Letting Users Choose Recommender Algorithms

Michael Ekstrand
(*Texas State University)*

Daniel Kluver, Max Harper, and Joe Konstan
(*GroupLens Research / University of Minnesota)*

pick a recommender

○ **"the peasant"**
non-personalized

○ "the bard"
based on movie group point allocation (configure)

○ **"the warrior"**
based on ratings

◉ **"the wizard"**
based on ratings

Research Objective

If we give users control over the algorithm providing their recommendations, what happens?

Why User Control?

- Different users, different needs/wants
  - Allow users to personalize the recommendation experience to their needs and preferences.

- Transparency and control may promote trust

Research Questions

- Do users make use of a switching feature?

- How much do they use it?

- What algorithms do they settle on?

- Do algorithm or user properties predict choice?

Relation to Previous Work

Paper you just saw: tweak algorithm output

We change the whole algorithm

Previous study (RecSys 2014): what do users perceive to be different, and say they want?

We see what their actions say they want

Outline

1. Introduction *(just did that)*
2. Experimental Setup
3. Findings
4. Conclusion & Future Work

Context: MovieLens

- Let MovieLens users switch between algorithms
- Algorithm produces:
  - Recommendations (in sort-by-recommended mode)
  - Predictions (everywhere)
- Change is persistent until next tweak
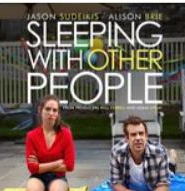- Switcher integrated into top menu

top picks   see more

MovieLens recommends these movies

The Lives of Others	Inside Job	The Imitation Game	Temple Grandin	Incendies	Star Wars: Episode V	Citizenfour	From the Earth to th
2006 R 137 min 📌	2010 PG-13 109 min 📌	2014 PG-13 113 min 📌	2010 108 min 📌	2010 R 130 min 📌	2015 124 min 📌	2014 R 114 min 📌	1998 720 min 📌

☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆

recent releases   see more

movies released in last 90 days

Sleeping with Other	Goodnight Mommy	The Visit	Legend	Listening	12 Rounds 3: Lockd	Colonia	Welcome to Leith
2015 101 min 📌	2015 100 min 📌	2015 PG-13 94 min 📌	2015 131 min 📌	2014 100 min 📌	2015 R 90 min 📌	2015 120 min 📌	2015 85 min 📌

☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆   ☰ ★★★★☆

298 ☆ ▾

top picks see more

MovieLens recommends these movies

The Lives of Others	Inside Job	The Imitation Game	Temple Grandin	Incendies	Star Wars: Ep
2006 R 137 min	2010 PG-13 109 min	2014 PG-13 113 min	2010 108 min	2010 R 130 min	2015 124 min

RATINGS AND RECOMMENDATIONS

You have rated 298 movies (click here for stats!). By rating more movies you improve your profile and recommendations.

You are using the **wizard** recommender. This recommender uses your ratings to determine which movies to recommend. It works by turning all users' ratings data into a small set of factors that capture the essential preference aspects of a movie or a user (it uses Simon Funk's implementation of the singluar value decomposition algorithm, for the technically minded and curious).

The MovieLens recommenders are powered by LensKit.

CHANGE YOUR RECOMMENDER

○ "THE PEASANT"
non-personalized

○ "THE BARD"
based on movie group point allocation (configure)

○ "THE WARRIOR"
based on ratings

● "THE WIZARD"
based on ratings

recent releases see more

movies released in last 90 days

Sleeping with Other	Goodnight Mommy	The Visit	Legend	Listening	12 Rounds 3: Lockd	Colonia	Welcome to Leith
2015 101 min	2015 100 min	2015 PG-13 94 min	2015 131 min	2014 100 min	2015 R 90 min	2015 120 min	2015 85 min

Algorithms

- Four algorithms
    - **Peasant:** personalized (user-item) mean rating
    - **Bard:** group-based recommender (Chang et al. CSCW 2015)
    - **Warrior:** item-item CF
    - **Wizard:** FunkSVD CF
- Each modified with 10% blend of popularity rank for top-N recommendation

pick a recommender

○ **"the peasant"**
non-personalized

○ "the bard"
based on movie group point allocation (configure)

○ **"the warrior"**
based on ratings

◉ **"the wizard"**
based on ratings

Experiment Design

- Only consider established users
- Each user randomly assigned an initial algorithm (not the Bard)
- Allow users to change algorithms
  - Interstitial highlighted feature on first login
- Log interactions

Users Switch Algorithms

- 3005 total users

- 25% (748) switched at least once

- 72.1% of switchers (539) settled on different algorithm
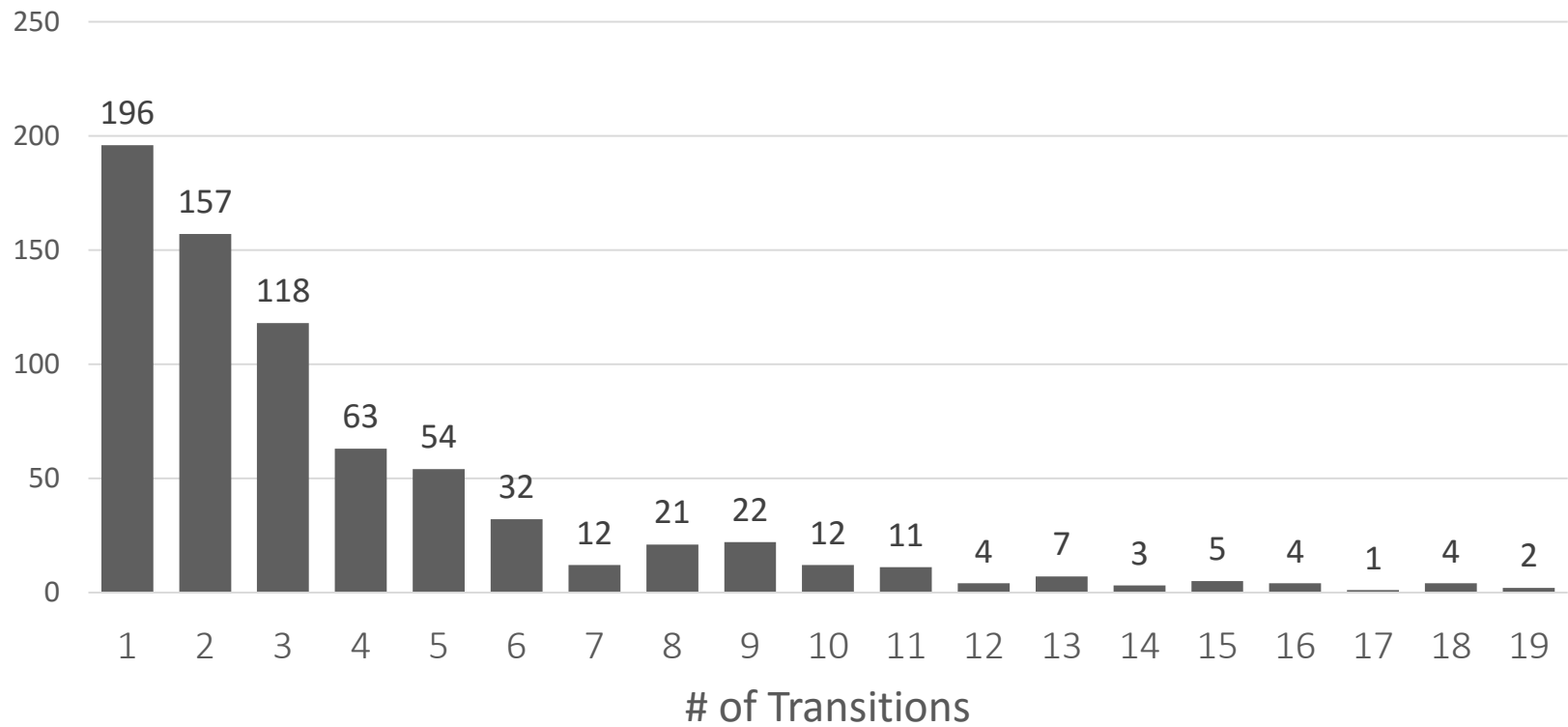
Finding 1: Users do use the control

Ok, so how do they switch?

- Many times or just a few?

- Repeatedly throughout their use, or find an algorithm and stick with it?

Switching Behavior: Few Times

Transition Count Histogram

Switching Beh.: Few Sessions

- Break *sessions* at 60 mins of inactivity
- 63% only switched in 1 session, 81% in 2 sessions
- 44% only switched in 1st session
- Few intervening events (switches concentrated)

Finding 2: users use the menu some, then leave it alone

I'll just stay here…

Question: do users find some algorithms more *initially satisfactory* than others?

Frac. of Users Switching
(all diffs. significant, χ^2 p<0.05)
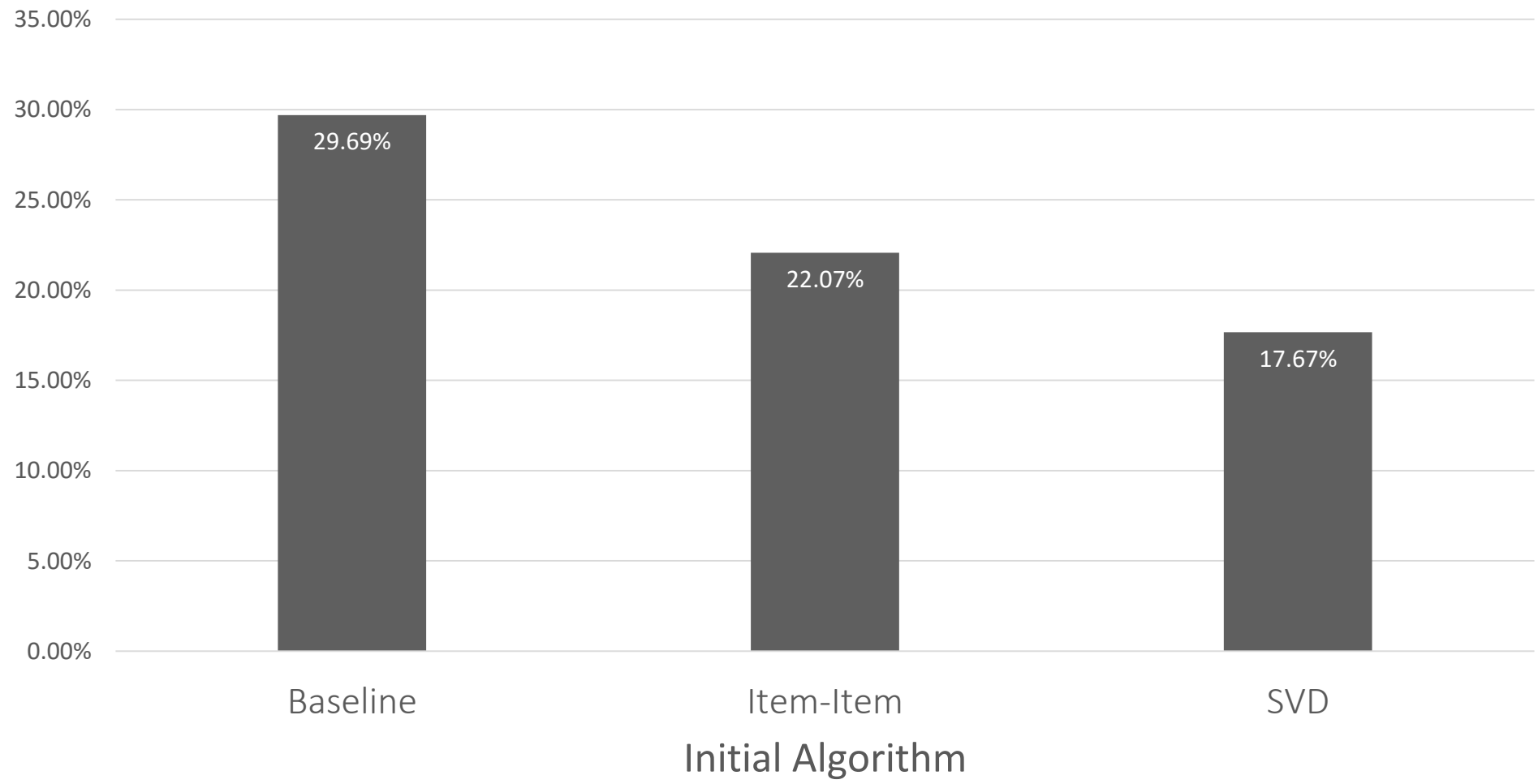
...or go over there...

Question: do users tend to find some algorithms more *finally satisfactory* than others?

…by some path

What do users do between initial and final?

- As stated, not many flips
- Most common: change to other personalized, maybe change back (A -> B, A -> B -> A)
- Users starting w/ baseline usually tried one or both personalized algorithms

Final Choice of Algorithm
(for users who tried menu)
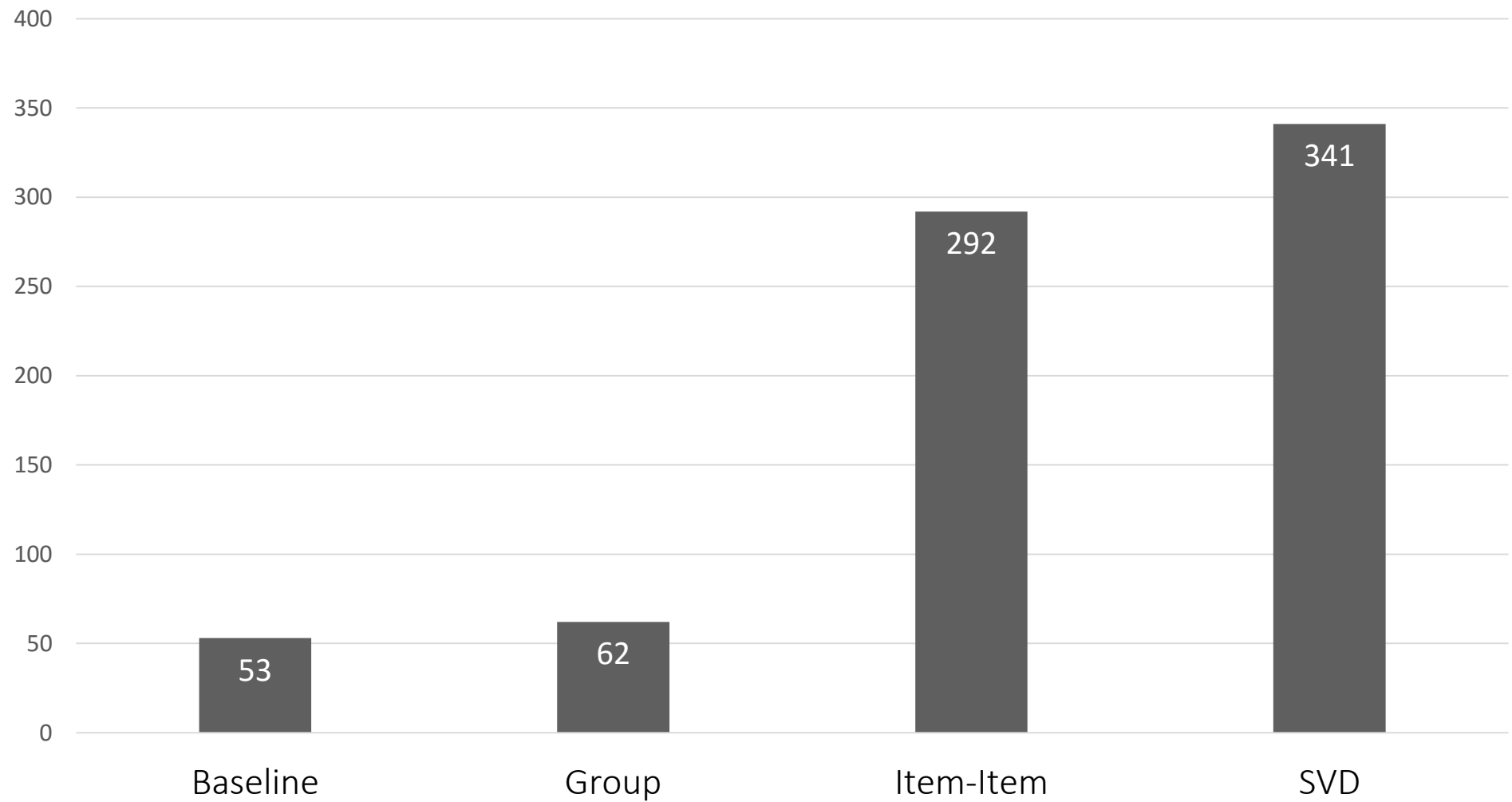
Algorithm Preferences

- Users prefer personalized (more likely to stay initially or finally)

- Small preference of SVD over item-item

- Caveat: algorithm naming may confound

Interlude: Offline Experiment

- For each user:
  - Discarded all ratings after starting experiment
  - Use 5 most recent pre-experiment ratings for testing
- Train recommenders
- Measure:
  - RMSE for test ratings
  - Boolean recall: is a rated move in first 24 recs?
  - Diversity (intra-list similarity over tag genome)
  - Mean pop. rank of 24-item list
- Why 24? Size of single page of MovieLens results
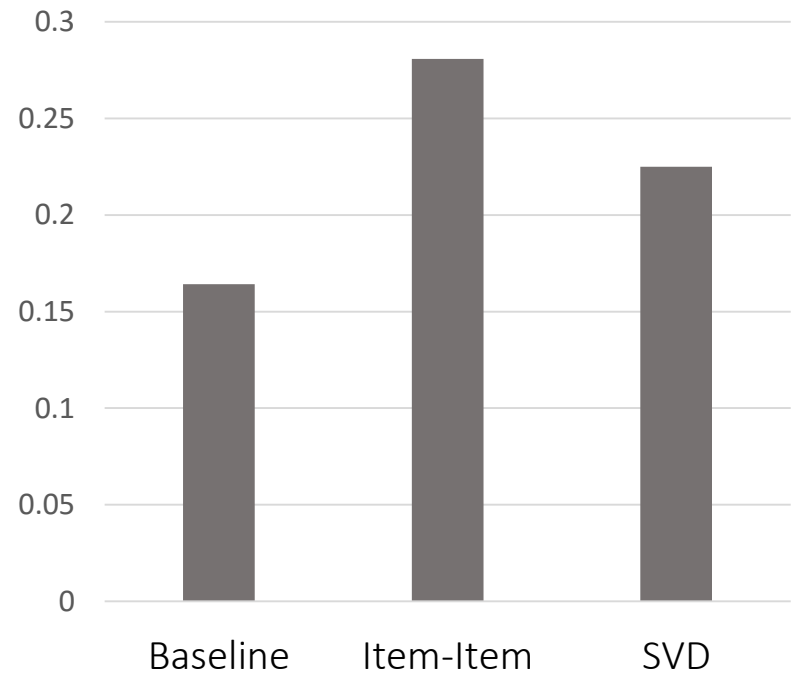
Algorithms Made Different Recs

- Average of 53.8 unique items/user (out of 72 possible)

- Baseline and Item-Item most different (Jaccard similarity)

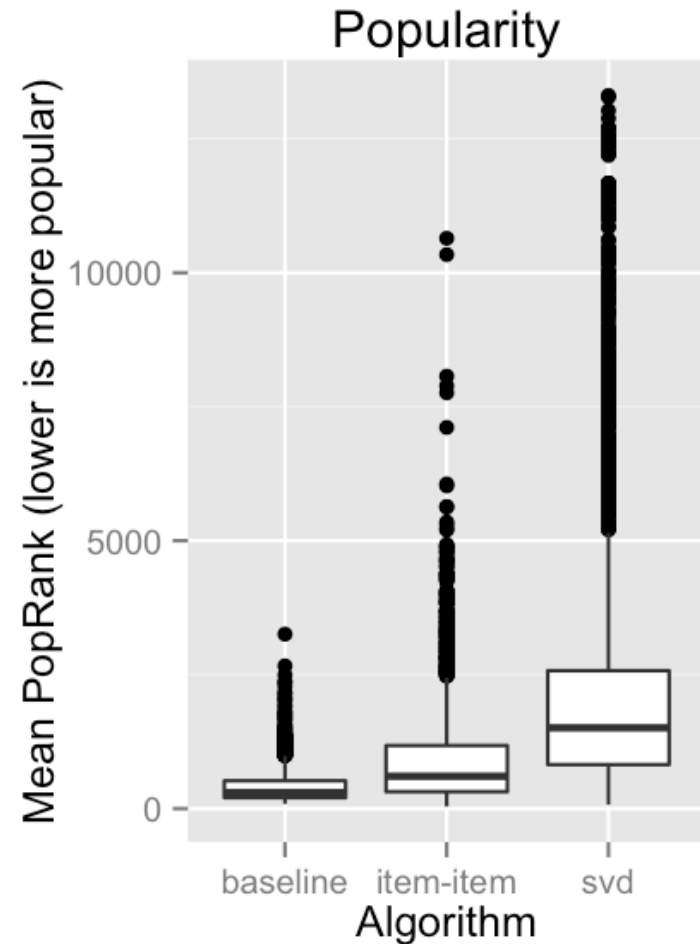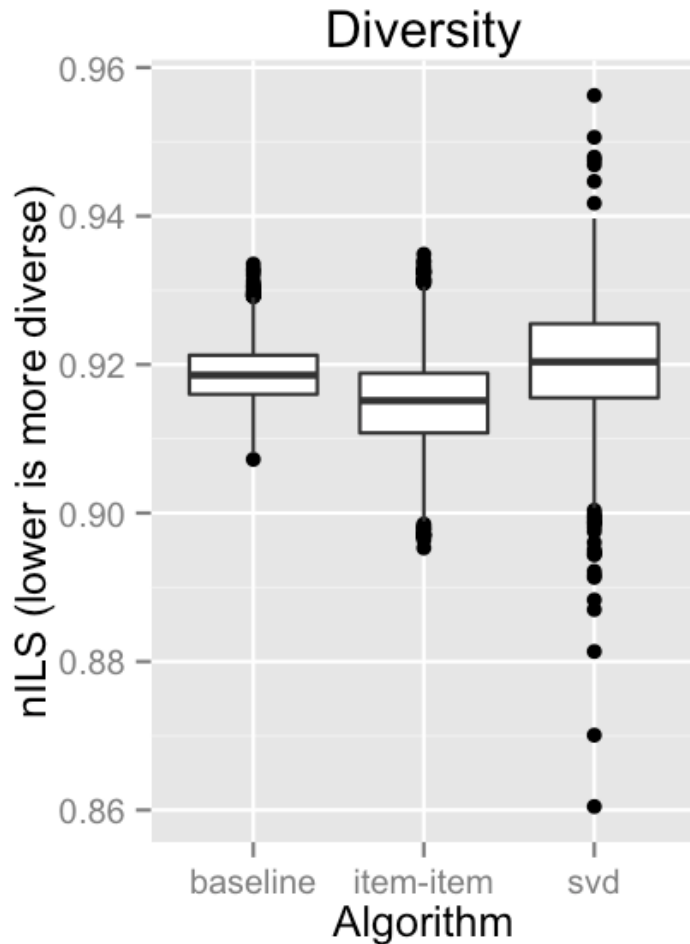- Accuracy is another story…

Algorithm Accuracy

RMSE
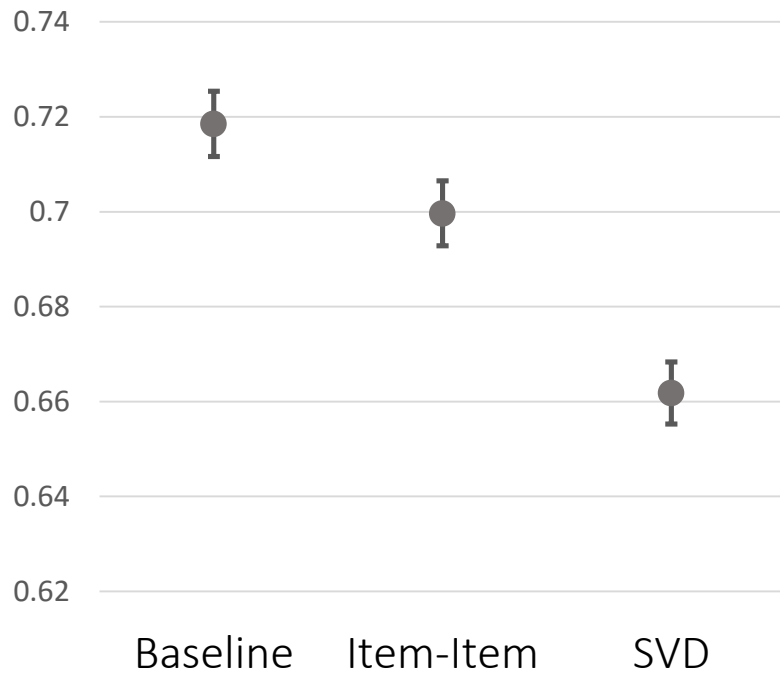
Boolean Recall

Diversity and Popularity

Not Predicting User Preference

- Algorithm properties do directly not predict user preference, or whether they will switch

- Little ability to predict user behavior overall
  - If user starts with baseline, diverse baseline recs increase likelihood of trying another algorithm
  - If user starts w/ item-item, novel baseline recs increase likelihood of trying
  - No other significant effects found

- Basic user properties do not predict behavior

What does this mean?

- Users take advantage of the feature

- Users experiment a little bit, then leave it alone

- Observed preference for personalized recs, especially SVD

- Impact on long-term user satisfaction unknown

Future Work

- Disentangle preference and naming

- More domains

- Understand impact on long-term user satisfaction and retention

Questions?