that, as Reed Hastings, the chief executive of Netflix, told me recently, "getting to 10 percent would certainly be worth well in excess of $1 million" to the company. The competition was announced in October 2006, and no one has won yet, though 30,000 hackers worldwide are hard at work on the problem. Each day, teams submit their updated solutions to the Netflix Prize Web page, and Netflix instantly calculates how much better than Cinematch they are. (There's even a live "leader board" ranking the top contestants.)

In March 2007, Bertoni decided he wanted to give it a crack. So he downloaded a huge set of data that Netflix put online: an enormous list showing how 480,189 of the company's customers rated 17,770 Netflix movies. When Netflix customers log into their accounts, they can rate any movie from one to five stars, to help "teach" the Netflix system what their preferences are; the average customer has rated around 200 movies, so Netflix has a lot of information about what its customers like and don't like. (The data set doesn't include any personal information — names, ages, location and gender have been stripped out.) So Bertoni began looking for patterns that would predict customer behavior — specifically, an algorithm that would guess correctly the number of stars a given user would apply to a given movie. A year and a half later, Bertoni is still going, often spending 20 hours a week working on it in his home office. His two children — 12 and 13 years old — sometimes sit and brainstorm with him. "They're very good with mathematics and algebra," he told me, chuckling. "And they think of interesting questions about your movie-watching behavior." For example, one day the kids wondered about sequels: would a Netflix user who liked the first two "Matrix" movies be just as likely to enjoy the third one, even though it was widely considered to be pretty dreadful?

Each time he or his kids think of a new approach, Bertoni writes a computer program to test it. Each new algorithm takes on average three or four hours to churn through the data on the family's "quad core"

Gateway computer. Bertoni's results have gradually improved. When I last spoke to him, he was at No. 8 on the leader board; his program was 8.8 percent better than Cinematch. The top team was at 9.44 percent. Bertoni said he thought he was within striking distance of victory.

But his progress had slowed to a crawl. The more Bertoni improved upon Netflix, the harder it became to move his number forward. This wasn't just his problem, though; the other competitors say that their progress is stalling, too, as they edge toward 10 percent. Why?

Bertoni says it's partly because of "Napoleon Dynamite," an indie comedy from 2004 that achieved cult status and went on to become extremely popular on Netflix. It is, Bertoni and others have discovered, maddeningly hard to determine how much people will like it. When Bertoni runs his algorithms on regular hits like "Lethal Weapon" or "Miss Congeniality" and tries to predict how any given Netflix user will rate them, he's usually within eight-tenths of a star. But with films like "Napoleon Dynamite," he's off by an average of 1.2 stars.

The reason, Bertoni says, is that "Napoleon Dynamite" is very weird and very polarizing. It contains a lot of arch, ironic humor, including a famously kooky dance performed by the titular teenage character to help his hapless friend win a student-council election. It's the type of quirky entertainment that tends to be either loved or despised. The movie has been rated more than two million times in the Netflix database, and the ratings are disproportionately one or five stars.

Worse, close friends who normally share similar film aesthetics often heatedly disagree about whether "Napoleon Dynamite" is a masterpiece or an annoying bit of hipster self-indulgence. When Bertoni saw the movie himself with a group of friends, they argued for hours over it. "Half of them loved it, and half of them hated it," he told me. "And they couldn't really say why. It's just a difficult movie."

Mathematically speaking, "Napoleon Dynamite" is a very significant problem for the Netflix Prize. Amazingly, Bertoni has deduced that this single movie is causing 15 percent of his remaining error rate; or to put it another way, if Bertoni could anticipate whether you'd like "Napoleon Dynamite" as accurately as he can for other movies, this feat alone would bring him 15 percent of the way to winning the $1 million prize. And while "Napoleon Dynamite" is the worst culprit, it isn't the only troublemaker. A small subset of other titles have caused almost as much bedevilment among the Netflix Prize competitors. When Bertoni showed me a list of his 25 most-difficult-to-predict movies, I noticed they were all similar in some way to "Napoleon Dynamite" — culturally or politically polarizing and hard to classify, including "I Heart Huckabees," "Lost in Translation," "Fahrenheit 9/11," "The Life Aquatic With Steve Zissou," "Kill Bill: Volume 1" and "Sideways."

