EXPERIMENTAL DESIGN AND PANDAS

LEARNING OBJECTIVES FOR TODAY

- Identify the business problem
- Types of Study: Cross sectional vs Longitudinal
- Titanic dataset on Kaggle
- Numpy and Pandas

PRE WORK & REVIEW

PRE WORK

Before this lesson, you should already be able to:

- Create, open and create a Jupyter Notebook

GITHUB?

Do you all have access to the github repo: https://github.com/alexperrier/gads

DATA SCIENCE WORKFLOW

1. **Identify** the Business Problem
2. **Acquire** Raw Data
3. **Parse and Mine** the Data: **data munging**
4. **Transform** the data: Feature engineering
5. **Select** and tune the Model: **Model Selection** and **Feature Selection**
6. **Present/ implement the results**: Visualization, deploy to production

LAST SESSION

ANY QUESTIONS FROM LAST CLASS?

EXPERIMENTAL DESIGN AND PANDAS

TODAY

Focus on

1. **Identify** the Business Problem
2. **Acquire** Raw Data
3. **Parse** the Data

With

- Numpy
- Pandas

TODAY

- [0 - 45mn] Experimental Design: Good questions - S.M.A.R.T. - Study types - Study Example
- [45mn - 1h] Titanic dataset
- [1h - 1:15] Numpy and Pandas Intro
- [1h15 - 1h45] Numpy and Pandas Code along
- [1h45 - end] LAB

ASKING GOOD QUESTIONS

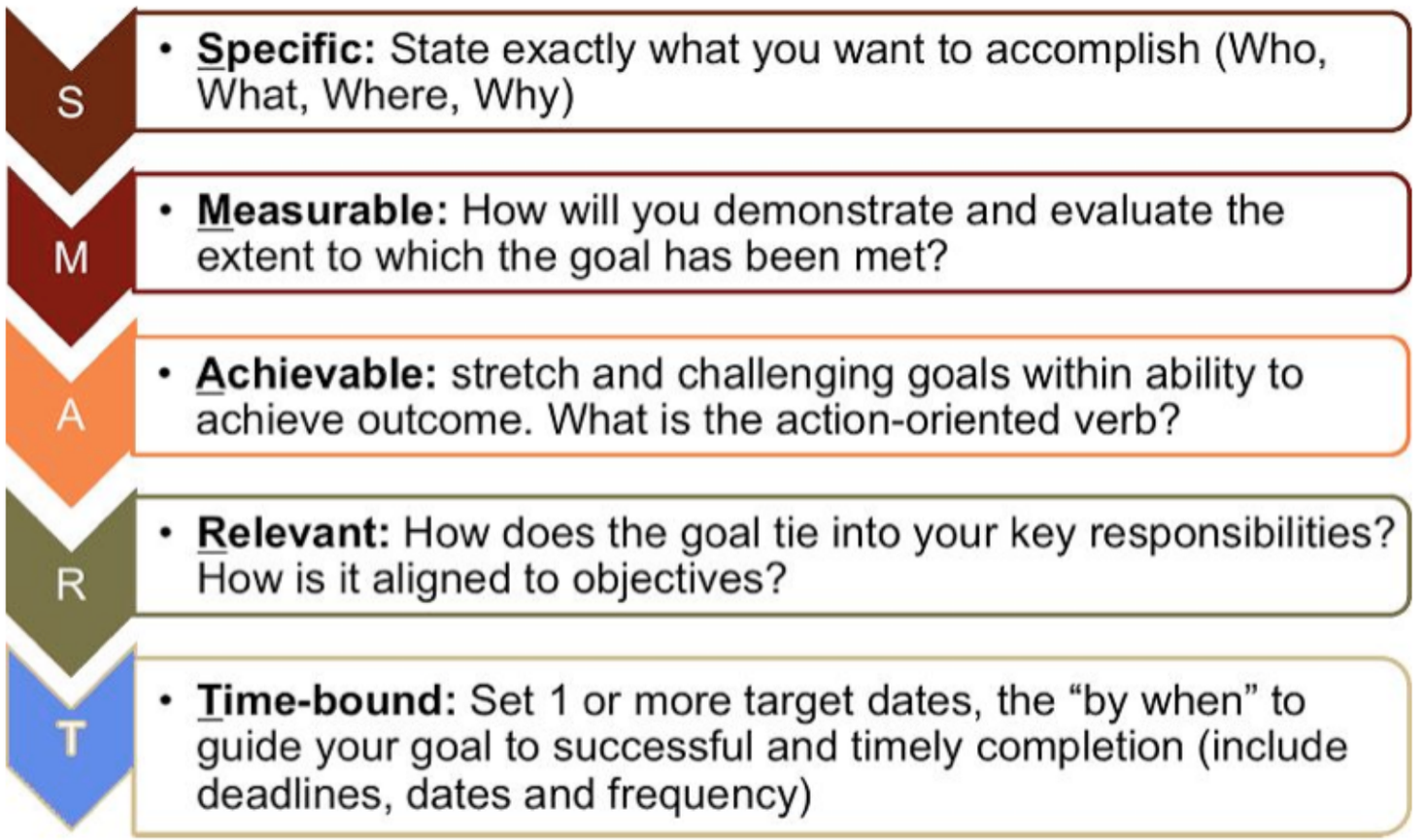
WHY DO WE NEED A GOOD QUESTION?

- “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?

- Goals are similar to the SMART Goals Framework.

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



S.M.A.R.T.

- 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.

EXAMPLE

EXAMPLE

Determine the association of foods in the home with child dietary intake.

Using one 24-hour recall from the cross-sectional NHANES 2009-2010, 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

S.M.A.R.T

- Measurable: Food, Daily Intake
- Attainable: Determining association, correlation not causation
- Reproducible: Anyone with the data and scripts / tools can reach the same conclusions
- Time bound: Using one 24-hour recall from NHANES 2007-2010

Context is key: Research, Sociology, Business, ...

HYPOTHESIS

EXERCISE

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.

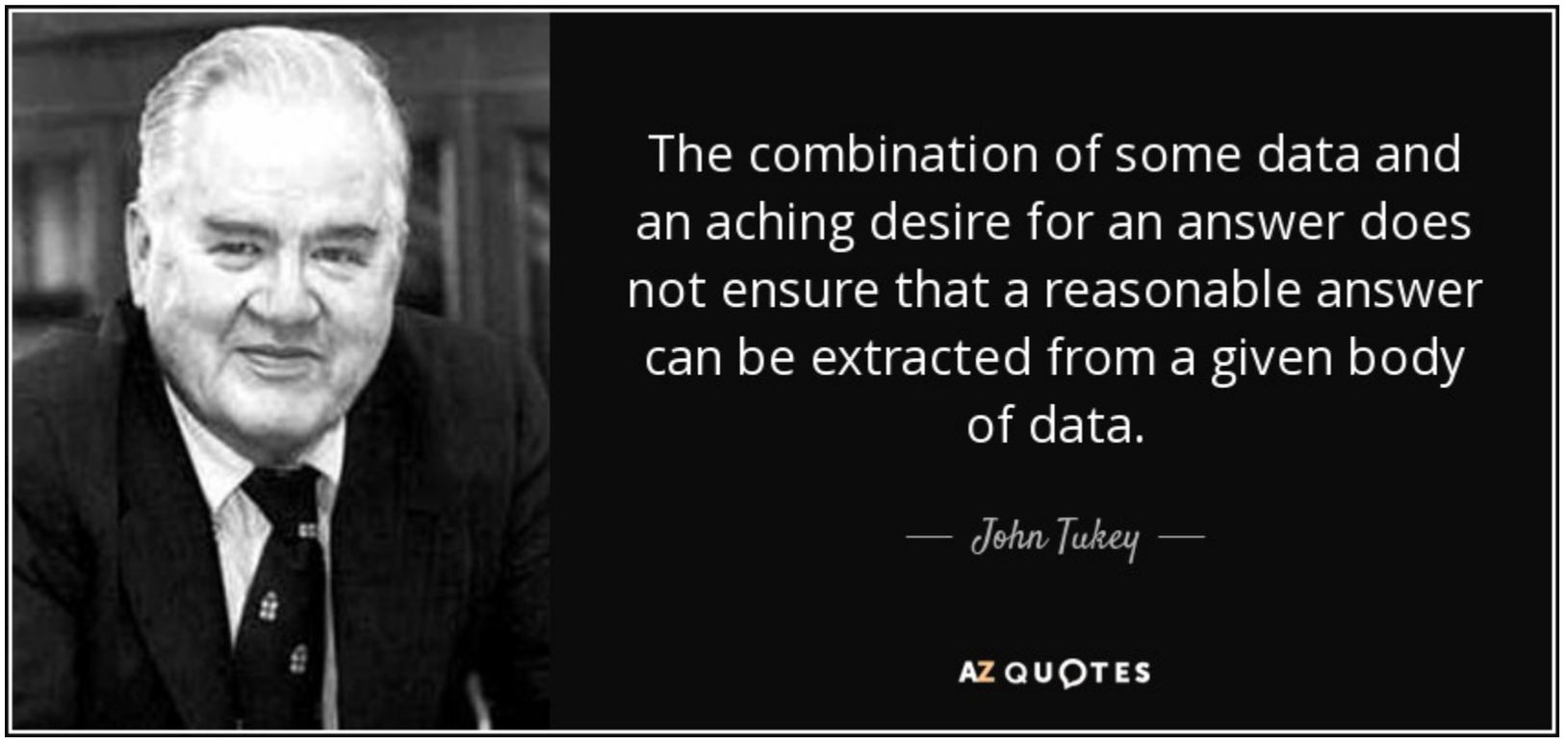
    b. Determine if the number of passengers on JetBlue, Delta and United domestic flights with carry-on luggage is associated with delayed take-off time using data from flightstats.com from January 2015- December 2015.

DELIVERABLE

Answers to the above questions

CROSS SECTIONAL VS LONGITUDINAL

JOHN TUKEY


The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
— John Tukey —
AZ QUOTES

STUDY TYPES

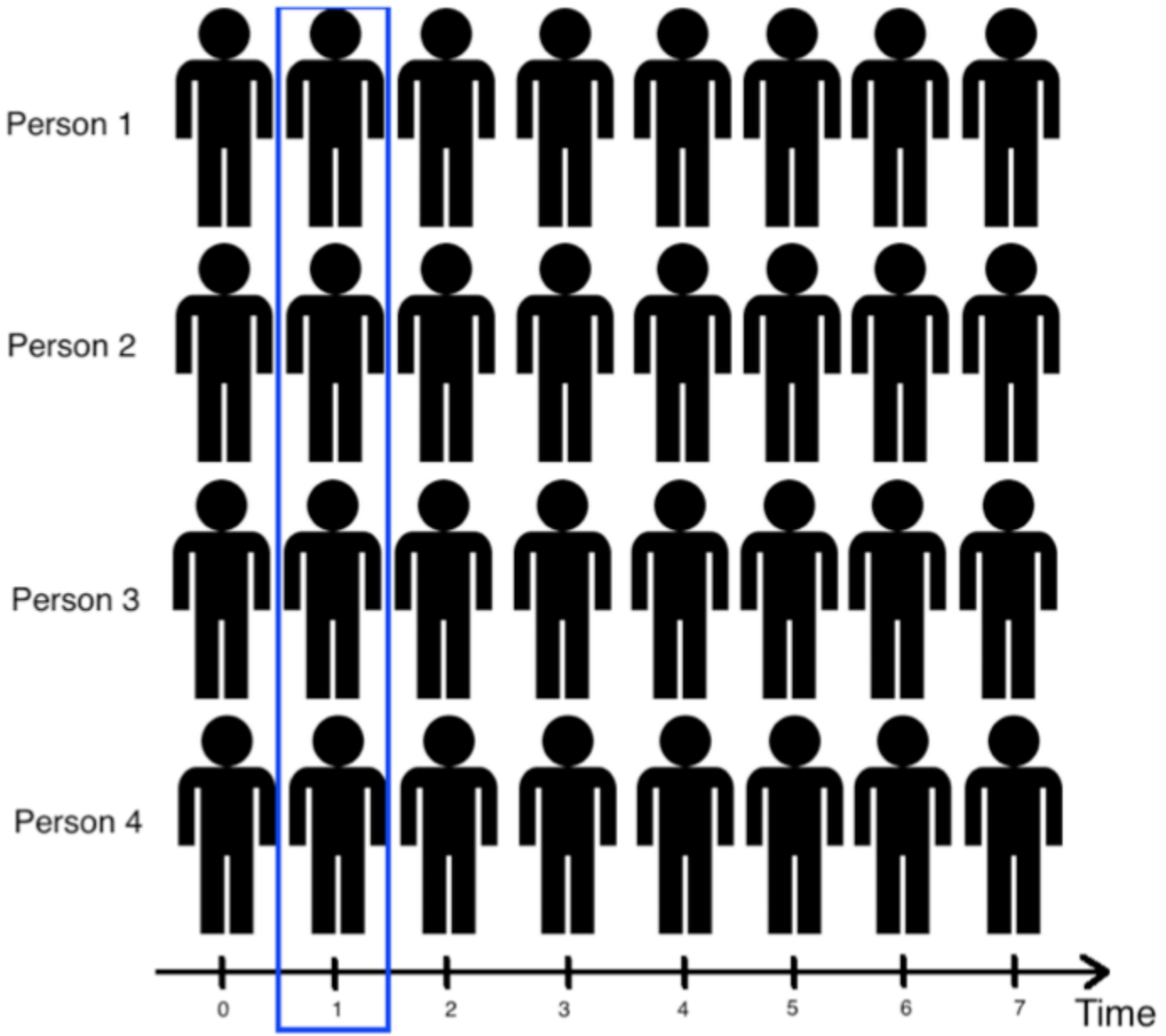
CROSS-SECTIONAL DATA

All information is determined at the same time: A snapshot.

LONGITUDINAL DATA

Time Series: The information is collected over a period of time: A recording. cross-sectional vs. longitudinal studies Both are observational studies.

CROSS-SECTIONAL DATA



CROSS-SECTIONAL DATA

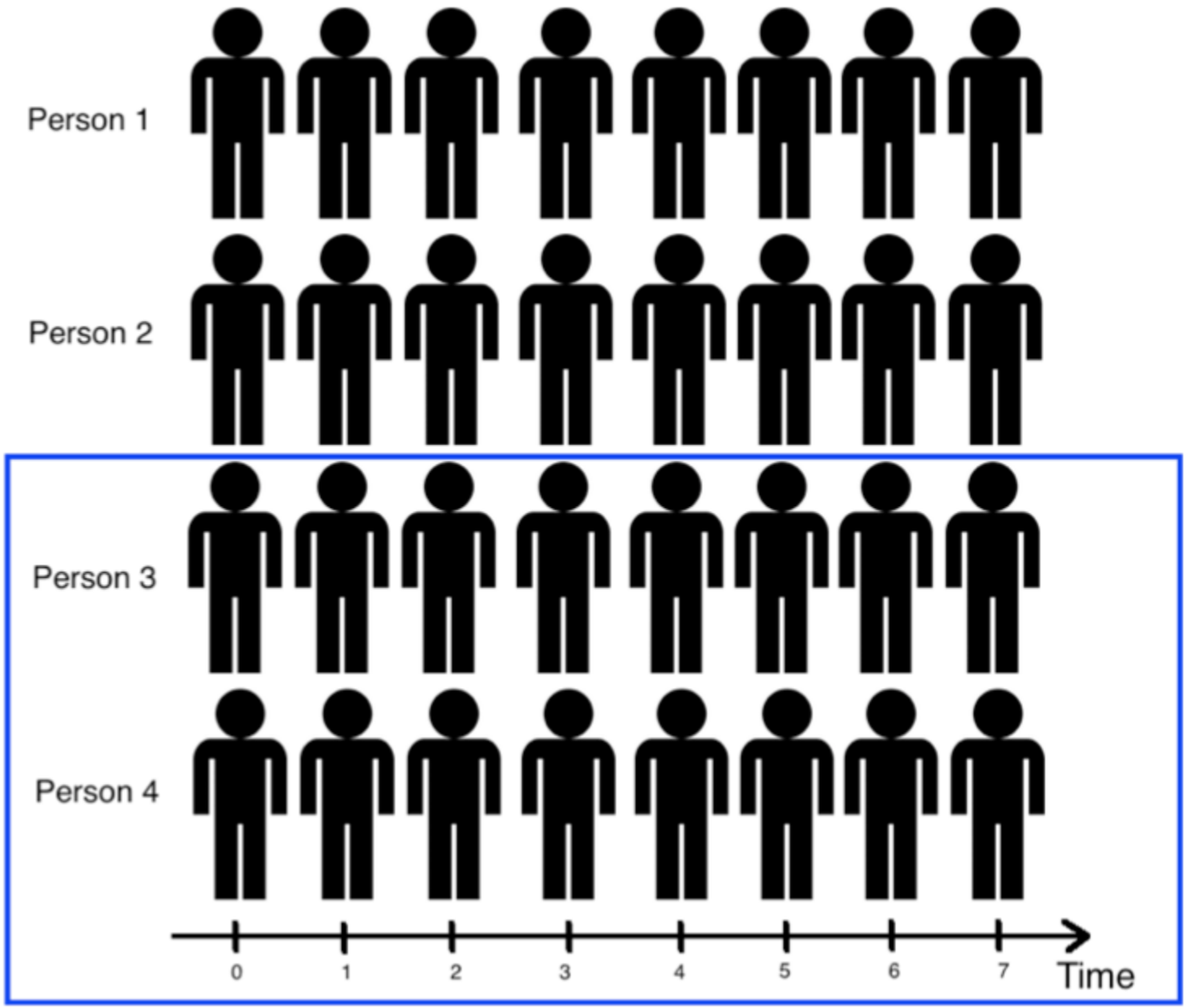
Strengths

- Population based
- Generalizable
- Reduced cost for data collection

Weaknesses

- Separation of cause and effect would be impossible (because causes precede their effects in time)
  - Some causal Inference methods would challenge that
- Variables/cases with long duration are over-represented

LONGITUDINAL DATA / TIME SERIES



LONGITUDINAL DATA / TIME SERIES

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

ACTIVITY 2: KNOWLEDGE CHECK

TWITTER STREAM?

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. What type of data is the flightstats 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?

DELIVERABLE

Answers to the above questions

ACTIVITY 3: 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

Review The SMART framework covers the “Identify” step of the data science workflow. Types of datasets: Cross-Sectional vs. Longitudinal Questions?

DATA DICTIONARIES AND DOCUMENTATION

Data dictionaries are often our primary source to help judge the quality of our data and also to understand how it is coded.

- If our gender variables are coded 0 and 1, how do we know which is male and which is female?
- Is your currency variable coded in dollars or euros?

examples:

- Iris dataset and here
  - (Sepal vs Petal)
- Titanic Dataset
- Boston Housing dataset

CODE ORGANIZATION

- One folder per lesson
  - data
  - py / code / notebook
  - doc / notes

```
(py34)[16:08][aperrier@:~/apps/gads/01_what_is_a_data_scientist(master)]$ tree
.
├── data
|   ├── Online\ Retail.csv
|   └── Online\ Retail.xlsx
├── notebooks
|   └── Online\ Retail.ipynb
└── slides
    ├── Lesson\ 01\ What\ is\ data\ Science\ Part\ 1.pdf
    ├── Lesson\ 01\ What\ is\ data\ Science\ Part\ 2.pdf
    └── jupyter.md

3 directories, 6 files
```

NUMPY & PANDAS

NUMERICAL PYTHON: NUMPY

numpy.org

Get started

Numpy brings decades of C math into Python!

Numpy provides a wrapper for extensive C/C++/Fortran codebases, used for data analysis functionality

N-DIMENSIONAL ARRAY OBJECT

- Array Creation
- Manipulations: resize, reshape, split, ...
- Questions: any?, all?
- Ordering: sort, max, argmin,
- Operations: sum, prod,
- Basic Statistics and Linear Algebra: mean, std, dot, ...

NUMPY: CODEALONG

Create new jupyter notebook

```
> jupyter notebook
> import numpy as np
```

PANDAS

Pandas uses a data structure similar to a spreadsheet: **Dataframes** A Dataframe contains rows and columns.

http://pandas.pydata.org

Even more features than numpy:

- advanced selection
- transformations
- DF to DF operations
- plotting

PANDAS

The strengths of pandas lie in

- reading in data
- manipulating rows and columns
- adjusting indices
- working with dates and time series
- sorting, grouping, re-ordering and general data munging
- dealing with missing values, etc., etc.

=> reads csv and excel files!

PANDAS: CODEALONG

Create new jupyter notebook

```
> jupyter notebook
> import numpy as np
```

LAB

- Check basic features, such as column names, number of observations
- Find and drop missing values
- Find basic stats like mean, max

The purpose of this lab is to get some practice working with Pandas.

http://localhost:8888/notebooks/gads/02_research_design_and_pandas/py/Pandas-Lab.ipynb

LESSON REVIEW

REVIEW

- Good Questions
- Study types: Cross sectional vs longitudinal
- Numpy
- Pandas

BEFORE NEXT CLASS

BEFORE NEXT CLASS

GET SOME PRACTICE WITH PANDAS

DataFrames are the bread and butter of the Data Scientist

- How to create a new dataframe?
- Change column names
- Reindex
- Create new columns from existing ones

READ

- Common Excel Tasks Demonstrated in Pandas
- Pandas Demo on the Quantitative Economics blog

KAGGLE - TITANIC

Open a kaggle account and go through the Titanic tutorial. Excellent practice. (don't

despair it's the journey that counts)

5 QUESTIONS ABOUT TODAY

EXIT TICKET

http://bit.ly/1PghQGv

LINKS

- SMART criteria
- cross-sectional vs. longitudinal studies
- Common Excel Tasks Demonstrated in Pandas
- Pandas Demo on the Quantitative Economics blog
- pandas.pydata.org
- www.numpy.org