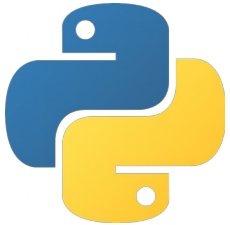


Workshop: **Introduction to Python**

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

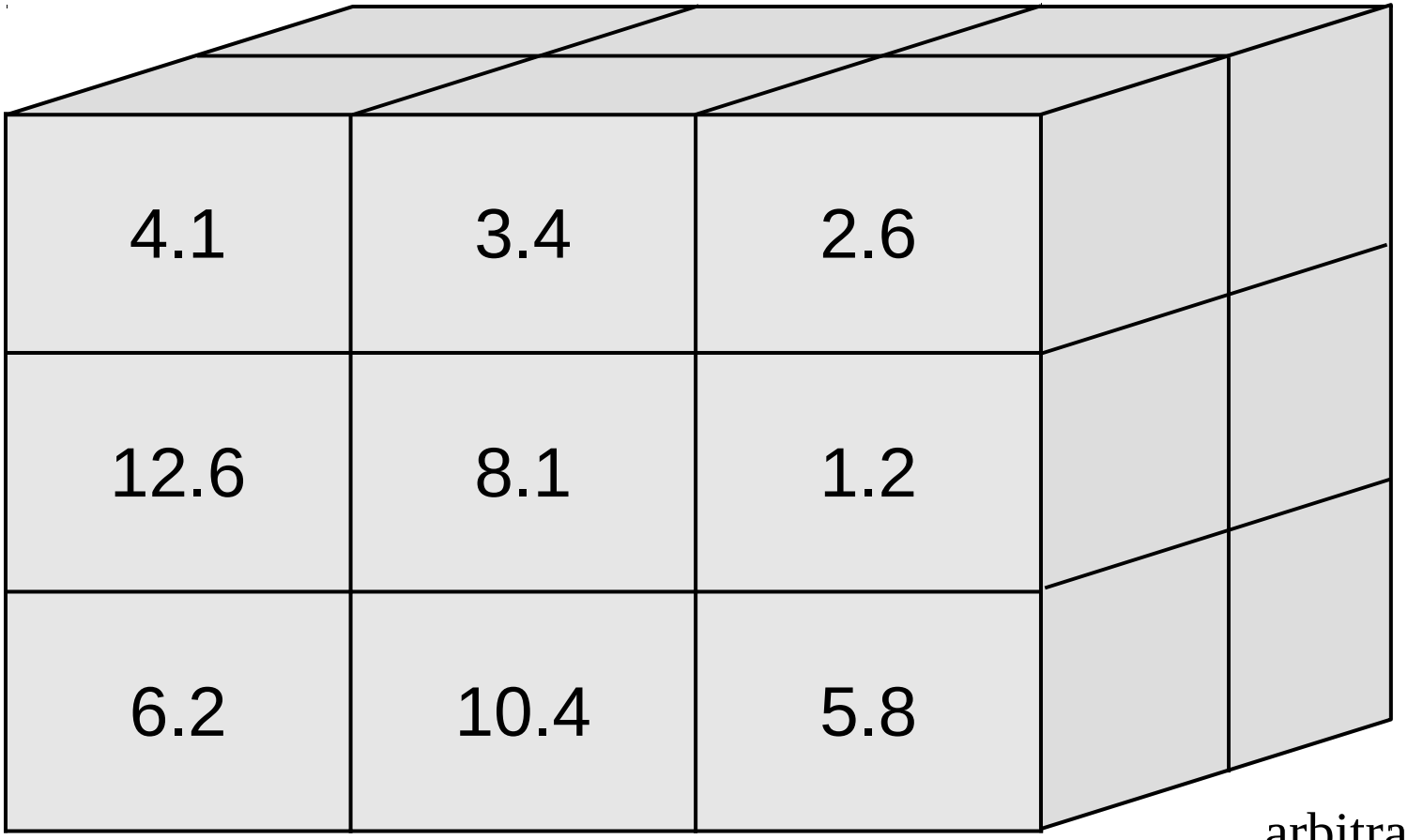
dtype = int32

numpy

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

dtype = complex

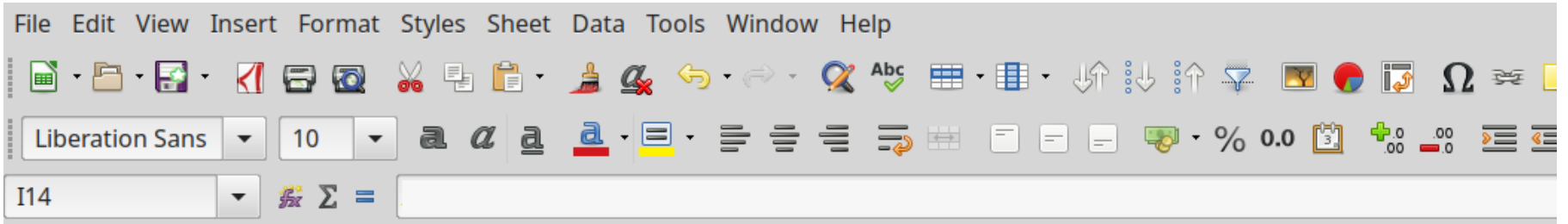
numpy



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)



	A	B	C	D	E	F
1	**Date**	**Salesperson**	**Item**	**Quantity**	**Cost**	**Total**
2	05/01/19	Mike	dishwasher	5	300	1500
3	04/24/19	Mike	washingmachine	4	500	2000
4	05/09/19	Mike	microwave	4	200	800
5	04/21/19	Mike	refrigerator	2	150	300
6	05/01/19	Mike	dishwasher	2	300	600
7	05/10/19	Mike	washingmachine	11	350	3850
8	05/19/19	Alice	microwave	2	500	1000
9	04/27/19	Alice	refrigerator	3	500	1500
10	05/20/19	Alice	microwave	11	350	3850
11	05/08/19	Alice	refrigerator	3	100	300
12	05/14/19	Alice	dishwasher	5	200	1000

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

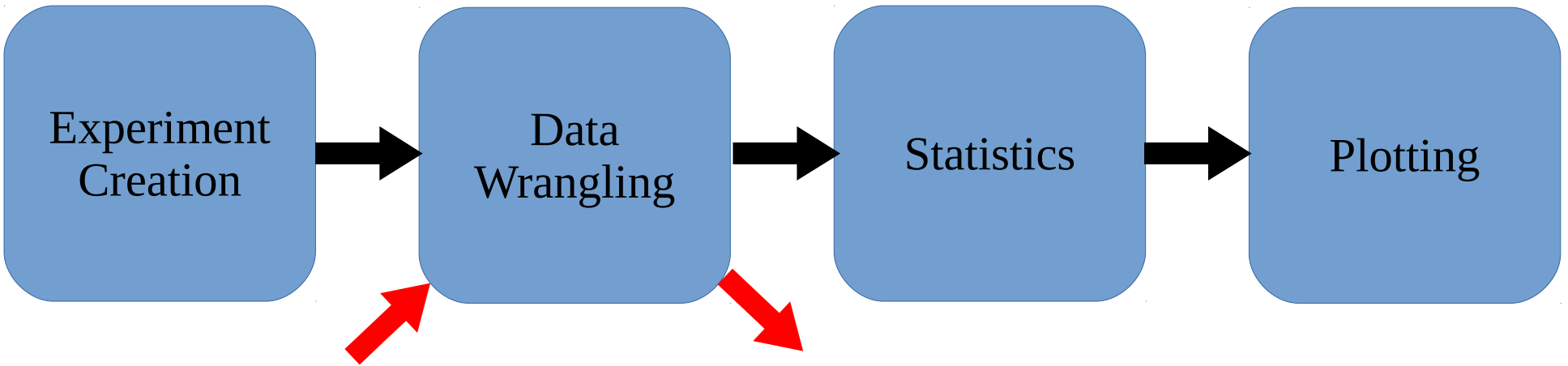
Wrangling

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