CS320: Value of Data & AI **Winter 2020**

Session 2

Lecturer(s): Steve Eglash *Scribe(s): Susannah Shattuck*

2.1 Goals of the Session

The subject of today's class is an examination of how Netflix, the streaming video content company, uses recommendation data and algorithms to create value. Netflix uses a variety of different technical approaches, primarily for content recommendation and search, and there are distinct negative and positive consequences to these approaches and their use at Netflix's scale. We will explore these consequences and then try to generalize our observations to other companies to understand the impact of recommendation algorithms across a variety of industries and sectors.

2.2 Netflix: Background Information

Netflix is, today, one of the largest and most successful entertainment and media companies; it is the seventh largest internet company by revenue in the world. Netflix's primary business model is a subscription streaming service, and the company has 150 million subscribers and $19 billion in annual revenue generated from monthly fees. The use of this subscriber model, as opposed to the ad-based revenue model that is much more commonly seen among data-based internet companies, makes Netflix somewhat unique amongst its peers.

Netflix's biggest cost is acquiring and/or creating content for its streaming platform. It is important to understand the costs at any company that we are examining, so that we better understand the specific challenges and opportunities that the business faces—we will do this for each company observed during this course.

Initially, when Netflix launched in 1999, the company provided a mail-based DVD rental service. Over the course of the 2000s and 2010s, Netflix transitioned to providing streaming

content online, and it eventually discontinued the DVD-based portion of its business. Today, Netflix has transitioned again to becoming a major producer of new content, having developed and launched X shows since X date.

Each time Netflix's business model evolved, the company faced new opportunities and threats in the market. From the start, however, working with data and algorithms has been a critical component of their business. We will examine how Netflix's evolving use of customer and third party data has informed its position in the market as a viewing platform for movies and television shows.

2.2.1 Netflix's Goals

Today, Netflix's executives might say that the goal of the company is to become the largest producer and distributor of television shows and movies. In order for Netflix to succeed at this goal, the company needs to be able to offer engaging content to its subscribers who come to the company's website for content. A corollary goal to this primary goal is to keep those subscribers on the website without getting distracted and dropping off to another website or source of entertainment.

2.2.2 Netflix's Challenges & Advantages

There are several critical challenges that Netflix faces in pursuit of its primary business goals. One key challenge is streaming quality and the required supporting infrastructure; consumers today are increasingly intolerant of slow or poor quality video streaming, and Netflix has had to invest a lot into its streaming infrastructure to support high quality and fast performance.

Another key challenge is that consumers are fickle. A subscriber visiting Netflix's website at any given time may be in a specific mood or have a specific need that is different the moods or needs indicated in that same user's past behavior. Being able to provide relevant content recommendations for its users, even as their desires are constantly changing, is a critical challenge that Netflix must solve in order to be successful in its primary goal.

One of the advantages Netflix has in the face of these challenges and increasingly tight competition in the video streaming marketplace is that it knows a lot about consumer

preferences for movies and television shows. Its access to data on these consumer preferences is a competitive advantage—and this is often the case for data-driven internet companies.

Other media groups are starting to catch on to Netflix's advantage when it comes to consumer data. The relationship that Netflix has with competitors in the content creation space has evolved over time as the company's business model has evolved; initially, content creators viewed Netflix as a way to access customers. As the dynamics of the media streaming industry changed, however, and Netflix moved into the content creation business, those initial partners started to view Netflix as more of a competitor. Industry dynamics are constantly evolving across the entire market, and data-driven internet companies are often key contributors to those shifting dynamics.

2.2.3 Netflix's Stakeholders

Netflix is beholden to a variety of stakeholders as a business; to understand the challenges and advantages of their business model, it's important to examine who those key stakeholders are and what their interests and needs are. Netflix's primary stakeholders are its customers, the subscribers who pay for its streaming service—and these will be the focus of this class' discussion.

Netflix is also beholden to its shareholders, as any publicly traded company is. Another key group of stakeholders are independent producers of content. This group values their relationship with Netflix particularly strongly, because Netflix provides them access to a wider audience in a way that traditional movie theaters and television might not.

On the infrastructure side, Netflix is beholden to ISP providers and must work closely with this group to deliver content more quickly and at high quality. Infrastructure for a streaming company like Netflix is particularly intensive—a large percentage of the world's internet traffic is related to video streaming.

Finally, Netflix must continue to attract and retain another key group of stakeholders: its employees. Its competitive advantages in the market are dependent in part on its ability to attract the best developers, engineers, and researchers to continue to develop cutting-edge approaches to the big technical problems it has to solve to achieve its goals. These

problems include not only infrastructure challenges but also the development of recommendation and search algorithms.

2.3 Recommendation Systems At Netflix

Netflix's success is dependent upon their ability to serve content to subscribers such that those subscribers stay engaged with Netflix as a platform. Netflix creates value by recommending personalized content to subscribers; data and recommendation algorithms are the driving forces behind this capability and, ultimately, the company's entire business model.

2.3.1 Data Sources

Netflix relies on a diverse range of data sources to generate its personalized recommendations. Key data sources include:

- Star ratings—Netflix subscribers can rate movies or television shows based on a five-star system, and each subscriber thereby builds a personal profile of preferences as they rate movies and television shows that they have seen;
- Content metadata—each movie or television show has a set of metadata that provide relevant information about the genre, cast, director, etc., which can be relevant for making a recommendation to a subscriber;
- Users' content queues—before Netflix switched primarily to streaming content, when it was mailing DVDs to subscribers, each subscriber had the ability to maintain a queue of movies and television shows they would like shipped next to them, and that queue could be a source of data about individual preferences;
- Source of content discovery on Netflix's website—subscriber behavior on Netflix.com can be used as a source of data, including where on the website a subscriber clicked on a particular piece of content;
- The content—a movie or television show is a collection of data points in and of itself, and that data can be used to recommend "other movies with car chase scenes," for example;
- Publicly available third party data—other relevant data, such as critics' reviews and social media response, may be used by Netflix to make recommendations;

- Temporal data—Netflix can make relevant recommendations based on different pieces of temporal or time-based data, such as what time of year it is (i.e. recommending holiday movies in December);

One challenge that many of these data sources present is that it is difficult to generate recommendations for a customer before they have interacted with the platform and provided indicators of their personal taste. "Cold starting" a subscriber with content recommendations from the moment of their first login is difficult, and recommendations become better (and easier) as the subscriber interacts with more content on the platform over time.

2.3.2 Types of Algorithms Used

While the purpose of this session is not to get into the technical details of how each of these different approaches works, below is a list of some of the key recommendation algorithms that Netflix uses to recommend content to its subscribers:

- Personal Video Ranker (PVR)
- Top-N
- Trending
- Continue Watching
- Video-Video Similarity (Sims)
- Genre
- Page Generation
- Evidence Selection

These algorithms combine statistical methods with machine learning, with a combination of supervised and unsupervised machine learning. We will cover these approaches in greater detail later in the course.

2.3.3 Complications with Recommendation Algorithms

There are three major complications that these recommendation algorithms pose, which Netflix is attempting to solve for:

- "Cold starting" a new user—because these algorithms rely so heavily on user behavior and input to provide accurate recommendations, generating recommendations for a new user without limited input data is very difficult;
- Different users, different moods—many subscribers share a Netflix account, and so recommendations to that one account might not be relevant for all of the subscribers who are using it, and additionally, even an individual subscriber has different moods, and therefore might not be interested in a certain type of recommendations from one day to the next; accounting for idiosyncracies in people's taste and desires is difficult;
- Optimizing around the right goals—Netflix's ultimate goal is to keep subscribers engaged, but depending on which metrics they pick as a proxy for this big-picture goal, they may achieve very different outcomes, and some of those outcomes might not actually be what's best for the subscriber or society as a whole (Netflix's CEO once said that Netflix's biggest competitor is "people's sleep");

The first and second challenges become technical problems that Netflix is attempting to solve through a variety of different product features and choices. For example, subscribers now have the ability to set "profiles" for different users who are sharing the same Netflix account, so that it is easier to recommend relevant content to each individual user within their own profile. Cold starting has become easier as Netflix has had more trend-related data upon which to base its recommendations, because it can recommend content based on behavior it has seen from all of its users in aggregate.

The last challenge, however, is a much more fundamental one that does not have a simple technical fix. Netflix runs many different experiments in how to optimize its algorithms to maximize monthly subscriber retention. Netflix is not alone in having to determine how to optimize performance of its algorithms. Thinking about the unintended consequences of how we articulate why and what we are doing is a critical theme of this course.

2.3.4 The Importance of Recommendation Algorithms to Netflix Today

Are accurate content recommendations as critical to Netflix's business model today as they were in the early days of the company? Today, Netflix produces high quality content that brings subscribers to the platform as much as content made by other creators; however, they were only able to achieve this status as content creator because of their success as a content discovery and viewing platform through personalized recommendations.

The answer to this question—whether content recommendations are integral to Netflix today—is dependent in part on subscribers' behavior when they visit the platform. If the typical subscriber is using Netflix as more of a library than a content discovery tool, recommendations are far less important to those subscribers staying engaged on the platform than having high quality, relevant content available to stream is. Additionally, if most subscribers who visit Netflix's platform are interested in watching the same small subset of Netflix's available content, it becomes much less important to have highly personalized recommendations for each subscriber; Netflix could simply point everyone to the same subset of top-performing content.

Additionally, back when Netflix was mailing DVDs to subscribers, the cost of watching something was high; you had to place your order and wait for the movie to arrive in the mail. Now that the cost of watching something on Netflix is relatively cheap—you can decide what to watch and whether you like it all within a span of five minutes, rather than several days—does the importance of relevant recommendations go down?

As Netflix has switched from solely streaming to content creation, its subscribers' preference and behavior data becomes even more valuable as a tool for understanding what content to create, rather than what content to recommend. This data becomes a critical competitive advantage for Netflix in the highly competitive and margins-challenged entertainment industry. The question, remains, however, as to whether it makes sense for Netflix to pursue content creation as a primary business model and revenue driver. Though there is not time for discussion of this topic in today's session, this could be a good project topic for the class.

In addition to the business challenges of content creation, there is also the question of whether algorithmically-generated (or facilitated) art is interesting and valuable. If we are only fed content that is based on inputs and analysis from our own preferences, we will never be pushed to watch or consume media that challenges our understanding of the world. The professors related this a bit to eating sugar—it is pleasant in the moment, but without the long-term rewards that we all prefer in a deeper way. This challenge is not only relevant for Netflix but also for any internet company that is using data to recommend content, like Facebook.

2.3.5 The Netflix Prize

In 2006, Netflix launched a competition to improve its content recommendation service. Dubbed "The Netflix Prize," this competition offered $1 million to the team that could generate the best improved performance to Netflix's core recommendation system. The company was less than ten years old at the time, having been formed in 1997.

Netflix publicly released a dataset of 100 million movie ratings contained by 500,000 subscribers between 1999 and 2005. The company also withheld a different dataset of subscribers' movie ratings to be used as a competition qualifying set (i.e. a test set).

Internally, at the time Netflix was using the *Cinematch* algorithm, using correlation to recommend similar movies to subscribers who had rated a certain movie highly. It was a regression-based method, and the figures of merit were the Root Mean Squared Error (RMSE) of the system's prediction against the user's actual rating and system throughput.

Contestants were required to make predictions on the test dataset. Upon submission, Netflix automatically calculated the RMSE immediately for half of the qualifying dataset, so contestants could enter their formulas on a regular basis and get results (so everyone knew how every other team was doing as the competition progressed). This approach offered the threat that contestants could learn how to game the system by developing formulas that specifically performed well on the hidden part of the dataset that was contained within the qualifying set, so Netflix kept a third portion of the dataset hidden for testing the final submissions, which would determine the winner.

The advantage of Netflix's approach to providing immediate test results was that it created a sense of community and results-sharing as the competition progressed. Splitting the dataset into these Probe, Quiz, and Test subsets also ensured that the competition was fair and created an open, collaborative environment for researchers to learn more about how to build efficient recommendation systems.

The dataset itself that was released for this competition was highly precious. It was an MxN matrix—a 2D array where each row was a subscriber, and each column was a movie, with each cell representing a star rating provided by the subscriber for a given movie. This dataset was a sparse matrix, in the sense that most of the cells in the database were empty; most subscribers had only viewed (and rated) a small fraction of all of the available movies.

Almost every submission to the Netflix Prize did worse than the *Cinematch* algorithm, with the exception of just a handful of submissions. The Leaders submitted an algorithm that performed six to eight per cent better than Netflix's recommendation service on the RMSE metric. This winning submission saw improvement over time, improving from the fall of 2006 to the summer of 2007. This improvement graph was a step-wise function; there were several moments when improvement jumped up to become significantly better.

The winning solution included a linear blend of multiple results, stochastic gradient descent, matrix factorization, collaborative filtering and neighborhood-based approaches, singular value decomposition, and neural networks. In the 13 years since the Netflix Prize, technology has changed significantly, so there are many new approaches to solving this problem that did not appear in the winning solution.

2.3.6 The De-anonymization Problem

The Netflix Prize dataset was not robustly anonymized, and therefore researchers found that using other readily available data, including public comments from the Internet Movie Database (IMDB), they could easily de-anonymize the dataset to assign personal identities to individual subscribers. The amount of information they needed to accurately identify people in the Netflix dataset was shockingly small, revealing a critical problem in anonymized datasets.

Micro-data, commonly called data, are different from statistical data, which relies on a large population to calculate the mean, median, and other statistical measures. Micro-data includes the actual records of individuals, rather than generalizations across a population that are aggregated from statistical data. The Netflix Prize dataset is a good example of a dataset of micro-data. Other examples include transaction records, recommendations and ratings, web browsing behavior, and search histories.

K-anonymity is a common approach for micro-data anonymity, in which some attributes are made public, and some attributes are protected. This approach does not, however, guarantee any privacy. The robustness of anonymization is an important theme we will return to as we look at other companies and business models over the course of the quarter.

Netflix did not anticipate this de-anonymization problem when they released the dataset for the Netflix Prize. They said, in their FAQ for the competition, that there was no risk

of an anonymization issue, because they had anonymized the dataset; however, they were embarrassingly wrong on this front.

Targeted de-anonymization presents a serious threat to privacy. If you're interested in doing further reading and study in this area, the professors can provide recommended reading and project ideas in differential privacy and related areas. One big question is this area is the question of data ownership; when we use a service, we often are giving that service implicit permission to use the data we generate, but we have not provided explicit consent in many cases. We will devote significant discussion to this question over the course of the quarter.

2.4 Generalizing Our Observations to Other Sectors

Netflix uses aggregate views on customer movie preferences to price, acquire, and develop video content, with the goal of maximizing subscriber engagement and minimizing monthly subscriber churn. Some key insights that we have discussed in studying Netflix's use of data and algorithms towards these end include the value of inverting customer data to focus on movie/show-level data, the importance of capturing long-tail demand through relevant recommendations, and the use of data and algorithms to create customer value through personalization.

There are several media industry trends that are relevant to Netflix's story, as well, and those include the digitization of media and the importance of content creation and consumption. As digital media enables "cheaper" consumption of content through online streaming, companies like Netflix need to find better ways to compete against other streaming sites and services.

We can broaden many of these ideas to other industries and companies. Inverting or transforming data to create new products, services, and value is a common theme across almost every data-driven internet company. We have not yet talked about advertisement as a critical component of the media industry's data-driven business models, as Netflix uses a subscriber model to generate revenue, but these same questions of consumer engagement and recommendations are critical in a discussion of online advertising.

Netflix's ability to use an existing dataset—its subscriber preferences and behavior—to create value is another broadly applicable theme. Netflix is very good at turning its existing

data into a valuable set of outputs—recommendation algorithms—but not every company succeeds at this task. Why are some companies better at using existing data than others? Some critical elements that *are* present within Netflix but might not be at other companies include a quick and effective feedback loop (the star rating system), a data-driven culture within the company from day one, a speed of capturing and creating deliverables from data, and monopoly power.

Netflix is also a great example of how digital platforms can disrupt old industries. Netflix's digital-first advantage means that it has a more efficient data exchange and thereby value extraction capabilities than many of its entertainment industry competitors, like traditional movie studios. We can apply these lessons to other industries in thinking about how e-commerce is disrupting retail, mobile apps are disrupting transportation, and digital marketplaces are disrupting finance and education.

2.5 A Preview

In the next lecture, we will discuss how to build products and companies based around data, with a discussion of the company DataBricks. We will touch on the technology and concerns for managing data.

References

Gomez-Uribe, Carlos, N. Hunt. The Netfix Recommender System: Algorithms, Business Value, and Innovation. 2015. https://dl.acm.org/doi/pdf/10.1145/2843948

Walker, Russell. Monetizing Big Data through Productization and Data Inverting. *From Big Data to Big Profits: Success with Data Analytics*. 2015.