GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

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 Jupyter Notebook to import, format, and clean using the Pandas library

COURSE

PRE-WORK

PRE-WORK REVIEW

- Create and open an Jupyter 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?

- Study design for analytics can use an adaption of the SMART framework.

S: **S**pecific

M: **M**easurable

A: **A**ttainable

R: **R**eproducible

T: **T**ime-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: Is the data available? Does it suit the needs of the study question? Will it be of good enough quality? Is there enough?
- 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.

DEMO

DIAGRAMMING AN AIM

EXAMPLE AIM

- Determine the association of foods in the home with child recommended 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 dietary 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 recommended 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

- Data has already been collected and is freely accessible.

- Study question is correlational in nature (easily examined).
  - Vs. Causational (much harder, typically not examined in data science)

- Collected by National Centre for Health **Statistics**
  - ASSUMED quality and quantity.

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.

    a. 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.

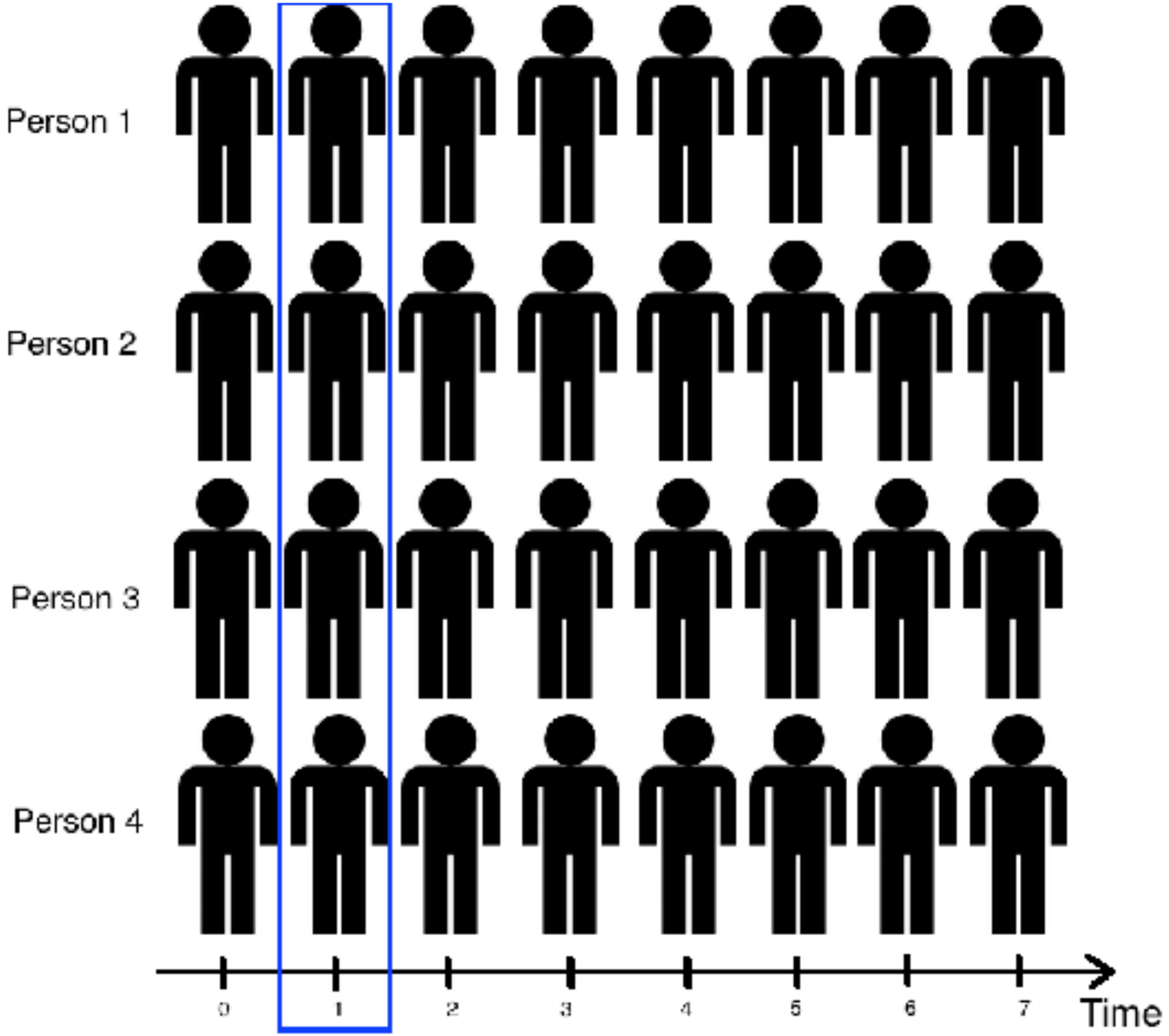
- Issues: There is no distinction between exposure and outcome

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

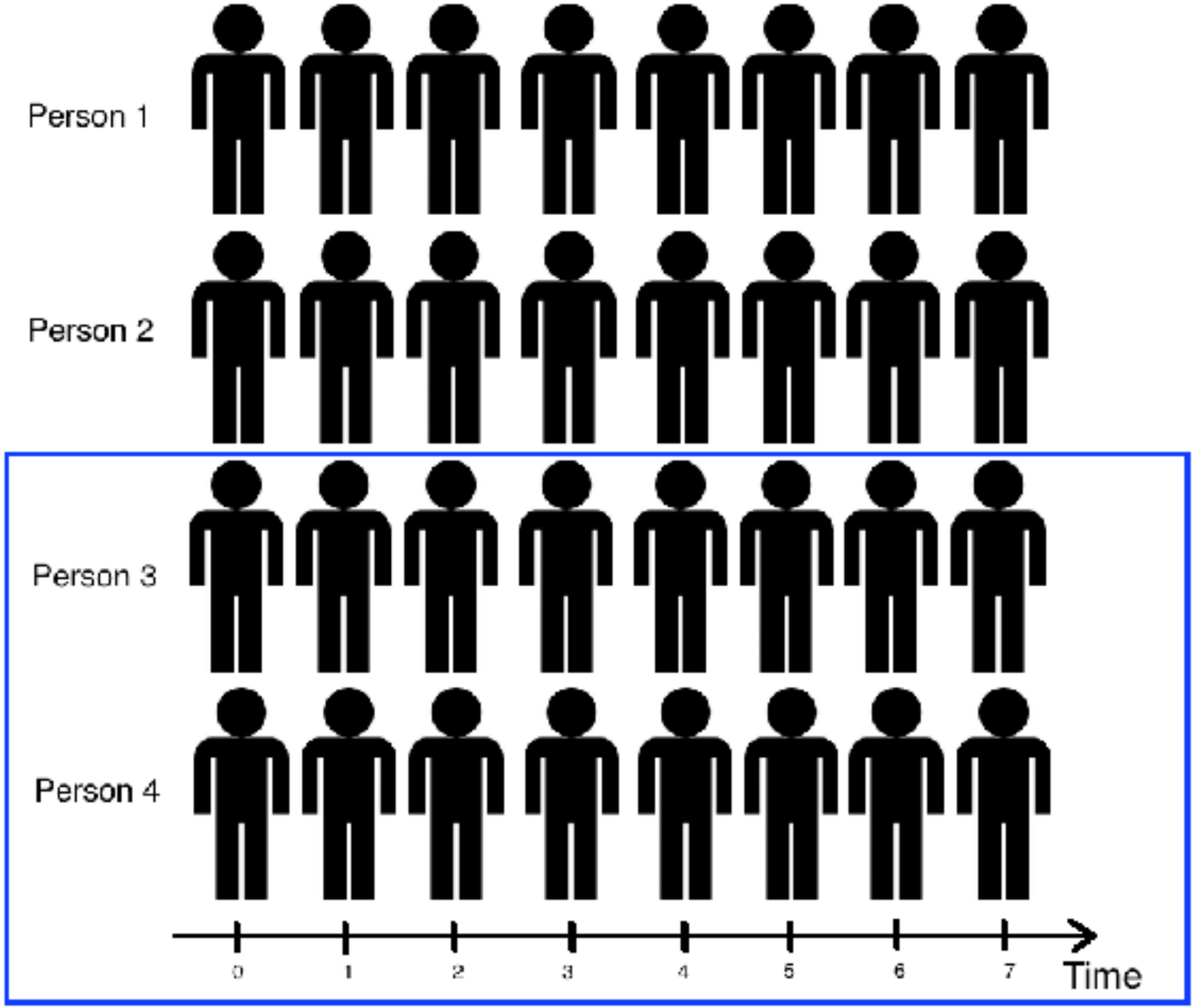
CROSS-SECTIONAL DATA



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



ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. What type of data is the flight delay data?
2. Determine if we can examine if the number of passengers on JetBlue, Delta and United domestic flights with carry-on luggage is associated with 15 minute delayed take-off time from January 2015- December 2015.
3. Can you create a cross-sectional analysis from a longitudinal data collection? How?

DELIVERABLE

Answers to the above questions

GUIDED PRACTICE

WRITE A RESEARCH QUESTION WITH RAW DATA

ACTIVITY: WRITE A RESEARCH QUESTION WITH RAW DATA

DIRECTIONS (10 minutes)

1. Individually, look at the data from Kaggle's Titanic competition and write a high quality research question.
2. Make sure you answer the following questions:
   a. What type of data is this, cross-sectional or longitudinal?
   b. What will we be measuring?
   c. What is the SMART aim for this data?
3. When finished, split into pairs and share your answers with each other.

DELIVERABLE

Research Question

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?

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 Jupyter 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

INTRO TO DATA DICTIONARIES AND DOCUMENTATION

- Data dictionaries help judge the quality of the data.
- They also help understand how it's coded.
  - Does gender = 1 mean female or male?
  - Is the currency dollarydoos or euros?
- Data dictionaries help identify any requirements, assumptions, and constraints of the data.
- They make it easier to share data.

DATA DICTIONARY EXAMPLE: KAGGLE TITANIC DATA

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard
Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances
Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

CODEALONG

NUMPY AND PANDAS INTRO

NUMPY AND PANDAS INTRO

- What are Numpy and Pandas?
  - Python packages
- Pandas is built on Numpy.
- Numpy uses arrays (lists) to do basic math and slice and index data.
- Pandas uses a data structure called a Dataframe.
- Dataframes are similar to Excel tables; they contain rows and columns.
  - A Pandas dataframe is similar to an R dataframe.

NUMPY AND PANDAS INTRO

	A	B	C	D
2014-01-01	0.731803	2.318341	-0.126191	-0.903675
2014-01-02	0.161877	-0.892566	0.967681	-1.514520
2014-01-03	0.776626	1.797420	0.916972	0.634322
2014-01-04	2.020242	-0.763612	1.239145	-0.919727
2014-01-05	0.772058	0.417369	-0.957359	-0.916665
2014-01-06	-1.670217	-3.249906	2.017370	1.674340

6 rows × 4 columns

NUMPY AND PANDAS INTRO

- With these packages, you can select pieces of data, do basic operations, calculate summary statistics.

- Follow along and code along as we learn about Numpy and Pandas.

DEMO

LAB WALKTHROUGH

LESSON 2 LAB WALKTHROUGH

- By the end of the lab, you will:
  - Check basic features of the data
  - Select and filter data
  - Find basic stats like mean and max
  - Be introduced to function documentation

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

- Project: Unit 1 (we'll go over it now)
  - Due **TUESDAY (27TH MARCH)**

- Bookmark and look over https://chrisalbon.com/ (great resource for DS in Python).