GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

Angelo Klin
Katra Analytics

LEARNING OBJECTIVES

- Define a problem and types of data
- Identify dataset types
- Define the Data Science Workflow
- Apply the Data Science Workflow in the context of Pandas
- Create a Jupyter Notebook to import, format and clean data using the Pandas library

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Create and open a Jupyter Notebook
- Complete the Python pre-work

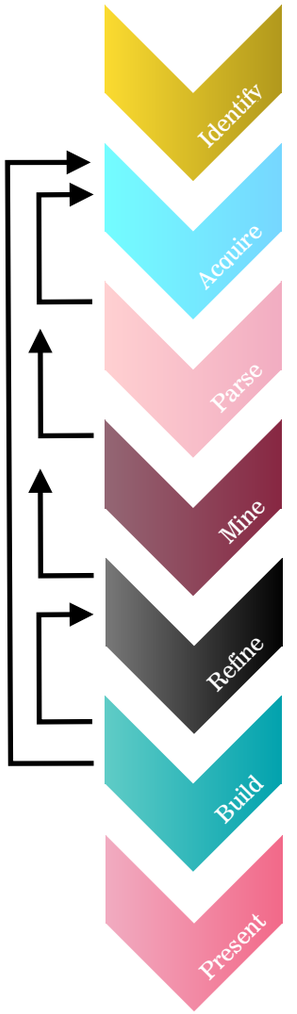
OPENING

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 are going to focus on steps 1 and 2 and some of 3
  1. Identify the Problem
  2. Acquire the Data
  3. Parse the Data

ASKING A GOOD QUESTION

WHY DO WE NEED A GOOD QUESTION?

- “A problem well stated is a problem half solved.”
  - Charles F. 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
  - **Specific**: State **exactly** what you want to accomplish (Who, What, Where, Why)
  - **Measurable**: How will you **demonstrate and evaluate** the extent to which the goal has been met?
  - **Attainable**: Stretch and challenging goals within the ability to achieve the outcome. What is the **action-oriented verb**?
  - **Relevant**: How does the goal tie into your key responsibilities? How is it aligned to objectives?
  - **Time-bound**: Set one or more target dates, the “by when” to guide your goal to successful and timely completion (include deadlines, dates and frequency)

SMART GOALS FRAMEWORK

Letter	Most Common	Alternatives
S	Specific	(Strategic and Specific)
M	Measurable	Motivating
A	Achievable	Agreed, Attainable, Action-oriented, Ambitious, Aligned with corporate goals
R	Relevant	Realistic, Resourced, Reasonable, (Realistic and Resourced), Results Based
T	Time-bound	Trackable, Time-based, Time/Cost limited, Timely, Time-sensitive, Timeframe

- https://en.wikipedia.org/wiki/SMART_criteria

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

DEMONSTRATION

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 the 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
  - Source: Dr. Amy Roberts' Dissertation

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 analysed
  - Using USDA recommendations as a gold-standard to measure the association
- The specific hypothesis and direction of the expected associations
  - Children will be more likely to meet their recommended intake level when a food is always available in their home

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

DIRECTIONS: 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 2016-December 2016

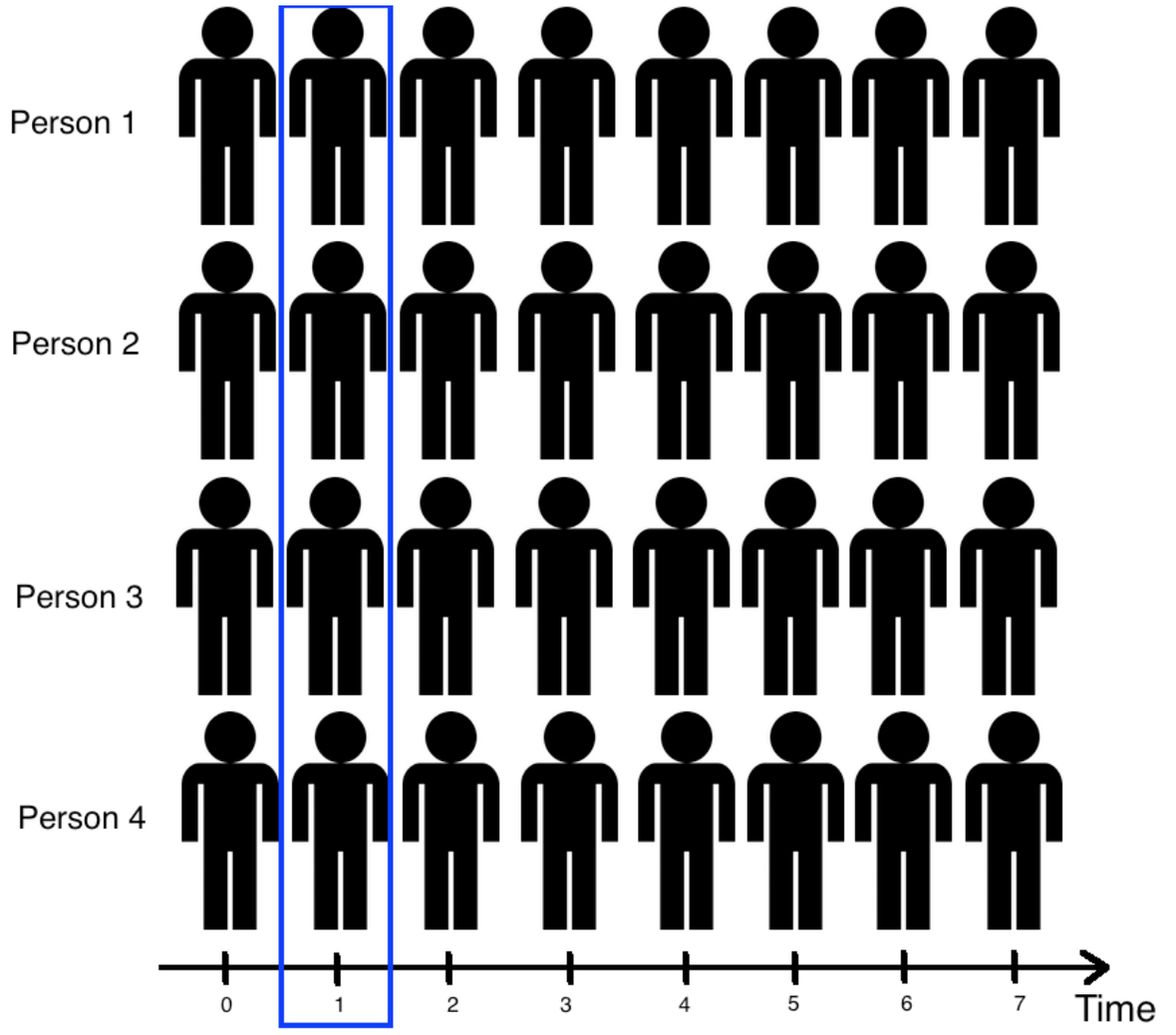
WHY DATA TYPES MATTER?

