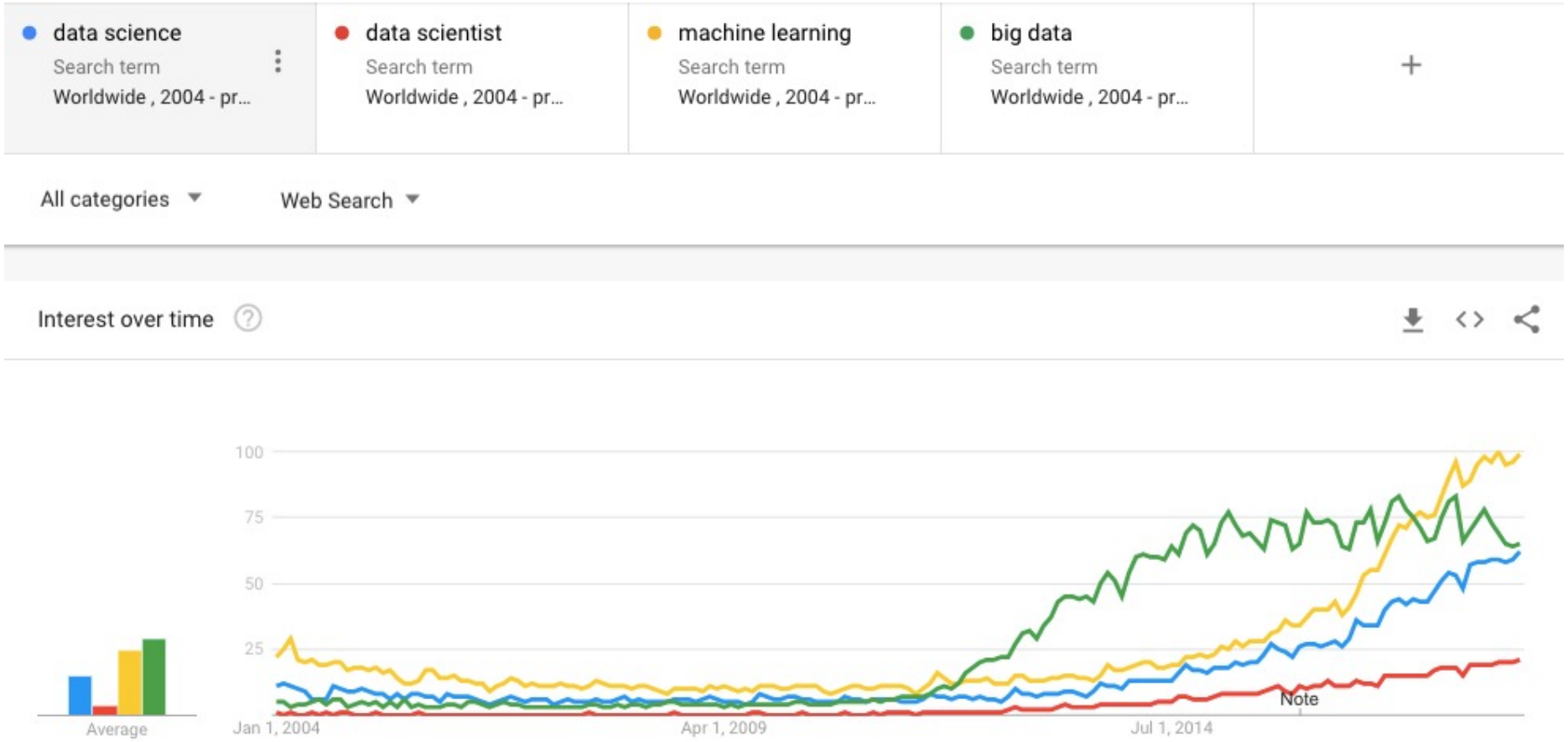
Data 601 @ UMBC

Welcome to Data 601



- Course Logistics
- What is Data Science?
- Software tools
- What are we not covering?
- Soft skills
- Homework

Syllabus and Course Logistics

- **Syllabus**: https://github.com/msaricaumbc/DS601_Fall22
- **Office Hours**: https://calendly.com/msarica1/15min
- **Email:** mehmet.sarica@umbc.edu
- **Schedule:** https://github.com/msaricaumbc/DS601_Fall22#weekly-schedule

Components of a good class

- How do you promote a safe classroom environment?
- How do you address the variety of backgrounds of participants?
- How do you encourage participation? Enable rest?
- What behavior fosters growth?

Activity: Think, then write

Ground rules

- Schedule: 7:10 – 9:40 (we may have breaks)
- I value being punctual (start of class, breaks, end of class)

- Don't apologize for asking a question or for not knowing something
- I find it acceptable for you to occasionally not participate

- Tell me if you cannot hear me or if you cannot understand me
- Slides/notes will be provided after lecture (Github)

- I value your feedback:
  - Direct: verbal
  - Indirect: anonymous question/comment sheets on your desk

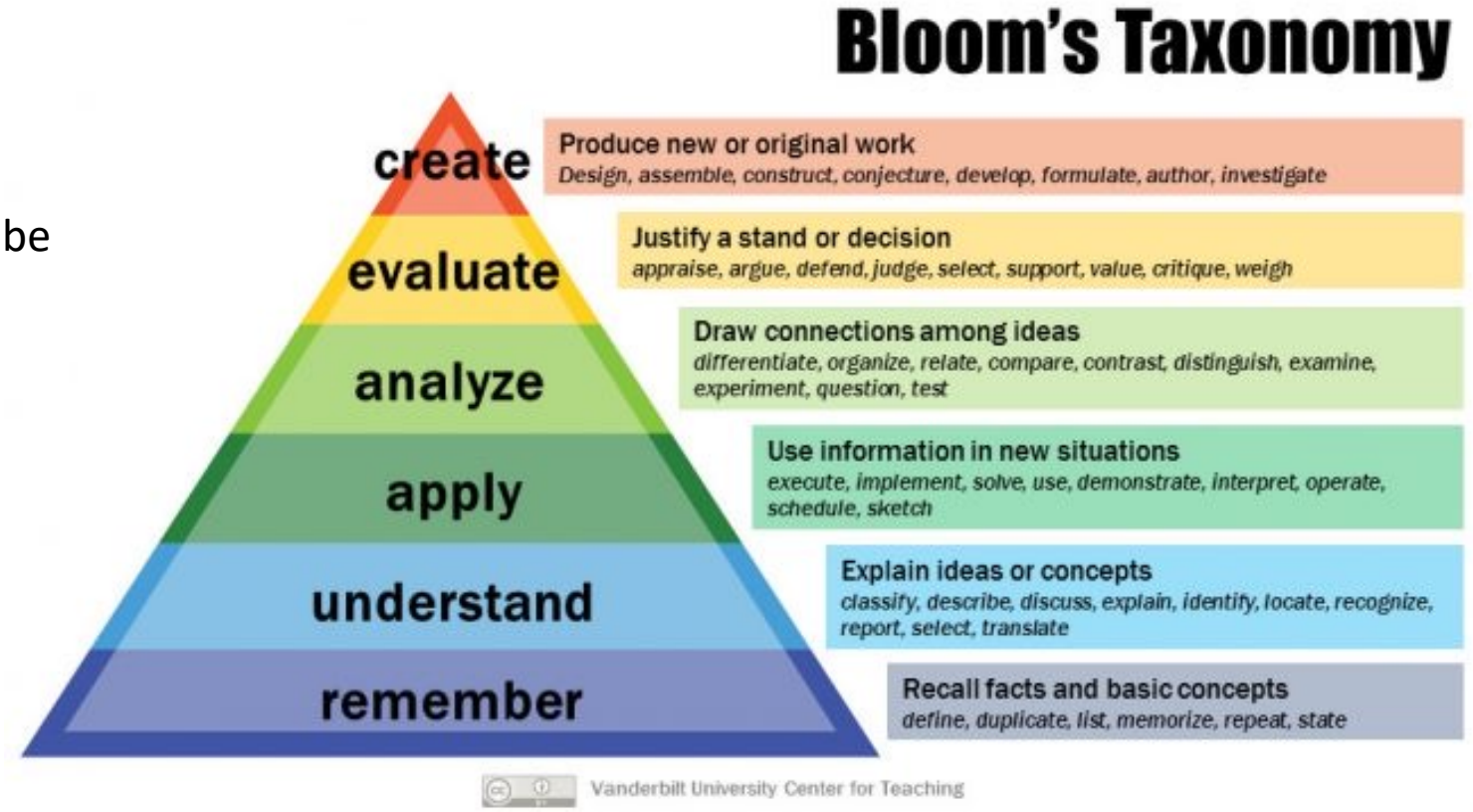
Grading

What you may care about for evaluation

- Attendance: 10%
  - Show up for class (*not always enough)
  - Participate in class exercises and surveys
- Homework: 40%
  - When homework is assigned, the due date will be provided
- Midterm Project: %20
- Final Project: 30%

Learning

What I care about conveying

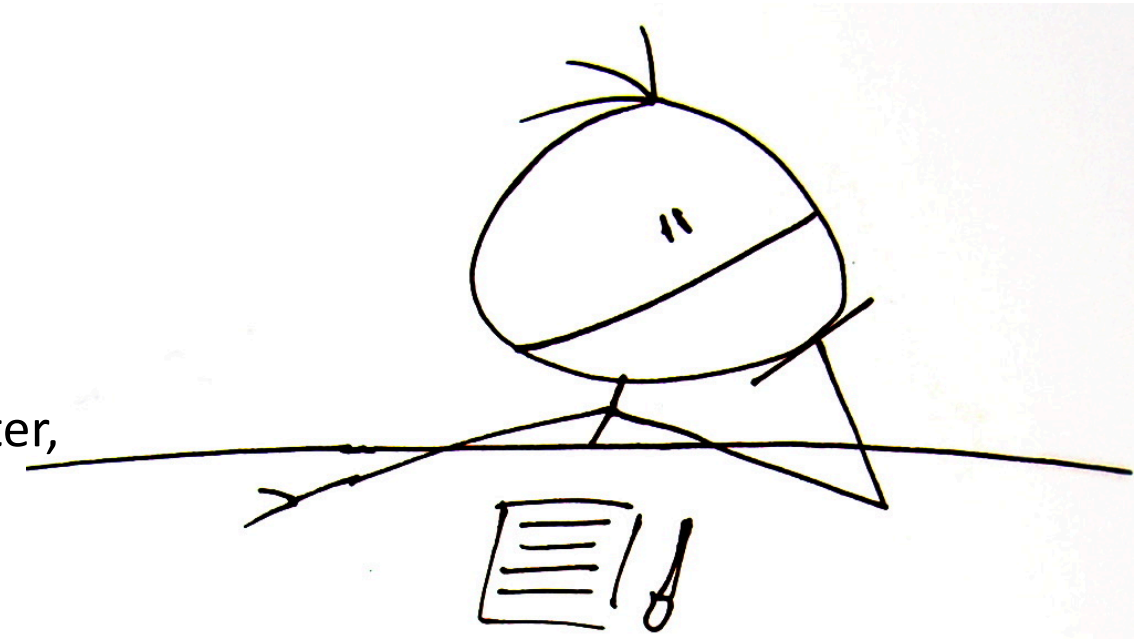


Log your assumptions and expectations

In-class exercise:

- Open a notepad(or a word document) on your computer, record the date.

- Write down your assumptions about this class
- Write down your expectations for this class



Activity: think and write

Store assumptions and expectations

We will revisit these notes later in the semester

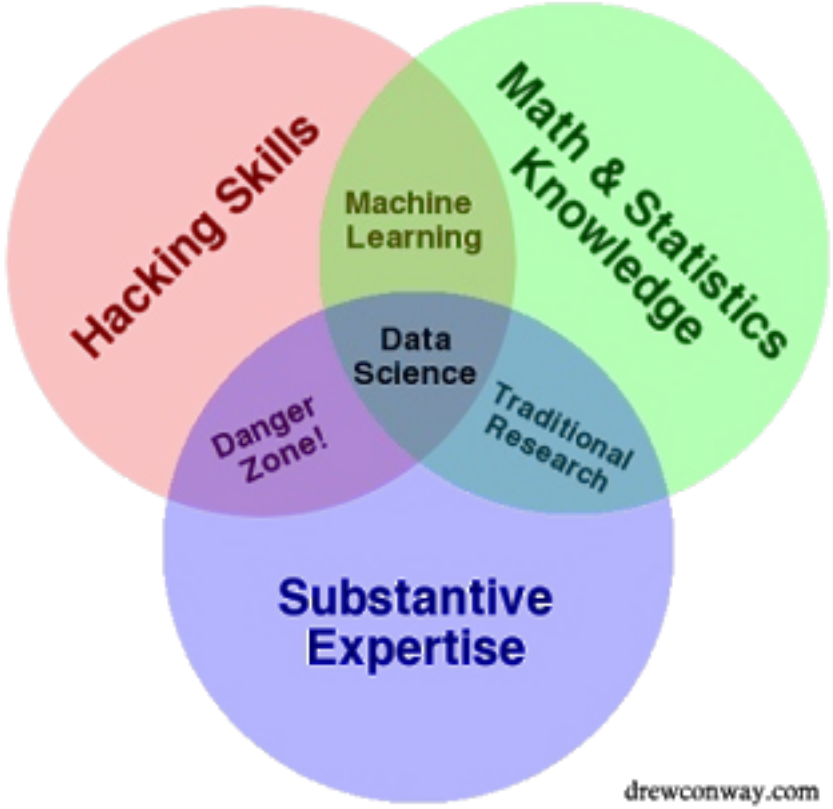
Store your note where it can be accessed later in the course

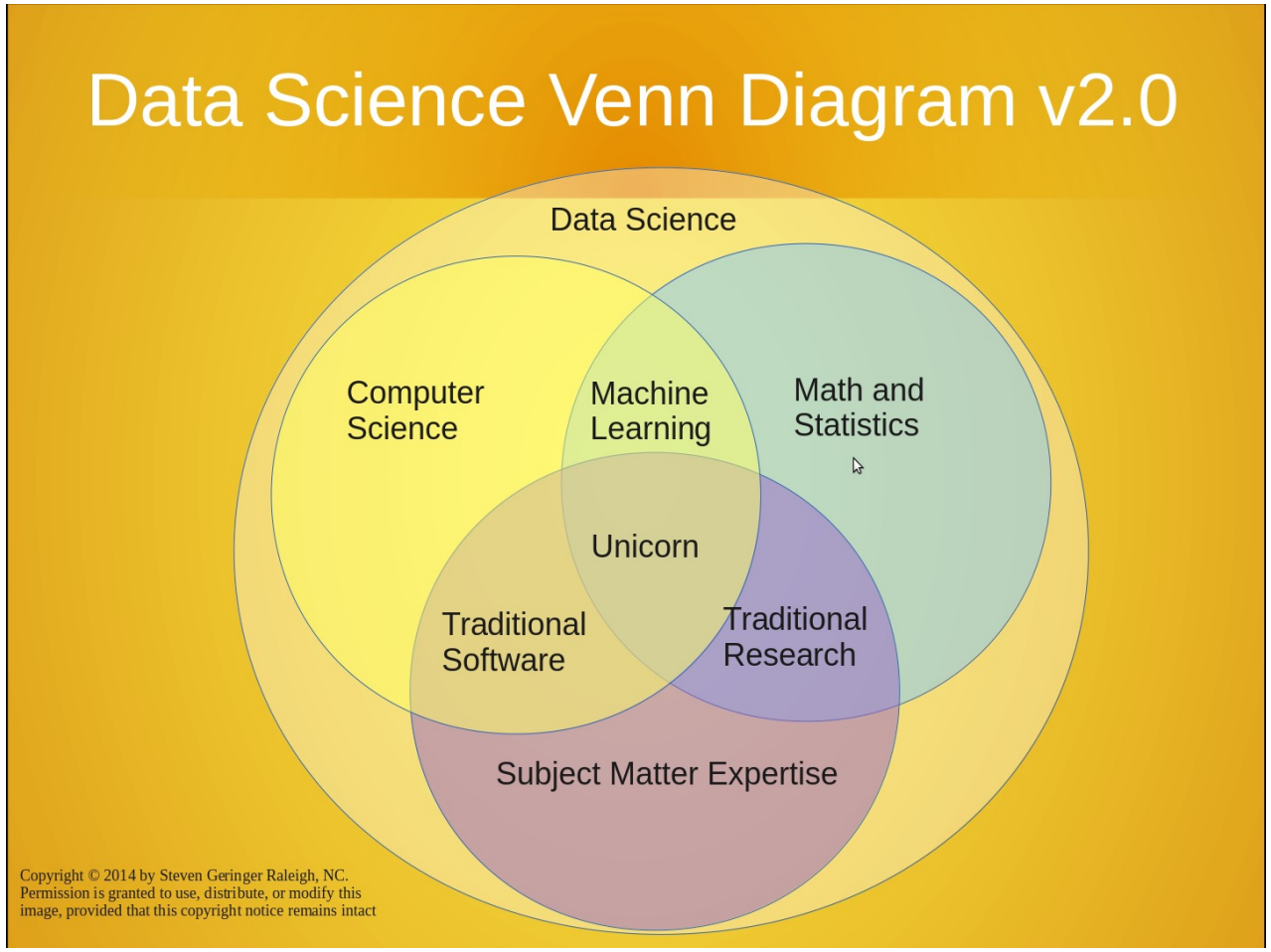
What do you want to learn in this class?