So this is the question that gently haunts the Netflix competition, as well as the recommendation engines used by other online stores like Amazon and iTunes. Just how predictable is human taste, anyway? And if we can't understand our own preferences, can computers really be any better at it?

IT USED TO BE THAT if you wanted to buy a book, rent a movie or shop for some music, you had to rely on flesh-and-blood judgment — yours, or that of someone you trusted. You'd go to your local store and look for new stuff, or you might just wander the aisles in what librarians call a stack search, to see if anything jumped out at you. You might check out newspaper reviews or consult your friends; if you were lucky, your

local video store employed one of those young cinéastes who could size you up in a glance and suggest something suitable.

The advent of online retailing completely upended this cultural and economic ecosystem. First of all, shopping over the Web is not a social experience; there are no clever clerks to ask for advice. What's more, because they have no real space constraints, online stores like Amazon or iTunes can stock millions of titles, making a stack search essentially impossible. This creates the classic problem of choice: how do you decide among an effectively infinite number of options?

But Web sites have this significant advantage over brick-and-mortar stores: They can track everything their customers do. Every page you visit, every purchase you make, every item you rate — it is all recorded. In the early '90s, scientists working in the field of "machine learning" realized that this enormous trove of data could be used to analyze patterns in people's taste. In 1994, Pattie Maes, an M.I.T. professor, created one of the first recommendation engines by setting up a Web site where people listed songs and bands they liked. Her computer algorithm performed what's known as collaborative filtering. It would take a song you rated highly, find other people who had also rated it highly and then suggest you try a song that those people also said they liked.

"We had this realization that if we gathered together a really large group of people, like thousands or millions, they could help one another find things, because you can find patterns in what they like," Maes told me recently. "It's not necessarily the one, single smart critic that is going to find something for you, like, 'Go see this movie, go listen to this band!' "

In one sense, collaborative filtering is less personalized than a store clerk. The clerk, in theory anyway, knows a lot about you, like your age and profession and what sort of things you enjoy; she can even read your current mood. (Are you feeling lousy? Maybe it's not the day for "Apocalypse Now.") A collaborative-filtering program, in contrast, knows very little about you — only what you've bought at a Web site and whether you rated it highly or not. But the computer has numbers on its side. It may know only a little bit about you, but it also knows a little bit about a huge number of other people. This lets it detect patterns we often cannot see on our own. For example, Maes's music-recommendation system discovered that people who like classical music also like the Beatles. It is an epiphany that perhaps make sense when you think about it for a second, but it isn't immediately obvious.

Soon after Maes's work made its debut, online stores quickly understood the value of having a recommendation system, and today most Web sites selling entertainment products have one. Most of them use some variant of collaborative filtering — like Amazon's "Customers Who Bought This Item Also Bought" function. Some setups ask you to actively rate products, as Netflix does. But others also rely on passive information. They keep track of your everyday behavior, looking for clues to your preferences. (For example, many music-recommendation engines — like the Genius feature on Apple's iTunes, Microsoft's Mixview music recommender or the Audioscrobbler program at Last.fm — can register every time you listen to a song on your computer or MP3 player.) And a few rare services actually pay people to evaluate products; the Pandora music-streaming service has 50 employees who listen to songs and tag them with descriptors — "upbeat," "minor key," "prominent vocal harmonies."

Netflix came late to the party. The company opened for business in 1997, but for the first three years it

offered no recommendations. This wasn't such a big problem when Netflix stocked only 1,000 titles or so, because customers could sift through those pretty quickly. But Netflix grew, and today, it stocks more than 100,000 movies. "I think that once you get beyond 1,000 choices, a recommendation system becomes critical," Hastings, the Netflix C.E.O., told me. "People have limited cognitive time they want to spend on picking a movie."

