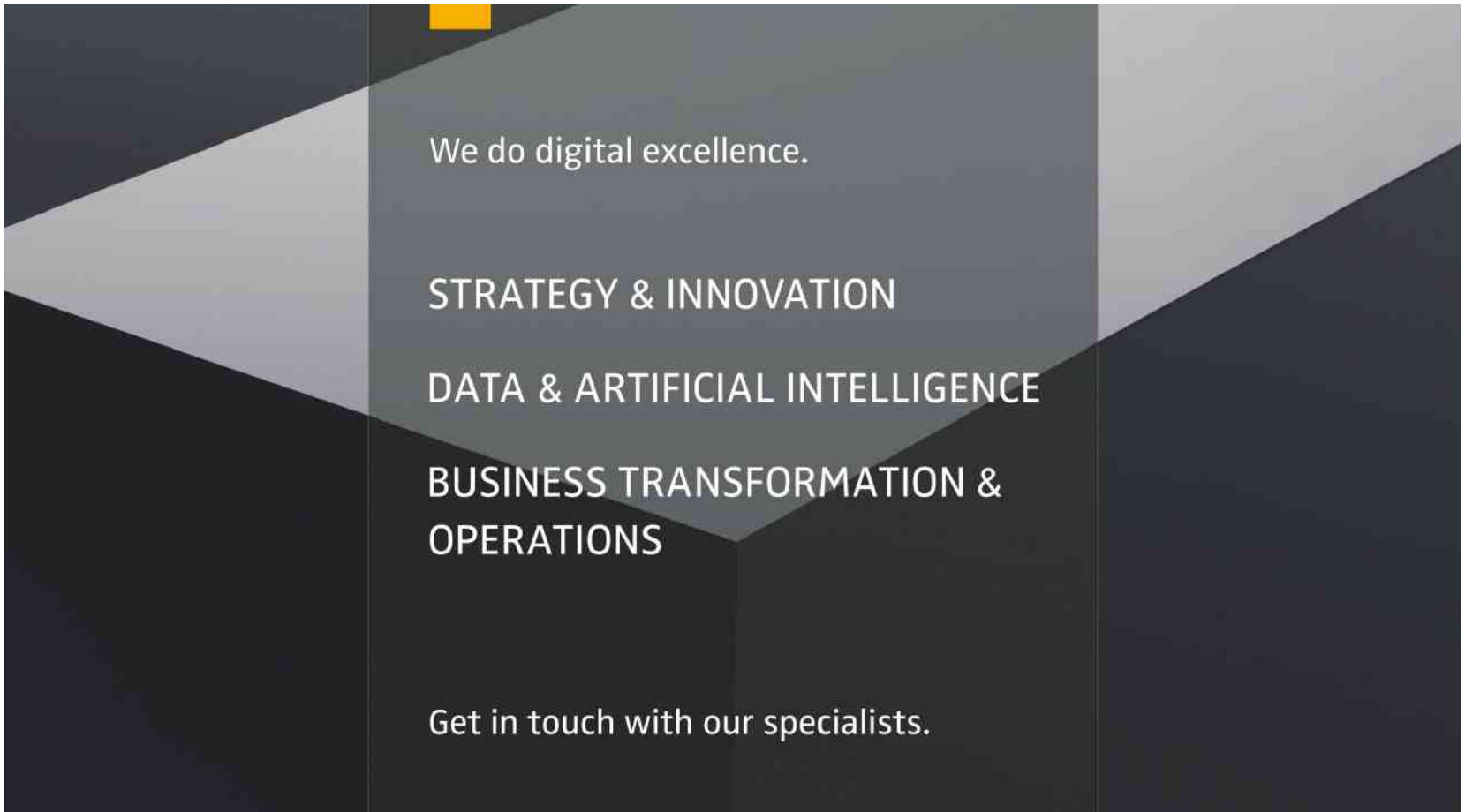
KŌNIGSWEG

Analytics with Pandas and Jupyterlab

ALEXANDER CS HENDORF FOR PYLADIES HAMBURG


KŌNIGSWEG
We do digital excellence.
STRATEGY & INNOVATION
DATA & ARTIFICIAL INTELLIGENCE
BUSINESS TRANSFORMATION & OPERATIONS
Get in touch with our specialists.

KŌNIGSWEG

LIKE THE PICTURES IN THE SLIDES?

COME IN THE BREAK AND MAKE ONE OF YOUR OWN!

Foto ändern

Mitglied von **PyData – 142 Gruppen**

PyData Südwest

Mannheim, Deutschland

737 Mitglieder · Öffentliche Gruppe

Organisiert von **Alexander C. S. Hendorf** und 5 andere

Teilen:

Über uns | Events | Mitglieder | Fotos | Diskussionen | Mehr

Gruppe verwalten | Event erstellen

Worum es bei uns geht

Welcome to PyData Südwest!

...

Mehr lesen

Bevorstehende Events (1)

Alle anzeigen

Organisatoren

Alexander C. S. Hendorf und 5 andere

Nachricht

Mitglieder (737)

Alle anzeigen


NUMFOCUS
[SPONSORED PROJECTS]
KÖNIGSWEG
IPy
IPython
pandas


KÖNIGSWEG

PyCon.DE

PyData
Berlin

Joint conference

October 14 to 16
at bcc, Berlin

Ticket sale opens April 6
CfP is open until May 5
Fin Aid is open until June 1

pycon.de. @pyconDE
berlin.pydata.org @pydataBER

KÖNIGSWEG

Alexander C. S. Hendorf

ah@koenigsweg.com

@hendorf

- Managing Partner & Principal Consultant Information Technology Consulting on AI & Data Science
- Python Software Foundation Fellow,
  Program Chair EuroPython, EuroSciPy, PyConDE & PyData, MongoDB Master Speaker Europa & USA MongoDB World New York / San José, PyCon Italy, CEBIT Developer World, BI Forum, IT-Tage FFM, PyData, PyParis, PyData Südwest & Frankfurt

Prepared?

- Follow instructions: git http://bit.ly/pandas-basel
- Make sure you have the latest version: *git pull*

KÖNIGSWEG

Introduction to Data Analytics with Pandas and Juypterlab

Jupyter

- Ecosystem
- Benefits of Jupyter
- Jupyter Notebooks
- Jupyterlab

Pandas

- Benefits of Pandas
- How to work with Pandas
- Visualisation

iPython

```
IPython 5.3.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.

In [1]: n = 100000

In [2]: import numpy as np

In [3]: %timeit np.sum(1. / np.arange(1., n) ** 2)
The slowest run took 8.35 times longer than the fastest. This could mean that an intermediate result is being cached.
1000 loops, best of 3: 186 µs per loop

In [4]: np.arange
```

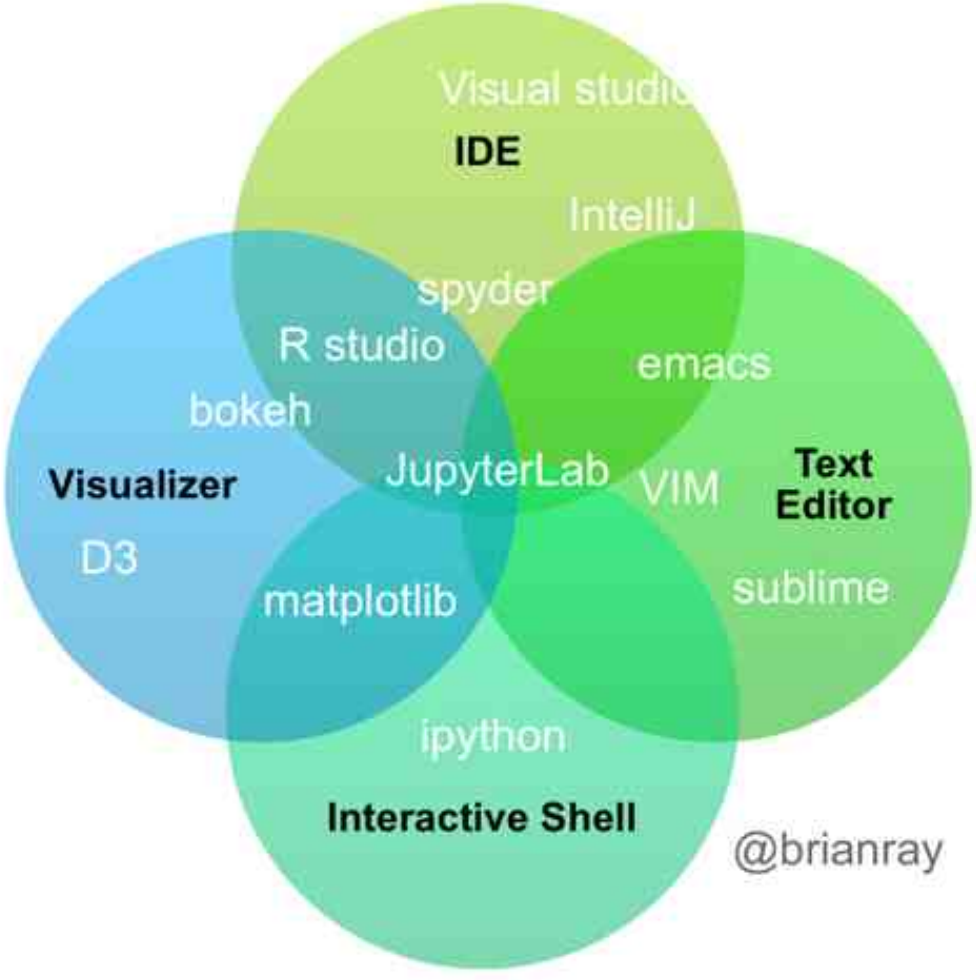
np.arange	np.arcsin	np.arctan2	np.argmin	np.argwhere	np.array2string	np.array_repr
np.arccos	np.arcsinh	np.arctanh	np.argpartition	np.around	np.array_equal	np.array_split
np.arccosh	np.arctan	np.argmax	np.argsort	np.array	np.array_equiv	np.array_str

Jupyter Notebooks

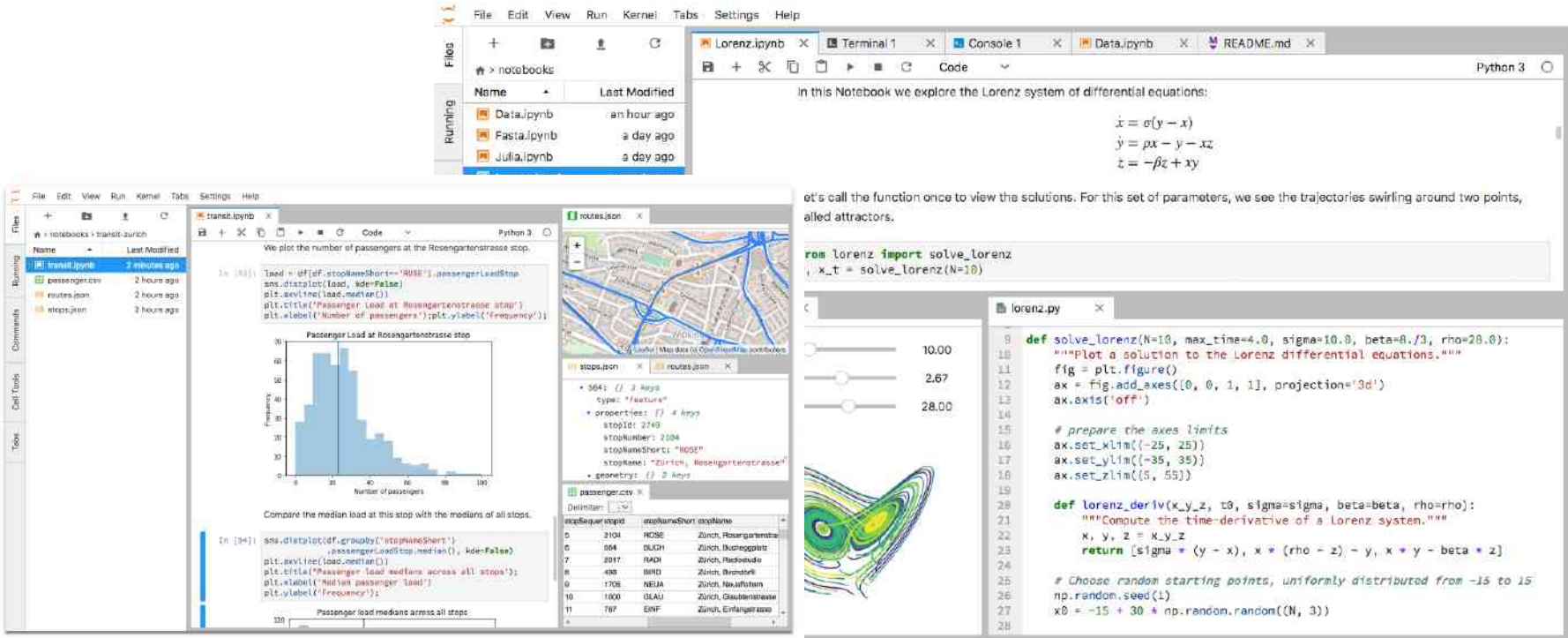
- Code, text (docs, background, research, references,...) und Visualisation
- Full programme / script
- IDE
- Explore iteratively
- Reproducable and customizeable
- Export to other formats(HTML, PDF, Python module,...)

Jupyterlab

- Was in the works since 2016
- Released in January 2018
- The *future* Jupyter Notebooks
- Compact DIE
- Extendable


Visual studio
IDE
IntelliJ
spyder
R studio
emacs
bokeh
Visualizer
JupyterLab
VIM
Text Editor
D3
matplotlib
sublime
ipython
Interactive Shell
@brianray

Hands on Jupyterlab



Fun Fact

–

„Anyone can learn Python,

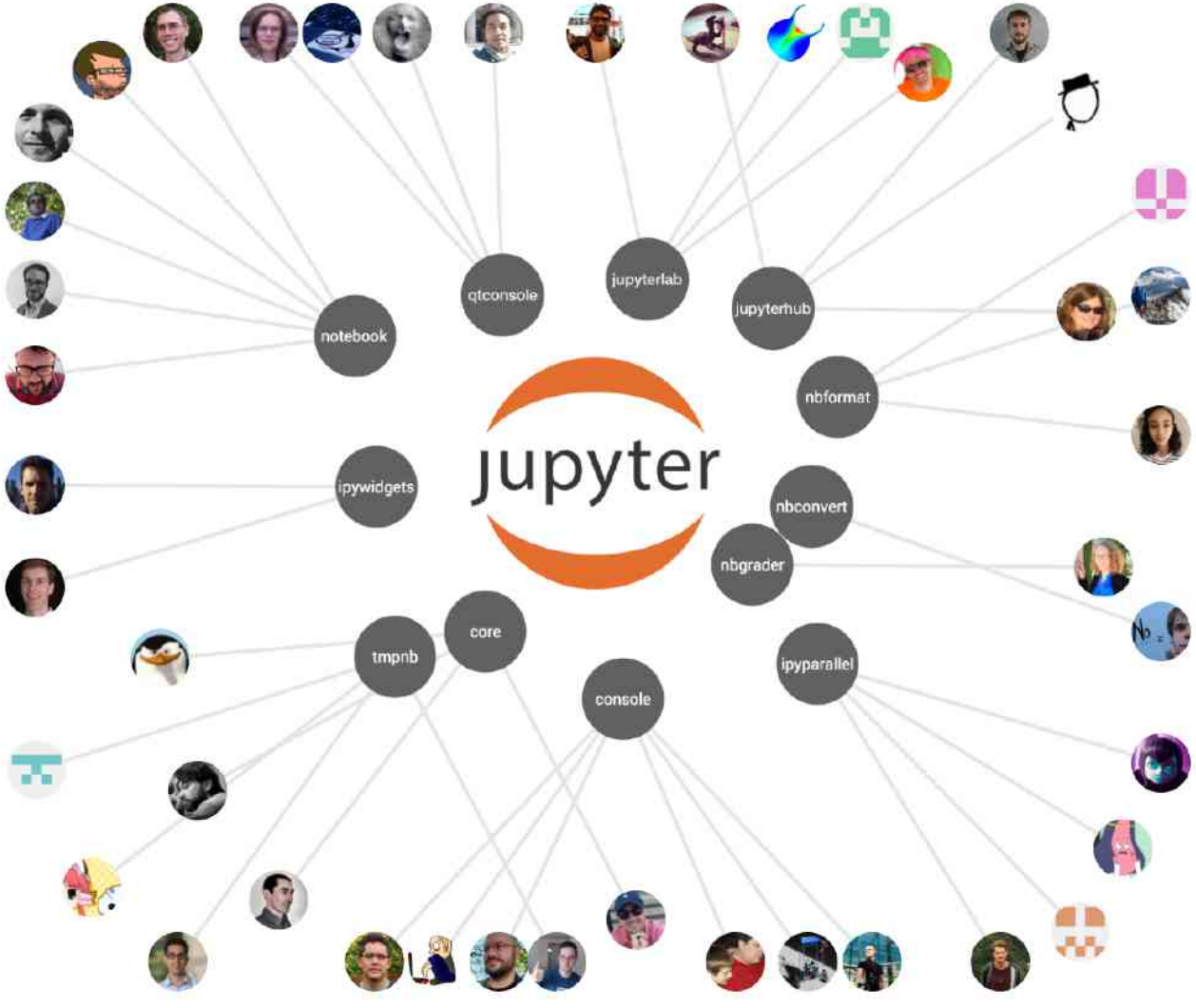
at least for Data Analytics.“

Business Fact

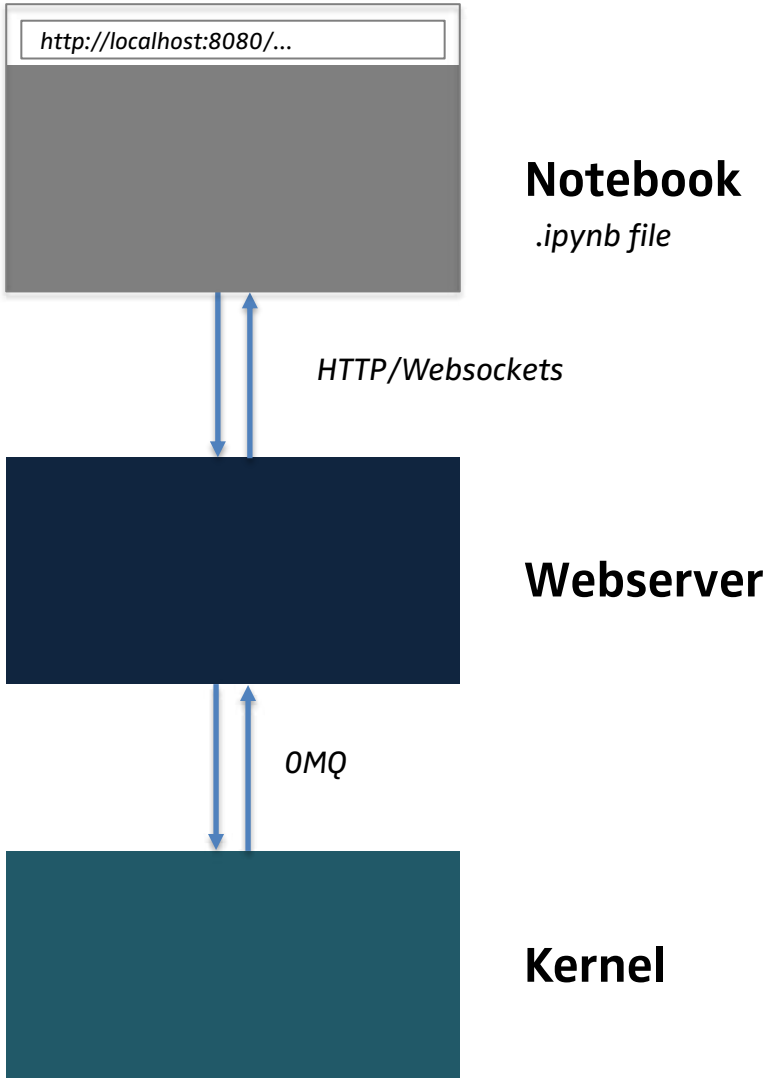
„Python is a perfect common language for a herterogeneous group.“

Jupyter

- *Jupyter supports Python, R, Julia...*
- *Language independent features:*
  - Notebook
  - Message queue
  - Qt-console
- *Open Source, modified BSD license*


jupyter
notebook
qtconsole
jupyterlab
jupyterhub
nbformat
ipywidgets
nbconvert
nbgrader
core
tmpnb
console
ipyparallel

Architecture



Jupyter Notebooks

- **Document:**
  - Executable code
  - Rich text elements: markdown, LaTex
  - Visualisations
- **Notebook App:**
  - server-client application allowing editing and running notebook documents via a web browser
- **Kernels:**
  - computational engine
- **Dashboard:**
  - manager

Anaconda Distribution

- Anaconda CPython distribution (covers 2.7 + 3.6)
- Package management *conda*
- 1000+ data-science libraries
- Ensures **packages are compatible** with each other
  (newer version of a package may have API changes)
- Provided by Continuum Analytics

Jupyterlab Extentions

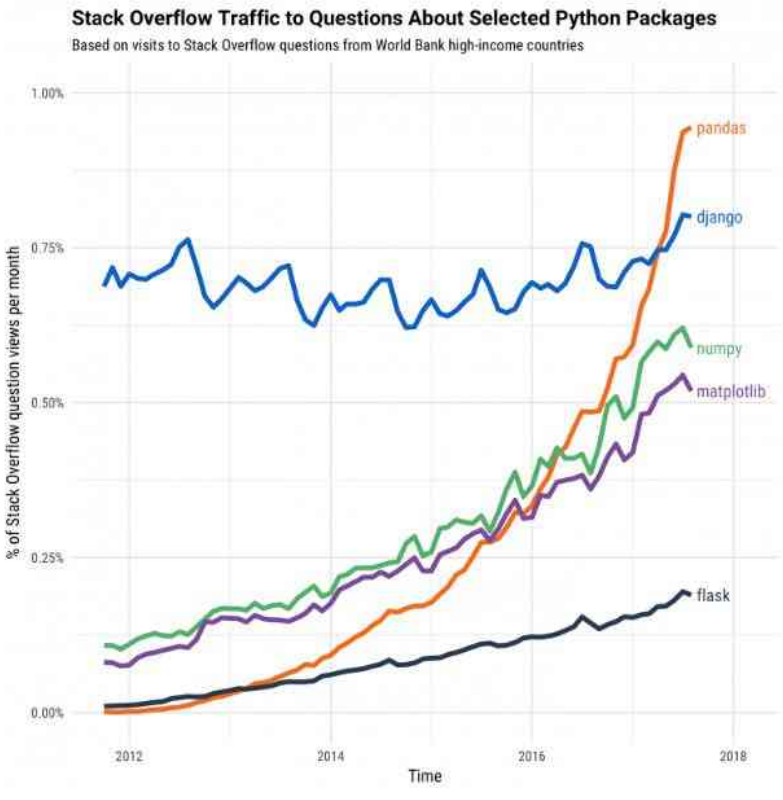
- JupyterLab is designed as an extensible environment
- Extensions can customize or enhance any part of JupyterLab
- Extensions are npm packages – JavaScript ;)
- Installation via Extension Manager or command line

KŌNIGSWEG

Pandas

- Open Source Python Library
- Praktische *'real-world*-Datenanalyse - schnell, effizient & einfach
- Lückenloser Datenanalyse Workflow (ohne Wechsel in z.B R)
- 2008 begonnen von Wes McKinney,
  nun PyData Stack bei Continuum Analytics ("Anaconda")
- Sehr Stabiles Projekt mit regelmäßigen Updates
- https://github.com/pydata/pandas

Development



Pandas Main Features

- Support for CSV, Excel, JSON, SQL, SAS, clipboard, HDF5,...
- Data cleansing
- Re-shape & merge data (joins & merge) & pivoting
- Data Visualisation
- Well integrated in Jupyter (iPython) notebooks
- Database-like operations
- Performant

NumPy under the hood

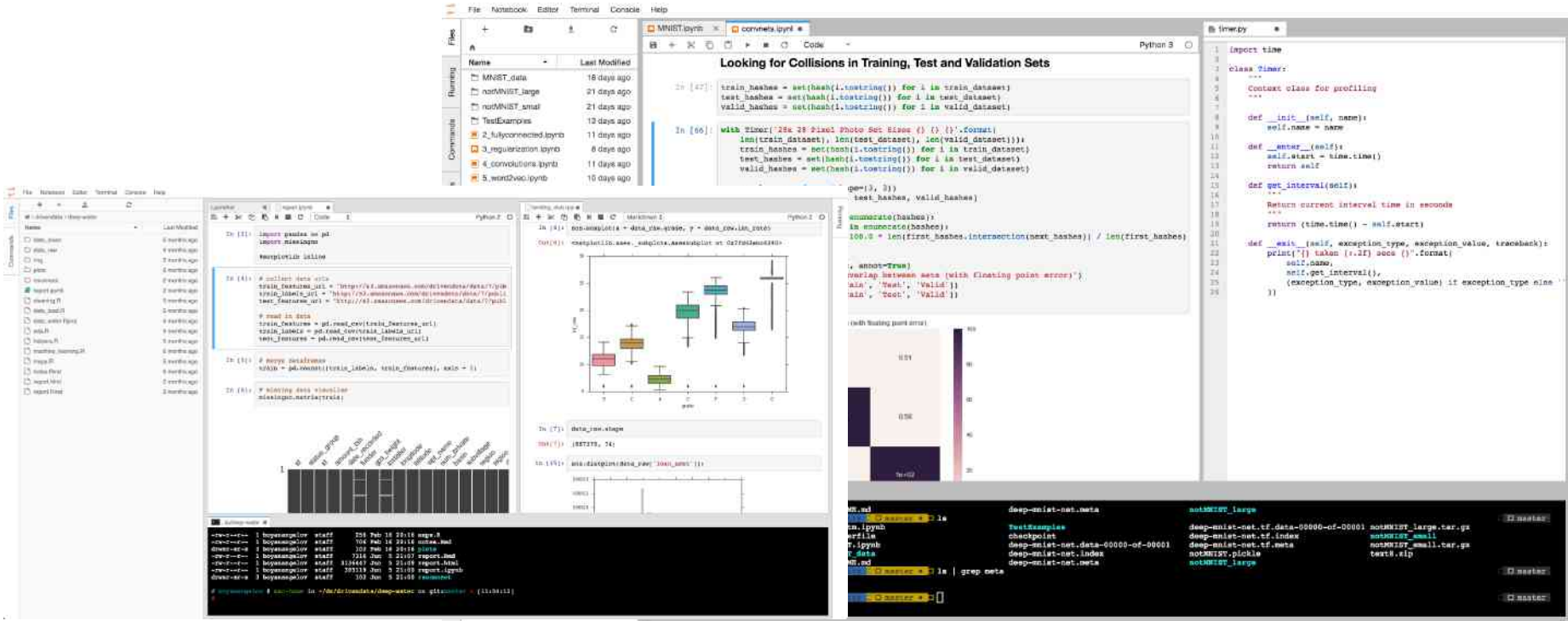
- Library for numerical operations in Python
- Typed Arrays
- Broadcasting

KÖNIGSWEG

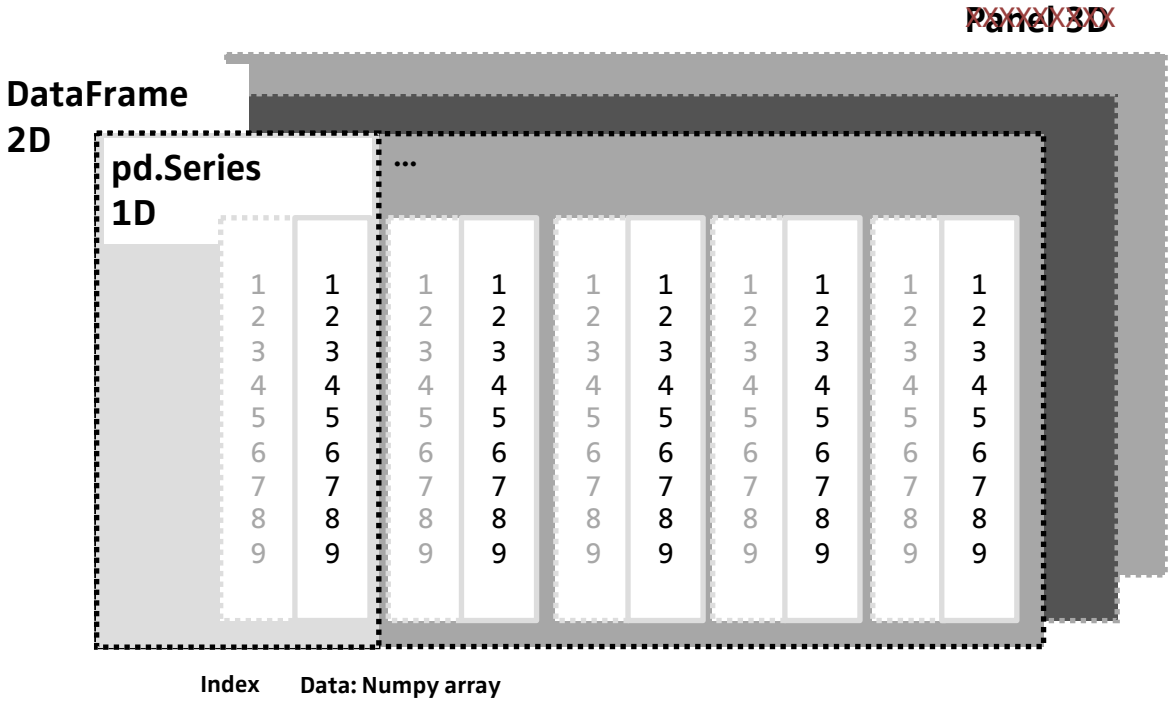
Hands on Pandas

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Structure



The Index

- Label of a DataSeries
- Immutable but replaceable
- One or more Dimensions
- Labels are not necessarily *unique*

Index Types

- Index
- MultiIndex
- DateTimeIndex
- TimeDelta
- IntervalIndex
- CategoricalIndex

Basic Stats & Aggregation

- **describe()**
- **Aggregation**
  - sum, count, custom fucntions,...
  - grouping
  - privoting
- **NaN (null) values and filler**

Visualization

- Close to the data / code
- Highly customizeable
- Many tools: matplotlib, seaborn, bokeh

Automation

— Use nbcovert

KÖNIGSWEG

Pandas Performance, Limits & Solutions

- Data sets 2-5GB
- stream processing via stepwise aggreagtion
- Dask for Distributed DataFrames
- Integration with pySpark, SciKit Learn
- Project Arrow

Jupyter Hub

- Jupyter as server for teams
- Collaboration
- Less overhead on local computers
- Access control

KÖNIGSWEG

Alexander C. S. Hendorf

ah@koenigsweg.com

@hendorf

- Managing Partner & Principal Consultant Information Technology Consulting on AI & Data Science
- Python Software Foundation Fellow,
  Program Chair EuroPython, EuroSciPy, PyConDE & PyData, MongoDB Master Speaker Europa & USA MongoDB World New York / San José, PyCon Italy, CEBIT Developer World, BI Forum, IT-Tage FFM, PyData, PyParis, PyData Südwest & Frankfurt

KŌNIGSWEG

Thank you!

Q & A

koenigsweg.com

ah@koenigsweg.com

@hendorf