CI CoE PILOT

Funded by the National Science Foundation Grant #1842042

Next Generation Cyberinfrastructure for Science: Cyberinfrastructure Center of Excellence Pilot for Large Facilities

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

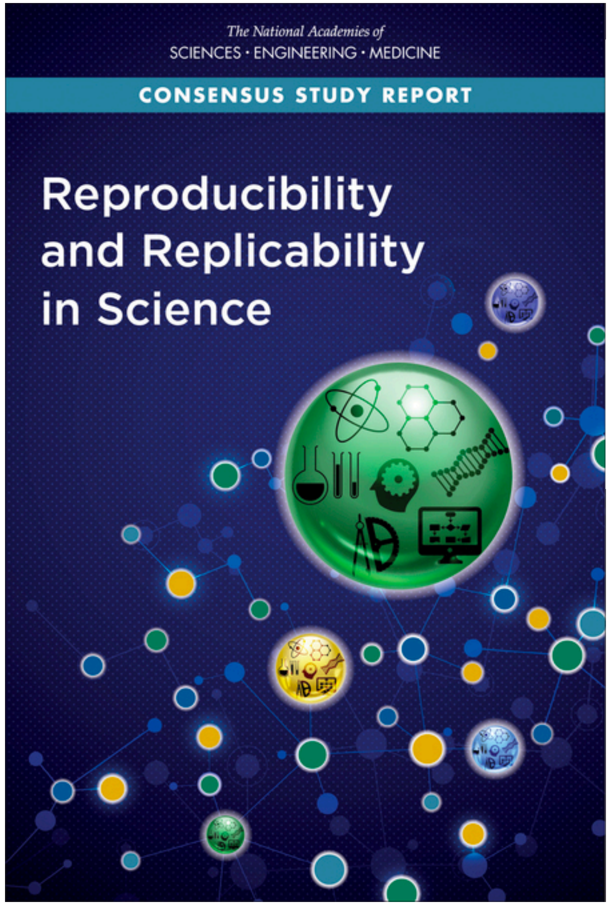
TRUSTED CI THE NSF CYBERSECURITY CENTER OF EXCELLENCE

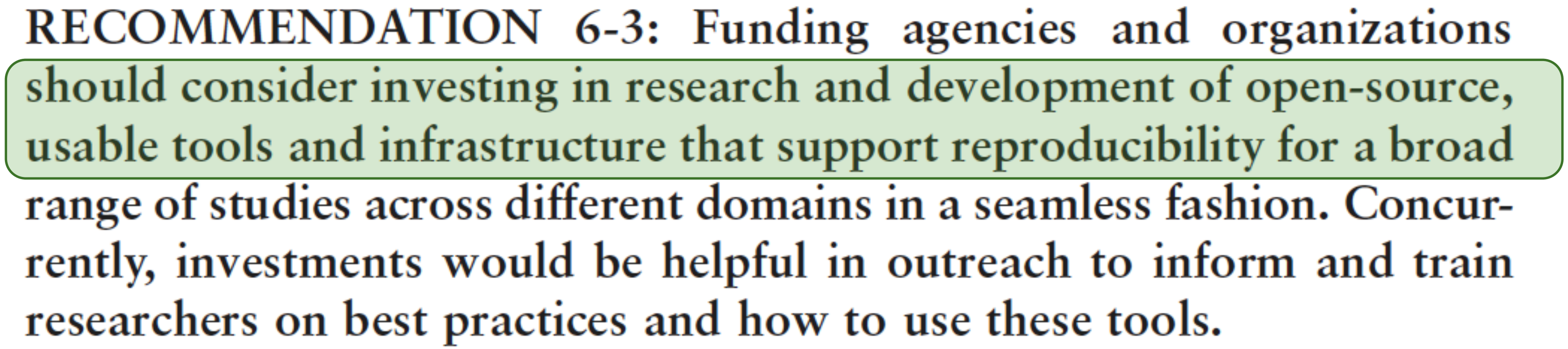
Cyberinfrastructure

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

2019 National Science Foundation Report


The National Academies of
SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
Reproducibility and Replicability in Science



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.

Victoria Stodden, Committee member

https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

CI CoE PILOT

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

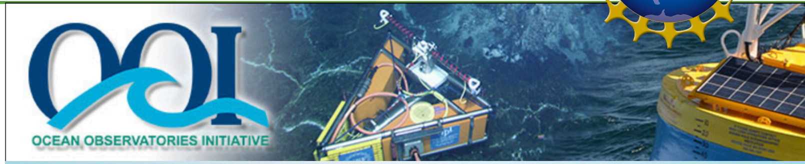
NSF

Searching for gravitational waves

Understanding ocean and coastal ecosystems

Looking for exoplanets

Studying climate



STATION PAPA
CABLED ARRAY
ENDURANCE ARRAY
IRMINGER SEA
PIONEER ARRAY
ARGENTINE BASIN
SOUTHERN OCEAN

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

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

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

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, *this expertise includes recommendations for supporting transparency, reproducibility*
6. Work with the LFs and the CI community on a blueprint for the CI CoE

Build partnerships:

- Trusted CI (identity management, trustworthy data): share personnel, common working groups
- 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

…workforce development

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

Data Storage, Curation and Preservation

Goals:

The goals of this working group is to compare and be able to consult on different data storage, curation and preservation technologies. **Current effort includes helping with metadata and applying schema.org schemas to data from large facilities.**

Charles Vardeman, UND

Team Members:

Pilot: Charles Vardeman, Valerio Pascucci, Steve Petruzza, Giorgio Scorzelli,
NEON: Christine Laney, Steve Jacobs, Tom Gulbransen, Jeremey Sampson

Data Storage, Curation and Preservation Activities

1. Implementation of Schema.org vocabulary markup within data portal landing pages to enable broader data discovery by search engines, mainly Google. Creation of templates based on ESIP science-on-schema.org best practice example in collaboration with Earthcube P418/P419 projects
2. Extension of schema.org vocabularies for earth sciences through ESIP/Earthcube P418/P419 geoshemas.org similar to bioschemas.org for the life sciences
3. Joint modeling -- "Vocamps" to extend out needed vocabularies to be published as linked data resources

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

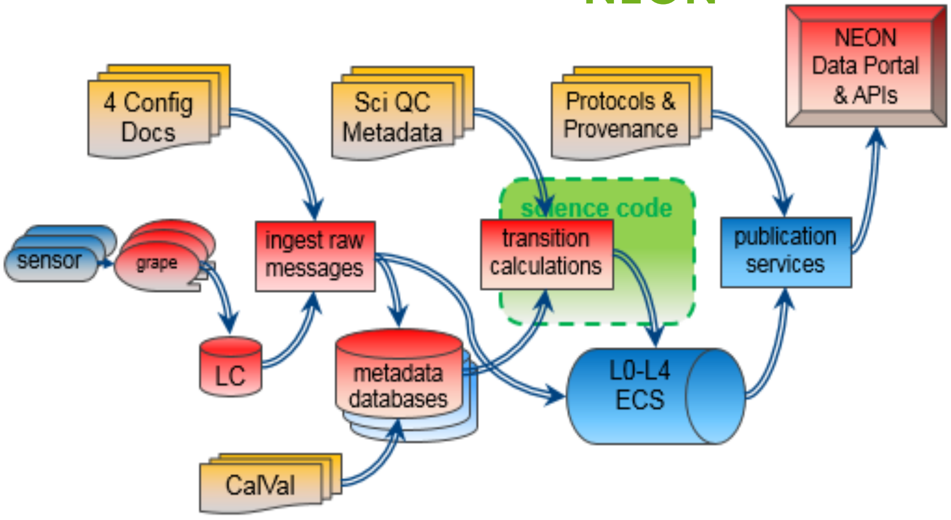
Contact: Ewa Deelman, deelman@isi.edu

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

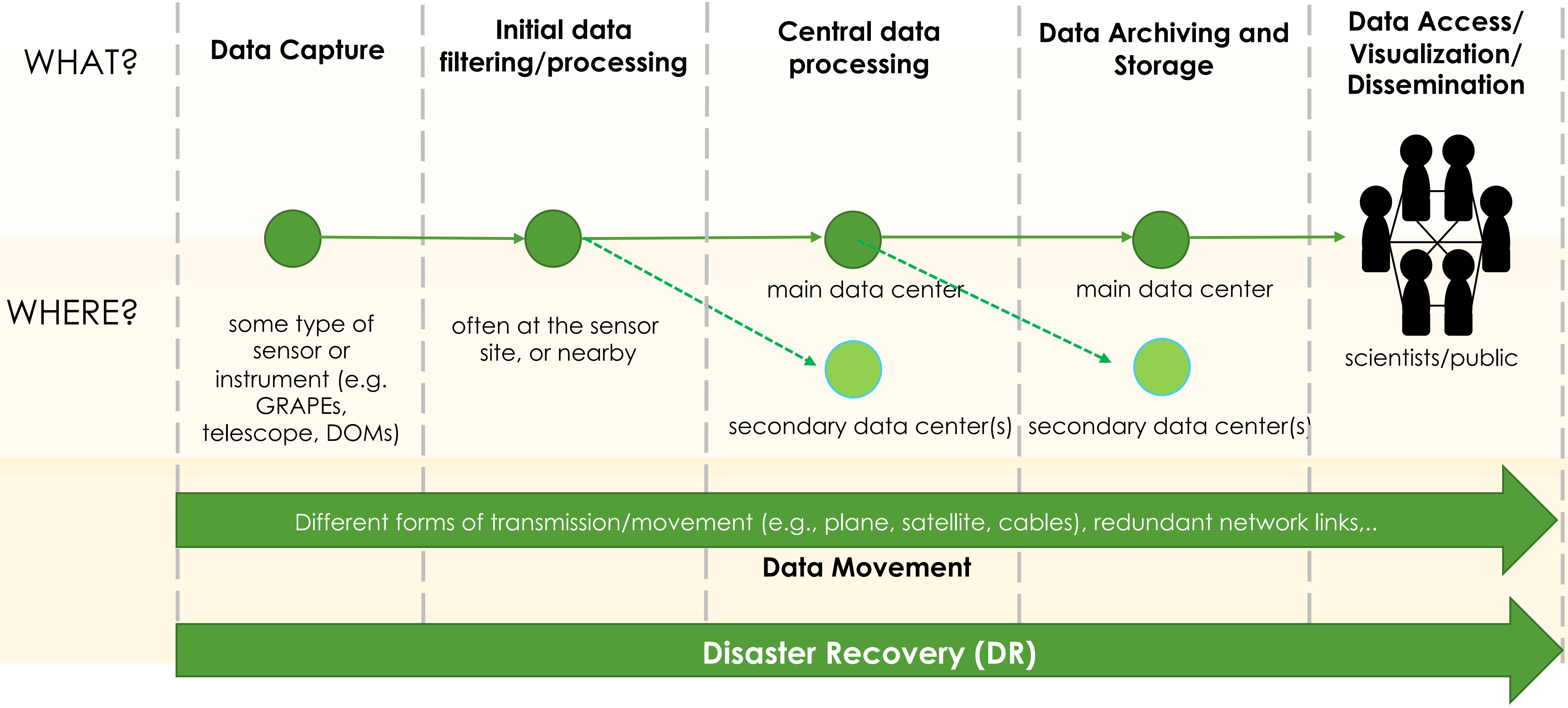
NEON



Slide courtesy of Tom Gulbransen, NEON

Data Lifecycle 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)

What services correspond to the data lifecycle stages?

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 CI services or service behavior are needed to support reproducibility?
  - Potentially set up a working group to share ideas and experiences

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: **identity management, trustworthy data calls (with Trusted CI)**
  - Facilitate virtual discussions, sessions at conferences, collect and share experiences, distill best practices
  - Discover opportunities for shared infrastructure
- **Community building:**
  - Identify related effort
  - 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)
   - CI workforce enhancement, training

Trustworthy Data Working Group

"<u>.. aims to provide guidance on data security for open science, to improve scientific productivity and trust in scientific results.</u> Open science relies on data integrity, collaboration, high performance computing, and scalable tools to achieve results, but currently lacks effective cybersecurity programs that address the trustworthiness of scientific data."

- Open group, calls every Monday 1pm PT/4pm ET
- Community survey

Led by Trusted CI: https://trustedci.org/2020-trustworthy-data

2019 LF CI Workshop Recommendations

2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure

Connecting Large Facilities, Connecting CI, Connecting People

http://facilitiesci.org

The workshop was funded by the National Science Foundation under grant #1933353

September 16-17, 2019
Alexandria, VA

- 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

To find out more

http://cicoe-pilot.org

ci-coe-pilot@isi.edu

Ewa Deelman deelman@isi.edu

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

Trustworthy Data WG: https://trustedci.org/2020-trustworthy-data