Cinematch was introduced in 2000, but the first version worked poorly — "a mix of insightful and boneheaded recommendations," according to Hastings. His programmers slowly began improving the algorithms. They could tell how much better they were getting by trying to replicate how a customer rated movies in the past. They took the customer's ratings from, say, 2001, and used them to predict their ratings for 2002. Because Netflix actually had those later ratings, it could discern what a "perfect" prediction would look like. Soon, Cinematch reached the point where it could tease out some fairly nuanced — and surprising — connections. For example, it found that people who enjoy "The Patriot" also tend to like "Pearl Harbor," which you'd expect, since they're both history-war-action movies; but it also discovered that they like the heartstring-tugging drama "Pay It Forward" and the sci-fi movie "I, Robot."

Cinematch has, in fact, become a video-store roboclerk: its suggestions now drive a surprising 60 percent of Netflix's rentals. It also often steers a customer's attention away from big-grossing hits toward smaller, independent movies. Traditional video stores depend on hits; just-out-of-the-theaters blockbusters account for 80 percent of what they rent. At Netflix, by contrast, 70 percent of what it sends out is from the backlist — older movies or small, independent ones. A good recommendation system, in other words, does not merely help people find new stuff. As Netflix has discovered, it also spurs them to consume more stuff.

For Netflix, this is doubly important. Customers pay a flat monthly rate, generally $16.99 (although cheaper plans are available), to check out as many movies as they want. The problem with this business model is that new members often have a couple of dozen movies in mind that they want to see, but after that they're not sure what to check out next, and their requests slow. And a customer paying $17 a month for only one movie every month or two is at risk of canceling his subscription; the plan makes financial sense, from a user's point of view, only if you rent a lot of movies. (My wife and I once quit Netflix for precisely this reason.) Every time Hastings increases the quality of Cinematch even slightly, it keeps his customers active.

But by 2006, Cinematch's improving performance had plateaued. Netflix's programmers couldn't go any further on their own. They suspected that there was a big breakthrough out there; the science of recommendation systems was booming, and computer scientists were publishing hundreds of papers each year on the subject. At a staff meeting in the summer of 2006, Hastings suggested a radical idea: Why not have a public contest? Netflix's recommendation system was powered by the wisdom of crowds; now it would tap the wisdom of crowds to get better too.

AS HASTINGS HOPED, the contest has galvanized nerds around the world. The Top 10 list for the Netflix Prize currently includes a group of programmers in Austria (who are at No. 2), a trained psychologist and Web consultant in Britain who uses his teenage daughter to perform his calculus (No. 9), a lone Ph.D. candidate in Boston who calls himself My Brain and His Chain (a reference to a Ben Folds song; he's at No. 6) and Pragmatic Theory — two French-Canadian guys in Montreal (No. 3). Nearly every team is working on the prize in its spare time. In October, when I dropped by the house of Martin Chabbert, a 32-year-old member of the Pragmatic Theory duo, it was only 8:30 at night, but we had to whisper: his four children,

including a 2-month-old baby, had just gone to bed upstairs. In his small dining room, a laptop sat open next to children's books like "Les Robots: Au Service de L'homme" and a "Star Wars" picture book in French.

"This is where I do everything," Chabbert said. "After the kids are asleep and I've packed the lunches for school, I come down at 9 in the evening and work until 11 or 12. It was very exciting in the beginning!" He laughed. "It still is, but with the baby now, going to bed at midnight is not a good idea."

Pragmatic Theory formed last spring, when Chabbert's longtime friend Martin Piotte — a 43-year-old electrical and computer engineer — heard about the Netflix Prize. Like many of the amateurs trying to win the $1 million, they had no relevant expertise. ("Absolutely no background in statistics that was useful," Piotte told me ruefully. "Two guys, absolutely no clue.") But they soon discovered that the Netflix competition is a fairly collegial affair. The company hosts a discussion board devoted to the prize, and competitors frequently help one another out — discussing algorithms they've tried and publicly brainstorming new ways to improve their work, sometimes even posting reams of computer code for anyone to use. When someone makes a breakthrough, pretty soon every other team is aware of it and starts using it, too. Piotte and Chabbert soon learned the major mathematical tricks that had propelled the leading teams into the Top 10.

