Interacting with Data

David M. Blei

COS424
Princeton University

February 12, 2008

Data are everywhere.

User ratings

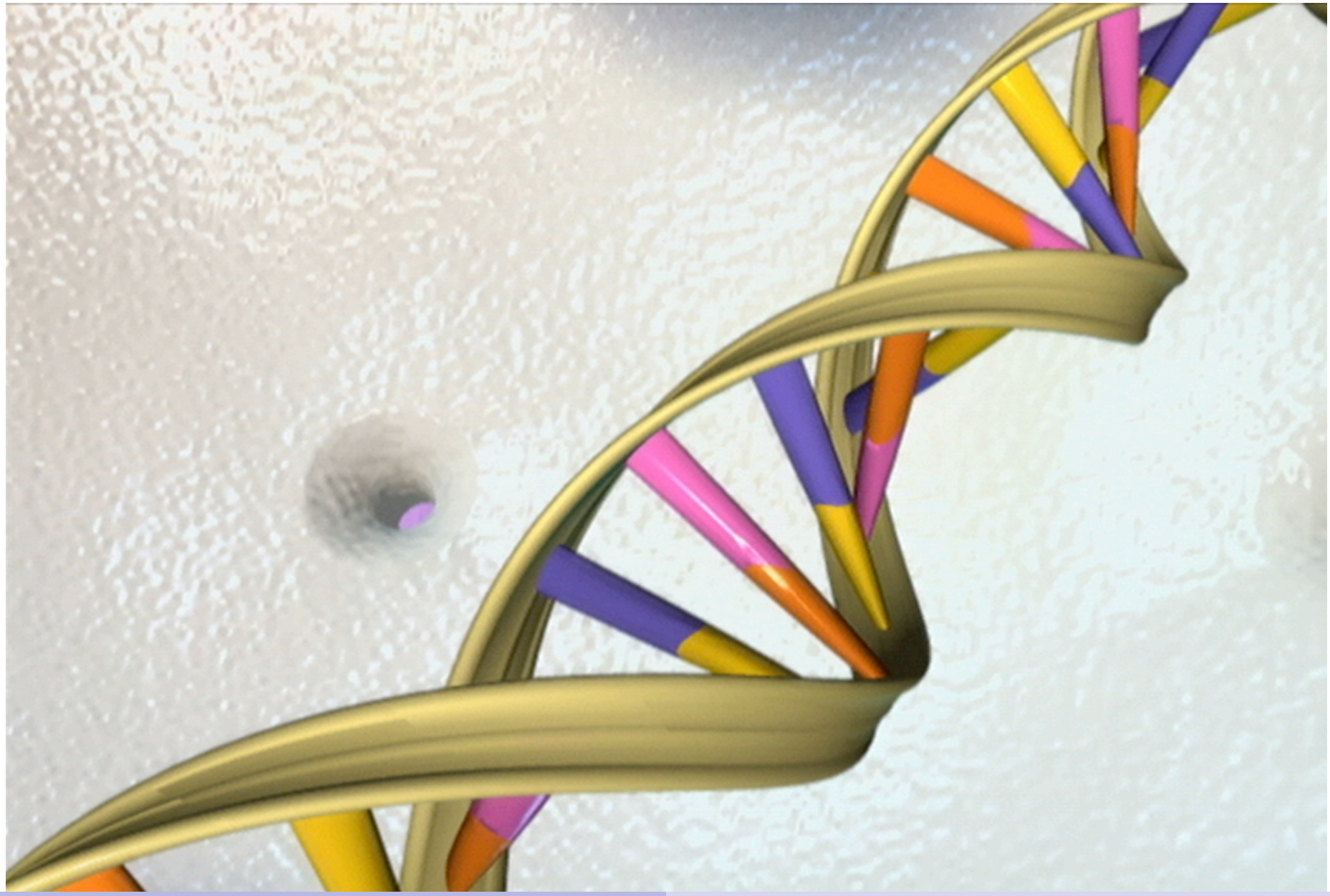
Ikiru (1952)	UR	Foreign	
Junebug (2005)	R	Independent	
La Cage aux Folles (1979)	R	Comedy	
The Life Aquatic with Steve Zissou (2004)	R	Comedy	
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure	
Lost in Translation (2003)	R	Drama	
Love and Death (1975)	PG	Comedy	
The Manchurian Candidate (1962)	PG-13	Classics	
Memento (2000)	R	Thrillers	
Midnight Cowboy (1969)	R	Classics	

Purchase histories

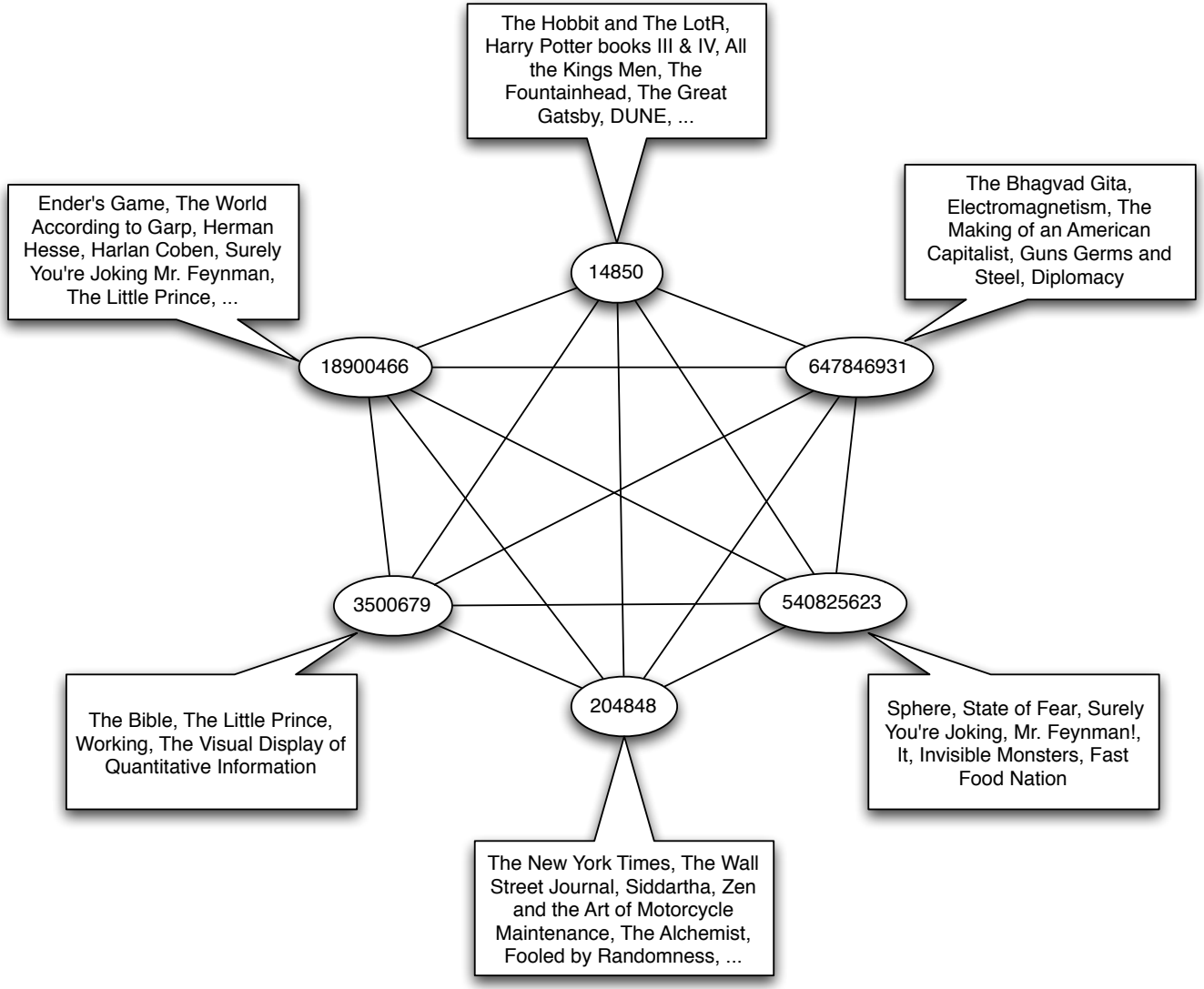
	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
1/1	**Stonyfield Farm Organic Lowfat Plain Yogurt** (32oz)		$3.59/ea	**$3.59**	
	Fruit				
3/3	**Anjou Pears** (Farm Fresh, Med)	1.76 lb	$2.49/lb	**$4.38**	
2/2	**Cantaloupe** (Farm Fresh, Med)		$2.00/ea	**$4.00**	**S**
	Grocery				
1/1	**Fantastic World Foods Organic Whole Wheat Couscous** (12oz)		$1.99/ea	**$1.99**	
1/1	**Garden of Eatin' Blue Corn Chips** (9oz)		$2.49/ea	**$2.49**	
1/1	**Goya Low Sodium Chickpeas** (15.5oz)		$0.89/ea	**$0.89**	
2/2	**Marcal 2-Ply Paper Towels, 90ct** (1ea)		$1.09/ea	**$2.18**	**T**
1/1	**Muir Glen Organic Tomato Paste** (6oz)		$0.99/ea	**$0.99**	
1/1	**Starkist Solid White Albacore Tuna in Spring Water** (6oz)		$1.89/ea	**$1.89**	

Document collections

Genomics



Social networks



Data are useful.

Will NetFlix user 493234 like Transformers?

Will NetFlix user 493234 like Transformers?

Group these images into 3 groups









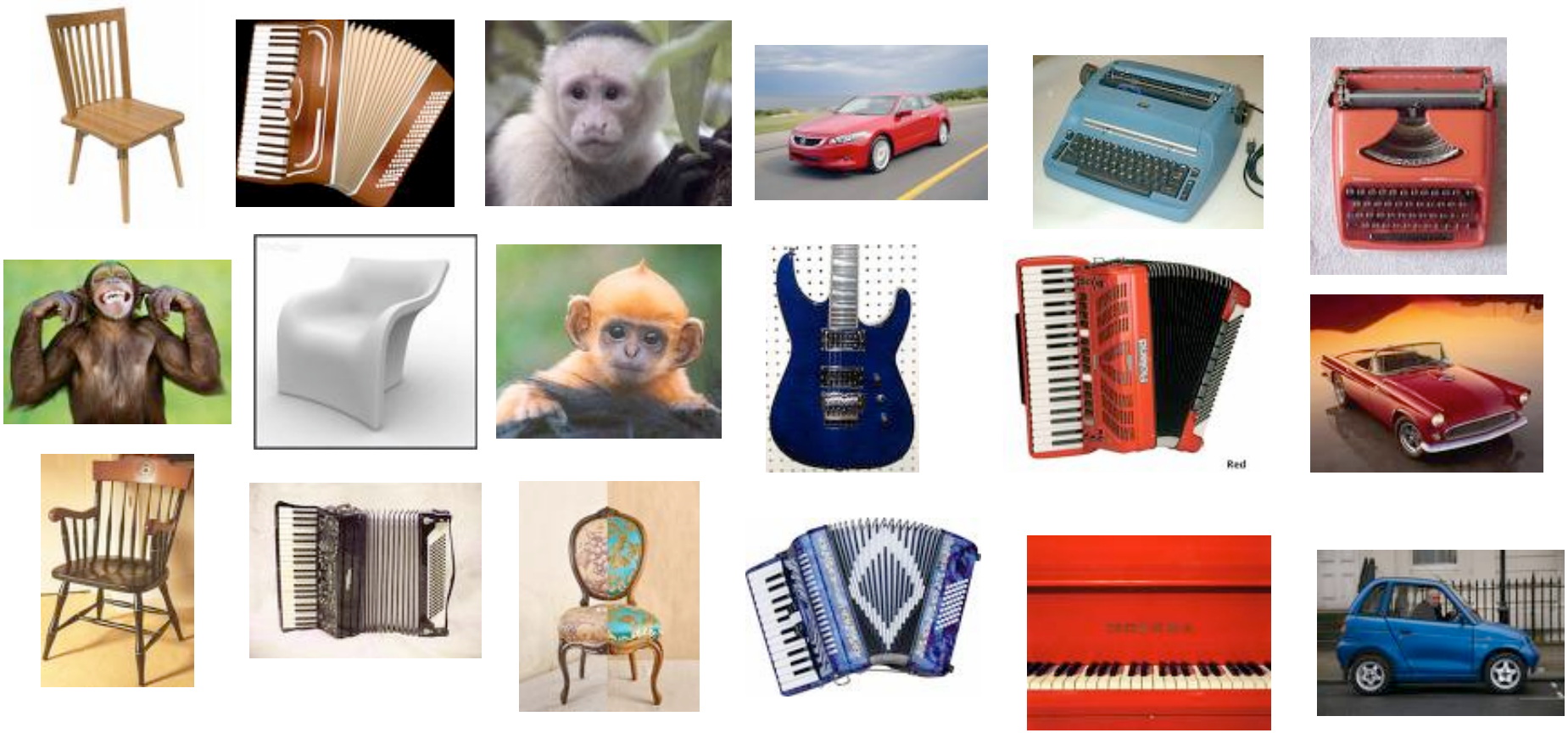
Group these images into 2 groups ... into 3 groups








Rank these images...



- ...according to relevance to `instrument`.
- ...according to relevance to `machine`

Is this spam?

Subject: CHARITY.
Date: February 4, 2008 10:22:25 AM EST
To: undisclosed-recipients:;
Reply-To: s.polla@yahoo.fr

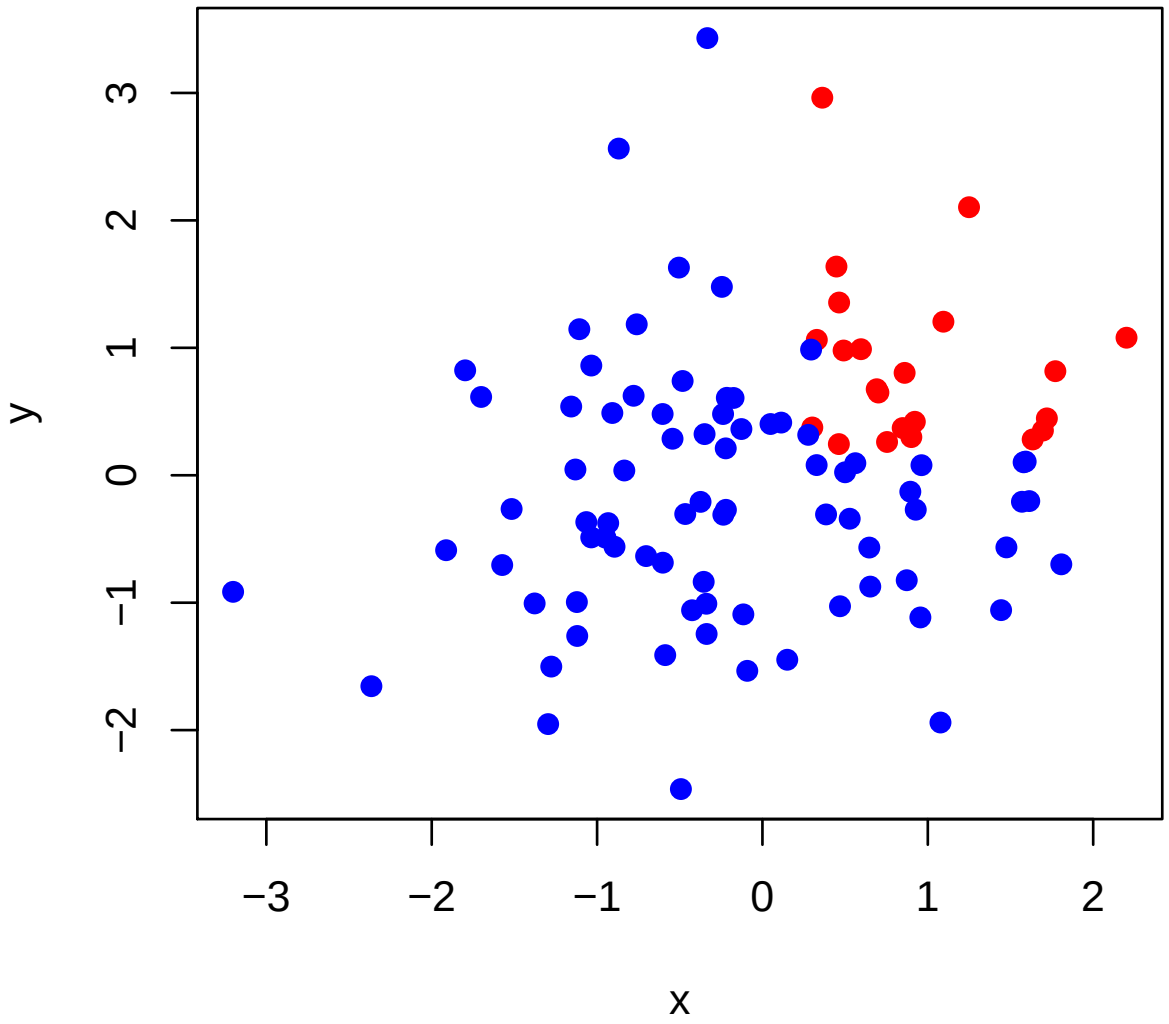
Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

How about this one?

From: [snipped]
Subject: Superbowl?
Date: January 30, 2008 8:09:00 PM EST
To: blei@cs.princeton.edu, [snipped]

Anyone interested in coming by to watch the game? Beer and pizza, I'd imagine. If anyone wants, we could get together earlier, play a board game or cards or roll up characters or something. Takers?

Label a new point



Data contain patterns.

COS424

- **Studying algorithms that find and exploit the patterns in data**
- These algorithms draw on ideas from
  - *machine learning*,
  - *artificial intelligence*
  - *applied statistics*
  - *optimization*
  - *probability theory*
- Applications include
  - natural science (e.g., genomics)
  - web technology (e.g., Google, NetFlix)
  - finance (e.g., stock prediction)
  - and many others

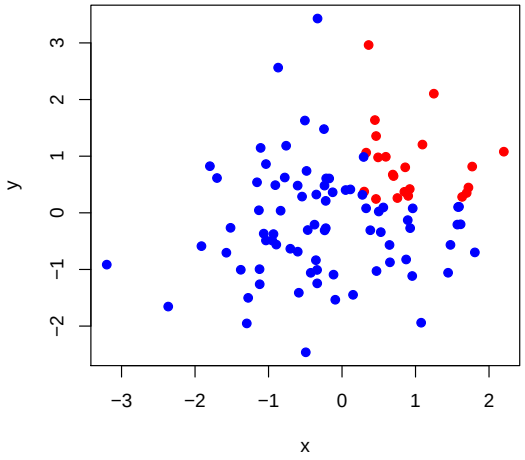
Basic idea behind everything we will study

Ikiru (1952)	UR	Foreign
Junebug (2005)	R	Independent
La Cage aux Folles (1979)	R	Comedy
The Life Aquatic with Steve Zissou (2004)	R	Comedy
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure
Lost in Translation (2003)	R	Drama
Love and Death (1975)	PG	Comedy
The Manchurian Candidate (1962)	PG-13	Classics
Memento (2000)	R	Thrillers
Midnight Cowboy (1969)	R	Classics

learning algorithm → predictor → 4.3 stars

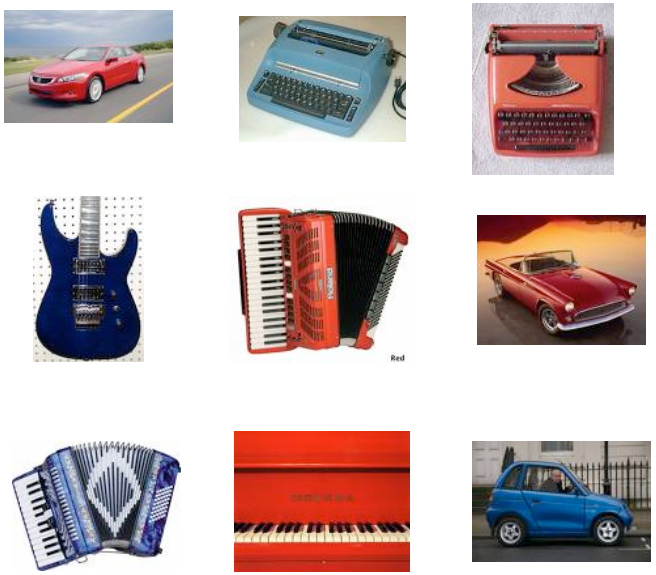
1. Take some data
2. Analyze it
3. Use it to do something

Supervised vs. unsupervised methods



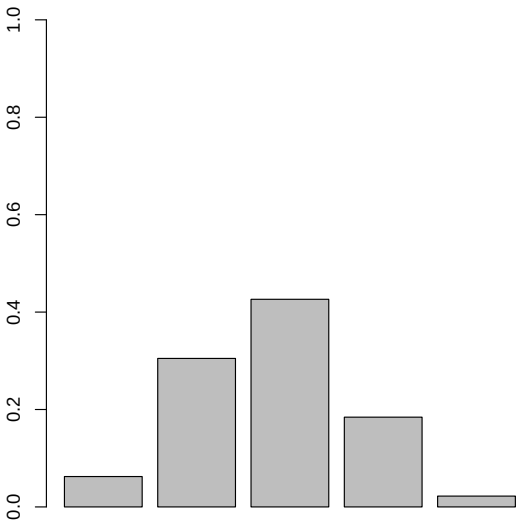
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email and automatically categorize it.

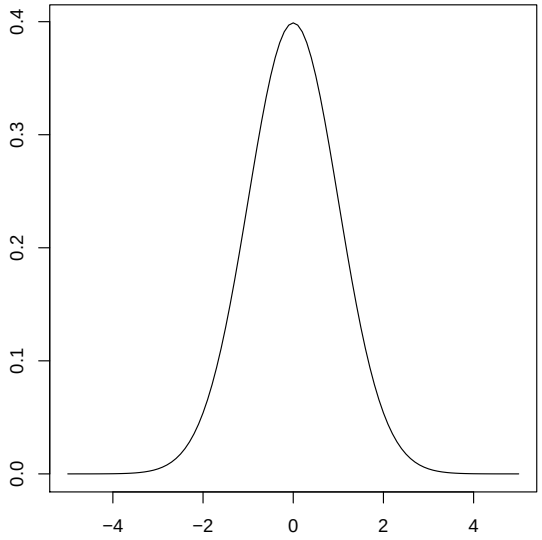
Supervised vs. unsupervised methods



- **Unsupervised methods** find **hidden structure** in data, structure that we can never formally observe.
- E.g., a museum has images of their collection that they want grouped by similarity into 15 groups.
- Unsupervised learning is more difficult to evaluate than supervised learning. But, these kinds of methods are widely used.

Discrete vs. continuous methods





- Discrete methods manipulate a finite set of objects
  - e.g., classification into one of 5 categories.
- Continuous methods manipulate continuous values
  - e.g.,prediction of the change of a stock price.

One useful grouping

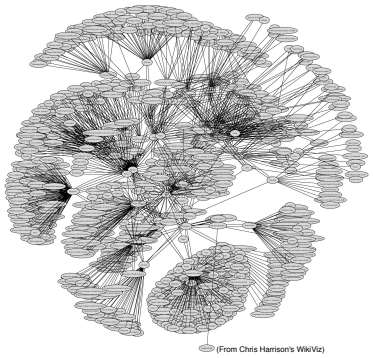
	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Data representation

$\longrightarrow \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle$

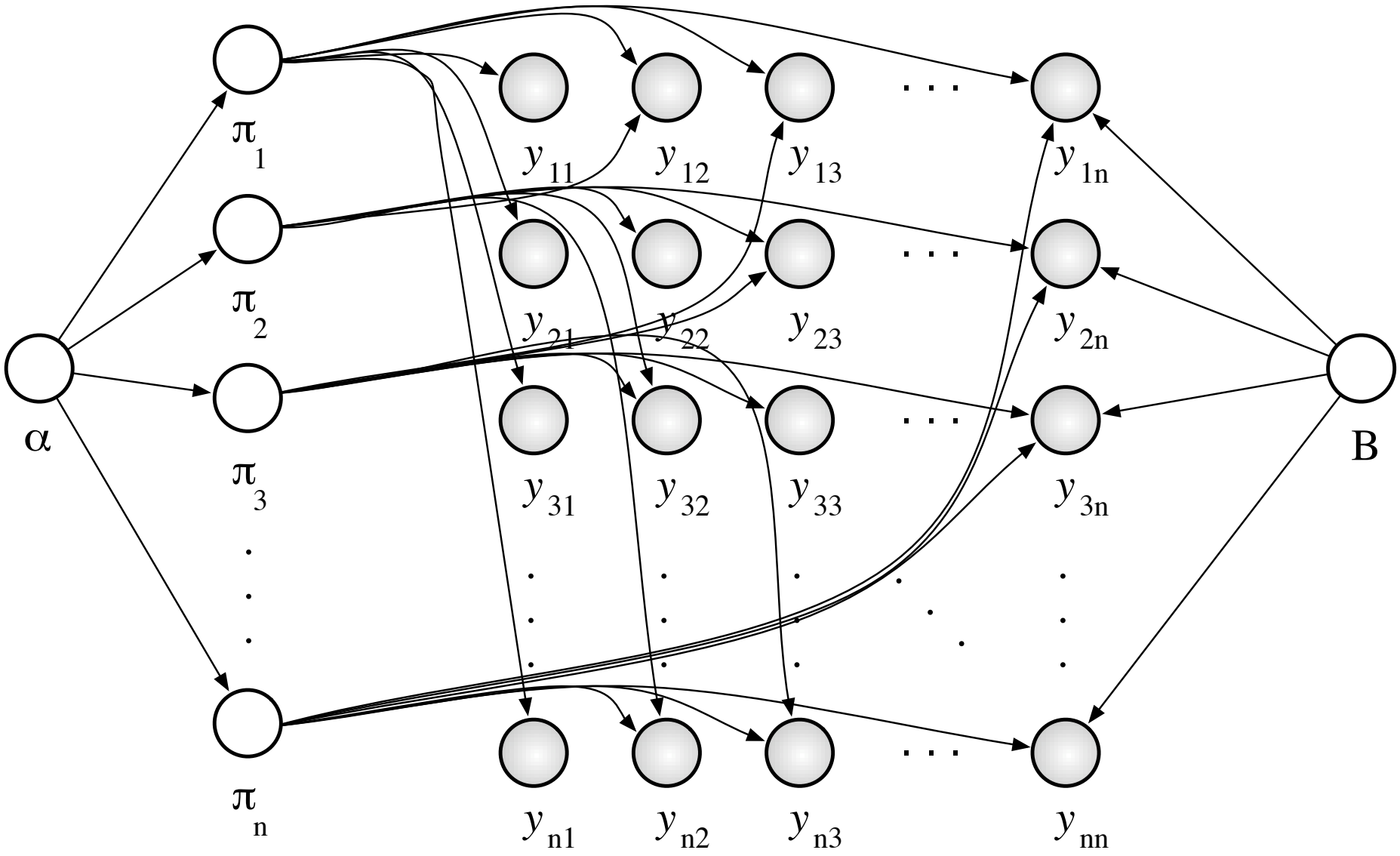
Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$

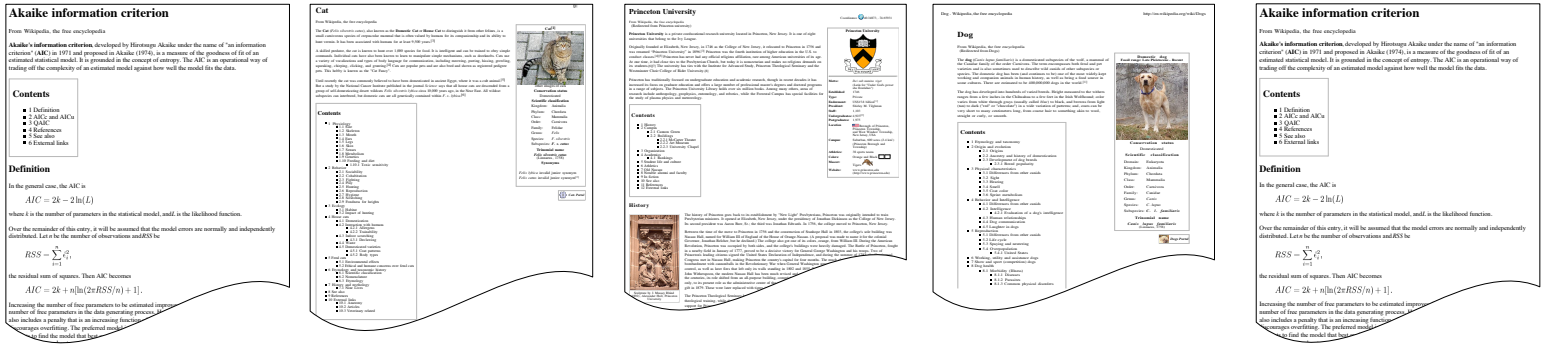


$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Probability models

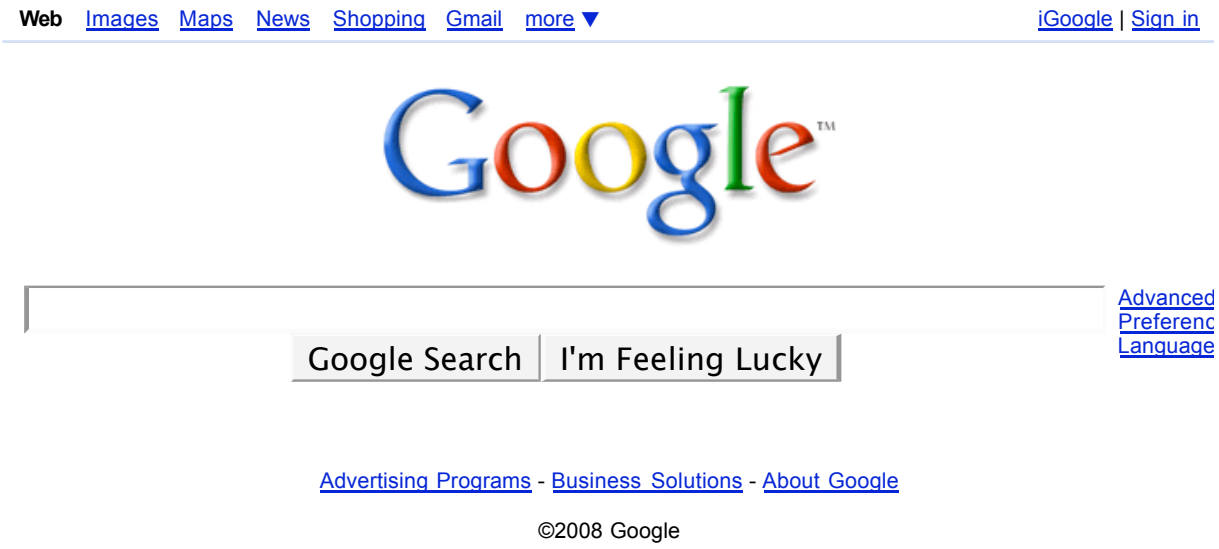


Understanding assumptions



- The methods we'll study make **assumptions** about the data on which they are applied. E.g.,
  - Documents can be analyzed as a sequence of words;
  - or, as a "bag" of words.
  - Independent of each other;
  - or, as connected to each other
- What are the assumptions behind the methods?
- When/why are they appropriate?

Computational efficiency

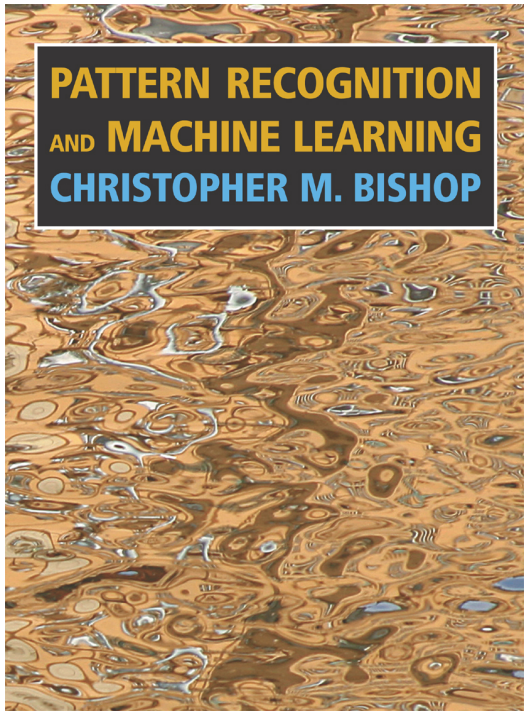


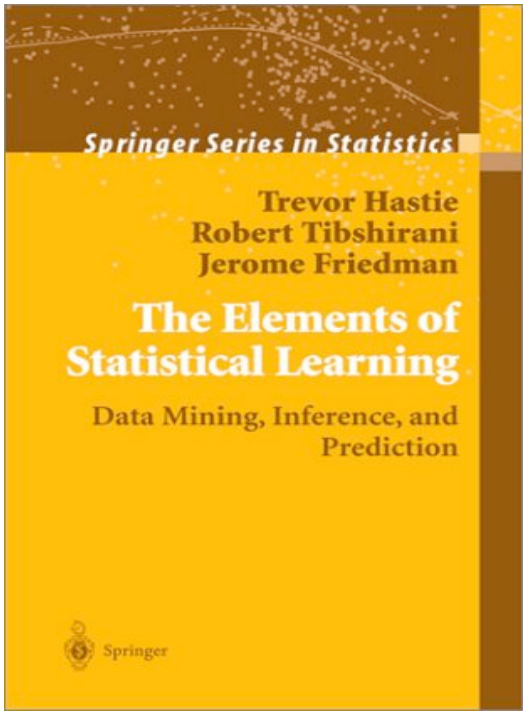
- What we can do with data depends on our computational constraints and on how much data we have.
- We need to understand these and tailor our methods to them. (This is connected to "understanding assumptions.")

Course requirements

- Attend and participate in lecture.
- Do the homework (about 65% of your grade).
- Write scribe notes.
- Prepare a final project (about 35% of your grade).

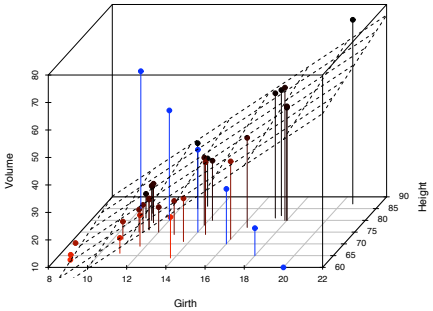
Course reading

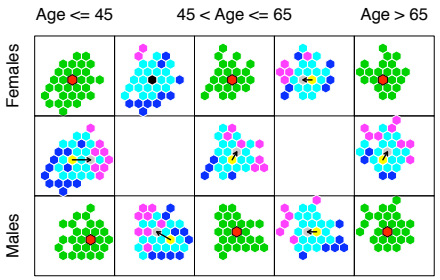


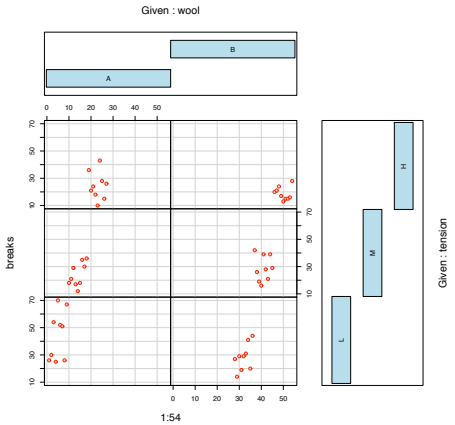


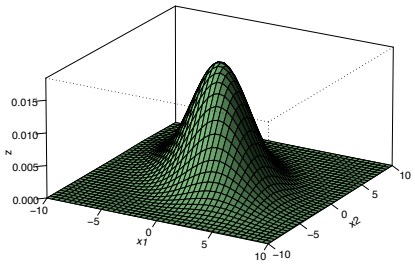
- We will provide reading materials.
- These two books are *excellent*.
- (In the future, Bishop will likely be required for this course.)

Homeworks









- Written and programming exercises
- Programming will be in R
- R is a great, free, open-source statistical programming
- The TAs will provide a tutorial for R in the next couple of weeks.
- (Proficiency in R will help you throughout your professional life.)
- See the course web-page for details on "late days."

Final Project

- The final project is the centerpiece of the course.
- Focused effort on a applied data analysis project
- Please try to work in pairs or groups of three.
- Example final projects from last year:
  - Analyzing the NetFlix competition data
  - Developing a wavelet-based clustering algorithm
  - Exploring *variational inference*, a general-purpose algorithm for learning probabilistic models

Course staff

- David Blei
  204 CS Building
  blei@cs.princeton.edu
  659-258-9907
  Office hours: by appointment
- Indraneel Mukherjee
  103C CS Building
  imukherj@cs.princeton.edu
  Office hours: Monday 6:30PM-8:30PM; AI lab (4th floor)
- Martin Suchara
  103A CS Building
  msuchara@cs.princeton.edu
  Office hours: Wednesday 6:30PM-8:30PM; AI lab (4th floor)

Contacting us

- Don't hesitate to contact us to discuss the material or anything else related to the course.
- **Preferred**: Use the course mailing list
  - Usually answered within 1 day by me, Martin, or Indraneel
  - Any kind of technical question
  - Many administrative questions
  - This way, everyone can benefit from the Q and A.
- If your query is more sensitive, then email the course staff separately. You will get a response within 2-3 days.
- If you need a response immediately, call me or stop by my office.

Tentative syllabus

- Probability and statistics review
- Classification (Naive Bayes, support vector machines, boosting)
- Clustering (K-means, agglomerative, mixture models)
- Sequential data (Hidden Markov models)
- Prediction (Linear regression, logistic regression, GLMs)
- Dimensionality reduction (PCA, Factor analysis)
- Continuous sequential data (Kalman filters)
- Advanced topics (Bayesian statistics, MCMC)
- Applications (Neuroscience, Vision, Information retrieval)