What is Probability and Statistics and Why Should You Care?

CS 3130/ECE 3530: Probability and Statistics for Engineers

January 7, 2025

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.

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:

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

Electrical Engineering:

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

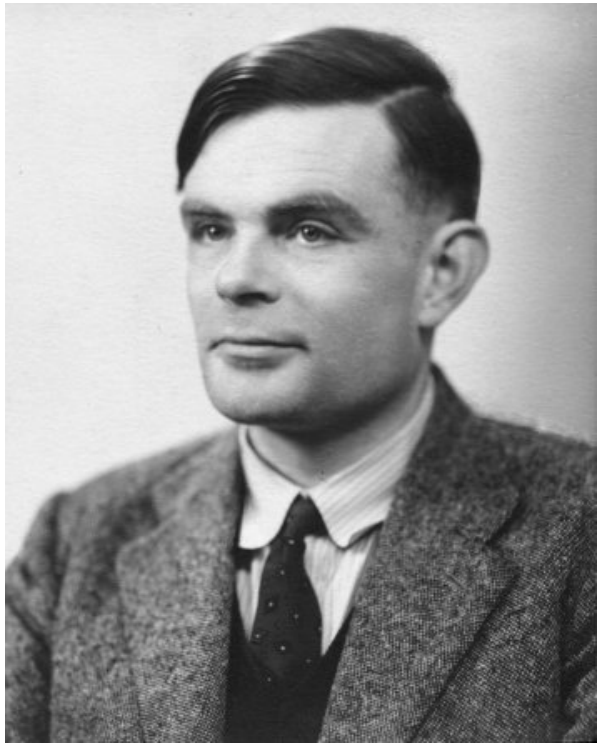
Applications of Probability and Statistics

General:

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

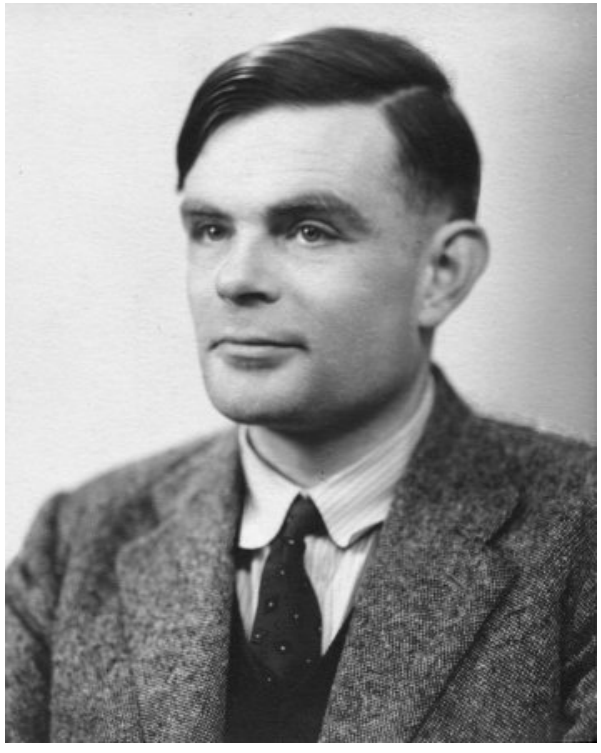
Alan Turing: Connecting CS and Probability

- “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

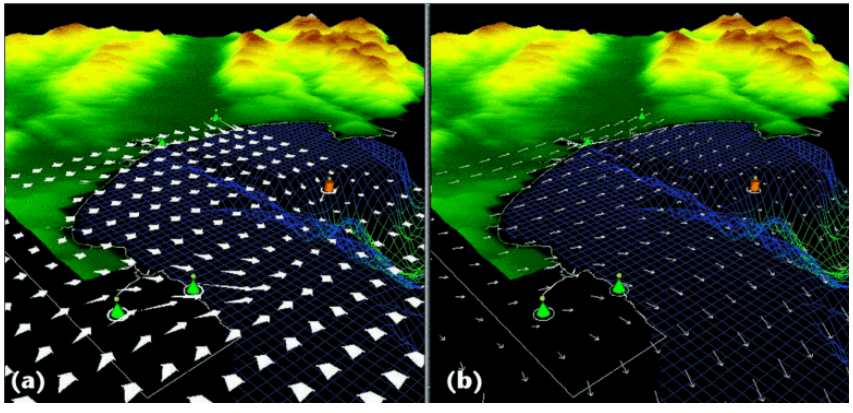
- Some algorithms benefit from using random steps rather than deterministic ones
- Example: QuickSort
  - One of the simplest & fastest sorting algorithms
  - Divide and Conquer: splits data based on **random** pivot
  - Takes $O(n \log n)$ time *in expectation*.

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: Visualization

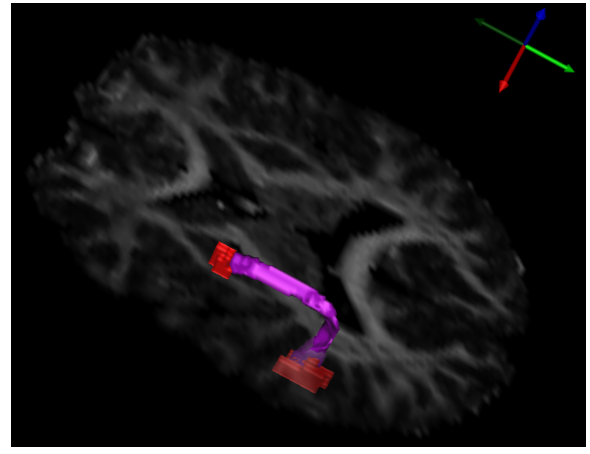
- Scientific data contains uncertainty
- Visualizations can be misleading as to “truth”
- Current research focuses on how to visualize uncertainty

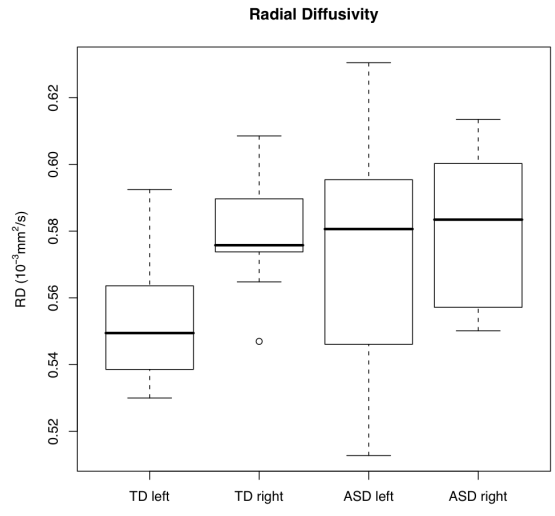


Johnson and Sanderson, IEEE Comp. Graph. and App., 2003

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

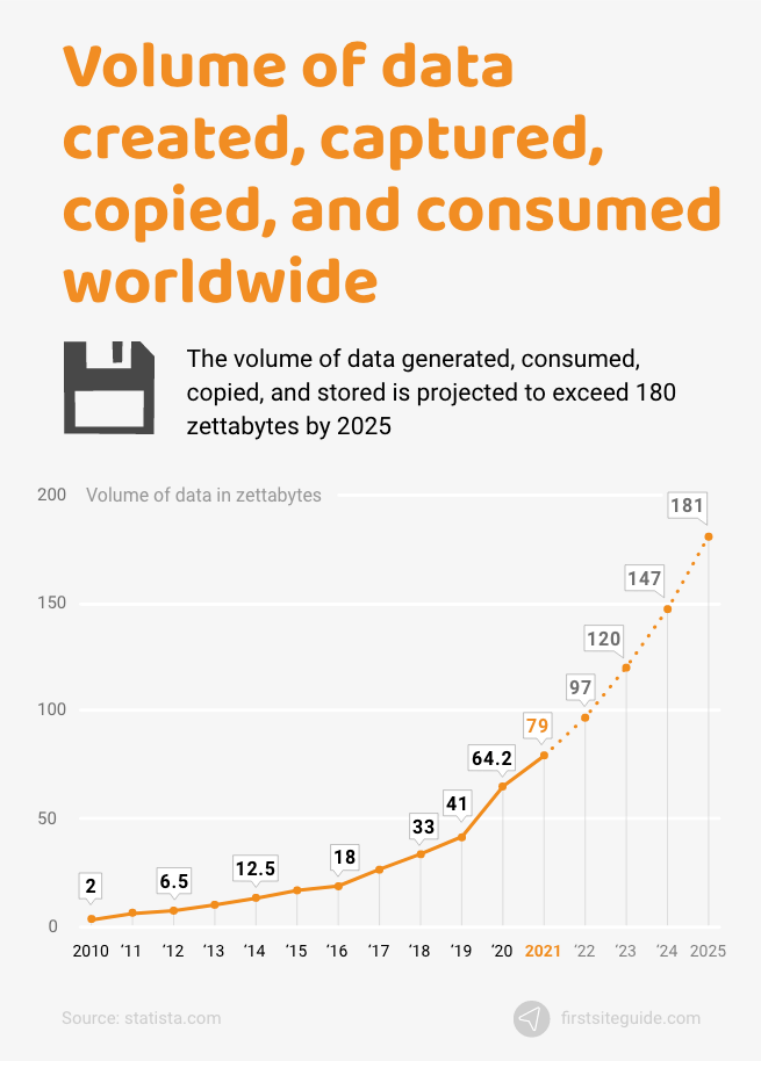




Fletcher et al, NeuroImage, 2010

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

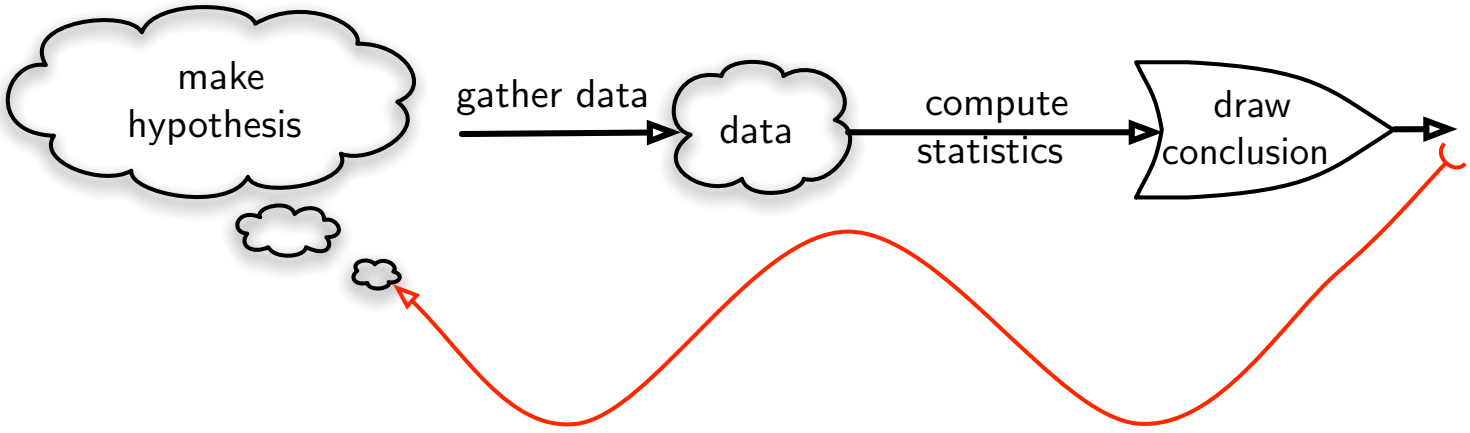
In 2005, there were **400,246,300,201** trees on Earth

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

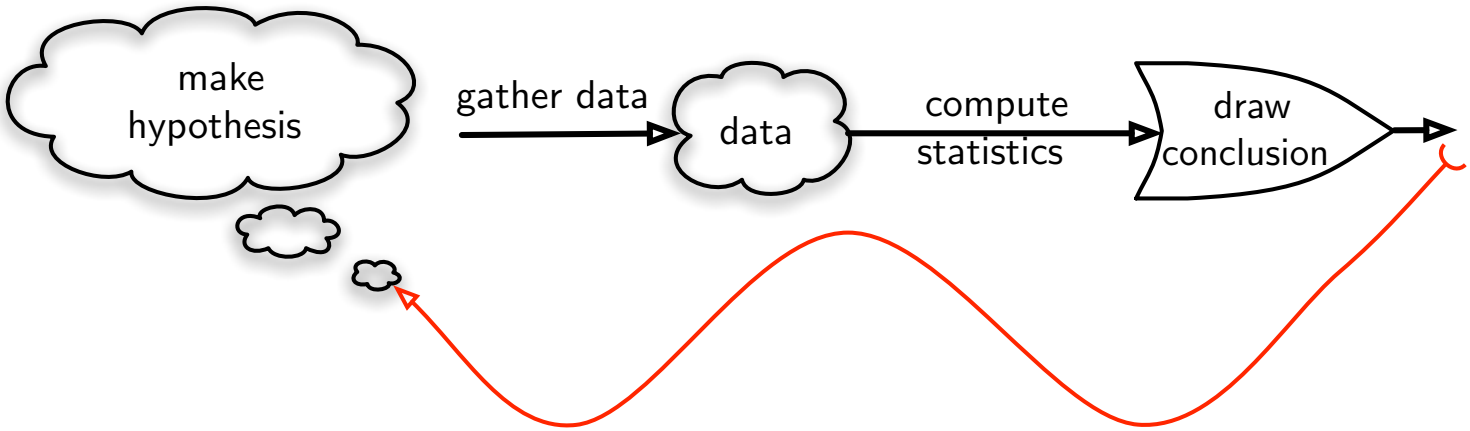
The Scientific Method



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

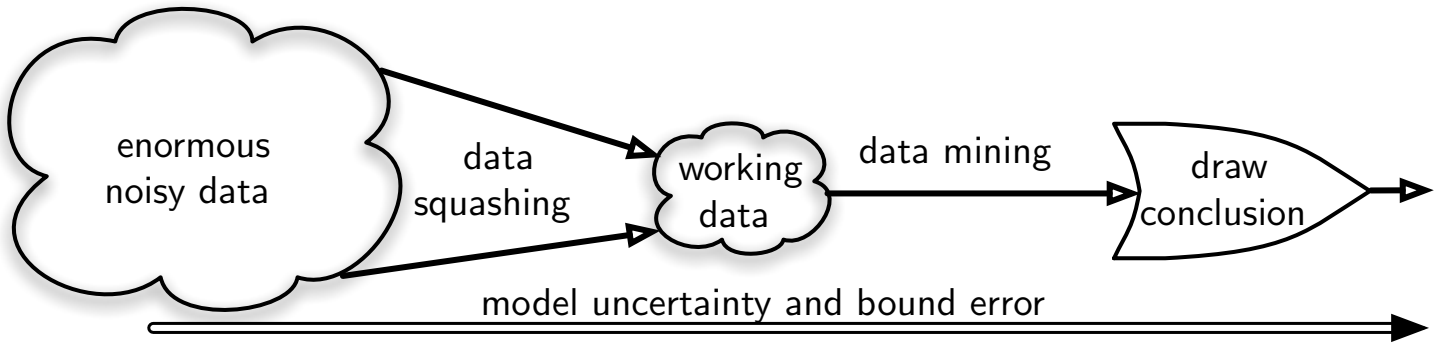


1. Define the question
2. Background research, observation
3. Formulate a hypothesis
4. **Design and run an experiment**
5. **Analyze the results**

Experimental measurements are noisy (randomness).

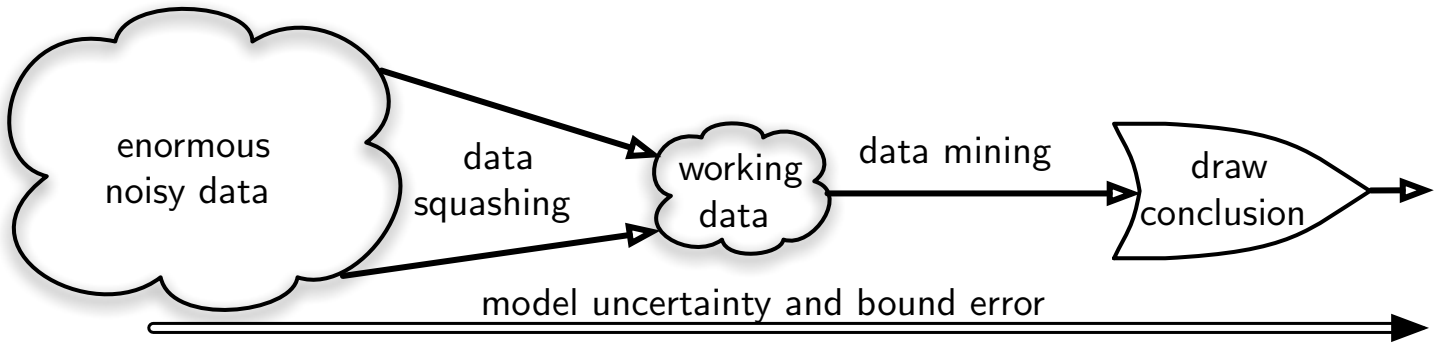
Statistics is critical in the last *two* steps!

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.

What You Should Do Now

1. Check out the class web page:
   `https://users.cs.utah.edu/~zhe/teach/cs3130.html`
2. Download the book
   (start reading Ch 1 & 2)