BEFORE WE GET STARTED

Download or clone today’s materials from GitHub

https://github.com/Dt1431/chi-ds-5-lesson-1

GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

David Turner

CEO, Waitbot Inc.

OUTLINE OF THE DAY

1. Welcome to GA / Course Information
2. Individual Introductions
3. Main Lesson (What is Data Science?)
4. Development Environment
5. Conclusions

ABOUT GENERAL ASSEMBLY

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.

ROAD TO SUCCESS

WE’RE ALL IN THIS TOGETHER.

Make the Most of your Experience!

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

STUDENT RESPONSIBILITY

As a self-directed program, we view students as a crucial part of the skill acquisition process.

GA	YOU
50%	50%

What Defines a Successful Grad?

HOMEWORK

(COMPLETE 80% OF HOMEWORK/LABS)

ATTENDANCE

no more than 3 absences

FINAL PROJECT

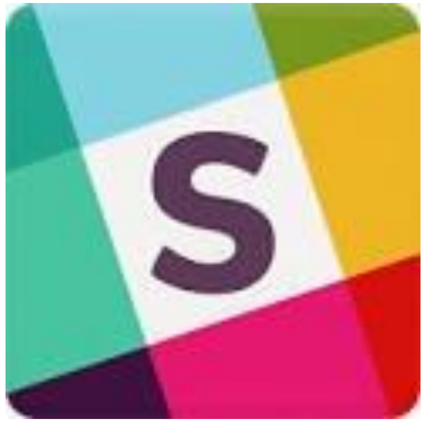
COMMUNITY ENGAGEMENT

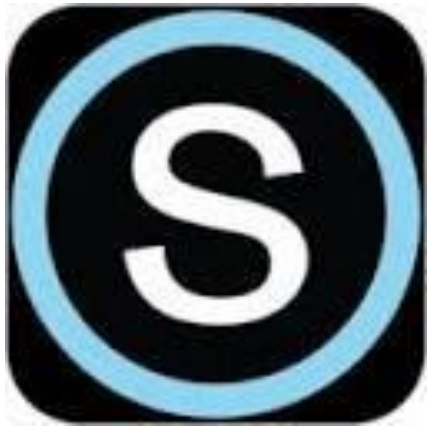
COURSE LOGISTICS

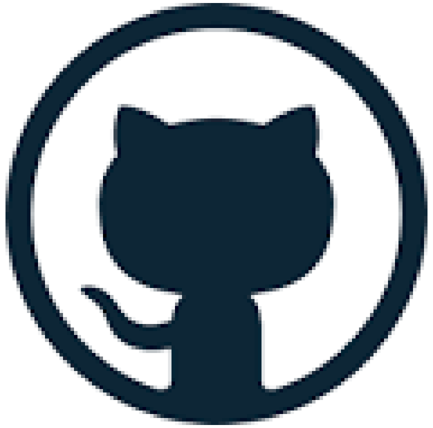
Feedback is a Value at GA

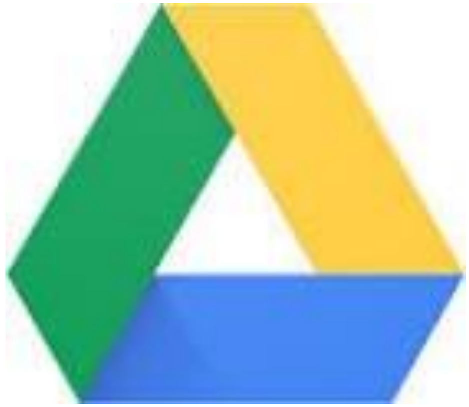
Exit Tickets	You’ll fill out a brief informal survey before you leave each lesson
MidCourse Feedback Survey	You’ll fill out a formal feedback survey in the middle of the course that we’ll review and implement changes as necessary
End of course Feedback Survey	You’ll fill out a formal feedback survey on the last night of class so we know how the course went for you
Informal Feedback	We encourage you to deliver feedback outside of these formal surveys to the instructional team and our course producer

Course Tools









Projects

- Homework/Unit Projects
  - 4 Unit Projects in Data Science
  - Each builds on top of skills learned previous
  - Assigned approximately ~2 weeks during first half of course
  - Full timeline available in the syllabus (main Github folder)

- Final Project
  - Address a data-related problem in your professional field
  - Acquire a real-world data set, form a hypothesis about it, clean, parse, and apply modeling techniques and data analysis principles
  - 5 structured assignments
  - Presentation of results and written report

CLASSROOM RULES & EXPECTATIONS

- Open and focused discussion is encouraged
  - Be mindful of giving everyone an equal chance to talk
  - Raise your hand before you speak
  - Zero tolerance for discrimination or harassment

- Laptops are a required part of the class
  - Used during the lab sessions
  - Must be closed during the lecture portion.
  - Take notes using pen and paper

CLASSROOM ACCOMODATIONS

- WiFi is provided by General Assembly
  - Network Name: SPACE
  - Password (lowercase): w0rk5pac3

- Restrooms are located near the elevator
  - Feel free to use at any time
  - Try to minimize distraction when entering/leaving

- Power outlets
  - Located in the middle of the room
  - Located on the left and right sides of the room

ASK AWAY!

INTRODUCTION

CLASS INTROUDCTIONS

ABOUT ME

- 15+ years experience building software products for the financial industry, consulting with non-profits, and launching innovative digital enterprises.
- CEO of Waitbot Inc., a smart city technology company dedicated to saving people time and making organizations more operationally efficient.
- Featured on NPR and BBC, and consulted by the White House Business Counsel.
- BS in Computer Engineering, and Masters in Business Administration & Public Policy each from the University of Michigan.

ABOUT YOU - ICEBREAKER

- What is your name?
- What do you hope to learn?
- What would be the name of your autobiography? And why?

INTRODUCTION

WHAT IS DATA SCIENCE?

LEARNING OBJECTIVES

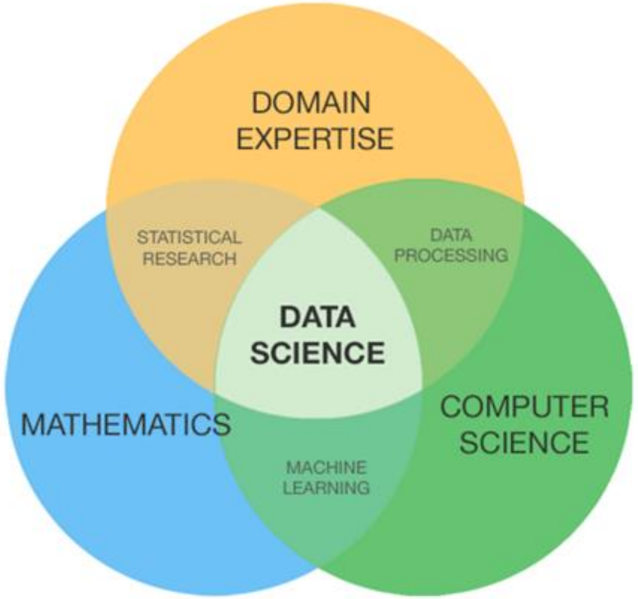
- Define data science and the data science workflow

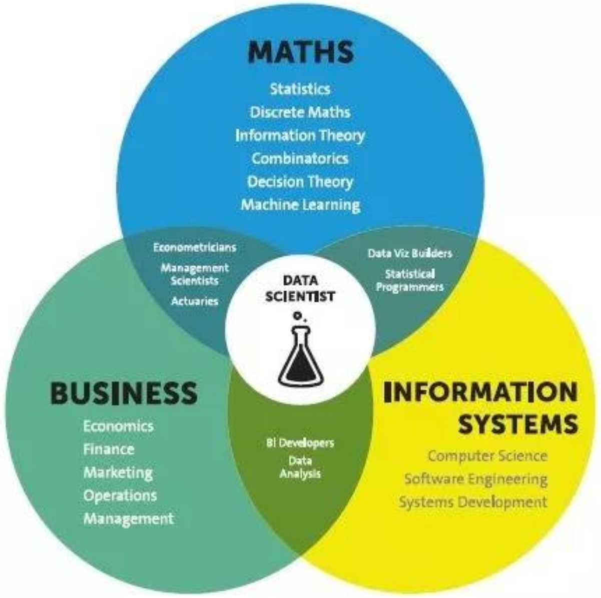
- Apply the data science workflow

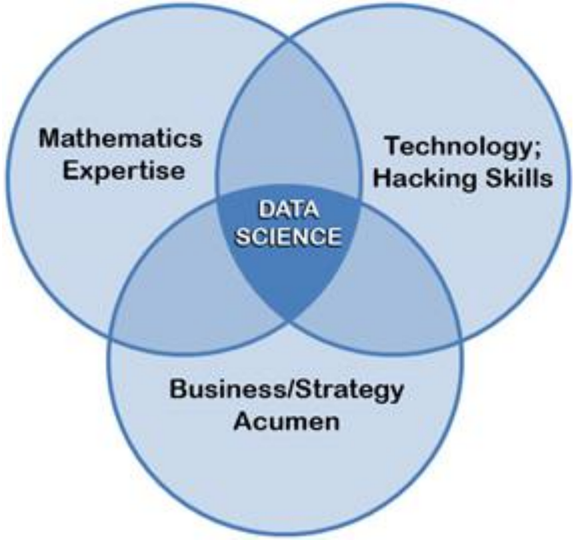
- Setup your development environment and review python basics

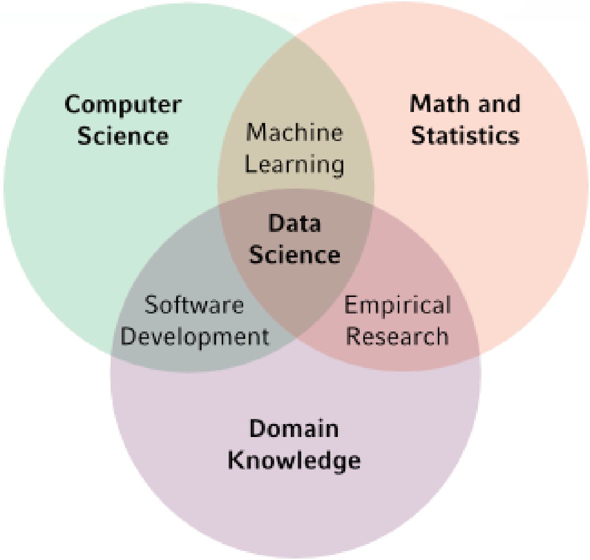
WHAT IS DATA SCIENCE?

- A set of tools and techniques for data
- Interdisciplinary problem-solving
- Multiple definitions
- Commonalities: Application of statistical and computational techniques to practical problems using the scientific method



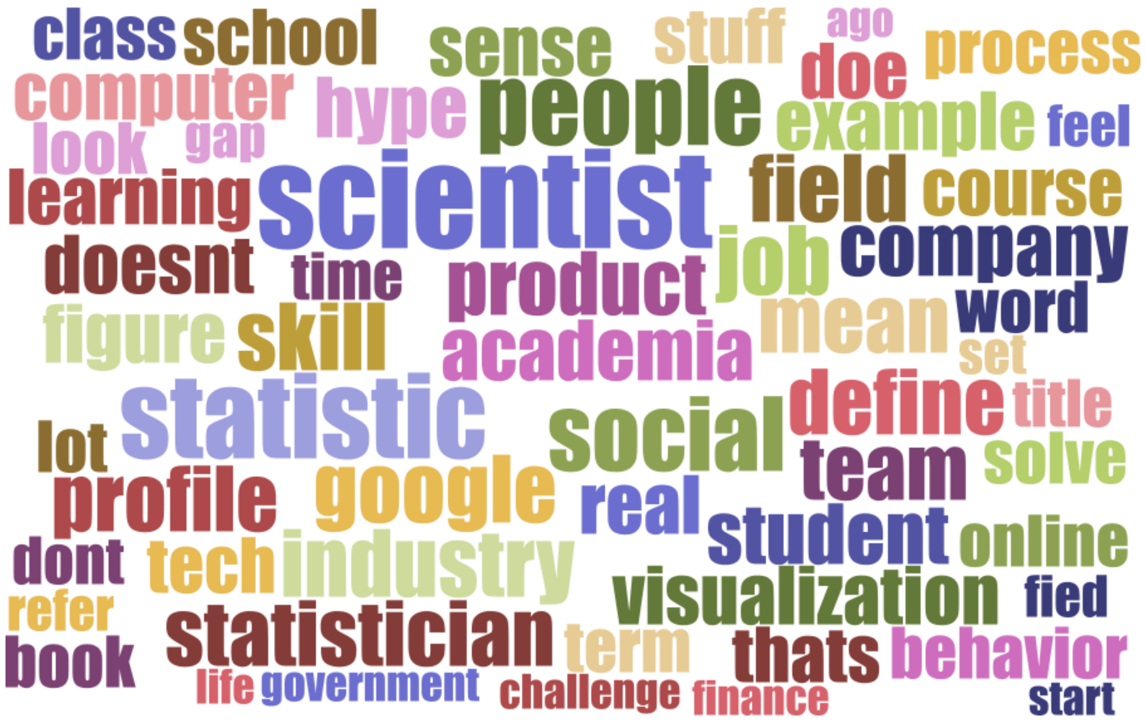






WHAT IS DATA SCIENCE?: Illustrated Example

WORD CLOUD OF
“DOING DATA SCIENCE: CHAPTER 1”

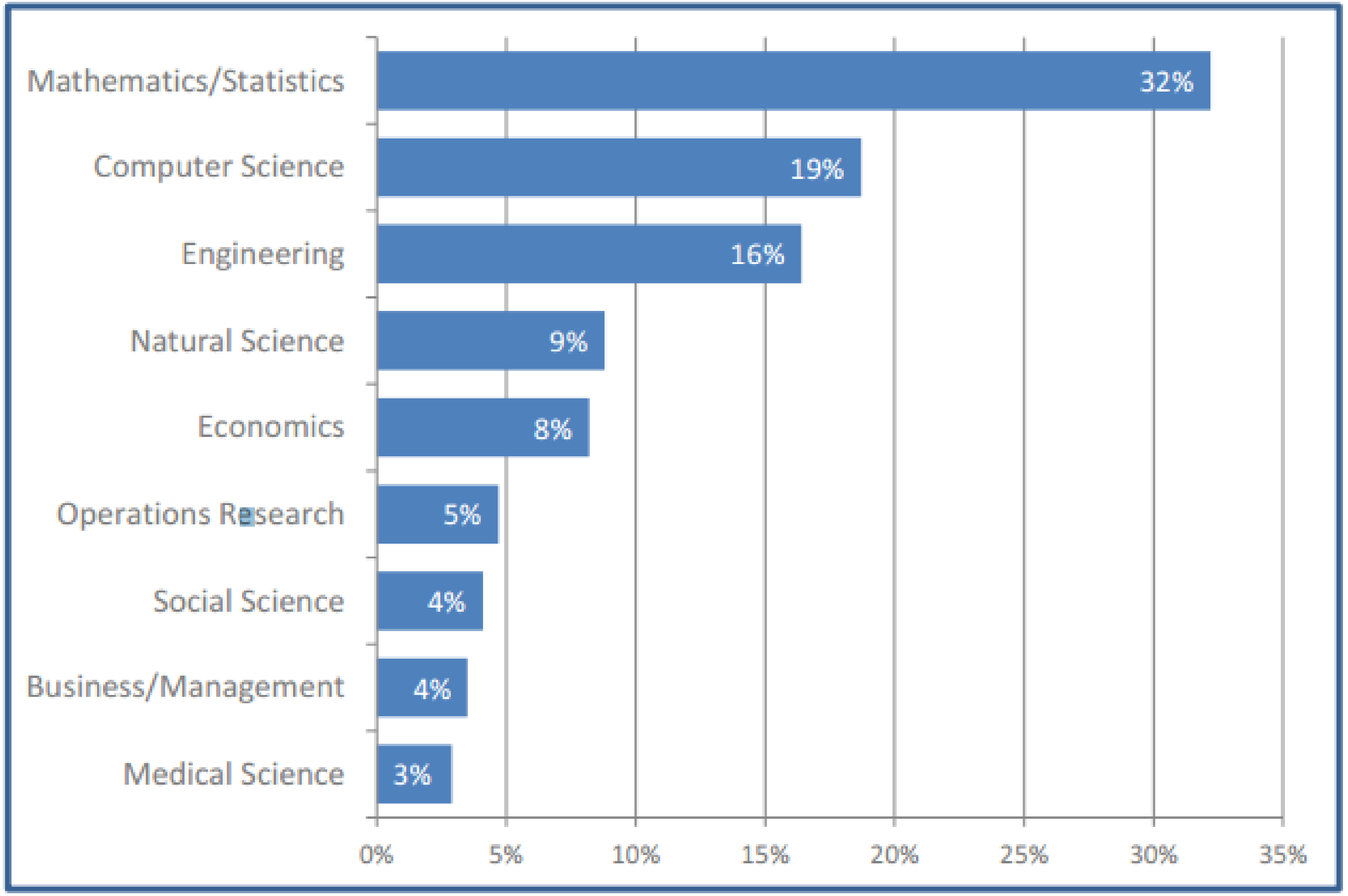


LATENT DIRICHLET ALLOCATION TOPIC MODEL OF
“DOING DATA SCIENCE: CHAPTER 1”

Topic 1	Topic 2	Topic 3
Scientist	Hype	Statistic
Social	Mean	Scientist
Student	Scientist	People
Academia	Teach	Job
Define	Statistician	Skill
Question	Term	Industry
Field	Course	Google
People	Feel	Profile
Sense	Machine	Team
Solve	Leaning	Product

WHO ARE DATA SCIENTISTS?

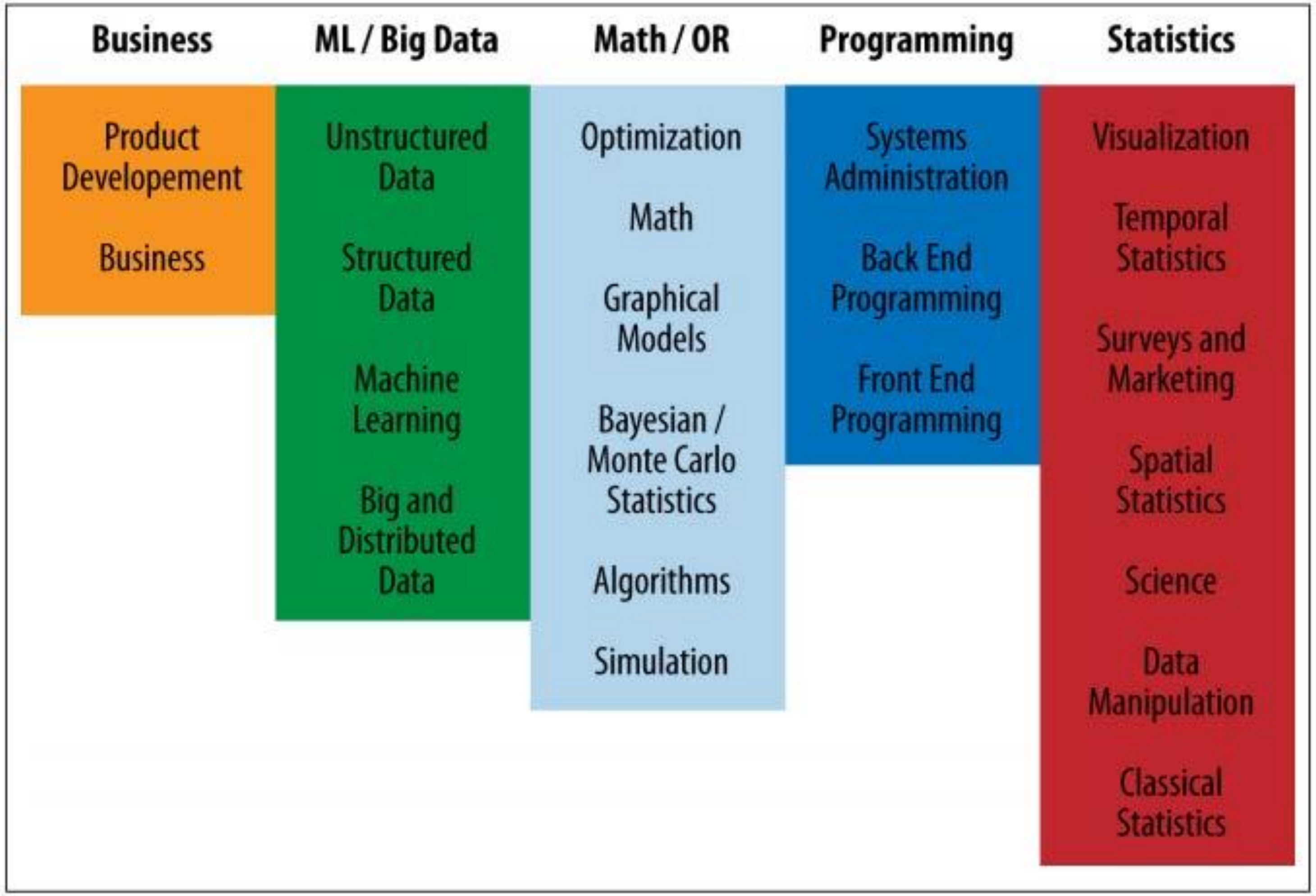
Figure 8. *Data Scientists by Area of Study*



Burtch Works Executive Recruiting study, 2014

WHAT ARE THE ROLES IN DATA SCIENCE?

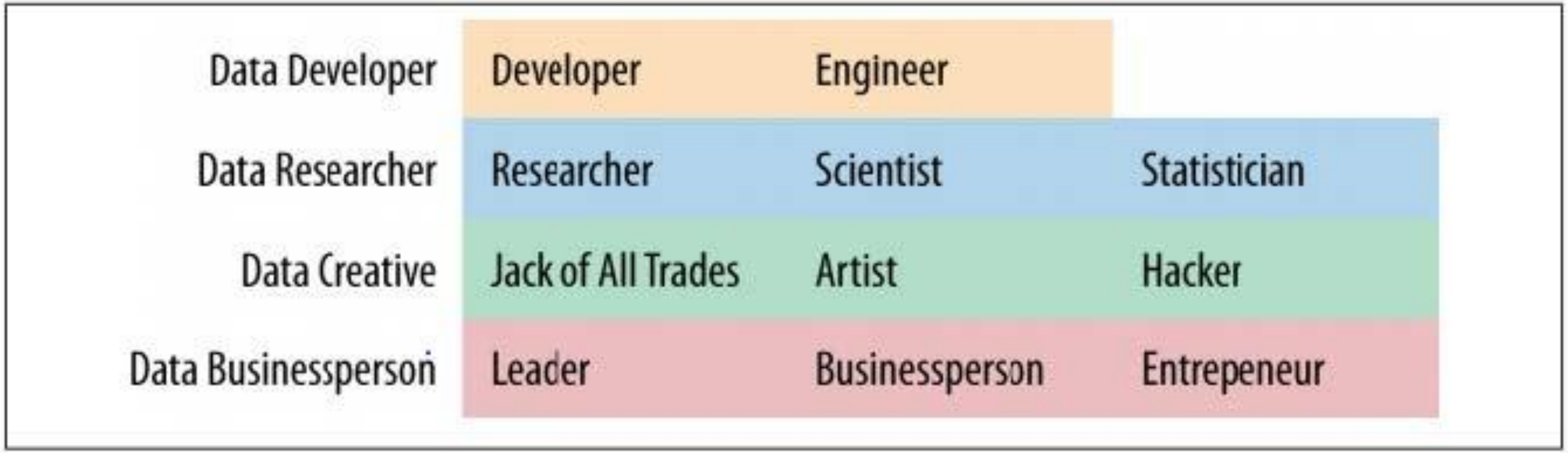
- Data Science involves a variety of skill sets, not just one.



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?

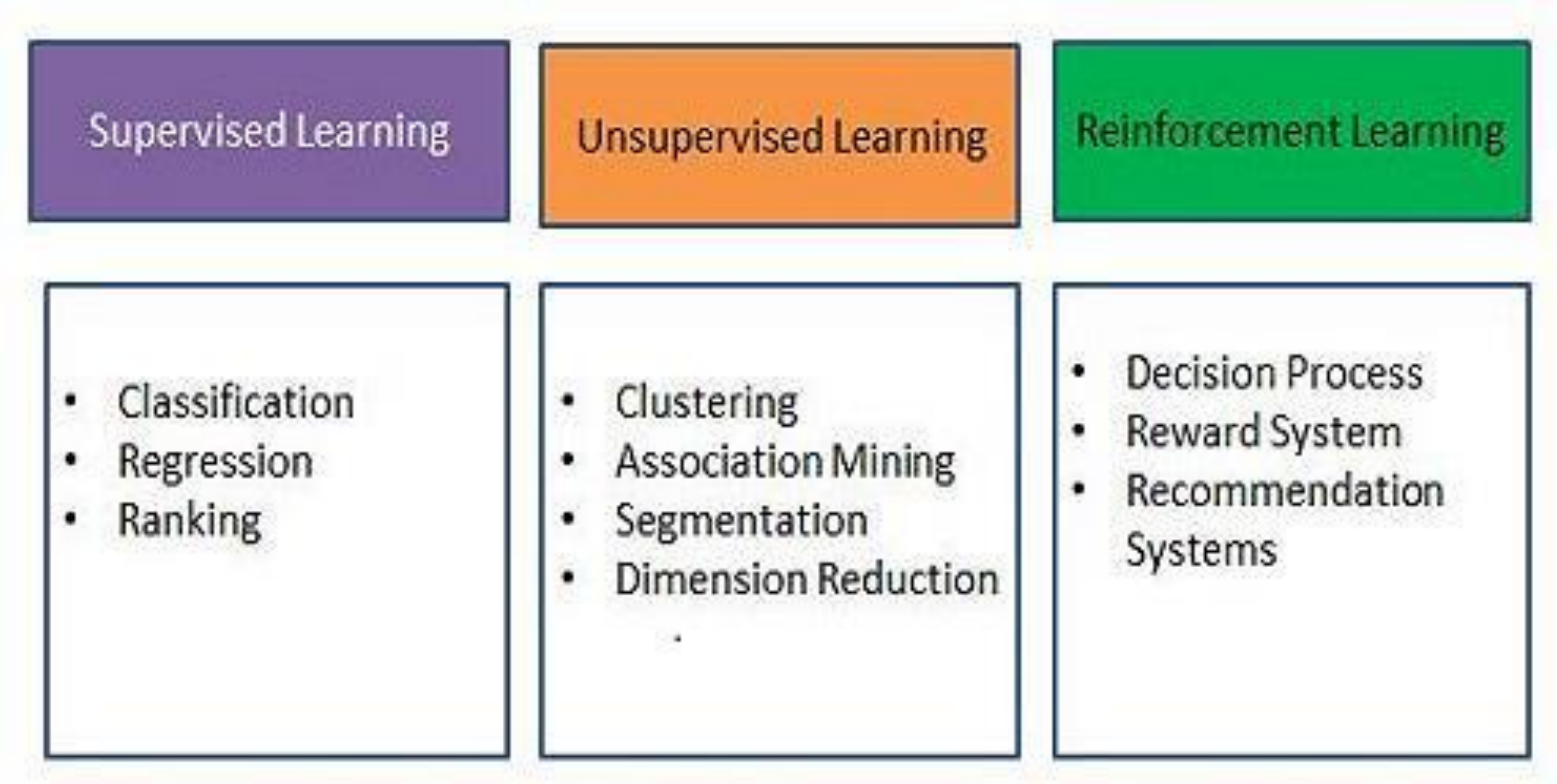
- Data Science involves a variety of roles, not just one.



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

WHAT 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

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Common questions answered by Data Scientists

  1. Is this A or B? (Classification / Binary Prediction)

  2. Is this A or B or C or D? (Recognition)

  3. Is this Unusual? (Anomaly Detection)

  4. How Much / How Many? (Regression / Quantative Prediction)

  5. How is this Data Organized? (Grouping / Dimension Reduction)

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this A or B? (Classification): Predict events that have two possible outcomes
  - Will this customer default on their loan?
  - Is this an image of a cat or a dog?
  - Will this customer click on the advertisement?
  - Will this team win the basketball game?
  - Is this mole malignant or benign?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this A or B or C or D? (Recognition): Predict which category a case belongs to
    - Which animal is in this image?
    - Which aircraft is causing this radar signature?
    - What is the topic of this news article?
    - What is the mood of this tweet?
    - Who is the speaker in this recording?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this Unusual? (Anomaly Detection): Determine if a phenomenon deviates from an expected range
    - Is this pressure reading unusual?
    - Is this internet message typical?
    - Is this combination of purchases very different from what this customer has made in the past?
    - Are these weather patterns normal for this century?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- How Much / How Many? (Prediction): Predict a quantitative outcome
  - What will the temperature be next Tuesday?
  - What will my fourth quarter sales in Portugal be?
  - How many kilowatts will be demanded from my wind farm 30 minutes from now?
  - How many new followers will I get next week?

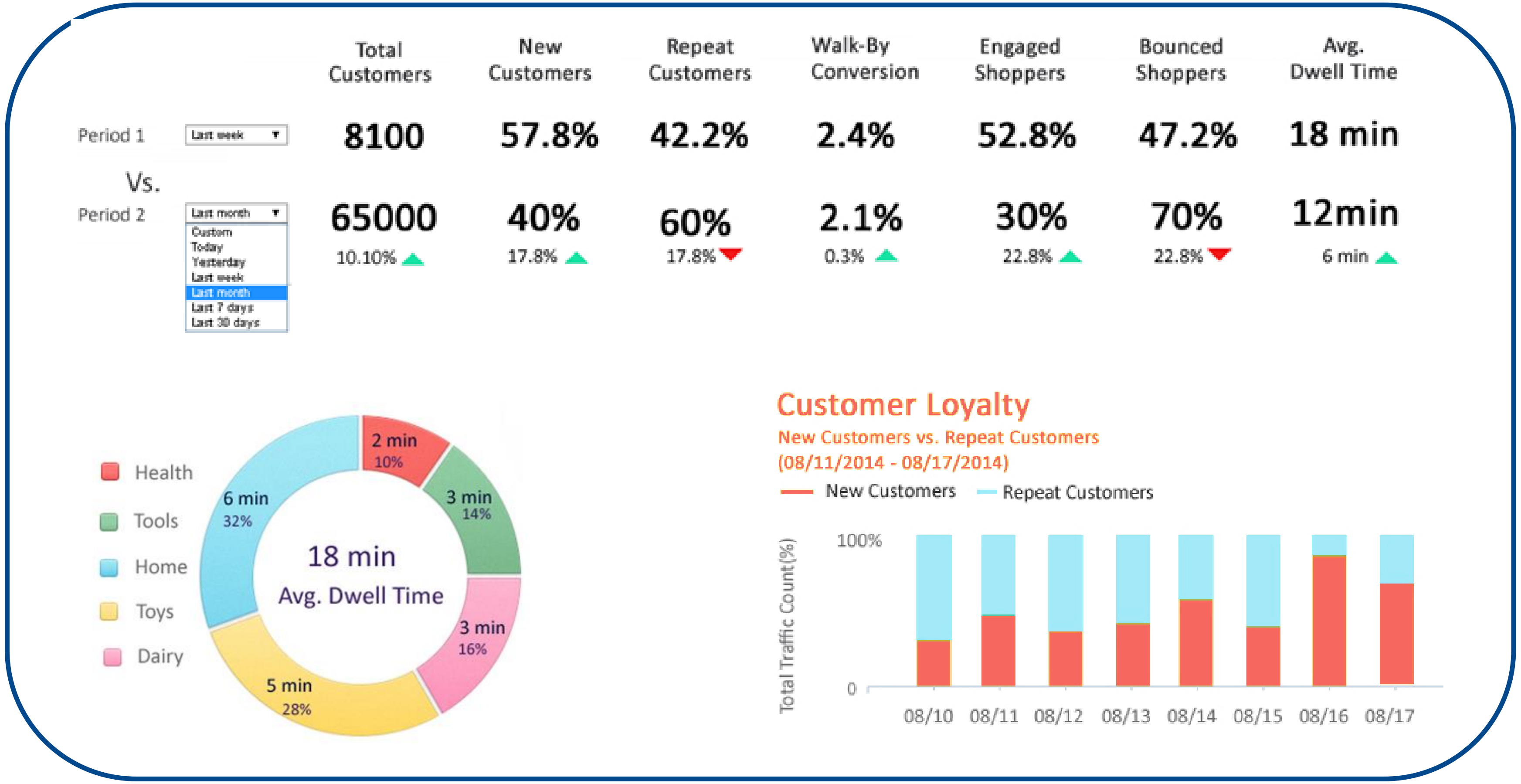
WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- How is this Data Organized? (Grouping): What are the categories or smaller dimensions within the data.
  - What are the different types of coffee drinkers?
  - Which viewers like the same kind of movies?
  - What kinds of car models does GM produce?
  - Are there common clusters of cable channels that customers tend to purchase together
  - What is a natural way to break these documents into five topics?

HOW DOES WAITBOT USE DATA SCIENCE?

Crowd Analytics

Store Statistics 08/11/2014 - 08/17/2014



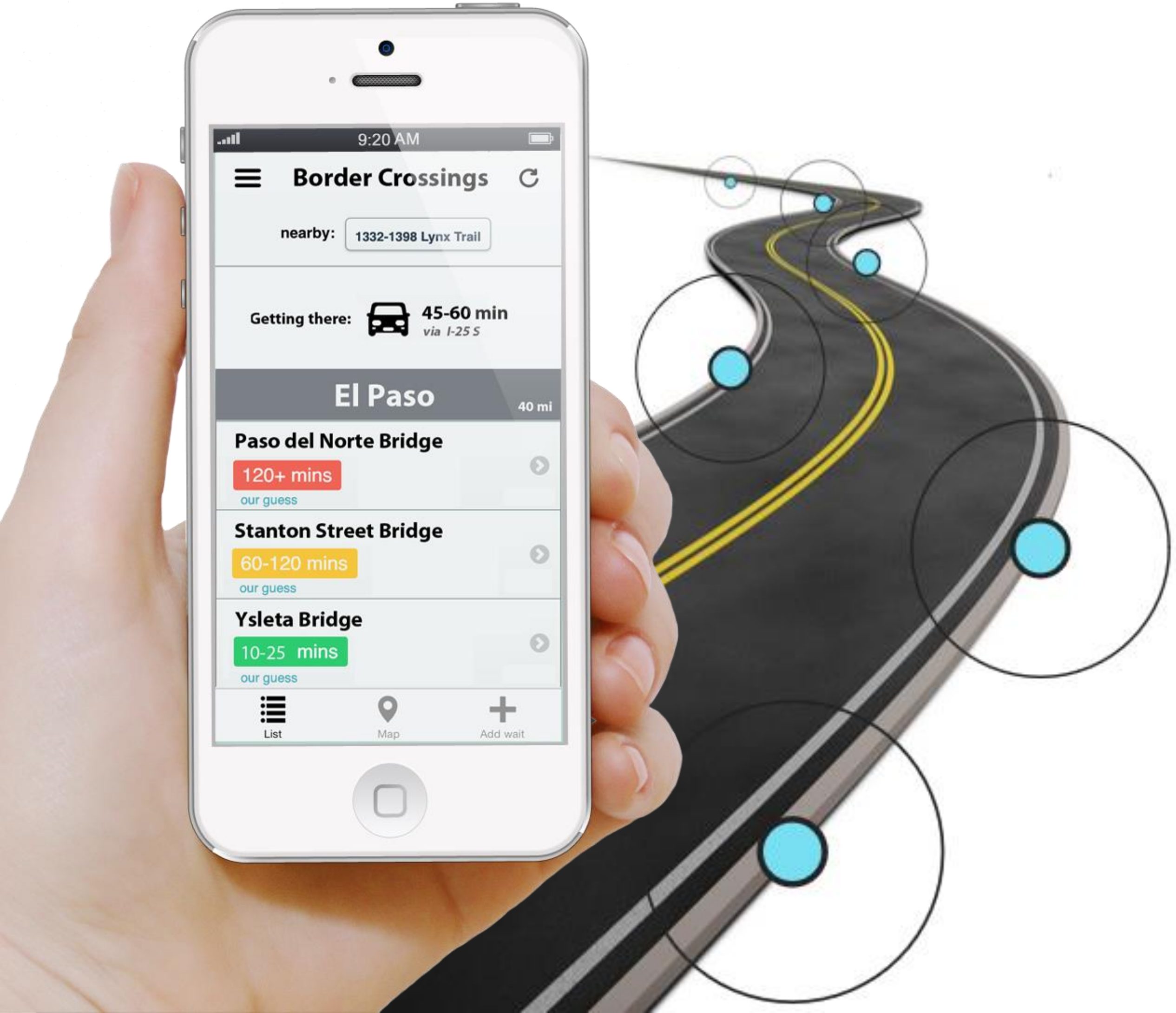
	Total Customers	New Customers	Repeat Customers	Walk-By Conversion	Engaged Shoppers	Bounced Shoppers	Avg. Dwell Time
Period 1 Last week ▼	8100	57.8%	42.2%	2.4%	52.8%	47.2%	18 min
Vs.							
Period 2 Last month ▼	65000	40%	60%	2.1%	30%	70%	12min
	10.10% ▲	17.8% ▲	17.8% ▼	0.3% ▲	22.8% ▲	22.8% ▼	6 min ▲

Customer Loyalty

New Customers vs. Repeat Customers (08/11/2014 - 08/17/2014)

HOW DOES WAITBOT USE DATA SCIENCE

Wait Time Predications



9:20 AM

Border Crossings

nearby: 1332-1398 Lynx Trail

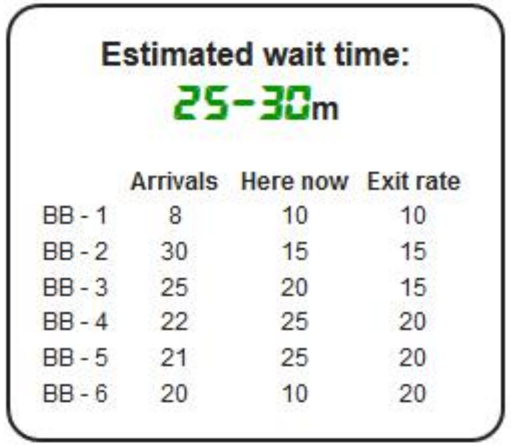
Getting there: 45-60 min via I-25 S

El Paso 40 mi

Paso del Norte Bridge
120+ mins
our guess

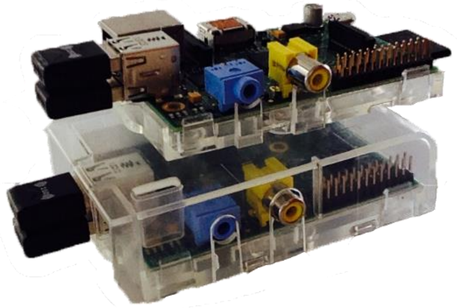
Stanton Street Bridge
60-120 mins
our guess

Ysleta Bridge
10-25 mins
our guess



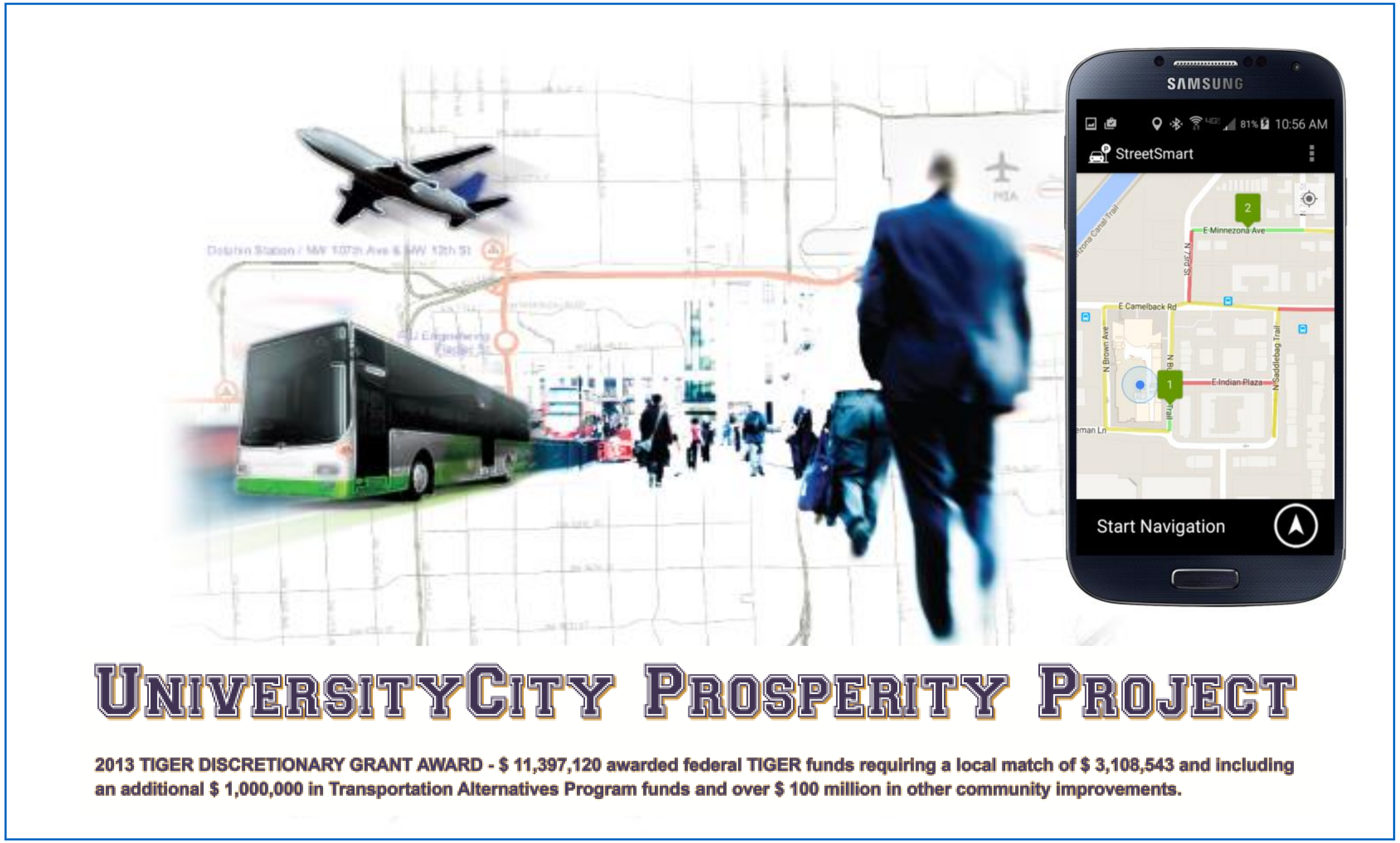
Estimated wait time:
25-30m

	Arrivals	Here now	Exit rate
BB - 1	8	10	10
BB - 2	30	15	15
BB - 3	25	20	15
BB - 4	22	25	20
BB - 5	21	25	20
BB - 6	20	10	20



HOW DOES WAITBOT USE DATA SCIENCE

Parking Availability



UNIVERSITYCITY PROSPERITY PROJECT

2013 TIGER DISCRETIONARY GRANT AWARD - $ 11,397,120 awarded federal TIGER funds requiring a local match of $ 3,108,543 and including an additional $ 1,000,000 in Transportation Alternatives Program funds and over $ 100 million in other community improvements.

WAITBOT'S TECHNOLOGY

Tracking sensors

App crowdsourcing

Big data analytics

CLASS ACTIVITY

DATA SCIENCE QUESTIONS

ACTIVITY: DATA SCIENCE QUESTIONS

EXERCISE

DIRECTIONS (10 minutes)

1. Break into pairs (person sitting next to you)

2. Pick a topic of interest to the both of you (e.g., music, finance, psychology, retail)

3. For each of the 5 kinds of data science questions, come up with a specific question you could ask for that topic.

    - Is this A or B? (Classification)

    - Is this A or B or C or D? (Recognition)

    - Is this Unusual? (Anomaly Detection)

    - How Much / How Many? (Prediction)

    - How is this Data Organized? (Grouping / Dimension Reduction)

WHO USES DATA SCIENCE?

NETFLIX

amazon.com

Google

FiveThirtyEight

DRAFT KINGS

WHO USES DATA SCIENCE?

- Can you think of others?

INTRODUCTION

THE DATA SCIENCE WORKFLOW

OVERVIEW OF THE DATA SCIENCE WORKFLOW

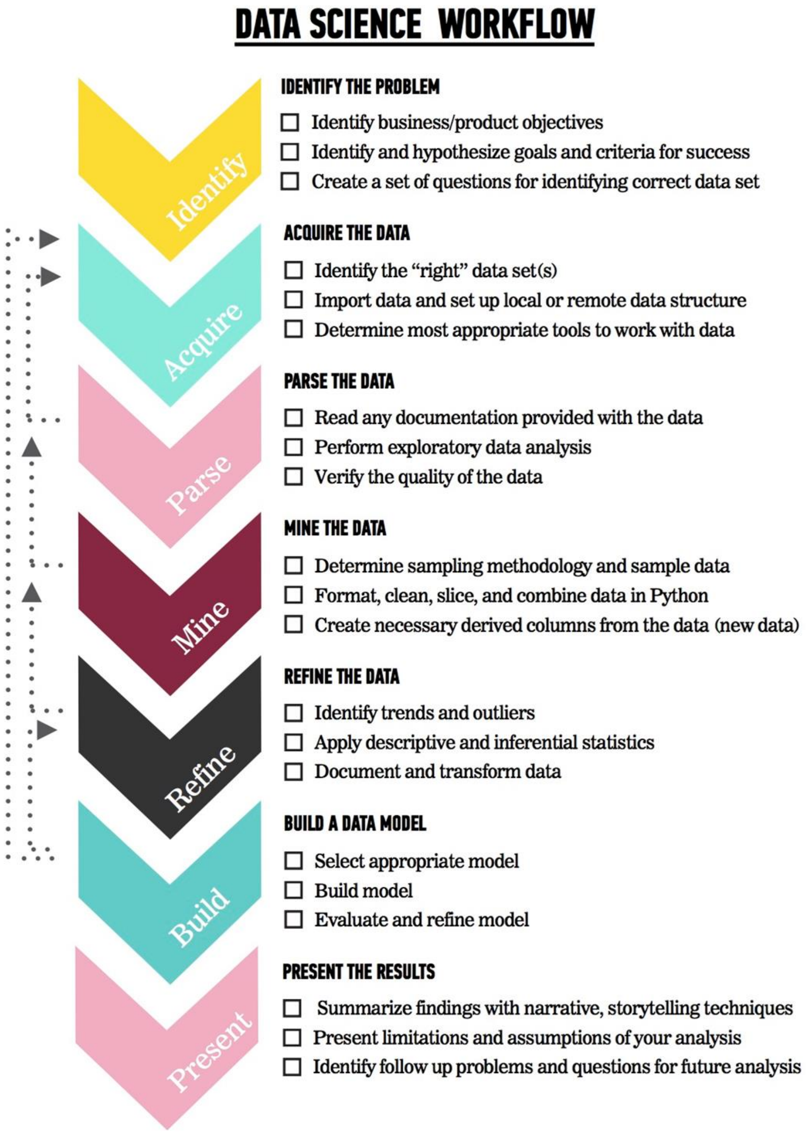
- What process does a data scientist follow?
- Similar to the scientific method
- Helps produce *accurate* and *reproducible* results
  - *Accurate*: Describes a true consistent phenomenon or finding
  - *Reproducible*: Others can follow your steps and get the same results

https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times/?utm_term=.0c5e0e8fe211

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

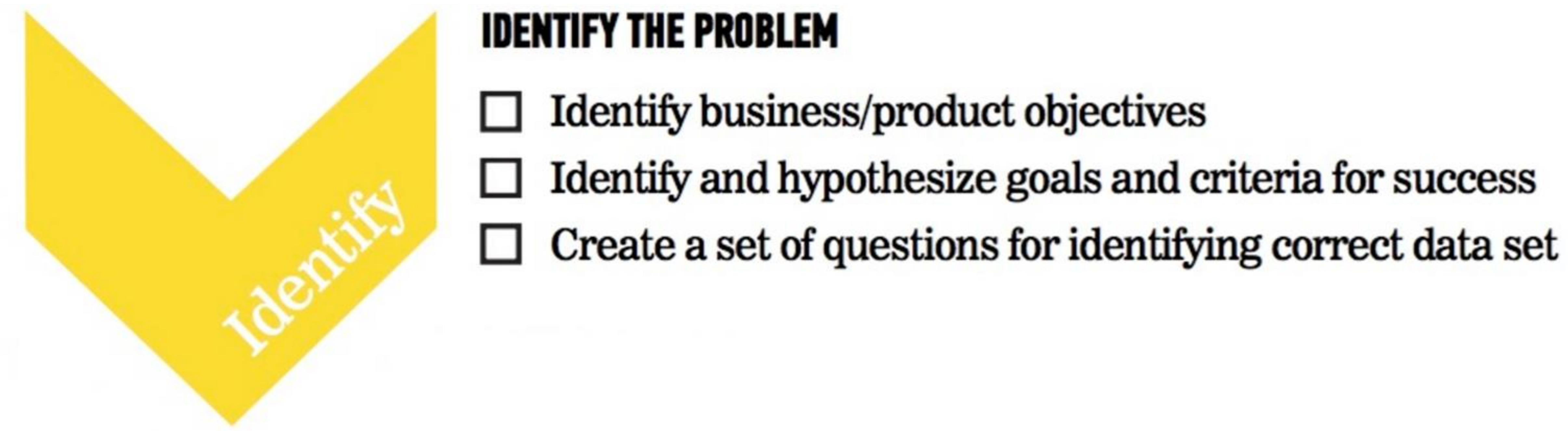
BUILD A DATA MODEL

□ 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

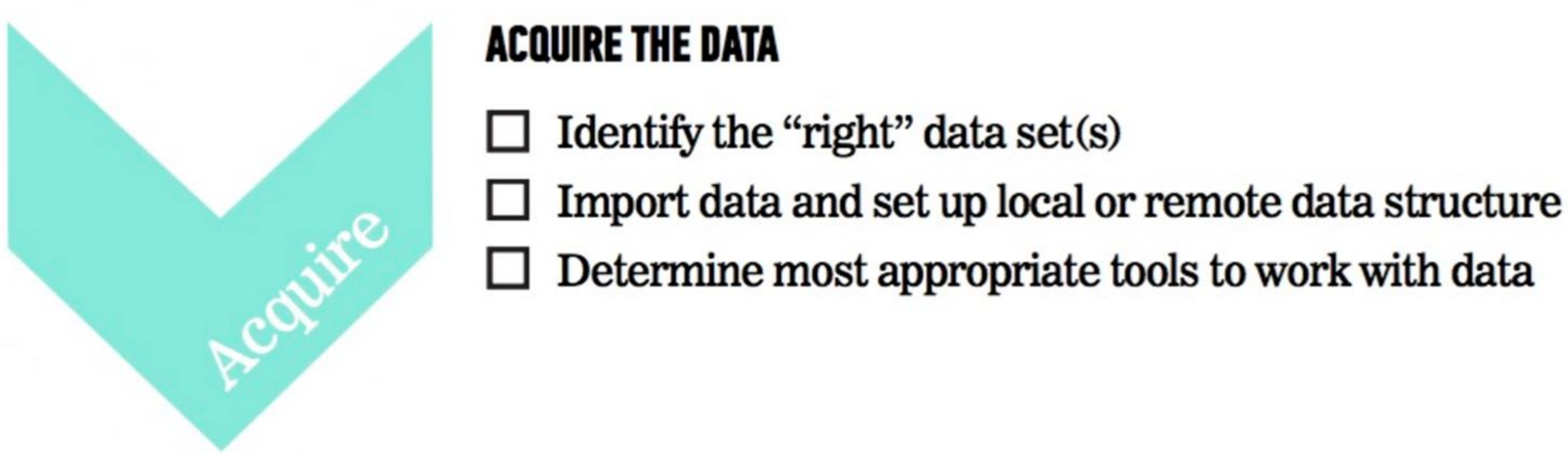
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

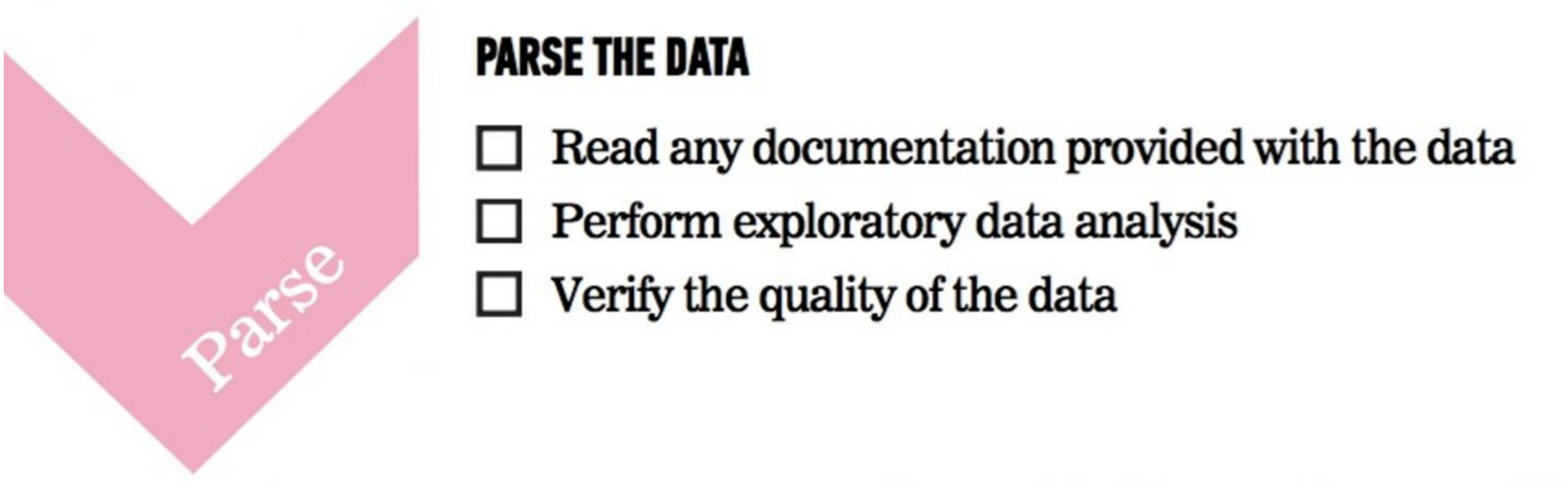
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

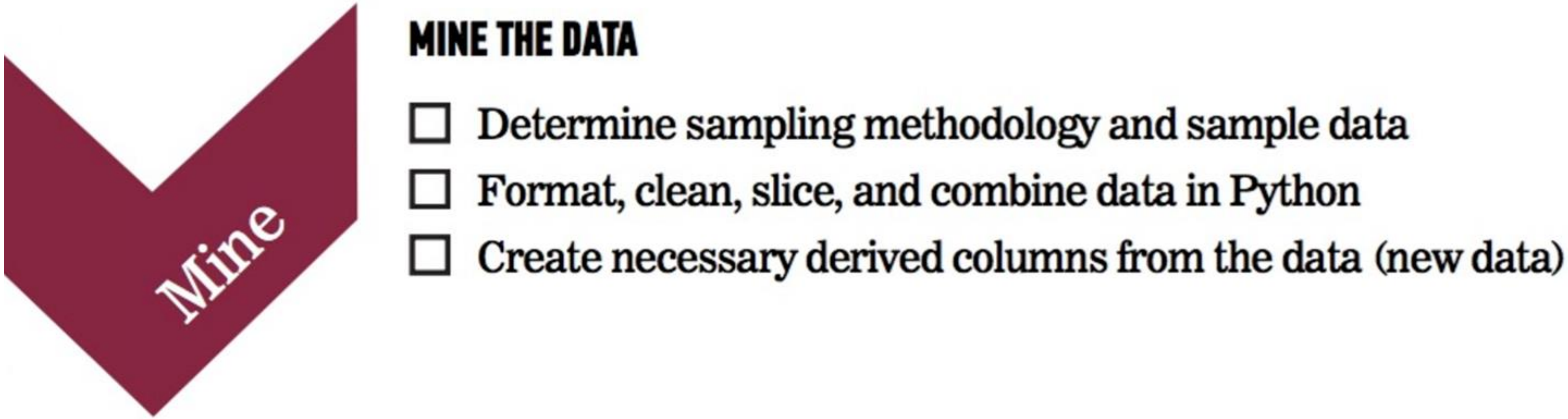
OVERVIEW OF THE DATA SCIENCE WORKFLOW



PARSE THE DATA

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

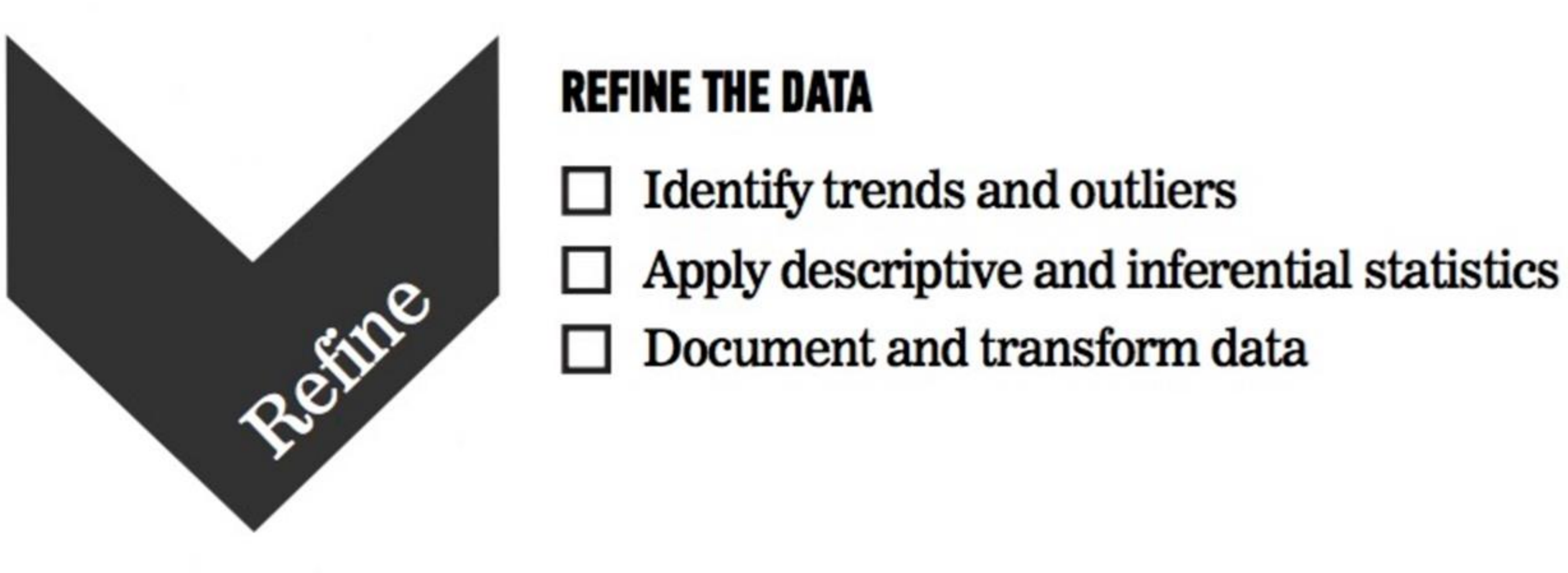
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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



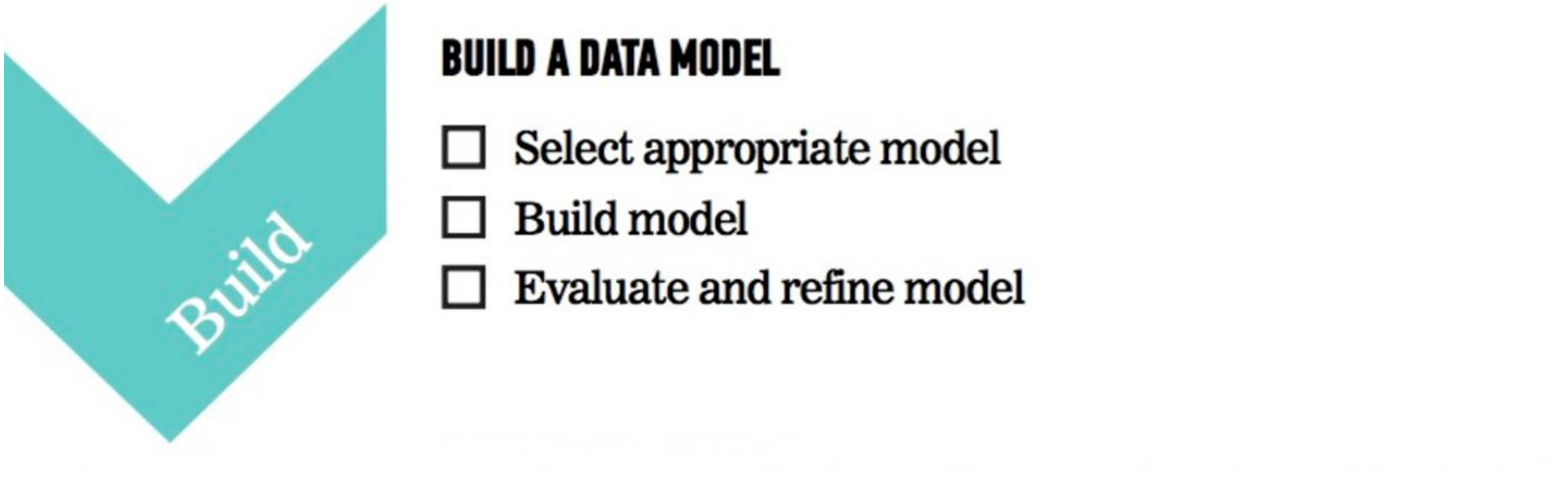
REFINE THE DATA

□ Identify trends and outliers

□ Apply descriptive and inferential statistics

□ Document and transform data

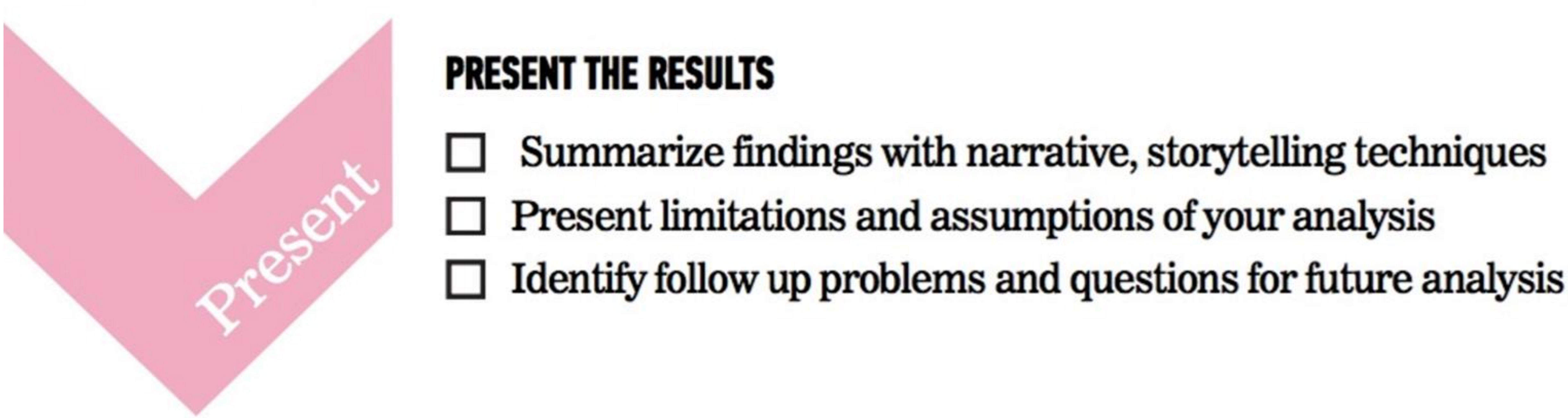
OVERVIEW OF THE DATA SCIENCE WORKFLOW



BUILD A DATA MODEL

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

OVERVIEW OF THE DATA SCIENCE WORKFLOW



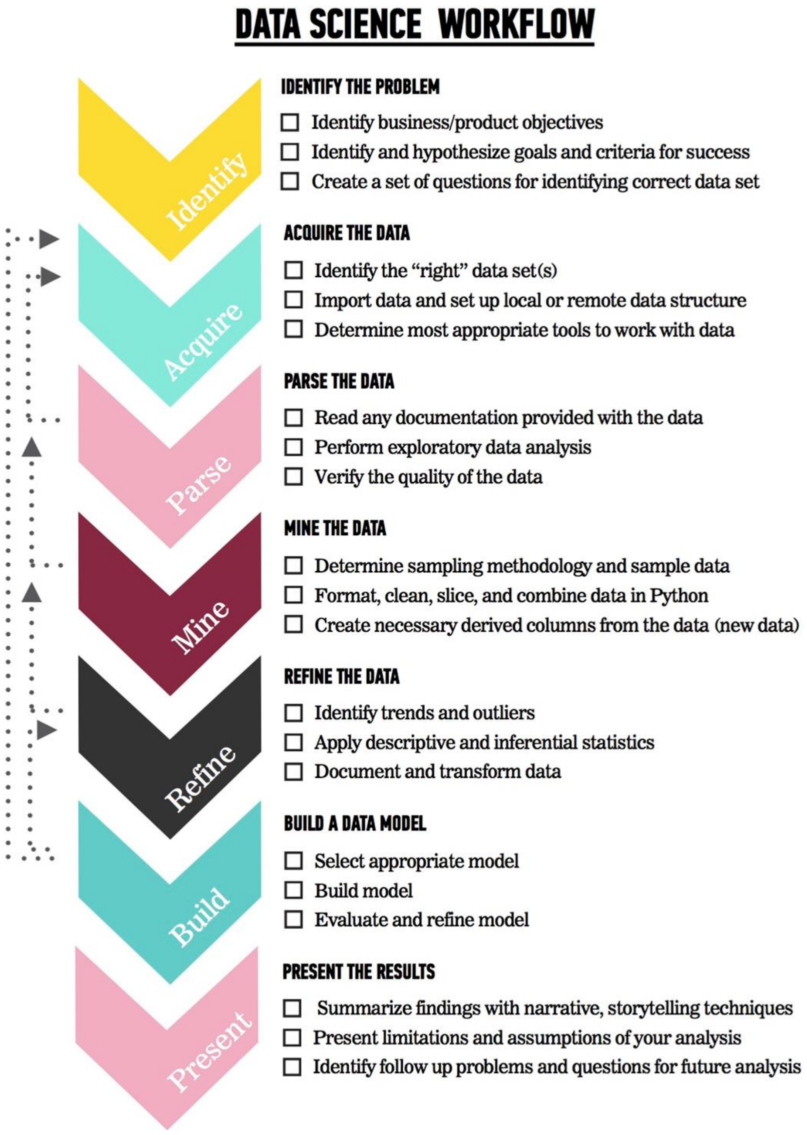
PRESENT THE RESULTS

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

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

INTRODUCTION

THE DATA SCIENCE WORKFLOW: NETFLIX EXAMPLE

LEARNING OBJECTIVES

- Define data science and the data science workflow
- Apply the data science workflow
- Setup your development environment and review python basics

NETFLIX EXAMPLE

- Problem Statement: In 2006, Netflix Prize held a competition, open to anyone, to develop an algorithm that could predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

- Netflix offered a $1 million prize to any person/team that could improve the accuracy of its own recommendation engine by at least 10%

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

NETFLIX 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.

NETFLIX 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?

NETFLIX 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?

NETFLIX EXAMPLE: PARSE THE DATA

- Secondary data = we didn’t directly collect it ourselves
- Example data dictionary

Variable	Description	Format
MovieID	A unique number indicating the movie	Categorical: Integer
CustomerID	A unique number indicating the customer who rated the movie	Categorical: Integer
Rating	Number of ‘stars’ assigned to a movie by a customer; integer from 1-5	Continuous: Integer
Title	English Language Title	Categorical: String
YearofRelease	Year a movie was released in the range [1890..2005].	Continuous: Integer

NETFLIX 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?

NETFLIX EXAMPLE: MINE THE DATA

- Think about sampling
- Get to know the data
- Explore outliers
- Address missing values
- Derive new variables (i.e. columns)

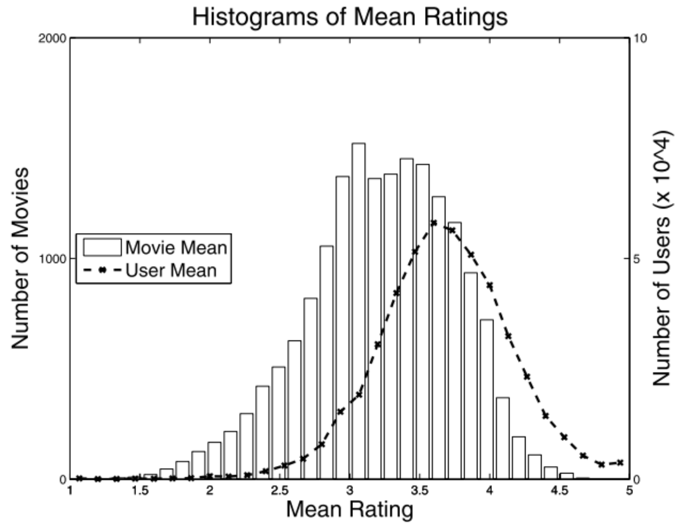
NETFLIX EXAMPLE: MINE THE DATA

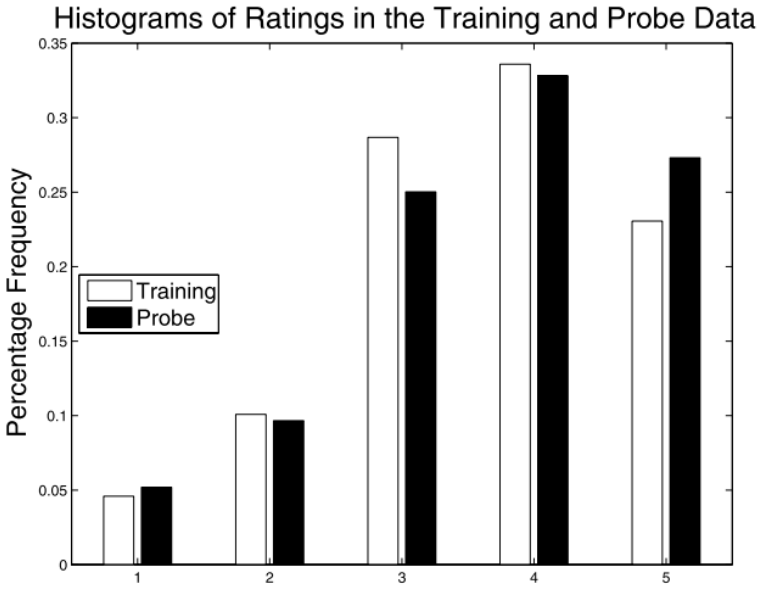
- Common steps while mining the data
    - Sample the data with appropriate methodology
    - Explore outliers and null values
    - Format and clean the data
    - Determine how to address missing values
    - Format and combine data; aggregate and derive new columns

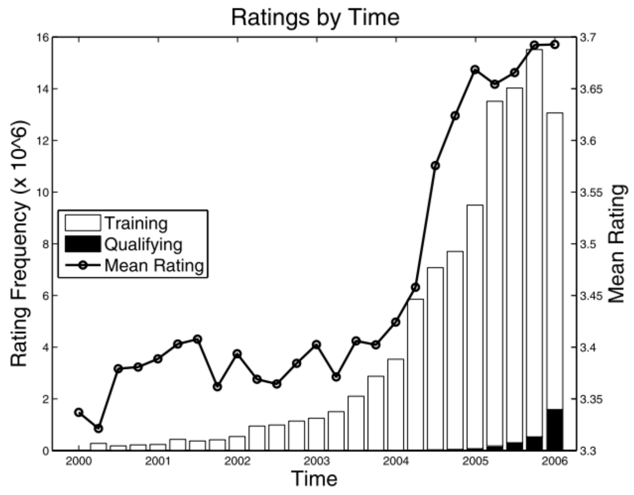
NETFLIX EXAMPLE: REFINE THE DATA

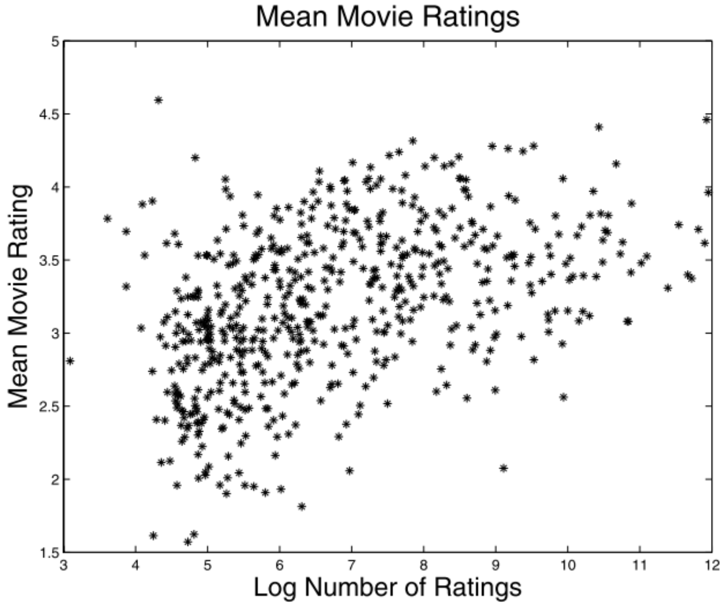
- Use descriptive statistics (mean, mode, standard deviation) to help:
    - Identifying trends and outliers
    - Deciding how to deal with outliers
    - Applying descriptive and inferential statistics
    - Determining visualization techniques for different data types
    - Transforming data

NETFLIX EXAMPLE: REFINE THE DATA









NETFLIX EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome

- Example models:
    - Linear regression where we predict a user’s movie rating using release date, the movie’s average rating, and user’s average rating

    - Decision tree where we predict if a movie will receive 5 stars from a user based on the number of 5 stars a user has already given

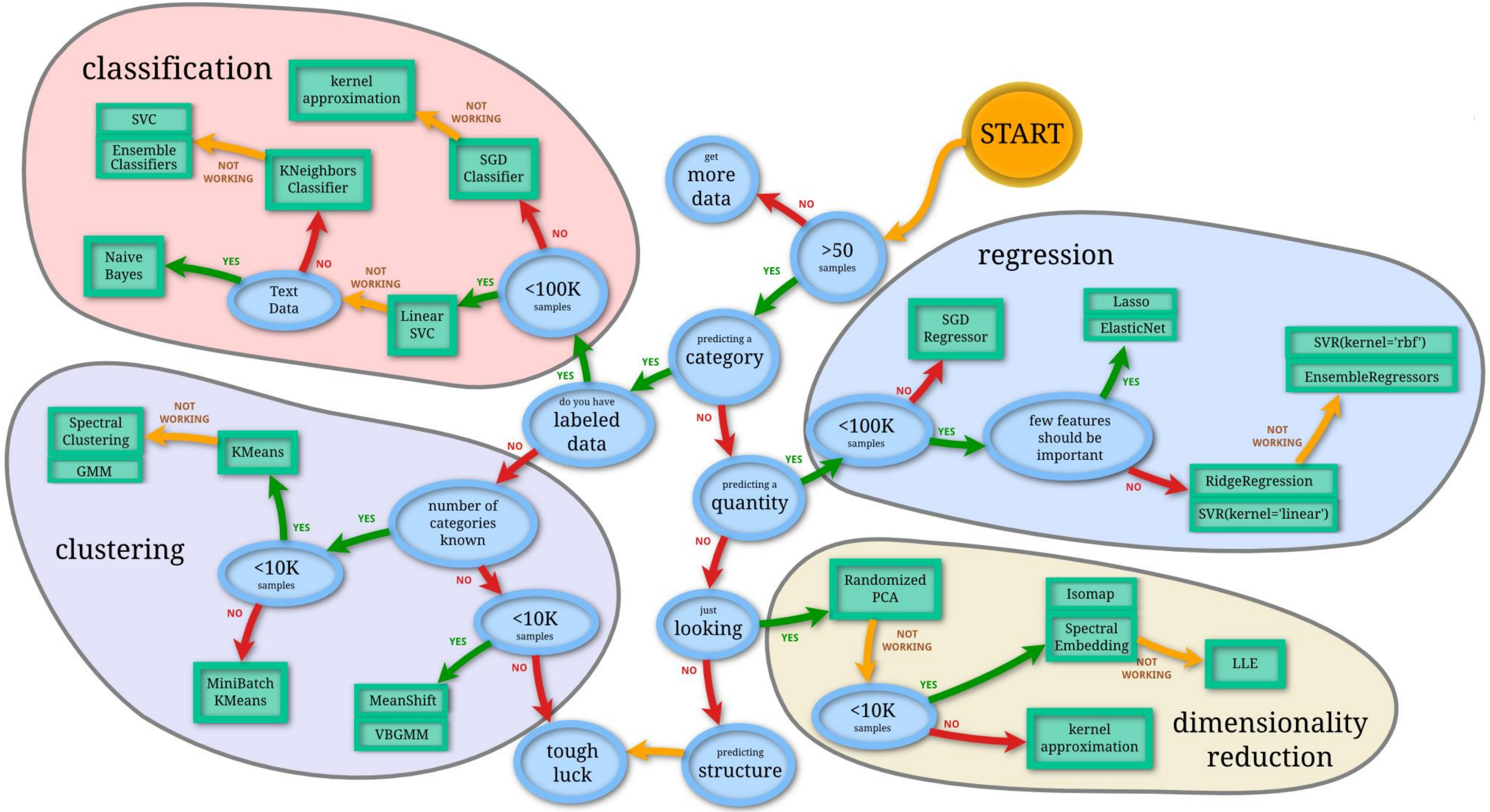
    - K-Nearest Neighbor where we predict what rating a movie will receive based on the rating of similarly titled movies

- Steps for model building

NETFLIX EXAMPLE: CREATE A DATA MODEL

- The steps for model building are:
  - Select the appropriate model
    - Depends on many factors (type of research question, type of outcome, type of predictors, number of variables, number of cases)
  - Build the model
    - Select variables and parameters that go into the model
  - Evaluate and refine the model
    - See how model performs on a sample of data set aside, and make changes to improve performance
  - Predict outcomes and action items

NETFLIX EXAMPLE: SELECT THE APPROPRIATE MODEL



NETFLIX EXAMPLE: PRESENT THE RESULTS

- You have to effectively communicate your results for them to matter!

- Make sure to consider your audience.

- A presentation for fellow data scientists will be drastically different from a presentation for an executive.

NETFLIX 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

NETFLIX EXAMPLE: PRESENT THE RESULTS

- Example presentations and infographics
  - http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=8901D4FBEB41F4E5670517227ABC92DD?doi=10.1.1.142.9009&rep=rep1&type=pdf
  - http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf

DEMO

ENVIRONMENT SETUP

LEARNING OBJECTIVES

- Define data science and the data science workflow
- Apply the data science workflow
- Setup your development environment and review python basics

GIT – VERSION CONTROL

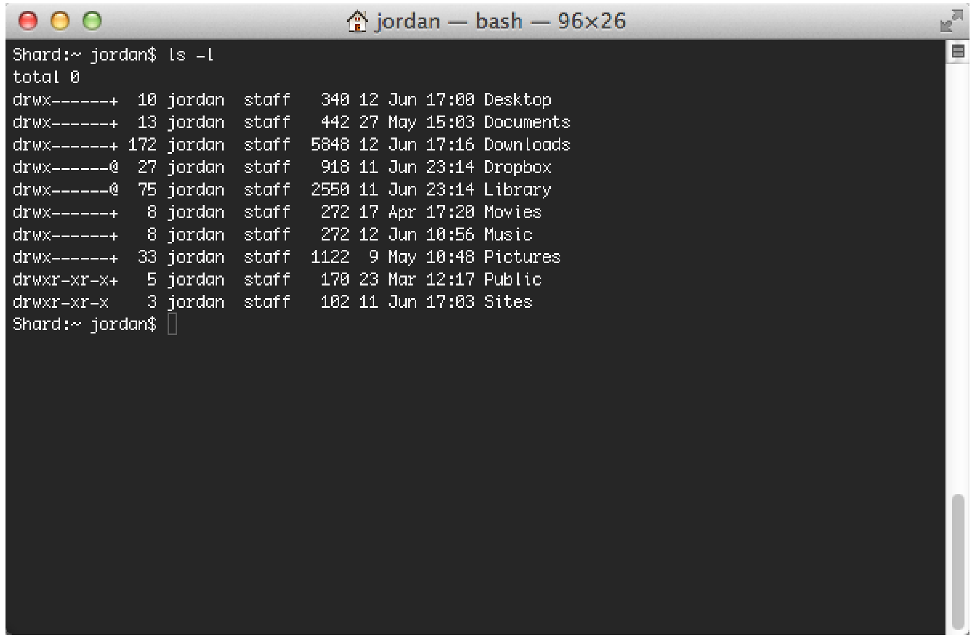
Got 15 minutes and want to learn about Git

- Version control is necessary when working on complex projects.
- Git is a way of tracking changes we've made to our programs and go back in time to fix errors.
- It is also a powerful tool for collaborating with colleagues allowing you to work on different aspects of the project simultaneously and merge all the changes together seamlessly
- There are many different ways to use these tools

TERMINAL / COMMAND PROMPT

Code Academy Command Line

- Download python packages
- Run python scripts and applications
- Connect to external data sources



```
Shard:~ jordan$ ls -l
total 0
drwx------+  10 jordan  staff   340 12 Jun 17:00 Desktop
drwx------+  13 jordan  staff   442 27 May 15:03 Documents
drwx------+ 172 jordan  staff  5848 12 Jun 17:16 Downloads
drwx------@  27 jordan  staff   918 11 Jun 23:14 Dropbox
drwx------@  75 jordan  staff  2550 11 Jun 23:14 Library
drwx------+   8 jordan  staff   272 17 Apr 17:20 Movies
drwx------+   8 jordan  staff   272 12 Jun 10:56 Music
drwx------+  33 jordan  staff  1122  9 May 10:48 Pictures
drwxr-xr-x+   5 jordan  staff   170 23 Mar 12:17 Public
drwxr-xr-x    3 jordan  staff   102 11 Jun 17:03 Sites
Shard:~ jordan$ 
```

- Common commands.

  cd

  pwd

  $home

  mkdir

  open

DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
  - Create a Github account (for homework)
  - Install Python 2.7 and Anaconda
  - Practice Python syntax, Terminal commands, and Pandas
- Jupyter (formally iPython) Notebook test and Python review

PYTHON – MANY THINGS TO MANY PEOPLE

- A productivity tool for data extraction and manipulation
- A data analysis and modeling toolkit
- A mobile or web application backend
- A first programing language
- A governing philosophy, seriously….

