

K Ŷ N I G S W E G



Git with instructions: <https://goo.gl/4cAfXP>

Introduction Data Analytics with Pandas and Jupyter

Workshop



Alexander C. S. Hendorf

- Senior Consultant Information Technology
- Program Chair EuroPython, PyConDE & PyData Karlsruhe,
PSF Managing Member
- Program Committee EuroSciPy, Percona Live
- MongoDB Masters / MongoDB Certified DBA
- Speaker Europa & USA MongoDB World New York / San José,
PyCon Italy, CEBIT Developer World, BI Forum, IT-Tage FFM,
PyData, PyParis



ah@koenigsweg.com
 @hendorf

Königsweg Strategieberatung

INNOVATIONS-STRATEGIE

Wir befähigen Sie mithilfe intensiver Assessments und konsequenter Implementierung zur Übersetzung Ihres innovativen Potenzials in messbaren Erfolg.

DIGITALE TRANSFORMATION

Wir führen Ihr Unternehmen sicher durch die Transformationsprozesse von Digitalisierung und Industrie 4.0.

BUSINESS INTELLIGENCE

Wir implementieren Data-Science-Verfahren, präzisieren Ihre Geschäftsprognosen und optimieren Ihre Prozesseffizienz.

CLUSTER MANAGEMENT

Wir verknüpfen High-Tech Start-ups mit relevanten Schlüsselakteuren, Investoren und Industrie-Partnern.

START-UP FINANZIERUNG

Wir unterstützen Start-ups und innovative Hightech Unternehmen bei der Realisierung von Eigenkapital-, Fremdkapital- sowie Mezzanine-Finanzierungen.

Introduction to Data Analytics with Pandas and Jupyter

Jupyter

-
- Ecosystem
 - Benefits of Jupyter
 - Jupyter Notebooks

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.argmax      np.array2string np.array_repr  
np.arccos      np.arcsinh     np.arctanh     np.argmax      np.around      np.array_equal   np.array_split  
np.arccosh     np.arctan      np.argmax      np.argsort     np.array      np.array_equiv  np.array_str
```

Hands on

The image shows two side-by-side Jupyter notebook interfaces running on localhost. Both notebooks are titled "data2day 2018".

Left Notebook Content:

- Section 1:** "Interaktive Datenanalyse n" (with a small 'n' icon).
 - Text: "3 minutes Markup Crash-Course".
 - Text: "Dies ist eine kleine Demo über die Möglichkeiten, unterschiedlicher Gewichtung können mit #".
 - Code cells:
 - In [1]: `1 import pandas as pd
2 import numpy as np
3 %matplotlib inline`
 - In [2]: `1 %config InlineBackend.figure_format = 'retina'`
 - In [3]: `1 df = pd.DataFrame(np.random.randn(1000, 2), columns=['a', 'b'])`
 - In [4]: `1 df['b'] = df['b'] + np.arange(1000)`
 - In [5]: `1 df['z'] = np.random.uniform(0, 3, 1000)`
 - In [6]: `1 df.plot.hexbin(x='a', y='b', C='z', reduce_C_function=np.max,
2 gridsize=25);`
 - Output: A hexbin plot showing a dense cloud of points colored by 'z' values from 0.5 to 2.5.

Right Notebook Content:

 - Section 2:** "Interaktive Datenanalyse mit Jupyter und Pandas".
 - Code cells:
 - In [1]: `1 import pandas as pd
2 import numpy as np
3 %matplotlib inline`
 - In [2]: `1 %config InlineBackend.figure_format = 'retina'`
 - In [3]: `1 df = pd.DataFrame(np.random.randn(1000, 2), columns=['a', 'b'])`
 - In [4]: `1 df['b'] = df['b'] + np.arange(1000)`
 - In [5]: `1 df['z'] = np.random.uniform(0, 3, 1000)`
 - In [6]: `1 df.plot.hexbin(x='a', y='b', C='z', reduce_C_function=np.max,
2 gridsize=25);`
 - Output: A hexbin plot showing a dense cloud of points colored by 'z' values from 0.5 to 2.5.

Jupyter Notebooks

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



Fun Fact

-

*„Anyone can learn Python,
at least for Data Analytics.“*

Business Fact

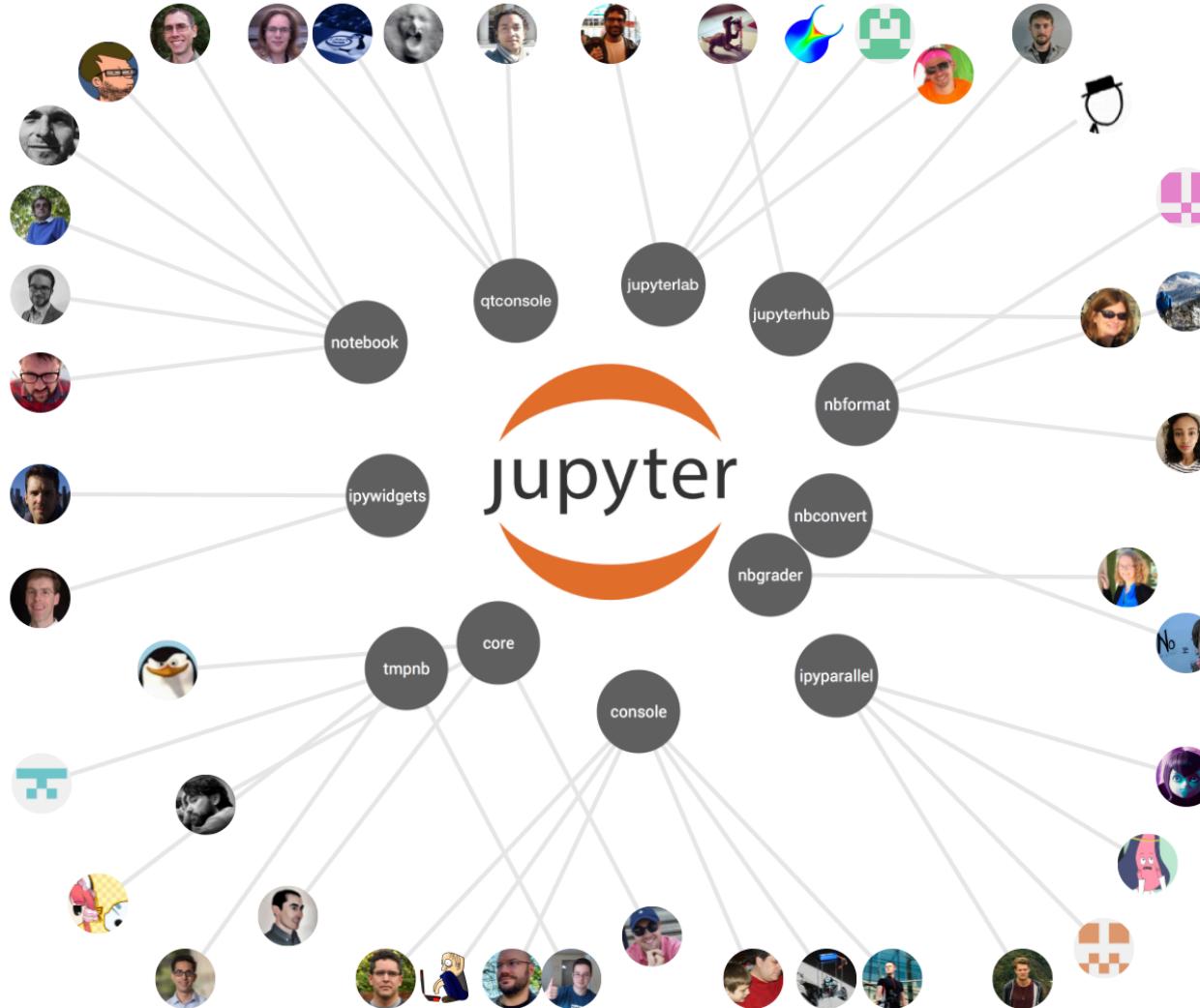
-

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

Jupyter

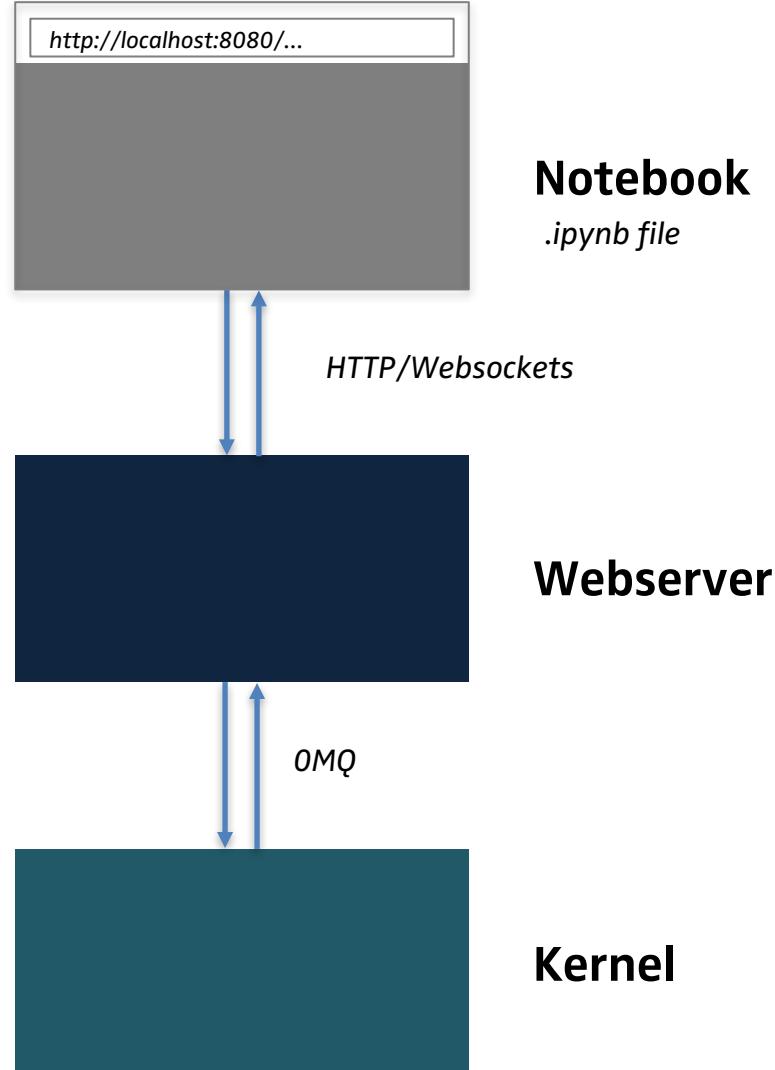
- *Jupyter supports Python, R, Julia...*
- *Language independent features:*
 - Notebook
 - Message queue
 - Qt-console
- *Open Source, [modified BSD license](#)*







Architektur



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

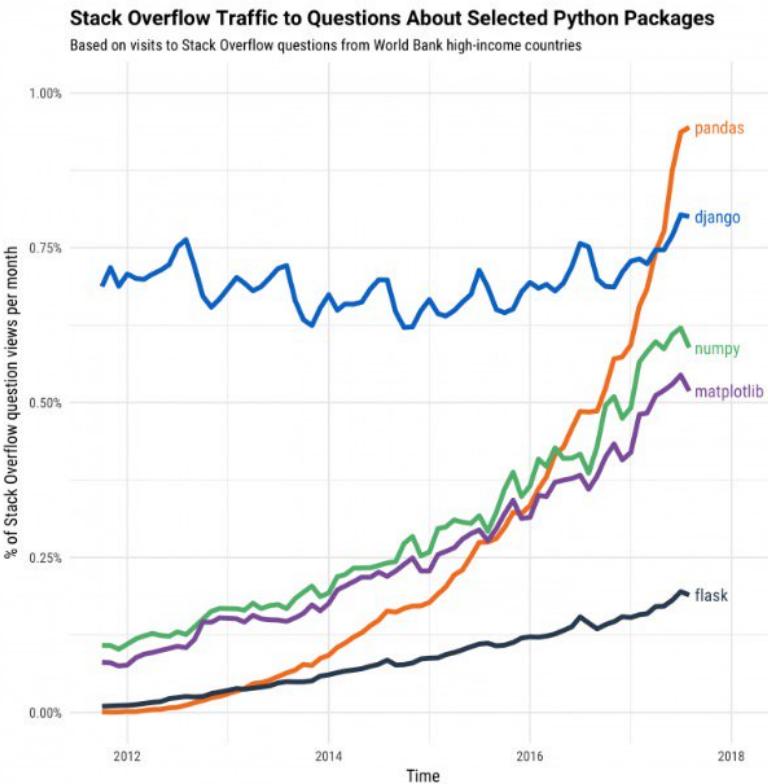


Pandas

- 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>



Entwicklung



Pandas Haupt-Funktionalitäten

- 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

Hands on

The image displays three separate Jupyter Notebook interfaces, each showing a different aspect of data analysis using Pandas.

- Top Notebook:** Titled "data2day" and "Interaktive Datenanalyse mit Jupyter und Pandas". It contains a section titled "Pandas: Daten einlesen und ausgeben".

```
In [2]: 1 import pandas as pd  
import as pd is a widely  
in our data directory we  
In [3]: 1 # make sure we  
2 print(pd.__ve  
3  
0.20.3 is accepta  
In [1]: 1 # a Jupyter m  
2 pwd  
Out[1]: u'/Users/hendorf/  
tebooks'  
In [6]: 1 # import  
2 sales_data =
```

Datenset

```
In [7]: 1 pd.set_option  
2 sales_data  
Out[7]:
```

	name	birthd
0	Pasquale	196
1	India	196
- Middle Notebook:** Titled "data2day" and "Interaktive Datenanalyse mit Jupyter und Pandas". It contains a section titled "Pandas: Daten Selektion & Indexe".

```
In [1]: 1 import numpy as np  
2 import pandas as pd  
3 import random
```

Series

```
In [4]: 1 series = pd.Series  
In [5]: 1 series  
Out[5]:
```

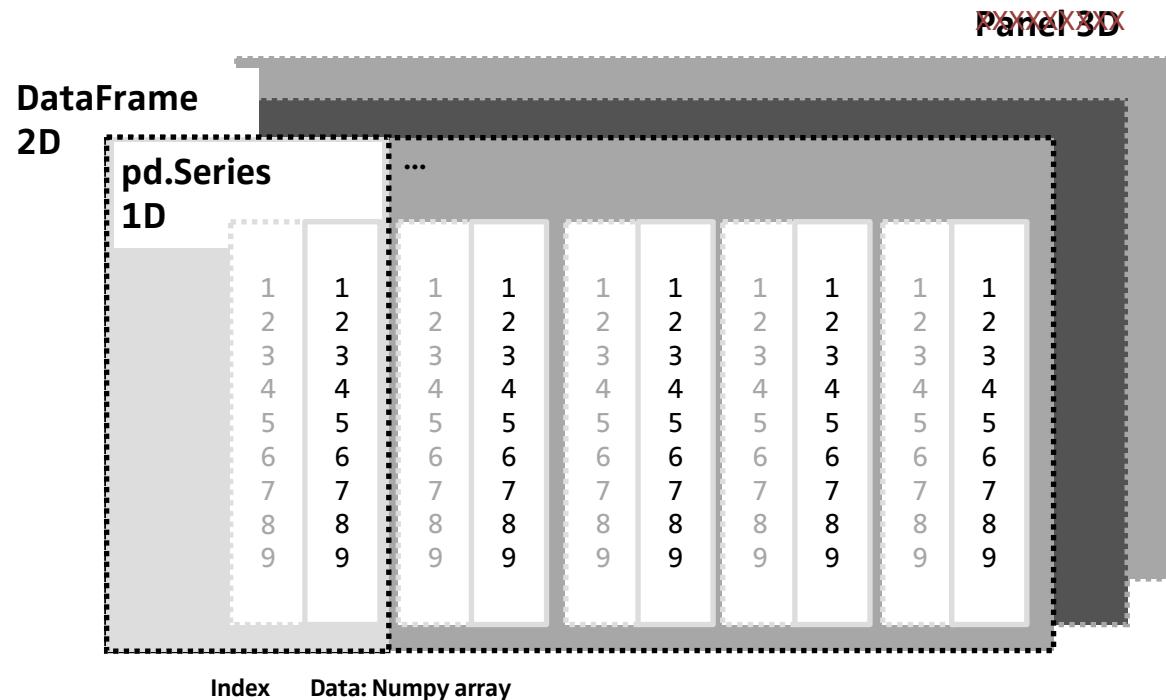
	0	1	2	3	4	5	6	7
0	3	62	75	83	47	43	39	16
- Bottom Notebook:** Titled "data2day" and "Interaktive Datenanalyse mit Jupyter und Pandas". It contains a section titled "Pandas: Daten Selektion & Indexe".

```
In [10]: 1 ax = tu.plot.bar()  
2 ax.axhline(tu['units'].median(), color='red', linestyle='solid')
```

Out[10]: <matplotlib.lines.Line2D at 0x11b020390>

A bar chart with blue bars representing the 'units' data. A red horizontal line is drawn across the chart at a value of approximately 2500, representing the median.

Struktur



The Index

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

Index Types

- `Index`
- `MultIndex`
- `DatetimeIndex`
- `Timedelta`
- `IntervalIndex`
- `CategoricalIndex`

Basic Stats & Aggregation

- `describe()`
- Aggregation
 - sum, count, custom functions,...
 - grouping
 - pivoting
- NaN (null) values and filler

Visualization

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

Automation

- Use nbconvert

Pandas Performance, Limits & Solutions

- Data sets 2-5GB
- stream processing via stepwise aggregation
- 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

Thank you!

Q & A

koenigsweg.com

ah@koenigsweg.com



@hendorf

