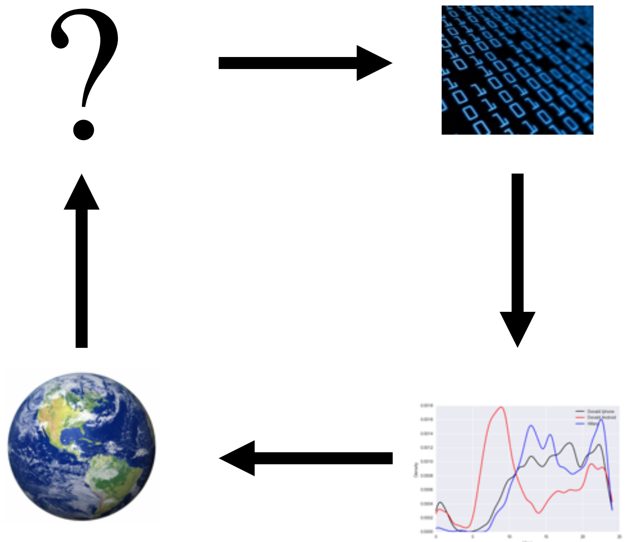
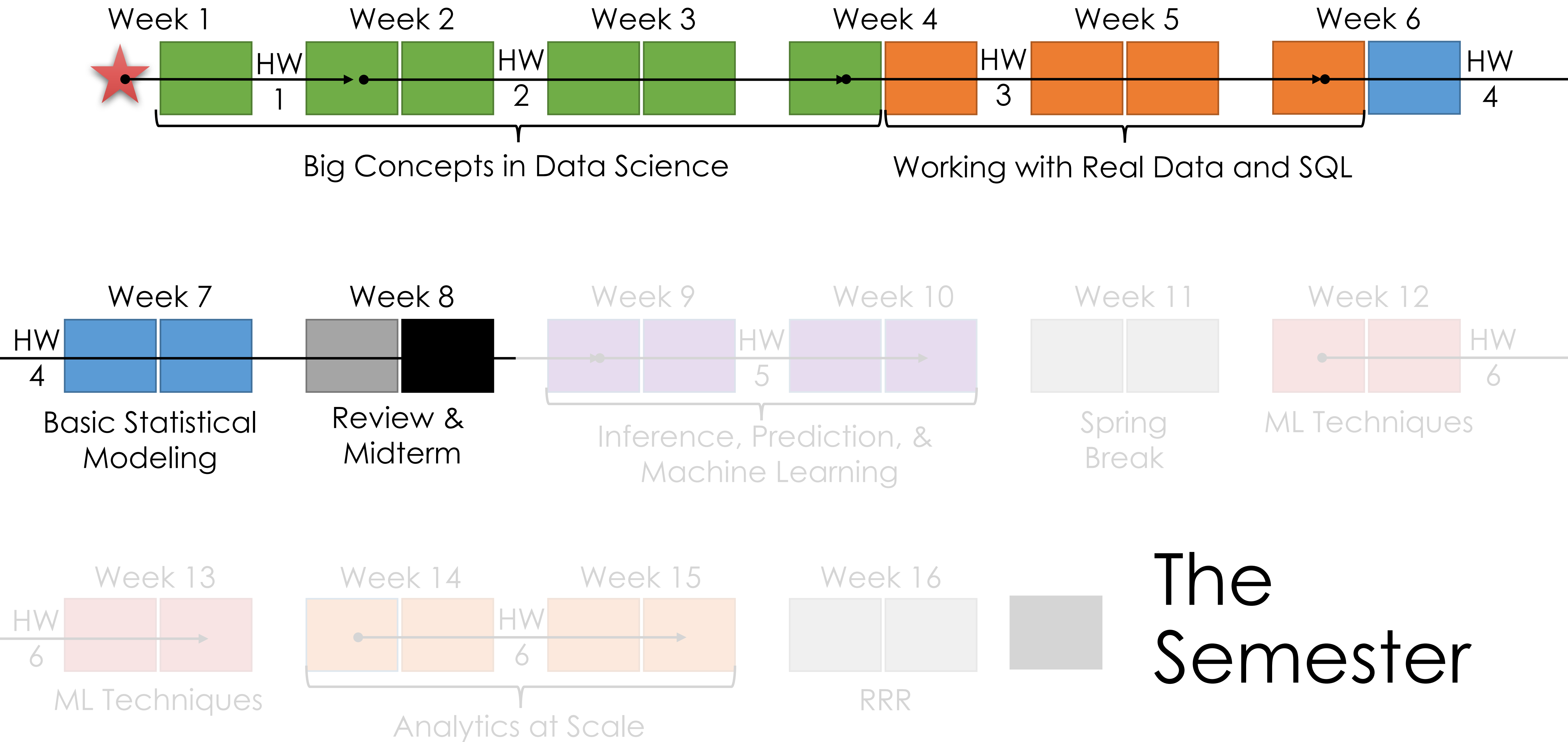
Midterm Review

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu





The Semester

http://www.ds100.org/sp17/syllabus

Introduction

Defining Data Science

The application of **data centric**, **computational**, and **inferential thinking** to

understand the world		*solve problems*
Science	**&**	**Engineering**

➢ *Data science is fundamentally <u>interdisciplinary</u>*

Reality of Data Science Today

- Data is often not that big
- Substantial time spent in data cleaning and exploration
  - Less time spent developing new models
- Wide range of tools: *SQL, R, Python, …*
- Data science workflow is iterative (the lifecycle)
- Discussed some ethical concerns of Data Science
- Explored Food Safety data (not covered on exam)

Question Formulation

Introduced QPR-V

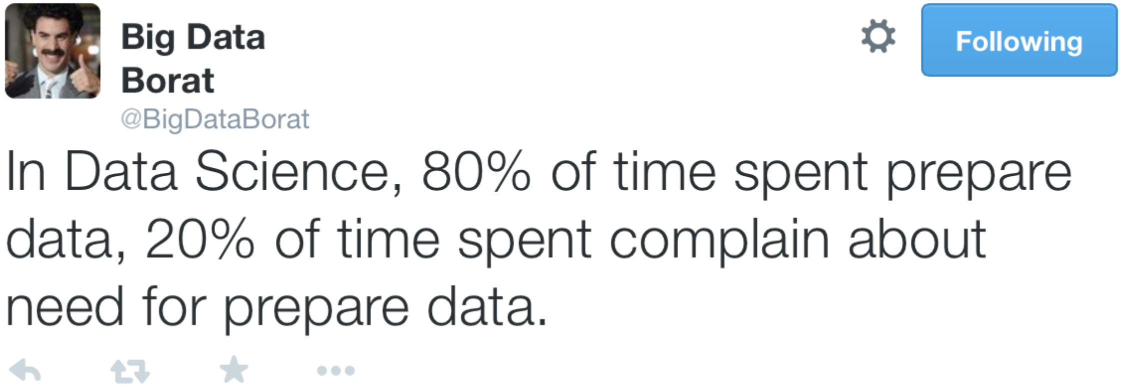
- **Q**uestion: *construct a well formed question*
  - If I study will I do well on the exam → If I review X material will I get a grade that is above average.
- **P**opulation: *identify the population in the question*
  - **Who** or **what** are we studying …
- **R**epresentative: *do the data reflect the population*
  - *Before collecting or analyzing the data*
  - *Depends on the collection process*
- **V**alidation:
  - verify conclusions through statistical inference and assess reproducibility

Data Collection and Sampling

- **Census:** the *complete population*
- **Survey:** *a sample of the population*
- **Observational Studies**: *data collected without direct intervention*
- **Randomization:** *mechanism to control for external factors*
  - **Simple random samples:** drawing data from the population uniformly at random
  - **Randomized Trial:** randomly assign subjects to treatment and control groups
    - gold standard in causal analysis

Data Wrangling

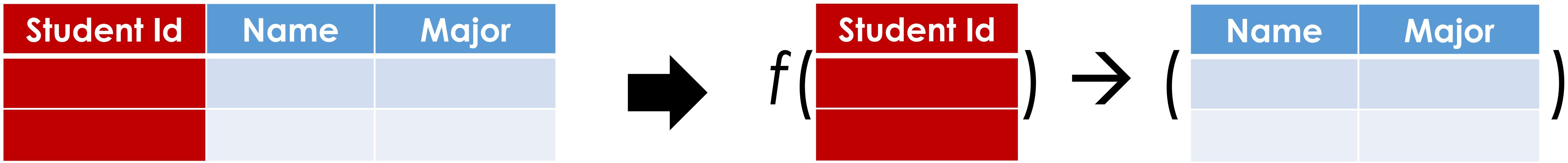
Data wrangling is the process of cleaning and transforming data to enable subsequent analysis.



- **Structure**: *the "shape" of a data file*
- **Granularity**: *how fine/coarse is each datum*
- **Faithfulness**: *how well does the data capture "reality"*
- **Temporality**: *how is the data situated in time*
- **Scope**: *how (in)complete is the data*

Primary Keys and Functional Dependencies

- A **primary key** is set of columns that uniquely identify each row. (e.g., *student_id*)
  - Can be composite: e.g. *(City, State)*
  - Each value occurs at most once in the key column(s).
- Functional Dependencies:



- Not just keys: (e.g., zipcode → State)

Types of Data

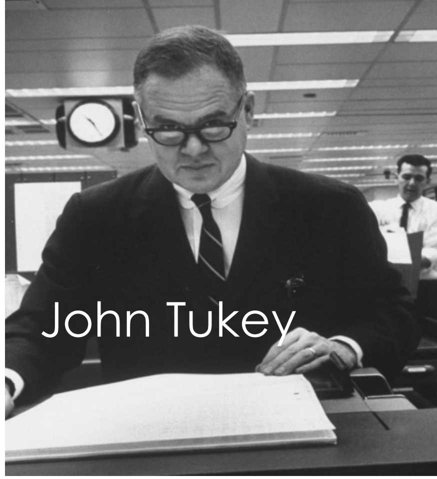
- Quantitative (Numeric)
  - Continuous (e.g., health care expenditure)
  - Discrete (e.g., number of siblings)
- Qualitative (Categorical)
  - Nominal (e.g., lane of traffic, country)
  - Ordinal (e.g., Yelp rating, education level)
- Think about how you might visualize each kind of data

Practice Types of Data

- Age
  - **Continuous**
- Homework assignments in a class
  - **Discrete**
- Political party affiliation
  - **Nominal**
- Exam Grade
  - Letter grade is **ordinal**
  - Score is **continuous**
- ZIP-code (e.g., 94703)
  - **Nominal**

Exploratory Data Analysis (EDA) and Visualization

Exploratory Data Analysis



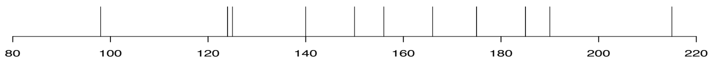
- Goals of EDA
  - **Validate** the **data collection** and preparation
  - **Confirm understanding** of the data
  - Search for **anomalies** or where data is **surprising**
- Iterative Exploratory Process
  - Analyze **summary statistics** and **data distributions**
  - **Transform** and **analyze relationships** between variables
  - **Segment data** across informative dimensions (granularity)
  - Use <u>**visualizations**</u> to build a deeper understanding

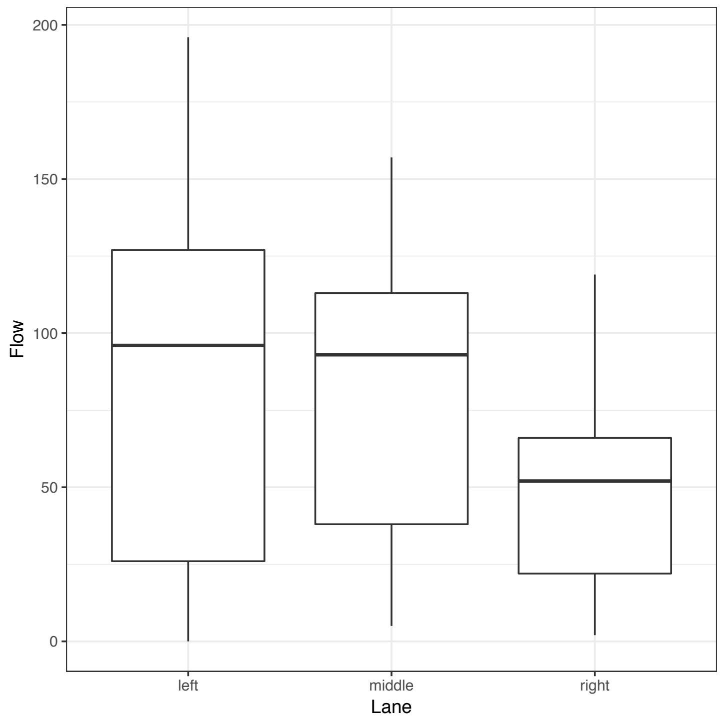
Several Case Studies in Class

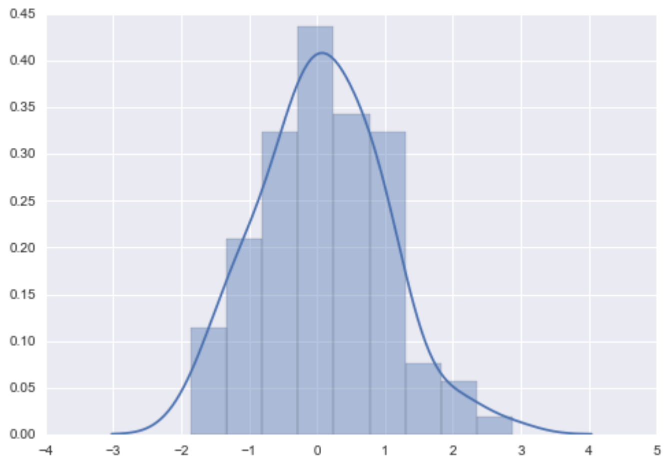
- Food safety in San Francisco
  - Recall heavy use of aggregation and spatial visualization
- In class TaFeng shopping data analysis
  - Data cleaning and outlier detection
  - You are required to be familiar with homework
- Freeway traffic analysis
  - Visualizing distribution of flow for each lane
- Baby Names in Pandas
  - Understanding popular baby names over time
- **You are not required to know these**
  - *reviewing may be helpful*

