Foundations of Data Science
Lecture 1

Rumi Chunara, PhD

CS3943/9223

Today

- What is Data Science?
- Data Handling
- R intro
- Statistics Review
- Polling YOU
- About the course

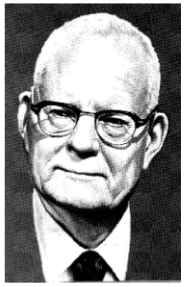
Why We've Analyzed Data Has Had Different Focuses Over Time

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

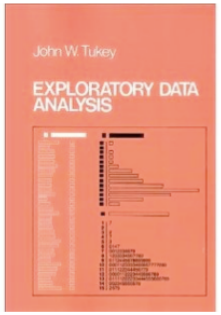
W.E. Demming



1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"



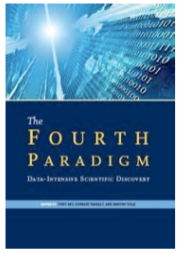
1989: "Business Intelligence"

Howard Dresner

1997: "Machine Learning"

2010:"The Data Deluge"

2007:"The Fourth Paradigm"



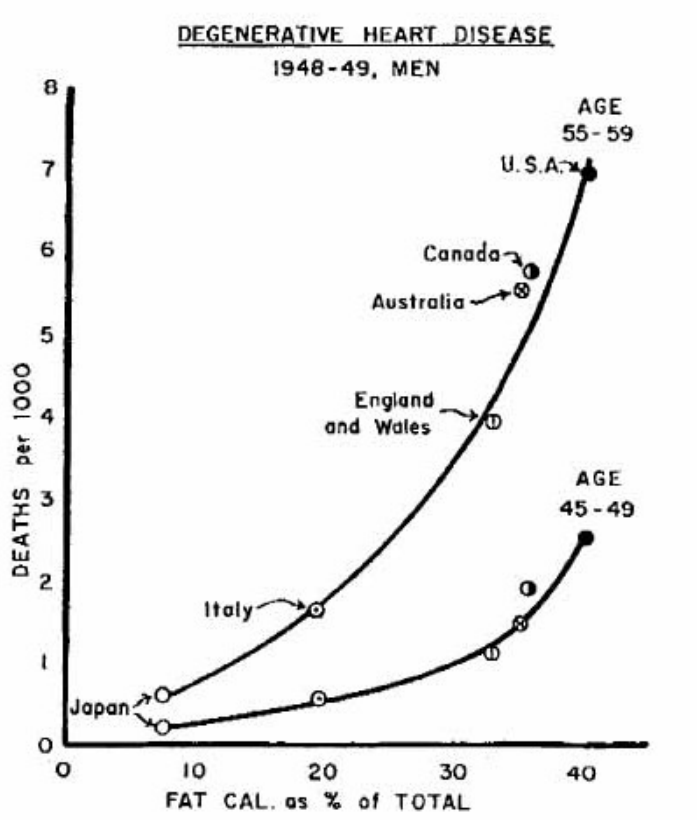
1996: Google

Google™

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

In General: Data Helps Solve Problems

- Seven Countries Study (Ancel Keys)
- 13,000 subjects total, 5-40 years follow-up.



In General: Data Helps Solve Problems

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.

In General: Data Helps Solve Problems

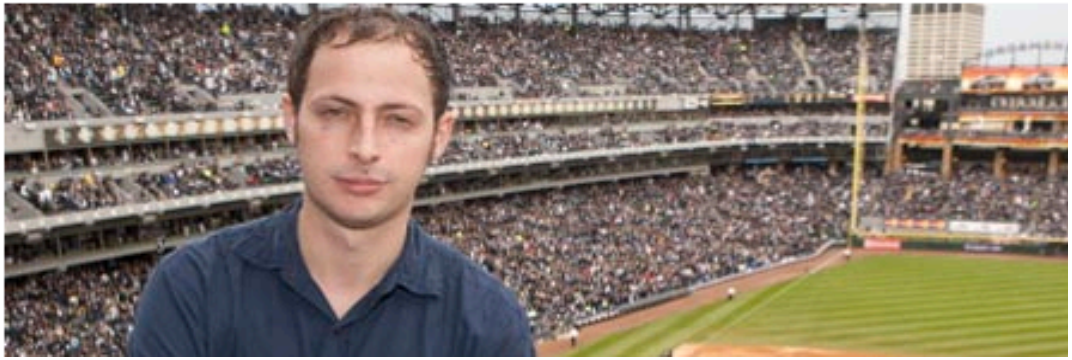
elections2012

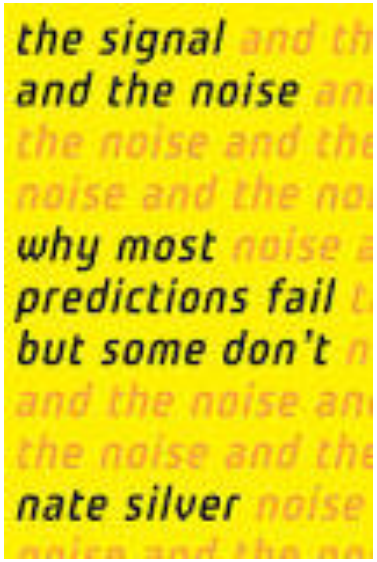
Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST





Data and Election 2012

- …that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, helped save the president's candidacy. In Chicago, the campaign recruited a team of behavioral scientists to build an extraordinarily sophisticated database

  …that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. The power of this operation stunned Mr. Romney's aides on election night, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

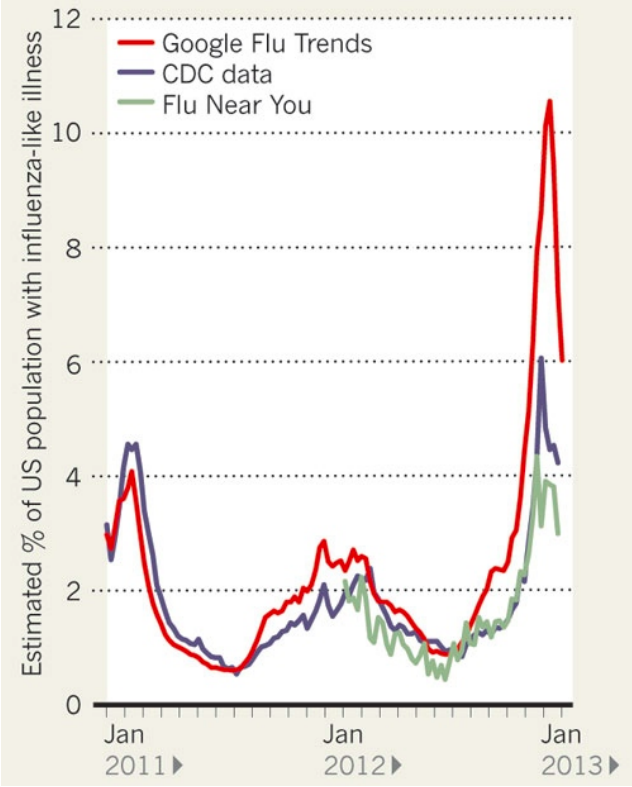
New York TImes, Wed Nov 7, 2012

More Data Brings New Challenges

nature International weekly journal of science

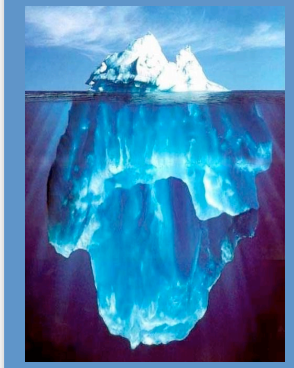
NATURE | NEWS

When Google got flu wrong



Data Sources

It's All Happening On-line



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

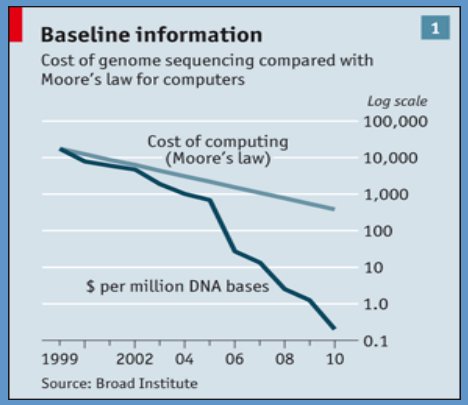
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

Health/Scientific Computing

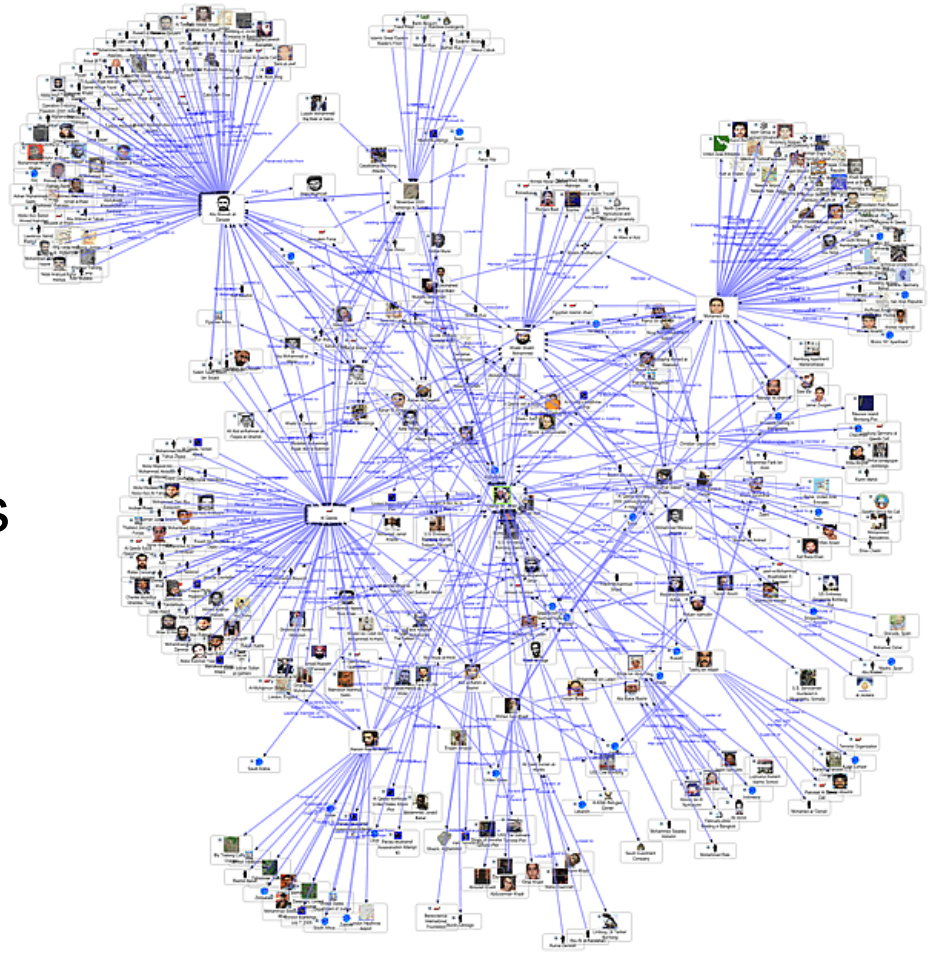


Graph Data

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)



Question Types

- **Simple (descriptive) Stats**
  - What are the genomic profiles of one group
- **Hypothesis Testing**
  - Is there a difference in movement of different people
- **Segmentation/Classification**
  - What are the common characteristics of customers
- **Prediction**
  - On what week will influenza peak in New York City?

Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]

1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

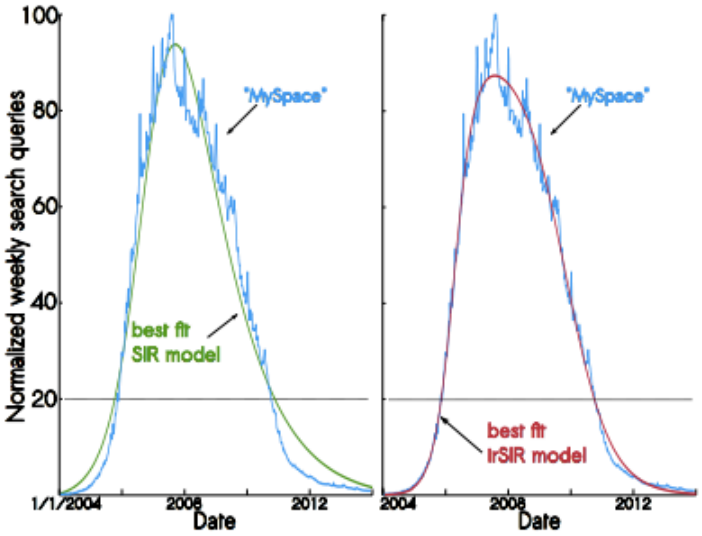
*** E-mail: Corresponding spechler@princeton.edu**

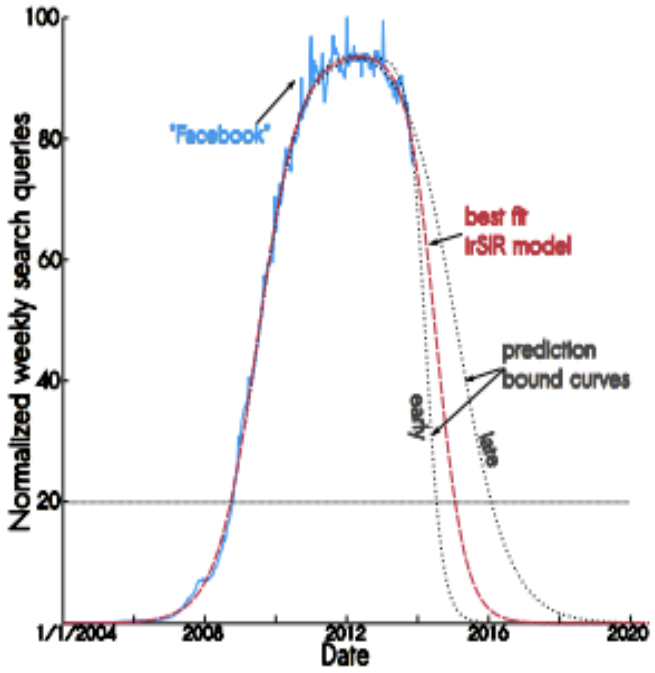
Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

Data Makes Everything Clearer?

Searches for
"MySpace"



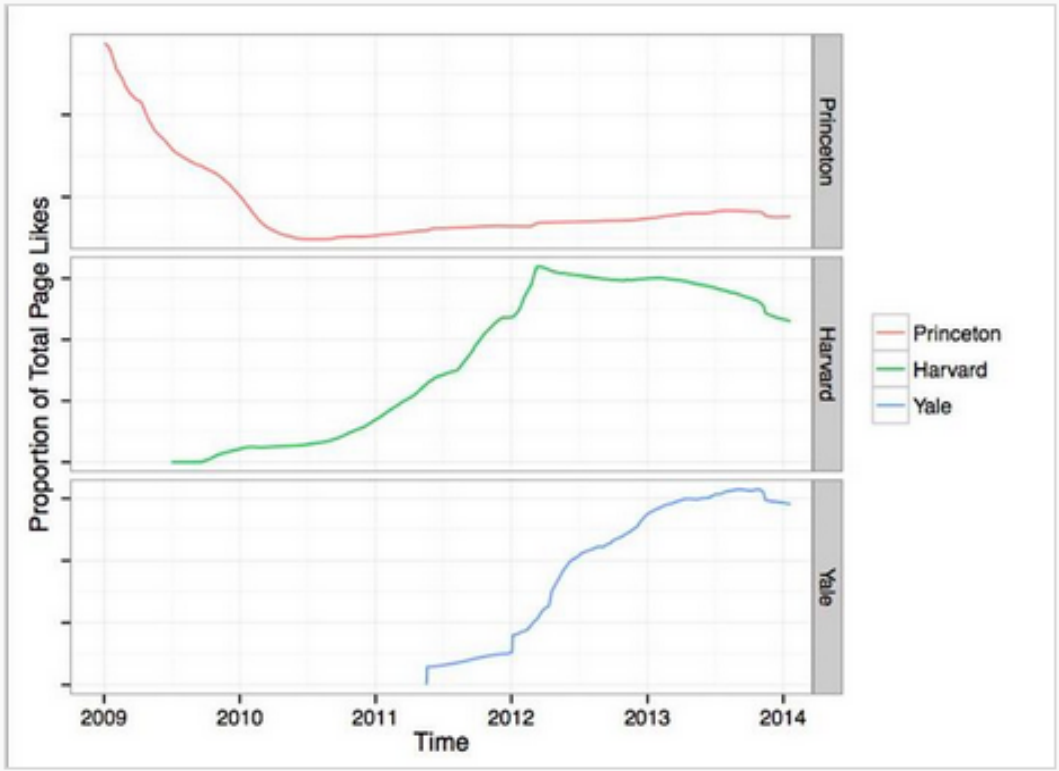


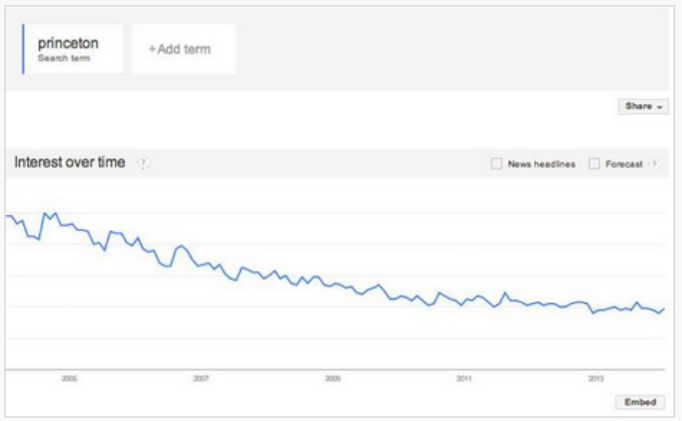
Searches for
"Facebook"

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/

"Big Data" is so 2012

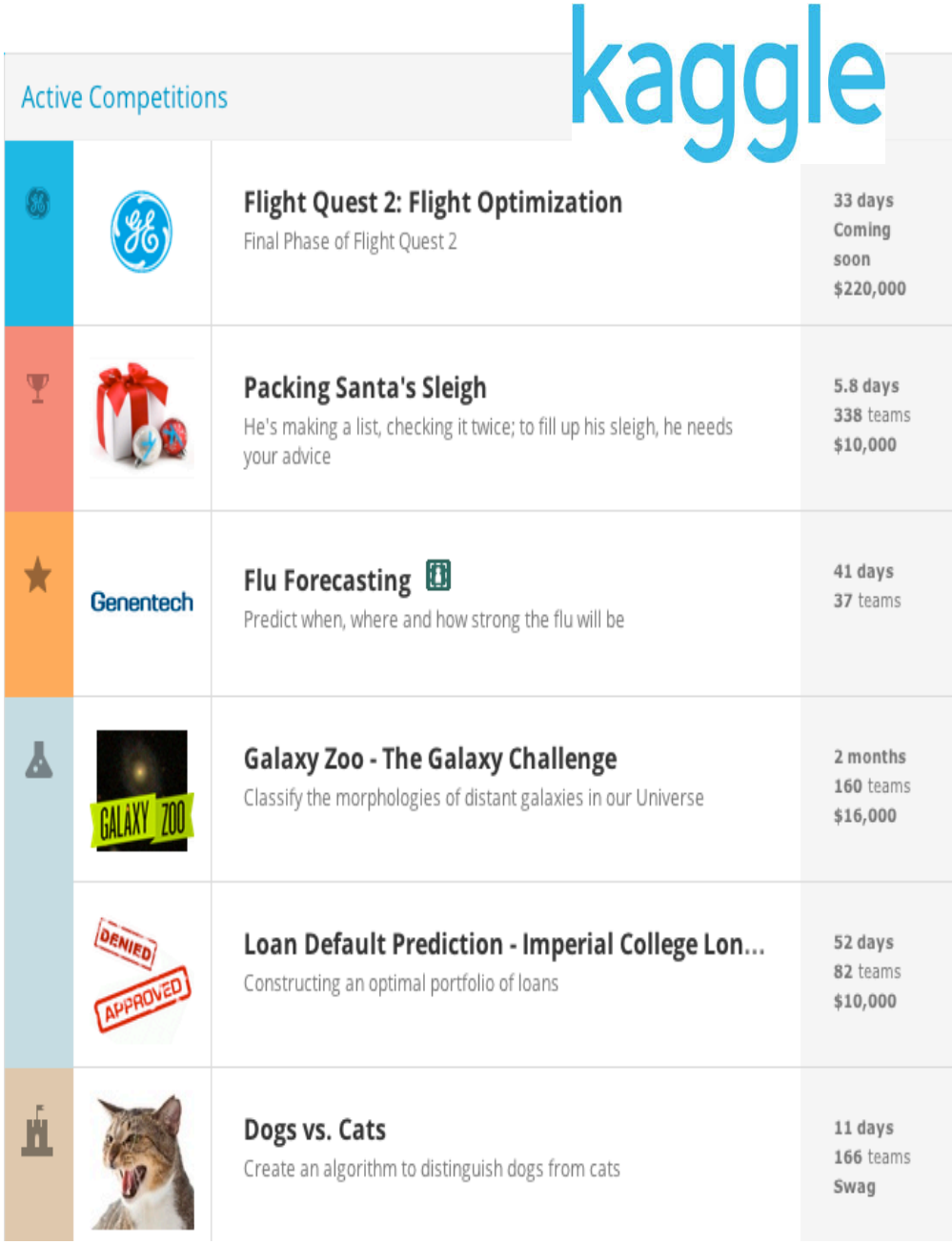
- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, 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?

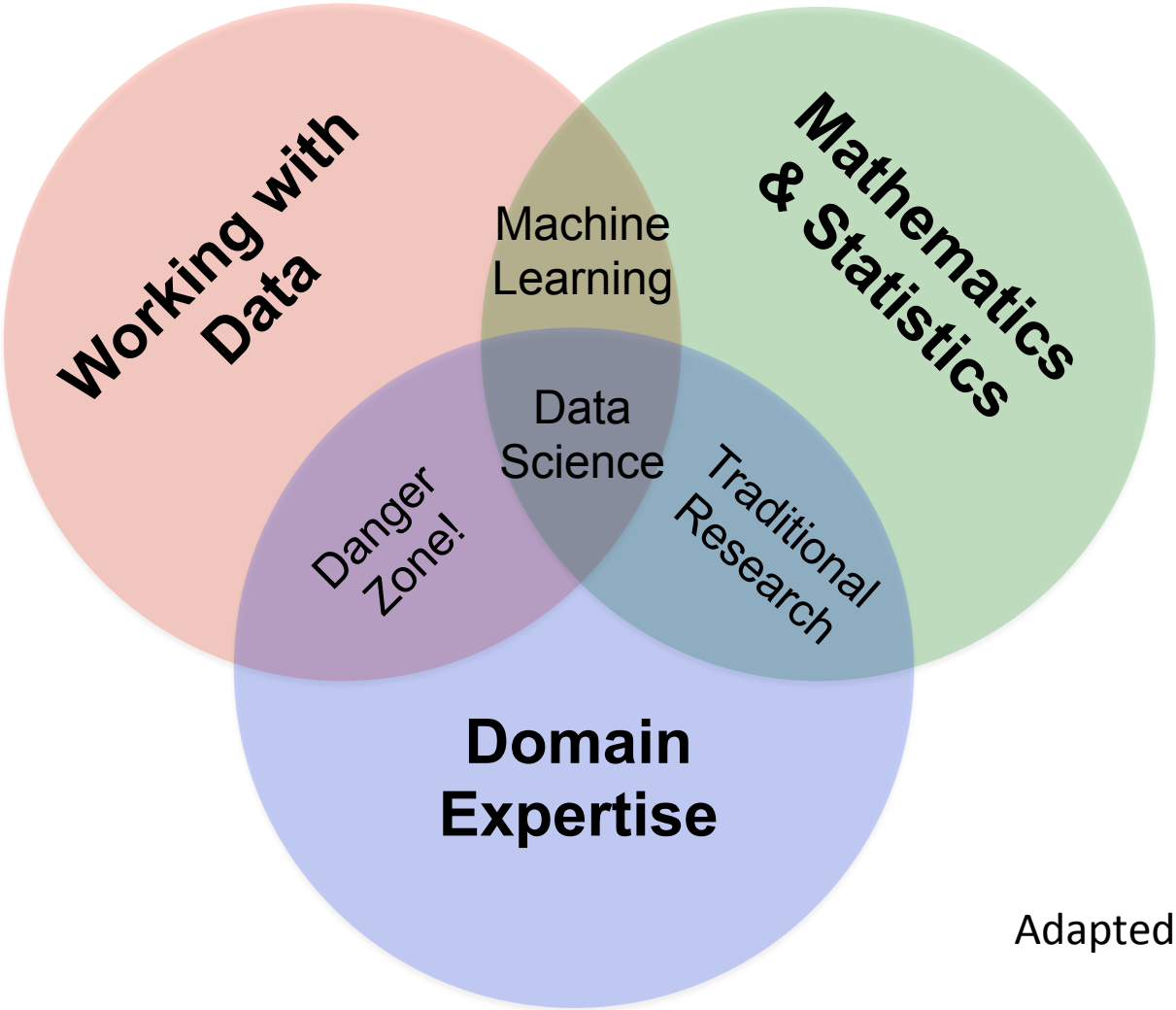
“Data Science” an Emerging Field

O’Reilly Radar report

Some recent DS Competitions



Data Science – One Definition

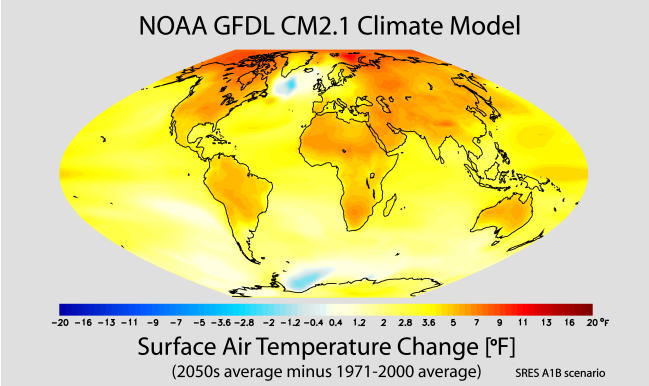


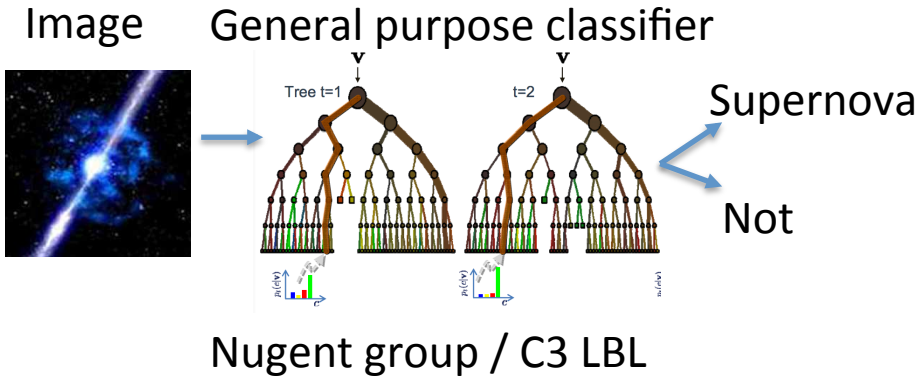
Adapted from Drew Conway

Databases vs. Data Science

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
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,…

Scientific Computing vs. Data Science





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

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

Machine Learning vs. Data Science

Machine Learning
Develop new (individual) models
Prove mathematical properties of models
Improve/validate on a few, relatively clean, small datasets
Publish a paper

Data Science
Explore many models, build and tune hybrids
Understand empirical properties of models
Develop/use tools that can handle massive datasets
Take action!

How to Learn Data Science?

- Masters Programs
- Get a *different* masters
- Work in Data Science

Doing Data Science

1) What Is A Data Scientist?
2) Data Science Workflow

What is A Data Scientist?

Zvi
@nivertech

Follow

"Data Scientist" is a Data Analyst who lives in California.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
140	40

9:55 PM - 14 Mar 2012

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
907	418

12:55 PM - 3 May 2012

Javier Nogales
@fjnogales

Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1

FAVORITES 5

9:08 AM - 27 Jan 2014

What is your definition?

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data "in production"**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

Today:

- Many mature off-the shelf tools for analysis exist
- Skills in **problem-solving**, **collaborating** and **communicating** are needed

Data Science Workflow

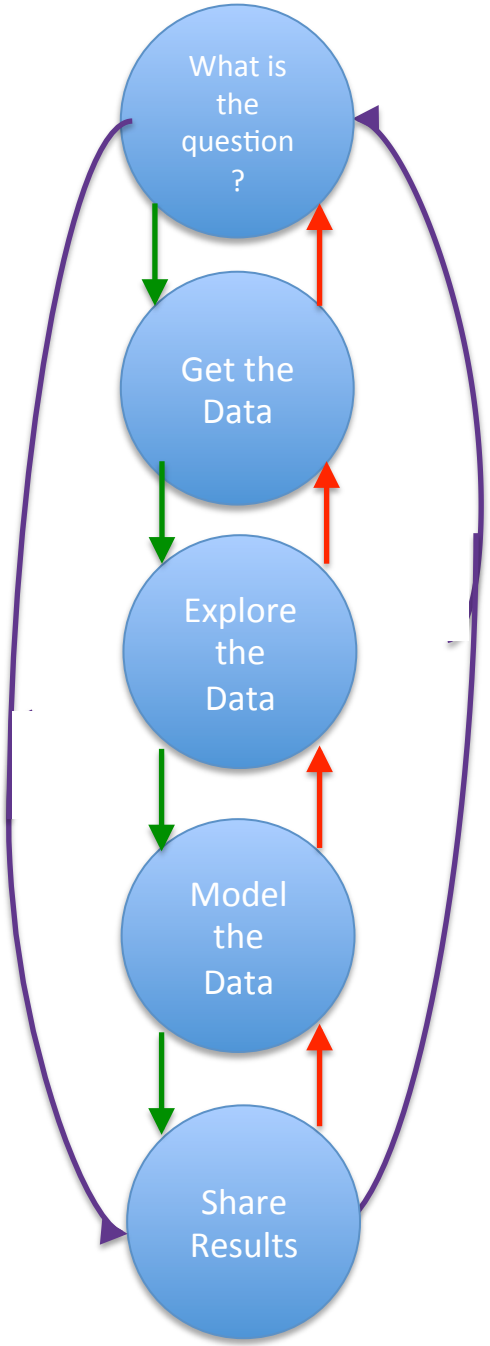
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Visualize the data
Are there **anomalies**?
Are there **patterns**?

Build a model
Fit the model
Validate the model

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?



Example: Predicting Neonatal Infection

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns and features in the data that predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

Example: Predicting Neonatal Infection

Problem: Processing disability claims at the Social Security Administration is a time-intensive process, with many claims taking over 2 years to adjudicate

Goal: Automate the approval of a subset of the "simplest" disability claims

Data: Free text in the claims form

Impact: Able to fully automate 20% of the simplest claims. Rating accuracy of the algorithm is higher than the average claims examiner.

Case Study: http://datamininglab.com/images/case-studies/ERI_Text_Mining_SSA_Claims_for_Disability_Approval.pdf

Jeff Hammerbacher's Model

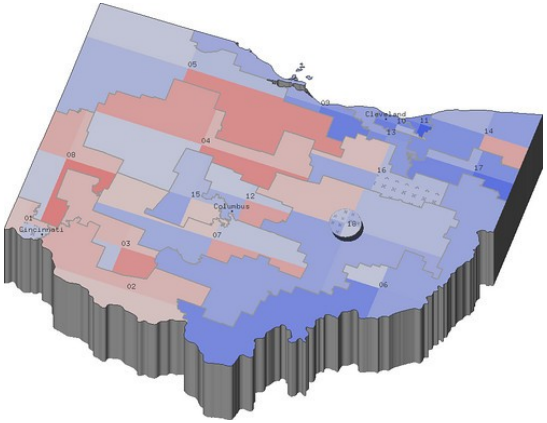
1. Identify problem + Hypothesis?

2. Instrument data sources

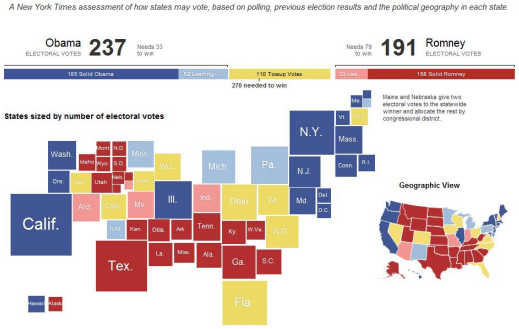
3. Collect data

4. Prepare data (integrate, transform, clean, filter, aggregate), Mulch, Dig around

5. Build model



6. Evaluate model



7. Communicate results

8. Scale?

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?)

About the Course

Grading

- In-class Quizzes (top 5) 30%
- Assignments (5) 30%
- Final Project 40%

Projects

Project teams should form and be reported by 2/23 (max 2).

Project proposals due by 3/15

Project presentations prior to due date to explain proposed approach and get feedback

You can choose a project topic, but we will also provide a list of suggested projects from around campus

You need:

- A clear problem statement (motivated by literature, personal contact, other resource)
- An accessible dataset
- Modeling plan + appropriate tools

Project Proposal

- What is the problem?
- How will you learn the background?
- What kinds of data will you use?
- Almost anything is OK, except other predictions.
- Numerical or text?
- What kind of model will you build?
- What assumptions are safe to make?
- Proposal should be ~1000 words + figures

About the Course

Staff Contact:

Instructor: Rumi Chunara, rumi.chunara@nyu.edu

Office hours: TH 9-10am (or by appointment).

TA:

Qi Wang: Fr 12-2 in 2 Metrotech, 10.054F
qiwang@nyu.edu

Use Piazza for questions…

Textbook and Assignment Submissions

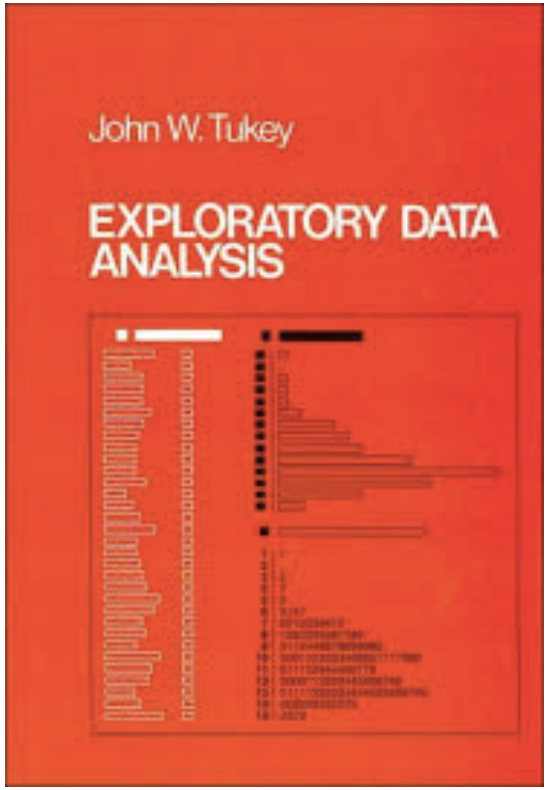
- Textbook: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* New York: Springer (available online for free)
- Other readings posted on Classes
- All assignments must be completed in R. Implement and comment your code so that anyone reading the file can reproduce the code easily (e.g. set the file path once at the beginning of the script where it can be easily changed).
- Save the code as an R markdown file, and upload it to NYU classes.

Getting to Know Data

- Techniques (we will cover)
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60's
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions



Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
  - We'll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
  - These are the techniques we'll leverage for Machine Learning and Prediction

Applying techniques

- Many questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
  - E.g. auto-encoders for image recognition neural nets.

Applying techniques

- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Latent Dirichlet Allocation

The Trouble with Summary Stats

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

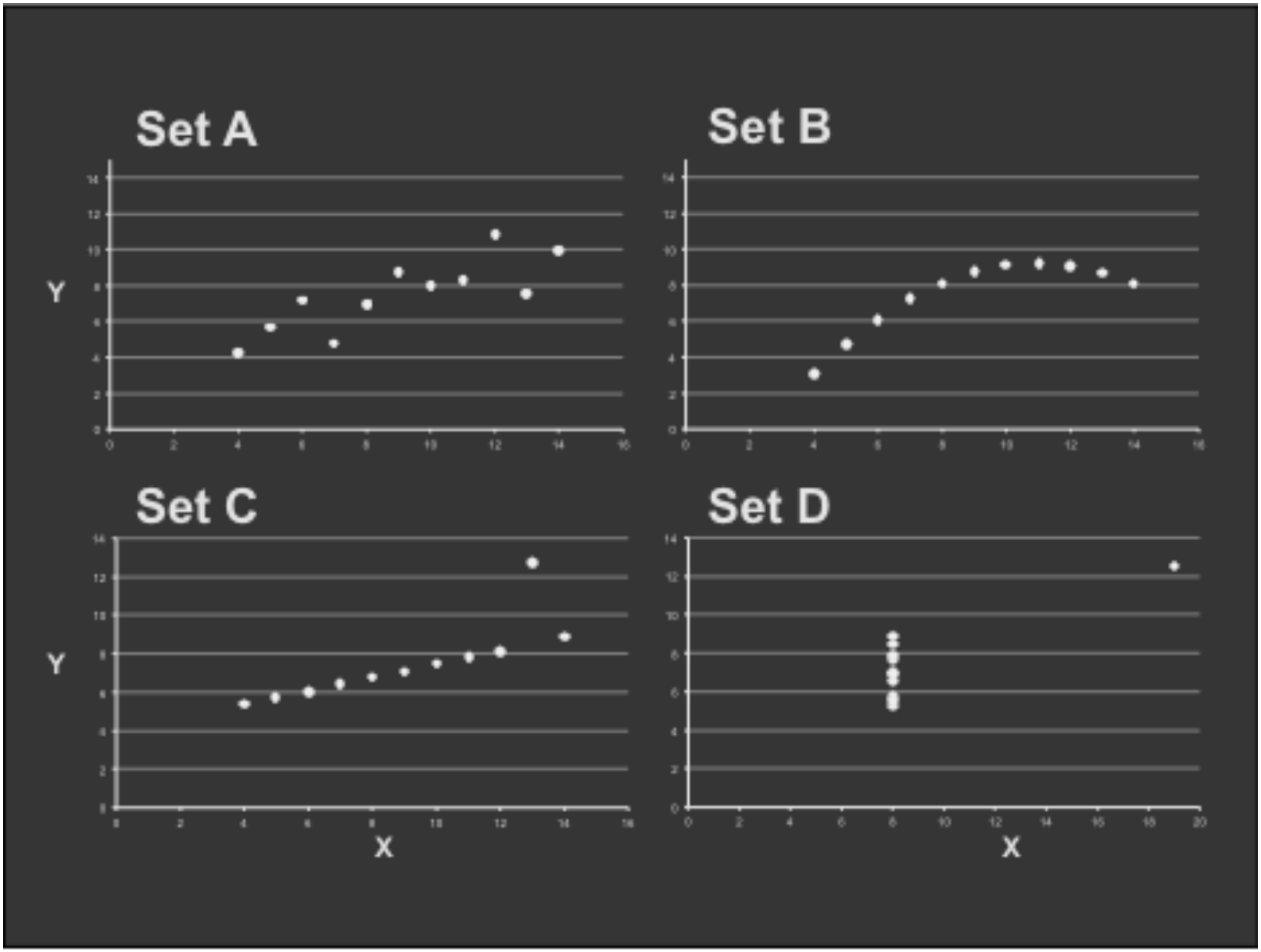
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

Looking at Data



The “R” Language

- An evolution of the “S” language developed at Bell labs for EDA.
- Idea was to allow interactive exploration and visualization of data.
- The preferred language for statisticians, used by many other data scientists.
- Features:
  - Probably the most comprehensive collection of statistical models and distributions.
  - CRAN: a very large resource of open source statistical models.

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

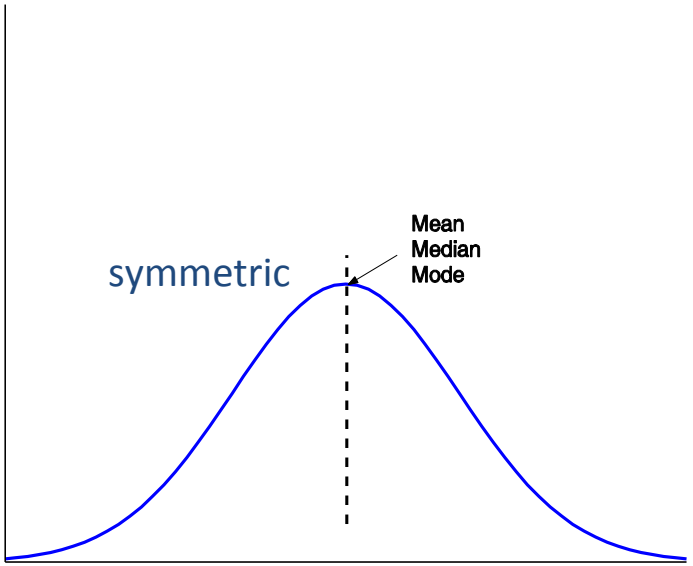
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

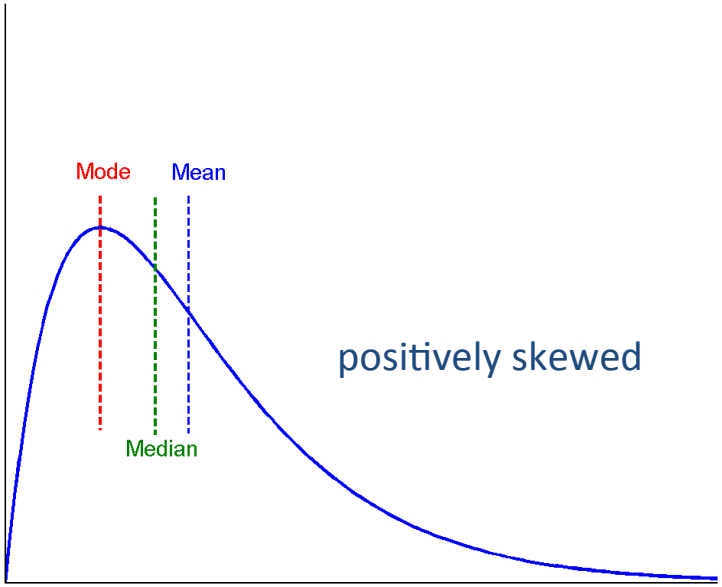
age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

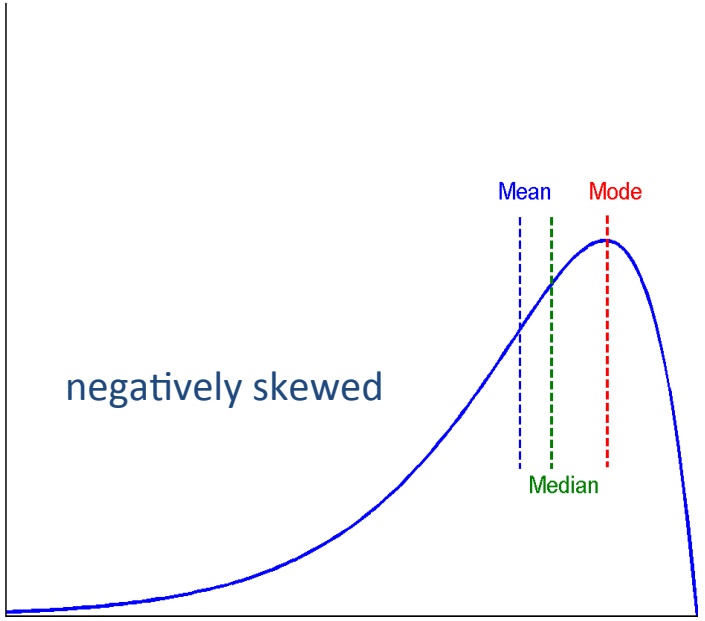
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







J

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

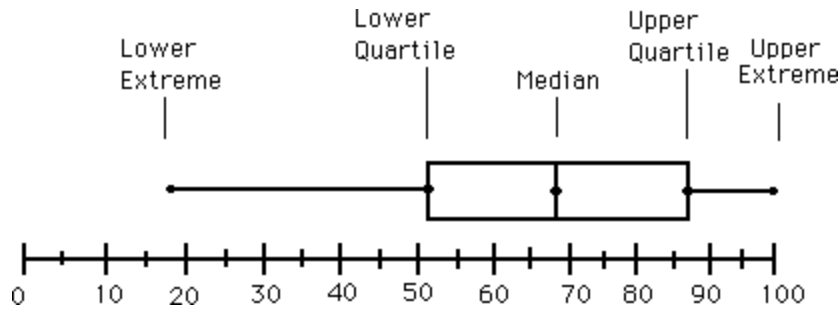
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance *s^2 (or σ^2)*

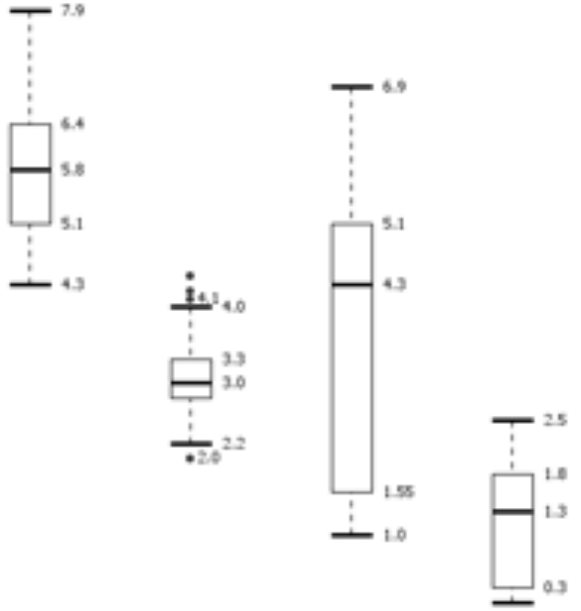
Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis arefrequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Boxplot Analysis

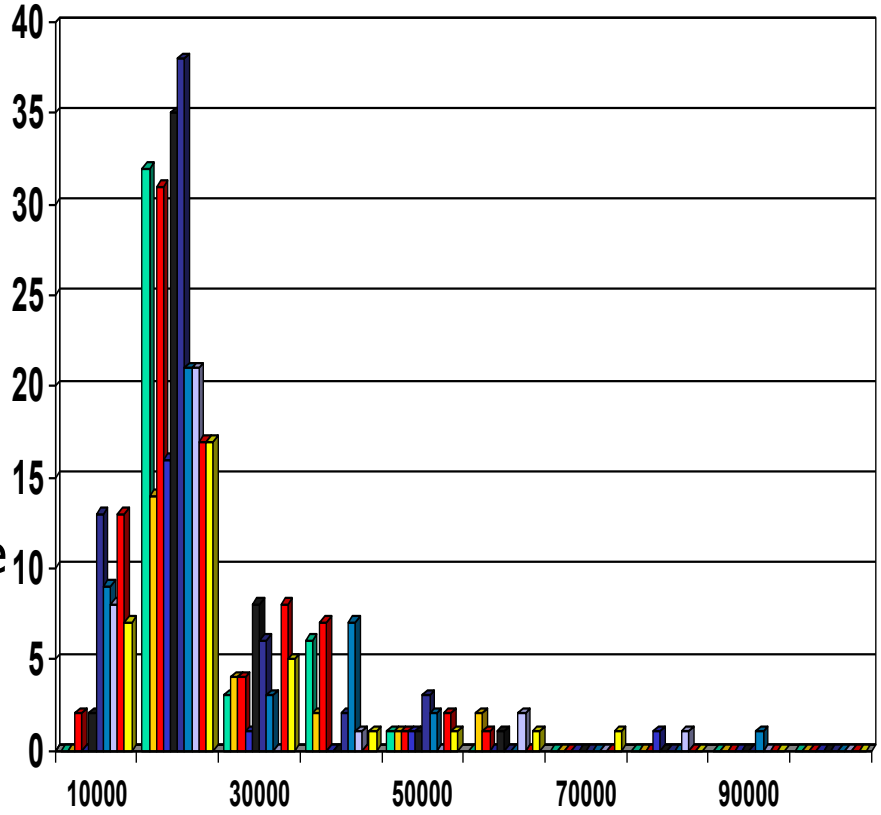


- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

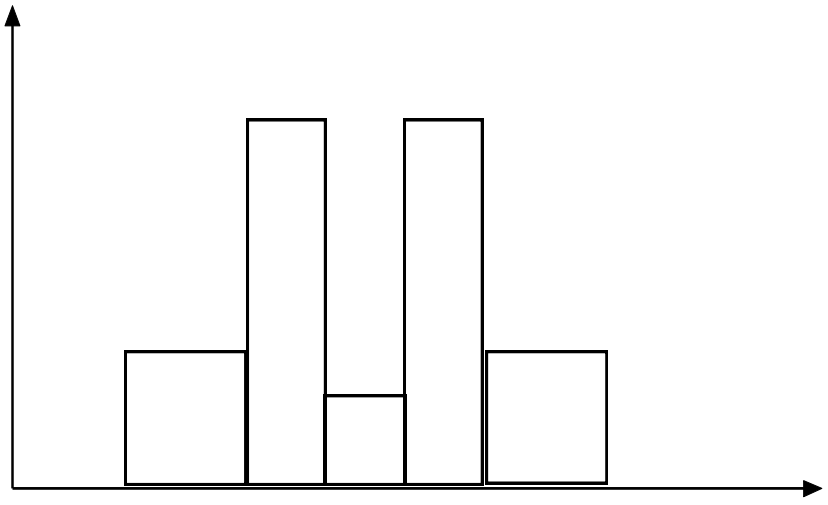


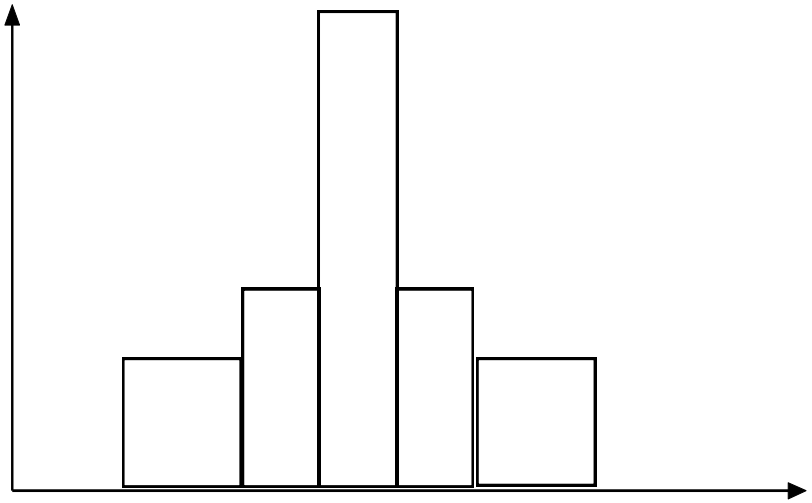
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

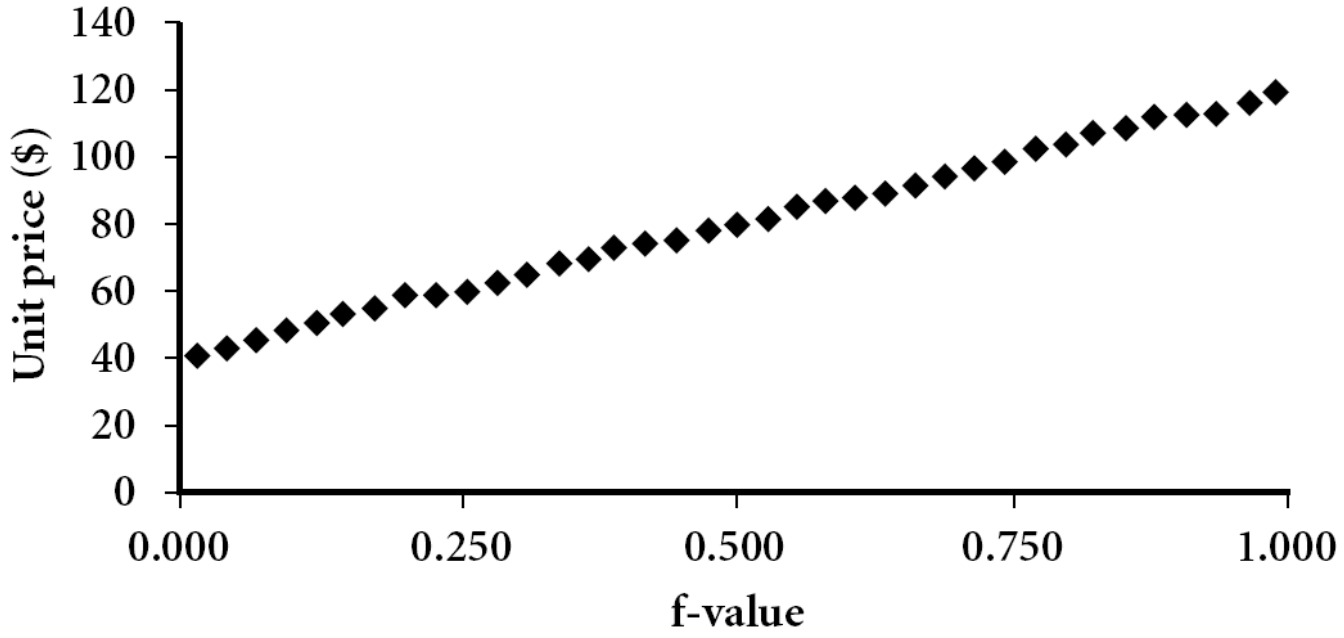




- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

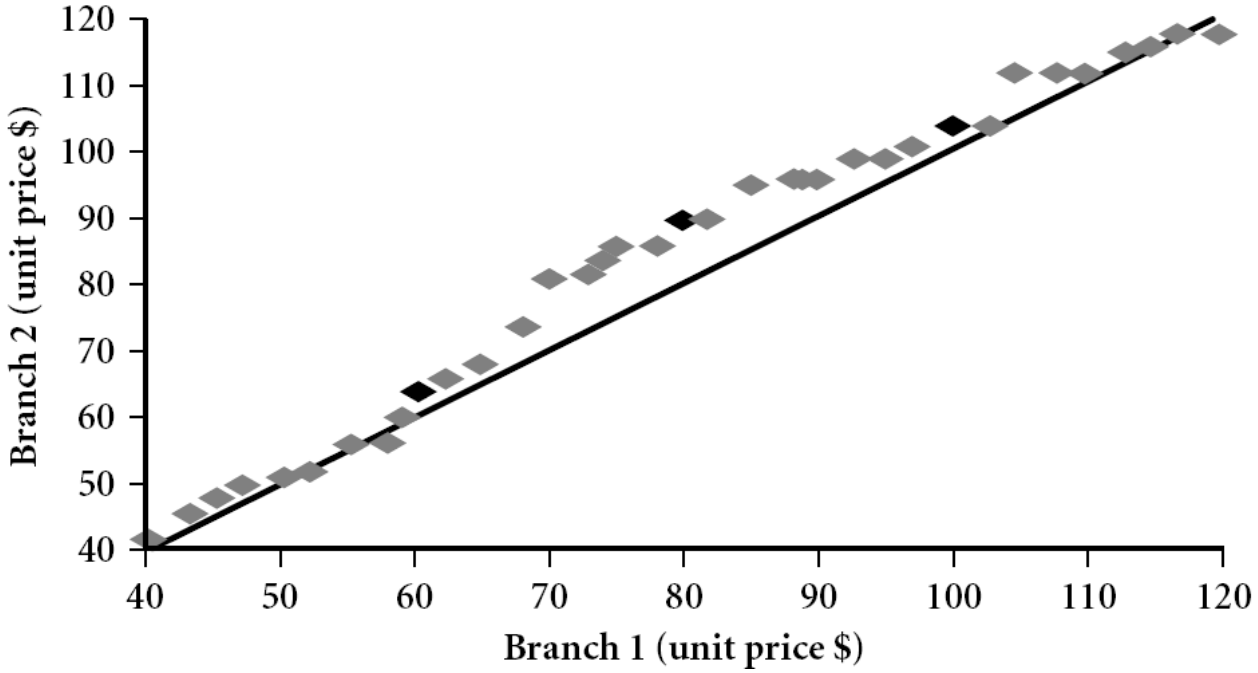
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



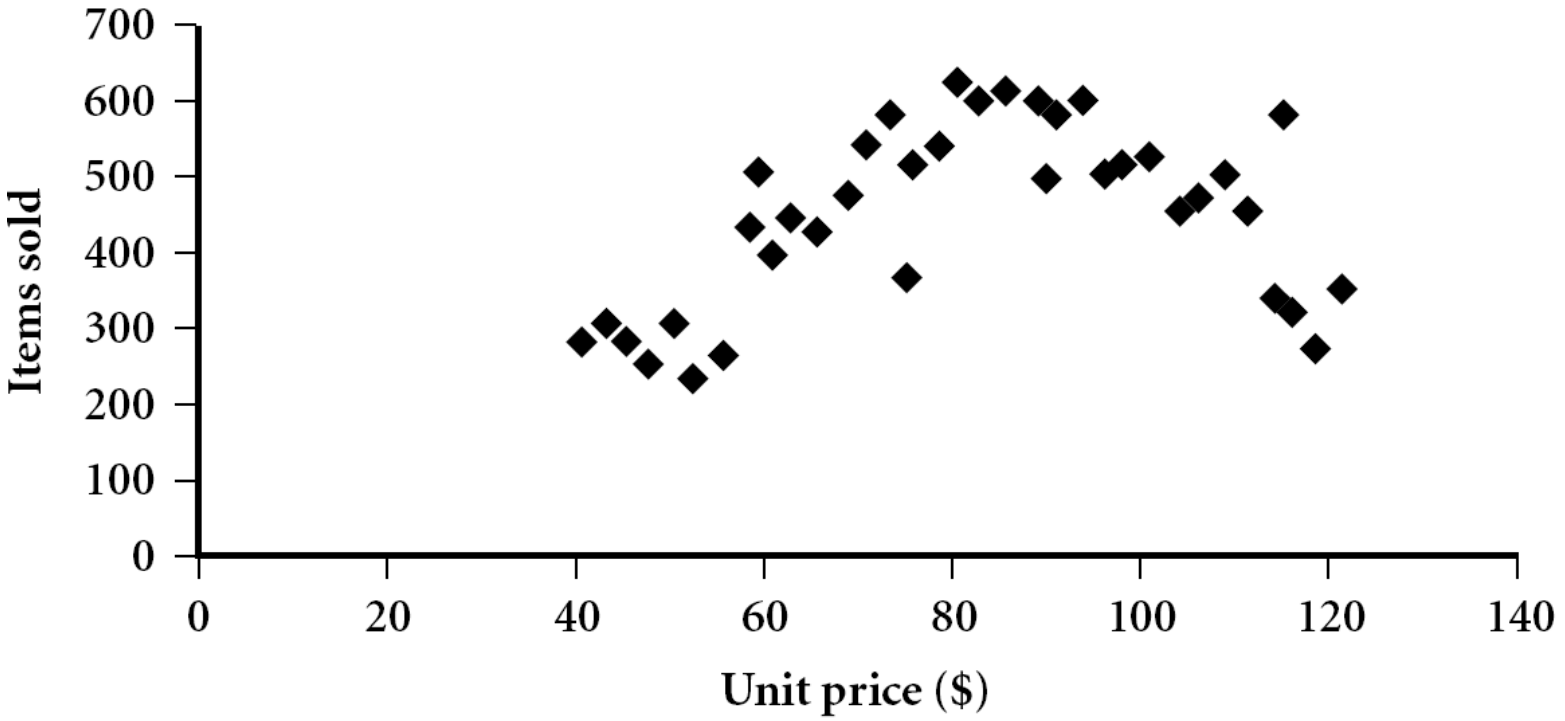
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

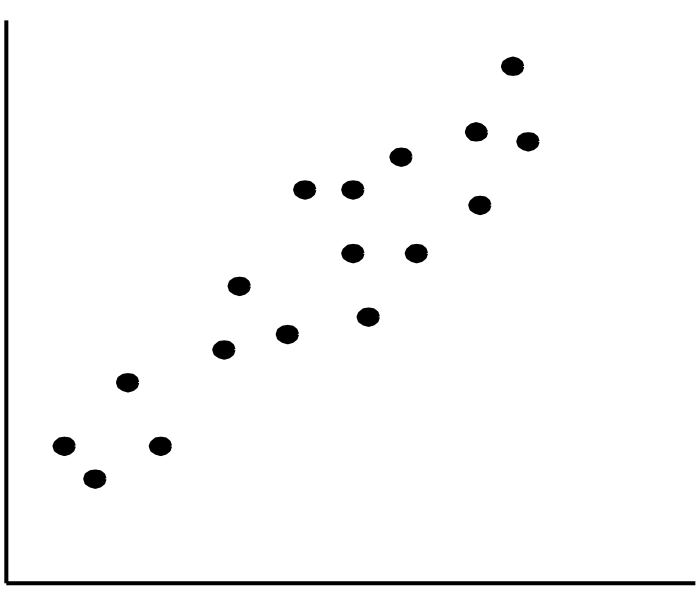


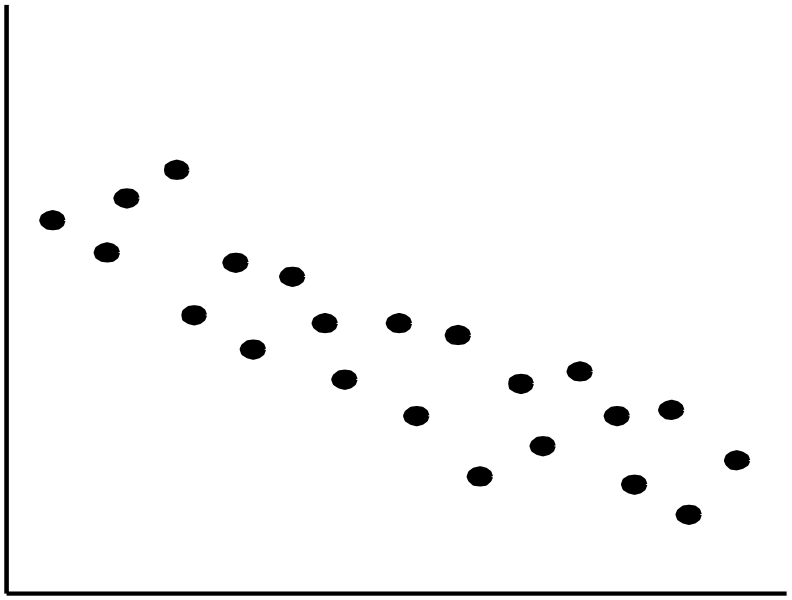
Scatter plot

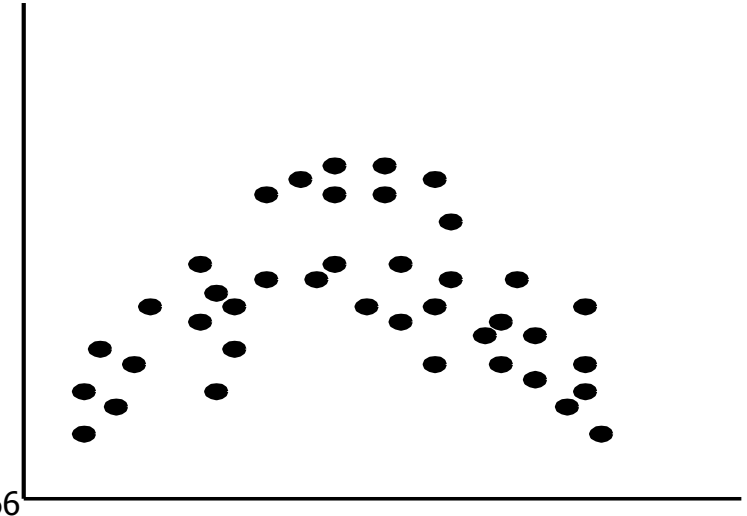
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

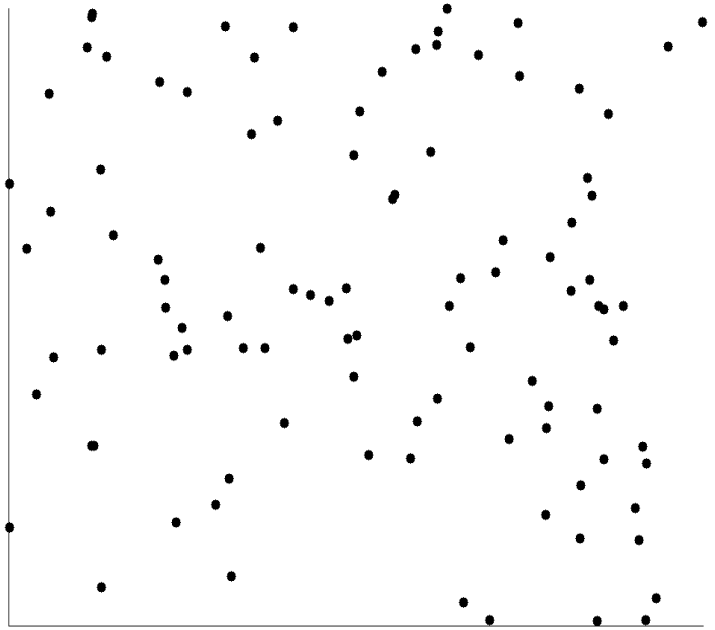


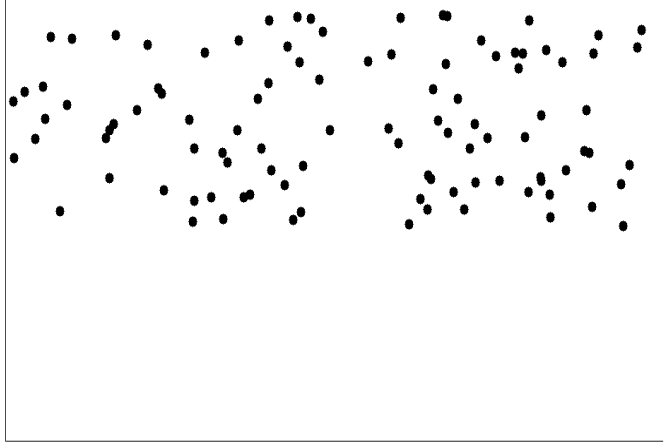


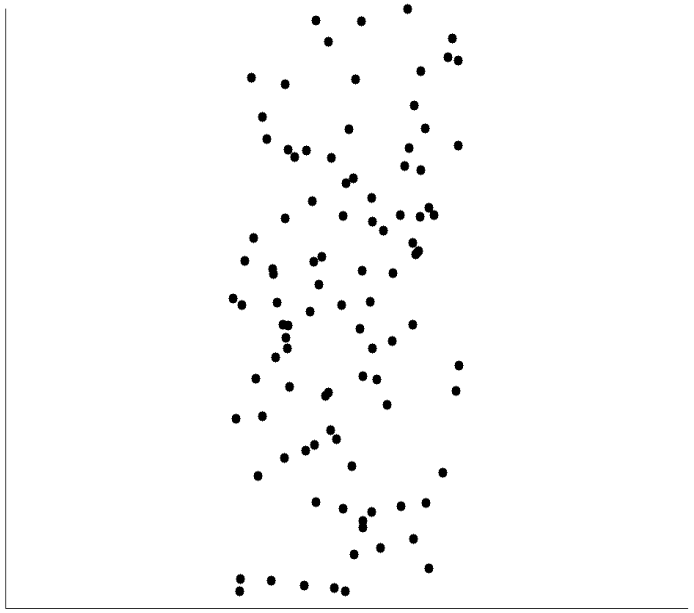


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

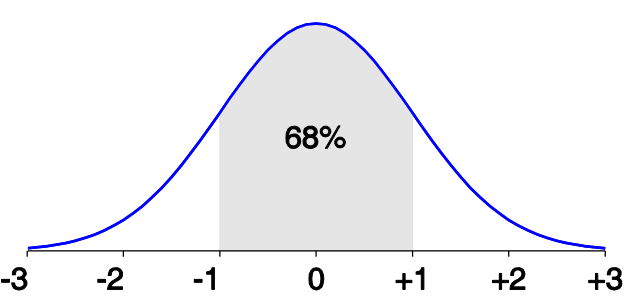


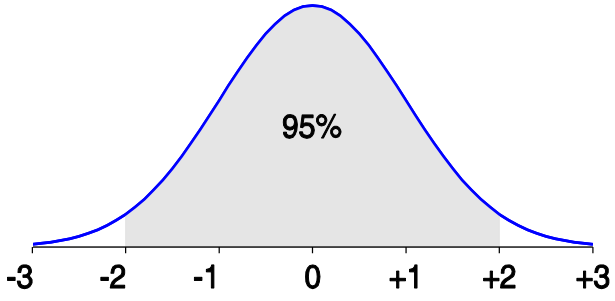


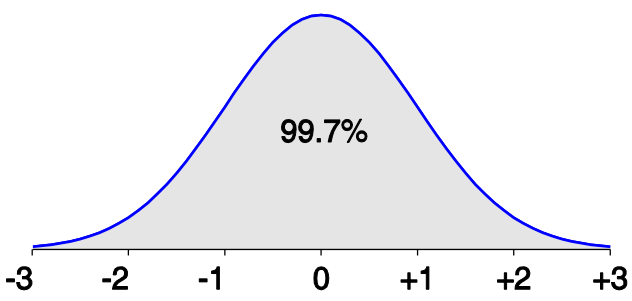


Properties of Normal Distribution Curve

- The normal (distribution) curve
  - mean = median = mode
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it





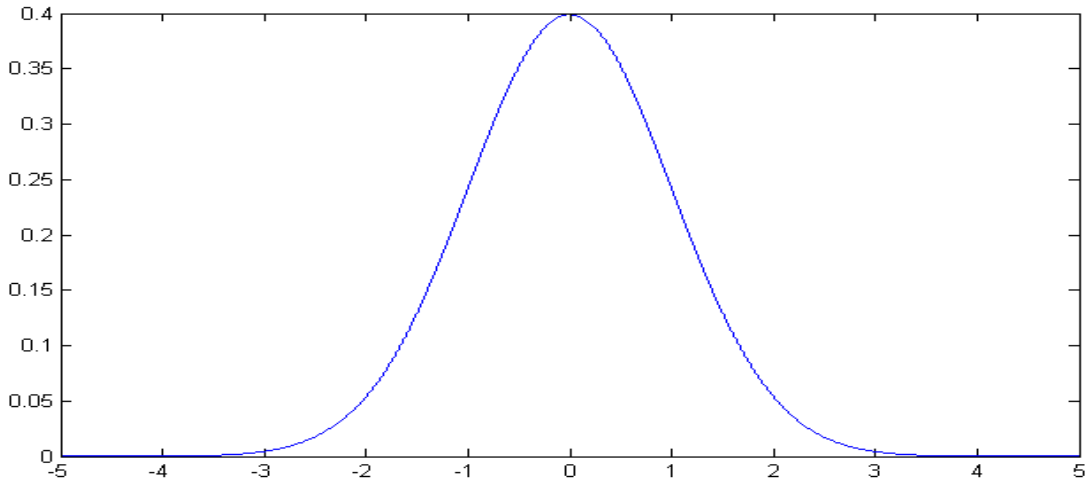


Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

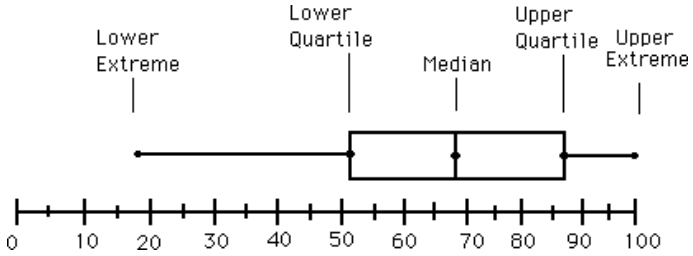
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



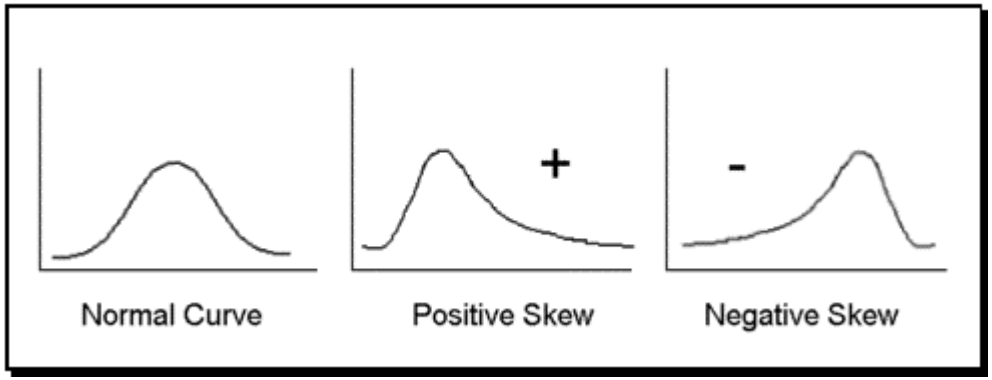
Correcting distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

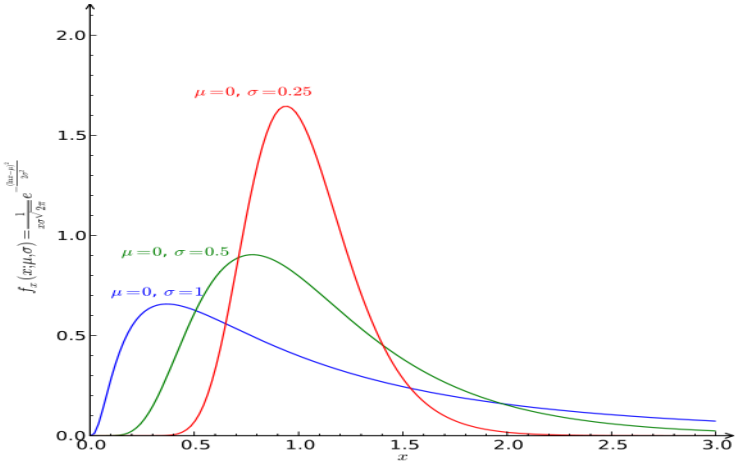


Correcting distributions

In many cases these distribution can be corrected before any other processing.

Examples:

- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X poisson with mean k and standard deviation: sqrt(k). Then sqrt(X) is approximately normally distributed with standard deviation = 1

Distributions

Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain "rate".
  - Observed frequency of a given term in a corpus.
  - Number of visits to a web site in a fixed time interval.
  - Number of web site clicks in an hour.
- **Exponential:** the interval between two such events.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or web site visits.
- **Binomial/Multinomial:** The number of counts of events (e.g. die tosses = 6) out of n trials.

- You should understand the distribution of your data before applying any model.

Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards.

Q: what's wrong with his conclusion?

* Example from Jeff Ullman/Anand Rajaraman

Rhine Paradox

He called back the “psychic” subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

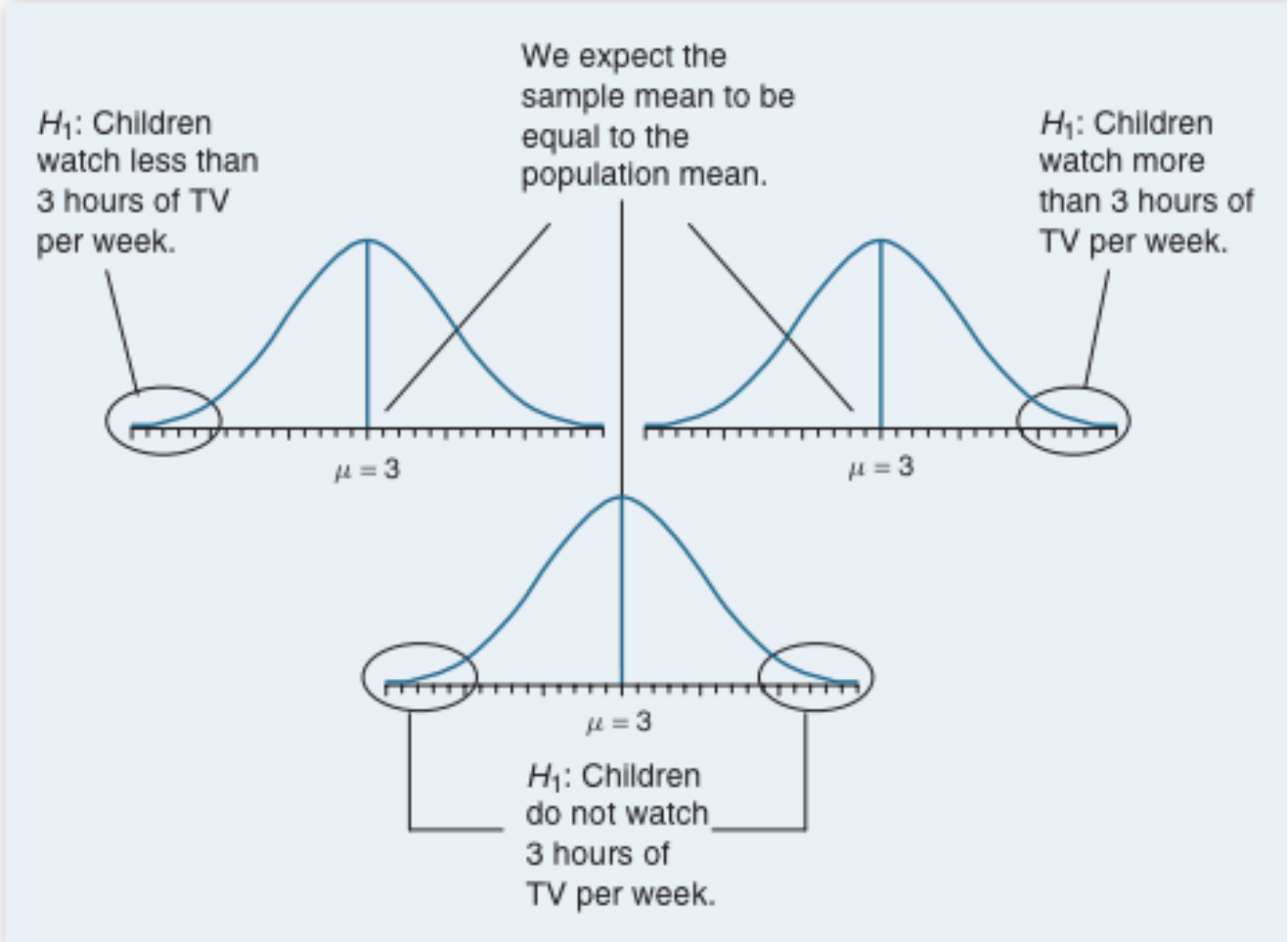
Hypothesis Testing

- We want to prove a hypothesis H_A, but its hard so we try to **disprove a null hypothesis H_0**.
- A **test statistic** is some measurement we can make on the data which is likely to be **big under H_A** but **small under H_0**.
- We chose a test statistic whose distribution we know if H_0 is true: e.g.
  - Two samples a and b, normally distributed, from A and B.
  - H_0 hypothesis that mean(A) = mean(B), test statistic is: s = mean(a) – mean(b).
  - s has mean zero and is normally distributed under H_0.
  - But its “large” if the two means are different.

Hypothesis Testing – contd.

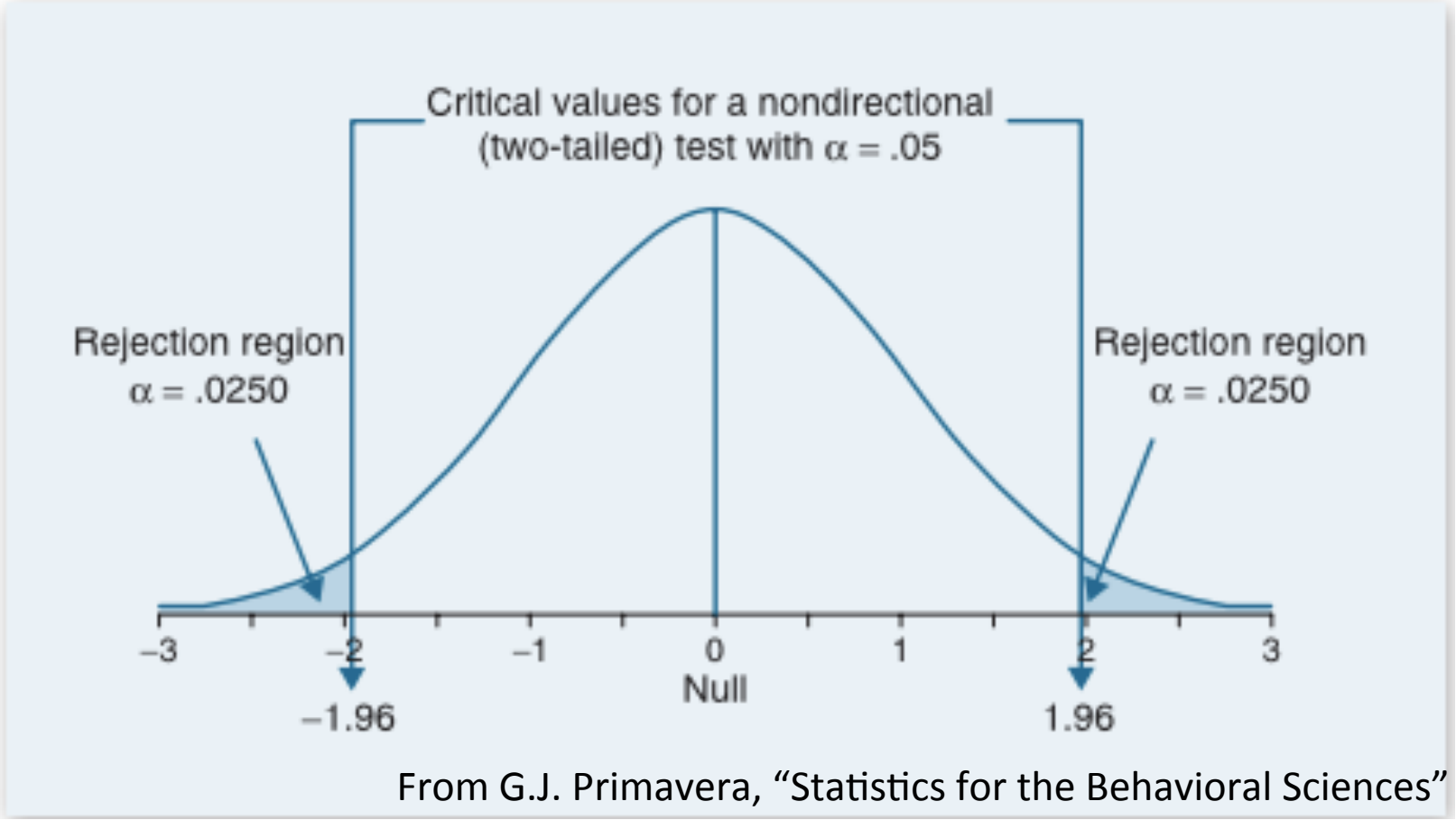
- s = mean(a) – mean(b) is our test statistic, H_0 the hypothesis that mean(A)=mean(B)
  - We reject if $Pr(x > s \mid H_0) < p$
  - p is a suitable "small" probability, say 0.05.

- This threshold probability is called a p-value.
  - P directly controls the false positive rate (rate at which we expect to observe large s even if is H_0 true).
  - As we make p smaller, the false negative rate increase – situations where mean(A), mean(B) differ but the test fails.
  - Common values 0.05, 0.02, 0.01, 0.005, 0.001



From G.J. Primavera, “Statistics for the Behavioral Sciences”

Two-tailed Significance



From G.J. Primavera, "Statistics for the Behavioral Sciences"

When the p value is less than 5% ($p < .05$), we reject the null hypothesis

Hypothesis Testing

		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

From G.J. Primavera, "Statistics for the Behavioral Sciences"

Three important tests (Parametric)

- **T-test:** compare two groups, or two interventions on one group

- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table" (continuous data)

- **ANOVA:** compare outcomes under several discrete interventions (multiple groups)

Test Statistics

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

T-test:

within-subjects (one group of individuals, two categories)

$$t = X / \sigma$$

Between-subjects design (two groups)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Chi-sq: (is an observation consistent with the data) O_i is an observed count, and E_i is the expected value of that count. It has a chi-squared distribution, whose p-values you compute to do the test.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Parametric Tests

All the tests so far are parametric tests that assume the data are **normally distributed**, and that the samples are **independent of each other and all have the same distribution** (IID).

They may be arbitrarily inaccurate is those assumptions are not met. Always make sure your data satisfies the assumptions of the test you're using. e.g. watch out for:

- Outliers – will corrupt many tests that use variance estimates.
- Correlated values as samples, e.g. if you repeated measurements on the same subject.
- Skewed distributions – give invalid results.

Non-parametric tests

These tests make no assumptions about the distribution of the input data, and can be used on very general datasets:

- K-S test
- Permutation tests
- Bootstrap confidence intervals

K-S test

The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

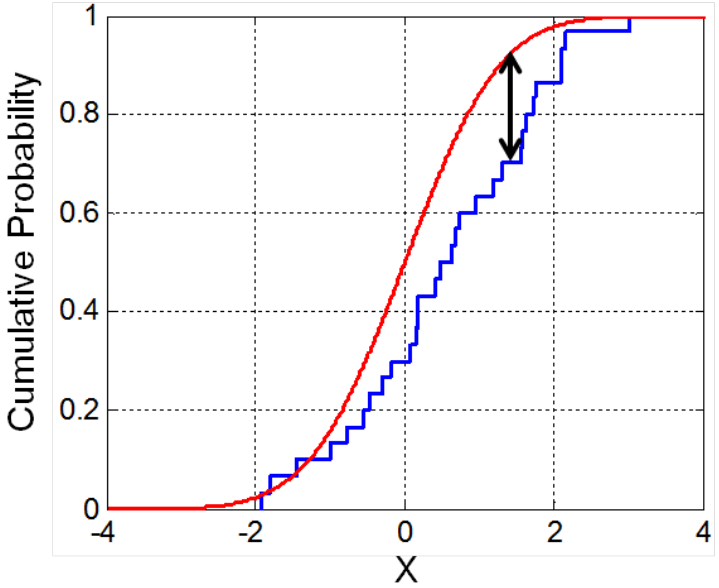
In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution.

In the **two-sided test**, two observed distributions are compared.

The K-S statistic is just the **max distance between the CDFs** of the two distributions.

While the statistic is simple, its distribution is not!

But it is available in most stat packages.



K-S test

The K-S test can be used to test **whether a data sample has a normal distribution** or not.

Thus it can be used as a sanity check for any common parametric test (which assumes normally-distributed data).

It can also be used to compare distributions of data values in a large data pipeline: **Most errors will distort the distribution of a data parameter and a K-S test can detect this**.

Non-parametric tests

Permutation tests

Bootstrap confidence intervals

- We wont discuss these in detail, but its important to know that non-parametric tests using one of the above methods exist for many forms of hypothesis.

- They make no assumptions about the distribution of the data, but in many cases are just as sensitive as parametric tests.

- They use computational cycles to simulate sample data, to derive p-value estimates approximately, and accuracy improves with the amount of computational work done.

Today, Recap

- What is Data Science?
- Data Handling
- Doing Data Science
- About the course
- Statistics Review
- Polling YOU

Next Time

- Intro to R: Basic Commands Graphics Indexing Loading Data Additional Graphical and Numerical Summaries
- Getting Data + APIs
- Data cleaning, sampling, processing
- Assignment 1 out