# Statistical Machine Learning: Collaborative Filtering

Daniel J. McDonald and Darren Homrighausen

Indiana University, Bloomington
and Colorado State University

mypage.iu.edu/∼dajmcdon

stat.colostate.edu/∼darrenho

April 2-3, 2013

# Collaborative filtering:
# How do I get you to buy lots of stuff you don't need?

## Recommendations for You in Sports & Outdoors

SIGG Cleaning Brush with Red Bristles
(21)
~~$11.39~~ $9.16
Why recommended?

UCO Sigg Bottle Clip Cap
(13)
$6.68
Why recommended?


Sigg Lifestyle Loop Top Water Bottle
(160)
$13.82 - $24.99
Why recommended?

› See more recommendations

## Recommendations for You in Kindle Store

The Count of Monte Cristo (annotated)
› Alexander Dumas
Kindle Edition
(122)
$0.99
Why recommended?


The Complete Works of Shakespeare
› William Shakespeare
Kindle Edition
(82)
$1.99
Why recommended?


THE HUNCHBACK OF NOTRE DAME
› Victor Hugo, Thomas Leclere, Vincent Leroger
Kindle Edition
(30)
$0.99
Why recommended?

› See more recommendations

# Recommender systems

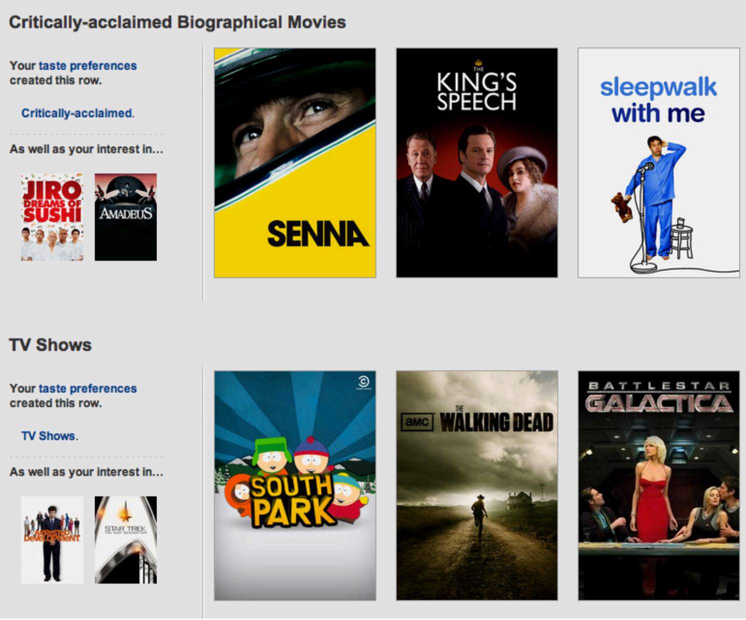
- I have lots of data on people who bought some stuff
- I want to get YOU to buy lots more stuff
- How do I figure out how to show you things you might want?


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.
FAQ | Forum | Netflix Home
© 1997-2009 Netflix, Inc. All rights reserved.

# References and sources

- Yehuda Koren, Robert Bell and Chris Volinsky (members of BellKor)
- Koren (2009), "The BellKor Solution to the Netflix Grand Prize."
- Koren and Bell (2009), "Advances in Collaborative Filtering."
- www.netflixprize.com
- Yifan Hu
- www.timelydevelopment.com

# THE NETFLIX PRIZE



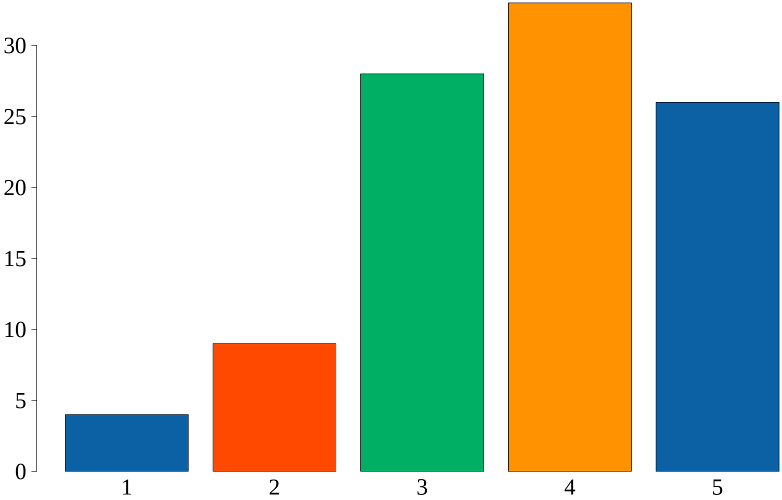
# Rating data

| Training data | | | Test data | | |
|---|---|---|---|---|---|
| User | Movie | Rating | User | Movie | Rating |
| 1 | 234 | 2 | 1 | 15 | ? |
| 1 | 849 | 5 | 2 | 27 | ? |
| 1 | 738 | 4 | 2 | 738 | ? |
| 2 | 383 | 3 | 2 | 215 | ? |
| 2 | 782 | 5 | 3 | 782 | ? |
| 3 | 614 | 1 | 3 | 2 | ? |
| | ⋮ | | | ⋮ | |

# The data

- Training data (∼8 GB publicly available)
  - 100 million ratings
  - 480,000 unique users
  - 17,700 movies
  - 6 years of data
- Test set
  - Last few ratings of each user (2.8 million)
  - Split into 2 groups (quiz, test)
  - Netflix Cinematch: RMSE 0.9514
  - $1 million prize to reduce error 10%
- Most of the "data" is missing: 99% of movie/user combinations are empty

# Rating distribution




Mean rating is 3.68.

# How did they win?

1. Baseline estimation
2. Latent factor models
3. $k$-nearest neighbors
4. Lots of tricks to get the last 1%

# Baseline estimation

# Baseline estimation

Average rating: 3.7

Joe's average rating: 3.5

Average rating for Men in Black: 4.1

Baseline: Joe would rate MIB $3.7 - 0.2 + 0.4 = 3.9$


# Baseline estimation

Call $r_{ui}$ the rating of the $u^{th}$ user on the $i^{th}$ movie.

Call the overall mean $\mu$, the baseline for user $u$, $b_u$ and the baseline for movie $i$, $b_i$.

Estimate $b_u$ and $b_i$ for all users $u$ and movies $i$ by solving

$$\min_{\mu,\mathbf{b}} \sum_{u,i} (r_{ui} - \mu - b_u - b_i)^2 + \lambda \left( \sum_u b_u^2 + \sum_i b_i^2 \right)$$

Lots of parameters. Avoid overfitting. Regularize!

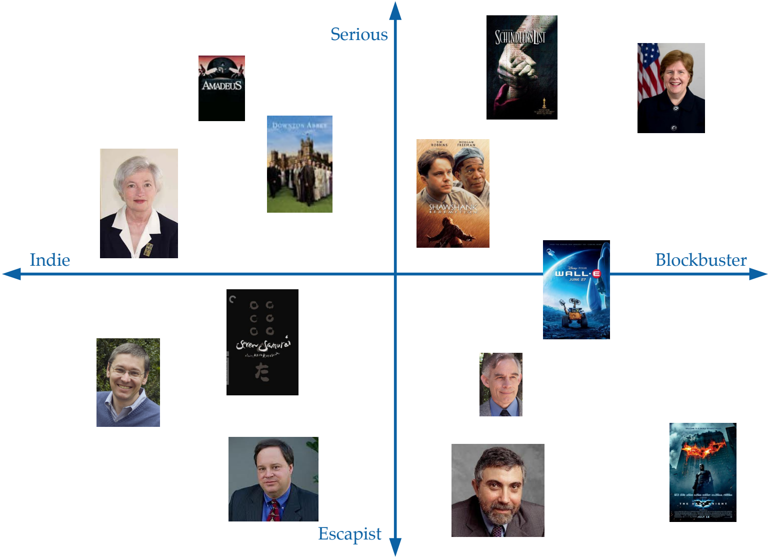
# Latent factors


Men in Black placed high on the "Alien slapstick" scale, but Joe was much lower

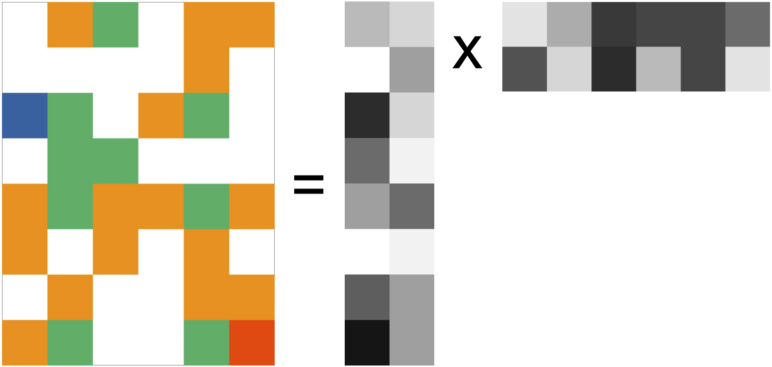
Adjust rating: $3.9 \to 3.6$

# Factor analysis

Decompose users and movies into user factors and movie factors



# Matrix factorization




Decompose the ratings matrix into user factors times movie factors.

Here a rank-2 decomposition. Simpler than a full-rank matrix.

Singular value decomposition undefined, impractical.

# More regularization

Modify the baseline method to incorporate low-dimensional factors. Say $d$-dimensional. ($d = 2$ on previous slide)

$q_i \in \mathbb{R}^d$ represents movie-specific attributes

$p_u \in \mathbb{R}^d$ represents user-specific attributes

$$\min_{\mathbf{p},\mathbf{q},\mu,\mathbf{b}} \sum_{u,i} (r_{ui} - \mu - b_u - b_i - q_i^\top p_u)^2 +$$
$$+ \lambda \left( \sum_u \left( b_u^2 + ||p_u||_2^2 \right) + \sum_i \left( b_i^2 + ||q_i||_2^2 \right) \right)$$

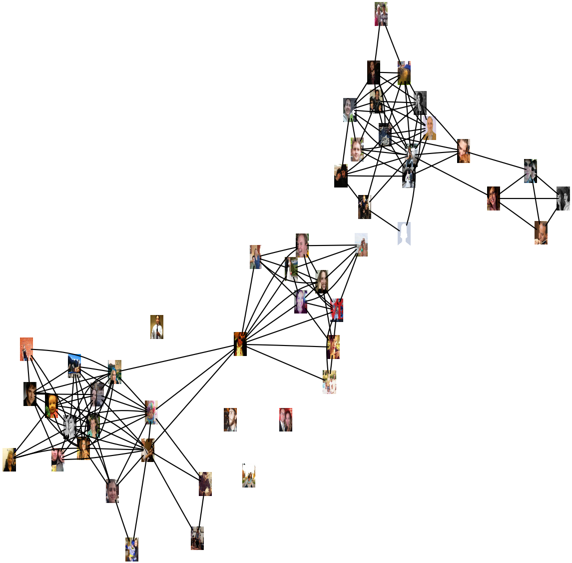
Can solve this optimization problem quickly in parallel with alternating least-squares

# Neighborhood methods

# $k$-NEAREST NEIGHBORS

- Very intuitive
- Common technique in nonparametric regression/classification
- Find some users like you (liked similar movies, demographics, etc.)
- Average their ratings to predict your rating

# $k$-NEAREST NEIGHBORS



# *k*-NEAREST NEIGHBORS

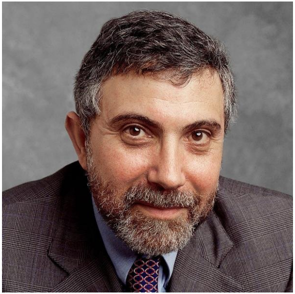
- Works on movies too (same director, lead actors, etc.)

# NEAREST NEIGHBORS

Joe and Paul are similar

Paul rated Men in Black 5 stars.

Adjust rating: $3.6 \rightarrow 4.1$



# $k$-NEAREST NEIGHBORS

Modify the method to incorporate neighborhood information.

$n_i$ is the average rating of similar users

$n_u$ is the average of similar movies

Here the tuning parameter is $k$, the number of neighbors you average over. Smaller $k$ is high variance but low bias.

Neighborhood methods are sensitive to the measure of distance.

# Making predictions better so as to win $$$

# Lessons from Netflix

- Winning entry combined these three methods with some really ingenious other ideas
- Overfitting is a tremendous disaster. Out of sample performance will be very poor
- Regularization is key
- The final model blended over 100 different individual models and totaled billions of parameters

*Predictive accuracy is substantially improved when blending multiple predictors. Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a single technique.*

*– Bell, Koren, Volinsky (2007)*