# Exploratory Data Analysis
# Part 1

# Today

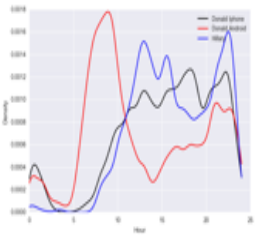
# Congratulations!


You have **collected** or **been given** a box of data?
What do you do next?

(B)Question & Problem Formulation

?

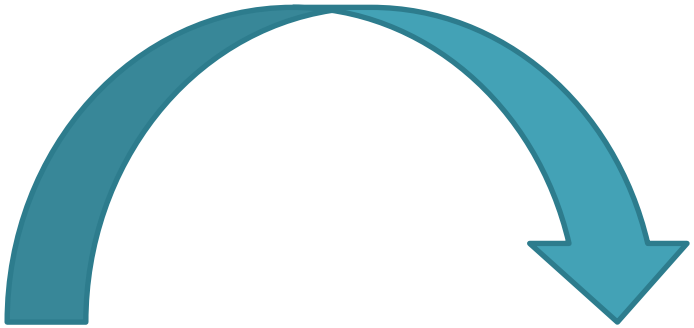
(IA) Data Acquisition



(EVA) Exploratory Data Analysis

(AVU) Prediction and Inference

# EVA



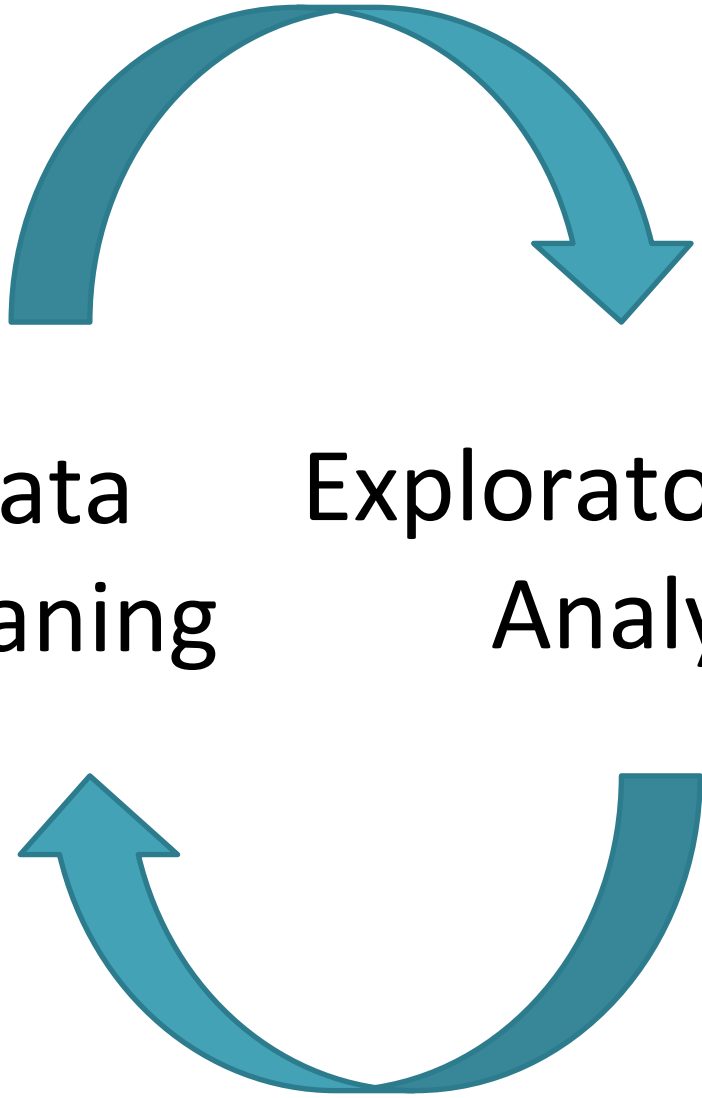
Data Cleaning

Exploratory Data Analysis

the infinite loop of data scienc

# Data Cleaning

- The process of transforming raw data to facilitate subsequent analysis
- Data cleaning often addresses
  - structure / formatting
  - missing or corrupted values
  - unit conversion
  - encoding text as numbers
  - …
- Sadly data cleaning is a big part of data science…






the infinite loop of data scienc

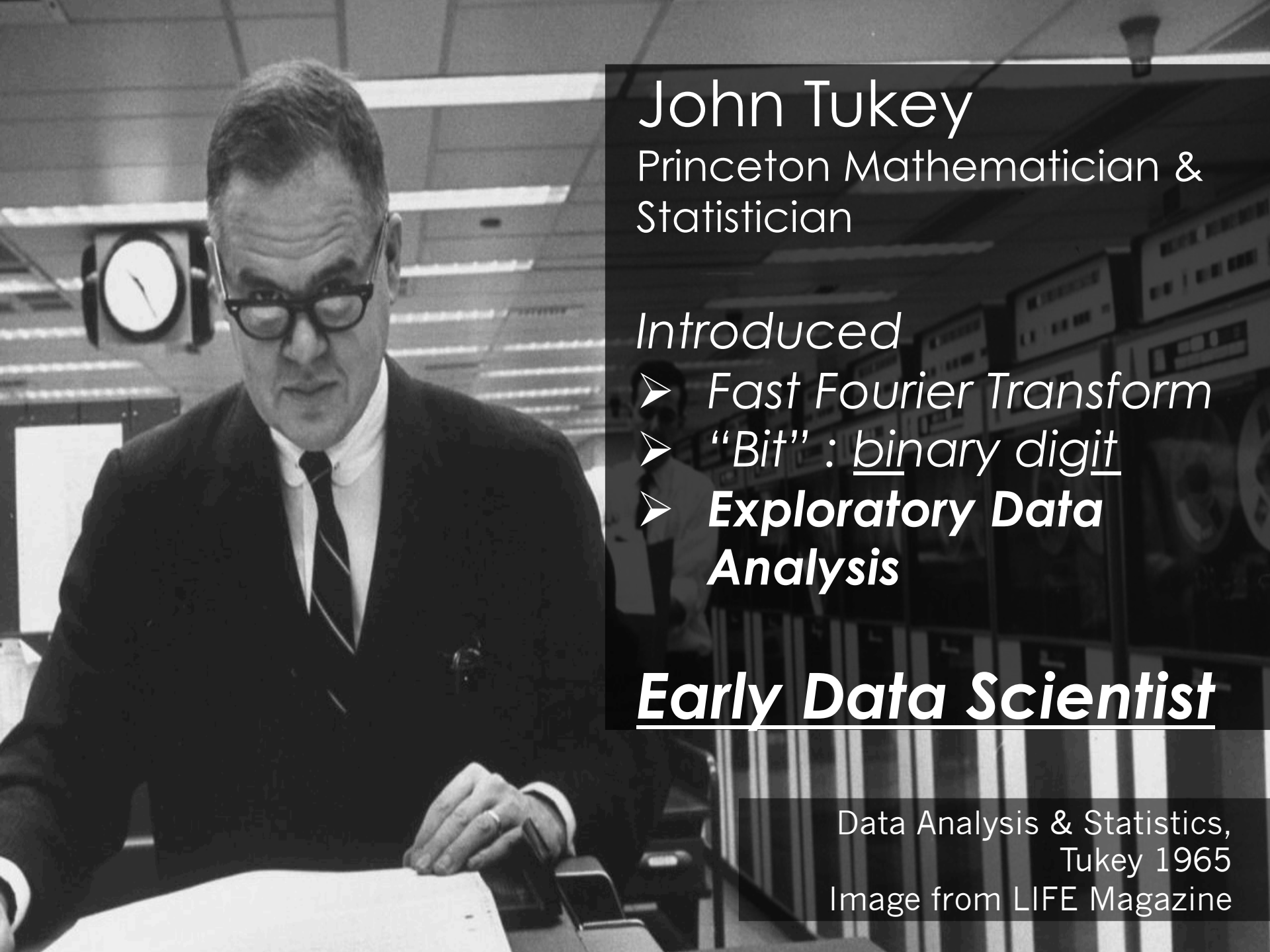
# Exploratory Data Analysis

*"Getting to know the data"*

The process of **transforming**, **visualizing**, and **summarizing** data to:

- Build/confirm understanding of the data and its provenance
- Identify and address potential issues in the data
- Inform the subsequent analysis
- discover *potential* hypothesis … (be careful)

- **EDA is an open ended analysis**
  - Be willing to find something surprising



# John Tukey

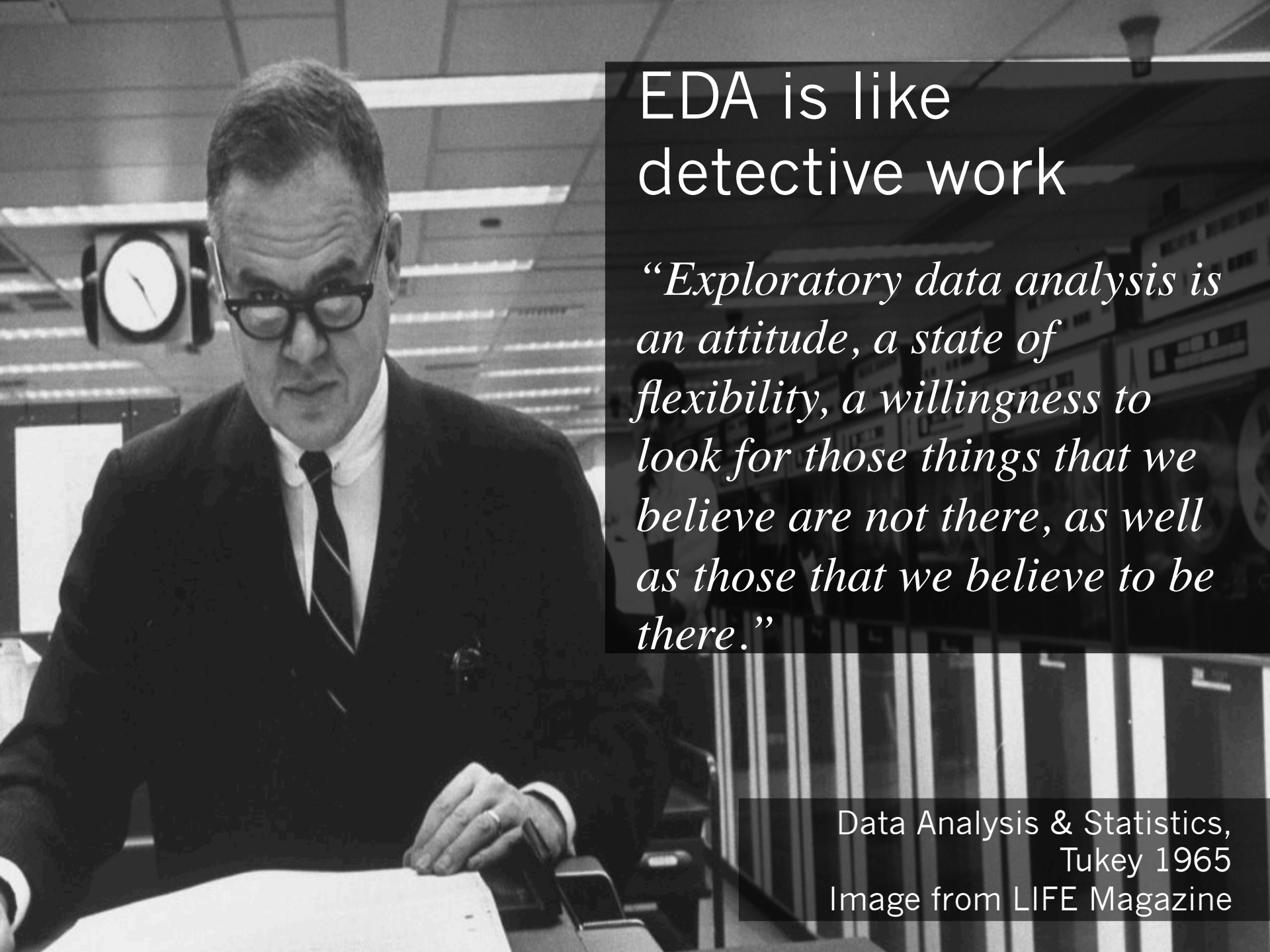
Princeton Mathematician & Statistician

*Introduced*

- *Fast Fourier Transform*
- *"Bit" : binary digit*
- ***Exploratory Data Analysis***

## ***Early Data Scientist***

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine



# EDA is like detective work

*"Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there."*

Data Analysis & Statistics,
Tukey 1965
Image from LIFE Magazine

# WHAT SHOULD WE LOOK FOR?

# Outline

- Exploratory Data Analysis
  - Chart types
  - Some important distributions
  - Important Statistical Tests

# 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

# Examples of Business Questions

- **Simple (descriptive) Stats**
  - "Who are the most profitable customers?"
- **Hypothesis Testing**
  - "Is there a difference in value to the company of these customers?"
- **Segmentation/Classification**
  - What are the common characteristics of these customers?
- **Prediction**
  - Will this new customer become a profitable customer? If so, how profitable?

adapted from Provost and Fawcett, "Data Science for Business"

# Applying techniques

- Most business 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

# 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

# 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 |

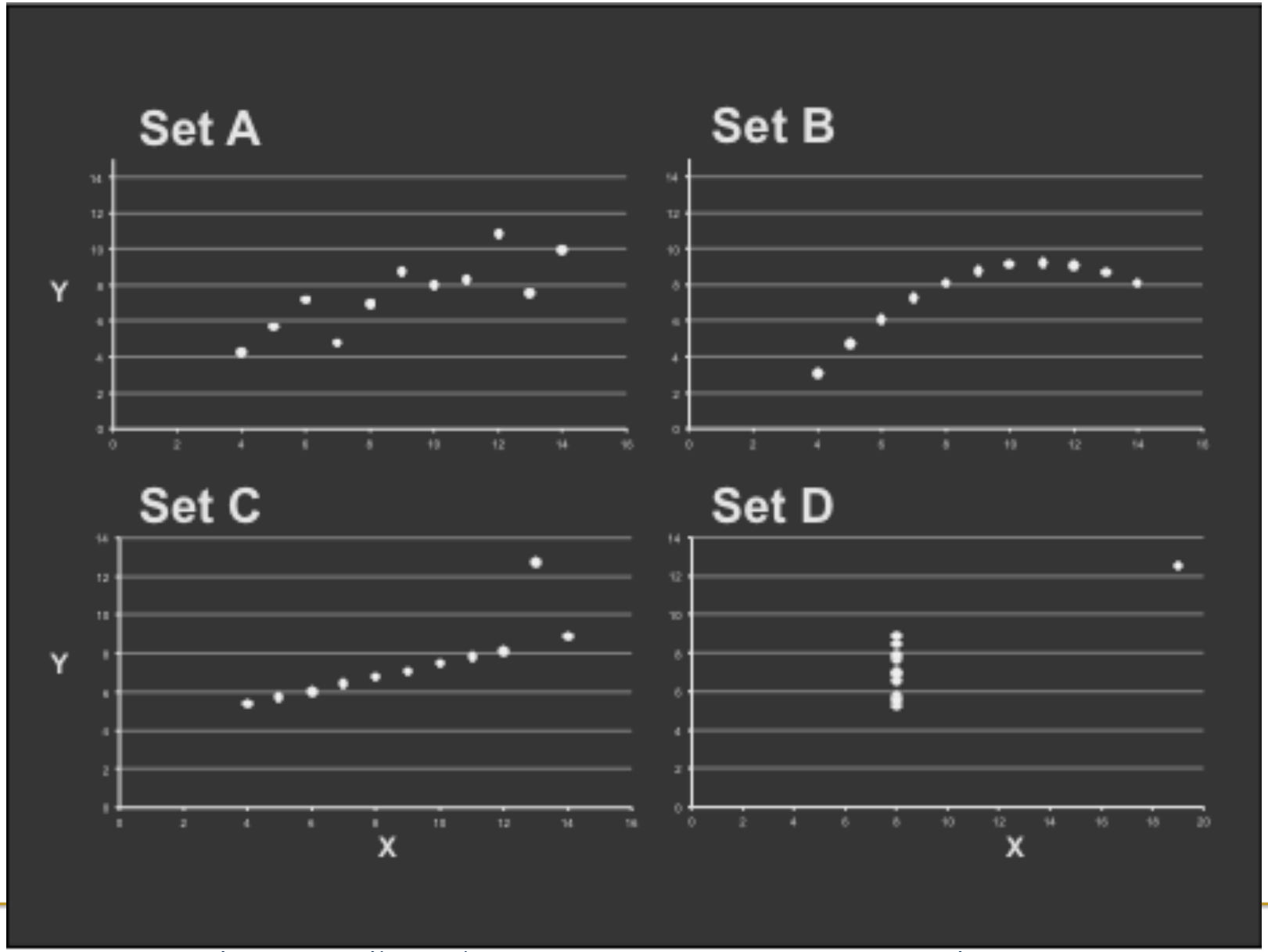
**Summary Statistics Linear Regression**

