GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

Tan Kwan Chong

Chief Data Scientist, Booz Allen Hamilton

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

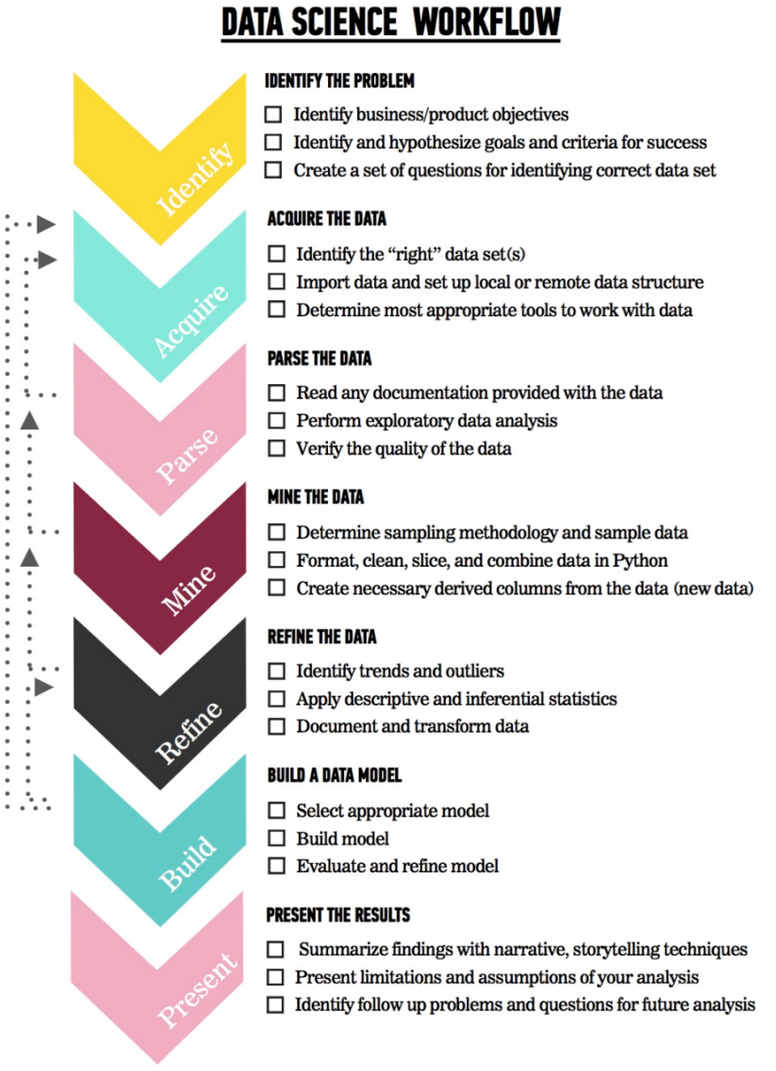
OPENING

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) during the lecture and touch on steps 2-4 in the lab

ASKING A GOOD QUESTION

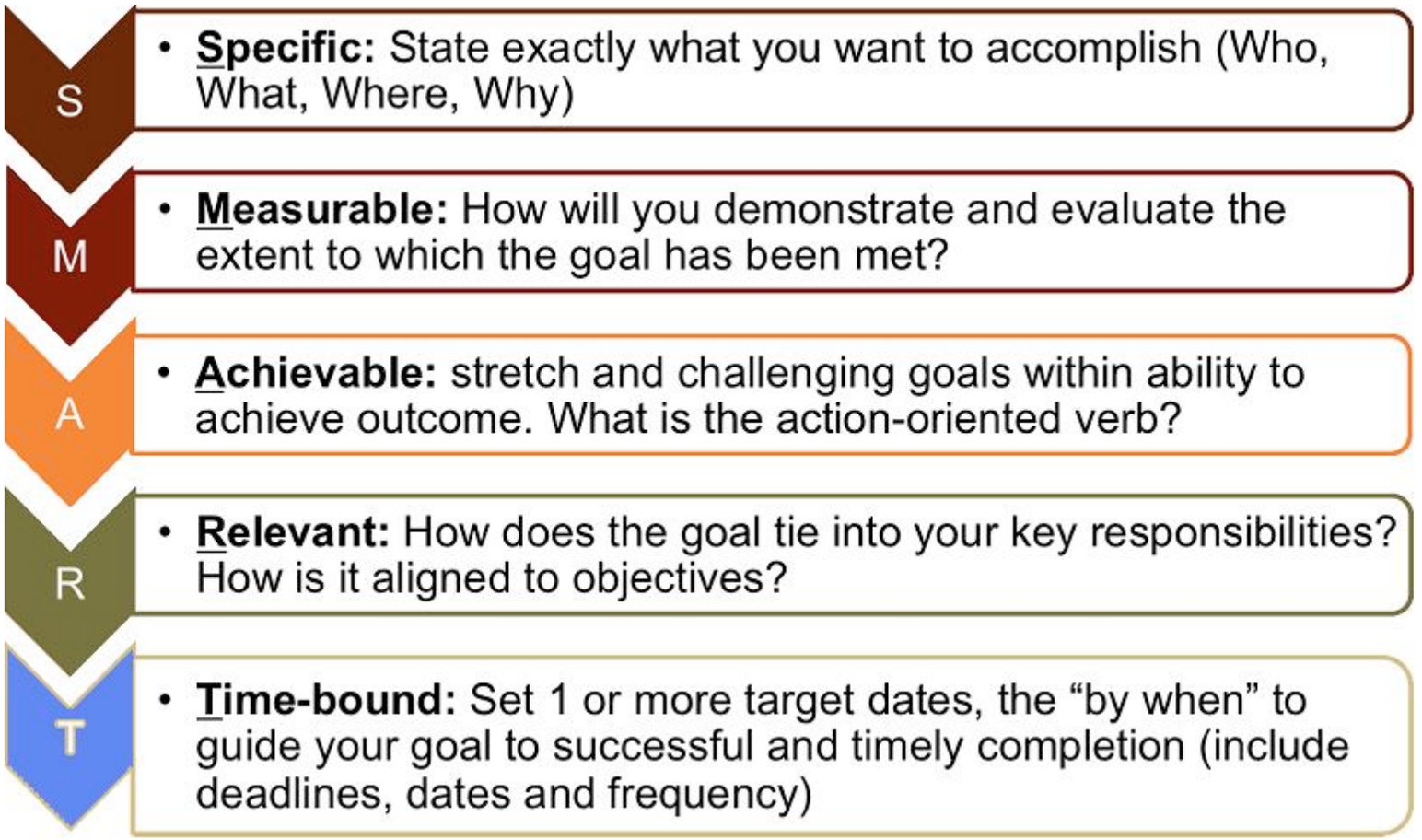
WHY DO WE NEED A GOOD QUESTION?

- “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?

- Analysis 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.

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.

EXAMPLE AIM

Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue

Teck-Hua Ho[a,b,1], Juin Kuan Chong[c], and Xiaoyu Xia[d]

[a]Office of the Deputy President (Research & Technology), National University of Singapore, Singapore 119077; [b]Haas School of Business, University of California, Berkeley, CA 94720; [c]National University of Singapore Business School, National University of Singapore, Singapore 119245; and [d]Department of Decision Sciences and Managerial Economics, Chinese University of Hong Kong Business School, Chinese University of Hong Kong, Shatin, NT, Hong Kong

Edited by George A. Akerlof, University of California, Berkeley, CA, and approved January 31, 2017 (received for review August 3, 2016)

Is there a link between the color of a taxi and how many accidents it has? An analysis of 36 mo of detailed taxi, driver, and accident data (comprising millions of data points) from the largest taxi company in Singapore suggests that there is an explicit link. Yellow taxis had 6.1 fewer accidents per 1,000 taxis per month than blue taxis, a 9% reduction in accident probability. We rule out driver difference as an explanatory variable and empirically show that because yellow taxis are more noticeable than blue taxis—especially when in front of another vehicle, and in street lighting—other drivers can better avoid hitting them, directly reducing the accident rate. This finding can play a significant role when choosing colors for public transportation and may save lives as well as millions of dollars.

car color | road safety | data science | transportation science | sensory perception

Accidents involving public transport are common and cause significant economic losses as well as loss of human life. Applying statistical analysis to a unique and comprehensive dataset we establish that a change in color can avert a significant number of taxi accidents, leading to a reduction in economic losses. Specifically, analysis of a complete set of accident records from the largest taxi operator in Singapore, which uses yellow and blue taxis, shows that yellow is safer than blue because yellow is more noticeable, with the result that potential accidents are avoided by other drivers' timely responses.

