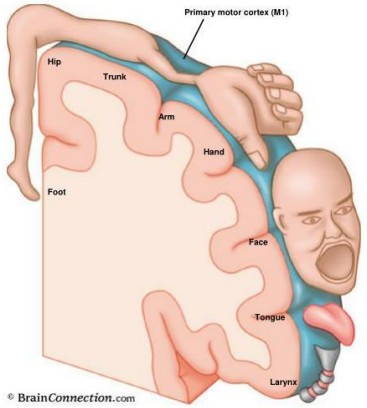
CSC196: Analyzing Data

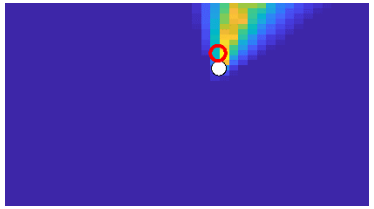
Course Introduction + Overview

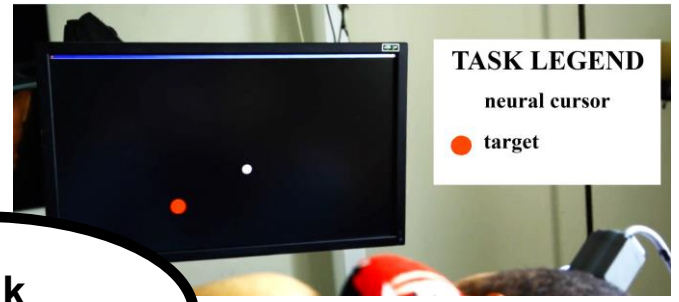
Jason Pacheco and Cesim Erten

Block 12: "Multiscale Semi-Markov Model"

