Data Mining
W4240 Section 001

Prof. Giovanni Motta

Columbia University, Department of Statistics

September 9, 2015

Course Information

Course: Data Mining

Number: STAT W4240, Section 001

Course Website: Courseworks, Piazza (for message board)

Instructor: Prof. Giovanni Motta
Email: gm2554@columbia.edu
Office hours: Friday 4-5pm

Teaching Assistant: Yixin Wang
Email: yw2539@columbia.edu
Office hours: Monday, Tuesday, Wednesday, Thursday, 8-8:30am

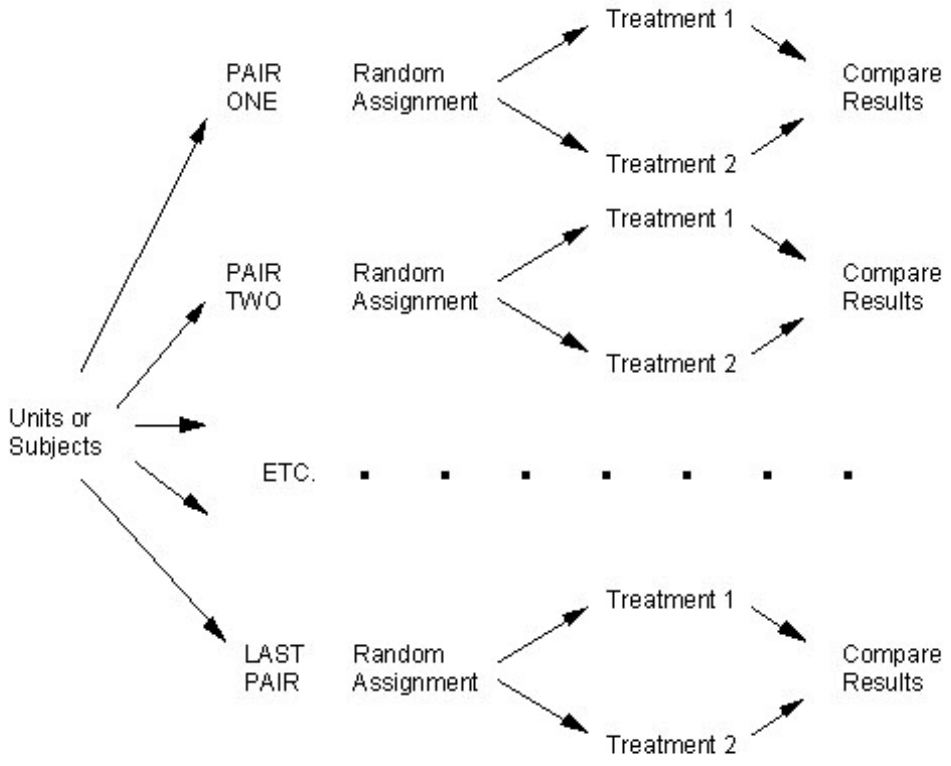
Outline

Today:

- Introduction to data mining
- Course details
- Course preparedness quiz

Data Challenges

In the classical model:



- carefully design experiments to answer a specific question
- try to get as much information as possible from data