The first major breakthrough came less than a month into the competition. A team named Simon Funk vaulted from nowhere into the No. 4 position, improving upon Cinematch by 3.88 percent in one fell swoop. Its secret was a mathematical technique called singular value decomposition. It isn't new; mathematicians have used it for years to make sense of prodigious chunks of information. But Netflix never thought to try it on movies.

Singular value decomposition works by uncovering "factors" that Netflix customers like or don't like. Say, for example, that "Sleepless in Seattle" has been rated by 200,000 Netflix users. In one sense, this is just a huge list of numbers — user No. 452 gave it two stars; No. 985 gave it five stars; and so on. But you could also think of those ratings as individual reactions to various aspects of the movie. "Sleepless in Seattle" is a "chick flick," a comedy, a star vehicle for Tom Hanks; each customer is reacting to how much — or how little — he or she likes "chick flicks," comedies and Tom Hanks. Singular value decomposition takes the mass of Netflix data — 17,770 movies, ratings by 480,189 users — and automatically sorts the films. The programmers do not actively tell the computer what to look for; they just run the algorithm until it groups together movies that share qualities with predictive value.

Sometimes when you look at the clusters of movies, you can deduce the connections. Chabbert showed me one list: at the top were "Sleepless in Seattle," "Steel Magnolias" and "Pretty Woman," while at the bottom were "Star Trek" movies. Clearly, the computer recognized some factor that suggests that someone who likes the romantic aspect of "Pretty Woman" will probably like "Sleepless in Seattle" and dislike "Star Trek." Chabbert showed me another cluster: this time DVD collections of the TV show "Friends" all clustered at the top of the list, while action movies like "Reindeer Games" and thrillers like "Hannibal" clustered at the bottom. Most likely, the computer had selected for "comic" content here. Other lists appear to group movies based on whether they lean strongly to the ideological right or left.

As programmers extract more and more values, it becomes possible to draw exceedingly sophisticated

correlations among movies and hence to offer incredibly nuanced recommendations. “We’re teasing out very subtle human behaviors,” said Chris Volinsky, a scientist with AT&T in New Jersey who is one of the most successful Netflix contestants; his three-person team held the No. 1 position for more than a year. His team relies, in part, on singular value decomposition. “You can find things like ‘People who like action movies, but only if there’s a lot of explosions, and not if there’s a lot of blood. And maybe they don’t like profanity,’ ” Volinsky told me when we spoke recently. “Or it’s like ‘I like action movies, but not if they have Keanu Reeves and not if there’s a bus involved.’ ”

MOST OF THE LEADING TEAMS competing for the Netflix Prize now use singular value decomposition. Indeed, given how quickly word of new breakthroughs spreads among the competitors, virtually every team in the Top 10 makes use of similar mathematical ploys. The only thing that separates their scores is how skillfully they tweak their algorithms. The Netflix Prize has come to resemble a drag race in which everyone drives the same car, with only tiny modifications to the fuel injection. Yet those tweaks are crucial. Since the top teams are so close — there is less than a tenth of a percent between each contender — even tiny improvements can boost a team to the top of the charts.

These days, the competitors spend much of their time thinking deeply about the math and psychology behind recommendations. For example, the teams are grappling with the problem that over time, people can change how sternly or leniently they rate movies. Psychological studies show that if you ask someone to rate a movie and then, a month later, ask him to do so again, the rating varies by an average of 0.4 stars. “The question is why,” Len Bertoni said to me. “Did you just remember it differently? Did you see something in between? Did something change in your life that made you rethink it?” Some teams deal with this by programming their computers to gradually discount older ratings.

