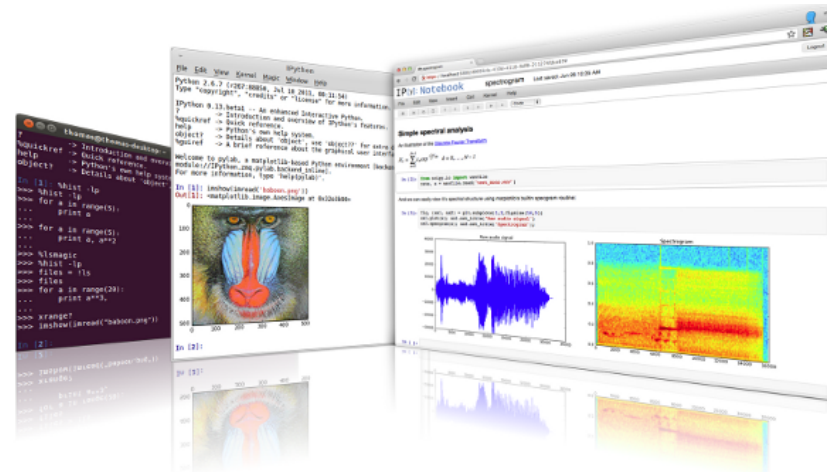




Datenanalyse mit IPython und Pandas

Dirk Loss, 2013-05-02

**„Data Scientist is now
the hottest job title in Silicon Valley.“**



Tim O'Reilly

Daten

Fokus: „Small/Medium Data“

- Performance-Messungen
- Logdateien
- Netzwerktraffic
- Source-Code Repositories
- ...

Pandas



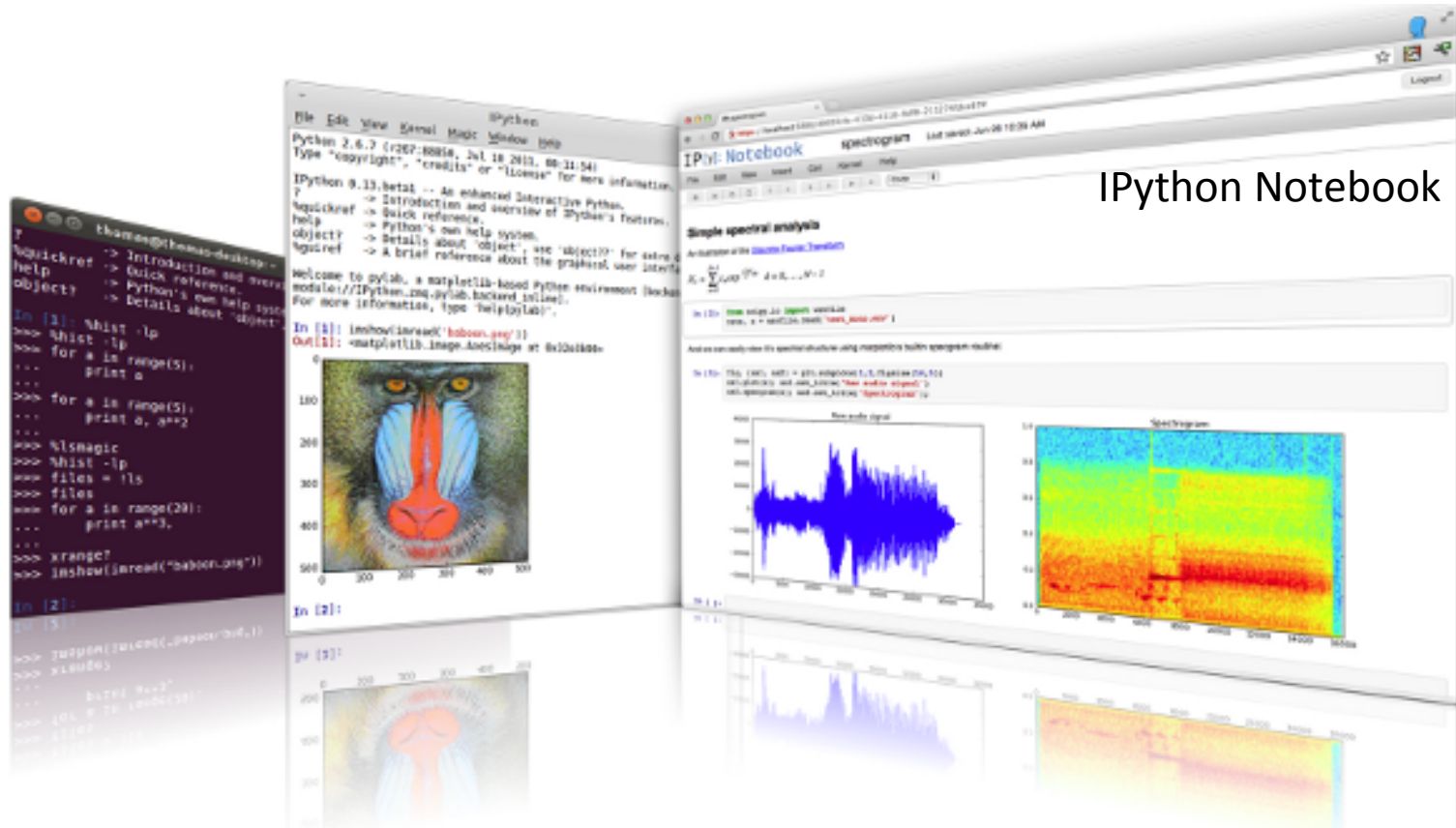
„Manipulation und Analyse mehrdimensionaler Daten“

	A	B	C
2013-05-01	4	foo	0.40
2013-05-02	6	bar	1.23
2013-05-03	3	boost	0.20
2013-05-04	3	100	-6.20

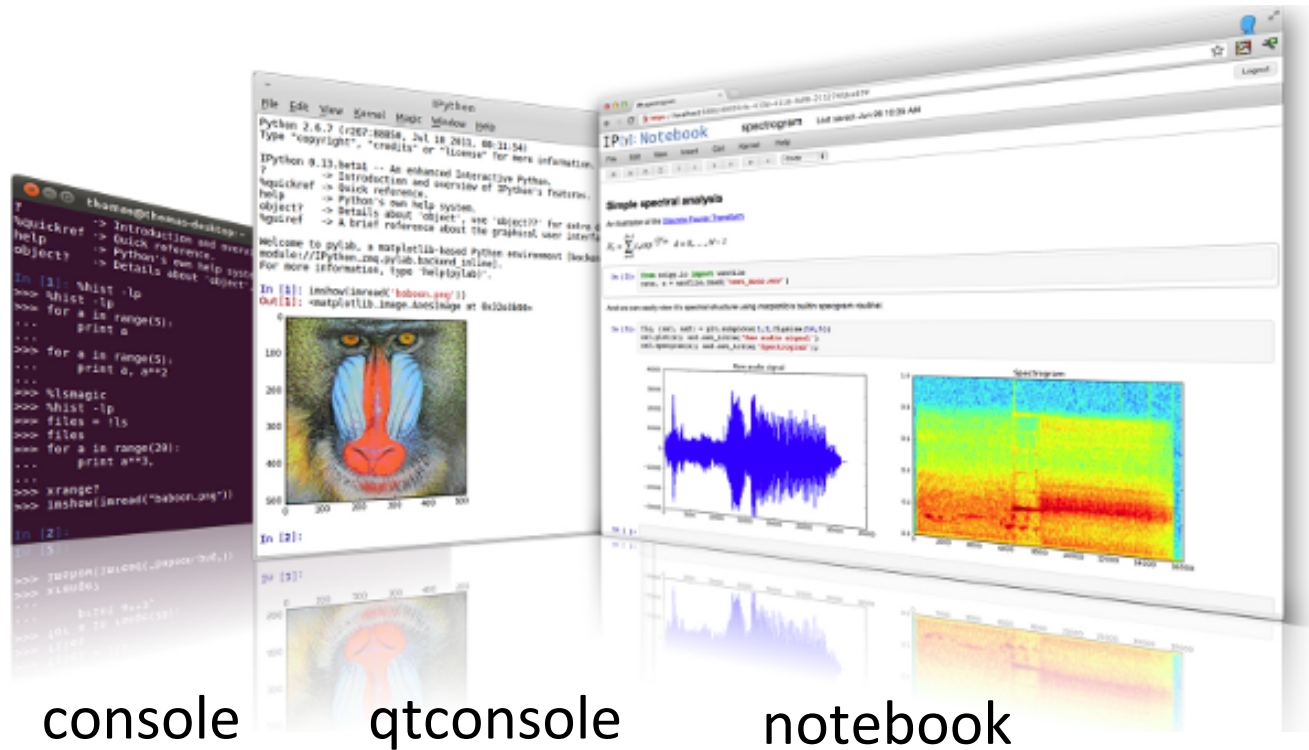
Time series Statistics Aggregation
Missing values Hierarchical indexes

IPython

„Interaktive Python-Arbeitsumgebung“



<http://ipython.org>



console

qtconsole

notebook

IPython

ZeroMQ

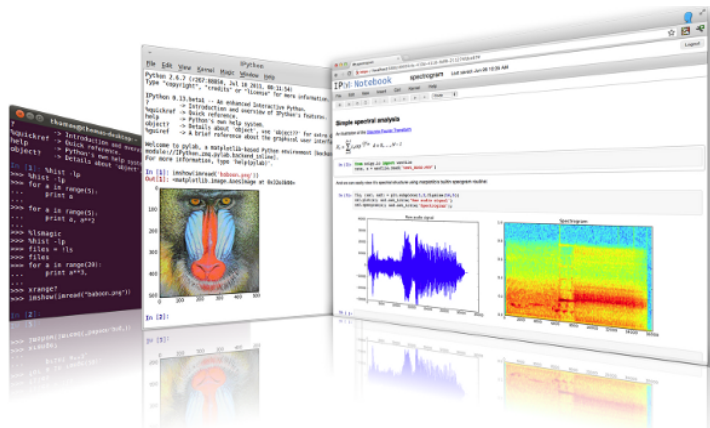
kernel



python™

Demo

IPython



IPython Notebook Viewer

A Simple way to share your IP[y]thon Notebook as Gists.

Share your own notebook, or browse others'

Go!

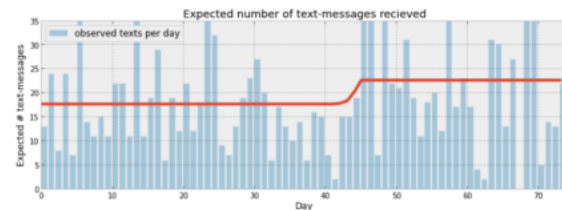
Probabilistic Programming

Why would I want samples from the posterior, anyways?

We will deal with this question for the remainder of the book, and it is an understatement to say we can perform amazingly useful things. For now, let's finishing with using posterior samples to answer the follow question: what is the expected number of texts at day t , $0 \leq t \leq 70$? Recall that the expected value of a Poisson is equal to its parameter λ , then the question is equivalent to what is the expected value of λ at time t ?

In the code below, we are calculating the following: Let i index a particular sample from the posterior distributions. Given a day t , we average over all λ_i on that day t , using $\lambda_{i,t}$ if $t < \tau_i$ else we use $\lambda_{2,t}$.

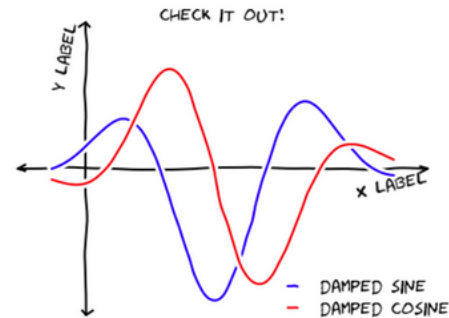
<matplotlib.legend.Legend at 0x1498ba20>



XKCD Plot With Matplotlib

XKCD plots in Matplotlib

Out [1]:



Sometimes when showing schematic plots, this is the type of figure I want to display. But drawing it by hand is a π matplotlib. The problem is, matplotlib is a bit too precise. Attempting to duplicate this figure in matplotlib leads to:

<http://nbviewer.ipython.org>

IP[y]: IPython

Interactive Computing

[Install](#) · [Docs](#) · [Videos](#) · [Notebook Viewer](#) · [News](#) · [Cite](#) · [Donate](#)

Sloan Foundation Grant

We are pleased to announce that the IPython project has received a \$1.15M grant from the Alfred P. Sloan foundation, that will support IPython development for the next two years (1/1/2013-12/31/2014). The grant, which is being made to the University of California, Berkeley and California Polytechnic State University, San Luis Obispo, will enable the project to focus on developing the IPython Notebook as a general tool for scientific and technical computing that is open, collaborative and reproducible.

Google™ Custom Search ×

MEMORIAL

John Hunter

1968–2012

J. Hunter Memorial Fund

VERSIONS

Stable

0.13.2 – April 2013

<http://ipython.org/sloan-grant.html>

Demo

Pandas

	A	B	C
2013-05-01	4	foo	0.40
2013-05-02	6	bar	1.23
2013-05-03	3	boost	0.20
2013-05-04	3	100	-6.20

Pandas DataFrame

The diagram illustrates the structure of a Pandas DataFrame. It features a header row for columns and a header column for the index. Arrows indicate the mapping from these headers to the data cells.

columns		foo	bar	baz	qux
index	A	0	x	2.7	True
	B	4	y	6	True
	C	8	z	10	False
	D	-12	w	NA	False
	E	16	a	18	False



Indexing in Python

s =

0	1	2	3	4	5	6	7	8	9
H	a	l	l	o		W	e	l	t
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1

s[0] == "H"

s[4] == "o"

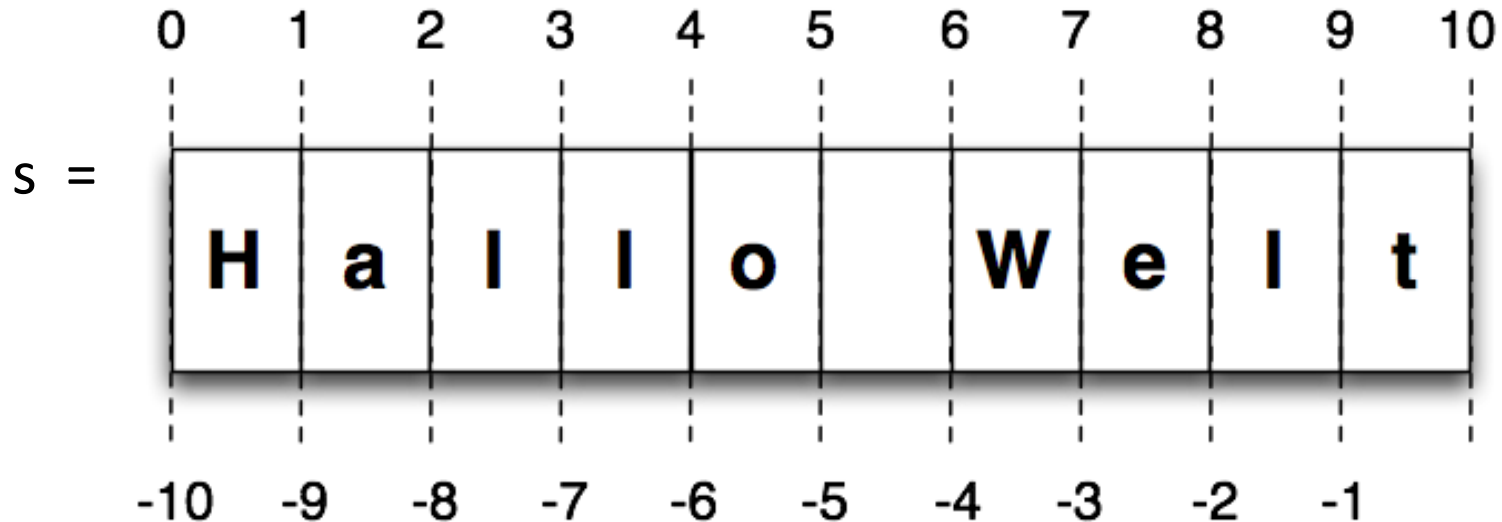
s[-1] == "t"

s[41] => IndexError

s[-12] => IndexError



Slicing in Python



`s[2:5] == "llo"`

`s[4:-5] == "o"`

`s[6:] == "Welt"`

`s[6:10000] == "Welt"`

`s[-3:] == "elt"`

`s[5:2] == ""`



IPython



SymPy



Weiterführende Infos

Agile Tools for Real World Data

Python for Data Analysis



O'REILLY®

Wes McKinney

IPython Videos

IPy Videos and Screencasts — x

ipython.org/videos.html

☆ 🔔 🗨️ 📄 ☰

IP[y]: IPython

Interactive Computing

[Install](#) · [Docs](#) · [Videos](#) · [Notebook Viewer](#) · [News](#) · [Cite](#) · [Donate](#)

Videos and Screencasts

Science And Python: retrospective of a (mostly) successful decade

A historical view of the co-evolution of IPython and the scientific Python stack (1h), that has been very well received. Delivered by [Fernando Perez](#) as a keynote presentation at the PyCon Canada 2012 conference in Toronto, it contains multiple demos of the workflows that IPython enables ([PDF slides](#)).

Science And Python: retrospective of a (mostly) successful decade



0:00:11 / 1:00:58

Google™ Custom Search

MEMORIAL

John Hunter
1968–2012
[J. Hunter Memorial Fund](#)

VERSIONS

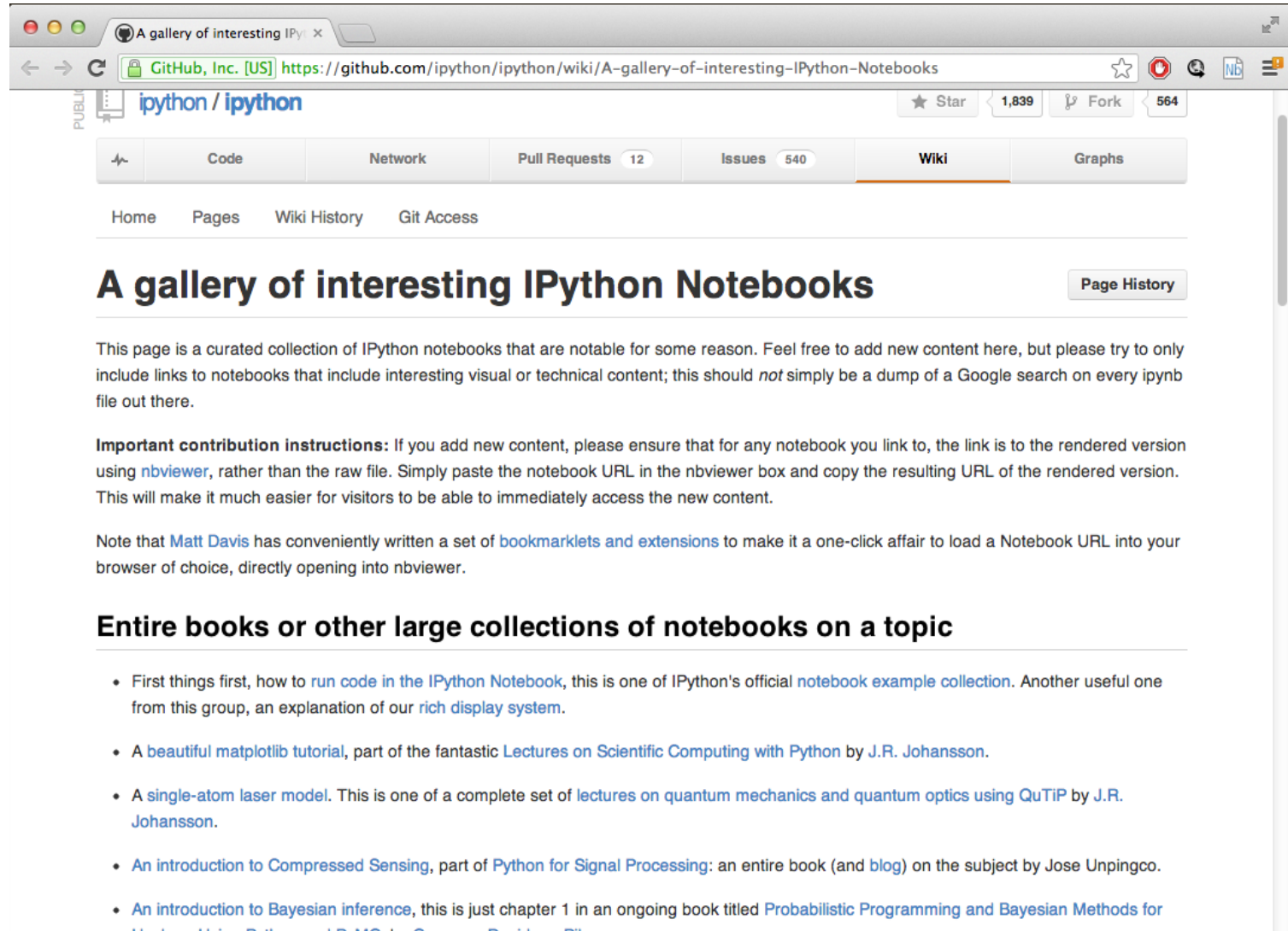
Stable
0.13.2 – April 2013
[Download](#)

Development
1.0.dev
[Github](#)

COMMUNITY

[Help Chat Room](#)
[Stack Overflow](#)
[Mailing list](#)
[Wiki](#)
[File a bug](#)

IPython Beispiel-Notebooks



The screenshot shows a web browser window displaying the GitHub repository page for "A gallery of interesting IPython Notebooks". The browser's address bar shows the URL: <https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks>. The repository name "ipython / ipython" is visible, along with 1,839 stars and 564 forks. The "Wiki" tab is selected, showing the title "A gallery of interesting IPython Notebooks" and a "Page History" button. The page content includes a description of the gallery, contribution instructions, and a list of recommended notebooks.

A gallery of interesting IPython Notebooks Page History

This page is a curated collection of IPython notebooks that are notable for some reason. Feel free to add new content here, but please try to only include links to notebooks that include interesting visual or technical content; this should *not* simply be a dump of a Google search on every ipynb file out there.

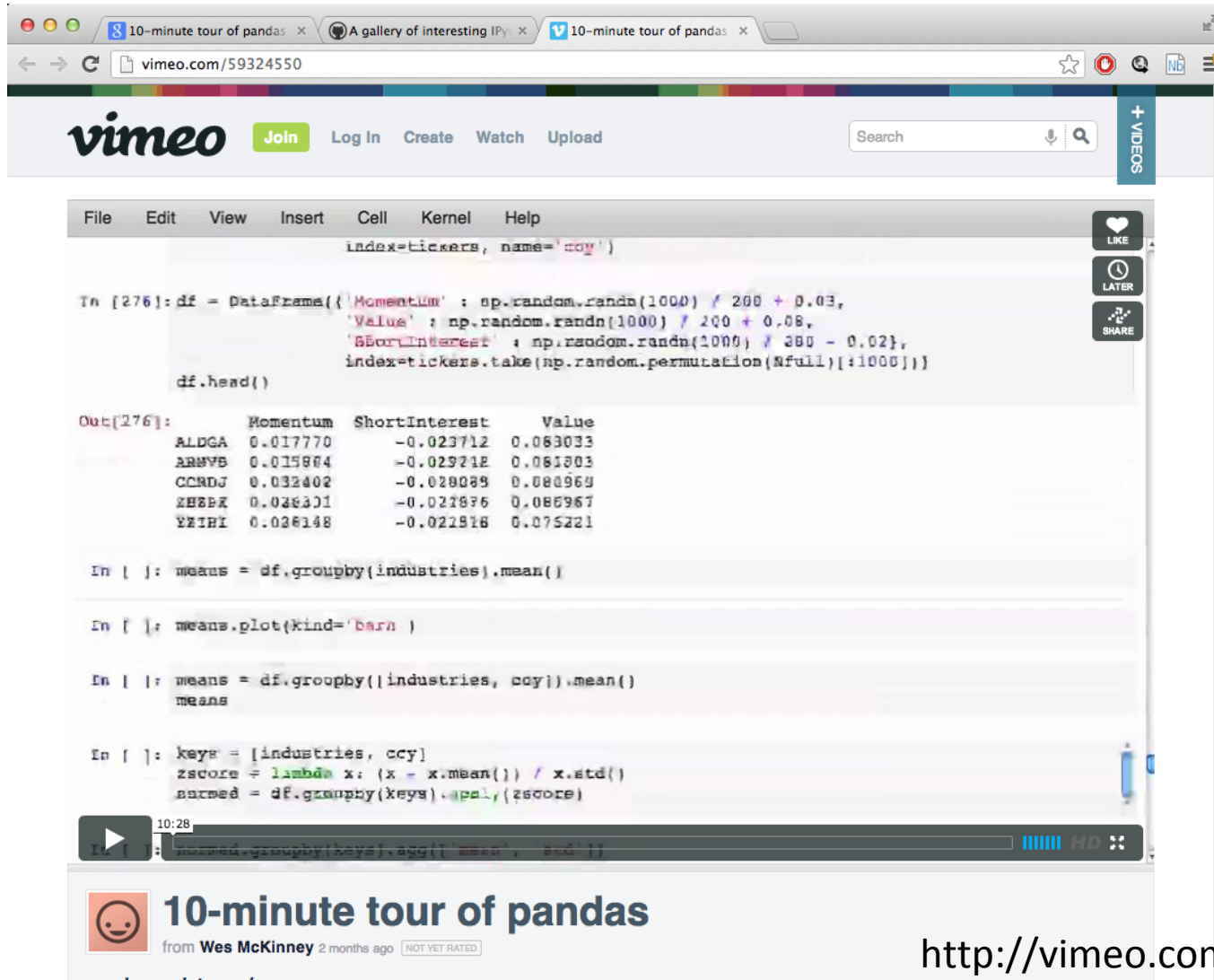
Important contribution instructions: If you add new content, please ensure that for any notebook you link to, the link is to the rendered version using [nbviewer](#), rather than the raw file. Simply paste the notebook URL in the nbviewer box and copy the resulting URL of the rendered version. This will make it much easier for visitors to be able to immediately access the new content.

Note that [Matt Davis](#) has conveniently written a set of [bookmarklets and extensions](#) to make it a one-click affair to load a Notebook URL into your browser of choice, directly opening into nbviewer.

Entire books or other large collections of notebooks on a topic

- First things first, how to [run code in the IPython Notebook](#), this is one of IPython's official [notebook example collection](#). Another useful one from this group, an explanation of our [rich display system](#).
- A [beautiful matplotlib tutorial](#), part of the fantastic [Lectures on Scientific Computing with Python](#) by J.R. Johansson.
- A [single-atom laser model](#). This is one of a complete set of [lectures on quantum mechanics and quantum optics using QuTiP](#) by J.R. Johansson.
- [An introduction to Compressed Sensing](#), part of [Python for Signal Processing](#): an entire book (and [blog](#)) on the subject by Jose Unpingco.
- [An introduction to Bayesian inference](#), this is just chapter 1 in an ongoing book titled [Probabilistic Programming and Bayesian Methods for Hackers Using Python and PyMC](#) by Cameron Davidson Pilon.

10-Minuten pandas Überblick



The screenshot shows a web browser window with the address bar displaying `vimeo.com/59324550`. The page features the Vimeo logo and navigation links: `Join`, `Log In`, `Create`, `Watch`, and `Upload`. A search bar is located on the right side of the header.

The video content displays a Jupyter Notebook interface with the following code and output:

```
index=tickers, name='ccy')

In [276]: df = DataFrame({'Momentum': np.random.randn(1000) / 200 + 0.03,
                          'Value': np.random.randn(1000) / 200 + 0.08,
                          'ShortInterest': np.random.randn(1000) / 200 - 0.02},
                          index=tickers.take(np.random.permutation(Nfull)[:1000]))

df.head()
```

Out[276]:

	Momentum	ShortInterest	Value
ALDGA	0.017770	-0.023712	0.083033
ARMV5	0.015864	-0.023212	0.081303
CCRDJ	0.032402	-0.028039	0.080969
ZHSEA	0.028301	-0.021876	0.085967
YZIHI	0.026148	-0.021816	0.075221

```
In [ ]: means = df.groupby(industries).mean()

In [ ]: means.plot(kind='bar')

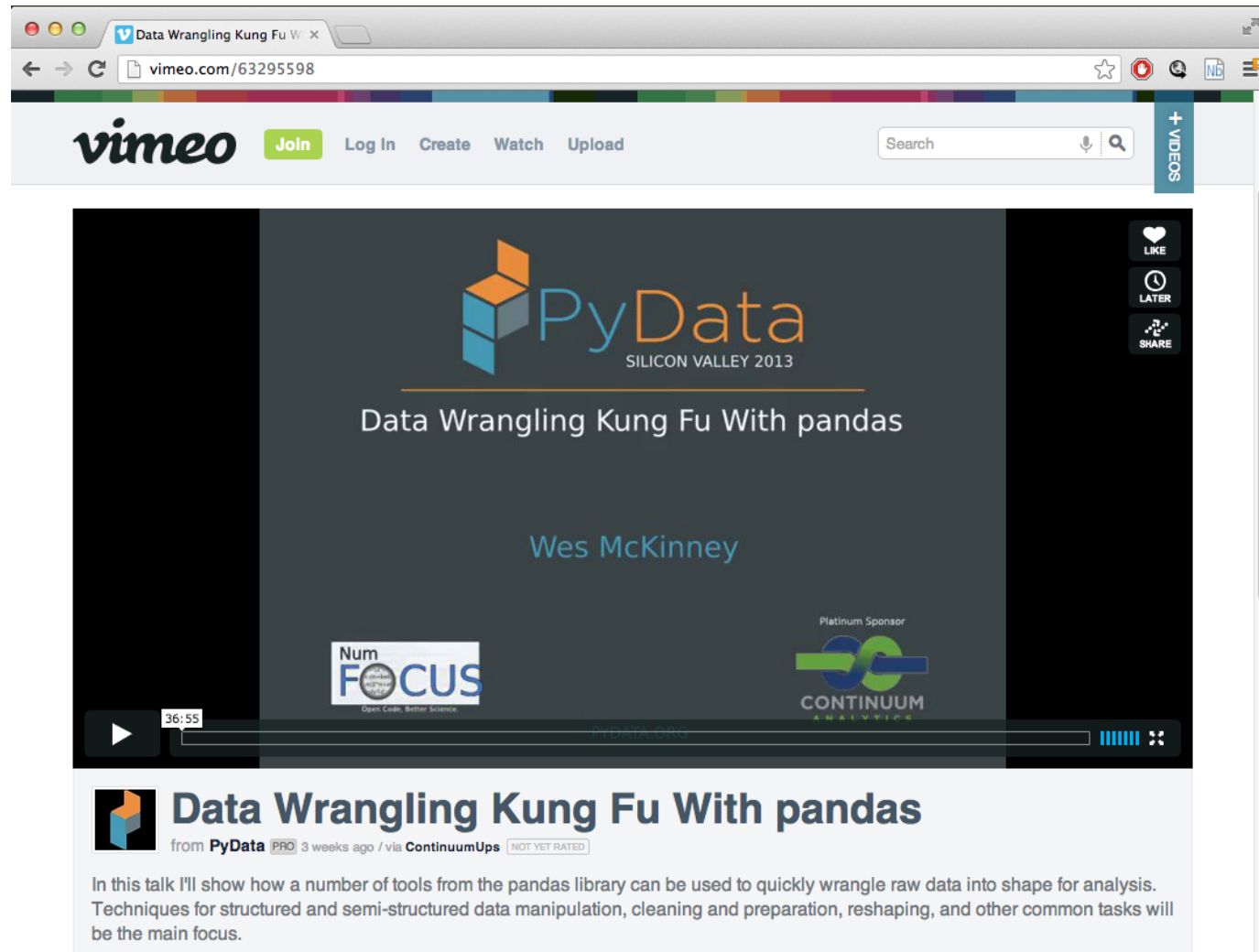
In [ ]: means = df.groupby([industries, ccy]).mean()
means

In [ ]: keys = [industries, ccy]
zscore = lambda x: (x - x.mean()) / x.std()
sorted = df.groupby(keys).apply(zscore)

In [ ]: sorted.groupby(keys).agg(['mean', 'std'])
```

The video player interface includes a play button, a progress bar showing 10:28, and a volume control. The video title is "10-minute tour of pandas" by Wes McKinney, posted 2 months ago. The URL `http://vimeo.com/59324550` is displayed at the bottom right.

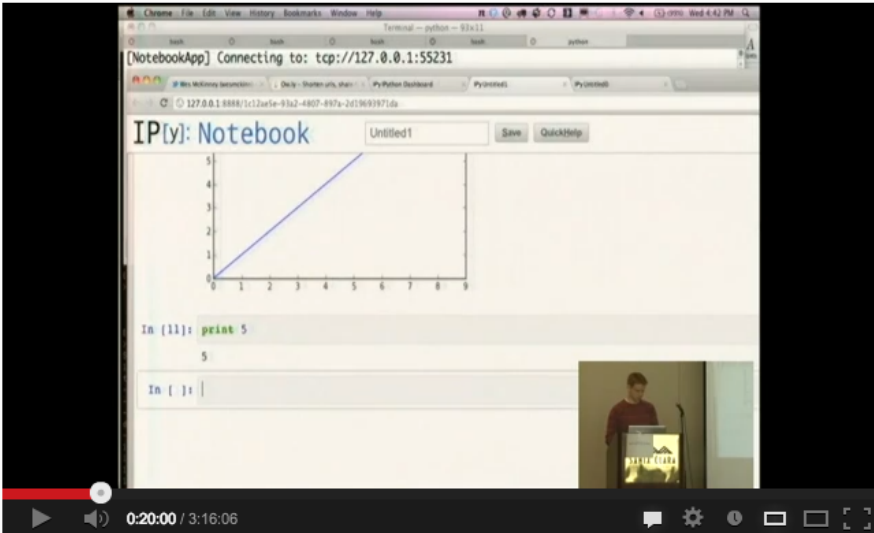
35-Minuten pandas Vortrag



The screenshot shows a web browser window with the address bar displaying `vimeo.com/63295598`. The Vimeo website header is visible, including the logo, navigation links (Join, Log In, Create, Watch, Upload), a search bar, and a '+ VIDEOS' button. The video player itself shows a dark background with the 'PyData SILICON VALLEY 2013' logo at the top. Below the logo, the title 'Data Wrangling Kung Fu With pandas' is displayed in white, followed by the speaker's name 'Wes McKinney' in blue. At the bottom of the video frame, there are logos for 'NumFOCUS' and 'CONTINUUM ANALYTICS'. The video player controls at the bottom show a play button, a progress bar at 36:55, and a volume icon. Below the video player, the video's title 'Data Wrangling Kung Fu With pandas' is repeated, along with its source 'from PyData PRO 3 weeks ago / via ContinuumUps' and a 'NOT YET RATED' badge. A short description follows: 'In this talk I'll show how a number of tools from the pandas library can be used to quickly wrangle raw data into shape for analysis. Techniques for structured and semi-structured data manipulation, cleaning and preparation, reshaping, and other common tasks will be the main focus.'

<http://vimeo.com/63295598>

3-Stunden IPython und Pandas



Data analysis in Python with pandas

NextDayVideo · 521 Videos

29.413

Abonnieren 5.969

Mag ich

Info Teilen Hinzufügen

Veröffentlicht am 09.03.2012

Wes McKinney

The tutorial will give a hands-on introduction to manipulating and analyzing large and small structured data sets in Python using the pandas library. While the focus will be on learning the nuts and bolts of the library's features, I als

Bjarne Stroustrup: The 5 Programming Languages You Need
von bigthink
119.100
ANGESAGT

What makes Python so AWESOME
von PythonItalia
23.805 Aufrufe

Boston Algorithmic Finance Meetup with Wes McKinney
von Dan Dunn
724 Aufrufe

Python in Big Data
von MarakanaTechTV
14.299 Aufrufe

Guido van Rossum on the History of Python
von droptalks
19.002 Aufrufe

Programming
von igoronline

IPython: Python at your fingertips
von NextDayVideo
45.620 Aufrufe

Tutorial: scikit-learn - Machine Learning in Python with Contributor
von MarakanaTechTV
11.350 Aufrufe

<http://www.youtube.com/watch?v=w26x-z-BdWQ>

Fazit

- IPython als Arbeitsumgebung
(nicht nur für Python)
- Pandas für die Datenanalyse
(insbesondere für Zeitreihen)

Dokumentation

ipython.org

pandas.pydata.org

nbviewer.ipython.org

matplotlib.org

sympy.org

Installation unter Ubuntu

- `sudo apt-get install
ipython-notebook`
- `sudo apt-get install
python-pandas
python-matplotlib
python-scipy
python-sympy
python-nose`