COGS 9 - Intro to Data Science (Discussion) 09/28/2022
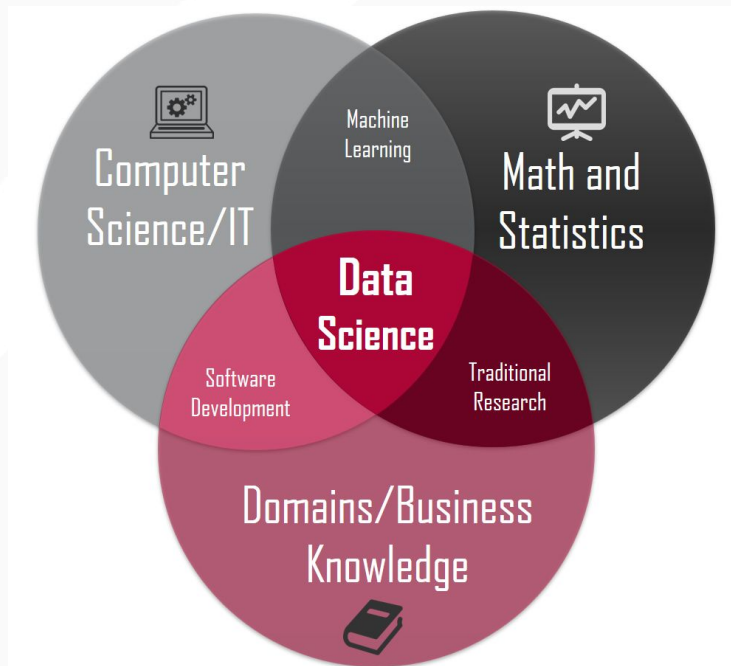
Ashwin Mishra, TA
Lindsey Gu, IA Jiesen Zhang, IA

What is this course about?

- Introduce students to the growing field of Data Science
- Provide students an overview of different fundamental parts of DS
- Help the students navigate the field and career opportunities

Donoho's Six-Part Definition

- Connects the discipline of DS to its 50+ years of history (John Tukey in 1960s)
- DS as an **extension** of statistics
- Traditional statistical analysis vs. cutting-edge tools and technologies
- Common Task Framework (e.g., Netflix Challenge)
- Six divisions for Greater Data Science (extension to the work of John Chambers and Bill Cleveland):
  a. Data Exploration and Preparation (eg, curating data, anomalies and outliers, data wrangling)
  b. Data Representation and Transformation (eg, Data storage in SQL and noSQL databases)
  c. Computing with Data (eg, Python and R to model and analyze data)
  d. Data Visualization and Presentation (eg, Advanced charts for EDA)
  e. Data modeling (eg, overlap of statistics and predictive modelling)
  f. Science about Data Science (eg, investigates the real-world work "in the wild")

Due date: Oct 06 11:59 pm
Late due date: Oct 08 11:59 pm (with 50% credit)

Assignment 1 FAQs

- How specific should my hypothesis be?
- How complicated should my DS question be?
- Do the citations in the background research be from scholarly sources only?
- What if I can't find the proper dataset?
- More?

Due date: October 14 11:59 pm

Group FAQs

- Small group size?
- Missing group members?
- Uncooperative group members?
- Unequal work contribution?
- More?