GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

K. Nathaniel Tucker

Quant, Google

LEARNING OBJECTIVES

- Define a problem and types of data
- Identify data set types
- Define the data science workflow
- Apply the data science workflow in the pandas context
- Create an iPython Notebook to import, format, and clean using the Pandas library

COURSE

PRE-EXITS

COURSE

PRE-WORK

PRE-WORK REVIEW

- Create and open an iPython Notebook
- Complete the Python pre-work

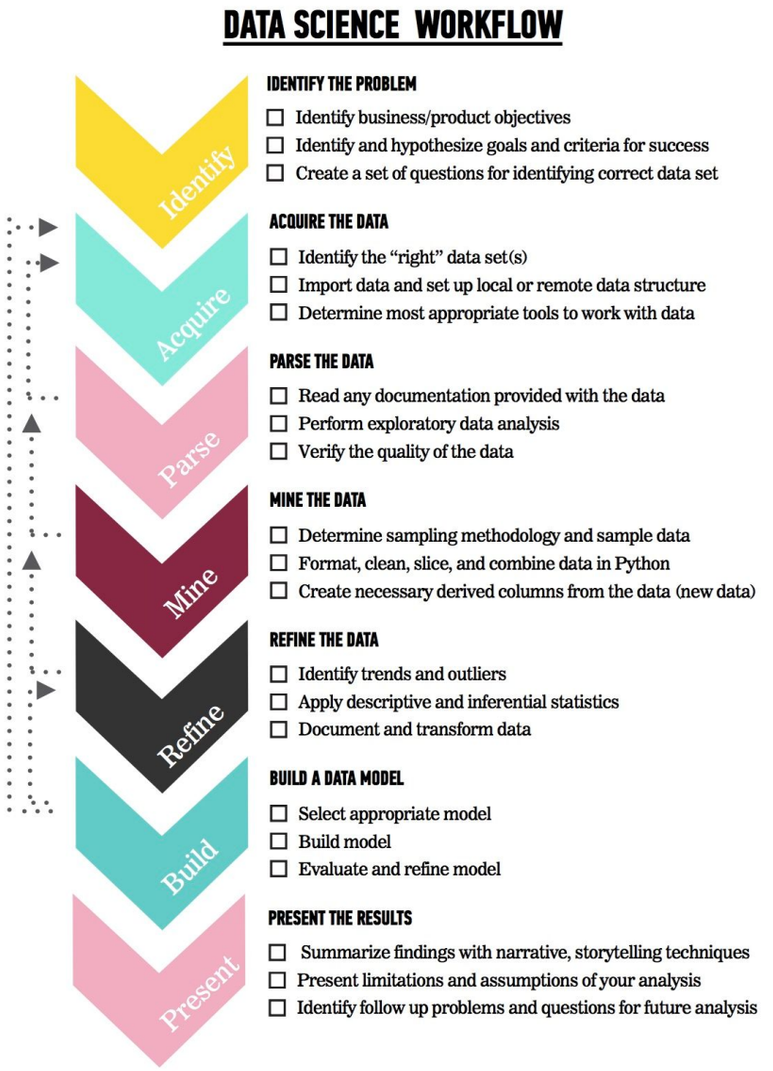
OPENING

EXPERIMENTAL DESIGN AND PANDAS

LET'S REVIEW 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



TODAY

- We’re going to focus on steps 1-2 (Identify the Problem and Acquire the Data).
- We’ll cover steps 3-5 in the next few classes

