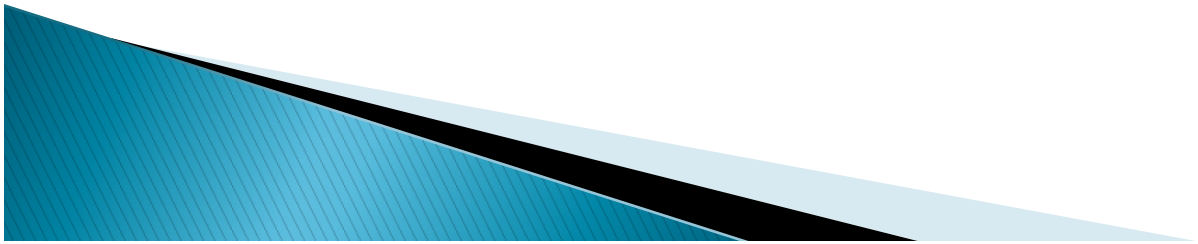
CSCI-UA.60-1 Database Design and Implementation

An Introduction to pandas

Prof Deena Engel
Department of Computer Science
deena.engel@nyu.edu

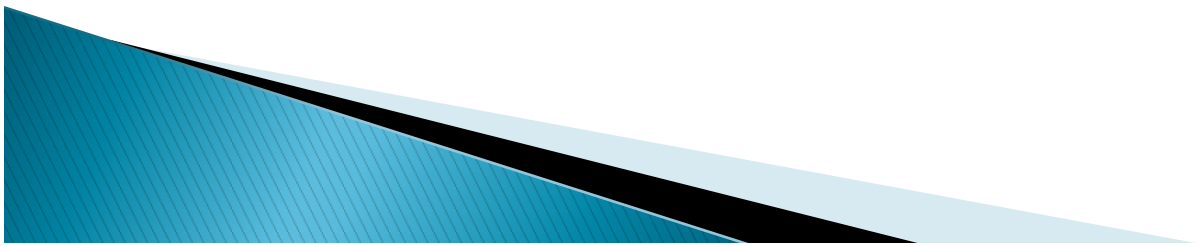
What is *pandas*?

- *pandas* stands for Panel Data System
- *pandas* is built on top of *NumPy*, *SciPy* and designed to work with *matplotlib*
- Open source



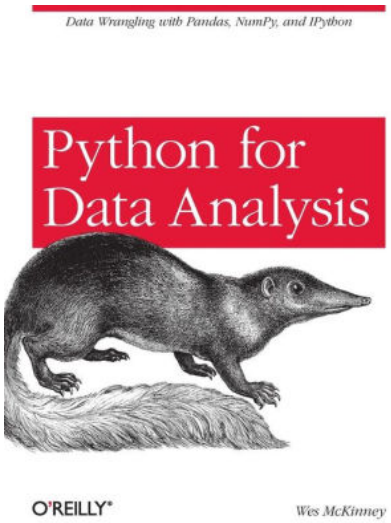
Why *pandas*?

- *pandas* was developed to provide functionality in Python for data analysis
- In practice, this means that you don't need to switch out to R, Matlab or another environment for basic data analysis



Developer goals

- *pandas* was initially developed by Wes McKinney



Python for Data Analysis by Wes McKinney (O'reilly,2013 and 2017)

Developer goals

- Data structures with labeled axes to support data alignment and prevent errors
- Integrated time series functionality
- Same data structures for time series data and non-time series data
- Flexible handling of missing data
- Merges and other relational operations found in SQL-based databases
- Arithmetic operations and summary operations would pass along the metadata (e.g. axes labels)

Python for Data Analysis by Wes McKinney (O'reilly,2013) - page 107

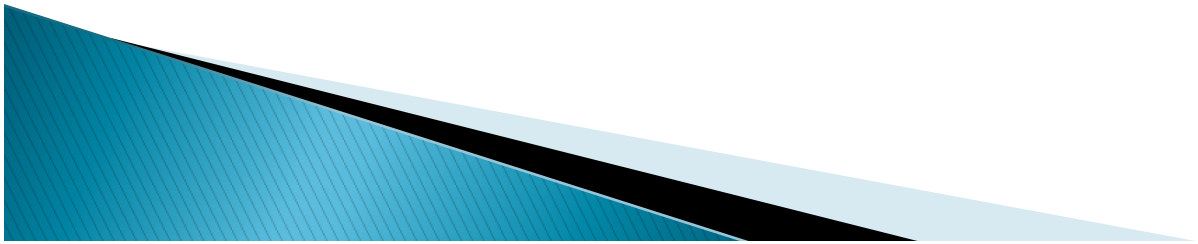
Getting started!

- *pandas* is included in your *Anaconda* installation.
- You can update *pandas* at the command line:

```
conda update pandas
```

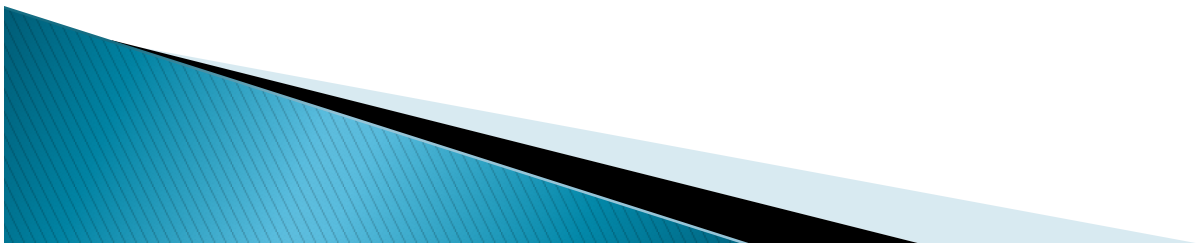
- By convention, import pandas and assign an alias of *pd*:

```
import pandas as pd
```



The pandas data structures

- Series
  - A one-dimensional array (list) of data
  - An associated one-dimensional array of data labels called its index.
  - If not otherwise noted, the index is like the index of a Python list: [0,1,2,3 …]
- DataFrame
  - A tabular, spreadsheet-like data structure containing rows and columns.
  - A dataFrame has both row and column indexes



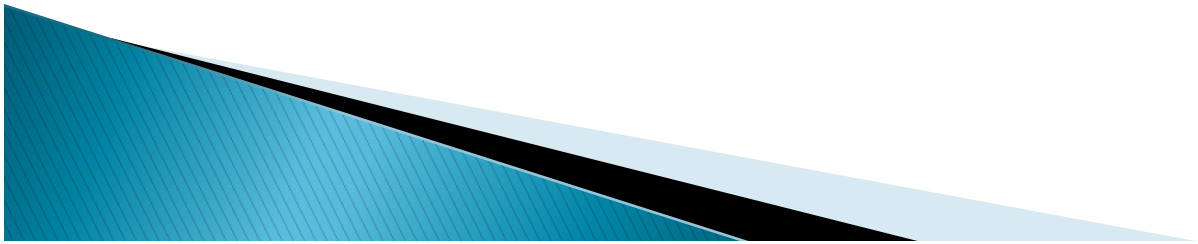
Getting your data into *pandas*

- Use a *pandas* to read in data:

```
pd.read_csv('file.csv')
```

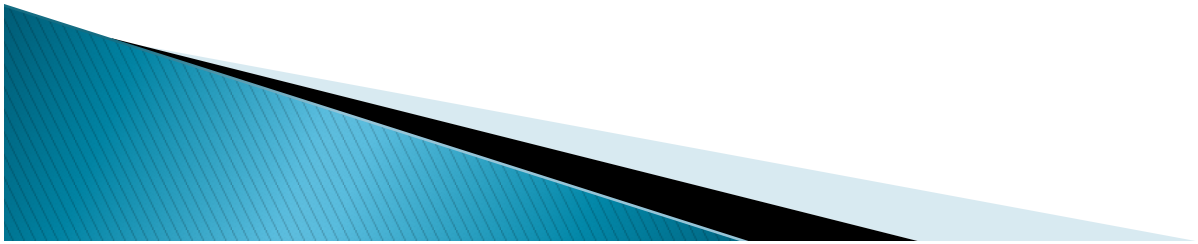
- In addition, you might consider `pd.read_clipboard` for tables from webpages and some of the other read methods

- See https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html



Getting your data into *pandas*

- Some of the parameters we will discuss in class:
  - `skiprows = n` – how many rows to skip at the top
  - `sep=","` – delimiter
  - `quotechar = '"'` – quote character
  - `nrows = n` – how many rows to read (e.g. read a small number of rows in a large file)
  - `encoding = 'utf8'` – for encoding



Descriptive and Summary Statistics

- `count`
- `min, max`
- `sum`
- `mean`
- `median`
- `var`
- `std`
- `pct_change`
- … and more: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html

Python for Data Analysis by Wes McKinney (O'reilly,2013) – page 135-6

Questions?

Send email to deena.engel@nyu.edu