CS 451/686-02 Data Mining Introduction

Fall 2016

Maria Daltayanni

part of the slides is credited to [mmds] and the [ISL] authors

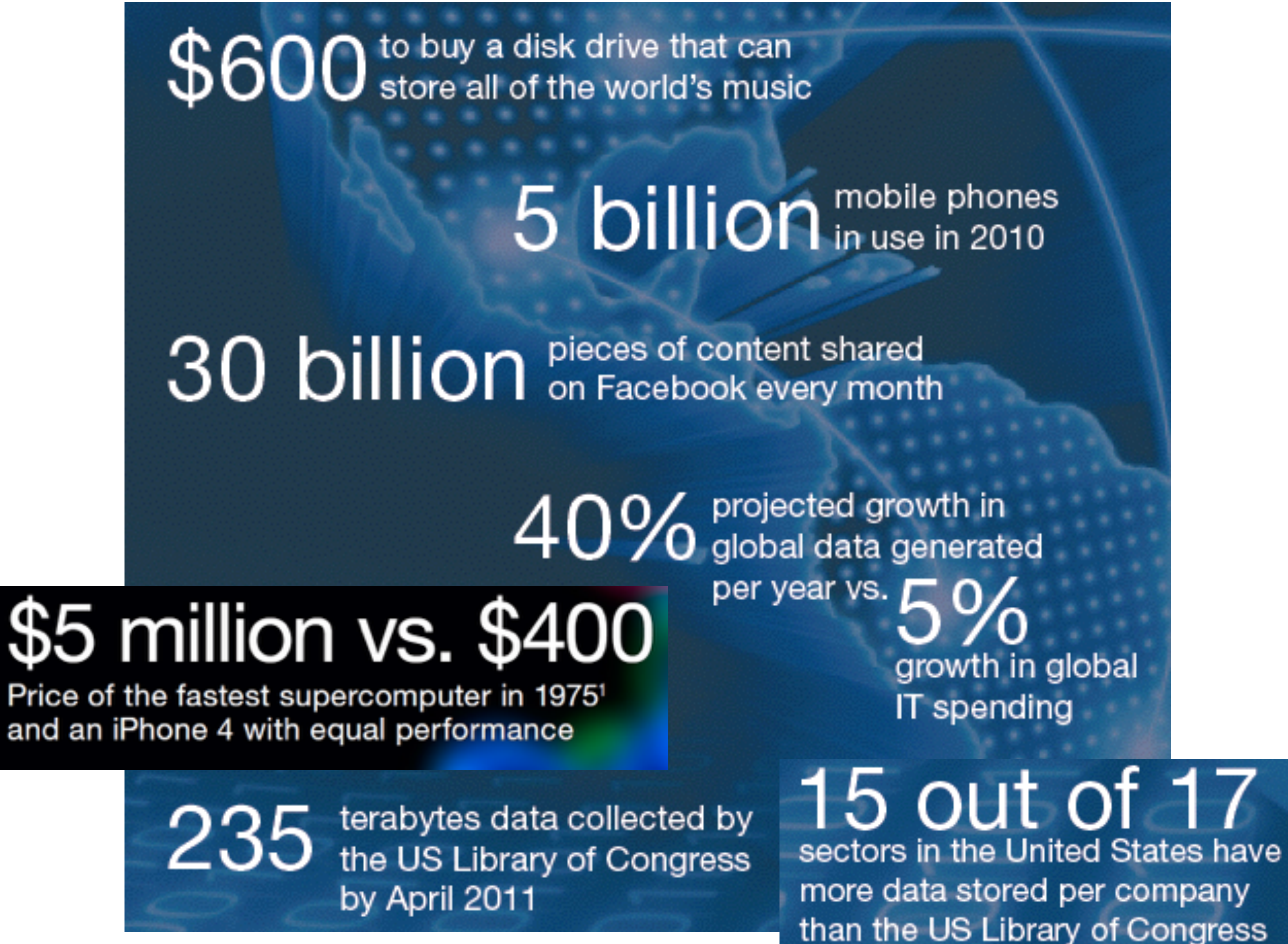
What is Data Mining?

Knowledge Discovery from Data

But:

- How BIG
- How PRICEY
- How VALUABLE

is the data?


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Data contains value and knowledge

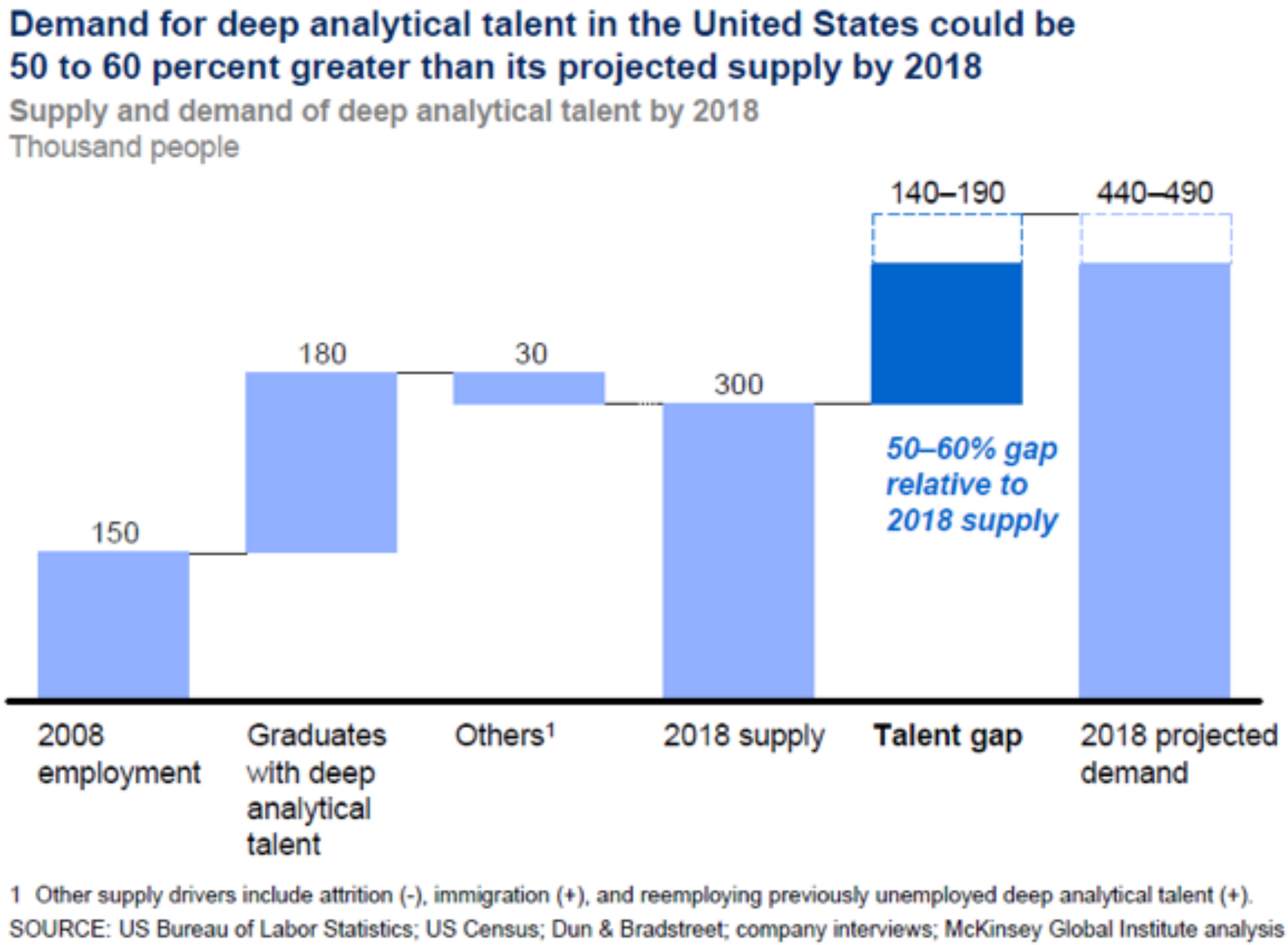
Data Mining

To extract the knowledge, data needs to be:

- **Stored**
- **Managed**
- **And ANALYZED ← this class**

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Good news: Demand for Data Mining



What is Data Mining?

- **Given lots of data**

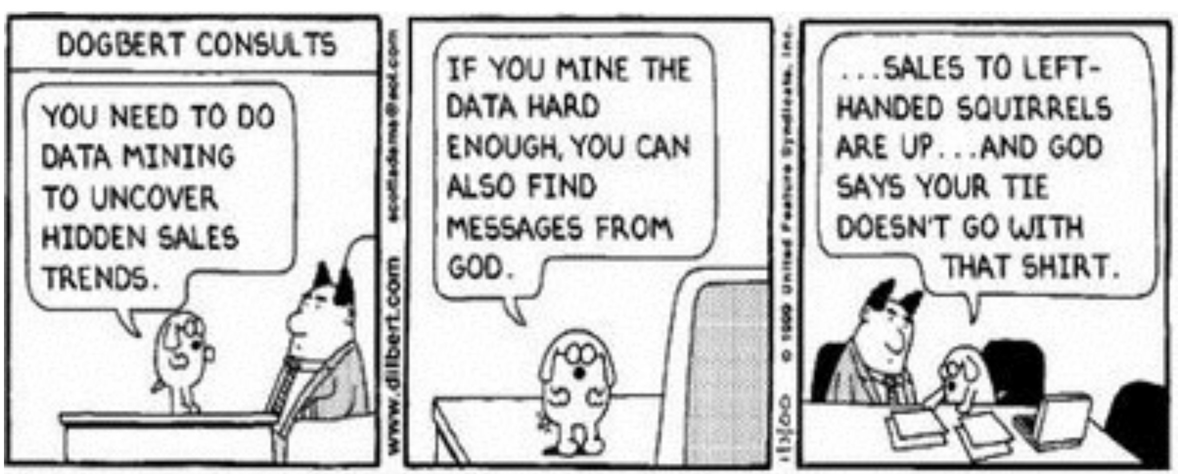
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless**
- Statisticians call it **Bonferroni’s principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

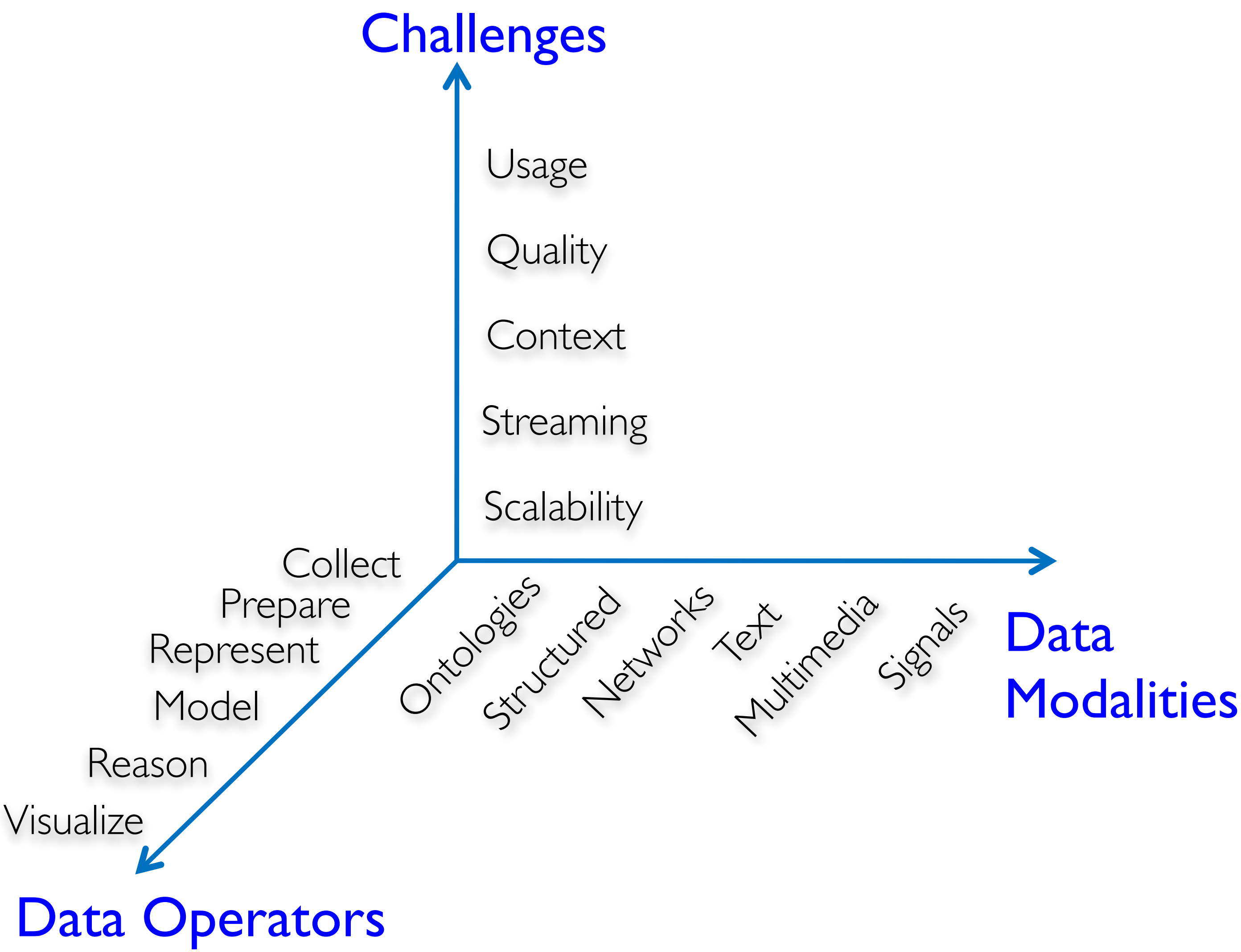


Meaningfulness of Analytic Answers

Example:

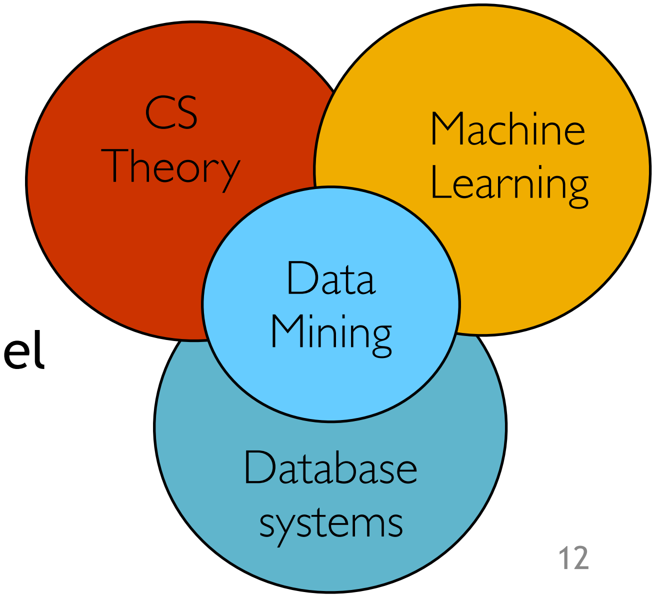
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **Expected number of "suspicious" pairs of people:**
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What matters when dealing with data?



Data Mining Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model
- **In this class we will do both!**



Statistical Learning vs Machine Learning

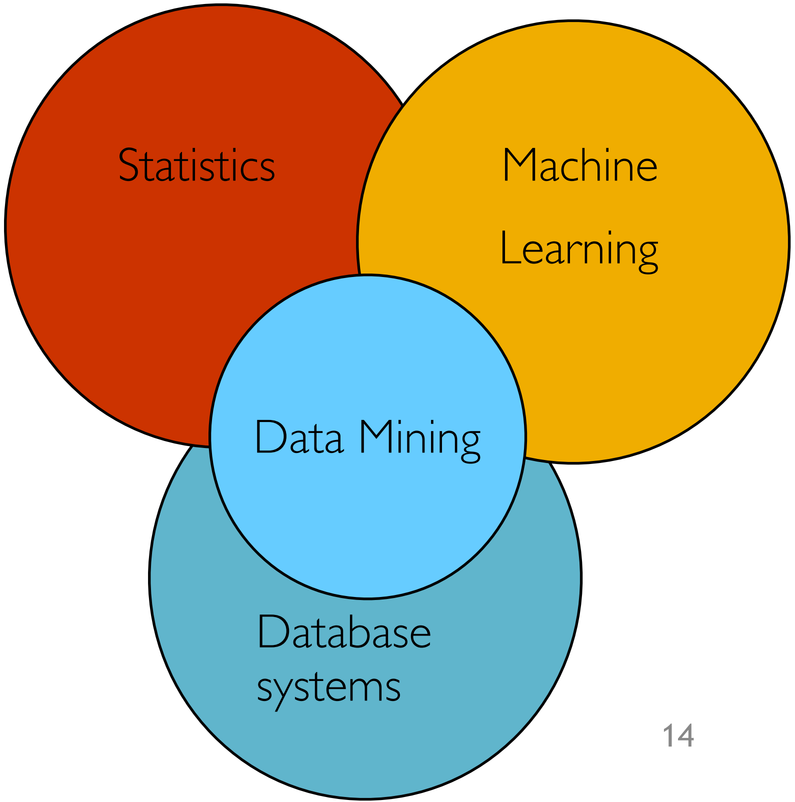
- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.

- **There is much overlap** — both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on **large scale** applications and **prediction accuracy**.
  - Statistical learning emphasizes **models** and their interpretability, and **precision** and **uncertainty**.

- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization"
- Machine learning has the upper hand in **Marketing**!

This class: CS 451/686-02

- **This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on**
- **Modeling**
- **Algorithms**
- **Computing architectures**
- Handling **large data**



What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - Single machine in-memory
  - MapReduce
  - Streams and online algorithms (tentative)

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
- **We will learn various "tools":**
  - Linear algebra (SVD, Rec. Sys., Communities)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

How it all fits together

High dim. data
- Locality sensitive hashing
- Clustering
- Dimensionality reduction

Graph data
- PageRank, SimRank
- Community Detection
- Spam Detection

Infinite data
- Filtering data streams
- Web advertising
- Queries on streams

Machine learning
- SVM
- Decision Trees
- Perceptron, kNN

Apps
- Recommender systems
- Association Rules
- Duplicate document detection

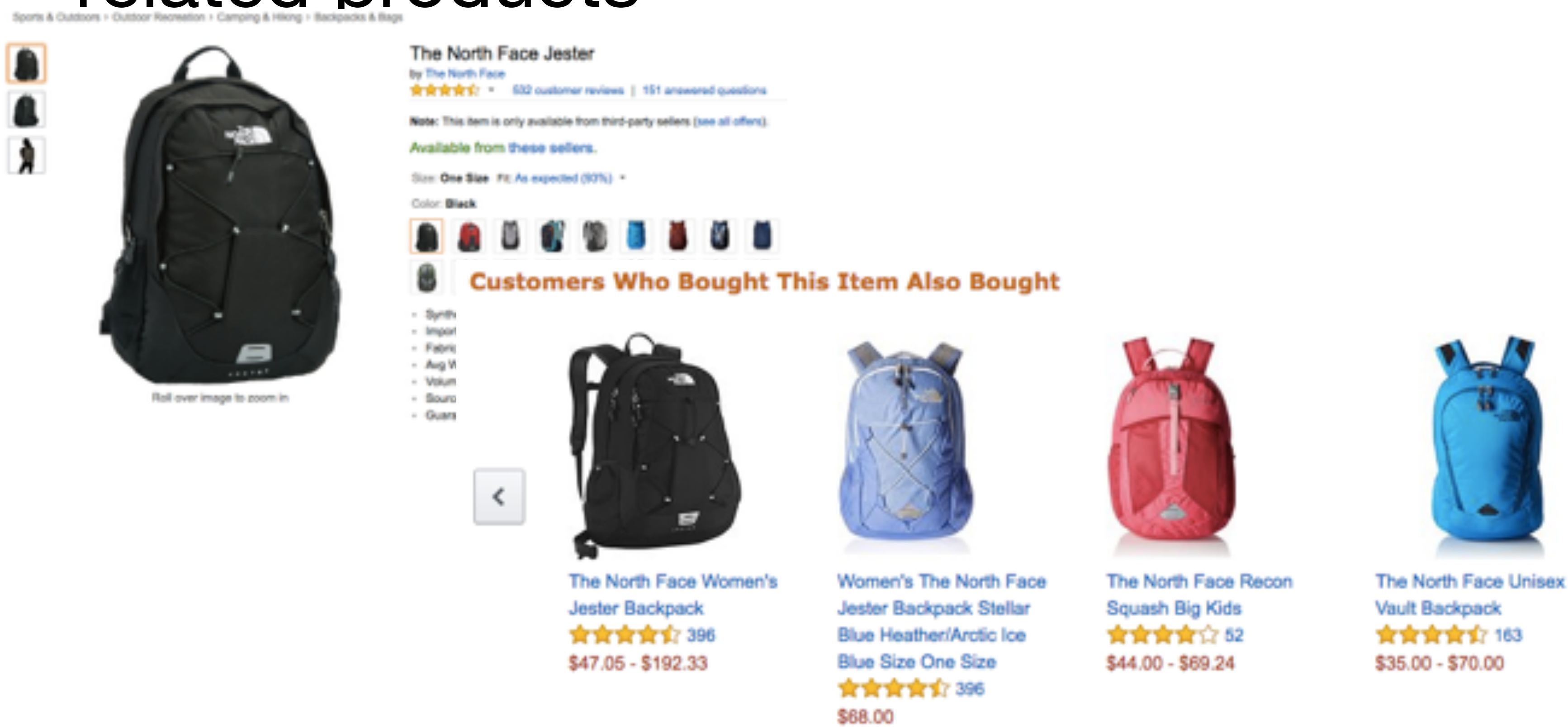
Any Real World problems?

- **Amazon** recommendations - association rules
- **Zillow** price estimate - regression
- **Gmail** spam - classification
- **Netflix** movie rating prediction - collaborative filtering
- **Facebook** news feed - classification
- **Yahoo!** news categories - clustering
- **Google** search - pagerank

… and many many more!

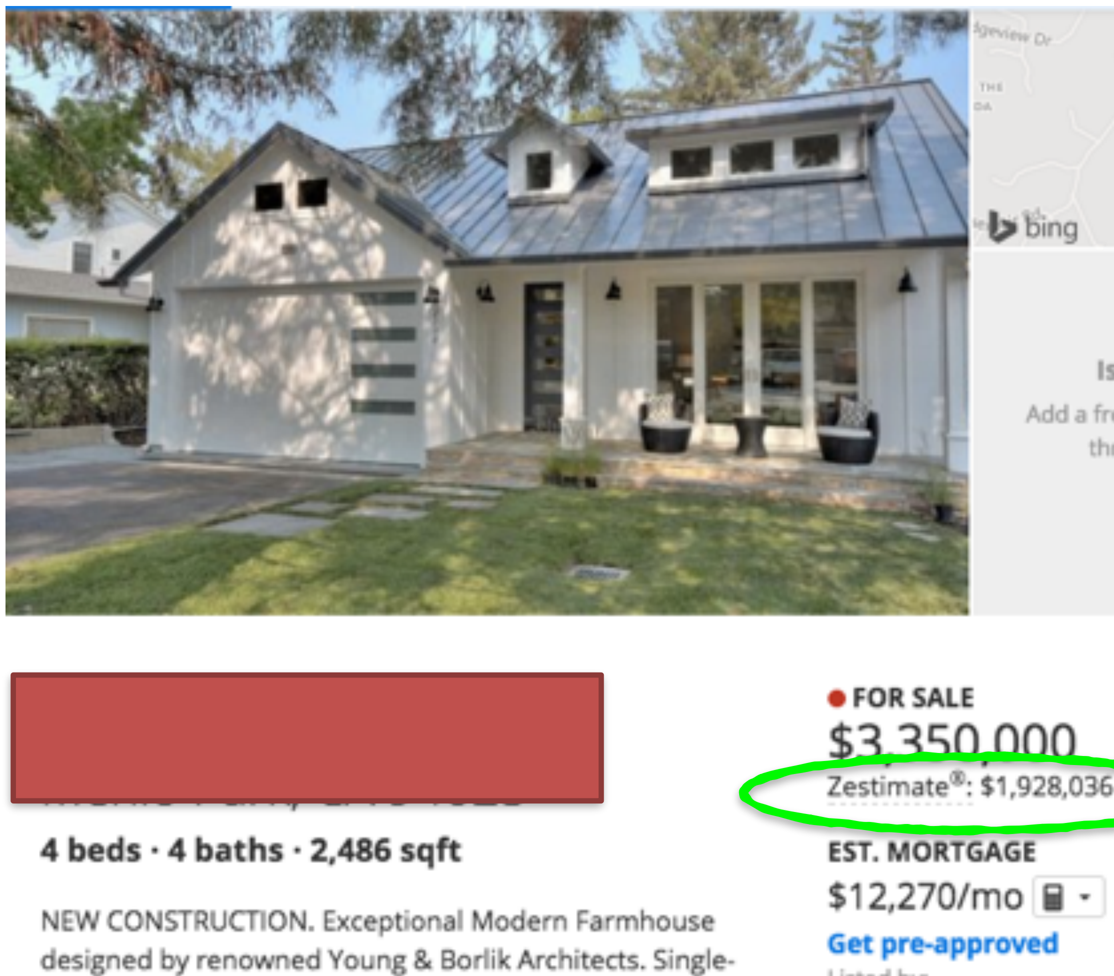
Amazon Recommendations

- Using Association rules to recommend related products



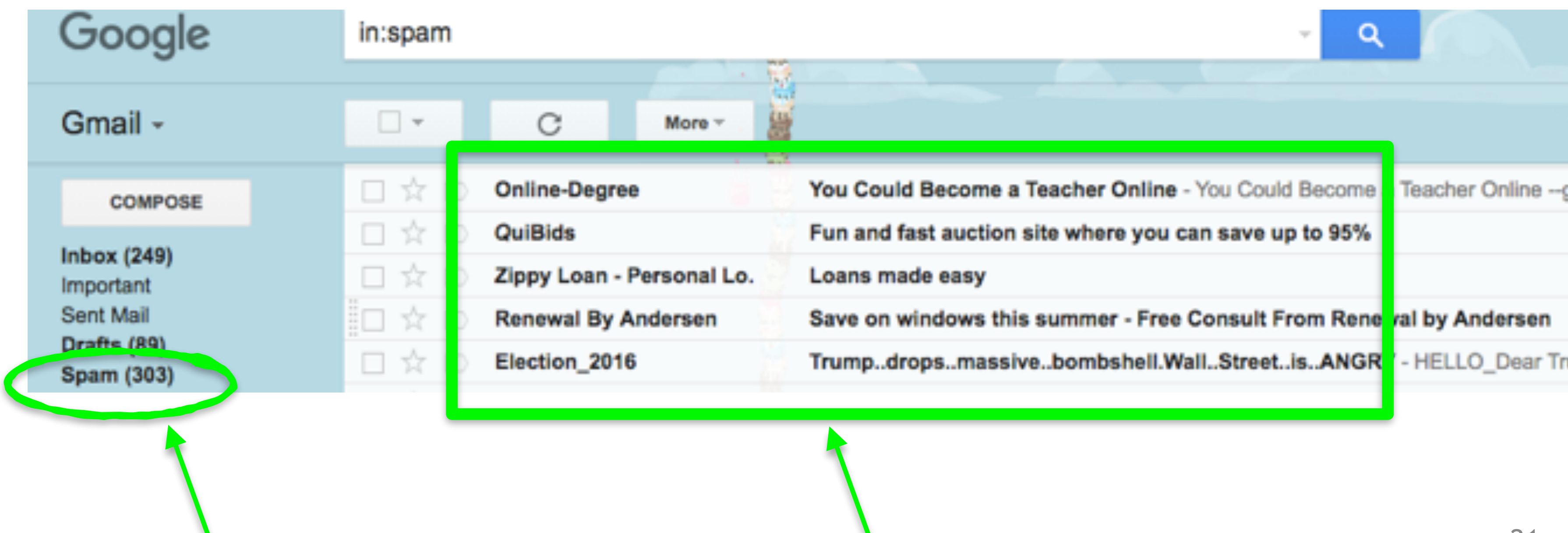
Zillow Price Estimate

- Using Regression to estimate the true value of a house



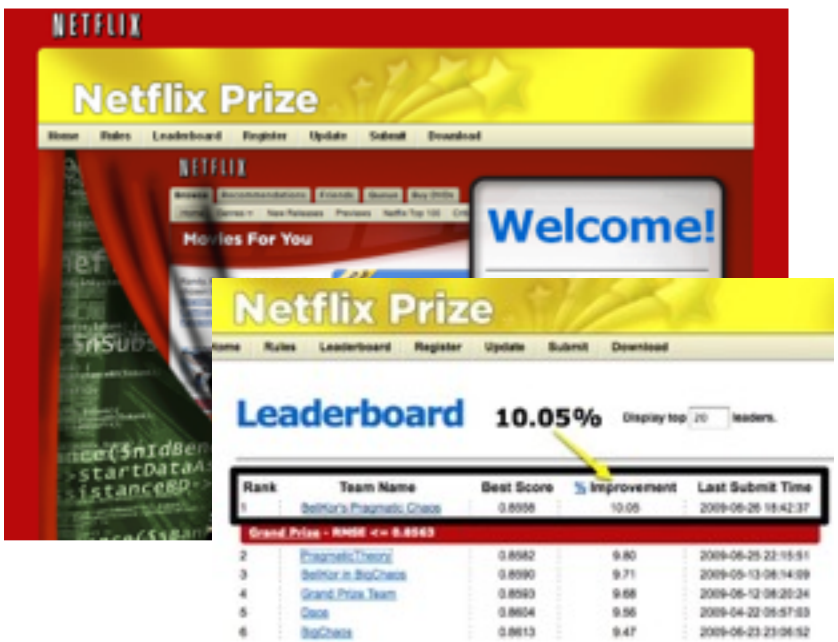
Gmail Spam

- Using Classification to detect spam emails
- Google categorizes email to trusted and spam. Spam is automatically dropped to trash.



Netflix Movie Rating Prediction

- Using Collaborative Filtering to predict user ratings for films: Netflix prize: 1M!
  - based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.



The Netflix Prize

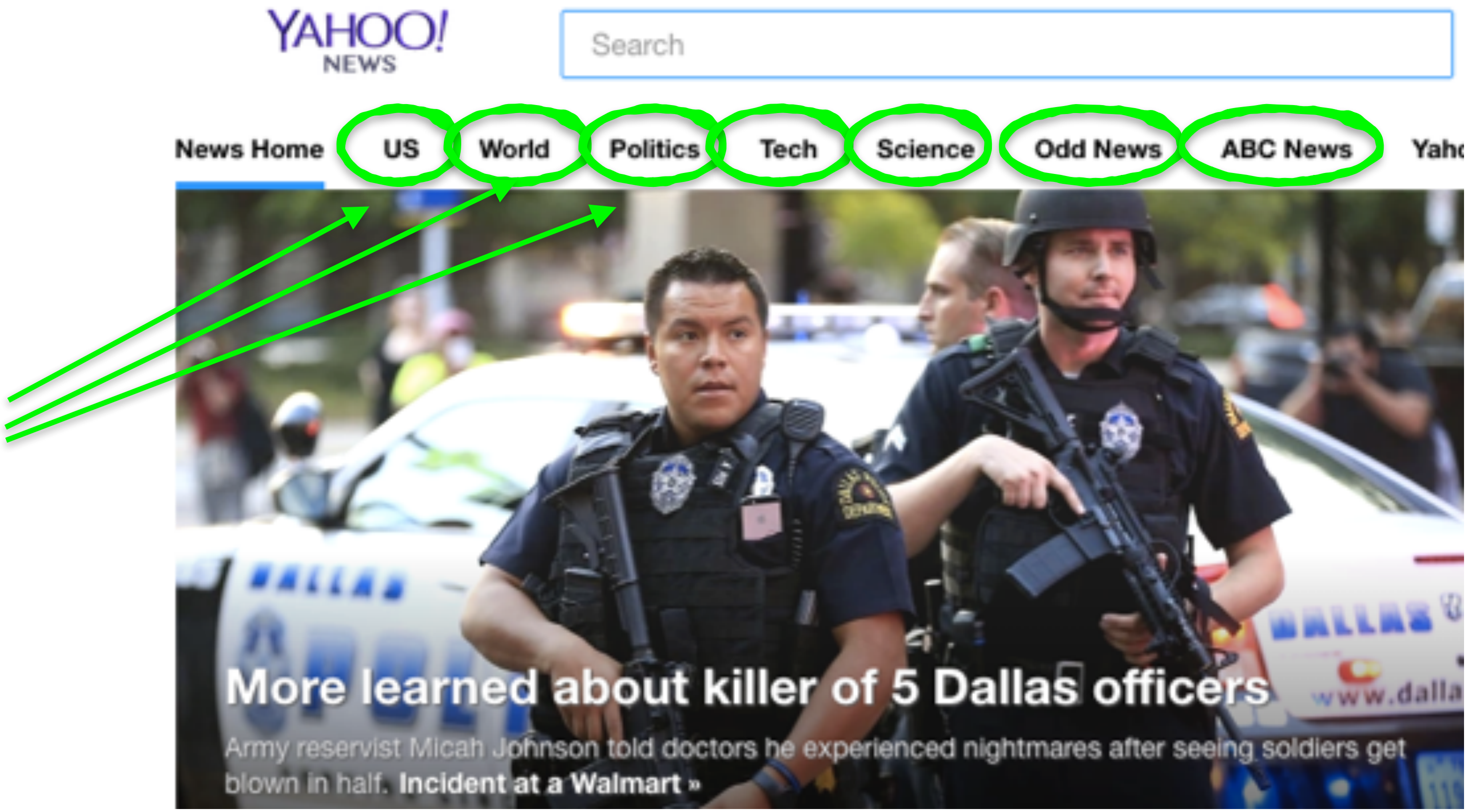
• competition started in October 2006. Training data is ratings for 18, 000 movies by 400, 000 Netflix customers, each rating between 1 and 5.
• training data is very sparse— about 98% missing.
• objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
• Netflix's original algorithm achieved a root MSE of 0.953.
The first team to achieve a 10% improvement wins one million dollars.
• is this a supervised or unsupervised problem?

Facebook News Feed

- Using Classification to select posts to display on main page

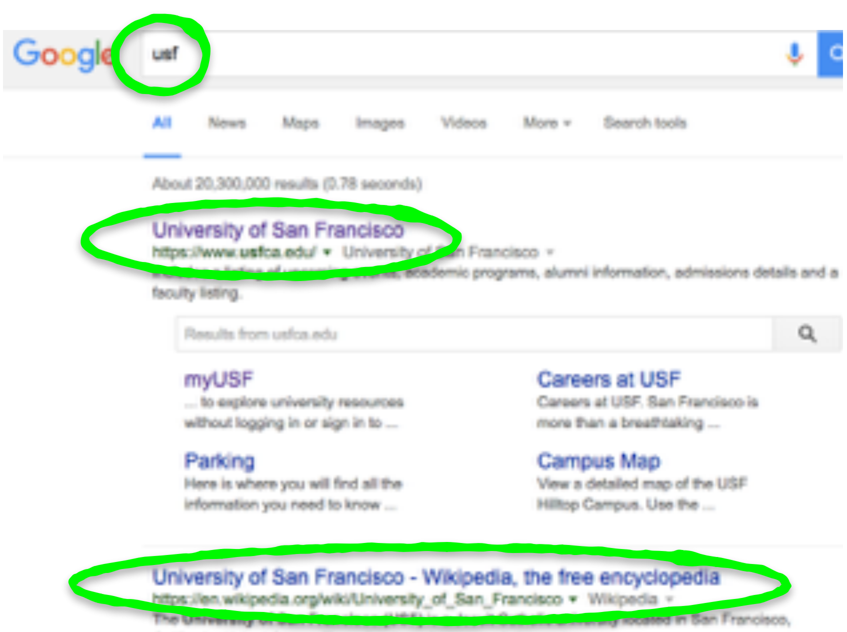
Yahoo! News Categories

- Using Clustering to organize news in categories (US, World, Politics, Tech, …)



Google Search

- Using Pagerank to rank webpages according to ***popularity*** and ***importance***
- Searching for "USF" returns 1) the official USFCA page, 2) the Wiki page, and 3) three pages from dons athletics, one of the most popular pages of USF



Knowledge discovered from data

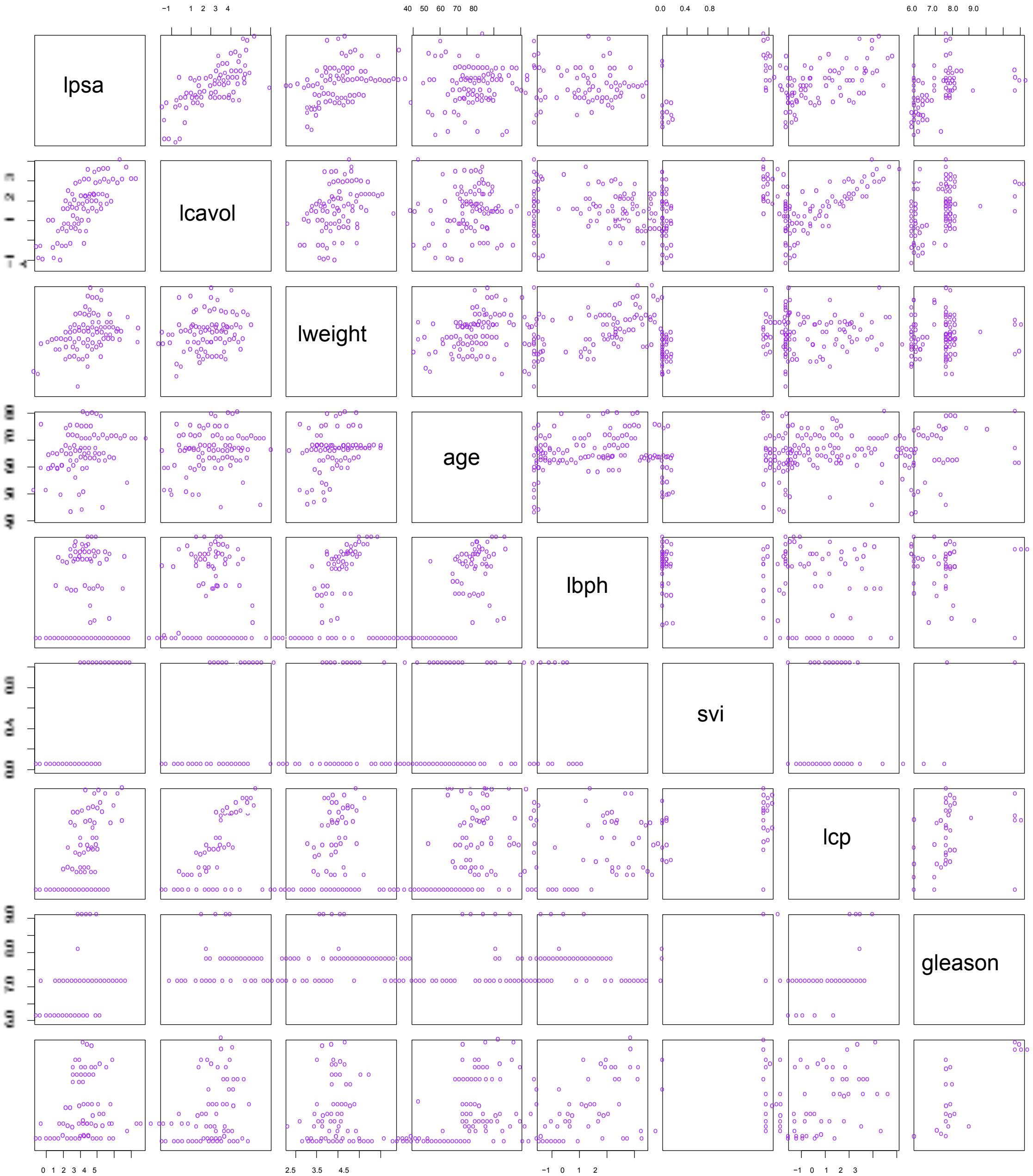
- Amazon **recommendations** - association rules
- Zillow **price estimate** - regression
- Gmail **spam** - classification
- Netflix **movie rating prediction** - collaborative filtering
- Facebook **news feed** - classification
- Yahoo! **news categories** - clustering
- Google **search** - page rank

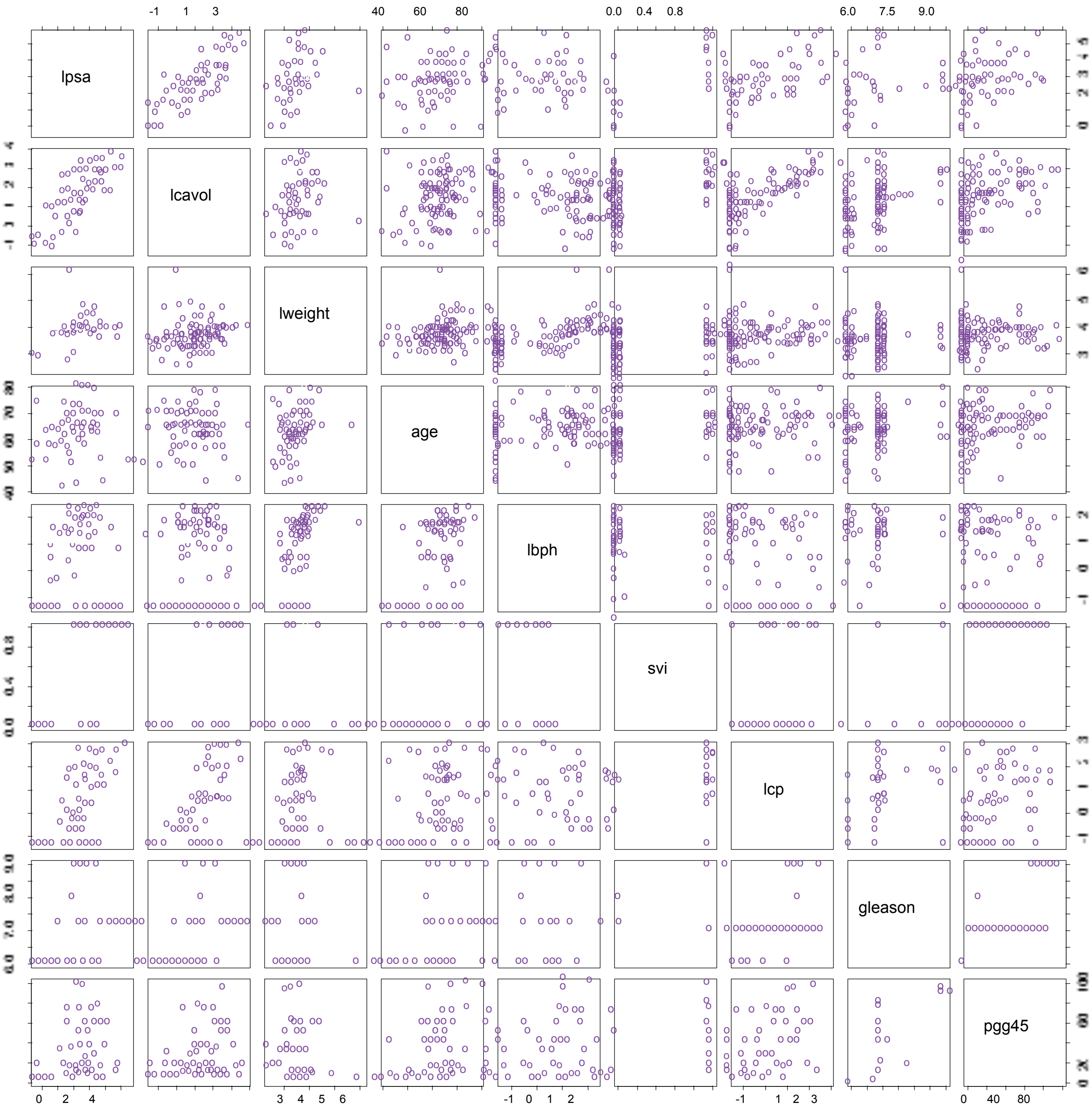
… and many more!

How do you want that data?

More Statistical Learning Problems

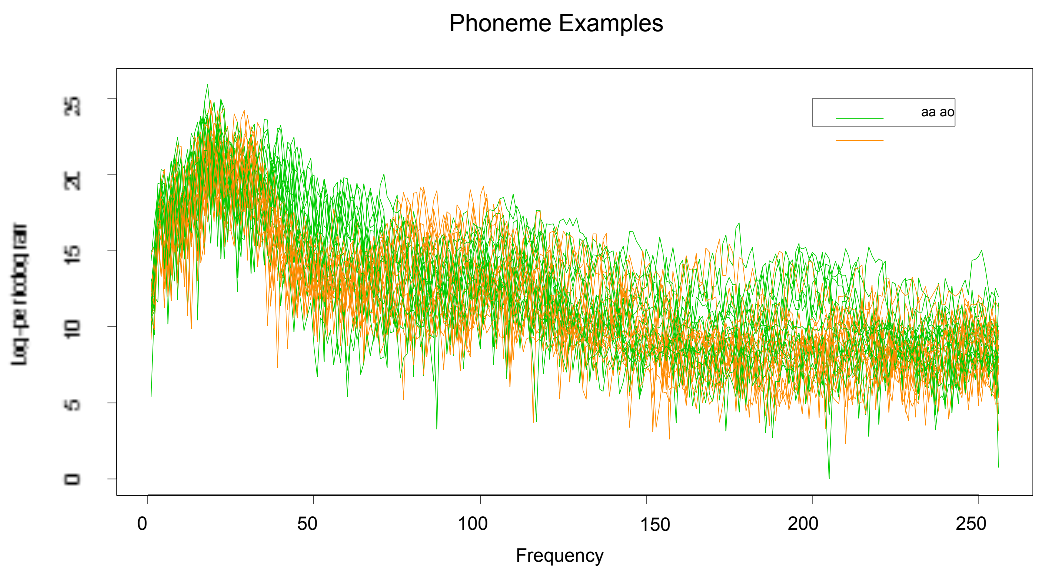
- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

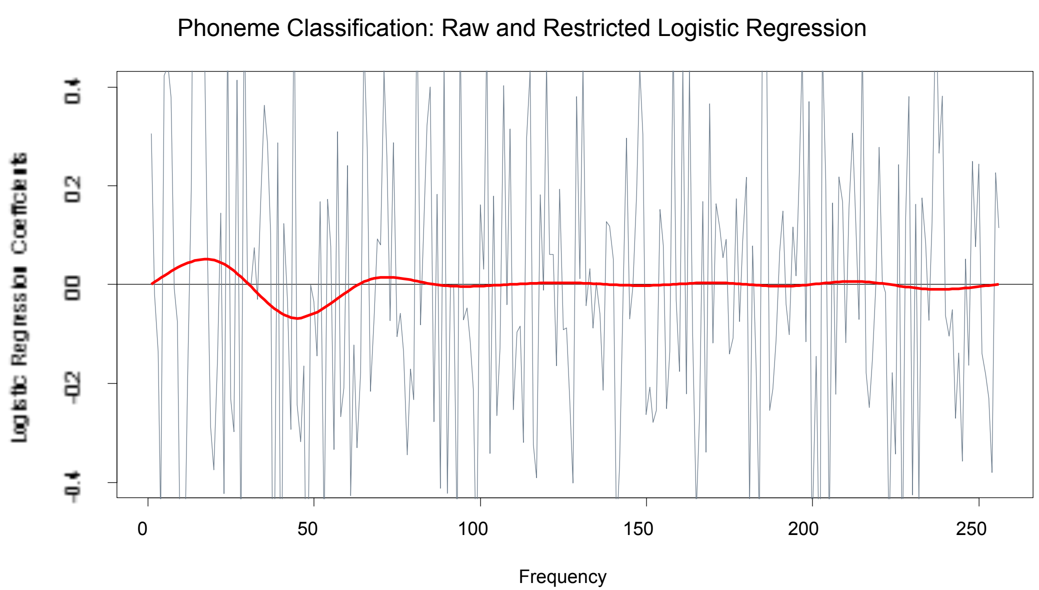

lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.


Phoneme Examples
aa ao
Log-periodogram
0
5
10
15
20
25
0
50
100
150
200
250
Frequency


Phoneme Classification: Raw and Restricted Logistic Regression
Logistic Regression Coefficients
-0.4
-0.2
0.0
0.2
0.4
0
50
100
150
200
250
Frequency

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

Spam Detection

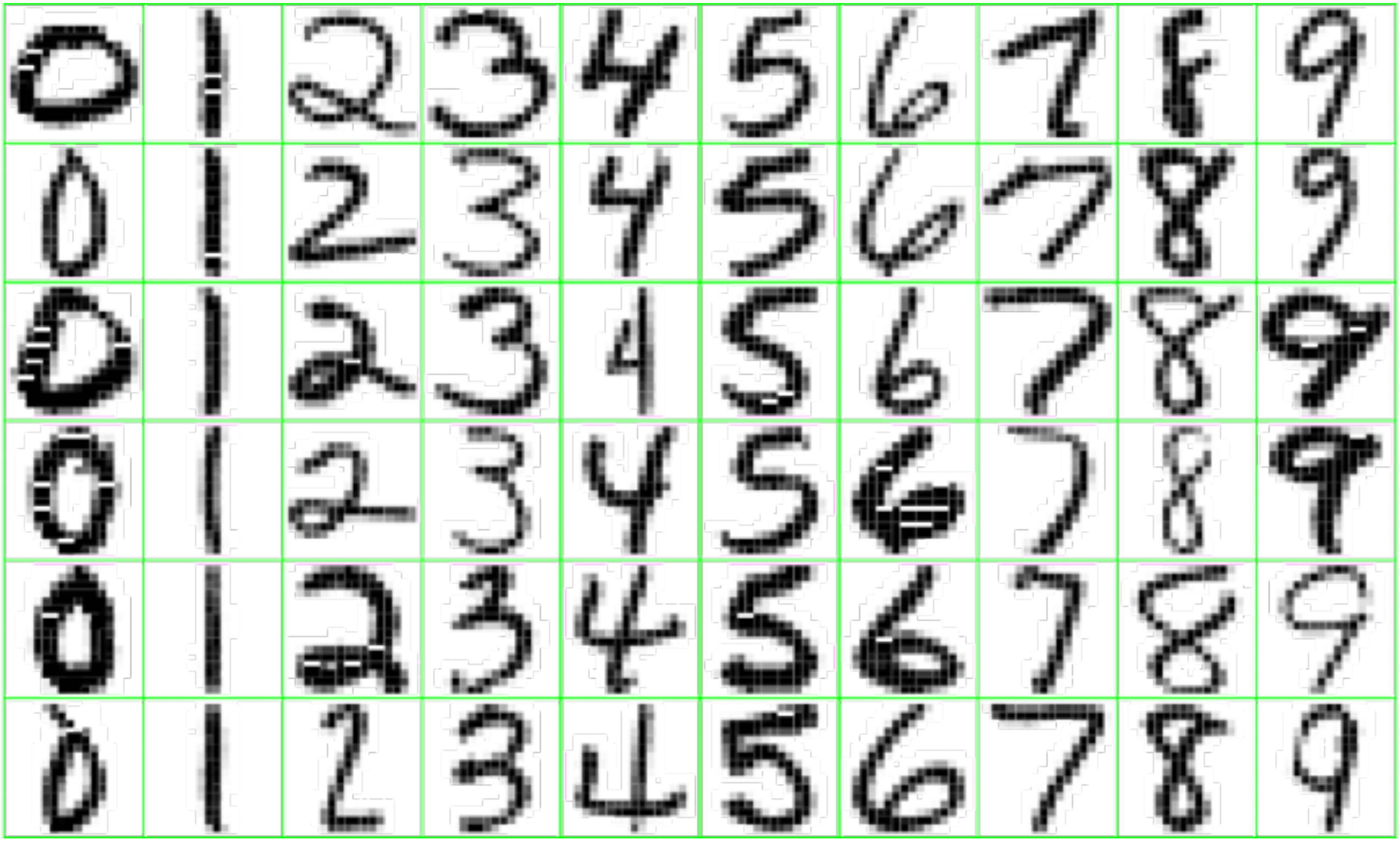
- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email*.*

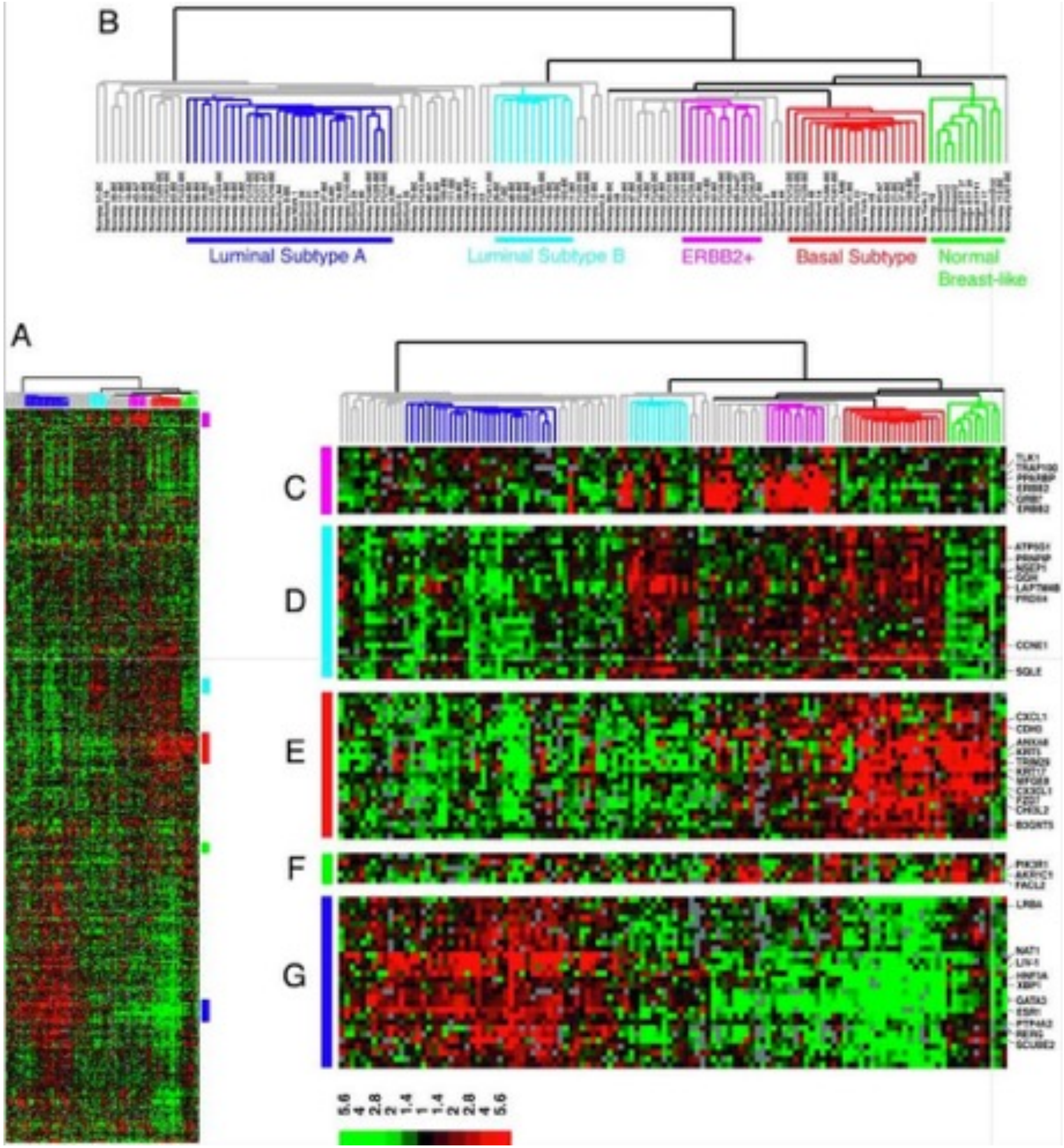
Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.



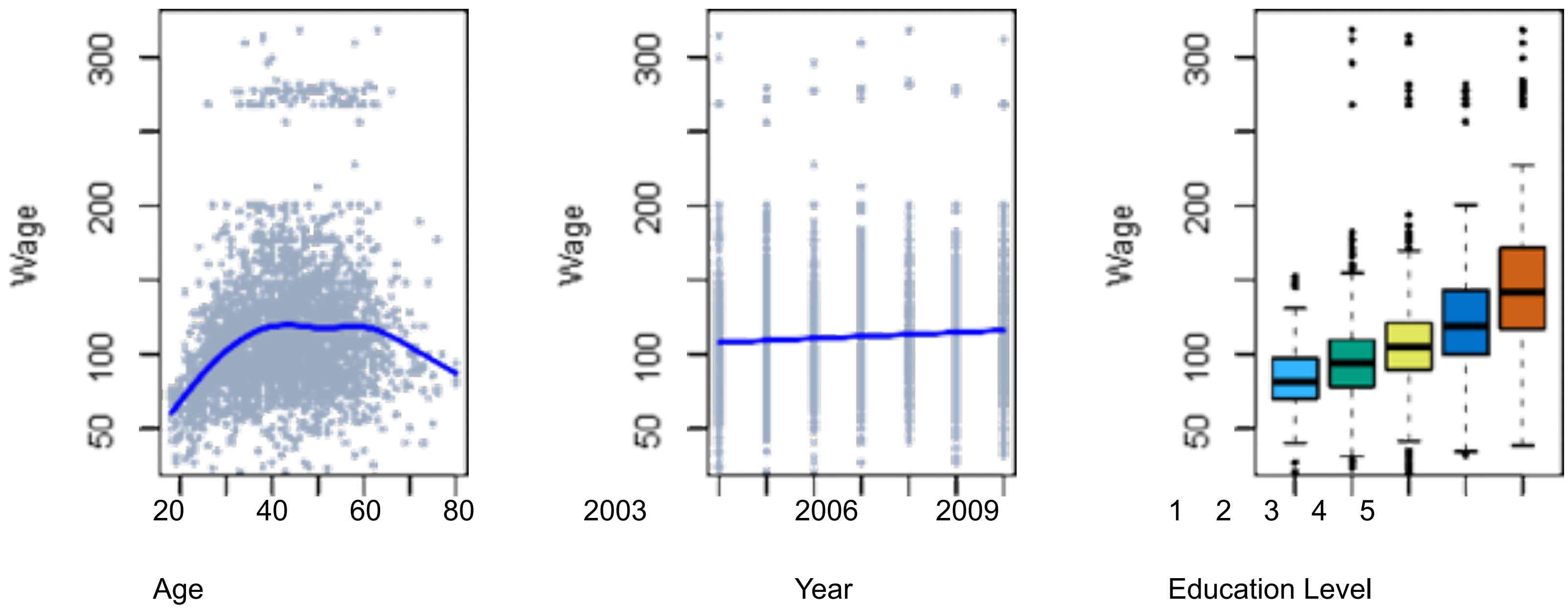
Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

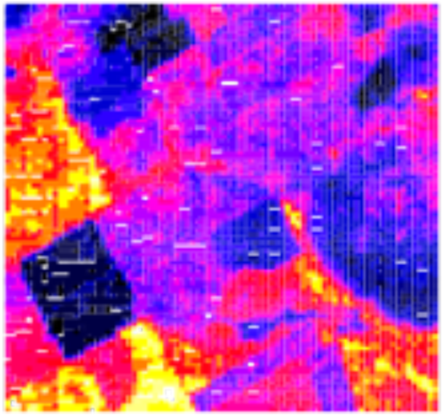


Income survey data for males from the central Atlantic region of the USA in 2009.

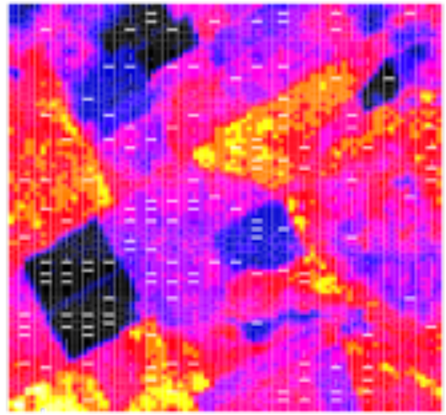
Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

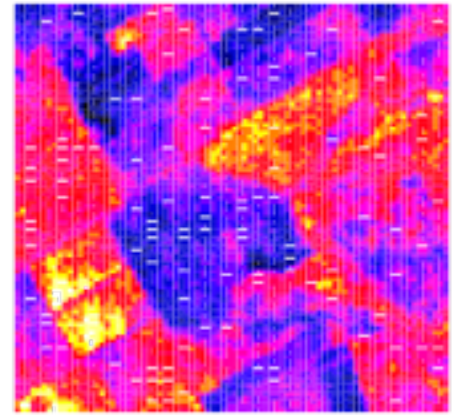
Spectral Band 1



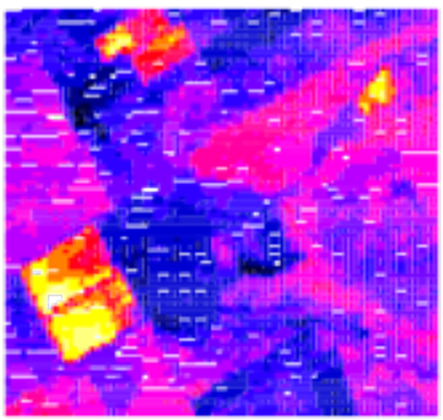
Spectral Band 2



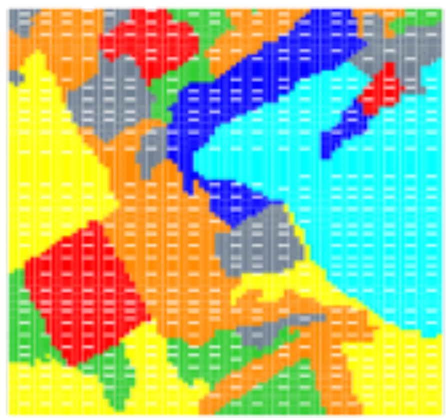
Spectral Band 3



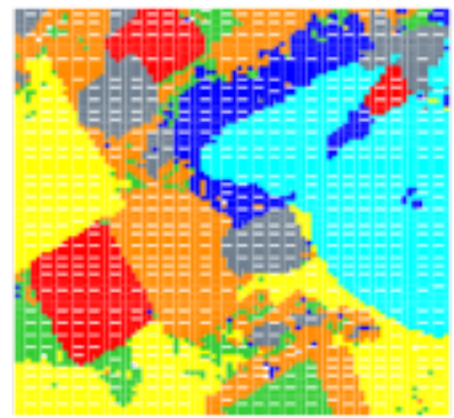
Spectral Band 4



Land Usage



Predicted Land Usage



Usage $\in$ {*red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil*}

The Supervised Learning Problem

Starting point:

• Outcome measurement Y (also called dependent variable, response, **target**).

• Vector of p predictor measurements X (also called inputs, regressors, covariates, **features**, independent variables).

• In the **regression** problem, Y is quantitative (e.g price, blood pressure).

• In the **classification** problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).

• We have **training data** (x1, y1), . . . ,(xN , yN). These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern **data scientist**.

Unsupervised Learning

No outcome variable, just a set of **predictors** (**features**) measured on a set of samples.

• objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
• difficult to know how well you are doing.
• different from supervised learning, but can be useful as a pre-processing step for supervised learning.