Funded by the National Science Foundation, Grant #2127548

CI Compass

CI Compass: The NSF Cyberinfrastructure (CI) Center of Excellence for Navigating the Major Facilities Data Lifecycle

Ewa Deelman
University of Southern California
Information Sciences Institute

NSF Advisory Committee for Cyberinfrastructure
April 28, 2022

USC Viterbi School of Engineering *Information Sciences Institute*

INDIANA UNIVERSITY

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

TEXAS TECH UNIVERSITY

renci

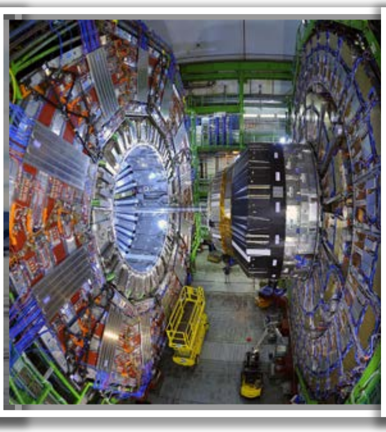
NSF Large/Major Facilities

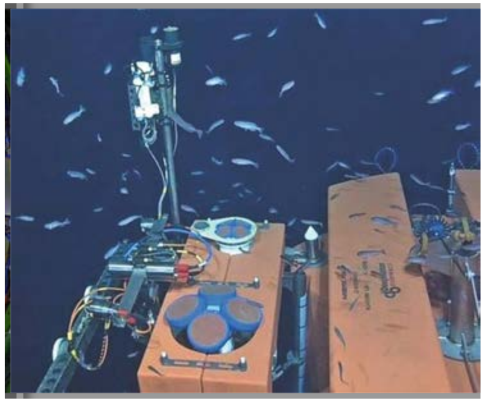
- Deliver data, modeling, computational, and physical capabilities to the broad research and engineering community, students, educators, and the public
- Highly diverse, complex, and heterogeneous
- Differ in types of data captured, scientific instruments used, data processing and analyses conducted, policies and methods for data sharing and use
- Rely on complex CI to the transform raw data into more interoperable and integration-ready data products that can be visualized, disseminated, and transformed into insights and knowledge.

IceCube

IRIS/SAGE

NEON


CMS

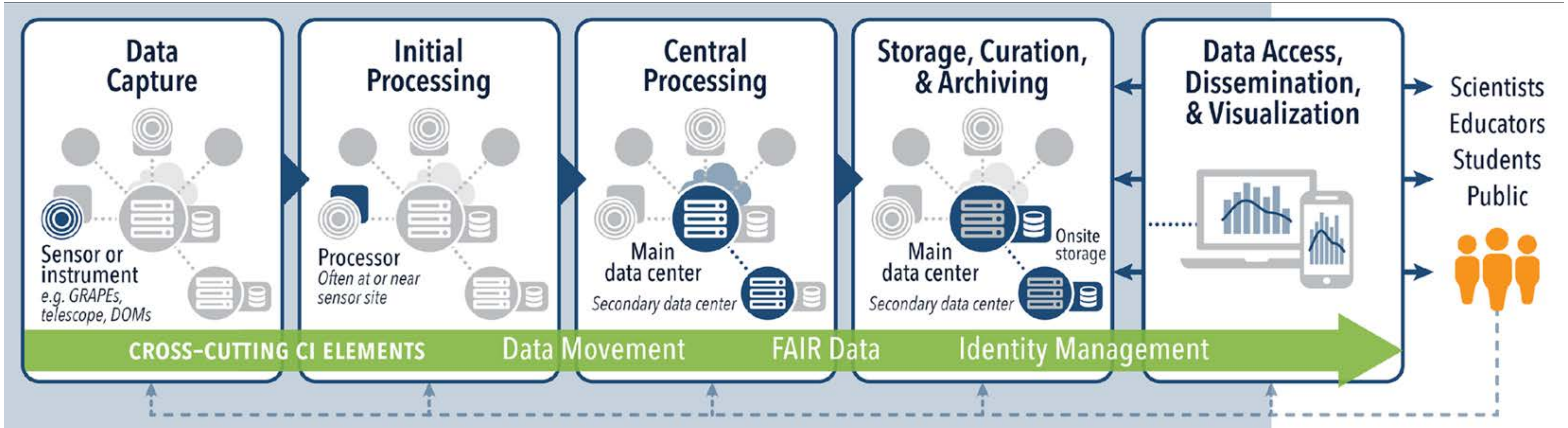

OOI

Mission

CI Compass

CI Compass provides expertise and active support to cyberinfrastructure practitioners at NSF Major Facilities in order to accelerate the data lifecycle and ensure the integrity and effectiveness of the cyberinfrastructure upon which research and discovery depend.

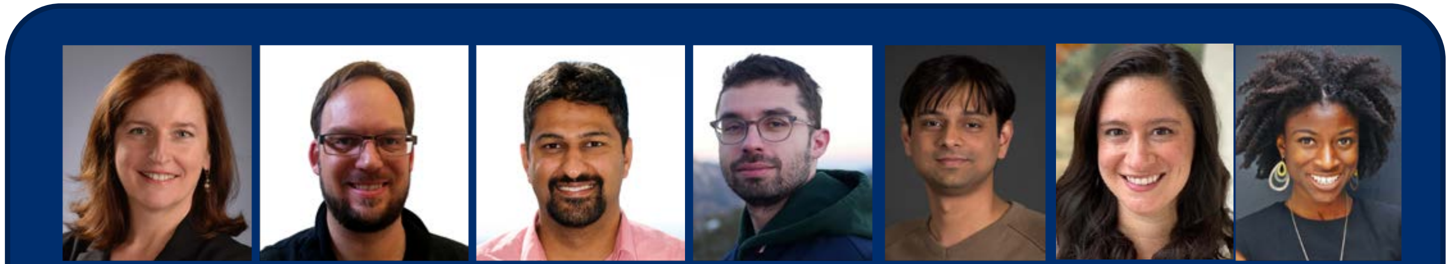
CI Compass Services focus on Major Facilities' Data Lifecycle



**Evaluate CI plans, Help architect new solutions, Develop proofs of concept,
Assess applicability/performance of existing solutions, Help leverage existing technologies**

CI Compass Team

Funded by the National Science Foundation, Grant #2127548



Automation, Resource Management, Workflows, Project Management

USC

- Ewa Deelman (PI)
- Mats Rynge
- Karan Vahi
- Loïc Pottier
- Rajiv Mayani
- Nicole VIrdone
- Ciji Davis

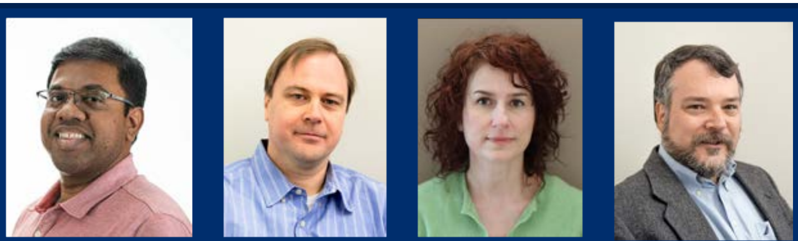
Texas Tech University

Kerk Kee Alex Olshansky

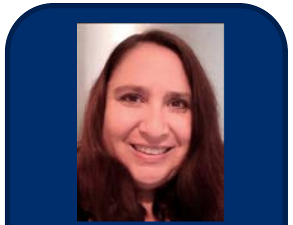
Communication & organization science

RENCI

- Anirban Mandal (co-PI)
- Ilya Baldin
- Laura Christopherson
- Erik Scott



Resource Management, Networking, Clouds, Social Science, Evaluation



Data Archiving

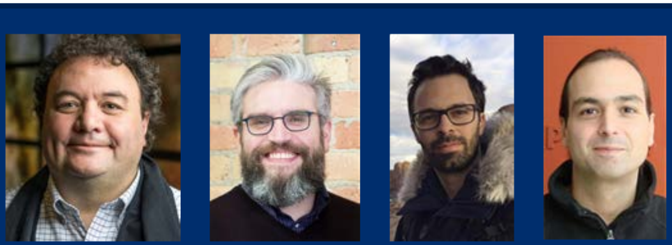
Indiana University

Angela Murillo (co-PI)

Workforce development, Sensors, operations, Semantic technologies, Communications and Outreach

University of Notre Dame

- Jarek Nabrzyski (Co-PI)
- Joanne Fahey
- Charles Vardeman
- Mary Gohsman
- Christina Clark
- Don Brower



Data management, visualization, clouds, CI deployment

University of Utah

- Valerio Pascucci (Co-PI)
- Rob Ricci
- Steve Petruzza
- Giorgio Scorzelli

CI Compass Team: Who we are

Deep expertise in several CI areas critical to the MFs

- Data management, data processing, visualization, archiving, semantic technologies
- Automation, resource management, workflows, sensors
- Networking, clouds, systems and infrastructure
- Large-scale CI deployment and operations, IdM
- Social science, understand the organization structures and culture of MFs

Experience in the management of CI projects

- Conceptualization, design phase, broad adoption
- Project Management and Evaluation
- Organizational science
- Communications & Outreach

Highly collaborative, strong history of working together and with the CS and CI Communities

- Many diverse community connections in astronomy, earth science, physics

Dedicated to the advancement of CI for science, engineering, and education

USC Viterbi School of Engineering Information Sciences Institute

INDIANA UNIVERSITY

UNIVERSITY OF NOTRE DAME

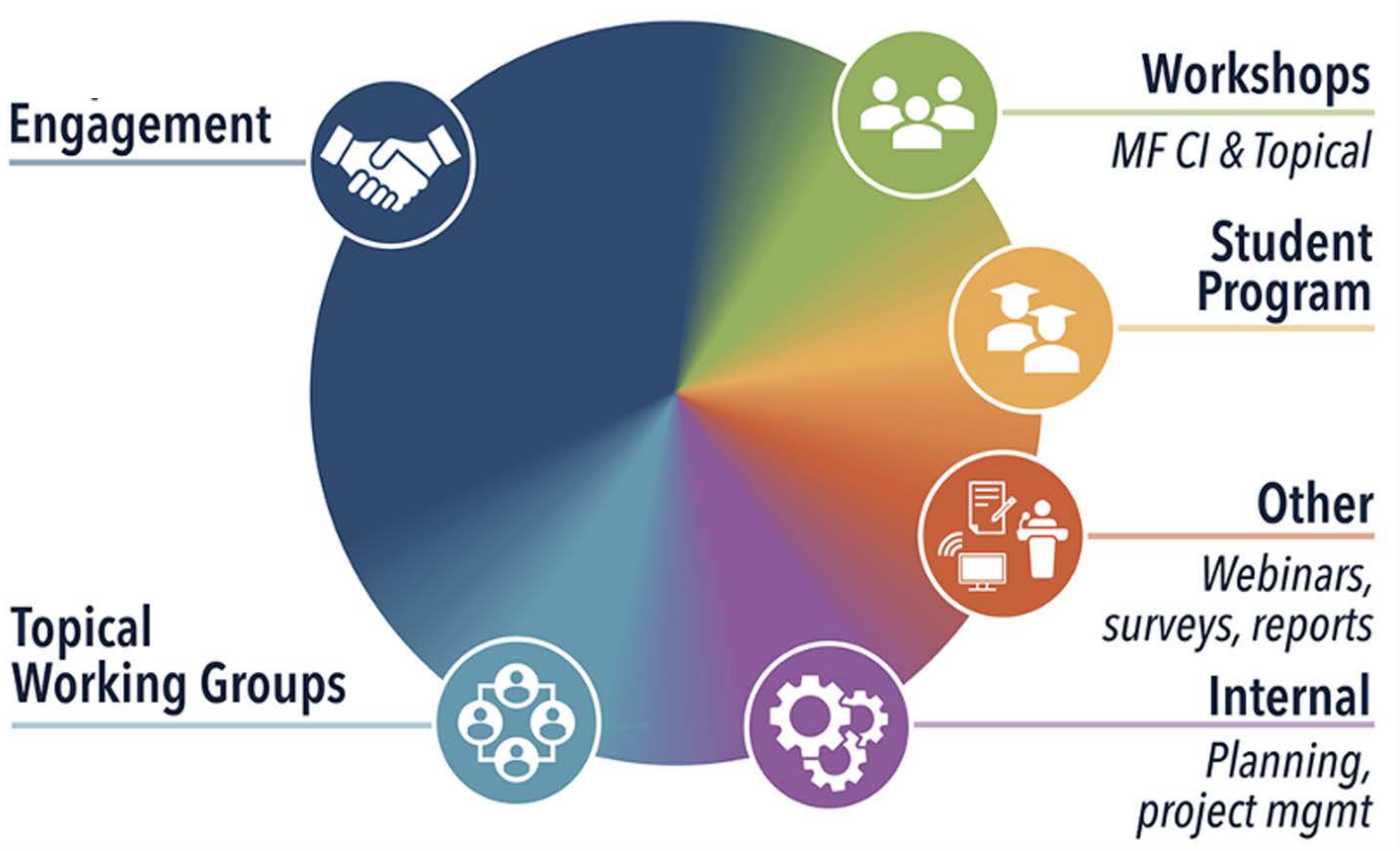
THE UNIVERSITY OF UTAH

TEXAS TECH UNIVERSITY

renci

Funded by the National Science Foundation, Grant #2127548

CI Compass activities: What we do



Engagement

Workshops

MF CI & Topical

Student Program

Other

Webinars, surveys, reports

Internal

Planning, project mgmt

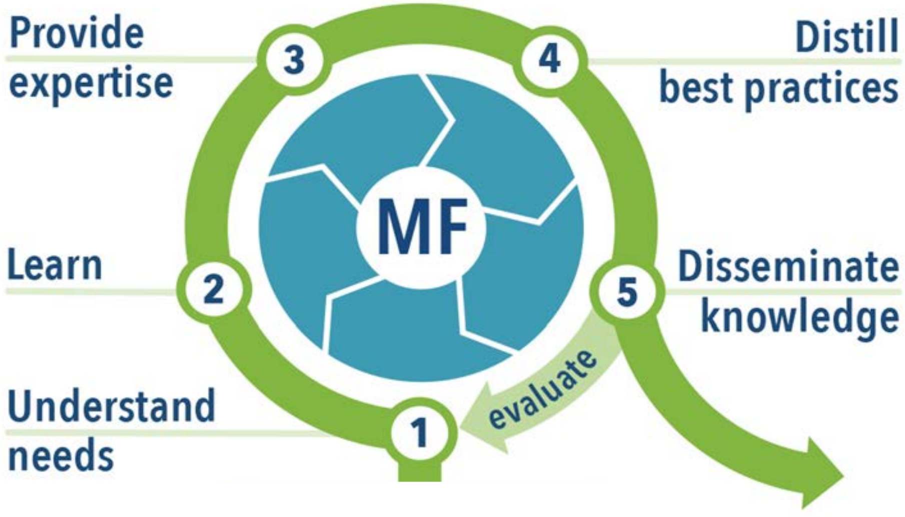
Topical Working Groups

Funded by the National Science Foundation, Grant #2127548

Overall CI Compass Strategy

1. Recognize the expertise, experience, and mission-focus of Major Facilities
2. Contribute knowledge and expertise to the MF DLC CI and enhance the overall NSF CI ecosystem.
3. Build expertise, not software
4. Build on existing knowledge, tools, community efforts
5. Leverage existing collaborations we are part of: PATh (Deelman, Sr. Personnel), ACCESS MATCH (user support, Deelman, Co-PI).
6. Build partnerships to leverage community expertise
   - Trusted CI: cybersecurity
   - Science Gateways Community Institute (SGCI): portals
   - Engagement and Performance Operations Center (EPOC): network utilization/optimization
   - Research Computing and Data Nexus, CI workforce development
   - Chameleon, cloud and edge-to-cloud experimentation and testing
   - Fabric, next generation networks experimentation and testing
7. Share knowledge, lessons learned, best practices with MFs, Partners, CI community

How we work?



