Syllabus: INDS 4997
Capstone in Data Science

Objectives:

- Familiarize students with the complete Data Science process from Data Gathering to Data Analysis
- Let students participate in a modern, agile project management process
- Allow students to develop a github portfolio for job hunting

Process:

We will study data on police activity at MPD. The ultimate goal is to identify issues in policing.

We will use Scrum as a methodology. Scrum has been successfully ported from Software Development to Data Science projects. Because you are not working full-time on the project, we will need to adapt Scrum.

You will meet shortly three times per week in a "stand-up" meeting where you will explain what you did for the project and what you are going to do until the next stand-up meeting. The project is divided into sprints of three weeks duration. At the end of each sprint, the project needs to be in a state that is useful. For the first sprint, at a minimum you have a running system that automatically gathers the data through web-scraping.

You will self-divide into two teams of up to ten persons.

Roles:

In Scrum there are three different roles.

Product Owner: Has a and communicates this vision about the final product. In a Data Science project, the product owner will also lead team discussions that will determine the questions and analysis performed in the later stages of the project. In our setting, the product owner is also responsible for a final testing and code review. The product owner is not part of the development team proper.

Scrum Master: A member of the development team who specializes in the scrum process. (This means if you are selected as scrum master, you will have to buy a couple of books on the Scrum process.) The scrum master convenes the team meetings, but is otherwise a developer as every-one else.

Developer: Every-one but the product owner is a member of the development team. A team manages itself through the frequent stand-up meetings. It evaluates itself during the sprint reviews.

GitHub Expert: All contribution need to be maintained on GitHub. The GitHub expert helps team members to organize their contributions.

Data Curator: Takes responsibility for data gathered and processed into a data-base.

Grading:

Product owners and scrum masters are graded by the instructor. Developers (including the scrum master) will be graded based on feedback. All developers will be given a budget of $4n$ points per rubric where n is the number of developers in the team. They anonymously assign points between 0 and 5 to all other developers. The grade received represents the average of points received: A: > 4 points average, B: > 3 points average, C: >2 points average. Failure otherwise.

The instructor will then multiply all team members points with a multiplier $\alpha, 0.5 \leq \alpha \leq 1.1$ reflecting the success of the project.

Class Activities

There will be mandatory lectures on project management.

Python

An almost complete treatment of Python for Data Science is found at https://tschwarz.mscs.mu.edu/Classes/PDS2021/index.html.