@quansightai

https://www.quansight.com

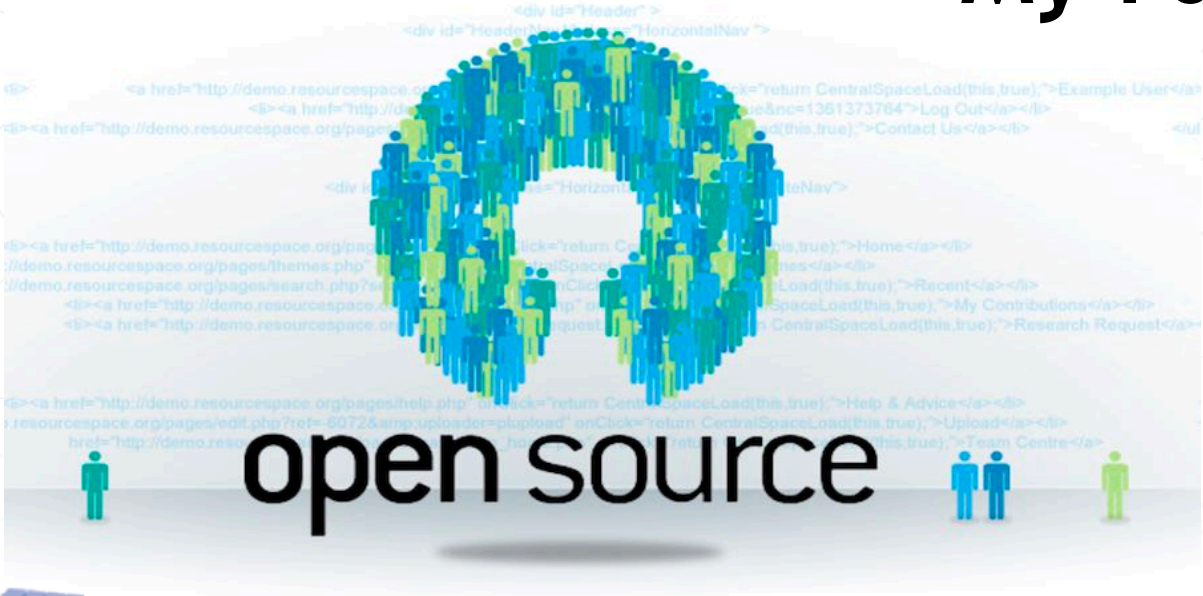
SciPy/PyData/NumFOCUS Communities: Past, Present, and Future

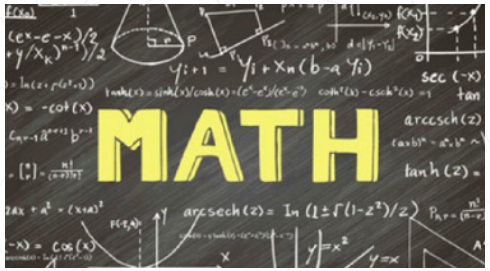
SciPy LATAM
Bogotá Colombia, Universidad de Los Andes
October 9, 2019

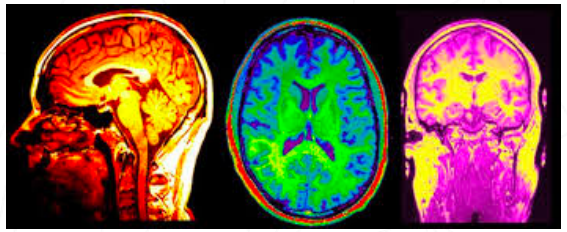
travis@quansight.com
@teoliphant

ANACONDA.

My Passions



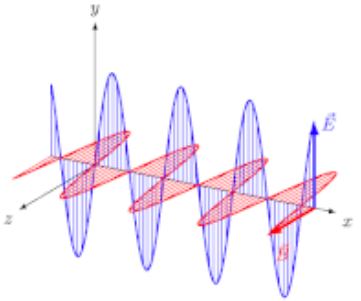




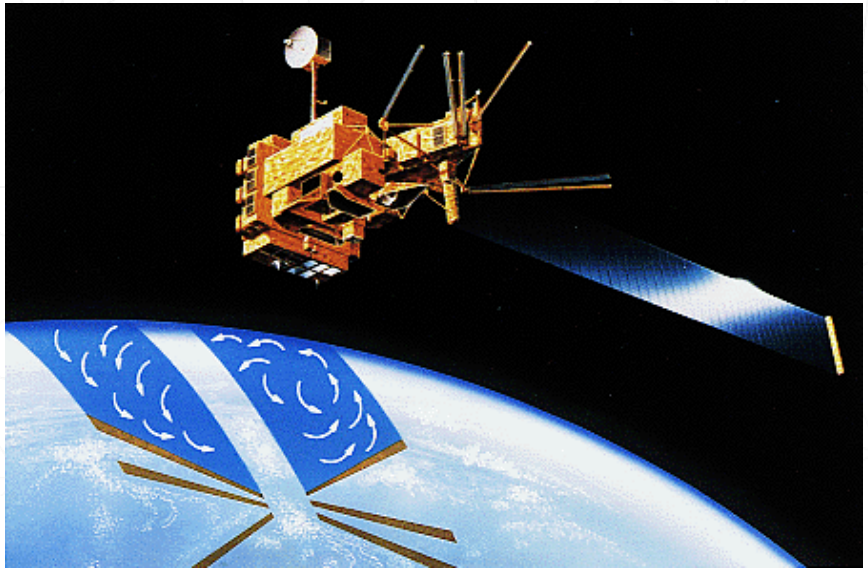
NumPy

python™

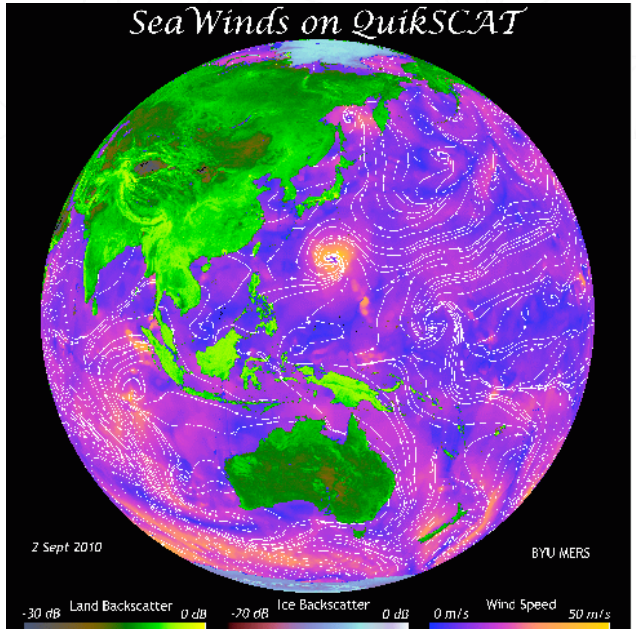
jupyter SciPy

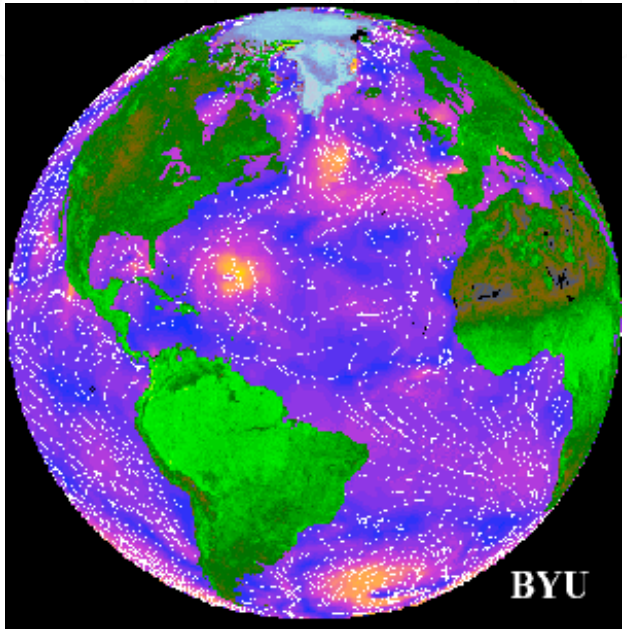


Started my career in computational science



Satellites Measure Backscatter





Computer Algorithms Produce Estimate of Earth Features

- Wind Speed
- Ice Cover
- Vegetation
- (and more)

MAYO CLINIC

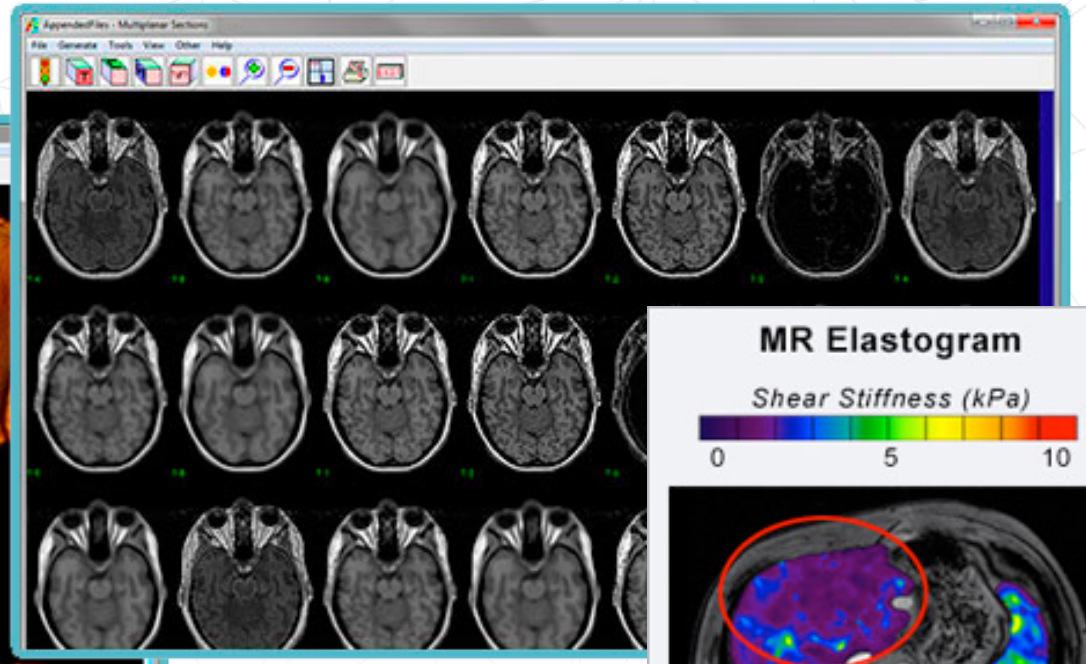
1996 - 2001

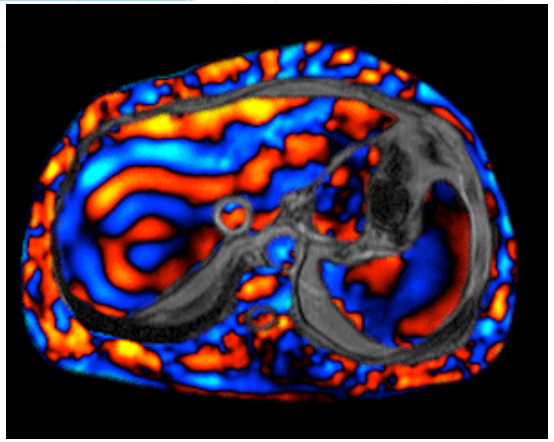
Bringing “SciFi” Medicine to Life since 1971

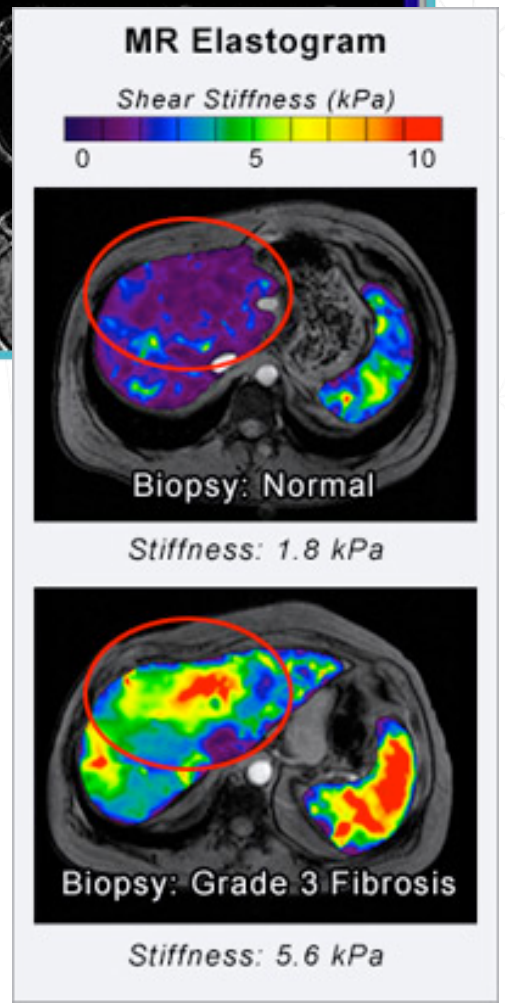
Richard Robb
Retired in 2015

Analyze 12.0

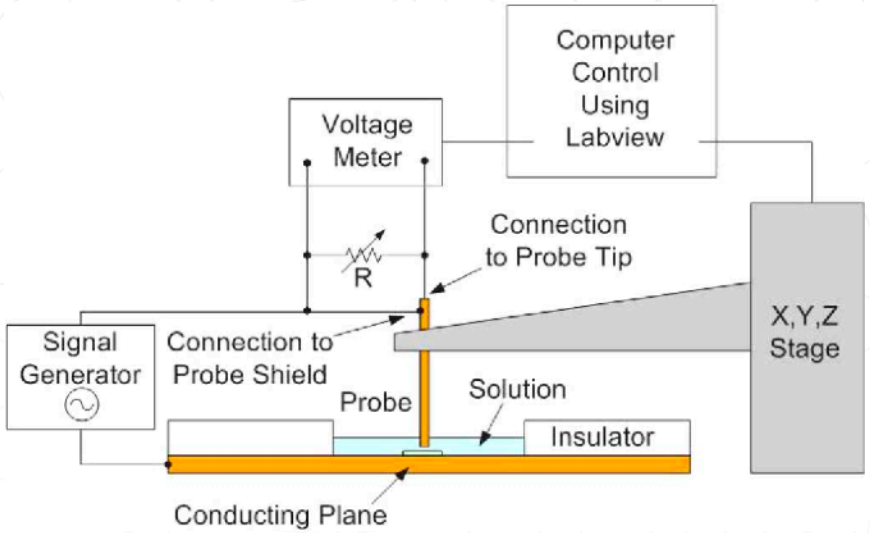
https://analyzedirect.com/

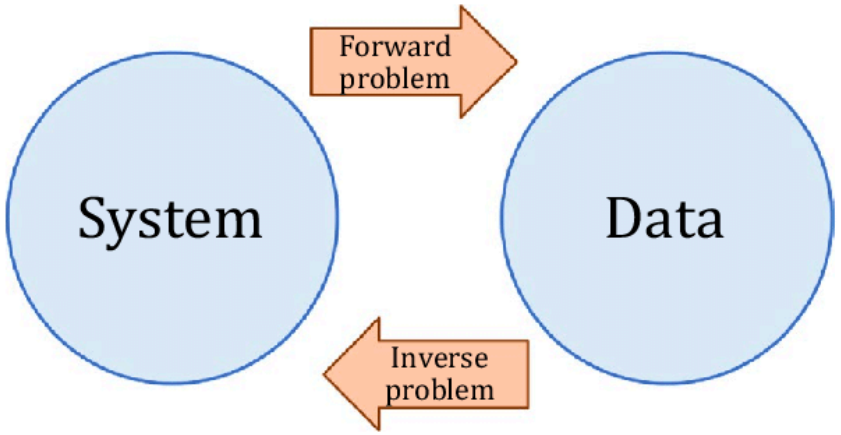


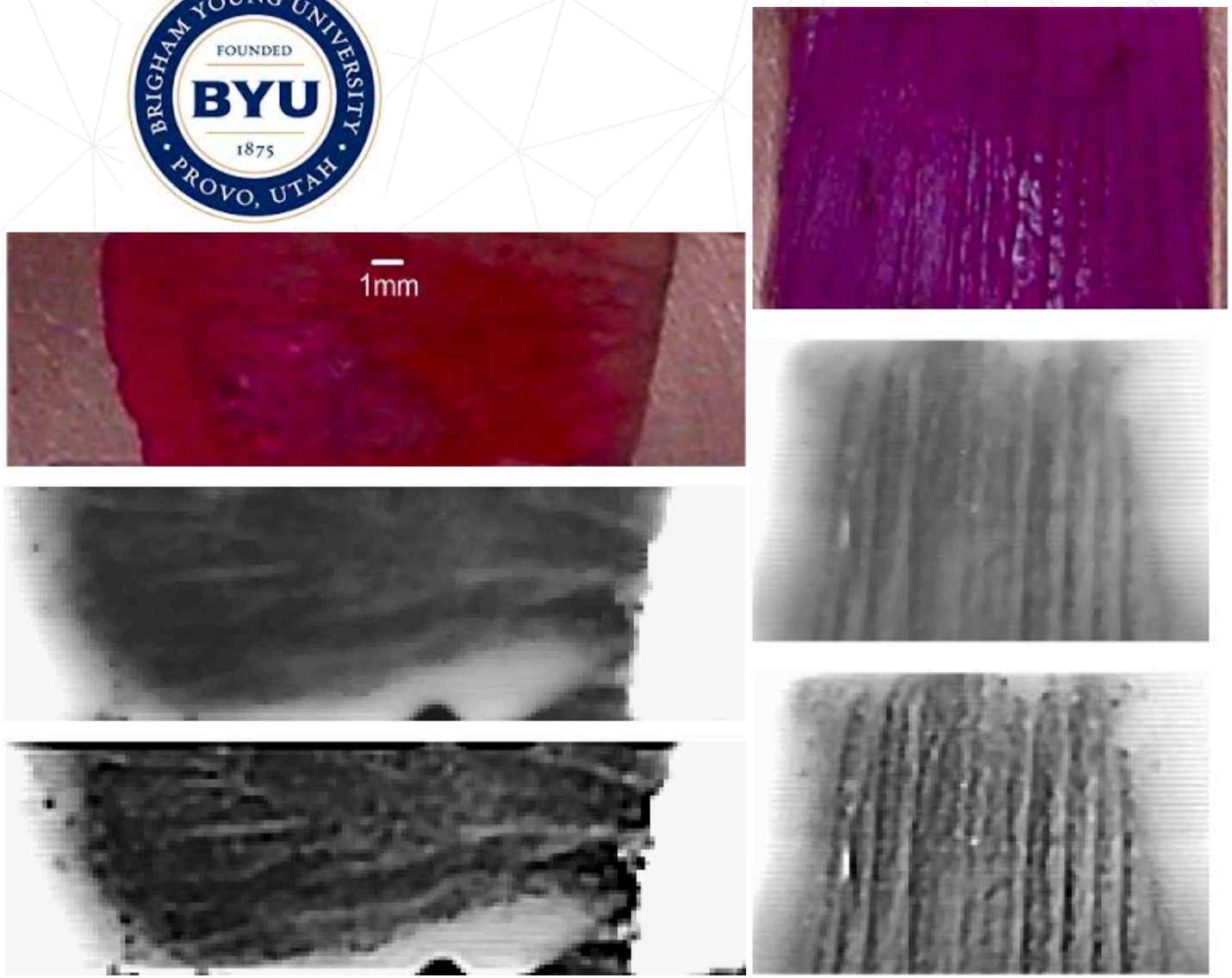




Professor at BYU



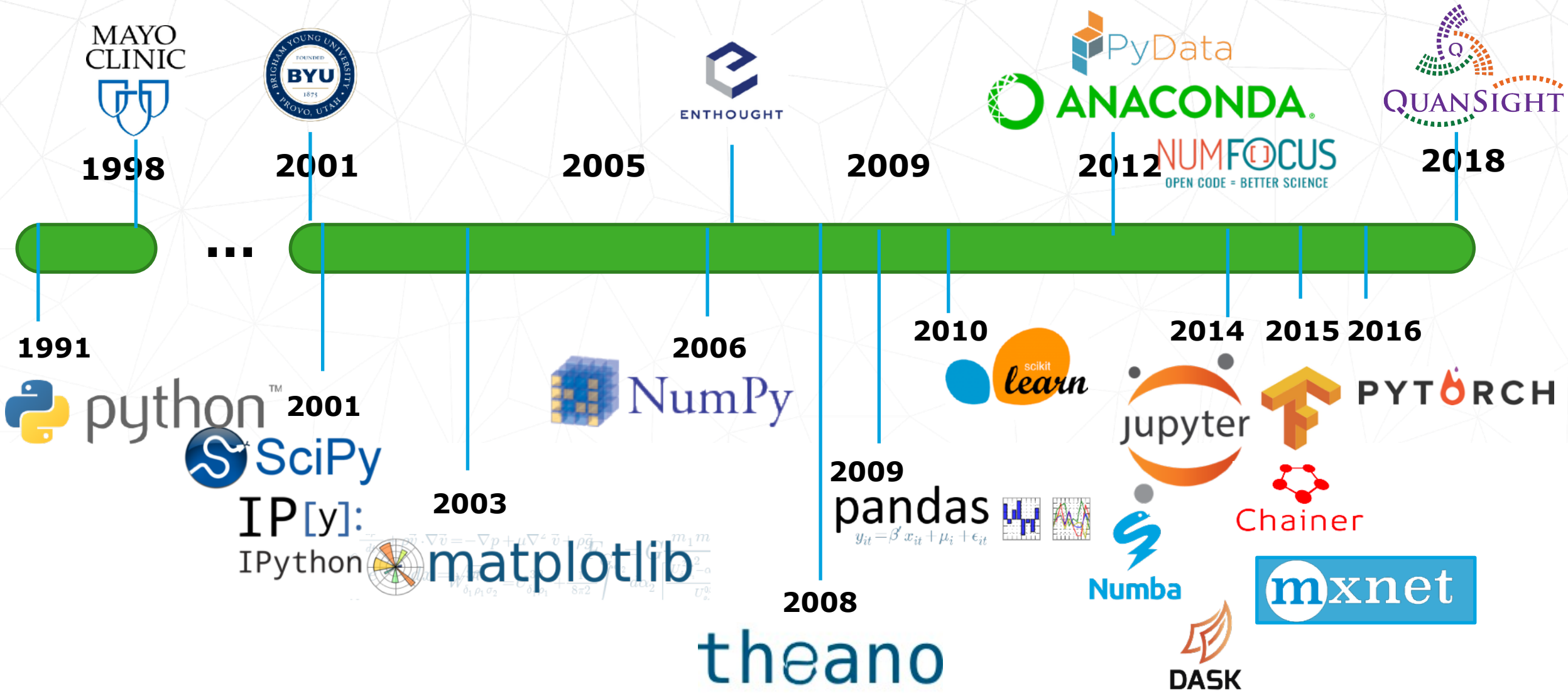




Scanning Impedance Imaging

My little “side projects” became my life

Python Data and ML/AI Time-Line

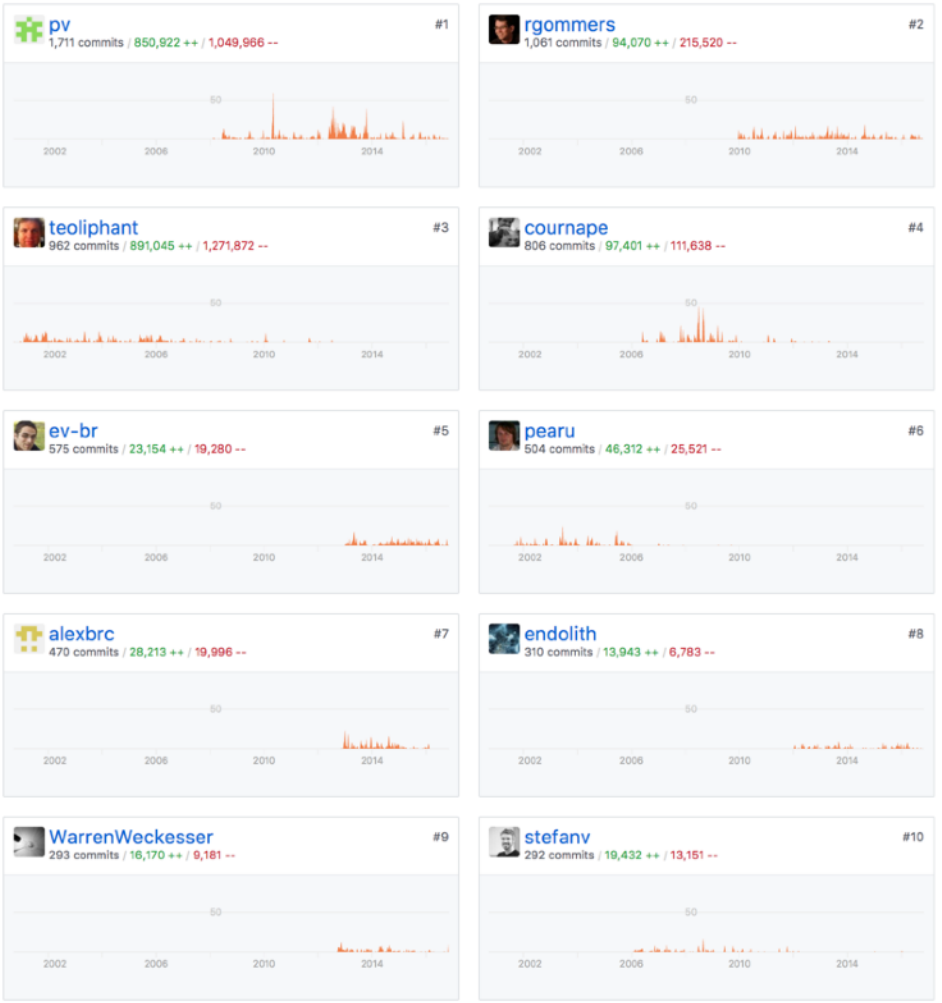


Where I started

Started as my graduate student "procrastination project" (as Multipack) in 1998 and became SciPy in 2001 with the help of Eric Jones and Pearu Peterson.

Used by: 128,495

108 releases, 766 contributors



SciPy

"Distribution of Python Numerical Tools masquerading as one Library"

Name	Description
cluster	KMeans and Vector Quantization
fftpack	Discrete Fourier Transform
integrate	Numerical Integration
interpolate	Interpolation routines
io	Data Input and Output
linalg	Fast Linear algebra
misc	Utilities
ndimage	N-dimensional Image processing

Name	Description
odr	Orthogonal Distance Regression
optimize	Constrained and Unconstrained Optimization
signal	Signal Processing Tools
sparse	Sparse Matrices and Algebra
spatial	Spatial Data Structures and Algorithms
special	Special functions (e.g. Bessel)
stats	Statistical Functions and Distributions

SciPy 1.0 — finally reached 1.0 in late 2017

- 2001: the first SciPy release
- 2005: transition to NumPy
- 2007: creation of scikits
- 2008: scipy.spatial module and first Cython code added
- 2010: moving to a 6-monthly release cycle
- 2011: SciPy development moves to GitHub

My active involvement

- 2011: Python 3 support
- 2012: adding a sparse graph module and unified optimization interface
- 2012: removal of scipy.maxentropy
- 2013: continuous integration with TravisCI
- 2015: adding Cython interface for BLAS/LAPACK and a benchmark suite
- 2017: adding a unified C API with scipy.LowLevelCallable; removal of scipy.weave
- 2017: SciPy 1.0 release

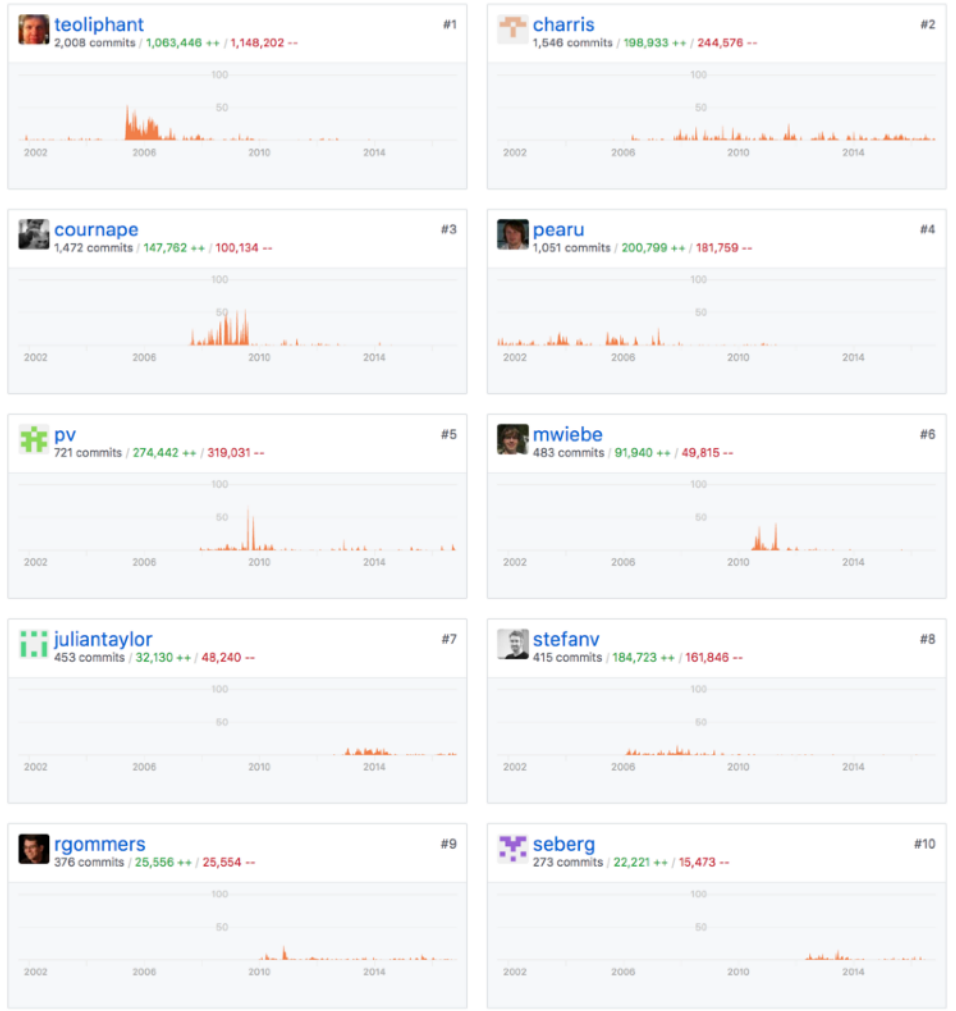
Others led

Where it led for me

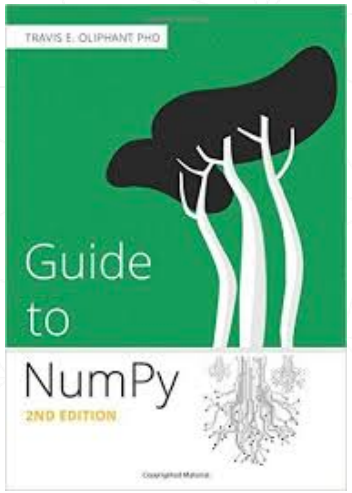
Gave up my chance at a tenured academic position in 2005-2006 to bring together the diverging array community in Python and unify Numeric and Numarray.

Used by: 254,856

159 releases, 827 contributors



NumPy: an Array Extension of Python

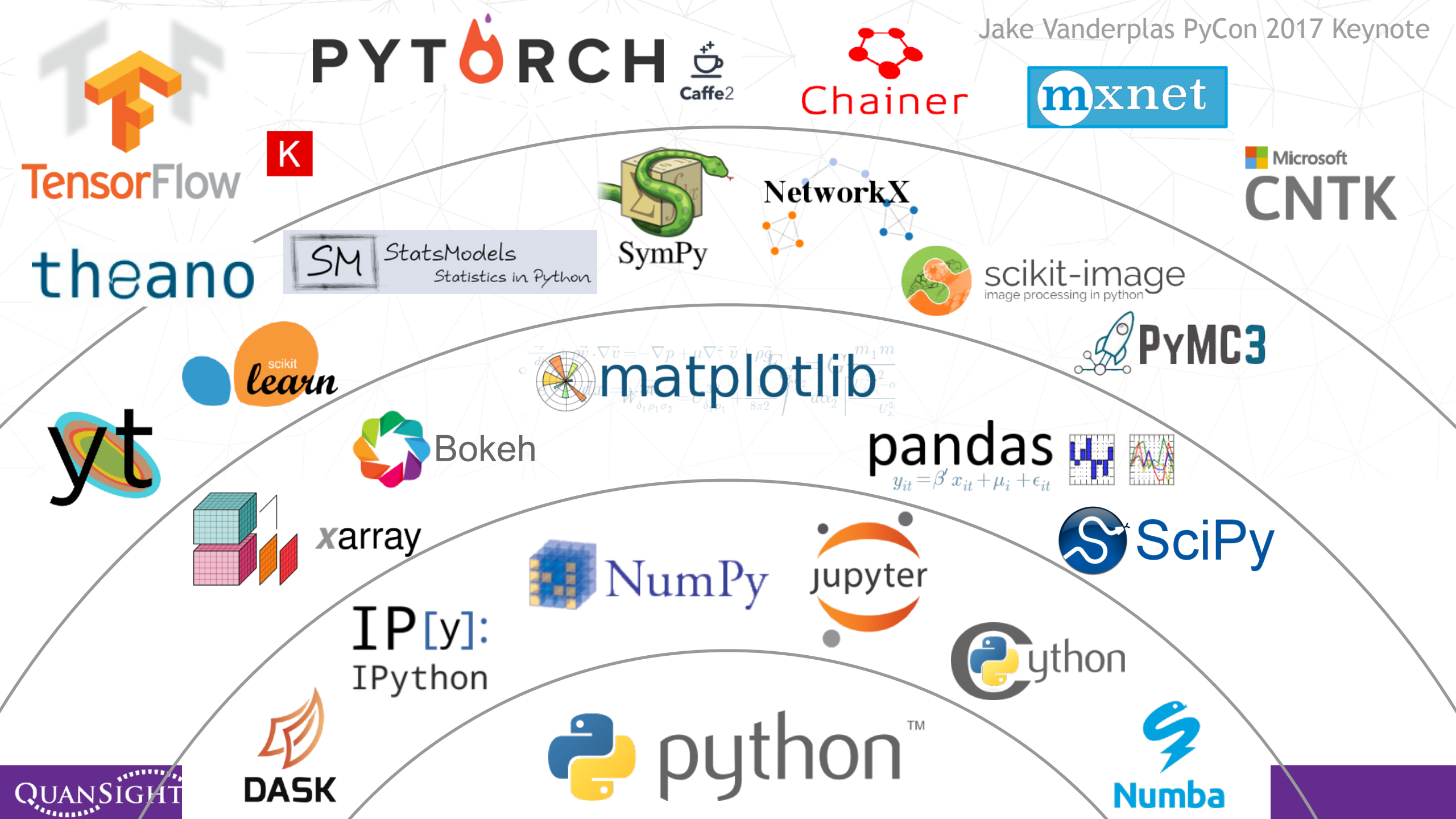


- Data: the array object
  - slicing and shaping
  - data-type map to bytes

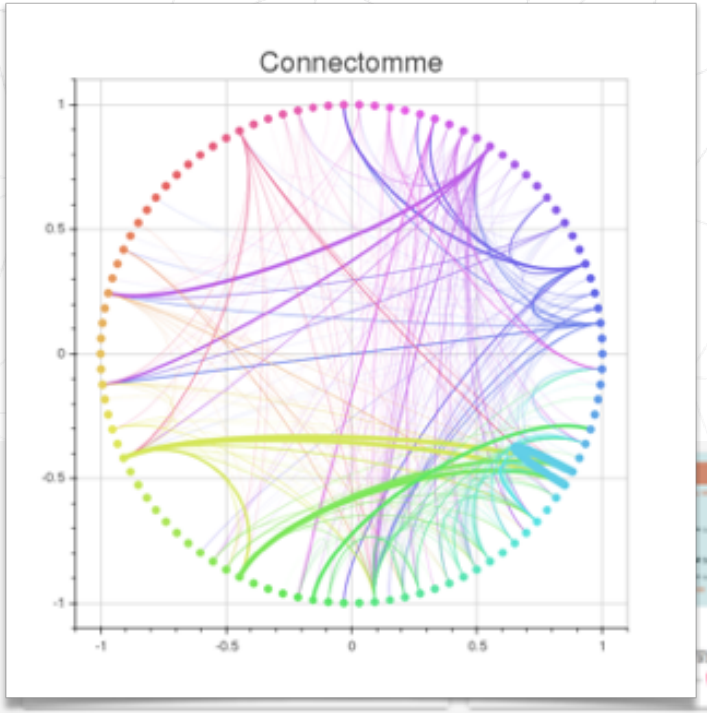
- Fast Math (ufuncs):
  - vectorization
  - broadcasting
  - aggregations

https://pydata.org/event-schedule/

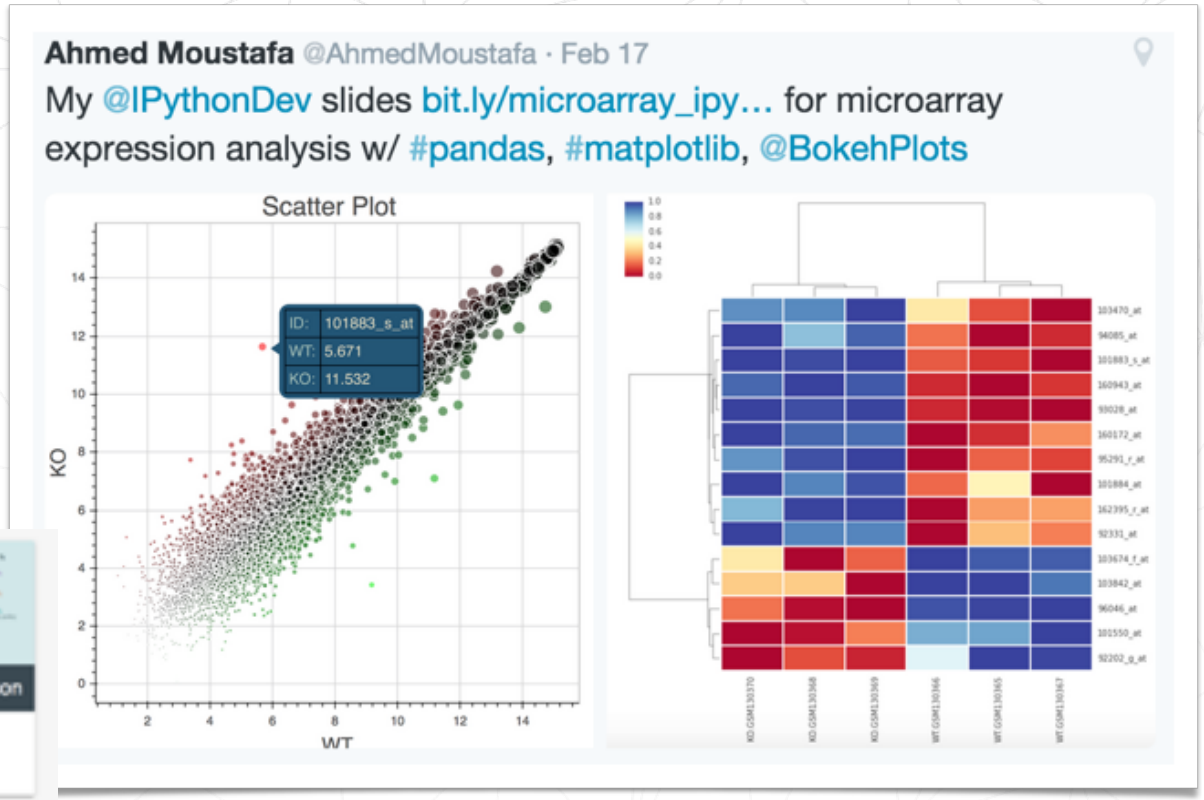
Jake Vanderplas PyCon 2017 Keynote



Bokeh — Data Visualization

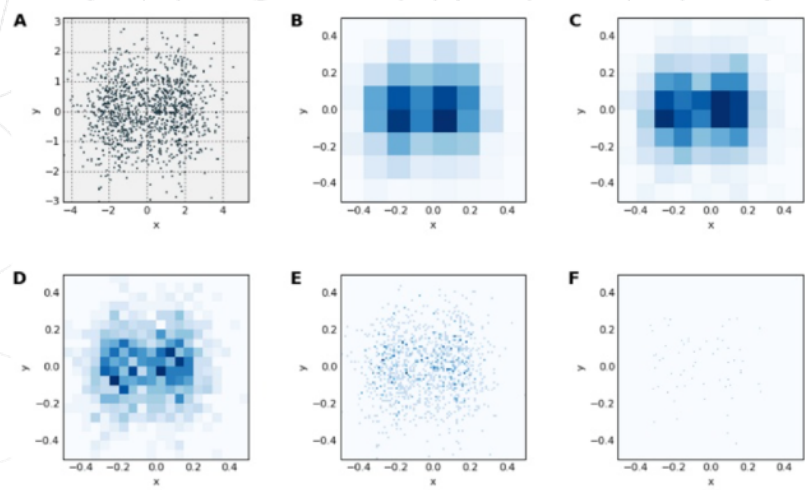


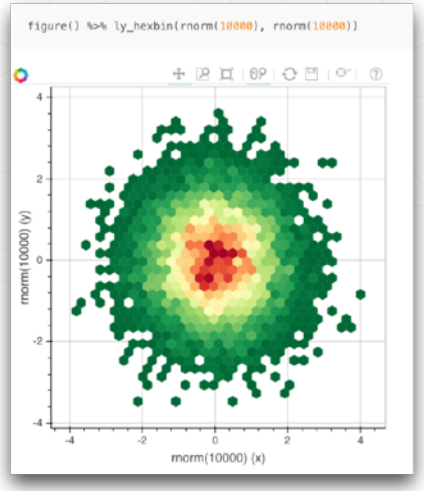
This site is a project of MIT's China Energy and Climate Project. The CECP is an alliance between the MIT Joint Program on the Science and Policy of Global Change and the Institute for Energy, Environment and Economy at Tsinghua University in Beijing, China. At MIT, the CECP is associated with and supported by the MIT Energy Initiative.

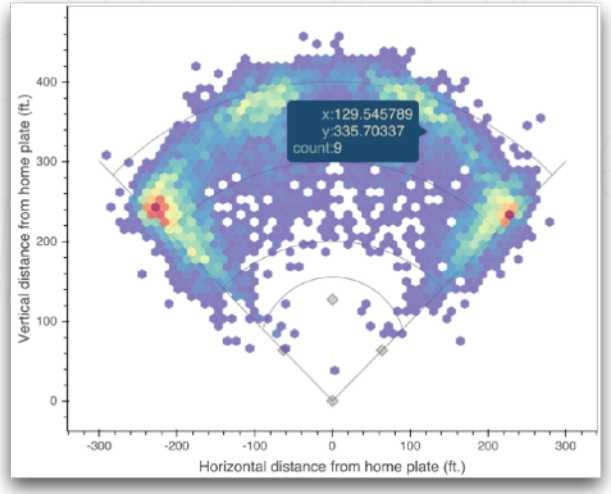


- Interactive viz, widgets, and tools
- Versatile high level graphics
- Streaming, dynamic, large data
- Optimized for the browser
- No Javascript
- With or without a server

Bokeh — Data Visualization

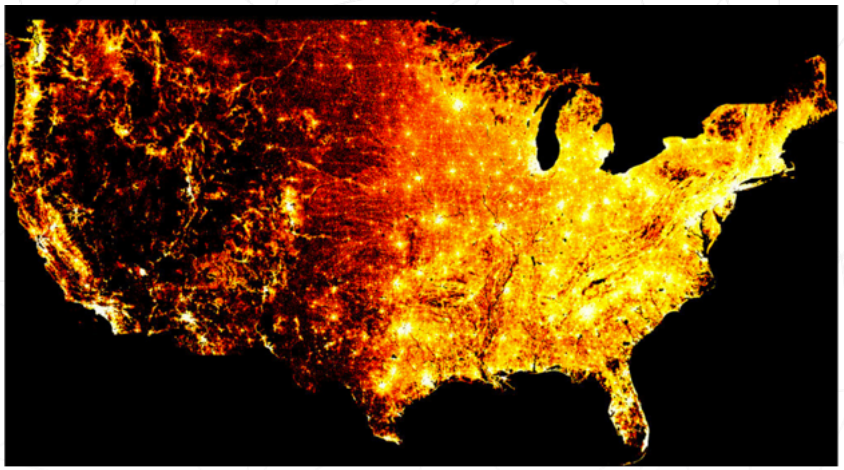




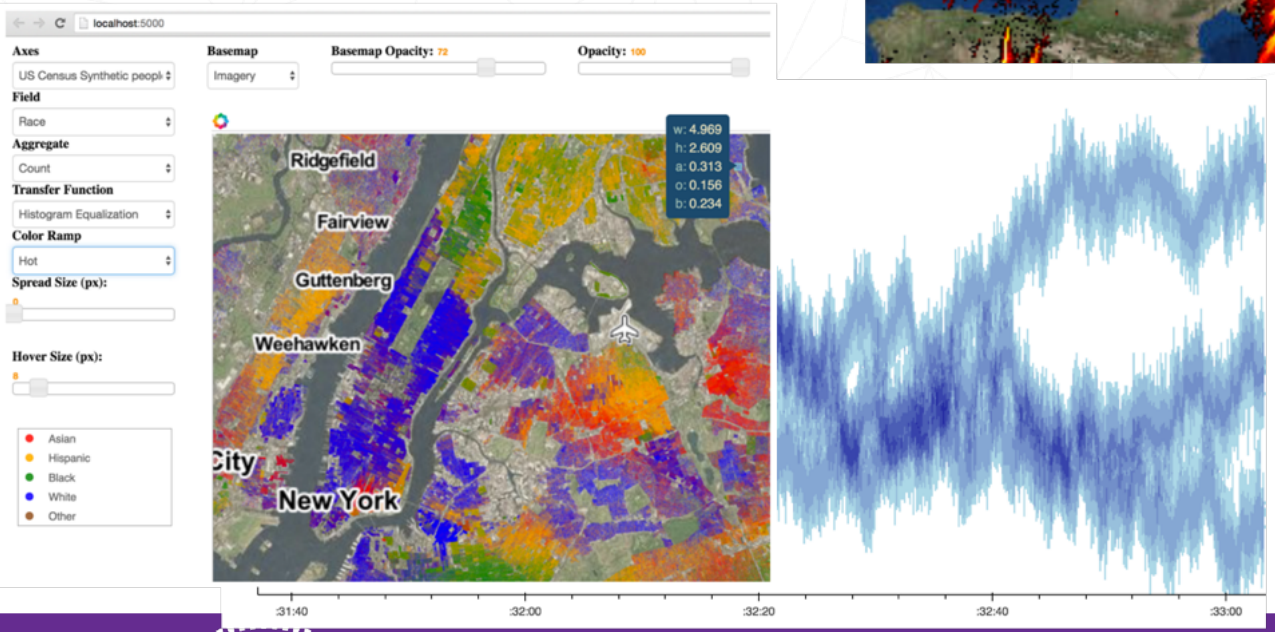


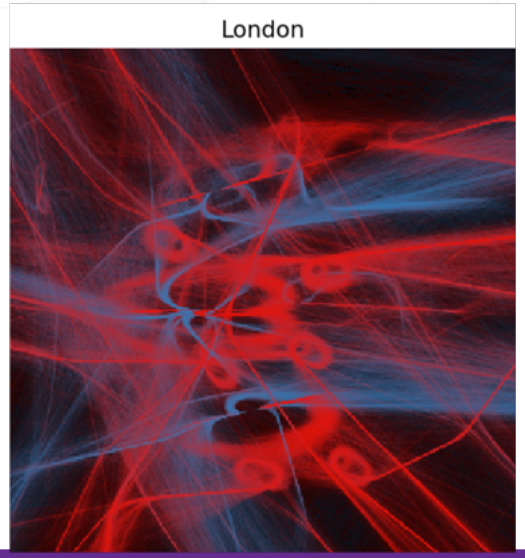
- Python interface
- R interface
- Smart plotting

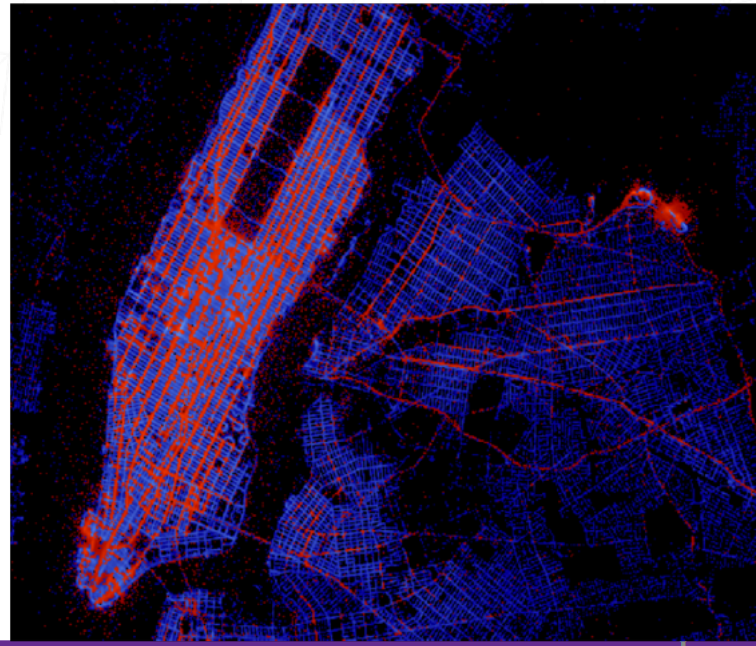
Datashader: Rendering a Billion Points of Data



- `datashader` provides a fast, configurable visualization pipeline for faithfully revealing even very large datasets
- Each of these visualizations requires just a few lines of code and no magic numbers to adjust by trial and error.





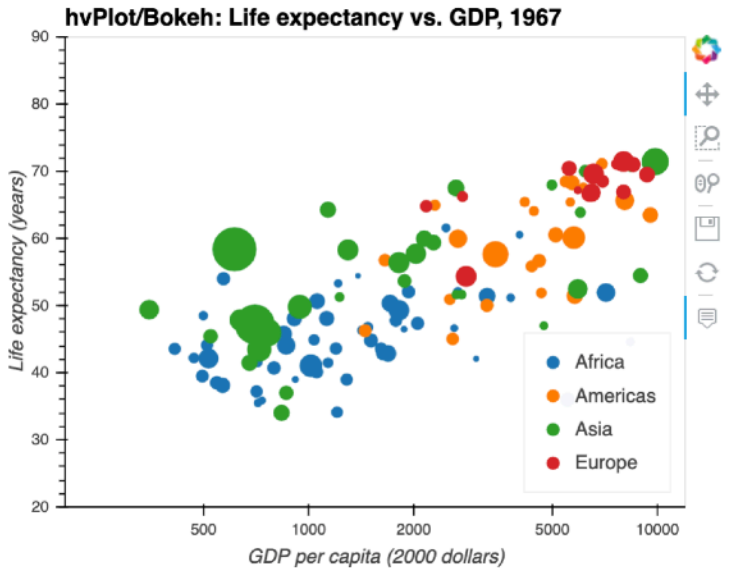


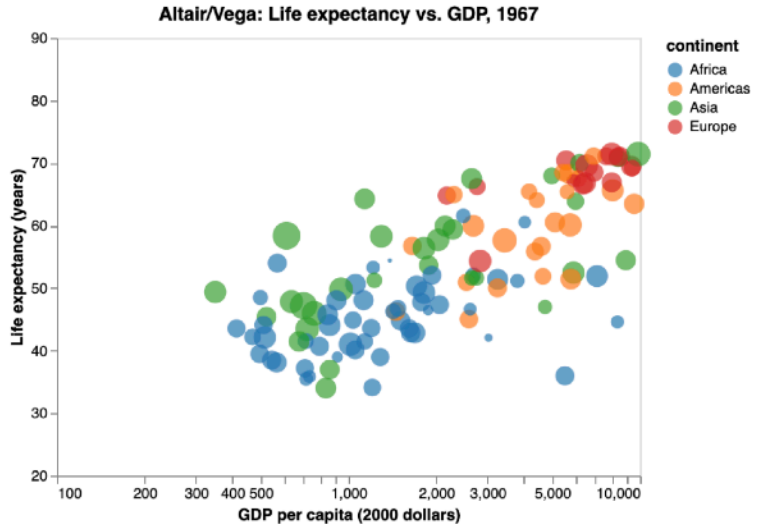
Plotting library comparison

The Panel library from PyViz lets you make widget-controlled apps and dashboards from a wide variety of plotting libraries and data types. Here you can try out five different plotting libraries controlled by a couple of widgets, for Hans Rosling's gapminder example.

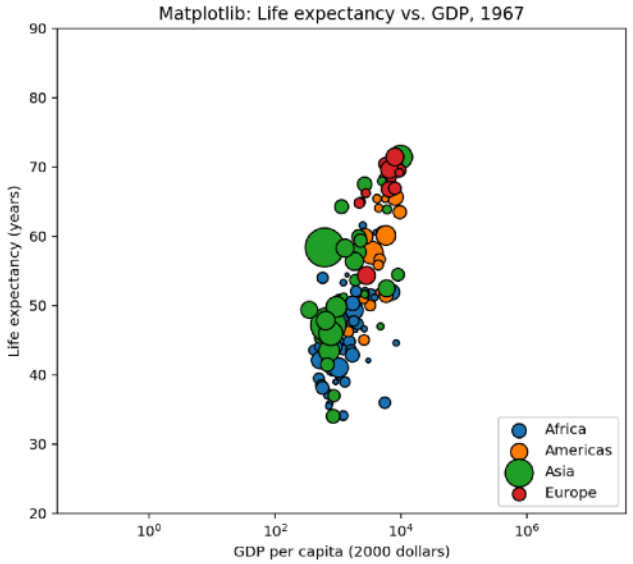
Year: **1967**

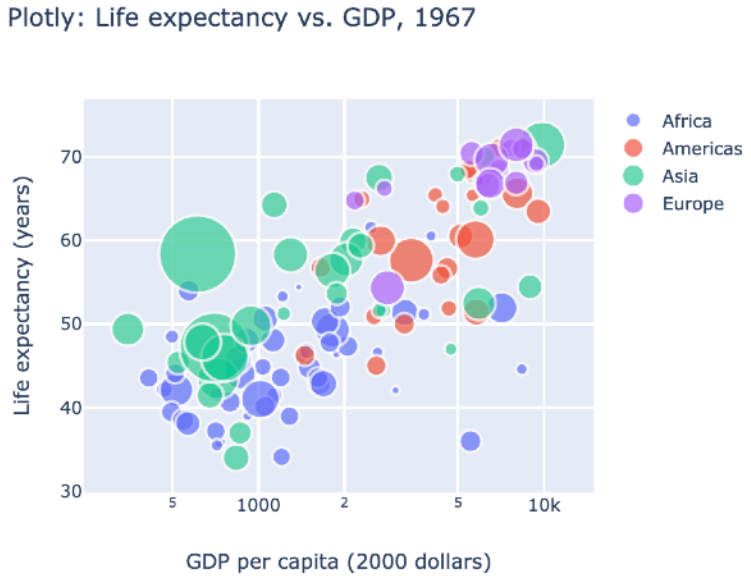
Show legend





Easy Dashboards!





Exploring Data in Jupyter Notebooks ——> easy Web Applications!

NYC Taxi Trips

Alpha: **0.70**

Cmap

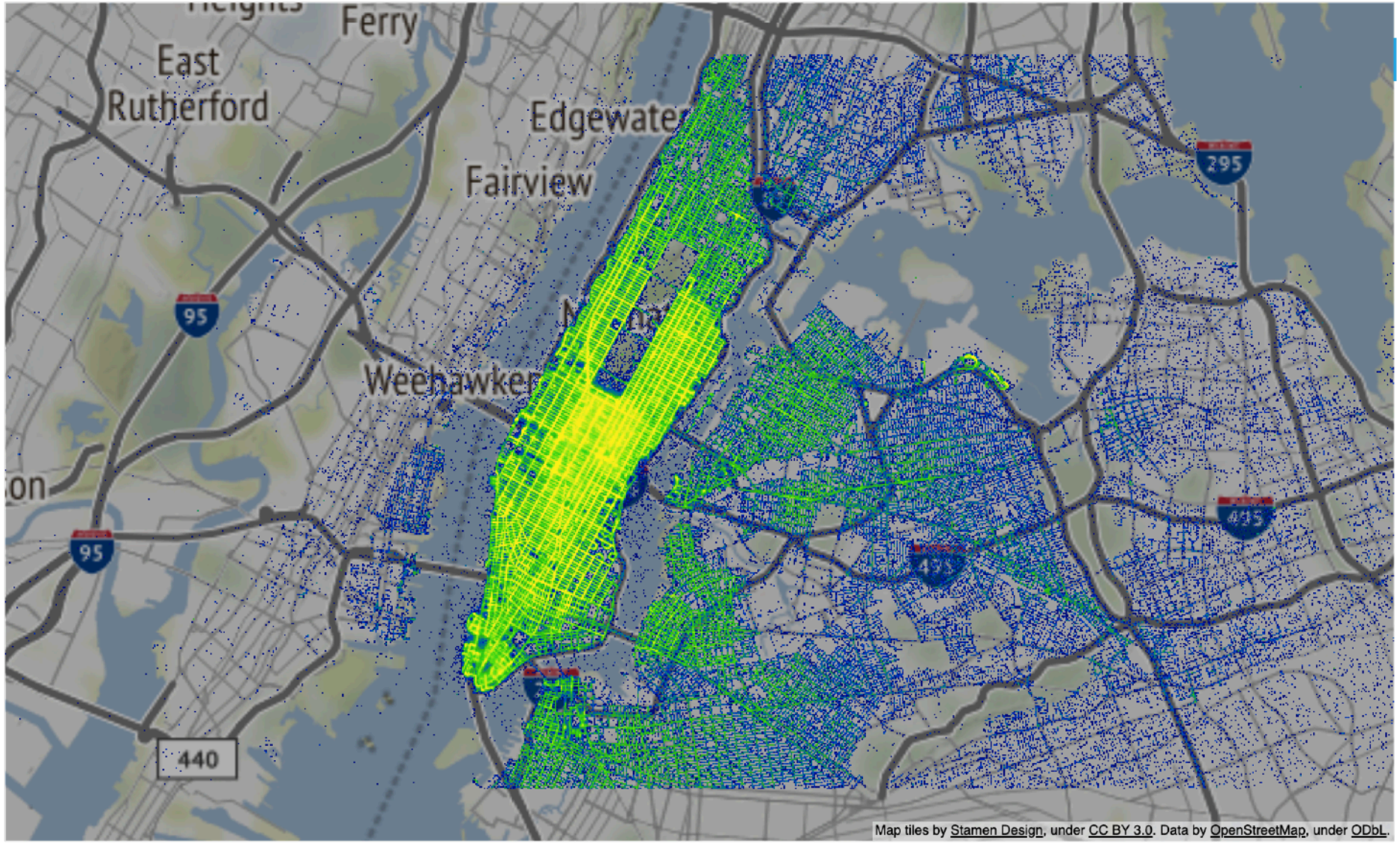
bgy

Hour: **0 .. 24**

Location

dropoff

http://panel.pyviz.org



Panel

an open source JIT compiler that translates a subset of Python and NumPy code into fast machine code.

http://numba.pydata.org

Accelerate Python Functions

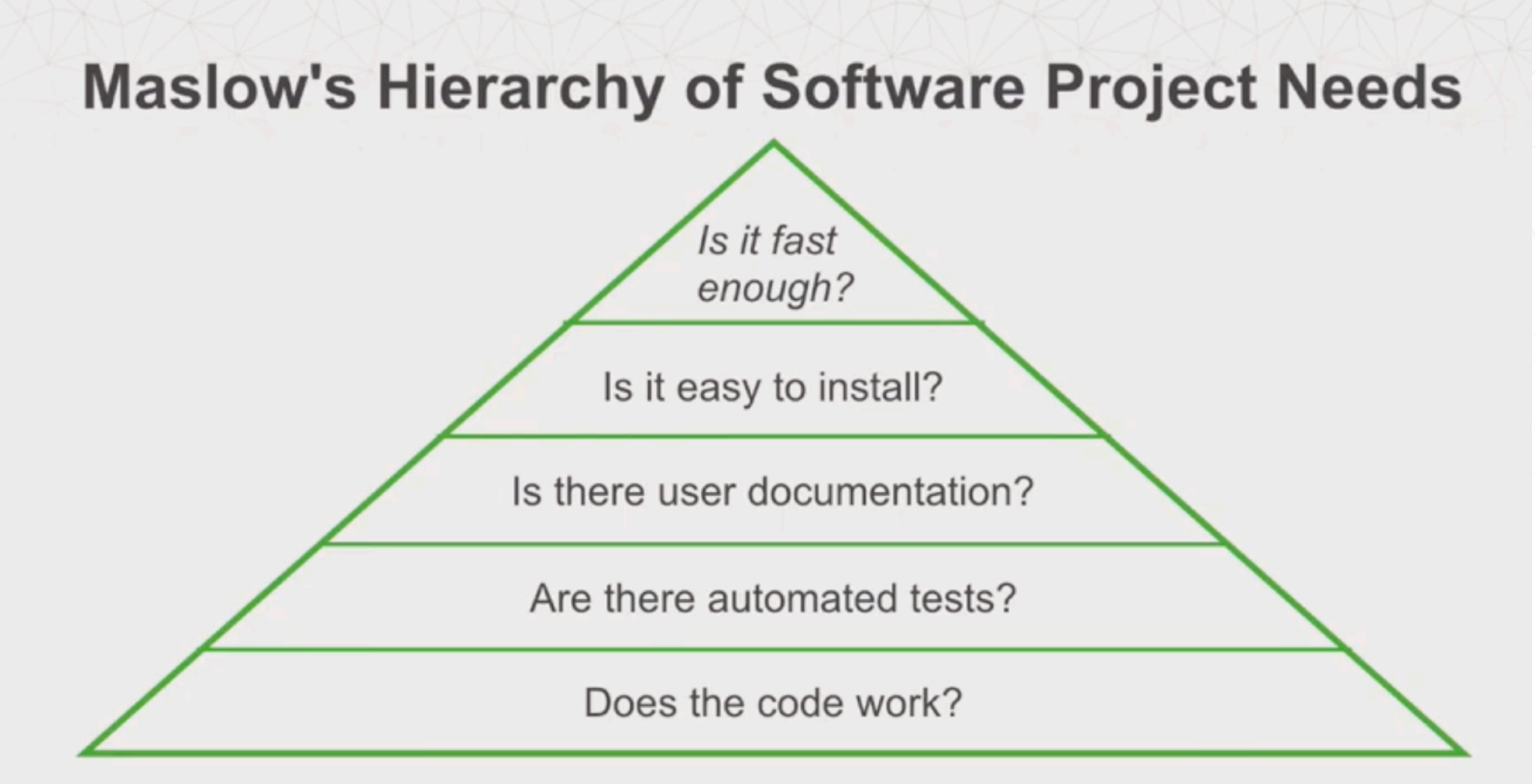
Numba translates Python functions to optimized machine code at runtime using the industry-standard LLVM compiler library. Numba-compiled numerical algorithms in Python can approach the speeds of C or FORTRAN.

You don't need to replace the Python interpreter, run a separate compilation step, or even have a C/C++ compiler installed. Just apply one of the Numba decorators to your Python function, and Numba does the rest.

```
from numba import jit
import random

@jit(nopython=True)
def monte_carlo_pi(nsamples):
    acc = 0
    for i in range(nsamples):
        x = random.random()
        y = random.random()
        if (x ** 2 + y ** 2) < 1.0:
            acc += 1
    return 4.0 * acc / nsamples
```

Before you try to optimize speed!



Stan Seibert, “How to Accelerate an Existing Codebase with Numba” SciPy 2019

DASK

Dask natively scales Python.

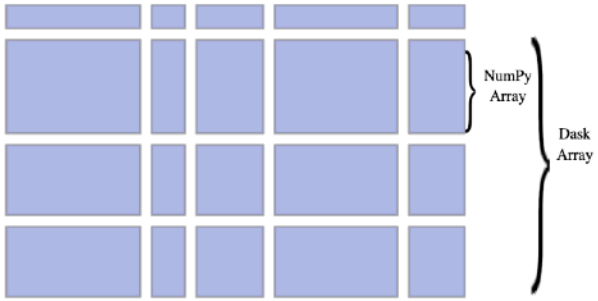
Dask provides advanced parallelism for analytics, enabling performance at scale for the tools you love

Integrates with existing projects

BUILT WITH THE BROADER COMMUNITY

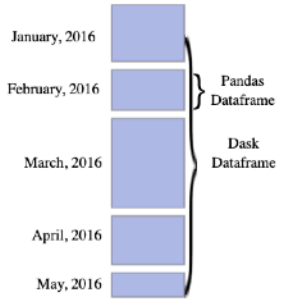
Dask is open source and freely available. It is developed in coordination with other community projects like Numpy, Pandas, and Scikit-Learn.

Numpy



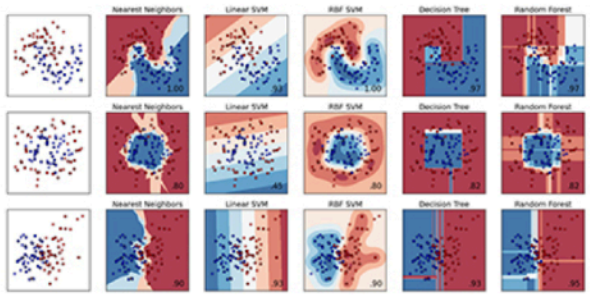
Dask arrays scale Numpy workflows, enabling multi-dimensional data analysis in earth science, satellite imagery, genomics, biomedical applications, and machine learning algorithms.

Pandas

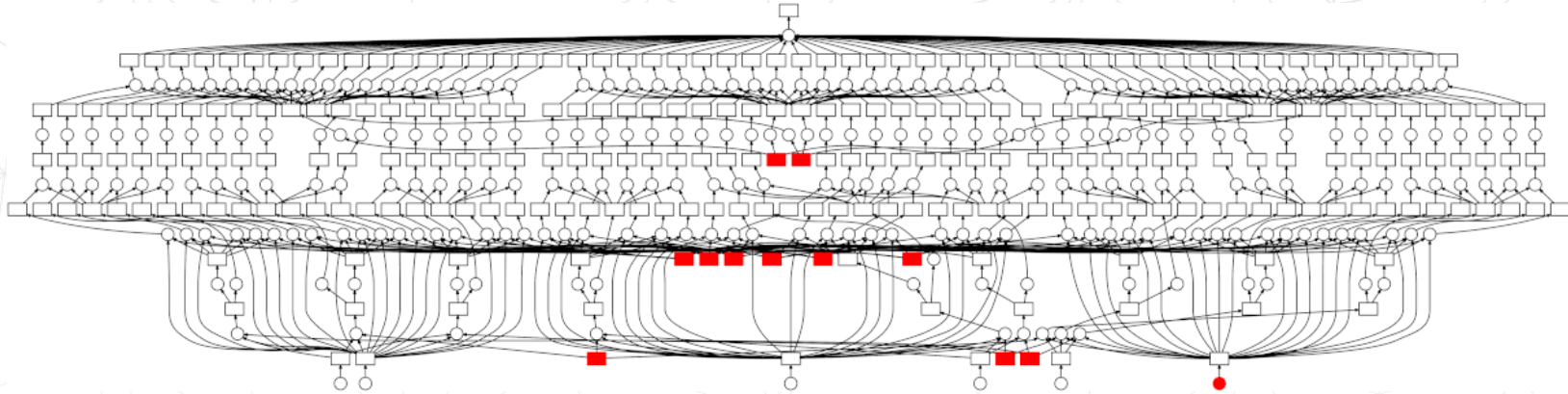


Dask dataframes scale Pandas workflows, enabling applications in time series, business intelligence, and general data munging on big data.

Scikit-Learn



Dask-ML scales machine learning APIs like Scikit-Learn and XGBoost to enable scalable training and prediction on large models and large datasets.



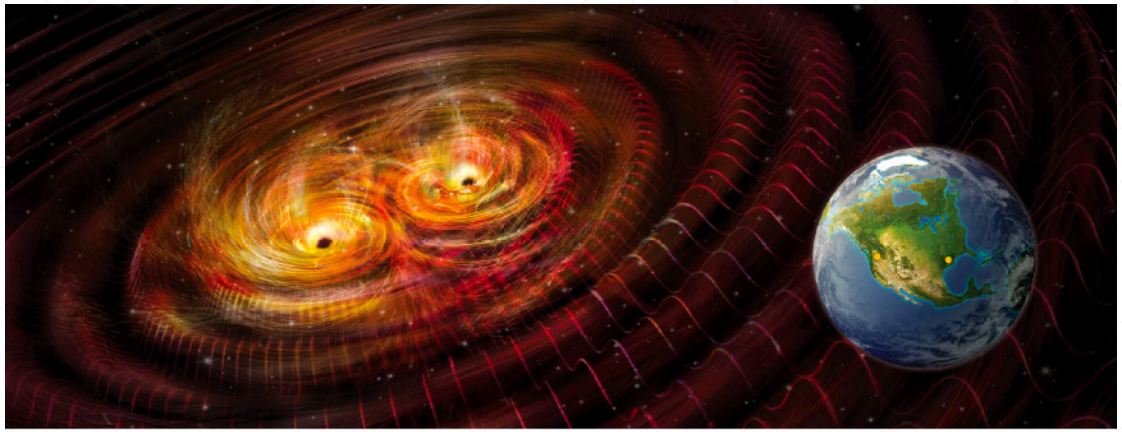
Scale up to clusters

OR JUST USE IT ON YOUR LAPTOP

Dask's schedulers scale to thousand-node clusters and its algorithms have been tested on some of the largest supercomputers in the world.

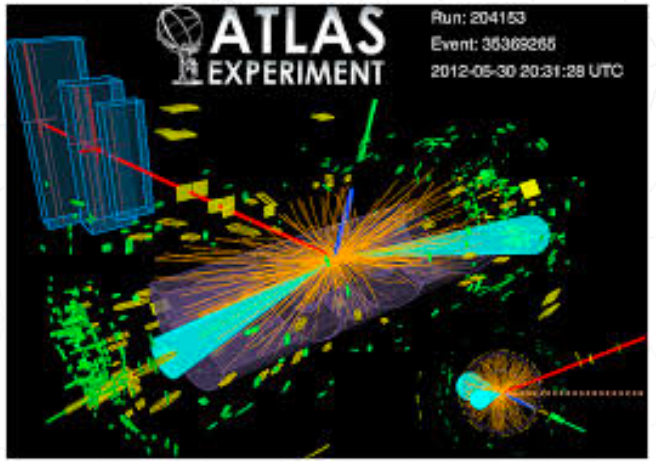
But you don't need a massive cluster to get started. Dask ships with schedulers designed for use on personal machines. Many people use Dask today to scale computations on their laptop, using multiple cores for computation and their disk for excess storage.

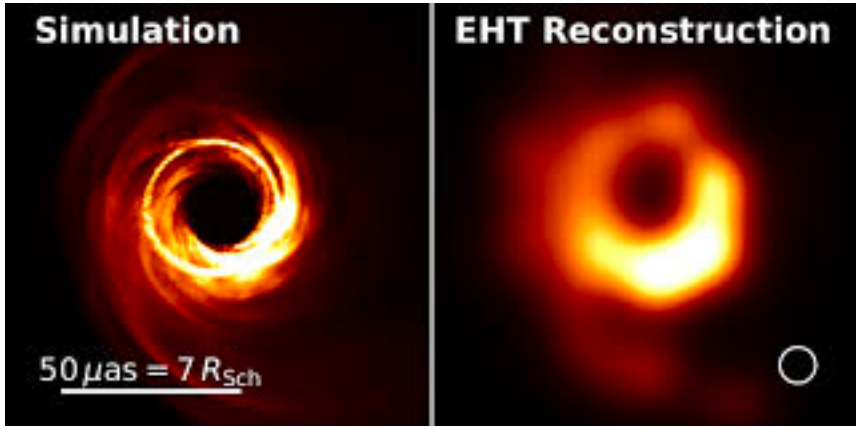
Huge Impact (from diverse efforts of 1000s)



LIGO : Gravitional Waves

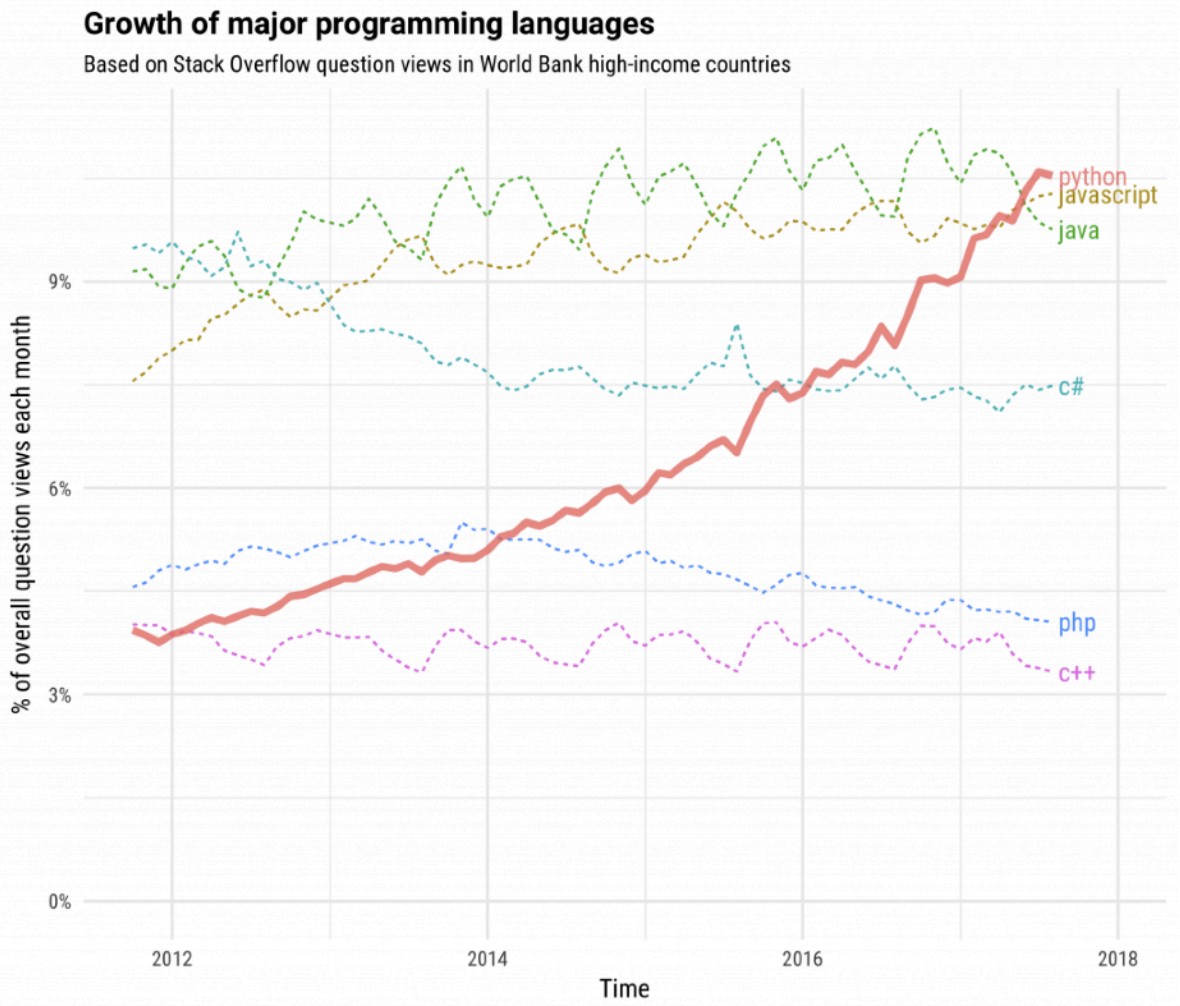
Higgs Boson Discovery

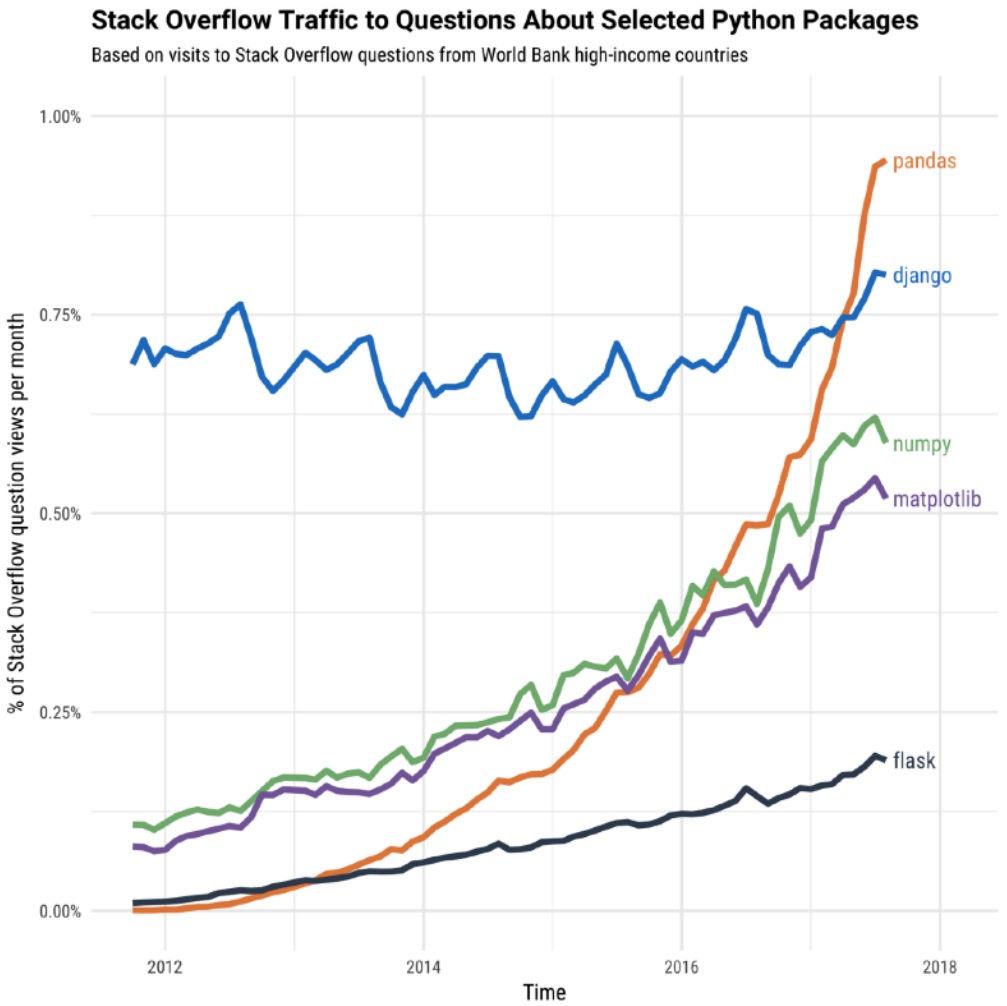




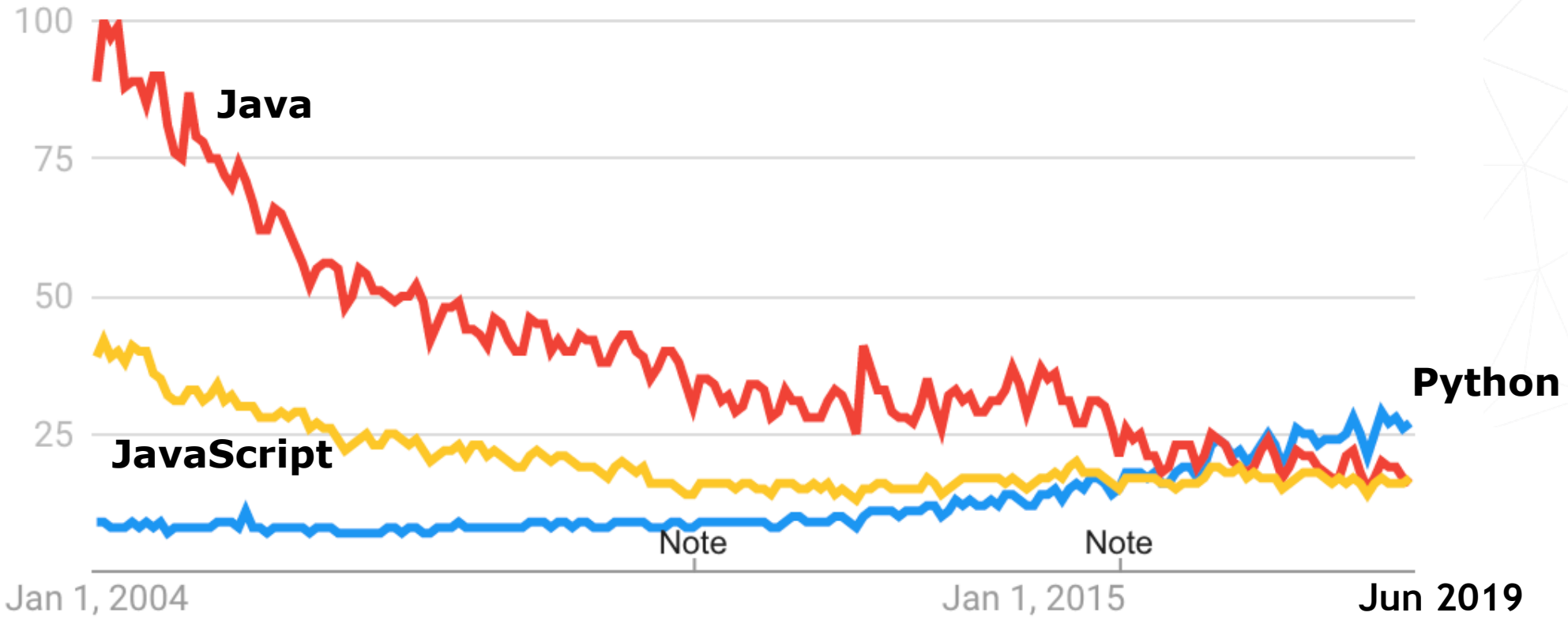
Black Hole Imaging

Python and in particular PyData keeps Growing

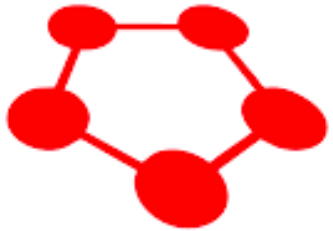




Google Search Trends



Several Deep Learning Libraries to choose



Chainer

Built on NumPy/SciPy

Recommended

TensorFlow

Recommended

Key Features Needed for any ML Library

For Training

- Ability to create chains of functions on n-dimensional arrays
- Ability to derive the derivative of the Loss-Function quickly (Automatic Differentiation)
- Key Loss Functions implemented
- Cross-validation methods
- An Optimization library with several useful methods
- Ability to compute functions on n-dimensional arrays on multiple hardware with highly parallel-execution

Missing from NumPy / SciPy and Scikit-Learn, but added by CuPy and Autograd

For Inference

- Ability to create chains of functions on n-dimensional arrays
- Ability to compute functions on n-dimensional arrays on multiple hardware

Most Libraries (other than Chainer) chose to re-implement NumPy and SciPy as they needed.

Reasons:

- Started with a legacy code in another language
- Had to work with other languages too (Node, Java, C++, Lua, *etc*.)
- Needed only a subset of functionality of NumPy / SciPy to build ML
- Needed GPU support
- Lacked familiarity with the NumPy / SciPy communities and how to engage with them

Working on this!

Result: Many competing similar choices for Deep Learning

SciPy Conference Early History

- First SciPy Conference held at CalTech in Summer of 2001 (~50 attendees)
- The first nine conferences were held in Pasadena, CA until the US conference site moved to Austin in 2010
- EuroSciPy started in 2008 in Leipzig
- SciPy.India started in 2009 in Trivandrum
- SciPy Latin America started in 2013
- SciPy Japan started in 2019

SciPy India 2009

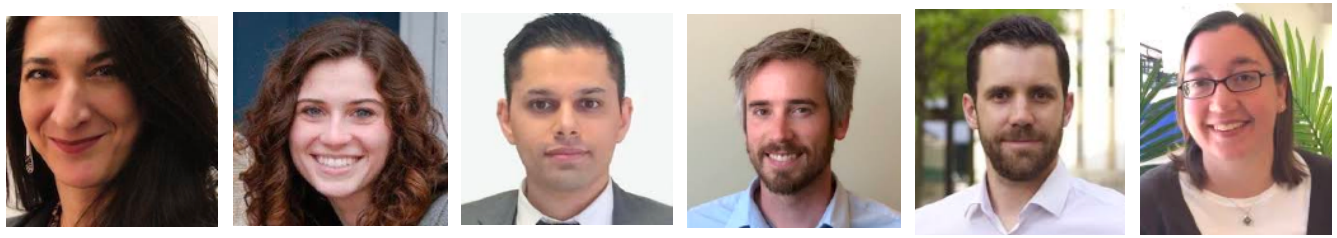
Non-profits an important part of fabric

NUMFOCUS

OPEN CODE = BETTER SCIENCE

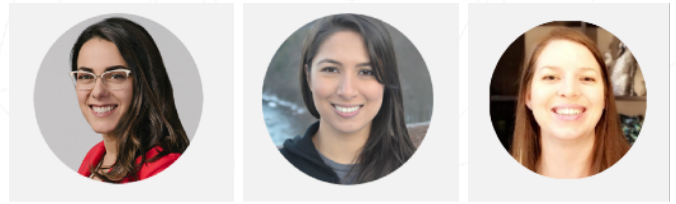
PyData

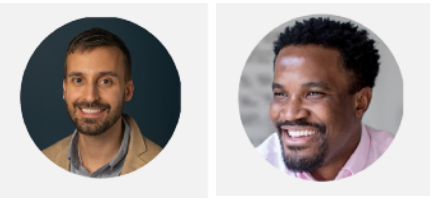
We founded in 2012 to unite communities from Jupyter, Numarray, Matplotlib, and the scikits together.

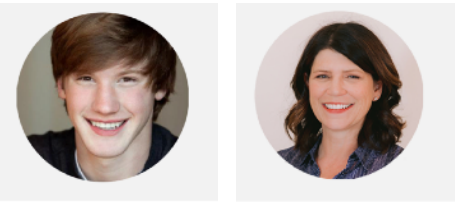


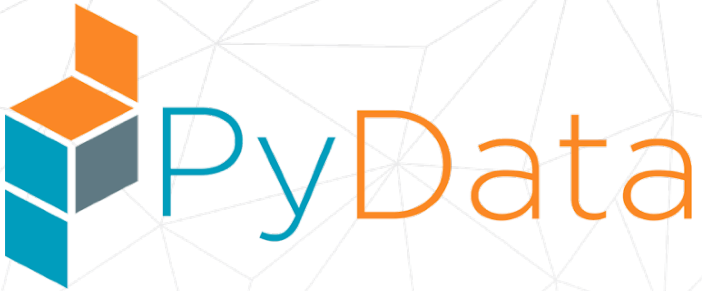
Board Members selected for 2018

2019 Staff









https://pydata.org/past-events/

https://pydata.org/event-schedule/

- 2012: Workshop at Google
- 2012: New York
- 2013: Silicon Valley, Boston, New York
- 2014: London, Berlin, Silicon Valley, New York
- 2015: Paris, Dallas, Berlin, London, Seattle, New York
- 2016: Amsterdam, Madrid, Florence, London, Berlin, Paris, Spain, San Francisco, Chicago, Carolinas
- 2017: Florence, Amsterdam, London, Barcelona, Luxembourg, Paris, Virginia, Berlin, Seattle, Rimini (Italy), Ossa (Poland), New York, New Delhi, Warsaw, Berlin, London, Budapest, San Luis, New York
- 2018: Oxford, Florence, London, New York, Lithuania, Seattle, Amsterdam, Berlin, Edinburgh, London, Delhi, New York, Cordoba, South Africa, New York, Los Angeles, Karlsruhe, Washington DC, Warsaw,
- 2019: Miami, Bogota, Florence, Amsterdam, Lithuania, Czech Republic, Basel, London, New Delhi, Cordoba, Berlin, New York, Budapest, Cambridge, Eindhoven, Los Angeles, Austin, Warsaw

A big part of my life is family!

My wife, Amy, and I have six children who are looking to us for support (emotional, logistical, financial, etc.)

I love participating in OSS but I **have** to do it sustainably and I am driven to help others participate sustainably as well.

Decided to start companies to sustain OSS

CONTINUUM ANALYTICS

renamed

ANACONDA®

Peter Wang

~17 million Anaconda users

Building new solutions

Key members of the original management team at Continuum created Quansight. We view NumFOCUS and Anaconda as our first (spin-out) organizations.

2012 CONTINUUM ANALYTICS

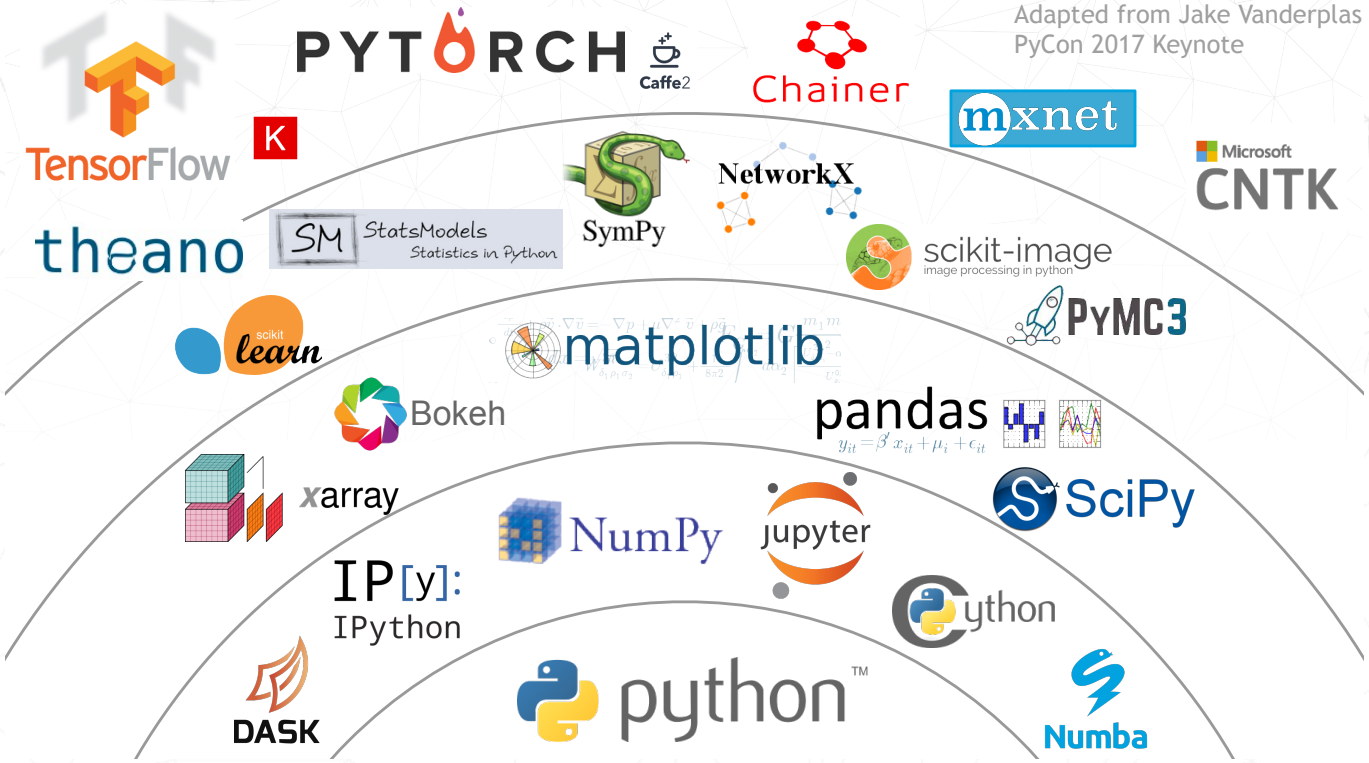
Replaced by → ANACONDA. 2015

2018 Quansight

Spin Out → OpenTeams

Spin Out → ?

QUANSIGHT LABS



Sustaining the Future

Open-source innovation and maintenance around the entire data-science and AI workflow.

- **NumPy ecosystem maintenance (PyData Core Team)**
- **Improve connection of NumPy to ML Frameworks**
- **GPU Support for NumPy Ecosystem**
- **Improve foundations of Array computing**
- **JupyterLab and JupyterHub**
- **Data Catalog standards**
- **Packaging (conda-forge, PyPA, etc.)**

Partnering with NumFOCUS and Ursa Labs (supporting Apache Arrow)

uarray — unified array interface for SciPy refactor
xnd — re-factored NumPy (low-level cross-language libraries for N-D (tensor) computing)

QUANSIGHT INITIATE

supporting

FairOSS

Bradden Blair

An early stage venture capital firm investing in startups that build on open-source technology and support the communities they depend on.

AUTO

VERTALO

SaveDay

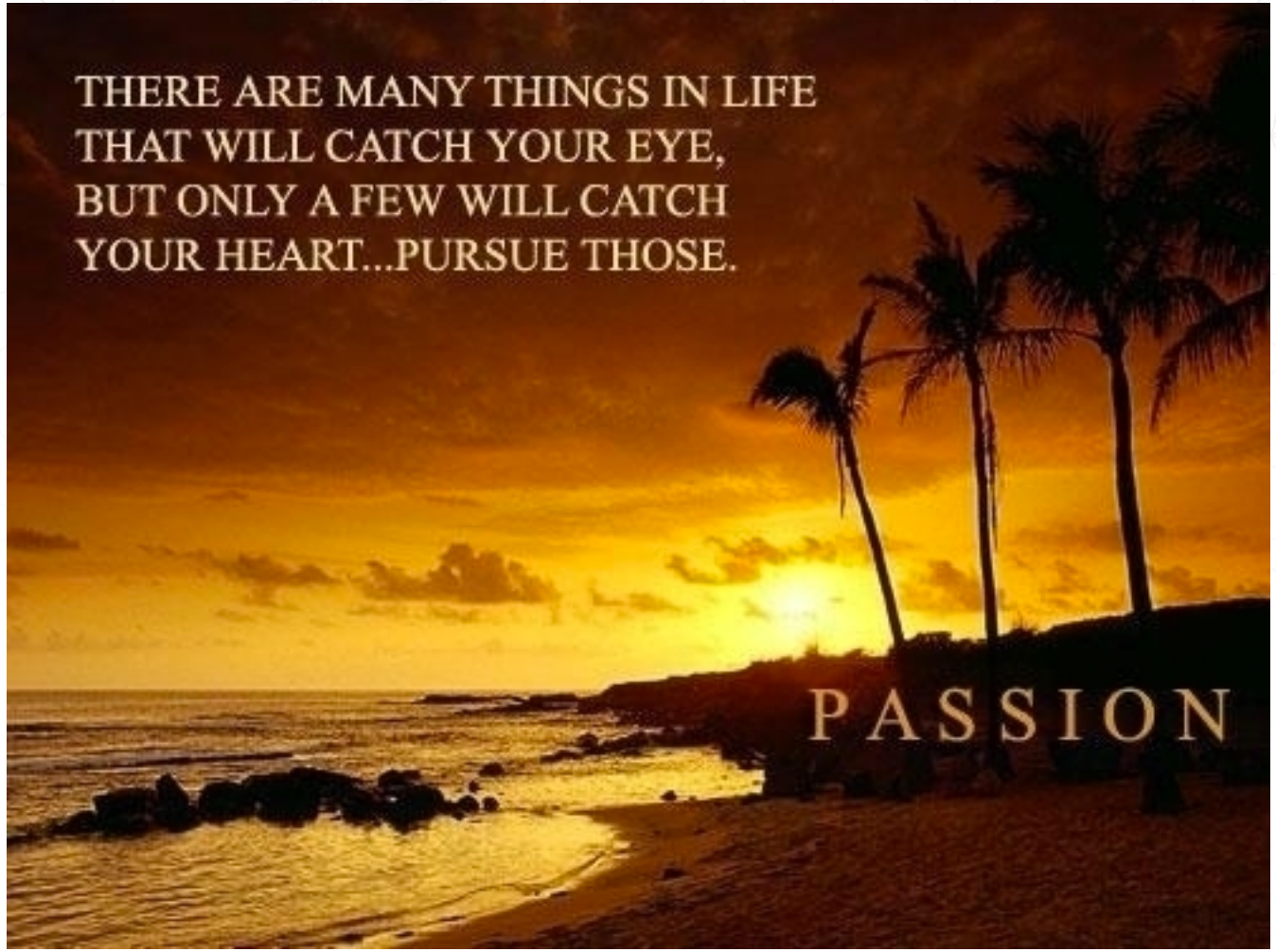
ANACONDA

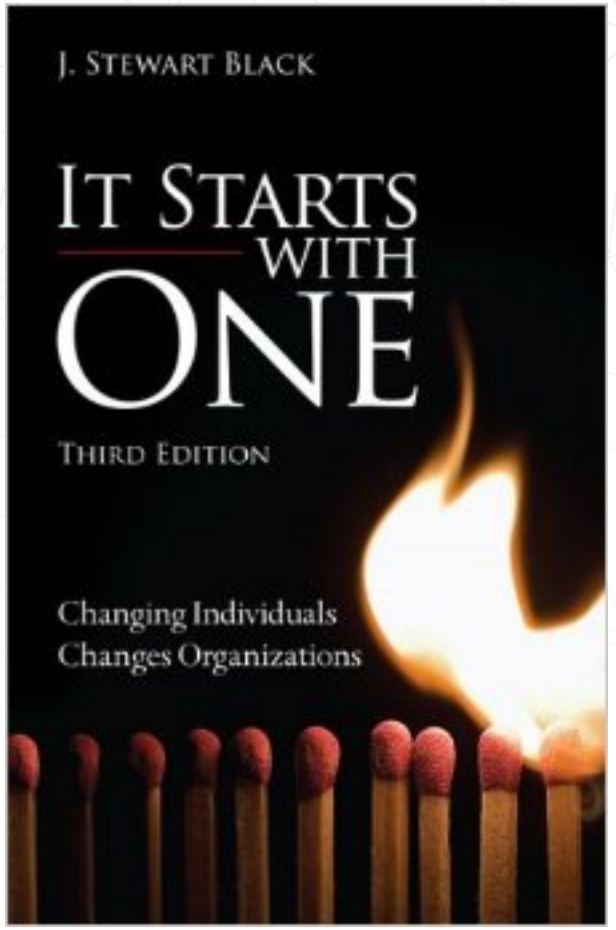
Tenavox

OpenTeams

How to get more Community!

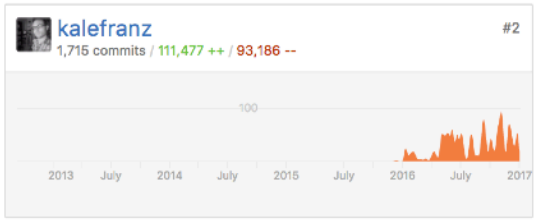
CULTIVATION OF COMMUNITY

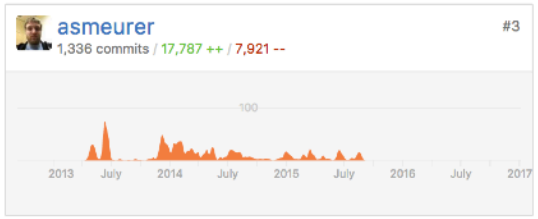


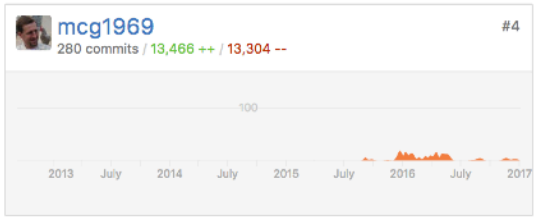


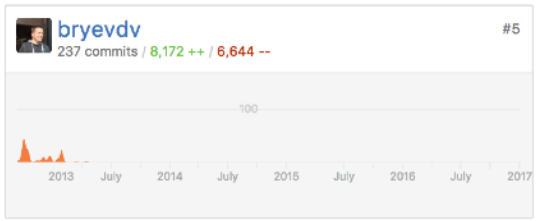
Great works are started by a small group usually 1-3 people).

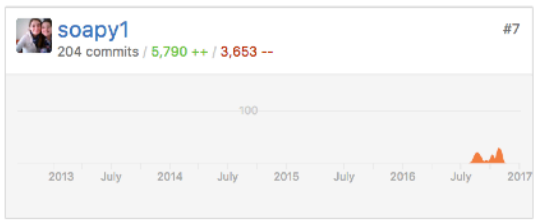
CONDA

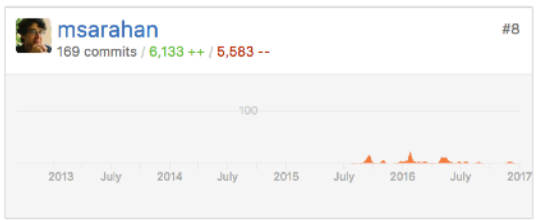




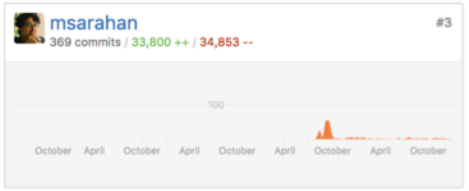


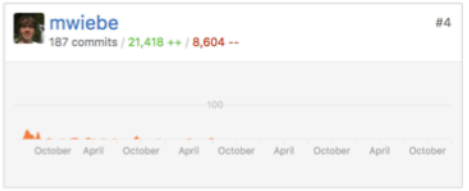


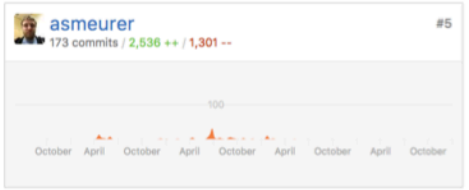


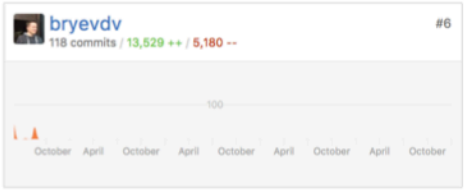


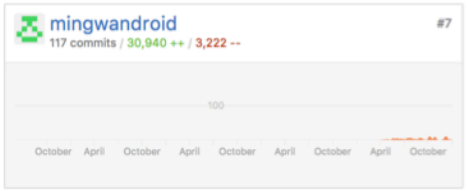
ANACONDA®

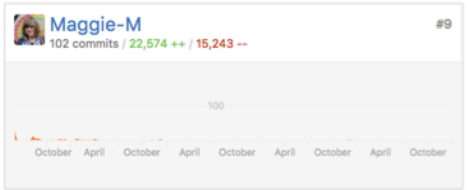












CULTIVATION OF COMMUNITY

Some of the Conda and Anaconda Team


Kale Franz


Michael Sarahan


Michael Grant

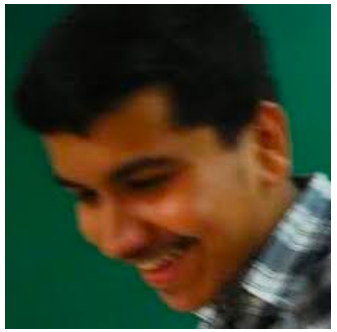

Bryan Van de Ven


Jonathan Helmus


Crystal Soja


Ilan Schnell


Ray Donnelly


Nehal Wani


Maggie Mari

CULTIVATION OF COMMUNITY

Over 1200 contributors!
Currently 8,093 pre-built packages!

There **are** better ways to do things. There is a **good**.

can unite theists, agnostics and atheists in common efforts.

What is the good : ???

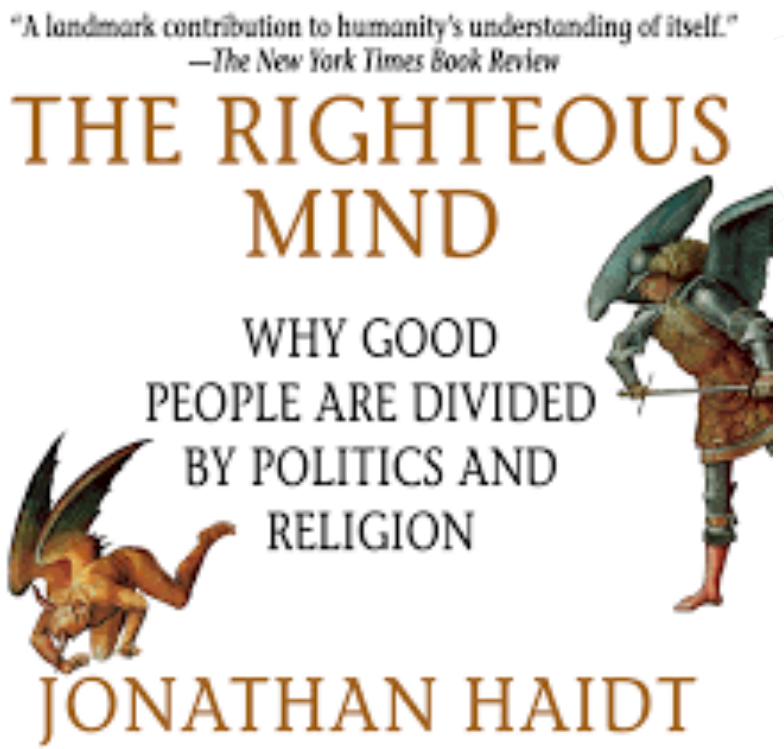
Possible working statement:

Ideas and actions that lead to lasting, self-consistent, and sustaining peace, happiness, self-direction, and prosperity — as decided by **all participants.**

You can only really think locally:

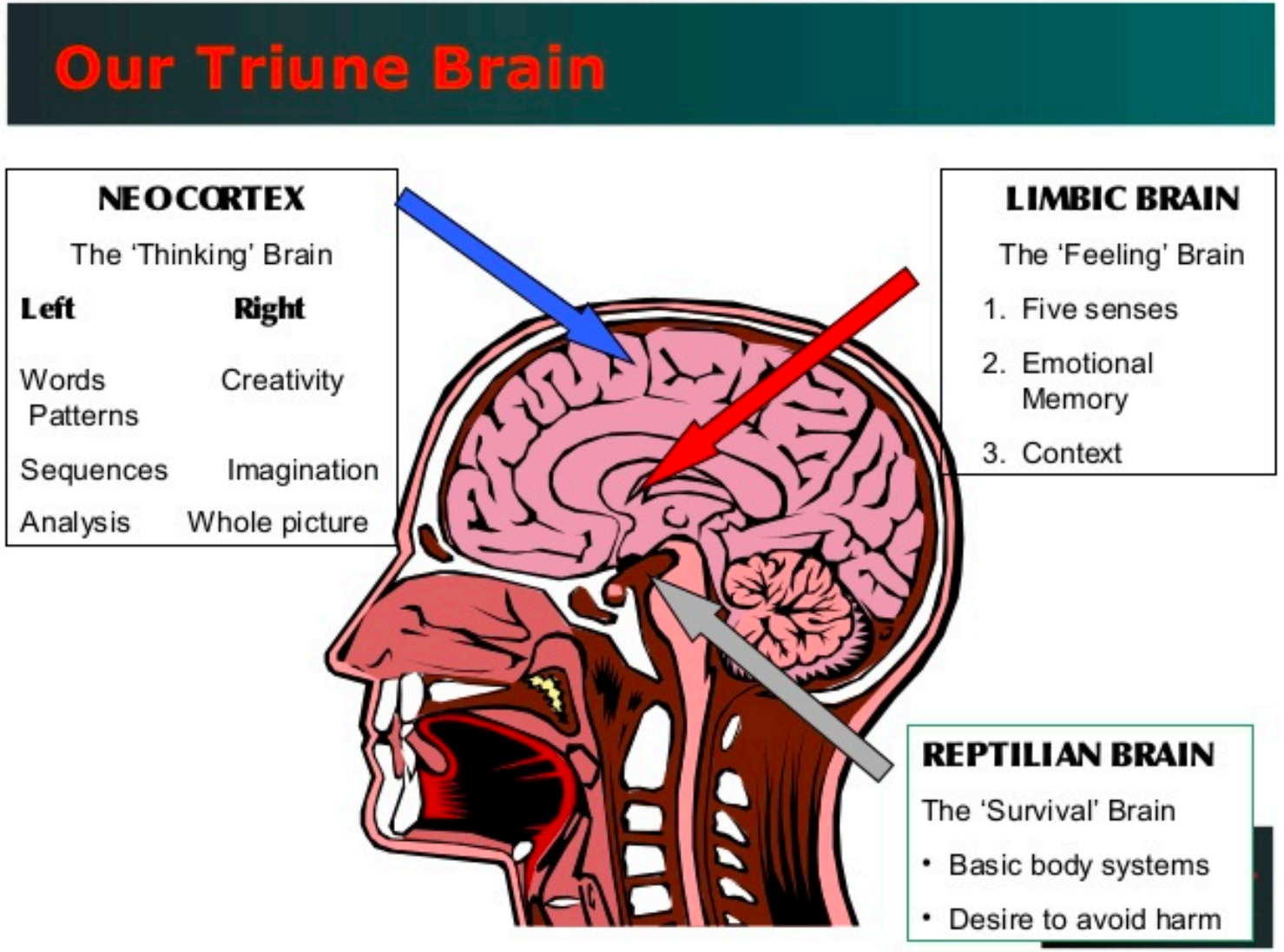
How do I serve the current and future users and developers of this project?

Excellent book:



Neuroscience Matters

- The brain is a complicated mix of many sub-systems and processes.
- We all share in a spectrum of “mental-illness” called the human experience.
- Learning to manage your particular experience (using all available resources: cognitive therapy, medication, improved eating, exercise, sleeping, etc.) is critical.

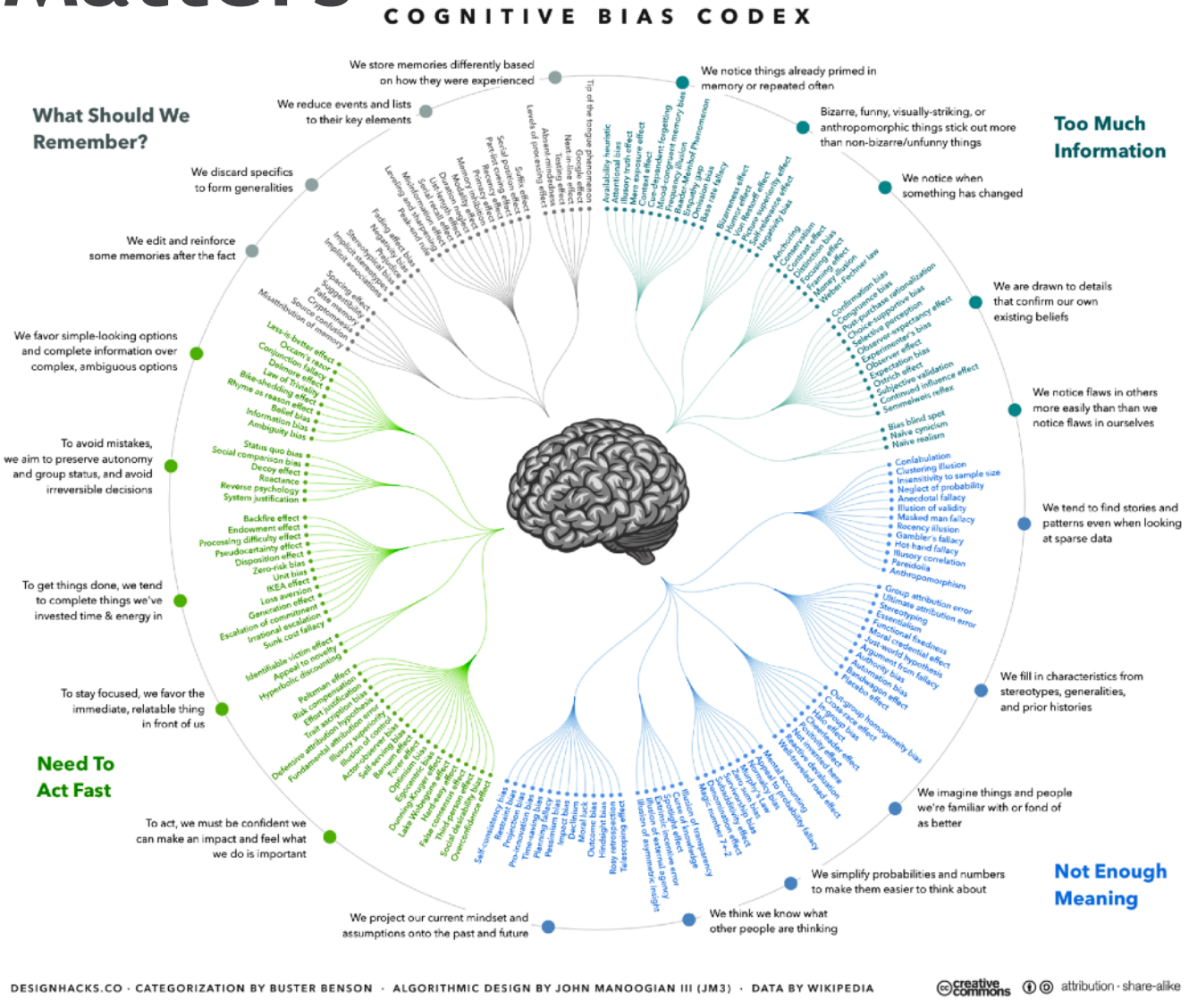


- File:Cognitive Bias Codex - 180+ biases, designed by John Manoogian III (jm3).jpg

Neuroscience Really Matters

Understanding the **cognitive biases** that emerged in our sub-systems from evolution helps us manage them.

Anchoring bias
Confirmation bias
Negativity bias
Bandwagon effect
Stereotypical bias
Blind Spot bias
Not invented here
…

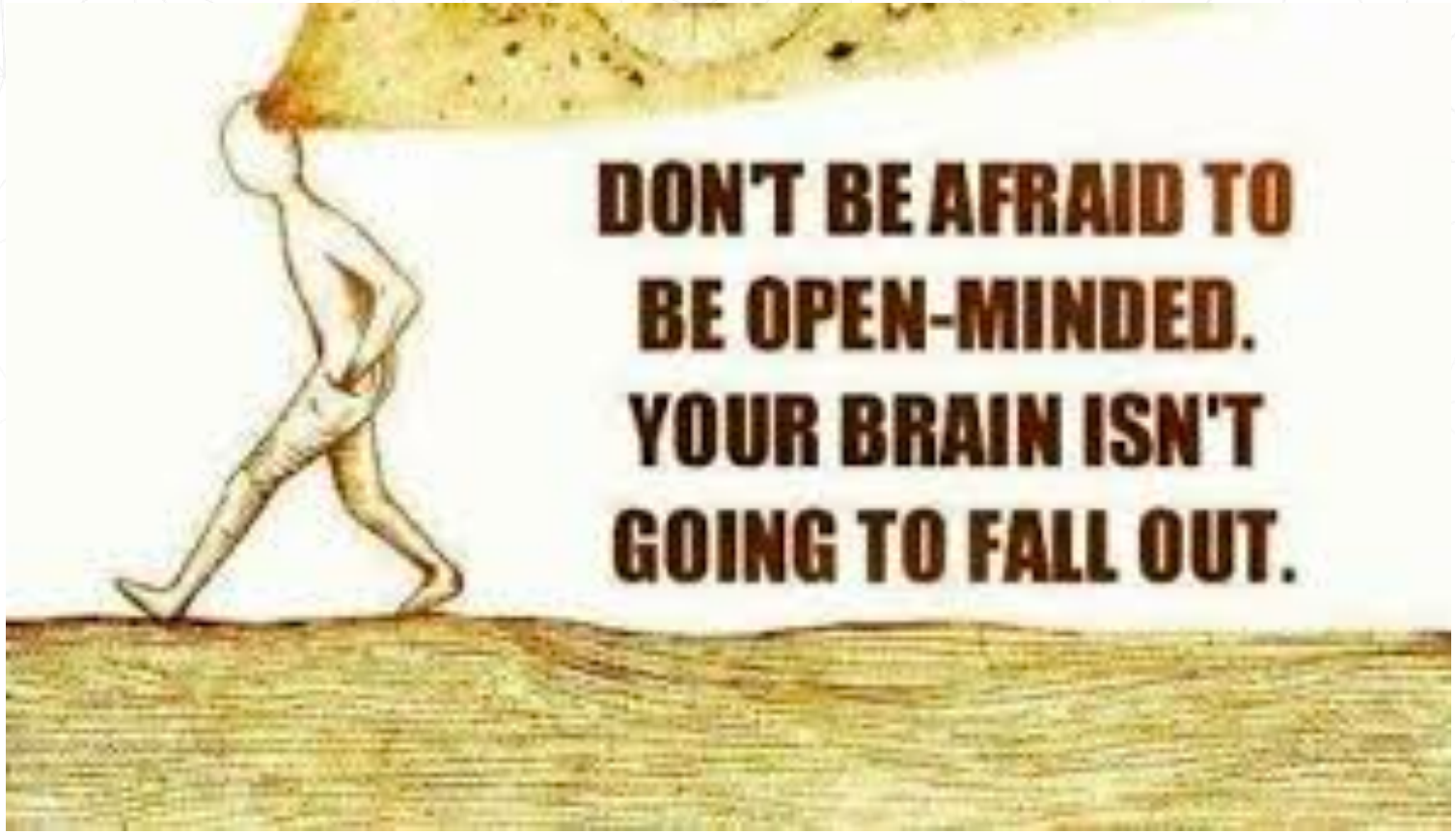


Life is a scientific experiment for you

We **all** must be open to recognizing our current ideas and thinking patterns need to change or at least adjust.

None of us have all the answers, but some people do have better answers and ways of thinking.

How and what you think determines what you do and who you are.



CULTIVATION OF COMMUNITY

Software communities are an intersection of humanity and technology.

But, they are affected more by who **we** are than what **technology** is.

It matters deeply:

- how we think about each other
- how we talk to and about each other,
- how we act towards each other.

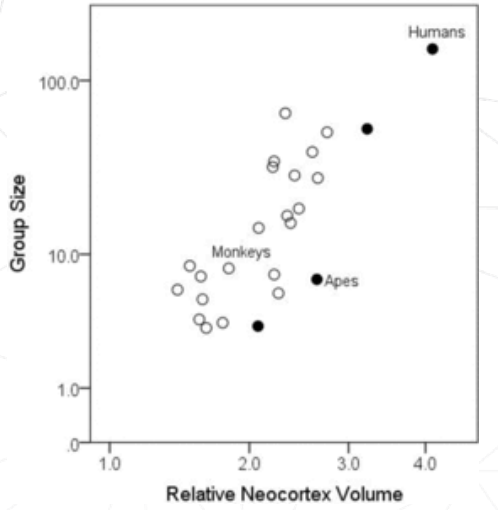
How we treat each other starts with how we treat ourselves

Individual habits matter
— your brain and body are all you have:

- good **sleep** patterns, good **diet** patterns, good **exercise** patterns
- repeated **reflections on the positive** things in your life and expressions of gratitude. This will train your feeling-brain to (remember the good things) and your life is what you remember
- **objectively learn from the negative** experiences with your thinking-brain but don't give them feeling-brain attention. Forgive and let go of any hard feelings — they just poison you.

Understand Dunbar numbers

"the bigger their brains, the larger their social groups"



Robin Dunbar

Humans have "layers" of social groups (limited by our brain capacity) — I think it's about the complexity of the model we can keep in our head about the people in our group.

Different interactions and communications are required at each of these levels —need for communication APIs

Group Size	Description
5	Best Friends
15	Working Group
50	Larger Group
150	Tribe

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

a : free from pretense or deceit : **frank**
b : easily detected or seen through : **obvious**
c : readily understood
d : characterized by visibility or accessibility of information especially concernin
business practices

- Tools facilitate this (GitHub, Slack, Gitter, Discourse, mailing-lists)
- Often a need to over-rotate to overcome negativity biases and confirmation biases of participants

COMMUNITY PRINCIPLES

How are decisions made?
Who are the people involved?
How do these evolve?

- Avoiding common pitfalls
  - Tyranny of the one
  - Back-room decision making
  - Exclusionary tendencies (unintentional or not, under-represented groups usually have something you don't understand holding them back from full participation)
  - Scape-Goating
- Good community governance **takes time and requires maintenance** — it's slower and often does not mix with other time-scales (enterprise, other software, etc.)
- Facilitation, and advocacy are often needed (face-time helpful)

COMMUNITY PRINCIPLES

The property of biological systems to remain diverse and productive indefinitely

Open Source is so important that it must be connected to a bustling marketplace of actors and agents.

All of the principles participate.

Trade is the foundation of peace and prosperity.

Connecting **Business to Open Source** is critical to sustainability.

Something we are working on to try and help.

Problem

OpenTeams

Commercial Organizations need commercial commitments and support

Open Source Projects need autonomy and independence

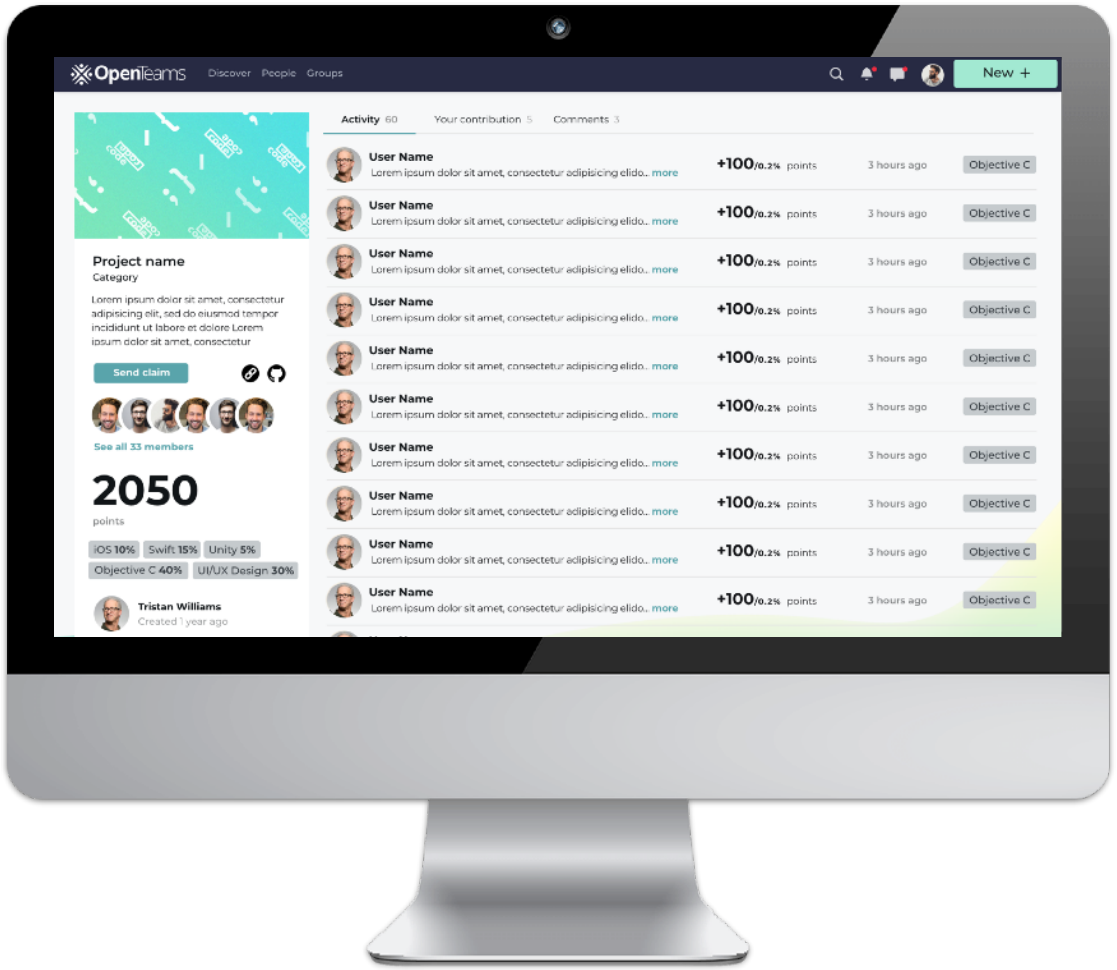
Open Source Contributors need credit and careers in OSS

OpenTeams

Works to ensure open-source contributors thrive professionally and financially.

Solution

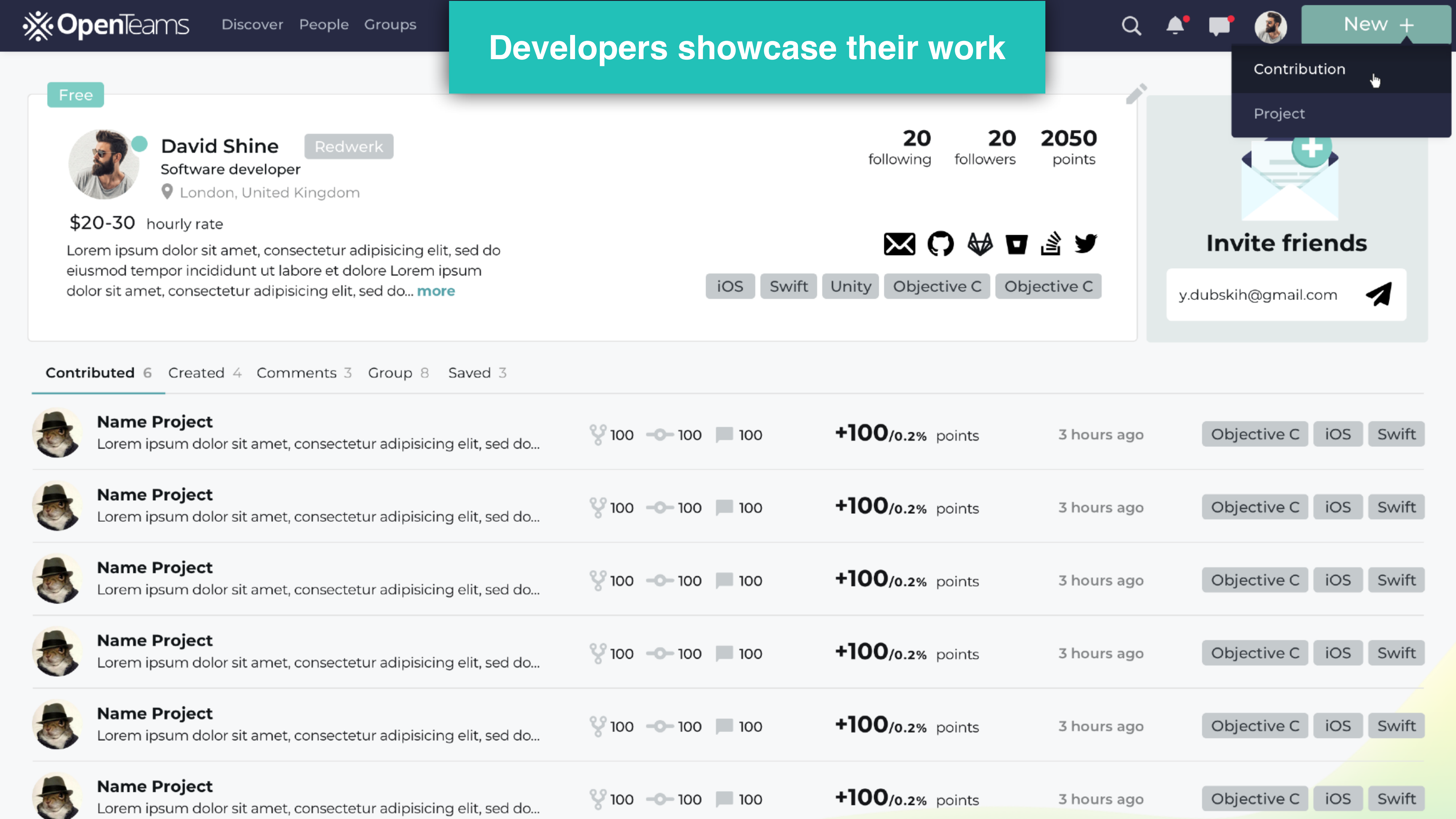
An online marketplace where companies connect with the people who write and lead open source

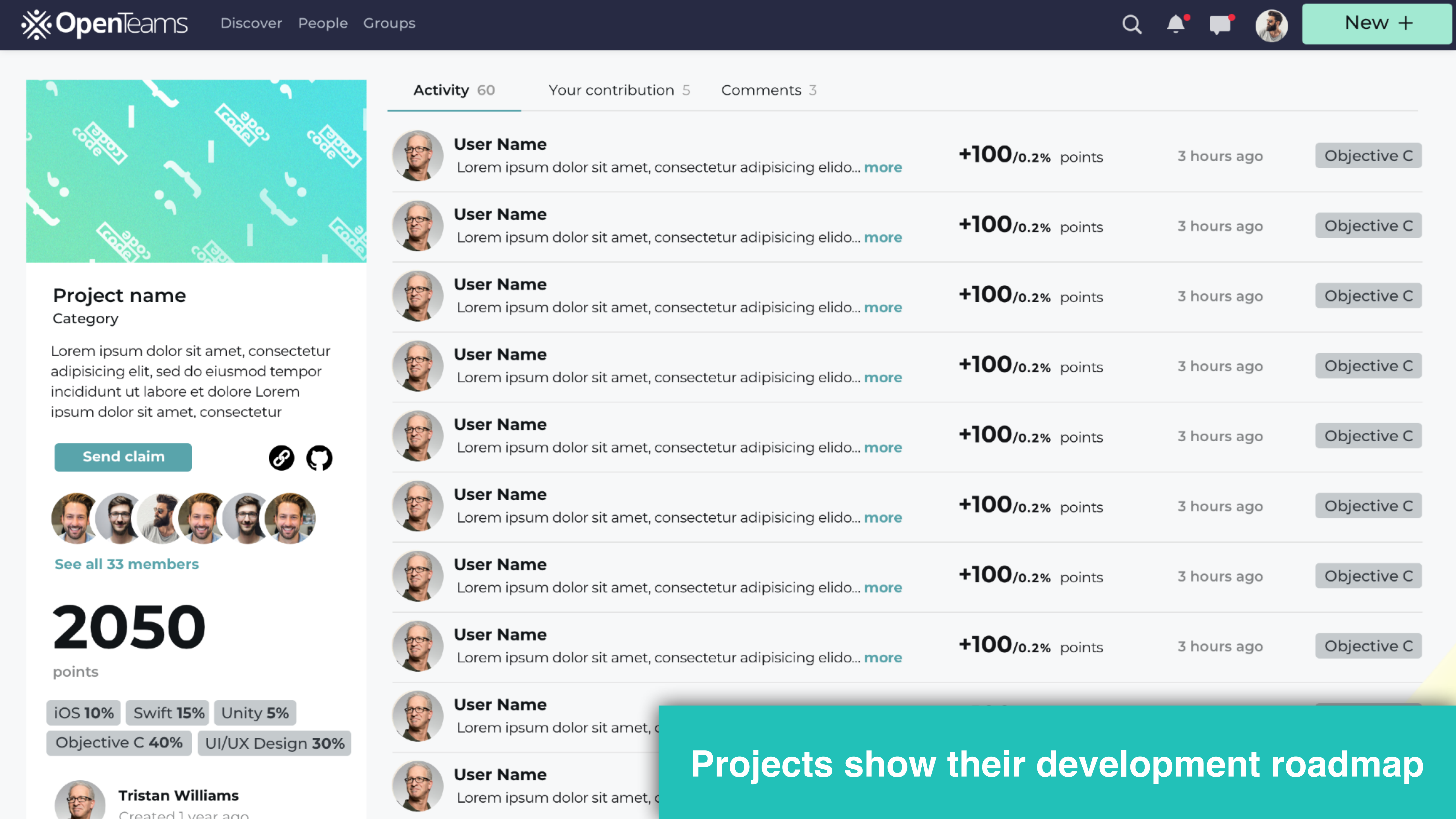


Open Source contributors
Find funding for projects

Commercial
Find and fund projects to meet their needs

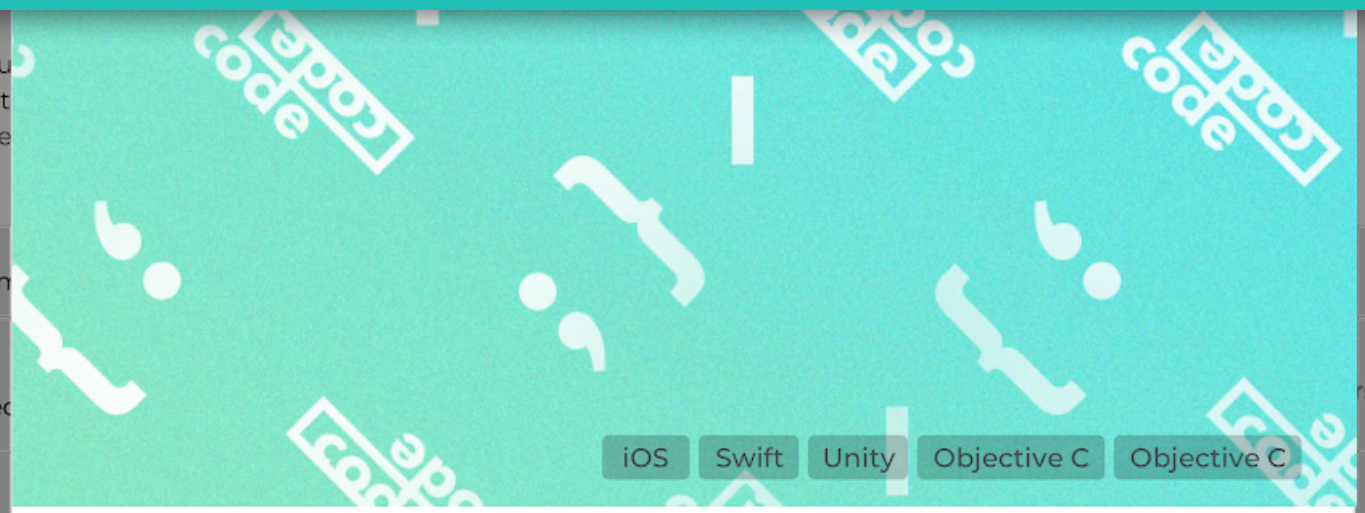
Developers showcase their work





Projects show their development roadmap

Organizations find and fund projects they depend on



Name Project

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore

Your contribution: **+50**/0.1% (Development) **+50**/0.1% (Design) 3 hours ago

100k points **24** contributors

Olivia Jordan
Created 18 hours ago

Andrew Newdigate
Updated 17 minutes ago

100 100 100 145 237

So what can you do?

1) Sign up on openteams.com and thank contributors and show how you use the tools.
2) Build your profile by making contributions to projects you care about, or creating a project you think is missing.
3) Contribute to the conversation about where your favorite open-source projects should develop (what features or integrations do you see missing).
4) Get in touch @teoliphant or travis@quansight.com especially if you are a C++ or Python or Javascript developer, or a technical writer.