Activity: verbal popcorn; record answers on board

- ~~Course Logistics~~
- What is Data Science?
- Software tools
- What are we not covering?
- Soft skills
- Homework

There's a lot to cover

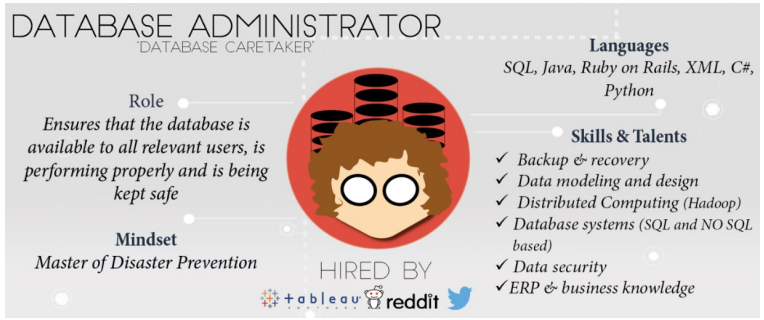


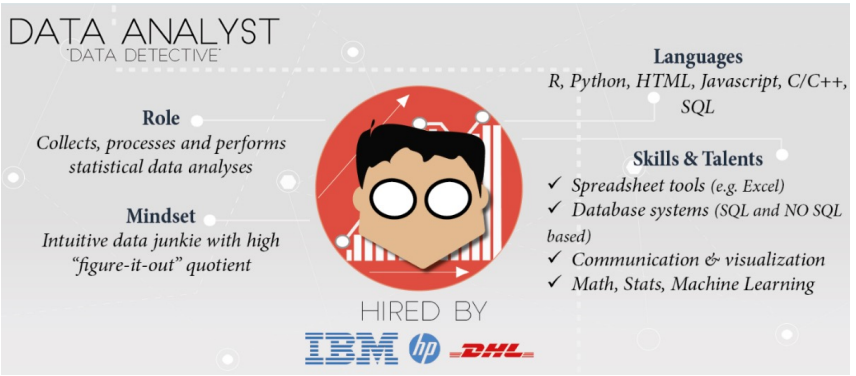


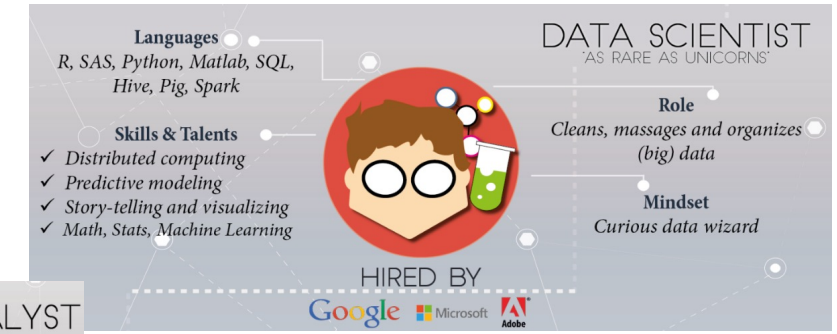
Suggested reading:

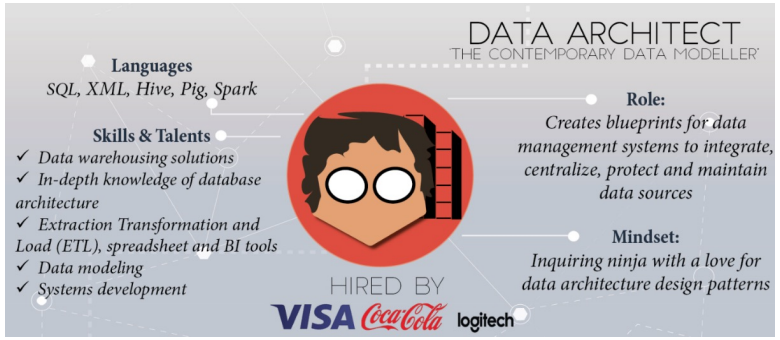
https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

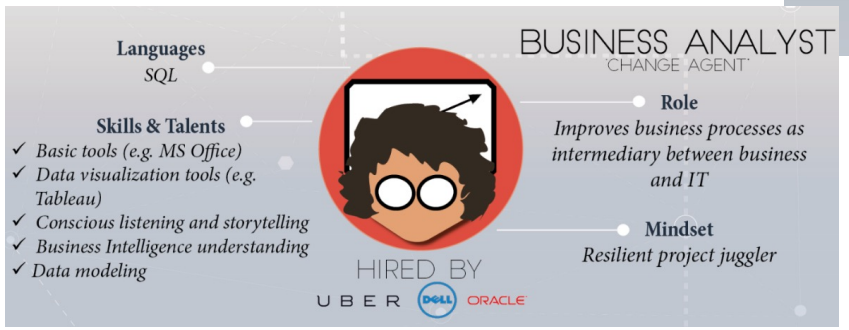
Skills and experience matter more than title and labels

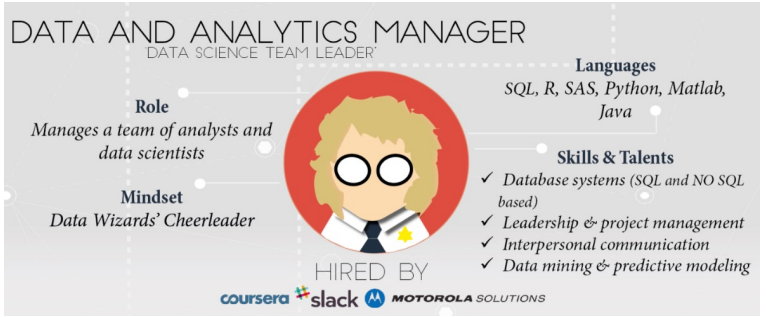












https://www.datacamp.com/community/tutorials/data-science-industry-infographic

Historical progression: data grooming, data mining, data scientist

Data science is an active field with lots of jargon

There will always be something you haven't heard of before.

- Know enough to be conversant with peers
- Be curious about new topics
- Research concepts and labels before using them

Reference: http://www.datascienceglossary.org/

Why learn data science?

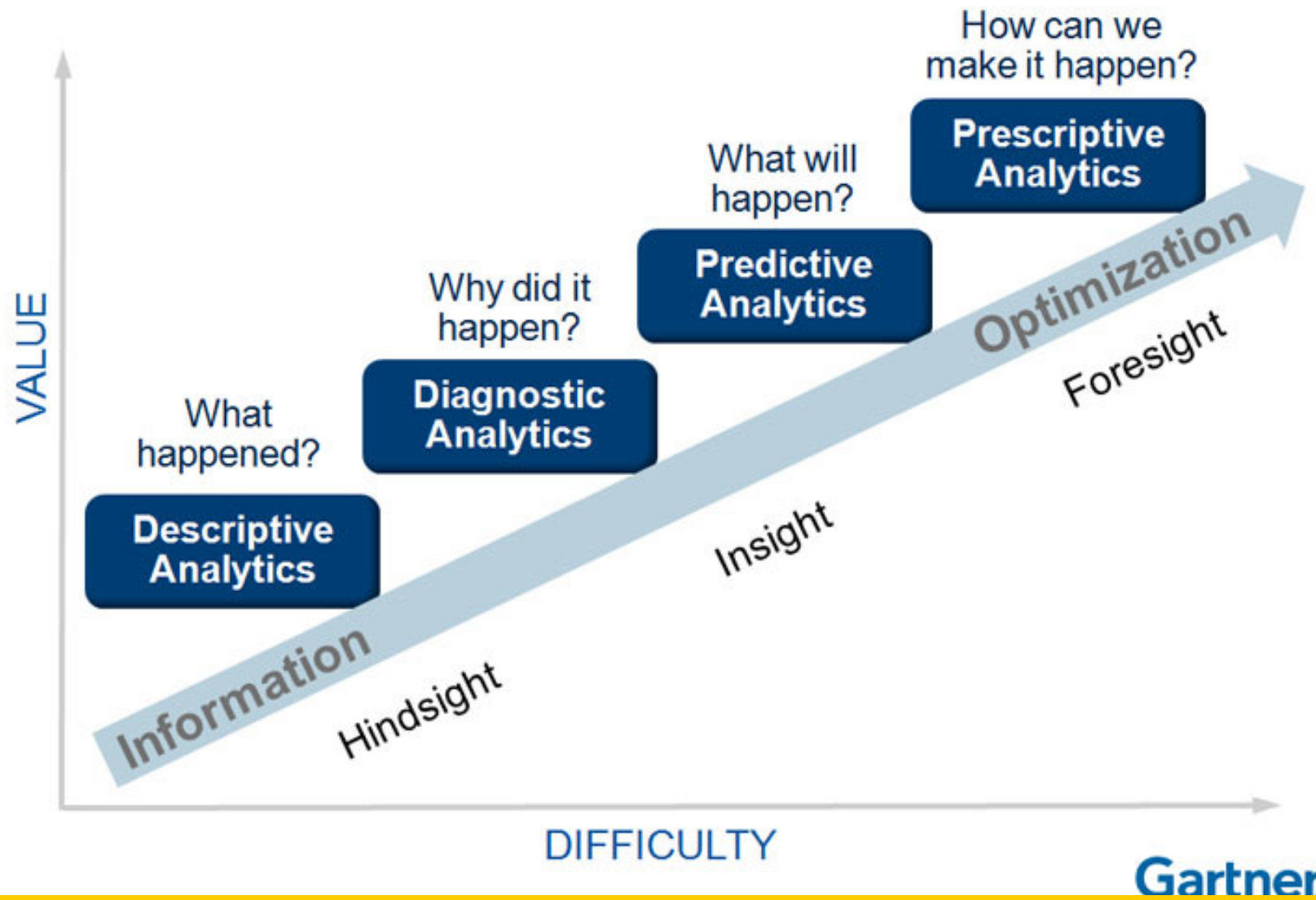
Explore: **identify patterns**

Predict: **make informed guesses**

Infer: **quantify what you know**

Motives:

- Make money
  - Employment
  - Promotion
- Help people
- Gain new knowledge



Large scale use cases with lots of data

- Google's search engine
- Recommendations from Amazon and Netflix
- Bank and Credit Card fraud detection
- Logistics (DHL, UPS) of fleet management
- Healthcare records from patients

Each depends on availability of compute and data

Assumption in this class

- In class we will assume you are a lone data scientist on an island with an internet connection.
- This is not the typical case -- you'll have coworkers, customers, bosses, competitors, collaborators, peers.

