Questions for Today

- Why am I excited about Data Science?
- What is Data Science?
- Who are we?
- What does it mean to be a data scientist today?
- Break
- What will I learn and how?
- Demo (who are you?)!

Slides from lecture available online at [http://ds100.org/sp18](http://ds100.org/sp18)

---

**Why am I excited about Data Science?**

Data is Changing the World

---

**Where should I eat?**

Where can I get the best burrito in SF?

Each ratings star added on a Yelp restaurant review translated to anywhere from a 5 percent to 9 percent effect on revenues.  
— Harvard Business School

Learn about eating the dangers of eating In SF in 2nd homework …

[http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business](http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business)

---

**Data Science is transforming Science**

Jim Gray  
Turing Award Winning Computer Scientist & Cal Alum.

---

**Data can help address climate change ...**

By tracking sales data on energy efficient appliances, data for climate action is helping guide urban campaigns to educate the general public and measure changes in purchasing behavior.

[http://www.dataforclimateaction.org](http://www.dataforclimateaction.org)
Jim Gray introduced the idea of the Fourth Paradigm of Science:
- Experimental
- Theoretical
- Simulation
- Data Intensive

Astronomy in the 4th Paradigm:
- Sloan Digital Sky Survey (SDSS)
- Database Systems
- Sky Server

Technology Trends:
- 2020s: Data Industry
  - Collect and sell information
- 2010s: Internet Industry
  - Online retailers and services
- 2000s: Software Industry
  - Sold computer software
- 1990s: Hardware Industry
  - Sold computers
- 1980s: Hardware Industry
  - Sold computers

The Darker Side of Data Science:
- Obscuring complex decisions
- Mortgage backed securities → market crash
- Teaching scores & job advancement
- Reinforcing historical trends and biases
- Hiring based on previous hiring data
- Recidivism and racially biased sentencing
- Social media, news, and politics
- We will touch on the ethics of data science throughout the class

But … I am optimistic:
- Knowledge is empowering
- Data science offers immense potential to address challenging problems facing society
- The future is in your hands and I believe
  You will use your knowledge for good.

… I am thrilled to teach Data 100!
What is Data Science?
The recurring question across industry and academia.

My Definition for Data Science
The application of data centric, computational, and inferential thinking to understand the world & solve problems in Science and Engineering.

Data science is fundamentally interdisciplinary.

Skills of Data Science
- Computer Science
- Statistics

Danger Zone!
Research Analyst
Data Scientist
Machine Learning
Statistics

Drew Conway’s Venn Diagram of Data Science

Who are we?
What does it mean to be a data scientist today?  
How can we answer this question?

There is a lot of excitement around Big Data  
... how big is the data?
What do they do?

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

- Developing Models
- Implementing ML Algorithms
- Visualization
- Exploratory Data Analysis (EDA)
- Researching Questions
- Writing Reports

Are the top items surprising?

Data Cleaning

What tools do they use?

- Programming Languages
- Machine Learning

Visualization

Implementing ML Algorithms

Developing Models

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

What do they do?

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

- Writing Reports
- Researching Questions
- Exploratory Data Analysis (EDA)

DATA CLEANING

53%

VISUALIZATIONS

58%

DATA ANALYSIS

69%

COMMUNICATING FINDINGS TO BUSINESS DECISION-MAKERS

61%

FEATURE EXTRACTION

43%

DATA PLATFORMS

30%

TEAM PROJECTS AND GUIDING

39%

COMMUNICATING WITH PEOPLE

20%

ETL

5%

PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

24%

ANALYTICS SOFTWARE

29%

DEVELOPING DATA PLATFORMS

30%

DATA ANALYSIS

69%

BASIC EXPLORATORY DATA ANALYSIS

69%

RESEARCH QUESTIONS ANALYSIS TO ANSWER

61%

ANALYTICS PROBLEMS TO BE SOLVED

47%

IDENTIFYING BUSINESS RESEARCH QUESTIONS

31%

CONDUCTING DATA ANALYSIS

30%

DEVELOPING PROTOTYPE MODELS

43%

DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

5%

DEVISING ALGORITHMS INTO PRODUCTION

36%

IMPLEMENTING MODELS/ALGORITHMS INTO PRODUCTION

39%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

24%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

20%

TEAM PROJECTS AND GUIDING

39%

COMMUNICATING WITH PEOPLE

20%

ETL

5%

PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

24%

ANALYTICS SOFTWARE

29%

DEVELOPING DATA PLATFORMS

30%

DATA ANALYSIS

69%

BASIC EXPLORATORY DATA ANALYSIS

69%

RESEARCH QUESTIONS ANALYSIS TO ANSWER

61%

ANALYTICS PROBLEMS TO BE SOLVED

47%

IDENTIFYING BUSINESS RESEARCH QUESTIONS

31%

CONDUCTING DATA ANALYSIS

30%

DEVELOPING PROTOTYPE MODELS

43%

DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

5%

DEVISING ALGORITHMS INTO PRODUCTION

36%

IMPLEMENTING MODELS/ALGORITHMS INTO PRODUCTION

39%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

24%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

20%

TEAM PROJECTS AND GUIDING

39%

COMMUNICATING WITH PEOPLE

20%

ETL

5%

PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

24%

ANALYTICS SOFTWARE

29%

DEVELOPING DATA PLATFORMS

30%

DATA ANALYSIS

69%

BASIC EXPLORATORY DATA ANALYSIS

69%

RESEARCH QUESTIONS ANALYSIS TO ANSWER

61%

ANALYTICS PROBLEMS TO BE SOLVED

47%

IDENTIFYING BUSINESS RESEARCH QUESTIONS

31%

CONDUCTING DATA ANALYSIS

30%

DEVELOPING PROTOTYPE MODELS

43%

DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

5%

DEVISING ALGORITHMS INTO PRODUCTION

36%

IMPLEMENTING MODELS/ALGORITHMS INTO PRODUCTION

39%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

24%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

20%

TEAM PROJECTS AND GUIDING

39%

COMMUNICATING WITH PEOPLE

20%

ETL

5%

PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

24%

ANALYTICS SOFTWARE

29%

DEVELOPING DATA PLATFORMS

30%

DATA ANALYSIS

69%

BASIC EXPLORATORY DATA ANALYSIS

69%

RESEARCH QUESTIONS ANALYSIS TO ANSWER

61%

ANALYTICS PROBLEMS TO BE SOLVED

47%

IDENTIFYING BUSINESS RESEARCH QUESTIONS

31%

CONDUCTING DATA ANALYSIS

30%

DEVELOPING PROTOTYPE MODELS

43%

DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

5%

DEVISING ALGORITHMS INTO PRODUCTION

36%

IMPLEMENTING MODELS/ALGORITHMS INTO PRODUCTION

39%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

24%

DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

20%

TEAM PROJECTS AND GUIDING

39%

COMMUNICATING WITH PEOPLE

20%

ETL

5%

PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

24%

ANALYTICS SOFTWARE

29%

DEVELOPING DATA PLATFORMS

30%

DATA ANALYSIS

69%

BASIC EXPLORATORY DATA ANALYSIS

69%

RESEARCH QUESTIONS ANALYSIS TO ANSWER

61%

ANALYTICS PROBLEMS TO BE SOLVED

47%

IDENTIFYING BUSINESS RESEARCH QUESTIONS

31%

CONDUCTING DATA ANALYSIS

30%

DEVELOPING PROTOTYPE MODELS

43%

DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

5%
What is their annual income?

SQL > R > Python

Cluster Analysis:
- Python > R: data scientists
- R > Python: analysts

Python users had higher salaries.

Highest Paid?
- Scala

 Salary Depends on Location and Industry
Intermission

5 Minute Break.

Ask a neighbor:
- What is your name?
-tabs or Spaces...?
- What do statisticians and pirates have in common?

Contemplate:
- What are the ethics of data science?
- Can data do harm?
- What do you want to get out of Data 100?

Pirates say

Important Administrative Reminders

- There will not be any labs or sections this week.
- We will be computing optimal assignments for lab & section
- complete the online section assignment poll
  - https://goo.gl/forms/YohOCvkrUia4zTeI2
- Signup for the DS100 Sp18 Piazza Page
  - https://piazza.com/berkeley/spring2018/ds100/home
- Homework 1 will go out next week and be due the following week.
- You may start to setup your Python environment!

What are your goals for DS100?

- What do you want to learn?
- How does this class fit into your future plans?

Our Goals

Prepare students for advanced Berkeley courses in data-management, machine learning, and statistics, by providing the necessary foundation and context

Enable students to start careers as data scientists by providing experience in working with real data, tools, and techniques.

Empower students to apply computational and inferential thinking to address real-world problems

What are the Prereqs. for Data 100

- Officially Listed Prerequisites:
  - Foundations in Data Science [Data8]
  - Computing (CS61a or CS88 or ... E7]
  - Calculus and Linear Algebra (Math 54 or EE16a or Stat 88)
- We will not be enforcing prerequisites
  - however you should be familiar with the material in these classes (especially Data8)
- Homework 1 will help verify your familiarity
  - Do Hw1 and skim the Data8 textbook:
    - https://www.inferentialthinking.com
What will I learn?

Topics covered in Data 100

- Data collection and sampling
- Data cleaning and manipulation
- Regular Expressions
- SQL and Enterprise Data Management
- Xpath and web-scraping
- Exploratory Data Analysis & Visualization
- Hypothesis Testing & Confidence Int.
- Model design & loss formulation
- Batch and Stochastic Gradient Descent
- Ordinary Least Squares Regression
- Logistic Regression
- Feature Engineering
- The Bias - Variance Tradeoff & overfitting
- Regularization & Cross validation

We will use **Real Data**

Homework, labs, and in class examples will build on real data:

- Twitter, Speeches, Scientific Data, Maps, Surveys, Images, ...

The data will be:

- **messy** and you will have to clean it
- **big(ish)** and you will have to be a little clever to process it
- **complicated** and you will have to learn about the domain

You will Learn How to Use Real Tools

- Focus on Python programming language
- We will use various different technologies
  - Jupyter notebooks, pandas, numpy, matplotlib, postgres, seaborn, scikit-learn, plotly, Dask, ...
- **We won’t** teach you everything ...
- **You will learn to** read documentation
- **You will learn to** teach yourself
- **BETA WARNING:** Things will break ...
- **You will learn how to** debug
- **You will learn how to** get help (on Piazza)

Reading and Reference Materials

No single great book (working on a Data 100 gitbook …)

- Lectures slides and screencasts will be available online
- Use online reference materials

We will occasionally (in a few lectures) reference a few ebooks

- Joel Grus. “Data Science from Scratch”  [ebook link]
- Cathy O’Neil and Rachel Schutt. “Doing Data Science” [ebook link]
- G. James, D. Witten, T. Hastie and R. Tibshirani. “An Introduction to Statistical Learning.”  [pdf link]
- Wes McKinney. “Python for Data Analysis”  [pdf link]

**Grades**

- [20%] 6 Homework assignments (drop the lowest)
- [10%] 2 Projects (multi-week homework’s)
- [10%] Labs (Graded on Completion)
- [5%] Vitamins (weekly online quizzes)
- [5%] In class participation
  - Participate in at least 18 of the lectures for full credit.
  - Using google forms or bcourses (bring a browser)
- [20%] 1 Midterm (in class)
- [30%] 1 Final
On Time Policy (don’t be late)

- **5 days** of “slip-time” to be used on homework/projects for unforeseen circumstances (e.g., get sick or deadline conflicts)
- After you have used your slip-time budget
  - **20%** per day for each late day
- If you are having trouble finishing assignments on time let us know!

Collaboration Policy: **Don’t Cheat!**

- Data Science is a collaborative activity
- You may discuss problems with friends
  - List their names at the top of your assignments
- We may periodically analyze the collaboration networks
- **You must write your solutions individually**

**Don’t Cheat**

- Content in the homework and vitamins will be on the midterm and final
- If you are struggling let us know so we can help!

Staying Up To Date

- All communication will be through Piazza
  - [https://piazza.com/berkeley/spring2018/cs100/home](https://piazza.com/berkeley/spring2018/cs100/home)
  - If you have questions about assignments
    - Try commenting on the appropriate discussion
  - Do not share your code publicly
  - If you have private question write a private post on Piazza
    - This will ensure a quick response
- We will also be updating the website with links to homework, lectures, and vitamins
  - [http://www.cs100.org/sp18/](http://www.cs100.org/sp18/)

Data Science Lifecycle

*High-level description of the data science workflow*

**Question / Problem Formulation**

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics of success?
Data Acquisition and Cleaning

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?

Obtain Data

Understand the Data

- How is our data organized and what does it contain?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

Exploratory Data Analysis & Visualization

What does the data say about the world?

Does it answer our questions or accurately solve the problem?

How robust are our conclusions and can we trust the predictions?

Understand the World

Predictions and Inference

Reports, Decisions, & Solutions

Data Science Demo