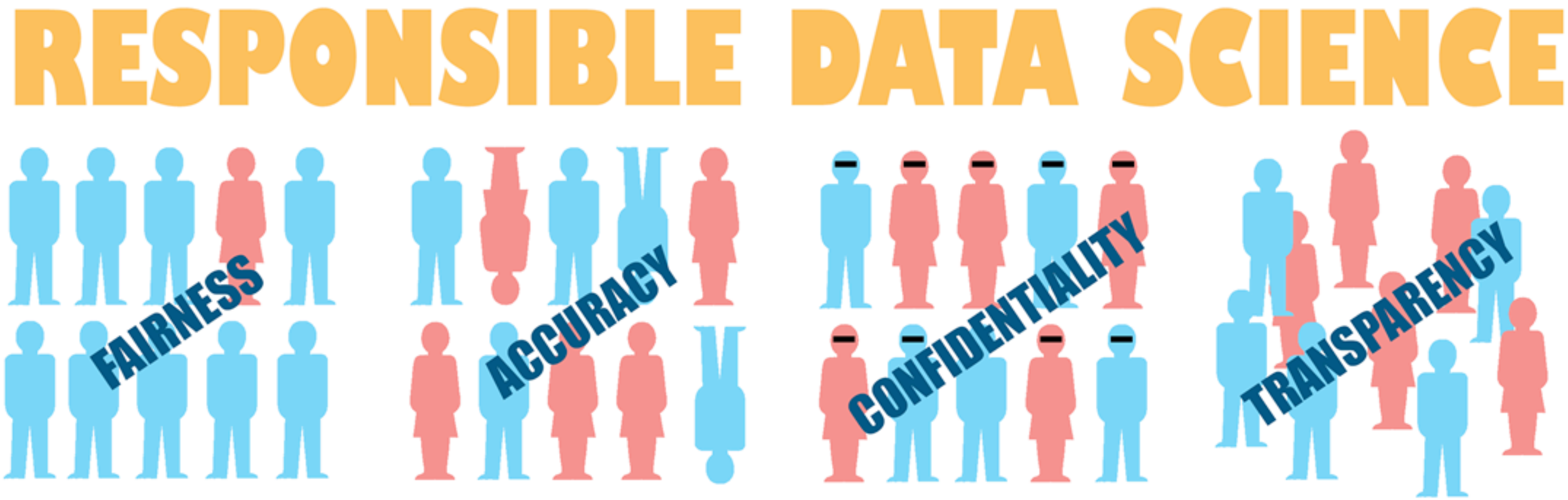
CMSC 290C: Responsible Data Science

Spring 2018

W 1:20-4:20

E2 506



Today

- Short Introduction
- Course Overview and Logistics
- Introductions
  - Find out who’s here, interests, and background
- Responsible Data Science 101, Abel Rodriguez

Introduction to Responsible Data Science

What is Responsible Data Science (RDS)?

- What do you think?

What is Responsible Data Science (RDS)?

- This is the question that we will unpack throughout the class!

Basic Components

- Literacy
  - Statistical Literacy
  - Computational Literacy
  - Domain Literacy

Statistical Literacy

- Understand the core assumptions behind most statistical inference procedures:
  - Sampling
    - Random sample
    - Independent, Identically Distributed (IID)
  - Stationarity
  - Differences between descriptive modeling, predictive modeling and causal modeling
- Common pitfalls
  - Overfitting
  - Curse of dimensionality
  - Need to quantify uncertainty

Computational Literacy

- Computational complexity
  - Sublinear, polynomial, exptime
  - Dimensionality reduction
  - Tree-width (applicable to constraint optimization & graphical models problems)
- Optimization
  - Formulation of optimization problems
  - Optimization algorithms
    - Dynamic programming, message passing, stochastic search
  - Performance: worst case, avg case, etc.
- Data representation
  - Tables, trees, graphs & relational
  - Data cleaning, deduplication
  - Normalization
  - Data provenance
  - Locality, indexes
- Distributed processing
  - Data vs. control
  - Fault tolerance
  - Performance variability
- Human in the Loop
  - Visualization
  - Active learning
  - Explanation
- Privacy, security

Domain Literacy

- Understanding the problem
  - Is it even a data science problem?
  - Requires dialog, collaboration, respect
  - Typically an iterative process
- Many compelling data science problems are about people
  - Understanding social science (psychology, sociology, communication studies, economics, geography, education, anthropology, etc.) becomes important
  - Emerging area of computational social science (CSS)
- Even domains that do not appear to be about people, have surprising structures that benefit from CSS ideas
  - Environmental problems, computational biology, material science, any experimental data

Emerging RDS Research Areas

- Privacy & Data Ownership
- Fairness, Accountability & Transparency
- Interpretability
- Reproducibility
- Ethics

Introductions

- **Basics:** Name, department, year, advisor, research topic if you have one
- **Where:** ugrad school/where you are from
- **Background:** familiarity with ML, responsible data science topics
- **Icebreaker question:** What is your best Santa Cruz/UC Santa Cruz tip or recommendation?

Course Etiquette

- Please arrive on time
- Laptops and cell phones:
  - It is long class, hard to pay attention
  - Please always think about your neighbors
  - Please strictly limit your use, it can be very distracting
- Participate, participate, participate!!

Course Structure

- Highly collaborative!
- We will be learning/developing this material together
- Truth in advertising: I am a newbie to responsible data science, most of this is new to me!

Logistics

- Webpage:
  - https://cmps290c-spring18-01.courses.soe.ucsc.edu/home
- For assigned papers, we will do QCRs
  - Question, Comment and Research idea
  - These should be posted to the class discussion page before the start of class
- Occasionally we will do BYOPs
  - Bring Your Own Paper
  - These are sort of a free-for all, where we all get a quick exposure to A LOT of ideas

Workload

- Weekly:
  - Attending class
  - Reading papers & doing QCRs & BYOPs
- Course project (can be done in groups)
  - Literature review or research project
  - Highly encouraged to choose a topic that aligns with your research
  - Structure:
    - Initial proposal (1 paragraph): 4/18
    - Midterm proposal (1 page): 5/9
    - Final project (poster/presentation): 6/6

Readings

- Next week, two short papers:
  - *Critical Questions for Big Data*, dana boyd & Kate Crawford. Information, Communication & Society, 2012. link
  - *Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency*, Hanna Wallach, Medium, 2014. link

- Next week, I will not be here, so we will have someone else lead the class

For Next Week:

- Please introduce yourself online as well. Go to the web forum, under Introductions.   Add:
  - Your name, a short handle (can be your first name)
  - Department, research area, advisor if you have one
  - Any notes about areas you are interested to see covered
  - Please do this by next Wed latest.
- Please do QCRs for the two papers by start of class next Wed