Introduction to Data Mining

CS 101, Spring 2013

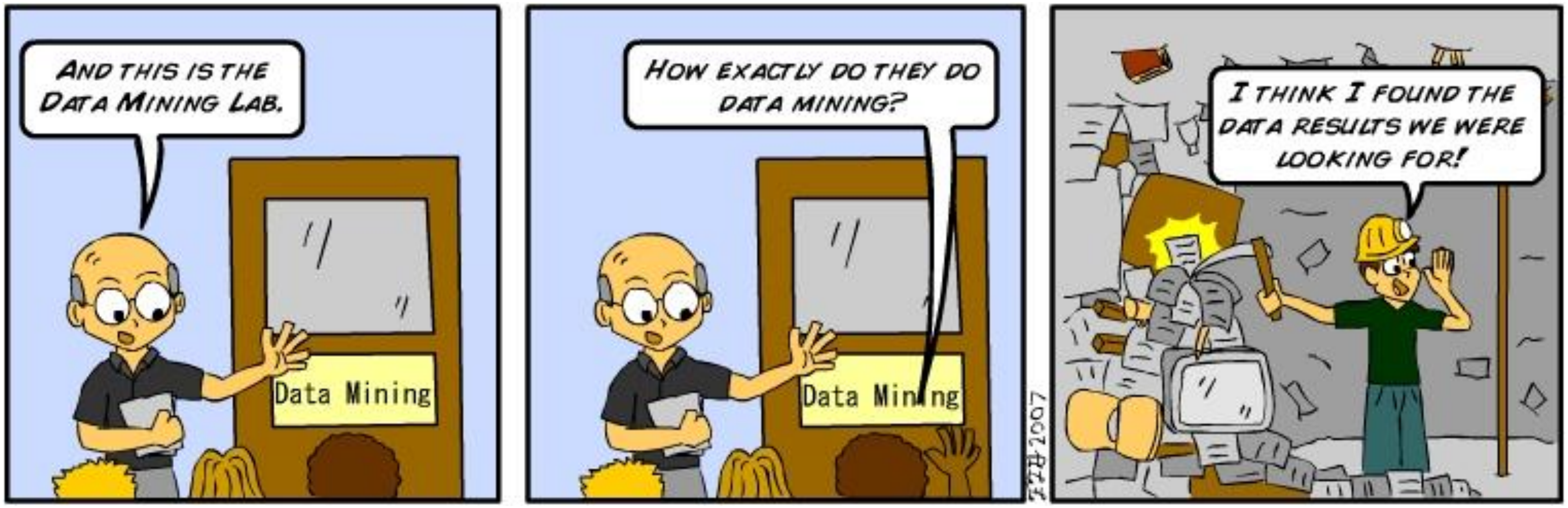
Huzefa Rangwala
Assistant Professor,
Computer Science
George Mason University
Email: rangwala@cs.gmu.edu
Website: www.cs.gmu.edu/~hrangwal

Slides are adapted from the available book slides developed by Tan, Steinbach and Kumar

Roadmap for Today

- Welcome & Introduction
- Introduction to Data Mining
  - Examples, Motivation, Definition, Methods
- A million $ competition
  - Recommender Systems

What do you think of data mining?



- Please could you write down examples that you know of or have heard of on the provided index card.

- Also write down your own definition.

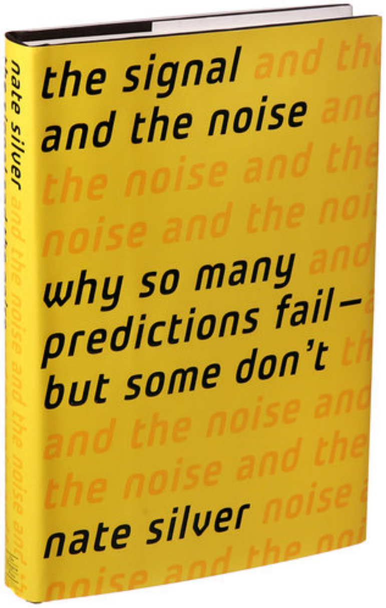
Election 2012 Data Mining

Inside the Secret World of the Data Crunchers Who Helped Obama Win

Read more:
http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/#ixzz2luhEmNcB

Mining Truth From Data Babel --- Nate Silver

http://www.nytimes.com/2012/10/24/books/nate-silvers-signal-and-the-noise-examines-predictions.html?_r=0



Data Deluge

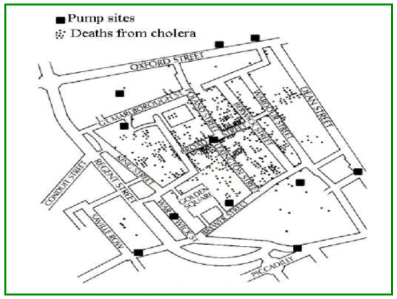
http://www.economist.com/node/15579717

Large-scale Data is Everywhere!

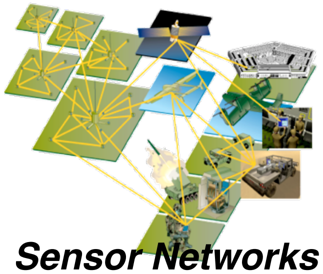
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Homeland Security

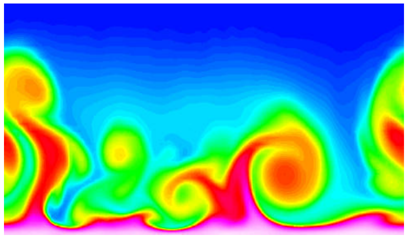
Business Data



Geo-spatial data



Sensor Networks

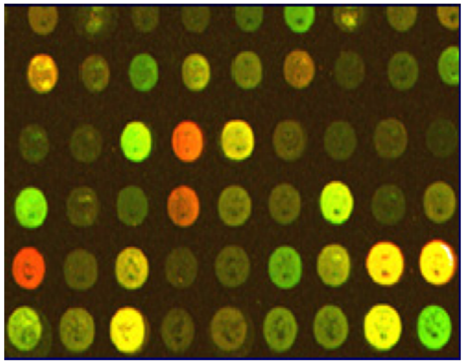


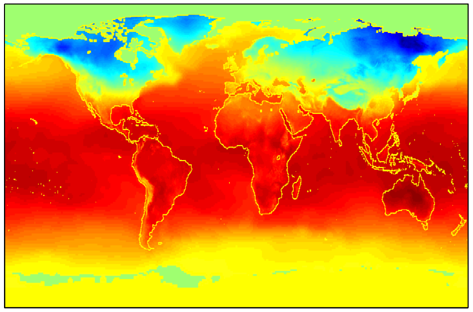
Computational Simulations

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has 2PB web data
    - Facebook has 400M active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon records 2M items/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint





- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over 1-petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

Mining Scientific Data - Fields

- Past decade has seen a huge growth of interest in mining data in a variety of scientific domains

- **Astroinformatics**
- **Neuroinformatics**
- **Quantum Informatics**
- **Health Informatics**
- **Evolutionary Informatics**
- **Veterinary Informatics**
- **Organizational Informatics**
- **Pharmacy Informatics**

- Social Informatics
- Ecoinformatics
- Geoinformatics
- Chemo Informatics

My Favorite Data Mining Examples

- Amazon.com, Google, Netflix
  - Personal Recommendations.
  - Profile-based advertisements.
- Spam Filters/Priority Inbox
  - Keep those efforts to pay us millions of dollars at bay.
- Scientific Discovery
  - Grouping patterns in sky.
  - Inferring complex life science processes.
  - Forecasting weather.
- Security
  - Phone Conversations, Network Traffic

Data Mining Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data (normally large databases)
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
- Part of the Knowledge Discovery in Databases Process.

What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

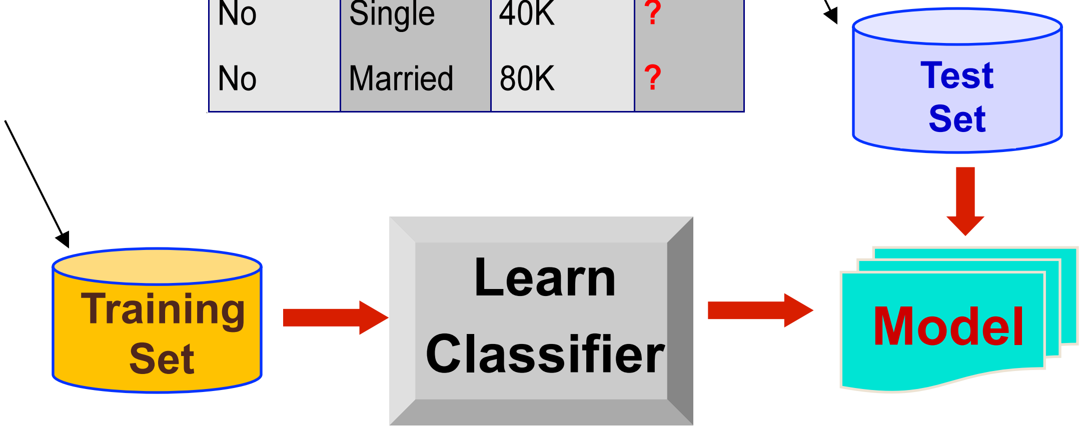
- What is Data Mining
  - Certain names are more prevalent in certain US locations (O' Brien, O' Rurke, O' Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,

Classification Example

categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



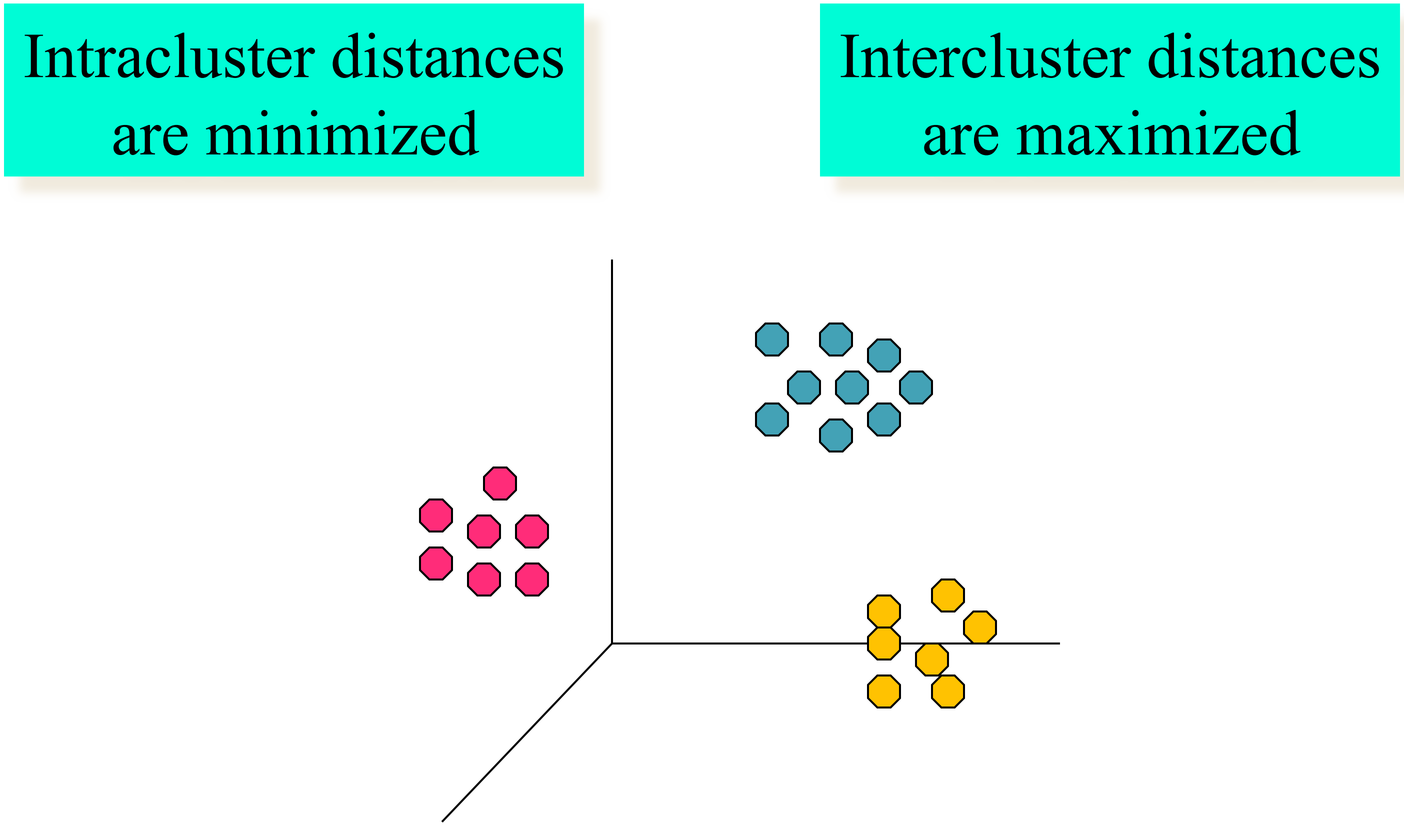
Test Set

Training Set

Learn Classifier

Model

Illustrating Clustering


Intracluster distances are minimized

Intercluster distances are maximized

- Euclidean Distance Based Clustering in 3-D space.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	***Total Articles***	***Correctly Placed***
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

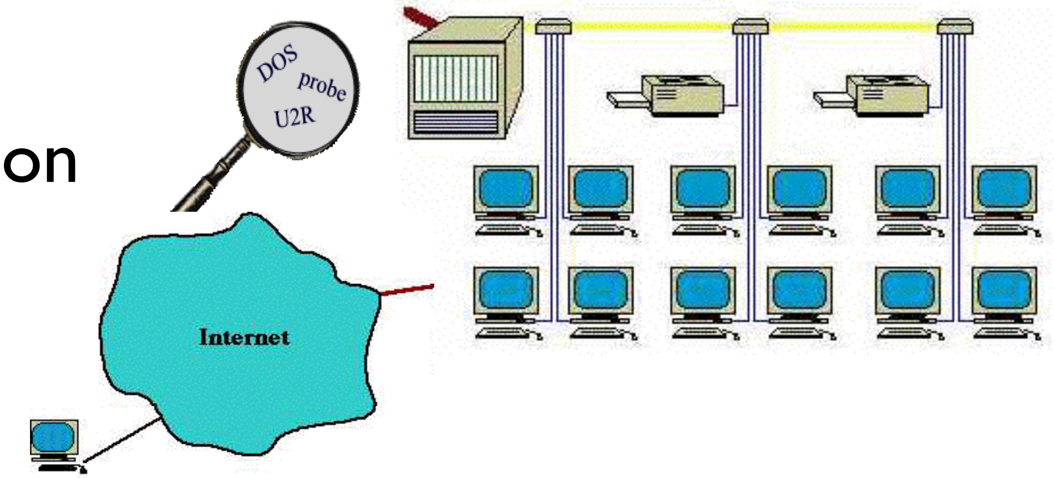
{Milk} --> {Coke}

Urban Legend ….

- Classic Association Rule Example:
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - Any plausible explanations ? ☺

Deviation/Anomaly Detection

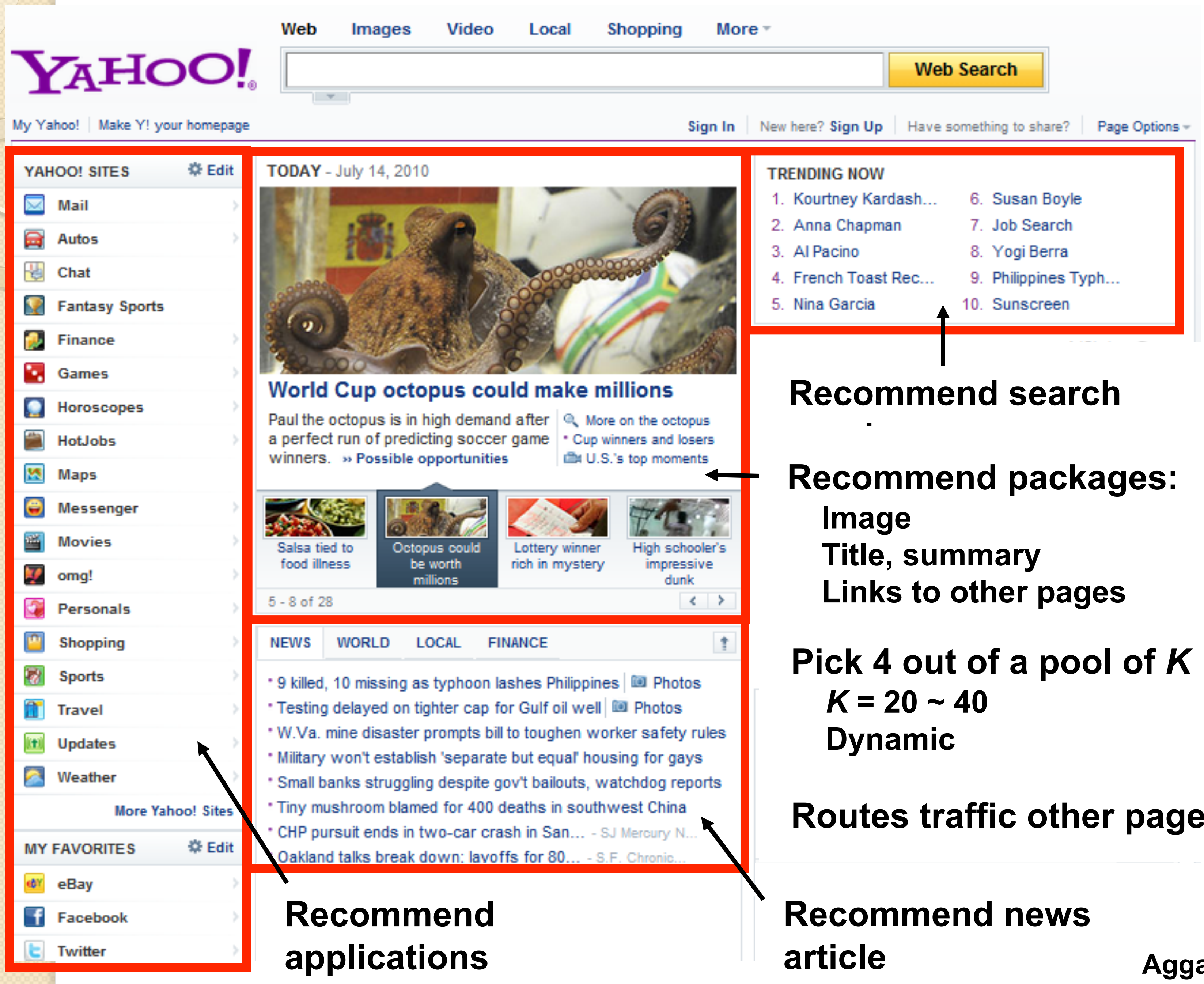
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



What else can Data Mining do ?

Dilbert

Recommender systems



TODAY - July 14, 2010

World Cup octopus could make millions

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

- More on the octopus
- Cup winners and losers
- U.S.'s top moments

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

5 - 8 of 28

TRENDING NOW

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

- 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... - SJ Mercury N...
- Oakland talks break down: layoffs for 80... - S.F. Chronic...

Recommend search

Recommend packages:
Image
Title, summary
Links to other pages

Pick 4 out of a pool of *K*
***K* = 20 ~ 40**
Dynamic

Routes traffic other pages

Recommend applications

Recommend news article

Aggarwal

Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

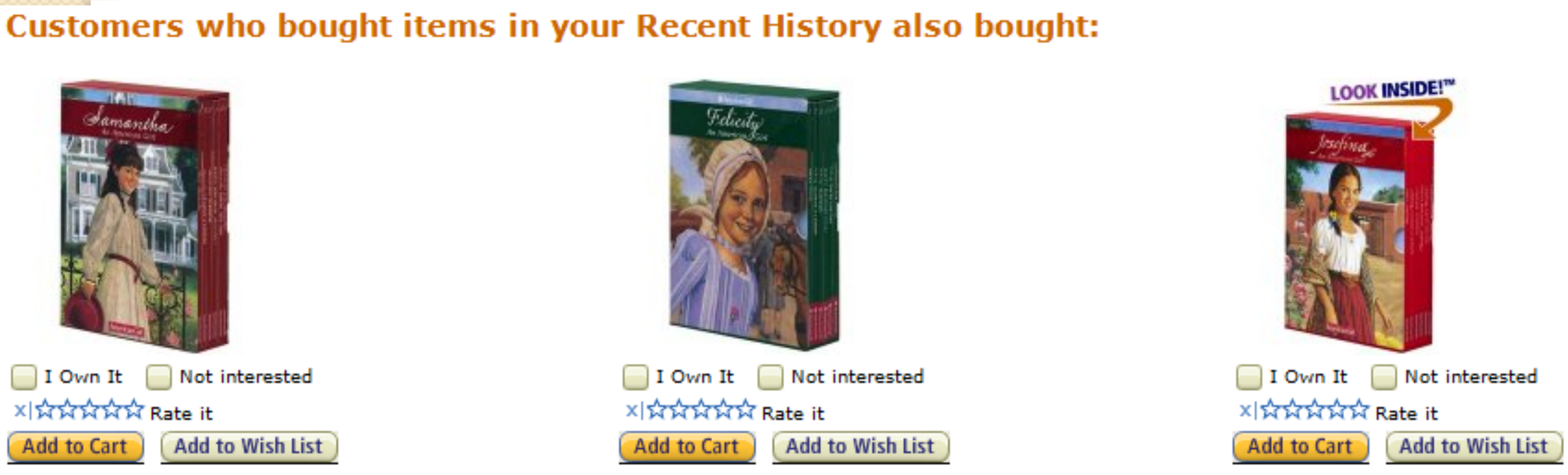
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!

