Getting Started

Week 1

DS198-003: Data Discovery Scholars Seminar
UC Berkeley - Computation, Data Science, and Society

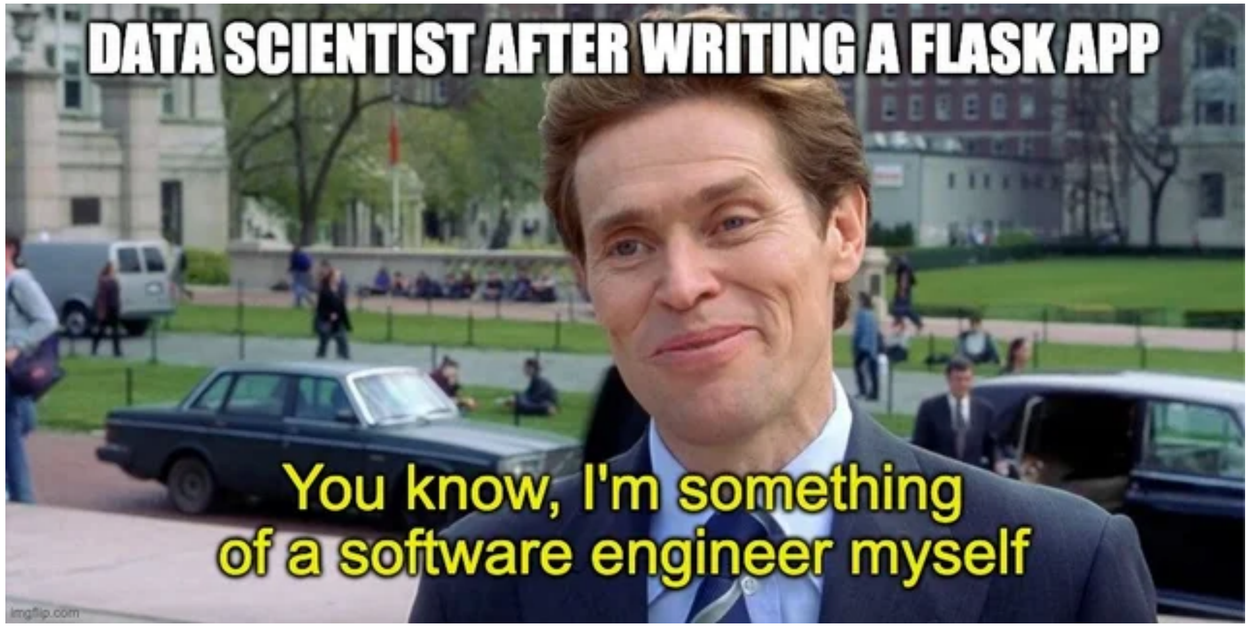
Spring 2022

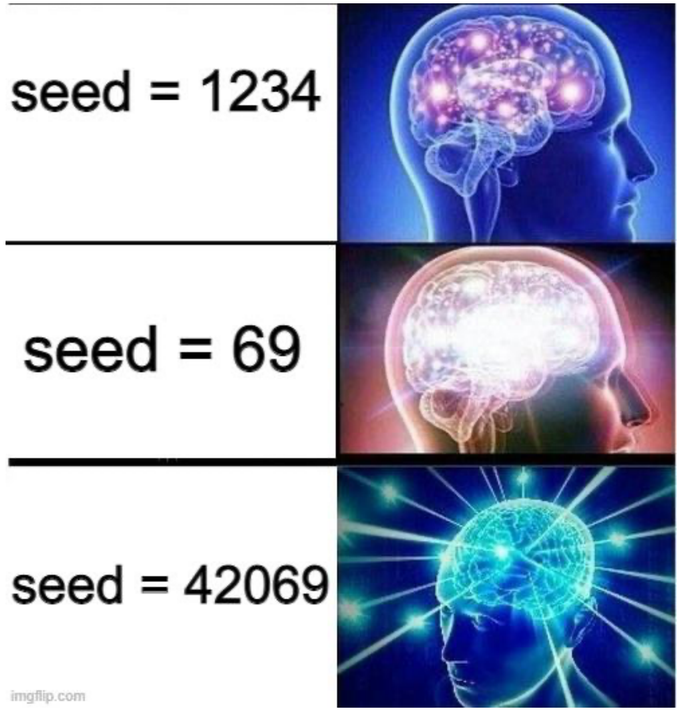
Today

1. Introductions
2. Discovery Scholar's Program
3. Planning & Managing Projects
4. Setting Goals

Meme of the Week

Memes of the Week





Introductions

Hi 👋 I’m Kevin Miao!

Introductions

- **M.S. Student** in **EECS** (mainly CS)
- *Hometown:* Eindhoven, the Netherlands 🇳🇱
- *Research Focus:* **Deep Learning & Explainable Artificial Intelligence (XAI)**
- *Hobbies:* Weight-lifting, Biking, Photography, Cooking, App Development, Loves Memes
- *Previously Taught:* Data 8 & CS61BL
- *Email: kevinmiao@cs.berkeley.edu*
- *Website:* kevin-miao.com

Your Turn! 🙋

Introductions

- Introduce yourself to the class:
  - Name
  - Year/Major(s)
  - Hometown
  - (Anti-ice-breaker) Ice-breaker you dislike
  - What were you doing during break?

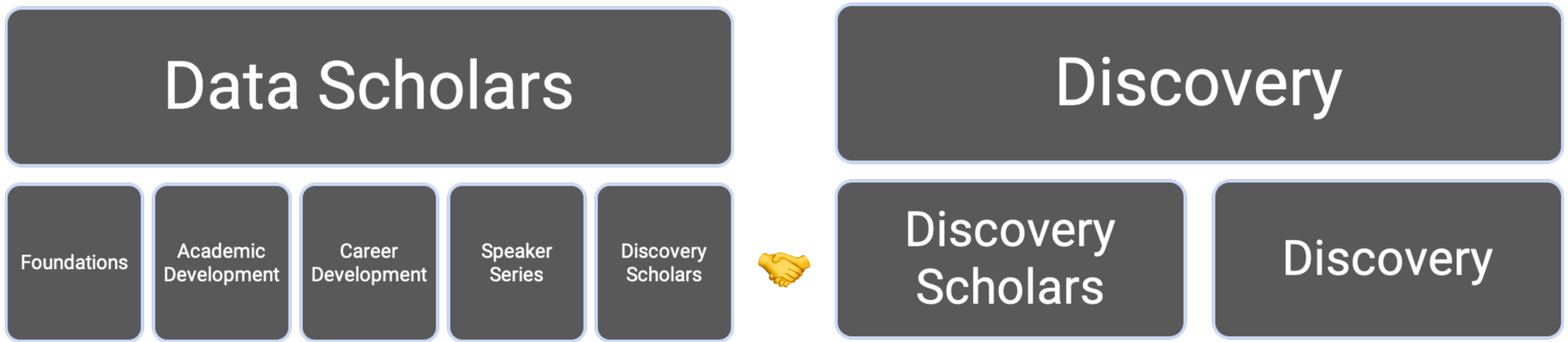
Data Scholar’s & Discovery

Discovery Scholar's Program

Collaboration between Data Scholar's & Data Discovery

- **Data Scholar's** is a program under the **Computing, Data Science, and Society** department that focuses on providing **historically underrepresented students** a community and resources to succeed.

- **Data Discovery** is a campus-wide research program that allows outside partners (companies/researchers/professors) to collaborate with students on projects



Course Overview

Course Overview

Class Structure

- 🏁 Goals of the class:
  - Personal/Career Development in Data Science
  - Project Management Skills
  - Technical Maturity
- Class used to be a 1.5 hour seminar with weekly homework assignments.
  - This iteration we will focus on keeping things small and flexible
  - **Concretely, we will have lots of flexibility and little to no heavy assignments.**

Course Overview

Administrivia: Assignments & Grading

- **Assignments**
  - Reflection Assignments (10%)
  - Final Presentation (10%)
  - Attendance & Participation (80%)
- **You need at least 80% to pass**

Course Overview

Administrivia: Extensions/Late-Policy

- **DSP**
  - Students with special accommodations should have the DSP office send a letter via AIM
  - DSP students will automatically get one day extension on all assignments (except for the presentations)
- **Non-DSP**
  - 2 slip days for all non-presentation assignments
  - Is there anything else going on? Feel free to reach out!

What is Data Science?

Project Management

Project Management

Data Science Lifecycle

- Data science lifecycle has many different forms/alternatives/derivatives but it boils down to the following five steps:
  - **Data Mining**: Obtaining data (SQL/MongoDB, ETL, Data Warehouse, Data Lake, Apache, Spark)
  - **Data Wrangling**: Detecting anomalies, making data human readable (Pandas/Hadoop)
  - **EDA:** Finding patterns/trends, visualize plots, feature engineering (Numpy, Matplotlib)
  - **Modeling**: Magic (we know it's math) Causal Inference, Supervised/Unsupervised Learning (Scikit-Learn, TensorFlow/Keras, XGBoost, R/CARET)
  - **Interpretation & Evaluation:** How well did my model perform?

Project Management

Data Science Lifecycle: What can go wrong?

- **Discussion Time**
  - Data Mining
  - Data Wrangling
  - EDA
  - Modeling
  - Interpretation & Evaluation

Project Management

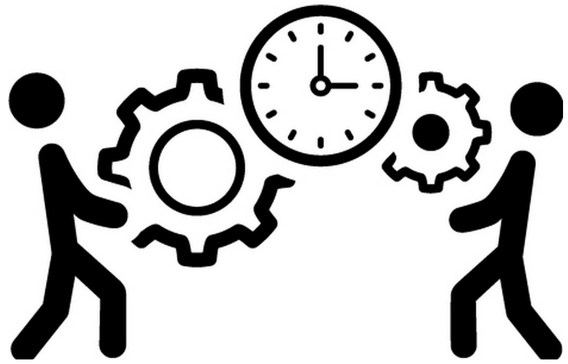
More Data Science Lifecycle

- The canonical lifecycle is **not** completely correct
- **Data Science** can be more diverse:
  - **Strategical Planning:** Thinking about project objectives, impact, problems you want to solve
  - **Constructing Data Pipeline:** Creating a whole architecture to accommodate data flow (Backend languages, Cloud Providers (GCP, AWS, Azure), Snowflake/SQL)
  - **Data Supervision**: Adding labels/annotations to your data
  - **Structuring Data**: Using Unsupervised Algorithms to make sense of the data
  - **Business Intelligence:** Providing data-driven recommendations based on data visualizations/findings
  - **Software Engineering:** Creating applications and monitoring the performance of these models

Project Management

Management Strategies

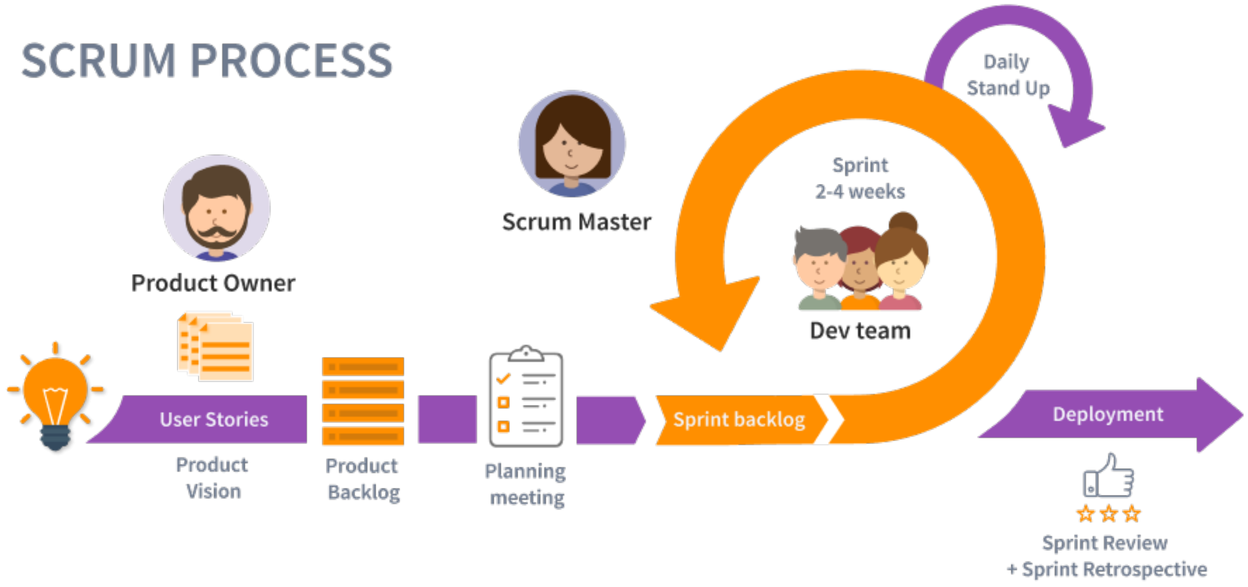
- Industry based management strategies to align teams
  - CMM
  - Waterfall
  - Agile/SCRUM
  - Hybrid
  - R&D



Project Management

Agile/SCRUM

- Agile is a widely adopted **software management** framework built on the foundations of **continuous integration** and **deployment**
  - **SCRUM** is a way of working in a team within the AGILE framework
    - **Process:** Divide project in multiple smaller pieces which are called **sprints**
  - **Tools:**
    - monday.com
    - notion.so



Project Management

Data Management

- As part of managing your project, you also need to manage your data.
    - How do I obtain data? Does it live in the cloud or not? Is it sensitive?
    - Create a guide
        - Who labeled the data?
        - Where did it come from?
        - What do all the columns mean?
        - When were the data collected?
    - What kind of data do we have structured (table) or unstructured (photos, text, video)?

Project Management

Tips

- Start by thinking about the problem statement
- Divide your project up in small chunks (mostly could be DS lifecycle stages) and tie them to a specific time range
- Be realistic about how much something takes
- Implement buffers
- Think about what resources you will need/stakeholders you will interact with for each phase
- **KEEP IT SIMPLE!**

K.I.S.S

Keep It Simple Stupid

“Great advice. Hurts my feelings every time.” - Dwight

Project Management

Optional Readings

- [Business Planning] https://towardsdatascience.com/how-data-driven-businesses-approach-scenario-planning-7ba3e9a89e79
- [DS Lifecycle] https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc
- [Project Management Strategies] https://neptune.ai/blog/data-science-project-management-in-2021-the-new-guide-for-ml-teams

Communication

Communication

- Communication is key; majority of issues revolve around communication
- Set expectations early-on in the semester
  - Who is responsible for what?
  - How do we communicate?
  - What is expected in terms of response time?
  - What are everyone’s commitments?
- Schedule routine check-in times
- Ask questions!!!!
- Document your code and keep it clean

Communication

Coding Etiquette

- Besides interpersonal communication, you have to make sure that your code is easy to understand too.
    - Understandable variable/function names
    - Write a `README.md`
        - Explain what each file does
    - Create a `requirements.txt` file
    - Perform/Ask for Code Reviews
    - Use git

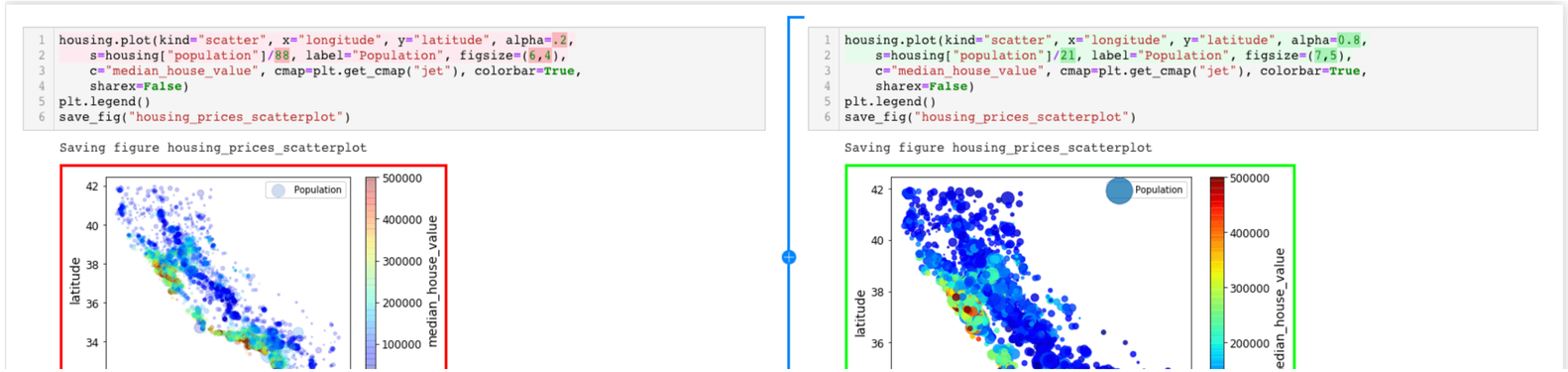
Communication

Code Reviews

- GitHub is amazing for code reviews!
- Create a `development` branch where you experiment
  - git branch development
  - git checkout -b development

Reviewing Code

- Use **reviewNB** where you can review someone else's code before you merge it onto the **master** using **pull requests**
- Things to focus on:
  - Runtime efficiency
  - Logic mistakes
  - Typos
  - Wrong variable names



Communication

What would you do?

Your teammate Alice has not been responding to your messages lately, when she does it’s super late. It seems like it is midterm season and you have to do the majority of the work.

Communication

What would you do?

Your project manager has put you in charge of performing a really niche analysis. However, it is extremely difficult. You have asked them for a lot of help already, but still don’t seem to get it. What would you do?

Communication

Optional Readings

- GitHub101
  - https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
- Communication
  - https://www.analyttica.com/the-art-of-communication-in-data-science-through-the-lens-of-experience/
  - https://towardsdatascience.com/communication-can-make-or-break-a-data-science-project-75ce3952de89
  - https://hbr.org/2014/06/how-to-give-your-team-feedback
- Code Reviews
  - https://www.reviewnb.com

Assignments

➡️ Actionable Items (due next class)

- **Submit Welcome Survey: https://tinyurl.com/3869awpj**
- **Weekly Assignment 1**
  - Your short reflection (max 1 page) needs to address the following questions:
    - What do you want from this seminar/Kevin?
    - Are you here to learn certain topics? How to conduct/discuss research? Do you want personal guidance? More or less focus on discussions?
    - What are your goals for yourself.