Another common problem is identifying overly punitive raters. If you’re a really harsh critic and I’m a much more easygoing one, your two-star rating may be equal to my four-star rating. To compensate, an algorithm might try to detect when a Netflix customer tends to hand out only one- or two-star ratings — a sign of a strict, pursed-lip customer — and artificially boost his or her ratings by a half-star or so. Then there’s the problem of movie raters who simply aren’t consistent. They might be evenhanded most of the time, but if they log into Netflix when they’re in a particularly bad mood, they might impulsively decide to rate a couple of dozen movies harshly.

TV shows, which are hot commodities on Netflix, present yet another perplexing issue. Customers respond to TV series much differently than they do to movies. People who loved the first two seasons of “The Wire” might start getting bored during the third but keep on watching for a while, then stop abruptly. So when should Cinematch stop recommending “The Wire”? When do you tell someone to give up on a TV show?

Interestingly, the Netflix Prize competitors do not know anything about the demographics of the customers whose taste they’re trying to predict. The teams sometimes argue on the discussion board about whether their predictions would be better if they knew that customer No. 465 is, for example, a 23-year-old woman in Arizona. Yet most of the leading teams say that personal information is not very useful, because it’s too crude. As one team pointed out to me, the fact that I’m a 40-year-old West Village resident is not very predictive. There’s little reason to think the other 40-year-old men on my block enjoy the same movies as I do. In contrast, the Netflix data are much more rich in meaning. When I tell Netflix that I think Woody Allen’s black comedy “Match Point” deserves three stars but the Joss Whedon sci-fi film “Serenity” is a

five-star masterpiece, this reveals quite a lot about my taste. Indeed, Reed Hastings told me that even though Netflix has a good deal of demographic information about its users, the company does not currently use it much to generate movie recommendations; merely knowing who people are, paradoxically, isn't very predictive of their movie tastes.

As the teams have grown better at predicting human preferences, the more incomprehensible their computer programs have become, even to their creators. Each team has lined up a gantlet of scores of algorithms, each one analyzing a slightly different correlation between movies and users. The upshot is that while the teams are producing ever-more-accurate recommendations, they cannot precisely explain how they're doing this. Chris Volinsky admits that his team's program has become a black box, its internal logic unknowable.

There's a sort of unsettling, alien quality to their computers' results. When the teams examine the ways that singular value decomposition is slotting movies into categories, sometimes it makes sense to them — as when the computer highlights what appears to be some essence of nerdiness in a bunch of sci-fi movies. But many categorizations are now so obscure that they cannot see the reasoning behind them. Possibly the algorithms are finding connections so deep and subconscious that customers themselves wouldn't even recognize them. At one point, Chabbert showed me a list of movies that his algorithm had discovered share some ineffable similarity; it includes a historical movie, "Joan of Arc," a wrestling video, "W.W.E.: SummerSlam 2004," the comedy "It Had to Be You" and a version of Charles Dickens's "Bleak House." For the life of me, I can't figure out what possible connection they have, but Chabbert assures me that this singular value decomposition scored 4 percent higher than Cinematch — so it must be doing something right. As Volinsky surmised, "They're able to tease out all of these things that we would never, ever think of ourselves." The machine may be understanding something about us that we do not understand ourselves.

Yet it's clear that something is still missing. Volinsky's momentum has slowed down significantly, as everyone else's has. There's some X factor in human judgment that the current bunch of algorithms isn't capturing when it comes to movies like "Napoleon Dynamite." And the problem looms large. Bertoni is currently at 8.8 percent; he says that a small group of mainly independent movies represents more than half of the remaining errors in the way of winning the prize. Most teams suspect that continuing to tweak existing algorithms won't be enough to get to 10 percent. They need another breakthrough — some way to digitally replicate the love/hate dynamic that governs hard-to-pigeonhole indie films.

"This last half-percent really is the Mount Everest," Volinsky said. "It's going to take one of these 'aha' moments."

SOME COMPUTER SCIENTISTS think the "Napoleon Dynamite" problem exposes a serious weakness of computers. They cannot anticipate the eccentric ways that real people actually decide to take a chance on a movie.

