

Embracing Open Data Science in your Organization

Christine Doig
Senior Data Scientist
Continuum Analytics



Agenda

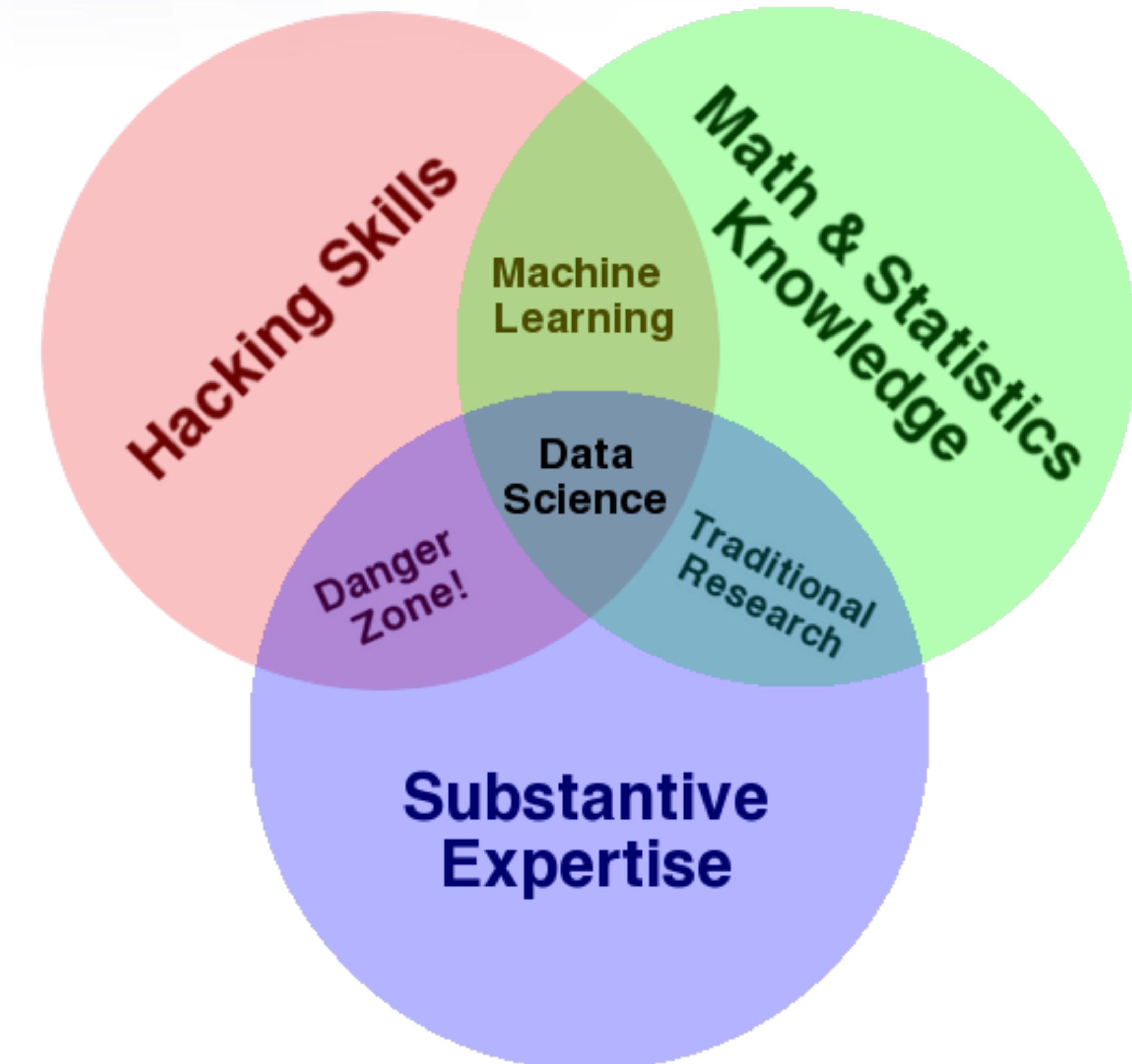
- Introduction to Data Science
- Data Science Challenges in Organizations
- Anaconda Distribution
- Anaconda Community Innovation
- Anaconda Enterprise Platform

INTRODUCTION TO DATA SCIENCE

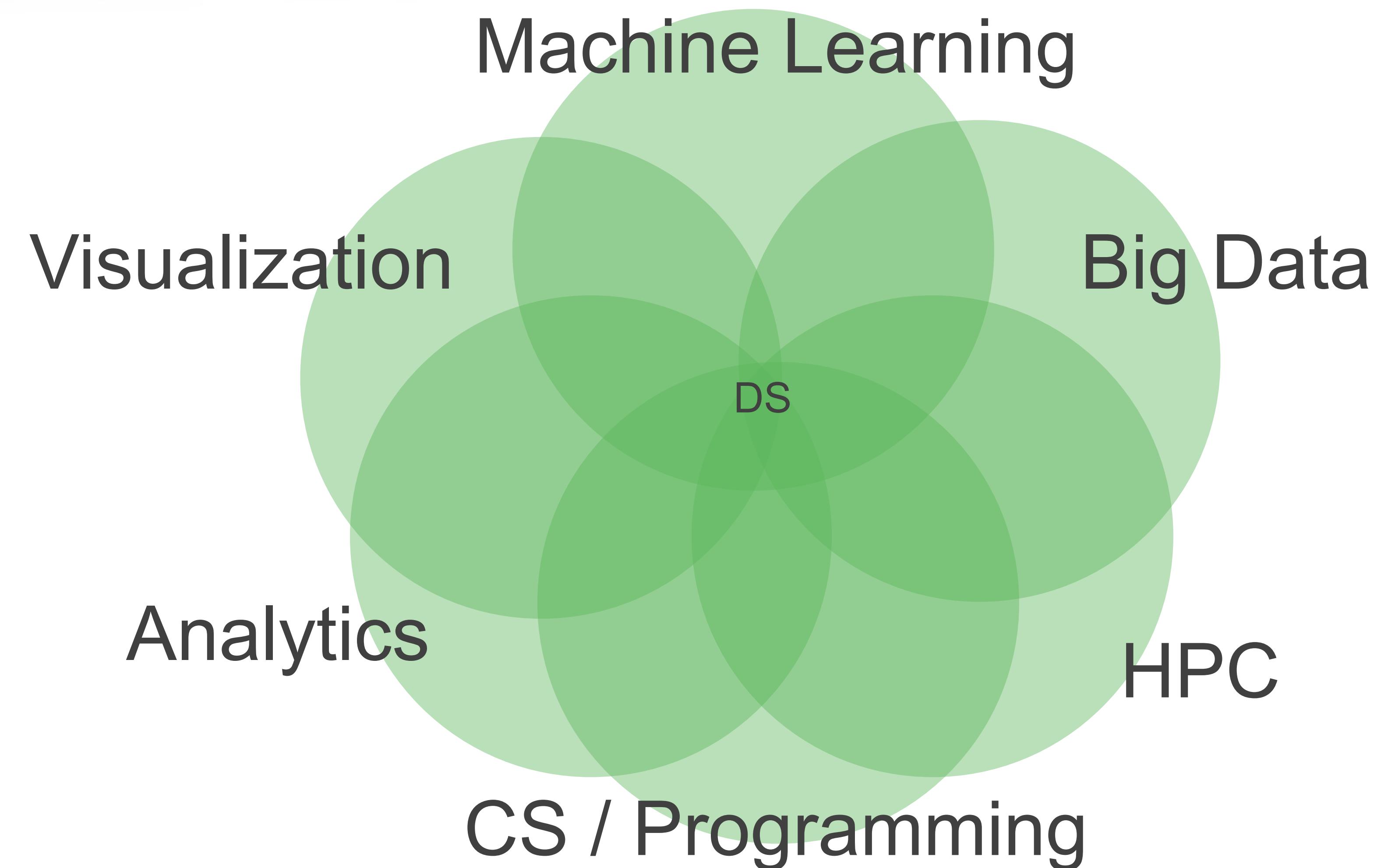


ANACONDA[®]

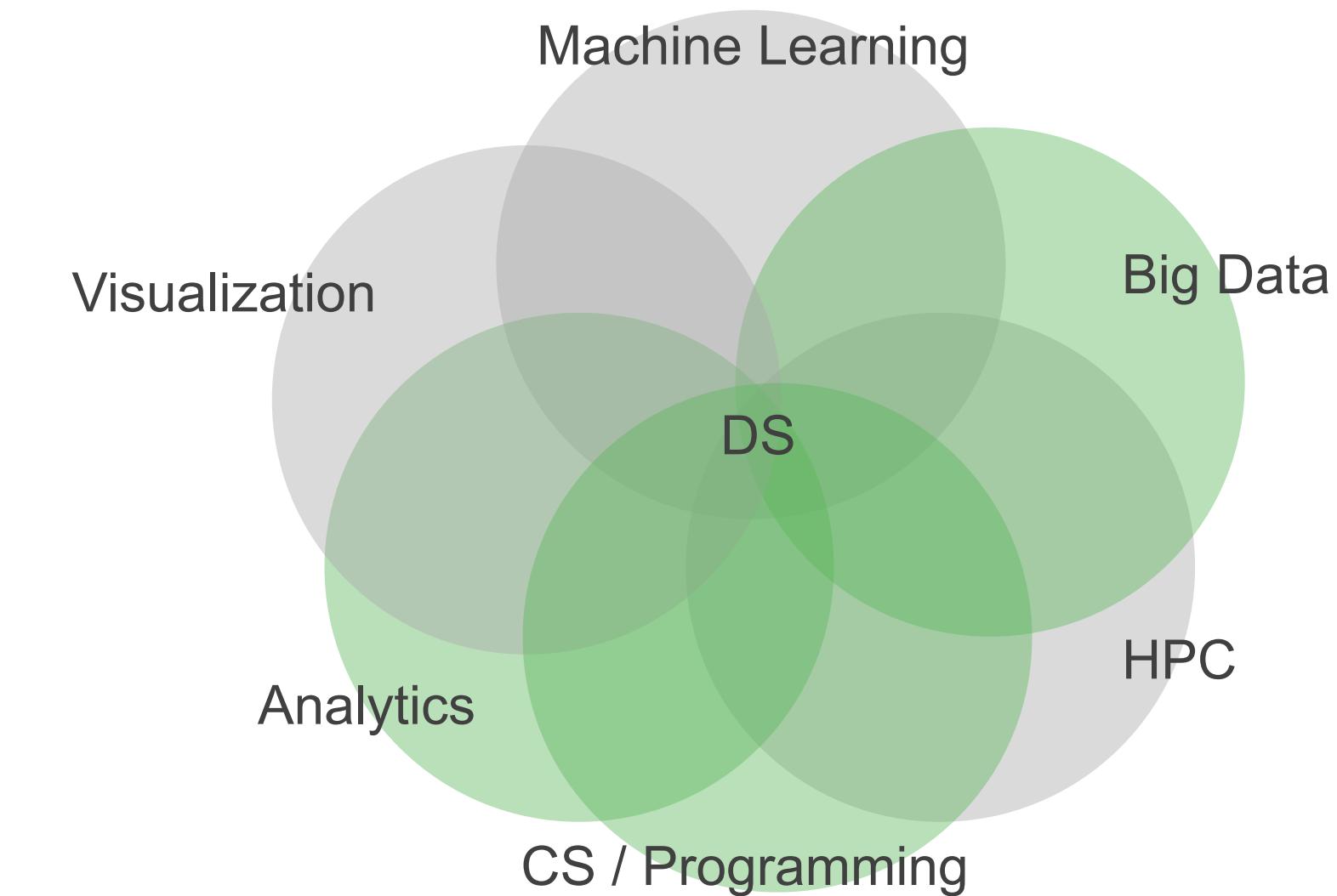
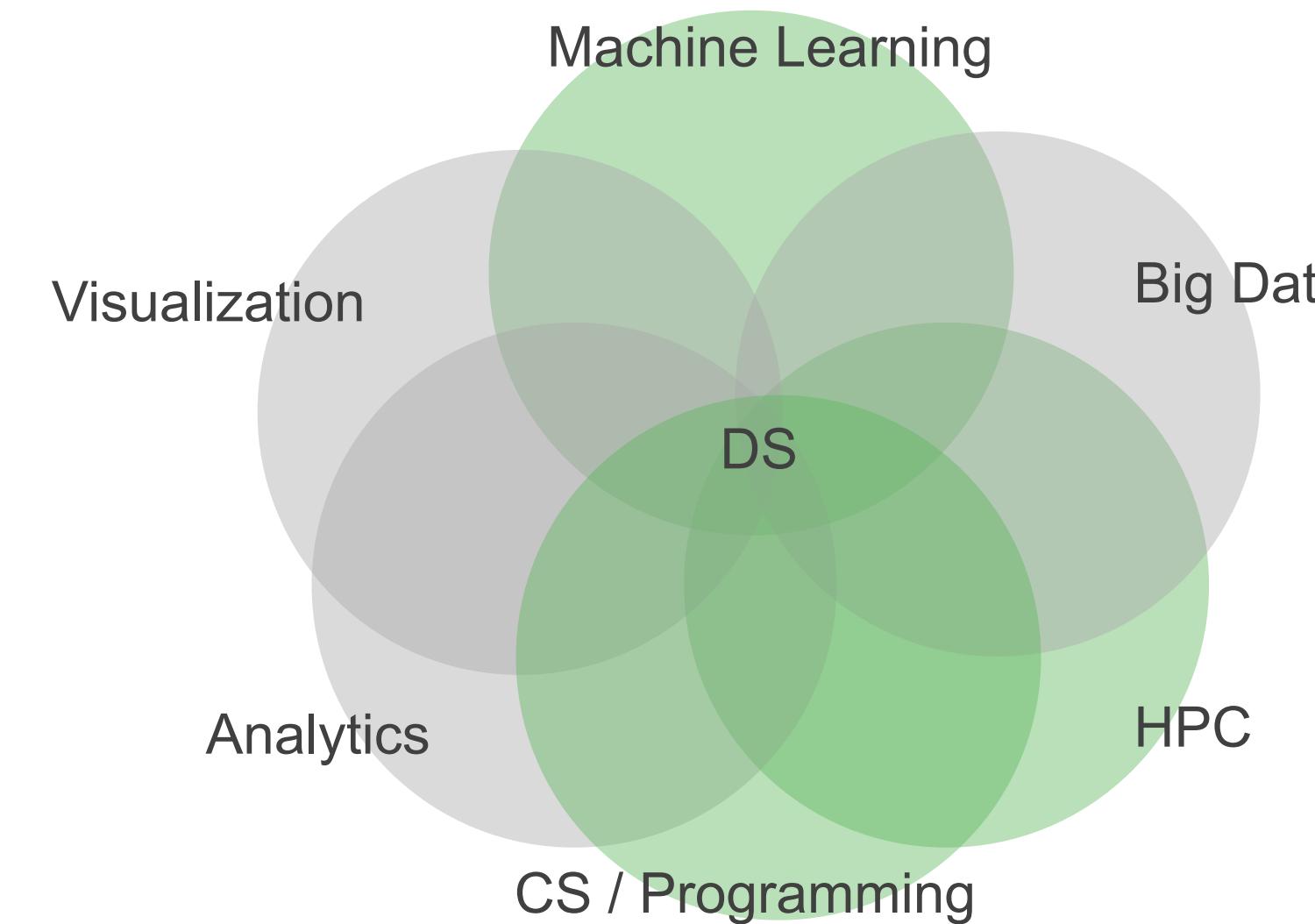
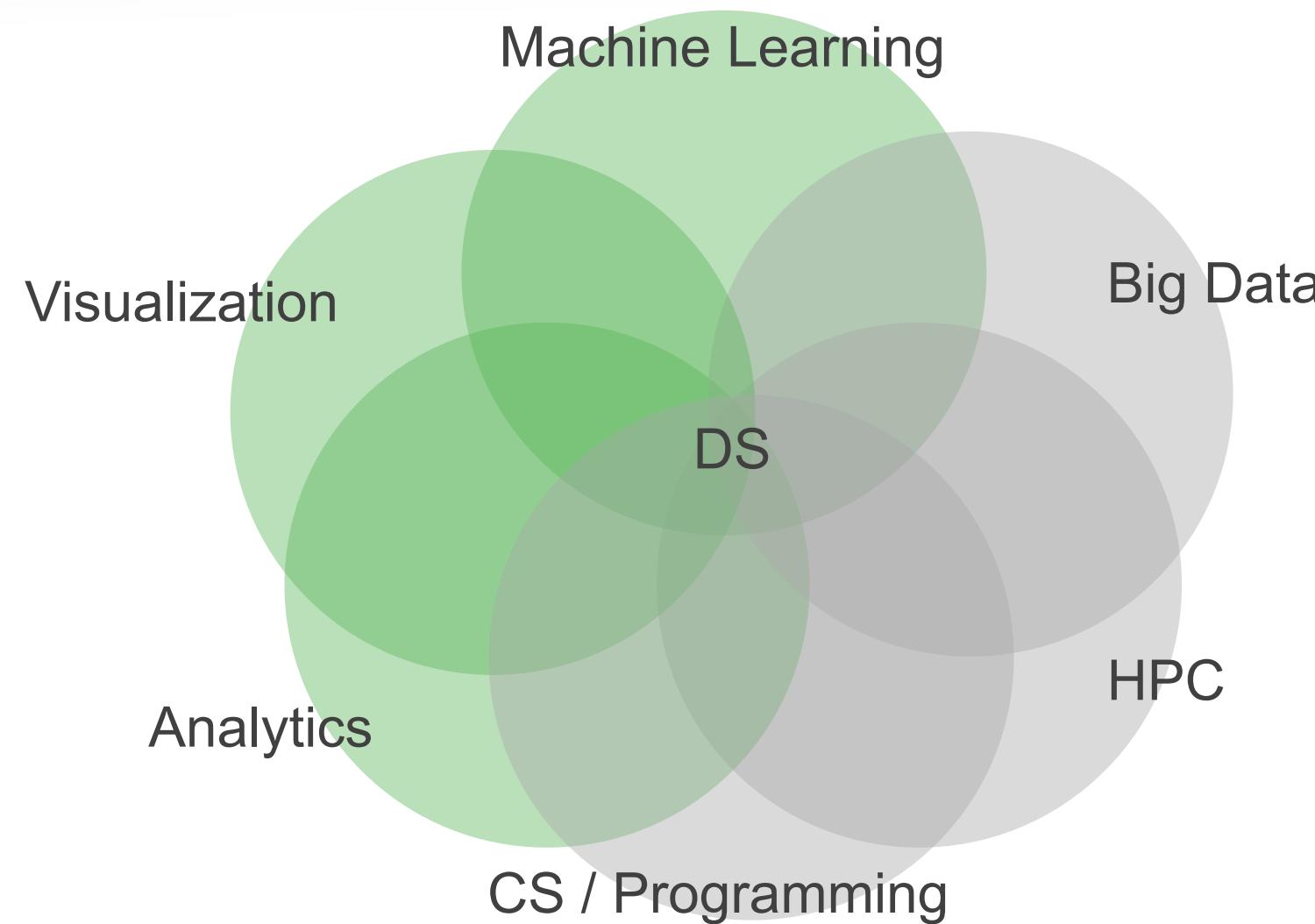
The Data Science Venn Diagram



The Data Science Venn Diagram Revisited



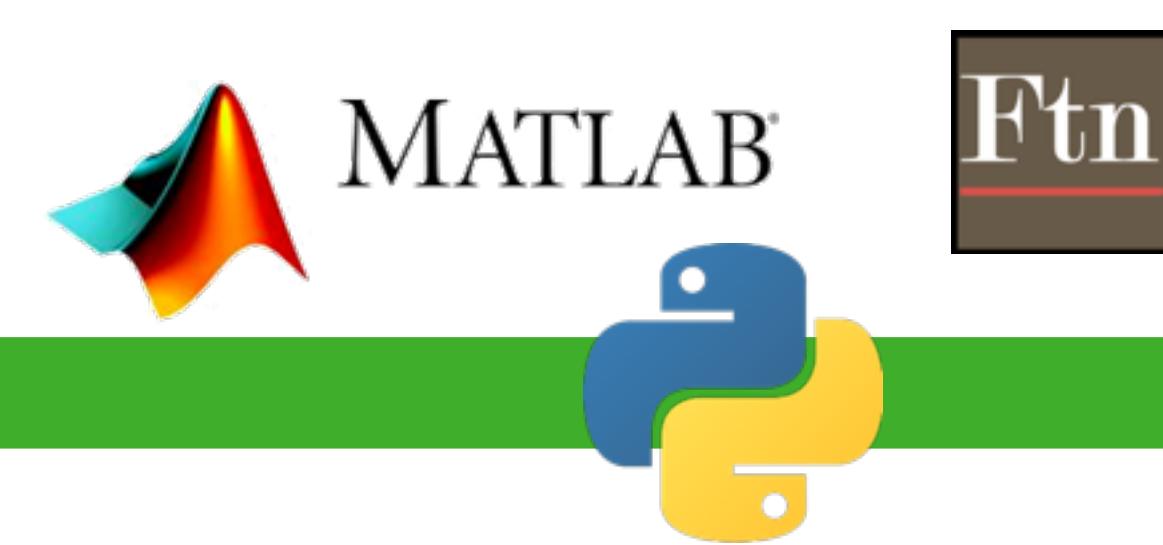
Data Scientist come with different skills and backgrounds



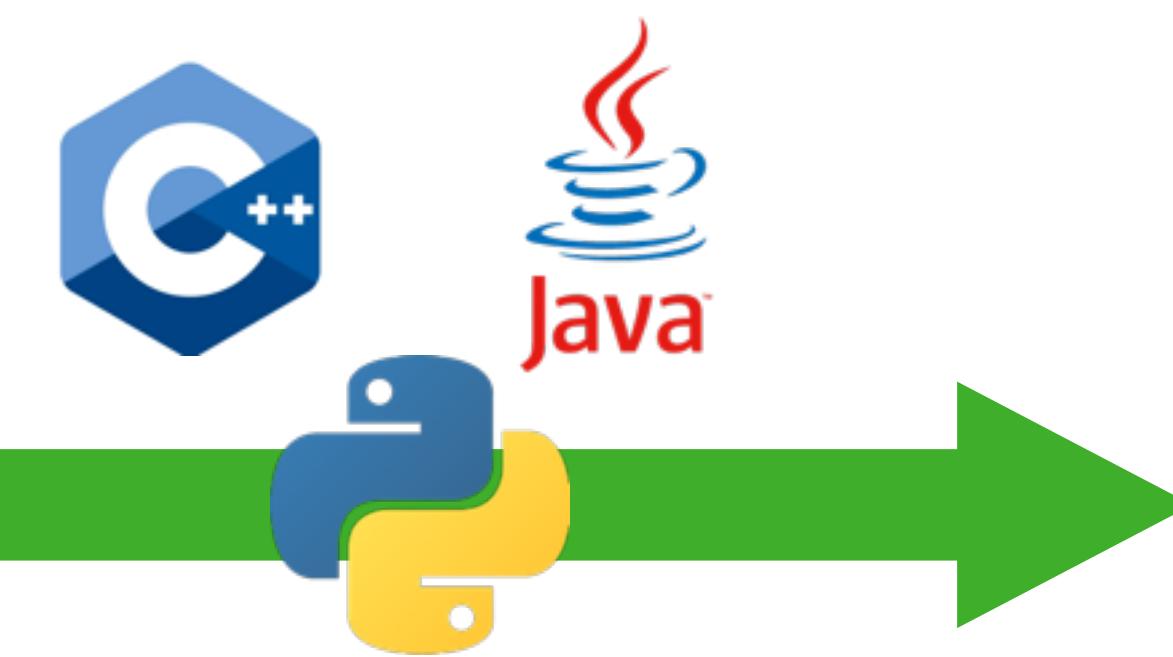
Statistician / Analyst



Research / Computational
Scientist



Developer / Engineer



Data Science in summary:

- is a team sport
- formed by team members with very diverse backgrounds
- both in terms of knowledge (CS, Statistics, Viz, ML...)
- and technology stacks (R, SAS, Python...)

How can companies organize efficiently in this environment?

Open Data Science

With an inclusive movement
that makes open source tools
for data science -- data, analytics, &
computation – easily work together as a
connected ecosystem

Open Data Science Vibrant and Growing Community

Python Community

30M+

Packages in Anaconda

720+

ANACONDA Downloads*

3M+

Spark Python Usage

50%+

R Community

16M+

Open Data Science means...

Availability

Innovation

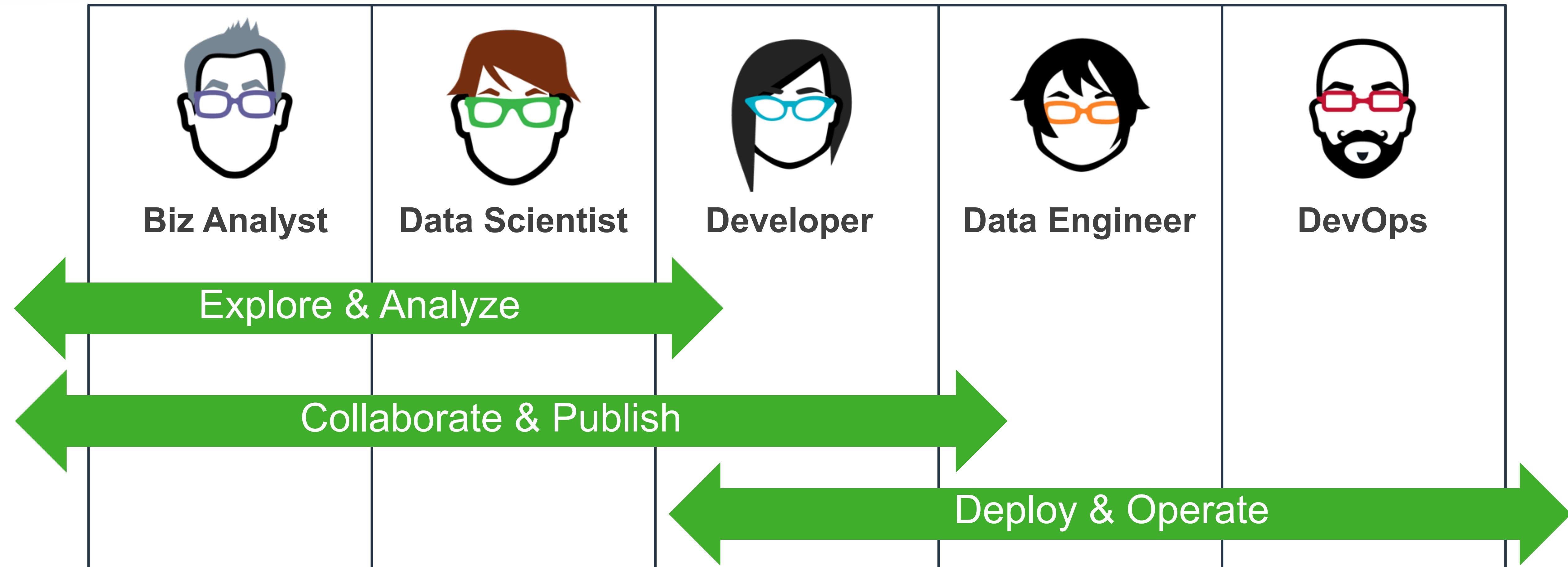
Interoperability

Transparency

For everyone in the data science team

OPEN DATA SCIENCE is the
FOUNDATION TO MODERNIZATION

Data Scientists are not the only player in the Data Science Team



Data Science assets



Biz Analyst

Spreadsheets
Reports
Presentations



Data Scientist

Notebooks
Scripts
Visualizations



Developer

Software packages
Web applications



assets

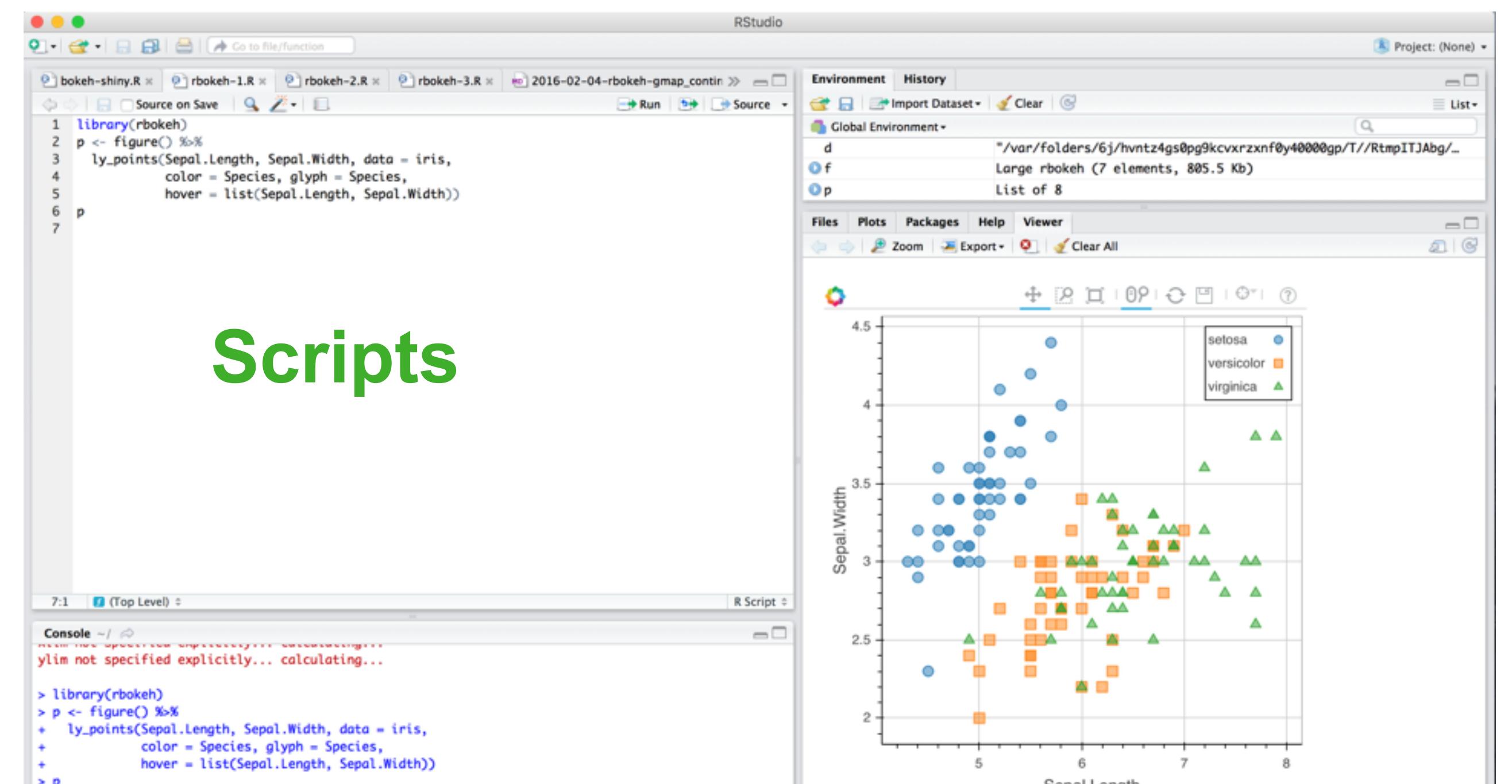
Data Scientist

The screenshot shows the Jupyter Notebook interface. On the left, a sidebar displays the "Welcome to the Notebook Server" message. In the main area, a notebook cell titled "Exploring the Lorenz System" contains text about the Lorenz system and its historical development. Below this, another cell shows a plot of the Lorenz attractor with several sliders for parameters σ , β , and ρ . The code for generating the plot is as follows:

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.), σ=(0.0,50.0), β=(0.,5), ρ=(0.0,50.0));
```

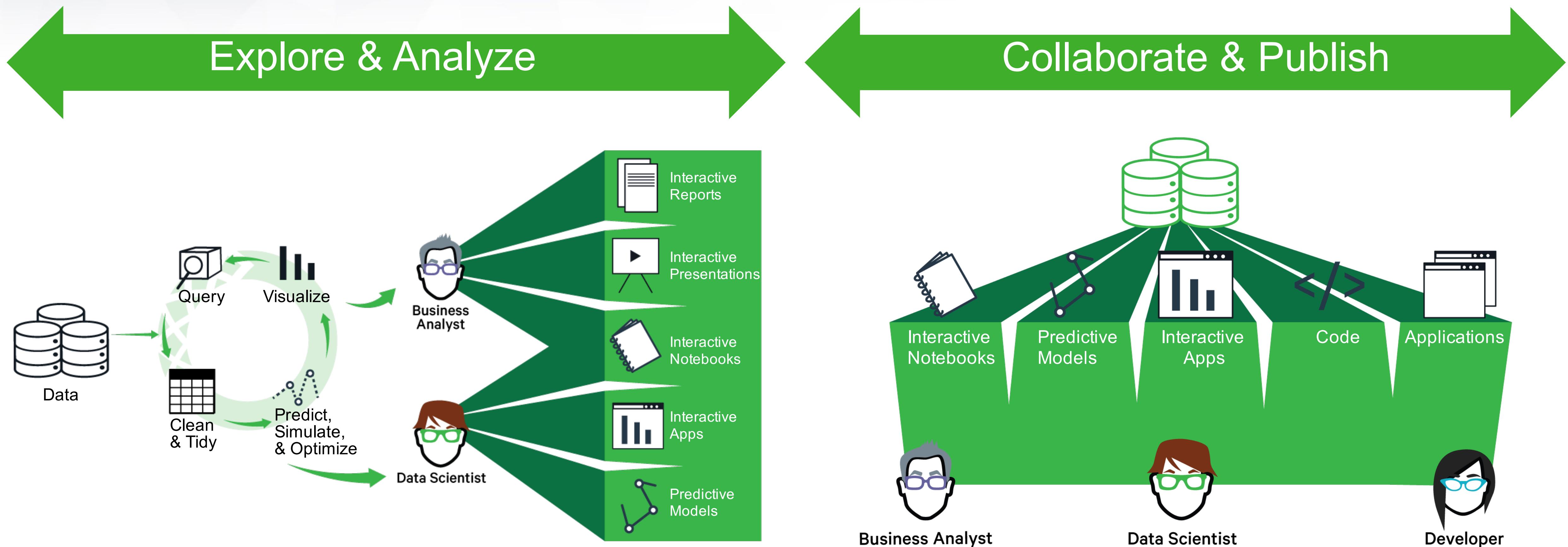
The plot shows a complex, double-lobed structure characteristic of the Lorenz system.

Notebooks

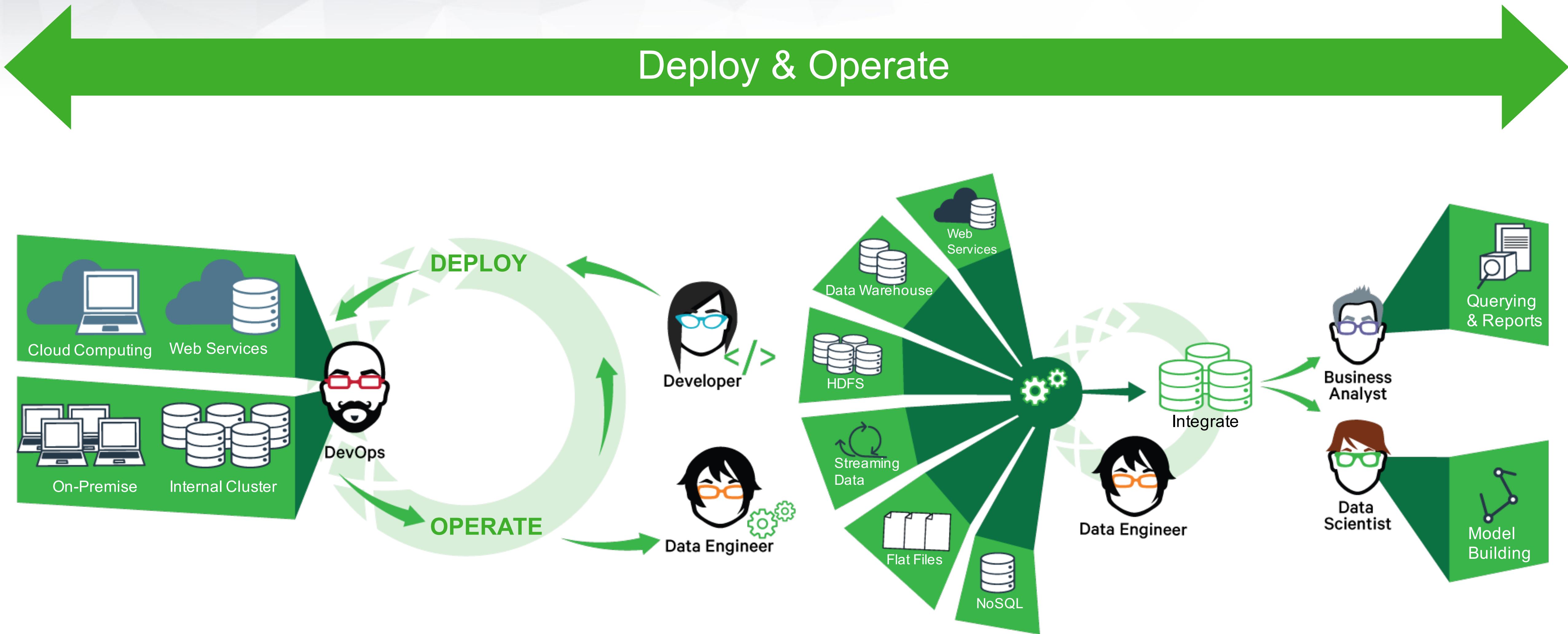


Scripts

Data Science workflows



Data Science workflows



DATA SCIENCE CHALLENGES IN ORGANIZATIONS



ANACONDA[®]

Challenges

- Manage reproducible heterogeneous Data Science environments
- Distribute, share and publish Data Science assets
- Get diverse data scientists (languages, tools, data models, assets...) to collaborate effectively
- Enable Data Scientists to easily leverage Big Data technologies
- Deploy data science assets into production applications
- Share insights with decision makers
- Enable Business Analysts and Managers to leverage Data Science

How are we solving those challenges through:

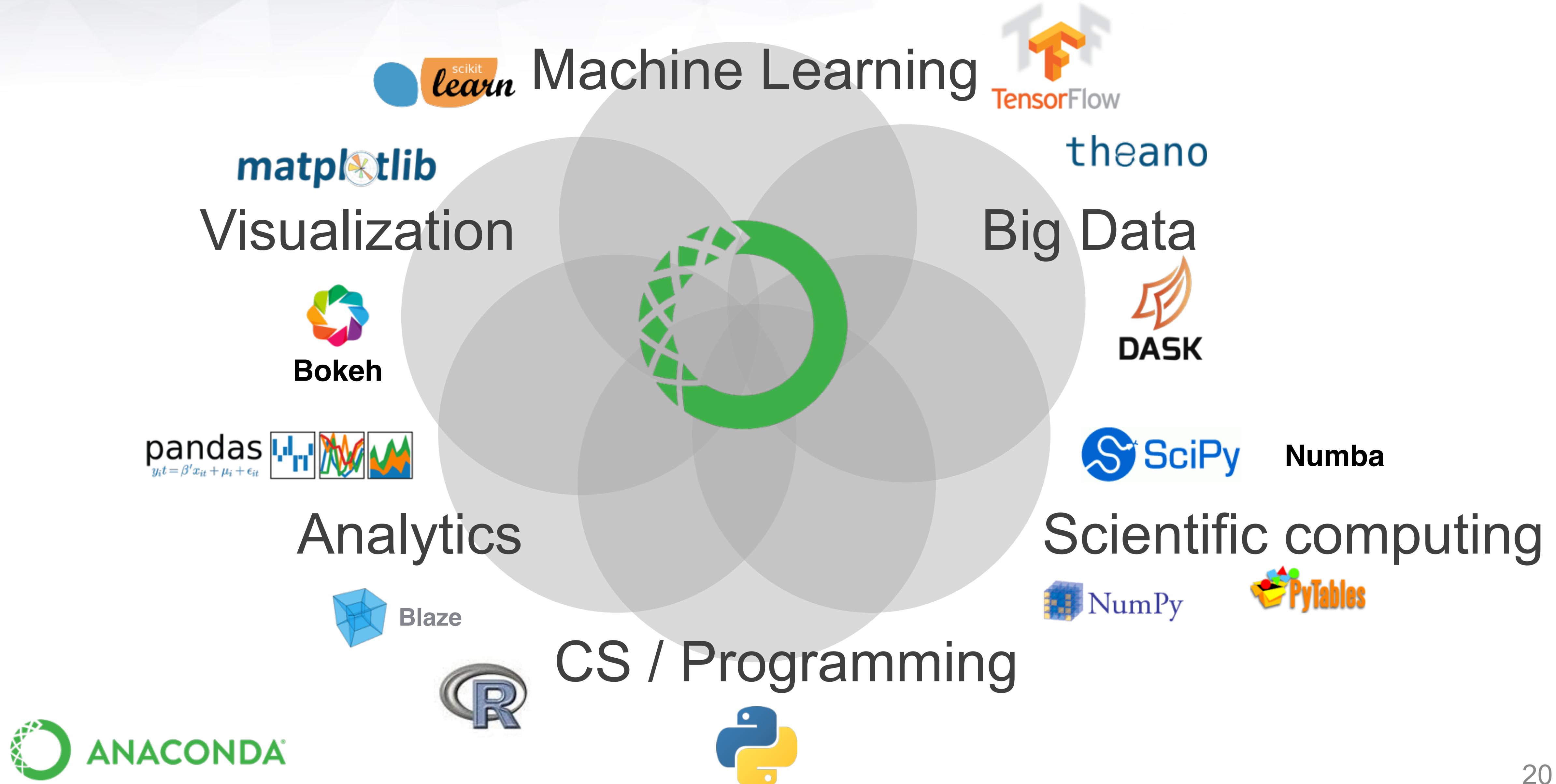
- Anaconda Distribution
- Anaconda Community Innovation
 - Jupyter, JupyterLab and extensions
 - Bokeh for interactive data visualizations
 - Datashader for large scale visualizations
 - Dask for parallel computing
 - Numba for high performance computing
- Anaconda Enterprise

ANACONDA DISTRIBUTION



ANACONDA[®]

The Distribution for Data Science



... with an amazing community!



Bill Santo
@BillSanto

Yes, life is much easier after switching from virtualenv to @ContinuumIO anaconda #python #datascience

RETWEETS
8 LIKES
10

6:10 PM - 18 May 2016



Chris Allison
@ToferC

Follow

If looking at #datascience, check out @ContinuumIO Anaconda distro. If just IDE @pycharm or sublime text are great @am_percival @Wikisteff

RETWEETS
3



4:23 PM - 2 Apr 2016



Eric Ma
@ericmjl

Following

@yvanscher I use #anaconda by @ContinuumIO, and never looked back.

LIKE
1

5:57 PM - 1 Mar 2016



Brian Okken
@brianokken

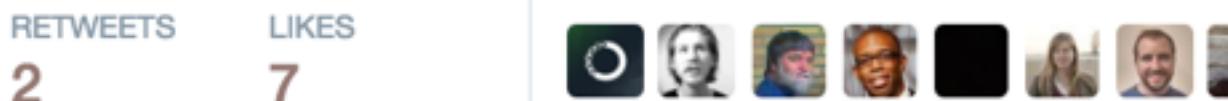
Follow

You forgot step 0. Download Anaconda from @ContinuumIO continuum.io/downloads

Trey Hunner @treyhunner

1. Install Python on Windows 10 🌎
2. Type "idle" into start menu thing 🔎
3. See nothing but "search the web" 😊
4. Cry 😢...

RETWEETS
2 LIKES
7



Yoakum
@Yoakum

Follow

I tried out Anaconda Python @ContinuumIO recently. I wish I would have done that sooner. Very nice to use. continuum.io/downloads

LIKE
1

6:14 AM - 10 Mar 2016

The screenshot shows the Continuum Analytics download page for Anaconda 4.1.1 on OSX. At the top, there are three tabs: "Download for Windows", "Download for OSX" (which is selected), and "Download for Linux". Below the tabs, the page title is "Anaconda 4.1.1" and the subtitle is "For OSX". A note states: "Anaconda is BSD licensed which gives you permission to use Anaconda commercially and for redistribution." There is a "Changelog" link. Under "Graphical Installer", instructions say: "1. Download the graphical installer
2. Double-click the downloaded .pkg file and follow the instructions". Under "Command Line Installer", instructions say: "1. Download the command-line installer
2. Optional: Verify data integrity with MD5 or SHA-256 [More info](#)
3. In your terminal window type one of the below and follow the instructions:
Python 3.5 version".

Python 3.5 version

GRAPHICAL INSTALLER (346M)

COMMAND-LINE INSTALLER (297M)

64-Bit

Python 2.7 version

GRAPHICAL INSTALLER (344M)

COMMAND-LINE INSTALLER (295M)

64-Bit

Download for free: www.continuum.io/downloads

Anaconda Distribution Glossary



PYTHON

NumPy, SciPy, Pandas, Scikit-learn, Jupyter / IPython, Numba, Matplotlib, Spyder, Numexpr, Cython, Theano, Scikit-image, NLTK, NetworkX and 150+ packages

conda

Anaconda distribution



PYTHON
conda

Miniconda

- **Anaconda distribution:** Python distribution that includes 150+ packages for data science (in the installer)
- **Miniconda:** Lightweight version of Anaconda, with just Python and conda.
- **Anaconda Cloud:** Cloud service to host and share public (free) and private data science assets
- **Anaconda Navigator:** Anaconda distribution UI to manage environments, launch applications and learn about what's happening in the community

Anaconda Navigator

The screenshot shows the 'My Applications' section of the Anaconda Navigator. It displays five applications: Jupyter notebook (version 4.1.0), IPython qtconsole (version 4.2.0), Spyder (version 2.3.8), glueviz (version 0.8.2), and orange-app (version 1.0.1). Each application has a 'Launch' button. The sidebar on the left includes links for Home, Environments, Learning, Community, Documentation, Developer Blog, and Feedback.

Launch applications



The screenshot shows the 'Environments' section of the Anaconda Navigator. It lists 222 packages available at the root level, including '_license', 'abstract-rendering', 'alabaster', 'anaconda', 'anaconda-client', 'anaconda-ui', 'apache-maven', 'appnope', 'appscript', 'argcomplete', 'astropy', 'atom', 'babel', and 'backports-abc'. The 'Community' section displays a grid of learning resources such as 'Conda', 'Learning == Creating (We're All in It Together) Keynote', 'The State of Analytics', 'Bokeh for Data Storytelling', 'Basic Bokeh Tutorial', 'Accelerating Scientific Code with Numba', 'Memex: Mining the Dark Web', and 'Numba'.

Manage environments
Learn about the
Anaconda community

CONDA

- **conda**: Cross-platform and language agnostic package and environment manager
- **conda-forge**: A community led collection of recipes, build infrastructure and distributions for the conda package manager
- **conda environments**: custom isolated sandboxes to easily reproduce and share data science projects
- **conda kapsel**: reproducible, executable project directories



```
$ conda install python=2.7  
$ conda install pandas  
$ conda install -c r r  
$ conda install -c conda-forge tensorflow
```

Install dependencies

```
name: myenv  
channels:  
- chdoig  
- r  
- foo  
dependencies:  
- python=2.7  
- r  
- r-ldavis  
- pandas  
- mongodb  
- spark=1.5  
- pip  
- pip:  
- flask-migrate  
- bar=1.4
```

```
$ conda env create  
$ source activate myenv
```

Manage multiple environments

```
$ conda kapsel run plot --show
```

Deploy an interactive visualization

What challenges does Anaconda Distribution solve?



PYTHON

NumPy, SciPy, Pandas, Scikit-learn, Jupyter / IPython, Numba, Matplotlib, Spyder, Numexpr, Cython, Theano, Scikit-image, NLTK, NetworkX and 150+ packages

conda

Anaconda distribution



Miniconda

- Easy to install on all platforms
- Language agnostic - Python, R, Scala...
- Trusted by industry leaders
- Trusted by the community - Large user base: 3M+ downloads
- BSD license
- Extensible - easily build, share and install proprietary libraries with Anaconda Cloud
- Allows isolated custom sandboxes with different versions of packages - conda environments
- Allows for easy encapsulation and deployment of data science assets - conda kapsel

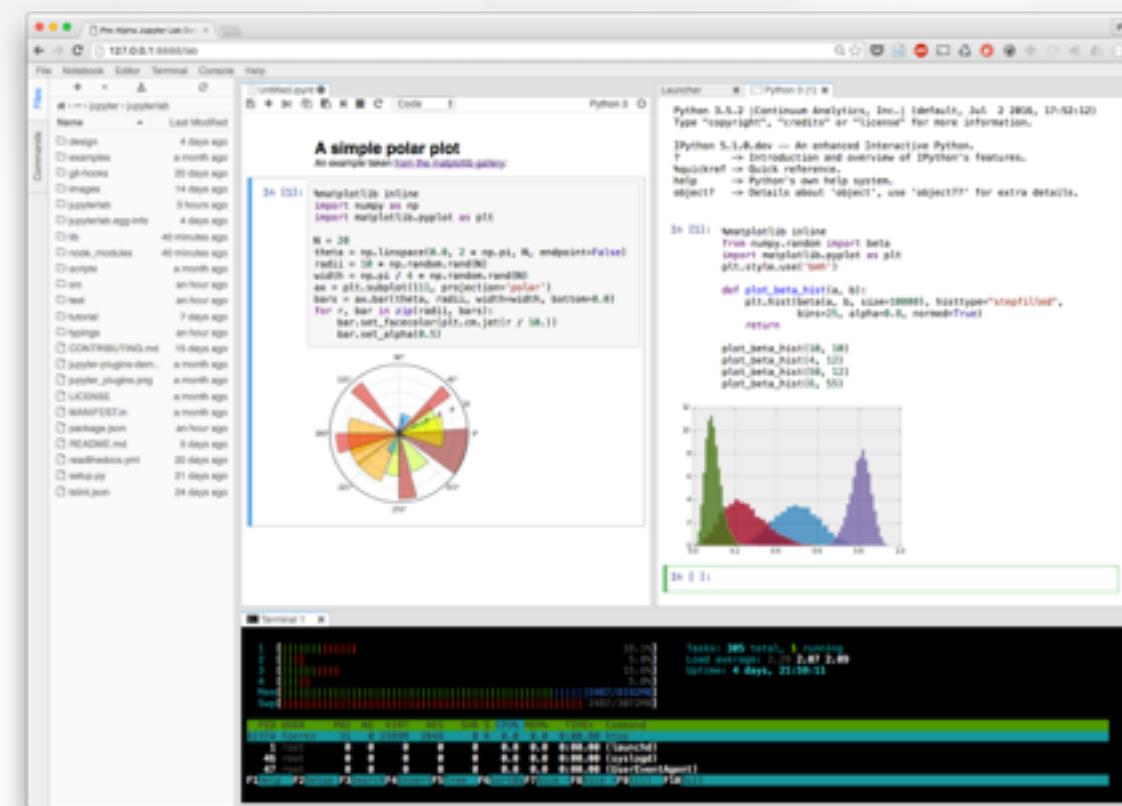
ANACONDA COMMUNITY INNOVATION



ANACONDA[®]

- Anaconda Distribution
- Anaconda Community Innovation
 - Jupyter, JupyterLab and extensions
 - Bokeh for interactive data visualizations
 - Datashader for large scale visualizations
 - Dask for parallel computing
- Anaconda Enterprise

Continuum Analytics contributions to the Python ODS ecosystem



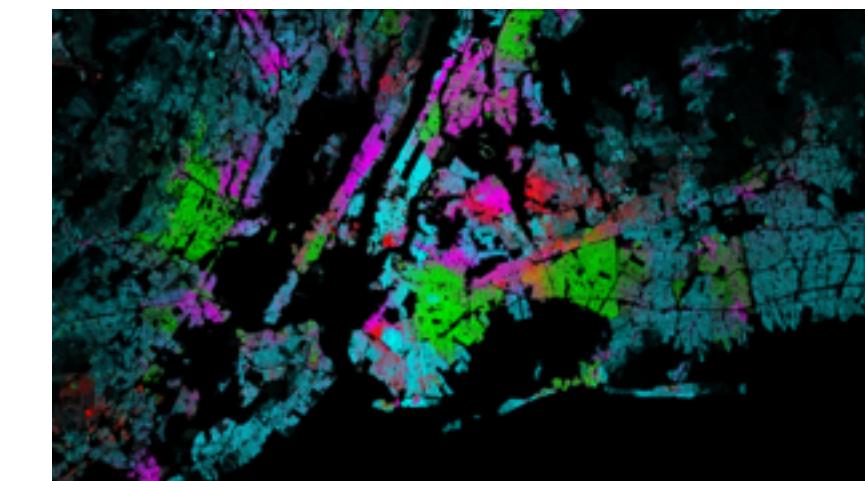
JupyterLab

- Next generation Data Science IDE



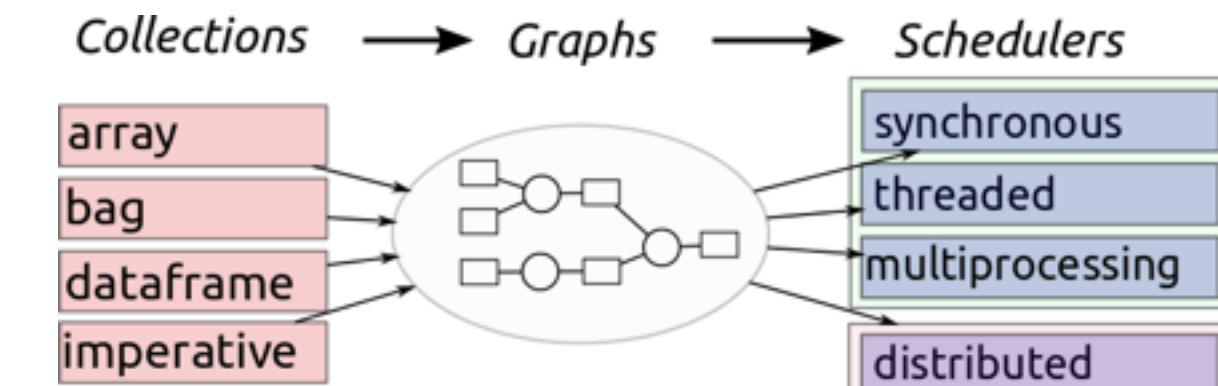
Bokeh

- Web interactive data visualizations (no JS)



Datashader

- Graphics pipeline system for creating meaningful representations of large amounts of data



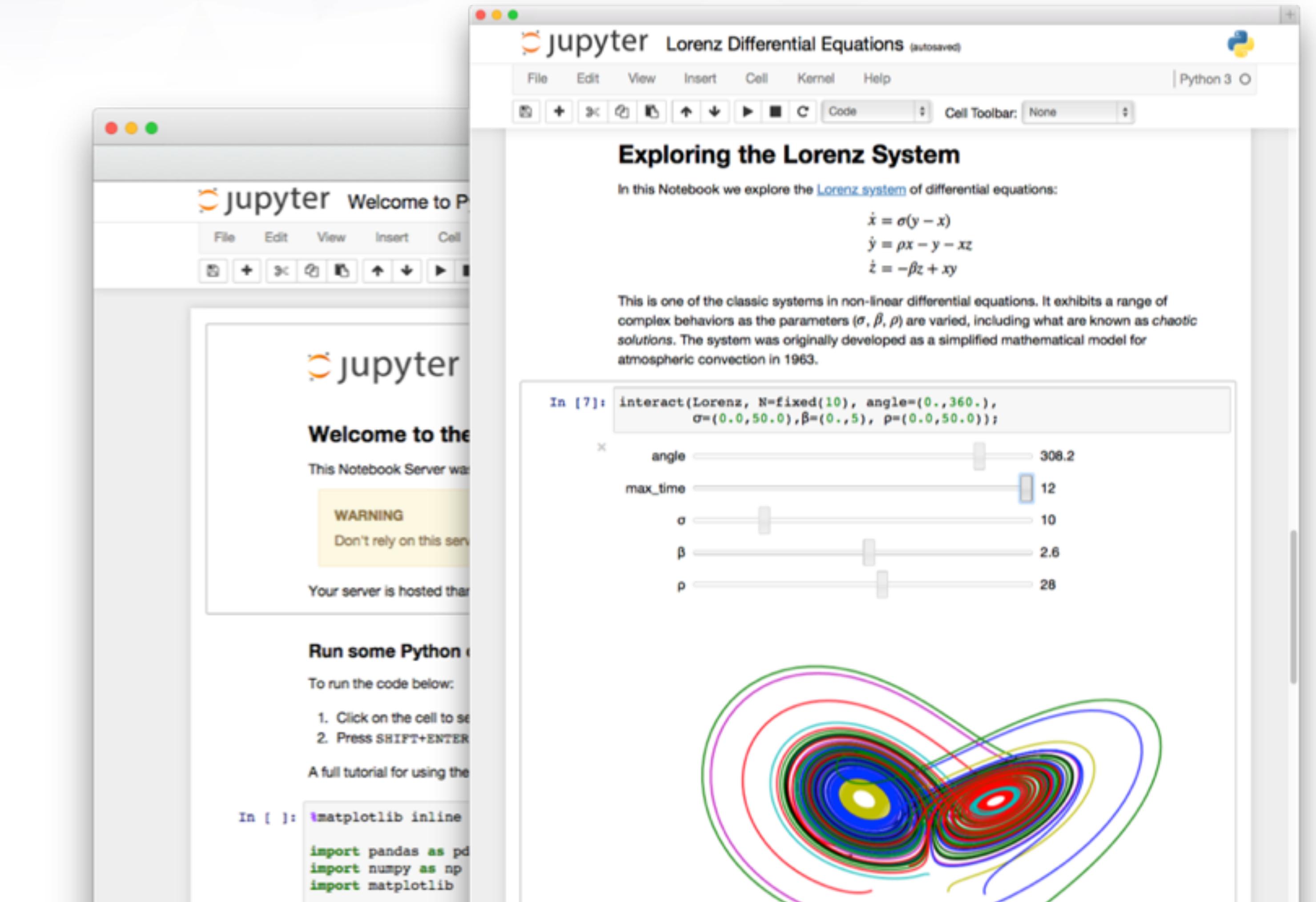
Dask

- Parallel computing framework

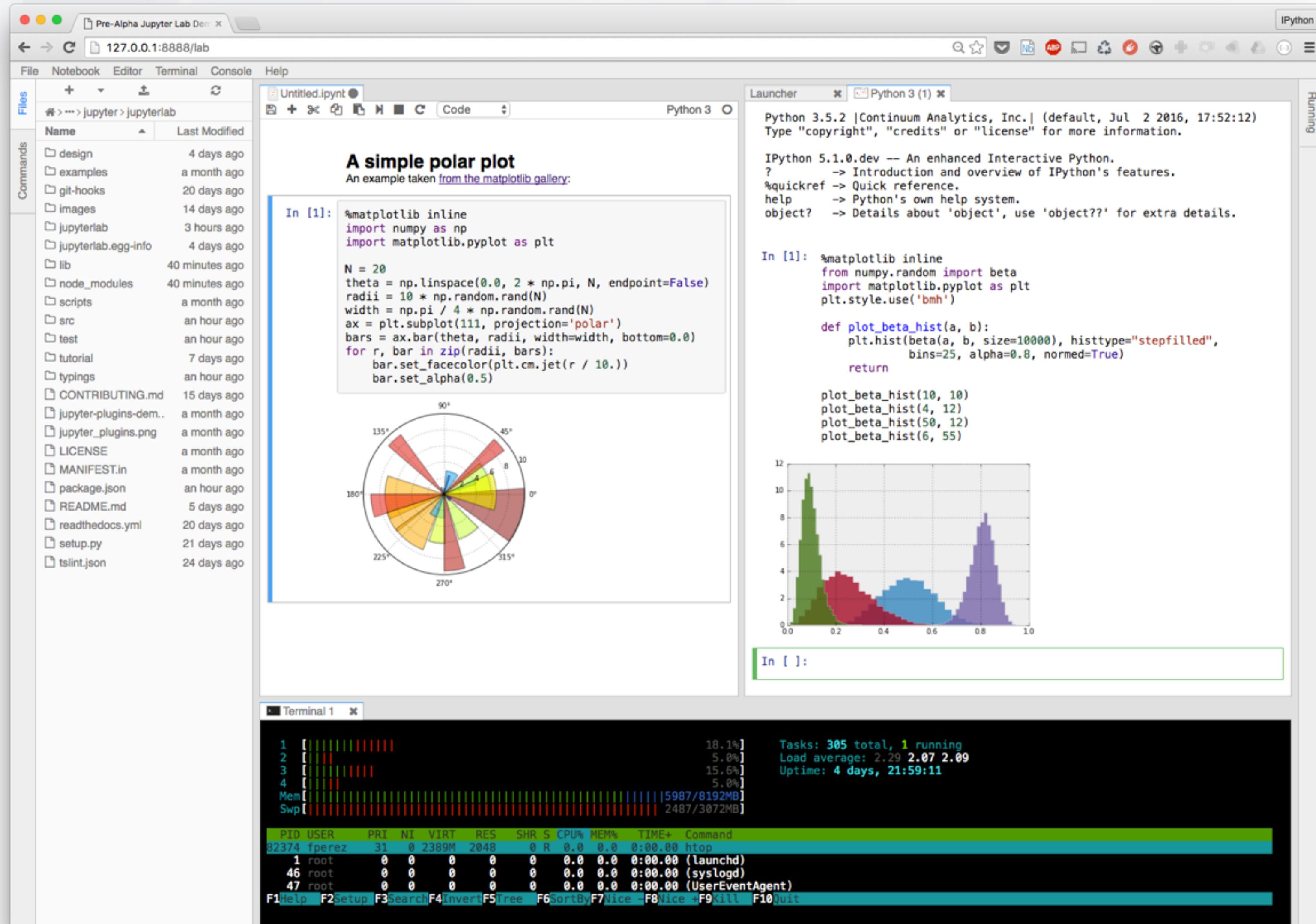
Jupyter Notebook

Web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.

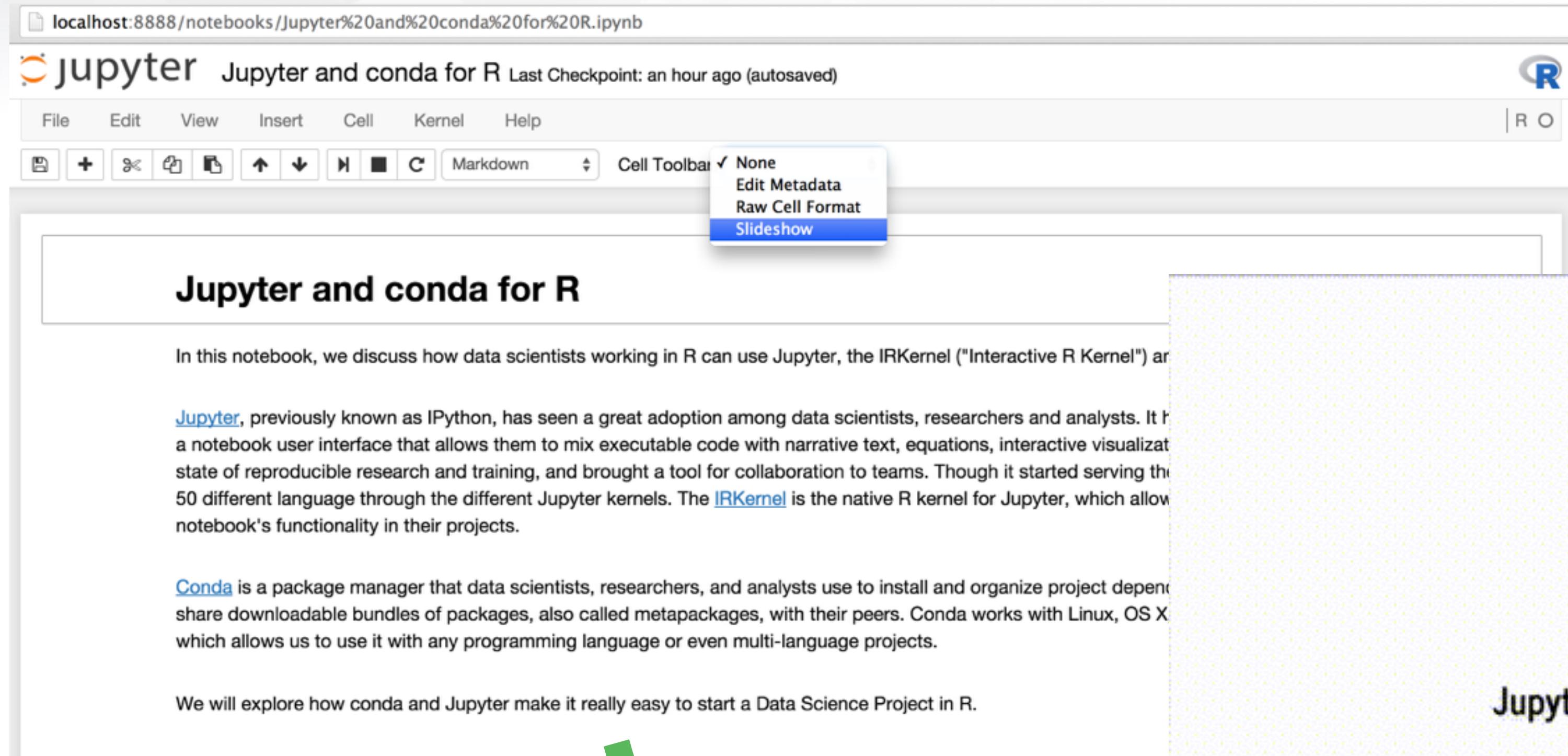
```
$ jupyter notebook
```



JupyterLab: the next generation



Sharing insights with decision makers



localhost:8888/notebooks/Jupyter%20and%20conda%20for%20R.ipynb

Jupyter Jupyter and conda for R Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Help

Cell Toolbar: None Edit Metadata Raw Cell Format Slideshow

Jupyter and conda for R

In this notebook, we discuss how data scientists working in R can use Jupyter, the IRKernel ("Interactive R Kernel") ar

Jupyter, previously known as IPython, has seen a great adoption among data scientists, researchers and analysts. It h a notebook user interface that allows them to mix executable code with narrative text, equations, interactive visualiz state of reproducible research and training, and brought a tool for collaboration to teams. Though it started serving the 50 different language through the different Jupyter kernels. The IRKernel is the native R kernel for Jupyter, which allow notebook's functionality in their projects.

Conda is a package manager that data scientists, researchers, and analysts use to install and organize project depend share downloadable bundles of packages, also called metapackages, with their peers. Conda works with Linux, OS X which allows us to use it with any programming language or even multi-language projects.

We will explore how conda and Jupyter make it really easy to start a Data Science Project in R.

Jupyter and conda for R

From text, code and visualizations directly to slides

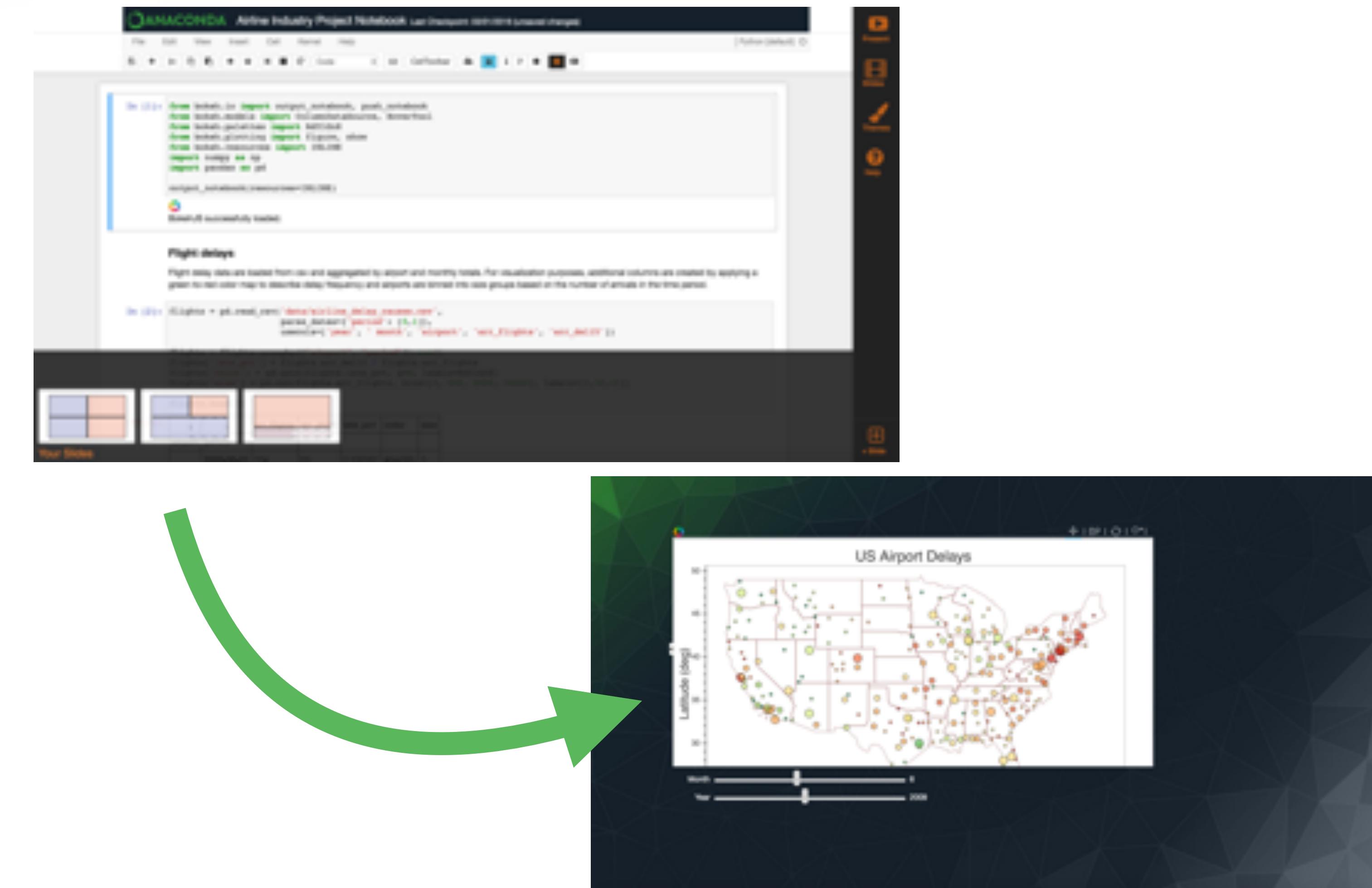


Jupyter: Extensions - nbpresent

*remix your Jupyter
Notebooks as
interactive slideshows
with a UI editor*

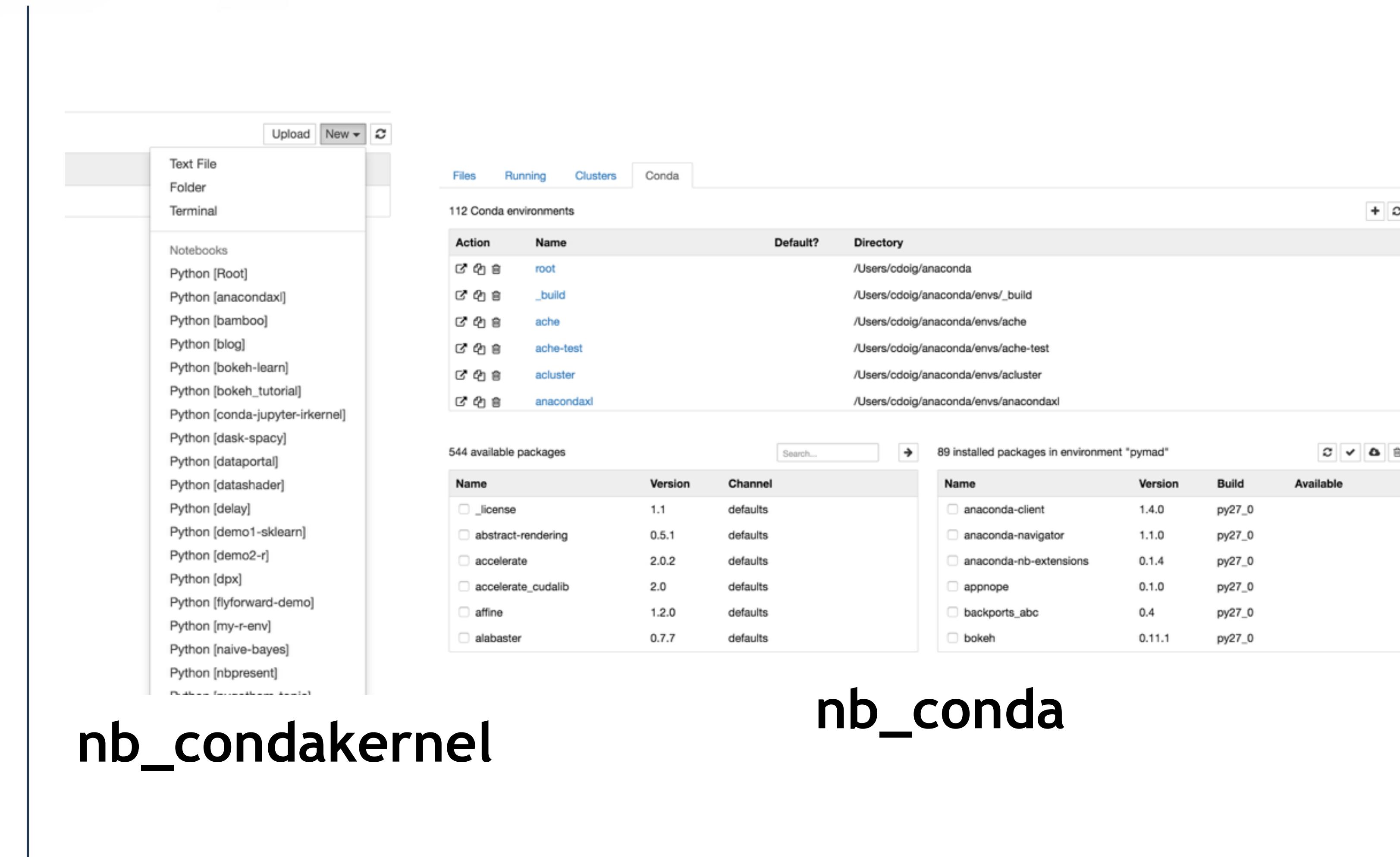
- Edit slides, layout and themes

```
conda install -c anaconda-nb-extensions nbpresent  
jupyter notebook
```



Jupyter extensions - anaconda-nb-extensions

- **nb_condakernel**: use the kernel-switching dropdown inside notebook UI to switch between conda envs
- **nb_conda**: help manage conda envs from inside file viewer of jupyter notebook



nb_condakernel

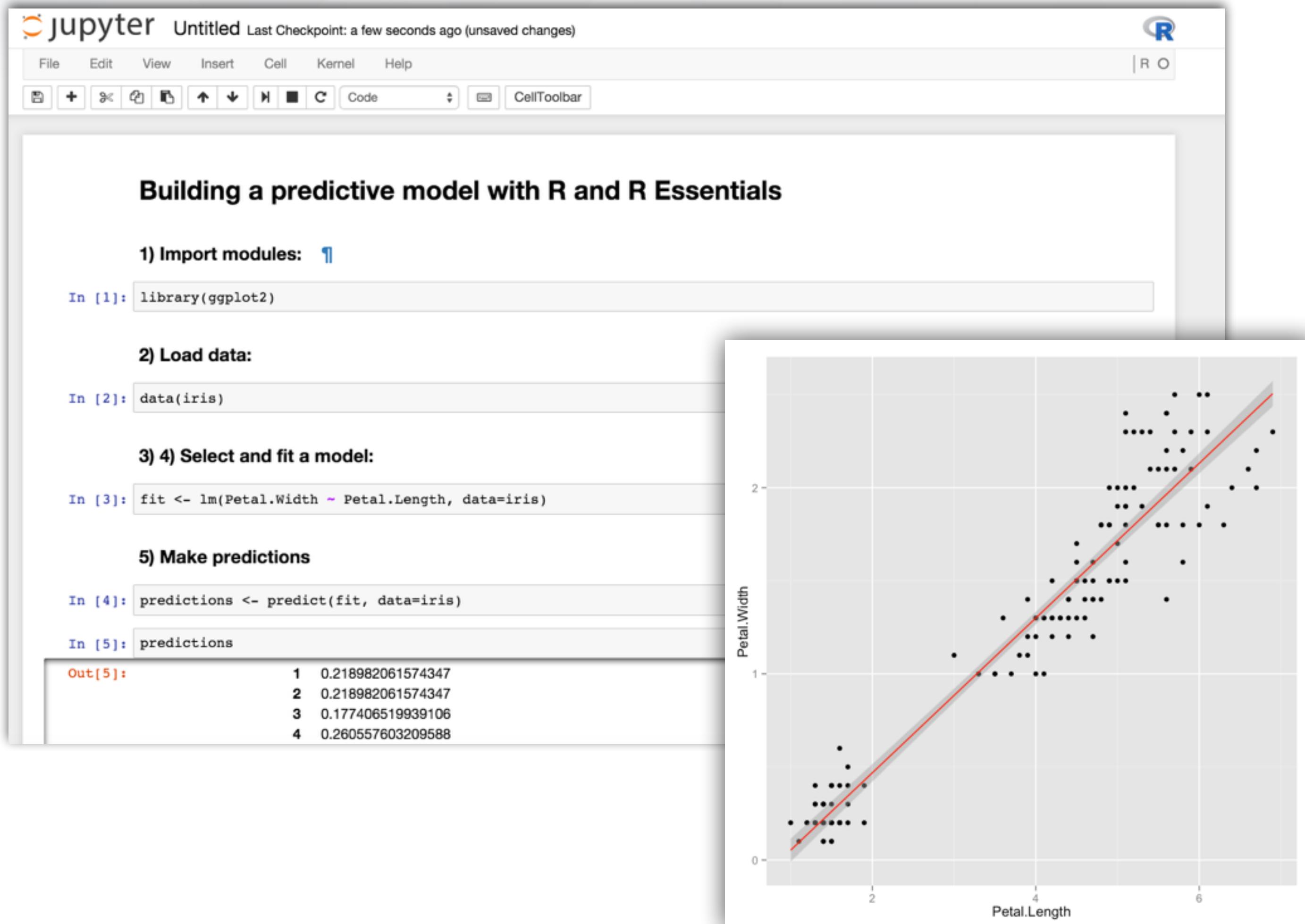
Action	Name	Default?	Directory
root	root		/Users/cdoig/anaconda
_build	_build		/Users/cdoig/anaconda/envs/_build
ache	ache		/Users/cdoig/anaconda/envs/ache
ache-test	ache-test		/Users/cdoig/anaconda/envs/ache-test
acluster	acluster		/Users/cdoig/anaconda/envs/acluster
anacondaxl	anacondaxl		/Users/cdoig/anaconda/envs/anacondaxl

Name	Version	Channel
_license	1.1	defaults
abstract-rendering	0.5.1	defaults
accelerate	2.0.2	defaults
accelerate_cudalib	2.0	defaults
affine	1.2.0	defaults
alabaster	0.7.7	defaults

Name	Version	Build	Available
anaconda-client	1.4.0	py27_0	
anaconda-navigator	1.1.0	py27_0	
anaconda-nb-extensions	0.1.4	py27_0	
appnope	0.1.0	py27_0	
backports_abc	0.4	py27_0	
bokeh	0.11.1	py27_0	

nb_conda

Jupyter: IRkernel



Trivial to get started writing R notebooks the same way you write Python ones.

```
conda config --add channels r
conda install r-essentials
jupyter notebook
```

Bokeh

0.11.1 ▾ INSTALLATION USER GUIDE GALLERY REFERENCE RELEASES DEVELOPER GUIDE

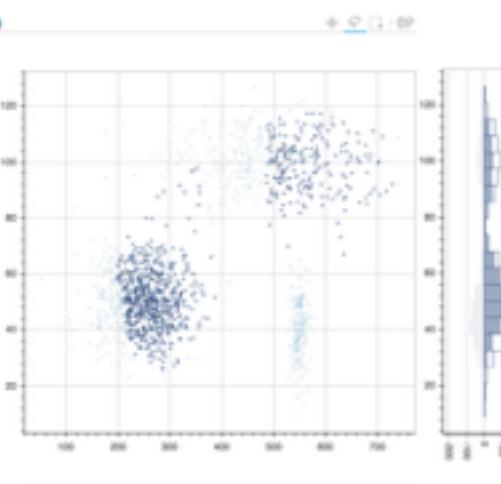
Gallery

Server App Examples

The examples linked below all show off usage of the Bokeh server. The Bokeh server provides a place where interesting things can happen—data can be updated to in turn update the plot, and UI and selection events can be processed to trigger more visual updates.



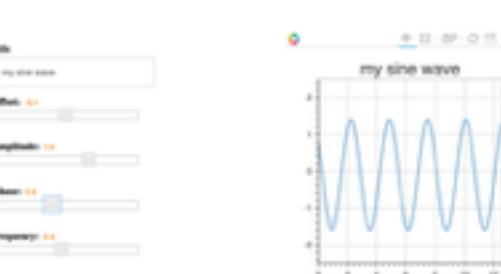
An interactive query tool for a set of IMDB data
Source code: [movies](#)



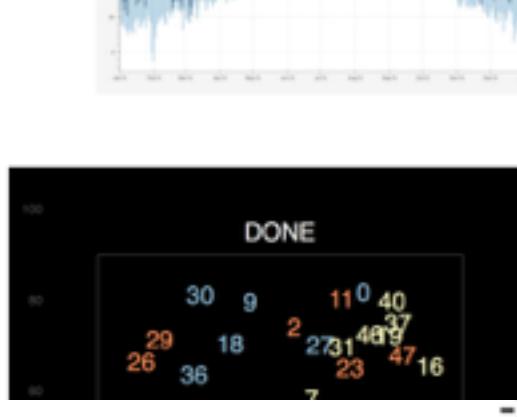
Shows axis histograms for selected and nonselected points in a scatter plot
Source code: [selection_histogram](#)



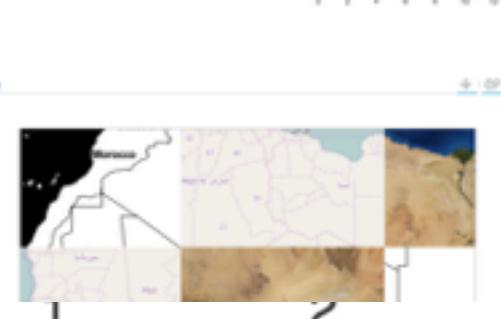
Interactive weather statistics for three cities.
Source code: [weather](#)



A basic demo that has sliders for controlling a plotted trigonometric function
Source code: [sliders.py](#)



An updating plot that demonstrates using timeout callbacks in Bokeh server apps
Source code: [timeout.py](#)



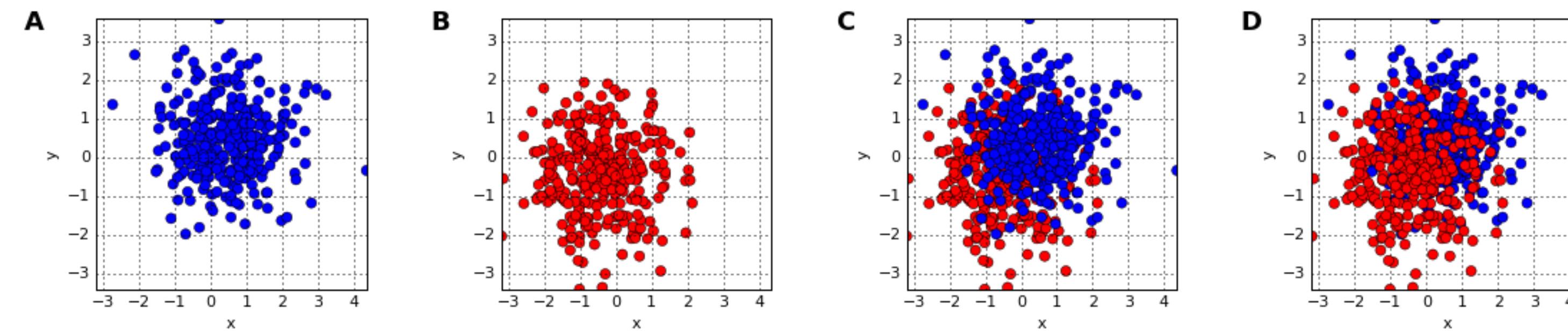
A user-defined extension showing randomized tiles.
Source code: [random_tiles](#)

Interactive visualization framework that targets modern web browsers for presentation

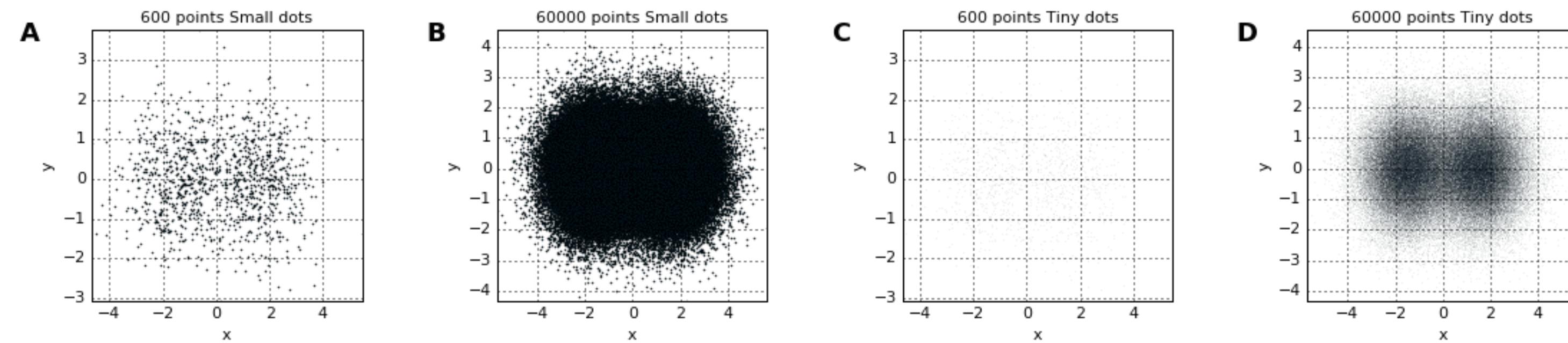
- No JavaScript
- Python, R, Scala and Lua bindings
- Easy to embed in web applications
- Server apps: data can be updated, and UI and selection events can be processed to trigger more visual updates.

Datashader - Plotting pitfalls

Overplotting:



Undersampling:

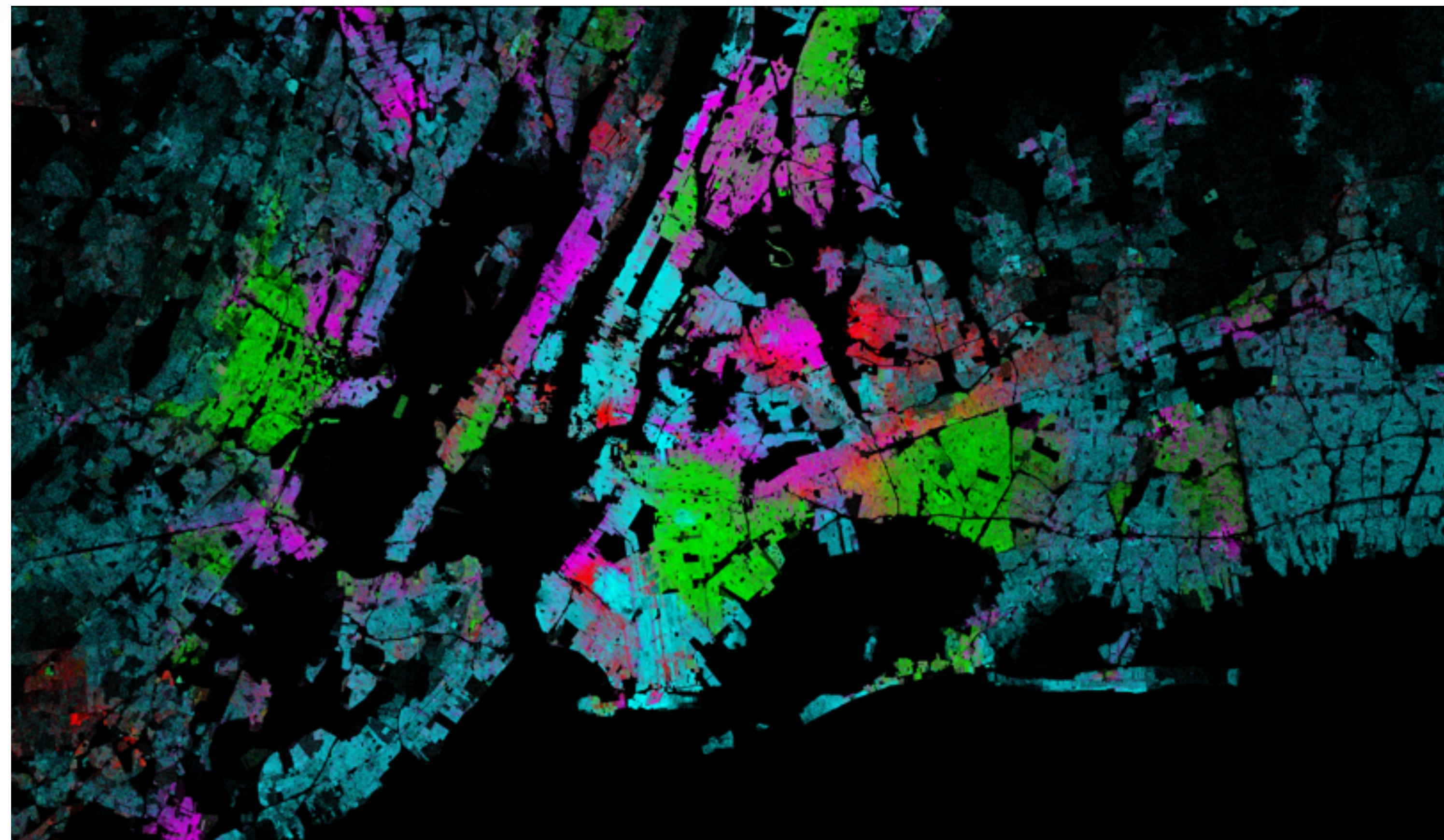


Datashader

*graphics pipeline system for
creating meaningful
representations of
large amounts of data*

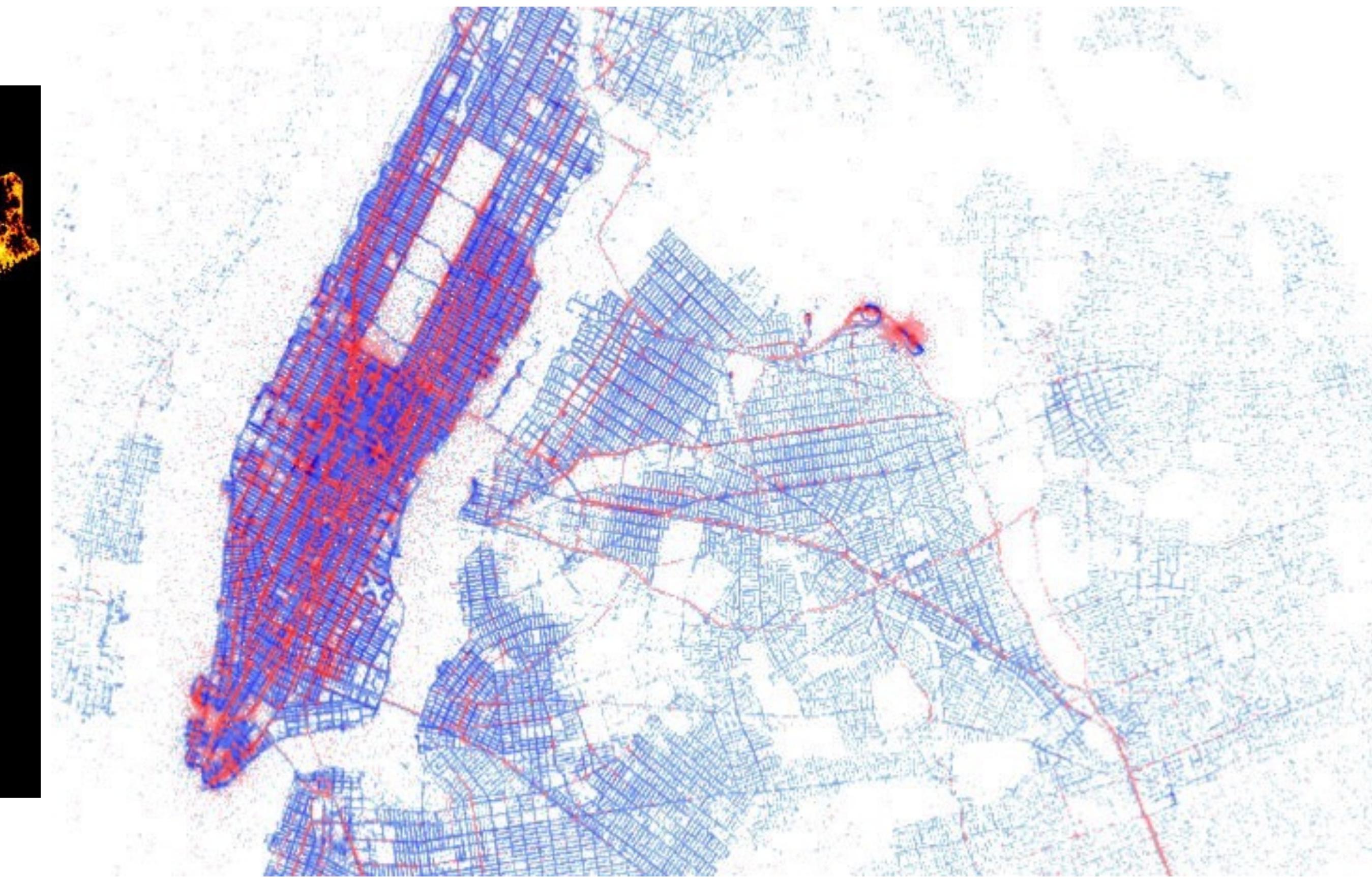
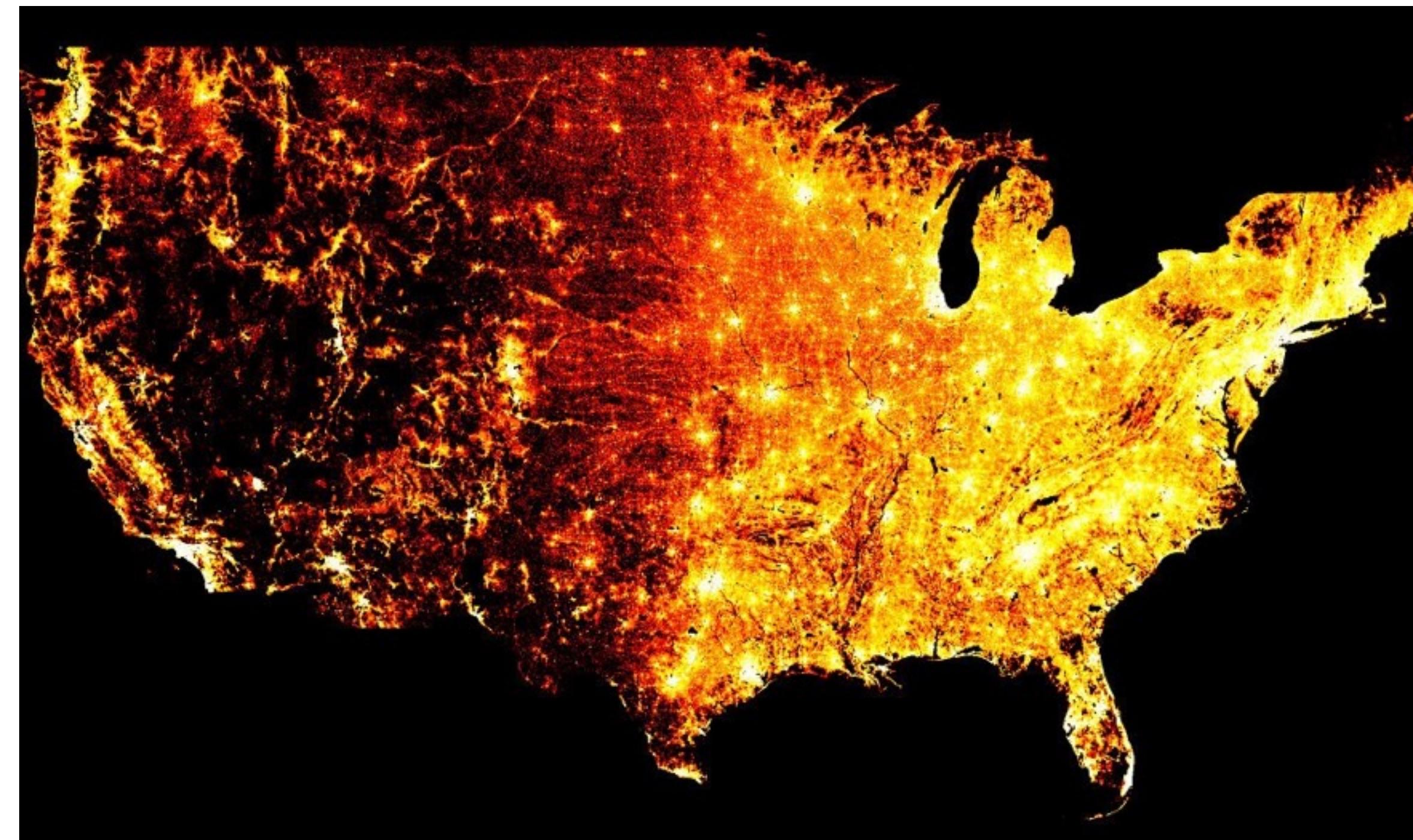
- Provides automatic, nearly parameter-free visualization of datasets
- Allows extensive customization of each step in the data-processing pipeline
- Supports automatic downsampling and re-rendering with Bokeh and the Jupyter notebook
- Works well with dask and numba to handle very large datasets in and out of core (with examples using billions of datapoints)

<https://github.com/bokeh/datashader>



NYC census data by race

Datashader



More examples:

<https://anaconda.org/jbednar/notebooks>

Dask: Scaling Data Analysis

One month CSV file ~ 2GBs

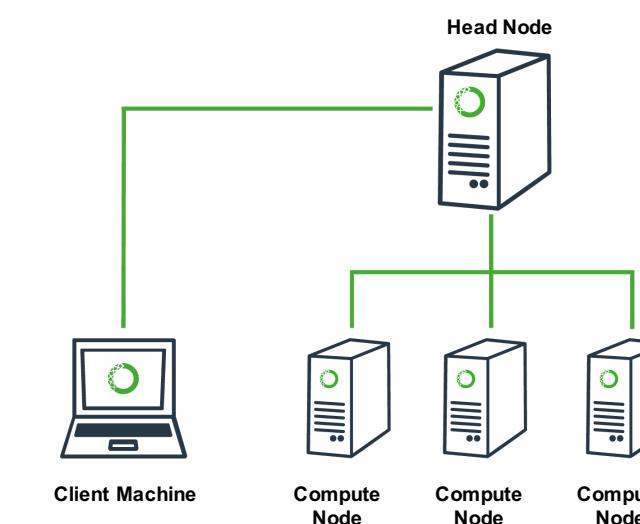


pandas
 $y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$



Scaling Data Analysis

Six month CSV file ~ 12GBs



Two years CSV files ~ 50GB

HDFS +

+ distributed

Dask Dataframes



```
>>> import pandas as pd
>>> df = pd.read_csv('iris.csv')
>>> df.head()

   sepal_length  sepal_width  petal_length
petal_width      species
0              5.1          3.5           1.4
0.2  Iris-setosa
1              4.9          3.0           1.4
0.2  Iris-setosa
```



```
>>> import dask.dataframe as dd
```

```
>>> ddf = dd.read_csv('*csv')
>>> ddf.head()

   sepal_length  sepal_width  petal_length
petal_width      species
0              5.1          3.5           1.4
0.2  Iris-setosa
1              4.9          3.0           1.4
0.2  Iris-setosa
2              4.7          3.2           1.3
-setosa
4.6          3.1           1.5
-setosa
5.0          3.6           1.4
-setosa
```

Dask dataframes look and feel like pandas dataframes, but operate on datasets larger than memory using multiple threads

```
>>> max_sepal_length_setosa = df[df.species ==
'setosa'].sepal_length.max()
5.799999999999999
```

```
>>> d_max_sepal_length_setosa = ddf[ddf.species ==
'setosa'].sepal_length.max()
>>> d_max_sepal_length_setosa.compute()
5.799999999999999
```

Distributed

Distributed is a lightweight library for distributed computing in Python.
It extends dask APIs to moderate sized clusters.

```
>>> from distributed import Executor, hdfs, progress
>>> e = Executor('127.0.0.1:8786')
>>> e
<Executor: scheduler=127.0.0.1:8786 workers=64 threads=64>

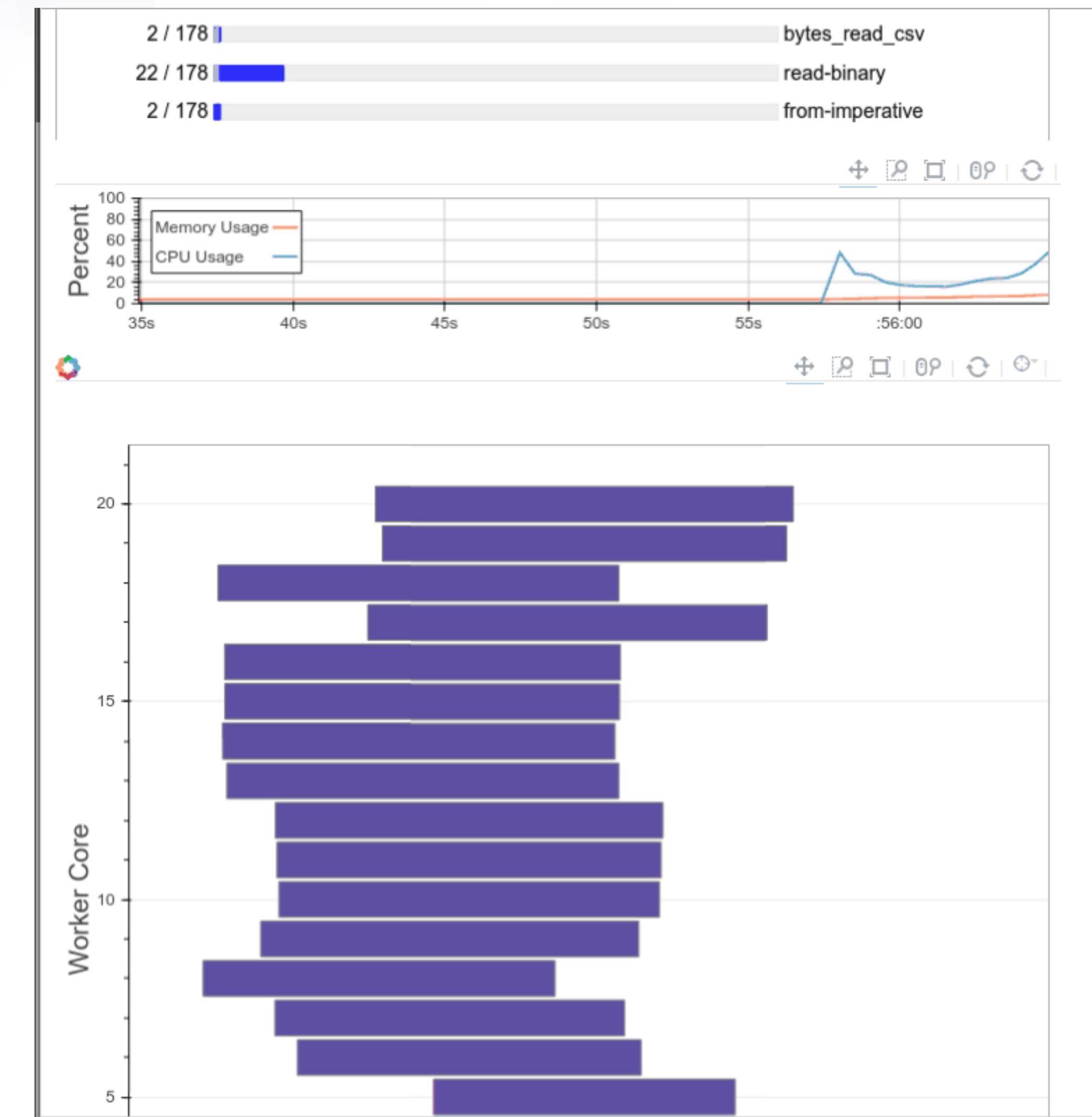
>>> nyc2014 = hdfs.read_csv('/nyctaxi/2014/*.csv',
...                         parse_dates=['pickup_datetime', 'dropoff_datetime'],
...                         skipinitialspace=True)

>>> nyc2015 = hdfs.read_csv('/nyctaxi/2015/*.csv',
...                         parse_dates=['tpep_pickup_datetime', 'tpep_dropoff_datetime'])

>>> nyc2014, nyc2015 = e.persist([nyc2014, nyc2015])
```

Web UI

Dask.distributed includes a web interface to help deliver information about the current state of the network helps to track progress, identify performance issues, and debug failures over a normal web page in real time.

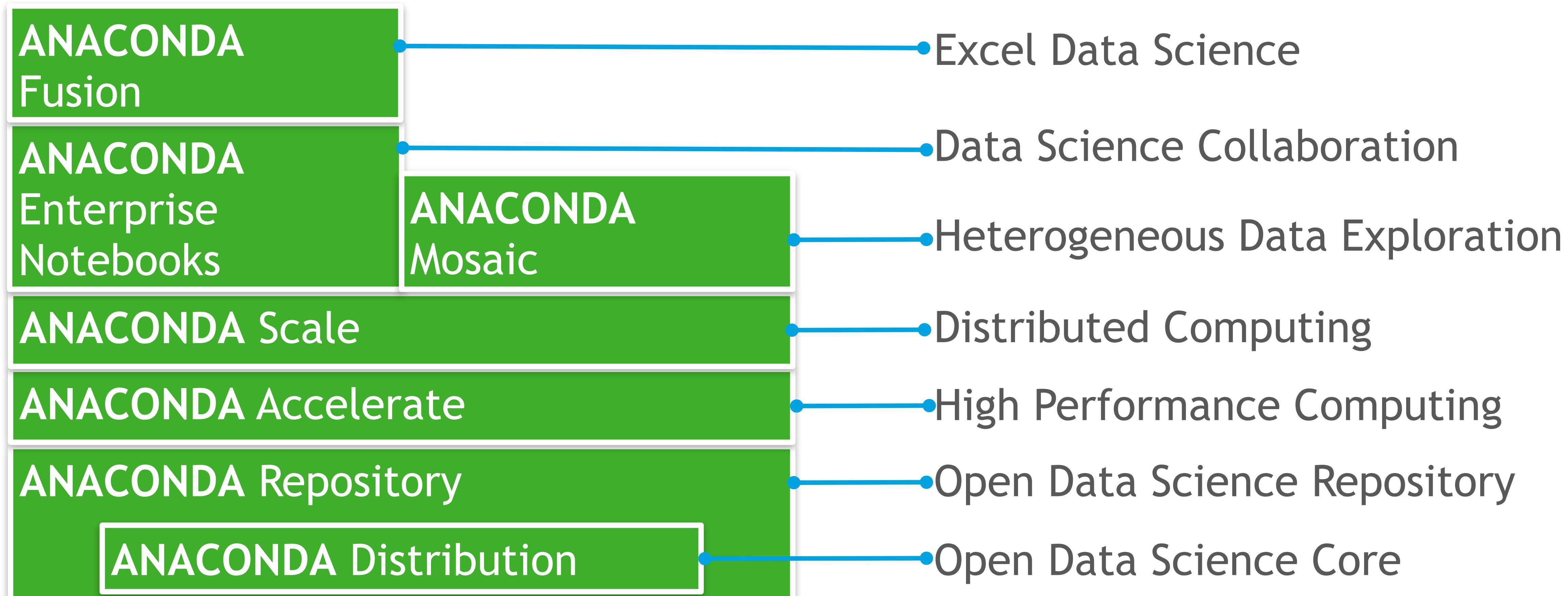


ANACONDA ENTERPRISE PLATFORM



ANACONDA®

ANACONDA platform



Profile

Packages View all (60)

- hdfs3 22 days and 8 hours ago
- libhdfs3 22 days and 9 hours ago
- tensorflow 22 days and 10 hours ago
- protobuf 22 days and 10 hours ago
- hs2client 22 days and 11 hours ago
- libhs2client 22 days and 11 hours ago
- toolchain 22 days and 18 hours ago
- xgboost 1 month and 4 days ago
- hdfs 1 month and 9 days ago
- fastavro 1 month and 9 days ago

Notebooks View all (48)

- yarn-blog-snippet 1 month and 9 days ago
- pyspark-nltk-json 2 months and 3 hours ago
- python-datasci-90m-start 2 months and 3 days ago
- anaconda-and-the-python-ecosystem 2 months and 16 days ago
- simple-pyspark-with-text-file 2 months and 29 days ago
- simple-nltk-on-hdfs 3 months and 23 hours ago
- simple-nltk-with-text-file 3 months and 23 hours ago
- high-performance-io 3 months and 4 days ago
- matlab_kernel 3 months and 5 days ago
- python-adv-cf16 6 months and 27 days ago

Contributor since Jul 31, 2013
Continuum Analytics Syracuse, NY

Organizations

Anaconda Repository

Log Out

Enter a filter term.

- iris_csv
- index_sq1
- nasdaq_lux
- nyse_lux
- nyc_taxi_hdf5
- all_stocks
- index_stocks
- tpo_orders_sq1
- GSPC_http
- cust_order
- vwap_index

df1

Operations: projection, merge, transform, relabel, concat, join, sort, select, distinct, rows, cols

Filter: x:tip_amount / x.passenger_count > 5

Plot Type: Shader, X: dropoff_x, Y: dropoff_y, App. Type: Sum

Agg. Field: passenger_count, Trans. Func.: Log, Dot size: 0.1

Plot: A scatter plot showing taxi pickup locations in New York City.

nyc_taxi_hdf5

Description: Open Dataset in New Notebook

URI: hdfstore://./databricks/examples/data/nyc_taxi.h5:dataset

Readme:

Creator: admin

Created: 16 days ago

Ancestors: nyc_taxi_hdf5

df1 Datashape:

Number of Rows:	10000
Columns:	
VendorID:	int64
tpep_pickup_datetime:	datetime
tpep_dropoff_datetime:	datetime
passenger_count:	int64
trip_distance:	float64
pickup_x:	float64
pickup_y:	float64
RateCodeID:	int64
store_and_fwd_flag:	string
dropoff_x:	float64
dropoff_y:	float64
payment_type:	int64
fare_amount:	float64
extra:	float64
mta_tax:	float64
tip_amount:	float64

Anaconda Mosaic

user

Projects (3)

- tanya / DeepLearningToKeepIceCreamCold (0 stars, 1 red dot)
- tanya / OnTime (4 stars, 1 green dot)
- tanya / RetailBankingTreasuriesForecast (0 stars, 1 red dot)

Contributing (2)

- simon / WeatherModel (0 stars, 1 red dot)

Top Tags

- python
- dask
- pandas
- bokeh
- blaze
- csv
- dataviz
- excel
- matplotlib
- odo

Anaconda Enterprise Notebooks

sec-demo - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do

Cells: H11, A1, B1, C1, D1, E1, F1, G1, I1, J1, K1

Data:

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Input:										
3	AAMC	AAPL									
4	AAMC	GOOG									
5	ZYNE	IBM									
6	Output:										
7	25.77178	69.285134	46.04256	4/1/2016							
8	24.50673	72.783928	44.20794	4/4/2016							
9	26.21665	68.645852	45.60325	4/5/2016							
10	28.11912	69.338864	40.61291	4/6/2016							
11	28.23309	70.659199	42.54809	4/7/2016							
12	27.94107	67.957801	37.59995	4/8/2016							
13	29.17468	72.767751	39.31971	4/11/2016							
14	27.79955	73.51119	40.02697	4/12/2016							
15	27.4695	69.990418	48.01223	4/13/2016							
16	25.00897	69.1234	41.30214	4/14/2016							
17	22.77024	63.676422	38.57985	4/15/2016							
18	28.18508	73.546795	40.32982	4/18/2016							
19	26.56918	69.274517	33.49575	4/19/2016							
20	23.06301	68.283159	37.26487	4/20/2016							
21	24.34234	70.289771	39.6918	4/21/2016							
22	25.63651	66.722213	40.87819	4/22/2016							
23	24.79597	69.96168	39.68989	4/23/2016							
24	24.87742	65.047669	39.45226	4/26/2016							
25	24.16071	71.007884	42.84937	4/27/2016							
26	23.94836	68.784003	42.25338	4/28/2016							
27	25.40819	71.347256	41.246	4/29/2016							
28	23.27046	73.082881	46.65864	5/2/2016							
29	24.37497	75.995785	46.22454	5/3/2016							
30	24.77613	74.637554	40.79139	5/4/2016							
31	22.50891	73.06938	35.36912	5/5/2016							
32	25.30971	73.471902	39.66066	5/6/2016							
33	24.44356	76.496492	37.40791	5/9/2016							
34	25.87467	79.727111	42.12388	5/10/2016							

Anaconda Fusion

CodeSheets

DATA

tech_tickers

make_tickchart

Magic Mode

Expert Mode

INPUTS

symbols: tech_tickers

metric: Oddlot Rate

Oddlot Rate by Ticker Symbol

Anaconda Fusion

Challenges revisited

- Manage reproducible Data Science environments
- Distribute Data Science assets
- Get diverse data scientists (languages, tools, data models, deliverables...) to collaborate effectively
- Enable Data Scientists to easily leverage Big Data technologies
- Deploy data science assets into production applications
- Share insights with decision makers

Learn more

<https://www.continuum.io/>

- Whitepapers: <https://www.continuum.io/whitepapers>
- Webinars: <https://www.continuum.io/webinars>
- Presentations: <https://www.continuum.io/presentations>
- Videos: <https://www.continuum.io/videos>

Thank you!

christine.doig@continuum.io
sales@continuum.io

Twitter: @ch_doig

