Using Big Data to Improve Product Quality

Thomas Debeauvais

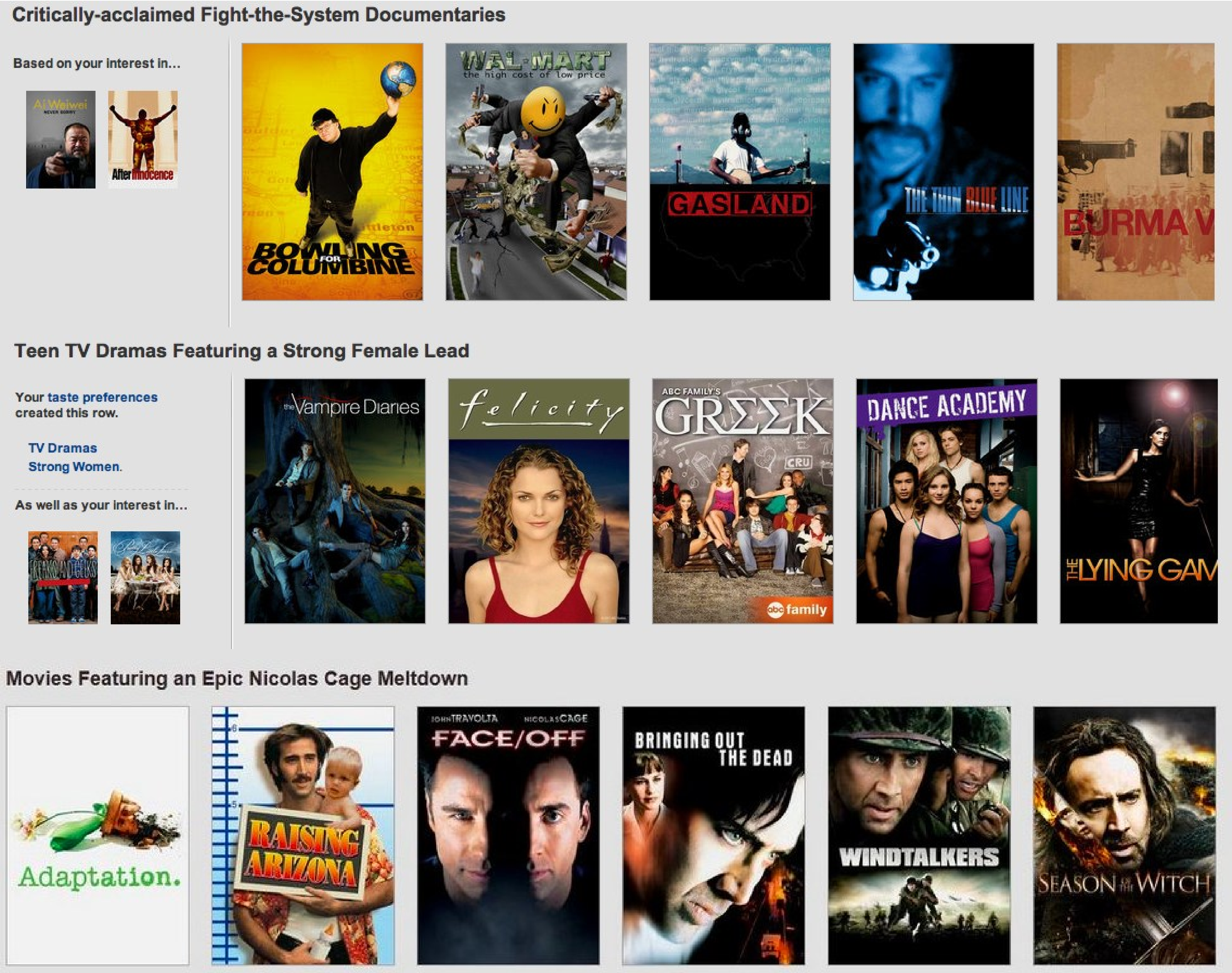
tdebeauv@uci.edu

Outline

- Netflix recommendations
- What is big data?
- Industry examples
- A technical example: singular value decomposition for recommender systems
- Thought exercises

NETFLIX

“Connecting people to the movies they love”



Critically-acclaimed Fight-the-System Documentaries

Based on your interest in...

Teen TV Dramas Featuring a Strong Female Lead

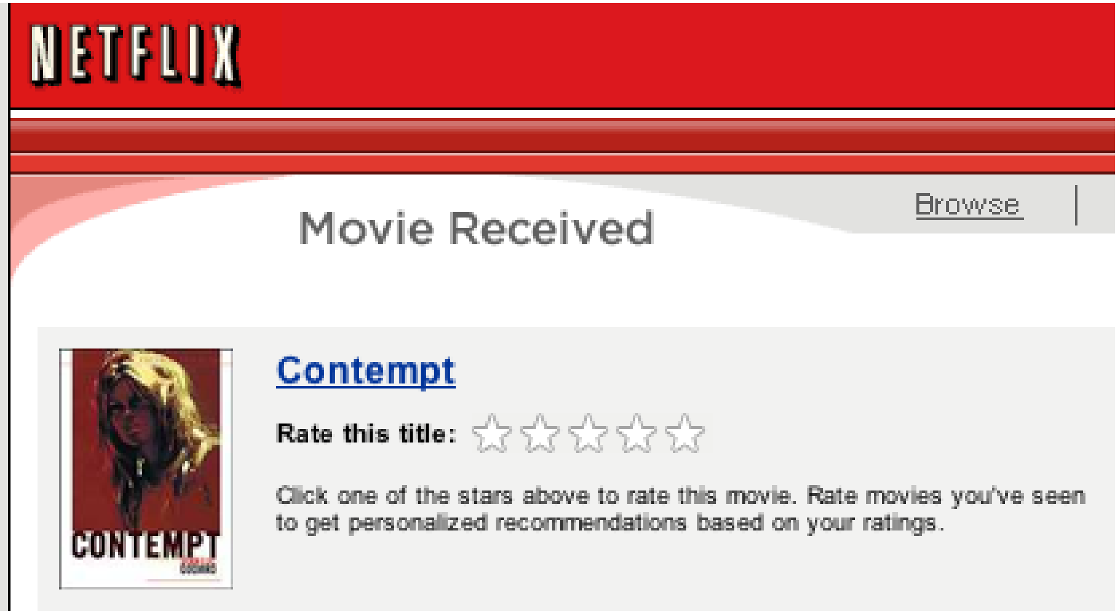
Your taste preferences created this row.

TV Dramas
Strong Women.

As well as your interest in...

Movies Featuring an Epic Nicolas Cage Meltdown

http://genresofnetflix.tumblr.com/

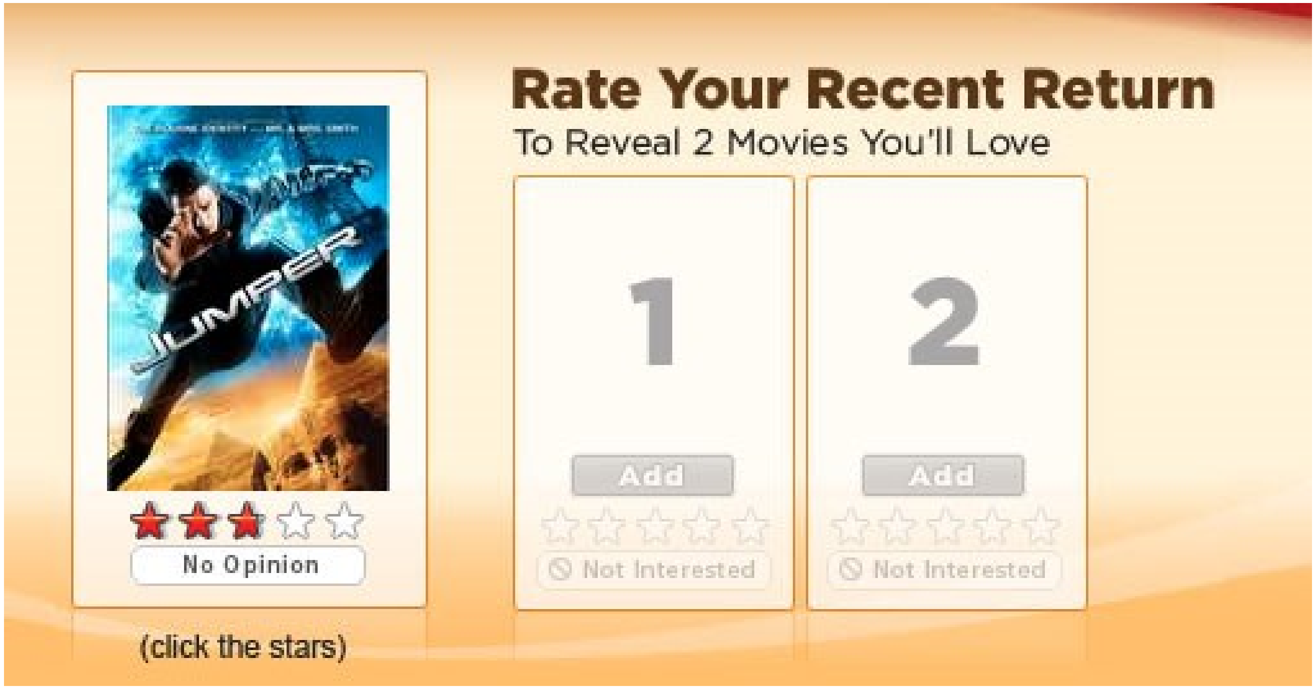


Movie Received

Contempt

Rate this title: ☆☆☆☆☆

Click one of the stars above to rate this movie. Rate movies you've seen to get personalized recommendations based on your ratings.



Rate Your Recent Return

To Reveal 2 Movies You'll Love

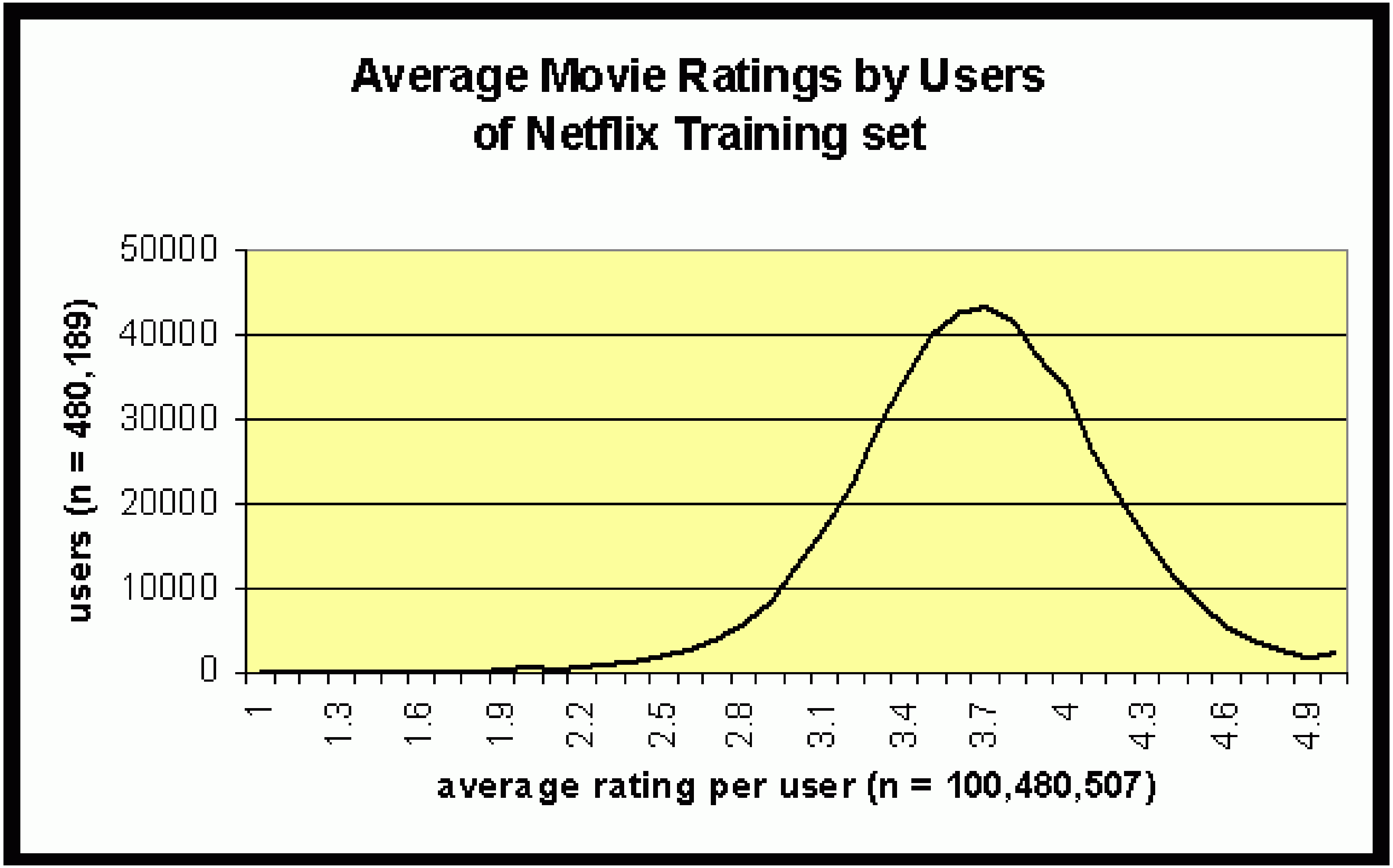
(click the stars)

Intuitions

1- People rate movies high

*They **generally** don’t bother rating bad movies*

People rate movies high



http://www.netflixprize.com/community/viewtopic.php?pid=5941#p5941

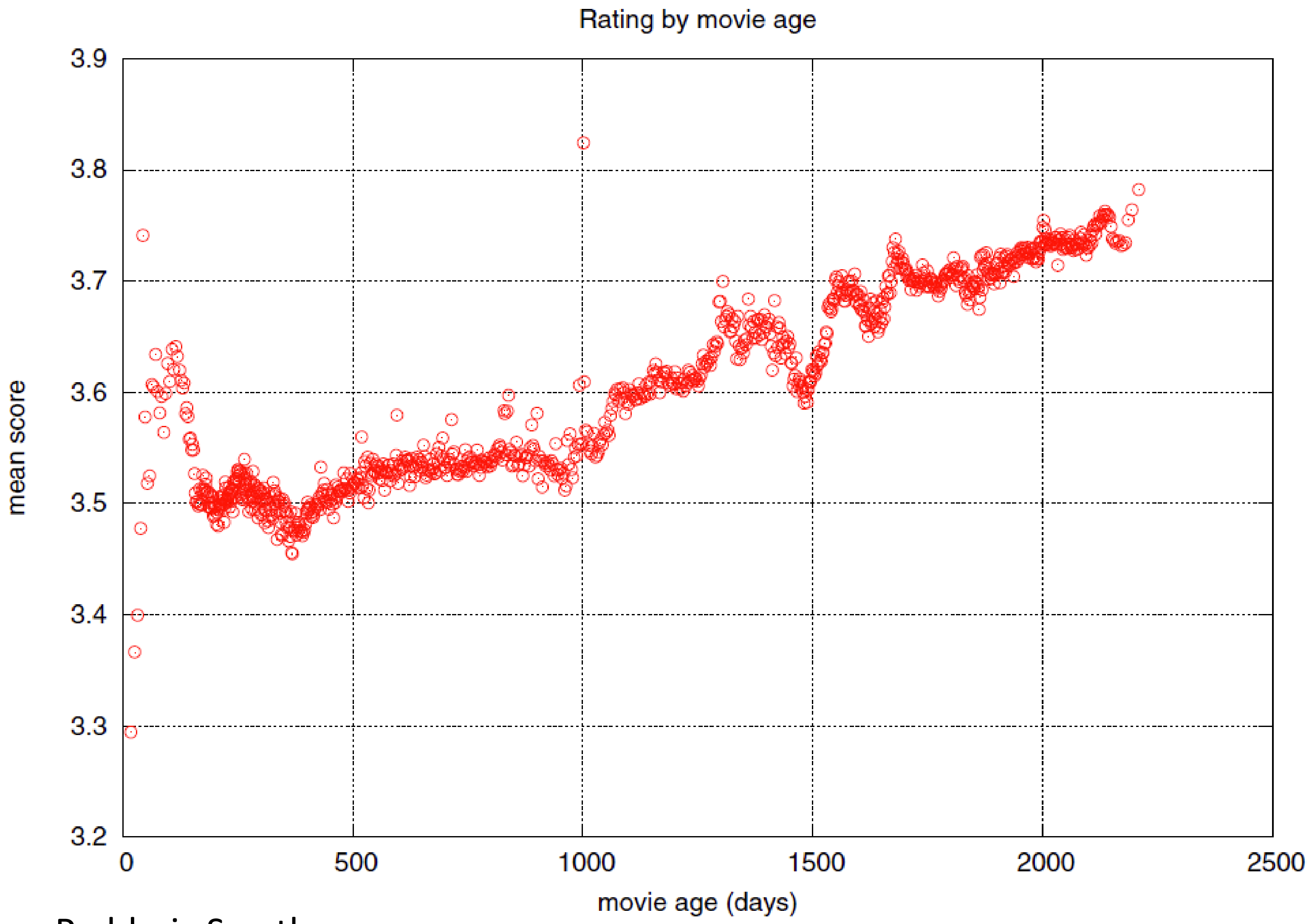
Intuitions

2- Older movies are rated higher

Nostalgia

Netflix keeps only the best of old movies

Older movies are rated higher

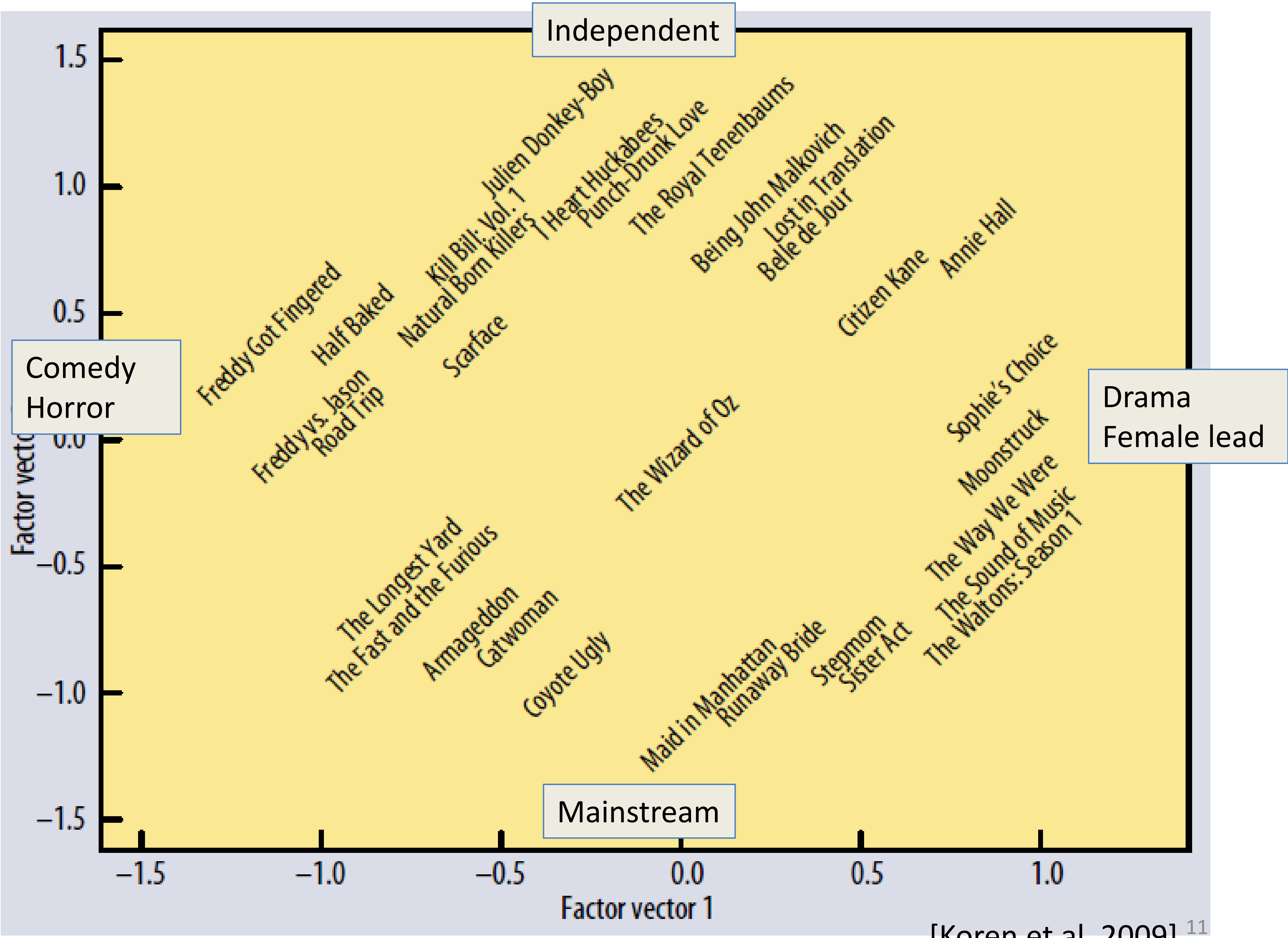


From Paddraic Smyth

Intuitions

MAGIC / FANTASY
DRAMA
COMEDY
SCI-FI
CARTOON / ANIMATION
HORROR
ROMANTIC
WESTERN
DETECTIVE
THRILLER
ACTION
ADVENTURE
N
E
W
S

3- Movies have genres



[Koren et al. 2009]

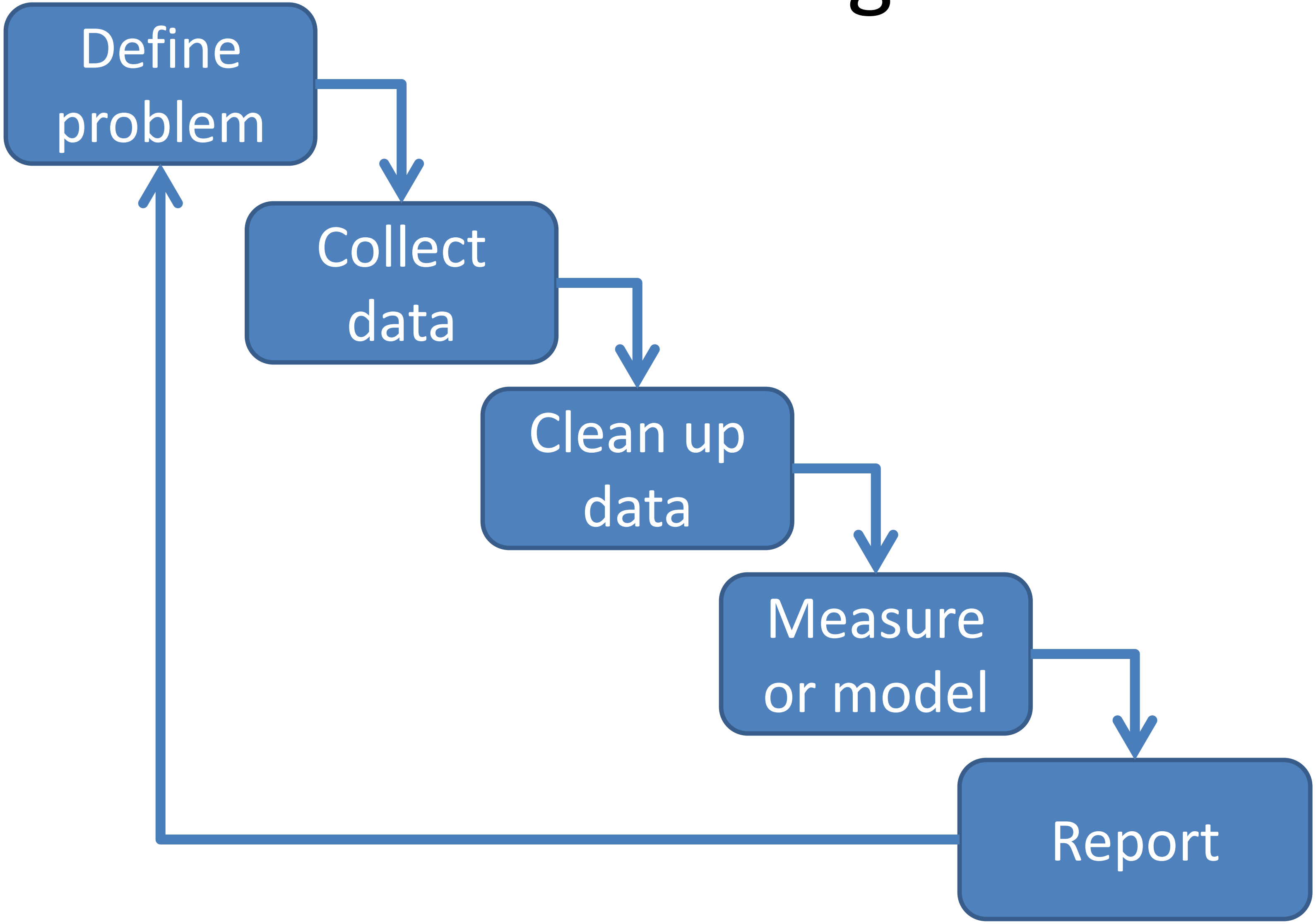
Take-away

- Confirm or reject intuitions with data

WHAT IS BIG DATA?

BIG DATA

Data mining



Define problem

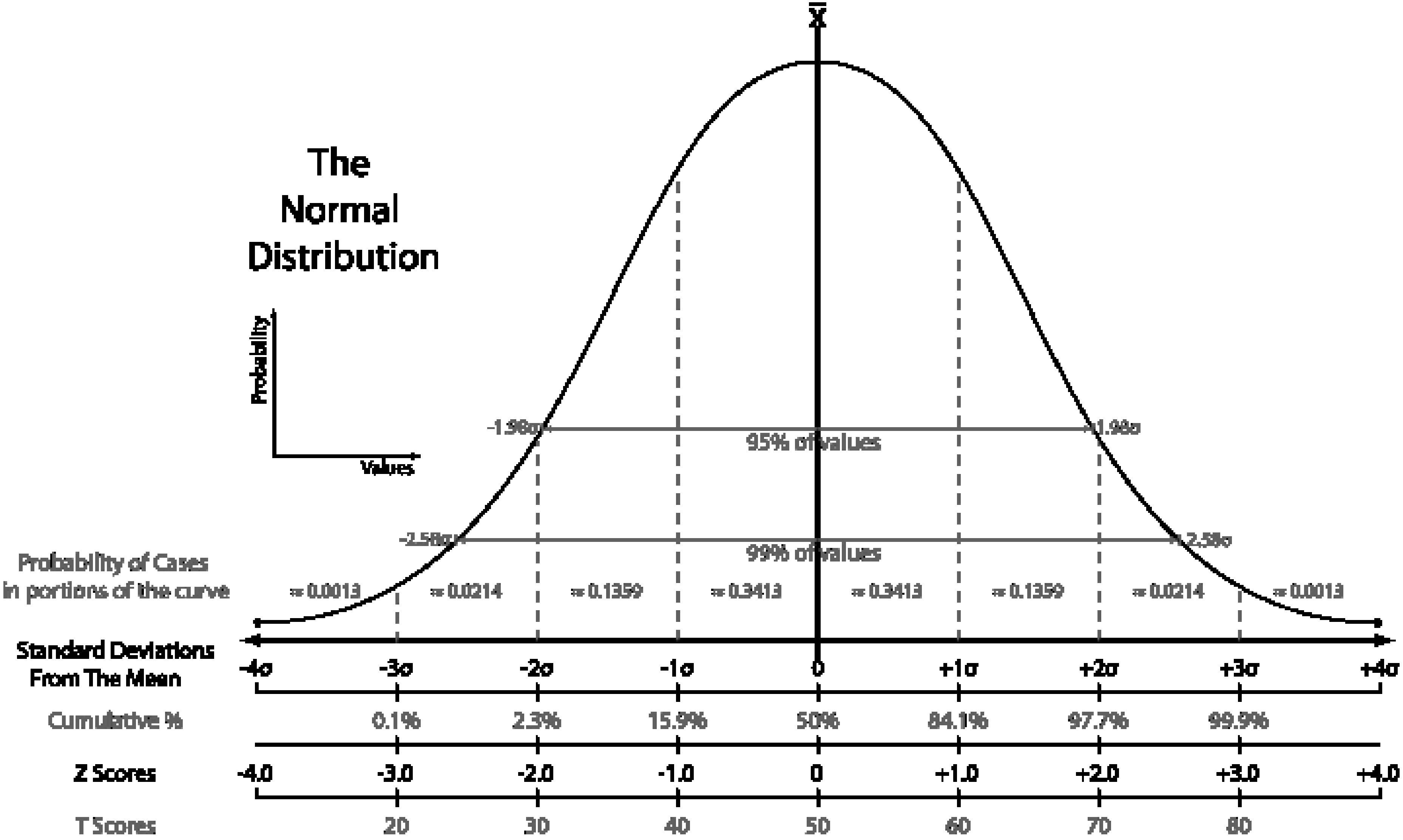
Collect data

Clean up data

Measure or model

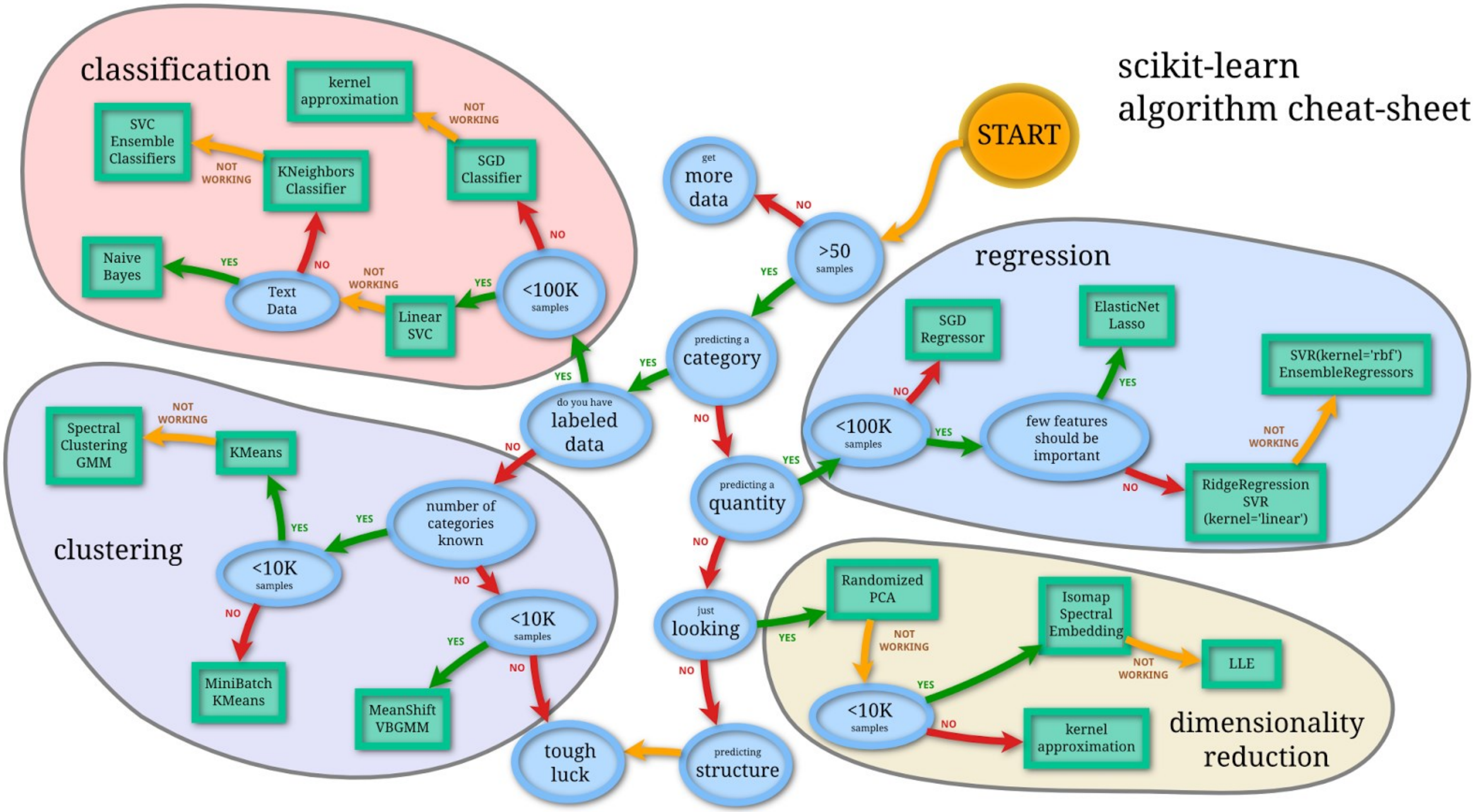
Report

Statistics



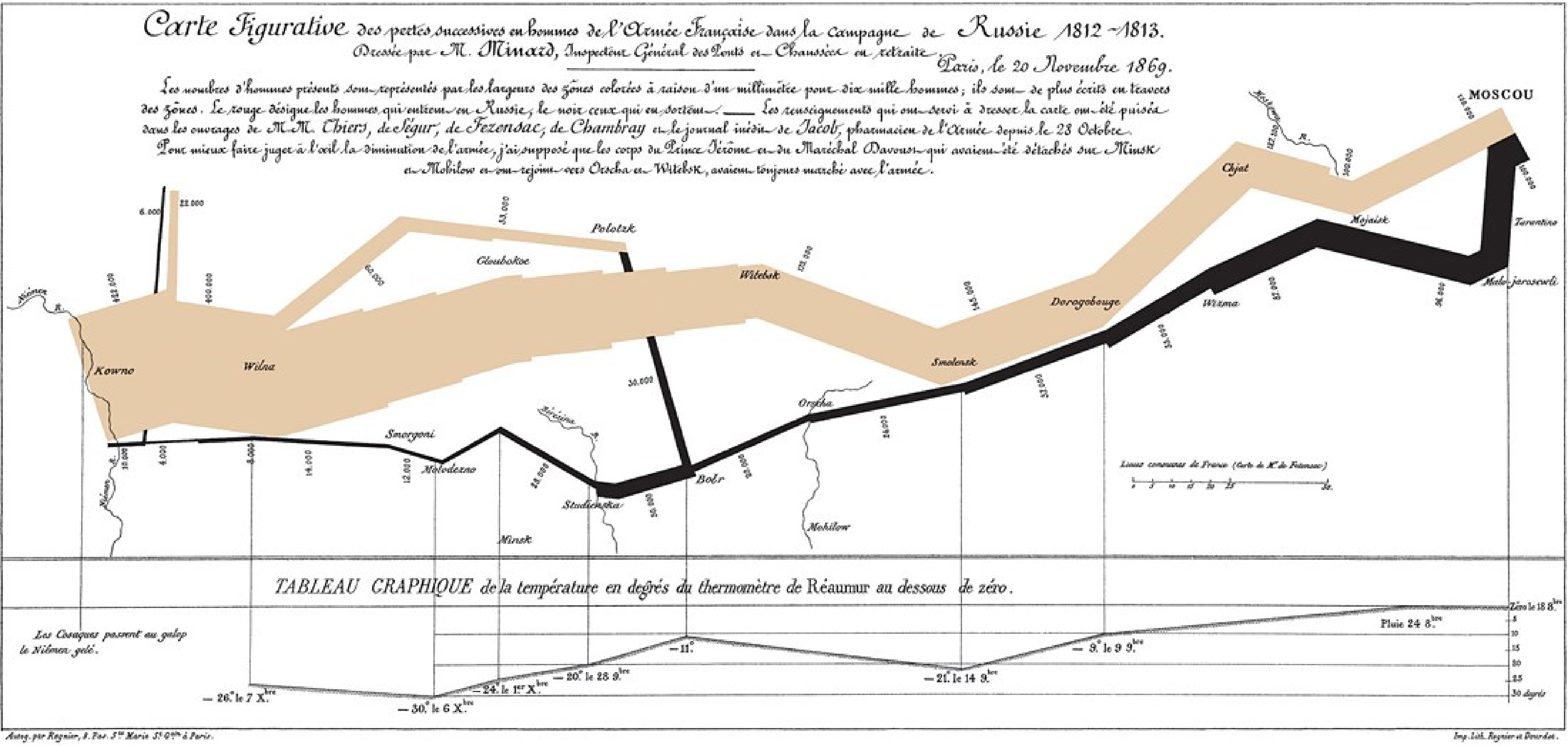
X̄
The Normal Distribution
Probability
Values
-1.96σ
1.96σ
95% of values
-2.58σ
2.58σ
99% of values
Probability of Cases in portions of the curve
≈ 0.0013
≈ 0.0214
≈ 0.1359
≈ 0.3413
≈ 0.3413
≈ 0.1359
≈ 0.0214
≈ 0.0013
Standard Deviations From The Mean
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative %
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Z Scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T Scores
20
30
40
50
60
70
80

Machine learning



http://n-chandra.blogspot.com/2013/01/picking-machine-learning-algorithm.html

Visualization



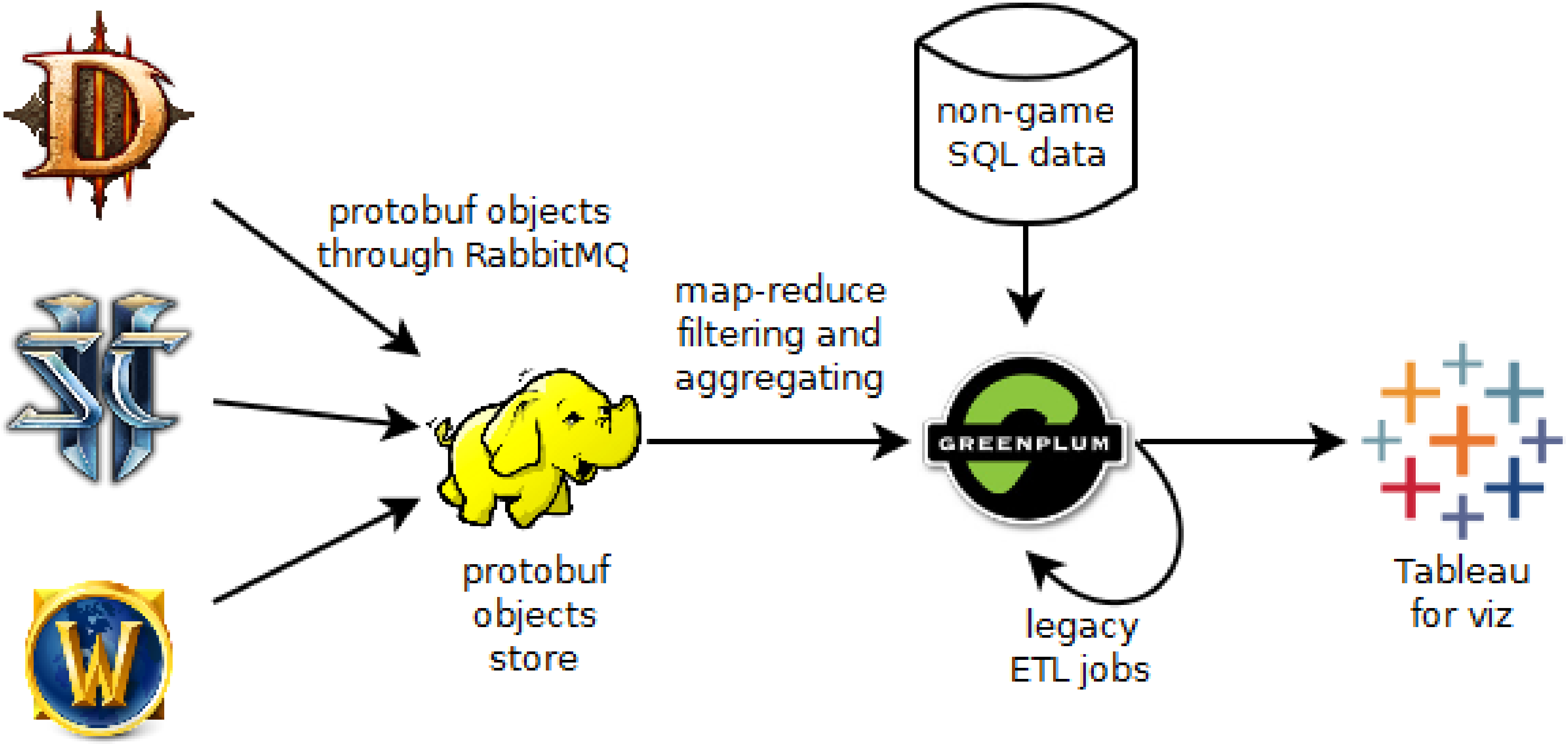
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

Data engineering

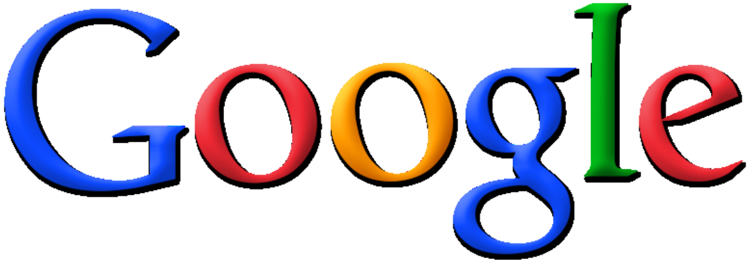


http://bowling-bash.blogspot.com/2013/10/blizzards-hadoop-platform.html

Take-aways

- Many tools and processes
- Some old, some recent
- Pick the right one for the problem at hand

INDUSTRY EXAMPLES



Google

trulia®
real estate search

Linked in

FredMeyer®

TARGET

TURNSTYLE

f

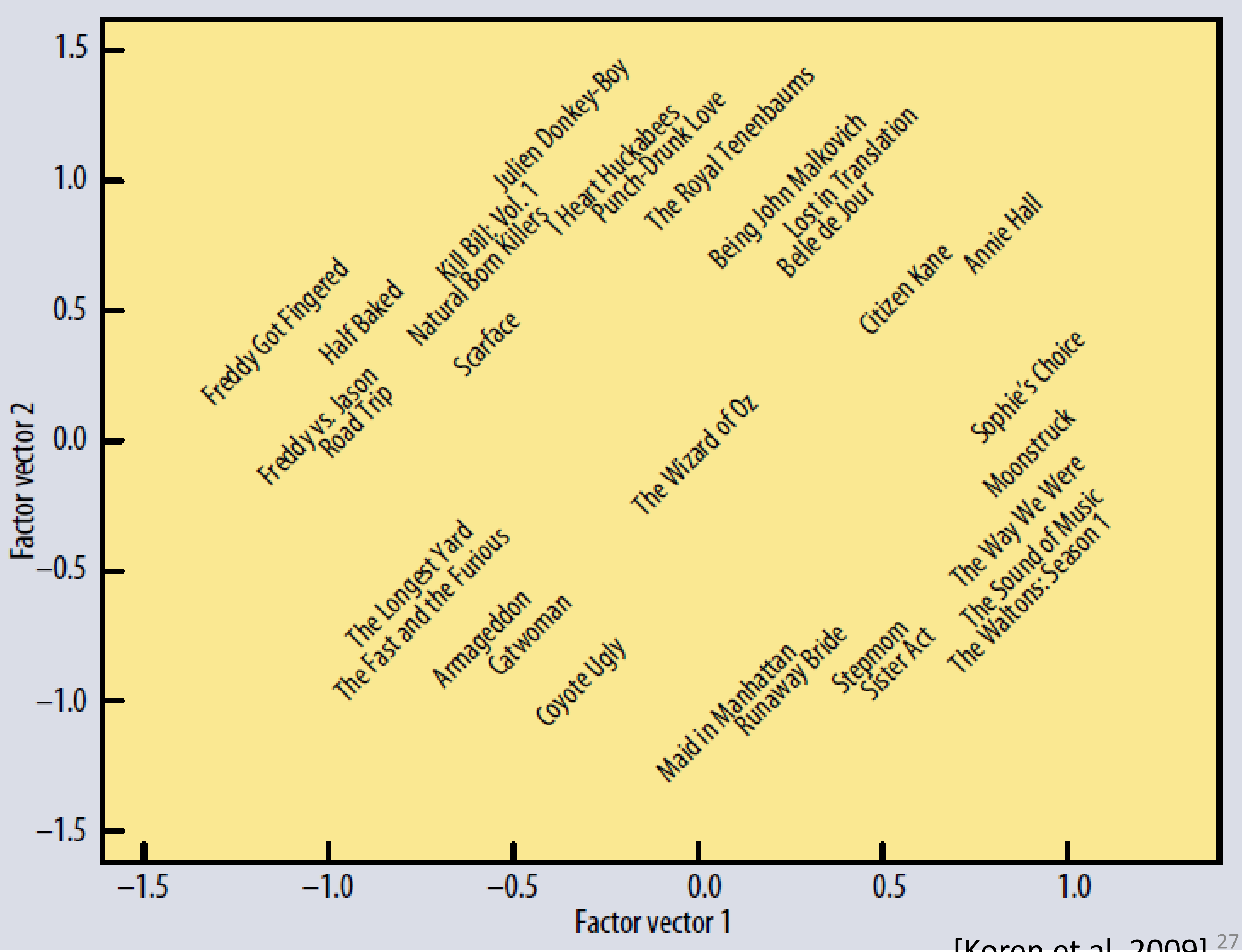
amazon.com®

NETFLIX

Take-aways

- Useful
- Everywhere
- Creepy?

SINGULAR VALUE DECOMPOSITION



[Koren et al. 2009]

Ratings data

	The Longest Yard	Citizen Kane	Fast & Furious	Kill Bill	Sound of Music
Alice	1	4	1		5
Bob	5		4	4	
Claire		4	2		4
Dan	4	1		3	
Eve		3			5
Frank	3	3	5	2	

(usually 99% sparse)

Singular Value Decomposition

	The Longest Yard	Citizen Kane	Fast & Furious	Kill Bill	Sound of Music
Alice	1	4	1	M	5
Bob	5	M	4	4	M
Claire	M	4	2	M	4
Dan	4	1	M	3	M
Eve	M	3	M	M	5
Frank	3	3	5	2	M

=

Users

*0.1	Factor 1	Factor 2
Alice	-4	7
Bob	-5	-4
Claire	-4	2
Dan	-4	-5
Eve	-4	0
Frank	-4	0

X

Tastes

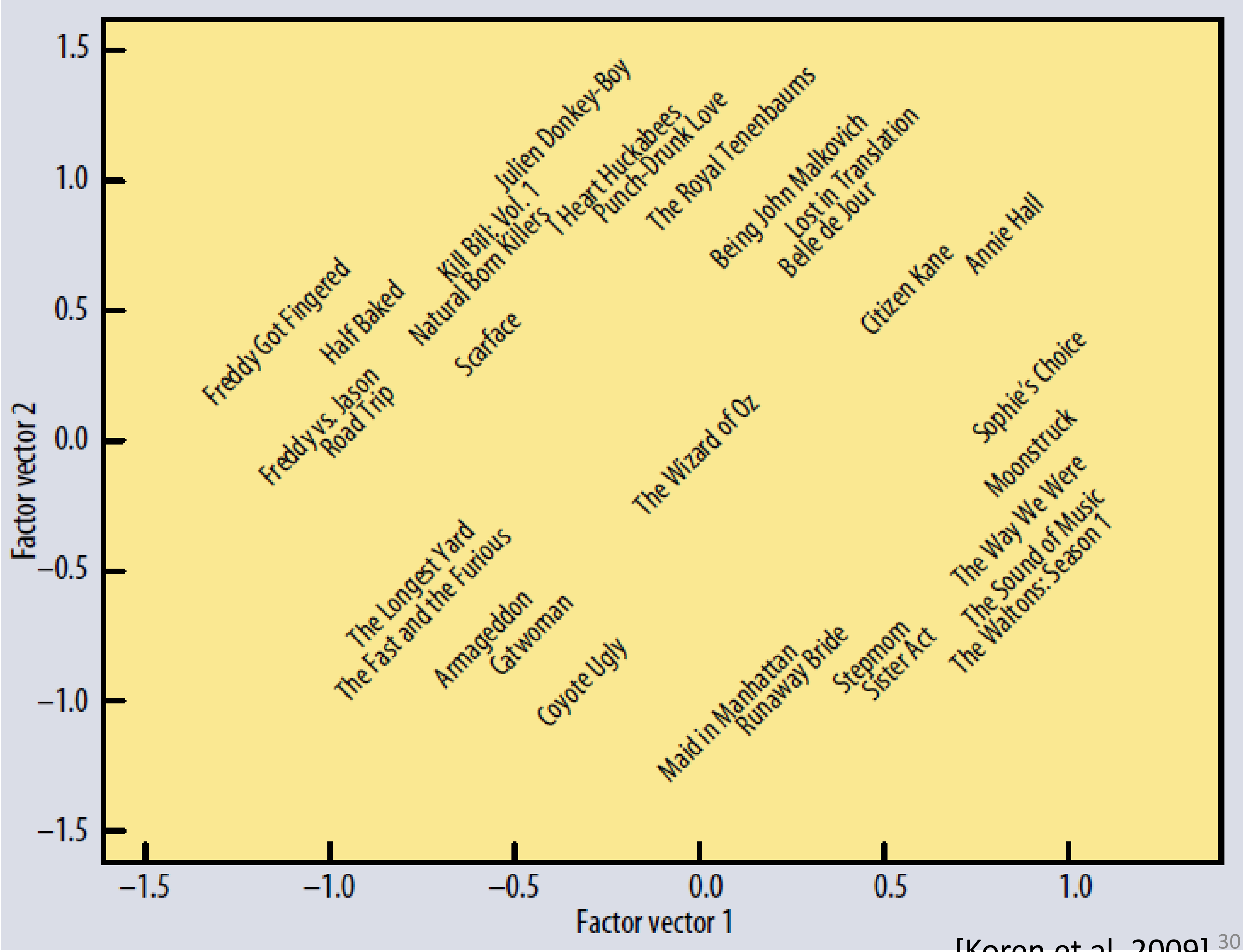
	F1	F2
F1	18	0
F2	0	4

X

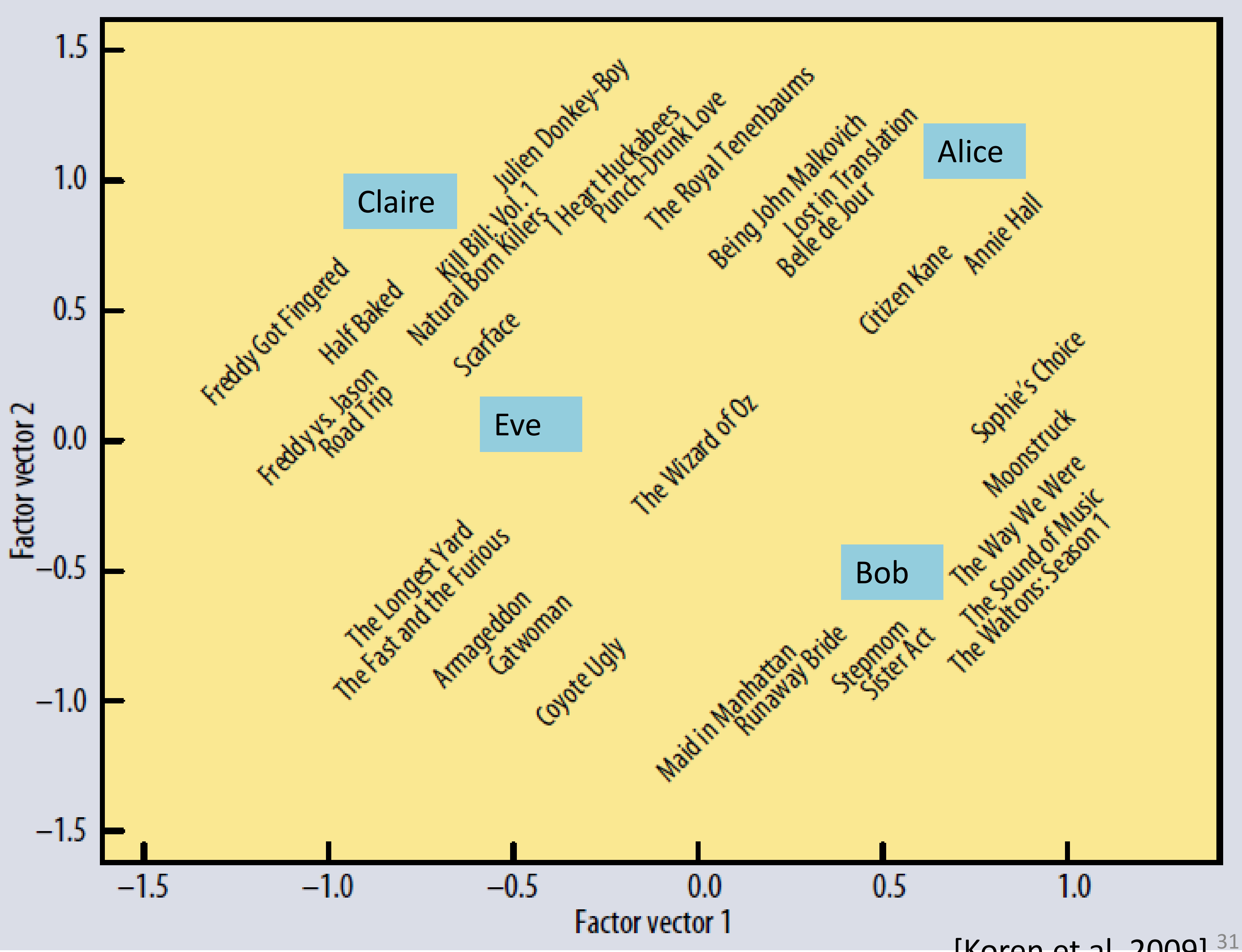
Movies

*0.1	The Longest Yard	Citizen Kane	Fast & Furious	Kill Bill	Sound of Music
F1	-4	-4	-4	-4	-5
F2	-5	4	-6	1	4

M = global average = 3.3



[Koren et al. 2009]



[Koren et al. 2009]

Densifying sparse matrices

Number of ratings x100

- Storage x100
- SVD is O(N^3), CPU x1M

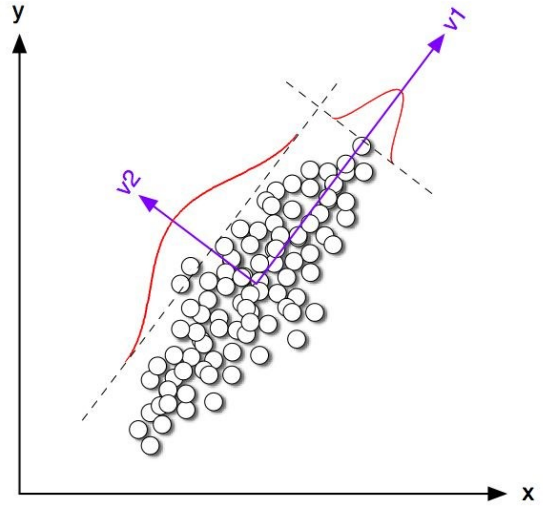
Minimize square error on known ratings using stochastic gradient descent

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(|| q_i ||^2 + || p_u ||^2)$$

	Armageddon	Citizen Kane	Fast & Furious	Kill Bill	Sound of Music
Alice	1	4	1	M	5
Bob	5	M	4	4	M
Claire	M	4	2	M	4
Dan	4	1	M	3	M
Eve	M	3	M	M	5
Frank	3	3	5	2	M

Similar techniques

- Latent Semantic Analysis
  - Movies -> documents, ratings -> tf-idf
- Latent Dirichlet Allocation
  - Cf "A theory of aspects as latent topics"
- Principal Component Analysis
  - Eigen value decomposition
  - Of the covariance matrix



Take-aways

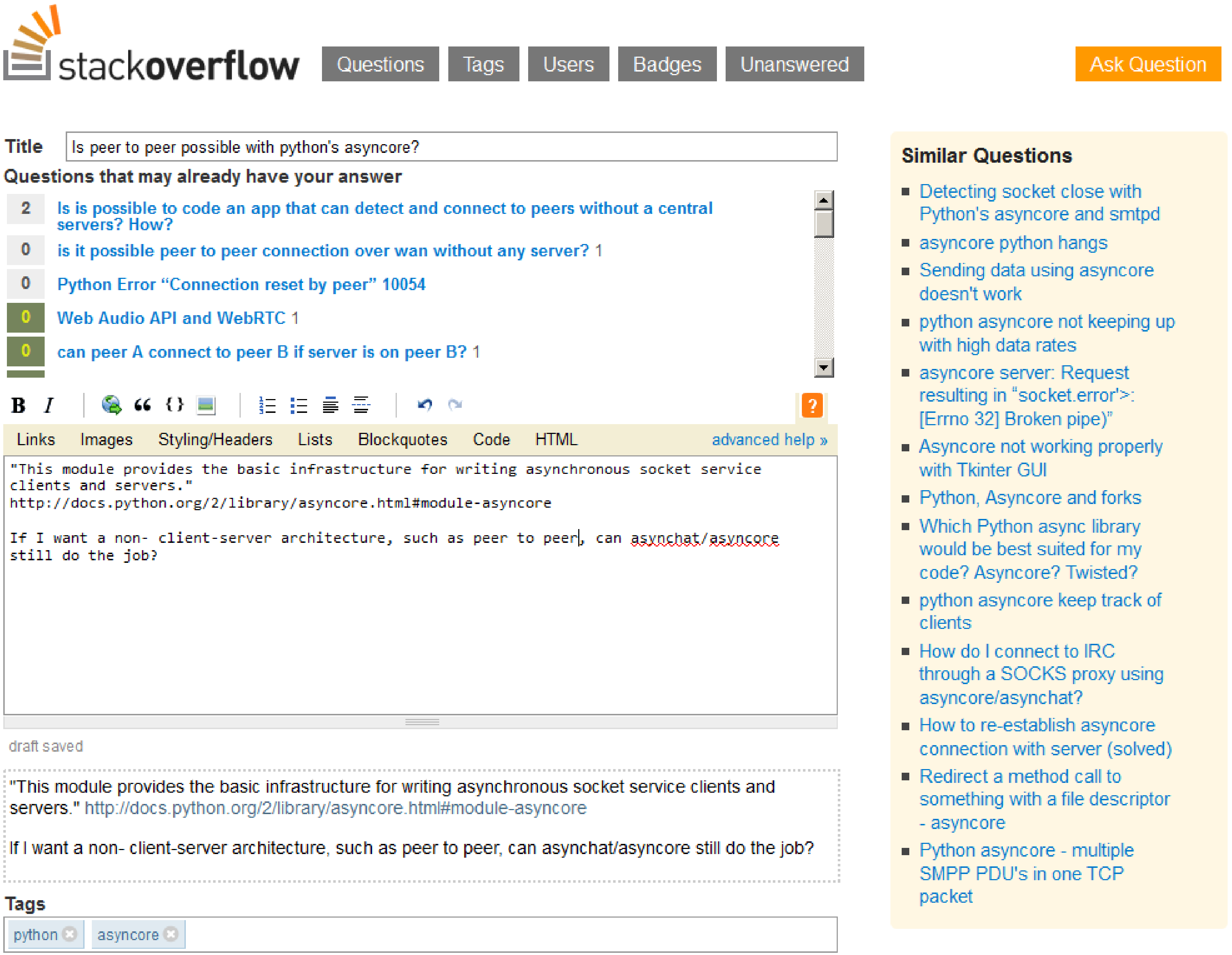
- Matrices
- Need practical methods
- The math escalates quickly
    - But you may need it …

HOW WOULD YOU DO …

Amazon

- “Frequently bought together”
  - Association rule mining
- Customer clustering
  - SVD

- (Stock management
  - Poisson process?)



Title: Is peer to peer possible with python's asyncore?

Questions that may already have your answer

2	Is is possible to code an app that can detect and connect to peers without a central servers? How?
0	is it possible peer to peer connection over wan without any server? 1
0	Python Error “Connection reset by peer” 10054
0	Web Audio API and WebRTC 1
0	can peer A connect to peer B if server is on peer B? 1

```
"This module provides the basic infrastructure for writing asynchronous socket service
clients and servers."
http://docs.python.org/2/library/asyncore.html#module-asyncore

If I want a non- client-server architecture, such as peer to peer, can asynchat/asyncore
still do the job?
```

draft saved

"This module provides the basic infrastructure for writing asynchronous socket service clients and servers." http://docs.python.org/2/library/asyncore.html#module-asyncore

If I want a non- client-server architecture, such as peer to peer, can asynchat/asyncore still do the job?

Tags

python asyncore

Similar Questions

- Detecting socket close with Python's asyncore and smtpd
- asyncore python hangs
- Sending data using asyncore doesn't work
- python asyncore not keeping up with high data rates
- asyncore server: Request resulting in “socket.error'>: [Errno 32] Broken pipe)”
- Asyncore not working properly with Tkinter GUI
- Python, Asyncore and forks
- Which Python async library would be best suited for my code? Asyncore? Twisted?
- python asyncore keep track of clients
- How do I connect to IRC through a SOCKS proxy using asyncore/asynchat?
- How to re-establish asyncore connection with server (solved)
- Redirect a method call to something with a file descriptor - asyncore
- Python asyncore - multiple SMPP PDU's in one TCP packet

Dating website

- Match %
  - SVD on weighted questions
    - “Would you rather be weird or normal?”= 99% weird
    - “Sex before marriage?”
  - SVD on tf-idf from profile essays
  - Cosine distance in SVD space
- Actual information in a profile
  - Everybody loves travelling, but they say it differently
  - “I love travelling!” = “travel” = low tf-idf
  - “Have a passport!” = “passport” = high tf-idf
  - Tf-idf is not always appropriate!

Thomas Debeauvais, tdebeauv@uci.edu

THANKS

References

- Padhraic Smyth’s slides on Netflix recsys http://www.ics.uci.edu/~smyth/courses/cs277/public_slides/recommender_systems_part2.pdf
- Koren et al. 2009: Matrix factorization techniques for recommender systems
- http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/
- http://online.wsj.com/news/articles/SB10001424052702303453004579290632128929194