Netflix Optimization: A Confluence of Metrics, Algorithms, and Experimentation

CIKM 2013, UEO Workshop

Caitlin Smallwood


Allegheny
Monongahela
Ohio River

TV & Movie Enjoyment Made Easy

Stream any video in our collection on a variety of devices for $7.99 a month


Over
800
Partner Products
Netflix Ready Devices
From: May 2008
To: May 2010
instant streaming ready
NETFLIX
Roku
Netflix Player by Roku
Microsoft
Xbox 360
LG
BD390
Sony
PlayStation 3
VIZIO
SV422XVT
Samsung
LED C8000
Nintendo
Wii
Apple
iPad

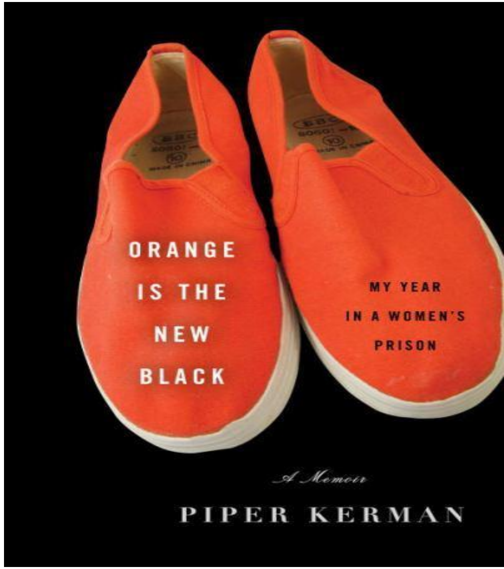
Content Partners

Kids

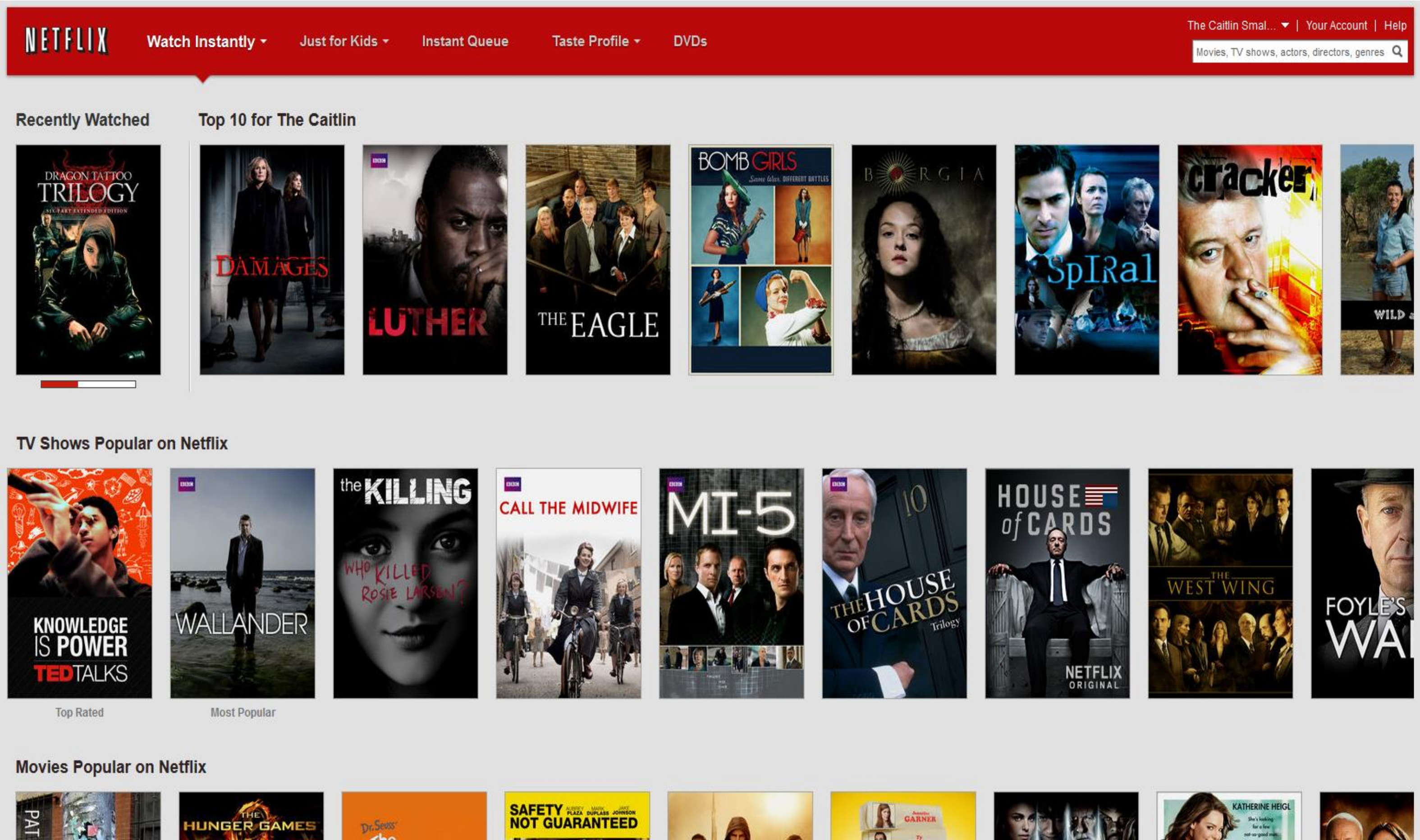
Original Content

HOUSE of CARDS

ARReSTeD DeVeLOPMeNT™



The UI


NETFLIX
Watch Instantly
Just for Kids
Instant Queue
Taste Profile
DVDs
The Caitlin Smal...
Your Account
Help
Movies, TV shows, actors, directors, genres
Recently Watched
Top 10 for The Caitlin
DRAGON TATTOO TRILOGY
DAMAGES
LUTHER
THE EAGLE
BOMB GIRLS
BORGIA
SpIRal
cracker
TV Shows Popular on Netflix
KNOWLEDGE IS POWER TEDTALKS
Top Rated
WALLANDER
Most Popular
the KILLING
WHO KILLED ROSIE LARSEN?
CALL THE MIDWIFE
MI-5
THE HOUSE OF CARDS Trilogy
HOUSE of CARDS
NETFLIX ORIGINAL
THE WEST WING
Movies Popular on Netflix
THE HUNGER GAMES
SAFETY NOT GUARANTEED
KATHERINE HEIGL

The Data

- Visitor data
- User Metadata
- Social
- Users’ Plays (streaming)
- Users’ Ratings
- Users’ Searches
- Device streaming performance
- Video Metadata
- Video Impressions

A few facts

- 40M members globally
- Ratings: 4M+/day
- Searches: 3M+/day
- Plays: 1B+/month

Metrics

“Engagement is a user’s response to an interaction that gains, maintains, and encourages their attention, particularly when they are intrinsically motivated”

- Jacques, 1996

User Engagement Measurement Techniques

- Self-reported or “explicit”
  - Satisfaction, likelihood to recommend, likelihood to use or re-use, self-reported usage, self-reported preferences
- Physical observation of users
  - User experience in-person research, eye tracking
- Behavioral observation of users
  - Analytics on behavioral data

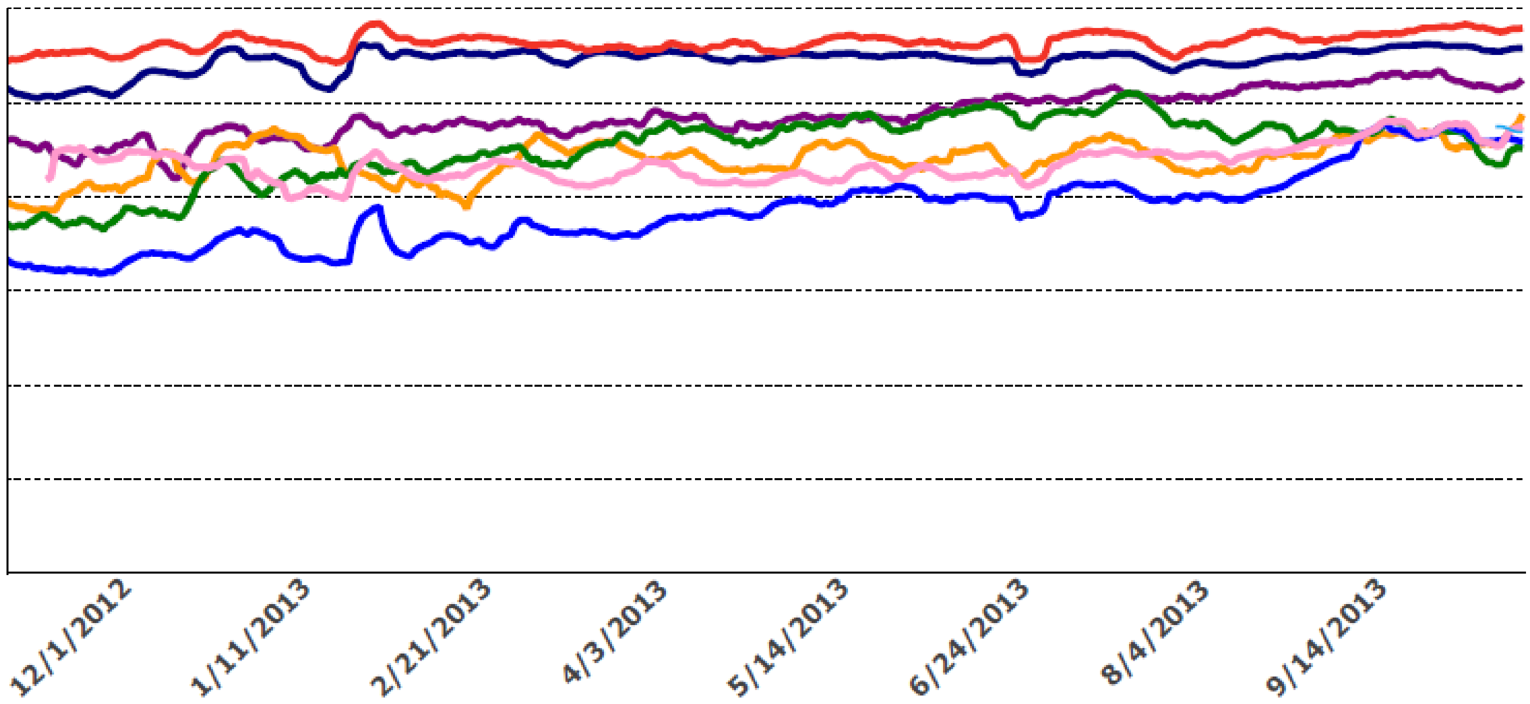
Common user engagement metrics

- Lifetime value (LTV)
- Retention
- Page views
- Time spent
- Number of distinct actions
- Recency of last visit/use
- Time between visits/uses

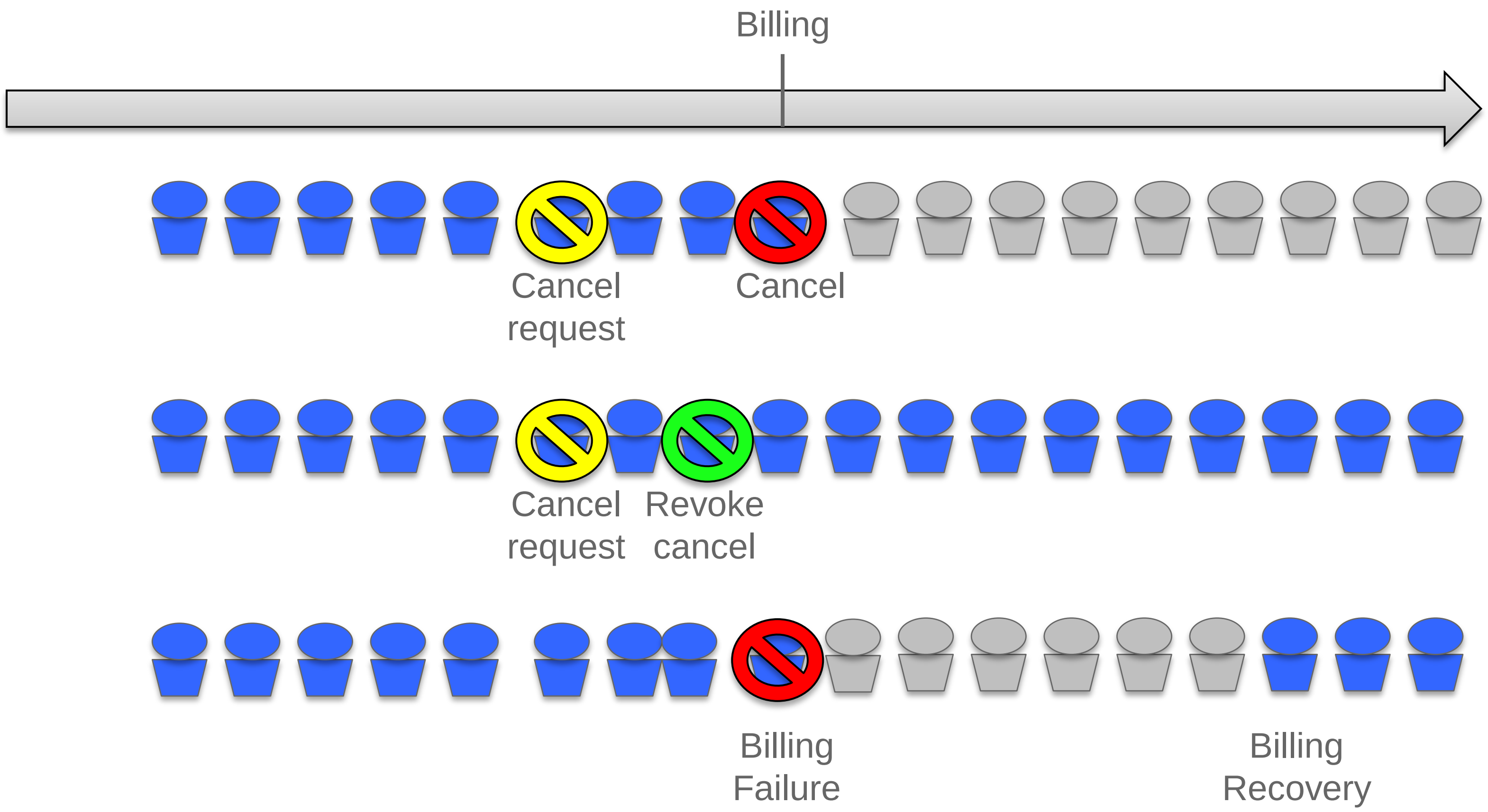
YOUR Engagement Metrics

- What's your business model?
  - Monthly subscription
- What do you want your customers to do?
  - Retain monthly (forever) because they enjoy the service
- What do your happiest, most valuable customers do?
  - Retain month over month…
  - and watch

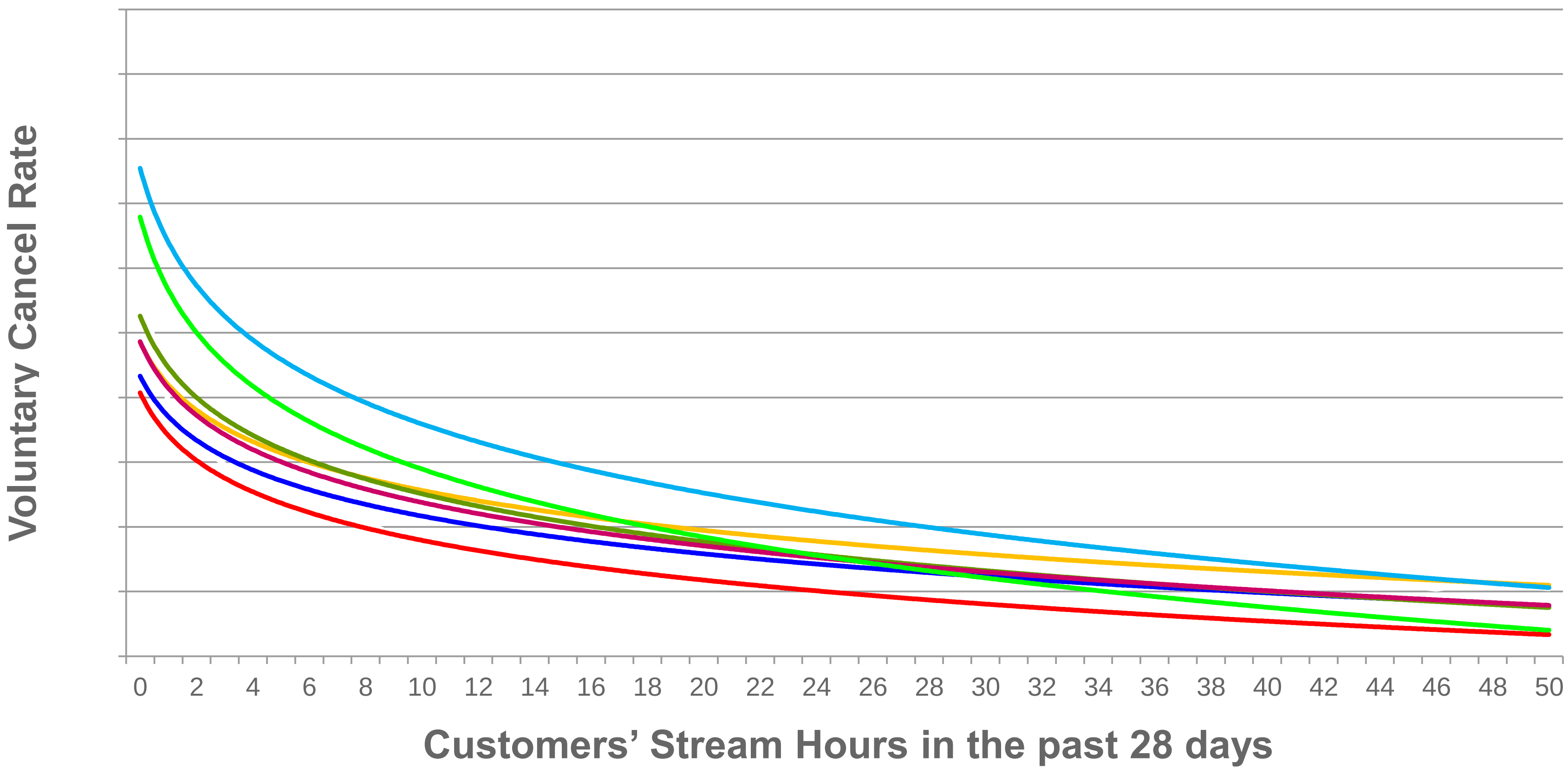
Monthly Retention



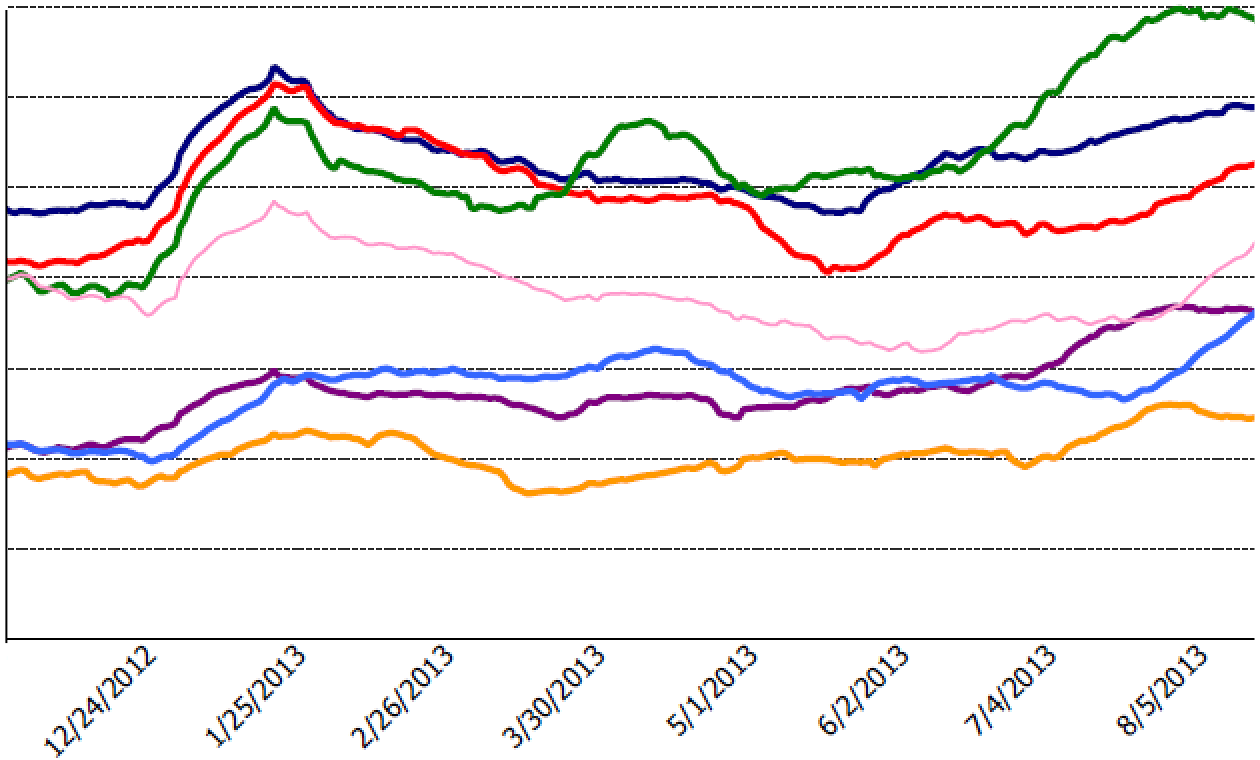
Specificity



Our most satisfied customers watch more



Median streaming hours per user

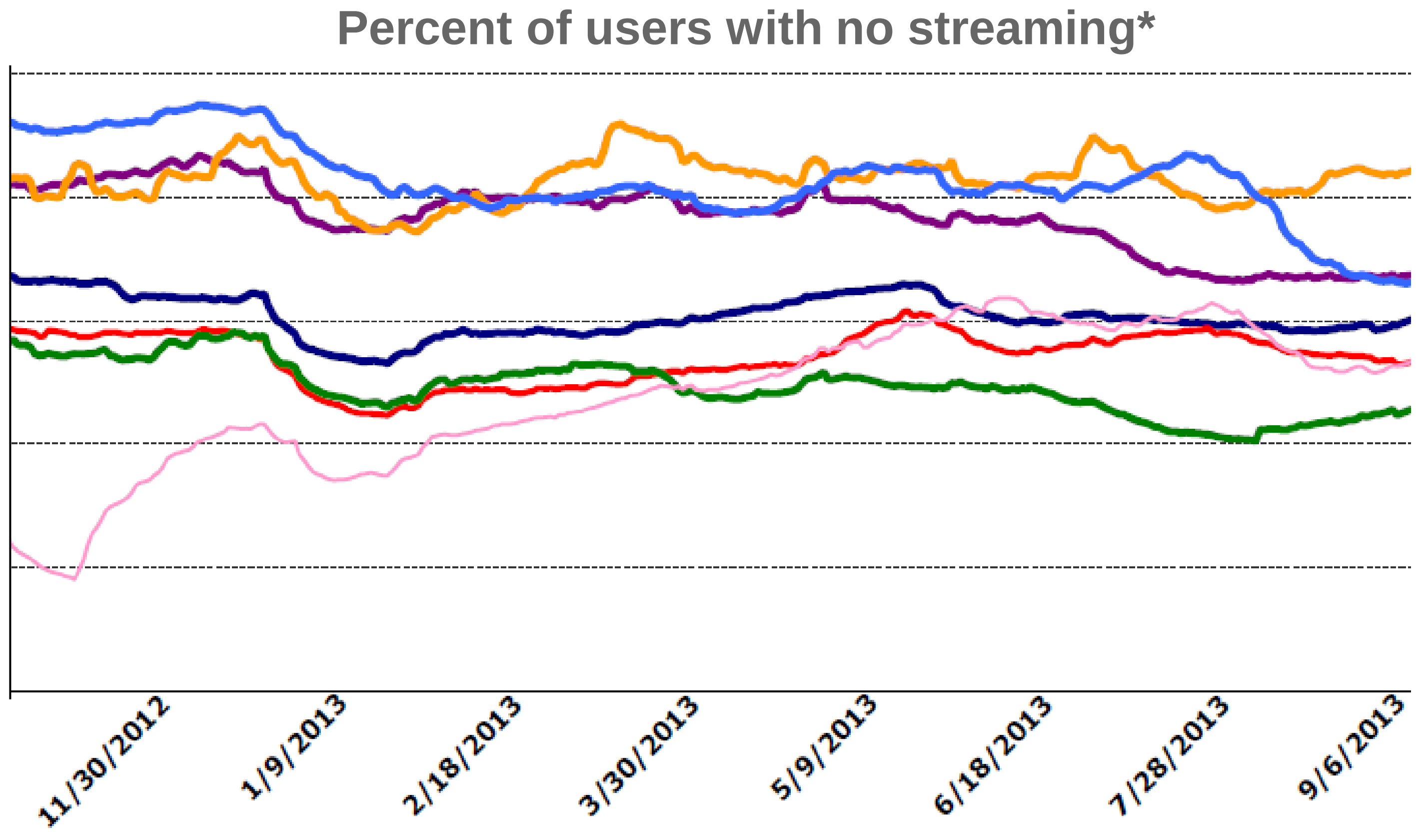


*Over 28-day period

User engagement = user granularity

- % of users who do x
- Medians – or better, distributions – of user-level volume measures

Negative metrics can also be useful



*Over 28-day period

One process for identifying engagement metrics

- Decide on criteria for a “good” metric
- Brainstorm metrics that might meet criteria
- Identify the best candidates
  - Predictive modeling or other analytic techniques
  - Expert judgment
  - Qualitative research
- Validate by trying to use the metric
  - Experiment measurement
  - Algorithm or model input
  - Trends

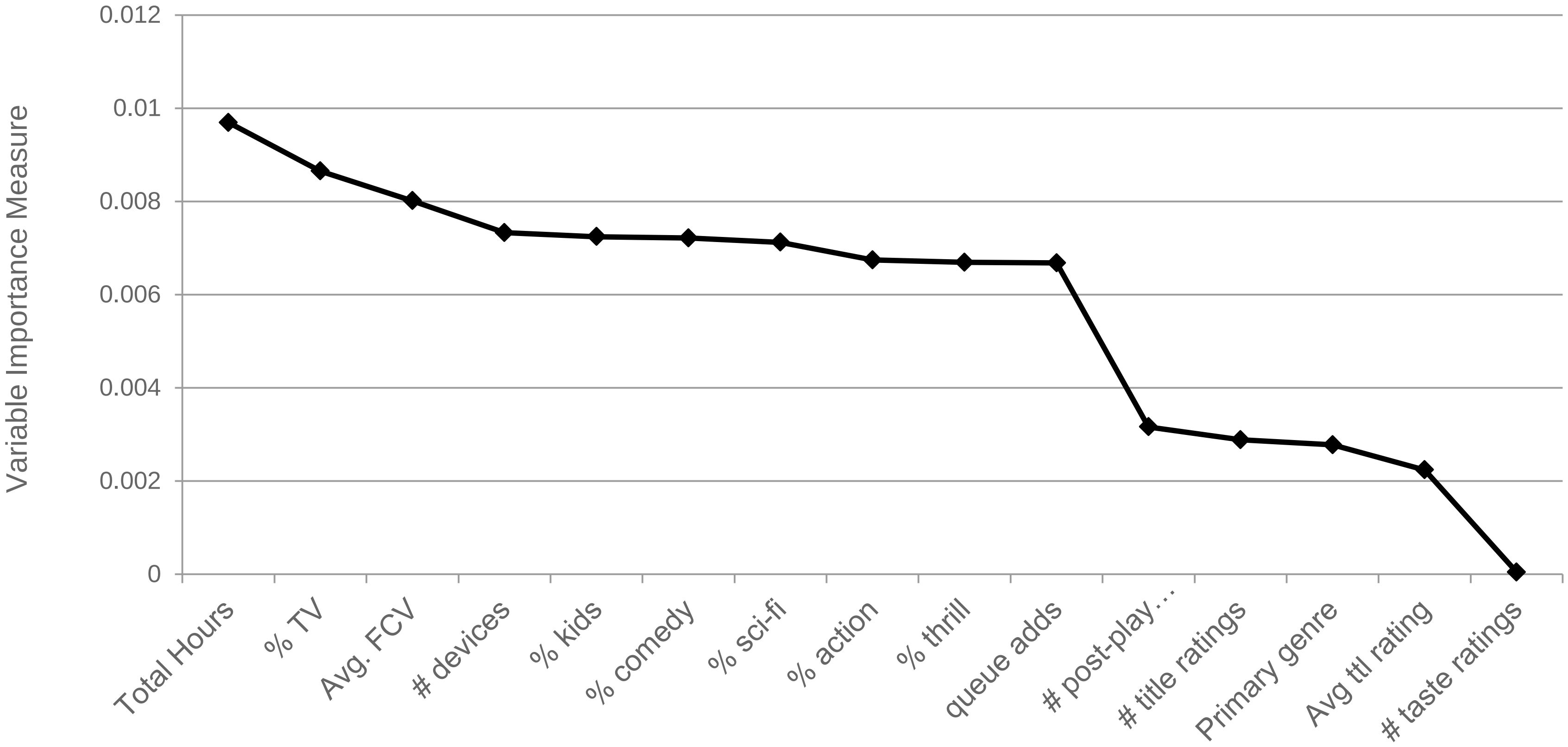
Some metric criteria suggestions

- Metric is correlated with core business metrics (conversion, retention)
  - and contributes unique predictive power beyond the other metrics?
- Metric is user-level or weights users properly toward core metrics
- Metric is actionable
- Metric shows differentiation

Brainstorm from all angles

- Variety/novelty, joy, trust, focused attention
- Positive and negative experiences
- What, who, how, why?
- Recency, Frequency, Monetization
- Short-term, long-term, changes over time
- Metric variants
- …

Example of ranking metrics' abilities to explain core business metrics (retention in this case)

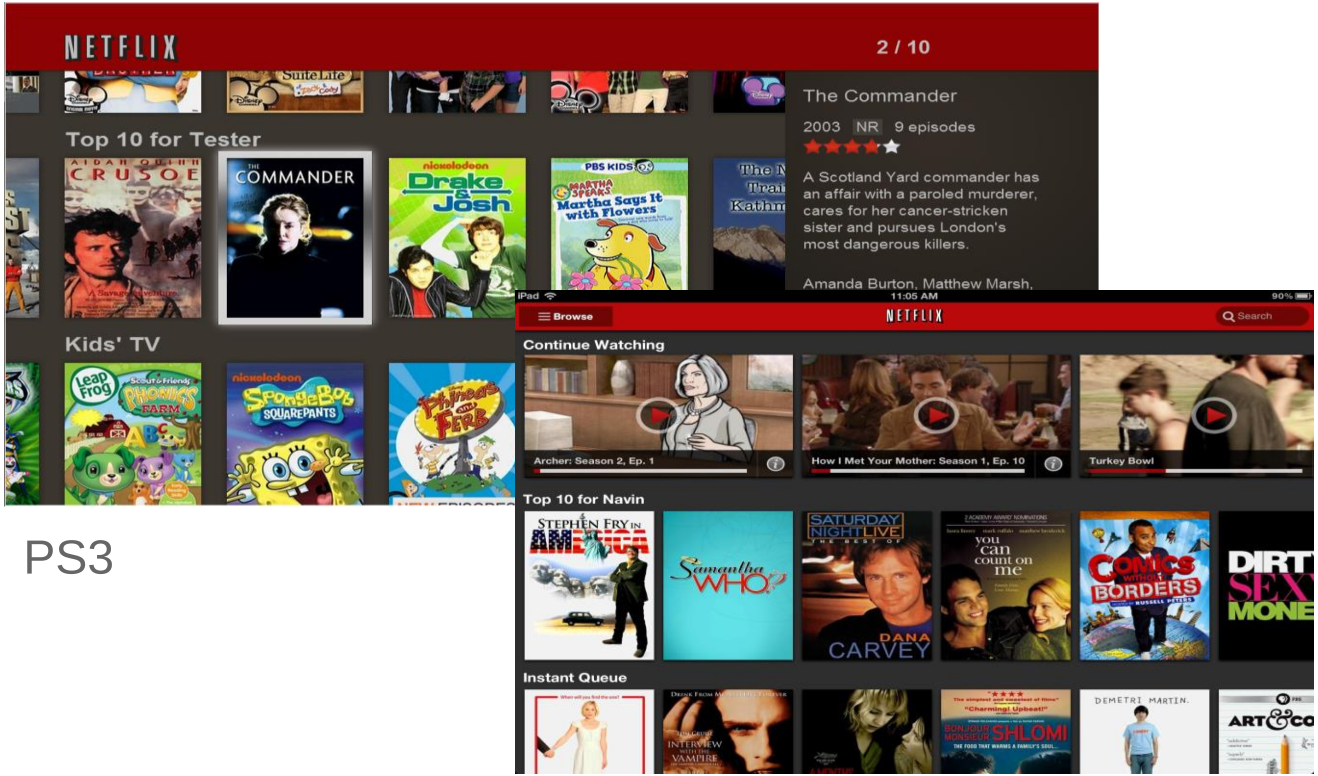


Algorithms

Algorithms for…

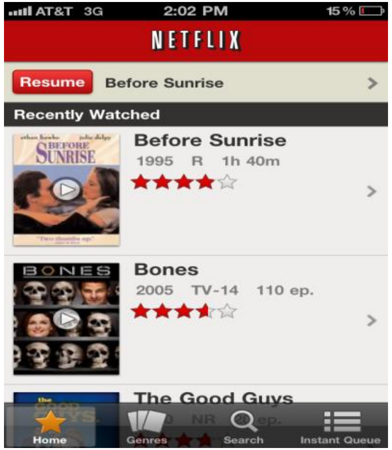
- Content recommendations
- Search results
- Streaming experience

80% of plays are based on recommendations



PS3

phone



tablet

Same algorithms power the recommendations on all devices

The Basics

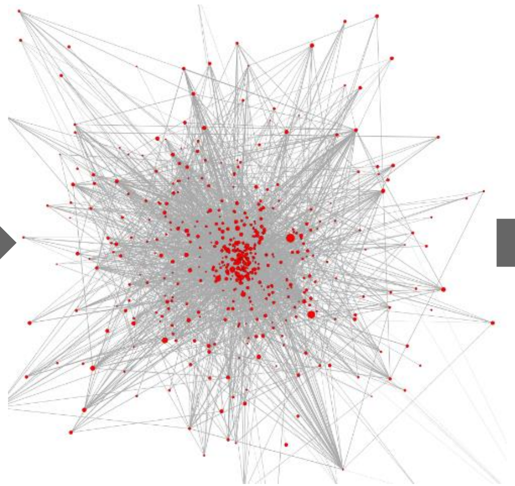
Data Inputs

Explicit member data
- Taste preferences
- Title ratings

Implicit member data
- Viewing history
- Queue adds
- Ratings

Non-personalized data
- Content library
- Title tags
- Popularity

Algorithms



Recommendations

- Rows
- Titles within rows

What the algorithms do

- Row selection
- Video ranking
- Video-video similarity
- User-user similarity
- Search recommendations

Also need to consider complex characteristics and tradeoffs such as:

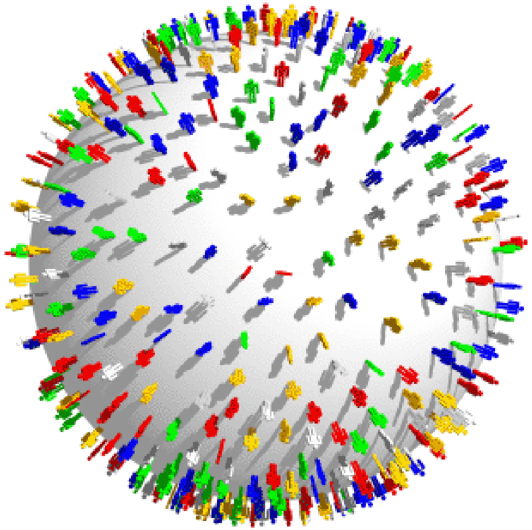
- Popularity vs personalization
- Diversity
- Novelty/Freshness
- Evidence

Probability, statistics, optimization, dynamic systems

- Hypothesis testing, estimation, delta method, bootstrapping
- Linear and generalized linear models
- Matrix factorization
- Markov processes
- Various clustering algorithms
- Bayesian models
- Latent Dirichlet Allocation
- L1 and L2 regularizations
- Association Rules
- Tree-based methods
- Bagging and boosting
- Vector spaces and the Mahalanobis distance
- …

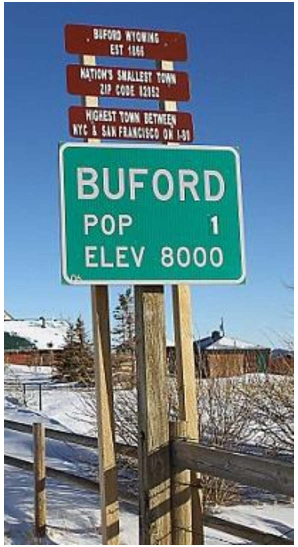
Source of Signals

Entire population



Segments

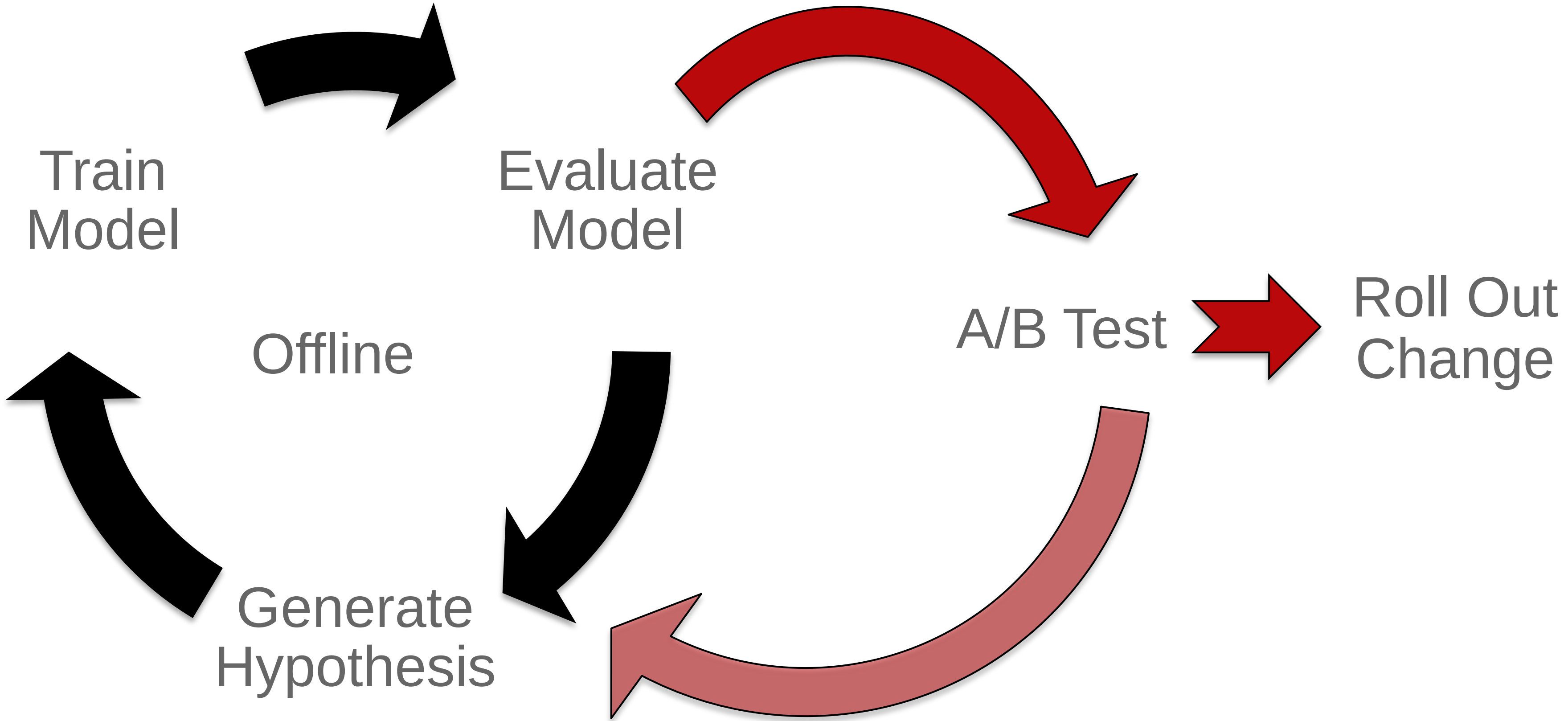
Individuals



Any can provide valuable signals

Evolution occurs through experimentation

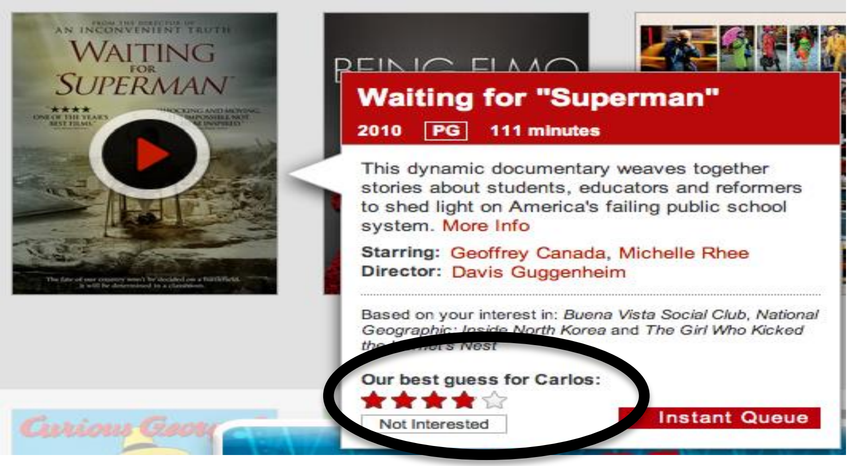
Faster Innovation Through Offline Testing



Offline Metrics

- Offline metrics help guide decisions on what to A/B test
  - Understand metric limitations and ignore as needed
- No offline set of metrics is predictive enough of cancelation rates
- Some metrics predict *local* algorithm metrics
  - In-line with the way algorithms are optimized

Root Mean Squared Error (RMSE)



$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Historically used to measure accuracy of predicted star ratings; good for offline optimization?

Why would RMSE improvement be a key driver to increase retention?

Not Interested

Our best guess for Carlos: **4.4 stars**

Average of 2,514,641 ratings: **4.4 stars**

vs.

Not Interested

Our best guess for Carlos: **3.2 stars**

Average of 23,041 ratings: **2.7 stars**

Personalized Video Ranking

Personalized Video Ranking

- TopN problem
- Natural metrics come from information retrieval:
  - Mean reciprocal rank
  - Precision
  - Recall
  - …
- But which correlate with cancelation rates and overall usage?

Interesting Challenges in Algorithms

- How do we develop recommender systems that directly optimize long term goals (user retention and overall consumption) offline?
- The effect of presentation bias
  - Can any offline metric help?
  - Can we remove this bias from our signals and algorithms?
- What's the best way to define the space of rows of videos?
- What's the best way to construct a page of recommendations?
- How can we best cold-start users and videos?

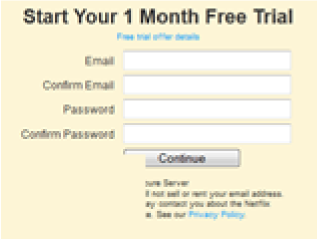
Experimentation

Controlled Experiment

Target population

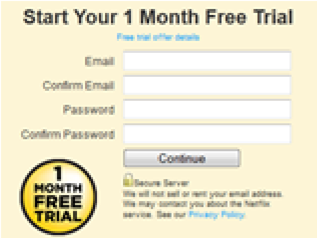
Random distribution

Version 'A'



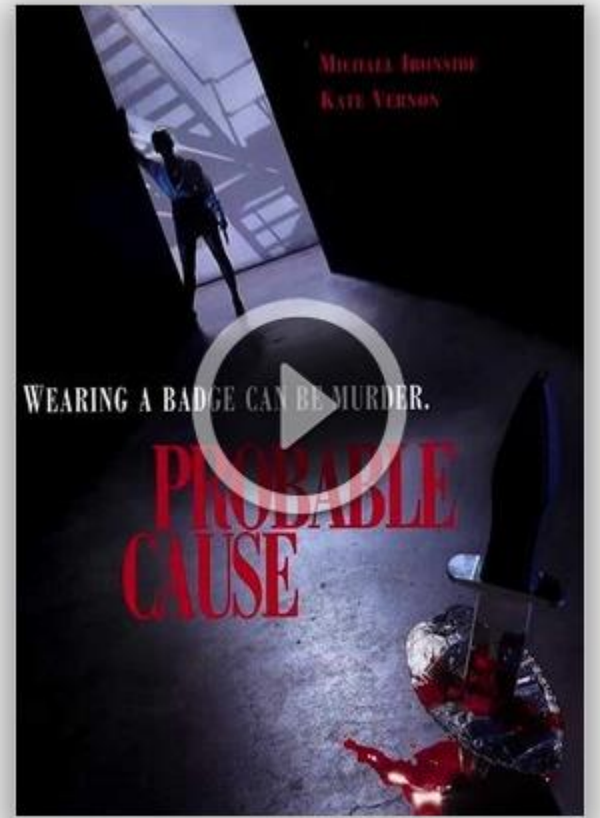
Identical except for the treatment being tested!

Version 'B'



Analyze & compare key metrics (with statistical confidence measures)

The Appeal: Causality



Probable Cause

1994 R 1hr 31m

Not Interested

Our best guess for Caitlin overall: **2.8 stars**

Average of 37,551 ratings: **3.3 stars**

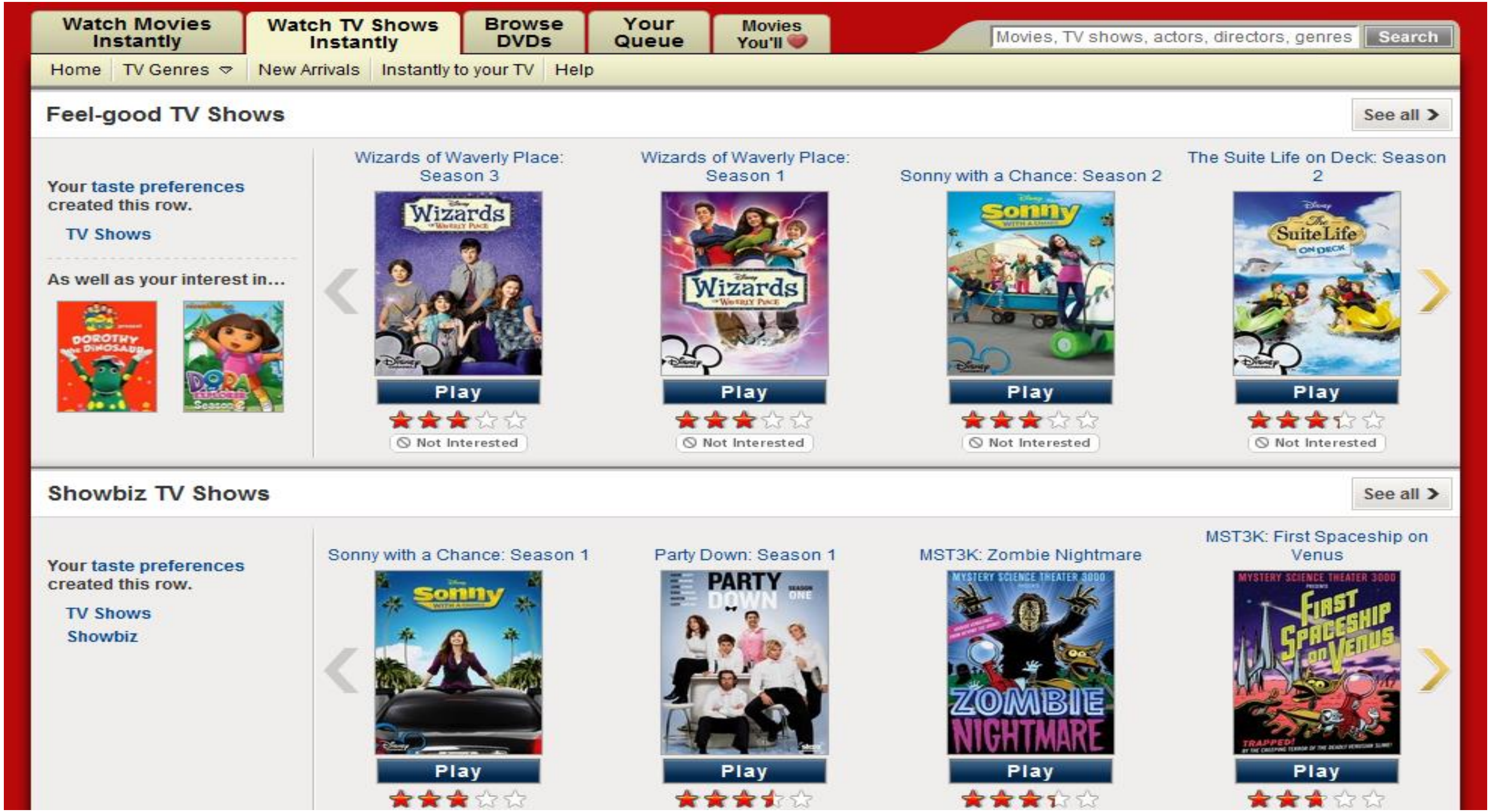
Police detective Gary Yanuck and his partner face a high-pressure engagement when they're tasked with nabbing a serial killer who's already offed a string of police officers and shows no sign of slowing down.

+ My List **Recommend to a friend**

Ingredients of great experimentation

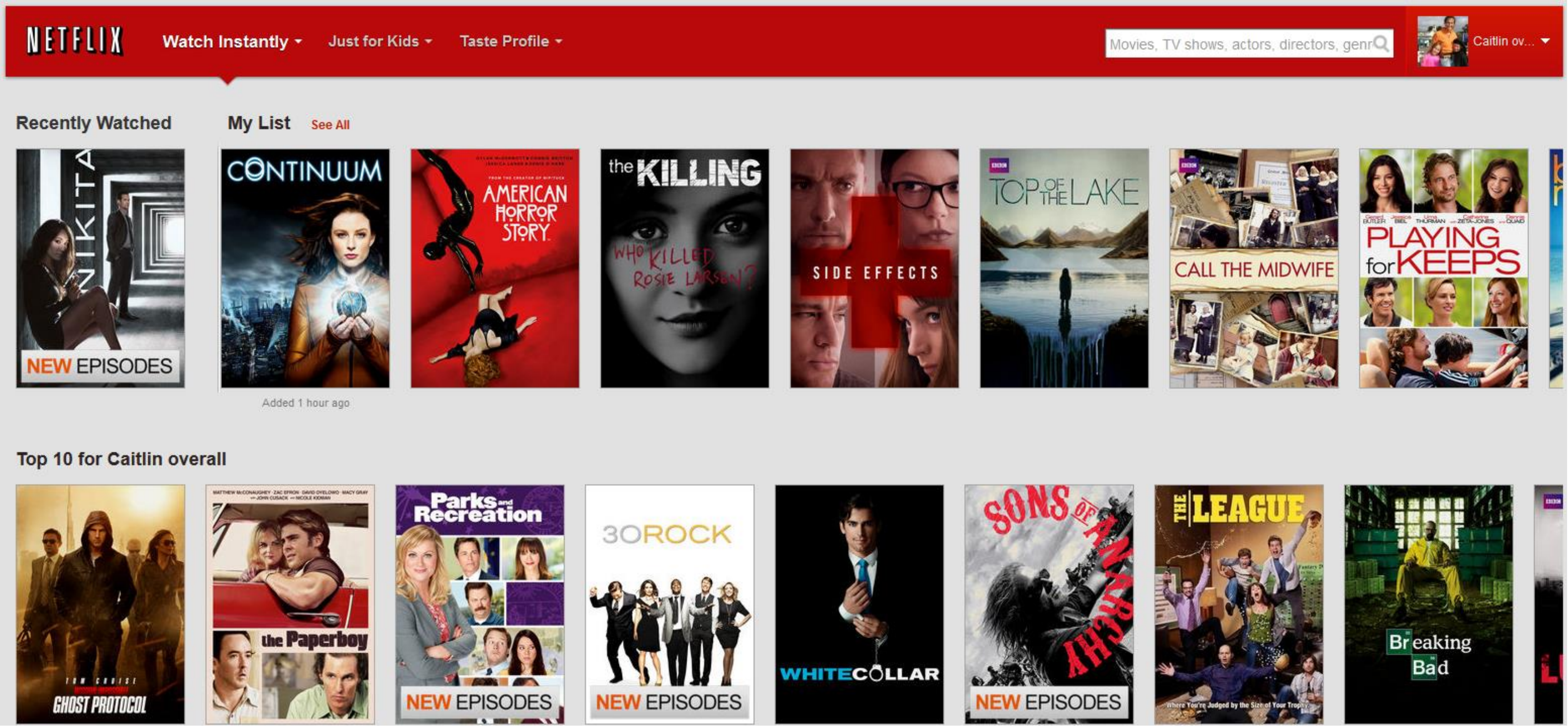
- Innovation and prioritization of impactful tests
- Experimental design (methodology, test cell design, sampling…)
- Execution of controlled experiment
- Accuracy (of data, engineering, statistics)
- Proper decision-making metrics & measurement techniques
- Pace & agility
- Interpretation and decision-making

2010



Watch Movies Instantly
Watch TV Shows Instantly
Browse DVDs
Your Queue
Movies You'll
Movies, TV shows, actors, directors, genres
Search
Home
TV Genres
New Arrivals
Instantly to your TV
Help
Feel-good TV Shows
See all
Your taste preferences created this row.
TV Shows
As well as your interest in...
Wizards of Waverly Place: Season 3
Wizards of Waverly Place: Season 1
Sonny with a Chance: Season 2
The Suite Life on Deck: Season 2
Play
Not Interested
Showbiz TV Shows
See all
Your taste preferences created this row.
TV Shows
Showbiz
Sonny with a Chance: Season 1
Party Down: Season 1
MST3K: Zombie Nightmare
MST3K: First Spaceship on Venus
Play

Now


NETFLIX
Watch Instantly
Just for Kids
Taste Profile
Movies, TV shows, actors, directors, genr
Caitlin ov...
Recently Watched
My List
See All
NIKITA
NEW EPISODES
CONTINUUM
AMERICAN HORROR STORY
the KILLING
WHO KILLED ROSIE LARSEN?
SIDE EFFECTS
TOP OF THE LAKE
CALL THE MIDWIFE
PLAYING for KEEPS
Added 1 hour ago
Top 10 for Caitlin overall
GHOST PROTOCOL
the Paperboy
Parks and Recreation
NEW EPISODES
30 ROCK
NEW EPISODES
WHITE COLLAR
SONS OF ANARCHY
NEW EPISODES
THE LEAGUE
Breaking Bad

Characteristics specific to Netflix testing

Challenges

- Sampling of new members has efficiency limitations
- Monthly billing cycles increase our testing timelines
- Breadth of devices and UIs impact pace of execution and add complexity across the ecosystem

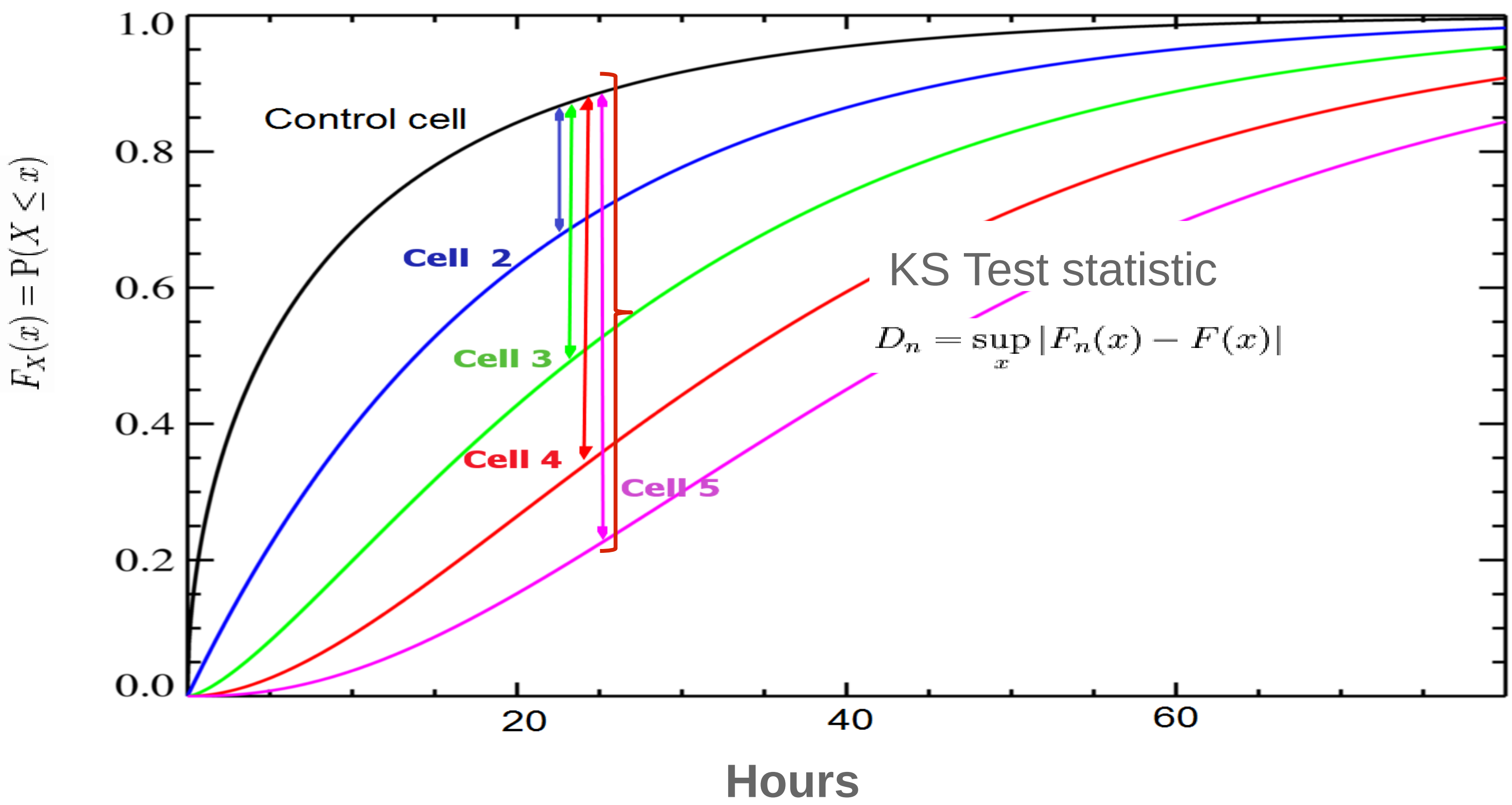
Assets

- Clear core metrics
- Member identification (logged-in, paying customers)
- Great data
- Bias toward product simplicity
- Culture of learning, openness, & debate
- Executive commitment & participation

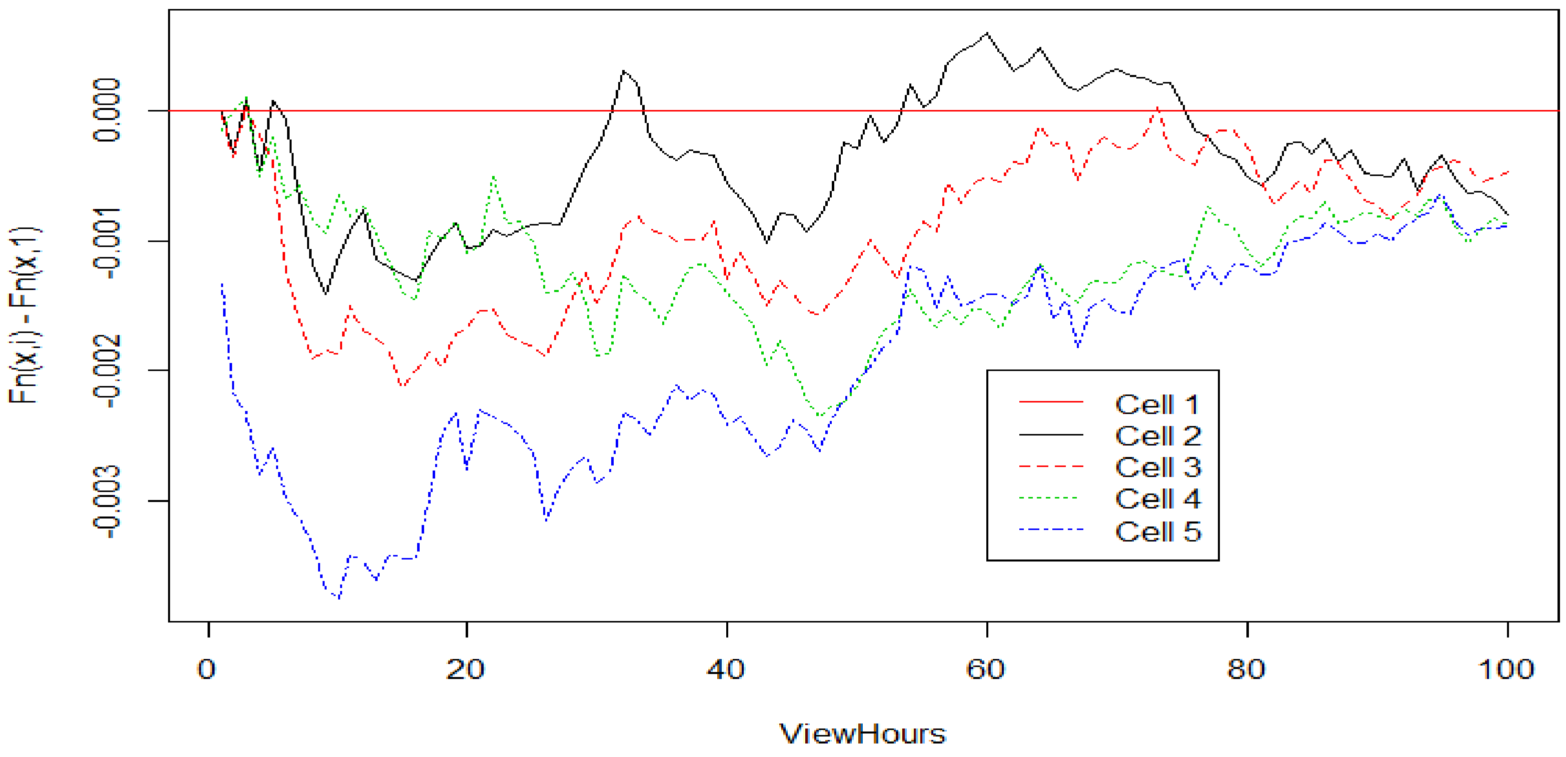
Metrics

- Cumulative Retention
- Streaming
- Many other “secondary” engagement metrics

Streaming measurement: Kolmogorov-Smirnov (KS) test

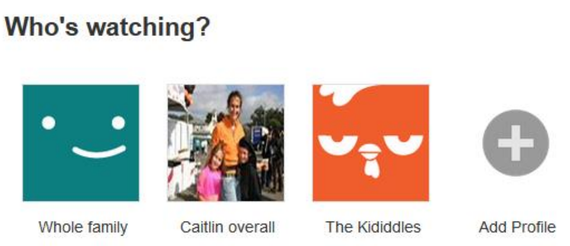


Streaming measurement: KS example



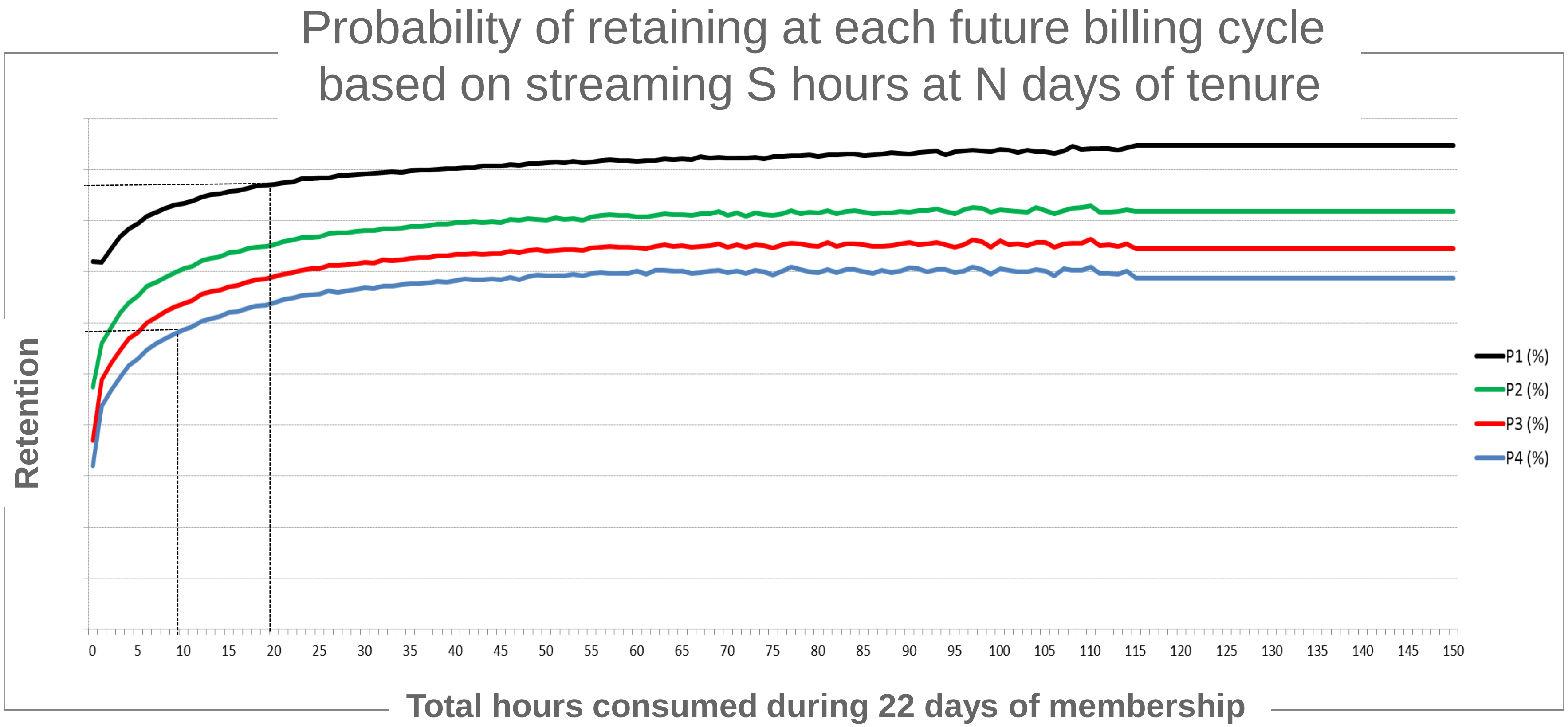
Streaming measurement: Thresholds with z-tests for proportions

Profiles win confirmation test



Display Cell		1	2
Cell Name		Profiles Enabled	Holdback
Comparison Cell	Set All	1	1
% Accounts with > 0 Hours		93.5%	93.5% 0.619
% Accounts with >= 1 Hour		89.1%	89.0% 0.440
% Accounts with >= 5 Hours		80.5%	80.4% 0.258
% Accounts with >= 10 Hours		72.6%	72.4% 0.089
% Accounts with >= 20 Hours		59.7%	59.2% 0.001
% Accounts with >= 40 Hours		40.7%	40.2% 0.000
% Accounts with >= 80 Hours		18.9%	18.5% 0.001

Streaming measurement:
Streaming score model

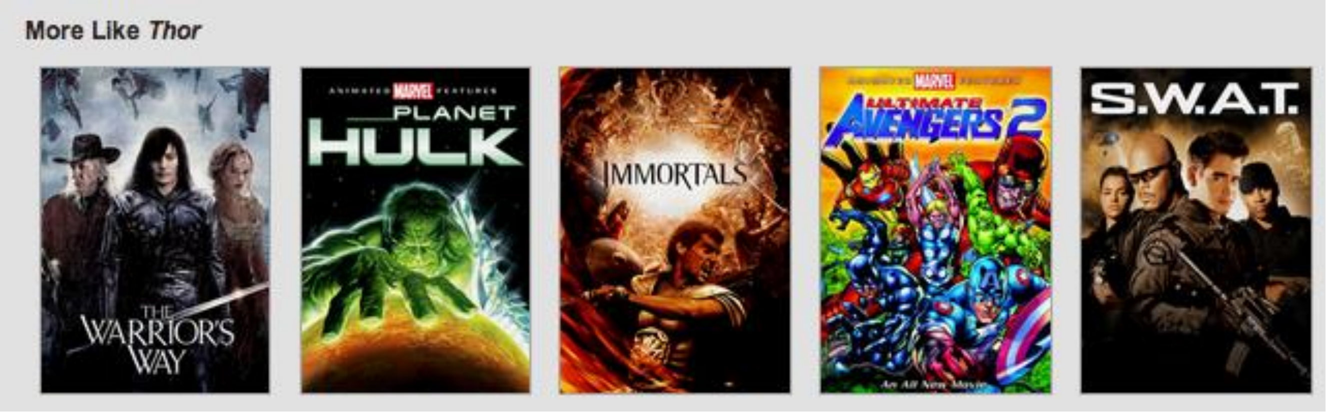


Challenges with “hours”

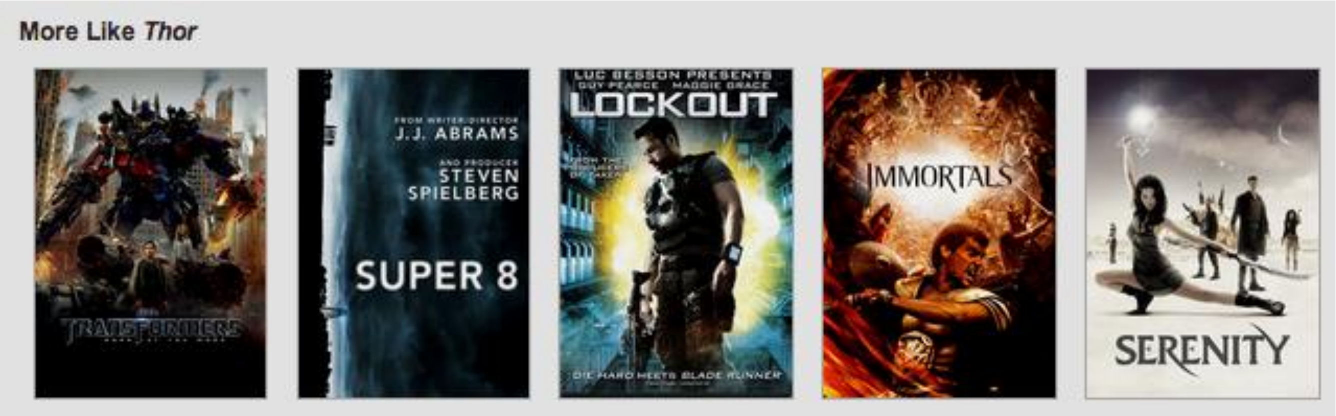
- Not all “hours” have equal value to customers
- TV vs features have dramatically different consumption rates
- Service is available after cancel request
- Timespan for hours measurement

“Similars Algorithm” Experiment

Algorithm A



Algorithm B



What should we measure in this test?

- Ideas?
- Retention & overall streaming
- CTR on Similars rows; Share of hours from similars rows?
  - Should we care about cannibalization?
- Horizontal position played?
- Should we measure whether the new algorithm generated results that were more “similar”?
- What does the customer expect out of the row based on its label?
- Did the customer enjoy the titles more even if he/she did not watch more in total hours?

How might we know whether a customer enjoyed a title?

- Gave it a high rating
  - But only a subset of users rate
- Came back to watch again
  - Different opportunity for a TV show vs movie
- Fraction of content viewed
  - $FCV = \frac{duration\ watched}{title\ runtime}$

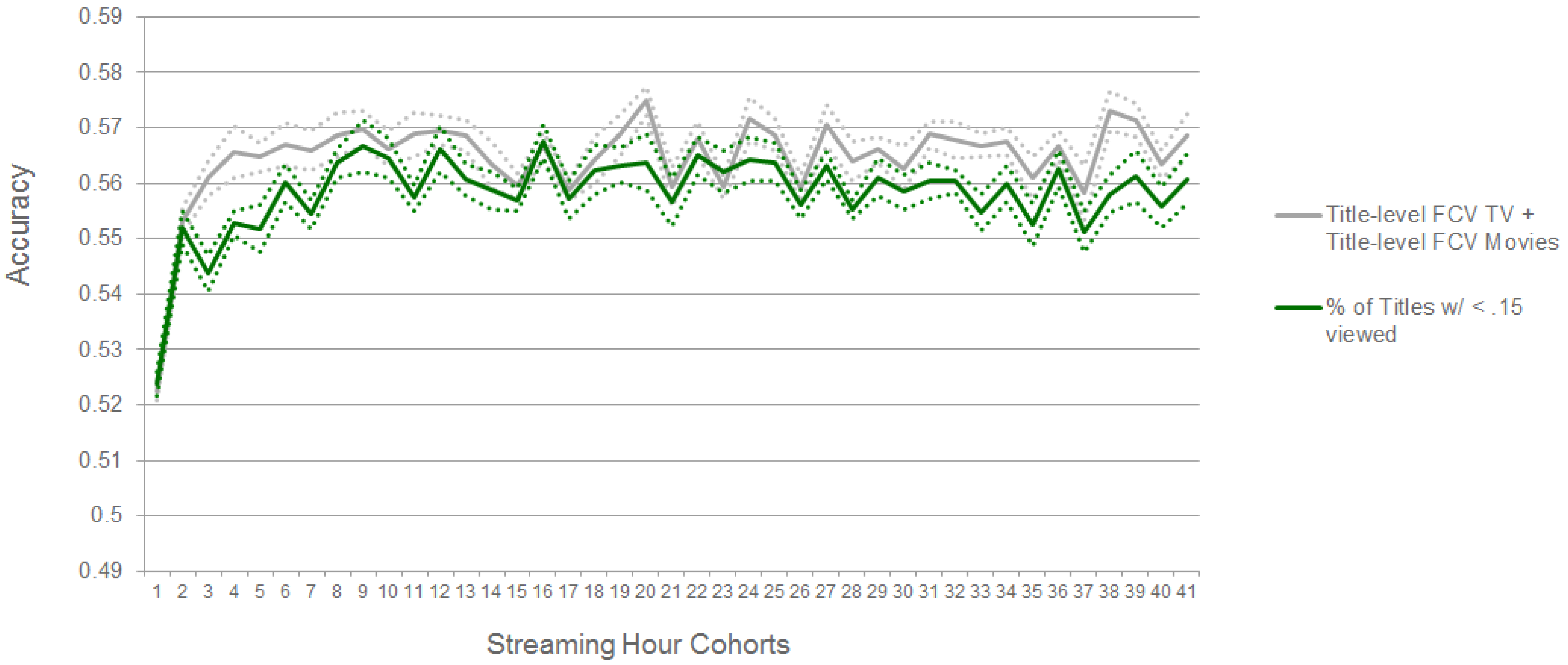
Fraction of Content Viewed ("FCV")

- Average FCV
- # of titles viewed/# of streaming hours
- % of Titles with FCV <= 15%
- % of sessions with FCV <= 15%
- % of Active Days with a Play >= 90%
- % of Play Days with a Full Play
- % of Hours from Browse Plays
- …

Variants for:
- TV vs movies
- Episode, season, show
- Timeframes
- How the title was found

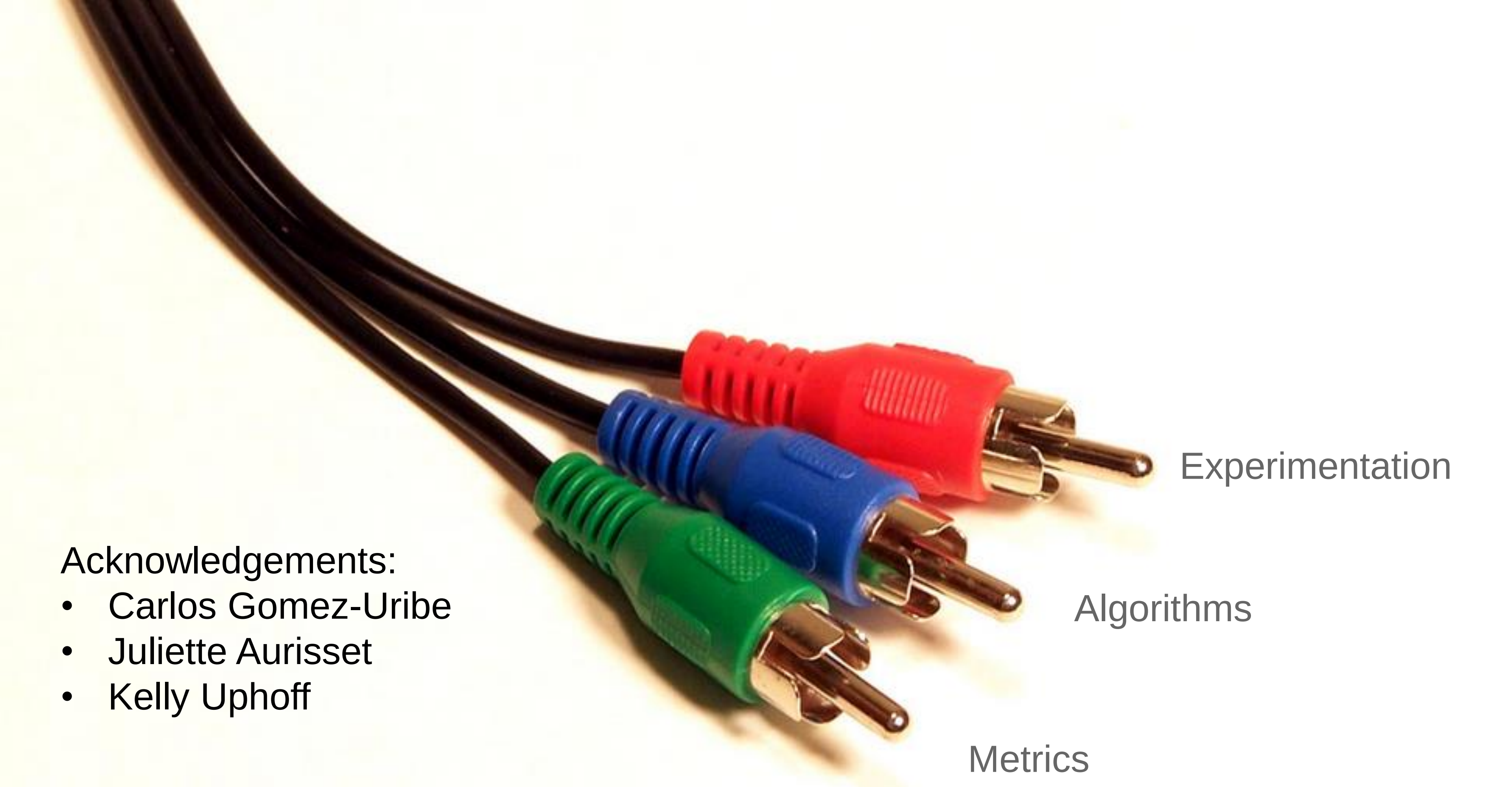
Nearly every engagement metric is highly correlated with total streaming hours

Best metric variants do provide some lift



Controlling for streaming hours, these metrics improve retention prediction

New metrics are often tested as algorithm input signals (and vice-versa)



Acknowledgements:
- Carlos Gomez-Uribe
- Juliette Aurisset
- Kelly Uphoff