Computational Thinking and Inferential Thinking

Foundations of Data Science

Michael Jordan
michael_jordan@berkeley.edu

University of California, Berkeley

November 2020

Our Vision

- All students should be able to make decisions based on data
- Data science education should be a conceptual blend of traditions in statistics and computer science, with a strong outward-looking component
- All students should be able to tackle significant problems, the complexity increasing with the students' background knowledge and interests

The Vision, Realized

- Foundations of Data Science (data8.org)
  - Connector courses
- Principles and Techniques of Data Science (data100.org)
- Probability for Data Science (prob140.org)
- Data, Inference, and Decisions (data102.org)
- Human Context and Ethics (https://data.berkeley.edu/file/2062)

Data, Inference, and Decisions

DS 102: Data, Inference, and Decisions
Lecture 1
Michael Jordan
University of California, Berkeley
306B

The Students

2019-2020

- 2800 in Data 8
- 1600 in Data 100
- 500 in Prob 140
- 500 in Data 102

The first Data Science majors graduated in Spring 2019.

Over 300 DS majors graduated in Spring 2020, along with several hundred who got the DS minor.

Data 8: *Foundations of Data Science*

- In the 2015-2016 academic year, we taught over 500 freshman-level students in our new Data Science course
  - – a wide range of students, pre-disciplinary
  - – it was a major success

- In the 2016-2017 academic year, we taught over 1000 students, and over 2500 in 2019-2020

- The course is a conceptual blend of traditions in statistics and computer science, with a strong outward-looking component (“connector courses”)

Some shapers of our perspective

- The increasingly prominent use of “data science” in job descriptions in the tech industry
  – and our undergraduates’ awareness of the term
- The rise of the “CS for All” movement
  – and its partial failure at many universities
- Success in teaching inference at the graduate level via resampling and sampling ideas
  – surely this can work at the undergrad level too?
- The explosive growth of iPython (Jupyter)
  – and related browser-based computing environments

Computational thinking

- Modern programming and problem-solving with a computer requires a new style of thinking
  - – the core concepts: abstraction, modularity, scalability, robustness, etc.
  - – concepts that are taught beautifully and successfully in freshman-level courses in computer science
- What’s missing in computational thinking:
  - – much of what we teach in our best statistics courses

Inferential thinking

- Analyzing data requires more than just computation, but it also requires a style of thinking
  - – consider the data-generating process behind the data
  - – being clear on the sampling pattern, with its possible biases
  - – understanding how to do causal inference
  - – understanding how to capture uncertainty

- What’s missing in inferential thinking:
  - – much of what we teach in our best computer science courses

Blending

- We believe that the blend of computational thinking and inferential thinking is:
  - powerful
  - engaging to students, by empowering them to solve real problems
  - what industry and government are really asking for when they refer to “data science” or “AI”
  - a kind of liberal arts for the 21st century

Analyzing data: three classical steps

- Formulate the question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and calculations
- Interpretation of the results in the language of the domain, ideally without statistical jargon

Classical teaching of data analysis

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and formulas
- Interpretation of the results in the language of the domain, without statistical jargon

The Data 8 way

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and computation
- Interpretation of the results in the language of the domain, without statistical jargon

The approach

- A course for freshmen
- No prerequisites

- Teach data science as a way of thinking:
  - computational thinking
  - inferential thinking
  - these are blended in every lesson

The Syllabus

- Cause and Effect
- Tables
- Data Types
- Census
- Histograms, Charts
- Functions
- Joins
- Iterations
- Chance
- Sampling
- Models
- Comparing Distributions
- Decisions and Uncertainty
- P-Values
- A/B Testing
- Causal Inference
- Confidence Intervals
- Designing Experiments
- Regression
- Classification

Three 2-week projects

- **Water use in California:** mapping the water districts and overlaying IRS taxable income data by zip code
- **Murder rates and the death penalty:** nonparametric inference; the importance of visualization
- **Classification of song lyrics:** hip-hop or country?

