GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

ANDREW WORSLEY

National Data Science Lead - Velrada

Lead Instructor (Part-time), Data Science - GA

Expertise in production-grade machine learning, high-performance cloud computing and statistical analysis.

generalassemb.ly/instructors/andrew-worsley/16428

www.linkedin.com/in/andrew-worsley/

INTRODUCTION

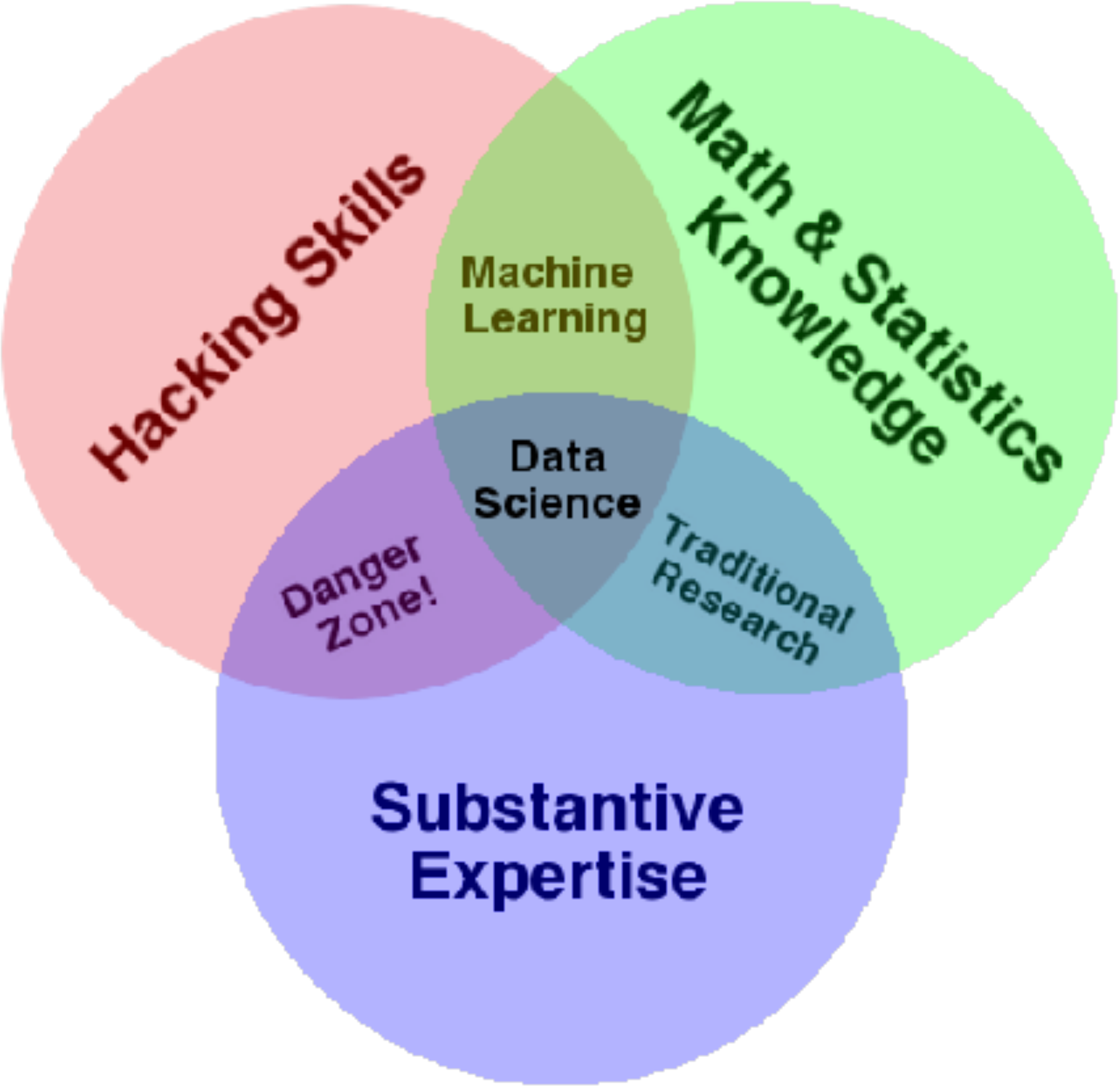
WHAT IS DATA SCIENCE?

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.

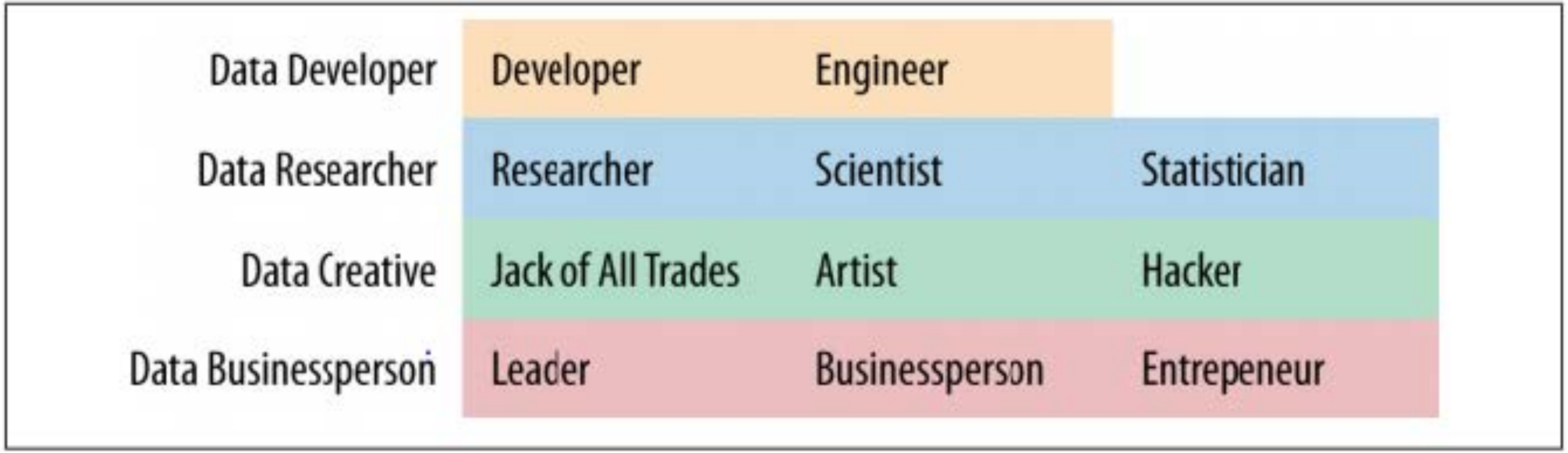
WHAT IS DATA SCIENCE?

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



WHAT ARE THE ROLES IN DATA SCIENCE?

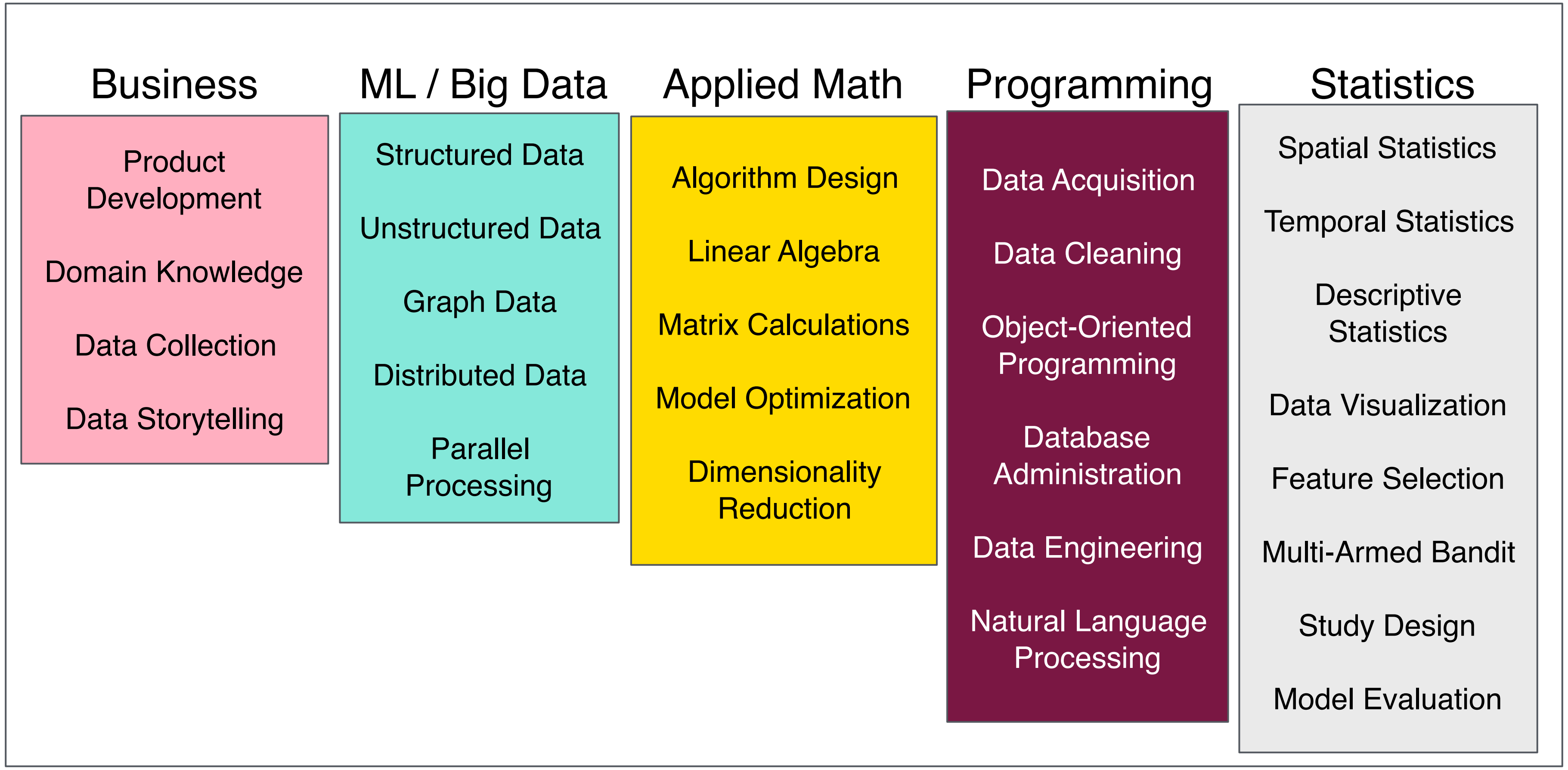
- Data Science involves a variety of roles, not just one.



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

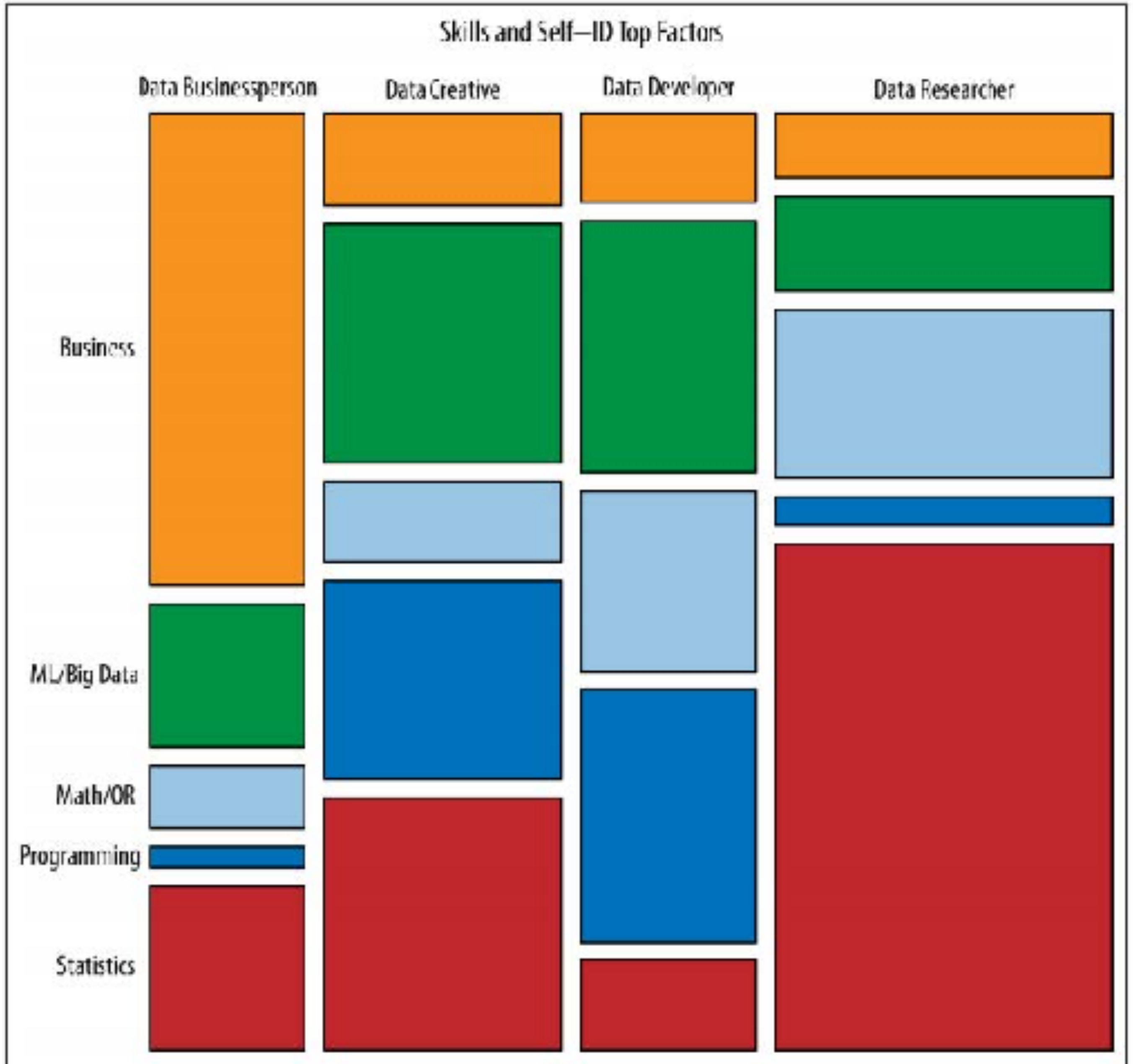
- Data Acquisition
- Data Cleaning
- Object-Oriented Programming
- Database Administration
- Data Engineering
- Natural Language Processing

Statistics

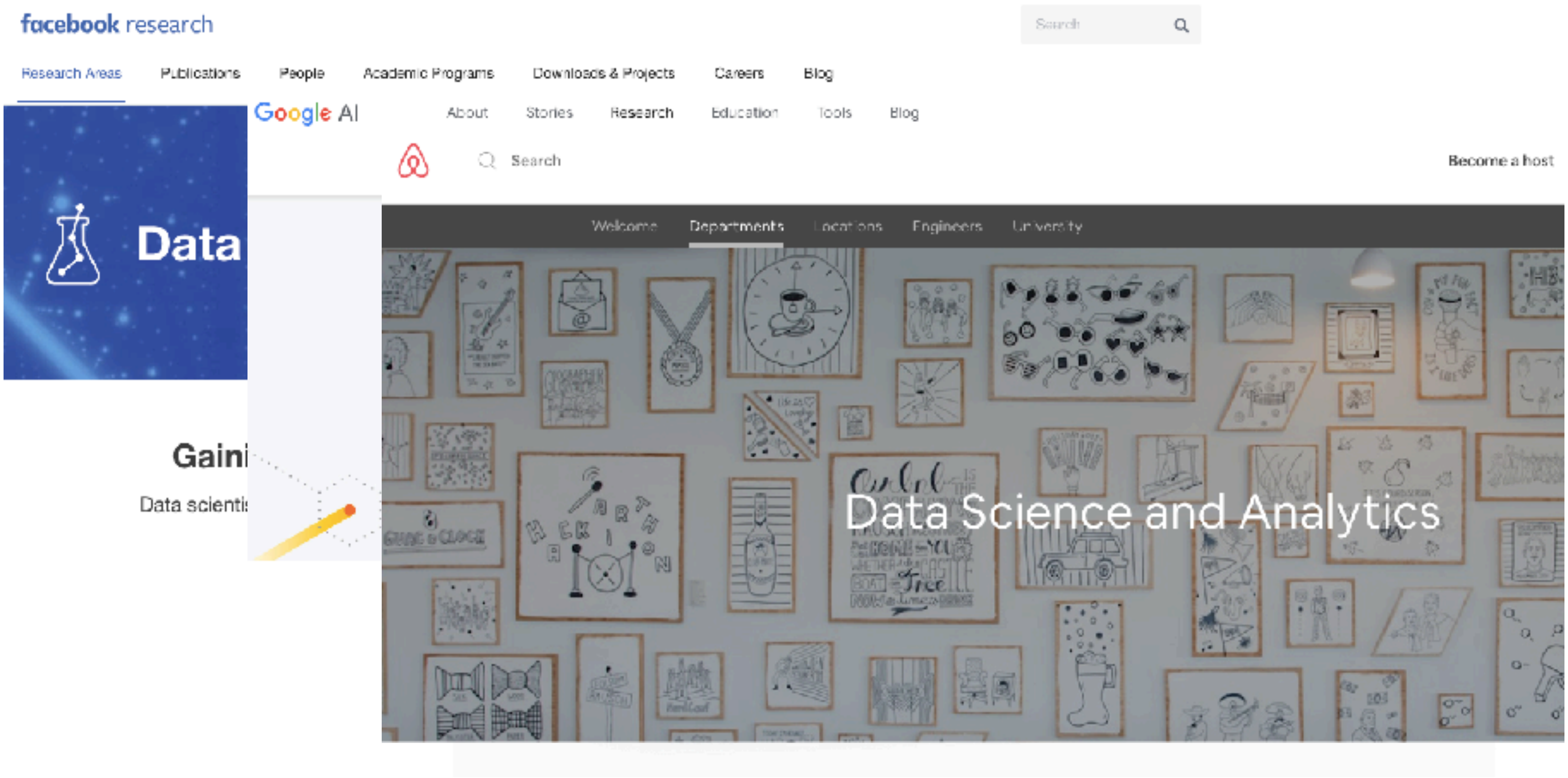
- Spatial Statistics
- Temporal Statistics
- Descriptive Statistics
- Data Visualization
- Feature Selection
- Multi-Armed Bandit
- Study Design
- Model Evaluation

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?



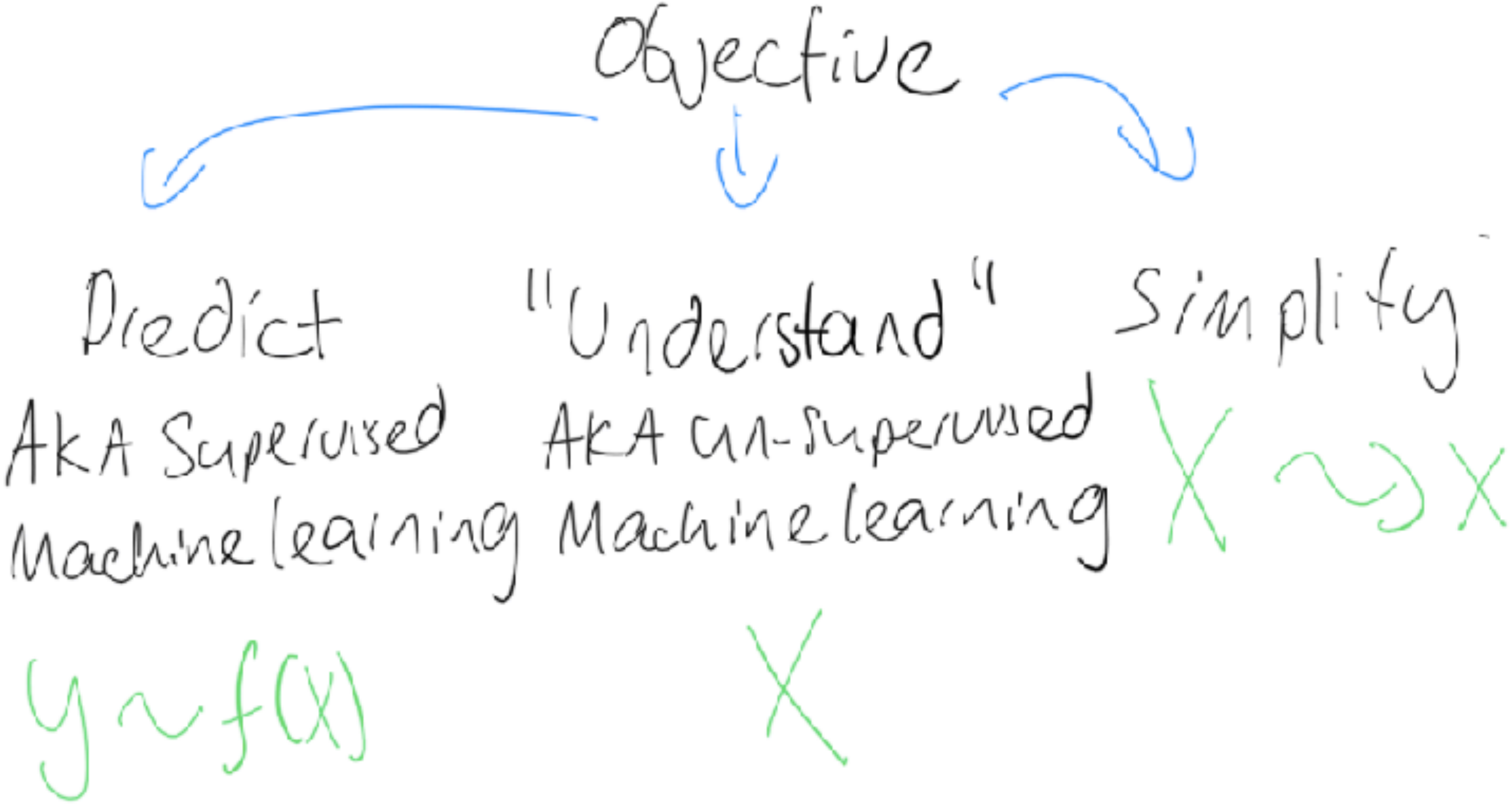
DATA SCIENCE A PRIORTIY



WHY NOW?

‘Big Data’ + Algorithms + **Computing Power** = Data Science

PROBLEMS WE SOLVE



Objective

Predict

AKA Supervised Machine learning

$y \sim f(X)$

"Understand"

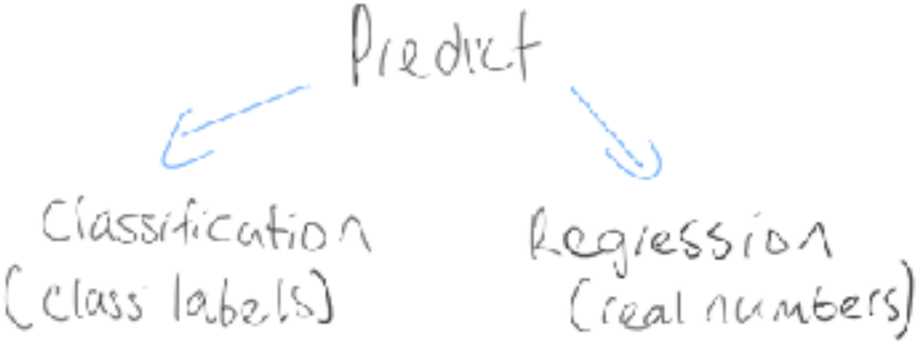
AKA un-supervised Machine learning

X

Simplify

$X \rightsquigarrow x$

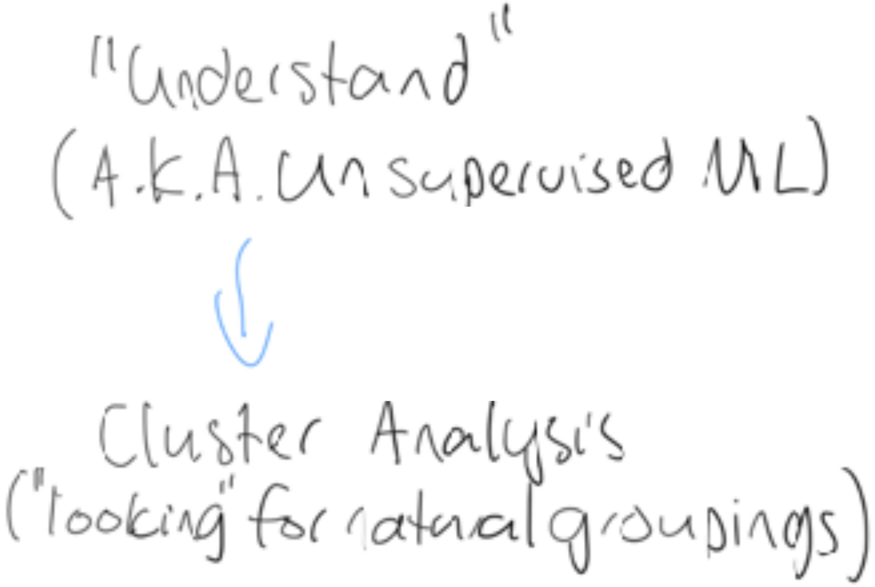
PROBLEMS WE SOLVE



Predict

Classification (class labels)

Regression (real numbers)



"Understand" (A.K.A. Unsupervised ML)

Cluster Analysis ("looking" for natural groupings)

Simplify (Reducing dataset complexity)

INTRODUCTION

THE DATA SCIENCE WORKFLOW

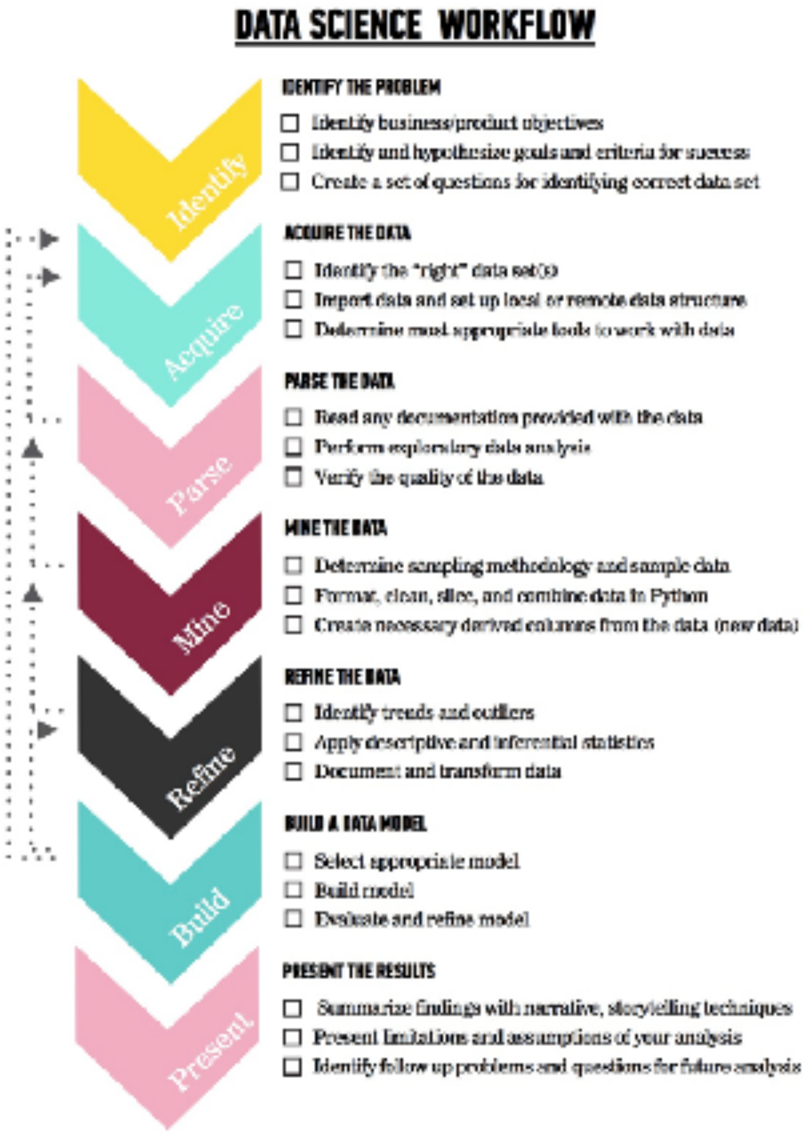
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

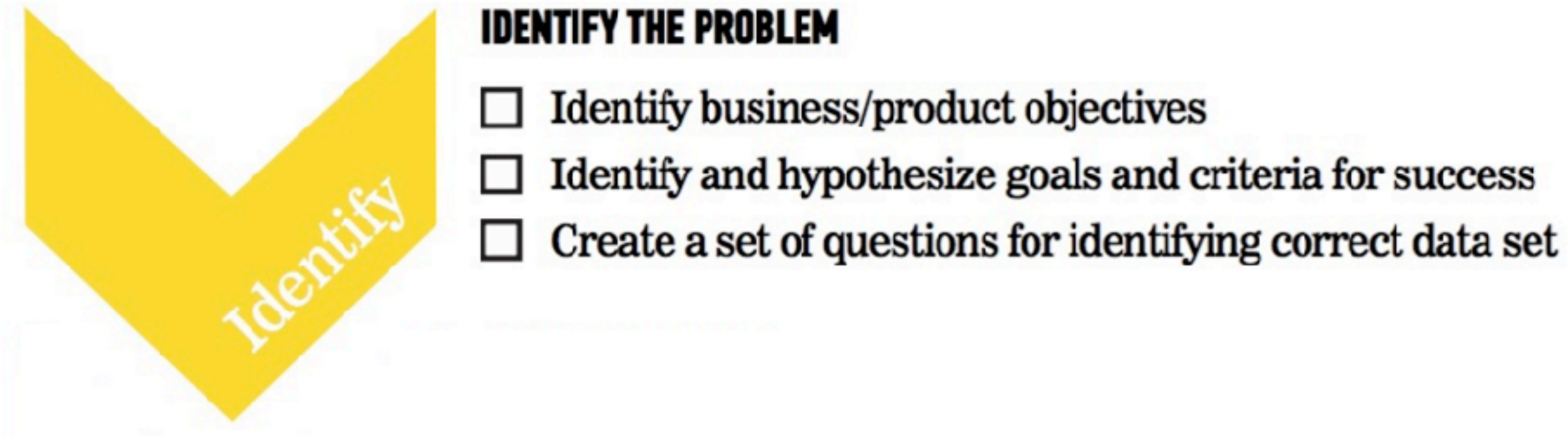
BUILD A DATA MODEL

- □ 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

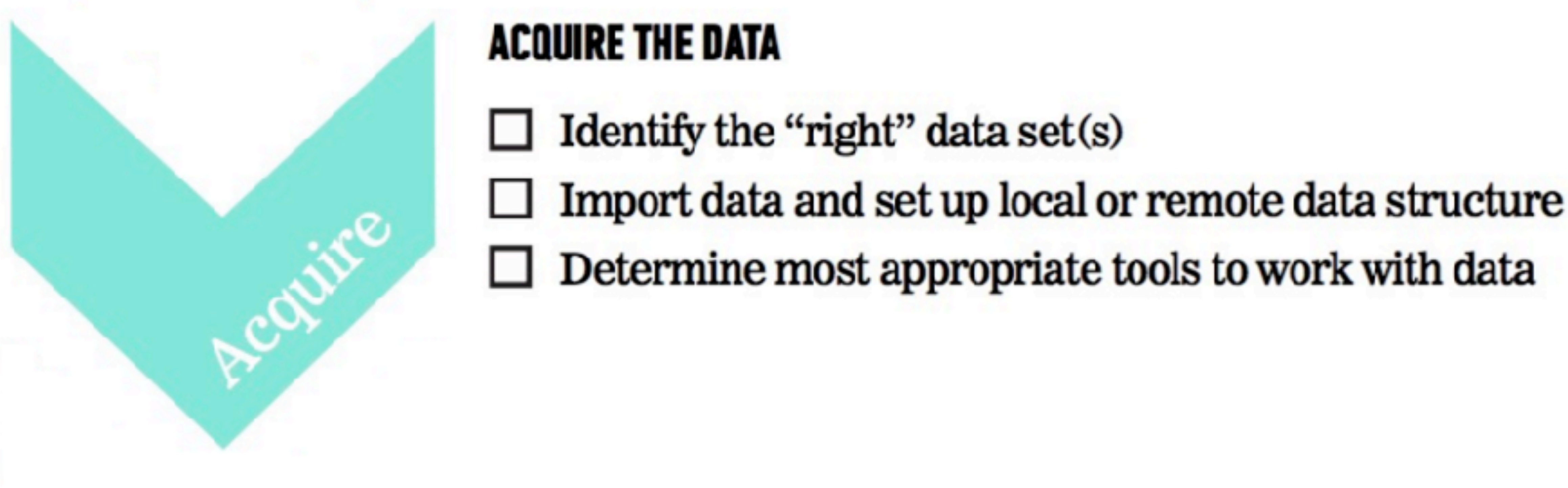
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

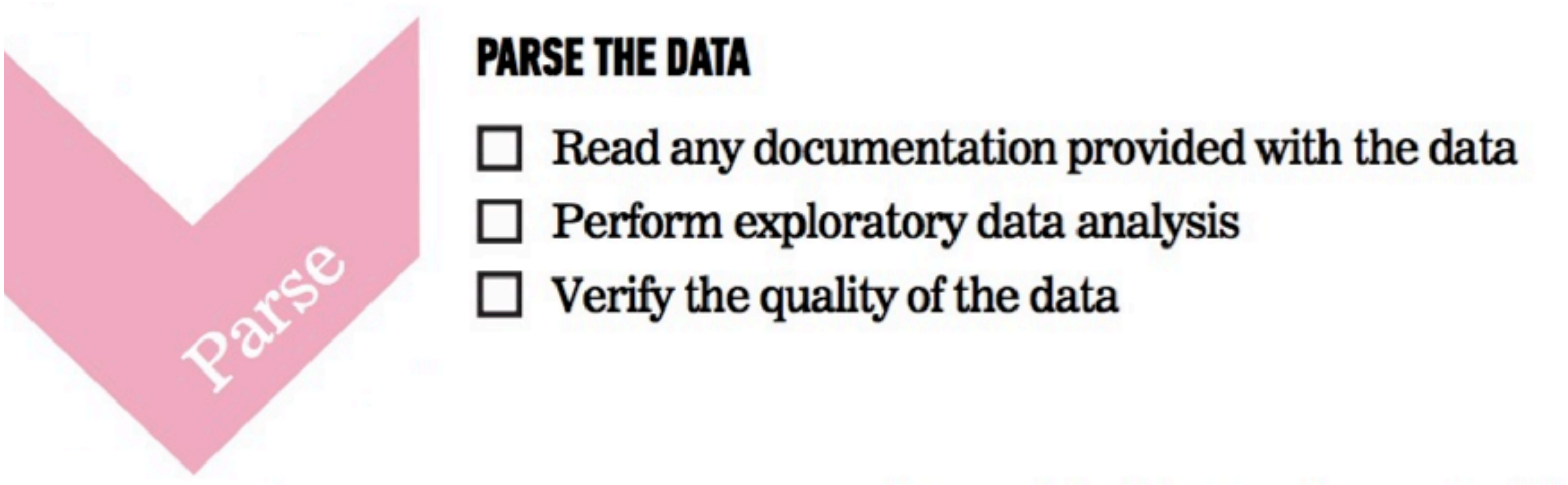
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

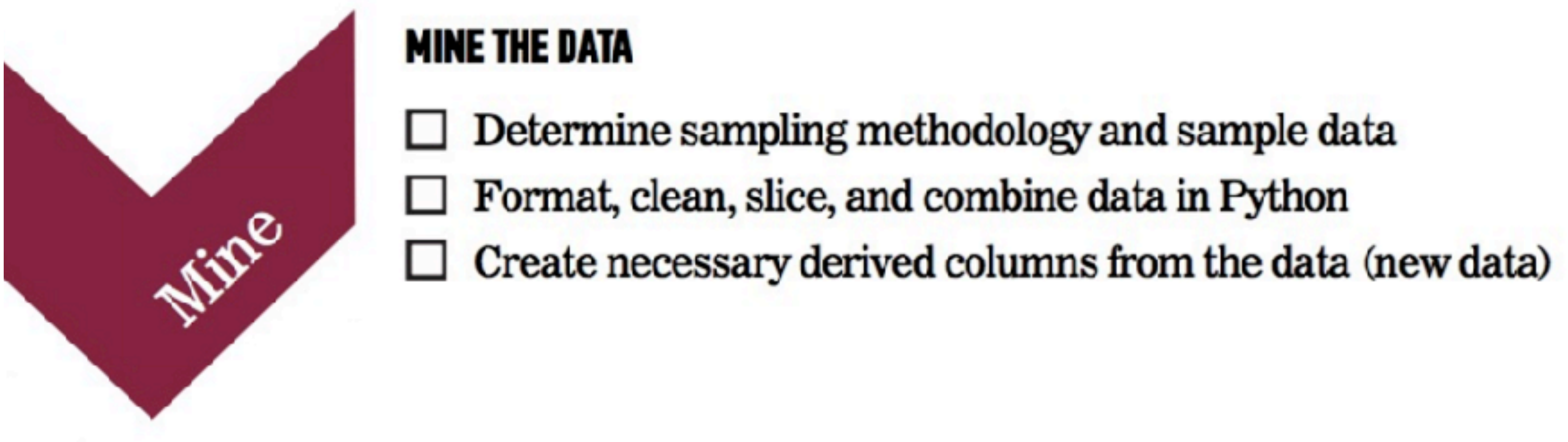
OVERVIEW OF THE DATA SCIENCE WORKFLOW



PARSE THE DATA

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

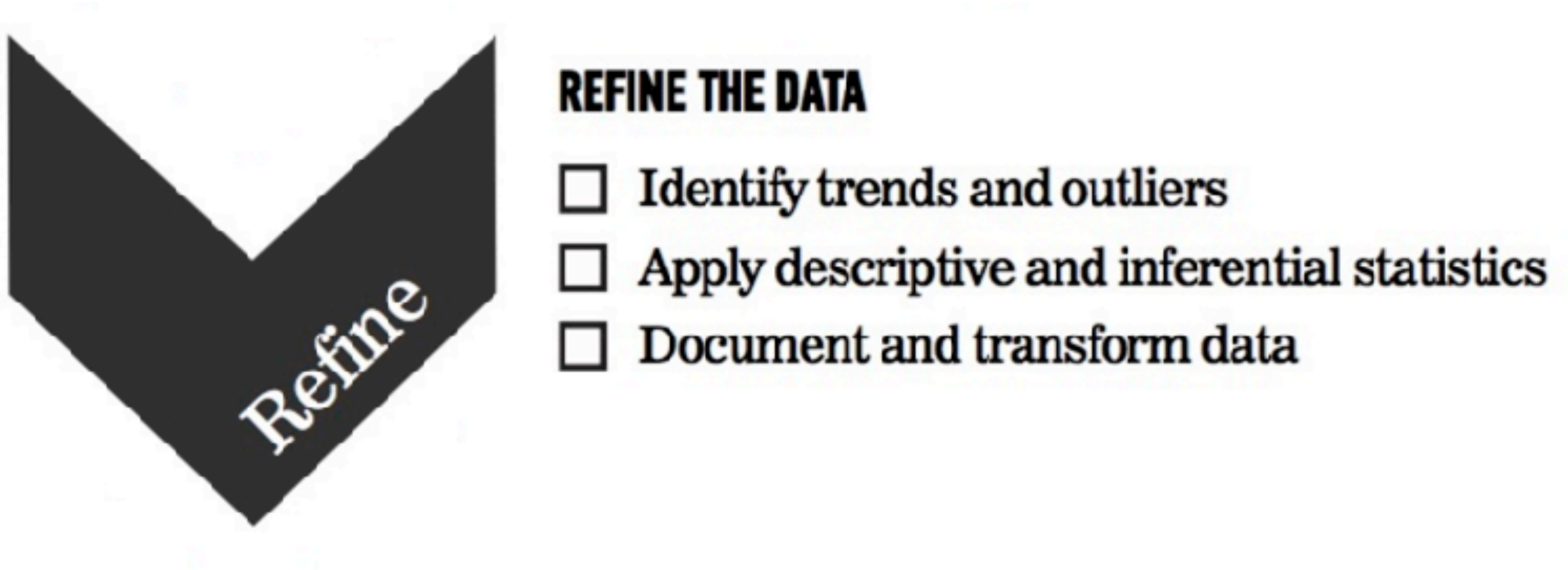
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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)

