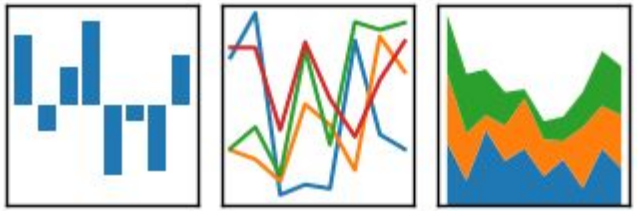
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



What is it ?

> Data Analysis Toolkit

Provides fundamental high-level building blocks for doing practical, **real word** data analysis in Python

Goals

- Open source, BSD-licensed
- High-performance
- Ease of use

What is it good at ?

Most notably

- High level container types
- IO / file formats
- Time series
- Groupby-split-apply combine
- Data handling
  - Reshaping
  - Merging, joining
  - Missing values
- Indexing
- And much more...

But

- pandas is not a statistics package

When can I use it ?

Typical Use Cases

- Flat and structured data, tabular/panel (not intrinsically n-dimensional)
- Fits in memory
- Interactive development

CSV, XLSX, SQL, HDF, … ✔

More dimensions?

- Try xarray, ...

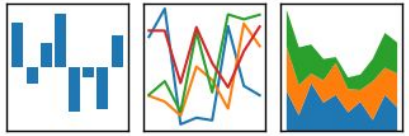
Out of memory?

- Bigger machine?
- Else: try dask, Apache Spark, ...

Where does it come from ?

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Builds on top of

- NumPy
- SciPy
- Matplotlib
- Cython
- and others...

Can you please show some code ?

Time for...

… a demo

Questions?

Thank you for your attention!

Claus Aichinger
claus.aichinger@gmail.com

Django Vienna 2019-03-19

Links

- https://pandas.pydata.org/

Pictures taken from

- https://github.com/pandas-dev/pandas
- https://scipy.org/
- https://xkcd.com/292/