PYTHON – THE ZEN OF PYTHON

https://zen-of-python.info/

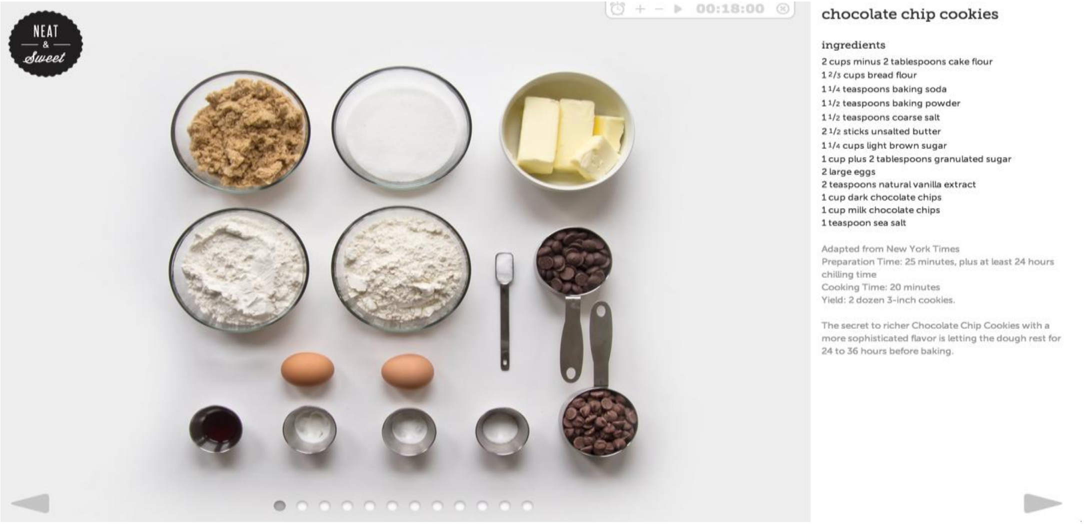
- Beautiful is better than ugly
- Simple is better than complex
- Complex is better than complicated
- Special cases aren't special enough to break the rules
- Although practicality beats purity
- In the face of ambiguity, refuse the temptation to guess
- Now is better than never
- Although never is often better than right now
- If the implementation is hard to explain, it's a bad idea
- If the implementation is easy to explain, it may be a good idea

PYTHON vs. OTHER LANGUAGES

- Initiative syntax, easy to read
- Large development community (i.e. you can google your problems)
- Plethora of open source packages/libraries
- Key technical considerations
    - Object oriented
    - Dynamic typing (i.e. don’t need to declare variable type)
    - Not as fast as C++, Java, but typically code is 3-5 shorter

PYTHON – SCRIPTS

- A Python script is a set of instructions (you write) which tell a computer what to do
- Scripts are typically "simple", have a start and finish time, and are started by a user or scheduled process



chocolate chip cookies

ingredients

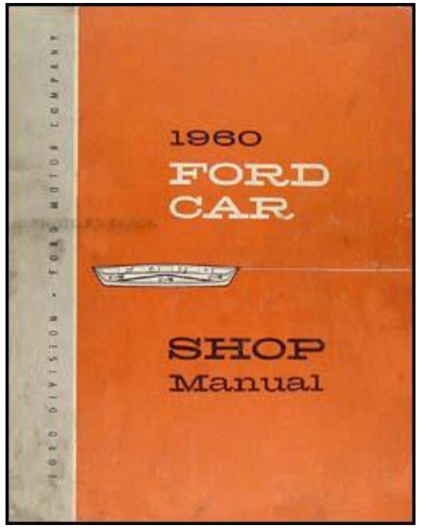
2 cups minus 2 tablespoons cake flour
1 2/3 cups bread flour
1 1/4 teaspoons baking soda
1 1/2 teaspoons baking powder
1 1/2 teaspoons coarse salt
2 1/2 sticks unsalted butter
1 1/4 cups light brown sugar
1 cup plus 2 tablespoons granulated sugar
2 large eggs
2 teaspoons natural vanilla extract
1 cup dark chocolate chips
1 cup milk chocolate chips
1 teaspoon sea salt

Adapted from New York Times
Preparation Time: 25 minutes, plus at least 24 hours chilling time
Cooking Time: 20 minutes
Yield: 2 dozen 3-inch cookies.

The secret to richer Chocolate Chip Cookies with a more sophisticated flavor is letting the dough rest for 24 to 36 hours before baking.

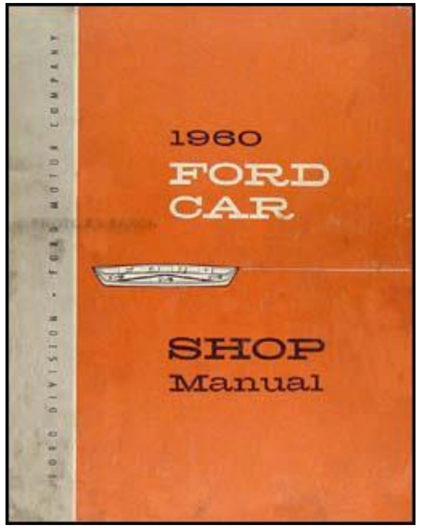
PYTHON – APPLICATIONS

- A Python application or program is an “instruction manual” capable of handling many situations.
- It is generally more complex than a script, comprises of multiple components, can run indefinitely with multiple user interactions.



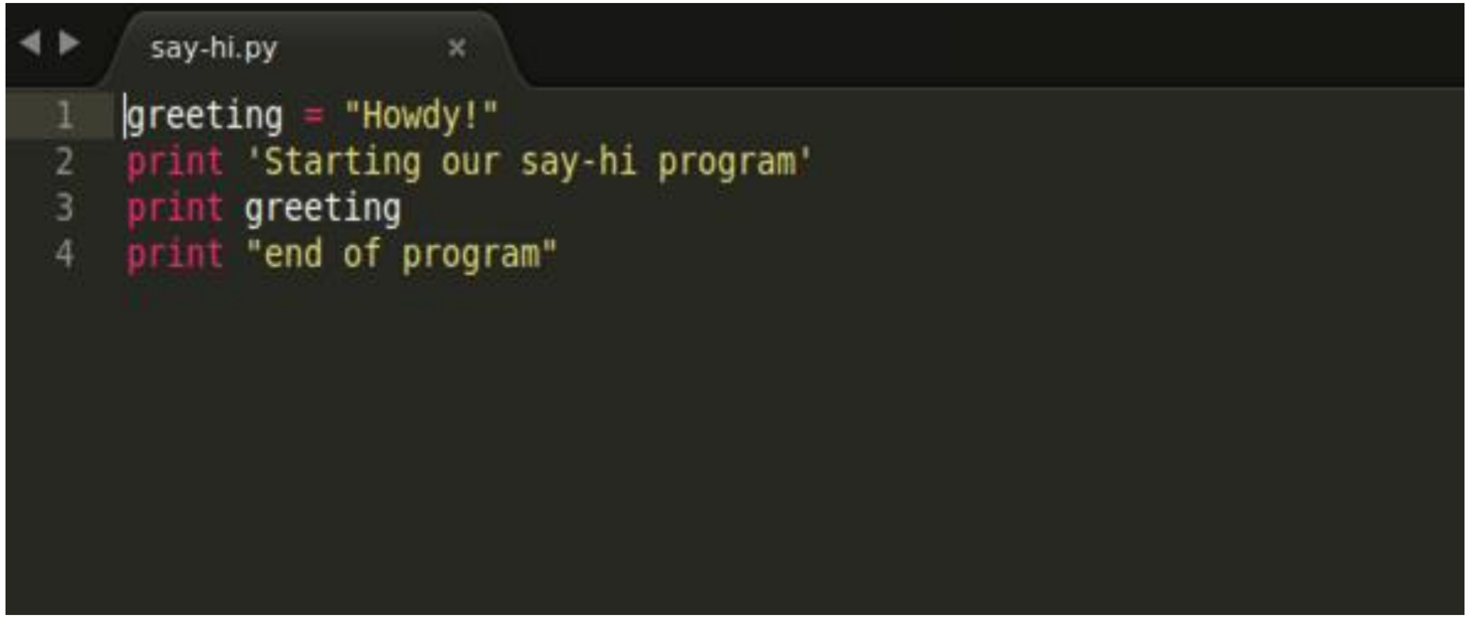
PYTHON – APPLICATIONS

- A Python application or program is an “instruction manual” capable of handling many situations.
- It is generally more complex than a script, comprises of multiple components, can run indefinitely with multiple user interactions.



WHERE DO I WRITE PYTHON?

- Any text editor technically, but we recommend using a text editor that highlights syntax. This makes debugging significantly easier



```
greeting = "Howdy!"
print 'Starting our say-hi program'
print greeting
print "end of program"
```

WHAT IS A PYTHON PACKAGE?

- A Python package (aka library) is collection of reusable code that can be added to your python scripts and applications.
- To download Python packages, we use a tool called pip

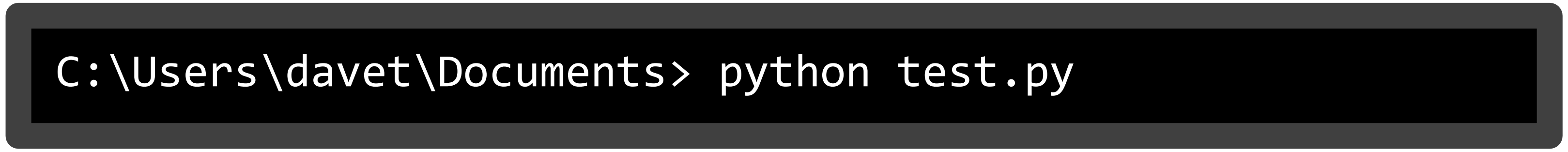
```
C:\Users\davet\Documents> pip install pandas
```

- After downloading, we use the import statement to include the package in our code

```
import pandas as pd
…
…
…
pd.read_excel(file_name_and_path, sheetname = worksheet)
```

HOW DO I EXECUTE MY PYTHON CODE?

- There are several ways to execute Python code
- Via command prompt

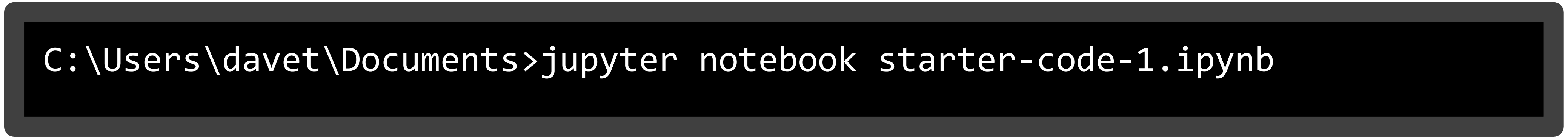


```
C:\Users\davet\Documents> python test.py
```

- Via an Integrated Development Environment
- Automatically via Windows Task Scheduler, Cron, etc.

JUPYTER (formally iPython) – A TOOL FOR LEARNING

- A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
- Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects."



```
C:\Users\davet\Documents>jupyter notebook starter-code-1.ipynb
```

DEV ENVIRONMENT SETUP

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

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

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?

DATA SCIENCE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

- Create Github account for uploading projects
- Read through final project instructions and start thinking about topic

Q & A

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET LINK:
https://docs.google.com/forms/d/1z2zkzWNeO2su32CDO9WQBnvHcpOslY9OF8IE_finfro/viewform?edit_requested=true#start=invite