Engagements

* Some work products are internal to an engagement

1-1 Engagements with MFs

- Identify a topic or topics that are important and not-yet fully solved by the Major Facility (MF)
- Form working groups/ embed in existing ones
- Conduct focused discussions, work together on particular challenges
- Work products: documents/papers, proofs of concept, schema implementations, demos
- Document and evaluate the collaboration and outcomes

Topical Working Groups

- Identify a topic that is important to a number of MFs
- Facilitate discussions, sessions at conferences, collect and share experiences, distill best practices

Community Building

- Share knowledge, build connections
- Host community activities: workshops, training
- Identify related efforts
- Help connect people, projects, and communities
- Collect information and disseminate information about the broad community activities and training opportunities

CI Compass

Engagements

Funded by the National Science Foundation, Grant #2127548

COMPLETED	ACTIVE	PLANNED/ IN DISCUSSION
▒ Arecibo	▒ NEON/NCAR	NCAR
░ ARF	LIGO	NAN/Midscale R-2
█ NEON	░ NOIRLAB	RDE/Midscale R-2
░ OOI	░ RCRV	
░ RCRV	█ SAGE/GAGE	

TIME SCALE

░	A few months
▒	About 1 year
█	More than 1 year

Credit: Arecibo Observatory

ENGAGEMENT	PARTNERS
Arecibo	EPOC, TACC Globus, UCF, IVOA
ARF, NOIRLab, RCRV	Trusted CI
SAGE/GAGE	Internet2

Regional Class Research Vessel (RCRV) Engagement:

Shipboard CI/network plan review

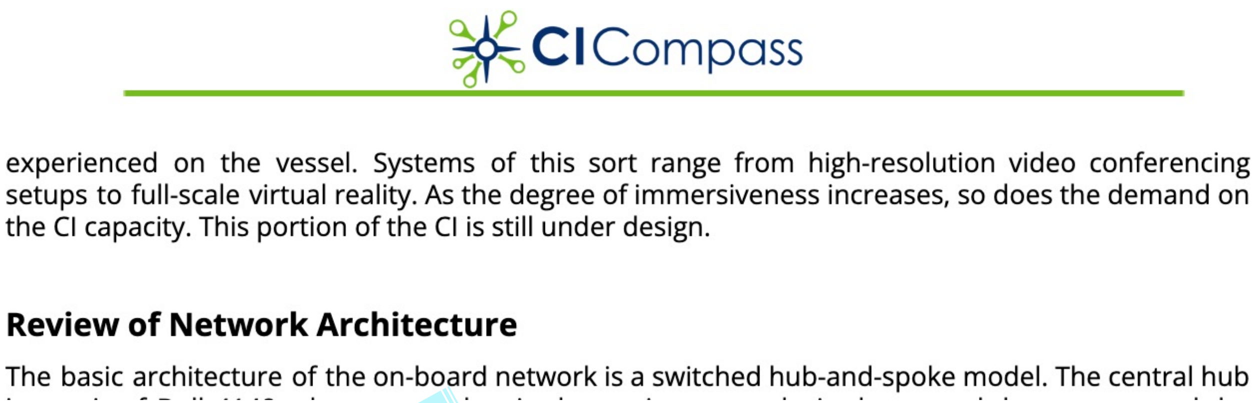
Transition to Operations Meeting Focus Session:
RCRV CYBERINFRASTRUCTURE

Regional Class Research Vessel Program- Datapresence and Engineering Support Center • Oregon State University

THE UNIVERSITY OF RHODE ISLAND GRADUATE SCHOOL OF OCEANOGRAPHY
Oregon State University College of Earth, Ocean, and Atmospheric Sciences
LUMCON
THE UNIVERSITY OF SOUTHERN MISSISSIPPI

Planned RCRV vessels

January - March 2021



CICompass

experienced on the vessel. Systems of this sort range from high-resolution video conferencing setups to full-scale virtual reality. As the degree of immersiveness increases, so does the demand on the CI capacity. This portion of the CI is still under design.

Review of Network Architecture

The basic architecture of the on-board network is a switched hub-and-spoke model. The central hub

"One of the primary concerns identified by the review was that the planned 1GbE switch ports in the ship's computer lab should be supplemented with 10 GbE and higher to support deployment of visiting equipment with high-speed network interfaces. "

From Chris Romsos, RCRV, OSU : "Thank you for identifying this as something to address now before delivery of the vessel. We planned for future upgrades like this and have sufficient fiber between the network core and the computer lab to support the upgrade…
Sikuliaq recently upgraded their edge switching throughout the vessel… . A nice piece of corroborating evidence there with Sikuliaq! "

Field report from ARF, RV Sikuliaq

Funded by the National Science Foundation, Grant #2127548

NEON/NCAR Engagement

Goals

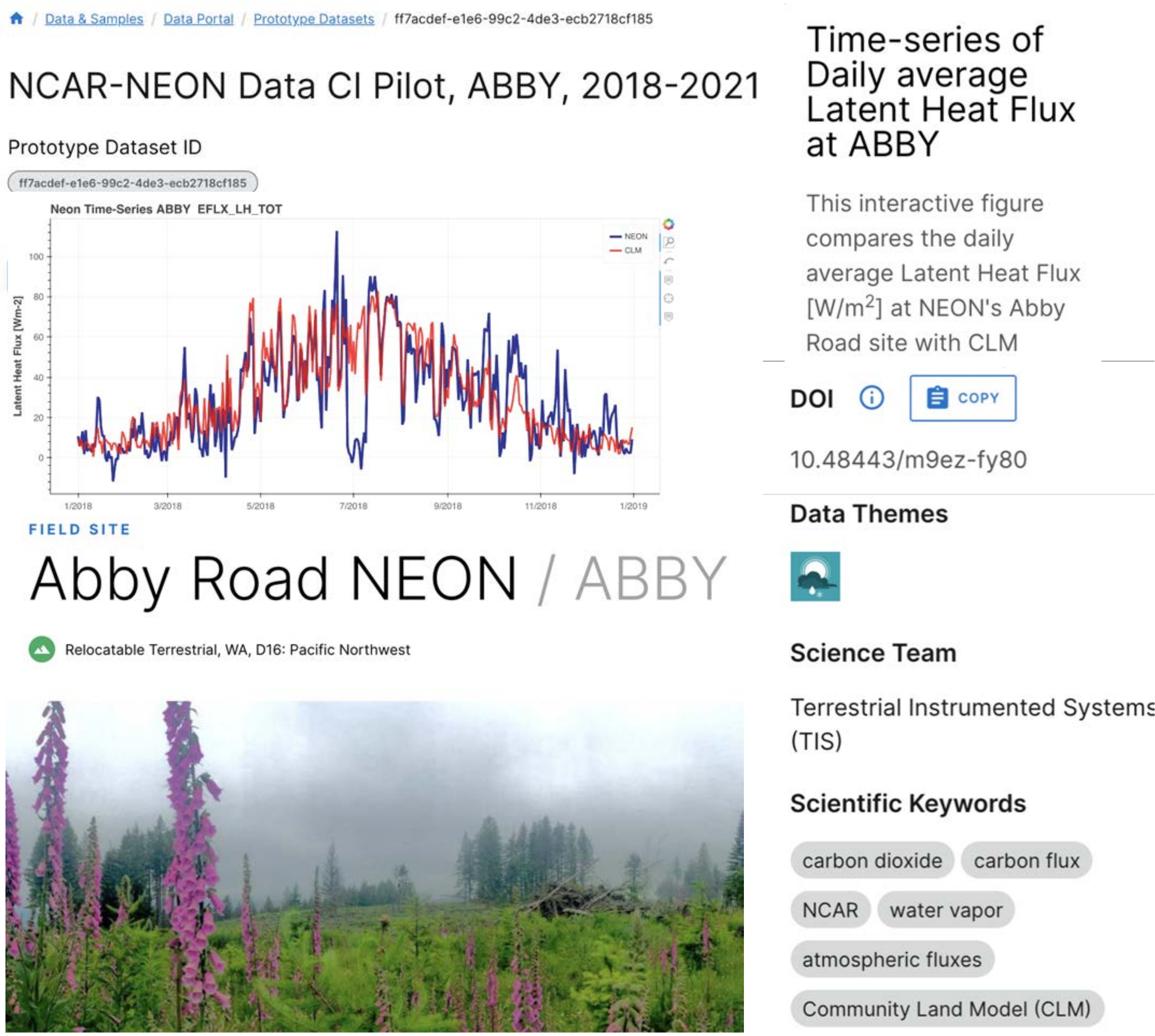
- Combine NEON ecosystem data with NCAR atmospheric and land modeling capabilities
- Inspire new discoveries with integrated data from NEON and NCAR modeling
- Use cloud technologies to enable data modeling and wide community access

---

- Consulted on cloud technologies, including containers
- Helped with container testing
- Consulted on FAIR data and visualization
- Learned about data management challenges for computational models

2020

https://www.neonscience.org/ncar-neon-community-collaborations



/ Data & Samples / Data Portal / Prototype Datasets / ff7acdef-e1e6-99c2-4de3-ecb2718cf185

NCAR-NEON Data CI Pilot, ABBY, 2018-2021

Prototype Dataset ID

ff7acdef-e1e6-99c2-4de3-ecb2718cf185

FIELD SITE

Abby Road NEON / ABBY

Relocatable Terrestrial, WA, D16: Pacific Northwest

Time-series of Daily average Latent Heat Flux at ABBY

This interactive figure compares the daily average Latent Heat Flux [W/m^2] at NEON's Abby Road site with CLM

DOI COPY

10.48443/m9ez-fy80

Data Themes

Science Team

Terrestrial Instrumented Systems (TIS)

Scientific Keywords

carbon dioxide carbon flux NCAR water vapor atmospheric fluxes Community Land Model (CLM)

SAGE/GAGE Common Cloud Platform (CCP) [Feb 2020 - …]

GAGE Operated by UNAVCO
UNAVCO
SAGE Operated by IRIS
IRIS

Project goal: Develop a Common Cloud Platform (CCP) for ingestion, archiving, curation, processing, and distribution of their data in a cloud environment in support of the combined SAGE/SAGE data services facility serving geodetic and seismic communities.

Engagement Phase 1 WGs	Engagement Phase 2 WGs	Engagement Phase 3 WGs
● Data Flows and Use Cases, ● Concept of Operations, ● High-level Requirements, ● Platform Design	● Data Collection, ● Data Archiving, ● Data Distribution, ● Cloud Provider Analysis, ● Process Orchestration, ● Identity Management	● GeoCrate Common Data Container/Framework, ● Metadata Handling System, ● Prototype System in Commercial Cloud

May 2020 February 2021 November 2021

CI Compass : Provide advice on different WG areas related to their DLC; Review system design and performance limitations; Conduct research into and documentation of CI best practices for CCP architecture design; Co-design architectural documents and solutions for data access, data ingest and processing, migration, storage tiering, and archiving. Observe, learn, and document a complex MF CI migration into Cloud and institutional merge process.

SAGE/GAGE Common Cloud Platform (CCP)

Data Collection

Data Processing

Data Distribution

Process Orchestration

Cloud Provider Analysis

Identity Management

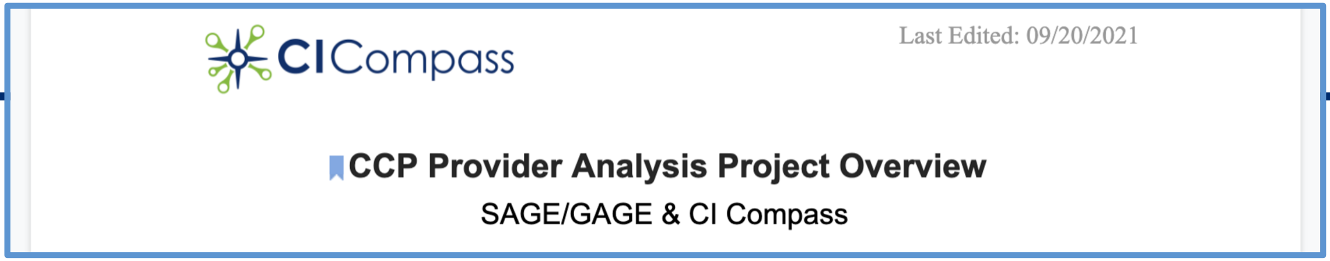
Data Identifier Schemes

Data Archiving

- Developed a Cloud Provider Analysis document advising CCP on a range of potential commercial, academic, or hybrid solutions that could provide the CCP data services at the lowest cost over a span of 5+ years, with a focus on risks *(Internal and Public versions)*
- Developed technical reports with advice on Data Storage Architecture Considerations: Cloud storage optimization, Block/File/Object storage design concerns, Database Design for geospatial data, FAIR data.. *(Internal and Public versions)*

REVIEW OF COST/RISK/BENEFIT ANALYSIS

- Is there a reasonable solution for an affordable cost?



Last Edited: 09/20/2021

CCP Provider Analysis Project Overview

SAGE/GAGE & CI Compass

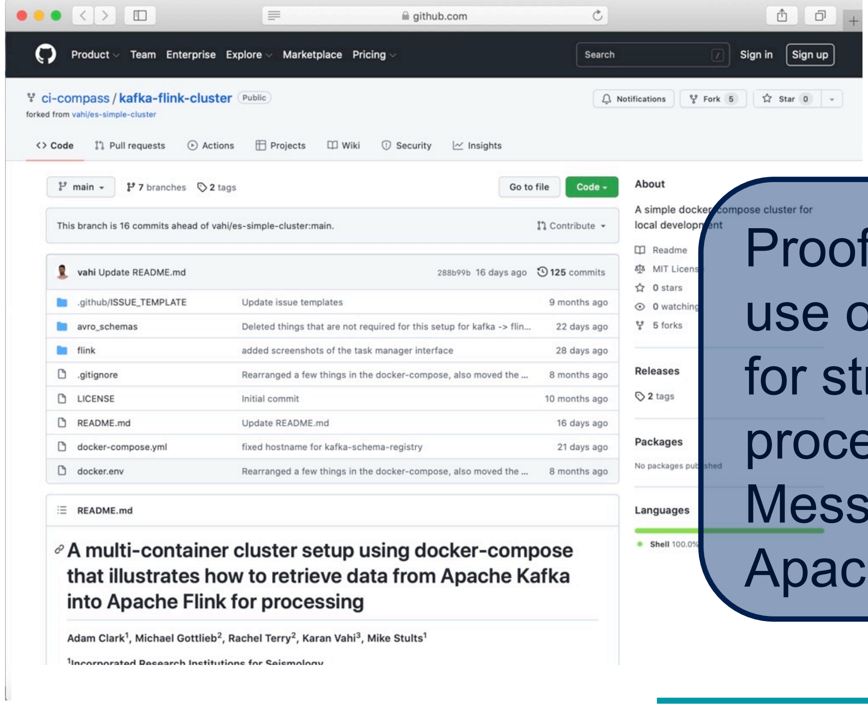
Feb ‘21

2021

Nov ‘21

Public Products

Funded by the National Science Foundation, Grant #2127548



This branch is 16 commits ahead of vahi/es-simple-cluster:main.

