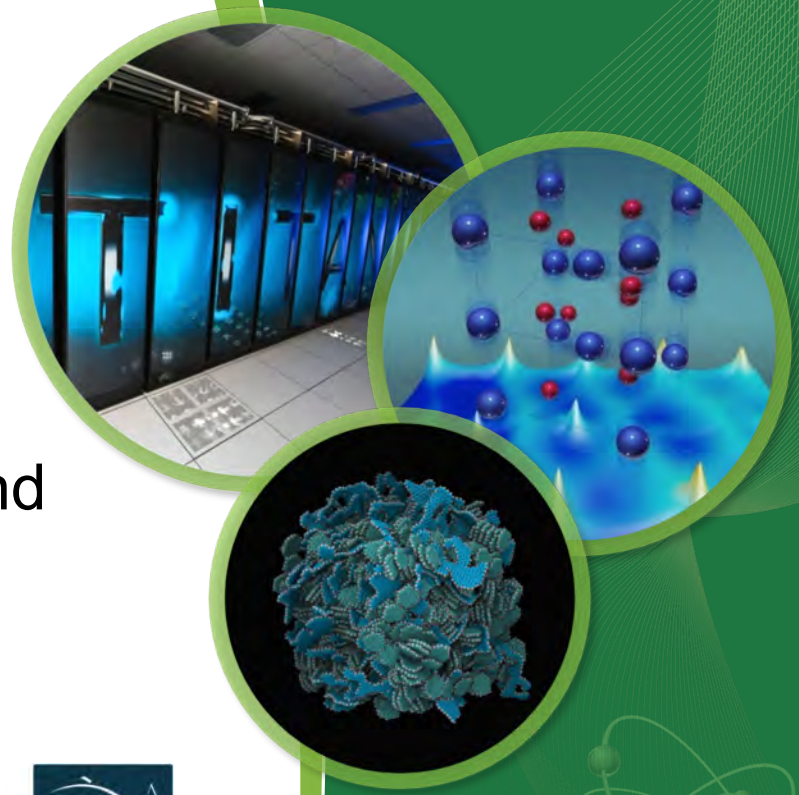
IDEAS: Software Productivity and Sustainability for the DOE

IDEAS productivity



David E. Bernholdt <*bernholdtde@ornl.gov*>, and
Anshu Dubey <*adubey@anl.gov*>

for the IDEAS Project Team

Argonne NATIONAL LABORATORY (Co-lead institution)

Sandia National Laboratories (Co-lead institution)

Los Alamos NATIONAL LABORATORY

BERKELEY LAB

Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory

UNIVERSITY OF OREGON

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

DOE's Take on Software Productivity and Sustainability

- ASCR is not interested in funding software engineering research
  - Identifying best practices and helping adapt them to HPC/CSE *is* in scope
- But they and others in DOE are interested in the overall scientific productivity of their researchers
  - And they recognize that software plays a key role in that
- So are willing to support work on enhancing developer productivity, software sustainability, etc.
  - At least from time to time…
  - At least in a couple of programs…

Supported Projects in the DOE World

- IDEAS = Interoperable Design of Extreme-Scale Application Software
- IDEAS-Classic
  - Funded by ASCR + BER Terrestrial Ecosystem Modeling
  - Improving interoperability of key DOE numerical libraries
- IDEAS-ECP
  - Funded by Exascale Computing Project (ECP)
- Common themes
  - Identifying, documenting, disseminating best practices
    - Better Scientific Software (http://bssw.io)
  - Engagement with individual dev teams
    - Identify development pain points, bottlenecks; guide and assist to improve
  - Outreach: tutorials, workshops, minisymposia, etc.

Lessons Learned

- There is a significant appetite for **training on software development practices**
  - Our webinar series *Best Practices for HPC Software Developers* gets consistently high participation
  - Also successful tutorials in various venues
  - **Don't underestimate the level of effort required to sustain such activities**
- **Awareness of existing resources** for software development is an important gap
  - *Better Scientific Software*, (http://bssw.io) was developed to address this
  - Intended to be community-based. We are providing the nucleus; we want and need contributions from the broad community!

- **Development teams are busy**
  - It can be hard for them to find time to interact, even if it is with people who want to help them
  - We're not abandoning 1:1 interactions…
  - But maybe more emphasis on small, easily digestible resources for use "on demand" and without having to coordinate with outsiders
- **Encourage discussion and recognition of software** in other venues
  - Workshops, publication opportunities, recognition and reward systems
- Combine all of the above to create a virtuous cycle!

moz://a

Work Open, Lead Open

#WOLO

Hello!
I'm Abby

Practice Lead,
Working Open
Mozilla Foundation

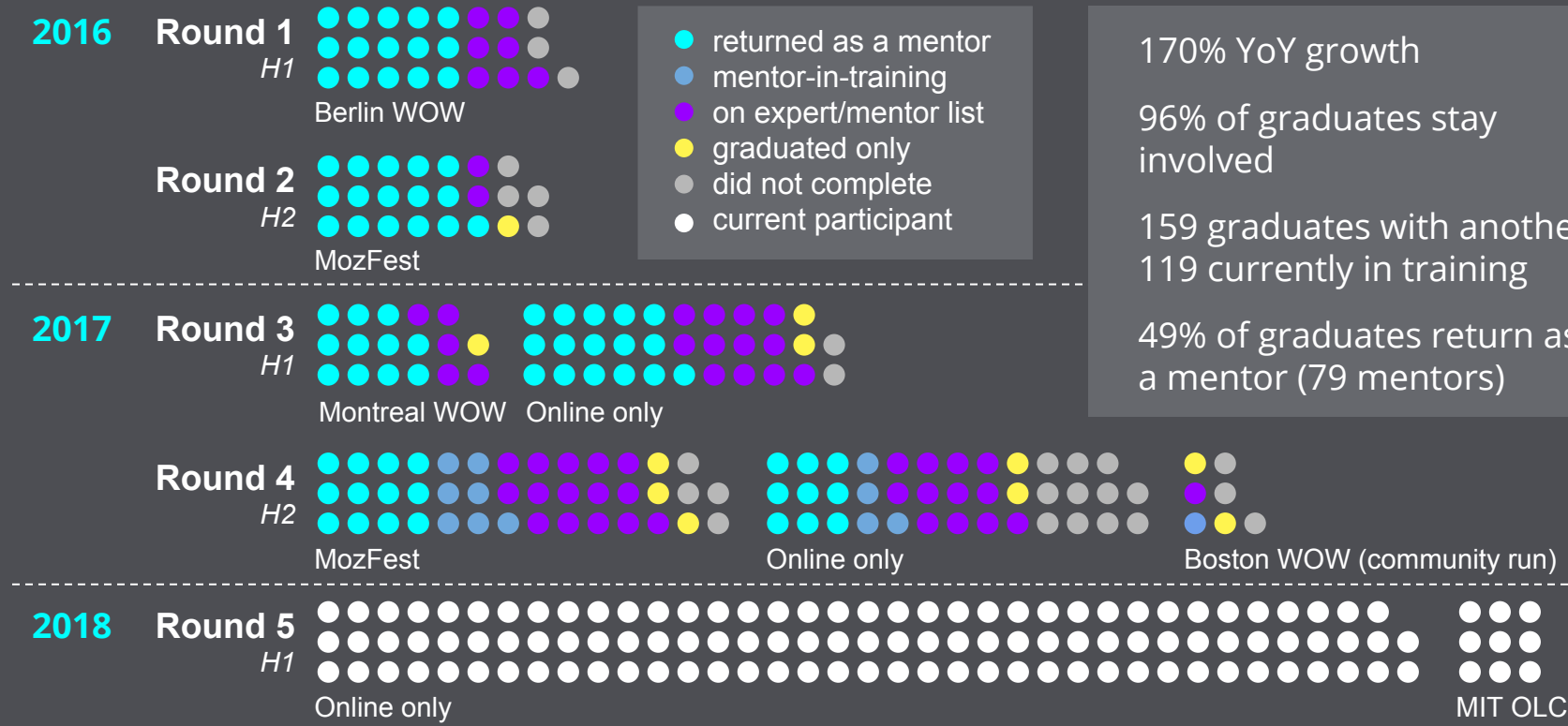
I want to grow a culture of openness in innovation and research

Mozilla Open Leaders

Mentorship & training on open practices

Fueling the Movement

We're growing & retaining our graduates



170% YoY growth

96% of graduates stay involved

159 graduates with another 119 currently in training

49% of graduates return as a mentor (79 mentors)

What do successful projects have?

Projects that grow & become sustainable need support in:

Training & Best Practices

Templates, case studies

Peer Support

Cohort, mentors

Resources

Time, funding, skills

Current Research in Open Source

- Gender bias in code reviews
- Community interaction types
- CHAOSS - Linux Foundation
- Project Bugmark

Open Leadership Framework

https://mzl.la/olf

moz://a

Thank you!

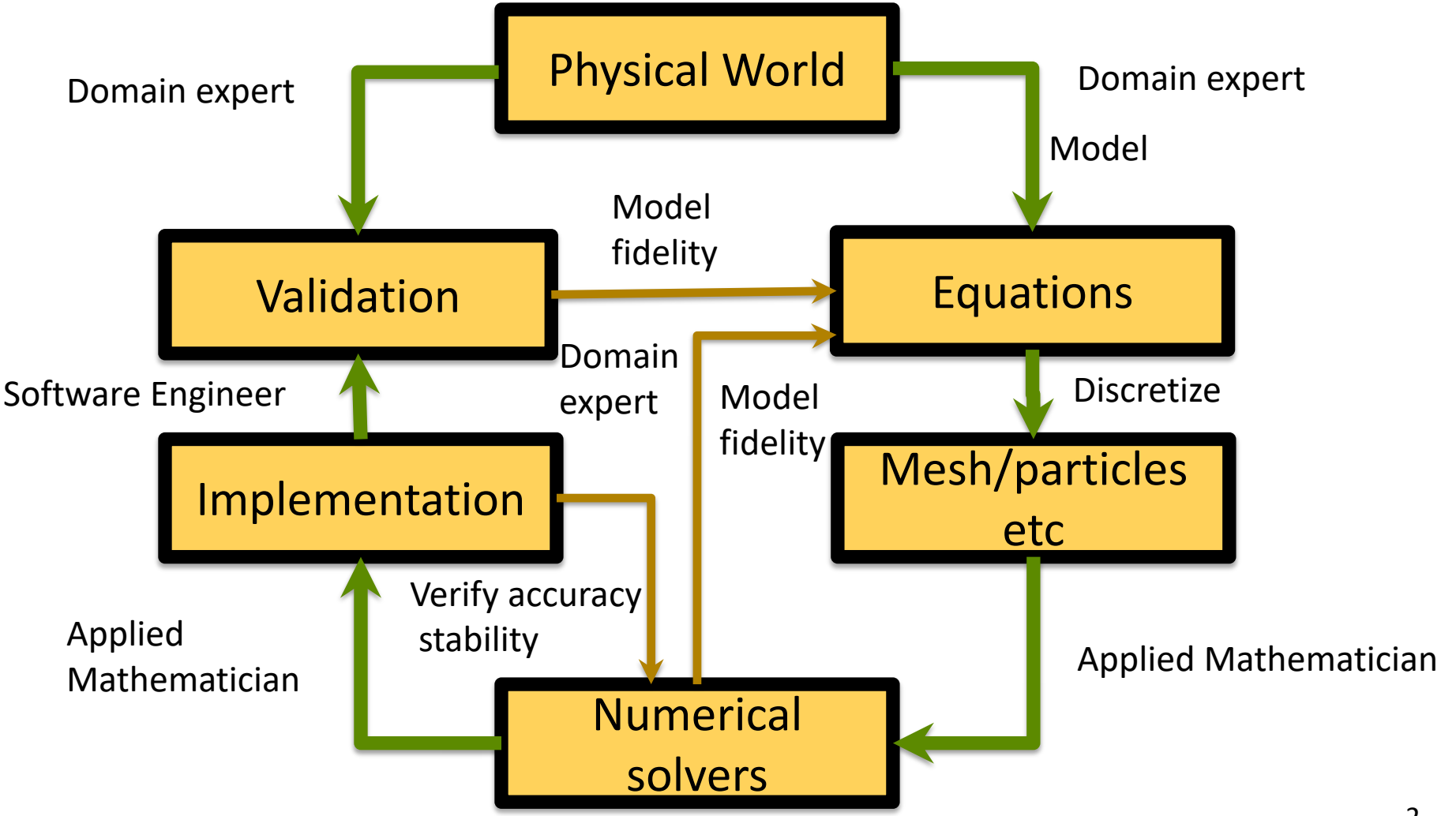
#WOLO // @abbycabs

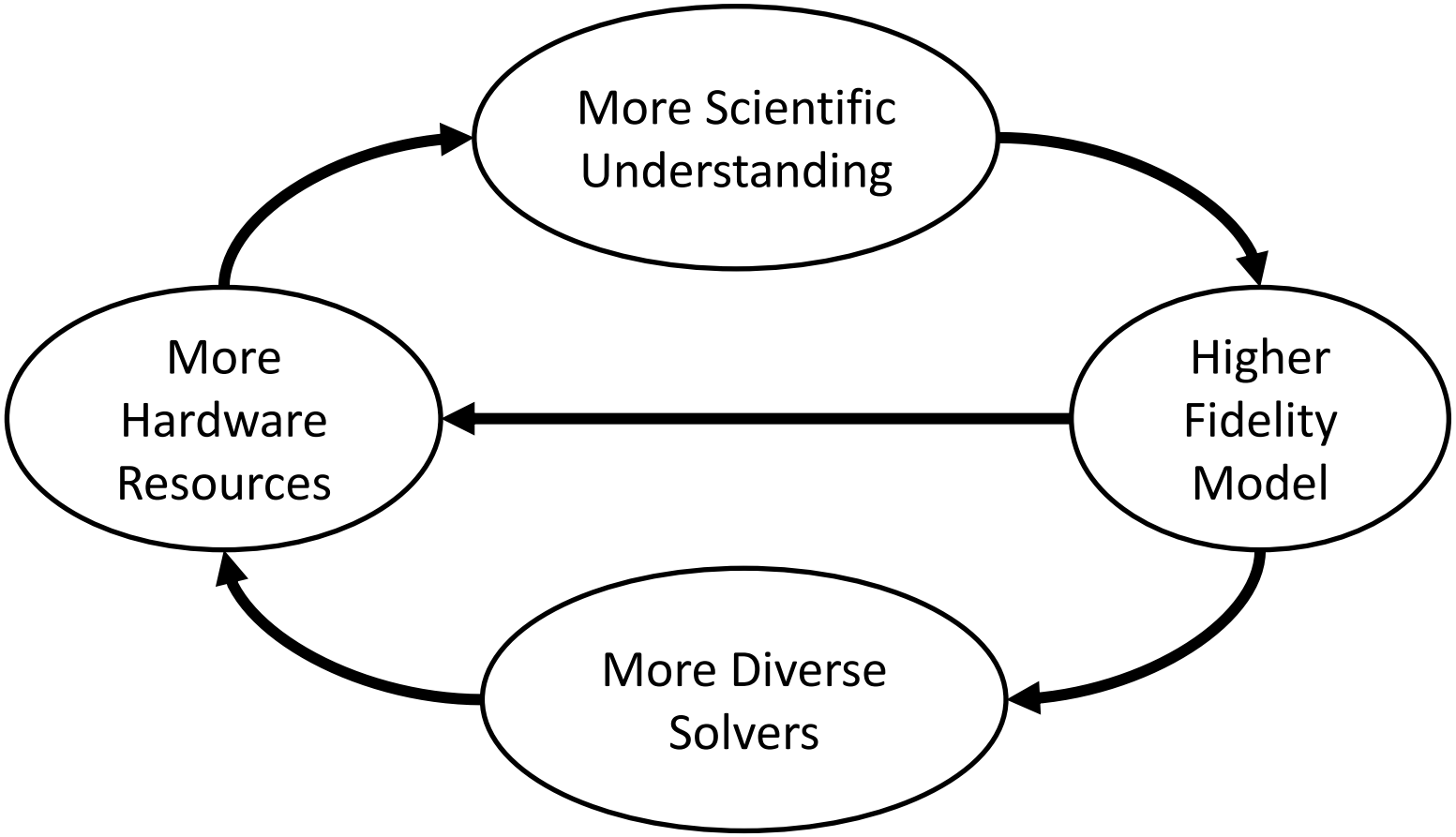
SCIENTIFIC IMPACT OF SOFTWARE DESIGN INVESTMENT

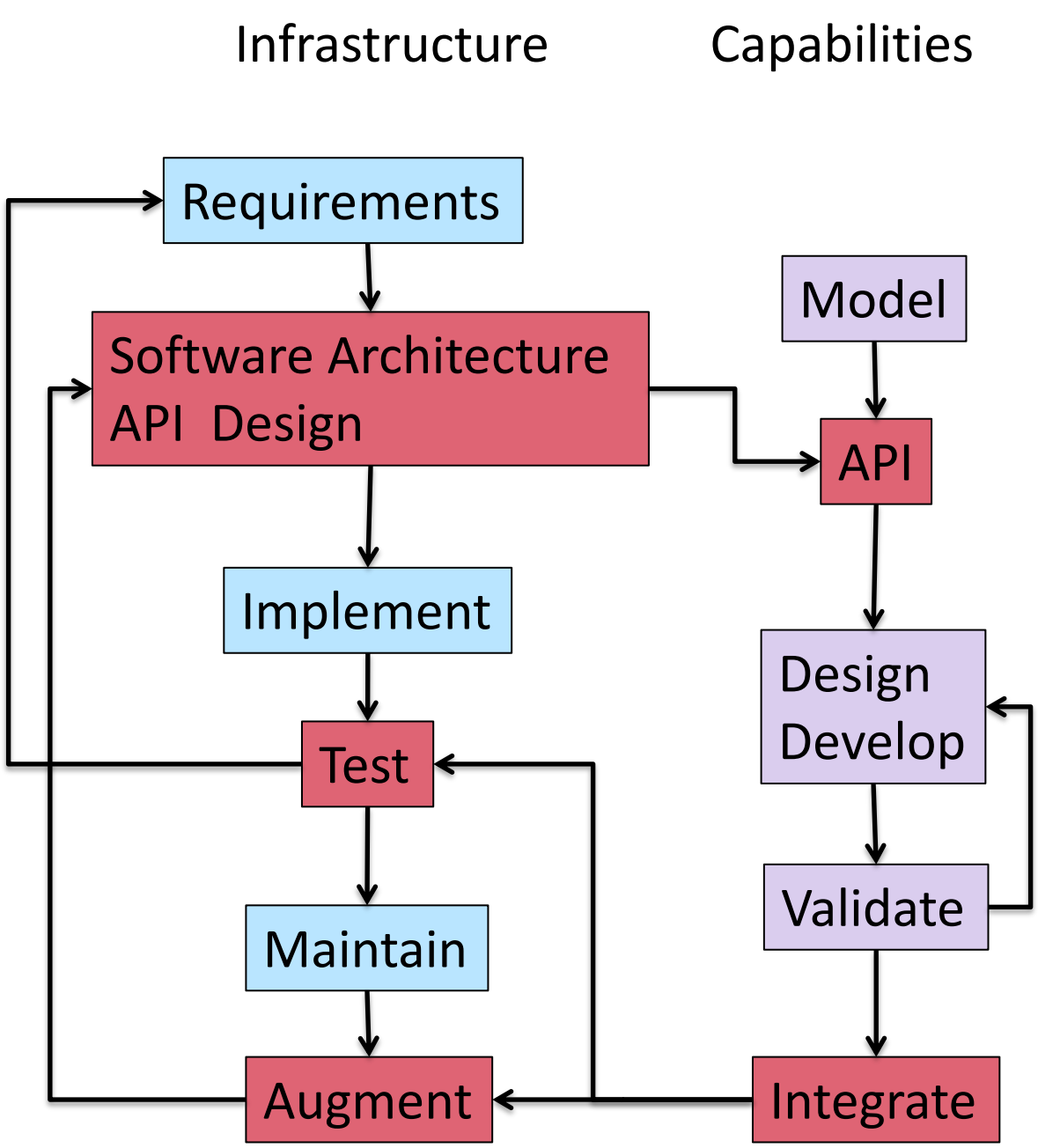
ANSHU DUBEY

URSSI First Workshop
April 10, 2018
Berkeley, CA

WORKFLOW AND EXPERTISE



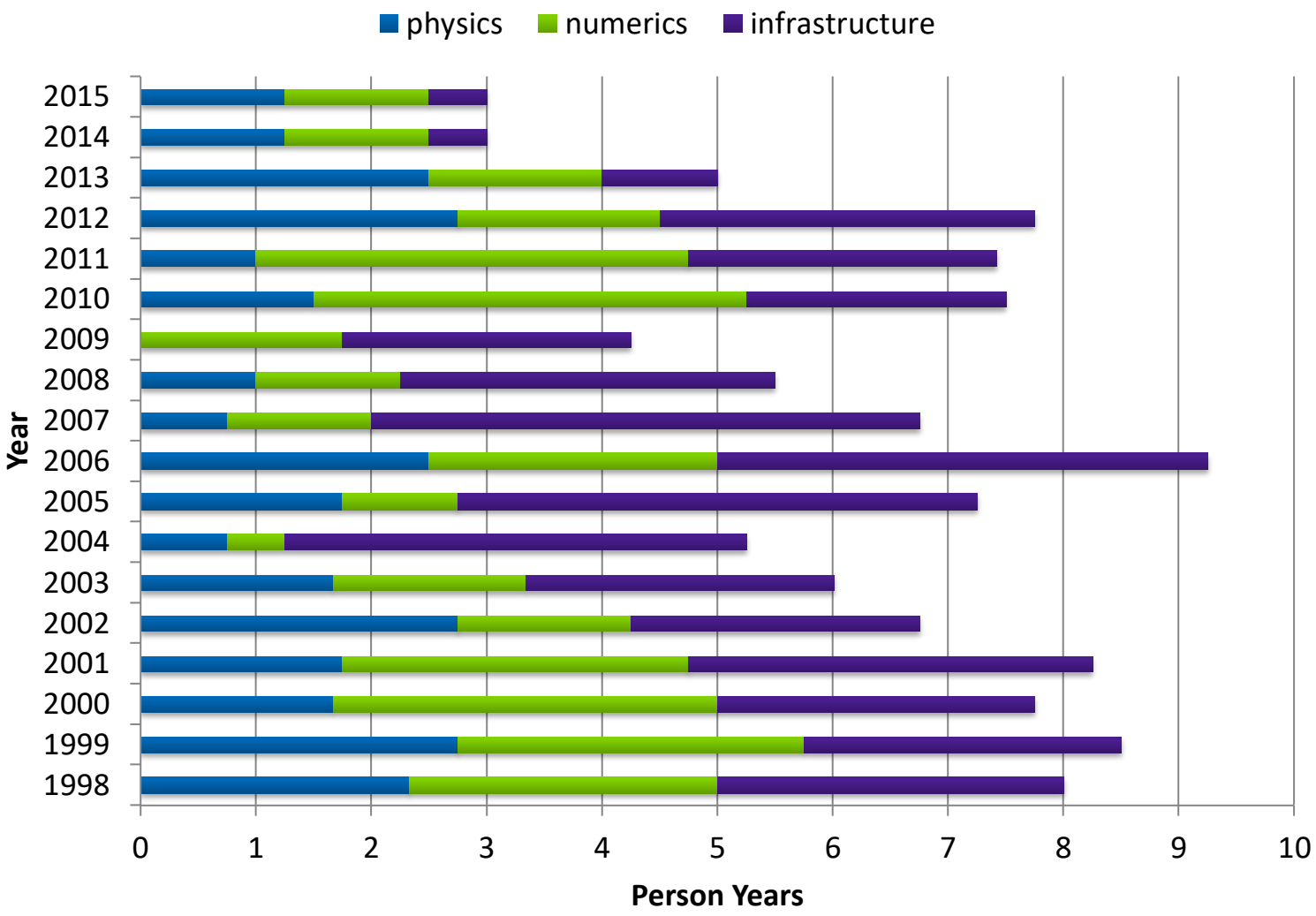

More Scientific Understanding
Higher Fidelity Model
More Diverse Solvers
More Hardware Resources


Infrastructure
Capabilities
Requirements
Software Architecture
API Design
Implement
Test
Maintain
Augment
Model
API
Design
Develop
Validate
Integrate

DESIGN INVESTMENT IMPACT

capabilities	categories	new community reached	year
base	all	thermonuclear astrophysics	2000
MHD	physics	reconnection, solar plasma	2002
particles	physics and infrastructure	cosmology	2003
multigrid	infrastructure	CFD	2008
Lagrangian Markers	infrastructure	FSI	2009
PIC	physics and infrastructure	plasma	2010
nuclear EOS, neutrino source terms and leakage	physics	core-collapse supernovae	2010-2012
3-T, conductivity Radiation, laser	physics and infrastructure	HEDP	2010-2011
sink particles	physics	star formation	2012

DESIGN INVESTMENT IMPACT

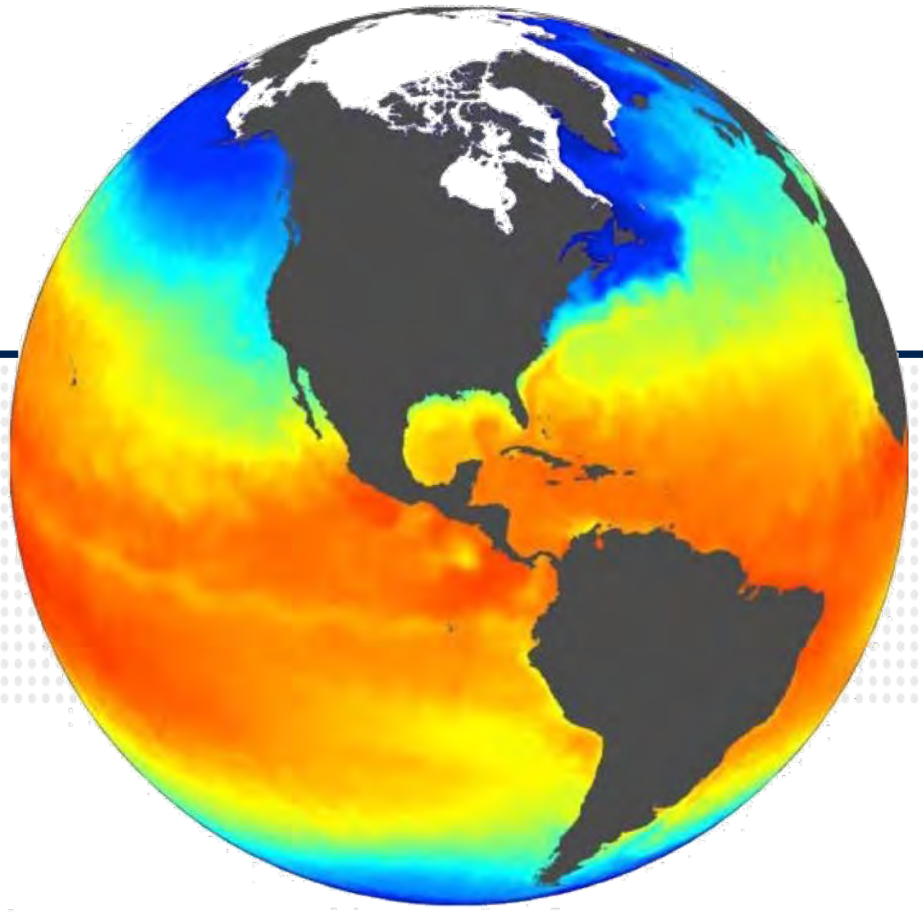


DESIGN INVESTMENT IMPACT

Domain	initial effort (person years)	total effort (person years)	release date (year)	first publication (year)	total publications
Supernova,Astro	16	60	2000	2000	427
Cosmology	2.5	20	2001	2001	257
Solar physics	0.5	5	2000	2002	70
HEDP	8	28	2010	2010	38
Core-collapse	1	5	2012	2008	20
CFD/FSI	3	5		2012	5

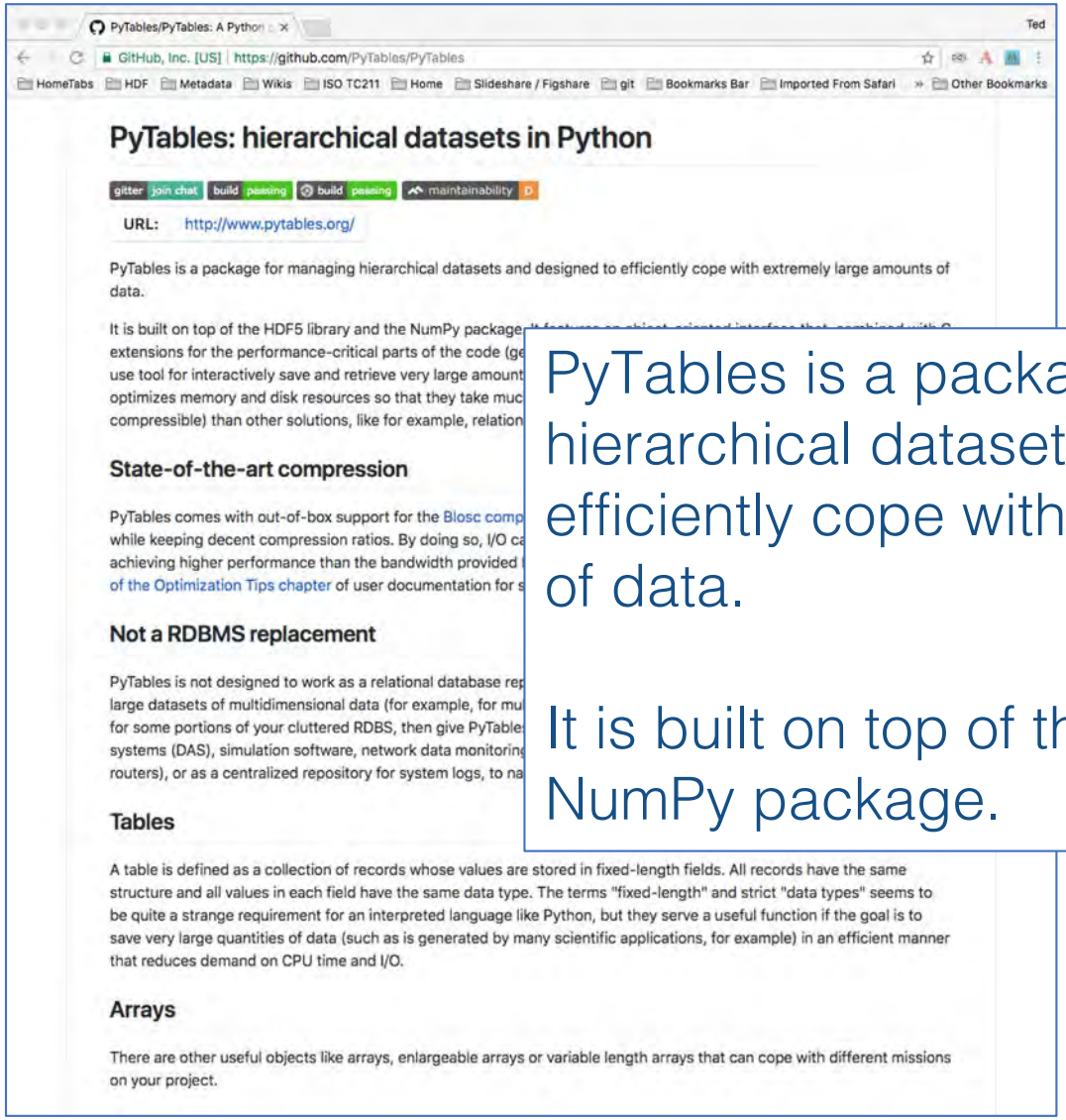
Table 4: Correlation between effort, duration released, and publications.

Sustainable Open-Source Tools for Sharing and Understanding Data



Ted Habermann
Director of Earth Science
The HDF Group

PyTables

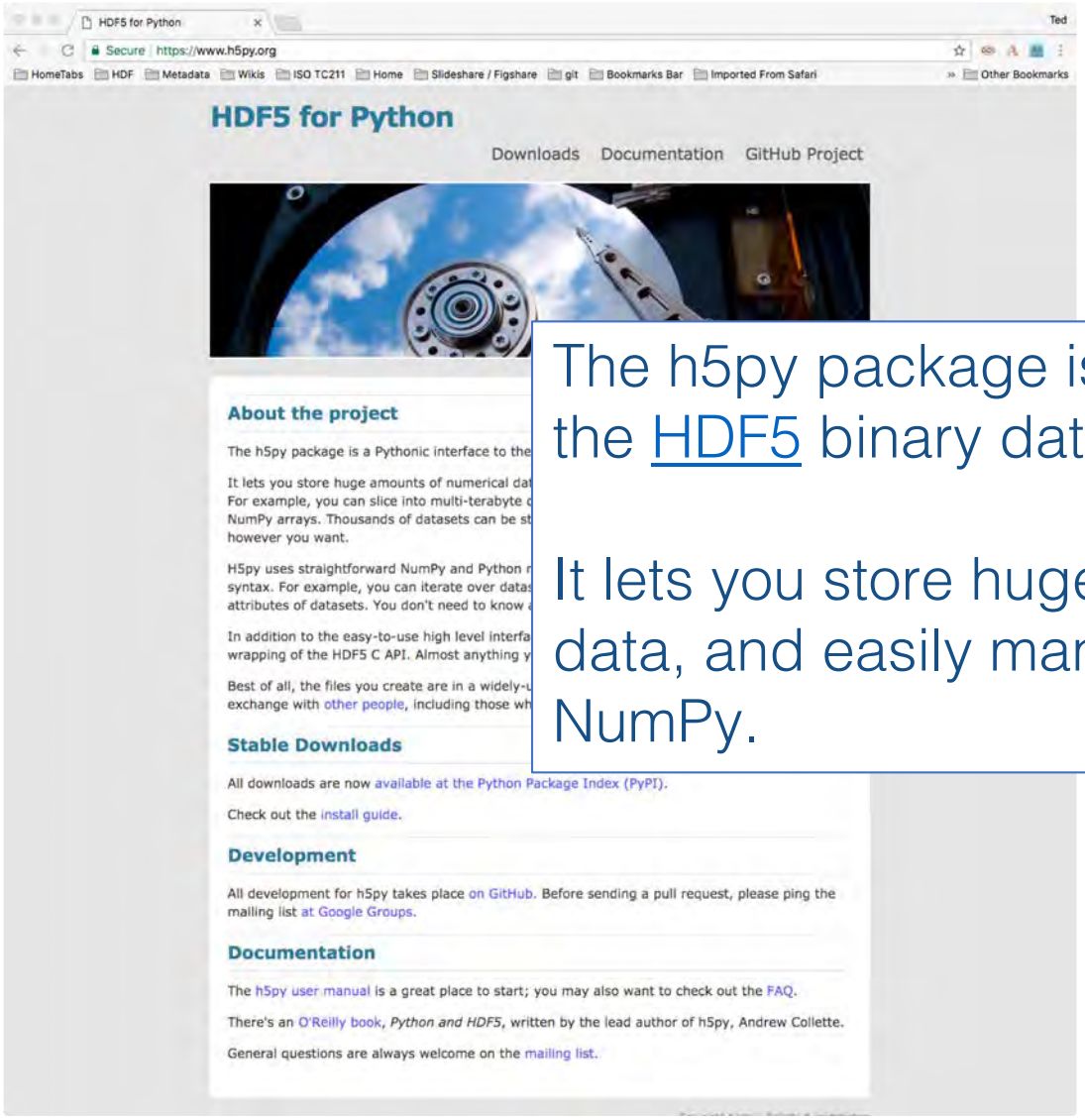


PyTables is a package for managing hierarchical datasets and designed to efficiently cope with extremely large amounts of data.

It is built on top of the HDF5 library and the NumPy package.

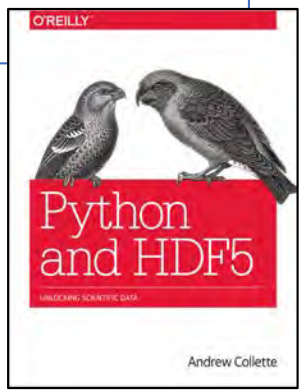
https://github.com/PyTables/PyTables

H5Py



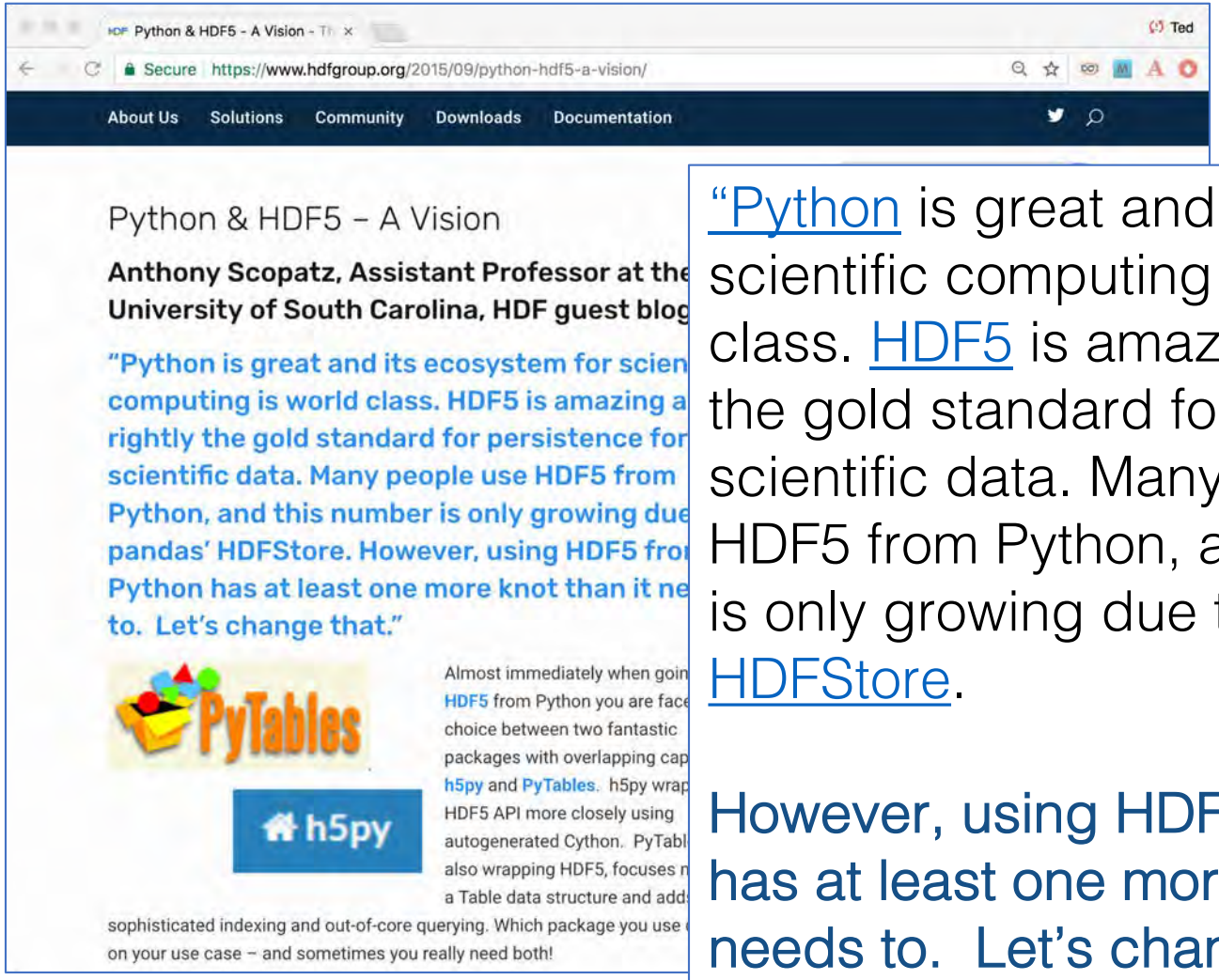
The h5py package is a Pythonic interface to the HDF5 binary data format.

It lets you store huge amounts of numerical data, and easily manipulate that data from NumPy.



https://www.h5py.org/

A Vision

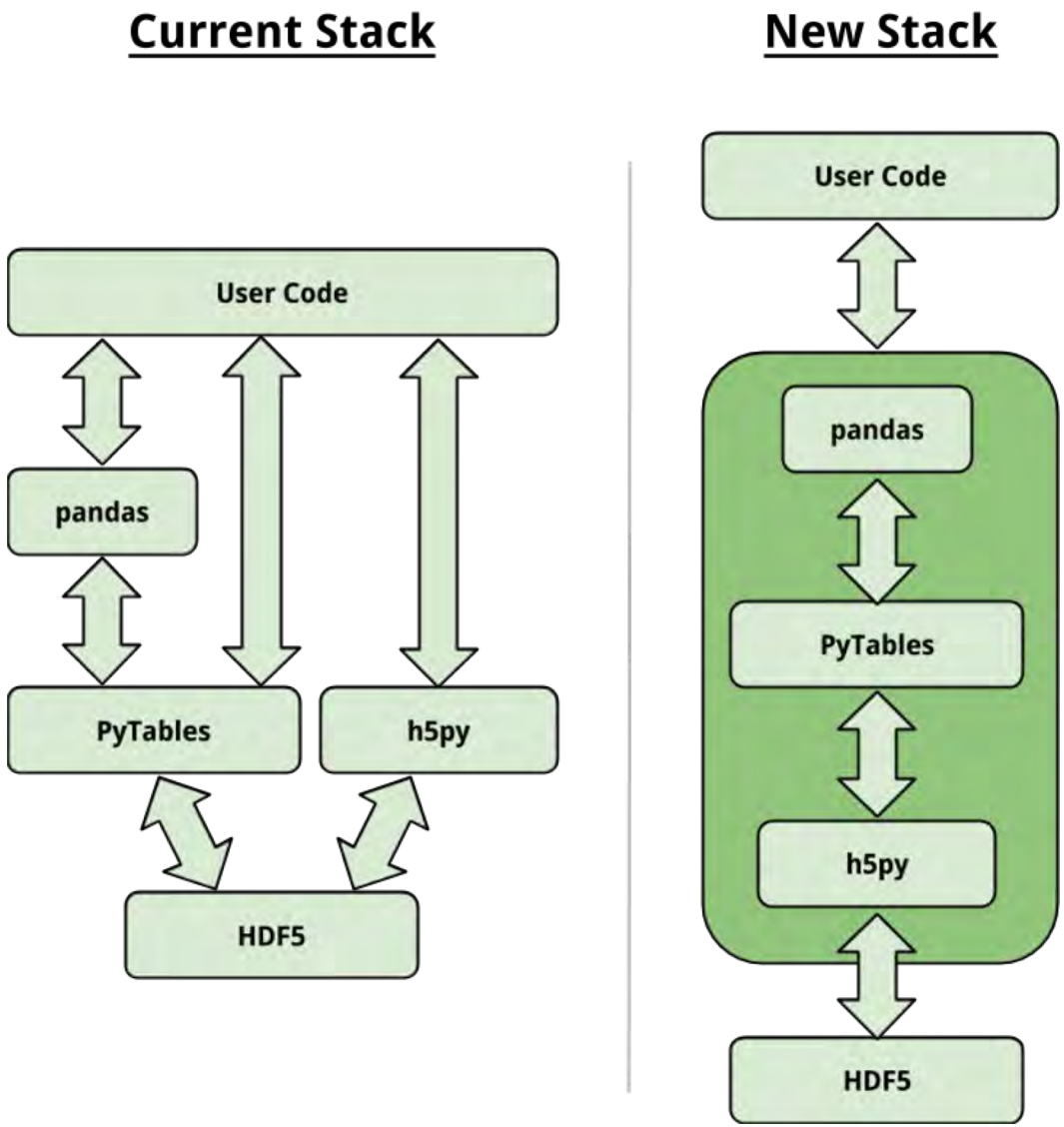


"Python is great and its ecosystem for scientific computing is world class. HDF5 is amazing and is rightly the gold standard for persistence for scientific data. Many people use HDF5 from Python, and this number is only growing due to pandas' HDFStore.

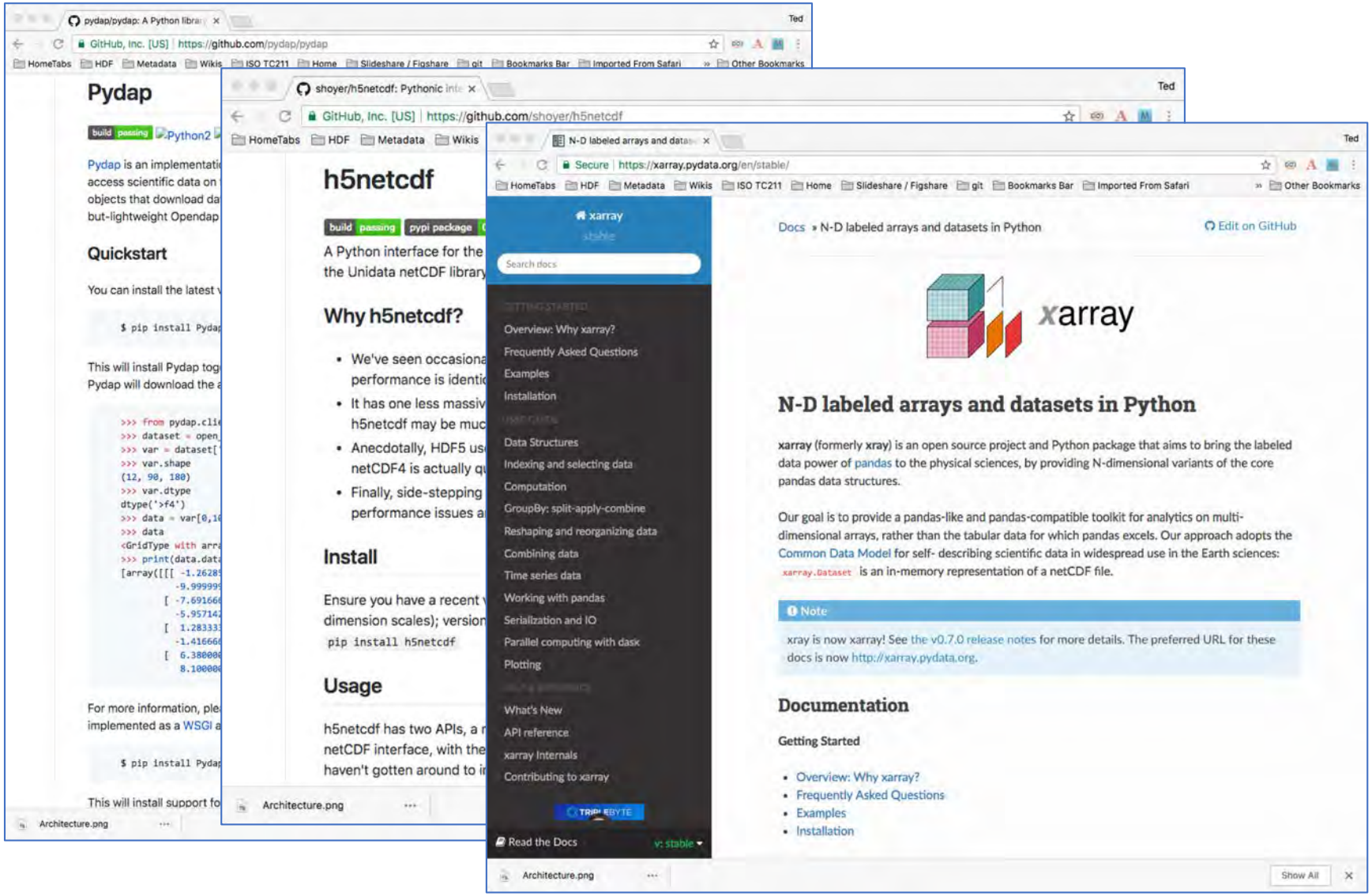
However, using HDF5 from Python has at least one more knot than it needs to. Let's change that."

https://www.hdfgroup.org/2015/09/python-hdf5-a-vision/

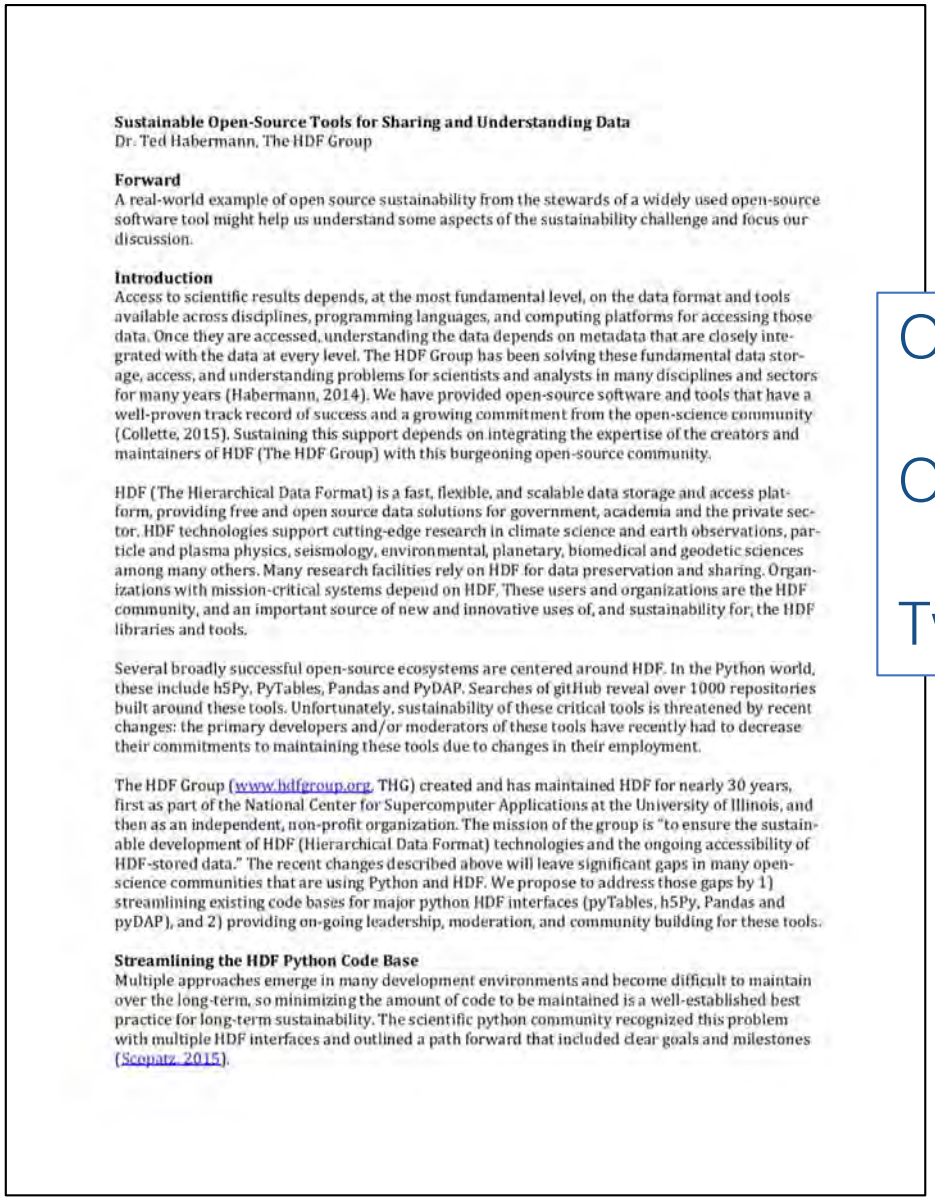
Architecture



Other winners



A Project



Sustainable Open-Source Tools for Sharing and Understanding Data
Dr. Ted Habermann, The HDF Group

Forward
A real-world example of open source sustainability from the stewards of a widely used open-source software tool might help us understand some aspects of the sustainability challenge and focus our discussion.

Introduction
Access to scientific results depends, at the most fundamental level, on the data format and tools available across disciplines, programming languages, and computing platforms for accessing those data. Once they are accessed, understanding the data depends on metadata that are closely integrated with the data at every level. The HDF Group has been solving these fundamental data storage, access, and understanding problems for scientists and analysts in many disciplines and sectors for many years (Habermann, 2014). We have provided open-source software and tools that have a well-proven track record of success and a growing commitment from the open-science community (Collette, 2015). Sustaining this support depends on integrating the expertise of the creators and maintainers of HDF (The HDF Group) with this burgeoning open-source community.

HDF (The Hierarchical Data Format) is a fast, flexible, and scalable data storage and access platform, providing free and open source data solutions for government, academia and the private sector. HDF technologies support cutting-edge research in climate science and earth observations, particle and plasma physics, seismology, environmental, planetary, biomedical and geodetic sciences among many others. Many research facilities rely on HDF for data preservation and sharing. Organizations with mission-critical systems depend on HDF. These users and organizations are the HDF community, and an important source of new and innovative uses of, and sustainability for, the HDF libraries and tools.

Several broadly successful open-source ecosystems are centered around HDF. In the Python world, these include h5Py, PyTables, Pandas and PyDAP. Searches of gitHub reveal over 1000 repositories built around these tools. Unfortunately, sustainability of these critical tools is threatened by recent changes: the primary developers and/or moderators of these tools have recently had to decrease their commitments to maintaining these tools due to changes in their employment.

The HDF Group (www.hdfgroup.org, THG) created and has maintained HDF for nearly 30 years, first as part of the National Center for Supercomputer Applications at the University of Illinois, and then as an independent, non-profit organization. The mission of the group is "to ensure the sustainable development of HDF (Hierarchical Data Format) technologies and the ongoing accessibility of HDF-stored data." The recent changes described above will leave significant gaps in many open-science communities that are using Python and HDF. We propose to address those gaps by 1) streamlining existing code bases for major python HDF interfaces (pyTables, h5Py, Pandas and pyDAP), and 2) providing on-going leadership, moderation, and community building for these tools.

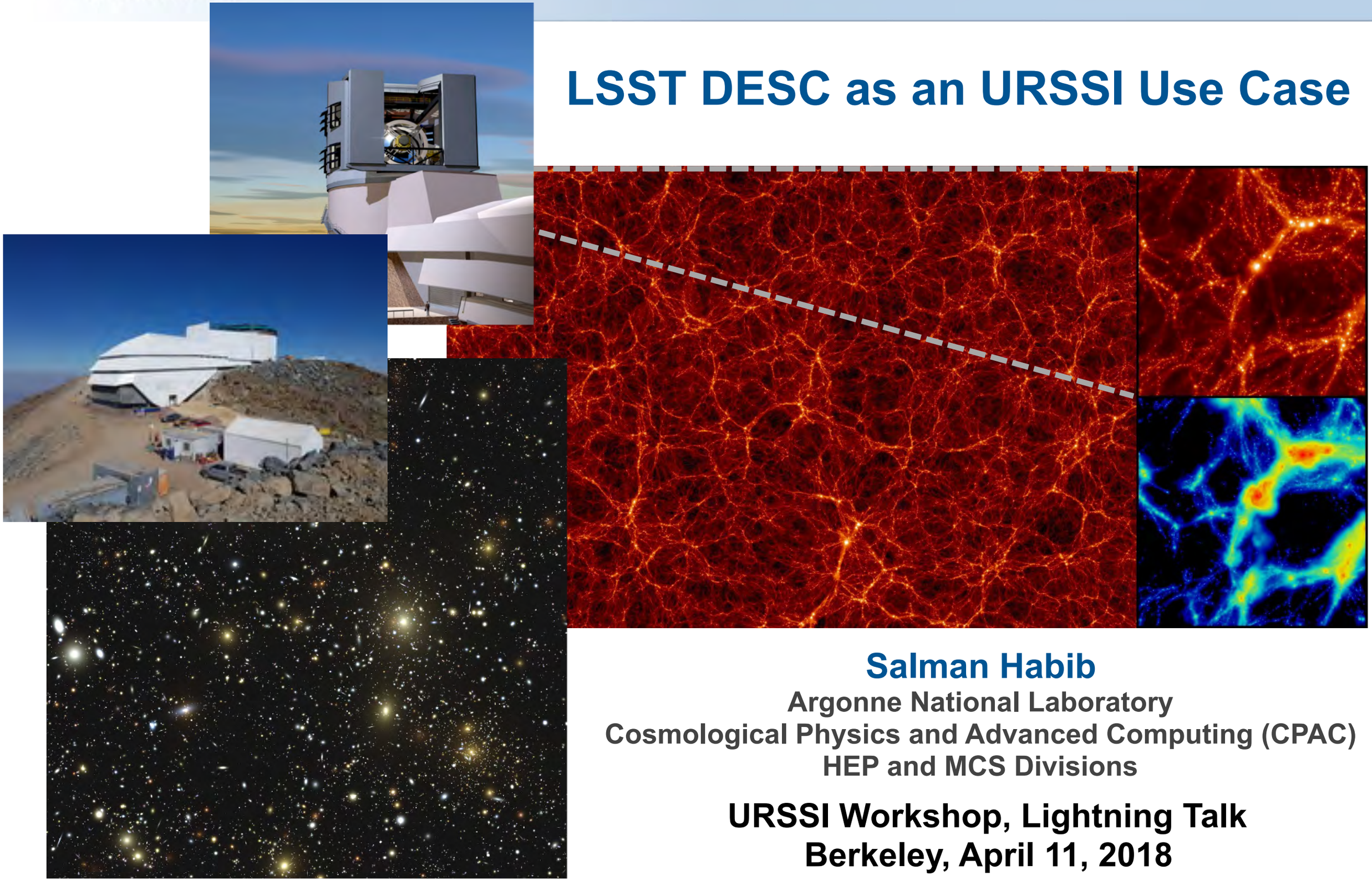
Streamlining the HDF Python Code Base
Multiple approaches emerge in many development environments and become difficult to maintain over the long-term, so minimizing the amount of code to be maintained is a well-established best practice for long-term sustainability. The scientific python community recognized this problem with multiple HDF interfaces and outlined a path forward that included clear goals and milestones (Scopatz, 2015).

One developer

One community expert

Two years

LSST DESC as an URSSI Use Case



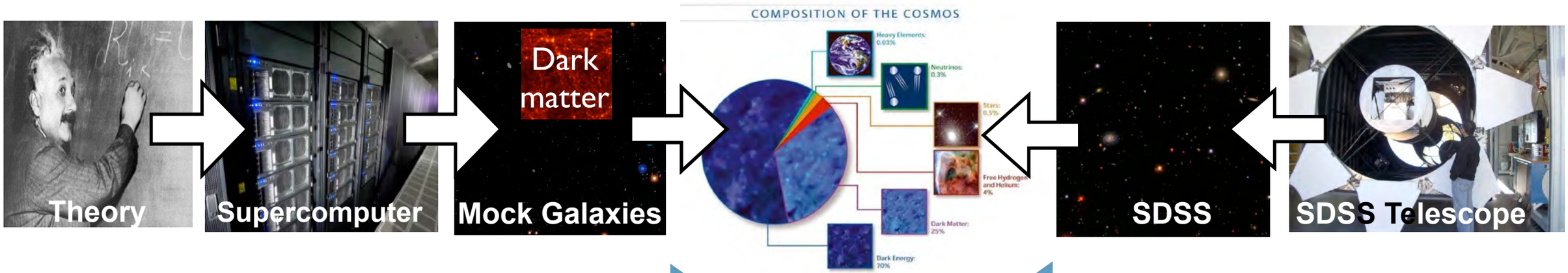
Salman Habib

Argonne National Laboratory
Cosmological Physics and Advanced Computing (CPAC)
HEP and MCS Divisions

URSSI Workshop, Lightning Talk
Berkeley, April 11, 2018

End-to-End Computing Paradigm

Simulated Data: 1) Large-scale simulation of the Universe, 2) Synthetic catalogs, 3) Statistical inference (cosmology); **Analysis:** Comparison with actual data



- **Multiple types and layers of computing**
  - Extreme-scale supercomputing for cosmological simulations
  - Large-scale data-intensive computing for synthetic catalogs (includes ML)
  - Medium-scale data-intensive computing for data reduction and analysis
  - “Many-user” Jupyter style querying and analysis for smaller projects

ECP
EXASCALE COMPUTING PROJECT

LSST Dark Energy Science Collaboration

- **Dark Energy Probes**
  - Supernovae, other transients
  - Galaxy clustering
  - Distribution of matter and its evolution (weak lensing)
  - Galaxy clusters
  - Cross-correlations with other surveys (multi-wavelength)
- **Data/Software Issues**
  - Roughly a PB or so per year of curated, reduced data, 10s of PB of raw data (starting roughly 2023)
  - 10-year survey
  - DESC will want to reprocess data
  - DESC will need its own analysis pipelines
  - As a software collaboration, DESC is underfunded and understaffed in critical areas
  - 700 members already, will have more than 1000 by the time data arrives
- **URSSI Possibilities**
  - Software design, planning, management in a sparse resource environment
  - Software standards implementation + training of a diverse group of researchers
  - Expect this to happen in an era of rapid hardware and software evolution
  - Need to raise community awareness of major problems ahead (not BAU!)

NUMFOCUS

OPEN CODE = BETTER SCIENCE

DISC

Diversity and Inclusion in Scientific Computing

Gina Helfrich, Ph.D.

Communications Director,
Program Manager for Diversity & Inclusion

How Diverse is the Open Source Scientific Computing Community?

Points Not Addressed:

- **Why** we care about diversity
- **Causes** of monoculture/lack of diversity in tech & open source
- The role of **inclusion** and how it is different from diversity

Steps to Tracking Diversity

1. Define diversity metrics
2. Survey/assess the community
3. Analyze results

1. Define diversity metrics

Potential diversity metrics:

- Gender
- Sexual Identity
- Race
- Ethnicity
- Age
- Disability Status
- Industry
- Experience Level

These are sensitive data points!
(especially taken together)

How do we collect this information accurately and responsibly?

We have to make ethical choices about how to collect, categorize, and report diversity data:

- Preserve anonymity
- Secure sensitive data
- Avoid further alienating people who already feel marginalized

If we ask about gender:

- Should we use a **text box** or **provide options**?
- Will **trans* & non-binary folks** feel **alienated** by the form of the question? Will they worry about being **outed**?
- Will **women** feel **tokenized**? Will reminding women of their gender **prime their answers** to other questions?

PyData

Conferences	Attendees	Speakers	Members	Groups	Countries
11	3,500	330	50,413	73	33

What does “diverse” mean in this context?

If we ask about race:

- How should we distinguish **race** from **ethnicity**?

- How do we **select and interpret racial categories** when looking at a community that stretches across countries and continents? (e.g. African-American, Afro-Brazilian, Black British, Black African)

2. Survey/Assess the Community

How do we collect this information accurately?

Make questions mandatory

How do we collect this information responsibly?

Make questions optional

Shortcut methods are problematic

Tools like genderize.io determine the gender of a first name based on historical data

- Assigning a gender to individuals **without directly asking them** is particularly problematic for trans*, non-binary, and gender non-conforming individuals.

3. Analyze Results

Existing, Recent Survey Data I'm Aware Of:

- **2017 GitHub Open Source Survey: 5,500 respondents (24 million people use GitHub across 200 countries.)**
  - 95% male
  - 3% female
  - 1% non-binary

Stack Overflow 2018 Developer Survey:

100,000 respondents

- **Gender**
  - 92.9% male
  - 6.9% female
  - 0.9% non-binary, genderqueer, or gender non-conforming

- **Race & Ethnicity**
  - 74.2% White or of European descent
  - 11.5% South Asian
  - 6.7% Hispanic or Latino/Latina
  - 5.1% East Asian
  - 4.1% Middle Eastern
  - 2.8% Black or of African descent
  - 0.8% Native American, Pacific Islander, or Indigenous Australian

Stack Overflow 2018 Developer Survey (Cont'd)

- **Sexual Orientation (first year this was asked)**
  - 93.2% Straight or heterosexual
  - 4.3% Bisexual or Queer
  - 2.4% Gay or Lesbian
  - 1.9% Asexual

- **Disability Status**
  - 8.5% I have a mood or emotional disorder (ex. depression, bipolar disorder)
  - 7.8% I have an anxiety disorder
  - 5.9% I have a concentration and/or memory disorder
  - 2.1% I identify as autistic / a person with autism

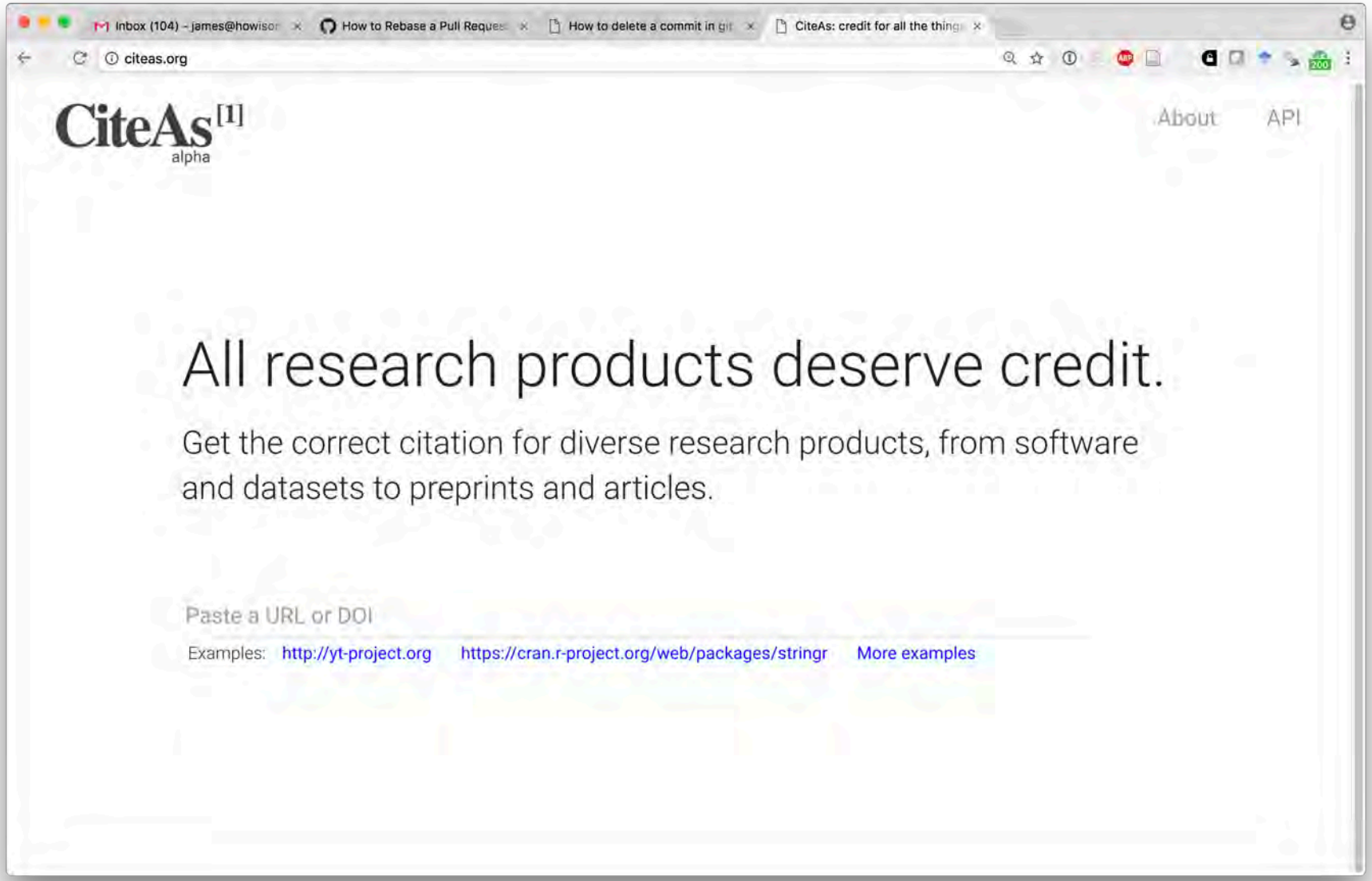
More at https://insights.stackoverflow.com/survey/2018/#demographics

gina@numfocus.org
@ginahelfrich

CiteAs.org

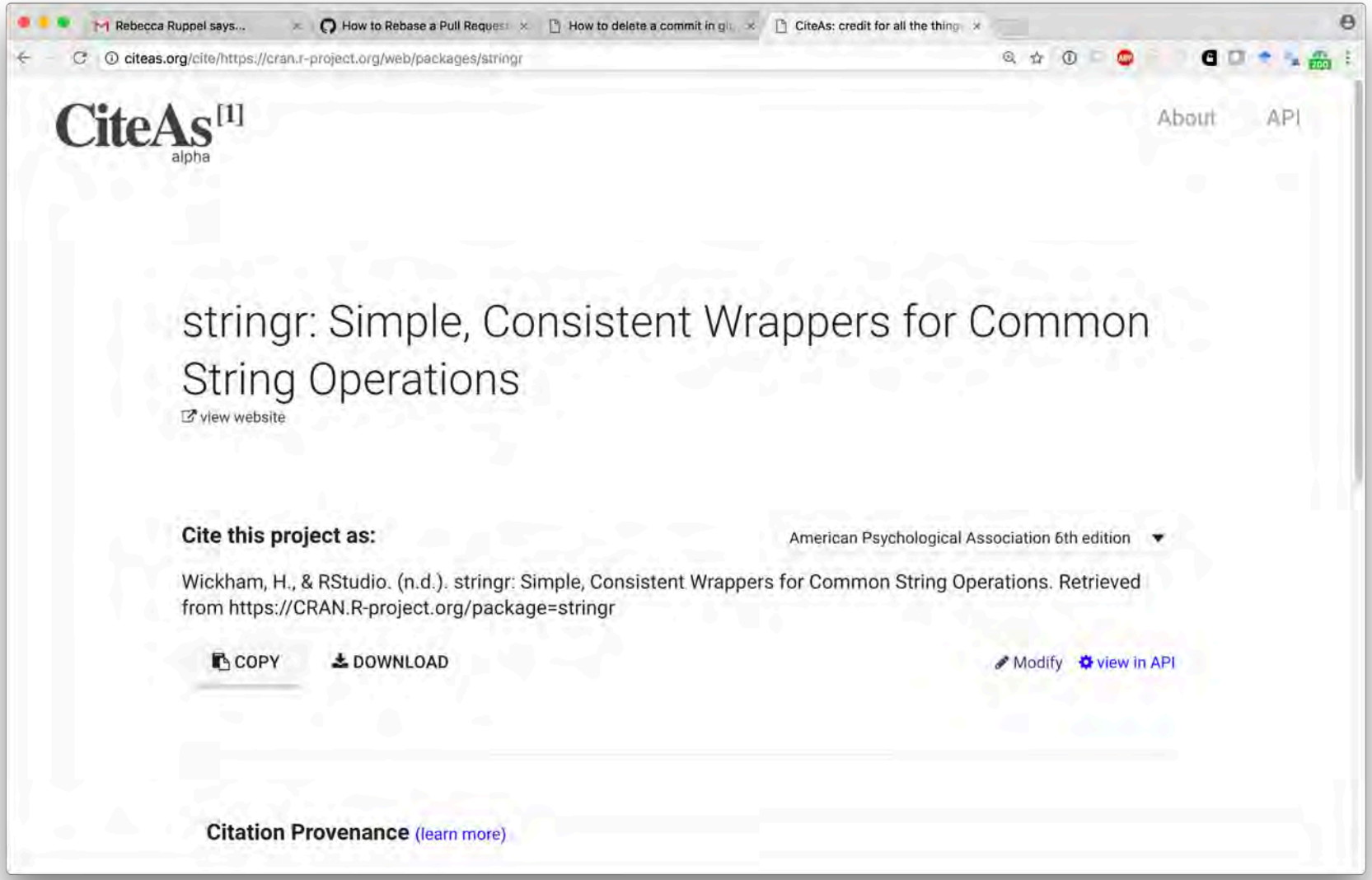
James Howison (UT Austin Information School)
Jason Priem and Heather Piwowar (ImpactStory.org)

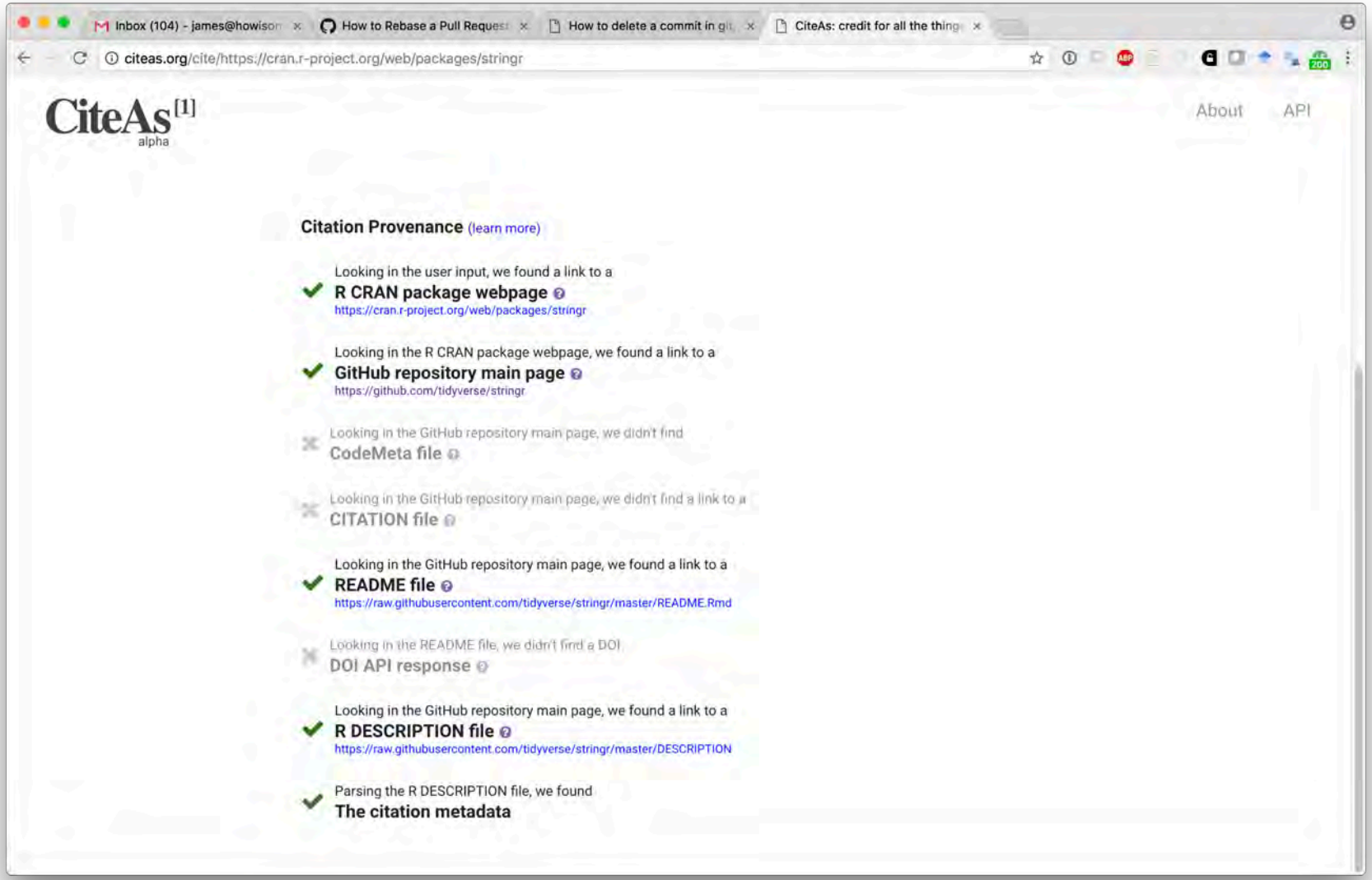
Funded by Sloan Foundation “Digital Science” via Josh Greenburg.


Inbox (104) - james@howison
How to Rebase a Pull Reques
How to delete a commit in git
CiteAs: credit for all the thing
citeas.org
CiteAs[1]
alpha
About
API
All research products deserve credit.
Get the correct citation for diverse research products, from software and datasets to preprints and articles.
Paste a URL or DOI
Examples: http://yt-project.org https://cran.r-project.org/web/packages/stringr More examples

CiteAs.org and Software ImpactStory

- Change science by helping those build software to make their case for impact.
- Map from software names, urls, to requested citations.
- Encourage projects to make clearer requests for citations
- Complement efforts for new standards, guidelines (e.g., FORCE11 principles).


Rebecca Ruppel says...
How to Rebase a Pull Reques
How to delete a commit in gi
CiteAs: credit for all the thing
citeas.org/cite/https://cran.r-project.org/web/packages/stringr
CiteAs[1]
alpha
About
API
stringr: Simple, Consistent Wrappers for Common String Operations
view website
Cite this project as:
American Psychological Association 6th edition
Wickham, H., & RStudio. (n.d.). stringr: Simple, Consistent Wrappers for Common String Operations. Retrieved from https://CRAN.R-project.org/package=stringr
COPY
DOWNLOAD
Modify
view in API
Citation Provenance (learn more)


Inbox (104) - james@howison
How to Rebase a Pull Reques
How to delete a commit in git
CiteAs: credit for all the thing
citeas.org/cite/https://cran.r-project.org/web/packages/stringr
CiteAs[1]
alpha
About
API
Citation Provenance (learn more)
Looking in the user input, we found a link to a
R CRAN package webpage
https://cran.r-project.org/web/packages/stringr
Looking in the R CRAN package webpage, we found a link to a
GitHub repository main page
https://github.com/tidyverse/stringr
Looking in the GitHub repository main page, we didn't find
CodeMeta file
Looking in the GitHub repository main page, we didn't find a link to a
CITATION file
Looking in the GitHub repository main page, we found a link to a
README file
https://raw.githubusercontent.com/tidyverse/stringr/master/README.Rmd
Looking in the README file, we didn't find a DOI
DOI API response
Looking in the GitHub repository main page, we found a link to a
R DESCRIPTION file
https://raw.githubusercontent.com/tidyverse/stringr/master/DESCRIPTION
Parsing the R DESCRIPTION file, we found
The citation metadata

A labeled dataset for finding software mentions

- Content analysis of publications to identify software “mentions”
- ~30 student coders read over 2,800 papers (so far): biomed, econ, astro.

  https://github.com/howisonlab/softcite-dataset
- Going to conduct machine learning to “distant read” the literature (and publish dataset for others to machine learn).
- Add Software Impact (in the literature) to ImpactStory.org

USSRI challenge

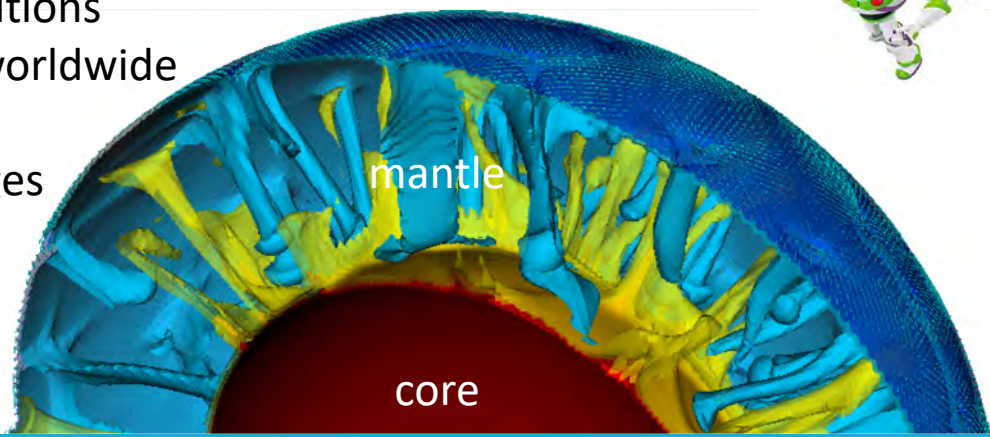
- Challenge:
  - What citations should projects request?
  - How should they make these requests?
  - How ought those building software make their case for contribution?
- Potential Solution:
  - CiteAs.org
- Help from URSSI?
  - Encourage formation of norms about citation requests
  - Overcome reticence to ask for citations among projects.

CIG

Advancing solid earth science through computational modeling

The **Computational Infrastructure for Geodynamics** is a community-driven organization that advances Earth science by **developing and disseminating software** for geophysics and related fields.

- 75+ member institutions
- 900+ participants worldwide
- 20 countries
- 33 Software packages
- Used worldwide

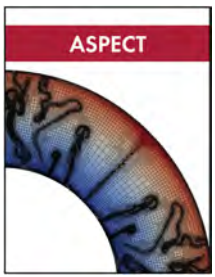


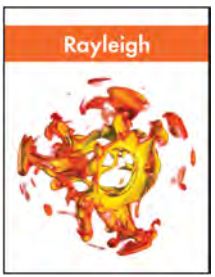
Software Development

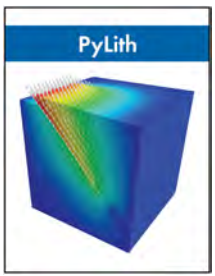
Successful CIG codes mix community and CIG development and support

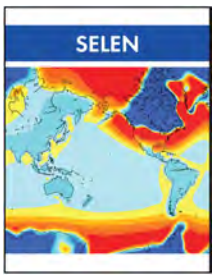
Entirely **CIG** funded

Community Contributed









deal.ii ASH PETSc

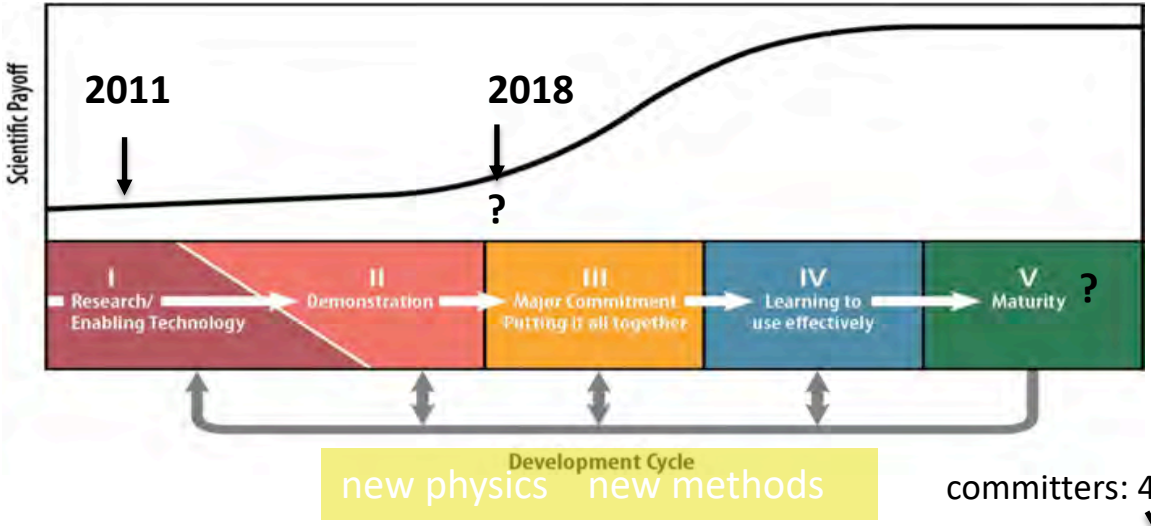
Building and sustaining teams includes
software engineering, domain expertise
and
social engineering

CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS

NSF

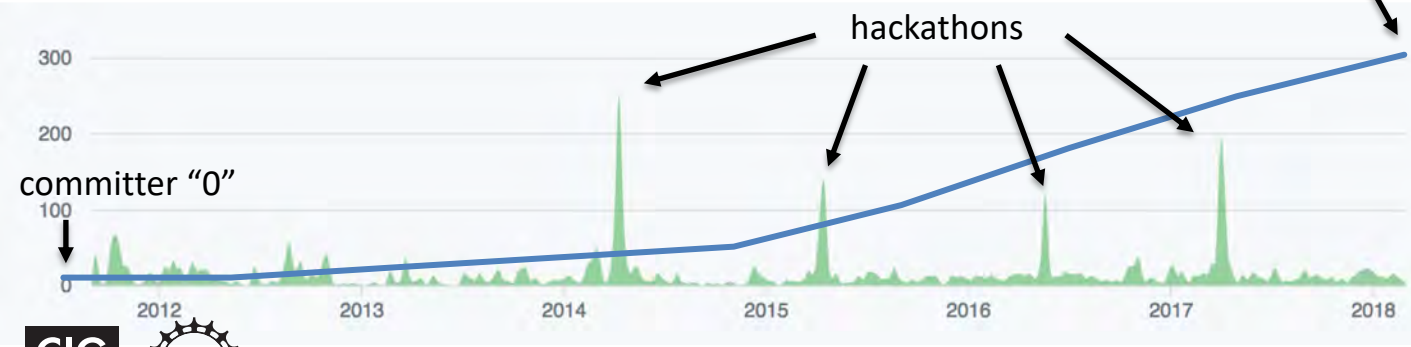
UCDAVIS UNIVERSITY OF CALIFORNIA

Challenge: Sustaining Communities



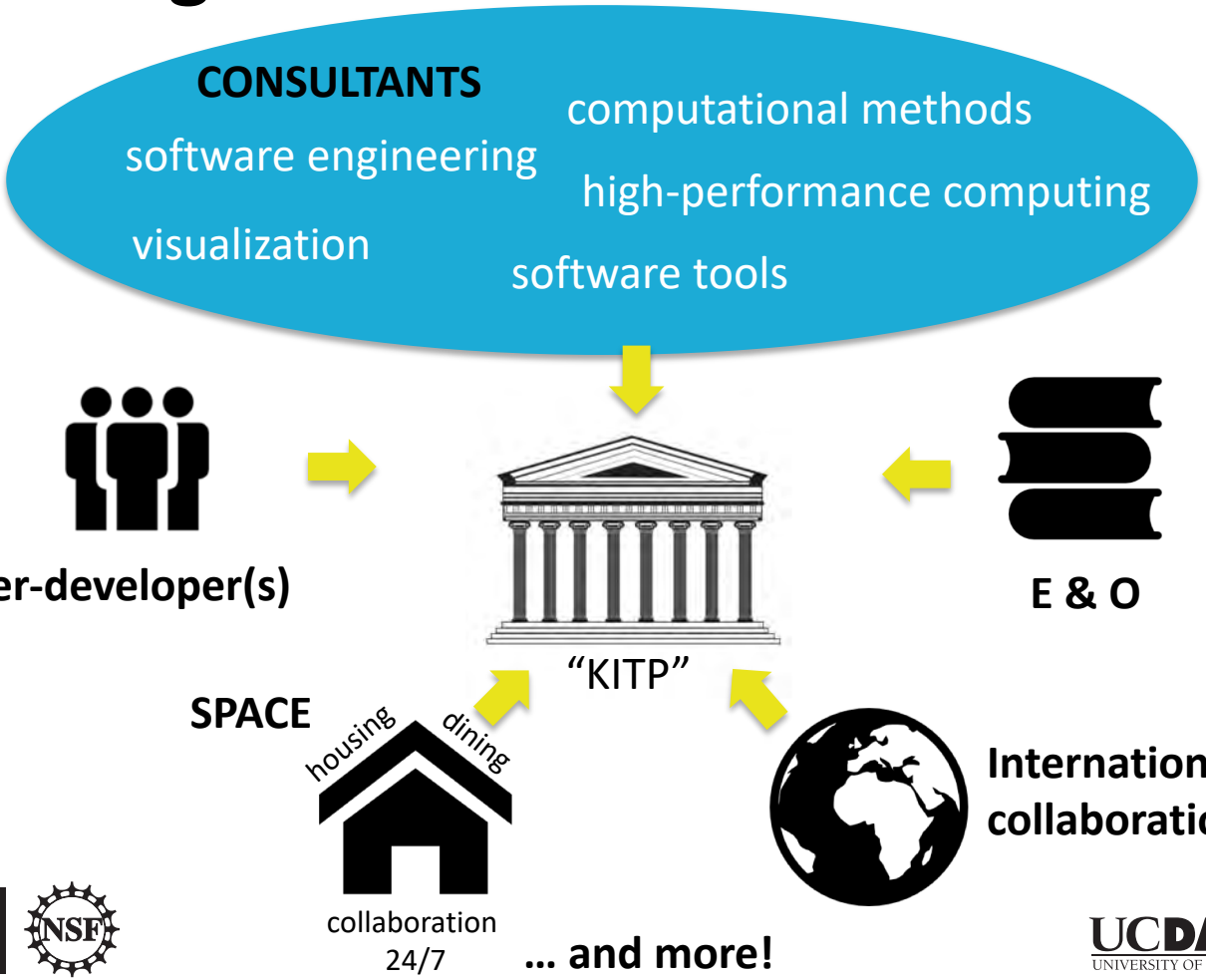
new physics new methods

committers: 44



UC DAVIS
UNIVERSITY OF CALIFORNIA

Creating the Nexus



CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS

NSF

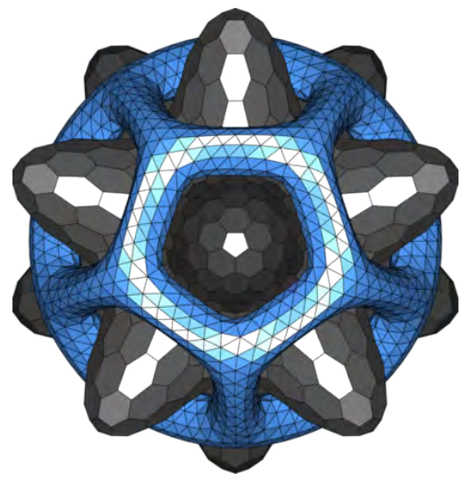
UCDAVIS UNIVERSITY OF CALIFORNIA

The Journal of Open Source Software

Arfon M. Smith, Kyle E. Niemeyer, Daniel S. Katz, Lorena A. Barba, George Githinji, Melissa Gymrek, Kathryn D. Huff, Christopher R. Madan, Abigail Cabunoc Mayes, Kevin M. Moerman, Pjotr Prins, Karthik Ram, Ariel Rokem, Tracy K. Teal, Roman Valls Guimera, and Jacob T. Vanderplas

https://joss.theoj.org

STScI | SPACE TELESCOPE SCIENCE INSTITUTE



A **developer friendly** journal for research software packages

> *A formal peer review process that is designed to **improve the quality of the software submitted.***

> *If your software is already well documented then paper preparation should take **no more than an hour.***

The Journal of Open Source S ×

Arfon

joss.theoj.org/papers/new

The Journal of Open Source Software

Editors Submit Papers About **Arfon Smith** · Sign out

Submit software for review

Before you submit

Please make sure you've read the submission instructions before submitting. In particular please make sure there is a `paper.md` present in your repository that is structured like this. We promise this will make things go much more quickly during the review process

Title

What's the title of this paper?

Repository address

What's the URL of your software?

Software version

e.g. v1.0.0

Suggested editor. View editors here »

Suggested editor

Description

Please give short (1-2 line) description of your software.

☐ **I certify that I am submitting software for which I am a primary author**

☐ **I confirm that I read and will adhere to the JOSS code of conduct**

Submit paper

Issues · openjournals/joss-revi ×

Arfon

GitHub, Inc. [US] | https://github.com/openjournals/joss-reviews/issues

This repository Search Pull requests Issues Marketplace Explore

openjournals / joss-reviews

Watch 30 Star 93 Fork 1

Code Issues 64 Pull requests 0 Projects 0 Insights Settings

Filters is:issue is:open Labels Milestones New issue

64 Open ✓ 449 Closed Author Labels Projects Milestones Assignee Sort

[PRE REVIEW]: pyneqsys: Solve symbolically defined systems of non-linear equations numerically Jupyter Notebook Python TeX pre-review
#529 opened 19 hours ago by whedon — 4

[REVIEW]: grapherator: A Modular Multi-Step Graph Generator review
#528 opened 2 days ago by whedon 0 of 18 — 4

[REVIEW]: reper - Genome-wide identification, classification and quantification of repetitive elements without an assembled genome review
#527 opened 2 days ago by whedon 0 of 18 — 3

[PRE REVIEW]: PyDMD: Python Dynamic Mode Decomposition Python Shell TeX pre-review
#526 opened 2 days ago by whedon — 19

[PRE REVIEW]: G^{3}M-f a global gradient-based groundwater modelling framwork pre-review
#525 opened 2 days ago by whedon — 5

[REVIEW]: MixEst: An Estimation Toolbox for Mixture Models review
#524 opened 3 days ago by whedon 0 of 18 — 3

[REVIEW]: ivporbit:An R package to estimate the instrumental variables probit model review
#523 opened 7 days ago by whedon 0 of 18 — 3

[REVIEW]: The Experiment Factory: Reproducible Experiment Containers review
#521 opened 19 days ago by whedon 18 of 18 — 3

[REVIEW]: arrgh: a Go interface to the OpenCPU R server system review — 14

[REVIEW]: Category Encoders: ×

Arfon

GitHub, Inc. [US] | https://github.com/openjournals/joss-reviews/issues/501

This repository Search Pull requests Issues Marketplace Explore

openjournals / joss-reviews

Watch 30 Star 93 Fork 1

Code Issues 64 Pull requests 0 Projects 0 Insights Settings

[REVIEW]: Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data #501

Edit New issue

Open **whedon** opened this issue 29 days ago · 10 comments

whedon commented 29 days ago • edited by desilinguist Owner

Submitting author: @wdm0006 (William McGinnis)
Repository: https://github.com/scikit-learn-contrib/categorical-encoding
Version: v1.2.5
Editor: @jakevdp
Reviewer: @desilinguist
Archive: Pending

Status

JOSS Under Review

Status badge code:

```
HTML: <a href="https://joss.theoj.org/papers/d57818316816a19a80112892c3d12ed7"><img src="ht
Markdown: [![status](https://joss.theoj.org/papers/d57818316816a19a80112892c3d12ed7/status.
```

Reviewers and authors:

Please avoid lengthy details of difficulties in the review thread. Instead, please create a new issue in

Assignees

jakevdp

Labels

review

Projects

None yet

Milestone

No milestone

Notifications

Unsubscribe

You're receiving notifications because you commented.

5 participants

Gala: A Python package for galactic dynamics

Adrian M. Price-Whelan

Article details

- View review »
- Download paper »
- Software repository »
- Software archive »

Submitted: 13 August 2017
Accepted: 08 October 2017

Cite as:

Price-Whelan, (2017), Gala: A Python package for galactic dynamics, Journal of Open Source Software, 2(18), 388, doi:10.21105/joss.00388

Status badge

JOSS 10.21105/joss.00388

JOSS
The Journal of Open Source Software

Gala: A Python package for galactic dynamics

Adrian M. Price-Whelan[1]

1 Lyman Spitzer, Jr. Fellow, Princeton University

DOI: 10.21105/joss.00388

Software

- Review
- Repository
- Archive

Summary

The forces on stars, galaxies, and dark matter under external gravitational fields lead to the dynamical evolution of structures in the universe. The orbits of these bodies are therefore key to understanding the formation, history, and future state of galaxies. The field of "galactic dynamics," which aims to model the gravitating components of galaxies to study their structure and evolution, is now well-established, commonly taught, and frequently used in astronomy. Aside from toy problems and demonstrations, the majority of problems require efficient numerical tools, many of which require the same base code (e.g., for performing numerical orbit integration).

`Gala` is an Astropy-affiliated Python package for galactic dynamics. Python enables wrapping low-level languages (e.g., C) for speed without losing flexibility or ease-of-use in the user-interface. The API for `Gala` was designed to provide a class-based and user-friendly interface to fast (C or Cython-optimized) implementations of common operations such as gravitational potential and force evaluation, orbit integration, dynamical transformations, and chaos indicators for nonlinear dynamics. `Gala` also relies heavily on and interfaces well with the implementations of physical units and astronomical coordinate systems in the `Astropy` package (Astropy Collaboration et al. 2013) (`astropy.units` and `astropy.coordinates`).

`Gala` was designed to be used by both astronomical researchers and by students in courses on gravitational dynamics or astronomy. It has already been used in a number of scientific publications (Pearson, Price-Whelan, and Johnston 2017) and has also been used in graduate courses on Galactic dynamics to, e.g., provide interactive visualizations of textbook material (Binney and Tremaine 2008). The combination of speed, design, and


cs-147.pdf
Arfon
PeerJ, Inc. [US] | https://peerj.com/articles/cs-147.pdf
PeerJ Computer Science

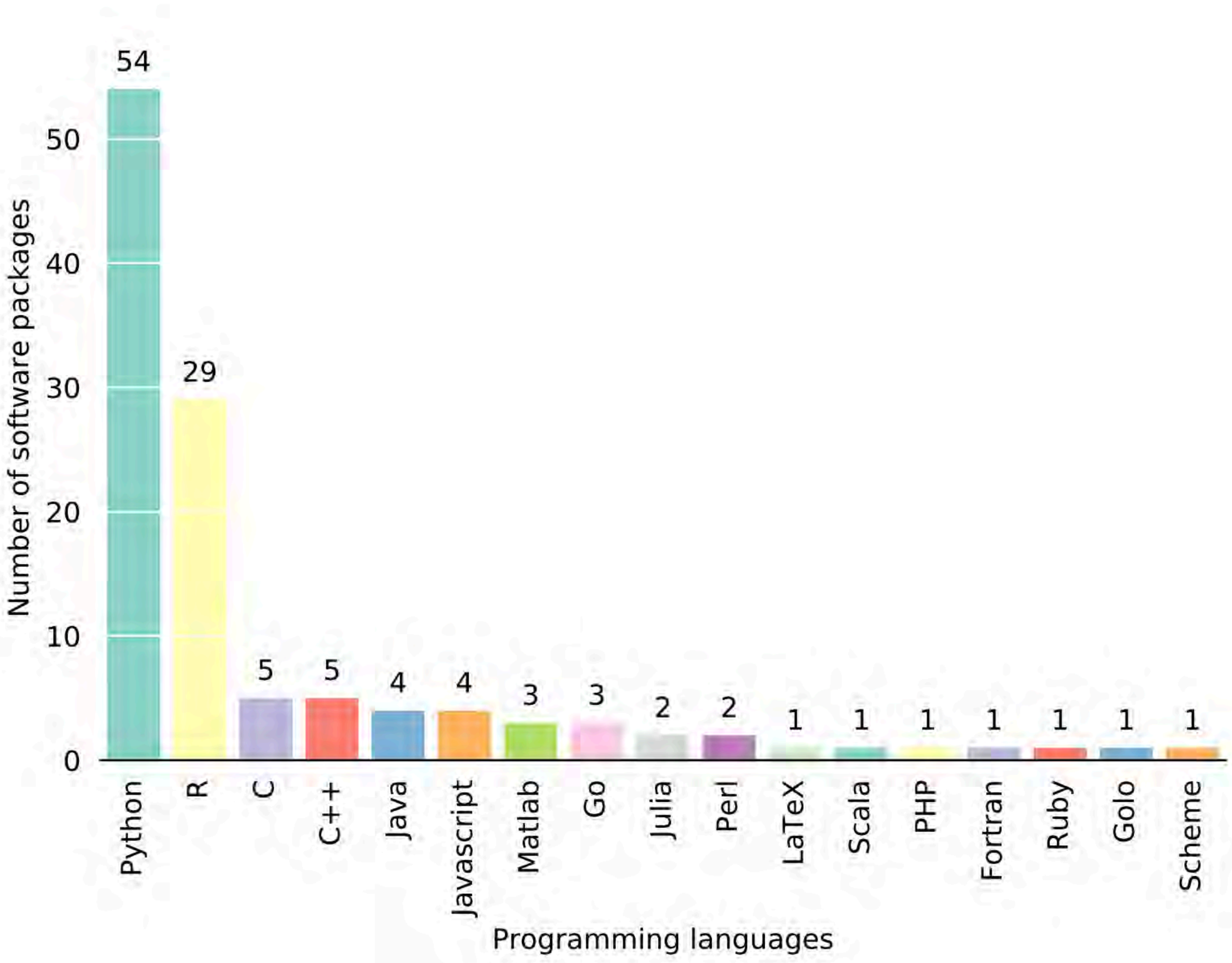
Journal of Open Source Software (JOSS): design and first-year review

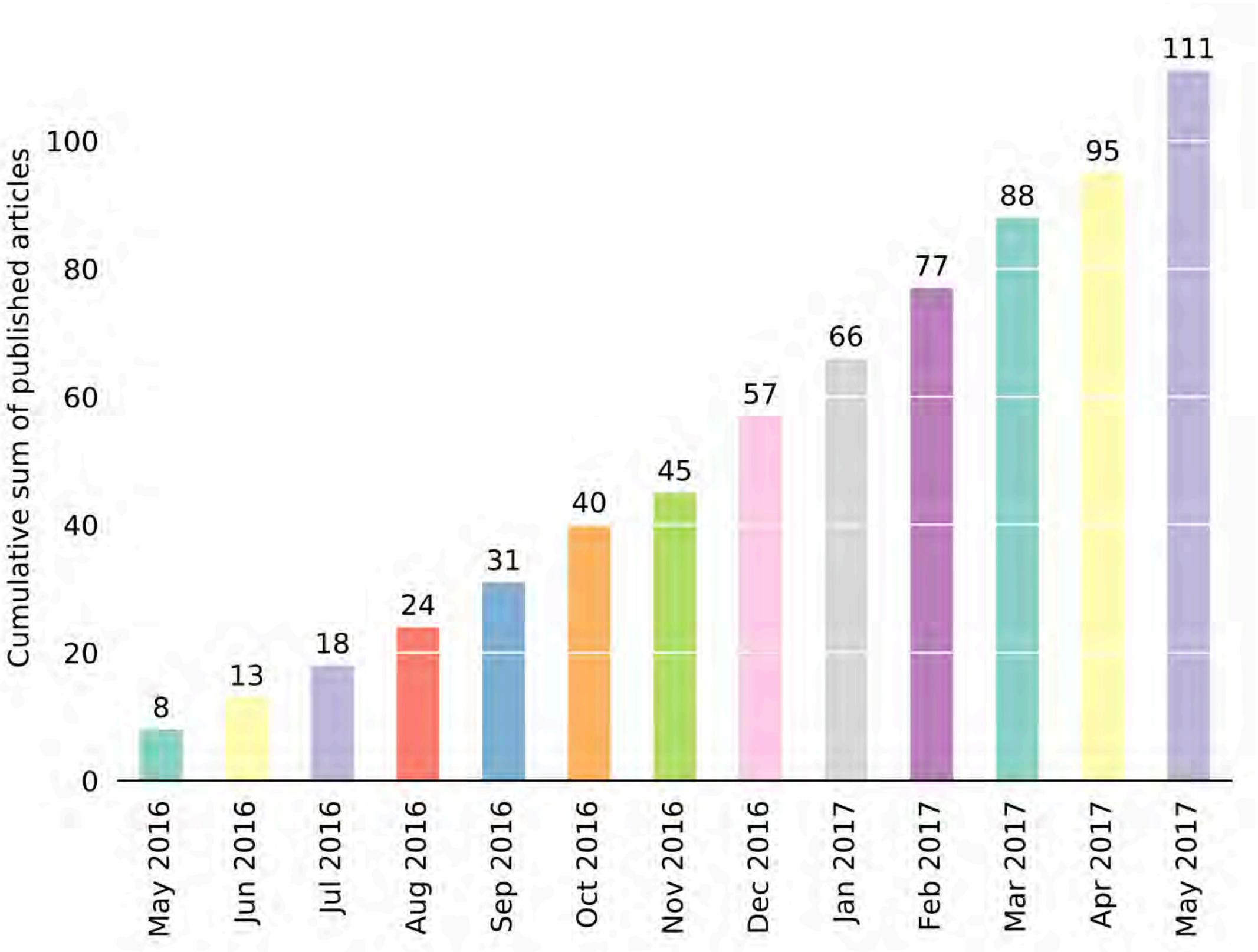
Arfon M. Smith[1], Kyle E. Niemeyer[2], Daniel S. Katz[3], Lorena A. Barba[4], George Githinji[5], Melissa Gymrek[6], Kathryn D. Huff[7], Christopher R. Madan[8], Abigail Cabunoc Mayes[9], Kevin M. Moerman[10,11], Pjotr Prins[12,13], Karthik Ram[14], Ariel Rokem[15], Tracy K. Teal[16], Roman Valls Guimera[17] and Jacob T. Vanderplas[15]

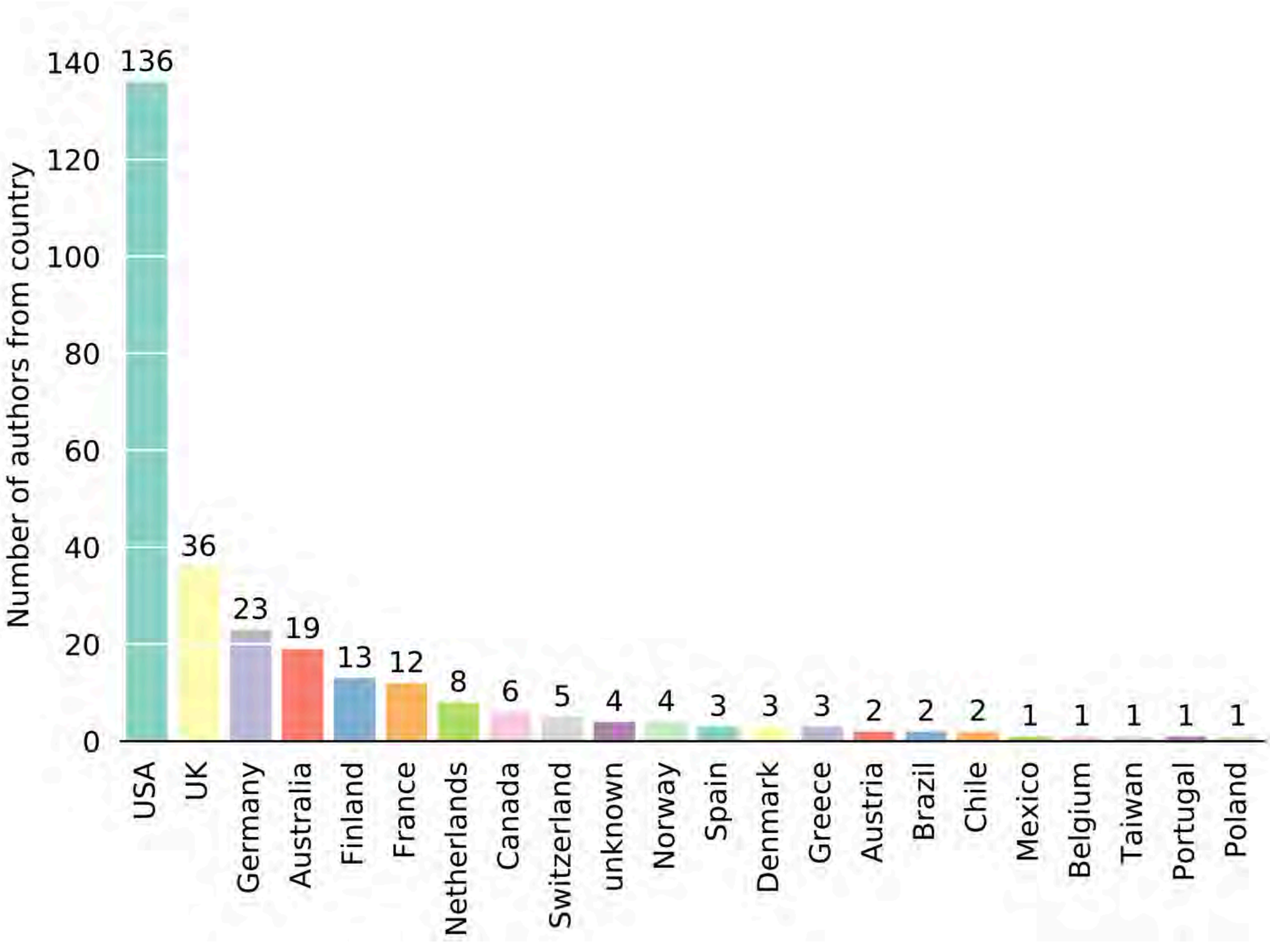
[1] Data Science Mission Office, Space Telescope Science Institute, Baltimore, MD, United States of America
[2] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, OR, United States of America
[3] National Center for Supercomputing Applications & Department of Computer Science & Department of Electrical and Computer Engineering & School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, United States of America
[4] Department of Mechanical & Aerospace Engineering, The George Washington University, Washington, D.C., United States of America
[5] KEMRI—Wellcome Trust Research Programme, Kilifi, Kenya
[6] Departments of Medicine & Computer Science and Engineering, University of California, San Diego, La Jolla, CA, United States of America
[7] Department of Nuclear, Plasma, and Radiological Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States of America
[8] School of Psychology, University of Nottingham, Nottingham, United Kingdom
[9] Mozilla Foundation, Toronto, Ontario, Canada
[10] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[11] Trinity Centre for Bioengineering, Trinity College, The University of Dublin, Dublin, Ireland
[12] University of Tennessee Health Science Center, Memphis, TN, United States of America
[13] University Medical Centre Utrecht, Utrecht, The Netherlands
[14] Berkeley Institute for Data Science, University of California, Berkeley, CA, United States of America
[15] eScience Institute, University of Washington, Seattle, WA, United States of America
[16] Data Carpentry, Davis, CA, United States of America
[17] University of Melbourne Centre for Cancer Research, University of Melbourne, Melbourne, Australia

Submitted 6 October 2017

https://peerj.com/articles/cs-147/


Number of software packages
0
10
20
30
40
50
54
29
5
5
4
4
3
3
2
2
1
1
1
1
1
1
1
Python
R
C
C++
Java
Javascript
Matlab
Go
Julia
Perl
LaTeX
Scala
PHP
Fortran
Ruby
Golo
Scheme
Programming languages


Cumulative sum of published articles
0
20
40
60
80
100
8
13
18
24
31
40
45
57
66
77
88
95
111
May 2016
Jun 2016
Jul 2016
Aug 2016
Sep 2016
Oct 2016
Nov 2016
Dec 2016
Jan 2017
Feb 2017
Mar 2017
Apr 2017
May 2017


Number of authors from country
140
120
100
80
60
40
20
0
136
36
23
19
13
12
8
6
5
4
4
3
3
3
2
2
2
1
1
1
1
1
USA
UK
Germany
Australia
Finland
France
Netherlands
Canada
Switzerland
unknown
Norway
Spain
Denmark
Greece
Austria
Brazil
Chile
Mexico
Belgium
Taiwan
Portugal
Poland

How to get career credit for research software in two easy steps...

Make something citable

1

Make it easy to be cited

2

“Make something citable”

> Upload your code to GitHub (or other public repository)
*> Create a DOI for repository via Zenodo/fig**share** etc.*
> Write a paper about your software

JOSS LOWERS

THE COST OF THIS

ISSN: 2352-7110

SoftwareX

› Open Access

Editors-in-Chief: Dr. Kate Keahey, Dr. Randall Sobie, Dr. David Wallom

› View Editorial Board

› Elsevier Physics homepage

SoftwareX aims to acknowledge the impact of software on today's research practice, and on new scientific discoveries in almost all research domains. *SoftwareX* also aims to stress the importance of the software developers who are, in part, responsible for this impact.

To this end, *SoftwareX* aims to support...

Read more

Most Downloaded | Recent Articles | Most Cited

RCrawler: An R package for parallel web crawling and scraping Salim Khalil | Mohamed Fakir

GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers Mark James Abraham | Teemu Murtola | ...

pysimm: A python package for simulation of molecular systems Michael E. Fortunato | Coray M. Colina

Journal of Open Research Soft ×

Arfon

Secure | https://openresearchsoftware.metajnl.com

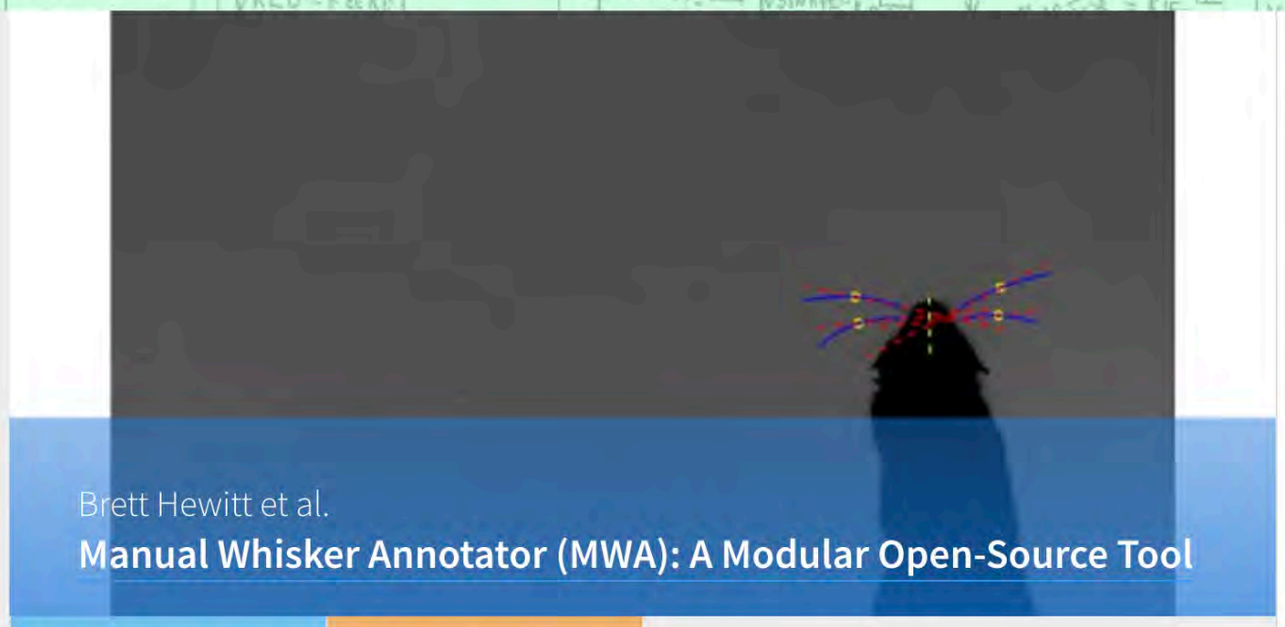
Home About Contact Content Research Integrity Search... Log in Register

Software Sustainability Institute

Journal of open research software

Start Submission Become a Reviewer

ubiquity press open access



Brett Hewitt et al.

Manual Whisker Annotator (MWA): A Modular Open-Source Tool

Follow on Twitter Follow Via RSS

About this Journal

The *Journal of Open Research Software* (JORS) features peer reviewed Software Metapapers describing research software with high reuse potential. We are working with a number of specialist and institutional repositories to ensure that the associated software is professionally archived, preserved, and is openly available. Equally importantly, the software and the papers will be citable, and reuse will be tracked.

JORS also publishes full-length research papers that cover different aspects of creating, maintaining and evaluating open source research software. The aim of the section is to promote the dissemination of best practice and experience related to the development and maintenance of reusable, sustainable research software.

LATEST ARTICLES POPULAR ARTICLES

RWebData: A High-Level Interface to the Programmable Web

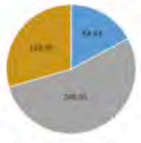
Matter — 21 Feb 2018

Share:

Fourth Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE4)

Katz et al. — 15 Feb 2018

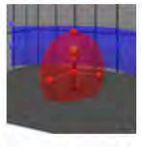
Share:



Tethys – A Python Package for Spatial and Temporal Downscaling of Global Water Withdrawals

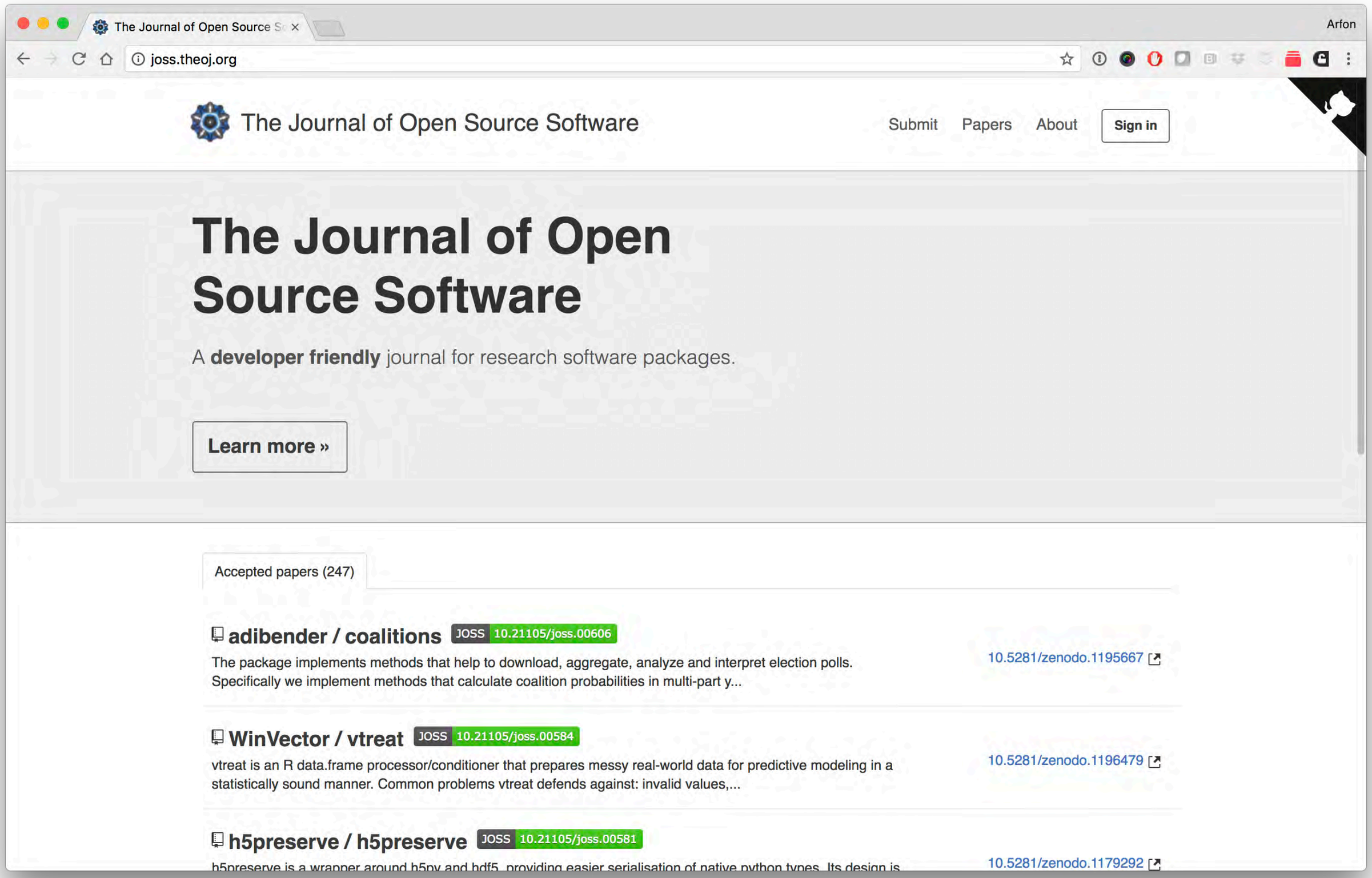
Li et al. — 09 Feb 2018

Share:



webMUSHRA — A Comprehensive Framework for Web-based Listening Tests

Schoeffler et al. — 05 Feb 2018


The Journal of Open Source Software
Arfon
joss.theoj.org
The Journal of Open Source Software
Submit
Papers
About
Sign in
The Journal of Open Source Software
A developer friendly journal for research software packages.
Learn more »
Accepted papers (247)
adibender / coalitions
JOSS 10.21105/joss.00606
The package implements methods that help to download, aggregate, analyze and interpret election polls. Specifically we implement methods that calculate coalition probabilities in multi-part y...
10.5281/zenodo.1195667
WinVector / vtreat
JOSS 10.21105/joss.00584
vtreat is an R data.frame processor/conditioner that prepares messy real-world data for predictive modeling in a statistically sound manner. Common problems vtreat defends against: invalid values,...
10.5281/zenodo.1196479
h5preserve / h5preserve
JOSS 10.21105/joss.00581
10.5281/zenodo.1179292

“Make it easy to be cited”

Documentation

Read the docs at emcee.readthedocs.io.

Attribution

Please cite Foreman-Mackey, Hogg, Lang & Goodman (2012) if you find this code useful in your research and add your paper to the testimonials list. The BibTeX entry for the paper is:

```
@article{emcee,
   author = {{Foreman-Mackey}, D. and {Hogg}, D.~W. and {Lang}, D. and {Goodman}, J.},
    title = {emcee: The MCMC Hammer},
  journal = {PASP},
     year = 2013,
   volume = 125,
    pages = {306-312},
   eprint = {1202.3665},
      doi = {10.1086/670067}
}
```

License

Encouraging citation of software – introducing CITATION files

By Robin Wilson, Fellow and postgraduate at the University of Southampton.

Put a plaintext file named CITATION in the root directory of your code, and put information in it about how to cite your software. Go on, do it now: it'll only take two minutes!

Software is very important in science – but good software takes time and effort that could be used to do other work instead. I believe that it is important to do this work, but to make it worthwhile, people need to get credit for their work, and in academia that means citations. However, it is often very difficult to find out how to cite a piece of software – sometimes it is hidden away somewhere in the manual or on the web-page, but often it requires sending an email to the author asking them how they want it cited. The effort that this requires means that many people don't bother to cite the software they use, and thus the authors don't get the credit that they need. We need to change this, so that software – which underlies a huge amount of important scientific work – gets the recognition it deserves.

As with many things relating to software sustainability in science, the R project does this very well: if you want to find out how to cite the R software itself you simply run the command:

```
citation()
```

If you want to find out how to cite a package you simply run:

```
citation(PROJECTNAME)
```

For example:

Tags

Robin Wilson
Community

Software and Research Blog

14-March-2018 - **Oxford Reproducibility Lectures** - By Ana Todorović, Oxford University In September 2017, we started the…

09-March-2018 - **What do we know about RSEs? Results from our international surveys** - By Olivier Philippe, Policy Researcher. Last year, the Software…

09-March-2018 - **What is lined up for Collaborations Workshop 2018?** - By Raniere Silva, Community Officer, Software Sustainability Institute.…

07-March-2018 - **Introducing NL-RSE: Bootstrapping the community of Research Software Engineers in the Netherlands** - By the Netherlands eScience Center This post was originally published at…

06-March-2018 - **Global Open Science Hardware Roadmap: Launching a revolution in access to scientific tools** - By Jenny Molloy, University of Cambridge. Introduction by Raniere Silva…

Latest News

08-March-2018 - **Google Summer of Code 2018:**

Search · filename:CITATION

Arfon

GitHub, Inc. [US] | https://github.com/search?utf8=✓&q=filename%3ACITATION&type=Code

filename:CITATION

Pull requests Issues Marketplace Explore

Repositories

Code 72K

Commits

Issues

Topics 312K

Wikis

Users

72,476 code results

Sort: Best match

SurajGupta/r-source – CITATION
Last indexed on 14 Sep 2016
src/library/base/inst/**CITATION**

stekhoven/missForest – CITATION
Last indexed on 15 Sep 2016
inst/**CITATION**

swcarpentry/2013-09-21-waterloo – CITATION
Last indexed on 16 Sep 2016
CITATION

swcarpentry/2014-01-18-ucb – CITATION
Last indexed on 16 Sep 2016
CITATION

rforge/pks – CITATION
Last indexed on 16 Sep 2016
pkg/inst/**CITATION**

rforge/rgnuplot – CITATION
Last indexed on 16 Sep 2016

Languages

Java	6,325
PHP	6,241
HTML	5,456
XML	2,879
Ruby	2,822
Smarty	2,635
Text	2,608
JSON	2,551
JavaScript	2,433
Python	2,299

Advanced search Cheat sheet

R

```
> citation('ggplot2')

To cite ggplot2 in publications, please use:

H. Wickham. ggplot2: elegant graphics for data analysis.
  Springer New York, 2009.

A BibTeX entry for LaTeX users is

@Book{,
    author = {Hadley Wickham},
    title = {ggplot2: elegant graphics for data analysis},
    publisher = {Springer New York},
    year = {2009},
    isbn = {978-0-387-98140-6},
    url = {http://had.co.nz/ggplot2/book},
  }
```

Python

```
> import astroML as aml

> aml.__citation__

@INPROCEEDINGS{astroML,
 author={{Vanderplas}, J.T. and {Connolly}, A.J.and {Ivezi{\'c}},
 {\v Z}. and {Gray}, A.},
 booktitle={Conference on Intelligent Data Understanding (CIDU)},
 title={Introduction to astroML: Machine learning for astrophysics},
 month={Oct.},
 pages={47 -54},
 doi={10.1109/CIDU.2012.6382200},
 year={2012}
}
```

“How URSSI could help”

*> Develop/document best practices for **receiving** career credit for software*
*> Develop/document best practices for **awarding** career credit for software*
> Strive to make JOSS unnecessary? (better fixes than just software papers)

Thanks

arfon@stsci.edu

Packaging a concrete recommendation for how URSSI can help

Berkeley Workshop

April 11, 2018

Todd Tannenbaum
Center for High Throughput Computing
Department of Computer Sciences
University of Wisconsin-Madison

CENTER FOR HIGH THROUGHPUT COMPUTING

HTCondor

University of Wisconsin Center for High Throughput Computing

HTCondor

- Open source distributed high throughput computing
- Schedule, provision, manage compute resources, containers, jobs, and workflows
- Primary objective: assist the scientific community with their high throughput computing needs
- Mature technology…

Mature… but actively developed

- Regular releases, both a stable (bug fixes only) and new features series
- Open source development model
- Evolve to meet the needs of the science community in an ever-changing computing landscape

	All Time	12 Month	30 Day
Commits:	39067	2349	141
Contributors:	152	21	10
Files Modified:	11588	1665	169
Lines Added:	12352208	444401	29395
Lines Removed	6810332	187595	7835

Source: https://www.openhub.net/p/condorproject

Software Development Activities

- Gather Requirements / Usability
- Design / Architect
- Implement
- Test
- Use
- Document
- Package and Distribute
- User Support / Engage

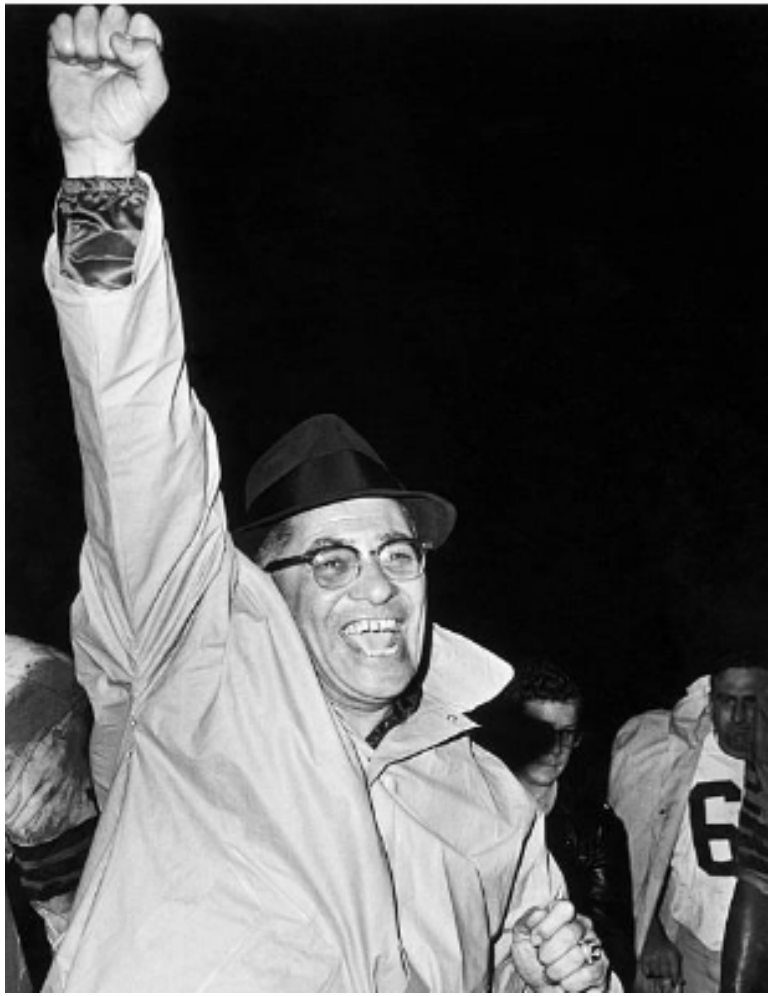
Packaging back in the day…

tar cvf condor_ver4.tar ./condor

cp condor_ver4.tar ~/public/ftp

Nowadays a lot more knowledge and effort required to package and distribute your software

"Is Not Enough" section of my talk



Tar is not enough… System Packaging

- **Several different formats**
  - RPM, DEB, SRC.RPM, SRC.DEB, Gentoo, …
  - (and beyond Linux: Chocolatey, MacPorts, MSI, …)
- Different packaging tool chains per format
- Different standards, requirements per distro
  - Packaging guides for Debian = 200+ pages!
- Package format **is not enough**, but **also distribution integration**
  - E.g. Systemd, SELinux, AppArmor, …

Package is not enough… distribution via repositories

- Repositories need to be
  - Secure: digitally signed, https, etc
  - One needed for each distribution major release
    - (each distribution has different dependency versions and names)
  - Potentially high bandwidth (CDN?)
- *What's wrong with official distro repo or EPEL?*

Package is not enough… distribution via repositories

- Repositories need to be
  - Secure: digitally signed, https, etc
  - One needed for each distribution major release
    - (each distribution has different dependency versions and names)
  - Potentially high bandwidth (CDN?)
- *What's wrong with official distro repo or EPEL?*
  - ***Metrics!!!! No download counts, 'phone home' disallowed***
  - Frequency of updates
  - Effort. rules / enforcement changes, emails, …

System package distribution is not enough…

- **Clouds and containers**
  - Amazon AMI
  - Docker hub image
  - Kubernetes POD / controller
  - Tomorrow? Rapidly evolving…

Even all this is not enough!

- **Language environment packaging**

  Example: HTCondor's Python API

  - Python Package Index (PyPI)
    - Easy for pure Python, much more involved if native code included
  - Anaconda

How can URSSI help?
Packaging Services for Science Software Community

- Have an institute service to package and distribute NSF funded software
  - Experts who know all the rules, packaging technologies, and set it up for me
  - Distribution servers (e.g. "NSF Science" YUM repos, Docker Hub, …)

Profit!

- Economies of scale
  - One person could package for hundreds of projects, consult on dozens of complex (SELinux) setups
- Concrete, actionable idea
- Natural separation of expertise
- Quantifiable cost savings benefit
- Software developers will love you
- Easier, flexible installation options = more users / ROI
- Potentially a significant weapon in battle for usage metrics, reproducibility
- "Free packaging" can serve as a carrot
  - Entry point to provide other institute services in the future

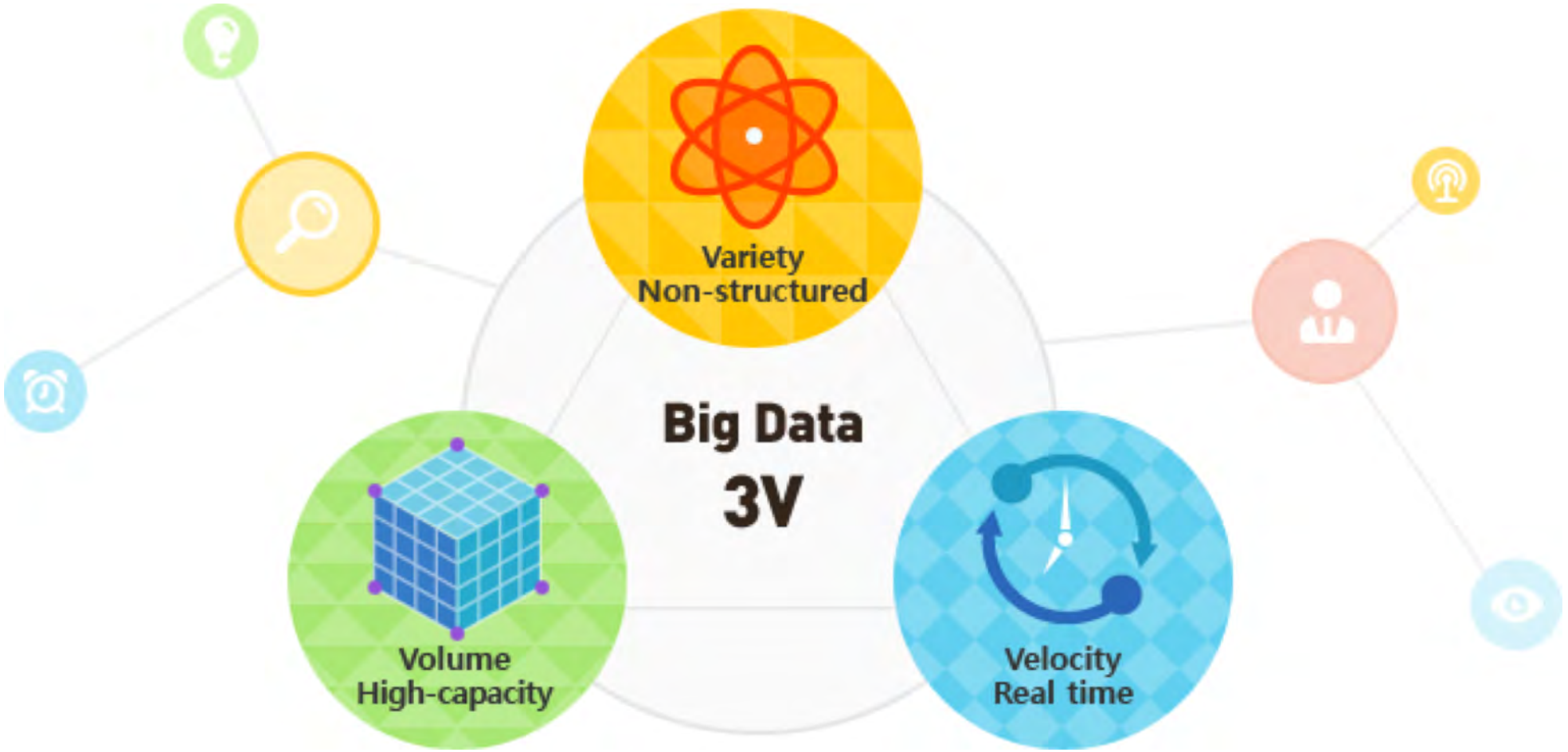
THE CARPENTRIES

@tracykteal | @thecarpentries

http://www.datacarpentry.org | http://www.software-carpentry.org

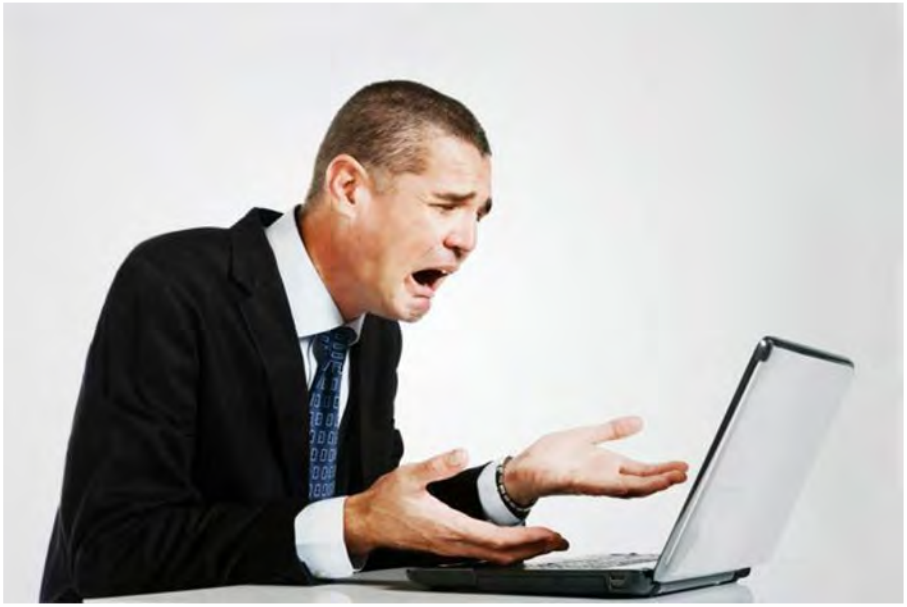
The life changing magic of data & software skills training

Our increasing capacity to collect data is changing research




"Big data" is for every researcher

Data pain

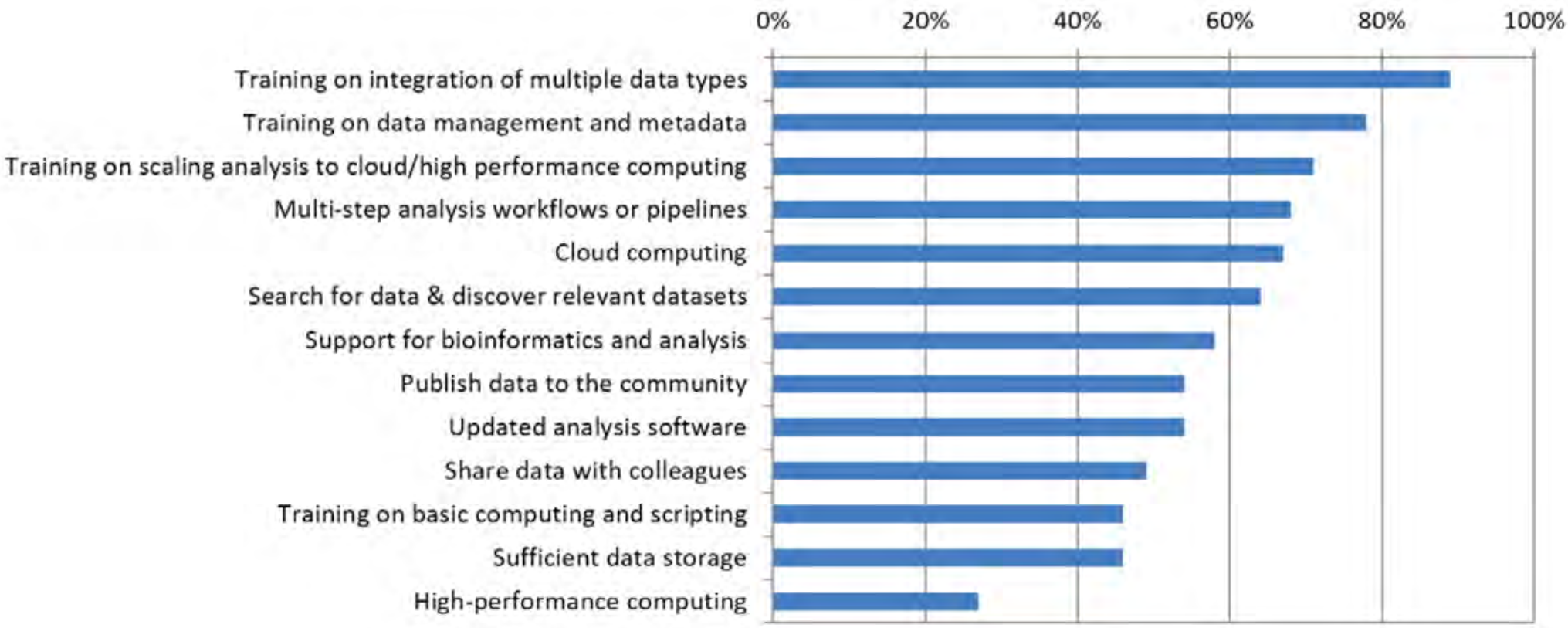


How do we scale data literacy along with data production?

Training is the missing piece

Most useful thing Bioinformatics Resource Australia could do is Offer Training

Current unmet needs



Barone L, Williams J and Micklos D. **Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators** (2017)

Where does this training come from?

software carpentry

THE CARPENTRIES

DATA CARPENTRY

Non-profit organization that develops curriculum, trains instructors and teaches workshops on the skills and perspectives to work effectively and reproducibly with software and data.

We teach

Data organization

Data ‘wrangling’

Data analysis & visualization:
R and Python

Software development practices: GitHub, shell

How do we do it?

Hands-on instruction

Collaboratively-developed Openly-licensed materials

DATA CARPENTRY

Short Introduction to Programming in Python

The Basics of Python

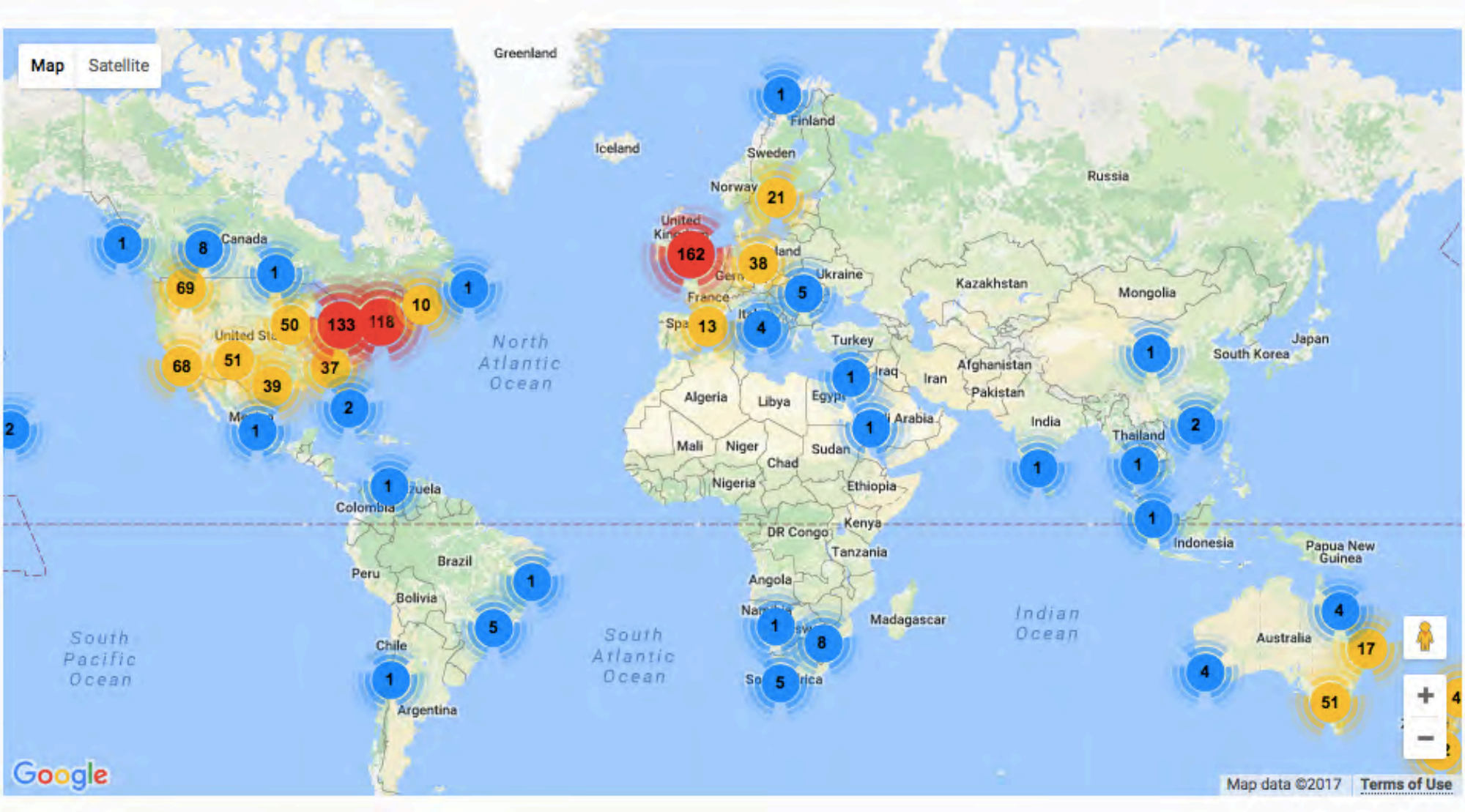
Python is a general purpose programming language, that supports rapid development of scripts and applications.

Python's main advantages:

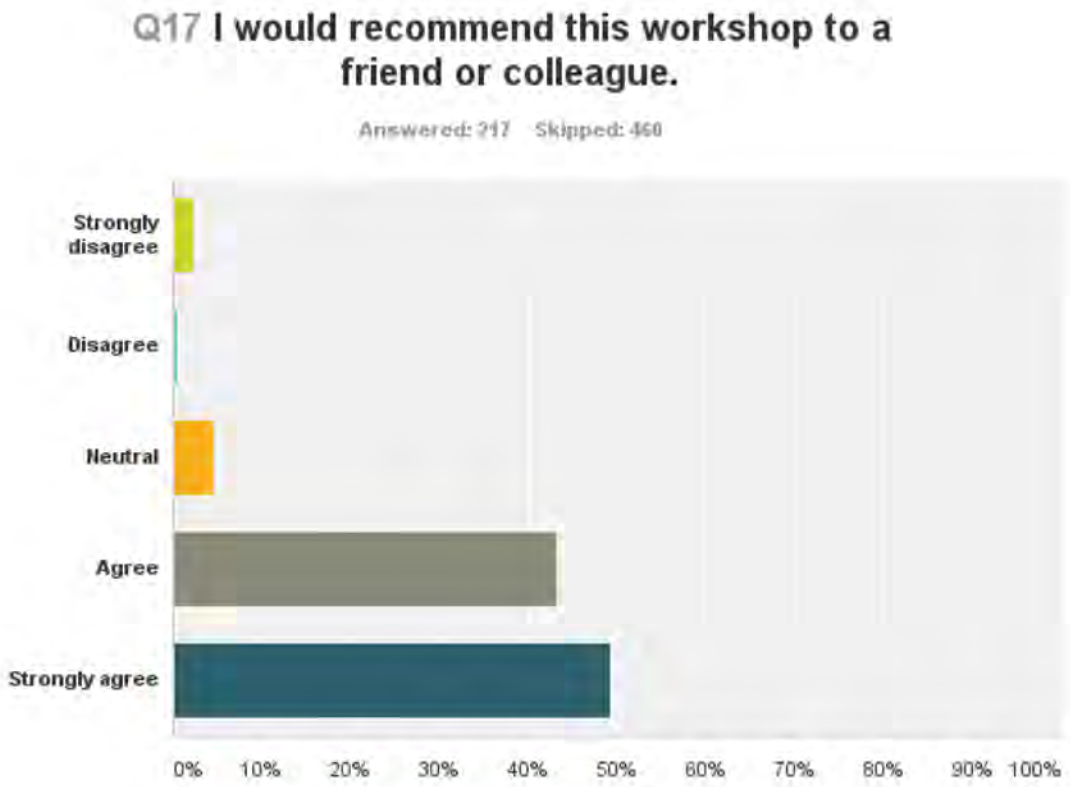
- Open Source software, supported by Python Software Foundation
- Available on all platforms
- "Batteries Included" philosophy - libraries for common tasks available in standard installation
- Supports multiple programming paradigms
- Very large community

Volunteer instructors

Impact

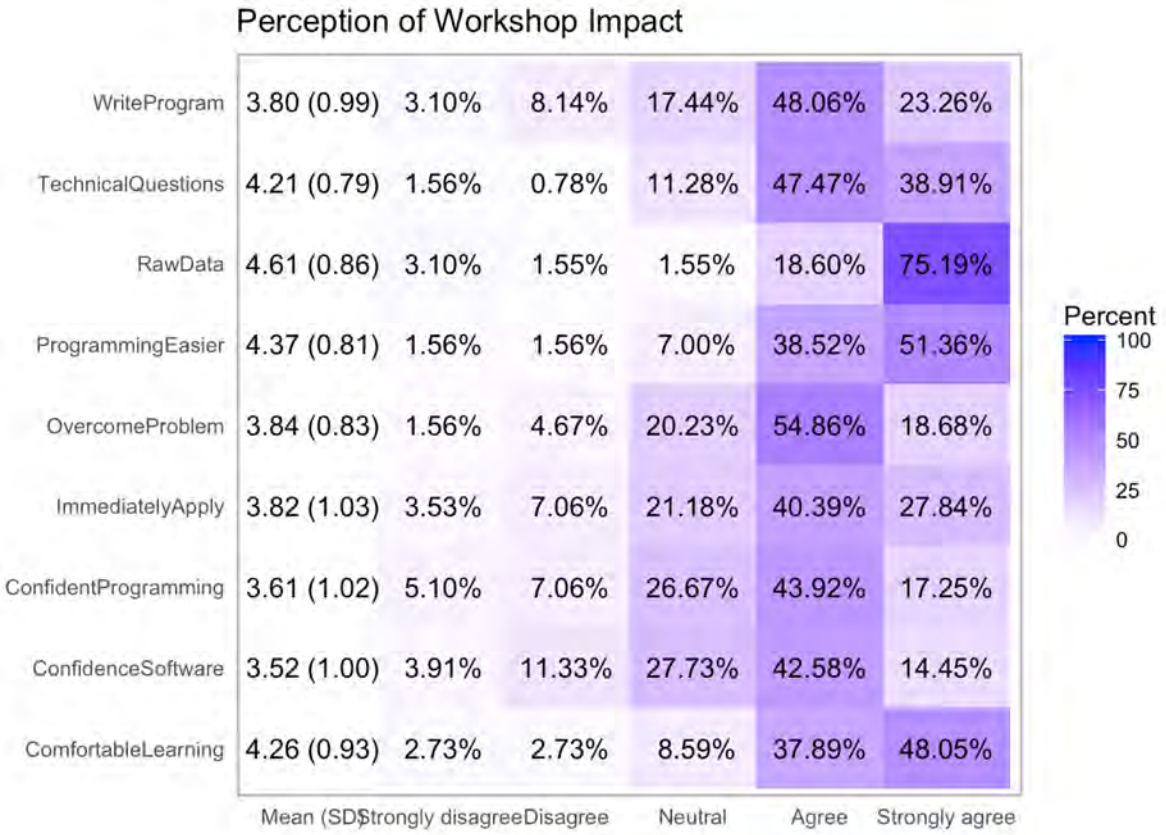

Map
Satellite
Greenland
Iceland
Finland
Sweden
Norway
Russia
Canada
United Kingdom
France
Ukraine
Kazakhstan
Mongolia
Japan
South Korea
Turkey
Iraq
Iran
Afghanistan
Pakistan
India
Thailand
Indonesia
Papua New Guinea
Australia
Algeria
Libya
Egypt
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Tanzania
Angola
Madagascar
Colombia
Peru
Brazil
Bolivia
Chile
Argentina
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
1
8
69
50
133
118
10
68
51
39
37
2
162
38
21
5
13
4
1
2
17
51
Google
Map data ©2017
Terms of Use

People like the workshops



Promoter Score	n	%
Detractor	14	5.511811
Passive	49	19.291339
Promoter	191	75.196850

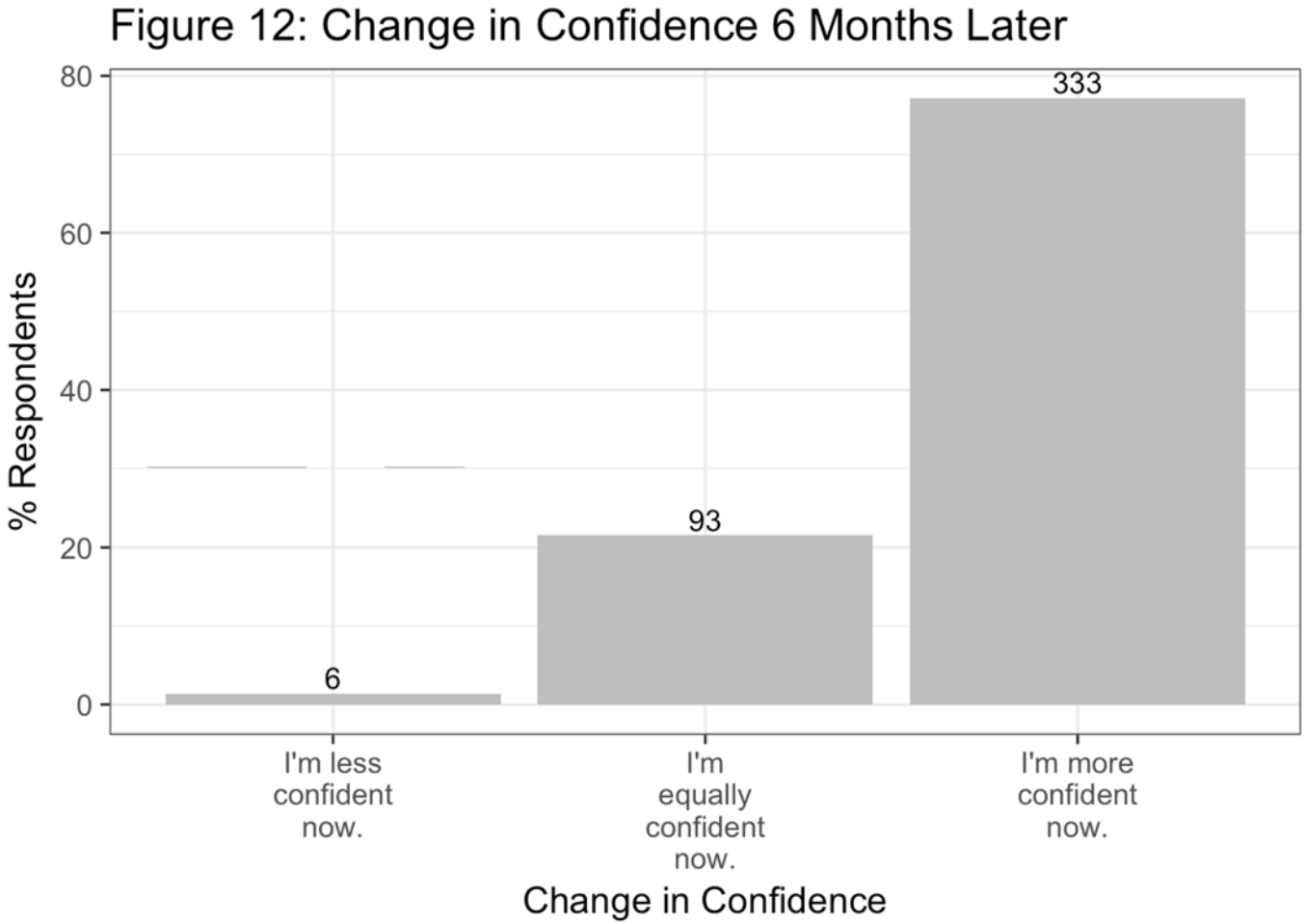
Short term survey



Paired Analyses Table

skill	mean_pre_feeling	sd_pre_feeling	mean_post_feeling	sd_post_feeling	n_pre	n_post	p.value
AnalysesEasier	4.104167	0.9139077	4.377778	0.7713786	144	135	0.0031887
OvercomeProblem	3.181818	1.0183967	3.807407	0.8149379	143	135	0.0000000
ProgrammingSoftware	2.542254	1.2182026	3.533333	0.9910043	142	135	0.0000000
RawData	4.416667	0.9274373	4.622222	0.8540218	144	135	0.0419499
SearchOnline	3.805556	0.8712805	4.281481	0.7595349	144	135	0.0000002
WriteScript	2.601399	1.3486554	3.792593	0.9470795	143	135	0.0000000

Increased confidence and change in perspective persists longer term



Get involved!

- Take a workshop
- Become an instructor
- Become a Member Organization
- Collaborate on training efforts

(lesson development, instructor training, teaching)

http://www.datacarpentry.org
http://www.software-carpentry.org

Thoughts on Entrepreneurship in URSSI

Michael Zentner

Director, HUBzero Platform

Incubator Leader, Science Gateways Community Institute

Unfunded Senior Personnel, URSSI

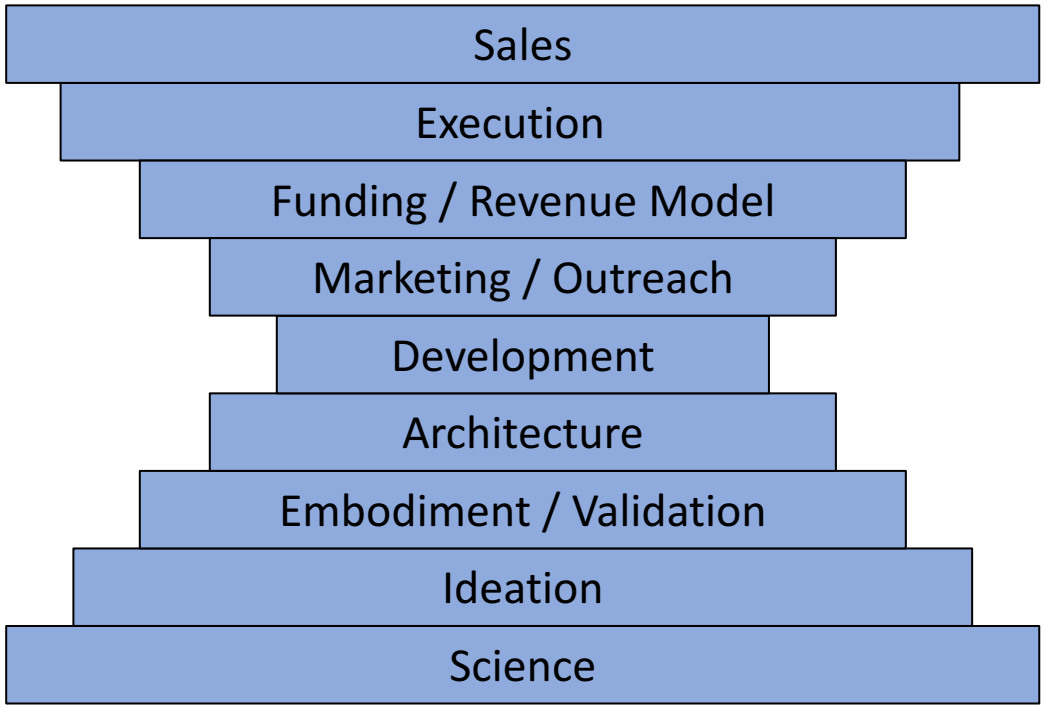
Entrepreneur

one who organizes, manages, and assumes the risks of a business or enterprise

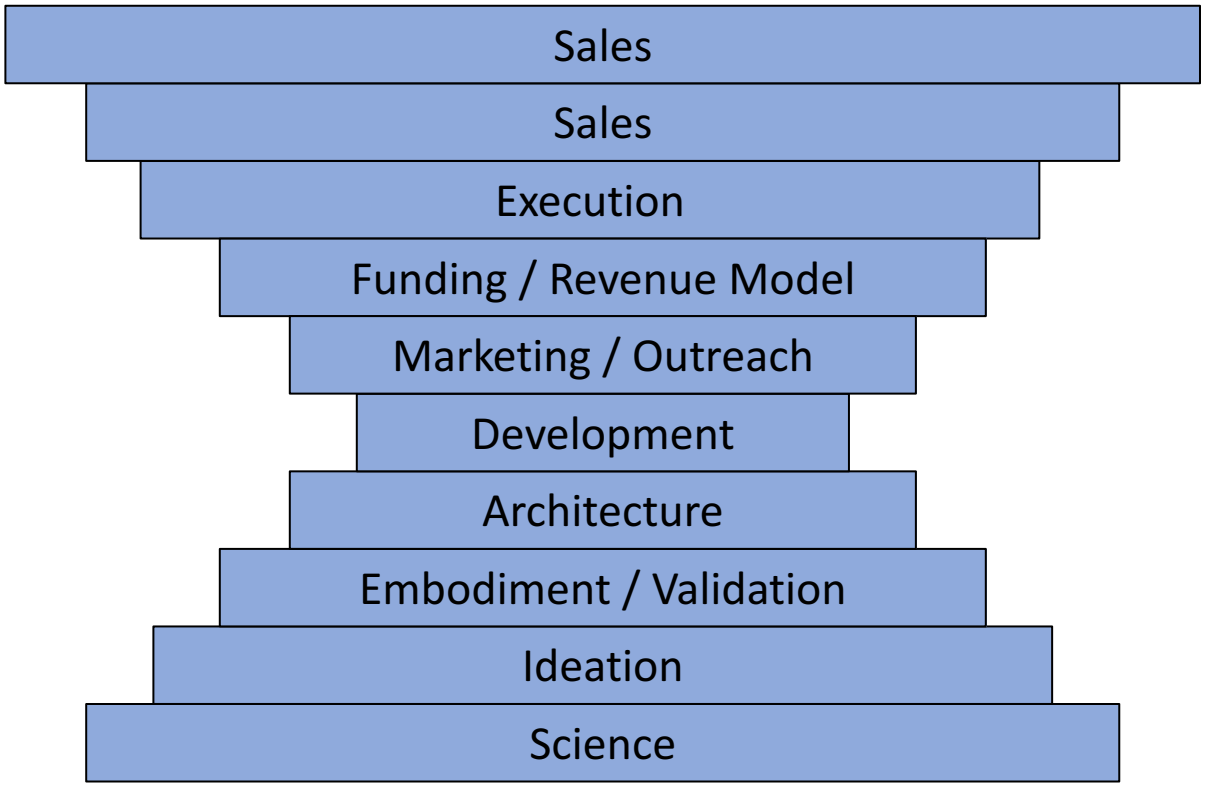
Entrepreneur

one who organizes, manages, and assumes the risks of a business or enterprise

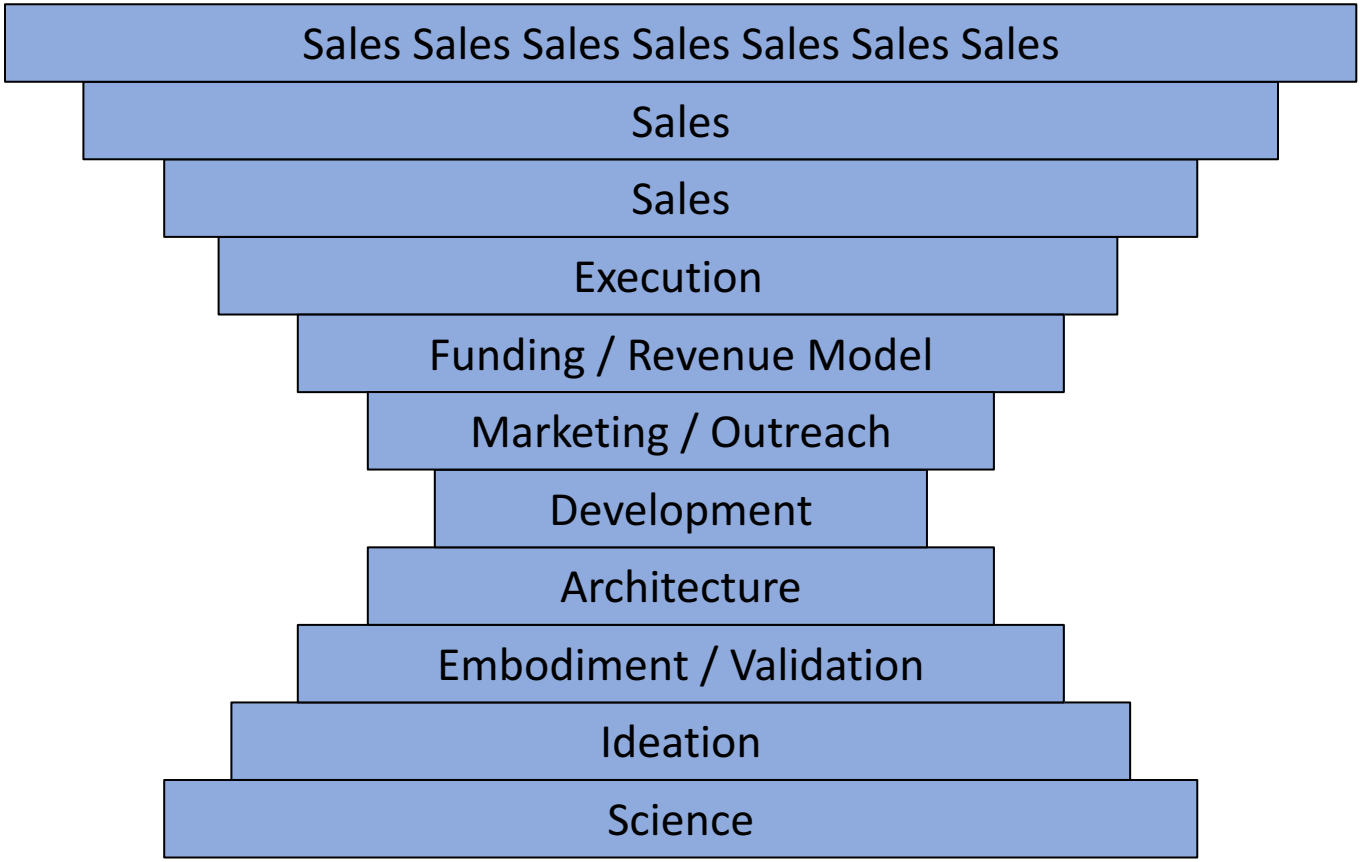
What is our role in overall value?



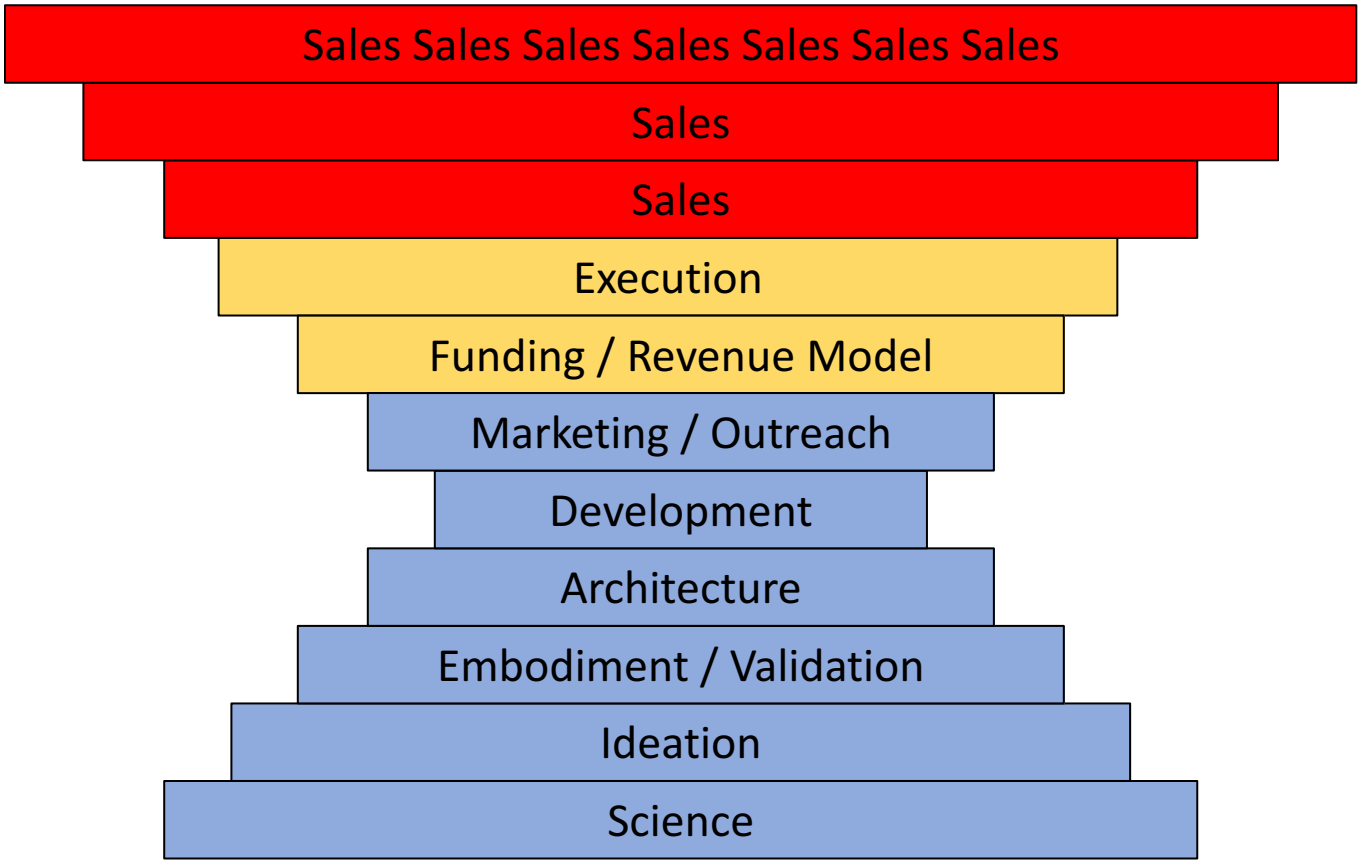
What is our role in overall value?



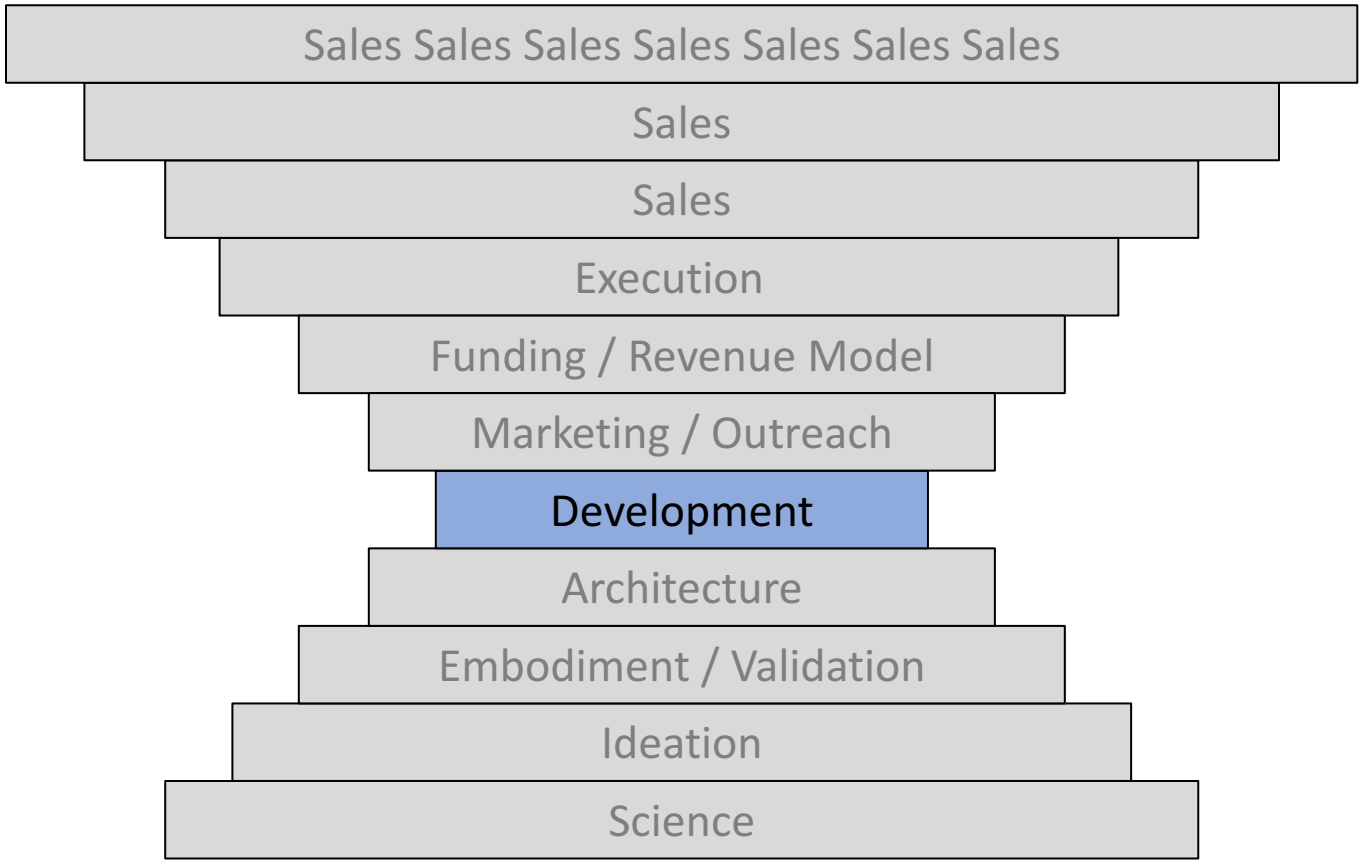
What is our role in overall value?



What is our role in overall value?



What is our role in overall value?



An institute might…

…teach how to participate as more fully rounded project stewards.

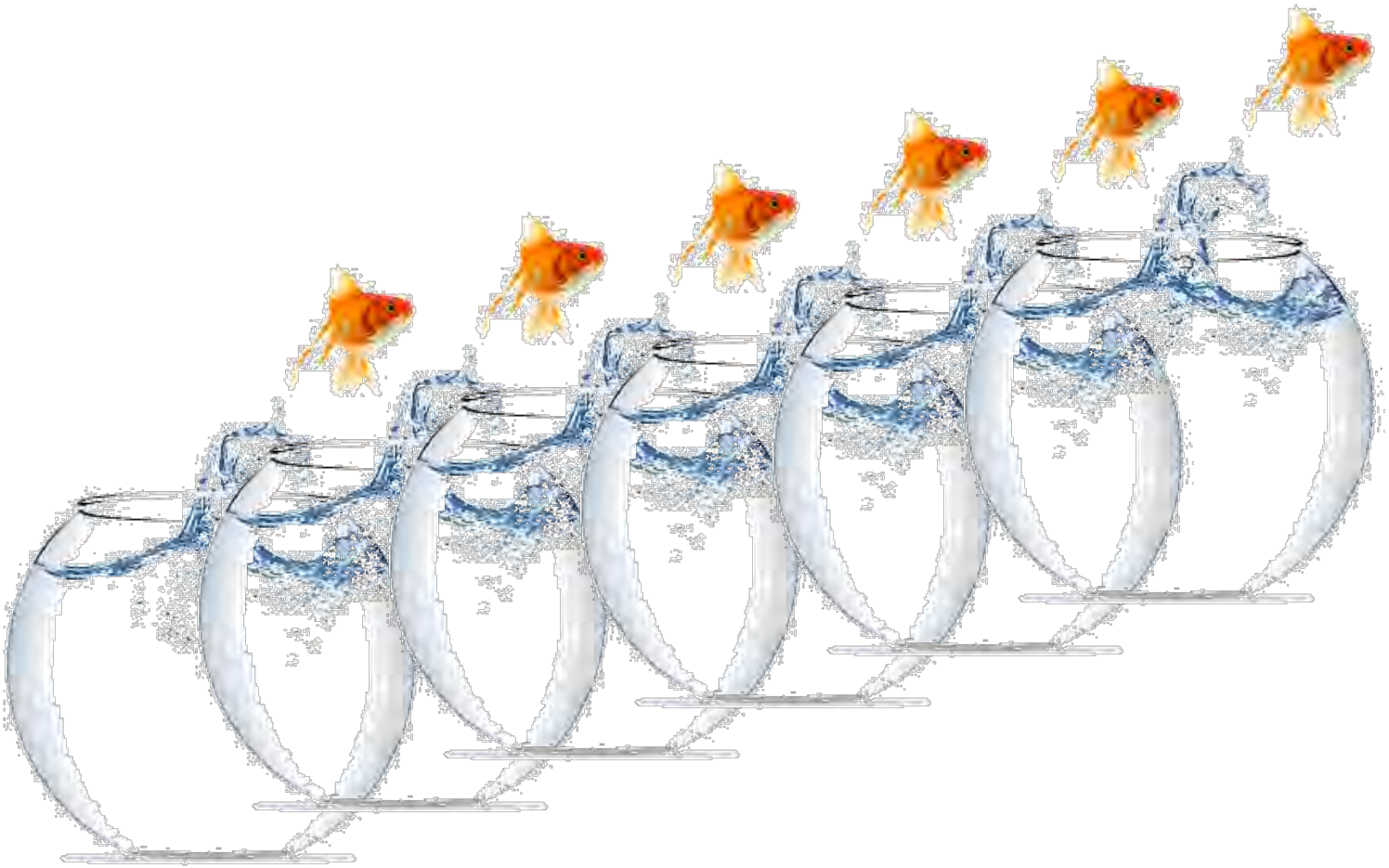
Our system…

Our system…

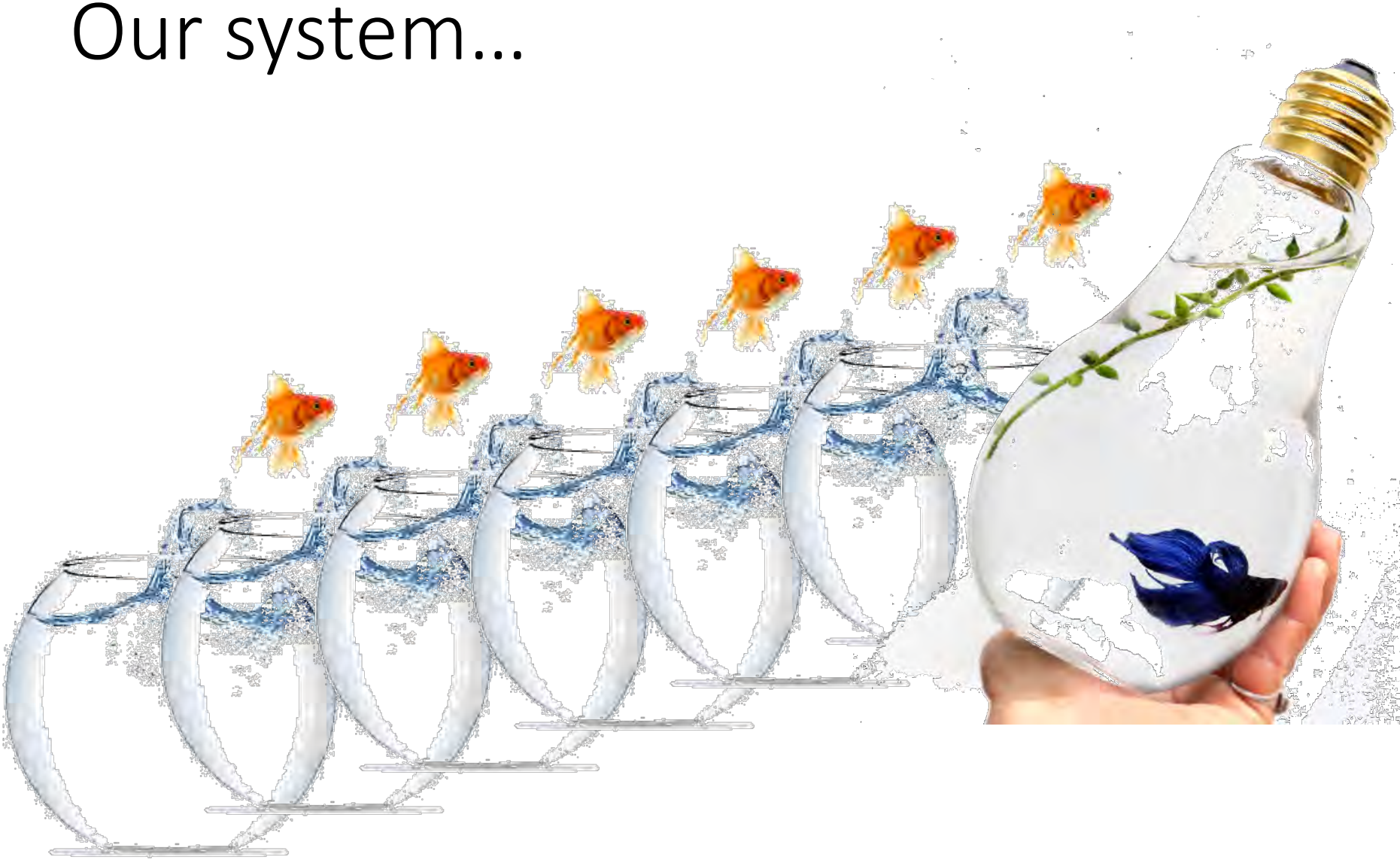
Our system…

Our system…

Our system…



Our system…



The dichotomy…

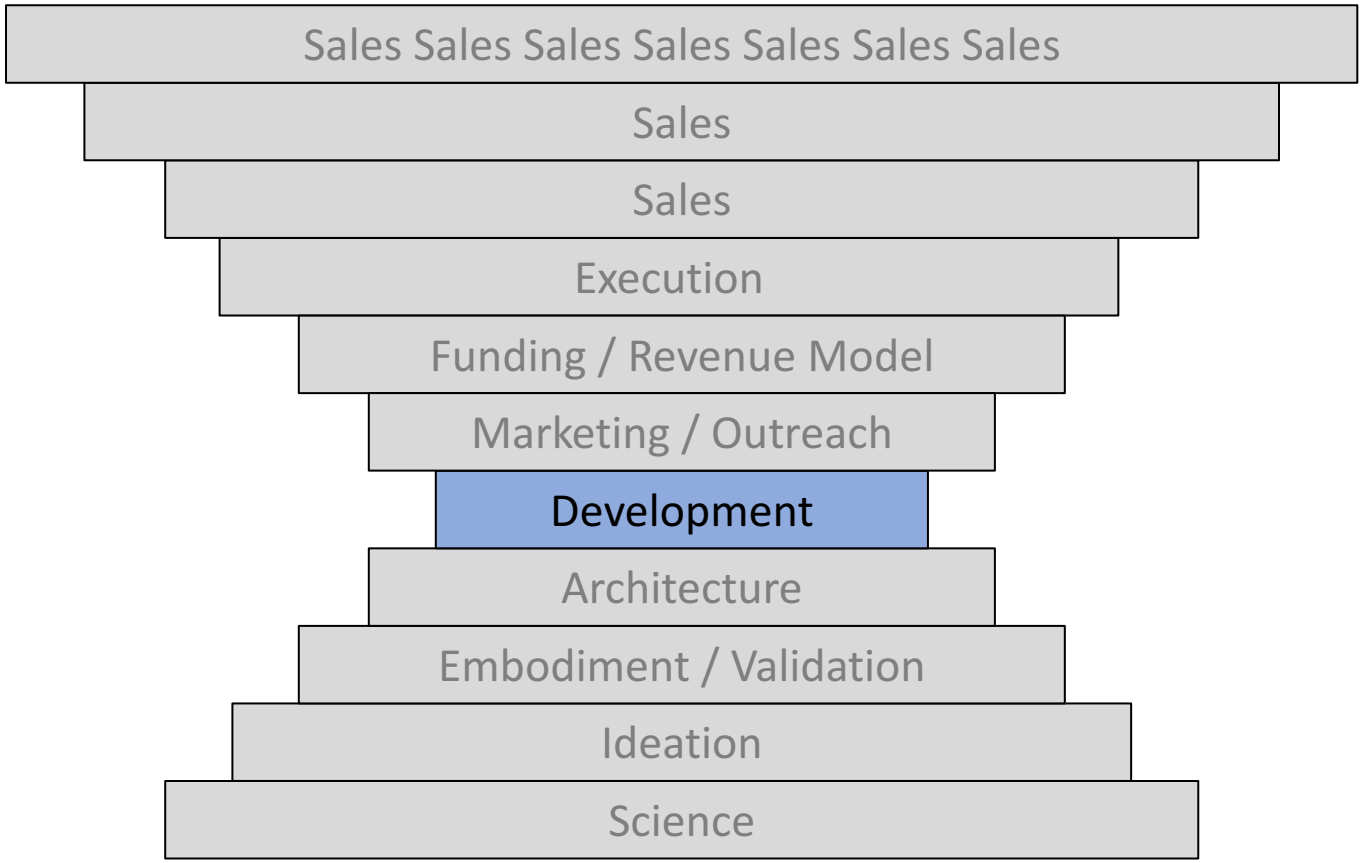
Such market forces drive innovation

Such market forces inhibit continuity

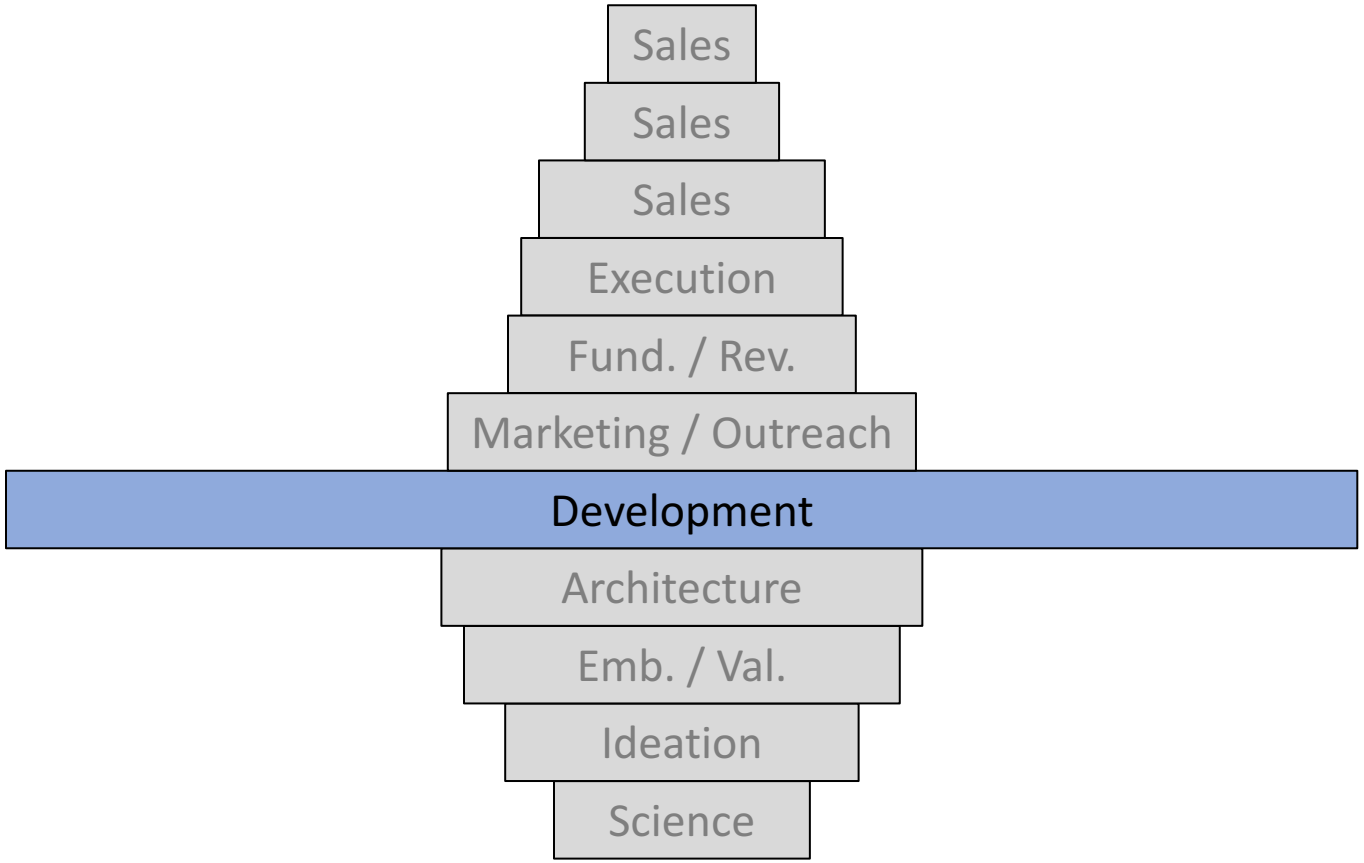
An institute might…

…proactively provide community to make us more aware of when we are hanging on to that which should change.

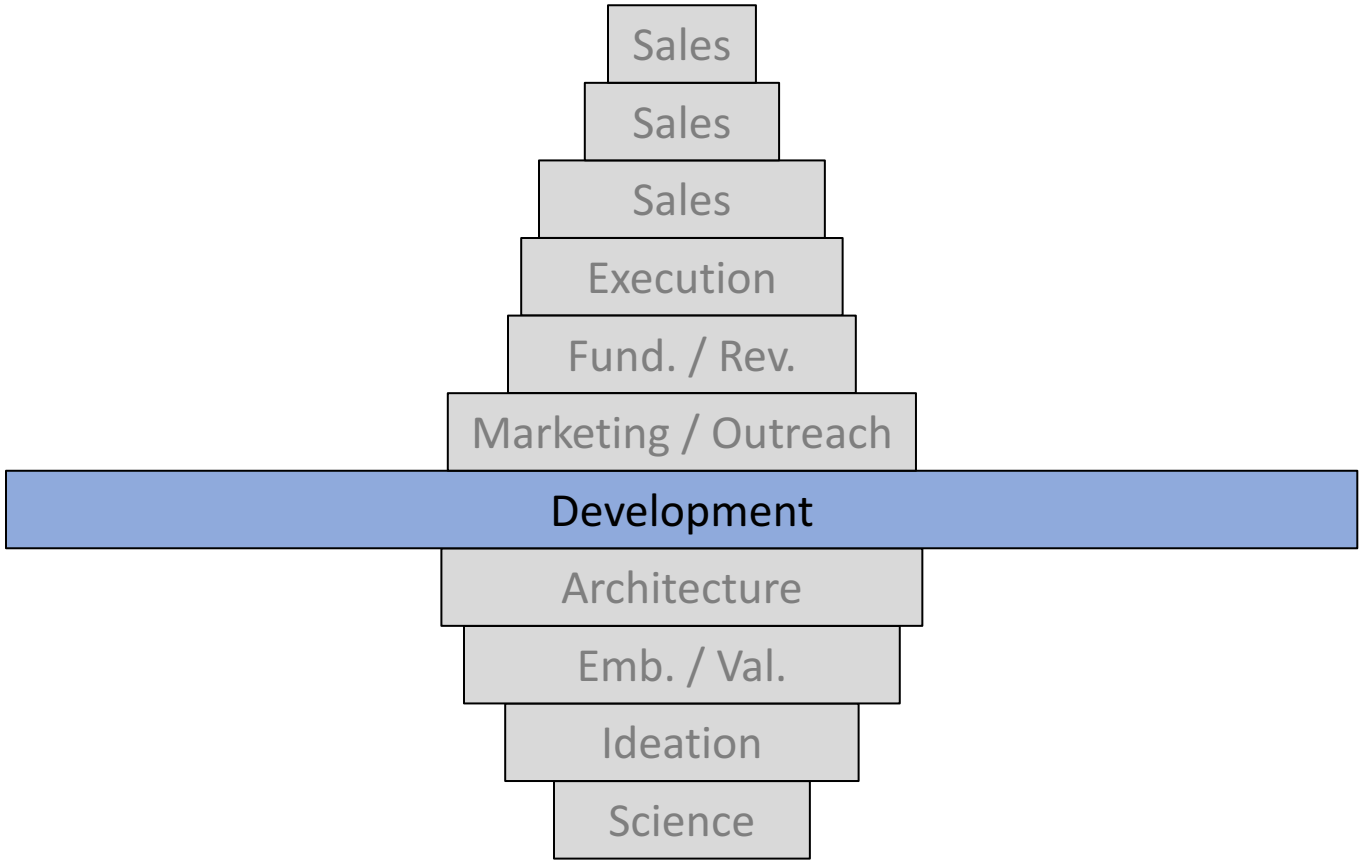
How would an entrepreneur think?



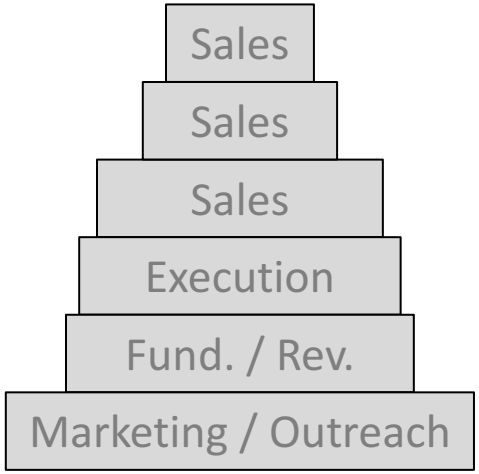
In quantity, our skills look like this:



This resembles a [vertical specialized] consulting company



An institute might…

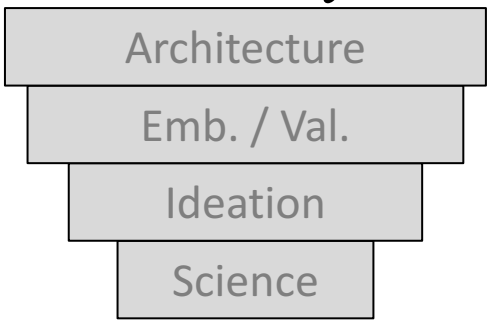


Specialize in this

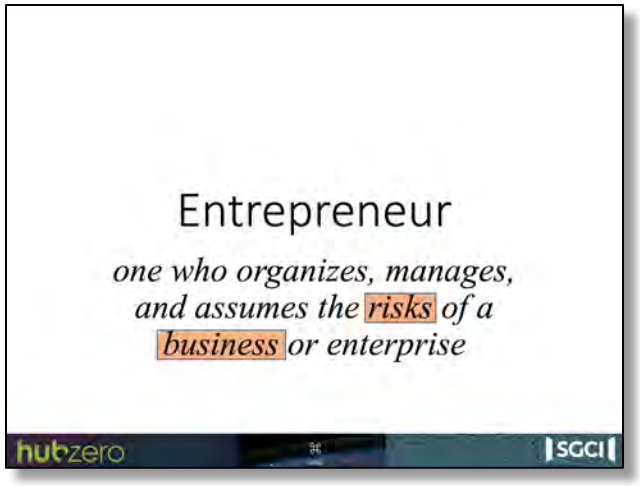
Development

Form a community of this that knows where efforts can lever other efforts based on vast experience & job-shop developers as a virtual organization to those opportunities

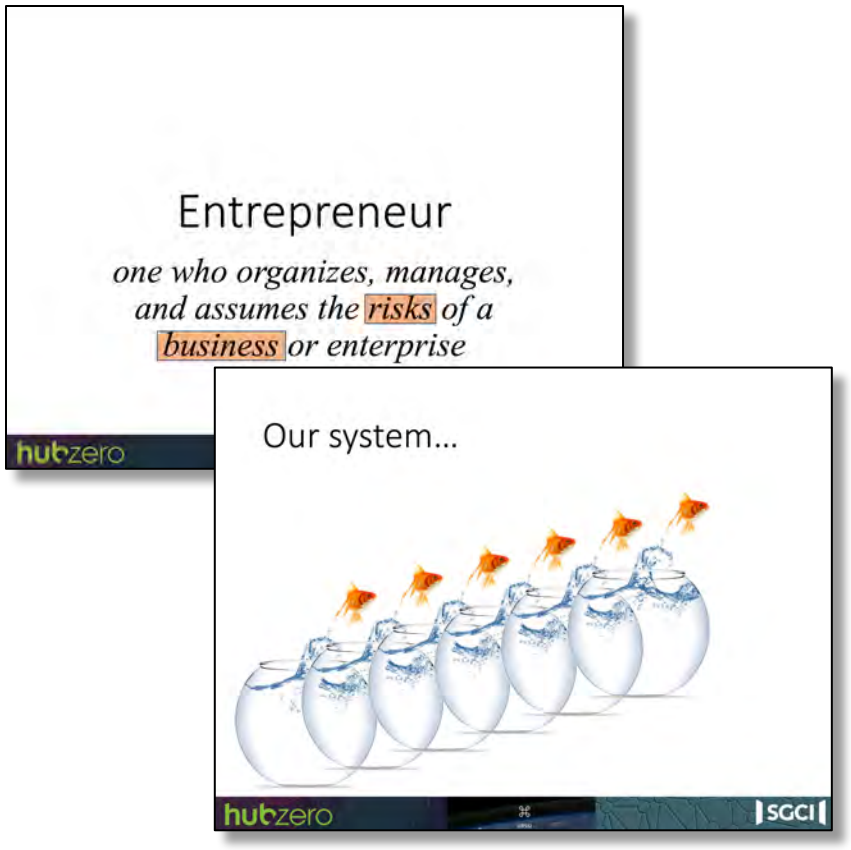
Reach these opportunities on behalf of its community



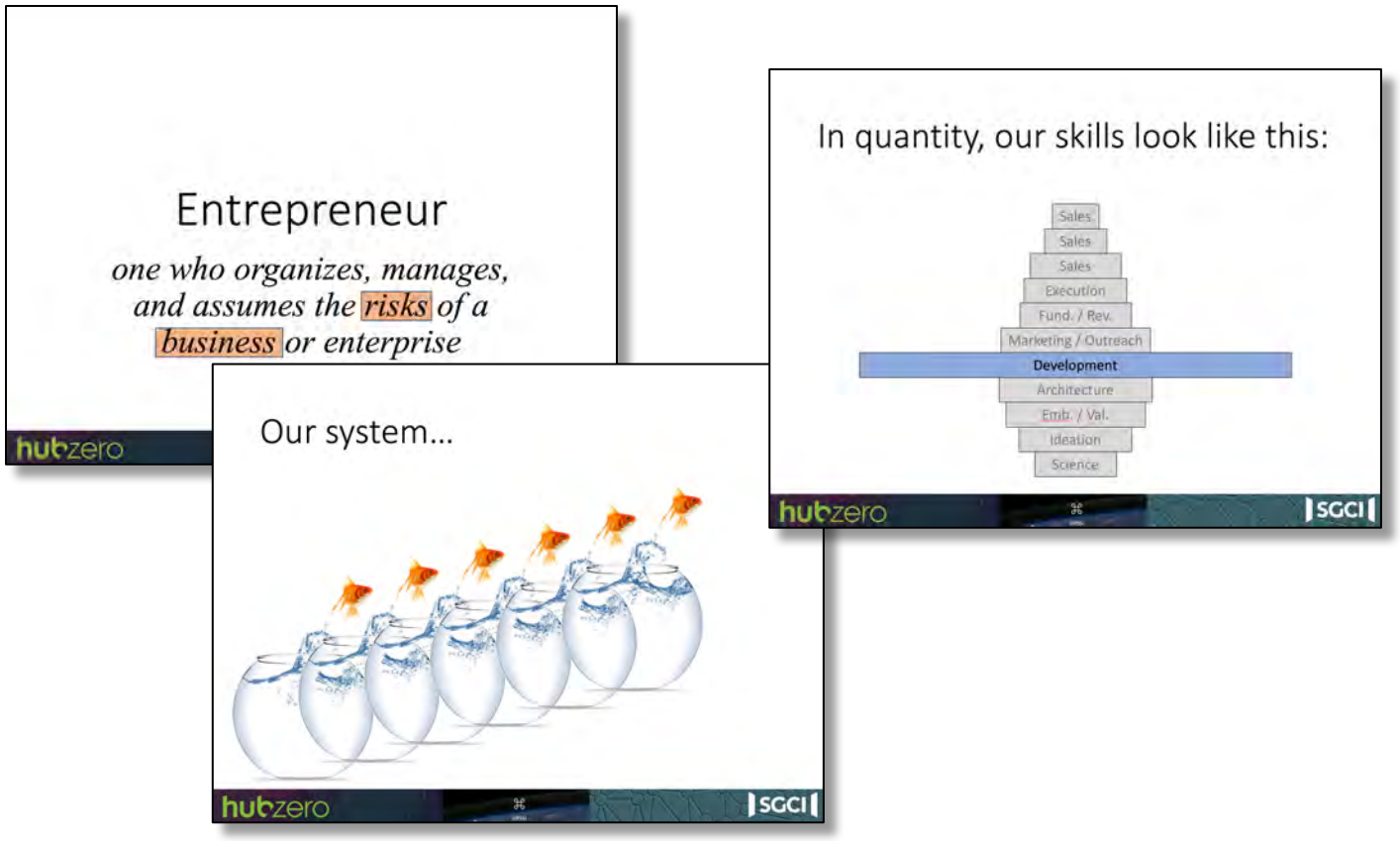
Do I believe what I say?



Do I believe what I say?



Do I believe what I say?



Do I believe what I say?

Entrepreneur

one who organizes, manages, and assumes the risks of a business or enterprise

hubzero

Our system…

hubzero SGCI

In quantity, our skills look like this:

hubzero

An institute might…

Specialize in this

Development

Form a community of this that knows where efforts can lever other efforts based on vast experience & job-shop developers as a virtual organization to those opportunities

Reach these opportunities on behalf of its community

hubzero SGCI