GENERAL ASSEMBLY

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

DATA SCIENCE

WELCOME TO GA!

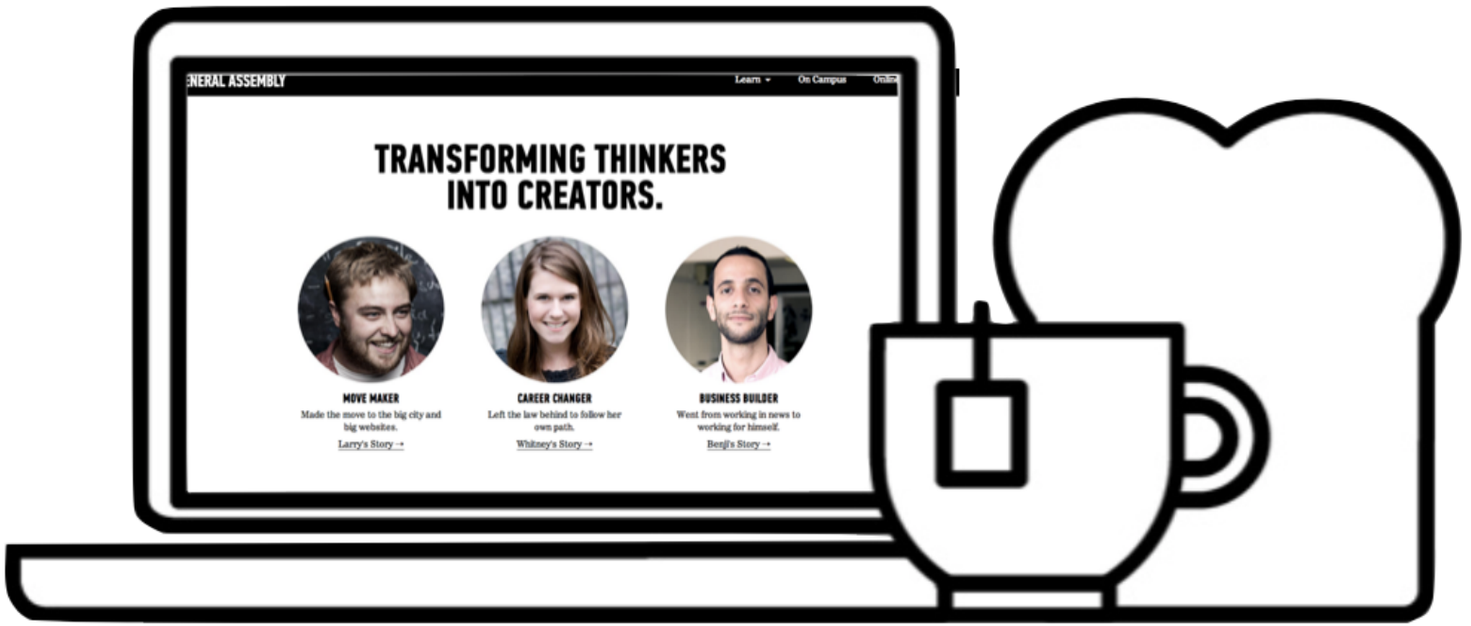
WELCOME TO GA!

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

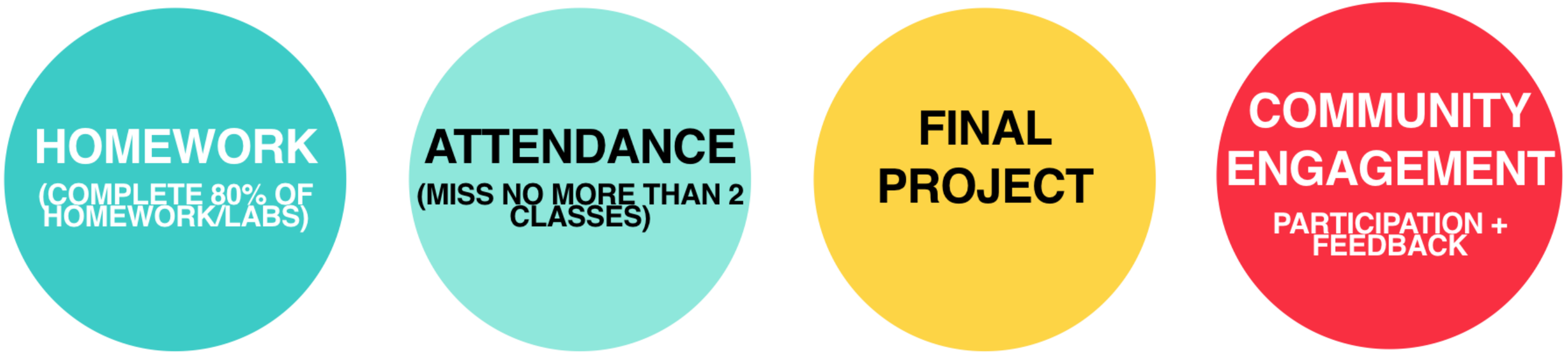
- General Assembly's mission is to build our community by transforming millions of thinkers into creators.

FEEDBACK/SUPPORT

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



GA GRADUATION REQUIREMENTS



FOREVER AND EVER

It's not just about altruism, your network is your most valuable asset

Alumni have started companies together and recruited other alumni to join their teams

