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

slavevoyages.org
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).
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.

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

AIRPLANS WORLD WARS NUCLEAR WEAPONS INTERNET

FOSSIL FUEL CO₂ EMISSIONS START RAPIDLY INCREASING
NORTHWEST PASSAGE OPENS

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

OPTIMISTIC SCENARIO CURRENT PATH

Image Source
Alex

Do you pronounce "cot" and "caught" the same?

Image Source

Joshua Katz, Dept of Statistics, NC State University
North Carolina Votes in 2016 Presidential Election

Counties by percentage of African American population

Counties by percentage of population with a Bachelors

Abby Draper & Sean Manning
Data are Everywhere
Data are everywhere

- **The complete works of William Shakespeare**
- Social sciences: sociology, political science, public health, economics, etc.
- Natural sciences: physics, astronomy, oceanography, biology, neuroscience, etc.
- Sports
2018 This Is What Happens In An Internet Minute

- Google: 973,000 Logins
- YouTube: 4.3 Million Videos Viewed
- Netflix: 266,000 Hours Watched
- Snapchat: 2.4 Million Snaps Created
- Instagram: 174,000 Scrolling Instagram
- Facebook: 18 Million Text Messages
- Amazon: 67 Voice-First Devices Shipped
- Tinder: 18 Million Emails Sent
- Gmail: 936,073 Views
- Spotify: 0.7 Million Songs Played
- Twitter: 481,000 Tweets Sent
- Pinterest: 1 Million Pins
- Netflix: 3.7 Million Search Queries
- Google Play: 27 Million Apps Downloaded
- Instagram: 36 Million Likes
- YouTube: 5.6 Million Comments
- Amazon: 25 Million Items Shipped
- Instagram: 8 Million People Following
- Facebook: 3 Million Posts Liked
- Spotify: 5 Million Songs Played

2019 This Is What Happens In An Internet Minute

- Google: 1 Million Logging In
- YouTube: 4.5 Million Videos Viewed
- Netflix: 694,444 Hours Watched
- Snapchat: 1 Million Snaps Created
- Instagram: 41.6 Million Messages Sent
- Facebook: 416 Million Messages
- Amazon: 390,030 People Tweeting
- Instagram: 1 Million Reviews
- Twitter: 1.4 Million Tweets Sent
- Spotify: 2 Million Songs Played
- Amazon: 877,000 People Viewing
- Netflix: 3.8 Million Search Queries
- WhatsApp: 1 Million Calls Made
- Instagram: 6.2 Million Photos Taken
- Facebook: 1 Million Likes
- Amazon: 1 Million Products Sold
- Spotify: 4.8 Million Gifs Served
- Amazon: 5 Million Downloads
- Pinterest: 3 Million Pin Likes
- Twitter: 3 Million Tweets

Created By:
@LoriLewis
@OfficiallyChadd
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 [a.k.a. machine learning], 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:**
   Empirical
Proceed with caution

- Algorithmic Bias Q&A with Cynthia Dwork
- When Discrimination is Baked into Algorithms
- Fairness, Accountability, & Transparency Conference
Goals of Data Science
Herb Simon: “Basic” vs. “Applied” Science

- **Basic science = Descriptive & 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 … predictions to be made from known values of some of the variables to unknown values of other variables.”
What are the goals of data science?

- **Exploration**: finding patterns in data
  - Descriptive statistics
    - Numerical summaries: tables
    - Visualizations (i.e., visual summaries): plots

- **Explanation**: providing reasons for patterns in data
  - Tell a causal story (e.g., smoking *causes* cancer)

- **Prediction**: predicting patterns in unseen data
  - Model potentially complex relationships in observed data, and use the model to make predictions about unobserved data
Data

- We might have data about middle-age, middle-class women (like me!) living in Providence, RI
- We might have a snapshot of these data, or the data set could be longitudinal (i.e., span multiple years)
- If the data concern women from, say, the 1950’s, we might even have labels: e.g., cause of death
Exploratory Goal of Data Science

- We can summarize the data by calculating the average age of death, the most popular cause of death, etc.
- With longitudinal data, we can plot weight, height, ..., over time
- Basic **tools** are descriptive statistics
  - Numerical summaries: tables
  - Visualizations (i.e., visual summaries): plots
Explanatory Goal of Data Science

- We learn features of the women who die of cancer (e.g., too much sun, not enough exercise, etc.), and who don’t

- Basic tools are statistical in nature:
  - Assume a statistical model: a “true” functional or distributional form
  - Use data/observations to estimate the parameters of the model
  - Where possible, use the model to draw causal conclusions
Predictive Goal of Data Science

- We learn a model (e.g., a function) that predicts (i.e., forecasts) whether a new individual (like me!) will die of cancer
- Some tools come from machine learning and optimization: E.g., learn a function that minimizes error in predictions
- Other tools are statistical in nature:
  - Assume a statistical model: a “true” functional or distributional form
  - Use data/observations to estimate the parameters of the model
  - Where possible, use the model to make inferences about the population
Methods of Data Science
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? (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? (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. User 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 Overview
Course Overview

1. **Descriptive Statistics**: Summarizing Data
   - No underlying model
   - No learning, statistical or otherwise
   - Just *Exploratory Data Analysis*

Examples

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

2. Classic Machine Learning: Summarizing Data
   - No underlying statistical model
   - Learning, so training on in-sample data
   - Prediction: Inductive, out-of-sample forecasting

Example Methods
   - Decision and regression trees
   - $k$-nearest neighbors
Course Overview (cont’d)

3. **Classic Statistics**
   - Law of Large Numbers
   - Central Limit Theorem
   - Confidence Intervals
   - Hypothesis Testing

**Example Applications**
   - Analyzing clinical trials to predict drug efficacy
   - Analyzing polling data to predict 2020 Democratic presidential candidate
Course Overview (cont’d)

3. Statistical Machine Learning (i.e., Estimation and 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
   ○ 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: Apply the model to generalize to out-of-sample data
Course Overview (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 Overview (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
  - Descriptive Statistics (measures of central tendency and dispersion)
  - Law of Large Numbers, Central Limit Theorem, etc.
  - Conditional Probability, Bayes’ Theorem, etc.
- Machine Learning
  - Classification
  - Regression
  - Clustering
- 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
  - TA-led discussion sections on most Fridays
  - Studios: collaborative hands-on activities

- **Take-home assignments**
  - Homework assignments, due every other week through Thanksgiving
  - One week mini project due right before Indigenous People’s Day
  - One month final project (the bulk of which you will do after Thanksgiving), in lieu of a final exam
Course Structure (cont’d)

- Lectures are conceptual, and can be theoretical at times
  - They are designed to introduce you a topic, generally, and at a high level
  - They include little explicit R instruction (except during programming week)
  - They often require thinking (indeed, you’ll notice me thinking aloud often)

- Studios and homeworks are hands on, and very practical
  - They are designed to help you work out details about a topic
  - They include explicit R instruction (sometimes, just “type this”; “type that”)
  - Sometimes, they (studios, especially) don’t require thinking
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>10%</td>
</tr>
<tr>
<td>Studios</td>
<td>20%</td>
</tr>
<tr>
<td>Homeworaks</td>
<td>35%</td>
</tr>
<tr>
<td>Final Project</td>
<td>35%</td>
</tr>
</tbody>
</table>
Office hours

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

When they are finalized, the TA’s office hours and locations will be posted on the course website calendar.
Collaboration Policy

Students are encouraged to collaborate with their peers in CSCI 0100. Studios are pair-programmed. For their own benefit, students should make a concerted effort to work with multiple partners over the course of the semester.

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, to assess whether or not there has been a violation of Brown's Academic Code.
Collaboration Policy (cont’d)

Even when collaborators are appropriately named on the students' handins, each *individual* student must be able to fully explain their solutions—including all code—to the course staff. Often students search the web for help with R, which is legitimate, as long as they can fully explain their submitted code to the course staff.

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 Hughes (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. We will support accommodations referred 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.
Course Laptop Use

Owning a laptop is neither required nor necessary to succeed in CSCI 100, so not owning a laptop does not preclude you from taking this course. Nonetheless, during some classes, such as sections and programming lectures, students may benefit from the use of a personal laptop. (Note that during other classes, the professor may expressly forbid the use of any personal devices.)

If you do not own a laptop, but would like access to one this semester, please contact the HTAs for assistance, assuming you are comfortable doing so. Otherwise, please feel free to reach out to Dean Elie, the Associate Dean for Financial Advising, for help purchasing a laptop, or the IT service center, to borrow a laptop.
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, please complete this survey, by 9 pm FRIDAY, September 6, 2019:

https://forms.gle/KKm6vPwuu54GJwTT6

It asks for your studio hour preferences, and also asks various other questions about your interests in data and otherwise.
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)
2019 TAs Favorite Visualizations
Round-trip time to the nearest clinic at 12 weeks pregnant.
The Daily Routines of Famous Creative People

Turns out great minds don’t think alike. Discover how some of the world’s most original artists, writers and musicians structured their day, based on ‘Daily Rituals’ by Mason Currey. Filter the different categories by toggling on or off, and hover over the colored bars to learn more about the daily routines.
On average, the pockets in women’s jeans are 48% shorter and 6.5% narrower than men’s pockets.
CDC Cause of Death in USA
"What actually causes death?"

Code: www.github.com/aaronpenne
Twitter: @aaronpenne
Aaron Penne © 2018

Based on in-depth analysis by H. Al-Jamaly, M. Siemers, O. Shen, and N. Stone at owenshen24.github.io/charting-death
2017 TAs Favorite Visualizations
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%

Best decade for male roles: 1930s

Vicars v tarts?
- 3

Fighting chance?
- 3

No sexual scenes and who, in the end, doesn't die on screen.

Image Source
Nina: Tree of Life

Image Source
2016 TAs Favorite Visualizations
Jackie

Data Mining Reveals the Six Basic Emotional Arcs of Storytelling
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
Eital

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.
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.
Extras
Politics

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

How Big Data Is Changing Healthcare

Bernard Marr, CONTRIBUTOR

I write about big data, analytics and enterprise performance  FULL BIO  View Full Bio

Opinions expressed by Forbes Contributors are their own.

If you want to find out how Big Data is helping to make the world a better place, there’s no better example than the uses being found for it in healthcare.
Industry

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

Delta Airlines introduces chips for smart luggage

BI Intelligence
© Aug. 24, 2016, 11:45 AM  A 2,807

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