Probabilistic Reasoning







Prior Knowledge

Uncertainty Quantification

Task Execution

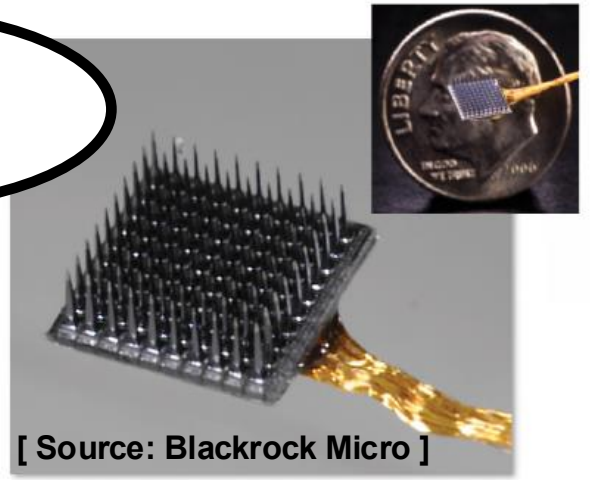
Representation

x_{t-1} x_t x_{t+1}

y_{t-1} y_t y_{t+1}

Inference

Information Gathering

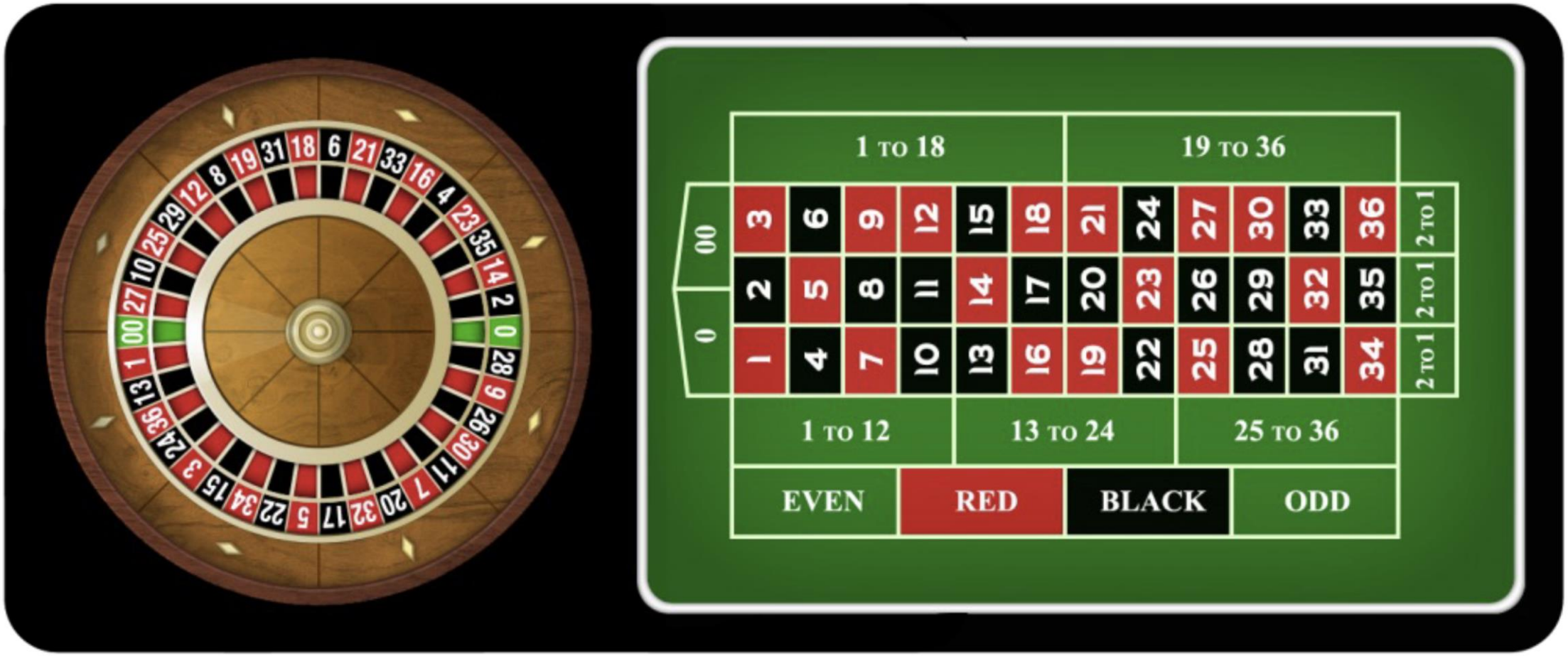


Spam Filtering

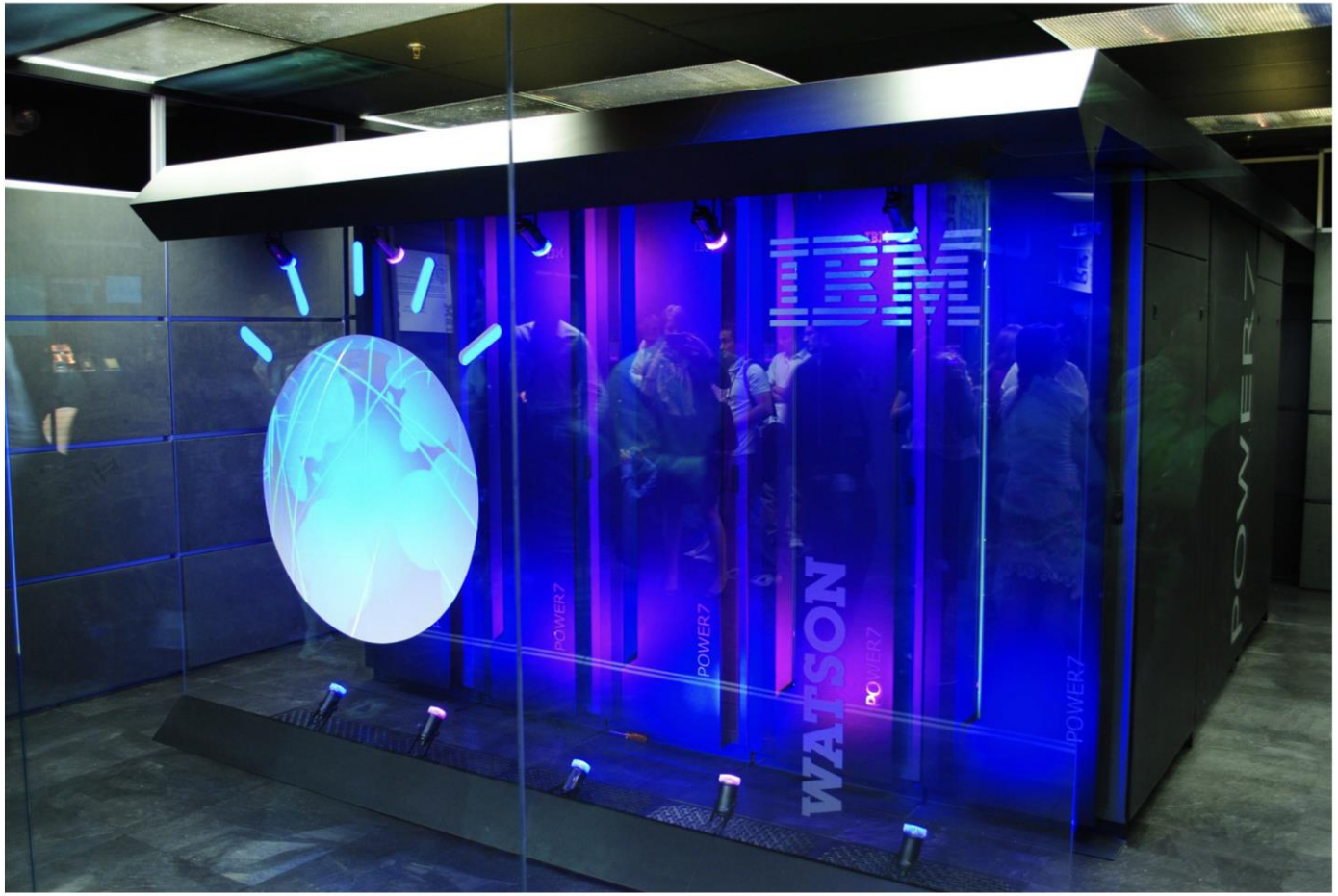
- Binary classification: Is this email useful (ham) or not (spam)?
- Noisy training data: Messages previously marked as spam
- Information: Probability that certain words are used in spam and non-spam emails
- Information: Probability that certain servers send spam

Image Source: https://www.ubisec.com/blog/what-is-email-filtering-and-how-email-filtering-boosts-productivity/

Maximizing Expected Reward

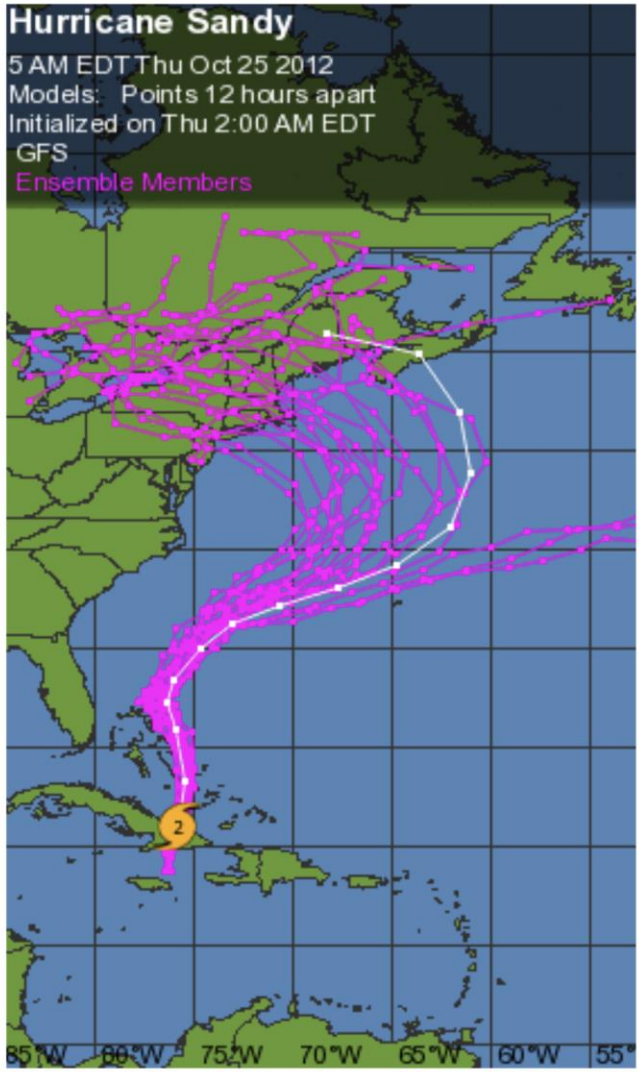


IBM’s Watson

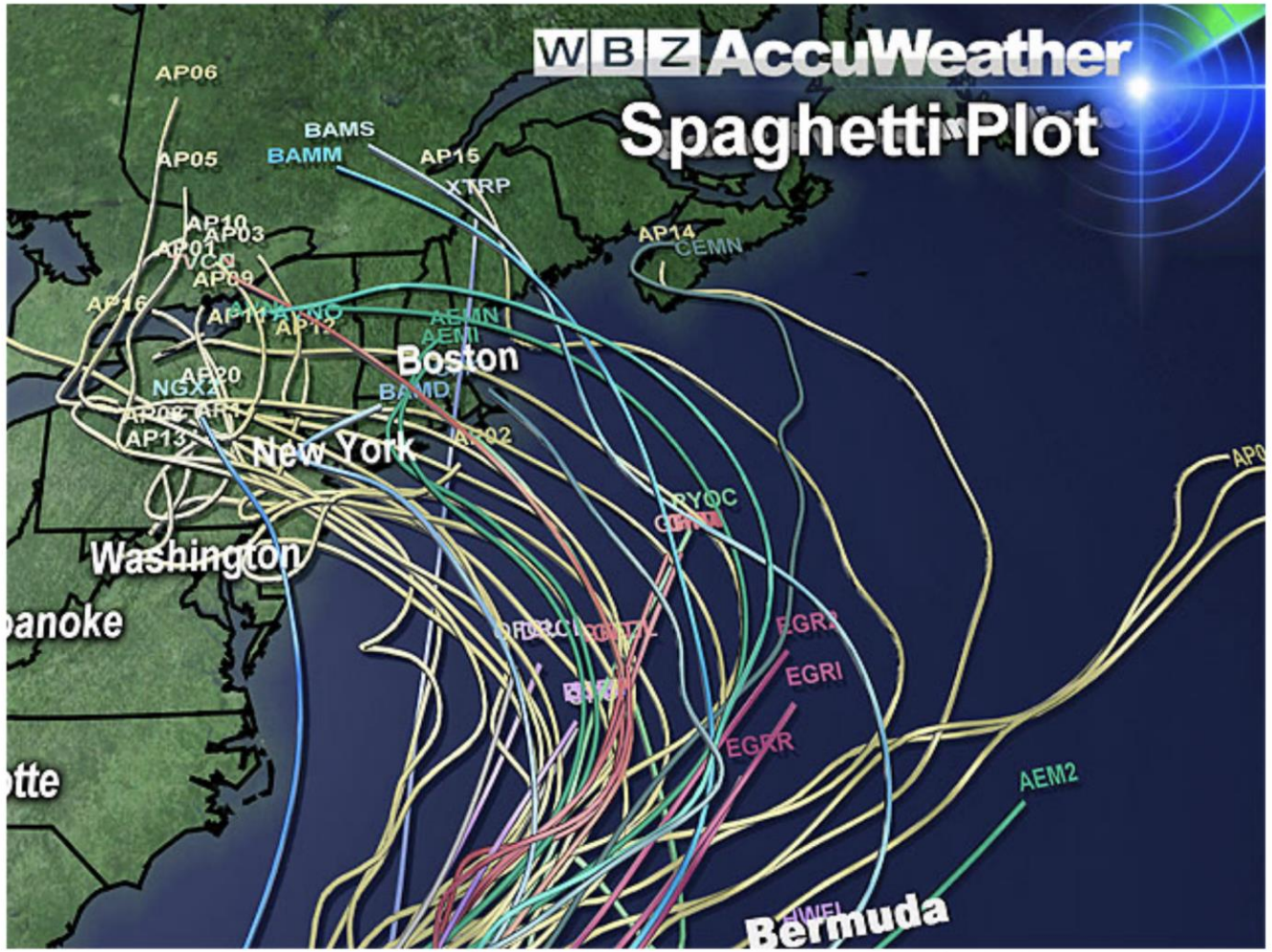

IBM
WATSON
POWER7


THINK
DENKE
$300,000
$1,000,000
$200,000
Ken
WATSON
BRAD

Monte Carlo Methods



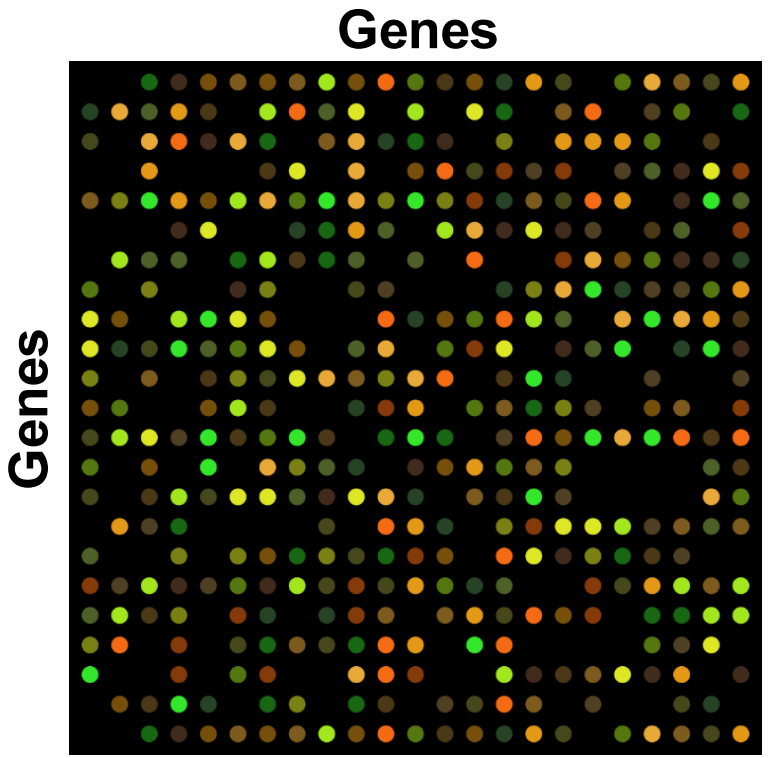
Weather Wisdom, Boston.com



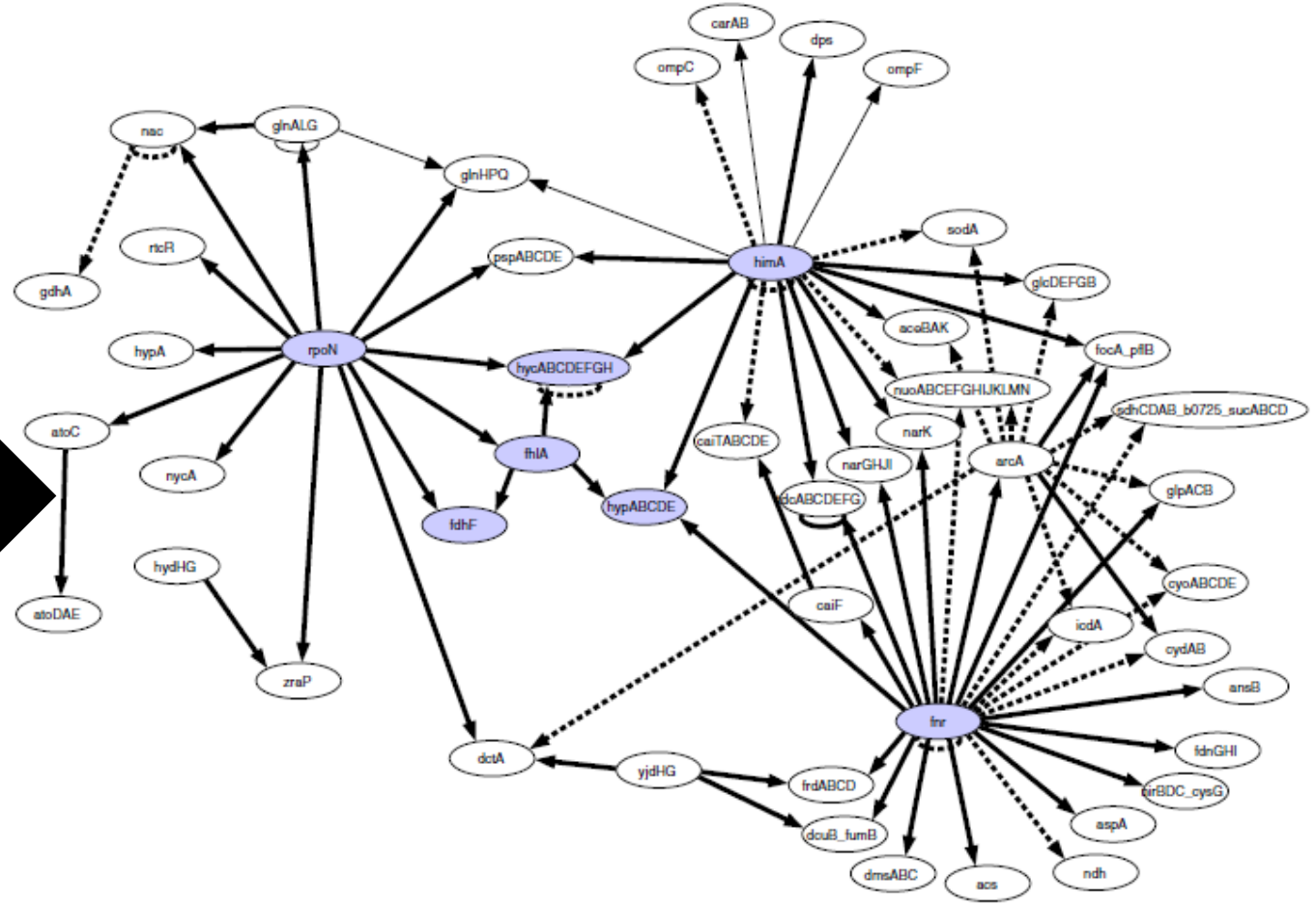
Hurricane Sandy made landfall in New Jersey on October 29, 2012.

Example: Gene Regulatory Network Inference

Gene Expression



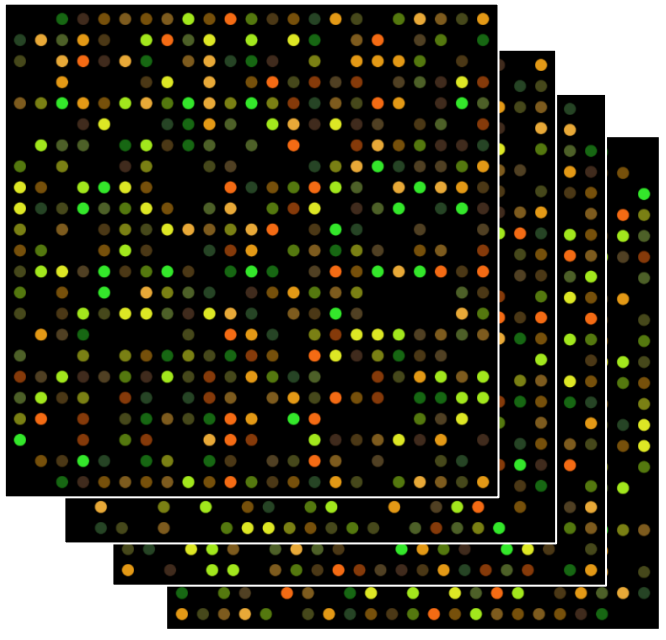
Regulatory Network



Goal: Estimate causal interaction network from expression data.

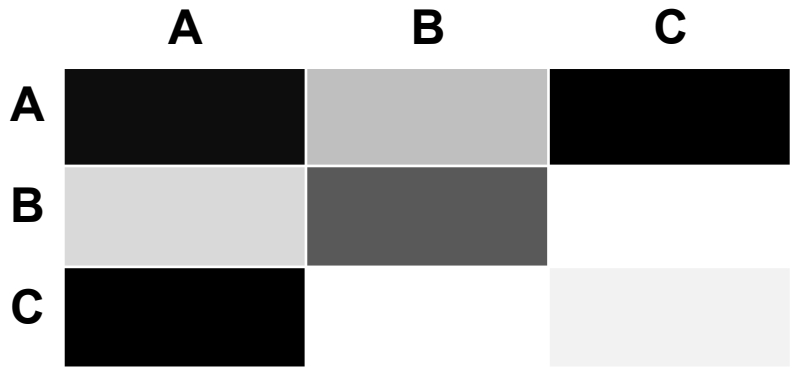
[Image: Bulcke et al., 2006]

Identifying Causality

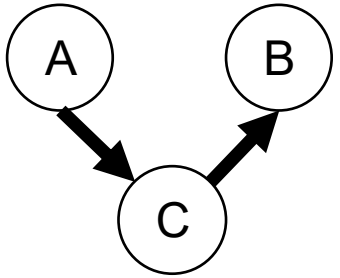


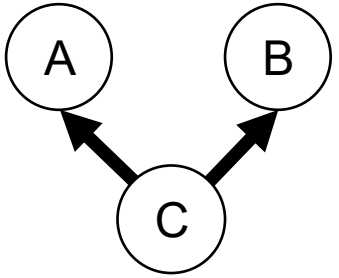
Dataset

Covariance Matrix

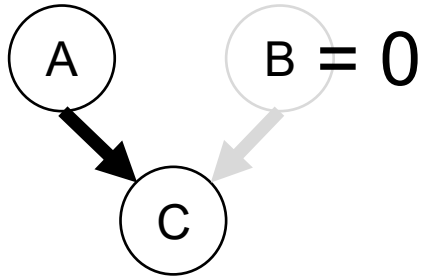


Possible Graphs



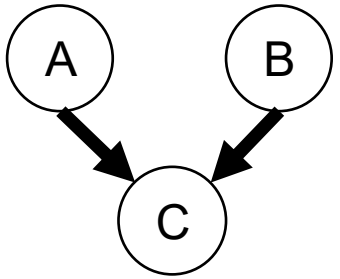


Cannot determine causality from correlations, need to perform active interventions …

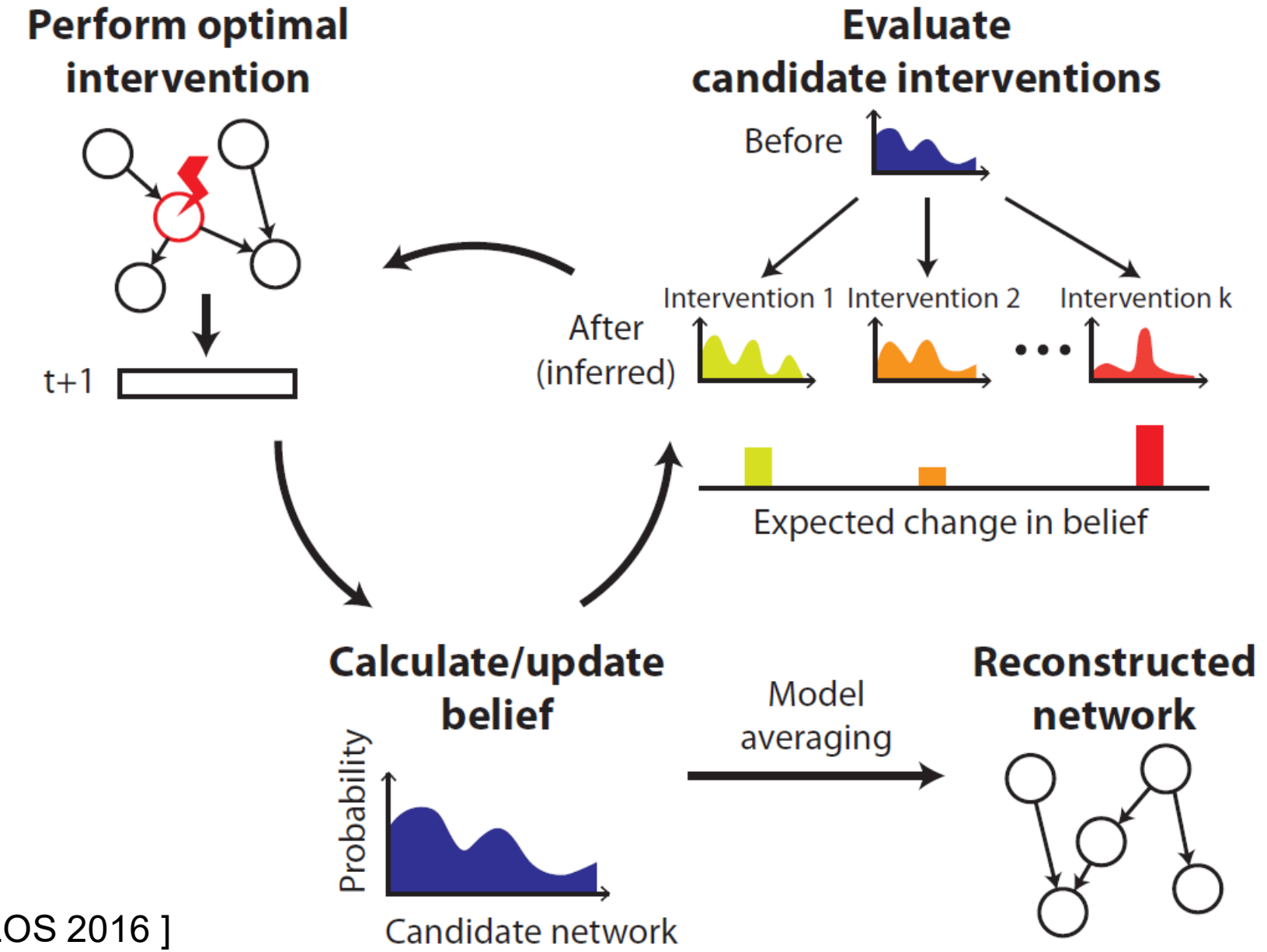


Clamp node to fixed value.

Gene Knockout

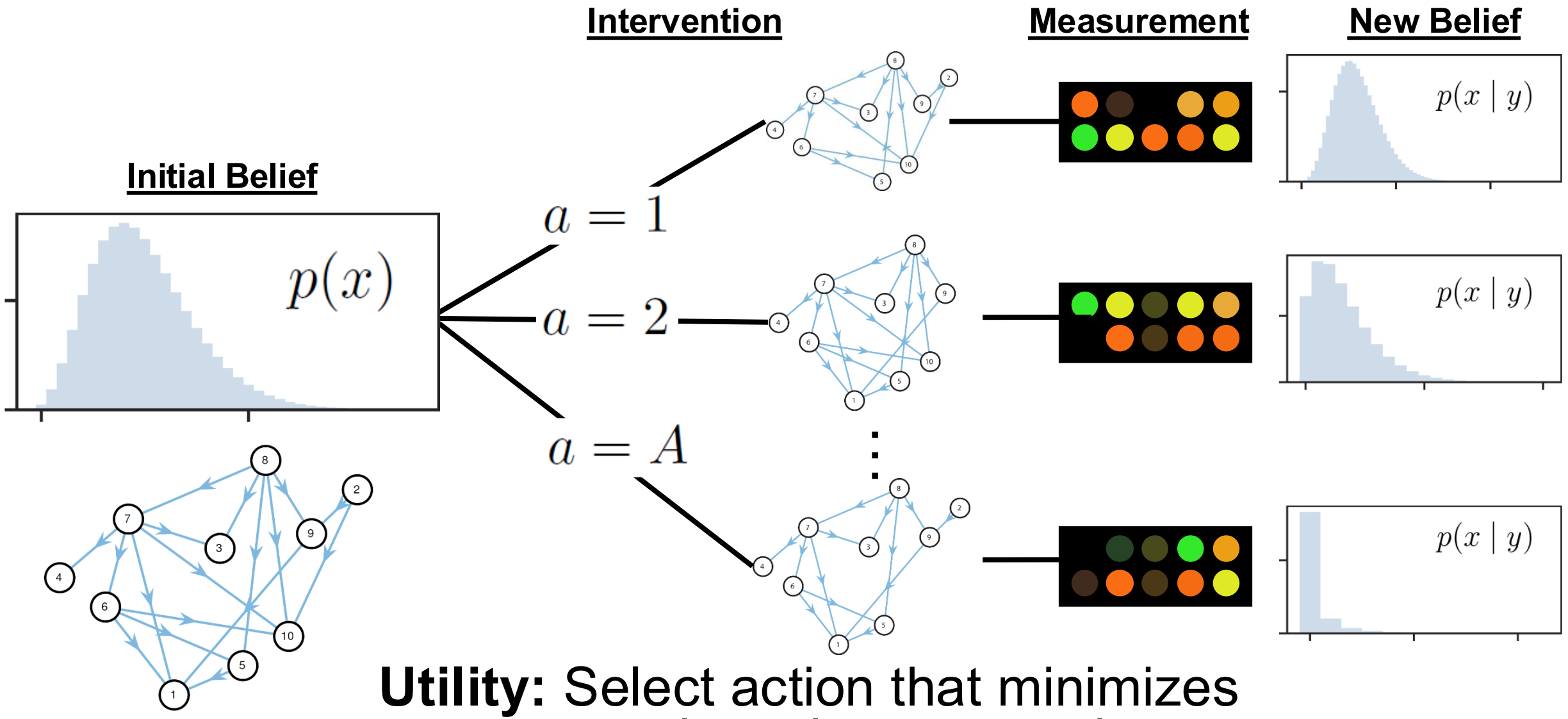


Optimal Experiment Design



[Source: Cho et al., PLOS 2016]

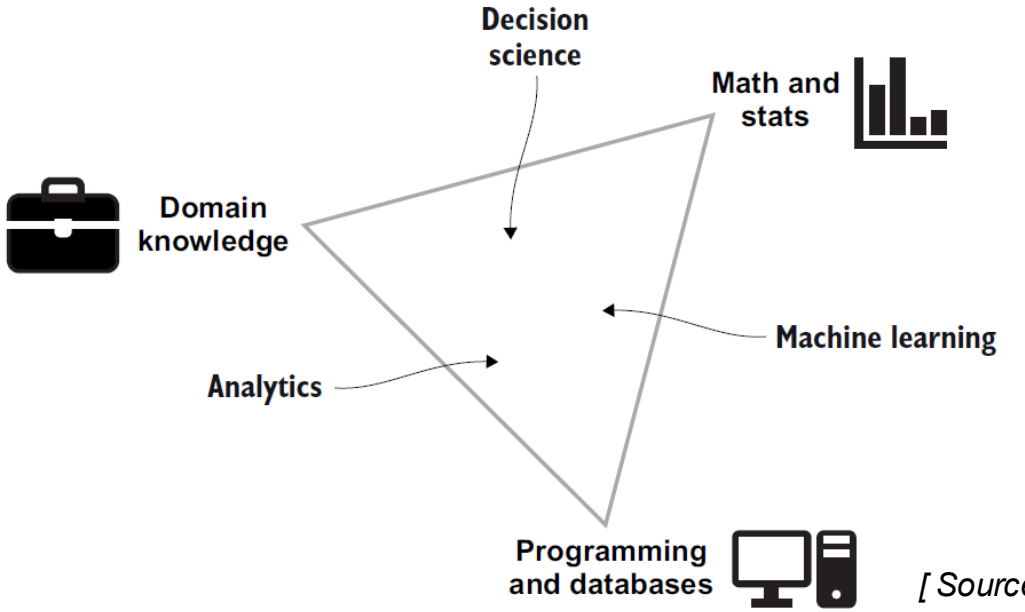
Choosing Experiments



Utility: Select action that minimizes uncertainty about network

Data Analysis

Definition “Data analysis” is the process of inspecting, cleaning, transforming, and modeling raw data to discover useful information.



[Source: Robinson, E. and Nolis, J.]

It involves techniques from statistics, computer science, and math to turn data into actionable insights for businesses, science, and various fields.

Summary of Topics

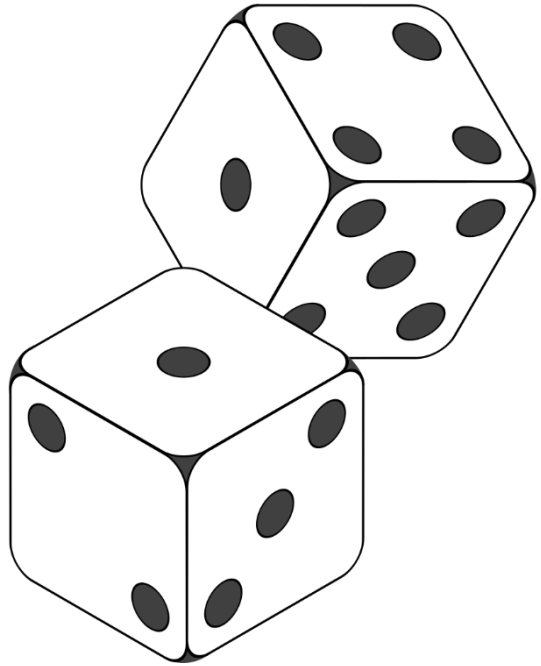
- Introduction to Statistics and Data Analysis
- Probability
- Random Variables and Probability Distributions
- Expectation and Moments of Random Variables
- Concepts of Calculus
- Continuous Probability
- Fundamental Sampling Distributions
- Statistical Estimation
- Bayesian Statistics

Probability and Statistics

Suppose we roll <u>two fair dice</u>…

- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?

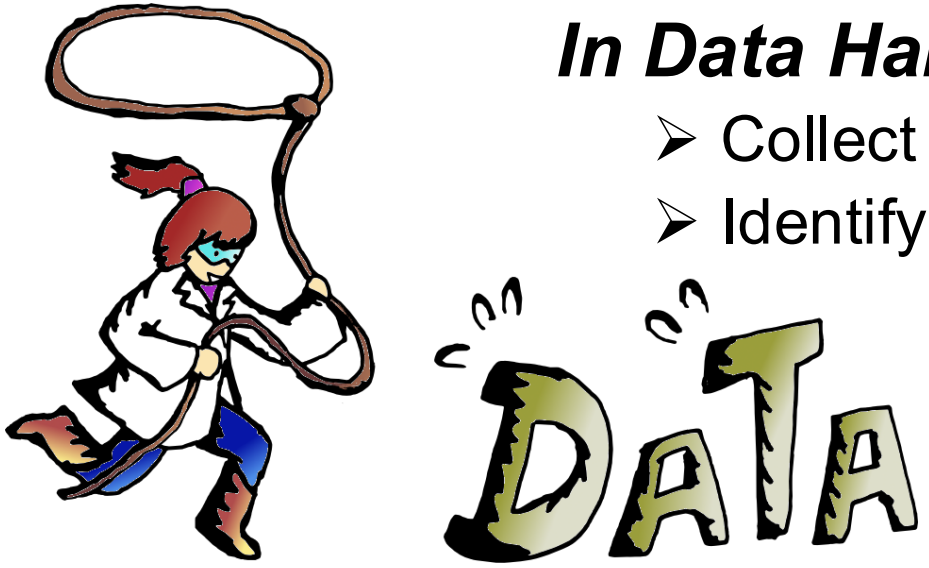
*… this is an **experiment** or **random process**.*



We will learn how to…

- Mathematically formulate outcomes and their probabilities?
- Describe characteristics of random processes
- Estimate unknown quantities (e.g. are the dice actually fair?)
- Characterize the uncertainty in random outcomes
- Identify and measure dependence among random quantities

Data Handling and Visualization



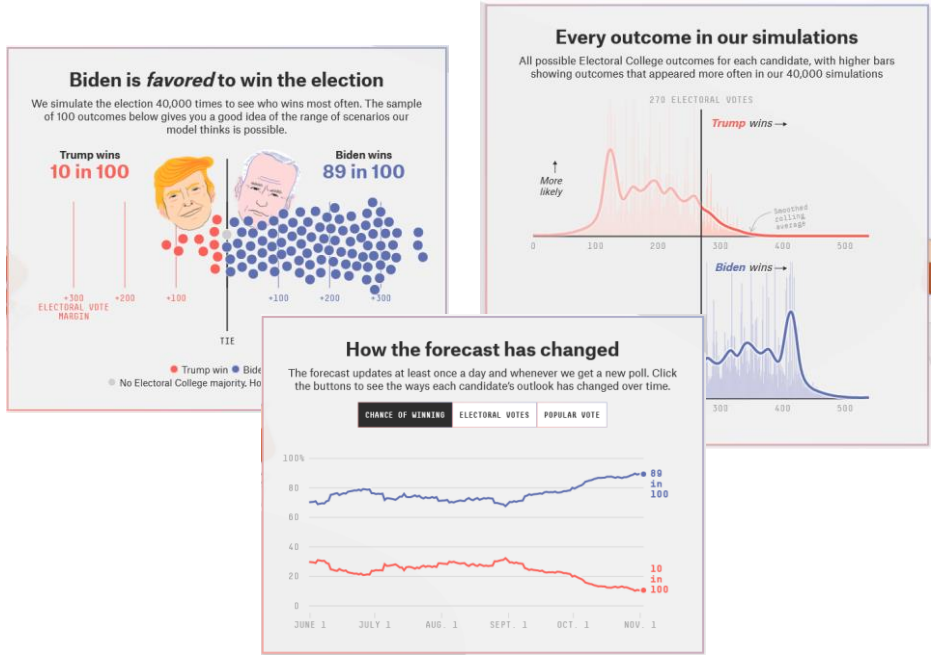
[Image Source: Code A Star]

In Data Handling we will learn to…

- Collect data through population sampling
- Identify and avoid biased population samples
- Clean data and correct errors
- Transform and preprocess data (*wrangling*)

In Data Visualization we will learn…

- Why visualization is important
- Exploratory data analysis
- Common forms of visualization
- Pitfalls and gotchas



Python Packages

Widely used in science / engineering applications. Contains multidimensional array data structures (ndarray), vectors, matrices, and functions to operate on them.

Comprehensive library for creating static, animated, and interactive visualizations.

We will focus on scipy.stats, which contains many probability distributions, summary statistics, correlation functions and statistical tests, and more.

Course Prerequisites

Programming

- CSC 110
- Python programming needed for homework assignments
- Ideally some exposure to Numpy

Math

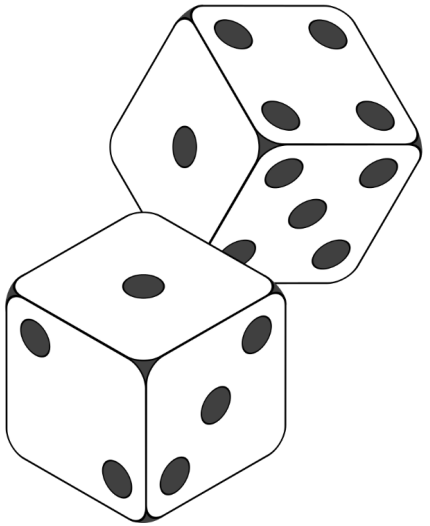
- Optional / Recommended: MATH 122B or MATH 125B
- Single variable calculus necessary for continuous probability
- We will review concepts of calculus as a refresher

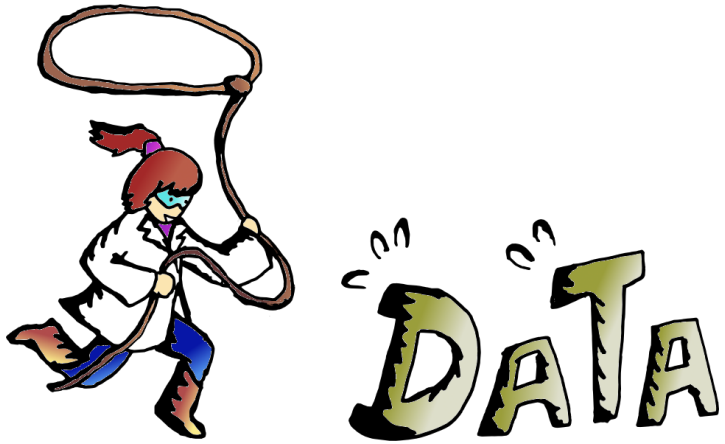
Course Overview: Resources

Resources accessible on course website
http://pachecoj.com/courses/csc196_spring26/

Specific resources

- Gradescope for assignment submission
- Piazza for **all** communication
- Readings and electronic textbooks
- Lecture slides

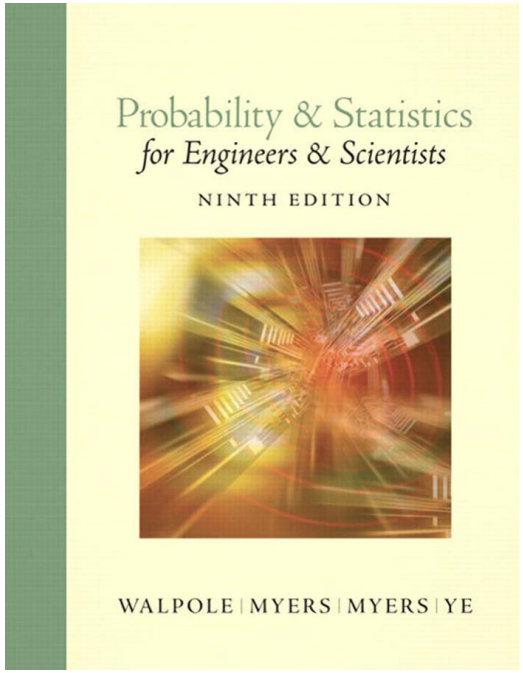




Every lecture accompanied by reading

- See course webpage for readings
- These will not be graded but are required
- Homeworks will incorporate material from readings

Textbook



Walpole et al. “Probability & Statistics for Engineers & Scientists 9th Ed.” 2012 (Available via D2L)

Additional readings on the course webpage

Course TAs

Your friendly course graduate TAs…

Yinan Li

yinanli@arizona.edu

Alonso Granados Baca

alonsog@arizona.edu

Assignments / Exams / Grading

11 Homeworks (worst dropped) + 2 Midterms (worst dropped) + Final Exam

Homeworks

- Generally, you will have 1 week per assignment
- There will be an assignment nearly every week
- Grades by one week after due date
- Some irregularity around holidays / exams
- No assignment over spring break

Grading Breakdown

- Homework: 20%
- Quizzes: 15%
- Midterm: 30%
- Final: 35%

First assignment out one week from today

- There will be an in-class quiz **most weeks**
- Questions will be from previous lecture slides
- No makeups so be sure to attend lecture

Late Policy

Late submissions impact other students and delays grading

But sometimes we need a little extra time…

- **No more than 1** assignment **no more than 1** day late without penalty
- All subsequent late assignments will receive a zero score
- D2L will accept late assignments but they will be flagged

If you are struggling with time…

- Notify us (Piazza) at least 24hrs before the deadline
- Submit the best version of what you have by the deadline
- In general I **will not** grant extra time, and will grade what has been submitted

*If you submit **all** assignments on time, it may benefit your final grade*

Academic Integrity

Assignments are to be done independently…

If I or the TA suspects you of having cheated

- You will be notified immediately
- We will have a conference where you can plead your case
- If I am not swayed then you receive a zero for the assignment
- There is an appeals process if you are confident in your case

Bottom line don't cheat

Office Hours

- Cesim : Mon / Wed : 2:15-3:15pm : Gould-Simpson Room 845
- Jason : Mon / Wed : 2:15-3:15pm : Gould-Simpson Room 707
- Yinan : Tue / Thurs : 10:00-11:00am : Zoom
- Alonso : Fri : 10:00am – 12:00pm : Zoom

Questions?