05: Short introduction to pandas

https://github.com/matthiaskoenig/itbtechtalks

Dr Matthias König
Humboldt University Berlin,
Institute for Theoretical Biology

pandas

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

Python data analysis library

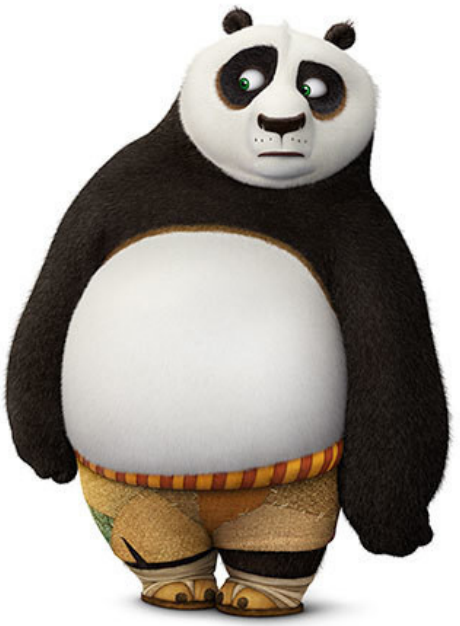
- **high-performance, easy-to-use data structures** and data analysis tools
- fundamental high-level building block for doing practical, **real world data analysis**

Applications

- **tabular data with heterogeneously-typed columns** (as in SQL table or excel spreadsheets)
- ordered/unordered (not necessarily fixed-frequency) **time series data**
- **arbitrary matrix data** (homogeneously or heterogeneously typed) with row and column labels
- any other form of observational/statistical data sets

Primary data structures

- Series (1D)
- DataFrame (2D)



https://pandas.pydata.org
https://pandas.pydata.org/pandas-docs/stable/10min.html

pandas

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

https://pandas.pydata.org
https://pandas.pydata.org/pandas-docs/stable/10min.html

Things that pandas does well

- easy handling of **missing data**
- **size mutability** (inserting/deleting)
- automatic and explicit **data alignment** (based on labels)
- powerful, flexible **group by** (split-apply-combine operations)
- **slicing, fancy indexing, subsetting**
- **merging** and **joining** data sets
- flexible **reshaping**
- **robust IO** tools (flat files, databases, excel, hdf5, …)
- **Time series**-specific functionality (moving window statistics)