

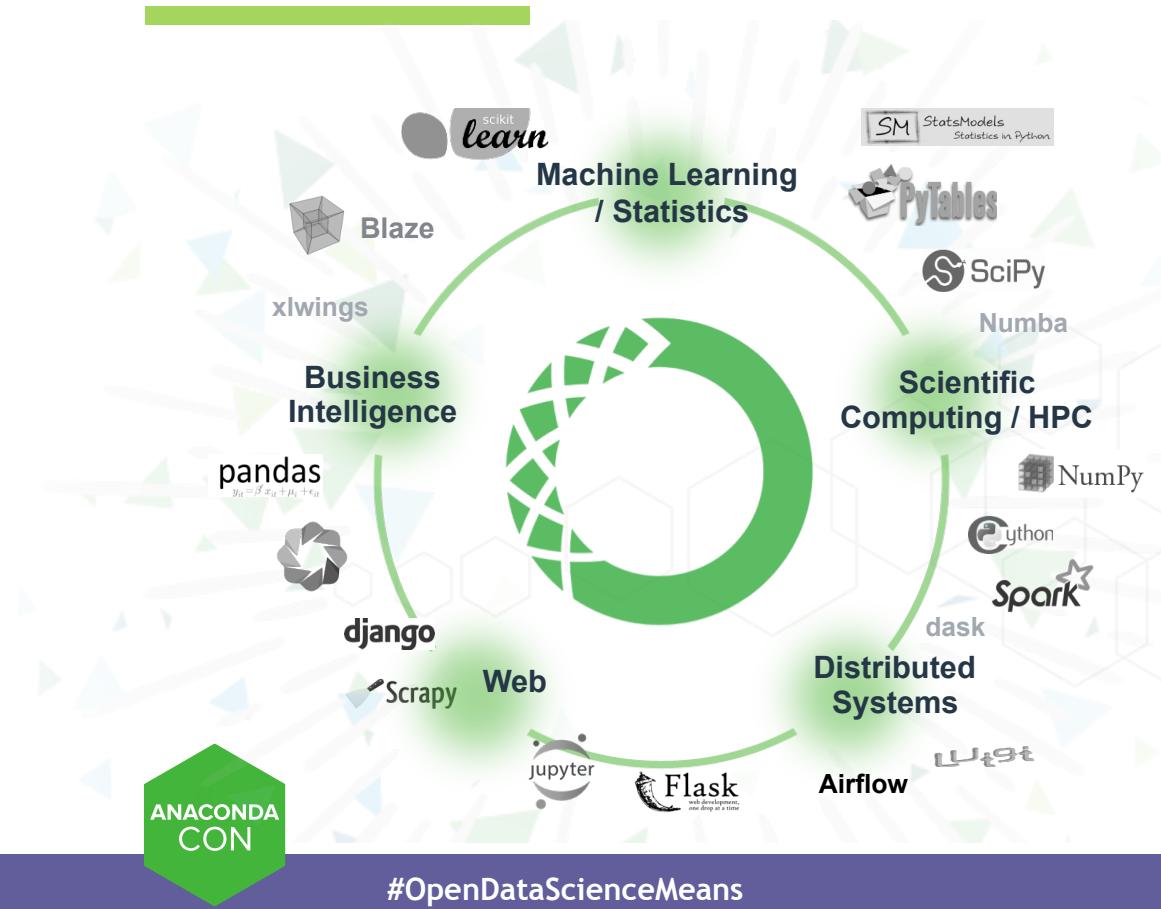


AUSTIN, TX | 2017

REACHING THE FULL POTENTIAL OF A DATA-DRIVEN WORLD

Growing the power of Anaconda and its Community

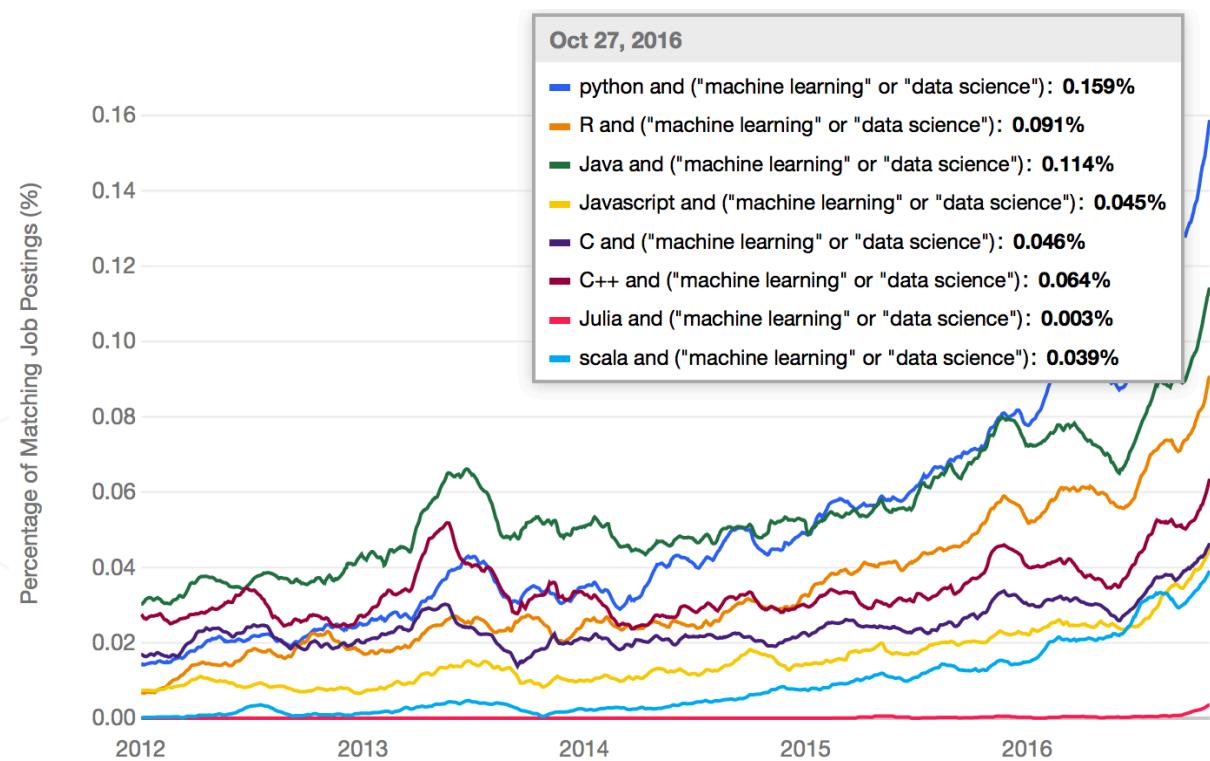
ANACONDA IS OPEN DATA SCIENCE



GROWTH OF DATA SCIENCE (AND ANACONDA)



#OpenDataScienceMeans
#AnacondaCON



By [Jean-Francois Puget](#), IBM.

How can
Anaconda and its
Community help?

How do we find
and build solid
foundations?

How do we turn
today's data
deluge into a
better life for us,
our children, our
grandchildren,
and beyond.



#OpenDataScienceMeans
#AnacondaCON

DATA DELUGE IS JUST BEGINNING



A FEW SOURCES OF INCREASING DATA



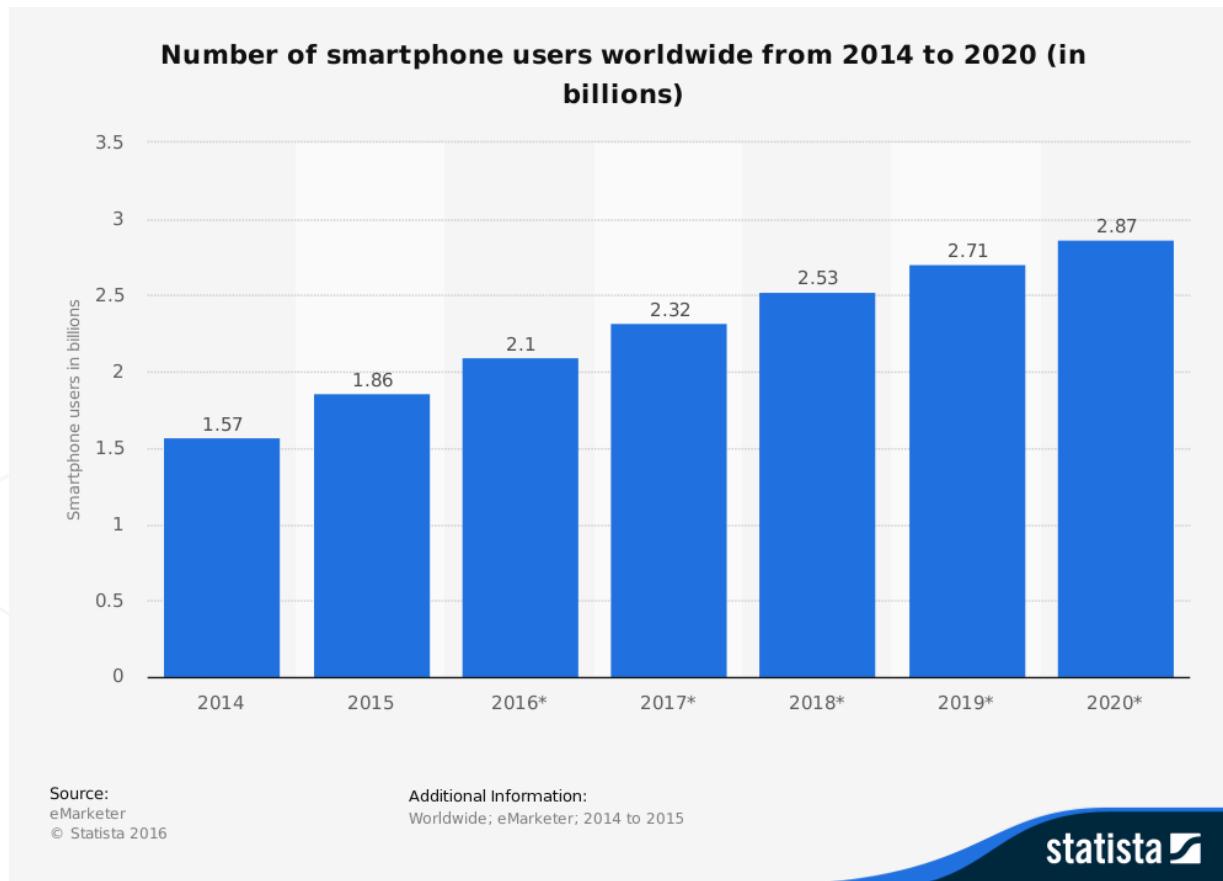
Mobile devices (including media)

Sensors (everything is collecting digital data at increasing rates)

Amazon Cloud Drive now offers unlimited storage for individuals



#OpenDataScienceMeans
#AnacondaCON



INTERNET OF THINGS...

Cars have many sensors collecting data for safety — the future is self-driving

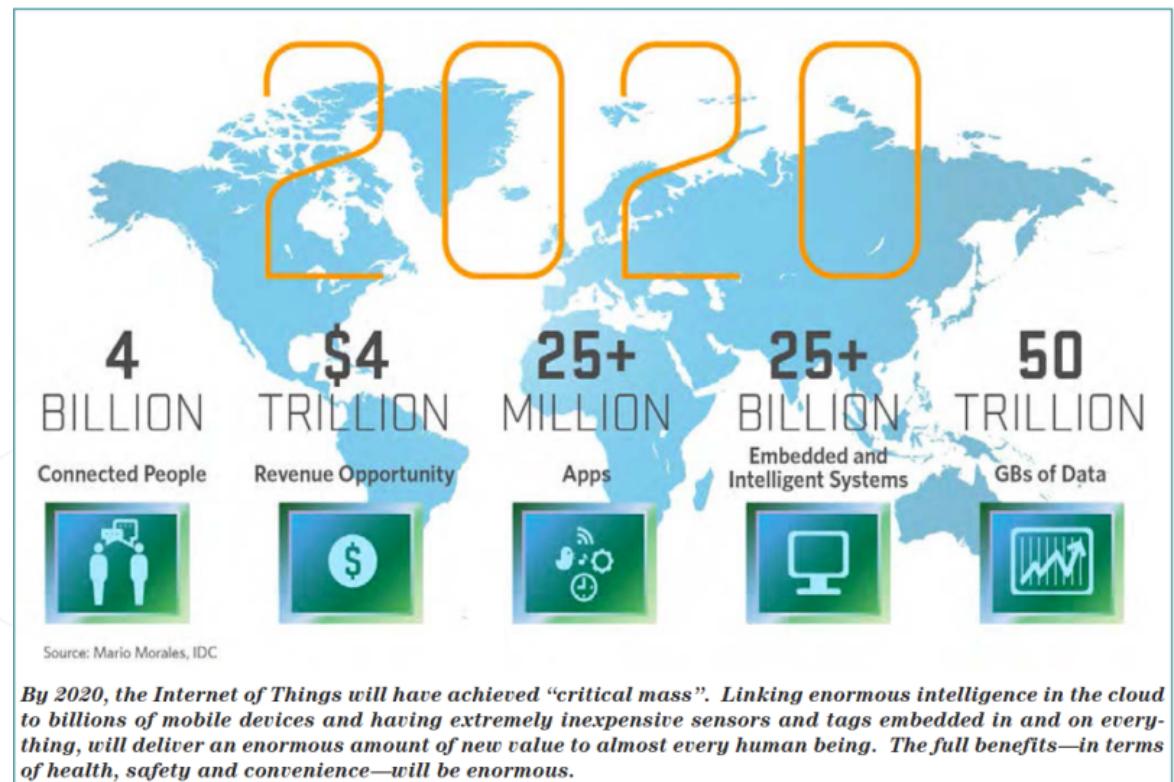
Automate homes

Improve lifetimes of household appliances



ANACONDA
CON

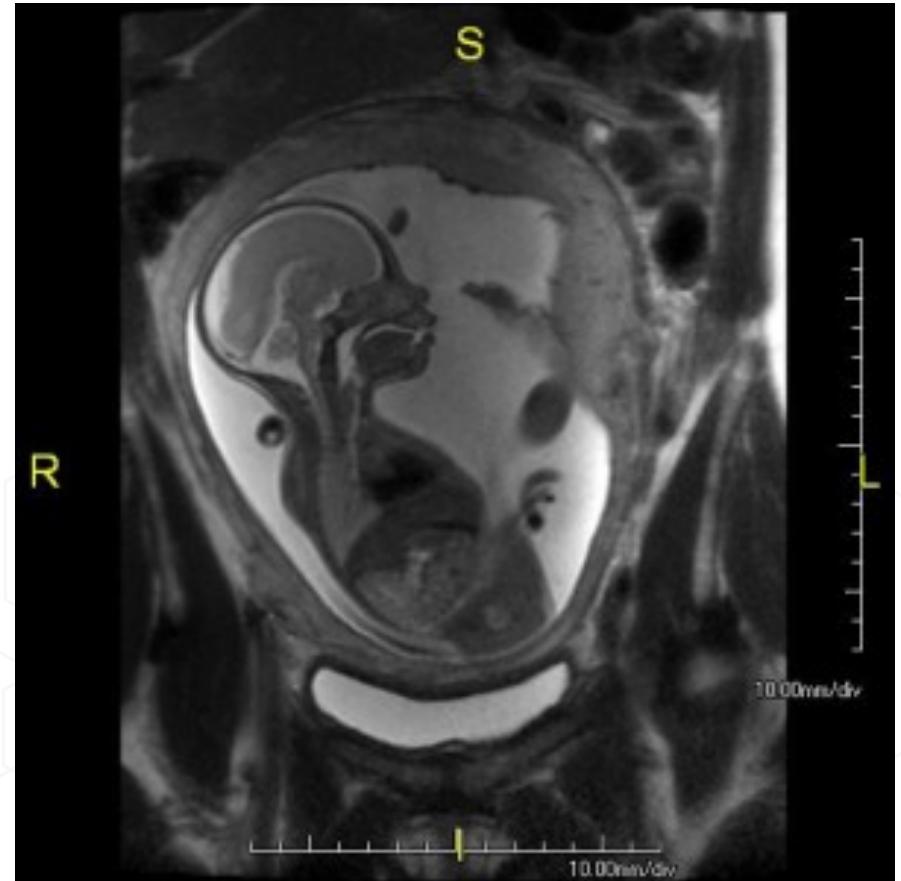
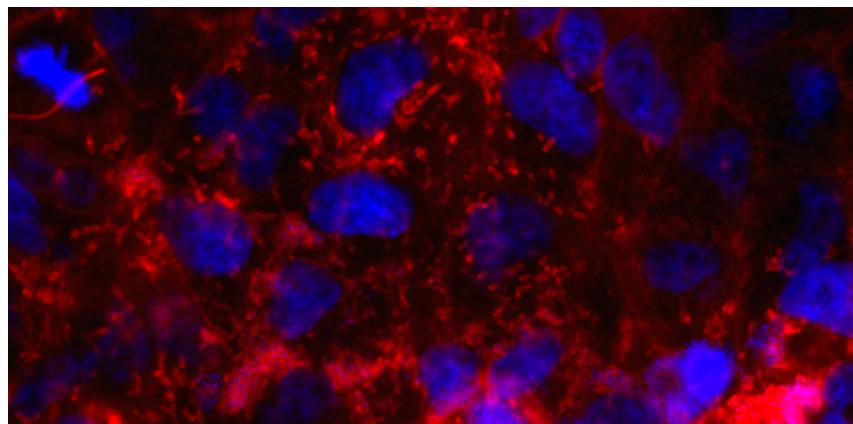
#OpenDataScienceMeans
#AnacondaCON



HEALTH DATA



Including the “Quantified Self”



ANACONDA
CON

#OpenDataScienceMeans
#AnacondaCON

© 2016 Continuum Analytics- Confidential & Proprietary

SOCIAL MEDIA



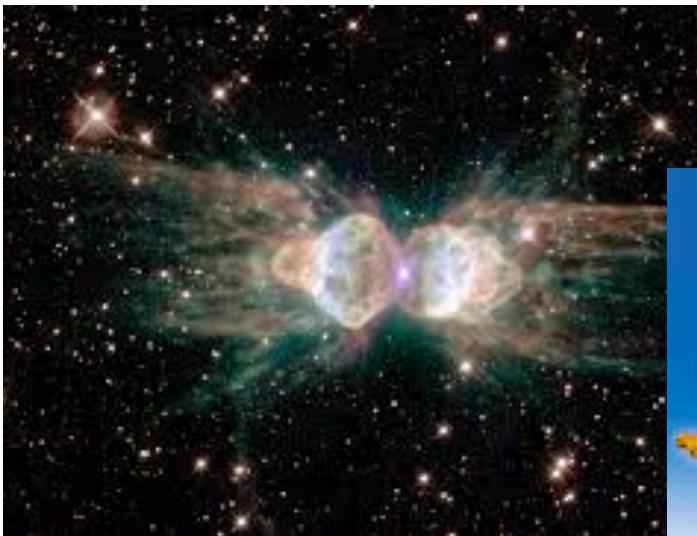
- 1/2 billion tweets per day
- 1 billion pieces of content shared each day on Facebook
- 1/2 billion users of Facebook (in a social graph)



#OpenDataScienceMeans
#AnacondaCON

SCIENTIFIC DATA

Astronomy (e.g. Hubble)



Big Physics (e.g. CERN)



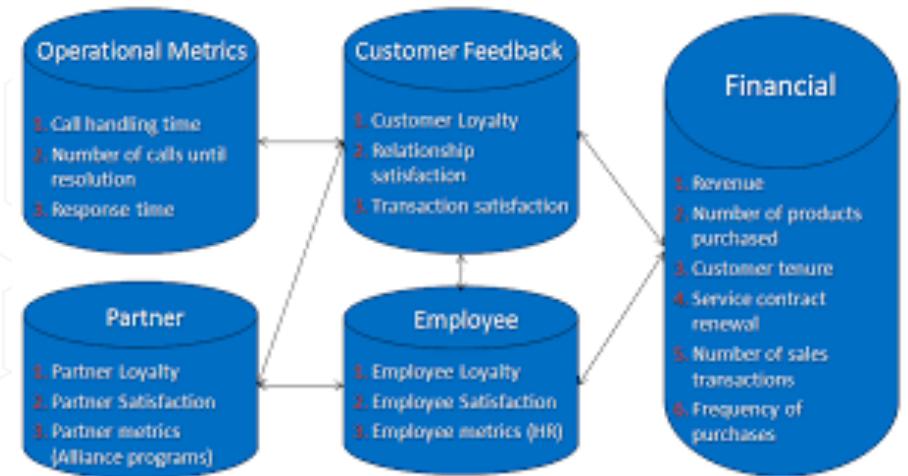
Satellite Imaging



#OpenDataScienceMeans
#AnacondaCON

BUSINESS DATA

- Customer information
 - Marketing
 - Sales
 - Account management
- Financial Information
- Models for your business
 - Great variety here.
 - Can conceptually use any data from any-source
- Whatever else you store in your ever-growing Databases.



DON'T FORGET THE “DARK DATA”

Dark Data: CSV, hdf5, npz, XLS, docx, logs, emails, and other files
in your company outside a traditional store

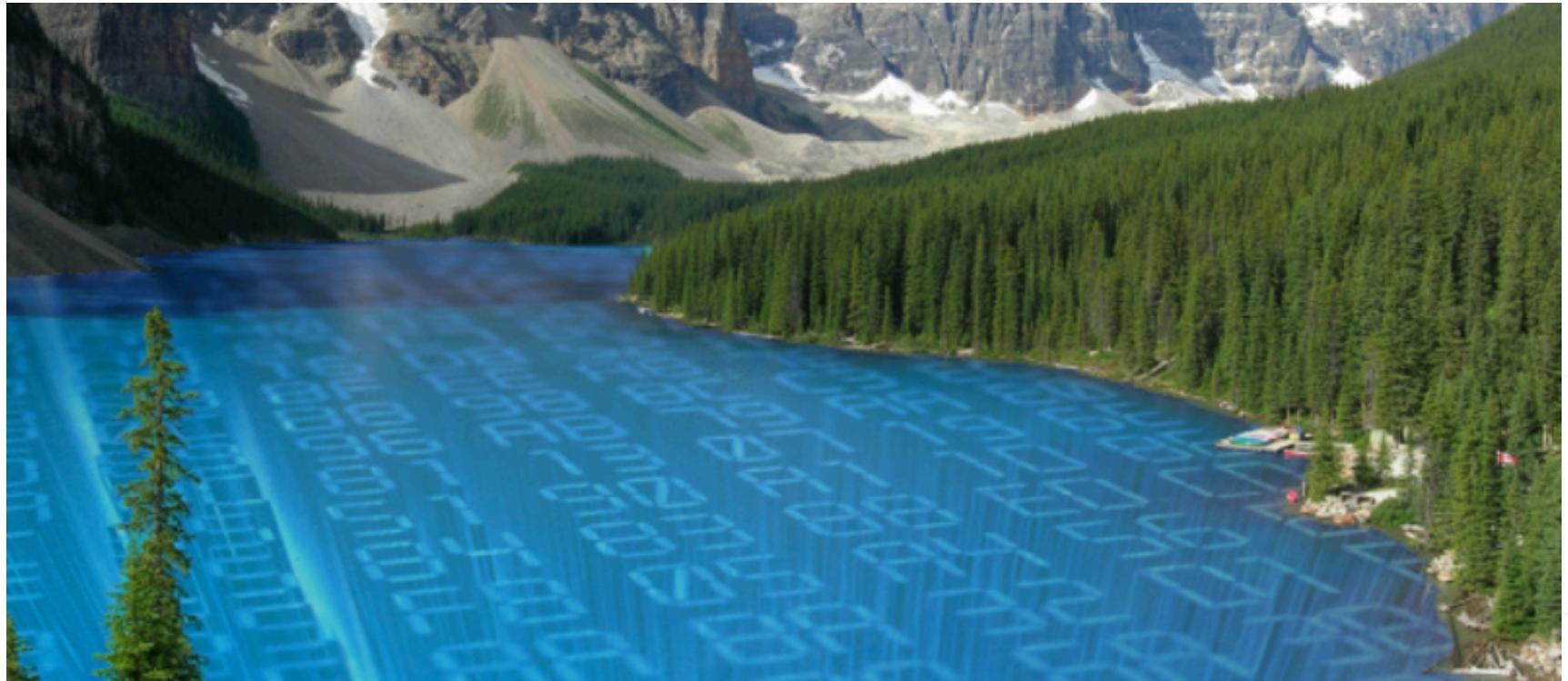
Information of extreme size, diversity and complexity



CON

#OpenDataScienceMeans
#AnacondaCON

SO YOU BUILD YOUR DATA LAKE



#OpenDataScienceMeans
#AnacondaCON

**Insight
Decisions
Actions
Results
Creations**

**Open Data
Science**

DATA



#OpenDataScienceMeans
#AnacondaCON

Applications

Dashboards

Notebooks

• • •

MicroServices

ANACONDA®



REACHING FULL POTENTIAL WITH ANACONDA

- Make “Code to Data” connection seamless and easy
- Amplify learning
- Shrink time from idea to production
- Choose the right algorithms for your data and question
- Cultivate and participate in community mission



#OpenDataScienceMeans
#AnacondaCON

CODE TO DATA

Data Silos are everywhere

Python is great glue to connect the Silos

Same data must be stored twice in memory for different languages because there are not common data-descriptions!

Blaze project still working to solve this!



#OpenDataScienceMeans
#AnacondaCON

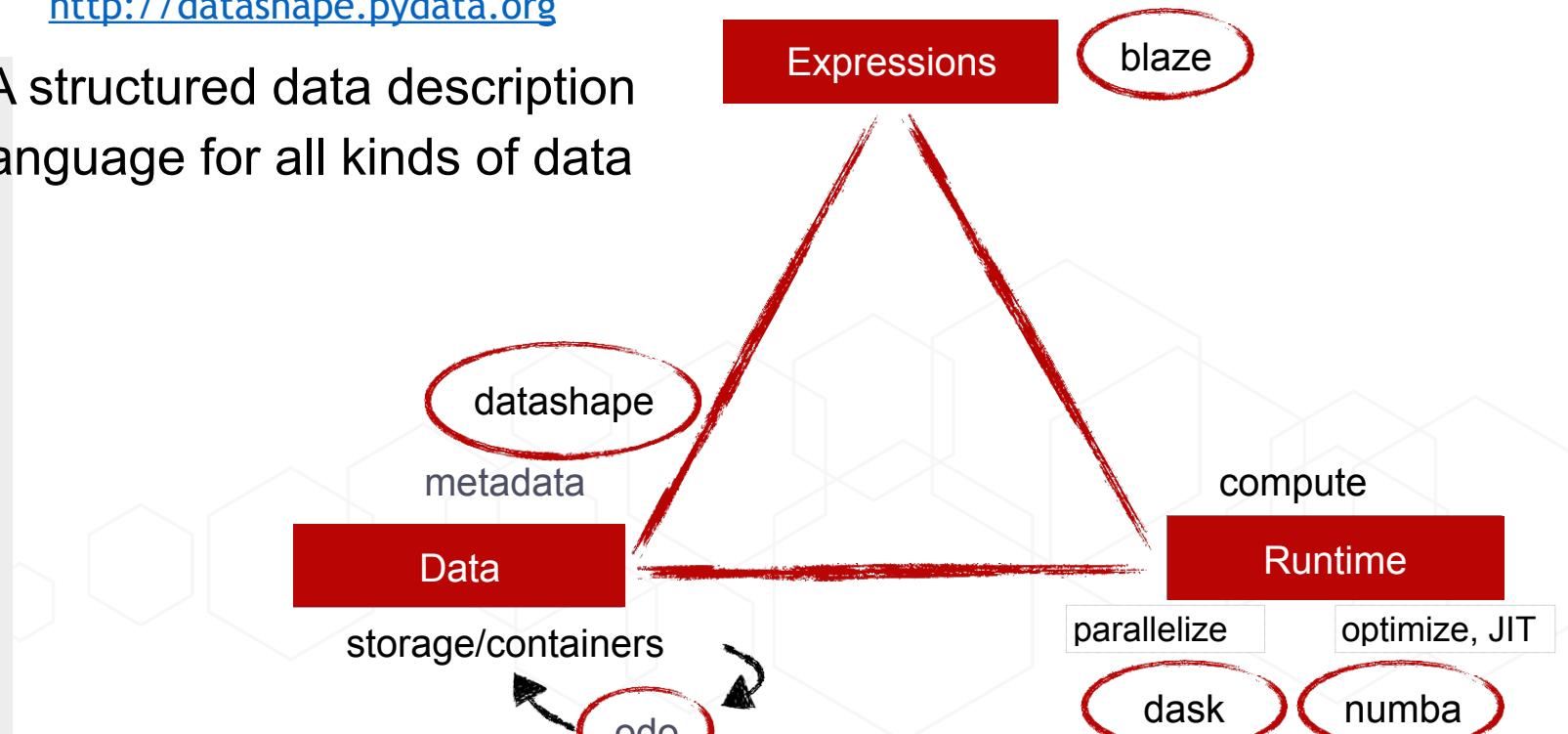


CODE TO DATA

<http://datashape.pydata.org>

```
{  
    flowersdb: {  
        iris: var * {  
            petal_length: float64,  
            petal_width: float64,  
            sepal_length: float64,  
            sepal_width: float64,  
            species: string  
        }  
    },  
    iriscsv: var * {  
        sepal_length: ?float64,  
        sepal_width: ?float64,  
        petal_length: ?float64,  
        petal_width: ?float64,  
        species: ?string  
    },  
    irisjson: var * {  
        petal_length: float64,  
        petal_width: float64,  
        sepal_length: float64,  
        sepal_width: float64,  
        species: string  
    },  
    irismongo: 150 * {  
        petal_length: float64,  
        petal_width: float64,  
        sepal_length: float64,  
        sepal_width: float64,  
        species: string  
    }  
}
```

A structured data description language for all kinds of data

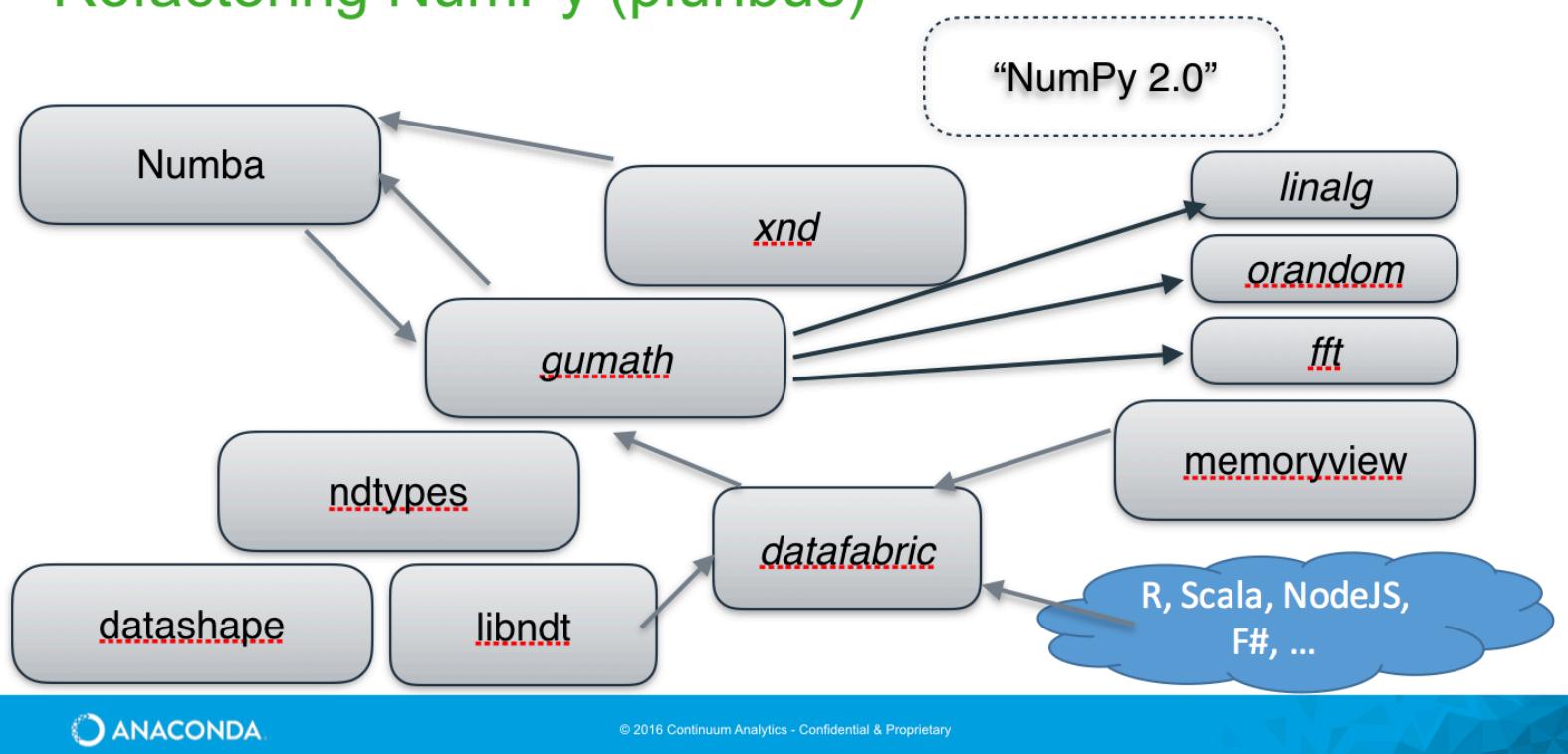


ANACONDA CON

#OpenDataScienceMeans
#AnacondaCON

CODE TO DATA

Refactoring NumPy (pluribus)



ANACONDA

© 2016 Continuum Analytics - Confidential & Proprietary



#OpenDataScienceMeans
#AnacondaCON

AMPLIFY LEARNING

190,000

**PROJECTED SHORTAGE
IN DATA SCIENTISTS
BY 2018**



Skills Assessment and Data Science Placement Program



#OpenDataScienceMeans
#AnacondaCON

Filling this gap with better tools
And streamlining education to be a data-scientist



AMPLIFY LEARNING



“I am still learning.”
-Michelangelo, age 87

Brain Plasticity continues well into old-age.

You can teach an “old-dog” new tricks – but it takes time and effort.

Often new initiatives have to wait for the “old guard” to retire in order for innovation to take root.

You can change the world by “embracing” learning throughout your life. No matter how much you already know.

ANACONDA
CON

#OpenDataScienceMeans
#AnacondaCON

AMPLIFY LEARNING



Skills Assessment and Data Science Placement Program

Formal “post-graduate” independent-study program with “on-the-job” learning and mentoring that takes people from where they are and improves their data-science, data-engineering, and quantitative programming skills.

Contact me if you:

1. want to enroll
2. want to “contract-to-hire” someone with Anaconda capabilities



#OpenDataScienceMeans
#AnacondaCON

AMPLIFY LEARNING



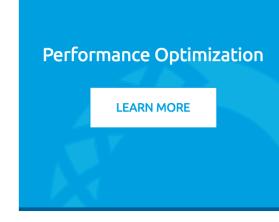
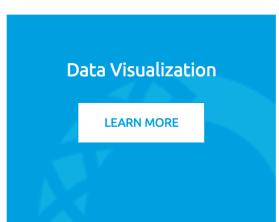
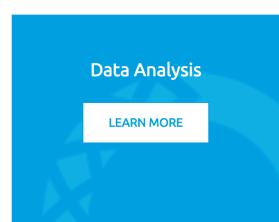
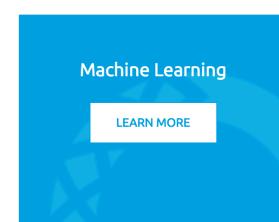
Data Science Classes

Our multi-day data science classes are ideal for analysts, data scientists and engineers with some prior programming experience. The classes are particularly useful for those wanting to improve their skills in building data science workflows.



Deep Dive Data Science Classes

Our Deep Dive classes are ideal for those wanting to learn more Python in an intensive one-day course. Participants should have completed introductory Python training or have strong prior experience with Python.



#OpenDataScienceMeans
#AnacondaCON

AMPLIFY LEARNING

Work to do on right curriculum from elementary school to graduate school!

- Basic Probability Theory and Bayes rule taught early (5th grade)
- Partial-derivatives with linear algebra (gradient without divergence and curl)
- Computing literacy
 - array-oriented concepts
 - basic data-structures
 - basic algorithms



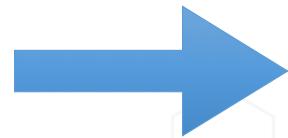
<http://brilliant.org>



#OpenDataScienceMeans
#AnacondaCON

SHRINKING TIME FROM IDEA TO PRODUCTION

- collaboration
- automation
- common platforms
- shared abstractions
- versioning
- authentication
- governance
- recognize it is iterative



i.e 5.x version of Anaconda Products!



#OpenDataScienceMeans
#AnacondaCON

RIGHT ALGORITHMS

- **Supervised Learning** — uses “labeled” data to train a model
 - Regression — predicted variable is continuous
 - Classification — predicted variable is discrete
- **Unsupervised Learning**
 - Clustering — discover categories in the data
 - Density Estimation — determine representation of data
 - Dimensionality Reduction — represent data with fewer variables or feature vectors
- Reinforcement Learning — “goal-oriented” learning (e.g. drive a car)
- **Deep Learning** — neural networks with many layers
- Semi-supervised Learning (use some labeled data for training)



#OpenDataScienceMeans
#AnacondaCON

RIGHT ALGORITHMS

Supervised Learning

\mathbf{X} Input Data or “feature vectors”

y Labels for training

θ Parameters that determine the model
Training is the process of estimating these.

\hat{y} Predicted outputs

$$y = f(\mathbf{x}, \theta)$$

$f(\cdot, \cdot)$ Learning model.

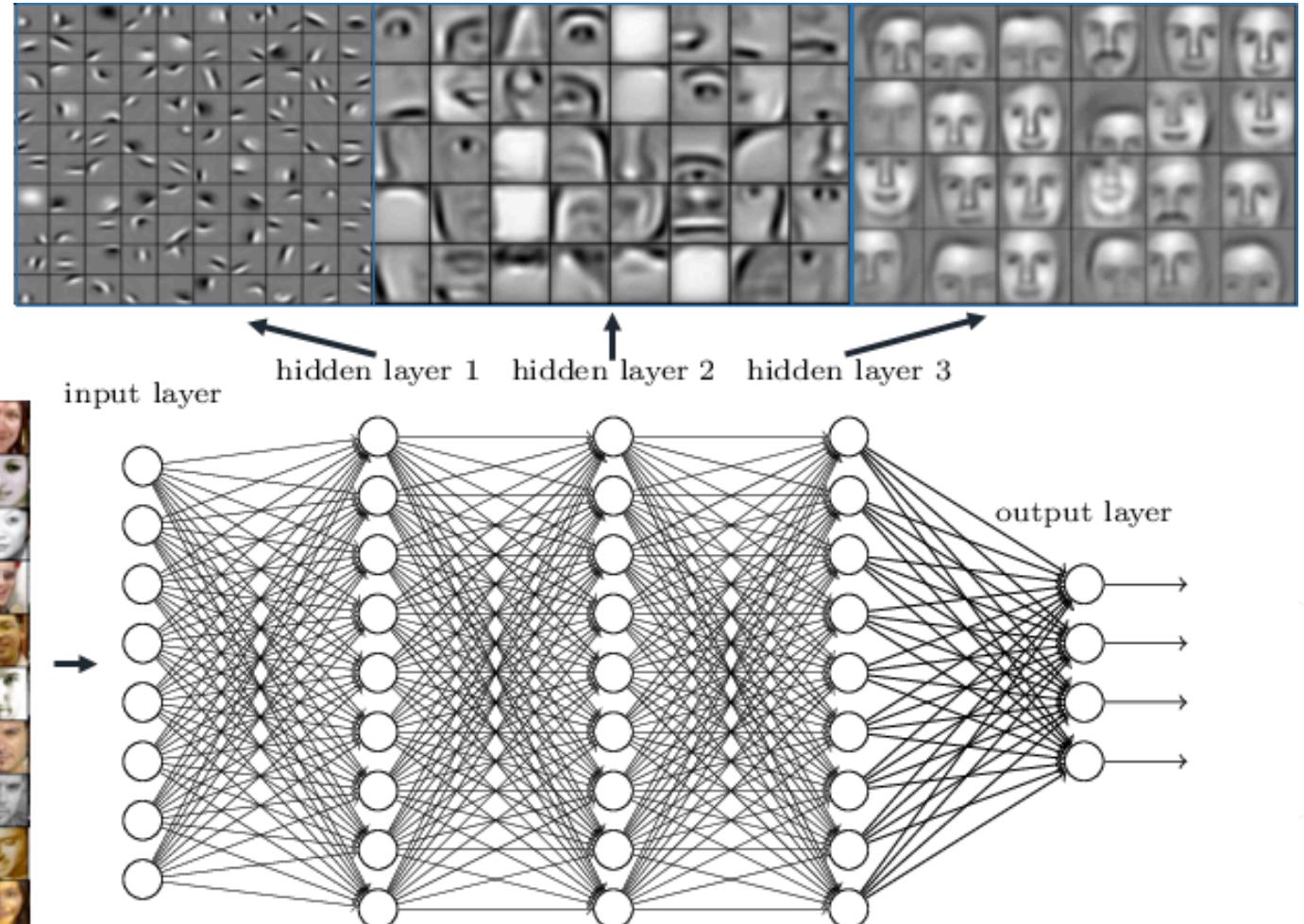
May be part of a family of models with **hyper-parameters** selecting the specific model

$$\hat{y} = f(\mathbf{x}, \theta)$$

ANACONDA
CON

#OpenDataScienceMeans
#AnacondaCON

Deep neural
networks learn
hierarchical feature
representations



ANACONDA
CON

#OpenDataScienceMeans
#AnacondaCON

Supervised Deep Learning

Same structure

X Input Data or “feature vectors”

y Labels for training

θ All the weights between layers

w_{ijk}

\hat{y} Predicted outputs

$$y = f(\mathbf{x}, \theta)$$

$$z_{ij} = g\left(\sum_k w_{ijk} z_{i-1k}\right)$$

$$g(u) = \frac{1}{1 + e^{-u}}$$

$$\hat{y} = f(\mathbf{x}, \theta)$$

ANACONDA
CON

#OpenDataScienceMeans
#AnacondaCON

Unsupervised Deep Learning

Auto-encoding

\mathbf{X} Input Data or “feature vectors”

$\hat{\mathbf{x}}$ Labels for training – set equal to Input

θ All the weights between layers

w_{ijk}

$$\hat{\mathbf{x}} = f(\mathbf{x}, \theta)$$

$$z_{ij} = g \left(\sum_k w_{ijk} z_{i-1k} \right)$$

$$g(u) = \frac{1}{1 + e^{-u}}$$

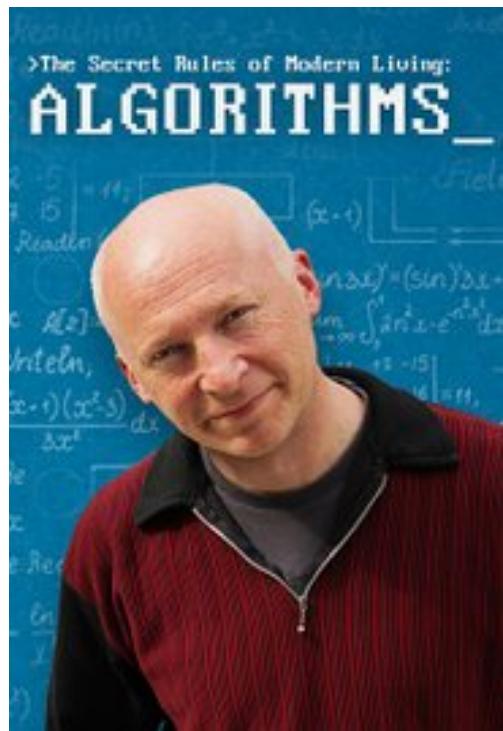
Auto-encoded network now represents the data in a lower-dimensional space

Outputs of hidden networks can be used as feature-vectors

Network can “de-noise” future inputs (project new inputs onto data-space)



RIGHT ALGORITHMS



There is no magic solution to your problem!



#OpenDataScienceMeans
#AnacondaCON

RIGHT ALGORITHMS

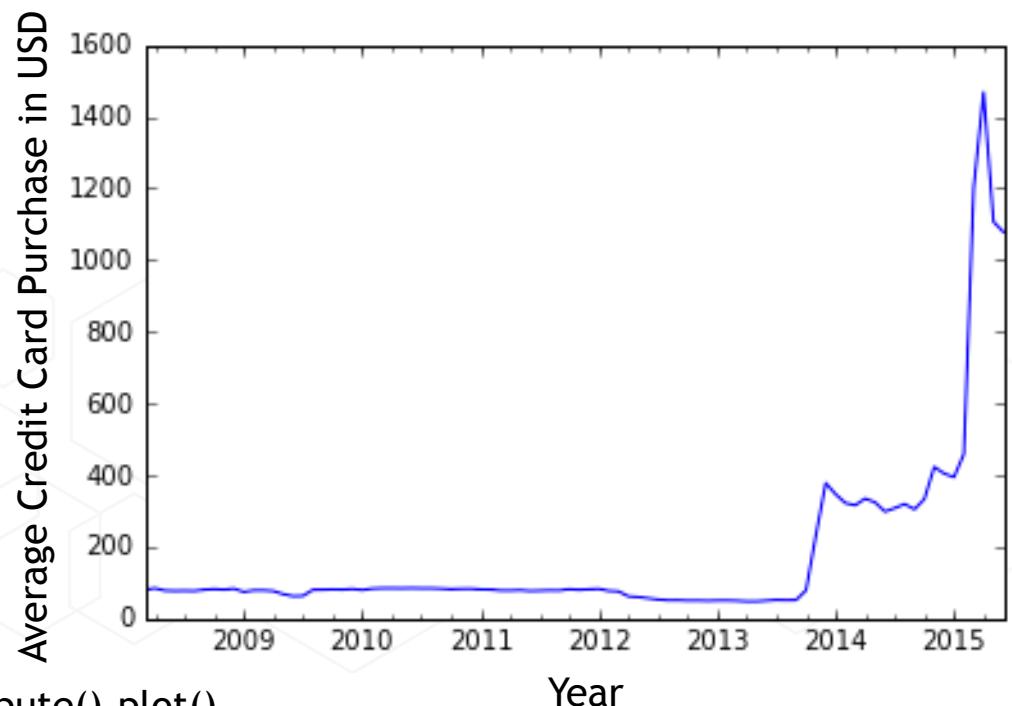
~2300 CSV files with >9million transactions over 8 years

4 nodes with
4 cores each

Use `read_csv` and some transformations in parallel to build a distributed data-frame (with ~2300 partitions, one for each file).

```
ddf = ddf.repartition(npartitions=100)  
ndf = ddf.set_index('DATE')  
ndf.persist()  
ndf.AMOUNT.resample('1M').mean().compute().plot()
```

Practical Parallelism for Scale



#OpenDataScienceMeans
#AnacondaCON

RIGHT ALGORITHMS

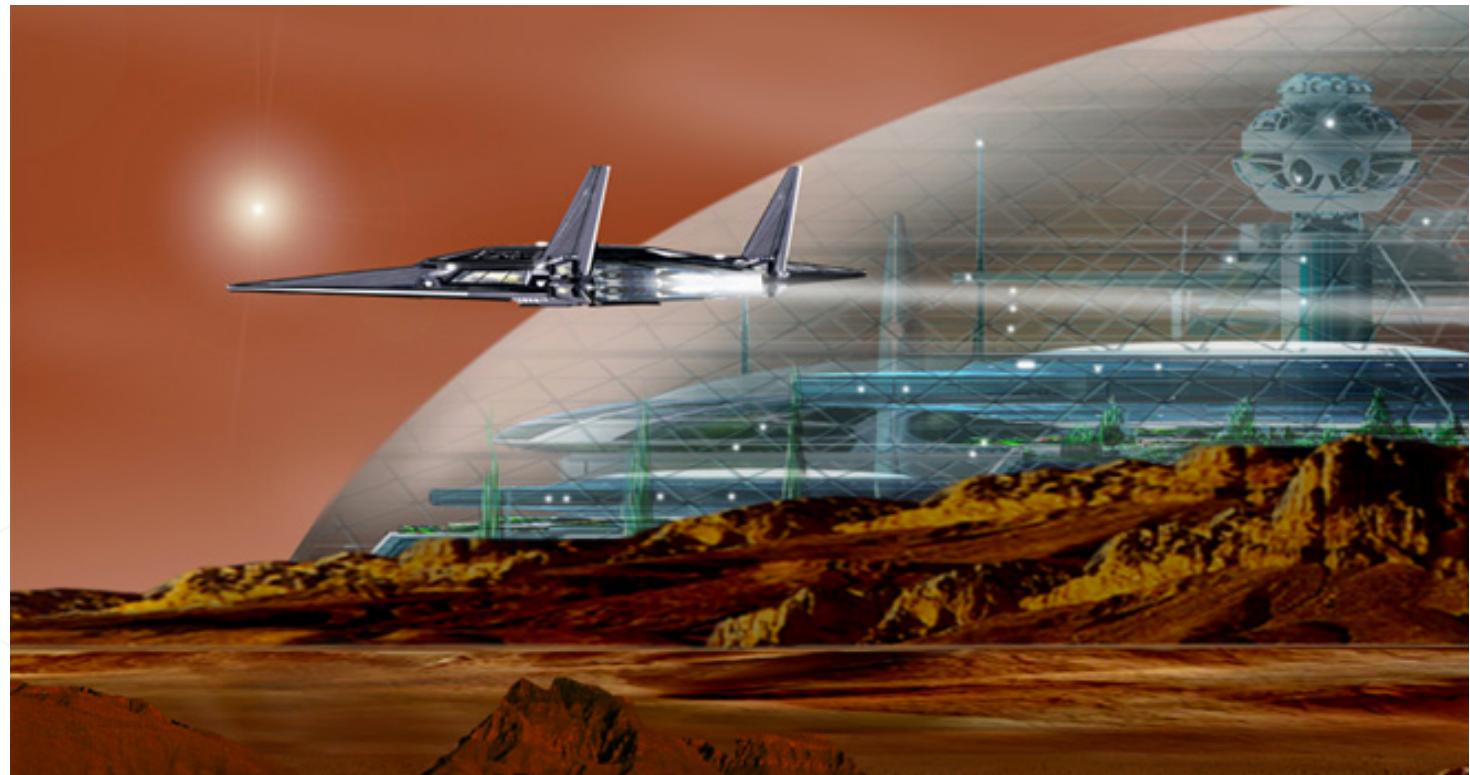
Dask feels modern.

Flexible parallelism

- machine learning
- advanced analytics and modeling
- advanced data munging

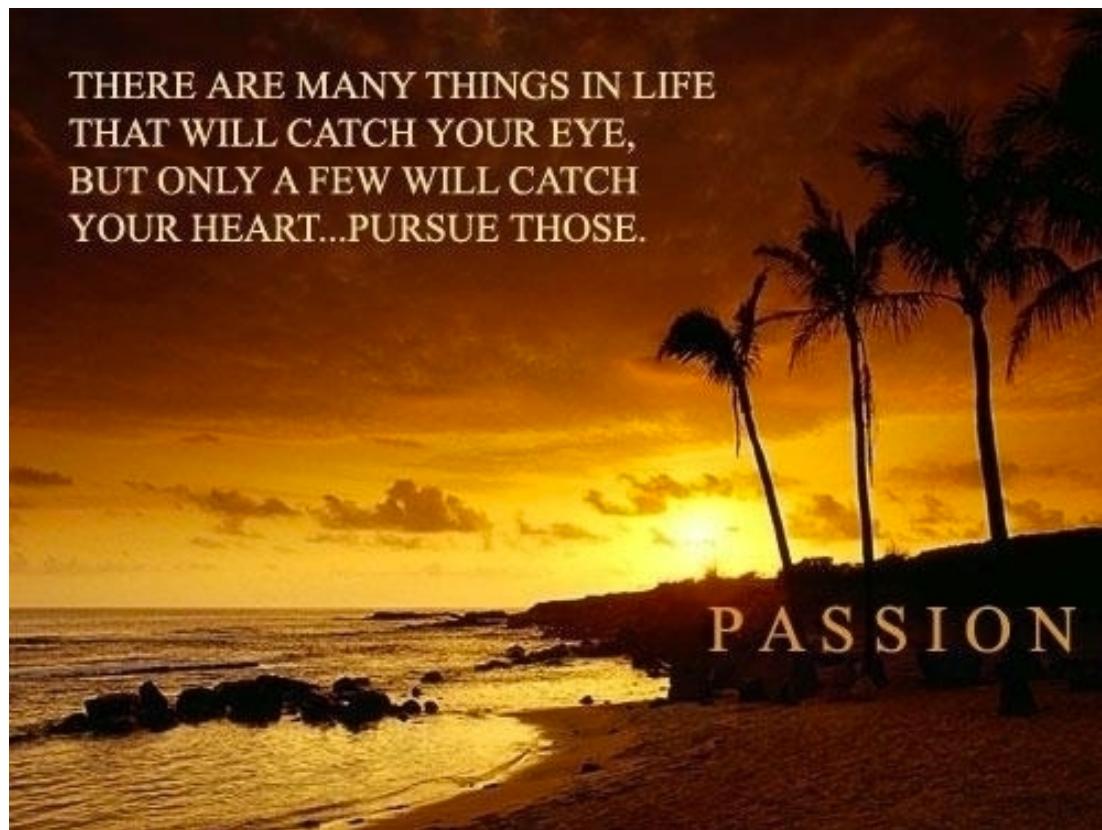


all an import away

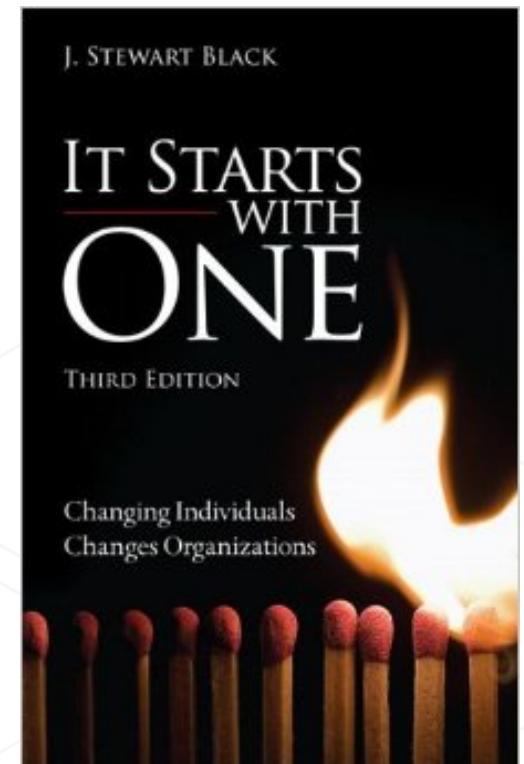


#OpenDataScienceMeans
#AnacondaCON

CULTIVATION OF COMMUNITY



#OpenDataScienceMeans
#AnacondaCON

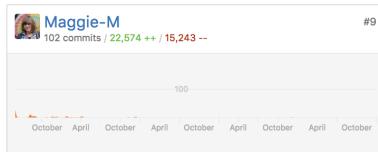


Great works are started by a small group usually 1-3 people).

CONDA

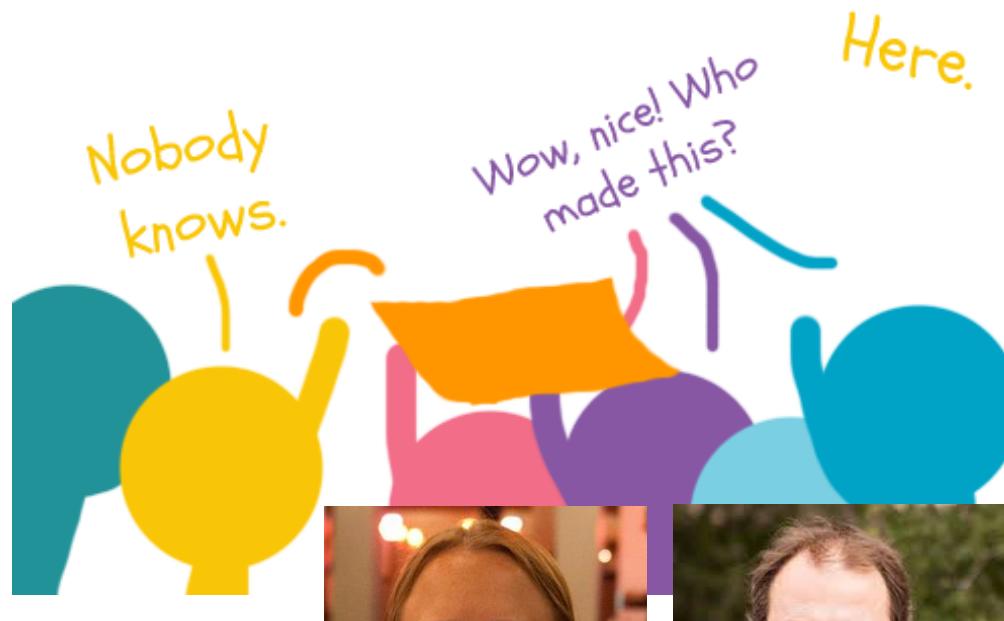


ANACONDA



#OpenDataScienceMeans
#AnacondaCON

CULTIVATION OF COMMUNITY



Crystal Soja



Ilan Schnell



Ray Donnelly

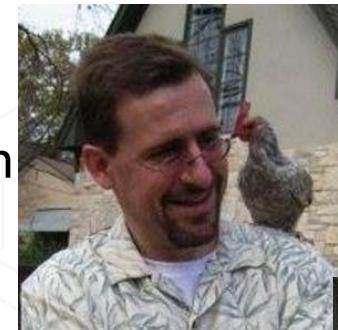


Kale Franz



Michael Sarahan

Some of the
Conda and
Anaconda Team



Bryan Van de Ven



Maggie Mari

Many others....

#OpenDataScienceMeans
#AnacondaCON

CULTIVATION OF COMMUNITY

Python community code of Conduct:

A member of the Python community is:

- Open
- Considerate
- Respectful



#OpenDataScienceMeans
#AnacondaCON

CULTIVATION OF COMMUNITY



**“Diversity is
the mix.
Inclusion is
making the mix
work,” Andrés
Tapia**

www.RedShoeMovement.com



#OpenDataScienceMeans
#AnacondaCON

COMMUNITY PRINCIPLES

Philia
φιλία

Greek word for “sisterly and brotherly love” – the fellowship that should exist between members of the same community.

“the central idea of φιλία is that of doing well by someone for her own sake, out of concern for her (and not, or not merely, out of concern for oneself). [... Thus] the different forms of φιλία [as listed above] could be viewed just as different contexts and circumstances in which this kind of mutual well-doing can arise”

— John M. Cooper



#OpenDataScienceMeans
#AnacondaCON

COMMUNITY PRINCIPLES

Seva
(Sewa)

A Sanskrit word meaning selfless sacrifice, volunteering for the community

“Helping out is not some special skill. It is not the domain of rare individuals. It is not confined to a single part of our lives. We simply heed the call of that natural impulse within and follow it where it leads us.”

— Ram Dass



#OpenDataScienceMeans
#AnacondaCON

Thiqa
(الثقة)

COMMUNITY PRINCIPLES

Arabic word for trust, belief, and confidence.

"The highest form a civilization can reach is a seamless web of deserved trust." "The right culture, the highest and best culture, is a seamless web of deserved trust." "Not much procedure, just totally reliable people correctly trusting one another. That's the way an operating room works at the Mayo Clinic."

— Charles T. Munger



#OpenDataScienceMeans
#AnacondaCON

COMMUNITY PRINCIPLES

sustainability

The property of biological systems to remain diverse and productive indefinitely

Open Data Science is so important that it must be connected to a bustling marketplace. The principles of community and trade are the foundation of a peaceful and wealthy world. Love, service, and trust are how you build a professional community of buyers and sellers that lifts everyone.



#OpenDataScienceMeans
#AnacondaCON

**Insight
Decisions
Actions
Results
Creations**

**Open Data
Science**

DATA



#OpenDataScienceMeans
#AnacondaCON

Applications

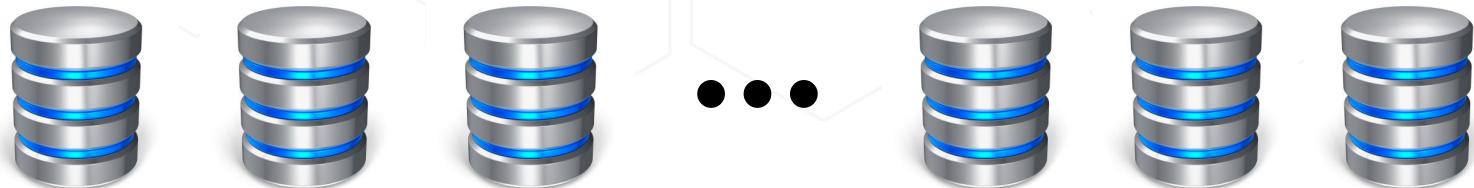
Dashboards

Notebooks

• • •

MicroServices

ANACONDA®



You are now part of the revolution!

Come back next year to
AnacondaCON '18

Will be in April in Austin.

Make the world better this
year with Anaconda!



#OpenDataScienceMeans
#AnacondaCON