vahi Update README.md 288b99b 16 days ago 125 commits

.github/ISSUE_TEMPLATE	Update issue templates	9 months ago
avro_schemas	Deleted things that are not required for this setup for kafka -> flin...	22 days ago
flink	added screenshots of the task manager interface	28 days ago
.gitignore	Rearranged a few things in the docker-compose, also moved the ...	8 months ago
LICENSE	Initial commit	10 months ago
README.md	Update README.md	16 days ago
docker-compose.yml	fixed hostname for kafka-schema-registry	21 days ago
docker.env	Rearranged a few things in the docker-compose, also moved the ...	8 months ago

About

A simple docker compose cluster for local development

Readme

MIT License

0 stars

0 watching

5 forks

Releases

2 tags

Packages

No packages published

Languages

Shell 100.0%

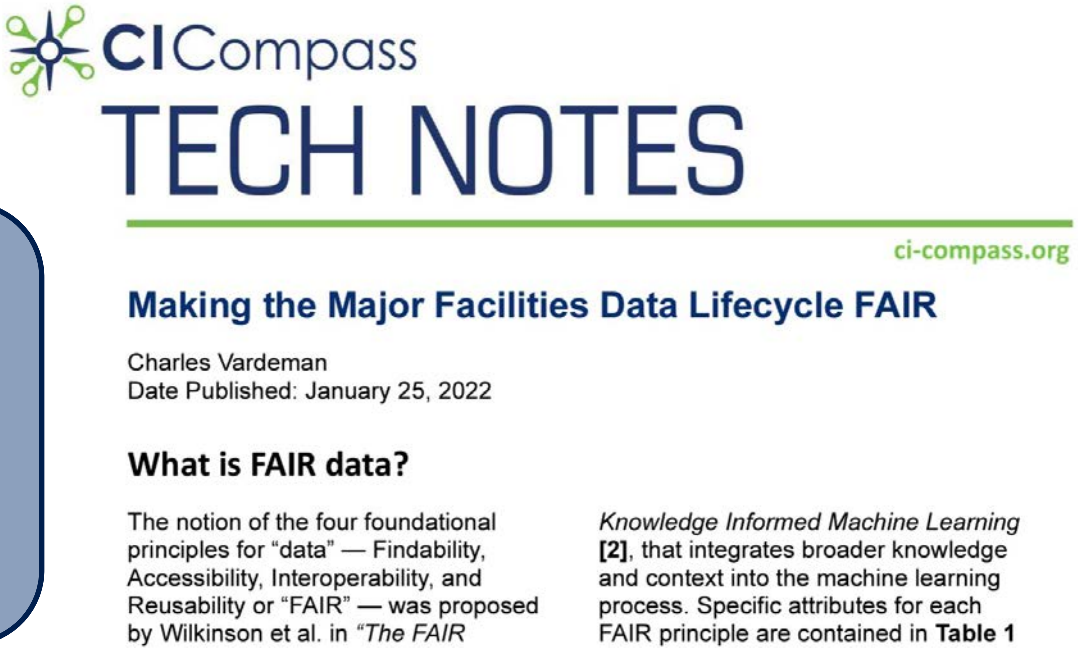
README.md

A multi-container cluster setup using docker-compose that illustrates how to retrieve data from Apache Kafka into Apache Flink for processing

Adam Clark[1], Michael Gottlieb[2], Rachel Terry[2], Karan Vahi[3], Mike Stults[1]

[1]Incorporated Research Institutions for Seismology

Proof of concept on use of Apache Flink for stream-based processing of Messages out of Apache Kafka



CI Compass

TECH NOTES

ci-compass.org

Making the Major Facilities Data Lifecycle FAIR

Charles Vardeman
Date Published: January 25, 2022

What is FAIR data?

The notion of the four foundational principles for "data" — Findability, Accessibility, Interoperability, and Reusability or "FAIR" — was proposed by Wilkinson et al. in *"The FAIR*

Knowledge Informed Machine Learning **[2]**, that integrates broader knowledge and context into the machine learning process. Specific attributes for each FAIR principle are contained in **Table 1**

CI Compass

TECHNICAL REPORT

Best Practices for Cloud Provider Analysis *

CI Compass

Last Edited: 07/01/2021

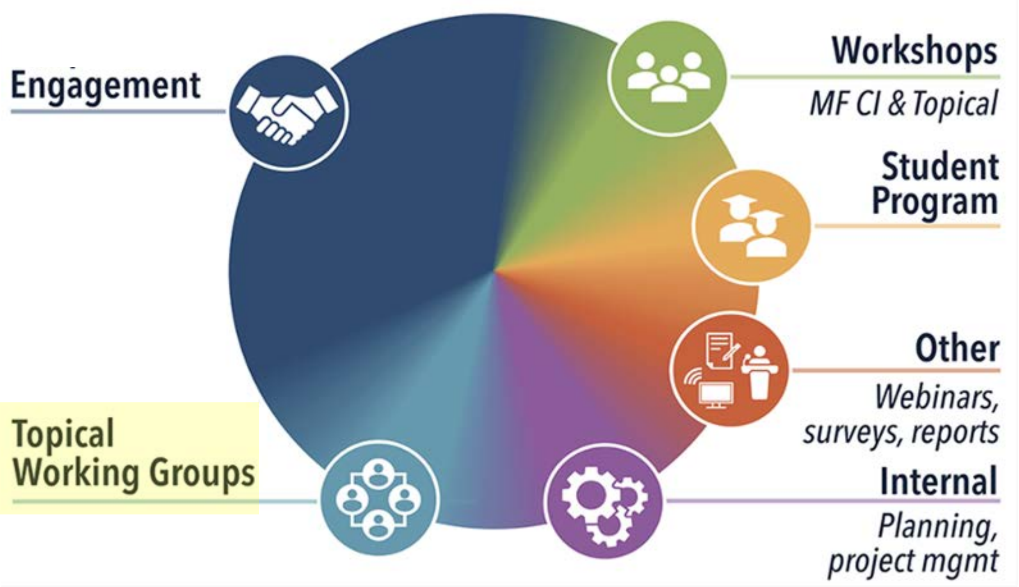
*

CI Compass Comments and Suggestions for Large Facility Cyberinfrastructure Design

Part 2: Data Storage Architecture Considerations

* In progress

CI Compass activities: Topical Working Groups



Identity Management Topical WG

Disseminate IdM information

- Quarterly meetings with speakers and discussions on topics relevant to MFs: e.g. CILogon
- Issues of identifying data usage and enabling reporting

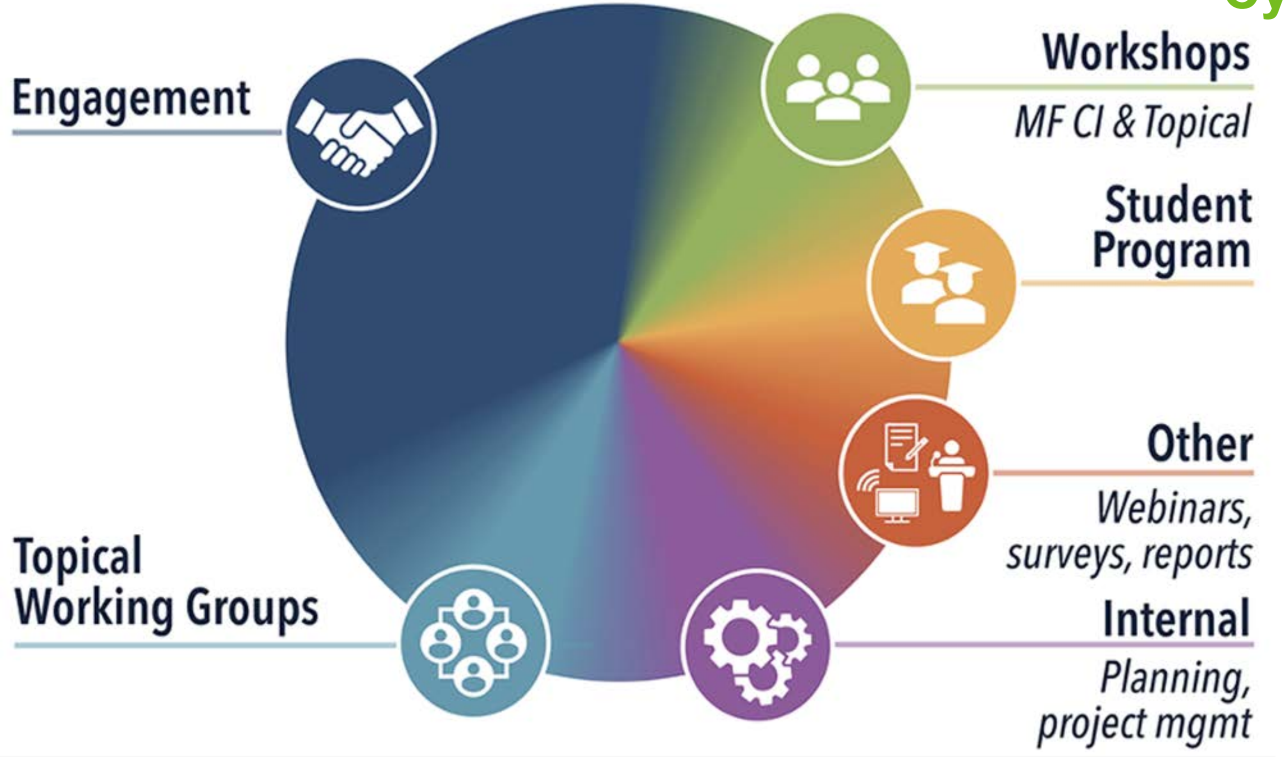
Cloud Infrastructure Topical WG

- Understand the current practices for clouds used by MFs
- Research alternative solutions and keep up to date with emerging cloud technologies
- Develop a general set of best practices that can inform the MFs

Workforce Development and FAIR data groups are being incubated based on CI4MFs workshop discussions

Funded by the National Science Foundation, Grant #2127548

Community Workshops



- *108 participants, including 35 in person in Redondo Beach*
- *Report due at the end of May 2022*

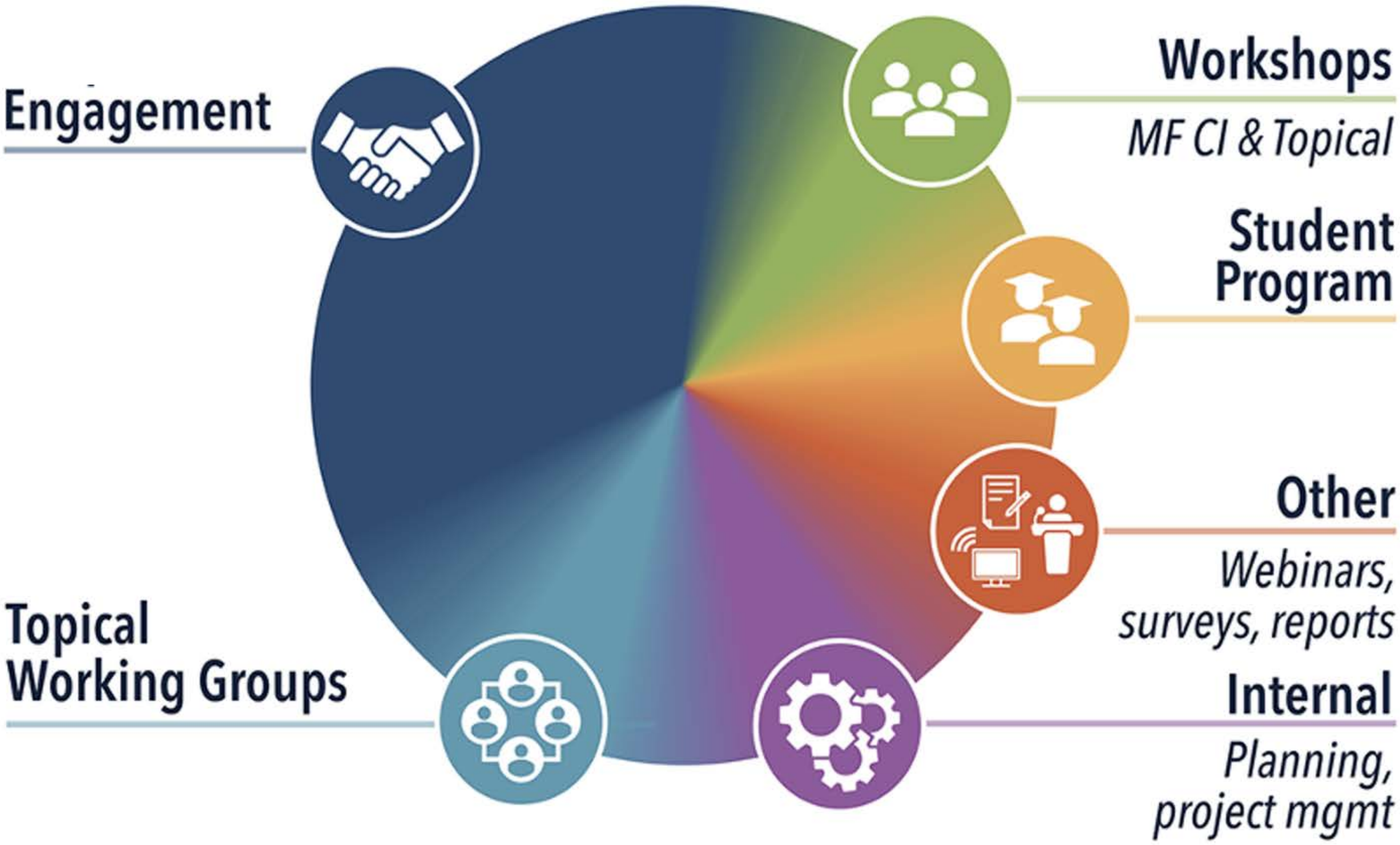
Cyberinfrastructure for Major Facilities Workshop

Getting Together, Working Together

March 1-2, 2022

- **Future of CI for MFs**
- **Cloud migration**
- **Making data FAIR**
- **CI workforce:**
  - **Developing and retaining talent**
  - **Developing resilience**

CI Compass activities: What we do



Engagement

Workshops

MF CI & Topical

Student Program

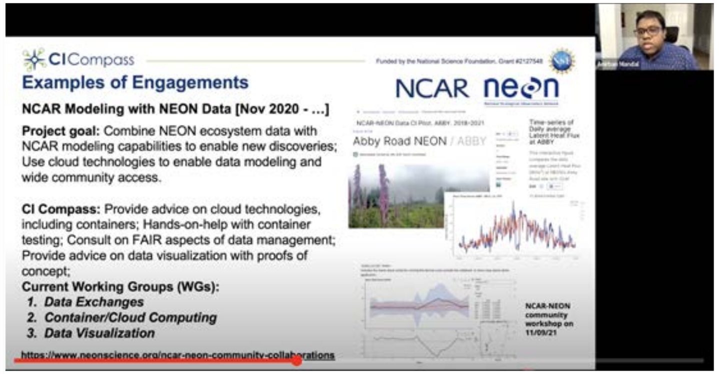
Other

Webinars, surveys, reports

Topical Working Groups

Internal

Planning, project mgmt



CI Compass

Funded by the National Science Foundation, Grant #2127548

Examples of Engagements

NCAR Modeling with NEON Data [Nov 2020 - …]

Project goal: Combine NEON ecosystem data with NCAR modeling capabilities to enable new discoveries; Use cloud technologies to enable data modeling and wide community access.

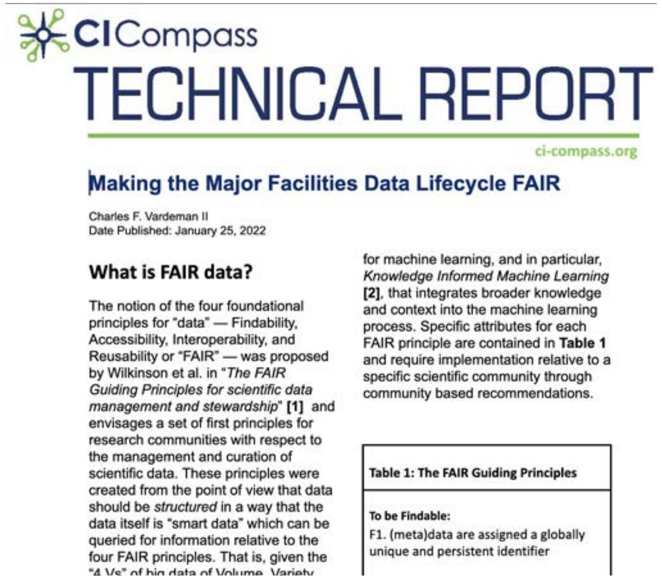
CI Compass: Provide advice on cloud technologies, including containers; Hands-on-help with container testing; Consult on FAIR aspects of data management; Provide advice on data visualization with proofs of concept;

Current Working Groups (WGs):

1. *Data Exchanges*
2. *Container/Cloud Computing*
3. *Data Visualization*

https://www.neonscience.org/ncar-neon-community-collaborations

NCAR
neon
Abby Road NEON / ABBY
Time-series of Daily average Latent Heat Flux at ABBY
NCAR-NEON community workshop on 11/09/21



CI Compass

TECHNICAL REPORT

ci-compass.org

Making the Major Facilities Data Lifecycle FAIR

Charles F. Vardeman II
Date Published: January 25, 2022

What is FAIR data?

The notion of the four foundational principles for "data" — Findability, Accessibility, Interoperability, and Reusability or "FAIR" — was proposed by Wilkinson et al. in "*The FAIR Guiding Principles for scientific data management and stewardship*" **[1]** and envisages a set of first principles for research communities with respect to the management and curation of scientific data. These principles were created from the point of view that data should be *structured* in a way that the data itself is "smart data" which can be queried for information relative to the four FAIR principles. That is, given the "4 Vs" of big data of Volume, Variety

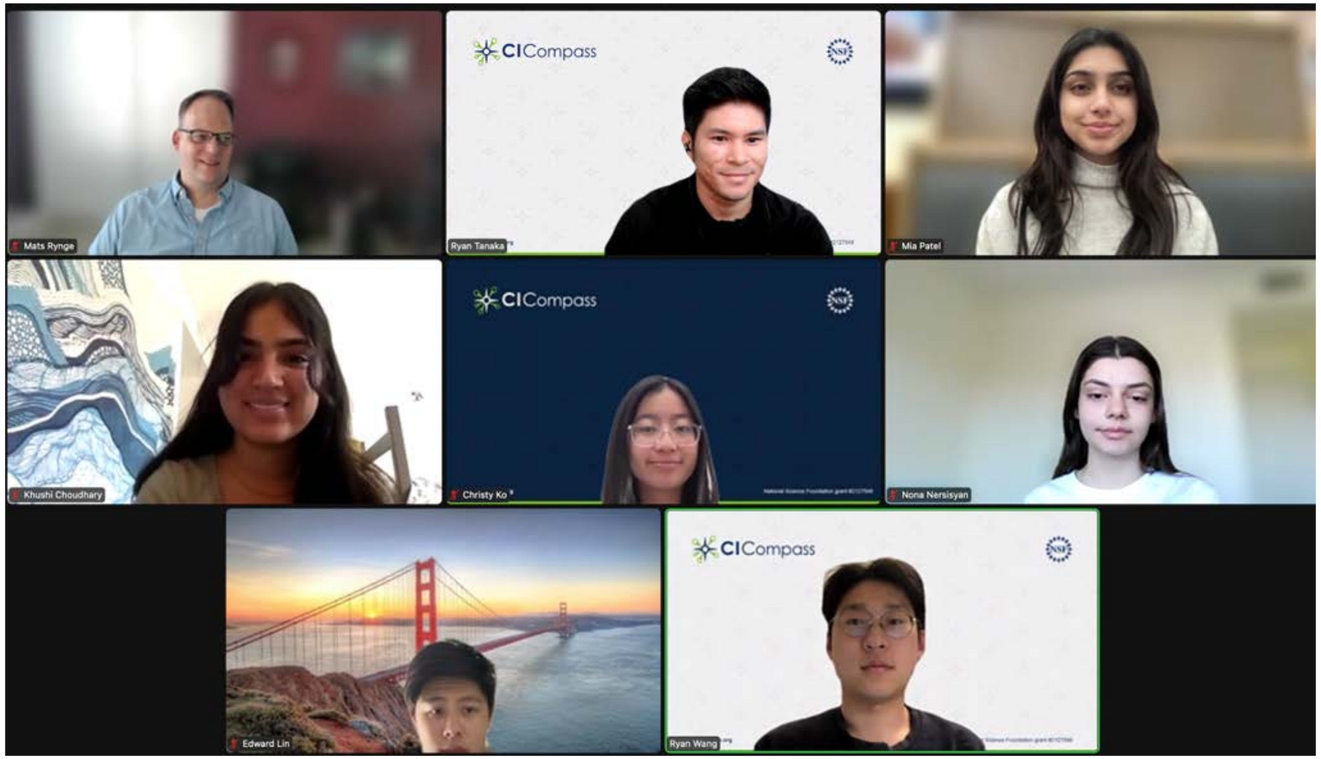
for machine learning, and in particular, *Knowledge Informed Machine Learning* **[2]**, that integrates broader knowledge and context into the machine learning process. Specific attributes for each FAIR principle are contained in **Table 1** and require implementation relative to a specific scientific community through community based recommendations.

Table 1: The FAIR Guiding Principles
To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier

Funded by the National Science Foundation, Grant #2127548

Undergraduate Student Fellowship Program

The program’s goal is to broaden student participation in CI research, development, deployment, and operations.



