What is Probability and Statistics and Why Should You Care?

CS 3130: Probability and Statistics for Engineers

August 22, 2017

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?

Definition

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

Statistics is used to:

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

Applications of Probability and Statistics

Computer Science:

- Machine Learning
- Data Mining
- Simulation
- Image Processing
- Computer Vision
- Computer Graphics
- Visualization
- Software Testing
- Algorithms

Electrical Engineering:

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

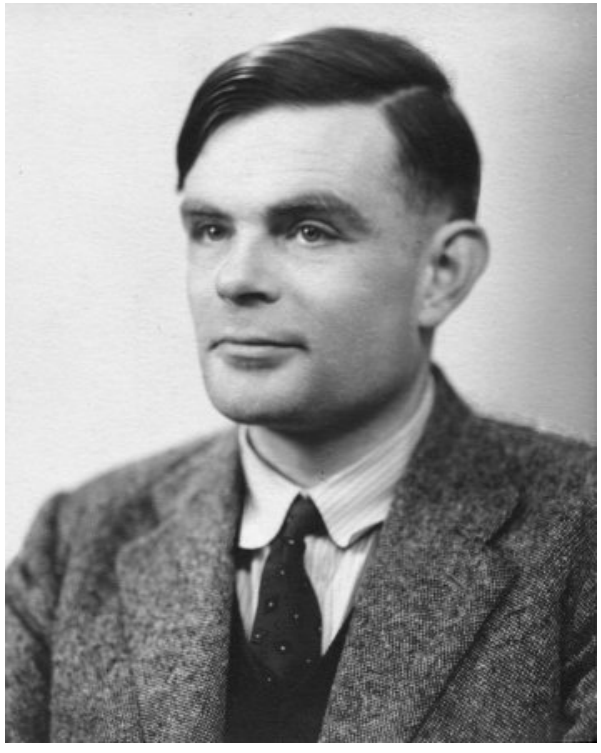
Applications of Probability and Statistics

General:

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

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.

Examples:

- Classification (recognition of faces or handwriting)
- Prediction (stock market, elections)
- Data mining

Application: Randomized Algorithms

- Some algorithms benefit from using random steps rather than deterministic ones
- Example: primality testing
  - Testing for all possible divisors is slow for large numbers
  - Instead test a random selection of divisors
  - Can be confident of primality up to a certain degree
- Example: stochastic optimization methods
  - Optimizations can get “stuck” in the wrong answer, depending on how they are initialized
  - Re-run the algorithm with several random initializations

Application: Computer Graphics

- Ray tracing models light photons bouncing around a scene
- Impossible to model *every* photon
- Monte Carlo ray tracing simulates a random selection of photons

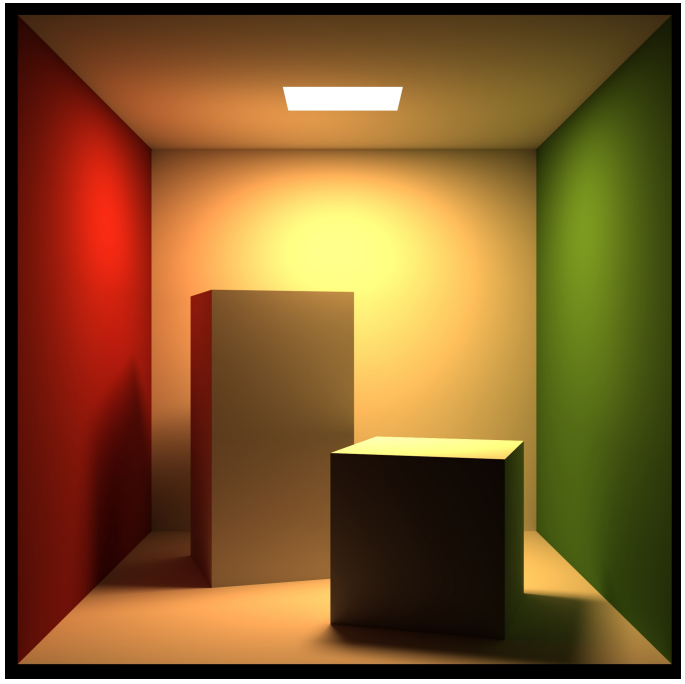
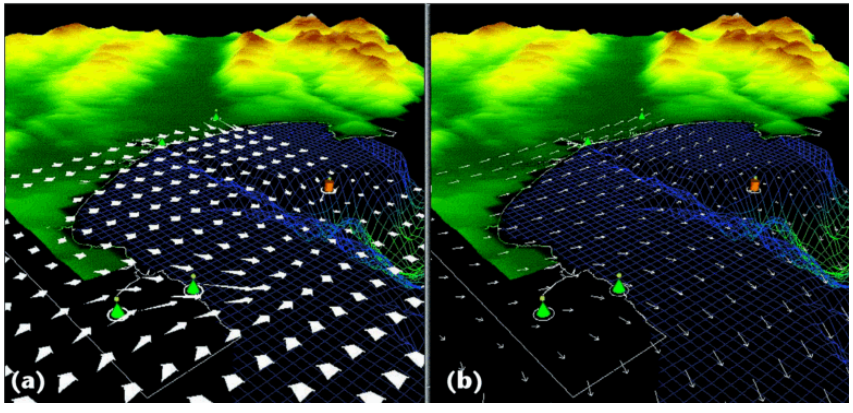


Image by Steve Parker (U of U)

Application: Visualization

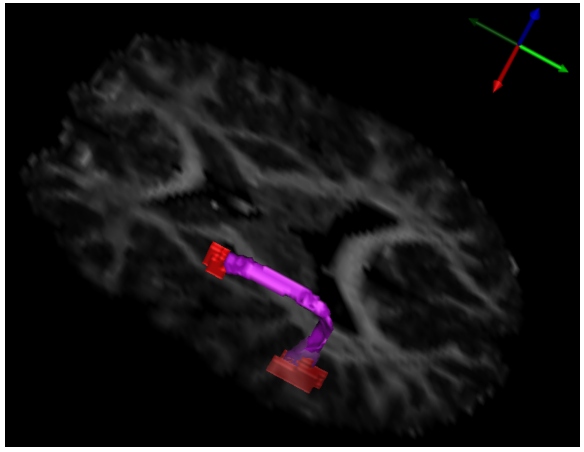
- Scientific data contains uncertainty
- Visualizations can be misleading as to “truth”
- Current research focuses on how to visualize uncertainty

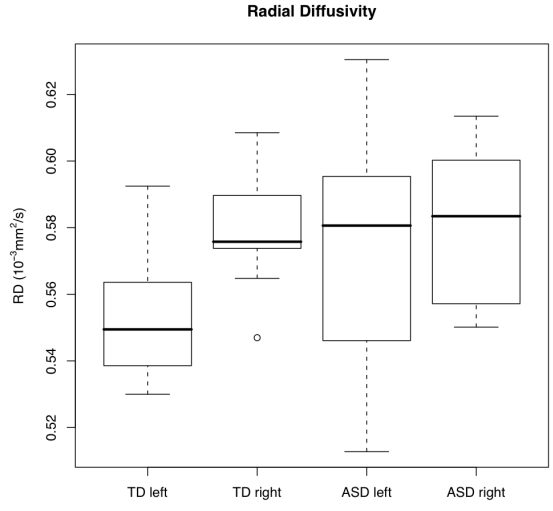


Johnson and Sanderson, IEEE Comp. Graph. and App., 2003

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

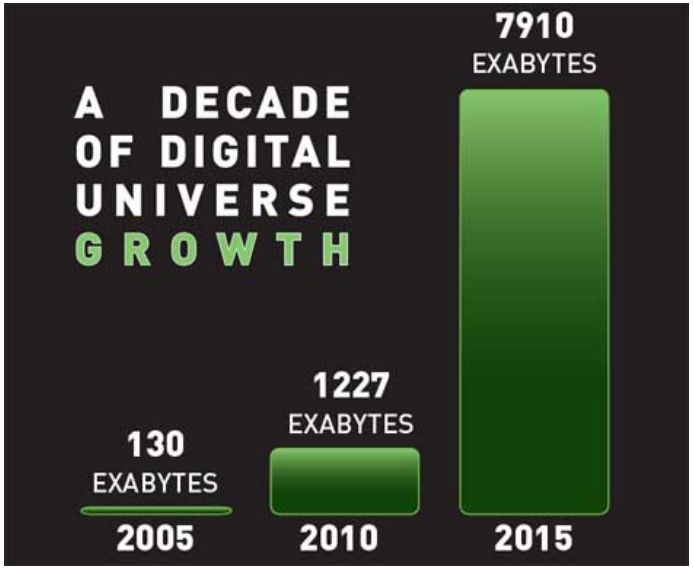




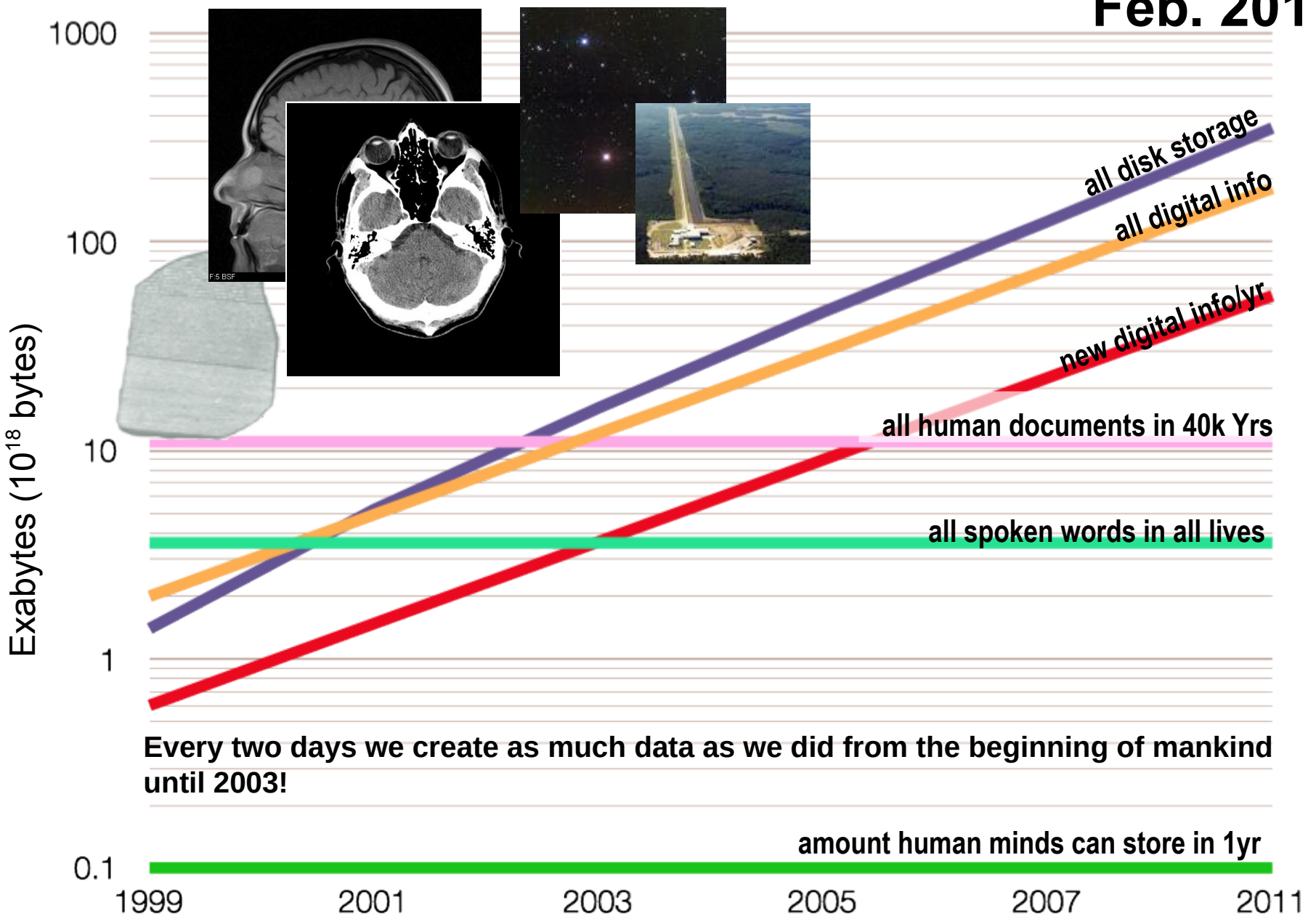
Fletcher et al, NeuroImage, 2010

“Big Data” and “Analytics”

- The amount of digital data is exploding!
- Big data analysis is statistics on steroids.
- Examples: social media, internet purchases, news articles, scientific data, medical data



Source: IDC/EMC Digital Universe Study



Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet
(slide by Chris Johnson)

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)

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!

What You Should Do Now

1. Check out the class web page
2. Sign up for the mailing list
3. Download the book
   (start reading Ch 1 & 2)
4. Download and install R and RStudio on your machine (take a look at R tutorial)