Yellow has been a popular color for taxis since 1907, when the Chicago Yellow Cab Company chose the color based on a survey conducted at the University of Chicago. The survey showed that yellow was the most noticeable color, which would make it easy for potential passengers to spot a yellow taxi in the sea of mass-produced black cars prevalent at the time (until 1914, "Japan Black" was the only paint color that would dry fast enough to be used in Ford's mass-production process). More than a century

demographic characteristics. These two datasets include millions of observations on the company's drivers and taxis, and accidents involving these taxis. The data from both datasets have been anonymized and are available in Datasets S1–S6.

The company uses yellow or blue for all its regular taxis (approximate colors are shown in Fig. 1).[†] The colors are the remnants of a 2002 merger that took place between two taxi companies, one of which used yellow and the other, blue. The company owns ~16,700 taxis in a ratio of one yellow to three blue (1y:3b), which translates to 4,175 yellow taxis and 12,525 blue ones. These account for 60% of the ~27,800 taxis in Singapore.[‡]

To control for the difference in the number of taxis used by the company (1y:3b), we calculated a normalized accident rate using the average number of accidents that occurred per 1,000 taxis

Significance

This paper examines the phenomenon that yellow taxis have fewer accidents than blue taxis. Statistical analysis of a unique and comprehensive dataset suggests that the higher visibility of the color yellow makes it easier for other drivers to avoid getting into accidents with yellow taxis, leading to a lower accident rate. This suggests that color visibility should play a major role in determining the colors used for public transport vehicles.

Author contributions: T.-H.H. and J.K.C. designed research; T.-H.H. and J.K.C. performed research; T.-H.H., J.K.C., and X.X. contributed new reagents/analytic tools; T.-H.H., J.K.C., and X.X. analyzed data; and T.-H.H., J.K.C., and X.X. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: dprhoth@nus.edu.sg.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1612551114/-/DCSupplemental.

Determine the association of the color of a taxi and its accident rate. Using a dataset on a randomly chosen sample of 20% (3,341 drivers) of the largest taxi company in Singapore which includes 30 months of data on their taxi contracts, accident records and basic demographic characteristics. We will examine if the color of the taxi increases the likelihood that a taxi will be involved in an accident on a given month.

HYPOTHESIS

- The color of the taxi has an impact on the accident rate. In particular, yellow taxis will have a lower accident rate than blue taxis because they are more noticeable.

SPECIFIC

- How data was collected:
  - The data was provided directly from the taxi company
- What data was collected:
  - Dataset is a randomly chosen sample of 20% (3,341 drivers) of the company's drivers which includes 30 months of data (January 2012 to June 2014) on their taxi contracts, accident records, and basic demographic characteristics

MEASURABLE

- A regression model will developed to examine the statistical significance (p-value) of the color of the taxi on the accident rate

ATTAINABLE

- The dataset obtained is a representative randomized sample of yellow and blue taxi drivers operated by the company
- The dataset contains information on the color of the taxi, demographic information on the driver, and accident records

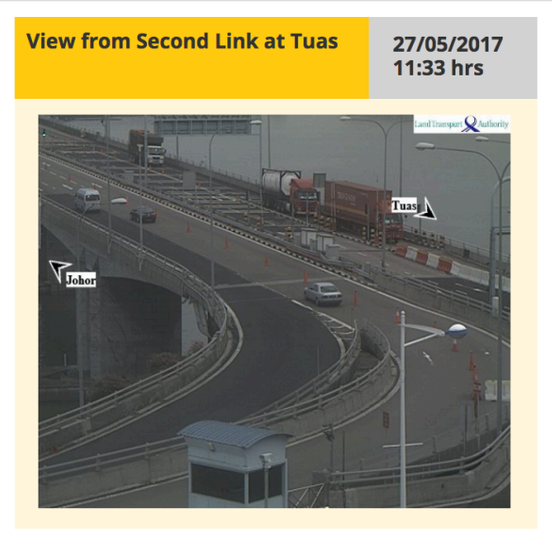
REPRODUCIBLE

- The dataset has been made available online for download by the authors of the research study

TIME BOUND

- Using 30 months (January 2012 to June 2014) of data across 3,341 drivers, we will determine the association of the color of taxi on accident rate

EXAMPLE AIM



Travelling across the causeway is subject to heavy traffic congestion which results in unnecessary time wastage. We want to create a model that can predict the likelihood of traffic congestion at a given day and time of day. We will collect traffic images at the checkpoints using the LTA API over a one week period at ten minute intervals. These images will be manually labeled and used to train a Convolutional Neural Network classifier. We will use 70% of the samples for training and the remainder for testing and desire for the accuracy of the classifier to be > 90%. If this accuracy is achieved, we will collect images for another eight weeks for the classifier to tag. For predictions, we will round the desired timing to the nearest 10 minute interval and take the average of the available sample points.

HYPOTHESIS

- There will be greater traffic congestion during peak commuting hours i.e. before and after office hours as well as during the weekends

SPECIFIC

- How data was collected:
  - Using the LTA API at a ten minute interval over nine weeks total
- What data was collected:
  - Image data from four traffic cameras at the two checkpoints
- How data will be analyzed:
  - One week of data will be manually tagged and used to train a classification model
  - The model will then be used to to make predictions on the remaining eight weeks of data which will then be used for the forecasting model

MEASURABLE

- We will set aside 30% of the first week data collected to be using as the testing set to verify the accuracy of the model
- If the accuracy of the model is greater than 90%, we will proceed to use it to tag the remaining data

ATTAINABLE

- A human is able to determine whether there is traffic congestion based on the camera image
- Convolutional Neural Networks have proven successful at high levels of accuracy for numerous image recognition and classification tasks

REPRODUCIBLE

- The LTA API is publicly accessible and other analysts can retrieve the images as well
- However, there might be differences in the accuracy of the classification of the model as well as the predictions of the forecasting model depending on the time period in which the images are collected

TIME BOUND

- Using one week of traffic image data from four causeway checkpoint cameras we will train a classifier and subsequently use that classifier to tag eight weeks of data for the forecasting model

CONTEXT IS IMPORTANT

- 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.

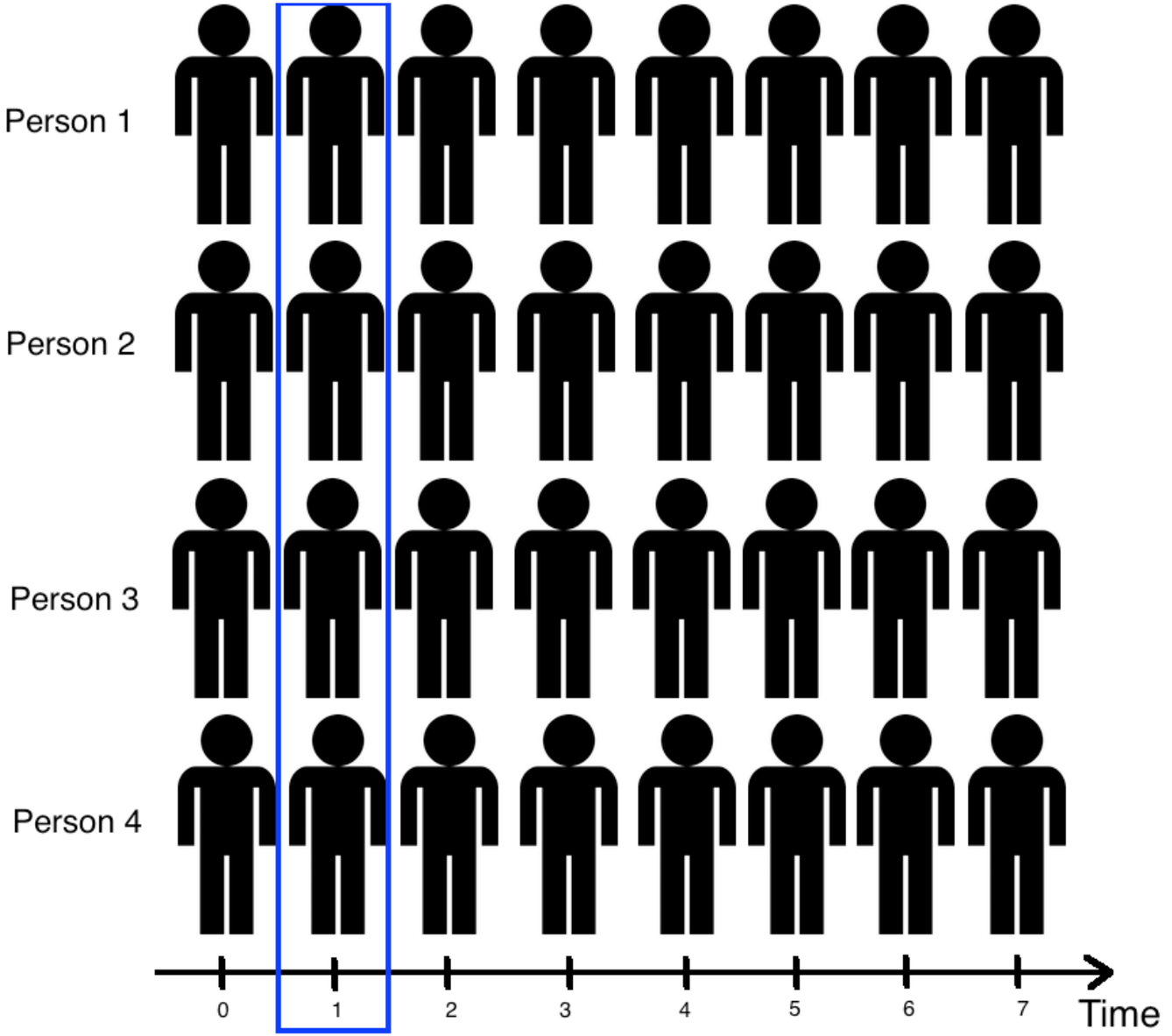
- Issues: There is no distinction between exposure and outcome

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



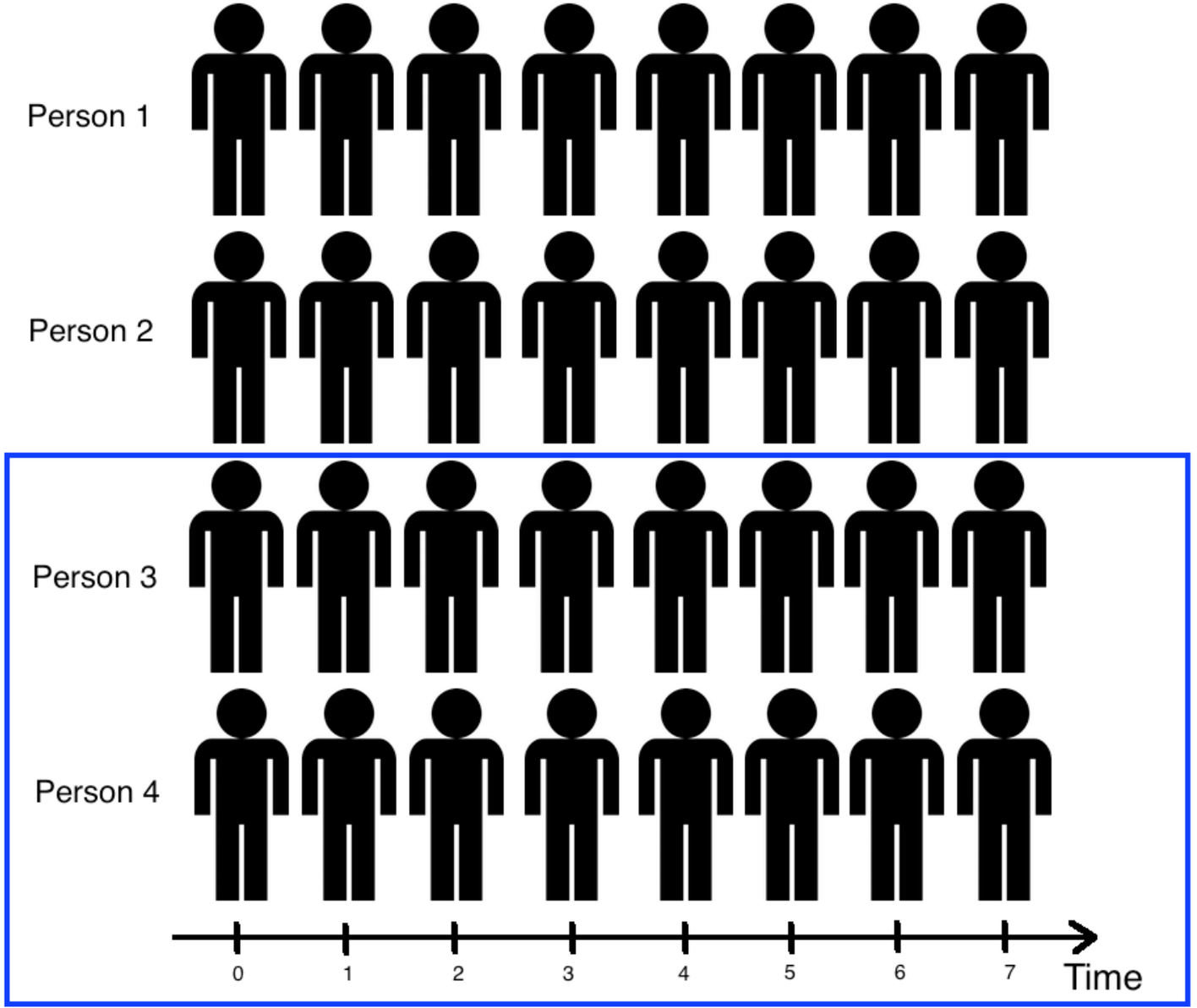
TIME SERIES/LONGITUDINAL DATA

- 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



GUIDED PRACTICE

WRITE A RESEARCH QUESTION WITH RAW DATA

ACTIVITY: WRITE A RESEARCH QUESTION WITH RAW DATA

EXERCISE

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

ACTIVITY: WRITE A RESEARCH QUESTION WITH RAW DATA

EXERCISE

Q: What type of data is this cross-sectional or longitudinal?
A: cross-sectional

Q: What will we be measuring
A: The association between being a woman or a child and survival on the Titanic.

Q: Write out a SMART aim for this data:
A: Using data from April 15, 1912, taken from the Titanic disaster, we will determine the association of gender, age (in years) and survival.

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, MINE

DATA SCIENCE WORKFLOW: ACQUIRE, PARSE, MINE

- For the remainder of class, we'll talk about steps 2 - 4 of the data science workflow: acquire, parse, and mine

ACQUIRE: GETTING AND IMPORTING DATA

- 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 dollars or euros?
- Data dictionaries help identify any requirements, assumptions, and constraints of the data.
- They make it easier to share data.

DATA DICTIONARY EXAMPLES

```
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.
```

DATA DICTIONARY EXAMPLES

```
# Metadata for Government Procurement
---
Identifier: '085dd6c3-387b-4661-ab26-02e422ad1286'
Name: 'government-procurement'
Title: 'Government Procurement'
Description: 'This dataset lists all open tenders put out by government
agencies since
  2015.'
Topics:
  - 'Finance'
Keywords:
  - 'GeBIZ'
  - 'procurement'
Publisher:
  Name: 'Ministry of Finance'
  Admin 1:
    Name: 'Charles Tan'
    Department: 'Performance & Resource Management'
    Email: 'Charles_TAN@mof.gov.sg'
Sources:
  - 'Ministry of Finance'
License: 'https://data.gov.sg/open-data-licence'
Frequency: 'Monthly'
Coverage: '2015-01-02 to 2017-01-31'
Last Updated: '2017-02-16T09:09:51.650638'
Resources:
  -
    Identifier: 'b9d8d509-5cb6-45dc-bb46-9508e670e3c2'
    Title: 'Government Procurement via GeBIZ'
    Url: 'https://storage.data.gov.sg/government-procurement/resources/
    government-procurement-via-gebiz-2016-11-22T11-45-37Z.csv'
    Format: 'CSV'
    Coverage: '2015-01-02 to 2017-01-31'
    Last Updated: '2016-11-22T11:45:37.105036'
```

```
    Url: 'https://storage.data.gov.sg/government-procurement/resources/
    government-procurement-via-gebiz-2016-11-22T11-45-37Z.csv'
    Format: 'CSV'
    Coverage: '2015-01-02 to 2017-01-31'
    Last Updated: '2016-11-22T11:45:37.105036'
    Schema:
      -
        Name: 'tender_no.'
        Title: 'Tender No.'
        Type: 'text'
        Sub Type: 'general'
      -
        Name: 'agency'
        Title: 'Agency'
        Type: 'text'
        Sub Type: 'general'
      -
        Name: 'tender_description'
        Title: 'Tender Description'
        Type: 'text'
        Sub Type: 'general'
      -
        Name: 'award_date'
        Title: 'Award Date'
        Type: 'datetime'
        Sub Type: 'date'
        Format: 'YYYY-MM-DD'
      -
        Name: 'tender_detail_status'
        Title: 'Tender Detail Status'
        Type: 'text'
        Sub Type: 'general'
```

MINE: CLEAN AND MANIPULATE THE DATA

- Format, clean, slice, and combine data
- Create necessary derived columns from the data

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.

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.

GEBIZ LAB

- In this lab, we will explore procurement data from Gebiz using Pandas and answer basic questions about the data

CONCLUSION

TOPIC REVIEW

REVIEW

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

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project 1

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET