What is Probability and Statistics and Why Should You Care?

CS 3130: Probability and Statistics for Engineers

January 9, 2023

What is Probability?

What is Probability?

Definition

Probability theory is the study of the mathematical rules that govern random events.

What is Probability?

Definition

Probability theory is the study of the mathematical rules that govern random events.

But what is randomness?

What is Probability?

Definition

Probability theory is the study of the mathematical rules that govern random events.

But what is randomness?

Informally, a **random event** is an event in which we do not know the outcome without observing it.

What is Probability?

Definition

Probability theory is the study of the mathematical rules that govern random events.

But what is randomness?

Informally, a **random event** is an event in which we do not know the outcome without observing it.

Probability tells us what we can say about such events, given our assumptions about the possible outcomes.

Reasoning About Unknown Events (e.g. Future)

What is Statistics?

What is Statistics?

Definition

Statistics is the application of probability to the collection, analysis, and description of random data.

What is Statistics?

Definition

Statistics is the application of probability to the collection, analysis, and description of random data.

Statistics is used to:

- **Design** experiments
- **Summarize** data
- **Draw conclusions** about the world
- **Explore** complex data

Applications of Probability and Statistics

Computer Science: Electrical Engineering:

- Machine Learning
- Data Mining
- Artificial Intelligence
- Simulation
- Image Processing
- Data Management
- Visualization
- Software Testing
- Algorithms

Applications of Probability and Statistics

Computer Science:

- Machine Learning
- Data Mining
- Artificial Intelligence
- Simulation
- Image Processing
- Data Management
- Visualization
- Software Testing
- Algorithms

Electrical Engineering:

- Signal Processing
- Telecommunications
- Information Theory
- Control Theory
- Instrumentation, Sensors
- Hardware/Electronics Testing

Applications of Probability and Statistics

General:

- Gambling

Applications of Probability and Statistics

General:

- Gambling (not recommended)

Applications of Probability and Statistics

General:

- Gambling (not recommended)
- Stock Market Analysis
- Politics
- Sports
- Demographics
- Medicine
- Economics

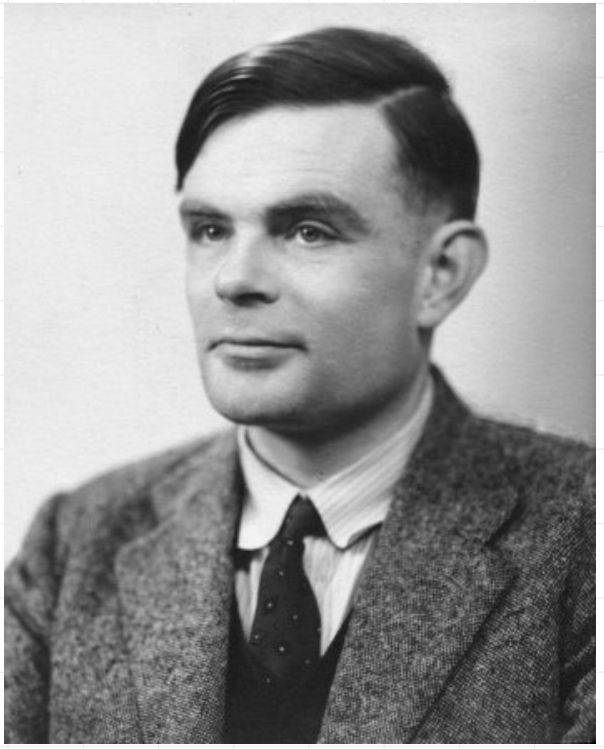
Applications of Probability and Statistics

General:

- Gambling (not recommended)
- Stock Market Analysis
- Politics
- Sports
- Demographics
- Medicine
- Economics
- All (Data) Sciences!!

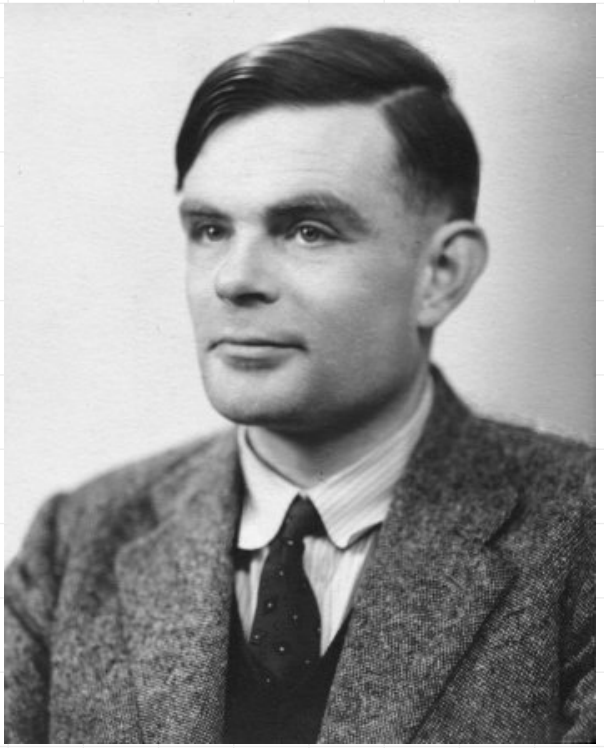
Alan Turing: Connecting CS and Probability

- "Father of Computer Science"
- Most famous for:
  - Computability, Turing machine
  - Stored-program computer
  - Turing test
  - WWII cryptanalysis



Alan Turing: Connecting CS and Probability

- "Father of Computer Science"
- Most famous for:
  - Computability, Turing machine
  - Stored-program computer
  - Turing test
  - WWII cryptanalysis
- Wrote a dissertation on probability theory!
- Turing used probability and statistics to crack Enigma



Application: Machine Learning

Machine Learning builds statistical models of data in order to recognize complex patterns and to make decisions based on these observations.

Core tasks:

- Classification (recognition of street signs or cancer)
- Prediction (elections, movie preferences)

Application: Randomized Algorithms

- Some algorithms benefit from using random steps rather than deterministic ones

Application: Randomized Algorithms

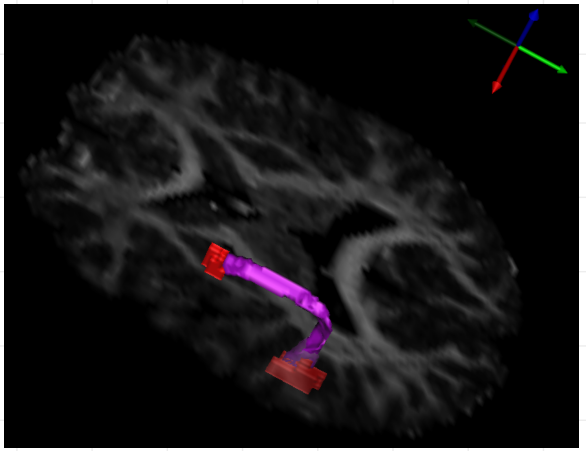
- Some algorithms benefit from using random steps rather than deterministic ones
- Example: QuickSort
  - One of the simplest & fastest sorting algorithms
  - Divide and Conquer: splits data based on **random** pivot
  - Takes $O(n \log n)$ time *in expectation*.

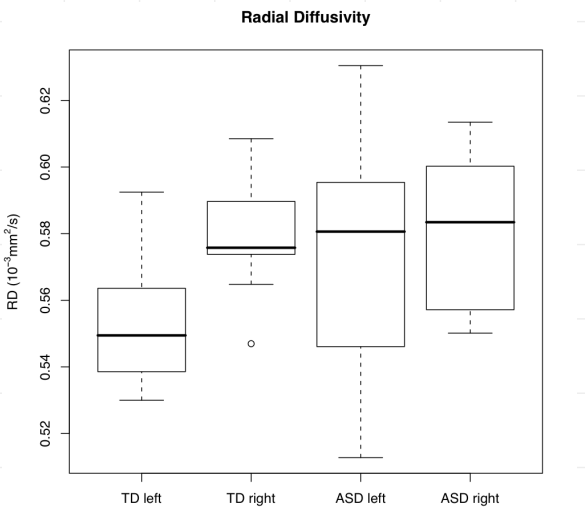
Application: Randomized Algorithms

- Some algorithms benefit from using random steps rather than deterministic ones
- Example: QuickSort
  - One of the simplest & fastest sorting algorithms
  - Divide and Conquer: splits data based on **random** pivot
  - Takes $O(n \log n)$ time *in expectation*.
- Example: stochastic optimization methods
  - Gradient descent optimizes cost functions: workhorse of machine learning
  - On large data sets (100s millions data points), just computing gradient is infeasible
  - Stochastic GD computes gradient on **random sample**: faster & more robust

Application: Medical Image Analysis

- Must deal with noisy image data
- Example: finding an anatomical structure in a 3D image
- Often includes statistical analysis of resulting data

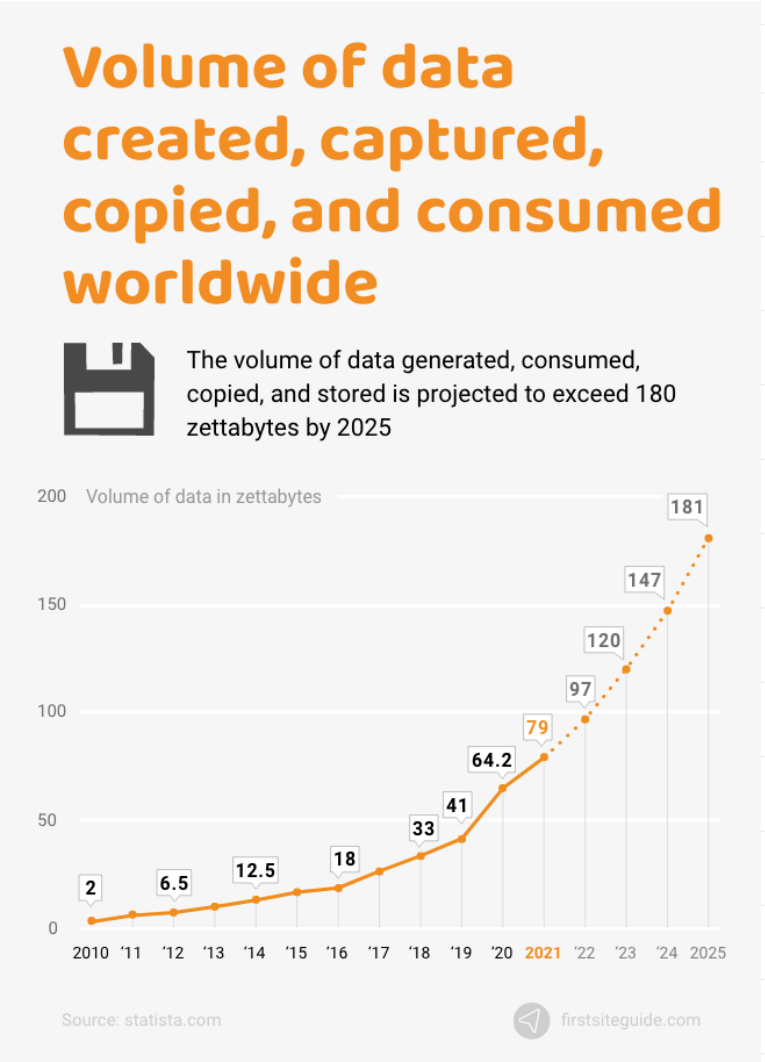




Fletcher et al, NeuroImage, 2010

Big Data & Analytics

- The amount of digital data is exploding!
- Big data analysis is statistics + scalable CS.
- coresets and sketches (often randomized)



How Much is an Exabyte?



How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately
500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

Thus it will take **530,785,562,327** trees to store an Exabyte of data

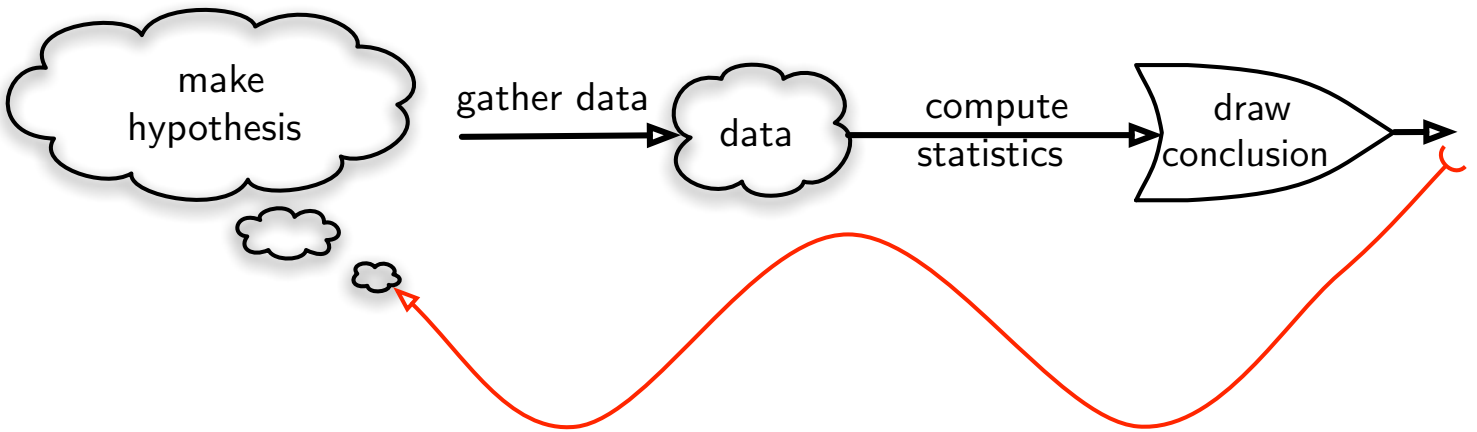
In 2005, there were **400,246,300,201** trees on Earth

We can store **.75** Exabytes of data using all the trees on the entire planet.

Sources: http://www.whatsabyte.com/ and http://wiki.answers.com
(slide by Chris Johnson)

Note: 1 Zettabyte is 1000 exabytes

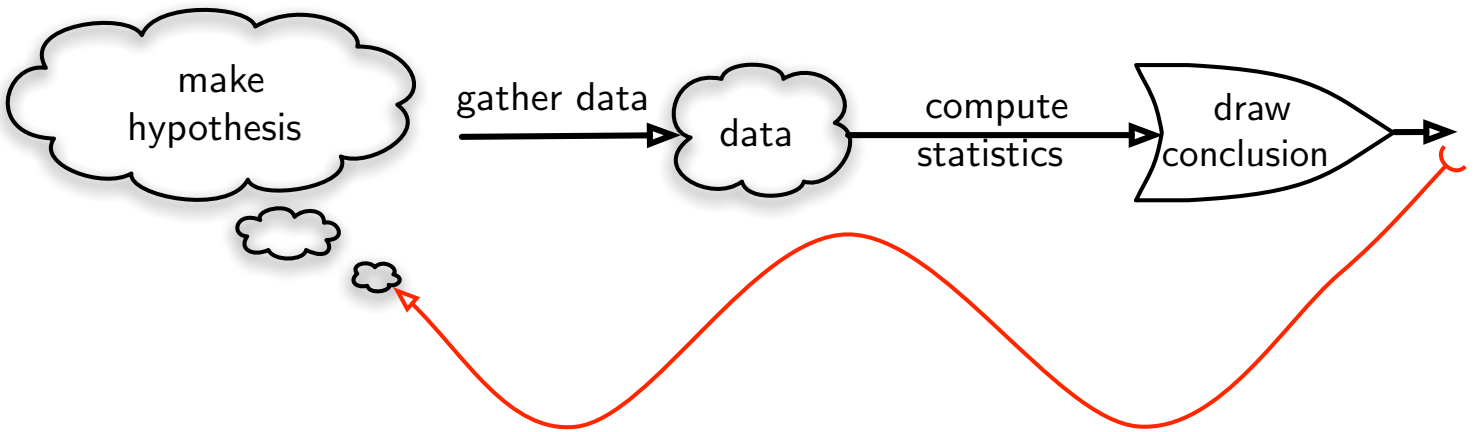
The Scientific Method



1. Define the question
2. Background research, observation
3. Formulate a hypothesis
4. **Design and run an experiment**
5. **Analyze the results**

Experimental measurements are noisy (randomness).

The Scientific Method

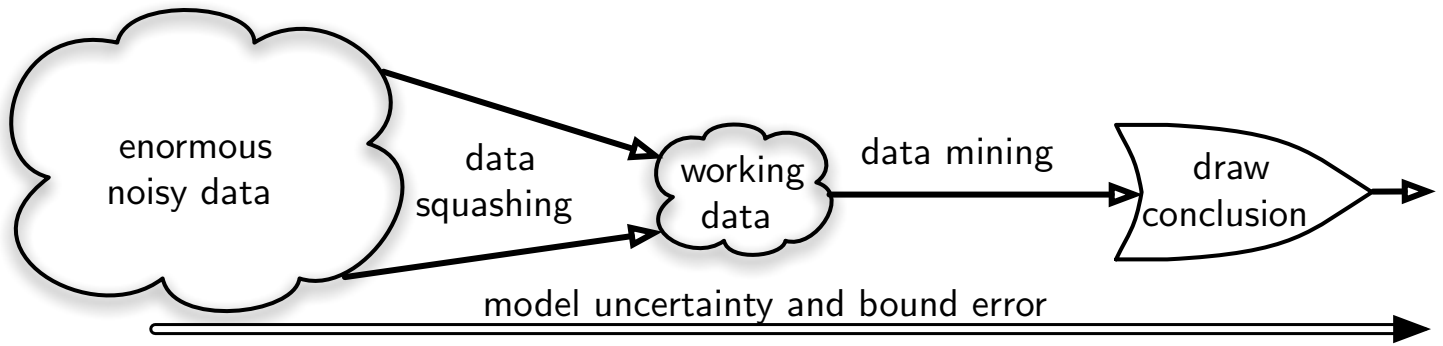


1. Define the question
2. Background research, observation
3. Formulate a hypothesis
4. **Design and run an experiment**
5. **Analyze the results**

Experimental measurements are noisy (randomness).

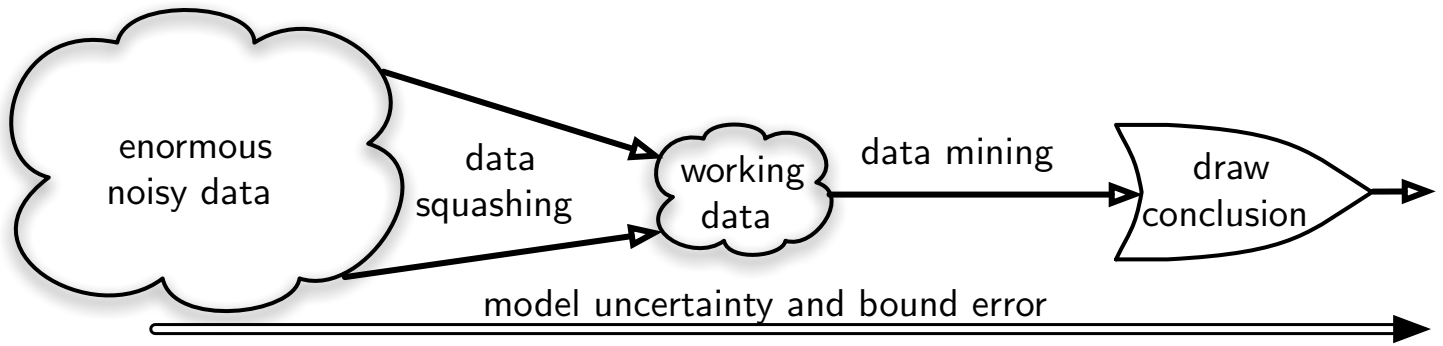
Statistics is critical in the last *two* steps!

Data Science



1. Process/squash enormous available data
2. Mine working data (calculate many statistics)
3. Analyze the results / Draw conclusions

Data Science



1. Process/squash enormous available data
2. Mine working data (calculate many statistics)
3. Analyze the results / Draw conclusions

Every step is subject to noise and involves statistics.

Prob/Stat — A Way of Thinking

1. Making decisions in uncertainty

Prob/Stat — A Way of Thinking

1. Making decisions in uncertainty
    1.1 “If I do A then X will happen but if I do B then Y will happen”

Prob/Stat — A Way of Thinking

1. Making decisions in uncertainty
   1.1 “If I do A then X will happen but if I do B then Y will happen” — Life is never works like this

Prob/Stat — A Way of Thinking

1. Making decisions in uncertainty
   - 1.1 “If I do A then X will happen but if I do B then Y will happen” — Life is never works like this
   - 1.2 “If I do A then X is more likely to happen but if I do B then X is less likely to happen”

Prob/Stat — A Way of Thinking

1. Making decisions in uncertainty
   - 1.1 “If I do A then X will happen but if I do B then Y will happen” — Life is never works like this
   - 1.2 “If I do A then X is more likely to happen but if I do B then X is less likely to happen”
2. Understanding scientific results when they are presented
   - 2.1 “Taking this test could save lives”
   - 2.2 “There is a correlation between eating X and getting disease Y”.

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.
- You test positive.

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.
- You test positive.
- What is the probability that you actually have COVID?

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.
- You test positive.
- What is the probability that you actually have COVID?
- What does this question mean?

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.
- You test positive.
- What is the probability that you actually have COVID?
- What does this question mean?
- What information do you need to know about the test to calculate this probability?

Example: COVID Testing

- You are asked to take a COVID test before a visit to a facility.
- You test positive.
- What is the probability that you actually have COVID?
- What does this question mean?
- What information do you need to know about the test to calculate this probability?
- What information do you need to know about COVID to calculate this probability?

About This Class — Learning Objectives

- Sample spaces and events, compute probabilities

About This Class — Learning Objectives

- Sample spaces and events, compute probabilities
- Random variables, reason about random processes

About This Class — Learning Objectives

- Sample spaces and events, compute probabilities
- Random variables, reason about random processes
- Aggregate properties of random events from large sample sizes — sampling and estimation

About This Class — Learning Objectives

- Sample spaces and events, compute probabilities
- Random variables, reason about random processes
- Aggregate properties of random events from large sample sizes — sampling and estimation
- Construct and analyze estimators of distributions

About This Class — Learning Objectives

- Sample spaces and events, compute probabilities
- Random variables, reason about random processes
- Aggregate properties of random events from large sample sizes — sampling and estimation
- Construct and analyze estimators of distributions
- Reason about hypothesis testing in context of experiments.

About This Class — Assessment

- Weekly quizzes 20%

About This Class — Assessment

- Weekly quizzes 20%
- Midterm 15%

About This Class — Assessment

- Weekly quizzes 20%
- Midterm 15%
- Final exam 25

About This Class — Assessment

- Weekly quizzes 20%
- Midterm 15%
- Final exam 25
- Homeworks 40%

About This Class — Assessment

- Quizzes last 15 minutes of each Thursday class.
  No exceptions. Lowest 3 grades will be dropped.
  You will need to have a computer/device in class.

About This Class — Assessment

- Quizzes last 15 minutes of each Thursday class. No exceptions. Lowest 3 grades will be dropped. You will need to have a computer/device in class.
- Midterm — in class on Feb 29. You must attend and take in person.

About This Class — Assessment

- Quizzes last 15 minutes of each Thursday class. No exceptions. Lowest 3 grades will be dropped. You will need to have a computer/device in class.
- Midterm — in class on Feb 29. You must attend and take in person.
- Final exam — Time/location in Unversity calendar. April 30, 3:30-5:30

About This Class — Assessment

- Quizzes last 15 minutes of each Thursday class. No exceptions. Lowest 3 grades will be dropped. You will need to have a computer/device in class.
- Midterm — in class on Feb 29. You must attend and take in person.
- Final exam — Time/location in Unversity calendar. April 30, 3:30-5:30
- Homeworks — due 10min before midnight, lose 10pts every 24 hours up to 20 points.

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks
  - May discuss the questions and approaches with other students or look at examples from interenet.

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks
  - May discuss the questions and approaches with other students or look at examples from interenet.
  - May not show or copy answers from other students or anywhere else (we will check).

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks
  - May discuss the questions and approaches with other students or look at examples from interenet.
  - May not show or copy answers from other students or anywhere else (we will check).
  - Turning in assignments is confirming that all code/answers were written by you.

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks
  - May discuss the questions and approaches with other students or look at examples from interenet.
  - May not show or copy answers from other students or anywhere else (we will check).
  - Turning in assignments is confirming that all code/answers were written by you.
  - Any online sources that you refer to or use should be explicitly cited in the HW submission.

About This Class — Collaboration and Cheating

- Tests and quizzes - on your own with only materials indicated.
- Homeworks
  - May discuss the questions and approaches with other students or look at examples from interenet.
  - May not show or copy answers from other students or anywhere else (we will check).
  - Turning in assignments is confirming that all code/answers were written by you.
  - Any online sources that you refer to or use should be explicitly cited in the HW submission.
- Violations of these rules will be considered an *honor violation*, resulting in a failing grade for the class.

About This Class — Programming in R

- You will self study the R language

About This Class — Programming in R

- You will self study the R language
- Use the R pages linked from course page

About This Class — Programming in R

- You will self study the R language
- Use the R pages linked from course page
- Find tutorials that suit your needs (e.g. Youtube or self paced).

About This Class — Programming in R

- You will self study the R language
- Use the R pages linked from course page
- Find tutorials that suit your needs (e.g. Youtube or self paced).
- Set aside several hours in the next week or so.

About This Class — Getting Help

- TA office hours (posted on web page) – Please take advantage of these!

About This Class — Getting Help

- TA office hours (posted on web page) – Please take advantage of these!
- Discussion board in Canvas. — Use these, but do not post answers to homework.

About This Class — Getting Help

- TA office hours (posted on web page) – Please take advantage of these!
- Discussion board in Canvas. — Use these, but do not post answers to homework.
- Send messages to course staff via Canvas.

What You Should Do Now

1. Check out the class web page:
   `www.cs.utah.edu/~whitaker/cs3130`

What You Should Do Now

1. Check out the class web page:
   `www.cs.utah.edu/~whitaker/cs3130`
2. Read the syllabus

What You Should Do Now

1. Check out the class web page:
   `www.cs.utah.edu/~whitaker/cs3130`
2. Read the syllabus
3. Download the book
   (start reading Ch 1 & 2)

What You Should Do Now

1. Check out the class web page:
   `www.cs.utah.edu/~whitaker/cs3130`
2. Read the syllabus
3. Download the book
   (start reading Ch 1 & 2)
4. Download and install R on your machine
   (take a look at R tutorial)