FOR INSTRUCTOR PURPOSES ONLY

INSTRUCTOR NOTES

- Insert Text Here

FOR INSTRUCTOR PURPOSES ONLY

MATERIALS

- Insert Text Here

FOR INSTRUCTOR PURPOSES ONLY

PRE-WORK

- Insert Text Here

GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Zack Peterson

Data Scientist

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

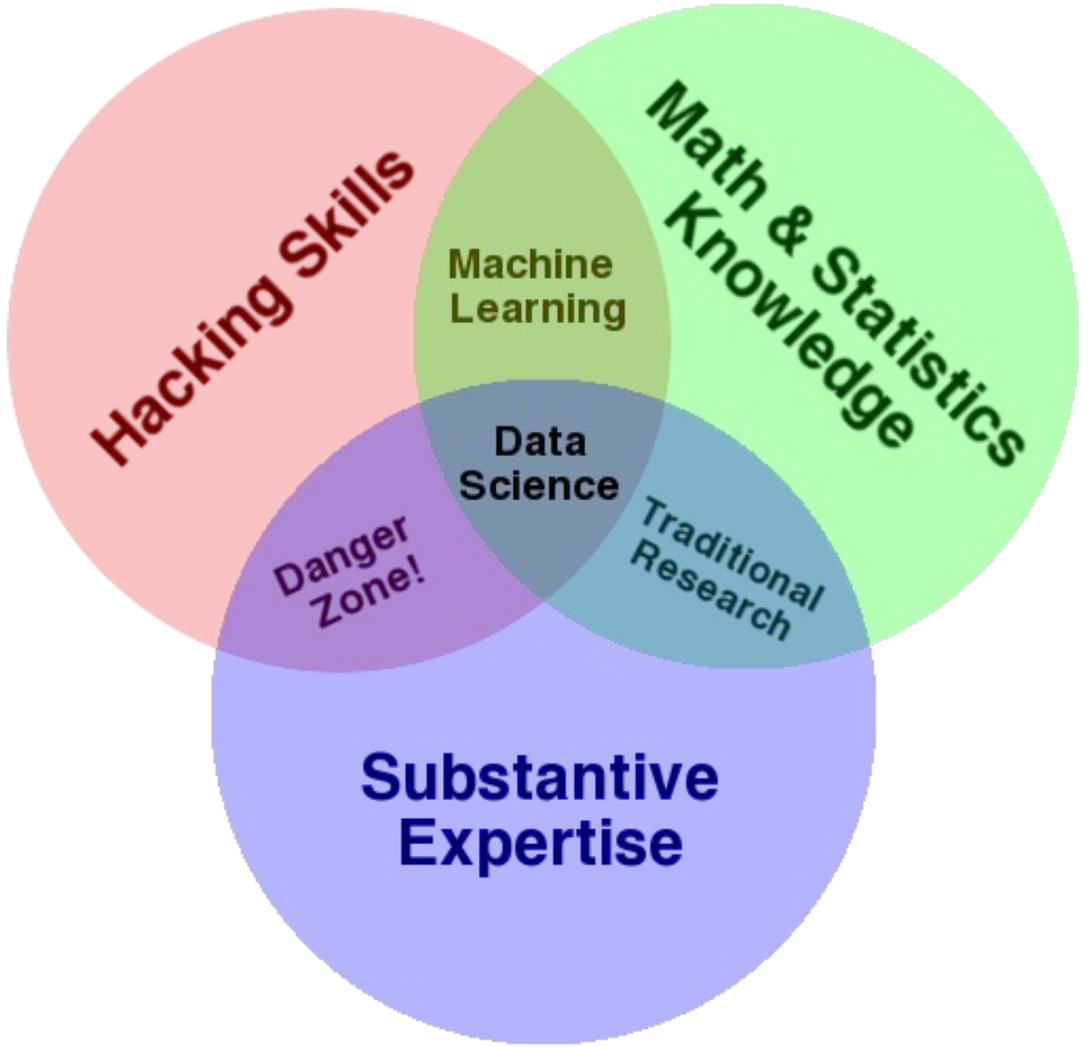
PRE-WORK REVIEW

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

WHAT IS DATA SCIENCE?
WHY 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

amazon.com

Google

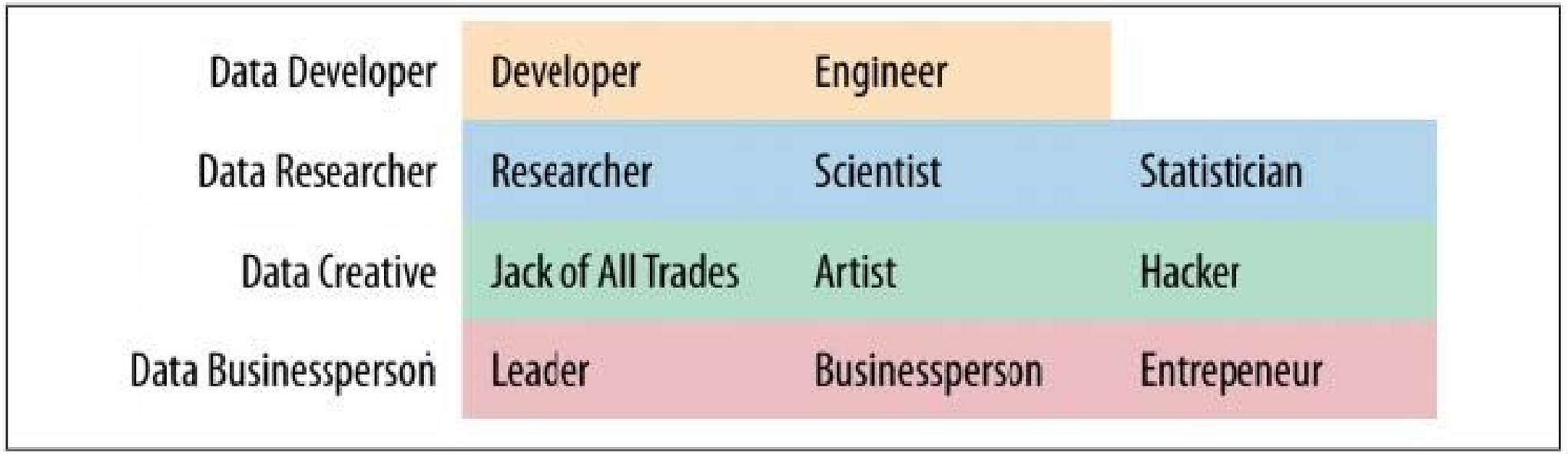
FiveThirtyEight

WHO USES DATA SCIENCE?

- Can you think of others?

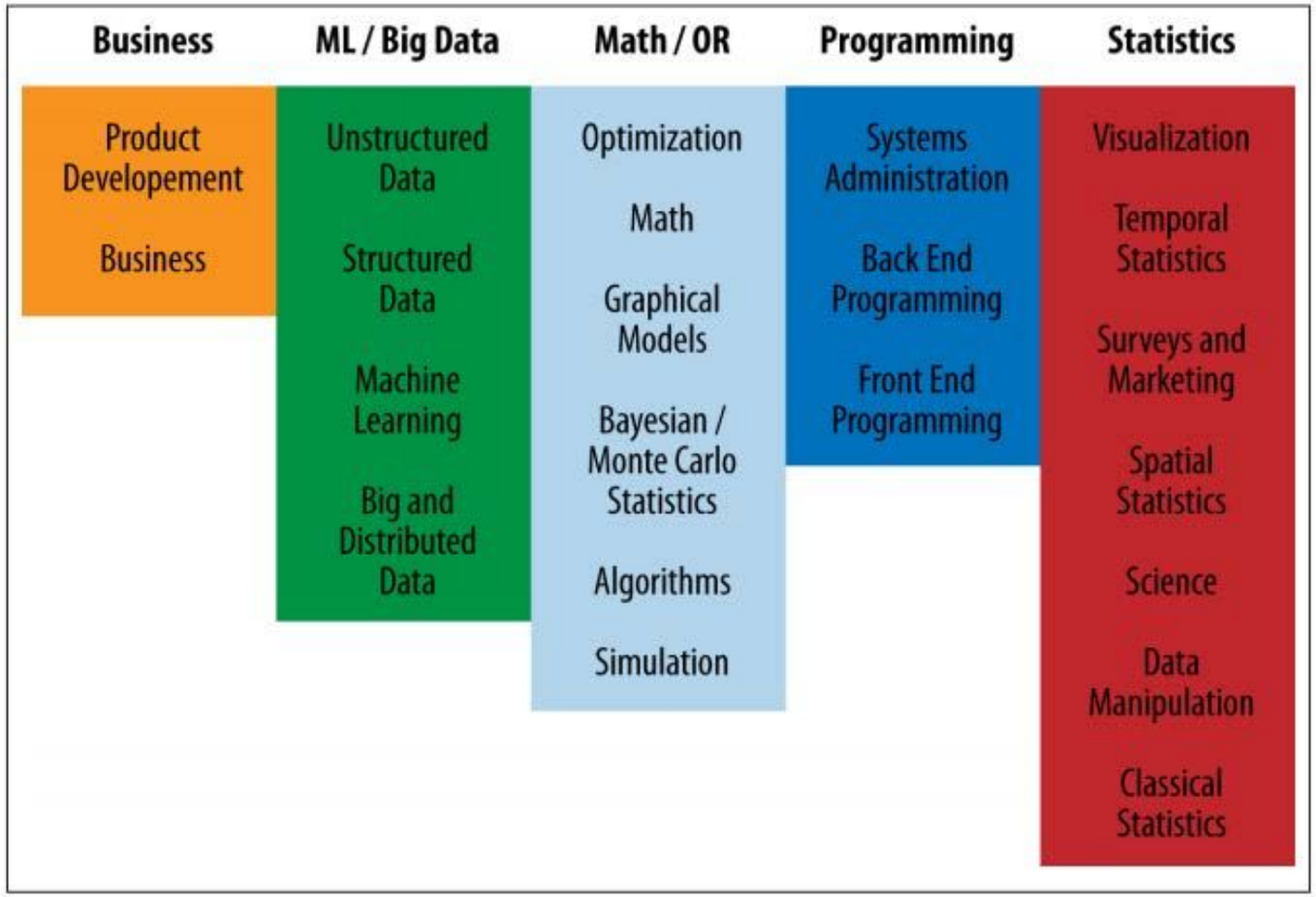
WHAT ARE THE ROLES IN DATA SCIENCE?

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



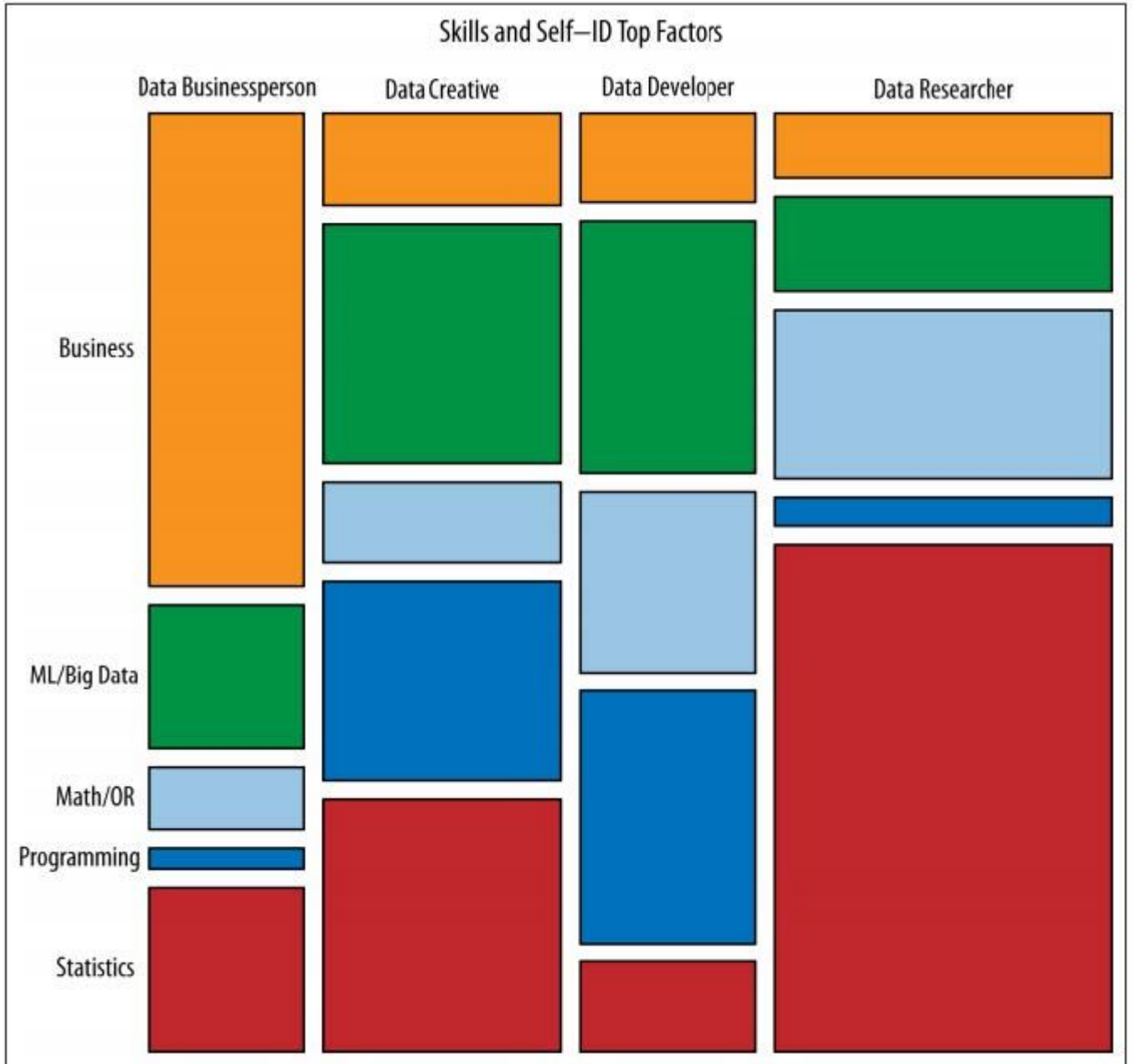
WHAT ARE THE ROLES IN DATA SCIENCE?

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

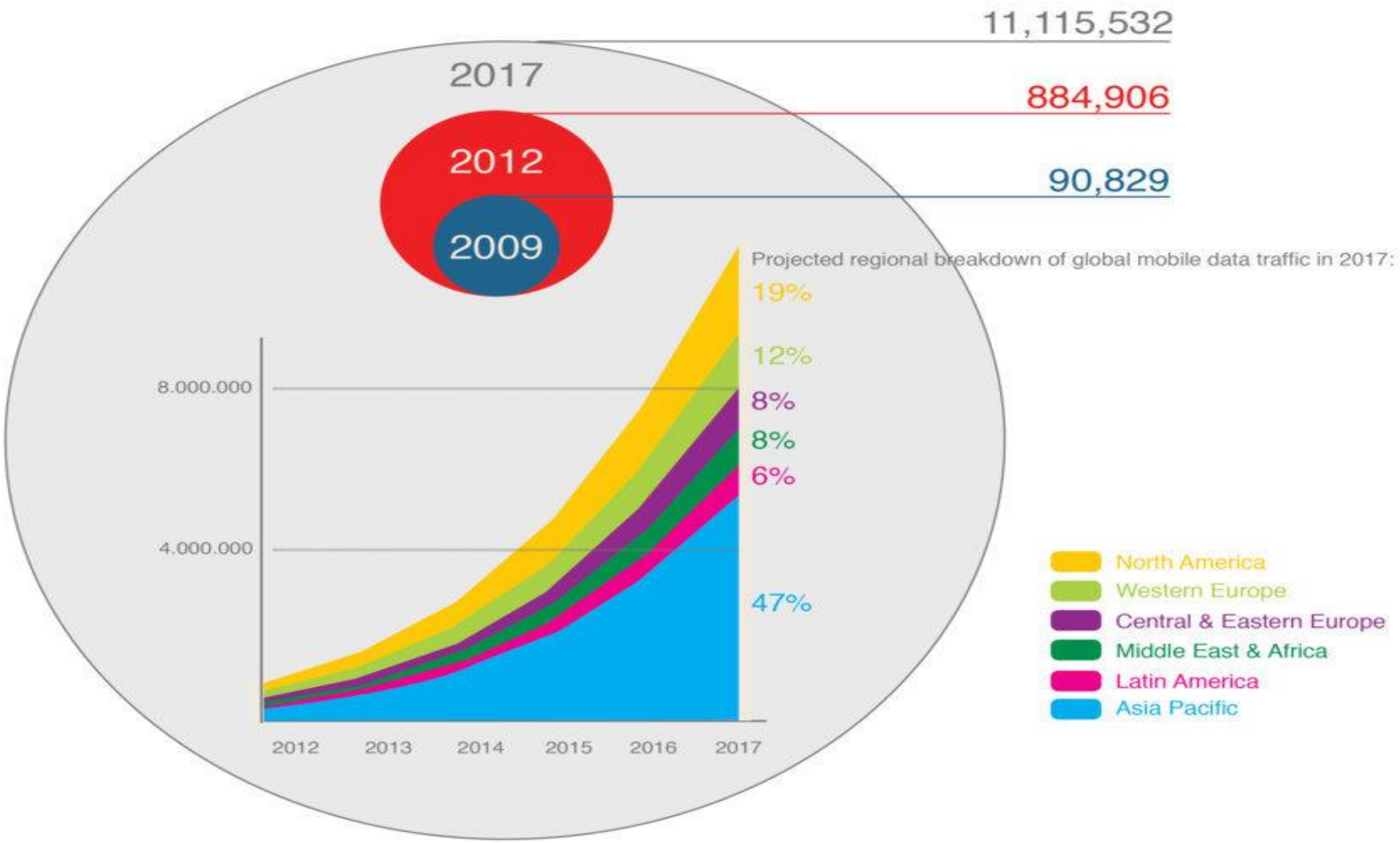


WHAT ARE THE ROLES IN DATA SCIENCE?

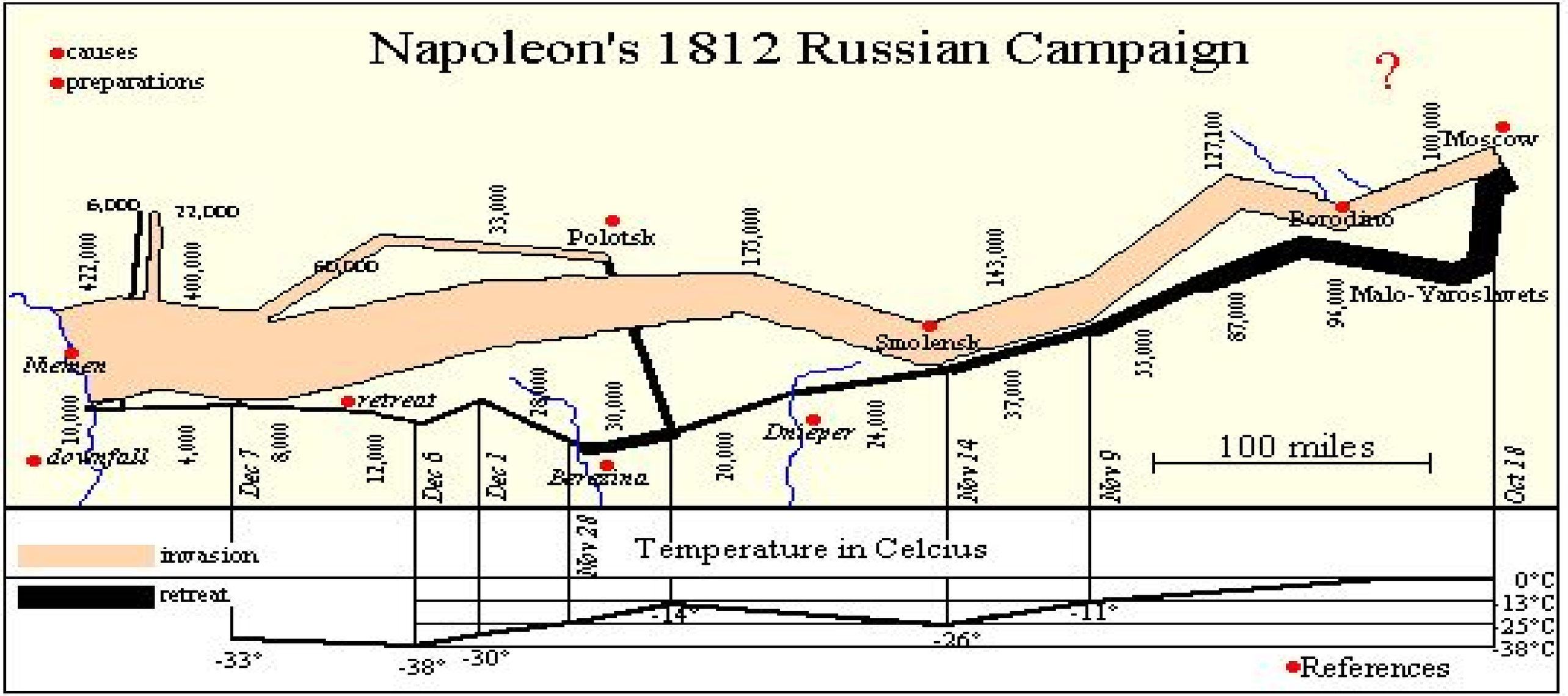
- These roles prioritize different skill sets.
- However, all roles involve some part of each skillset.
- Where are your strengths and weaknesses?



Global Mobile Data - Traffic growth & forecast (terabytes per month)



Data Scientists tell stories.



WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Data Scientists tend to use machine learning algorithms to address problems

QUIZ

DATA SCIENCE BASELINE

ACTIVITY: DATA SCIENCE BASELINE QUIZ

EXERCISE

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

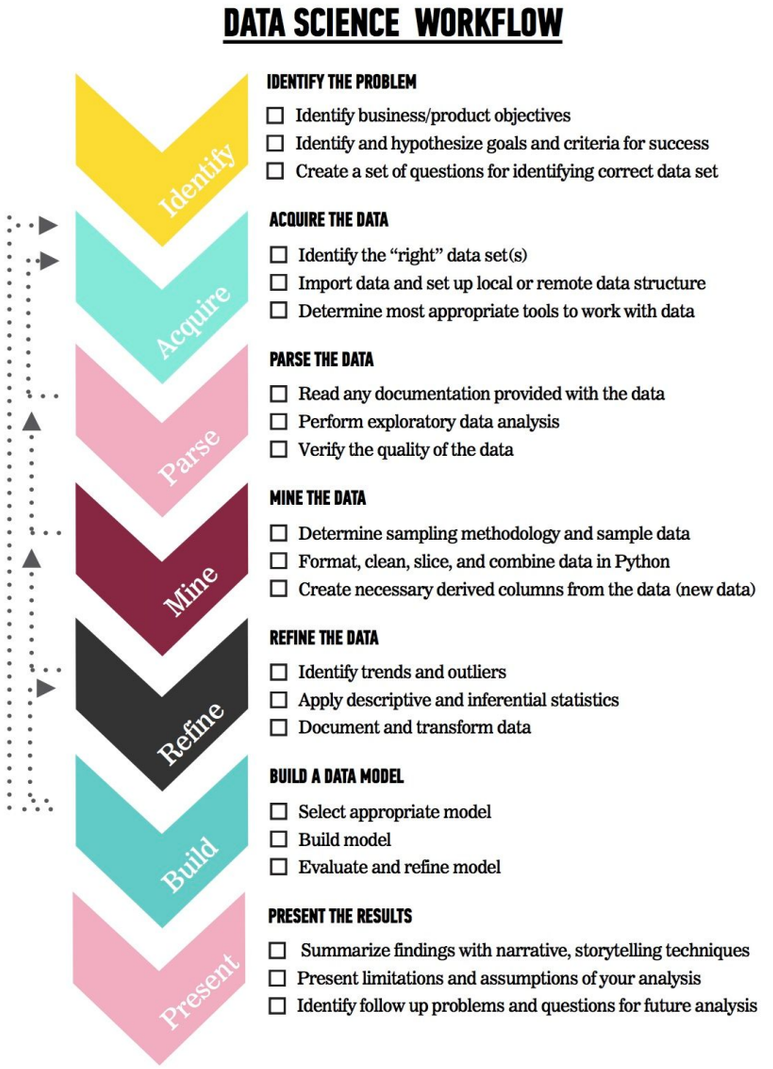
Activity

Form groups of 3-4 and organize these slides in the proper order.

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

Credit Card Fraud Example

- Who has had an experience with credit card fraud?
- How can we use Data Science to help mitigate fraud?
- We can use the Data Science workflow to work through this problem.

Fraud: IDENTIFY THE PROBLEM

- Someone steals your card and spends money.
- Identify and hypothesize goals and criteria for success.
- Create a set of questions to help you identify the correct data set.

Fraud: ACQUIRE THE DATA

- Does the data exist internally, externally, or do we have to create it?

- Learn about limitations of the data.
  - Is there enough data?

Fraud EXAMPLE: PARSE THE DATA

- Example data dictionary

Variable	Description	Type of Variable
Transaction Time	Time of Transaction	Date Time
Amount	Amount of Transaction	Numerical
Location	Business of Transaction	Categorical
Average Transaction	Average customer transaction	Numerical

Fraud 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 data types are the variables?
  - Are there outliers? Are there trends?

Fraud EXAMPLE: MINE THE DATA

- Think about sampling
- Address missing values
- Derive new variables (i.e. columns)

Fraud EXAMPLE: REFINE THE DATA

- Use statistics and visualization to identify trends

- Example of basic statistics
  - Mean
  - Median
  - Mode
  - Standard Deviation

Fraud 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/scaling data

Fraud EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome
- Example model statement: “We completed a logistic regression using Python that calculates the probability that a transaction is fraudulent”
- Evaluate and refine the model

Fraud 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.

Fraud 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

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

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 or just Jupyter Notebook

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