

WELCOMETO DATA SCIENCE

Sri Kanajan

WELCOME TO DATA SCIENCE

LEARNING OBJECTIVES

- Describe the roles and components of a successful learning environment
- Define data science and the data science workflow
- Apply the data science workflow to meet your classmates
- Setup your development environment and review python basics

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Define basic data types used in object-oriented programming
- Recall the Python syntax for lists, dictionaries, and functions
- Create files and navigate directories using the command line interface

DATA SCIENCE

WELCOME TO GAI

FEEDBACK/SUPPORT

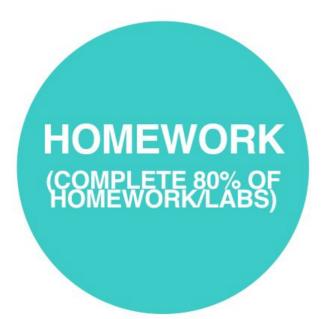
- Access to EIRs: office hours, in class support
- Exit Tickets
- Mid-Course Feedback
- End of Course Feedback



Exit Tickets

- ▶ Meant to help me help you
- But please first try to come directly to me to address your issues and concerns before simply adding complaints to your reviews. Steps in the below order:
 - 1.Ask me questions during class for things that you don't understand. I will assume you understand it unless stated otherwise!
 - Come to office hours (after class)
 - 3.Add question to the "parking lot" and/or slack
 - 4.Do the home works, review your notes, try to search for the answer yourself online. Struggling is a sign that you are learning.
 - 5. If all fails, then go ahead and add it to the review and I can address it directly with you
- Remember that a review for you just takes a few minutes but I have to deal with it for the entire class. Accountability goes both ways! Don't complain, if you don't understand something, let us know and we are here to help!

GA GRADUATION REQUIREMENTS





FINAL PROJECT

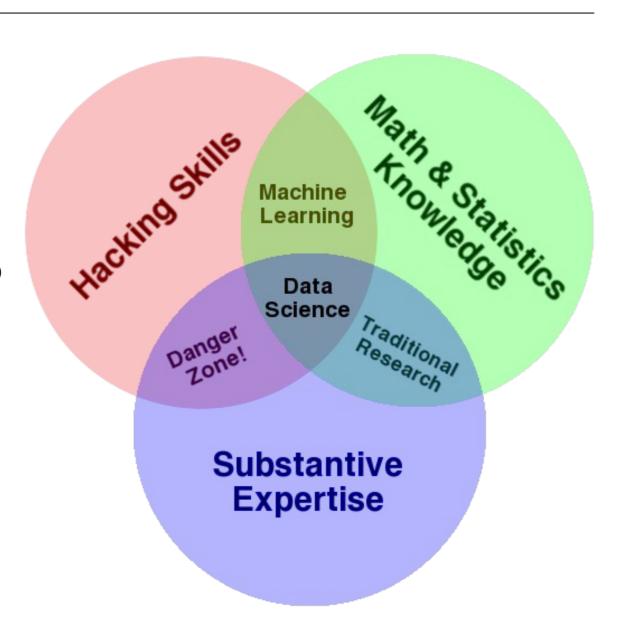


INTRODUCTION

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

- A set of tools and techniques for data
- Interdisciplinary problem-solving
- Application of scientific techniques to practical problems



WHO USES DATA SCIENCE?

NETFLIX







♥ FiveThirtyEight



WHO USES DATA SCIENCE?

Can you think of others?

Position Title	Mathematics /Statistics (e.g. P-value analysis, AB testing)	Database Querying (SQL)	Algorithms (e.g. Supervised learning, Entity Resolution)	Software Engineering (e.g. Python, Java, Object Oriented)	Big Data/Systems Engineering (e.g. Spark, HBase, Hadoop)	Soft Skills/Domain Expertise (E.g. public speaking, presentation skills)
Product Data Scientists	Medium	Medium	Medium	High	High	Medium
Data Engineering	Low	Medium	Low	High	High	Low
Data Scientist	High	Medium	High	Low	Low	High
Business Intelligence Data Scientists	Medium	High	Medium	Low	Low	High
Data Analyst	Low	High	Low	Low	Low	High

DATA SCIENCE BASELINE

ACTIVITY: DATA SCIENCE BASELINE QUIZ



DIRECTIONS (10 minutes)

- 1. Form groups of three.
- 2. Answer the following questions.
 - a. True or False: Gender (coded male=0, female=1) is a continuous variable.
 - b. According to the table on the next slide, admit is the _____
 - i. Outcome
 - ii. Predictor
 - c. Draw a normal distribution, binomial distribution and exponential distribution
 - d. True or False: Linear regression is an unsupervised learning algorithm.

ACTIVITY: DATA SCIENCE BASELINE QUIZ



```
admit gre gpa prestige

0 380 3.61 3

1 660 3.67 3

1 800 4 1

1 640 3.19 4

0 520 2.93 4

1 760 3 2

1 560 2.98 1

0 400 3.08 2

1 540 3.39 3

0 700 3.92 2
```

This dataset is from a UCLA graduate school admission college.

INTRODUCTION

THE DATA SCIENCE WORKFLOW

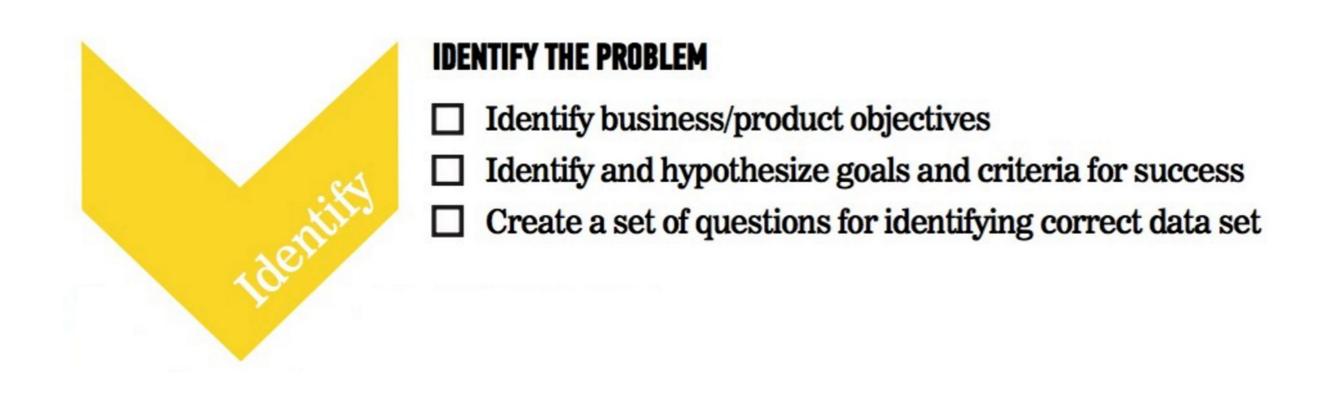
- A methodology for doing Data Science
- Similar to the scientific method
- Helps produce *reliable* and *reproducible* results
 - Reliable: Accurate findings
 - *Reproducible*: Others can follow your steps and get the same results

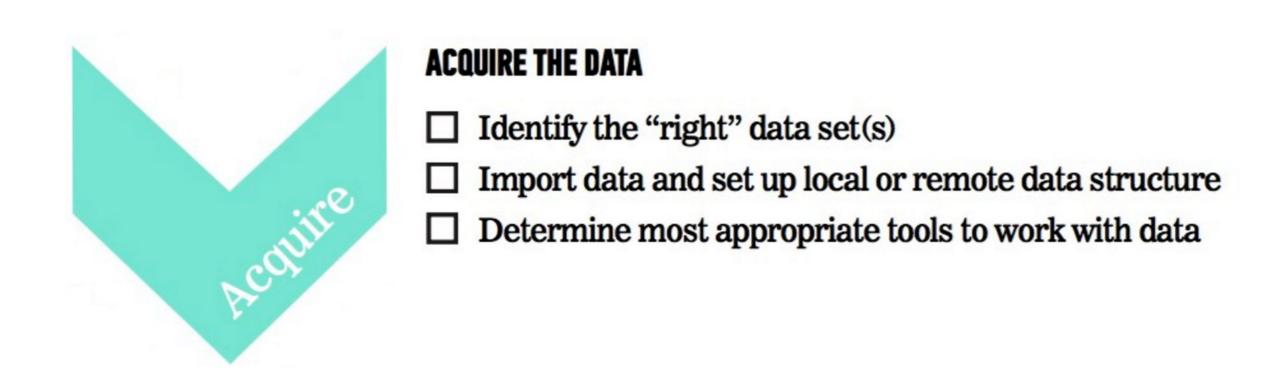
The steps:

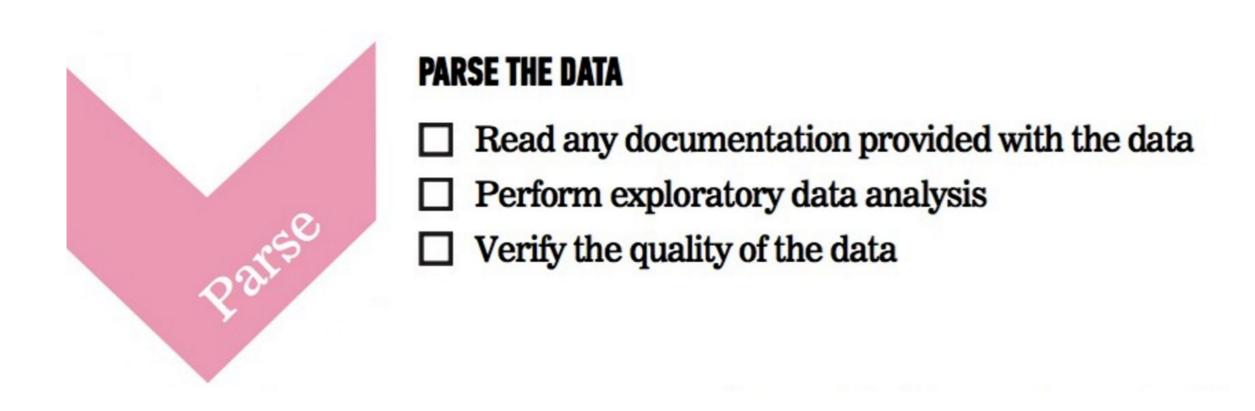
- 1. Identify the problem
- 2. Acquire the data
- 3. Parse the data
- 4. Mine the data
- 5. Refine the data
- 6. Build a data model
- 7. Present the results

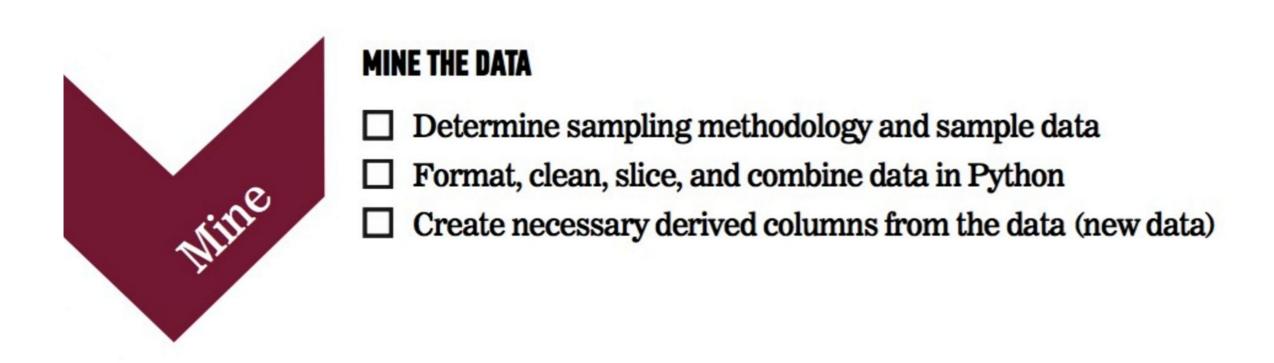
DATA SCIENCE WORKFLOW

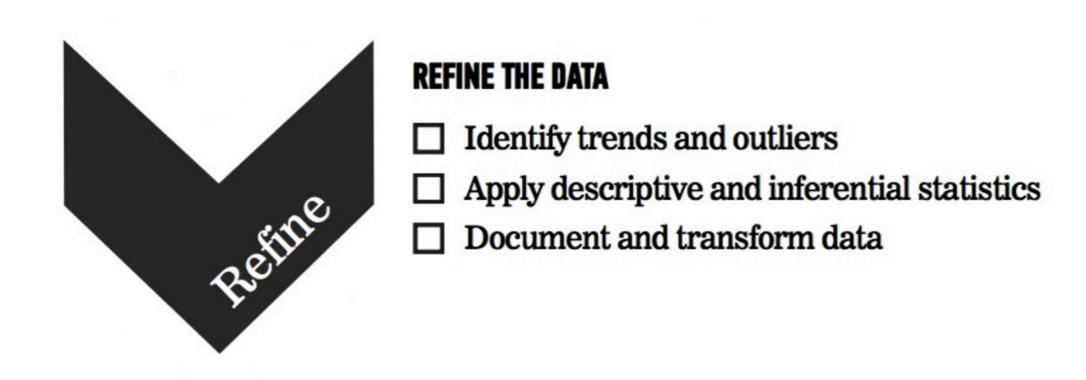


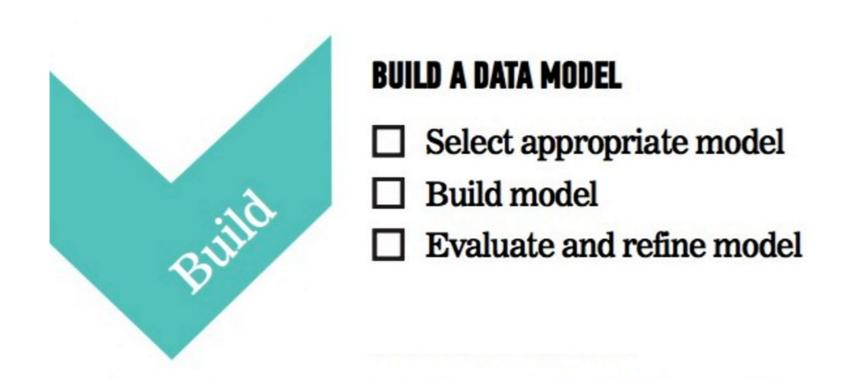














PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

EXAMPLE: DATA DICTIONARY

Example data dictionary

Variable	Description	Type of Variable	
Profession	Title of the account owner	Categorical	
Company Size	1- small, 2- medium, 3- large	Categorical	
Location	Planet of the company	Categorical	
Days Since Last Delivery	Integer	Continuous	
Number of Deliveries	Integer	Continuous	

EXAMPLE DATA VISUALIZATION

- Example presentations and infographics
 - → 512 Paths to the White House
 - 2015 NFL Predictions

GUIDED PRACTICE

DATA SCIENCE WORK FLOW

ACTIVITY: DATA SCIENCE WORKFLOW



DIRECTIONS (25 minutes)

- 1. Divide into 4 groups, each located at a whiteboard.
- 2. **IDENTIFY**: Each group should develop 1 research question they would like to know about their classmates. Create a hypothesis to your question. Don't share your question yet! (5 minutes)
- 3. **ACQUIRE**: Rotate from group to group to collect data for your hypothesis. Have other students write or tally their answers on the whiteboard. (10 minutes)
- 4. **PRESENT**: Communicate the results of your analysis to the class. (10 minutes)
 - a. Create a narrative to summarize your findings.
 - b. Provide a basic visualization for easy comprehension.
 - c. Choose one student to present for the group.

DELIVERABLE

Presentation of the results

ENVIRONMENT SETUP

DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
 - Create a Github account
 - Install Python 2.7 and Anaconda
 - Practice Python syntax, Terminal commands, and Pandas
- iPython Notebook test and Python review

DEV ENVIRONMENT SETUP

- Test your new setup using the lesson 1 starter code available at https://git.generalassemb.ly/NYC-DAT-6-20/NYC-DAT-6-20/blob/master/lessons/lesson-01/code/starter-code/starter-code-1.ipynb in the Github repo
- Ask your classmates and instructor for help if you have problems!

CONCLUSION

REVIEW

CONCLUSION

- You should now be able to answer the following questions:
 - What is Data Science?
 - What is the Data Science workflow?
 - How can you have a successful learning experience at GA?

DATA SCIENCE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

Project: Begin work on Project 1

WELCOME TO DATA SCIENCE

Q&A

WELCOME TO DATA SCIENCE

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET