

Building a Cyberinfrastructure Center of Excellence

Ewa Deelman, USC (PI)

Co-PIs:

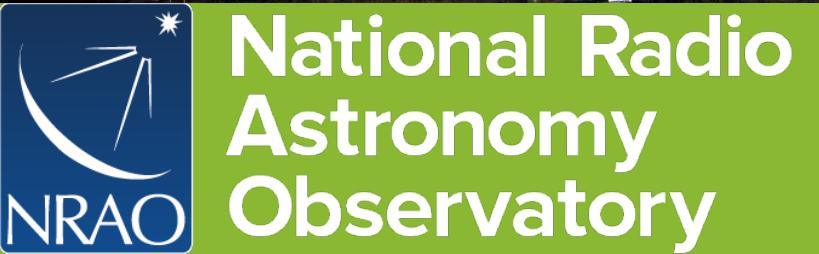
Anirban Mandal, RENCI

Jarek Nabrzyski, Notre Dame University

Valerio Pascucci and Rob Ricci,
University of Utah

Cyberinfrastructure “consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and **people**, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible.”¹

¹ Craig A. Stewart, et al. 2010. “What is cyberinfrastructure?” SIGUCCS '10. ACM, New <http://doi.acm.org/10.1145/1878335.1878347>



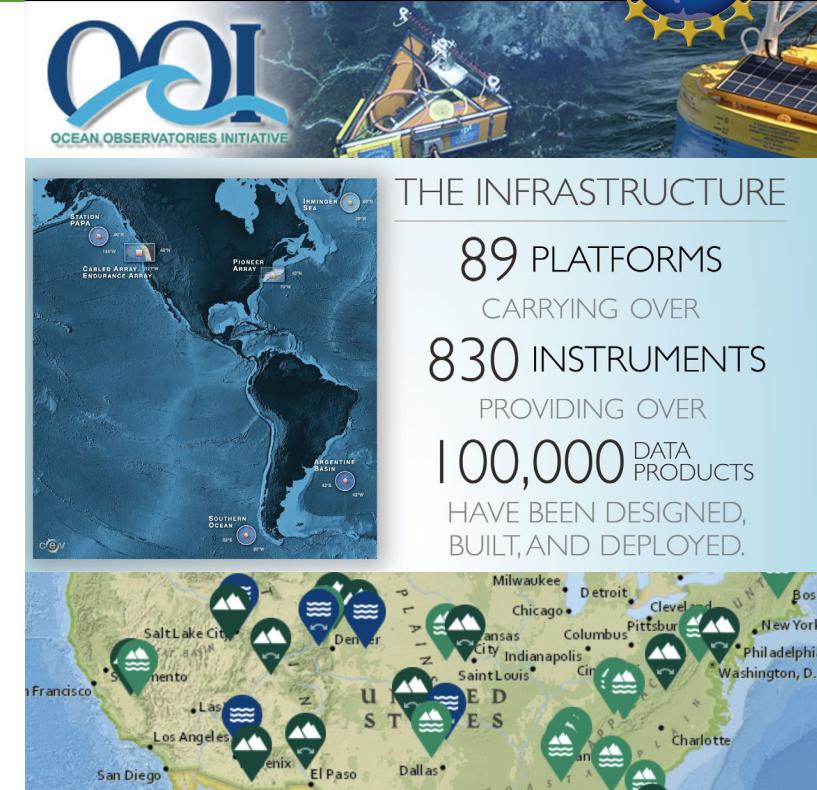
**Searching for
gravitational
waves**

**Understanding ocean
and coastal
ecosystems**

**Looking for
exoplanets**

Studying climate

The National Ecological Observatory Network: Open data
to understand how our aquatic and terrestrial ecosystems are
changing.



Manish Parashar (PI and Chair), Rutgers University and OOI
Stuart Anderson, LIGO

Ewa Deelman, USC

Valerio Pascucci, University of Utah

Donald Petrasick, LSST

Ellen M. Rathje, NHERI



September 2017 Workshop report at <http://facilitiesci.org/>

- Establish a **center of excellence** (following a model similar to the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure, CTSC) as a resource providing expertise in CI technologies and effective practices related to large-scale facilities as they conceptualize, start up, and operate.
- Foster the creation of a facilities' CI community and establish mechanisms and resources to enable **the community to interact, collaborate, and share**.

Develop a model and a plan for a Cyberinfrastructure Center of Excellence

- Dedicated to the enhancement of CI for science
- Platform for knowledge sharing and community building
- Forum for discussions about CI sustainability and workforce development and training
- Key partner for the establishment and improvement of large-scale projects with advanced CI architecture designs
- Partnering with other community efforts (TrustedCI, ResearchSOC, SGCI, OSG,...) to support science

<http://cicoe-pilot.org/>

2018- 2021

USC

Ewa Deelman (PI)
Mats Rynge
Karan Vahi
Loïc Pottier
Rafael Ferreira da Silva
Wendy Whitcup



Automation, Resource Management, Workflows, Project Management

RENCI

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



Resource Management, Networking, Clouds, Social Science

University of Notre Dame

Jarek Nabrzyski (Co-PI)

Jane Wyngaard

Charles Vardeman

Mary Gohsman



University of Utah

Valerio Pascucci, Rob Ricci (Co-PIs)

Marina Kogan

Steve Petruzza



Indiana University

Angela Murillo

Data Archiving



Trusted CI
Susan Sons
Josh Drake



Workforce development, Sensors, operations, Semantic technologies

Data management, visualization, clouds, large-scale CI deployment, Crisis Informatics, Social Computing

Cybersecurity

1. Recognize the expertise, experience, and mission-focus of Large Facilities
2. Engage with and learn from current LFs CI
3. Build on existing knowledge, tools, community efforts
 - Avoid duplication, seek providing added value,
4. Build expertise, not software
5. Develop proof of concepts that can enhance particular LF's CI
 - Keep a separation between our efforts and the LF's CI developments
6. Work with the LFs and the CI community on a blueprint for the CI CoE

Build partnerships:

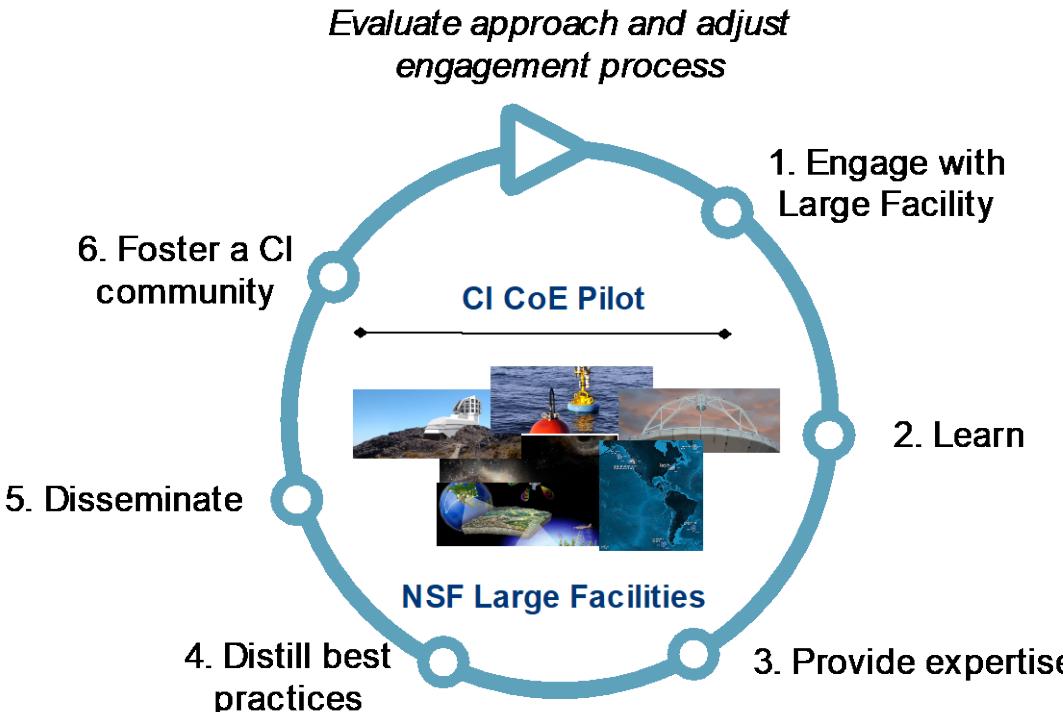
- Trusted CI (identity management): share personnel
- Open Science Grid (data and workload management): share expertise
- Science Gateways Community Institute (portals): share expertise
- Campus Research Computing Consortium (CaRCC): workforce development

Advisory Board

- **Stuart Anderson**, Caltech, LIGO
- **Pete Beckman**, ANL, Northwestern University
- **Tom Gulbransen**, Battelle, NEON
- **Bonnie Hurwitz**, University of Arizona
- **Miron Livny**, University of Wisconsin, Madison, OSG
- **Ellen Rathje**, University of Texas at Austin
- **Von Welch**, Indiana University, Trusted CI
- **Michael Zentner**, UCSD, SGCI



Developing and improving Engagement Model



Process for Engagement with a Facility

- Engage at the management level, potentially seek introductions from NSF PO, participate in meeting (LF Workshop, LF CI Workshop)
- Initial virtual technical group discussions to define possible avenues of engagement
- In person meeting/virtual with a number of technical personnel
- Identify topics for engagement
- Set up working groups
- Follow up email and conference call discussions focused on particular topics/working groups
- Bigger group discussions/checkpointing
- Reports of engagement, gather feedback from the project engaged

National Ecological Observatory Network



NEON provides a coordinated national system for monitoring critical ecological and environmental properties at multiple spatial and temporal scales.

...transformative science
development

...workforce



20 ecoclimatic domains
distinct landforms,
vegetation, climate, and
ecosystem dynamics.

Terrestrial sites:
terrestrial plants, animals, soil,
and the atmosphere,

Aquatic sites: aquatic
organisms, sediment and
water chemistry,
morphology, and hydrology.

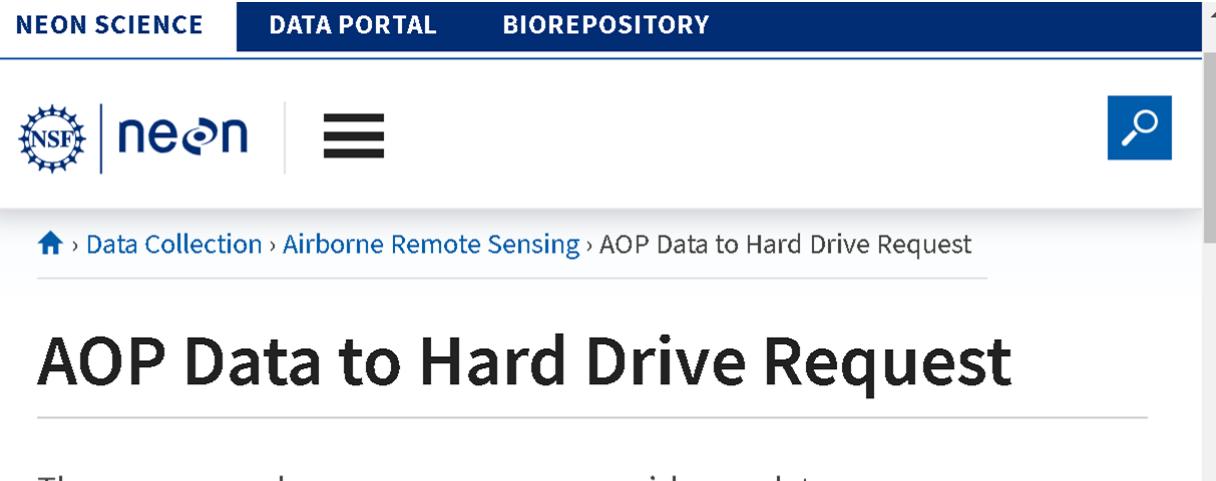
Data collection over 30 years

27 Relocatable terrestrial sites

13 Relocatable aquatic sites

Before

NEON SCIENCE DATA PORTAL BIOREPOSITORY



The screenshot shows the NEON Data Portal interface. At the top, there are tabs for 'NEON SCIENCE', 'DATA PORTAL' (which is selected), and 'BIOREPOSITORY'. Below the tabs is the NEON logo and a search icon. The main content area shows a breadcrumb navigation path: 'Data Collection > Airborne Remote Sensing > AOP Data to Hard Drive Request'. The title 'AOP Data to Hard Drive Request' is prominently displayed. The text below states: 'There are several ways users can access airborne data:' followed by a bulleted list.

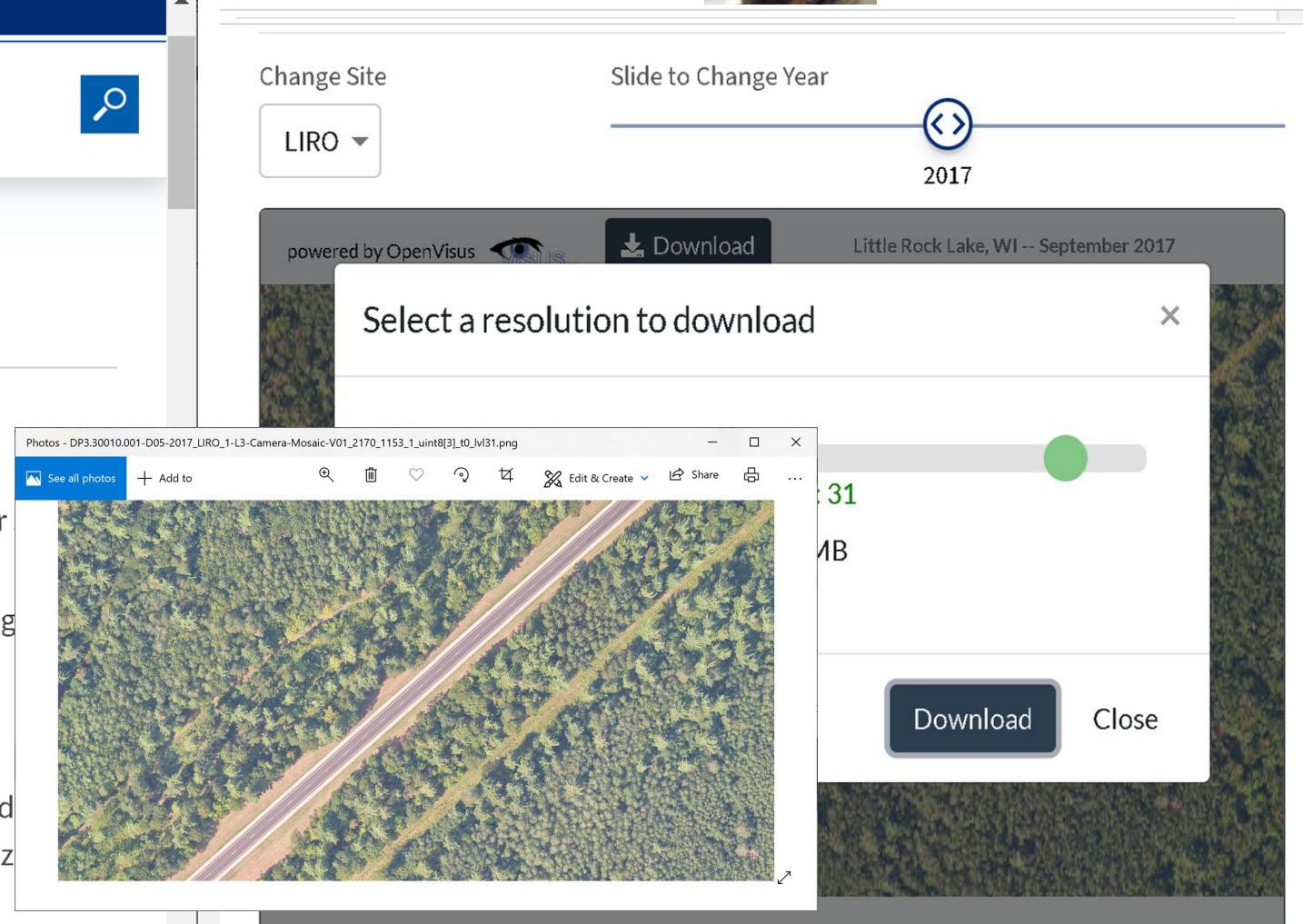
- Download the data from the [NEON data portal](#) (recommended for amounts of data)
- Programmatically access the data with the [NEON Data API](#) or using the [NEON Utilities](#) GitHub repo (>1 GB downloads)
- Mail in a hard drive to receive your data

Please fill out the form below if you are interested in receiving a hard drive of AOP data, and we will respond with a recommended hard drive size as well as mailing instructions.

After



Steve Petruzza, Utah



The screenshot shows a modern, interactive data access interface. At the top, there are buttons for 'Change Site' (set to 'LIRO') and 'Slide to Change Year' (set to '2017'). Below this is a preview image of an aerial photograph of a road through a forest. A modal dialog box is open, titled 'Select a resolution to download'. It shows a preview of the image and includes a resolution selection slider (set to 31 MB) and a download button. The background shows a larger version of the same aerial image.

Working group	Goals	Products
Data Capture	Develop demonstrators and comparisons of the multiple architectures for data capture at the sensor to data deposition in a repository	<ul style="list-style-type: none"> Proof of Concept: architecture demo on github: https://github.com/cicoe/SensorThingsGost-Balena
Data Processing	Provide support and distill best practices for workflows and services related to the processing of data.	<ul style="list-style-type: none"> Paper: "Exploration of Workflow Management Systems Emerging Features from Users Perspectives" (Workshop on Big Data Tools)
Data Storage, Curation, & Preservation	Compare and be able to consult on different data storage, curation and preservation technologies.	<ul style="list-style-type: none"> Document: Competency questions based on scenarios that domain experts may use Google dataset search for NEON dataset discovery Presentation: at ESIP on schema.org Small containerized prototype of publishing neon vocabularies as linked data and linked data connection
Identity Management	Understand current practice in authentication and authorization and help mature practice across the NSF Large Facilities.	<ul style="list-style-type: none"> Production deployment: Connection to CI Logon NEON data download (using existing university / organization credentials) https://cert-data.neonscience.org/home Paper: NEON IdM Experiences (NSF Cybersecurity Summit)

CI CoE Pilot Benefits to NEON Thus Far

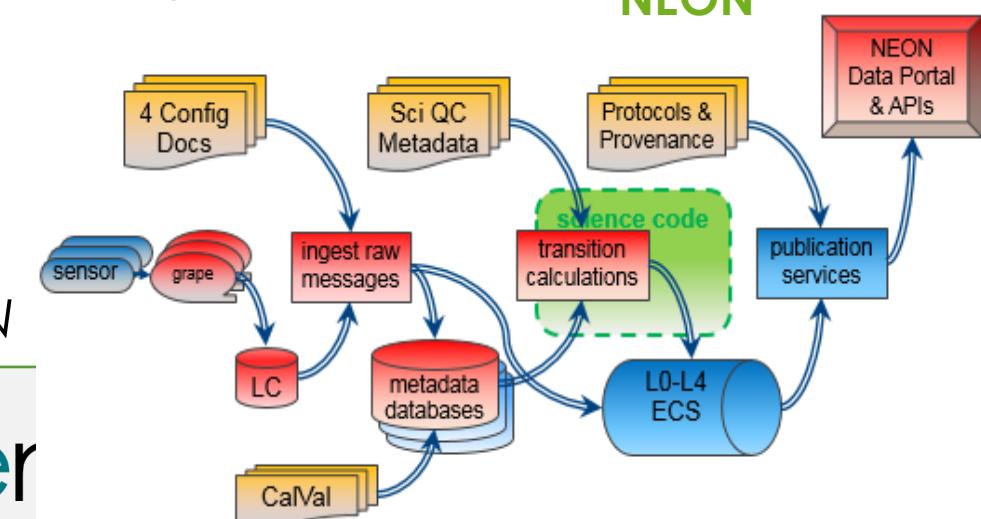
- Short ramp-up due to receptivity/readiness to change
- Broadened network of expert CI colleagues
- Major upgrade to Data Portal's remote sensing visualization
- Accelerated Data Portal completion plan
- Affirmed strategies for workflow, messaging, & DR
- Raised critical mass of attention on semantics & schema.org
- Excited software developers
- Escalated accountability of CI
- More coming

Slide courtesy of Tom Gulbransen, NEON



Tom Gulbransen

NEON



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

CI CoE Pilot Engagement

- Consult on cloud technologies, including data, containers, etc.
- Helped inform NEON/NCAR's proposal to NSF
- Expect to engage further if/when funded

2020, ~2 months

Goal: Provide CI expertise on a Common Cloud Platform (CCP) to pilot the migration of their collective data resources and services from individual in-house solution to common Cloud

Engagement

- Embedded in a number of CCP working groups
- Provided high-level schematics and guidance for contents, organization and structuring of the CCP **high-level requirements** draft
- Presented existing best practices for designing a **ConOps** document
- Discussed CI **best practices** for various aspects of **cloud architecture design** for CCP
- Drafted a companion document to inform CPP **platform design** (cloud migration considerations for data, resource orchestration, cloud storage, messaging, supporting FAIR data principles, Identity Management (in collaboration with Trusted CI)

SAGE: Seismological Facilities for the Advancement of Geoscience
GAGE: Geodetic Facilities for the Advancement of Geoscience

April 2020 -

- **Deep engagement:**
 - Identify a topic that is important and not-yet fully solved by the LF,
 - Conduct focused discussions, mix of virtual and in-person presence, hands-on work
 - Includes an engagement template that defines scope, sets expectations, identifies products
 - Work products: documents/papers, proof of concepts, schema implementations, demos
- **Topical discussions:**
 - Identify a topic that is important to a number of LFs
 - Facilitate virtual discussions, sessions at conferences, collect and share experiences, distill best practices
 - Discover opportunities for shared infrastructure
- **Community building: bringing in new members to the CI CoE Pilot effort**
 - Identify related efforts
 - Collect information and disseminate information about the broad community activities
 - Maintain a living resource for community information
- **Each engagement has a working group with a leader and a set of work products.**

Identity Management WG (in Collaboration with Trusted CI)

Goal: Disseminate IDM information

- Monthly meetings with speakers and discussions on topics relevant to LFs: e.g. CILogon
- Engagements, primarily focusing on federated identity management
- Issues of identifying data usage and enabling reporting
- ARF and GAGE, exploring solutions and developing demonstrations
- SCIMMA, contributing to their Identity Access Management prototype



2020 Cyberinfrastructure/Cybersecurity Workshop
August 18 and August 20, 2020
<https://cics-workshop.org>

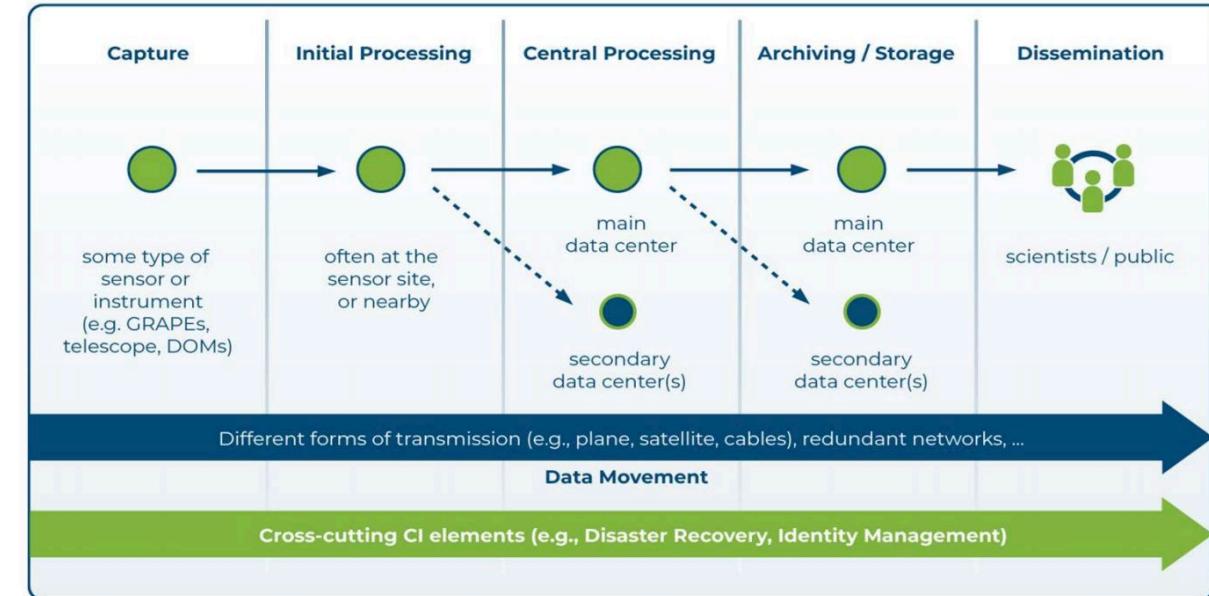
Engaging with the community to discuss issues of:

- Big data visualization
- Cloud migration
- ConOps
- Fair data
- LF data lifecycle
- Science workflows
- Workforce development

Data Lifecycle

Goal: Understand the LF DLC to design CI CoE

22 representatives from 9 LFs participated in the interviews: CHESS, GAGE, IceCube, LHC-CMS, NHERI/Designsafe, NHERI/RAPID, NOIR, OOI, and SAGE.



Workforce Development

Goal: Understand how large facilities attain, develop, and grow their workforce, and identify challenges they face in doing so

15 representatives from 12 LFs: ARF, Arecibo, CHESS, IceCube, NHERI (Converge, Designsafe, NCO, SimCenter), NHMFL, NOIR, NRAO, OOI

IN PROGRESS

Mission

CI CoE provides expertise and active support to cyberinfrastructure practitioners at NSF Large Facilities in order to accelerate the data life cycle and ensure the integrity and effectiveness of the cyberinfrastructure upon which research and discovery depends.

- Continue Engagement (deep, working groups, community)
- Finalize the Blueprint for the CI CoE
- Propose the CI CoE to NSF

<http://cicoe-pilot.org/>

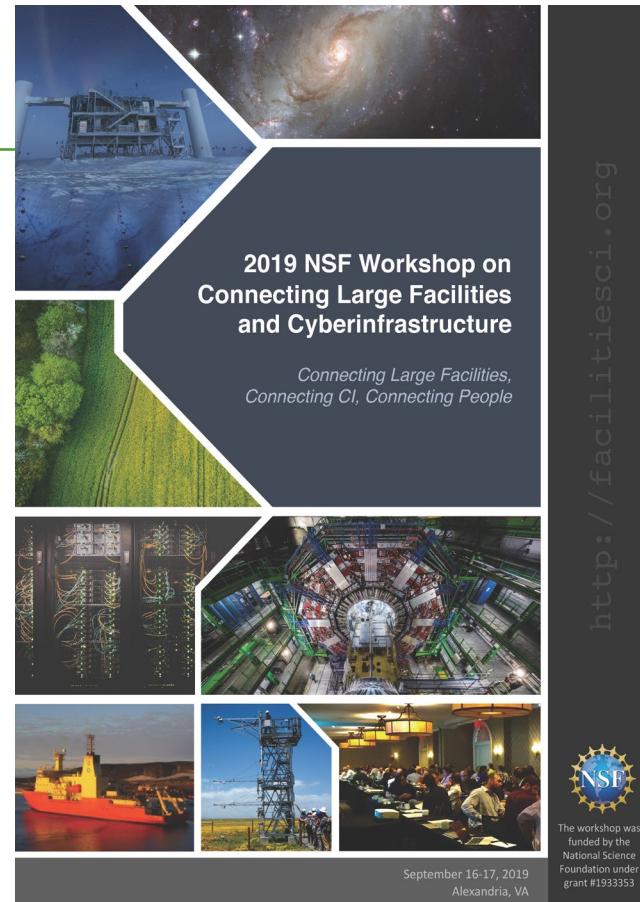




NSF workshop on Smart CI (2020)



<http://smartci.sci.utah.edu/>



NSF workshops on CI for Large Facilities (2017,
2019)

<http://facilitiesci.org>