UNIVERSITY of WASHINGTON

Software Design for Data Science

Introduction

Melissa Winstanley
University of Washington
January 3, 2023

Me: your teacher

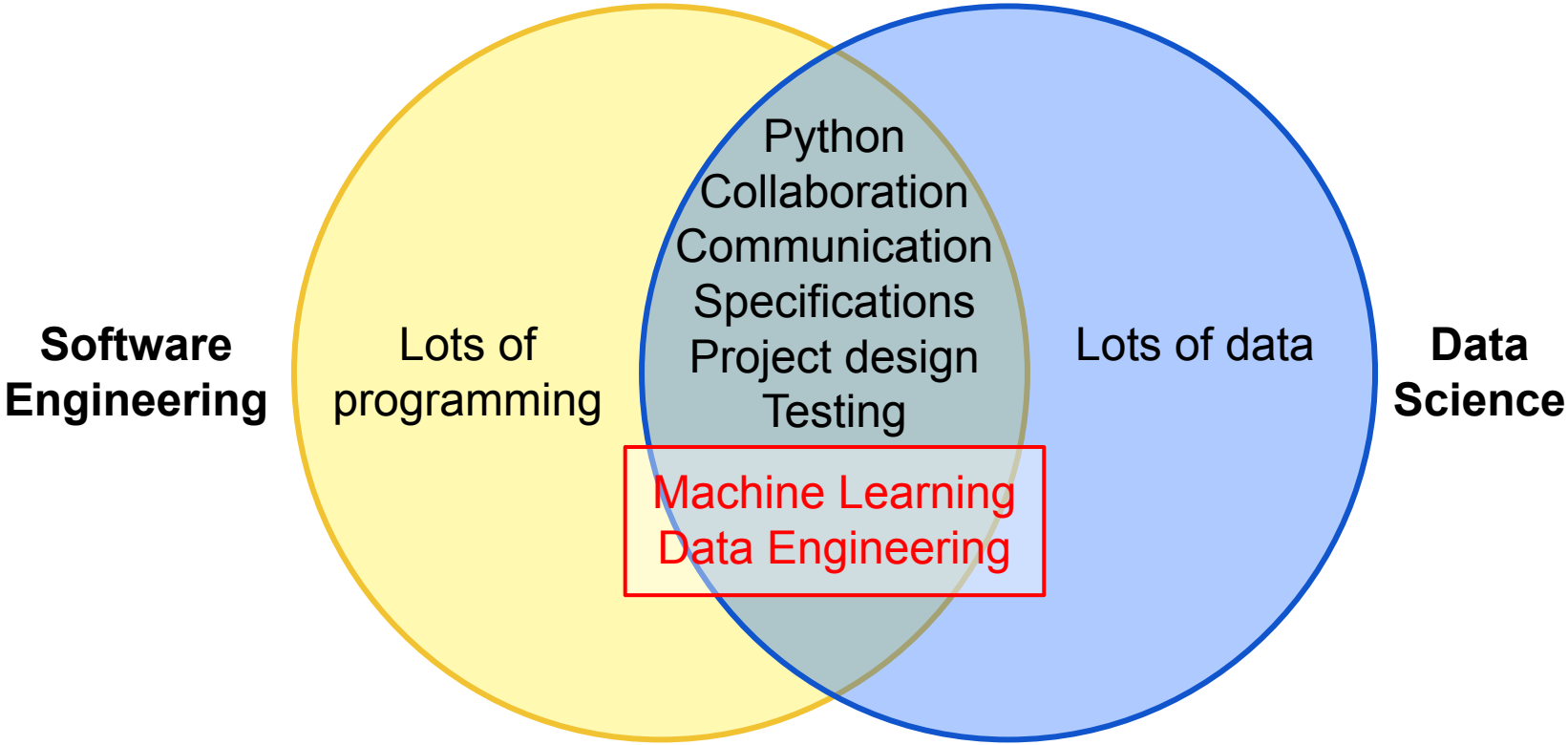
Google

Your teachers

Me: software engineer
TAs: data scientists

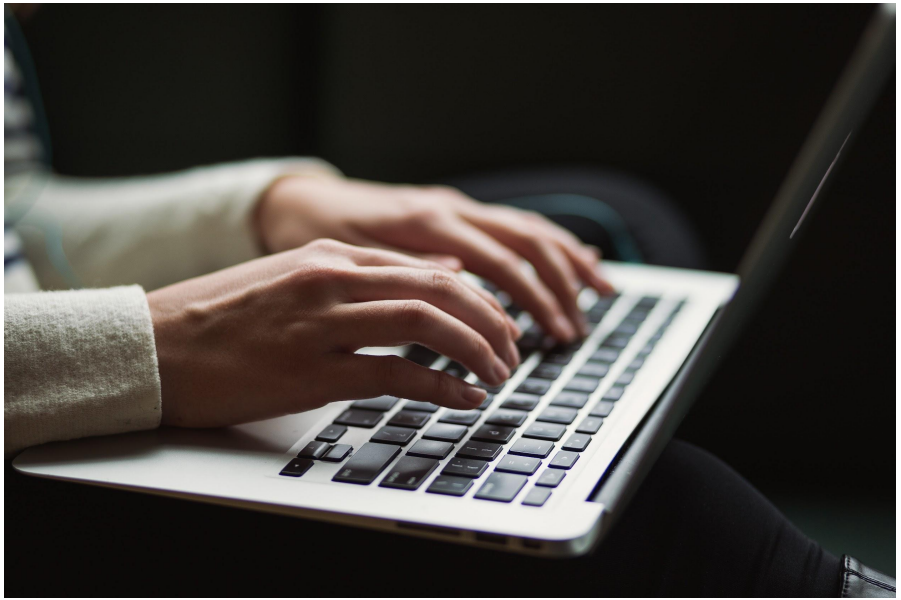
You: data scientists

Software Engineering vs Data Science



This course

- Practical
- Hands-on
- Useful after graduation
- Focus on HOW to learn, not just the skills



Course Website

https://uwdata515.github.io/

Linked from MyUW

+ Canvas for grades, surveys

+ Ed Discussions

Course Project

- Collaborative software engineering experience
- Teams of 3 to 4 with 4 being optimal
- Develop project using version control

Course Project

Collaborative software engineering experience

- Design (use cases, component specification)
- Documentation (how to, docstrings)
- Style (PEP8, pylint)
- Coding, testing & milestones
- Standup & code reviews

Project Type 1: Answer “Research” Questions

Problem statement: Answer two to three questions of business or scientific relevance

- Use a Jupyter notebook and supporting python files

Example

- Climate Police: Analyze effects of pollution on the planet.

Project Type 2: Create Reusable Data

Problem statement: Create data repository with tools (e.g., search, visualization, analytics)

Example

Car2Know: Provide car rental data to users of Car2Go (e.g., for planning trips)

Project Type 3: Create a Tool

Problem statement: Solve a problem common to many users

- Don’t reinvent the wheel

Example

BioReactor Data Logging – Monitor and publish data from BioReactor experiments

Getting Started

Students present statements of interest

Gather with like-minded students

Verify the project idea

Size the effort

Things to Think About

- Topics of interest
  - Is there an unmet need (i.e. no code already exists)?
  - Is there only commercial software available for a task?
  - What is the potential user base?
- Data you have access to NOW
  - How much you've used the data
  - Code you have to access the data
  - How clean the data are

Verify the Project Idea

- Is there an unmet need (i.e. no code already exists)?
- Clarity about the project type?
- Consensus on the problem being solved.
- Do you have data that can solve the problem?

More on the Data

- At least two non-trivial data sets
- Data need to be combined, joined, merged, etc. to answer the scientific questions
- Have access to the data NOW!

Some Public Data

http://drugbank.ca

http://toxnet.nlm.nih.gov

https://data.seattle.gov/Transportation/Traffic-Flow-Counts/7svg-ds5z

https://www.divvybikes.com/data

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

https://www.kaggle.com

Pronto bike data

American Fact Finder Data

European union data (World bank)

Russian federation data (World bank)

China data (World bank)

Data! Data! Data!

- At least two non-trivial data sets
- Data need to be combined, joined, merged, etc.

Think about your data NOW!

Project Ideation

Over the first few weeks:

- What areas are you interested in? E.g. social good or a job demo.
- What data are available in that space?
- What tools already exist in that space?
- What type of project is this? (answer research question, create reusable data, create a tool, other?)
- **Volunteer to give a one slide, 5 minute project idea pitch at the start of class!**

Academic Integrity

- **Software development is a highly collaborative endeavor**
- We expect you to collaborate, but your work is your own
- In software, there is rarely one correct solution to any problem
  - Standing around a white board brainstorming is OK
  - Directly copying code someone else in your class is not OK
- The point is for you to learn the concepts and copying answers can defeat that point