Come in We're LEARNING

WELCOME TO DATA SCIENTIST

Please write your name on your whiteboard and say hello to your new classmates.

Wi-fi: GA-Guest
pw: yellowpencil

Your Instructional Team

Zack Peterson

Lead Instructor
Data Scientist,
Wells Fargo

George McIntire

Mario Carrillo

Student Services at studentservicesSF@ga.co

Course logistics

- Access to tools
- Feedback about the course
- Enrollment and finances
- Graduation certificates

Campus questions

- GA Facilities
- GA events outside of class
- Discounts for other courses

Others you may see

EMILY PEEK

Instructor Coach

JESSICA REPAS

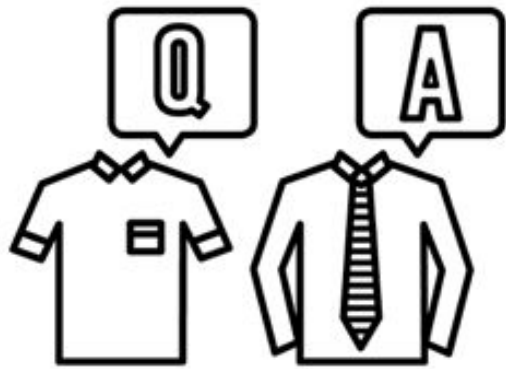Front Lines Lead

TJ THOMANDER

Education Programs
Producer

VANESSA OHTA

Education Programs
Manager

Let's get to know each other

STRUCTURE
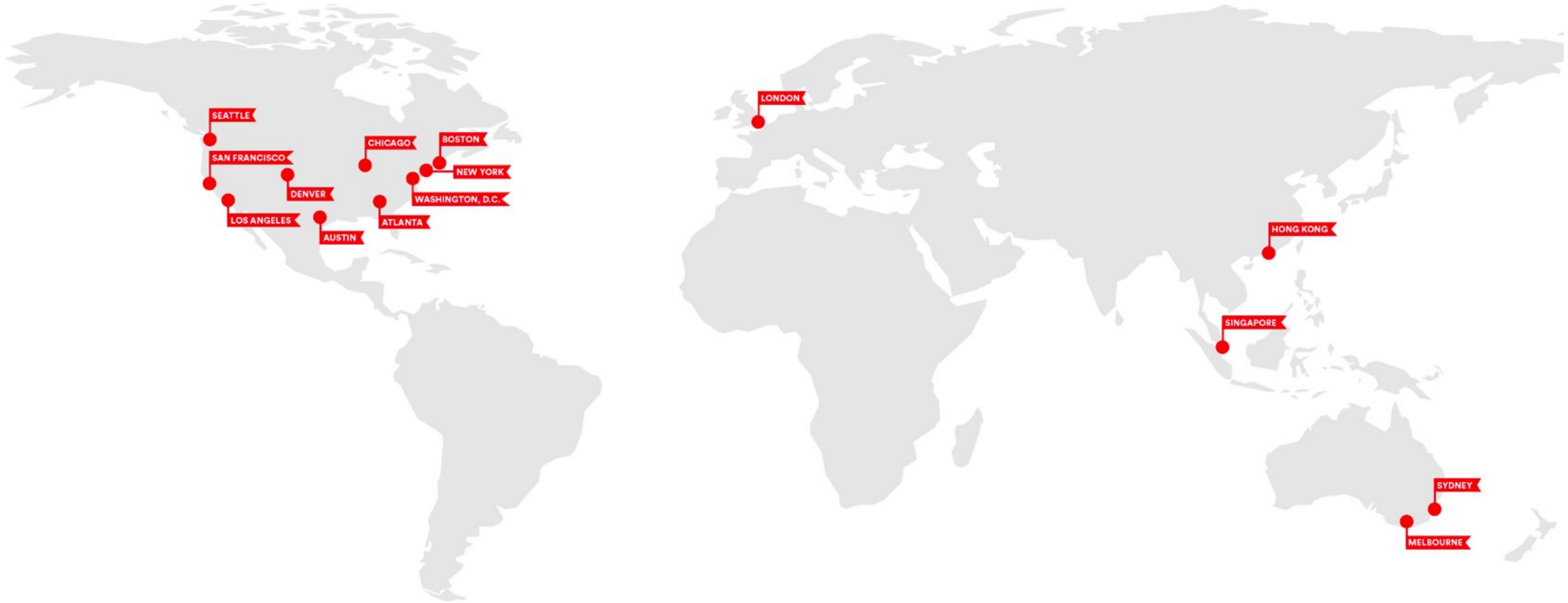
PAIRS

**INTROS: 2 MIN
SHARING: 15 MIN**

OBJECTIVES

1. Take 5 minutes to get to know your neighbor by finding out:

   a. Their name

   b. Why they are taking this course

   c. A guilty pleasure

2. Be prepared to introduce your neighbor to the rest of the room

WHAT IS GENERAL ASSEMBLY?

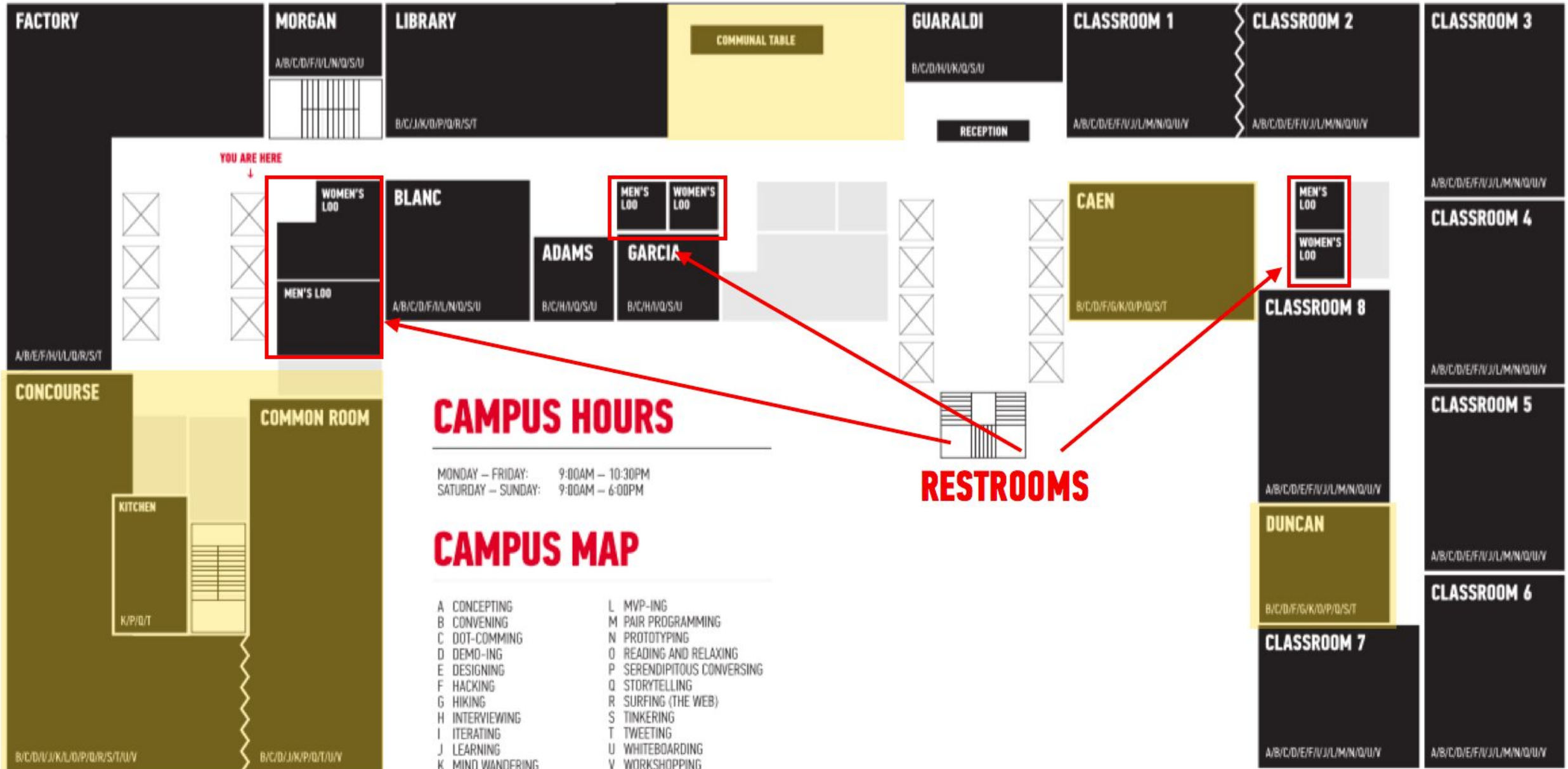# 15+ campuses around the world

Come work on campus.

We're open:
8am - 10pm, Monday to Friday
10am - 6pm, Saturday and Sunday

FACTORY

A/B/C/D/F/I/L/N/Q/S/U

MORGAN

A/B/C/D/F/I/L/N/Q/S/U

B/C/J/K/O/P/Q/R/S/T

LIBRARY

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

CLASSROOM 4

A/B/C/D/E/F/I/J/L/M/N/Q/U/V

YOU ARE HERE

WOMEN'S LOO

MEN'S LOO

A/B/E/F/H/I/L/Q/R/S/T

BLANC

A/B/C/D/F/I/L/N/Q/S/U

ADAMS

B/C/H/I/Q/S/U

GARCIA

B/C/H/I/Q/S/U

MEN'S LOO   WOMEN'S LOO

CAEN

B/C/D/F/G/K/O/P/Q/S/T

MEN'S LOO

WOMEN'S LOO

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

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

RESTROOMS

CONCOURSE

COMMON ROOM

KITCHEN

K/P/Q/T

B/C/D/I/J/K/L/O/P/Q/R/S/T/U/V

B/C/D/J/K/P/Q/T/U/V

PUBLIC USE SPACES

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

Have a question about...
- the campus?
- lost and found?
- loaner equipment?
- free coffee and snacks?

Come here to talk to Front Lines and they will help you out.

Course dates

Data Science {Enter Instance Number Here}

Course dates:

