Implicit Feedback and Performance Evaluation in Recommender Systems

Shay Ben Elazar **Mike Gartrell** **Noam Koenigstein** **Gal Lavee**

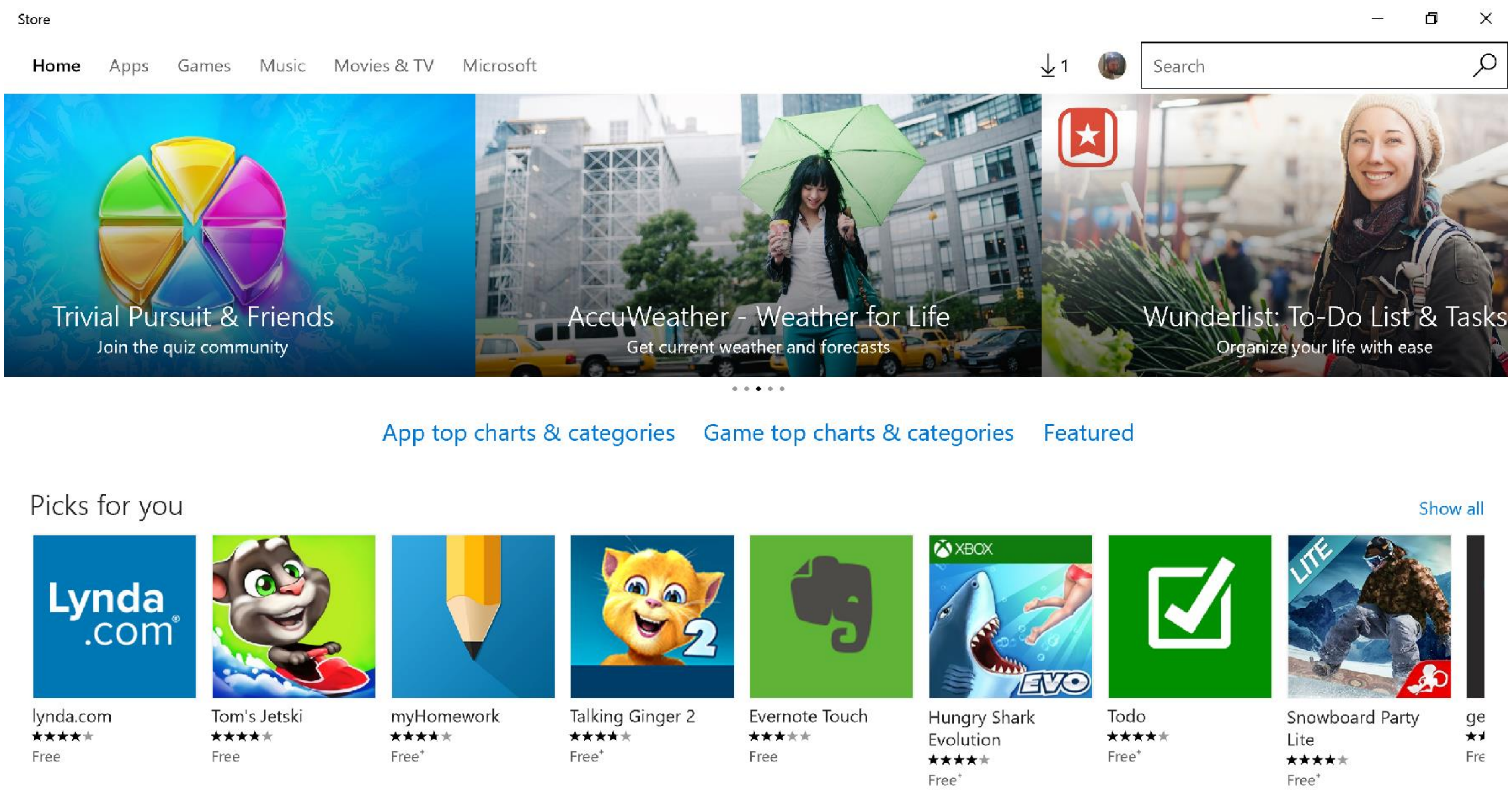
Microsoft

Agenda

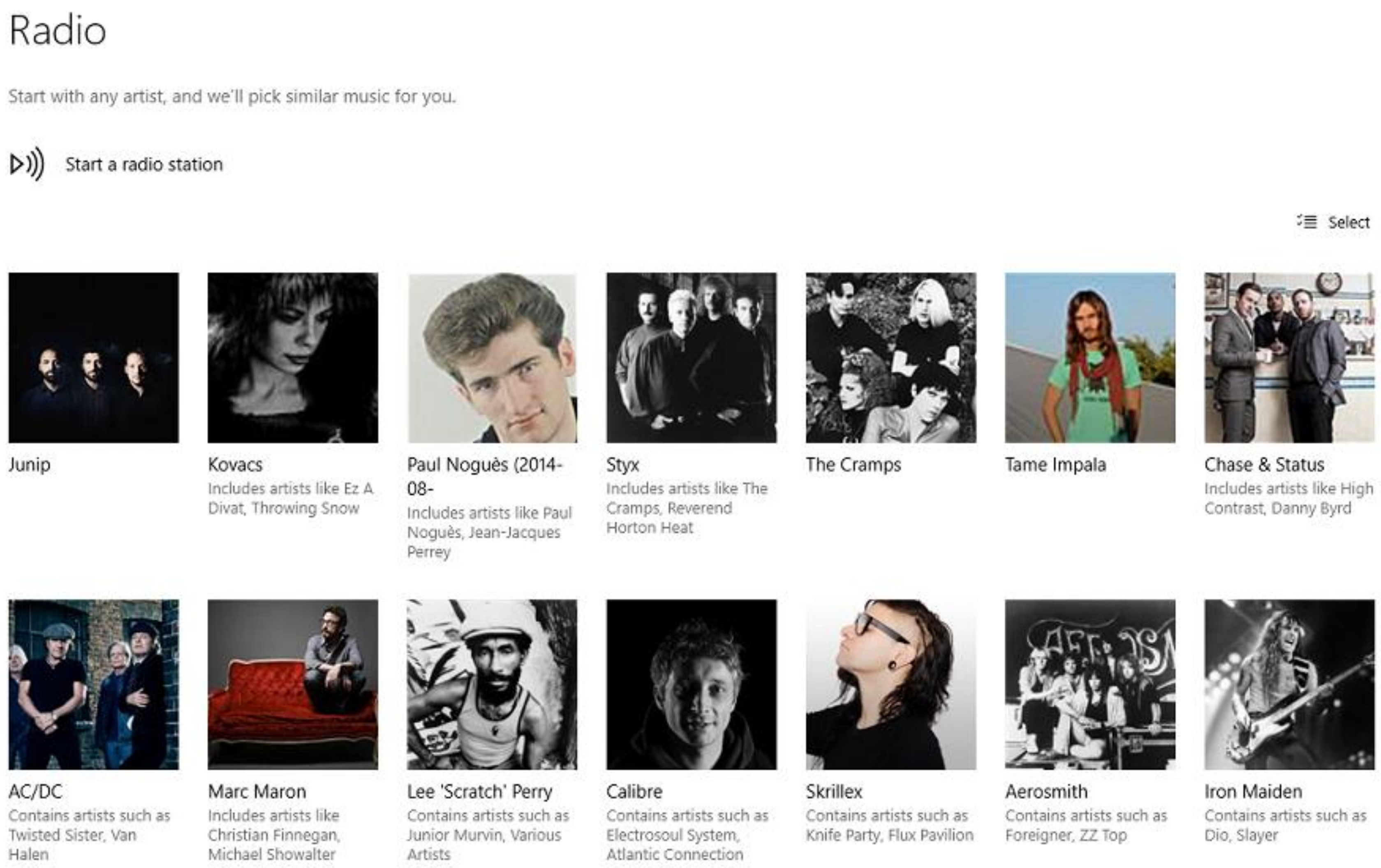
- Intro Universal Store Recommendations
- Extreme Classification with Matrix Factorization
- Offline Evaluation Techniques
- Online Evaluation
- The Gap
- Bridging The Gap…

Microsoft Universal Store Recommendations

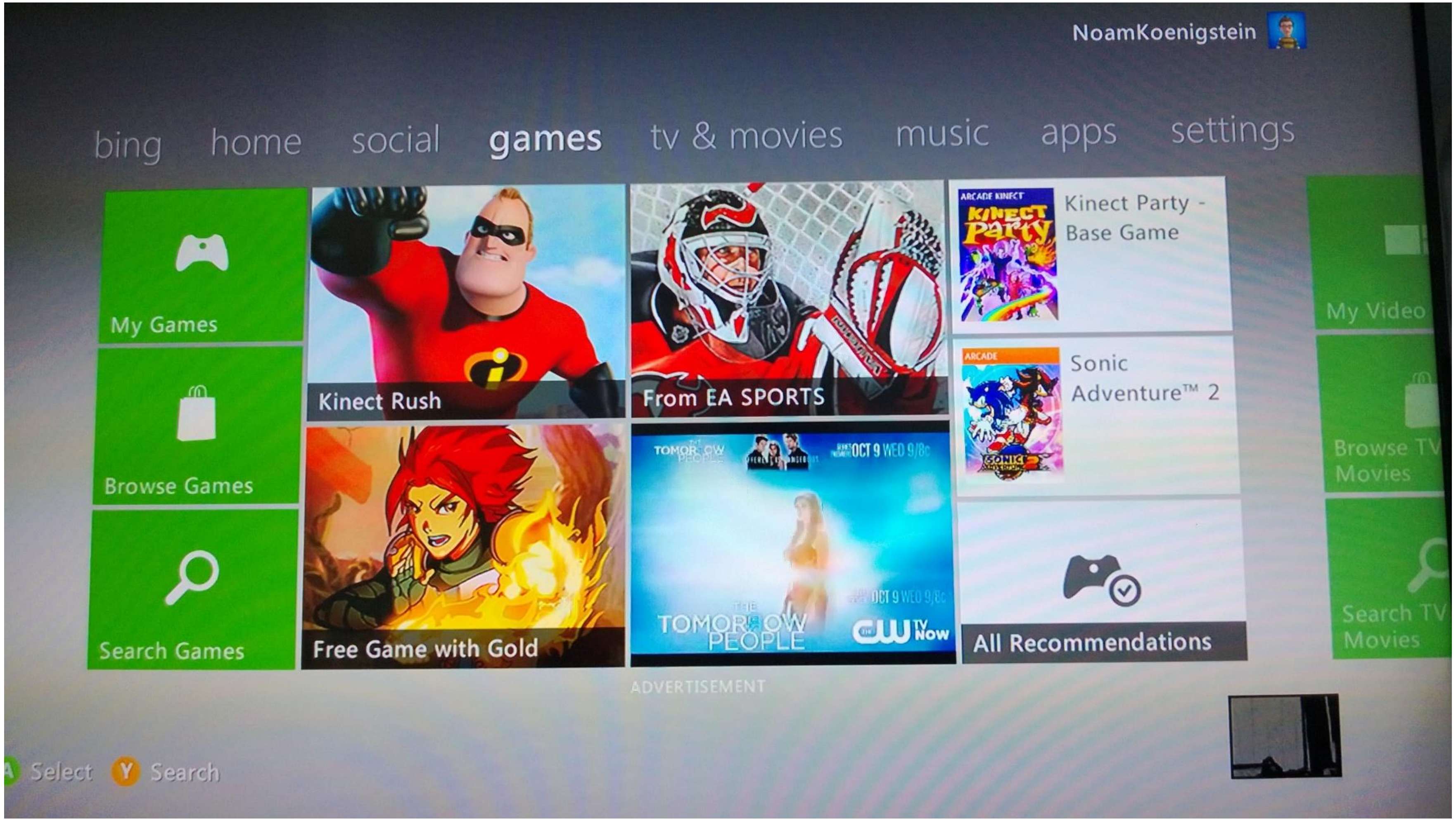
Windows Store



Groove Music



Xbox

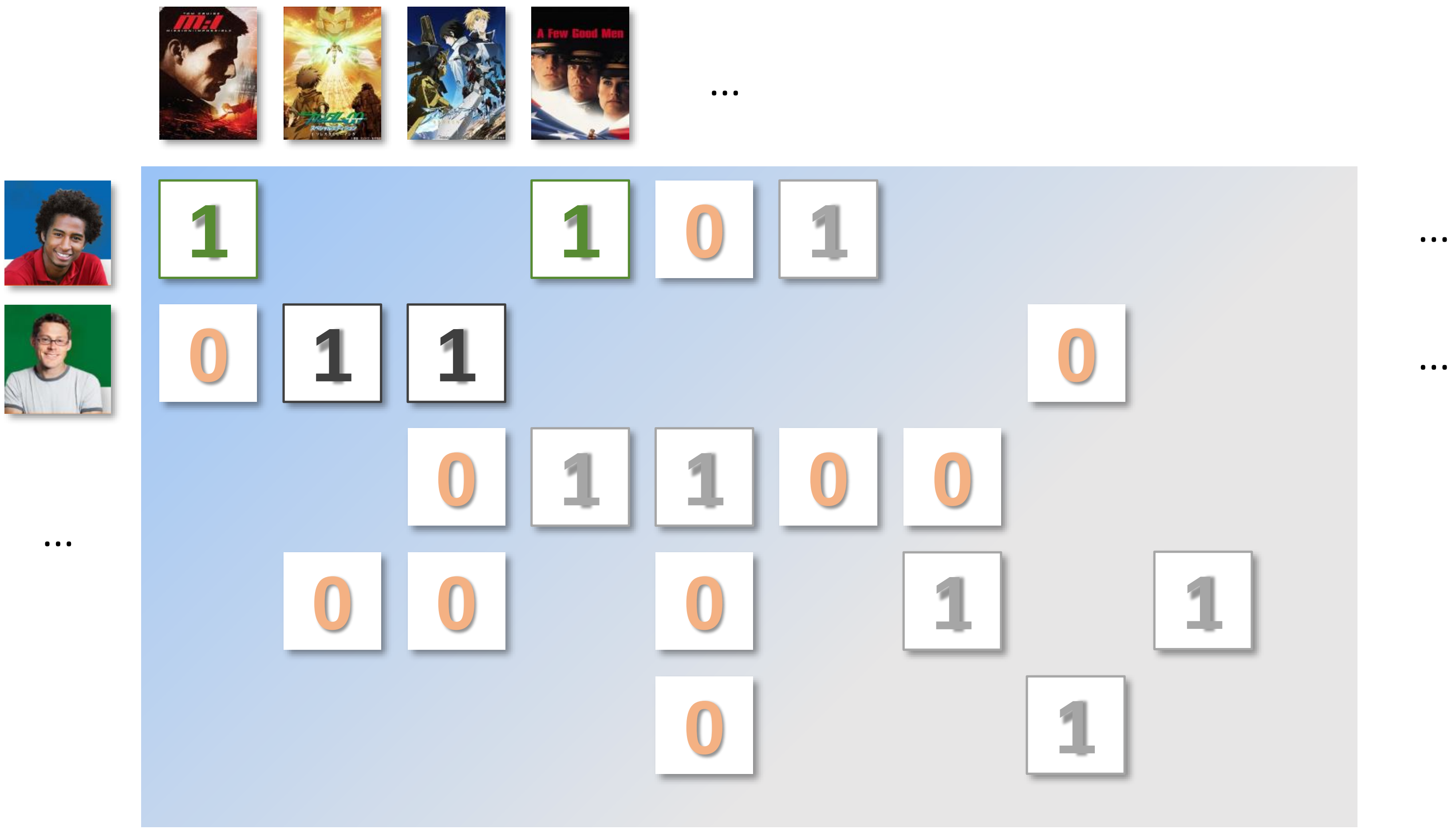


NoamKoenigstein
bing
home
social
games
tv & movies
music
apps
settings
My Games
Kinect Rush
From EA SPORTS
Kinect Party - Base Game
My Video
Browse Games
Sonic Adventure™ 2
Browse TV Movies
Free Game with Gold
All Recommendations
Search Games
Search TV Movies
ADVERTISEMENT
Select
Search

Extreme Classification with Matrix Factorization

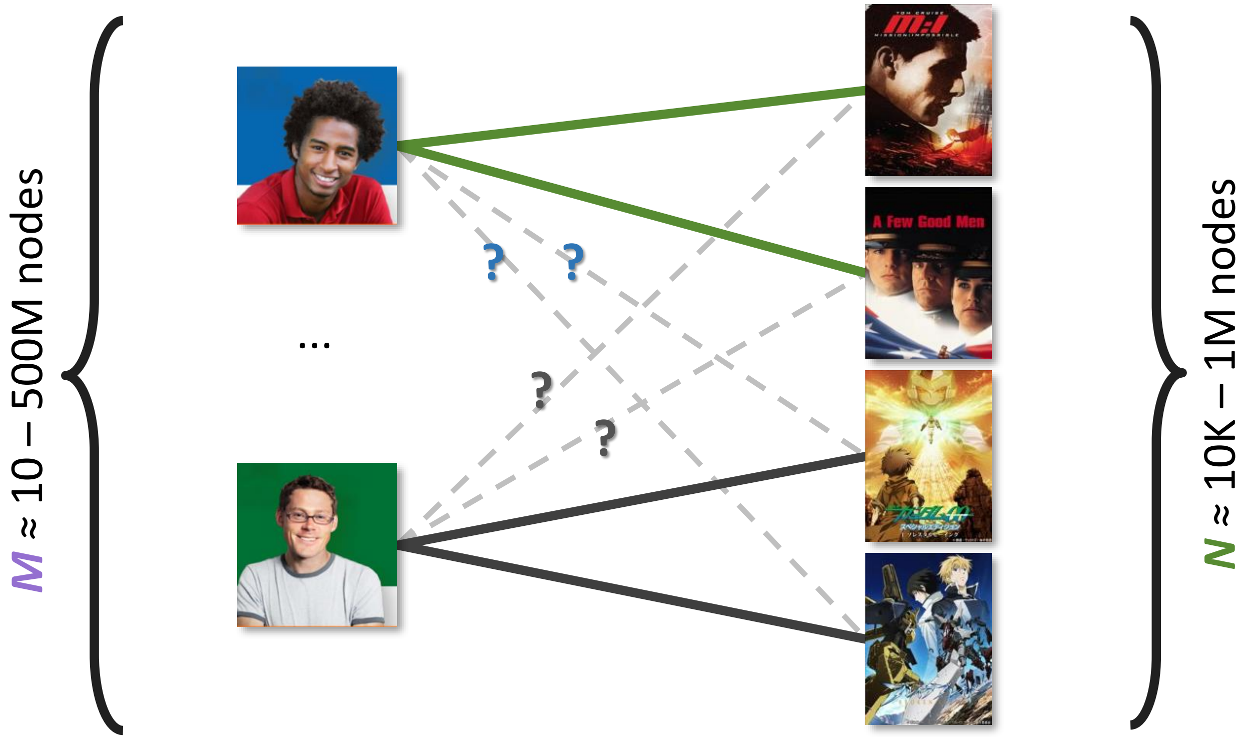
History: Netflix Prize

Two-class data – Extreme Classification



One-class data

Problem formulation



Bipartite graph → We care about ? = *p*(***link***)

Fully Bayesian model based on Variational Bayes optimization

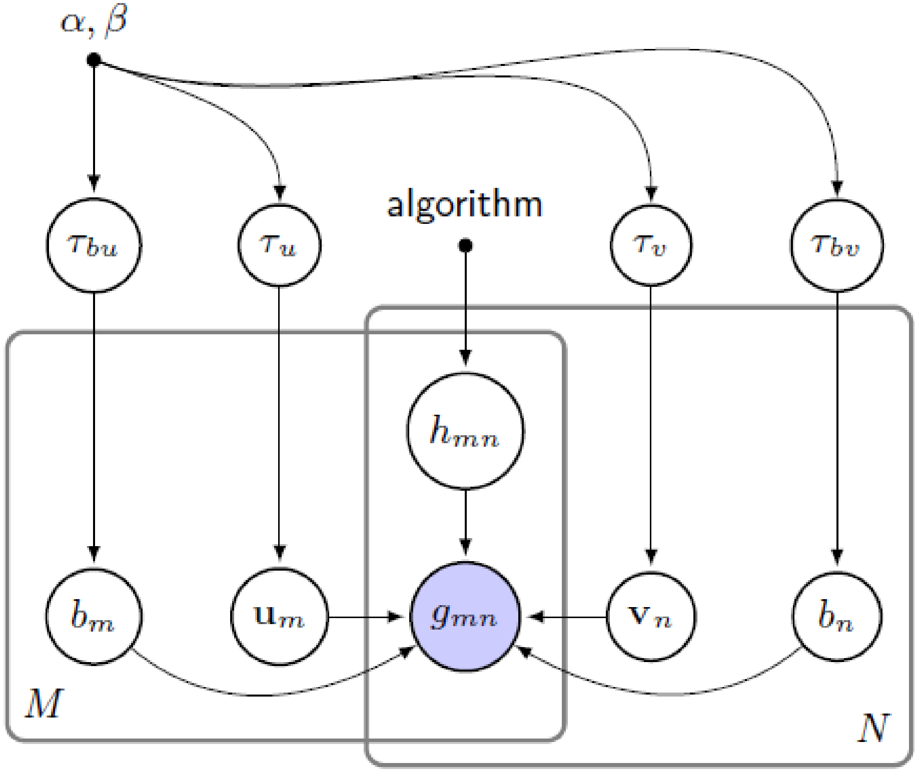


Figure 2: The graphical model for observing graph G connecting M user with N item vertices. The prior on the hidden graph **H** is algorithmically determined to resemble the type of the observed graph.

Offline Evaluation Techniques

$RMSE$- Root Mean Square Error

RMSE is computed by averaging the square error over all user item pairs,$(u, i) \in \mathcal{R}$

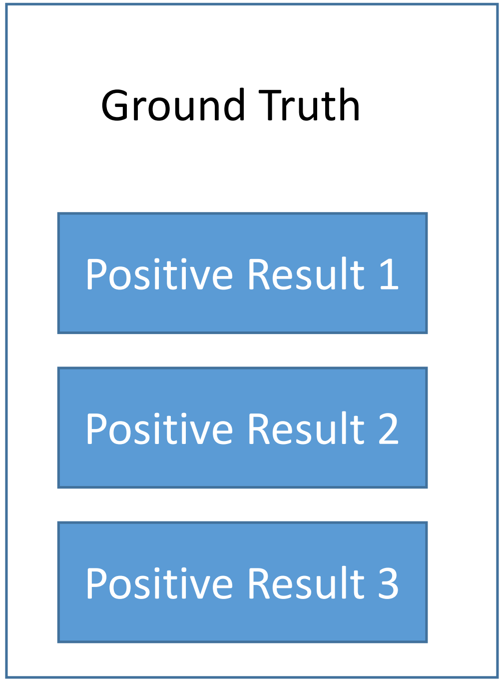
$$RMSE = \sqrt{\frac{1}{|\mathcal{R}|} \sum_{(u,i) \in \mathcal{R}} SE_{ui}}$$

$wRMSE$- Weighted Root Mean Square Error

This variant of RMSE is achieved by assigning each data point a weight, w_{ui}, based on its importance.

$$RMSE = \sqrt{\frac{1}{\sum w_{ui}} \sum_{(u,i)\in\mathcal{R}} w_{ui} \cdot SE_{ui}}$$

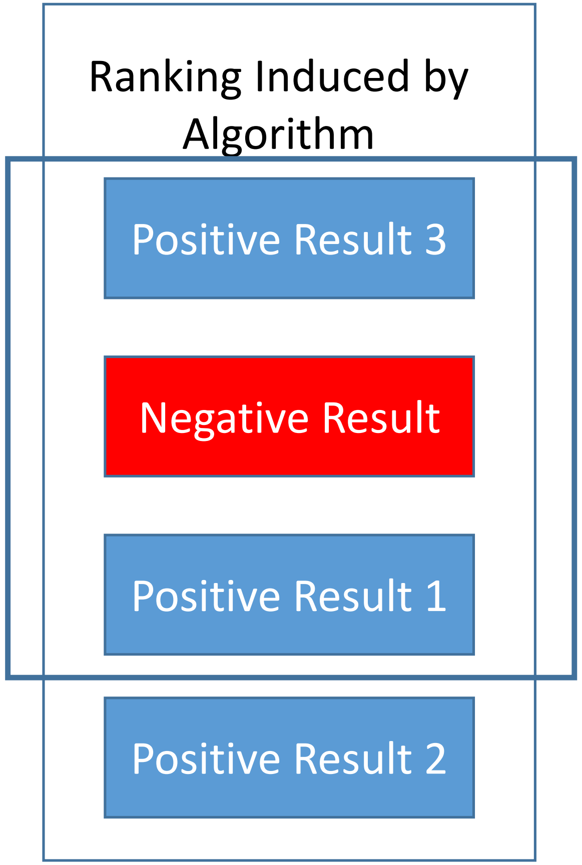
Precision@k / Recall@k



$$\boldsymbol{k = 3}$$

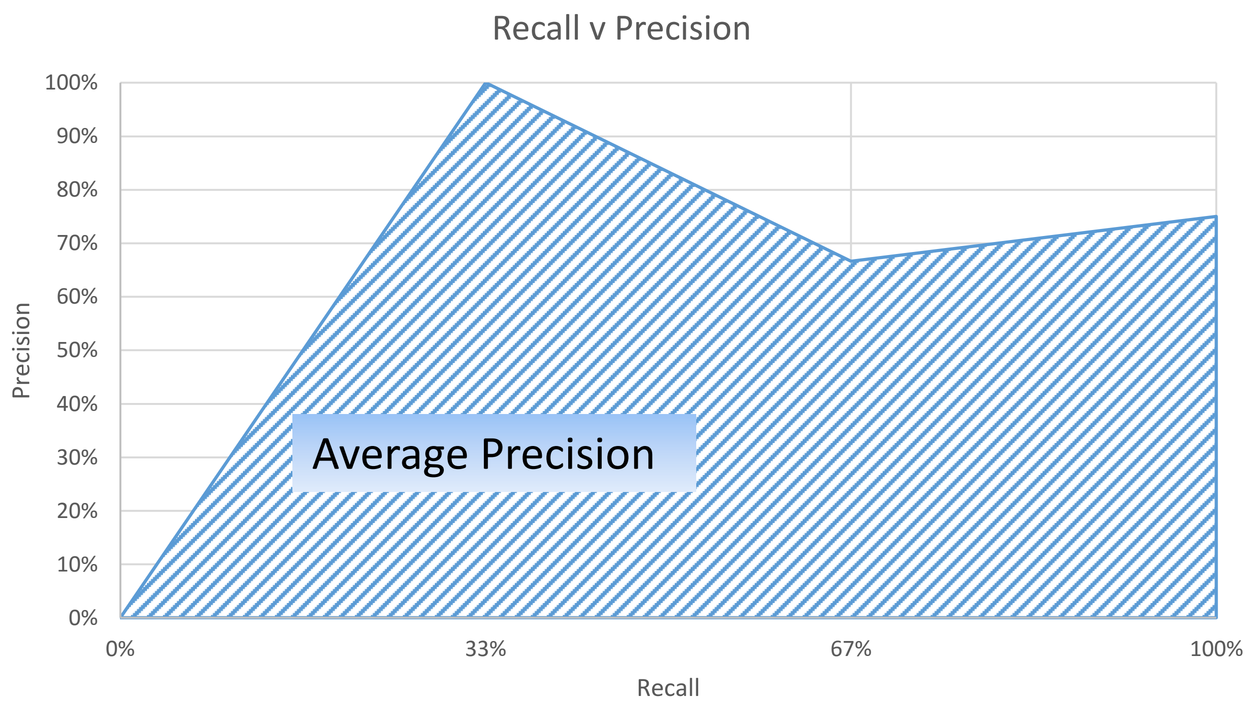
$$precision@k = \frac{2}{3}$$

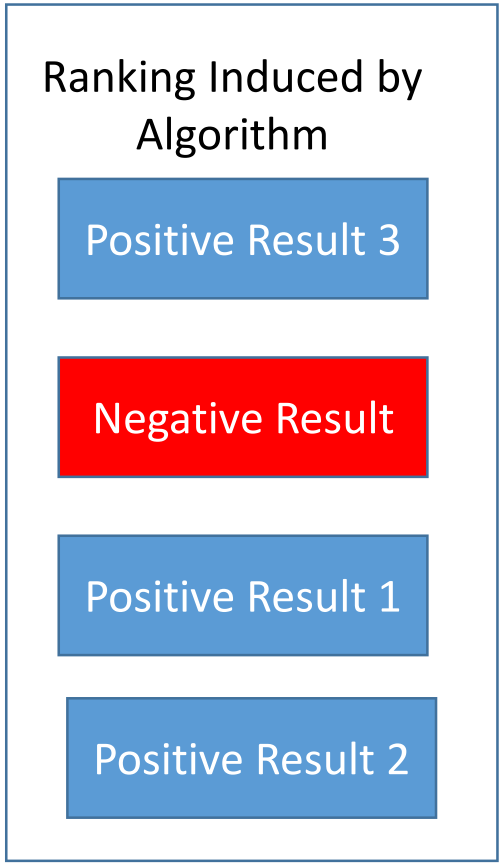
$$recall@k = \frac{2}{3}$$



Mean Average Precision

We can plot precision as a function of recall





$NDCG$ – *Normalized Discounted Cumulative Gain*

The relevance is discounted by $\gamma_i = \frac{1}{\log_2(i+1)}$ and the sum @ *k is normalized by* its upper bound – the $IDCG$

Ground Truth

Positive Result 1
Relevance: 5

Positive Result 2
Relevance: 3

Positive Result 3
Relevance: 1

$$\boldsymbol{k = 3}$$

$$DCG@k = \frac{1}{log_2(1+1)} + 0 + \frac{5}{log_2(3+1)} = 3.5$$

$$IDCG@k = \frac{5}{log_2(1+1)} + \frac{3}{log_2(2+1)} + \frac{1}{log_2(3+1)} = 7.39$$

$$\boldsymbol{NDCG@k = \frac{3.5}{7.39} = 0.47}$$

Ranking Induced by Algorithm

Positive Result 3

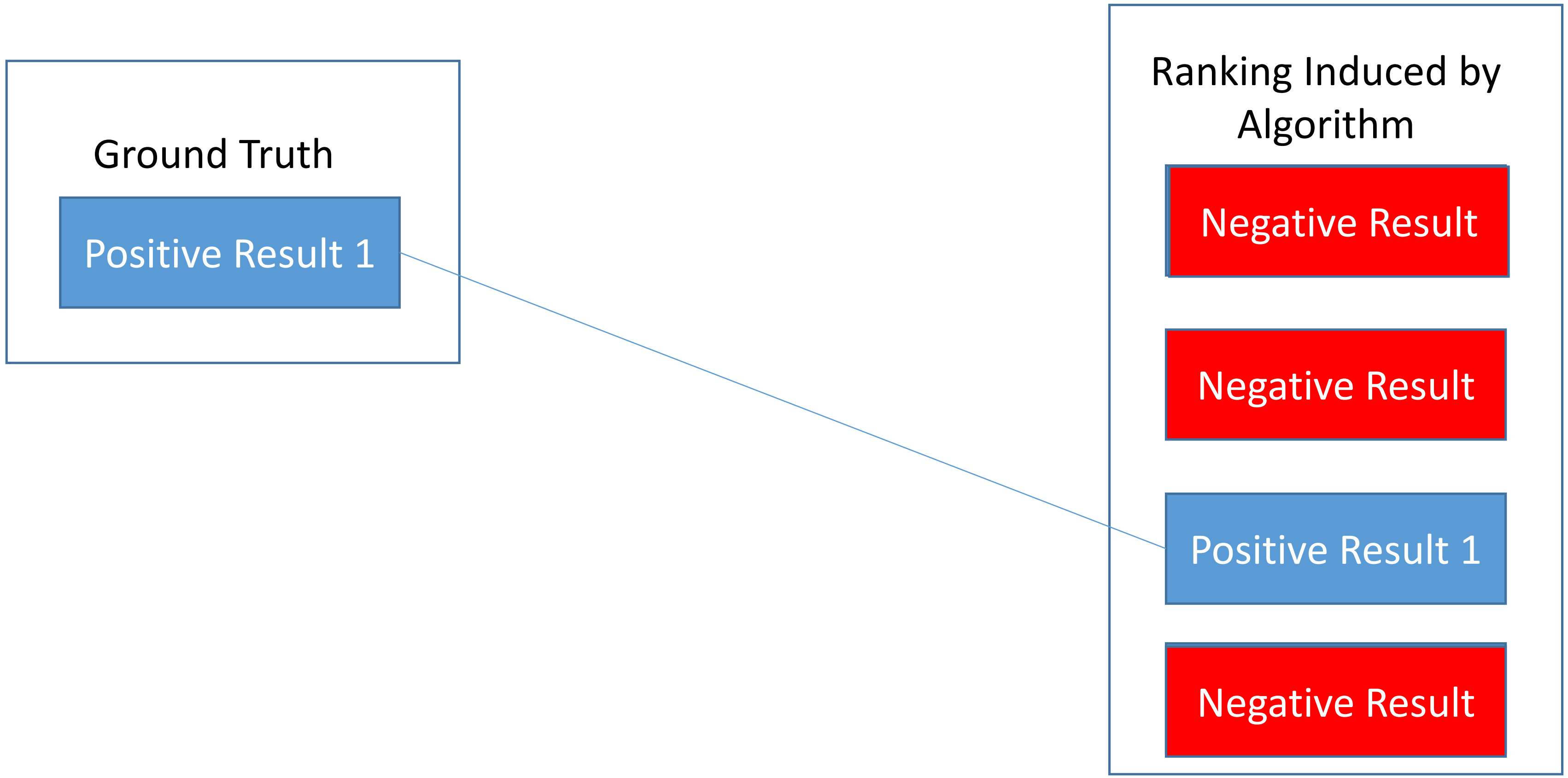
Negative Result

Positive Result 1

Positive Result 2

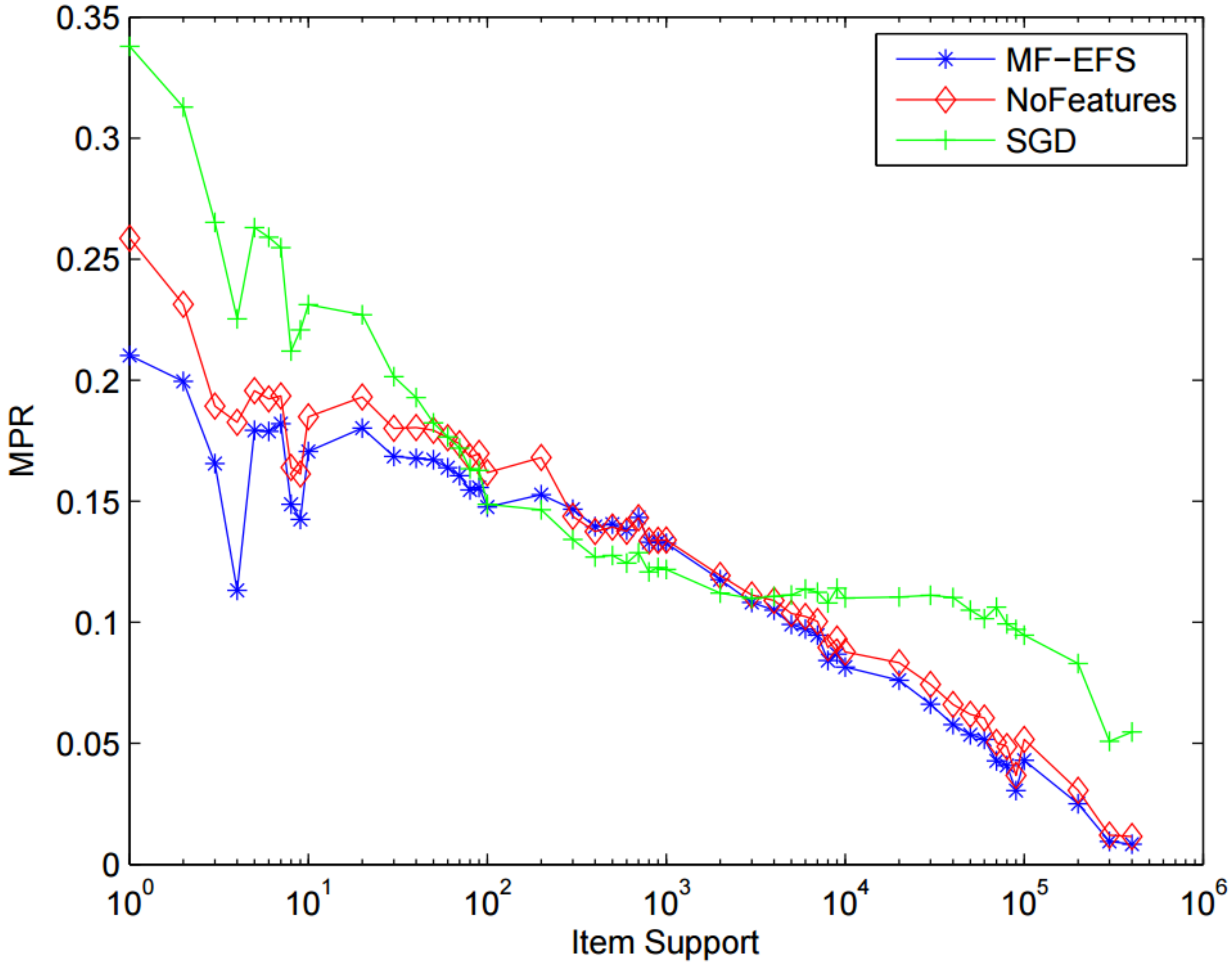
MPR- Mean Percentile Rank

Sometimes there is only one "positive" items in the test set…



$\boldsymbol{rank_i = 3}$
$\boldsymbol{MPR = 0.5}$

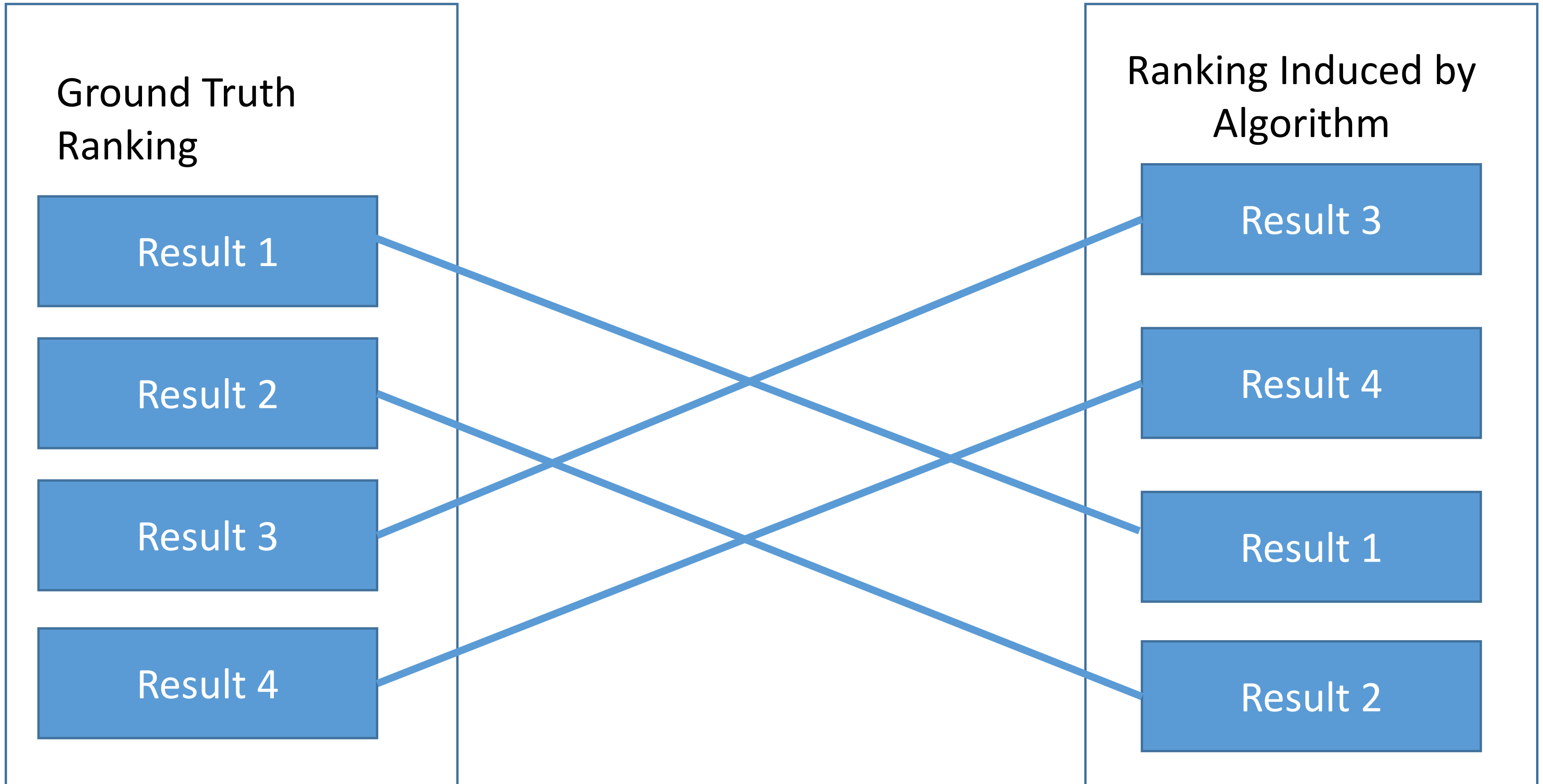
MPR in Xbox



(a) Xbox Movies

Spearman's Rho Coefficient

In scenarios where we want to emphasize the full ranking we may compare the ranking of the algorithm to a reference ranking



$$r_1 - \hat{r}_1 = 1 - 3$$

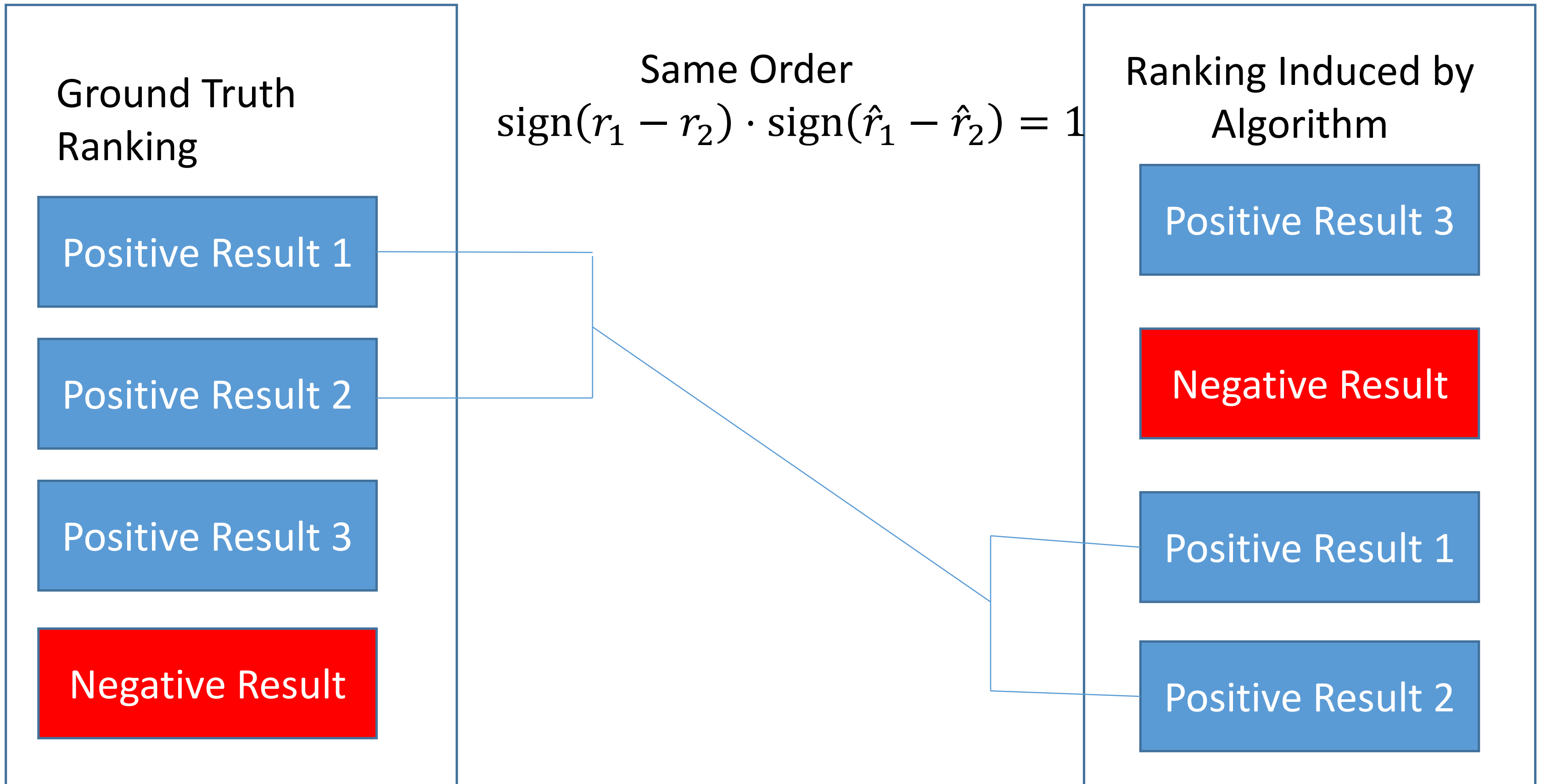
$$r_2 - \hat{r}_2 = 2 - 4$$

$$r_3 - \hat{r}_3 = 3 - 1$$

$$r_4 - \hat{r}_4 = 4 - 2$$

Kendall’s Tau Coefficient

In scenarios where we want to emphasize the full ranking we may compare the ranking of the algorithm to a reference ranking



Offline Techniques – Open Questions

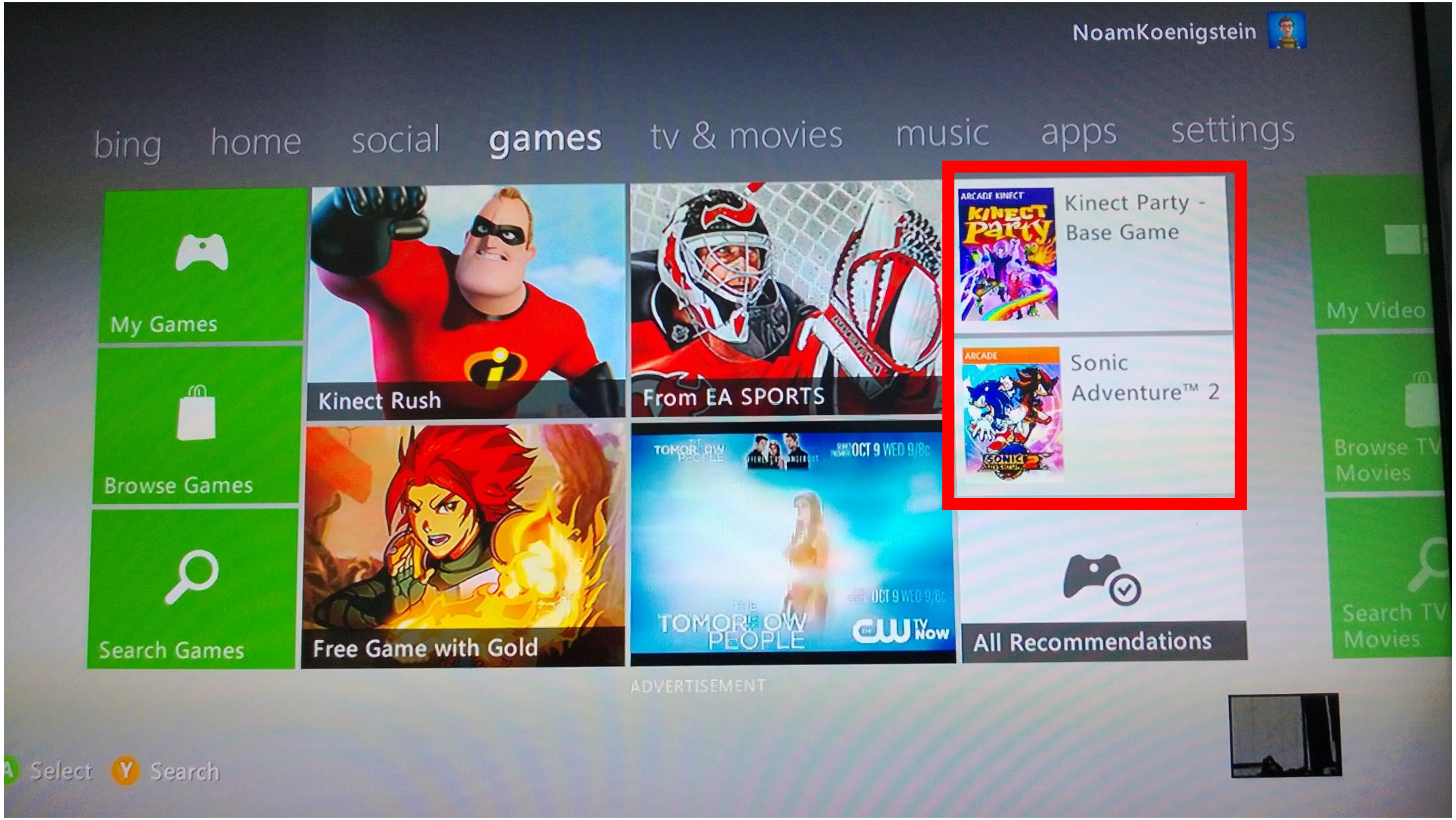
- How do we measure the importance/ relevance of the positive items?
  - Long tail items are more important. But how do we quantify?
  - How many items do we care to recommend?
- Should the best item be the first item?
  - Maybe the best item should be in the middle?
- What about diversity?
- What about contextual effects?
- What about items fatigue?

Online Experimentation

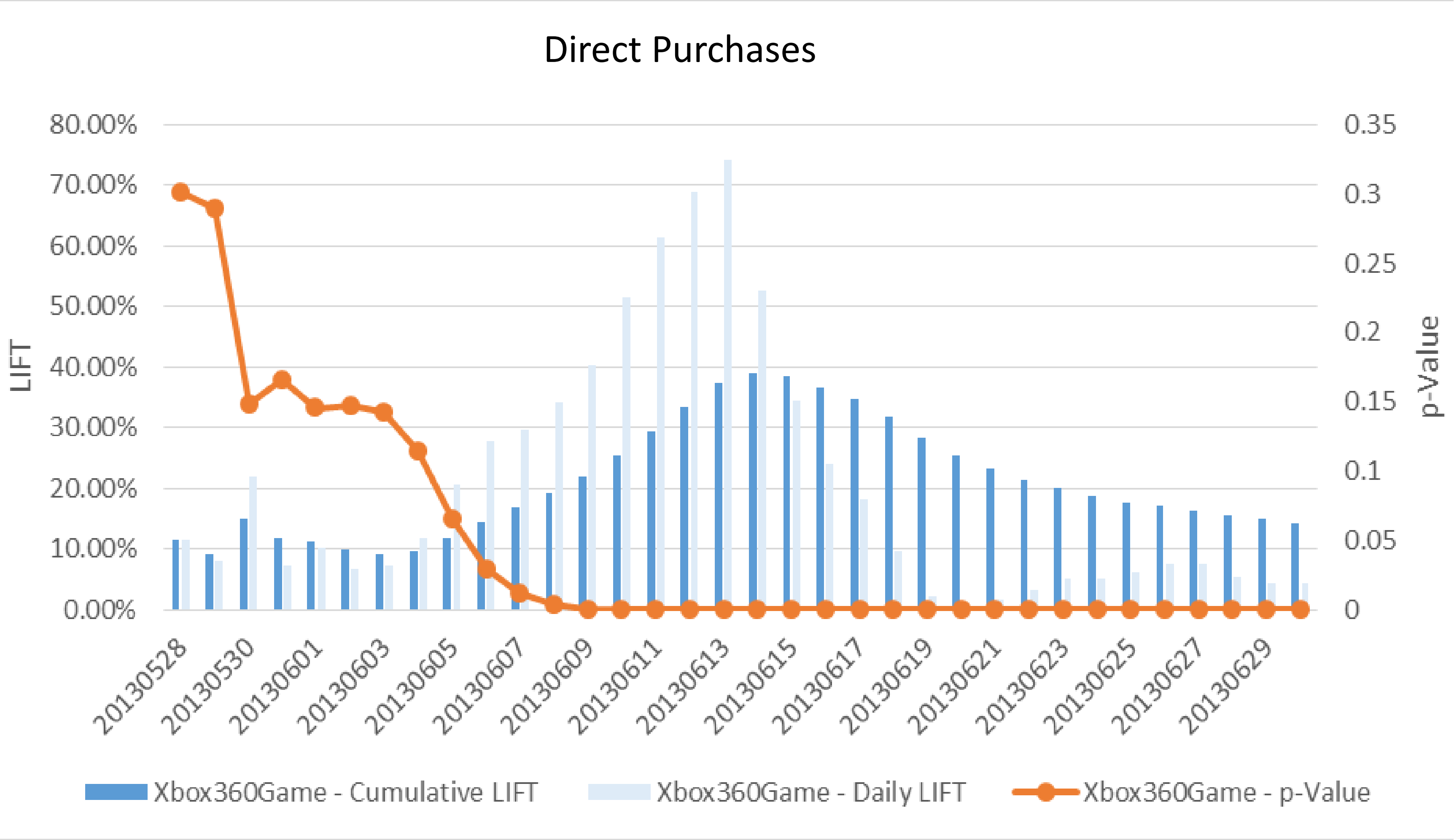
Online Experiments

- Randomized controlled experiments
- Measure KPIs (Key Performance Indicator) directly
- Can compare several variants simultaneously
- The ultimate evaluation technique!

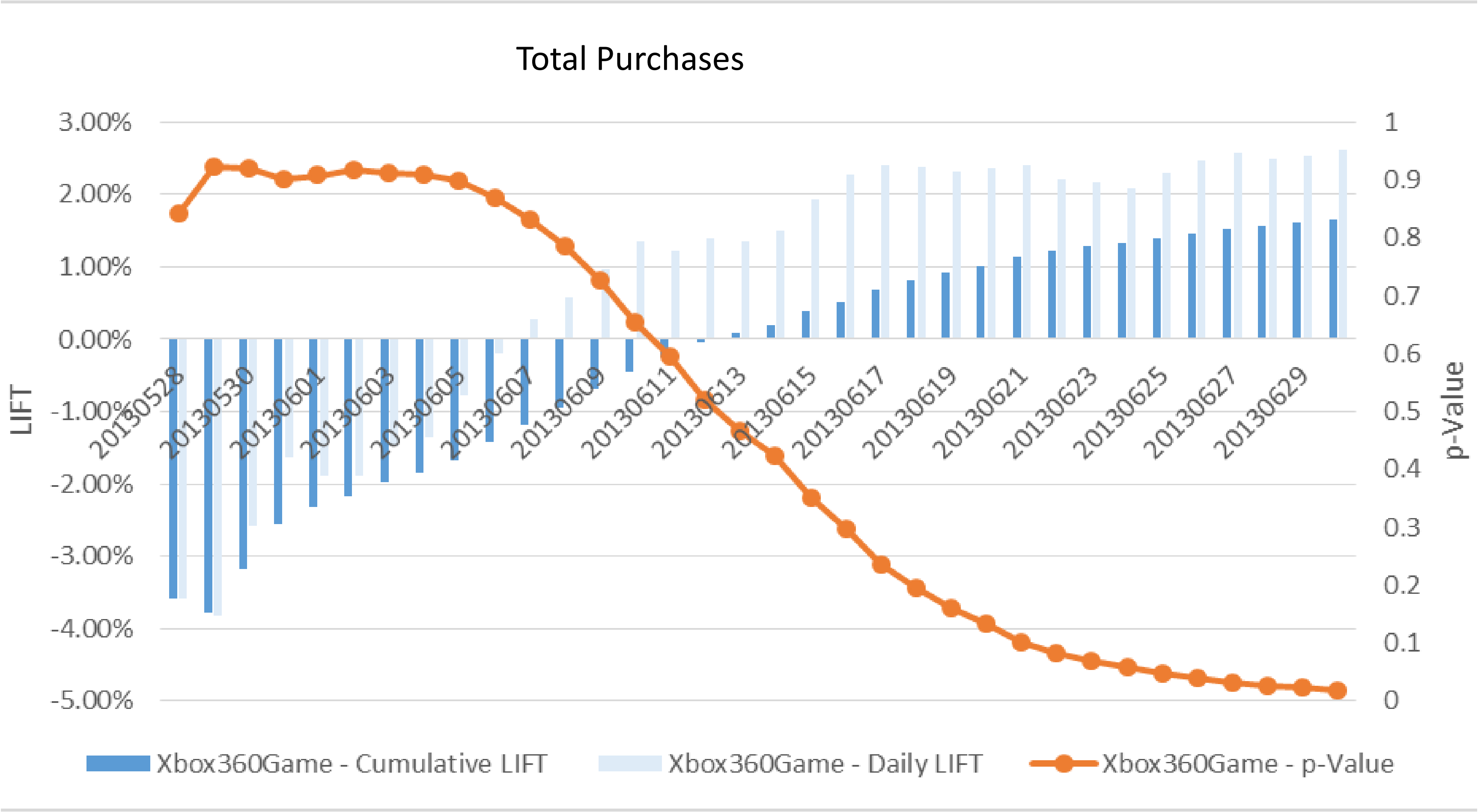
Online Experiments in Xbox



Game Purchase



Total Game Purchase



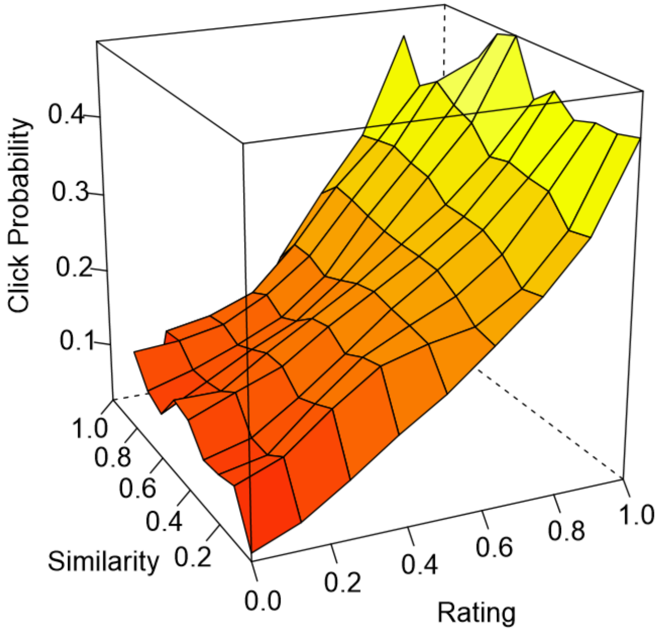
Experimentation Caveats

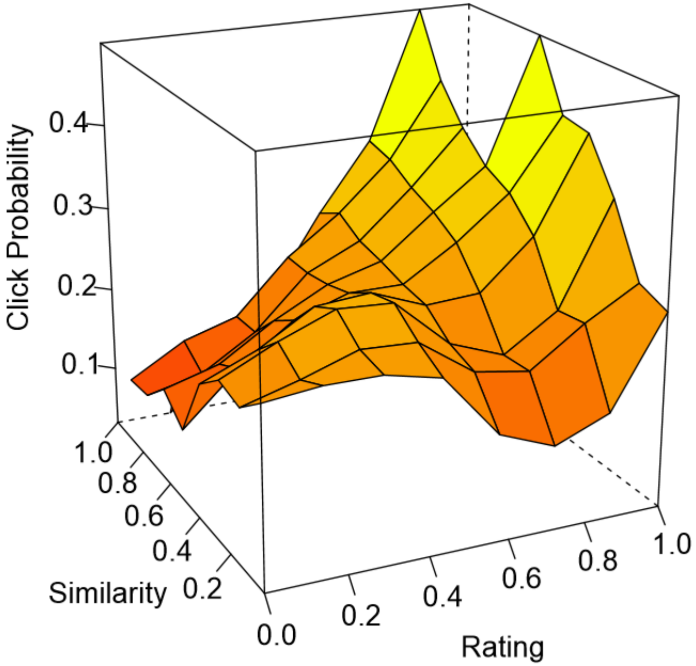
- What KPIs to measure?
- How long to run the experiment?
- External factors may influence the results
- Cannibalization is hard to account for

- Expensive to implement
- Can’t compare algorithms before “lighting up”

The Gap

Accuracy and Diversity Interactions





Characterizing The Offline / Online Evaluation Gap

- Overemphasis of popular items
- List recommendations (diversity, item position)
- Freshness/ Fatigue
- Contextual information is not fully utilized

- Learning from historical data lets you predict the future. But what we really care about is changing the future!

Bridging The Gap

Mitigating Evaluation Techniques

- Domain experts / focus groups
- Internal user studies
- Off-policy evaluation techniques

Off Policy Evaluation - Example

$V_h^\pi(S)$- The expected reward of a policy h given data S from a “logging policy” π.

$$V_h^\pi(S) = \frac{1}{|S|} \sum_{(x,a,r)\in S} \frac{r \cdot \mathbb{I}[h(x) == a]}{\max(\hat{\pi}(x|a), \tau)}$$

where S denotes the set of context-action-reward tuples available in the logs

Caveats of Off-policy Evaluation

- Need to formulate everything in terms of a policy
- Needs sufficient support
- Becomes very difficult when your policies are time dependent

Thank you!

We are looking for postdoc researchers to join us in Israel…

Email: RecoRecruitmentEmail@microsoft.com

Microsoft