GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Mart van de Ven | Dickson Kwong | Alex Anzola Jürgenson

Data Scientists, Droste

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

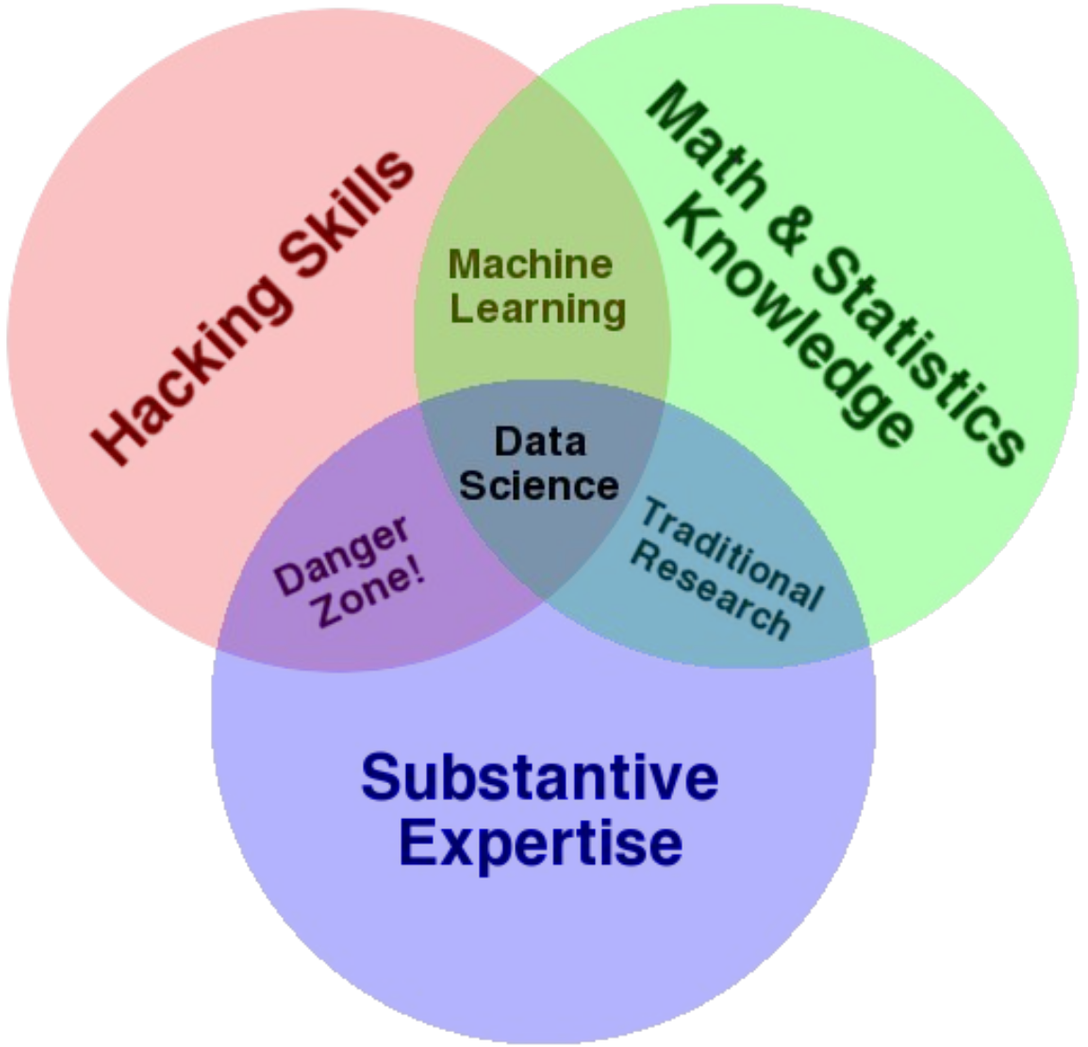
DATA SCIENCE

WELCOME TO GA!

WHAT IS DATA SCIENCE?

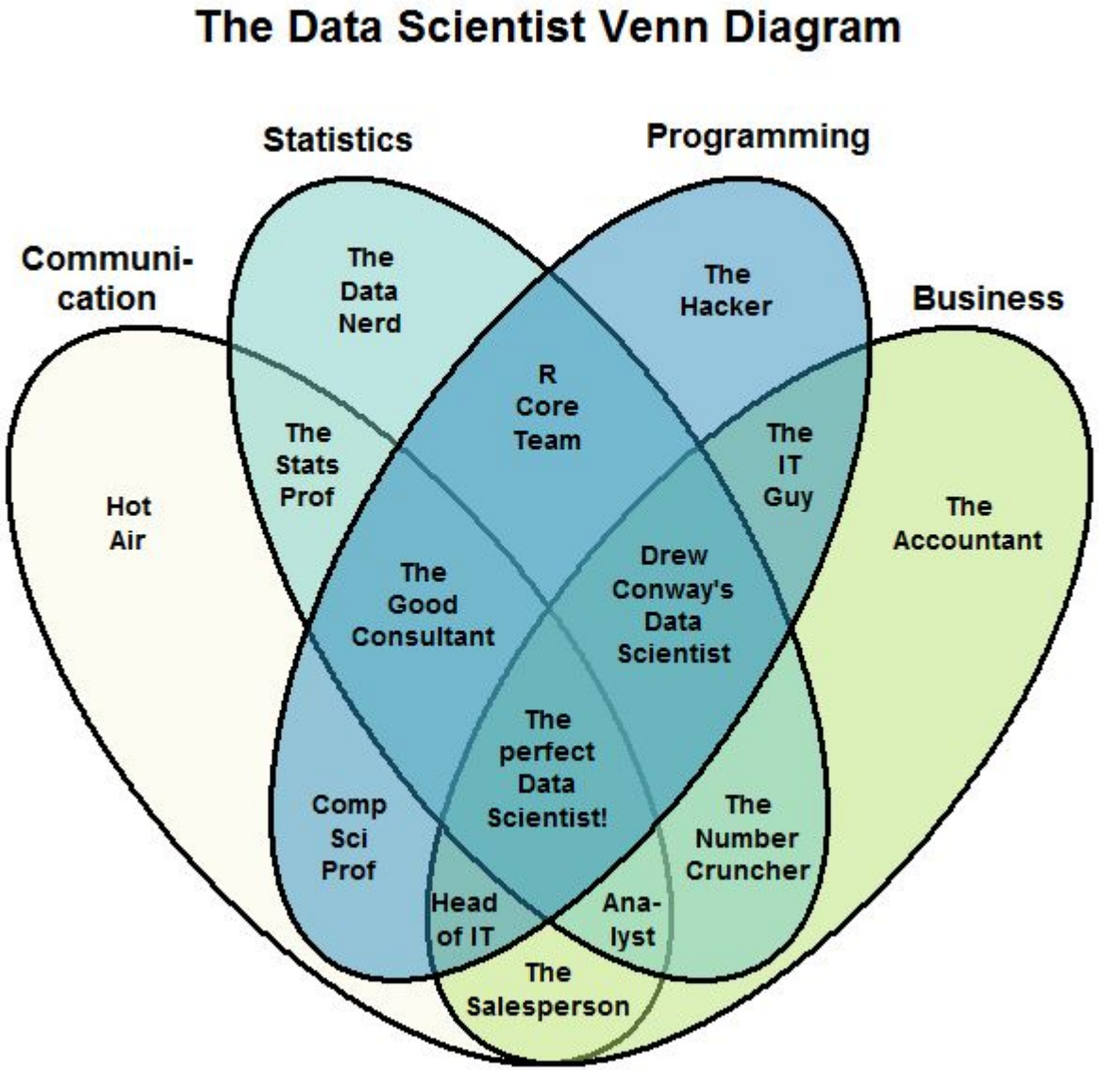
WHAT IS DATA SCIENCE?

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



WHAT IS DATA SCIENCE?

- A developing field with lots definitions of what data science is
- For our purposes, we will take Data Science to be an approach to finding intelligence in data with machine learning methods



WHO USES DATA SCIENCE?

NETFLIX

amazon.com

Google

FiveThirtyEight

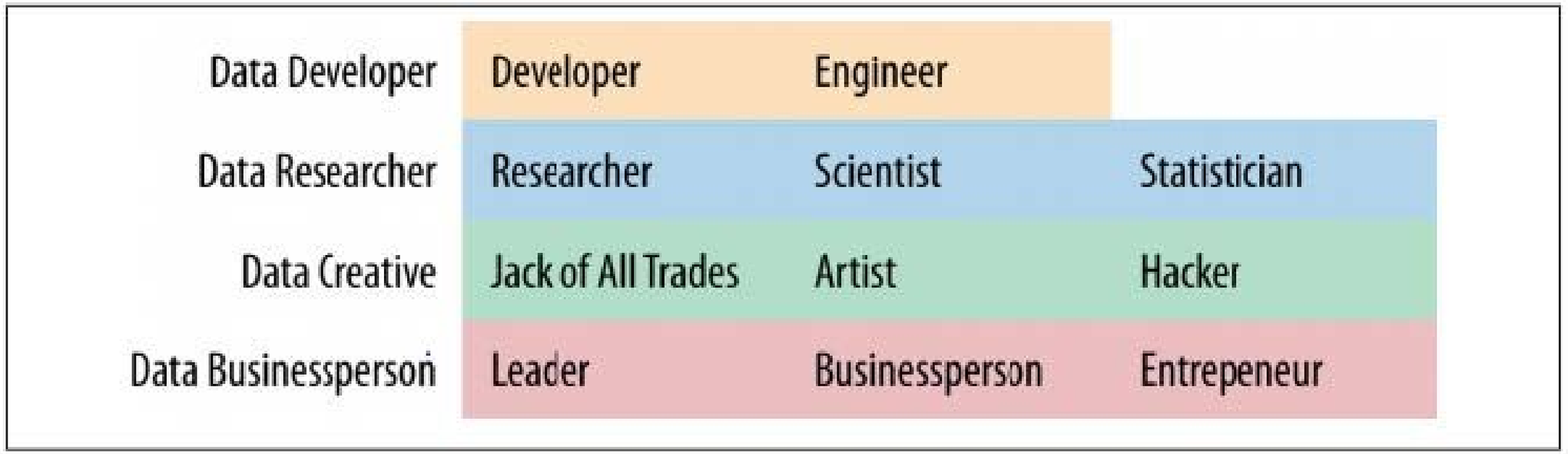
WHO USES DATA SCIENCE?

- Can you think of others?


ima...
021 3 1

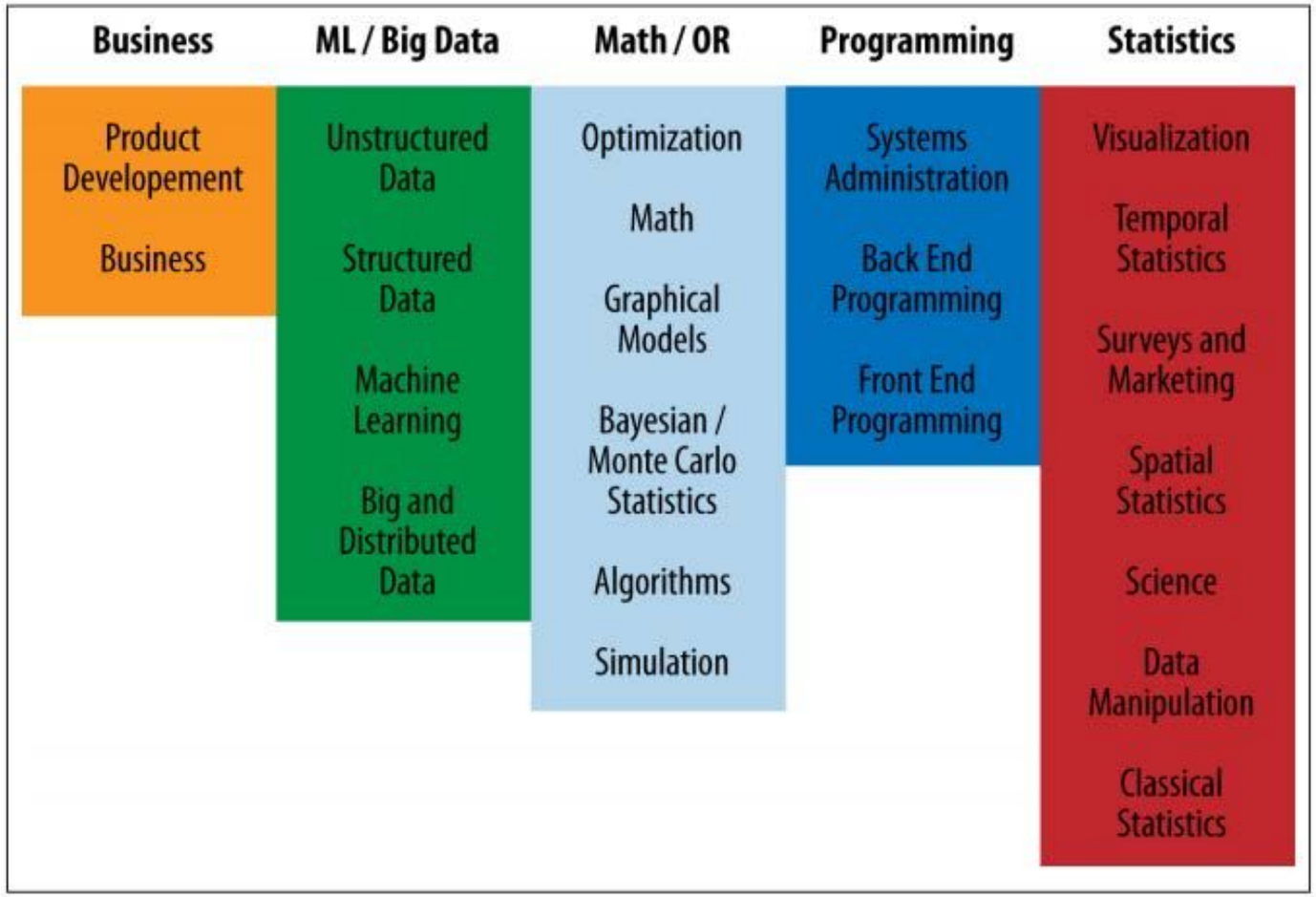
WHAT ARE THE ROLES IN DATA SCIENCE?

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



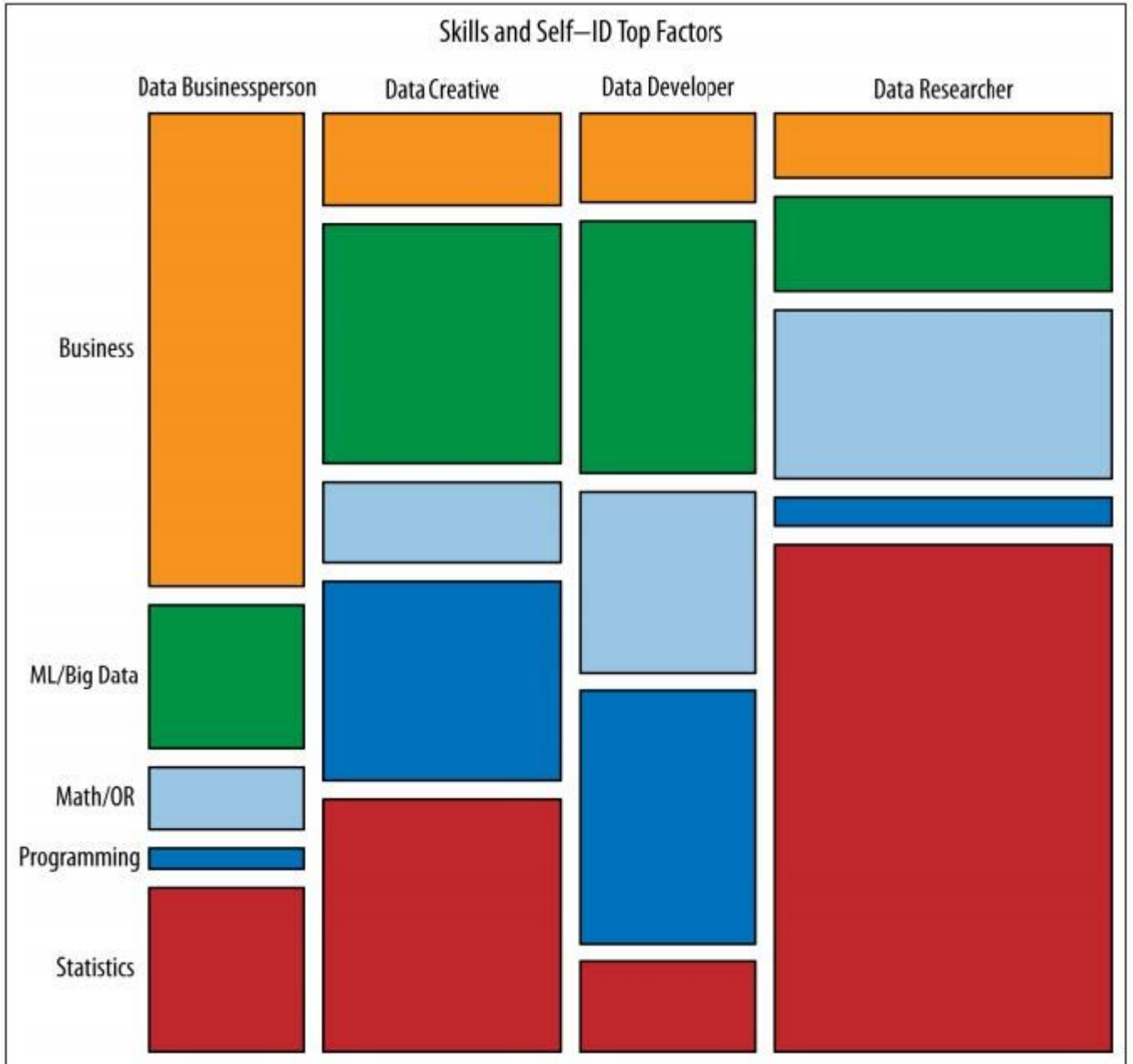
WHAT ARE THE ROLES IN DATA SCIENCE?

- 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

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. According to the table on the next slide, BMI is the _____
      i. Outcome
      ii. Predictor
      iii. Covariate
   c. Draw a normal distribution
   d. True or False: Linear regression is an unsupervised learning algorithm.
   e. What is a hypothesis test?

ACTIVITY: DATA SCIENCE BASELINE QUIZ

EXERCISE

Table 3. Adjusted mean[a] (95% confidence interval) of BMI and serum concentration of metabolic biomarkers in American adults by categories of weekly frequency of fast-food or pizza meals, NHANES 2007–2010

BMI or serum biomarker	*Weekly frequency of fast-food or pizza meals*				P[b]
	0 Time	*1 Time*	*2–3 Times*	*⩾4 Times*	
BMI[c], *kg m*$^{-2}$					
All ($N=8169$)	27.5 (27.1, 27.8)	27.9 (27.6, 28.2)	28.9 (28.4, 29.4)	28.8 (28.3, 29.2)	< 0.0001
Men ($n=4002$)	27.9 (27.4, 28.3)	28.0 (27.6, 28.4)	28.5 (28.0, 29.0)	28.6 (28.2, 29.0)	0.05
Women ($n=4167$)	27.2 (26.8, 27.6)	27.7 (27.3, 28.1)	29.3 (28.6, 29.9)	29.0 (28.1, 29.8)	< 0.0001
Total cholesterol, mg dl^{-1} ($N=8236$)	199 (197, 202)	198 (196, 200)	199 (196, 201)	198 (196, 201)	0.5
HDL-cholesterol[c], *mg dl*$^{-1}$					
All ($n=8236$)	54 (53, 55)	53 (52, 54)	52 (51, 53)	51 (50, 52)	< 0.0001
Men ($n=4042$)	48 (47, 49)	48 (47, 49)	48 (46, 49)	46 (45, 47)	0.003
Women ($n=4194$)	60 (59, 61)	58 (57, 60)	56 (55, 57)	56 (54, 58)	0.001
LDL-cholesterol[d], *mg dl*$^{-1}$					
All ($n=3604$)	113 (111, 116)	117 (113, 120)	113 (110, 116)	114 (110, 118)	0.6
< 50 Years ($n=2151$)	107 (105, 110)	112 (109, 116)	111 (107, 114)	108 (104, 112)	0.8
⩾50 Years ($n=1453$)	123 (118, 129)	126 (121, 131)	118 (113, 123)	129 (122, 137)	0.5
Triglycerides, mg dl^{-1} ($n=3659$)	103 (98, 109)	103 (99, 108)	110 (106, 115)	110 (104, 117)	0.2
Fasting glucose[c], *mg dl*$^{-1}$					
All ($n=3668$)	99 (98, 100)	99 (98, 100)	99 (98, 100)	99 (98, 100)	0.5
Men ($n=1750$)	102 (101, 104)	102 (101, 104)	101 (99, 102)	101 (99, 102)	0.1
Women ($n=1918$)	97 (95, 98)	95 (94, 97)	97 (96, 99)	98 (96, 101)	0.2
Glycohemoglobin, % ($N=8234$)	5.42 (5.39, 5.44)	5.39 (5.36, 5.42)	5.39 (5.36, 5.42)	5.40 (5.37, 5.44)	0.2

Abbreviations: BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NHANES, National Health and Nutrition Examination Surveys. [a]Adjusted means were computed from multiple linear regression models with each biomarker as a continuous dependent variable. All biomarkers (except BMI, total- and HDL-cholesterol) were log-transformed for analysis; therefore, the back-transformed values for LDL-cholesterol, triglycerides, fasting glucose and glycohemoglobin are geometric means and their 95% confidence intervals. Independent variables included: frequency of fast-food meals (0, 1, 2–3 and ⩾4 times), age (20–39, 40–59 and ⩾60), sex, race/ethnicity (non-Hispanic white, non-Hispanic black, Mexican-American and other), poverty income ratio (⩽1.3, >1.3–3.5, ⩾3.5 and unknown), years of education (< 12, 12, some college and ⩾college), serum cotinine (continuous), hours of fasting before phlebotomy, (continuous), physical activity (none, tertiles of MET minutes/week), alcohol-drinking status (never drinker, former drinker, current drinker and unknown). *N* refers to observations used in the regression model for each biomarker. [b]*P*-value for the Sattherwaite-adjusted F test for frequency of fast-food meals as a continuous variable. [c]Significant interaction of fast-food meals with sex ($P_{interaction} < 0.05$; thus, the results are stratified by sex [d]Significan interaction of frequency of fast-food meals with age ($P_{interaction} < 0.05$); thus, the results are stratified by age categories.

THE DATA SCIENCE WORKFLOW

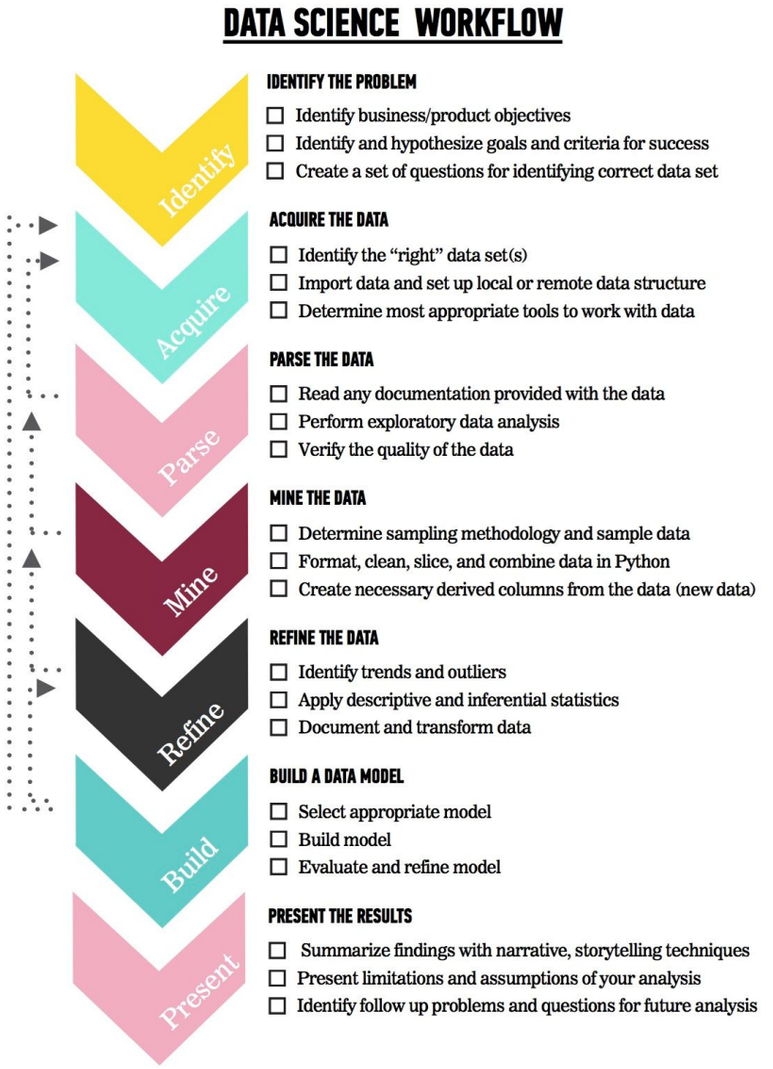
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

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 (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 GitHub Desktop
  - Install Python 2.7 with 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