- {___}days and {___}days, 6:30pm - 8:30pm
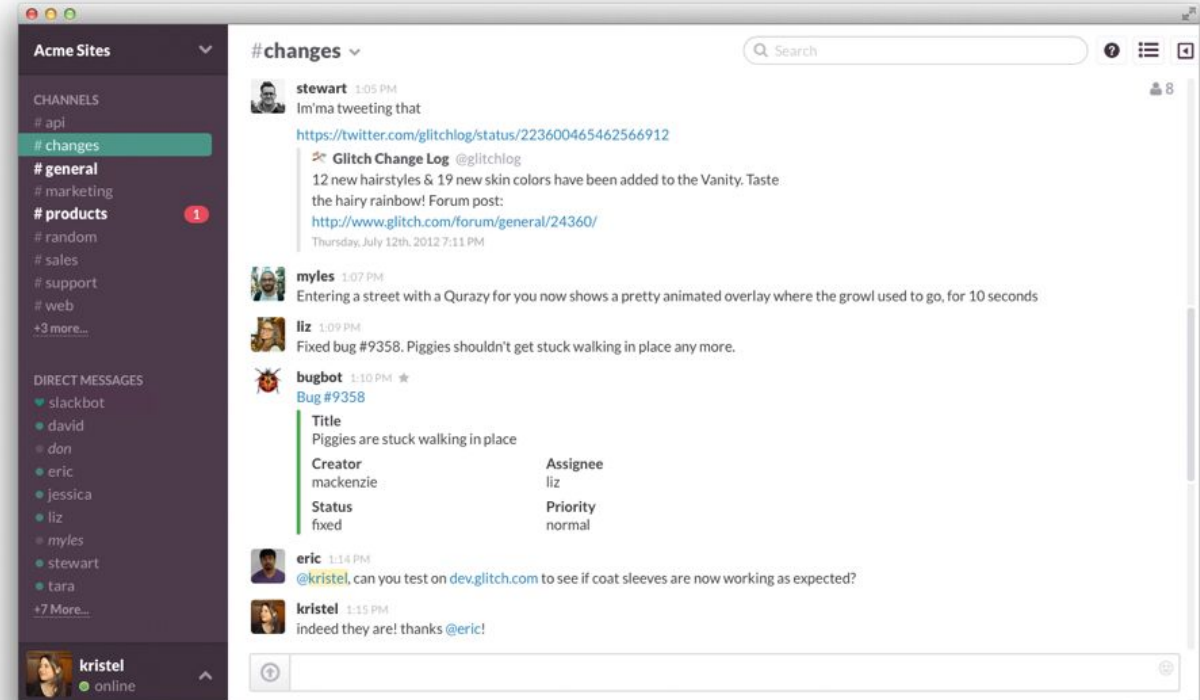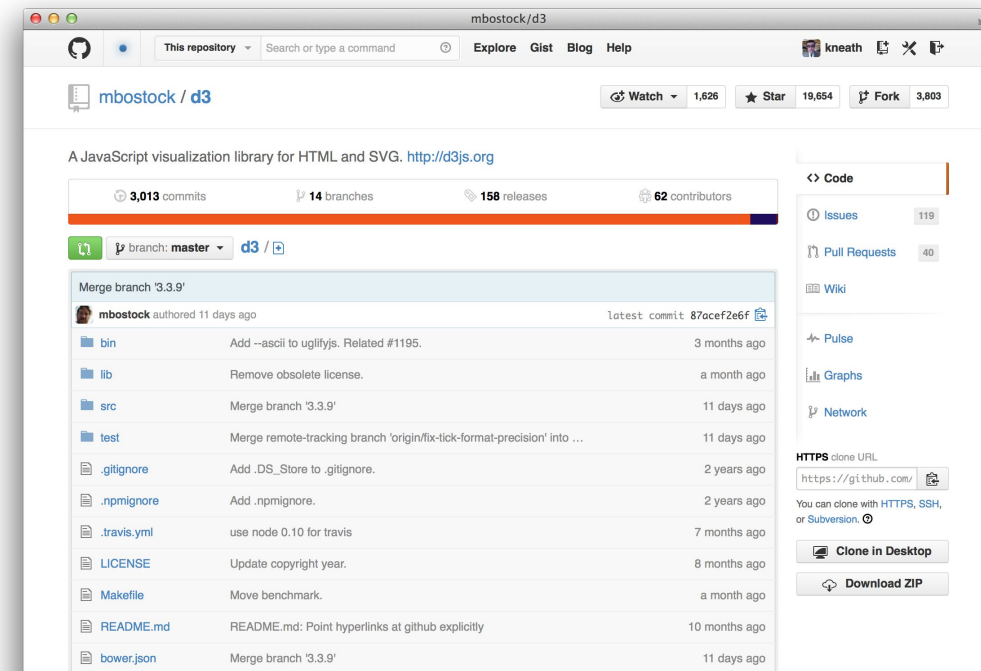- {Start Date - End Date}

No class on:

- {Enter holidays here}

STUDENT EXPERIENCE

Slack

All course communication with each other and
instructors will happen here.

Github

Github will have all the course resources you need:
sample code, assignments, and lesson decks.

Classroom culture

Let's all agree to:
- Treat each other with respect
- Avoid bringing distractions into class
- Add more here....

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

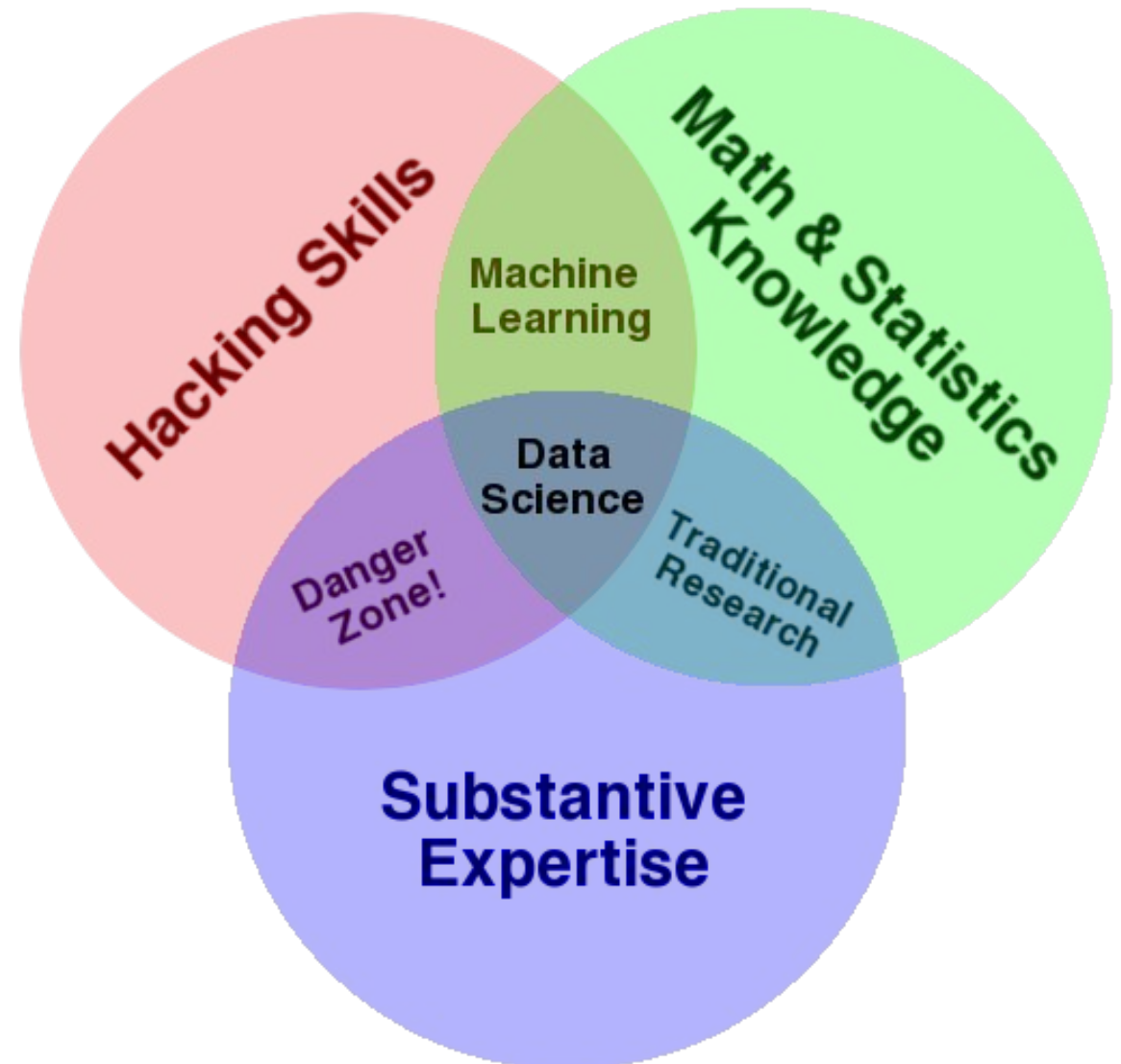‣ Setup your development environment and review python basics

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?

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.

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

WHAT ARE THE ROLES IN DATA SCIENCE?

‣ Data Science involves a variety of skill sets, not just one.
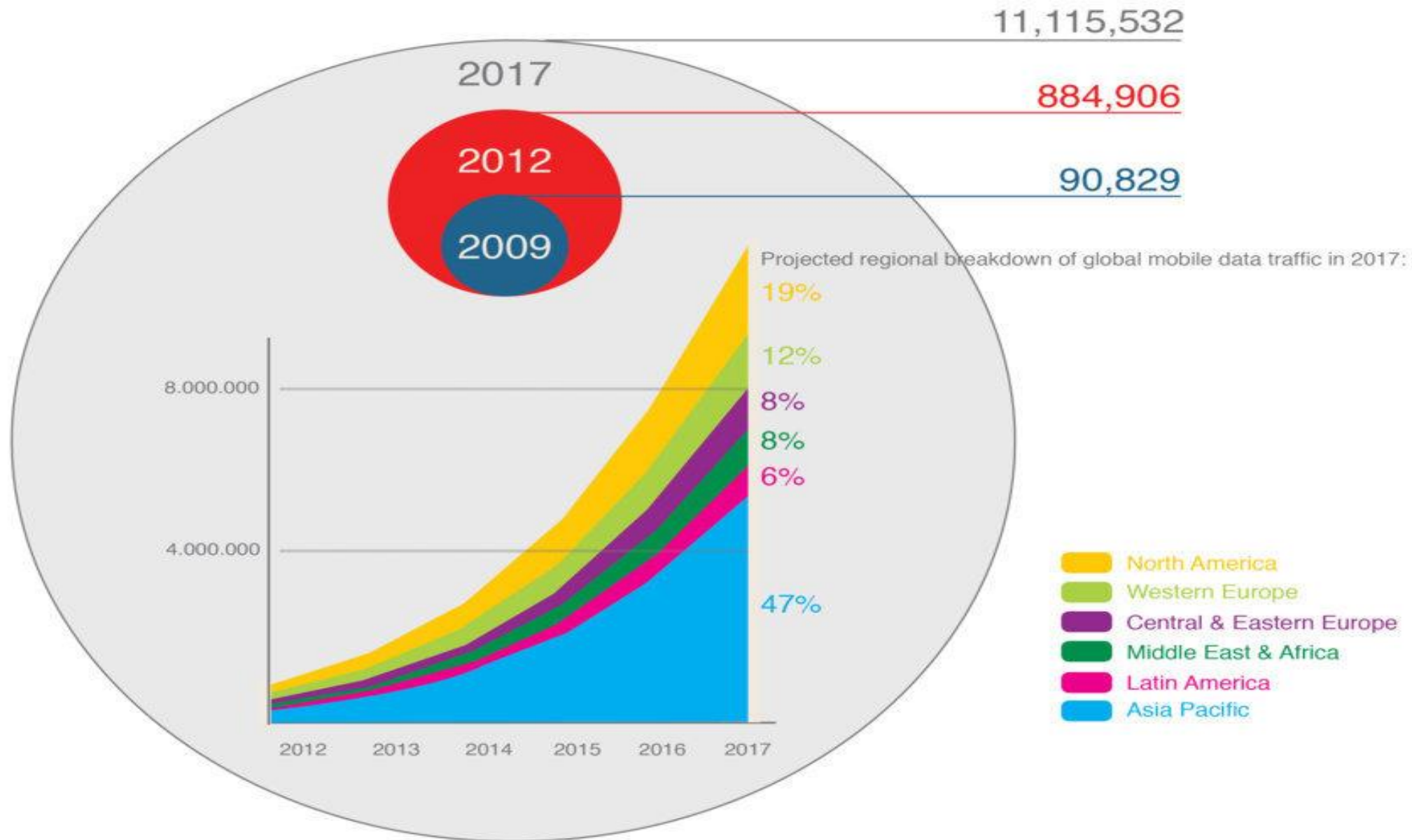
WHAT ARE THE ROLES IN DATA SCIENCE?

‣ These roles prioritize different skill sets.

‣ However, all roles involve some part of each skillset.

‣ Where are your strengths and weaknesses?

Skills and Self—ID Top Factors
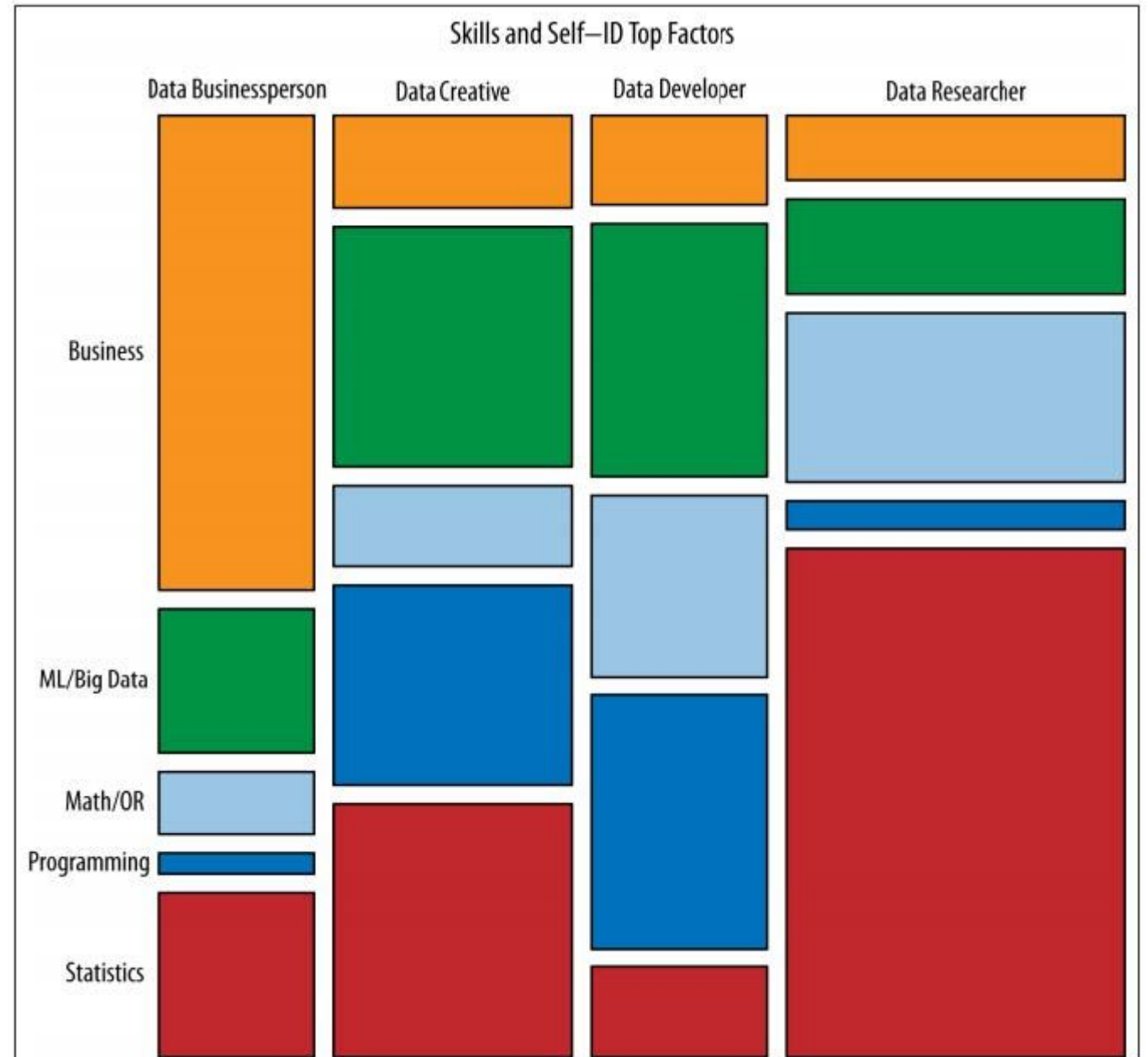
Global Mobile Data - Traffic growth & forecast (terabytes per month)

11,115,532

884,906

90,829

2017

2012

2009

Projected regional breakdown of global mobile data traffic in 2017:

19%

12%

8%

8%

6%

47%

8.000.000

4.000.000

2012   2013   2014   2015   2016   2017

- North America
- Western Europe
- Central & Eastern Europe
- Middle East & Africa
- Latin America
- Asia Pacific

Data Scientists tell stories

Napoleon's 1812 Russian Campaign

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

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

Supervised Learning	Unsupervised Learning	Reinforcement Learning
• Classification • Regression • Ranking	• Clustering • Association Mining • Segmentation • Dimension Reduction	• Decision Process • Reward System • Recommendation Systems

DATA SCIENCE BASELINE

ACTIVITY: DATA SCIENCE BASELINE QUIZ

DIRECTIONS (10 minutes)

EXERCISE

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

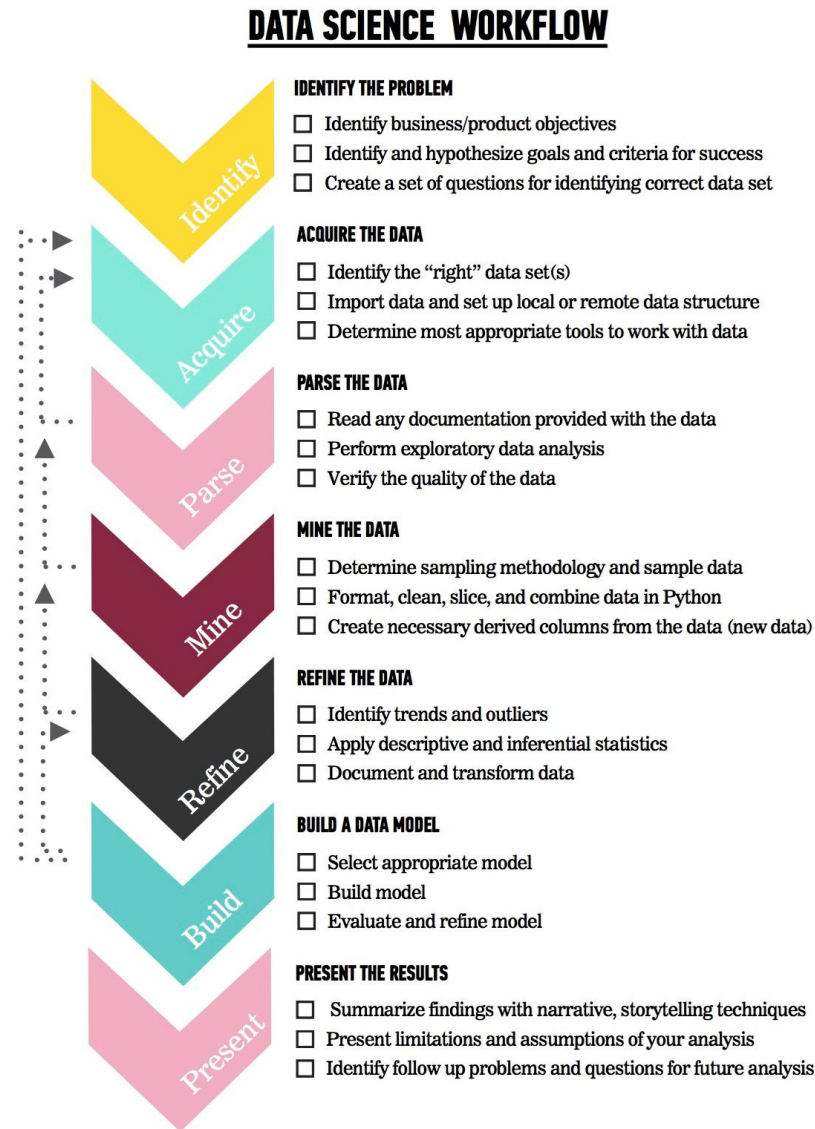   ‣ *Reproducible*: Others can follow your steps and get the same results

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

DATA SCIENCE WORKFLOW

Identify

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

Acquire

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

Parse

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

Mine

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)

Refine

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

Build

BUILD A DATA MODEL
- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

Present

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

Identify

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

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

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

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

Refine

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

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Build

BUILD A DATA MODEL

☐ Select appropriate model

☐ Build model

☐ Evaluate and refine model

OVERVIEW OF THE DATA SCIENCE WORKFLOW

Present

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 and costs banks 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.

DATA SCIENCE WORK FLOW

ACTIVITY: DATA SCIENCE WORKFLOW

EXERCISE

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

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?

DATA SCIENCE
BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

‣ Project: Begin work on Project 1

Q & A

WELCOME TO DATA SCIENCE

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET