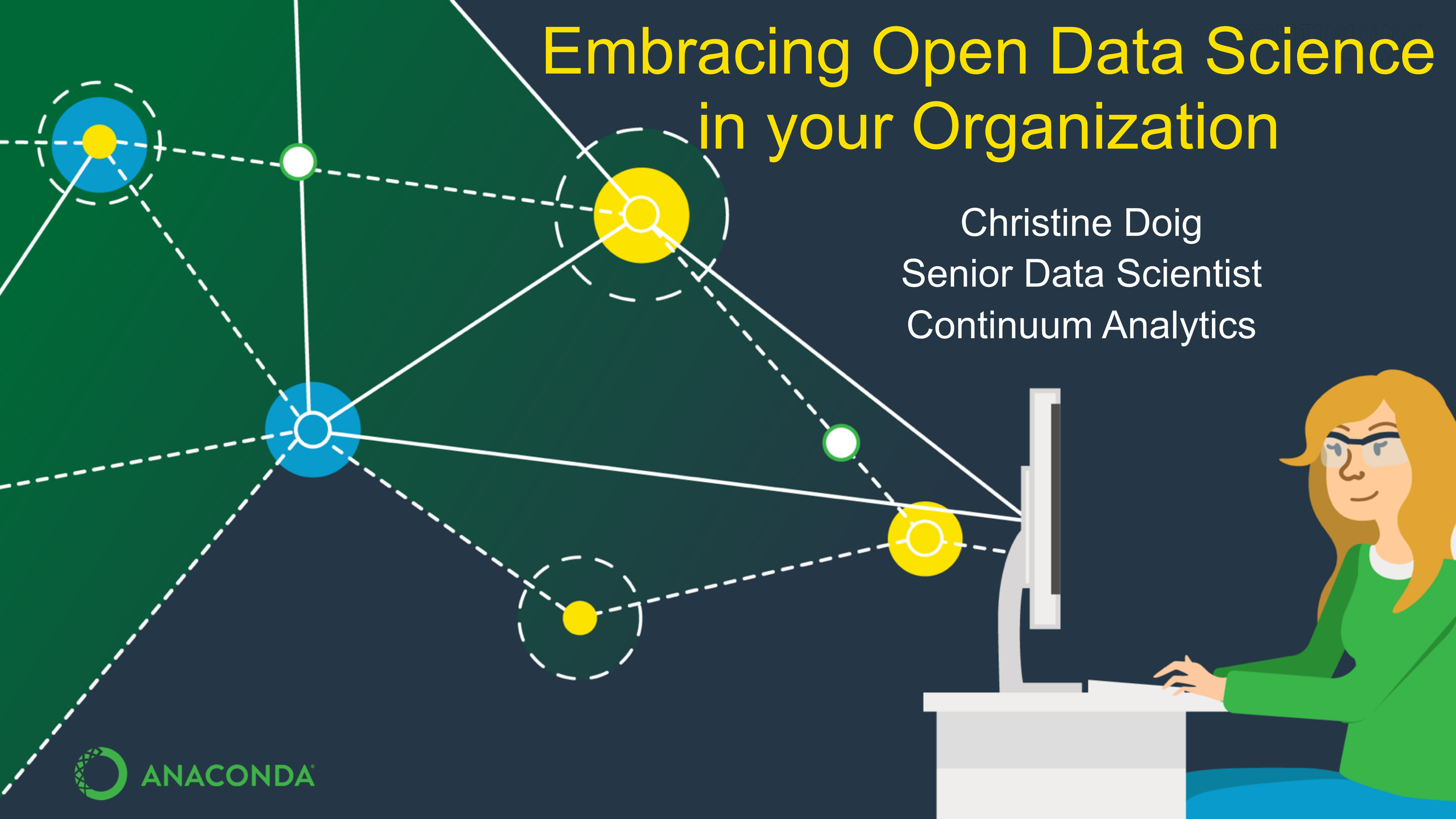
Embracing Open Data Science in your Organization

Christine Doig
Senior Data Scientist
Continuum Analytics



ANACONDA®

CONTINUUM® ANALYTICS

Agenda

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

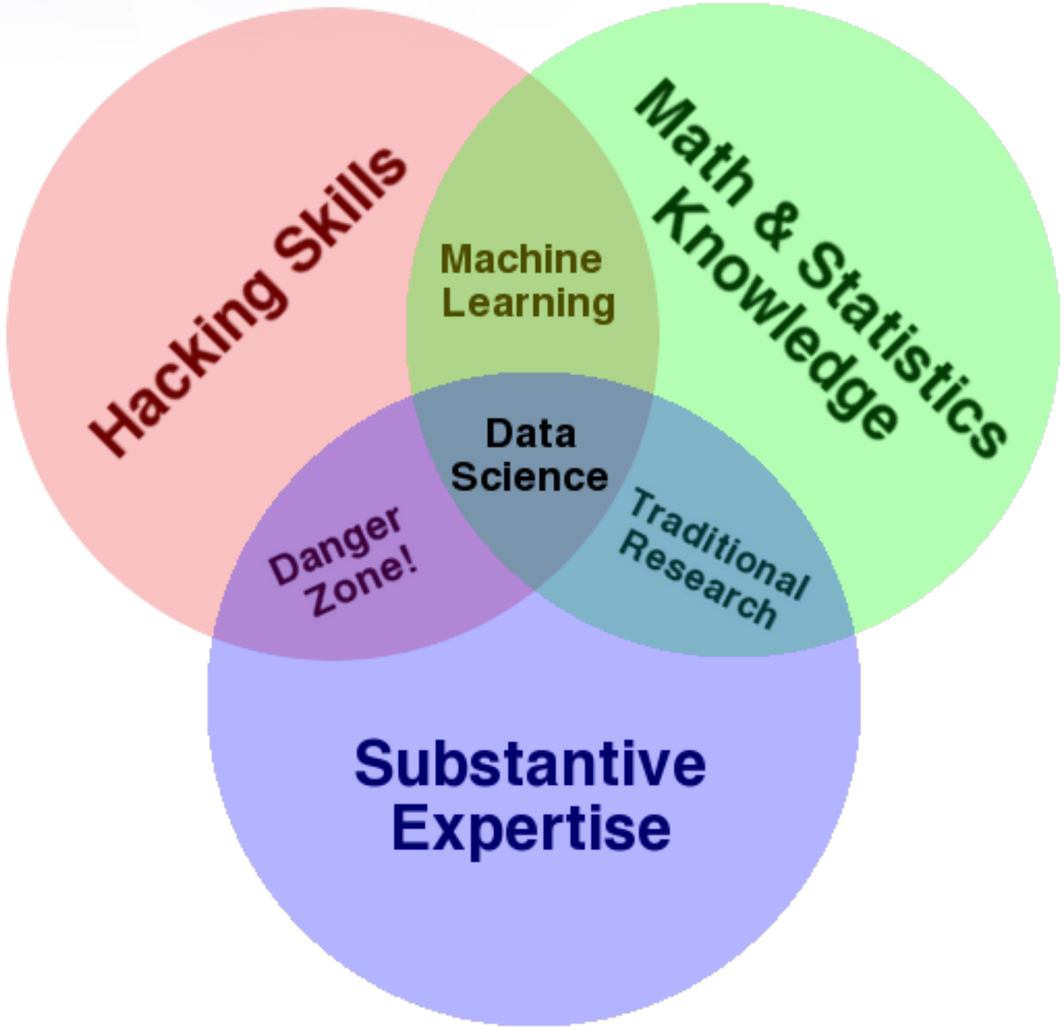
ANACONDA®

CONTINUUM®
ANALYTICS

INTRODUCTION TO DATA SCIENCE

ANACONDA®

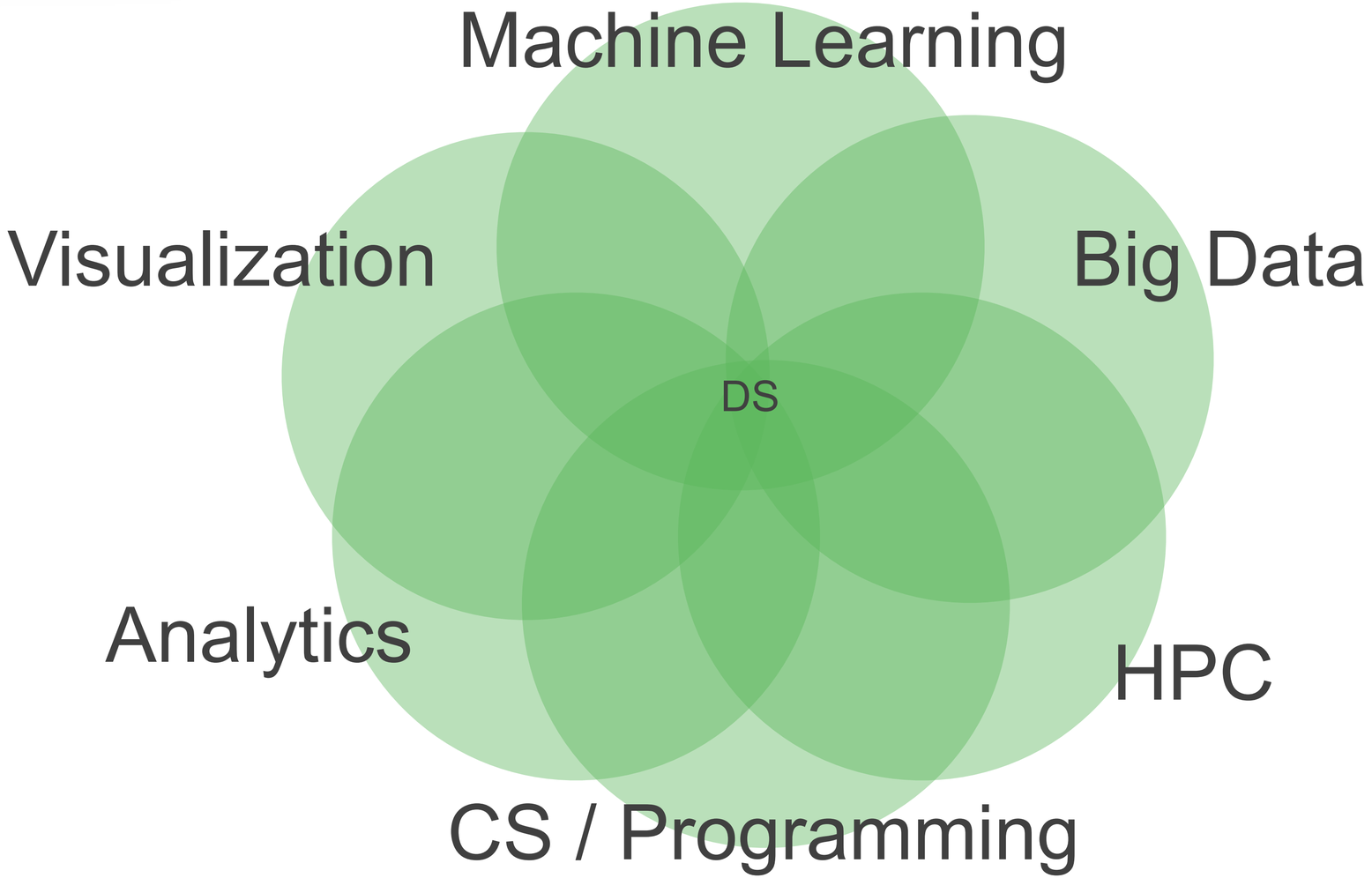
The Data Science Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

CONTINUUM® ANALYTICS

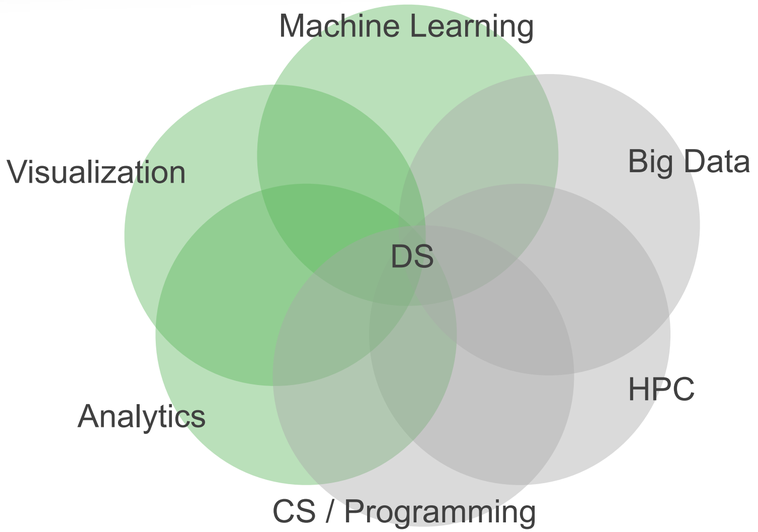
The Data Science Venn Diagram Revisited

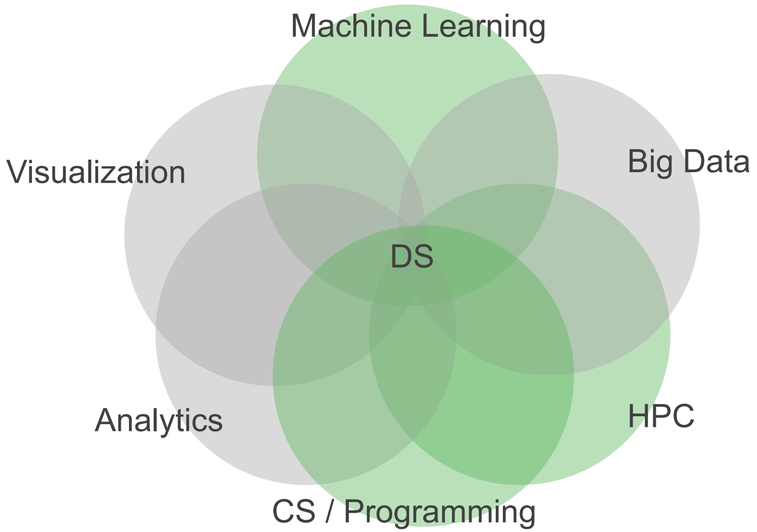


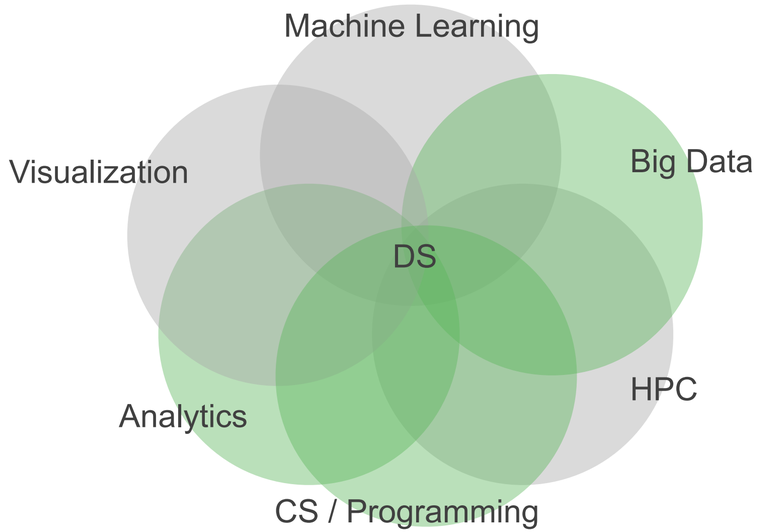
ANACONDA®

CONTINUUM ANALYTICS

Data Scientist come with different skills and backgrounds



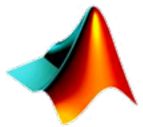




Statistician / Analyst

Research / Computational Scientist

Developer / Engineer



CONTINUUM ANALYTICS

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?

ANACONDA

CONTINUUM ANALYTICS

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

ANACONDA

© 2016 Continuum Analytics- Confidential & Proprietary

Open Data Science
Vibrant and Growing Community

Python Community
30M+

Packages in Anaconda
720+

ANACONDA Downloads*
3M+

R Community
16M+

Spark Python Usage
50%+

© 2016 Continuum Analytics- Confidential & Proprietary

* As of Dec 2015. Another 2.7M download YTD

CONTINUUM ANALYTICS

Open Data Science means…

Availability

Innovation

Interoperability

Transparency

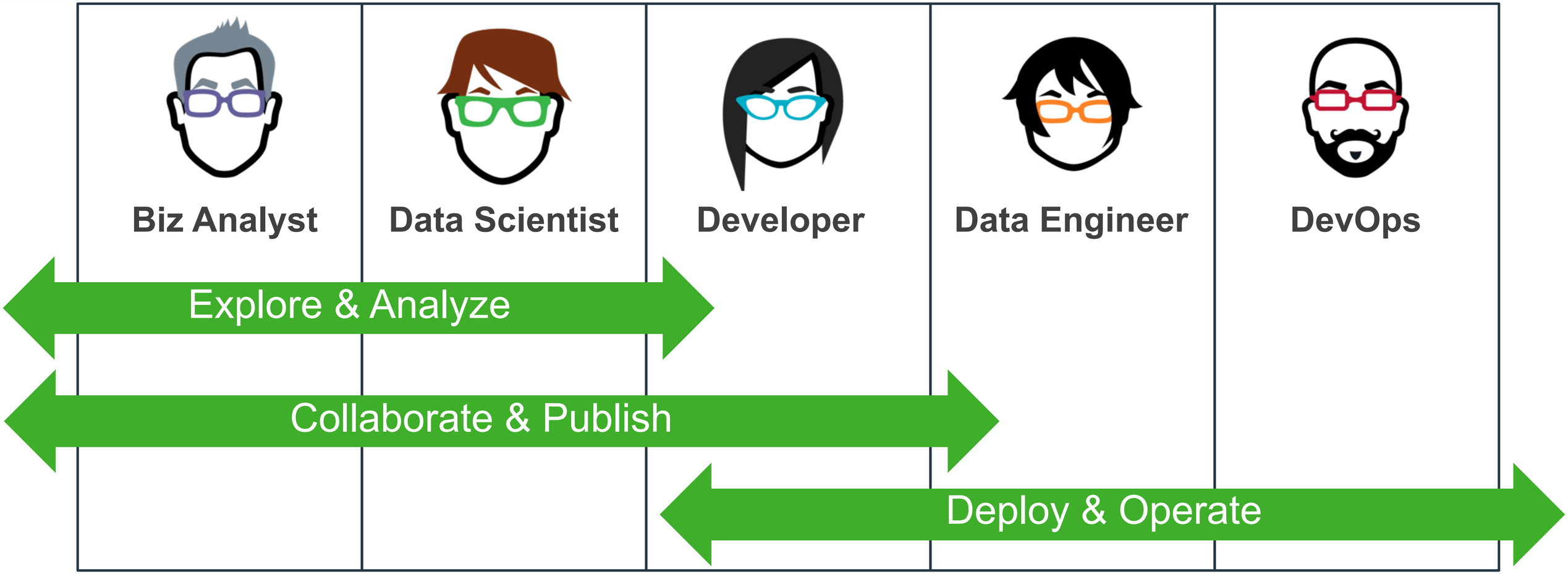
For everyone in the data science team

OPEN DATA SCIENCE is the
FOUNDATION TO MODERNIZATION

ANACONDA

© 2016 Continuum Analytics- Confidential & Proprietary

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



© 2016 Continuum Analytics- Confidential & Proprietary

Data Science assets

CONTINUUM ANALYTICS

Biz Analyst

Spreadsheets
Reports
Presentations

Data Scientist

Notebooks
Scripts
Visualizations

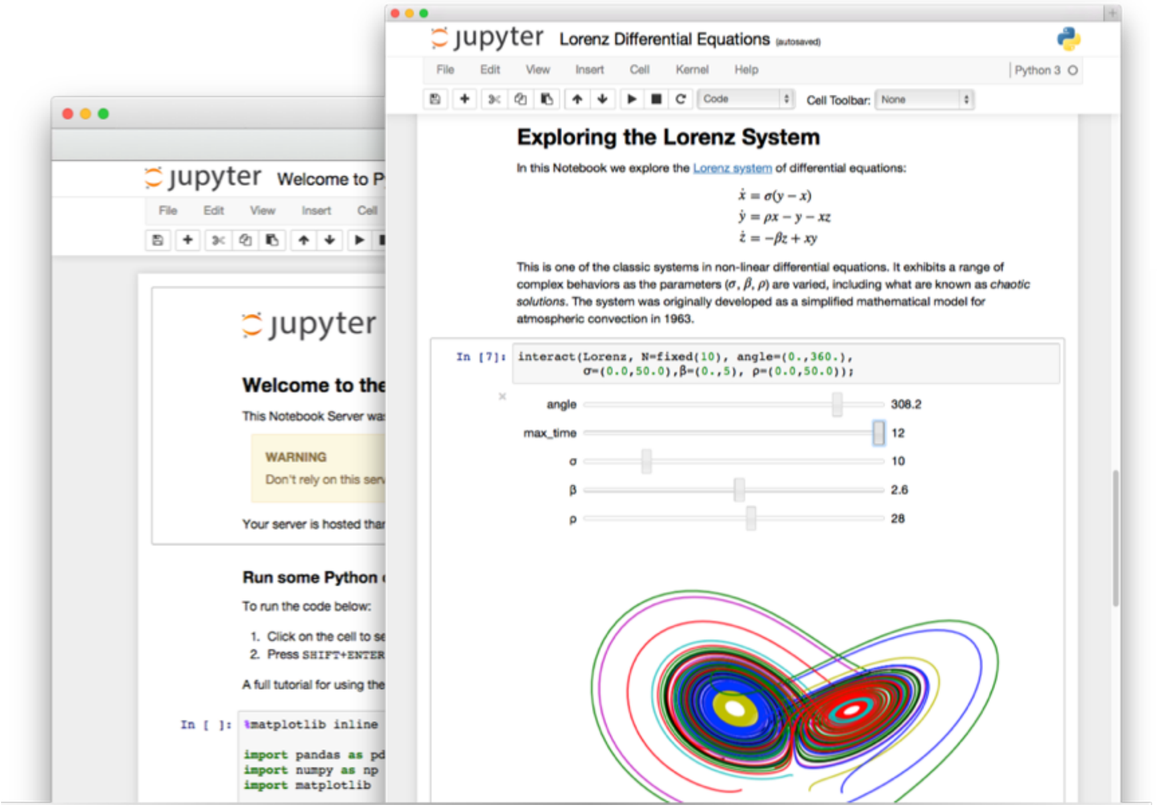
Developer

Software packages
Web applications

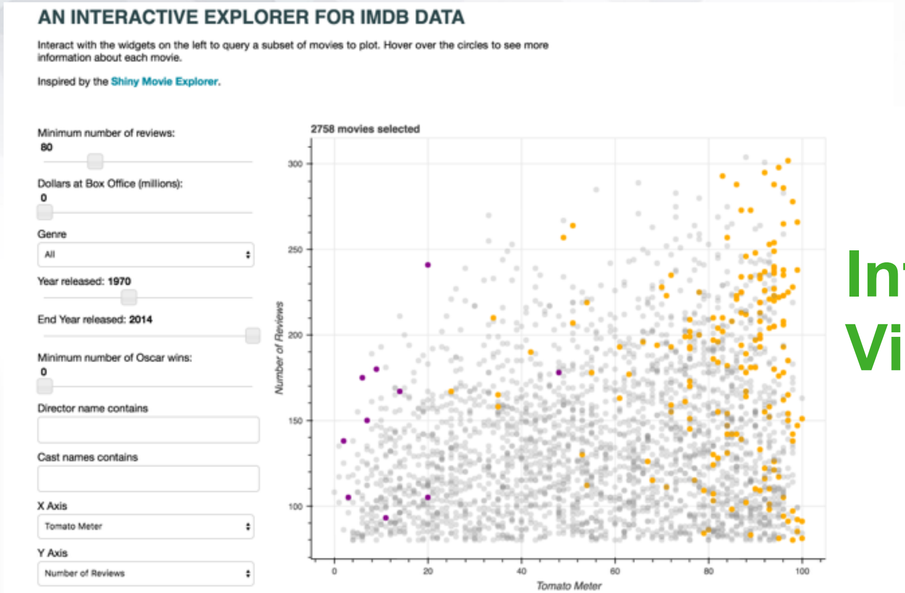
ANACONDA

assets

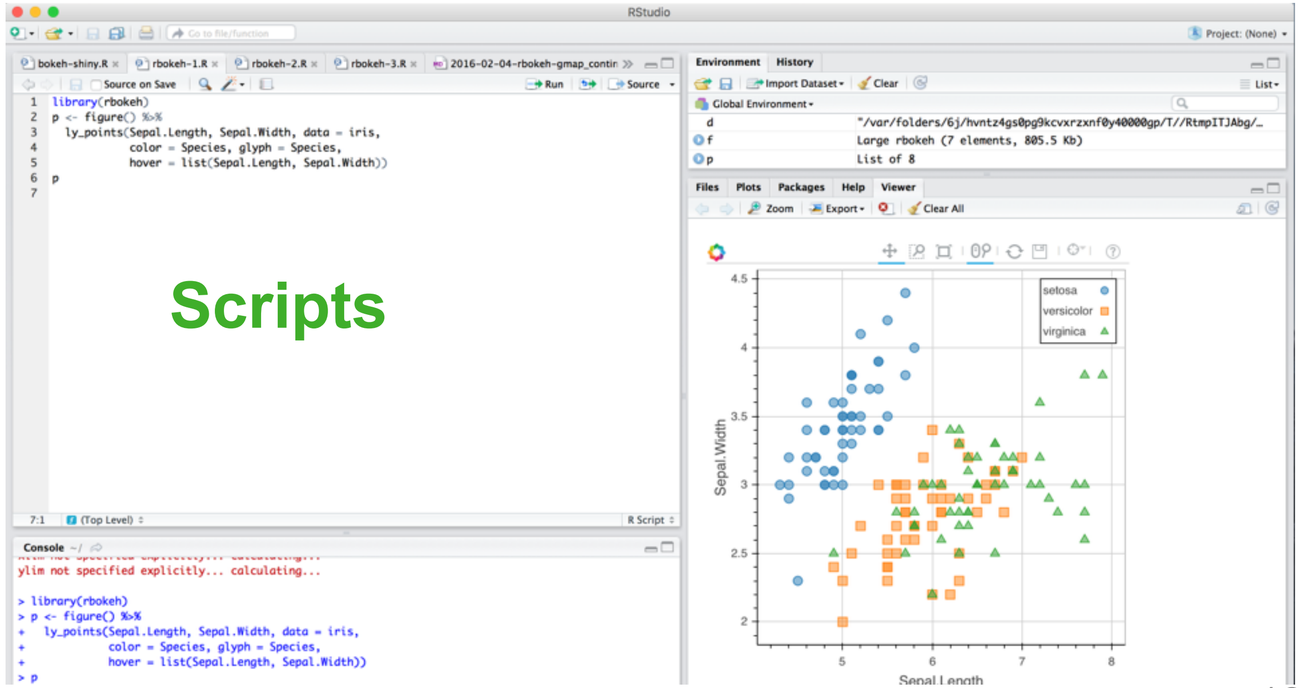
Data Scientist



Notebooks



Interactive Data Visualizations



Scripts

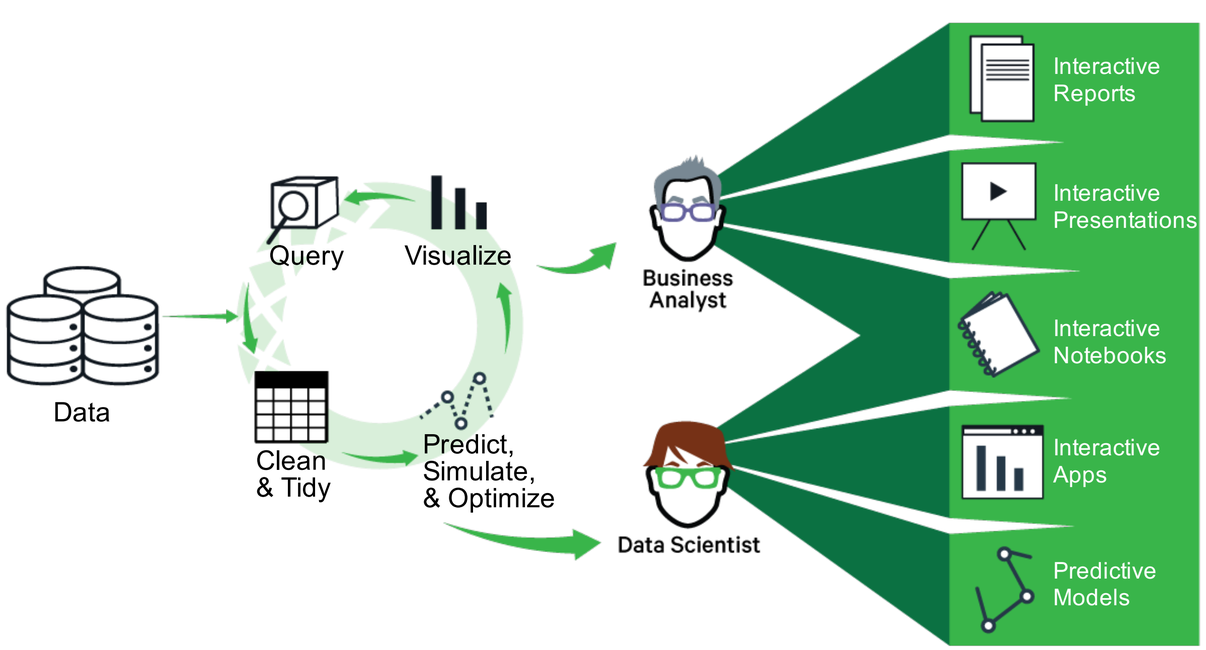
ANACONDA®

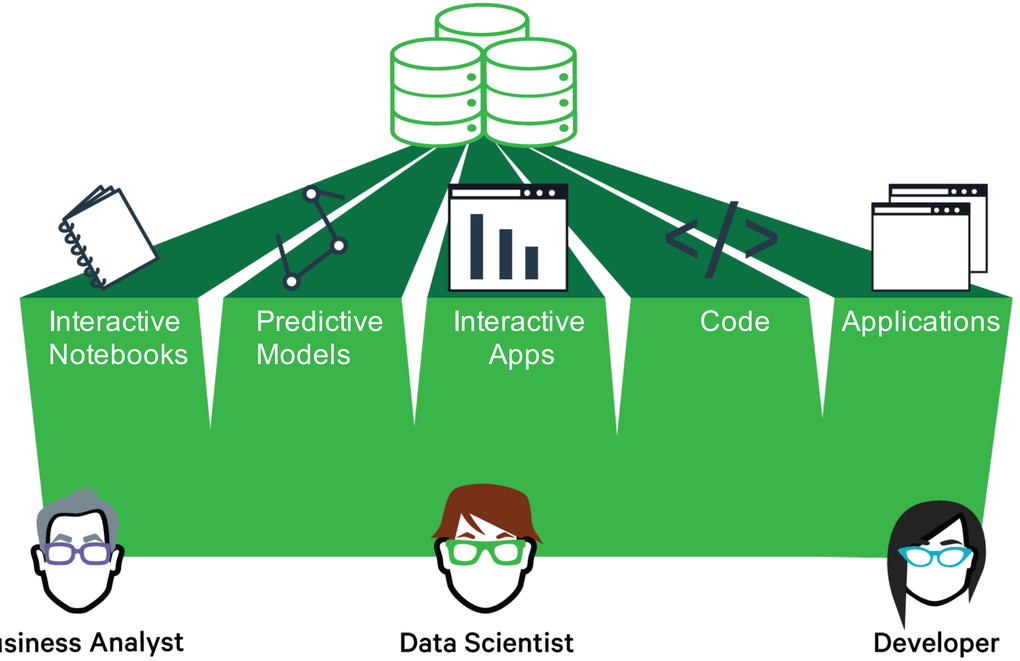
Data Science workflows

CONTINUUM ANALYTICS

Explore & Analyze

Collaborate & Publish



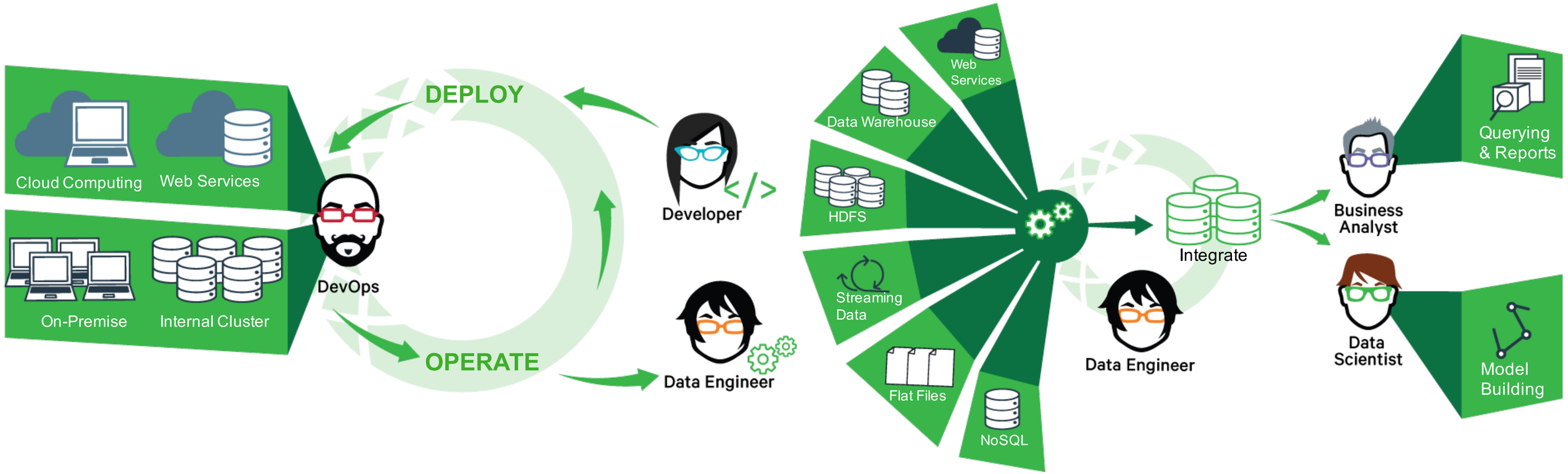


ANACONDA

Data Science workflows

CONTINUUM ANALYTICS

Deploy & Operate



ANACONDA

CONTINUUM® ANALYTICS

DATA SCIENCE CHALLENGES IN ORGANIZATIONS

ANACONDA®

CONTINUUM ANALYTICS

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

ANACONDA

CONTINUUM ANALYTICS

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

CONTINUUM®
ANALYTICS

ANACONDA DISTRIBUTION

ANACONDA®

The Distribution for Data Science

CONTINUUM ANALYTICS

Machine Learning

Visualization

Big Data

Bokeh

Numba

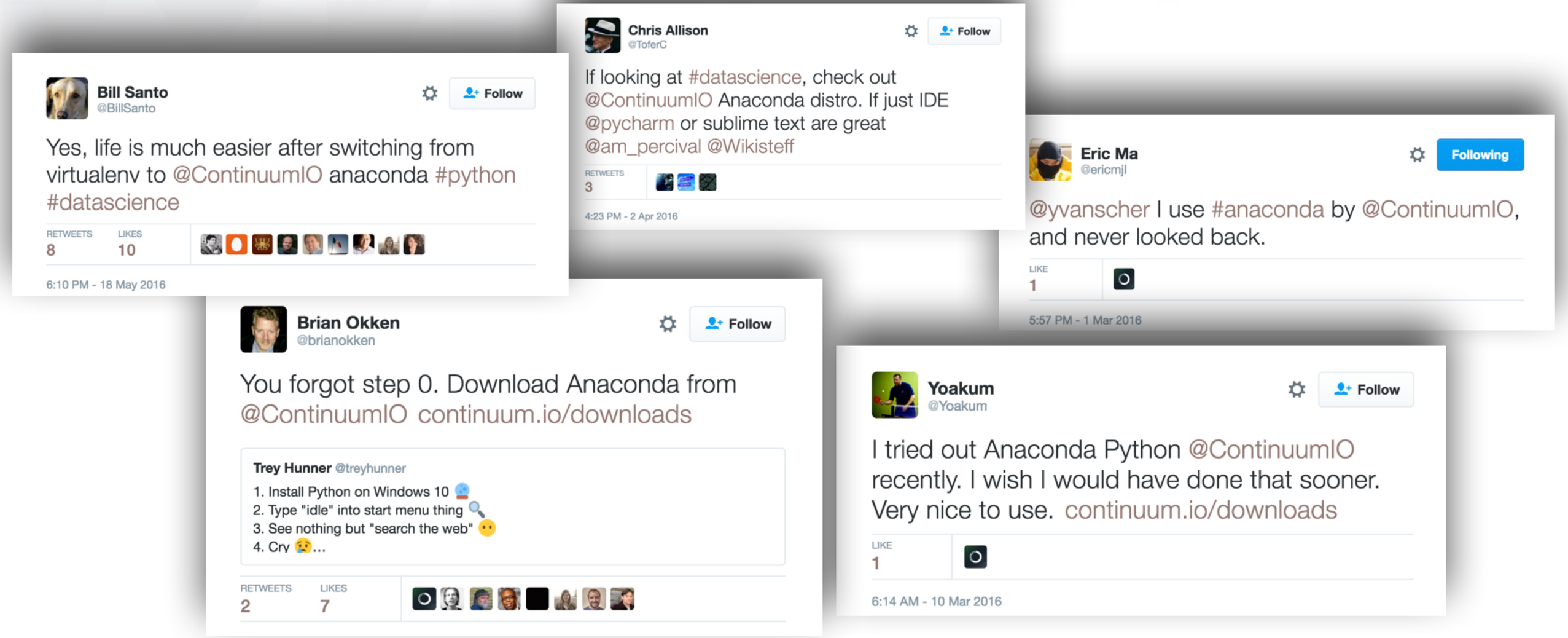
Analytics

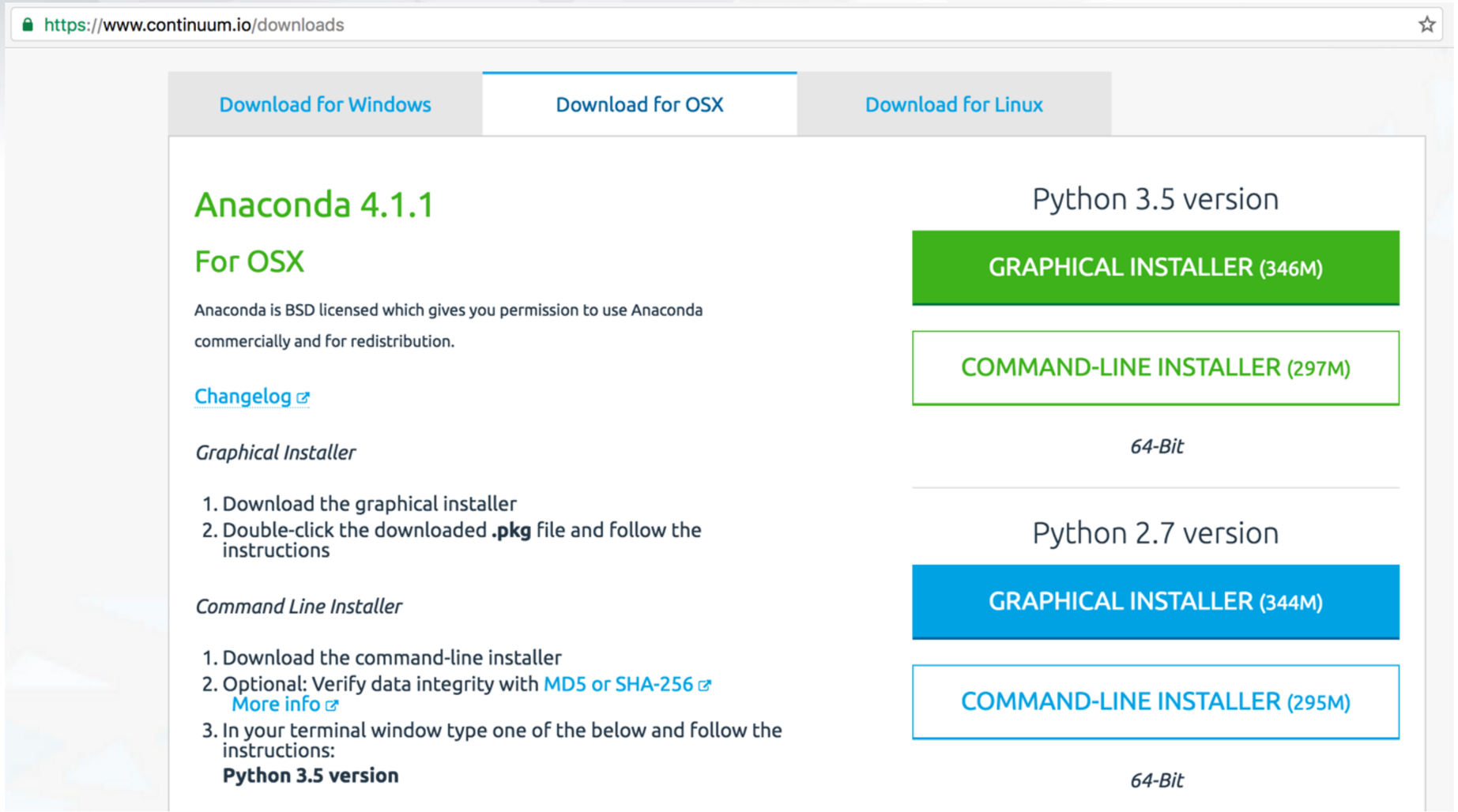
Scientific computing

Blaze

CS / Programming

… with an amazing community!





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

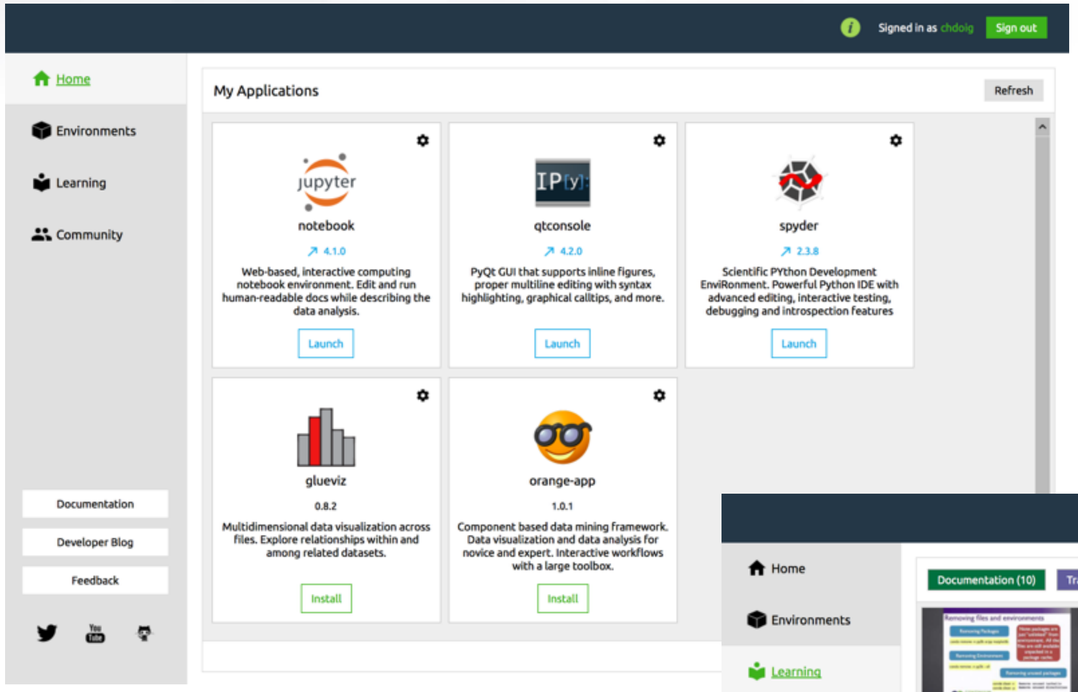
conda

Miniconda

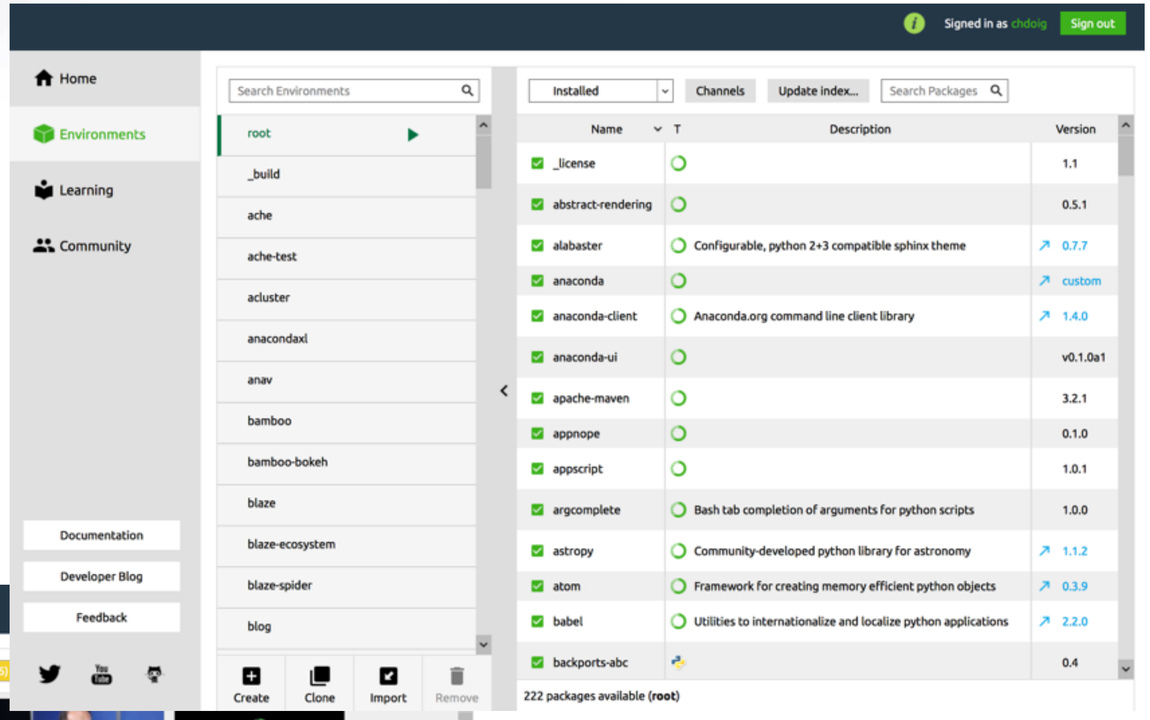
- **Anaconda distribution**: Python distribution that includes 150+ packages for data science (in the installer)
- **Miniconda**: Lightweight version of Anaconda, with just Python and conda.
- **Anaconda Cloud:** Cloud service to host and share public (free) and private data science assets
- **Anaconda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

Anaconda Navigator

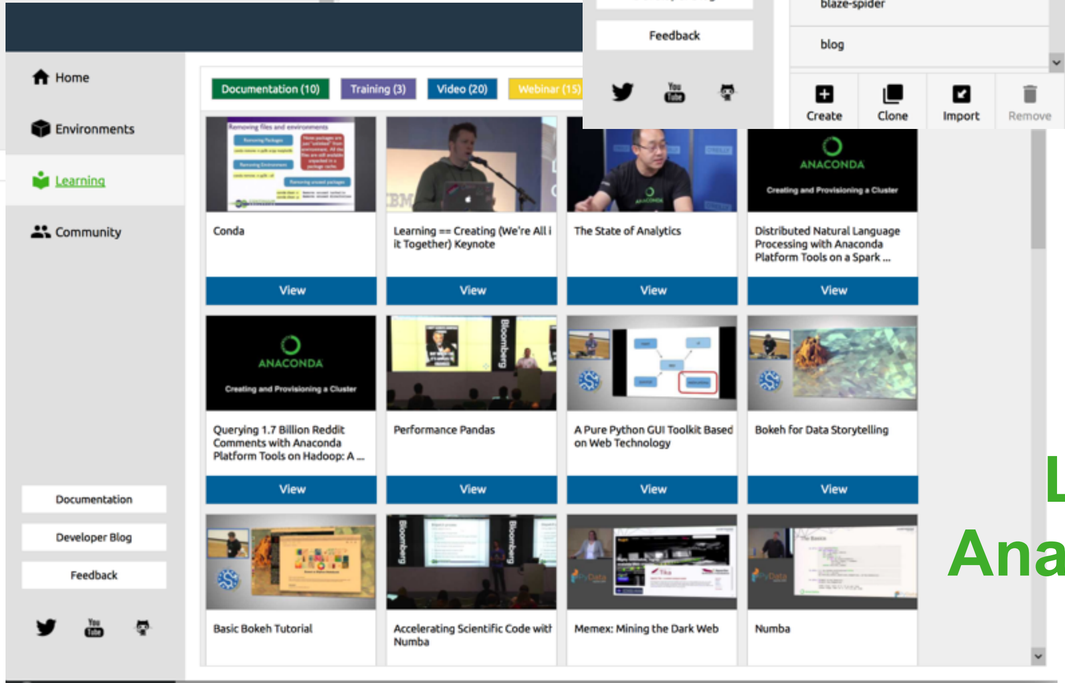
CONTINUUM ANALYTICS



Launch applications



Manage environments



Learn about the Anaconda community

ANACONDA

- **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

CONTINUUM ANALYTICS

CONDA

```
$ 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
dependecies:
  - 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

ANACONDA

CONTINUUM ANALYTICS

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

PYTHON

conda

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

CONTINUUM®
ANALYTICS

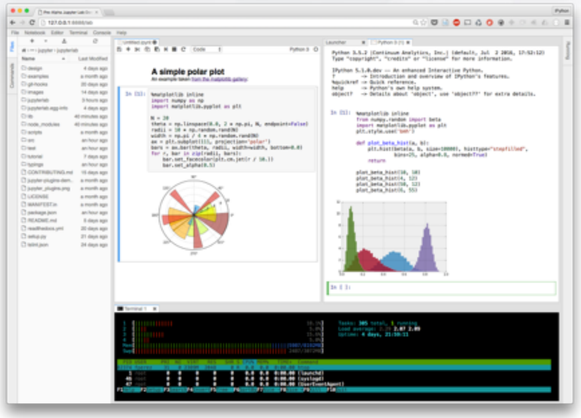
ANACONDA
COMMUNITY INNOVATION

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

Continuum Analytics contributions to the Python ODS ecosystem

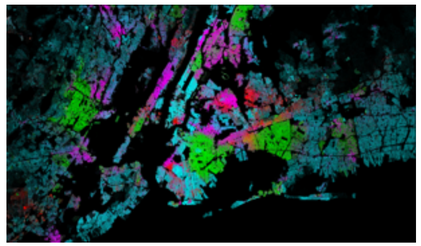


JupyterLab

- Next generation Data Science IDE

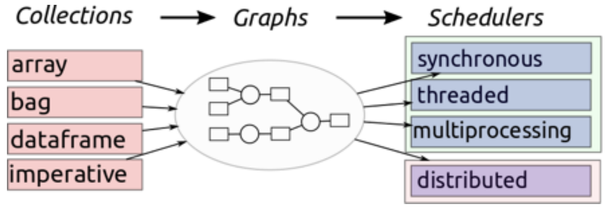
Bokeh

- Web interactive data visualizations (no JS)



Datashader

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



Dask

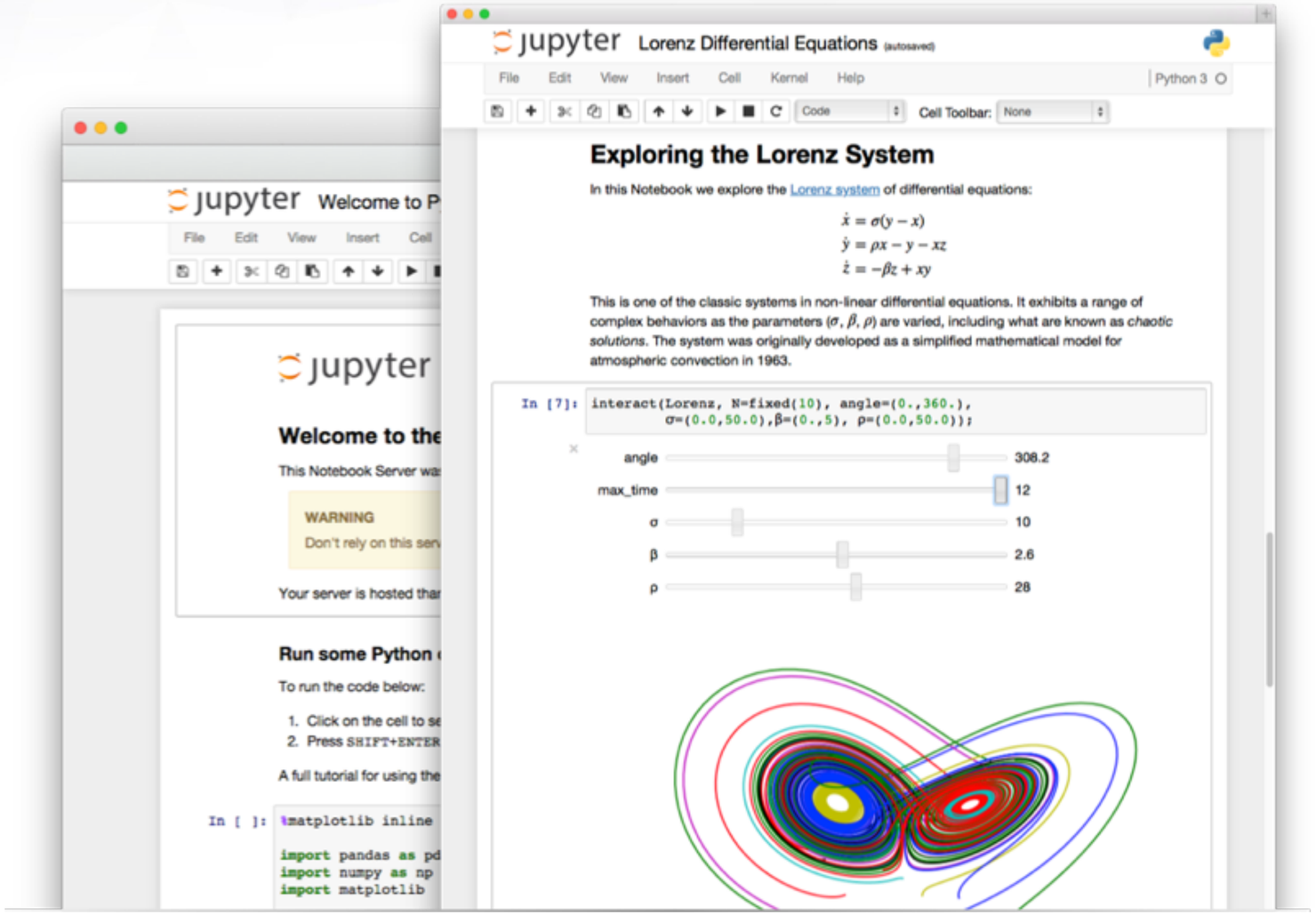
- Parallel computing framework

ANACONDA

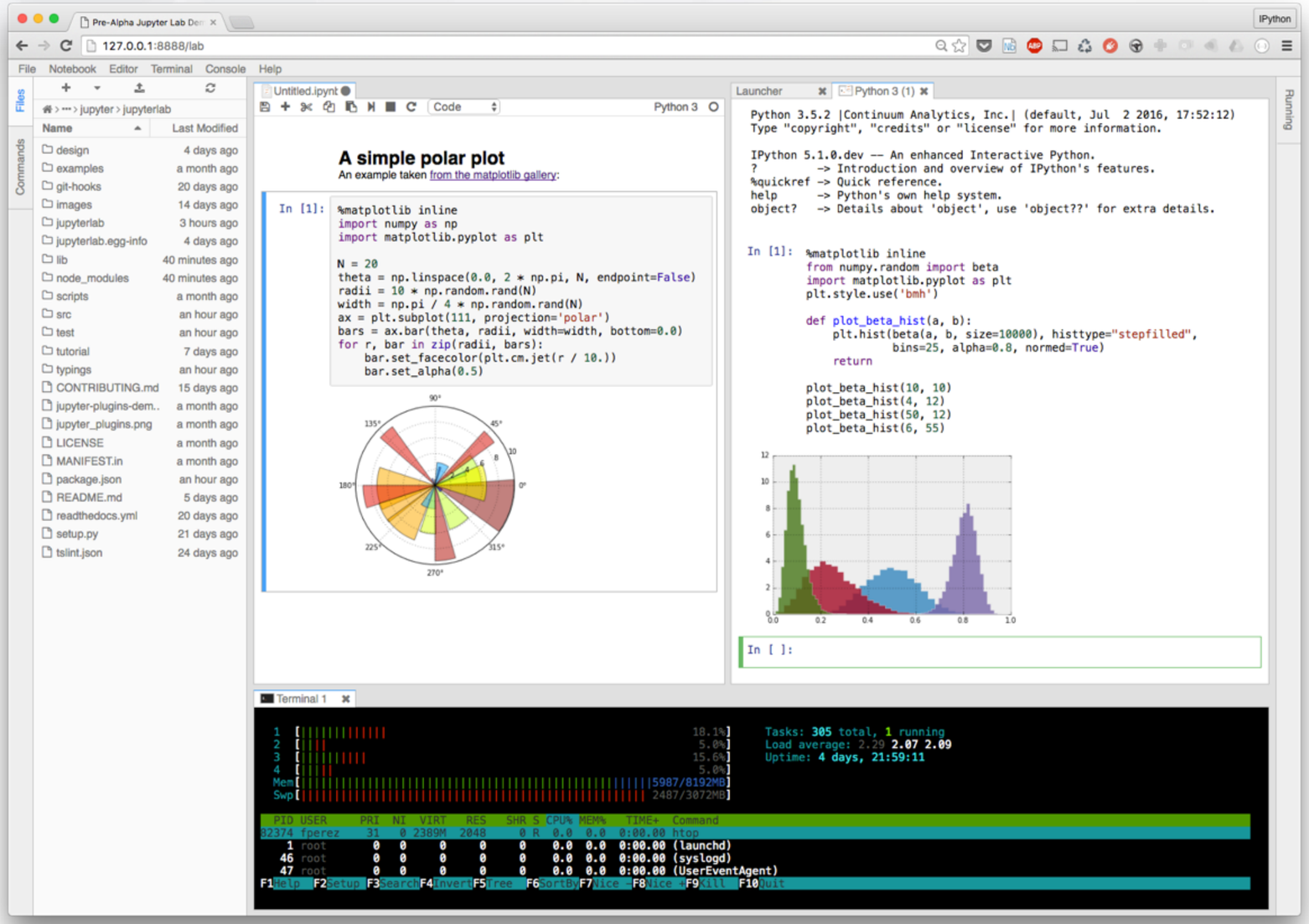
Jupyter Notebook

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

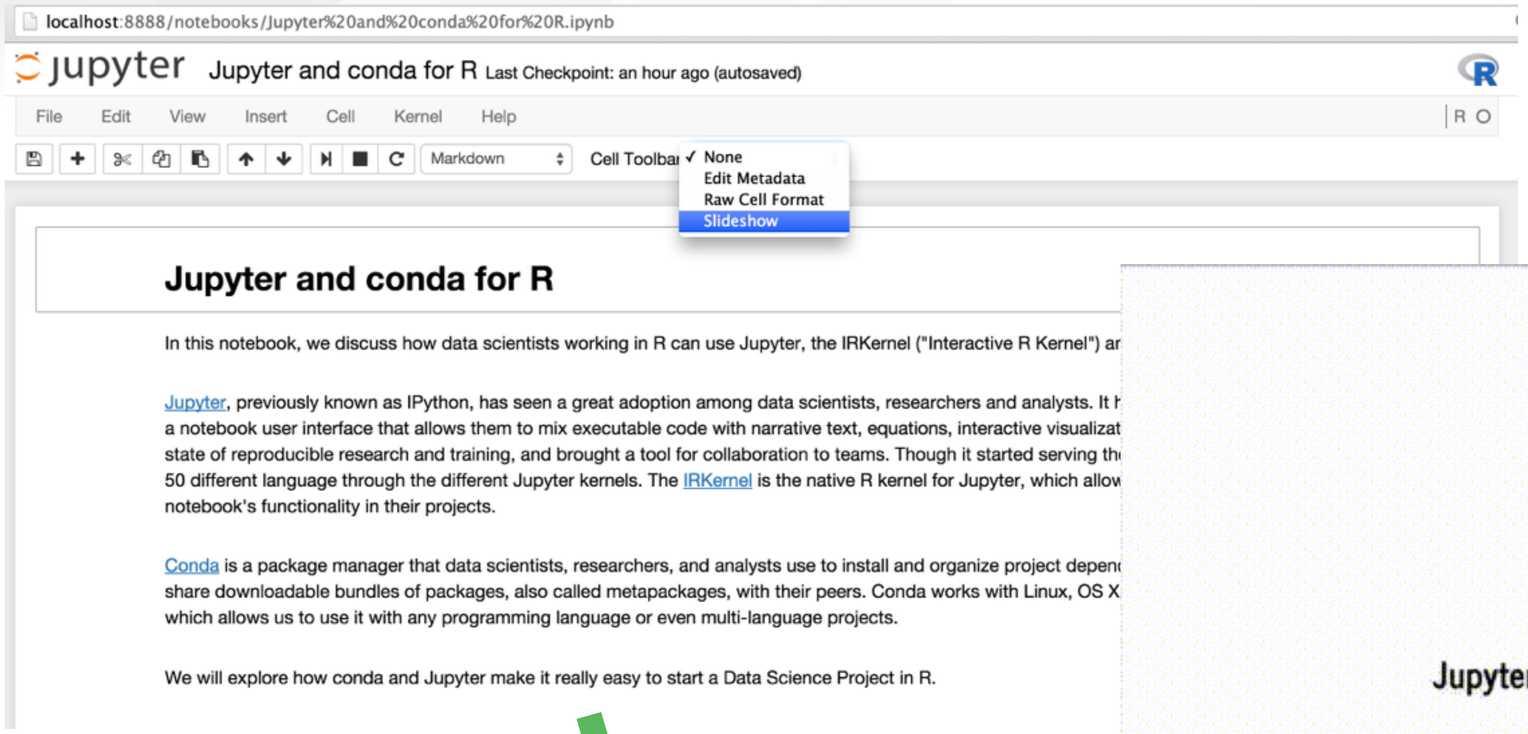
```
$ jupyter notebook
```



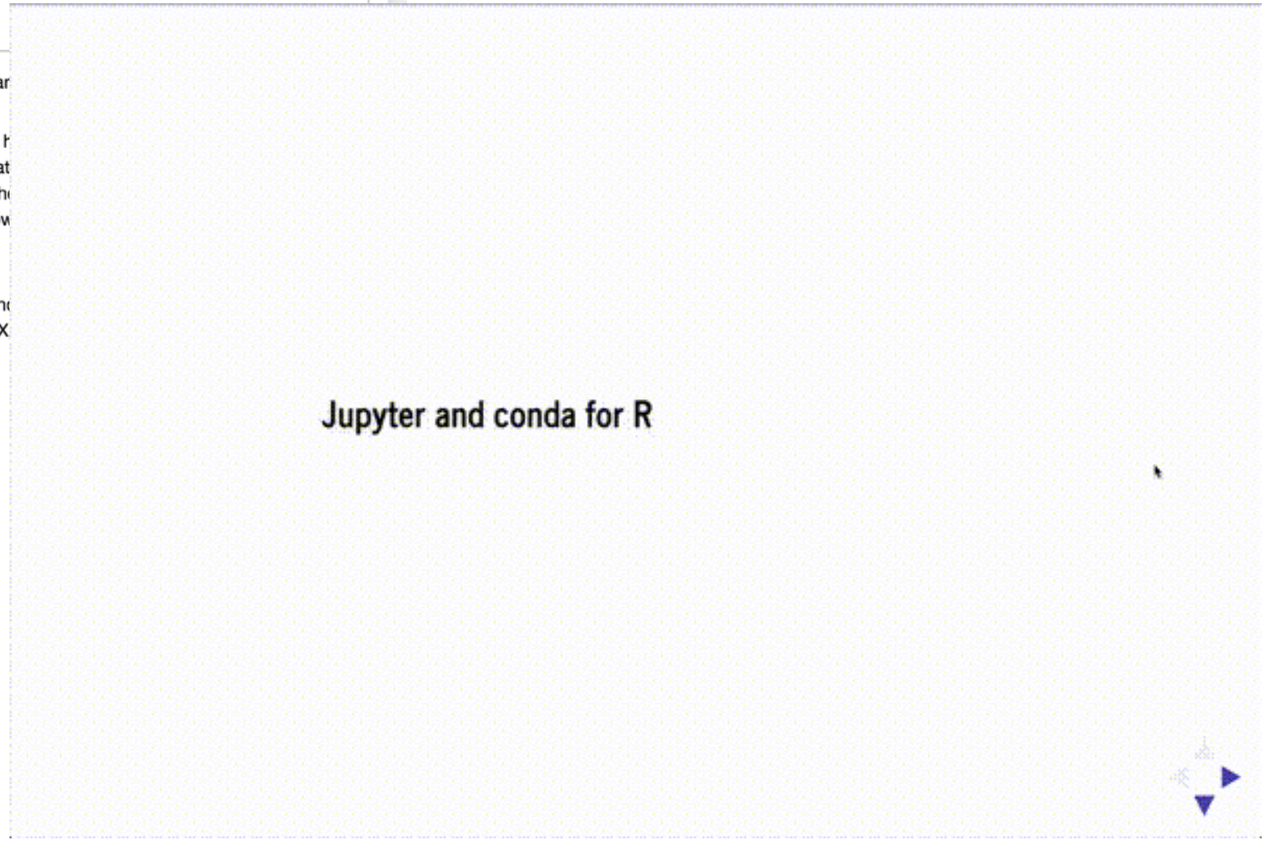
JupyterLab: the next generation



Sharing insights with decision makers



From text, code and visualizations directly to slides

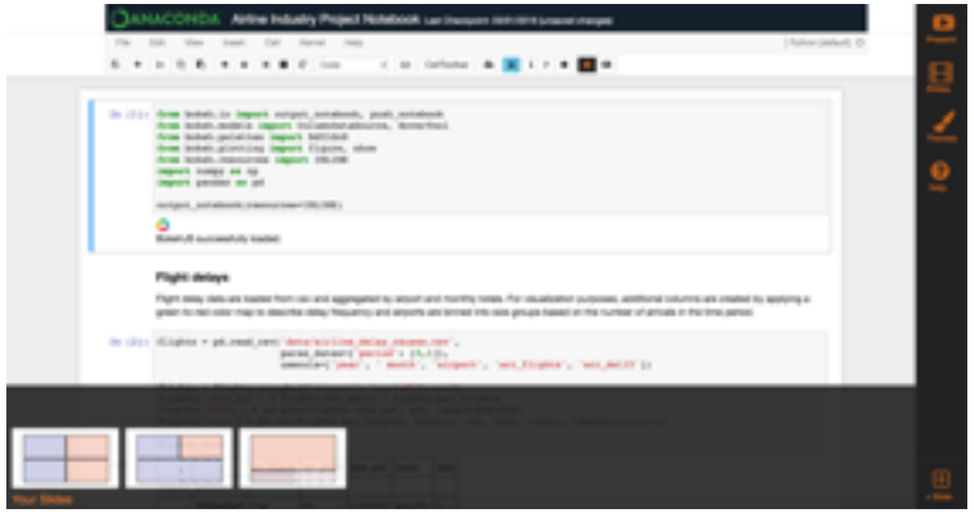


Jupyter: Extensions - nbpresent

remix your Jupyter Notebooks as interactive slideshows with a UI editor

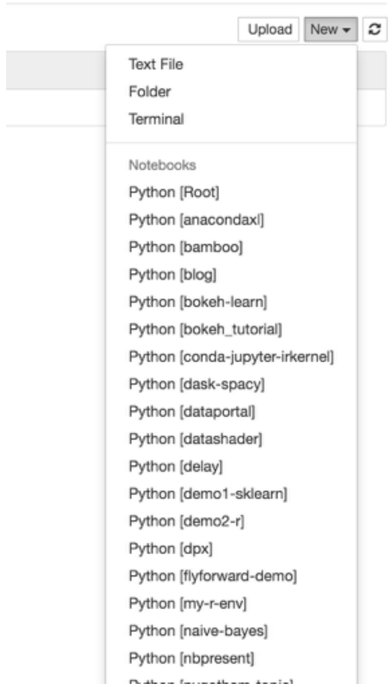
- Edit slides, layout and themes

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

jupyter notebook
```

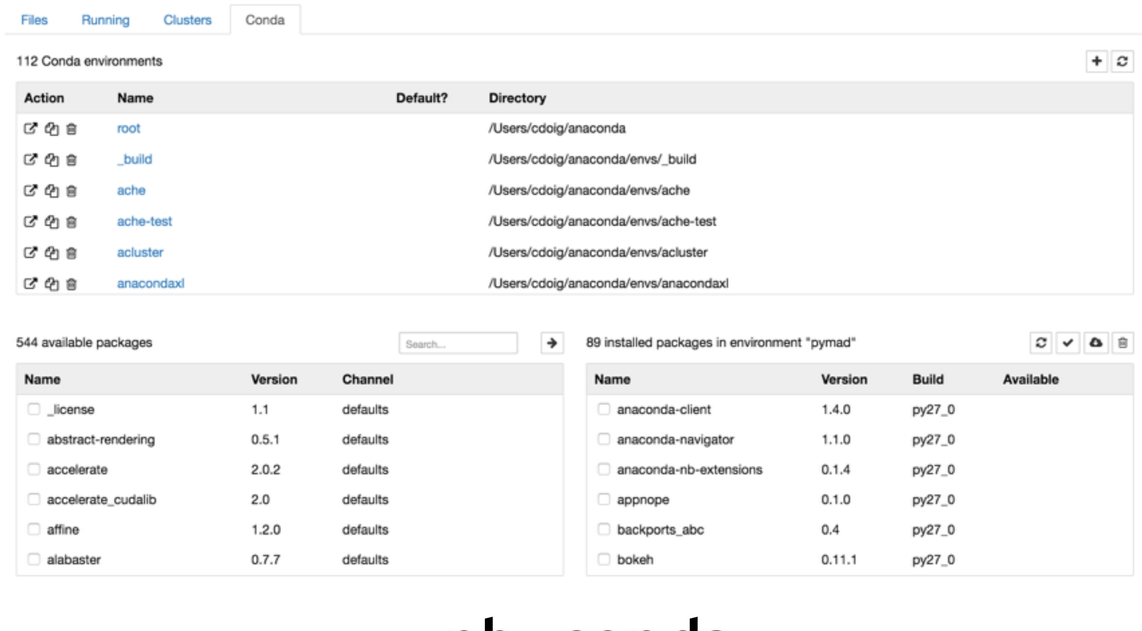


Jupyter extensions - anaconda-nb-extensions

- **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 jupter notebook

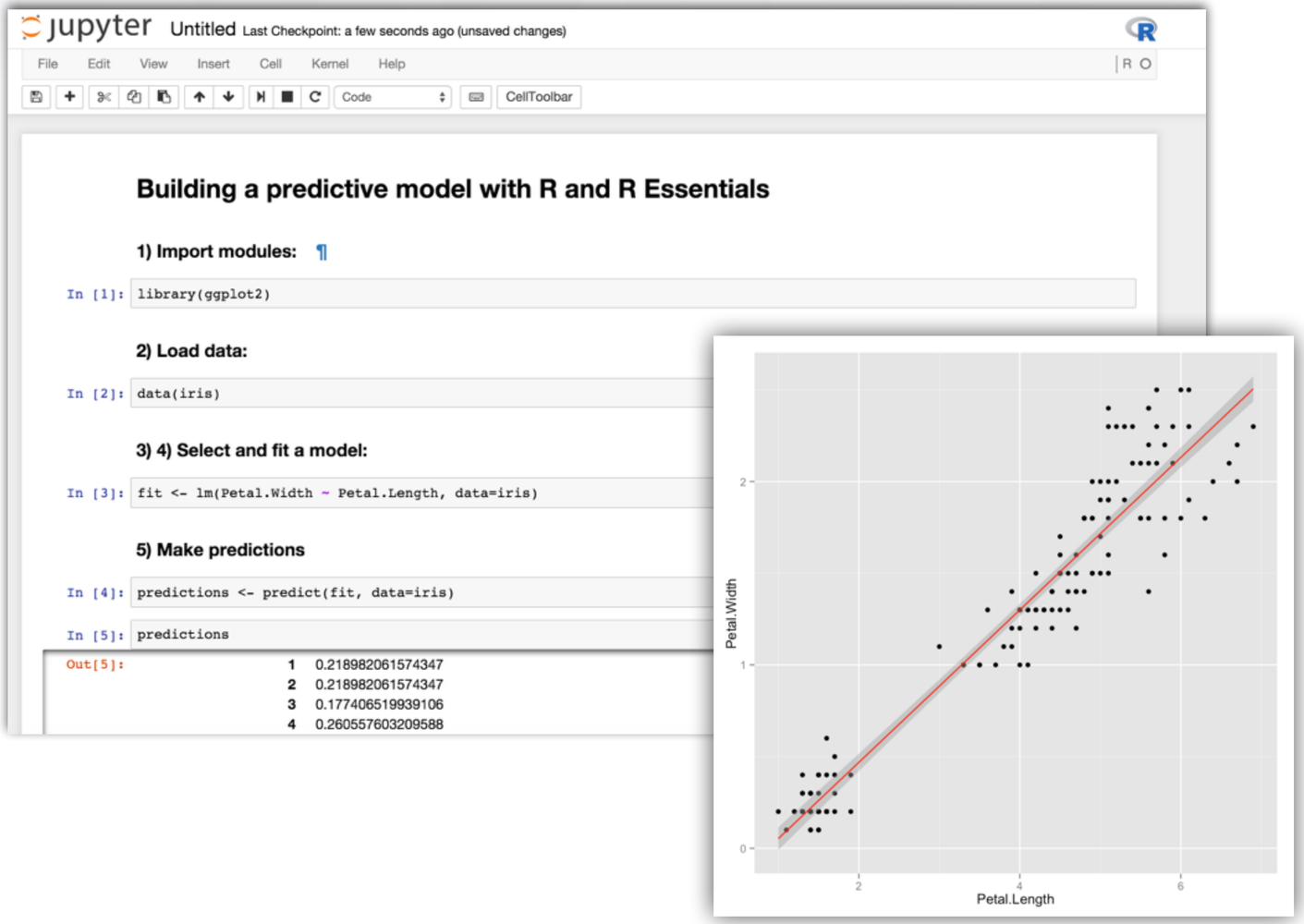


nb_condakernel



nb_conda

Jupyter: IRkernel

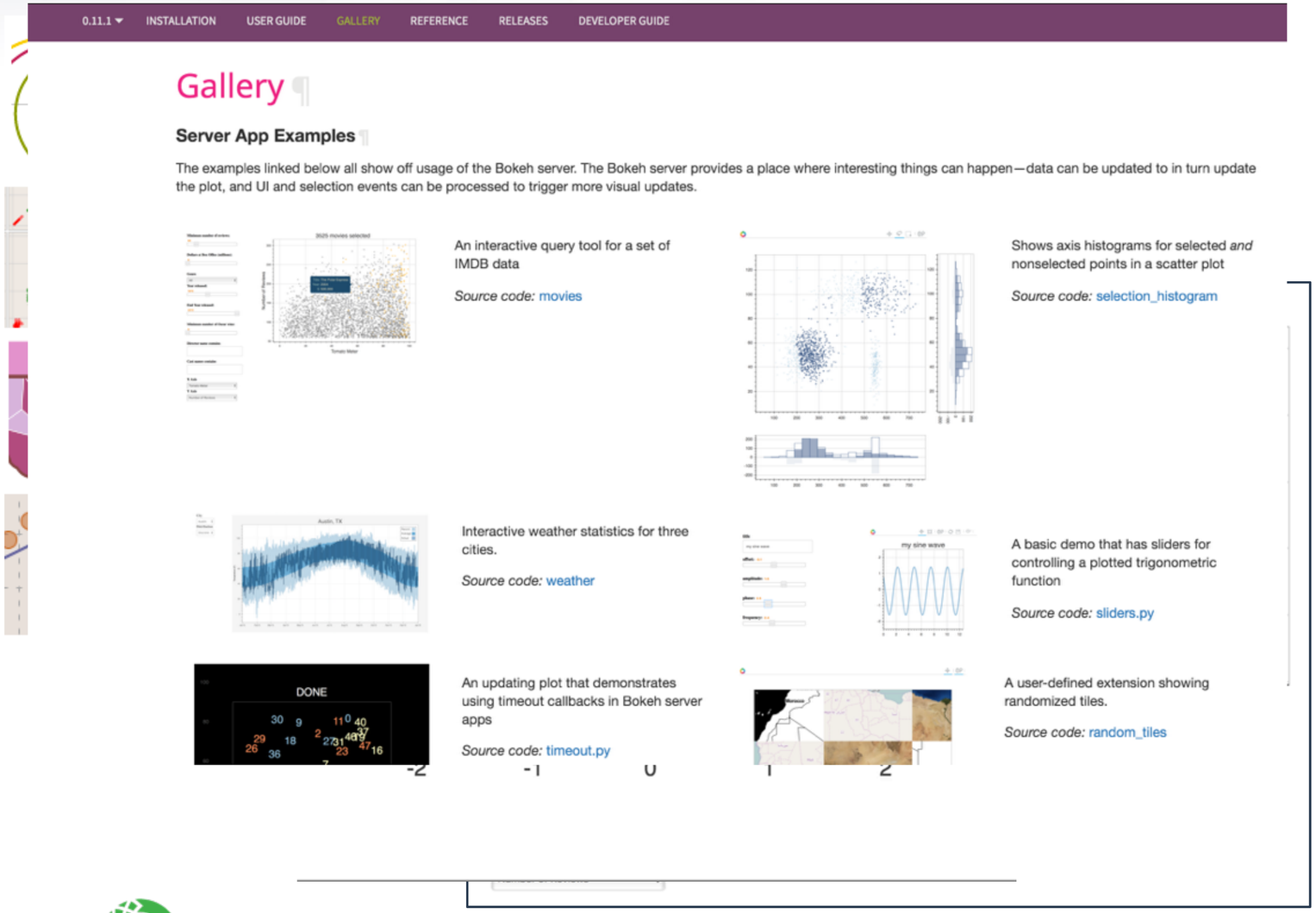


Trivial to get started writing R notebooks the same way you write Python ones.

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

https://www.continuum.io/blog/developer/jupyter-and-conda-r

Bokeh



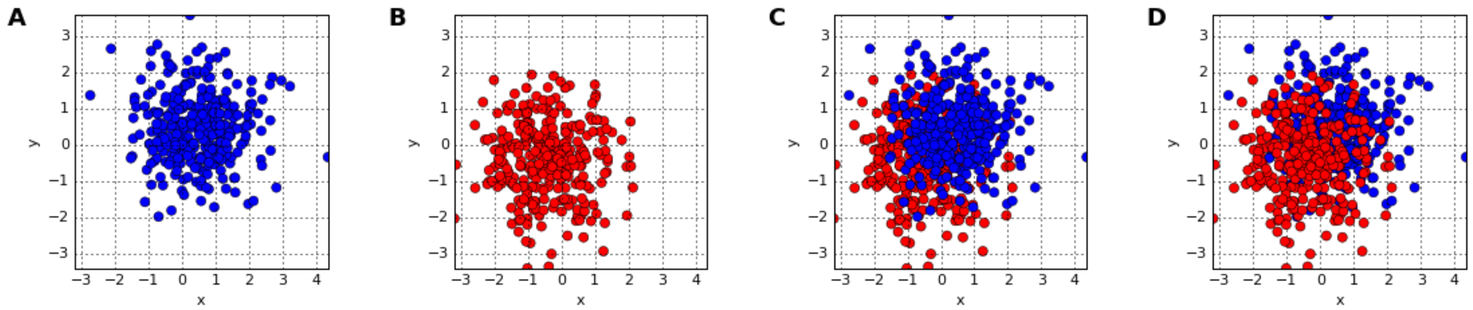
Interactive visualization framework that targets modern web browsers for presentation

- 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.

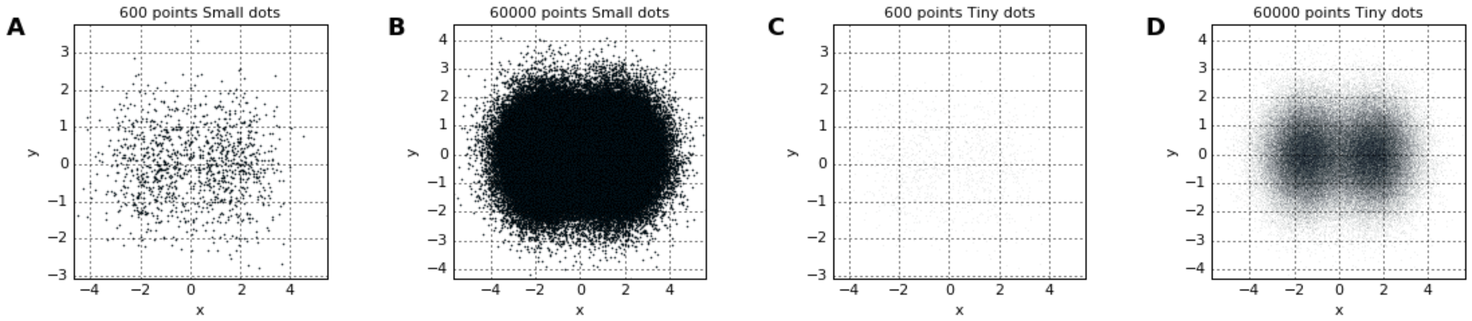
http://bokeh.pydata.org/en/latest/

Datashader - Plotting pitfalls

Overplotting:



Undersampling:



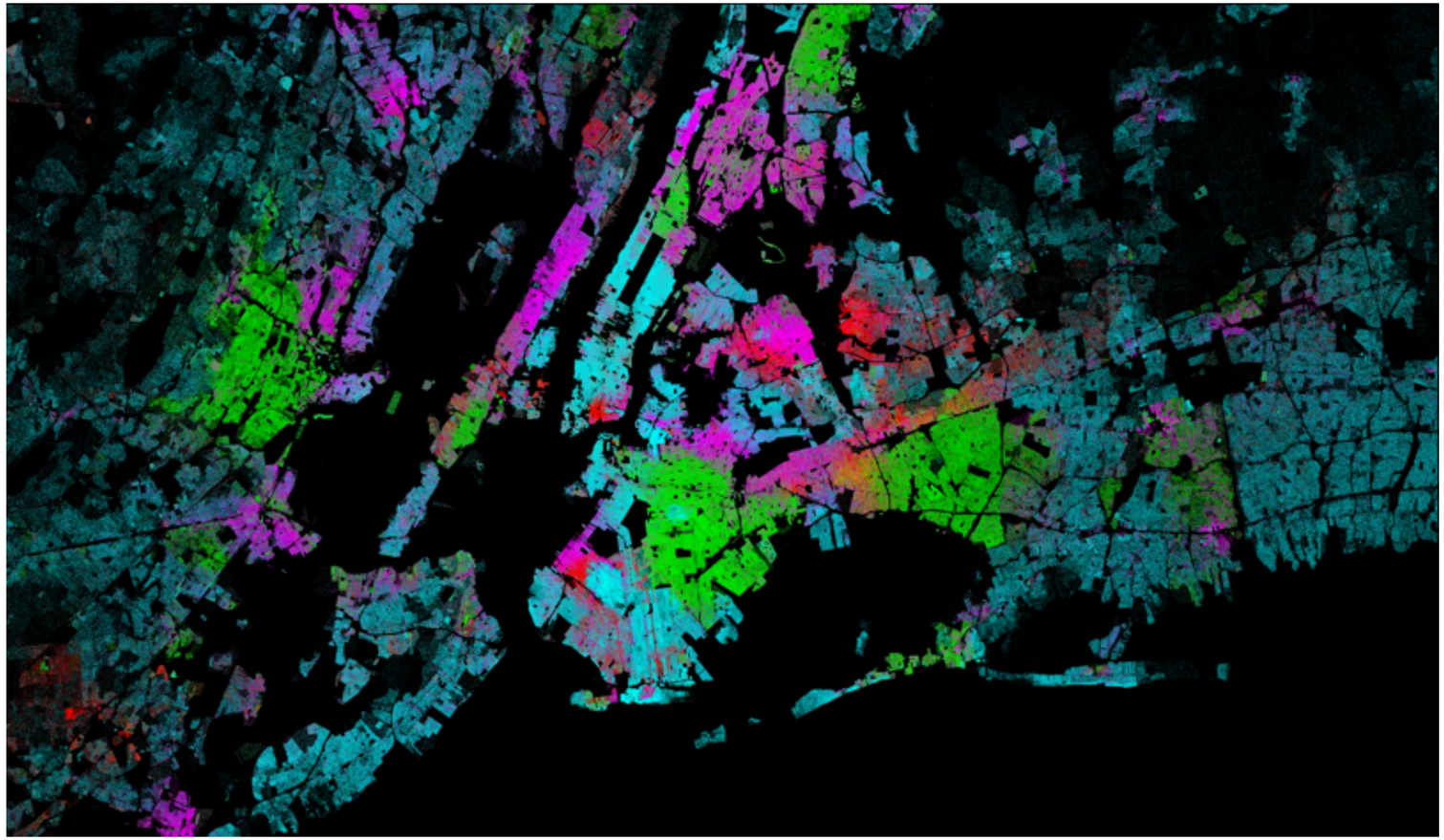
https://anaconda.org/jbednar/plotting_pitfalls/notebook

Datashader

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

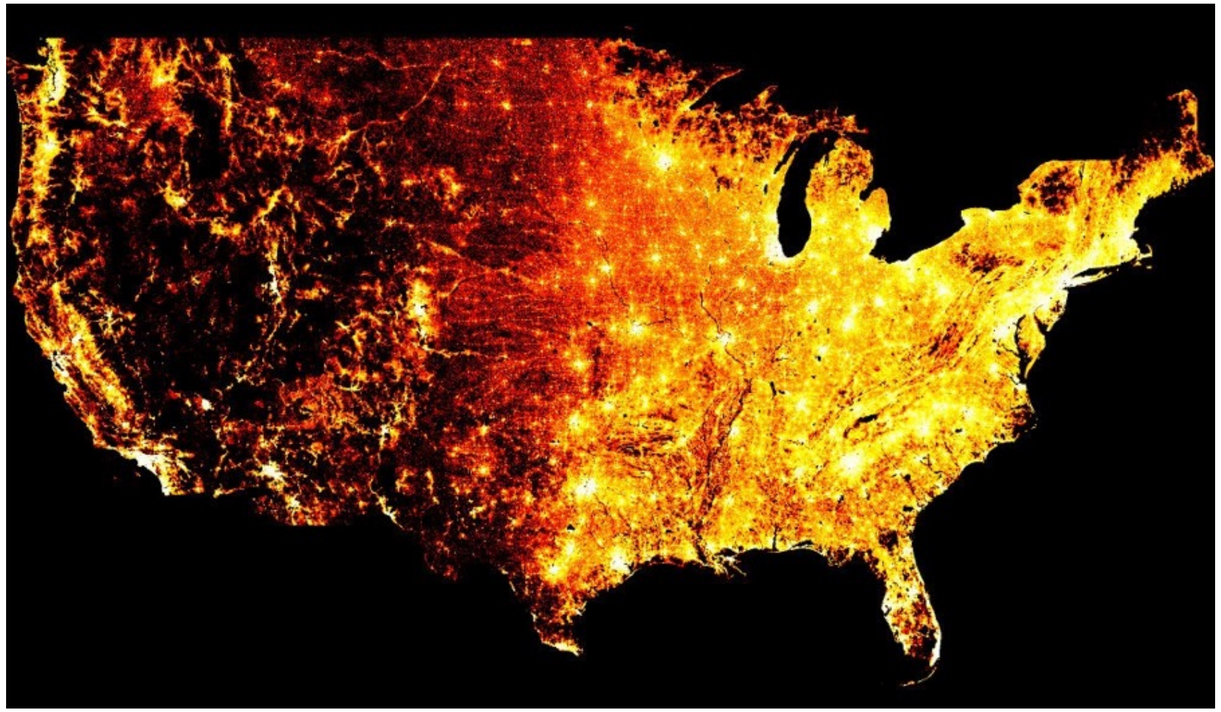
- Provides automatic, nearly parameter-free visualization of datasets
- Allows extensive customization of each step in the data-processing pipeline
- Supports automatic downsampling and re-rendering with Bokeh and the Jupyter notebook
- Works well with dask and numba to handle very large datasets in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

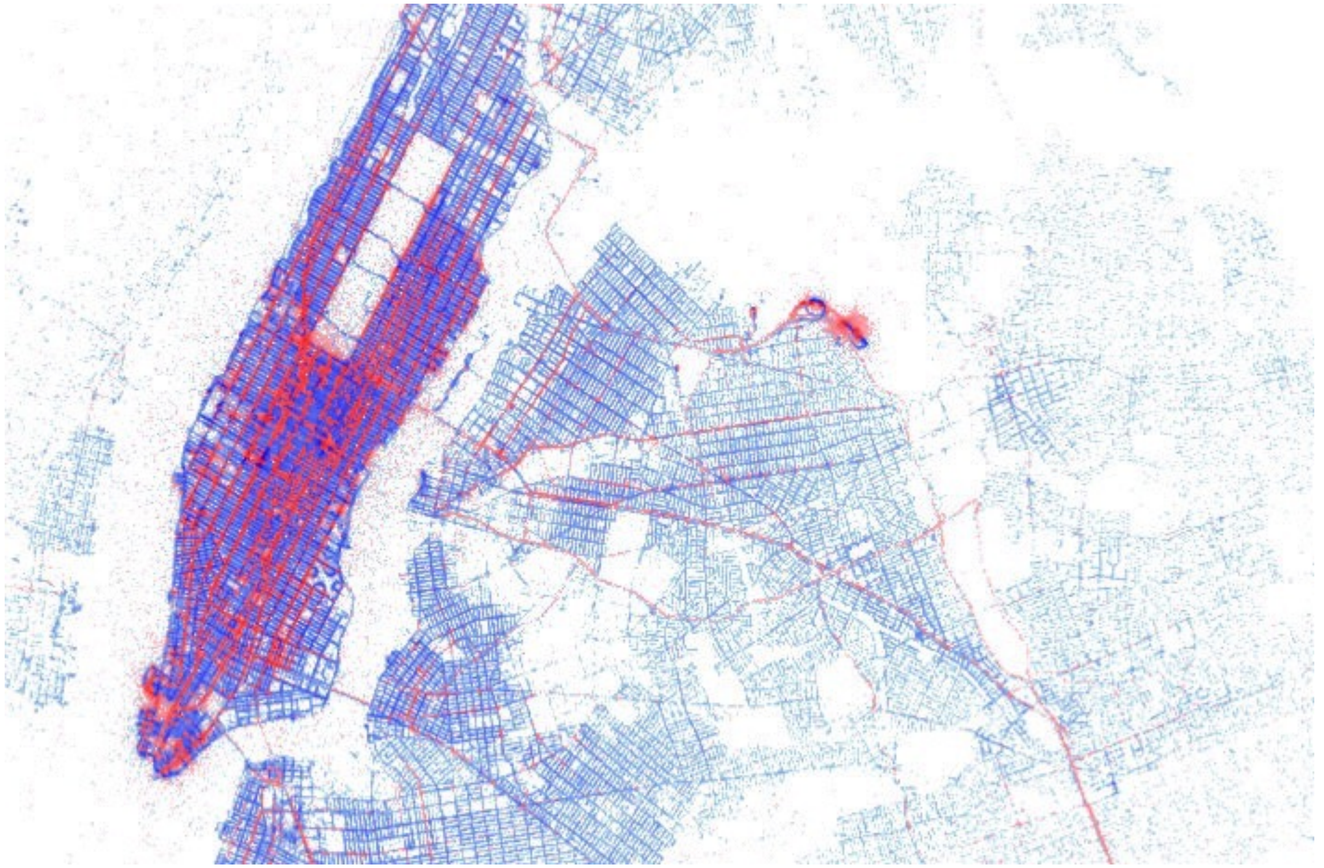
https://github.com/bokeh/datashader



NYC census data by race

Datashader





More examples:

https://anaconda.org/jbednar/notebooks

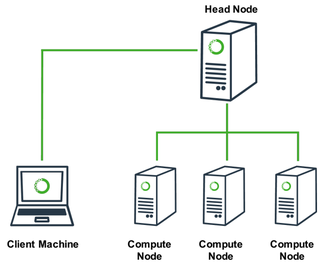
CONTINUUM ANALYTICS

Dask: Scaling Data Analysis

One month CSV file ~ 2GBs

Six month CSV file ~ 12GBs

Two years CSV files ~ 50GB



HDFS +

DASK

+ distributed

Scaling Data Analysis

ANACONDA

Dask Dataframes

CONTINUUM ANALYTICS


```
>>> 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
```

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()

5.7999999999999998
```

```
>>> 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
3           4.6          3.1           1.5
   -setosa
4           5.0          3.6           1.4
   -setosa
...

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

>>> d_max_sepal_length_setosa.compute()

5.7999999999999998
```

Distributed

CONTINUUM ANALYTICS

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

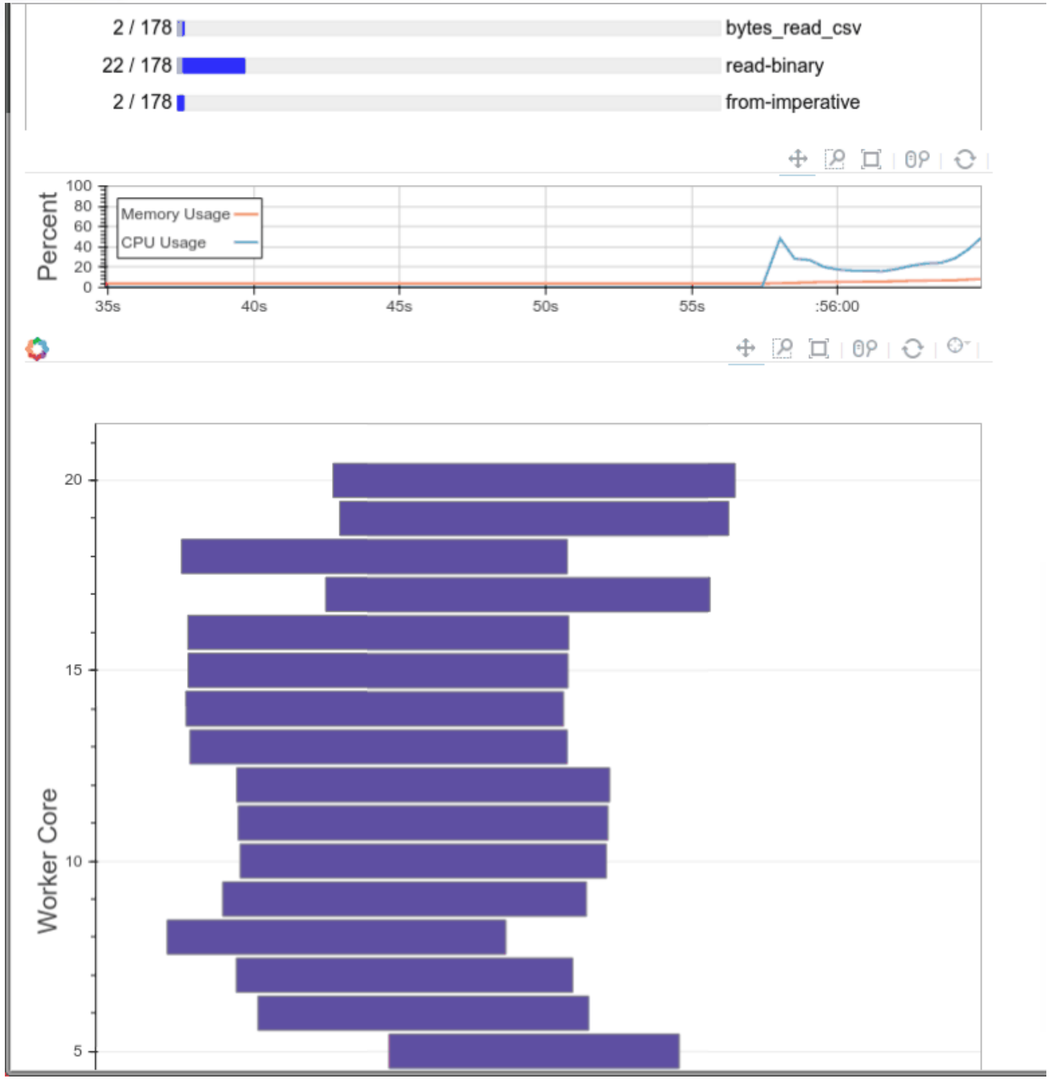
```
>>> from distributed import Executor, hdfs, progress
>>> e = Executor('127.0.0.1:8786')
>>> e
<Executor: scheduler=127.0.0.1:8786 workers=64 threads=64>

>>> 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])
```

ANACONDA

http://distributed.readthedocs.io/en/latest/

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.



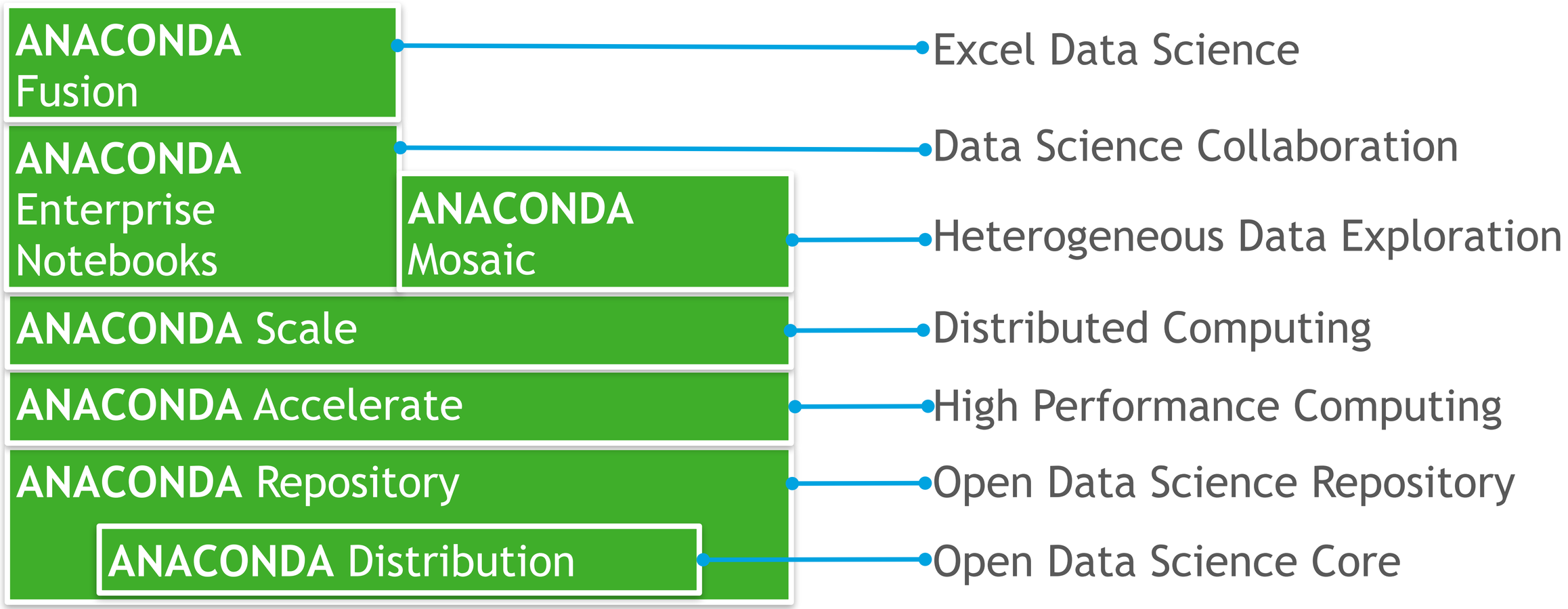
CONTINUUM®
ANALYTICS

ANACONDA ENTERPRISE PLATFORM

ANACONDA®

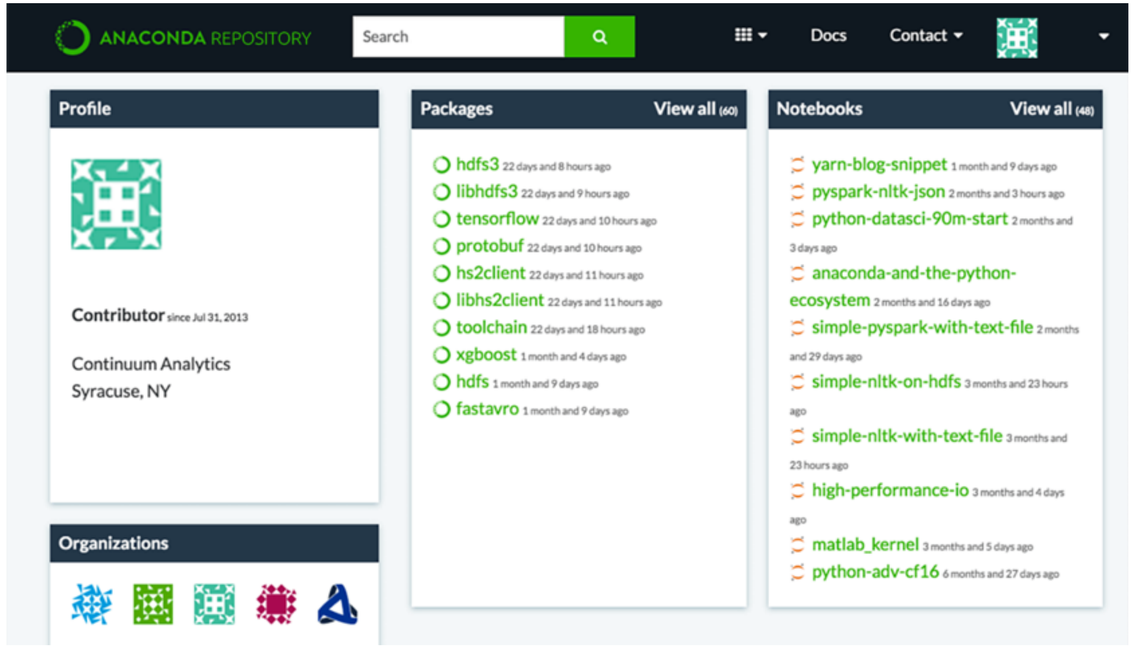
CONTINUUM ANALYTICS

ANACONDA platform

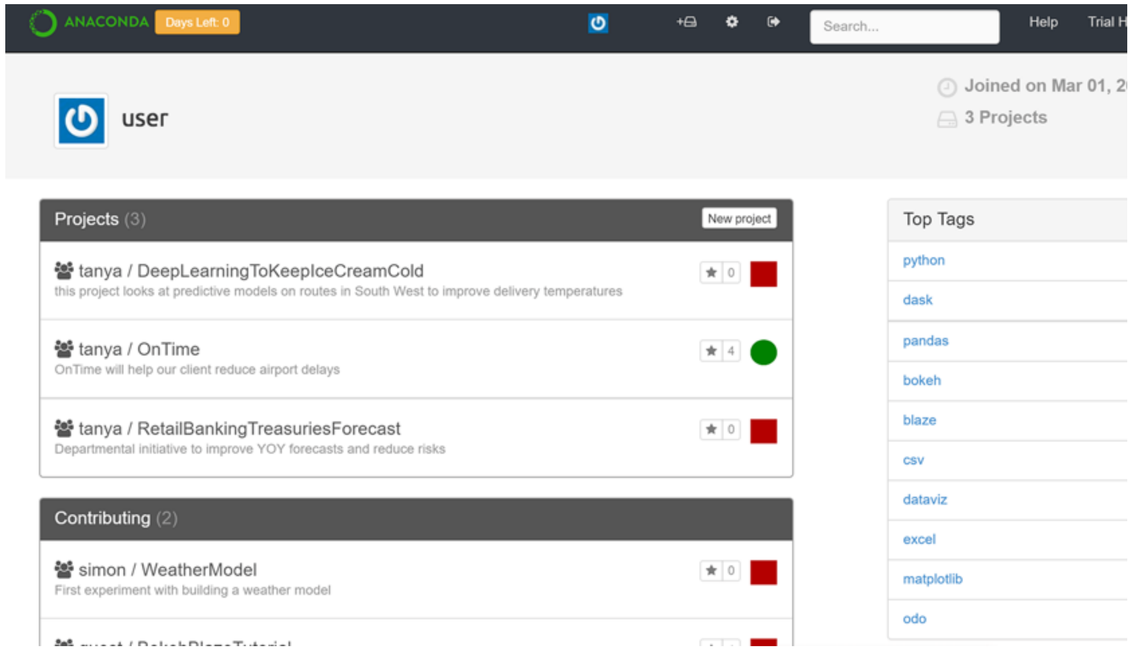


ANACONDA

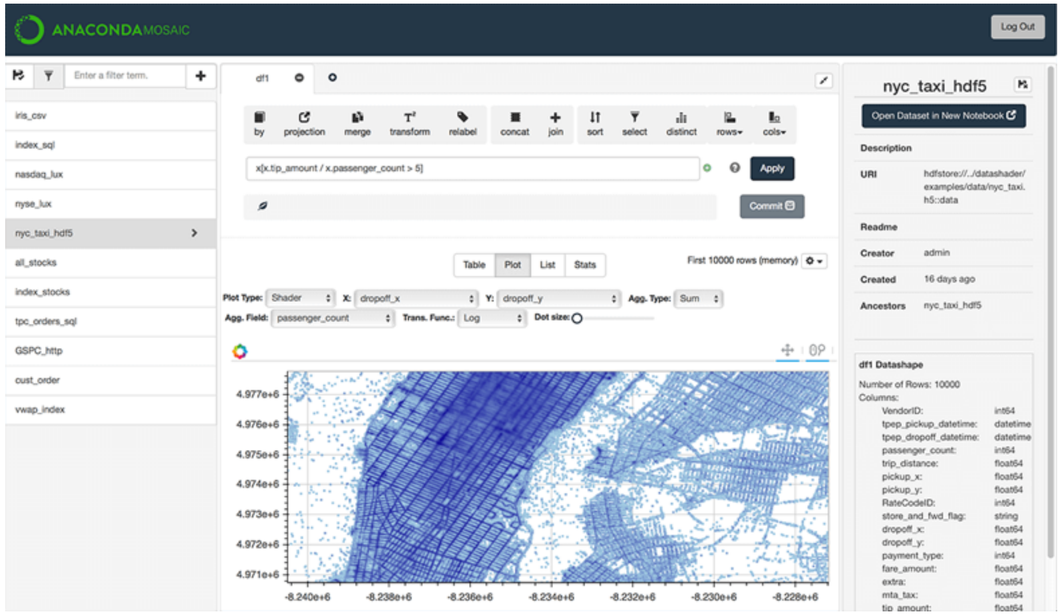
© 2016 Continuum Analytics- Confidential & Proprietary



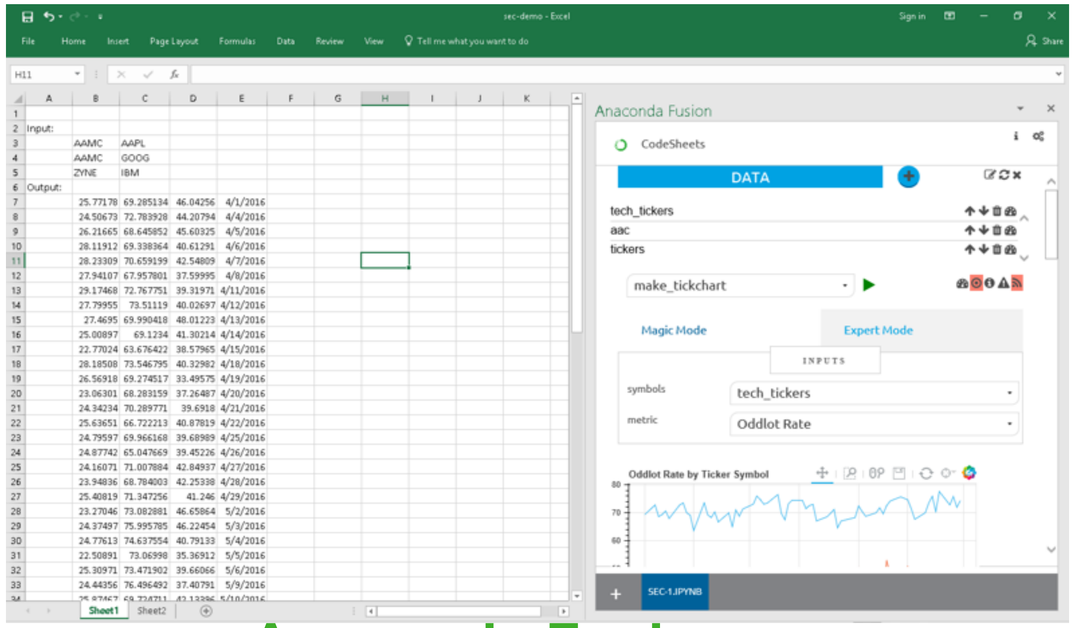
Anaconda Repository



Anaconda Enterprise Notebooks



Anaconda Mosaic



Anaconda Fusion

CONTINUUM® ANALYTICS

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

ANACONDA®

CONTINUUM ANALYTICS

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

ANACONDA

CONTINUUM ANALYTICS

Thank you!

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

Twitter: @ch_doig