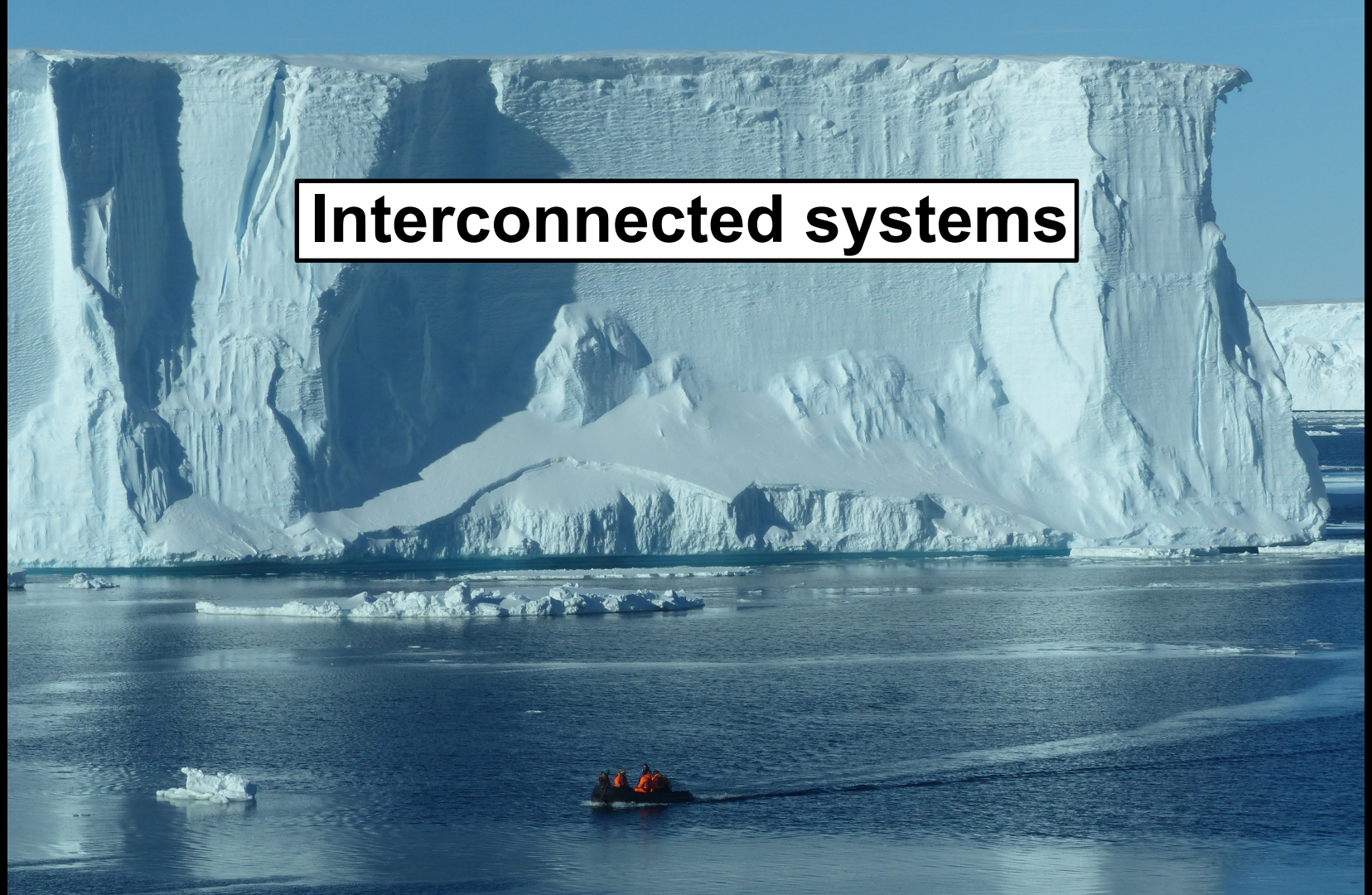
CryoCloud JupyterHub for NASA Cryosphere communities: Open science as a process, not a product

Tasha Snow[1], Joanna Millstein[2], Wilson Sauthoff[1], Jessica Scheick[3], Wei Ji Leong[4], James Colliander[5,6], James Munroe[6], Fernando Perez[7], Denis Felikson[8], Tyler Sutterley[9], Matt Fisher[10], Facundo Sapienza[7], Ellianna Abrahams[8], Whyjay Zheng[11], Matthew Siegfried[1]

[1]Mines
[2]MIT/WHOI
[3]UNH
[4]Development Seed
[5]2i2c
[6]UBC
[7]Berkeley
[8]NASA Goddard
[9]APL, UW
[10]NSIDC
[11]National Central University

CryoCloud


Interconnected systems

Interconnected researchers

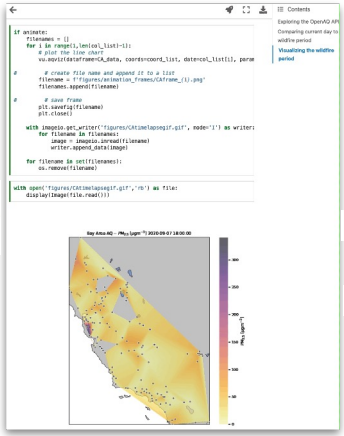
Kontur

Inclusive

Accessible

Reproducible

Transparent



Impactful

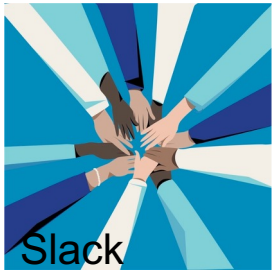
Open science, community building, and infrastructure for _____ communities

Defining open-source science and values

Models for serving the community

CryoCloud JupyterHub for Cryosphere communities

Lessons learned and outlook



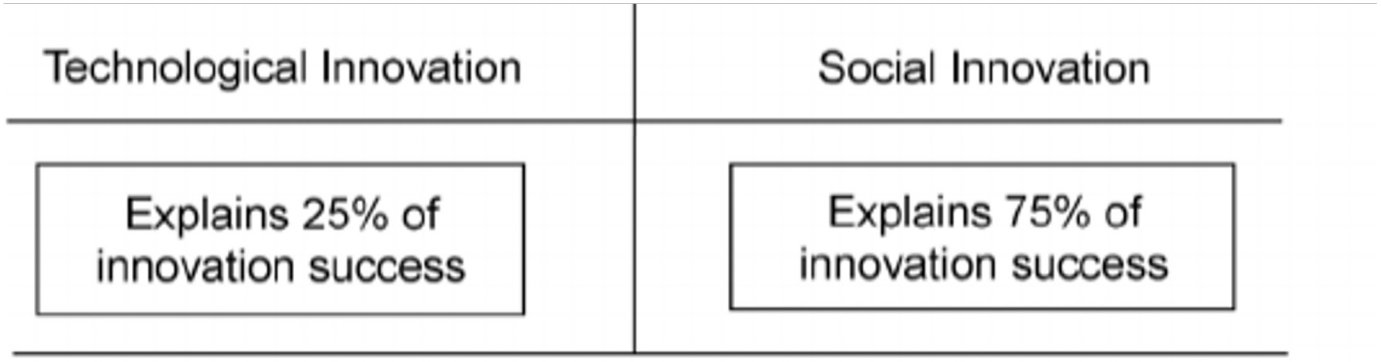
The problem

Science has success in discovery, some limitations in impact

- Participation of all society
- More informed decision-making
- Better localized knowledge and decision-making

The opportunity

Advances in technological & social innovation can boost science



Source: Erasmus University: Competition and Innovation monitor (2006)

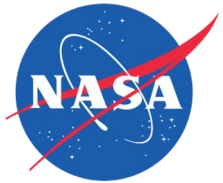
Necessity

For a high performing science ecosystem to address global challenges

- Misinformation
- Climate change
- Policy making

Science done in a fundamentally more open way is the future

Principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity.



Open-source science at the forefront

OPEN-SOURCE SCIENCE INITIATIVE

Overview Why Do Open Science? Transform to Open Science (TOPS)

Open-Source Science Initiative

NASA is making a long-term commitment to building an inclusive open science community over the next decade. Open-source science is a commitment to the open sharing of software, data, and knowledge (algorithms, papers, documents, ancillary information) as early as possible in the scientific process. The principles of open-source science are to make publicly funded scientific research transparent, inclusive, accessible, and reproducible. Advances in technology, including collaborative tools and cloud computing, help enable open-source science, but technology alone is insufficient. ***Open-source science requires a culture shift to a more inclusive, transparent, and collaborative scientific process, which will increase the pace and quality of scientific progress.***

To help build a culture of open science, **NASA is championing a new initiative: the Open-Source Science Initiative (OSSI)**. OSSI is a comprehensive program of activities to enable and support moving science towards openness, including policy adjustments, supporting open-source software, and enabling cyberinfrastructure. OSSI aims to implement NASA's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024, which was developed through community input.

OPEN **(TRANSPARENT)** SCIENCE
scientific process and results should be visible, accessible, and understandable

OPEN **(ACCESSIBLE)** SCIENCE
data, tools, software, documentation, and publications should be accessible to all (FAIR)

OPEN **(INCLUSIVE)** SCIENCE
process and participants should welcome participation by and collaboration with diverse people and organizations

OPEN **(REPRODUCIBLE)** SCIENCE
scientific process and results should be open such that they are reproducible by members of the community

Transform to Open Science (TOPS)

From 2022 to 2027, TOPS will accelerate the engagement of the scientific community in open science practices through events and activities aimed at:

- Lowering barriers to entry for historically excluded communities
- Better understanding how people use NASA data and code to take advantage of our big data collections
- Increasing opportunities for collaboration while promoting scientific innovation, transparency, and reproducibility.

The TOPS mission is aligned with recommendations from NASA's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024, the National Academies reports on open science, reproducibility, and scientific software, and the 2021 UNESCO draft Recommendation on Open Science synthesis report.

Engage in the open science community on the Transform to Open Science (TOPS) GitHub

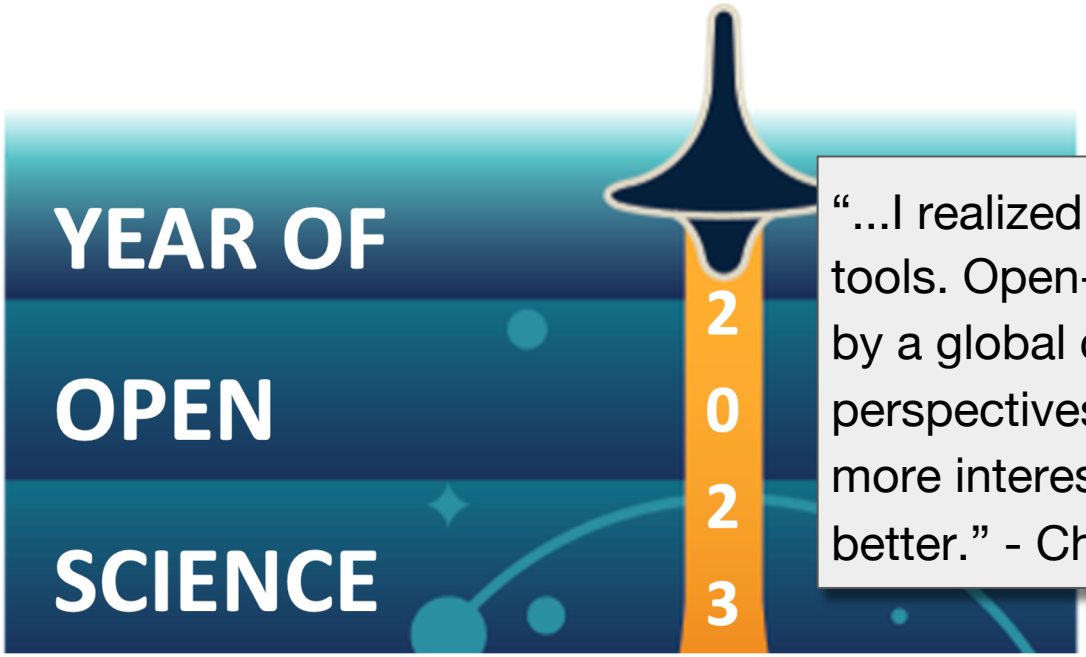
TOPS
TRANSFORM to OPEN SCIENCE

TOPS. (2021). Creative Commons Attribution 4.0 License. https://doi.org/10.5281/zenodo.5225076

Open Science Curricula: *OpenCore*

https://nasa.github.io/Transform-to-Open-Science/

NASA (and US+): 2023 the Year of Open Science



"...I realized that open science isn't just about tools. Open-science innovation is being driven by a global community with diverse perspectives. The scientific questions are more interesting and nuanced, the solutions better." - Chelle Gentemann

open.science.gov

U.S. National Library of Medicine

Open science values

Intellectual generosity

Intellectual humility

Right to participate in science

Everyone deserves to be treated with dignity and respect

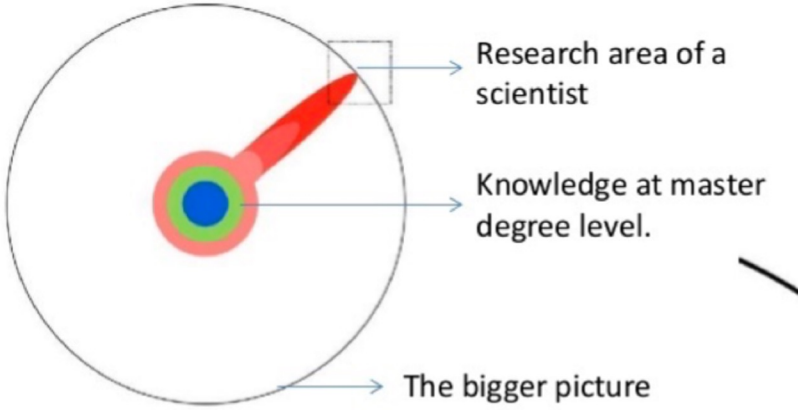
Intellectual generosity

Sharing ideas, advancing other's understanding

Reduce competition and enhance collaboration

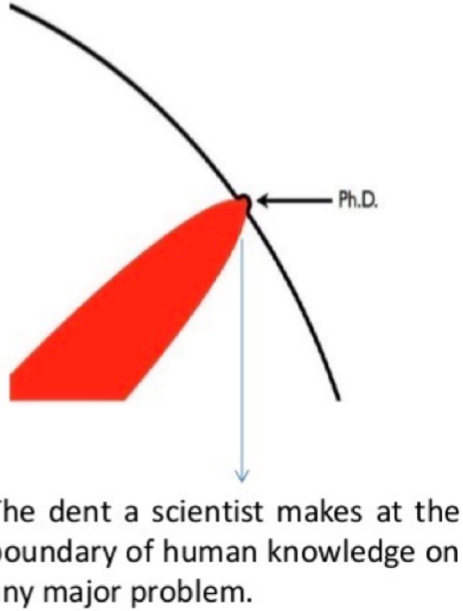
LinkedIn

Intellectual humility



matt.might.net

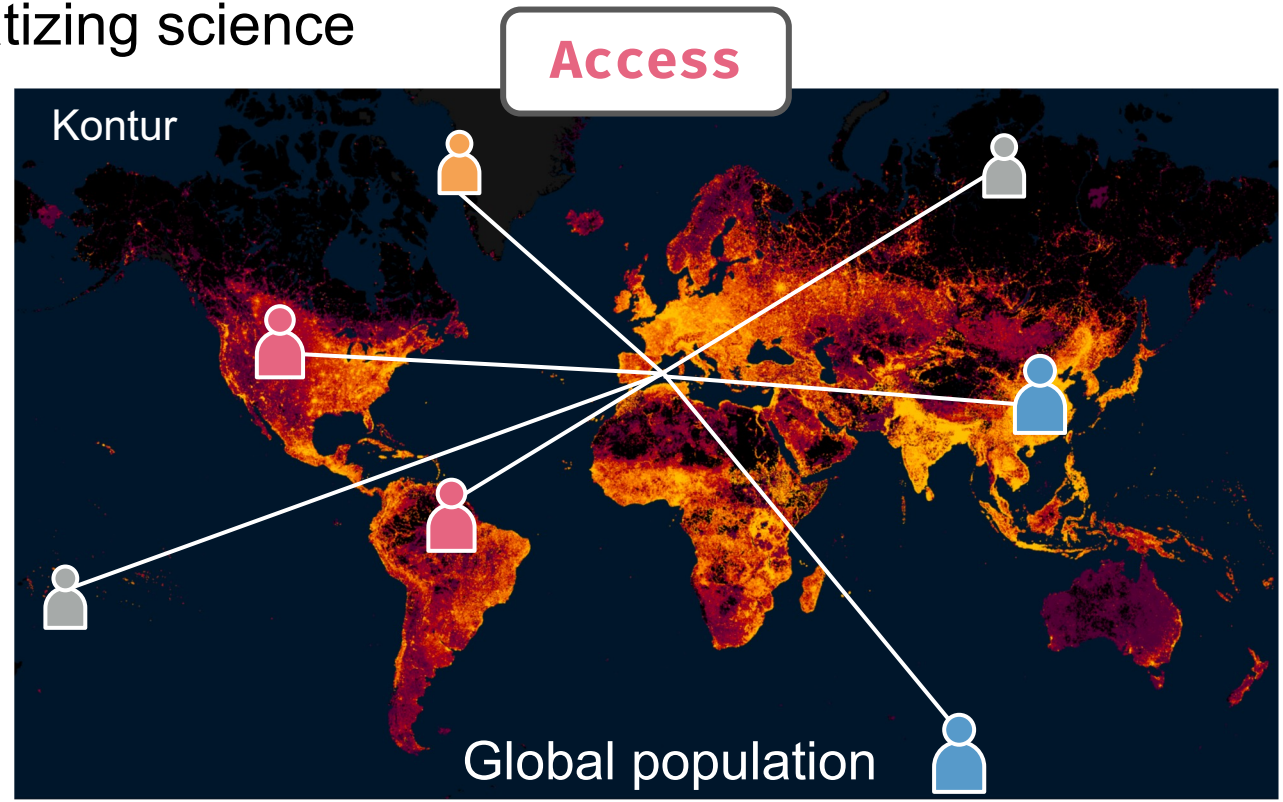
Our contributions are small relative to the body of knowledge



Give and receive criticism with grace

Right to participate in science

Democratizing science



Everyone deserves to be treated with dignity and respect

Objective and constructive discourse

Open science values

Intellectual generosity

Intellectual humility

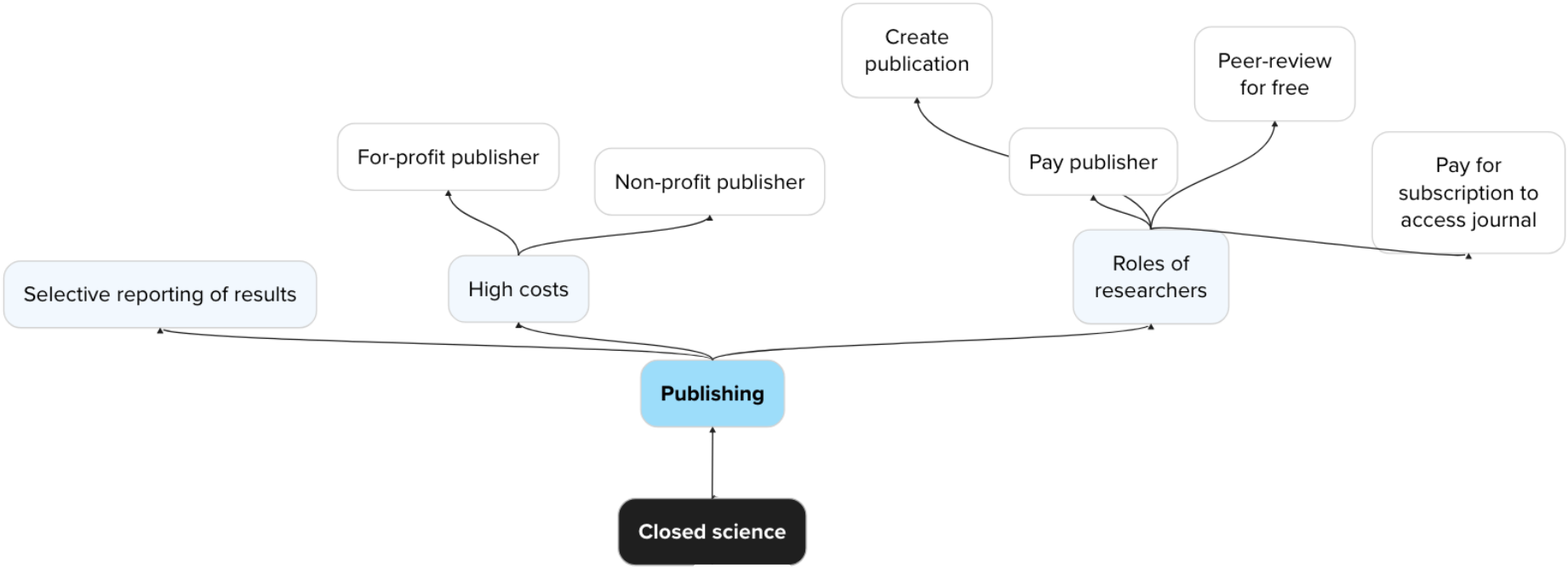
Right to participate in science

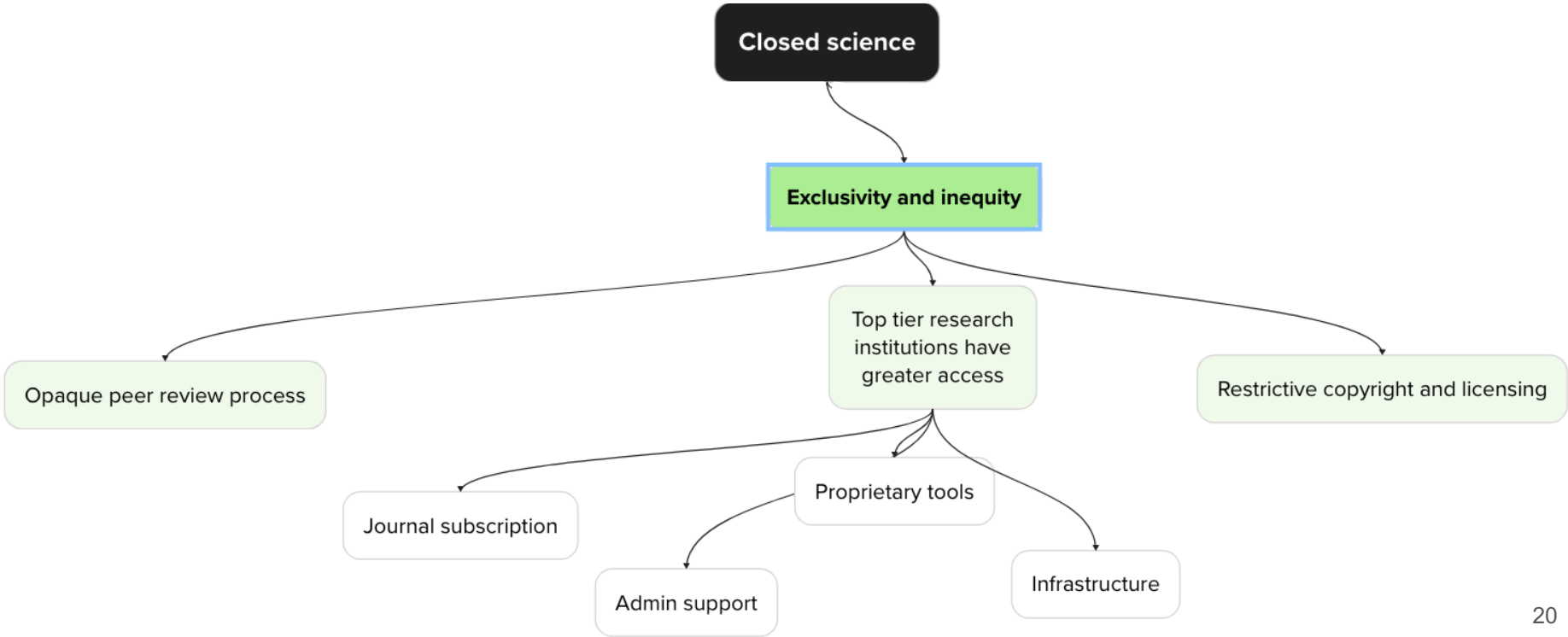
Everyone deserves to be treated with dignity and respect

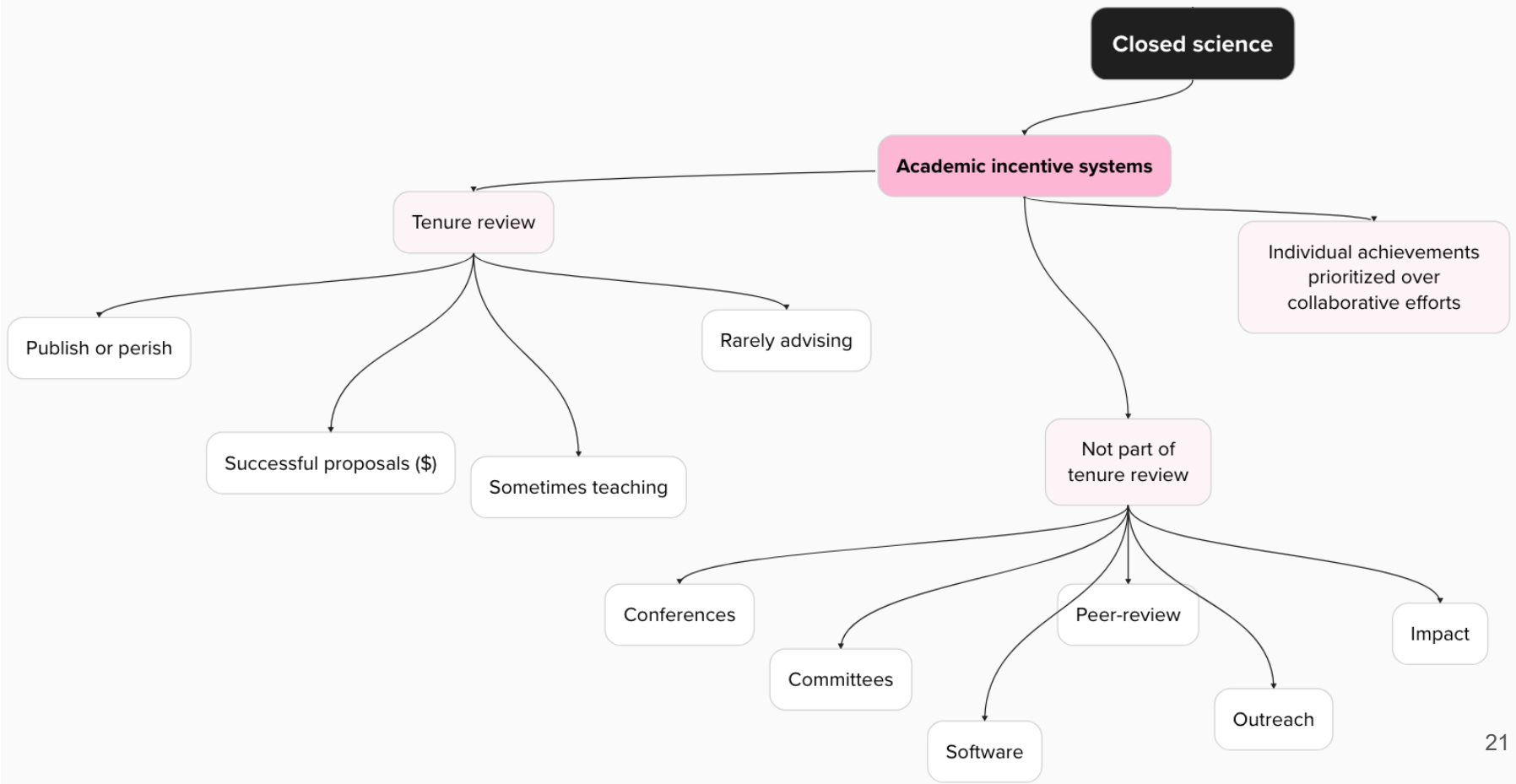
Open science as a process, not a product

The challenge:

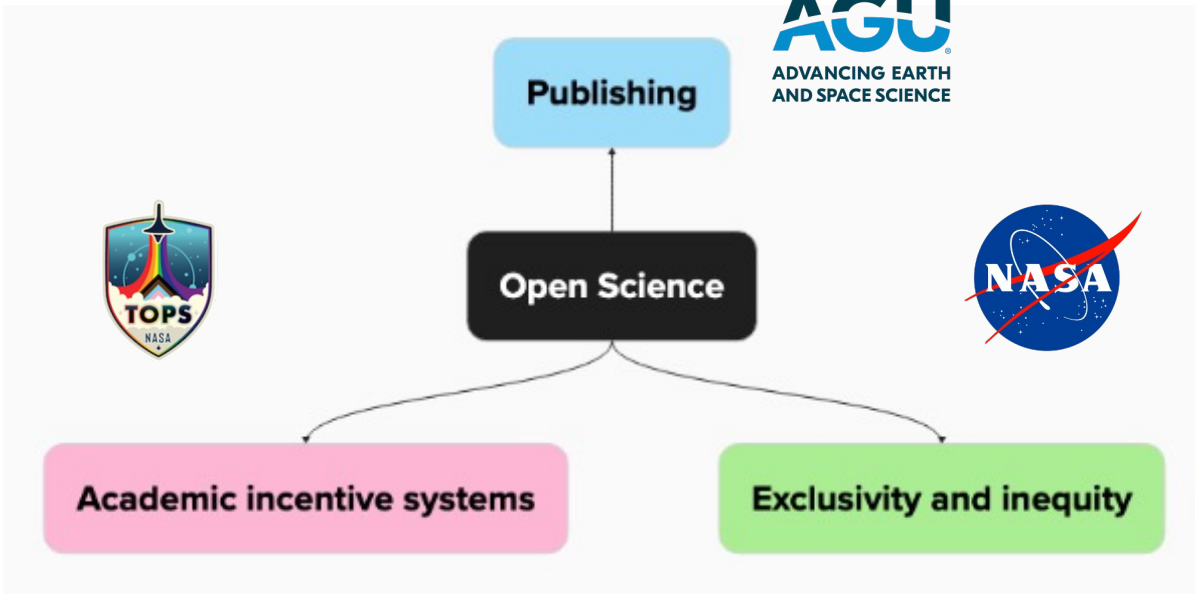
Closed science


Create publication
Peer-review for free
For-profit publisher
Pay publisher
Non-profit publisher
Pay for subscription to access journal
Roles of researchers
Selective reporting of results
High costs
Publishing
Closed science

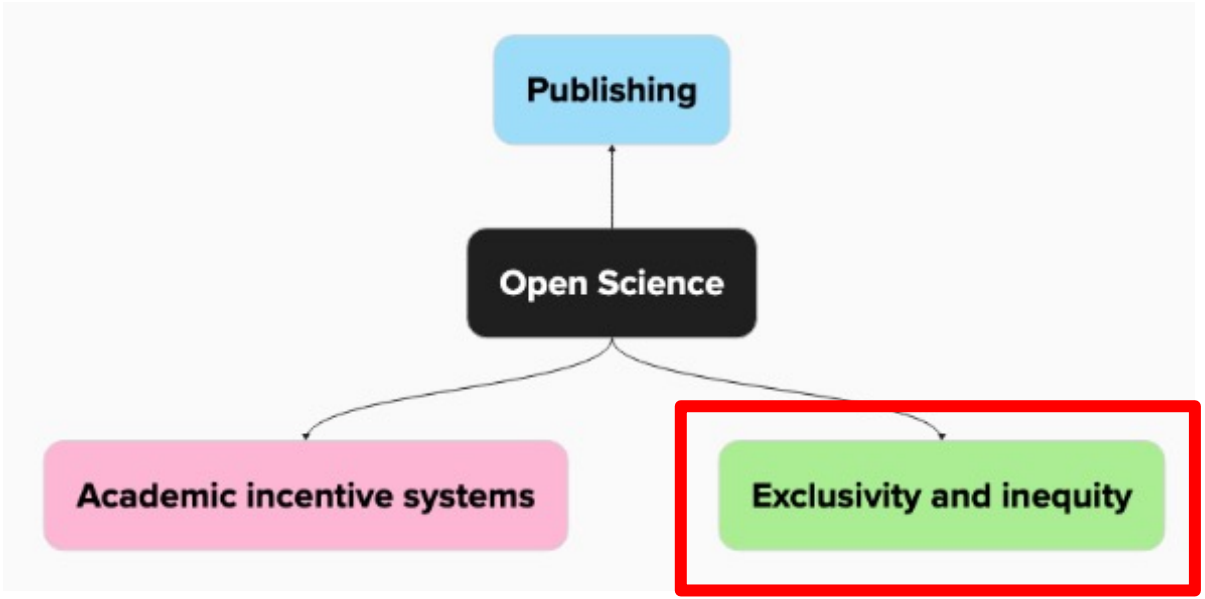

Closed science
Exclusivity and inequity
Opaque peer review process
Top tier research institutions have greater access
Restrictive copyright and licensing
Journal subscription
Proprietary tools
Admin support
Infrastructure


Closed science
Academic incentive systems
Tenure review
Individual achievements prioritized over collaborative efforts
Publish or perish
Rarely advising
Successful proposals ($)
Sometimes teaching
Not part of tenure review
Conferences
Committees
Peer-review
Impact
Outreach
Software

Finding solutions



Our piece of the puzzle



CryoCloud

Solutions through computing

Research requires computing capabilities

Research computing requires infrastructure that may not be trivial to access

Ai

Adobe

X

W

Arc GIS ESRI

QGIS

python™

R

MATLAB®

We need certain computing resources to produce impactful results

Produce solutions

Conduct research

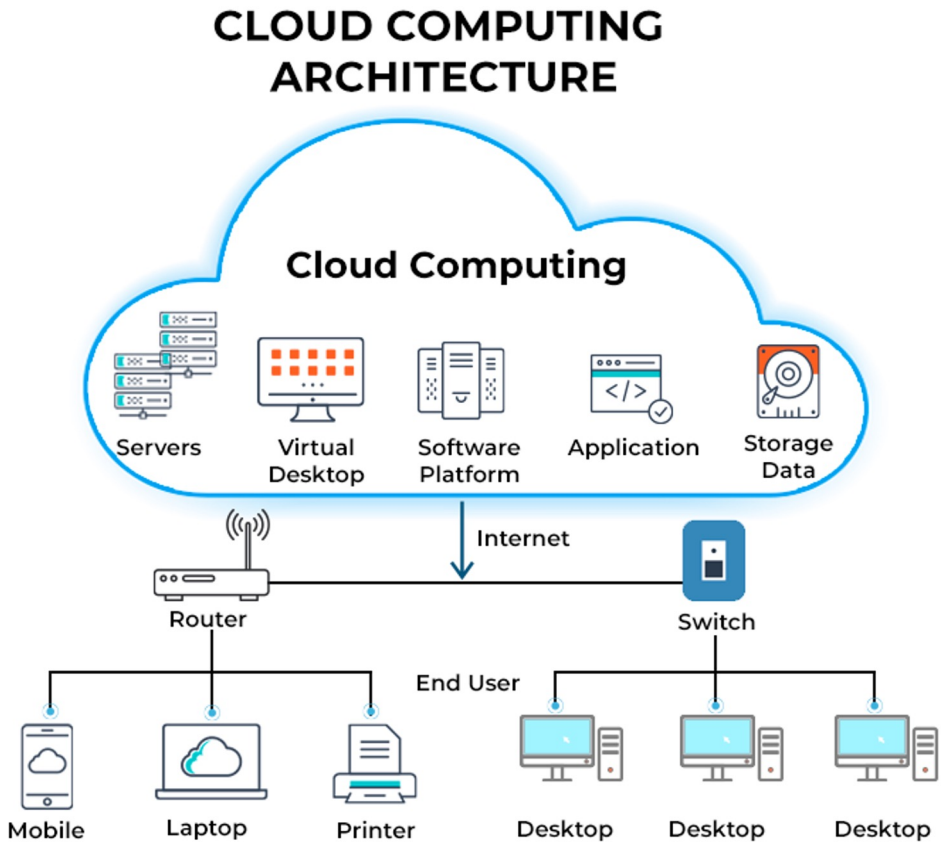
ChatGPT

On usable timeframes

ChatGPT

Collaborative, reproducible, open science in the cloud

What is the cloud?



AGU Advances

RESEARCH ARTICLE
10.1029/2020AV000354

Science Storms the Cloud

C. L. Gentemann[1,2], C. Holdgraf[3,4], R. Abernathey[3,5], D. Crichton[6], J. Colliander[3,7,8], E. J. Kearns[9], Y. Panda[3], and R. P. Signell[10]

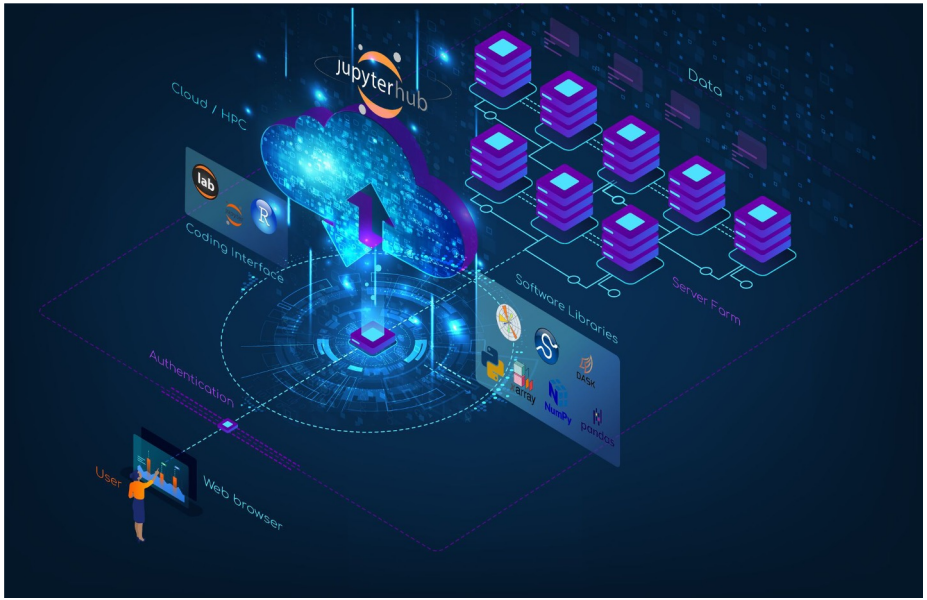
ChelleGentemann

Key Points

Science stands at the cusp of a new, open science, cloud-enabled era

Advances in data, software, and computing are enabling transformational, interdisciplinary science, **changing the realm of possible questions**

Deliberately designed open science communities can **advance science and inclusivity simultaneously**



https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020AV000354

AGU Advances

RESEARCH ARTICLE
10.1029/2020AV000354

Science Storms the Cloud

ChelleGentemann

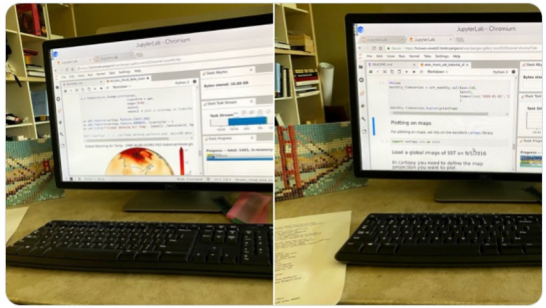
Key Points

- Science stands … cloud-enabled e…
- Advances in dat… enabling transf… changing the re…
- Deliberately des… advance scienc…

Process massive amounts of data from a $36 computer…

Chelle-ter in Place
@ChelleGentemann

On my kids @Raspberry_Pi running @TeamKano #OpenSource OS I'm analyzing @GCPcloud #cmip6 climate data and @awscloud MUR SST from @podaac. A $36 computer running processes on both AWS and GCP with over 80 workers and 245GB. #openscience! @NASAEarth



1:18 PM · Mar 1, 2020 · Twitter Web App

…or from your cell phone or on a flight

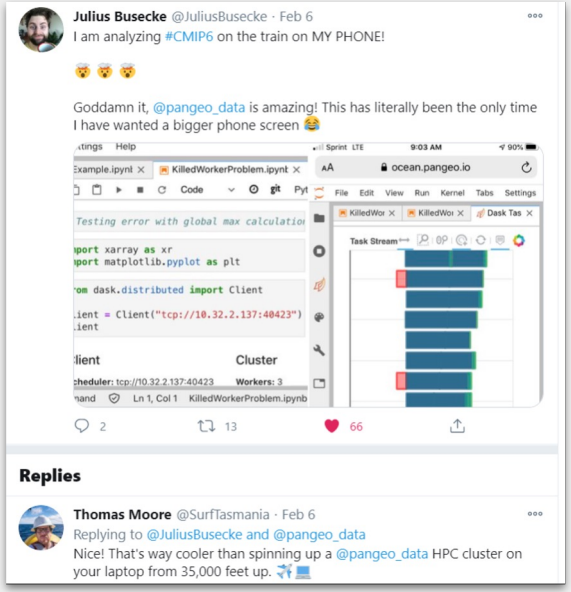


Image credits: Chelle Gentemann

https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020AV000354

The Digital Watering Hole (in the cloud)

Image: diana_robinson

An opportunity shaped by:

- Open, FAIR and CARE Data
- Scalable computation next to the data
- Modular tools for exploration/narrative

To tackle challenges that

- go beyond disciplinary silos…
- require analysis of *really* big data
- integration of disparate data…
- *participation* of disparate, diverse communities…
- to ultimately connect with society and impact critical decision making.

Proofs of concept

Exoplanet analysis: two orders of magnitude faster in the cloud

UCAR
NASA 3rd Eddy Symposium

Retention of an exoplanet's atmosphere as it orbits close to a star

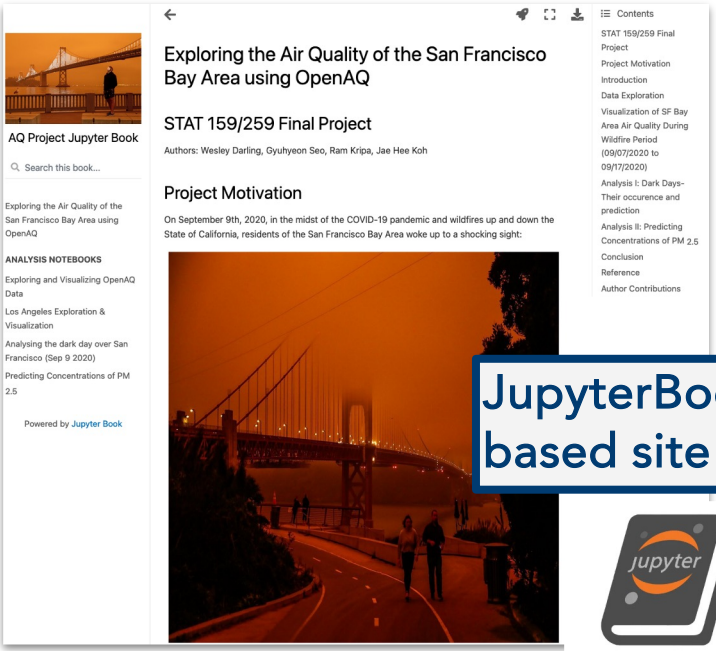
Analysis per planet:

- 5 hours on local computer
- 5-10 min in the cloud

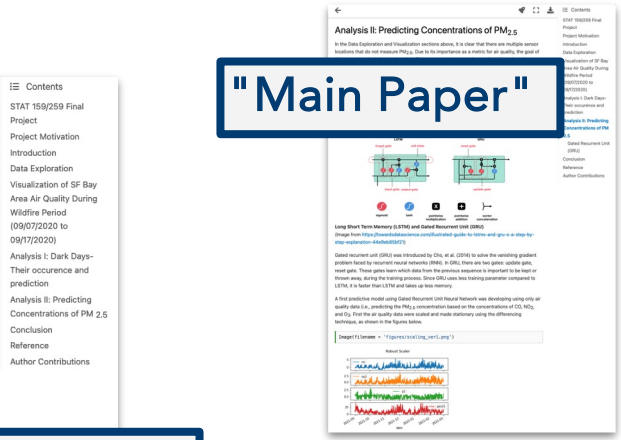
With a few lines of code that gave cloud access to data archives

Stat 159 by Fernando Pérez (course site)

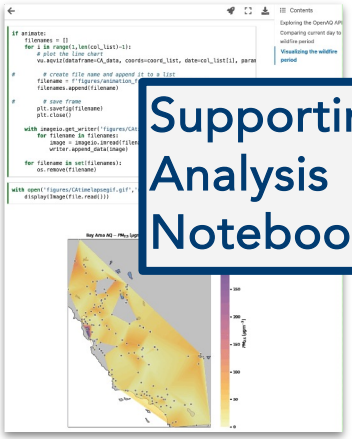
Berkeley
UNIVERSITY OF CALIFORNIA



JupyterBook based site

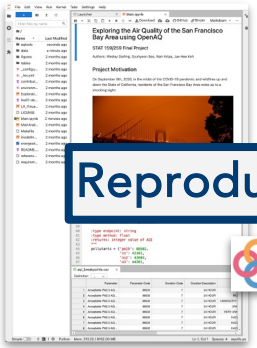


"Main Paper"



Supporting Analysis Notebooks

Supporting code & tests



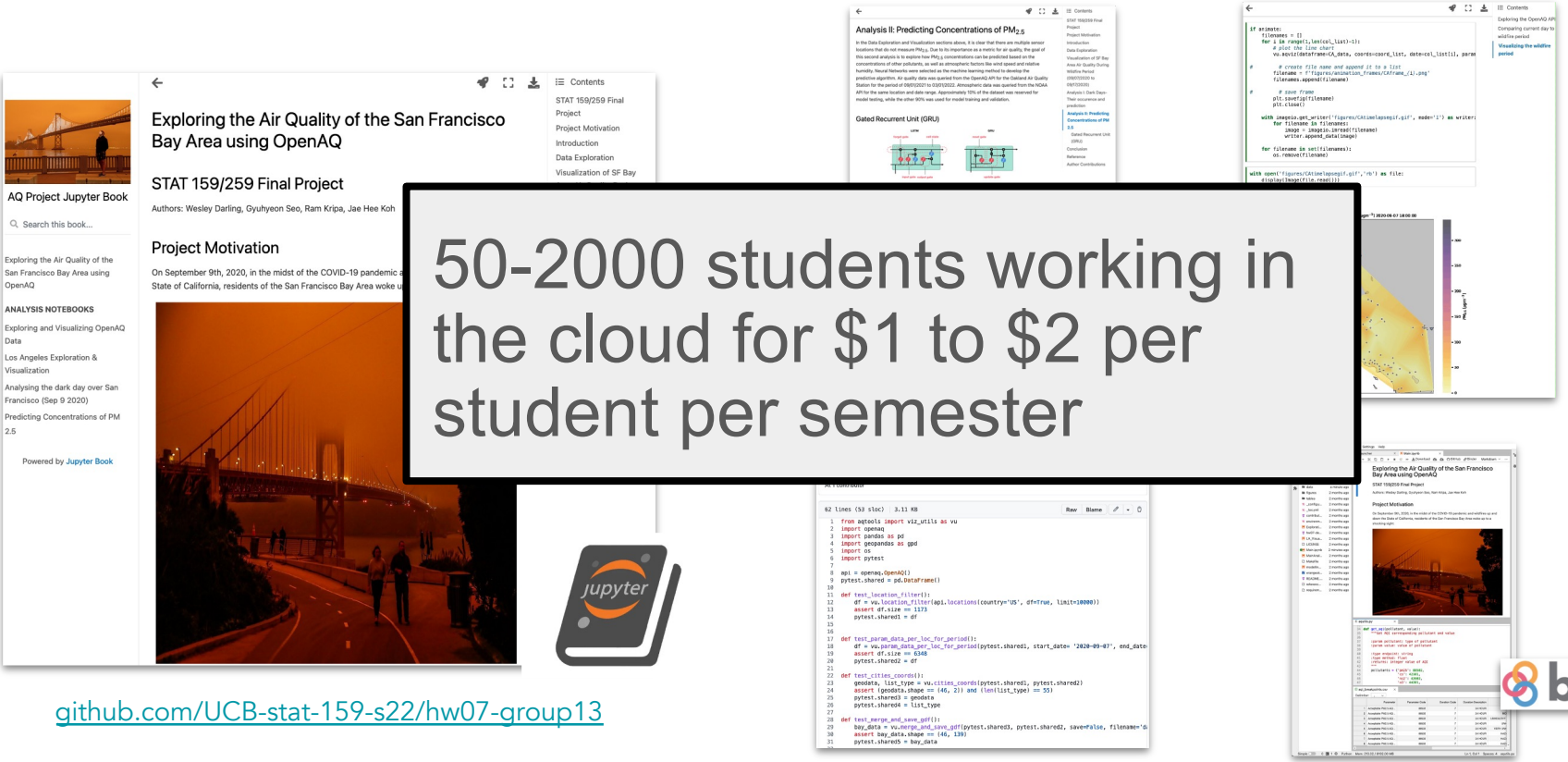
Reproducibility

binder

github.com/UCB-stat-159-s22/hw07-group13

Stat 159 by Fernando Pérez (course site)

Berkeley
UNIVERSITY OF CALIFORNIA



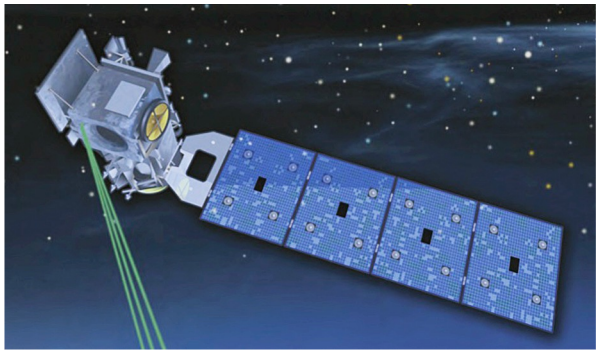
github.com/UCB-stat-159-s22/hw07-group13

New spaces and organizational models are needed

Cloud computing and ICESat-2 science

Cloud computing and open science concerns from the May 2022 ICESat-2 Science Team Meeting

- Non-intuitive pricing structures, computing options, infrastructure
- Poor documentation
- Costly to use
- Not obviously more collaborative or faster

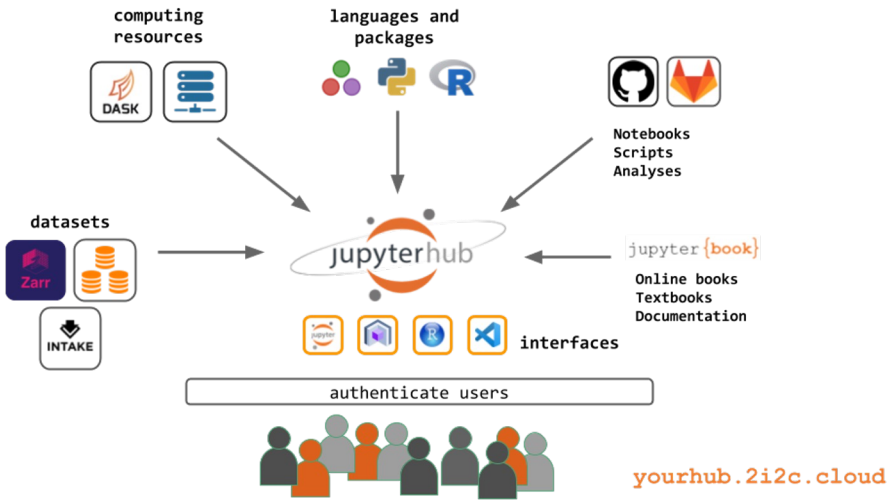


This didn't ring true to our experience in the cloud!

2i2c.org

The International Interactive Computing Collaboration

- **Non-profit.**
- **Service provider** for interactive computing infrastructure.
- An R&D team that **contributes back to open source** communities.



Berkeley Computing, Data Science, and Society

jupyter

binder

PANGEO

Callysto

SYZYGY.CA

Funding (Open Science Program)

CryoCloud: A cloud-computing platform with *bumpers*

Goal: Simple and cost effective managed cloud environment for training and transitioning new users to cloud workflows and determining community best practices

Built and developed for cryosphere scientists by software professionals at **2i2c** to make it possible to:

- Process data faster
- Minimize downloading
- Democratize science

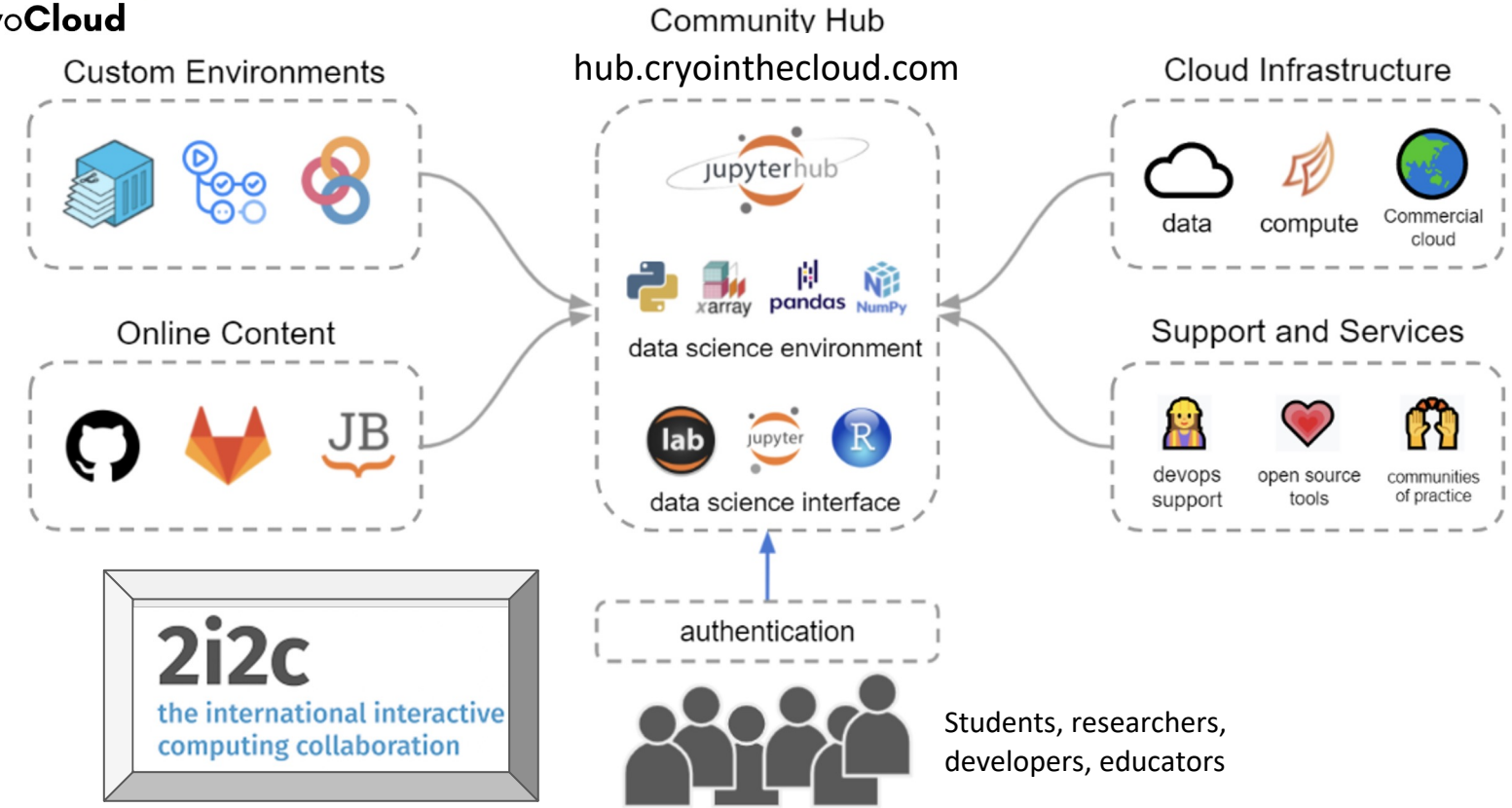
pixabay.com

CryoCloud: A cloud-computing platform with *bumpers*

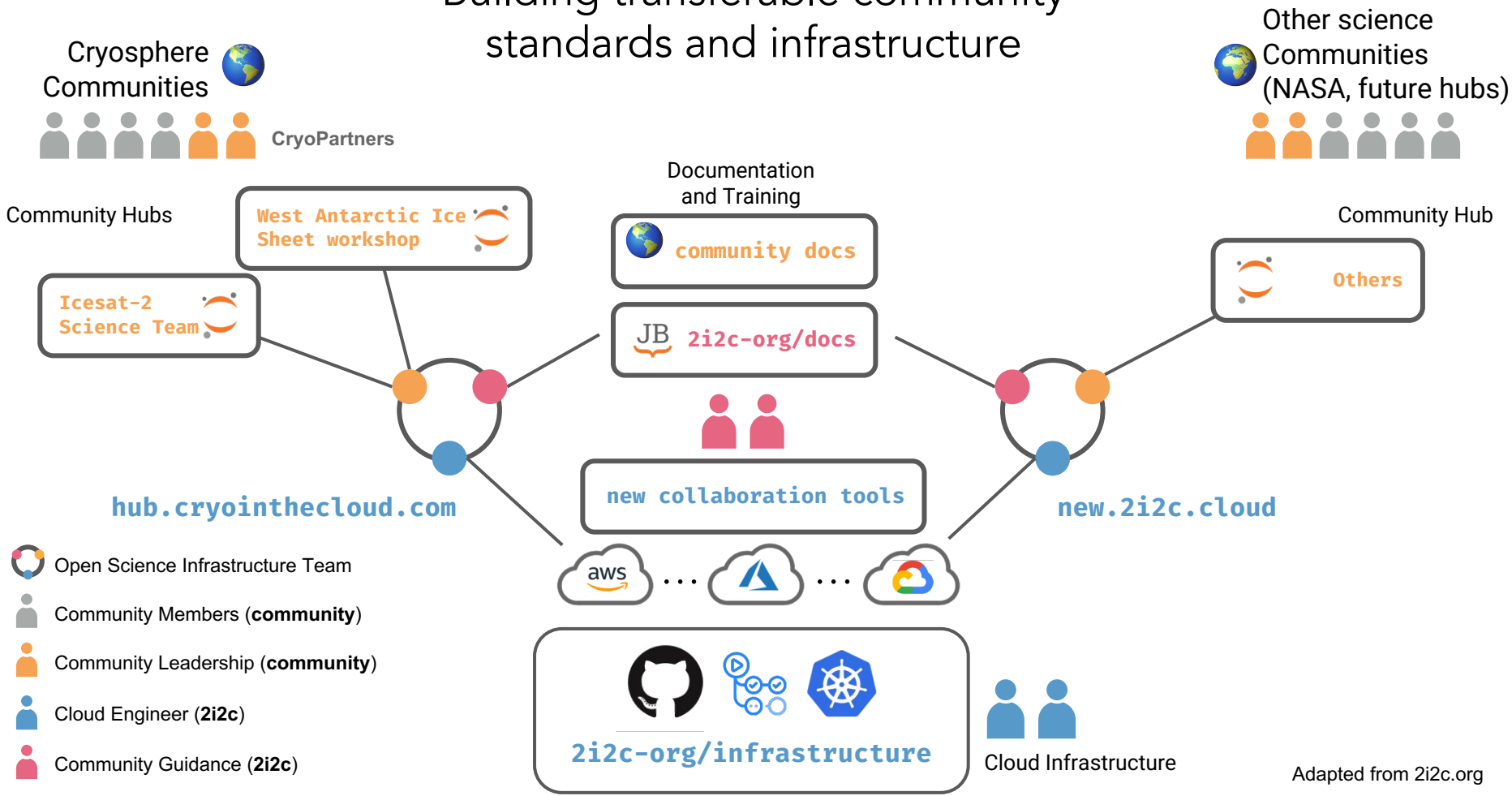
- Persistent for (at least) three years
- Small servers (32 Gb / 4 CPU) for all users, option to bring own cloud credits to access larger servers
- New tool development
  - Personal cost-monitoring tool
  - Intra- and inter-hub collaboration tools
- Helping 2i2c scale with community surveys, feedback, and guidance

pixabay.com

CryoCloud: accelerate discovery and enhance collaboration



Building transferable community standards and infrastructure



Adapted from 2i2c.org

CryoCloud documentation & community building

cryointhecloud.com

- CryoCloud Github: github.com/cryointhecloud

CryoCloud JupyterBook

All the content! More about us, resources, training, and tutorials all found here!

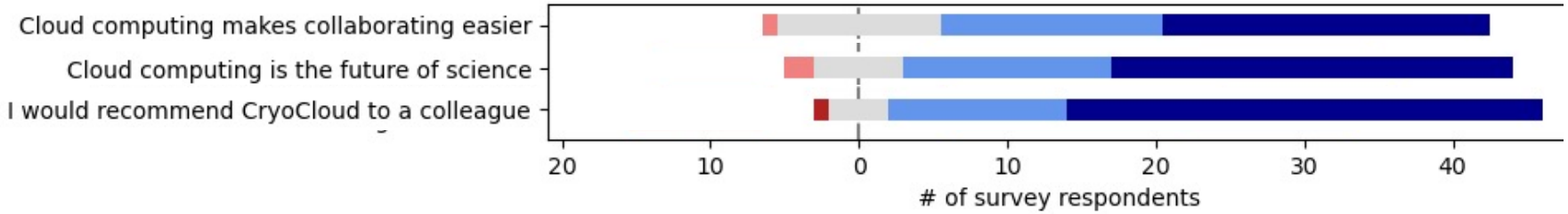
CryoCloud JupyterHub

Get onto the cloud. Our shared cloud platform for NASA Cryosphere communities.

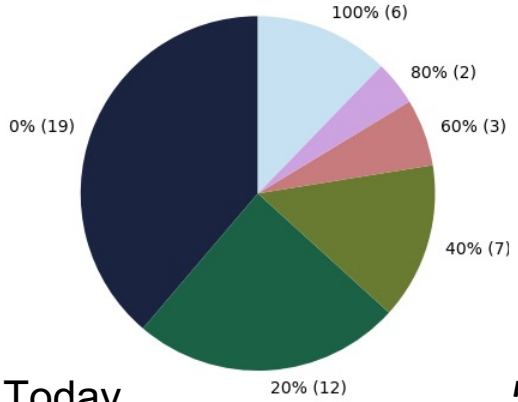
- CryoCloud Slack
- Community office hours
- Training, tutorials, and resources
- Bring in Cryosphere communities and share in infrastructure ideation and construction

Advantages to a community platform

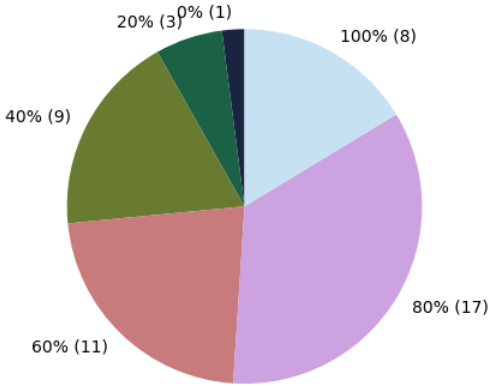
CryoCloud community feedback integral to building community best practices



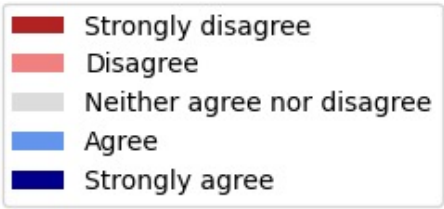
What percentage of your research computing to you perform in the cloud



Today



5 years from now



CryoCloud helps accelerate science and makes open science easy

Faster – No manual data handling, faster data read-ins

Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

CryoCloud helps accelerate science and makes open science easy

Faster – No manual data handling, faster data read-ins

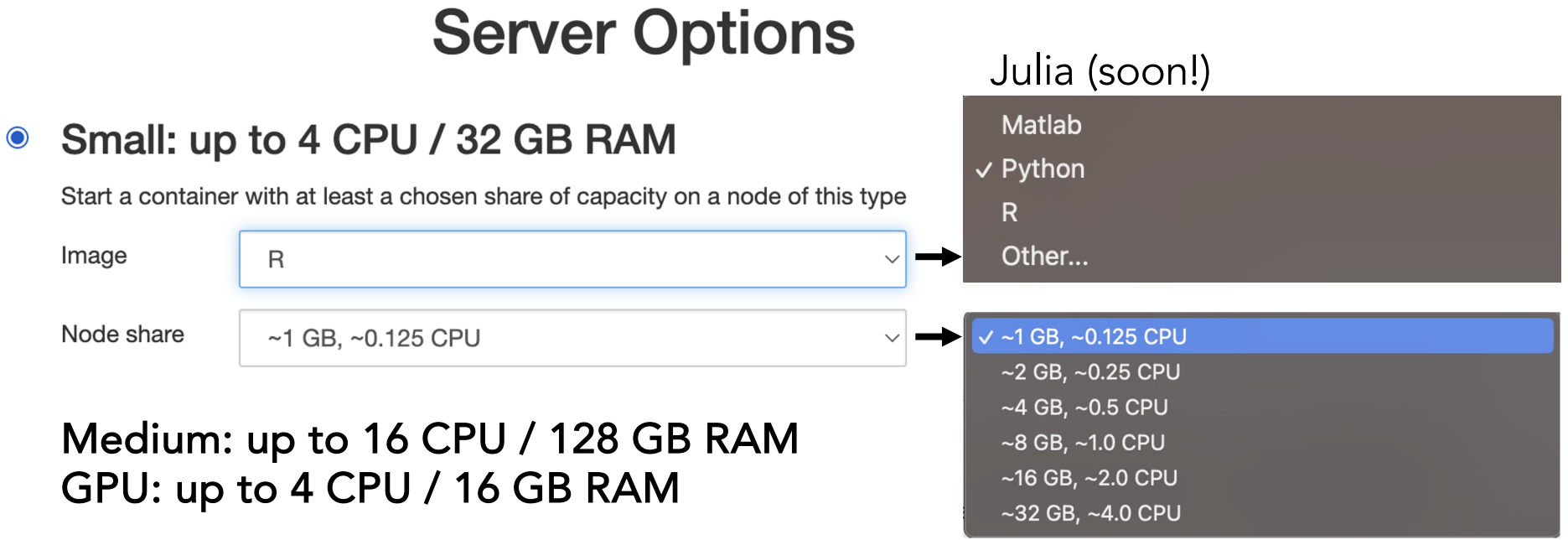
Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

Versatile choice of computer language and server size



Medium: up to 16 CPU / 128 GB RAM
GPU: up to 4 CPU / 16 GB RAM

CryoCloud helps accelerate science and makes open science easy

Faster – No manual data handling, faster data read-ins

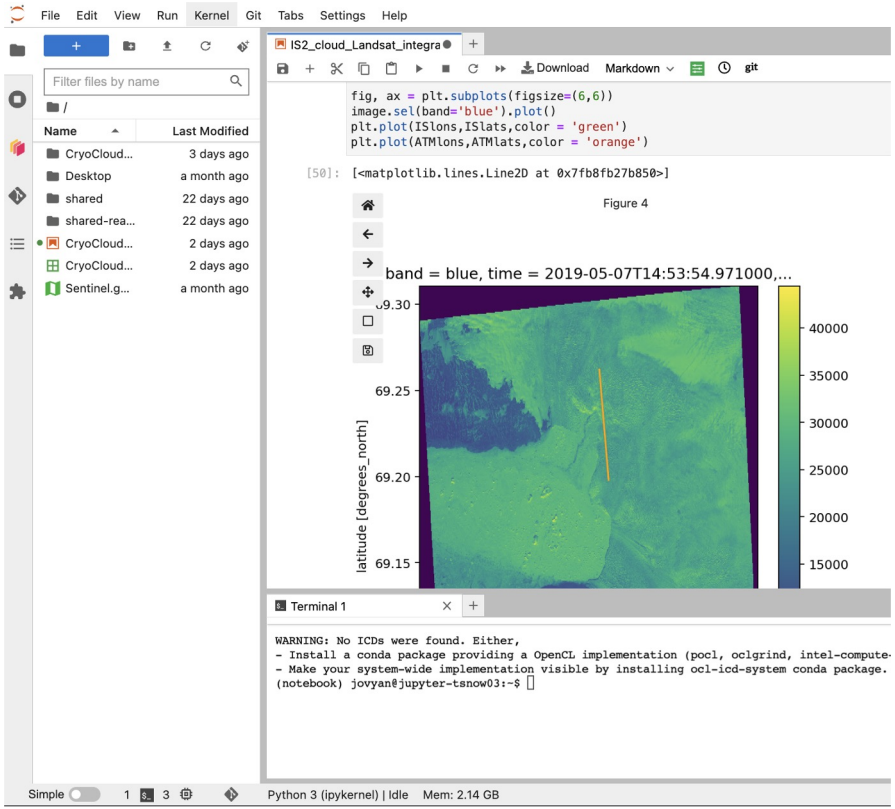
Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

Different kinds of users in one place accelerates feedback and collaboration



Datasets

ICESat-2
ICE, CLOUD, AND LAND ELEVATION SATELLITE-2

NASA MODIS
Moderate Resolution Imaging Spectroradiometer

LANDSAT
LANDSAT-8
USGS • NASA

Tools and Developers

xarray

icepyx

SR

earth access

earthaccess
A Python Library for NASA Earthdata

CryoCloud helps accelerate science and makes open science easy

Faster – No manual data handling, faster data read-ins

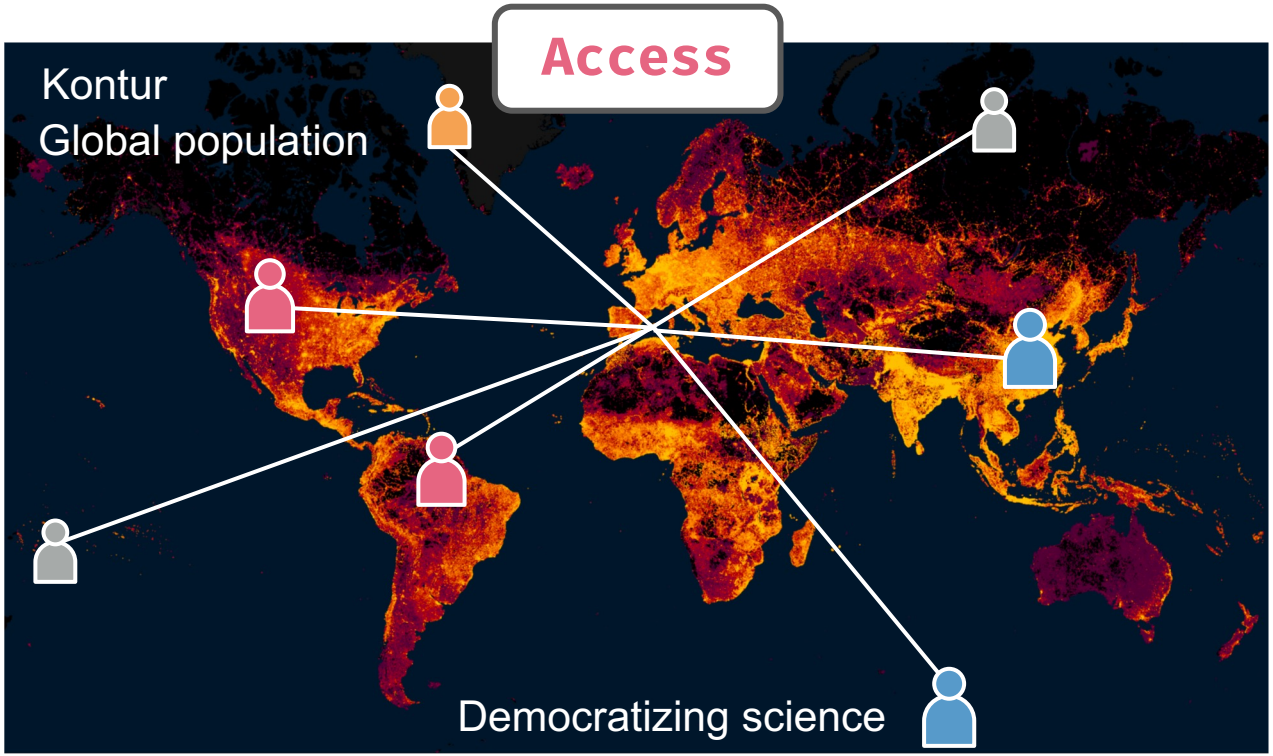
Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

Early careers, underrepresented groups, and non-R1 academic institution researchers benefit the most



CryoCloud helps accelerate science and makes open science easy

Faster – No manual data handling, faster data read-ins

Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

A consistent platform provides a place where individuals and communities can rapidly mobilize when need arises

- ICESat-2 Science Team pre-launch training: Dec 2022
- FOGGS: March 2023
- ICESat-2 Hackweek: Aug 2023
- WAIS: Sept 2023
- ICESat-2 Science Team Meeting: Oct 2023
- AGU Year of Open Science: Dec 2023

Norwegian ML Workshop (Apr), QGreenland Workshop (May), NSIDC User Working Group (Sept), GeoSmart Hackweek (Oct)

Cost to run a workshop in the cloud

Case study: QGreenland Workshop

Standing up own hub for 1 month	With CryoCloud
$4500 for hub	$75 in cloud credits
$75 in cloud credits	2 hours to 2 work days - Science and/or tech experts to advise on user & infrastructure needs
Science experts to advise in constraining hub needs	**Shared resource community model** – expertise and technology
2-4 work weeks - Build and maintain hub	

+

Cost to run a workshop in the cloud

Case study: QGreenland Workshop

Standing up own hub for 1 month	With CryoCloud
~~$4500 for hub~~	$75 in cloud credits
$75 in cloud credits	2 hours to 2 work days - Science and/or tech experts to advise on user & infrastructure needs
Science experts to advise in constraining hub needs	**Shared resource community model** – expertise and technology
~~2-4 work weeks - Build and maintain hub~~	

QGreenland + CryoCloud

Cost to run a workshop in the cloud

Case study: QGreenland Workshop

Standing

loud

$4500 for hu

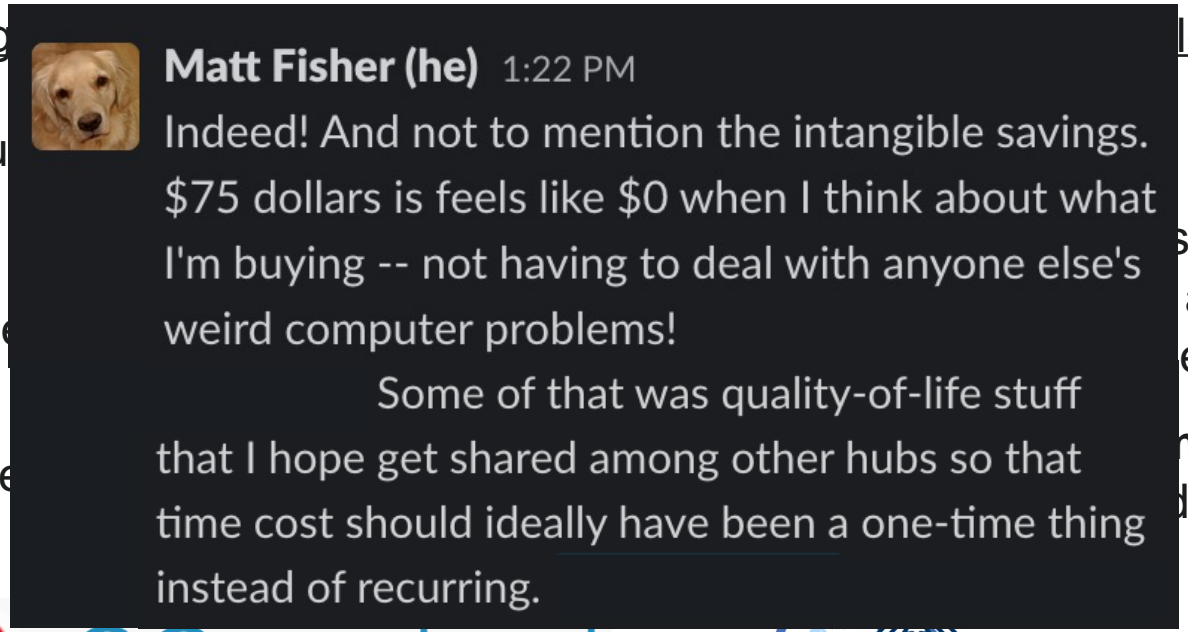
$75 in cloud

Science expe
hub needs

2-4 work wee

s - Science
advise on
eeds

nmunity
d technology



QGreenland + CryoCloud

Self-sustaining ecosystem generated once have a critical mass of experts who drank the coolaid

Faster – No manual data handling, faster data read-ins

Easy to use, customizable – same software as on personal computer

Collaboration made easy – co-coding, shared tools

Eliminates technology bottlenecks – shared challenges

No software expertise needed – cloud-computing as a service

Ongoing challenges

Data not in the cloud or cloud-optimized formats

Boundaries of what a community is

Imagining new social and logistical structures

- Models for contributing to sustain the infrastructure and costs
- Interconnection between computing systems

Vision for the future

Open science values

Intellectual generosity

Intellectual humility

Right to participate in science

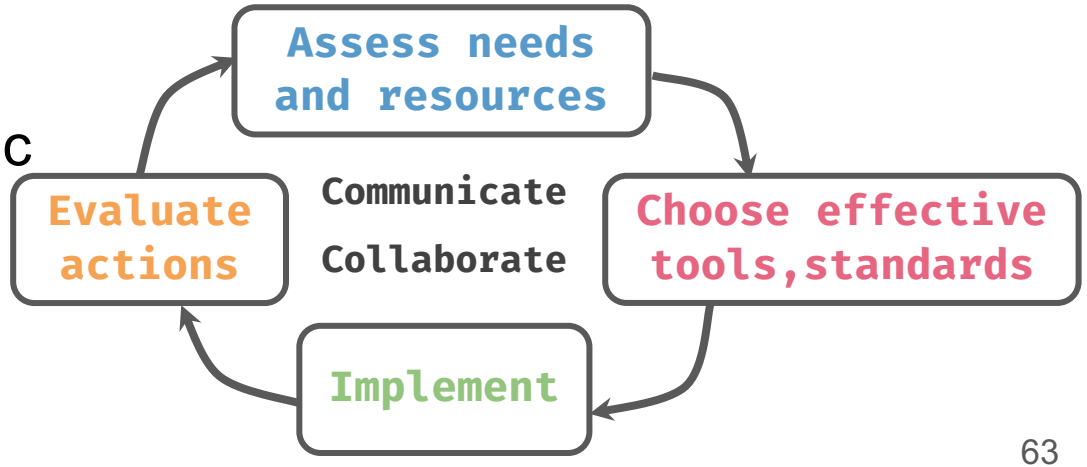
Everyone deserves to be treated with dignity and respect

open.science.gov

Guiding practices

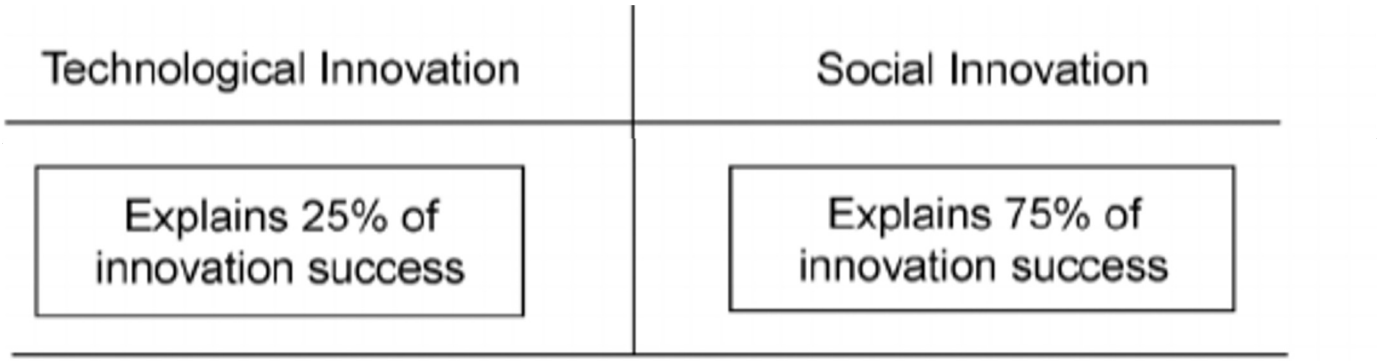
Avoid vendor lock-in

Active engagement of scientific community in building new community models



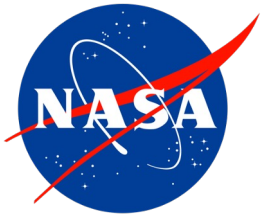
Driving social and technological innovation forward

Open science as a process, not a product



Source: Erasmus University: Competition and Innovation monitor (2006)


Interconnected systems
Interconnected researchers



cryointhecloud.com

Thank You!

tsnow03.github.io

@tsnow03 @TashaMSnow tsnow@mines.edu