GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Mart van de Ven | Dickson Kwong | Alex Anzola Jürgenson

Data Scientists, Droste

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

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

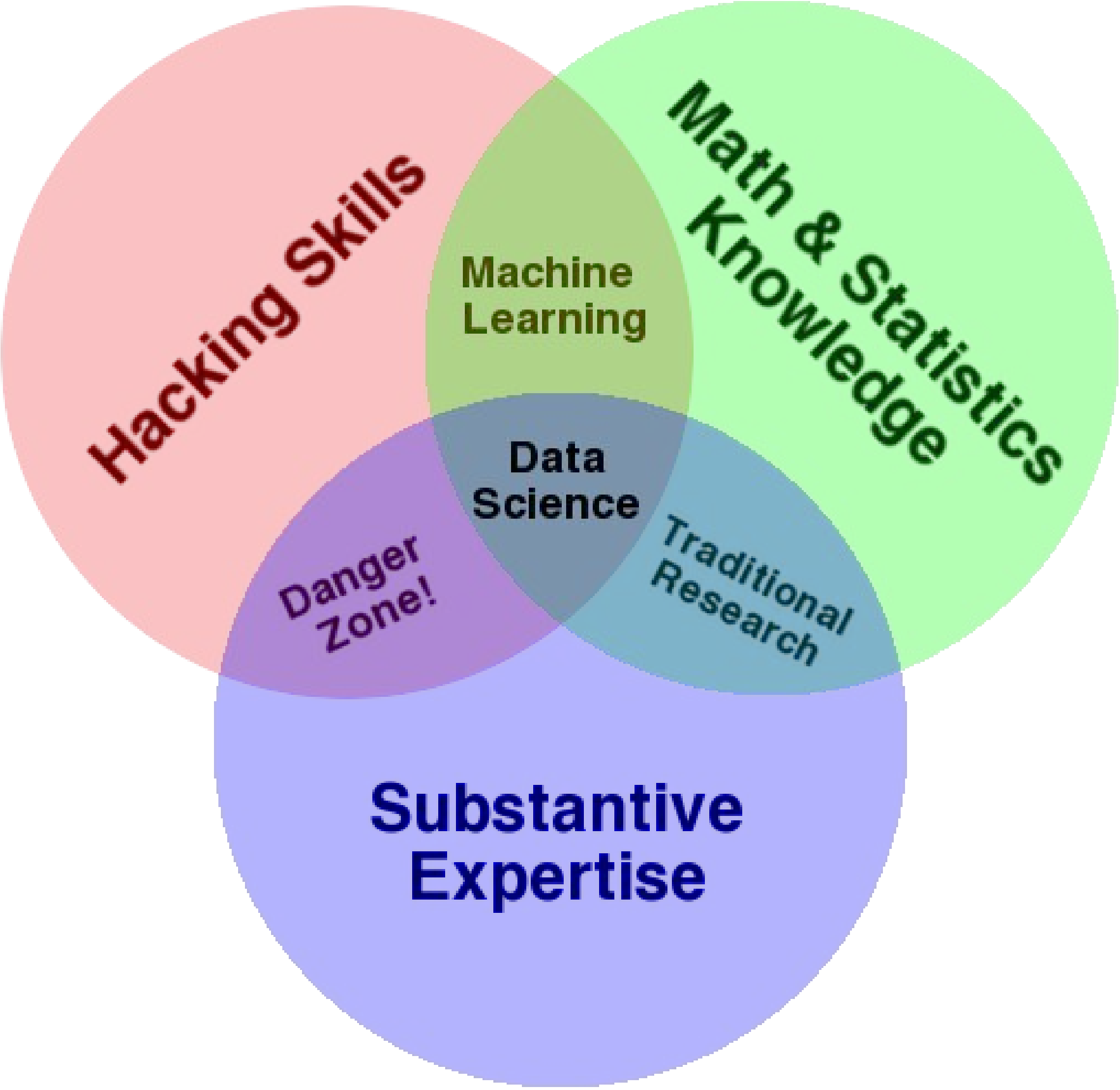
WELCOME TO GA!

INTRODUCTION

WHAT IS DATA SCIENCE?