ASKING A GOOD QUESTION

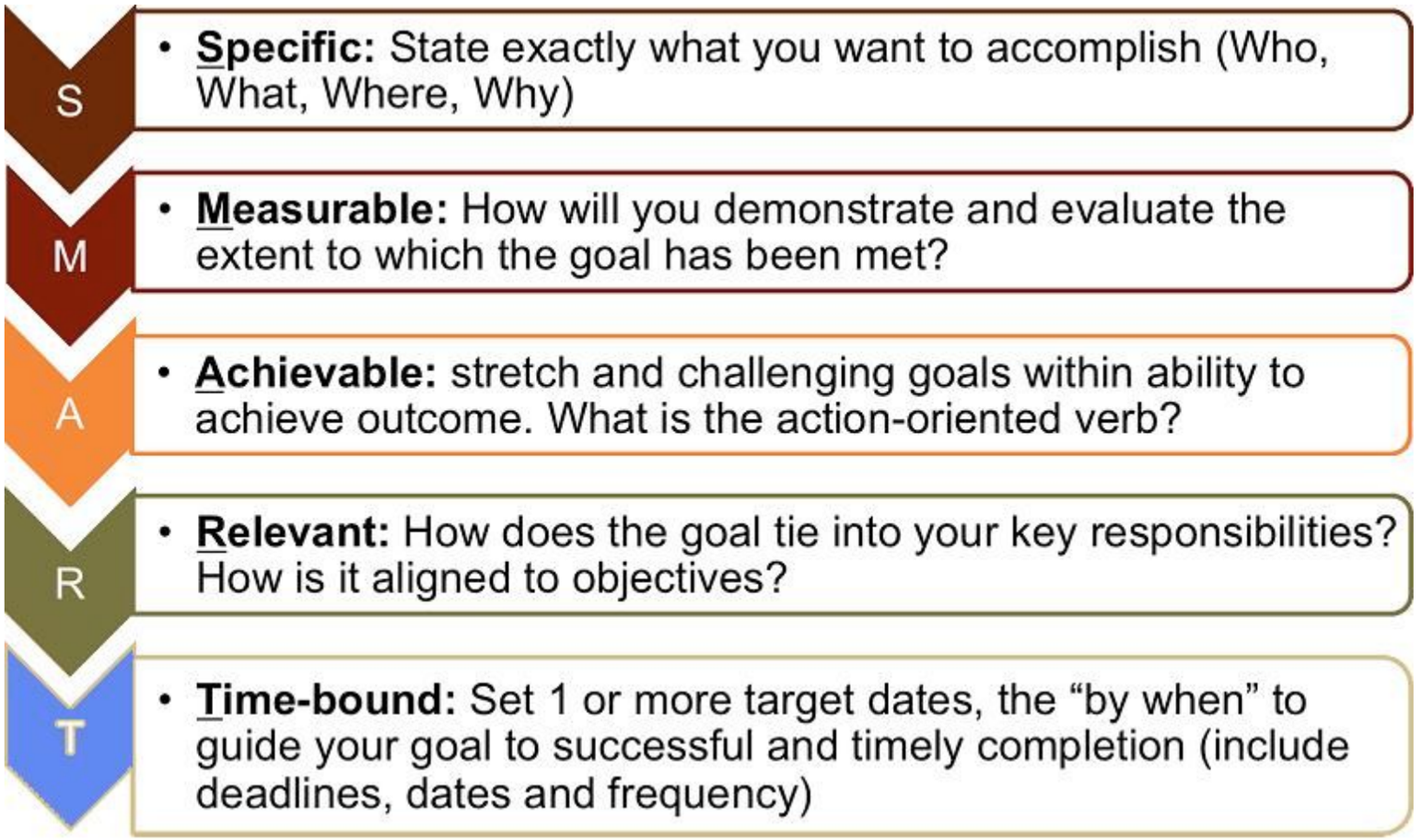
WHY DO WE NEED A GOOD QUESTION?

- “A problem well stated is half solved.” -Charles Kettering
- Sets yourself up for success as you begin analysis
- Establishes the basis for reproducibility
- Enables collaboration through clear goals

WHAT IS A GOOD QUESTION?

- Goals are similar to the SMART Goals Framework.

