GITHUB; CLOUD COMPUTING; BASIC STATS

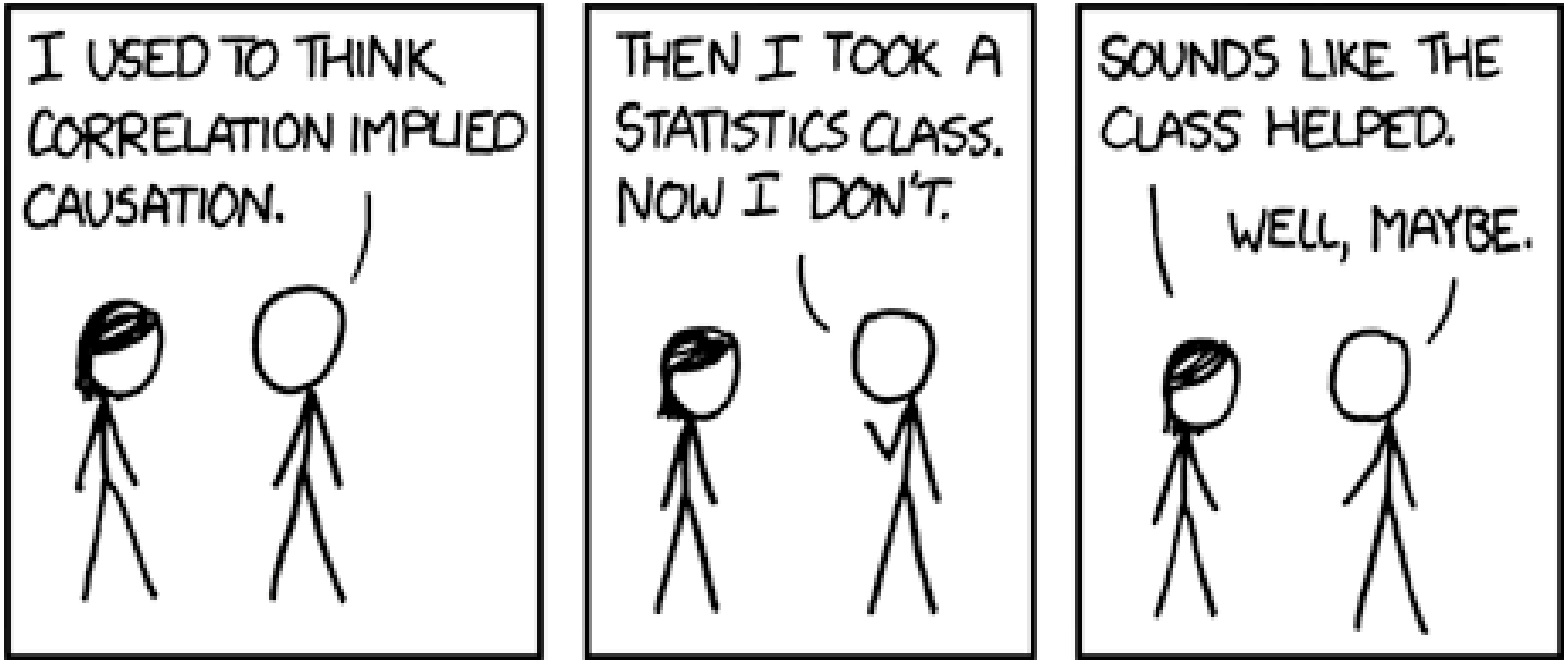
Prof. Mohammad Hajiaghayi & Dr. Arefeh Nasri
Wiki & Linkedin: @Mohammad Hajiaghayi
Twitter:@MTHajiaghayi
YouTube:@hajiaghayi [PLEASE SUBSCRIBE]
Instagram:@mhajiaghayi

DATA/MSML602: Lecture 5
Principles of Data Science
TuTh 6:00pm – 8:45pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

TODAY'S CLASS

- A Primer on Basic Statistics


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Biases
  - Causation
  - Misuse of Statistics

TERMS/DEFINITIONS

- Types of data:
  - Quantitative: Discrete vs Continuous
  - Categorical (e.g., Zipcodes): no inherent order among the values

- Populations vs Samples
  - Population: any set of objects or units under consideration
    - *N* typically represents the total number of observations in the population
  - Sample: a subset of the data
    - *n* typically used to represent the number of samples

- Statistics typically works with samples, and tries to infer things about the overall population

DESCRIPTIVE STATISTICS

- Statistics used to summarize or describe a set of observations
  - Rather than learning about the population that sample came from

- Central Tendency
  - Mean: numerical average of things
    - Is heavily influenced by outliers
  - Median: the middle value after ordering them
    - Considered the most representative
  - Mode: most frequently occurring value
    - Useful for categorical data, where mean and median are not meaningful

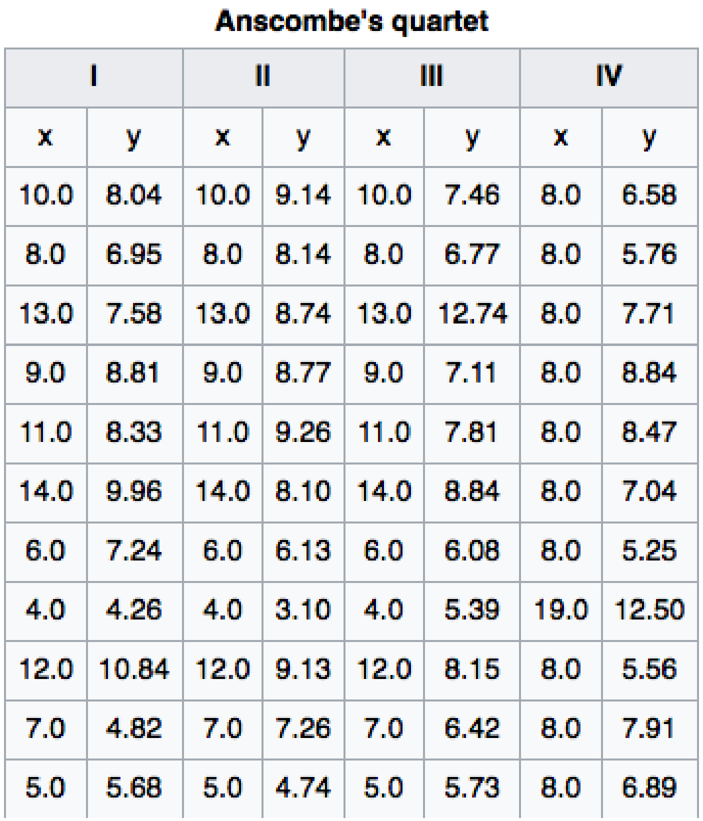
DESCRIPTIVE STATISTICS

- Statistics used to summarize or describe a set of observations
  - Rather than learning about the population that sample came from

- Dispersion or Variability
  - Range of values (min, max)
  - Standard Deviation:
    - Provides an estimate of the average difference of each value from the mean
    - Slightly different formulas for sample vs population
  - Kurtosis:
    - A measure of "tailedness"
  - Skewness
    - A measure of "asymmetry"
  - Different ways to quantify these

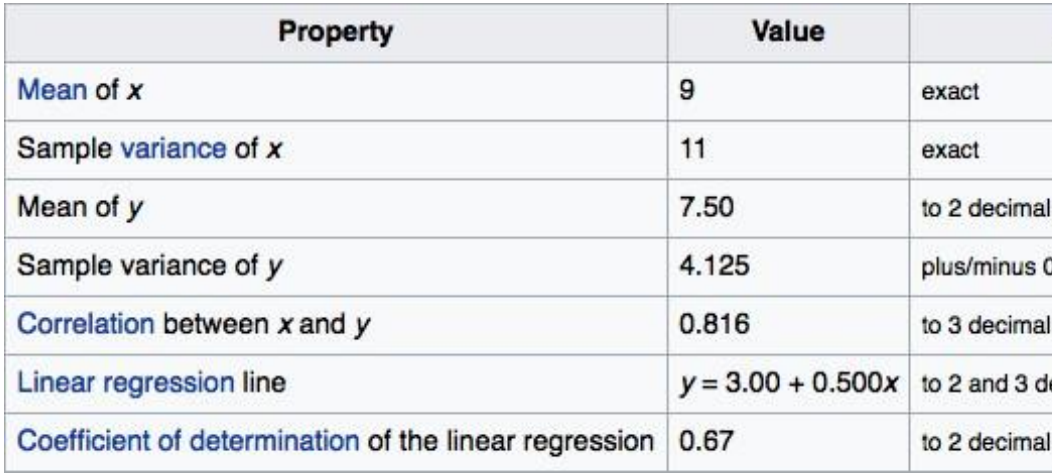
ANSECOMBE'S QUARTET

None of the descriptive statistics can distinguish between the four

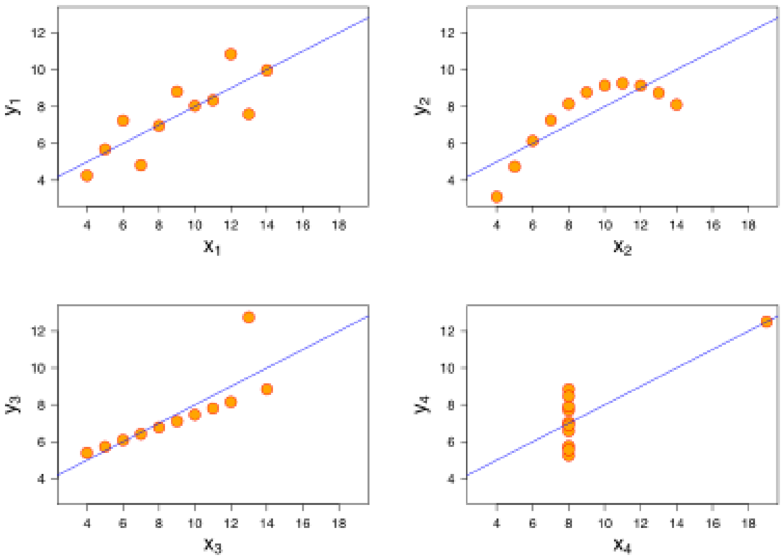


Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



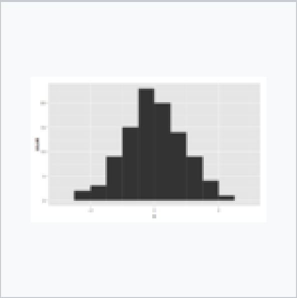
Property	Value	
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal
Sample variance of *y*	4.125	plus/minus 0
Correlation between *x* and *y*	0.816	to 3 decimal
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 d
Coefficient of determination of the linear regression	0.67	to 2 decimal

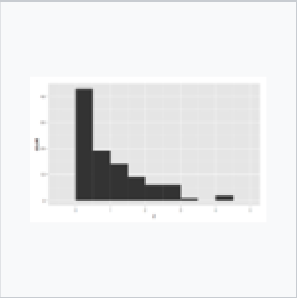


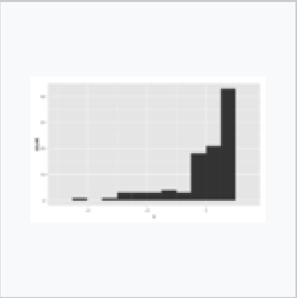
DESCRIPTIVE STATISTICS

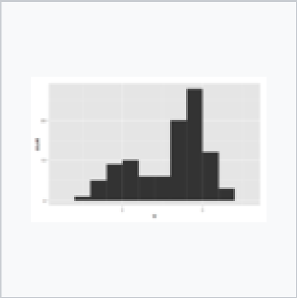
- Statistics used to summarize or describe a set of observations
  - Rather than learning about the population that sample came from

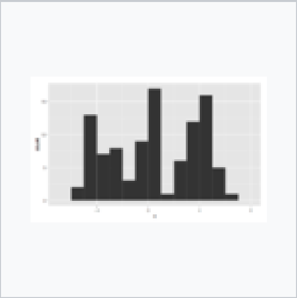
- Shape of Distributions
  - E.g., create buckets and plot number of values in each bucket (also called "histograms")

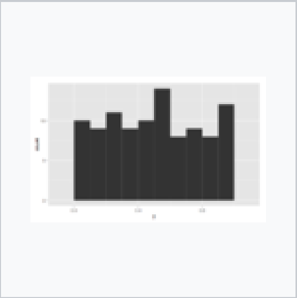

Symmetric, unimodal


Skewed right


Skewed left


Bimodal


Multimodal


Symmetric

  - Other types of plots to describe multi-variate data

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Biases
  - Causation
  - Misuse of Statistics

INFERENTIAL STATISTICS

- Process of deducing properties of an underlying probability distribution by analyzing sample

- Typically requires building a statistical model of the process that generated the data, and deducing different things like:
  - a point estimate;
  - or a confidence interval;
  - or rejection of a hypothesis;
  - or clustering or classification of data points into groups.

- E.g., using polling to predict who will win an election

- Requires some use of Probability Theory

CONFIDENCE INTERVALS

- *Point estimate:* a single statistics or parameter that we are trying to estimate for a population
  - e.g., mean of the population
- When we give a confidence interval [a, b] with confidence level 95%:
  - If we were to repeatedly draw samples from the population, 95% of the time, the true mean of the population will lie in the range computed from the sample
  - Confidence is in the method to compute the interval, not the interval itself
- Factors:
  - Higher confidence level --> bigger interval
  - Higher variability --> bigger interval
  - Higher number of samples --> smaller interval
    - Inverse square root relationship

NORMAL DISTRIBUTION

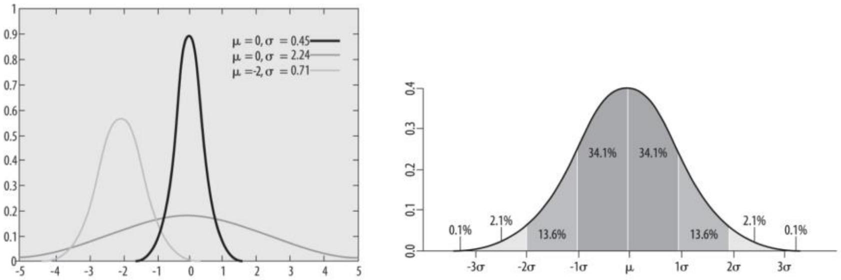
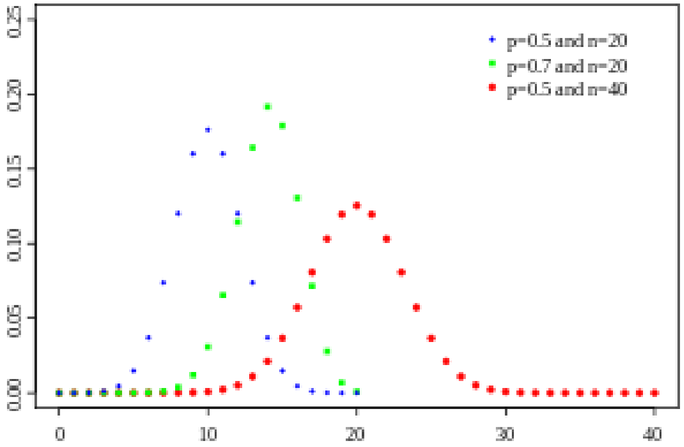


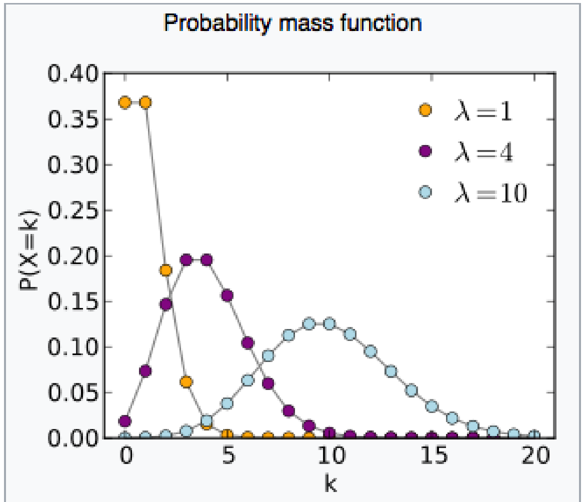
Figure 3.1: (left) All normal distributions have the same shape but differ to their μ and σ: they are shifted by μ and stretched by σ. (right) Percent of data failing into specified ranges of the normal distribution.

- 99.7% values will fall within 3 standard deviations (around the mean)
  - 95% for 2 standard deviations; 68% for 1
- **Central Limit Theorem**: As sample size approaches infinity, distribution of sample means will follow a normal distribution irrespective of the original distribution

Binomial distribution P(#heads) for a coin

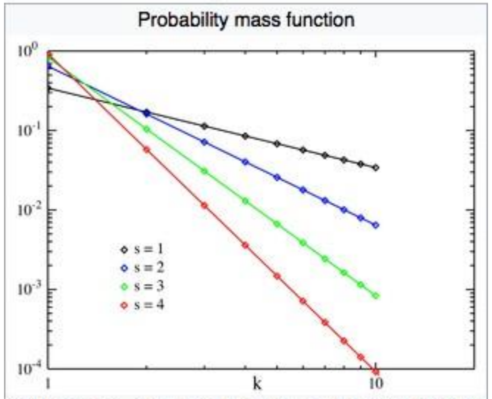


Poisson


Probability mass function

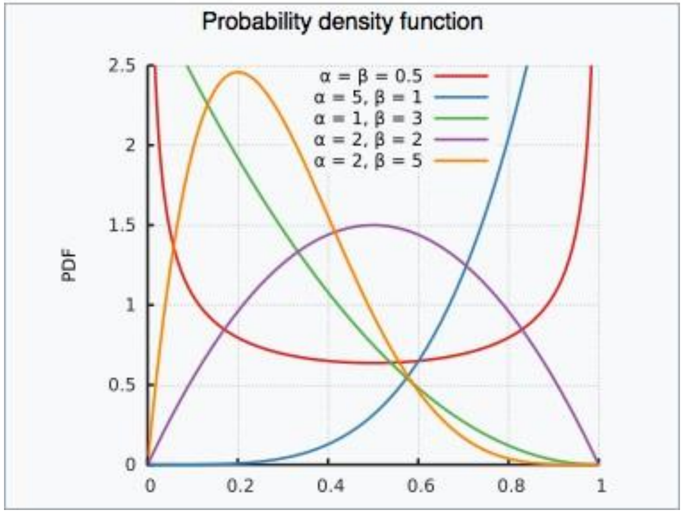
The horizontal axis is the index *k*, the number of occurrences. *λ* is the expected number of occurrences. The vertical axis is the probability of *k* occurrences given *λ*. The function is defined only at integer values of *k*. The connecting lines are only guides for the eye.

Zipf's law


Probability mass function

Zipf PMF for *N* = 10 on a log–log scale. The horizontal axis is the index *k* . (Note that the function is only defined at integer values of *k*. The connecting lines do not indicate continuity.)

Beta


Probability density function

List of probability distributions

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Hypothesis Testing
  - Biases
  - Causation
  - Misuse of Statistics

HYPOTHESIS TESTING

- Accepting or rejecting a statistical hypothesis about a population

- H_0: *null hypothesis*, and H_1: the *alternative hypothesis*
  - Mutually exclusive and exhaustive
  - H_0 can never be proven to be true, but can be rejected
- Statistical significance: probability that the result is not due to chance
- Example: Deciding if a coin is fair
  - http://20bits.com/article/hypothesis-testing-the-basics

HYPOTHESIS TESTING

- Process:
  - Decide on H_0 and H_1
  - Decide which *test statistic* is appropriate
    - Key question: what is the distribution of the test statistic over samples?
  - Select a significance level (\sigma), a probability threshold below which the null hypothesis will be rejected -- typically 5% or 1%.
  - Compute the observed value of the test statistic t_obs from the sample
  - Compute **p-value**: the probability that the test statistic took that value by chance
    - Use the distribution above to compute the *p-value*
  - Reject the null hypothesis if the *p-value* < *\sigma*

HYPOTHESIS TESTING

- Type 1 Error: Rejected null hypothesis by mistake
- Type 2 Error: Accepted null hypothesis by mistake
- Common Test Statistics
  - One-sample Tests: Appropriate when comparing a sample to the underlying population
  - Two-sample Tests: When comparing two samples, e.g., experimental and control
- Wikipedia article has a lot more details on different types of tests
  - Along with history of hypothesis testing, debates about it, criticisms, etc.

SCIENTIFIC METHOD: STATISTICAL ERRORS

- Nature Article
- P values not as reliable as many scientists assume
- Issues:
  - A p-value of 0.01 does not mean 99% probability of the hypothesis being true -- in fact the probability of false alarm may be 11% or higher
  - p-hacking: cherry picking data points etc., to get the p-values; repeating experiments if they fail till you get the result
- Much discussion/debate about this issue in recent years

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Hypothesis Testing
  - Biases
  - Causation
  - Misuse of Statistics

SAMPLING BIASES

- Sampling effective at reducing the data you need to analyze
- Ideally you want **random** sample
  - Otherwise you need to account for bias, which can be tricky
- Bias in sampling: need to be very careful when generalizing inferences drawn from a sample
  - Even for random samples

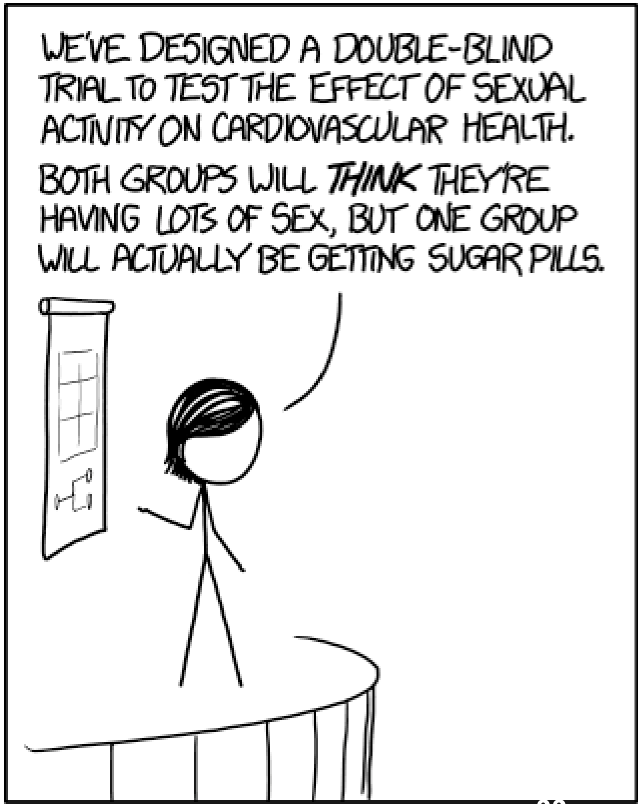
- Questions to ask: How was the sample selected? Was it truly random? Potential biases? How were questions worded? How is missing data/attrition handled? Was the sample size large enough?

SOME POTENTIAL SOURCES OF BIASES

- Sample Bias
  - Selection bias: some subjects more likely to be selected
  - Volunteer bias: people who volunteer are not representative
  - Nonresponse bias: people who decline to be interviewed
- Survey/Response Bias
  - Interviewer bias
  - Acquiescence bias
  - Social desirability bias: people are not going to admit to embarrassing things
- Also watch out for:
  - Confirmation bias
  - Anchor bias

SOME POTENTIAL SOURCES OF BIASES

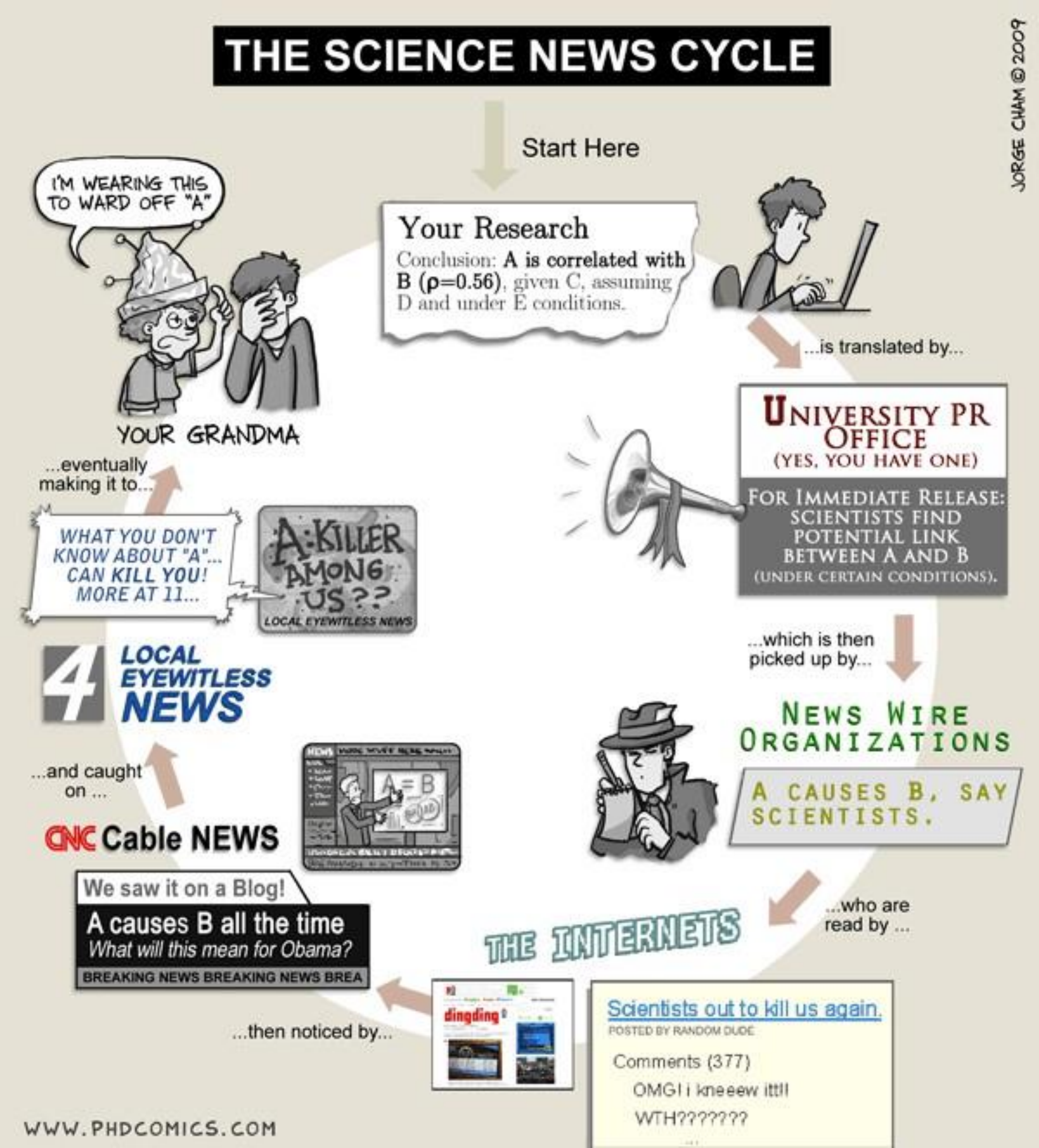
- Gold Standard: Randomized Clinical Trials
  - Some people receive "treatment", others in a "control" group
  - Picked randomly to take care of all confounding factors
  - Problems:
    - Ethically feasible only if **clinically equipoise**
      - Can't ask some people to smoke to figure out the effects of smoking
    - Very expensive and cumbersome
    - Impossible in many cases
- Recall: Recent Facebook experiment on emotions


WE'VE DESIGNED A DOUBLE-BLIND TRIAL TO TEST THE EFFECT OF SEXUAL ACTIVITY ON CARDIOVASCULAR HEALTH.
BOTH GROUPS WILL *THINK* THEY'RE HAVING LOTS OF SEX, BUT ONE GROUP WILL ACTUALLY BE GETTING SUGAR PILLS.

THE LIMITATIONS OF BLIND TRIALS

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Hypothesis Testing
  - Biases
  - Causation
  - Misuse of Statistics


THE SCIENCE NEWS CYCLE
JORGE CHAM © 2009
Start Here
Your Research
Conclusion: A is correlated with B (ρ=0.56), given C, assuming D and under E conditions.
...is translated by...
UNIVERSITY PR OFFICE
(YES, YOU HAVE ONE)
FOR IMMEDIATE RELEASE: SCIENTISTS FIND POTENTIAL LINK BETWEEN A AND B (UNDER CERTAIN CONDITIONS).
...which is then picked up by...
NEWS WIRE ORGANIZATIONS
A CAUSES B, SAY SCIENTISTS.
...who are read by ...
THE INTERNETS
dingding
Scientists out to kill us again.
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!!
WTH???????
...then noticed by...
We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA
CNC Cable NEWS
A=B
...and caught on ...
4 LOCAL EYEWITLESS NEWS
WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...
A: KILLER AMONG US??
LOCAL EYEWITLESS NEWS
...eventually making it to...
YOUR GRANDMA
I'M WEARING THIS TO WARD OFF "A"
WWW.PHDCOMICS.COM

DETERMINING CAUSATION

- Bradford Hill's Criteria: widely accepted in the modern era as useful guidelines for investigating causality in epidemiological studies
  - Strength: how large is the association
  - Consistency across different samples
  - How specific
  - Cause should precede effect (temporality)
  - Biological gradient (increase dose → increase association)
  - Plausibility
  - Coherence
  - Experiment
  - Consideration of alternate explanations

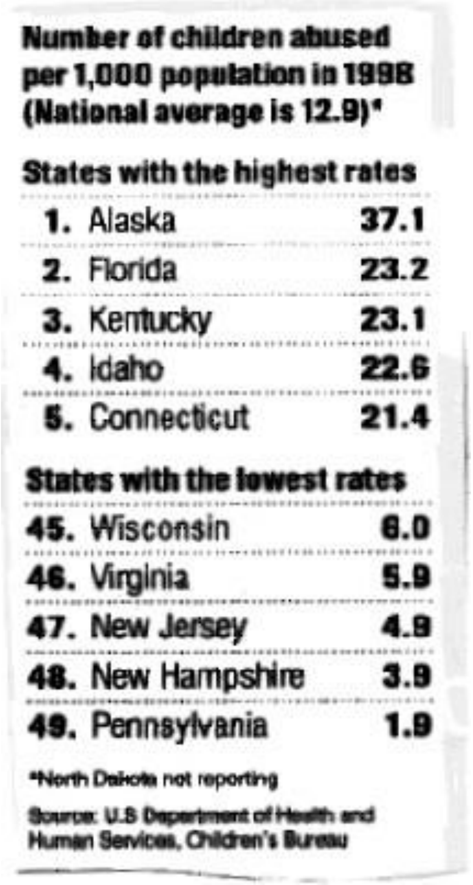
Article to explain the Criteria

TODAY'S CLASS

- A Primer on Basic Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Hypothesis Testing
  - Biases
  - Causation
  - Misuse of Statistics

MISUSE OF STATISTICS

- This famous, but old book on statistics goes into detail about How to lie with statistics



Number of children abused per 1,000 population in 1998 (National average is 12.9)*

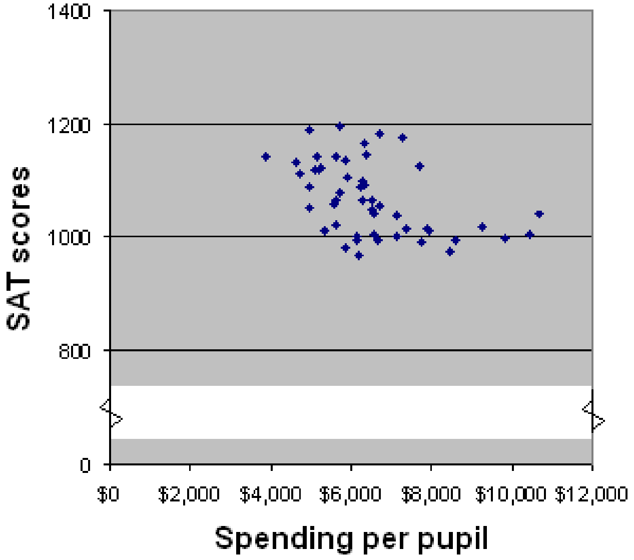
States with the highest rates		
1.	Alaska	37.1
2.	Florida	23.2
3.	Kentucky	23.1
4.	Idaho	22.6
5.	Connecticut	21.4

States with the lowest rates		
45.	Wisconsin	6.0
46.	Virginia	5.9
47.	New Jersey	4.9
48.	New Hampshire	3.9
49.	Pennsylvania	1.9

*North Dakota not reporting

Source: U.S Department of Health and Human Services, Children's Bureau

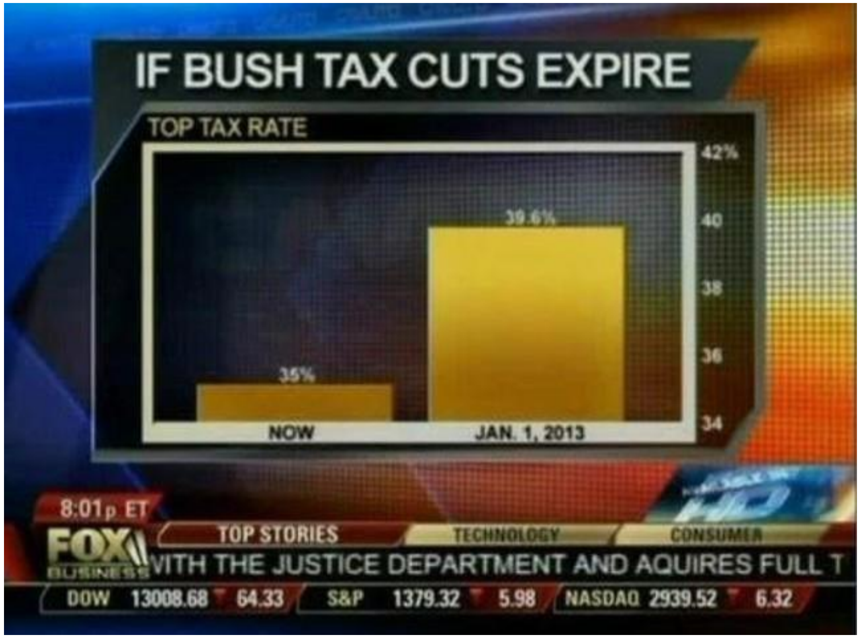
Spending per Pupil and SAT Scores by State, 1998

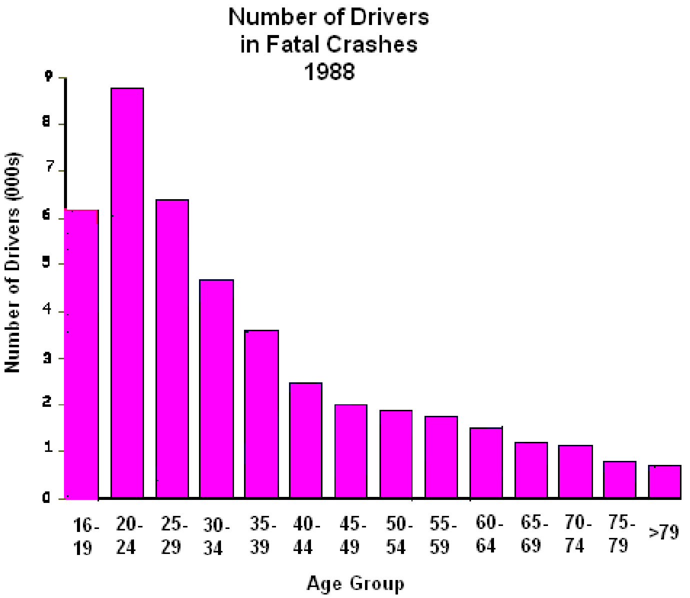


SAT Scores, 1998

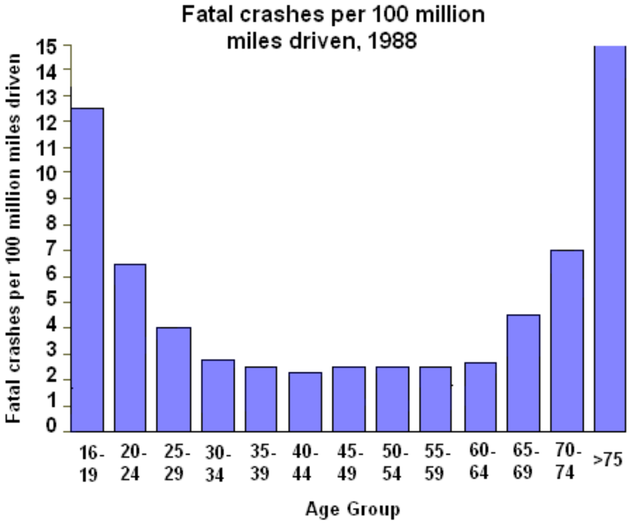
State	Verbal	Math	Participation Rate
North Dakota	590	599	5%
New Jersey	497	508	79%

BEWARE OF CHARTS !



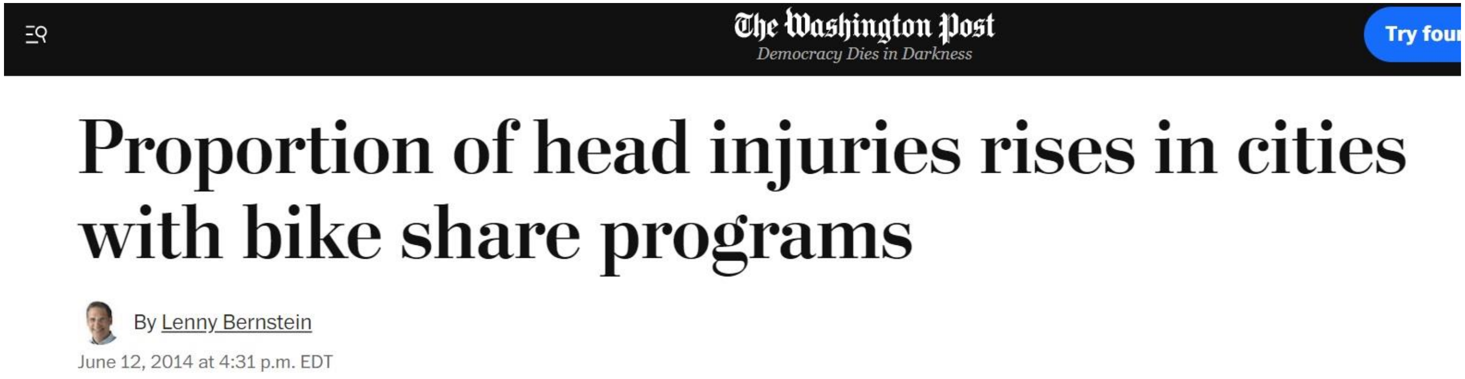


Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.



Graph is based on data from this study: Williams, Allan F., Pl.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

NEWSPAPERS EVEN MORE



Proportion of head injuries rises in cities with bike share programs

By Lenny Bernstein

June 12, 2014 at 4:31 p.m. EDT

- Source
- A Washington Post article says: *In the first study of its kind, researchers from Washington State University and elsewhere found a 14 percent greater risk of head injuries to cyclists associated with cities that have bike share programs. In fact, when they compared raw head injury data for cyclists in five cities before and after they added bike share programs, the researchers found a 7.8 percent increase in the number of head injuries to cyclists.*
- Actually: head injuries declined from 319 to 273, and overall injuries declined from 757 to 545
  - So the **proportion** of head injuries went up !!

DIGGING UP SOURCES CAN BE TRICKY

- Many claims, even well-known ones, don't stand up to scrutiny

- *6 feminist myths that will not die…*
  - *In the United States, 22%–35% of women who visit hospital emergency rooms do so because of domestic violence.*
  - E.g. wage gap between men and women (*"Women earn 77 cents for every dollar a man earns—for doing the same work."*)

- More than 80% of Dentists recommend Colgate (Colgate was stop saying that in UK)

- Many health-related statements are based on unclear data

WRAP-UP

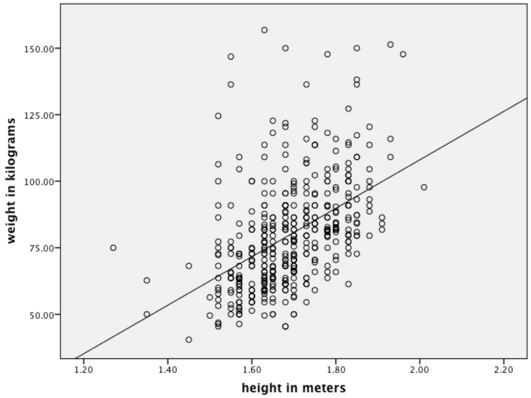
- A Primer on Basic Statistics
- Will post some resources to follow up
- Quiz: Find as many sources of bias in the survey as you can!
- Next: A high-level overview of Statistical Techniques and ML

OVERVIEW

- Statistical models vs machine learning algorithms
  - Used somewhat interchangeably
- Some key terms:
  - **Supervised vs Unsupervised Learning**
    - In former, given additional attributes that we want to predict/classify
      - **Classification**: e.g., handwritten digit recognition
      - **Regression**: e.g., predicting stock prices
    - In latter, the goal is to deduce structure in the given *unlabeled* data
      - **Clustering**: e.g., group customers based on their buying patterns
      - **Association rule mining**: e.g., if you buy *A*, you are likely to buy *B*
      - **Dimensionality reduction**: project the data onto a few dimensions to make it simpler to visualize/classify etc.
    - Semi-supervised Learning: a combination of labelled and unlabelled data

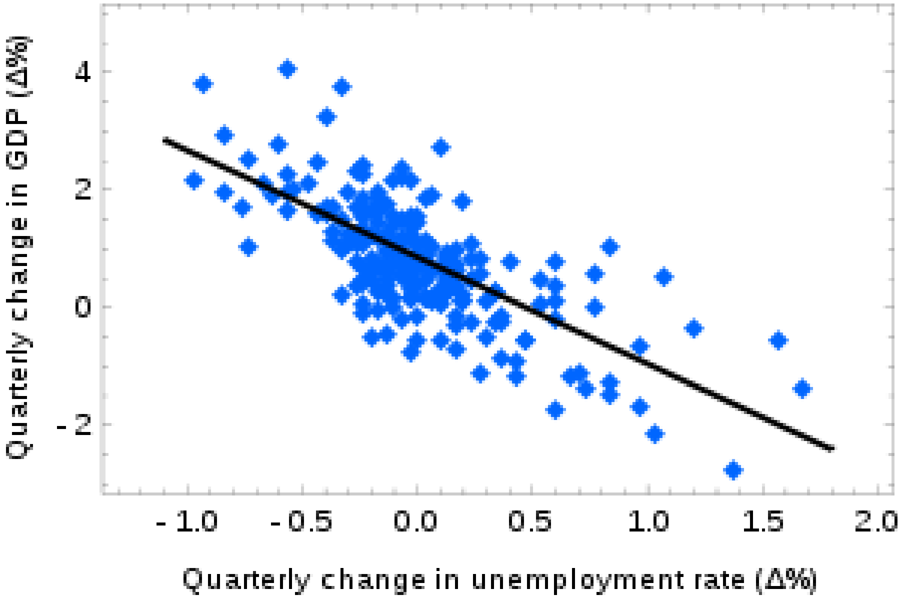
OVERVIEW

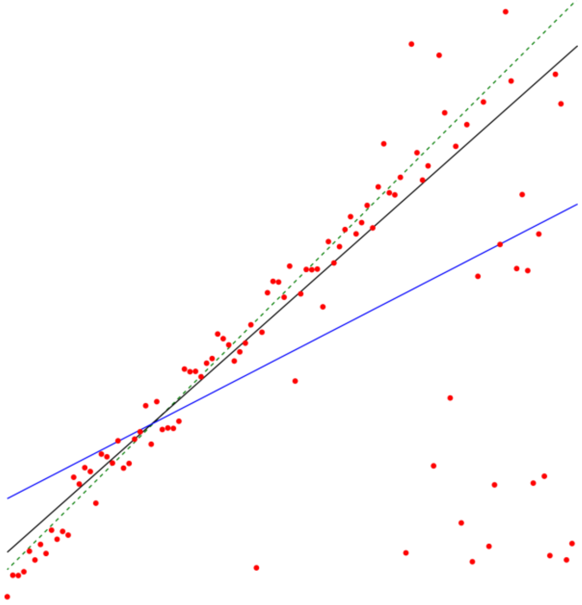
- Some key terms:
  - **Training vs testing datasets**
    - Use the former to learn, the latter to test the learned model
  - **Feature Selection**:
    - Selecting a subset of all possible features to use
    - Most data contain too many features, most of which are redundant
  - **Regularization**:
    - Simplifying solutions produced by a technique by penalizing complexity
  - **Overfitting**:
    - Learned model starts capturing the random error or noise instead of the underlying relationship
    - Typically happens when model is too complex (e.g., too many parameters given the data)
- A key challenge for a data scientist:
  - Choosing which method/algorithm to use for a specific task



LINEAR REGRESSION

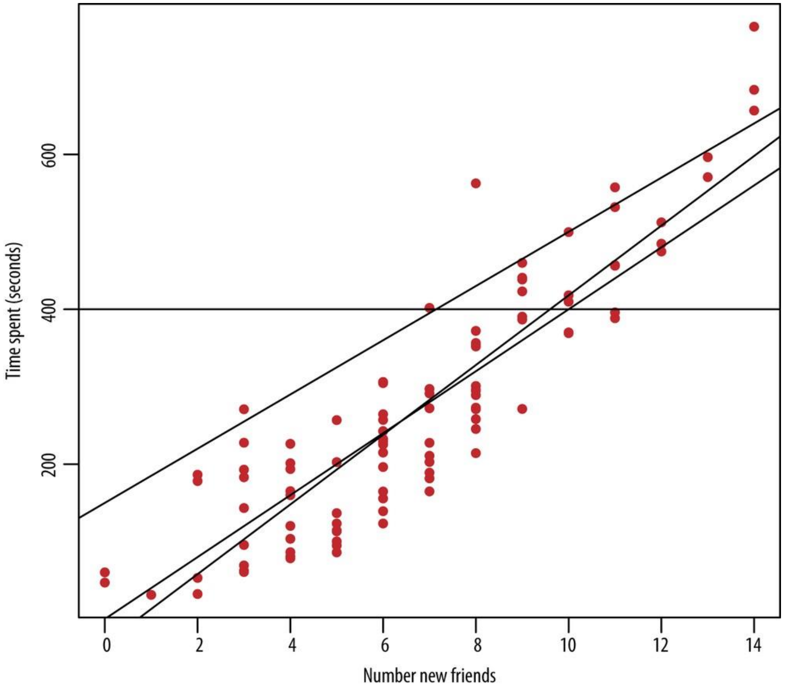
- Goal: to predict the value of a *dependent* (or response, or outcome) variable using one or more *independent* (or explanatory) variables
- Captures the relationship using a linear equation

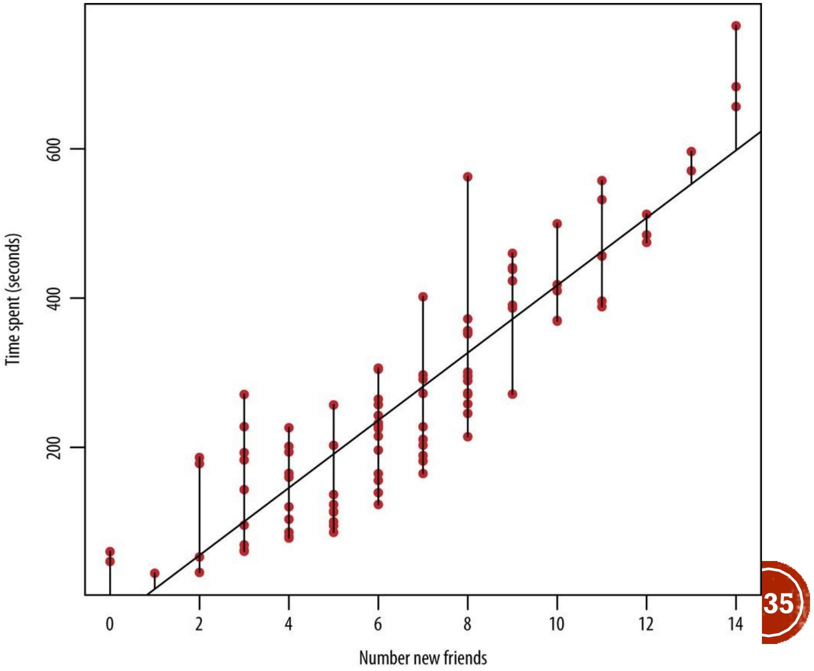




LINEAR REGRESSION II

- How to find the best fit?
  - Ordinary least squares: minimize the squared distance from the line
  - Ridge regression: add a penalty term for the sizes of the coefficients





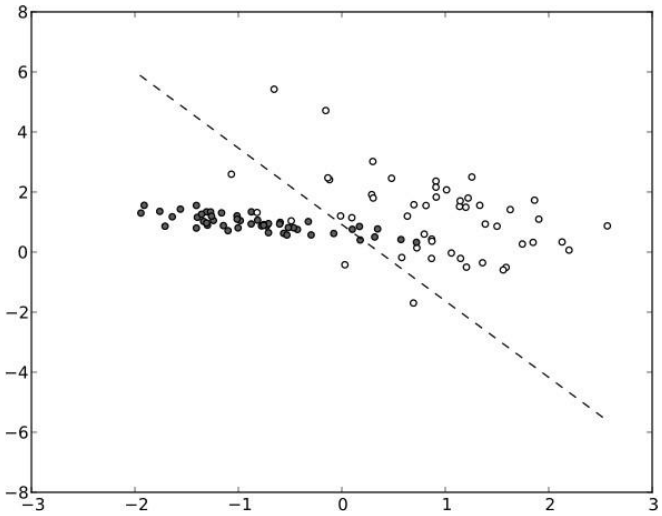
LINEAR REGRESSION III

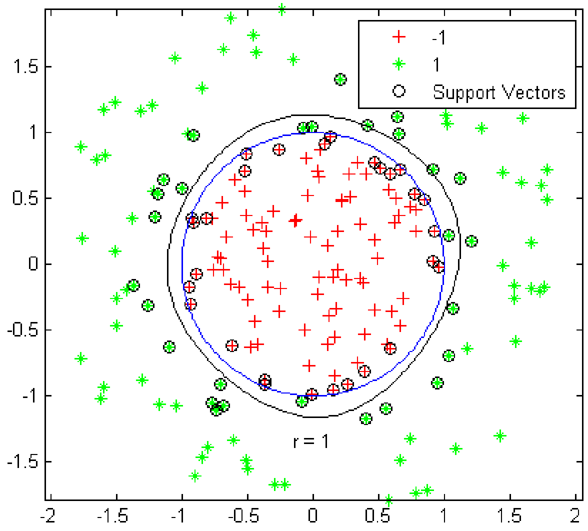
- Transformations
  - The *linearity* assumption is only about the parameters
  - We can take arbitrary transformations of the independent variables themselves
  - The following is a linear regression
    - r is the response variable, x, y, z, are the independent variables
    - r = c_1 x^2 + c_2 x + c_3 + c_4 y^3 + c_5 log(z) + c_6 x * y + c_7 x * y^2 * log(z)
- Assumption of independence
  - Things break down if the predictor variables are not independent, or if the errors are not independent

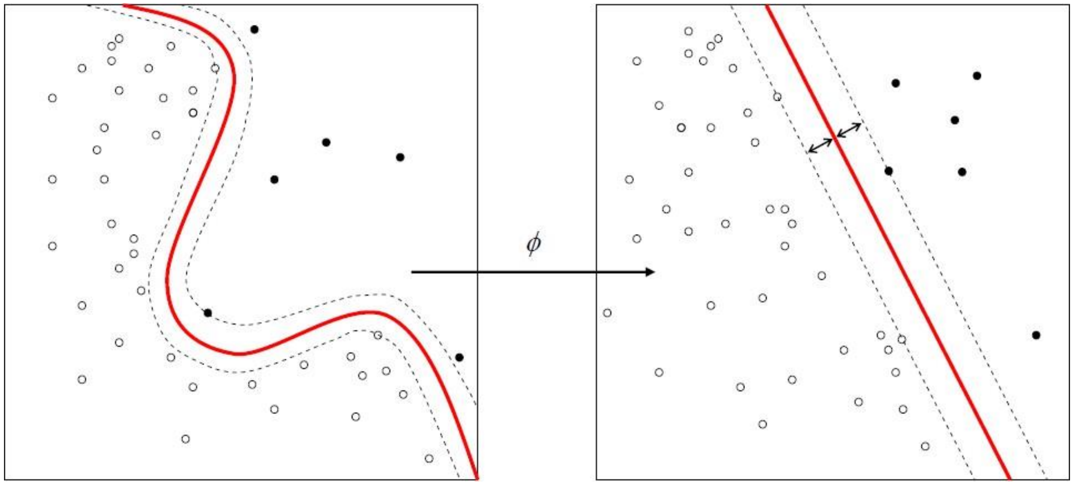
CLASSIFICATION I

- One of the most common ML tasks
- Examples: spam vs non-spam emails; diagnostics; high risk vs low risk customers (for loans); …
- Typical problem setting:
  - Given a set of features (also called independent variables), and a set of samples (training data) along with labels (dependent variable) for those samples,
  - Learn a model to use for future data points
  - Labels are categorical, not numerical
- For numerical features, visualizing can often suggest how to classify

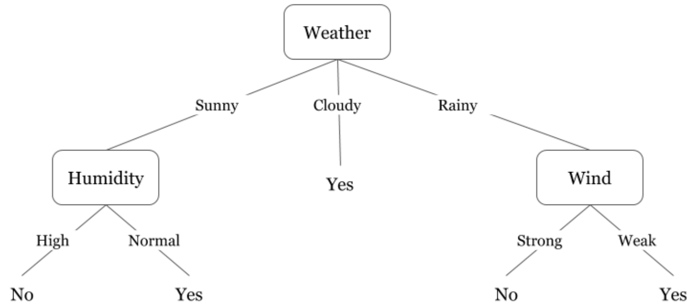
CLASSIFICATION II







CLASSIFICATION: DECISION TREES



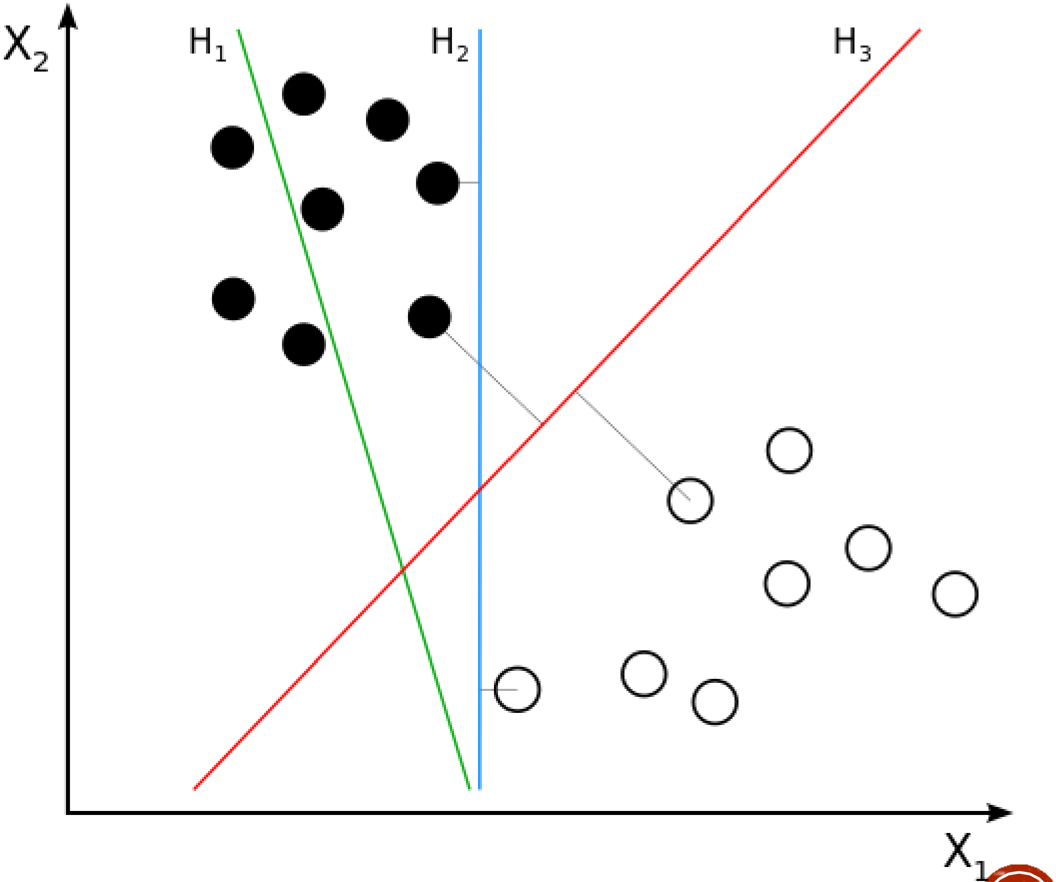
- Very intuitive and easy-to-use classification models
  - Works for categorical and numerical features
- Many techniques developed for learning them over large volumes of data
- Key advantage: **Interpretability**
- Key disadvantage: Not rich enough to give good fits
- Random Forest
  - Basically a collection of Decision Trees

CLASSIFICATION: LOGISTIC REGRESSION

- An example of a classification model, not a regression model
- Logistic function:
  - $F(t) = 1/ (1 + e^{-t})$
  - Always between 0 and 1
  - Can be interpreted as a probability
- Say: features are: x, y, and response variable is r which takes two values (true and false)
- Consider the formula:
  - $F = 1 / (1 + e^{-(c\_1\ x + c\_2\ y + c\_3)})$
  - Treat it as a probability that the response variable is true or false
- Learning problem: from training data, estimate the coefficients c_i's

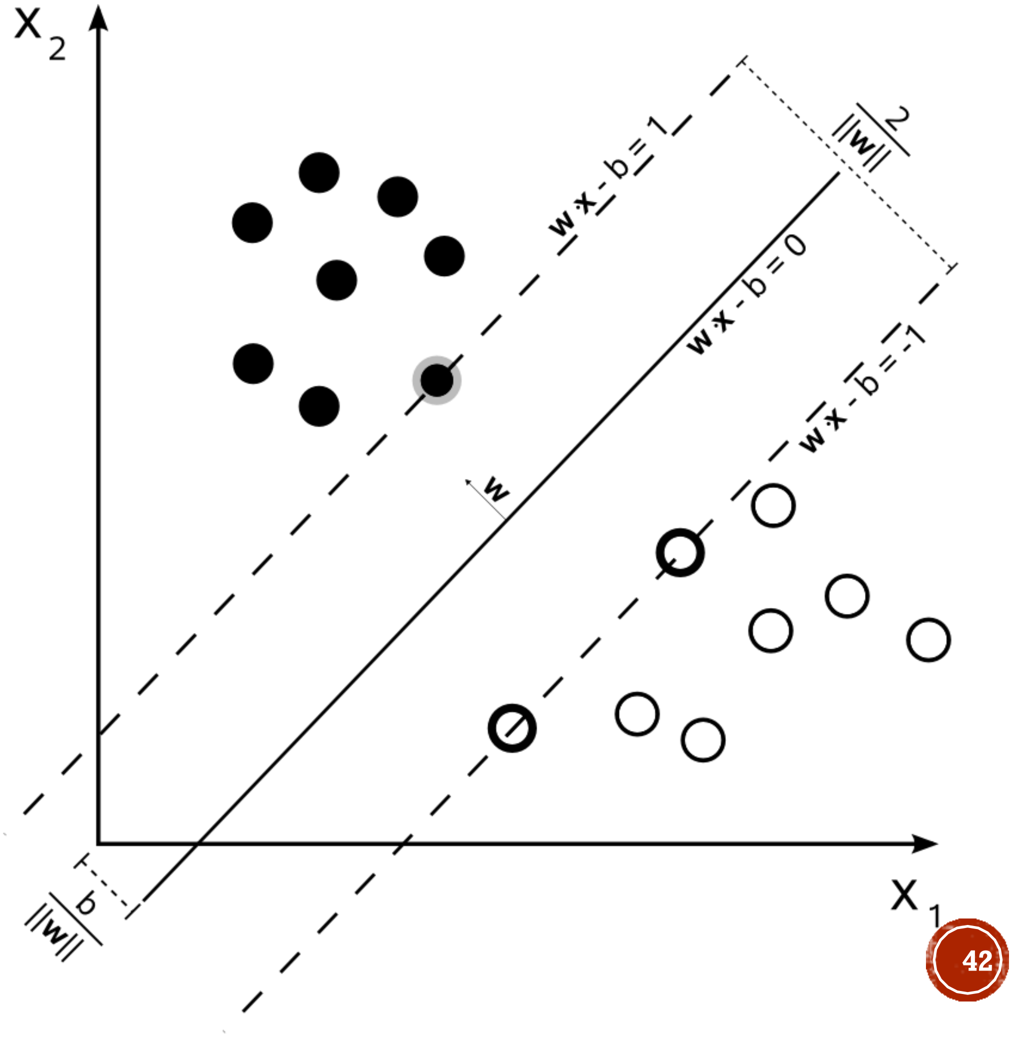
CLASSIFICATION: SUPPORT VECTOR MACHINES

- One of the most popular classification techniques
  - Very powerful, and flexible
  - Also called max-margin classifier
- Maximally separate the two classes with a hyperplane
  - In the example, H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.



CLASSIFICATION: SUPPORT VECTOR MACHINES

- Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.



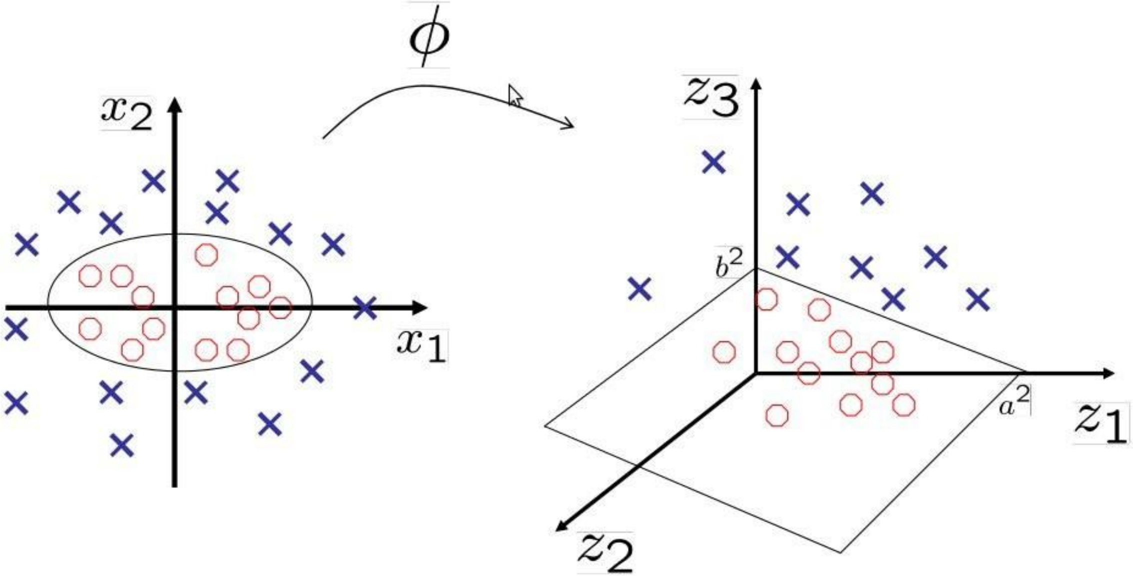
X_2
X_1
$\mathbf{w}\cdot\mathbf{x} - b = 1$
$\mathbf{w}\cdot\mathbf{x} - b = 0$
$\mathbf{w}\cdot\mathbf{x} - b = -1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}$
$\frac{b}{\|\mathbf{w}\|}$

CLASSIFICATION: SUPPORT VECTOR MACHINES

- What if the two classes are not linearly separable ??
- Soft-margin SVMs:
  - Allow for mislabeled samples, i.e., samples that lie on the wrong side of the hyperplane
  - But penalize them
  - **New optimization goal**: Find a hyperplane with high margin but with only a few mislabelled examples
- Finding the optimal separating hyperplane (with a soft-margin one) is quite efficient
  - Between O(n^2) and O(n^3), where n is the number of training samples
  - Work well for high-dimensional data as well

CLASSIFICATION: SVM KERNEL TRICK

- Can be used to construct non-linear classifiers
  - An ellipse in the original 2D space becomes a hyperplane in the 3D transformed space
    - i.e., we can learn a linear classifier in the 3D space, and transform it back

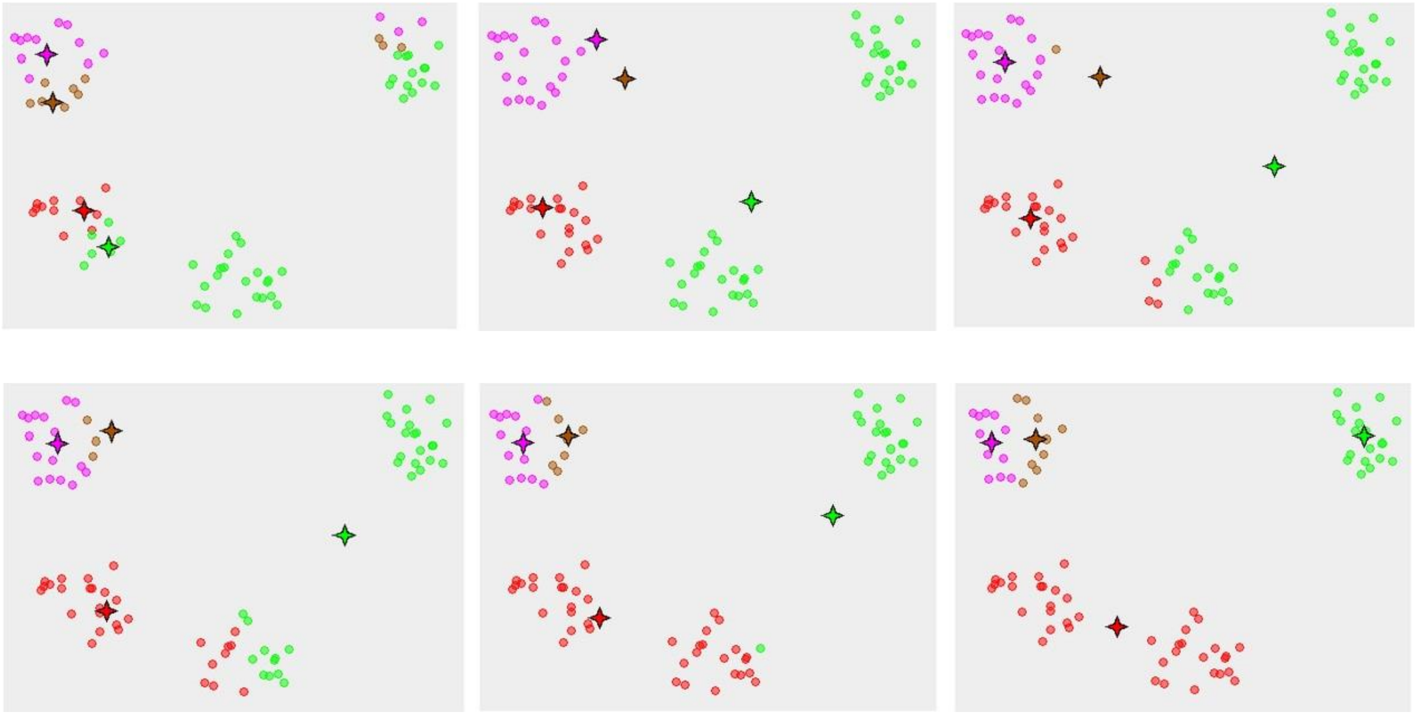


$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \longrightarrow \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

CLUSTERING

- Goal is to group the samples into clusters so that the objects in a cluster are more similar to each other than to objects outside
  - Somewhat vague definition -- different ways to formalize it
- *K-means* Clustering:
  - A heuristic algorithm that iteratively adjusts centroids till it converges

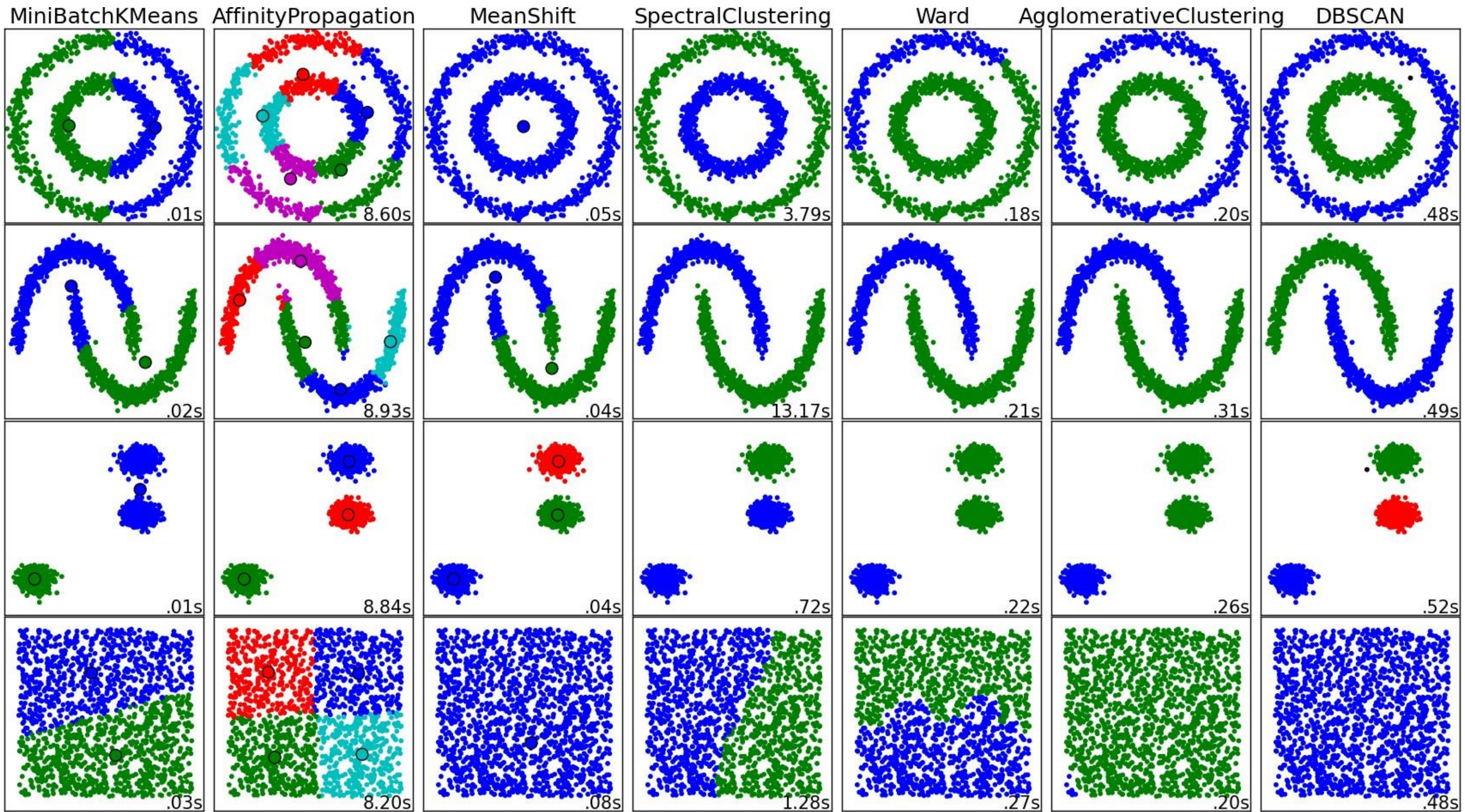


CLUSTERING

- Hierarchical Clustering
  - Build nested clusters by merging (bottom-up) or splitting (top-down)
  - Agglomerative Clustering: a bottom-up strategy
    - Start with each data point in its own cluster
    - Merge the closest two clusters
    - Repeat till you are left with one cluster
- DBSCAN
  - Clusters are high-density areas separated by low-density areas
  - So clusters can be of any shape (unlike K-Means)

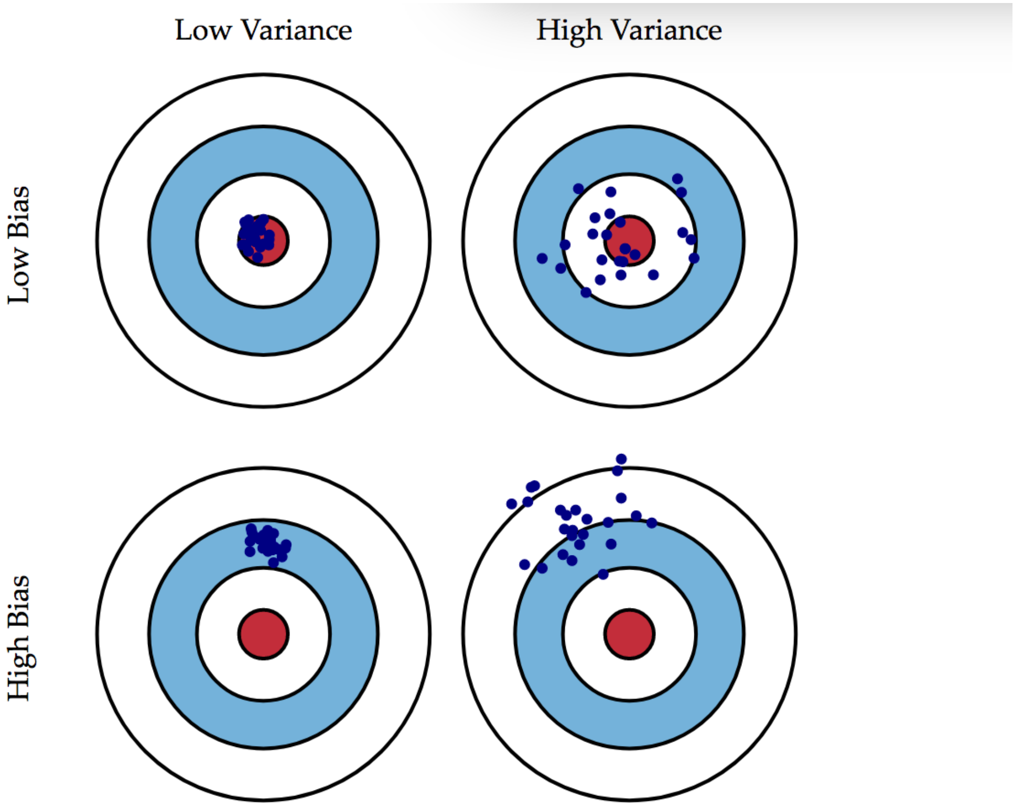
CLUSTERING

- From Sci-Kit Guide:



BIAS-VARIANCE TRADEOFF

- Balancing between:
  - Under-fitting due to an biased or impoverished model
  - Overfitting due to capturing noise



http://scott.fortmann-roe.com/docs/BiasVariance.html

SUMMARY

- For each of the tasks, many other techniques
  - Neural networks or k-nearest neighbors for classification

- Need to choose the right technique and parameters for a specific scenario

- Bias-variance Tradeoff (Dilemma)