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)

NSF

**Searching for gravitational waves**

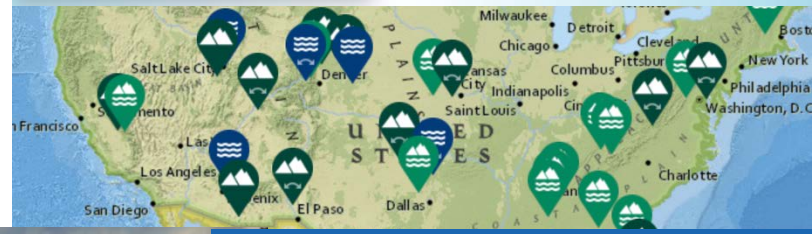
**Understanding ocean and coastal ecosystems**

**Looking for exoplanets**

**Studying climate**

THE INFRASTRUCTURE

89 PLATFORMS
CARRYING OVER
830 INSTRUMENTS
PROVIDING OVER
100,000 DATA PRODUCTS
HAVE BEEN DESIGNED, BUILT, AND DEPLOYED.




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

# Recognizing the importance of CI in 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/

- **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**.

# 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.

CI CoE PILOT

# 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
Erik Scott
Paul Ruth

*Resource Management, Networking, Clouds, Social Science*

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 Team

NSF

## 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

# 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**, SDSC

# 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

- Data Life Cycle and Disaster Recovery
- Data Capture
- Data Processing
- Data Storage/Curation/Preservation
- Data Visualization/Dissemination
- Identity management
- Engagement with Large Facilities

# Data Life Cycle for LFs

Anirban Mandal, lead

| | Data Capture | Initial data filtering/processing | Central data processing | Data Archiving and Storage | Data Access/ Visualization/ Dissemination |
|---|---|---|---|---|---|
| WHAT? | | | | | |
| WHERE? | some type of sensor or instrument (e.g. GRAPEs, telescope, DOMs) | often at the sensor site, or nearby | main data center; secondary data center(s) | main data center; secondary data center(s) | scientists/public |

Different forms of transmission/movement (e.g., plane, satellite, cables), redundant network links,..

**Data Movement**

**Disaster Recovery (DR)**

# 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 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<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 |

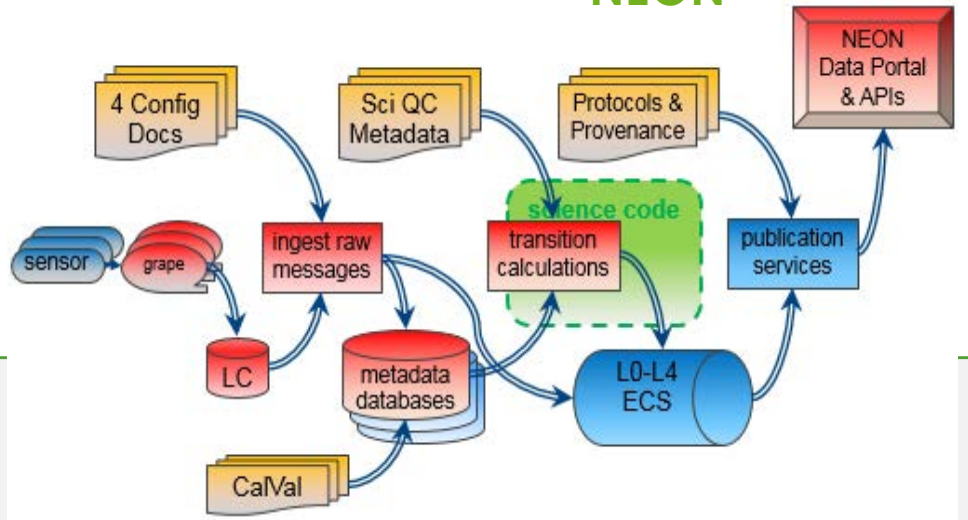
USC Viterbi School of Engineering Information Sciences Institute
University of Notre Dame
The University of Utah
renci
Trusted CI The NSF Cybersecurity Center of Excellence
Contact: Ewa Deelman, deelman@isi.edu

## 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

# Preliminary Engagement with ARF

- In-person meetings with selected ARF members, with focus on CI aspects
  - Shipboard technical support team (Scripps) and other ARF personnel
  - At community events like the NSF Cybersecurity Summit
- Ongoing engagement as part of Trusted CI effort
  - Identity Management (IdM)
  - LF Security team
- Invited R2R to provide an overview to CICoE Pilot team during project call in October
  - Understood data management operations undertaken by R2R on behalf of the ARF.
  - Data delivery to R2R from ship cruises: Issues and data arrival windows.
  - Data storage and archiving at various facilities including NOOA NCEI, Amazon.
  - Data QC and data processing activities.
  - Data movement issues.
  - Data dissemination using R2R portal.

# Lessons Learnt for CICoE Pilot During Preliminary Engagement with ARF

- Gained understanding of the **data life cycle for ARF CI** operations with respect to R2R.
- Learnt the **importance of real-time data delivery**, processing and QC for robust science.
- Learnt about significant progress made by the ARF on **standardization of instruments** on-board for uniform data collection, and continued improvement in CI design for newer ships (e.g. RCRV).
- Learnt the **critical role of Marine Technicians**, including the wide range of their functions and required training; learnt the complexity of ship operations, e.g. ship scheduling.
- Learnt about **systematic workforce development activities** undertaken by ARF, e.g. **MATE** program.
- Learnt the **unique aspects of CI functions for ARF** because the ships are mobile platforms with long cruises operating under harsh environmental conditions.

# CI Challenges for the ARF

- Difficult operating conditions of the ARF fleet result in some **CI Technology Challenges**
  - **Networking challenges:** more sat. bandwidth desirable but budget constraints; Cloud solutions not feasible.
  - **Real-time transmission** for QC and rapid turnaround: not implemented for most vessels.
  - **Robust science**: Delayed QC makes it difficult to ensure that the quality of data collected by RVs is adequate and accurate for scientific studies.
  - **Diversity of sensors**, some of which might be mitigated in future designs, is still a challenge.
  - **Obsolescence**: Computing/storage/network equipment bought years ago might be deployed years later.
- The ARF also faces some **non-technical/policy/operational CI Challenges**
  - **CI plan:** There are no/limited definitions for CI best practices for the ARF; No clear data management plan.
  - Need to design/deploy/plan for IT services/**CI solutions both on Ship Ops side and Tech Services side**.
  - No dedicated funding stream for **CI activities in** the **NSF solicitation**.
  - ARF is not well positioned to respond to **possible future regulatory** pushes with respect to CI.
  - Current organizational structure of ARF prevents effective communication regarding CI across Ship Operators / ARF entities, which might lead to potential duplication of activities.

# How Can CICoE Pilot and ARF Work Together to Address CI Challenges?

- Work together to **develop a CI and data management plan** for the ARF
  - Help to create a **Data Life Cycle (DLC) model** and CI services supporting DLC stages for the ARF by leveraging CICoE's experience in developing similar DLC models for other LFs.
  - Help ARF articulate CI requirements and planning, which will *increase ROI and shared efficiency, and decrease TCO* across ship operators. Help identifying commonalities in CI requirements and solutions across ship operators.
  - Help in designing a CI plan that is **cross-departmental, consistent and continuing**.
- CICoE can help **make connections with other LFs** facing similar issues by communicating and potentially helping ARF to leverage CI solutions from other LFs that faced similar pain points, and vice versa.
  - E.g. in data collection: there maybe some lessons learned we can distill to share with other LFs or bring in some experiences from NEON or other LFs/projects/communities (e.g. ESIP, RDA).

# How Can CICoE Pilot and ARF Work Together to Address CI Challenges?

- CICoE can **provide expertize in specific technical aspects** by researching CI best practices and potentially prototyping
  - E.g. Standardization of shipboard instrumentation data collection (schema, frequency, common language and formats) - an overview of the best practices in this area will inform ARF.
  - E.g. Research on approaches for real time data delivery and feedback, including evaluation of data delivery with messaging systems with high-latency communication channels.
  - E.g. Identity Management as part of existing Trusted CI engagement.
- CICoE can help articulating the importance of CI for ARF by providing an **external balanced view** to stakeholders and funding agencies.
- This is our current understanding and we are **open to feedback and other ideas !!!**

http://cicoe-pilot.org

ci-coe-pilot@isi.edu

Ewa Deelman deelman@isi.edu
Anirban Mandal anirban@renci.org

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