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

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

**[Anscombe 73]**

# Looking at Data



# Data Presentation

- Data Art

# 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.

Chart examples from Jeff Hammerbacher’s CS194 class

# Chart types

- Single variable
  - Dot plot
  - Jitter plot
  - Error bar plot
  - Box-and-whisker plot
  - Histogram
  - Kernel density estimate
  - Cumulative distribution function

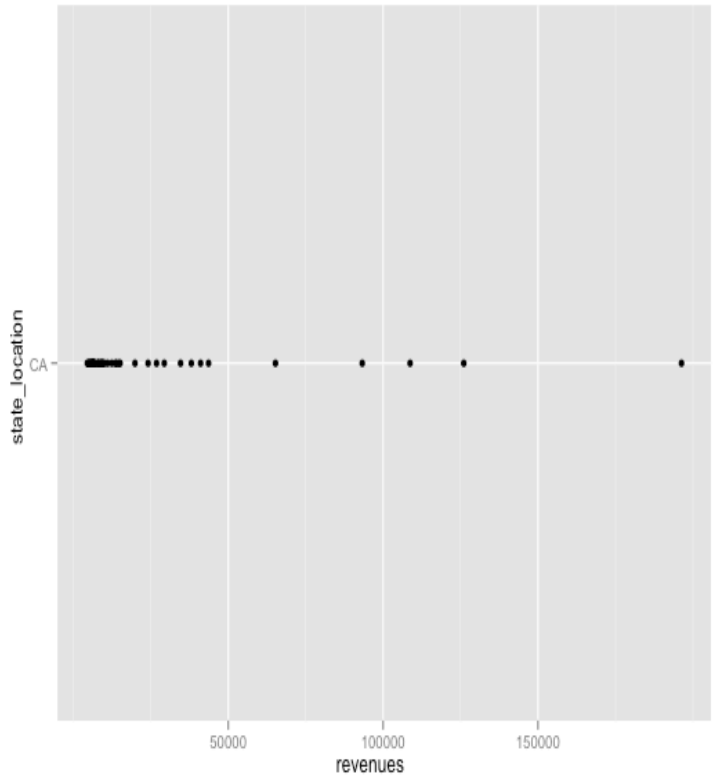
(note: examples using qplot library from R)

Chart examples from Jeff Hammerbacher's 2012 CS194 class

# Chart types

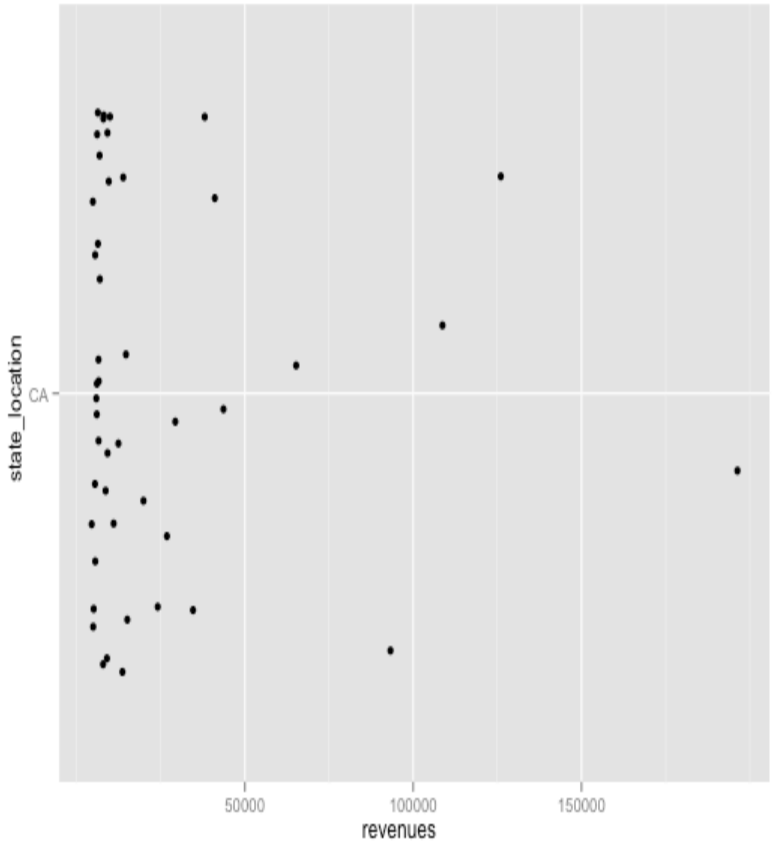
- **Dot plot**

```
> f500.ca <- subset(f500, state_location == "CA")
> f500.ca$state_location <- factor(f500.ca$state_location)
> qplot(revenues, state_location, data=f500.ca)
```



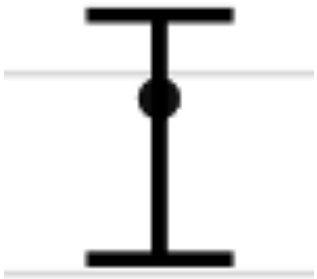
# Chart types

- **Jitter plot** `> qplot(revenues, state_location, data=f500.ca, geom="jitter")`
- Noise added to the y-axis to spread the points



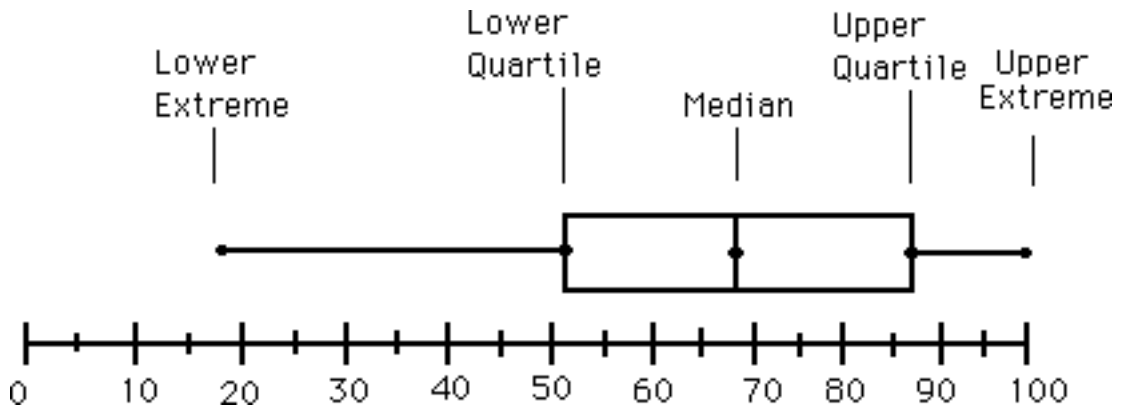
# Chart types

- **Error bars:** usually based on confidence intervals (CI). 95% CI means 95% of points are in the range, so 2.5% of points are above or below the bar.
- Not necessarily symmetric:



# Chart types

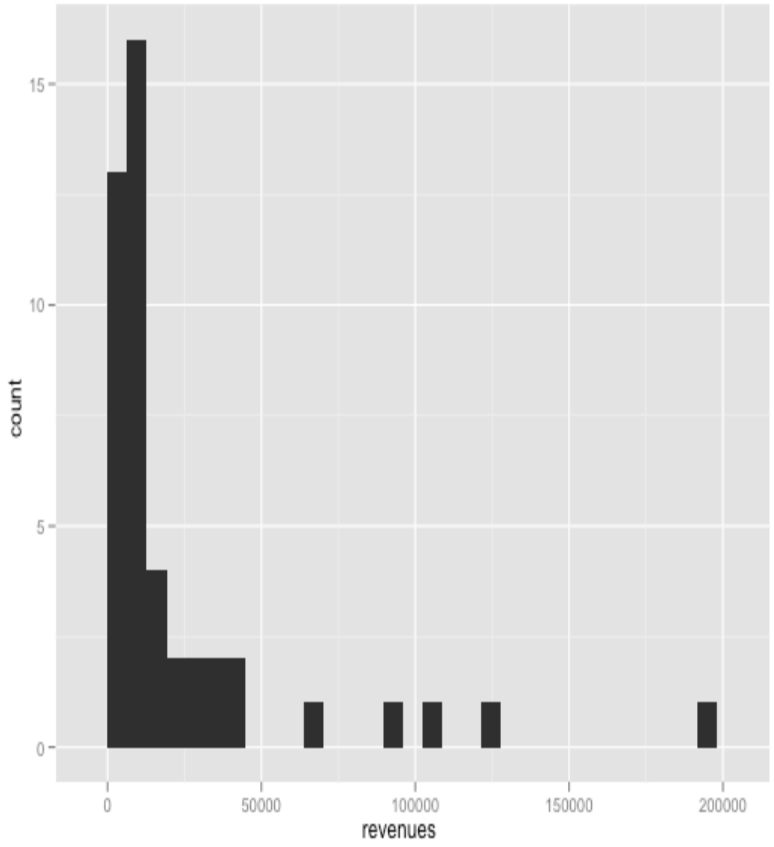
- **Box-and-whisker plot** : a graphical form of 5-number summary (Tukey)



# Chart types

- **Histogram**

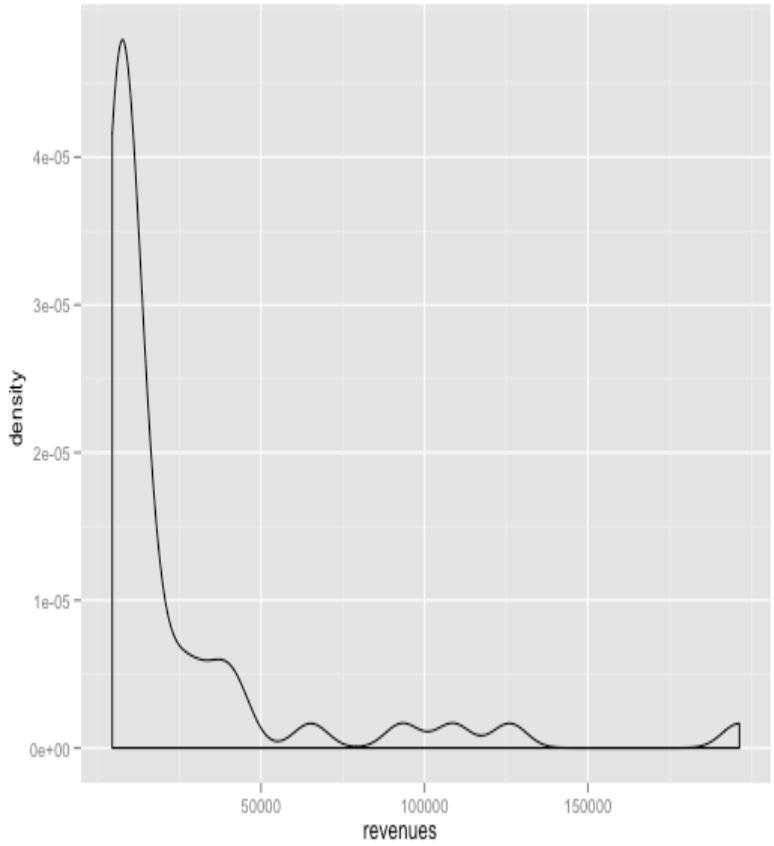
```
> qplot(revenues, data=f500.ca, geom="histogram")
stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



# Chart types

- **Kernel density estimate**

```
> qplot(revenues, data=f500.ca, geom="density")
```



# Chart types

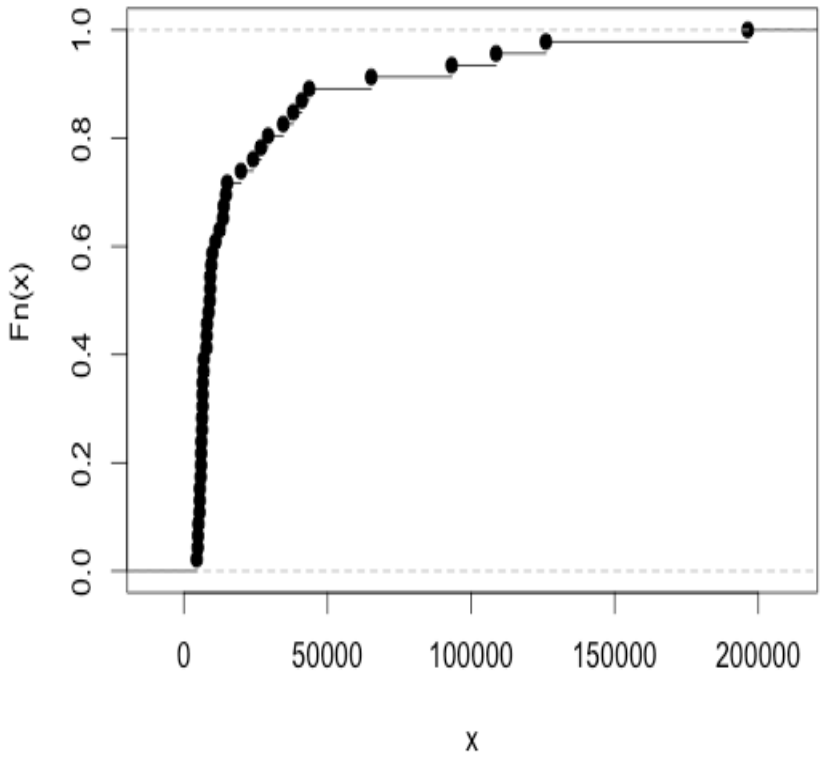
- Histogram and Kernel Density Estimates
  - Histogram
    - Proper selection of bin width is important
    - Outliers should be discarded
  - KDE (like a smooth histogram)
    - Kernel function
      - Box, Epanechnikov, Gaussian
    - Kernel bandwidth

# Chart types

- **Cumulative distribution function**

```
> plot(ecdf(f500.ca$revenues))
```

- Integral of the histogram – simpler to build than KDE (don't need smoothing)

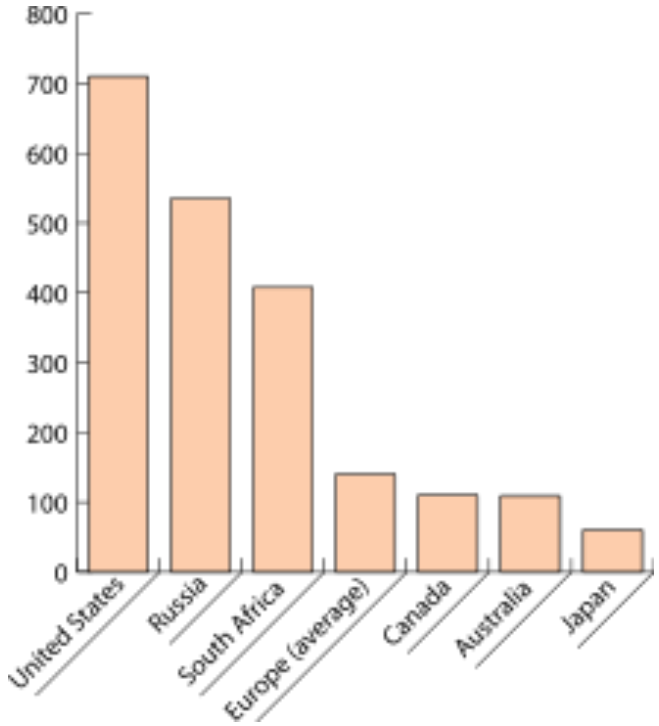


# Chart types

- Two variables
  - Bar chart
  - Scatter plot
  - Line plot
  - Log-log plot

# Chart types

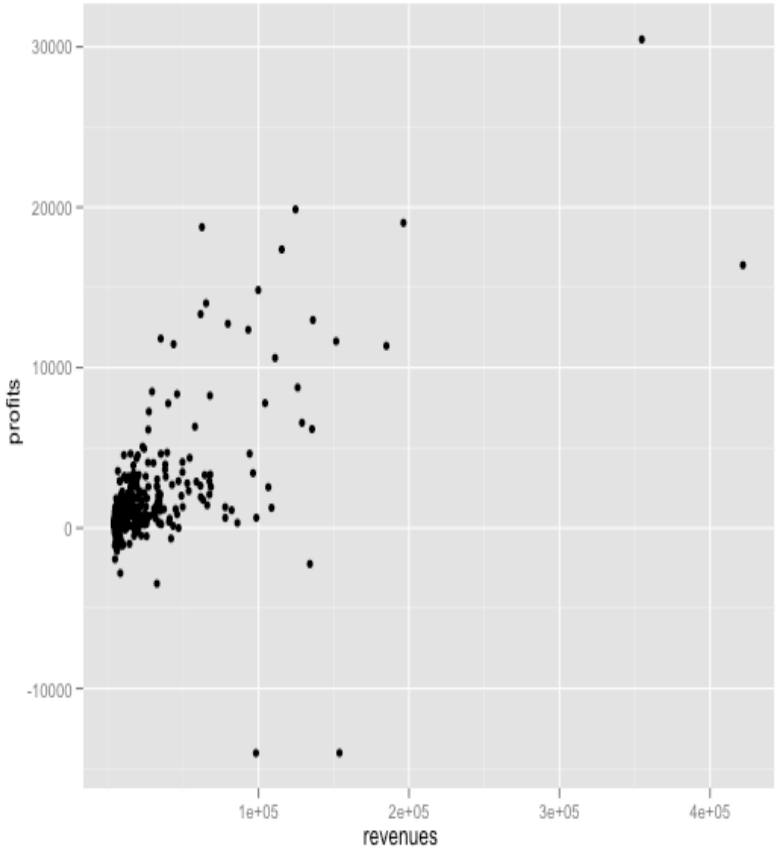
- **Bar plot:** one variable is discrete



# Chart types

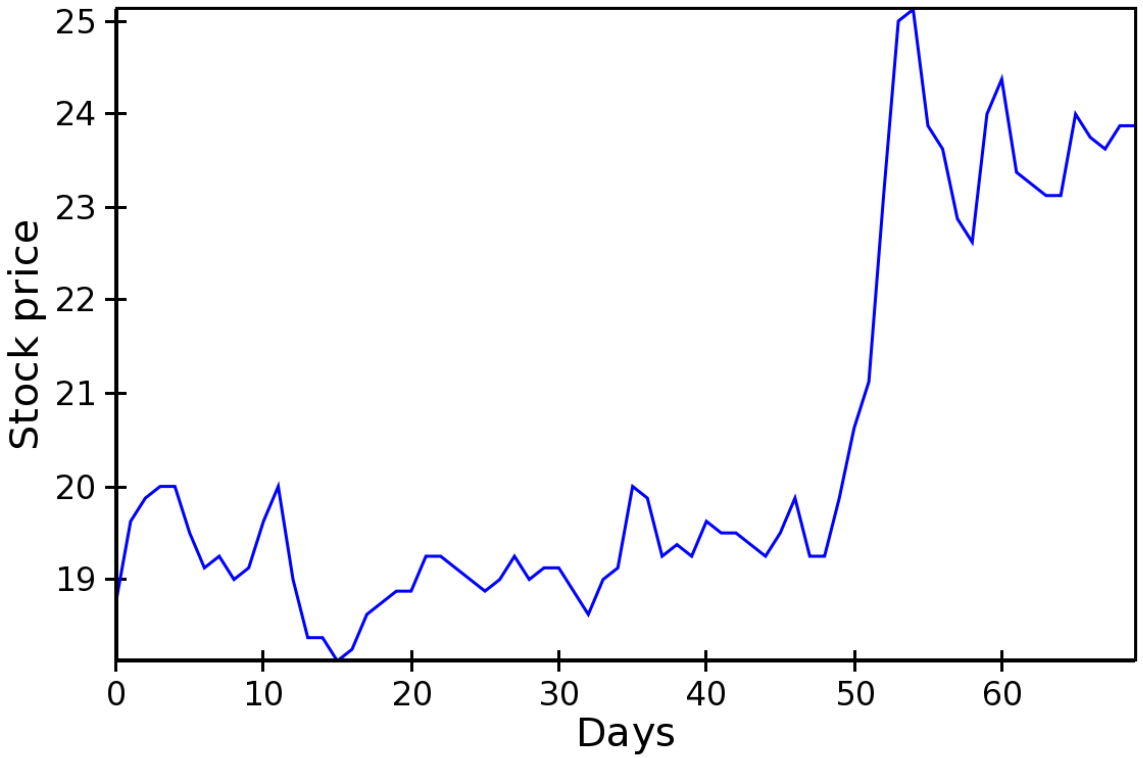
- **Scatter plot**

```
> qplot(revenues, profits, data=f500)
```



# Chart types

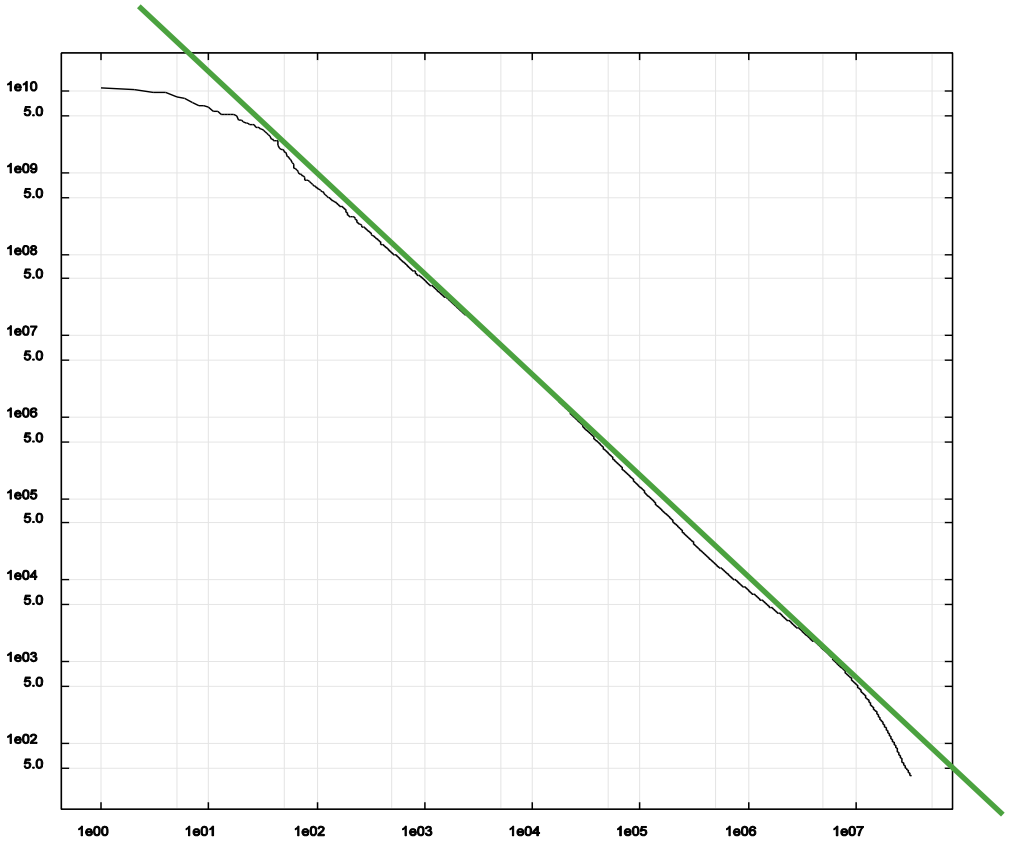
- **Line plot**



# Chart types

- **Log-log plot:** Very useful for power law data

Frequency of words in tweets




slope ~ -1

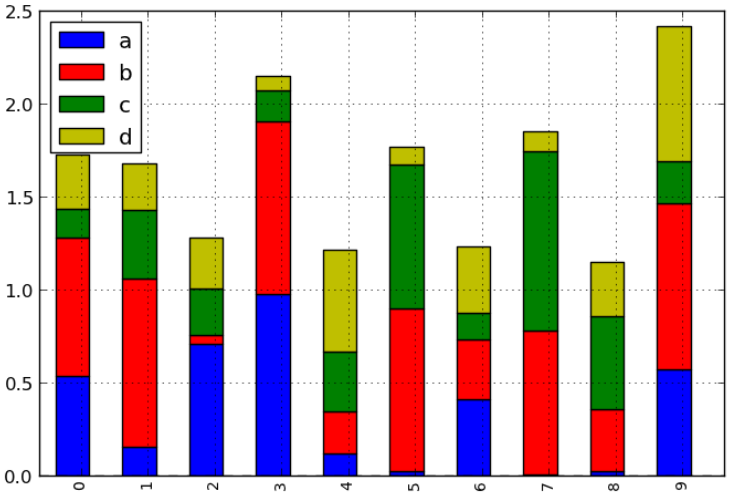
Rank of words in tweets, most frequent to least:
I, the, you,…

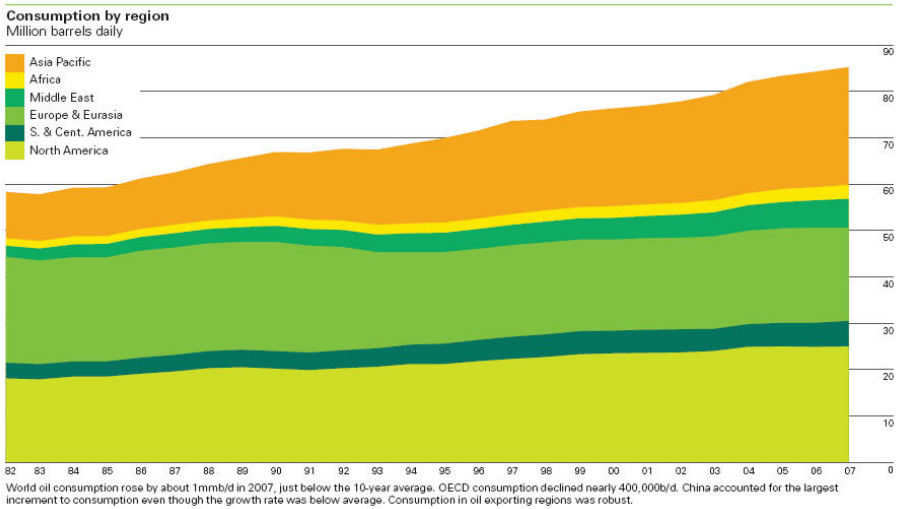
# Chart types

- More than two variables
  - Stacked plots
  - Parallel coordinate plot

# Chart types

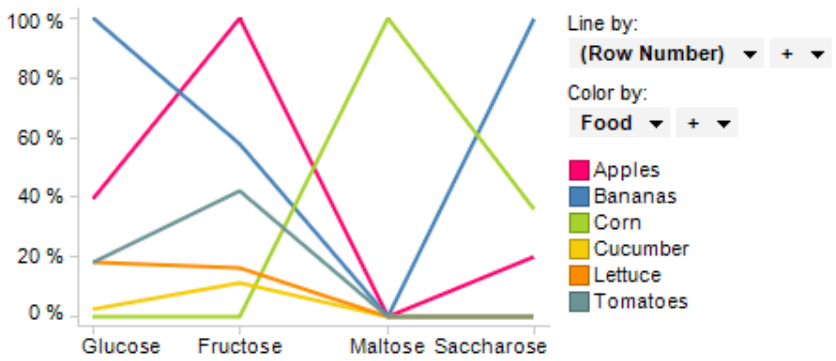
- **Stacked plot:** stack variable is discrete:

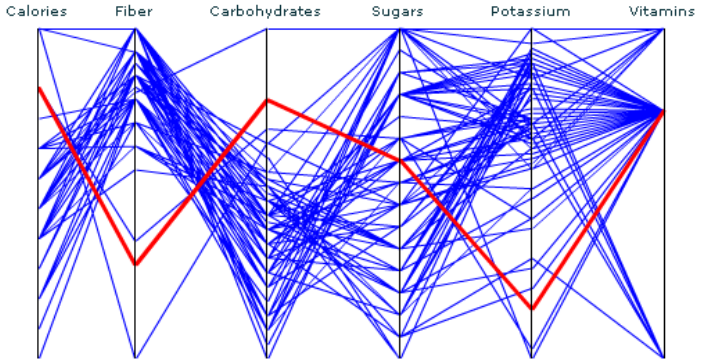


# Chart types

- **Parallel coordinate plot:** one discrete variable, an arbitrary number of other variables:



# 5-minute break

# Normal Distributions, Mean, Variance

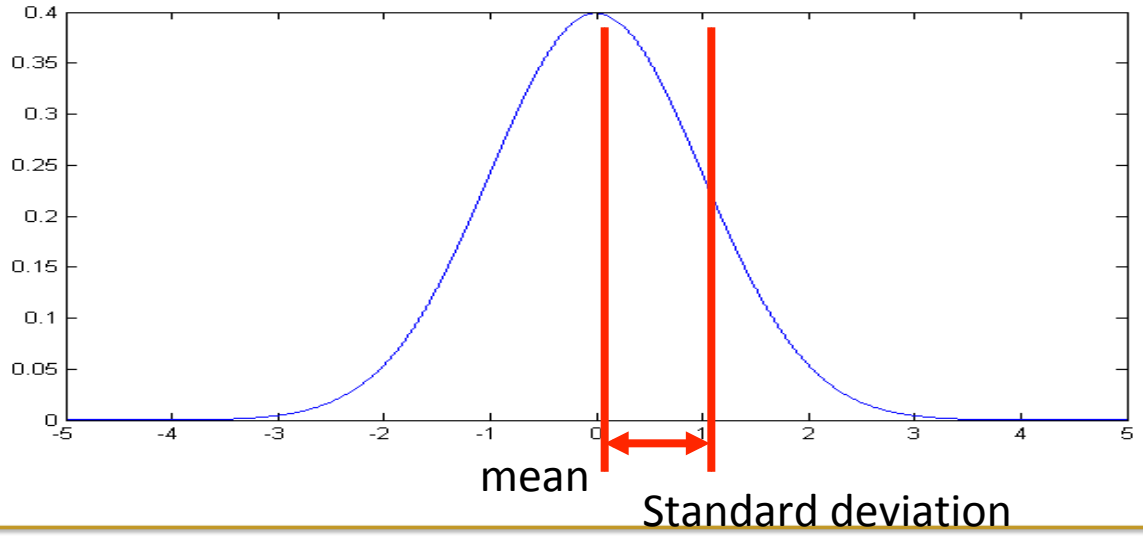
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

$$Var(X)=1/n \sum i=1 \uparrow n \blacksquare (X \downarrow i -X)\uparrow 2$$

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.

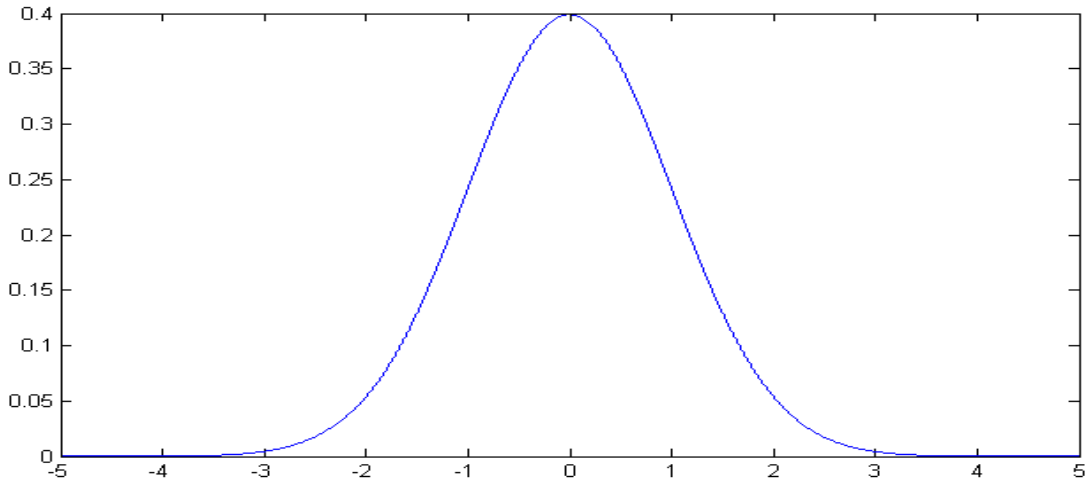


# Central Limit Theorem

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.

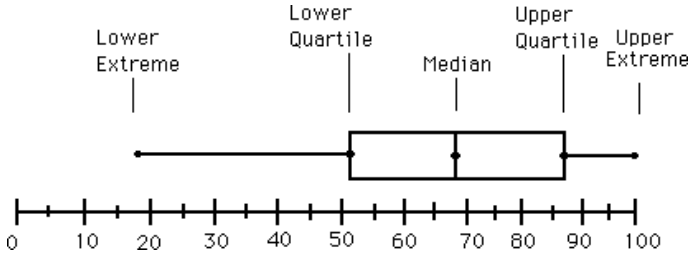
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



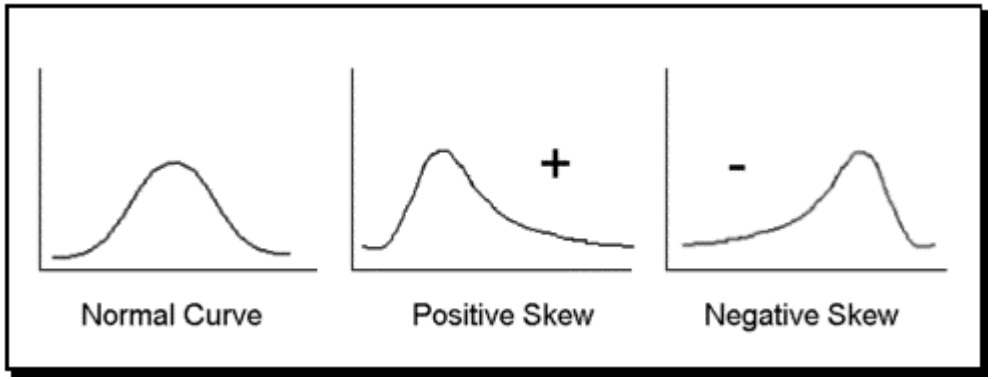
# Correcting distributions

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

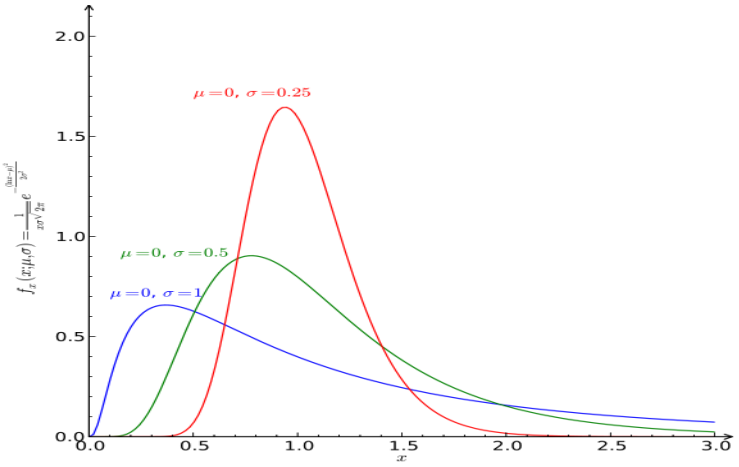


# Correcting distributions

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 sdev. sqrt(k). Then sqrt(X) is approximately normally distributed with sdev 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

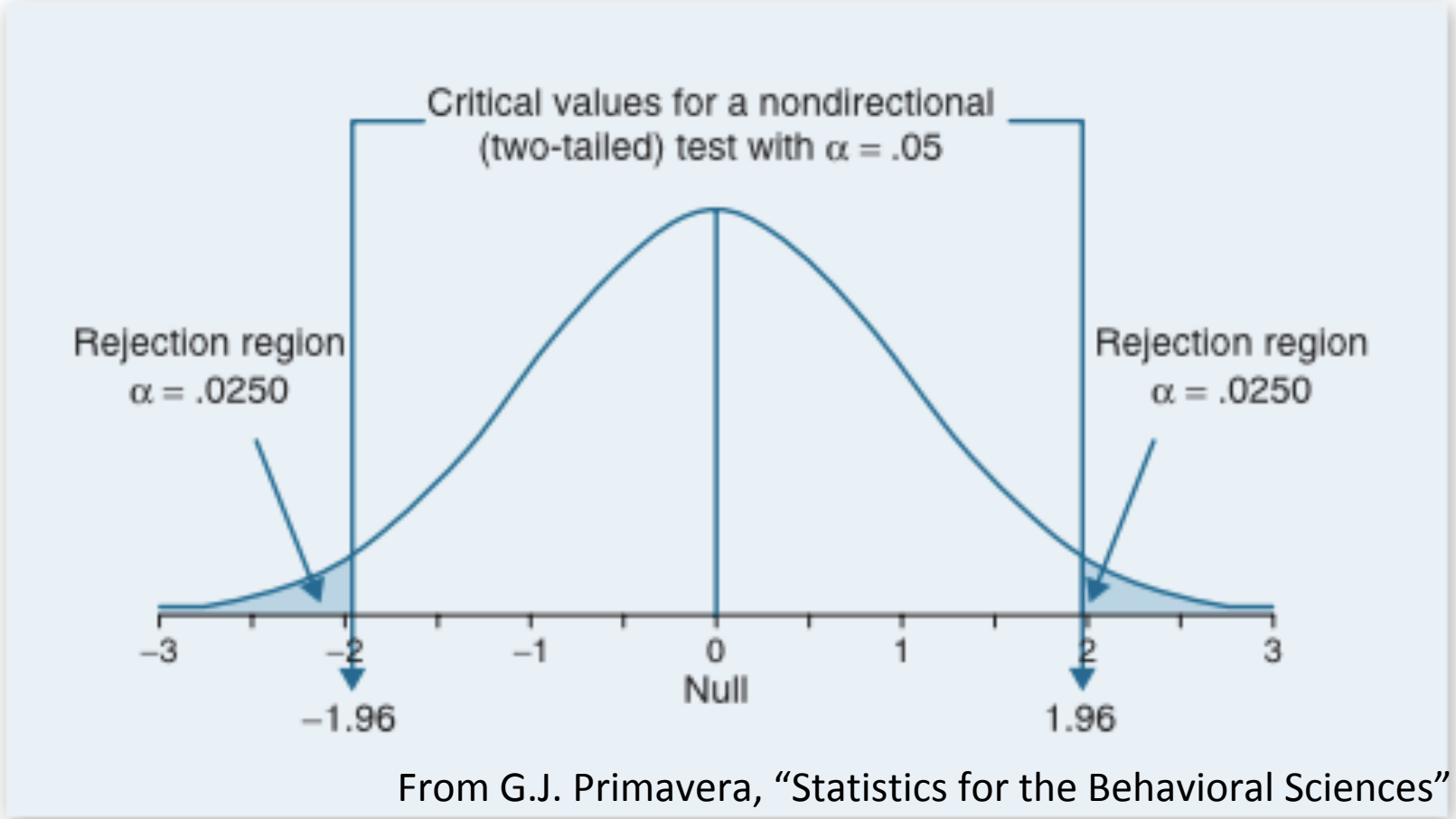
He called back the “psychic” subjects and had them do the same test again. They all failed.

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

# p Value

- A p value is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true.
- In many cases: when the p value is less than 5% ($p < .05$), we reject the null hypothesis
  - Note this means that 1 out of 20 times we incorrectly reject the null hypothesis
  - Do “green jelly beans cause acne?” (see XKCD)

# 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

# Three important tests

- **T-test:** compare two groups, or two interventions on one group.
- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table".
- **ANOVA:** compare outcomes under several discrete interventions.

# T-test

**Single-sample:** Compute the test statistic:

$$t = \bar{X} / \sigma$$

where $\bar{X}$ is the sample mean and $\sigma$ is the sample standard deviation, which is the square root of the sample variance Var(X).

If X is normally distributed, t is **almost** normally distributed, but not quite because of the presence of $\sigma$ .

You use the single-sample test for **one group** of individuals in **two conditions**. Just subtract the two measurements for each person, and use the difference for the single sample t-test.

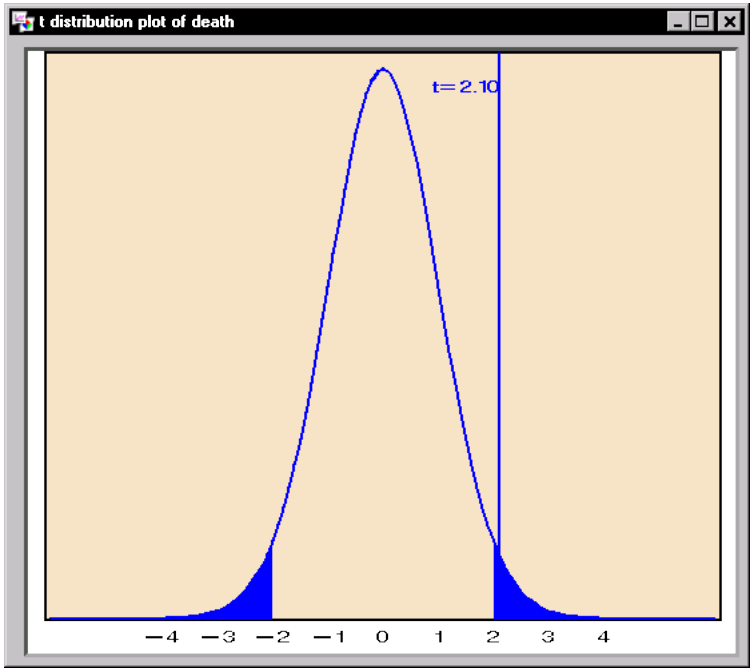
This is called a **within-subjects** design.

# T-statistic and T-distribution

- We use the t-statistic from the last slide to test whether the mean of our sample could be zero.
- If the underlying population has mean zero, the t-distribution should be distributed like this:

- The area of the tail beyond our measurement tells us how likely it is under the null hypothesis.

- If that probability is low (say < 0.05) we reject the null hypothesis.



# Two sample T-test

In this test, there are **two samples** $X\downarrow 1$ and $X\downarrow 2$ . A t statistic is constructed from their sample means and sample standard deviations:

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

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

You should try to understand the formula, but you shouldn't need to use it. most stat software exposes a function that takes the samples $X\downarrow 1$ and $X\downarrow 2$ as inputs directly.

This design is called a **between-subjects** test.

# Chi-squared test

Often you will be faced with discrete (count) data. Given a table like this:

| | Prob(X) | Count(X) |
|---|---|---|
| X=0 | 0.3 | 10 |
| X=1 | 0.7 | 50 |

Where Prob(X) is part of a null hypothesis about the data (e.g. that a coin is fair).

The CHI-squared statistic lets you test whether an observation is consistent with the data:

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

$O_i$ 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.

# Fisher's exact test

In case we only have counts under different conditions

| | Count1(X) | Count2(X) |
|---|---|---|
| X=0 | a | b |
| X=1 | c | d |

We can use Fisher's exact test (n = a+b+c+d):

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Which gives the probability directly (its not a statistic).

# One-Way ANOVA

ANOVA (ANalysis Of VAriance) allows testing of multiple differences in a single test. Suppose our experiment design has an independent variable Y with four levels:

Y

| Primary School | High School | College | Grad degree |
|---|---|---|---|
| 4.1 | 4.5 | 4.2 | 3.8 |

The table shows the mean values of a response variable (e.g. avg number of Facebook posts per day) in each group.

We would like to know in a single test whether the response variable depends on Y, at some particular significance such as 0.05.

# ANOVA

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

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

The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.

We can look up the F-statistic value in a cumulative F-distribution (similar to the other statistics) to get the p-value.

ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

# Closing Words

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 assumption 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.

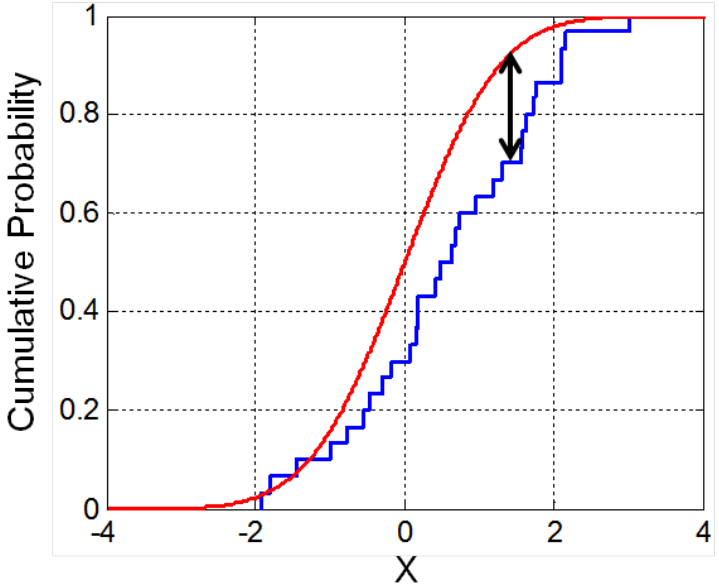
In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution.

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.

# The Need for Models

"All models are wrong, but some models are useful."
George Box

- Data represents the traces of the real-world processes.
- Two sources of randomness and uncertainty:
  1) those underlying the process themselves
  2) those associated with the data collection methods
- To simplify the traces into something more comprehensible you need:
  - mathematical models or functions of the data -> Statistical estimators

# More on Models

- N is size of population
- n is sample size (subset of the population)

- Getting the subset (i.e. sampling) can introduce "bias" leading to incorrect conclusions

# Probability Distributions

- Natural processes tend to generate measurements whose empirical shape could be approximated by mathematical functions with a few parameters that could be estimated from the data.

# Note on ML Algos vs. Stat Models

- Techniques and underlying concepts in common
- Difference in goals/use:
  - ML Algos – goal: predict or classify with high accuracy.
    - basis of many data products
  - Models – get at the underlying generative process
- "Black box" vs. "White box"
- Dealing with uncertainty (at the heart of stats)
- Distributions vs. non-parametic approaches

# EDA Part 2
# Stats and Featurization

# Outline

- Statistics
  - Measurement
  - Hypothesis Testing
- Featurization
  - Feature selection
  - Feature Hashing
- Visualizing Accuracy

# Measurement

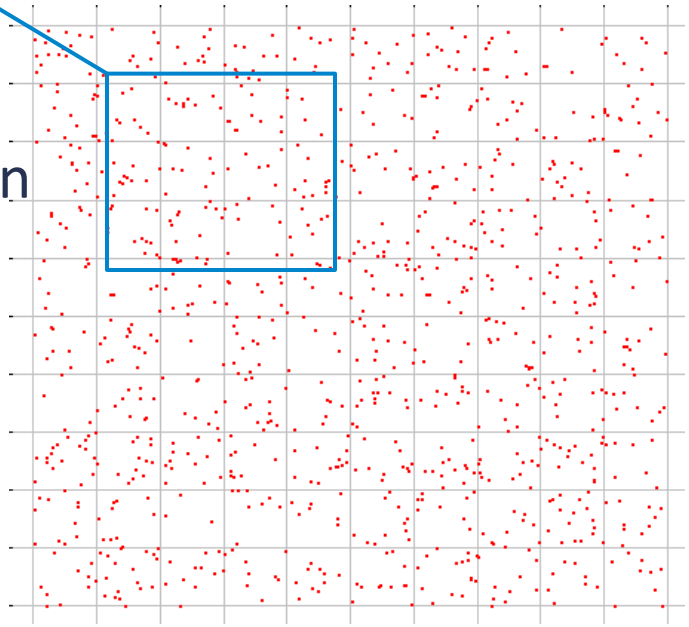
- **Measurement:** We often want to measure properties of data or models. For the data:
  - **Basic properties:** Min, max, mean, std. deviation of a dataset.
  - **Relationships:** between fields (columns) in a tabular dataset, via scatter plots, regression, correlation etc.
- And for models:
  - **Accuracy:** How well does our model match the data (e.g. predict hidden values)?
  - **Performance:** How fast is a ML system on a dataset? How much memory does it use? How does it scale as the dataset size grows?

# Measurement on Samples

- Many datasets are **samples** from an **infinite population**.
- We are most interested in **measures on the population**, but we have access only to a **sample** of it.

A sample measurement is called a **"statistic".** Examples:

- Sample min, max, mean, std. deviation

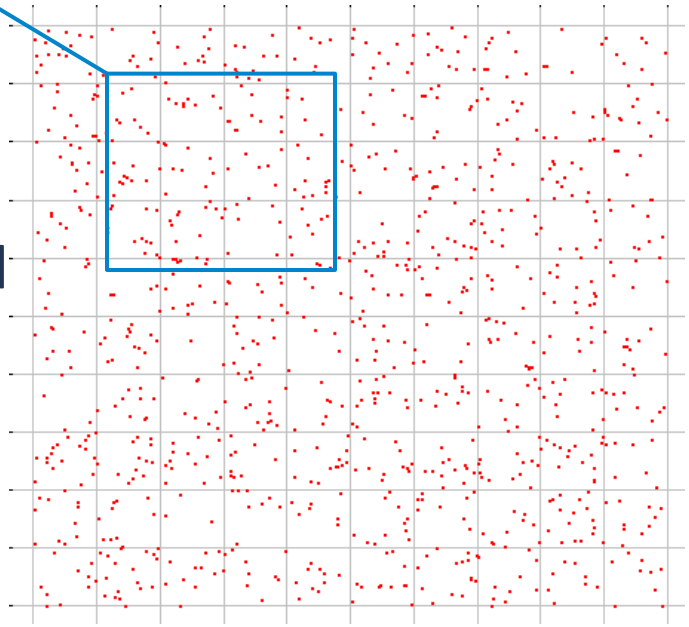


# Measurement on Samples

- Many datasets are **samples** from an **infinite population**.
- We are most interested in **measures on the population**, but we have access only to a **sample** of it.

That makes measurement hard:

- Sample measurements are "noisy," i.e. vary from one sample to the next
- Sample measurements may be biased i.e. systematically be different from the measurement on the population.

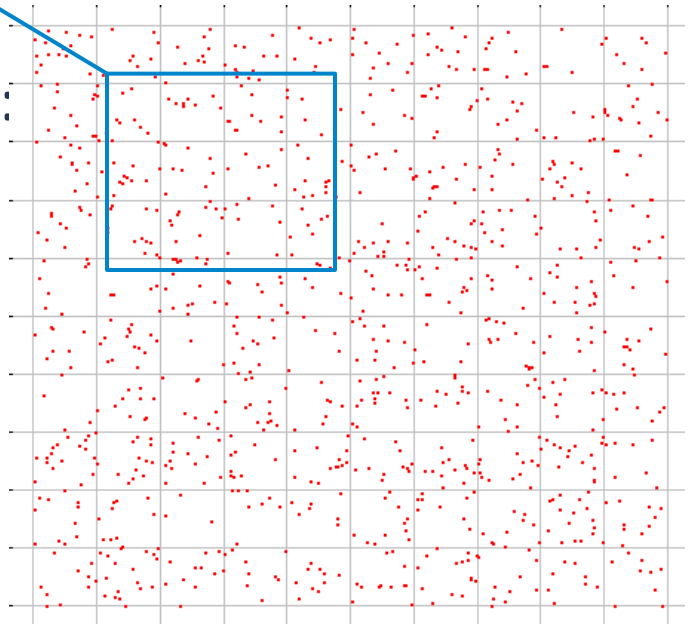


# Measurement on Samples

- Many datasets are **samples** from an **infinite population**.
- We are most interested in **measures on the population**, but we have access only to a **sample** of it.

That makes measurement hard:

- Sample measurements have **variance**: variation between samples
- Sample measurements have **bias**, systematic variation from the population value.



# Examples of Statistics

Unbiased:

- Sample mean (sample of n values) $x = \sum i=1 \uparrow n \blacksquare x \downarrow i \ / n$
- Sample median ($k^{th}$ largest in 2k-1 values)

Biased:

- Min
- Max
- Sample variance $\sigma \uparrow 2 = \sum i=1 \uparrow n \blacksquare (x - x\ ) \uparrow 2 \ / n$
- (but this does correctly give population variance in the limit as $n \rightarrow \infty$)

For biased estimators, the bias is usually worse on small samples.

# Statistical Notation

We'll use upper case symbols "$X$" to represent random variables, which you can think of as draws from the entire population.

Lower case symbols "$x$" represent particular samples of the population, and subscripted lower case symbols to represent instances of a sample: $x \downarrow i$

# Measurement

- Statistics
  - Measurement
  - **Hypothesis Testing**
- Featurization
  - Feature selection
  - Feature Hashing
- Visualizing Accuracy

# 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 $H_0$**.

# Hypothesis Testing

Example:

- We suspect that a particular coin isn't fair.
- We toss it 10 times, it comes up heads every time…
- We conclude it's not fair, why?
- How sure are we?

Now we toss a coin 4 times, and it comes up heads every time.

- What do we conclude?

# Hypothesis Testing

- We want to prove a hypothesis $H_A$ **(the coin is biased),** but its hard so we try to **disprove a null hypothesis $H_0$ (the coin is fair)**.

- A **test statistic** is some measurement we can make on the data which is likely to be **big under $H_A$** but **small under $H_0$**.
  **the number of heads after k coin tosses – one sided**
  **the difference between number of heads and k/2 – two-sided**

- **Note:** tests can be either one-tailed or two-tailed. Here a two-tailed test is convenient because it checks either very large or very small counts of heads.

# Hypothesis Testing

- Another example:
  - 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.

* - We need to use the fact that the sum of two independent, normally-distributed variables is also normally distributed.

# Hypothesis Testing – contd.

- s = mean(a) – mean(b) is our test statistic, $H_0$ the null hypothesis that mean(A)=mean(B)
  - We reject if $Pr(x > s \mid H_0) < p$, i.e. the probability of a statistic value **at least as large as** s, should be small.
  - 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

# Hypothesis Testing

- Compare an experimental group and a control group
- $H_0$: Null Hypothesis
  - No difference between the groups
- $H_A$: Alternative Hypothesis
  - Statistically significant difference between the groups
- "difference" defined in terms of some **test statistic**
  - Different means (e.g., t-test), different variances (e.g., F-test)
- Groups defined through careful experimental design
  - randomized, blinded, double-blinded
- Examples:
  - "The new ad placement produces more click-throughs"
  - "This treatment produces better outcomes"

# More on Hypothesis Testing

- Null Hypothesis is given the benefit of the doubt (e.g., innocent until proven guilty).
- Alternative Hypothesis directly contradicts the Null Hypothesis
- "Step 1: State the hypotheses."
- "Step 2: Set the criteria for a decision."
- "Step 3: Compute the test statistic."
- "Step 4: Make a decision."

# Hypothesis Testing

| | | Decision | |
|---|---|---|---|
| | | **Retain the null** | **Reject the null** |
| **Truth in the population** | True | CORRECT<br>$1 - \alpha$ | TYPE I ERROR<br>$\alpha$ |
| | False | TYPE II ERROR<br>$\beta$ | CORRECT<br>$1 - \beta$<br>POWER |

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

# Non-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

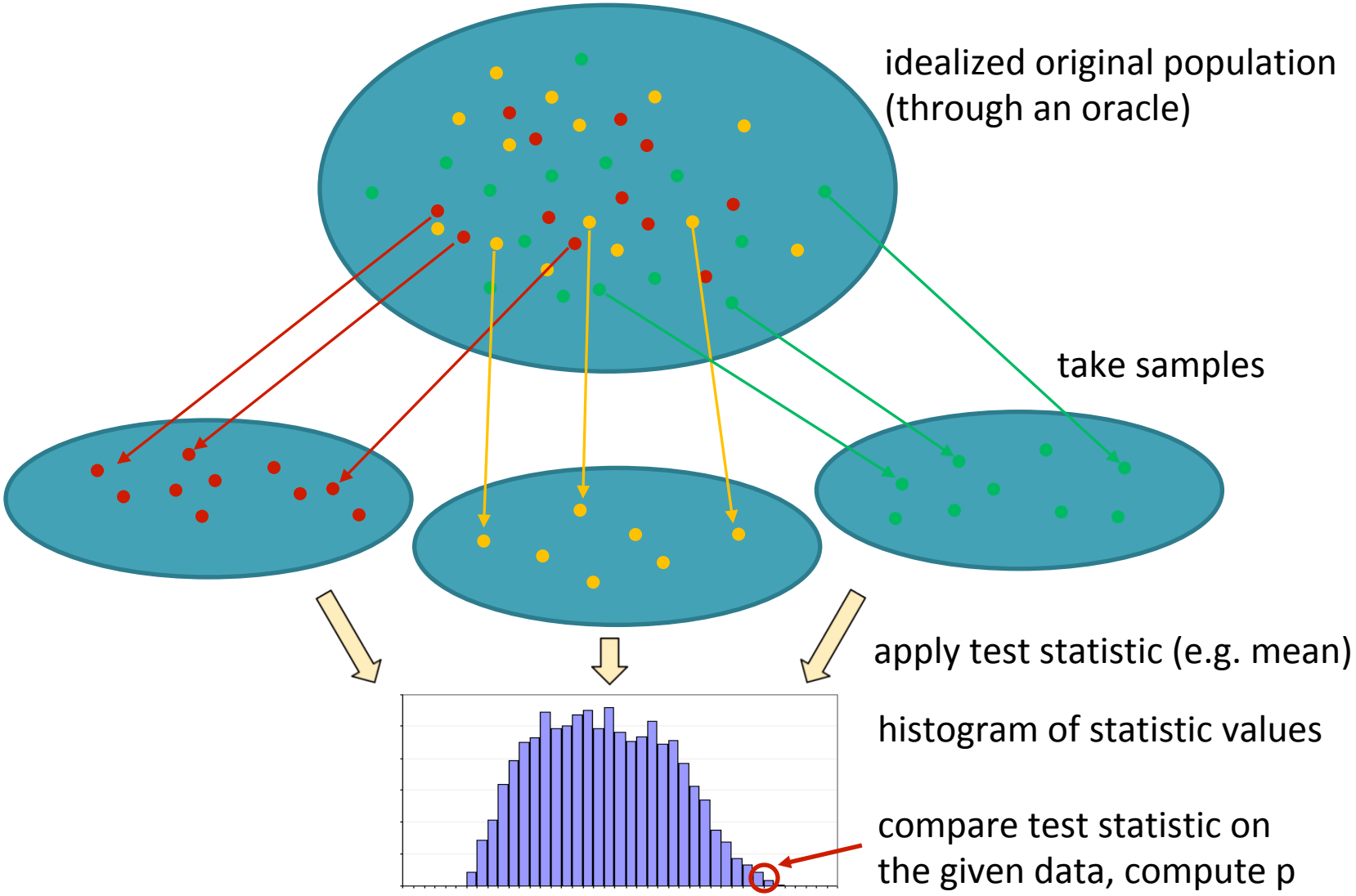
These tests make no assumption about the distribution of the input data, and can be used on very general datasets:

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

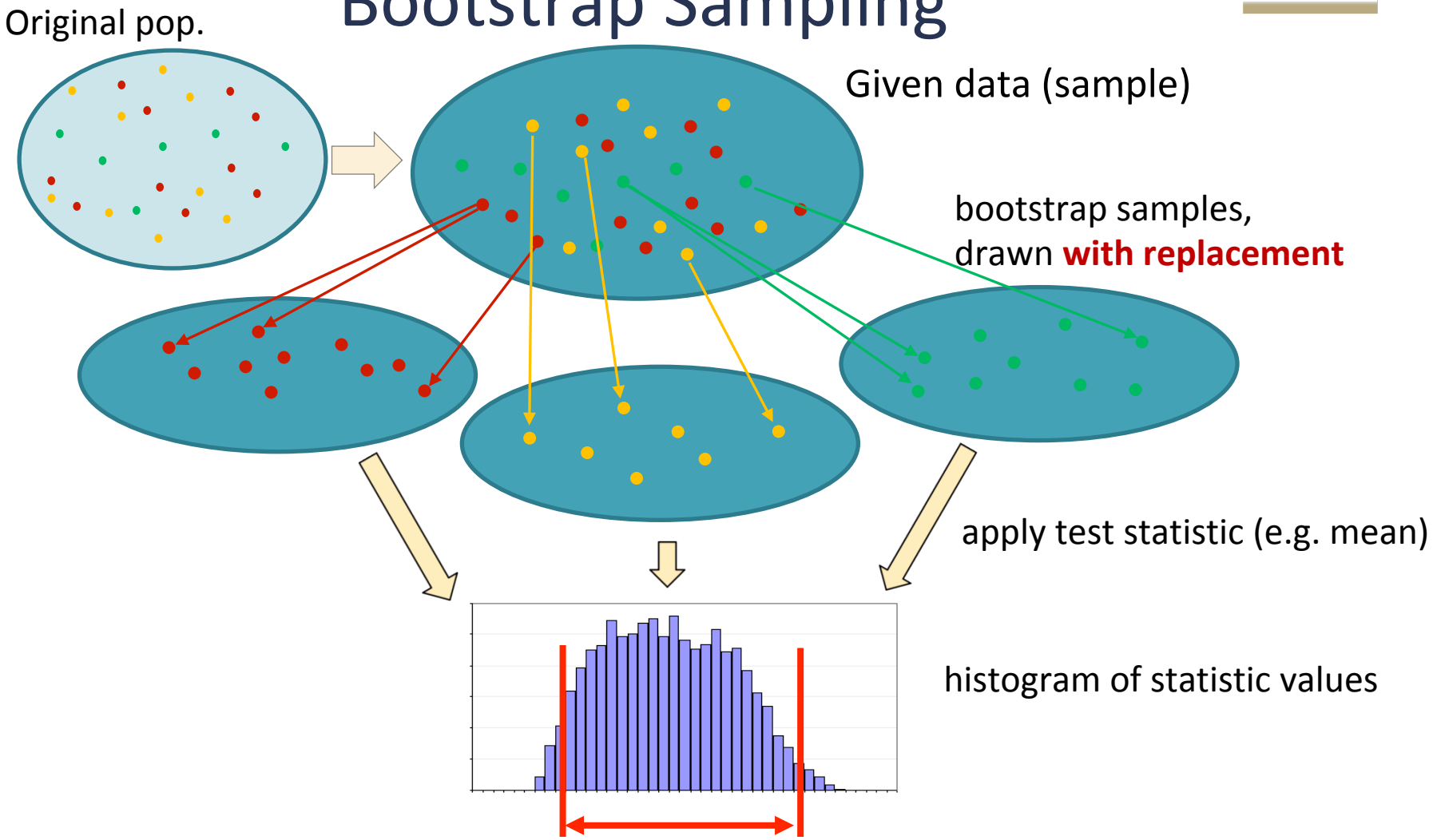
# Bootstrap samples

- Often you have only one sample of the data, but you would like to know how some measurement would vary across similar samples (i.e. the variance or histogram of a statistics).
- You can get a good approximation to related samples by "resampling your sample".
- This is called bootstrap sampling (by analogy to lifting yourself up by your bootstraps).
- For a sample S of N values, a bootstrap sample is a set $S_B$ of N values drawn with replacement from S.

# Idealized Sampling



# Bootstrap Sampling




The region containing 95% of the samples is a 95% confidence interval (CI)

# Bootstrap Confidence Interval tests

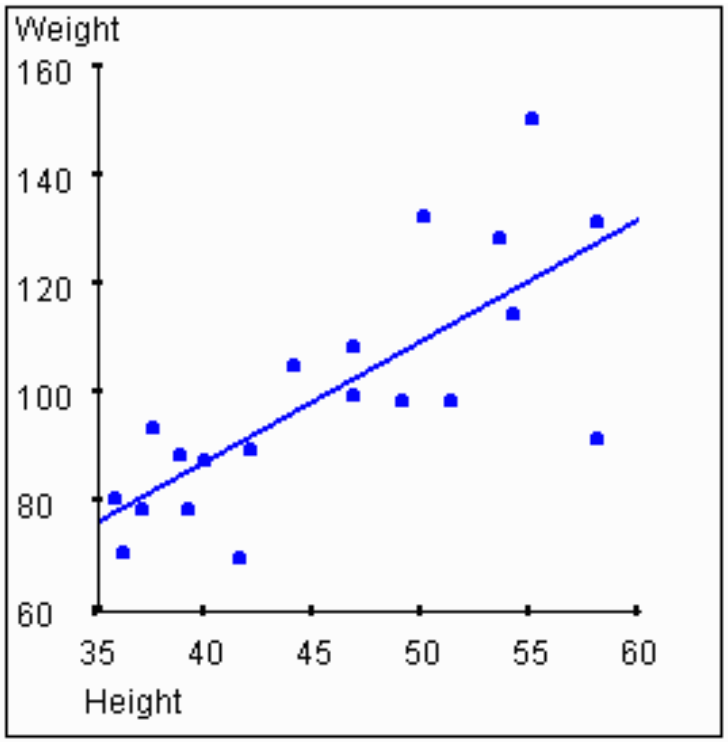
Then a test statistic outside the 95% Confidence Interval (CI) would be considered significant at 0.05, and probably not drawn from the same population.

e.g. Suppose the data are differences in running times between two algorithms. If the 95% bootstrap CI does not contain zero, then original distribution probably has a mean other than zero, i.e. the running times are different.

We can also test for values other than zero. If the 95% CI contains only values greater than 2, we conclude that the difference in running times is significantly larger than 2.

# Bootstrap Test for Regression

- Suppose we have a single sample of points, to which we fit a regression line?
- How do we know whether this line is “significant”? And what do we mean by that?

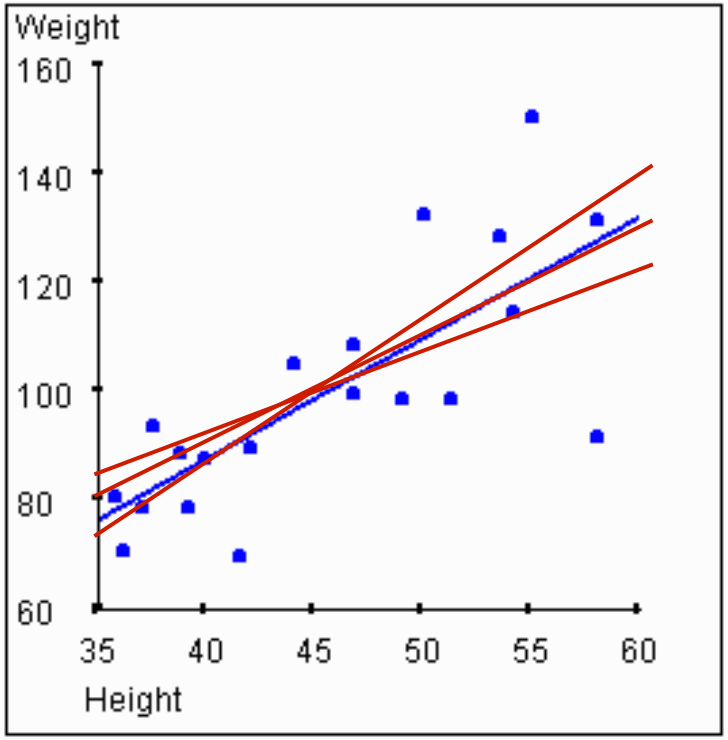


# Bootstrap Test for Regression

ANS: Take bootstrap samples, and fit a line to each sample.

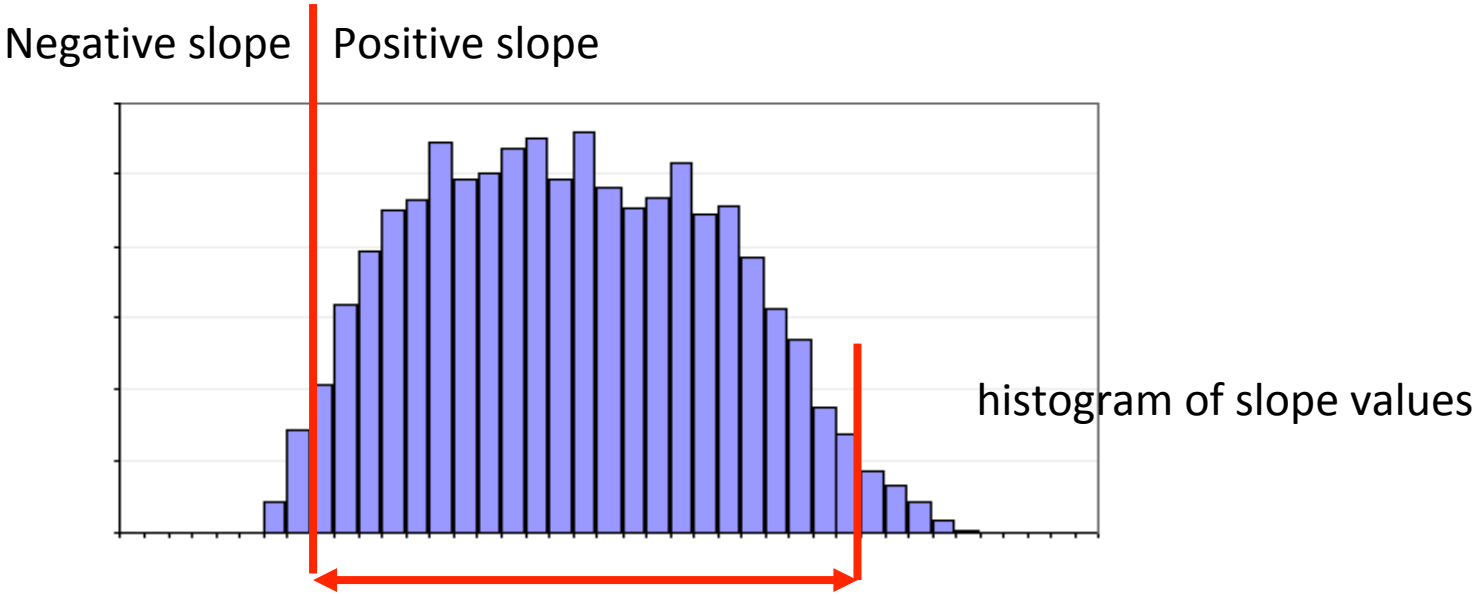
The possible regression lines are shown below:

What we really want to know is “how likely is a line with zero or negative slope”.



# Bootstrap Test for Regression

- ANS: Take bootstrap samples, and fit a line to each sample.
- The possible regression lines are shown below:
- What we really want to know is “how likely is a line with zero or negative slope”.




The region containing 95% of the samples is a 95% confidence interval (CI)

# Outline

- Statistics
  - Measurement
  - Hypothesis Testing
- **Featurization – train/test/validation sets**
  - Feature selection
  - Feature Hashing
- Visualizing Accuracy

# Train-Test-Validation Sets

- When making measurements on a ML algorithm, we have additional challenges.
- With a sample of data, any model fit to it models both:
  1. Structure in the **entire population**
  2. Structure in the **specific sample not true of the population**

1. is good because it will generalize to other samples.

2. is bad because it wont.

**Example:** a 25-year old man and a 30-year old woman.

- Age predicts gender perfectly. (age < 27 => man else woman)
- Gender predicts age perfectly. (gender == man => 25 else 30)

Neither result generalizes. This is called **over-fitting**.

# Train-Test-Validation Sets

**Train/Test split:**

- By (randomly) partitioning our data into train and test sets, we can avoid biased measurements of performance.
- The model now fits a **different sample** from the measurement.
- ML models are trained only on the training set, and then measured on the test set.

**Example:**

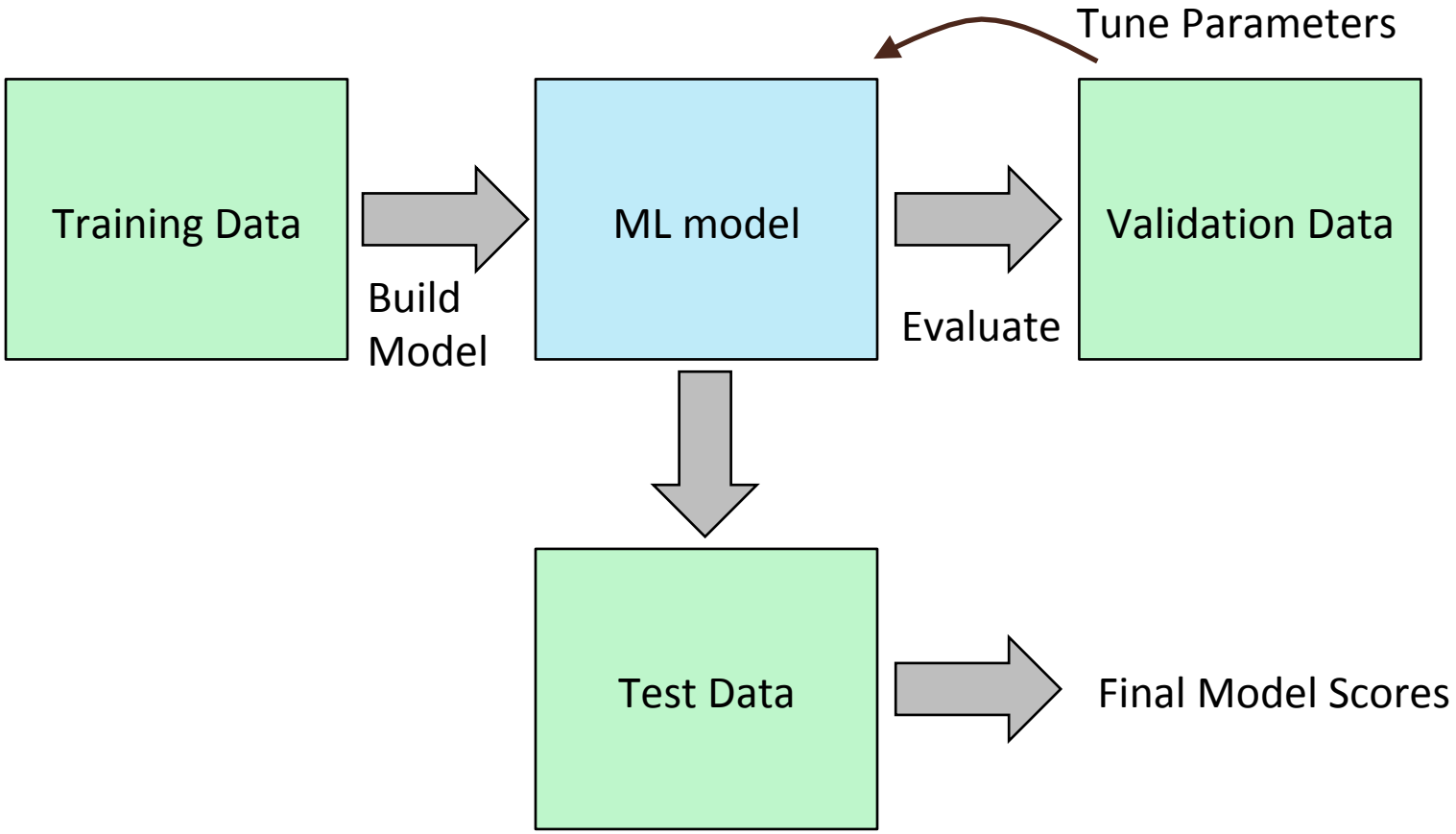
- Build a model of age/gender based on the man/woman above.
- Now select a test set of 40 random people (men + women).
- The model will fail to make reliable predictions on this test set.

# Validation Sets

- Statistical models often include “tunable” parameters that can be adjusted to improve accuracy.
- You need a test-train split in order to measure performance for each set of parameters.
- But now you’ve used the test set in model-building which means the model might over-fit the test set.
- For that reason, its common to use a third set called the ***validation set*** which is used for parameter tuning.

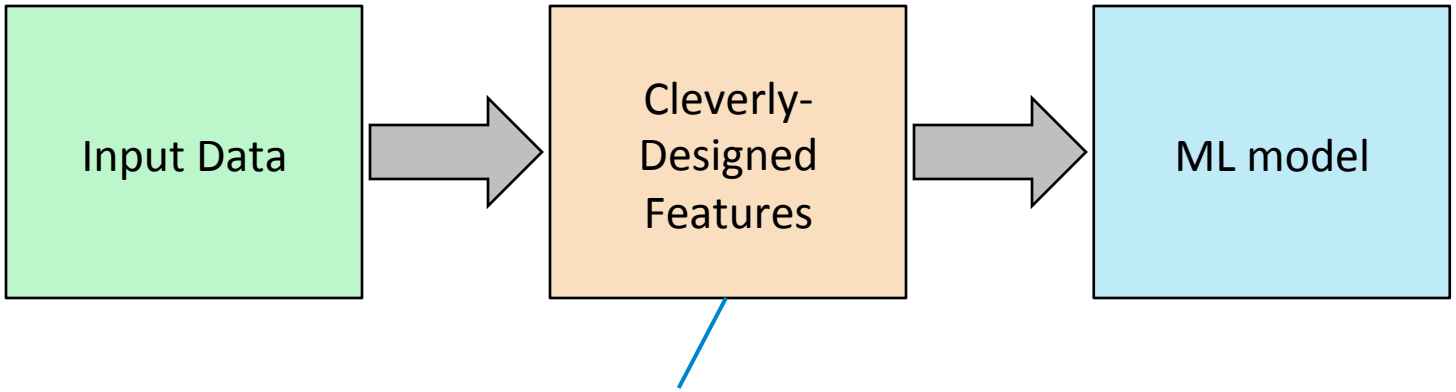
- A common dataset split is 60-20-20 training/validation/test

# Model Tuning



# A Brief History of Machine Learning

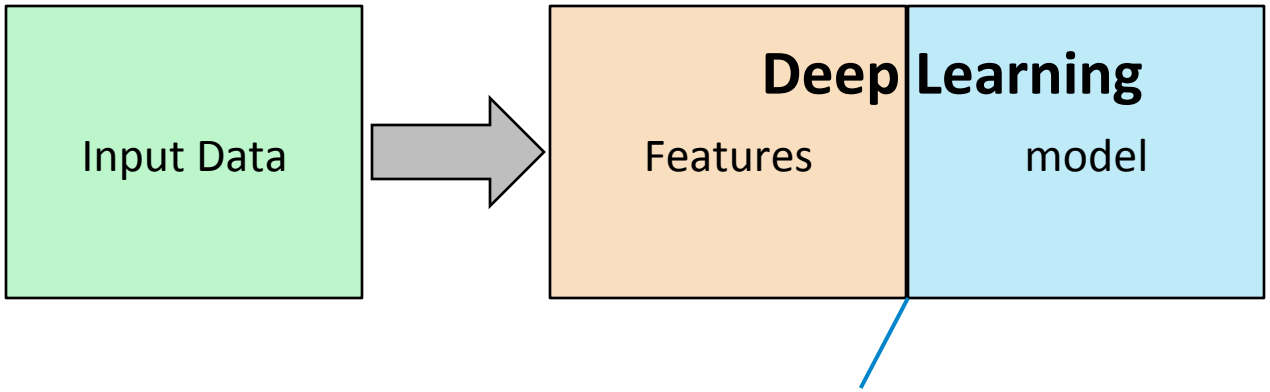
- Before 2012*:




* Before publication of Krizhevsky et al.'s ImageNet CNN paper.

# A Brief History of Machine Learning

- After 2012:



# A Brief History of Machine Learning

- But this (pre-2012) picture is still typical of many pipelines.
- We'll focus on one aspect of feature design: feature selection, i.e. choosing which features from a list of candidates to use for a ML problem.

# Method 1: Ablation

- Train a model on features $(f_1, ..., f_n)$, measure performance $Q_0$
- Now remove a feature $f_k$ and train on $(f_1, ..., f_{k-1}, f_{k+1}, ..., f_n)$, producing performance $Q_1$.

- If performance $Q_1$ is significantly worse than $Q_0$, keep $f_k$ otherwise discard it.

Q: How do we check if "$Q_1$ is significantly worse than $Q_0$"

If

# Method 1: Ablation

- Train a model on features $(f_1, ..., f_n)$, measure performance $Q_0$
- Now **remove a feature** $f_k$ and train on $(f_1, ..., f_{k-1}, f_{k+1}, ..., f_n)$, producing performance $Q_1$.

- If performance $Q_1$ is **significantly worse** than $Q_0$, keep $f_k$ otherwise discard it.

Q: How do we check if "$Q_1$ is significantly worse than $Q_0$"

- If we know $Q_0$, $Q_1$ are normally-distributed with variance $\sigma$ we can **do a t-test**.

# Method 1: Ablation

- Train a model on features $(f_1, ..., f_n)$, measure performance $Q_0$
- Now remove a feature $f_k$ and train on $(f_1, ..., f_{k-1}, f_{k+1}, ..., f_n)$, producing performance $Q_1$.

- If performance $Q_1$ is significantly worse than $Q_0$, keep $f_k$ otherwise discard it.

Q: How do we check if "$Q_1$ is significantly worse than $Q_0$"

- Do **bootstrap sampling** on the training dataset, and compute $Q_0$, $Q_1$ on each sample.
- Then use an appropriate statistical test (e.g. a CI) on vectors of $Q_0$ $Q_1$ values generated by bootstrap samples.

# Method 1: Ablation

**Question:** Why do you think ablation starts with all the features and removes one-at-a-time rather than starting with no features, and adding one-at-a-time?

# Method 2: Mutual Information

Mutual information measures the extent to which **knowledge of one feature influences the distribution of another** (the classifier output).

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

Where U is a random variable which is 1 if term $e_t$ is in a given document, 0 otherwise. C is 1 if the document is in the class c, 0 otherwise. These are called indicator random variables.

Mutual information can be used to rank features, the highest will be kept for the classifier and the rest ignored.

# Method 3: CHI-Squared

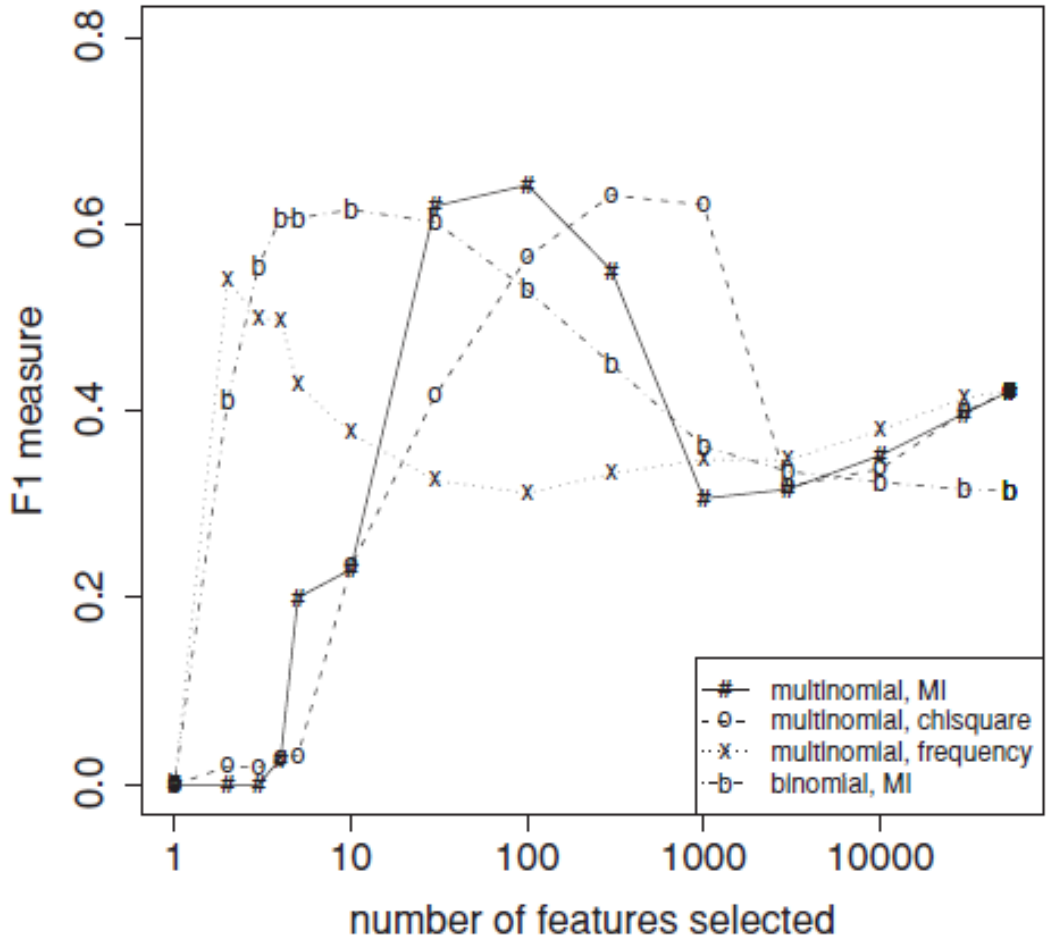
CHI-squared is an important statistic to know for comparing count data.

Here it is used to measure **dependence between word counts in documents and in classes**. Similar to mutual information, terms that show dependence are good candidates for feature selection.

CHI-squared can be visualized as a test on contingency tables like this one:

| | Right-Handed | Left-Handed | Total |
|---|---|---|---|
| Males | 43 | 9 | 52 |
| Females | 44 | 4 | 48 |
| Total | 87 | 13 | 100 |

# Example of Feature Count vs. Accuracy



# Outline

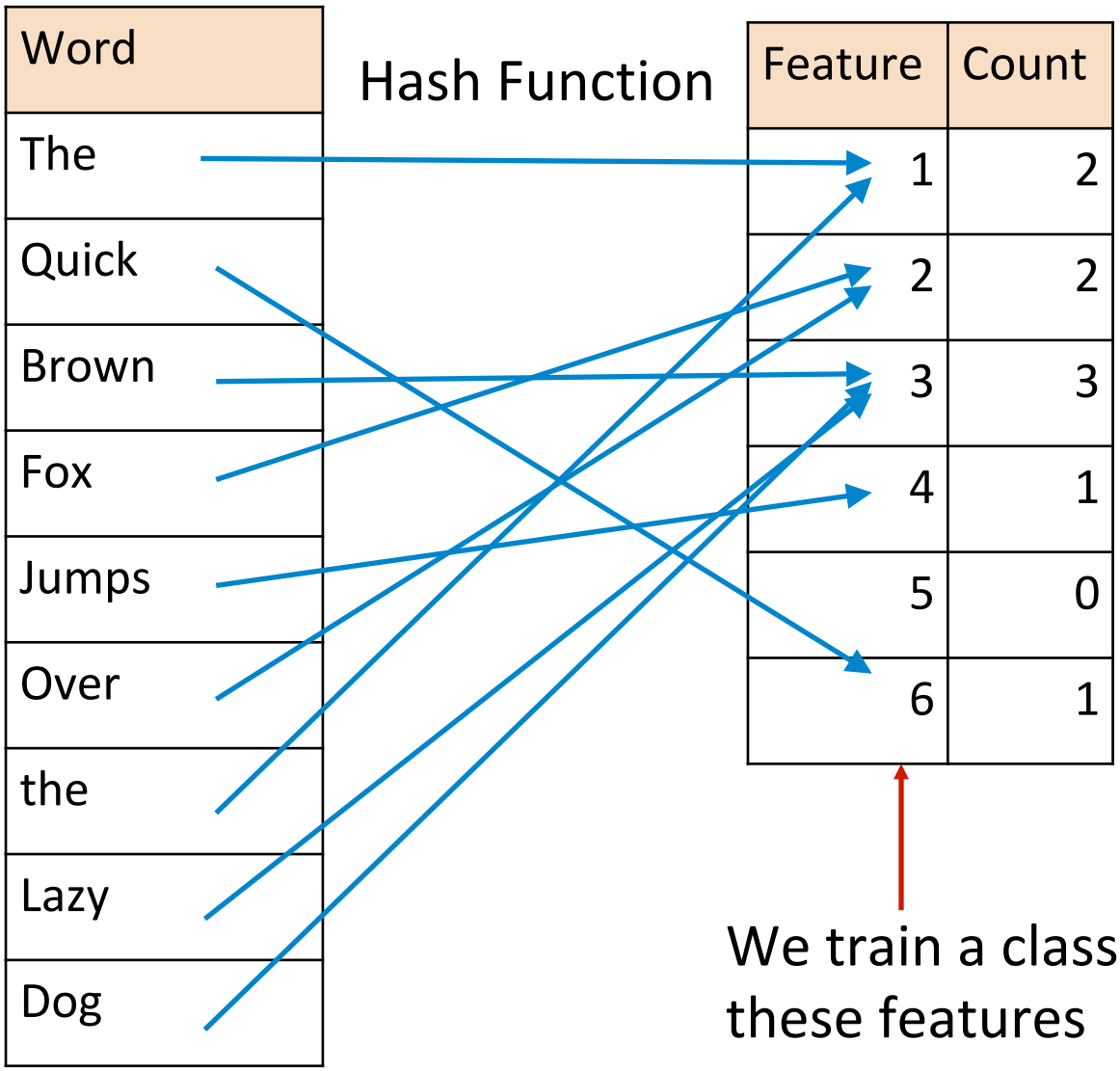
- Statistics
  - Measurement
  - Hypothesis Testing
- Featurization
  - Feature selection
  - **Feature Hashing**
- Visualizing Accuracy

# Feature Hashing

Challenge: many prediction problems involve very, very rare features, e.g. URLs or user cookies.

- There are billions to trillions of these, too many to represent explicitly in a model (or to run feature selection on!)
- Most of these features are not useful, i.e. don't help predict the target class.
- A small fraction of these features are **very** important for predicting the target class (e.g. user clicks on a BMW dealer site has some interest in BMWs).

# Feature Hashing




Feature table much smaller than feature set.

We train a classifier on these features

# Feature Hashing

- Feature 3 receives “Brown”, “Lazy” and “Dog”.
- The first two of these are not very salient to the category of the sentence, but “Dog” is.
- Classifiers trained on hashed features often perform surprisingly well – although it depends on the application.
- They work well e.g. for add targeting, because the false positive cost (target dog ads to non-dog-lovers) is low compared to the false negative cost (miss an opportunity to target a dog-lover).

# Feature Hashing and Interactions

- One very important application of feature hashing is to **interaction features**.
- Interaction features (or just interactions) are tuples (usually pairs) of features which are treated as single features.
- E.g. the sentence “the quick brown fox…” has interaction features including: “quick-brown”, “brown-fox”, “quick-fox” etc.
- Interaction features are often worth “more than the sum of their parts” e.g. “BMW-tires,” “ipad-charger,” “school-bags”
- There are $N^2$ interactions among N features, but very few are meaningful. Hashing them produces many collisions but most don’t matter.

# Outline

- Statistics
  - Measurement
  - Hypothesis Testing
- Featurization
  - Feature selection
  - Feature Hashing
- **Visualizing Accuracy**

# Why not to use “accuracy” directly

The simplest measure of performance would be the fraction of items that are correctly classified, or the “accuracy” which is:

$$\frac{tp + tn}{tp + tn + fp + fn}$$

(tp = true positive, fn = false negative etc.).

But this measure is dominated by the larger set (of positives or negatives) and favors trivial classifiers.

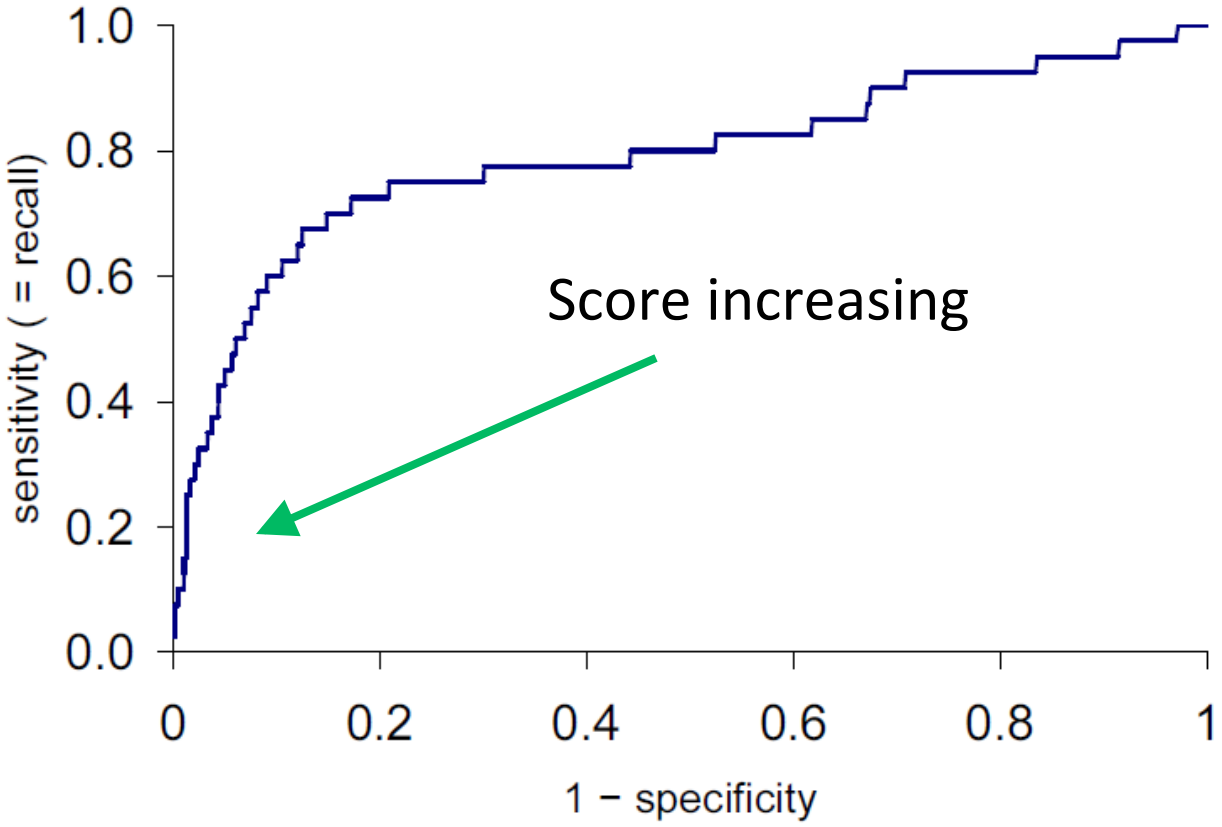
e.g. if 5% of items are truly positive, then a classifier that always says “negative” is 95% accurate.

# ROC plots

ROC is Receiver-Operating Characteristic. ROC plots

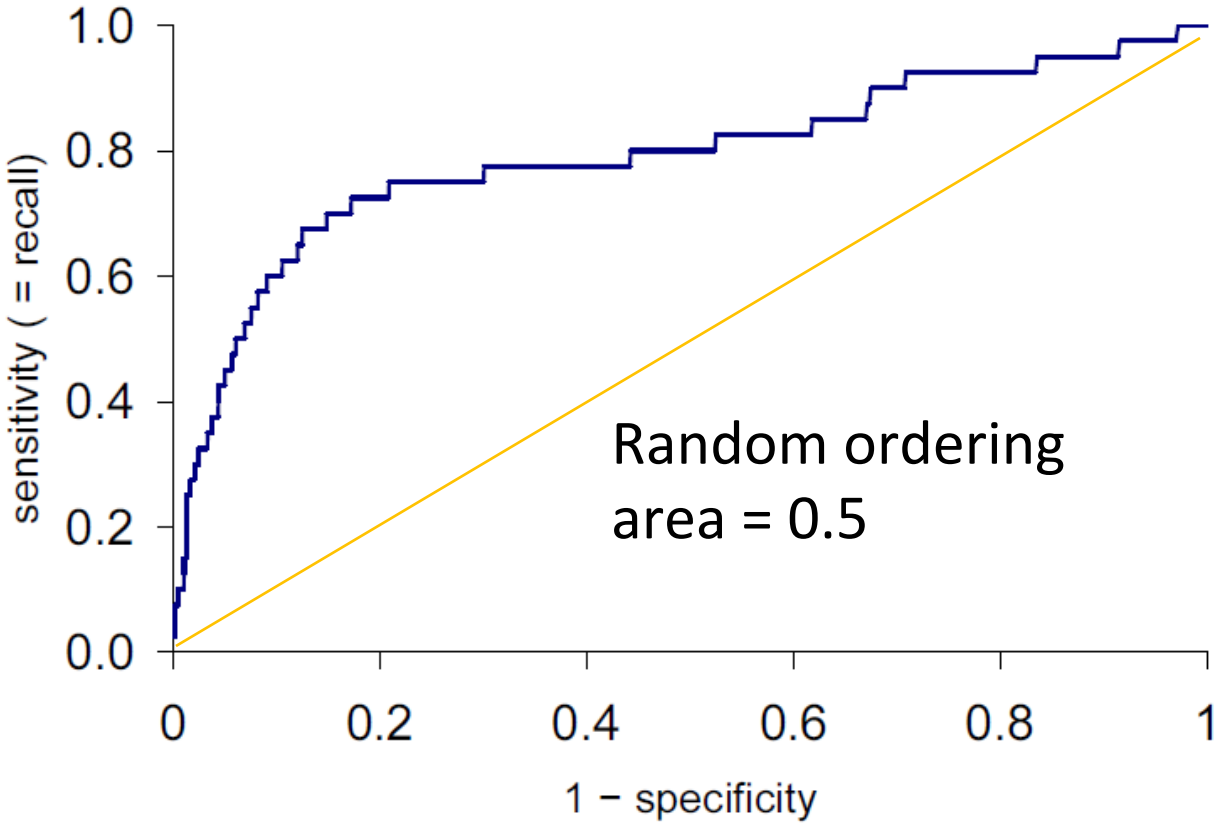
Y-axis: true positive rate = tp/(tp + fn), same as recall

X-axis: false positive rate = fp/(fp + tn) = 1 - specificity



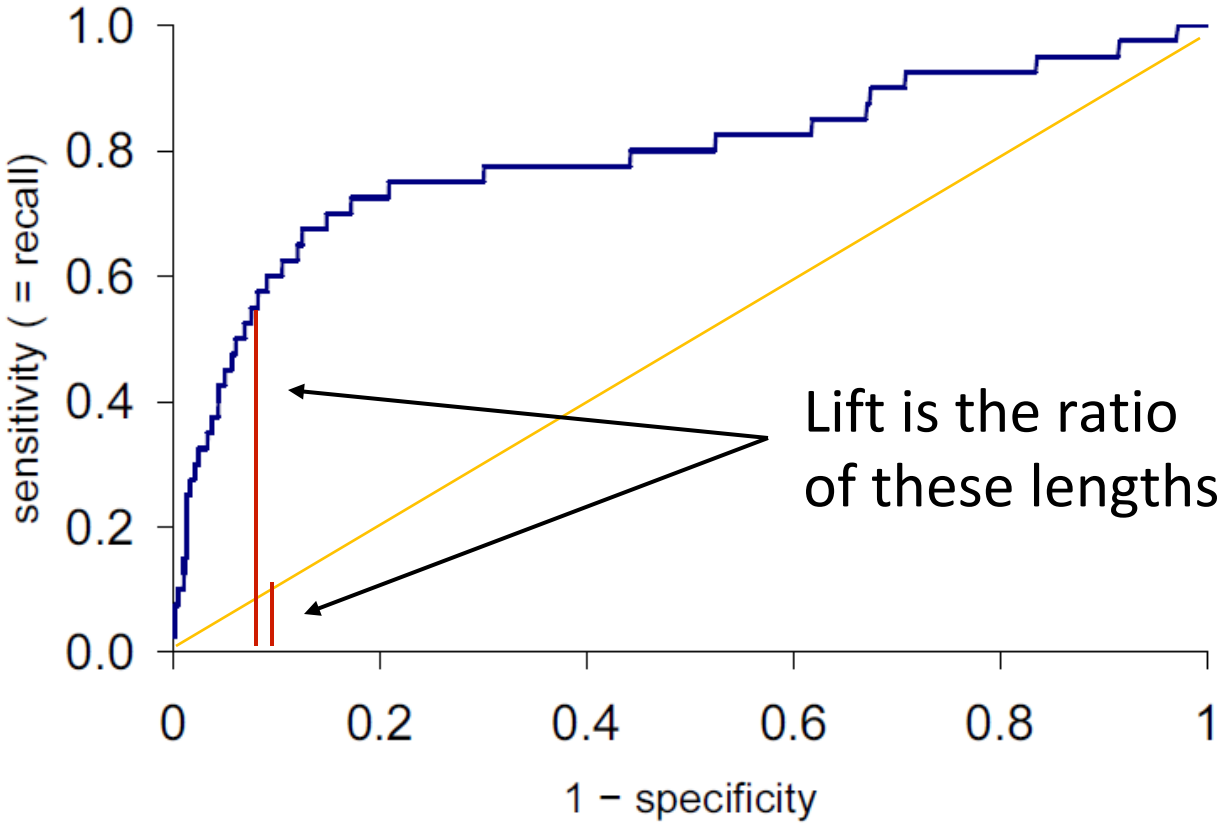
# ROC AUC

ROC AUC is the “Area Under the Curve” – a single number that captures the overall quality of the classifier. It should be between 0.5 (random classifier) and 1.0 (perfect).



# Lift Plot

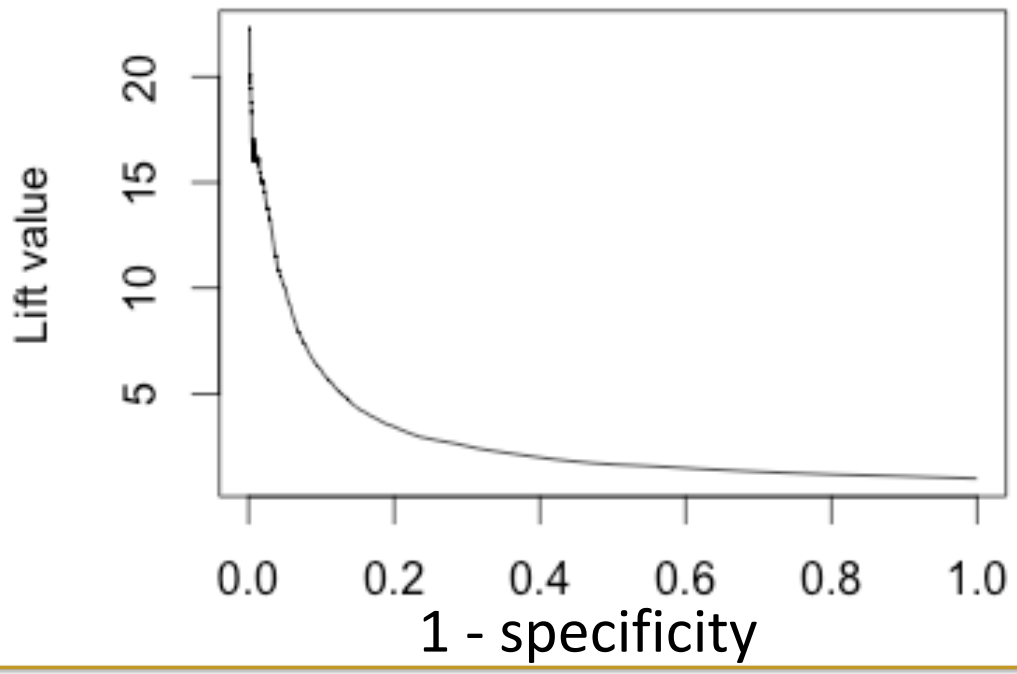
A variation of the ROC plot is the lift plot, which compares the performance of the actual classifier/search engine against random ordering, or sometimes against another classifier.



# Lift Plot

Lift plots emphasize initial precision (typically what you care about), and performance in a problem-independent way.

Note: The lift plot points should be computed at regular spacing, e.g. 1/00 or 1/1000. Otherwise the initial lift value can be excessively high, and unstable.



# Summary

- Statistics
  - Measurement
  - Hypothesis Testing
- Featurization
  - Feature selection
  - Feature Hashing
- Visualizing Accuracy