BE 562: Computational Biology: Genomes, Networks, Evolution

Course Instructors

- Lecturer
  - James Galagan
  - LSEB 1002
  - jgalag@bu.edu
- TA
  - Yuan Yin
  - yin@bu.edu

Logistics

- Lectures

  Tu/Th 1:30-3:00, Room LSE B03

- Recitations

  Fri 9:05-9:55, Room LSE B03

- Course Website

  Blackboard 8 at learn.bu.edu

  You must be registered to access site and receive group emails

- Office Hours

  Please contact us to make appointments

Goals of Course

- Introduction to Computational Biology
  - Understand *how* algorithms work - Emphasize concepts over algebra
  - Recognize connection between different concepts and applications
  - Exposure to current research topics
- Hands on experience
  - Computational problem solving – formulating biological questions as computational problems
  - Programming – implementing useable solutions
  - Hands on experience with genomic datasets
  - Research: Final Projects

Course Outline

- First Half: Foundations
  - Core computational problems and concepts
  - String matching, DB Searching, HMMs, Gene Prediction, Clustering, Classification, Molecular Evolution
- Midterm 1
- Second Half: Frontiers
  - Current Research Topics
  - CRFs, Comparative Genomics, Gene Regulation, Metabolic Modeling, Systems & Synthetic Biology
- Mini-Midterm 2
- Final Project
  - Your own research topic

Grading

Problem Sets (35%)	Midterms (25%)	Final Project (35%)	Scribing (5%)

- 3 Problem Sets
  - Each on 12-15% of grade
- Exams
  - One in-class **midterm** **October 22**
  - We will work through all the problems in the next two lectures
  - A **mini-midterm**, in class, **November 26**
- Final Project
  - Introduction to research in computational biology (7 weeks!)
  - Includes peer-reviewed NIH-style proposal and much feedback
  - Presentations **December 5,10**
- Collaboration Policy
  - Collaboration allowed (except on exams), but you must:
  - Work independently on each problem before discussing it
  - Write solutions on your own
  - Acknowledge sources and collaborators. No outsourcing.

Problem Sets

- Due Mondays by 8 pm
  - Email to TA
  - Late Policy: Flexible +/- a few hours
  - More than few hours requires prior arrangement (except special circumstances, etc)
- 3 Problem Sets (ea. 12-15% of grade)
  - Typically 4-6 problems per assignment (except PS1)
  - Both theoretical and programming problems
- Programming
  - You can program in any language you like as long as it works
  - Recommend Python, Perl, or Matlab
  - Example code will be in Python or Matlab
  - Comment your code!
  - There is a lot of coding required!
- First Homework out Today – <u>Due September 16</u>

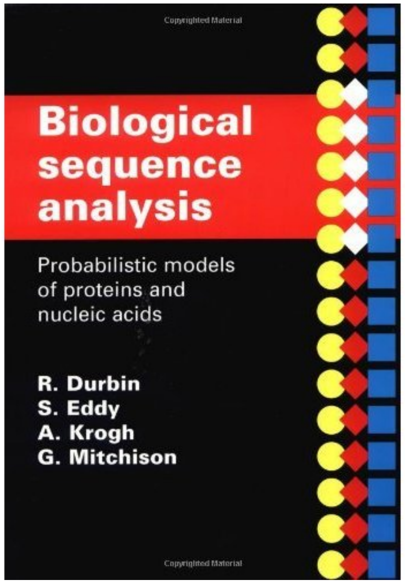
Final Project

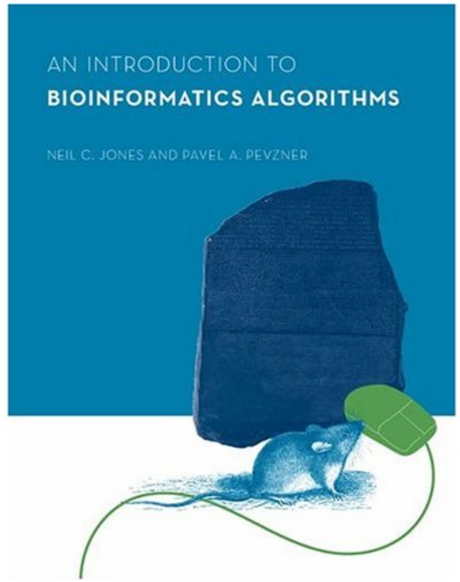
- The goal of the final project is to help prepare you for original research in computational biology.
- The Problem
  - Frame a biological question computationally
  - Gathering relevant literature and datasets
  - Solving it using new algorithms and ML techniques
  - Interpreting the results biologically
- The Process
  - Prepare a research proposal (fellowships/grants)
  - Review peer proposals (peer-review)
  - Receive feedback and revise your proposal (part of life)
  - Present your research orally in scientific audience (the fun part)
  - Writing up your results in a scientific paper (the other fun part)
- We will provide guidance on all these steps