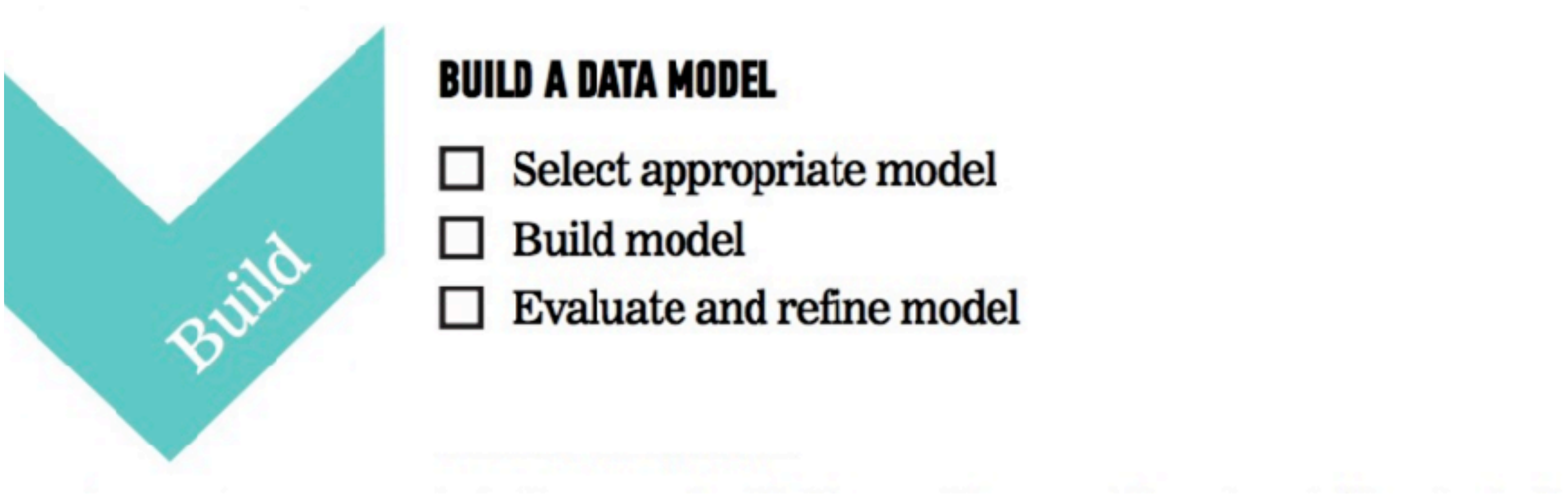
OVERVIEW OF THE DATA SCIENCE WORKFLOW



REFINE THE DATA

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

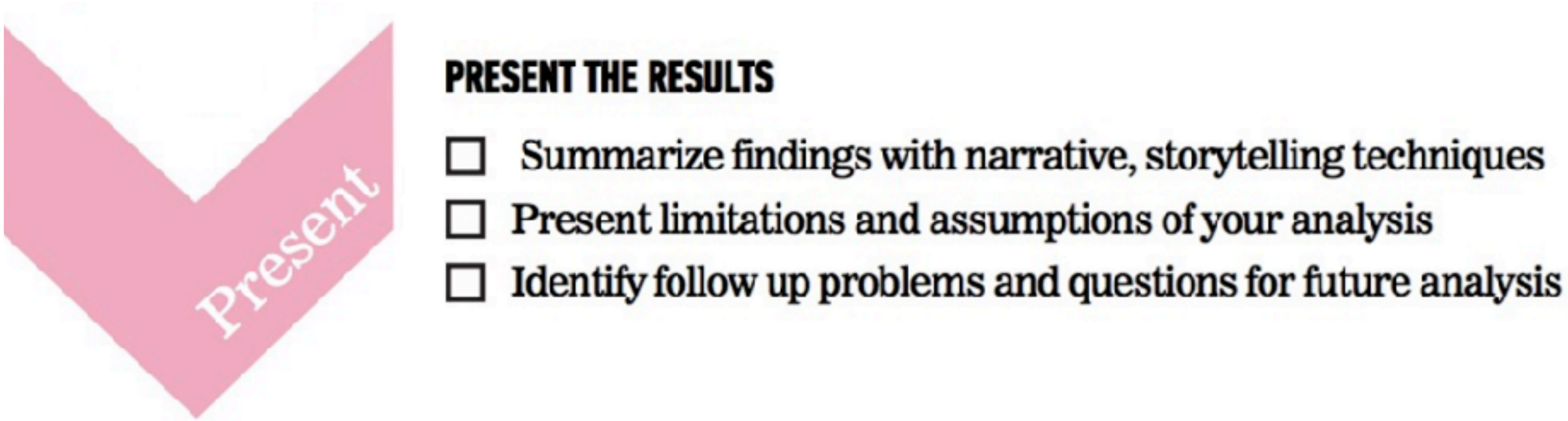
OVERVIEW OF THE DATA SCIENCE WORKFLOW



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

EXAMPLE DATA SCIENCE TOOLBOX

SQL

LETS CODE!