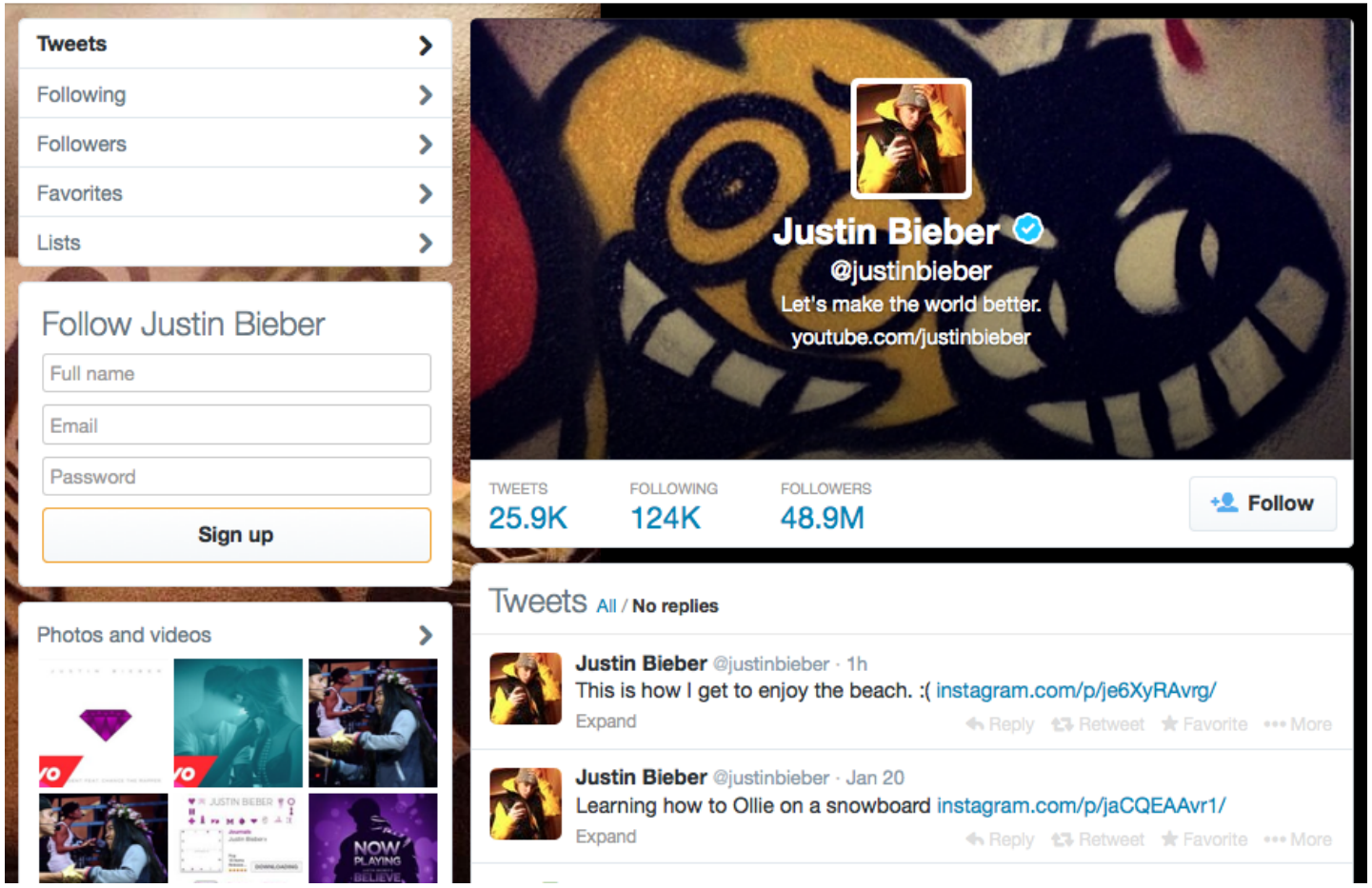
Data Challenges

In the classical model:

- get careful and expensive data (10's to 1,000's of observations)
- still the most common approach (outside the internet/marketing)

Data Challenges

Also, the internet leads to new problems:



- have observational data
- will it help me answer my question?

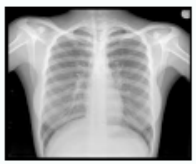
Data Challenges

Also, the internet leads to new problems:

- have *a lot* of data
- how can we make sense of it?

Data Science

- Data are everywhere.
- Data are not new.

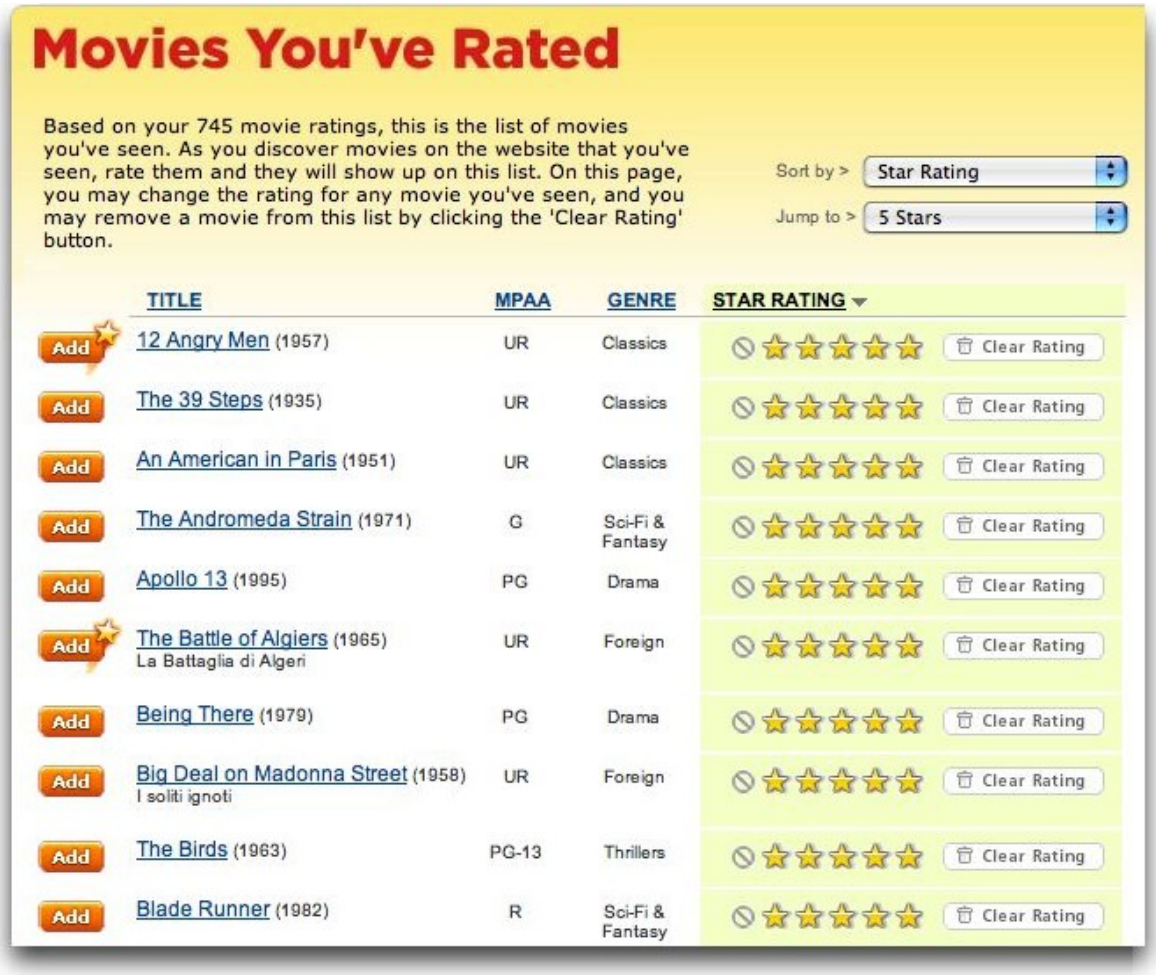
Field	Scarce Data / Easy Questions	Abundant Data / Hard Questions	Beyond
	Dawn of public markets (Dutch East India, etc.)	20th Century public markets CAPM, B.S. option pricing (Markowitz, many others)	21st Century public markets Algorithmic trading, ...
	X-ray imaging (Rontgen, others)	Computed Tomography Projection methods, Radon Transform (Vallebona, Hounsfield)	MRI, Ultrasound, PET,...

- All data can be interesting (big, small, and in between).

Shopping Histories

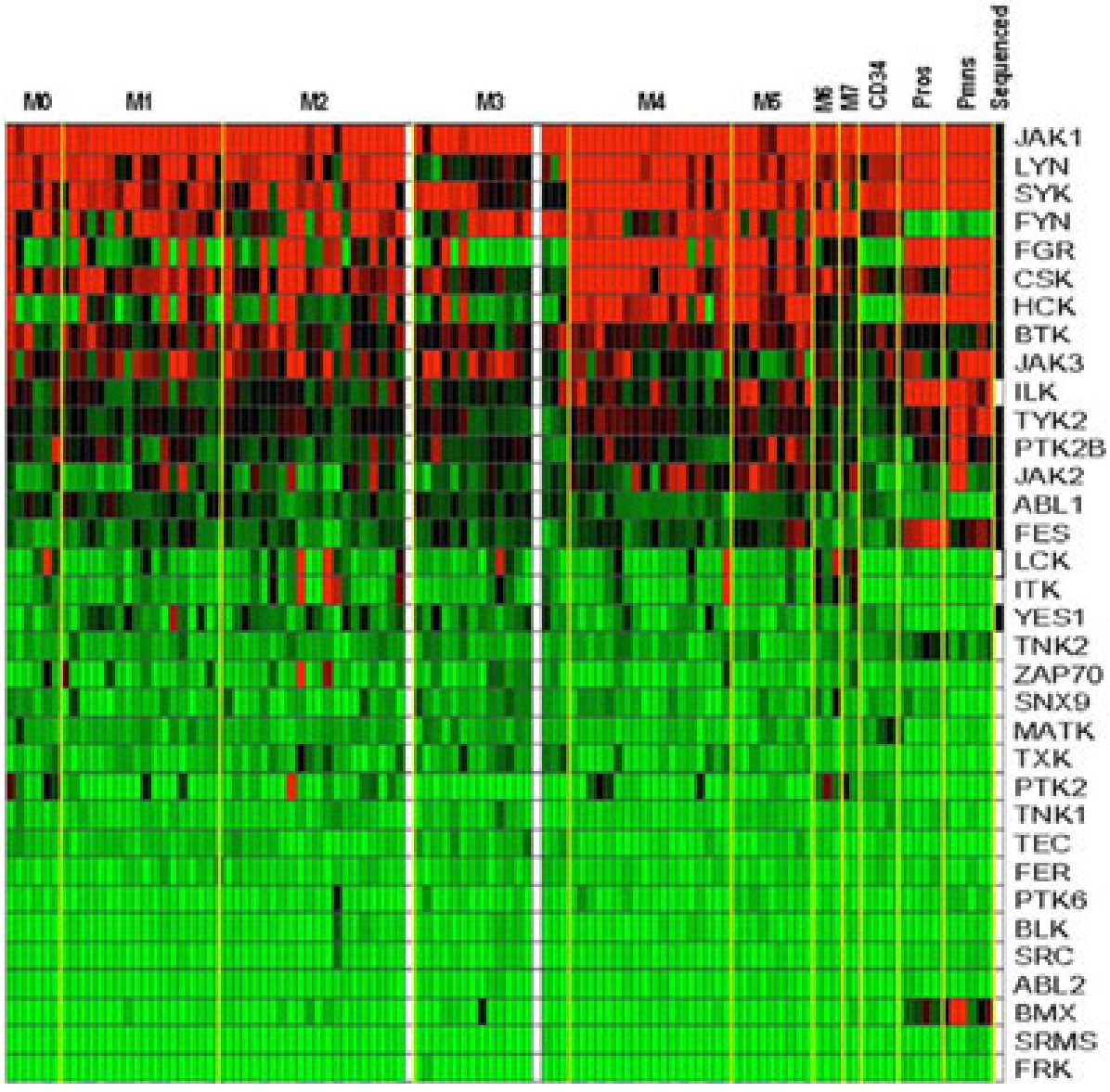
Order Date	Item	Price
August 31, 2012	Girl Genius Omnibus Volume One: Agatha Awakens	$8.97
August 25, 2012	Ito En Oi Ocha Japanese Green Tea, 16.9-Ounce Bottles (12 Pack)	$19.92
August 24, 2012	R in a Nutshell: A Quick Desktop Reference	$33.90
August 21, 2012	Creative HS-800 Fatal1ty Gaming Headset	$3225
August 21, 2012	Ensign Peak Everyday Duffel Bag, Gray	$10.59
August 16, 2012	Lucky Peach Issue 4	$7.79
July 25, 2012	Ito En Oi Ocha Japanese Green Tea, 16.9-Ounce Bottles (12 Pack)	$19.92
July 25, 2012	Feliway – Refil, 48 ml	$19.68
July 23, 2012	DRINKWELL Original Filters 12 pack	$16.95

User Ratings



Document Collections

Genomics


M0
M1
M2
M3
M4
M5
M6
M7
CD34
Pros
Pmns
Sequenced
JAK1
LYN
SYK
FYN
FGR
CSK
HCK
BTK
JAK3
ILK
TYK2
PTK2B
JAK2
ABL1
FES
LCK
ITK
YES1
TNK2
ZAP70
SNX9
MATK
TXK
PTK2
TNK1
TEC
FER
PTK6
BLK
SRC
ABL2
BMX
SRMS
FRK

Finance


Dow Jones Composite Average (^DJA) - DJI
Add to Portfolio
4,372.78 20.68 (0.47%) 1:48PM EDT
Enter name(s) or symbol(s)
GET CHART
COMPARE
EVENTS
TECHNICAL INDICATORS
CHART SETTINGS
RESET
Aug 29, 2012 2:59 PM - 3:04 PM EDT: ^DJA 4416.03
4,420
4,410
4,400
4,390
4,380
4,370
© 2012 Yahoo! Inc.
Tue Aug 28, 2012
Wed Aug 29
Thu Aug 30
Fri Aug 31
Tue Sep 4
Volume: 571,700
8.0M
6.0M
4.0M
2.0M
1D
5D
1M
3M
6M
YTD
1Y
2Y
5Y
Max
FROM: Aug 28 2012
TO: Sep 4 2012
-0.97%
1980
1985
1990
1995
2000
2005
2010

Data can help us solve specific problems.

How should these pictures be placed into 3 groups?









How should these pictures be placed into groups? How many groups should there be?

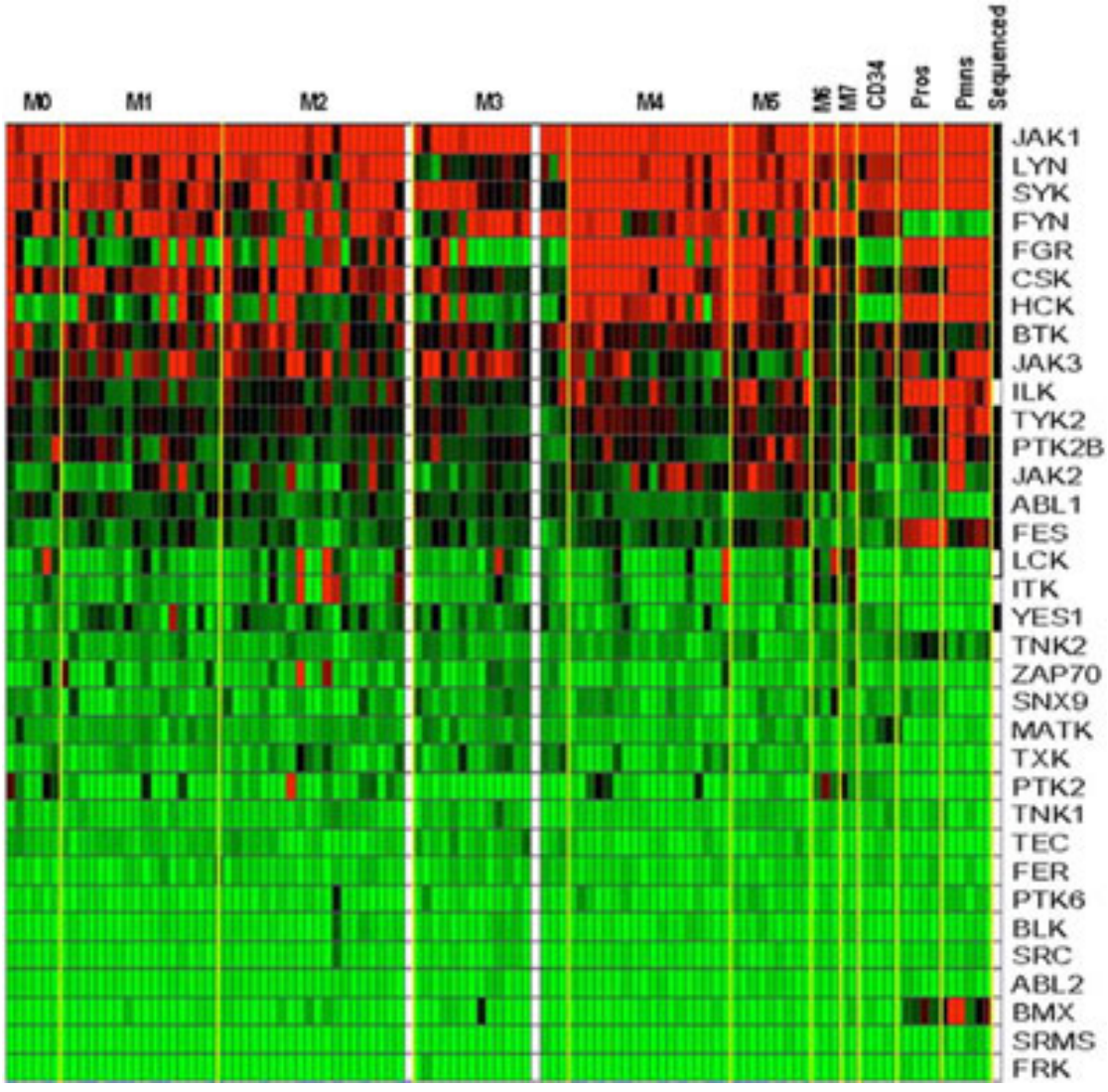








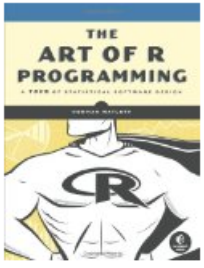
Which genes are associated with a disease? How can expression values be used to predict survival?



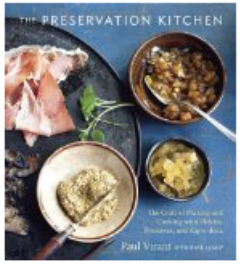
What items should Amazon display for me?

Books

Page 2 of 20 (Start over)

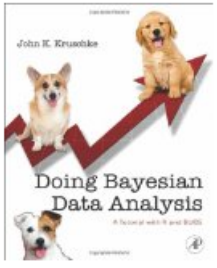


The Art of R Programming
Norman S. Matloff
★★★★½ (22)
Paperback
~~$39.95~~ $24.32
Why recommended?

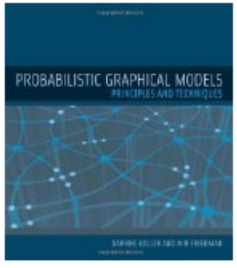


The Preservation Kitchen
Kate Leahy
★★★★½ (19)
Hardcover
~~$29.99~~ $18.74
Why recommended?

Agatha H. and the Clo...
Phil Foglio
★★★★★ (9)
Hardcover
~~$24.99~~ $16.44
Why recommended?



Doing Bayesian Data A...
John K. Kruschke
★★★★★ (18)
Hardcover
~~$89.95~~ $64.15
Why recommended?



Probabilistic Graphic...
Daphne Koller
★★★★☆ (13)
Hardcover
~~$99.00~~ $84.25
Why recommended?

› See all recommendations in Books

Is this spam?

hi backpackers,

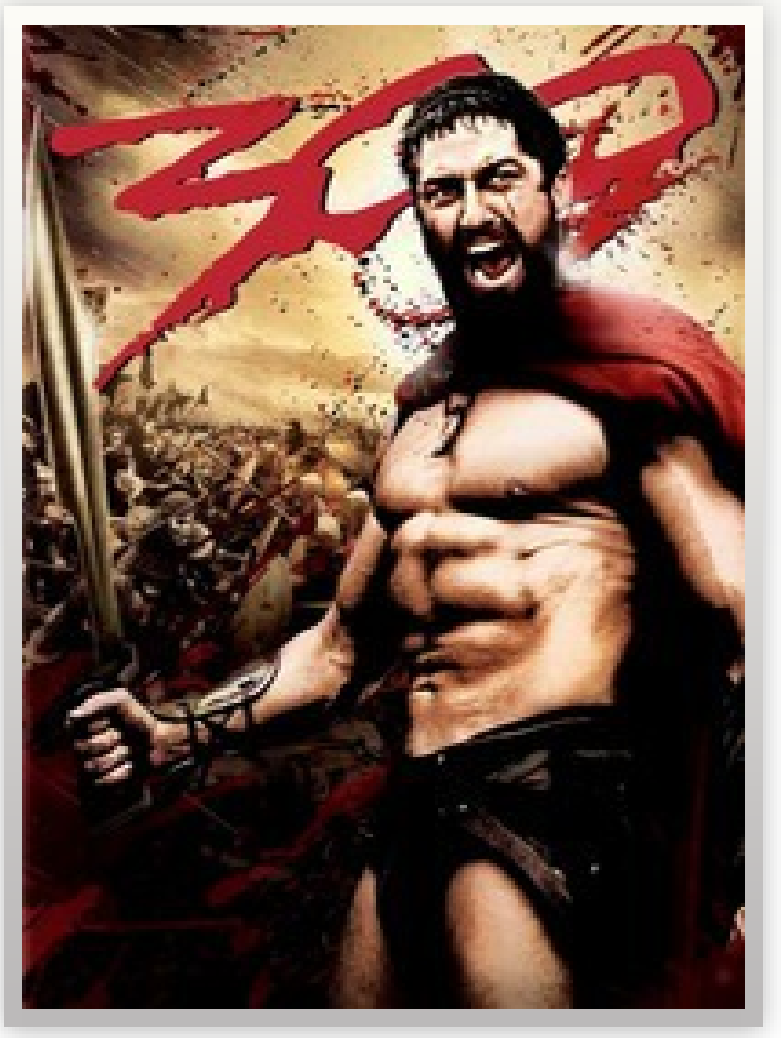
i saw that close to my hotel there is a pub with bowling (it's on market between 9th and 10th avenue). are you up to it? i think it is about 20 years i haven't played... if you like the idea what about 8.30 there?

otherwise any suggestion welcome. i can survive another 20 years without bowling.

What about this?

```
Enter for a chance to win a trip to Universal Orlando
Resort To celebrate the arrival of Dr.  Seuss The
Lorax on Movies On Demand on August 21st, were
offering you a chance to win a trip to Universal
Orlando Resort.  Head to www.facebook.com/twc
starting 8/21 for more details.  Enter now
```

Will I like *300*?



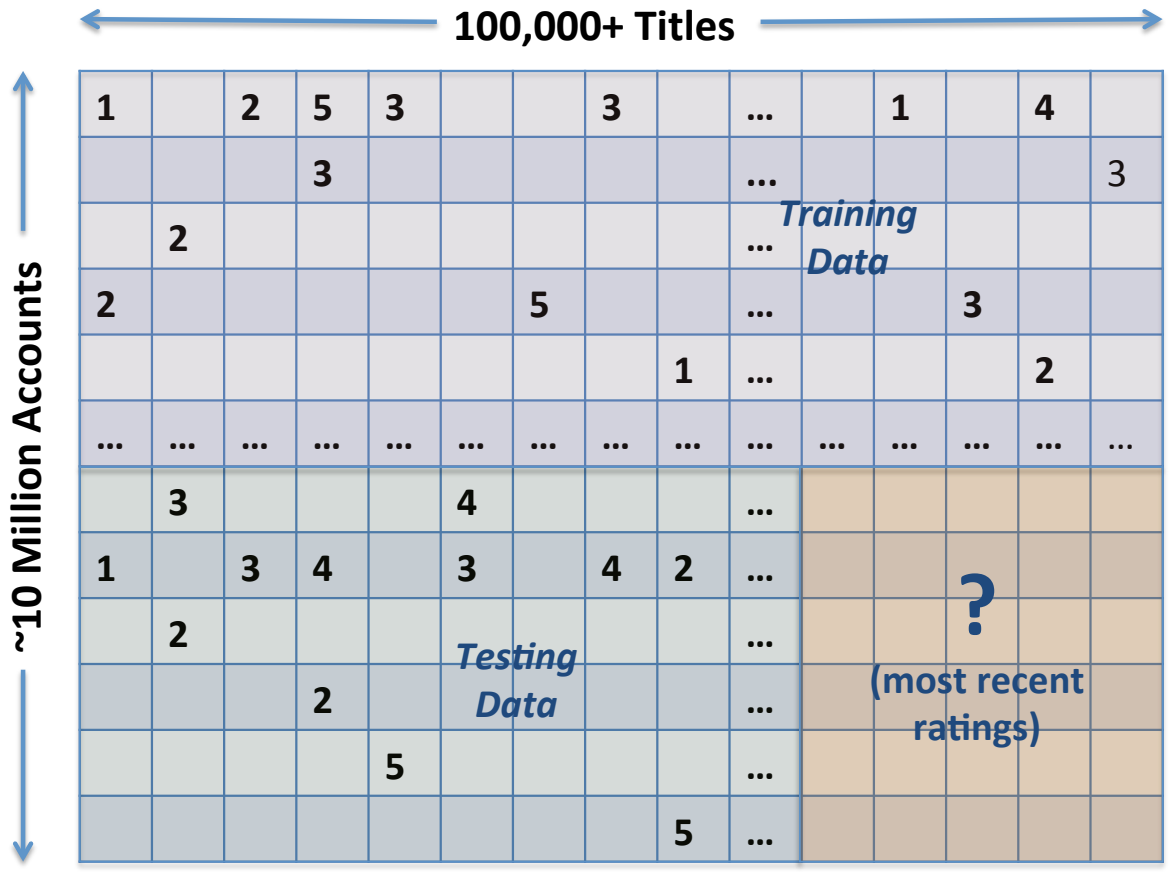
Will I like *300*? How would Netflix know?

★★★★★

★★★★☆

★★★★☆

How would Netflix know? What makes them think their prediction is accurate?



Data can help us solve these problems.

W4240: Data Mining

What is data mining?

- "The process of extracting patterns from data." (Wikipedia)

W4240: Data Mining

What is data mining?

- "The process of extracting patterns from data." (Wikipedia)
- "Searching large volumes of data looking for patterns that accurately predict behavior in customers and prospects." (Adobe)

W4240: Data Mining

What is data mining?

- "The process of extracting patterns from data." (Wikipedia)
- "Searching large volumes of data looking for patterns that accurately predict behavior in customers and prospects." (Adobe)
- "The practice of compiling information about Internet users by tracking their motions through Web sites, recording the time they spend there, what links they click on and other details, usually for marketing purposes." (Consumer Privacy Guide)

W4240: Data Mining

In this class you will study algorithms that exploit patterns in data, using:

- machine learning
- statistics
- computer science
- data mining

Applications include:

- natural science (e.g. genomics)
- image processing
- digital humanities
- finance
- many others

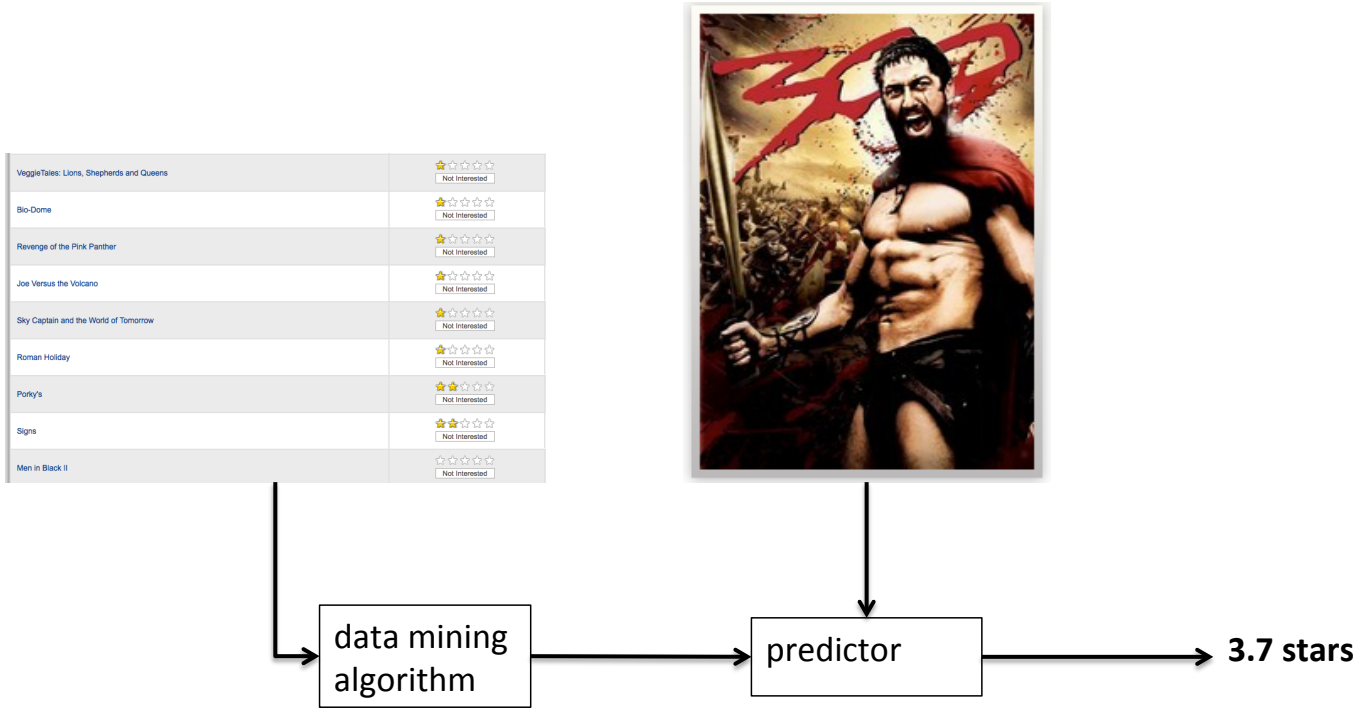
W4240: Data Mining

In this class you will study algorithms that exploit patterns in data.

- Goal: learning how to think about problems in data mining
- You will learn a set of data analysis tools, including
  - how to use them
  - the assumptions they make
  - their capabilities and limitations

Formula for Class

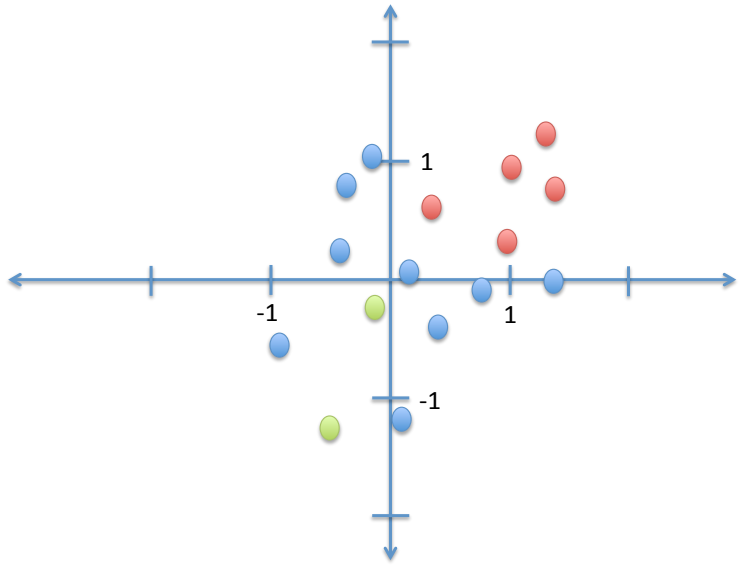
1. Get data.
2. Analyze it to get a pattern.
3. Use the pattern to do something.



Basic Ideas Used Throughout Class

- Supervised learning
- Unsupervised learning
- Discrete data
- Continuous data
- Computational efficiency

Supervised Learning

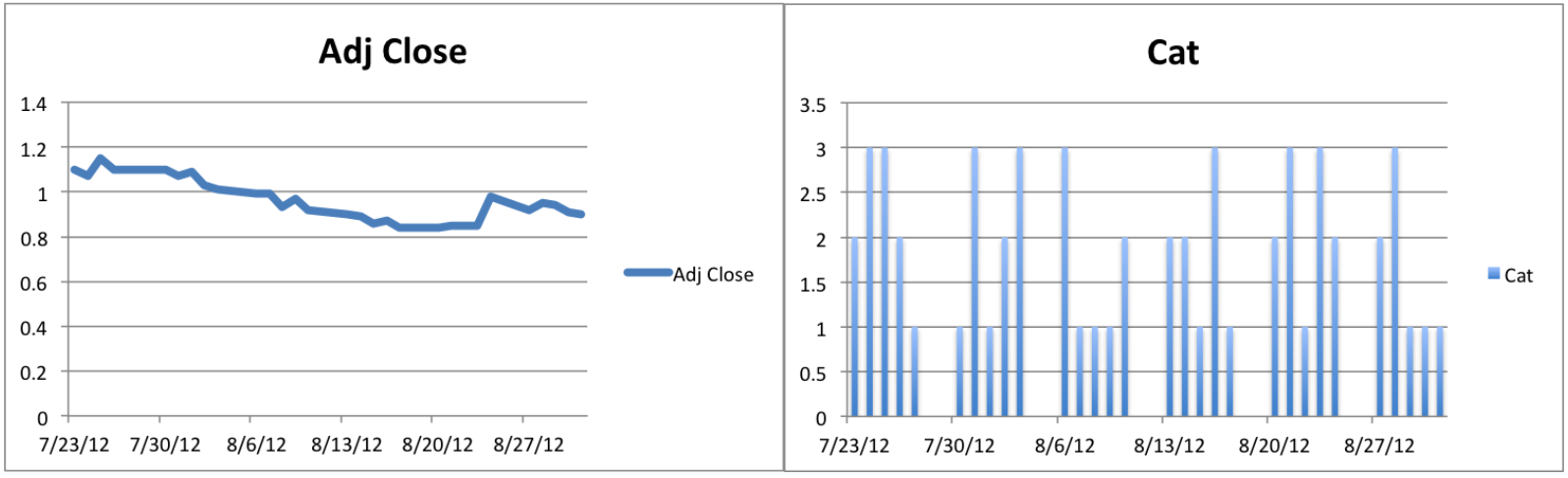


- **Supervised data** have labels or response values
- Find patterns in the **fully observed** data and apply them to partially observed data
- Example: a collection of emails labeled "spam" and "ham"
  - find a rule to separate spam from ham in labeled data
  - apply to new (unlabeled) emails

Unsupervised Learning

- **Unsupervised data** do not have labels
- We want to find a hidden structure that is never fully observed
- Example:
  - find groups in a collection of images
  - develop a set of topics to describe a corpus of documents
- Evaluation is difficult

Discrete vs. Continuous Data



- **Continuous variables** have continuous values (e.g. in [0,1] or $(-\infty, \infty)$)
  - stock returns
  - temperature
- **Discrete variables** take only a finite set of values (e.g. categories)
  - "spam" and "ham"
  - "fraud", "not fraud", and "unclear"

Problem Types (Generally)

	Continuous	**Categorical**
Supervised	Regression	Classification
Unsupervised	Dimension Reduction	Clustering

Course Details.

Books

Readings come from the following texts:

- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning* Springer, 2014.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition.* Springer, 2009.
- Torgo, L. *Data Mining with R.* CRC Press, 2011.

Other useful books:

- Adler, J. *R in a Nutshell: A Desktop Quick Reference.* O'Reilly Media, 2010.
- Bishop, C. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006.
- Witten, I. H., Frank, E. and Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman, 2011.
- Wu, X. and Kumar, V., eds. *Top Ten Algorithms in Data Mining.* CRC Press, 2009.

Communications

To pool resources, class discussions will be *online* on Piazza at:
`https://piazza.com/class/hq76cfqlygz6no?cid=6`

- course materials, like data sets and lectures, will be posted on Courseworks
- discussion for both sections is on Piazza
- if you have a question, post it on Piazza
- questions about course: try the book first! If the answer isn't there, as the forum. If the forum does not have an answer, go to office hours and ask the TA.

Grading

- **Homeworks (30 %).** Homeworks contain both written and `R` data analysis elements.
- **Midterm Exam (30 %).** This exam is closed-notes and closed-book.
- **Final Exam (40 %).** This exam is closed-notes and closed-book.

Homework: Due online on Courseworks **BEFORE CLASS**.

Late Work: No late work is accepted.

Academic Integrity

Cheating is not tolerated.

Homework:

- you may work in groups to solve problems, but the writeup and code must be your own
- if you violate this policy, you will get a 0 on the homework at issue ($\approx 1/3$ of a letter grade)

Any violations may also be reported to the GSAS office of Academic Integrity

R

We will be using R to implement data mining algorithms.

Why R?

- R is free, and many statistics courses at Columbia use it
- R is considered standard software for statisticians
- MATLAB is considered standard software for engineers
- python is considered standard software for ...

R

We will be using R to implement data mining algorithms.

Why R?

- R is free, and many statistics courses at Columbia use it
- R is considered standard software for statisticians
- MATLAB is considered standard software for engineers
- python is considered standard software for ...
- Why not?

R for W4240

Course expectations:

- part of the course is becoming proficient in R
- there will be some R taught in lecture
- however, you will only learn a programming language through practice

Practicalities:

- next lecture will be an intro to R
- install the latest version on your computer and follow along
- download from cran.r-project.org

R for W4240

Homework 1 is designed as an intro to R and data manipulation.

I am having trouble with R. What should I do?

- search the internet
- search the internet
- search the internet
- ask your friends
- post on the discussion board on Piazza
  - search threads for your question
  - if not posted, start a new thread with your question
  - participation (posting good questions and solutions) may help if your grade is borderline
- go to office hours!

Homework

Homework will include exercises from the book and some external data-based questions:

- we will implement 2 projects over the semester in the homework: facial recognition for the Yale Faces B data set and document analysis for the Federalist Papers
- facial recognition:
  - homeworks 1, 2 & 3, before midterm
  - learn methods for image processing
  - learn PCA and k-nearest neighbors
  - learn Discriminant Analysis and naive Bayes
- document analysis:
  - homeworks 4, 5 & 6 after midterm
  - learn methods for text mining
  - learn Bootstrap, Subset Selection, and Shrinkage
  - learn Trees, boosting, SVM and clustering

Course Preparedness

We assume that you know:

- calculus
- linear algebra
- introductory probability and statistics
- how to code elementary algorithms without drama

You can get a reasonable grade in the class without all of that background, but it will require a **significant** amount of work beyond the **significant** amount of work the class asks of you.

For Next Time

- get familiar with Courseworks and Piazza
- download `R` and bring your computer.
- get HW01

Now: quiz

- I am passing out a quiz for you to assess your readiness. Solutions will be posted on Courseworks at the end of the week. (note: this should remind you that this class will be about math and programming)

- Let's first read through the explanation together.