November 8, 2023

The Importance of Research Cyberinfrastructure Expertise in Modern Scientific Research

Dr. Blake Joyce
Senior Scientific Consultant

bioteam.net

BioTeam
Accelerate Science

Talk Overview

A (brief) description of how I came into the research computing space

- Descriptions of new roles emerging in research computing
  - This will not be exhaustive
  - There's a lot of new roles developing!
- Discussion of some services those people provide in a research computing setting

A Quick Definition of Research Cyberinfrastructure (CI)

- Here we'll be talking about research computing resources
  - Analysis and data management software
  - Networking equipment for data transfer
  - Computing hardware
- I'll define research cyberinfrastructure (CI) as anything that isn't a local computer (laptop, desktop)
  - I would normally include these in research CI
  - But here we need a convenient term for anything beyond these resources

A Portrait of the Artist as a Young Scientist

I am a Joyce after all…

quest.utk.edu/2012/blake-joyce/

Quest

Research, Scholarship, and Cr

HOME ABOUT TOPICS SCHOLARS BOOKS ARCHIVES

SCHOLARS

Blake Joyce

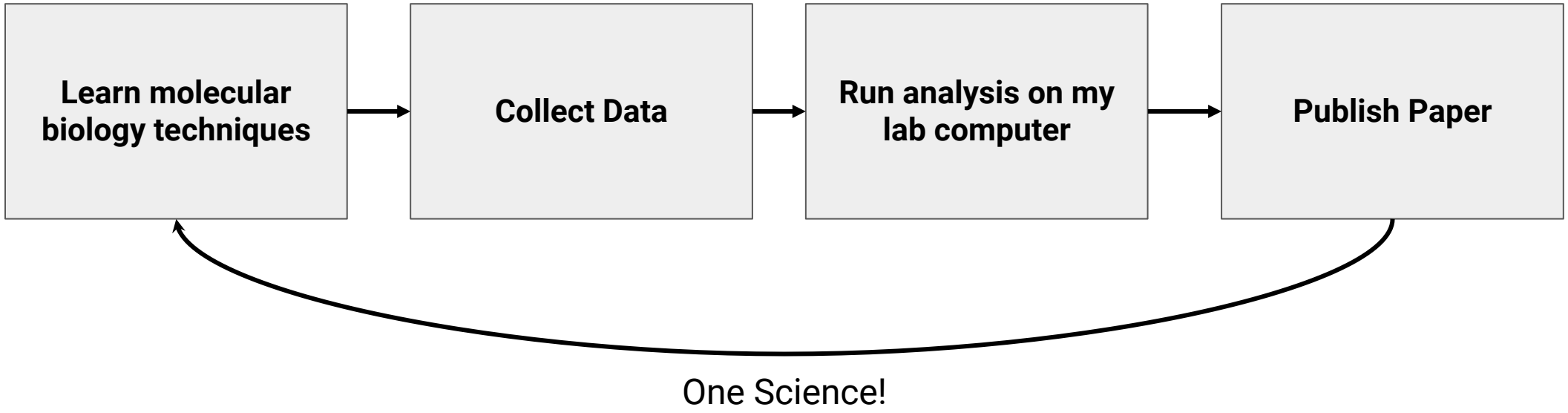
Blake Joyce

Blake Joyce, graduate student in the Department of Plant Sciences, successfully completed his MA project on fern biotechnology and then made a significant leap into biofuels for his PhD work in the area of plant molecular genetics. During the past three years, Blake has published sever papers, including an invited review paper on biofuels that was published last year online in *Biotechnology Advances*. In addition, he has been an invited speaker to several international conferences and received competitive funding fellowships to attend these meetings. In just the past 12 months, these include meetings in San Diego, CA, Raleigh, NC, and Vienna, Austria.

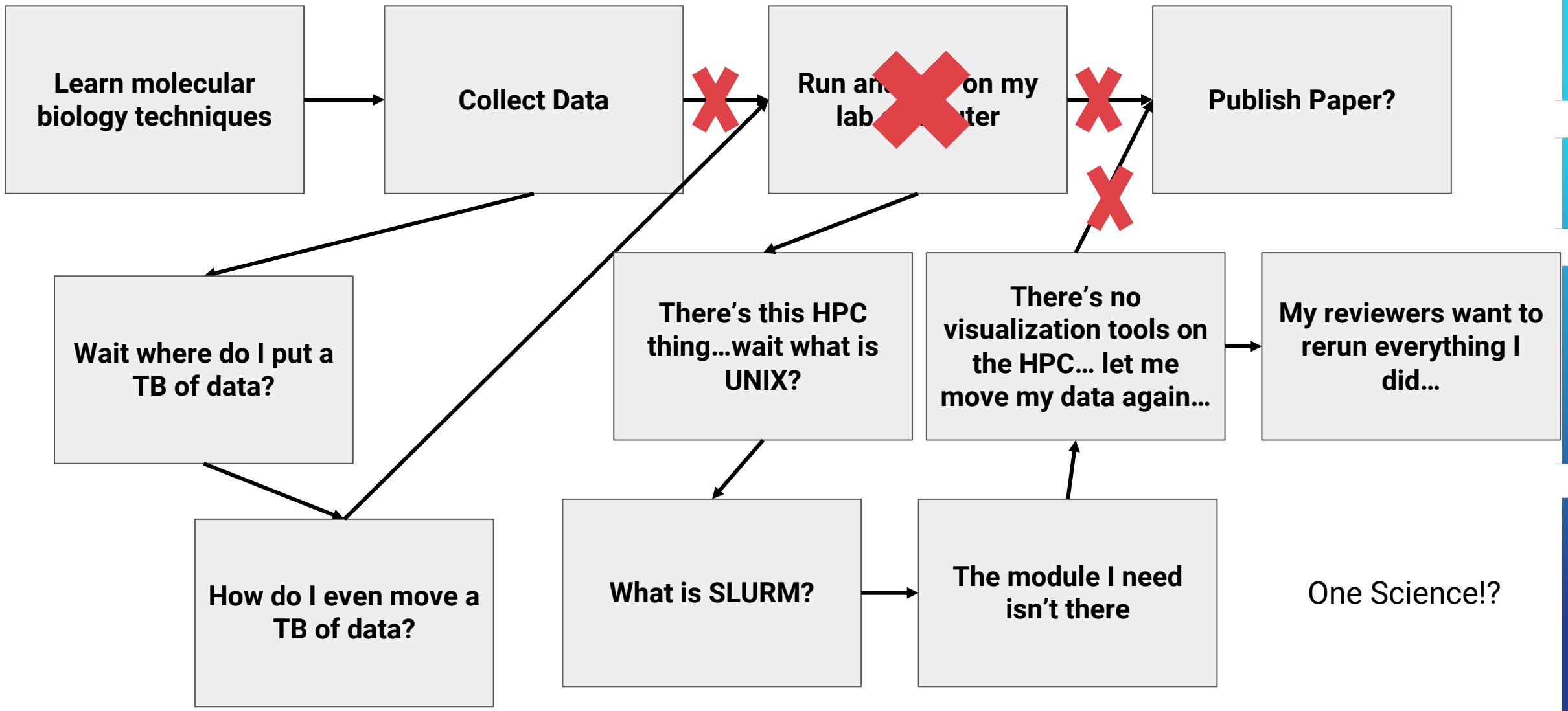
For Blake's PhD project, he has literally searched the world for plants that produce compounds that have the appearance of gasoline and diesel fuel. These include the diesel tree (Brazil) and the petroleum nut (Philippines). He has worked to fund his own research and travel, including being the motive force of a $80,000 Science Alliance grant to work with ORNL researchers to analyze the fuel characteristics of these plant compounds, a $30,000 Sun Grant project on lemongrass and palmeroso with a University of Wyoming professor, and most recently a $100,000 award from UTIA that pulls together a great number of researchers from th Philippines to Palestine and also a private company to pursue the genomics and biochemistry of these unique fuel producing plants.

We want to understand exactly how these plants produce compounds that look like gasoline an diesel so we can transfer those properties to bioenergy crops that can be grown in Tennessee.

The Traditional Research Approach: Everyone Learn Everything!



But There is a Lot of Data in Research Now…



One Science!?

The Problem with that Traditional Thinking

- Researchers CAN learn anything (probably)
- But you really want researchers doing research
  - That includes understanding the data being used!
  - That includes understanding what analysis tools are doing to their data!
- Asking researchers to get a research PhD + a comp sci PhD drives talent away from research
- The few unicorns that can do many different disciplines (well) will be rare
- Having to context switch from research to software to infrastructure usually leads to brittle code and single-use infrastructure

It's Complicated to Be a Researcher (Now?)

- Computing is a matter of fact in modern research
- Research is probably not suddenly complicated (it always was)
- Instead: **Research (and CI as a result) is becoming a team sport**
  - Data Scientists and Data Engineers (difference?)
  - Research CI Facilitators to help other researchers learn how to use computing resources
  - Research Software Engineers to support software and 'middleware' code
  - Cyberinfrastructure Professionals to support evolving research infrastructure

If you want to go fast,
go alone.

If you want to go far,
go together.

Data Scientists

“Use computational and mathematical tools and create workflows to analyze data to create knowledge for a domain of research”

Van Tuyl, S. (Ed). (2023). Hiring, Managing, and Retaining Data Scientists and Research Software Engineers in Academia: A Career Guidebook from ADSA and US-RSE. Zenodo. doi.org/10.5281/zenodo.8329337

Instant Digital Transformation: Just Add Science

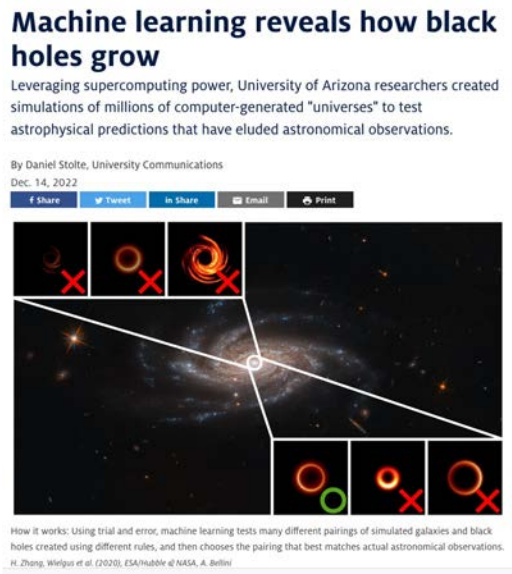
- I won't spend a lot of time defining data science; it's still evolving
- My understanding: originally an alternate term for computer science
- I take a 'you'll know it when you see it' approach here

So let's talk some examples of data science instead

Van Tuyl, S. (Ed). (2023). Hiring, Managing, and Retaining Data Scientists and Research Software Engineers in Academia: A Career Guidebook from ADSA and US-RSE. Zenodo. doi.org/10.5281/zenodo.8329337

The Answer to Life, Millions of Universes, and Everything

- Each simulation is the universe forming with 12 million galaxies inside for 400 million years
- Needed massive computing resources to search a very large parameter space



Machine learning reveals how black holes grow

Leveraging supercomputing power, University of Arizona researchers created simulations of millions of computer-generated "universes" to test astrophysical predictions that have eluded astronomical observations.

By Daniel Stolte, University Communications

Dec. 14, 2022

Share Tweet in Share Email Print

How it works: Using trial and error, machine learning tests many different pairings of simulated galaxies and black holes created using different rules, and then chooses the pairing that best matches actual astronomical observations.

H. Zhang, Wielgus et al. (2020), ESA/Hubble & NASA, A. Bellini

Virtual 'Universe Machine' Sheds Light on Galaxy Evolution

By creating millions of virtual universes and comparing them to observations of actual galaxies, a UA-led research team has made discoveries that present a powerful new approach for studying galaxy formation.

By Daniel Stolte, University Communications

Aug. 9, 2019

Share Tweet in Share Email Print

Peter Behroozi, Risa Wechsler, Andrew Hearin, Charlie Conroy. 2018. UniverseMachine: The Correlation between Galaxy Growth and Dark Matter Halo Assembly from z=0-10. doi.org/10.48550/arXiv.1806.07893

As Above, So Below

- Biology has nearly limitless inward universes
  - Human genome is 1064 ‘Harry Potter Equivalents’ (text from all 7 books)
  - We each have somewhere between 15-70 trillion cells
- The research CI associated with universe simulation or genetic simulation is fundamentally the same as drug design, docking simulations, etc

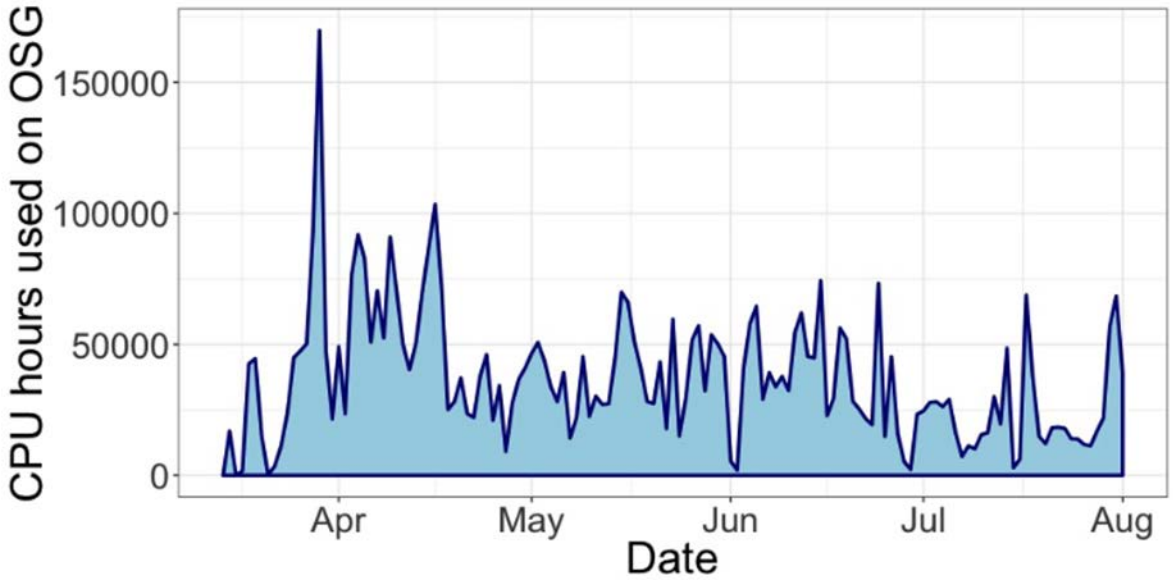


Figure 5: Number of CPU hours used per day on OSG [9] using the Pegasus [8] workflow. CPU hours are equal to wall hours as each job uses one CPU. Over five and a half months, a total of 5.19 million CPU hours were used on the OSG. On average, ~37,000 CPU hours were used per day with a standard deviation of ~25,000 CPU hours per day. The number of CPU hours used per day varies based on the availability of OSG resources. Data obtained from the Open Science Grid Connect dashboard (https://gracc.opensciencegrid.org/dashboard/snapshot/8FQeqrYWIQ7Dpp1wwYcvQ6mgR1Nq6C3o) and are available on the GitHub (https://agladstein.github.io/SimPrily/).

Ariella L. Gladstein, Consuelo D. Quinto-Cortés, Julian L. Pistorius, David Christy, Logan Gantner, Blake L. Joyce. SimPrily: A Python framework to simplify high-throughput genomic simulations. SoftwareX: 7: 2018, 335-340,ISSN 2352-7110, doi.org/10.1016/j.softx.2018.09.003

CI Professionals

“If I ran a Research Computing group like Enterprise IT: I should be fired.

If I ran an Enterprise IT group like a Research Computing group: I should be fired.”

Mike Bruck, retired Director of RC, University of Arizona

Research CI vs Enterprise IT

Berente 2017 was the first publication I saw do two key things:

- Outline the difference between Research CI and Enterprise IT
  - Both have a purpose and a place but they are not the same thing
  - Enterprise IT is traditionally more understood
- Champion that CI Professionals should have bespoke career paths
  - CI Professionals were (are still?) scarce in the market
  - Usually have familiarity with research (and/or advanced degrees)
  - There was no pipeline to train and engage CI Professionals

Berente, N., Howison, J., Cutcher-Gershenfeld, J., King, J.L., Barley, S.R. and Towns, J.,"Professionalization in Cyberinfrastructure" (February 15, 2017). ssrn.com/abstract=3138592

CI Professionals

- 2020 CI Workforce Development Workshop (NSF funded)
- Recommendations in 12 different topics for workforce development
- Need to define a career path for CI Professionals

A Traditional Role but Definitions Vary…

- CI Professionals build research CI
  - Networking
  - Computing hardware
  - Data storage
  - Data center infrastructure: power, cooling, ‘computer gardening’
- The assumption seems to be building physically, but that has changed in the era of cloud computing
- Traditionally referred to as ‘System Administrators’ or ‘Engineers’
- The Campus Research Computing Consortium (CaRCC) Research Computing and Data Services (RCD) Professionalization workgroup started a Human Resources Job Family Matrix (carcc.org/rcd-professionalization/)

carcc.org/rcd-professionalization/hr-job-framework-bgnd/

DevOps? Infrastructure as Code?

- CI professionals do not necessarily have to build physical infrastructure these days
  - ‘Cloud computing’ means you can rent/buy infrastructure and turn it into whatever you need
  - There is a lot of hybridization of infrastructure and job roles happening
- This advent allows for a lot of interesting new activities:
  - Scripting entire analysis infrastructures with Terraform and Ansible
  - Network monitoring, log aggregation, and optimization with Splunk and others
  - Greater automation of software deployment and versioning with containers and module definitions like Spack

Research Software Engineers

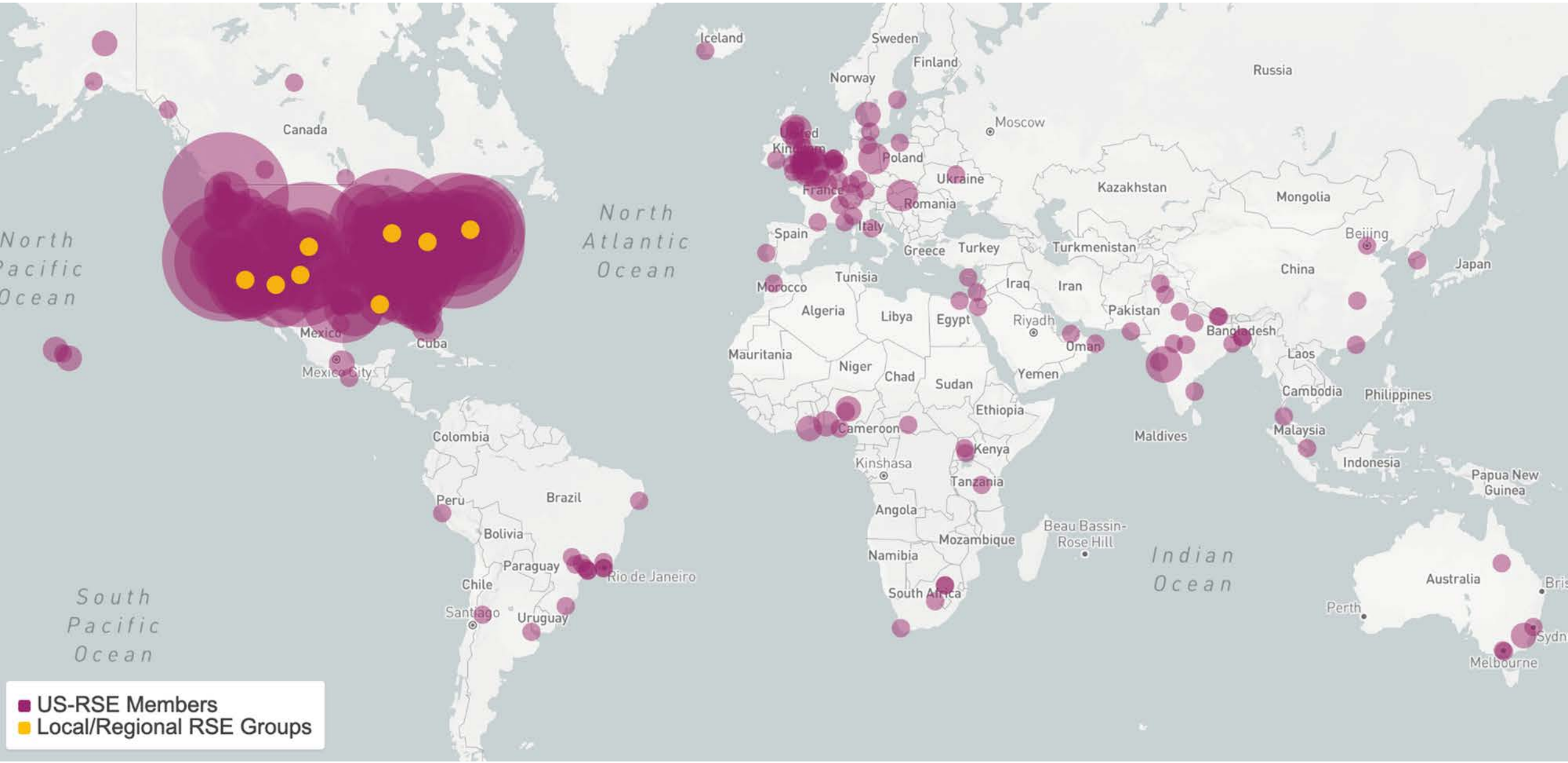
“Research Software Engineers encompass those who regularly use expertise in programming to advance research.”

US RSE Association

Research Software Engineers

- The designation started March 2012 at a Software Sustainability Institute event
- This position has existed for awhile, but there was not a uniform name for it.
- April 2022 US-RSE Community Building Workshop
- October 2023 the 1st US-RSE Conference

US RSE Association: us-rse.org/
RSE History: society-rse.org/about/history/
US RSE Conference: us-rse.org/usrse23/


Iceland
Sweden
Finland
Norway
Russia
Moscow
Canada
Poland
Ukraine
Kazakhstan
Mongolia
Romania
Spain
Italy
Beijing
Greece
Turkey
Turkmenistan
China
Japan
Tunisia
Morocco
Iraq
Iran
Algeria
Libya
Egypt
Riyadh
Pakistan
Bangladesh
Oman
Mauritania
Laos
Niger
Chad
Sudan
Yemen
Cambodia
Philippines
Ethiopia
Maldives
Cameroon
Malaysia
Kenya
Kinshasa
Indonesia
Tanzania
Papua New Guinea
Angola
Beau Bassin-Rose Hill
Mozambique
Namibia
Indian Ocean
Australia
South Africa
Perth
Melbourne
Mexico
Cuba
Mexico City
Colombia
Peru
Brazil
Bolivia
Paraguay
Rio de Janeiro
Chile
Santiago
Uruguay
North Atlantic Ocean
North Pacific Ocean
South Pacific Ocean
US-RSE Members
Local/Regional RSE Groups

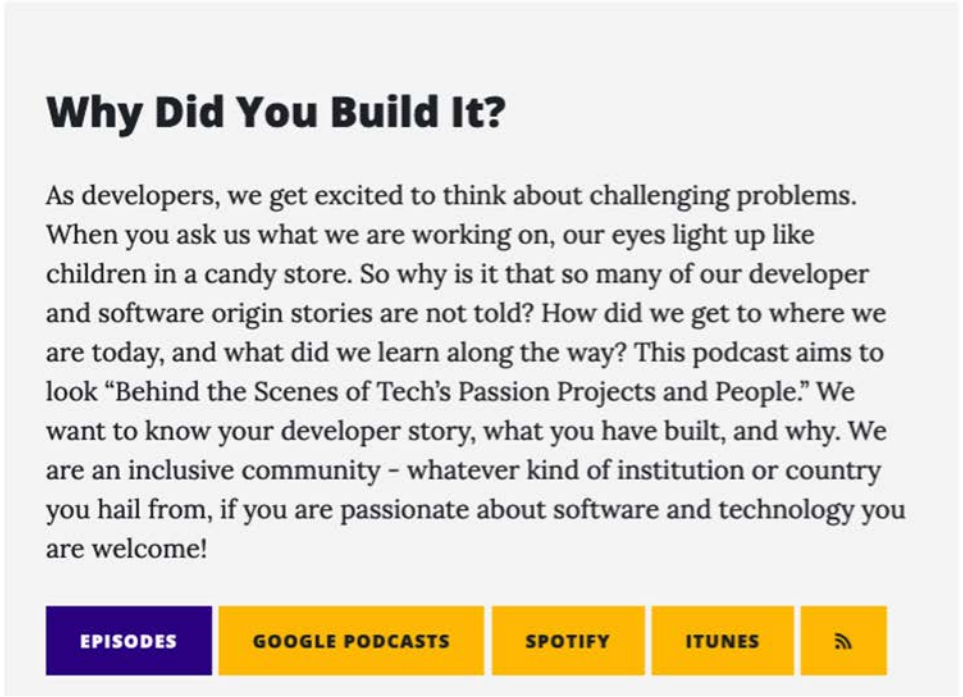
RSE Growth and Diversity Mirrors the Work

- Software development
- Code optimization
- Automated workflows with NextFlow, Snakemake, others
- Dashboards for data democratization
- Advanced and interactive data visualizations
- Containerization for nomadic workflows (and maybe reproducibility)

The list goes on and on and on….

Developer Stories

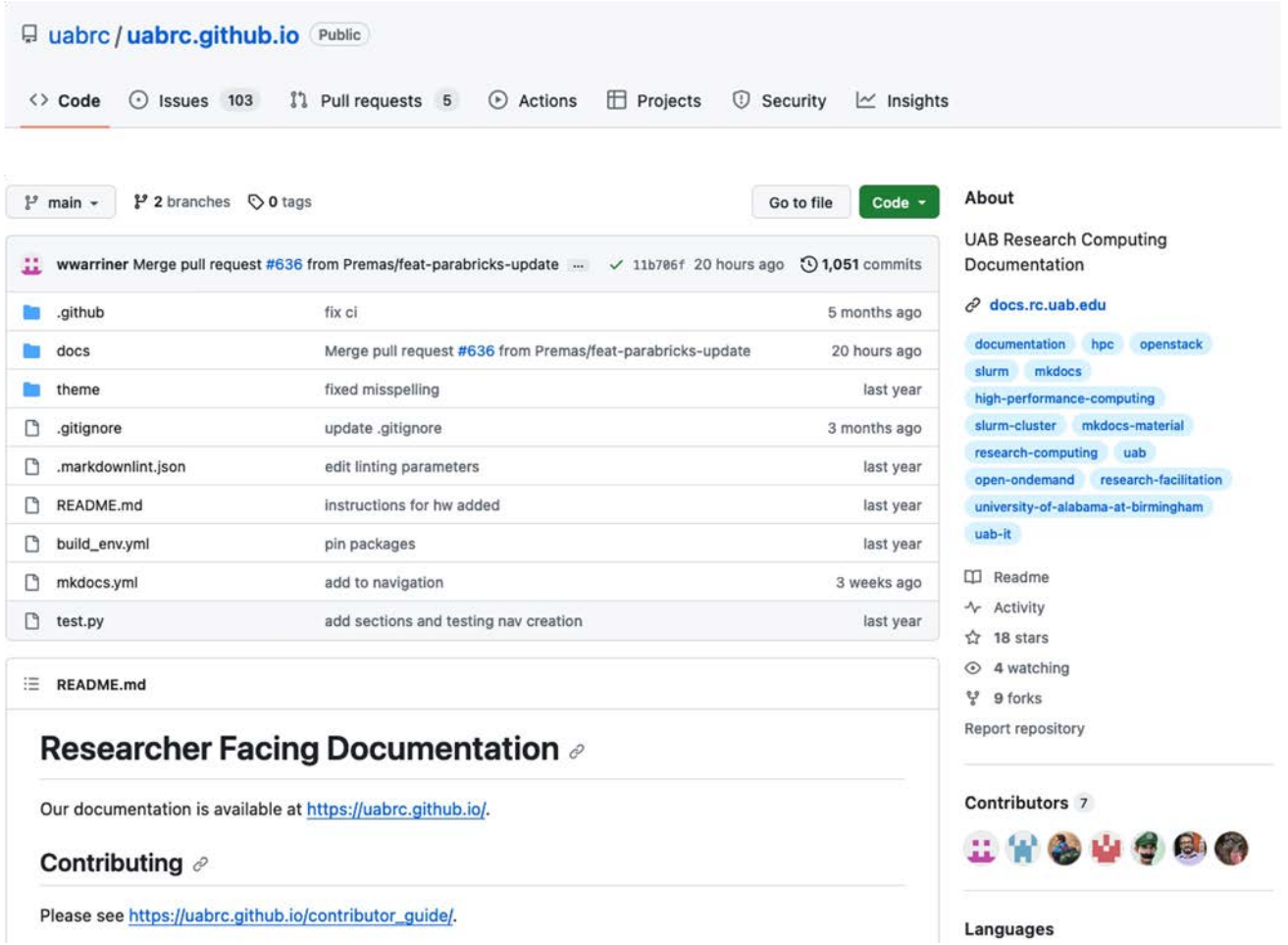
I highly recommend Developer Stories (rseng.github.io/devstories/)



Shameless plug: rseng.github.io/devstories/2020/blake-joyce/

Automated, Scripted, Written Documentation!

- GitHub actions to rebuild documentation after every pull request
- Automatic linting and filters
- We actually had random researchers submitting PRs
- VSCode + Markdown + MkDocs (mkdocs.org)



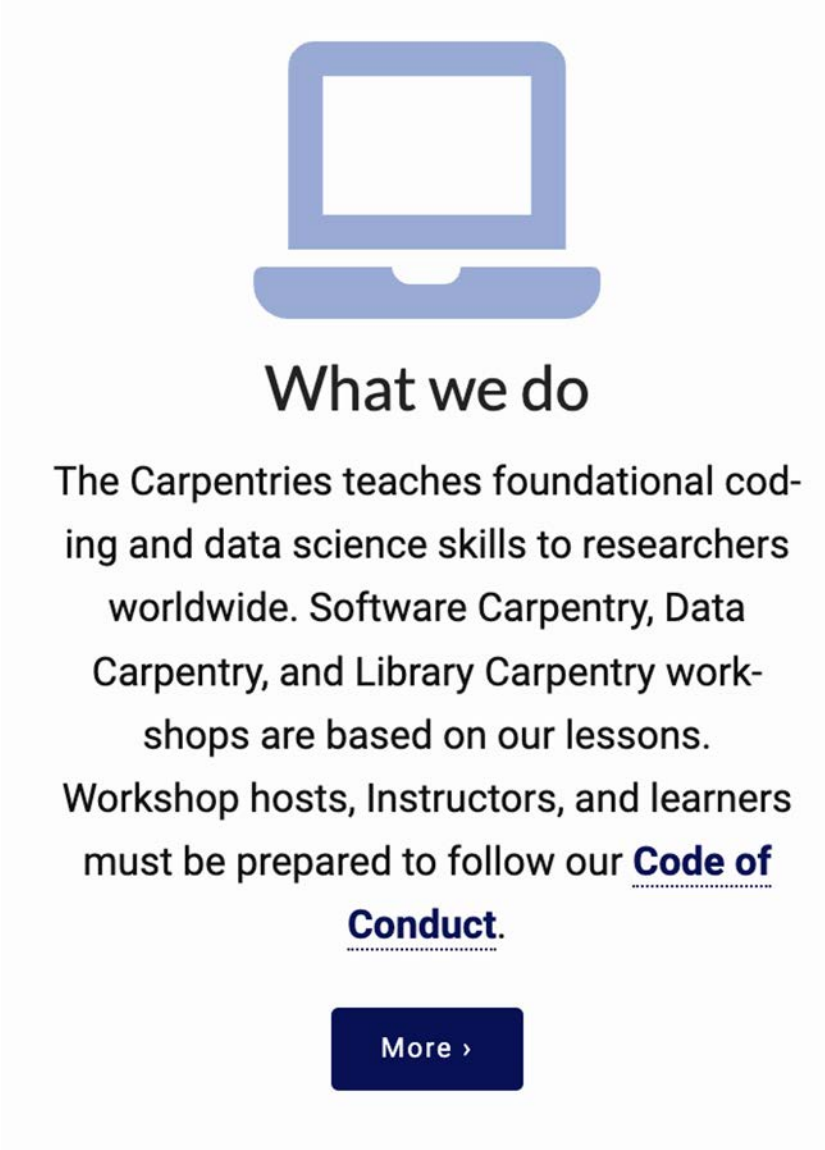
github.com/uabrc/uabrc.github.io

CI Facilitators

Work with diverse research domains to onboard, collaborate, and identify methods to help them leverage CI resources.

CI Facilitators

- Sometimes called Research Computing Facilitators
- I would classify the (Software) Carpentries were early facilitators
- A group of research and computer science volunteers that taught other researchers about code
- “We teach foundational coding and data science skills to researchers worldwide.”



Democratization is Powerful

- AI/ML/DL has been around for awhile
- ChatGPT wasn't really a revolutionary development
- ChatGPT became an overnight sensation because it democratized AI
  - A huge explosion in applications and time saving by users
  - Automated menial tasks so people can spend time on more important ones
  - Still requires an expert human in the loop to correct mistakes and strange results
- Imagine the same effect, but democratizing Research CI in an organization

Case Study: Here's 7 people, go help 2000 researchers

- Actual words said to me when I became a director
- Every manager here will know these things, but I had to learn them empirically
- General operating rules for helping researchers in a 285:1 ratio:
  - Define services and make them available (directly and asynchronously)
  - Define what you are not resourced to do … yet
  - Always have a prioritized budget ready to define new things you'd like to do

And Always Have Metrics on Hand

- UArizona RC ran about $11.2 million in computing equipment overall
- When I left, 105 departments were using the services (nearly everyone)

+68.6%
Unique Researchers
Using RC Services

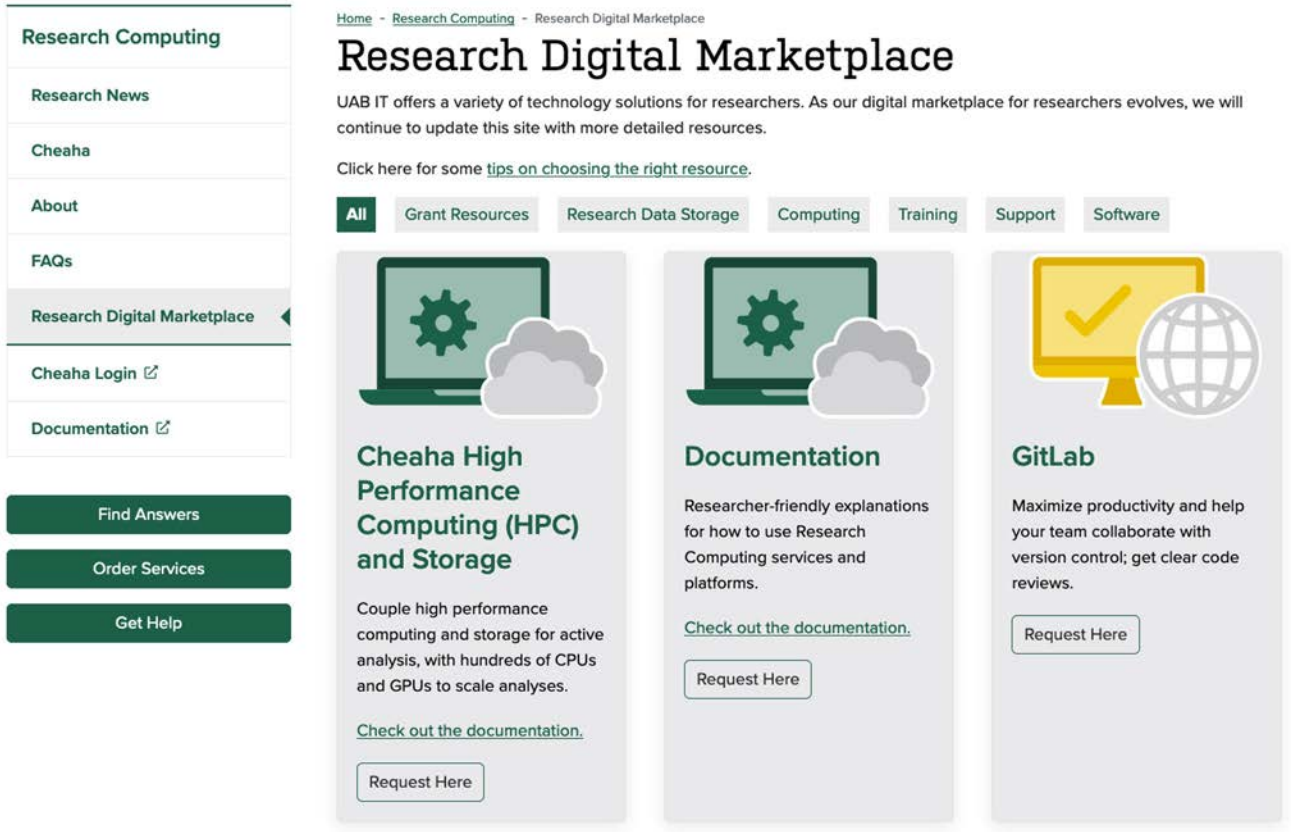
652 in FY 2018 → 2,269 in FY 2021

270 in FY 2018 → 725 in FY 2021

+74.4%
Unique Lab Groups
Using RC Services

Show Rather than Tell: research CI marketplaces

- Facilitators are doing their job best when no one knows they have done anything at all
- Distilling CI services and providing a place for asynchronous learning is critical
- After a year our community increased from **584 to 716 active researchers (+22.6%)**



uab.edu/it/home/research-computing/research-digital-marketplace

Roles that are Starting to Emerge

We apologize for the inconvenience.

This section is the least structured because these roles seem to change daily.

Data Management and Data Lifecycle

I am following this area most intently; for a few reasons:

- Analysis-ready data is crucial for modern research (AI/ML/DL)
- Getting data ready takes the right people and the right CI
- The role traditionally fell on research leadership:
  - Leadership can set policies, but are not prepared for data management
  - Everyone is realizing that tera/petabytes ***need dedicated staff expertise***
  - There needs to be a combination of CI, software, research, and management experience on the team

Can We Save Lives with AI? bioteam.net/events/webinar-can-we-save-lives-with-ai-data-readiness-in-healthcare-and-life-sciences/

If you missed the first webinar in this series:

Can We Save Lives with AI? Data Readiness in Healthcare and Life Sciences

Hosted by Dr. Ari Berman, CEO at BioTeam

Watch the recording at bioteam.net/events

I Have Heard a Lot of Potential Roles and Titles

- Data-Product Manager
- Chief Data Officer
- Metadata Generator
- Data Engineer
- Database Architect
- Data Manager
- Data Analyst

- These often feel like a subset of other job titles (*e.g.* data scientist, CI professional)
- Maybe that is the point though:
  - The other job titles and roles are colors to be mixed together
  - Some combinations of roles leads to a new title to fit that need

Chief Data Officer: hbr.org/2013/10/are-you-ready-for-a-chief-data-officer
Data-Product Manager: hbr.org/2022/10/why-your-company-needs-data-product-managers

Research Project Managers

- Frequently lumped in with traditional management structures
- But I argue there are fundamental and important differences
- Do you need someone to:
  - Outline a timeline and critical paths for a project to get completed?
  - Assess compliance and risks for non-compliance? (*e.g.* data security)
  - Communicate, plan, and manage change to CI, services, or policies?
  - Find improvements for chaotic work processes?
- Well then you need a Research Project Manager
  - The trouble is I cannot seem to find a group that trains these people
  - The Project Management Institute kind of has a research section? Sorta?

Leading and Managing CI Groups

- Leadership, management, and the business aspects of Research CI Organizations is a whole other talk
  - Assessment of research needs and developing new service catalogs
  - Creating budgets based on staffing and technology market analysis
  - Metrics and assessment of existing research computing practices
- But there are some community resources available
  - Research Software Engineers: Creating a Career Path—and a Career (zenodo.org/records/10073233)
  - The ADSA and US RSE Hiring, Managing, and Retaining DS and RSEs in Academia book (doi.org/10.5281/zenodo.8329337)

Research Software Engineers: Creating a Career Path—and a Career https://zenodo.org/records/10073233
Van Tuyl, S. (Ed .) . (2023). Hiring, Managing, and Retaining Data Scientists and Research Software Engineers in Academia: A Career Guidebook from ADSA and US-RSE. Zenodo. doi.org/10.5281/zenodo.8329337

The Business of CI: Buy-in and Windfall Queues

- Researchers wanted a way to increase their cluster computing resources (time, number of jobs, number of GPUs)
- A windfall system allowed the community to use cycles when the resources were idle
- And it allowed us to pivot to organizational priorities when needed

Arizona Universities Join Research Computing Fight Against COVID-19

July 20, 2020

Tri-U Research Computing on COVID-19

*UANews published an article recently about the contributions of Arizona's three state universities to the **Folding@home project**.*

The University of Arizona, in partnership with research computing centers at Arizona State University and Northern Arizona University, is contributing research computing resources to a worldwide effort to advance COVID-19 research.

Arizona's three state universities are participating in the national **Folding@home project**, which relies on volunteers' idle computing power to run protein modeling computations that help researchers learn more about how to cure or treat certain diseases.

it.arizona.edu/news/arizona-universities-join-research-computing-fight-against-covid-19

Visit us at SC23 in Denver—booth 1787.

Questions?

BioTeam

Accelerate Science