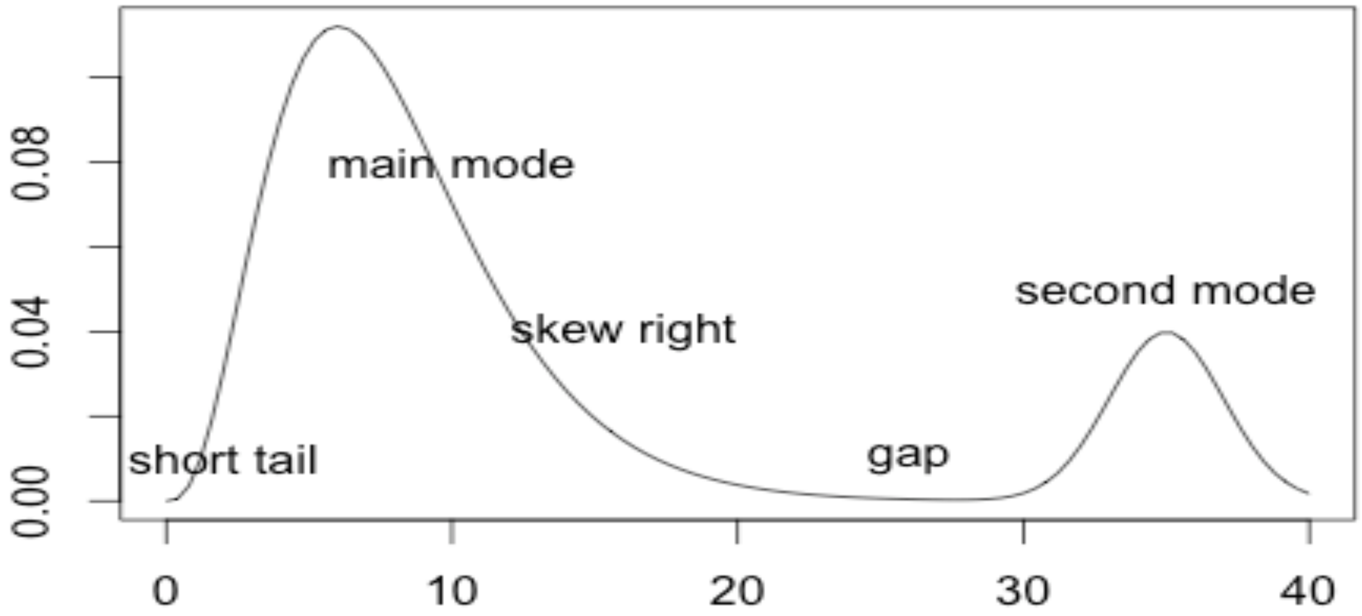
Visualizing Distributions

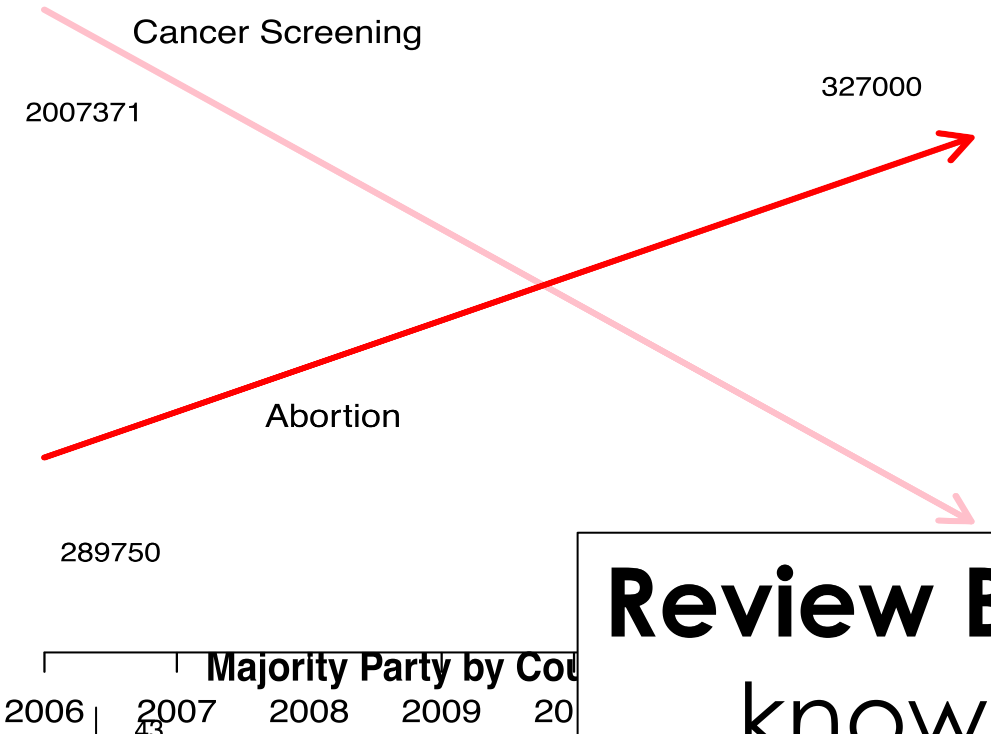
- Rug Plots, Box Plots, Histograms, Smoothed Estimators (e.g., KDEs)
- Terminology
  - Modes, Skew, Tails, Gaps, Outliers

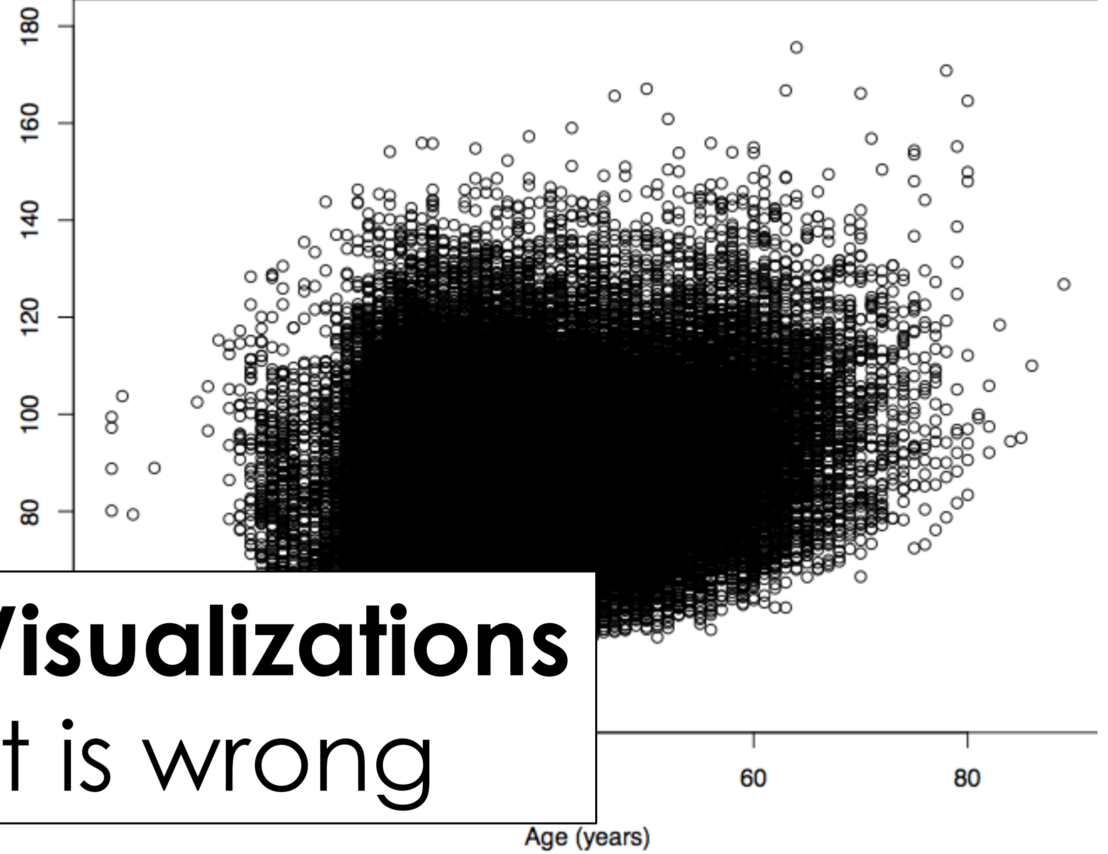






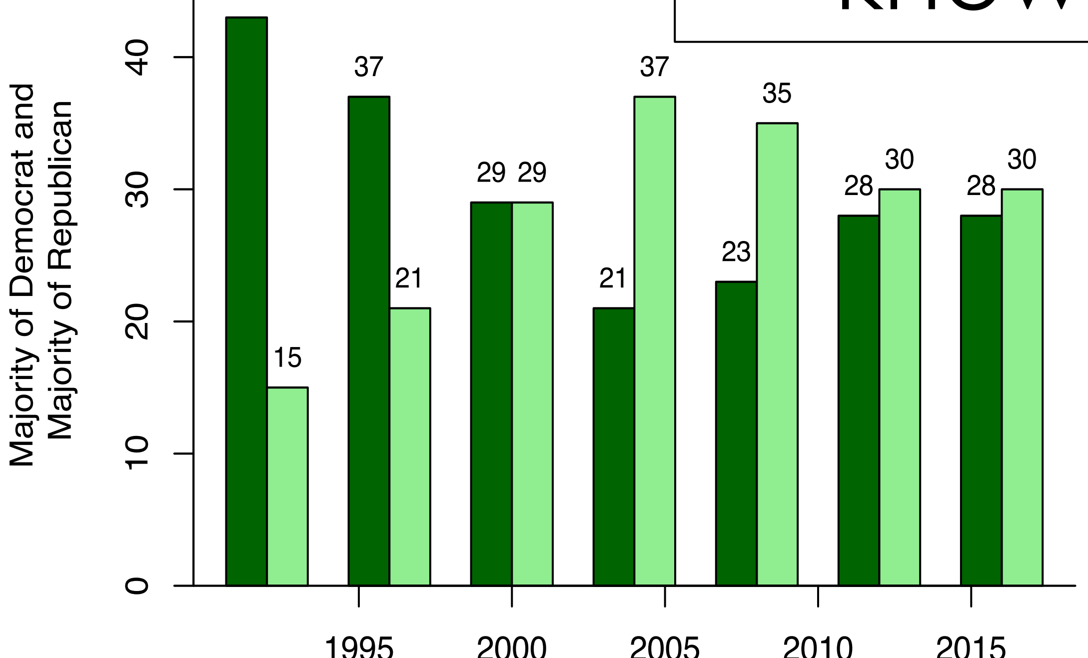


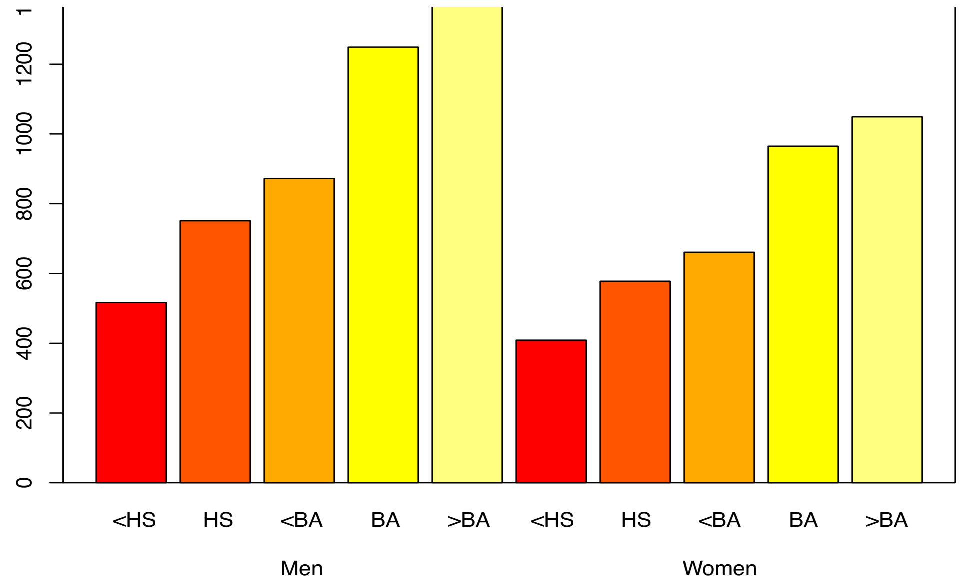




Review Bad Visualizations

know what is wrong

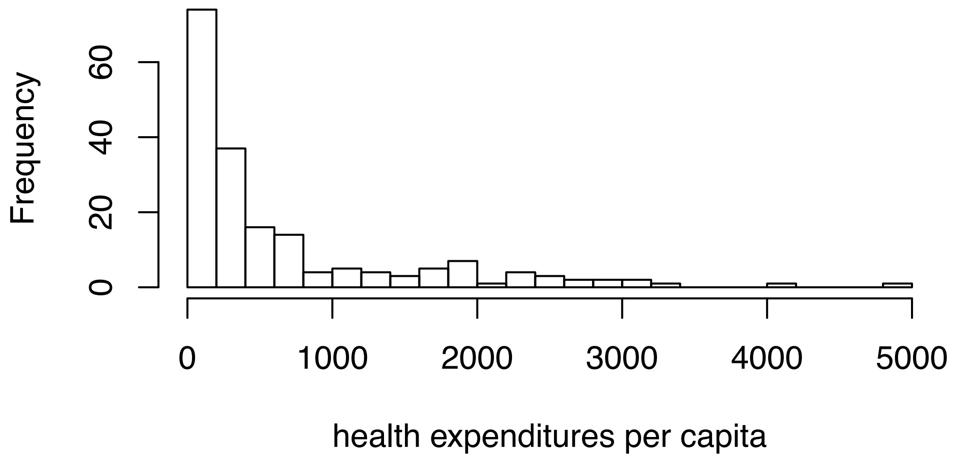


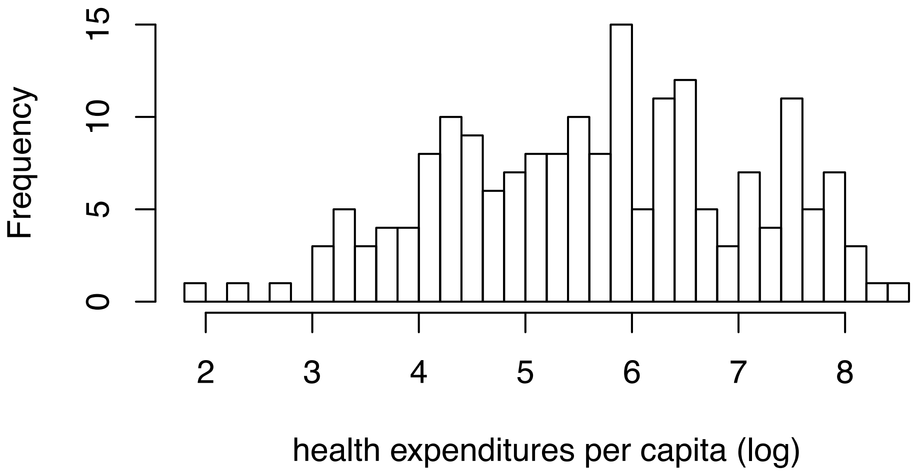


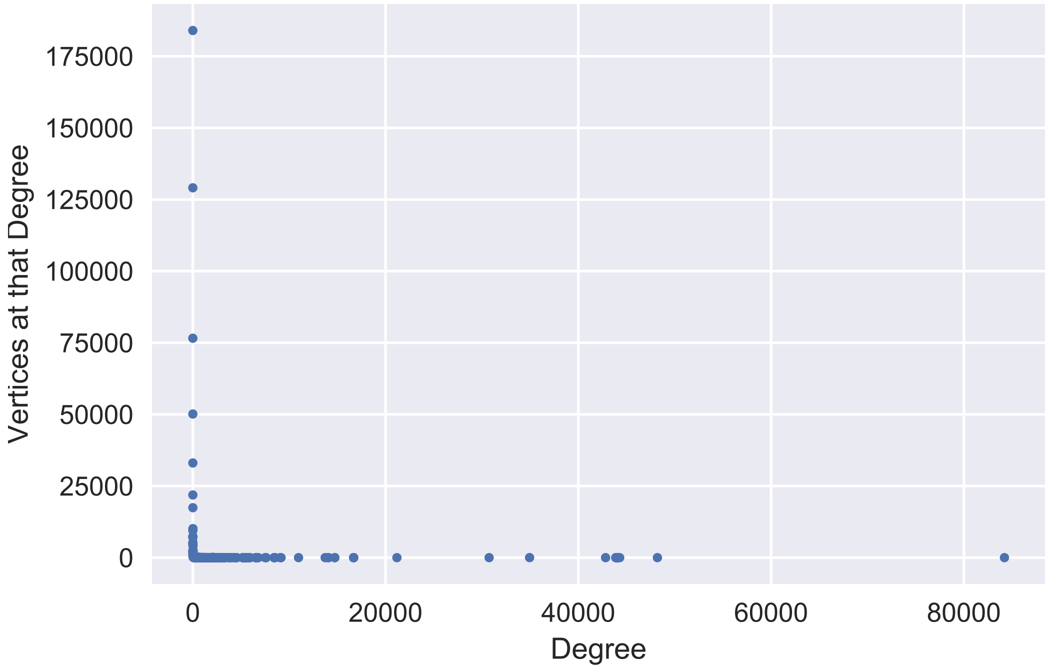
Techniques of Visualization

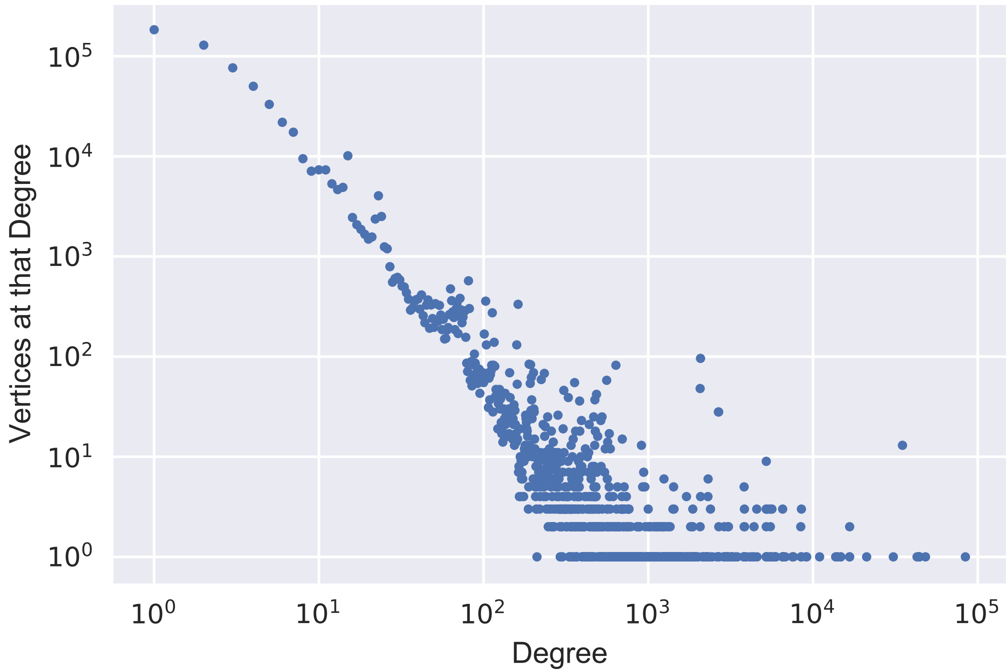
- **Scale:** ranges of values and how they are presented
  - Units, starting points, zoom, …
- **Conditioning:** breakdown visualization across a dimensions for comparison (e.g., income as function of education conditioned on gender)
- **Perception**
  - **Length**: encode relative magnitude (best for comparison)
  - **Color**: encode conditioning and additional dimensions and
- **Transformations:** to linearize relationships highlight important trends(e.g., log-y & log-log plots)
  - Symmetrize distribution
  - Linearize relationships
- Avoid stacking, chart junk, and over plotting

Log Transformations





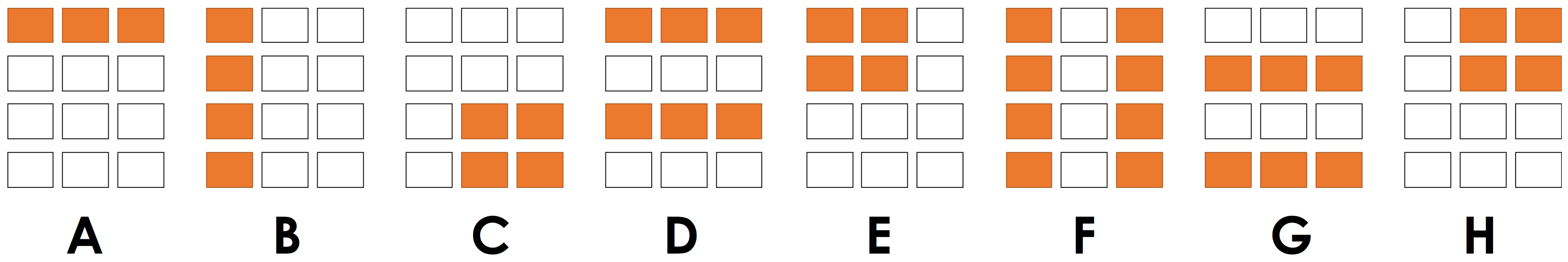




Numpy + Pandas

Numpy

- Review basic slicing commands and Boolean indexing



- **Key functions:** *sum, mean, variance, arange*
- *You will **not** be asked to **write complex python** programs on the exam.*
  - *You may need to **read** complex python expression…*

Pandas

- Review *column selection* and *Boolean slicing on rows*
- Be familiar with usage of basic commands:
  - *sort_values, head, read_csv*
- Review ***groupby*** and **merge** syntax:
  - *df.groupby([‘state’, ‘gender’])[[‘age’, ‘height’]].mean()*
    - *Know about mean(), sum(), count()*
  - dfA.merge(dfB, on=‘key’, how=‘outer’)
    - be comfortable with *inner* and *outer* joins
- Understand rough usage of basic plotting commands
  - plot, barh, histogram …

Prediction and Inference

Prediction and Inference

- Generalize beyond the data
  - **make predictions:** *how much of this product will we sell?*
  - **infer properties of the population:** *what is the distribution of heights of all humans?*
- Relies on **Models** and **Assumptions**
  - **Models:** capture essential trends, laws, or patterns
    - make predictions + reveal relationships
  - **Assumptions:** about the data and it's relationship to the quantities of interest
    - e.g., past reflects the future
- Fundamental tool of science: *test hypotheses*

Machine Learning

- **Defn:** study of algorithms & programs that improve through experience
  - **Key enabling technology**: voice rec., spam detection, online advertising, fraud detection, content recommendation, self driving cars …
  - *Learning* a function/program from **Input(X) → Output(Y)**
    - make predictions (ML & Stats)
    - make inferences about the population (Stats)

- **Taxonomy of ML problems:**
  - **Supervised Learning:** *given **input** (e.g., image) and **output** (e.g., Label)*
    - **Classification:** output is nominal (e.g., spam/ham)
    - **Regression:** output is continuous (e.g., price)
  - **Unsupervised Learning:** *given just the **input** (e.g., image)*
    - **Clustering:** output is nominal (e.g., day or night)
    - **Dimensionality Reduction:** output is continuous (e.g., time of year)
  - **Reinforcement Learning:** *given **input** and **rewards***
    - Game AIs and robotic controllers

Practice Examples

- Will it rain tomorrow? (Data: rainfall & season)
  - Answer: Supervised, Classification
- How much will it rain? (Data: rainfall & season)
  - Answer: Supervised, Regression
- What are the kinds of micro-climates in SF? (Data: Rainfall)
  - Answer: Unsupervised, Clustering
- Plot the micro-climates in 2D according to their similarity?
  - Answer: Dimensionality Reduction
- Translate this sentence to Korean?
  - Answer: … sequence of classification problems … (not on exam)

Relational Algebra and SQL

Relational Terminology

- **Database:** Set of Relations
- **Relation:** a table
  - **Schema:** the metadata including names, attribute names and types (and optional constraints)
  - **Instance:** tuples that satisfy the schema
- **Attribute**: a column
- **Tuple:** a row
- **Database Schema:** the set of schemas of its relations.

Relational Algebra Operators

Unary Operators: *operate on* ***single*** *relation instance*

- **Projection (π):** *Retains only desired columns* (vertical)
- **Selection (σ):** *Selects a subset of rows (horizontal)*
- **Renaming (ρ):** *Rename attributes and relations.*

Binary Operators: *operate on* ***pairs*** *of relation instances*

- **Union ($\cup$):** *Tuples in r1 or in r2.*
- **Intersection ($\cap$):** *Tuples in r1 and in r2.*
- **Set-difference (—):** *Tuples in r1, but not in r2.*
- **Cross-product ($\times$):** *Allows us to combine two relations.*
- **Joins ($\bowtie_\theta$, $\bowtie$):** *Combine relations that satisfy predicates*

Extensions: *Not in the original relational algebra*

- $\gamma_{\text{age, AVG(rating)}}$ **(Sailors):** groupby operator

Practice Question

Boats(bid,bname,color)
Sailors(sid, sname, rating, age)
Reserves(sid, bid, day)

- For each boat color what is the average rating of sailors over 32 that reserved a boat of that color?

$$\gamma_{\text{color, AVG(rating)}}(\sigma_{\text{age>32}}(\text{Sailors}) \bowtie_{\text{sid}} \text{Res} \bowtie_{\text{bid}} \text{Boats})$$

- Names of all pairs of sailors that reserved the same boat

$$\sigma_{\text{sname1 != sname}}(\pi_{\text{sname1, sname}}(\rho_{\text{T(name1,bid)}}(\pi_{\text{sname,bid}}(\text{Sailors} \bowtie_{\text{sid}} \text{Res})) \bowtie_{\text{bid}} \text{Res} \bowtie_{\text{sid}} \text{Sailors}))$$

SQL: Structured Query Language

- Understand the basic structure of SQL queries:
  - Be able to fill in the blanks
- Know base operators and agg.
  - AND, OR, <>, IS NULL, AVG, COUNT, COUNT(DISTINCT…)
- You should also know
  - **WITH r1(c1, c2) AS** (SELECT …
- Look over HW4 …

SELECT …
FROM …
[**WHERE** …]
[**GROUP BY** …]
[**HAVING** …]
[**ORDER BY** …]
[**LIMIT** …];

Practice Question

Boats(<u>bid</u>,bname,color)
Sailors(<u>sid</u>, sname, rating, age)
Reserves(<u>sid</u>, <u>bid</u>, <u>day</u>)

- For each boat color what is the average rating of sailors over 32 that reserved a boat of that color?

```
SELECT      color, AVG(rating)
FROM        Boats B, Reserves R, Sailors S
WHERE       B.bid = R.bid AND R.sid = S.sid
            AND S.age >32
GROUP BY    color;
HAVING
ORDER BY
LIMIT
```

Practice Question

Boats(bid,bname,color)
Sailors(sid, sname, rating, age)
Reserves(sid, bid, day)

➢ Names of all pairs of sailors that reserved the same boat

```
SELECT    DISTINCT S1.name, S2.name
FROM      Sailors S1, Reserves R1,
          Sailors S2, Reserves R2
WHERE     S1.sid = R1.sid AND R2.sid = S2.sid
          AND R1.bid = R2.bid
          AND S1.name < S2.name
```

Extra Study Suggestions

- You should be comfortable with filling in SQL expressions

- What will not be covered on the midterm
  - Window functions
  - User defined functions and aggregates
  - Table and View creation

Probability, Maximum Likelihood, and Priors

Rules of Probability

Ω: set of all possible outcomes from the chance process

A and B: collections of outcomes, AKA an event

1. $P(\Omega) = 1$
2. $0 \leq P(A) \leq 1$
3. If A and B disjoint, then $P(A \text{ or } B) = P(A) + P(B)$

- If B is contained in A, then $P(B) \leq P(A)$
- $P(A^c) = 1 - P(A)$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- $P(A \text{ and } B) = P(A)\, P(B)$ ***if A and B are independent***
- **Conditional Probability:** $P(A|B) = P(A \text{ and } B)/P(B)$
  - $P(A \text{ and } B \text{ and } C) = P(A \mid B, C)\, P(B|C)\, P(C)$

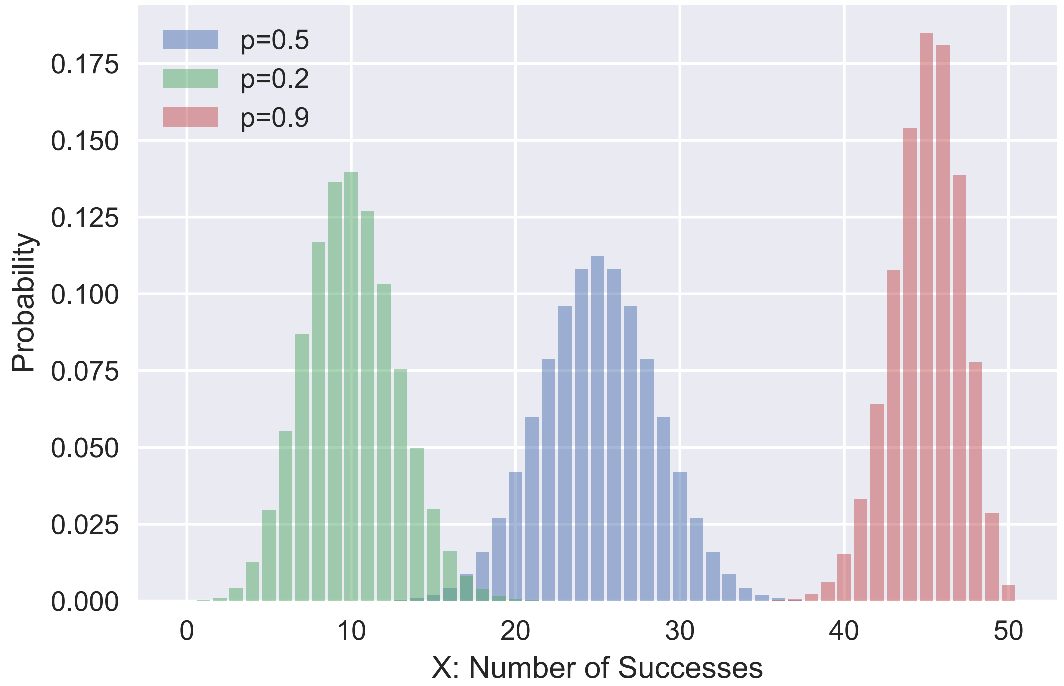
Distributions

- ***X ~ Bernoulli(p)*** distribution → X ∈ {0,1}:
  - Example: flipping a coin (or a thumb tack)

$$\mathbf{Prb}(X = k) = p^k(1-p)^{(1-k)} = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

- ***X ~ Binomial(n,p)*** distribution → X∈ {0,1, …, n}:
  - Number of times the coin lands heads in *n* flips

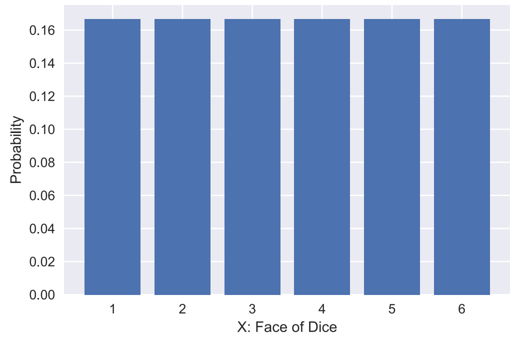
$$\mathbf{Prb}(X = k) = \binom{n}{k} p^k(1-p)^{(n-k)}$$



Distributions Continued

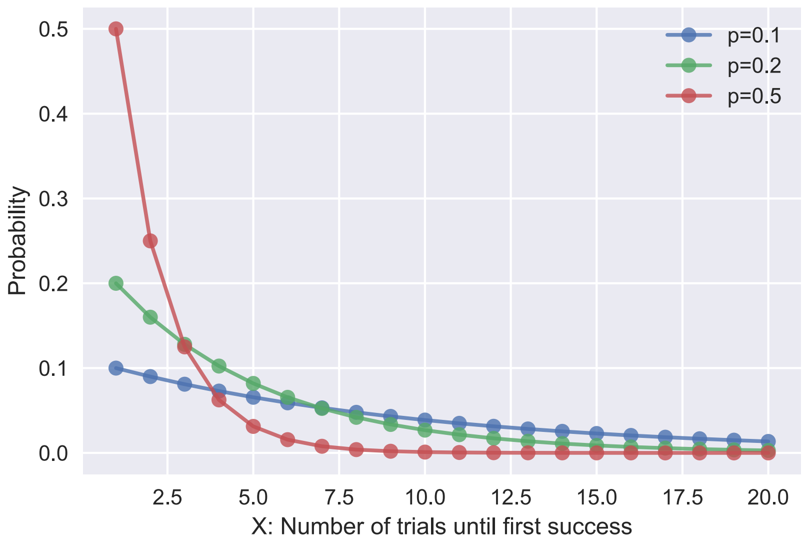
- ***X ~ DiscreteUniform(a,b)*** distribution → $X \in \{a, a+1, \ldots, b\}$:
  - Roll a fair dice

$$\mathbf{Prb}(X = k) = \frac{1}{b - a + 1}$$

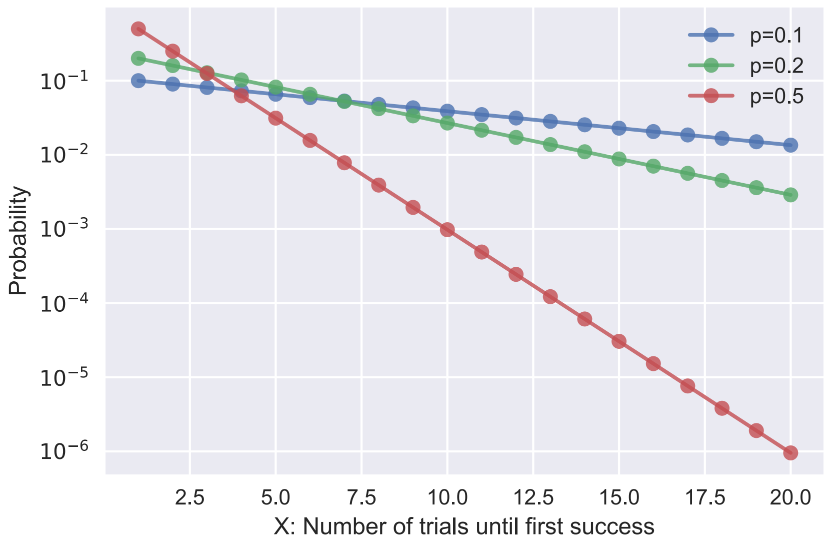


- ***X ~ Geometric()*** distribution → $X \in \{1,2,3,\ldots\}$:
  - Number of times to flip a coin until it first lands heads

$$\mathbf{Prb}(X = k) = p(1 - p)^{k-1}$$



Log y



Distributions Continued

- You will not be required to know
  - Continuous distributions (e.g., Beta, Normal, Uniform)
    - We will use them later and they may be final ☺
  - Hypergeometric
  - Posterior estimation and conjugacy

Summarizing Distributions

- **Expectation**

$$\mathbb{E}[X] = \sum_{x \in \Omega} x\, \mathbf{P}(x)$$

Properties of Expectations

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

- **Variance**

$$\mathbf{Var}[X] = \sum_{x \in \Omega} (x - \mathbb{E}[X])^2\ \mathbf{P}(x)$$

$$= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Properties of the Variance

$$\mathbf{Var}[aX + b] = a^2\mathbf{Var}[X]$$

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

$$\text{if independent}$$

Method of Maximum Likelihood

- How do we estimate the parameters of a model?
  - Maximize the likelihood of the data under the model

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \underbrace{\mathbf{P}\left(\mathcal{D} \,|\, \theta\right)}_{\text{Likelihood often written } \mathcal{L}(\theta)}$$

- How do we determine P(D | θ)?
  - Often assume Independent and Identically Distributed Data (IID)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} \mathbf{P}\underbrace{(X = x_i \,|\, \theta)}_{\text{Likelihood for each record}}$$

  - Where *D* is the list of obs. $(x_1, \ldots, x_n)$

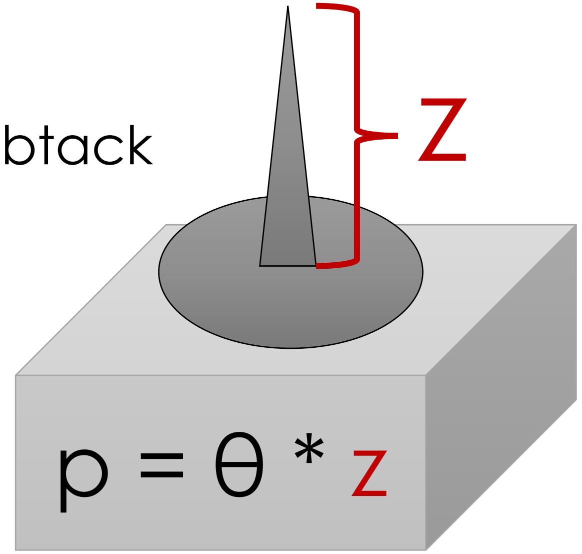
- How do we determine P(D | θ)?
  - Often assume Independent and Identically Distributed Data (IID)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} \underbrace{\mathbf{P}(X = x_i \,|\, \theta)}_{\text{Likelihood for each record}}$$

  - Where *D* is the list of obs. $(x_1, \ldots, x_n)$
- Need to define the likelihood for each record (modeling!):
  - Example:

Probability needle lands facing up (heads)

Thumbtack



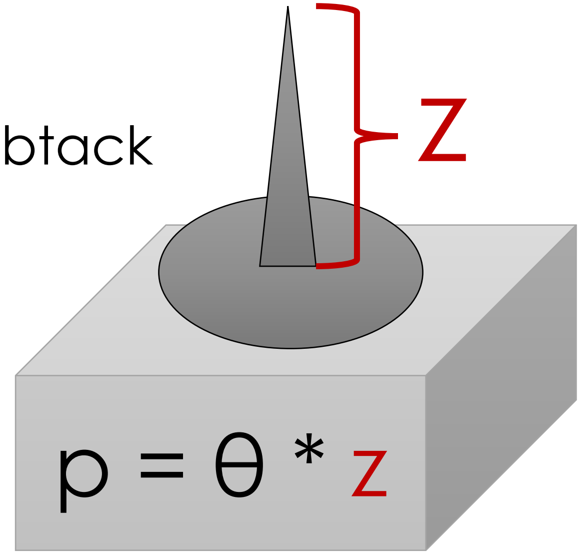
Assume z ∈ [0,1]

$$X \sim \mathbf{Bernoulli}(\theta z)$$

$$P(X = x \,|\, \theta) = (\theta z)^x (1 - \theta z)^{1-x}$$

- Need to define the likelihood for each record (modeling!):
  - Example:

Thumbtack



Assume z ∈ [0,1]

Probability needle lands facing up (heads)

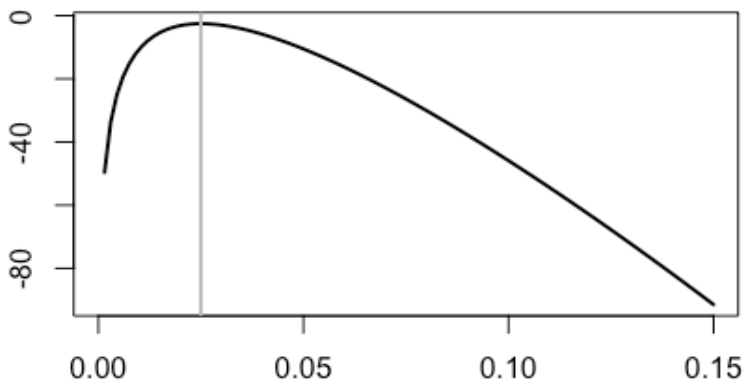
$$X \sim \mathbf{Bernoulli}(\theta z)$$

$$P(X = x \mid \theta) = (\theta z)^x (1 - \theta z)^{1-x}$$

$$\hat{\theta}_{\mathrm{M}LE} = \arg\max_{\theta} \prod_{i=1}^{n} \mathbf{P}(X = x_i \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} (\theta z)^{x_i} (1 - \theta z)^{1-x_i}$$

Take the log (why?)

$$= \arg\max_{\theta} \sum_{i=1}^{n} x_i \log(\theta z) + (1 - x_i) \log(1 - \theta z)$$



$$\hat{\theta}_{\mathrm{M}LE} = \arg\max_\theta \prod_{i=1}^{n} \mathbf{P}(X = x_i \,|\, \theta) = \arg\max_\theta \prod_{i=1}^{n} (\theta z)^{x_i} (1 - \theta z)^{1-x_i}$$

Take the log (why?)

$$= \arg\max_\theta \sum_{i=1}^{n} x_i \log(\theta z) + (1 - x_i) \log(1 - \theta z)$$

- Maximize by computing the derivative

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(\theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} x_i \log(\theta z) + (1 - x_i) \log(1 - \theta z)$$

$$= \sum_{i=1}^{n} x_i \frac{\partial}{\partial \theta} \log(\theta z) + (1 - x_i) \frac{\partial}{\partial \theta} \log(1 - \theta z)$$

- Maximize by computing the derivative

$$\frac{\partial}{\partial\theta}\log\mathcal{L}(\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n} x_i\log(\theta z) + (1-x_i)\log(1-\theta z)$$

$$= \sum_{i=1}^{n} x_i\frac{\partial}{\partial\theta}\log(\theta z) + (1-x_i)\frac{\partial}{\partial\theta}\log(1-\theta z)$$

$$= \sum_{i=1}^{n} x_i\frac{z}{\theta z} + (1-x_i)\frac{-z}{1-\theta z}$$

From calculus ("chain rule"):

$$\frac{\partial}{\partial\theta}\log f(\theta) = \frac{1}{f(\theta)}\frac{\partial}{\partial\theta}f(\theta)$$

- Set equal to zero and solve for θ:

$$\sum_{i=1}^{n} x_i\frac{z}{\theta z} + (1-x_i)\frac{-z}{1-\theta z} = 0$$

- Set equal to zero and solve for θ:

$$\sum_{i=1}^{n} x_i \frac{z}{\theta z} + (1 - x_i)\frac{-z}{1-\theta z} = 0$$

- Algebra

$$\frac{1}{\theta}\underbrace{\sum_{i=1}^{n} x_i}_{S} = \frac{z}{1-\theta z}\left(n - \underbrace{\sum_{I=1}^{n} x_i}_{S}\right) \Rightarrow$$

$$\frac{s}{\theta} = \frac{z}{1-\theta z}(n-s)$$

$$\frac{s}{n-s} = \frac{z\theta}{1-\theta z}$$

$$\frac{s/n}{1-s/n} = \frac{z\theta}{1-\theta z}$$

- Solution

$$\hat{\theta}_{\mathrm{MLE}} = \frac{1}{nz}\sum_{x\in\mathcal{D}} x$$

How does this relate to the Bernoulli?

Bayes Rule and Priors

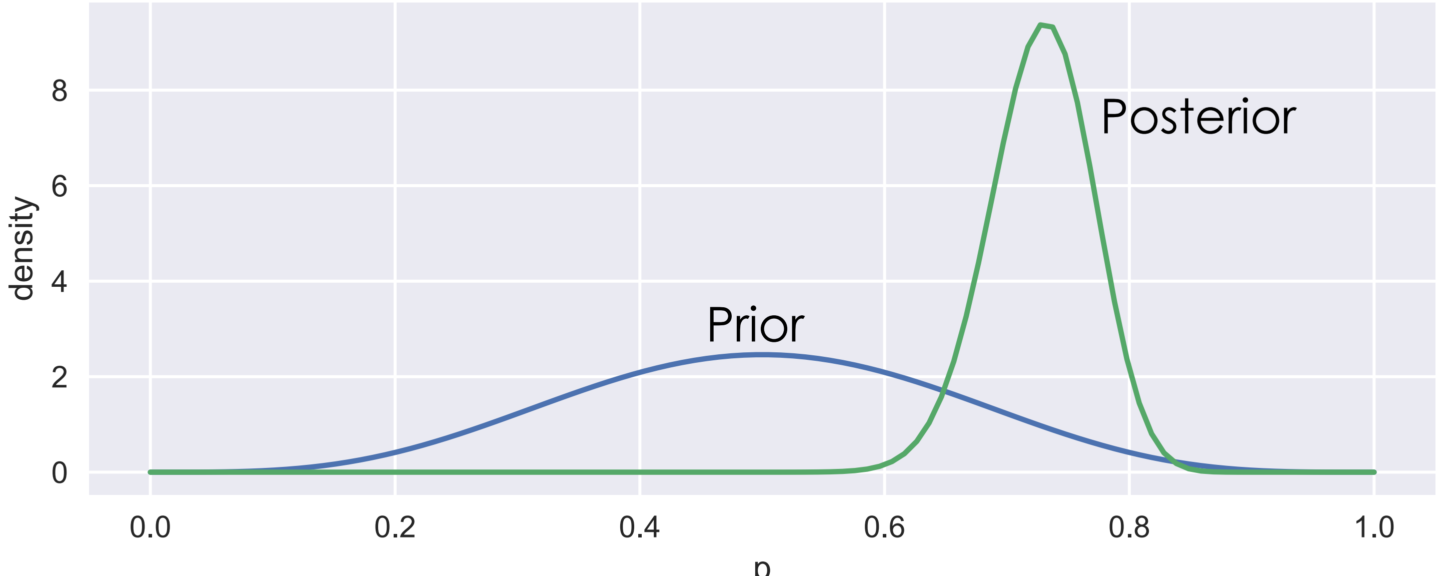
Posterior

Likelihood

Prior

$$\mathbf{P}(\theta \mid \mathcal{D}) = \frac{\mathbf{P}(\mathcal{D} \mid \theta)\mathbf{P}(\theta)}{\mathbf{P}(\mathcal{D})}$$

- Used to estimate a distribution over possible parameters:
  - **Likelihood** determines likelihood of the data under the model
  - **Prior** encodes our prior knowledge about the model
  - **Posterior** is our updated distribution over parameters



Exam Review Review

- Extra Study suggestions
  - Review lectures and section discussions
  - Look over practice questions
  - Go over this review lecture once more
- Exam Details (see Piazza post for details)
  - ~80 Minutes
  - ~1 page (front and back) answer sheet
  - Allowed one page (front and back) cheat sheet
- Good Luck!