CI CoE PILOT

Cyberinfrastructure Center of Excellence Pilot

Funded by the National Science Foundation
Grant #1842042

Ewa Deelman, USC (PI)

Co-PIs:
Anirban Mandal, RENCI

Jarek Nabrzyski, Notre Dame University

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

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 "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

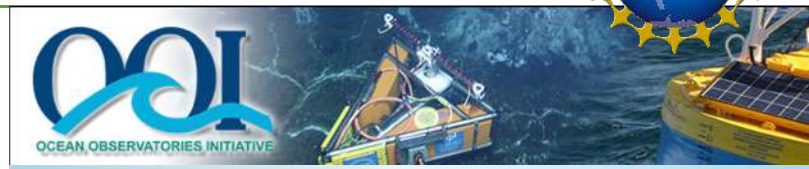
CI is a critical component of Science: Large Facilities (LFs)

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

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.

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

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








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

Slide courtesy of Tom Gulbransen, NEON

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

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 • **Document:** Filled out template example (IceCube) • **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"
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 • **Presentation**: at ESIP on schema.org • Small containerized **prototype** of publishing neon vocabularies as linked data and linked data connection

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

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 • **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 • **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 • **Presentations:** SCIMMA project meeting, 2019 LF meeting, PEARC'19, LF CI Workshop, Cybersecurity Summit'19 • **Paper:** Invited e-Science 2019 paper

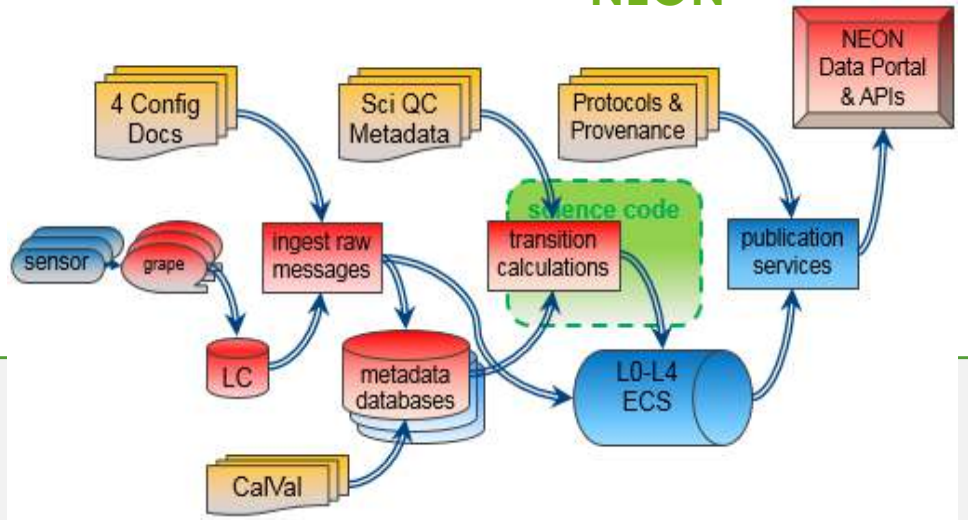
USC Viterbi School of Engineering Information Sciences Institute
University of Notre Dame
The University of Utah
renci
Trusted CI The NSF Cybersecurity Center of Excellence
Contact: Ewa Deelman, deelman@isi.edu

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

Tom Gulbransen

NEON



Slide courtesy of Tom Gulbransen, NEON

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 (e.g. workflow management)
6. Co-existence of old and new systems, making for a heterogeneous CI landscape

Engagement with Large Facilities Future Plans

1. Reaching out to other Large Facilities
   - Deep engagement, topical discussions, community building
2. Gathering feedback on the data life cycle abstraction
3. Mapping the data life cycle 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

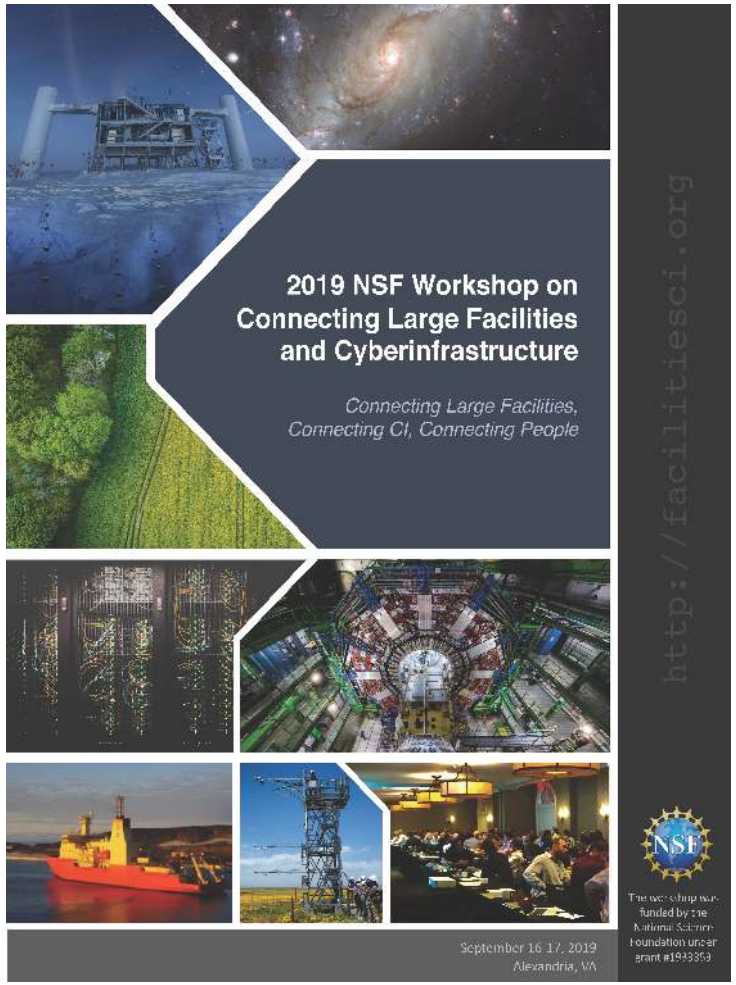
Engagement with Large Facilities
Accomplishments and Work Products

- **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: 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.**

Engagement with Large Facilities
Future Plans

1. Developing a blueprint for the CI CoE:
   a. Community needs
   b. Areas of focus
2. Reaching out to other large facilities
3. Gathering feedback on the data lifecycle abstraction
4. Mapping the data lifecycle to CI capabilities and services
5. Defining new working groups and discussion topics
   - Identity management (in collaboration with Trusted CI)
   - Trustworthy Data, looking into data integrity (in collaboration with Trusted CI)
   - CI workforce enhancement, training

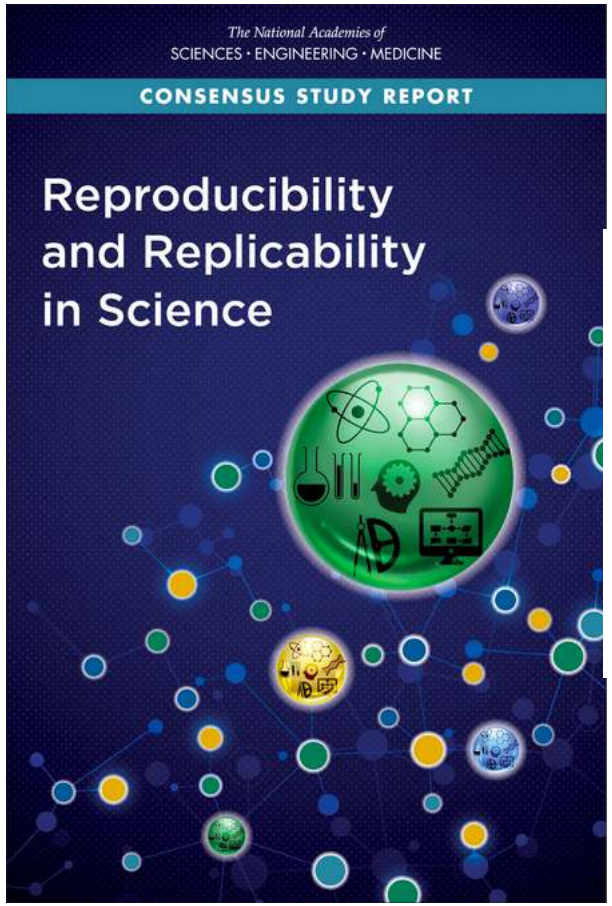
2019 LF CI Workshop Recommendations



- Create opportunities for CI discovery and sharing of existing solutions, services, training resources amongst the LFs as well as CI projects.
- Create a common location of knowledge about CI best practices with system descriptions, architectures, use cases, and core system tools.

http://facilitiesci.org

2019 National Science Foundation Report



RECOMMENDATION 6-3: Funding agencies and organizations should consider investing in research and development of open-source, usable tools and infrastructure that support reproducibility for a broad range of studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

Role of Reproducibility in CI CoE

- Examining issues of reproducibility in the context of the data lifecycle
  - What are the challenges and approaches to reproducibility within the LFs?
  - What services are used today to enhance reproducibility?
  - What services or service behavior are needed to support reproducibility?
  - Potentially set up a working group to share ideas and experiences

http://cicoe-pilot.org

ci-coe-pilot@isi.edu

Ewa Deelman deelman@isi.edu

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