Data Mining

CS57300
Purdue University

Jan 9, 2018

Bruno Ribeiro

Introduction

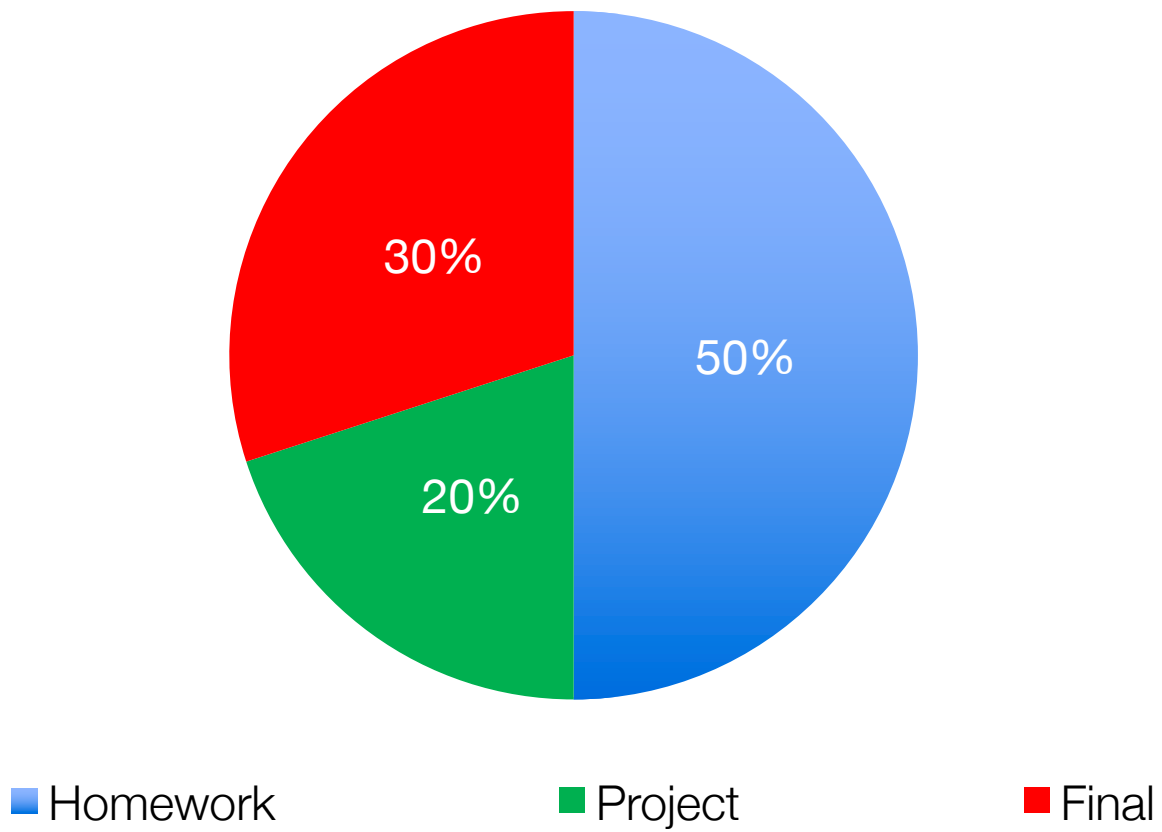
- Course overview
- What is data mining?
- Data mining process

Course overview

Logistics

- Time and location: Tue Thu 9:00am-10:15am, Wetherill Lab of Chemistry 320
- Instructor: Bruno Ribeiro
  ribeiro@cs.purdue.edu, LWSN 2142C,
  office hours: Thusdays 10:30am-11:30am
- Teaching assistants:
  Linjie Li. Office hours: HASS G50, Thursdays 1:30-2:30 PM.
  Akhil Israni. Office hours: HASS G50, Friday 3:00-4:00PM.
  Anoop Santhosh. Office hours: HASS G50, Monday 9:00-10:00AM.
- Webpage: https://www.cs.purdue.edu/~ribeirob/courses/Spring2018
- All communications via Piazza (see course webpage)
- Prerequisites: introductory statistics course (e.g., STAT 516), basic programming skills (e.g., CS381, STAT598G)
  - CS57300 Assumes you know basic probability and statistics, linear algebra, Python programming (see syllabus for topics list)

Course Grade Distribution



Homework (50%) Workload

- Assignments include written/math exercises, programming assignments in **Python 3**
- **Datasets** given in assignments are real data (dirty), that is, missing data, incomplete data, wrong data
  - **Cleaning** the data is part of the learning experience
- 5 or 6 homeworks, lowest grade removed from HW average
  - ABSOLUTELY NO extensions
  - ABSOLUTELY NO late homeworks
- Homework usually due 6:00am
  - Only TWO PDF uploads on Blackboard

Homework Competitions

- Sometimes you will compete with each other
- Grades: Curve of best prediction ranking

kaggle

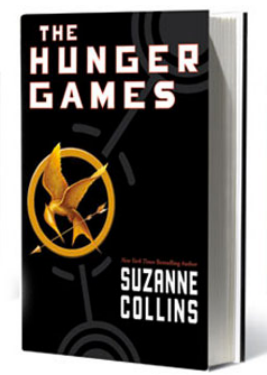
PUBLIC LEADERBOARD

This leaderboard is calculated on approximately 30% of the test data so the final standings may be different.

* in the money

#	Δ1w	Team Name	RMSLE	Entries	Last Submission UTC (Best Submission - Last)
1	-	Opera Solutions *	0.455469	34	Mon, 24 Oct 2011 22:18:48 (-3.9d)
2	-	Petterson & Caetano @ NICTA *	0.456237	72	Thu, 06 Oct 2011 22:24:24 (-13.6d)
3	-	Market Makers	0.456384	130	Mon, 19 Sep 2011 23:12:43 (-20.1d)
4	-	Larry_tempe	0.456764	21	Mon, 24 Oct 2011 22:08:35 (-11.9d)
5	-	SD_John	0.456765	46	Thu, 13 Oct 2011 00:29:42
6	-	lily	0.457018	37	Tue, 11 Oct 2011 05:12:08 (-11.2d)

Brought to you by



Final & Project

- Final (30%) covers all material
- Project (20%):
  - We will soon post a set of projects
  - Projects are individual (no groups)
  - If you have your own project (research), you must talk to me by next week

Textbook

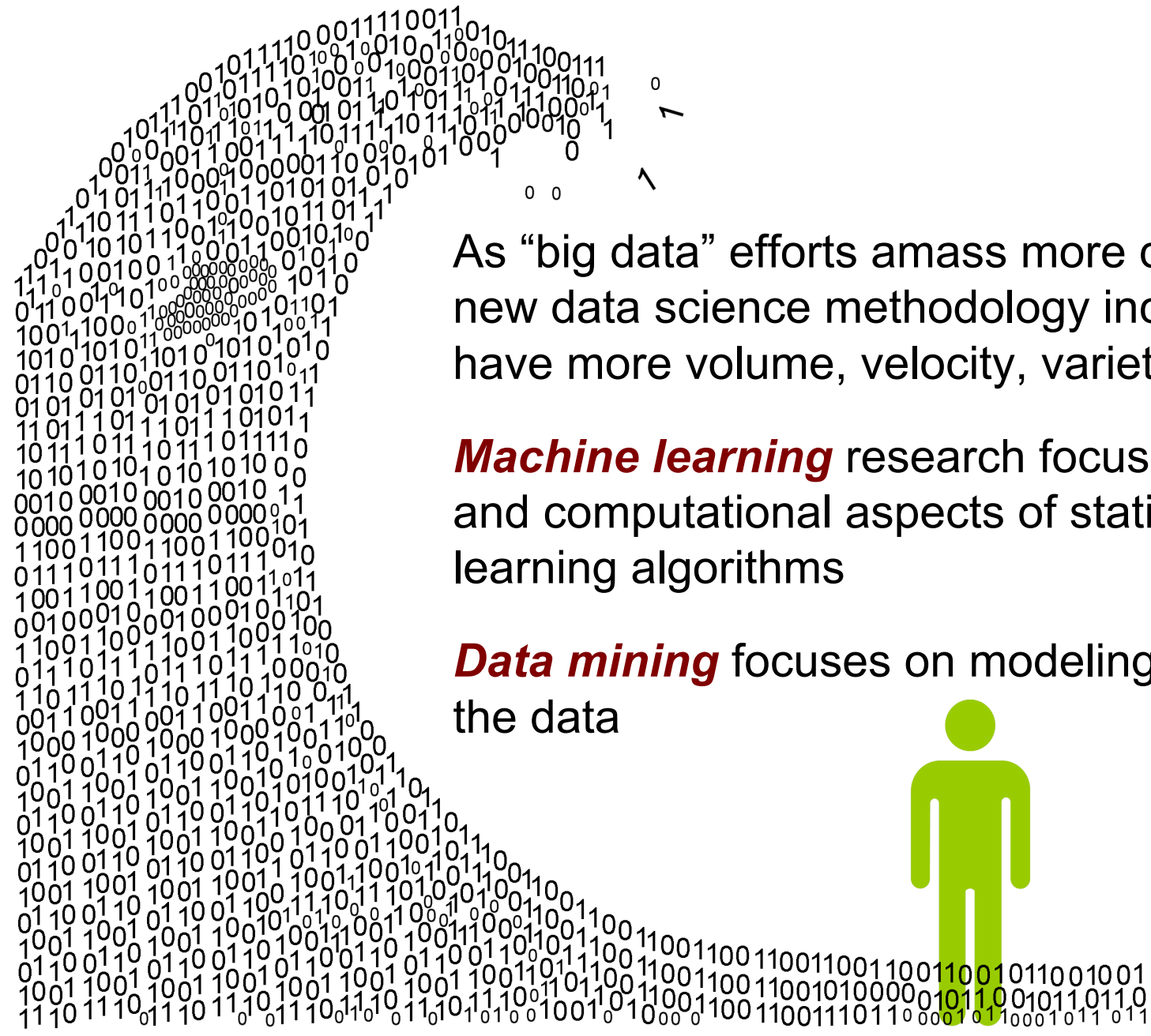
- (Required) David J. Hand; Heikki Mannila; Padhraic Smyth, Principles of Data Mining, FREE with PUID
- The texts below are recommended but not required:
  - David J. Hand; Heikki Mannila; Padhraic Smyth, Principles of Data Mining, FREE with PUID
  - James, Witten, Hastie, and Tibshirani, Introduction to Statistical Learning..
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning.
  - C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

Course goals

- Identify key elements of data mining systems and the knowledge discovery process
- Understand how algorithmic elements interact
- Recognize various types of data mining tasks
- Familiarity with standard models/algorithms
- Implement and apply basic algorithms
- Understand how to evaluate performance

What is Data Mining?

The data revolution



As “big data” efforts amass more data… the need for new data science methodology increases. Data today have more volume, velocity, variety, etc.

Machine learning research focuses on the theoretical and computational aspects of statistical models and learning algorithms

Data mining focuses on modeling and understanding the data

Standard view of data mining process

Data

Selection

Target data

Preprocessing

Processed data

Model

Knowledge

Interpretation evaluation

Patterns

Mining

Topics

- See website

Data Mining: Not Just Applying Known Techniques

“World Map” in 1459

- Shown biased and incomplete (Columbus et al. 1492)
- Data analysis has similar problems

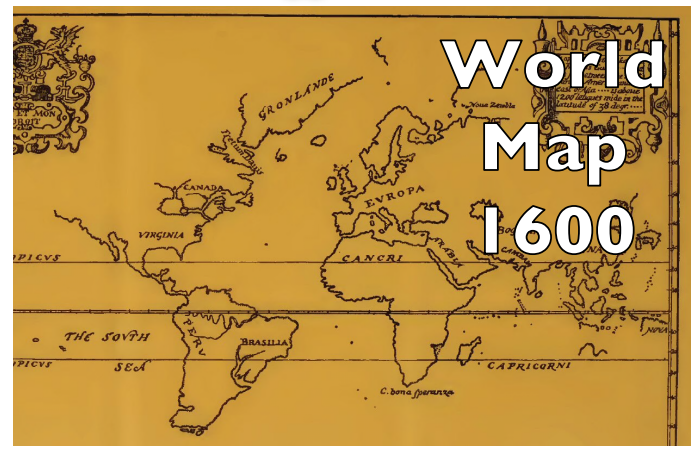
source: Wikipedia

The Fra Mauro world map (1459)

Data Analysis without Understanding the Data

Tools to Tailored to Data

More Precise Analysis



Example

- In the U.S., high school classrooms with the highest average SAT scores generally have at most 15 students
  - Why?
  - Statistical fluctuations:
    - E.g.: divide a 30 student classroom randomly into 10 classes with 3 students each.
    - Clearly, one of these classrooms will have much higher SAT scores than the others
    - Can we compare the teacher effectiveness in schools with small classes?

World-changing applications: charities, healthcare, politics, and, yes, advertising

Really, What is Data Mining?

- **Data mining** *is the science of discovering structure and making predictions in data sets*
- **Discovering structure (Unsupervised Learning)**
  - E.g., given observations $X_1, \ldots X_n$, learn some underlying structure
- **Making predictions (Supervised Learning)**
  - E.g., given observations $(X_1, Y_1), \ldots (X_n, Y_n)$, predict Y_i from X_i

Finding Likely Hypotheses from Data

- Supervised Learning

  Observations $\{(X_i, Y_i)\}_{i=1}^n$

  - X_i: e.g., user age, height, income; Y_i: e.g., user defaults on loan
  - Hypothesis h: $f_{\theta_h}(Y_i; X_i)$

- Unsupervised Learning

  Observations $\{(X_i)\}_{i=1}^n$

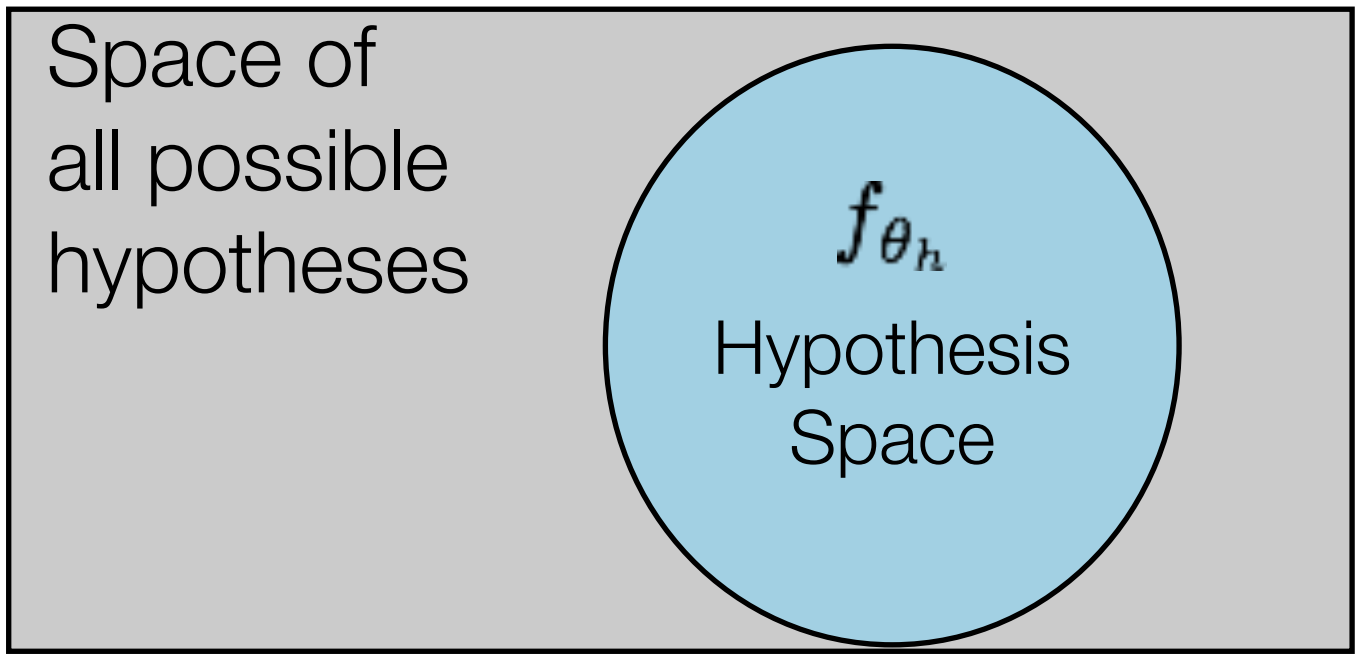
  - X_i: e.g., user age, height, income, health, ...
  - Hypothesis h: $g_{\theta_h}(X_i)$
  - θ_h might explain hidden "groups" (millennials, retirees, ...)

The Quest for a Good Hypothesis

- Need to understand the data
  - Missing data?
  - Hidden effects influencing observations?
    - Are observations biased?
    - Are observations independent?

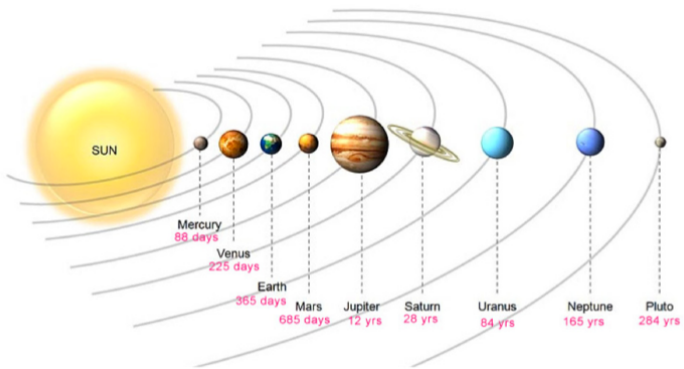
How to Find a Good Hypothesis?

- If we understand the data we can predict and explain it
  - Need to form a hypothesis about the data
  - How do we find a good hypothesis?
  - Issue: testing multiple hypotheses



Our Hypotheses Will Be Mostly About Non-intervention Predictions

- Data is often cold: already collected, no intervention possible
  - But… Newtonian celestial mechanics is also about correlations, and very successful



  - A hypothesis is t the ”future” (unseen data)
  - These conclusions rarely imply causation

E.g., Classroom SAT Scores Problem

- How to pose the high school SAT score problem as a hypothesis test problem?
  - Say, we want to know if classroom *i* is the best classroom according to its SAT scores
  - But, for each student, SAT scores are random (same student taking the test at different days will have different scores)
  - We will NEVER treat observed data as noise-free

- Data mining is a lot about formulating the question statistically:
  - X_{ijt} – SAT score of student j in classroom i at test t
    - We will often assume $\{X_{ijt}\}_{jt}$ are independent random variables with the same unknown distribution (with average μ_i)
    - $P(\mu_i > \max_{k \neq i} \mu_k \mid \{X_{ijt}\}_{ijt})$

…what is the probability classroom i will has the best (intrinsic) SAT average

Given the SAT score data we have seen so far…

Population Sampling

- X_{ijt} – SAT score of student j in classroom i at test t is a sample from the a population of SAT scores in classroom i

- More formally…

Modeling uncertainty

- Necessary component of almost all data analysis
- Approaches to modeling uncertainty:
  - Fuzzy logic
  - Possibility theory
  - Rough sets
  - **Probability *(focus in this course)***

Probability

- Probability theory *(some disagreement)*
  - Concerned with interpretation of probability
  - 17th century: Pascal and Fermat develop probability theory to analyze games of chance
- Probability calculus *(universal agreement)*
  - Concerned with manipulation of mathematical representations
  - 1933: Kolmogorov states axioms of modern probability

Probability basics

- Basic element: **Random variable**
  - Mapping from a property of objects to a variable that can take one of a set of possible values
  - *X* refers to random variable; *x* refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite set of possible values;
    Continuous RV can take any value within an interval
  - Boolean: e.g., Warning (is there a storm warning? = <yes, no>)
  - Discrete: e.g., Weather is one of <sunny,rainy,cloudy,snow>
  - Continuous: e.g., Temperature

Probability basics

- **Sample space (S)**
  - Set of all possible outcomes of an experiment
- **Event**
  - Any subset of *outcomes* contained in the sample space S
  - When events ***A*** and ***B*** have no outcomes in common they are said to be *mutually exclusive*

Examples

Random variable(s)	**Sample space**
One coin toss	H, T
Two coin tosses	HH, HT, TH, TT
Select one card	2♥, 2♦, ..., A♣ (52)
Play a chess game	Win, Lose, Draw
Inspect a part	Defective, OK
Cavity and toothache	TT, TF, FT, FF

Axioms of probability

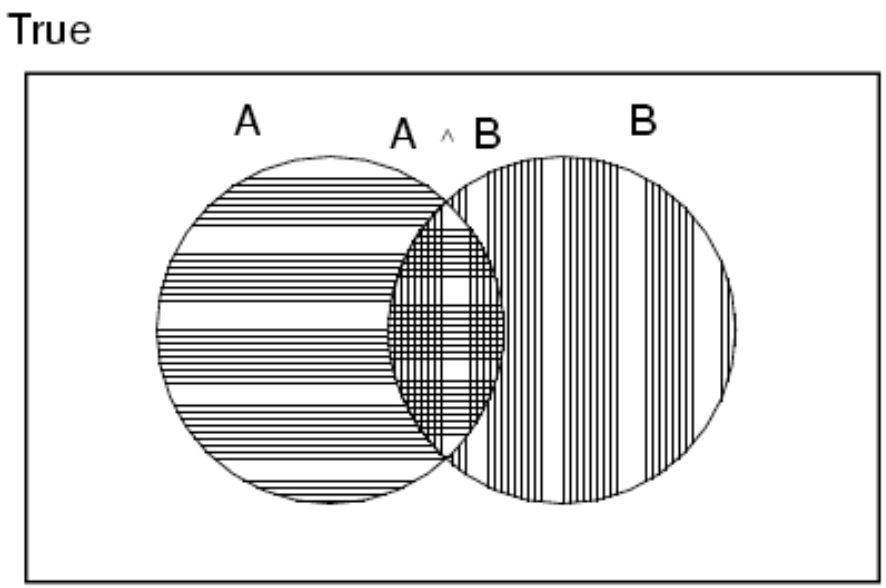
- For a sample space S with possible events $\mathbf{A_S}$, a function that associates real values with each event A is called a ***probability function*** if the following properties are satisfied:

  1. $0 \leq P(A) \leq 1$ for every A
  2. $P(S) = 1$
  3. $P(A_1 \cup A_2 \ldots \cup A_{n \in S}) = P(A_1) + P(A_2) + \ldots + P(A_n)$

     if $A_1, A_2, \ldots, A_n$ *are pairwise mutually exclusive events*

Implications of axioms

- For any events ***A, B*** in universe S
  - P(*A*) = 1 - P(S*A*)
  - P(true) = 1 and P(false) = 0
  - If A and B are mutually exclusive then P(*A* ∩ *B*) = 0
  - P(*A* U *B*) = P(*A*) + P(*B*) - P(*A* ∩ *B*)



Interpreting probabilities

- Meaning of probability is focus of debate and controversy
- Two main views: Frequentist and Bayesian

Frequentist view

- Dominant perspective for last century
- Probability is an **objective** concept
  - Defined as the frequency of an event occurring under repeated trials in “same” situation
  - E.g., number of heads in repeated coin tosses
- Restricts application of probability to repeatable events

Calculating probabilities (frequentist)

- Frequentist view
  - Let n be the number of times an experiment is performed
  - Let n(A) be the number of outcomes in which A occurs
  - Then as $n \to \infty$ $P(A) = n(A) / n$
- When the various outcomes of an experiment are equally likely, the task of computing probability reduces to counting
  - Let N be size of sample space (i.e., number of simple outcomes)
  - Let N(A) be the number of outcomes contained in A
  - Then: $P(A) = N(A) / N$

Example

- Roll two 6-sided dice. What is the probability that the result sums to 8?
  - P = num ways event can occur / possible outcomes
- What is the size of the sample space?
  - 6*6=36
- How many events involve the two dice summing to 8?
  - {2,6},{3,5},{4,4},{5,3},{6,2} = 5
- Overall probability?
  - 5/36 = 0.139

Permutations and combinations

- An **ordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **permutation** of size k
  - The number of permutations of size k that can be constructed from the n objects is:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

- An **unordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **combination** of size k
  - The number of combinations of size k that can be constructed from the n objects is:

$$C_{k,n} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Example

- An urn contains ten balls, six of which are red and four of which are white.
  Five balls are drawn at random. What is the probability of drawing three red and two white balls?

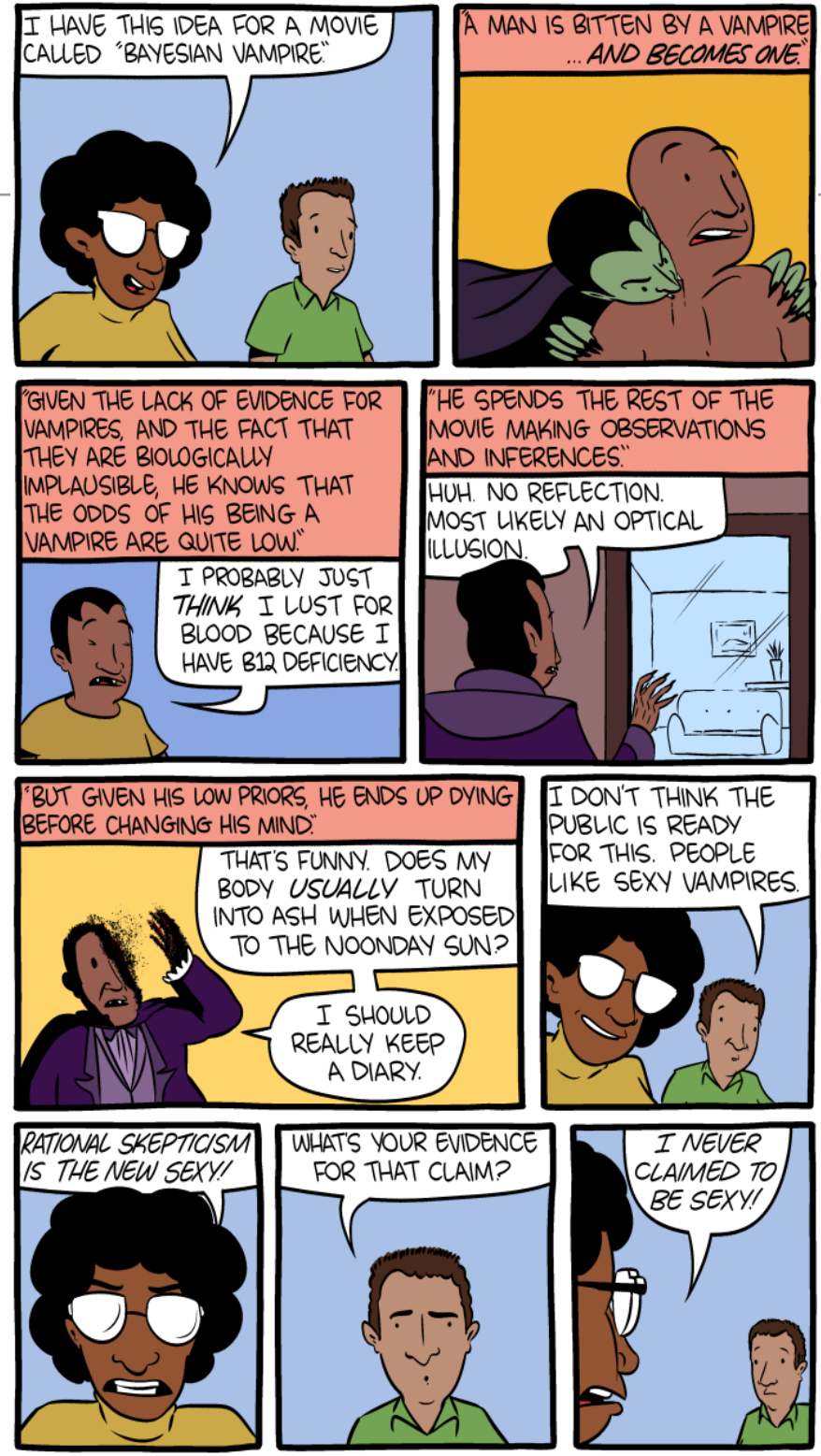
$$\frac{C_{3,6} \cdot C_{2,4}}{C_{5,10}} = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{6!}{3!3!}\frac{4!}{2!2!}\frac{5!5!}{10!}$$

- An urn contains five balls, numbered from 1 to 5. Three balls are drawn at random. What is the probability that we draw the sequence 3, 4, 1?

$$\frac{1}{P_{3,5}} = \frac{(5-3)!}{5!}$$

Bayesian view

- Increasing importance over last decades
  - Due to increase in computational power that facilitates previously intractable calculations
- Probability is a **subjective** concept
  - Defined as individual degree-of-belief that event will occur
  - E.g., belief that we will have another snow storm tomorrow
- Begin with prior belief estimates and update those by conditioning on observed data



Calculating probabilities: Bayesian

- *Begin with prior belief estimates*: P(A)
  - E.g., After the Seahawks won their conference, Vegas casinos believed the Seahawks were likely to win the Superbowl over the Patriots:
    P(S wins)=0.525, P(P wins)=0.475
- Observe data
  - But then Vegas observed a heavy majority of the betters (80%) chose the Patriots, which is unlikely given their current belief
- *Update belief by conditioning on observed data*
  P(A|data) = P(data|A) P(A) / P(data)
  - So they updated their belief to increase the the Patriots's chance of a win:
    P(S wins | betting) = P(betting | S wins) P(S wins) / P(betting) = 0.50
- Even when the same data is observed, if people have different priors, they can end up with different posterior probability estimates P(A|data)

Bayesian vs. frequentist

- Bayesian central tenet:
  - Explicitly model all forms of uncertainty
  - E.g., Parameters, model structure, predictions
- Frequentist often model same uncertainty but in less-principled manner, e.g.,:
  - Parameters set by cross-validation
  - Model structure averaged in ensembles
  - Smoothing of predicted probabilities
- Although interpretation of probability is different, underlying calculus is the same

Probability distribution

- **Probability distribution** *(i.e., probability mass function or probability density function)* specifies the probability of observing every possible value of a random variable
- Discrete
  - Denotes probability that *X* will take on a particular value:

$$P(X = x)$$

- Continuous
  - Probability of any particular point is 0, have to consider probability within an interval:

$$P(a < X < b) = \int_a^b p(x)dx$$

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

  E.g., P(Weather, Warning) = a 4 × 2 matrix of values:

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- Every question about events can be answered by the joint distribution

Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., P(warning=Y | snow=T) = 0.4
  - Complete conditional distributions specify conditional probability for all possible combinations of a set of RVs:
    P(warning | snow) =
    {P(warning = Y | snow = T), P(warning = N | snow = T)},
    {P(warning = Y | snow = F), P(warning = N | snow = F)}
- If we know more, then we can update the probability by conditioning on more evidence
  - e.g., if Windy is also given then P(warning | snow, windy) = 0.5

Conditional probability

- Definition of conditional probability:

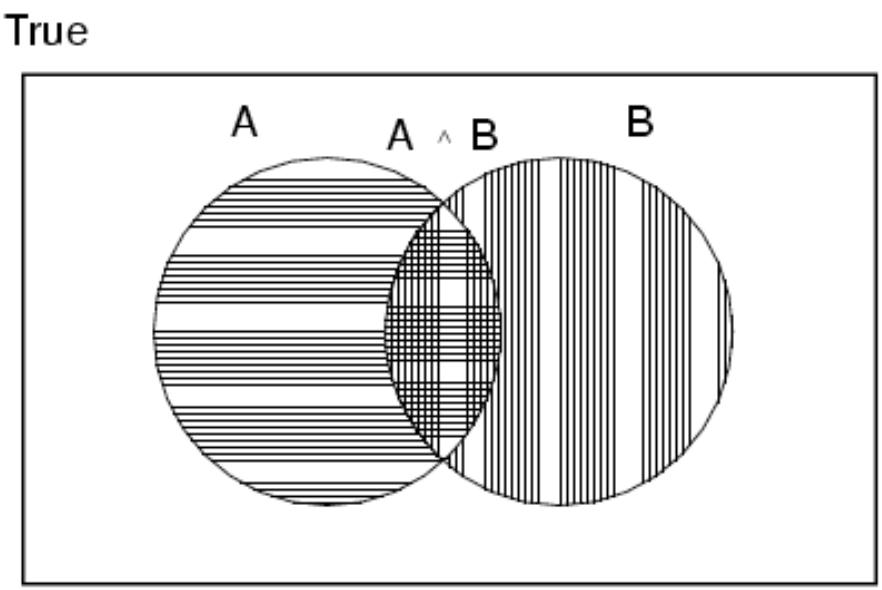
$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

- **Product rule** gives an alternative formulation:

$$P(A \cap B) = P(A|B)P(B)$$
$$= P(B|A)P(A)$$

- **Bayes rule** uses the product rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Example

- Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

- Example: What is P(sunny | warning = Y)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Conditional probability

- **Chain rule** is derived by successive application of product rule:

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= P(X_n | X_1, \ldots, X_{n-1}) P(X_1, \ldots, X_{n-1}) \\
&= P(X_n | X_1, \ldots, X_{n-1}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_1, \ldots, X_{n-2}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events
- Marginalization:

$$P(A) = \sum_{b \in B} P(A, b)$$

$$= \sum_{b \in B} P(A|b)P(b)$$

- Example: What is P(cloudy)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Independence

- A and B are independent iff:
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)P(B)$
  - *Knowing B tells you nothing about A*
- Examples
  - Coin flip 1 and coin flip 2?
  - Weather and storm warning?
  - Weather and coin flip=H?
  - Weather and election?

Conditional independence

- Two variables *A* and *B* are **conditionally** independent given *Z* iff for all values of *A, B, Z:*

$$P(A, B \mid Z) = P(A \mid Z)\, P(B \mid Z)$$

- *Note: independence does not imply conditional independence or vice versa*

Example 1

- **Conditional independence does not imply independence**
- Gender and lung cancer are not independent
  $P(C \mid G) \neq P(C)$
- Gender and lung cancer are conditionally independent given smoking
  $P(C \mid G, S) = P(C \mid S)$
- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

Example 2

- **Independence does not imply conditional independence**
- Sprinkler-on and raining are independent
  $P(S \mid R) = P(S)$
- Sprinkler-on and raining are not conditionally independent given
  grass is wet
  $P(S \mid R, W) \neq P(S \mid R)$
- Why? Because once we know the grass is wet, if it's not raining, then the explanation for the grass being wet has to be the sprinkler

Expectation

- Denotes the expected value or mean value of a random variable X

$$E[X] = \sum_x x \cdot p(x)$$

- Discrete

$$E[X] = \int_x x \cdot p(x) dx$$

- Continuous

$$E[h(X)] = \sum_x h(x) \cdot p(x)$$

- Expectation of a function

$$E[aX + b] = a \cdot E[X] + b$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- X = {0, 1, 2, 3}
- P(X=0) = 1/8; P(X=1) = 3/8; P(X=2) = 3/8; P(X=3) = 1/8
- What is the expected value of X, E[X]?

$$E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})$$
$$= \frac{3}{2}$$

Variance

- Denotes the squared deviation of X from its mean

- Variance

$$Var(X) = E[(x - E[X])^2]$$
$$= E[X^2] - (E[X])^2$$

- Standard deviation

$$\sigma = \sqrt{Var(X)}$$

- Variance of a function

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- X = {0, 1, 2, 3}

$$E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})$$
$$= \frac{3}{2}$$

- What is the variance of X, Var(X)?

$$Var(X) = \left(\left[0 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) + \left(\left[1 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[2 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[3 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right)$$
$$= \left(\frac{9}{4} \cdot \frac{1}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{9}{4} \cdot \frac{1}{8}\right)$$
$$= \frac{3}{4}$$

Questions?