Example of how class ≠ real world

- This class will not use competitive grading. (Imagine if it were.)
- As an employee at a company, you may be competing for a bonus or promotion

--> consequence: personal and organizational politics factor into the work environment

- ~~Course Logistics~~
- ~~What is Data Science?~~
- Software tools
- What are we not covering?
- Soft skills
- Homework

Most popular tool in data science

- what do you think the most popular software tool in data science is?

https://www.trippinsights.com/2018/01/04/milling-about/

Activity: interactive survey

Most popular tool in data science

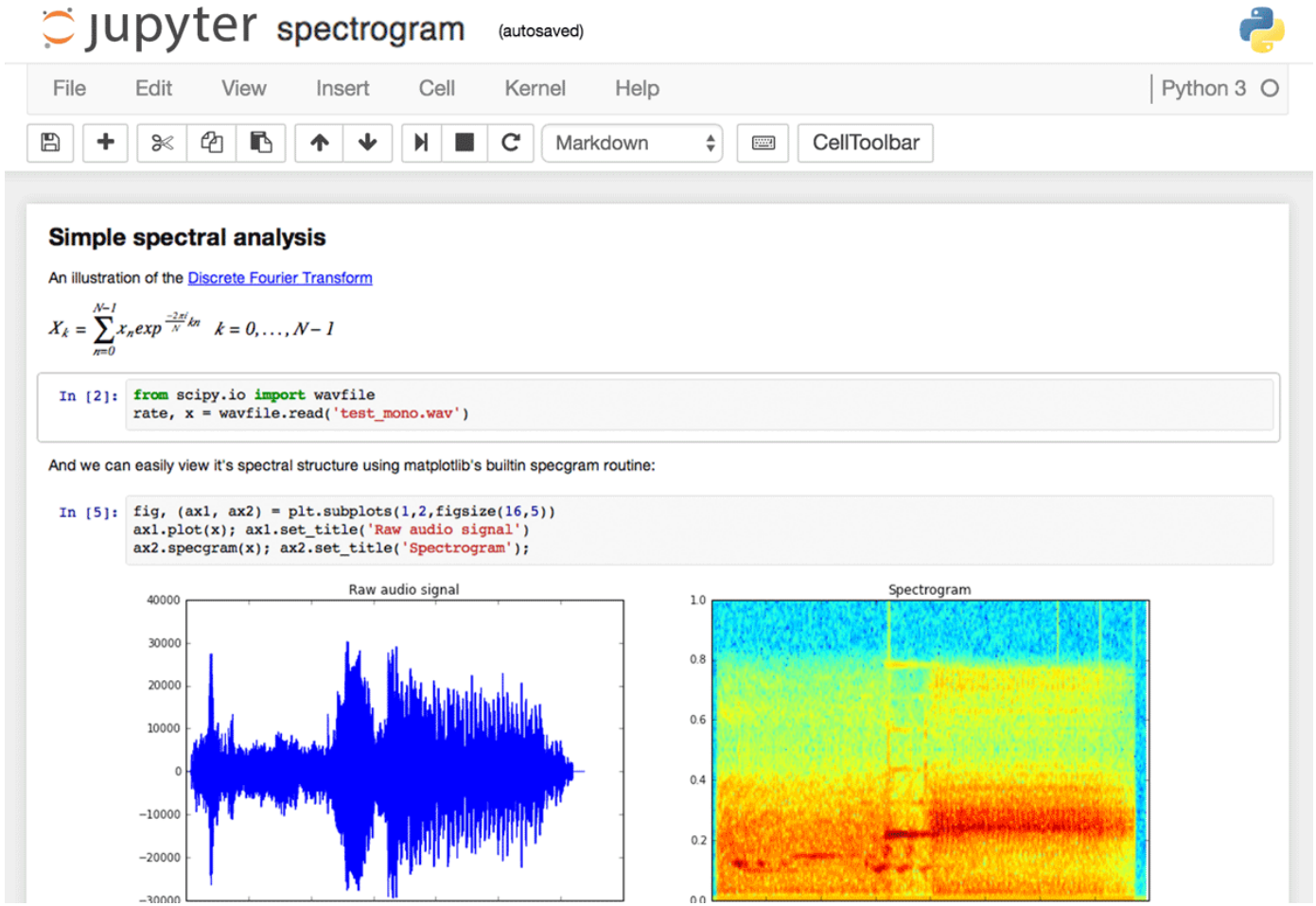
The most popular tool in data science is **Excel**.

- Excel is attractive because it's flexible and capable -- it can store data, the transform, and the visualization (no context switch).
- However, as an analytic ages in Excel and is increasingly tailored, it becomes increasingly brittle.
- Also, in the list of capabilities I didn't include documentation.

Most popular tool in Data 601: *Python*

Interface to Python in Data 601: *Jupyter*

Write in the chat box if you have **used** a Jupyter notebook



Why Jupyter + Python for Data 601?

Jupyter is useful for

- Exploration of data (*jargon*: EDA = exploratory data analysis)
- Documenting your activities (to enable reproducibility)
- Figuring out which software is relevant, which algorithms to use, which software libraries are useful
- Visualizing results

And both Jupyter and Python are free!
And both are widely used!

Python and Jupyter do not cover every use case

- For sufficiently large data sets, Jupyter and Python are not the right tool
- For sufficiently complex analytics, Jupyter and Python are not the right tool

Speed and security are typically not your priority during exploration

Knowing when to invest in switching tools is a skill

Evaluate trade-offs of flexibility and security and speed for a given scale

Relevance of infrastructure to data science

Usual explanation when replicating analysis:

1. Get this data
2. (*Documentation*) Apply this transformation to get result

No explanation of

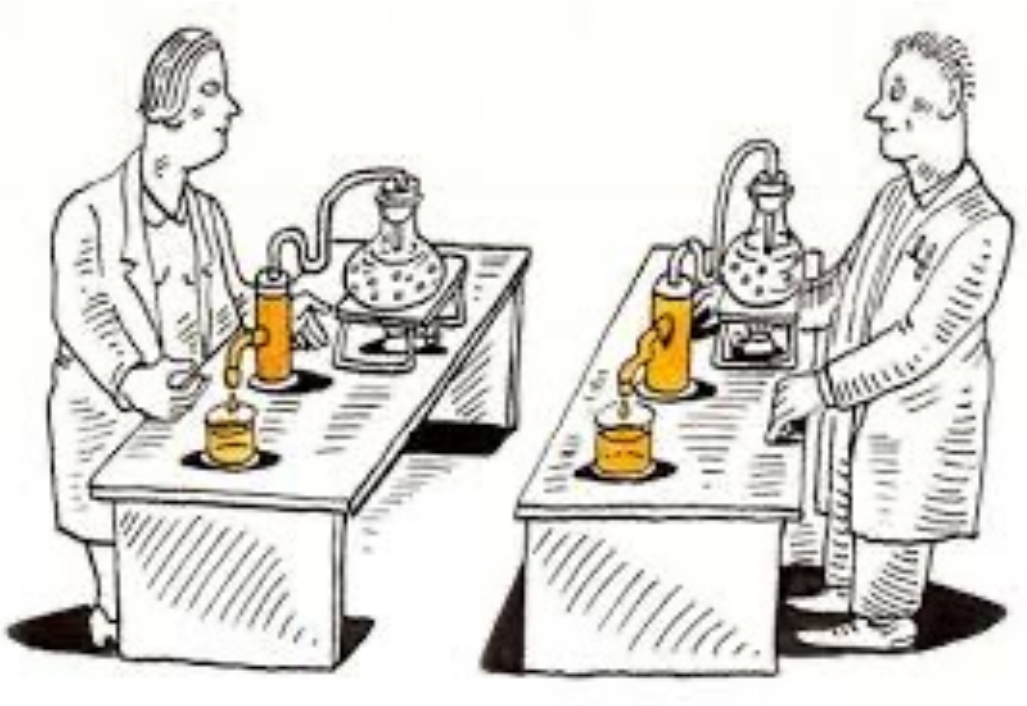
- software used
- software versions
- configurations
- Implementation details

Digital archeology:
Suppose you are to diagnose why someone else's approach doesn't yield same results
Suppose they did their work 20 years ago

Infrastructure in data science to enable Reproducibility and Portability

In addition to data and analysis, implementation and environment matters

1. Use this Operating System
2. Install this software
3. Configure software this way
4. Add these packages
5. Get this data in this format
6. Run analysis against data
7. Create plots
8. Generate report



Best practices: Version control

- Reproducibility applies to your own attempts (not just other people)
- Regardless of how you develop analytics, you'll be creating or editing software and documents.
- [*lesson*] Regardless of how you implement best practices, avoid inventing solutions for which someone else already provided a path.

Suggested resource: https://try.github.io/

Have you used software for version control?

Examples: git, svn, hg

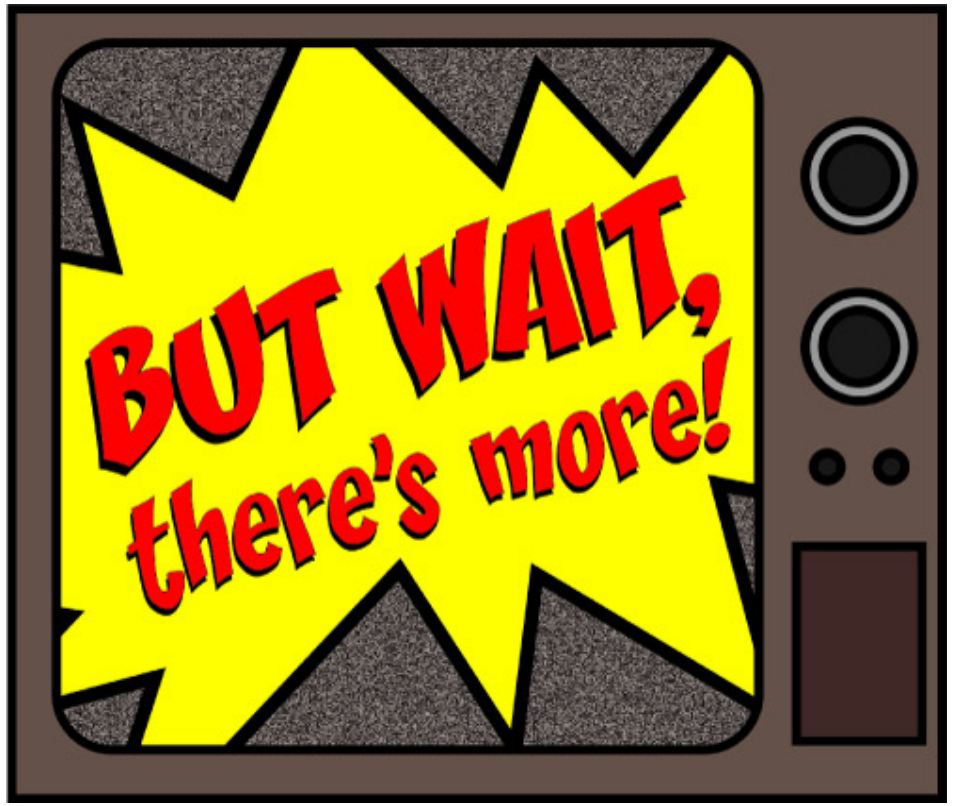
- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- What are we not covering?
- Soft skills
- Homework

Processes external to data science

Exploration may be enough and the effort terminates

Processes external to data science

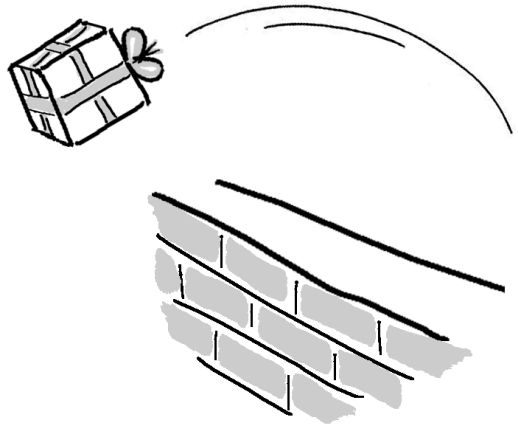
~~Exploration may be enough and the effort terminates~~



Additional refinement is often needed; data science is often merely the start of an investment

Not covered: machine learning

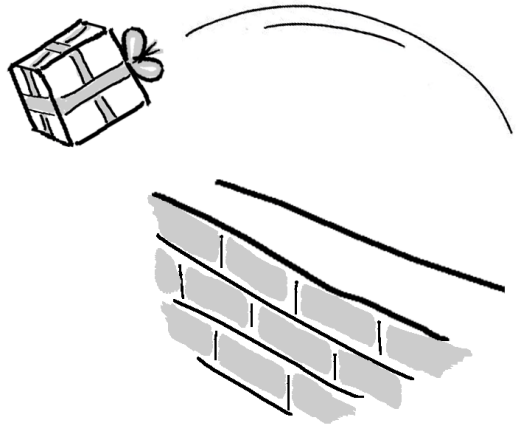
- 602 Introduction to Machine Learning covers Machine Learning
  - If you want to try out, we can do it in one of the last sessions



See http://dev2ops.org/2010/02/what-is-devops/

Not covered: product integration

- There's a complex network of dependencies (i.e. software engineers, managers) of which data science is one component.
- Downstream consumers of your output are likely to be software developers who use containers and support users.
- This class is focused on the data science; not with integration.



See http://dev2ops.org/2010/02/what-is-devops/

Not covered: security

We focus on data science techniques; these do not emphasize secure design of software.

- Let’s Look at some Python code

Figure 1

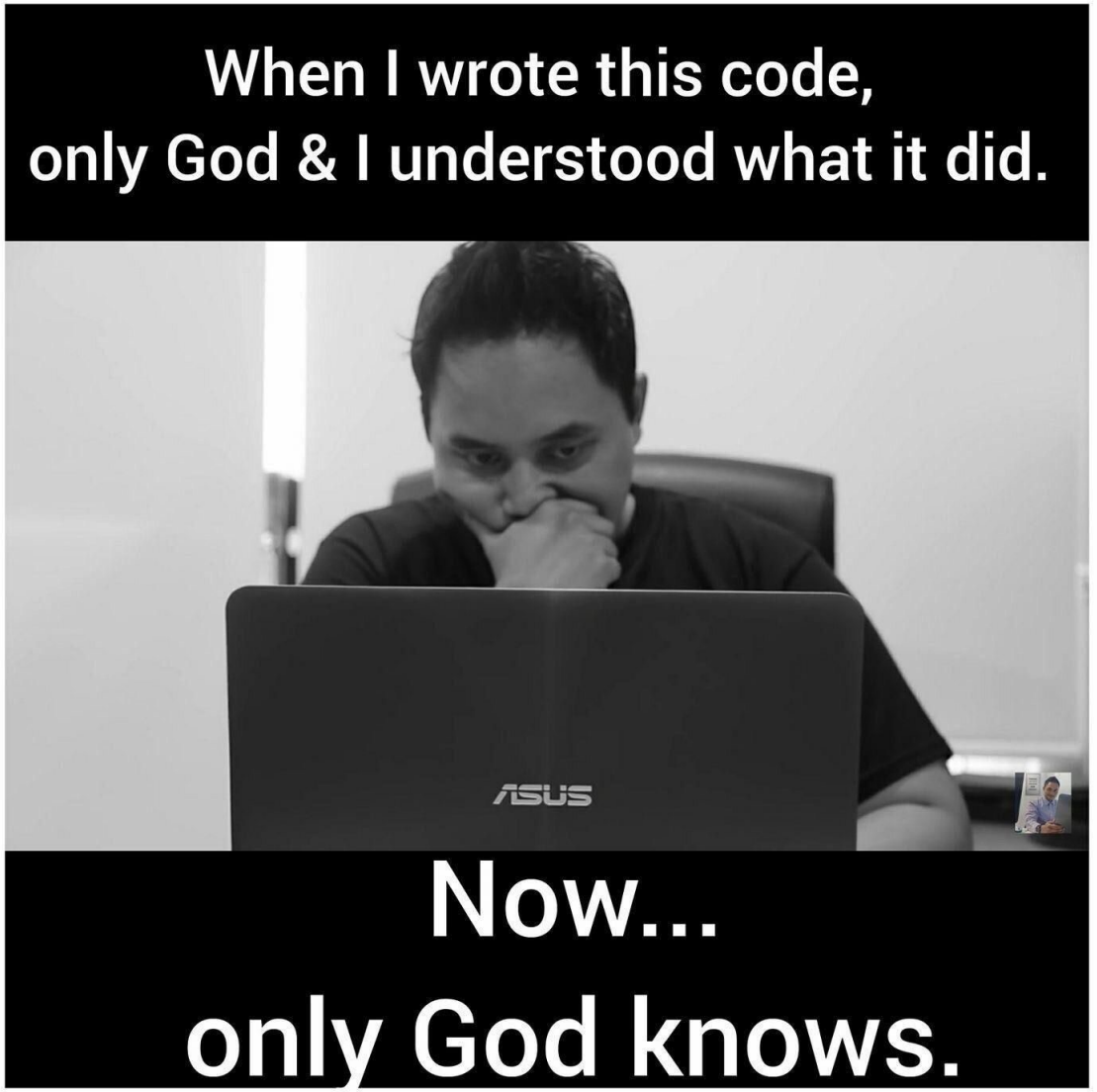
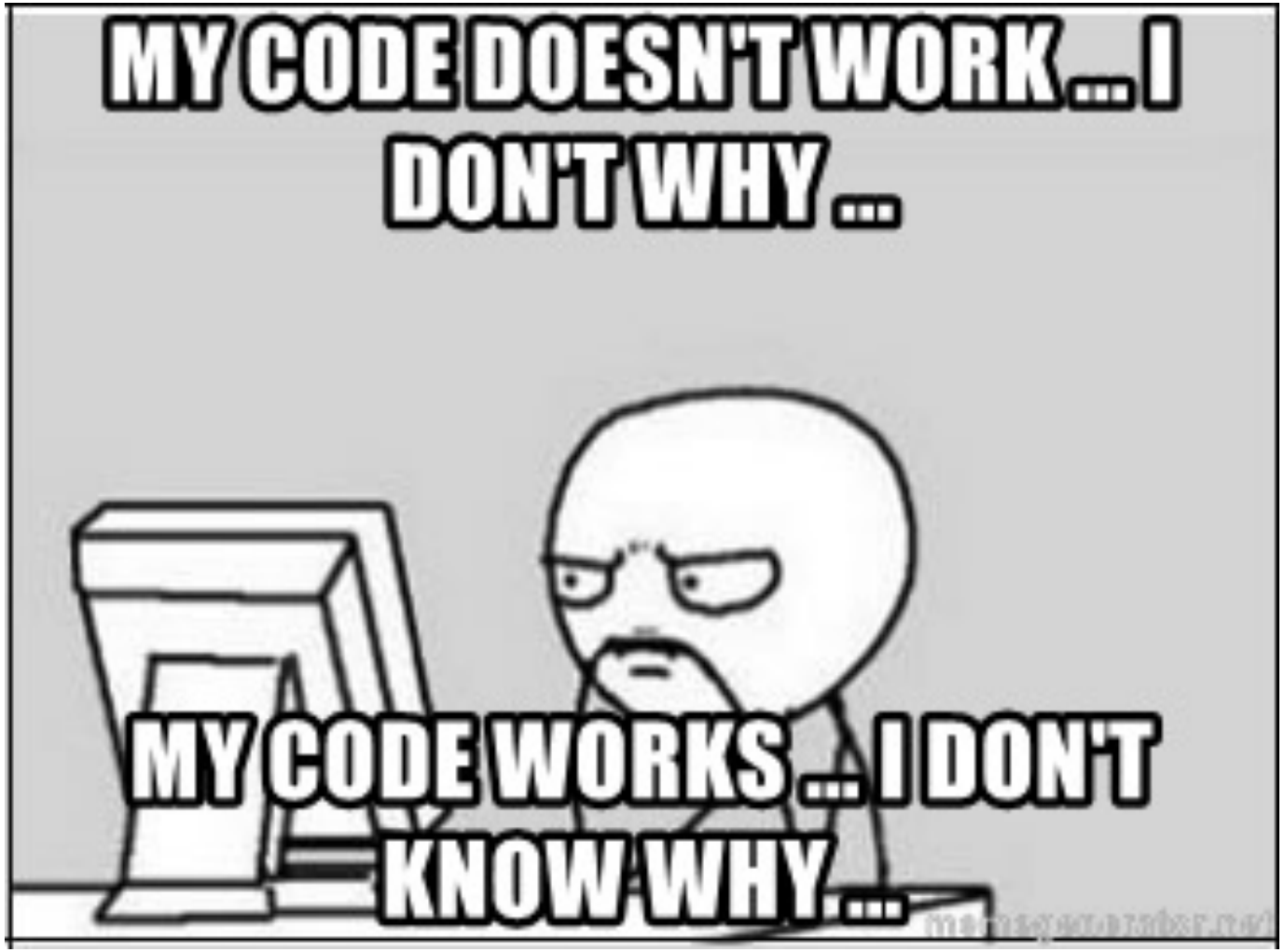
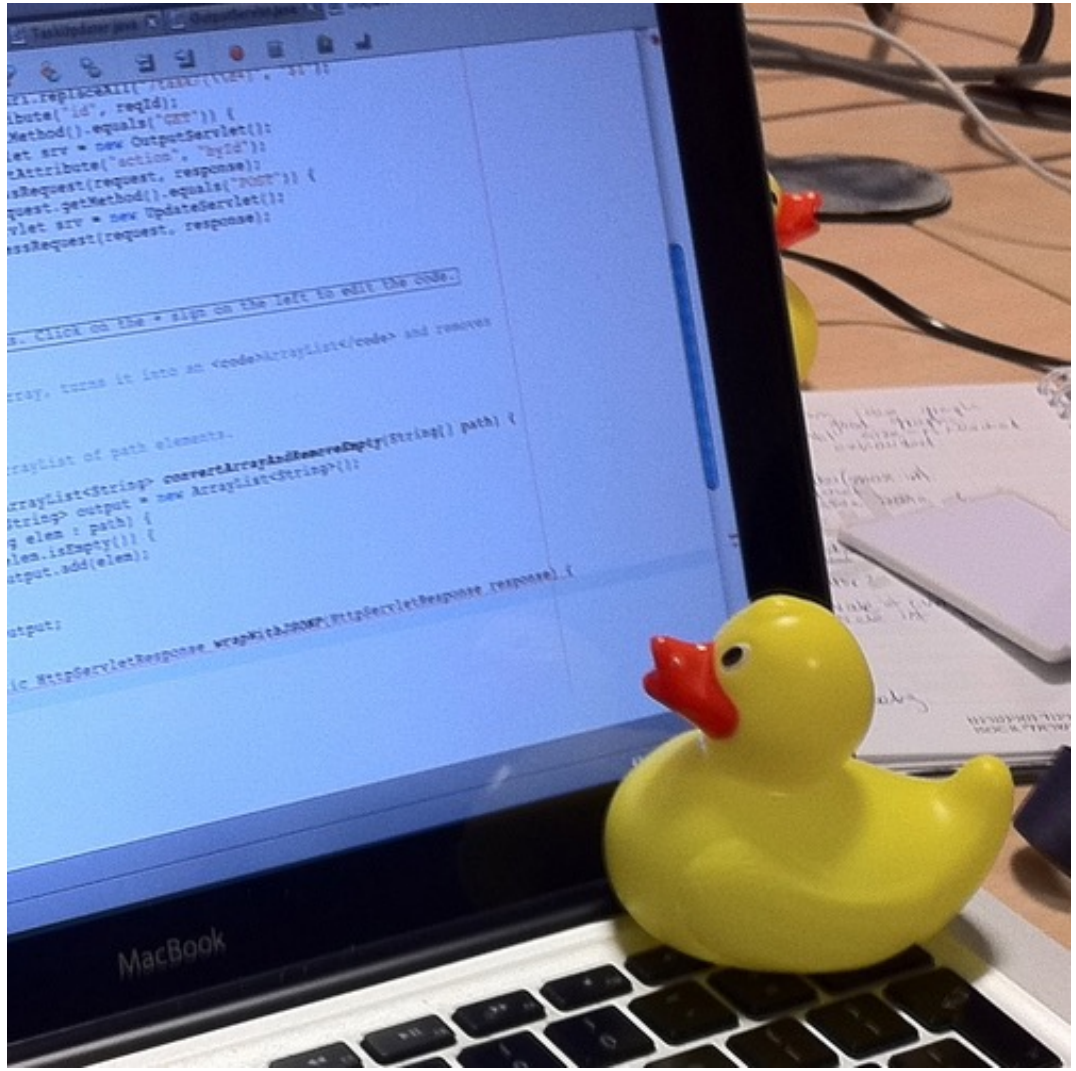

When I wrote this code,
only God & I understood what it did.
ASUS
Now...
only God knows.

Figure 2



Rubber Duck Debugging



https://en.wikipedia.org/wiki/Rubber_duck_debugging

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- Soft skills
- Homework

Data Science is more than Math and Software

Human interaction in data science

- Discovering stakeholders
- Negotiating with data owners
- Customer engagement

https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

Iterating with customers

- As a data scientist, you'll often be working for someone other than yourself.
- Expect under-specified requirements from customers. Iterate.
- Provide incomplete solutions rather than waiting until the product is perfect.

https://en.wikipedia.org/wiki/Minimum_viable_product

When to persist,
When to change course,
When to seek help

Try attacking the challenge for 30 minutes

Then seek help or do something else for a while

https://en.wikipedia.org/wiki/Pomodoro_Technique

Pro-tip when seeking help

How to ask well-formed questions:

https://stackoverflow.com/help/how-to-ask

[Intentional sidetrack to StackOverflow.]

Ask technical questions:

- *Poor*: "I don't understand Python dictionaries" (--> online tutorials)
- *Better*: "When is it appropriate to use a key-value pair?"

- *Poor*: If I submitted this assignment as is, what score would I get?
- *Better*: I am planning to submit the attached assignment, but currently there's an error in the third cell. I've searched online but don't find any references to the error message. Can you provide guidance?

Emotions in Data Science

- As a data scientist, most of your time will be spent in a desert of uncertainty, frustration, and doubt.

- There will be rare short-lived interspersed spikes of excitement and happiness due to events like getting a new dataset, creating a new analytic, getting a new result, or being thanked by a stakeholder.

This experience is normal and does not go away.

See also the psychology of slot machines

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- ~~Soft skills~~
- Homework

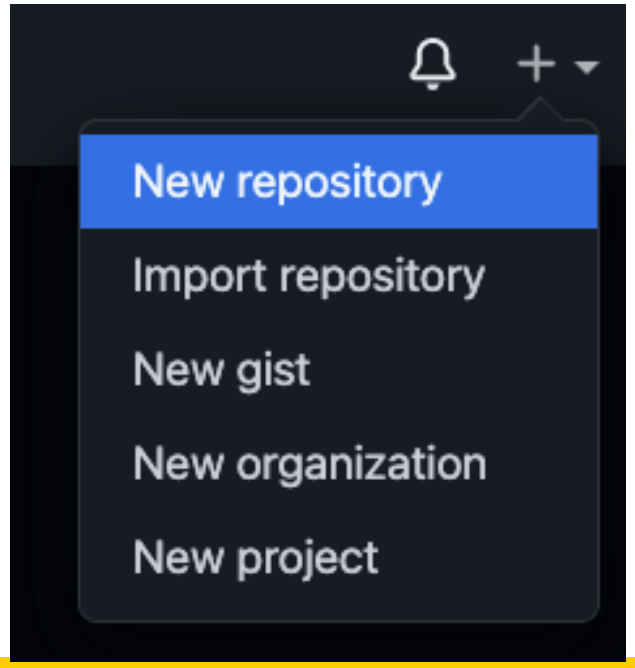
Tasks & Homework

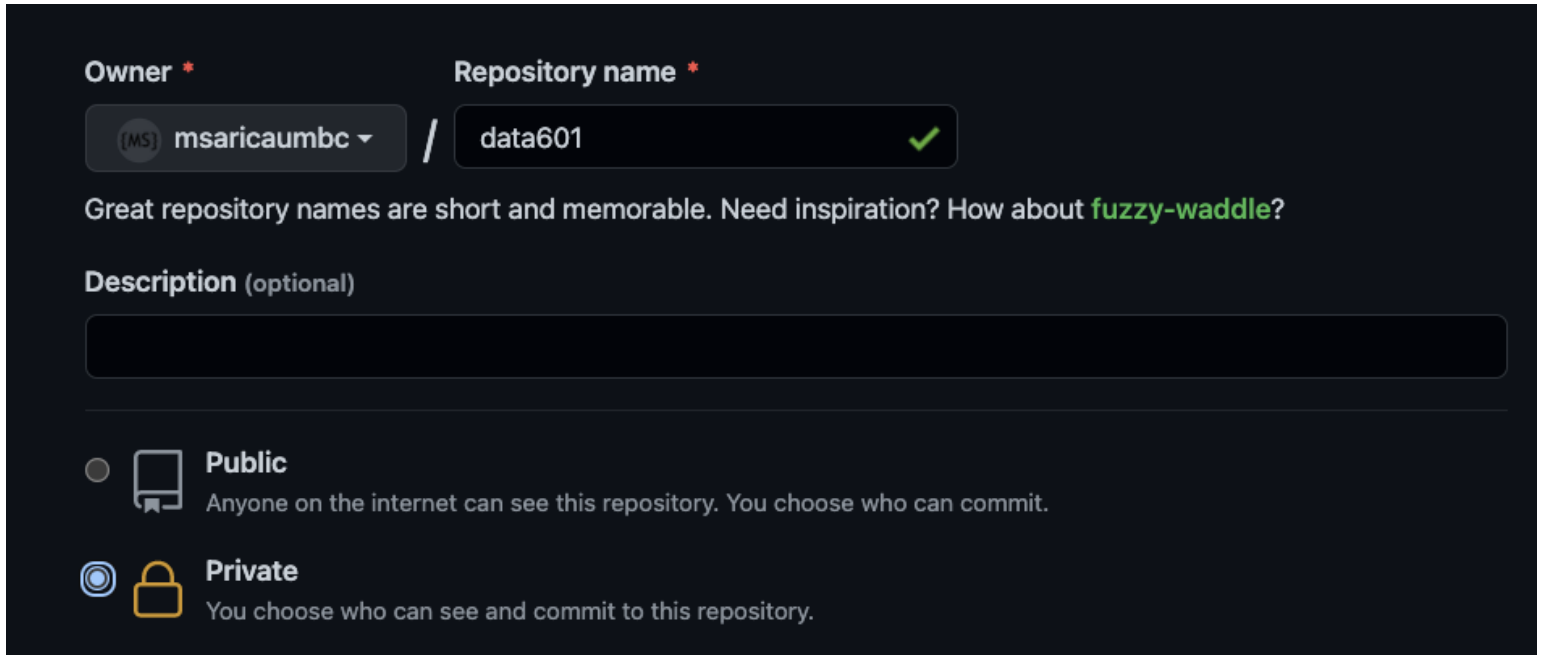
- Install Anaconda
  - Anaconda individual edition - install
  - How to install Anaconda
- Create a github account

Create a new repository

- https://github.com/UMBC-Data-Science/Welcome#github

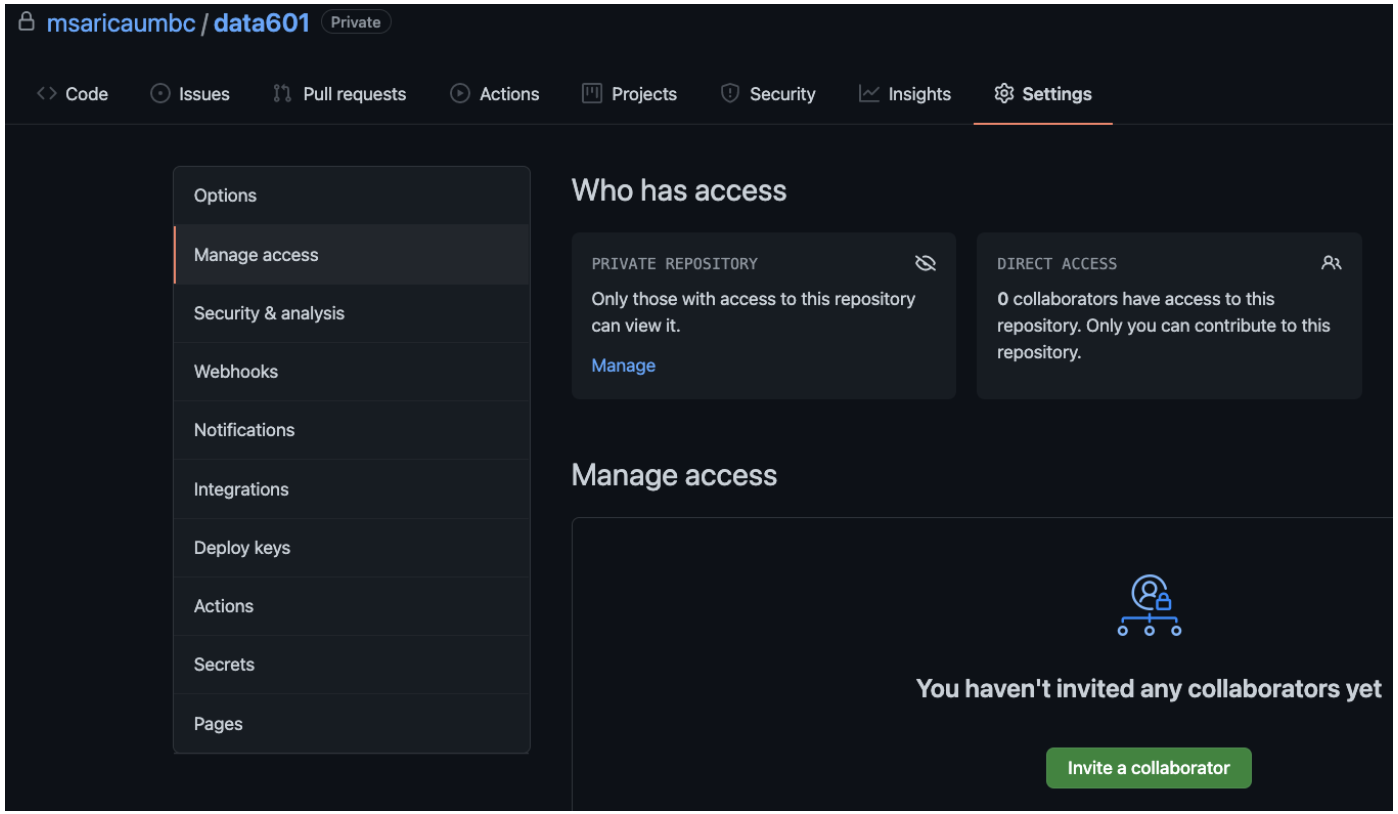
1. Goto github.com and create an account (you may skip this step if you already have an account)
2. Create a new PRIVATE repository & name it Data601.





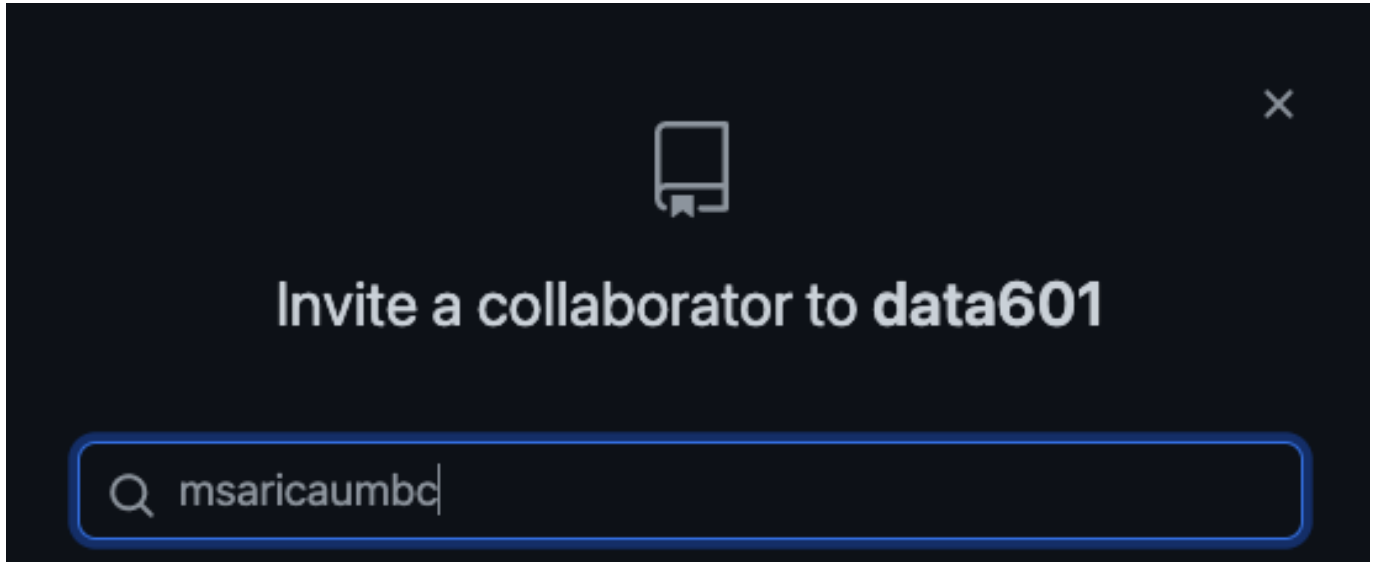
Create a new repository

3. Goto Settings & click Manage access from the left panel. (It may ask you to enter your password).

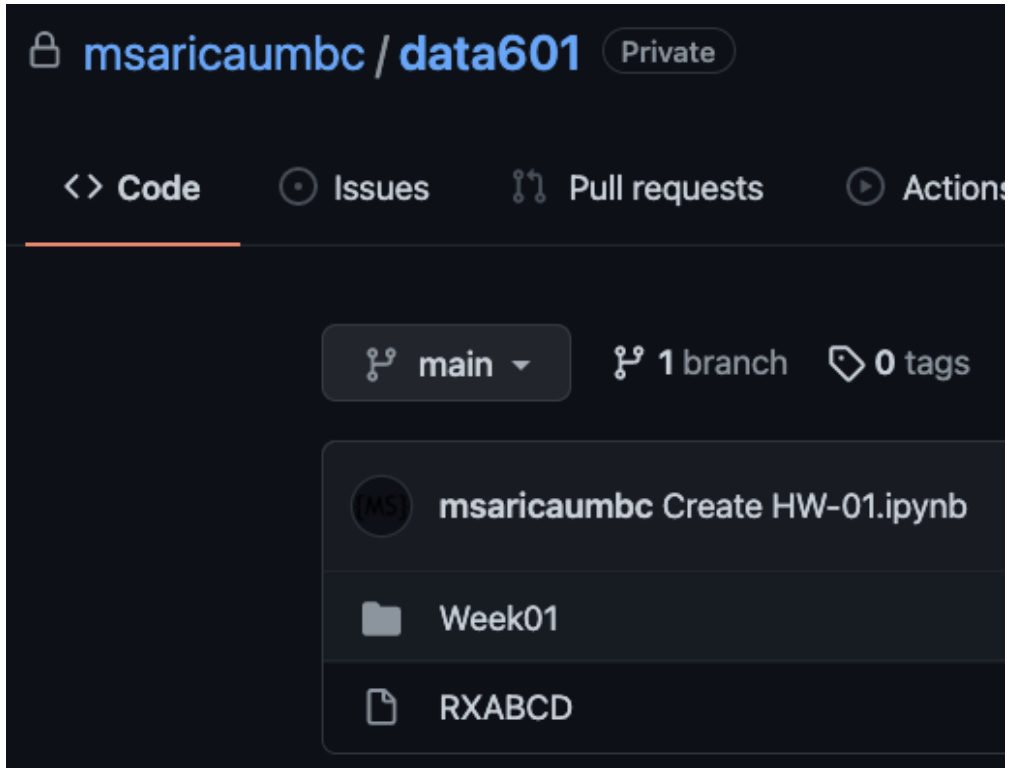


Create a new repository

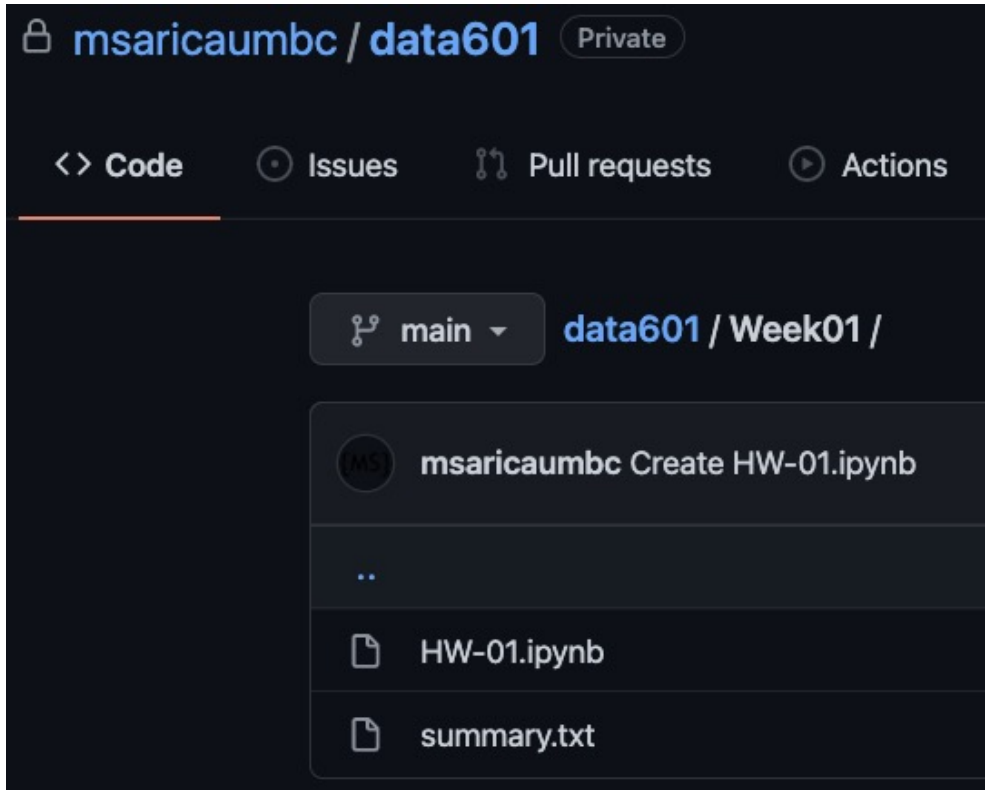
4. Click Invite collaborator button and enter msaricaumbc
5. Click invite collaborator button and enter xxxxxx. (TA)



Your repo should look like the following!



.gitignore



Homework

- Blackboard: Introduce yourself
- Create an empty file with **your student id number** in your repository's root folder
  - Eg: RXABC
- Create a **.gitignore** file and add following lines in this file:

  *.**

  **.csv*
- Create a folder in your github repository **week01**
- Read: First 10 pages of "50 years of data science"

https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734

  - Write a half page summary of the text explaining what is data science save as **summary.txt**
  - Make sure that there are no typos and your summary is grammatically correct!
  - Make sure it's a txt file: not pdf, not word
- Complete HW-01.ipynb (and put your answer in week1 folder)

Materials: https://drive.google.com/drive/folders/1f8OBdc1u2lWKQi5M9Z19UJOUP-N3_v9m