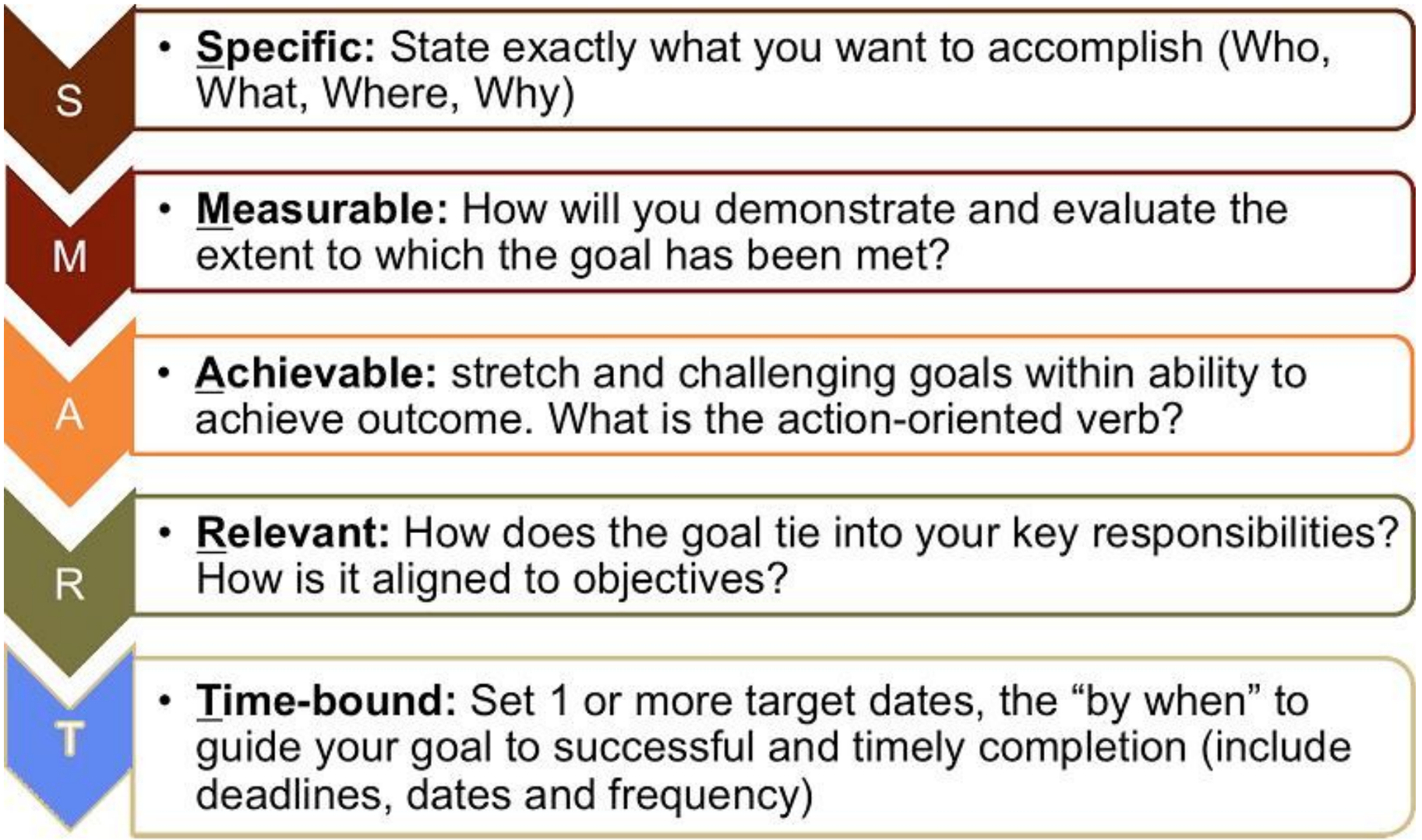
- S: specific
- M: measurable
- A: attainable
- R: reproducible
- T: time-bound



WHAT IS A GOOD QUESTION?

- Specific: The dataset and key variables are clearly defined.
- Measurable: The type of analysis and major assumptions are articulated.
- Attainable: The question you are asking is feasible for your dataset and is not likely to be biased.
- Reproducible: Another person (or future you) can read and understand exactly how your analysis is performed.
- Time-bound: You clearly state the time period and population for which this analysis will pertain.

WHAT IS A GOOD QUESTION?



DEMO

DIAGRAMMING AN AIM

EXAMPLE AIM

- Determine the association of foods in the home with child dietary intake. Using one 24-hour recall from the cross-sectional NHANES 2007-2010 we will determine the factors associated with food available in the homes of American children and adolescents. We will test if reported availability of fruits, dark green vegetables, low fat milk or sugar sweetened beverages available in the home increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for that food.

HYPOTHESIS

- Children will be *more likely* to meet the USDA recommended intake level when food is always available in their home compared to *rarely or never*.

SPECIFIC

- How data was collected:
  - 24-hour recall, self-reported
- What data was collected:
  - Fruits, dark green vegetables, low fat milk or sugar sweetened beverages, always vs. rarely available
- How data will be analyzed:
  - Using USDA recommendations as a gold-standard to measure the association
- The specific hypothesis & direction of the expected associations:
  - Children will be more likely to meet their recommended intake level

MEASURABLE

- Determine the association of foods in the home with child dietary intake.
- We will test if the reported availability of certain foods increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for food.

ATTAINABLE

- Cross-sectional data has inherent limitations; one of the most common is that causal inference is typically not possible.

- Note that we are determining association, not causation.

REPRODUCIBLE

- With all the specifics, it would be straightforward to pull the data from NHANES and reproduce the analysis.

TIME BOUND

- Using one 24-hour recall from NHANES 2007-2010, we will determine the factors associated with food available in the homes of American children and adolescents.

CONTEXT IS IMPORTANT

- The previous example laid out research goals.
- In a business setting, you will need to articulate business objectives.
- Example: Success for the Netflix recommendation engine may be if 70% of customers over the age of 18 select a movie from the recommended queue during Q3 of 2015.
- Regardless of setting, start your question with the SMART framework to help achieve your objectives.

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Which of the following uses the SMART framework? Why? What is missing?

    a. I am looking to see if there is an association with number of passengers with carry on luggage and delayed take-off time.

    b. Determine if the number of passengers on JetBlue, Delta and United domestic flights with carry-on luggage is associated with delayed take-off time using data from flightstats.com from January 2015- December 2015.

DELIVERABLE

Answers to the above questions

WHY DATA TYPES MATTER

- Different data types have different limitations and strengths.
- Certain types of analyses aren’t possible with certain data types.

CROSS-SECTIONAL DATA

- All information is determined at the same time; all data comes from the same time period.

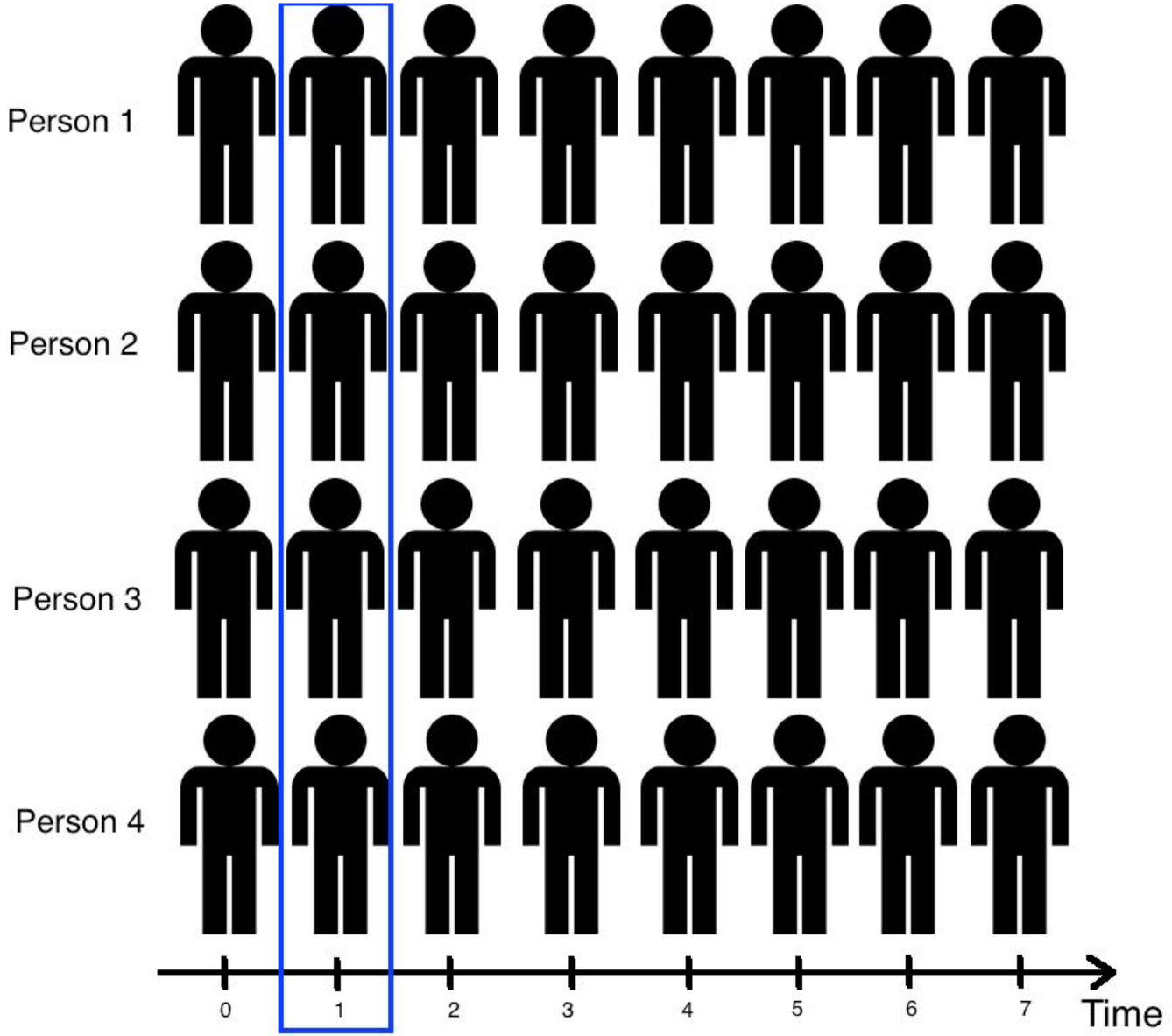
- Issues: There is no distinction between exposure and outcome

CROSS-SECTIONAL DATA

- Strengths
  - Often population based
  - Generalizability
  - Reduce cost compared to other types of data collection methods

- Weaknesses
  - Separation of cause and effect may be difficult (or impossible)
  - Variables/cases with long duration are over-represented

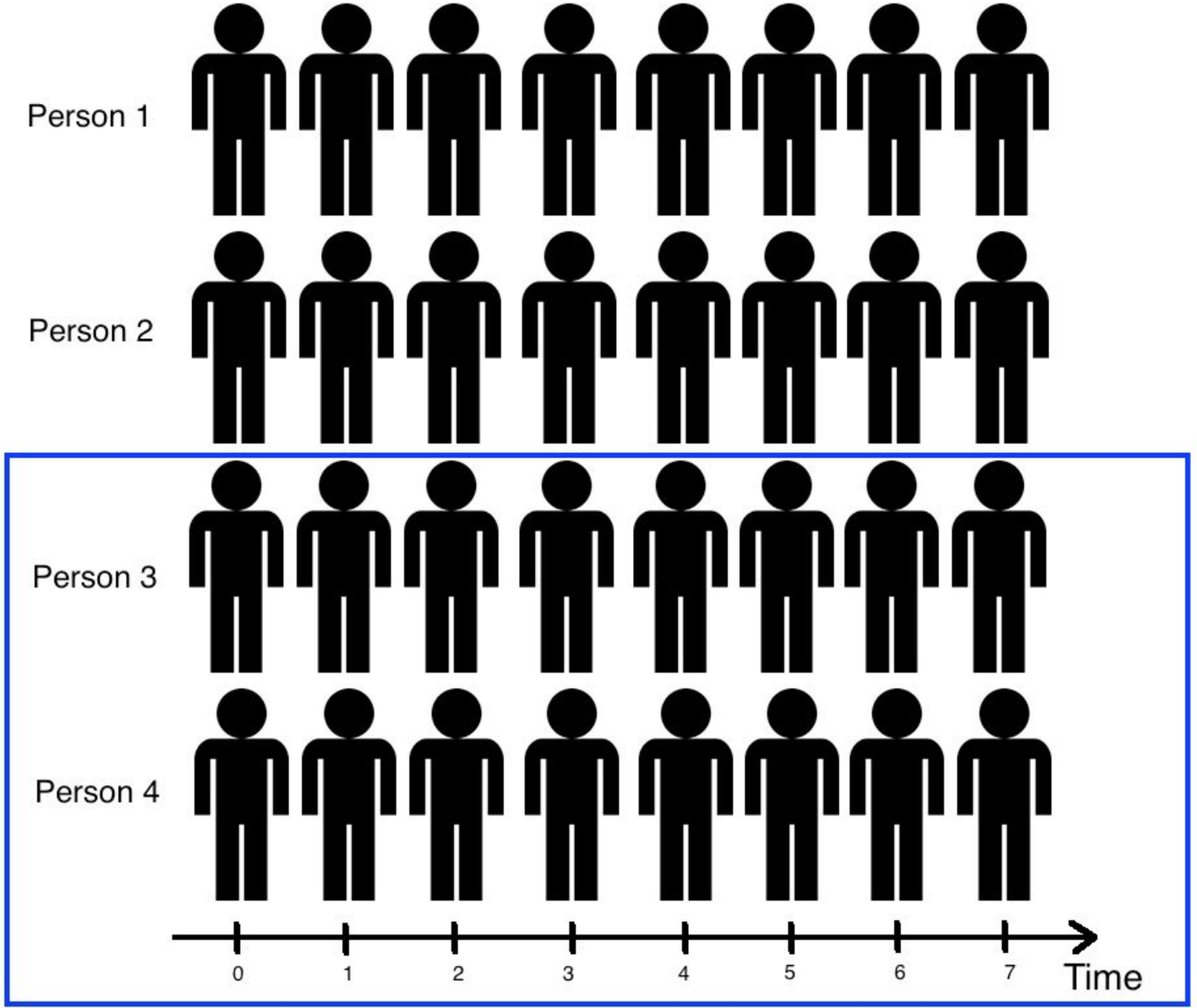
CROSS-SECTIONAL DATA



TIME SERIES/LONGITUDINAL DATA

- The information is collected over a period of time
- Strengths
  - Unambiguous temporal sequence - exposure precedes outcome
  - Multiple outcomes can be measured
- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data

TIME SERIES/LONGITUDINAL DATA



REVIEW

SMART

SMART REVIEW

- The SMART framework covers the “Identify” step of the data science workflow.
- Types of datasets: cross-sectional vs. time series/longitudinal
- Questions?

INTRODUCTION

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

- For the remainder of class, we'll talk about steps 2 & 3 of the data science workflow: acquire and parse
- We'll be using iPython Notebook
- First a demo, then a codealong
- Finally, some hands on practice in a lab

DEMO

WALKTHROUGH ACQUIRE & PARSES WITH PANDAS

ACQUIRE

- Where we determine if we have the “right” dataset for our problem
- Questions to ask:
  - What type of data is it, cross-sectional or longitudinal?
  - How well was the data collected?
  - Is there much missing data?
  - Was the data collection instrument validated and reliable?
  - Is the dataset aggregated?
  - Do we need pre-aggregated data?

LOGISTICS OF ACQUIRING YOUR DATA

- Data can be acquired through a variety of sources
- Web (Google Analytics, HTML, XML)
- File (CSV, XML, TXT, JSON)
- Databases (SQL, NOSQL, etc)
- Today, we'll use a CSV (comma separated file)

PARSE: UNDERSTANDING YOUR DATA

- You need to understand what you're working with.
- To better understand your data
  - Create or review the data dictionary
  - Perform exploratory surface analysis
  - Describe data structure and information being collected
  - Explore variables and data types

CODEALONG

NUMPY AND PANDAS INTRO

DEMO

LAB WALKTHROUGH

LESSON 2 LAB WALKTHROUGH

- In this lab, you will merge two datasets: ozone and data.
- By the end of the lab, you will:
  - Merge datasets
  - Check basic features of the data
  - Find and drop missing values
  - Find basic stats like mean and max

CONCLUSION

TOPIC REVIEW

REVIEW

- Let’s go through the lab. Any questions?
- Today, we’ve talked about
  - Defining a problem
  - Types of data
  - Acquiring and parsing data
  - Using Pandas

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE: Lesson 4

- Project: Unit 1

LESSON

CREDITS

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET