APS Workshop: **Introduction to Python**
San Francisco, CA, 24 May 2018

Data Wrangling

Christian C. Luhmann
Stony Brook University

Wrangling

- Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one “raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

- This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses.

en.wikipedia.org/wiki/Data_wrangling

Relevant Packages

- numpy
- pandas
- matplotlib

Relevant Packages

- numpy
  - Matrix representation
  - Linear algebra
  - Fast

numpy

4.1	3.4	2.6
12.6	8.1	1.2
6.2	10.4	5.8

dtype = float

numpy

4	3	2
12	8	1
6	10	5

dtype = int32

numpy

4+9j	3+6j	2+4j
12+4j	8+9j	1+8j
6+8j	10+6j	5+2j

dtype = complex

numpy

arbitrary # of dims

Relevant Packages

- numpy
  - Matrix representation
  - Linear algebra
  - Fast
- pandas
  - R-style dataframe
  - Best for a mixture of heterogenous data types (e.g., subject #, name, DOB)

Dataframe

PowerBI_Test_Data.xlsx - Excel

Stock Name	Symbol	Shares	Purchase Price	Cost Basis	Current Price	Market Value	Gain/Loss	Dividend/share	Annual Yield
Apple	AAPL	100	$90.00	$9,000.00	$144.13	$14,413.27	$14,269.14	$2.28	1.58%
Microsoft	MSFT	200	$32.00	$6,400.00	$65.57	$13,114.14	$13,048.57	$1.56	2.38%
Salesforce	CRM	150	$25.00	$3,750.00	$82.57	$12,385.50	$12,302.93	$0.00	0.00%
Oracle	ORCL	250	$50.00	$12,500.00	$44.56	$11,138.75	$11,094.20	$0.64	1.44%
Hewlett Packard Enterprise	HPE	500	$18.00	$9,000.00	$17.69	$8,842.50	$8,824.82	$0.26	1.47%
Alphabet	GOOG	100	$225.00	$22,500.00	$833.36	$83,336.00	$82,502.64	$0.00	0.00%
Intel	INTC	200	$22.00	$4,400.00	$36.07	$7,213.00	$7,176.94	$1.09	3.02%
Cisco	CSCO	225	$18.00	$4,050.00	$33.24	$7,478.78	$7,445.54	$1.16	3.49%
Qualcomm	QCOM	185	$65.00	$12,025.00	$56.48	$10,447.88	$10,391.40	$2.12	3.75%
Amazon	AMZN	50	$800.00	$40,000.00	$897.64	$44,882.00	$43,984.36	$0.00	0.00%
Redhat	RHT	100	$95.00	$9,500.00	$86.26	$8,626.00	$8,539.74	$0.00	0.00%
Facebook	FB	1000	$17.00	$17,000.00	$141.64	$141,640.00	$141,498.36	$0.00	0.00%
Twitter	TWTR	500	$45.00	$22,500.00	$14.61	$7,302.55	$7,287.94	$0.00	0.00%

Dataframe

	A	B	C	D	E	
1	Stock Name	Symbol	Shares	Purchase Price	Cost Basis	Curr
2	Apple	AAPL	100	$90.00	$9,000.00	
3	Microsoft	MSFT	200	$32.00	$6,400.00	
4	Salesforce	CRM	150	$25.00	$3,750.00	
5	Oracle	ORCL	250	$50.00	$12,500.00	
6	Hewlett Packard Enterprise	HPE	500	$18.00	$9,000.00	

Relevant Packages

- numpy
  - Matrix representation
  - Linear algebra
  - Fast
- pandas
  - R-style dataframe
  - Best for a mixture of heterogenous data types (e.g., subject #, name, DOB)
  - Lots of slicing and dicing options
- matplotlib
  - Matlab-style plotting

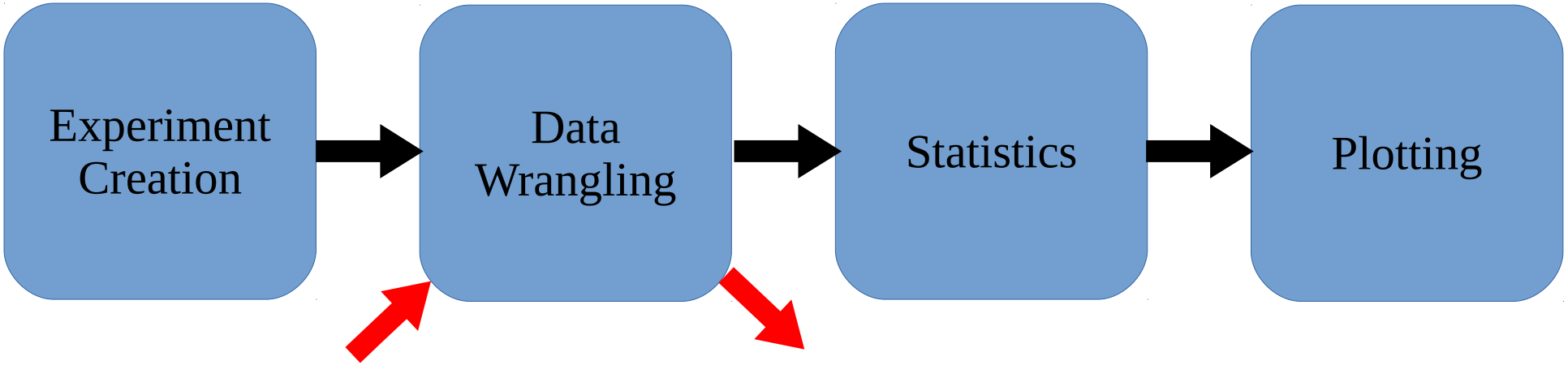
Wrangling

So let’s go wrangle some data

Pandas

- Pandas can read/write a variety of data formats...
  - CSV
  - JSON
  - HTML
  - Local clipboard
  - MS Excel
  - HDF5 Format
  - Feather Format
  - Parquet Format
  - Msgpack
  - Stata
  - SAS (read only)
  - Python Pickle Format
  - SQL
  - Google Big Query

The Pipeline



Take-homes

- You've now seen some **data wrangling** done in Python

- You've seen some of the functionality that **relevant packages** provide
  - pandas
  - jupyter (notebook)
  - matplotlib

- You have some sense of the **flexibility** provided by these tools

Outline

1. Overview
2. Ways of using Python
3. Python basics
4. Data set overview
5. Data wrangling
6. Statistics
7. Plotting
8. Experiment creation