GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

K. Nathaniel Tucker

Quant, Google

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

WELCOME TO GA!

WELCOME TO GA!

- General Assembly is a global community of individuals empowered to pursue the work we love.

- General Assembly's mission is to build our community by transforming millions of thinkers into creators.

FEEDBACK/SUPPORT

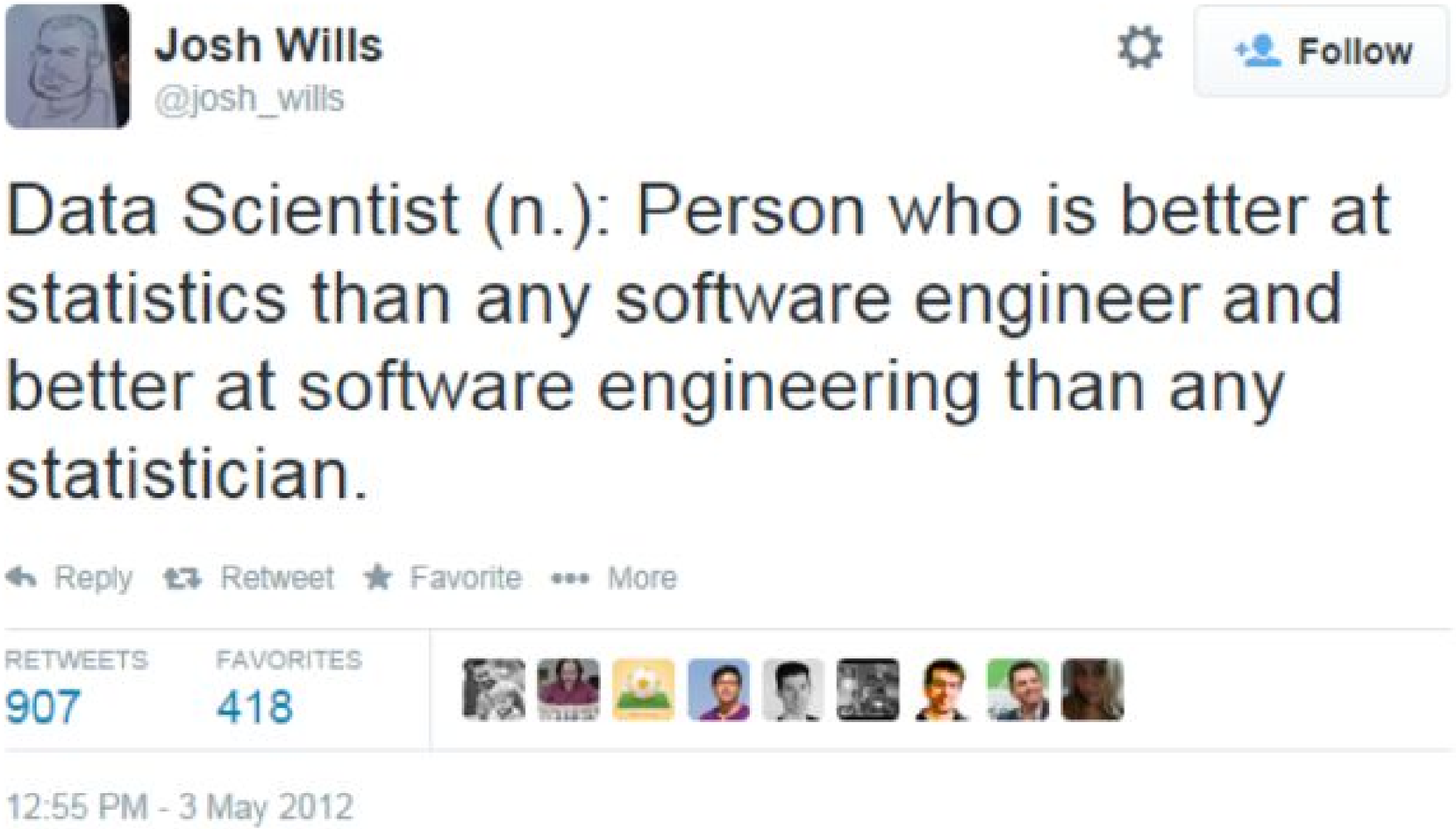
- Access to Instructional Team: office hours, in class support
- Exit Tickets
- Mid-Course Feedback
- End of Course Feedback

GA GRADUATION REQUIREMENTS

WHAT IS DATA SCIENCE?

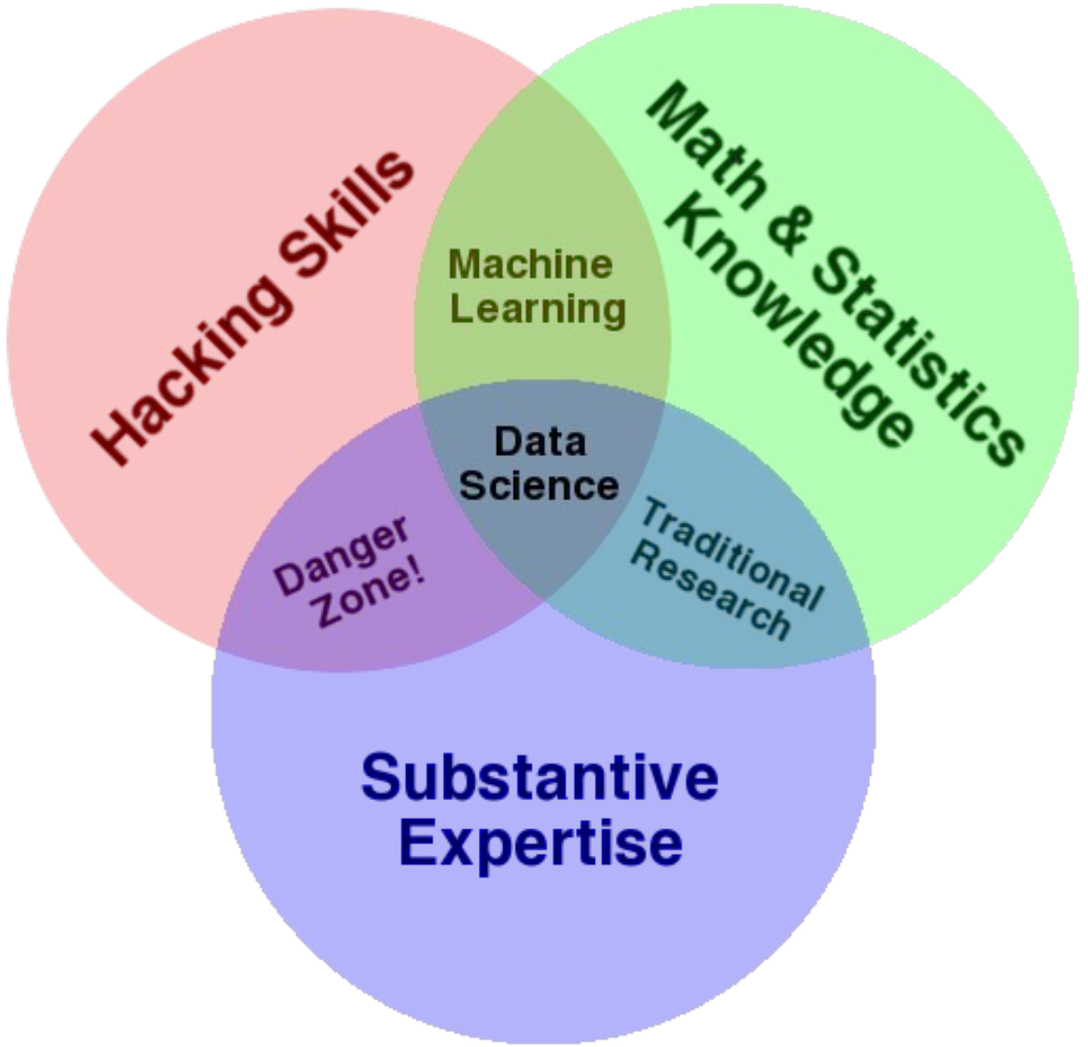
Who is a Data Scientist

Who is a Data Scientist



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

amazon.com

Google

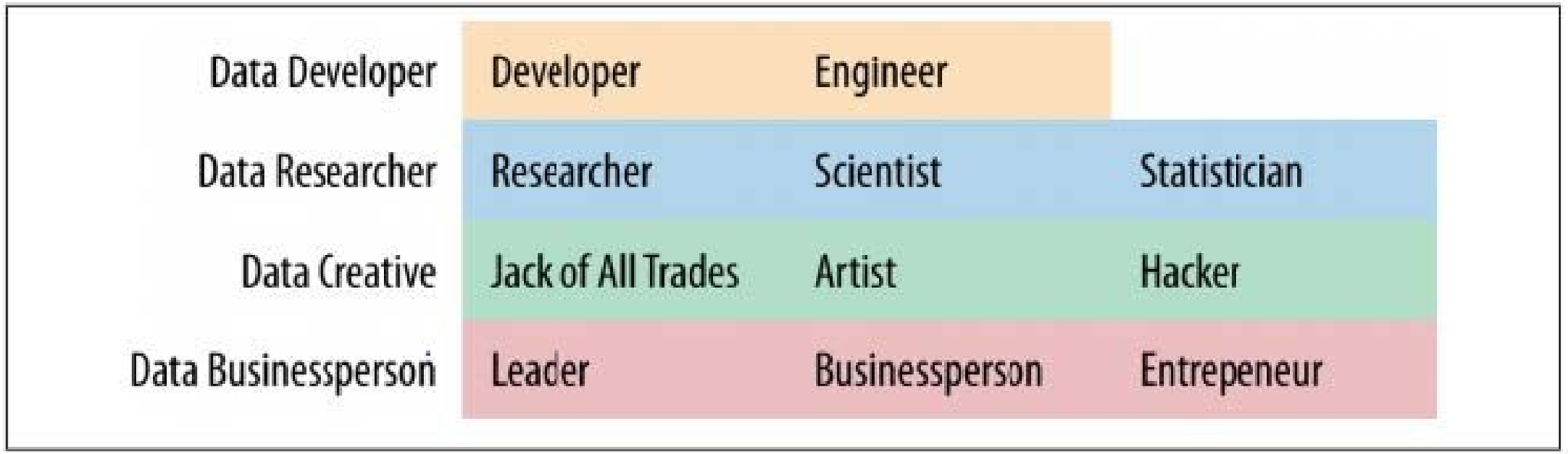
FiveThirtyEight

WHO USES DATA SCIENCE?

- Can you think of companies that use DA? And how so?

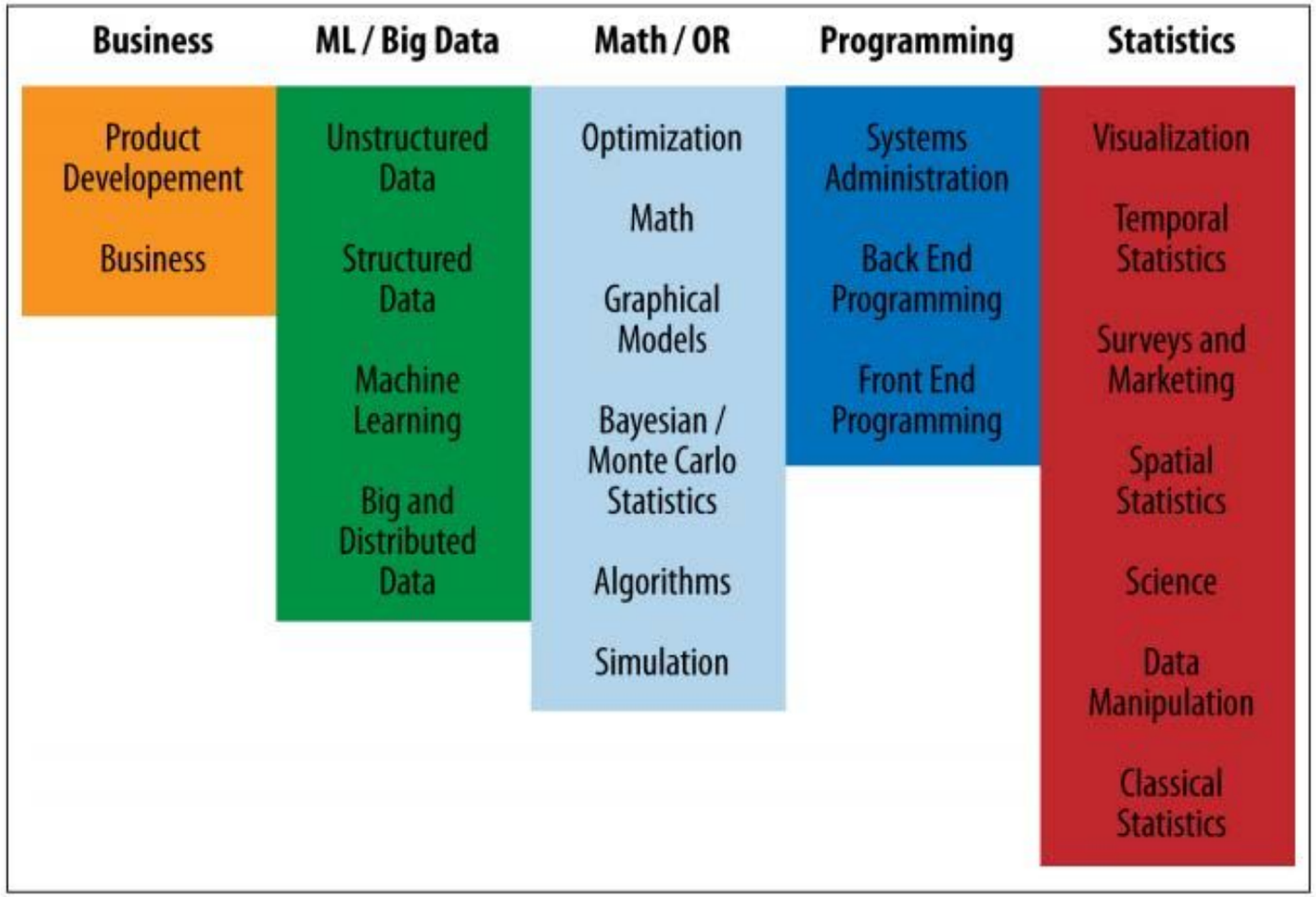
WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of roles, not just one.



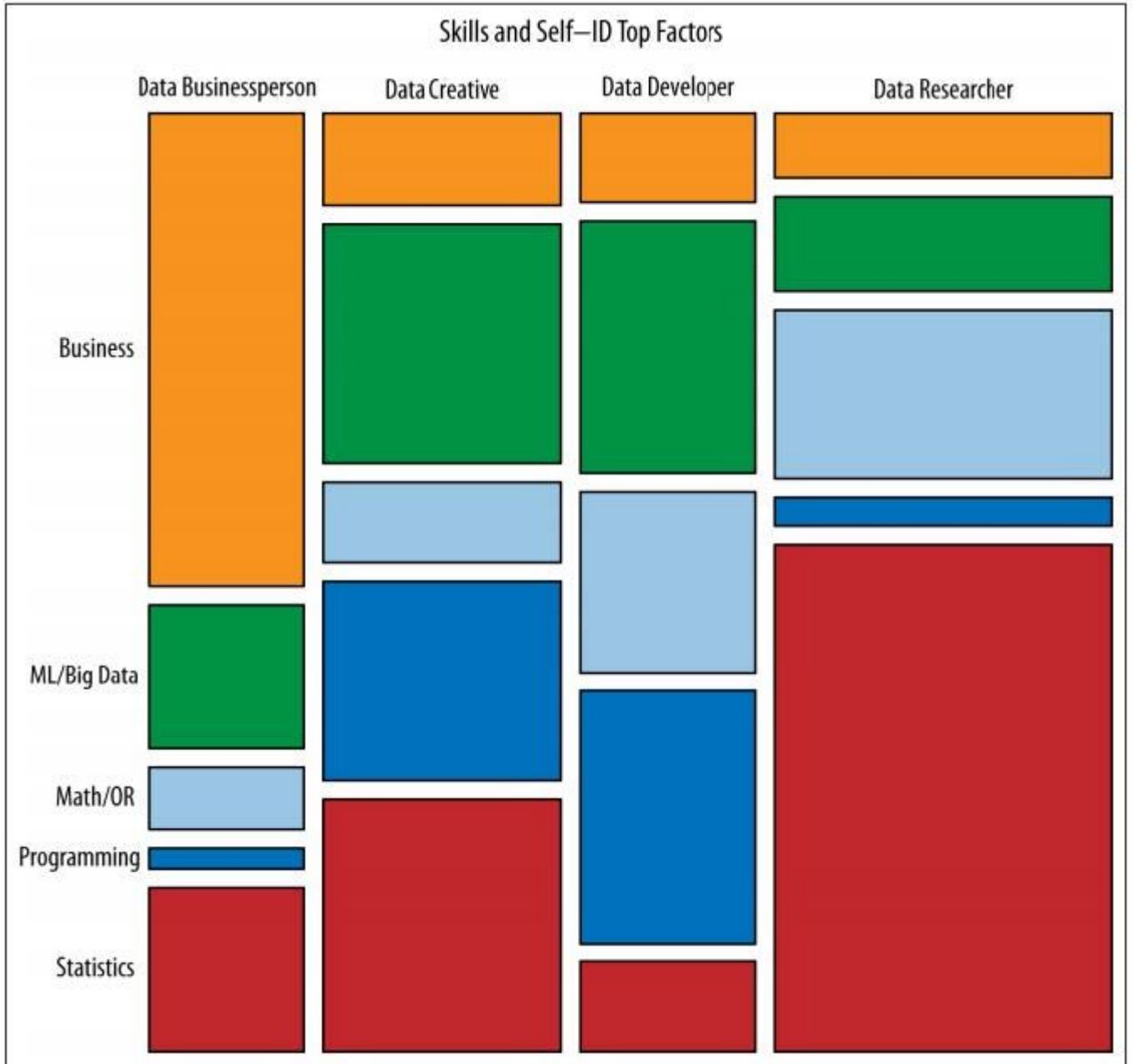
WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of skill sets, not just one.



WHAT ARE THE ROLES IN DATA SCIENCE?

- These roles prioritize different skill sets.
- However, all roles involve some part of each skillset.
- Where are your strengths and weaknesses?



QUIZ

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. Draw a normal distribution
   c. True or False: Linear regression is an unsupervised learning algorithm.
   d. What is a hypothesis test?

THE DATA SCIENCE WORKFLOW

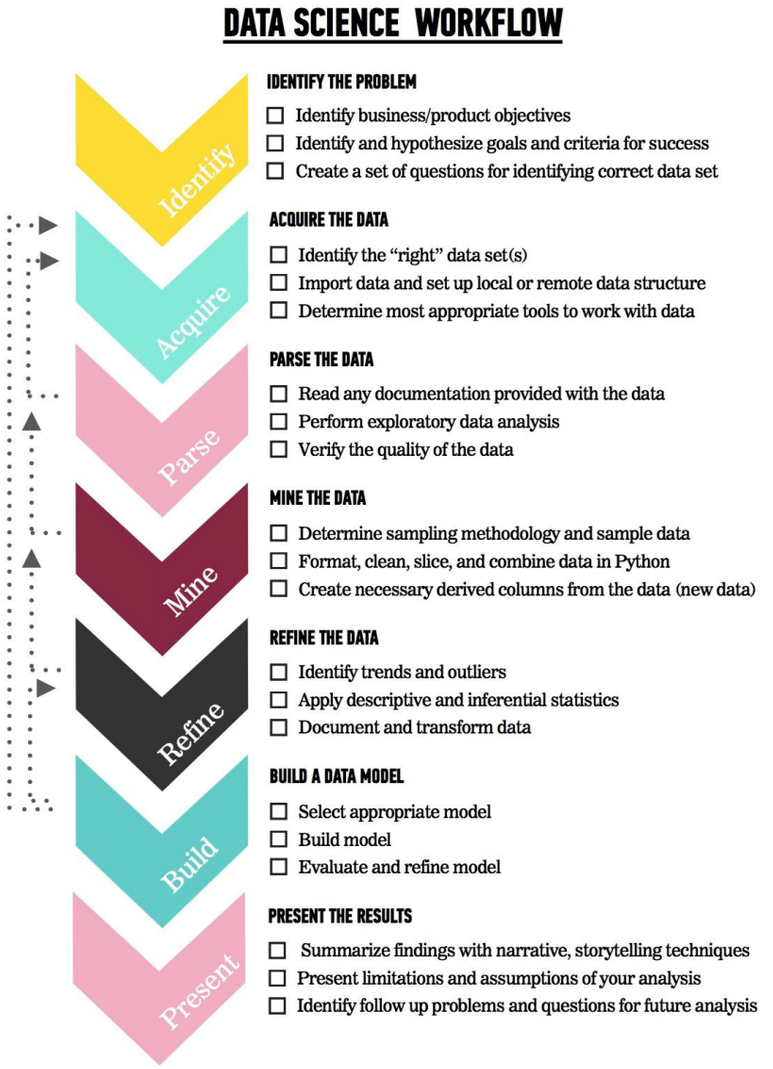
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

OVERVIEW OF THE DATA SCIENCE WORKFLOW

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



OVERVIEW OF THE DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

OVERVIEW OF THE DATA SCIENCE WORKFLOW

ACQUIRE THE DATA

- ☐ Identify the “right” data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

OVERVIEW OF THE DATA SCIENCE WORKFLOW

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

OVERVIEW OF THE 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

FUTURAMA EXAMPLE

- Problem Statement: "Using Planet Express customer data from January 3001-3005, determine how likely previous customers are to request a repeat delivery using demographic information (profession, company size, location) and previous delivery data (days since last delivery, number of total deliveries)."

- We can use the Data Science workflow to work through this problem.

FUTURAMA EXAMPLE: IDENTIFY THE PROBLEM

- Identify the business/product objectives.
- Identify and hypothesize goals and criteria for success.
- Create a set of questions to help you identify the correct data set.

FUTURAMA EXAMPLE: ACQUIRE THE DATA

- Ideal data vs. data that is available
- Learn about limitations of the data.
- What data is available for this example?
- What kind of questions might we want to ask about the data?

FUTURAMA EXAMPLE: ACQUIRE THE DATA

- Questions to ask about the data
  - Is there enough data?
  - Does it appropriately align with the question/problem statement?
  - Can the dataset be trusted? How was it collected?
  - Is this dataset aggregated? Can we use the aggregation or do we need to get it pre-aggregated?

FUTURAMA EXAMPLE: PARSE THE DATA

- Secondary data = we didn't directly collect it ourselves
- 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

Vocab Alert

Categorical & Continuous Values

FUTURAMA EXAMPLE: PARSE THE DATA

- Questions to ask while parsing
  - Is there documentation for the data? Is there a data dictionary?
  - What kind of filtering, sorting, or simple visualizations can help understand the data?
  - What information is contained in the data?
  - What data types are the variables?
  - Are there outliers? Are there trends?

FUTURAMA EXAMPLE: MINE THE DATA

- Think about sampling
- Get to know the data
- Explore outliers
- Address missing values
- Derive new variables (i.e. columns)

FUTURAMA EXAMPLE: MINE THE DATA

- Common steps while mining the data
  - Sample the data with appropriate methodology
  - Explore outliers and null values
  - Format and clean the data
  - Determine how to address missing values
  - Format and combine data; aggregate and derive new columns

FUTURAMA EXAMPLE: REFINE THE DATA

- Use statistics and visualization to identify trends

- Example of basic statistics

Variable	Mean (STD) or Frequency (%)
Number of Deliveries	50.0 (10)
Earth	50 (10%)
Amphibios 9	100 (20%)
Bogad	100 (20%)
Colgate 8	100 (20%)
Other	150 (30%)

FUTURAMA EXAMPLE: REFINE THE DATA

- Descriptive stats help refine by
  - Identifying trends and outliers
  - Deciding how to deal with outliers
  - Applying descriptive and inferential statistics
  - Determining visualization techniques for different data types
  - Transforming data

FUTURAMA EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome
- Example model statement: “We completed a logistic regression using Statsmodels v. XX. We calculated the probability of a customer placing another order with Planet Express.”
- Steps for model building

FUTURAMA EXAMPLE: CREATE A DATA MODEL

- The steps for model building are
  - Select the appropriate model
  - Build the model
  - Evaluate and refine the model
  - Predict outcomes and action items

FUTURAMA EXAMPLE: PRESENT THE RESULTS

- You have to effectively communicate your results for them to matter!
- Ranges from a simple email to a complex web graphic.
- Make sure to consider your audience.
- A presentation for fellow data scientists will be drastically different from a presentation for an executive.

FUTURAMA EXAMPLE: PRESENT THE RESULTS

- Key factors of a good presentation include
  - Summarize findings with narrative and storytelling techniques
  - Refine your visualizations for broader comprehension
  - Present both limitations and assumptions
  - Determine the integrity of your analyses
  - Consider the degree of disclosure for various stakeholders
  - Test and evaluate the effectiveness of your presentation beforehand

FUTURAMA EXAMPLE: PRESENT THE RESULTS

- Example presentations and infographics
  - 512 Paths to the White House
  - Who Old Are You?
  - 2015 NFL Predictions

GUIDED PRACTICE

DATA SCIENCE WORK FLOW

ACTIVITY: DATA SCIENCE WORKFLOW

DIRECTIONS (20 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! (3 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. (8 minutes)
4. **PRESENT**: Communicate the results of your analysis to the class. (9 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

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

History Alert

CLI

DEMO

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 */lessons/lesson-1/code/starter-code/lesson1-starter-code.ipynb* in the Github repo

- Ask your classmates and instructor for help if you have problems!

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?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project: Begin work on Project 1

WELCOME TO DATA SCIENCE

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

Bonus Material

Finding the Truth

Prediction and Estimation

- Future vs. Past and Present
- Really the same question