Favorite Visualization: Amy

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

slavevoyages.org
Favorite Visualization: Julia

“200 Years of Immigration to the US”
Favorite Visualization: Isha

Providence
San Francisco
5.4m² Arctic ice

https://shameplane.com/
Favorite Visualization: Rutvik

Network Graph of the relationships between the characters in HBO’s ‘The Wire’
Favorite Visualization: Justin

Benefits of Electric Cars
Favorite Visualization: Krishi

The number of moves it takes a knight to get around the chess board
Favorite Visualization: Jay

Graphing the history of philosophy

Image Source
Favorite Visualization: Serdar

**Global GDP 2021**

Gross domestic product (GDP) serves as a barometer for a country’s economic health. It measures the total market value of final goods and services produced in a country during a given year.

Together, the U.S. and China account for 42% of global GDP. Here is GDP by country according to IMF estimates.

**Global GDP 2021**
Favorite Visualization: Aditya
Student Projects
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
What Makes a Winner

Kriyana Reddy
Jason Traum
The Politics and Economics of Development

Francis Baviera Maloney
Katya Scocimara
Data are Everywhere
Data are Everywhere

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

Airlines
● Price setting
● Route planning
● Revenue management
● 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.

The system will leverage RFID tags connected to each bag that will be scanned by Delta workers, and notifications of the bag’s whereabouts will be pushed to the bag owner via a mobile application. RFID technology has been around for decades and has long been used to track parcels.

Previously, Delta used barcodes to track the
2021: This is What Happens in an Internet Minute

- 1.4 Million Facebook Scrolling
- 21.1 Million Facebook Texts Sent
- 9,132 LinkedIn Connections Made
- 28,000 Netflix Subscribers Watching
- $1.6 Million Spent Online
- 3.4 Million Snapchat Snaps Created
- 69 Million WhatsApp Messages Sent
- 695,000 Instagram Stories Shared
- 197.6 Million Emails Sent
- 2 Million Tinder Swipes
- 5,000 Google Home Downloads
- 200,000 People Tweeting
- 2 Million Amazon Alexa Devices Shipped
- Created by: LoriLewis & OfficiallyChadd
2020: This Is What Happens In An Internet Minute

- 1.3 Million Facebook Logging In
- 19 Million Texts Sent
- 4.7 Million YouTube Videos Viewed
- 694,444 Scrolling Instagram
- 194,444 People Tweeting
- 1.6 Million Swipes
- 190 Million Emails Sent
- 1.2 Million Views
- 1,400 Downloads
- 305 Smart Speakers Shipped
- 2.5 Million Images Viewed
- 2.5 Million Snaps Created
- $1.1 Million Spent Online
- 764,000 Hours Watched
- 400,000 Apps Downloaded

Created By:
@LoriLewis
@OfficiallyChadd
Information Explosion

- Volume
- Variety
- Velocity

Ability to Analyze

Analysis Gap

Image Source
Information Explosion

- Volume
- Variety
- Velocity

Analysis Gap

Ability to Analyze

Data Science

Communication

Management

Sociology

Data

Domain

Thinking

Informatics

Statistics

Computing

Image Source

Image Source
Observation by Michael Franklin
(University of Chicago 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: 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?

● **Description**: describing patterns in data
  ○ Descriptive statistics
    ■ Numerical summaries: tables
    ■ Visualizations (i.e., visual summaries): plots

● **Explanation**: explaining patterns in data
  ○ Tell a causal story (e.g., smoking *causes* cancer)
  ○ Tell an effect story (e.g., the effect of smoking on health)

● **Prediction**: predicting patterns in unseen data
  ○ Model potentially complex relationships in observed data, and use the model to make predictions about unobserved data
What are the goals of data science?

- Abductive Reasoning
- Inductive reasoning
- Deductive reasoning
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
Descriptive Goal of Data Science

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

- We learn a causal model that is intended to explain which features a woman possesses may cause her to die of cancer.

- Some **tools** come from machine learning and optimization:
  - Assume a machine learning model: e.g., a “true” functional form
  - Learn a function that minimizes error in predictions
  - Prioritize the model’s interpretability over its accuracy

- Other **tools** are statistical in nature:
  - Assume a statistical model: e.g., a “true” distributional form
  - Use data/observations to estimate the parameters of the model
  - Use the model to make causal inferences, where possible
Predictive Goal of Data Science

- We learn a model that predicts the likelihood that a woman characterized by a certain set of features will die of cancer.

- Some **tools** come from machine learning and optimization:
  - Assume a machine learning model: e.g., a “true” functional form
  - Learn a function that minimizes error in predictions
  - Prioritize accuracy. Function may be very complex.

- Other **tools** are statistical in nature:
  - Assume a statistical model: e.g., a “true” distributional form
  - Use data/observations to estimate the parameters of the model
  - Use the model to make predictions about new data/observations
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. Method implementation
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
How do you do Data Science? (CSCI 0100)

1. Define goal
2. Find and prepare data
3. Exploratory data analysis
4. Choose variables and methods (i.e., build models)
5. Evaluate, validate, and model selection
6. Report (explanations or predictions)
Course Overview
Course Overview

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

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

2. **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 election outcomes
Course Overview (cont’d)

3. Classic Machine Learning
   ○ Assume a functional form
   ○ 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)

4. Statistical Machine Learning (i.e., Estimation and Inference)

  ○ Assume an underlying statistical model of a population
    ■ Selects a few key variables of interest
    ■ Might describe how they relate to one another
    ■ Might make assumptions about how they are distributed

  ○ Estimate the parameters of the model, using in-sample data
    ■ Example estimators: sample mean, sample variance, etc.
    ■ Example techniques: maximum likelihood, maximum a posteriori, etc.

  ○ 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., still relevant, 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 privacy and provenance, etc.
Course Administration
Learning Outcomes

1. Students should become proficient in the programming basics of R and RStudio
2. Students should learn to apply data-science concepts to develop and assess data models
3. Students should learn to communicate data findings effectively, orally, visually, and in writing
Goal of CSCI 0100

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 or intend to be a CS concentrator, other Brown courses are better suited to your level/needs, like CSCI 1951A (Available Spr 2022).
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 some Fridays
  ○ Studios: collaborative hands-on activities

● Take-home assignments
  ○ Homework assignments, due every other week through Thanksgiving
  ○ One week mini project due around 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
(Tentative) Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studios</td>
<td>30%</td>
</tr>
<tr>
<td>Homeworks</td>
<td>35%</td>
</tr>
<tr>
<td>Mini Project</td>
<td>10%</td>
</tr>
<tr>
<td>Final Project</td>
<td>25%</td>
</tr>
</tbody>
</table>
Late Policy

Students are granted three free late days, which can be applied, as needed, over the course of the semester to homework assignments and the mini-project, but not to the final project.

In the unfortunate circumstance that the three free late days are all used up, late day penalties will apply: -10% within 24 hours, -25% within 48, and -50% within 72. No assignments will be accepted more than 72 hours beyond their due date.

Extensions may be granted by the professor in extreme circumstances. If you are ill, please visit health services before requesting an extension. If you are under any other sort of duress, please seek advice from a dean.
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 or gender 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.
Office hours

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

Once they are finalized, the TA’s office hours and locations will be posted on the course website calendar.
Final bit of logistics
Survey

If you plan to take this class, even if you are already registered, please complete this survey, by 12pm MONDAY, September 12: https://forms.gle/Jzw5i2XcU3DoKiP69

Just for fun, please complete this survey as well, also by 12pm MONDAY, September 12: https://forms.gle/nyvu5c1M1toNtRTf8
If you are 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 CS dept machines

   (Login with your Brown email address)

4. Sign up for Gradescope: https://www.gradescope.com/
   (Login using `School Credentials` and select Brown University)

Course code: 57RR43
Studio 0

Studio 0 is a take-home assignment.

It involves reading our course policies, signing the course collaboration policy, installing the requisite software, etc.

It is due on Wednesday, September 14 at 10:59 a.m..
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