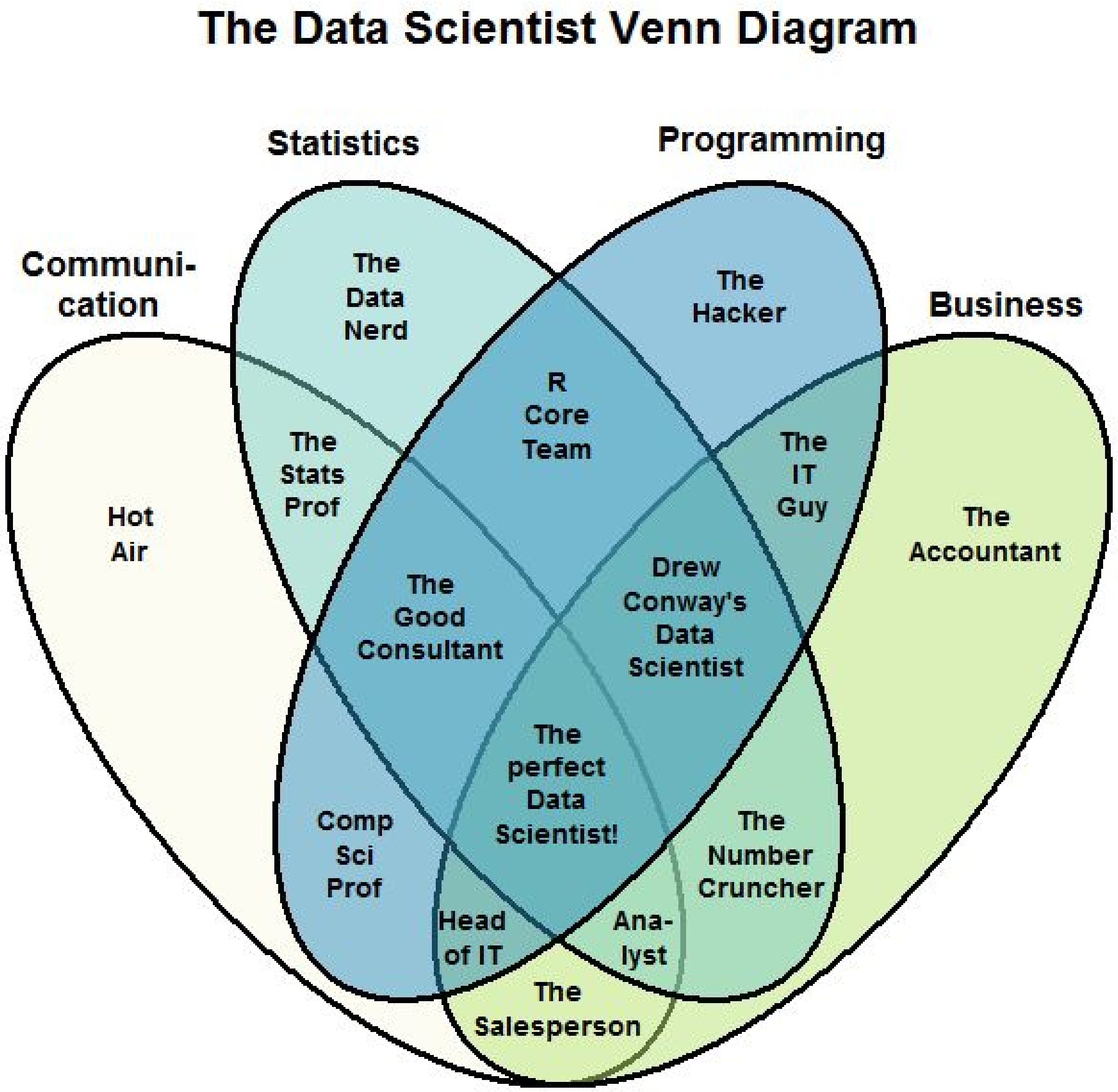
WHAT IS DATA SCIENCE?

- A set of tools and techniques for data
- Interdisciplinary problem-solving
- Application of scientific techniques to practical problems



WHAT IS DATA SCIENCE?

- A developing field with lots definitions of what data science is
- For our purposes, we will take Data Science to be an approach to finding intelligence in data with machine learning methods



WHO USES DATA SCIENCE?

FiveThirtyEight

DRAFT
KINGS

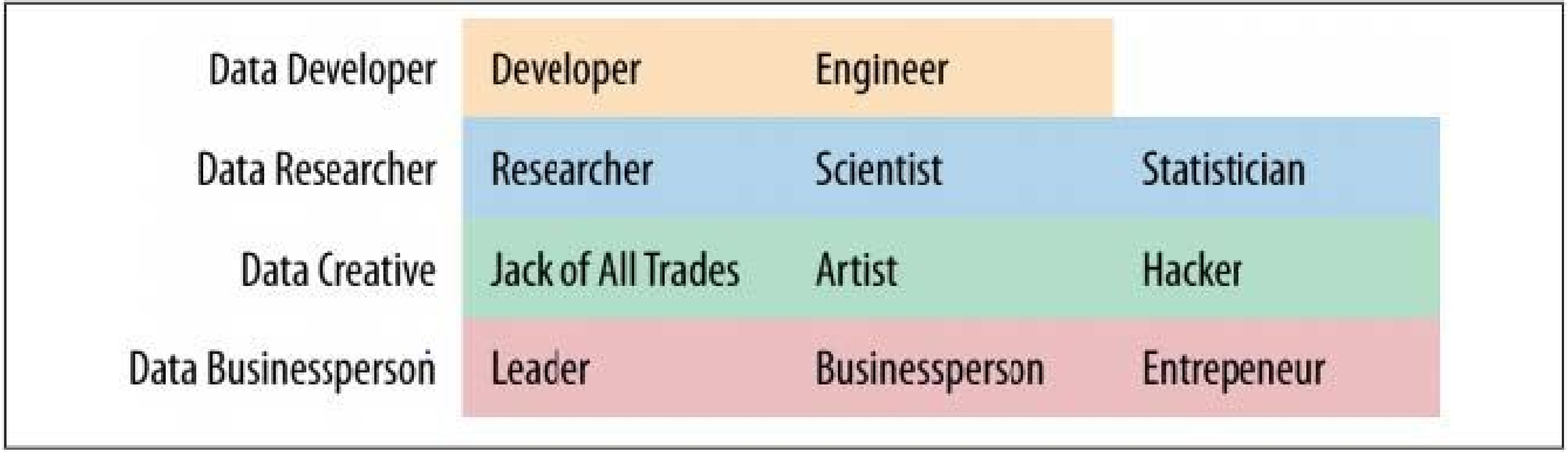
WHO USES DATA SCIENCE?

- Can you think of others?

ima...
021 3 1

WHAT ARE THE ROLES IN DATA SCIENCE?

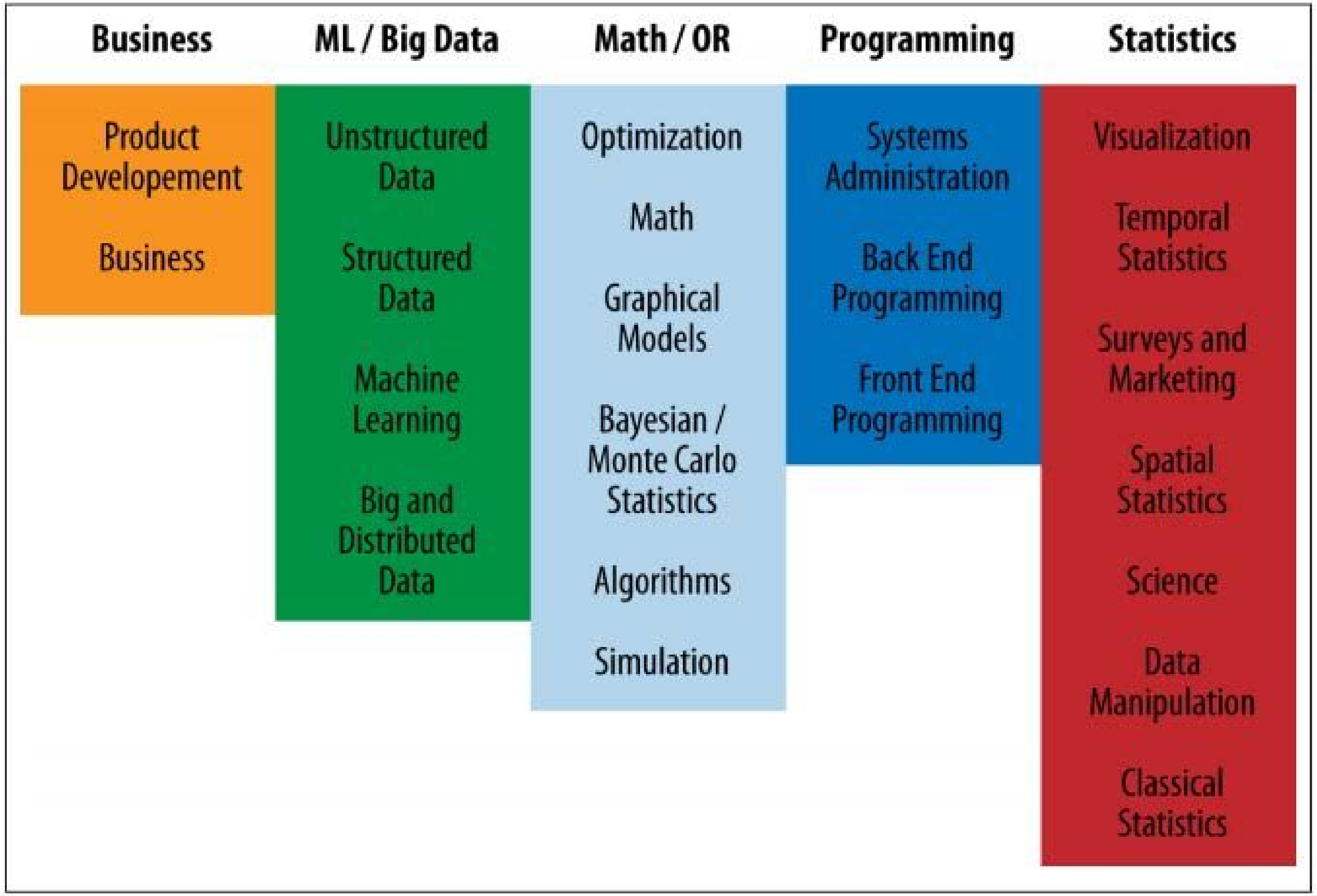
- Data Science involves a variety of roles, not just one.



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.



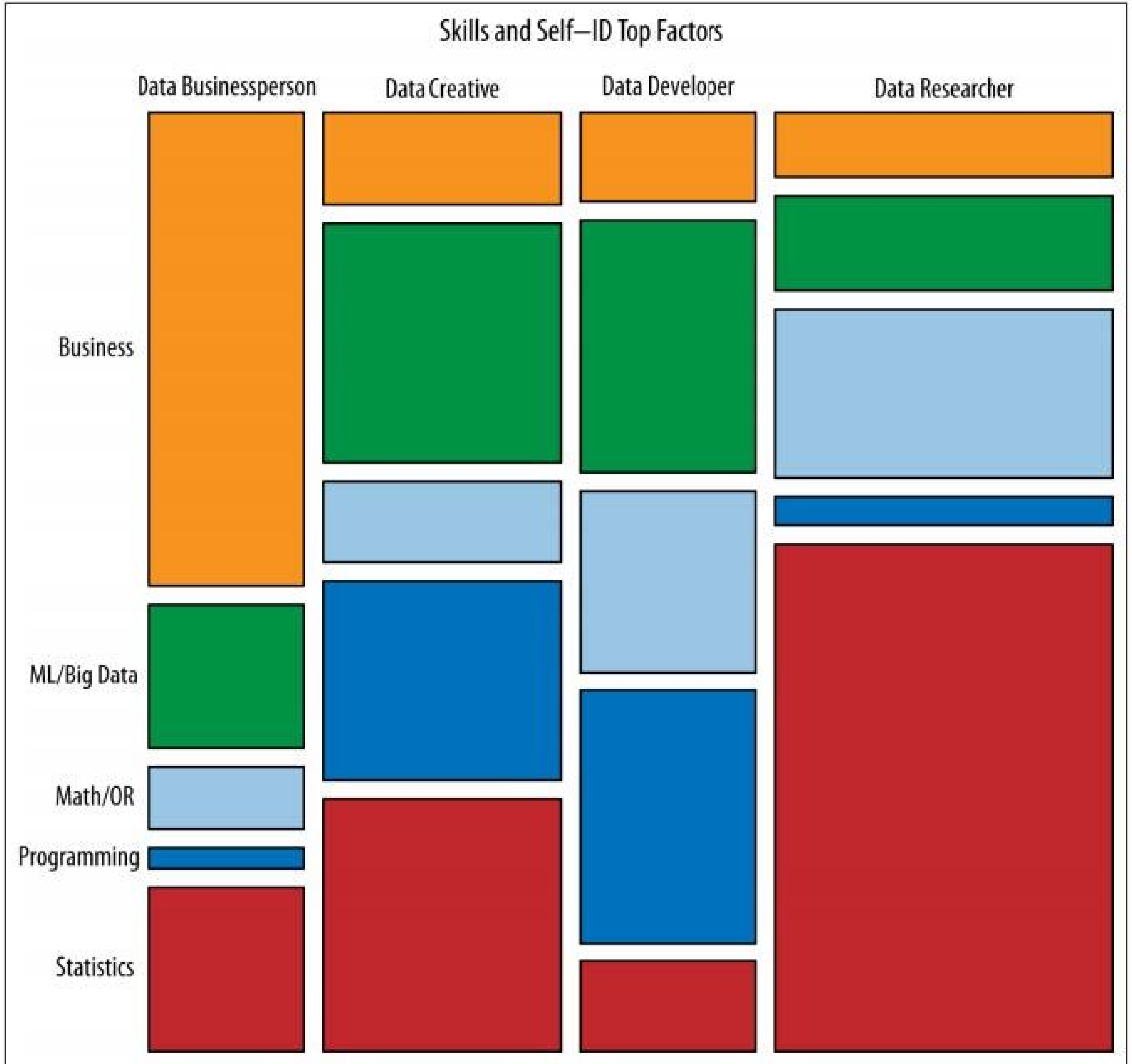
Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and 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?



INTRODUCTION

THE DATA SCIENCE WORKFLOW

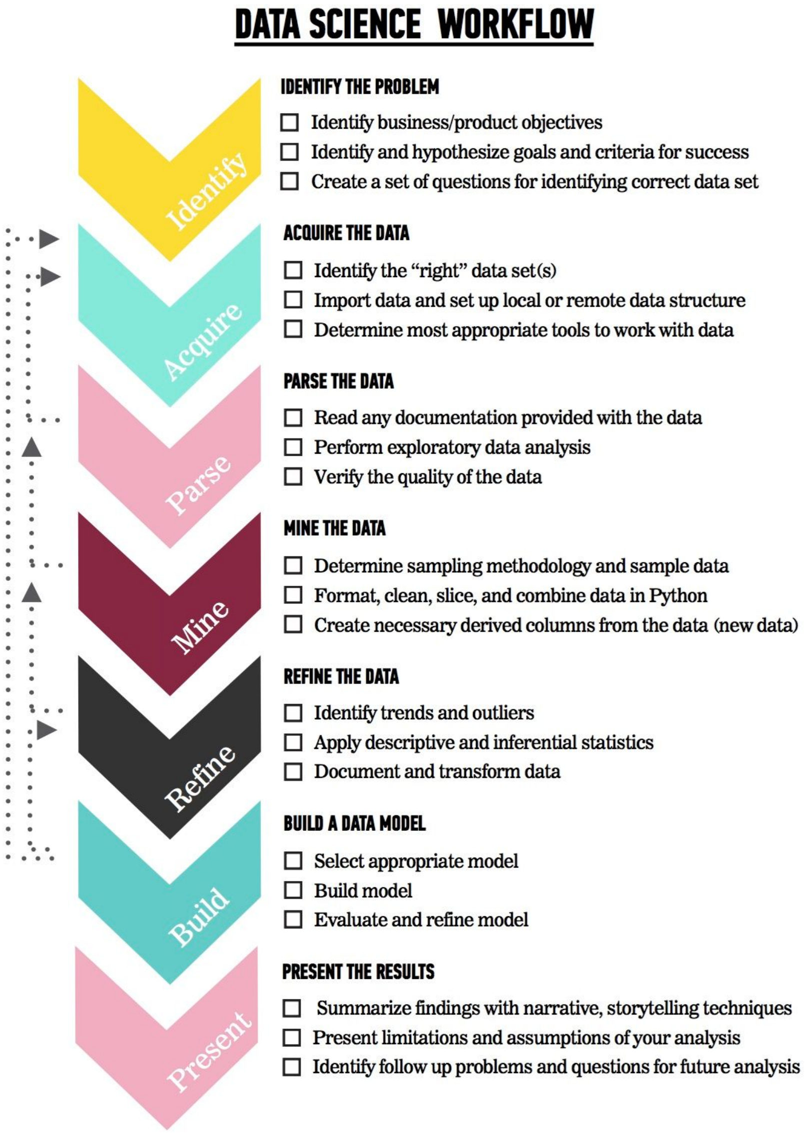
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

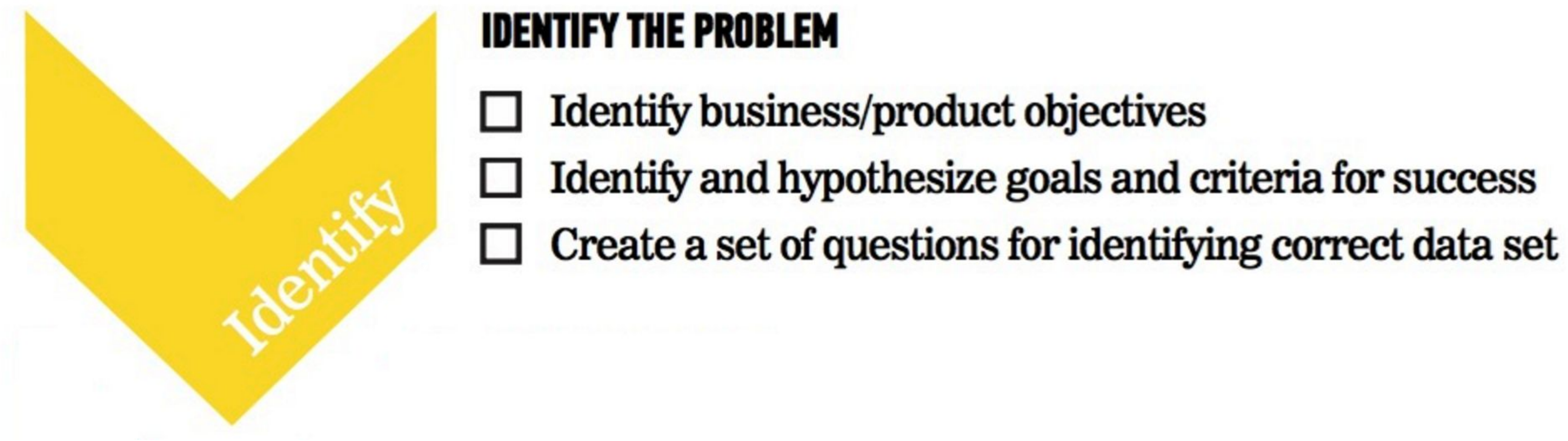
BUILD A DATA MODEL

- □ 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

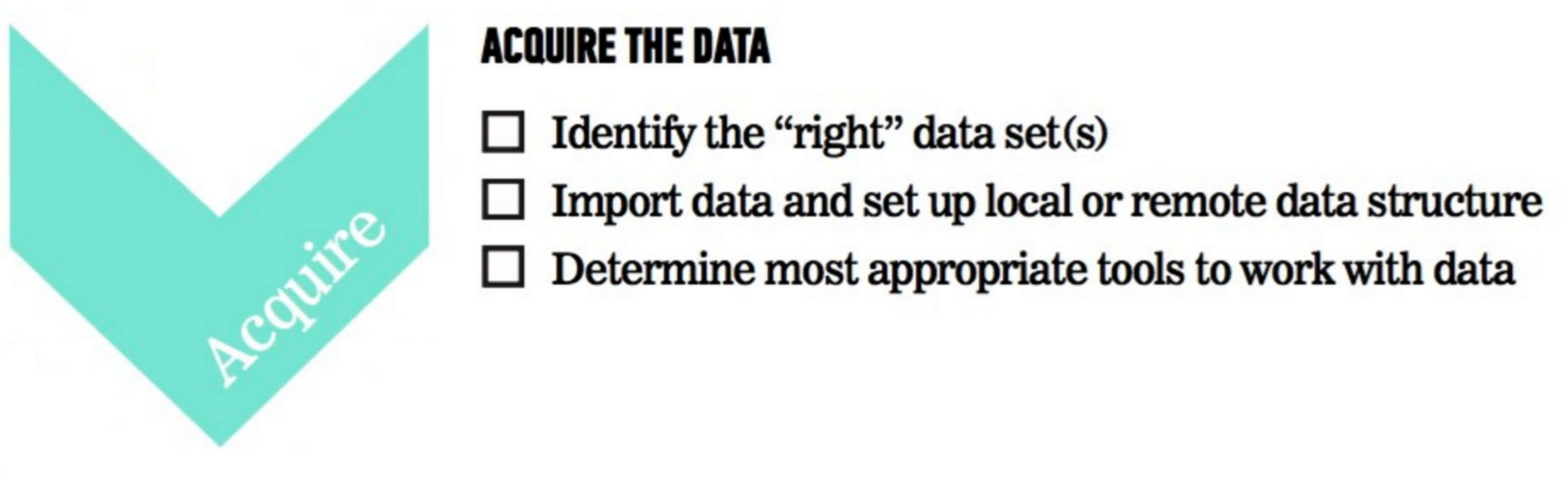
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

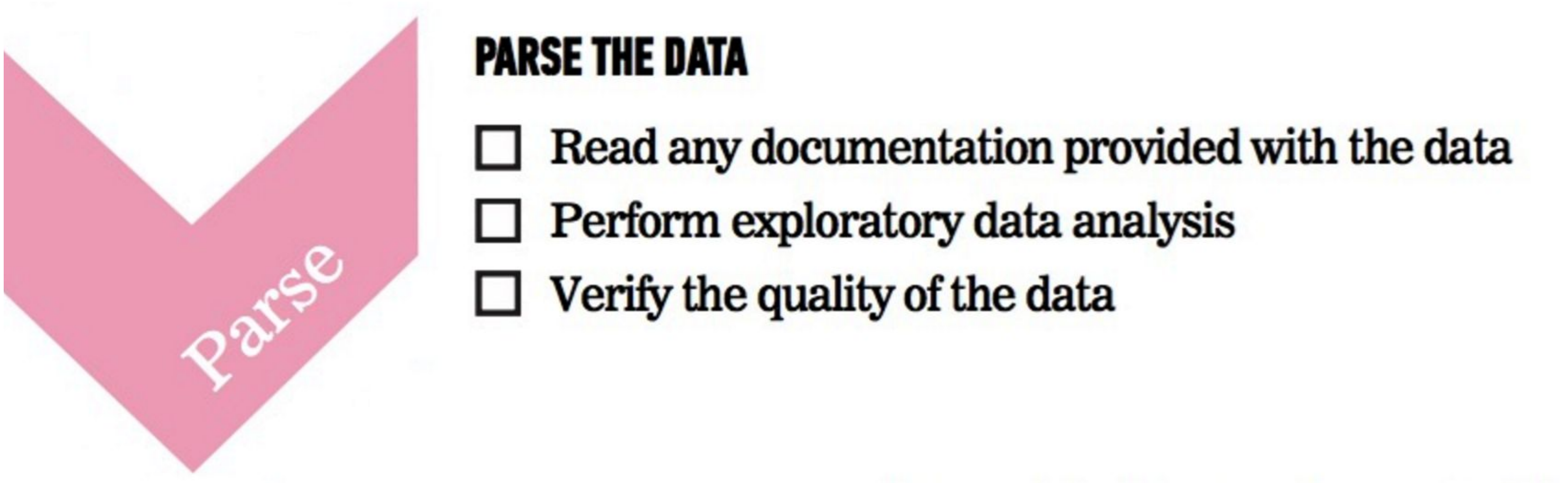
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

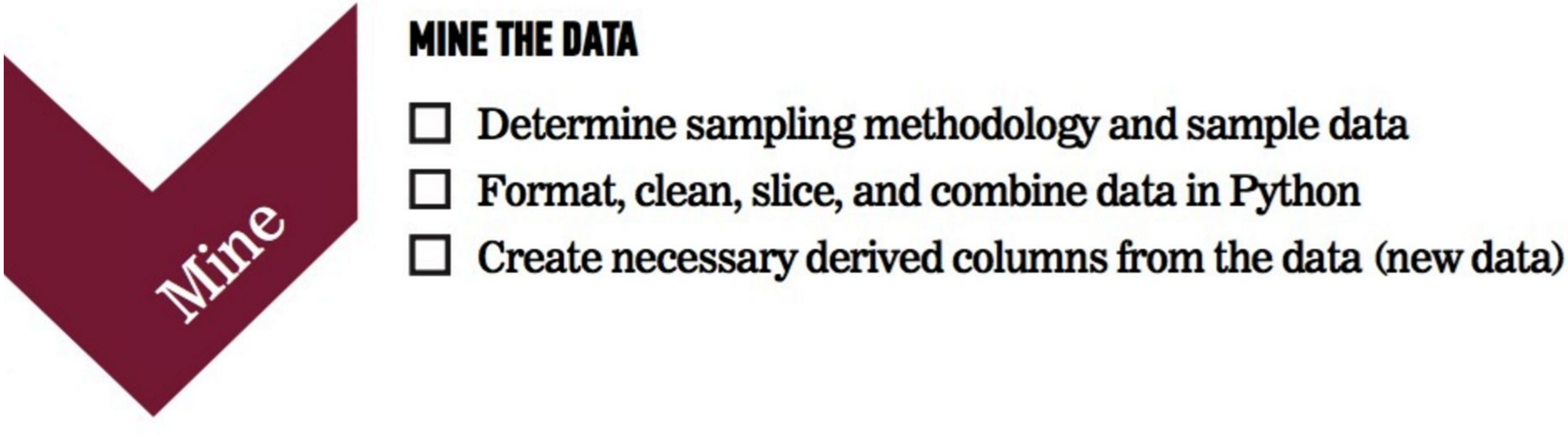
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

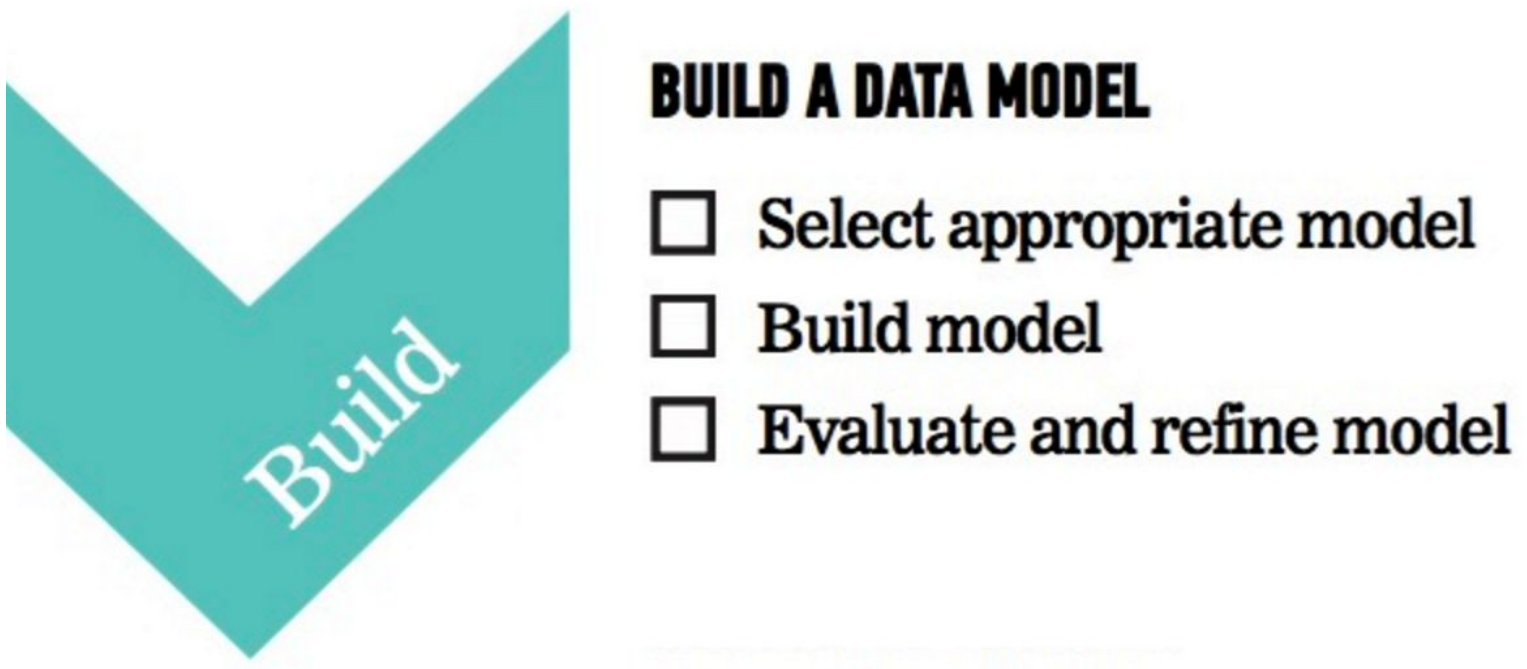
□ Create necessary derived columns from the data (new data)

OVERVIEW OF THE DATA SCIENCE WORKFLOW

REFINE THE DATA

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

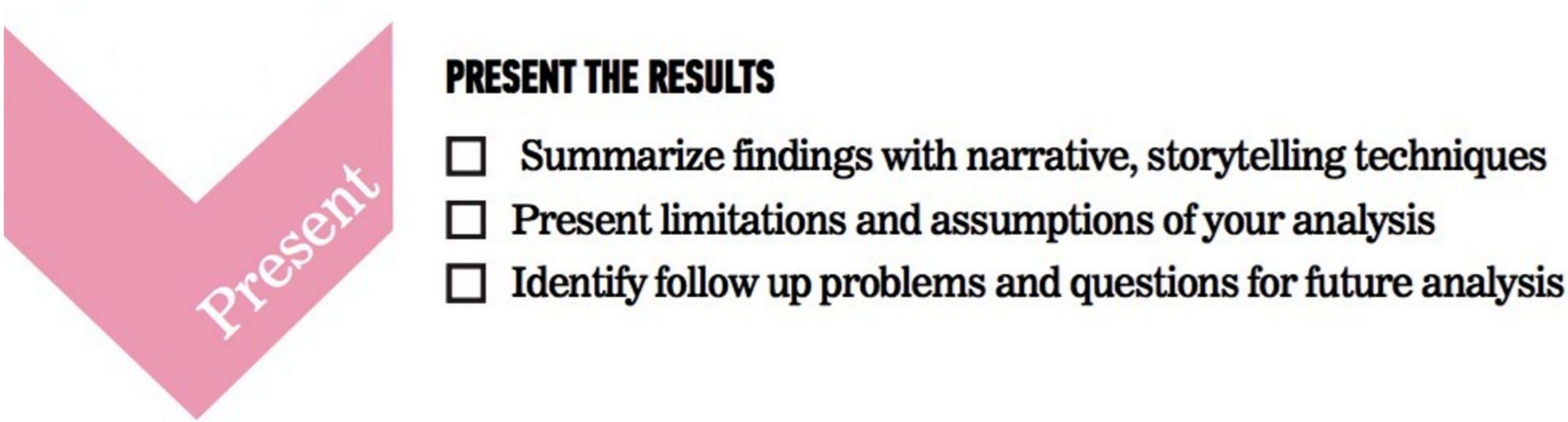
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

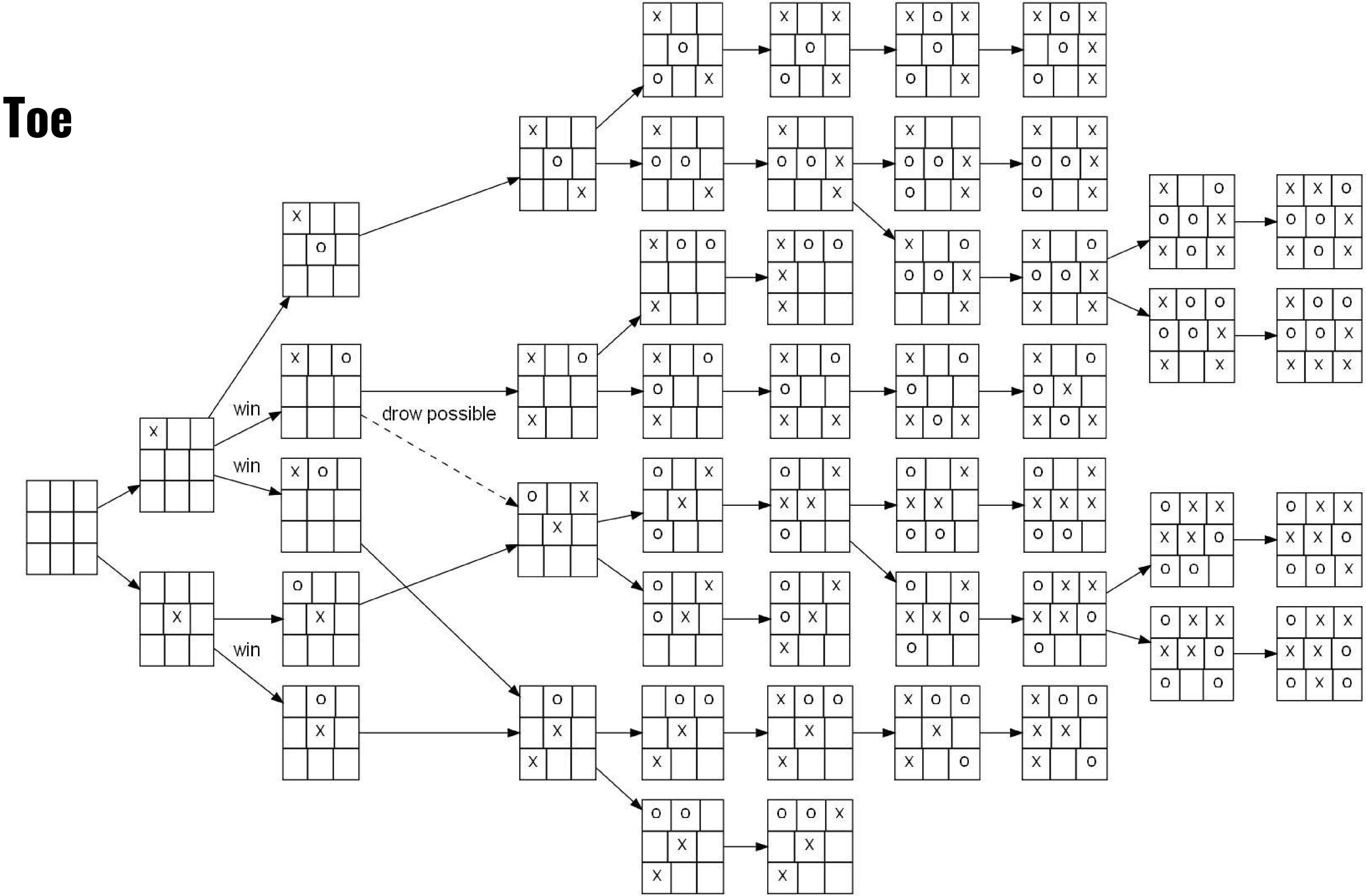
GUIDED PRACTICE

DATA SCIENCE WORK FLOW

GUIDED PRACTICE

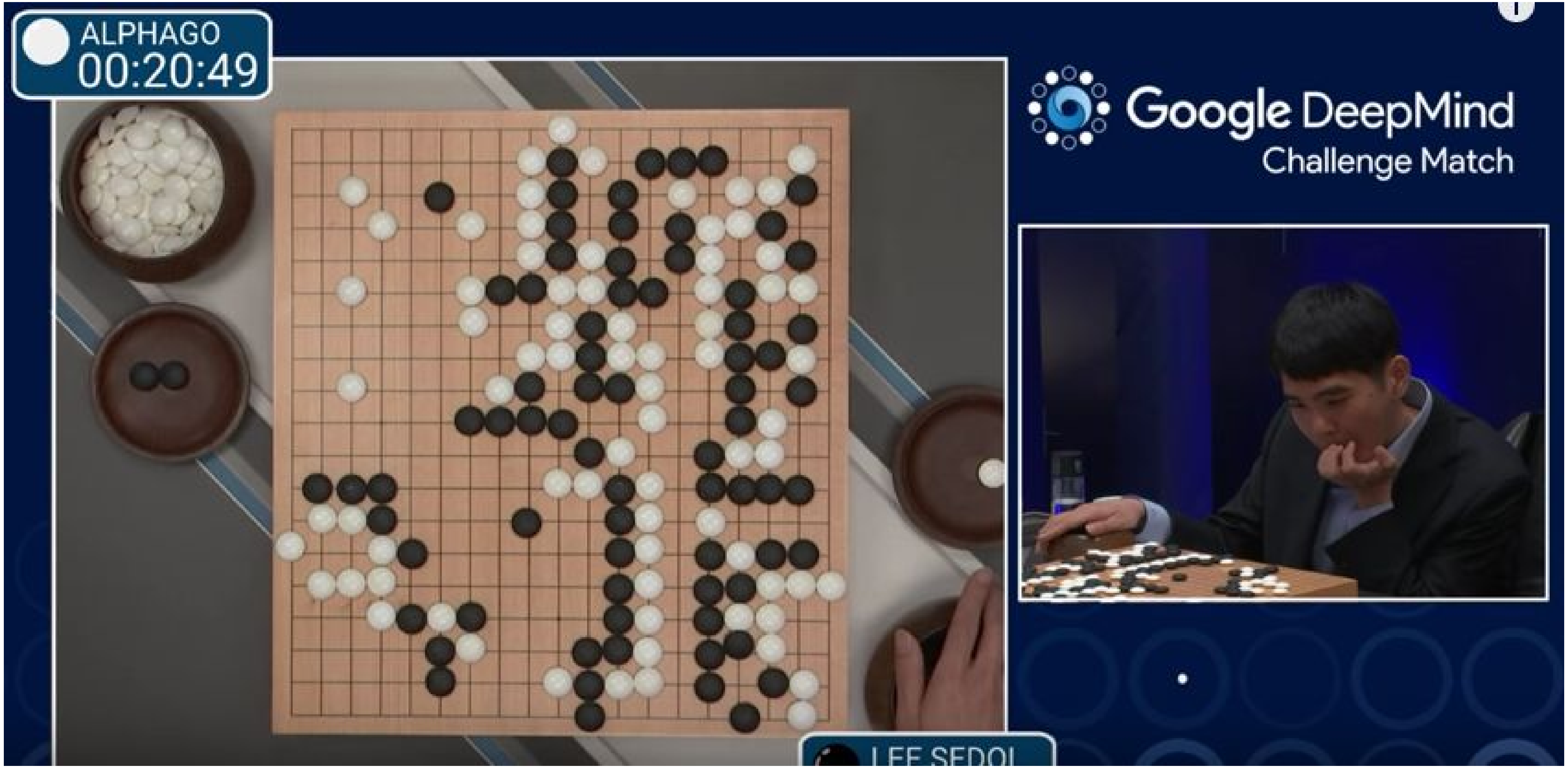
CREATING A TAK AI

Tik Tak Toe



Tak

Go



ACTIVITY: THE 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

DEMO

ENVIRONMENT SETUP

DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
  - Create a Github account, Install GitHub Desktop
  - Install Python 2.7 with Anaconda
  - Practice Python syntax, Terminal commands, and Pandas
- iPython Notebook test and Python review

DEV ENVIRONMENT SETUP - GITHUB

- Create a Github Account (github.com)
- Install GitHub Desktop (Win/OSX) / GitKraken (Linux)
- Follow your instructors

tijptjik

DicksonK

AlexAnzolaJ

- Check out gist.github.com with your instructor :)

DEV ENVIRONMENT SETUP - CONDA

- Install Python 2.7 with Anaconda
- Open a terminal, and copy paste:

```
pip install plotly cufflinks watermark
```

- Test your jupyter notebook server

```
jupyter notebook
```

- Check out the content for lesson 1 starter code available at */lessons/lesson-1/code/starter-code/lesson1-starter-code.ipynb* in the Github repo

DEV ENVIRONMENT SETUP - CLASS FLOW

ONE TIME

- Clone the DS_HK_XX repo with the GitHub Desktop Client
- Create a folder to store your personal copies of the class files, call it something like DS_HK_alex

EACH TIME

- Sync your Repo with the GitHub Client
- Copy the files from DS_HK_XX repo into DS_HK_alex
- Start the Jupyter Notebook Server in DS_HK_alex

DEV ENVIRONMENT SETUP - FIXING JUPYTER

Copy over the `jupyter_notebook_config.py` file from the class materials you have downloaded into:

- OSX/Linux: `~/.jupyter`
- Windows : `%PROGRAMDATA%\jupyter\`

Now from your Jupyter Notebook index page, open

- `install.verification.ipynb`

Click ‘Cell’ → ‘Run All’ and make sure it renders the 3D bubble plot.

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

- Project: Begin work on Project 1

Q & A