PEARC22 BOF

July 12, 2022

Reproducibility and Trustworthiness of Scientific Research

H. Birali Runesha, Ana Trisovic, Sandra Gesing and James Wilgenbusch

Our BoF today …

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Questions and answers

BoF notes: https://tinyurl.com/ynpz3yks

Goals

This BoF will discuss opportunities and challenges for developing support services to expand the user base, lower barriers for capturing artifacts while doing research, and brainstorm how to work as a community towards a concerted effort to build an ecosystem of tools to support reproducibility.

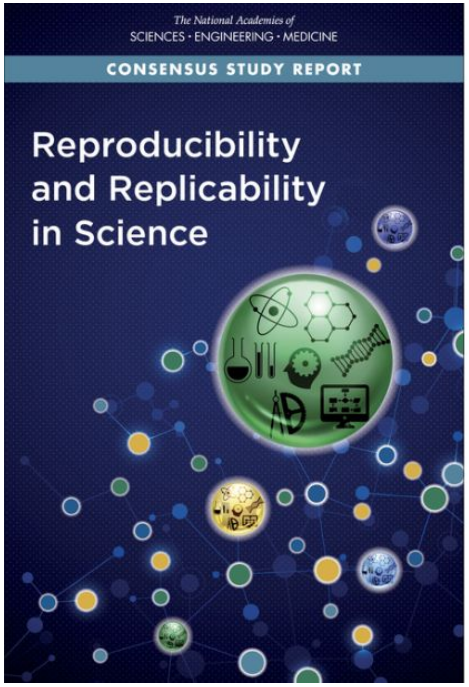
BoF notes: https://tinyurl.com/ynpz3yks

Reproducibility and Replicability

- **Reproducibility**: Obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis

  -> not working on **numerical reproducibility**

- **Replicability:** obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data



Source: National Academies of Sciences, Engineering and Medicine. 2019. **https://doi.org/10.17226/25303**

Improving Trustworthiness of Computational Results: Opportunities for the NSF Office of Advanced Cyberinfrastructure to address recommendations from the National Academies Report on Reproducibility

"OAC Reproducibility Opportunities Report"
Draft for Comment Summary

Wolfgang Bangerth, Juliana Freire, Patrick Heimbach, Michael Heroux (chair), Ivo Jimenez, Ellen Rathje, Hakizumwami Runesha, Victoria Stodden

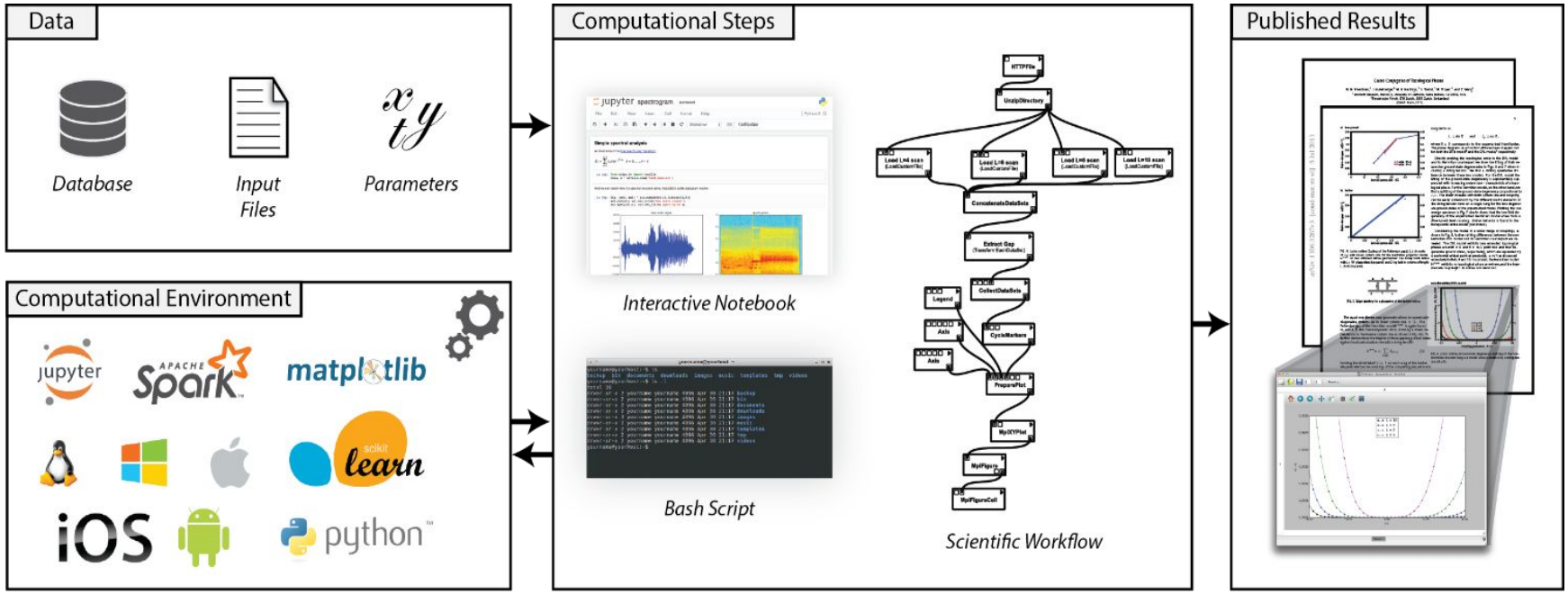
Version for community comment:

https://docs.google.com/document/d/1d7kJ28-m8xxtrXQbTodKfFmDiR11uJto1jFb2h_w7bY/edit?usp=sharing

Vision for Trustworthy Computational Science

We look toward a future for computational science where all computational results are reproducible, including those from pipelines across multiple teams. Effective and efficient reproducibility will enable qualitative advances in science and make possible a new level of demonstrable trust in scientific results and outcomes.

Source: Michael Heroux presentation 2021-09-24



PROVENANCE

Description of Data + Computational Steps + Description of Environment

REPRODUCIBLE EXPERIMENT

SOURCE: Fernando Chirigati & Juliana Freire

What are the questions?

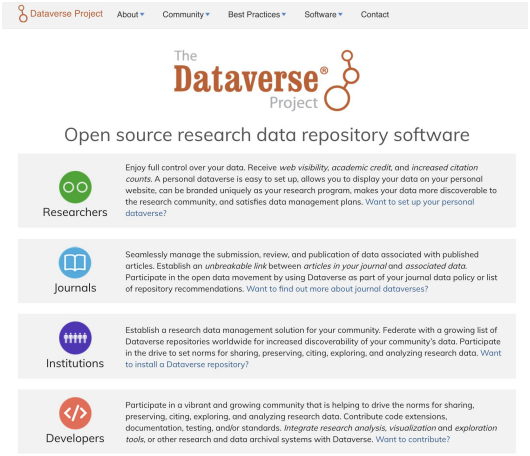
- Can you **reproduce the results** of a scientific research in a published paper?
- Do you **have enough information** to allow you to reproduce the results?
- Research **takes time** before getting results. Do we collect enough information **while doing research** to facilitate the reproducibility of the final results?
- Reproducibility is **hard and can be labor intensive**. How do we minimize the manual effort required to put together the artifact to be shared?
- etc.

Do existing tools and repositories fully address the reproducibility question? Are we using what is available today?

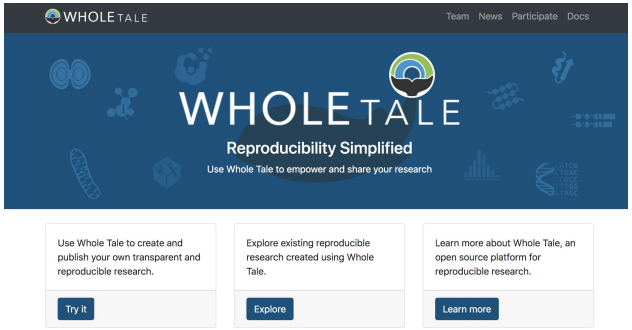
Examples of existing tools

- **Project jupyter**: web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- **Containers:** software to package applications allowing them to be portable to any system running a Linux OS. It captures necessary system dependencies and vastly help with reproducibility
- **Github/Gitlab**: a web-based version-control and collaboration platform for software developers
- **Globus:** software for transferring and sharing files. It is also used to build applications and gateways
- **Digital Object identifier (DOI):** is a persistent identifier or handle used to identify digital objects uniquely.
- etc.

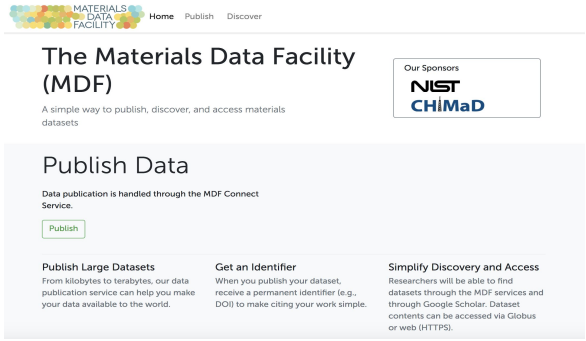
Example of initiatives



dataverse.org

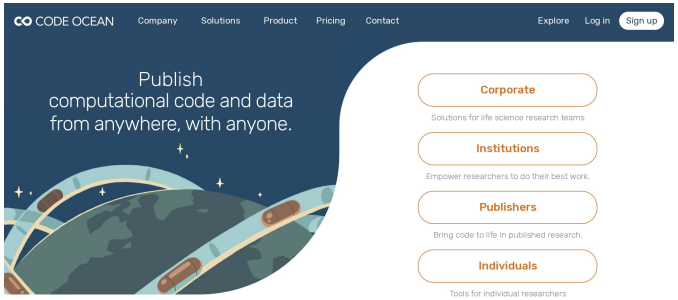


wholetale.org

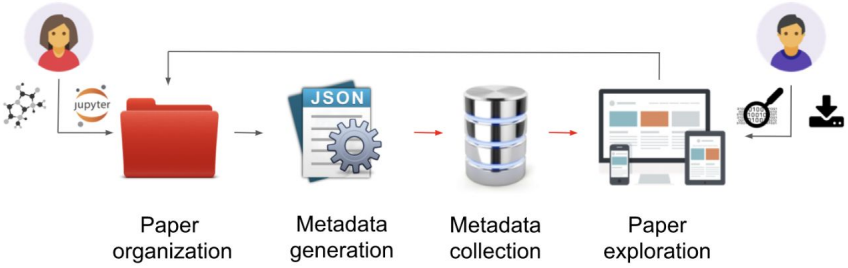


materialsdatafacility.org

etc.



codeocean.com



qresp.org

Observations

- Some existing tools are addressing **parts of the reproducibility question**
- some are focusing on reproducible papers, i.e. allowing **interaction with final results** of research, rather than **an entire workflow**
- some projects focus on **project management** rather than computational reproducibility
- Some require users to **manually** define workflows of their research (not automatically)
- eb-based tools that support the use of containers
- the artifact needed for reproducibility is put together **after the research is completed**
- Some are enhanced **repositories.**

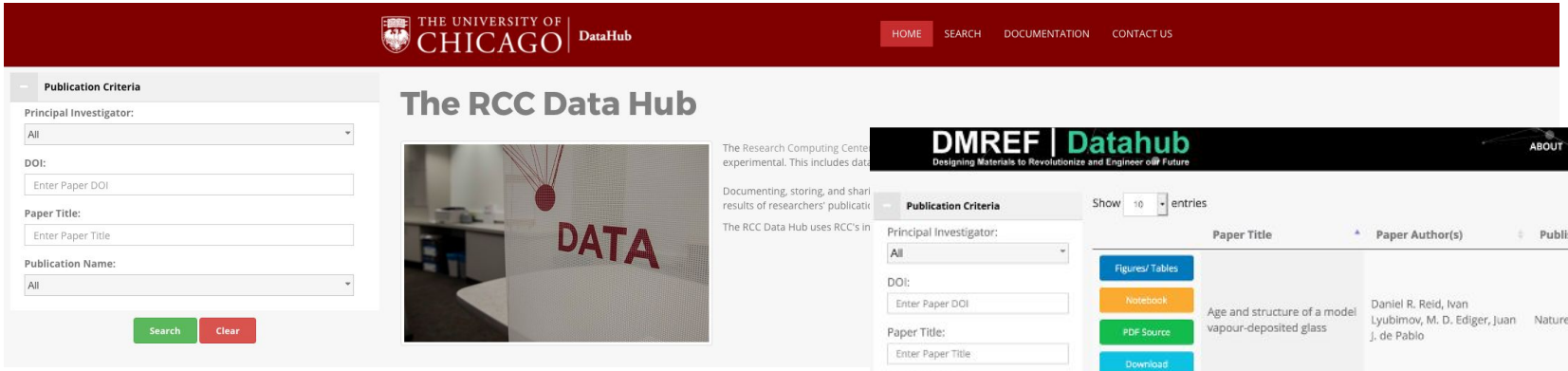
RCC Data Hub

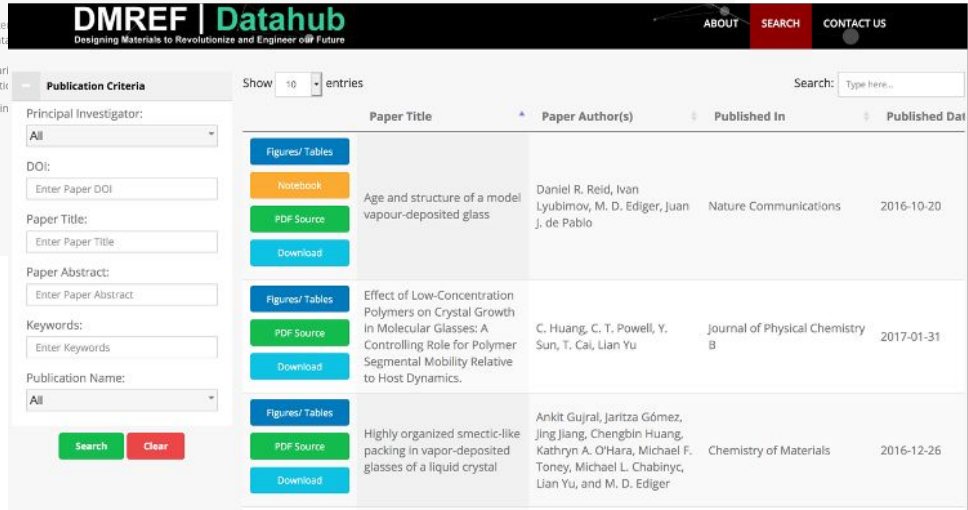
https://datahub.rcc.uchicago.edu/

A **data portal** to search, view and download workflows, tools, documentation, and all data sets needed to **reproduce the results of a scholarly work.**

Aims:

- **Advance the openness of all scientific data** produced throughout the life cycle of a project for compliance with funded research grants and accelerated productivity.
- **Increase the integrity and reproducibility of scientific results.**





What are we trying to achieve?

As a community, can we work together to build an ecosystem of tools and service to help researchers capture and collect metadata/info about the methodology, data, software, tools, platform, etc. , associated with results, while doing their research with the least amount of effort?

Rethinking how to support Research Computing and training

We need to …

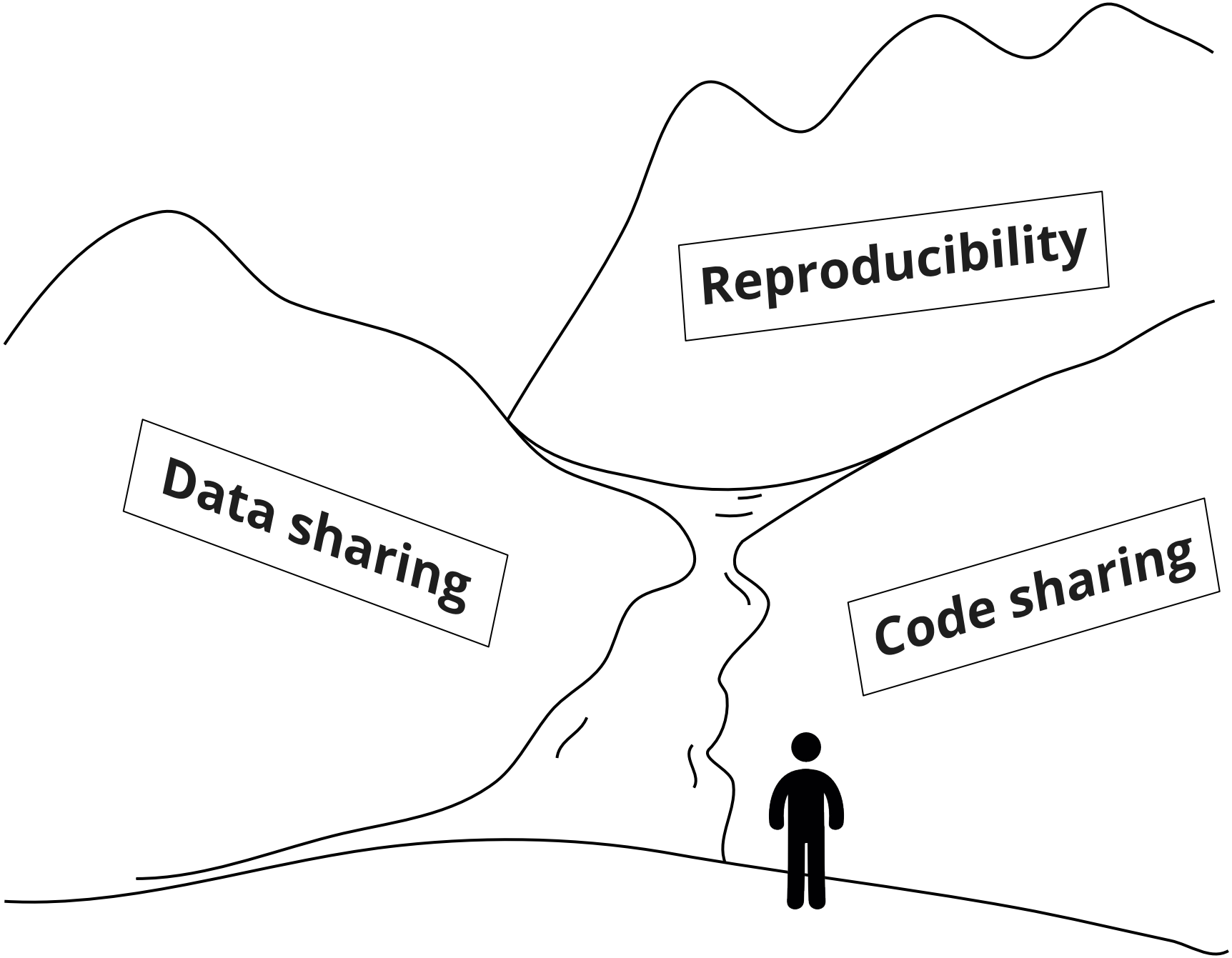
- platforms and tools that facilitate reproducibility
- change how we currently conduct research
- develop new training and services

Our BOF today …

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Questions and answers

Prerequisites: FAIR + R

- Reproducibility requires some level of sharing (doesn't need to be open access)
- FAIR principles - guidelines for sharing data and, more recently, research software
  - Data repositories and software repositories implement features that meet FAIR principles
  - Repositories aim to facilitate computational reproducibility


Reproducibility
Data sharing
Code sharing

Research repositories

Provenance capture

Encapsulation tools

Workflow engines

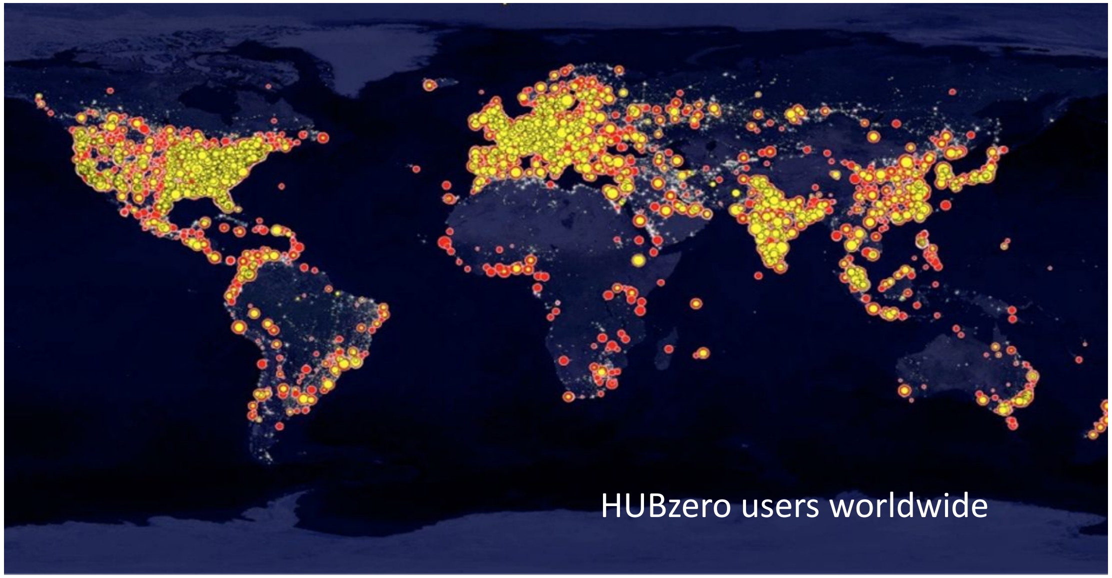
Science Gateway Technologies

- Widely used complete frameworks (HUBzero, Open Science Framework, Galaxy, Globus Data Portal, etc.)
- [illegible]

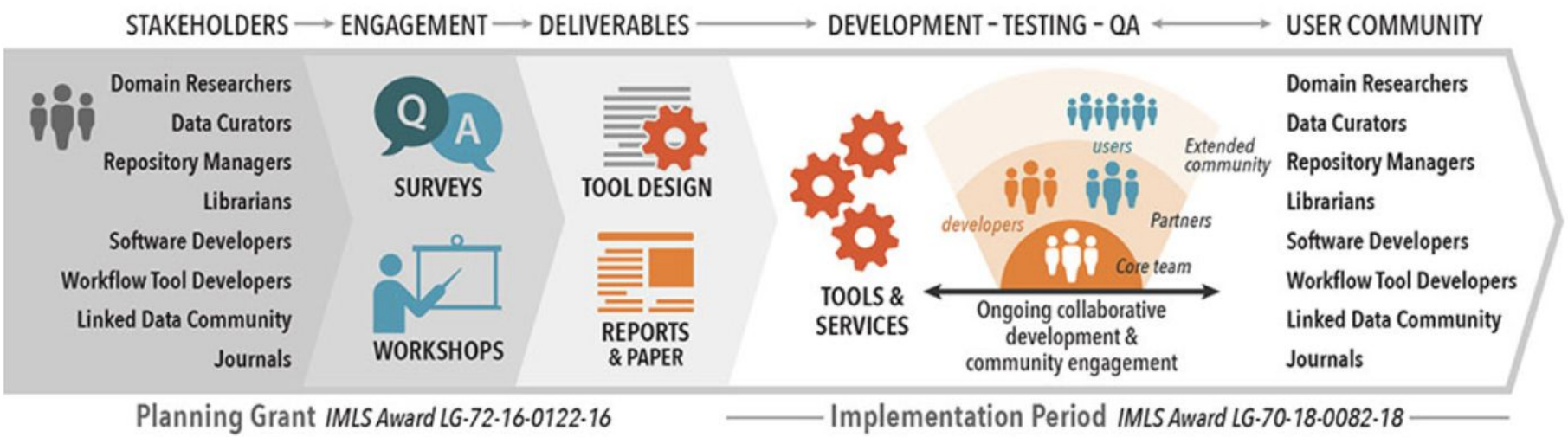
Predestined for sharing data and computational methods and reproducibility

in ONE instance

Sharing between different instances and technologies is complex



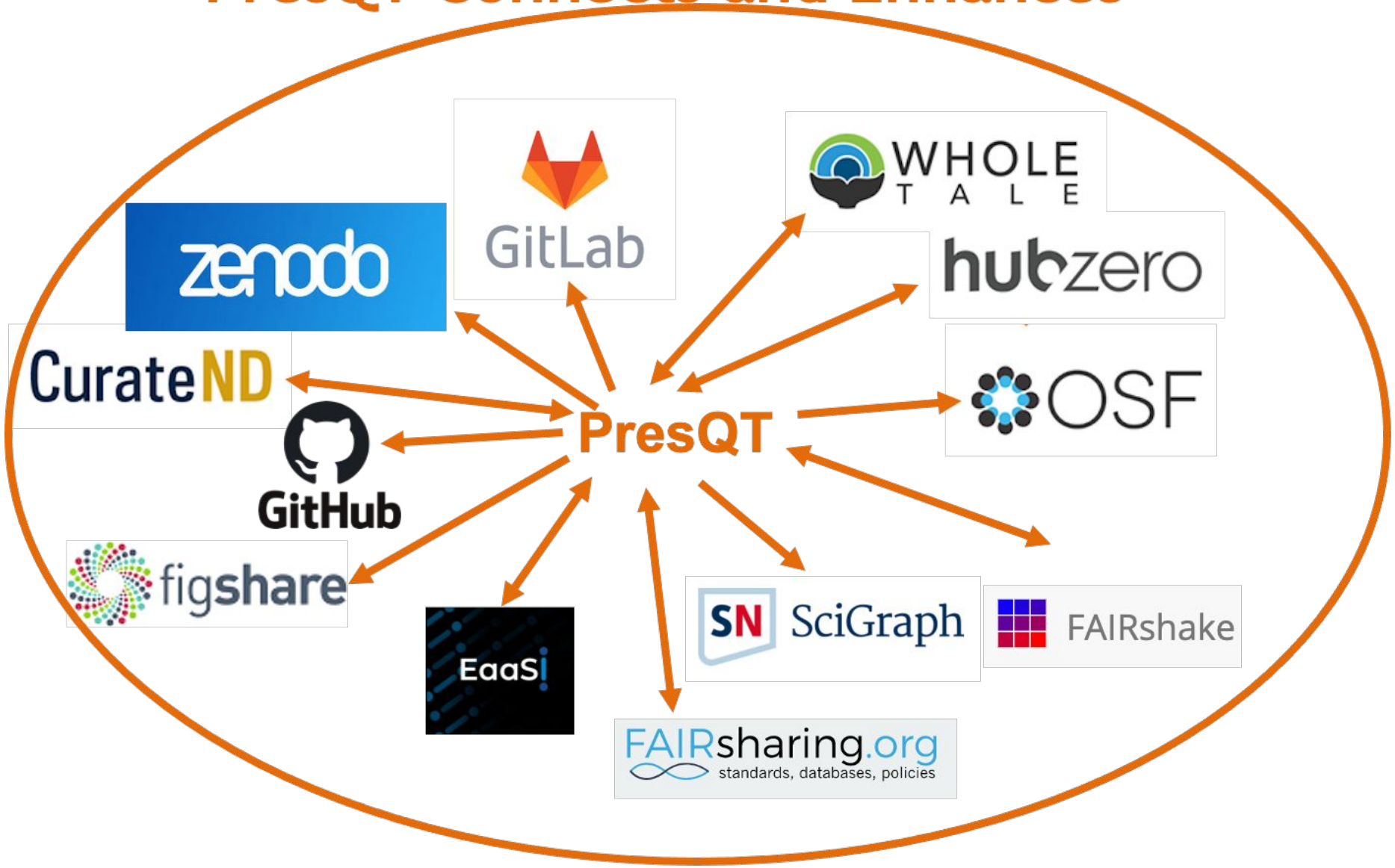
PresQT Connects and Enhances



Concept

- **not** standalone solutions
- partner systems and services easily integrable via RESTful APIs and services
- user-centered open design and collaborative development

PresQT Connects and Enhances

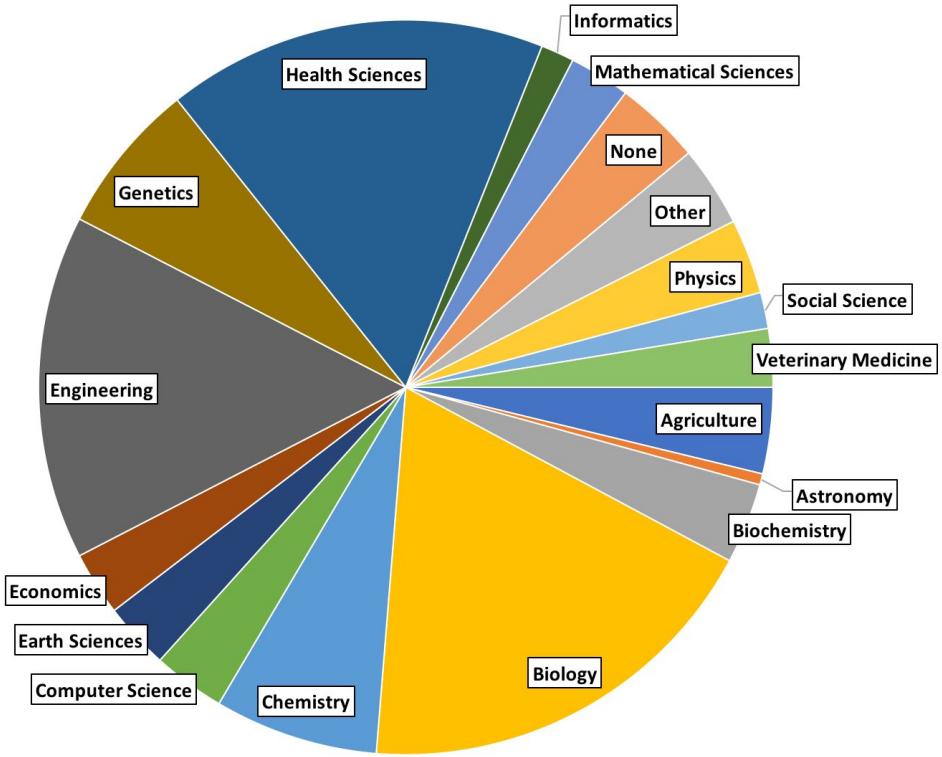


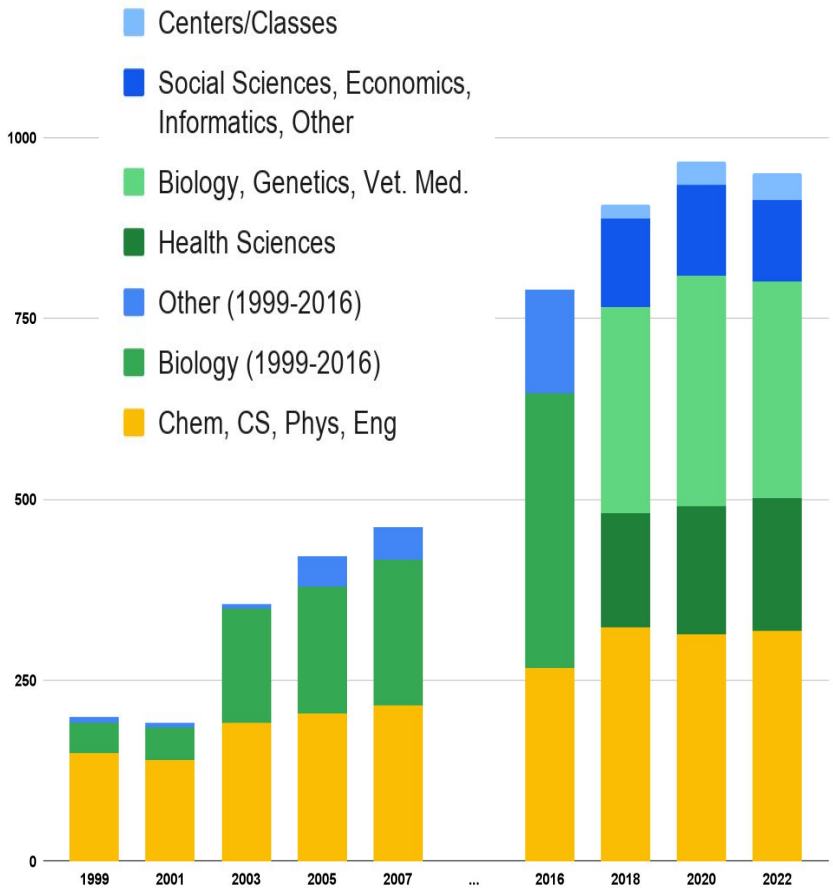
https://presqt-prod.crc.nd.edu/ui/

https://presqt.readthedocs.io/en/latest/

The Long Tail of Research Computing

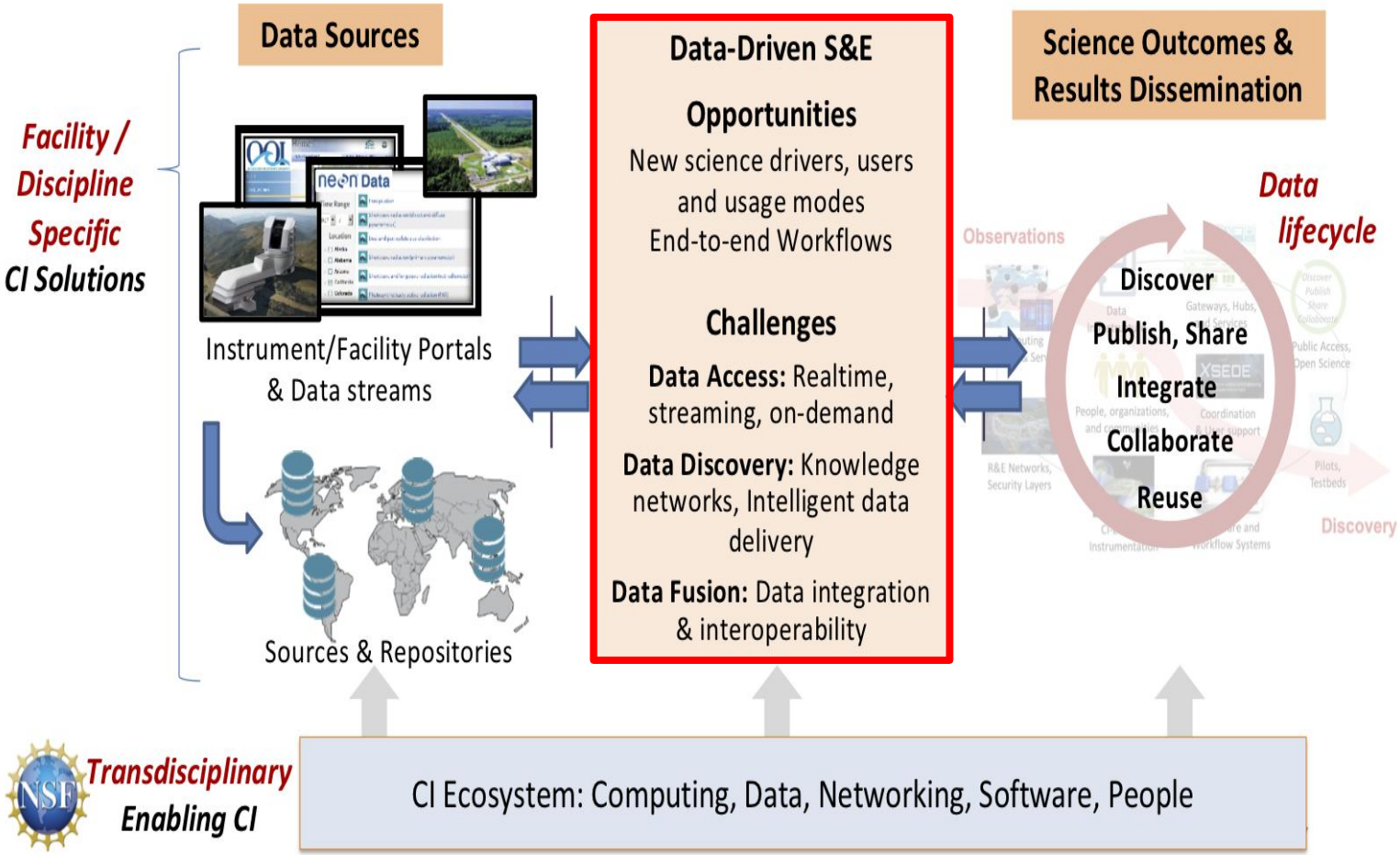
Groups in 2020: 900 User Groups
4,555 Active users





Biggest increasing in Life Sciences

Data-Intensive Discovery Pathways – The "missing middle"



*Slide Credit: Manish Parashar, Office Director of the Office of Advanced Cyberinfrastructure (OAC) at the National Science Foundation (NSF) presented at the Fall Midwest Big Data Hub All hands Meeting, October 30, 2019.

Developing the "Missing Middle"
System-wide View of the Research Computing

Data Generators and Repositories

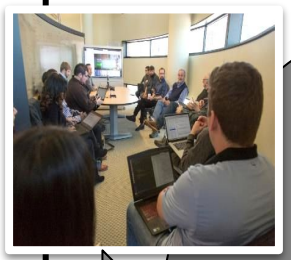
AgroInformatic Institute - GEMS

Masonic Institute for the Developing Brain

Science Outcomes & Result Sharing

Univ of MN Genomic Center

UMN Core Facilites



Research Computing

Publications

UMN Institutes & Centers

Center for Magnetic Resonance Research

MSI

UMII

Workshops, Conferences, and Training

U-Spatial

Curated Data Resources

National Institute of Mental Health Data Archive

Global & Local Data Archives

McGill Centre for Integrative Neuroscience

Institute for Health Informatics

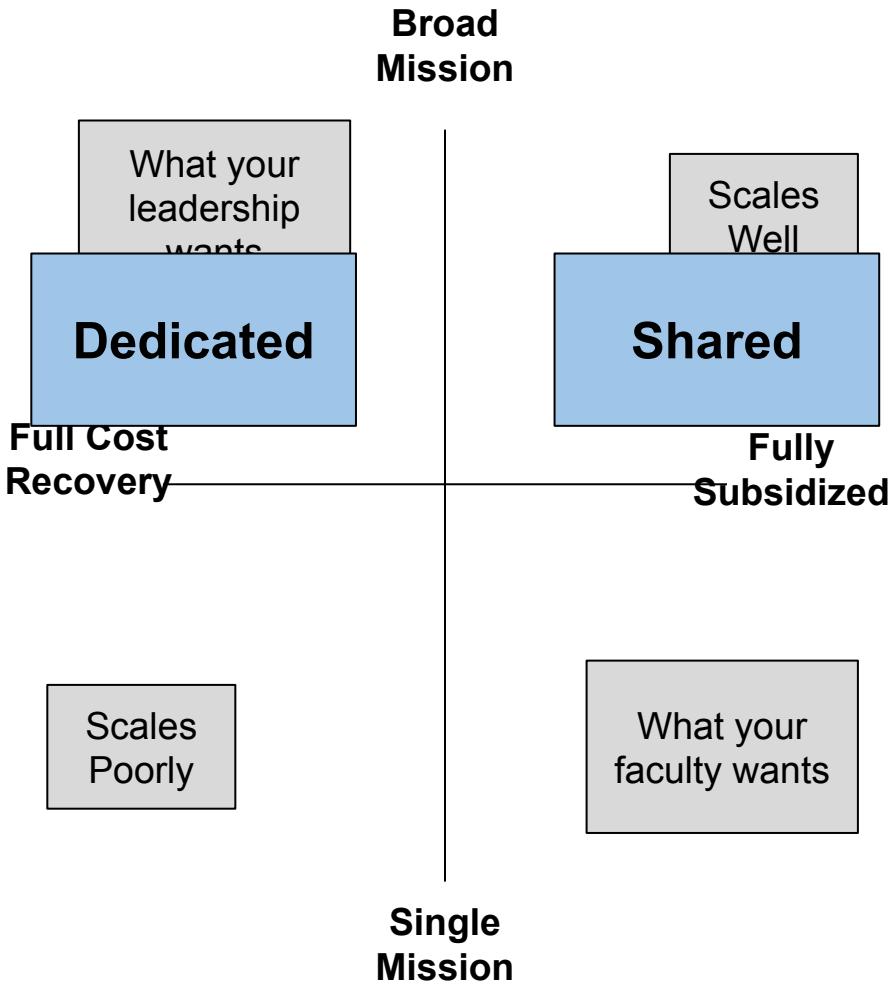
Tools & Platforms for Analysis & Sharing

Key Ingredients to Reproducibility

Things

Scalable, multipurpose systems infrastructure

- **Shared** - Available to everyone on a first come first serve basis
- **Dedicated** - Dedicated to single research group or project

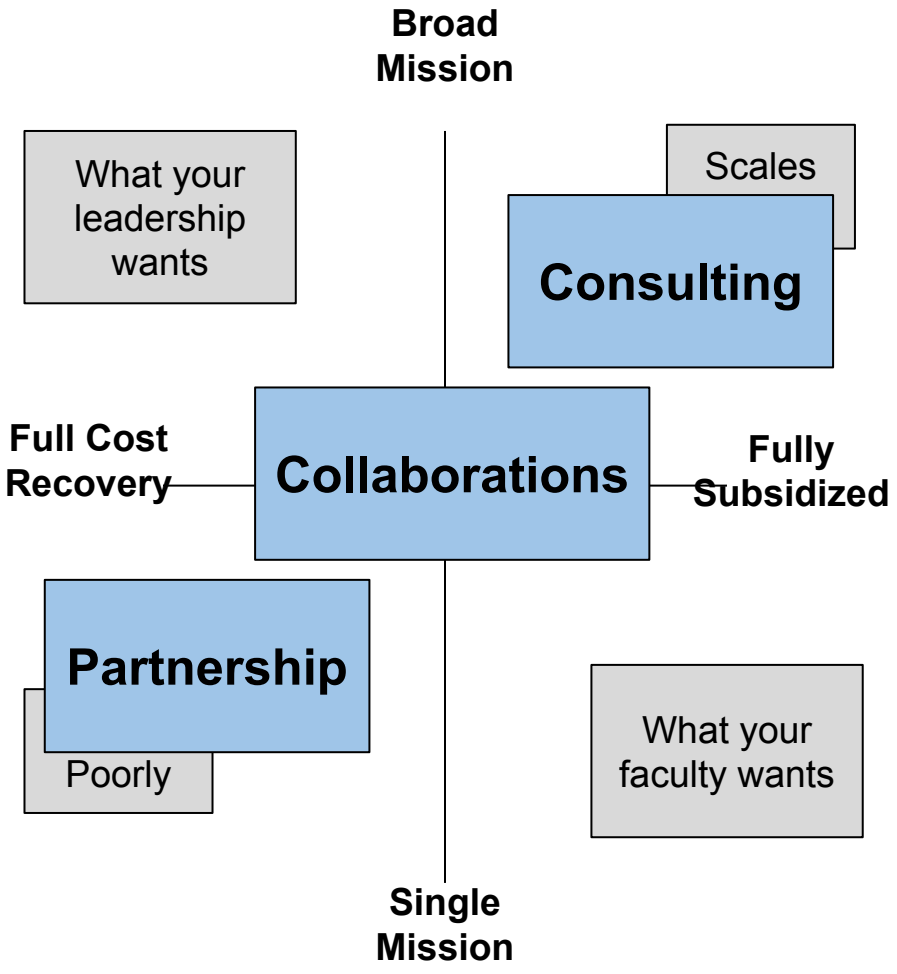


Key Ingredients to Reproducibility

People

Scalable, multipurpose staff experts

- **Consultations** - No direct cost. Short engagements (<1 hour)
- **Collaborations** - Shared investments in emerging or locally specialized areas of research. Weeks to months of investment
- **Partnerships** - Cost recovery, well defined agreements, and longer term commitments for funding areas of research that are generally considered to be core UMN strengths.



Our BOF today …

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Discussion – Questions and answers