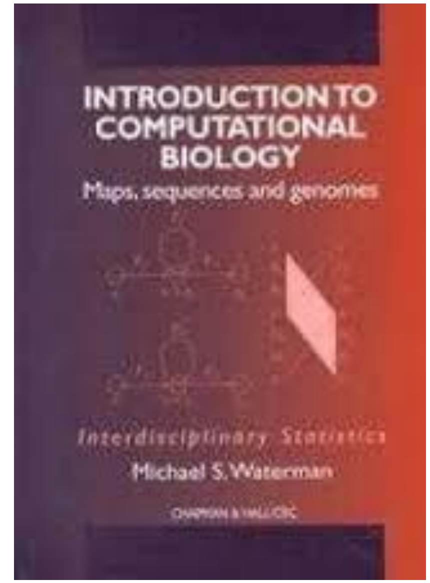
Lecture Scribing

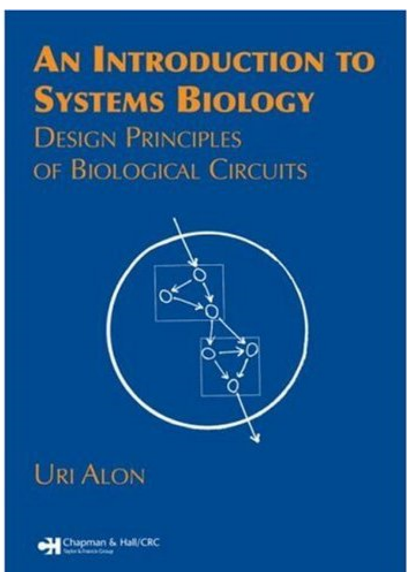
- Each lecture will have a dedicated scribe who will take notes on the lecture
  - Please sign up to scribe for lecture on the sheet being passed around
- Build on notes from previous years
  - Available on course website
  - Very mature for most lectures – just needs continued polishing
- First draft of scribe notes due 2 days after lecture
  - We will review and provide feedback
- Final draft of scribe notes due 6 days after lecture
- Counts for 5% of your grade

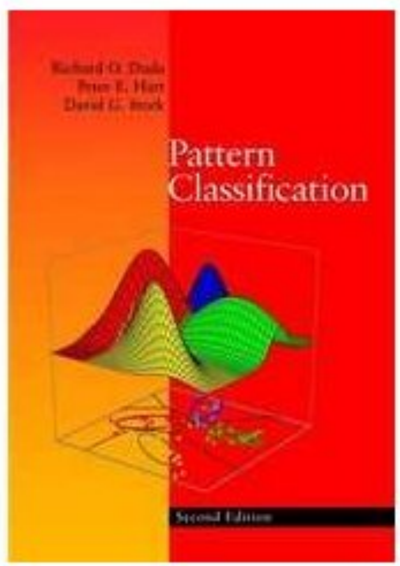
Reference Books

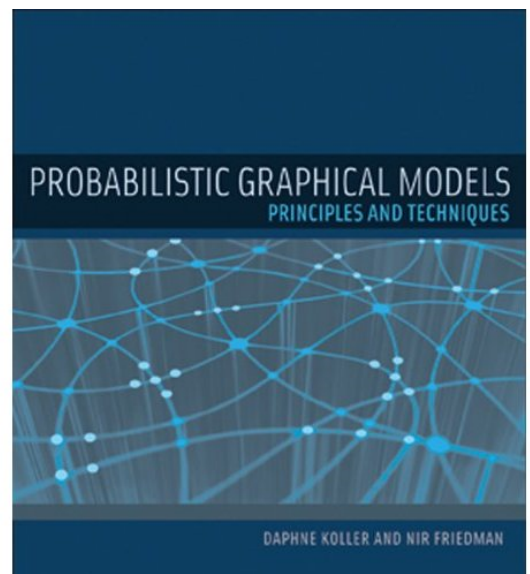




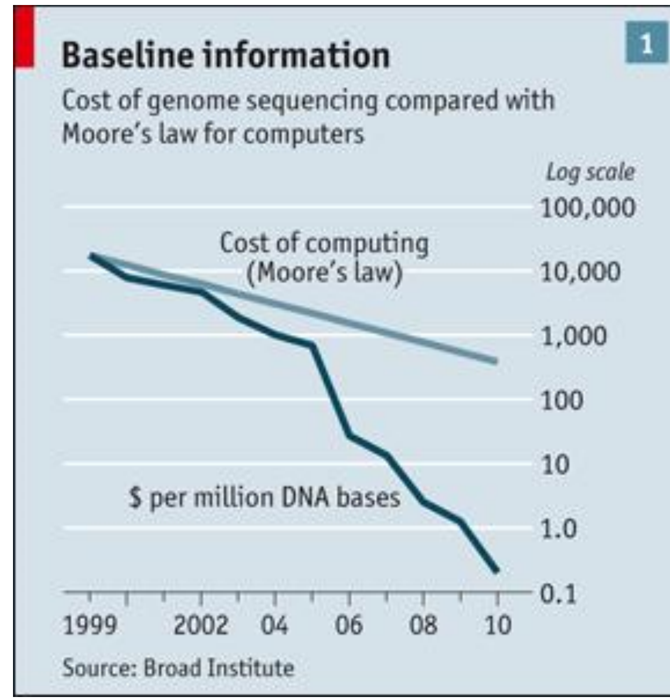








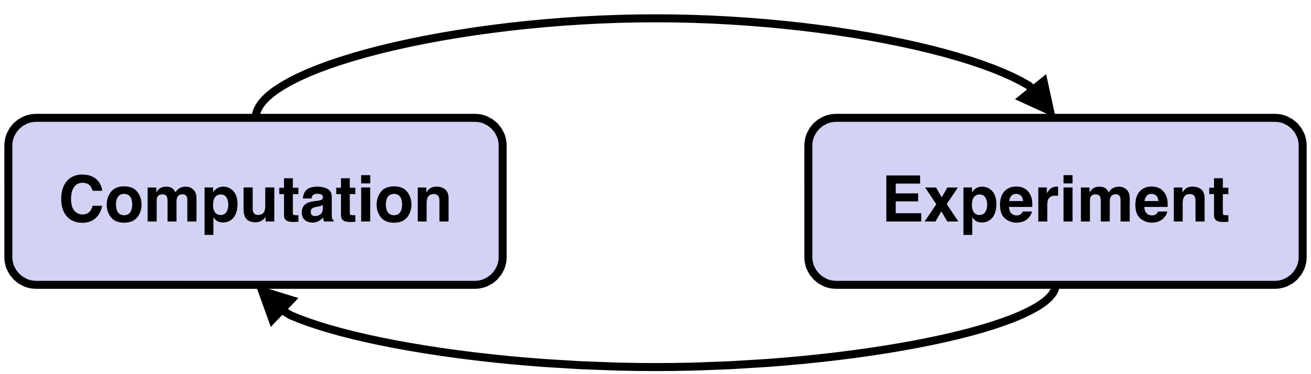
Why Computational Biology?



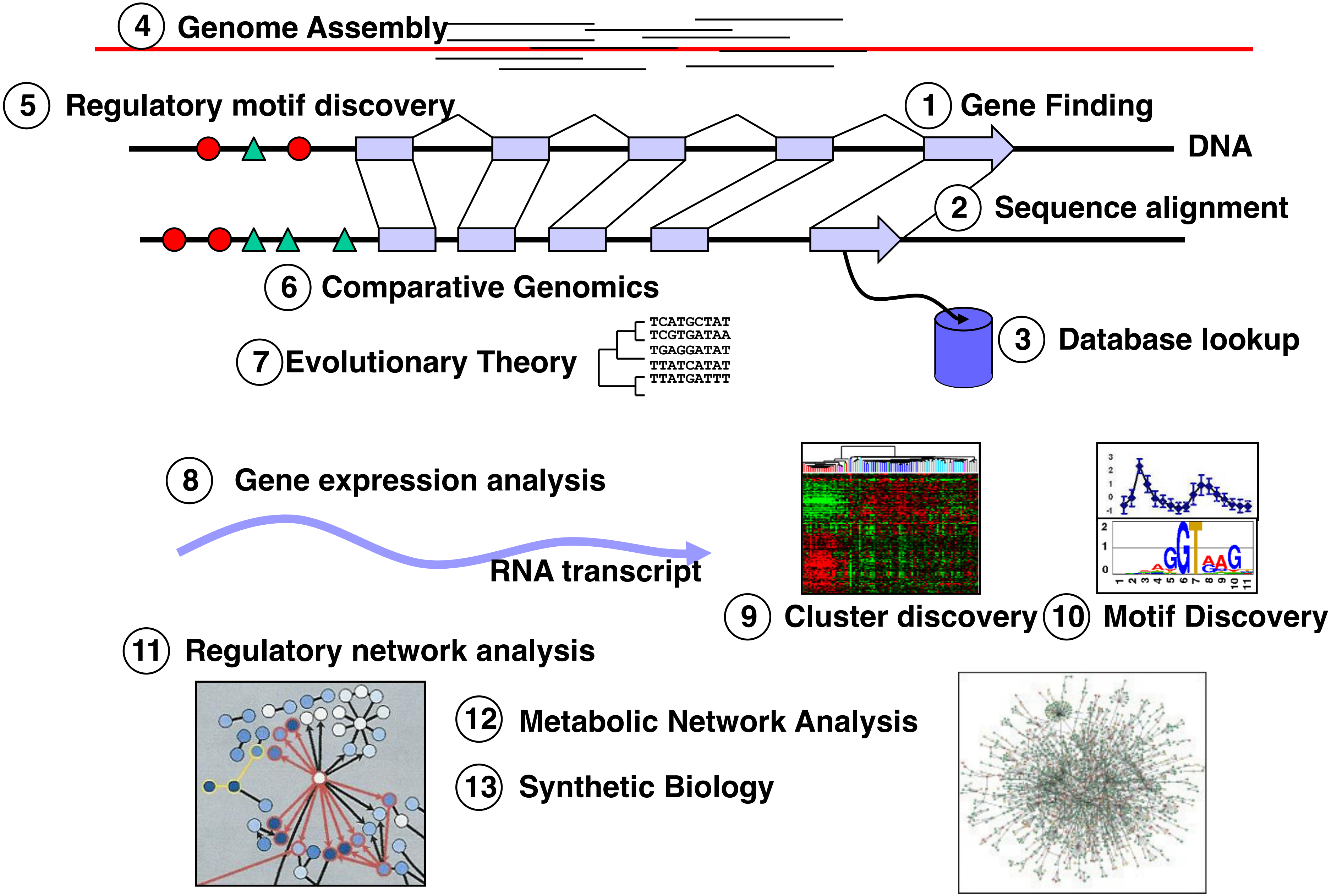
15 years ago it was challenging to sequence a gene
Now we are profiling genes, proteins, etc…
We are awash in data

Biology as Computation

- We can *be* engineers and computer scientist *doing* biology
- We can treat biology as a *computational discipline – information driven*
- Computational analysis of raw data
- Computational understanding, modeling, & interpretation



Challenges in Computational Biology



Lec	Date	Topic	Homework	Week
1	Tuesday, September 03, 2019	Intro Part 1: Probability and Statistics	PS1 Due 9/16	1
2	Thursday, September 05, 2019	Intro Part 2: Probability and Statistics		
3	Tuesday, September 10, 2019	Sequence Alignment		2
	Thursday, September 12, 2019	Bioinformatics Retreat - No lecture		
4	Tuesday, September 17, 2019	Sequence Alignment Part 2	PS2 Due 9/30	3
5	Thursday, September 19, 2019	Clustering		
6	Tuesday, September 24, 2019	Classification		4
7	Thursday, September 26, 2019	Regulatory Motifs/Gibbs Sampling/EM		
8	Tuesday, October 01, 2019	HMMs1 - Evaluation / Parsing	PS3 Due 10/13	5
9	Thursday, October 03, 2019	HMMs2 - PosteriorDecoding/Learning		
10	Tuesday, October 08, 2019	Phylogenetics		6
11	Thursday, October 10, 2019	Molecular Evolution, and Measures of Selection (Tennessen)		
	Tuesday, October 15, 2019	Substitute Monday	Project Proposal Due 11/4	7
12	Thursday, October 17, 2019	Generalized HMMs and Gene Prediction		
13	Tuesday, October 22, 2019	Midterm		8
14	Thursday, October 24, 2019	Midterm Solutions Review		
15	Tuesday, October 29, 2019	Midterm Solutions Review 2		9
16	Thursday, October 31, 2019	Metabolic Modeling 1 - Introduction to FBA		
17	Tuesday, November 05, 2019	Metabolic Modeling 2 - Applications and Incorporation of Omics Data	Project Reviews Due 11/8	10
	Thursday, November 07, 2019	No Lecture		
18	Tuesday, November 12, 2019	Bayesian Networks	Revised Proposal Due 11/18	11
19	Thursday, November 14, 2019	Sampling Methods and Bayesian Networks		
20	Tuesday, November 19, 2019	DREM	Final Project Writeup Due 12/12	12
21	Thursday, November 21, 2019	Conditional Random Fields		
22	Tuesday, November 26, 2019	Mini-Midterm		13
	Thursday, November 28, 2019	Thanksgiving		
23	Tuesday, December 03, 2019	Guest Lecture (TBD)		14
24	Thursday, December 05, 2019	Final Presentations - Part I		
25	Tuesday, December 10, 2019	Final Presentations - Part 2 (9am - Recitation Room)		

But we have to start at the beginning…..

Probability and Statistics Review

Why Probability and Statistics?

- Biological data is noisy
- Probability provides a calculus for manipulating models
- Not limited to yes/no answers – can provide "degrees of belief"
- Many common computational tools based on probabilistic models

Probability and Statistics

- **Probability** – a mathematical framework for calculating with uncertainties
- **Statistics** – applied probability: solutions to specific problems derived using probability calculus

What do probabilities *mean*?

Flipping a coin: what does P(H) = 0.5 mean?

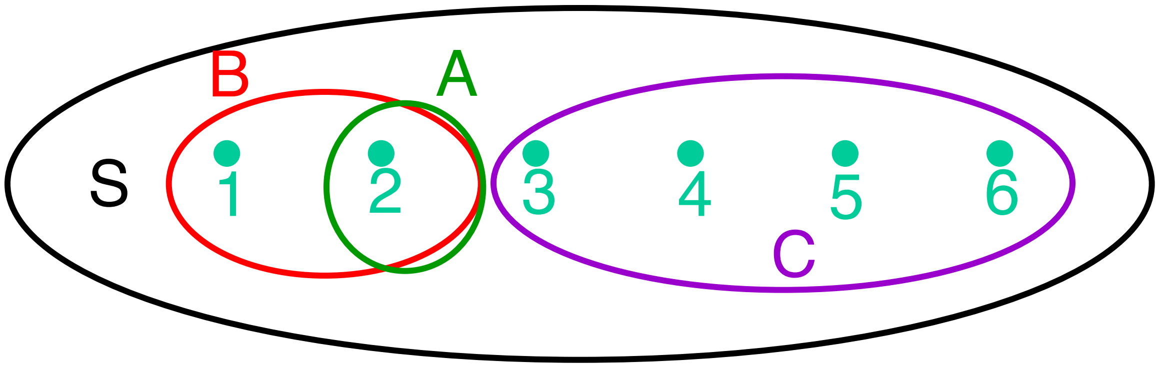
Frequentist

- *Relative frequency* of occurrence - given 100 coin tosses, ~50/100 will be heads

Bayesian

- *Probability of event* – on the next coin toss, 50% chance of heads
- *Belief in event* – given all available data, we have 50% belief in H on next toss

Sample Space and Events



Roll a dice:

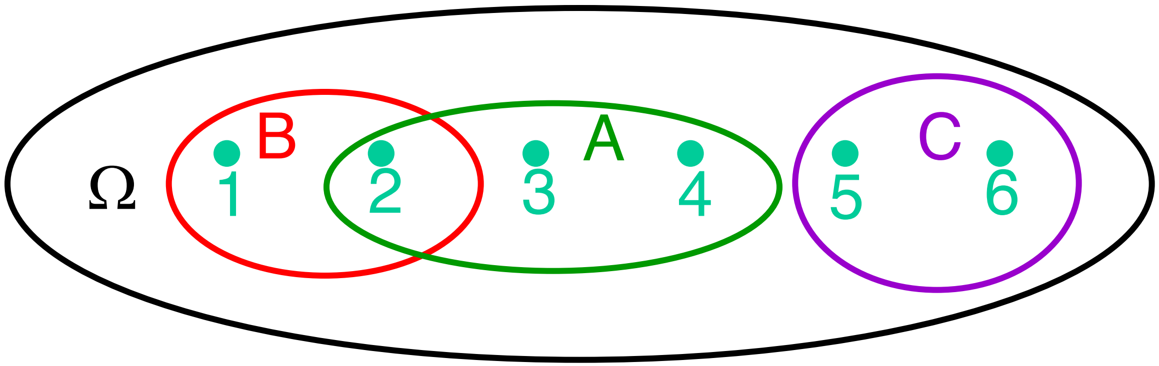
Event A: get a 2

Event B : get < 3

Event C : get > 2

- Sample Space Ω: All conceivable outcomes.
- Sample Point: One conceivable outcome.
- Event : A subset of conceivable outcomes.

Set Theory Algebra

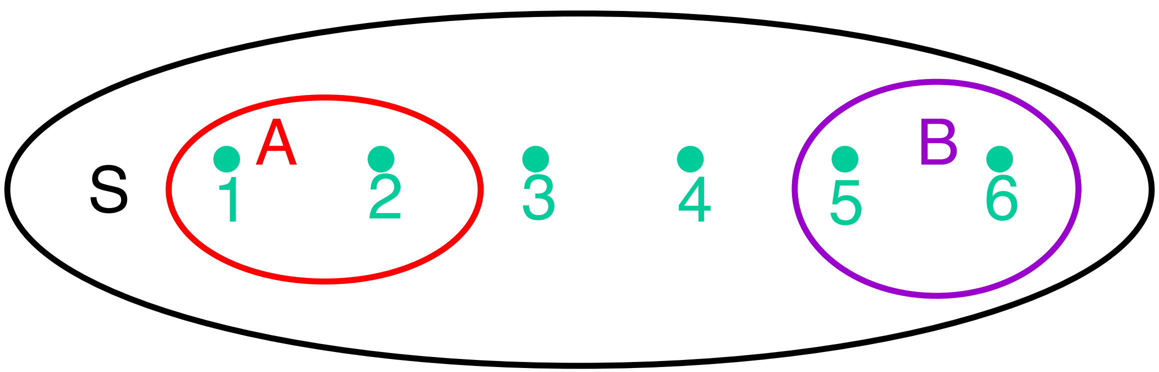


- Union: $D = A \cup B = \{1,2,3,4\}$
- Intersection: $D = A \cap B = \{2\}$
- Complement $A^C = \{1,5,6\}$
- Null Event: $D = A \cap C = \emptyset$, A and C *disjoint*

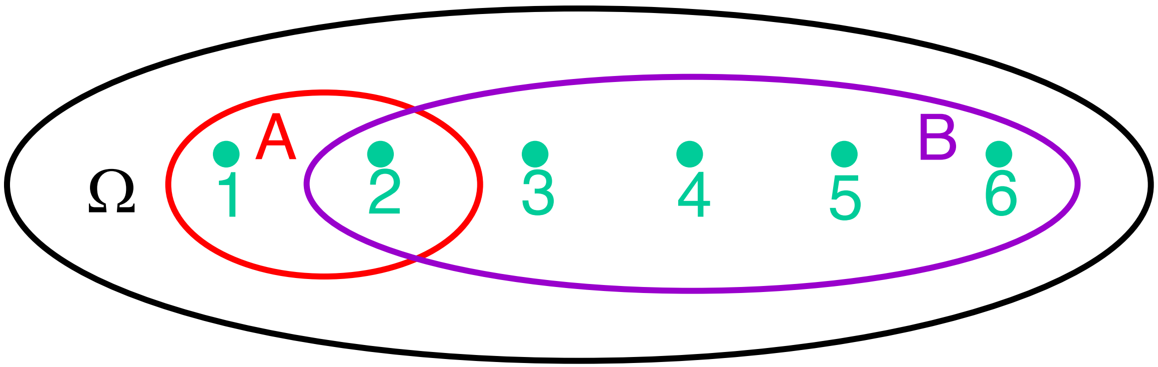
Probability Axioms

A probability measure assigns a real number to each subset of Ω such that:

1. $P(\Omega) = 1$
2. If $A \subset \Omega$, then $P(A) \geq 0$
3. If A and B disjoint, $P(A \cup B) = P(A) + P(B)$



The Addition Law



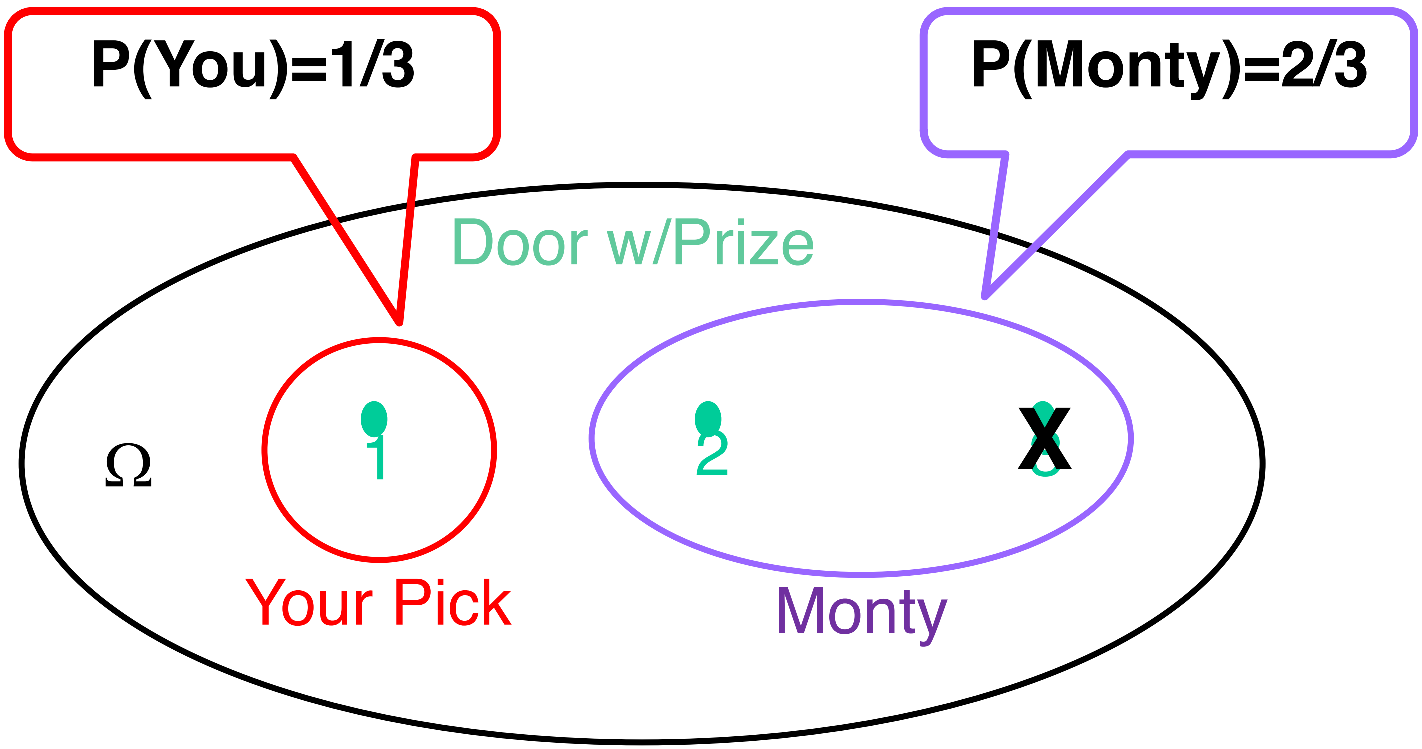
$$P(A \cup B) = ?$$

$$= P(A) + P(B) - P(A \cap B)$$

You will prove this using the axioms in your homework

Monty Hall Problem

Do You Switch Doors? Why/Why Not?



Random Variables

- A random variable, X, represent a set of events
- A probability distribution, P(X), maps each event in X to a real number between 0 and 1
- A joint probability distribution, P(A,B,…C), maps all possible conjunctions of events in {A,B,..C} to a real number between 0 and 1

Example: X1 and X2 are the outcomes of flipping two coins

P(X1)	P(X1, X2)
{X1} = {H,T}	{X1,X2} = {HH,HT,TH,TT}
with p=.50 each	with p=.25 each

If you only have P(X1,X2) can you calculate P(X1)?

Example

P(A,B) assigns a probability to joint event A,B.

Study 100 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

A \ B	Sick	OK
Pos	35	15
Neg	5	45

How do we get P(A,B)?

Normalization

P(A,B) assigns a probability to joint event A,B.

Study 100 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

A \ B	Sick	OK
Pos	.35	.15
Neg	.05	.45

How do we get P(A,B)?

Normalizing means simply dividing by the total so that $P(\Omega)=1$

Marginalization

P(A,B) assigns a probability to joint event A,B. But sometimes we just want P(A)

Study 100 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

A \ B	Sick	OK
Pos	.35	.15
Neg	.05	.45

What are

P(Pos)
P(Neg)

Marginalization

P(A,B) assigns a probability to joint event A,B. But sometimes we just want P(A)

Study 150 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

B / A	Sick	OK
Pos	.35	.15
Neg	.05	.45

What are

P(Pos)
P(Neg)

P(Pos)=0.5
P(Neg)=0.5

Marginalization

P(A,B) assigns a probability to joint event A,B. But sometimes we just want P(A)

For a discrete variable

$$\boxed{P(A) = \sum_{\text{all } B} P(A,B)}$$

Conditional Probability

We are often interested in the probability of one event (A) given another event (B)

This is called the *conditional probability* of A given B

P(A I B)

Conditional Probability

Study 100 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

B / A	Sick	OK
Pos	35	15
Neg	5	45

What is

P(Sick|Pos)?

 # ways to be positive = 50
 # ways to be positive and sick = 35
35/50=0.7

Conditional Probability

Study 100 Patients

Pos = Positive Disease Test
Neg = Negative Test
Sick, Ok = Health Status

A \ B	Sick	OK
Pos	.35	.15
Neg	.5	.45

What is

P(Sick|Pos)?

Normalized
P(Pos) = .50
P(Pos,Sick) = .35
P(Sick|Pos) = .35/.50=0.7

Conditional Probability

In general, if we are given P(A,B) how do we calculate the conditional probability of A given B, P(A|B)?

$$\boxed{P(A \mid B) = \frac{P(A,B)}{P(B)}}$$

The Chain Rule

$$\frac{P(A,B)}{P(B)} = P(A \mid B)$$

$$\Downarrow$$

$$\boxed{P(A,B) = P(A \mid B)P(B)}$$

You must know this!

But what if...

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

In this case the probability of A does not depend on B

A and B are independent variables

$$\boxed{A \perp B}$$

Independence

If $A \perp B$, then

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

$$P(A,B)=P(A)P(B)$$

All of this follows from the chain rule

Eugene…

Eugene…

P(Description|Dean)>P(Description|Truck Driver)

This is called the Likelihood of the data

P(Truck Driver)>>P(Dean)

These are the Prior Probabilities

What we want are

P(Dean|Description) & P(Truck Driver|Description)

Bayes Rule

Likelihood

Prior

$$P(Class \mid Feature) = \frac{P(Feature \mid Class)P(Class)}{P(Feature)}$$

Posterior

Evidence

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

Bayes Rule

Evaluate evidence

Belief before evidence

$$P(Class \mid Feature) = \frac{P(Feature \mid Class)P(Class)}{P(Feature)}$$

Belief after evidence

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

Bayes Rule

Evaluate evidence

Belief before evidence

$$P(Class \mid Feature) = \frac{P(Feature \mid Class)P(Class)}{P(Feature)}$$

Belief after evidence

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

Bayes Rule

You should be able to see this follows from the chain rule…

$$P(Class \mid Feature) = \frac{P(Feature \mid Class)P(Class)}{P(Feature)}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

Rules You Need to Know

$$P(\Omega) = 1$$

If $A \subset \Omega$, then $P(A) \geq 0$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(A) = \sum_{\text{all } B} P(A,B)$$

$$\text{If } A \perp B,\ P(A \mid B) = P(A)$$

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

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