Welcome to CompSci 116: Foundations of Data Science

Jeff Forbes

January 10, 2019

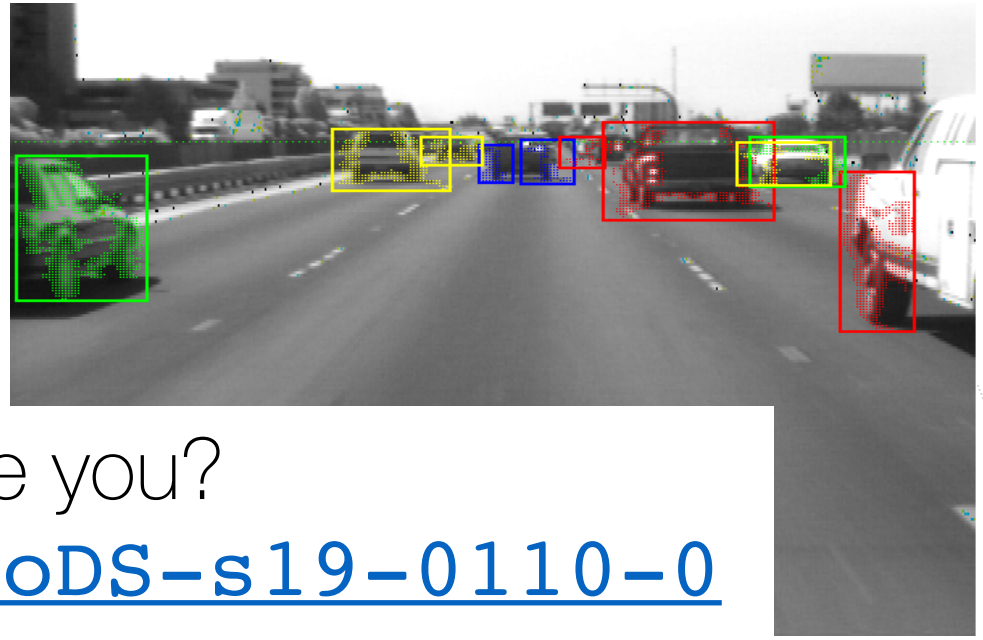
Plan For The Day (PFTD)

- Be able to articulate whether FoDS is the right course for you, in terms of being able to complete it with understanding
- Be able to describe what data science is and what the implications are of the work of data scientists
- Be able to explain what work is expected: in-class team work, homework, projects, and exams

Acknowledgements

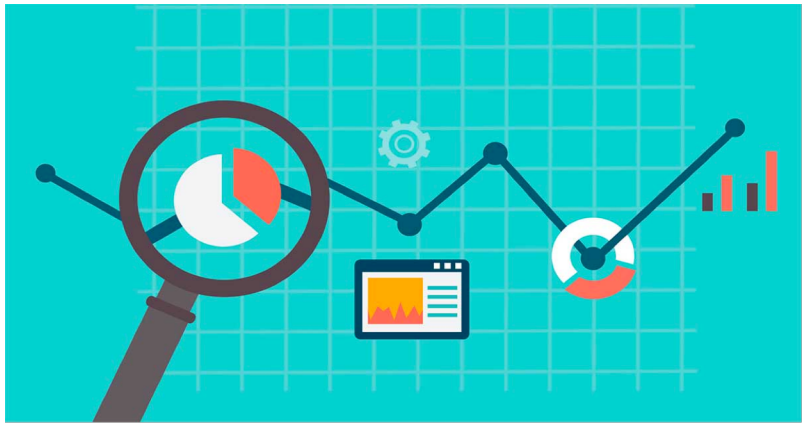
- Adaption of Data 8
  - Ani Adhikari, John DeNero, and David Wagner + a lot of staff at the University of California, Berkeley.
  - Materials used with permission
- NAS Report on Envisioning the Data Science Discipline
- Slides from Mike Franklin & Data Science in the 21st Century at CRA Snowbird 2016
- Thanks to Mine Cetinkaya-Rundel, Kristin Stephens-Martinez, Max Bartlett, & Jose San-Martin

Who am I?



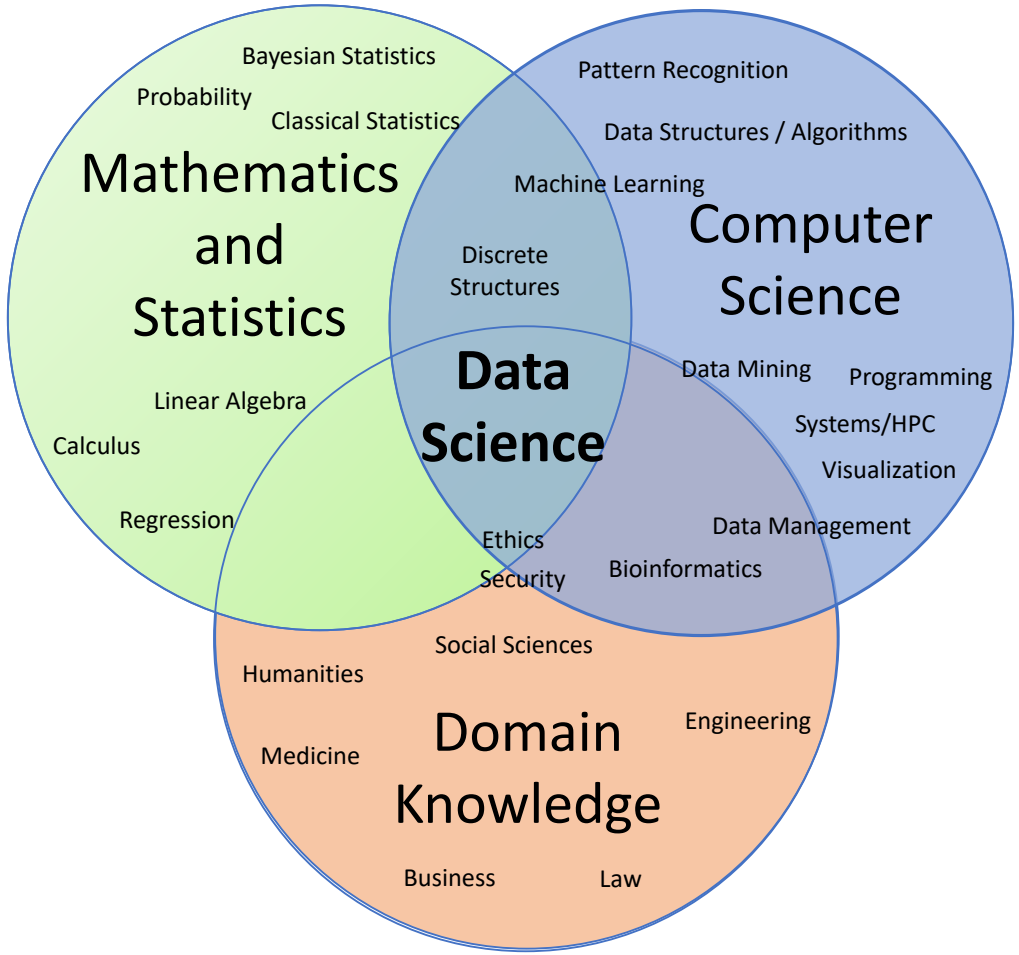
Who are you?

http://bit.ly/FoDS-s19-0110-0



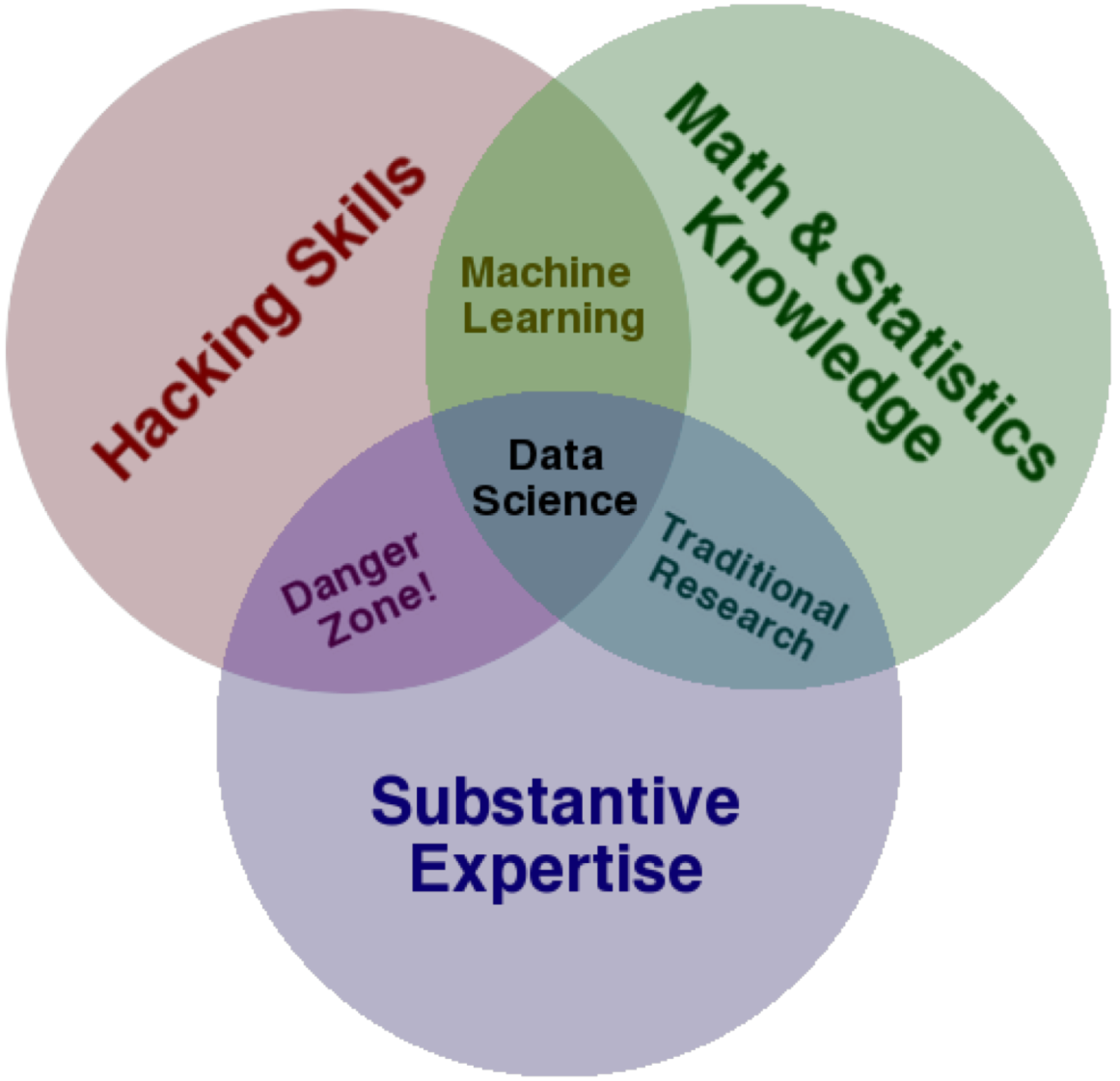
What is Data Science?

- From the ACM Taskforce on Data Science Curricula
- Draws from many different disciplines



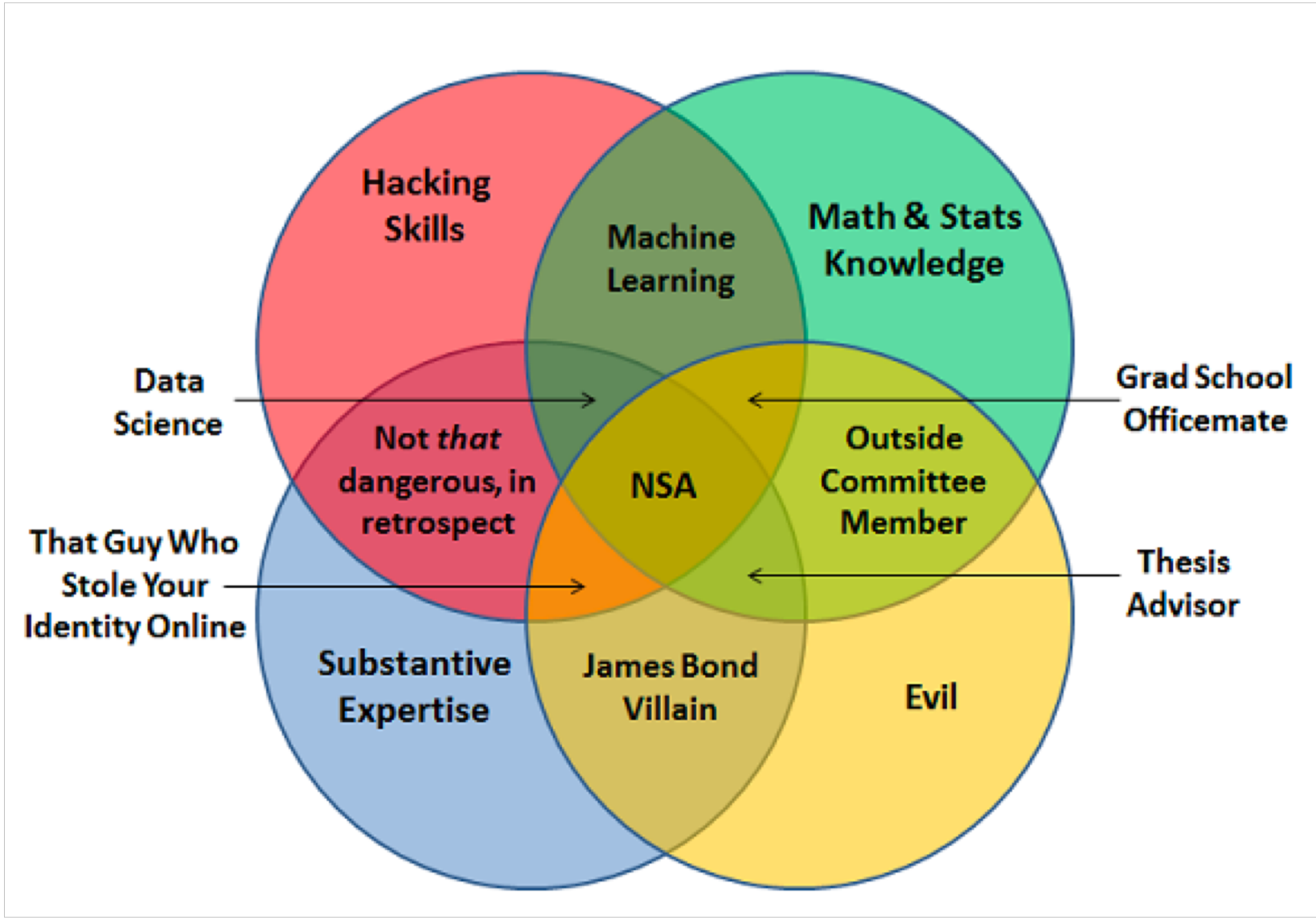
What is Data Science?

- From Drew Conway
- Skill-based



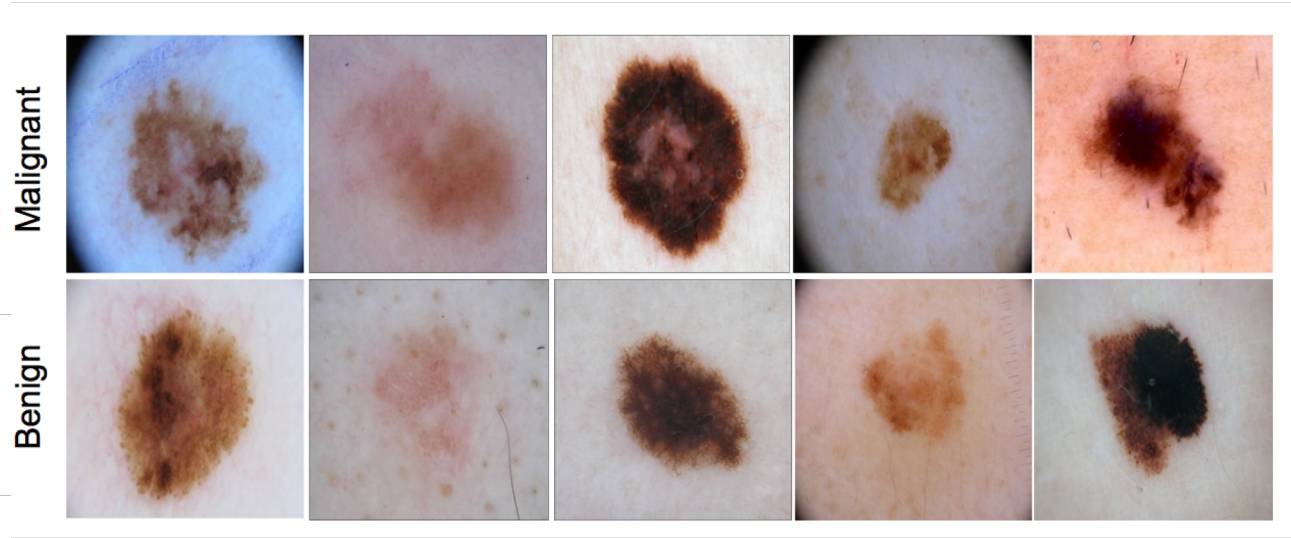
What is Data Science?

- From Joel Grus
- Don't use data science for evil



Applications of Data Science

- **Medicine**: Melanoma detection
  - [Codella et al 2017]
- **Business**:



- **Smart Cities:**

What applications of data science interest you?

http://bit.ly/FoDS-s19-0110-1

What does a data scientist do?

- From Peter Skomoroch
- The Best Job in America for the past three years?
- A global shortage in data scientists?
- Lots of buzz and buzzwords?

Ethical & Social Implications of Data

- Fairness
  - Consider equity and avoid bias that may be inherent in data sets
  - **Example**: Sentencing practices for criminal justice
- Validity
  - Data set should contain accurate and relevant information. Context matters!
  - **Example**: Survivor bias in analyzing STEM degree production
- Data confidence
  - Don’t draw stronger-than-appropriate conclusions
  - **Example**: Stock market predictions
- Privacy
  - Must be good stewards of data
  - Consider how data is collected and analyzed

Latanya Sweeney

- **Prof. Government and Technology @ Harvard**
- **Former CTO of the FTC**

I am a computer scientist with a long history of weaving technology and policy together to remove stakeholder barriers to technology adoption. My focus is on "computational policy" and I term myself a "computer (cross) policy" scientist. I have enjoyed success at creating technology that weaves with policy to resolve real-world technology-privacy clashes.

http://latanyasweeney.org/

- k-Anonymity: each subject cannot be distinguished from at least k-1 others
- Identify 87% of US population using (dob,zip,gender).

What is Foundations of Data Science?

Drawing *useful* conclusions from data using computation

- **Exploration**
  - Identifying patterns in information
  - Uses visualizations
- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization
- **Prediction**
  - Making informed guesses
  - Uses machine learning

Course Details

- https://www.cs.duke.edu/courses/compsci116/spring19/
- Thursdays: Active Lecture
  - Reading assigned
- Tuesdays: Team-Based Learning in Lab
- Midterm Exams: **2/19** & **4/18**
- Weekly Homework: Discuss with your team but submit individually
- 3 Projects: Work in pairs
- Final Project: Work in pairs – Present on **5/3**

Team-Based Learning

- Why?
  - Facilitate collaboration
  - Problem solving accompanied by group interaction promotes learning
- Do reading outside of class
- Readiness Assurance
  1. Individual
  2. Team
- Application-focused Exercise
- End of Semester: Peer Evaluation

How will you learn?

- Learn by doing
  - Learn computing concepts by doing interesting things on data
  - Learn statistical concepts by observing what's interesting
  - Learn domain knowledge just in time

- Minimal setup: Jupyter Notebooks (https://jupyterhub.cs.duke.edu)

Is FoDS the right course for you?

- Yes if:
  - You're interested in gaining quantitative (QS) or computational skills.
  - You want to understand and develop points of view based on the analysis of data as well as evaluate arguments made by others
- Probably not if:
  - You've already taken a number of computer science and statistics courses. CompSci 216 – Everything Data may be more appropriate?
  - You've already taken Stat 199
  - Want a course that satisfies an elective requirement for Stats or CompSci
- Ask me!

What's next?

- Review Chapter 1 of *Computational and Inferential Thinking*

- Tell a friend
  - There's still space!