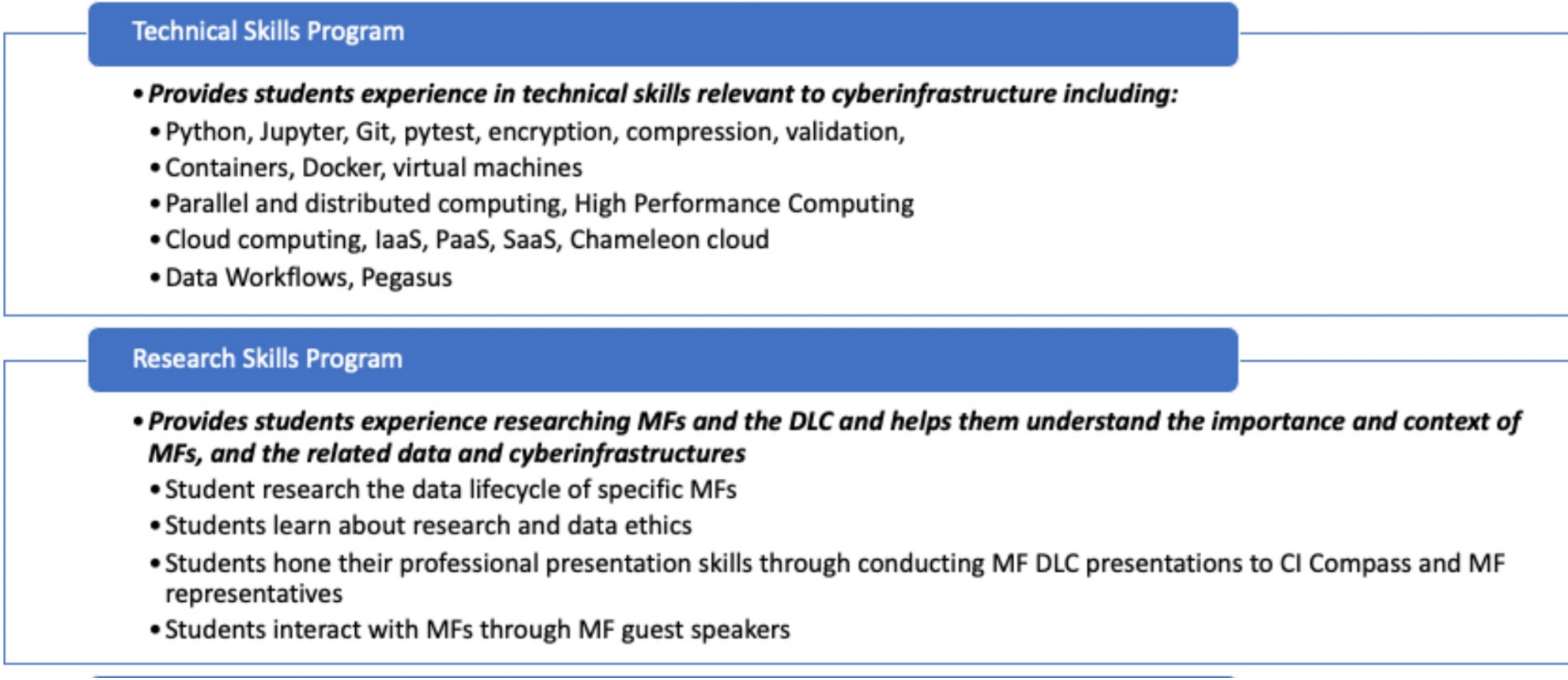
February 28, 2022, Topic: **Software Best Practices II, Containers!**

Year 1: Create a Pilot Framework

Creating program protocols, procedures, and guidelines including:

- Student recruitment, application, and selection,
- Student onboarding and training,
- Student activities with CI Compass, MFs, and other student interns, and
- Pilot program

Curriculum



Technical Skills Program

- ***Provides students experience in technical skills relevant to cyberinfrastructure including:***
  - Python, Jupyter, Git, pytest, encryption, compression, validation,
  - Containers, Docker, virtual machines
  - Parallel and distributed computing, High Performance Computing
  - Cloud computing, IaaS, PaaS, SaaS, Chameleon cloud
  - Data Workflows, Pegasus

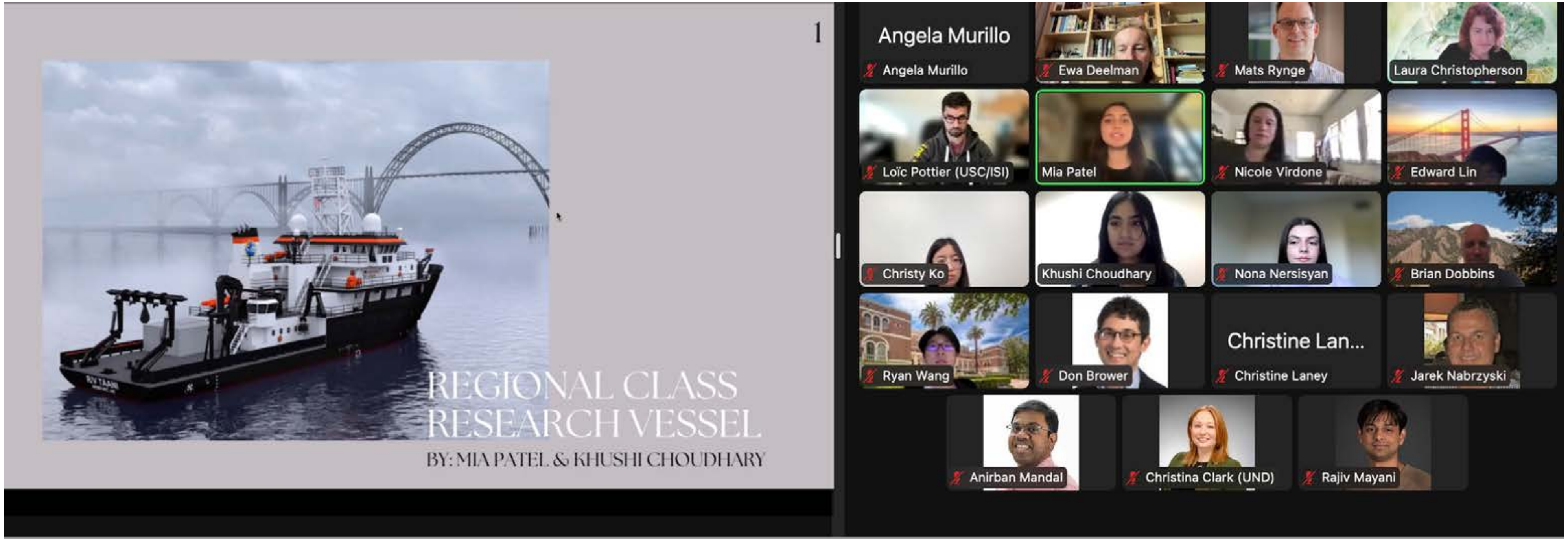
Research Skills Program

- ***Provides students experience researching MFs and the DLC and helps them understand the importance and context of MFs, and the related data and cyberinfrastructures***
  - Student research the data lifecycle of specific MFs
  - Students learn about research and data ethics
  - Students hone their professional presentation skills through conducting MF DLC presentations to CI Compass and MF representatives
  - Students interact with MFs through MF guest speakers

CI Compass

Funded by the National Science Foundation, Grant #2127548

Student Fellowship Program (Year 1 Pilot)



1

REGIONAL CLASS RESEARCH VESSEL

BY: MIA PATEL & KHUSHI CHOUDHARY

April 13, 2022 **Student presentations with CI Compass and MF participants**

Undergraduate Students in CS: 5 at USC, 1 at UND

- Spring with CI Compass practice and directed research, Summer Internships: working on projects related to MFs

Years 2+ Hope to expand to external institutions and mentors, Summer's at MFs

Funded by the National Science Foundation, Grant #2127548

Join the Conversation!

To learn more about CI Compass services, leadership, news, upcoming events and our resource library,
please visit **ci-compass.org**

Contact the CI Compass Team with questions or requests by emailing

contact@ci-compass.org

Connect on social media

Twitter
Follow **@CICompass**

LinkedIn
Connect with us **linkedin.com/company/ci-compass**

YouTube
Subscribe to our channel
CI Compass