- Different data types have different limitations and strengths
- Certain types of analyses are not 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



CROSS-SECTIONAL DATA

- Strengths
  - Often population-based
  - Generalisability
  - 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

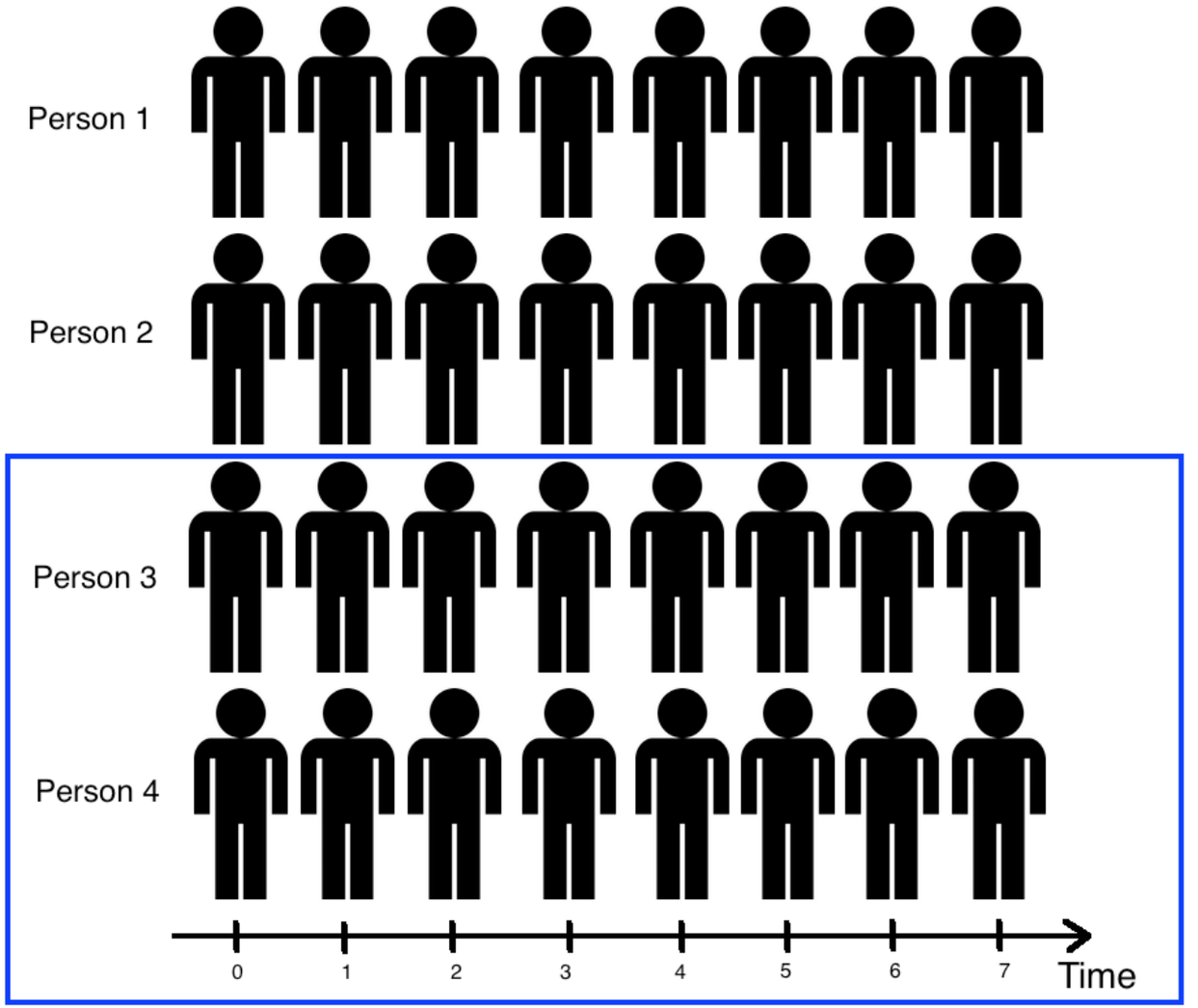
TIME-SERIES / LONGITUDINAL DATA

- 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

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (5 MINUTES)

1. What type of data is the flight stats data?
2. 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.
3. Can you create a cross-sectional analysis from a longitudinal data collection? How?

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

REVIEW

SMART

SMART REVIEW

- The SMART framework covers the “Identify” step of the Data Science Workflow
- Types of datasets: cross-sectional versus time-series / longitudinal
- Questions?

DATA SCIENCE WORKFLOW: ACQUIRE AND PARSE

DATA SCIENCE WORKFLOW: ACQUIRE AND PARSE

- For the remainder of class, we will talk about steps 2 and 3 of the Data Science Workflow: Acquire and Parse
- We will be using Jupyter Notebook
- First a demonstration, then a code along
- Finally, some hands-on practice in a lab

DEMONSTRATION

WALKTHROUGH ACQUIRE AND PARSE 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 many missing data?
  - Was the data collection instrument validated and reliable?
  - Is the dataset aggregated?
  - Do we need pre-aggregated data?

LOGISTICS OF ACQUIRING 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 will use a CSV (comma separated value)

PARSE: UNDERSTANDING YOUR DATA

- You need to understand what you are working with
- To better understand your data
  - Create or review the Data Dictionary (AKA Code Book)
  - Perform Exploratory Data Analysis
  - Describe data structure and information being collected
  - Explore variables and data types

INTRODUCTION TO DATA DICTIONARIES AND DOCUMENTATION

- Data Dictionaries help judge the quality of the data
- They also help understand how it is coded
  - Does `gender = 1` mean female or male?
  - Is the currency dollars 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 Fiancés 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
neighbours in a village, however, the definitions do not
support such relations.
```

INTRODUCTION TO DATA CODE BOOKS

- The US National Epidemiological Survey on Alcohol and Related Conditions (NESARC) is a survey designed to determine the magnitude of alcohol use and psychiatric disorders in the US population. It is a representative sample of the non-institutionalised population 18 years and older.
  - NESARC codebook (.pdf)

- The Mars Craters Study, presents a global database that includes over 300,000 Mars craters 1 km or larger that were created between 4.2 and 3.8 billion years ago during a period of heavy bombardment (i.e. impacts of asteroids, proto-planets, and comets).
  - Mars Crater codebook (.pdf)

CODE ALONG

NUMPY AND PANDAS INTRODUCTION

NUMPY AND PANDAS INTRODUCTION

- **NumPy** (pronounced “Numb Pie”) is an extension to the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.
  - NumPy site
  - NumPy on Wikipedia
- **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
  - pandas site
  - pandas on Wikipedia

NUMPY AND PANDAS INTRODUCTION

- 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 analogous to spread sheets and database tables
  - they contain rows and columns

NUMPY AND PANDAS INTRODUCTION

```
import numpy  as np
import pandas as pd
df = pd.DataFrame(np.random.randn(12, 4),
                  index   = pd.date_range(start   = '2017-1-1',
                                          periods = 12,
                                          freq    = 'M'),
                  columns = list('ABCD'))

df
                   A         B         C         D
2016-01-31  0.303552  1.091355 -1.334006  0.650964
2016-02-29  0.185949 -0.376847 -0.659672  0.803404
2016-03-31  0.062719  2.018625  0.761613  0.414411
```

NUMPY AND PANDAS INTRODUCTION

- With these packages, you can select pieces of data, do basic operations and calculate summary statistics
- Follow along and code along as we learn about Numpy and Pandas
  - ~/lessons/lesson-02/code/numpy-and-pandas.ipynb

NUMPY AND PANDAS INTRODUCTION

- We often have to merge data together, correct missing data and plot our findings

- Once again, follow and code along

DEMONSTRATION

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
- ~/lessons/lesson-02/code/starter/starter-2.ipynb

REVIEW

REVIEW

- Today we have talked about
  - Defining a problem
  - Types of data
  - Acquiring and Parsing data
  - Using Pandas

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project
  - Unit Project 1

Q & A

EXIT TICKETS

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

Exit Ticket Link

What's the lesson number?	02
What was the topic of the lesson?	Experimental Design and Pandas