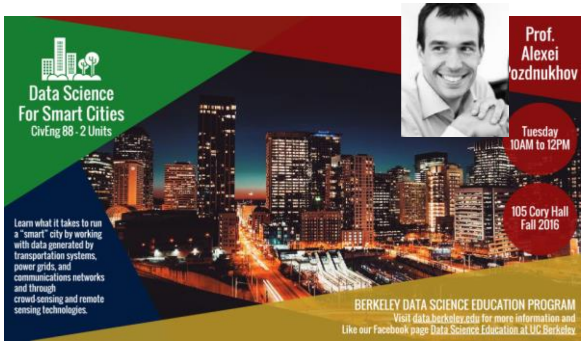
“Connector” Courses

- Data science in a particular domain
- Much smaller: 5-50 students
- Half the time commitment of a typical undergraduate course
- Can be taken concurrently with Data 8, or later

Connector Courses, 2015-2016

- Data Science for Smart Cities
- Data Science and the Mind
- Data Science in Ecology and the Environment
- Exploring Geospatial Data
- How Does History Count?
- Data and Ethics
- Health, Human Behavior, and Data
- Race, Policing, and Data Science
- Literature and Data
- Matrices and Graphs in Data Science
- Computational Structures in Data Science
- Probability and Mathematical Statistics for Data Science

Connections Fall 2016



Data Science For Smart Cities
CivEng 88 - 2 Units

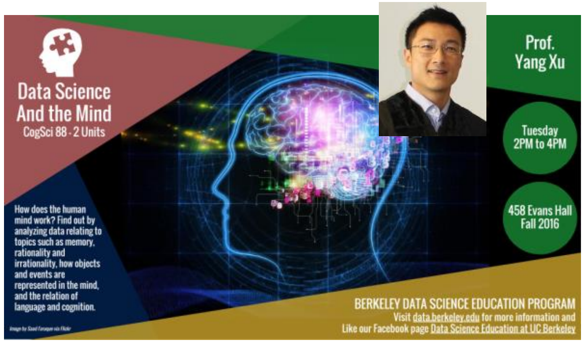
Learn what it takes to run a "smart" city by working with data generated by transportation systems, power grids, and communications networks and through crowd-sensing and remote sensing technologies.

Prof. Alexei Pozdnukhov

Tuesday 10AM to 12PM

105 Cory Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Data Science And the Mind
CogSci 88 - 2 Units

How does the human mind work? Find out by analyzing data relating to topics such as memory, rationality and irrationality, how objects and events are represented in the mind, and the relation of language and cognition.

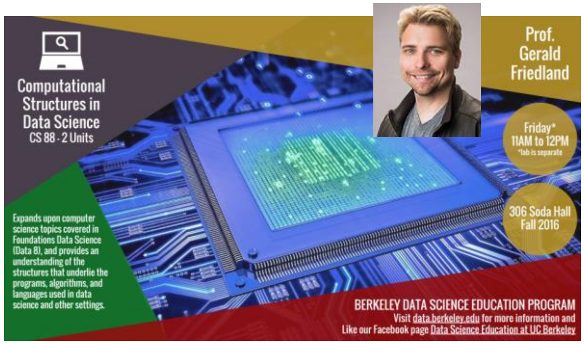
Image by Saad Faruque via Flickr

Prof. Yang Xu

Tuesday 2PM to 4PM

458 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Computational Structures in Data Science
CS 88 - 2 Units

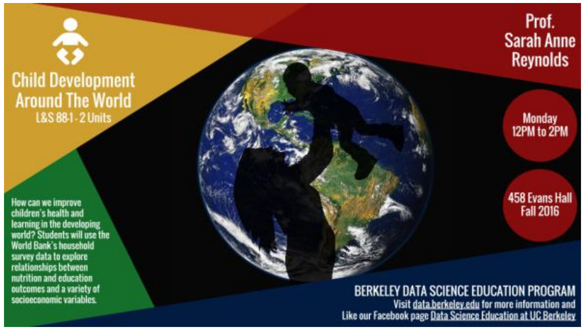
Expands upon computer science topics covered in Foundations Data Science (Data 8), and provides an understanding of the structures that underlie the programs, algorithms, and languages used in data science and other settings.

Prof. Gerald Friedland

Friday* 11AM to 12PM
*lab is separate

306 Soda Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Child Development Around The World
L&S 88-1 - 2 Units

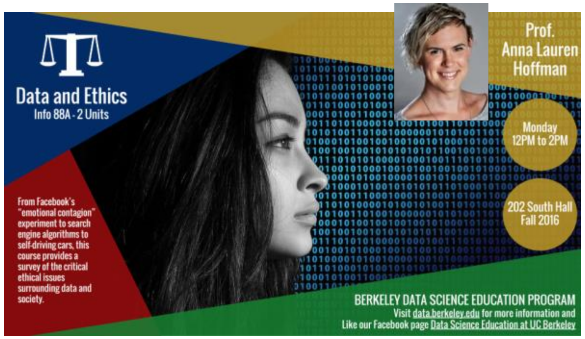
How can we improve children's health and learning in the developing world? Students will use the World Bank's household survey data to explore relationships between nutrition and education outcomes and a variety of socioeconomic variables.

Prof. Sarah Anne Reynolds

Monday 12PM to 2PM

458 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Data and Ethics
Info 88A - 2 Units

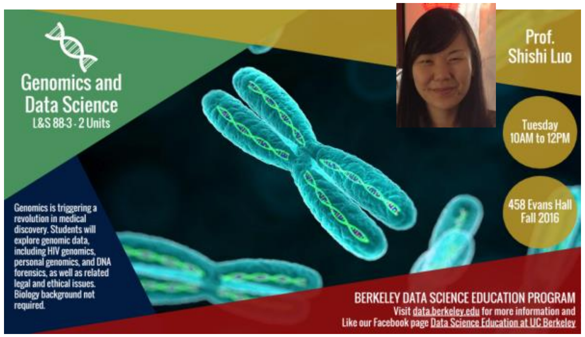
From Facebook's "emotional contagion" experiment to search engine algorithms to self-driving cars, this course provides a survey of the critical ethical issues surrounding data and society.

Prof. Anna Lauren Hoffman

Monday 12PM to 2PM

202 South Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Genomics and Data Science
L&S 88-3 - 2 Units

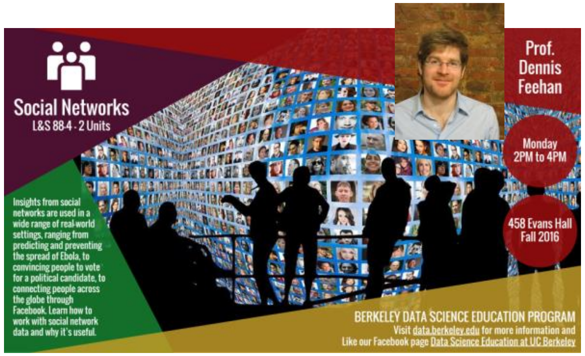
Genomics is triggering a revolution in medical discovery. Students will explore genomic data, including HIV genomics, personal genomics, and DNA forensics, as well as related legal and ethical issues. Biology background not required.

Prof. Shishi Luo

Tuesday 10AM to 12PM

458 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Social Networks
L&S 88-4 - 2 Units

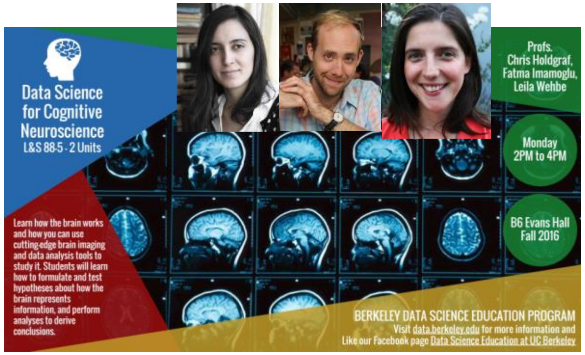
Insights from social networks are used in a wide range of real-world settings, ranging from predicting and preventing the spread of Ebola, to convincing people to vote for a political candidate, to connecting people across the globe through Facebook. Learn how to work with social network data and why it's useful.

Prof. Dennis Feehan

Monday 2PM to 4PM

458 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Data Science for Cognitive Neuroscience
L&S 88-5 - 2 Units

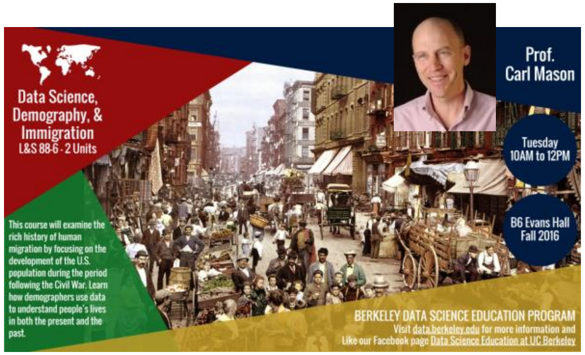
Learn how the brain works and how you can use cutting-edge brain imaging and data analysis tools to study it. Students will learn how to formulate and test hypotheses about how the brain represents information, and perform analyses to derive conclusions.

Profs. Chris Holdgraf, Fatma Imamoglu, Leila Wehbe

Monday 2PM to 4PM

B6 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley



Data Science, Demography, & Immigration
L&S 88-6 - 2 Units

This course will examine the rich history of human migration by focusing on the development of the U.S. population during the period following the Civil War. Learn how demographers use data to understand people's lives in both the present and the past.

Prof. Carl Mason

Tuesday 10AM to 12PM

B6 Evans Hall Fall 2016

BERKELEY DATA SCIENCE EDUCATION PROGRAM
Visit data.berkeley.edu for more information and
Like our Facebook page Data Science Education at UC Berkeley

Data 8 Spring 2016 survey

- About 450 students, 418 responses
  - 42% women
  - 17% underrepresented ethnic or racial minority
  - 52 different intended majors
- Does the course fulfill a major requirement?

Yes:	8.6%
Maybe:	10.4%
I don’t know:	10.6%
No:	70.4%

Always DataBears

Of the roughly 1210 students in Foundations of Data Science in Fall 2018, about 70 worked on the Spring 2019 offering of the course and its connectors, as lab assistants, tutors, graders, and course developers.

Environment and course materials

- Jupyter notebooks; Python 3

- JupyterHub
  - Multi-user server for Jupyter notebooks
  - Browser-based computation in the cloud

Course materials: data8.org

Source: https://github.com/data-8

Data, Inference, and Decisions

DS 102: Data, Inference, and Decisions
Lecture 1
Michael Jordan
University of California, Berkeley
306B

Data 102: *Data, Inference and Decisions*

- In the 2019-2020 academic year, we initiated a new senior-level course in Data Science course

- The course is again a conceptual blend of ideas from statistics and computer science, with a strong outward-looking component

- But it also brings in economics and control theory, offering an integrated perspective on decision-making under uncertainty

The Syllabus

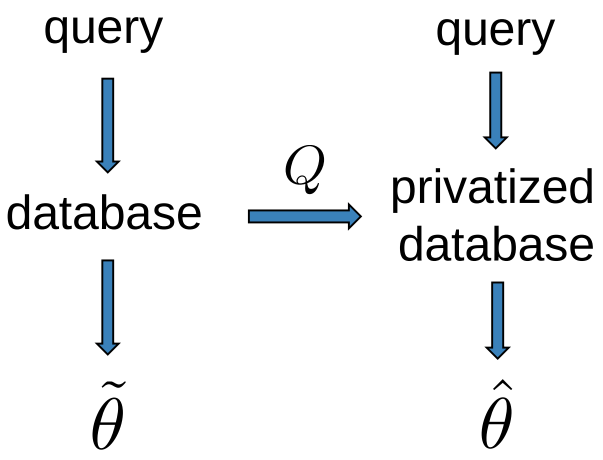
- False Discovery Rate Control
- Online False Discovery Rate Control
- Bayesian and Frequentist Decision-Making
- Probability Interpretation of Linear and Logistic Models
- Bayesian Hierarchical Models
- Importance Sampling and Gibbs Sampling
- Bootstrap
- Concentration Inequalities
- Causal Inference
- Experimental Design
- Bandits: Greedy and UCB Algorithms
- Thompson Sampling
- Matching Markets
- Game Theory
- Dynamic Programming and Q-Learning
- Introduction to LQR and Control Theory
- Privacy

Example: Privacy

Privacy and Data Analysis

- Individuals are not generally willing to allow their personal data to be used without control on how it will be used and now much privacy loss they will incur
- “Privacy loss” can be quantified (e.g.) via differential privacy
- We want to trade privacy loss against the value we obtain from “data analysis”
- The question becomes that of quantifying such value and juxtaposing it with privacy loss

Privacy



query

query

Q

database

privatized database

$\tilde{\theta}$

$\hat{\theta}$

Classical problem in differential privacy: show that $\hat{\theta}$ and $\tilde{\theta}$ are close under constraints on Q

Inference

$$\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] \\
P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[d] \\
\theta & \tilde{\theta}
\end{tikzcd}$$

Classical problem in statistical theory: show that $\tilde{\theta}$ and θ are close under constraints on S

Privacy and Inference

$$\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] & \text{query} \arrow[d] \\
P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[r, "Q"] \arrow[d] & \text{privatized database} \arrow[d] \\
\theta & \tilde{\theta} & \hat{\theta}
\end{tikzcd}$$

The privacy-meets-inference problem: show that θ and $\hat{\theta}$ are close under constraints on Q and on S

Conclusion

- Data Science involves blending computational thinking with inferential thinking

- Education in Data Science involves teaching this blend, in the context of solving real-world, substantive problems

- I believe that Data Science is a core part of a modern liberal education

Computational Thinking and Inferential Thinking

Foundations of Data Science

Michael Jordan
michael_jordan@berkeley.edu

University of California, Berkeley

November 2020