Collaborative filtering

- Recommend items based on past transactions of users
- Analyze relations between users and/or items
- Specific data characteristics are irrelevant
  - Domain-free: user/item attributes are not necessary
  - Can identify elusive aspects

amazon.com



Customers who bought items in your Recent History also bought:

☐ I Own It ☐ Not interested
☆☆☆☆☆ Rate it
Add to Cart Add to Wish List

☐ I Own It ☐ Not interested
☆☆☆☆☆ Rate it
Add to Cart Add to Wish List

☐ I Own It ☐ Not interested
☆☆☆☆☆ Rate it
Add to Cart Add to Wish List

Movie rating data

Training data

user	movie	score
1	21	1
1	213	5
2	345	4
2	123	4
2	768	3
3	76	5
4	45	4
5	568	1
5	342	2
5	234	2
6	76	5
6	56	4

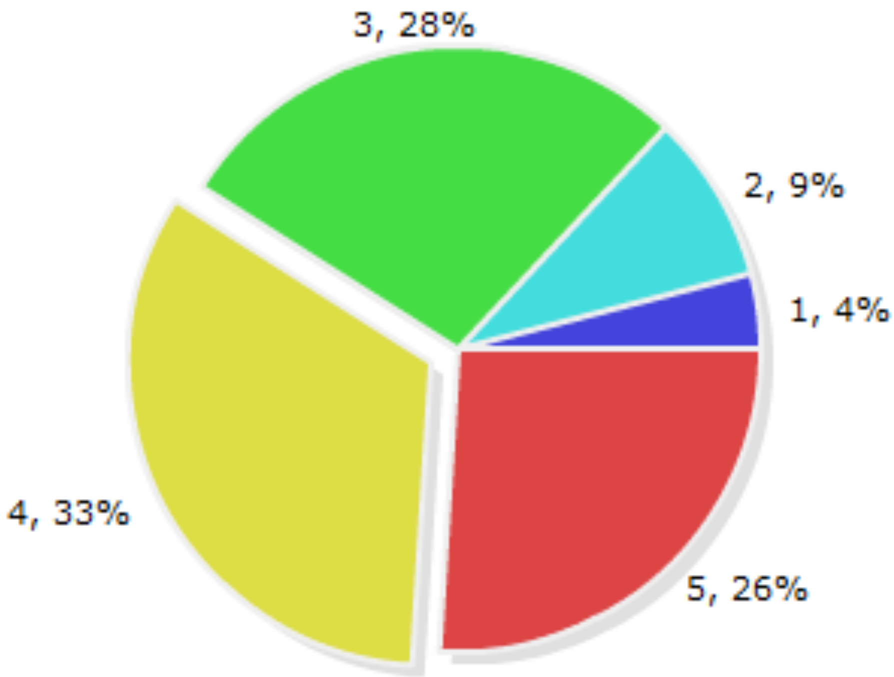
Test data

user	movie	score
1	62	?
1	96	?
2	7	?
2	3	?
3	47	?
3	15	?
4	41	?
4	28	?
5	93	?
5	74	?
6	69	?
6	83	?

Netflix Prize

- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005
- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation criterion: root mean squared error (RMSE)
  - Netflix Cinematch RMSE: 0.9514
- Competition
  - 2700+ teams
  - $1 million grand prize for 10% improvement on Cinematch result
  - $50,000 2007 progress prize for 8.43% improvement

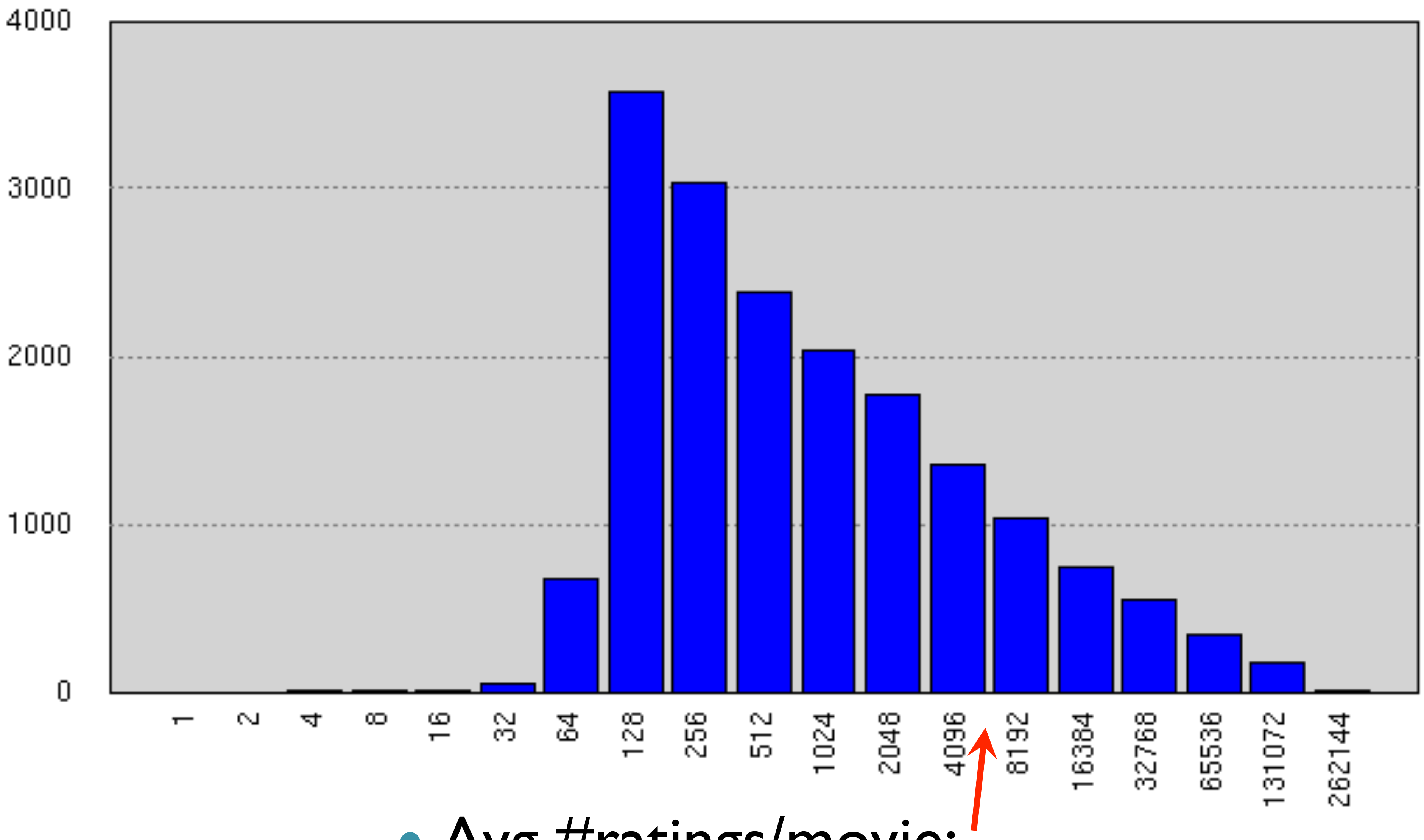
Overall rating distribution



- Third of ratings are 4s
- Average rating is 3.68

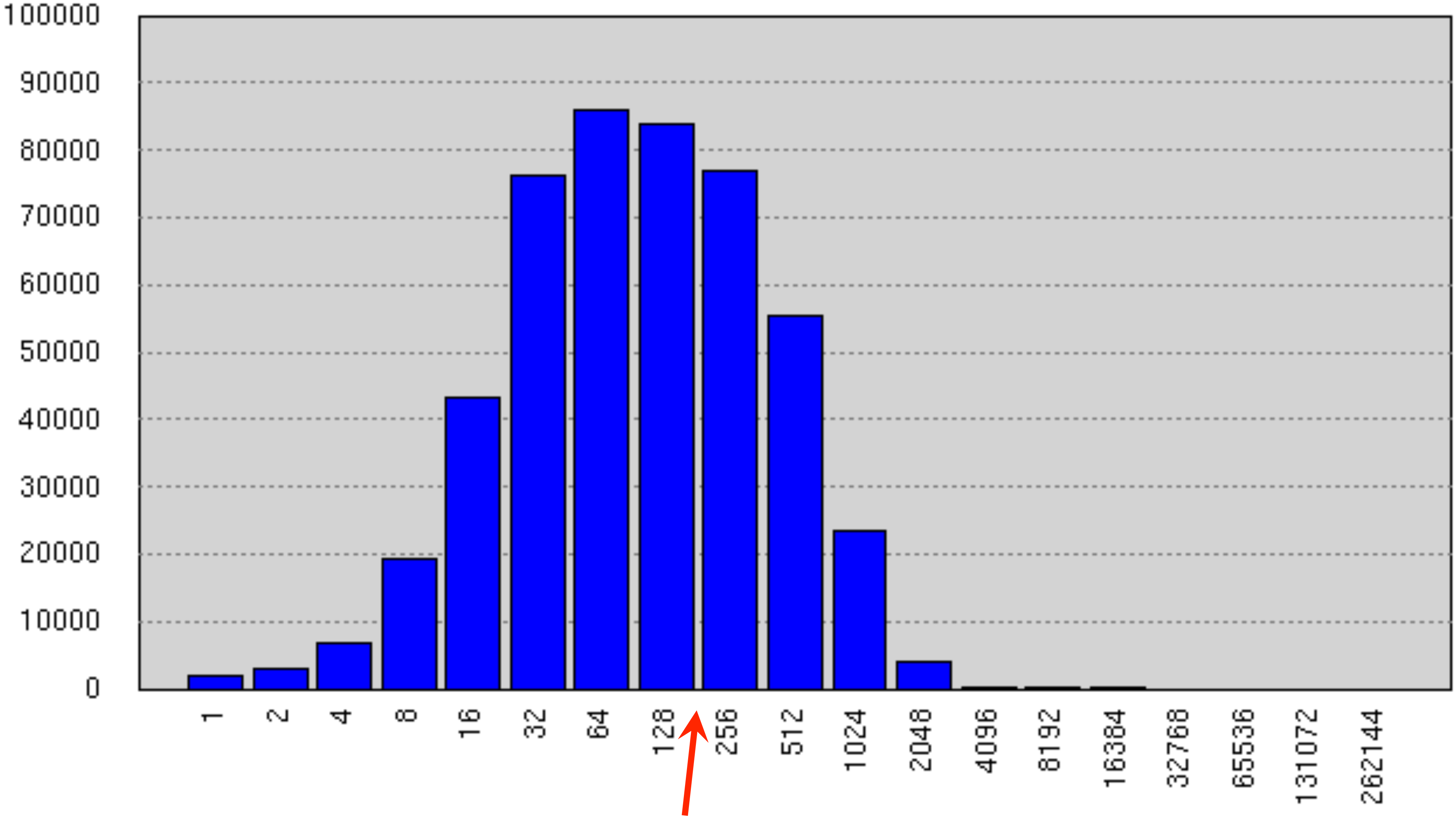
From TimelyDevelopment.com

#ratings per movie



- Avg #ratings/movie: 5627

#ratings per user

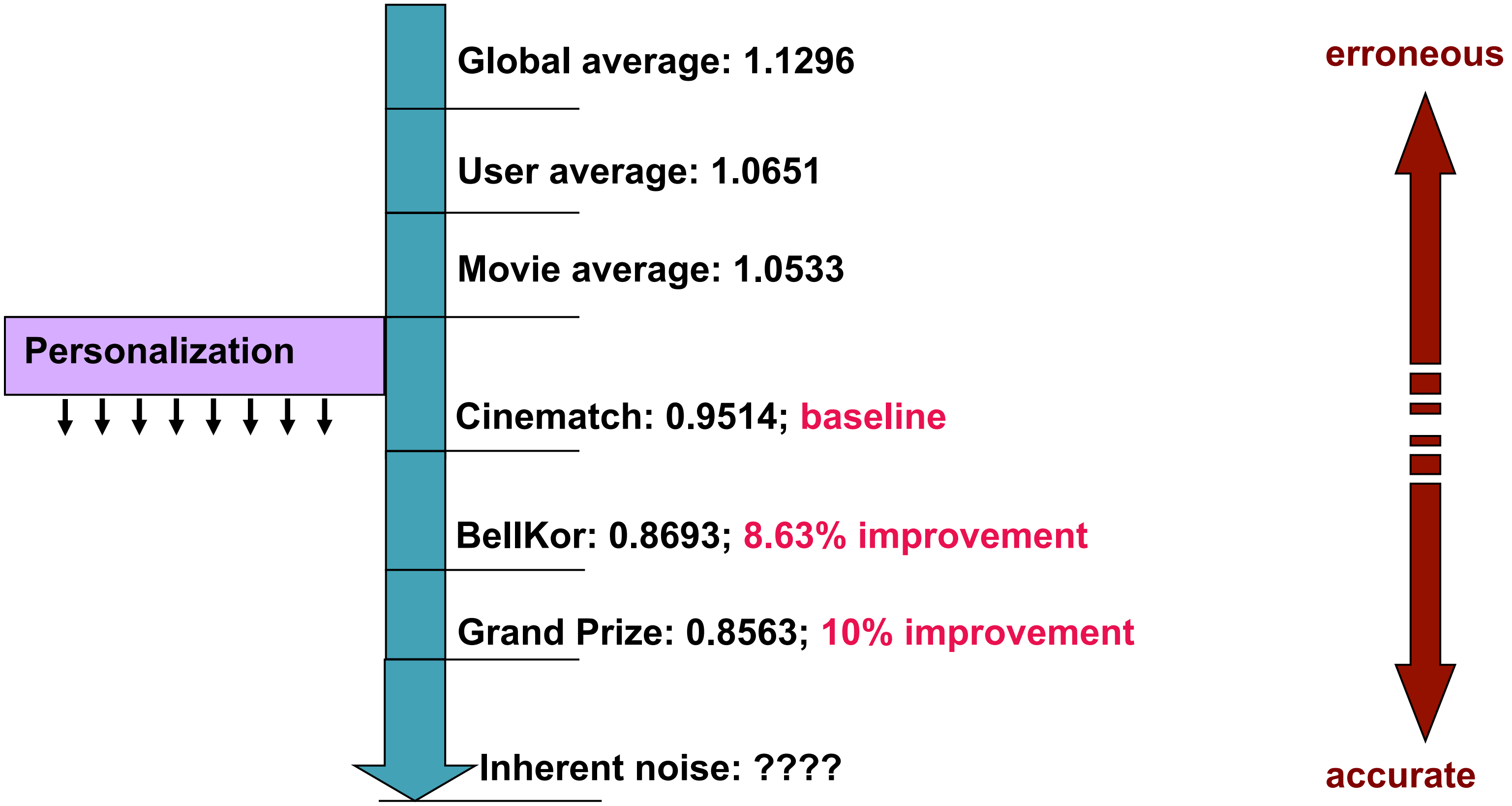


- Avg #ratings/user: 208

Most loved movies

Title	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	150676
Finding Nemo	4.415	139050
Raiders of the Lost Ark	4.504	117456
Forrest Gump	4.299	180736
Lord of the Rings: The Fellowship of the ring	4.433	147932
The Sixth Sense	4.325	149199
Indiana Jones and the Last Crusade	4.333	144027

Important RMSEs



Personalization

Global average: 1.1296

User average: 1.0651

Movie average: 1.0533

Cinematch: 0.9514; baseline

BellKor: 0.8693; 8.63% improvement

Grand Prize: 0.8563; 10% improvement

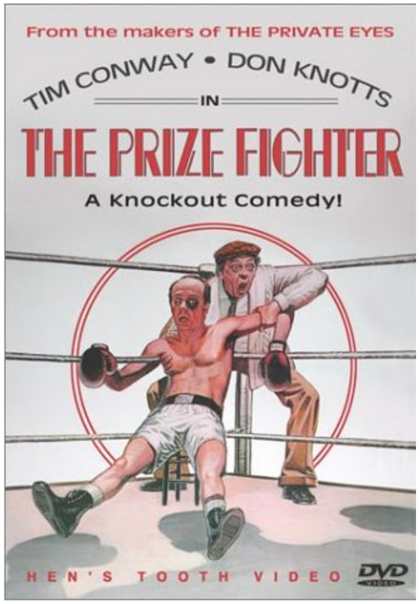
Inherent noise: ????

erroneous

accurate

Challenges

- Size of data
  - Scalability
  - Keeping data in memory
- Missing data
  - 99 percent missing
  - Very imbalanced
- Avoiding overfitting
- Test and training data differ significantly



movie #16322

k-NN

users

movies

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

☐ **- unknown rating**

🟨 **- rating between 1 to 5**

k-NN

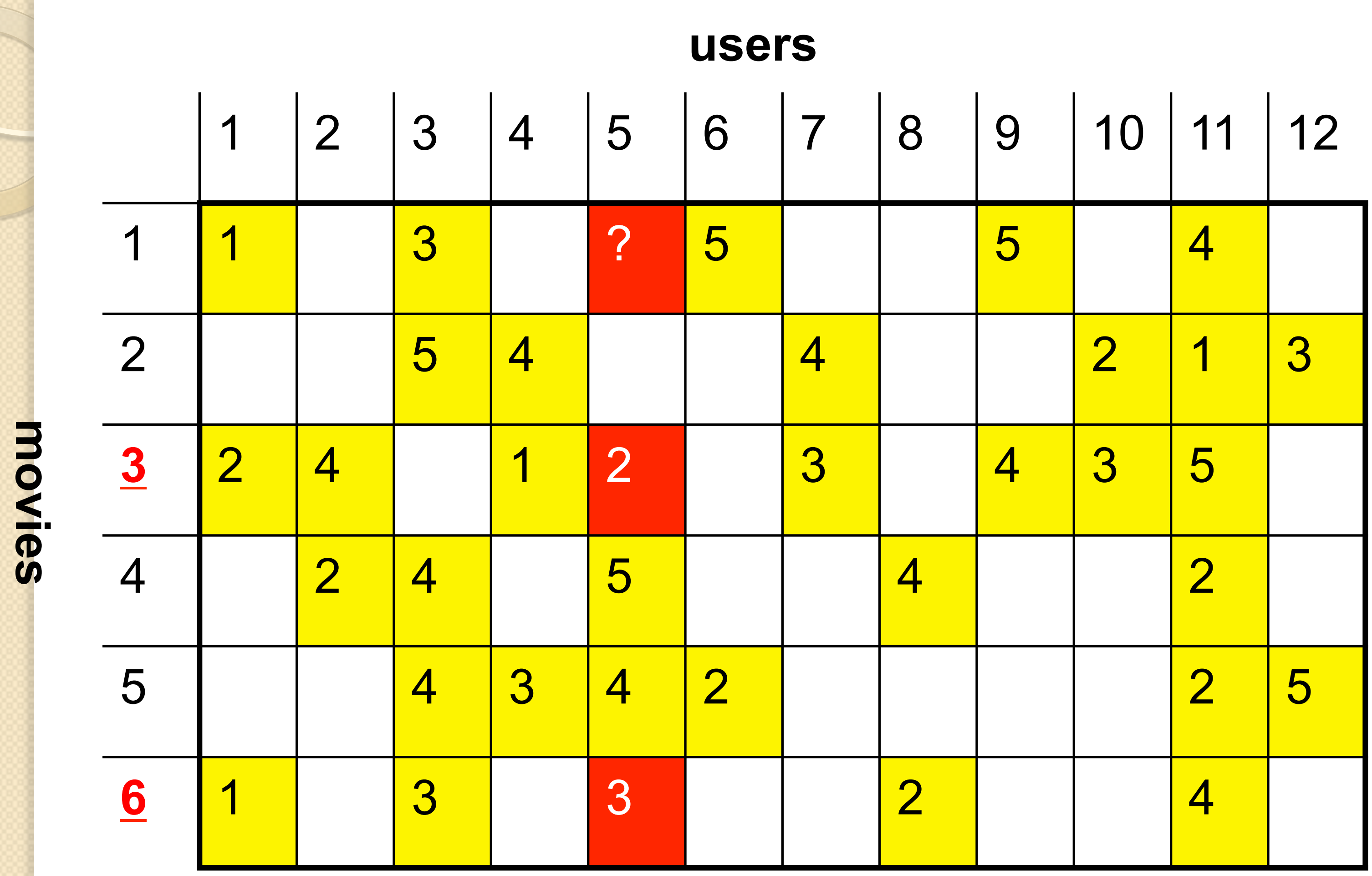
users

movies

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

🟥 **- estimate rating of movie 1 by user 5**

k-NN



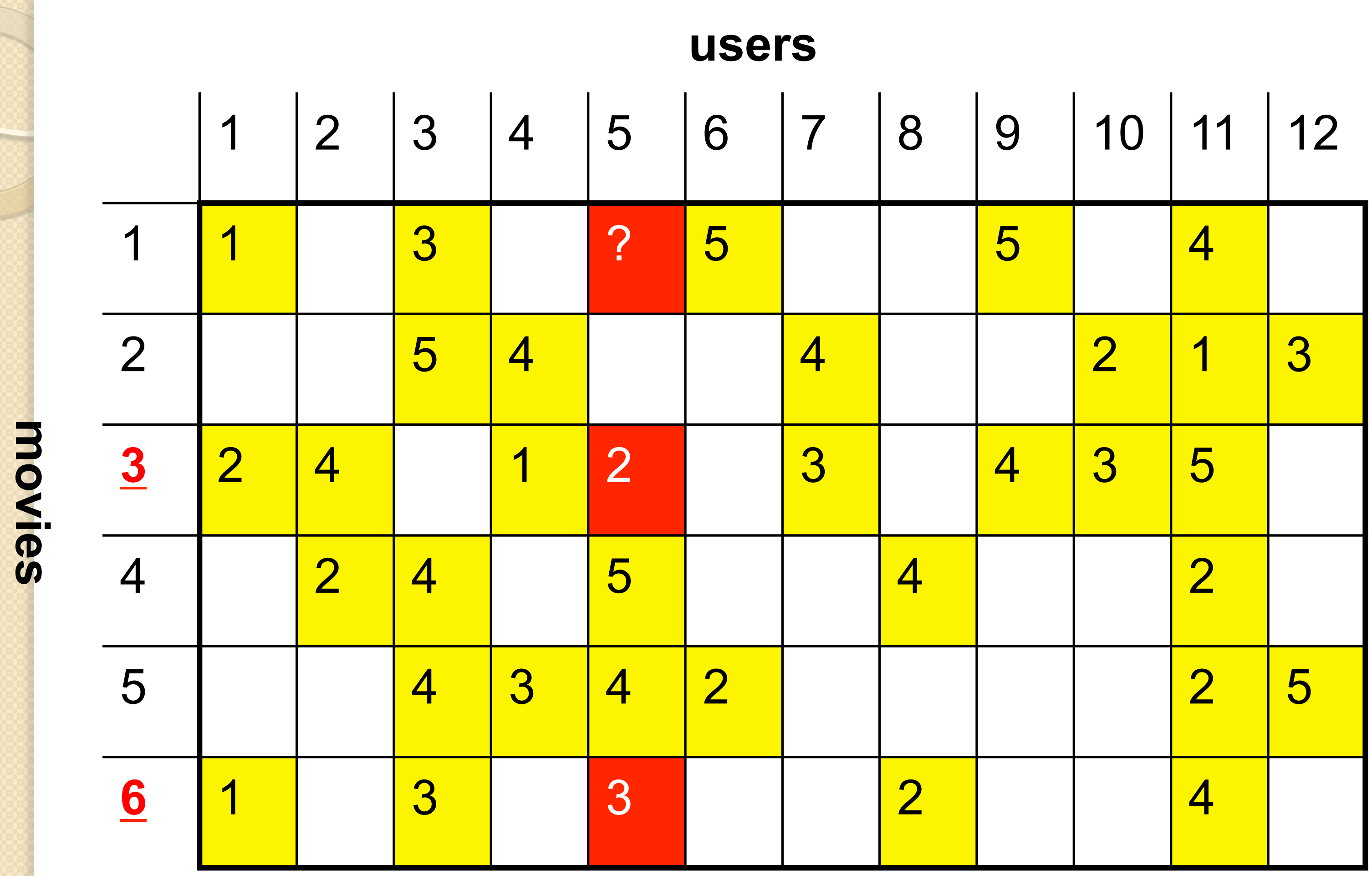
users

movies

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Neighbor selection:
Identify movies similar to 1, rated by user 5

k-NN



users

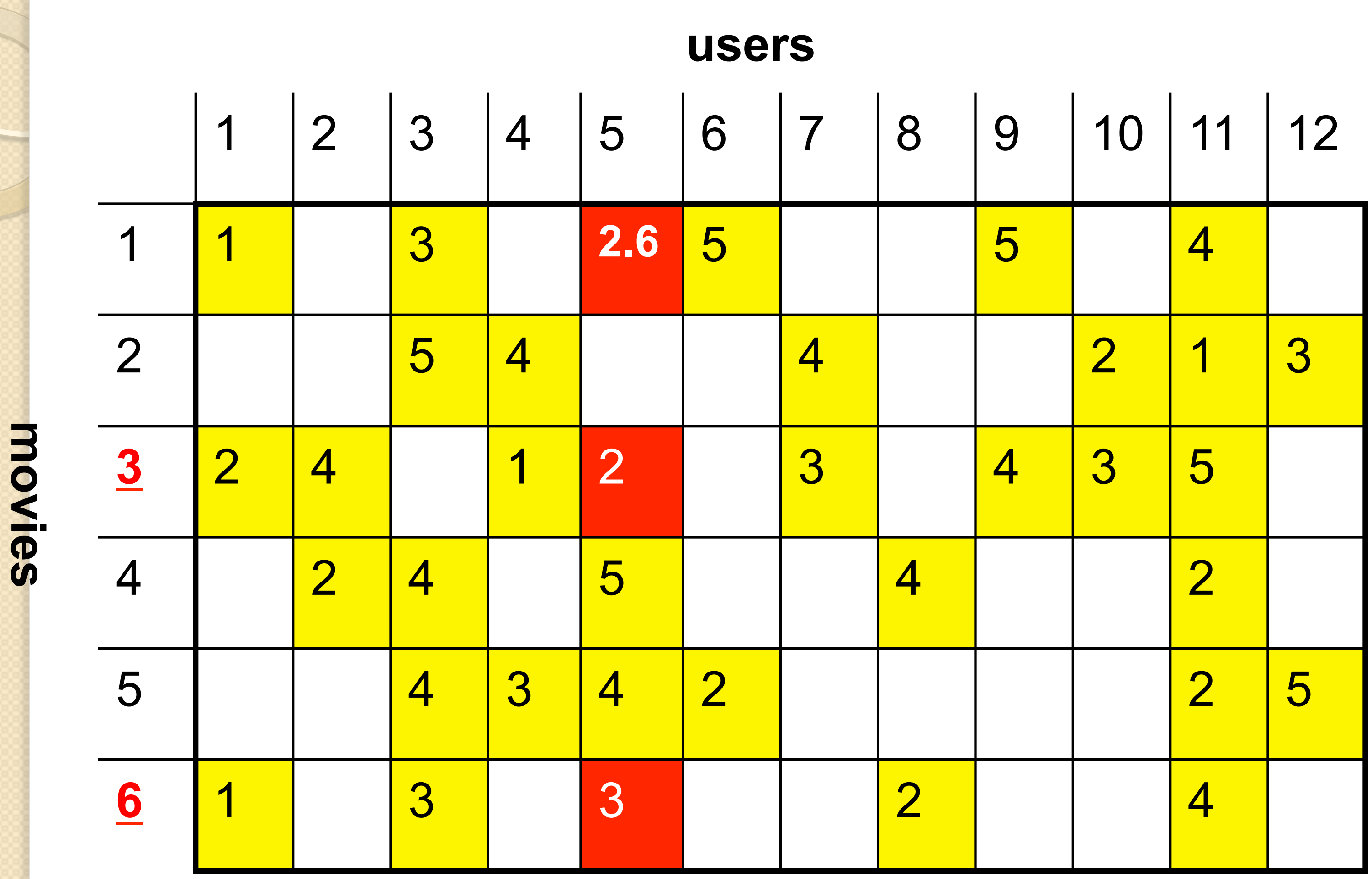
movies

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Compute similarity weights:

$s_{13}=0.2, s_{16}=0.3$

k-NN



users

movies

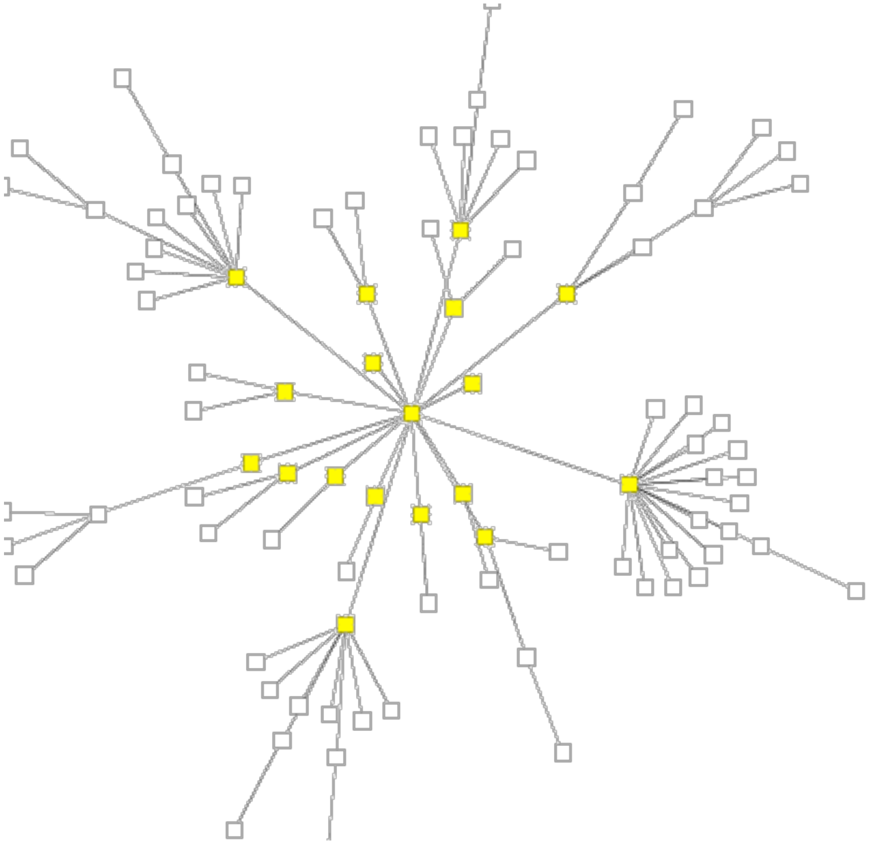
	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		**2.6**	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Predict by taking weighted average:

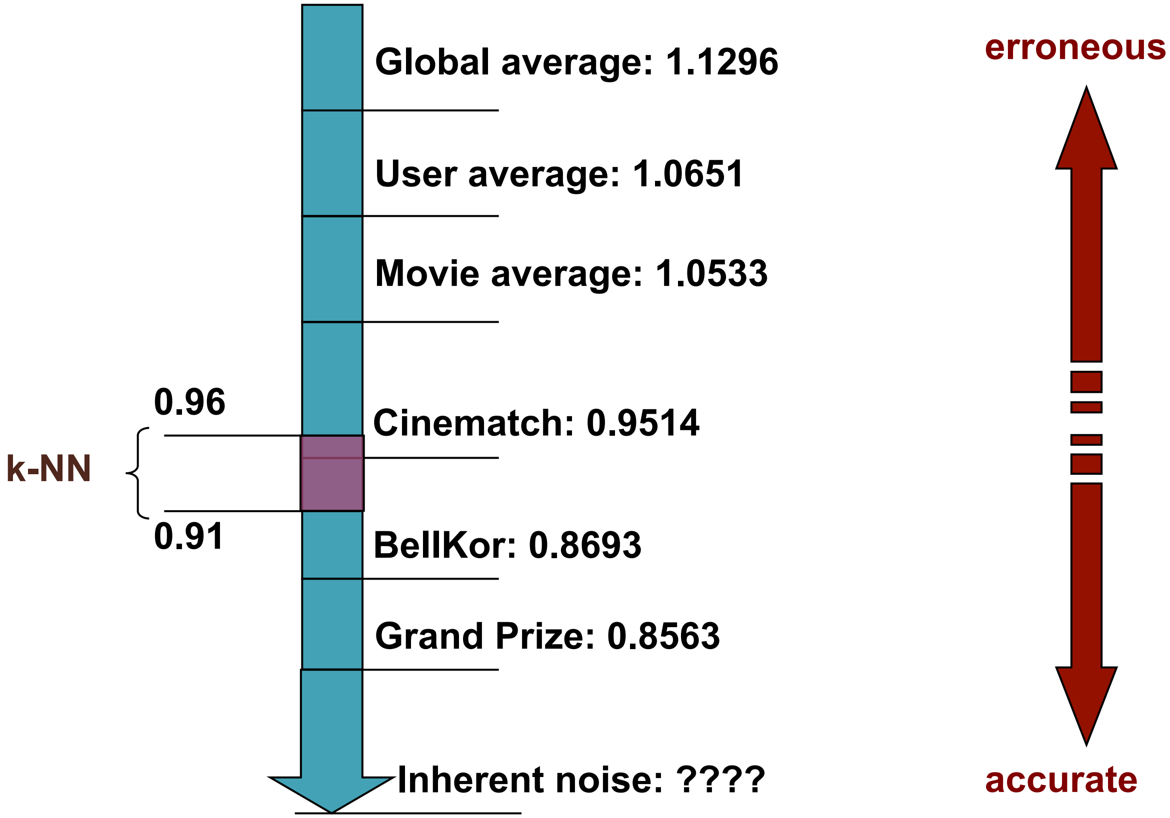
$(0.2*2+0.3*3)/(0.2+0.3)=2.6$

Properties of k-NN

- Intuitive
- No substantial preprocessing is required
- Easy to explain reasoning behind a recommendation
- Accurate?



k-NN on the RMSE scale



Global average: 1.1296

User average: 1.0651

Movie average: 1.0533

Cinematch: 0.9514

BellKor: 0.8693

Grand Prize: 0.8563

Inherent noise: ????

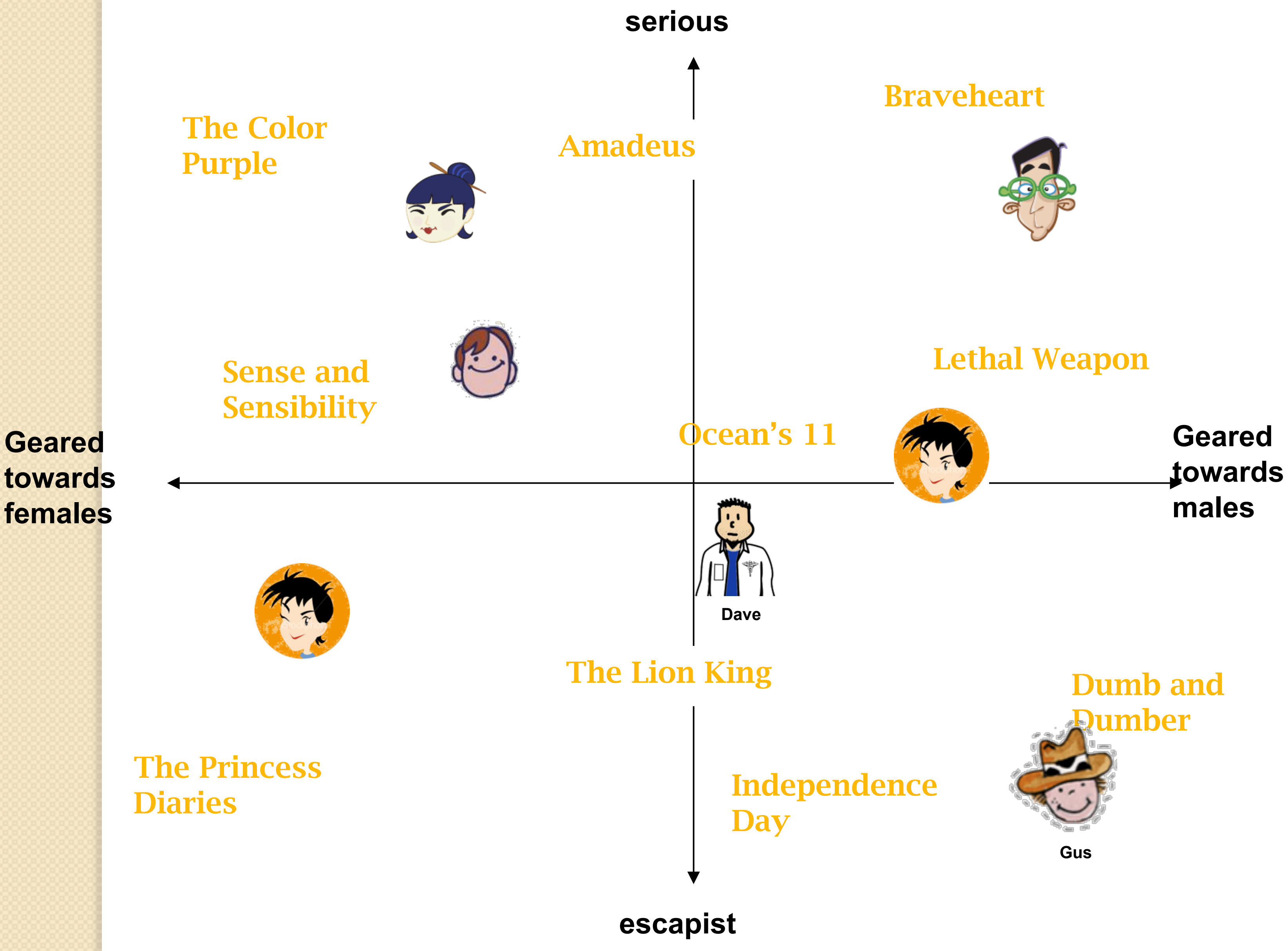
k-NN - Common practice

1. Define a similarity measure between items: s_{ij}
2. Select neighbors -- N(i;u): items most similar to i, that were rated by u
3. Estimate unknown rating, r_{ui}, as the weighted average:

$$r_{ui} = b_{ui} + \frac{\sum_{j \in N(i;u)} s_{ij} \left(r_{uj} - b_{uj} \right)}{\sum_{j \in N(i;u)} s_{ij}}$$

baseline estimate for r_{ui}

Latent factor models



serious

The Color Purple

Amadeus

Braveheart

Sense and Sensibility

Lethal Weapon

Ocean's 11

Geared towards females

Geared towards males

Dave

The Lion King

Dumb and Dumber

The Princess Diaries

Independence Day

Gus

escapist

Latent factor models

users

items

1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

~

~

items

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

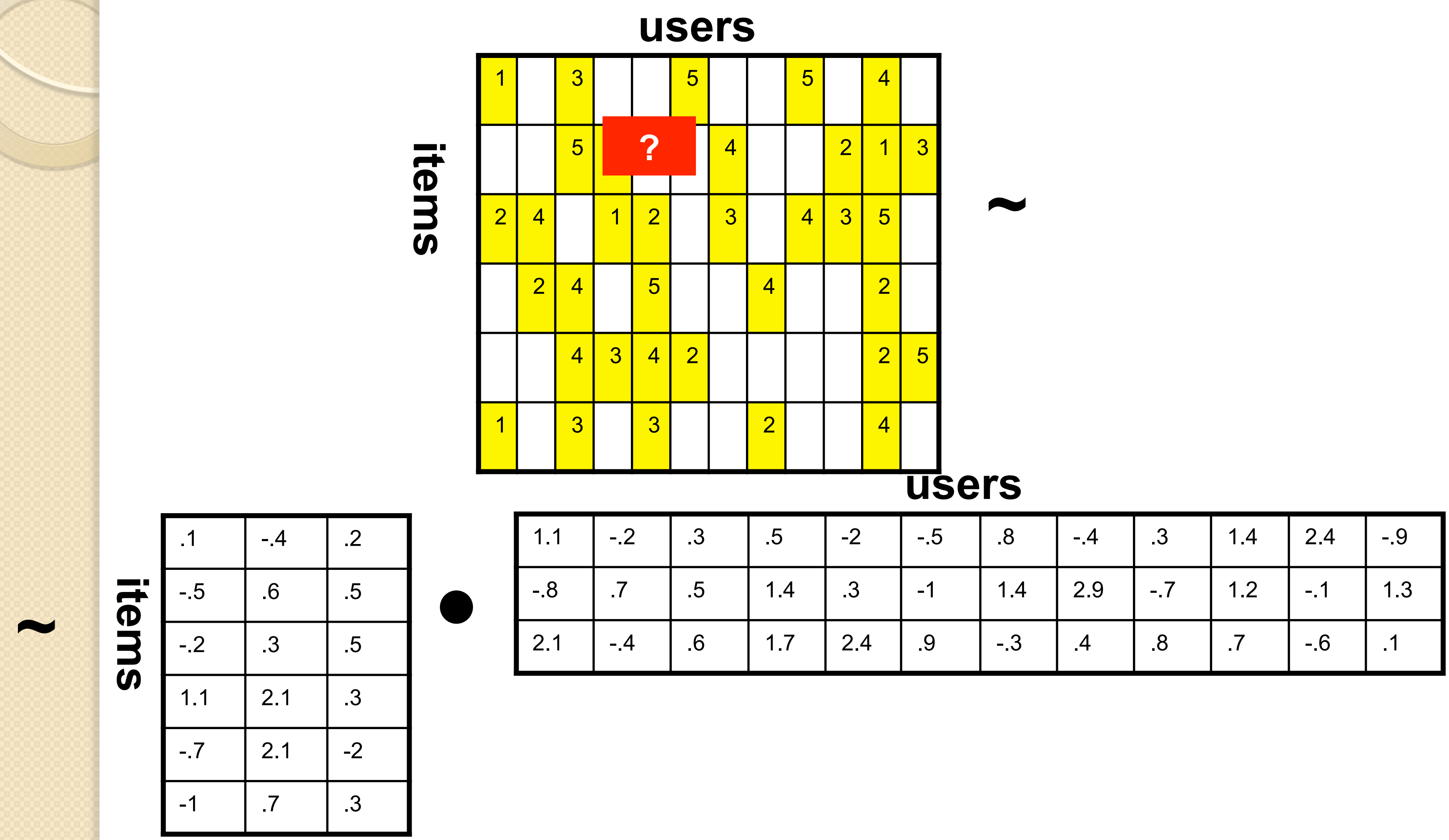
●

users

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

A rank-3 SVD approximation

Estimate unknown ratings as inner-products of factors:



users

items

1		3			5			5		4	
		5		?		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

~

~

items

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

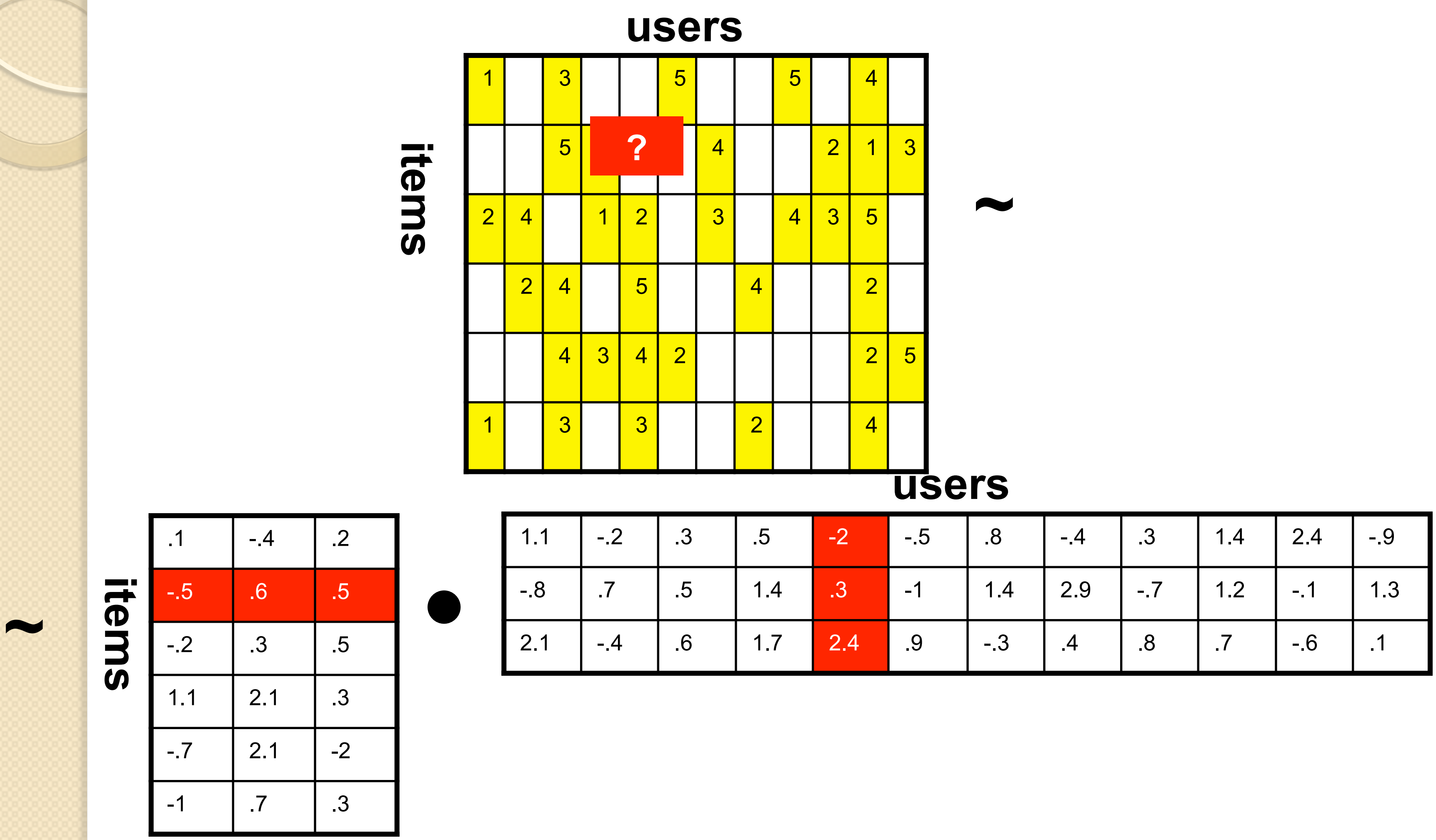
●

users

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

A rank-3 SVD approximation

Estimate unknown ratings as inner-products of factors:



users

items

1		3			5			5		4	
		5		?		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

~

~ items

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

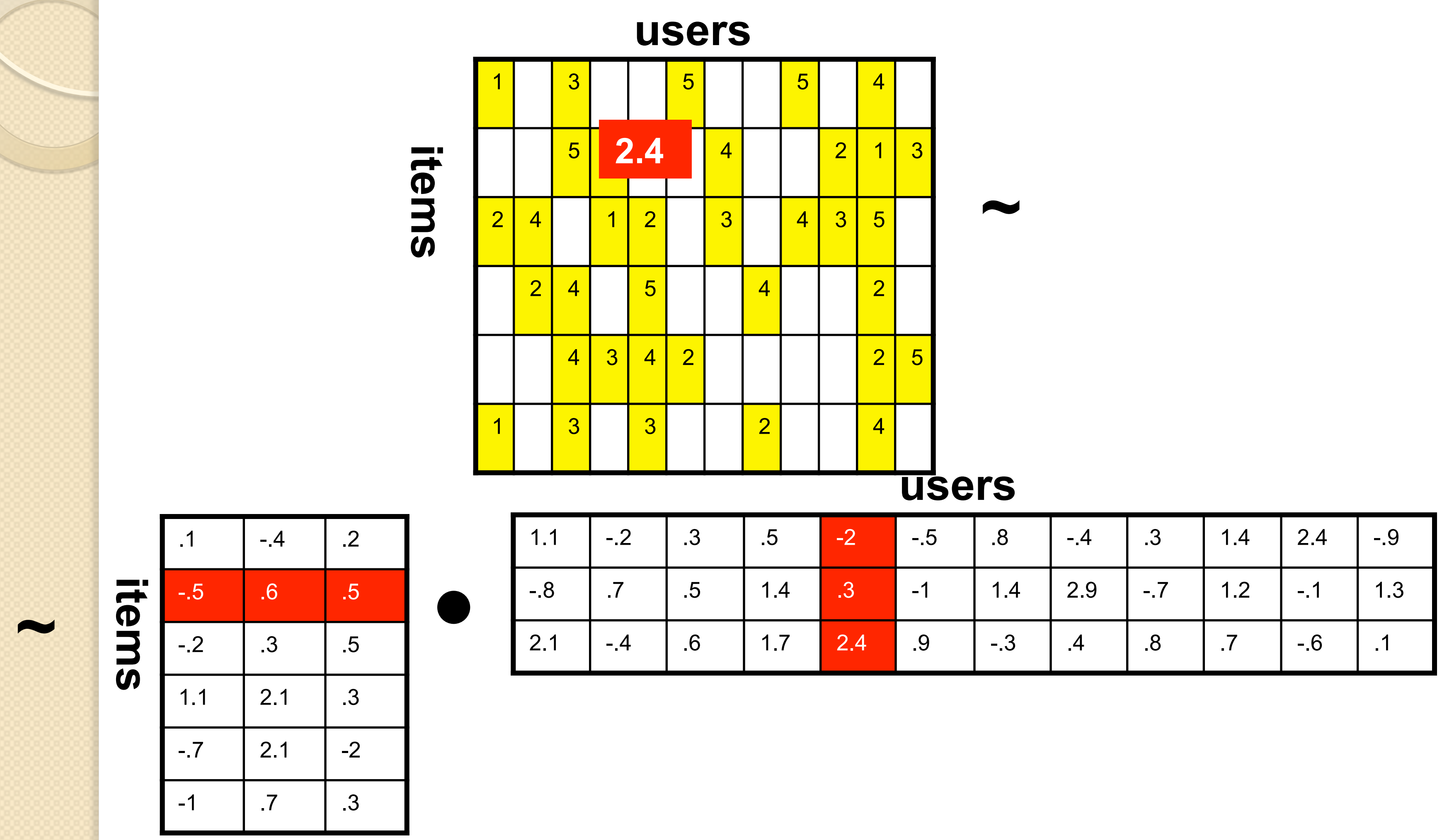
●

users

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

A rank-3 SVD approximation

Estimate unknown ratings as inner-products of factors:



users

items

1		3			5			5		4	
		5		2.4		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

~

~

items

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

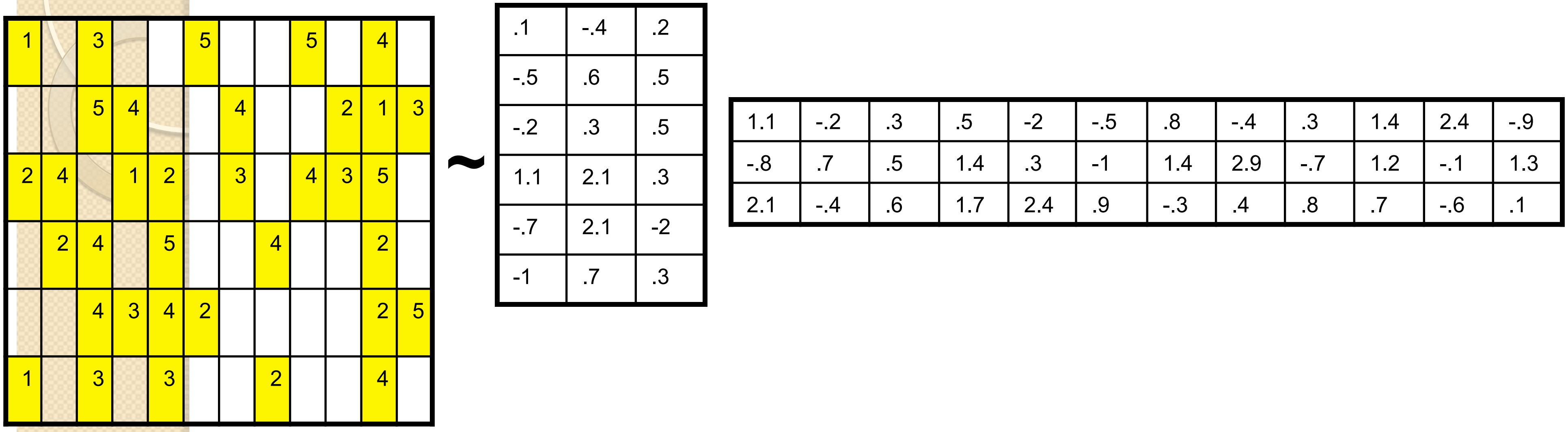
●

users

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

A rank-3 SVD approximation

Latent factor models



1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

~

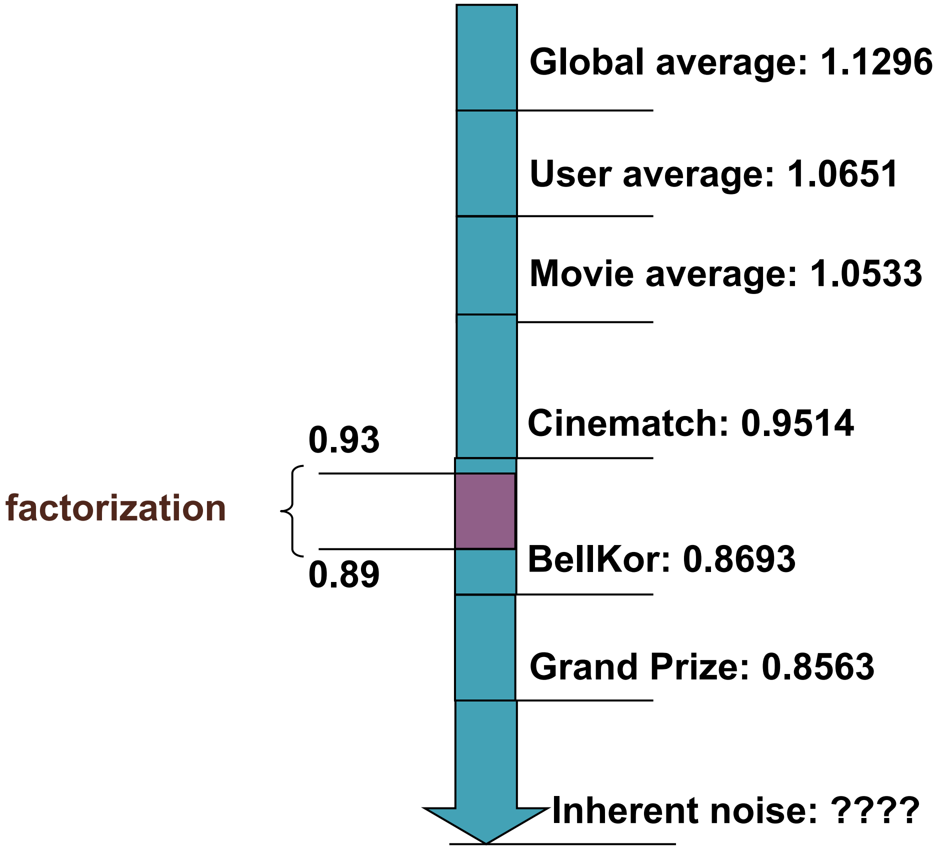
.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

Properties:

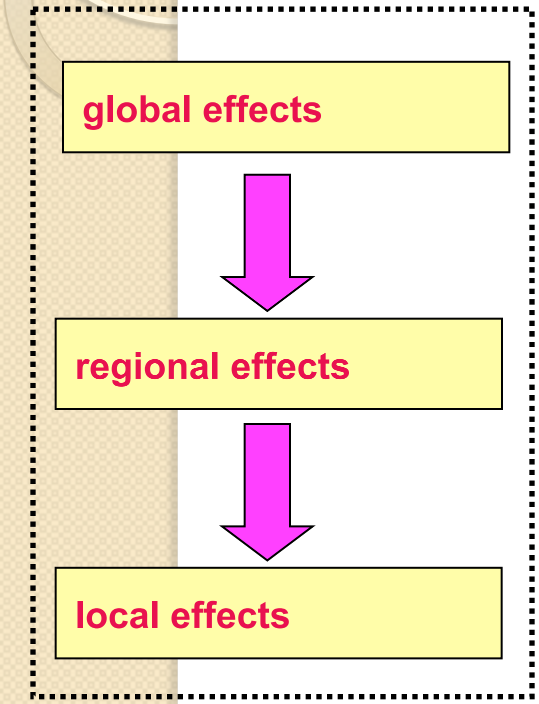
- SVD isn’t defined when entries are unknown ➔ use specialized methods
- Very powerful model ➔ can easily overfit, sensitive to regularization
- Probably most popular model among contestants
  - 12/11/2006: Simon Funk describes an SVD based method
  - 12/29/2006: Free implementation at timelydevelopment.com

Factorization on the RMSE scale



Combining multi-scale views

Residual fitting

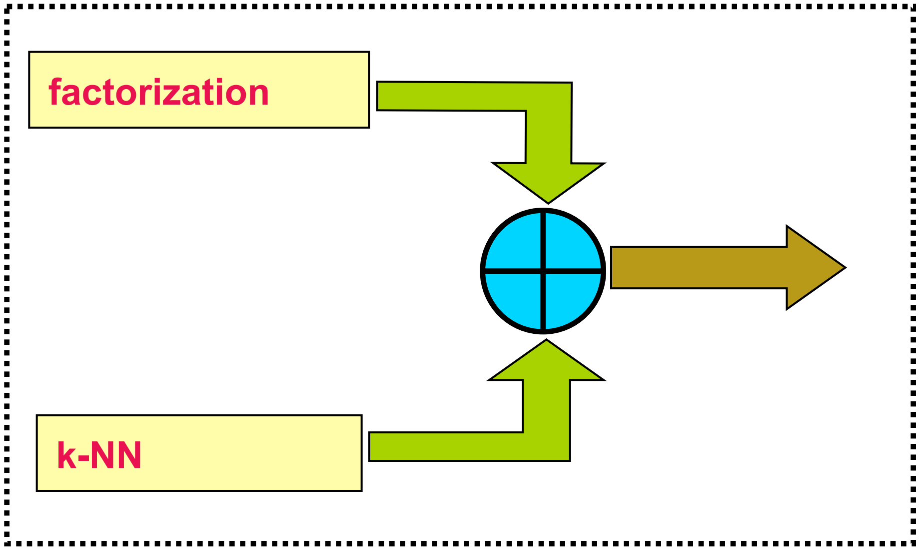


global effects

regional effects

local effects

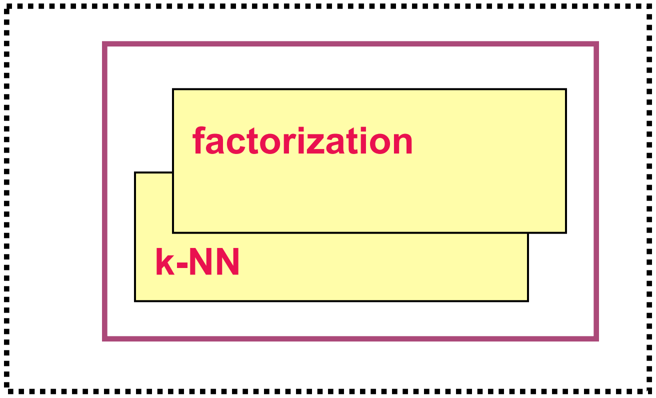
Weighted average



factorization

k-NN

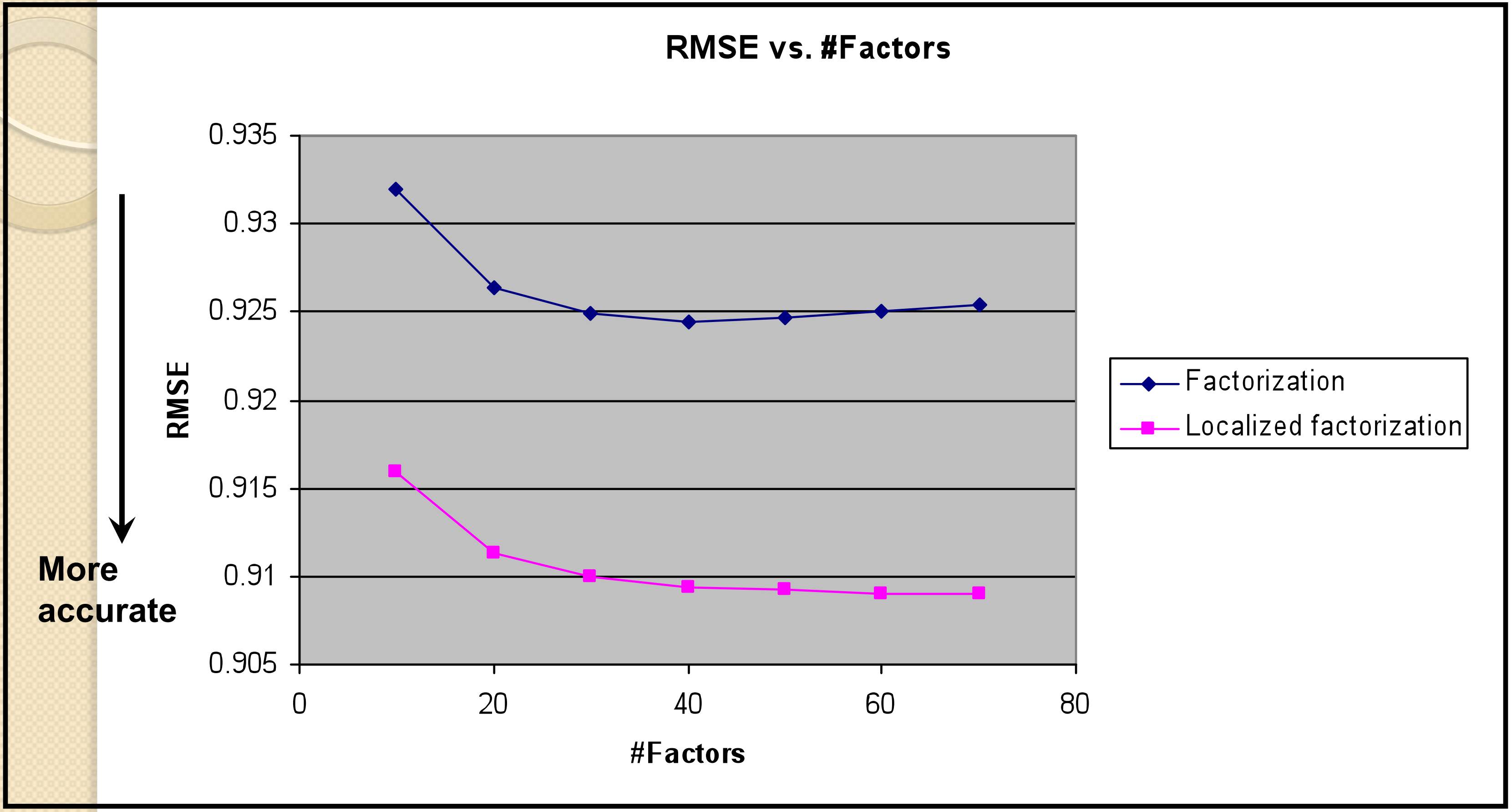
A unified model



factorization

k-NN

Results on Netflix Probe set

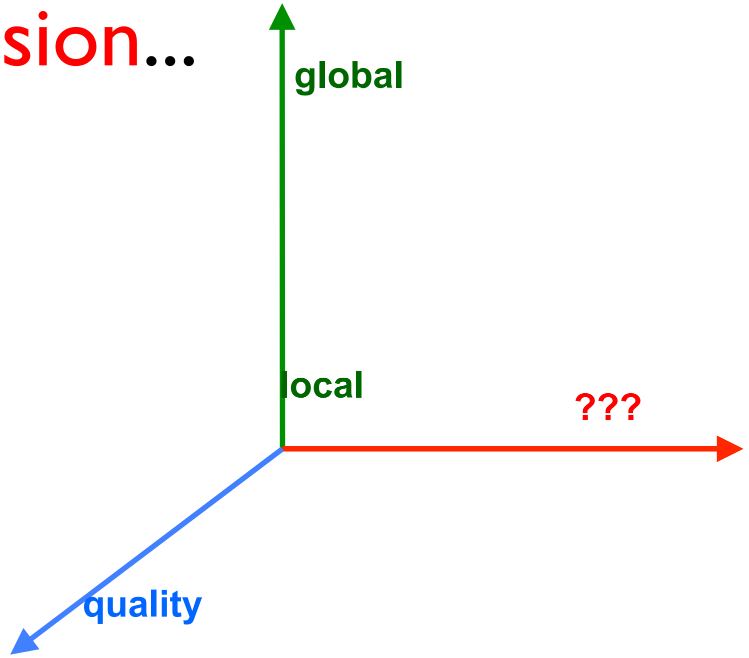


RMSE vs. #Factors

More accurate

Seek alternative perspectives of the data

- Can exploit movie titles and release year
- But movies side is pretty much covered anyway...
- It's about the users!
- Turning to the third dimension...



Lessons

What it takes to win:

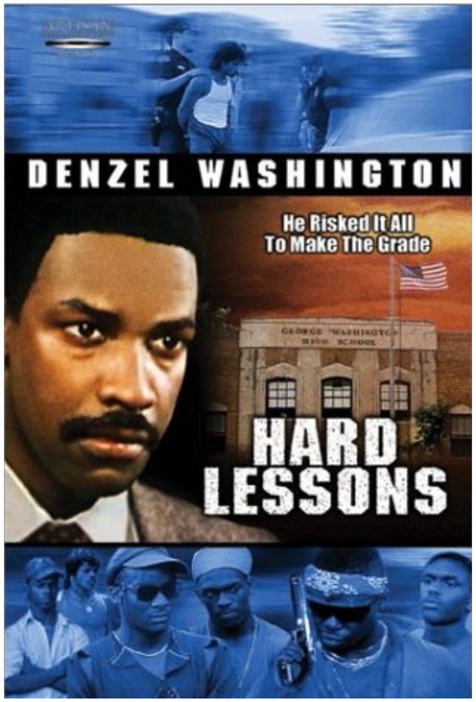
1. Think deeper – design better algorithms
2. Think broader – use an ensemble of multiple predictors
3. Think different – model the data from different perspectives

At the personal level:

1. Have fun with the data
2. Work hard, long breath
3. Good teammates

Rapid progress of science:

1. Availability of large, real life data
2. Challenge, competition
3. Effective collaboration



movie #13043