Big Data and Data Science in the Browser

Global Big Data Conference
Santa Clara, 01. September 2015

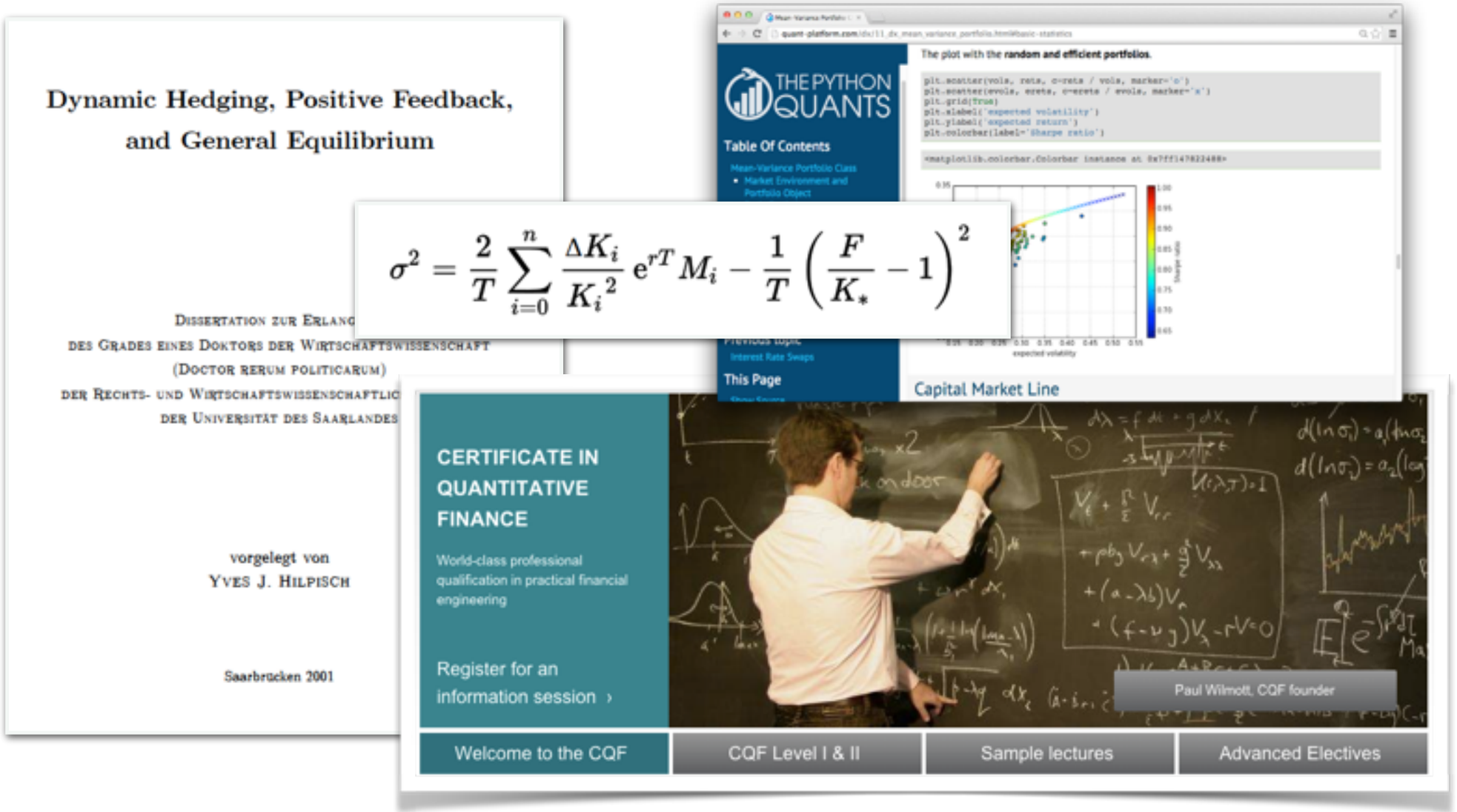
Yves Hilpisch | The Python Quants GmbH

Yves Hilpisch – http://hilpisch.com

Python Entrepreneur

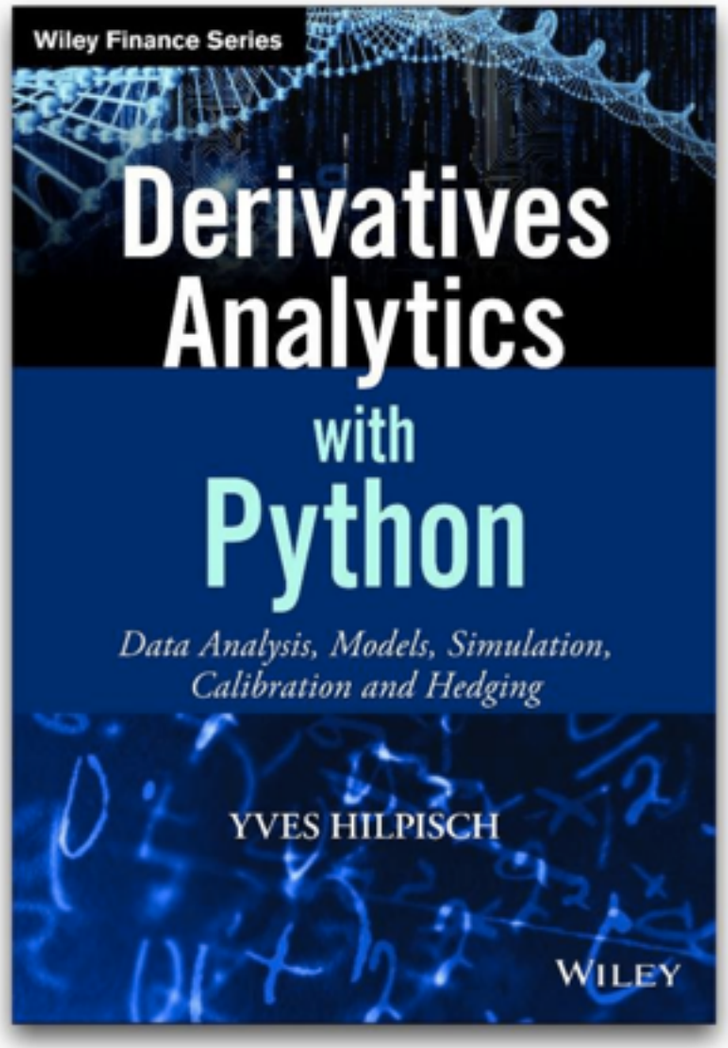
Yves Hilpisch – http://hilpisch.com

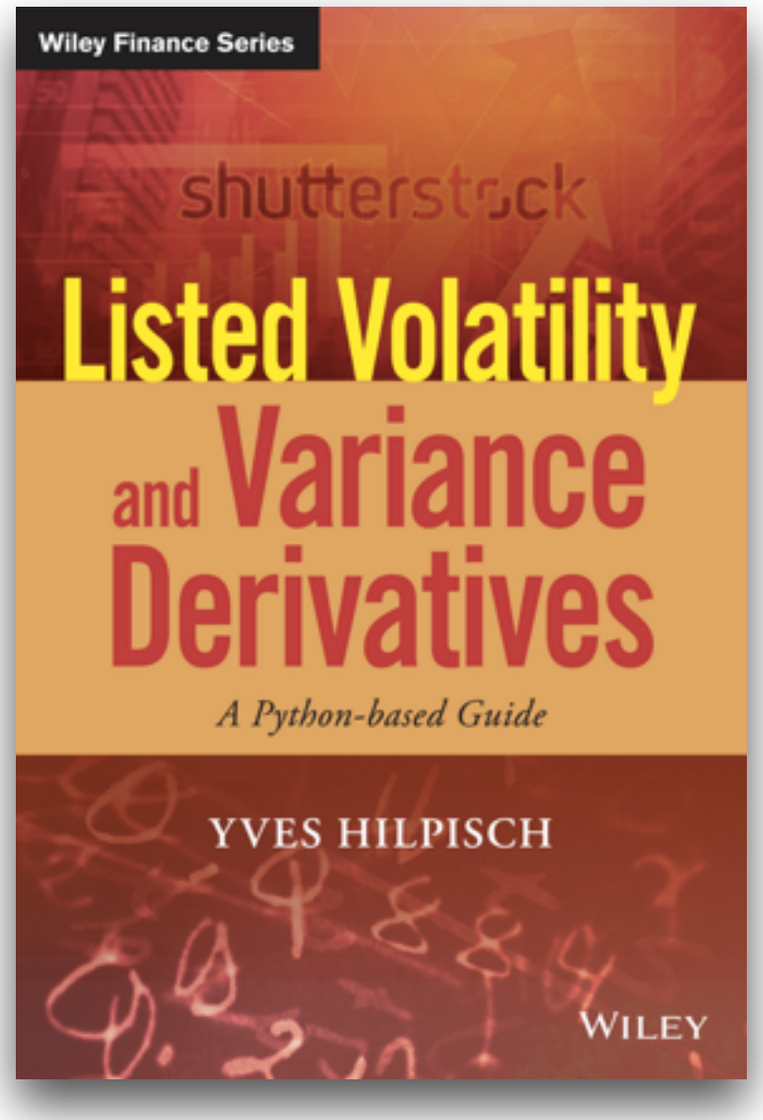
Quant & Lecturer



Yves Hilpisch – http://hilpisch.com

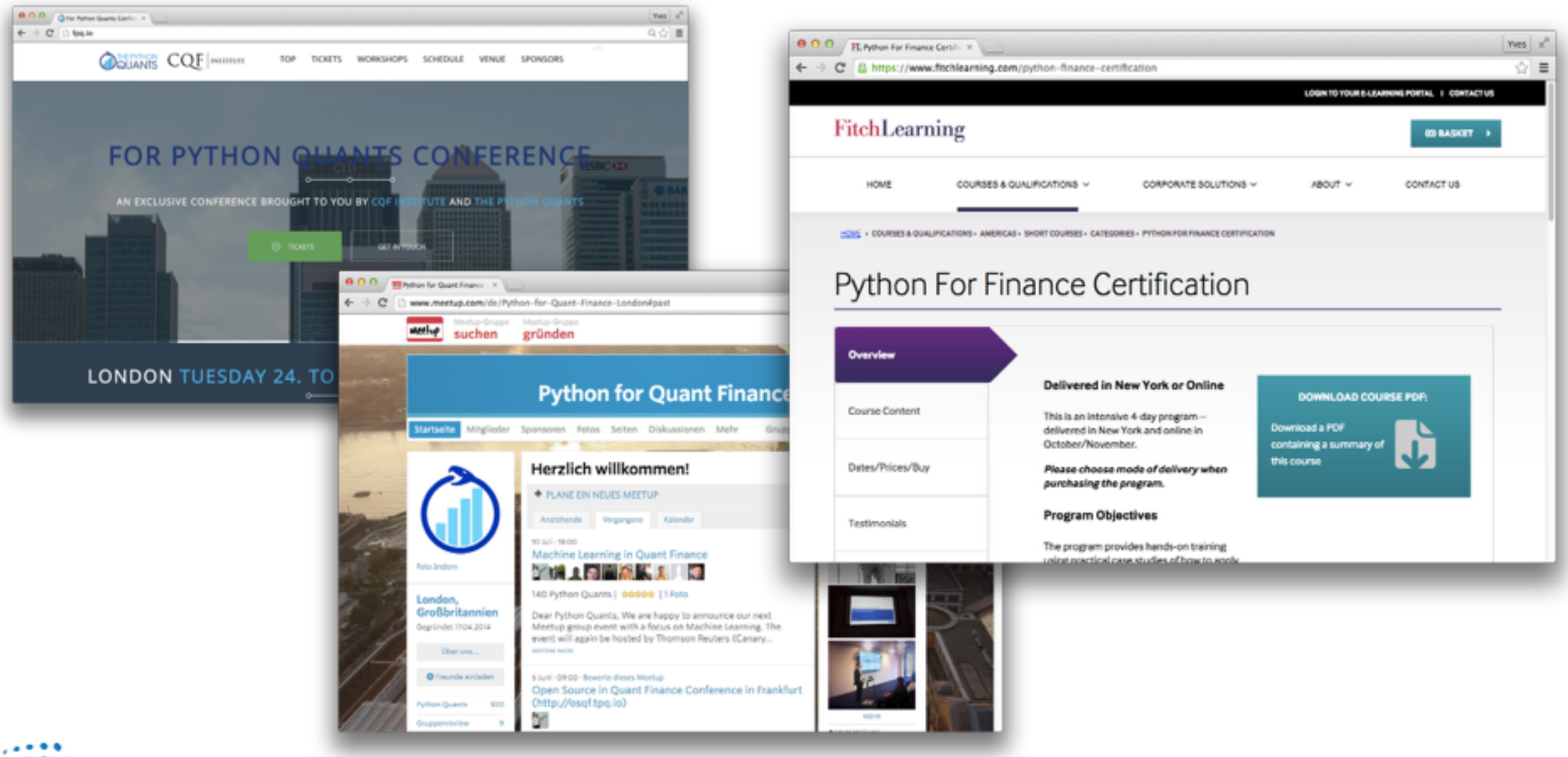
Author





The Python Quants — http://tpq.io

Events, Training & Conferences



I. **Open Source Data Science**
II. Data Science in the Browser
III. Benefits and Use Cases

Data Analytics

Data analytics is a top priority of almost any organisation

"Companies will spend an average of $7.4M on data-related initiatives over the next twelve months , with enterprises investing $13.8M, and small & medium businesses (SMBs) investing $1.6M.

80% of enterprises and 63% of small & medium businesses (SMBs) already have deployed or are planning to deploy big data projects in the next twelve months.

83% of organizations are prioritizing structured data initiatives as critical or high priority in 2015, and 36% planning to increase their budgets for data-driven initiatives in 2015."

Source: http://www.forbes.com

Mega Trends

Mega trends that influence data science

Today's standard is "open source", even for key technologies.

Dynamic communities shape the way knowledge is transmitted

More and more data sets are "open and free".

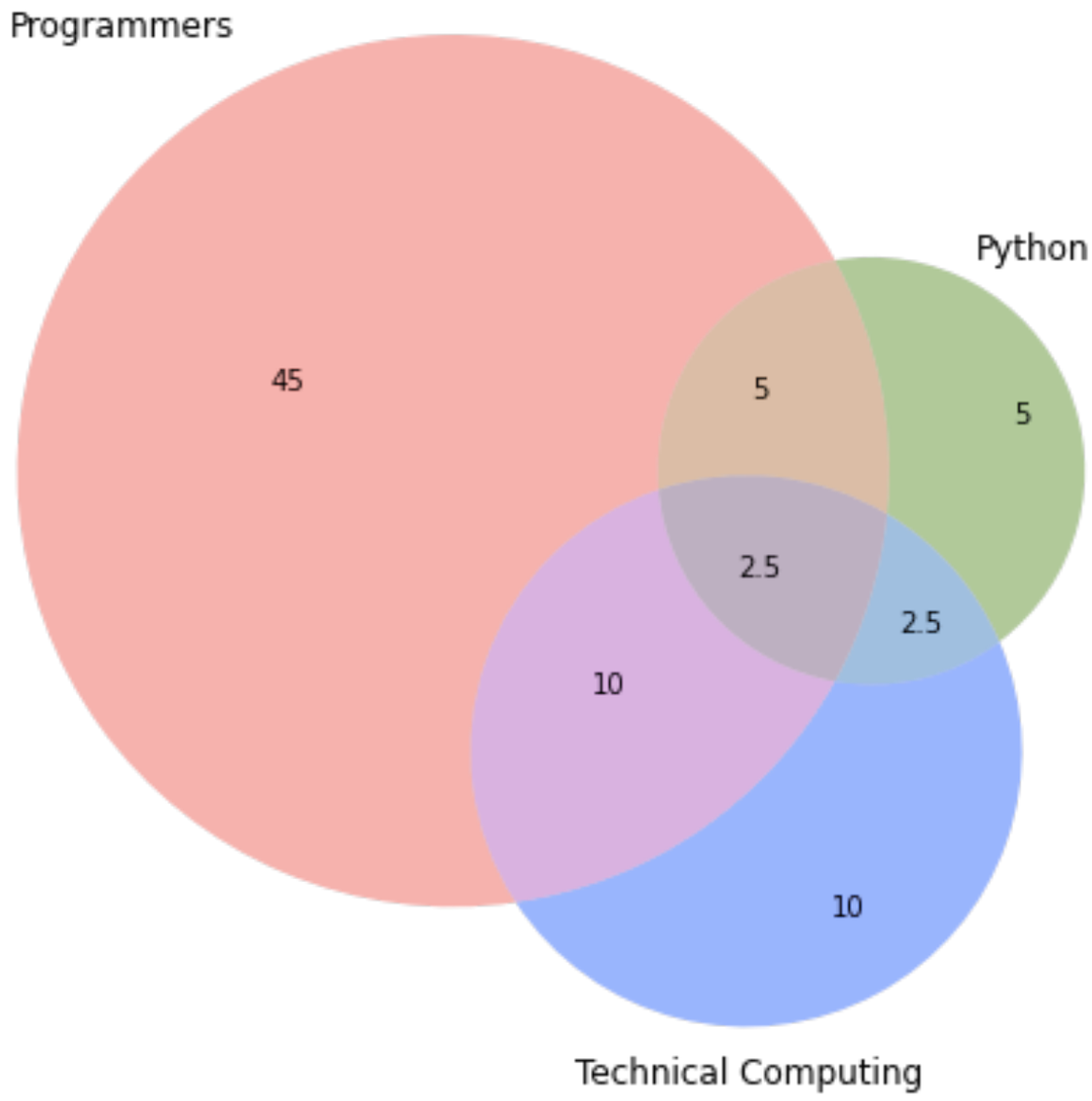
Complex analytics work flows are coded in the browser.

Individuals and institutions store more and more data in the cloud.

Infrastructure is a standardized commodity, billed by the hour.

Data Scientists and Engineers

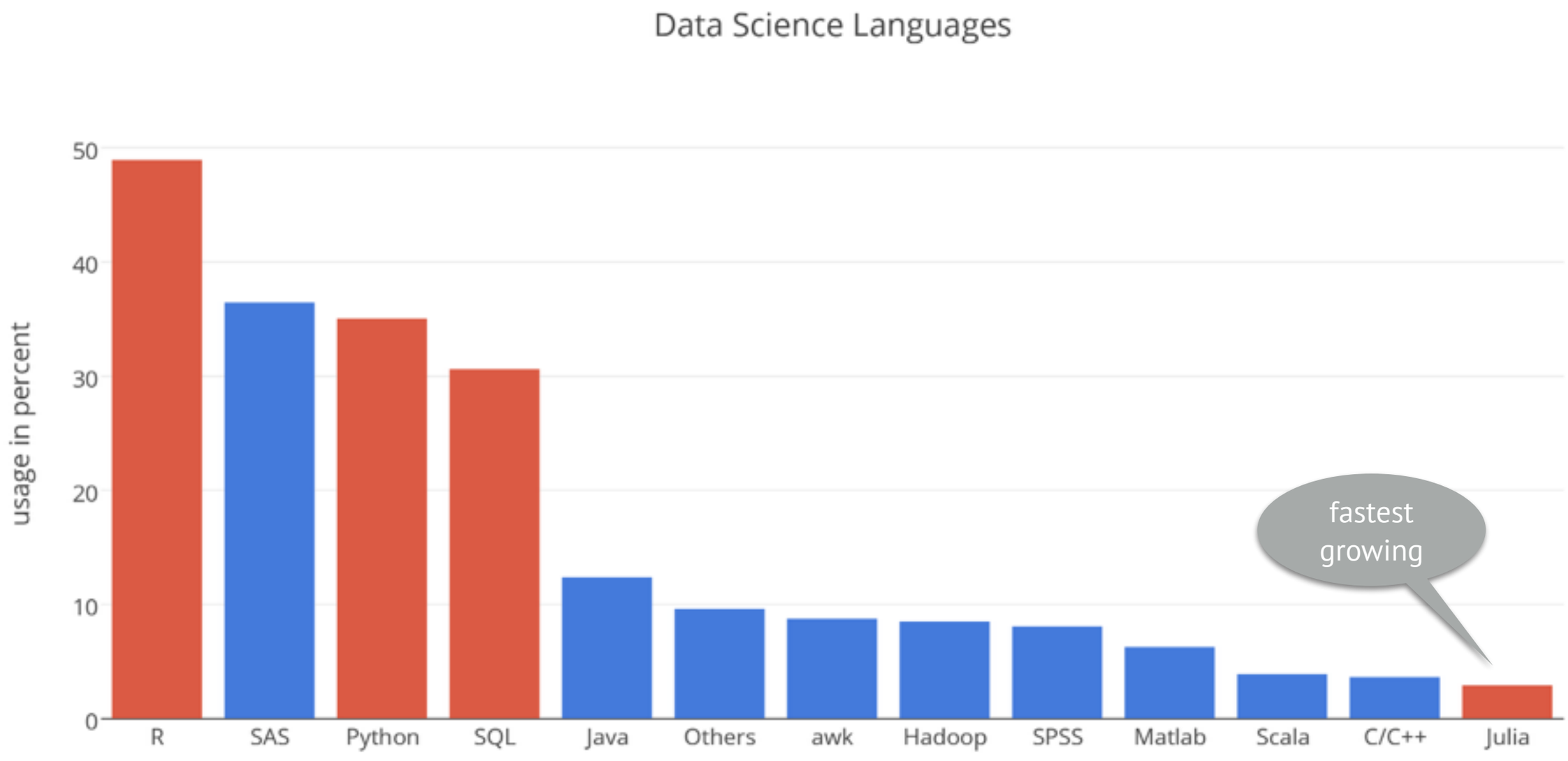
There are about 10mn people in technical computing



Source: diverse Web resources; in mn people

Languages

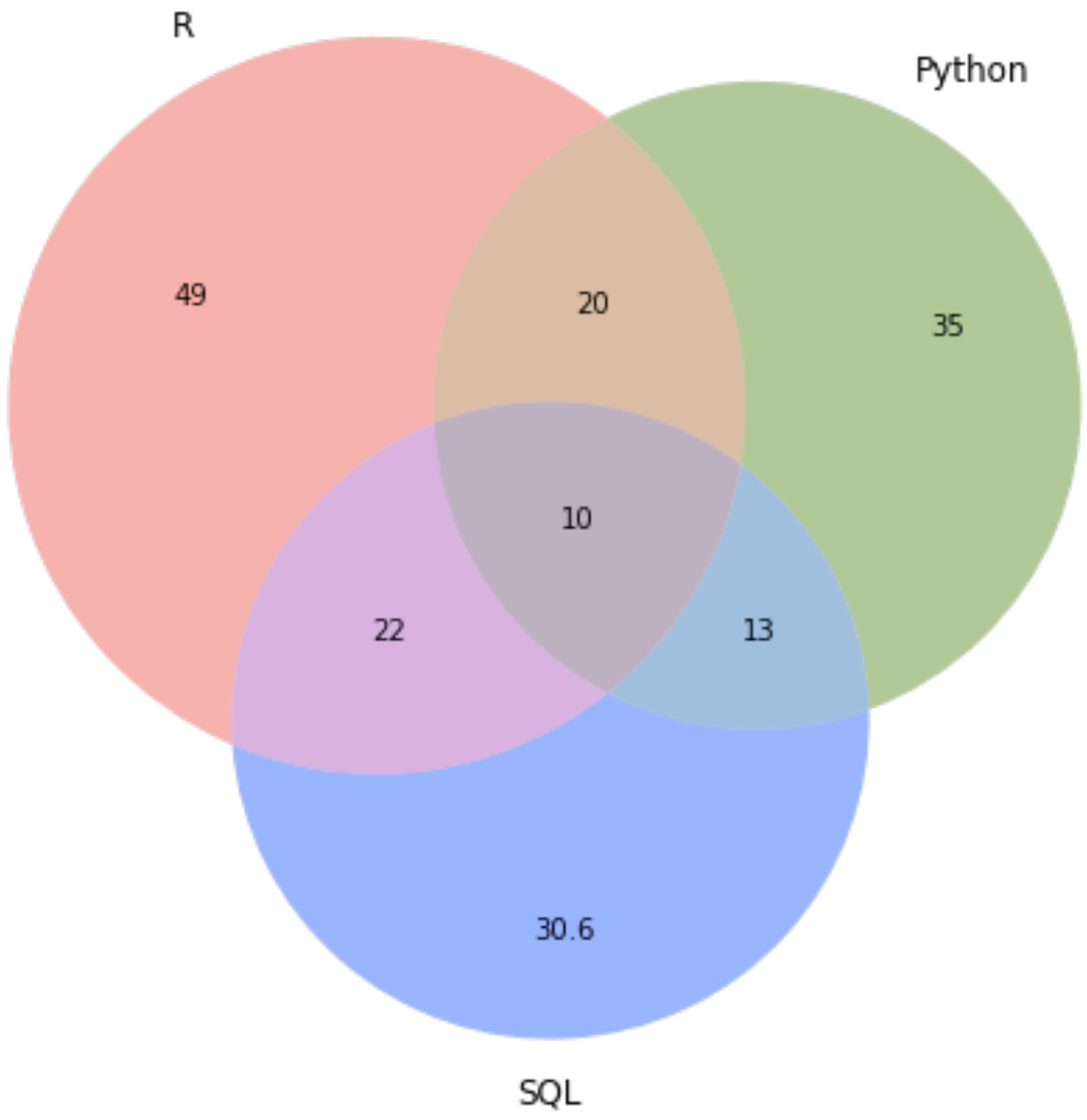
Open Source languages dominate data science these days



Poll data from August 2014. Source: http://www.kdnuggets.com

Multilinguism

One language is hardly ever enough



Poll data from August 2014; usage in %. Source: http://www.kdnuggets.com

The Problem

Obstacles to using open source software for data science

Open Source
fast changing environment

Vendors & Partners
almost no vendors that provide help & support

Libraries
huge amount of libraries to manage

Tools
multitude of useful standalone tools

Deployment
complex, lengthy, costly, risky

Maintenance
how to update, maintain infrastructure?

Diverse End Users
computer & data scientists as well as domain experts

Training
how to train and re-train people?

Start
where and how to start, who to talk to?

I. Open Source Data Science

II. Data Science in the Browser

III. Benefits and Use Cases

The Solution

Open source data science technologies in your browser

The Infrastructure

Delivery based on modern, secure & scalable infrastructure

The Approach

Do not reinvent the wheel

“Absorb what is useful, discard what is not, and add what is uniquely your own.”

—Bruce Lee

datapark.io

Comprehensive toolbox for data scientists







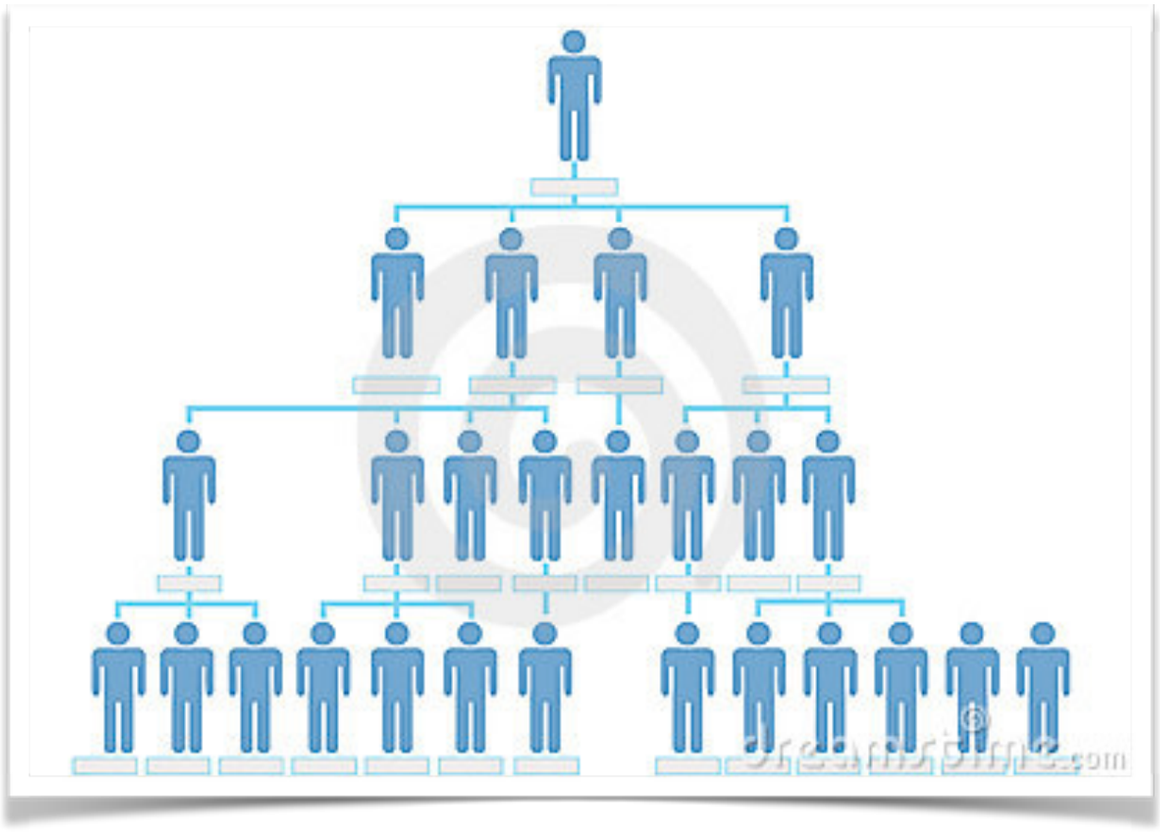





Standard tools and technologies quants and data scientists know and love.

User Management

datapark adds sophisticated user management to the mix



Using the unique, decades long developed and matured user and rights & role management of Linux as the basis (“bottom-up approach”)

Adding standardized features for team sharing and public sharing.

Open as Guideline

Being open in all directions

“Only standards, easy in, easy out, fully integrated.”

Jupyter Notebook, upload, download (eg “zip all”),
integrated with Dropbox, multiple sharing options,
Web folder, deployable anywhere ...

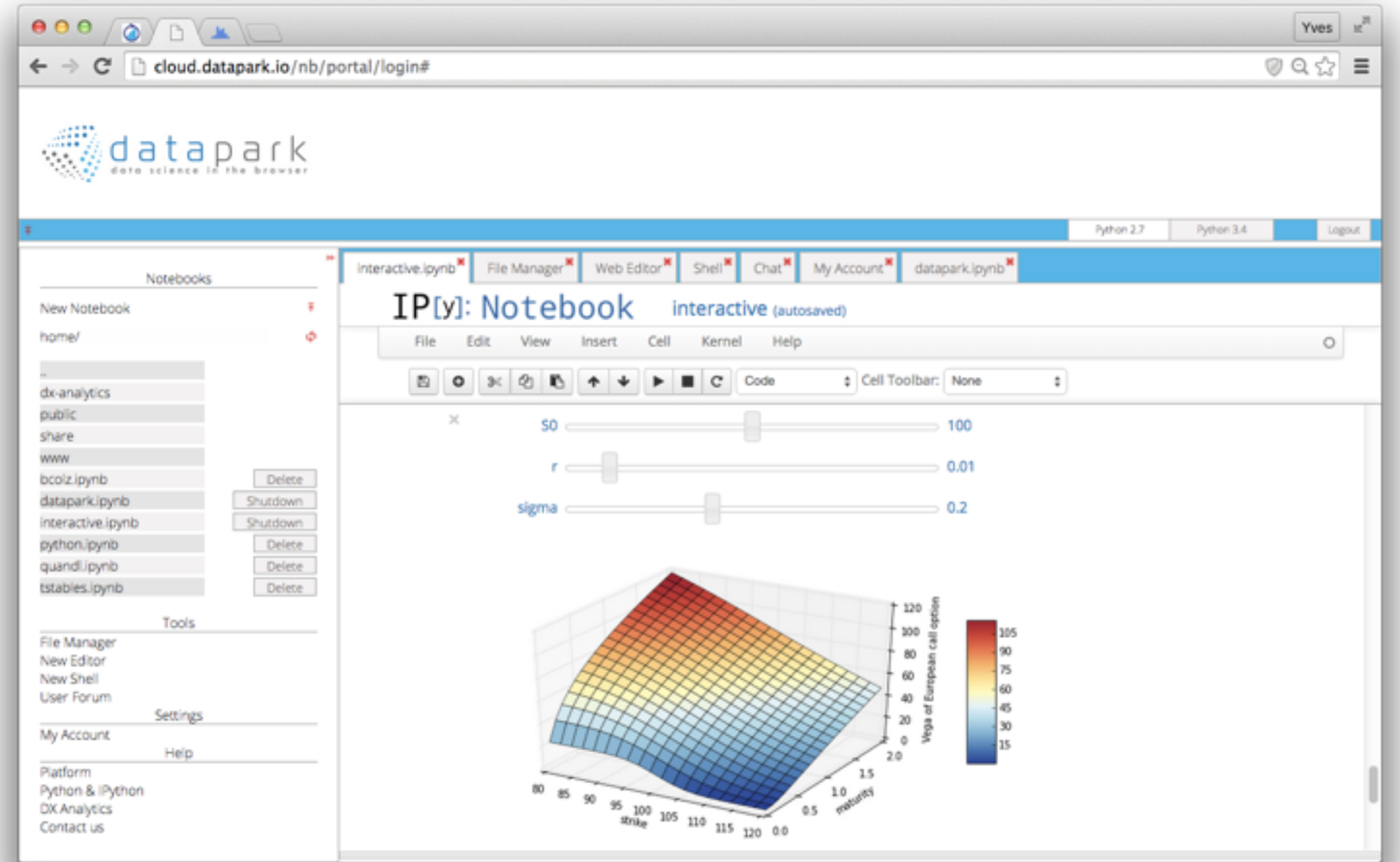
Browser-based Data Science

datapark capitalizes on new Web technologies and tools

1. **Generation: Move Data Around** — data analytics started by moving data from one place to another, analyzing it locally and moving results back to the remote data source
2. **Generation: Move Code Around** — moving tons of data is costly and time consuming; moving small code sets is less costly and faster
3. **Generation: Don't Move Anything** — the Browser and Web technologies allow to work directly and in real-time on the infrastructure where data and code are stored (replacing e.g. remote ssh access)

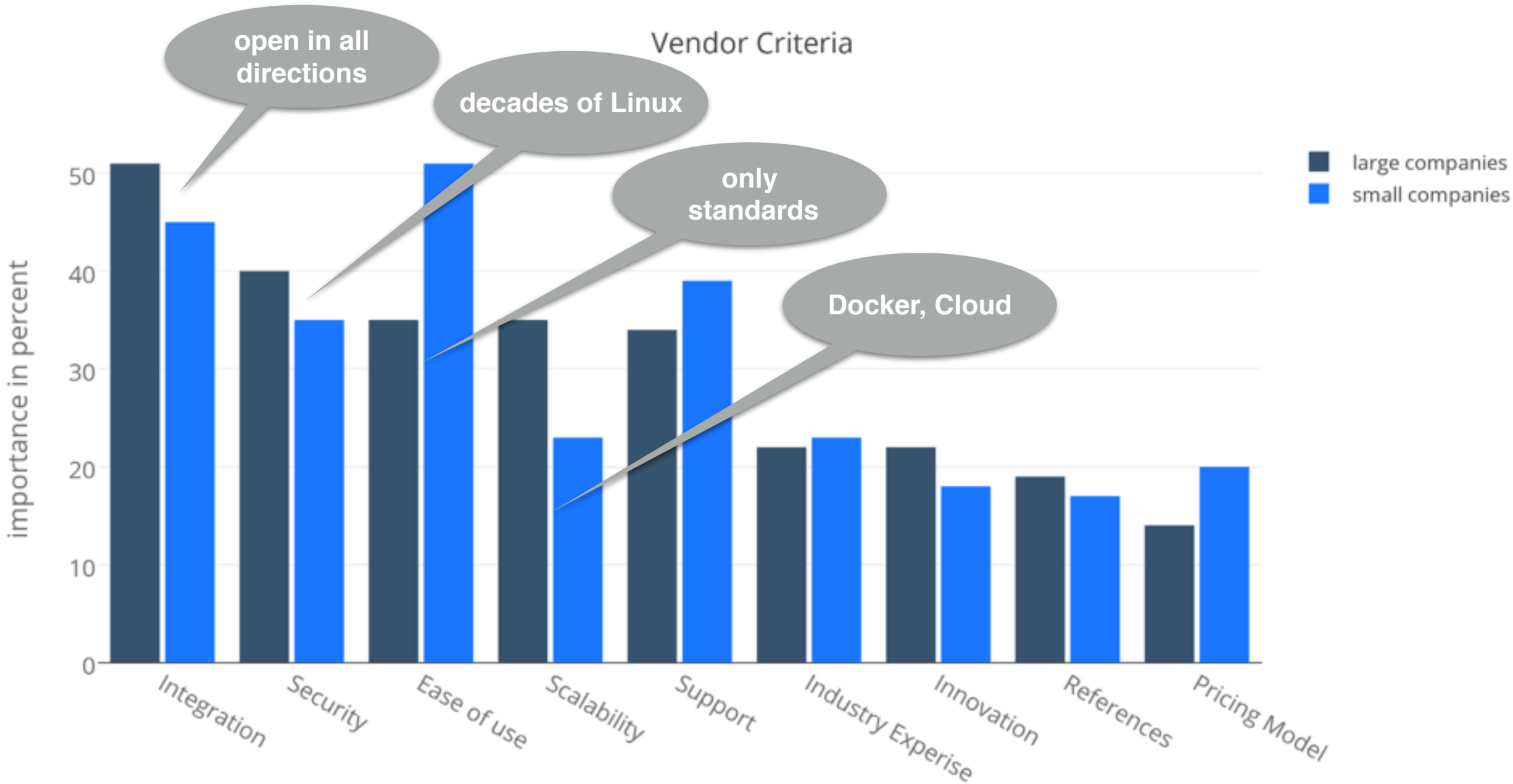
The Result

Bringing the best of Open Source together in the browser



Vendor Criteria in Data Analytics

Integration, security, ease of use & scalability important



Source: 2015 Big Data Analytics Survey (Summary Slides)

I. Open Source Data Science
II. Data Science in the Browser
III. Benefits and Use Cases

Benefits Illustrated

From easy deployment to sharing, publishing and AaaS

Deployment
A single deployment step that only takes between 30 mins to a few hours brings a complete, multi-user data science platform

Analytics and Sharing
Working on data analytics problems and sharing documents, data sources and results with colleagues & others — making use of Jupyter Notebooks, public folder, email functionality & more

Publishing
Converting, for example, Jupyter Notebooks to HTML documents or HTML5 presentations — and publishing them on datapark.io

AaaS and Notebook Hosting
Allowing for collaborative, reproducible analytics work-flows — providing the data, code and the execution environment

Web App Deployment
Developing and deploying full-fledged (Web) applications — from prototypes to full deployment of applications on the same platform and infrastructure

Shipping Data Science Toolbox
datapark is deployed via Docker containers that can run on any Linux based infrastructure — e.g. consultants can bring this toolbox and deploy it on clients' premises (behind firewalls)

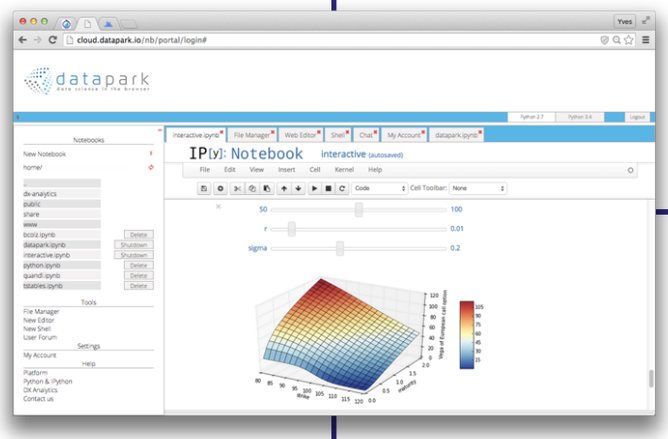
Use Cases for datapark.io

From teaching to data science to AaaS to a social app store

Teaching Programming & Data Science

FitchLearning

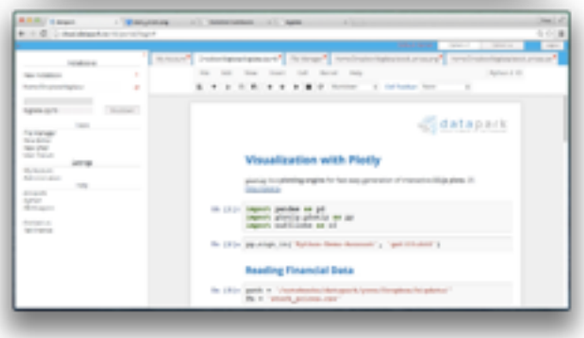
Data Science Platform in Academic Institutions and Corporations



Analytics-as-a-Service for OS Projects and Proprietary Data and Code

eurex

Market Place for Ideas, Projects, Apps etc. (“Social Data Science”)

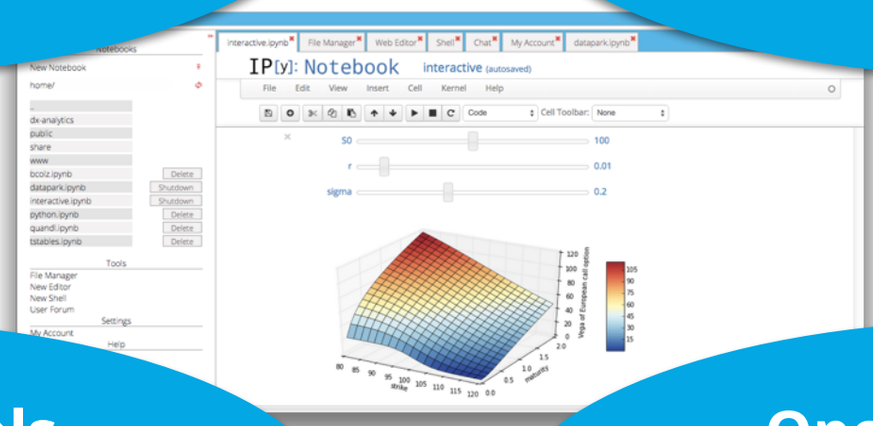


Data Science in the Browser

... based on Open Source and Standards

The Best of Open Source for Data Science

Powerful Infrastructure (Linux, Anaconda, Docker, ...)



Powerful Tools (Jupyter, ACE, Shell w/ eg Git, File Manager)

Open Standards (Py, R, Julia, IPYNB, Linux FS, Dropbox, ...)

Just try it.

http://datapark.io

Give us feedback.

team@datapark.io

datapark

data science in the browser

Dr. Yves J. Hilpisch

datapark.io | team@datapark.io | @dataparkio

The Python Quants GmbH