

Software Design for Data Science

Project

Melissa Winstanley
University of Washington
January 18, 2024



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

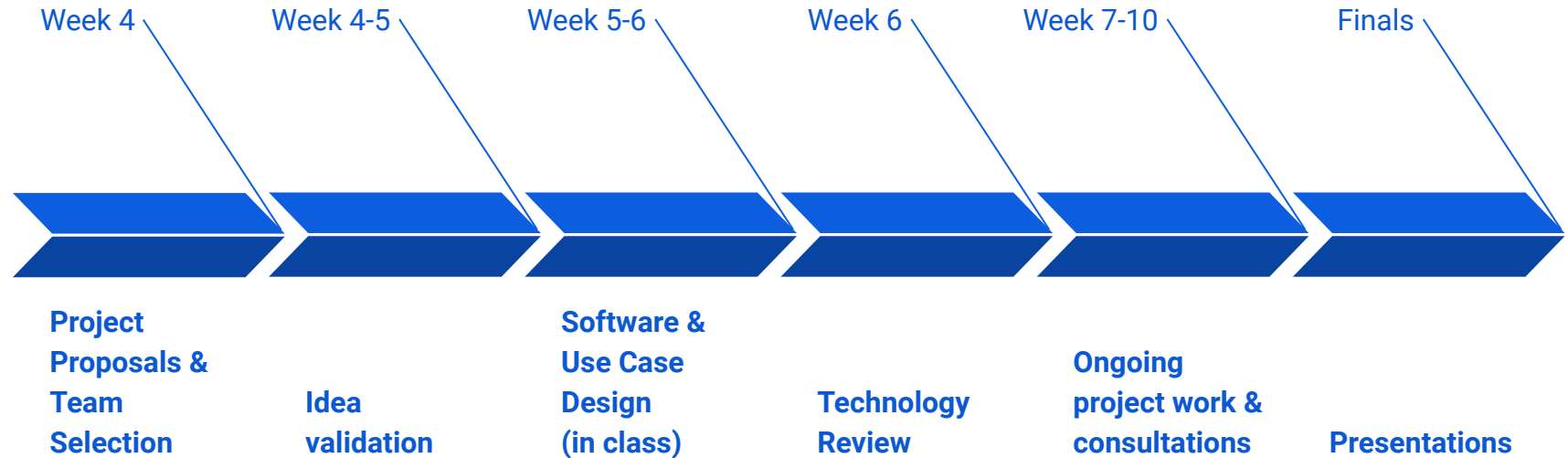
Example

[BioReactor Data Logging](#) – Monitor and publish data from BioReactor experiments

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!

Project Timeline



Week 4: Project Proposals

Share what you're passionate about and convince others to work with you:

- 1 slide (PowerPoint, Keynote, or Google Slides)
- Title
- Project type (research, reusable data, tool)
- Short pitch - what will you do, why it is cool
- Your 2+ data sources (could be tentative)
- Your name

Due by 12pm next Thursday via Canvas (I need time to compile the deck!)

In-class presentation of your proposal (2 minutes)

Participation points! If you are unable to attend, please let me know.

Week 4: Team Selection

During project proposals:

- Take notes on what projects sound interesting to you

After project proposals:

- You'll have time in class to talk to each other and form teams around a project proposal
- 3-4 people per team - 1 person submits names via Canvas
- If you'd prefer, I'm happy to help find a group for you, particularly if your team has just 1-2 people
- If you have to miss class
 - Let me know
 - I will help you find a project team

Week 4-5: Idea Validation

- Agree as a team on what the project is
 - Clarity about the project type
 - Consensus on the problem being solved
- Validate that the project is feasible and large enough
 - Is there an unmet need (i.e. no code already exists)?
 - Do you have data that can solve the problem?
 - Will this project take about 5-6 weeks of effort for 3-4 people to complete?
- Create a git repository for your team
 - README.md with result of the idea validation
 - One person submits the repository link via Canvas

Week 5-6: Software & Use Case Design (in class)

In-class exercise to design your project:

- Who are the users? What do they know?
- What information do users want from the system?
- Use cases - how users interact with the system

After class: complete and submit the design in your GitHub repository by 2/15

Week 6: Technology Reviews

In-class presentation addressing a choice of library.

Stay tuned!

Week 7-10: Project Work

- Weekly standups in class
- Collaboration outside of class - code reviews & pull requests
- Deliver on the milestones you've defined
 - Functionality
 - Documentation
 - Style
 - Testing

Finals: Presentations

- 8 minute oral presentations
 - All group members should present a part
 - Background
 - Data
 - Use cases
 - Component design
 - Demo
 - Lessons learned
- Slides + demo

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 \(World bank\)](#)

Project Proposal Examples

Just to get your mind working



Weather Wizard for Farmers

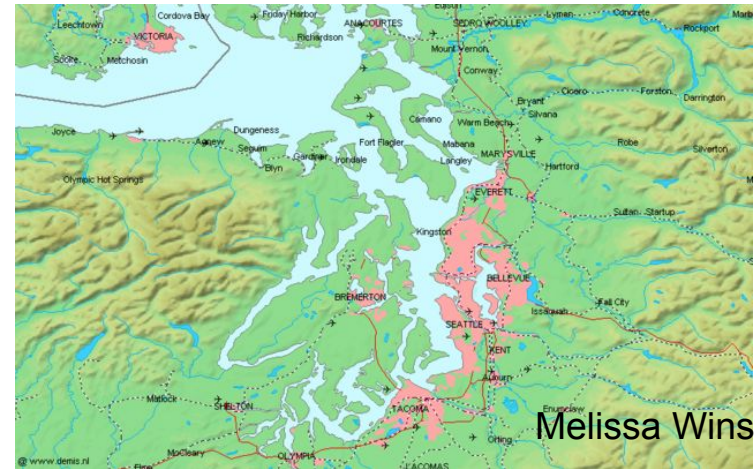
- Project output: reusable data (a simple web app)
 - Give it your location (zip code?)
 - Use a machine learning model to predict future temperature and precipitation for your zip code
 - It will predict your last frost and whether it will be a warmer/cooler or wetter/dryer year
 - It will predict which crops you (the farmer) should plant to make the most money
- Data
 - NOAA data
 - Historical weather (temperature/precipitation) by location
 - El Niño/Southern Oscillation history
 - Atlantic/Pacific Multidecadal Oscillation history
 - Solar Cycle history
 - USDA data
 - State agricultural output data by crop



Melissa Winstanley

Tide Me Over!

- Project output: research & a tool (to make map visualizations)
 - Research how tides affect water changes in Puget Sound via mooring buoy
 - Depth, oxygen, chlorophyll, and salinity
 - Build map visualizations with a time slider to show the changes
 - Include a search feature to narrow in on a particular location
- Data
 - NOAA historical tide data
 - King County mooring buoy data



Melissa Winstanley

Fantastical Basketball*



- Project output: reusable data (a simple web app)
 - Give today's best fantasy lineup for the NBA
 - Daily/weekly/whole season
 - Option to select best=points or best=money
 - Show how the tool has performed for the previous week's worth of games, relative to existing fantasy players
- Data
 - NBA historical data from basketball-reference.com (or NBA site)
 - Day-by-day projection data from fantasydata.com (or other fantasy site)

** This is a VERY HARD PROBLEM. The output of the project is not a predictor that does better than existing players, but one that at least is not terrible.*