CI CoE PILOT

Building a Cyberinfrastructure Center of Excellence

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

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.

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

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

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

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

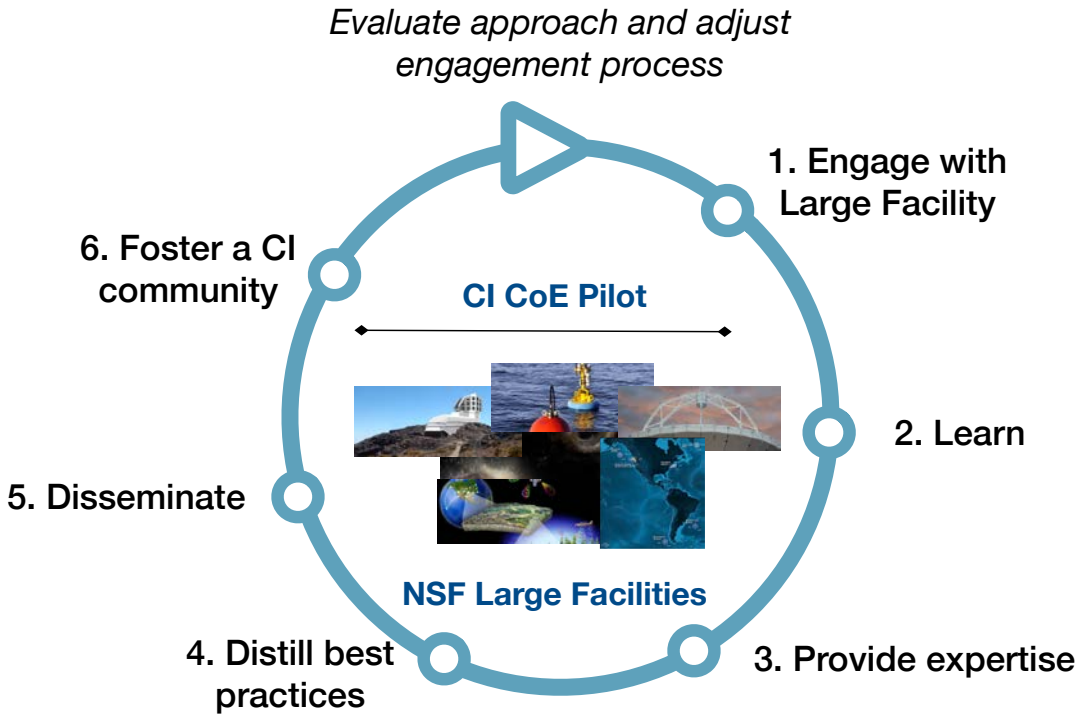







Engagement Methodology

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)
- 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

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 (DLC) and Disaster Recovery (DR)

Goals:

The goals of this working group are to (1) understand the current practices for Disaster Recovery (DR) for the NEON facility and other large facilities (LF) by studying the architectural elements of the CI used by LFs, and (2) 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.

Team Members:

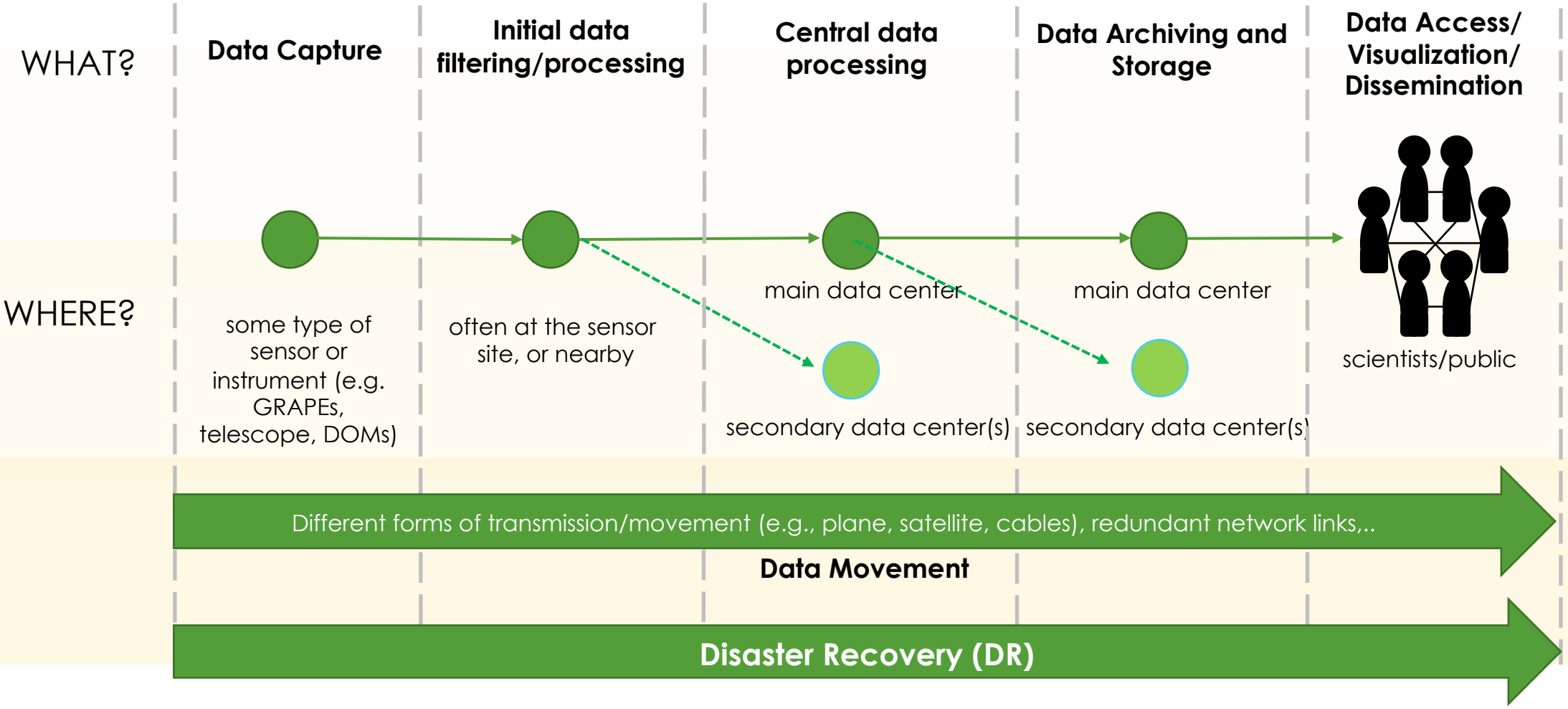
Pilot: Anirban Mandal, Laura Christopherson, Erik Scott, Ilya Baldin, Paul Ruth

NEON: Tom Gulbransen, Phil Harvey, Steve Jacobs

IceCube: Benedikt Riedel

Data Life Cycle for LFs

Anirban Mandal, lead



Data Life Cycle and Disaster Recovery Accomplishments and Work Products

- A preliminary version of **a generalizable DLC model** for LFs
  - Based on engagements with NEON and IceCube and initial literature survey of OOI and LSST facilities.
- **Taxonomy of CI services**, architectures and functionalities that support different DLC stages for four LFs.
- **DR effective processes templates** for NEON and IceCube
  - Can be used by these LFs for DR planning during future procurements/ enhancements and infrastructure planning.
- A DR effective processes template that can be adapted by other LFs using the NEON/IceCube DR planning guides as examples.

Data Life cycle and Disaster Recovery Lessons Learned

- **DLC** is **ONE** way to **learn, reason and catalog the CI functionalities** at each stage of data operation for LFs.
- DLC abstraction helps reasoning about
  - What *services are offered* by each DLC stage ?
  - What *CI architectural elements* support each DLC stage ?
- There are both fundamental *commonalities and differences* across LFs for DLC.
- *Devil is in the details* for both for DLC and DR; Many a time, specific elements or types of data are prioritized.
- *Heterogeneity of data processing* – the priorities of processes handling the data differs according to the type of data. *Heterogeneity of CI tools and stacks.*

Data Capture

Goals:

This working group focuses on the multiple technology stacks required and challenges involved in: (i) data capture at the sensor front end, (ii) pre-process data at the edge, (iii) transport of data from sensors to central processing and archive sites, and (iv) deployment & maintenance of large scale remote sensor systems. The goals of this group are to provide demonstrators and comparisons of the multiple architectures that might be used in accomplishing the above.

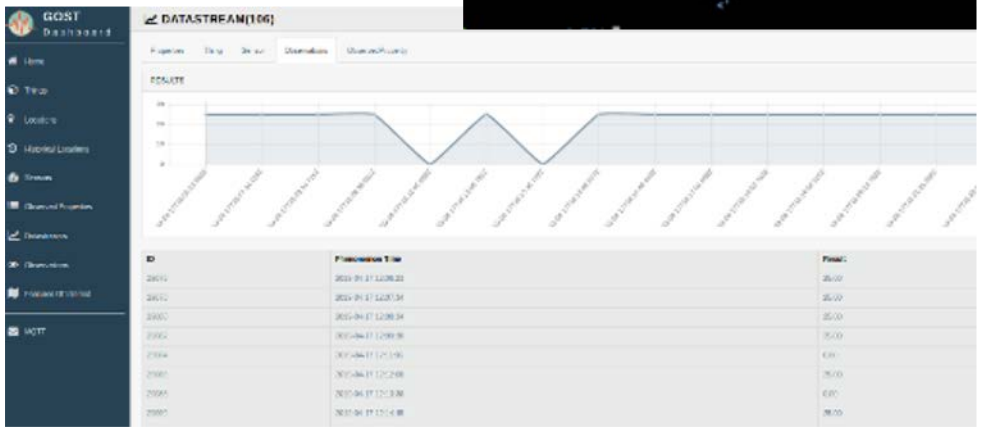
Team Members:

Pilot: Jane Wyngaard, Charles Vardeman II, Robert Ricci

NEON: David Barlow, Steve Jacobs, Thomas Gulbransen, Christine Laney, Laura Leyba-Newton, Santiago Bonarrigo, Dan Allen, John Staarmann

Data Capture
Accomplishments and Work Products

1. Data Movement Architecture
   a. Design Review and discussions
   b. Tech demonstrator 1:
      i. SensorThings (Simple linked metadata)
      ii. Gost (web gui based on Sensor Things)
      iii. Balena (containerised embedded systems)



1. Data Movement Architecture
    a. Lessons for CoE Pilot:
        i. Data Capture/sensor data Engagement Template
        ii. LF operations, processes, constraints
        iii. Hardware development, deployment, maintenance constraints
        iv. Gaps in standardised tooling for this niche

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.

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

Data Storage, Curation and Preservation Lessons Learned

1. Leveraging broader community through organizations like ESIP that can carry forward work beyond engagement period
2. Inclusion of other NSF projects that are working on similar activities during engagement is helpful in exchanging experience as well as creating contact points that persist beyond engagement
3. The need of concrete examples to elucidate potentially complicated implementation details
4. In presenting results at ESIP, broader earth science data community has many questions with the details of "how" to implement schema.org and the relationship of schema.org to existing vocabularies and ontologies in community use

Data Storage, Curation and Preservation Future Plans

1. Joint talk (CI-CoE, EarthCube P419, NEON) at AGU on creation of geoschemas.org
2. Continuation of modeling activities under ESIP Geoschemas Cluster
3. Harmonization of Neon terms in ENVO and SWEET under ESIP Semantic Harmonization Cluster
4. Possible extension of metadata linked data service prototype to include broader alignment to community ontologies and vocabularies
5. Use of metadata linked data service an example for other facilities as part of ESIP geoschemas.org effort

Data Visualization and Dissemination

Goals:

This working group focus on understanding the access, visualization and user interaction workflows in Large Facilities. In particular, we look at how the users explore and interface with the data (e.g., via APIs) for visualization and analysis purposes. Our goal is to learn best practices and provide solutions to improve the access and usability of the available data.

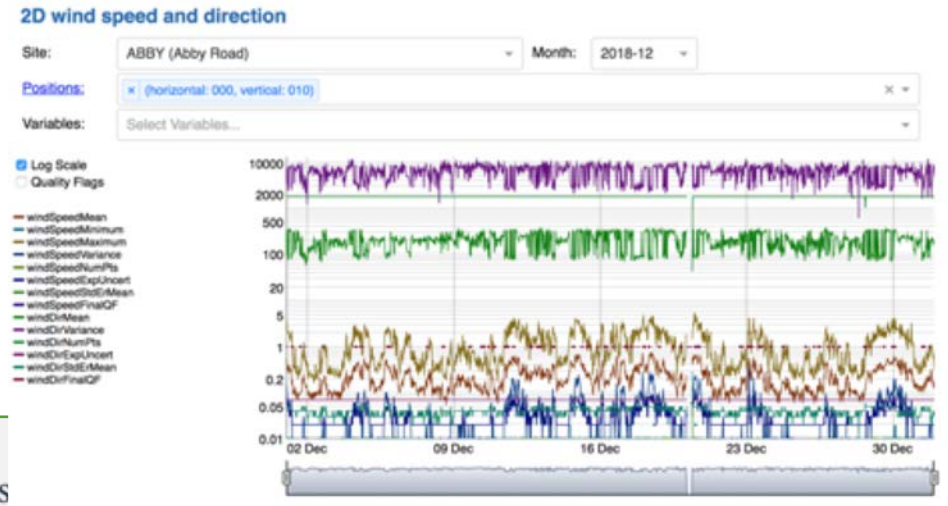
Team Members:

PILOT: Steve Petruzza, Valerio Pascucci, Giorgio Scorzelli, Attila Gyulassy, Timo Bremer, Charles Vardeman II, Robert Ricci

NEON: Christine Laney, Steve Jacobs Chris Clark, Jeremy Sampson, Steve Stone, Tom Gulbransen, Leslie Goldman, Ivan Lobo-padilla, Tristan Goulden, David Hulslander

NEON AOP data access

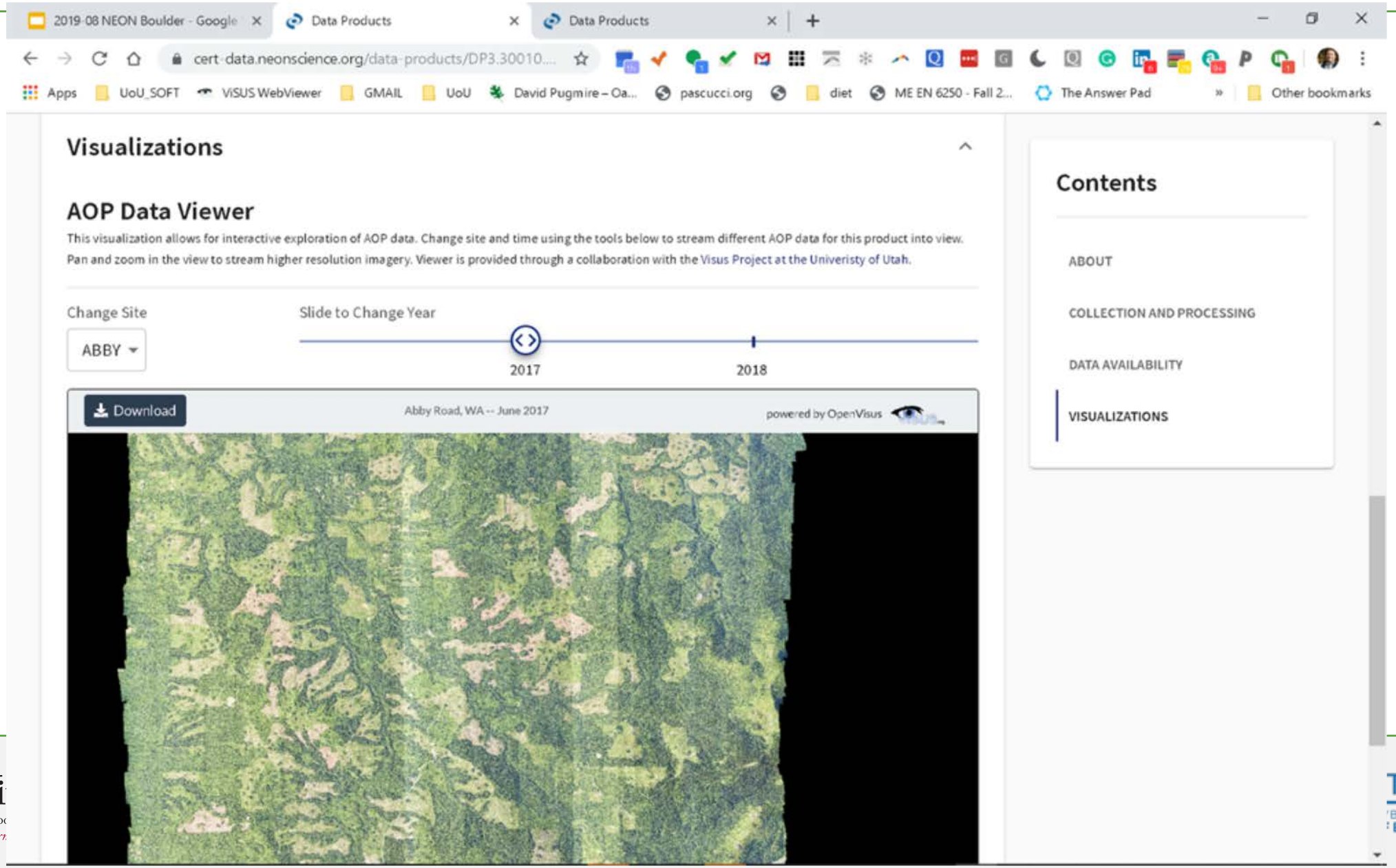
- NEON has a large amount of data that is shared with the community through their **data portal**
- There exist **APIs** to download those data in bulk
  (per site, per year, per data product, now also by area)
- For some data, such as sensor measurements, the portal provides an **interactive** navigation system
- For others, like **Airborne Observation Platforms data**, there is a long list of image files…
- There is a need to present all AOP data interactively, where the users can preview, navigate, and select/access/download the data they need



