Welcome to CS100

Professor Amy Greenwald
Before we begin...

This course is NOT for CS concentrators. It does not count towards CS concentration credit. If you are a CS concentrator, please give up your seat.

For the remaining slots, we will favor (not necessarily in this order):

1. Seniors
2. First-year students, who are exploring Brown, and could potentially become CS concentrators, but have no real intention of doing so at this point
3. Masters students in Economics pursuing the “data-driven public policy” track
Let’s Get Started
How to win an Oscar

We analyse the characters played by every winner of an "actor/actress in a leading role" Oscar to see which parts will most reliably lead to glory.

Facial hair?
17%
18%

Vicars v tarts?
3
4

Fighting chance?
3

Who is North American?
British
Irish
European

Who works as a soldier?
Lawman
Monarch
Politician

Who participates in no sexual scenes?

And who, in the end, doesn't die on screen?
A TIMELINE OF EARTH'S AVERAGE TEMPERATURE
SINCE THE LAST ICE AGE GLACIATION

WHEN PEOPLE SAY “THE CLIMATE HAS CHANGED BEFORE,”
THESE ARE THE KINDS OF CHANGES THEY'RE TALKING ABOUT.

START
-4°C  -3°C  -2°C  -1°C  0°C  +1°C  +2°C  +3°C  +4°C

TEMPERATURE

COMPARSED TO THE
MEDIAN PRO AVERAGE

AT THE START OF OUR TIMELINE, 22,000 YEARS AGO,
EARTH IS 4°C COLDER THAN DURING THE LATE 20TH CENTURY.

20000
BC

1900

PRESENT DAY

Fossil fuel CO₂ emissions
start rapidly increasing

Northwest Passage opens

2000

2016

2100

BEST-CASE SCENARIO
ASSUMING IMMEDIATE,
MAMAVE ACTION TO
LIMIT EMISSIONS

OPTIMISTIC
SCENARIO

CURRENT PATH

AIRPLANES
WORLD WARS
NUCLEAR WEAPONS
INTERNET
Do you pronounce “cot” and “caught” the same?

Joshua Katz, Dept of Statistics, NC State University

Source
Nina

Tree of Life: Visualize Evolutionary Data

Interactive Web Site:
https://www.evogeneao.com/explore/tree-of-life-explorer
Amy

Visualization of the displacement of 10 million enslaved Africans over the course (3+ centuries) of the Atlantic slave trade.

slavevoyages.org
Data Mining Reveals the Six Basic Emotional Arcs of Storytelling
Luke

World Flight Routes

- Visualizing airplane routes is an example of creating art from data
- Airline routes reflect modern borders between countries and regions
How people in the US spend their days, based on a 2008 survey
Quantifying the effect of gun violence, in stolen years

It is hard to comprehend what 10,000 lost lives means, but this visualization enables the observer to better understand the scope of what is being lost as a result of gun violence in the US.
This visualization is a Twitter map during the 2013 NBA Playoffs. Red dots represent mentions of the Miami Heat, and white dots, the San Antonio Spurs.
Monica

Doctor and statistician Hans Rosling created a dynamic visualization plotting income vs. life expectancy for 200 countries over 200 years. Each country is represented by a dot whose size is proportional to some other metric (in this case, population).

Youtube video

Interactive website
This video depicts the number of military and civilian deaths during WWII, and compares these numbers to deaths associated with other events, past and present.
Alex

How does Hamilton, the non stop, hip-hop Broadway sensation tap rap's master rhymes to blur musical lines?

WSJ used a custom algorithm to explore and visualize rhymes in the musical Hamilton and its influences.
Data are Everywhere
Data are everywhere

- The complete works of William Shakespeare
- Social sciences: sociology, political science, economics, etc.
- Natural sciences: physics, astronomy, oceanography, biology, neuroscience, etc.
- Sports
Data Deluge

What happens in an internet minute in 2016?

- Facebook: 701,389 Facebook logins
- Netflix: 69,444 hours watched
- WhatsApp: 20.8 million messages
- YouTube: 2.78 million video views
- Uber: 1,389 Uber rides
- Tinder: 972,222 signups
- Google: 2.4 million search queries
- Spotify: 38,052 hours of music
- Vine: 1.04 million Vine loops
- LinkedIn: 38,194 posts to Instagram
- Twitter: 387,222 new Twitter accounts
- App Store: 51,000 app downloads
- Snapchat: 527,760 photos shared
- Amazon: $203,596 in sales

Total: 60 seconds
Data Deluge

2016: What happens in an Internet Minute?
- Facebook: 701,389 Facebook logins, 20.8 million messages
- Netflix: 69,444 hours watched
- Uber: 1,389 Uber rides
- YouTube: 2.78 million video views
- Tinder: 972,222 swipes
- Google: 38,052 hours of music
- Spotify: 1.04 million new listeners
- Amazon: 303,596 sales
- Vine: 38,184 new loops
- Instagram: 347,222 posts
- LinkedIn: 120 new LinkedIn accounts

2017: What happens in an Internet Minute?
- Facebook: 854,166 Facebook logins, 29.2 million messages
- Netflix: 80,880 hours watched
- Uber: 3,833 Uber rides
- YouTube: 4.1 million video views
- Tinder: 152,777 swipes
- Google: 3.5 million search queries
- Spotify: 46,875 hours of music
- Amazon: 11,788 hours streamed
- Instagram: 41,666 posts
- Vine: 350,000 new tweets
Information Explosion

- Volume
- Variety
- Velocity

Analysis Gap

Ability to Analyze
Observation by Michael Franklin
(Berkeley Computer Science Professor)

- 1970’s: the confluence of electrical engineering and maths led to the birth of the field of Computer Science
- 2010’s: the confluence of computer science and statistics, together with relevant domain knowledge, is prompting the growth of a new field called Data Science
The fourth scientific paradigm

1. Theoretical
2. Experimental
3. Computational
4. Data-driven
Proceed with caution

- Algorithmic Bias Q&A with Cynthia Dwork
- When Discrimination is Baked into Algorithms
Goals of Data Science
What are the goals of data science?

● **Exploration**: finding patterns in the data
  ○ Descriptive (i.e., summary) statistics
  ○ Visualization

● **Explanation**: statistical inference
  ○ Draw conclusions about *causality* for a population by observing the population in its entirety, or small samples of the population

● **Prediction**: machine learning + optimization
  ○ Learn a function that describes (potentially complex) relationships in data, and use it to make predictions about new data
Herb Simon: “Basic” vs. “Applied” Science

- Basic science = Explanatory goals
  - To know: i.e., “to describe the world”
  - To understand: i.e., “to [explain] phenomena”

- Applied science = Predictive goals
  - “Laws connecting sets of variables allow inferences or predictions to be made from known values of some of the variables to unknown values of other variables.”
Data Science Models

- Data science/statistical methods enable us to build models of populations, having observed (perhaps only small) samples
- Example application: We might have data about middle-age, middle-class women like myself living in Providence, RI, some of whom die of cancer, and others who don’t
Explanatory Goal of Data Science

- **Explanatory**: we build a model that attempts to explain why some of these women die of cancer (e.g., too much sun, not enough exercise, etc.), and why others don’t

- **Basic tools** are statistical in nature: assume an underlying model (e.g., a “true” distribution), and where possible use data to infer a causal model (e.g., parameters of that distribution)
Predictive Goal of Data Science

- **Predictive**: we build a model (i.e., a function) that predicts whether a new individual (like me!) will die of cancer
- **Basic tools** come from machine learning and optimization: E.g., learn a function that minimizes error in predictions
Methods of Data Science
How do you do Data Science? (Peter Huber)

1. Inspection
2. Error Checking
3. Modification
4. Comparison
5. Modeling and model fitting
6. Simulation
7. What-if analyses
8. Interpretation
9. Presentation of conclusions
How do you do Data Science? (Colin Mallows)

1. Identify data to collect and its relevance to your problem
2. Statistical specification of the problem
3. Method selection
4. Analysis of method
5. Interpret result for non-statisticians
How do you do Data Science? (Ben Fry)

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact
How do you do Data Science? (Galit Shmueli)

1. Define goal
2. Design study and collect data
3. Prepare data
4. Exploratory data analysis
5. Choose variables
6. Choose methods
7. Evaluate, validate, and model selection
8. Use model and report

Source: https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf
How do you do Data Science? (CS 100)

1. Prepare data
2. Exploratory data analysis
3. Choose variables and methods (i.e., build models)
4. Evaluate, validate, and model selection
5. Report (explanations or predictions)
Course Structure
Course Structure

1. **Descriptive Statistics**: Summarizing Data
   - No underlying statistical model
   - No learning/estimation; no inference
   - Just *Exploratory Data Analysis*

Examples

- Histograms, conditional histograms
- Measures of central tendency
- Measures of dispersion
- Curve fitting
Course Structure (cont’d)

2. **Classic Machine Learning**: Summarizing Data
   - No underlying statistical model
   - Arguably learning, but no estimation, and hence no quantification of uncertainty
   - **Prediction**: Inductive, out-of-sample inference

**Example Methods**
- Decision and regression trees
- $k$-nearest neighbors and LOESS
3. **Statistics** (central limit theorem, law of large numbers, etc.)

4. **Statistical Machine Learning** (i.e., **Estimation & Inference**)
   - Assume an underlying statistical model of a population
     - Selects a few key variables of interest
     - Might make assumptions about how they are distributed
     - Might describe how they relate to one another
   - Learning: Estimate true parameters of the model, using in-sample data
     - Example estimators: sample mean, sample variance, etc.
     - Example techniques: maximum likelihood, maximum a posteriori
   - Inference: Generalize using those parameters to the entire population
     (including any out-of-sample data)
Course Structure (cont’d)

Model desiderata
- Plausible
- Interpretable
- Simple (“the simplest explanation is best”)
- Generalizable (i.e., applicable well beyond any sample)

Model checking is key!
“All models are wrong, but some are useful.” -- George Box
Course Structure (cont’d)

- Data cleaning (yuk!)
- Data visualization (fun!)

- Structured, as well as unstructured, data
  - Text, maps, social networks, etc.

- Algorithm bias, data ethics and privacy, etc.
Course Administration
Goal of CS 100

To endow students with a basic set of computational skills that will enable them to process data, and ultimately glean meaningful information from them.
What will students learn in this course?

- **Probability and Statistics**
  - Probability Distributions, Descriptive Statistics, etc.
  - Law of Large Numbers, Central Limit Theorem, etc.
  - Conditional Probability, Bayes’ Theorem, etc.

- **Machine Learning**
  - Clustering
  - Regression
  - Classification

- **Tools**
  - Spreadsheets, R, and Markdown
Who does Data Science?

- Statisticians
- Computer Scientists
- Domain Experts (e.g., Economists, Biologists, etc.)
- Really...everyone!
Who is this course for?

Really...everyone!

Everyone who wants to learn to process any part of the myriad of data that are currently being collected by both the private and public sector about our daily lives.

Caveat: if you are a CS concentrator, other Brown courses are better suited to your level/needs, like CSCI 1951A.
What do students need to know in advance?

NOTHING!

This course has no prerequisites.

Course Structure

● Meetings
  ○ Lectures on Mondays and Wednesdays and Fridays
  ○ TA-led discussion sections from time-to-time
  ○ Studios: collaborative hands-on activities

● Weekly readings
  ○ Many online references
    ■ Seeing Theory, A Visual Introduction to Stats
  ○ Optional Textbooks
    ■ The Cartoon Guide to Statistics
    ■ Naked Statistics, by Charles Wheelan
Grading

- Take-home assignments
  - Homework assignments, due every other week through Thanksgiving
  - One final project (the bulk of which you will do after Thanksgiving), in lieu of a final exam

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td>5%</td>
</tr>
<tr>
<td>Studios</td>
<td>25%</td>
</tr>
<tr>
<td>Homeworks</td>
<td>40%</td>
</tr>
<tr>
<td>Final Project</td>
<td>30%</td>
</tr>
</tbody>
</table>
Office hours

Amy’s office hours are Thursdays 12-1, or by appointment. Her office number is CIT 383.

The TA’s office hours are:

- 4-6 pm Saturday
- 7-9 pm Monday
- 4-6 pm Thursday

TA hours are held in CIT 203.
Collaboration Policy

Students are encouraged to collaborate with their peers in CSCI 0100. Studios are pair-programmed, with each student finding a different partner each week. When working on homework assignments, students may consult one another; but students are required to list the names of all students with whom they discussed an assignment on their submitted work.

Unnatural similarities among students’ submissions with other students whose names are not listed will be forwarded to the Dean of the College’s office for review.

If you have any questions about this policy, please ask the course staff for clarification. Not understanding our policy is not grounds for not abiding by it.
Diversity and Inclusion

The computer science department is committed to diversity and inclusion, and strives to create a climate conducive to the success of women, students of color, students of all (or no) sexual orientations, and any other students who feel marginalized for any reason.

If you feel you have been mistreated by another student, or by any of the course staff, please feel free to reach out to one of the CS department’s Diversity and Inclusion Student Advocates, or to Professor Greenwald, Professor Doeppner (DUS), or Professor Cetintemel (the CS department chair).

We, the CS department, take all such complaints seriously.
Accommodations

If you feel you have any disabilities that could affect your performance in the course, please contact SEAS, and ask them to contact the course staff.

We will support accommodations recommended by SEAS.
Harassment

Please review Brown’s Title IX and Gender Equity Policy.

If you feel you might be the victim of harassment (in this course or any other), you may seek help from any of the resources listed here.
Jargon
Jargon

Perhaps for practical reasons, all fields are full of jargon.

Never in this classroom or in studio should you hesitate to ask for clarification if you do not understand some bit of jargon used by the professor, a TA, or any of your fellow students.

No one understands all jargon. Please do not be embarrassed to ask questions when you are confused by terminology.
Big Data

“Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations.”

Oxford Dictionary

N.B. This course is concerned primarily with small data. Additional tools, beyond those taught in this course, are necessary to manipulate big data.
Data Mining
Extracting comprehensible information from data

Data Munging/Wrangling/Jujitsu
Converting data from one "raw" form into another form, which is often cleaner and more structured
Predictive Modeling

Building a statistical model of unknown behavior

Predictive Analytics

Making predictions about unknown future events
Final bit of logistics
Survey

If you plan to take this class, even if you are already registered, you must complete this survey, which asks some demographic questions, and other questions about your interests in data and otherwise, by 9 pm FRIDAY, September 8, 2017:

http://tinyurl.com/cs100form
If you end up taking this class, be sure to:

1. Visit the course website
   
   http://www.cs.brown.edu/courses/cs100

2. Register for the course so you can login to the Brown CS server

3. Sign up for Piazza (instructions are on the course website)

4. Go to the CIT first floor to get an iClicker (ask the at the desk), and then register your iClicker on Canvas
Politics

- Predict elections
- Study demographics
- Campaign managers study voters and target their messages accordingly

Source: www.fivethirtyeight.com
The role of big data in medicine

Technology is revolutionizing our understanding and treatment of disease, says the founding director of the Icahn Institute for Genomics and Multiscale Biology at...
Industry

Airlines
● Price setting
● Route planning
● Supply chain optimization
● Frequent flyer program design

Delta Airlines introduces chips for smart luggage

This story was delivered to BI Intelligence IoT Briefing subscribers. To learn more and subscribe, please click here.

Delta Airlines announced it will be releasing a new system that uses RFID chips placed on passengers’ bags to track their location, NBC News is reporting.

The airline hopes that this will help solve the problem of lost baggage, which costs airlines thousands of dollars per year across the globe.

Southwest Airlines Uses Big Data To Deliver Excellent Customer Service