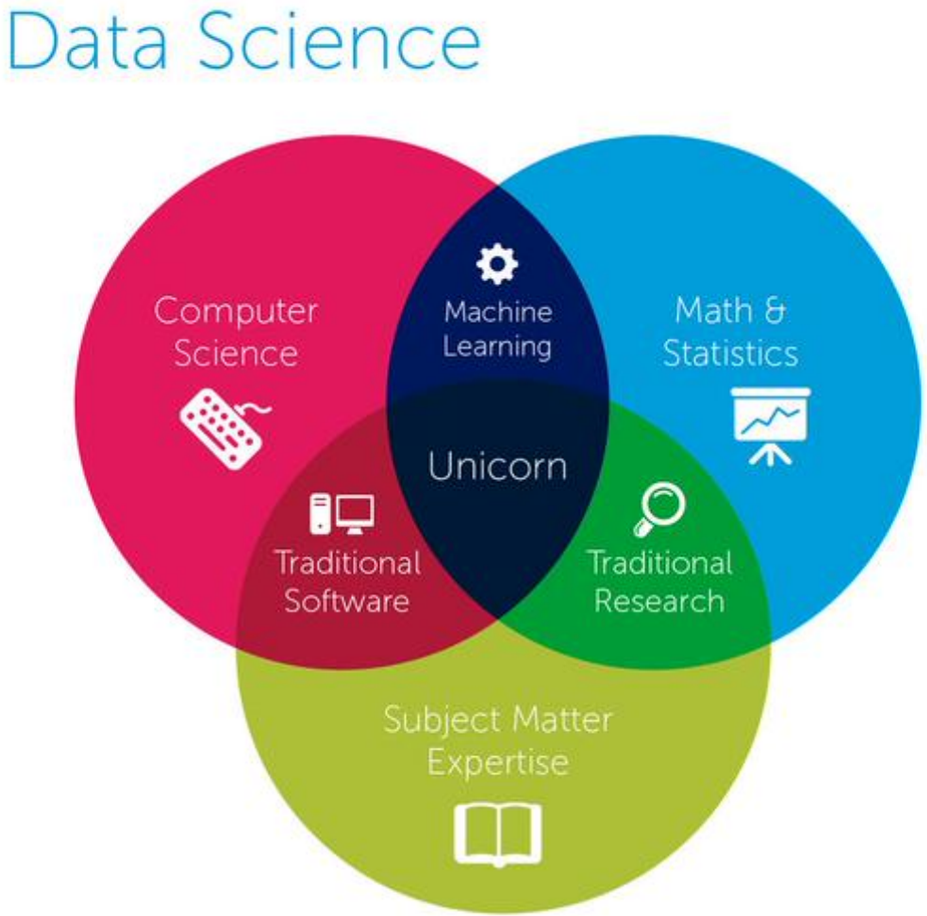
You're part of the alumni community forever

We can't wait to have you back on campus

WHAT IS DATA SCIENCE?

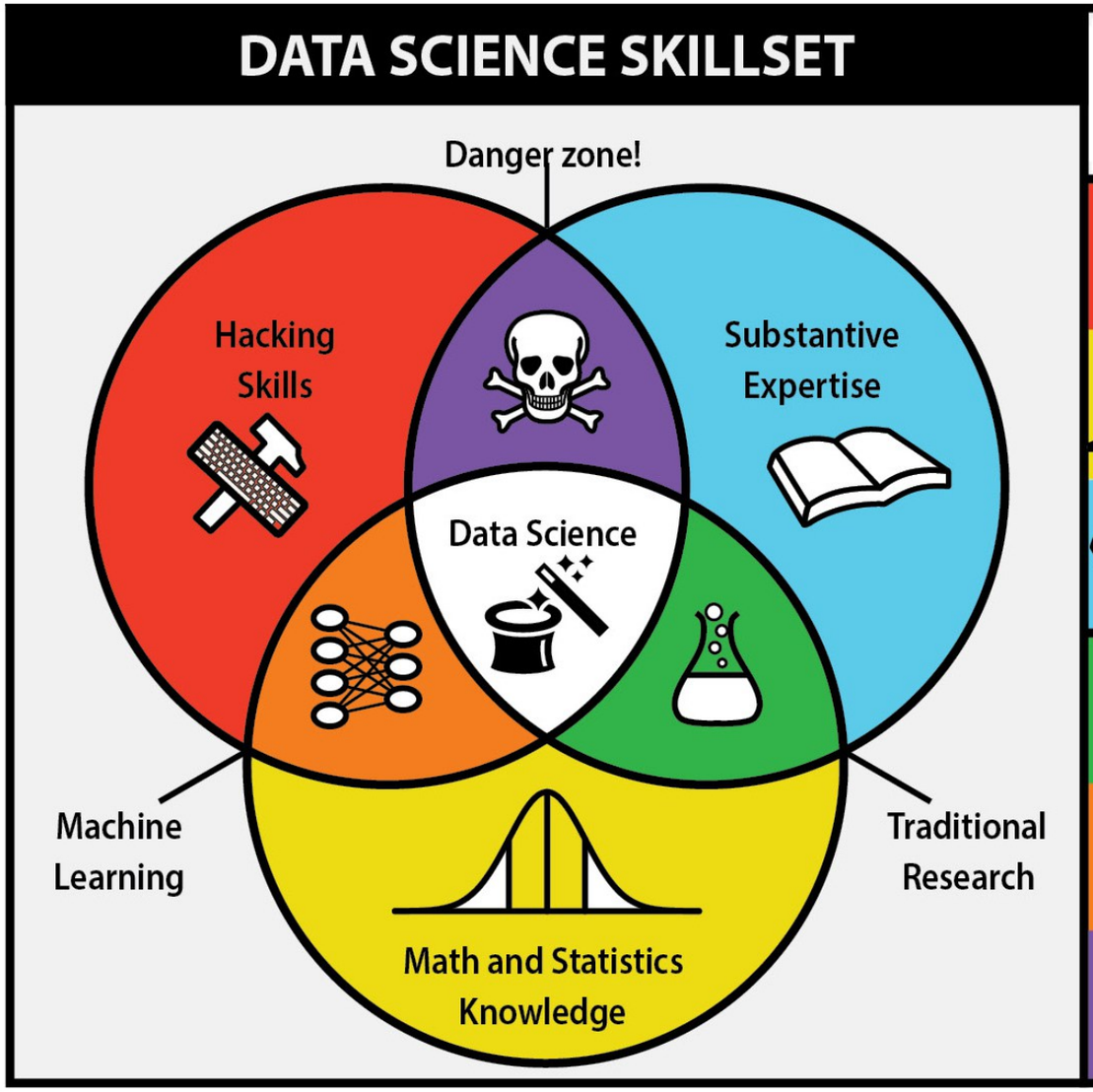
WHAT IS DATA SCIENCE?

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



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

WHAT IS DATA SCIENCE?



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

WHO USES DATA SCIENCE?

NETFLIX

amazon.com

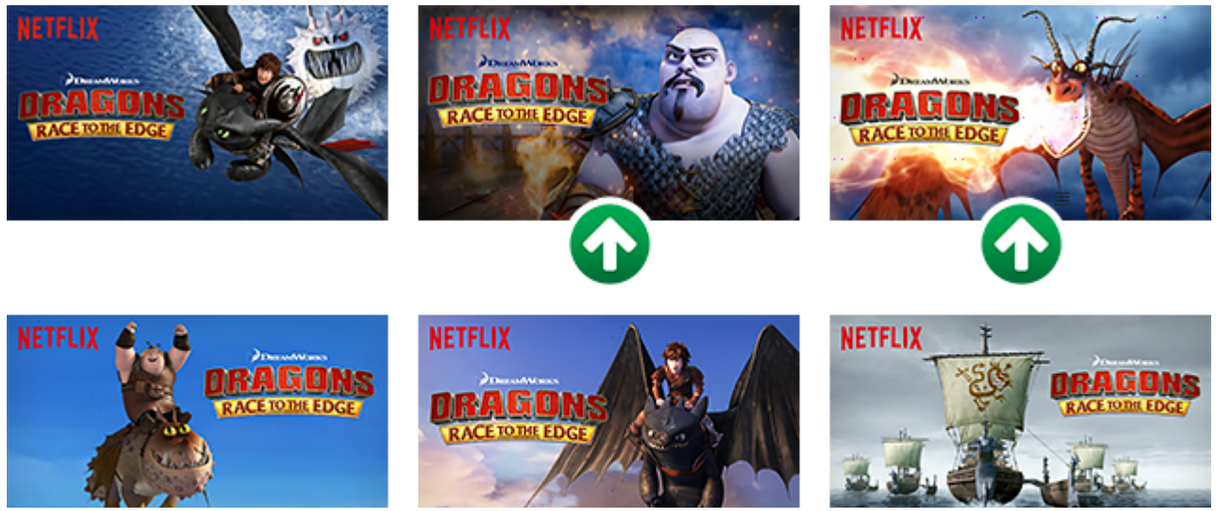
Google

FiveThirtyEight

WHO USES DATA SCIENCE?

The Switch

Red light, green light: New Audis will predict the time until that stoplight turns green Article



Innovations

Netflix reveals what images hook viewers on new shows Article

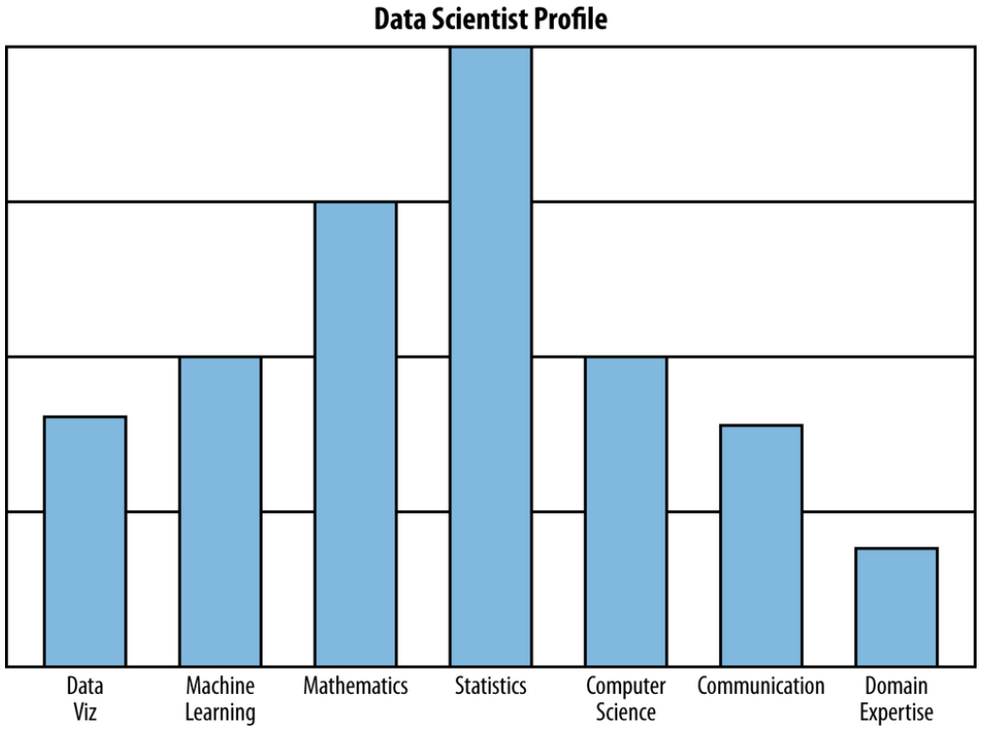
WHO USES DATA SCIENCE?

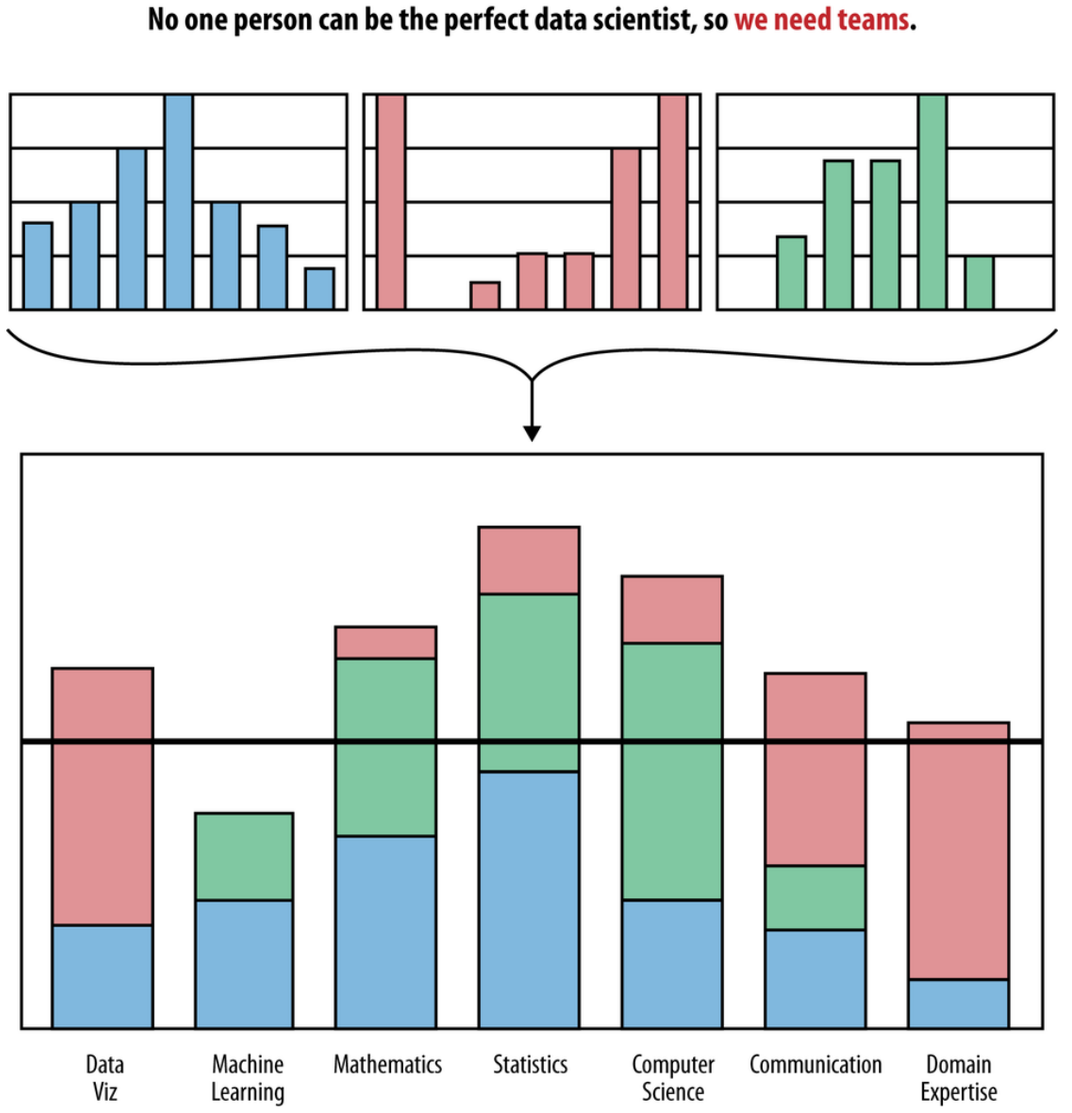
- Can you think of others?

WHAT IS DATA SCIENCE?

Data Science Roles

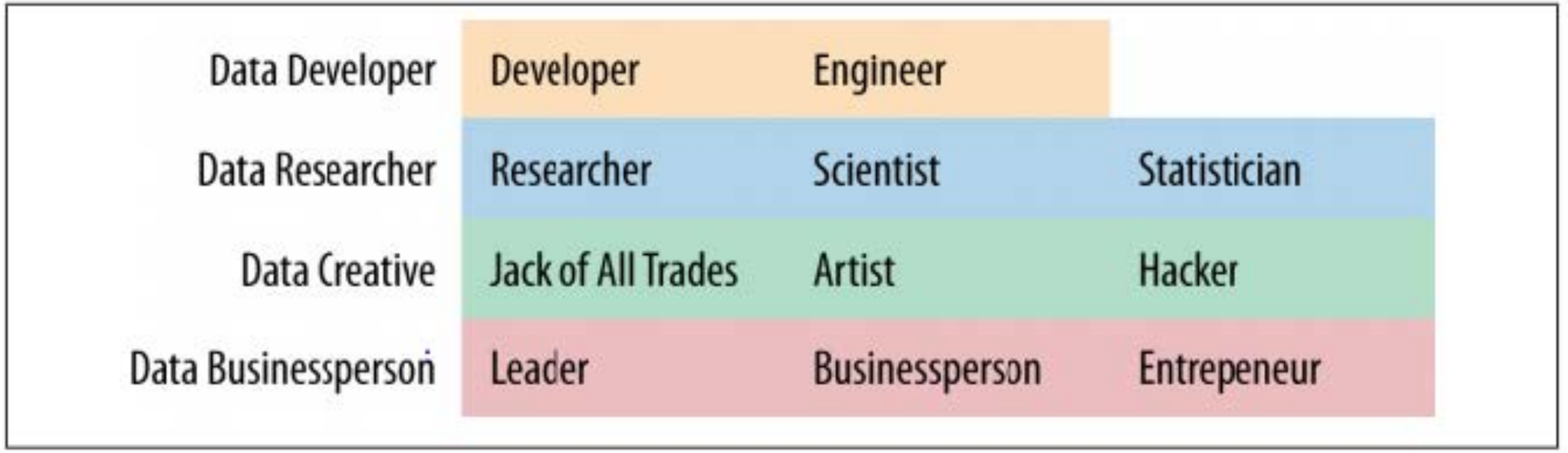
WHAT IS DATA SCIENCE?





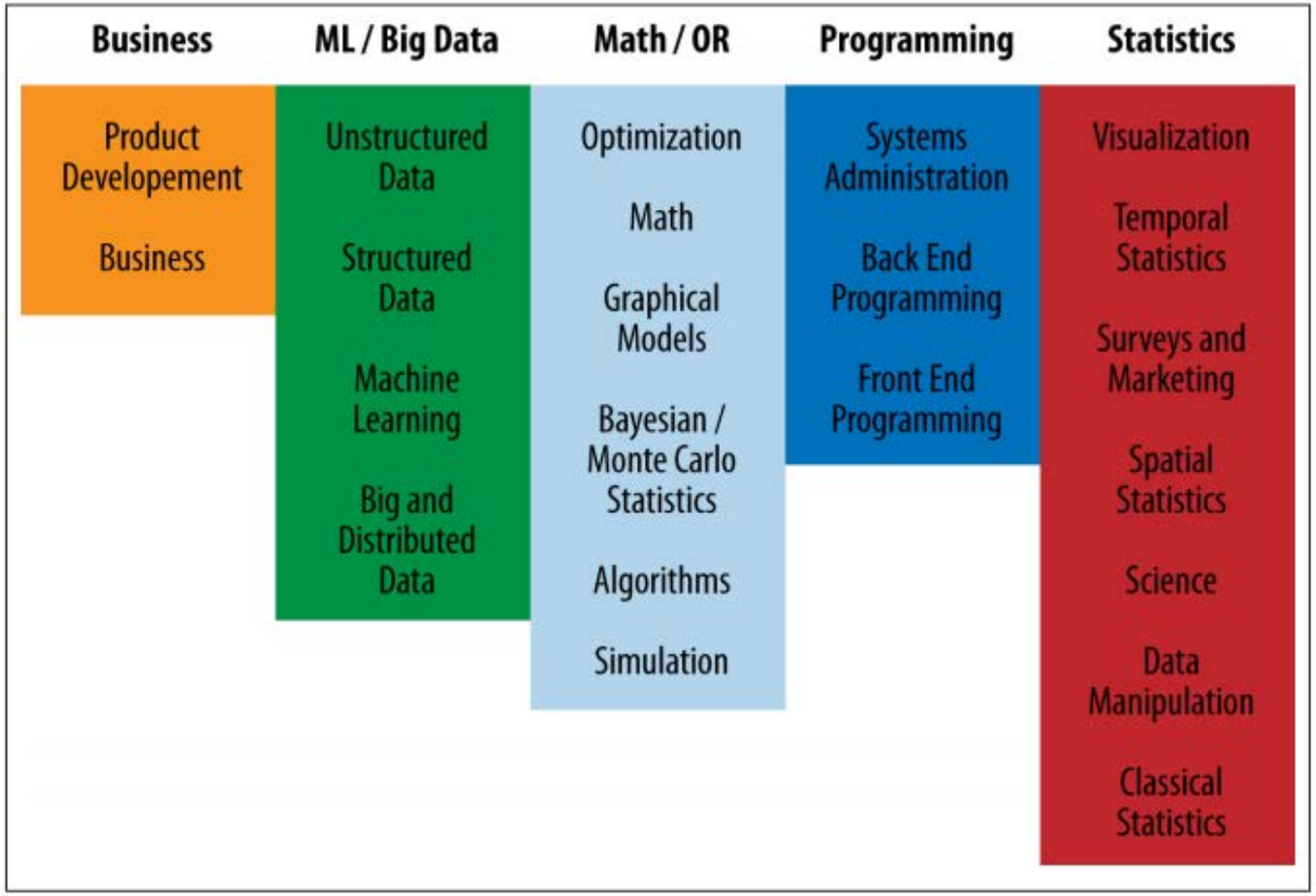
WHAT ARE THE ROLES IN DATA SCIENCE?

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



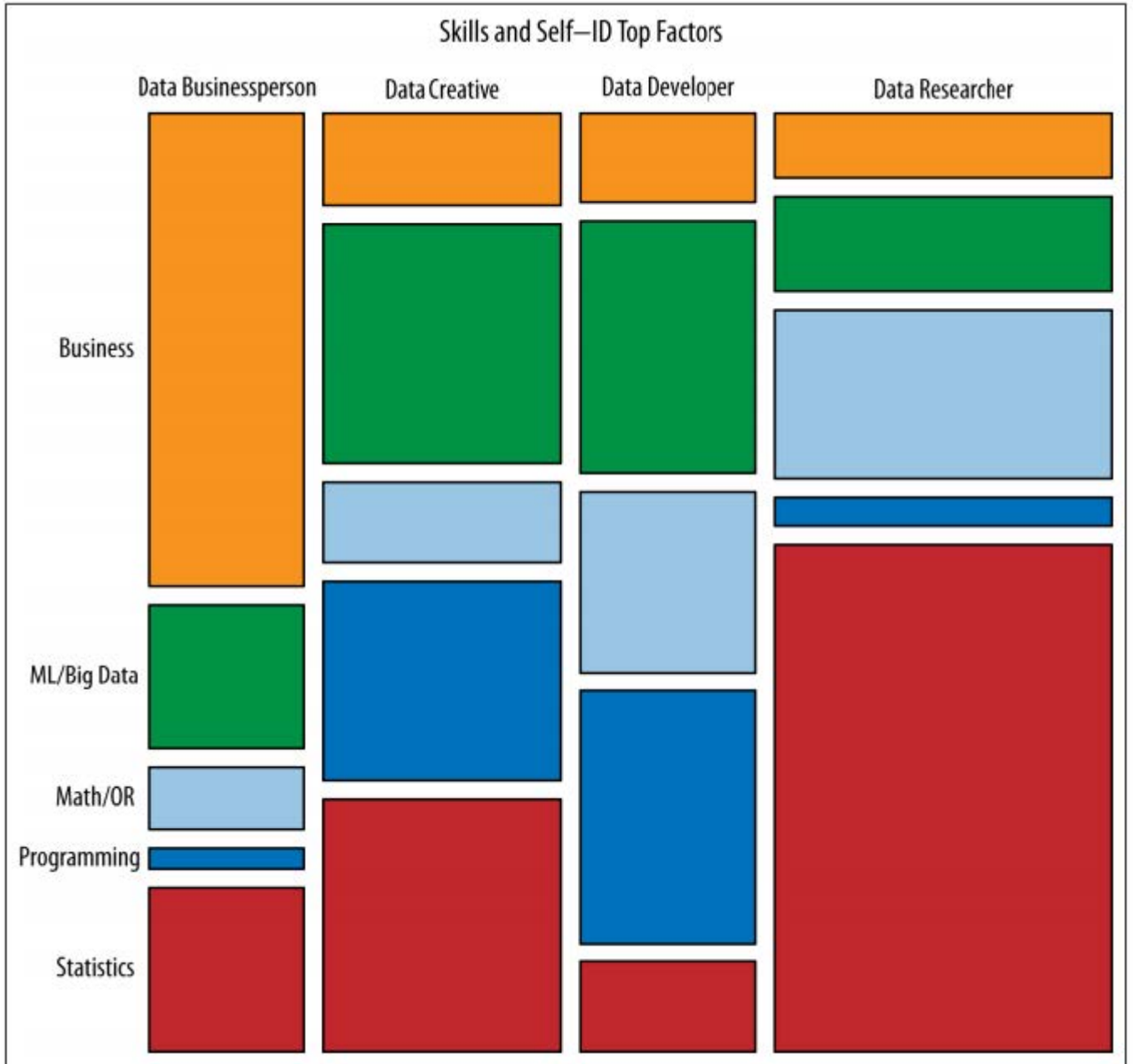
WHAT ARE THE ROLES IN DATA SCIENCE?

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

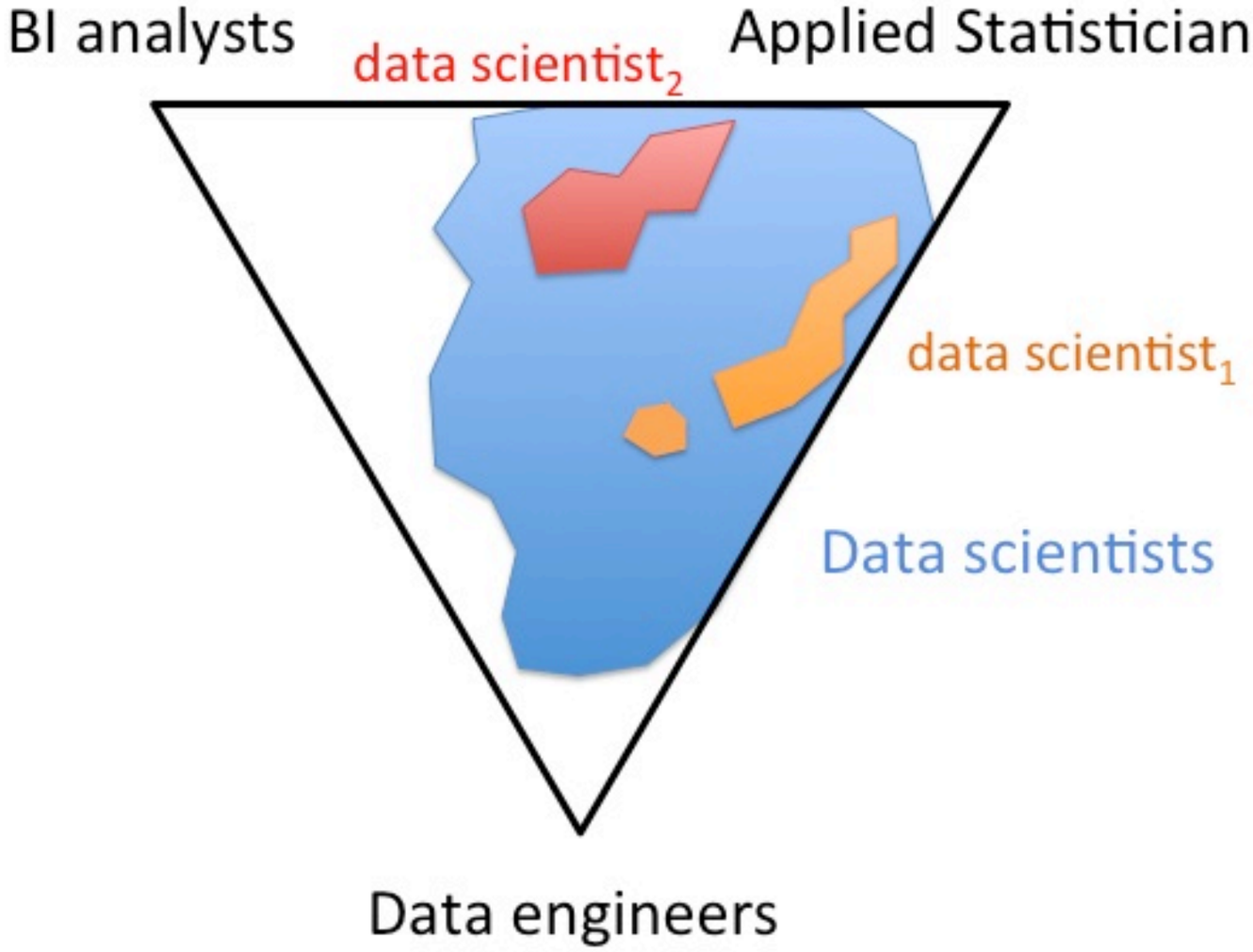


WHAT ARE THE ROLES IN DATA SCIENCE?

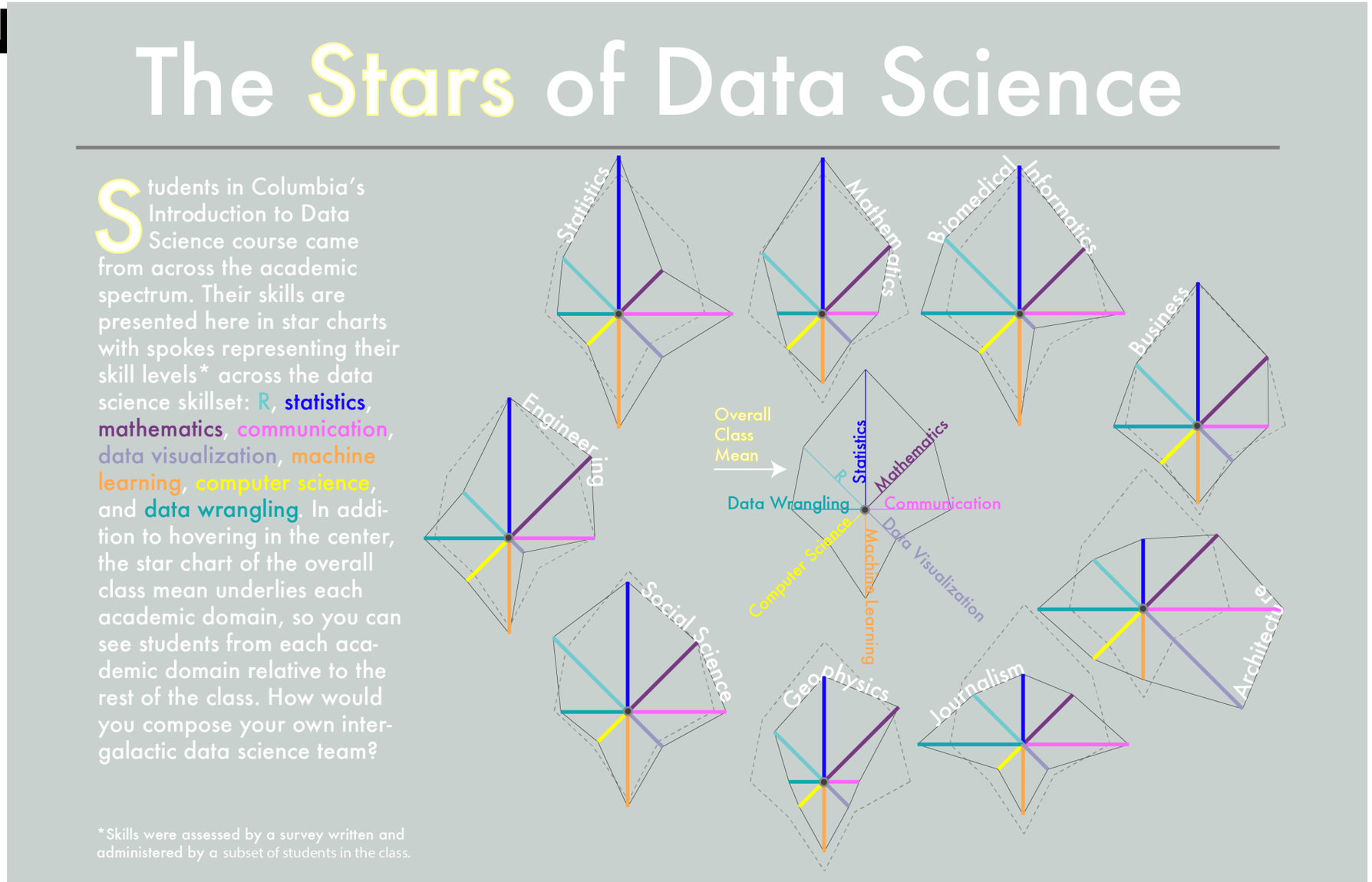
- These roles prioritize different skill sets.
- However, all roles involve some part of each skillset.
- Where are your strengths and weaknesses?



WHAT ARE THE ROLES IN DATA SCIENCE?



WHAT ARE THE ROLES IN DATA SCIEN



THE DATA SCIENCE WORKFLOW

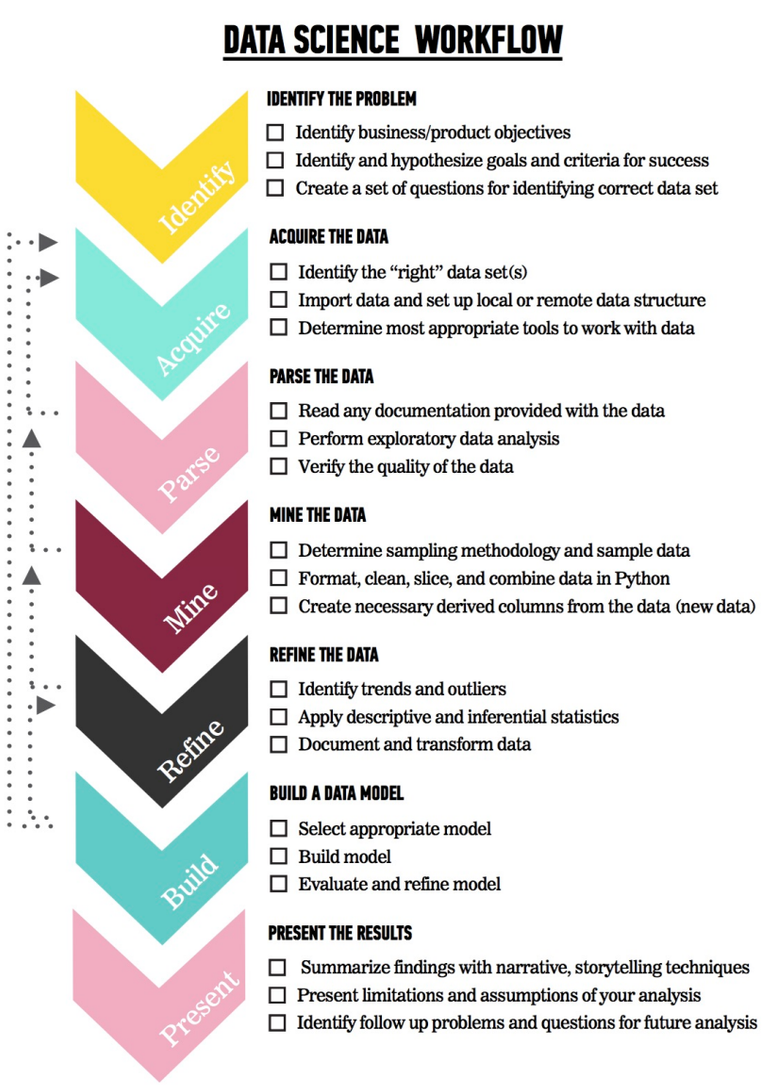
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

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

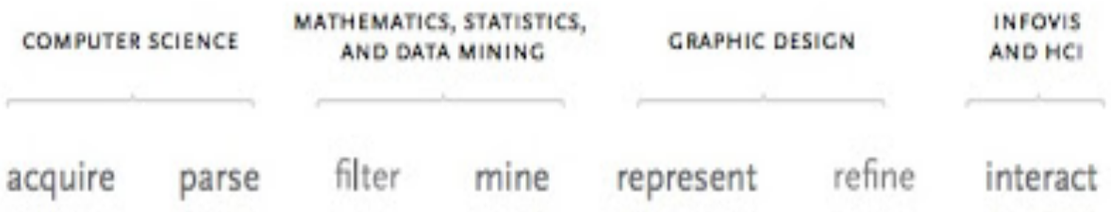
OVERVIEW OF THE DATA SCIENCE WORKFLOW

PRESENT THE RESULTS

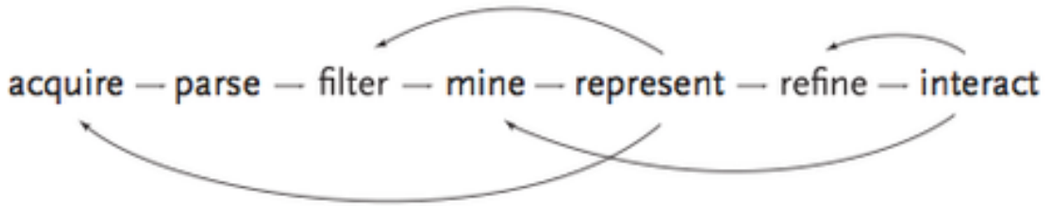
- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

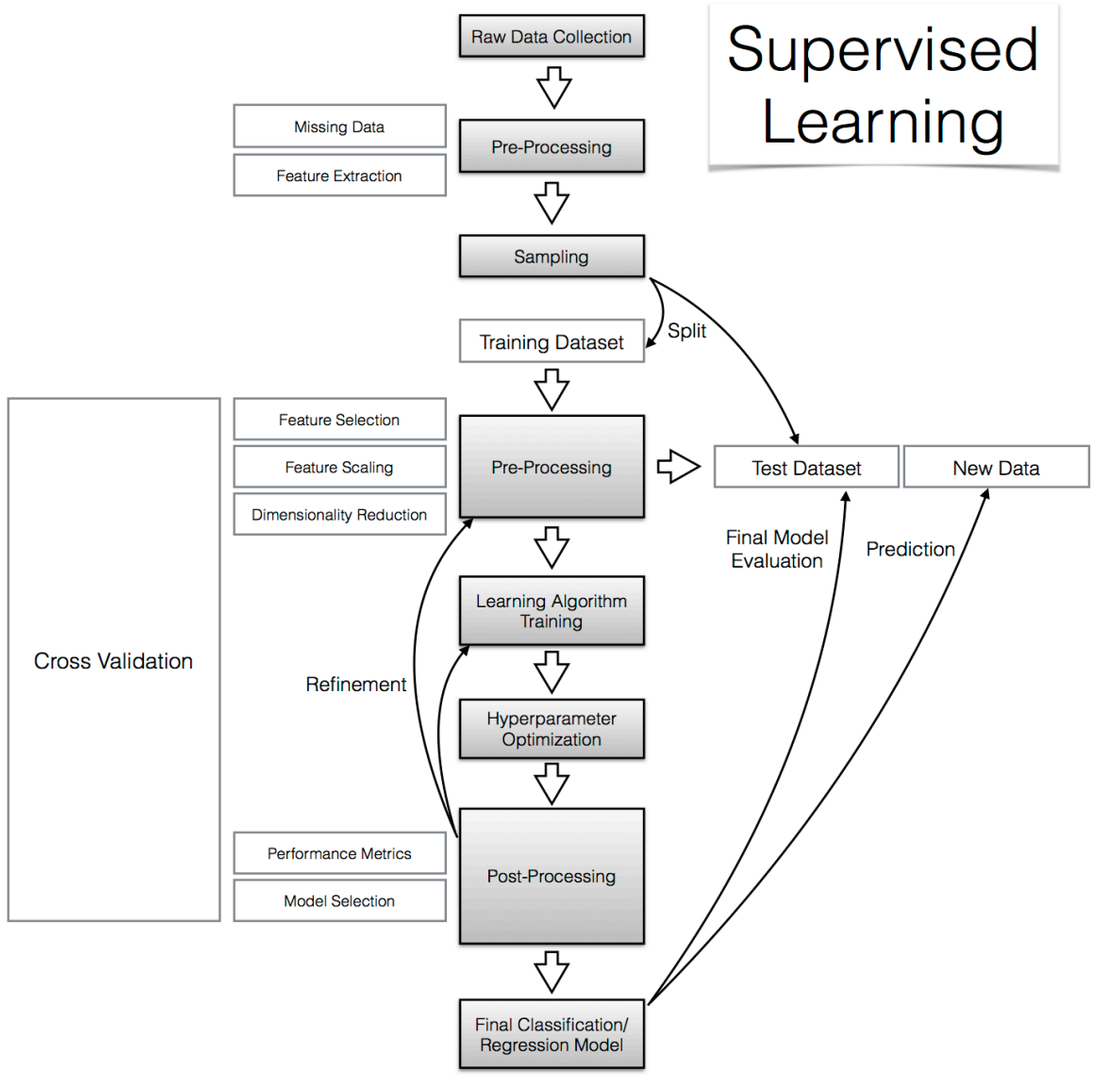
OVERVIEW OF THE DATA SCIENCE WORKFLOW

Interdisciplinary



Recursion





Sebastian Raschka 2014
This work is licensed under a Creative Commons Attribution 4.0 International License.

GUIDED PRACTICE

DATA SCIENCE WORK FLOW EXERCISE

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

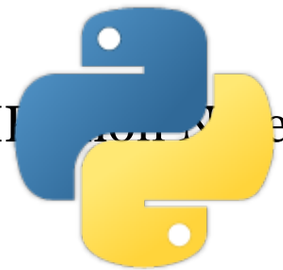
DEMO

ENVIRONMENT SETUP

DEV ENVIRONMENT SETUP

- A Brief Intro to Our Environment

- II ebook Test an



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

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

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET