GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	Research Design and Pandas	Lesson 2
	Statistics Fundamentals I	Lesson 3
	Statistics Fundamentals II	Lesson 4
	Flexible Class Session	Lesson 5
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

Today's Class

LAST CLASS

WHAT DID WE LEARN?

- ✓ Meet & Greet
- ✓ What’s data science?
- ✓ The data science workflow
- ✓ Environment setup: Anaconda, Jupyter, and Spyder
- ✓ Case study: NYC traffic analysis

Any questions on LAB 1 – Home Practice?

LAST CLASS

ANNOUNCEMENTS

- We need to talk. Reserve your 1:1 on doodle

  https://doodle.com/poll/nymgqzrwq263vqiz

- Fill your exit ticket!

thinking about final projects and where to get data

Will we be working on our projects in class or are they entirely "homework"?

How to do the ipython notebook test?

How much of neural networks are we going to explore?

I can't see lesson 1 on website

Are there any limitations to using python for data science compared to other programming languages?

Others?

LEARNING OBJECTIVES

- Manage your development environment and files
- Define and Identify a problem and types of data
- Apply the data science workflow in the pandas context
- Create an Notebook to import, format, and clean using the Pandas

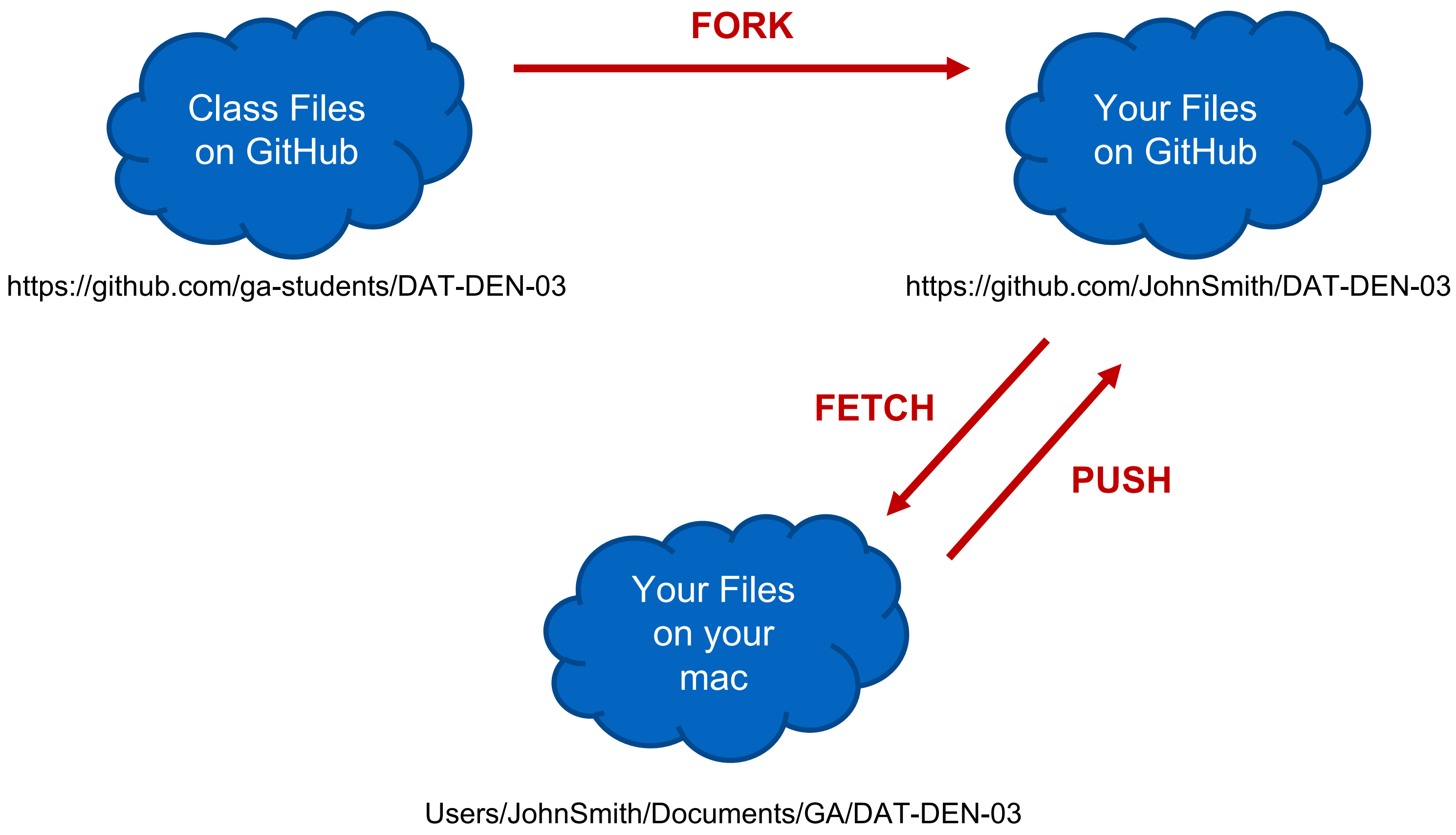
PRE-WORK

GITHUB FILES MANAGEMENT

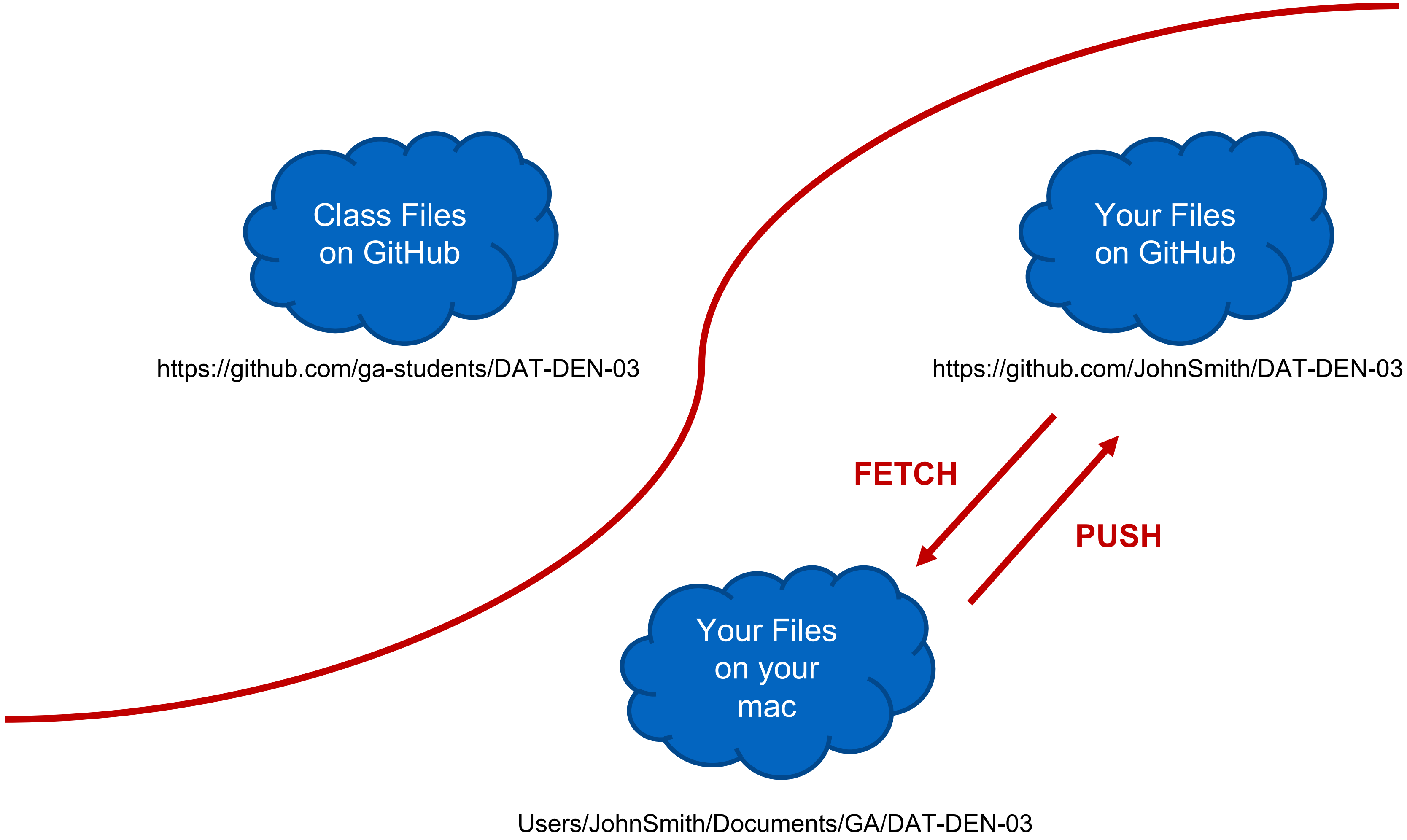
DID YOU INSTALL AND COMPLETE THE FOLLOWING?

- Joined Slack and the class repository
- Anaconda
- Python 2.7
- Atom (Optional)
- GitHub Account
- GitHub Desktop

HOW ARE WE GOING TO MANAGE OUR FILES?


Class Files
on GitHub
FORK
Your Files
on GitHub
https://github.com/ga-students/DAT-DEN-03
https://github.com/JohnSmith/DAT-DEN-03
FETCH
PUSH
Your Files
on your
mac
Users/JohnSmith/Documents/GA/DAT-DEN-03

WHAT HAPPENS AFTER THE CLASS FILES ARE UPDATED?



Class Files on GitHub

https://github.com/ga-students/DAT-DEN-03

Your Files on GitHub

https://github.com/JohnSmith/DAT-DEN-03

FETCH

PUSH

Your Files on your mac

Users/JohnSmith/Documents/GA/DAT-DEN-03

HOW TO KEEP YOUR GITHUB UPDATED?

Synch to the class GitHub few hours after class using your Terminal.

```
git clone git@github.com/JohnSmith/DAT-DEN-03.git

cd /Users/665066/Documents/GitHub/DAT-DEN-03
git remote add upstream git://github.com/ga-students/DAT-DEN-03.git
git fetch upstream
git commit -m ”.” (if there is any change)
git pull upstream master
git push (to keep your online Github account synch with your local files)
```

Create and modify notebooks and python files…

THIS CLASS: EXPERIMENTAL DESIGN AND PANDAS

LEARNING OBJECTIVES

- ~~Manage your development environment and files~~
- Define and Identify a problem and types of data
- Apply the data science workflow in the pandas context
- Create an Notebook to import, format, and clean using the Pandas

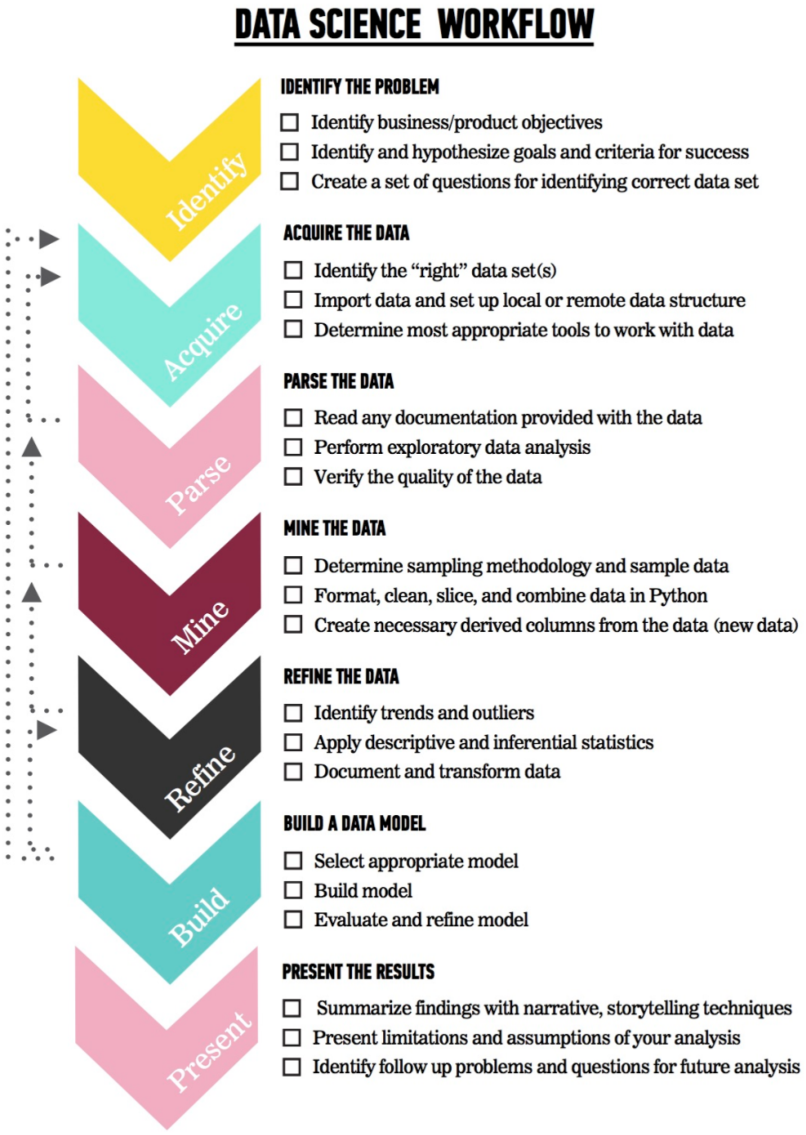
ANY QUESTIONS?

LET'S REVIEW THE DATA SCIENCE WORKFLOW

The steps:

TODAY

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



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- □ Identify business/product objectives
- □ Identify and hypothesize goals and criteria for success
- □ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- □ Identify the "right" data set(s)
- □ Import data and set up local or remote data structure
- □ Determine most appropriate tools to work with data

PARSE THE DATA

- □ Read any documentation provided with the data
- □ Perform exploratory data analysis
- □ Verify the quality of the data

MINE THE DATA

- □ Determine sampling methodology and sample data
- □ Format, clean, slice, and combine data in Python
- □ Create necessary derived columns from the data (new data)

REFINE THE DATA

- □ Identify trends and outliers
- □ Apply descriptive and inferential statistics
- □ Document and transform data

BUILD A DATA MODEL

- □ Select appropriate model
- □ Build model
- □ Evaluate and refine model

PRESENT THE RESULTS

- □ Summarize findings with narrative, storytelling techniques
- □ Present limitations and assumptions of your analysis
- □ Identify follow up problems and questions for future analysis

INTRODUCTION

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? SMART

- **S**pecific: The dataset and key variables are clearly defined.
- **M**easurable: The type of analysis and major assumptions are articulated.
- **A**ttainable: The question you are asking is feasible for your dataset and is not likely to be biased.
- **R**eproducible: Another person (or future you) can read and understand exactly how your analysis is performed.
- **T**ime-bound: You clearly state the time period and population for which this analysis will pertain.

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 (10 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

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

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

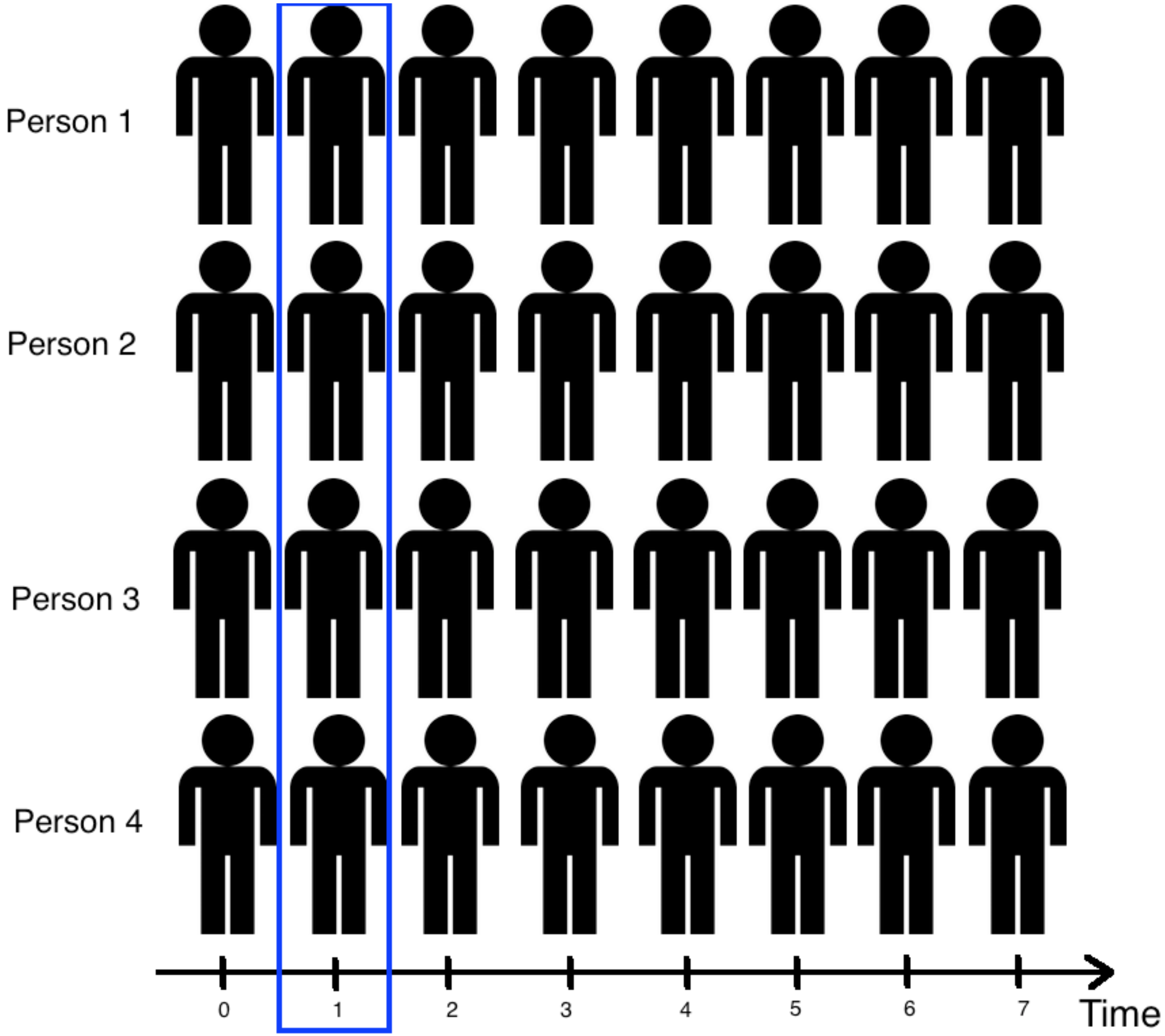
b. Determine if the number of passengers on JetBlue, Delta and United domestic flights with carry-on luggage is associated with delayed take-off time using data from flightstats.com from January 2015- December 2015.

- **S**pecific: The dataset and key variables are clearly defined.
- **M**easurable: The type of analysis and major assumptions are articulated.
- **A**ttainable: The question you are asking is feasible for your dataset and is not likely to be biased.
- **R**eproducible: Another person (or future you) can read and understand exactly how your analysis is performed.
- **T**ime-bound: You clearly state the time period and population for which this analysis will pertain.

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



CROSS-SECTIONAL DATA

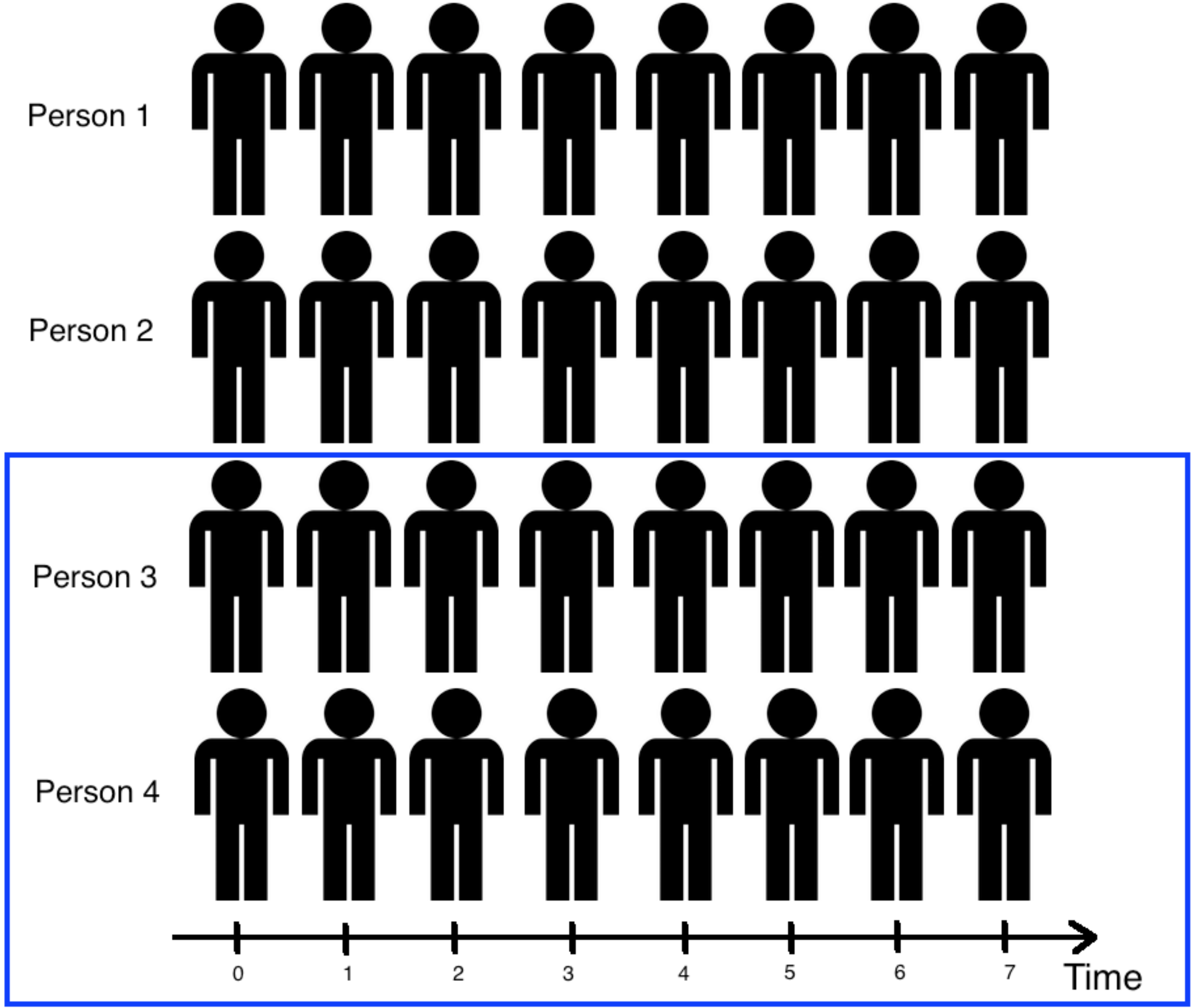
- All information is determined at the same time; all data comes from the same time period.

- 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

TIME SERIES/LONGITUDINAL 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

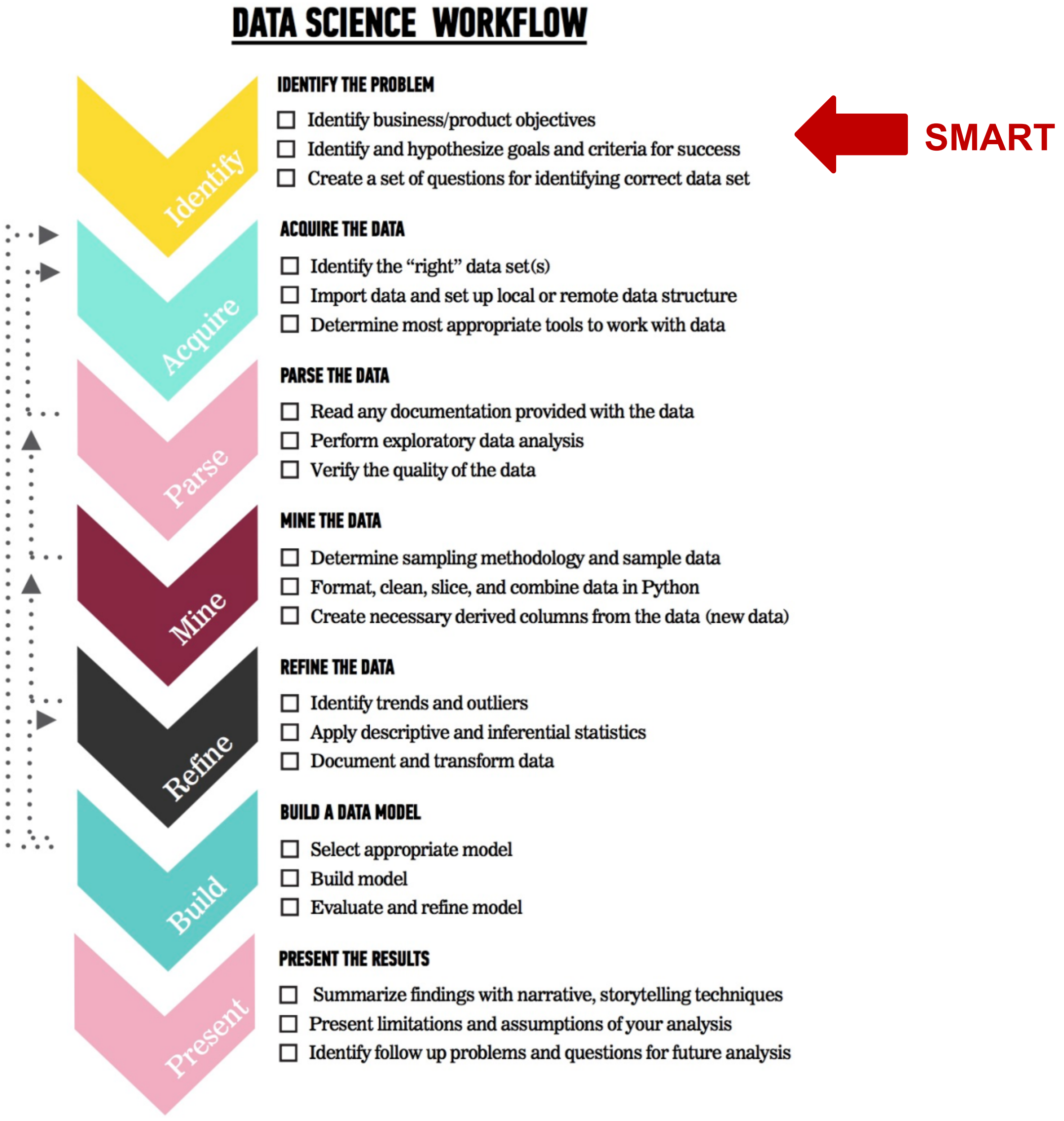
- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data

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

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the “right” data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

THIS CLASS: EXPERIMENTAL DESIGN AND PANDAS

LEARNING OBJECTIVES

- ~~Manage your development environment and files~~
- ~~Define and Identify a problem and types of data~~
- Apply the data science workflow in the pandas context
- Create an Notebook to import, format, and clean using the Pandas

ANY QUESTIONS?

5 min Break

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)

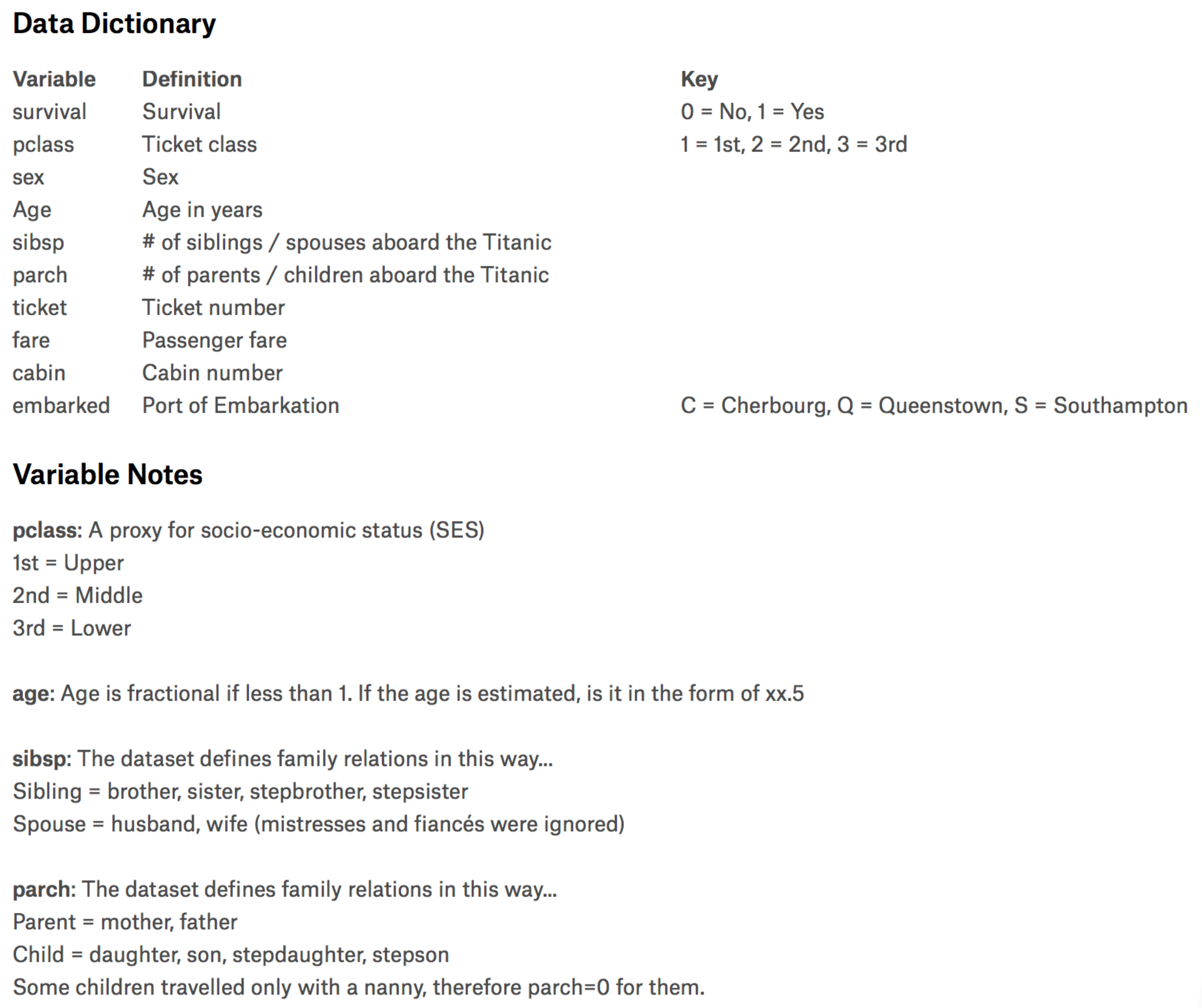
PARSE: UNDERSTANDING YOUR DATA

- 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

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 EXAMPLE:



Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

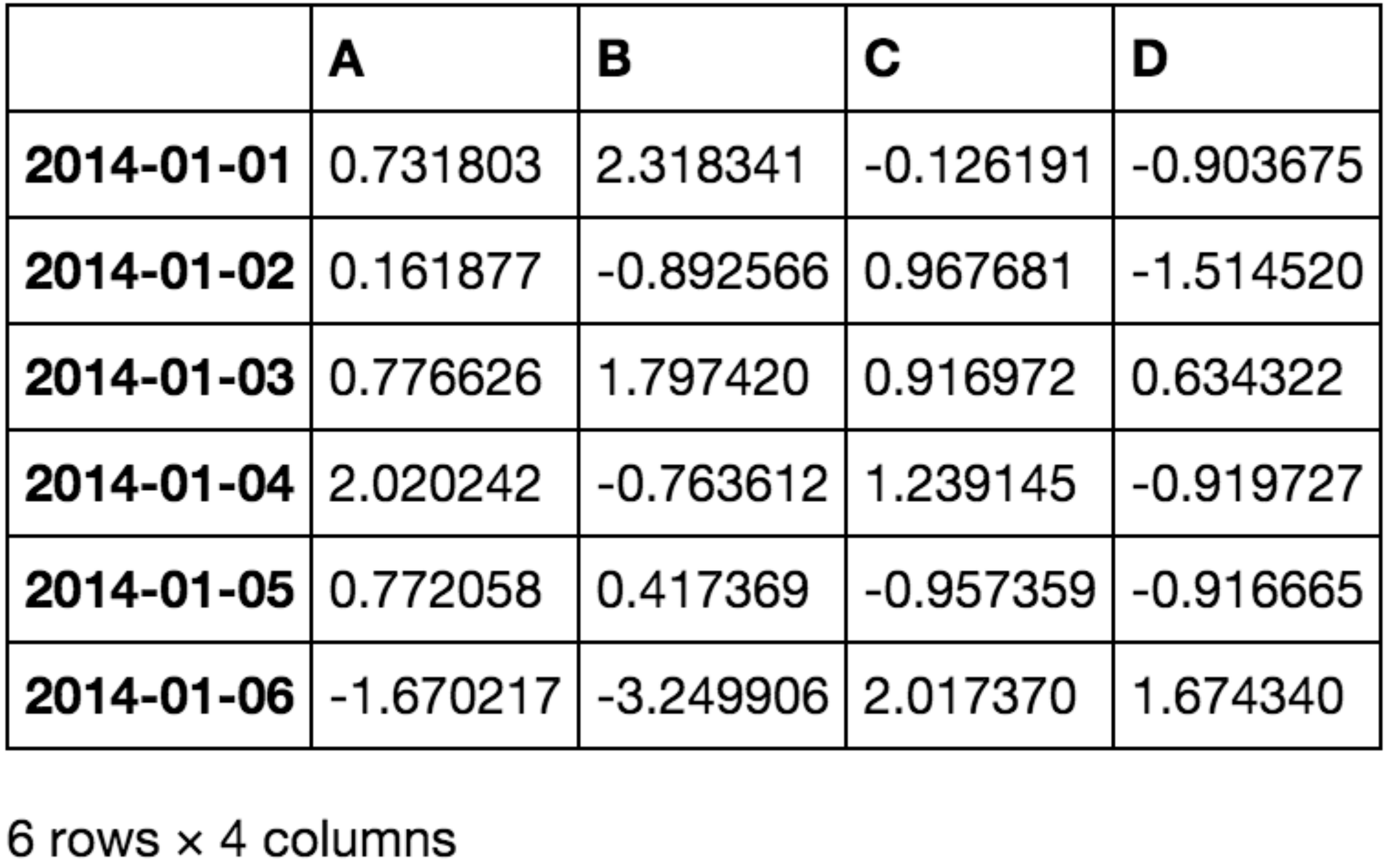
CODEALONG

NUMPY AND PANDAS INTRO

NUMPY AND PANDAS INTRO

- What are Numpy and Pandas? Python packages
- Pands 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



	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.

NUMPY AND PANDAS INTRO

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

- Once again, follow and code along.

5 min Break

DEMO

LAB WALKTHROUGH

LESSON 2 IN-CLASS LAB WALKTHROUGH

- 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

THIS CLASS: EXPERIMENTAL DESIGN AND PANDAS

LEARNING OBJECTIVES

- ~~Manage your development environment and files~~
- ~~Define and Identify a problem and types of data~~
- ~~Apply the data science workflow in the pandas context~~
- ~~Create an Notebook to import, format, and clean using the Pandas~~

ANY QUESTIONS?

CONCLUSION

TOPIC REVIEW

REVIEW

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

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project: Unit 1
- Lab 2 – Home Practice

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	~~Research Design and Pandas~~	~~Lesson 2~~
	Statistics Fundamentals I	Lesson 3
	Statistics Fundamentals II	Lesson 4
	Flexible Class Session	Lesson 5
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

Today's Class

LESSON

Q & A

Let’s talk about Class Schedule

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET