The Atlantic

TECHNOLOGY

How Netflix Reverse-Engineered Hollywood

To understand how people look for movies, the video service created 76,897 micro-genres. We took the genre descriptions, broke them down to their key words … and built our own new-genre generator.

ALEXIS C. MADRIGAL JANUARY 2, 2014

The Atlantic's Netflix-Genre Generator

Feel-Good Assassination Sci-Fi Movies From the 1960s About Couples

If you use Netflix, you've probably wondered about the specific genres that it suggests to you. Some of them just seem *so specific* that it's absurd. Emotional Fight-the-System Documentaries? Period Pieces About Royalty Based on Real Life? Foreign Satanic Stories from the 1980s?

If Netflix can show such tiny slices of cinema to any given user, and they have 40 million users, how vast did their set of "personalized genres" need to be to

each and every microgenre that Netflix's algorithm has ever created.

Through a combination of elbow grease and spam-level repetition, we discovered that Netflix possesses not several hundred genres, or even several thousand, but 76,897 unique ways to describe types of movies.

There are so many that just loading, copying, and pasting all of them took the little script I wrote more than 20 hours.

We've now spent several weeks understanding, analyzing, and reverse-engineering how Netflix's vocabulary and grammar work. We've broken down its most popular descriptions, and counted its most popular actors and directors.

To my (and Netflix's) knowledge, no one outside the company has ever assembled this data before.

RECOMMENDED READING

The Internet Is Starting to Turn on MLMs

KAITLYN TIFFANY

Facebook Is a Doomsday Machine

ADRIENNE LAFRANCE

Dressed for the Plague. No, Not This One.

KAITLYN TIFFANY

What emerged from the work is this conclusion: Netflix has meticulously

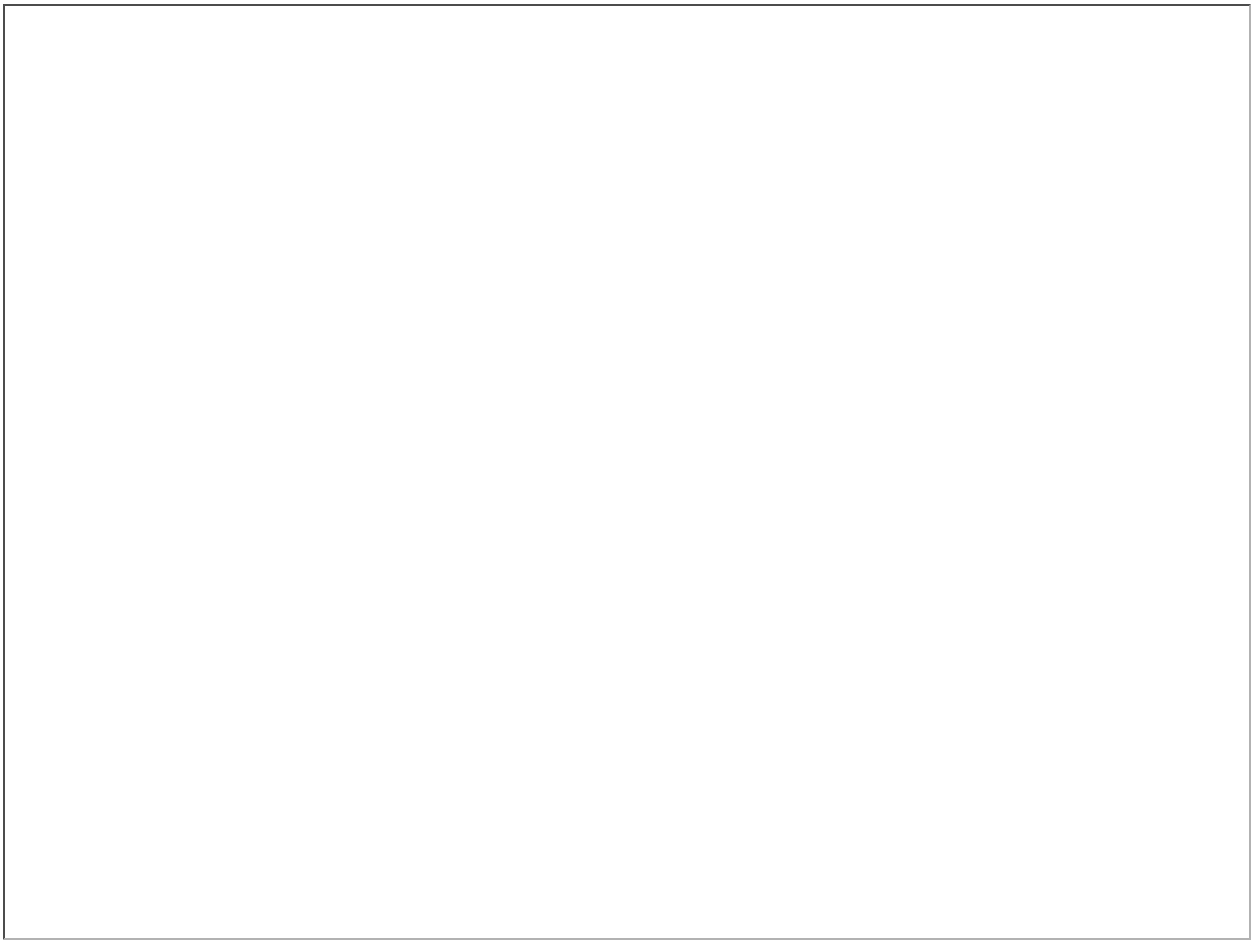
surface manifestation of this deeper database.

Netflix cooperated with my quest to understand what they internally call "altgenres," and made VP of product innovation Todd Yellin, the man who conceived of the system, available for an in-depth interview. Georgia Tech professor and *Atlantic* contributing editor, Ian Bogost, worked closely with me recreating the Netflix grammar, and he programmed the magical genre generator above.

If we reverse-engineered Yellin's system, it was Yellin himself who imagined a much more ambitious reverse-engineering process. Using large teams of people specially trained to watch movies, Netflix deconstructed Hollywood. They paid people to watch films and tag them with all kinds of metadata. This process is so sophisticated and precise that taggers receive a 36-page training document that teaches them how to rate movies on their sexually suggestive content, goriness, romance levels, and even narrative elements like plot conclusiveness.

They capture dozens of different movie attributes. They even rate the moral status of characters. When these tags are combined with millions of users viewing habits, they become Netflix's competitive advantage. The company's main goal as a business is to gain and retain subscribers. And the genres that it displays to people are a key part of that strategy. "Members connect with these [genre] rows so well that we measure an increase in member retention by placing the most tailored rows higher on the page instead of lower," the company revealed in a 2012 blog post. The better Netflix shows that it knows you, the likelier you are to stick around.

And now, they have a terrific advantage in their efforts to produce their own content: Netflix has created a database of American cinematic predilections. The data can't tell them *how* to make a TV show, but it can tell them *what* they should be making. When they create a show like *House of Cards*, they aren't guessing at what people want.



. Illustration by Darth.

Operation Scrape All the Data

This journey began when I decided I wanted a comprehensive list of Netflix microgenres. It seemed like a fun story, though one that would require some fresh thinking, as many other people had done versions of it.

I started on Twitter, asking my followers to submit the categories that showed up for them on Netflix to a shared document. "To my knowledge, no such list exists, but obviously one should," I wrote. "And then we can see what Netflix is really doing to us."

That call for help yielded about 150 genres, which seemed like a lot, relative to your average Blockbuster (RIP). But it was at that point that Sarah Pavis, a writer

number at the end of the web address.

That is to say, http://movies.netflix.com/WiAltGenre?agid=*1* linked to "African-American Crime Documentaries" and then http://movies.netflix.com/WiAltGenre?agid=*2* linked to" Scary Cult Movies from the 1980s." And so on.

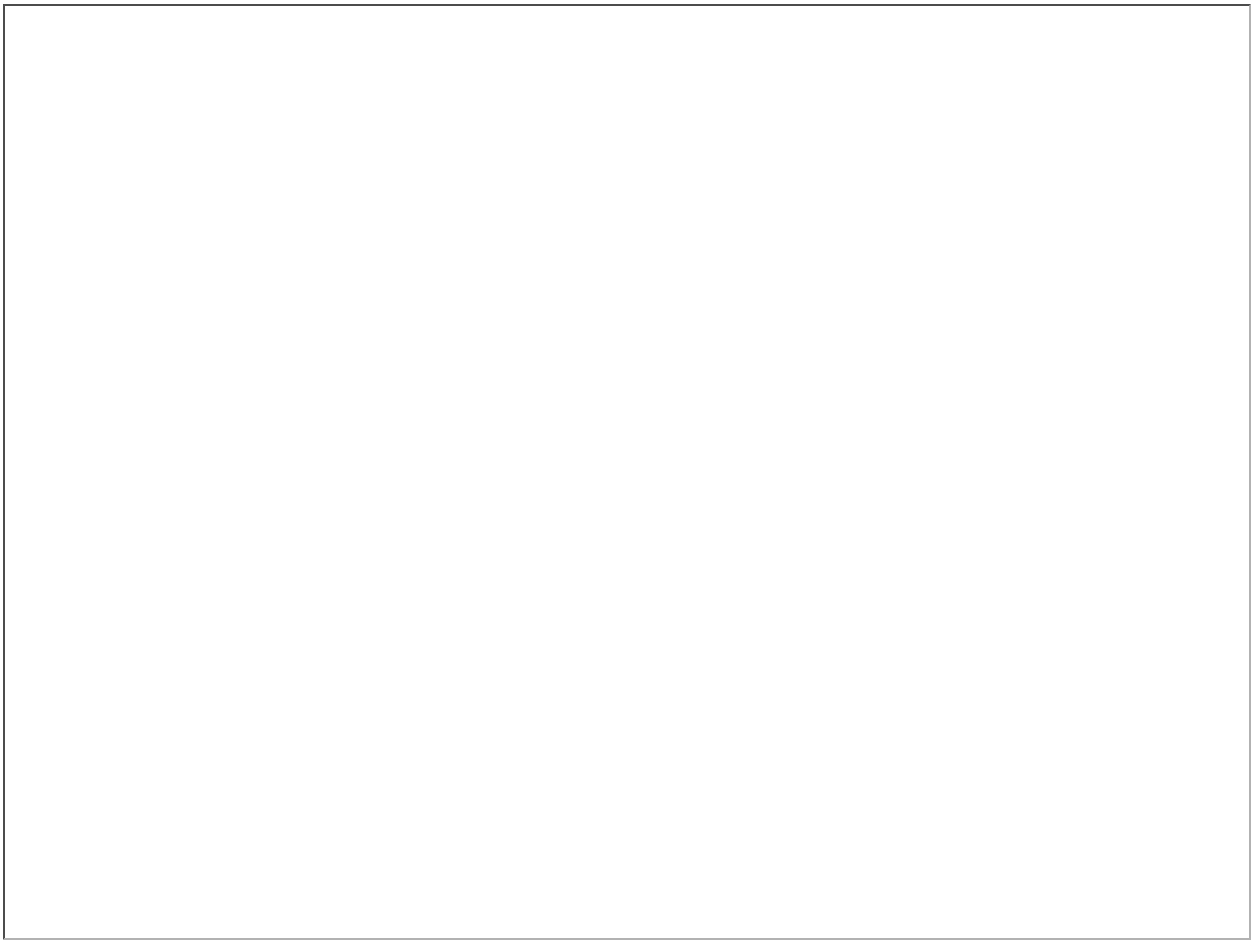
After walking through a few dozen URLs, I began to try out what seemed like arbitrarily high numbers. 1000: Movies directed by Otto Preminger. 3000: Dramas Starring Sylvester Stallone. 5000! Critically-Acclaimed Crime Movies from the 1940s. 20000! Mother-Son Movies from the 1970s. There were a lot of blanks in the data, but the entries extended into the 90,000s.

This database probing told me three things: 1) Netflix had an absurdly large number of genres, an order of magnitude or two more than I had thought, 2) it was organized in a way that I didn't understand, and 3) there was no way I could go through all those genres by hand.

But I also realized there was a way to scrape all this data. I'd been playing with an expensive piece of software called UBot Studio that lets you easily write scripts for automating things on the web. Mostly, it seems to be deployed by low-level spammers and scammers, but I decided to use it to incrementally go through each of the Netflix genres and copy them to a file.

After some troubleshooting and help from Bogost, the bot got up and running and simply copied and pasted from URL after URL, essentially replicating a human doing the work. It took nearly a day of constantly running a little Asus laptop in the corner of our kitchen to grab it all.

Imaginary



. Illustration by Darth.

As the software ran, I began to familiarize myself with the data. I randomly selected a snippet, so you can see what the raw genre data looks like:

> Emotional Independent Sports Movies
> Spy Action & Adventure from the 1930s
> Cult Evil Kid Horror Movies
> Cult Sports Movies
> Sentimental set in Europe Dramas from the 1970s
> Visually-striking Foreign Nostalgic Dramas
> Japanese Sports Movies
> Gritty Discovery Channel Reality TV
> Romantic Chinese Crime Movies

> Violent Suspenseful Action & Adventure from the 1980s
> Time Travel Movies starring William Hartnell
> Romantic Indian Crime Dramas
> Evil Kid Horror Movies
> Visually-striking Goofy Action & Adventure
> British set in Europe Sci-Fi & Fantasy from the 1960s
> Dark Suspenseful Gangster Dramas
> Critically-acclaimed Emotional Underdog Movies

The first thing that I noticed was that not every genre had streaming movies attached to it. The reason for that is the streaming catalog rotates and the genres that I was looking at represented the total possible universe of different genres, not just the ones that people were being shown on that particular day in this particular geography (the United States). So, right now, category 91,300, "Feel-good Romantic Spanish-Language TV Shows" doesn't show me anything I can stream. But category 91,307, "Visually Striking Latin American Comedies" has two movies and category 6,307, "Visually Striking Romantic Dramas" has 20.

So this is the main caveat to keep in mind as we go through this data: The existence of a genre in the database doesn't precisely correspond to the number of movies that Netflix has in its vaults. All the genre's existence means is that, based on an algorithm we'll get into later, there are some movies out there that fit the description.

As the thousands of genres flicked by on my little netbook, I began to see other patterns in the data: Netflix had a defined vocabulary. The same adjectives appeared over and over. Countries of origin also showed up, as did a larger-than-expected number of noun descriptions like Westerns and Slashers. There were ways of saying where the idea for the movie came from ("Based on Real Life" "Based on Classic Literature") and where the movies were set ("Set in Edwardian Era"). Of course, there were the various time periods, as well—from the 1980s, and so on—and references to children ("For Ages 8 to 10").

As the hours ticked by, the Netflix grammar—how it pieced together the words to form comprehensible genres—began to become apparent as well.

If a movie was both romantic *and* Oscar-winning, Oscar-winning always went to the left: *Oscar-winning* Romantic Dramas. Time periods always went at the end of the genre: Oscar-winning Romantic Dramas *from the 1950s*.

The single-word adjectives (such as romantic) could basically just pile up, though, at least to a point: Oscar-winning *Romantic Forbidden-Love* Movies.

And the content-area categories were generally tacked onto the end: Oscar-winning Romantic Movies *about Marriage*.

In fact, there was a hierarchy for each category of descriptor. Generally speaking, a genre would be formed out of a subset of these components:

> Region + Adjectives + Noun Genre + Based On… + Set In… + From the… + About… + For Age X to Y

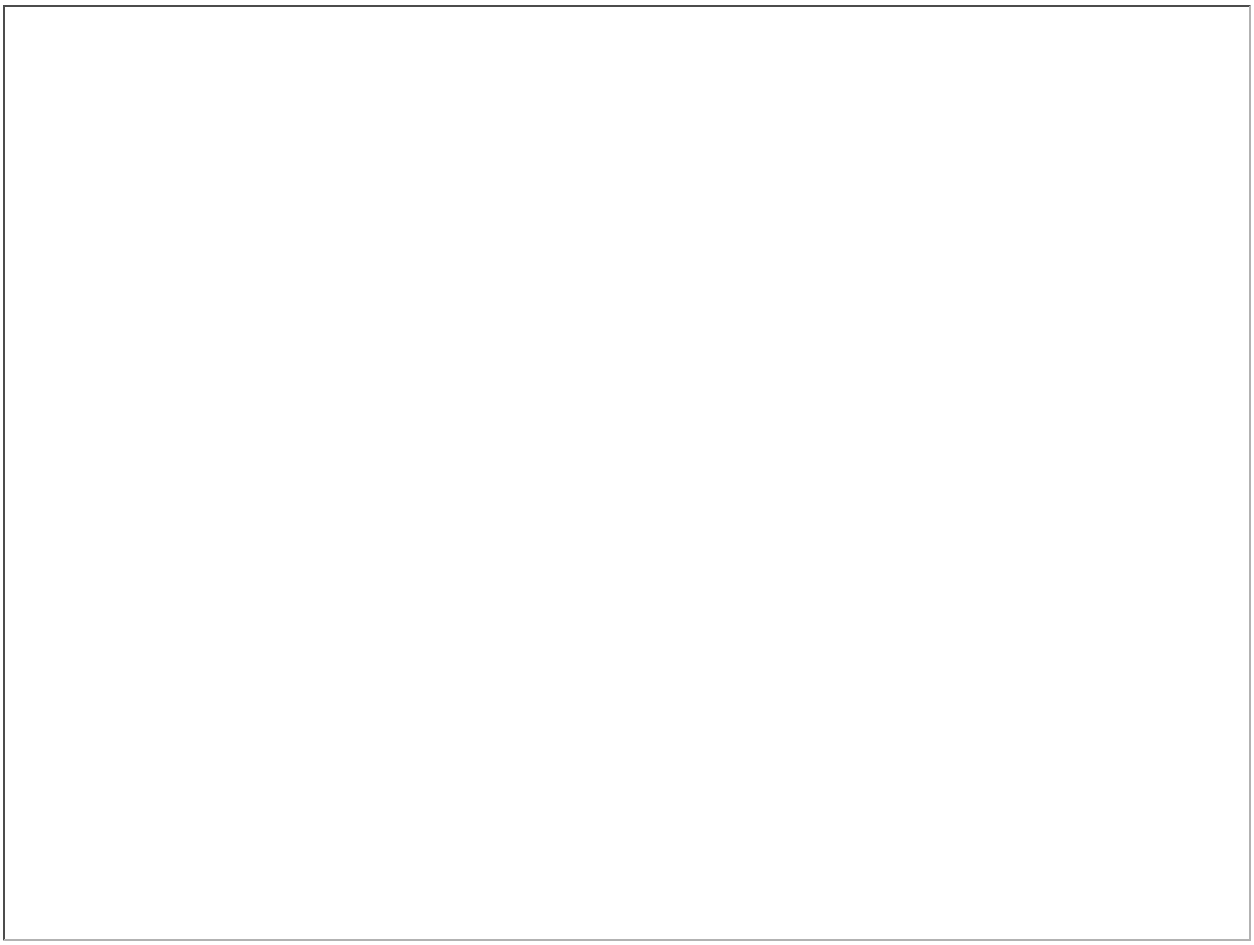
And, of course, there were all the genres that are for movies or TV shows starring or directed by certain individuals.

But that was it. All 76,897 genres that my bot eventually returned, were formed from these basic components. While I couldn't understand that mass of genres, the atoms and logic that were used to create them were comprehensible. I could fully wrap my head around the Netflix system.

I should note that the success of my bot had made me giddy by this point. A few Netflix categories put together are funny and intriguing. What could we do with 76,897 of them?!

And it was then that Ian Bogost, my colleague, suggested that we build the generator you see at the top of this article.

Imaginary



. Illustration by Darth.

Decoding Netflix's Grammar

To build a generator, however, our understanding of the grammar needed to get precise. I turned to another piece of software called AntConc, a freeware program maintained by a professor in Japan. It's generally used by linguists, digital humanities scholars, and librarians for dealing with corpuses, large amounts of text. If you've ever played with Google's Ngram tool, then you've seen at least one of the capabilities of AntConc.

text that forms Netflix's database, for example.

So, it becomes trivial to create a list of the top 10 ways that Netflix likes to describe movies in their personalized genres.

Or you can have it count the appearance of all 3-word phrases that begin with "from" and that would output the top decades in Netflix genres, with the 1980s rightfully and expectedly on top. When you're looking for an '80s movie, nothing else will do, you know?

By searching for phrases beginning with "Set in" I found all the locations mentioned in genres:

genre descriptions. Netflix has content "for kids" generally, as well as for ages 0 to 2, 0 to 4, 2 to 4, 5 to 7, 8 to 10, 8 to 12, and 11 to 12.

I took all of this data about Netflix's vocabulary and I created one large spreadsheet. Separately, I calculated the top actors, directors, and creators, and stashed those in a separate file.

Ian then took these spreadsheets and created several different grammars. The first and easiest method just lets lots of adjectives pile up and throws all the different descriptors into the mix very often. That's the *GONZO* setting in the generator. It outputs amazing stuff that you immediately want to copy and paste to your friends like:

- Deep Sea Father-and-Son Period Pieces Based on Real Life Set in the Middle East For Kids
- Assassination Bounty-Hunter Secret Society Dramas Based on Books Set in Europe About Fame For Ages 8 to 10
- Post-Apocalyptic Comedies About Friendship

Gosh, those are good, no? The second you read one, don't you just want that movie to exist? Can't you just imagine it? All that to say, Gonzo, for me, is *films that should exist* but won't. Or at least pitches that should exist and might soon.

Then, we scaled back the fun stuff, allowing only a few adjectives into the titles. Suddenly, we found ourselves staring at the extant movie-production logic of the Hollywood studios. Basically: endless recombination of the same few themes.

- Classic Action Movies
- Family-Friendly Westerns
- Buddy Period Pieces

That's the Hollywood button. (And that's Hollywood.)

- Raunchy Absurd Slashers
- Fight-the-System Political Love Triangle Mysteries
- Chilling Action Movies About Royalty

As we worked on the generator, I could tell someone had gone down this road before. A single human brain had had to make the decisions that we had. How many adjectives? How long should they be? And even more basic: what should the adjectives be? Why cerebral and not brainy? Why differentiate between gory and violent?

As a writer, I kept asking myself: why are the adjectives *just right*? Mind-bending and sandal-and-sword (you know, Conan!) and Twisty Tale and Rogue-Cop and Mad Scientist and Underdog and Feel-Good and Understated.

The words themselves were carefully chosen. By whom?

There were questions we still had, too. From a *Los Angeles Times* article, we knew the basics of tagging. But how did the tags relate to Netflix's "personalized genres"? What algorithm converted this mass of tags into precisely 76,897 genres?

If most people attempting to understand Netflix's genres were like the classic blind man trying to comprehend an elephant, I felt like I could see the front half of the beast, perhaps, but not the whole thing. I needed someone to explain the back end.

So, after I'd secured my data, I called up Netflix's PR liaison, a Dutch guy named Joris Evers who keeps a miniature windmill on his desk. I told him we had to talk.

After I filled him in on what we'd done, I waited to hear his reaction, wondering if I was about to have my Netflix account permanently canceled. Instead, he said, "And now you want to come in and talk to Todd Yellin, I guess?"

Yellin is Netflix's VP of Product and the man responsible for the creation of

guided the creation of all the systems.

Yes, of course I wanted to meet Yellin. He had become my Wizard of Oz, the man who made the machine, the human whose intelligence and sensibility I'd been tracking through the data.

At our interview, Yellin turned to me and said, "I've been waiting for someone to bubble up like this for years."

* * *

On the day I visited Netflix in Los Gatos, California, a lesser-known Silicon Valley town, there was a recycling center fire spewing toxins all across the Bay Area. The sky turned strange colors and the smell of burning plastic crept into one's nostrils.

Netflix is housed in a huge Italianate building that looks like a converted spa:

It feels oddly like a movie set, except everybody is doing the wrong thing, like if you showed up at a Universal Studios backlot and it turned out to be a branch office of Charles Schwab. They should be lounging by a pool, eating olives and drinking rose, but instead they're typing in vast and admirably adult rows of cubicles.

Yellin had some of the misplaced Hollywood feel, too. Intelligent, quick, and energetic, he feels like a producer, which makes sense as he's been, by his own accounting, "on all sides of the movie industry." Physically, he bears a remarkable resemblance to the actor Michael Kelly, who plays Doug Stamper, chief of staff to Frank Underwood (Kevin Spacey) in Netflix's original series *House of Cards*.

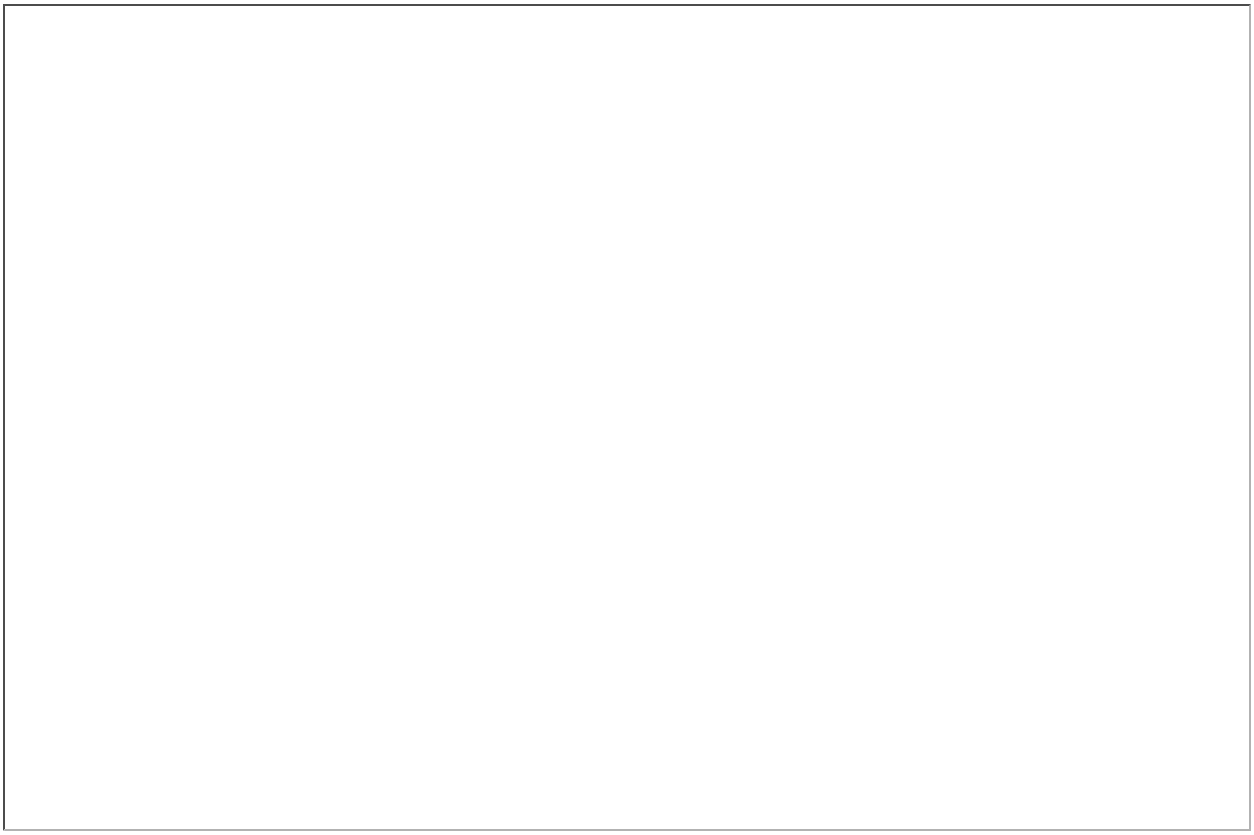
He seems like a guy who can make things work.

As we sit down in a conference room, I pull out my computer and begin to show off the genre generator we built. I walk him through my spreadsheets and show him all the text analysis we've done.

Though he seems impressed at our nerdiness, he patiently explains that we've merely skimmed one end-product of the entire Netflix data infrastructure. There is *so much more data* and a whole lot more intelligence baked into the system than we've captured.

Here's how he told me all the pieces fit together.

"My first goal was: tear apart content!" he said.



Todd Yellin at Netflix headquarters.

How do you systematically dismember thousands of movies using a bunch of different people who all need to have the same understanding of what a given microtag means? In 2006, Yellin holed up with a couple of engineers and spent months developing a document called "Netflix Quantum Theory," which Yellin now derides as "our pretentious name." The name refers to what Yellin used to call "quanta," the little "packets of energy" that compose each movie. He now prefers the term "microtag."

The Netflix Quantum Theory doc spelled out ways of tagging movie endings, the "social acceptability" of lead characters, and dozens of other facets of a movie. Many values are "scalar," that is to say, they go from 1 to 5. So, every movie gets a romance rating, not just the ones labeled "romantic" in the personalized genres. Every movie's ending is rated from happy to sad, passing through ambiguous.

That's the data at the base of the pyramid. It is the basis for creating all the altgenres that I scraped. Netflix's engineers took the microtags and created a syntax for the genres, much of which we were able to reproduce in our generator.

To me, that's the key step: It's where the human intelligence of the taggers gets combined with the machine intelligence of the algorithms. There's something in the Netflix personalized genres that I think we can tell is not fully human, but is *revealing* in a way that humans alone might not be.

For example, the adjective "feel good" gets attached to movies that have a certain set of features, most importantly a happy ending. It's not a direct tag that people attach so much as a computed movie category based on an underlying set of tags.

The only semi-similar project that I could think of is Pandora's once-lauded Music Genome Project, but what's amazing about Netflix is that its descriptions of movies are foregrounded. It's not just that Netflix can show you things you might like, but that it can tell you *what* kinds of things those are. It is, in its own weird way, a tool for introspection.

That distinguishes it from Netflix's old way of recommending movies to you, too. The company used to trumpet the fact that it could kind of predict how many stars you might give a movie. And so, the company encouraged its users to rate movie after movie, so that it could take those numeric values and develop a taste profile for you.

They even offered a $1 million prize to the team that could design an algorithm that would improve the company's ability to predict how many stars users would give movies. It took years to improve the algorithm by a mere 10 percent.

The prize was awarded in 2009, but Netflix never actually incorporated the new models. That's in part because of the work required, but also because Netflix had decided to "go beyond the 5 stars," which is where the personalized genres come in.

families and viral plagues. We wanted to put in more language," Yellin said. "We wanted to highlight our personalization because we pride ourselves on putting the right title in front of the right person at the right time."

And nothing highlights their personalization like throwing you a very, very specific altgenre.

So why aren't they ultraspecific, which is to say, super long, like the gonzo genres that our play generator can create?

Yellin said that the genres were limited by three main factors: 1) they only want to display 50 characters for various UI reasons, which eliminates most long genres; 2) there had to be a "critical mass" of content that fit the description of the genre, at least in Netflix's extended DVD catalog; and 3) they only wanted genres that made syntactic sense.

We ignore all of these constraints and that's precisely why our generator is hilarious. In Netflix's real world, there are no genres that have more than five descriptors. Four descriptors are rare, but they do show up for users: *Scary Cult Mad-Scientist Movies from the 1970s*. Three descriptors are more common: *Feel-good Foreign Comedies for Hopeless Romantics*. Two are widely used: *Steamy Mind Game* Movies. And, of course, there are many ones: *Quirky Movies*.

A fascinating thing I learned from Yellin is that the underlying tagging data isn't just used to create genres, but also to increase the level of personalization in all the movies a user is shown. So, if Netflix knows you love Action Adventure movies with high romantic ratings (on their 1-5 scale), it might show you that kind of movie, without ever saying, "Romantic Action Adventure Movies."

"We're gonna tag how much romance is in a movie. We're not gonna tell you how much romance is in it, but we're gonna recommend it," Yellin said. "You're gonna get an action row and it may have more or less romance in it based on what we know about you."

As Yellin talked, it occurred to me that Netflix has built a system that really only

serving you up filmed entertainment.

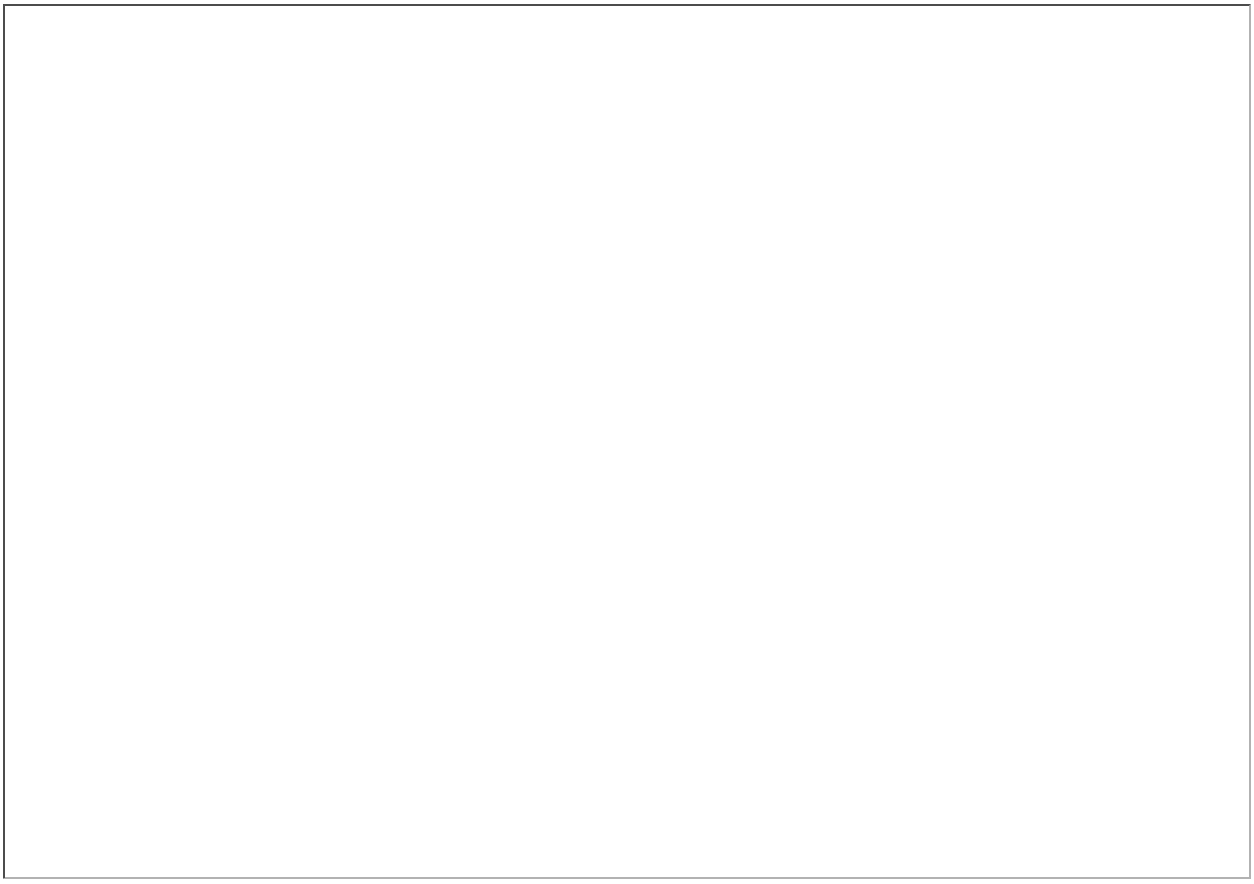
Which makes its hybrid human and machine intelligence approach that much more impressive. They could have purely used computation. For example, looking at people with similar viewing habits and recommending movies based on what they watched. (And Netflix does use this kind of data, too.) But they went beyond that approach to look at the *content itself*.

"It's a real combination: machine-learned, algorithms, algorithmic syntax," Yellin said, "and also a bunch of geeks who love this stuff going deep."

As a thought experiment: Imagine if Facebook broke down individual websites according to a 36-page tagging document that let the company truly understand what it was people liked about *Atlantic* or *Popular Science* or *4chan* or *ViralNova*?

It might be impossible with web content. But if Netflix's system didn't already exist, most people would probably say that it couldn't exist either.

The Perry Mason Mystery



Raymond Burr in *Please Murder Me*.

As our interview concluded, I pulled my computer back out and showed Yellin this one last chart. Take a good look at it. Something should stand out.

Sitting atop the list of mostly expected Hollywood stars is Raymond Burr, who starred in the 1950s television series *Perry Mason*. Then, at number seven, we find Barbara Hale, who starred opposite Burr in the show.

How can Hale and Burr outrank Meryl Streep and Doris Day, not to mention Samuel L. Jackson, Nicholas Cage, Fred Astaire, Sean Connery, and all these other actors in the top few dozen?

Raymond Burr
Bruce Willis
George Carlin
Jackie Chan
Andy Lau
Robert De Niro
Barbara Hale
Clint Eastwood
Gene Autry

Michael Caine
Roy Rogers
Sean Connery
Burt Reynolds
Charles Bronson
Dolph Lundgren
Harrison Ford
John Cusack
Ken Shamrock
Lance Henriksen

Tommy Lee Jones
Val Kilmer
Anderson Silva
Buster Keaton
Eric Roberts
Fred Williamson
Jean-Claude Van Damme
Michael Madsen
Mickey Rourke

Cary Grant
Elvis Presley
Fred Astaire
John Wayne
Rutger Hauer
Samuel L. Jackson
Steven Seagal
Sylvester Stallone
Smiley Burnette
Tom Berenger
Wesley Snipes

It's not that the list is nonsensical. That would be easy. We'd simply say: Netflix's actor-based genre-creation doesn't make much sense. But that's not the case at all. The rest of the actors at the top of the list make a lot of sense, even if it does not precisely reflect the top box-office earners.

Take a look at this list of the top 15 directors, too. Since you probably don't recognize his name, Christian I. Nyby II directed several *Perry Mason* made-for-TV movies in the 1980s. (His father, Christian I. Nyby, directed episodes of the original series, too!)

Christian I. Nyby II
Manny Rodriguez
Takashi Miike
Woody Allen
Ernst Lubitsch
Jim Wynorski
John Woo
Joseph Kane
Norman Taurog
Peter Jackson
Akira Kurosawa
Ingmar Bergman
R.G. Springsteen
Ridley Scott
Roger Corman

No, the strange thing is that these lists seem pretty spot-on, *except for this weird*

doesn't mean that Netflix users are having these movies pop up all the time. They are much more likely to get Action Movies Starring Bruce Willis.

But, then, why have all these genres?

> Mysteries starring Raymond Burr
> Movies starring Raymond Burr
> Dramas starring Raymond Burr
> Thrillers starring Raymond Burr
> Suspenseful Movies starring Raymond Burr
> Suspenseful Dramas starring Raymond Burr
> Cerebral Thrillers starring Raymond Burr
> Cerebral Dramas starring Raymond Burr
> Cerebral Suspenseful Dramas starring Raymond Burr
> Cerebral Mysteries starring Raymond Burr
> Cerebral Suspenseful Movies starring Raymond Burr
> Cerebral Movies starring Raymond Burr
> Murder Mysteries starring Raymond Burr
> Understated Movies starring Raymond Burr
> Understated Suspenseful Dramas starring Raymond Burr
> Understated Suspenseful Movies starring Raymond Burr
> Understated Mysteries starring Raymond Burr
> Understated Thrillers starring Raymond Burr
> Understated Dramas starring Raymond Burr

What was the deal? I asked Yellin.

Actually, I had a theory, which I told him. "In the DVD days, Perry Mason fans ordered a ton of Perry Mason, one after the other after the other," I said. "It created sufficient demand that you guys thought there should be categories."

That is not an accurate theory, Yellin told me. That's just not how it worked.

On the other hand, no one — not even Yellin — is quite sure why there are so

I tried on a bunch of different names for the Perry Mason thing: ghost, gremlin, not-quite-a-bug. What do you call the something-in-the-code-and-data which led to the existence of these microgenres?

The vexing, remarkable conclusion is that when companies combine human intelligence and machine intelligence, some things happen that we cannot understand.

"Let me get philosophical for a minute. In a human world, life is made interesting by serendipity," Yellin told me. "The more complexity you add to a machine world, you're adding serendipity that you couldn't imagine. Perry Mason is going to happen. These ghosts in the machine are always going to be a by-product of the complexity. And sometimes we call it a bug and sometimes we call it a feature."

Perry Mason episodes were famous for the reveal, the pivotal moment in a trial when Mason would reveal the crucial piece of evidence that makes it all makes sense and wins the day.

Now, reality gets coded into data for the machines, and then decoded back into descriptions for humans. Along the way, humans ability to understand what's happening gets thinned out. When we go looking for answers and causes, we rarely find that *aha!* evidence or have *the* Perry Mason moment. Because it all doesn't actually make sense.

Netflix may have solved the mystery of what to watch next, but that generated its own smaller mysteries.

And sometimes we call that a bug and sometimes we call it a feature.