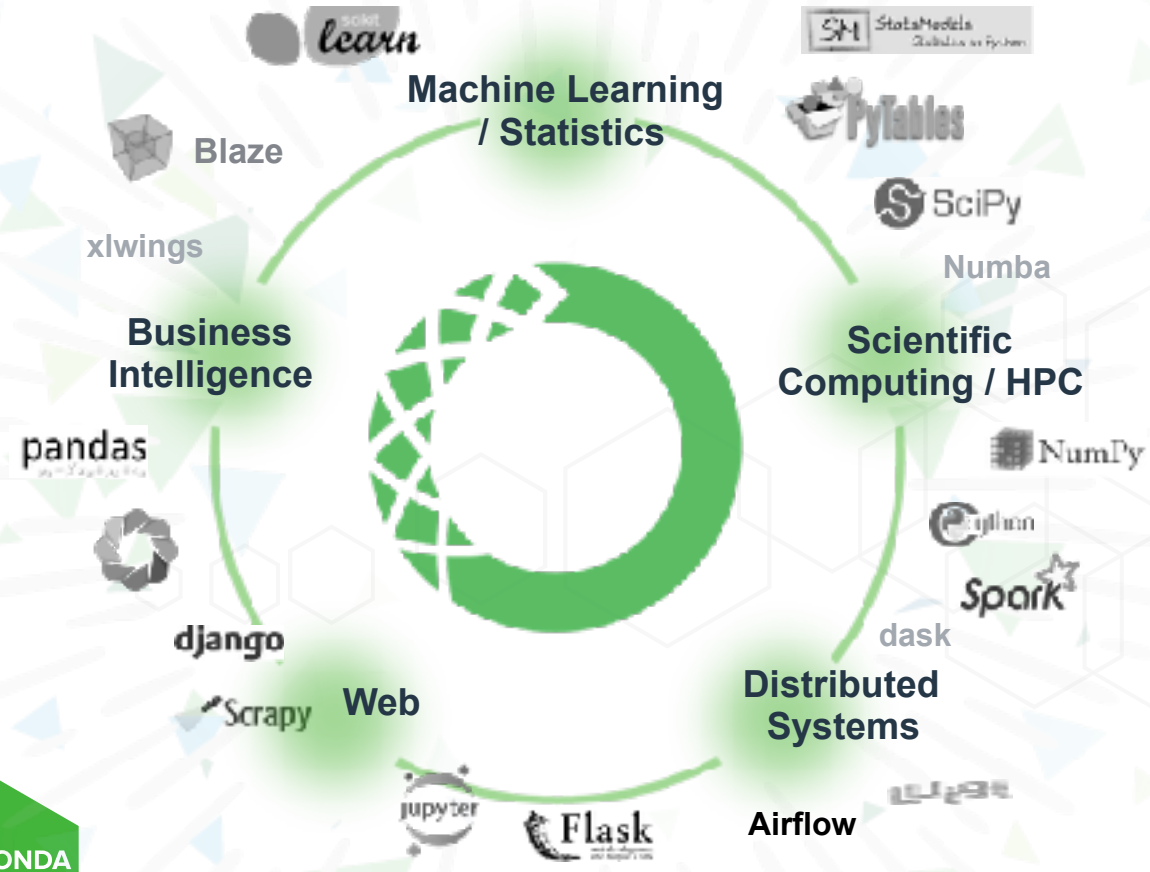

ANACONDA
CON
AUSTIN, TX | 2017
REACHING THE FULL POTENTIAL
OF A DATA-DRIVEN WORLD
Growing the power of Anaconda and its Community

ANACONDA IS OPEN DATA SCIENCE



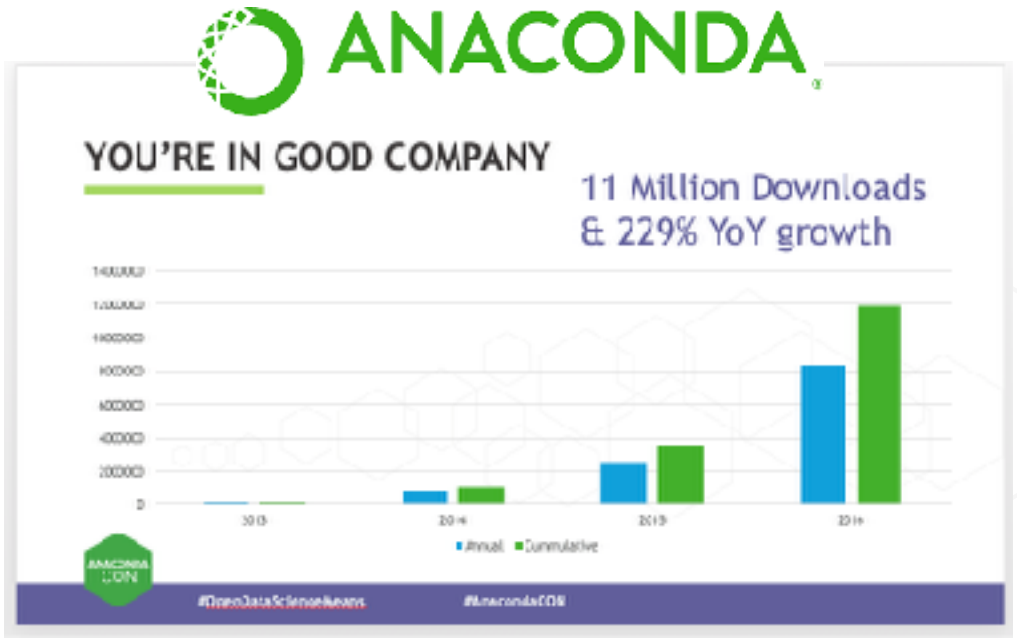
ANACONDA

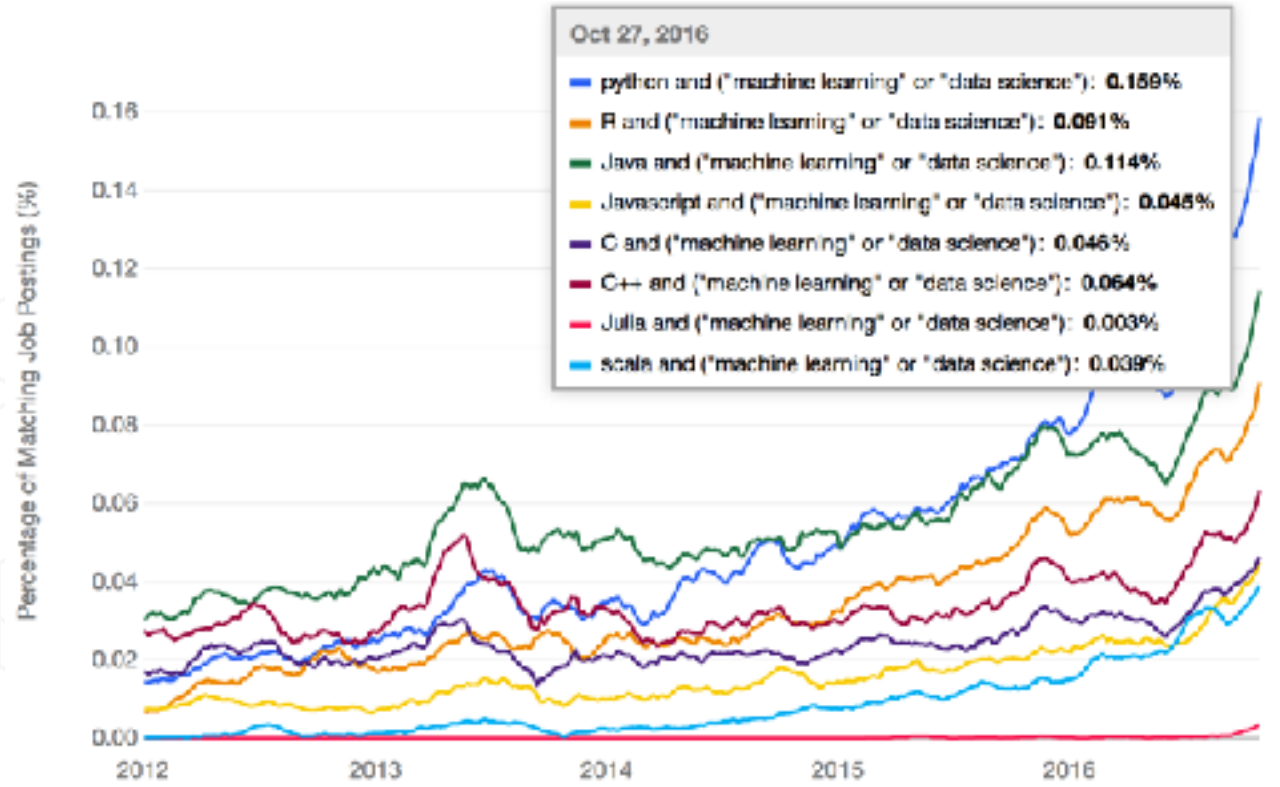
IS COMMUNITY

ANACONDA CON

#OpenDataScienceMeans
#AnacondaCON

GROWTH OF DATA SCIENCE (AND ANACONDA)





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

DATA DELUGE IS JUST BEGINNING

How can Anaconda and its Community help?

How do we find and build solid foundations?

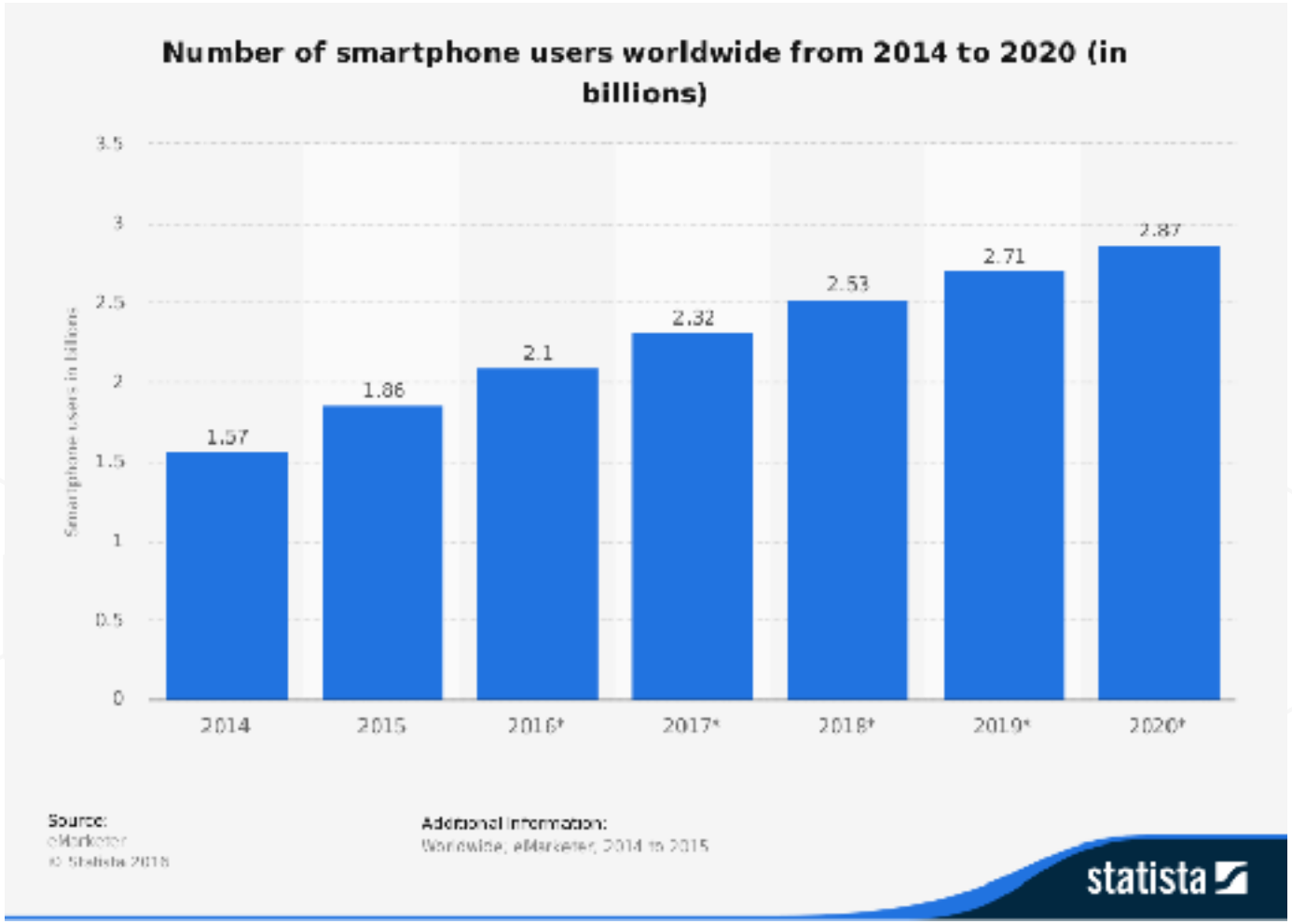
How do we turn today's data deluge into a better life for us, our children, our grandchildren, and beyond.

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

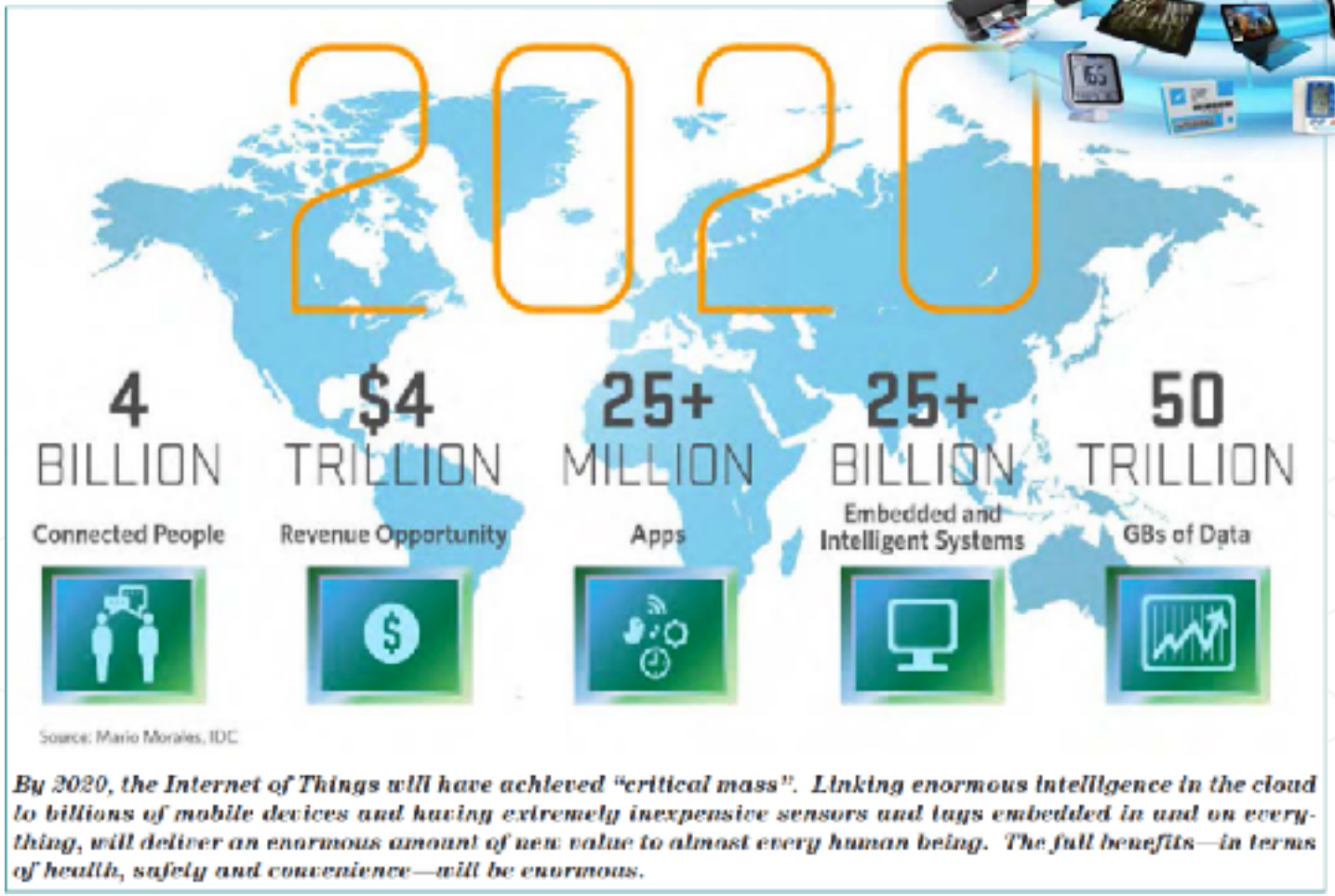


INTERNET OF THINGS…

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

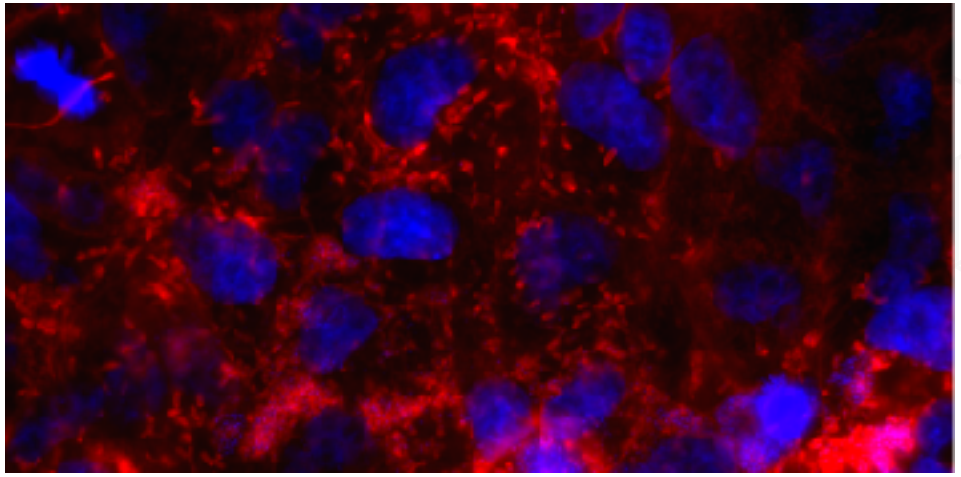
Automate homes

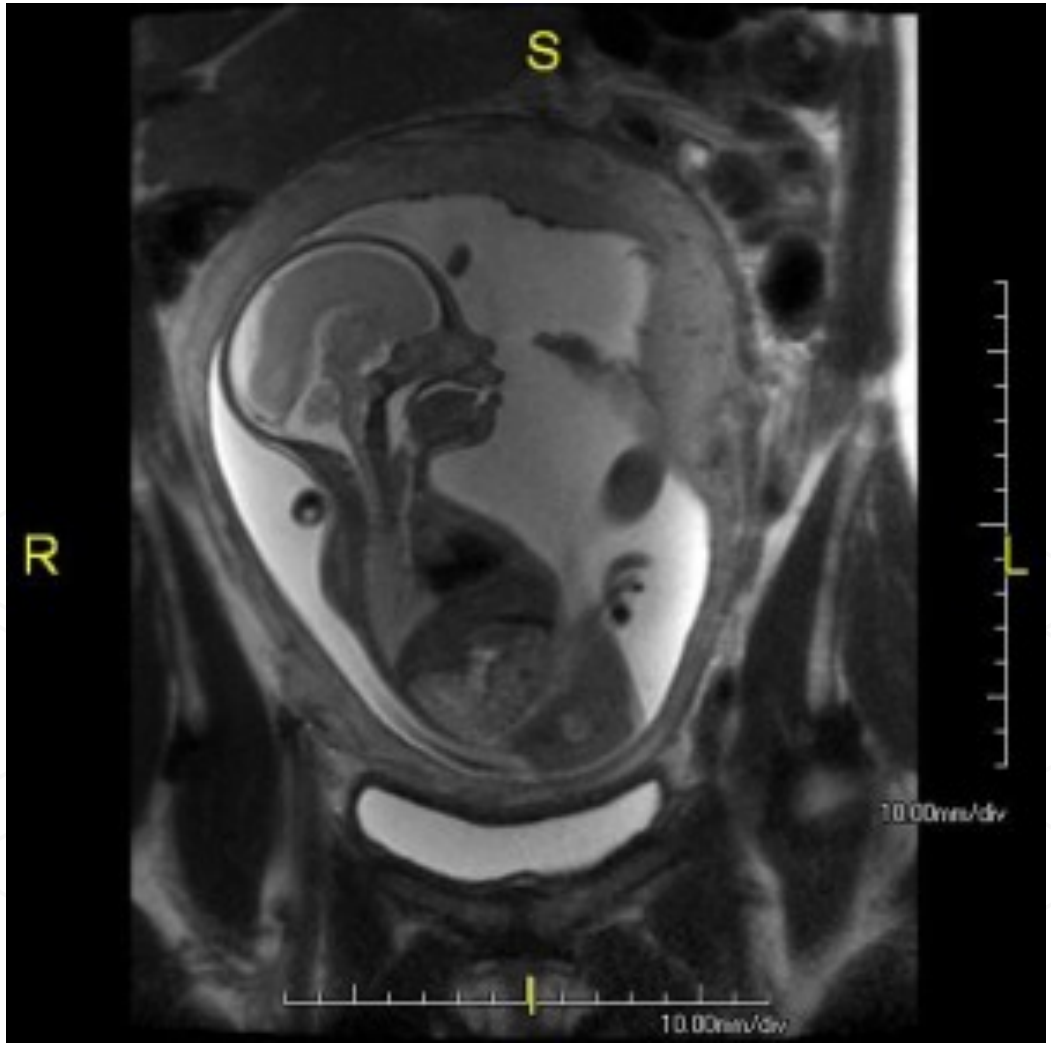
Improve lifetimes of household appliances



HEALTH DATA

Including the “Quantified Self”





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)

SCIENTIFIC DATA

Astronomy (e.g. Hubble)

Satellite Imaging

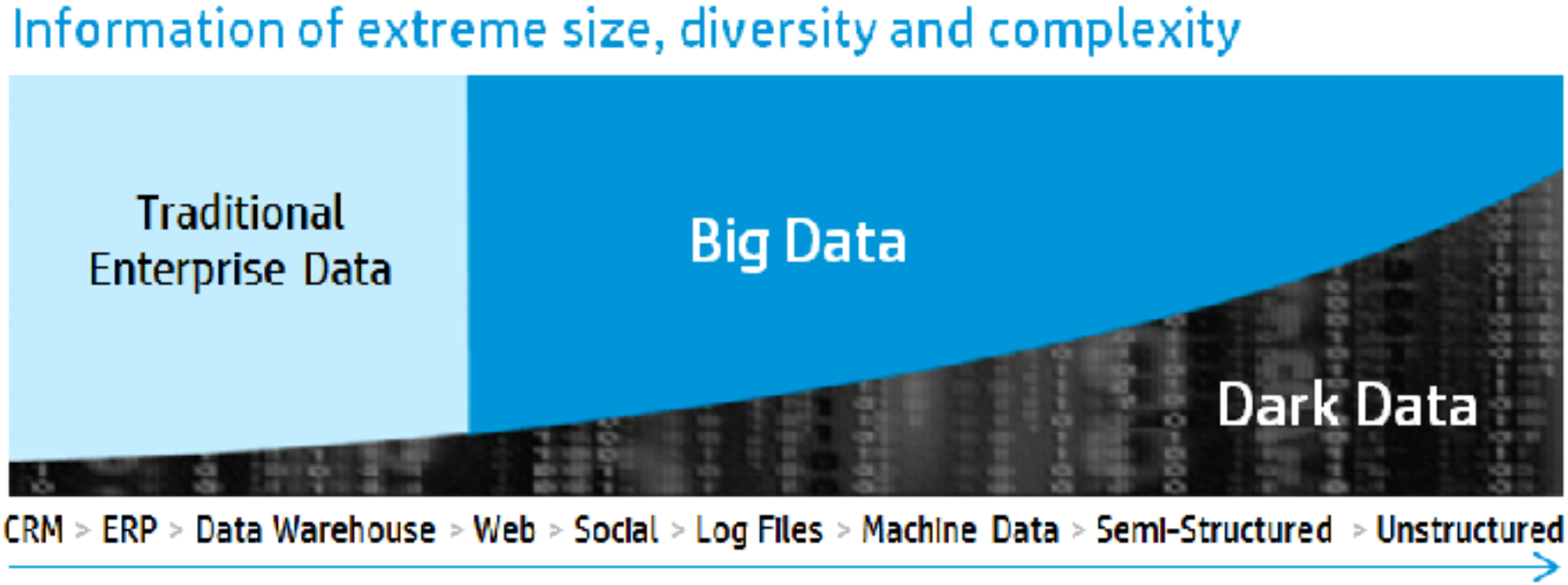
Big Physics (e.g. CERN)

BUSINESS DATA

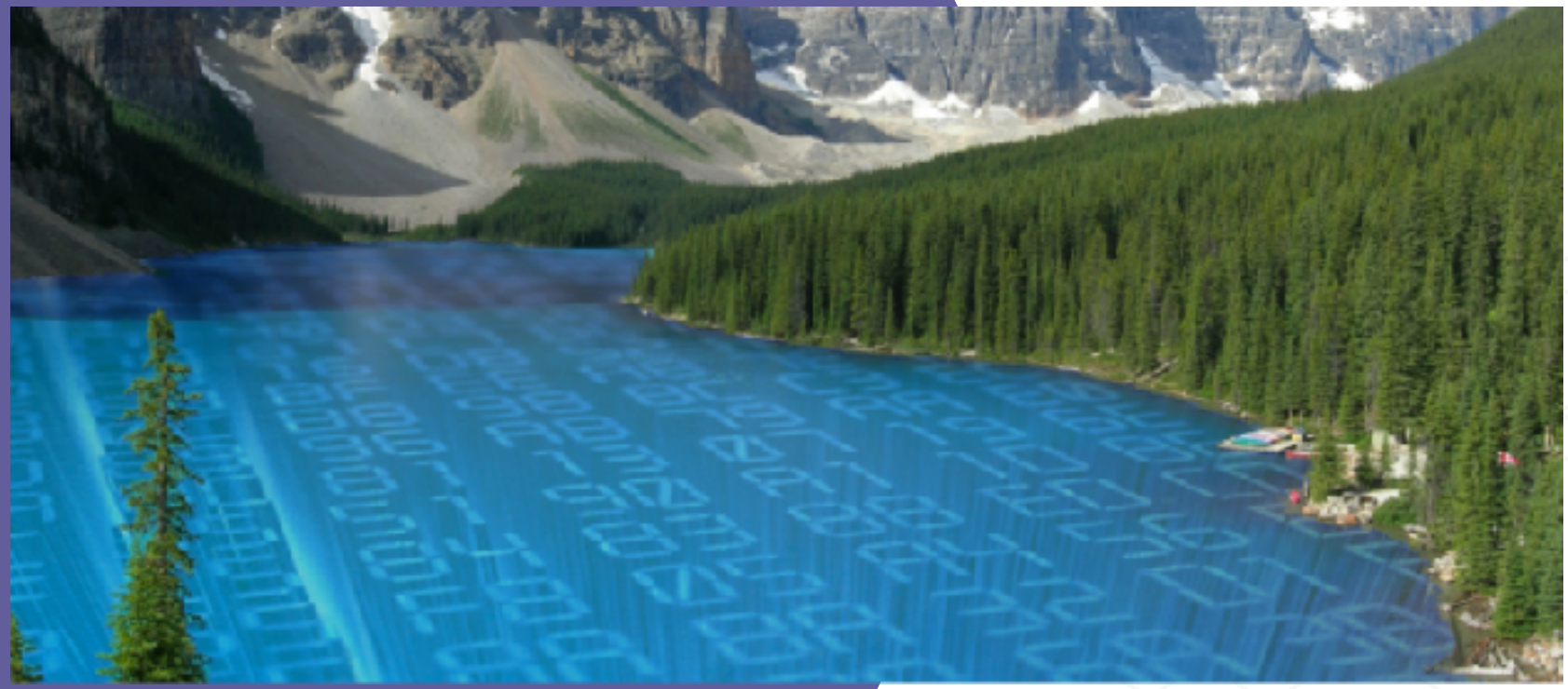
- Customer information
  - Marketing
  - Sales
  - Account management
- Financial Information
- Models for your business
  - Great variety here.
  - Can conceptually use any data from any-source
- Whatever else you store in your ever-growing Databases.

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



SO YOU BUILD YOUR DATA LAKE



ANACONDA CON

#OpenDataScienceMeans
#AnacondaCON

Insight
Decisions
Actions
Results
Creations

Open Data Science

DATA

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

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!

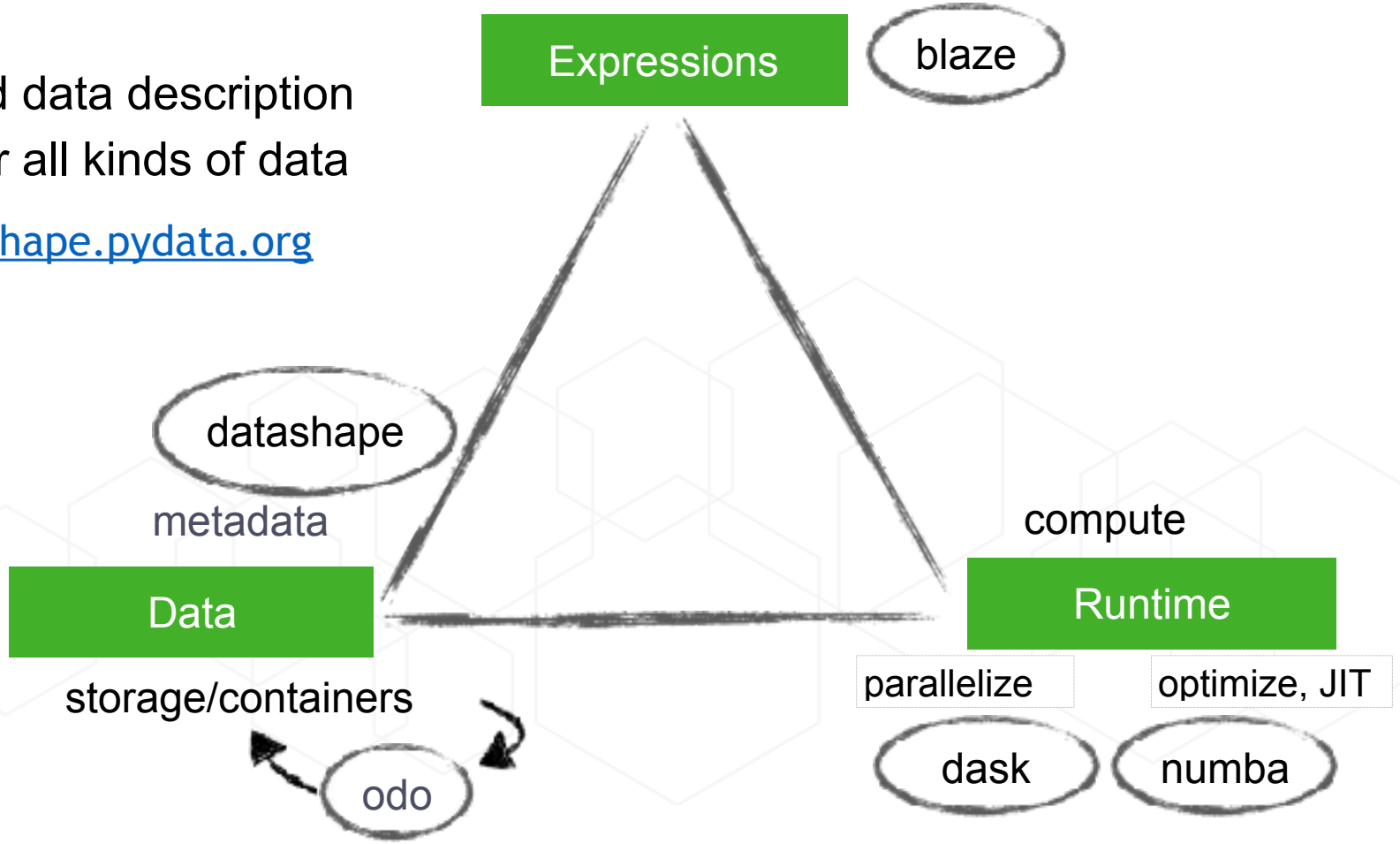
Blaze project still working to solve this!

CODE TO DATA

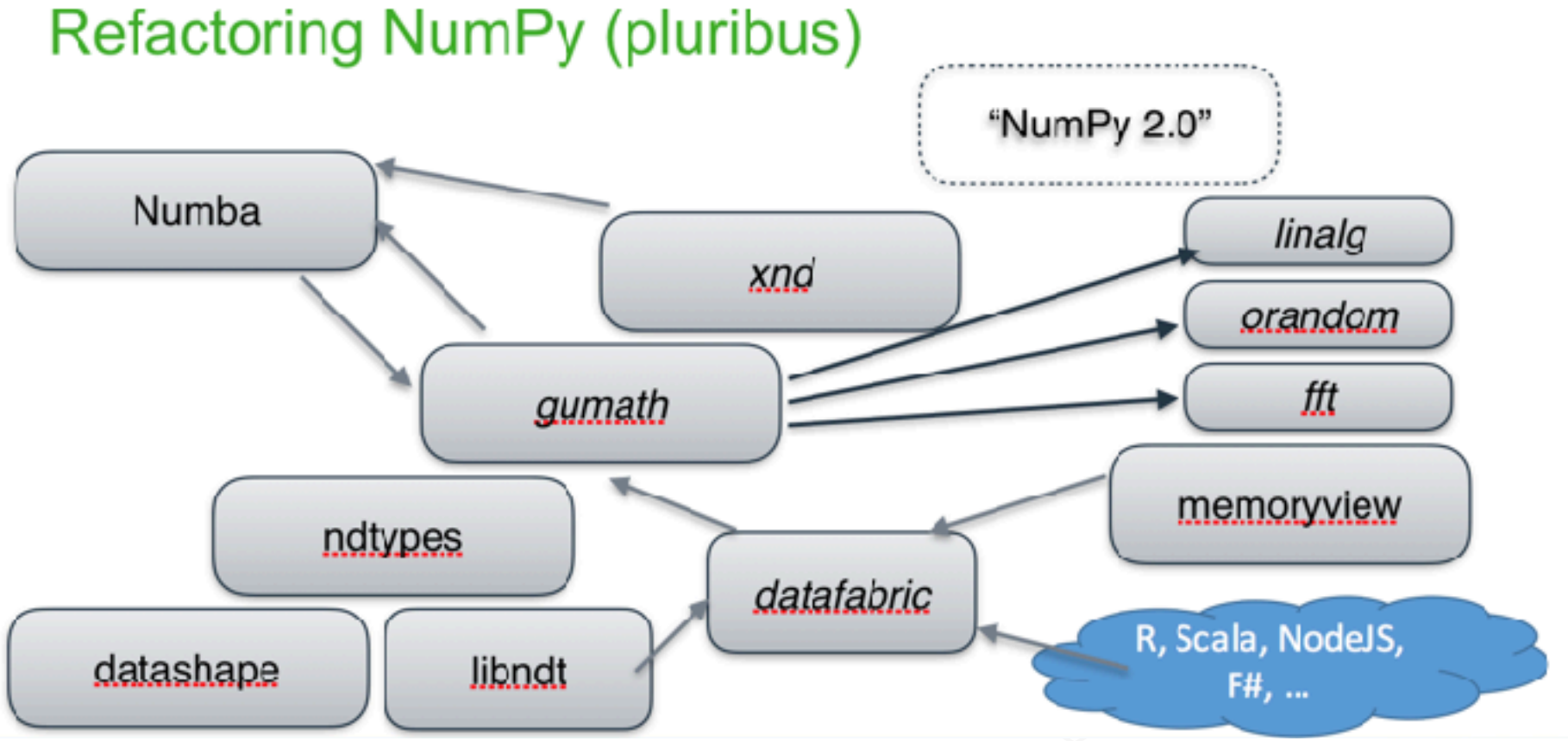
```
{
  flowersdb: {
    iris: var * {
      petal_length: float64,
      petal_width: float64,
      sepal_length: float64,
      sepal_width: float64,
      species: string
      }
    },
  iriscsv: var * {
    sepal_length: ?float64,
    sepal_width: ?float64,
    petal_length: ?float64,
    petal_width: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

http://datashape.pydata.org



CODE TO DATA


Refactoring NumPy (pluribus)
"NumPy 2.0"
Numba
xnd
linalg
orandom
gumath
fft
memoryview
ndtypes
datafabric
datashape
libndt
R, Scala, NodeJS,
F#, ...

AMPLIFY LEARNING

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

190,000
PROJECTED SHORTAGE IN DATA SCIENTISTS BY 2018

DataCamp

galvanize

ANACONDA

Skills Assessment and Data Science Placement Program

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.

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

AMPLIFY LEARNING

ANACONDA®

On-Site Training

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.

3 Days
Python for Software Professionals Using Anaconda
LEARN MORE

4 Days
Data Science Using Anaconda
LEARN MORE

4 Days
Scientific Computing Using Anaconda
LEARN MORE

4 Days
Python for Finance Professionals Using Anaconda
LEARN MORE

4.5 Days Partner Course
Practical Python Programming
LEARN MORE

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.

Machine Learning
LEARN MORE

Data Analysis
LEARN MORE

Data Visualization
LEARN MORE

Performance Optimization
LEARN MORE

Best Software Practices
LEARN MORE

ANACONDA CON

AMPLIFY LEARNING

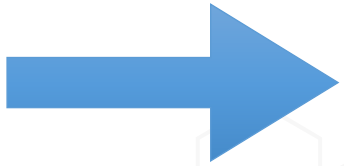
Work to do on right curriculum from elementary school to graduate school!

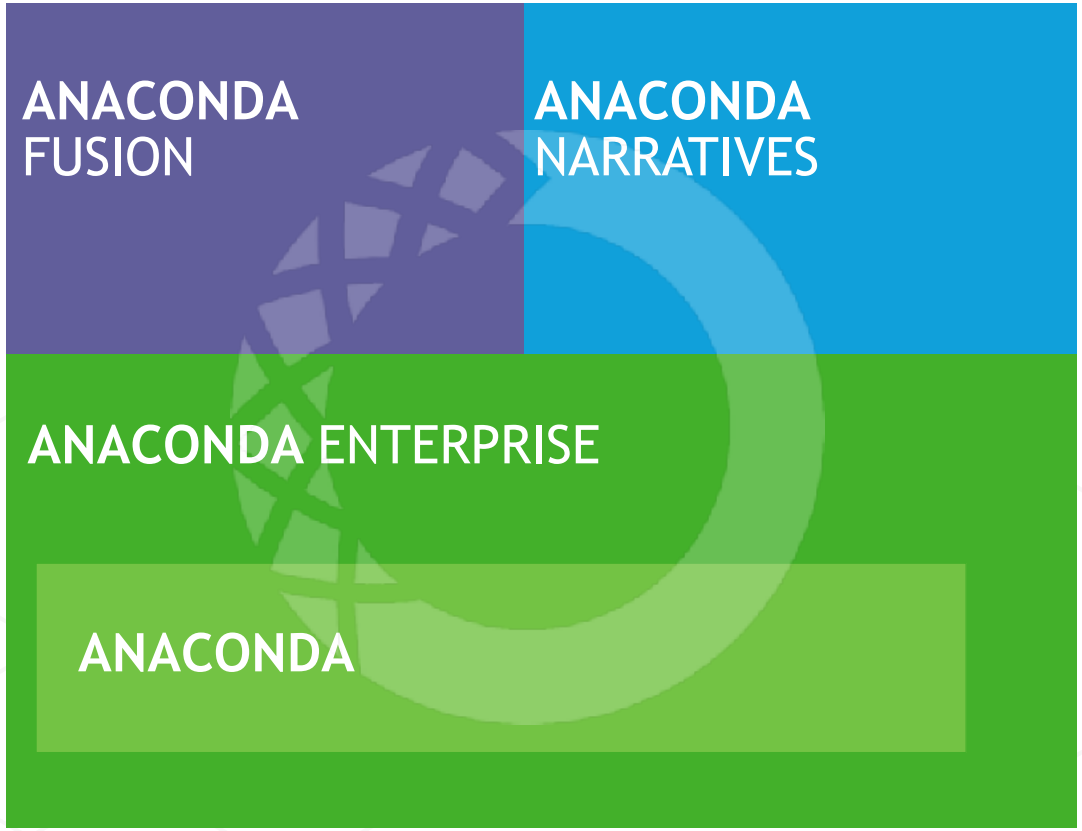
- Basic Probability Theory and Bayes rule taught early (5th grade)
- Partial-derivatives with linear algebra (gradient without divergence and curl)
- Computing literacy
  - array-oriented concepts
  - basic data-structures
  - basic algorithms

http://brilliant.org

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!

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)

RIGHT ALGORITHMS

Supervised Learning

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

$\mathbf{x}$ 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

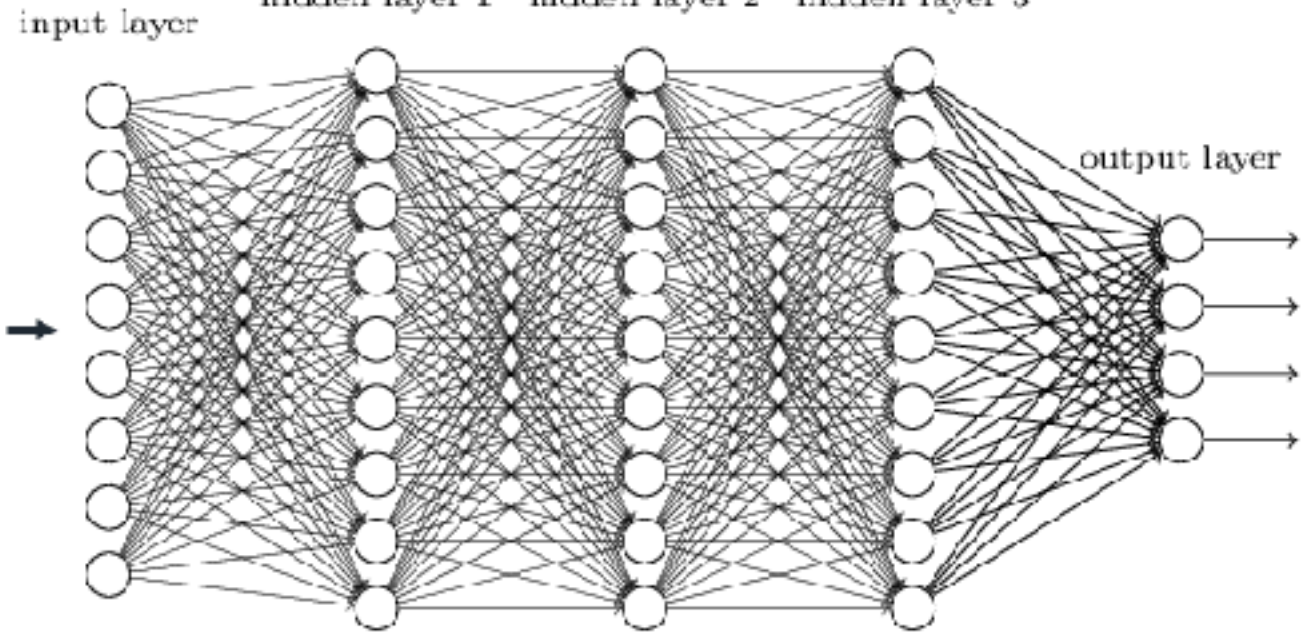
$f(\cdot, \cdot)$ Learning model.

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

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

Deep neural networks learn hierarchical feature representations

hidden layer 1 hidden layer 2 hidden layer 3



Supervised Deep Learning

Same structure

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

$\mathbf{x}$ Input Data or “feature vectors”

y Labels for training

θ All the weights between layers

$$w_{ijk}$$

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

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

$\hat{y}$ Predicted outputs

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

Unsupervised Deep Learning

Auto-encoding

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

$\mathbf{x}$ Input Data or “feature vectors”

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

θ All the weights between layers

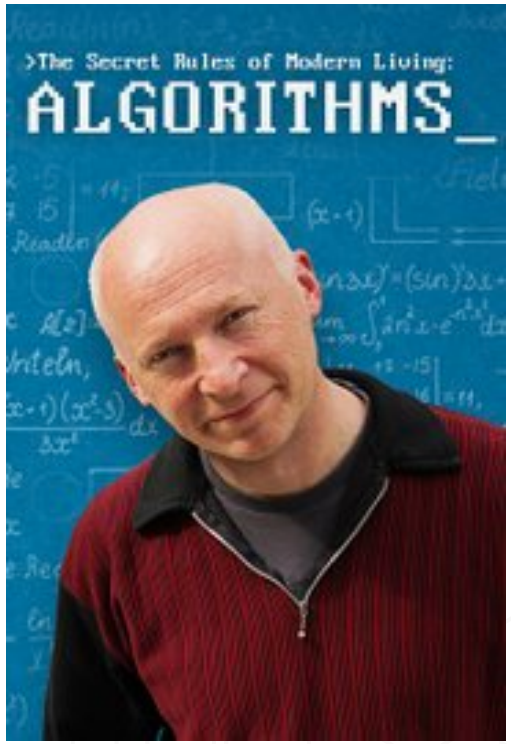
$$z_{ij} = g\left(\sum_k w_{ijk} z_{i-1k}\right)$$

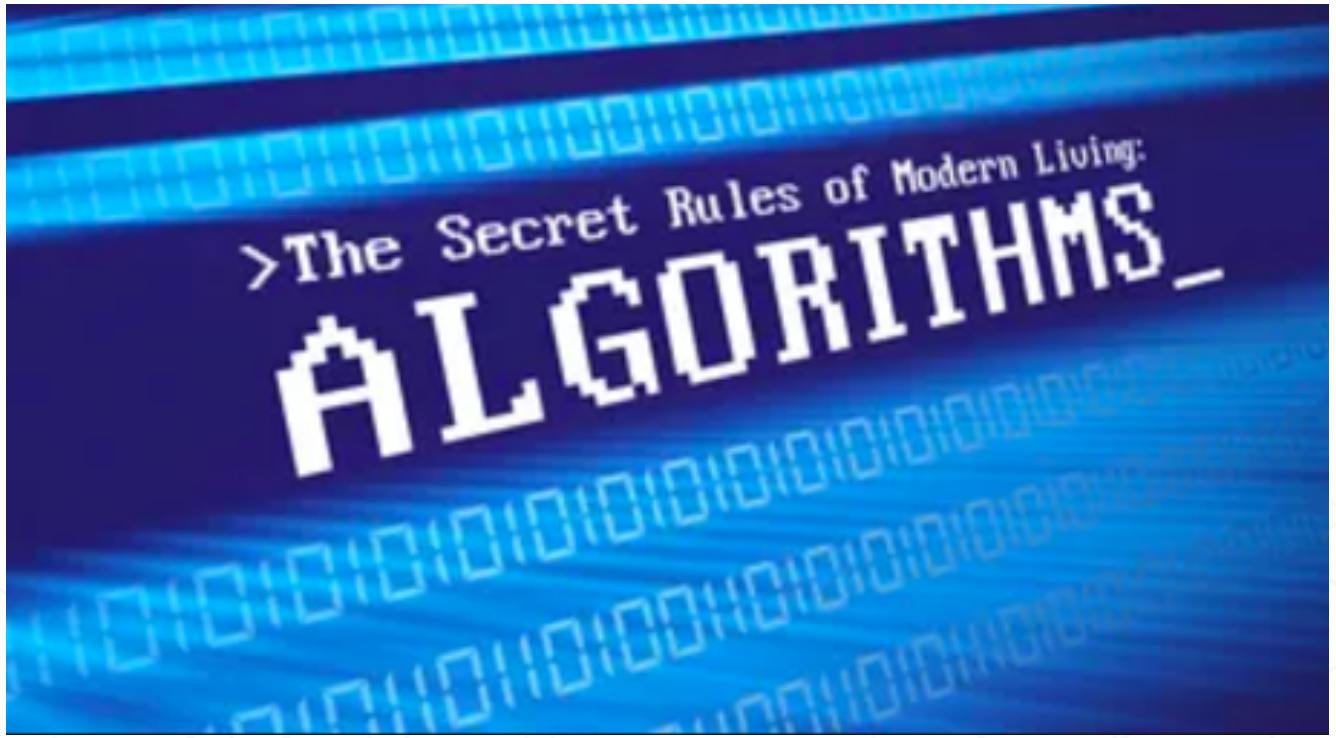
$$g(u) = \frac{1}{1 + e^{-u}}$$

$$w_{ijk}$$

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

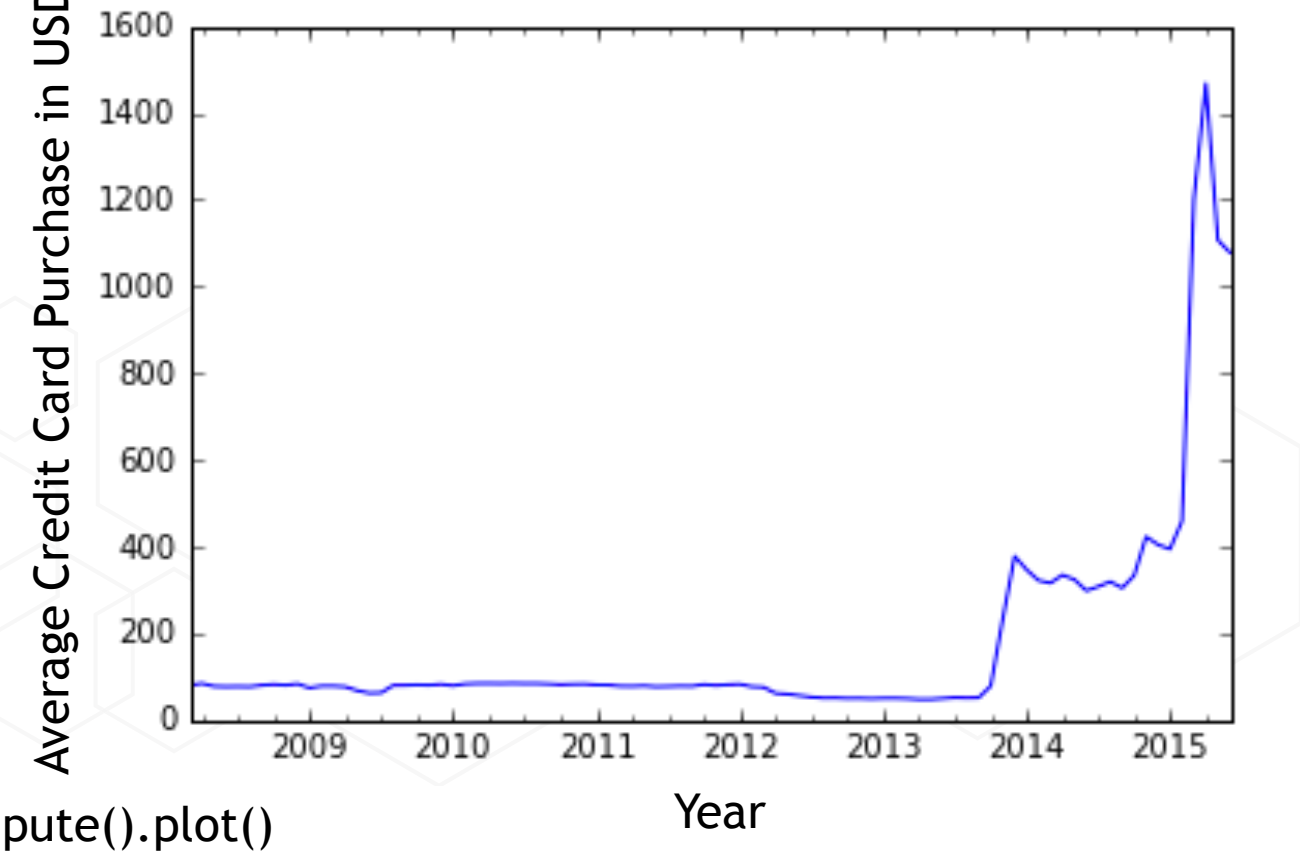
RIGHT ALGORITHMS

Practical Parallelism for Scale

~2300 CSV files with >9million transactions over 8 year from Ashley Madison “hack”

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()
```

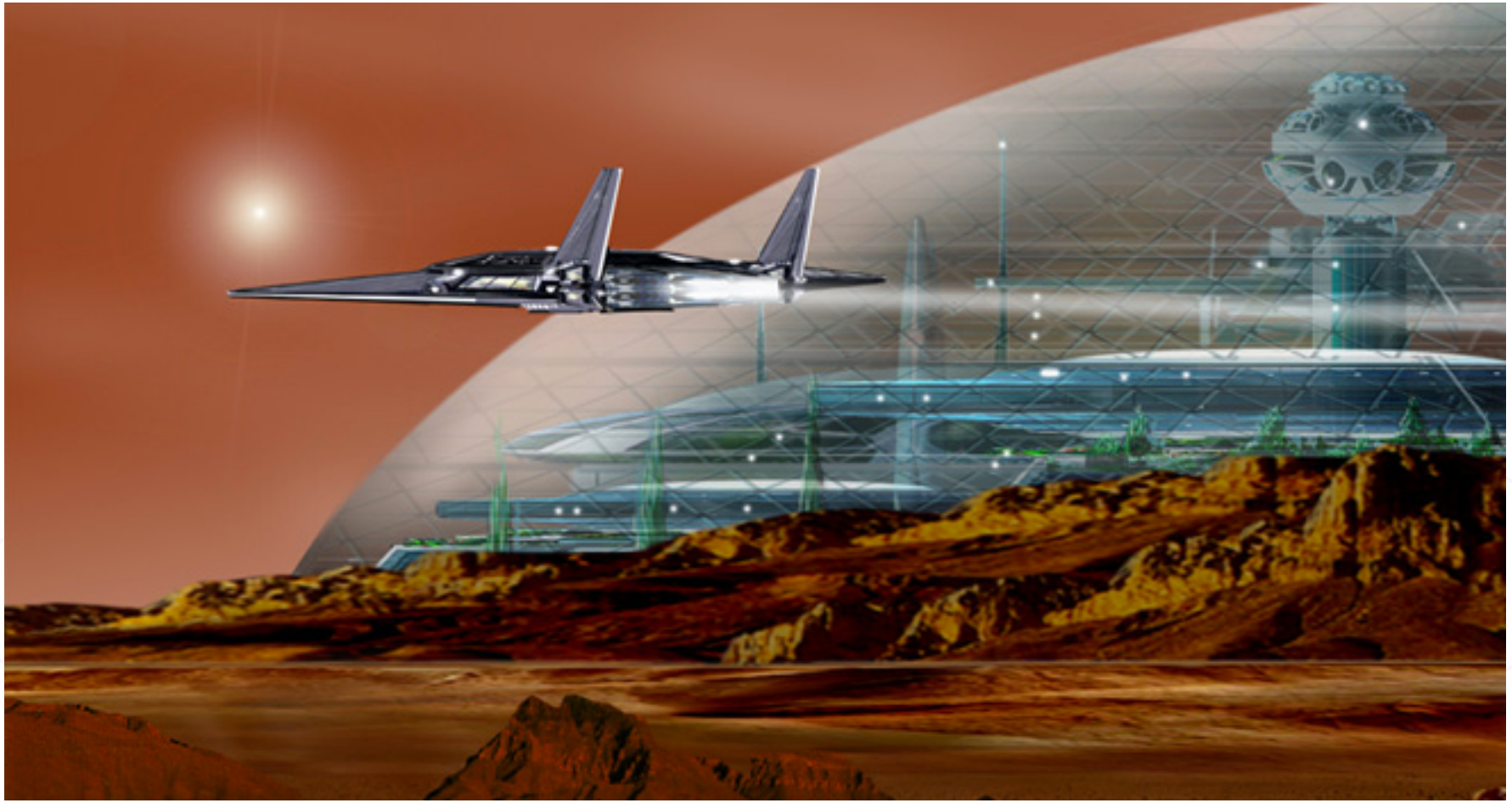
RIGHT ALGORITHMS

Dask feels modern.

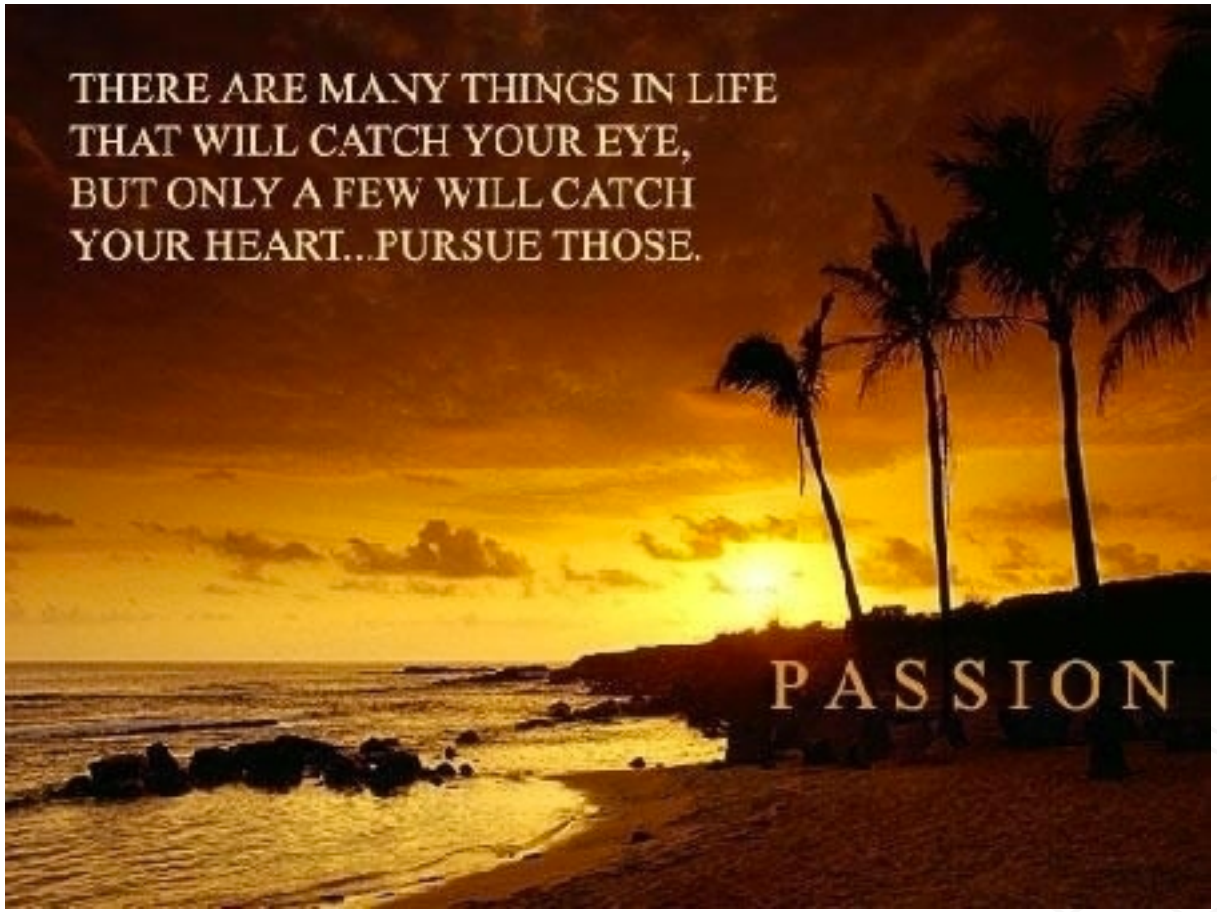
Flexible parallelsim

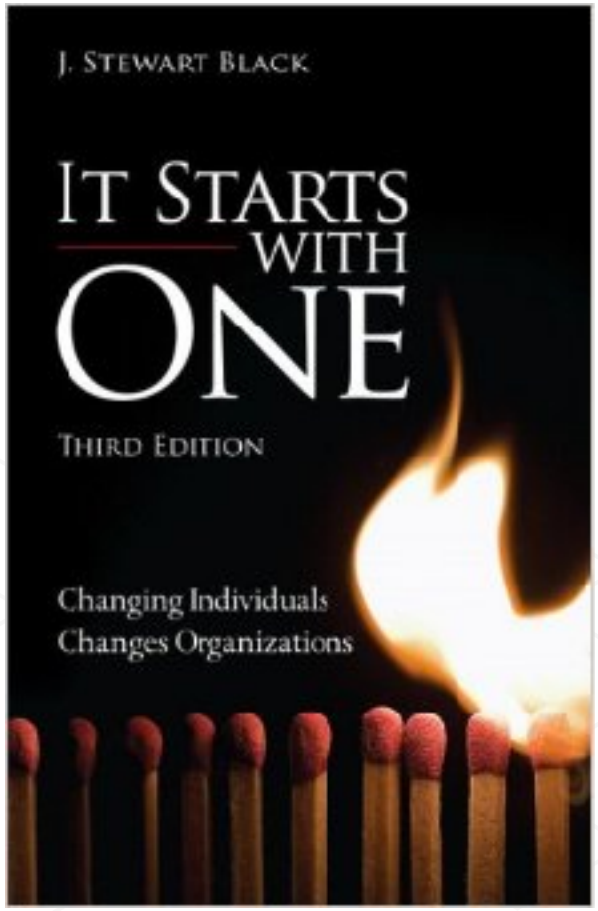
- machine learning
- advanced analytics and modeling
- advanced data munging

all an import away



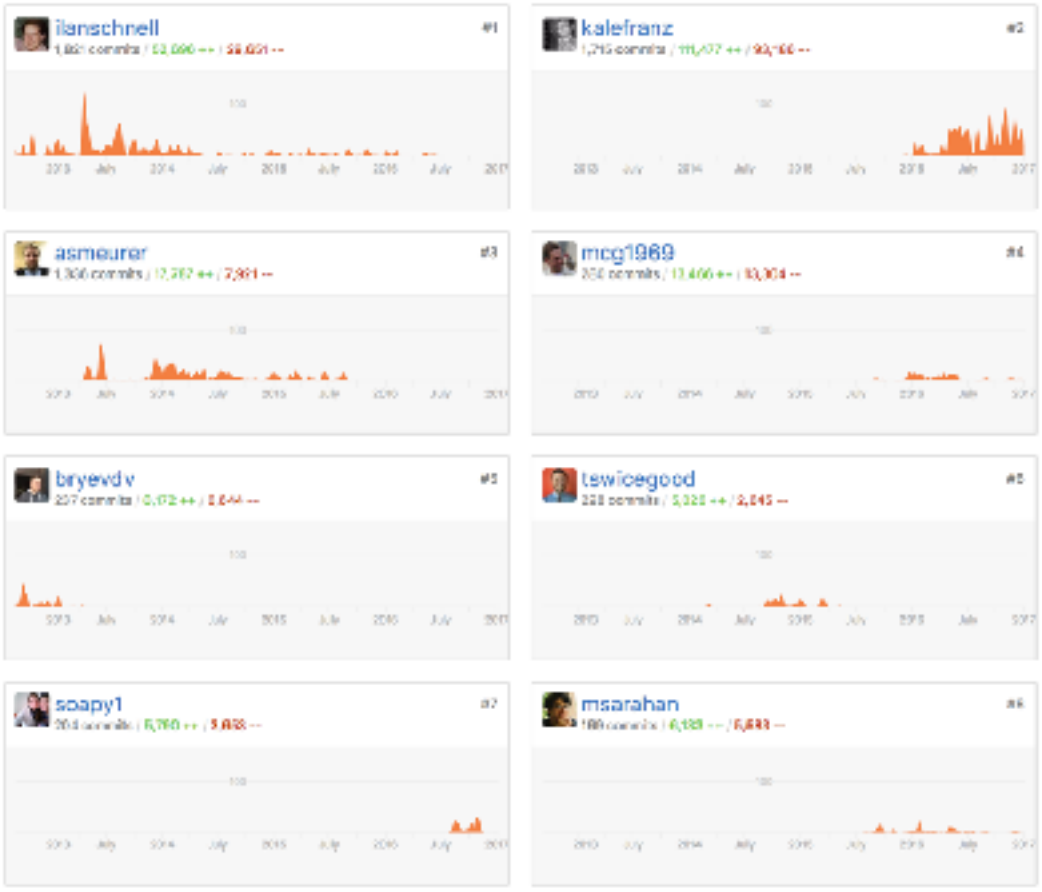
CULTIVATION OF COMMUNITY



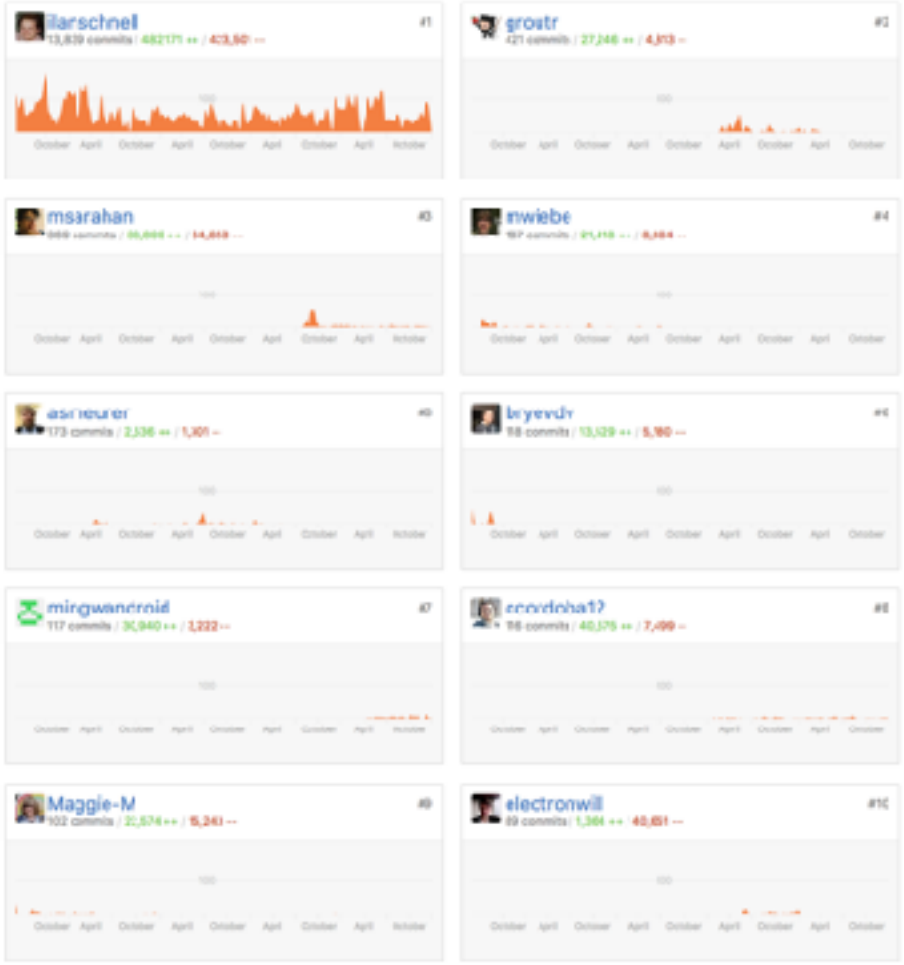


Great works are started by a small group usually 1-3 people).

CONDA



ANACONDA.



ANACONDA CON

#OpenDataScienceMeans
#AnacondaCON

CULTIVATION OF COMMUNITY

Some of the Conda and Anaconda Team

Crystal Soja

Ilan Schnell

Ray Donnelly

Kale Franz

Bryan Van de Ven

Many others....

Michael Grant

Michael Sarahan

Maggie Mari

CULTIVATION OF COMMUNITY

Python community code of Conduct:

A member of the Python community is:

- Open
- Considerate
- Respectful

CULTIVATION OF COMMUNITY

"Diversity is the mix. Inclusion is making the mix work," Andrés Tapia

www.RedShoeMovement.com

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

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

COMMUNITY PRINCIPLES

Thiqa
(الثقة)

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

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.

Insight
Decisions
Actions
Results
Creations

Open Data Science

DATA

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!

ANACONDA CON