

Week 13

Agenda

- What is a recomendar system?
- Early recommender systems: Content vs Behavior
- Collaborative Filtering
- Netflix Prize: Ensembles and MF
- How most companies do recommendation today
- What the few top companies (e.g. Netflix, Amazon, Spotify) are doing today

AMAZON CIRCA 2005

Frequently Bought Together



Price For All Three: \$258.02

Add all three to Cart

- This item: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)** by Trevor Hastie
- [Pattern Recognition and Machine Learning \(Information Science and Statistics\)](#) by Christopher M. Bishop
- [Pattern Classification \(2nd Edition\)](#) by Richard O. Duda

Customers Who Bought This Item Also Bought



All of Statistics: A Concise Course in Statistical Inference by Larry Wasserman

4.5 stars (8) \$60.00



Pattern Classification (2nd Edition) by Richard O. Duda

4.5 stars (27) \$117.25



Data Mining: Practical Machine Learning Tools and Techniques by Ian H. Witten and Eibe Frank

4.5 stars (29) \$41.55



Bayesian Data Analysis, Second Edition (Texts in Statistical Science) by Andrew Gelman, John Carlin, Hal Stern, and Donald Rubin

4.5 stars (10) \$56.20



Data Analysis Using Regression and Multimethod Modeling by Andrew Gelman and Jennifer Hill

NETFLIX

Your Account | Buy | Redeem Gift | Help

Add Queue

Your Queue

"MissedMask" was moved to the top of your Queue.

55 movies

See Related Suggestions

At Home

Movie Title	Star Rating	MPAA	Genre	Shipped	Est. Arrival
1 Pirates of the Caribbean: Dead Man's Chest	★★★★½	PG-13	Action & Adventure	04/26/07	04/27/07
2 Crash	★★★★½	R	Drama	04/13/05	04/20/05

Get another movie for only \$2.50 and we'll send it faster! [Upgrade to the 2 of 4 Now \(Additional fees apply\)](#)

Reset viewing history

DVDs in Your Queue

Priority	Movie Title	Star Rating	MPAA	Genre	Availability	Remove To Top
1	MissedMask	★★★★½	PG	Family	Now	<input type="checkbox"/>
2	Home's Missing Cards	★★★★★	PG	Children & Family	Now	<input type="checkbox"/>

Update Your Queue

MCDONALD'S BITES ON BIG DATA WITH \$300 MILLION ACQUISITION



McDonald's Corp., in its largest acquisition in 20 years, is buying a decision-logic technology company to better personalize menus in its digital push.

The world's biggest restaurant chain is spending more than \$300 million on Dynamic Yield Ltd., according to a person familiar with the matter. With the new technology, McDonald's restaurants can vary their electronic menu boards' display of items, depending on factors such as the weather—more coffee on cold days and McFlurries on hot days, for example—and the time of day or regional preferences. The menus will also suggest add-on items to customers.



berkeley mids academic calendar



All

Shopping

News

Maps

Images

More

Settings

Tools

About 9,440 results (0.67 seconds)

I School Feeds | UC Berkeley School of Information

<https://www.ischool.berkeley.edu/feeds> ▾

Apr 13, 2018 - Academic calendar for students in on-campus degree programs. (From the UC ... Add MIDS & MICS Academic Calendar to my calendar.

Academic Calendar < University of California, Berkeley

guide.berkeley.edu/academic-calendar/ ▾

For PDFs of current and future **Berkeley Academic Calendars**, visit the Calendar page on the Office of the Registrar website. See the instructions at the bottom of ...



Search



Pop Music Recommended videos for you



Kings Of Leon - Hands To Myself (Selena Gomez cover...)

BBCRadio1VEVO
5.2M views • 1 year ago



Fiona Apple - I Walk a little faster (underwater)

Fiona Apple Rocks
105K views • 1 month ago



Harry Styles - The Chain (Fleetwood Mac cover) in the BBC Radio 1 Live Lounge

BBCRadio1VEVO
10M views • 10 months ago



First Aid Kit - Running Up That Hill (Kate Bush Cover)...

Sheen Gekoo
44K views • 2 weeks ago

Personal Data Marketplaces

The image shows a mobile application interface with a light blue header bar. The first section, titled "Basic Info", contains two items: "Graduate Degree" (with a "Demographic > Education > Graduate Degree" link) and "Gen X" (with a "Demographic > Age > Lifestyles > Gen X" link). Below this is a large empty white box. The second section, titled "Location & Neighborhood", includes a location pin icon. The third section, titled "Professional Interests", includes a briefcase icon. The fourth section, titled "Hobbies & Interests", includes a bowling pin icon. The fifth section, titled "What Others Know About You", includes an information icon.



DMPs: Personal data marketplaces

- Bluekai.com/registry
- Discuss.
- How do you feel about them?
- Which datasets about yourself do you assume are public?

Personal Data Marketplaces



TiVo: A Buyer's Guide

DATA TYPES: TV Viewership

TiVo Data 101

TiVo's expertise spans multi-screen functionality and back-end services that support linear television, video on demand (VOD), mobile apps, streaming, etc. Our experience in the market puts TiVo in a unique position to offer superior data for targeting television viewers and measuring sales impact. Our data segments power any planning, modeling, or reporting you can imagine. Our viewership data is ready to be manipulated by sophisticated data science teams familiar with processing large datasets.

Description of Data Types:

Flat Files or TV Viewership data – Raw data processed by a buyer's internal data teams.

Syndicated TV Segments – Viewership segments that span across verticals and viewership attributes.

Campaign Segments – Ad-exposed segments used for an attribution ad matched to digital campaigns.

Custom Segments – Ability to create segments not readily available in data stores.

Collection Methodology:

TiVo's TV Viewership data is sourced from TiVo's retail set-top boxes and from MVPDs who use TiVo's software. For our segments, we deploy a 1:1 deterministic match to 1st and 3rd party data with our owned and licensed set-top box data. Experian is our leading partner to execute the 3rd party match process based on name/address. For our raw data, matching occurs at household level via Experian with each device viewership mapped to each household. While we don't model our raw data, our segments can be modeled in-house via a 3rd party partner or within the data stores.

Use Our Data For:

- Targeting cross-platform
- Optimizing frequency
- Extending reach
- Improving campaign KPIs
- Appending TV data to digital campaigns

Personal Data Marketplaces

Evite: A Buyer's Guide

Evite Data 101

With more than 22 million registered users and over 25,000 invitations sent each hour, Evite is the top online invitation and social planning website.

Launched in 1998, Evite is headquartered in Los Angeles.

Description of Data Types

Segment/Definition

- Presence of Child in Household
 - » Host of Kids Birthday
 - » Babys first Bday
 - » Kids Corner
 - » Kids themes
 - » Halloween for Kids
- Age of Child in Household
 - » Title scrape for numbers (1st, first, First) in Host of Birthday for Kids
 - » Age assumed to be 1 for Host of Baby's first
- Recent Movers/Furnishers
 - » Host of Housewarming party
- Bride
 - » Host of Save the Date
- Wedding Attendee
 - » Host and Guest of Wedding/Engagement
 - » Bridal Shower
 - » Bachelor
 - » Bachelorette Party
 - » Save the Date

IDEA

evite®

- Pre-Natal/Expecting
 - » Host or Guest of Baby Shower
- Upcoming Birthday
 - » Host or Guest of Birthday for Her
 - » Birthday for Him
- Recent Graduate
 - » Host or Guest of Graduation
- Sports Enthusiast
 - » Host or Guest of Sports/Leagues
- Home Entertainers
 - » Host or Guest of Hostess Party
 - » Dinner Party
 - » Cocktail Party
 - » House Party
 - » BBQ/Pool Party
 - » Pot Luck
 - » Game Night
- Travel intenders
 - » Hosts or Guest of Trips/Getaways
- Halloween
 - » Host or Guest of Halloween Party
- Winter Holidays
 - » Host or Guest of Winter Holiday Party
- Super Bowl
 - » Host or Guest of The Big Game Party
- Thanksgiving
 - » Host or Guest of Thanksgiving
- Religious
 - » Host or Guest of Religious Event

< 50 >

< 51 >

Personal Data Marketplaces

TECH

Vizio nears \$17 million settlement for TV data-tracking lawsuit

The deal is still subject to final approval

By Chaim Gartenberg | @cgartenberg | Oct 4, 2018, 1:30pm EDT

f t SHARE

The image shows a Vizio television screen displaying a news article from The Verge. The article is titled "Vizio nears \$17 million settlement for TV data-tracking lawsuit". Below the title, it says "The deal is still subject to final approval". The author is Chaim Gartenberg (@cgartenberg) and the date is Oct 4, 2018, 1:30pm EDT. There are social sharing icons for Facebook, Twitter, and a "SHARE" button. The TV screen itself displays a travel-related app (Conde Nast Traveler) and a grid of streaming service icons including Netflix, Amazon Prime, Hulu, and others.

Personal Data Marketplaces

“One of the biggest challenges in protecting privacy is that many of the violations are invisible,” Cook writes. “For example, you might have bought a product from an online retailer—something most of us have done. But what the retailer doesn’t tell you is that it then turned around and sold or transferred information about your purchase to a ‘data broker’—a company that exists purely to collect your information, package it, and sell it to yet another buyer.

“The trail disappears before you even know there is a trail. Right now, all of these secondary markets for your information exist in a shadow economy that’s largely unchecked-out of sight of consumers, regulators, and lawmakers. Let’s be clear: You never signed up for that. We think every user should have the chance to say, ‘Wait a minute. That’s my information that you’re selling, and I didn’t consent.’”



Taste Domains



Taste Domains

- Early research focused on 'taste' domains, particularly movies, music, and books.
- In these domains, 'finding' often involved suggestions from friends or tastemakers
- Researchers created collaborative filtering and other approaches as means of emulating this process
- Usually cast as ratings prediction problem in part because its relatively easy to collect ratings data

1996 MovieLens (Minnesota)

- <http://en.wikipedia.org/wiki/MovieLens>
- Research project collected ratings on movies, etc
- Very early Amazon and Netflix strongly influenced by this
- User based CF: Find k-nearest users and use their ratings
- 2001 Item based CF: Find k nearest items to those items a user prefers.
http://files.grouplens.org/papers/www10_sarwar.pdf
- Still popular recommendation algorithm

Movielens (Late 90's)

m o v i e l e n s
helping you find the *right* movies

Welcome aibc
You're the 24th visitor in the past hour.

So far you have rated **15** movies.
MovieLens needs at least **15** ratings from you to generate predictions.
Please rate as many movies as you can from the list below.

Your Rating		Movie Information
???	Not seen	Beneath the Planet of the Apes (1970) Action, Sci-Fi
???	Not seen	Gift, The (2000) Thriller
???	Not seen	Great Muppet Caper, The (1981) Children, Comedy
???	Not seen	Heaven Can Wait (1978) Comedy
★★★☆	4.0 stars	Hitch (2005) Comedy, Romance
???	Not seen	Kate & Leopold (2001) Comedy, Romance
???	Not seen	Muppets Take Manhattan, The (1984) Children, Comedy, Musical
???	Not seen	Police Academy 4: Citizens on Patrol (1987) Comedy
???	Not seen	Saturday Night Fever (1977) Comedy, Drama, Romance
???	Not seen	Teenage Mutant Ninja Turtles II: The Secret (1991) Action, Children, Fantasy

To get a new set of movies click the [next>](#)

m o v i e l e n s - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección <http://movielens.umn.edu/search/searchPhrase=&action=newSearch&hiddenParam=1&genre=All&date=All&domain=All&genreSearch=Search+Genre%2FDate%21>

Welcome dus@infovis.net
You've rated **48** movies.
You're the 24th visitor in the past hour.

★★★★★ = Must See
★★★★ = Will Enjoy
★★★★ = It's OK
★★★★ = Fairly Bad
★★★★ = Awful

[Home](#) | [Manage Buddies](#) | [Your Account](#) | [Help](#) | [Logout](#)

You've searched for **all titles**.
Found 7233 movies, sorted by **Prediction**.
Genres: All | Exclude Genres: None
Dates: All | Domain: All | Format: All | Language: All
[Show Printer-Friendly Page](#) | [Download Results](#) | [Suggest a Title](#)

Page 1 of 483 | Go to page: [1](#) ... [96](#) ... [192](#) ... [288](#) ... [384](#) ... [480](#) ... [last](#) | [page 2>](#)

Predictions for you	Your Ratings	Movie Information	Wish List
★★★★★	Not seen	Tainted (1998) info imdb Comedy, Thriller	<input type="checkbox"/>
★★★★★	Not seen	Friday Night Lights (2004) info imdb Action, Drama	<input type="checkbox"/>
★★★★★	Not seen	Harry Potter and the Prisoner of Azkaban (2004) info imdb Adventure, Children, Fantasy	<input type="checkbox"/>
★★★★★	Not seen	Spider-Man 2 (a.k.a. Spiderman 2) (2004) info imdb Action, Fantasy, Sci-Fi, Thriller	<input type="checkbox"/>
★★★★★	Not seen	Finding Nemo (2003) DVD , VHS , info imdb Adventure, Animation, Children, Comedy	<input type="checkbox"/>
★★★★★	Not seen	X-Men 2 (a.k.a. X2: X-Men United) (2003) DVD , VHS , info imdb Action, Adventure, Sci-Fi	<input type="checkbox"/>
★★★★★	Not seen	Oliver Twist (1948) info imdb Adventure, Crime, Drama	<input type="checkbox"/>
★★★★★	Not seen	Raiders of the Lost Ark (1981) DVD , info imdb Action, Adventure	<input type="checkbox"/>
★★★★★	Not seen	Indiana Jones and the Last Crusade (1989) DVD , info imdb Action, Adventure	<input type="checkbox"/>

m o v i e l e n s
helping you find the *right* movies

Shortcuts Search

Search Titles Use selected buddies!

Search by Genre/Date Use selected buddies!

Advanced Search

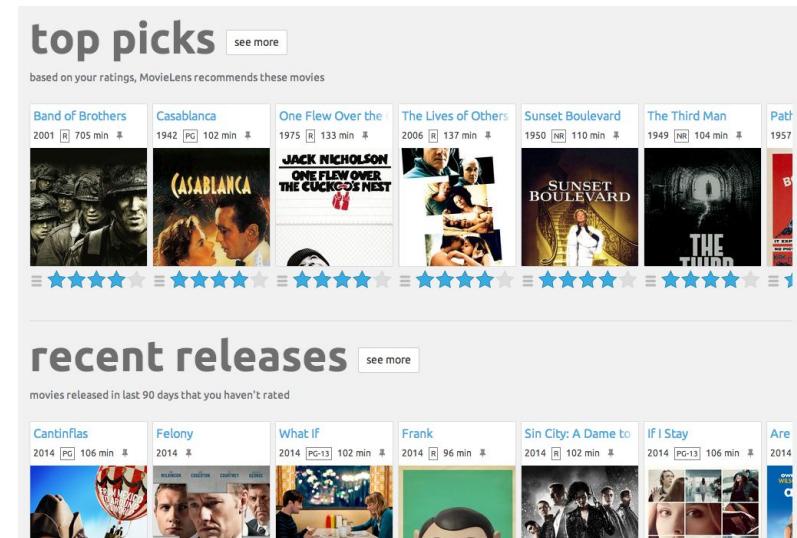
Select Buddies Test Buddy [What are buddies?](#)

Movielens (Today)

Stable benchmark dataset. 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users.
Released 1/2009.

- [README.html](#)
- [ml-10m.zip](#) (size: 63 MB, [checksum](#))

Permalink: <http://grouplens.org/datasets/movielens/10m/>



The image shows two sections from the MovieLens website: "top picks" and "recent releases".

top picks: A section titled "top picks" recommends six movies based on user ratings. Each movie card includes the title, year, rating (e.g., R), runtime, and a small thumbnail image. Below each card is a row of five blue stars, indicating the average rating for that movie.

Movie	Year	Rating	Runtime
Band of Brothers	2001	R	705 min
Casablanca	1942	PG	102 min
One Flew Over the Cuckoo's Nest	1975	PG	133 min
JACK NICHOLSON ONE FLEW OVER THE CUCKOO'S NEST	1975	PG	133 min
The Lives of Others	2006	R	137 min
Sunset Boulevard	1950	NR	110 min
The Third Man	1949	NR	104 min
Path	1957	NR	104 min

recent releases: A section titled "recent releases" shows movies released in the last 90 days that haven't been rated. Each movie card includes the title, year, rating, runtime, and a small thumbnail image.

Movie	Year	Rating	Runtime
Cantinflas	2014	PG	106 min
Felony	2014	NR	102 min
What If	2014	PG-13	102 min
Frank	2014	R	96 min
Sin City: A Dame to Kill For	2014	R	102 min
If I Stay	2014	PG-13	106 min
Are We There Yet?	2014	NR	104 min

Recommending Houses

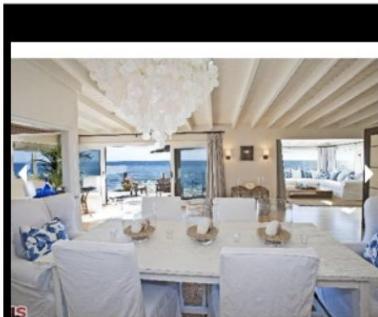
Get suggested homes just for you BETA

Step 1: Select the place you like best in Malibu, CA

[Pick a new city ▲](#)

Malibu, CA

[Pick City](#)



\$7,995,000 4 bd / 4 ba / 2,642 sqft

24604 Malibu Rd
Malibu, CA 90265

[Select](#)



\$2,595,000 5 bd / 5,481 sqft

30010 Andromeda Ln
Malibu, CA 90265

[Select](#)



\$1,999,000 6 bd / 8 ba / 5,012 sqft

4400 Encinal Canyon Rd
Malibu, CA 90265

[Select](#)



\$23,800,000 7 bd / 8 ba

26848 Pacific Coast Hwy
Malibu, CA 90265

[Select](#)

Recommending Houses

Get suggested homes just for you BETA

Finding your suggestions
Don't see anything you like? Start a new search and like homes in the results.

Thumbnail	Price	Bedrooms	Bathrooms	Sq Ft	Action
	\$35,000,000	7	13	14,395 sqft	
	\$22,995,000	8	8	10,979 sqft	
	\$37,500,000	6	9	11,722 sqft	
	\$21,500,000	5	9	9,000 sqft	
	\$19,950,000	5	8	13,927 sqft	
	\$18,850,000	6	10	12,002 sqft	
	\$17,900,000	7	9	13,167 sqft	
	\$22,500,000	7	9	16,000 sqft	
	\$47,500,000	12	15	23,649 sqft	
	\$21,995,000	7	8	9,737 sqft	

Search vs. Discovery

- Real estate search used to involve looking at classifieds in newspaper. Sites like Trulia are the online equivalent.
- Real estate recommendation used to (and still does) involve agents passing along listings. These recommendations get better as the agent learns more about the homebuyers.
- There are no real estate sites that emulate agent recommendations particularly well

Listing Discovery

- Important decision in one's life
- Only partly a taste domain--utility also plays a role (e.g. space, commute distance)
- Decision making happens over a period of time
- Tastes sometimes evolve during search process
- Social component to both (family, agent, sometimes friends)

Recommending Houses

Your Homes

Recommendations 23 Followed 10 Liked 11 Hidden 200 From My Agent 121 From My Wife 1,123 Recently Viewed

These are homes you and your agent, Alice Agent, are sharing.

The image shows three house listings from a real estate platform. Each listing includes a thumbnail image, price, address, and basic details. A callout box with a quote from 'Alice Agent' is overlaid on the first listing.

Listing 1: You and your wife may like this home because it's totally unique, unlike all of these homes on this mock.
-Alice Agent

Listing 2: \$899,000 / 2 bd / 2 ba / 1,215 sqft
235 Berry St. #611 San Francisco, CA 94105 South of Market (SoMA)

Listing 3: \$899,000 / 2 bd / 2 ba / 1,215 sqft
235 Berry St. #611 San Francisco, CA 94105 South of Market (SoMA)

Listing 4: \$899,000 / 2 bd / 2 ba / 1,215 sqft
235 Berry St. #611 San Francisco, CA 94105 South of Market (SoMA)

Netflix DVD Recommendations

Movies You've Rated

Based on your 745 movie ratings, this is the list of movies you've seen. As you discover movies on the website that you've seen, rate them and they will show up on this list. On this page, you may change the rating for any movie you've seen, and you may remove a movie from this list by clicking the 'Clear Rating' button.

TITLE	MPAA	GENRE	STARS	ACTIONS
12 Angry Men (1957)	UR	Classics		Add
The 39 Steps (1935)	UR	Classics		Add
An American in Paris (1951)	UR	Classics		Add
The Andromeda Strain (1971)	G	Sci-Fi & Fantasy		Add
Apollo 13 (1995)	PG	Drama		Add
The Battle of Algiers (1965) La Battaglia di Algeri	UR	Foreign		Add
Being There (1979)	PG	Drama		Add
Big Deal on Madonna Street (1958) I soliti ignoti	UR	Foreign		Add
The Birds (1963)	PG-13	Thrillers		Add
Blade Runner (1982)	R	Sci-Fi & Fantasy		Add



Welcome [dustaninfo.net](#)
You've rated 48 movies.
You're the 24th visitor in the past hour.

movielens - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Busqueda Favoritos Multimedia
DIRECCIÓN: http://movielens.unn.edu/search/searchPhrase=Baction=newSearch&hiddenParam=1&genre>AllGenre=&allDomain=&allGenreSearch=Search+Genre%2Fdate%21

Shortcuts Search

Search Titles
 Use selected buddies!

Search by Genre/Date
All Genres Domain: All movies Use selected buddies!

Advanced Search
 Test Buddy

Predictions Your Ratings Movie Information Wish List

You've searched for all titles. Found 7233 movies, sorted by Prediction. Genres: All | Exclude Genres: None. Dates: All | Domain: All | Format: All | Language: All. Show Printer-Friendly Page | Download Results | Suggest a Title. Page 1 of 483 | Go to page: 1...96...192...288...384...480...last page 2>

PREDICTION	MOVIE	INFO	IMDB
★★★★★	Tainted (1998)	info	imdb
★★★★★	Friday Night Lights (2004)	info	imdb
★★★★★	Harry Potter and the Prisoner of Azkaban (2004)	info	imdb
★★★★★	Spider-Man 2 (a.k.a. Spiderman 2) (2004)	info	imdb
★★★★★	Finding Nemo (2003)	DVD, VHS, info	imdb
★★★★★	X-Men 2 (a.k.a. X2: X-Men United) (2003)	DVD, VHS, info	imdb
★★★★★	Oliver Twist (1948)	info	imdb
★★★★★	Raiders of the Lost Ark (1981)	DVD, info	imdb
★★★★★	Indiana Jones and the Last Crusade (1989)	DVD, VHS, info	imdb

Netflix DVD Recommendations

The screenshot shows two main sections. On the left, a yellow header reads "Movies You've Rated". Below it is a table listing ten movies with columns for Title, MPAA rating, and genre. Each row has an "Add" button with a star icon. On the right, a white box titled "Rating Activity" displays statistics: # of Ratings: 745, # Favorite Genres: 0, # Recommendations: 428, and # of Reviews Written: 5. Below these stats is a 5-star rating scale with a "Clear Rating" button. The main area lists the same ten movies from the table, each with a 5-star rating scale and a "Clear Rating" button.

TITLE	MPAA	GENRE
Add ★ 12 Angry Men (1957)	UR	Classics
Add The 39 Steps (1935)	UR	Classics
Add An American in Paris (1951)	UR	Classics
Add The Andromeda Strain (1971)	G	Sci-Fi & Fantasy
Add Apollo 13 (1995)	PG	Drama
Add ★ The Battle of Algiers (1965) La Battaglia di Algeri	UR	Foreign
Add Being There (1979)	PG	Drama
Add Big Deal on Madonna Street (1958) I soliti ignoti	UR	Foreign
Add The Birds (1963)	PG-13	Thrillers
Add Blade Runner (1982)	R	Sci-Fi & Fantasy

2002 Netflix DVDs

- Mailed 1-3 DVDs to users
- Brand based in part on ratings and recommendations
- Users could rate any movie, not just those recently watched
- Users were told recommendations get better with more ratings

Search and Discovery

- Amazon historically incorporated recommendations into search results, including to extend results
- Recommendations used at different stages of transaction (E.g. Product page, cart)

Amazon Recommendations (2003)

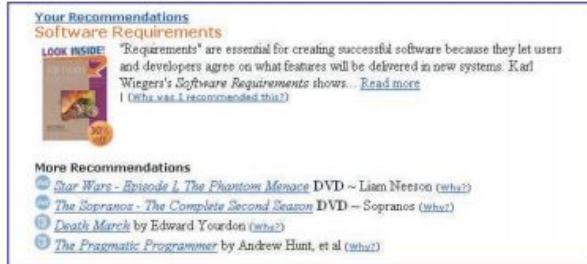


Figure 1. The "Your Recommendations" feature on the Amazon.com homepage. Using this feature, customers can sort recommendations and add their own product ratings.



Figure 2. Amazon.com shopping cart recommendations. The recommendations are based on the items in the customer's cart: The Pragmatic Programmer and Physics for Game Developers.

2003 Amazon

- Many people at time identified recommendations with Amazon
- Item-Item CF (<http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf>)
- Compares to User-Item based CF (as in GroupLens)
- Pointed to scalability: compute similarities between items offline, usually nightly. This is preferred as item similarities change less frequently than user similarities
- Use item similarities to recommend similar items to those a user has shown preference towards
- Item preference a weighted combination of views, ratings, cart adds, and purchases
- Added blended exponential decay to account for time
- Item similarities can also be used in non-personalized fashion ("users who considered this, also considered that...")
- Not fully real-time and suffers from cold-start problem
- Flawed problem setup: not directly optimizing sales

Amazon Recommendations

Item-based Collaborative Filtering Recommendation Algorithms

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl

GroupLens Research Group/Army HPC Research Center
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455
{sarwar, karypis, konstan, riedl}@cs.umn.edu

Appears in WWW10, May 1-5, 2001, Hong Kong.

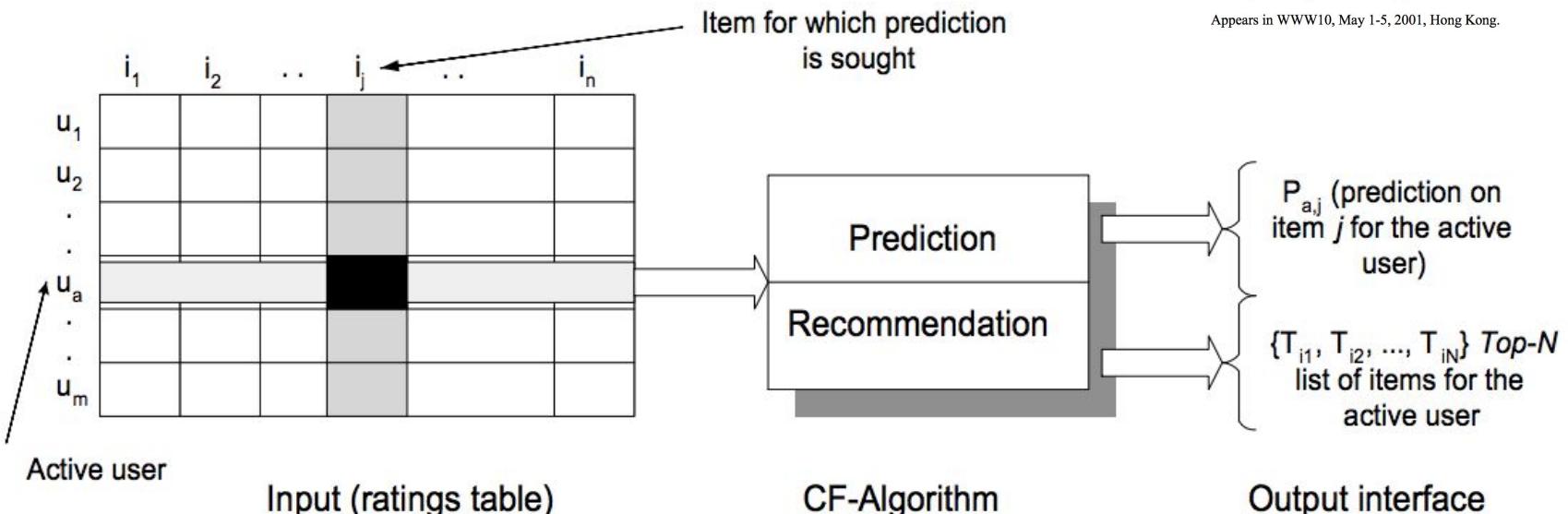


Figure 1: The Collaborative Filtering Process.

Amazon Recommendations

Item-based Collaborative Filtering Recommendation Algorithms

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl

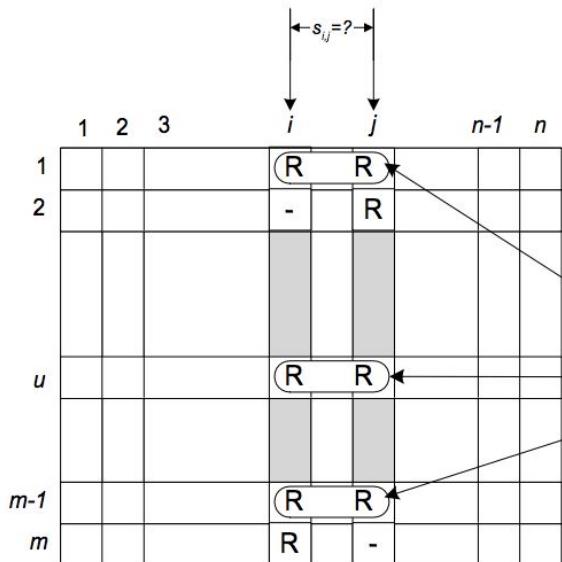
GroupLens Research Group/Army HPC Research Center

Department of Computer Science and Engineering

University of Minnesota, Minneapolis, MN 55455

{sarwar, karypis, konstan, riedl}@cs.umn.edu

pearls in WWW10, May 1-5, 2001, Hong Kong.



$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_j)^2}}.$$

Item-item similarity is computed by looking into co-rated items only. In case of items *i* and *j* the similarity s_{ij} is computed by looking into them. Note: each of these co-rated pairs are obtained from different users, in this example they come from users 1, *u* and *m*-1.

Figure 2: Isolation of the co-rated items and similarity computation

Amazon Recommendations

Item-based Collaborative Filtering Recommendation Algorithms

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl

GroupLens Research Group/Army HPC Research Center
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455
(sarwar, karypis, konstan, riedl)@cs.umn.edu

Appears in WWW10, May 1-5, 2001, Hong Kong.

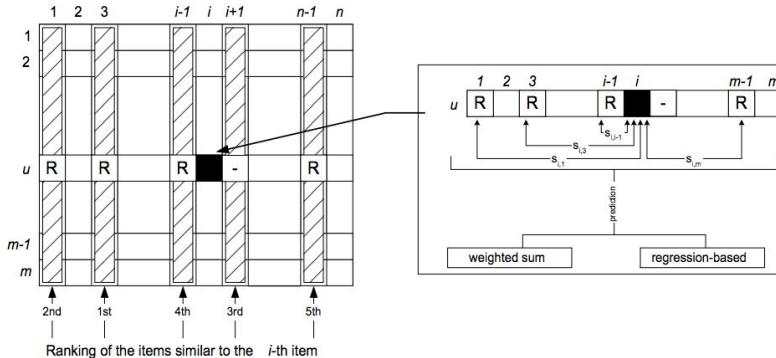


Figure 3: Item-based collaborative filtering algorithm. The prediction generation process is illustrated for 5 neighbors

the corresponding user average from each co-rated pair. Formally, the similarity between items i and j using this scheme is given by

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}}.$$

Here \bar{R}_u is the average of the u -th user's ratings.

Cold Start Problem

How do we handle new items?

How do we handle new users?

Pandora Recommendations



The Music Genome Project

Track Length (Min:Sec)	0:19
Head (Principal Melody)	
Fixed-to-Improvized [0-4]	1.0
Number of Secondary Themes [0-4]	0.0
Span Narrow-to-Wide [0-4]	4.5
Lyrical-to-Angular [0-4]	3.5
Melodic Rhythm Intensity Lo-to-Hi [0-4]	3.0
Contour Mono-to-Melismatic [0-4]	3.5
Phrase Repetitive-to-Thru [0-4]	2.5
Ornamentation [0-4]	0.0
Presentation Single-to-Ensemble [0-4]	1.0
Presentation Unison to Chordal [0-4]	1.0
Antiphony [0-4]	0.0
Counterpoint [0-4]	0.0
Harmony	
Modal [0-2]	4.0
Minor-to-Major [0-4]	1.5
Diatomic-to-Chromatic [0-4]	2.0
Overall Resonance Lo-to-Hi [0-4]	2.0
Chordal Patternning [0-5]	2.0
Chordal Rhythm Slow-to-Fast [0-4]	2.5
Pedal Point [0-5]	3.0
Fermi and Arrangement	
Multi-Sectioned [0-2]	0.0
Head-Solo-Head-to-Thru Composed [0-4]	1.0
12-bar Blues [0-2]	0.0
"Song" Form (ABA) [0-2]	2.5
Breaks [0-4]	0.2
Intro Incidental-to-Dominant [0-1-4]	1.0
Intro Faded-to-In Tempo [0-1-5]	5.0
Rhythmic Tempo, Meter, Feel, Groove	
Primary Tempo BPM [0-200]	172.0
Cut time [0-6]	0.0
Triple (3/4,3/8,9/8) [0-4]	0.0
Compound Double (6/8,12/8) [0-4]	0.0
Odd (5/4,7/4) [0-4]	5.0
Discernibility Lo-to-Hi [0-4]	5.0
Number of Shifts [0-4]	0.0
Swing or Shuffle [0-5]	5.0
Swing to Shuffle [0-1-4]	1.0
Swung Sixteenths [0-2]	0.0
Subdivisions Lo-to-Hi [0-4]	3.0
Latin [0-2]	0.0
Double-Time [0-3]	0.0
Back-beat Prominence [0-4]	0.0
Rhythmic Ostinato-Based [0-5]	4.0
Motion-Inducing Lo-to-Hi [0-5]	3.5
Rhyth. Temp. Simple-to-Complex [0-3]	3.0
Syncopation Level Low-to-High [0-4]	3.0
Rhythmic Loops (Pre-recorded) [0-4]	0.0
Head, Harmony, Form, Rhythms	1
Low, Vocal, Ensemble, Improv, A, Improv, B	2
Sax, Clarinet, Tuba, Harmonica, Trumpet	3
Trombone, Valve, Vibes, Flute, Organ, Synthesizer, Accordion, Ocarina	4
Bass, Gtr, Bass, Drums, Perc., Horn, Bass, String Inst., Acoustic Gtr, And Inst.	5
Lyric, Sound, Consistency, Durability	6
Guitar, Overall	7

This analysis is NOT approved and should be excluded from the database.



datascience@berkeley

The Netflix Prize

- Training data
 - 100,480,507 ratings
 - 480,189 users
 - 17,770 movies
 - <user, movie, date of grade, grade>
 - <integer, integer, date, 1–5>
- Test data
 - 2,817,131 ratings
 - <integer, integer, date>
- Objective: minimize RMSE:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2}$$

Most Active Users

User ID	# Ratings	Mean Rating
305344	17,651	1.90
387418	17,432	1.81
2439493	16,560	1.22
1664010	15,811	4.26
2118461	14,829	4.08
1461435	9,820	1.37
1639792	9,764	1.33
1314869	9,739	2.95

Copyright AT&T

Data about the Movies

Most Loved Movies	Avg rating	Count
The Shawshank Redemption	4.593	137812
Lord of the Rings : The Return of the King	4.545	133597
The Green Mile	4.306	180883
Lord of the Rings : The Two Towers	4.460	150678
Finding Nemo	4.415	139050
Raiders of the Lost Ark	4.504	117456

Most Rated Movies

Miss Congeniality
Independence Day
The Patriot
The Day After Tomorrow
Pretty Woman
Pirates of the Caribbean

Highest Variance

The Royal Tenenbaums
Lost In Translation
Pearl Harbor
Miss Congeniality
Napoleon Dynamite
Fahrenheit 9/11

The Netflix Prize



The Netflix Prize

SVD for Rating Prediction

- User factor vectors $p_u \in \Re^f$ and item-factors vector $q_v \in \Re^f$
- Baseline $b_{uv} = \mu + b_u + b_v$ (user & item deviation from average)
- Predict rating as $\hat{r}_{uv} = b_{uv} + p_u^T q_v$
- **SVD++** (Koren et. Al) asymmetric variation w. implicit feedback

$$\hat{r}_{uv} = b_{uv} + q_v^T \left(|R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} (r_{uj} - b_{uj}) x_j + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_j \right)$$

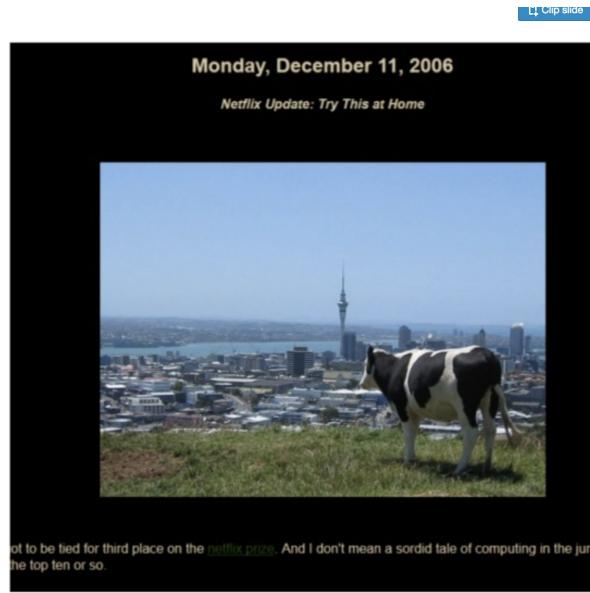
- Where
 - $q_v, x_v, y_v \in \Re^f$ are three item factor vectors
 - Users are not parametrized, but rather represented by:
 - $R(u)$: items rated by user u
 - $N(u)$: items for which the user has given implicit preference (e.g. rated vs. not rated)

The Netflix Prize

Simon Funk's SVD

- One of the most interesting findings during the Netflix Prize came out of a blog post
- Incremental, iterative, and approximate way to compute the SVD using gradient descent

NETFLIX



Clustering

- Item-item CF didn't work well
- <http://gigapan.com/gigapans/65469>

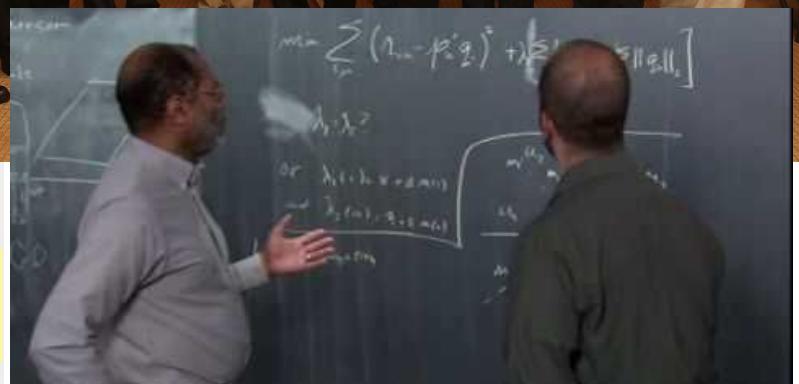
Matrix Factorization Approaches

- 'Simon Funk' released his matrix factorization solution, which supported data sparsity. Used in Netflix (for predicted rating) today
- Dim reduction on sparse data using SVD
- <http://sifter.org/~simon/journal/20061211.html>
- Later approaches added better formalization and incremental improvements

The Netflix Prize

What Happened

- October 2, 2006: contest launched—\$1 million first prize
 - Cinematch RMSE = 0.9514
 - Naïve "average rating" RMSE = 1.0540
- October 8, 2006: "WXYZ" team beats Cinematch
- June 2007: 20,000 teams in competition (150 countries)
- September 2007: "BellKor" RMSE = 0.8728 (\$50k)
- September 2008: "BellKor" RMSE = 0.8616 (\$50k)
- July 2009: two teams hit 10% margin—no more entries
- September 2009: "BellKor's Pragmatic Chaos" wins \$1 million with RMSE of 0.8567. Same RMSE matched by "The Ensemble," but submission made 20 minutes later. Tiebreaker went to BellKor's Pragmatic Chaos



The Netflix Prize

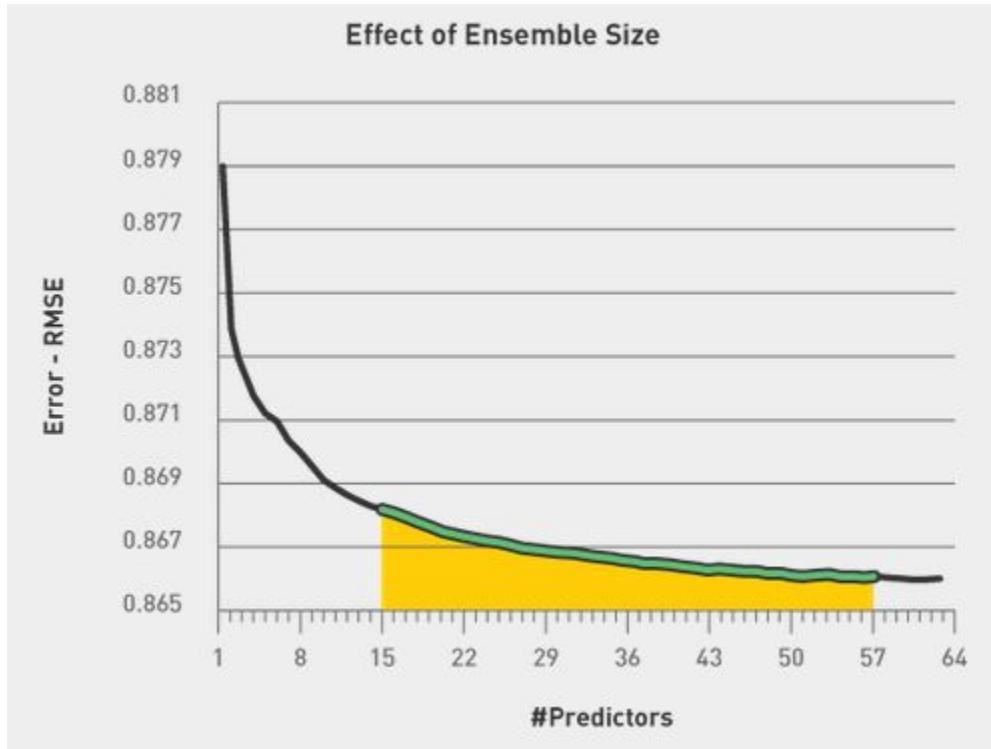
Netflix: The Winner (contd)

- Final product used ensemble methods, which combined results from several algorithms.
 - Neighborhood methods
 - Matrix decomposition methods
 - Regression
 - Boltzmann machines
 - *Et cetera*
- Combining results is topic of another lecture!
- Gradient-boosted decision trees combine 500+ models.

What Other Factors Matter?

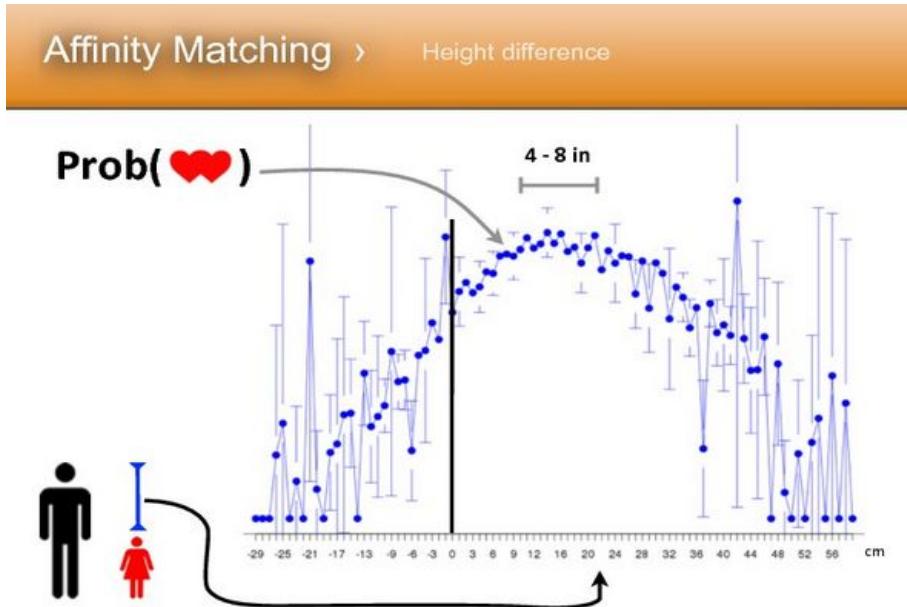
- The number of other people who have rated the movie
- The (square root of the) number of days since first rating
- The number of days since the movie's first rating
- Plus a bunch of others!

The Netflix Prize



Dating Recommendations - How would you do it?

EHarmony



EHarmony

- Straightforward supervised learning approach to building a recommender system.
- Focus is on feature engineering
- 320 attributes (e.g. romantic, height, photo zoom, food preferences)
- Predict likelihood of 'successful match' using logistic regression (Vowpal Wabbit, GBM)
- For more info
<https://ieondemand.com/divisions/big-data/presentations/data-science-of-love-1>

Job Recommendation - How would you do it?

The Recommender Ecosystem

Jobs You May Be Interested In



Talent Match



CAP



Similar Profiles



Companies

Recommendations, similar companies search, peer companies, and company browse maps, company products and services browse maps



Related search



Network updates



Profile browse maps



Jobs browse maps



Ad matching engine

$\mu\text{CTR} = \beta(\text{member}, \text{creative}, \text{advertiser}, \text{context}, \text{inventory}, \text{DCTR})$

Connections



Events You May Be Interested In



Groups

Recommendations, similar groups search



Similar jobs



Referral Engine



News



News Recommendation - How would you do it?

News Recommendation



Prismatic (Acquired by MS in 2015)

- Apple news is almost identical
- Problem: Provide personal news feed
- Popular on mobile devices as way to kill time (e.g. on bus)
- Mobile usage impacts implicit data

ML Approach

- Doesn't use Netflix Prize style collaborative filtering
- Step 1: Featurization using Topic Modeling (25K+ topics)
- Step 2: Directly optimize likelihood to read using logistic regression

Netflix Today

Goodbye



& Cinematch

Hello



&

% Match

Why?

+200% ratings volume

Clear link to personalization

Music Recommendation

Duo Mix

Duo Mix is a playlist of songs that combines the music you and the other member of your Duo plan listens to.

It comes preloaded with songs based on genres, artists, and songs that you both have played on your own and updates the more you play.

Tip: Don't have Premium Duo? Check out [Daily Mix](#) instead.



Netflix Today

House of Cards

★★★★★ 2013 TV-MA 1 Season HD 5.1

Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

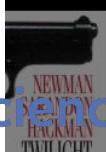


NETFLIX

Because you watched Orange Is the New Black



Because you watched Red Lights

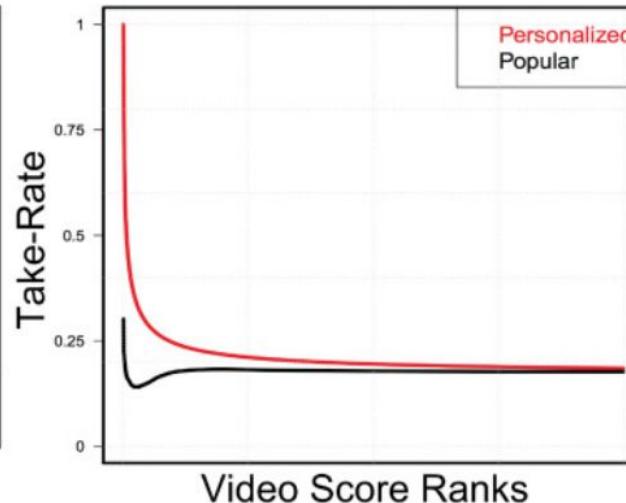
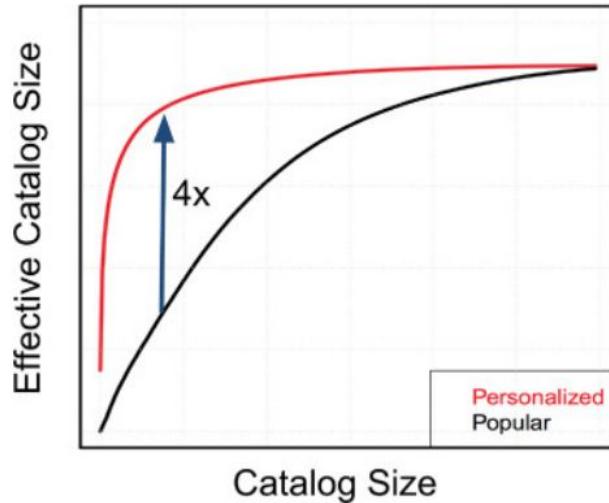


datascience@berkeley

Netflix Today



Netflix Today



The Netflix Recommender System: Algorithms, Business Value, and Innovation

CARLOS A. GOMEZ-URIBE and NEIL HUNT, Netflix, Inc.

Netflix Today



Netflix Today

Ranking

- Ranking = **Scoring + Sorting + Filtering** bags of movies for presentation to a user
 - **Goal:** Find the best possible ordering of a set of *videos* for a *user* within a specific *context* in real-time
 - **Objective:** maximize consumption
 - **Aspirations:** Played & “enjoyed” titles have best score
 - Akin to CTR forecast for ads/search results
- **Factors**
 - Accuracy
 - Novelty
 - Diversity
 - Freshness
 - Scalability
 - ...

Netflix Today

Ranking

- Popularity is the obvious baseline
- Ratings prediction is a clear secondary data input that allows for personalization
- We have added many other features (and tried many more that have not proved useful)
- What about the weights?
 - Based on A/B testing
 - Machine-learned

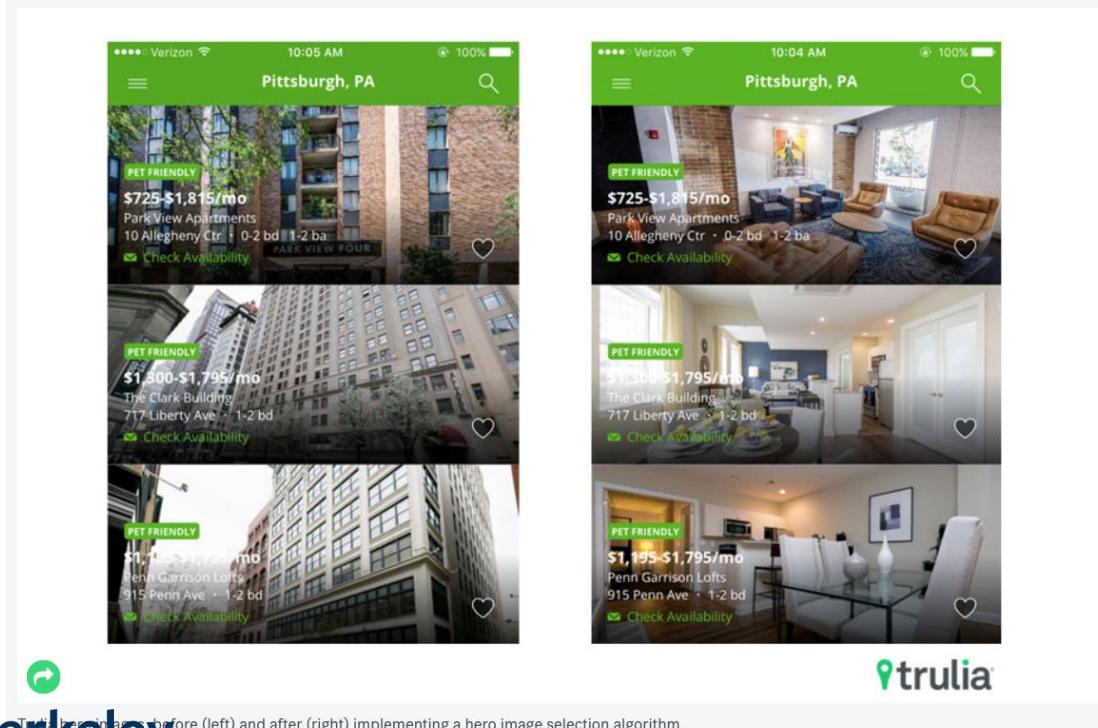
Netflix Today

Learning to Rank Approaches

1. Pointwise
 - Ranking function minimizes loss function defined on individual relevance judgment
 - Ranking score based on regression or classification
 - Ordinal regression, Logistic regression, SVM, GBDT, ...
2. Pairwise
 - Loss function is defined on pair-wise preferences
 - Goal: minimize number of inversions in ranking
 - Ranking problem is then transformed into the binary classification problem
 - RankSVM, RankBoost, RankNet, FRank...

Learned Presentation

What Makes a Photo Click: Selecting Hero Images with Deep Learning



To the left: images before (left) and after (right) implementing a hero image selection algorithm.

Learned Presentation

Using Deep Learning to automatically rank millions of hotel images

At idealo.de we trained two Deep Neural Networks to assess the aesthetic and technical quality of images 😊😊😊



6.52 (1.44)



5.58 (1.37)



5.53 (1.42)



5.04 (1.35)



4.92 (1.48)



4.29 (1.45)

Aesthetic MobileNet predictions



8.04 (2.11)



4.61 (2.75)



1.92 (1.53)



5.73 (2.85)



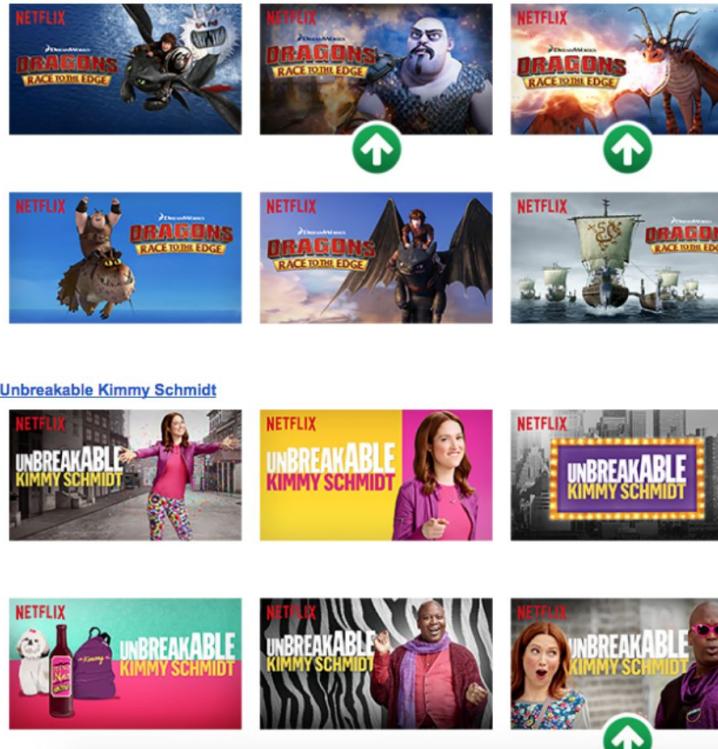
4.31 (2.7)



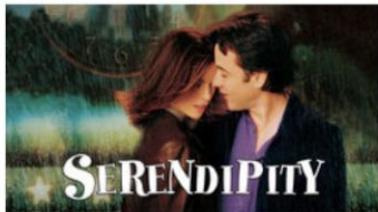
4.22 (2.78)

Personalized Presentation

Here are some screenshots from the tool that we use to track relative artwork performance.
Dragons: Race to the Edge: the two marked images below significantly outperformed all others.



Personalized Presentation



Personalized Presentation



stacia i. brown
@slb79

Follow

Other Black @netflix users: does your queue do this? Generate posters with the Black cast members on them to try to compel you to watch? This film stars Kristen Bell/Kelsey Grammer and these actors had maaaaybe a 10 cumulative minutes of screen time. 20 lines between them, tops.

NETFLIX
LIKE FATHER

2018 TV-MA 1h 43m

▶ PLAY

After she's left at the altar, a workaholic advertising executive ends up on her Caribbean honeymoon cruise with her estranged father.

Cast: Kristen Bell, Kelsey Grammer, Seth Rogen

9:22 PM - 17 Oct 2018

428 Retweets 793 Likes

140 428 793



stacia i. brown @slb79 · Oct 18, 2018



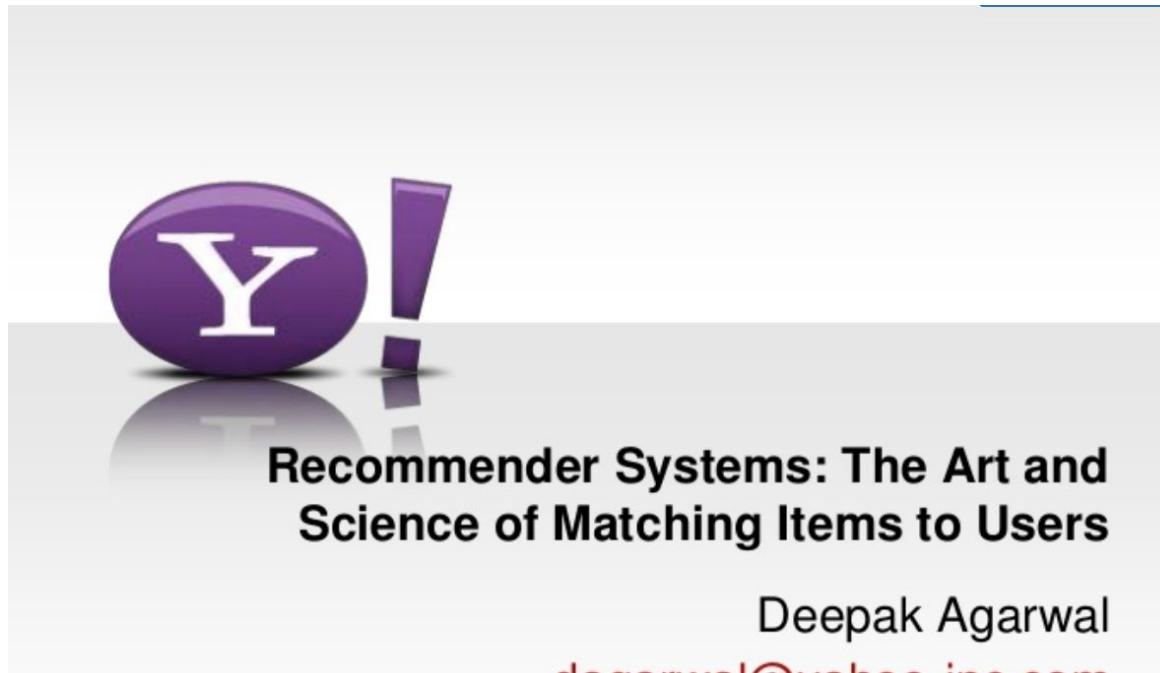
UCLA Week 6 – Google Slides
Replying to @slb79

Just did another cursory scroll of suggested watches and the posters they gave me.



"We don't ask members for their race, gender or ethnicity so we cannot use this information to personalize their individual Netflix experience. The only information we use is a member's viewing history. In terms of thumbnails, these do differ and regularly change. This is to ensure that the images we show people are useful in deciding which shows to watch."

Bandits



Bandits

Clip slide

Content Optimization: Match articles to users

My Yahoo | Make Y! your homepage

Web Images Video Local Shopping More

YAHOO!

Sign In | New here? Sign Up | Have something to share? | Page Options

TODAY - July 14, 2010

World Cup octopus could make millions

The octopus is in high demand after a perfect run of predicting soccer game winners. > Possible opportunity

Salsa heat to food illness Octopus could be worth millions Lottery winner rich in mystery High soccer's impressive dunk

9 - 8 of 28

NEWS WORLD LOCAL FINANCE

9 killed, 10 missing as typhoon lashes Philippines | Photos Testing delayed on tighter cap for Gulf oil well | Photos W-Va. mine disaster prompts bill to toughen worker safety rules Military won't establish 'separate but equal' housing for gays Small banks struggling despite gov't bailouts, watchdog reports Tiny mushroom blamed for 400 deaths in southwest China CHP pursuit ends in two-car crash in San... - S.J. Mercury News Oakland talks break down; layoffs for 80... - S.F. Chron... Stanford grad student dies in Yosemite... - Mountain View... NBA - NHL - MLB - Tennis - Golf - Soccer - NASCAR

updated 01:49 am More News Popular Buzz

TRENDING NOW

1. Kourtney Kardashian
2. Anna Chapman
3. Al Pacino
4. French Toast Rec...
5. Nina Garcia
6. Susan Boyle
7. Job Search
8. Yogi Berri
9. Philippines Typh...
10. Sunscreen

Anything you want, you got it with Ultimate Rewards.

• YOUR CHOICE OF REWARDS • NO FLIGHT RESTRICTIONS • UNIQUE EXPERIENCES

Click to see <your reward>

UR ultimate rewards. CHASE

Go To UltimateRewards.com

Ultimate Rewards - All Feedback

Must-see music news & features

Britney lays down the law with kids Lady Gaga photo & Beatles fans

'Idol' runner-up gets cosmetic work Kelle Pickler's new retro '40s video

1 of 4

DAILY OFFERS

Mortgage rates low as 3.32% APR

data science@berkeley

Neek Agarwal @LinkedIn '11

Bandits

Back to Yahoo! front page

The screenshot shows the Yahoo! homepage from July 14, 2010. A red box highlights the main news area. Inside this box, four numbered arrows point to specific elements: 1 points to the headline "World Cup octopus could make millions"; 2 points to the image of the octopus; 3 points to the text "Octopus could be worth millions"; and 4 points to the text "High schooler's impressive dunk". To the right of this highlighted area, there is a sidebar titled "TRENDING NOW" with a list of nine items, and a section titled "Recommend articles:" with three sub-points: "Image", "Title, summary", and "Links to other pages". Below the main news area, there is a news ticker and a "NEWS" tab. At the bottom, there are sections for "WORLD", "LOCAL", and "FINANCE". The footer features a "M FAVORITES" section and a "Sign In" button.

YAHOO!

My Yahoo! | Make Y! your homepage

YAHOO! SITES Edit

1 TODAY - July 14, 2010

2 3 4

World Cup octopus could make millions

Paul the octopus is in high demand after a perfect run of predicting soccer game winners! » Possible opportunities

Salsa tied to food illness Octopus could be worth millions Lottery winner rich in mystery High schooler's impressive dunk

5 - 8 of 28

NEWS WORLD LOCAL FINANCE

* 9 killed, 10 missing as typhoon lashes Philippines | Photos
* Testing delayed on tighter cap for Gulf oil well | Photos
* W.Va. mine disaster prompts bill to toughen worker safety rules
* Military won't establish 'separate but equal' housing for gays
* Small banks struggling despite govt bailouts, watchdog reports
* Tiny mushroom blamed for 400 deaths in southwest China
* CHP pursuit ends in two-car crash in San... - SJ Mercury N...
* Oakland talks break down; layoffs for 80... - S.F. Chronic...

TRENDING NOW

1. Kourtney Kardash...
2. Anna Chapman
3. Al Pacino
4. French Toast Rec...
5. Susan Boyle
6. Job Search
7. Yogi Berra
8. Philippines Typh...
9. Philippines Typh...

Recommend articles:

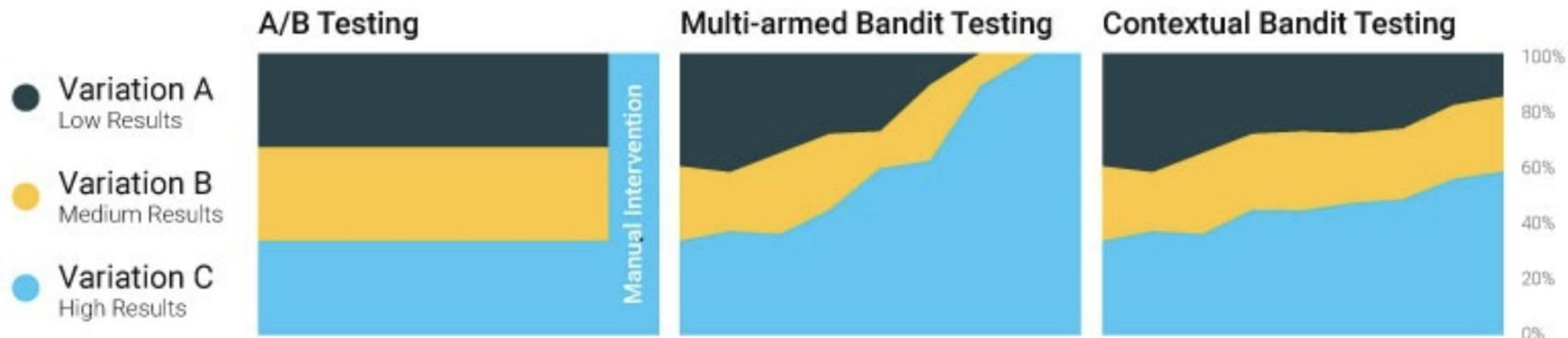
- Image
- Title, summary
- Links to other pages

For each user visit,
Pick 4 out of a pool of K

Routes traffic to other pages

47

Bandits



Bandits

Epsilon-Greedy Example



	1	2	3	4	5	6	7	8	9
1	1		0			1		0	
2		0			0			0	
3			0	1			0		

Observed
Reward

2/4
(greedy)

0/3

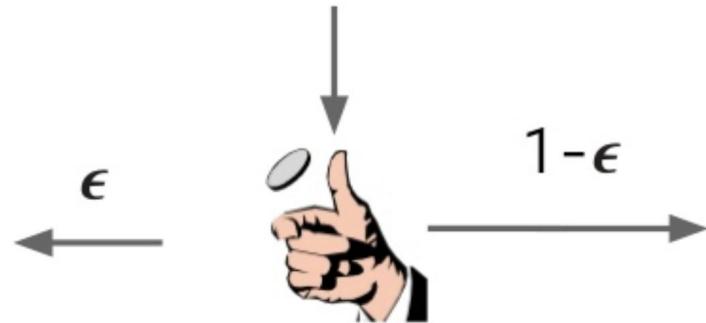
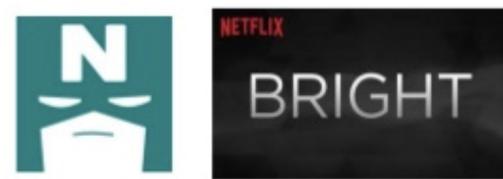
1/3

Bandits



- **Context:** Member, device, page, etc.

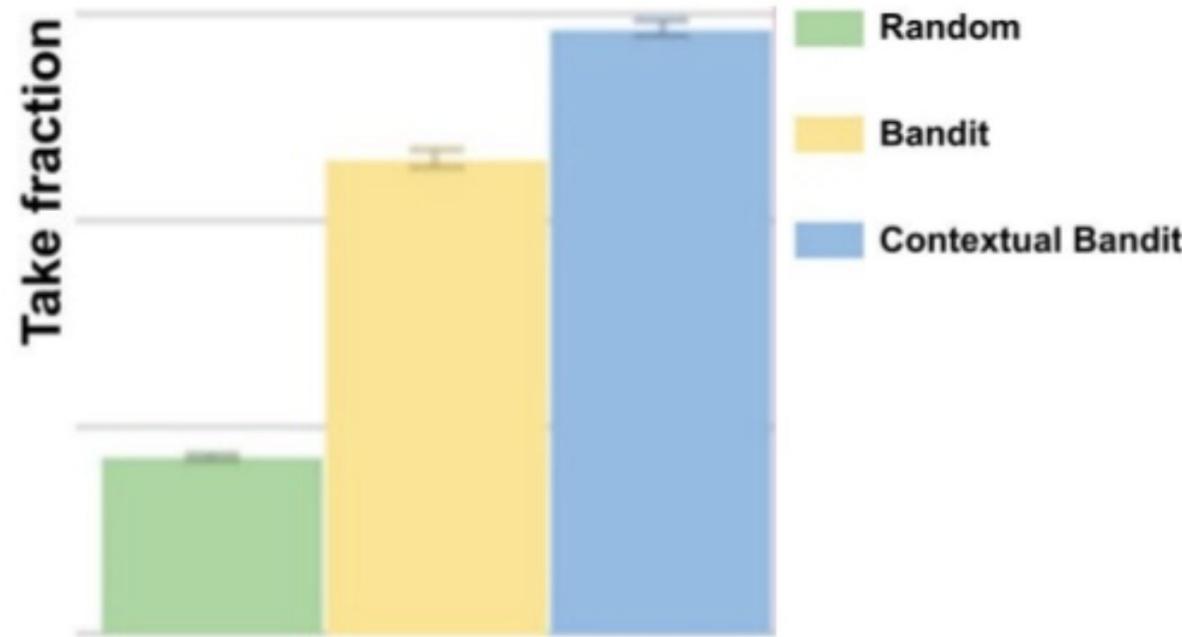
Bandits



Personalized
Image



Bandits



Frontier: Deep Latent Factors

Variational Autoencoders

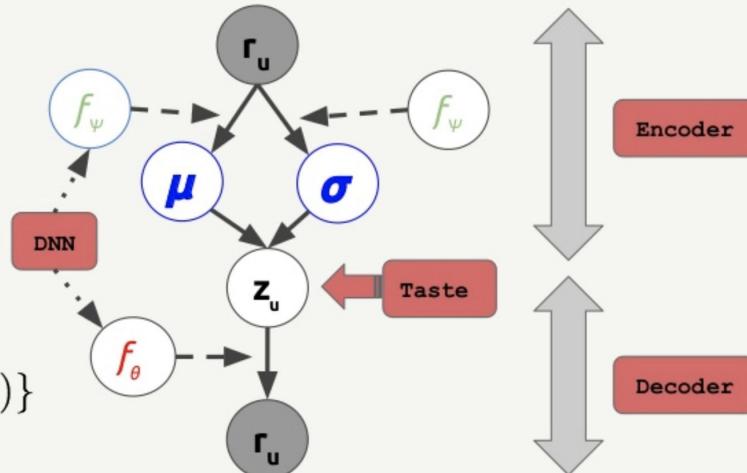
Inference model:

$$q_\psi(\mathbf{z}_u \mid \mathbf{r}_u) = \mathcal{N}(\mu_\psi(\mathbf{r}_u), \text{diag}\{\sigma_\psi^2(\mathbf{r}_u)\})$$

Generative model:

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathbf{I}_K), \pi(\mathbf{z}_u) \propto \exp\{f_\theta(\mathbf{z}_u)\}$$

$$\mathbf{r}_u \sim \text{Multi}(N_u, \pi(\mathbf{z}_u))$$



Liang et al. (2018), Variational Autoencoders for Collaborative Filtering, WWW.

Frontier: Whole Page Optimization

Beyond Ranking: Optimizing Whole-Page Presentation

Yue Wang^{1*}, Dawei Yin², Luo Jie³, Pengyuan Wang², Makoto Yamada^{2,4},
Yi Chang², Qiaozhu Mei^{1,5}

¹Department of EECS, University of Michigan, Ann Arbor, MI, USA

²Yahoo Labs, 701 First Avenue, Sunnyvale, CA, USA

³Snapchat, Inc., 64 Market St, Venice, CA, USA

⁴Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

⁵School of Information, University of Michigan, Ann Arbor, MI, USA





LATEST NEWS

July 26, 2018: Tentative lists of accepted [short](#) and [long](#) papers for RecSys 2018 are online!

July 12, 2018: This year's RecSys conference will feature a record number of six [tutorials](#), taking place on Oct 2, prior to the main conference.

July 10, 2018: Acceptance notifications have been sent out: It was very competitive this year, with 18% acceptance rate for long and 25% for short papers.

July 4, 2018: [Registration](#) for RecSys 2018 is open!

June 12, 2018: The call for [late-breaking results \(posters\)](#) is out now, and the call for [demos](#) has been updated!

June 05, 2018: RecSys 2018 will feature three exciting [keynotes](#) by Elizabeth F. Churchill (Google), Lise Getoor (UCSC), and Christopher Berry (Canadian Broadcasting Corporation)!

May 30, 2018: The tutorial submission deadline has been extended to next Monday, June 4, 2018!

SHORTCUTS TO CONFERENCES

- [RecSys 2018 \(Vancouver\)](#)
- [RecSys 2017 \(Como\)](#)
- [RecSys 2016 \(Boston\)](#)
- [RecSys 2015 \(Vienna\)](#)
- [RecSys 2014 \(Silicon Valley\)](#)
- [RecSys 2013 \(Hong Kong\)](#)
- [RecSys 2012 \(Dublin\)](#)
- [RecSys 2011 \(Chicago\)](#)
- [RecSys 2010 \(Barcelona\)](#)
- [RecSys 2009 \(New York\)](#)
- [RecSys 2008 \(Lausanne\)](#)
- [RecSys 2007 \(Minnesota\)](#)

Final Thoughts?