



# The NETFLIX Recommender System

## Algorithms, Business Value and Innovation



This is where the paper  
presentation begins





# I About The Paper

- The paper was published in 2015 by Carlos A. Gomez-Urbe and Neil Hunt from Netflix Inc.
- This paper covers up the Netflix Recommender System, its Algorithms, Business Value and Innovative effects.
- Netflix is a company with numerous huge competitors. Therefore, the paper does not reveal the exact mechanisms of the algorithms.



# | Introduction

- Netflix lies at the intersection of the Internet and storytelling. The main goal is to come up with a system that works as an Internet television.
- By the time that paper was published, Netflix had more than 65 million members who stream more than 100 million hours of movies and TV shows per day.
- The Netflix Recommender System does not rely on only one algorithm but rather a collection of different algorithms.



# I THE NETFLIX RECOMMENDER SYSTEM

Each of the algorithms in the recommender system relies on statistical and machine learning techniques.

## Algorithms

- Personalized Video Ranker: PVR
- Top-N Video Ranker
- Trending Now
- Continue Watching
- Video-Video Similarity
- Page Generation: Row Selection and Ranking
- Evidence Selection
- Search



Including both supervised (classification, regression) and unsupervised approaches (dimensionality reduction through clustering or compression)

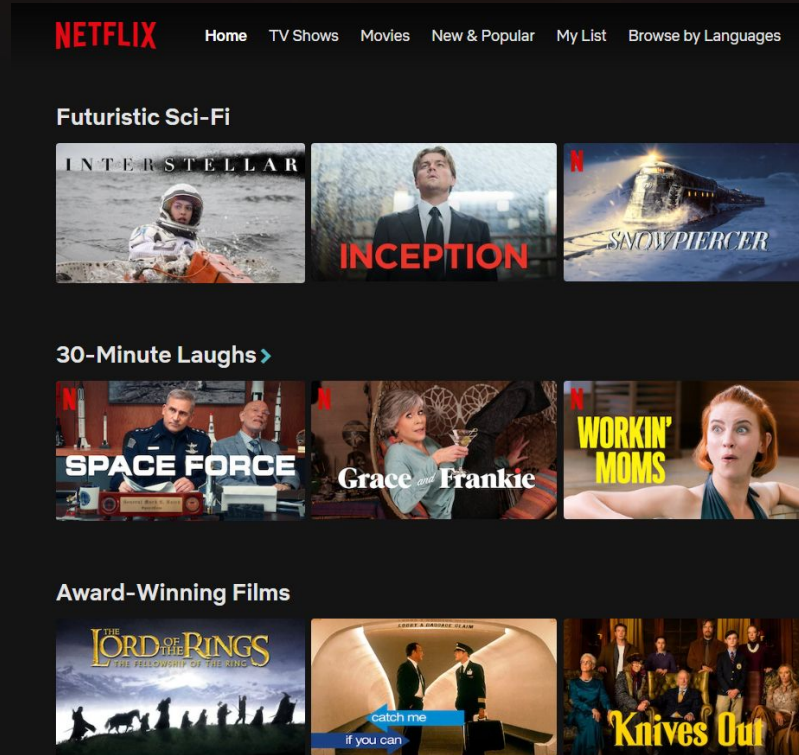


# | Personalized Video Ranker: PVR

- The videos in a given row typically come from a single algorithm.
- Because PVR algorithm is used widely by the Netflix Recommender System, it must be good at general-purpose relative rankings throughout the entire catalog; this limits how personalized it can actually be.
- PVR works better when personalized signals are blended with a pretty healthy dose of (unpersonalized) popularity.



# | Personalized Video Ranker: PVR



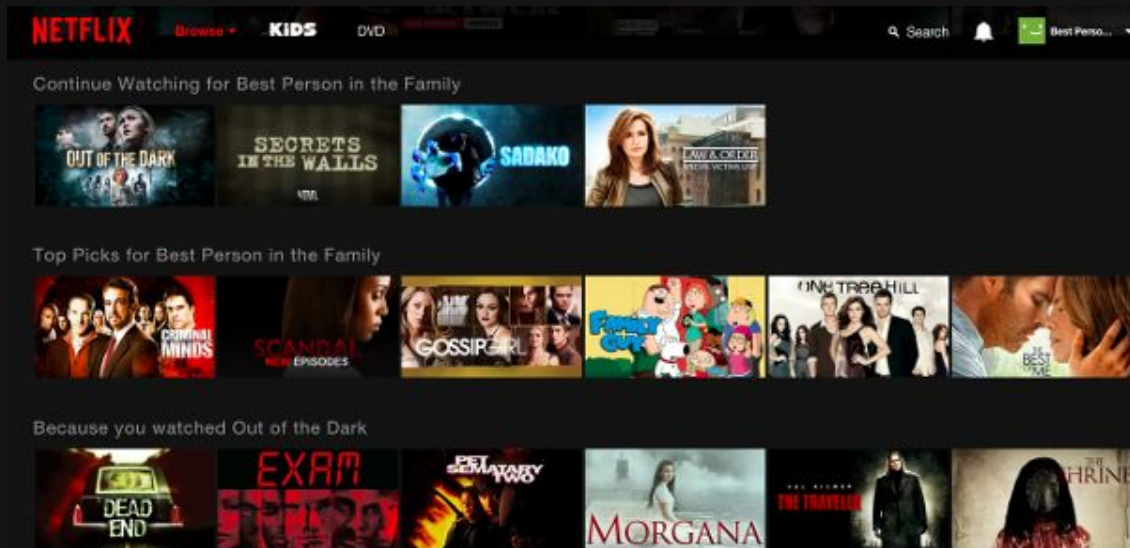


# | Top-N Video Ranker

- The goal of this algorithm is to find the best few personalized recommendations in the entire catalog for each member.
- The Top-N ranker is optimized and evaluated using metrics and algorithms that look only at the head of the catalog ranking that the algorithm produces, rather than at the ranking for the entire catalog (as is the case with PVR).



# | Top-N Video Ranker








# | Trending Now

- There are two types of trends that this ranker identifies nicely:
  - Those that repeat every several months (e.g., yearly) yet have a short term effect when they occur, such as the uptick of romantic video watching during Valentine's Day in North America.
  - One-off, short-term events, for example, a big hurricane with an impending arrival to some densely populated area, being covered by many media outlets, driving increased short-term interest in documentaries and movies about hurricanes and other natural disasters.





# | Trending Now






**NETFLIX** Home TV Shows Movies New & Popular My List Browse by Languages



**LITTLE WOMEN**

 Play  More Info

Trending Now



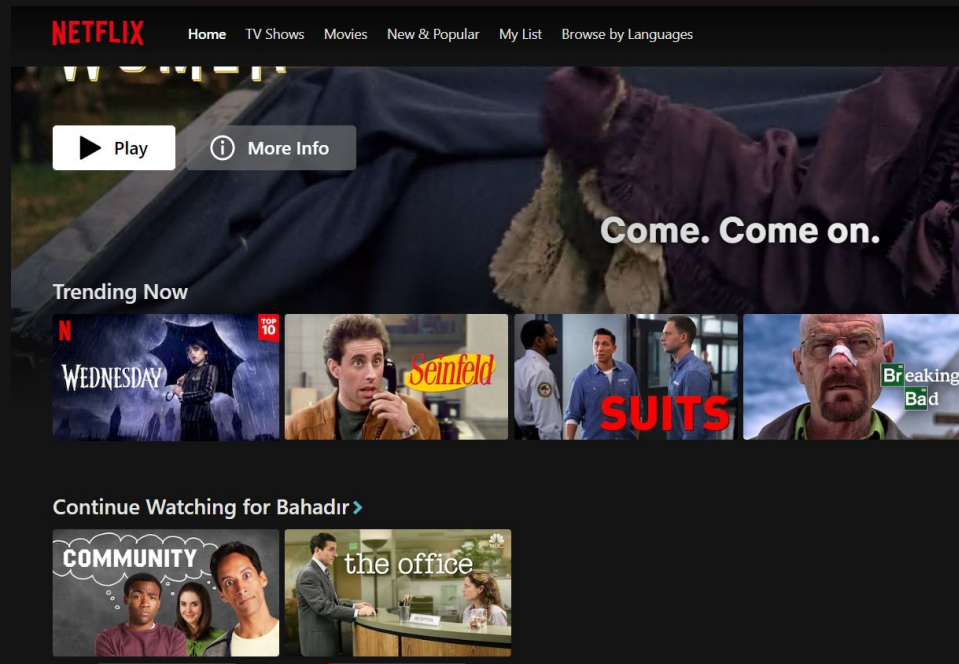


# I Continue Watching

- The Continue Watching ranker sorts the subset of the recently viewed titles based on their best estimate or whether the member intends to resume watching or rewatch, or whether the member has abandoned something not as interesting as anticipated.
- The signals that were used include:
  - The time elapsed since viewing,
  - The point of abandonment (mid-program vs. beginning or end),
  - Whether different titles have been viewed since, and
  - The devices used.



# | Continue Watching





# | Video-Video Similarity




- A Because You Watched (BYW) row anchors its recommendations to a single video watched by the member. The video-video similarity algorithm, which is referred to as 'sims', drives the recommendations in these rows.
- The sims algorithm is an unpersonalized algorithm that computes a ranked list of videos. -the similars- for every video in the catalog.
- Even though the sims ranking is not personalized, the choice of which BYW rows make it onto a homepage is personalized.






# I Video-Video Similarity

**NETFLIX** Home TV Shows Movies New & Popular My List Browse by Languages




**Because you watched The Ranch**



**Kids & Family Movies**



**Futuristic Sci-Fi**





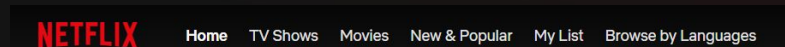
# | Page Generation: Row Selection and Ranking

- The page generation algorithm uses the output of all the algorithms already described to construct every single page of recommendations, taking into account the relevance of each row to the member as well as the diversity of the page.
- Before 2015, a rule-based approach that would define what type of row (e.g., genre row, BYW row) would go in each vertical position of the page was being used. This page layout was used to construct all homepages for all members. Today, a fully personalized and mathematical algorithm that can select and order rows from a large pool of candidates to create an ordering optimized for relevance and diversity is being used.





# | Page Generation: Row Selection and Ranking



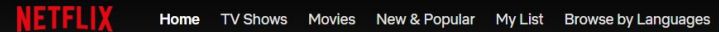
## Because you watched The Ranch



## Kids & Family Movies



## Futuristic Sci-Fi



## Futuristic Sci-Fi



## 30-Minute Laughs >



## Award-Winning Films







# | Evidence Selection

- An evidence can be thought as all the information of the content shown on the top left of the page, including:
  - Predicted star rating,
  - The synopsis,
  - Awards
  - Cast
  - Metadata
  - Images
- Evidence selection algorithms evaluate all the possible evidence items that can be displayed for every recommendation, to select the few that will be most helpful to the member viewing the recommendation.



# | Evidence Selection

**NETFLIX ORIGINAL**  
**STRANGER THINGS**

95% Match 2017 2 Seasons 4K Ultra HD 5.1

When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces and one strange little girl.

*Winona Ryder, David Harbour, Matthew Modine*  
TV Shows, TV Sci-Fi & Fantasy, Teen TV Shows



### Popular on Netflix



### Recently Watched





# | Search






- The search experience is built around several algorithms:
  - One algorithm attempts to find the videos that match a given query, for example, to retrieve **Frenemies** for the partial query “**fren**”.
  - Another algorithm predicts interest in a concept given a partial query, for example identifying the concept **French Movies** for the query “**fren**”.
  - A third algorithm finds video recommendations for a given concept, for example, to populate the videos recommended under the concept French Movies.
- The search algorithms combine:
  - play data,
  - search data,
  - metadatato arrive at the results and the recommendations being offered.



# | Search

usual

Titles related to The Usual Suspects



fren



Titles related to French Movies



French Stewart  
Katie French  
Edward Lewis French  
Dawn French  
Patricia French  
Harold French

Back

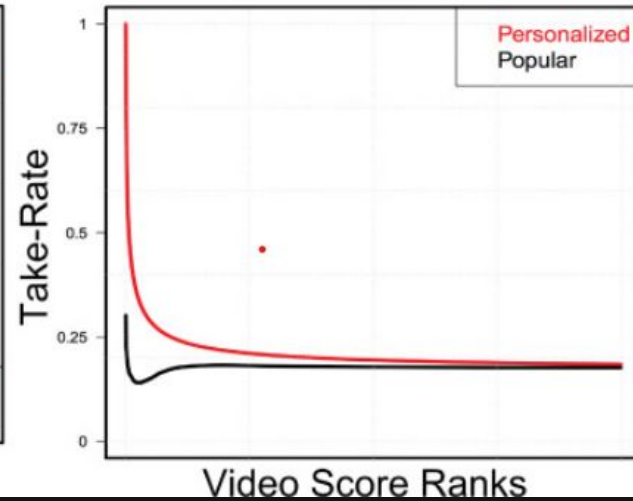
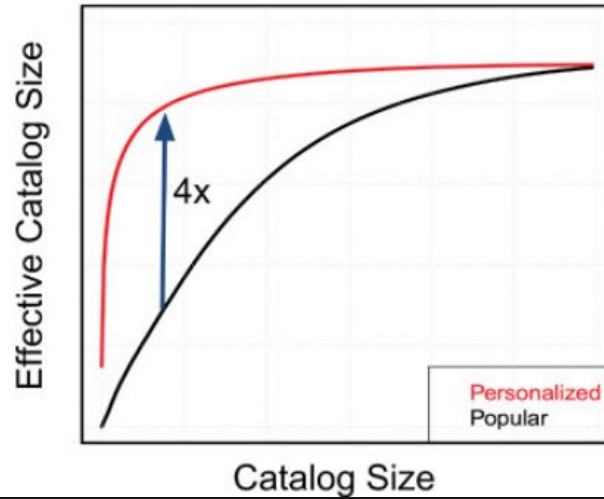


# | BUSINESS VALUE

- The Netflix Recommender System helps them win **moments of truth**: when a member starts a session and they try to help that member find something engaging within a few seconds, preventing abandonment of their service for an alternative entertainment option.
- The effective catalog size (ECS) is a metric that describes how spread viewing is across the items in the catalog. This metric is being used to describe the effects of personalization.
- One metric that gets at this is the take- rate—the fraction of recommendations offered resulting in a play.



# | BUSINESS VALUE





# | A/B TESTING

A/B tests are randomized, controlled experiments, to compare the medium-term engagement with Netflix along with member cancellation rates across algorithm variants. Algorithms that improve these A/B test metrics are considered better.

- Choosing Metrics for A/B Testing
- Test Cell Sizes for Statistical Validity
- Nuances of A/B Testing
- Alternative Metrics
- Test Audience
- Faster Innovation Through Offline Experiments
- Estimating Word-of-Mouth Effects



# I Choosing Metrics for A/B Testing

- Revenue is proportional to the number of members, and three processes directly affect this number:
  - The acquisition rate of new members,
  - Member cancellation rates, and
  - The rate at which former members rejoin
- A/B tests randomly assign different members to different experiences that referred to as **cells**. For example, each cell in an A/B test could map to a different video similars algorithm, one of which reflects the default (often called “production”) algorithm to serve as the **control cell** in the experiment—other cells in the test are the **test cells**. Then let the members in each cell interact with the product over a period of months, typically 2 to 6 months. Finally, the resulting data is analyzed to answer several questions about member behavior from a statistical perspective.





# | Test Cell Sizes for Statistical Validity

- Statistics are used as a guide to whether they have enough data to conclude that there is a difference in an A/B test metric across cells.
- As an example, suppose that we find that after two months, a fraction  $p_c$  and  $p_t$  of members in the control and test cell of an A/B test with 2 cells are still Netflix members, with  $\Delta = p_t - p_c > 0$ . Intuitively, we should trust the observed delta more the more members we have in the test. But how many members are enough to trust the test result?



# | Test Audience

- We typically test algorithm changes on two groups of members:
  - Existing members
  - New members
- Netflix prefer to test on new members because they have not experienced a different version of the product before; thus, their responses tend to be indicative of the effectiveness of the alternative versions of the algorithm rather than the change from old to new, yielding cleaner measurements.
- A disadvantage is that they have fewer new members, only as many signups as they get during the time period when they allocate new members into a test. Another disadvantage is that they offer new members a one-month-free trial, so they see few cancellations before this free month expires and cannot measure accurate retention rates until one month after the last new member in the test joined Netflix.

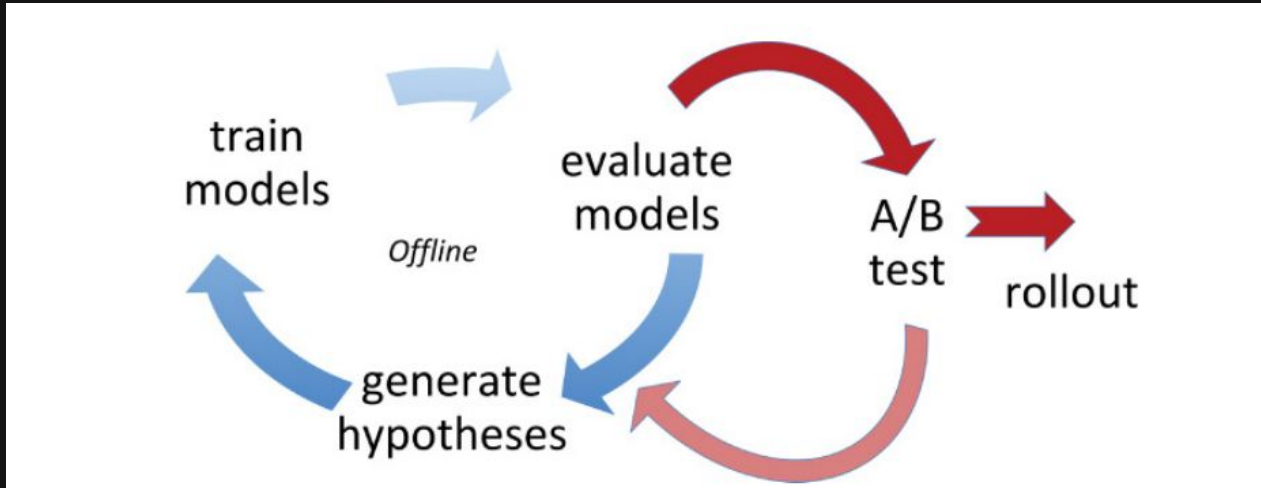


# | Faster Innovation Through Offline Experiments

- The time scale of the A/B tests might seem long, especially compared to those used by many other companies to optimize metrics, such as click-through rates. This is partly addressed by testing multiple variants against a control in each test; thus, rather than having two variants, A and B, they typically include 5 to 10 algorithm variants in each test.
- To really speed up innovation, they also rely on a different type of experimentation based on analyzing historical data. This offline experimentation changes from algorithm to algorithm, but it always consists of computing a metric for every algorithm variant tested that describes how well the algorithm variants fit previous user engagement.



# | Faster Innovation Through Offline Experiments





# | Estimating Word-of-Mouth Effects

- Improving the experience for members might be expected to generate stronger word-of-mouth; this, by definition, has influence beyond the boundaries of an A/B test cell, thus is hard to measure.



# I KEY OPEN PROBLEMS

Some of their main current open problems revolve around A/B testing, others around the recommendation algorithms themselves.

- Better Experimentation Protocols
- Global Algorithms
- Controlling for Presentation Bias
- Page Construction
- Member Coldstarting
- Account Sharing
- Choosing the Best Evidence to Support Each Recommendation



# | Better Experimentation Protocols

- They want to have a better alternative to offline experimentation that allows us to iterate just as quickly, but that is more predictive of A/B test outcomes.
- Another possibility is developing new offline experiment metrics that are more predictive of A/B test outcomes.
- They are also interested in general improvements to their A/B testing, for example, effective variance reduction methods to conduct experiments with higher resolution and fewer noisy results, or new A/B engagement metrics that are even more highly correlated with retention rates.



# | Global Algorithms

- Today, they group countries into regions that share very similar catalogs, yet have a big enough member base that generates enough data to fit all the necessary models.
- They then run copies of all of their algorithms isolated within each region. Rather than scaling this approach as they offer their service around the world, they are developing a single global recommender system that shares data across countries.





# I Controlling for Presentation Bias

- They have a system with a strong positive feedback loop, in which videos that members engage highly with are recommended to many members, leading to high engagement with those videos, and so on. Yet, most of their statistical models, as well as the standard mathematical techniques used to generate recommendations, do not take this feedback loop into account. In their opinion, it is very likely that better algorithms explicitly accounting for the videos that were actually recommended to their members, in addition to the outcome of each recommendation, will remove the potential negative effects of such a feedback loop and result in better recommendations.



# | Member Coldstarting

- Netflix Recommender System does a satisfactory job helping members with a large Netflix history, but not so for new members, about whom they know little.
- Thus, they are always interested in finding better models and signals to improve the recommendations for new members, to increase their engagement and their retention rates.
- Today, their member coldstart approach has evolved into a survey given during the sign-up process, during which they ask new members to select videos from an algorithmically populated set that they use as input into all of their algorithms.



# | Account Sharing

- A large percentage of profiles are used by multiple people in the household. Their recommender system has, by necessity, evolved through years of A/B testing to deliver a mix (union) of suggestions necessary to provide good suggestions to whichever member of the household may be viewing (owner, spouse, children) at any time, but such amalgamated views are not as effective as separated views.



# I CONCLUSION

- They have described the different algorithms that make up the Netflix recommender system, the process that we use to improve it, and some of our open problems.
- The recommender systems can democratize access to long-tail products, services, and information, because machines have a much better ability to learn from vastly bigger data pools than expert humans, thus can make useful predictions for areas in which human capacity simply is not adequate to have enough experience to generalize usefully at the tail.



**THANK YOU!**