Recommending for People

PIReT

BOISE STATE UNIVERSITY

MICHAEL D. EKSTRAND

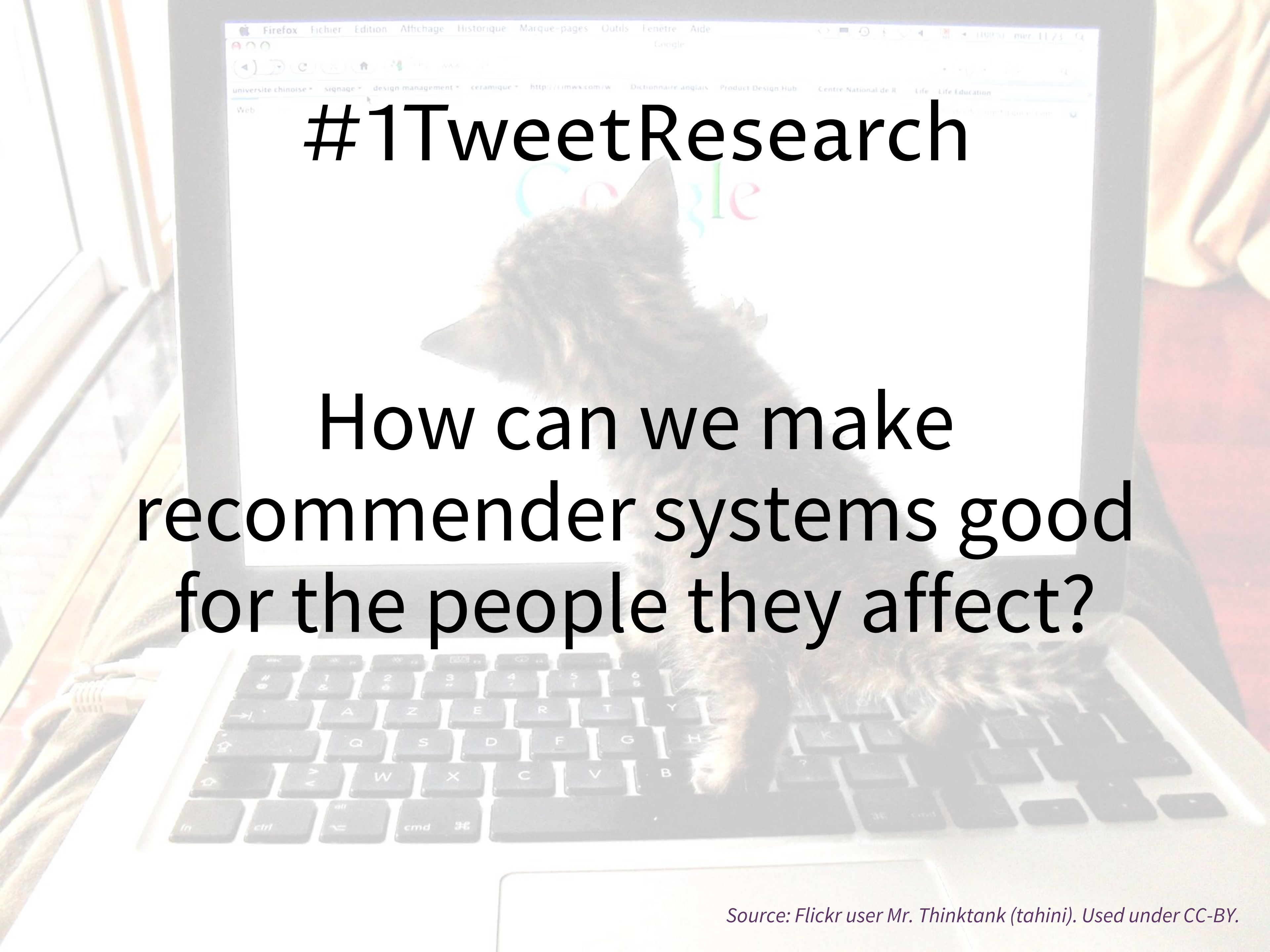
People and Information Research Team

Dept. of Computer Science, Boise State University

bit.ly/RecPeopleAN16

The PIReTs

People and Information Research Team



#1TweetResearch

How can we make recommender systems good for the people they affect?

Source: Flickr user Mr. Thinktank (tahini). Used under CC-BY.

Background

Tools and Instrumentation

Offline Recommender Errors

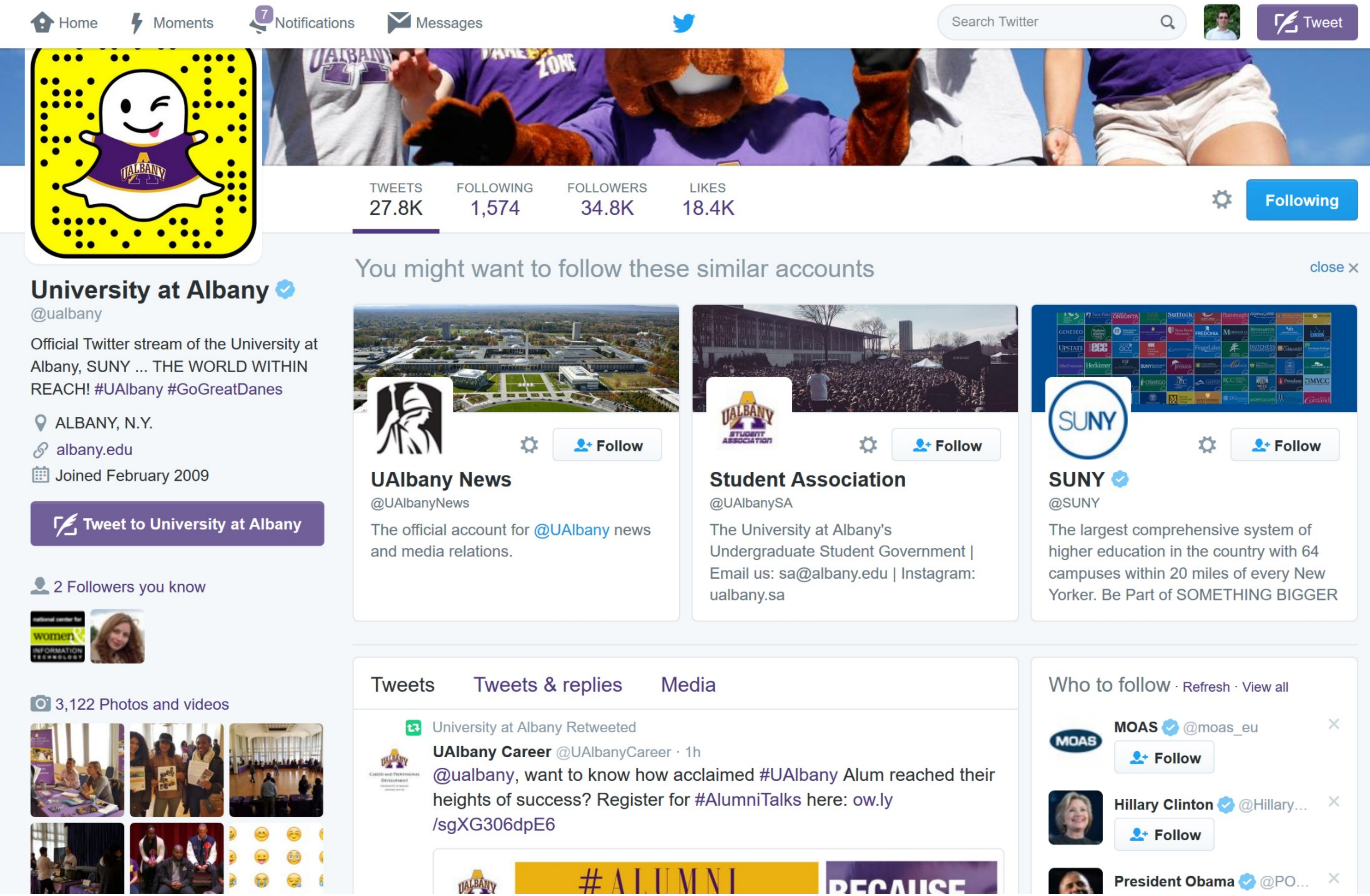
User Perception of Recommendations

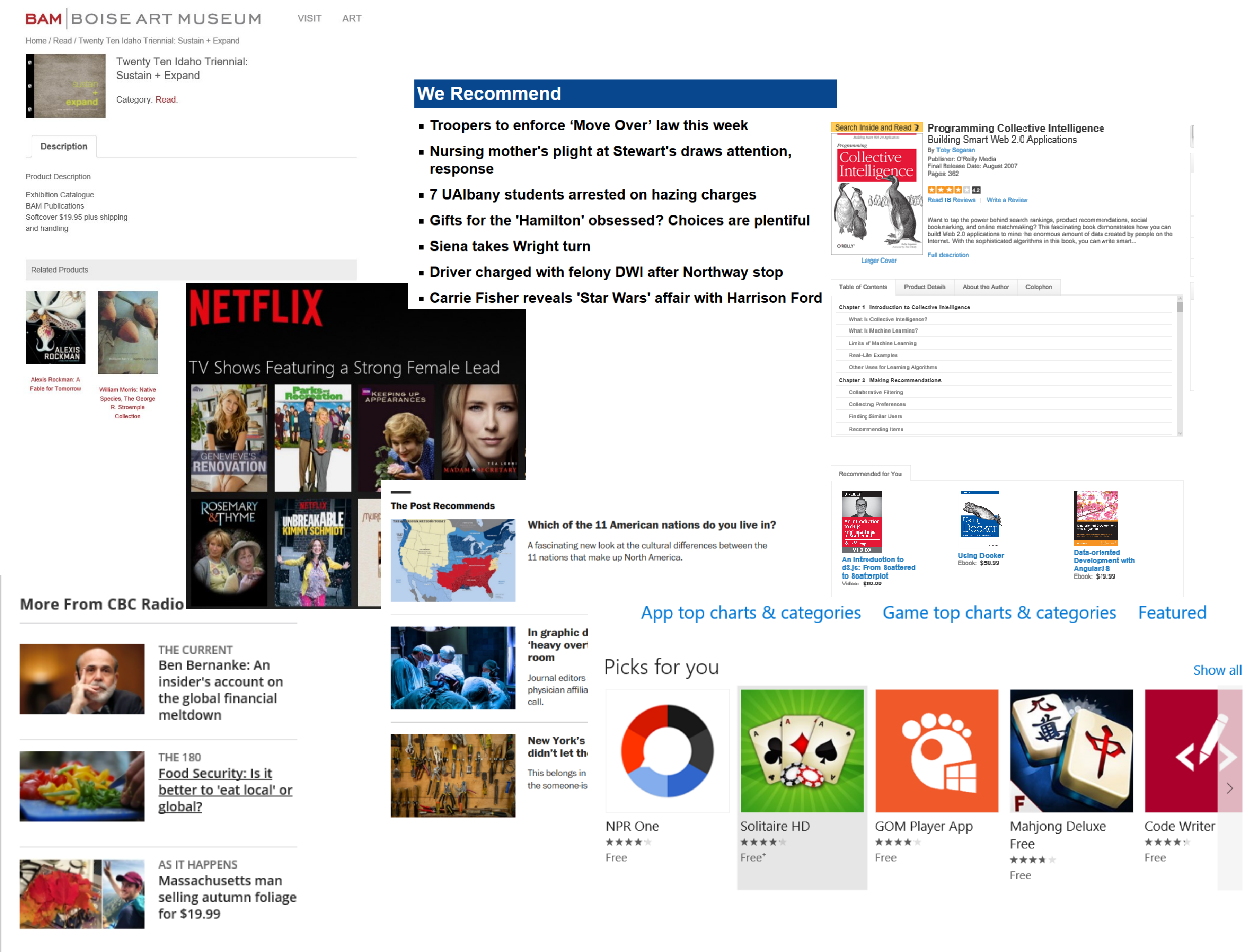
User Behavior in Recommender Choice

Recommendation in Context

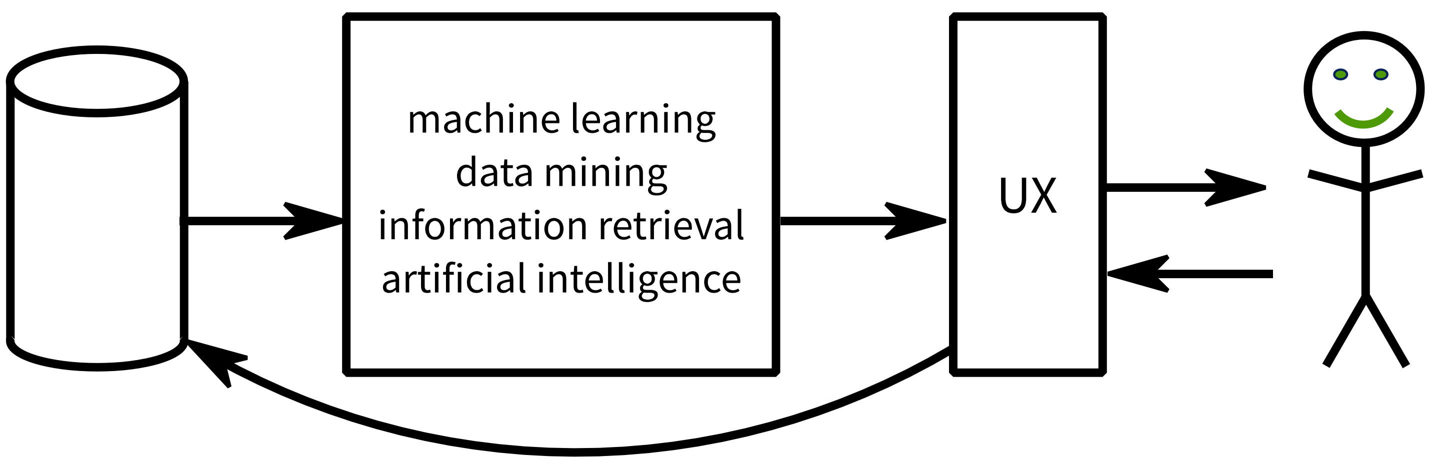
Wrapup

- **Background**
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Recommendation in Context
- Wrapup


Home
Moments
7
Notifications
Messages
Search Twitter
Tweet
TWEETS
27.8K
FOLLOWING
1,574
FOLLOWERS
34.8K
LIKES
18.4K
Following
University at Albany
@ualbany
Official Twitter stream of the University at Albany, SUNY ... THE WORLD WITHIN REACH! #UAlbany #GoGreatDanes
ALBANY, N.Y.
albany.edu
Joined February 2009
Tweet to University at Albany
2 Followers you know
3,122 Photos and videos
You might want to follow these similar accounts
close
Follow
UAlbany News
@UAlbanyNews
The official account for @UAlbany news and media relations.
Follow
Student Association
@UAlbanySA
The University at Albany's Undergraduate Student Government | Email us: sa@albany.edu | Instagram: ualbany.sa
Follow
SUNY
@SUNY
The largest comprehensive system of higher education in the country with 64 campuses within 20 miles of every New Yorker. Be Part of SOMETHING BIGGER
Tweets
Tweets & replies
Media
University at Albany Retweeted
UAlbany Career @UAlbanyCareer · 1h
@ualbany, want to know how acclaimed #UAlbany Alum reached their heights of success? Register for #AlumniTalks here: ow.ly /sgXG306dpE6
Who to follow · Refresh · View all
MOAS @moas_eu
Follow
Hillary Clinton @Hillary...
Follow
President Obama @PO...


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
GENEVIEVE'S RENOVATION
Parks and Recreation
KEEPING UP APPEARANCES
MADAM SECRETARY
ROSEMARY & THYME
UNBREAKABLE KIMMY SCHMIDT
We Recommend
Troopers to enforce 'Move Over' law this week
Nursing mother's plight at Stewart's draws attention, response
7 UAlbany students arrested on hazing charges
Gifts for the 'Hamilton' obsessed? Choices are plentiful
Siena takes Wright turn
Driver charged with felony DWI after Northway stop
Carrie Fisher reveals 'Star Wars' affair with Harrison Ford
Programming Collective Intelligence
Building Smart Web 2.0 Applications
By Toby Segaran
Publisher: O'Reilly Media
Final Release Date: August 2007
Pages: 362
Read 18 Reviews
Write a Review
Full description
Larger Cover
Table of Contents
Product Details
About the Author
Colophon
Chapter 1 : Introduction to Collective Intelligence
What is Collective Intelligence?
What is Machine Learning?
Limits of Machine Learning
Real-Life Examples
Other Uses for Learning Algorithms
Chapter 2 : Making Recommendations
Collaborative Filtering
Collecting Preferences
Finding Similar Users
Recommending Items
Recommended for You
An Introduction to d3.js: From Scattered to Scatterplot
Using Docker
Data-oriented Development with AngularJS
The Post Recommends
Which of the 11 American nations do you live in?
A fascinating new look at the cultural differences between the 11 nations that make up North America.
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
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; Pera, 2014]
- 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 [Rendle, 2009]

Learning about Users

Look at what they do

Implicit Feedback

Behavioral A/B Testing

Created by Luis Prado
from Noun Project

Listen to what they say

Explicit Feedback

User surveys

Focus groups

Created by Sarah Abraham
from Noun Project

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
- Recommendation in Context
- Wrapup

LensKit

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

LensKit

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, Boise 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.
[Ekstrand and Ludwig, 2016]

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Recommendation in Context
- Wrapup

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

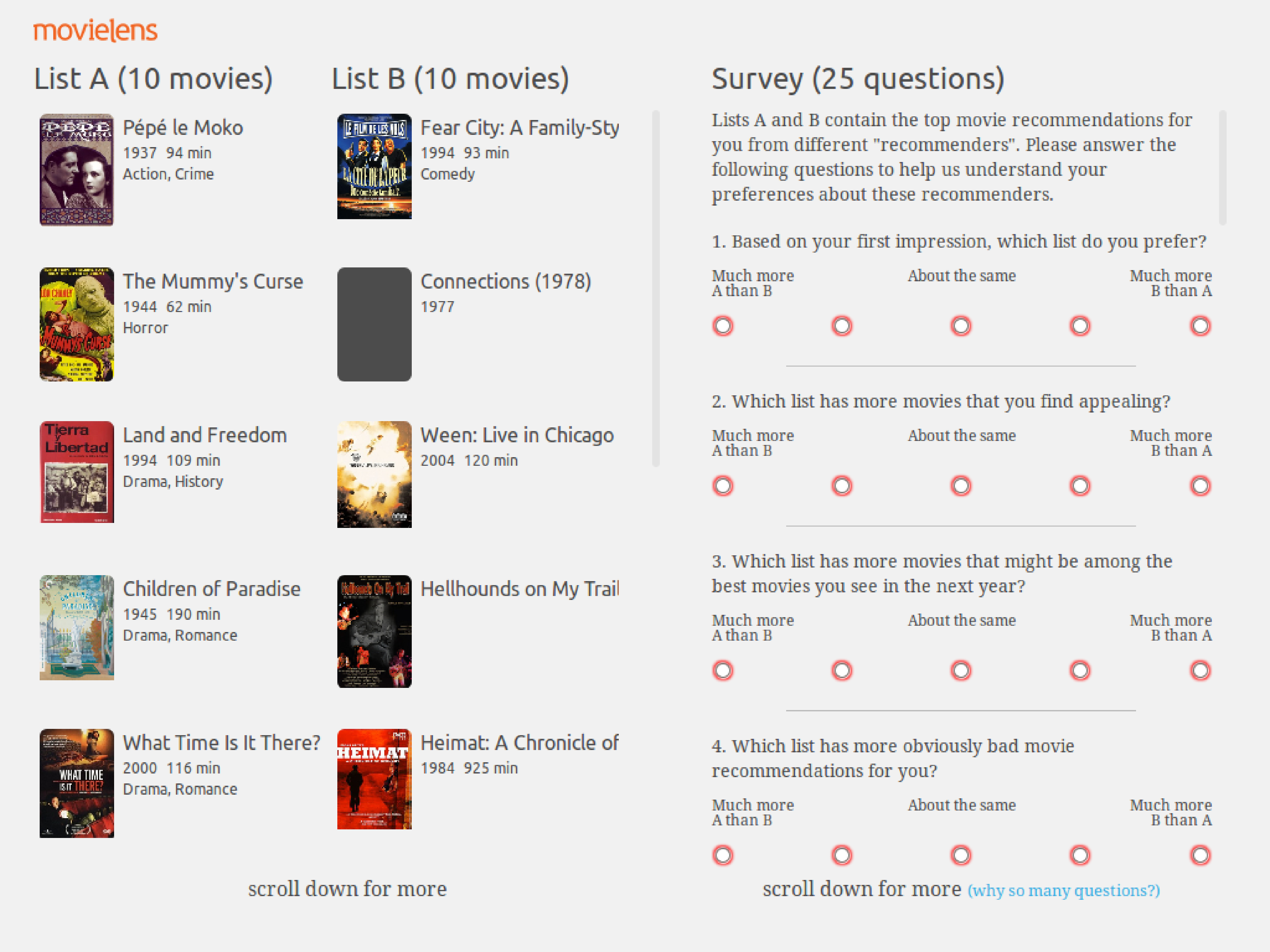
Offline Recommender Errors

User Perception of Recommendations

User Behavior in Recommender Choice

Recommendation in Context

Wrapup


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?

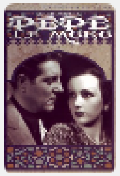
With GroupLens, Martijn Willemsen

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

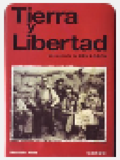
movielens

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

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?)

Experiment Features

Joint evaluation: users compare 2 lists

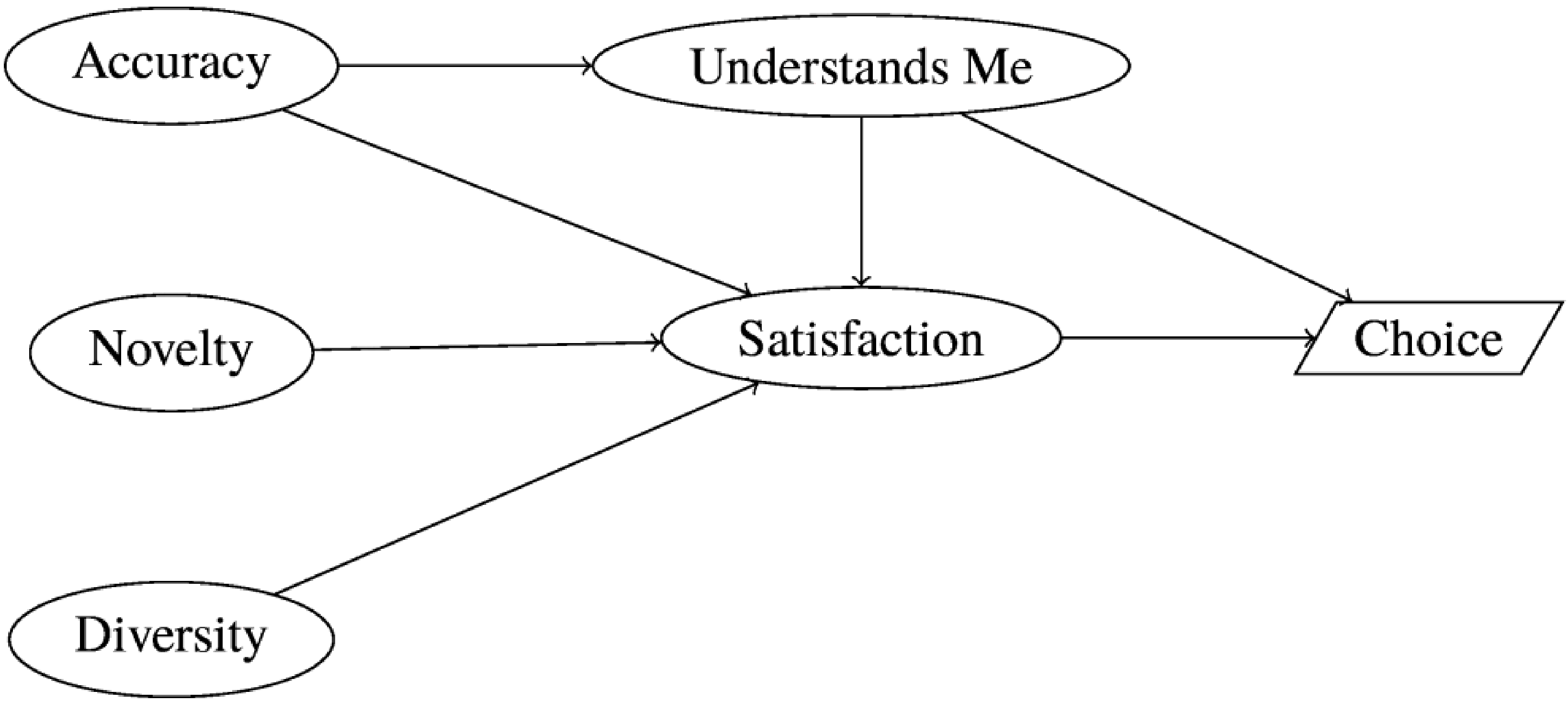
- enables more subtle distinctions than separate eval
- harder to interpret

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



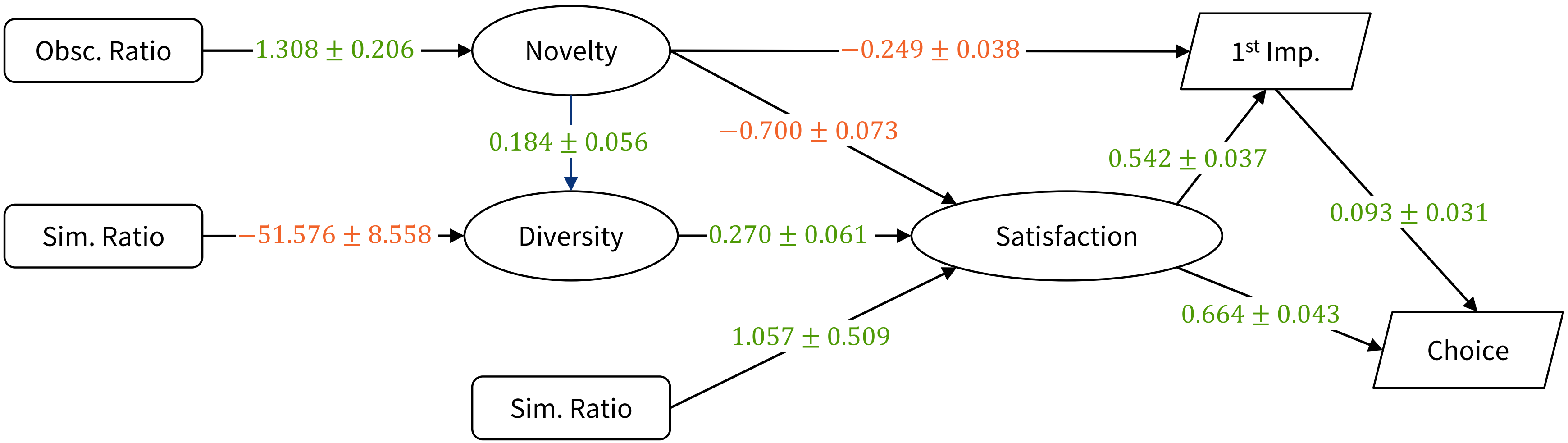
Response Summary

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%**

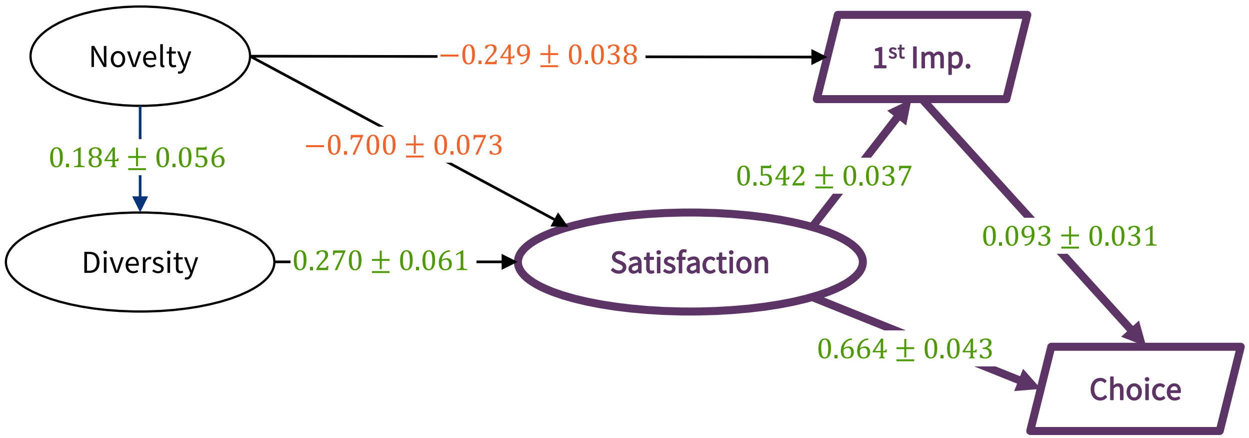
bold is significant ($p < 0.001$, H_0: ${}^{b}/_{n} = 0.5$)

Measurement Model



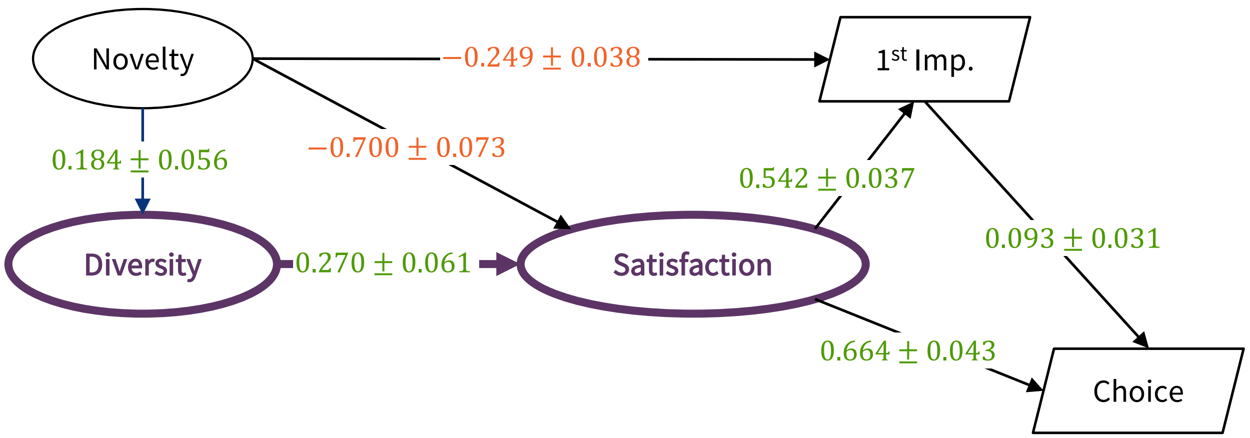
- Multi-level linear regression
- Direction comes from theory
- All measurements relative: positive is ‘more B than A’
- *Accuracy*, *Understands Me* folded into *Satisfaction*

Choice: Satisfaction



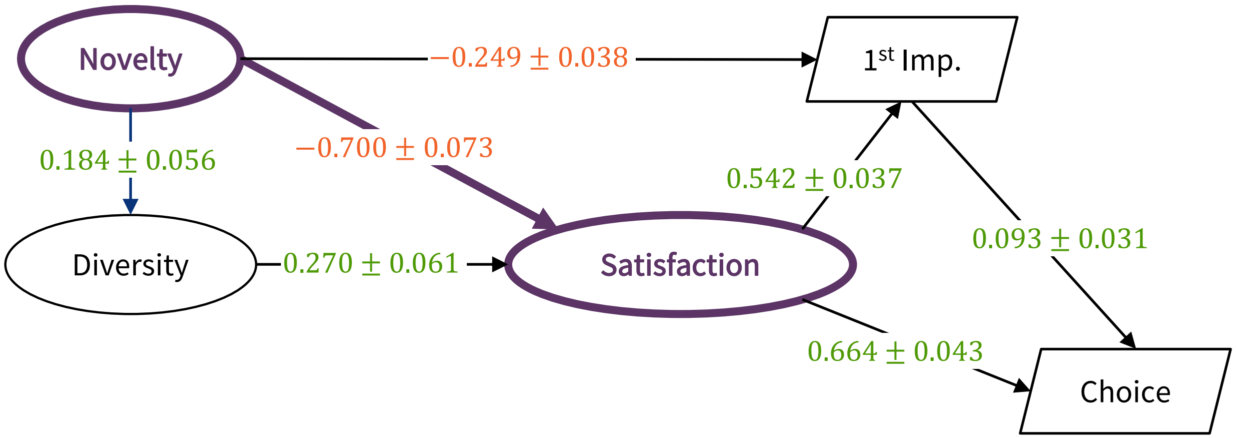
Satisfaction positively affects impression and choice

Choice: Diversity



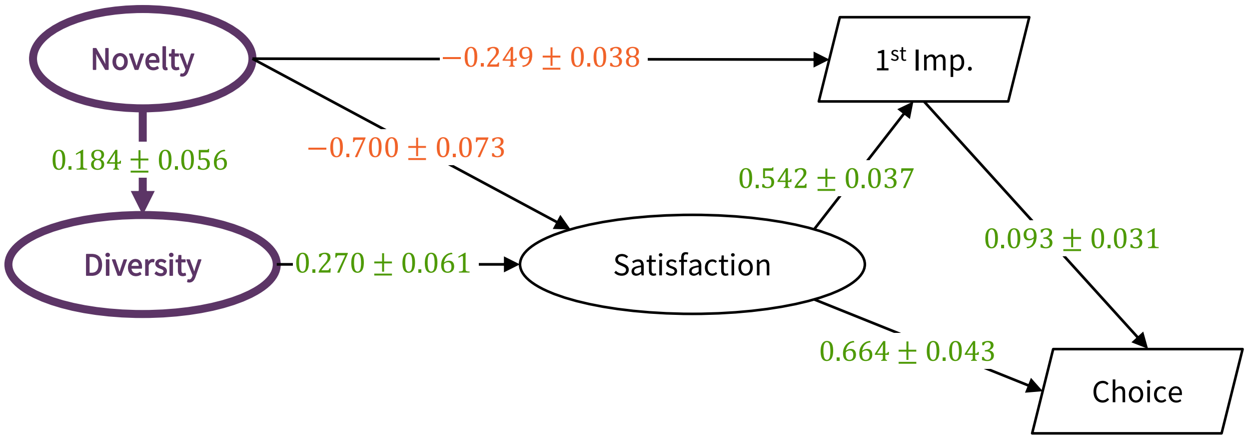
Diversity positively affects satisfaction and choice

Choice: Novelty



Novelty hurts satisfaction (and choice)

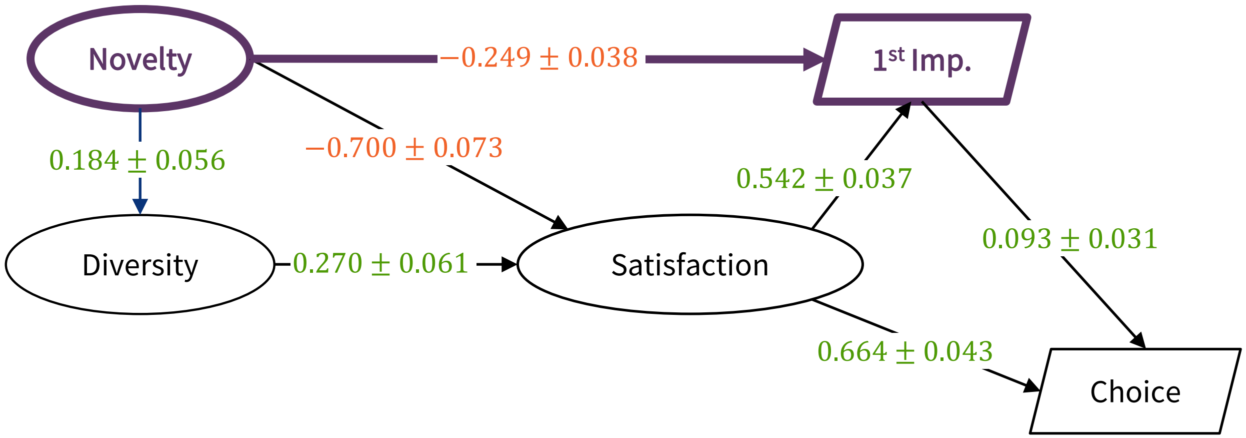
Novelty and Diversity



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

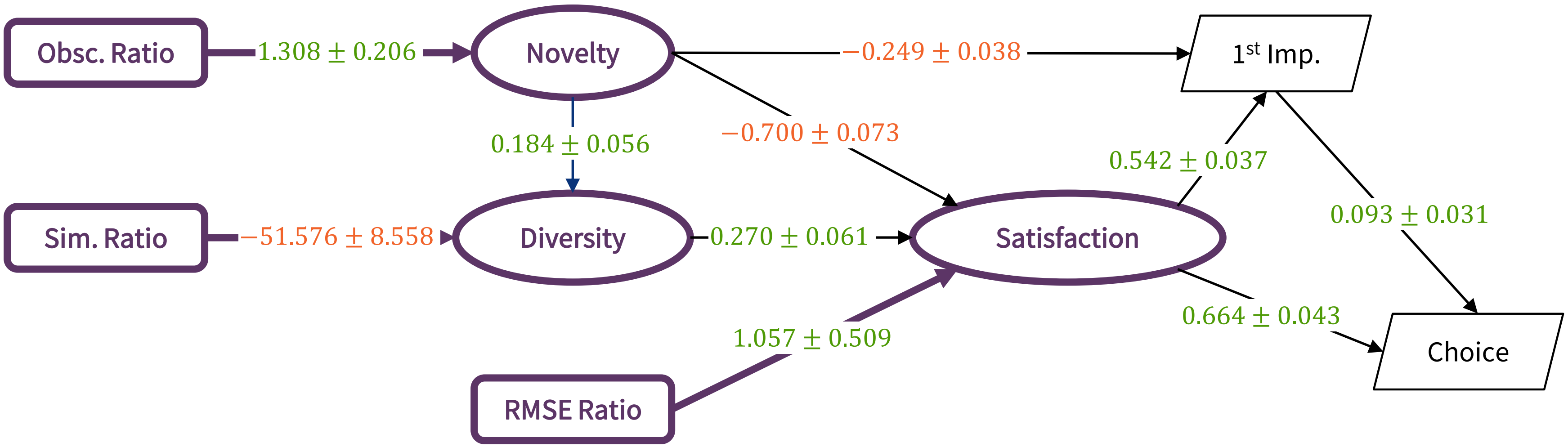
- Pairwise comparisons are difficult to interpret
- Method: re-interpret as 3 between-subjects pseudo-experiments:

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



- 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

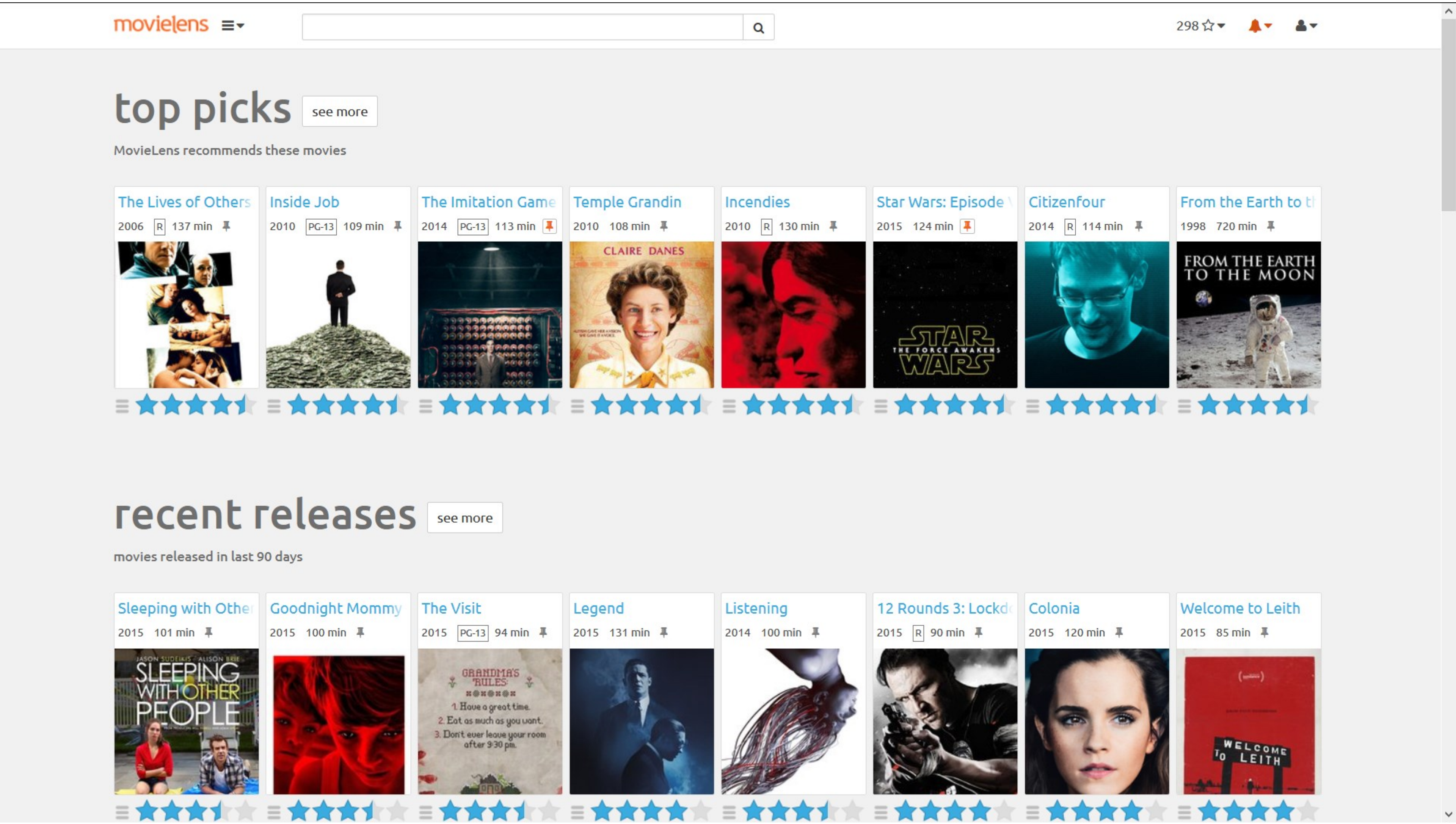
More research needed, of course

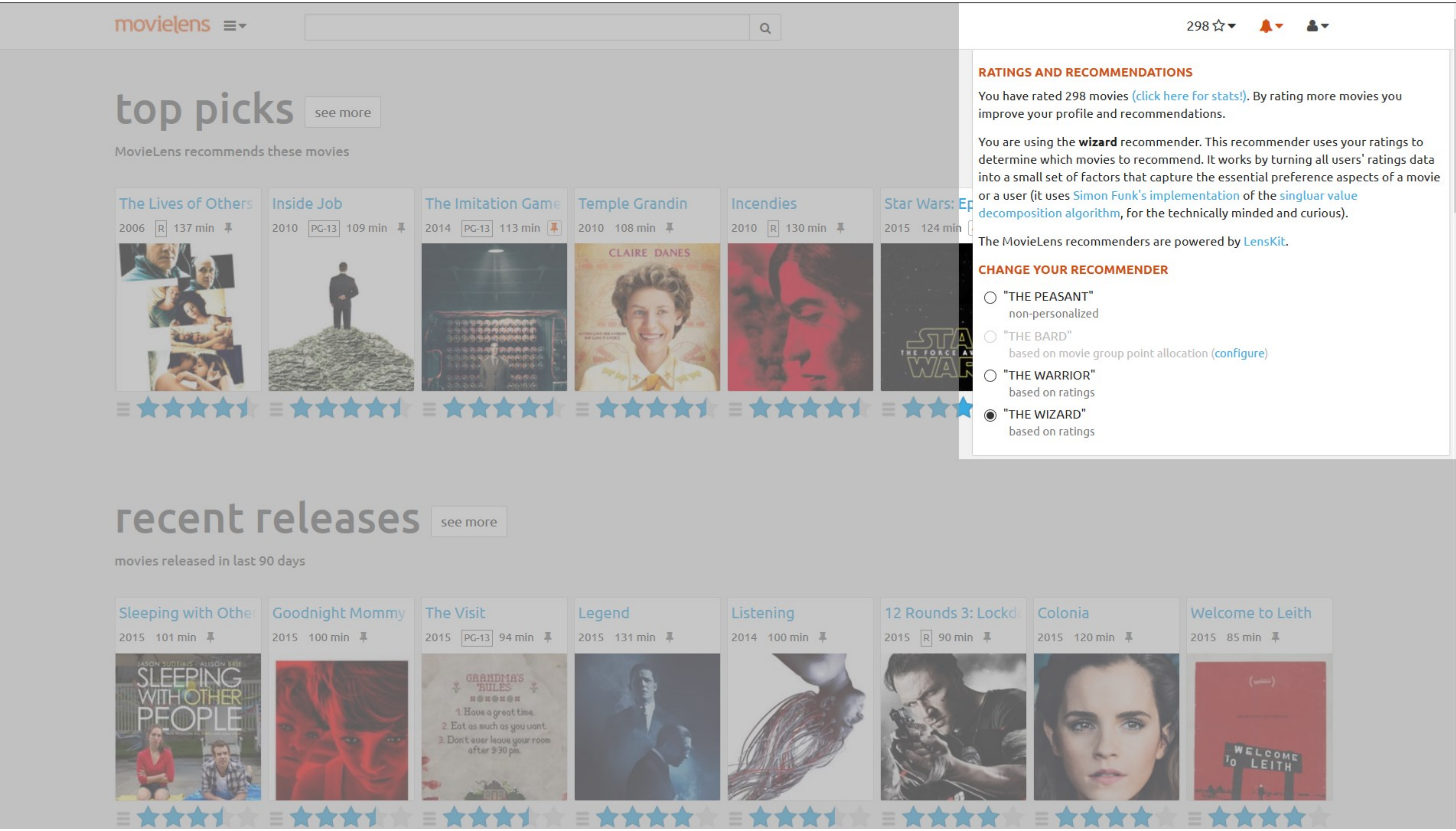
- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Recommendation in Context
- Wrapup

Giving Users Control
Ekstrand et al., RecSys 2015

Let's do it live!

- 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
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

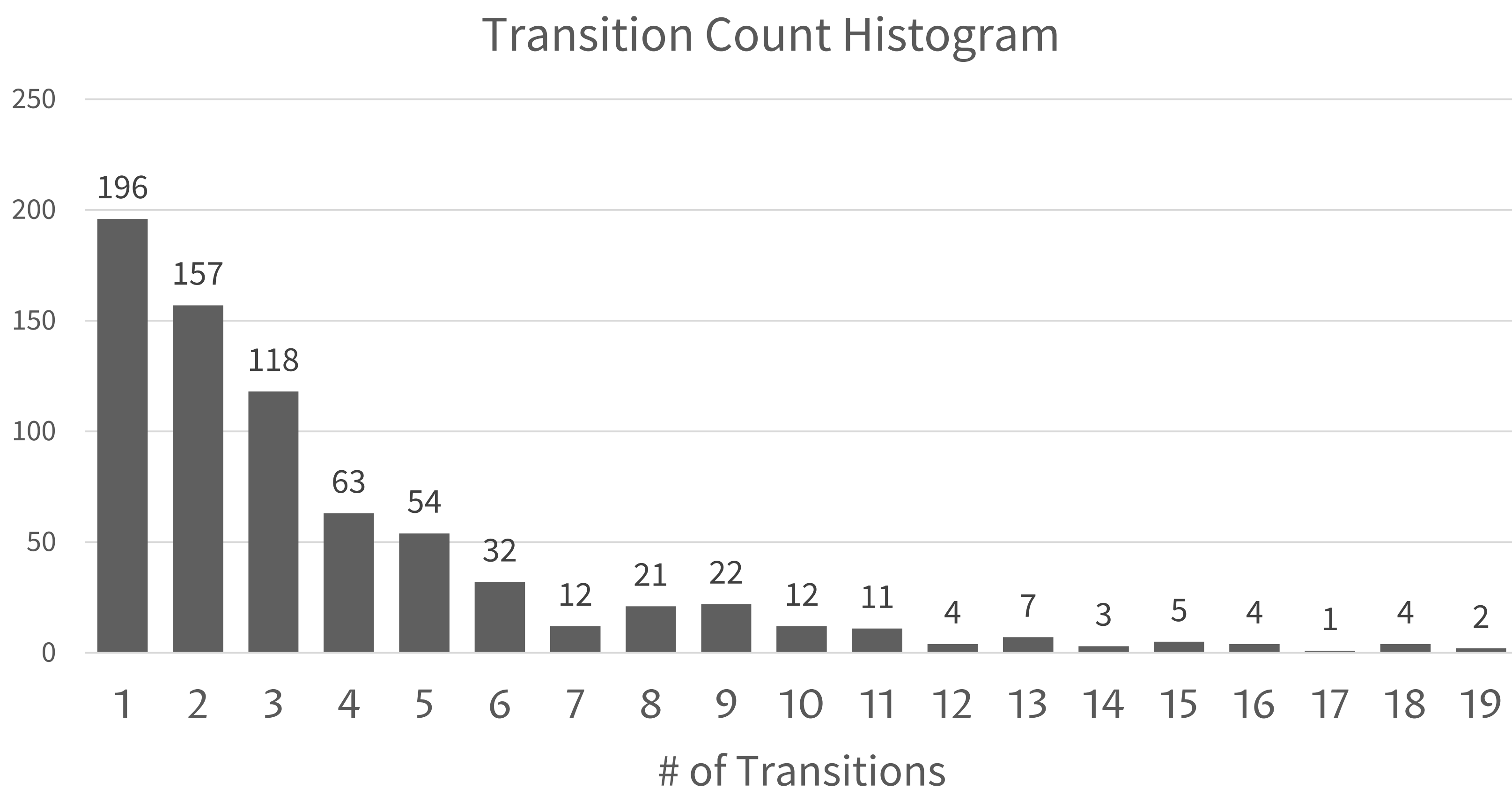

movielens
298
top picks
see more
MovieLens recommends these movies
The Lives of Others
2006 R 137 min
Inside Job
2010 PG-13 109 min
The Imitation Game
2014 PG-13 113 min
Temple Grandin
2010 108 min
CLAIRE DANES
Incendies
2010 R 130 min
Star Wars: Ep
2015 124 min
recent releases
see more
movies released 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
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

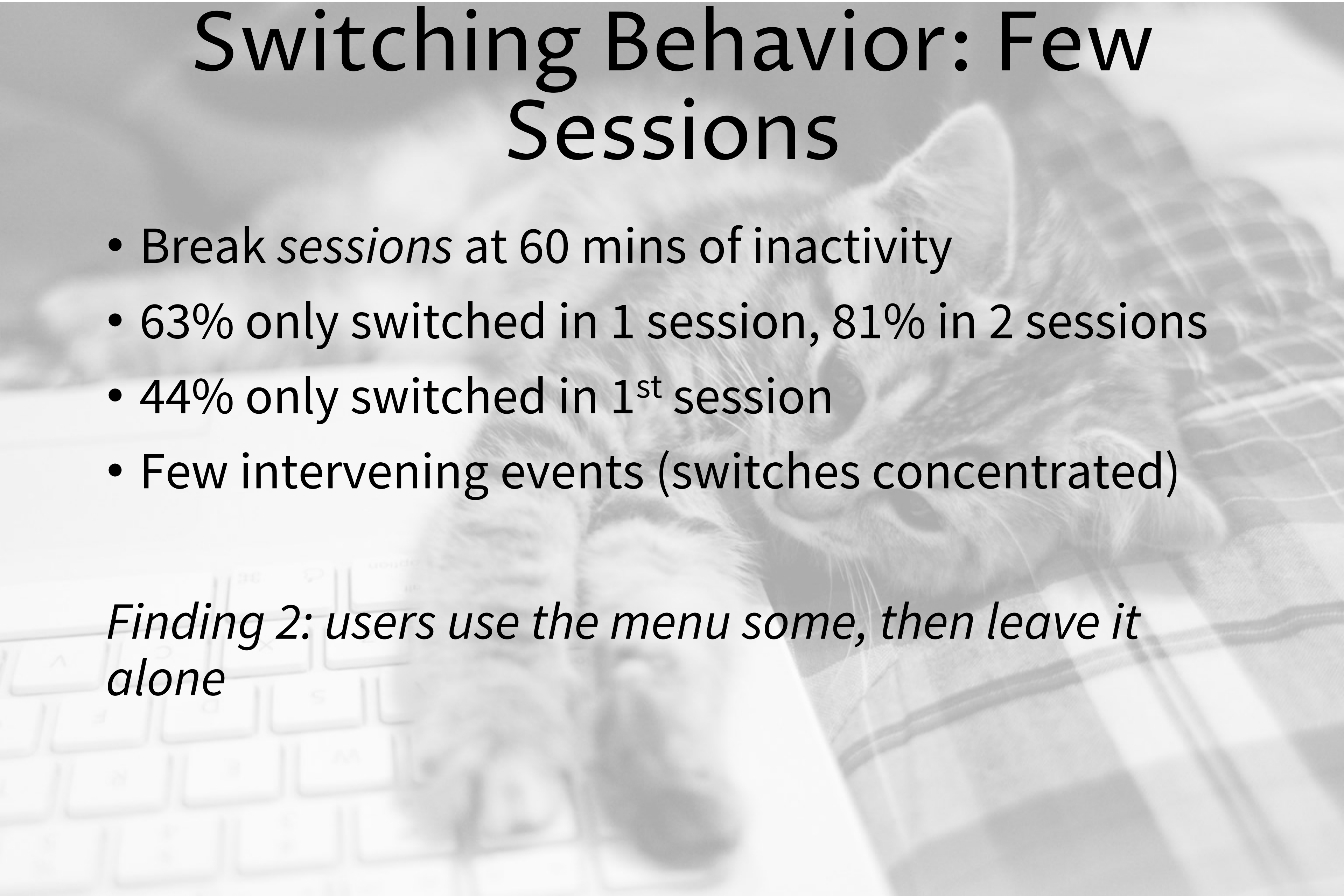
Users Switch Algorithms

- 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 (some)

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

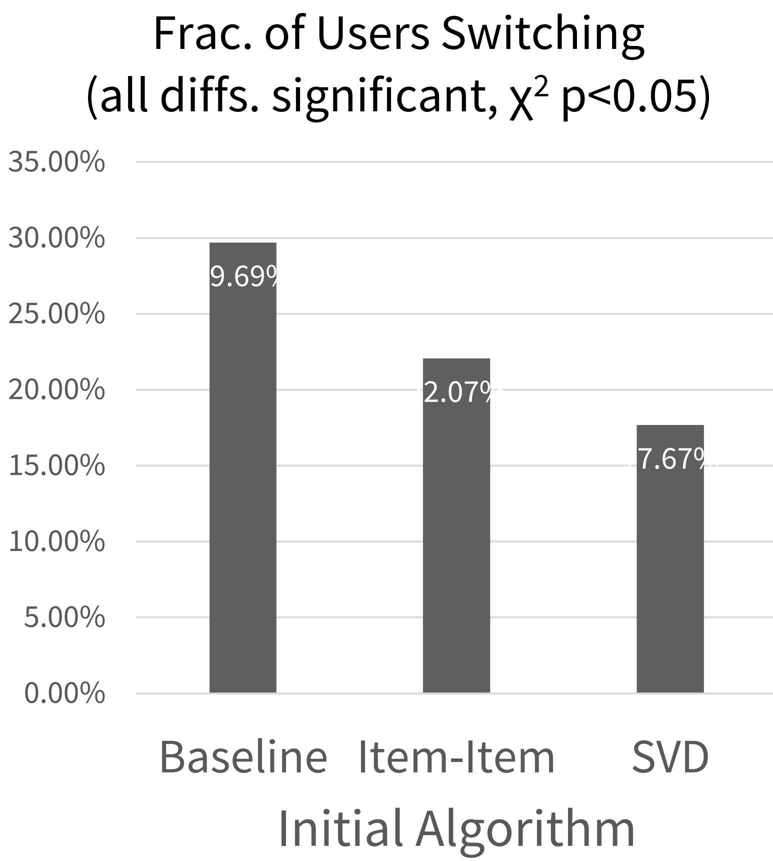
Source: Flickr user Ryan Forsythe. Used under CC-BY-SA.

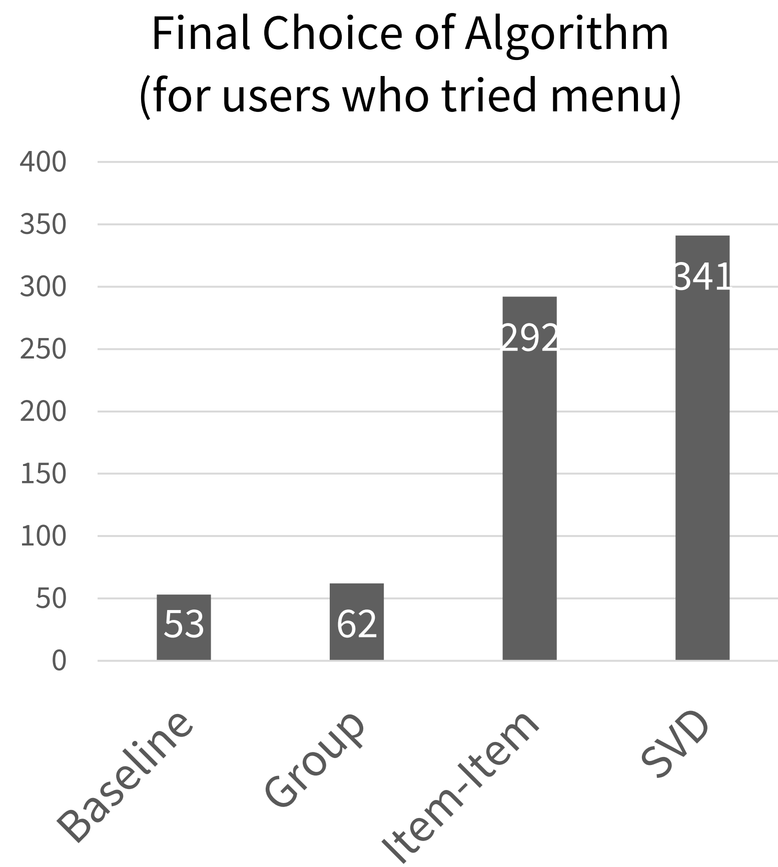
Algorithm Preferences

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





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

To Recap

3 studies, similar questions, similar outcomes

- Item-item and SVD very similar
- Different recommenders better in different cases
- Consistent theme across experimental settings

Opportunity to tailor to user needs.

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Recommendation in Context
- Wrapup

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

Fair Recommendation

- Fairness in machine learning and data mining is gaining research attention
- My questions:
  - What does it mean for a recommender to be fair?
  - Does the recommender *exacerbate*, *perpetuate*, or *mitigate* bias in its input?
  - How does the recommender react to user responses to its recommendations over time?
  - Can, and should, we build notions of fairness or representation into its logic?

Strong Impact

- Facebook and Google can swing elections
- News feed content, search results affect thought
- Visibility of issues or people in hands of recommender
  - Do films w/ lead actors of color sell as well?
  - If they don't, do studios make them?
- Recently: data mining affecting prison sentences

Questions

RQ1

Can we observe gender bias in users' book reading?

RQ2

Can we observe gender bias in recommendations?

RQ3

Does recommender propagate bias?

Methods

- Use BookCrossing book rating data
- Link with OpenLibrary for book metadata
- Run author names through `gender-detect`
  - Yes, this is broken. We know.
- Infer distribution of bias with hierarchical Bayesian model
  - Deals with differing user profile sizes
  - Will be augmenting with set & ranking fairness metrics

Early Results

Gender bias in book reading?

Yes, but mild and high-variance

Gender bias in recommendations?

Looks like yes, still need to tease out confounds

Propagate bias?

Not really

Future Work

- Improving analysis
- Improving demographic data
- More domains
- Feedback loop

Interdisciplinary Conversation

- CS alone cannot fix these problems
- Goal: contribute to interdisciplinary conversation
  - Data on current situation, impact of systems
  - Characterize response under hypothetical conditions
  - Provide testing ground to predict impact of proposed policy, technology, or interventions
- Dialogue with lawyers, ethicists, sociologists, psychologists, political scientists, etc.

- Background
- Tools and Instrumentation
- Offline Recommender Errors
- User Perception of Recommendations
- User Behavior in Recommender Choice
- Recommendation in Context
- Wrapup

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
- Bias in recommenders

Beyond Behaviorism

How can we **engage users** in recommender

evaluation

operation

design

to enable **great systems** that **meet users' needs** in accordance with **their values?**

Participatory Design for Recommenders

Limits of Behavioral Observation

Neil Hunt, RecSys '14 keynote:

NetFlix's metrics cannot distinguish between an enriched life and addiction.

Learning about Users

Look at what they do

Implicit Feedback

Behavioral A/B Testing

Created by Luis Prado
from Noun Project

Listen to what they say

Explicit Feedback

User surveys

Focus groups

Created by Sarah Abraham
from Noun Project

If they disagree?

Whose Values are Built For?

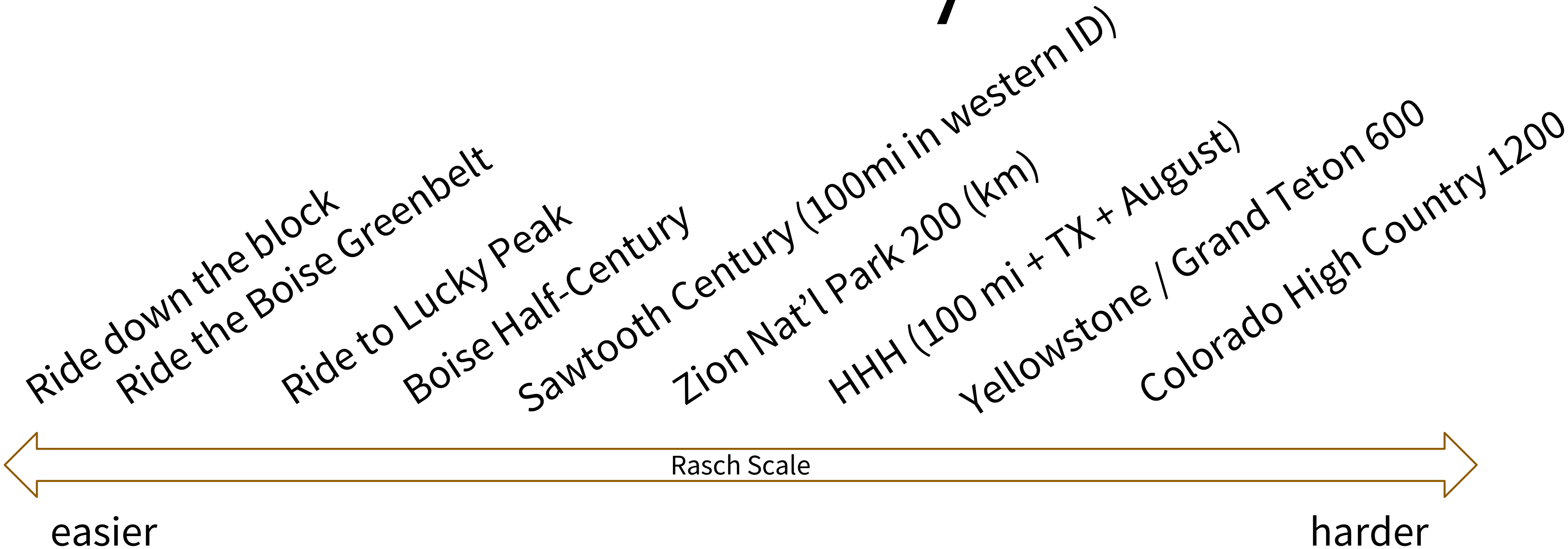
Many stakeholders, each with values:

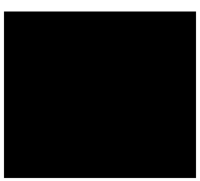
Shareholders
Management
Developers
Users

What values are embedded in the system?
Whose values are embedded in the system?

Behavior will not tell you values.

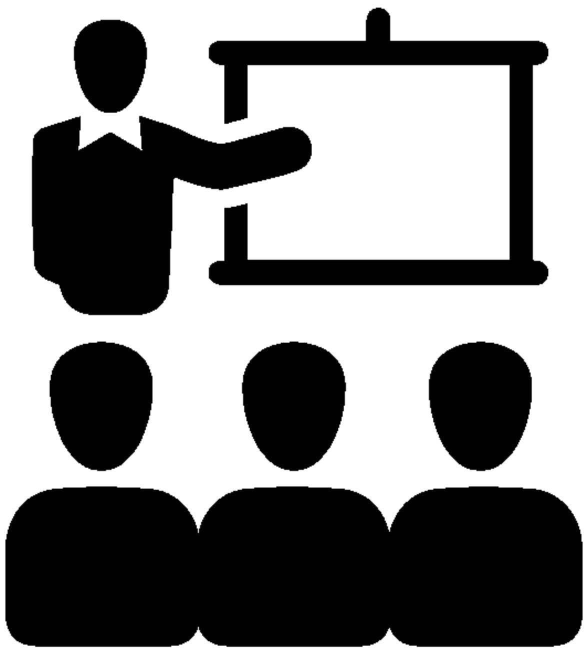
Difficulty


Ride down the block
Ride the Boise Greenbelt
Ride to Lucky Peak
Boise Half-Century
Sawtooth Century (100mi in western ID)
Zion Nat'l Park 200 (km)
HHH (100 mi + TX + August)
Yellowstone / Grand Teton 600
Colorado High Country 1200
Rasch Scale
easier
harder



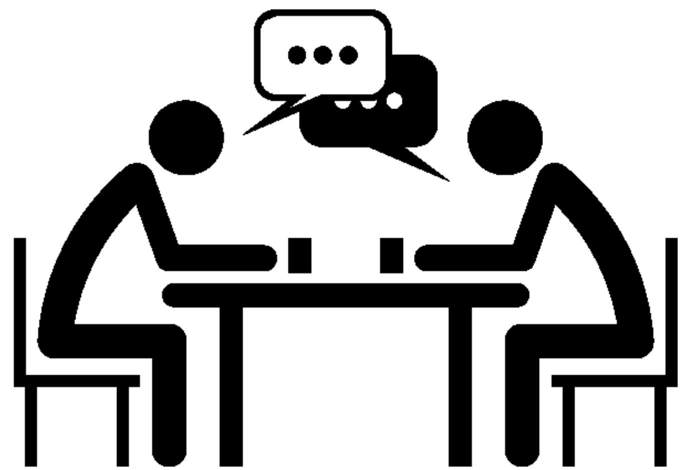
Reciprocity (Franklin, 1989)

Broadcast …



Created by Delwar Hossain
from Noun Project

… or conversation?



Created by Michael V. Suriano
from Noun Project

Thank you

Also thanks to:

- *Collaborators (PIReT, GroupLens, Martijn Willemsen)*
- *NSF*

PIReT

BOISE STATE UNIVERSITY

http://bit.ly/RecPeopleAN16