CI CoE PILOT

# Cyberinfrastructure Center of Excellence Pilot: Connecting Large Facilities Cyberinfrastructure

**Ewa Deelman**,

University of Southern California, PI

Co-PIs:

**Anirban Mandal**, RENCI

**Jarek Nabrzyski**, Notre Dame University

**Valerio Pascucci** and **Rob Ricci**,
University of Utah

Funded by the National Science Foundation
Grant #1842042

USC Viterbi School of Engineering
*Information Sciences Institute*

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

renci

TRUSTED CI
THE NSF CYBERSECURITY CENTER OF EXCELLENCE

# Cyberinfrastructure

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

CI CoE PILOT

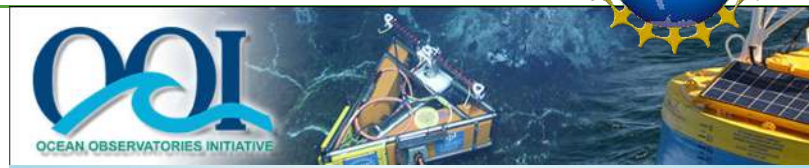
# CI is a critical component of Science: Large Facilities (LFs)

**Searching for gravitational waves**

**Understanding ocean and coastal ecosystems**

**Looking for exoplanets**

**Studying climate**






NRAO **National Radio Astronomy Observatory**


neon National Ecological Observatory Network

USC Viterbi School of Engineering *Information Sciences Institute*

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

renci

TRUSTED CI THE NSF CYBERSECURITY CENTER OF EXCELLENCE

# 2017 Workshop on Cyberinfrastructure for Large Facilities

**Manish Parashar** (PI and Chair), Rutgers University and OOI
**Stuart Anderson**, LIGO
**Ewa Deelman**, USC
**Valerio Pascucci**, University of Utah
**Donald Petravick**, LSST
**Ellen M. Rathje**, NHERI


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

- Understand **best practices** of current CI architecture and operations at the large facilities.
- Identify common requirements and solutions as well as CI elements that can **be shared across facilities.**
- Enable CI developers to most effectively target CI needs and the **gaps** of large facilities.
- Explore opportunities for **interoperability** between the large facilities and the science they enable.
- Develop guidelines, mechanisms, and processes that can assist future large facilities in constructing and **sustaining their CI.**
- Explore **mechanisms and forums** for evolving and sustaining the conversation and activities initiated at the workshop.
- Generate recommendations that can serve as inputs to current and future NSF CI related programs.

USC Viterbi
School of Engineering
*Information Sciences Institute*

# Workshop Key Recommendations



- **Establish a center of excellence** (following a model similar to the NSF-funded Trusted CI) 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**.
- Support the creation of a **curated portal and knowledge base** to enable the discovery and sharing of CI-related challenges, technical solutions, innovations, best practices, personnel needs, etc., across facilities and beyond.
- Establish structures and resources that bridge the facilities and that can strategically address **workforce development, training, retention, career paths, and diversity**, as well as the overall career paths for CI-related personnel.

# Questions We Posed


- Are we ready to build a CI community?
- How do we build a CI community?
- How do we enhance collaborations across large facilities and CI projects?
- How do we capture knowledge, effective practices in a way that is relevant, evolving, and impactful?
- How do we maintain and enhance/increase the CI talent pool?

# Project Team

## USC

Ewa Deelman
Mats Rynge
Karan Vahi  Loïc Pottier
Rafael Ferreira da Silva
Ryan Mitchell

*Automation, Resource Management, Workflows*

## RENCI

Anirban Mandal
Ilya Baldin
Laura Christopherson
Paul Ruth
Erik Scott

*Resource Management, Networking, Clouds, Social Science*

# Project Team

**University of Notre Dame**
Jarek Nabrzyski
Jane Wyngaard
Charles Vardeman



*Workforce development, Sensors, Semantic technologies*

**University of Utah**
Valerio Pascucci, Rob Ricci,
Marina Kogan
Steve Petruzza




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

**Trusted CI**
Susan Sons
Ryan Kiser


*Cybersecurity*

USC Viterbi School of Engineering Information Sciences Institute
University of Notre Dame
The University of Utah
renci
Trusted CI The NSF Cybersecurity Center of Excellence

CI CoE PILOT

# Project Goals


Award #1842042

## 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
- Key partner for the establishment and improvement of Large Facilities with advanced CI architecture designs
- Grounded in re-use of dependable CI tools and solutions
- Forum for discussions about CI sustainability and workforce development and training
- Pilot a study for a CI CoE through close engagement with NEON and further engagement with other LFs and large CI projects.

10/2018– 9/2020

# Advisory Board

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








# Overall Strategy

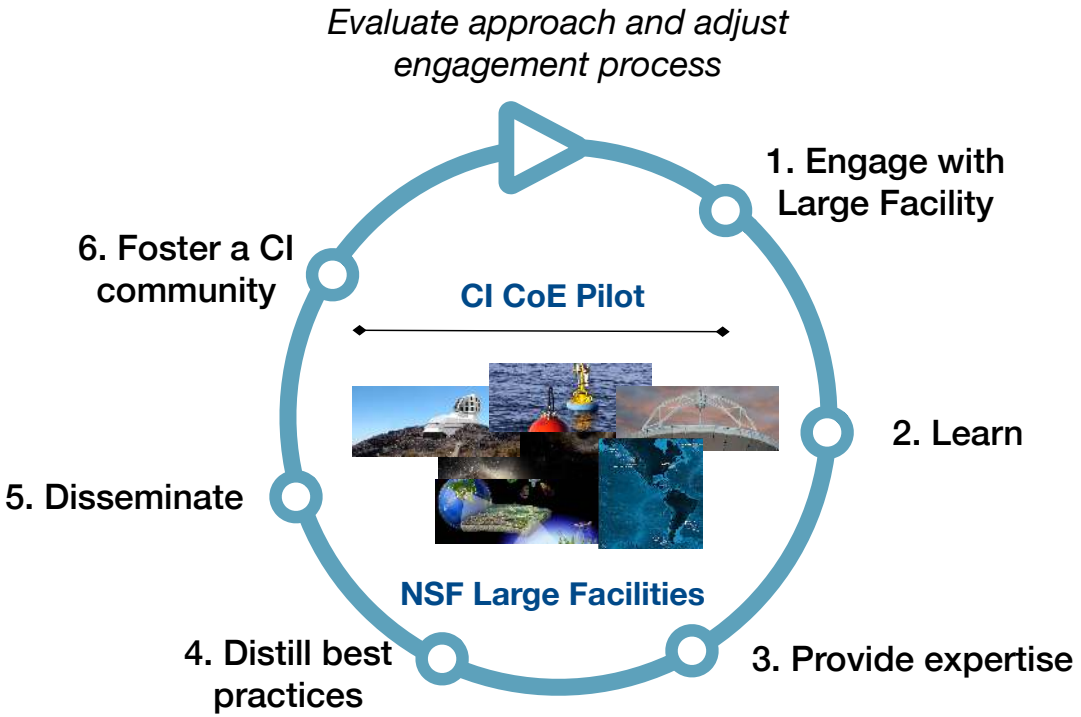
1. Recognize the expertise, experience, and mission-focus of Large Facilities
2. Engage with and learn from current LFs CI
3. Build on existing knowledge, tools, community efforts
   -Avoid duplication, seek providing added value,
4. Prototype solutions that can enhance particular LF's CI
   -Keep a separation between our efforts and the LF's CI developments
5. Build expertise, not software
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
- Campus Research Computing Consortium (CaRCC): workforce development

# Engagement Methodology

## 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, Trusted CI)
- Initial virtual technical group discussions to define possible avenues of engagement
- In person meeting with a number of technical personnel
- Identity 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 Mission

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

…transformative science development

…workforce








*Slide courtesy of Tom Gulbransen, NEON*

**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

# NEON Engagement, October 2018-

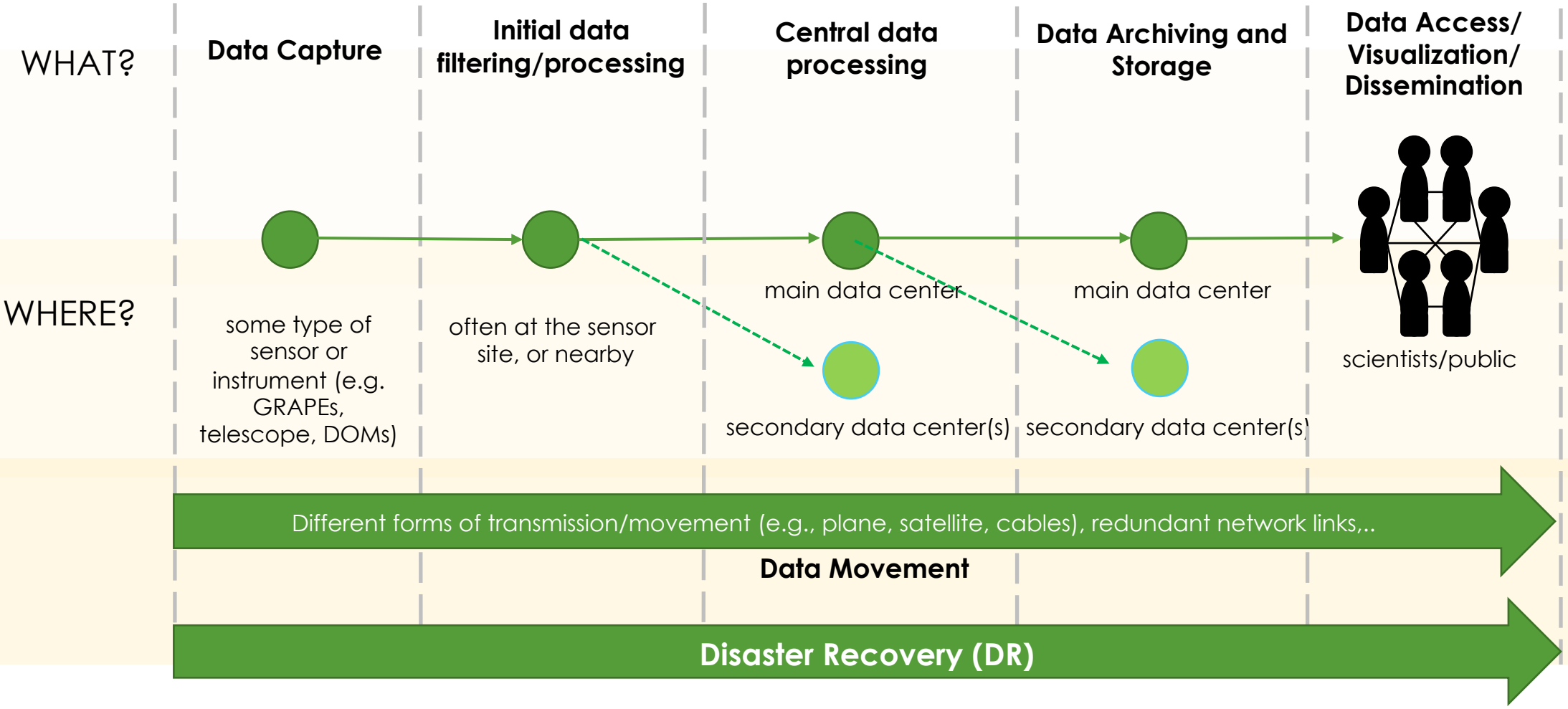


- Engagement facilitated by NSF
- Engagement Goals:
  - Increase Pilot's understanding of NEON's cyberinfrastructure architecture and operations
  - Increase NEON's understanding of the Pilot's goals and expertise
  - Select & scope mutually beneficial opportunities to prototype or learn from CI methods
- Engagement Process
  - In-person management meeting
  - NEON shared a number of design documents
  - Team conference calls
  - Meeting with NEON
    - November 2018: Identified topics and formed working groups
    - August 2019: took stock, summarized


- Data Life Cycle and Disaster Recovery
- Data Capture
- Data Processing
- Data Storage/Curation/Preservation
- Data Visualization/Dissemination
- **Identity management**
  - **2:30pm presentation: "NEON and CI CoE Pilot vs. Identity Management, a story": Jeremy Sampson, Ryan Kiser, Terry Fleury**
- Engagement with Large Facilities

# Data Lifecycle for LFs

Anirban Mandal, lead



# Existing Disaster Recovery (DR) Strategies for Some LFs across the Data Life Cycle

| | Data Capture | Processing/ Filtering | Data Movement | Data Archiving and Storage | Data Access |
|---|---|---|---|---|---|
| **NEON** | GRAPES can buffer up to 1 month of data.<br>Replace GRAPE if fails. | No failover for compute - onsite or offsite. | If 2nd data center is built, we might see some replication there (TBD). | Replication (cloud) using Wasabi for backup of ECS.<br>Plans for 2nd data center in Wyoming (for just data replication). | No existing DR strategies.<br>Fail overs?<br>Availability guarantees?<br>SLA? |
| **OOI** | Replacement?<br>How long is this data kept in retrieval locations (e.g., Pacific City)? How much is buffered or cached? | West Coast isn't used for processing but could be.<br>Has plans for failover. | Redundant network links between East and West. | West Coast replicates data from East Coast.<br>No automatic failover, but has plans. | Failovers planned for user access.<br>Availability guarantees?<br>SLA? |
| **IceCube** | Replace with a spare. | Good separation between remote and central processing.<br>Distributed processing provides resilience. | Different ways to transmit - plane, satellite, Internet.<br>GridFTP to both DESY and NERSC. | At least 4 copies of data in different locations: 1 copy kept at the South Pole, 1 each in UW, DESY and NERSC. | Availability guarantees?<br>SLA?<br>Varies based on the caching solution (e.g. xrootd). |
| **LSST** | Base Facility has a copy of data.<br>Significant buffering planned for anticipated network failures. | Multiple facilities do processing of different types.<br>No failover or redundant processing capabilities. | Redundant connection from BASE to NCSA.<br>Protection against network failures for Summit to Base and from Base to NCSA | 3 copies of data reside in different places: Base facility in Chile, NCSA, CC-IN2P3 (France) | Different means of access including different Data Access Centers, via web, via APIs.<br>Availability guarantees?<br>SLA? |
| **Strategies for LFs** | • Caching/buffering<br>• Backup copies<br>• Replace with a spare | • Failover compute sites? | • Plans for failover<br>• Redundant connections | • Data replication<br>• Backup services | • Automatic failovers for data access<br>• Multiple data access points |

# Data Presentation – New interactive data access

## Before

NEON SCIENCE | DATA PORTAL | BIOREPOSITORY

neon

› Data Collection › Airborne Remote Sensing › AOP Data to Hard Drive Request

### AOP Data to Hard Drive Request

There are several ways, users can access airborne data:

- 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 of AOP data, and we will respond with a recommended hard drive siz well as mailing instructions.

## After

Steve Petruzza, Utah

Change Site: LIRO

Slide to Change Year: 2017

powered by OpenVisus | Download | Little Rock Lake, WI -- September 2017

Select a resolution to download

Download | Close

Photos - DP3.30010.001-D05-2017_LIRO_1-L3-Camera-Mosaic-V01_2170_1153_1_uint8[3]_t0_lvl31.png

See all photos | Add to | Edit & Create | Share

# Working Groups and Products co-lead by the Pilot and NEON

| 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 | • **Prototype**: architecture demo on github: https://github.com/cicoe/SensorThingsGost-Balena |
| **Data Life Cycle & Disaster Recovery** | Develop a general set of DR requirements and policies that can inform the LFs about best practices for DR and how those can be adapted for specific facilities. | • **Document:** Disaster recovery template<br>• **Document:** Filled out template example (IceCube)<br>• **Webinar**: Best Practices for NSF Large Facilities: Data Life Cycle and Disaster Recovery Planning |
| **Data Processing** | Provide support and distill best practices for workflows and services related to the processing of data. | • **Paper:** "Exploration of Workflow Management Systems Emerging Features from Users Perspectives" (in submission) |
| **Data Storage, Curation, & Preservation** | Compare and be able to consult on different data storage, curation and preservation technologies. | • **Document**: Competency questions based on scenarios that domain experts may use Google dataset search for NEON dataset discovery<br>• **Presentation**: at ESIP on schema.org<br>• Small containerized **prototype** of publishing neon vocabularies as linked data and linked data connection |

# Working Groups and Products

| Working group | Goals | Products |
|---|---|---|
| **Data Visualization & Dissemination** | Understand the access, visualization and user interaction workflows in large facilities. Distill best practices and provide solutions to improve the access and usability of the available data. | • **Document** describing AOP data visualization cyberinfrastructure<br>• **Online demo and video**: Visualizing AOP Data-- https://cert-data.neonscience.org/data-products/DP3.30010.001 |
| **Identity Management** | Understand current practice in authentication and authorization and help mature practice across the NSF Large Facilities. | • **Production deployment**: Connection to CI Logon NEON data download (using existing university / organization credentials) https://cert-data.neonscience.org/home<br>• **Paper:** NEON IdM Experiences (NSF Cybersecurity Summit) |
| **Engagement with Large Facilities** | Engage with Large Facilities and other large cyberinfrastructure projects to foster knowledge and effective practice sharing; 2) define avenues of engagement, modes of engagement, and plan community activities. | • **Document**: LF engagement template<br>• **Presentations:** SCIMMA project meeting, 2019 LF meeting, PEARC'19, LF CI Workshop, Cybersecurity Summit'19<br>• **Paper:** Invited e-Science 2019 paper |

USC Viterbi School of Engineering Information Sciences Institute
University of Notre Dame
The University of Utah
renci
Trusted CI The NSF Cybersecurity Center of Excellence
Contact: Ewa Deelman, deelman@isi.edu

# Engagement with Large Facilities
# Lessons Learned

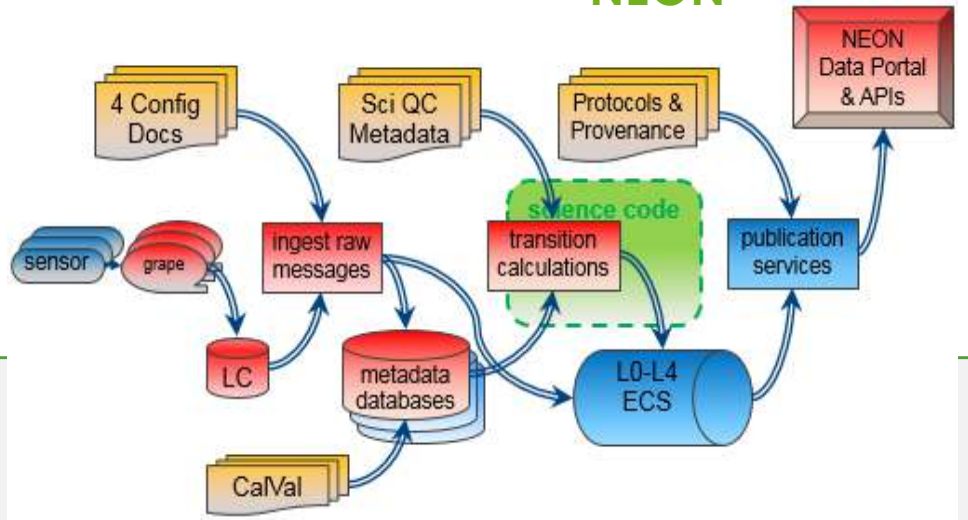

1. Importance of f2f discussions, building relationships and trust
2. Benefits of formalizing the engagement: expectation, timelines, resources to use
3. Importance of LF priorities and challenges, importance of good timing
4. Organizing work around working groups and work products
5. Be open to learn about what works, don't fix it (workflow management)
6. Co-existence of old and new systems, making for a heterogeneous CI landscape

## 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

**Tom Gulbransen**

**NEON**




*Slide courtesy of Tom Gulbransen, NEON*

# Expanding Engagement with Large Facilities

- **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, prototypes, 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:**
  - **Connect CI professionals**
  - Collect information and disseminate information about the broad community activities
  - Maintain a living resource for community information
  - Develop new partnerships
- **Each engagement has a working group with 1-2 leaders and a set of work products.**

# 2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure


**Technical:**

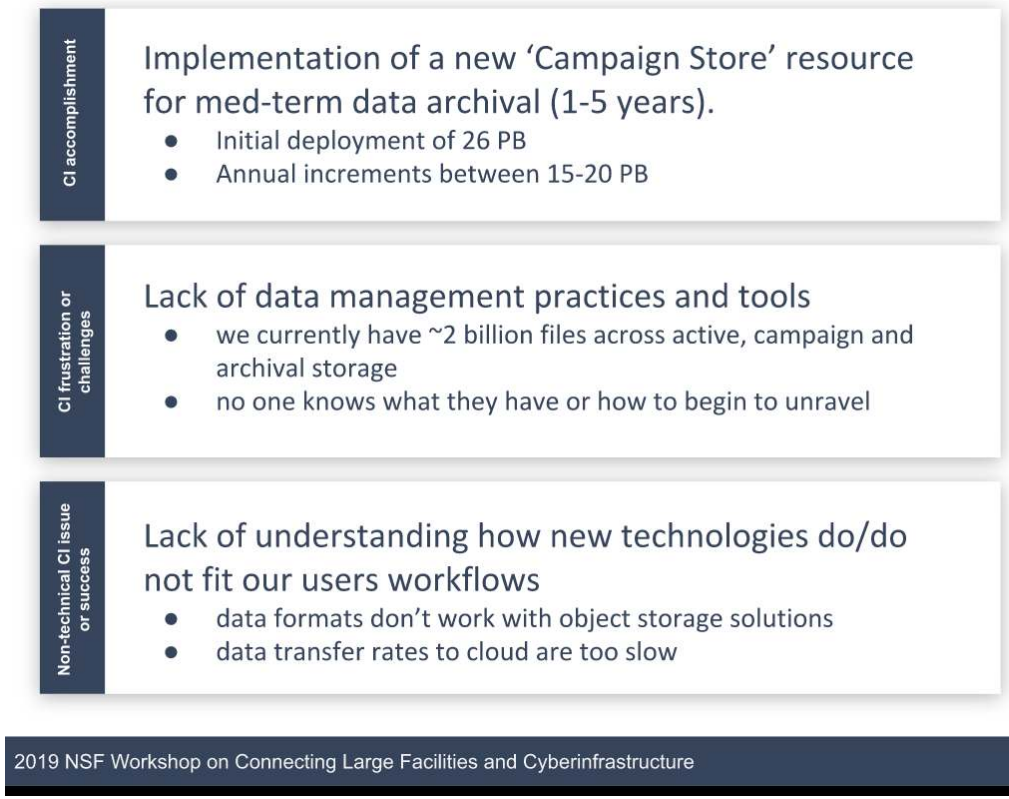
- What are the CI challenges that need to be addressed to support LF science?
- Where does LF CI end and the user CI begin (issues of data sharing, reproducibility)?
- Can we better utilize current CI investments?
- What are the opportunities to share CI services?

**Socio-technical:**

- What are the opportunities for collaboration amongst LFs and other Large CI projects?
- What are the non-technical issues that influence CI development and how they can be collaboratively addressed?
- Enhancing the CI workforce: what are the challenges and solutions?
- How can we build a CI community: what are the impediments and opportunities?

# Connecting LFs, CI, People

- CI Calling Cards (**61**):
  - Biggest CI accomplishment,
  - Biggest CI frustration or challenge
  - Non-technical frustration or accomplishment when building CI
  - **We will make them searchable and expand**





Pamela Hill
pjg@ucar.edu
NCAR


HPC Storage Architectures

Project URL

http://facilitiesci.org

# Engagement with Large Facilities
# Future Plans

1. Reaching out to other large facilities
2. Gathering feedback on the data lifecycle abstraction
3. Mapping the data lifecycle to CI capabilities and services
4. Discovering opportunities for CI sharing
5. Defining new working groups and discussion topics
   - Broadening the disaster recovery discussion
   - Data archiving and preservation
   - CI workforce enhancement, training


http://cicoe-pilot.org

ci-coe-pilot@isi.edu

Ewa Deelman deelman@isi.edu

- Connecting LF CI workshop, 2019: http://facilitiesci.org