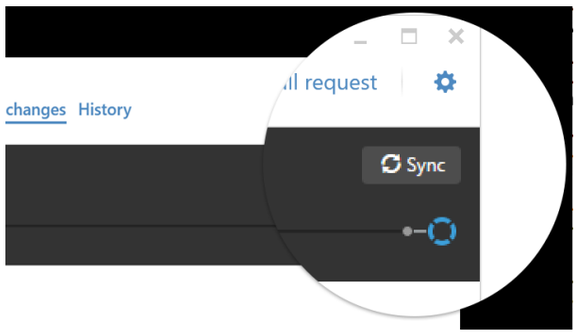
BEFORE WE GET STARTED

Have a Github account:
https://github.com/join

Download the Github desktop client
https://desktop.github.com

Clone the today's materials from Github (green button "Clone or download")

https://github.com/Dt1431/chi-ds-5-lesson-2



GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

David Turner
CEO, Waitbot Inc.

LEARNING OBJECTIVES

- Define a problem and types of data
- Different ways to acquire data
- Parse, refine, and mine in the pandas context
- Use Jupyter Notebook to import, format, and clean using the Pandas library

COURSE

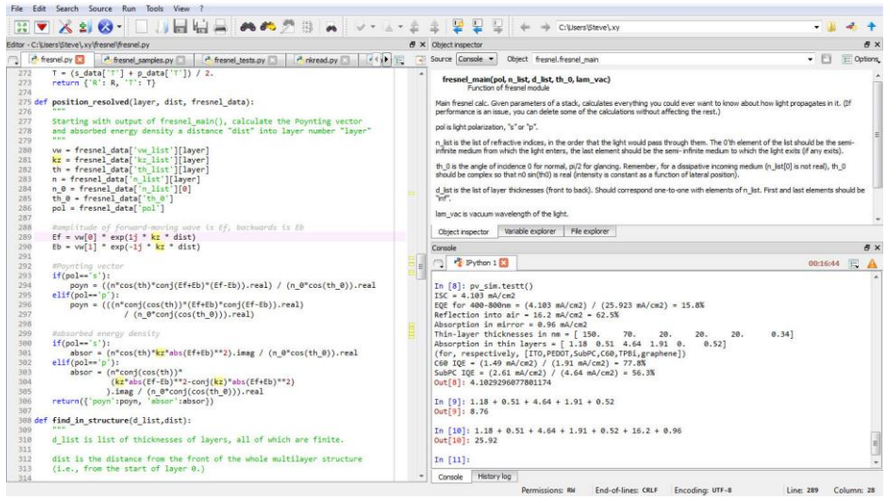
PRE-WORK

PRE-WORK REVIEW

- Create and open a Jupyter Notebook
- Understand the Spyder Interface

Jupyter Notebook

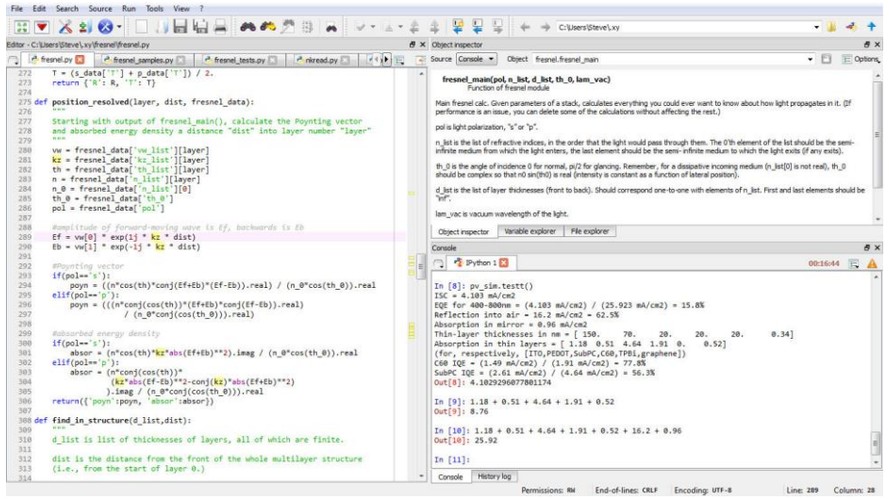
- An interactive document that can execute Python code within the same document
- Shareable with others. Convertible to different formats
- Intended for demonstration/education/ replication
- Not intended for efficient coding and analysis (Spyder is preferred)



Spyder

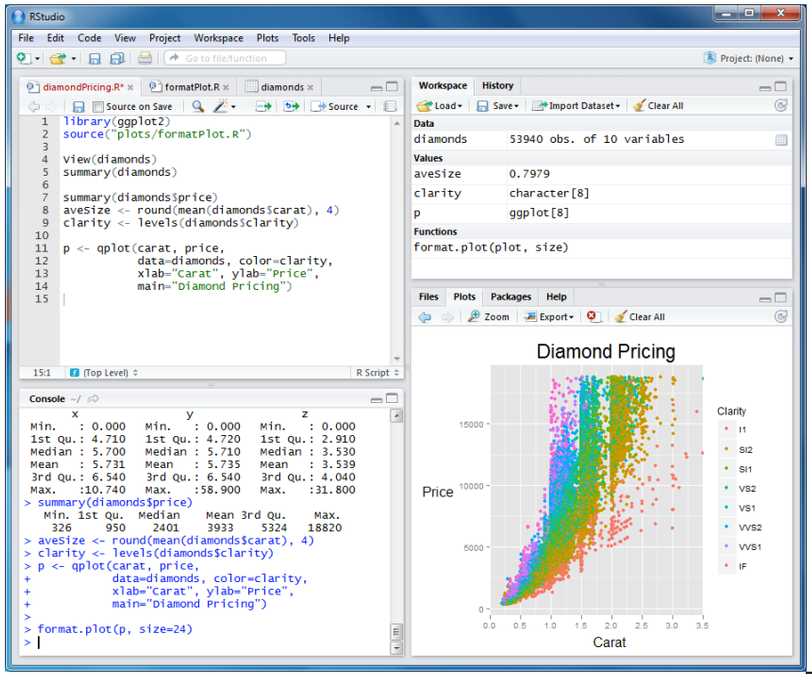
Spyder

- Interface for Python that allows you to
  - Write/Edit multiple files
  - Execute code interactively
  - Inspect and Visualize objects
- Tabular / Modular design
- Complete the workflow all within the same window
- Intended for quickly conducting data parsing and analysis
- Not intended for demonstration/education (Jupyter is preferred)

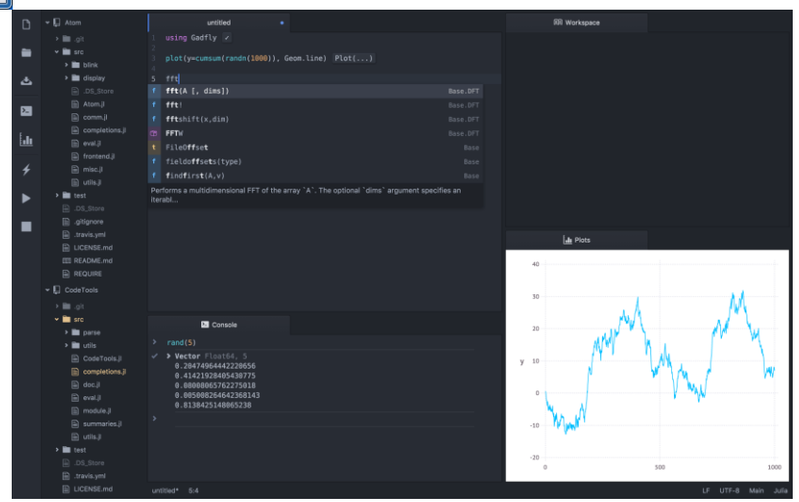


Spyder

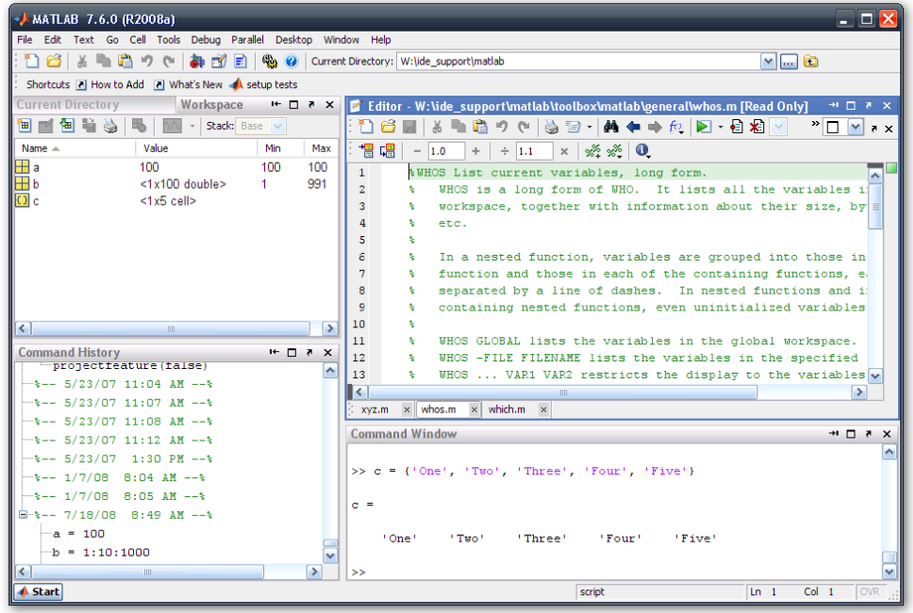
Spyder vs. Popular Programming Environments



Rstudio (R)



Juno (JULIA)



MATLAB

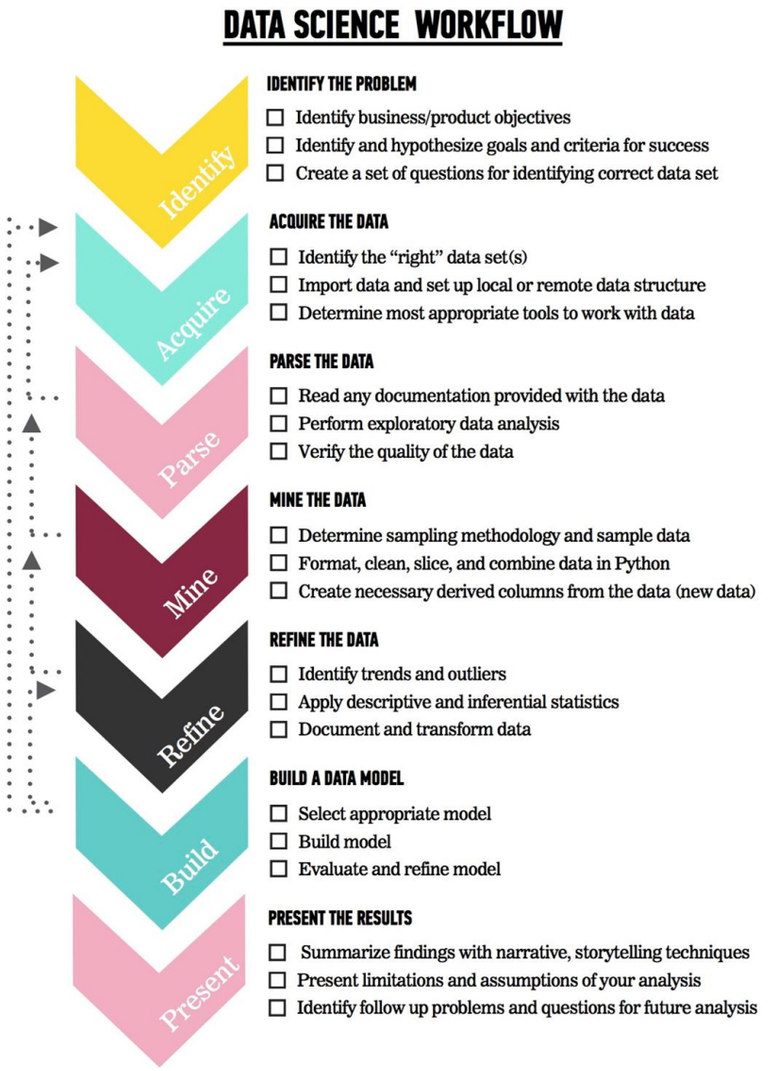
OPENING

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 1-4 (Identify the Problem and Acquire the Data and Parse the Data and Mine the Data).

- We’ll cover steps 5-6 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
  - Workflow is sequential and cumulative
  - Downstream consequences add up

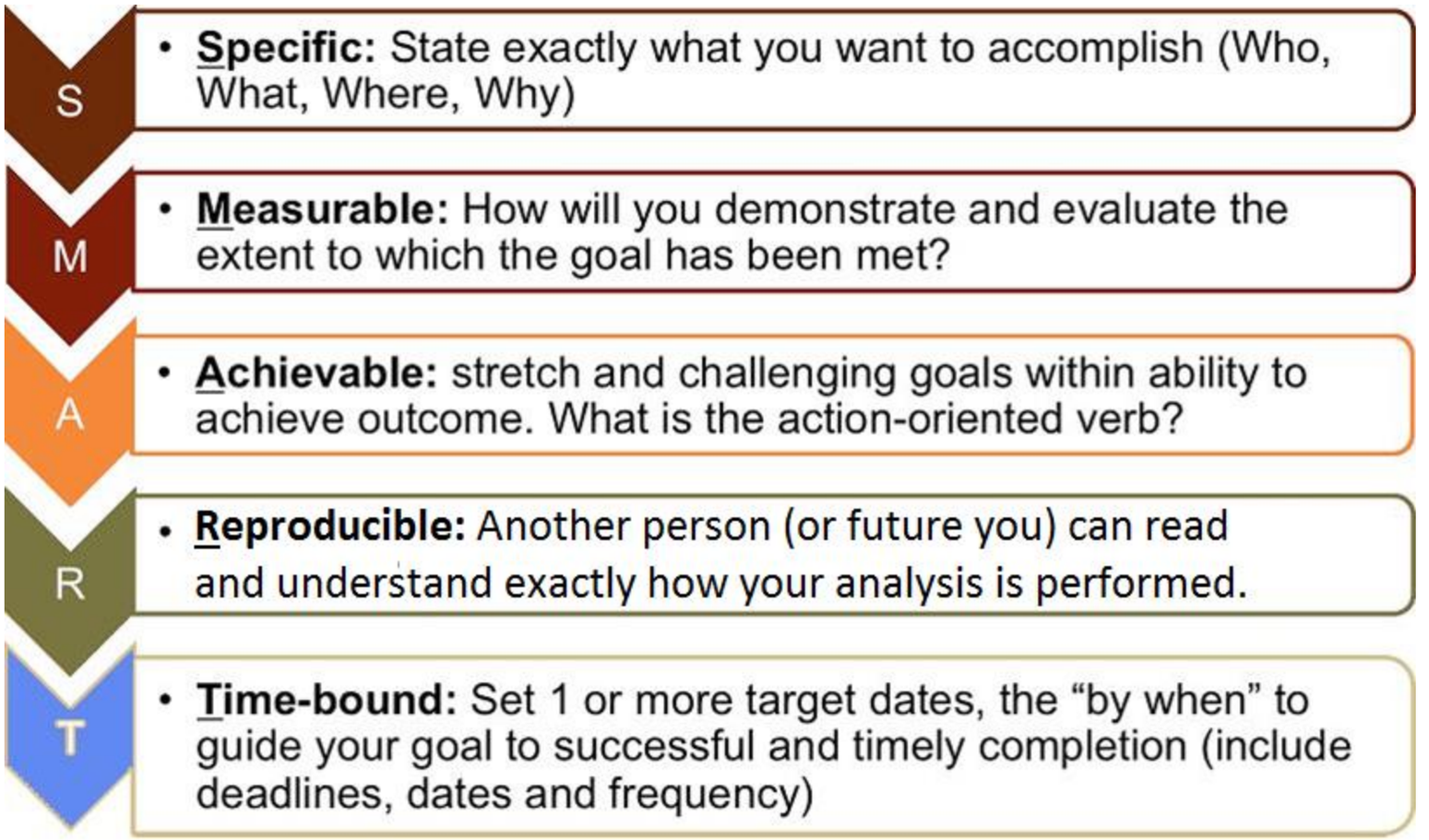
- Establishes the basis for reproducibility
  - More specific and clear (less room for misinterpretation)

- Enables collaboration through clear goals
  - Easier to build off of previous research
  - Easier to find others to work with you

WHAT IS A GOOD QUESTION?

- Goals are similar to the SMART Goals Framework.

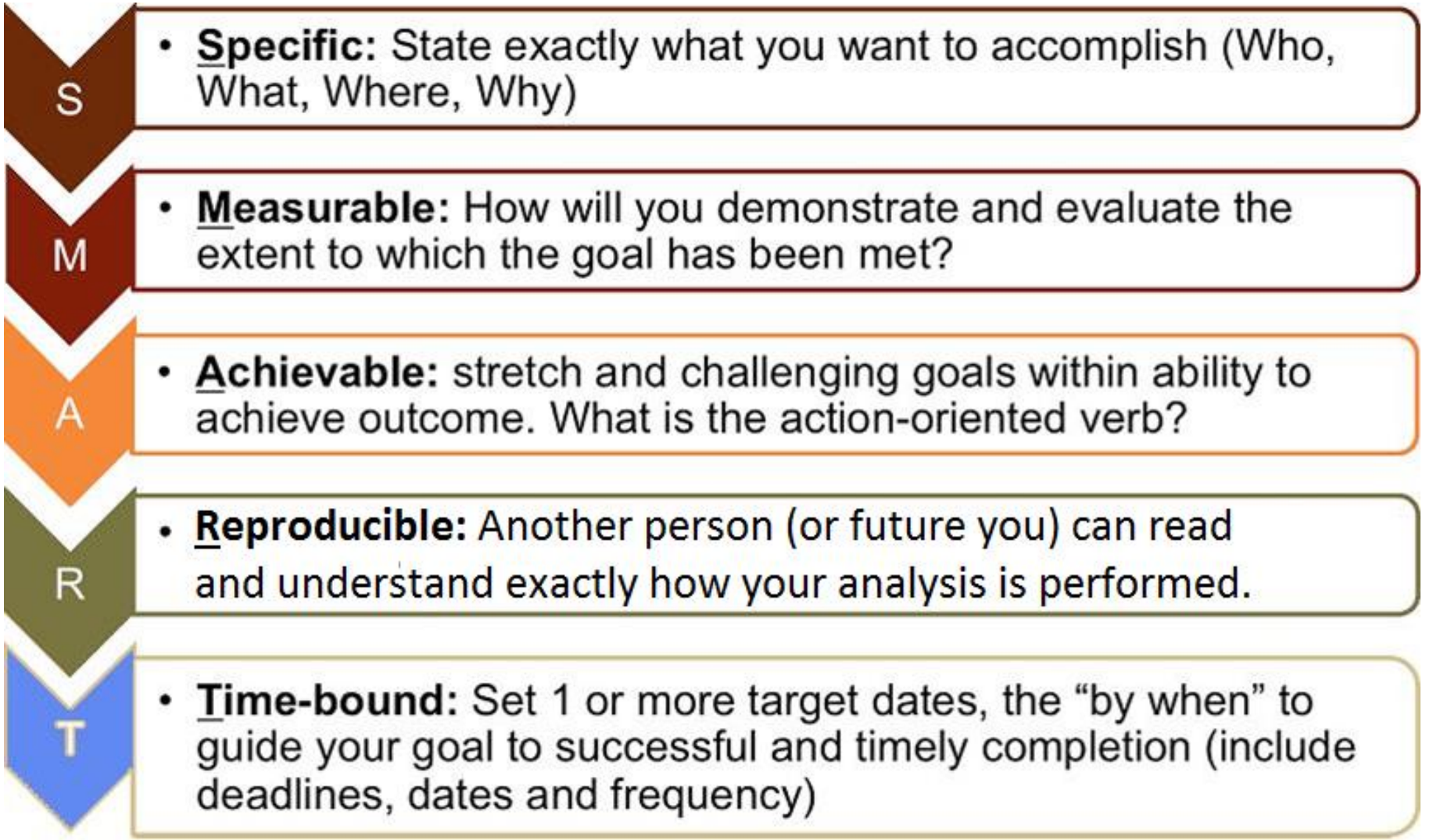
- S: specific
- M: measurable
- A: attainable
- R: reproducible
- T: time-bound



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 PROBLEM AND AIM

- What is 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 predictive of how much American children and adolescents eat. 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

- Different ways to measure:
  - Physical device (ruler, micrometer, lumiter, thermometer)
  - Biological / Physiological
  - Self-report (recollection, perception, attitudes)
  - Observer Coding (perception)
  - Behavior (click, eye gaze, movement)
  - Language (text, speech)

- 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.

- Need to measure
  - USDA recommended dietary intake for food
  - Child dietary intake
  - Food presence / amount

ATTAINABLE

- Is there access to this data?
  - Already collected?
  - If not, is collecting possible?

- Are there proxys (indirect) ways to get the data?

- Using one 24-hour recall from the cross-sectional NHANES 2007-2010: http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm

REPRODUCEABLE

- Can the analysis be replicated and validated?

- Need to be able to have the results checked and re-attempted both in your dataset and in the future
  - If not, then conclusion is based on faith
  - Need to have verification

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

TIME BOUND

- Is there a clear time-point and duration specified for the collection of the data

- 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

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (2 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

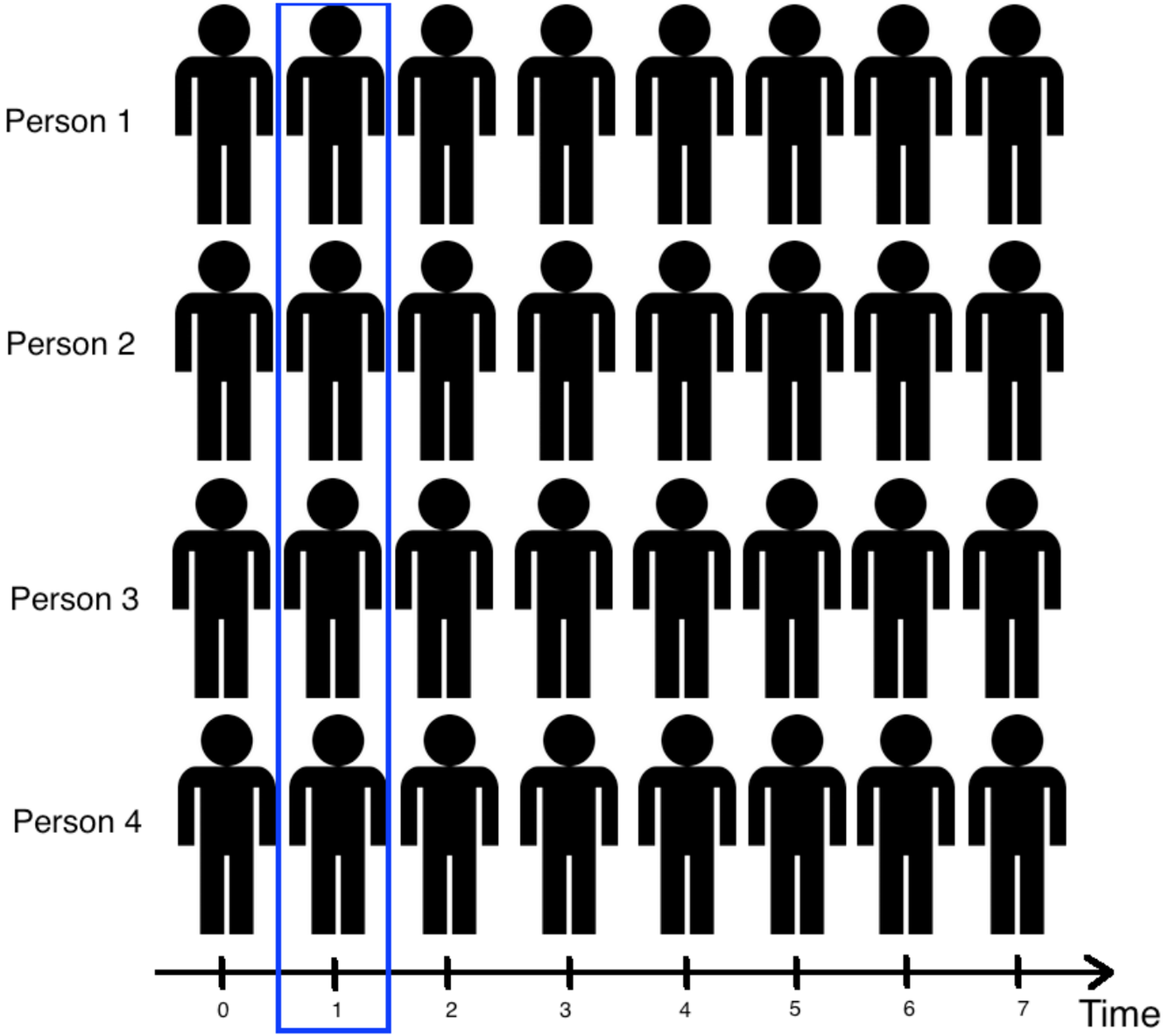
- All data/information is collected at the same time; all data comes from the same time period.
- Example: You survey 100 people about their music preferences. You don’t follow up with any of them after the first survey.
- Example: You take a sample of 100 different beers from a production plant one day. Each beer is measured only one

CROSS-SECTIONAL DATA

- Strengths
  - Often population based / Generalizability (broader samples usually)
  - Reduce cost compared to other types of data collection methods

- Weaknesses
  - Separation of cause and effect may be difficult (or impossible)
    - Example: Measure how much people exercise and their stress
      - You find a strong association. People who exercise more have lower stress
      - Which came first? Low Stress -> Exercising or Exercising -> Lower Stress or Neither
  - Variables/cases with long duration are over-represented

CROSS-SECTIONAL DATA



Note: Measurement across many different people happens at the same point in time.

Each person is only measured once.

CROSS-SECTIONAL DATA

- Easy way to remember!
  - If you have 1 measurement across many different people/items/cases, then its probably cross-sectional
  - If people/objects are measured at only one point time, it’s probably cross-sectional

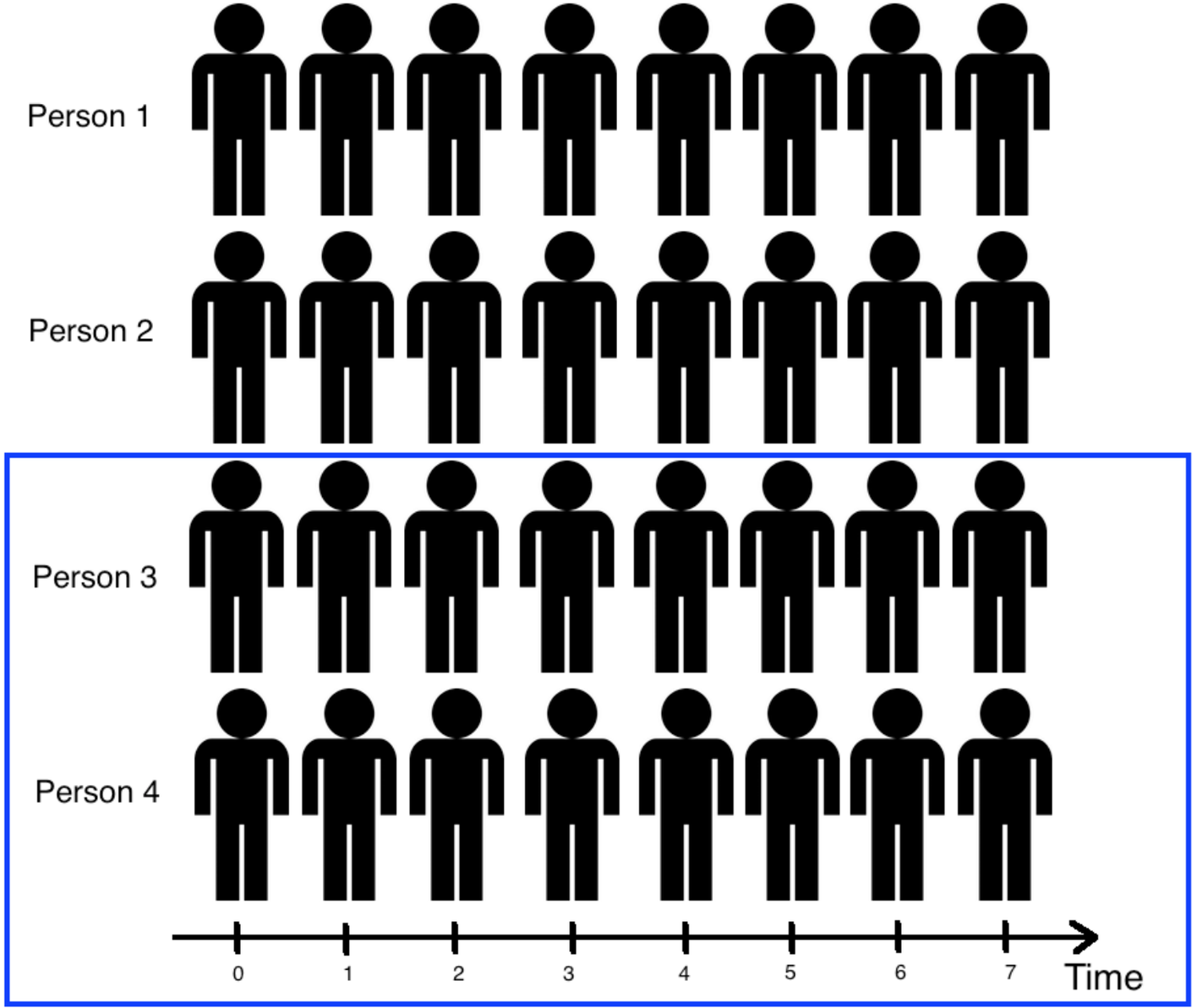
TIME SERIES/LONGITUDINAL DATA

- The information is collected over a period of time

- Strengths
  - Unambiguous temporal sequence – you know which variable came first
  - Multiple outcomes can be measured (you can measure the average outcome, as well as change over time/stability)

- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data (You need to get people/objects to stay/remain available)

TIME SERIES/LONGITUDINAL DATA



Note: Measurement happens at different points in time.

Each person is measured multiple times.

LONGITUDINAL DATA

- Easy way to remember!
  - If you have many measurements for the same people/items/cases, then its probably longitudinal
  - If the same person/object was measured more than once, its probably longitudinal

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Break up into pairs and come up with one research question that are cross-sectional

2. Come up with a research question that is longitudinal.

DELIVERABLE

Answers to the above questions

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 & 4 of the data science workflow: acquire, parse, and mine
- 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

LOGISTICS OF ACQUIRING YOUR DATA

- Data can be acquired through a variety of sources

- Locally Available Formatted/Structured Files (CSV, XML, TXT, XLS)
  - Open entire file with file reader and load into programming environment)
  - Provided by company, connection, archive/repository (Github), logs

- Remotely Hosted Data (APIs, Webpages/HTML, Images)
  - Send out request and parse into structured format

- Databases (SQL, NOSQL, Filemaker, Access, etc.)
  - Connect to database and query data

- Today, we'll use a CSV (comma separated file)

- Next class we cover acquiring data through the web

ACQUIRE

- Where we determine if we have the “right” dataset for our problem

- Questions to ask (things you want to optimize):
    - Does it match the variables of interest and format (cross-sectional vs. longitudinal)?
        - Are variables available elsewhere (merge together later)
        - Are there proxy variables (you want body fat %, settle for BMI)
    - How well was the data collected?
    - Is there much missing data? (more complete = better)
    - Was the data collection instrument validated (accurate) and reliable (consistent)?
    - Is the dataset at the correct level (individual or grouped/aggregated)

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:
  - Poor Quality: Baby, Young, Adult, Old
  - Higher Quality: Age in years
  - Even Higher Quality: Date of Birth

- 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. Necessary for assessing the quality of the data and know how to parse/mine the data.

- They make it easier to share data and reproduce results.

DATA DICTIONARY EXAMPLE: 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

PARSING AND MINING

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 sheets; they contain rows and columns.
  - People/Cases/Objects are the rows
    - 10 rows indicated there are 10 people in the sample(Usually)
  - Variables (Predictors and Outcomes) are the columns
    - 5 columns indicates there are 5 variables (For cross-sectional data)

NUMPY AND PANDAS INTRO

Numpy is a library for easily working with numerical data

Pandas is library for representing data in a structure conducive for analysis

	Height	Weight	Age	Sex
Person 1	65	130	21	M
Person 2	71	156	42	M
Person 3	62	150	66	F
Person 4	68	163	43	F
Person 5	69	172	59	M

Dataframes are typically referenced in the following convention:

Dataframe[row,column]

For example Dataframe[“Person 1”, “Height”] = 65

You can also do it by the index (starting at 0):
Dataframe[0,0] = 65
Dataframe[1,0] = 75

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.

NUMPY AND PANDAS INTRO

- Once we get out data, it's not yet ready to analyze.

1) Load dataset into environment
2) Clean the data (correct errors and issues)
3) Create new variables (average together, take absolute differences)
4) Scale /Transform data (change the range and distribution of values)
5) Merge data together (if bringing together two datasets)

- Once again, follow and code along.

DEMO

LAB WALKTHROUGH

LESSON 2 LAB WALKTHROUGH

- In this lab, you will go through the data refinement process using the Titanic dataset.
- 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

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET LINK:
http://bit.ly/2nxjvB2

Lesson: 2
Topic: Experimental Design & Pandas