GENERAL ASSEMBLY

# WELCOME TO DATA SCIENCE

*Trevor Lindsay*

*Data Scientist, Facebook*

# 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.

# YOUR INSTRUCTOR


**Trevor Lindsay**

Lead Instructor
Data Scientist, Facebook
trevor.lindsay@yahoo.com


# STUDENT SERVICES


**Matt Jones**

Student Experience Associate

studentservicesSF@ga.co

Course Logistics + Campus Questions

# OTHERS YOU MAY SEE

**RAY HSIA**
Instructor Manager

**NIÑA PINEDA**
Front Lines Lead

**VANESSA OHTA**
Senior
Instruction Manager

# LET'S GET TO KNOW EACH OTHER

- Introduce yourself (name, work experience, where in the Bay Area you live, etc.)

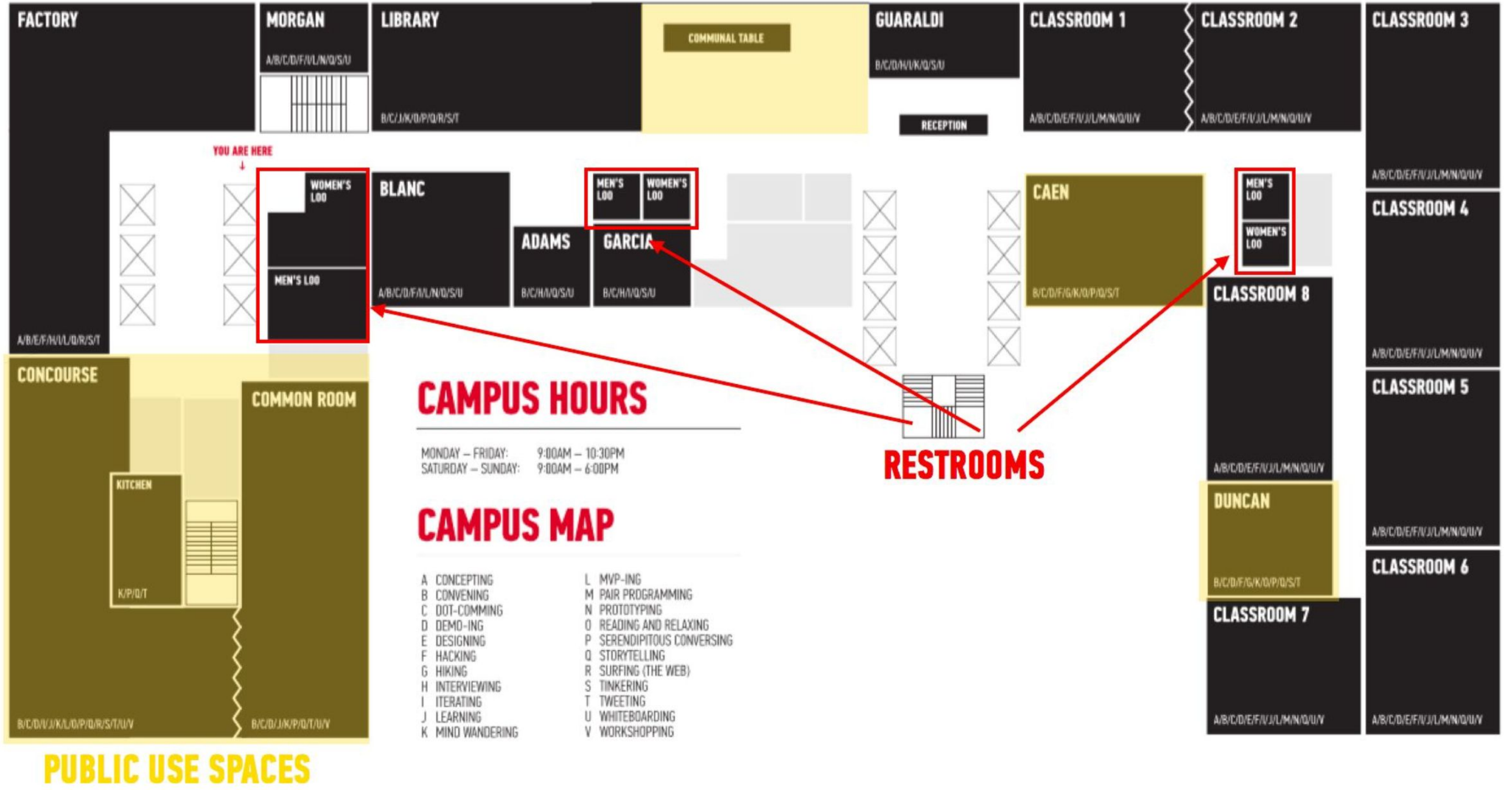
- Tell us why you are taking this course and what you hope to get out of it

- Open your preferred music app (Spotify, Apple Music, etc.) and share the last song you listened to

# COME WORK ON CAMPUS!

- Hours
  - 8am - 10pm, Monday to Friday
  - 9am - 6pm, Saturday and Sunday

- Reception can help with:
  - Loaner equipment
  - Lost and found
  - Free coffee and tea


FACTORY
A/B/E/F/H/I/L/Q/R/S/T
MORGAN
A/B/C/D/F/I/L/N/Q/S/U
LIBRARY
B/C/J/K/O/P/Q/R/S/T
COMMUNAL TABLE
GUARALDI
B/C/D/H/I/K/Q/S/U
RECEPTION
CLASSROOM 1
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
CLASSROOM 2
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
CLASSROOM 3
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
YOU ARE HERE
WOMEN'S LOO
MEN'S LOO
BLANC
A/B/C/D/F/I/L/N/Q/S/U
ADAMS
B/C/H/I/Q/S/U
MEN'S LOO
WOMEN'S LOO
GARCIA
B/C/H/I/Q/S/U
CAEN
B/C/D/F/G/K/O/P/Q/S/T
MEN'S LOO
WOMEN'S LOO
CLASSROOM 4
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
CLASSROOM 5
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
CLASSROOM 6
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
CONCOURSE
KITCHEN
K/P/Q/T
B/C/D/I/J/K/L/O/P/Q/R/S/T/U/V
COMMON ROOM
B/C/D/J/K/P/Q/T/U/V
CAMPUS HOURS
MONDAY – FRIDAY: 9:00AM – 10:30PM
SATURDAY – SUNDAY: 9:00AM – 6:00PM
CAMPUS MAP
A CONCEPTING
B CONVENING
C DOT-COMMING
D DEMO-ING
E DESIGNING
F HACKING
G HIKING
H INTERVIEWING
I ITERATING
J LEARNING
K MIND WANDERING
L MVP-ING
M PAIR PROGRAMMING
N PROTOTYPING
O READING AND RELAXING
P SERENDIPITOUS CONVERSING
Q STORYTELLING
R SURFING (THE WEB)
S TINKERING
T TWEETING
U WHITEBOARDING
V WORKSHOPPING
RESTROOMS
CLASSROOM 8
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
DUNCAN
B/C/D/F/G/K/O/P/Q/S/T
CLASSROOM 7
A/B/C/D/E/F/I/J/L/M/N/Q/U/V
PUBLIC USE SPACES

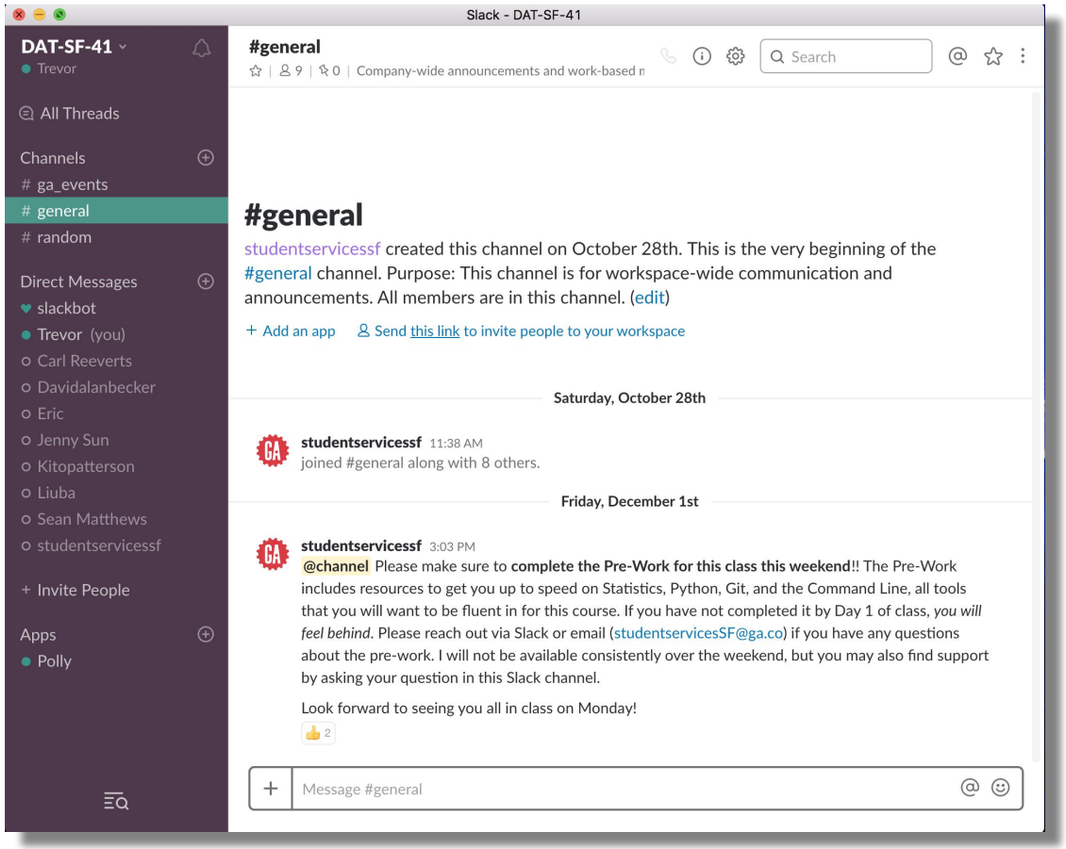
DATA SCIENCE

# Student Experience

# SLACK


## dat-sf-41.slack.com

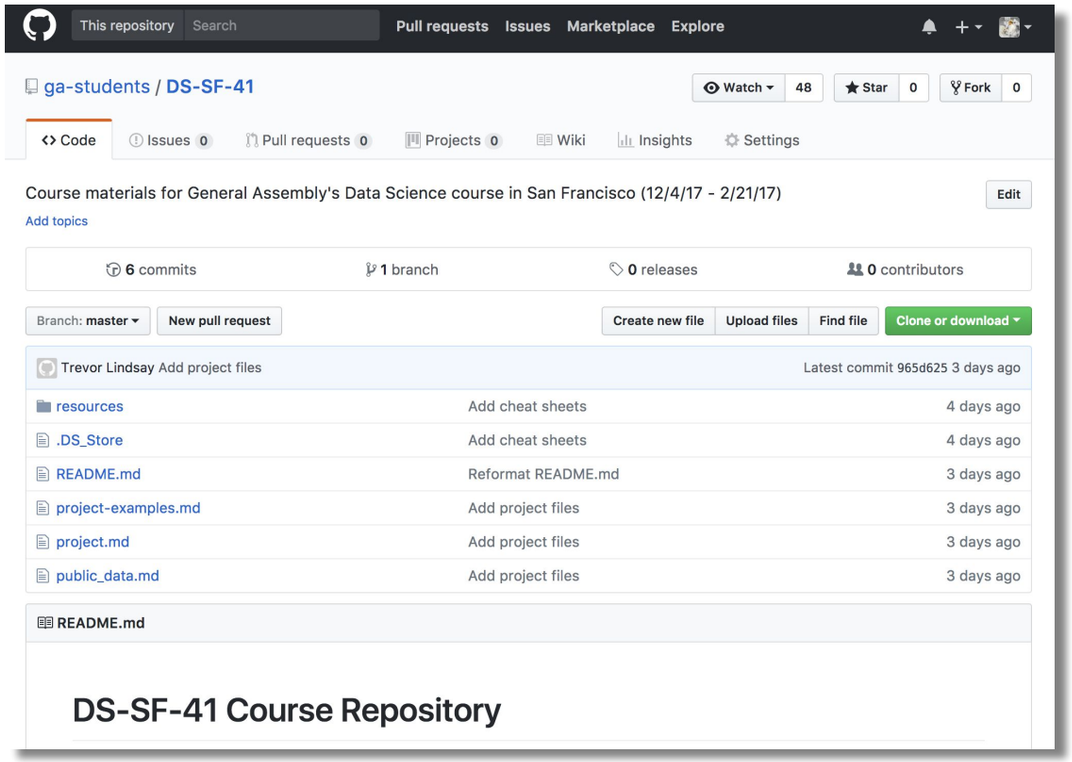
All course communication with each other and instructors will happen here



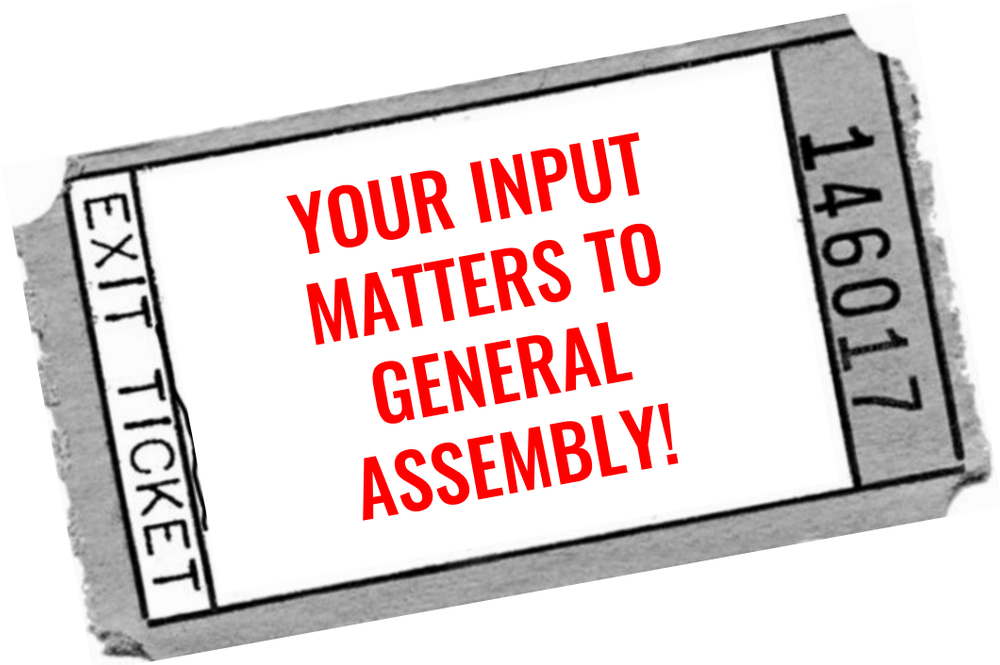
# GITHUB


## github.com/ga-students/DS-SF-41

All the course resources (sample code, assignments, slides) you need will be stored here.



# EXIT TICKETS




## https://tinyurl.com/dat-sf-41

At the end of each class, please fill out an exit ticket to provide feedback on your experience

# CLASSROOM CULTURE


Let’s all agree to:

- Treat each other with respect
- Avoid distractions during class
- Actively participate
- Arrive on time

# INSTRUCTOR PHILOSOPHY

- Seek an optimal pace
- Facilitate healthy and active learning environment
- Involve everyone
- Communicate early and often
  - Feedback should be a busy two-way street

# CONTENT PHILOSOPHY

- Alignment with industry
- Understand foundational principles of theories and ideas
- Balance depth with breadth
- Engaging and interesting data
  - Examples of real world data science
- Course project

# KEYS TO SUCCESS

- Effort > prior knowledge
- Ask questions **early** and **often**
- Keep track of what you've learned and overall progress
- Use each other
- Be patient
- Time management

# CLASS STRUCTURE

- Lecture + Theory: Introduction to material
  - 45 min - 1 hr
- Interactive Coding: Demonstrate material
  - 1 hr
- Class Work: Practice material on real data
  - 30 min - 1 hr
- 5-10 min breaks to digest the material and reset

# CLASS STRUCTURE (CONT.)

- All of the course material is in the Github repo
- Almost all of the classes take place in Jupyter Notebooks
- Office hours from 5:30pm - 6:30pm
- Assignments (required for graduation)
  - 2 homeworks
  - 1 final project
  - 4 small unit projects (optional)
- Use Slack to communicate with classmates and me
- If you're late or have to miss class, please inform me at your earliest convenience

# GA GRADUATION REQUIREMENTS

DATA SCIENCE

# PRE-WORK

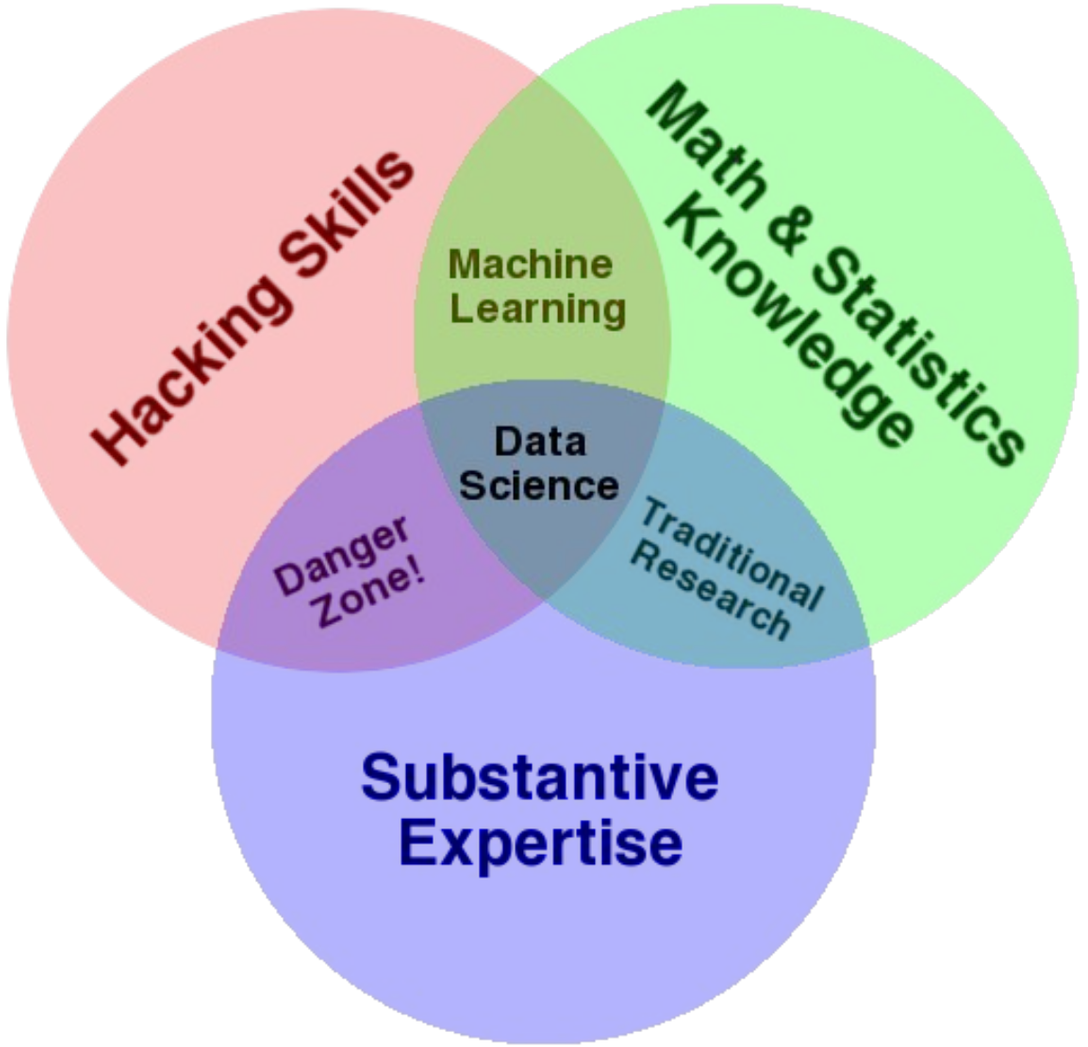
# PRE-WORK REVIEW

- 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

# 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 IS A DATA SCIENTIST?

## Data Scientist:

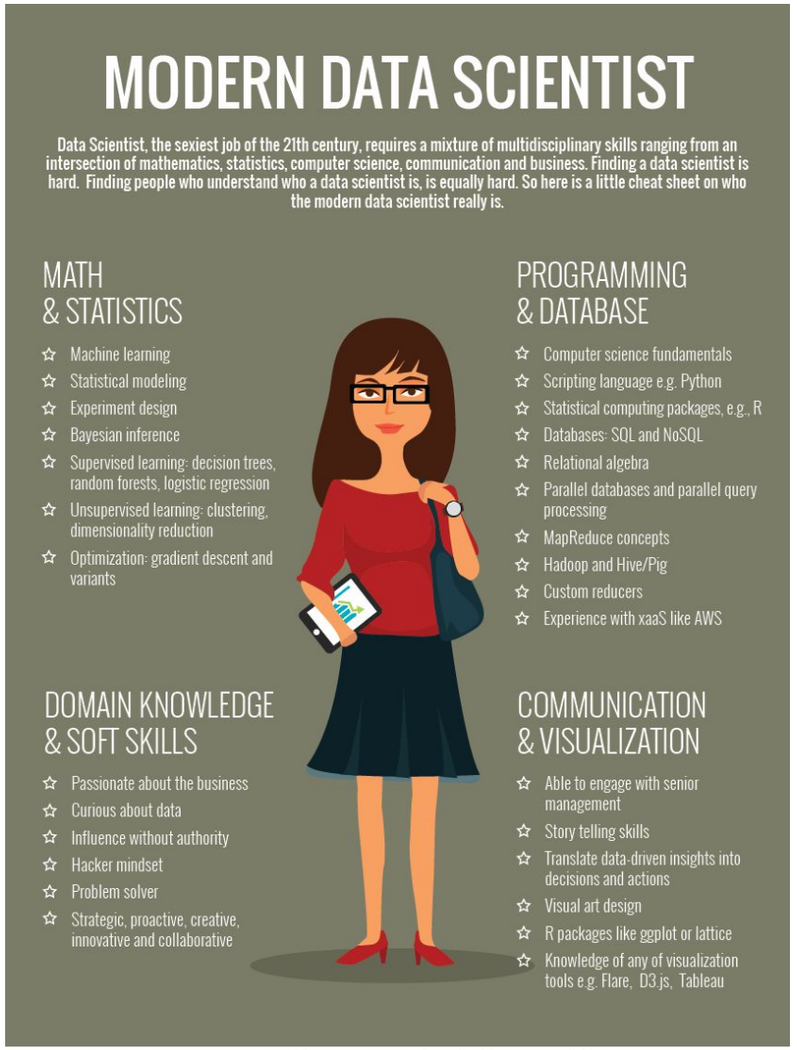
### *The Sexiest Job of the 21st Century*

***Meet the people who can coax treasure out of messy, unstructured data.***

*by Thomas H. Davenport and D.J. Patil*

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."





# WHO IS A DATA SCIENTIST?

"...unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data."
*- DJ Patil, Former U.S. Chief Data Scientist*

"...someone who blends, math, algorithms, and an understanding of human behavior with the ability to hack systems together to get answers to interesting human questions from data."
*- Hilary Mason, Data Scientist in Residence at Accel*

"...better at statistics than any software engineer and better at software engineering than any statistician."
*- Unknown*

# DATA SCIENCE IS EVERYWHERE

NETFLIX

amazon.com

FiveThirtyEight


Google

Walmart

Save money. Live better.

# FACEBOOK DATA SCIENTIST

- Apply your expertise in quantitative analysis, data mining, and the presentation of data to see beyond the numbers and understand how our users interact with both our consumer and business products
- Partner with Product and Engineering teams to solve problems and identify trends and opportunities
- Inform, influence, support, and execute our product decisions and product launches
- Work across four main areas:
  - Product Operations
  - Exploratory Data Analysis
  - Product Leadership
  - Data Infrastructure

# GOOGLE DATA SCIENTIST

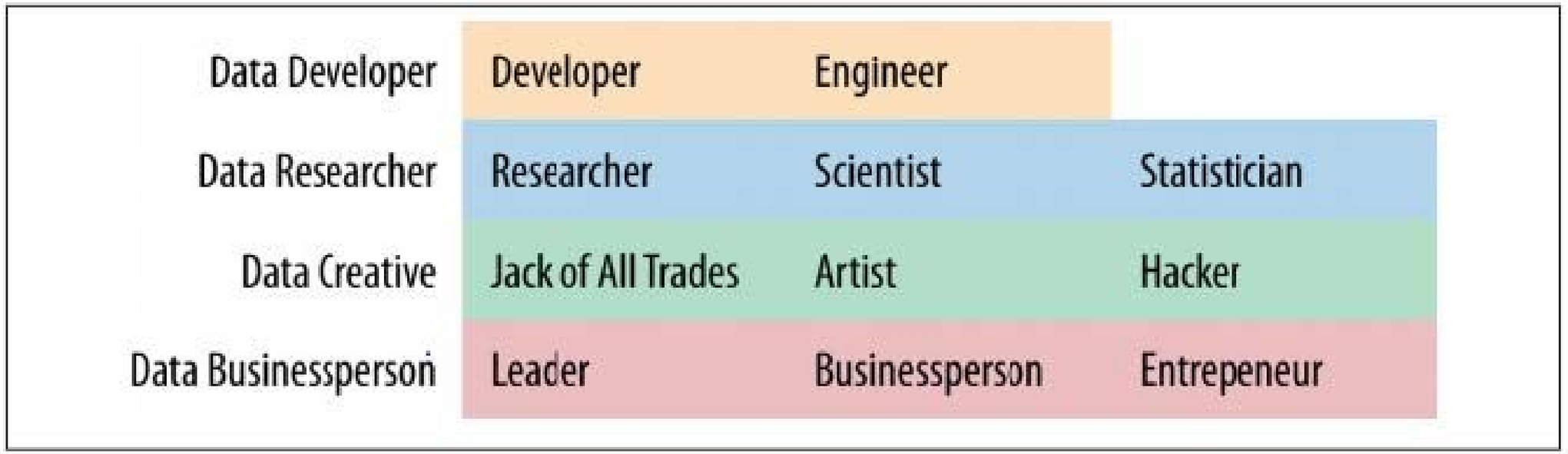
- Evaluate and improve Google's products by collaborating with a multi-disciplinary team of engineers and analysts on a wide range of problems
- Bring analytical rigor and statistical methods to the challenges of measuring quality, improving consumer products, and understanding the behavior of end-users, advertisers, and publishers
- Work with large, complex data sets. Solve difficult, non-routine analysis problems, applying advanced analytical methods as needed.
- Conduct end-to-end analysis that includes data gathering and requirements specification, processing, analysis, ongoing deliverables, and presentations.

# UBER DATA SCIENTIST

- Familiarity with technical tools for analysis - Python (with Pandas, etc.), R, SQL
- Programming chops - demonstrable familiarity (work experience, Github account) with programming concepts. Python skills and previous software engineering background a plus
- Research mindst - ability to structure a project from idea to experimentation to prototype to implementation
- Driven and focused self-starters, great communicators, amazing follow-through - you passionately pursue your work and love the responsibility of being individually empowered
- A preference for quality over quantity - you get the math right and aspire to build the right solution; you like a team that holds each other to a high bar

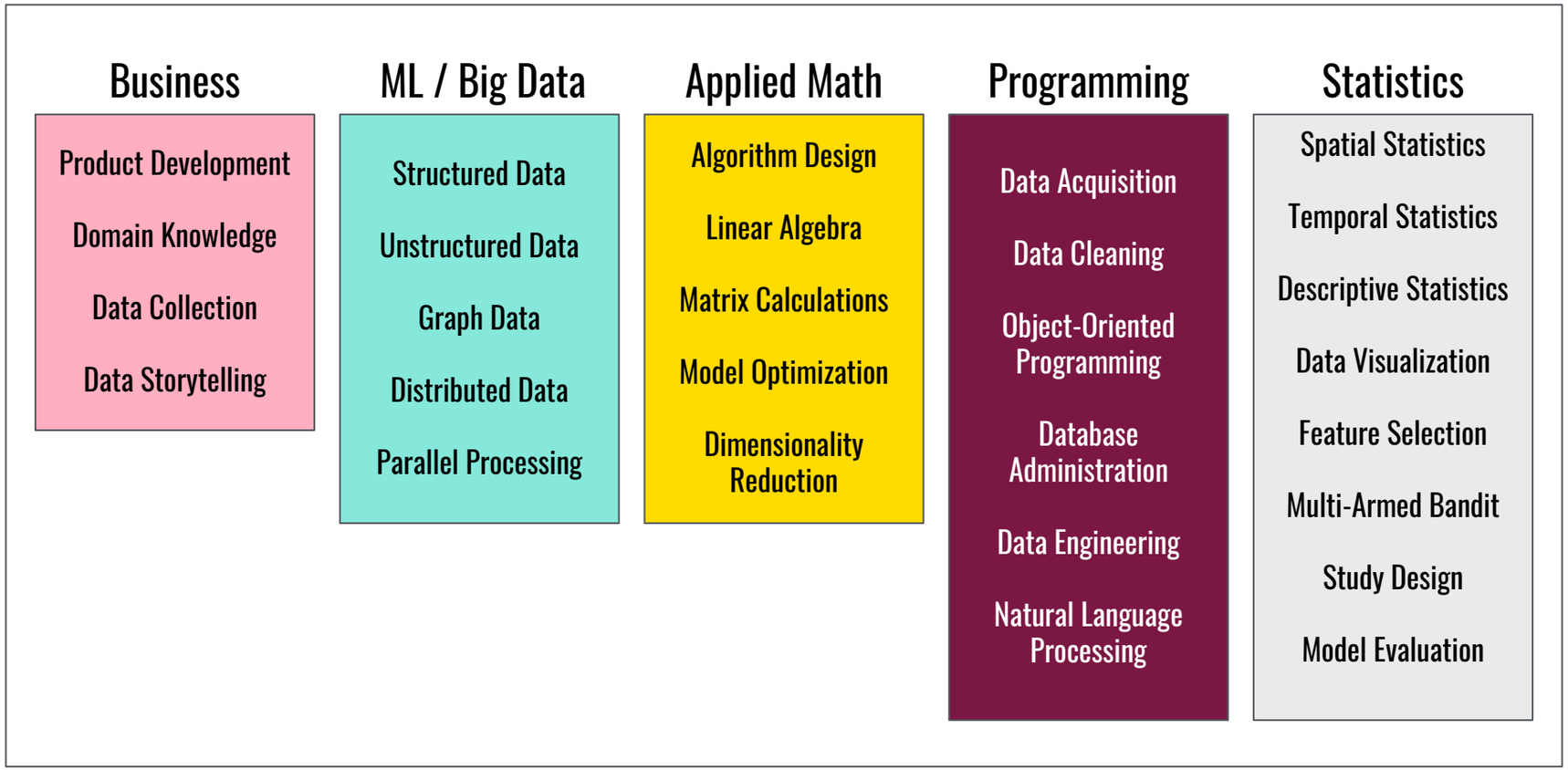
# WHAT ARE THE ROLES IN DATA SCIENCE?

- 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.



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

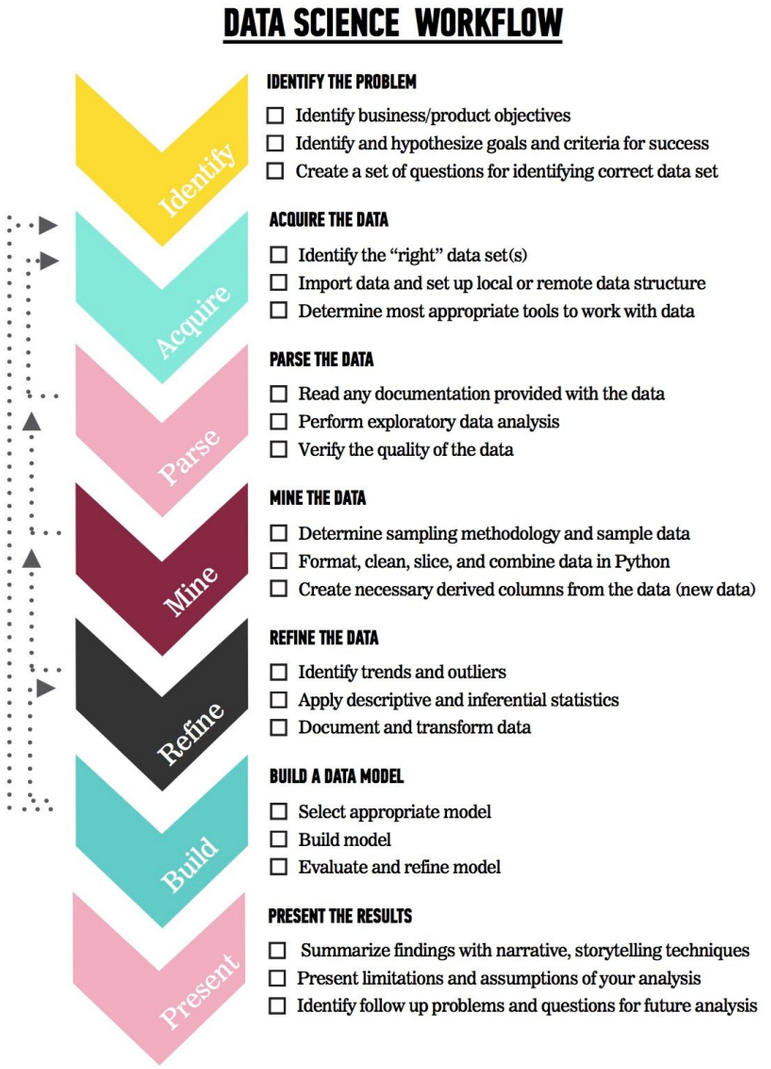
# OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

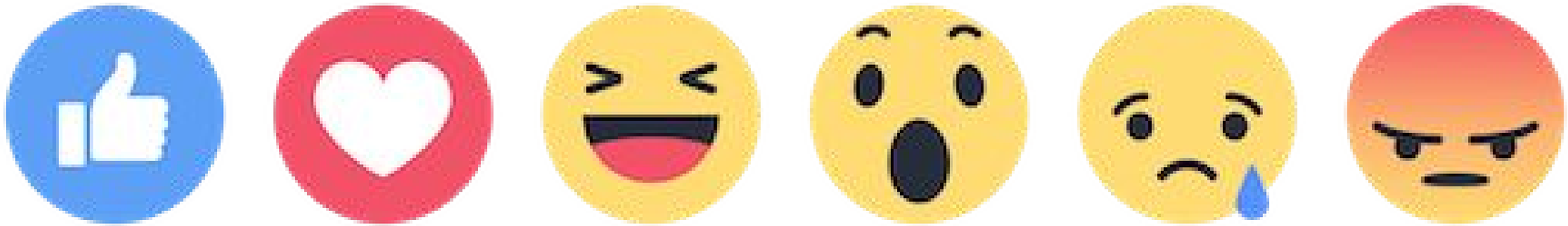
- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

# EXAMPLE #1: PREDICTING PAYMENTS FRAUD

- **Problem:** Some advertisers attempt to use hacked accounts or stolen credit cards to run bad ads on Facebook
- **Goal:** Find the bad actors before their ads are delivered to users

# EXAMPLE #2: LAUNCHING A NEW PRODUCT FEATURE

- **Problem:** Users want more ways than just clicking “like” to respond to content in their news feed
- **Goal:** Build and test additional “reactions”



GUIDED PRACTICE

# DATA SCIENCE WORK FLOW

# ACTIVITY: DATA SCIENCE WORKFLOW


## DIRECTIONS (25 minutes)

1. Divide into pairs, 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

# EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

## 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?