Recommending for People

MICHAEL EKSTRAND

NOVEMBER 16, 2015

#1TweetResearch

How can we make the real world of intelligent information systems good for its inhabitants?

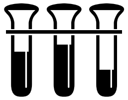
The Real World of Technology

Ursula Franklin’s 1989 Massey Lectures

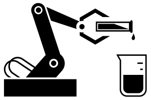
Technology is not just artifacts. Rather:

- It is process
- It affects people
- It is a product of volition, was designed, could be designed other ways

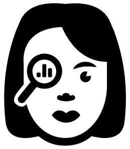
Must understand people and social structures surrounding our technology.



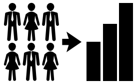
- Tools and Instrumentation



- Offline Recommender Errors



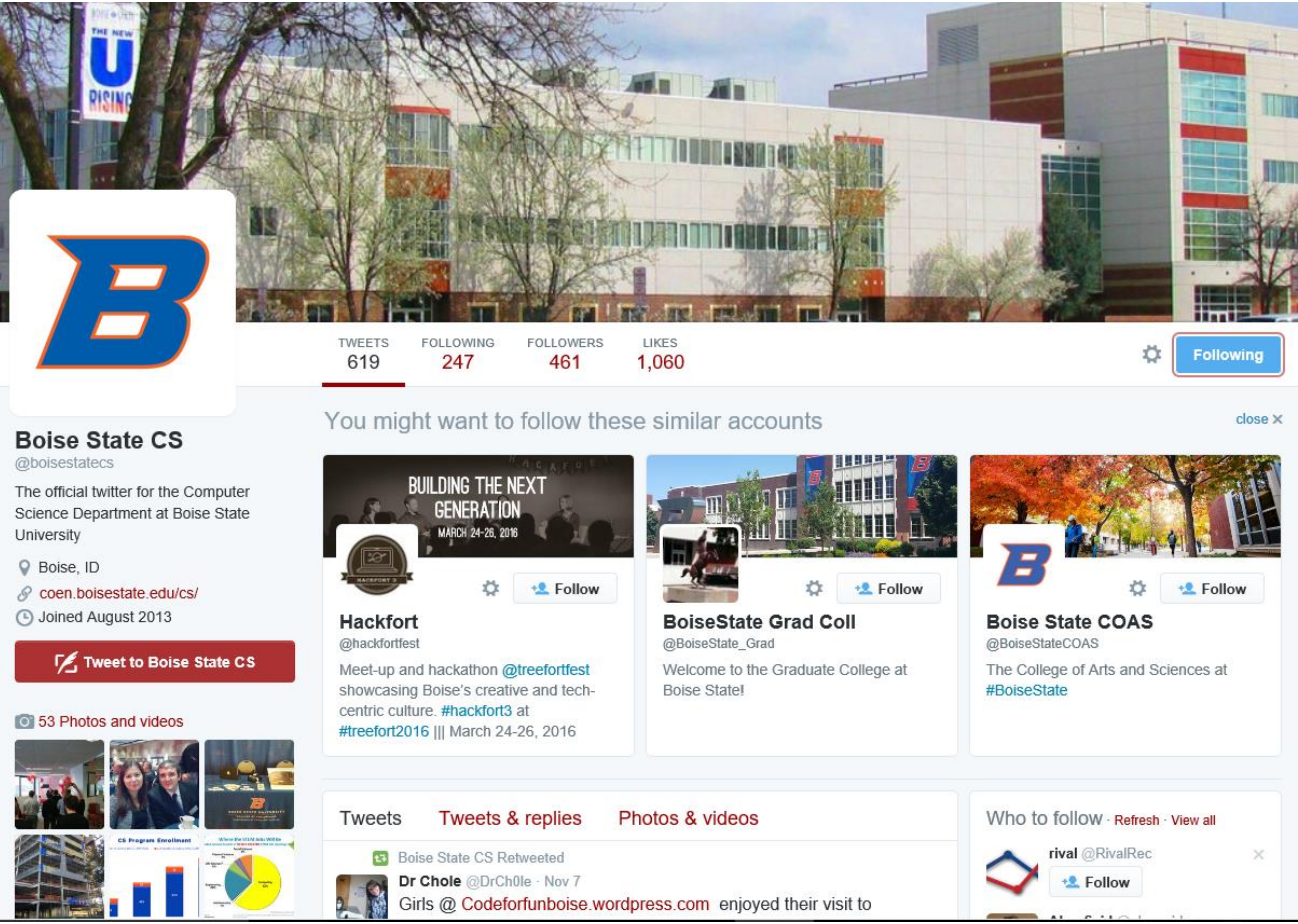
- User Perception of Recommendations

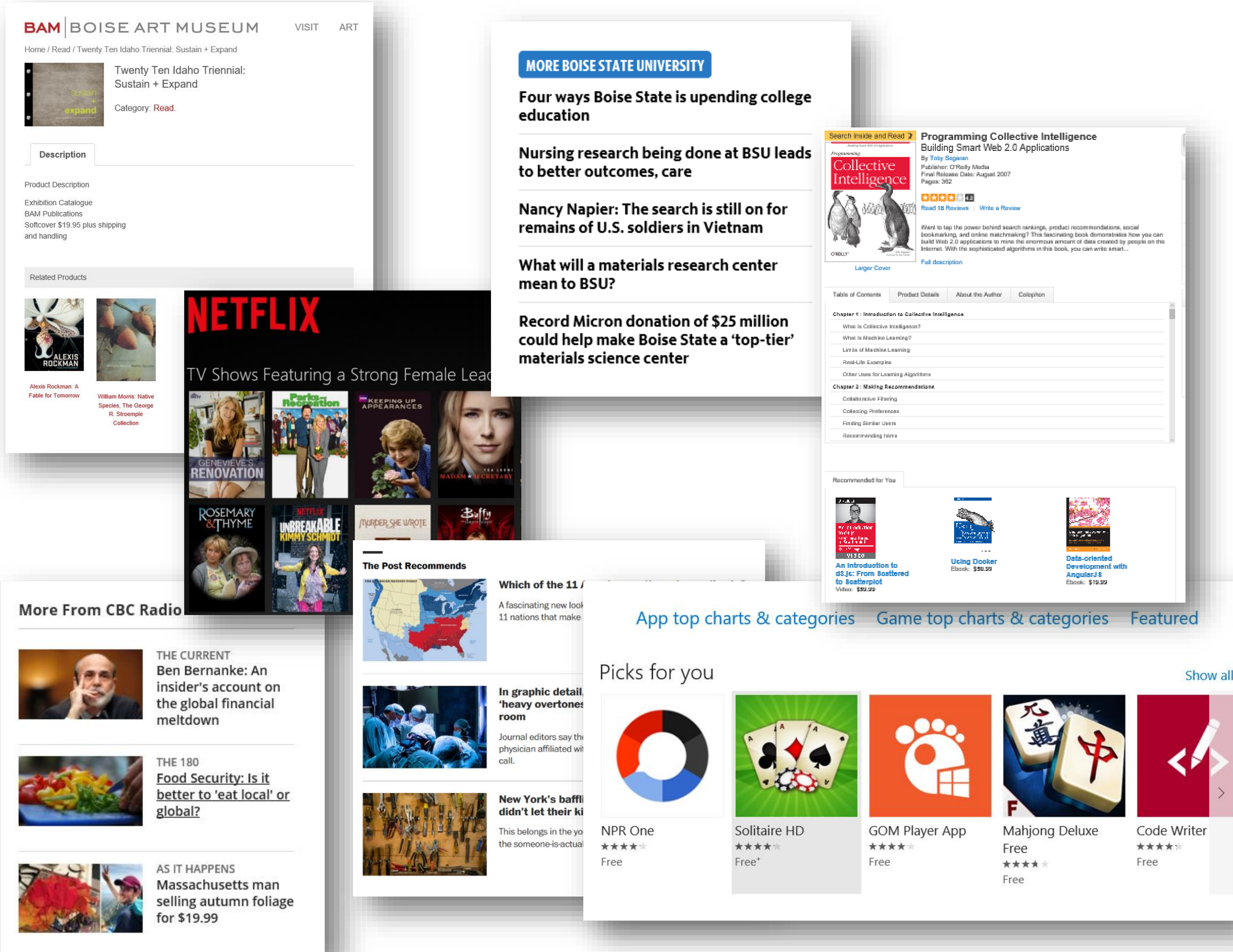


- User Behavior in Recommender Choice

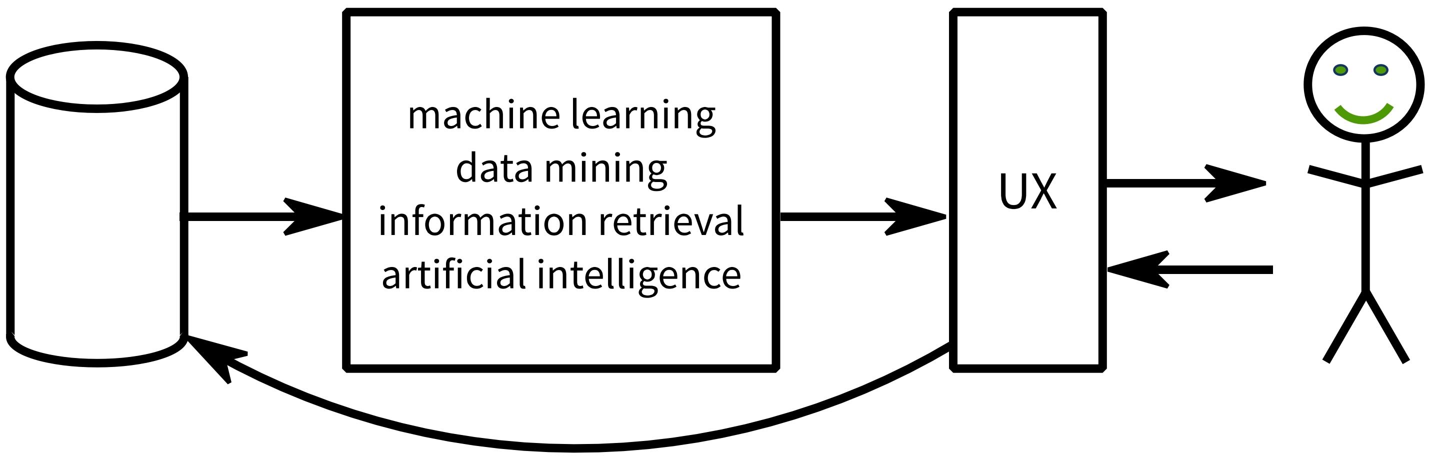
- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Agenda and Future Work

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Agenda and Future Work


THE NEW U RISING
TWEETS 619
FOLLOWING 247
FOLLOWERS 461
LIKES 1,060
Following
You might want to follow these similar accounts
close ×
Boise State CS
@boisestatecs
The official twitter for the Computer Science Department at Boise State University
Boise, ID
coen.boisestate.edu/cs/
Joined August 2013
Tweet to Boise State CS
53 Photos and videos
BUILDING THE NEXT GENERATION
MARCH 24-26, 2016
Follow
Hackfort
@hackfortfest
Meet-up and hackathon @treefortfest showcasing Boise's creative and tech-centric culture. #hackfort3 at #treefort2016 ||| March 24-26, 2016
Follow
BoiseState Grad Coll
@BoiseState_Grad
Welcome to the Graduate College at Boise State!
Follow
Boise State COAS
@BoiseStateCOAS
The College of Arts and Sciences at #BoiseState
Tweets
Tweets & replies
Photos & videos
Boise State CS Retweeted
Dr Chole @DrCh0le · Nov 7
Girls @ Codeforfunboise.wordpress.com enjoyed their visit to
Who to follow · Refresh · View all
rival @RivalRec
Follow


BAM | BOISE ART MUSEUM
VISIT
ART
Home / Read / Twenty Ten Idaho Triennial: Sustain + Expand
Twenty Ten Idaho Triennial: Sustain + Expand
Category: Read.
Description
Product Description
Exhibition Catalogue
BAM Publications
Softcover $19.95 plus shipping and handling
Related Products
Alexis Rockman: A Fable for Tomorrow
William Morris: Native Species, The George R. Stroemple Collection
NETFLIX
TV Shows Featuring a Strong Female Lead
MORE BOISE STATE UNIVERSITY
Four ways Boise State is upending college education
Nursing research being done at BSU leads to better outcomes, care
Nancy Napier: The search is still on for remains of U.S. soldiers in Vietnam
What will a materials research center mean to BSU?
Record Micron donation of $25 million could help make Boise State a 'top-tier' materials science center
Programming Collective Intelligence
Building Smart Web 2.0 Applications
Table of Contents
Product Details
About the Author
Colophon
Recommended for You
More From CBC Radio
THE CURRENT
Ben Bernanke: An insider's account on the global financial meltdown
THE 180
Food Security: Is it better to 'eat local' or global?
AS IT HAPPENS
Massachusetts man selling autumn foliage for $19.99
The Post Recommends
App top charts & categories
Game top charts & categories
Featured
Picks for you
Show all
NPR One
Free
Solitaire HD
Free*
GOM Player App
Free
Mahjong Deluxe Free
Free
Code Writer
Free

Recommender Architecture



Common Approaches

- Non-personalized
- Content-based [Balabanović, 1997; others]
- Collaborative filtering
  - User-based [Resnick et al., 1994]
  - Item-based [Sarwar et al., 2001]
  - Matrix factorization [Sarwar et al., 2000; Funk, 2006]
- Hybrid approaches [Burke, 2002]
- Learning to Rank

Evaluating Recommenders

Many measurements:

- ML/IR-style experiments with data sets
  - Measure error of predicting user ratings (RMSE, MAE)
  - Measure accuracy of retrieving user’s rated/liked/purchased items (P/R, MAP, MRR, NDCG)
- User studies and surveys
- A/B testing in the field
  - Engagement metrics
  - Business metrics

Research Goals

Premise: Algorithms perform differently

No reason to think one size fits all! [McNee et al., 2006]

Questions: How do they differ…

… in objectively measurable output?

… in subjective perception of output?

… in user preference (observed and articulated)?

… in impact on users and community?

Objective: So we can build a better world of technology

Background

Tools and Instrumentation

Offline Recommender Errors

User Perception of Recommendations

User Behavior in Recommender Choice

Agenda and Future Work

LensKit

An open-source toolkit for **building**, **researching**, and **learning about** recommender systems.

LensKit

Ekstrand et al., 2011

build

- prototype and study recommender applications
- deploy research results in live systems

research

- reproduce and validate results
- new experiments with old algorithms
- research algorithms with users
- make research easier
- provide good baselines

learn

- open-source code
- study production-grade implementations

LensKit in Use

- Engine behind user-facing recommenders
  - MovieLens, ~3K users/month
  - BookLens, built into Twin Cities public libraries
  - Confer system for CHI/CSCW
- Supports education
  - Coursera MOOC (~1000 students)
  - Recommender classes @ UMN, TX State
- Used in research (> 20 papers)

Algorithm Architecture

Principle

Build algorithms from reusable, reconfigurable components.

Benefits

- Reproduce many configurations
- Try new ideas by replacing one piece
- Reuse pieces in new algorithms

Enabled by *Grapht*, our Java dependency injector.

Evaluator

- Cross-validate rating data sets
- Train and measure recommenders
- Many metrics
  - Predict: RMSE, MAE, nDCG (rank-accuracy)
  - Top-N: nDCG, P/R@N, MRR
  - Easy to write new metrics
- Optimized: reuses common algorithm components

Research Outcomes

- Public, open-source software, v. 3.0 coming soon
- Direct publications
  - Software presented in RecSys 2011 paper and demo
  - Paper on Grapht under review for *J. Object Technology*
- Supported additional research on recommender interfaces (Kluver et al., 2012; Nguyen et al., 2013)
- Used by various systems and researchers

Ongoing Work

- Finishing LensKit 3.0 with simplified tooling, better integration
- Re-launching programming portion of MOOC
- Improving efficiency of algorithms, evaluator
- Several student projects
  - Efficient strategies for tuning hyperparameters
  - Understanding and improving performance over time
  - Documenting current best practices and making them accessible defaults

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Agenda and Future Work

When Recommenders Fail

Ekstrand and Riedl, RecSys 2012

When do algorithms make mistakes?

Do different algorithms make different mistakes?

Do different algorithms perform better for different users?

Data and Setting

- MovieLens (http://movielens.org)
  - Movie recommendation service & community
  - 2500-3000 unique users/month
  - Extensive tagging features
- Snapshots of rating database publicly available
  - ML-10M: 10M 5-star ratings of 10K movies by 70K users
  - Also: ML-100K, ML-1M, ML-20M

Algorithms Considered

- User-based collaborative filtering (User-User)
- Item-based collaborative filtering (Item-Item)
- Matrix factorization (FunkSVD)
- Tag-based recommendations (Lucene)
- Personalized user-item mean baseline (Mean)

Outcomes

Counting *mispredictions* ($|p - r| > 0.5$) gives different picture than prediction error.

Consider per-user fraction correct and RMSE:

- Correlation is 0.41
- Agreement on best algorithm: 32.1%
- Rank-consistent for overall performance

Marginal Correct Predictions

Q1: Which algorithm has the most successes ($\epsilon \leq 0.5$)?

Qn+1: Which has the most successes where 1…n failed?

Algorithm	# Good	%Good	Cum. % Good
ItemItem	859,600	53.0	53.0
UserUser	131,356	8.1	61.1
Lucene	69,375	4.3	65.4
FunkSVD	44,960	2.8	68.2
Mean	16,470	1.0	69.2
Unexplained	498,850	30.8	100.0

Lessons Learned

- Algorithms make different mistakes
- Looking at ‘was wrong?’ can yield different insight then aggregating error
- Different users have different best algorithms
- Room to pick up additional signal

Background

Tools and Instrumentation

Offline Recommender Errors

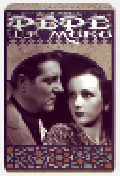
User Perception of Recommendations

User Behavior in Recommender Choice

Agenda and Future Work

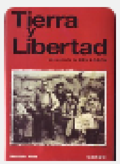
movielens

List A (10 movies)



Pépé le Moko
1937 94 min
Action, Crime

The Mummy's Curse
1944 62 min
Horror



Land and Freedom
1994 109 min
Drama, History

Children of Paradise
1945 190 min
Drama, Romance

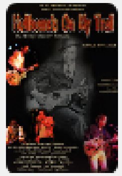
What Time Is It There?
2000 116 min
Drama, Romance

List B (10 movies)

Fear City: A Family-Sty
1994 93 min
Comedy

Connections (1978)
1977

Ween: Live in Chicago
2004 120 min



Hellhounds on My Trail

Heimat: A Chronicle of
1984 925 min

scroll down for more

Survey (25 questions)

Lists A and B contain the top movie recommendations for you from different "recommenders". Please answer the following questions to help us understand your preferences about these recommenders.

1. Based on your first impression, which list do you prefer?

Much more A than B — About the same — Much more B than A

2. Which list has more movies that you find appealing?

Much more A than B — About the same — Much more B than A

3. Which list has more movies that might be among the best movies you see in the next year?

Much more A than B — About the same — Much more B than A

4. Which list has more obviously bad movie recommendations for you?

Much more A than B — About the same — Much more B than A

scroll down for more (why so many questions?)

Research Questions
Ekstrand et al., RecSys 2014

RQ1

How do subjective properties affect choice of recommendations?

RQ2

What differences do users perceive between lists of recommendations produced by different algorithms?

RQ3

How do objective metrics relate to subjective perceptions?

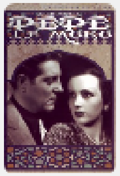
With GroupLens, Martijn Willemsen

Experiment Design

- Each user was assigned 2 algorithms
  - User-User
  - Item-Item
  - FunkSVD
- Users answered comparative survey
  - Initial ‘which do you like better?’
  - 22 questions
    - ‘Which list has more movies that you find appealing?’
    - ‘much more A than B’ to ‘much more B than A’
  - Forced choice selection for future use

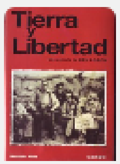
movielens

List A (10 movies)



Pépé le Moko
1937 94 min
Action, Crime

The Mummy's Curse
1944 62 min
Horror



Land and Freedom
1994 109 min
Drama, History

Children of Paradise
1945 190 min
Drama, Romance

What Time Is It There?
2000 116 min
Drama, Romance

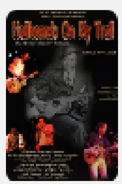
scroll down for more

List B (10 movies)

Fear City: A Family-Sty
1994 93 min
Comedy

Connections (1978)
1977

Ween: Live in Chicago
2004 120 min



Hellhounds on My Trail

Heimat: A Chronicle of
1984 925 min

Survey (25 questions)

Lists A and B contain the top movie recommendations for you from different "recommenders". Please answer the following questions to help us understand your preferences about these recommenders.

1. Based on your first impression, which list do you prefer?

Much more A than B		About the same		Much more B than A

2. Which list has more movies that you find appealing?

Much more A than B		About the same		Much more B than A

3. Which list has more movies that might be among the best movies you see in the next year?

Much more A than B		About the same		Much more B than A

4. Which list has more obviously bad movie recommendations for you?

Much more A than B		About the same		Much more B than A

scroll down for more (why so many questions?)

Experiment Features

Joint evaluation: users compare 2 lists

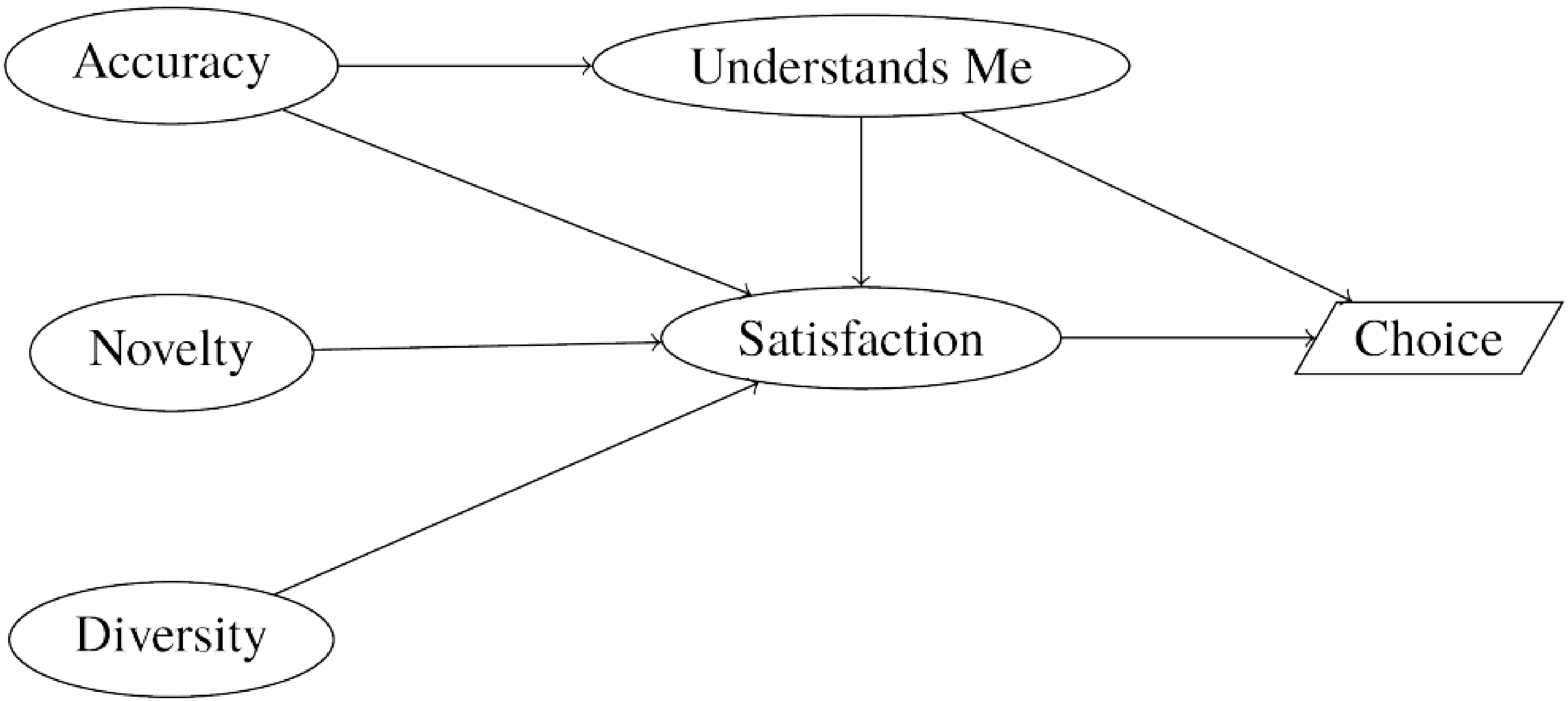
- enables more subtle distinctions than separate eval
- harder to interpret

Factor analysis: 22 questions measure 5 factors

- more robust than single questions
- **structural equation model** tests relationships

New problem: SEM on joint evaluation

Hypothesized Model



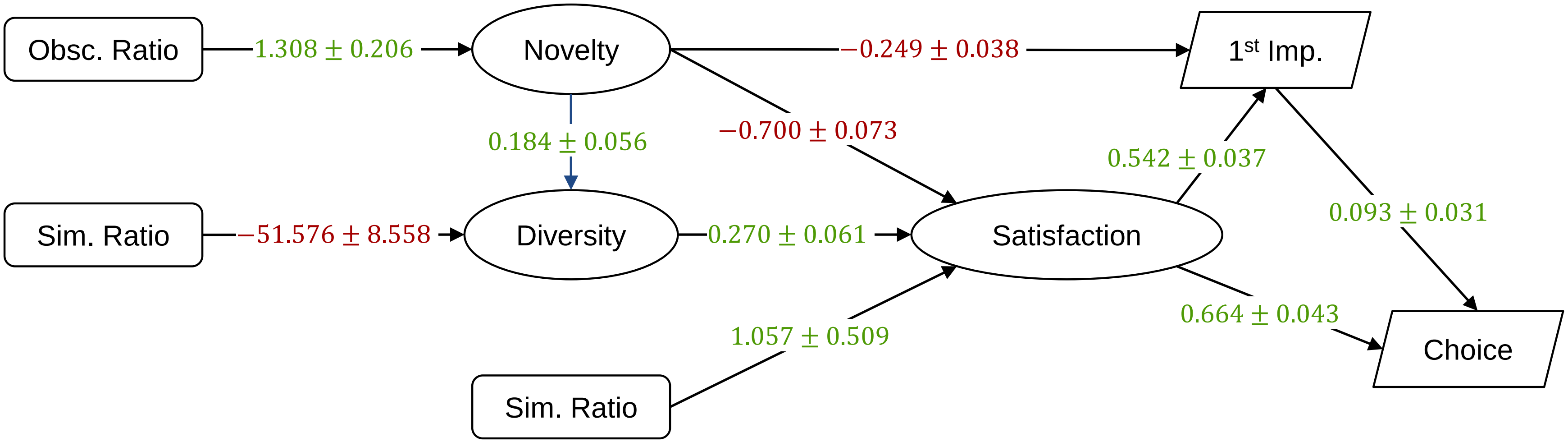
Response Summary

582 users completed

Condition (A v. B)	*N*	Pick *A*	Pick *B*	% Pick *B*
I-I v. U-U	201	144	57	**28.4%**
I-I v. SVD	198	101	97	49.0%
SVD v. U-U	183	136	47	**25.7%**

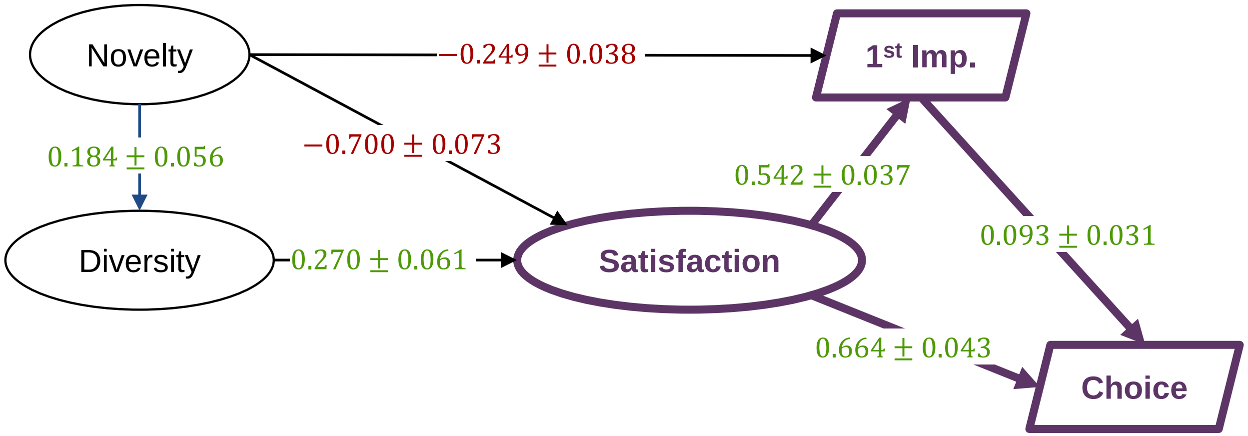
bold is significant ($p < 0.001$, $H_0\colon {}^{b}/_{n} = 0.5$)

Measurement Model



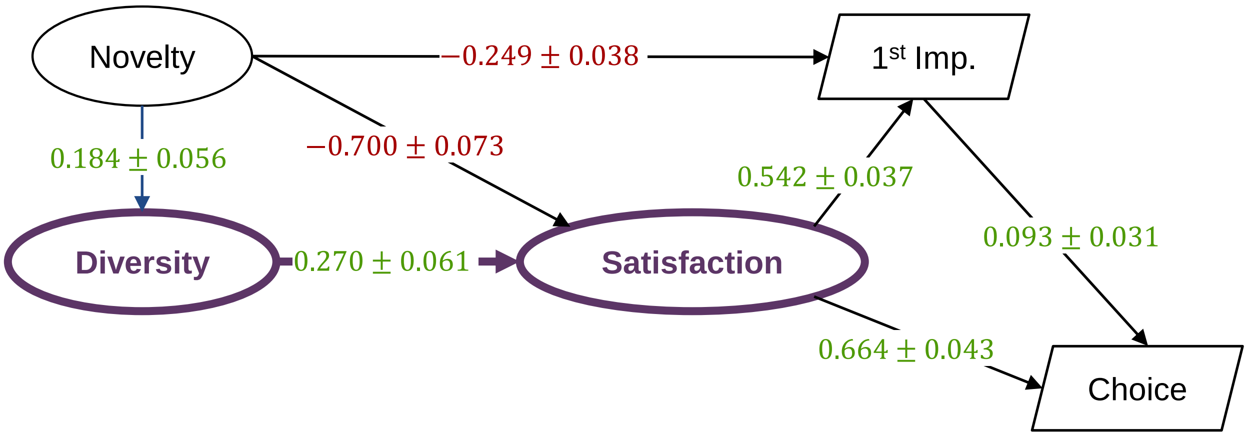
- Multi-level linear regression
- Direction comes from theory
- All measurements relative: positive is ‘more B than A’
- Accuracy, Understands Me folded into Satisfaction

Choice: Satisfaction



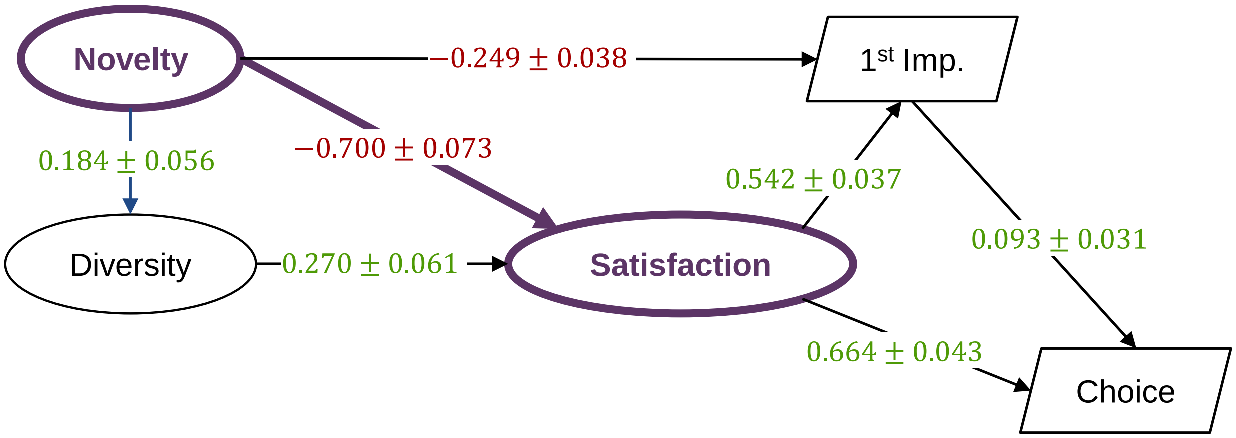
Satisfaction positively affects impression and choice

Choice: Diversity



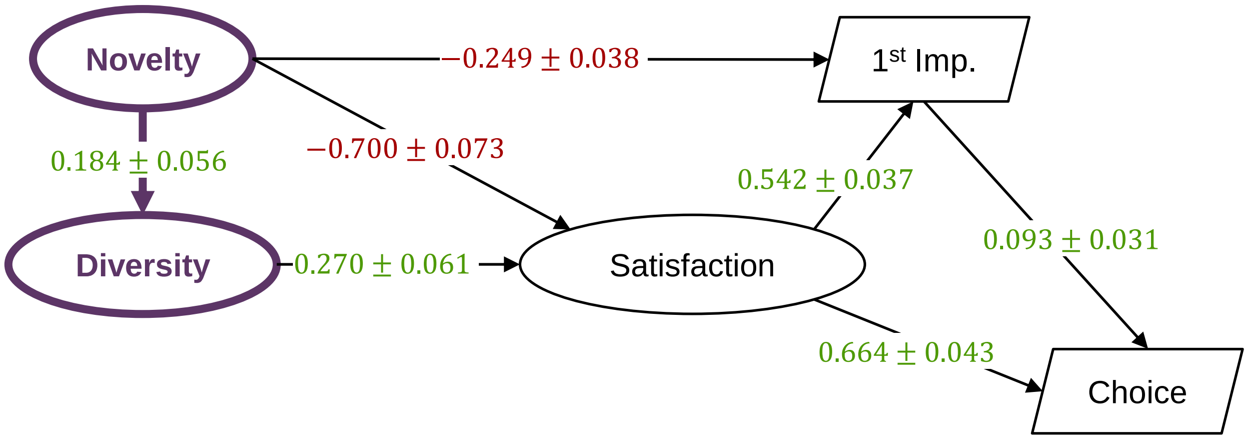
Diversity positively affects satisfaction and choice

Choice: Novelty



Novelty hurts satisfaction and choice

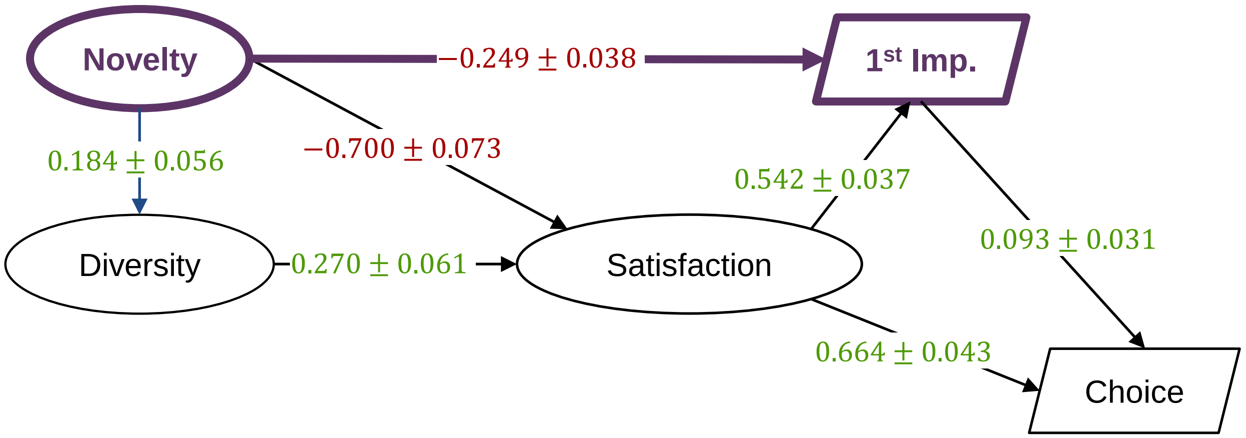
Novelty and Diversity



Novelty improves diversity

Impact on satisfaction outweighed by direct negative effect

Novelty and Impression



Novelty has direct negative impact on 1st impression

Implications

Context: choosing an algorithm to provide recs

- Novelty boosts diversity, but hurts algorithm impression
- Negative impact of novelty diminishes with close scrutiny
  - Can recommender get less conservative as users gain experience?
- Diversity has positive impact on user satisfaction
- Diversity does not trade off with *perceived* accuracy

RQ2: Algorithm Differences

- Pairwise comparisons are difficult to interpret
- Method: re-interpret as 3 between-subjects pseudo-experiments:

Baseline	Tested	% Tested > Baseline
Item-Item	SVD	48.99
	User-User	28.36
SVD	Item-Item	51.01
	User-User	25.68
User-User	Item-Item	71.64
	SVD	74.32

RQ2 Summary

- User-user more novel than either SVD or item-item
- User-user more diverse than SVD
- User-user's excessive novelty decreases for experienced (many ratings) users
- Users choose SVD and item-item in roughly equal measure
- Results consistent with raw responses

RQ3: Objective Properties

Measure objective features of lists:

Novelty

obscurity (popularity rank)

Diversity

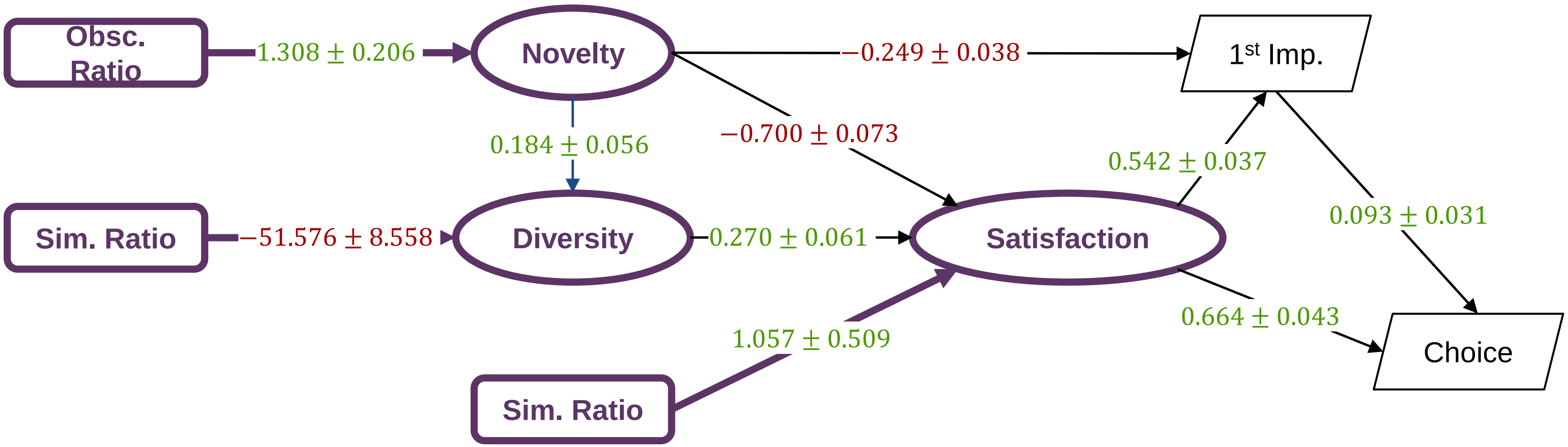
intra-list similarity

Sim. metric: cosine over tag genome (Vig)

Accuracy/Sat

RMSE over last 5 ratings

Model with Metrics



- Each metric correlates with its subjective factor
- Metric impact entirely mediated by subjective factors
- Algorithm condition still significant – metrics don't capture all

Summary

- Novelty has complex, largely negative effect
  - Exact use case likely matters
  - Complements McNee's notion of *trust-building*
- Diversity is important, mildly influenced by novelty.
  - Tag genome measures perceptible diversity best, but advantage is small.
- User-user loses (likely due to obscurity), but users are split on item-item vs. SVD
- Consistent responses, reanalysis, and objective metrics

Refining Expectations

- Commonly-held offline beliefs:
  - Novelty is good
  - Diversity and accuracy trade off
- Perceptual results (here and elsewhere):
  - Novelty is complex – be careful
  - Diversity and accuracy both achievable

More research needed, of course

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Agenda and Future Work

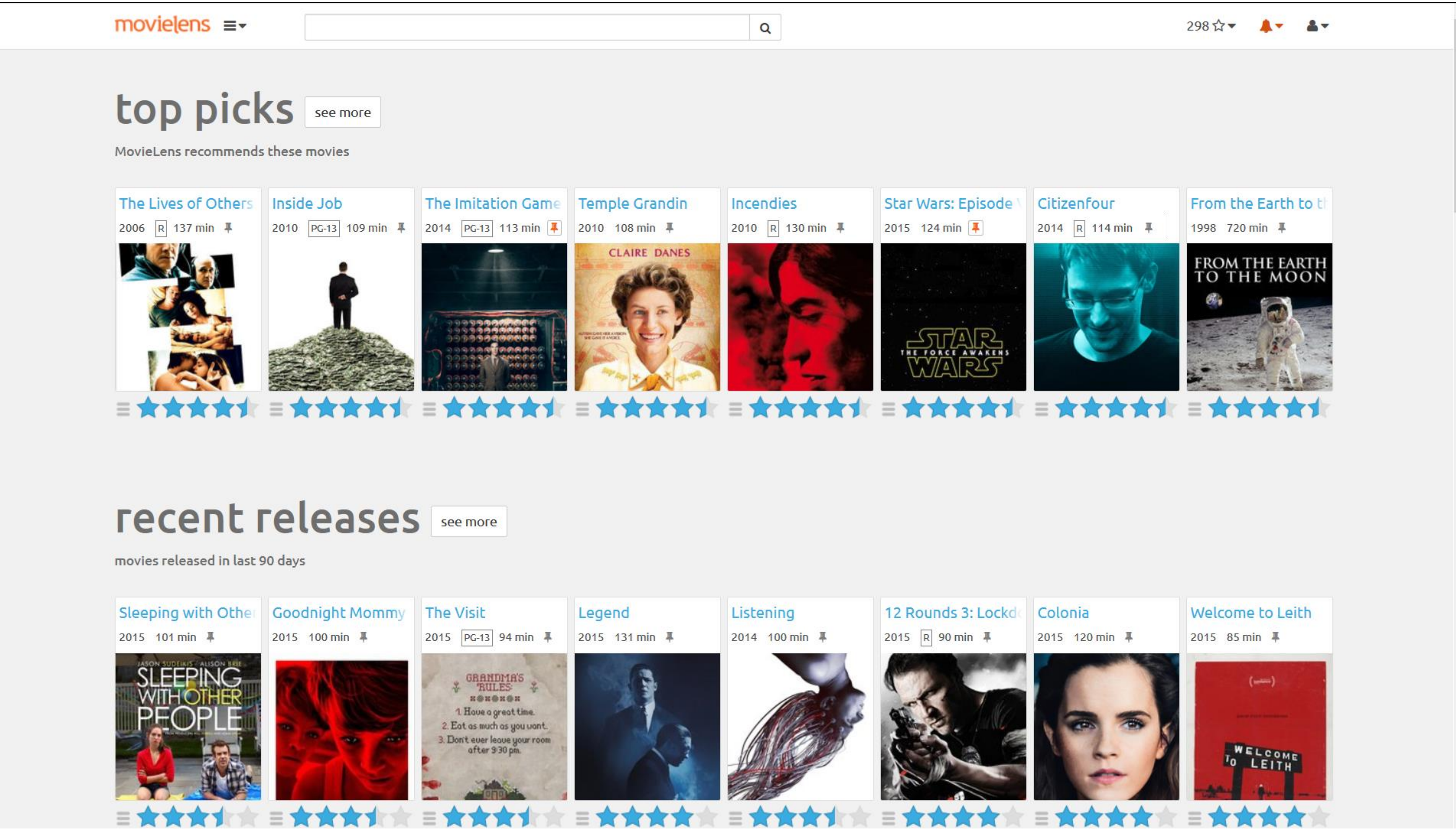
Giving Users Control

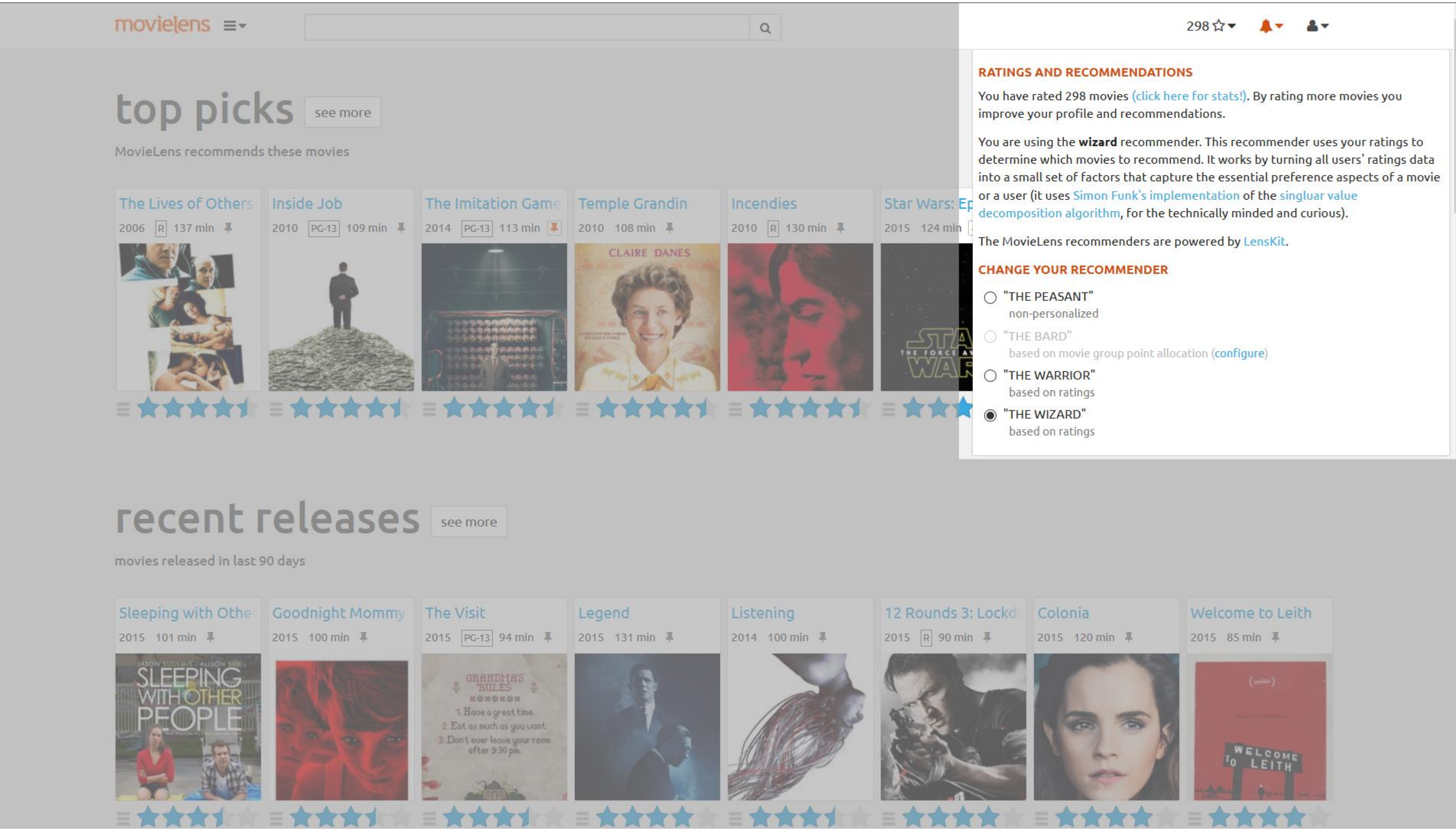
Ekstrand et al., RecSys 2015

- We have:
  - Analyzed performance on offline data
  - Asked users what they want
- What happens when we just let them pick in actual use?

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?


movielens
298
top picks
see more
MovieLens recommends these movies
The Lives of Others
2006 R 137 min
Inside Job
2010 PG-13 109 min
The Imitation Game
2014 PG-13 113 min
Temple Grandin
2010 108 min
CLAIRE DANES
Incendies
2010 R 130 min
Star Wars: Episode
2015 124 min
STAR WARS
THE FORCE AWAKENS
Citizenfour
2014 R 114 min
From the Earth to t
1998 720 min
FROM THE EARTH TO THE MOON
recent releases
see more
movies released in last 90 days
Sleeping with Othe
2015 101 min
JASON SUDEIKIS ALISON BRIE
SLEEPING WITH OTHER PEOPLE
Goodnight Mommy
2015 100 min
The Visit
2015 PG-13 94 min
GRANDMA'S RULES:
1. Have a great time.
2. Eat as much as you want.
3. Don't ever leave your room after 9:30 pm.
Legend
2015 131 min
Listening
2014 100 min
12 Rounds 3: Lockd
2015 R 90 min
Colonia
2015 120 min
Welcome to Leith
2015 85 min
WELCOME TO LEITH



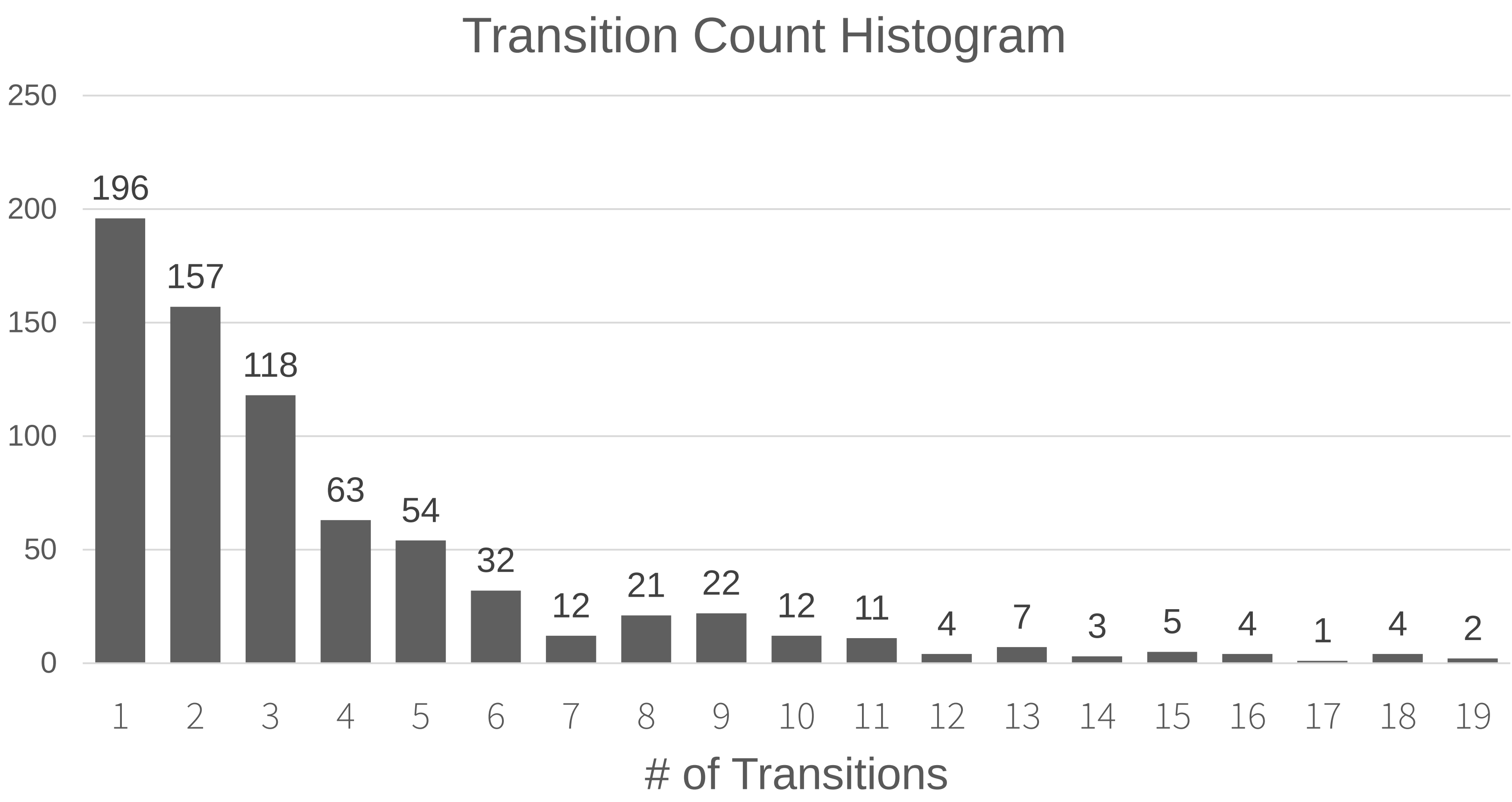
movielens
298
top picks
see more
MovieLens recommends these movies
The Lives of Others
2006 R 137 min
Inside Job
2010 PG-13 109 min
The Imitation Game
2014 PG-13 113 min
Temple Grandin
2010 108 min
CLAIRE DANES
Incendies
2010 R 130 min
Star Wars: Ep
2015 124 min
RATINGS AND RECOMMENDATIONS
You have rated 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 Othe
2015 101 min
SLEEPING WITH OTHER PEOPLE
Goodnight Mommy
2015 100 min
The Visit
2015 PG-13 94 min
GRANDMA'S RULES
1. Have a great time.
2. Eat as much as you want.
3. Don't ever leave your room after 9:30 pm.
Legend
2015 131 min
Listening
2014 100 min
12 Rounds 3: Lockd
2015 R 90 min
Colonia
2015 120 min
Welcome to Leith
2015 85 min
WELCOME TO LEITH

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

Switching Behavior: Few Times



Switching Behavior: 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

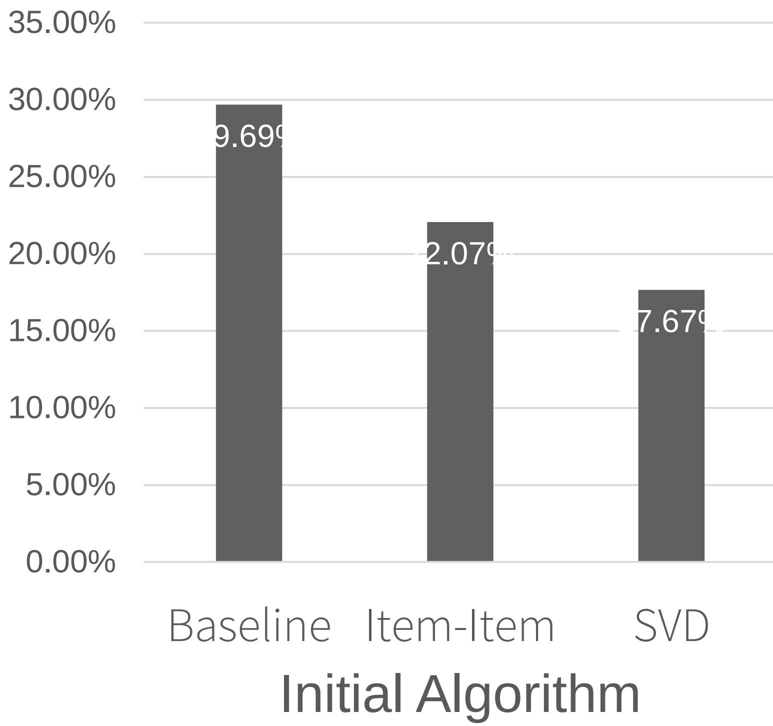
Algorithm Preferences

Q1: do users find some algorithms more *initially satisfactory* than others?

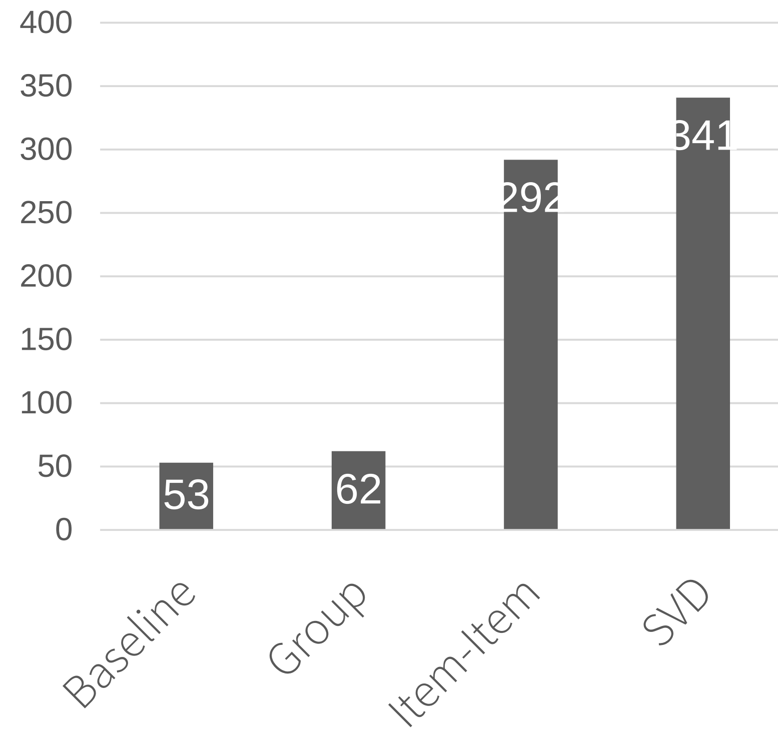
Q2: do users tend to find some algorithms more *finally satisfactory* than others?

Algorithm Preference

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



Final Choice of Algorithm
(for users who tried menu)



Down the garden path…

What do users do between initial and final states?

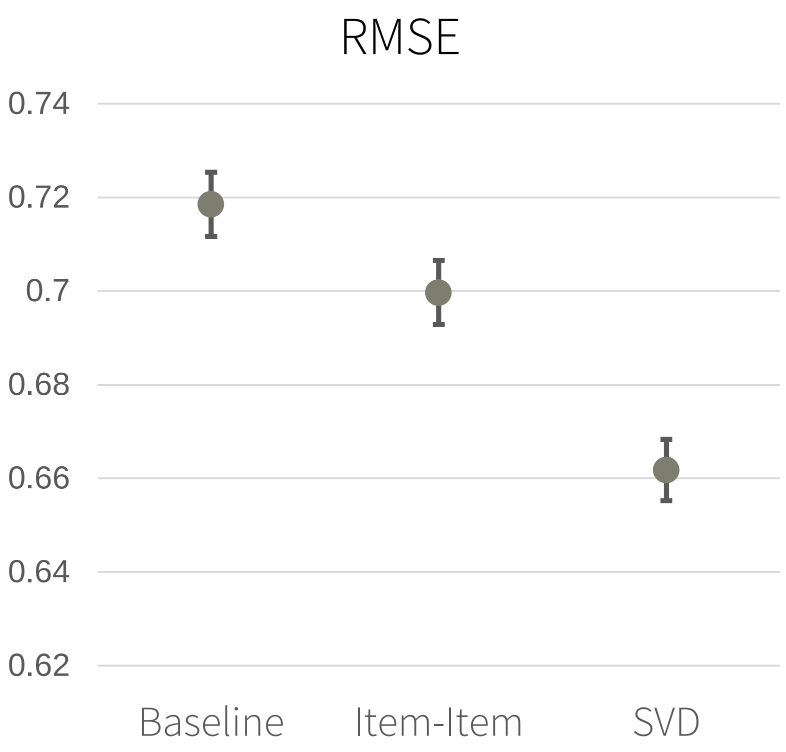
- As stated, not many flips
- Most common: change to other personalized, maybe change back (A -> B, A -> B -> A)
- Users starting w/ baseline usually tried one or both personalized algorithms

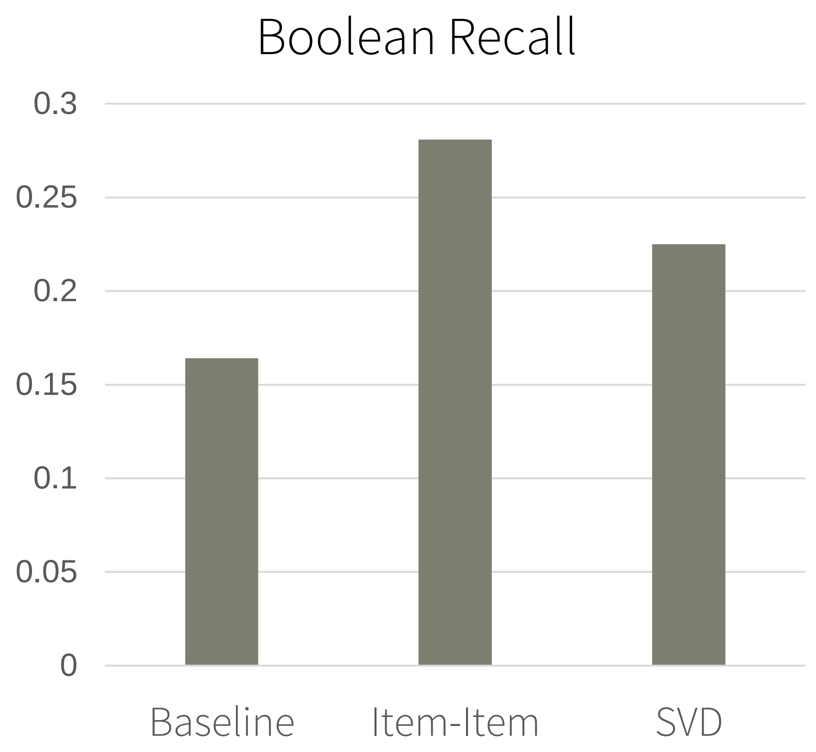
Algorithms Made Different Recs

Analyzed recommender behavior for users offline.

- 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





Measured over attempts to predict or recommend last 5 items user rated before entering experiment.

Not Predicting User Preference

- Algorithm properties do directly not predict user preference, or whether they will switch
- Little ability to predict user behavior overall
- 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

Ongoing Work

3 studies, similar questions, similar outcomes

- Item-item and SVD very similar
- Different recommenders better in different cases

Goal:

- Integrate findings
- Analyze behavior data from survey users
- Analyze user properties more deeply

Background

Tools and Instrumentation

Offline Recommender Errors

User Perception of Recommendations

User Behavior in Recommender Choice

Agenda and Future Work

Core Ideas

How can we make the real world of intelligent information systems good for its inhabitants?

Have seen:

- User-centric offline evaluation
- User surveys
- User behavior studies

So far, individual users in static scenarios.

Interactive Recommendation

Goal: recommender-user collaboration for building collections (bibliographies, film lists, etc.)

Idea:

- Recommenders provide suggestions, critique other recommendations
- User decides what to add
- Recommenders and meta-recommender learn and improve

Broadening the Lens

- How do recommenders affect their users *as a group*?
- How do recommenders affect their users *with relation to other users*?
- How do recommenders interact with their broader sociotechnical context?
  - Biased input data
  - Assumptions made in algorithm design
  - Legal and ethical implications of outputs

Agenda Summary

- Ongoing work
  - LensKit development, continuing to promote reproducible research
  - User-centric examination of recommendation techniques, mapping user and task suitability
  - Collaboration with psychology
- New directions
  - Interactive recommendation to support novel tasks
  - Studying social impact of recommenders

Thank you

Also thanks to:

- *Collaborators (GroupLens, Martijn Willemsen)*
- *NSF for funding Ph.D studies*
- *Texas State for supporting current work*