UNIVERSITY of WASHINGTON

DATA 515A

Software Engineering for Data Scientists
Project Overview

Joseph Hellerstein[1,2], Bernease Herman[1,2]

[1]eScience Institute

[2]Computer Science Engineering

The University of Washington

April 14, 2020

Objectives For Today

- Introduce expectations for the project
    - Sample the landscape of possible projects

Objectives For Next Tuesday

- Initiate team formation – matching process, sometimes we call it project "dating"

Class project overview

- Collaborative software engineering experience
  - Teams of 3 to 4 with 4 being optimal
    - Prefer *teams with diversity*
  - Develop project in Git w/ GitHub
    - Not Google docs or Dropbox

GitHub

Class project overview

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

http://uwseds.github.io/projects.html

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.

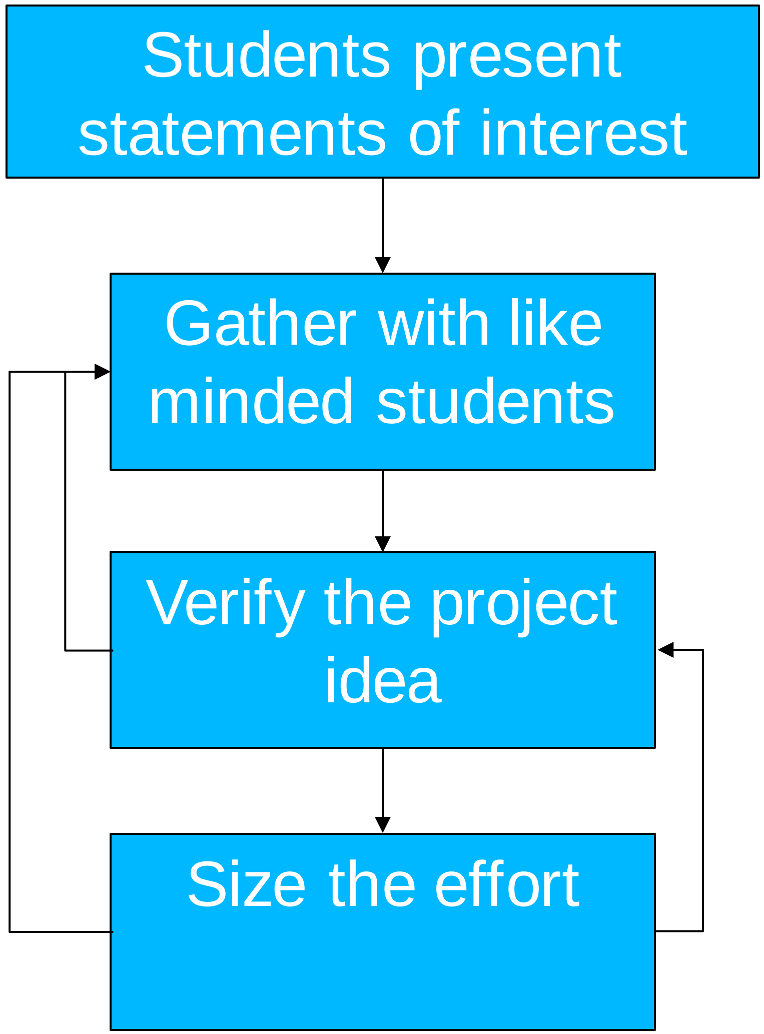
Capstone Project Type 2: *Create Reusable Data*

- 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 re-invent the wheel
- Example
  - BioReactor Data Logging – Monitor and publish data from BioReactor experiments

Project Matching Overview



Students present statements of interest

Gather with like minded students

Verify the project idea

Size the effort

Things to Think About

- Topics of interest
- 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)

Some Third Party Tools

- What third party tools can / might you leverage?
  - Sci-Kit Learn
    - http://scikit-learn.org/stable/
  - Natural Language Toolkit
    - https://www.nltk.org/
  - Bokeh
    - http://bokeh.pydata.org/en/latest/

Grading Rubric

Projects will be evaluated based on the following criteria:

- Project structure
- Quality of the documentation (especially the functional specification and design specification)
- Uses at least two data sources
- Code quality
- Test coverage
- Quality of the example of using the package (in the examples folder of the project repository)
- Implements continuous integration (e.g., via travis-CI), and all tests pass.
- Completeness of the setup.py script
- Creativity and technical challenge

Data! Data! Data!

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

Think about your data NOW!

Project Updates

What is your data?

You should have 2 datasets in hand!

Who are your users?

General public? Scientists? Analysts?

What questions are users trying to answer?

What are the use cases (user-system interactions) to answer their questions?

What issues are there ("known unknowns") with building your system?

Project Ideation

- In class exercise to get ideas flowing (10mins)
  - 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?)
  - At the end of this time, we'll share a few ideas and Joe & Bernease can give feedback.

Projects

Hellerstein & Herman, 2020