Software Engineering for Data Scientists

Capstone Project

David Beck[1,2], Joseph Hellerstein[1,3], Jake VanderPlas[1,4]

[1]eScience Institute

[2]Chemical Engineering

[3]Computer Science Engineering

[4]Astronomy

University of Washington

March 13, 2017

Why a Capstone Project?

- Apply the technical knowledge you acquire
  - Programming, engineering
- Gain experience with collaboration and collaboration tools

Capstone Overview

- Collaborative software engineering experience
  - Teams of 2 to 3
  - Develop project in Git w/ GitHub

Scope

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

Data! Data! Data!

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

Think about your data NOW!

Some Public Data

- Pronto bike data
- American Fact Finder Data
- European union data (World bank)
- Russian federation data (World bank)
- China data (World bank)

Capstone Project Type 1: *Answer Questions*

- Problem statement: Answer two to three questions of business or scientific relevance
  - Use a Jupyter notebook and supporting python files
- Example – Pronto Data
  - What is the effort for bicycle redistribution based on bike in-flow vs. out-flow?
  - How does weather affect in-flow and out-flow?
  - What is a good plan for bike redistribution (minimizing trips and trucks)?

Capstone Project Type 2: *Create Reusable Data*

- Problem statement: Create data repository with tools (e.g., search, visualization, analytics)
- Examples
  - FDA recall data
  - Housing price and crime data

Workflow

1. Form a team!
2. Verify the project idea by iterating on
   - Problem statement, data
3. Catalog the data (size, dimensionality)
4. Sketch out a design
   - Include 3rd party pieces
5. Prepare a 3 to 5 minute presentation
   - Describe steps 2 - 3