The Cinematch system, like any recommendation engine, assumes that your taste is static and unchanging. The computer looks at all the movies you've rated in the past, finds the trend and uses that to guide you. But the reality is that our cultural tastes evolve, and they change in part because we interact with others. You hear your friends gushing about "Mad Men," so eventually — even though you have never had any particular interest in early-'60s America — you give it a try. Or you go into the video store and run into a

particularly charismatic clerk who persuades you that you really, really have to give "The Life Aquatic With Steve Zissou" a chance.

As Gavin Potter, a Netflix Prize competitor who lives in Britain and is currently in ninth place, pointed out to me, a computerized recommendation system seeks to find the common threads in millions of people's recommendations, so it inherently avoids extremes. Video-store clerks, on the other hand, are influenced by their own idiosyncrasies. Even if they're considering your taste to make a suitable recommendation, they can't help relying on their own sense of what's good and bad. They'll make more mistakes than the Netflix computers — but they're also more likely to have flashes of inspiration, like pointing you to "Napoleon Dynamite" at just the right moment.

"If you use a computerized system based on ratings, you will tend to get very relevant but safe answers," Potter says. "If you go with the movie-store clerk, you will get more unpredictable but potentially more exciting recommendations."

Another critic of computer recommendations is, oddly enough, Pattie Maes, the M.I.T. professor. She notes that there's something slightly antisocial — "narrow-minded" — about hyperpersonalized recommendation systems. Sure, it's good to have a computer find more of what you already like. But culture isn't experienced in solitude. We also consume shows and movies and music as a way of participating in society. That social need can override the question of whether or not we'll like the movie.

"You don't want to see a movie just because you think it's going to be good," Maes says. "It's also because everyone at school or work is going to be talking about it, and you want to be able to talk about it, too." Maes told me that a while ago she rented a "Sex and the City" DVD from Netflix. She suspected she probably wouldn't really like the show. "But everybody else was constantly talking about it, and I had to know what they were talking about," she says. "So even though I would have been embarrassed if Netflix suggested 'Sex and the City' to me, I'm glad I saw it, because now I get it. I know all the in-jokes."

Maes suspects that in the future, computer-based reasoning will become less important for online retailers than social-networking tools that tap into the social zeitgeist, that let customers see, in Facebook fashion, for example, what their close friends are watching and buying. (Potter has an even more intriguing idea. He says he thinks that a recommendation system could predict cultural microtrends by monitoring news events. His research has found, for example, that people rent more movies about Wall Street when the stock market drops.) In the world of music, there are already several innovative recommendation services that try to analyze buzz — by monitoring blogs for repeated mentions of up-and-coming bands, or by sifting through millions of people's playlists to see if a new band is suddenly getting a lot of attention.

Of course, for a company like Netflix, there's a downside to pushing exciting-but-risky movie recommendations on viewers. If Netflix tries to stretch your taste by recommending more daring movies, it also risks annoying customers. A bad movie recommendation can waste an evening.

Is there any way to find a golden mean? When I put the question to Reed Hastings, the Netflix C.E.O., he told me he suspects that there won't be any simple answer. The company needs better algorithms; it needs breakthrough techniques like singular value decomposition, with the brilliant but inscrutable insights it enables. But Hastings also says he thinks Maes is right, too, and that social-networking tools will become

more useful. (Netflix already has one, in fact — an application that lets users see what their family and peers are renting. But Hastings admits it hasn't been as valuable as computerized intelligence; only a very small percentage of rentals are driven by what friends have chosen.) Hastings is even considering hiring cinephiles to watch all 100,000 movies in the Netflix library and write up, by hand, pages of adjectives describing each movie, a cloud of tags that would offer a subjective view of what makes films similar or dissimilar. It might imbue Cinematch with more unpredictable, humanlike intelligence.

"Human beings are very quirky and individualistic, and wonderfully idiosyncratic," Hastings says. "And while I love that about human beings, it makes it hard to figure out what they like."

Clive Thompson, a contributing writer for the magazine, writes frequently about technology.

Copyright 2008 The New York Times Company