GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Mason Gallo

LEARNING OBJECTIVES

- Describe the roles and components of a successful learning environment
- Define data science and the data science workflow
- Setup your development environment and review python basics

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Did everyone receive and complete pre-work?
- Basic Python and stats

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

TRANSFORMING THINKERS INTO CREATORS.

MOVE MAKER
Made the move to the big city and big websites.
Larry's Story →

CAREER CHANGER
Left the law behind to follow her own path.
Whitney's Story →

BUSINESS BUILDER
Went from working in news to working for himself.
Benji's Story →

GA GRADUATION REQUIREMENTS

HOMEWORK

(COMPLETE 80% OF HOMEWORK/LABS)

ATTENDANCE

(MISS NO MORE THAN 2 CLASSES)

FINAL PROJECT

COMMUNITY ENGAGEMENT

PARTICIPATION + FEEDBACK

FOREVER AND EVER

BUILD YOUR NETWORK

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

FIND OPPORTUNITIES

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

13,000+ STRONG

You're part of the alumni community forever

PERKS!

We can't wait to have you back on campus

DATA SCIENCE

MEET YOUR TEAM!

YOUR INSTRUCTOR: MASON GALLO

WHO AM I

- Data Scientist
- Open Source Contributor
- ML Researcher - Educational Technology
- Data Science Instructor @ GA

Washington University in St.Louis

HAVAS helia. creative with data

Georgia Tech

GA GENERAL ASSEMBLY

HAVAS WORLDWIDE

OmnicomMediaGroup OMP phd

Duke Clinical Research Institute
From Thought Leadership to Clinical Practice

Columbia University

YOUR ASSOCIATE INSTRUCTOR: PAUL SINGMAN

WHO AM I

- Data Engineer
- Former Actuary
- Insight Fellow

Wharton

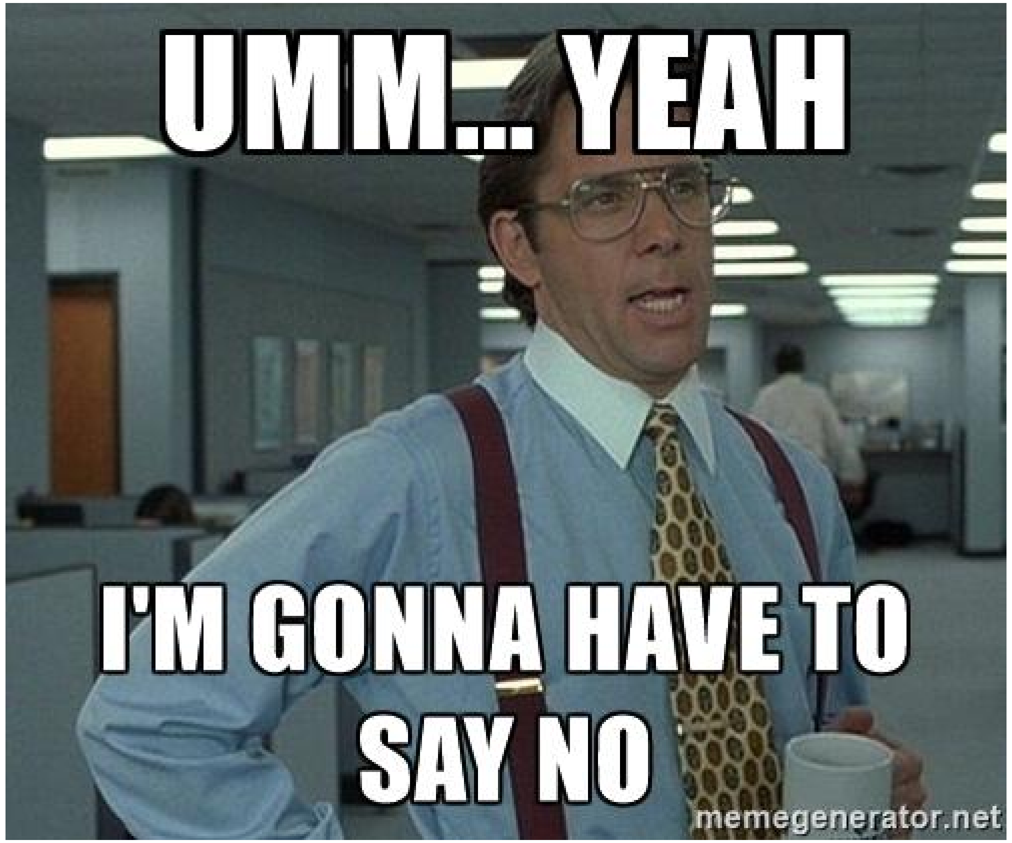
NEW YORK LIFE

GA GENERAL ASSEMBLY

MY PHILOSOPHY

- If you’re not sure, please ask!
- If you’re still stuck, we’ll revisit together
- Don’t forget your fellow students
- Participate!!!
- Breaks
- This stuff is hard for everyone
- Don’t forget Paul’s office hours!

TO MASTER DATA SCIENCE?



BIG PICTURE STRATEGY

- Good amount of breadth
- Just enough depth
- Intuition first
- Make you sweat a little
- Baseline for the future!

ICE BREAKER

TELL THE CLASS

- Your name
- 1 sentence description of your background
- 1 sentence description of why you are taking data science
- Your guilty pleasure

DEMO

ENVIRONMENT SETUP

DEV ENVIRONMENT SETUP

- This is extremely important

- Little bit of pain now so we can focus on data science moving forward!

NOTE ABOUT TECH SUPPORT

- Mac is preferred environment and support will be provided
- All in-class examples and graphics made from my computer (Mac)
- Windows IT support is NOT guaranteed
- If using work laptop, become friends with your IT team

PRE-WORK

- All check for the pre-work PDF sent before class
- Instructional team to personally confirm each student
- Please let us know if you did not receive the PDF

SLACK

- You should have already received a Slack invite
- Confirm that you’re all set and always have Slack open during class

GITHUB

- If you didn’t already, create a GitHub account
- Enter your GitHub name into Slack so we can grant you access to course materials!
- Note: git is outside the scope of this course but highly recommended

COURSE MATERIALS

- Bookmark this page for course materials
- Download zip vs git clone
- If you’ve never used Git, you will need to make sure you stay updated
- Remember to always “save as” your work!

https://github.com/ga-students/DAT-NYC-45

INTRODUCTION

OUR TOOLS

POPULAR TEXT EDITORS

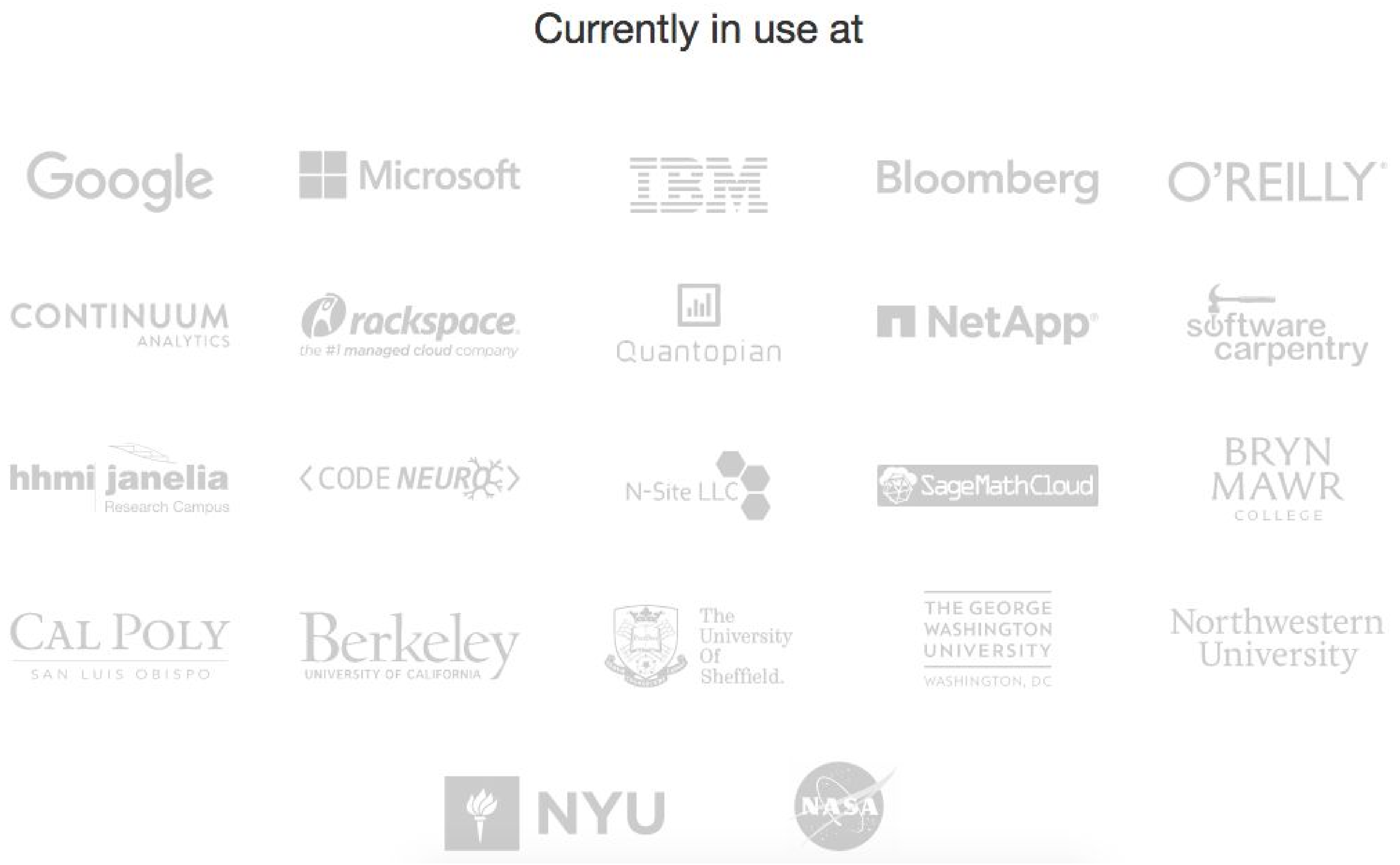
- SublimeText: http://www.sublimetext.com/3
- Atom: https://atom.io

JUPYTER NOTEBOOK

Typical Python workflow:

- Prototype/ideate in a “notebook”
- Share notebook with colleagues
- Use notebook for presentations
- If/when ideas go to production, export the notebook as a .py file
- Use text editor (such as Sublime Text) to edit .py files
- Use command line to run your .py files

JUPYTER NOTEBOOK IN USE EVERYWHERE

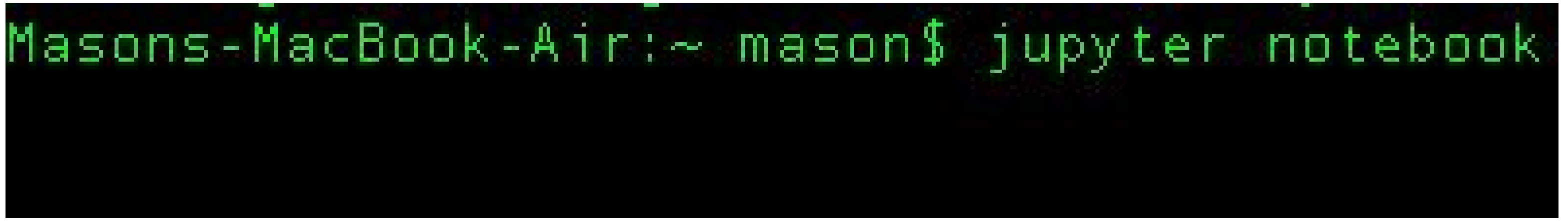


Currently in use at

JUPYTER NOTEBOOK INSTALLED VIA ANACONDA

If you installed Anaconda like I recommended, you already have it!

- Open your terminal / powershell
- Type “jupyter notebook” (without the quotes) and press enter
- Put your class materials in an easy to locate place, like Documents



```
Masons-MacBook-Air:~ mason$ jupyter notebook
```

- Once the browser window opens, navigate to the course materials and open DAT-NYC-45/classes/01/notebook_intro.ipynb

DEV ENVIRONMENT SETUP

- Test your new setup using the lesson 1 starter code available at */classes/01/code/starter-code/lesson1-starter-code.ipynb* in the Github class repo

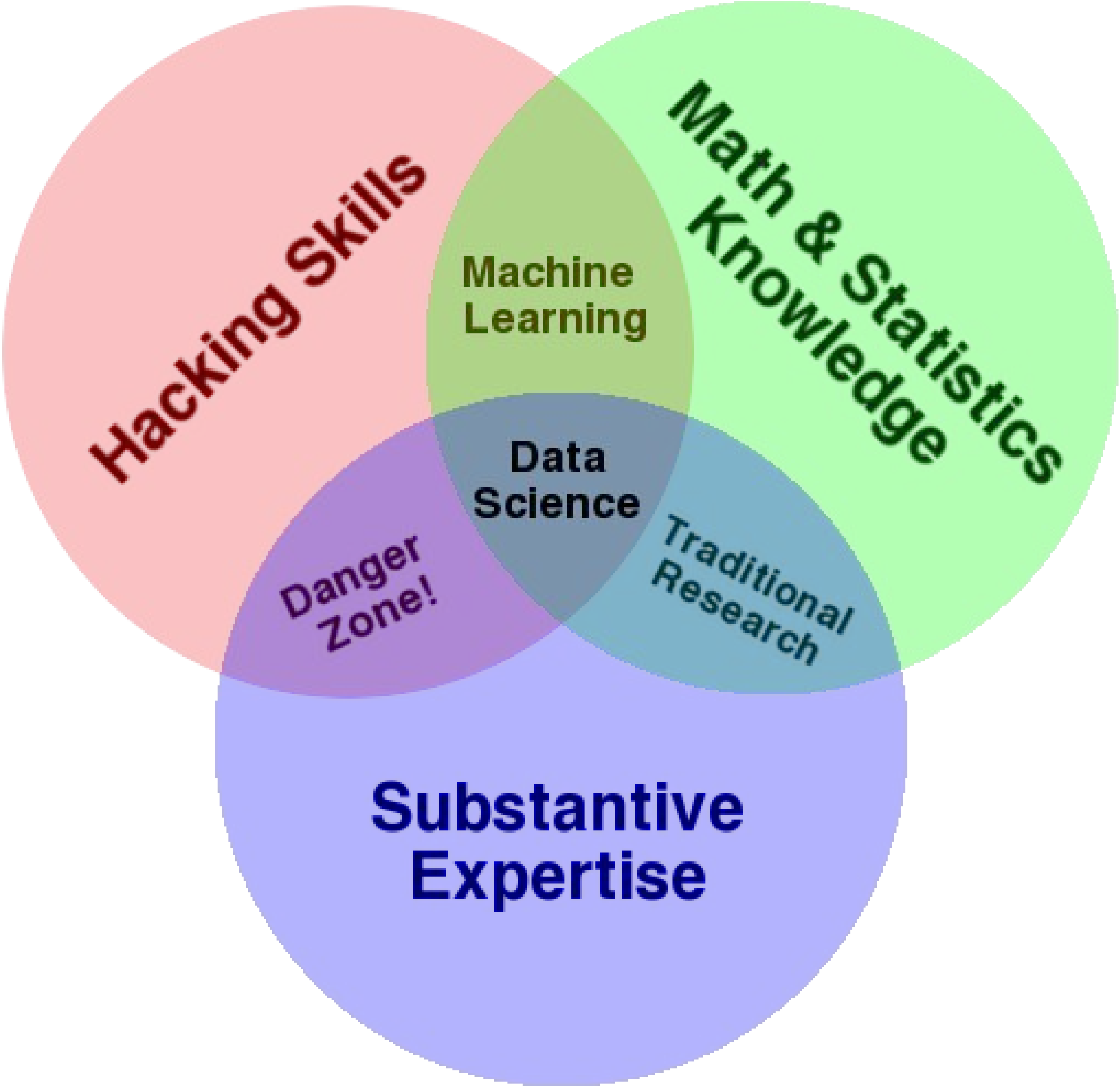
- Ask your classmates and instructor for help if you have problems!

INTRODUCTION

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 USES DATA SCIENCE?

NETFLIX

amazon.com®

Google

FiveThirtyEight

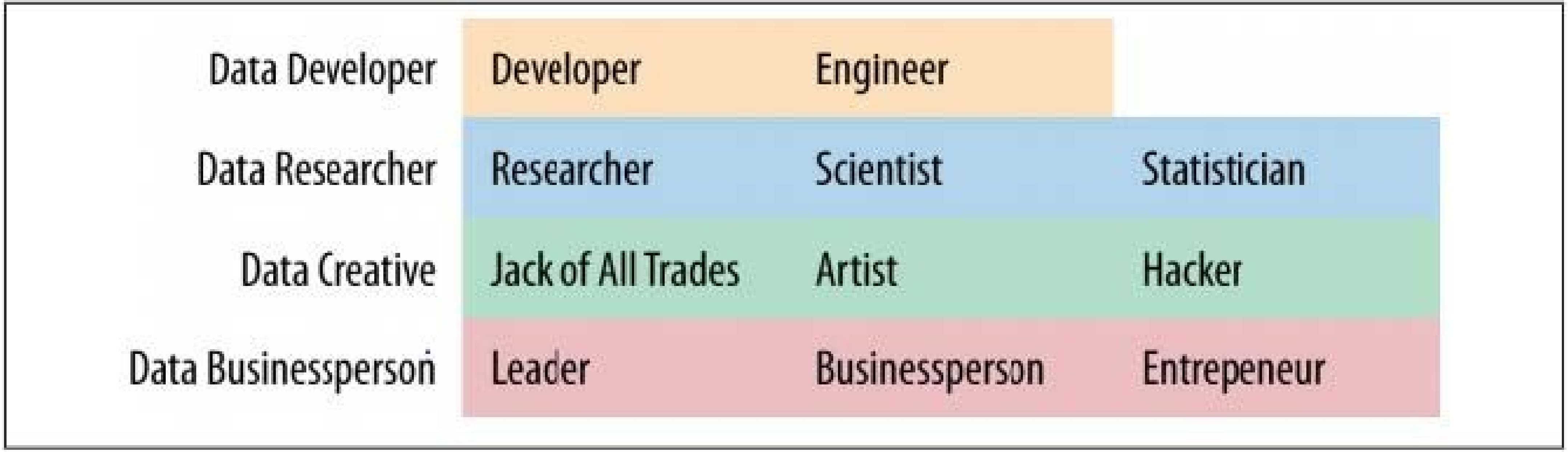
DRAFT KINGS

WHO USES DATA SCIENCE?

- Can you think of others?

WHAT ARE THE ROLES IN DATA SCIENCE?

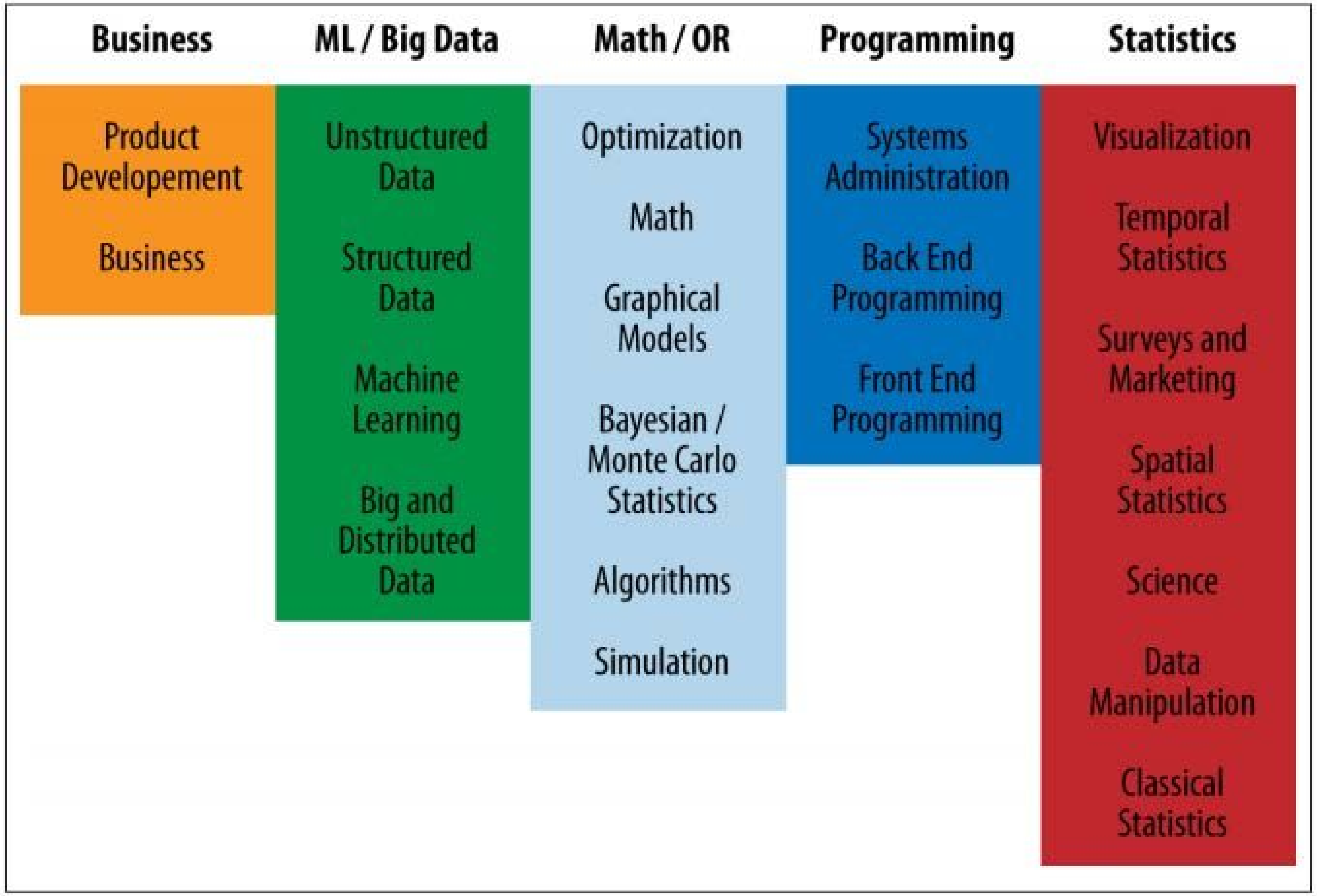
- Data Science involves a variety of roles, not just one.



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.



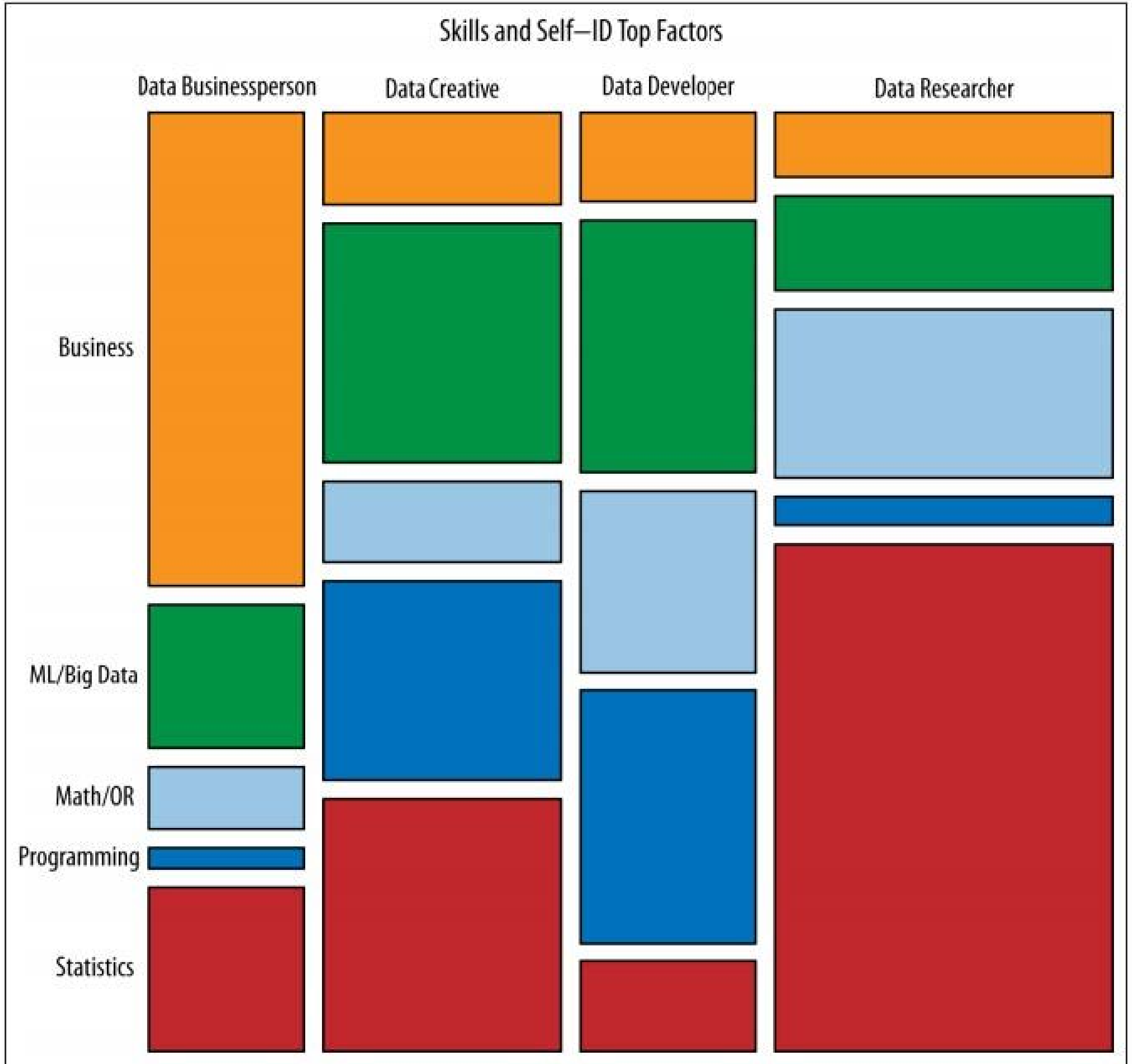
Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

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

1. Form groups of three.
2. Answer the following questions on your tables
   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 an outlier?
   e. What’s the difference between classification and regression?

INTRODUCTION

THE DATA SCIENCE WORKFLOW

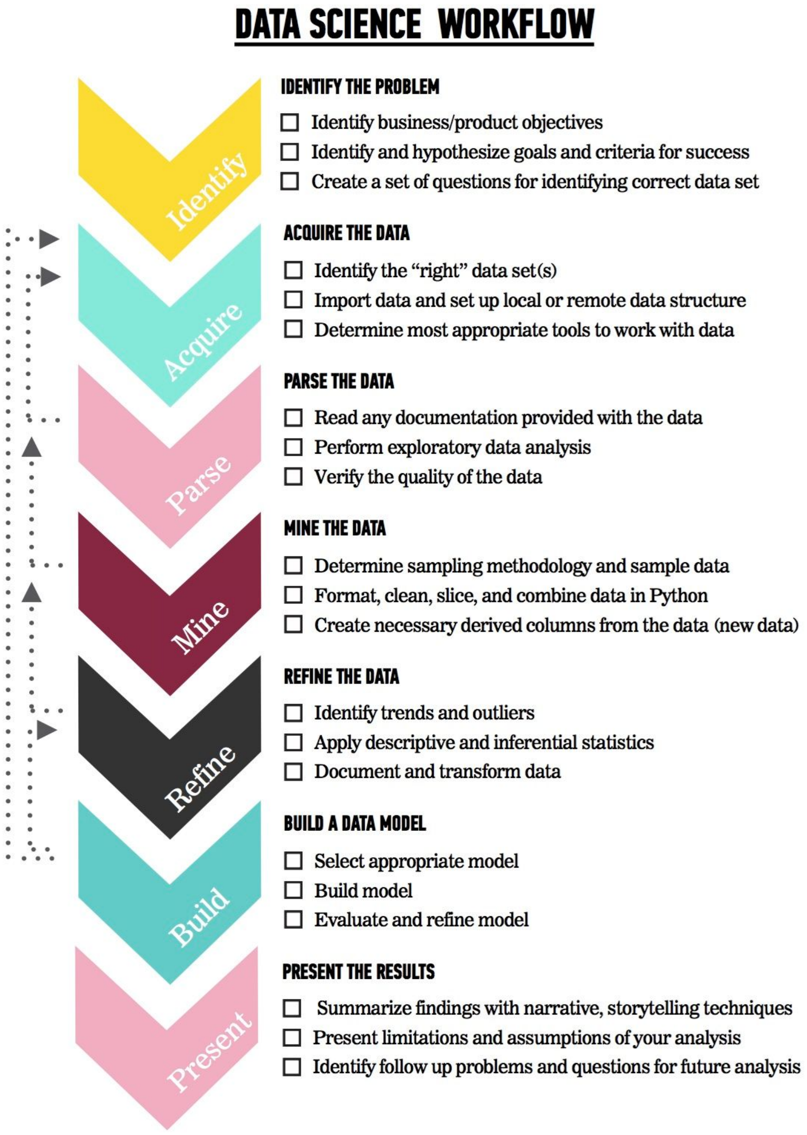
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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



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

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 THE DATA

- □ Read any documentation provided with the data
- □ Perform exploratory data analysis
- □ Verify the quality of the data

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 THE DATA

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

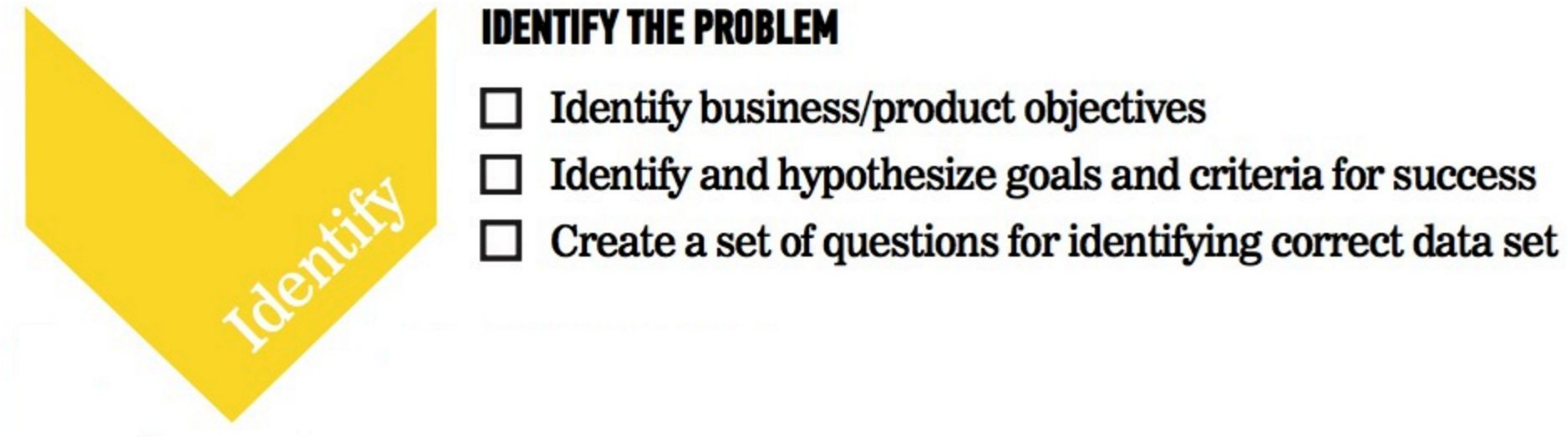
BUILD A DATA MODEL

- □ Select appropriate model
- □ Build model
- □ Evaluate and refine model

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



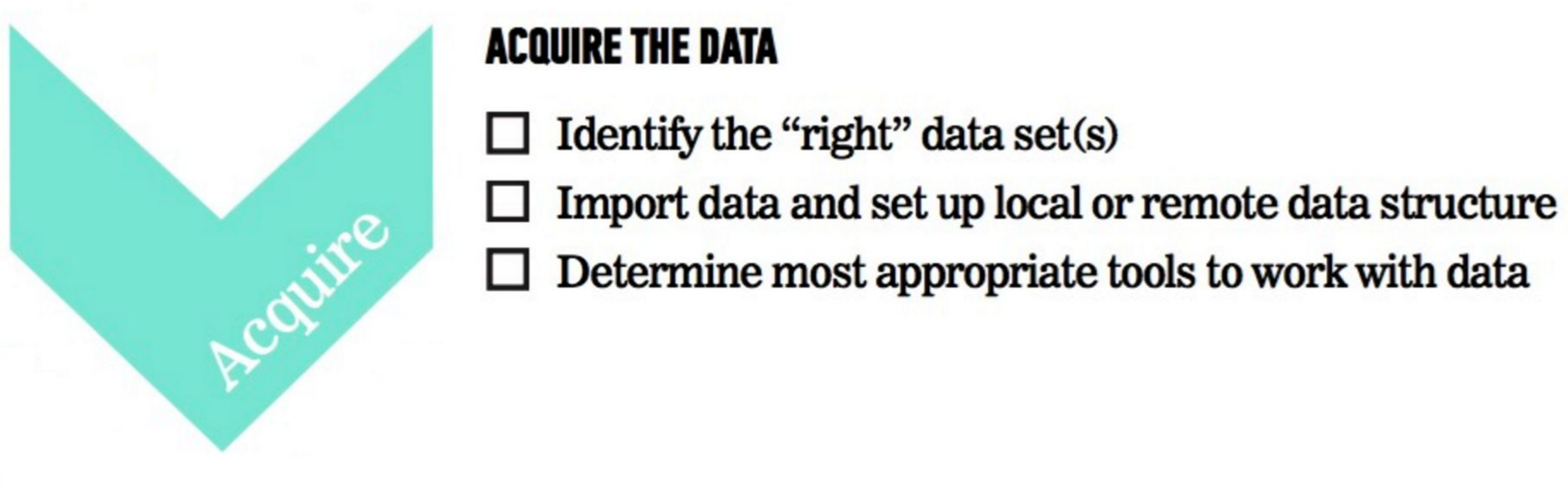
IDENTIFY THE PROBLEM

□ 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



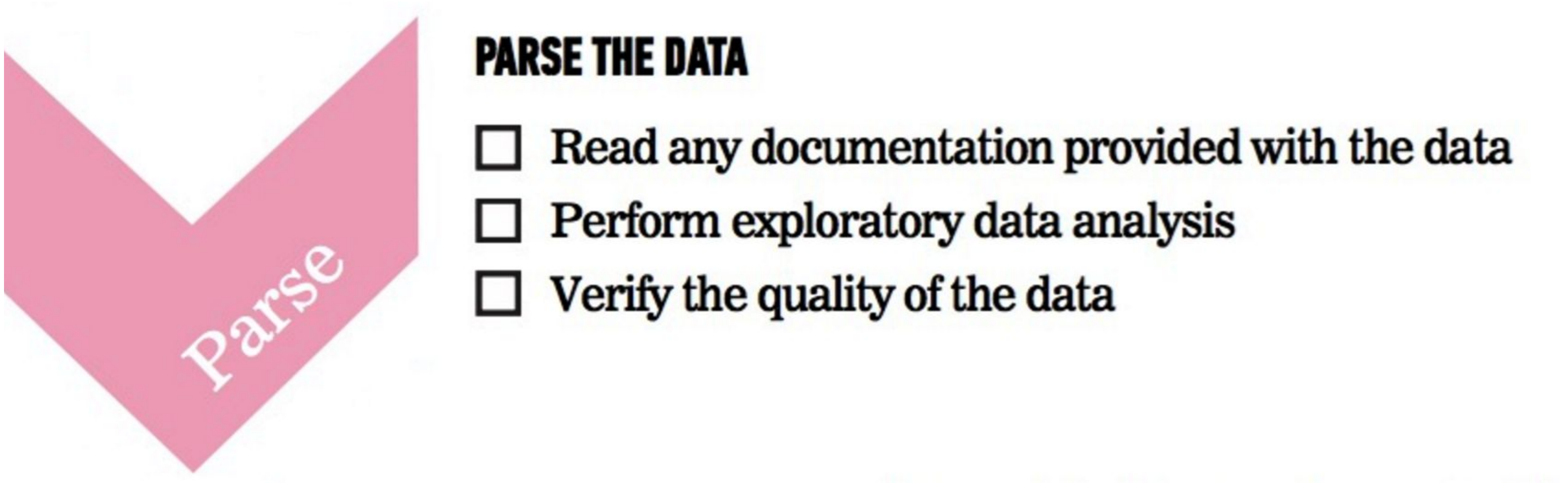
ACQUIRE THE DATA

□ Identify the "right" data set(s)

□ Import data and set up local or remote data structure

□ Determine most appropriate tools to work with data

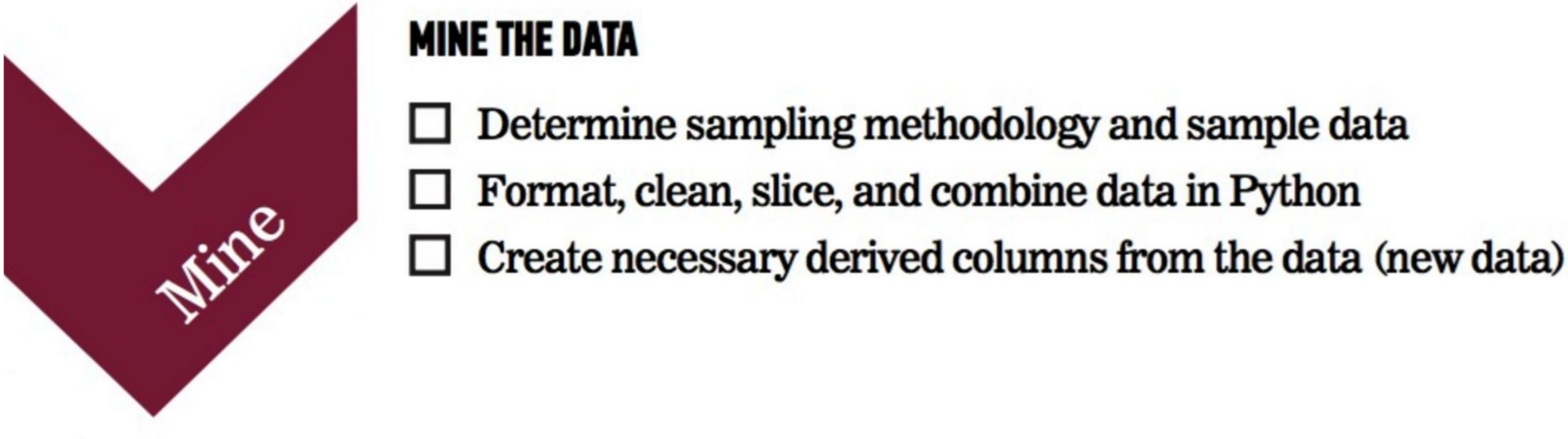
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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



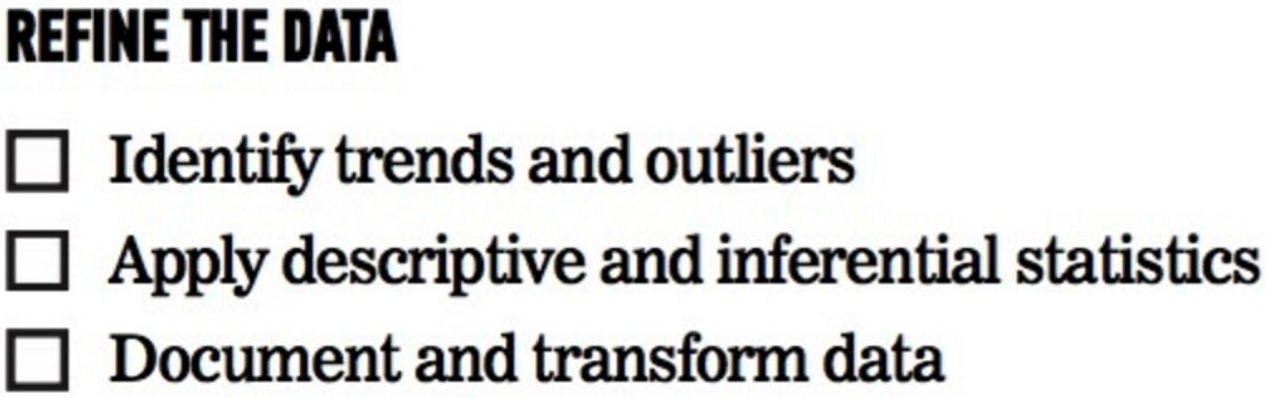
MINE THE DATA

□ Determine sampling methodology and sample data

□ Format, clean, slice, and combine data in Python

□ Create necessary derived columns from the data (new data)

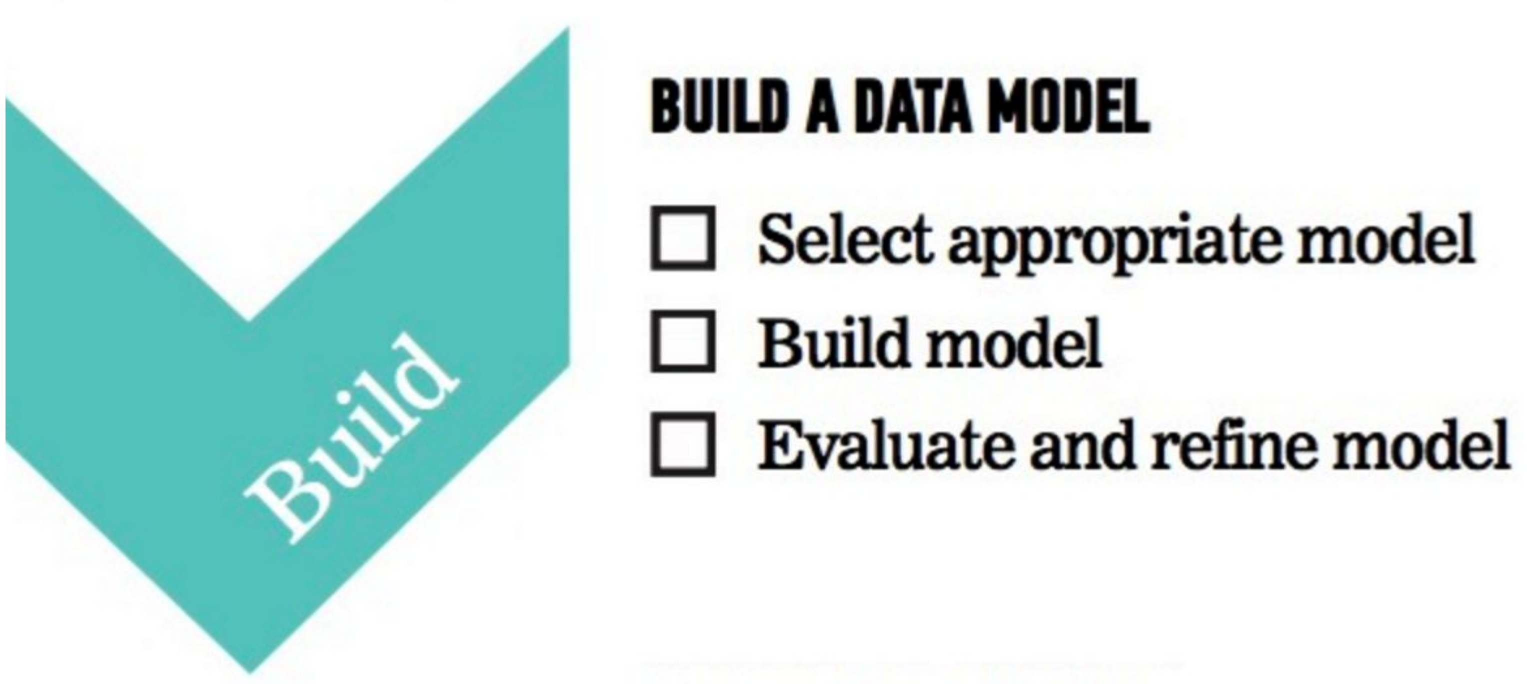
OVERVIEW OF THE DATA SCIENCE WORKFLOW



REFINE THE DATA

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

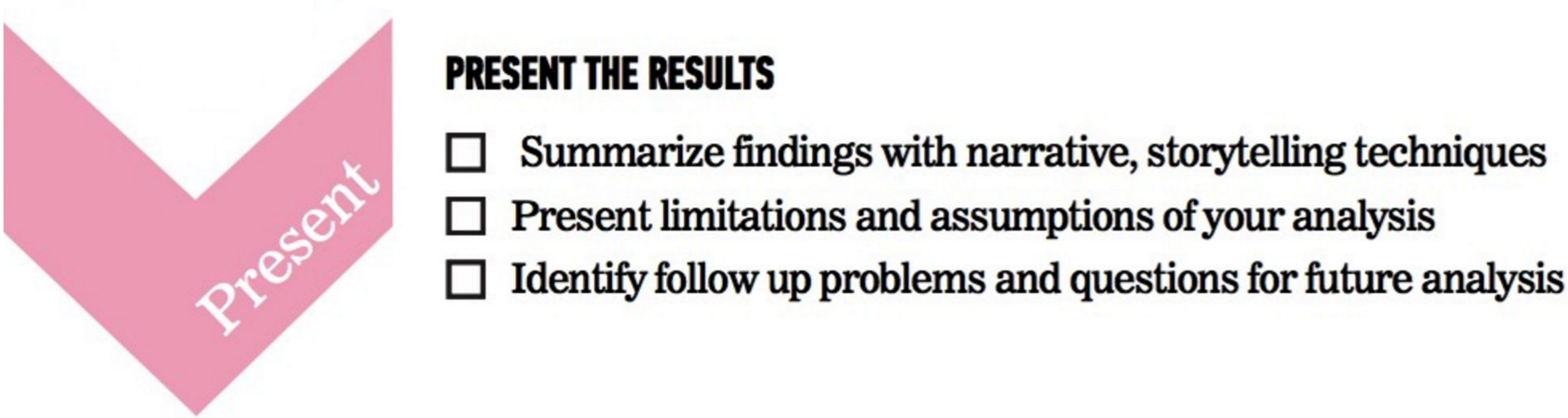
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

MACY’S EXAMPLE

- Problem Statement: “Using credit card transaction data from the past 2 years at Macy’s, determine the factors that lead to increased customer basket size.”

- We can use the Data Science workflow to work through this problem.

MACY'S EXAMPLE: IDENTIFY THE PROBLEM

- The objective is basket size $

- Do we need to describe the relationship between predictors and basket size? Do we simply need to make a prediction for each customer visit?

- Create a set of questions to help you identify the correct data set.

MACY'S EXAMPLE: ACQUIRE THE DATA

- Ideal data vs. data that is available: credit card / loyalty card purchases

- Learn about limitations of the data: what about cash purchases?

- What data is available for this example?

- What kind of questions might we want to ask about the data?
    - Representative of the general population?

MACY’S EXAMPLE: ACQUIRE THE DATA

- Questions to ask about the data
  - Is there enough data? % of total purchases / revenue?
  - Does it appropriately align with the question/problem statement?
  - Can the dataset be trusted? How was it collected?
  - Do we have customer level data? How is the data grouped?

MACY'S 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
Gender	0- male, 1- female	Categorical
Location	Zip code	Categorical
Days Since Last Purchase	Integer	Continuous
Age	Integer	Continuous

MACY'S EXAMPLE: MINE THE DATA

- Think about sampling: can we take a random sample? What about timing?

- Get to know the data

- Explore outliers: extremely large or small basket sizes?

- Address missing values: any trends in what is missing?

- Derive new variables (i.e. columns)

MACY'S EXAMPLE: REFINE THE DATA

- Use statistics and visualization to identify trends

- Example of basic statistics

Variable	**Mean (STD) or Frequency (%)**
Age	45.7
Gender	70% Female
Days Since Last Purchase	22.4

MACY'S 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

MACY'S EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome

- Example model statement: "We performed ridge regression to predict the customer basket size at Macy's using credit card transactional data."

MACY'S 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

MACY'S 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.

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?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Familiarize with Unit Project 1 (12/15)
- Python fundamentals with Think Python: http://greenteapress.com/thinkpython/thinkpython.pdf

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET