ICERM

Mathematics in Data Science Workshop

NETFLIX

Personalized Page Generation using Data, Science, and Algorithms

Justin Basilico
Page Algorithms Engineering

July 30, 2015

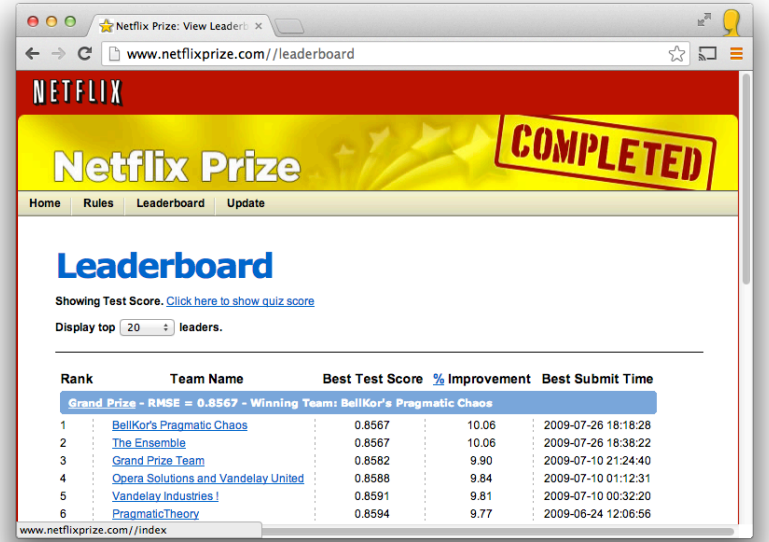
@JustinBasilico

NETFLIX

Netflix Prize

COMPLETED

- Interested in high-quality recommendations
- Proxy question:
  - Accuracy in predicted rating
  - Measured by root mean squared error (RMSE)
  - Improve by 10% = $1 million!
- Data size:
  - 100M ratings (back then "almost massive")

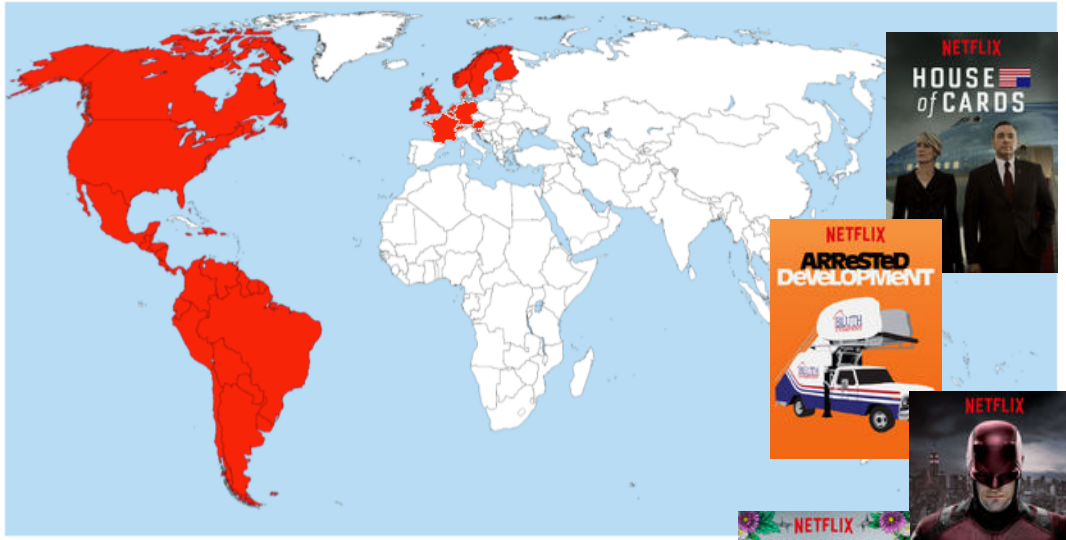


$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

NETFLIX

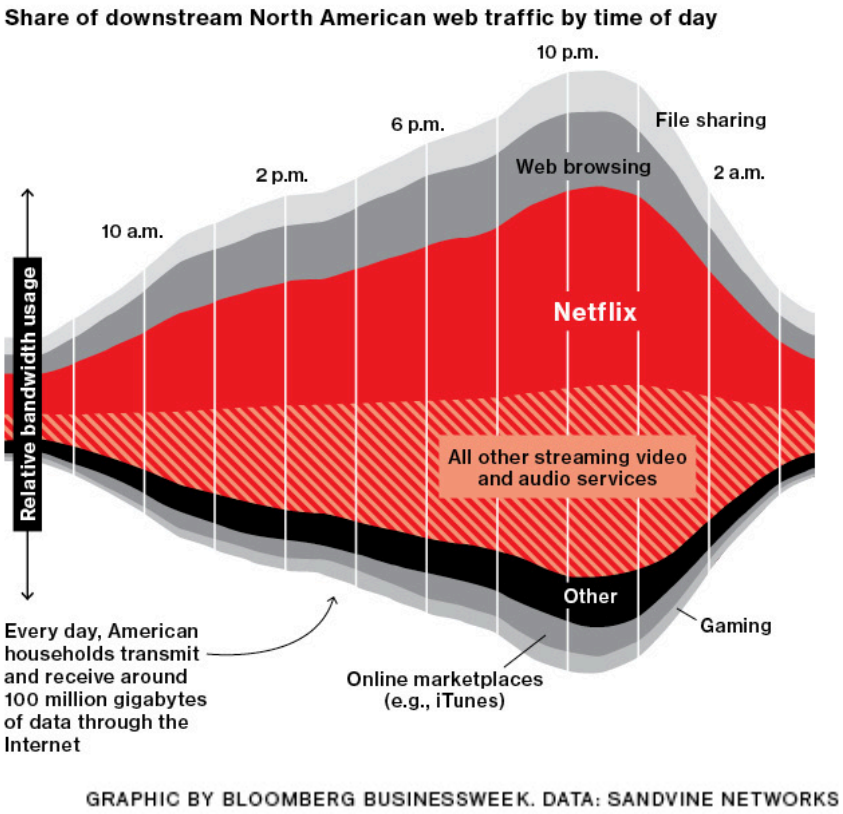
Change of focus

2006



2015

Netflix Scale



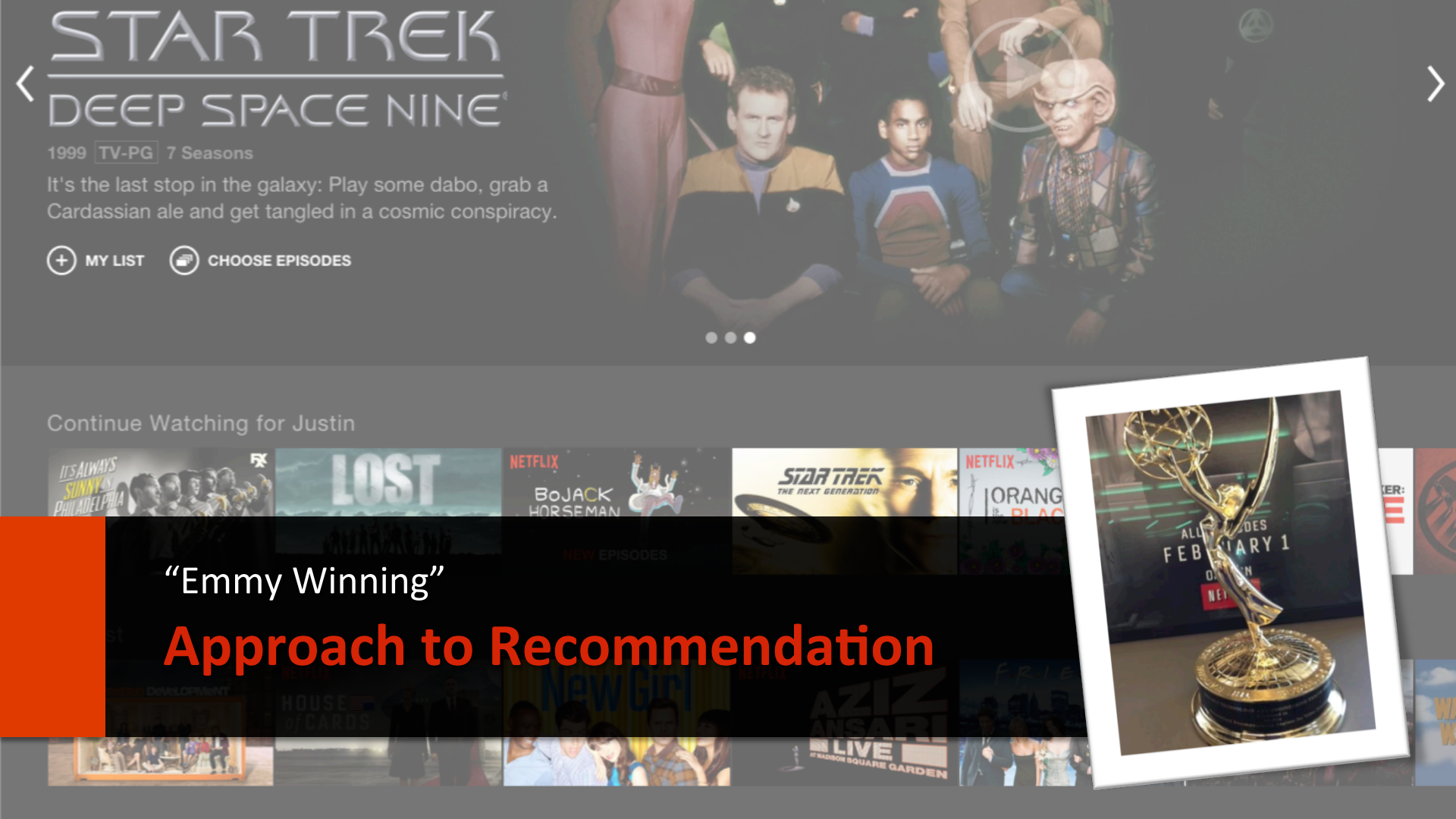
- > 65M members
- > 50 countries
- > 1000 device types
- Hours: > 3B/month
- Plays: > 100M/day
- Log 100B events/day
- 36.5% of peak US downstream traffic

Where we use Data Science?

- **Personalization**
- Search
- *Product experimentation*
- Content buying
- Content delivery
- Streaming quality
- Marketing effectiveness
- Predicting retention
- Fraud detection
- …

Data-driven Company

NETFLIX



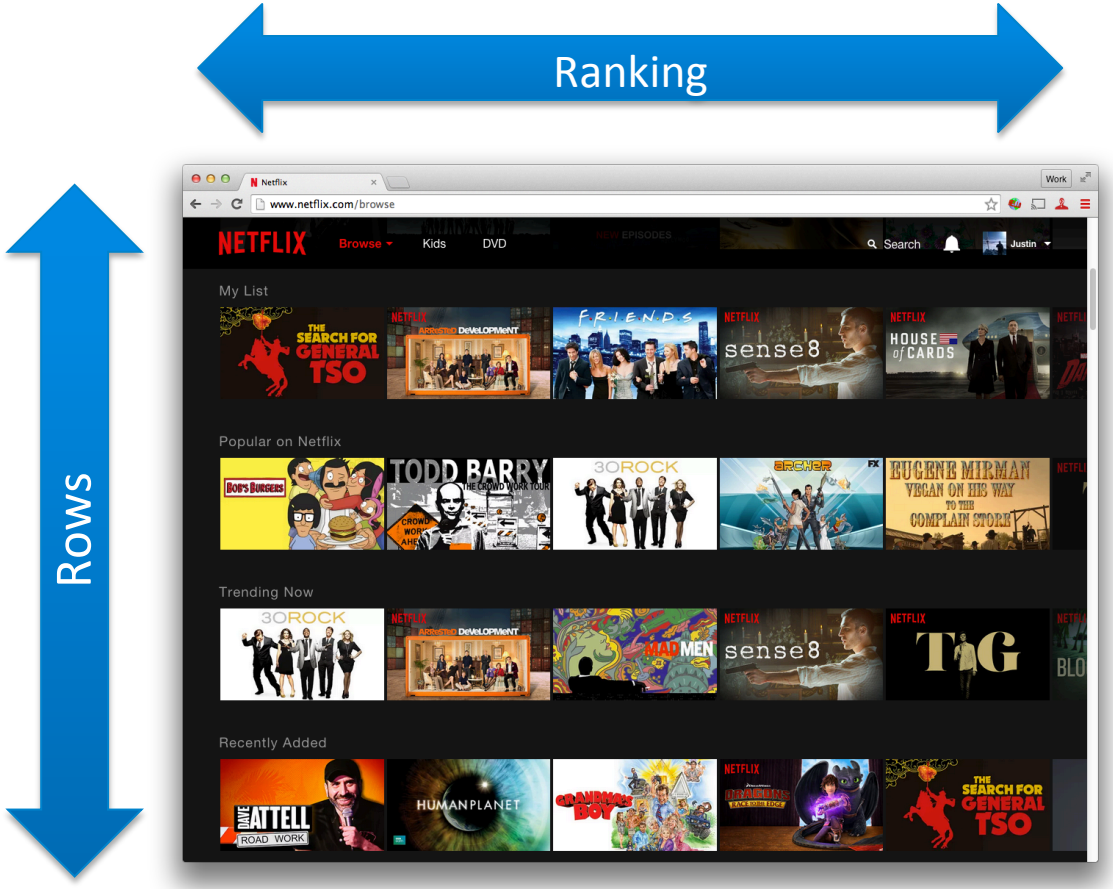
"Emmy Winning"

Approach to Recommendation

Goal

Help members find content to watch and enjoy to maximize member satisfaction and retention

Everything is a Recommendation

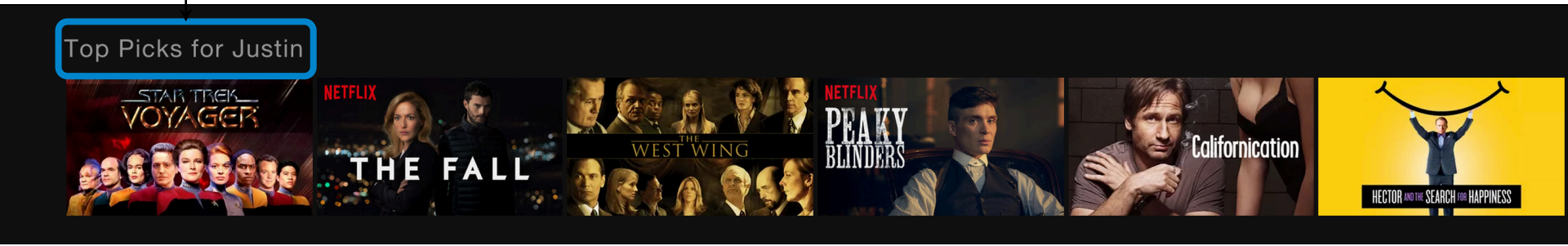


Over 80% of what people watch comes from our recommendations

Recommendations are driven by **Machine Learning**

Top Picks

Personalization awareness

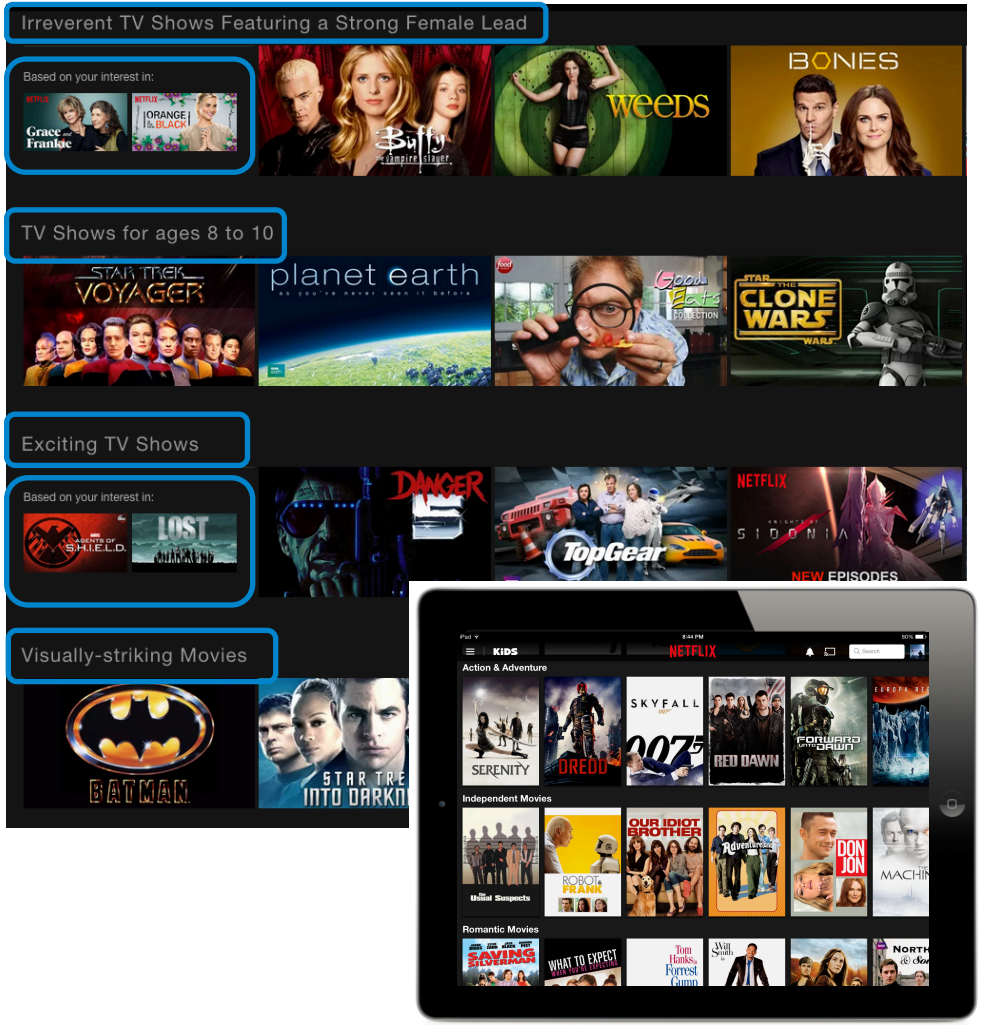


Diversity

Personalized genres

- **Genres** focused on user interest
  - Derived from tag combinations
  - Provide context and evidence

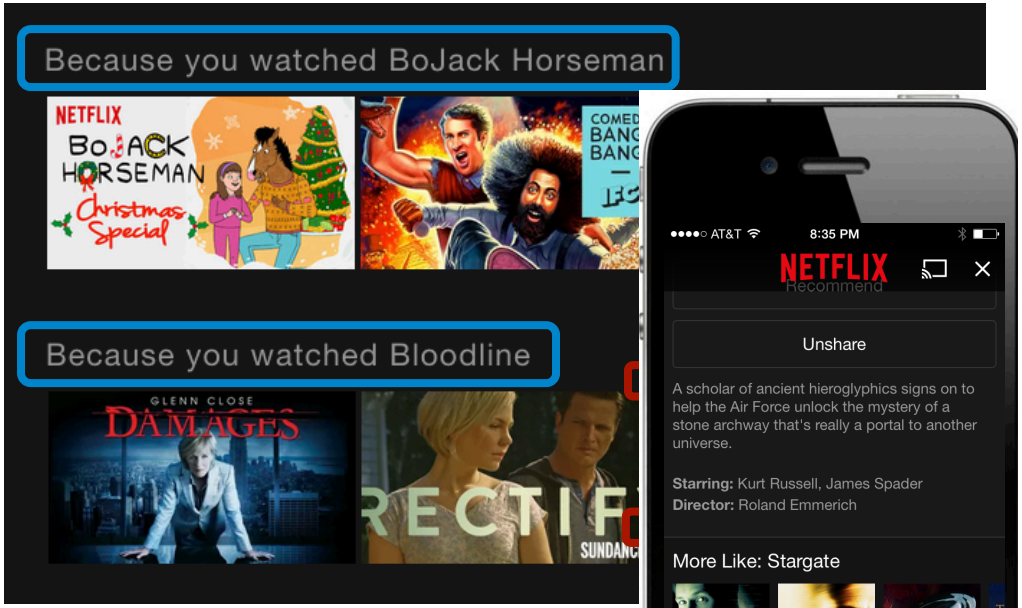
- How are they generated?
  - **Implicit**: Based on recent plays, ratings & other interactions
  - **Explicit:** Taste preferences

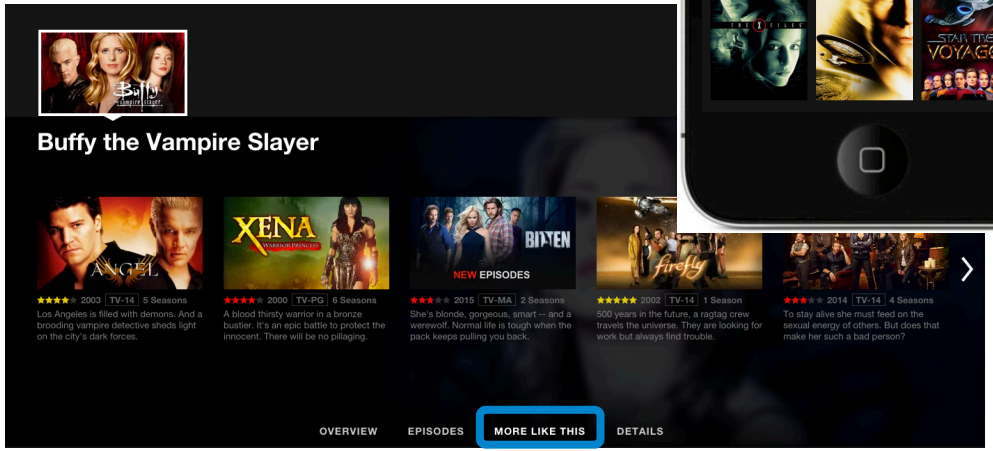


NETFLIX

Similarity

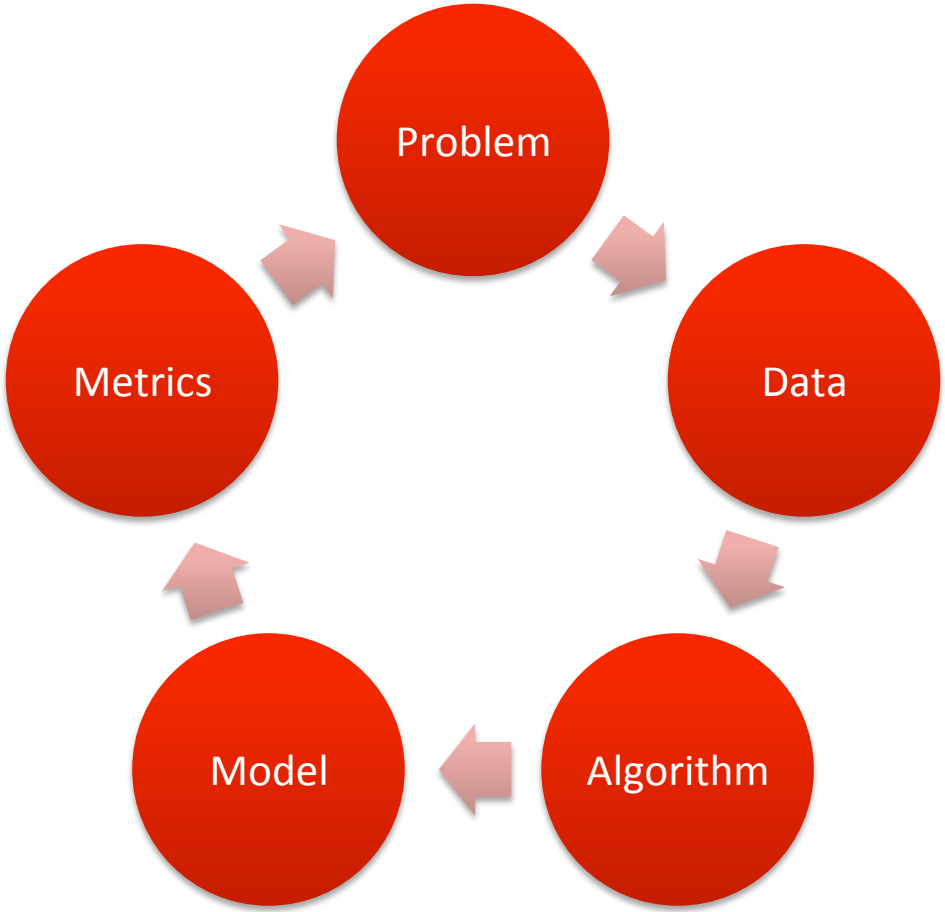
- Recommend videos similar to one you've liked
- "Because you watched" rows
- Pivots
  - Video information page
  - In response to user actions (search, list add, …)





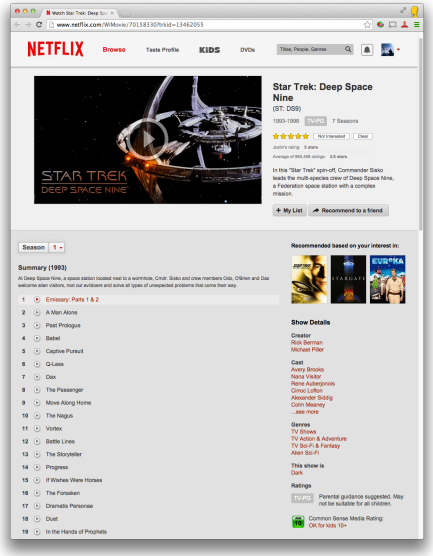
Science and Data

Machine Learning Approach

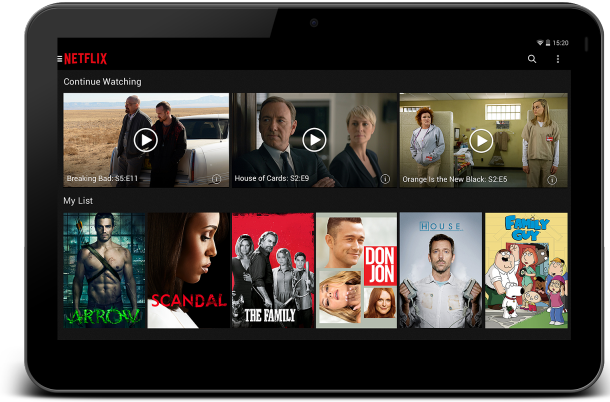


Data

- Plays
  - Duration, bookmark, time, device, …
- Ratings
- Metadata
  - Tags, synopsis, cast, …
- Impressions
- Interactions
  - Search, list add, scroll, …



ARRESTED DEVELOPMENT



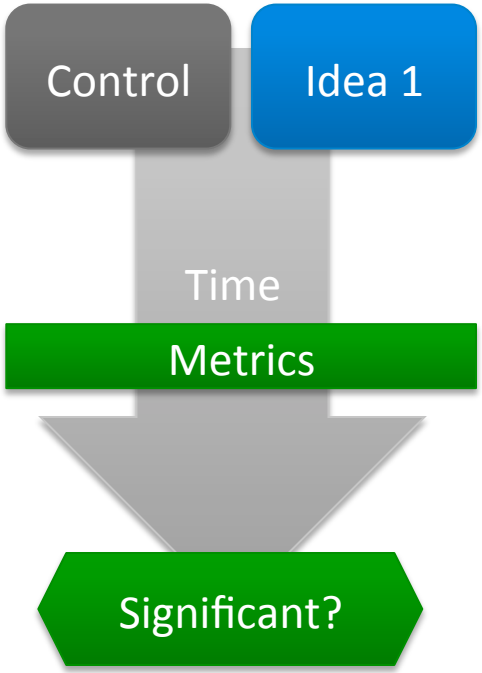
NETFLIX

Models & Algorithms

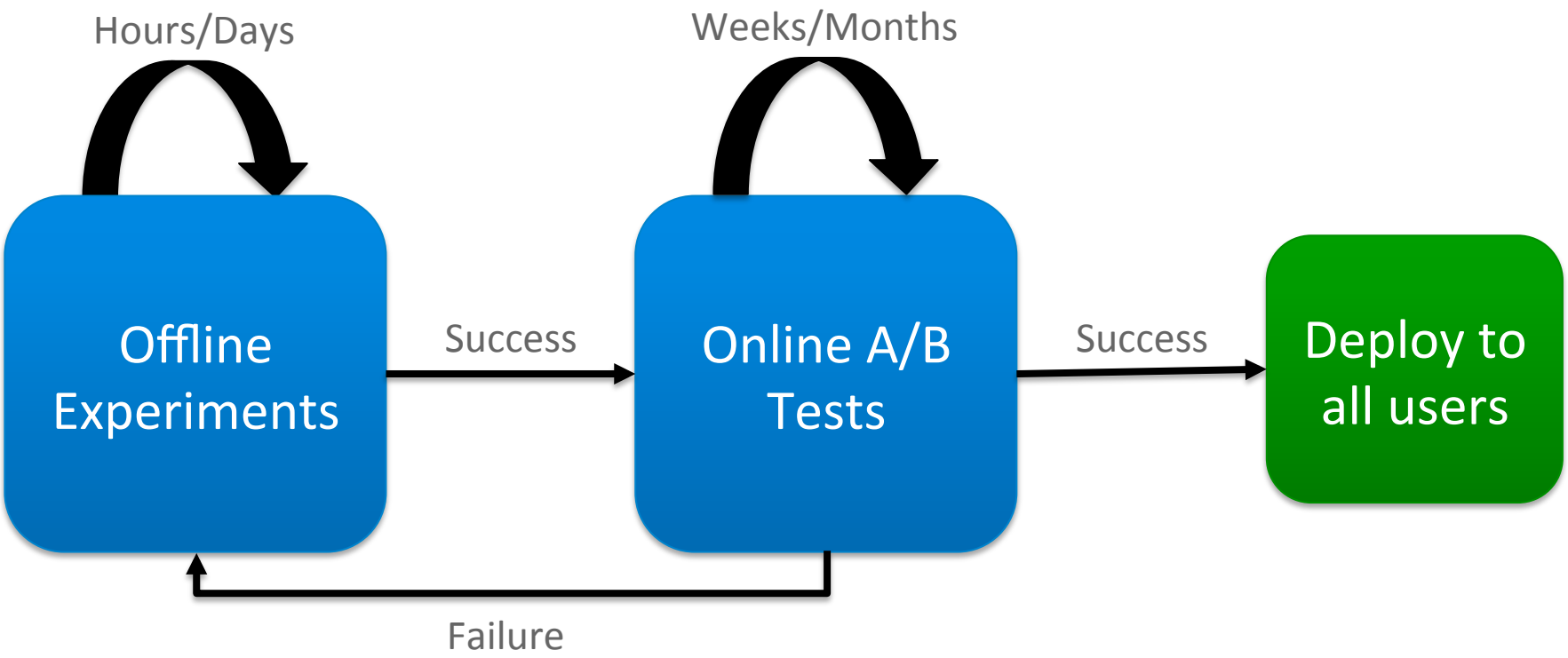
- Regression (Linear, logistic, elastic net)
- SVD and other Matrix Factorizations
- Factorization Machines
- Restricted Boltzmann Machines
- Deep Neural Networks
- Markov Models and Graph Algorithms
- Clustering
- Latent Dirichlet Allocation
- Gradient Boosted Decision Trees/ Random Forests
- Gaussian Processes
- …

A/B Testing

- Randomized, controlled experiments
- Start with hypothesis for potential improvement
- Take random sample of users (say ~500k)
  - Assign half to control group (existing experience)
  - Assign other half to new experience
- Collect business-relevant metrics over time (say 1-2 months)
  - Retention
  - Engagement (hours of plays)
- Compare outcomes with statistical test
  - Significant improvement? Roll out new experience
  - No? Use insight to inform next hypothesis



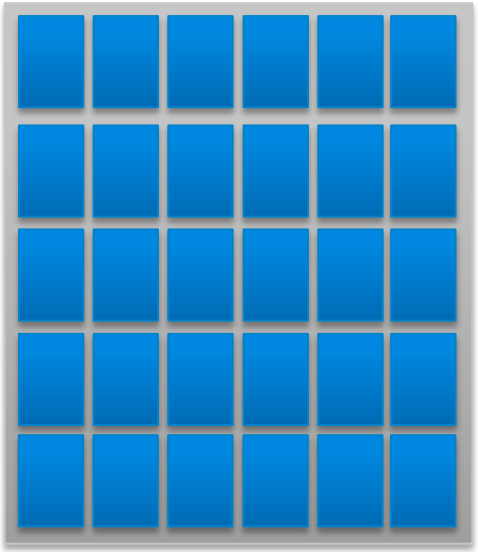
Offline-Online Testing Process



Evolution of our Recommendation Approach

Rating

Ranking

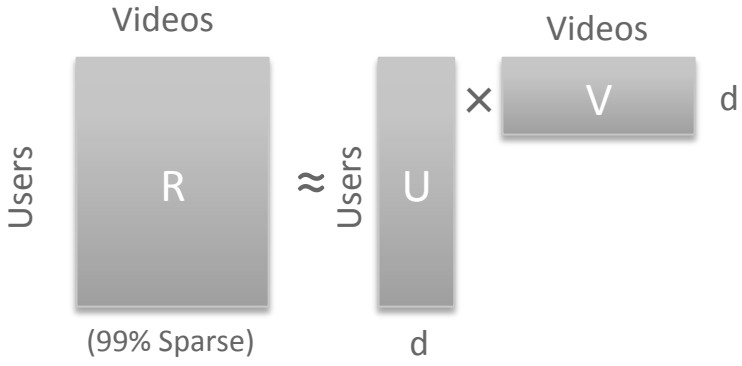


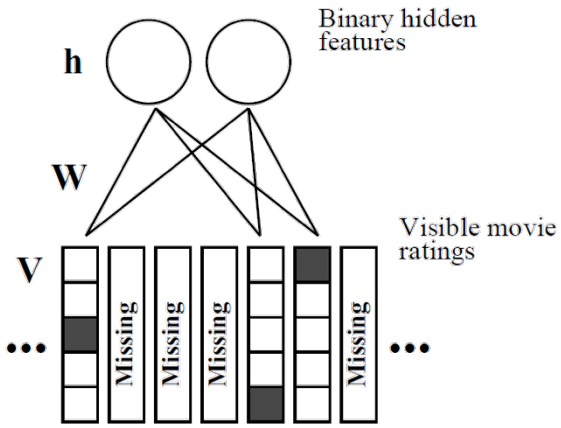
Page Generation

Personalized Rating Prediction

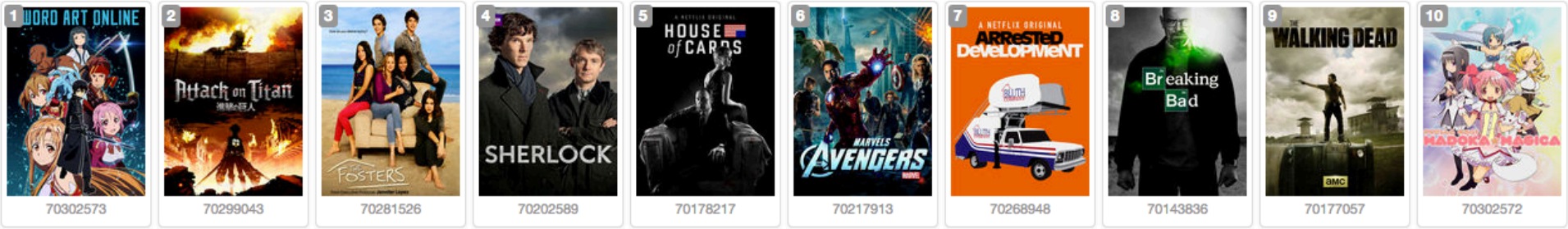
Rating Prediction

- Based on first year progress prize
- Top 2 algorithms
- Matrix Factorization (SVD++)
- Restricted Boltzmann Machines (RBM)
- Ensemble: Linear blend





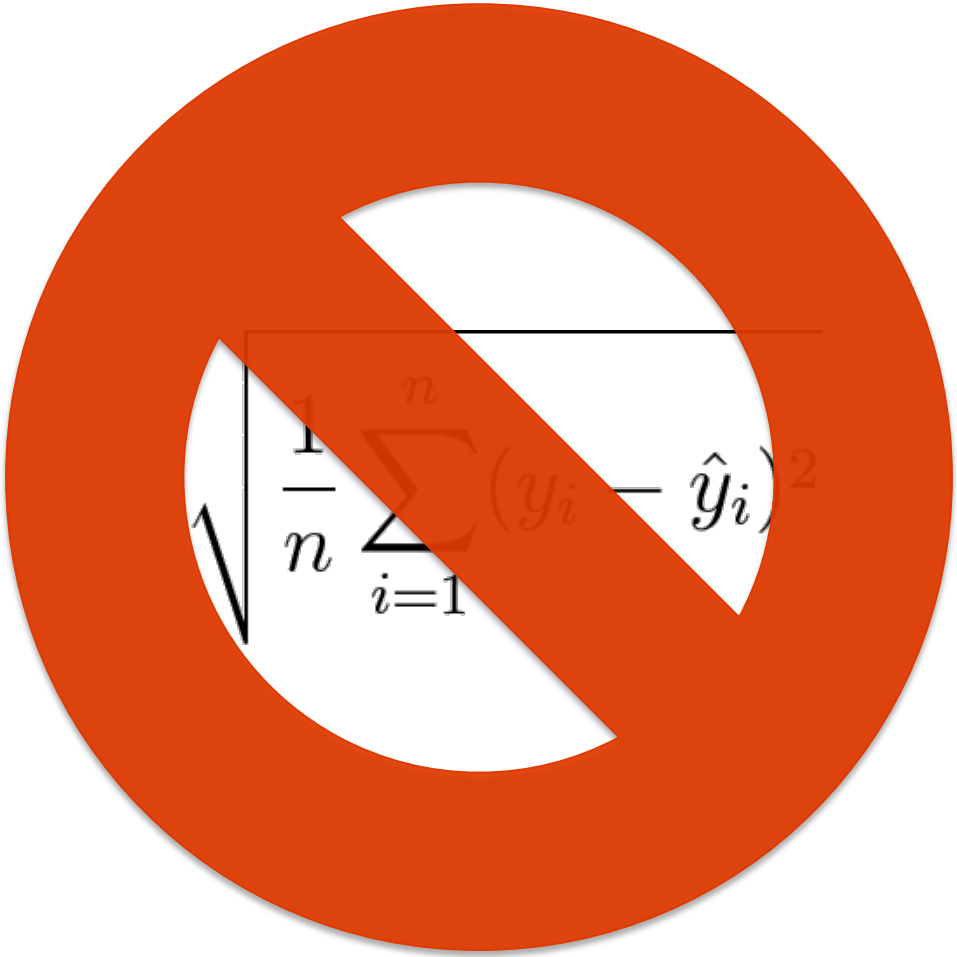
Ranking by ratings



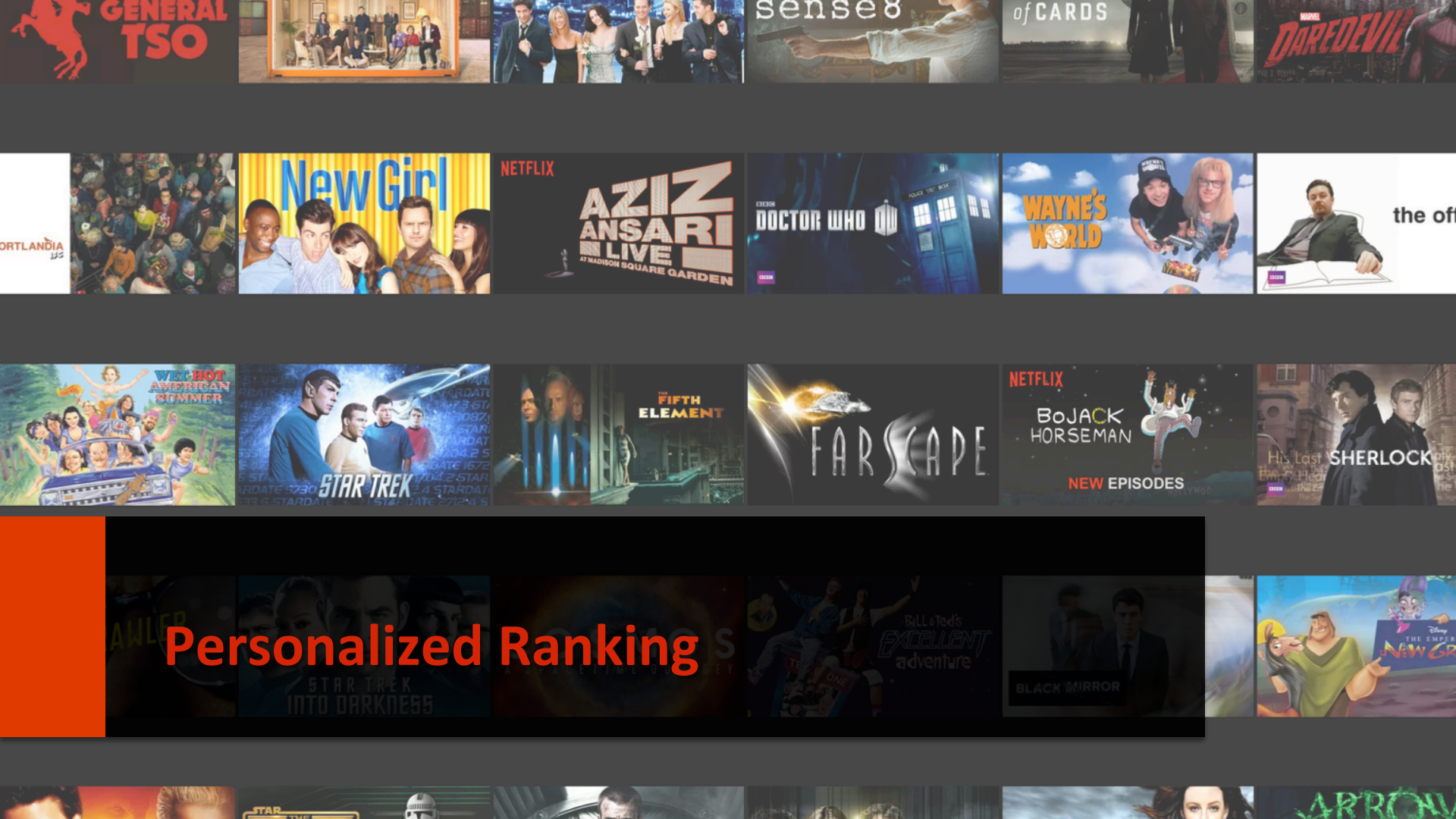
4.7 4.6 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5

Niche titles

High average ratings… by those who would watch it

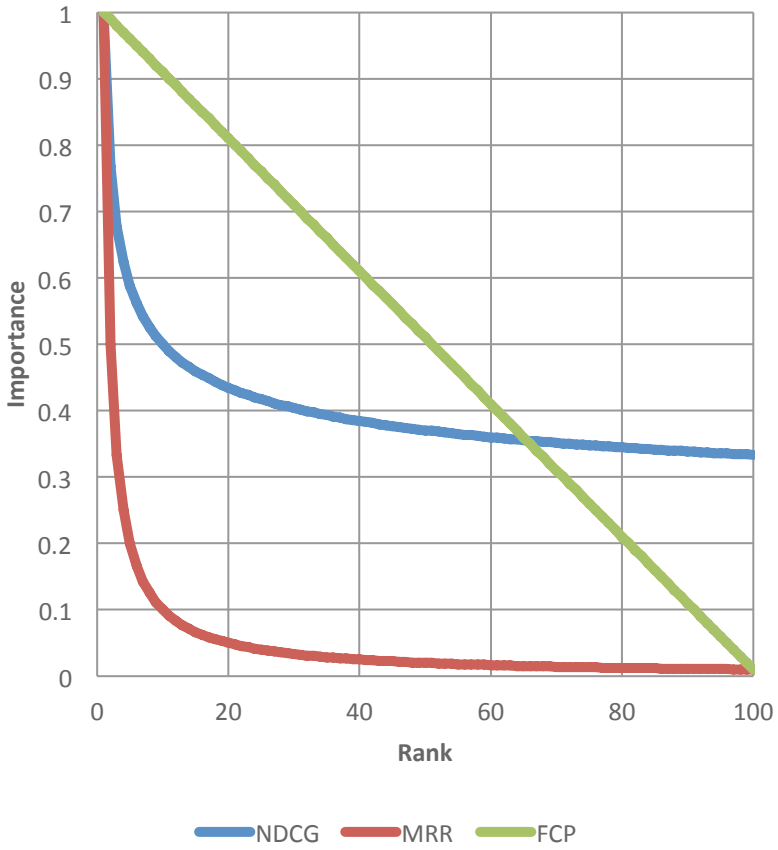

$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Personalized Ranking



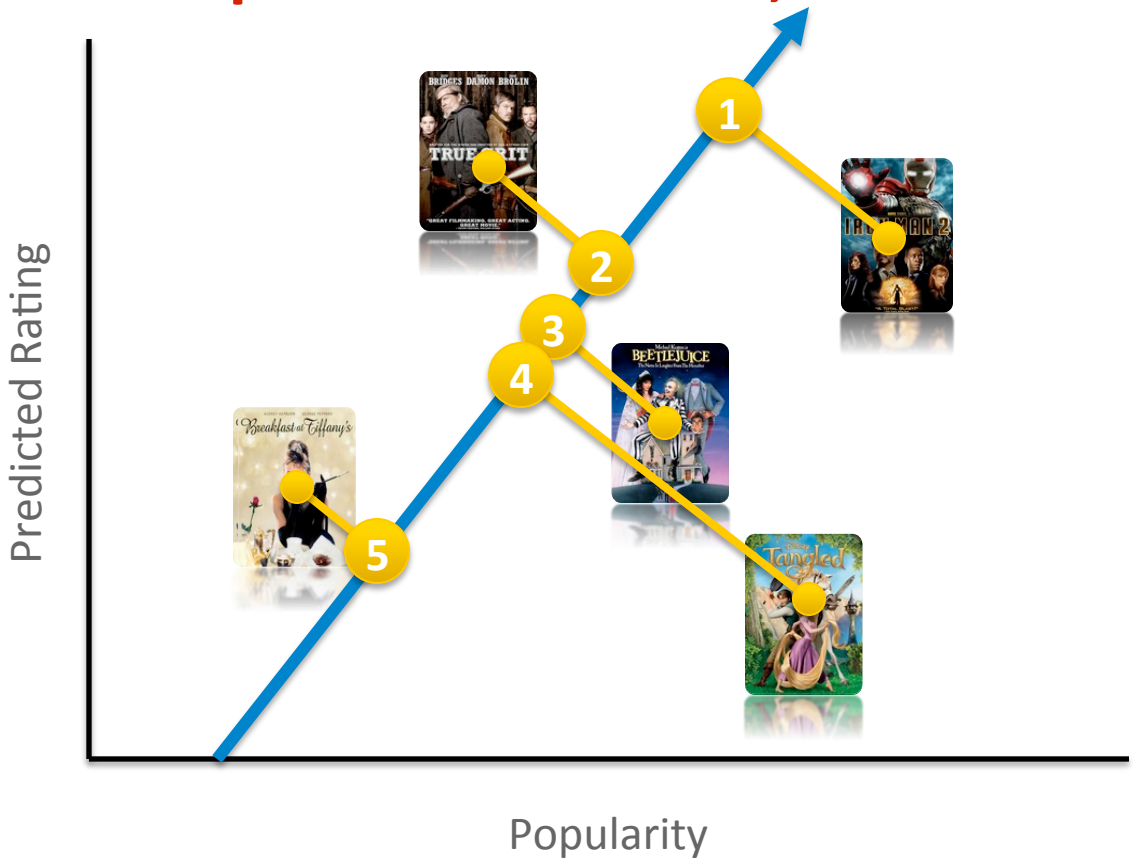
Learning to Rank

- Approaches:
  - **Point-wise**: Loss over items (Classification, ordinal regression, MF, …)
  - **Pair-wise**: Loss over preferences (RankSVM, RankNet, BPR, …)
  - **List-wise**: (Smoothed) loss over ranking (LambdaMART, DirectRank, GAPfm, …)
- Ranking quality measures:
  - Recall@N, Precision@N, NDCG, MRR, ERR, MAP, FCP, …



NETFLIX

Example: Two features, linear model

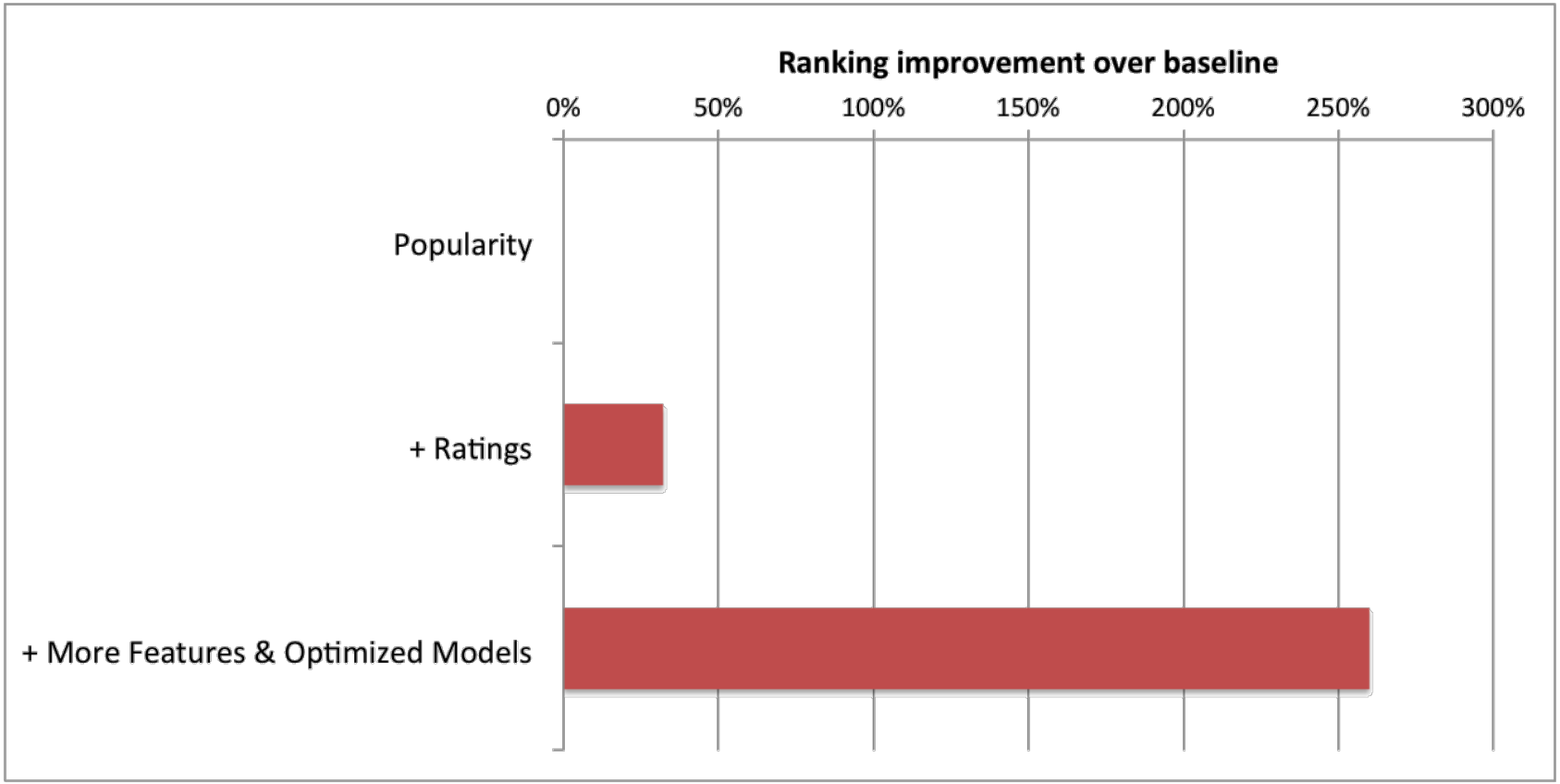


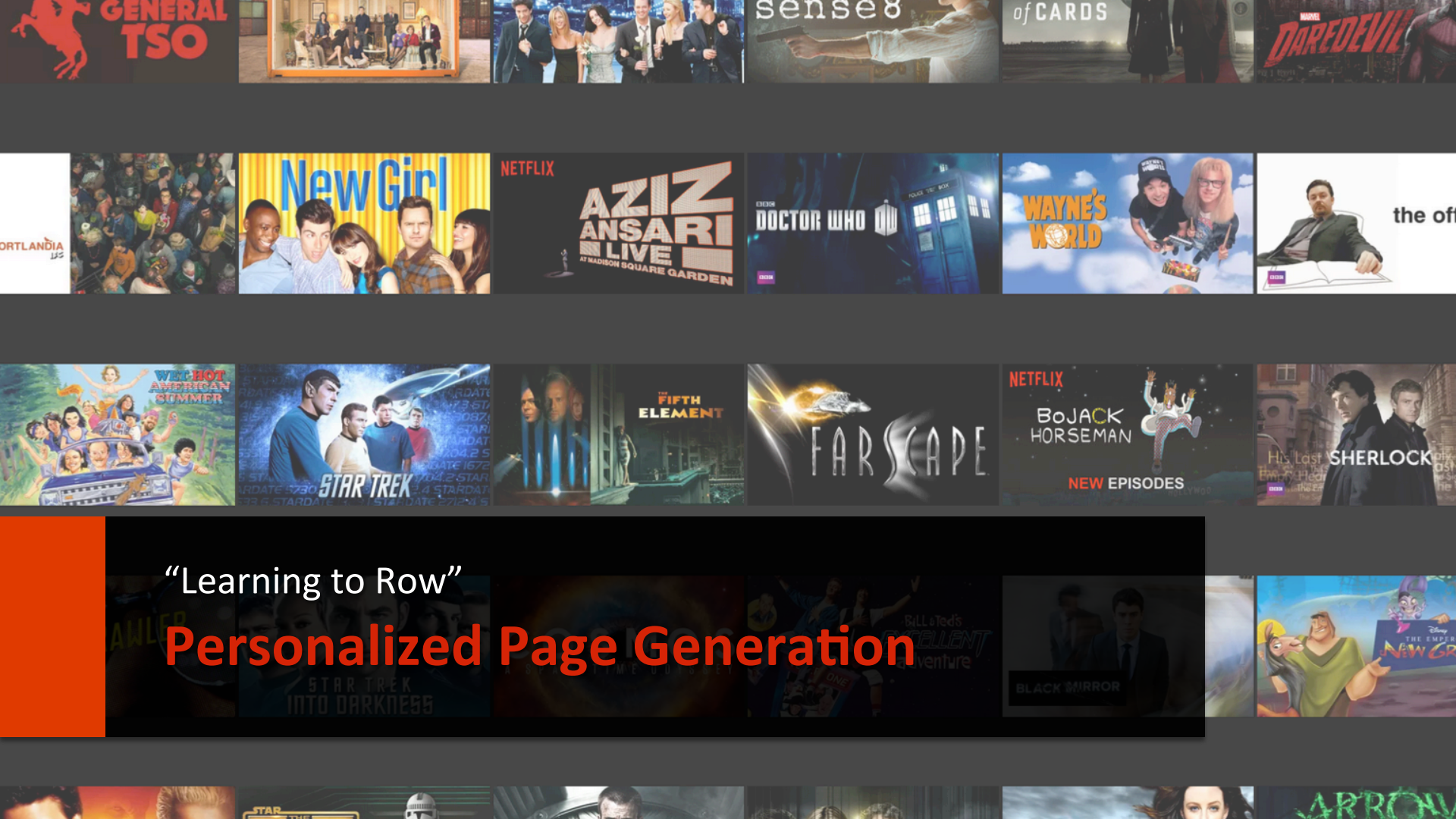
Linear Model:

$$f_{\text{rank}}(u,v) = w_1\, p(v) + w_2\, r(u,v)$$

Final Ranking

Personalized Ranking



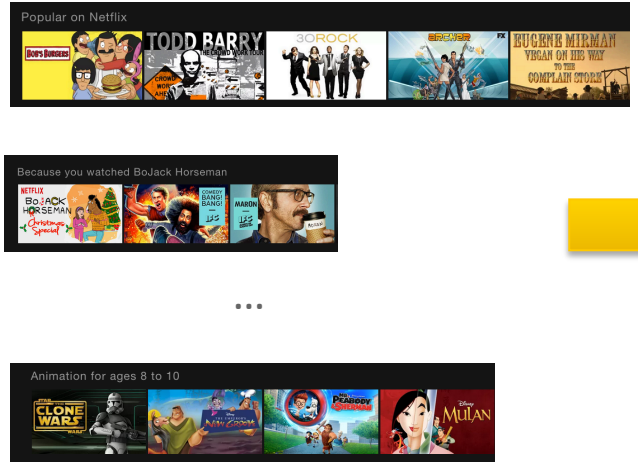


"Learning to Row"

Personalized Page Generation

Page-level algorithmic challenge

10,000s of possible rows



Variable number of possible videos per row (up to thousands)

1 personalized page

10-40 rows

per device

Balancing a Personalized Page

Accurate vs. Diverse

Discovery vs. Continuation

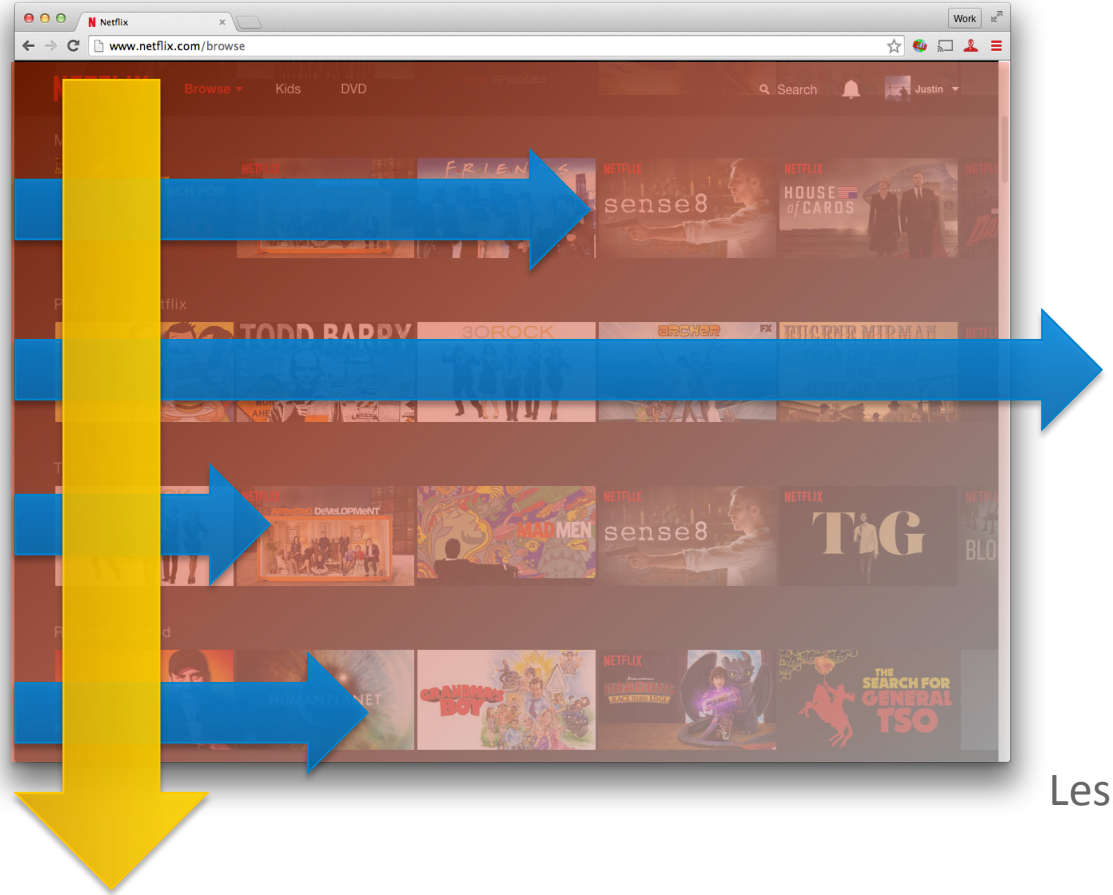
Depth vs. Coverage

Freshness vs. Stability

Recommendations vs. Tasks

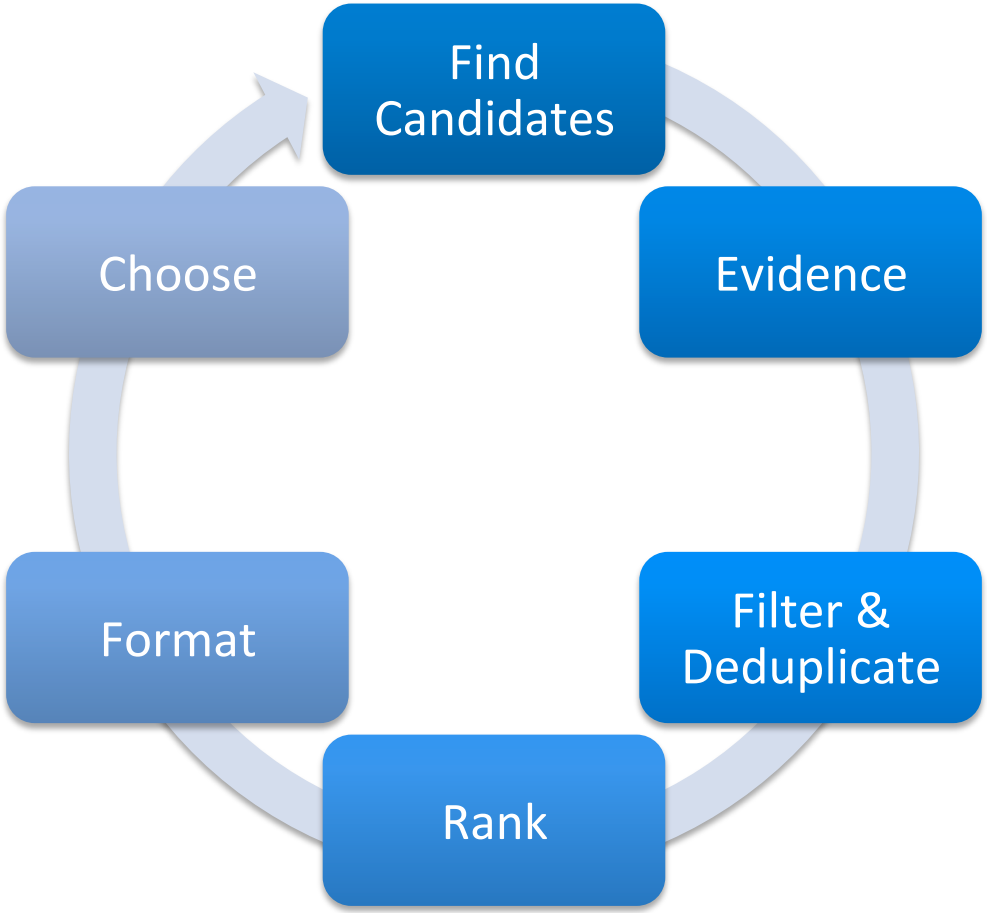
Browsing Modeling

More likely to see



Less likely

Process for creating rows

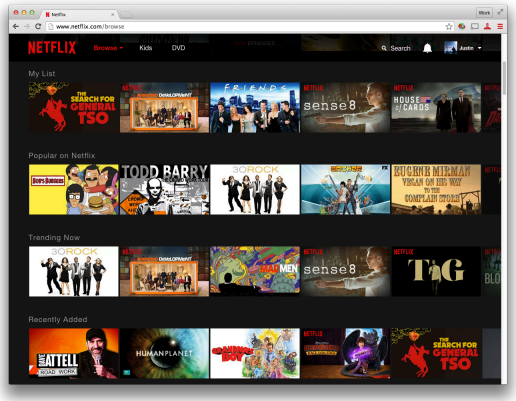


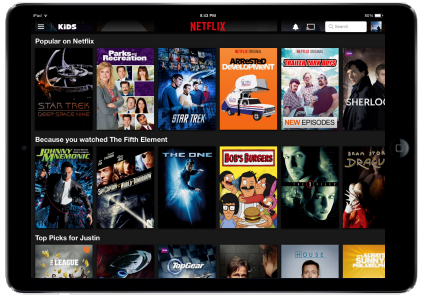
NETFLIX

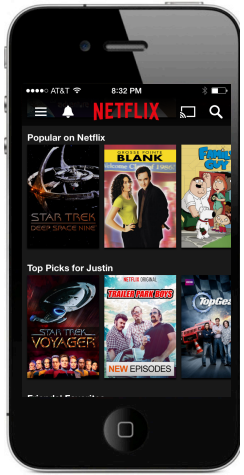
Building a page algorithmically

Approaches

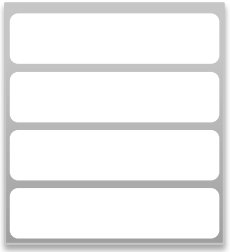
1. Rule-based
2. Row-independent
3. Stage-wise
4. Page-wise







Rule-based Approach



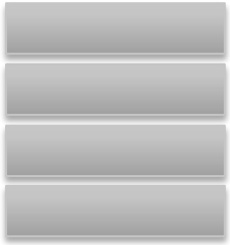
- Create rules for location of rows
- Non-personalized layout
- Optimize by AB testing

- Example:
  - First: Continue Watching (if any)
  - Next: Top picks
  - Next: Popular on Netflix
  - Next 3: Because you watched (if any)
  - Next 3: Genres
  - Next: Recently added
  - …

- Pros:
  - Backed up by AB testing
  - Easy to compute
  - Very familiar page layout
- Cons:
  - One size fits nobody
  - Local optima
  - Hard to add new kinds of rows
  - Rules get complicated
  - Still need to pick best rows within groupings

NETFLIX

Row-independent Approach



- Treat row selection as a standard ranking problem
- Greedily rank rows by $f(r \mid u, d)$
- Sort and take top n

- Approaches:
  - Top-n recommendation algorithms

- Pros:
  - Available algorithms
  - Fast
  - High accuracy
- Cons:
  - No consideration of diversity
  - People watch videos, row interaction is indirect

Stage-wise Approach

- Use information about already picked rows
- Pick next rows by $f(r \mid u, d, p_{1:n})$

- Approaches:
  - Train scoring model
  - Greedy
  - Submodular optimization
  - K-step look-ahead

- Pros:
  - Can balance accuracy and diversity
  - Follows navigation pattern
- Cons:
  - May need to evaluate each row multiple times from deduping
  - Still greedy
  - Hard to get training data

Page-wise approach

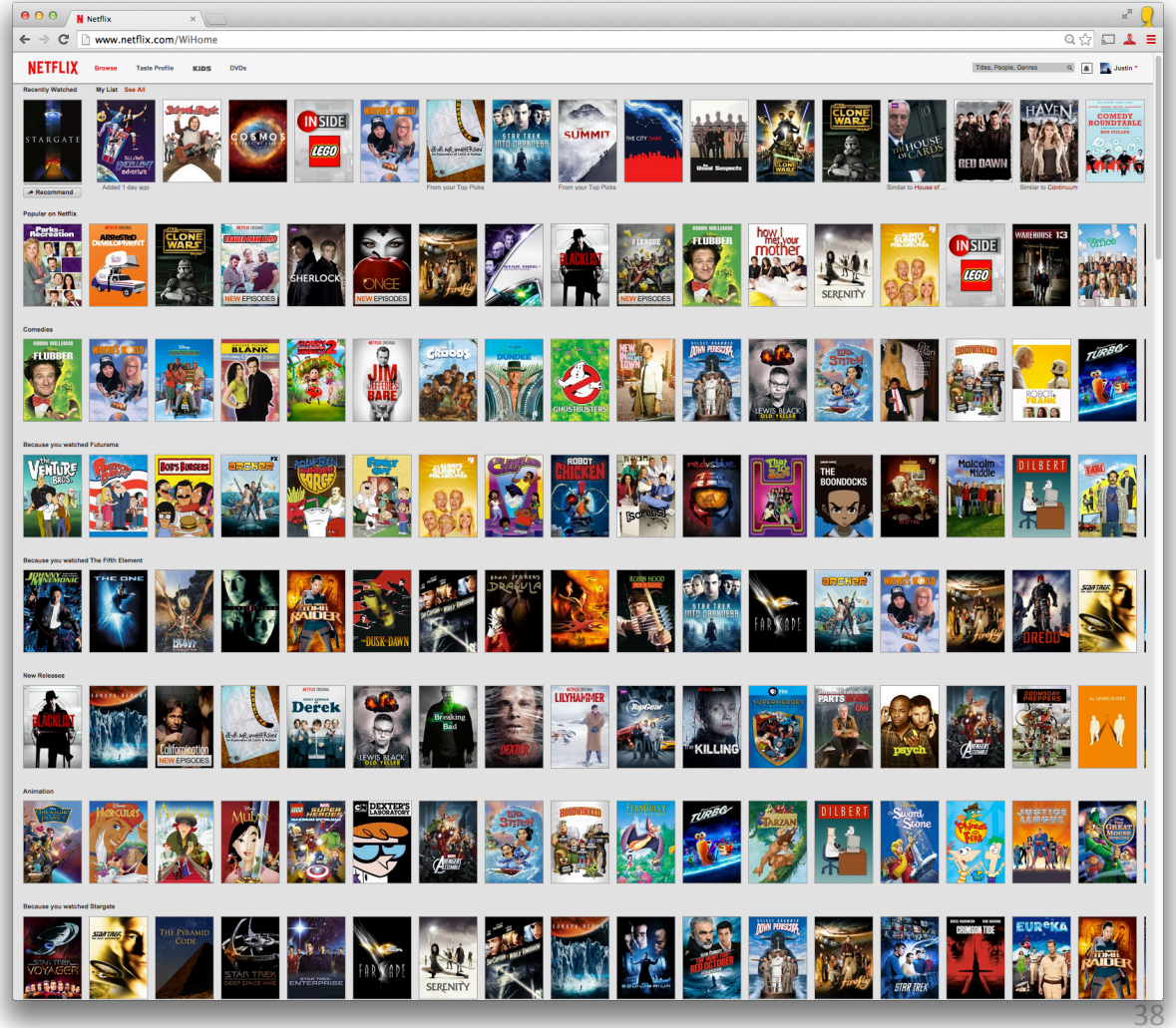
- Optimize the whole page
- Potentially re-rank within rows
- Total page fitness $f(p \mid u, d)$

- Approaches:
  - Constrained optimization
  - Approximate integer programming

- Pros:
  - Includes information about entire page
  - Can include diversity and deduping into objective
  - Able to synthesize rows
- Cons:
  - Computationally expensive
  - Huge space of possible pages

Row Features

- Quality of items
- Features of items
- Quality of evidence
- User-row interactions
- Item/row metadata
- Recency
- Item-row affinity
- Row length
- Position on page
- Title
- Diversity
- Similarity
- Freshness
- …

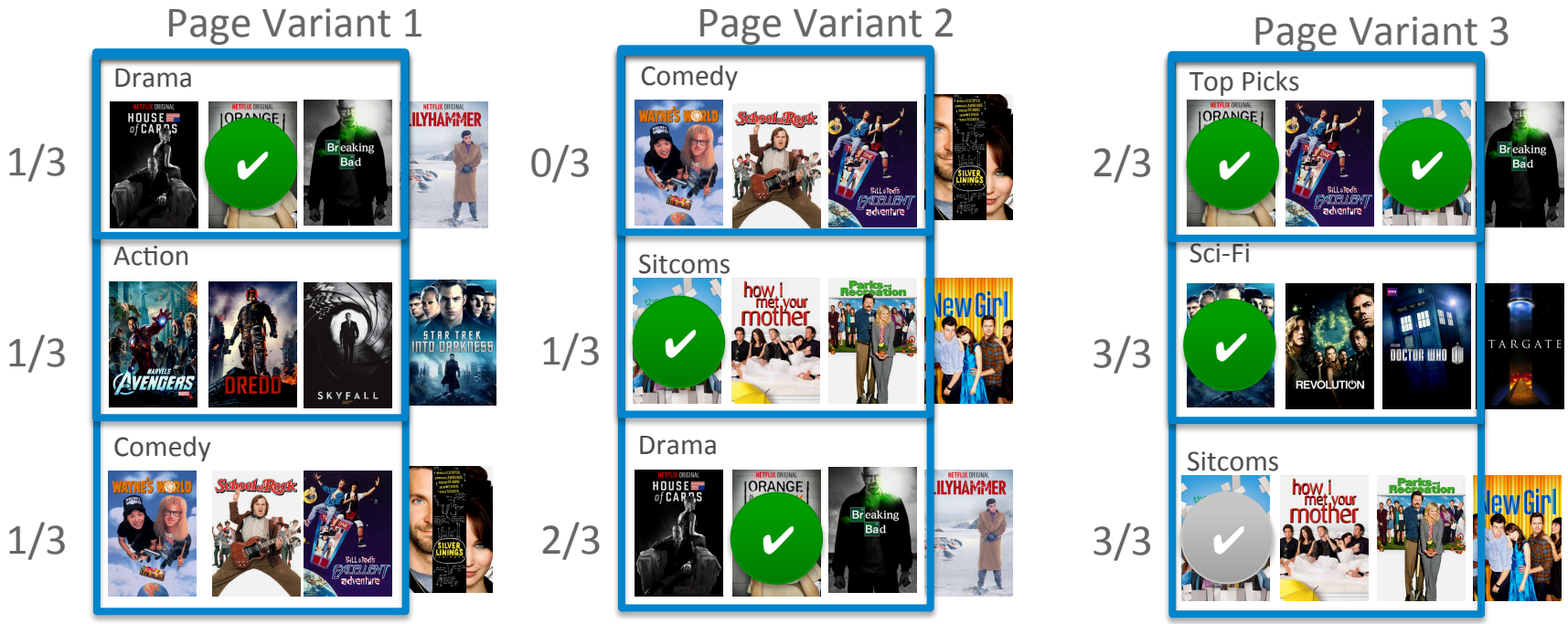


NETFLIX

Page-Level Metrics

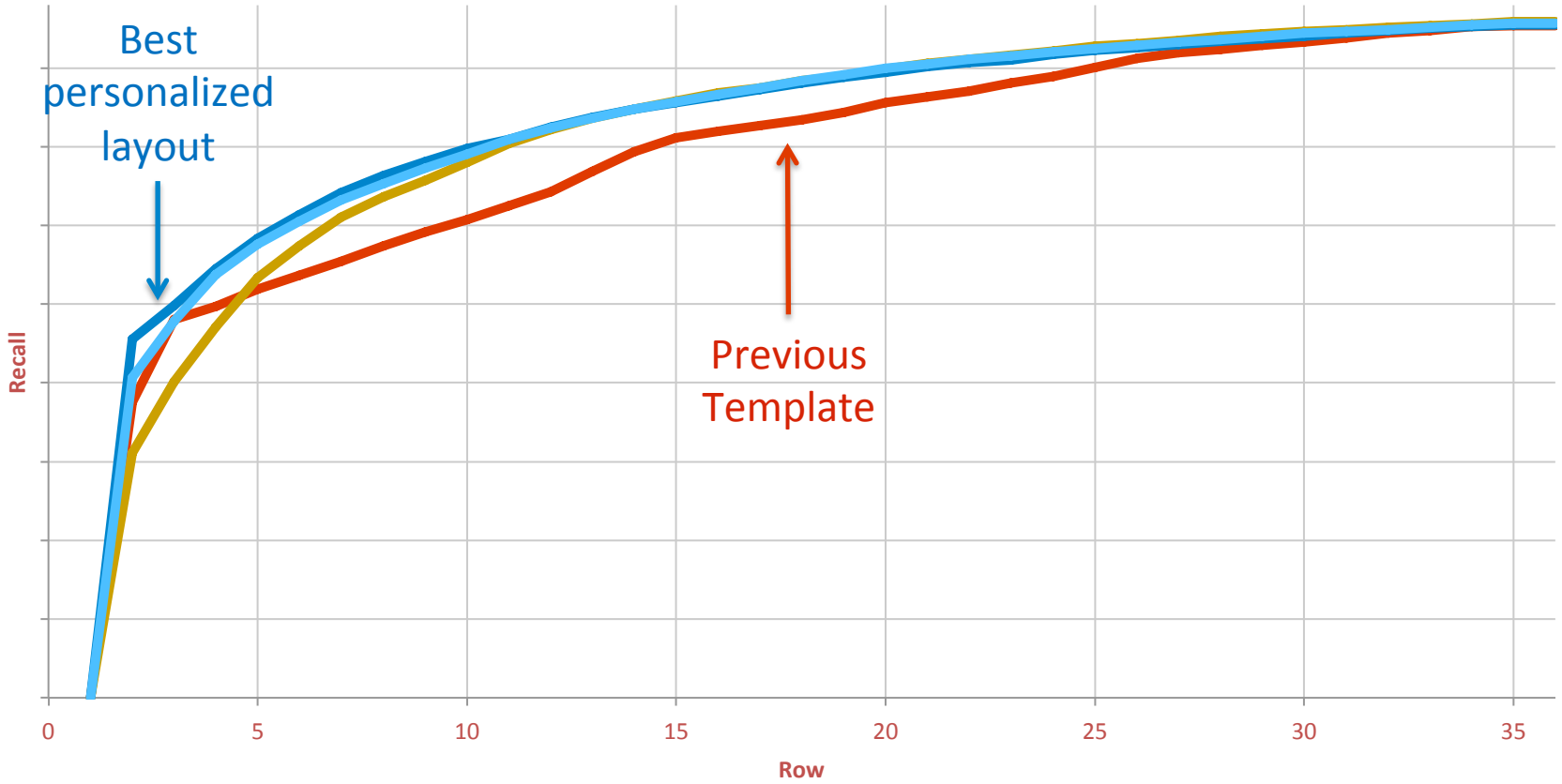
- How do you measure the quality of the homepage?
  - Ease of discovery
  - Diversity
  - Novelty
  - …
- Challenges:
  - Position effects
  - Row-video generalization
- Approach: 2D versions of ranking quality metrics

Simple example: Recall @ row-by-column

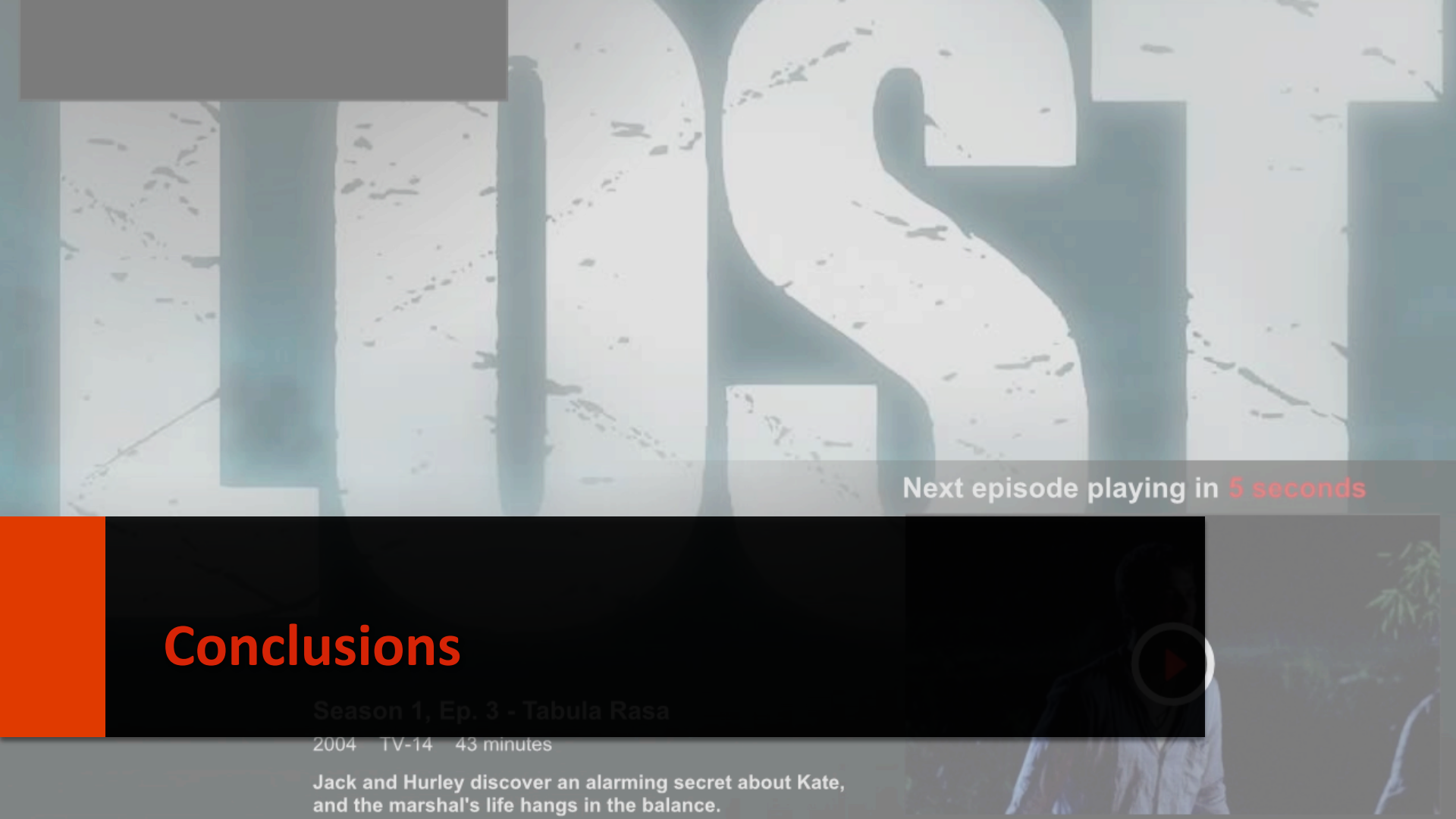


User Watched

Offline Results: Recall @ row-by-column

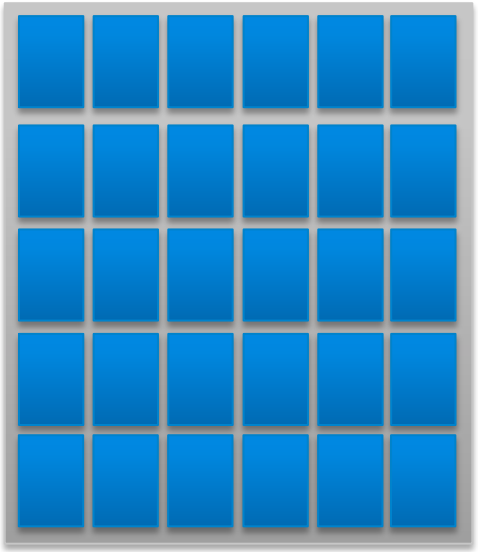


NETFLIX



Conclusions

Evolution of Recommendation Approach



Rating

Ranking

Page Generation

Research Directions

Presentation effects

Full-page optimization

Context Awareness

Evidence Selection

Global Algorithms

Cold start

Thank You

Justin Basilico
jbasilico@netflix.com
@JustinBasilico