KŌNIGSWEG

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

Introduction Data Analytics with Pandas and Juypter

Workshop

KÖNIGSWEG

Alexander C. S. Hendorf

ah@koenigsweg.com
@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

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.

KÖNIGSWEG

Introduction to Data Analytics with Pandas and Juypter

Jupyter

- Ecosystem
- Benefits of Jupyter
- Jupyter Notebooks

Pandas

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

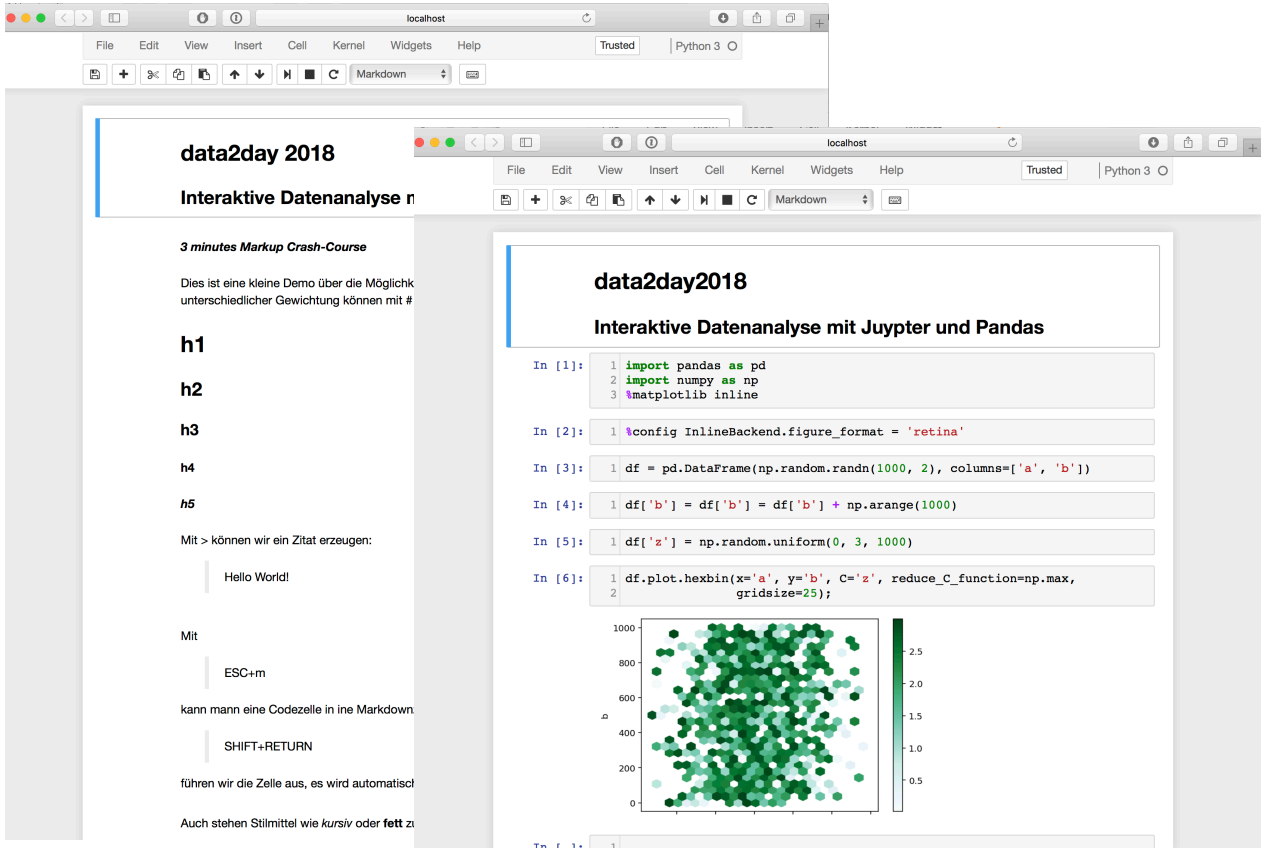
iPython

```
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.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

KÖNIGSWEG

Hands on



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,...)

Fun Fact

–

„Anyone can learn Python,

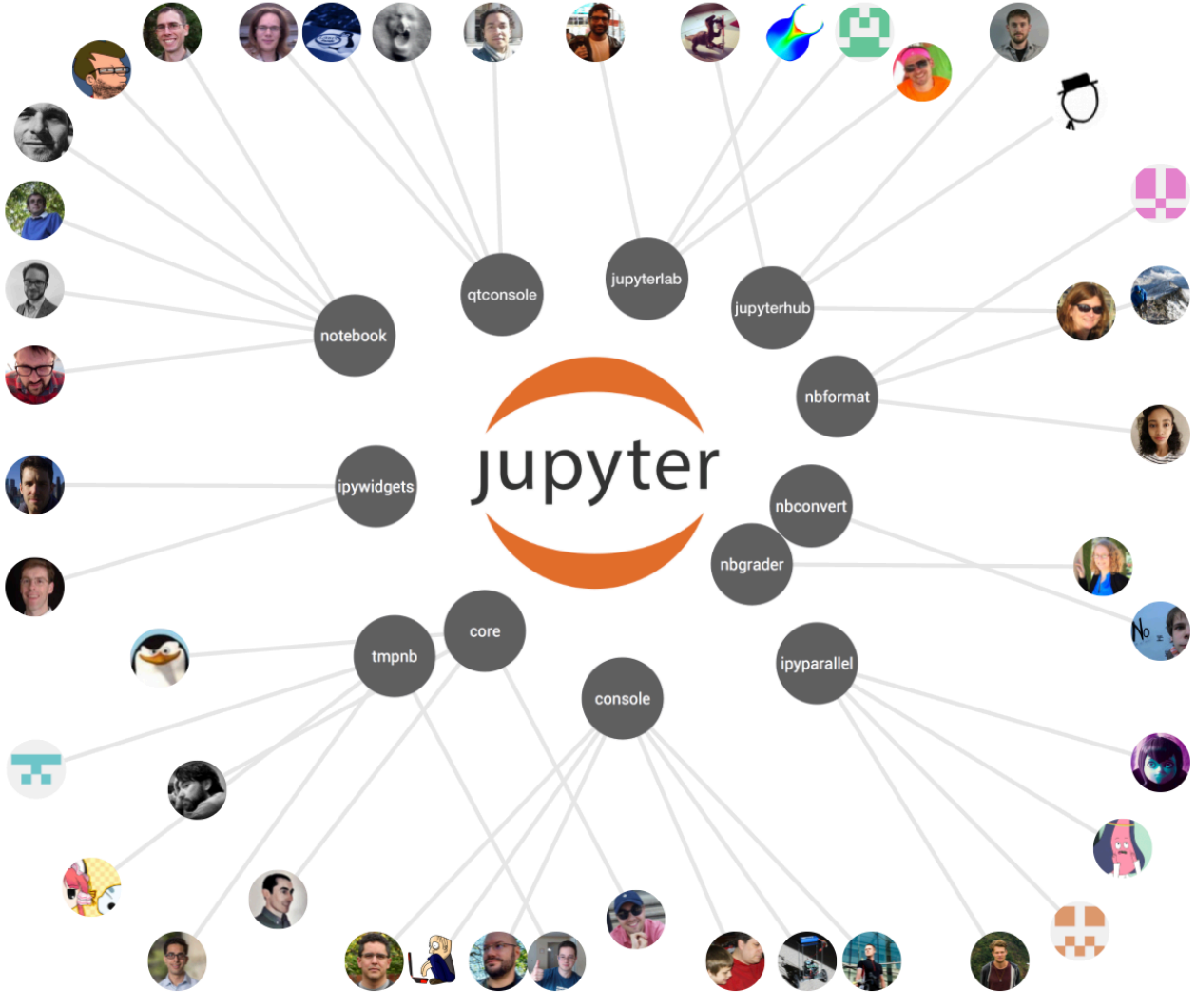
at least for Data Analytics.“

Business Fact

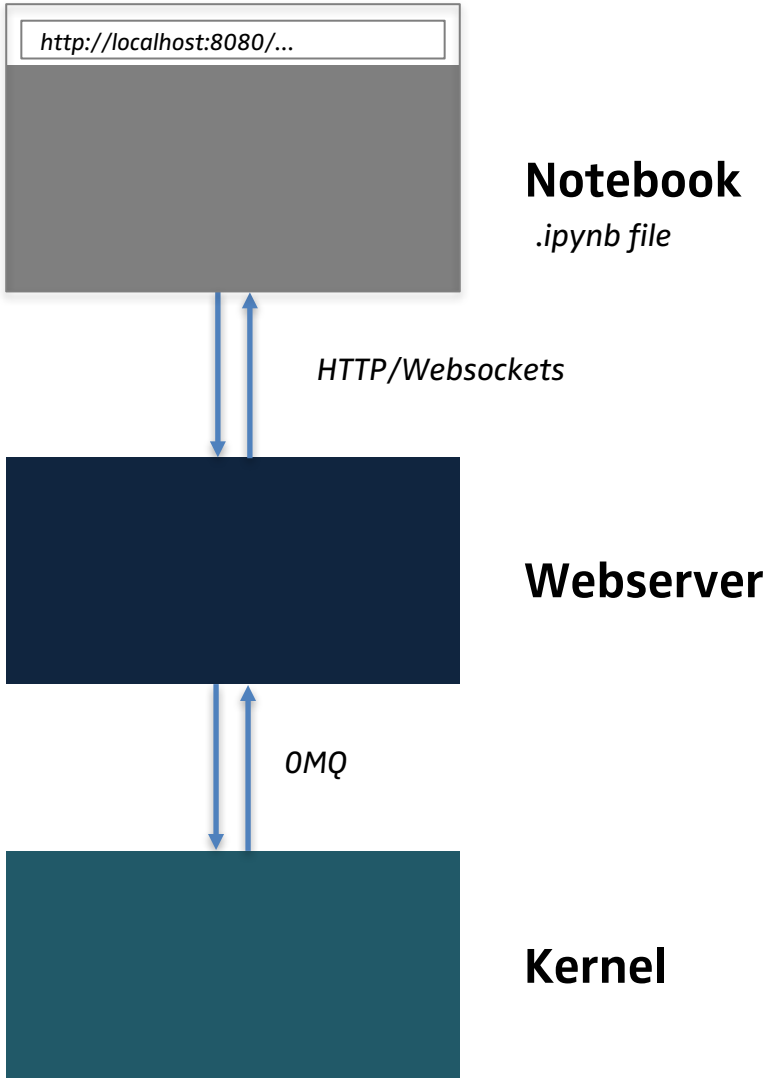
„Python is a perfect common language for a herterogeneous group.“

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
tmpnb
core
console
ipyparallel

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

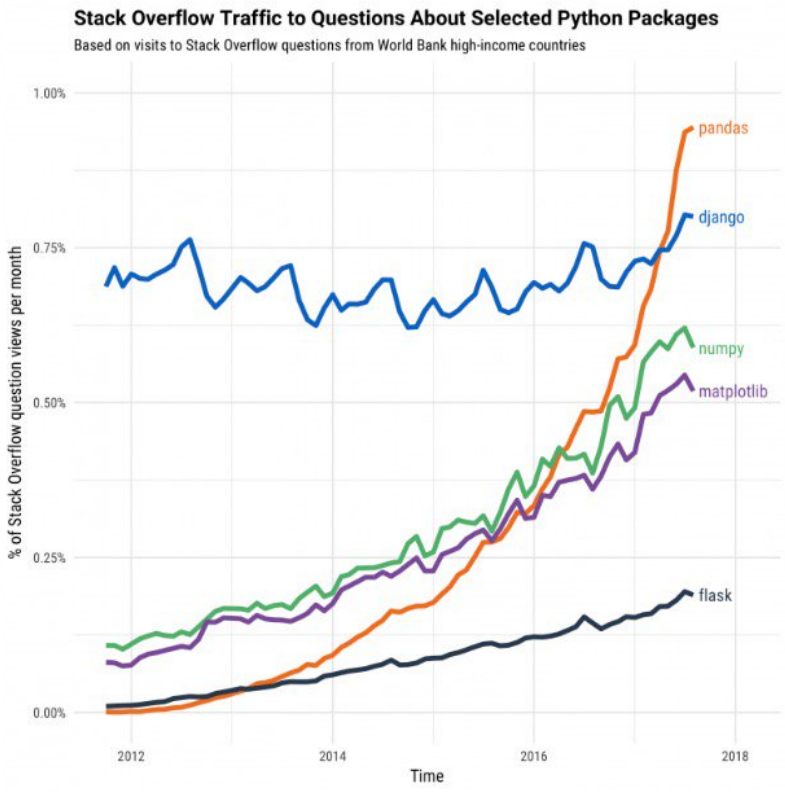
Anaconda Distribution

- 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

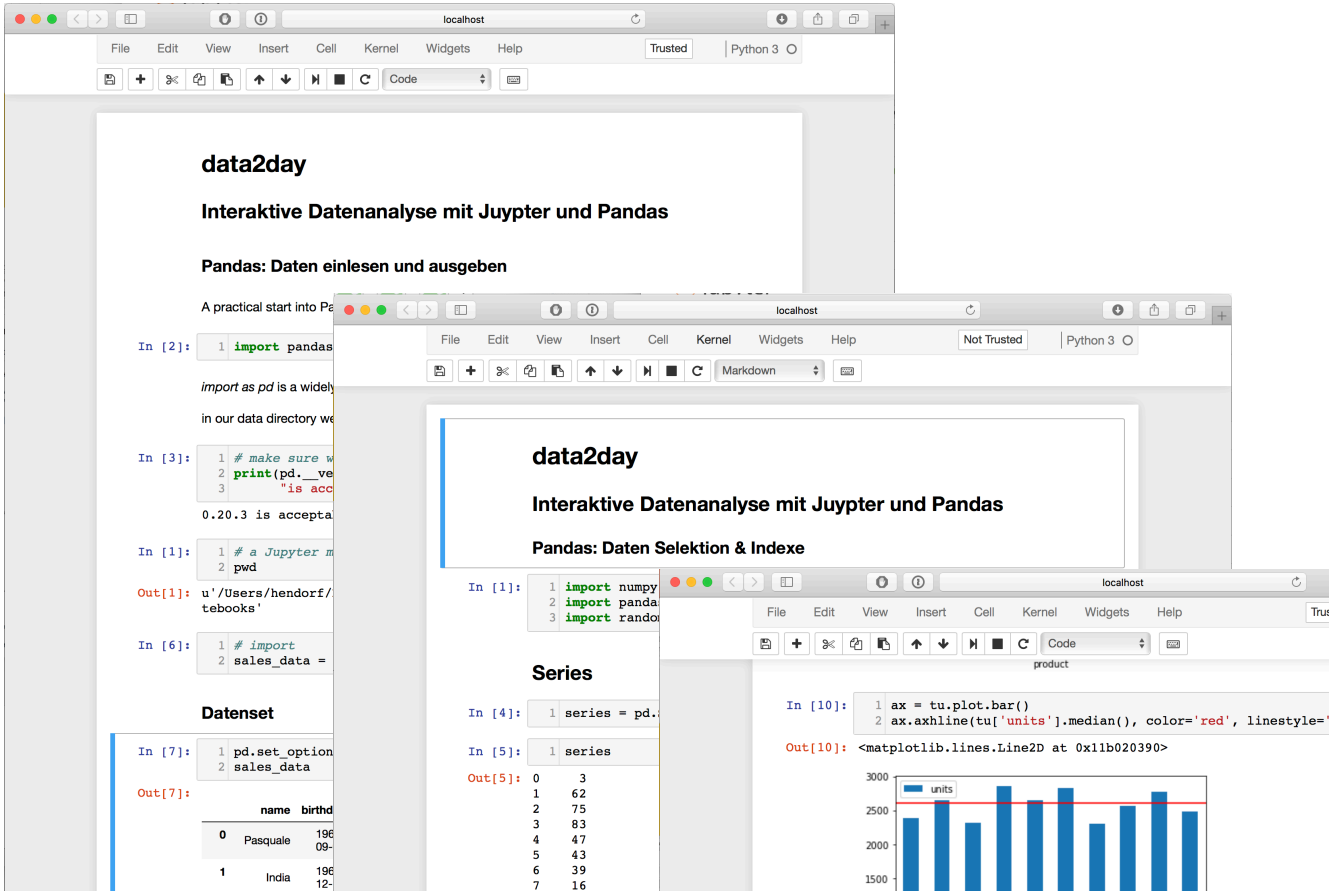
- Support for CSV, Excel, JSON, SQL, SAS, clipboard, HDF5,...
- Data cleansing
- Re-shape & merge data (joins & merge) & pivoting
- Data Visualisation
- Well integrated in Jupyter (iPython) notebooks
- Database-like operations
- Performant

NumPy under the hood

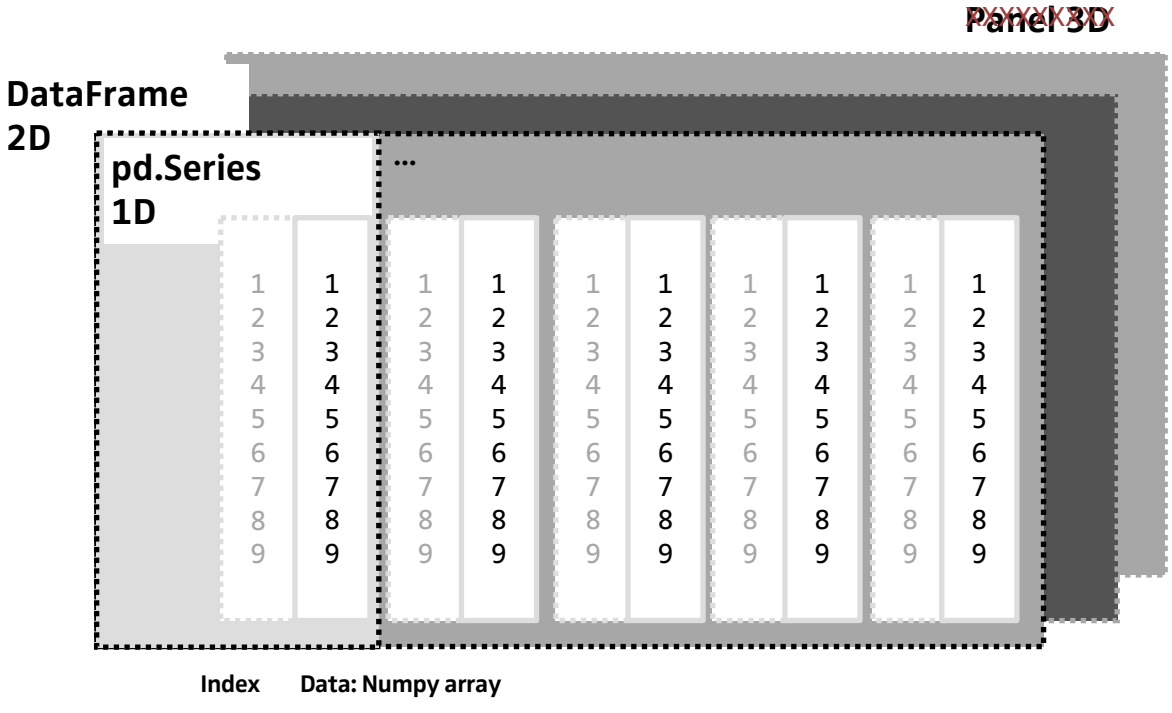
- Library for numerical operations in Python
- Typed Arrays
- Broadcasting

KÖNIGSWEG

Hands on



Strukur



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

Pandas Performance, Limits & Solutions

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

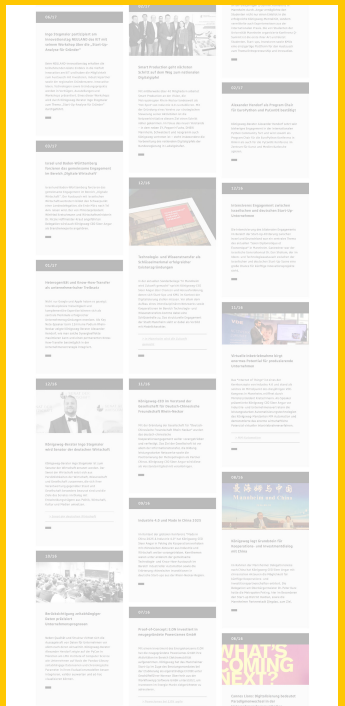
Jupyter Hub

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

KÖNIGSWEG

Thank you!

Q & A



koenigsweg.com

ah@koenigsweg.com

@hendorf