UNIVERSITY of WASHINGTON

Introduction to Data Science: Logistics

Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute

How this course is organized

- A “guided tour” of important trends and technologies
- A “deep dive” into selected must-know algorithms, techniques, and technologies
- A set of hands-on assignments to deliver specific skills and experiences

Prequisites

- We assume
  - some prior programming experience in some language
  - “muscle memory” with basic college statistics
  - some exposure to databases and database concepts
- One assignment will require writing SQL
- Two assignments will require writing Python
- One (optional) assignment will involve processing ~1TB of data using Amazon Web Services
  - You will pay for these resources, should you choose to complete the assignement
- One assignment will involve solving a prediction problem on kaggle.com using whatever tools you wish.
- Some understanding of distributed systems will be helpful, but not required

Learning Objectives

- The ability to describe the landscape of data science concepts, tools, algorithms, and technologies
- Hands-on experience in data manipulation, analysis and prediction
- You will be an “advanced beginner” in a variety of data science topics

Course Philosophy

- The skills needed by a data scientist span a variety of different areas
  - statistics, programming, databases, systems, visualization
- The traditional organization of topics is not ideal
  - It is difficult to acquire introductory-level knowledge in all areas
  - Cross-cutting concepts and abstractions are obscured
- Our goal: Expose and simplify the underlying commonalities between these areas

Non-goals for this course

- You will not emerge an expert in statistics
  - Though you will apply basic statistical methods
- You will not emerge an expert in machine learning
  - Though you will be familiar with some important concepts and will have the chance to exercise them
- You will not emerge an expert in databases and NoSQL
  - Though you will understand the concepts they share and know how to apply them
- You will not emerge an expert in R, Python, MapReduce, or SQL
  - Though you will use all of these in assignments

Quizzes and Assignments

- Short “finger exercise” quizzes after most video segments
- A set of full-length offline assignments
- Some assignments will be graded via peer assessment

My Background

- BS Industrial and Systems Engineering, GA Tech 99
- Consulting 99-01
  - Deloitte, Microsoft, Siebel, Schlumberger, Verizon
- Phd, Computer Science, Portland State University 01-07
- NSF Science and Technology Center for Coastal Margin Observation and Prediction (CMOP) 07-09
  - Data Architect, Research Scientist
- University of Washington 09-present
  - Affiliate Assistant Professor
    Computer Science & Engineering
  - Director of Research, Scalable Data Analytics
    eScience Institute