GENERAL ASSEMBLY

WELCOME TO THE DATA SCIENCE PART TIME COURSE

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 programming basics

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

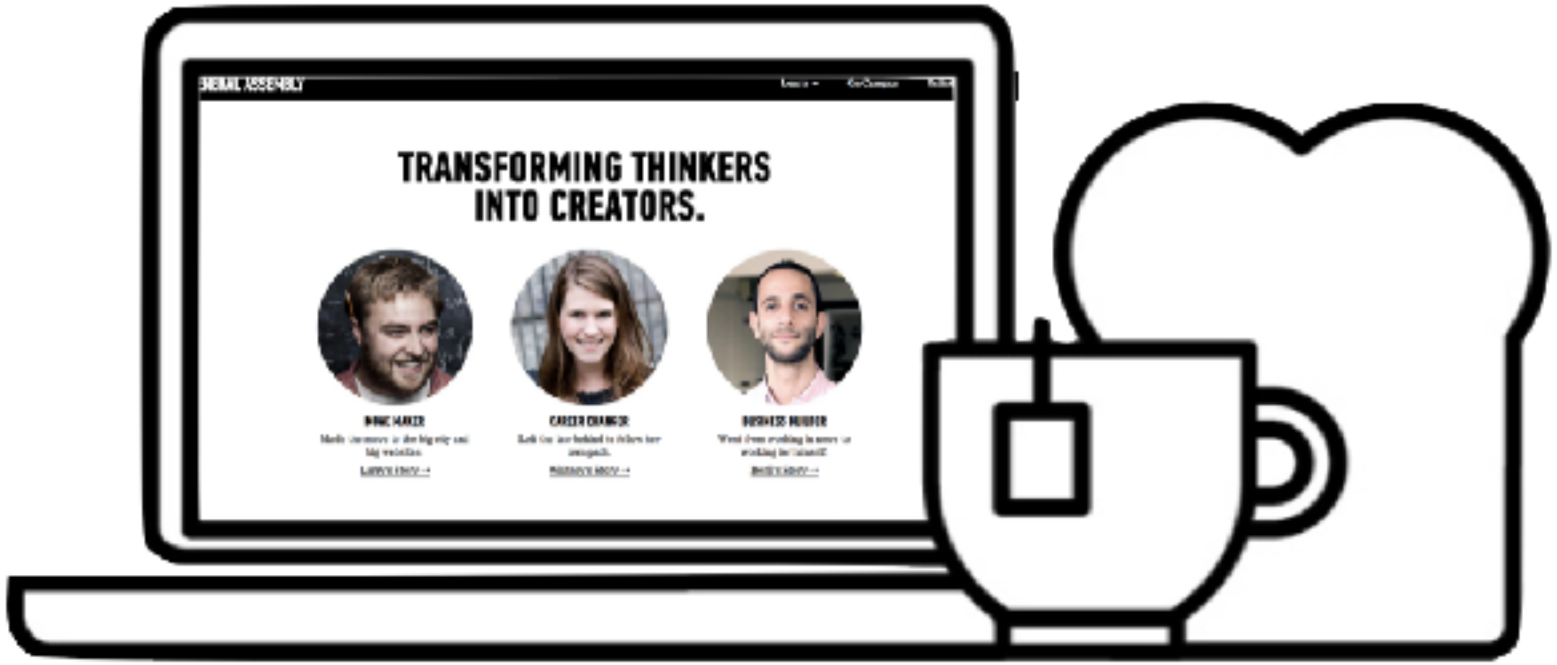
- Define basic data types used in object-oriented programming
- Recall the Python syntax for lists, dictionaries, and functions
- Create files and navigate directories using the command line interface

DATA SCIENCE

WELCOME TO GA!

FEEDBACK/SUPPORT

- Slack
- Exit tickets
- Mid-Course Feedback
- End of Course Feedback



GA GRADUATION REQUIREMENTS

HOMEWORK

(COMPLETE 80% OF HOMEWORK/LABS)

ATTENDANCE

(MISS NO MORE THAN 2 CLASSES)

FINAL PROJECT

COMMUNITY ENGAGEMENT

PARTICIPATION + FEEDBACK

WHO ARE WE?

Klaudia Magda

Klaudia holds a BSc in Control Engineering and Robotics and a MSc of Computer Science from polish top-universities.

Her expertise covers Machine Learning, Statistics and Data Manipulation. Her main tools are R, SQL and Python.

Klaudia has over 5 years experience in teaching and over 1 year experience in consulting environments, where she works with data.

WHO ARE WE?

Andrew Worsley

Andrew has held senior positions in some of New Zealand's largest and most innovative companies. He has co-authored and contributed to several academic papers and is activity involved in Auckland's tech meetup scene. His interests include Bayesian statistics, machine learning, cloud computing and AI.

WHO ARE YOU?

INTRODUCTION

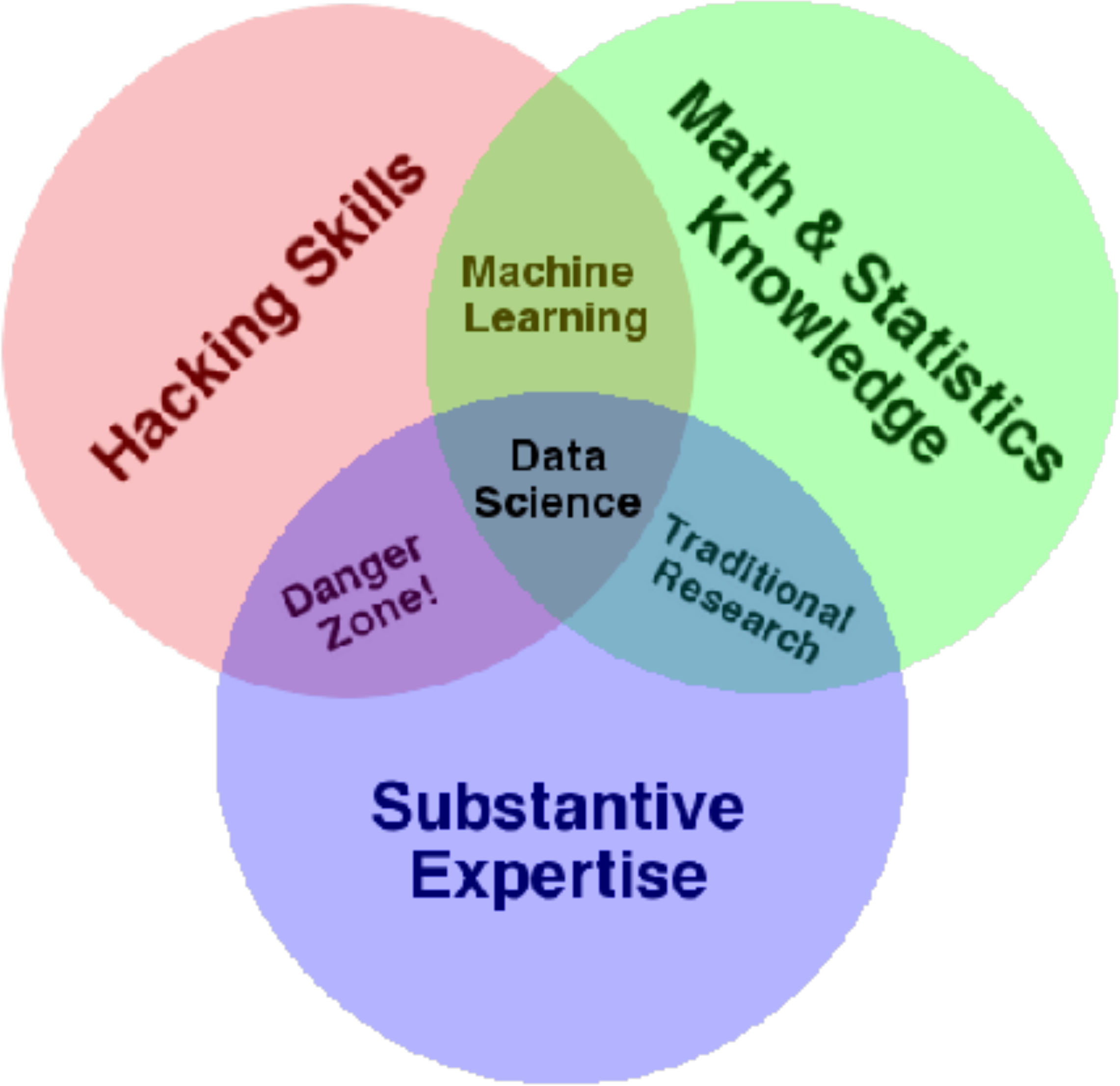
WHAT IS DATA SCIENCE?

WHAT IS A DATA SCIENTIST?

- “Data Scientist’ is a Data Analyst who lives in California”
- “A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."
- Someone who can collect, statistically explore and analyse data in an efficient and reproducible manner… but who can also translate from Dataese to Peoplese. Oh, and something something machine learning.

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

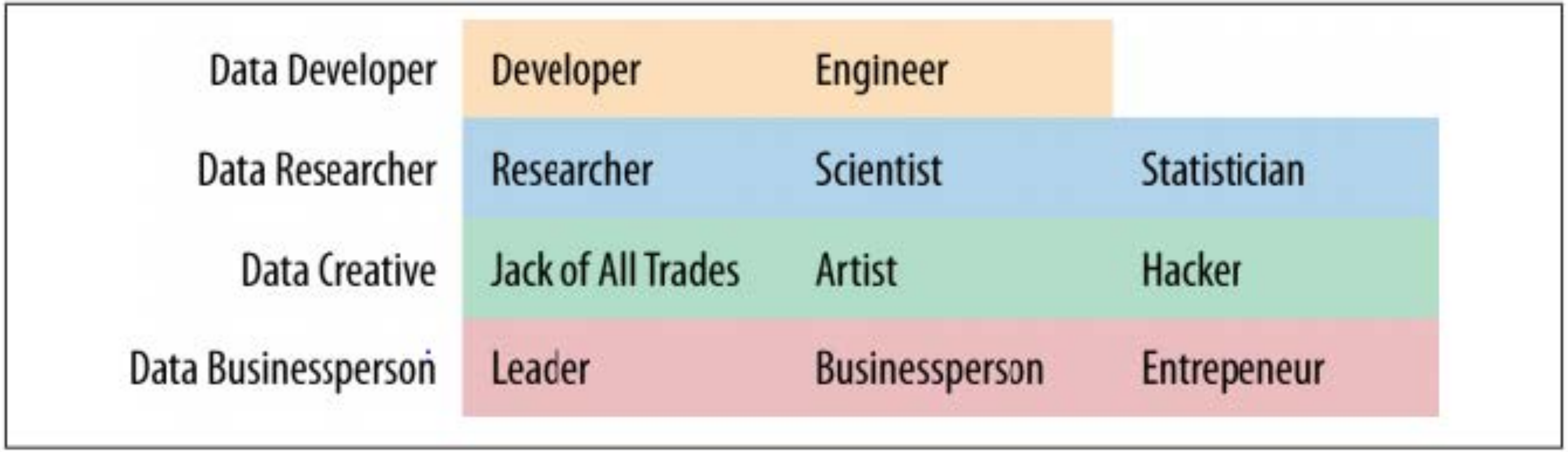
Google

FiveThirtyEight

DRAFT KINGS

WHAT ARE THE ROLES IN DATA SCIENCE?

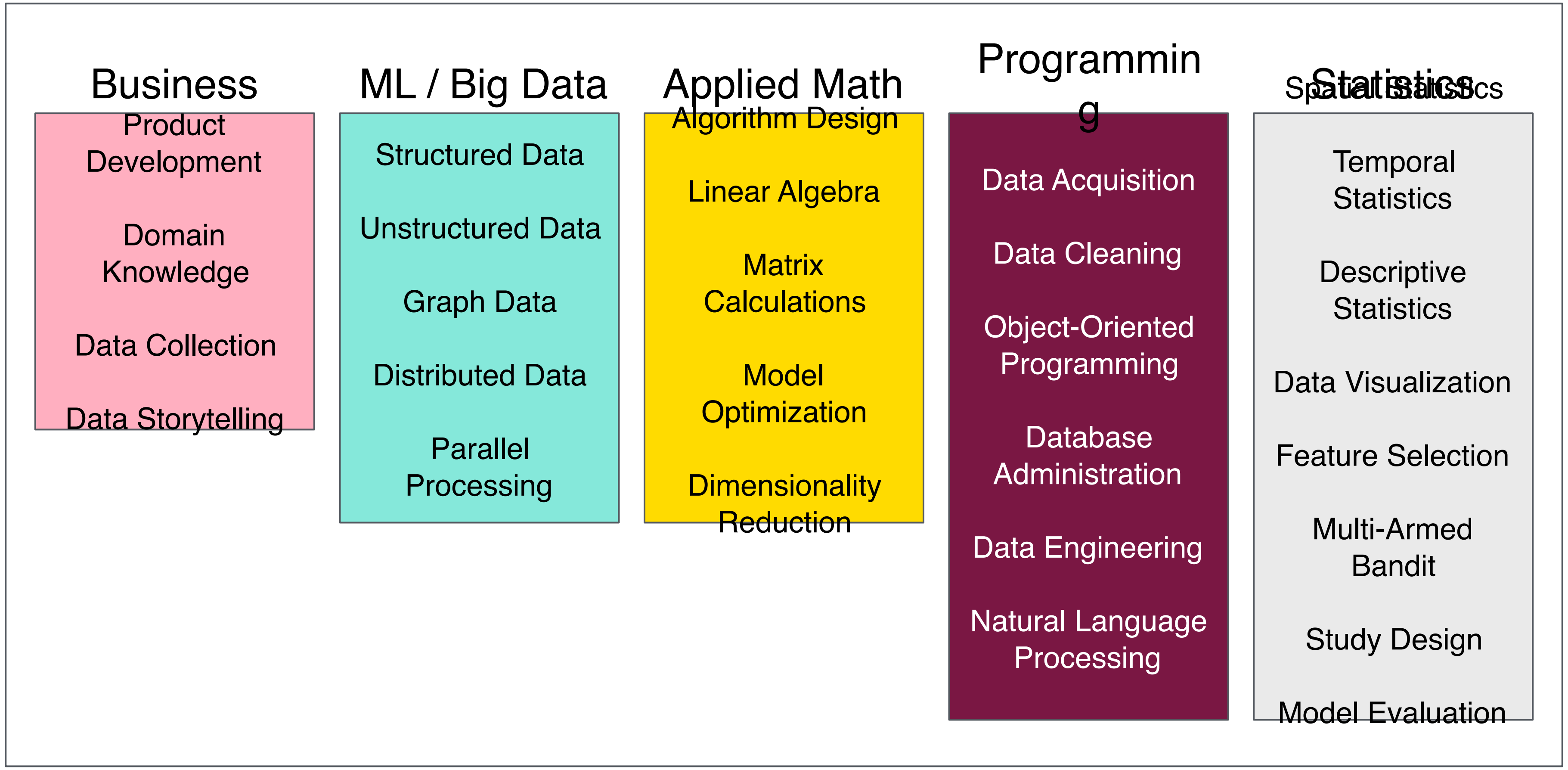
- Data Science involves a variety of roles, not just one.



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

Product Development

Domain Knowledge

Data Collection

Data Storytelling

ML / Big Data

Structured Data

Unstructured Data

Graph Data

Distributed Data

Parallel Processing

Applied Math

Algorithm Design

Linear Algebra

Matrix Calculations

Model Optimization

Dimensionality Reduction

Programming

Data Acquisition

Data Cleaning

Object-Oriented Programming

Database Administration

Data Engineering

Natural Language Processing

Statistics

Spatial Statistics

Temporal Statistics

Descriptive Statistics

Data Visualization

Feature Selection

Multi-Armed Bandit

Study Design

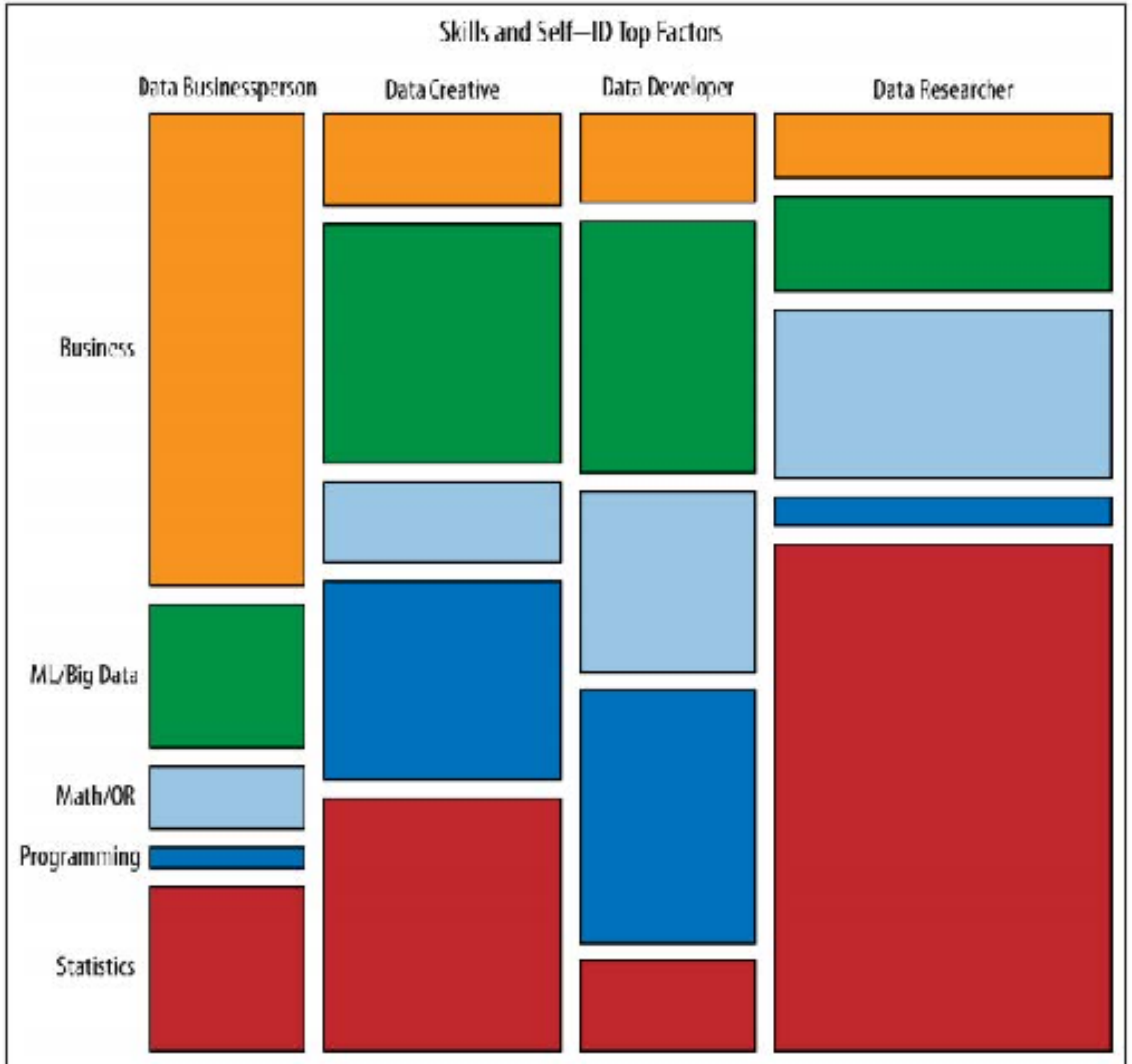
Model Evaluation

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?



INTRODUCTION

THE DATA SCIENCE WORKFLOW

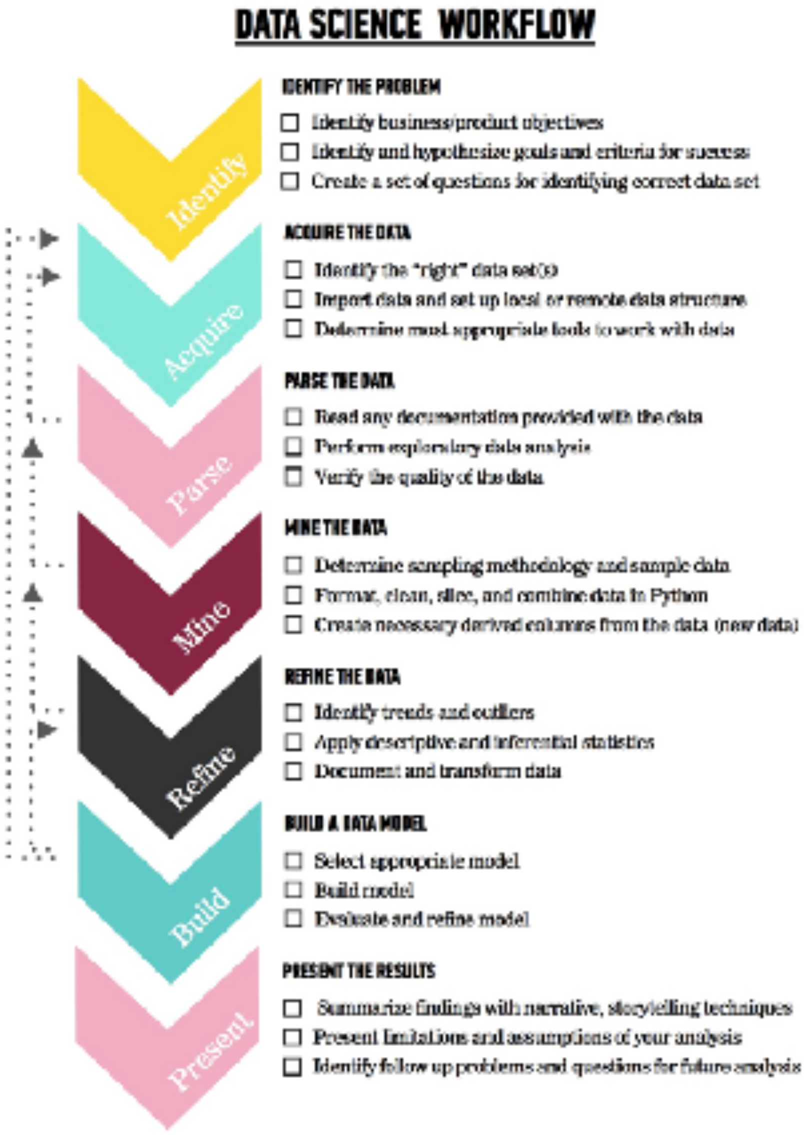
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

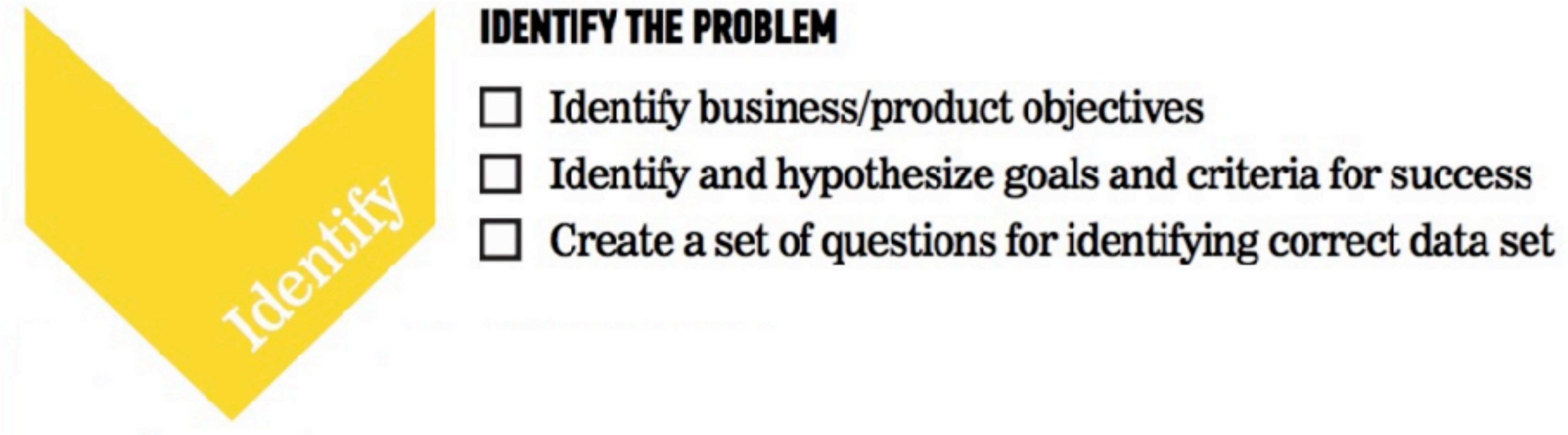
BUILD A DATA MODEL

□ 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

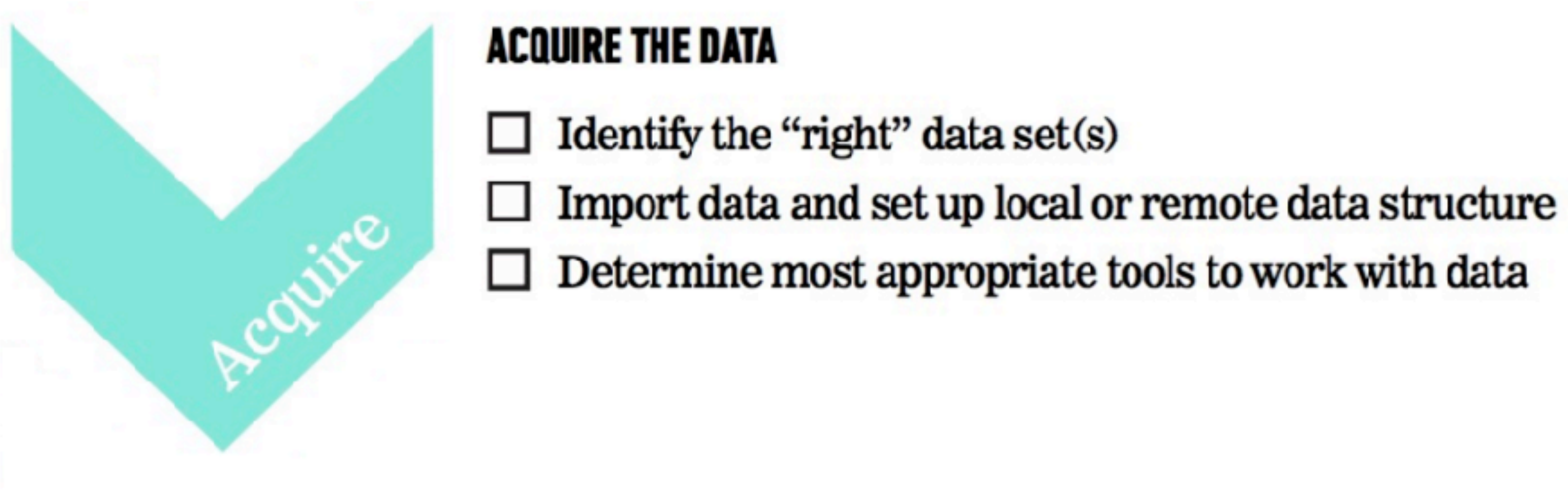
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

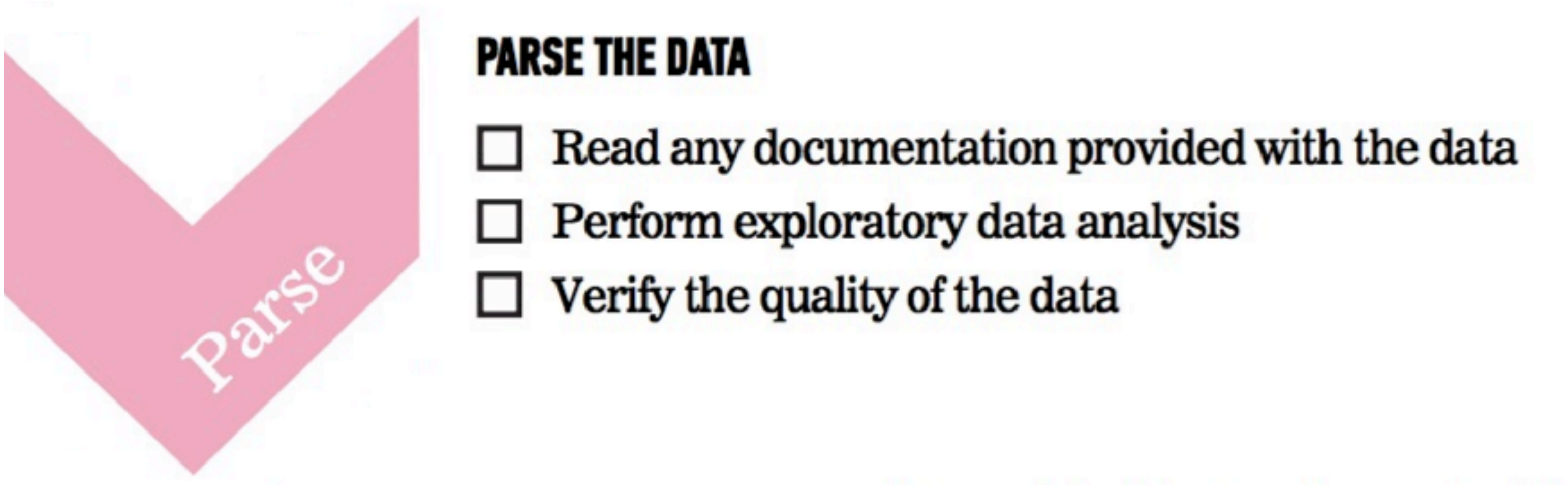
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

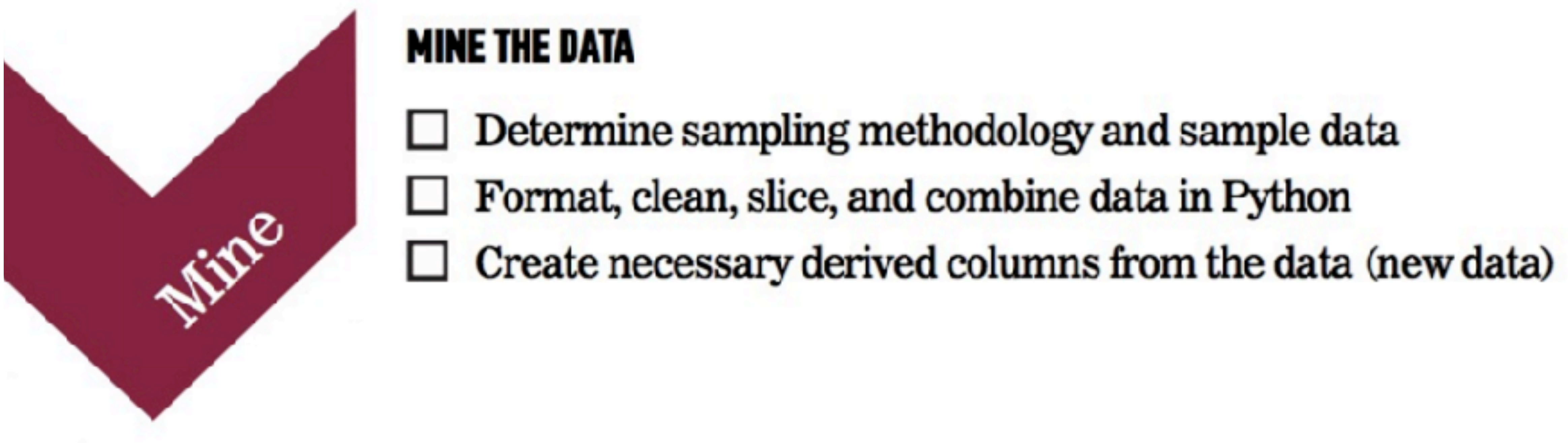
OVERVIEW OF THE DATA SCIENCE WORKFLOW



PARSE THE DATA

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

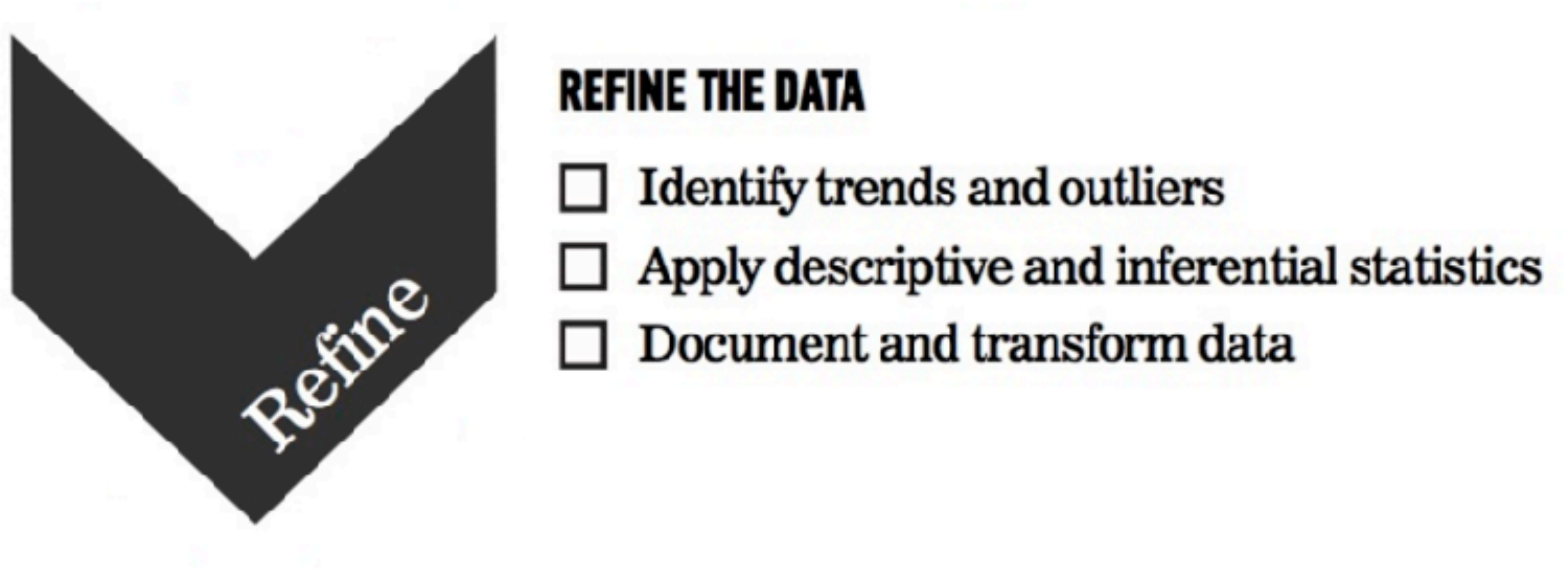
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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)

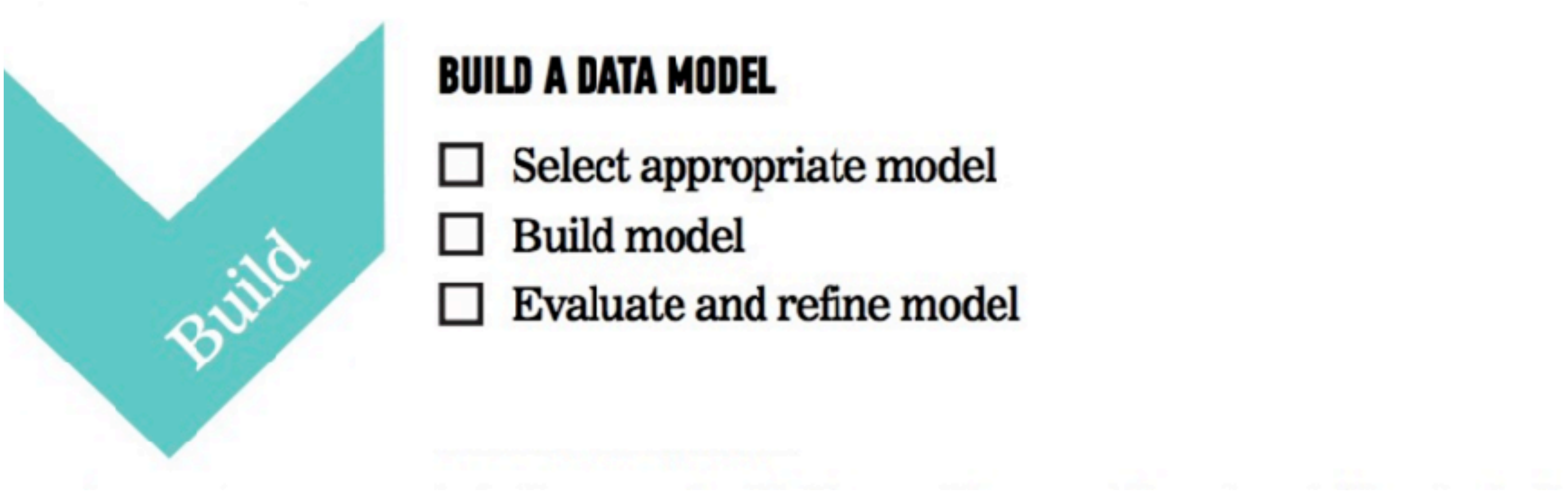
OVERVIEW OF THE DATA SCIENCE WORKFLOW



REFINE THE DATA

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

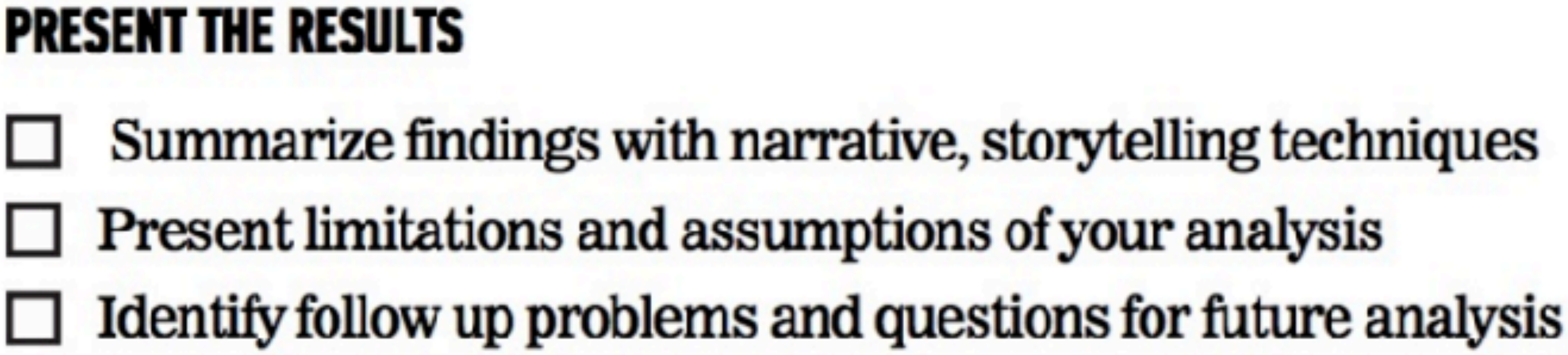
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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. 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 method
  - 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

DEMO

ENVIRONMENT SETUP

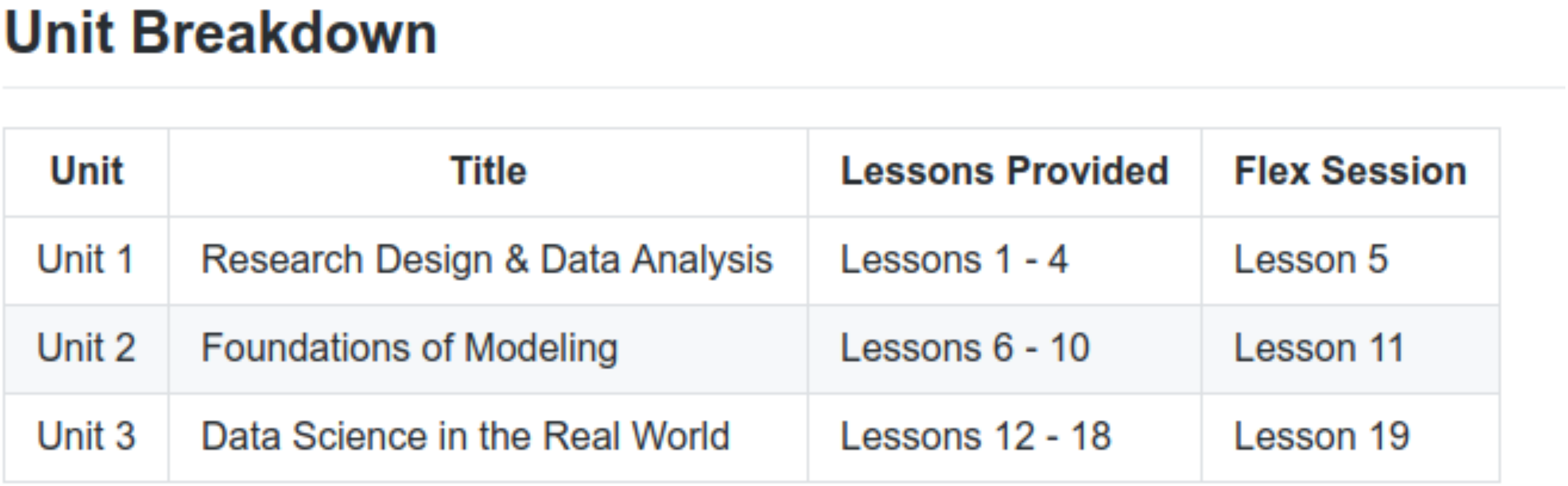
DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
  - Create a Github account
  - Install Python 3.x and Anaconda
  - Practice Python syntax, Terminal commands, and Pandas
- iPython Notebook test and Python review

CLASS STUFF

COURSE OVERVIEW

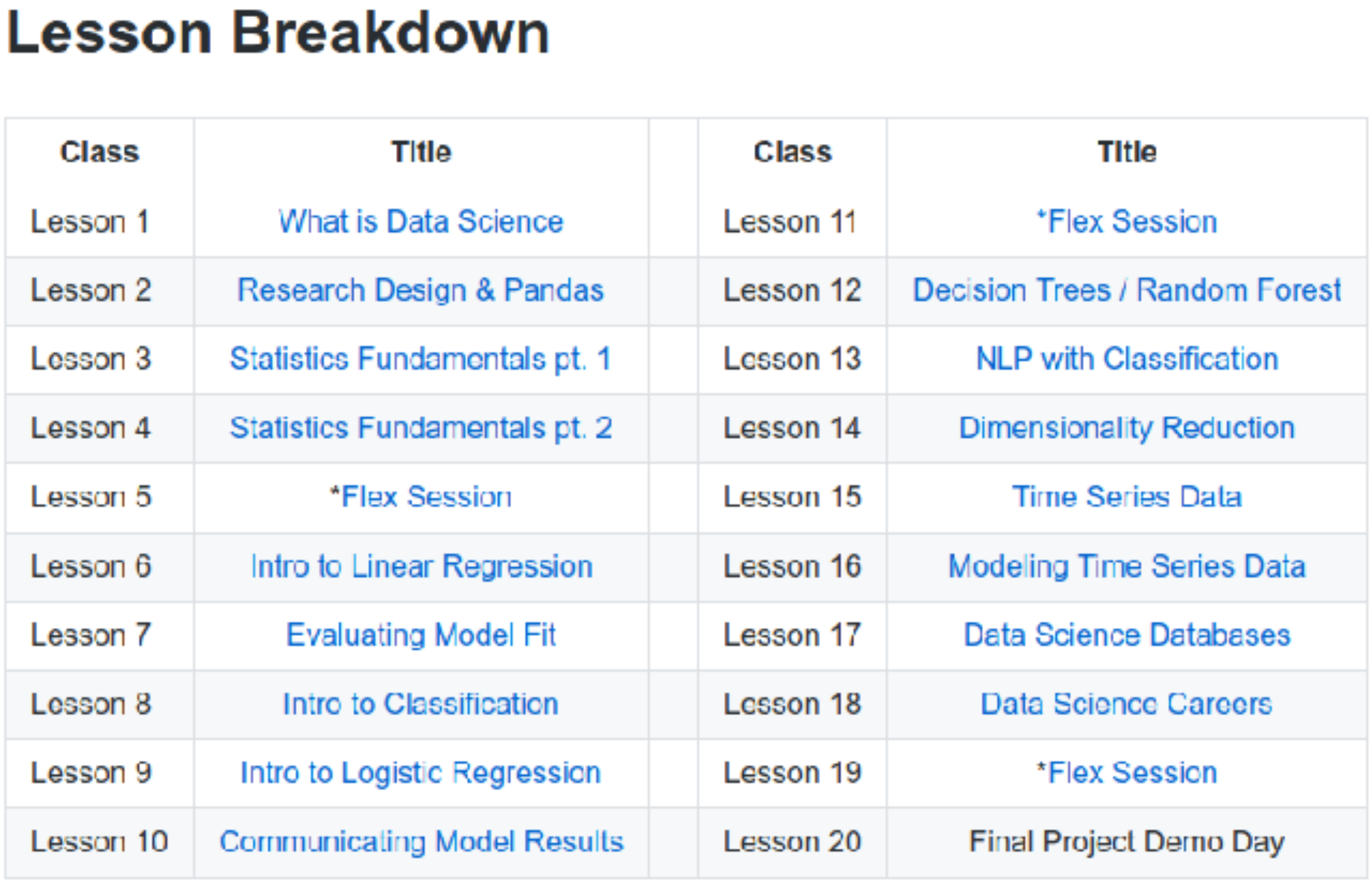
UNITS



Unit Breakdown

Unit	Title	Lessons Provided	Flex Session
Unit 1	Research Design & Data Analysis	Lessons 1 - 4	Lesson 5
Unit 2	Foundations of Modeling	Lessons 6 - 10	Lesson 11
Unit 3	Data Science in the Real World	Lessons 12 - 18	Lesson 19

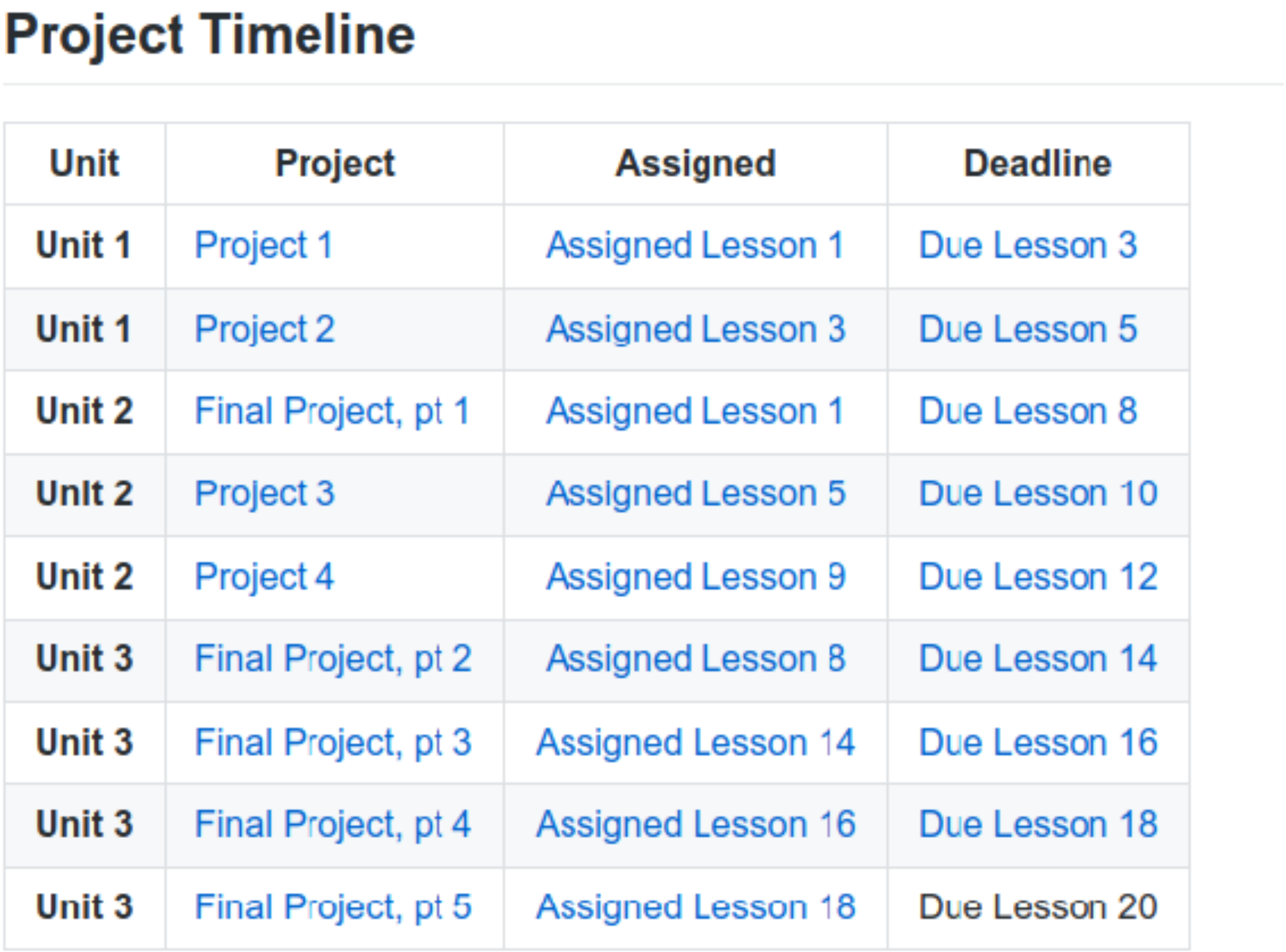
LESSONS



Lesson Breakdown

Class	Title		Class	Title
Lesson 1	What is Data Science		Lesson 11	*Flex Session
Lesson 2	Research Design & Pandas		Lesson 12	Decision Trees / Random Forest
Lesson 3	Statistics Fundamentals pt. 1		Lesson 13	NLP with Classification
Lesson 4	Statistics Fundamentals pt. 2		Lesson 14	Dimensionality Reduction
Lesson 5	*Flex Session		Lesson 15	Time Series Data
Lesson 6	Intro to Linear Regression		Lesson 16	Modeling Time Series Data
Lesson 7	Evaluating Model Fit		Lesson 17	Data Science Databases
Lesson 8	Intro to Classification		Lesson 18	Data Science Careers
Lesson 9	Intro to Logistic Regression		Lesson 19	*Flex Session
Lesson 10	Communicating Model Results		Lesson 20	Final Project Demo Day

PROJECTS



Project Timeline

Unit	Project	Assigned	Deadline
Unit 1	Project 1	Assigned Lesson 1	Due Lesson 3
Unit 1	Project 2	Assigned Lesson 3	Due Lesson 5
Unit 2	Final Project, pt 1	Assigned Lesson 1	Due Lesson 8
Unit 2	Project 3	Assigned Lesson 5	Due Lesson 10
Unit 2	Project 4	Assigned Lesson 9	Due Lesson 12
Unit 3	Final Project, pt 2	Assigned Lesson 8	Due Lesson 14
Unit 3	Final Project, pt 3	Assigned Lesson 14	Due Lesson 16
Unit 3	Final Project, pt 4	Assigned Lesson 16	Due Lesson 18
Unit 3	Final Project, pt 5	Assigned Lesson 18	Due Lesson 20

REVIEW

WHAT ARE THESE COMMANDS (AND WHERE DO THEY COME FROM)?

ls	seq	import
mkdir	float	touch
git clone	len	def
str.replace	range	git add

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE - LESSON 3 (next Monday)

- Project: Begin work on Project 1