Lecture-1

Course Title: Data Science

Course Code: 17B11CI611

L-T-P Scheme: 3-1-0 Credit: 4

Course coordinator: Dr. Ajay Kumar

Content

- About the course
- Grading policy
- Learning objectives of this course
- Introduction to Data Science

About the course

- A mixture of theory and practice
- Introductory, broad overview of subjects
- Focus on practical aspects, but not on ever-changing technology and tools
- Seminar style - I am here to learn as well as to teach

Continued…

- I used many sources in preparing for this course:
  - Practical Data Science using R by Zumel and Mount
    - http://www.manning.com/zumel/
  - Data Mining with R: Learning with Case Studies, by Torgo
    - http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR/
  - An Introduction to Data Science, Version 3, by Stanton
    - http://jsresearch.net/
  - Monte Carlo Simulation and Resampling Methods for Social Science, by Carsey and Harden
    - http://www.sagepub.com/books/Book241131/reviews?course=Course14&subject=J00&sortBy=defaultPubDate%20desc&fs=1#tabview=title
  - Machine Learning with R by Lantz
    - http://www.packtpub.com/machine-learning-with-r/book

Continued…

- Required:
  - Data Science from Scratch (**DSS**) by Joel Grus
  - **Free** e-book: Think Stats (**TS**) by Allen B. Downey. PDF | website

Continued…

- A Simple Introduction to Data Science, by Burlingame and Nielsen
  - http://newstreetcommunications.com/businesstechnical/a_simple_introduction_to_data_science
- Ethics of Big Data, by Davis
  - http://shop.oreilly.com/product/0636920021872.do
- Privacy and Big Data, by Craig and Ludloff
  - http://shop.oreilly.com/product/0636920020103.do
- Doing Data Science: Straight Talk from the Frontline, by O'Neil and Schutt
  - http://shop.oreilly.com/product/0636920028529.do

Continued…

- Lots of places to learn more about R
  - All of the sources on the first slide have R code available
  - Comprehensive R Archive Network (CRAN)
    - http://cran.r-project.org/manuals.html
  - Springer Textbooks Use R ! Series
    - http://www.springer.com/series/6991
  - Online search tool Rseek
    - http://www.rseek.org/
  - The RStudio site
    - http://www.rstudio.com/
  - The Odum Institute's online course
    - http://www.odum.unc.edu/odum/contentSubpage.jsp?nodeid=670

Grading policy

- 25% for internal including 5% for attendance
- T1- 15 %
- T2- 25%
- T3-35%

- I reserve the right to slightly adjust the weights of internal marks components if necessary

Learning objectives of this course

- After taking this course,

....you should understand the full data science pipeline, and be familiar with programming tools to accomplish the different portions

... be able to collect data from unstructured sources and store it using appropriate structure such as relational databases, graphs, matrices, etc

... know to explore and visualize your data

... be able to analyze your data rigorously using a variety of statistical and machine learning approaches

Course Overview

Topics covered (subject to change)

- Data Science Introduction
- Current landscape of perspectives
- Populations and samples
- Statistical modelling, probability distributions,
- Regression, fitting a model
- Introduction to data science programming using R
- Introduction to Machine Learning
- Data understanding and Data Preparation
- Feature Selection and Model Evaluation
- Supervised Modeling
- Unsupervised Modeling

Continued…

- Deep Learning
- Hypothesis testing,
- kernel methods and SVMs
- boosting, clustering
- dimensionality reduction
- recommender systems
- probabilistic models
- scalable ML
- **Visualization: basic visualization and data exploration**
- **data presentation** and interactivity……

Introduction to Data Science

- Data Science – Why all the excitement?
  - examples
- Where does data come from
- So what is Data Science
- Doing Data Science
- About the course
  - what we'll cover
  - data science first, big data later

Continued…

1935: "The Design of Experiments"

R.A. Fisher 

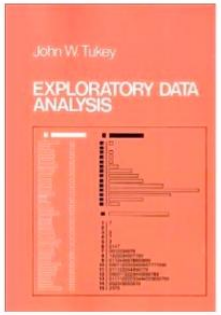
1939: "Quality Control"

W.E. Demming 

1958: "A Business Intelligence System"

 Peter Luhn

1977: "Exploratory Data Analysis"

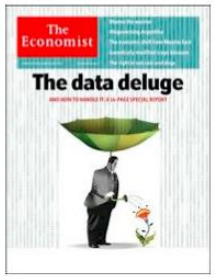


1989: "Business Intelligence"

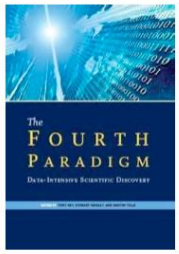
Howard Dresner 

1997: "Machine Learning"

2010:"The Data Deluge"



2007:"The Fourth Paradigm"



2009:"The Unreasonable Effectiveness of Data"

1996: Google

Google™

Abridged Version of Jeff Hammerbacher's timeline for CS 194, 2012

Data Science: Why all the Excitement?

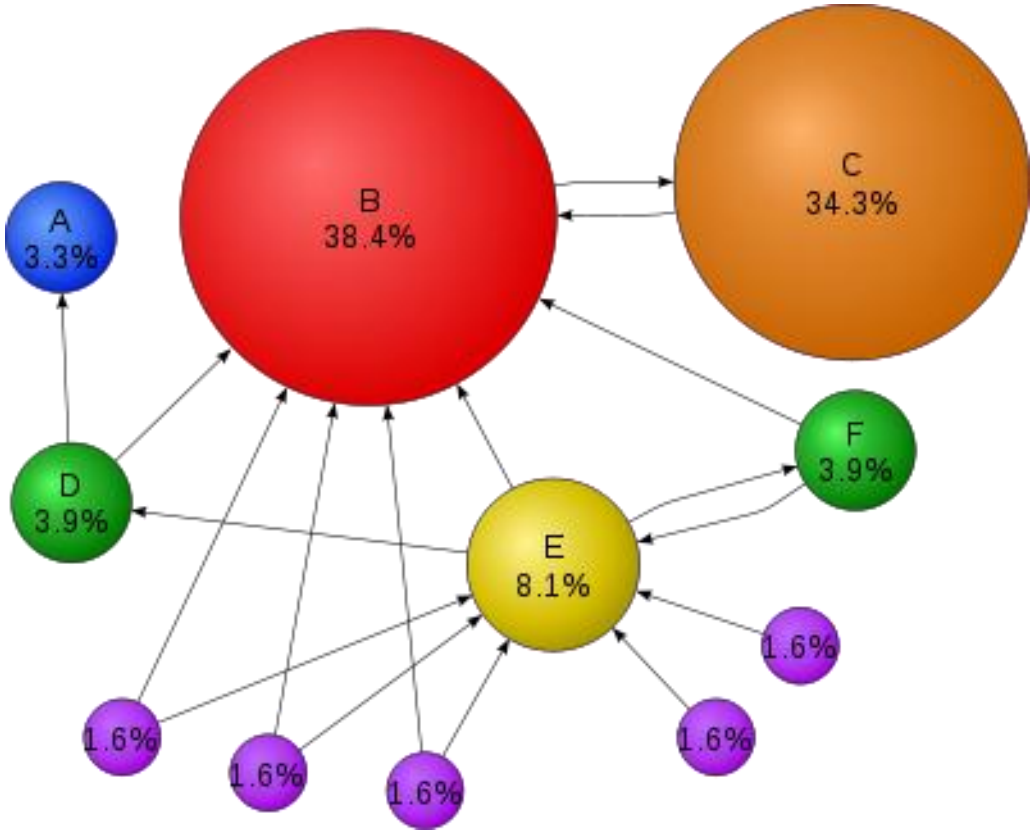
e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Continued…

-Page rank: The web as a behavioral dataset



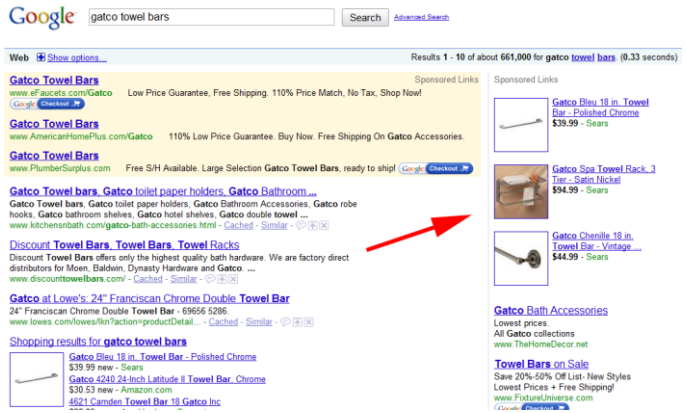
Continued...

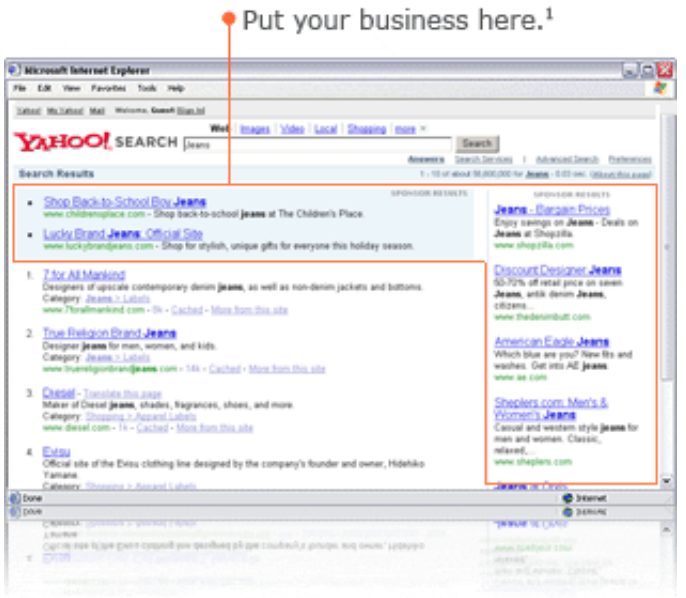
DB size = 50 billion sites

Google server farms
2 million machines (est)

1998 – sponsored search

Overture

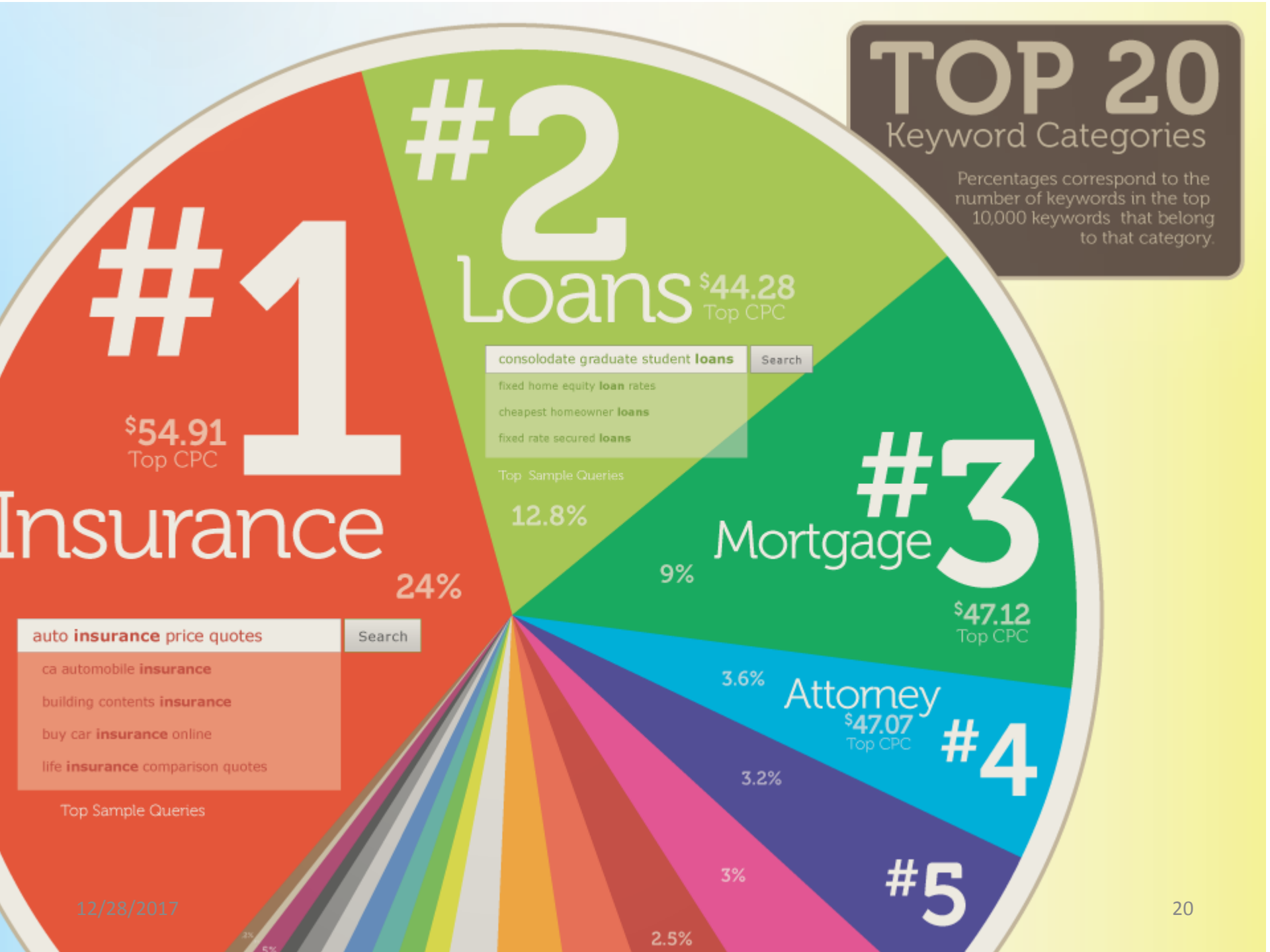


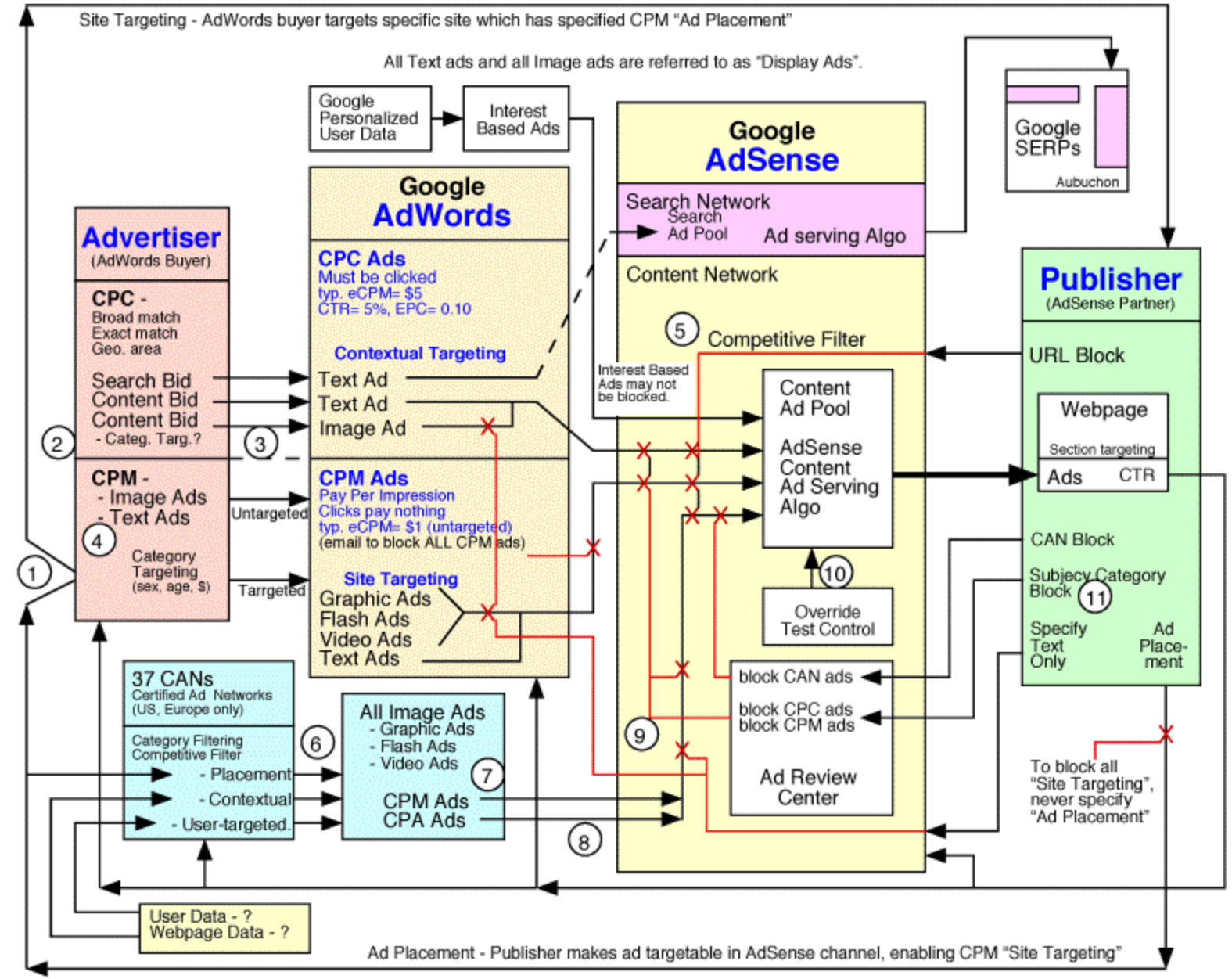


2002

Sponsored search

- Google revenue around $50 bn/year from marketing, 97% of the companies revenue.
- Sponsored search uses an auction – a pure competition for marketers trying to win access to consumers.
- In other words, a competition for **models** of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.
- There are around 30 billion search requests a month. Perhaps a **trillion events** of history between search providers.


TOP 20
Keyword Categories
Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.
#1
$54.91
Top CPC
Insurance
24%
auto insurance price quotes
Search
ca automobile insurance
building contents insurance
buy car insurance online
life insurance comparison quotes
Top Sample Queries
#2
Loans
$44.28
Top CPC
consolodate graduate student loans
Search
fixed home equity loan rates
cheapest homeowner loans
fixed rate secured loans
Top Sample Queries
12.8%
#3
Mortgage
9%
$47.12
Top CPC
3.6%
Attorney
$47.07
Top CPC
#4
3.2%
#5
3%
2.5%

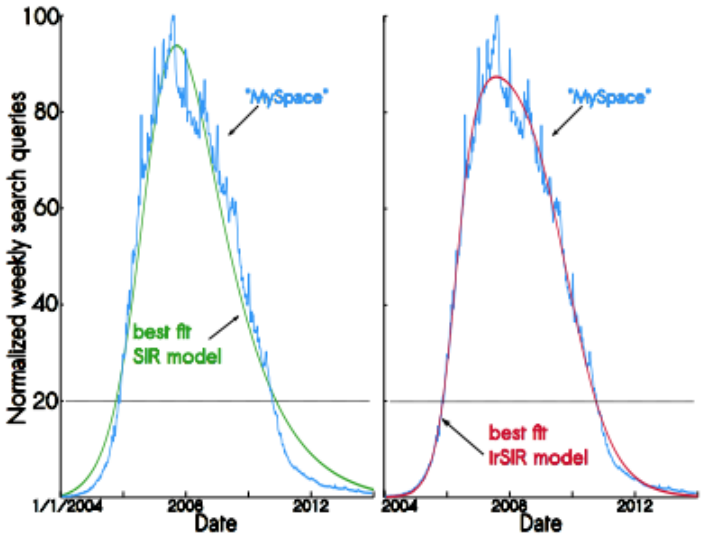


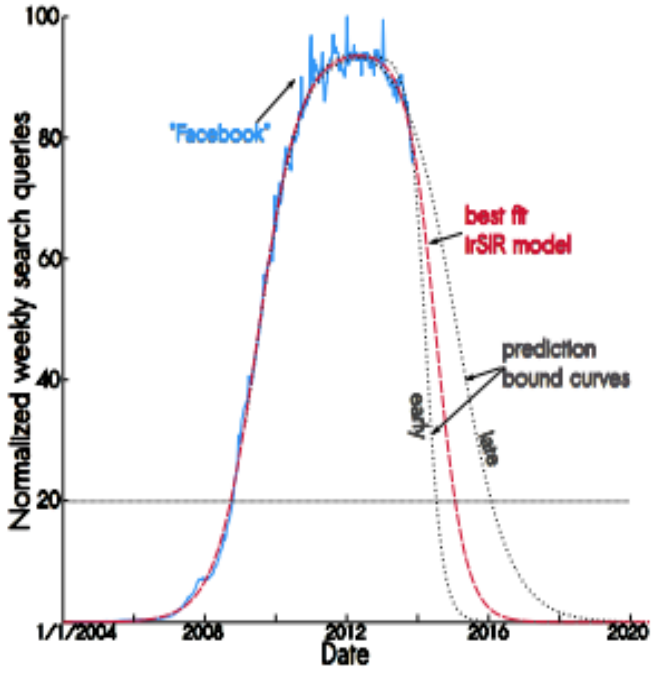
Source: http://www.vaughns-1-pagers.com/internet/google-certified-networks.htm

filename = adsense-ad-options.gif.v17

Data Makes Everything Clearer

Searches for "MySpace"



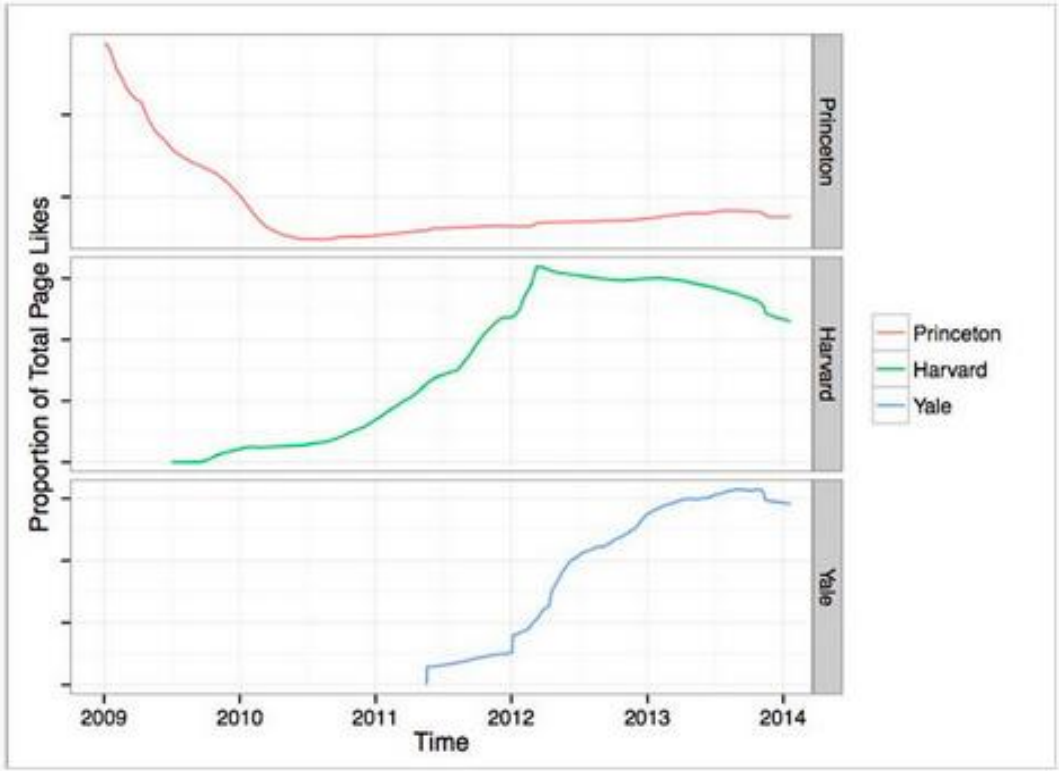


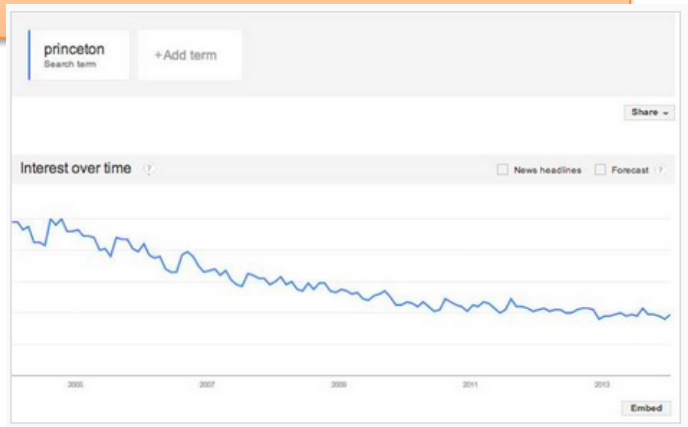
Searches for "Facebook"

Figure 3: **Data for search query "Myspace" with best fit (a) SIR and (b) irSIR models overlaid.** The search query data are normalized such that the maximum data point corresponds to a value of 100.

Data Makes Everything Clearer

In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:





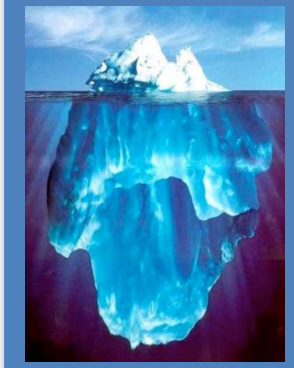
and based on Princeton search trends:

"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…

http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/

Data Sources

It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

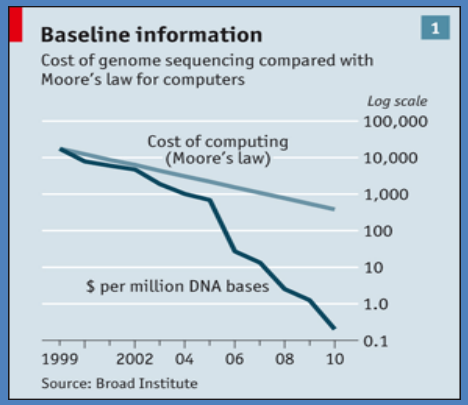
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

Health/Scientific Computing

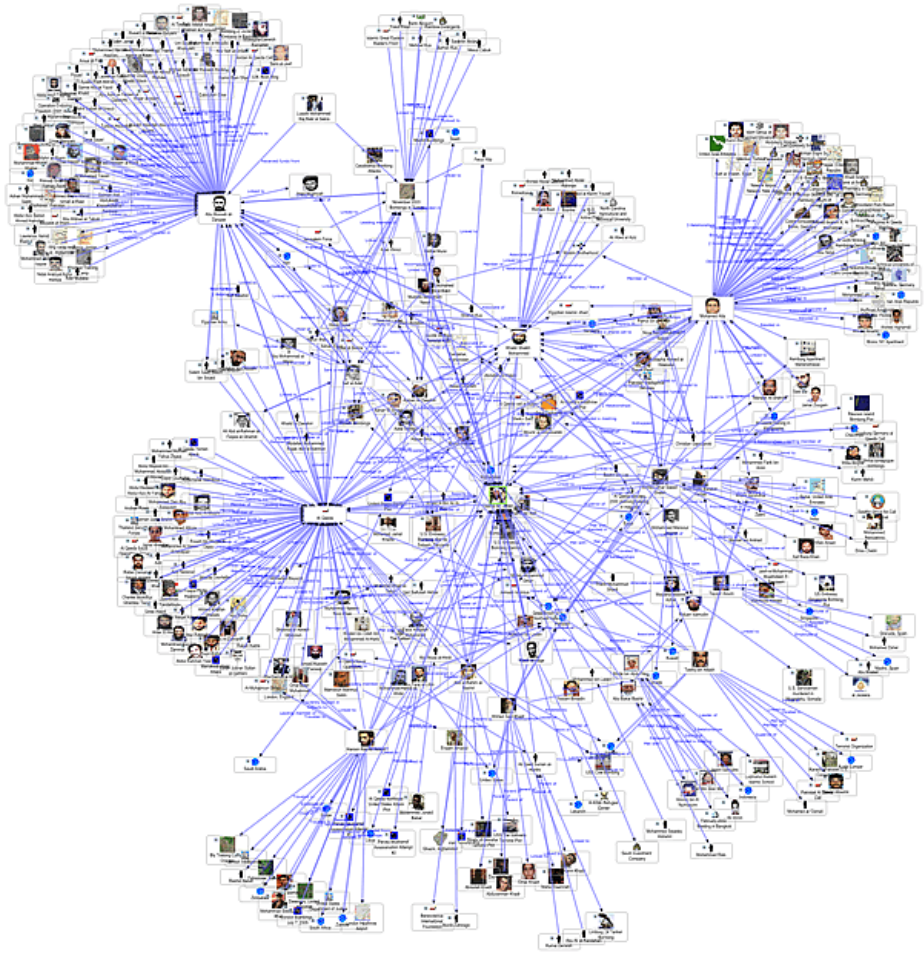


Graph Data

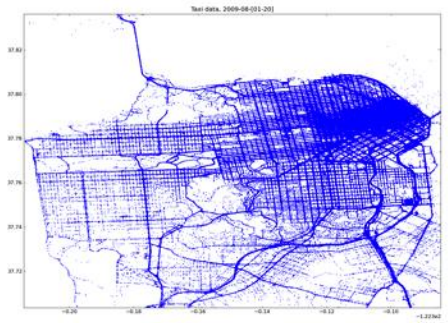
Lots of interesting data has a graph structure:

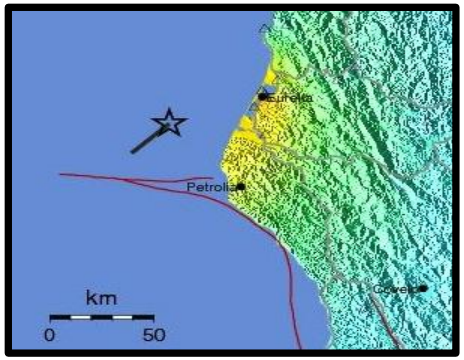
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

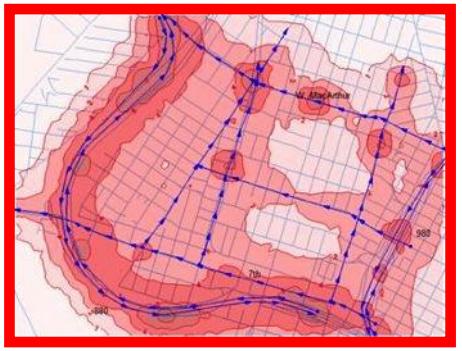
Some of these graphs can get quite large (e.g., Facebook* user graph)



What can you do with the data?







Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

From Alex Bayen, UCB

"Data Science" is so 2016

- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"

DATA SCIENCE – WHAT IS IT?

“Data Science” an Emerging Field

O’Reilly Radar report

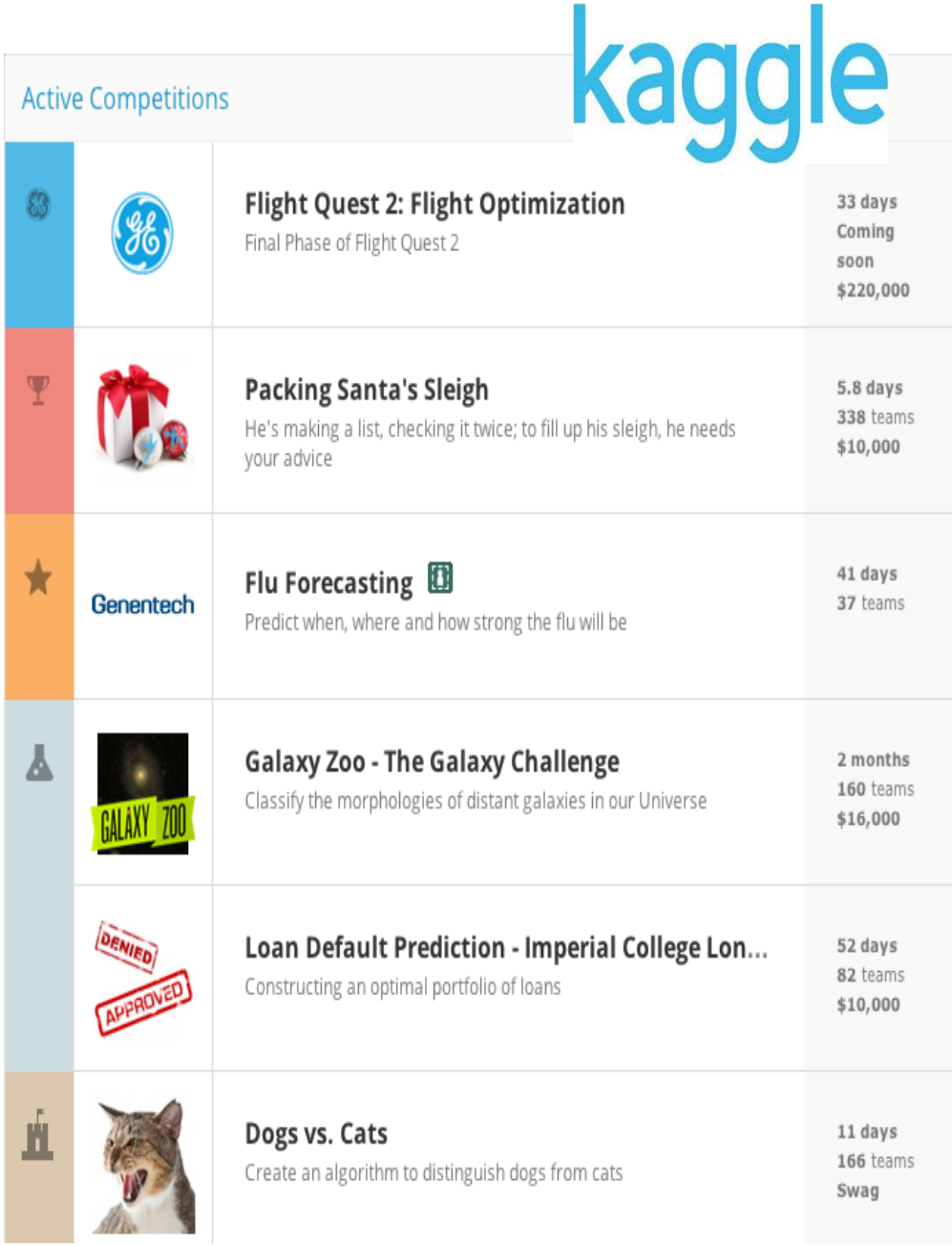
Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

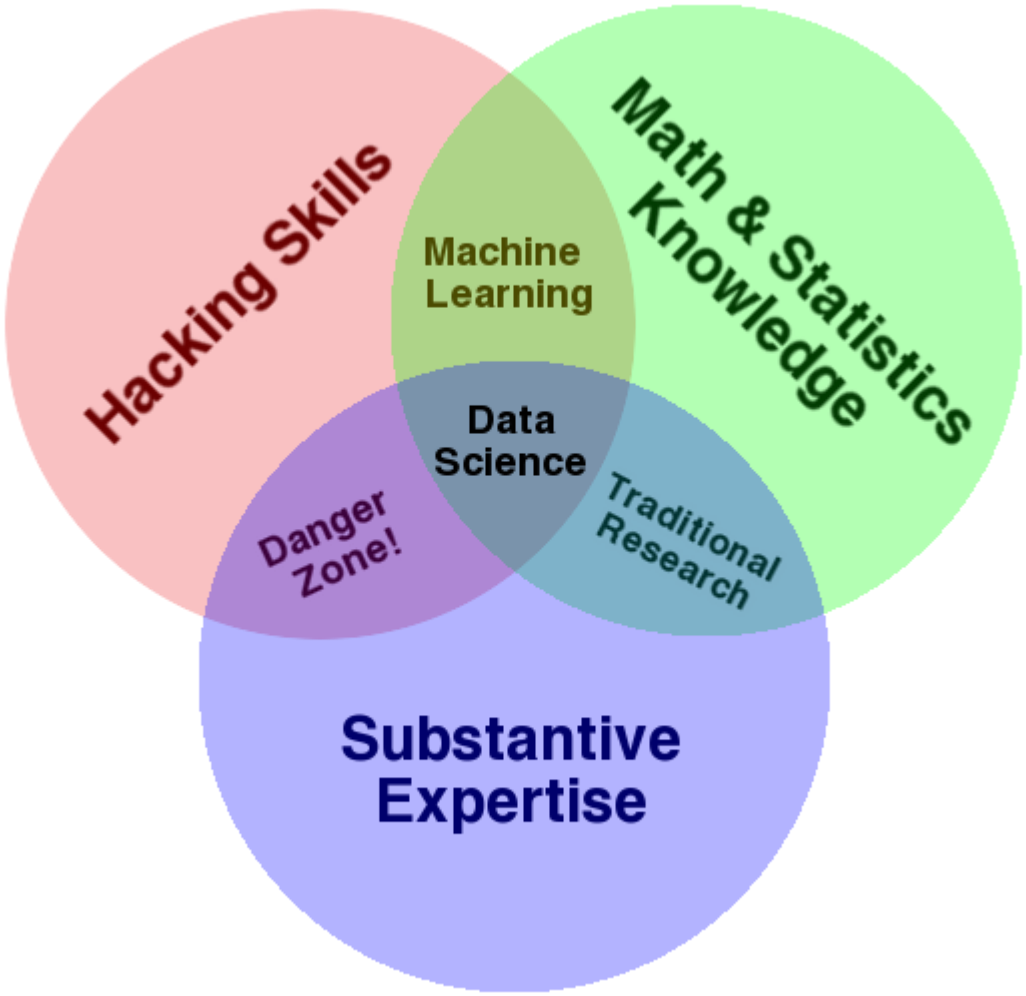
Goal of Data Science

Turn data into data products.

Some recent ML Competitions



Data Science – One Definition



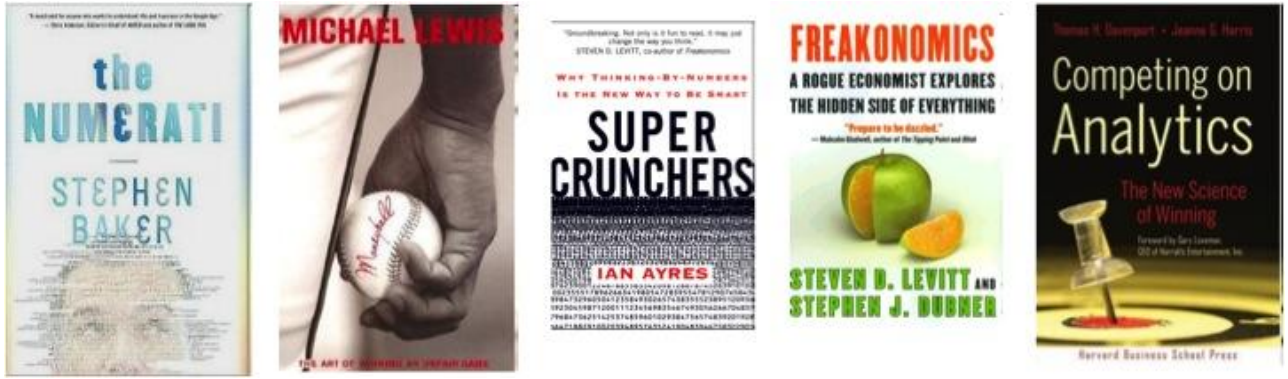
Contrast: Databases

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,…

ACID = Atomicity, Consistency, Isolation and Durability CAP = Consistency, Availability, Partition Tolerance

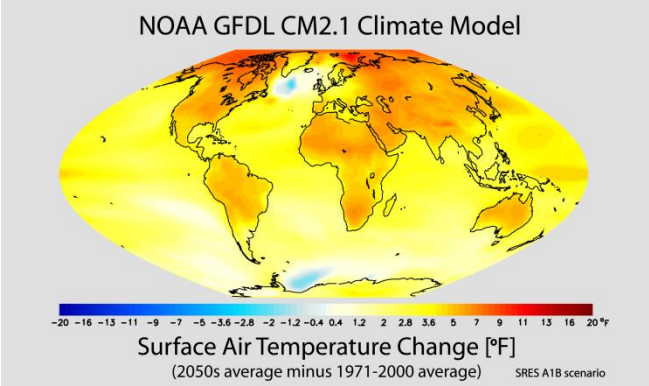
Contrast: Databases

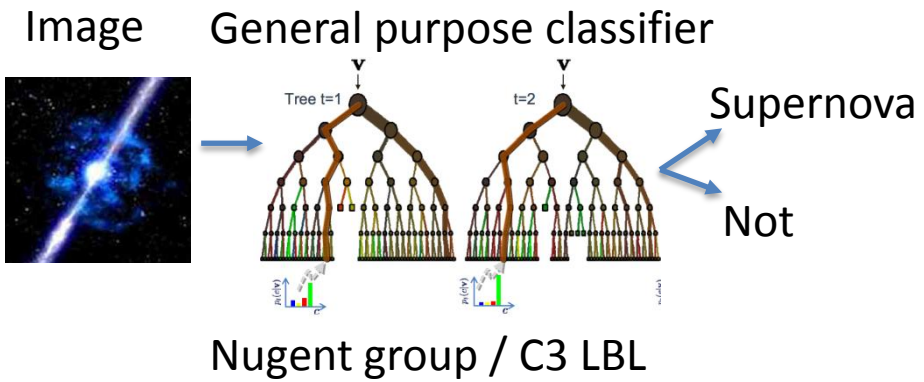
Databases	Data Science
Querying the past	Querying the future



Business intelligence (**BI**) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Contrast: Scientific Computing



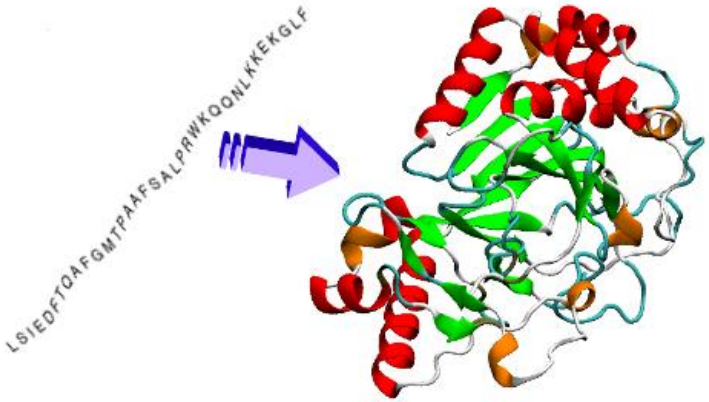


Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach
General inference engine replaces model
Structure not related to problem
Statistical models handle true randomness, and **unmodeled complexity**.
Run on cheaper computer Clusters (EC2)

Contrast: Computational Science

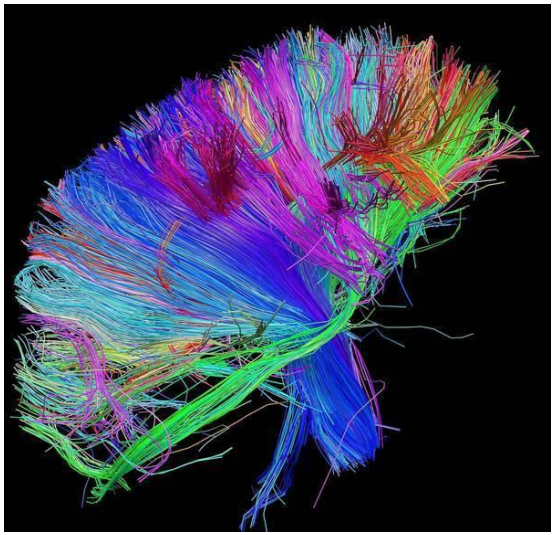
CASP: A Worldwide, Biannual Protein Folding Contest



Quark
Rich, Complex Energy Models
Faithful, Physical Simulation

Raptor-X
Data-intensive, general ML models
Feature-based inference
Conditional Neural Fields

Brain Mapping: Allen Institute, White House, Berkeley

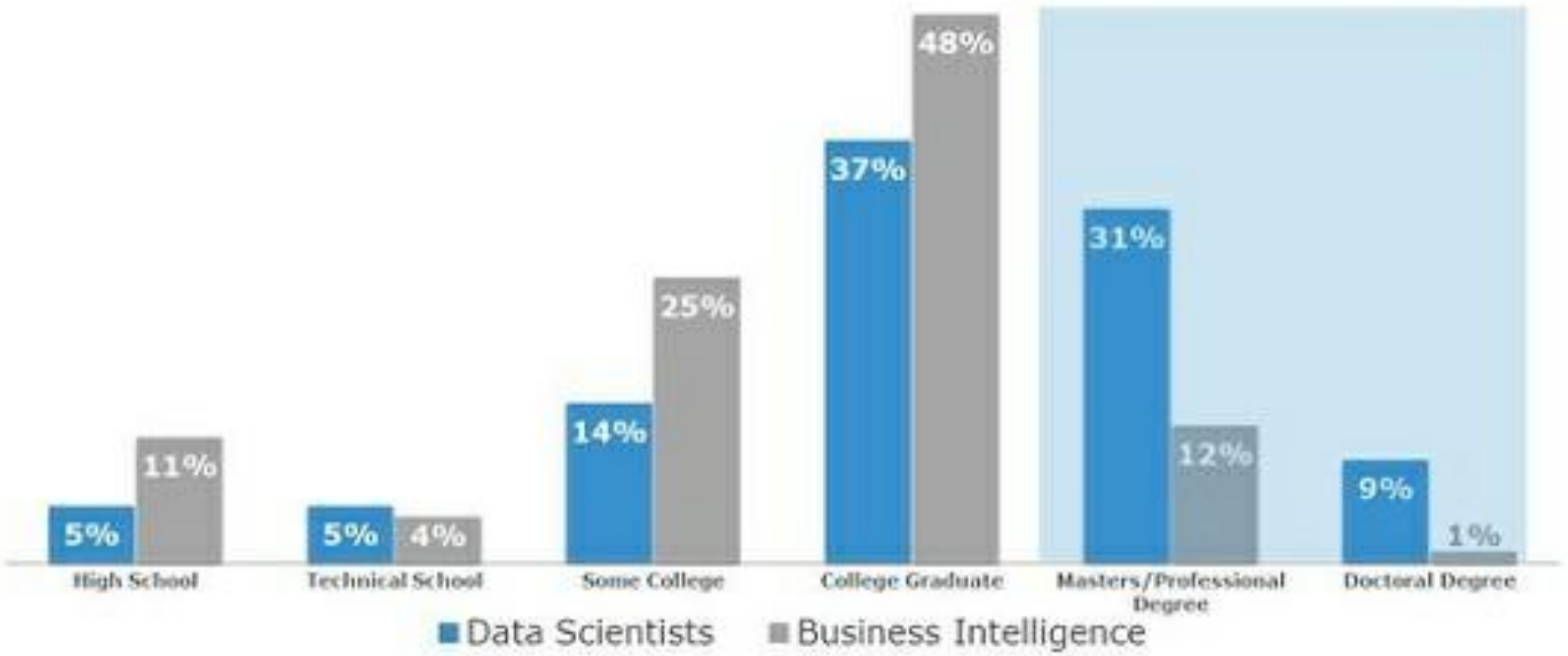


Techniques (Massive ML)
Principal Component Analysis
Independent Component Analysis
Sparse Coding
Spatial (Image) Filtering

Contrast: Machine Learning

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

Data science requires greater education



40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

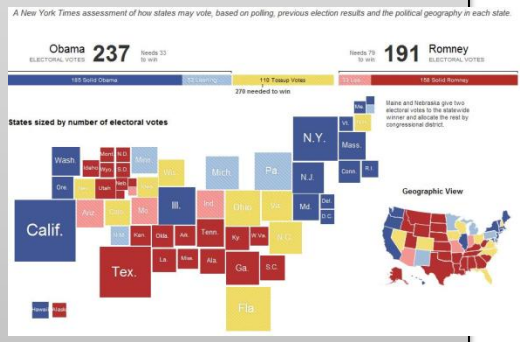
DOING DATA SCIENCE

Ben Fry's Model

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

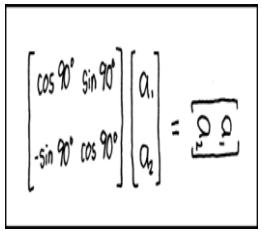
1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results



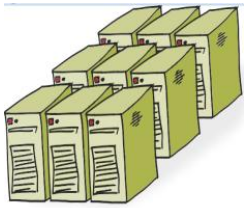
Data Scientist's Practice

Digging Around in Data

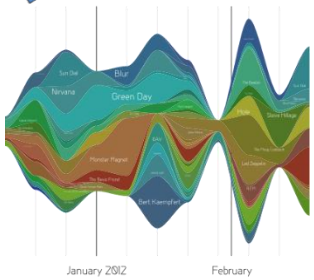
Clean, prep



Hypothesize Model



Large Scale Exploitation



Evaluate Interpret

What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

Machine Learning is Everywhere…

- Handwriting Recognition
- Speech Recognition
- Automatic translation
- Credit-card fraud detection
- Image Classification
- Social Networks Analysis (community detection)
- Movie / product / article recommendations
- Autonomous cars
- ….

What is Machine Learning?

Wikipedia:

Machine Learning, a branch of artificial intelligence, concerns the construction and study of systems that can **learn** from data.

Tom Mitchell (1998): A computer program is said to **learn from experience**, with respect to some task and some performance measure if its performance, as measured by the performance measure, **improves with experience.**

“Child Learning”

Action	Reaction	Lesson
Touching hot stove	aching hand	Do not touch again
Playing with toys	Fun	Continue playing
Running in to the road	Screaming parent	Don’t run to roads
Running in the house	Fun	Run in the house
Eating chocolate	Fun	Search for chocolate
Eating too much chocolate	Stomach ache	Don’t eat too much
Saying bla bla	No Reaction	Try variations
Saying daddy	Overexcited parents	Do that again

Learning from Examples

What is “Dangerous”?

Typical Machine Learning Tasks

No two machine learning tasks are identical, but still there are common prototypes:

- **Supervised Learning**
  - Learning from labeled examples (for which the answer is known)
- **Unsupervised Learning**
  - Learning from unlabeled examples (for which the answer is unknown)
- Semi-supervised Learning
  - Learning from both labeled and unlabeled examples
- Active Learning
  - Learning while interactively querying for labels of examples
- Reinforcement Learning
  - Learning by trial and feedback, like the "child learning" example

Typical Machine Learning Tasks

Supervised Learning

Estimate an **unknown result**, given explicit values of some explaining variables ("features").

Estimate it based on a set of observations for which **both the result and the explaining variables are known** ("training set").

This may be prediction ("it's difficult to give forecasts, especially about the future") or estimation.

Typical Machine Learning Tasks

Supervised Learning

Example 1: What will be the annual spend of my clients?

- **The unknown result**: the annual spend (this is a prediction)
- **Explaining variables ("features")**: Client's details (e.g., domain, size, purchase history)
- **Training set**: The annual spend in past years, with respect to the client's data available so far (at the beginning of that year)

Typical Machine Learning Tasks

Supervised Learning

Two main tasks are considered in Supervised Learning:

- **Regression:** the unknown result is a **numerical value** (e.g., annual spend)
- **Classification:** the unknown result is a **class relation** (e.g., the activity)

Regression and classification have different objective measures, and often different algorithms.

Typical Machine Learning Tasks

Unsupervised Learning

Given explicit values of some variables (pre-defined set), extract interesting **patterns** that appear in the data, or provide an **insightful representation** of the data inherent distribution.

Typical Machine Learning Tasks

Unsupervised Learning

Example: Market Segmentation

- Input data: Clients information
- Objective: Identify what types of clients are there?
  - This objective is known as 'Clustering' or 'Cluster Analysis'

Supervised Learning

X_1	X_2	X_3	…	X_{n-2}	X_{n-1}	X_n	Y
$x_{1,1}$	$x_{2,1}$	$x_{3,1}$	…	$x_{n-2,1}$	$x_{n-1,1}$	$x_{n,1}$	y_1
$x_{1,2}$	$x_{2,2}$	$x_{3,2}$	…	$x_{n-2,2}$	$x_{n-1,2}$	$x_{n,2}$	y_2
. . .	. . .	. . .	… … …	. . .	. . .	. . .	
$x_{1,m-1}$	$x_{2,m-1}$	$x_{3,m-1}$	…	$x_{n-2,m-1}$	$x_{n-1,m-1}$	$x_{n,m-1}$	y_{m-1}
$x_{1,m}$	$x_{2,m}$	$x_{3,m}$	…	$x_{n-2,m}$	$x_{n-1,m}$	$x_{n,m}$	y_m

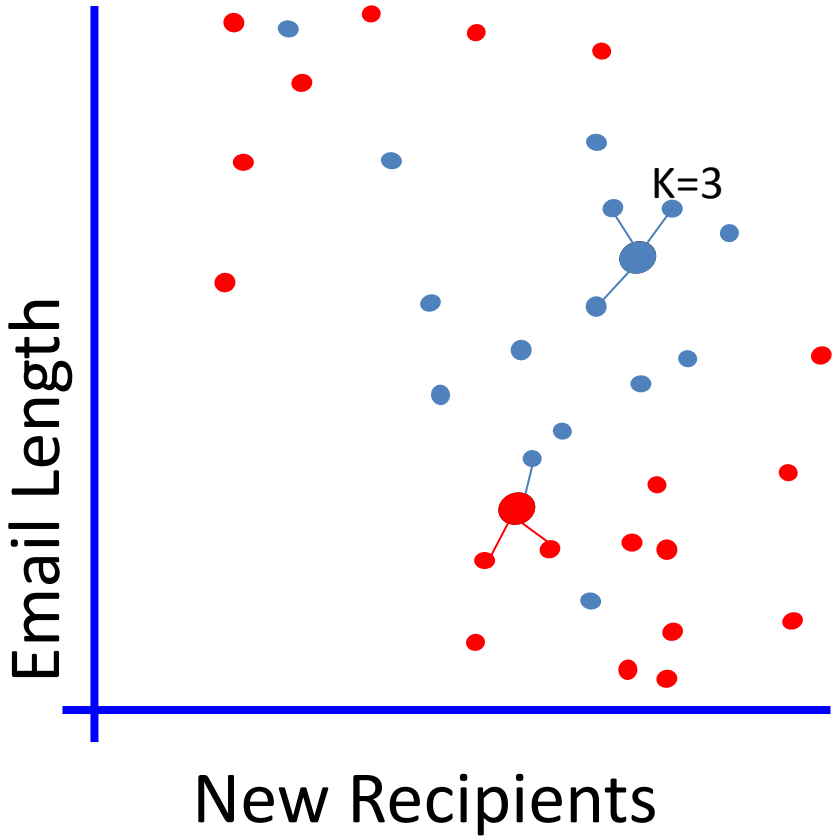
- Uses a set of labeled examples with **known answer** ("training set")
- Success is evaluated on a separate set of examples ("test set").

  Various success criteria may be considered:

  - For classification: Accuracy, Recall, Precision…
  - For regression: MSE, RMSE,…

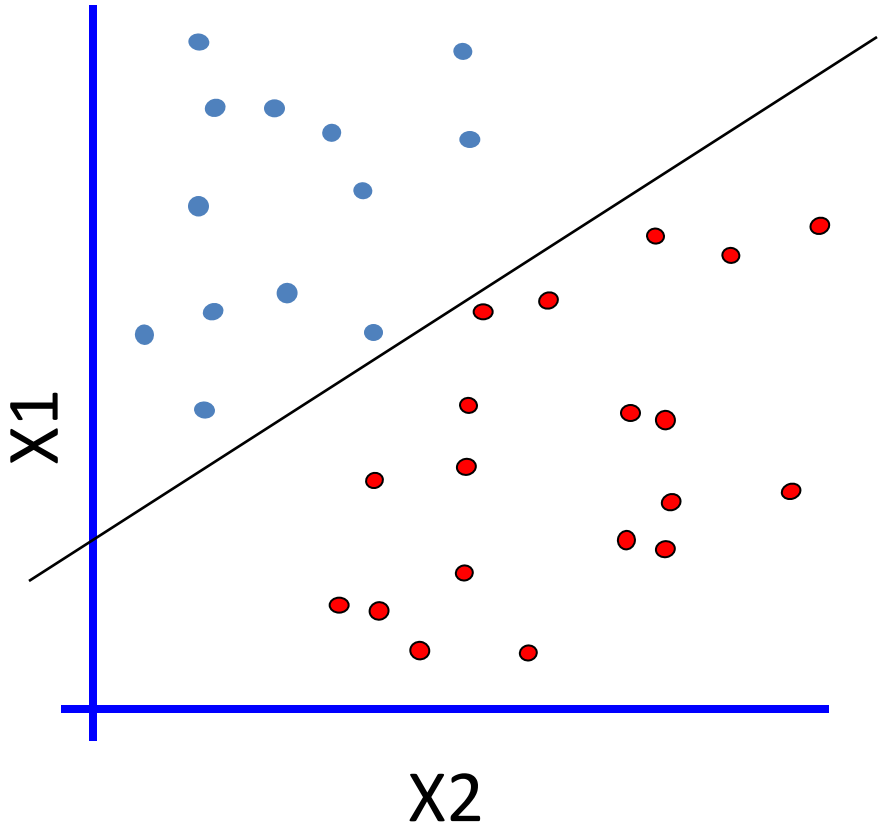
Lazy Learner: k-Nearest Neighbors

Identifying spam emails



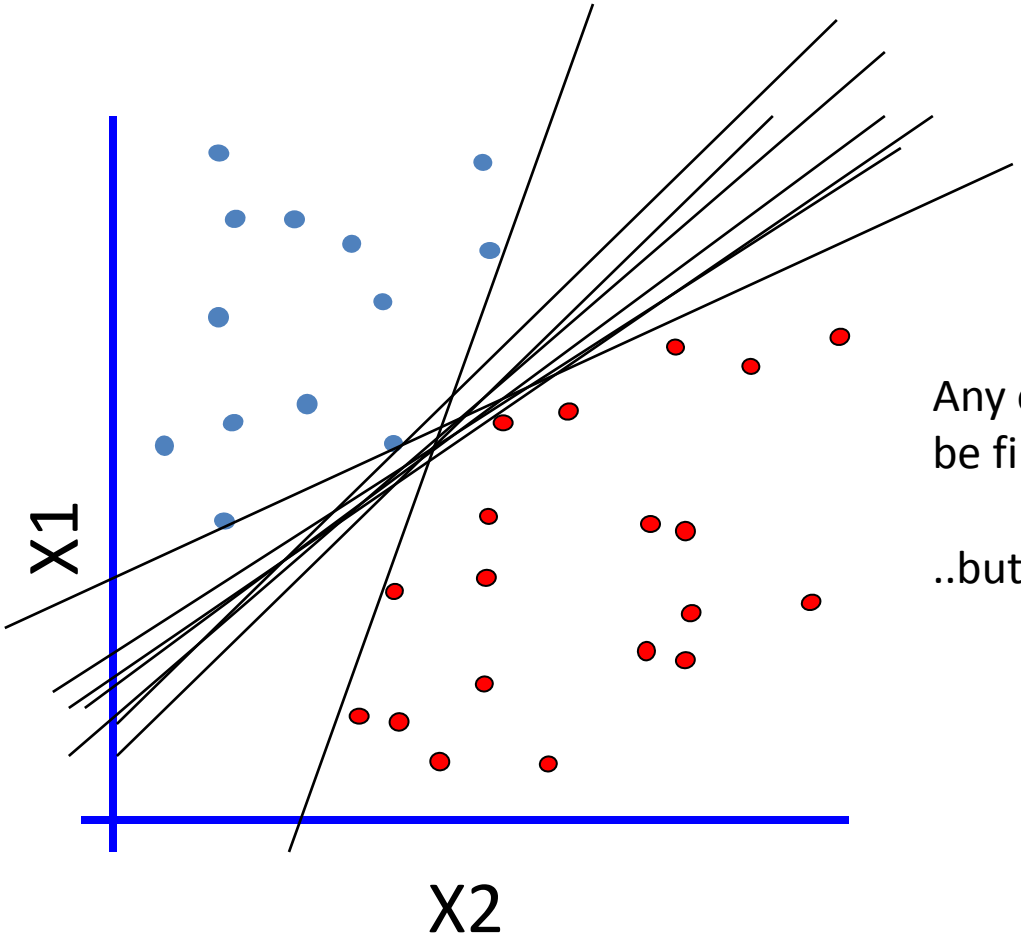
- What should be k?
- Which distance measure should be used?
- Computation

Linear Classifiers



How would you classify this data?

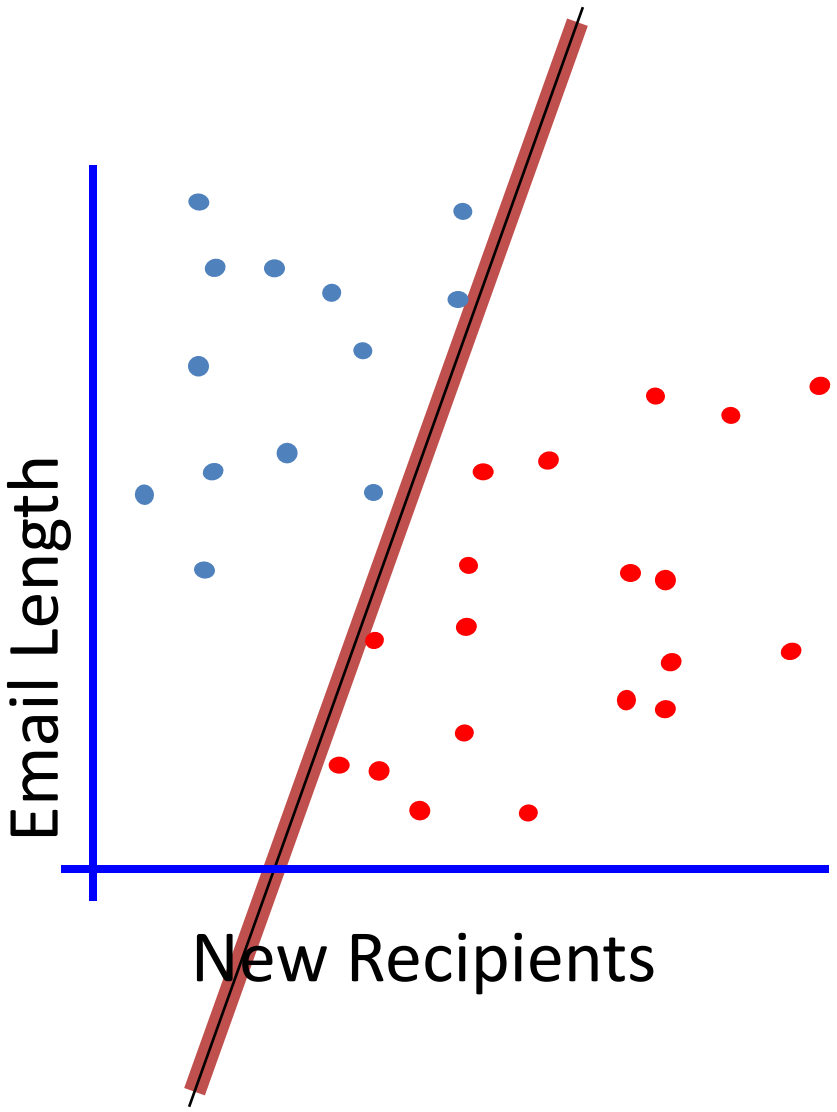
Linear Classifiers



Any of these would be fine..

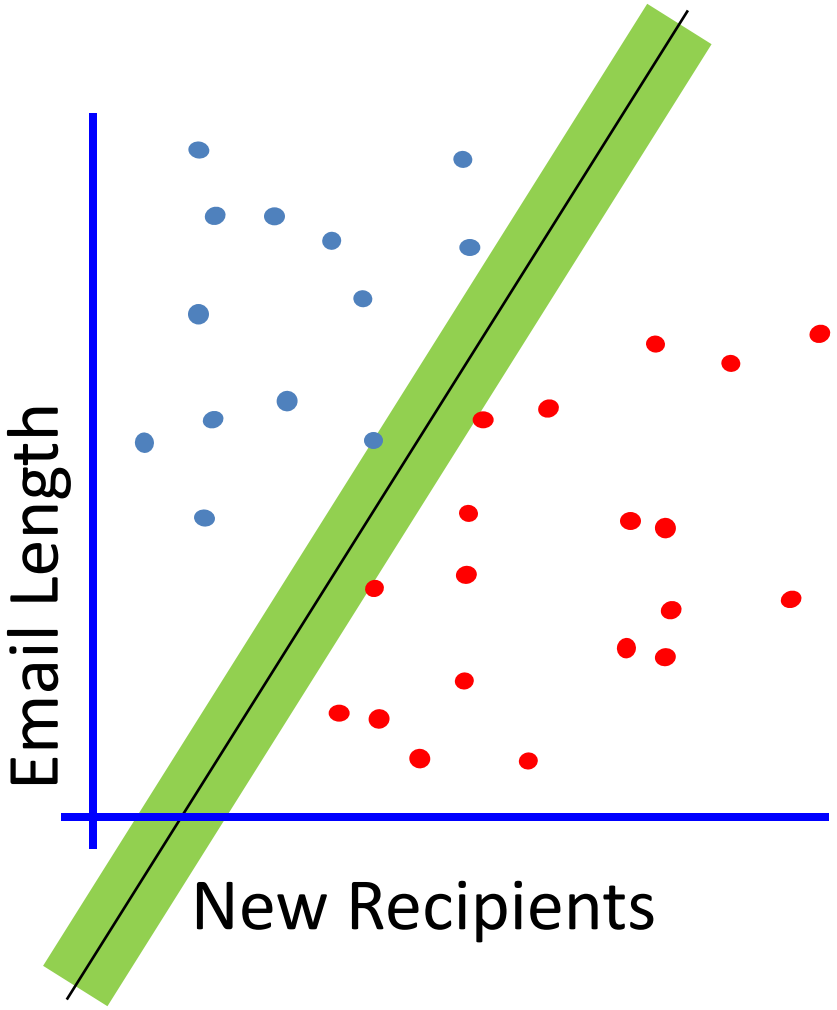
..but which is best?

Maximum Margin



Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a data point.

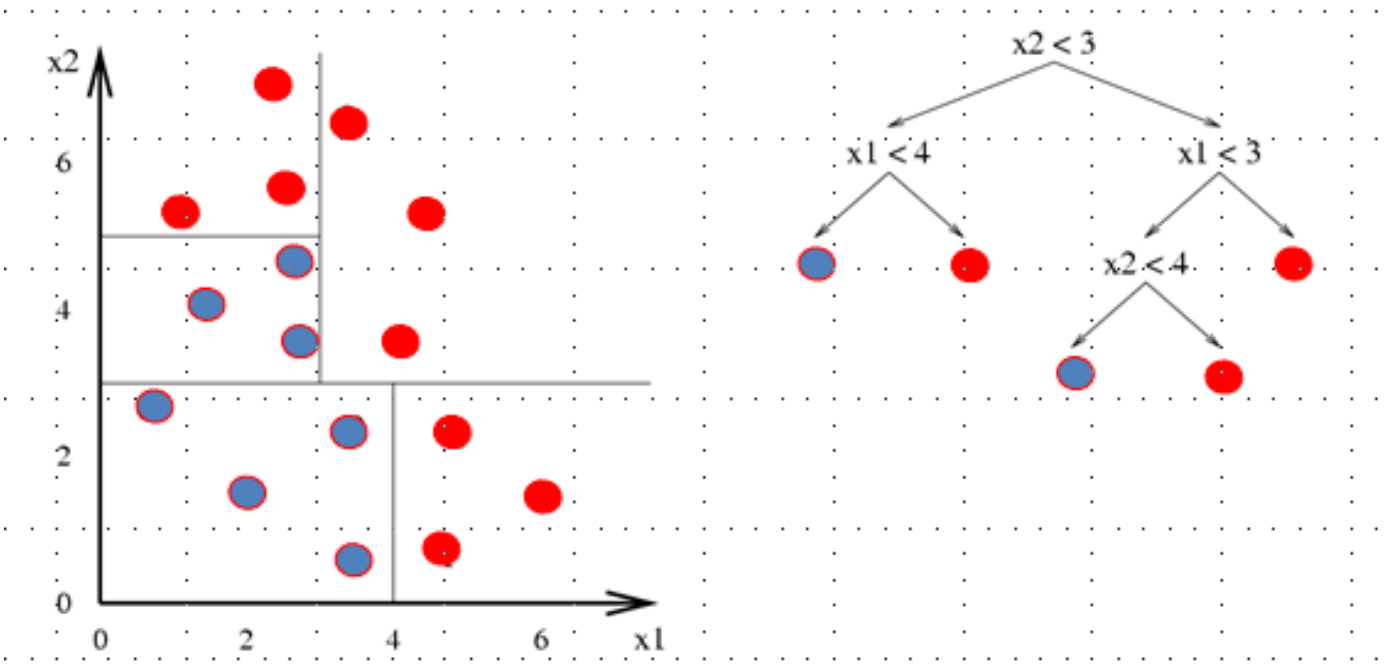
Maximum Margin



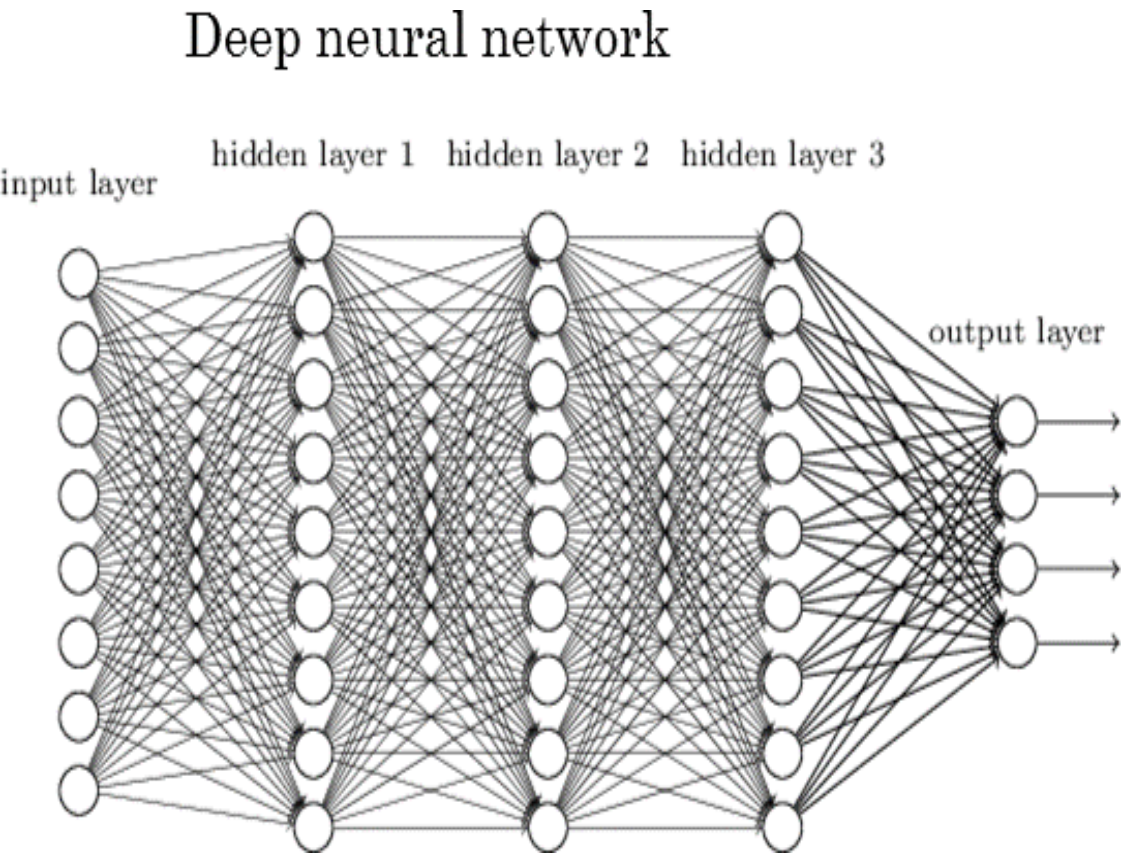
The maximum margin linear classifier is the linear classifier with the maximum margin.
This is found by the SVM algorithm (Support Vector Machine)

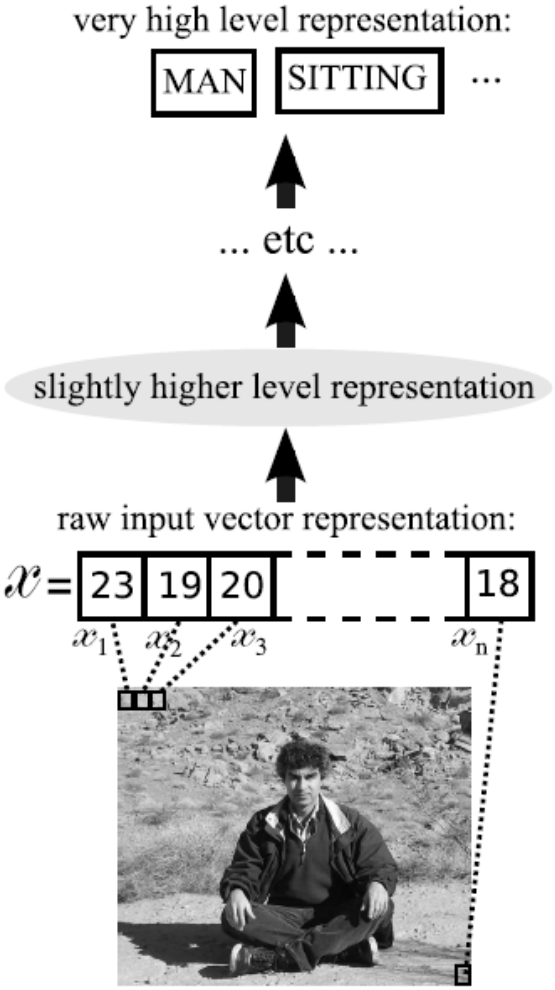
Decision tree

- A flow-chart-like tree structure
- Internal node denotes a test on one of the features
- Branch represents an outcome of the test
- Leaf nodes represent class labels



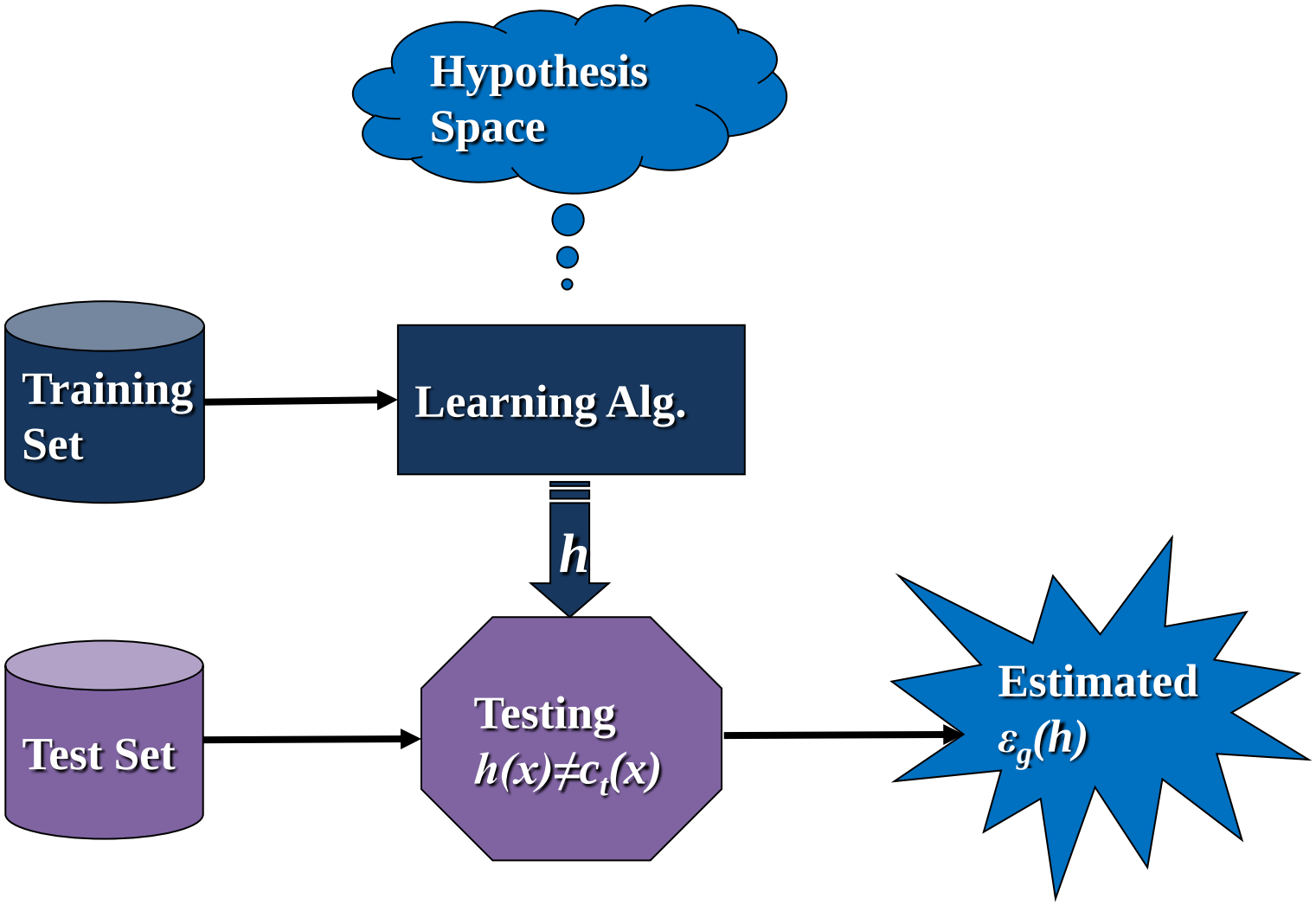
DEEP NEURAL NETWORKS





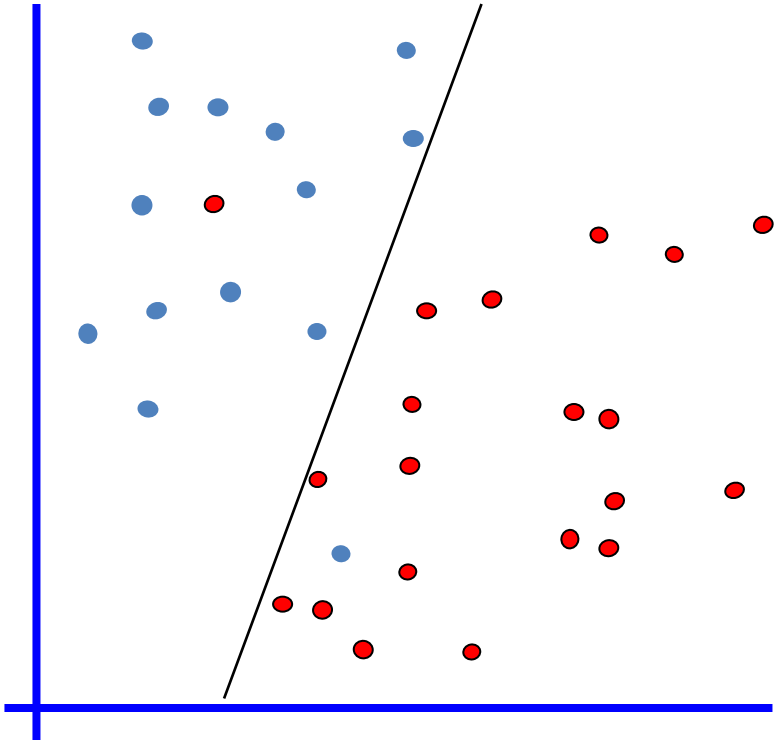
Bengio, 2009

Block Diagram of a Supervised Learning System



Evaluating What's Been Learned

1. Test set
2. Cross Validation

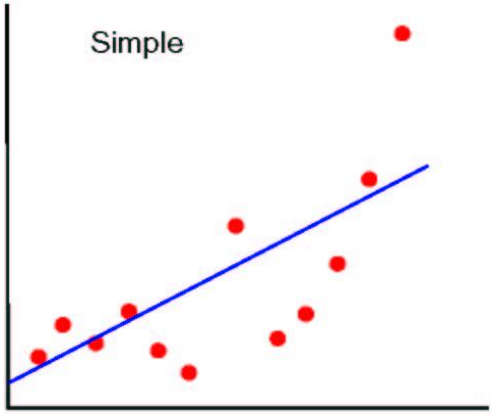


Confusion Matrix

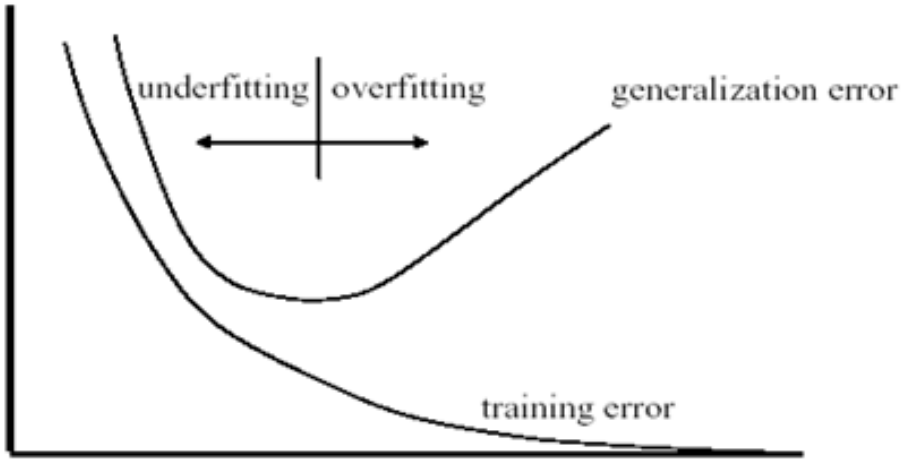
Classified As

Actual		Blue	Red
	Blue	7	1
	Red	0	5

Regression Learning Example



Overfitting and Underfitting



Overfitting: The model learns the training set too well – it over fits the training set such that it cannot generalize to new instances.

Underfitting: the model is too simple, both training and test errors are large

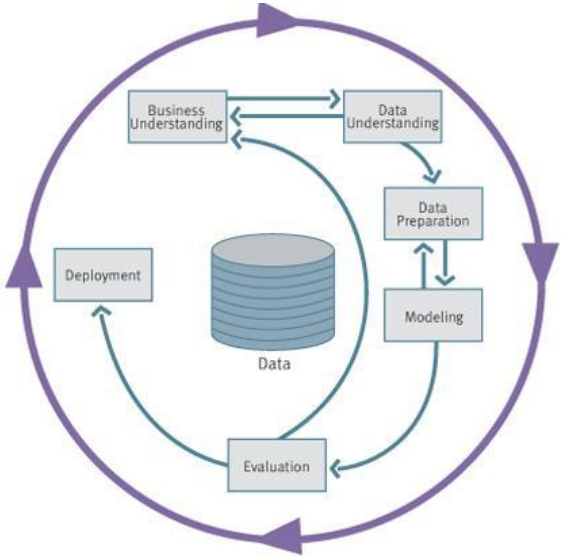
CRISP-DM Methodology

- CRISP-DM stands for **Cross Industry Standard Process for Data Mining**
  - Conceived in 1996-7 by SPSS, Teradata, Daimler, NCR and OHRA
  - IBM is the primary corporation that embraced and incorporated it in its SPSS Modeler product
  - CRISP-DM defines a methodology for ML/DM projects

CRISP-DM Methodology

CRISP-DM breaks the process of data mining into six major phases

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



The sequence of the phases is not strict and moving back and forth between different phases may be required

Summary

We briefly discussed today:

- What is Machine Learning
- Typical Machine Learning tasks
- Supervised Learning:
  - Learning means Generalization
  - Overfitting and Underfitting
  - Simple learning paradigms
  - Training vs. Testing
  - Classification and Regression
- CRISP-DM