AOP data

Interoperability

- Explored solutions to integrate in the same visualization multiple “tile” sources
- Proof of concept of use AOP data and Google Earth
- New version of data format and server will allow to visualize AOP data in their geographical context


2019-08 NEON Boulder - Google
Data Products
Data Products
cert-data.neonscience.org/data-products/DP3.30010....
Apps
UoU_SOFT
ViSUS WebViewer
GMAIL
UoU
David Pugmire – Oa...
pascucci.org
diet
ME EN 6250 - Fall 2...
The Answer Pad
Other bookmarks
Visualizations
AOP Data Viewer
This visualization allows for interactive exploration of AOP data. Change site and time using the tools below to stream different AOP data for this product into view. Pan and zoom in the view to stream higher resolution imagery. Viewer is provided through a collaboration with the Visus Project at the Univeristy of Utah.
Change Site
ABBY
Slide to Change Year
2017
2018
Download
Abby Road, WA -- June 2017
powered by OpenVisus
Contents
ABOUT
COLLECTION AND PROCESSING
DATA AVAILABILITY
VISUALIZATIONS

Data Processing

Goals:

The data processing group focuses on workflows and services related to processing of data, for example transforming raw sensor data from sensors to more specific data products.

Team Members:

Pilot: Ryan Mitchell, Loïc Pottier, Mats Rynge, Karan Vahi

NEON: Steve Jacobs

- Engaged with Neon to understand their data processing pipelines and issues encountered.
- Early in the engagement, Neon decided to switch to a data-driven workflow system (Pachyderm) from earlier task-driven system (AirFlow).

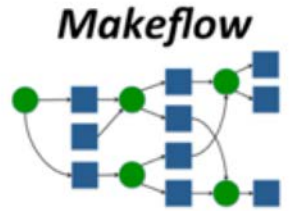
The group took a broader view of the problem and tried to answer

As a new user, regarding my workflow specification, which WMS should I choose?

Workflow Management System comparison study:

- Task-driven with Pegasus, Makeflow and Airflow
- Data-driven with Pachyderm

Test case: Weather Radar Analysis Workflow



Data Processing Lessons Learned

- No One Size Fits All
- New paradigms of computing gaining traction that are not normally associated with scientific data processing
  - Data Driven workflows with a strong need of reproducibility
- New Use-cases
  - big data analytics
  - large-scale science
  - machine learning
- Newer Technologies and usages gaining adoption
  - Cloud
  - Containers

Identity Management (IdM)

Goals:

The Identity Management working group focuses on understanding current practice in authentication and authorization, and helping to mature practice across the NSF Large Facilities. We do this by maintaining and sharing awareness of best practices as well as current policy and technology options for implementing those practices. Through direct engagements with specific facilities, participation in the NSF Cybersecurity Summit and other events, and publishing experience papers, case studies, and other artifacts, we hope to accelerate the exchange of lessons learned in Identity Management across the NSF Large Facility ecosystem.

Team Members:

Pilot: Susan Sons, Ryan Kiser, Terry Fleury*

NEON: Christine Laney, Jeremy Sampson, Steve Jacobs

Terry Fleury's time is funded via the CI CoE's partnership with TrustedCI, the NSF Cybersecurity Center of Excellence, ACI-1547272.

Identity Management (IdM) Activities

1. The IdM team worked with NEON to evaluate options for updating their data portal to more modern IdM practice.

2. We worked with NEON to help them manage the migration to their chosen solution (Auth0), both from a technical standpoint and in terms of managing user expectations and experience.

3. Captured process and lessons learned to benefit the wider community.

- Much of this took a consulting format, which worked to help keep the main activities/expertise-building solid within NEON. This is key to keeping NEON from becoming dependent on CI CoE Pilot.

Identity Management (IdM) Accomplishments and Work Products

1. NEON has a production-level Auth0-based authentication system for the data portal.

2. NEON and CI CoE Pilot are collaboratively authoring an experience paper to capture what we've done and what we've learned, as well as some open questions to help this become a community-wide learning activity.
3. Presentation at the NSF Cybersecurity Summit

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" (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 • **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

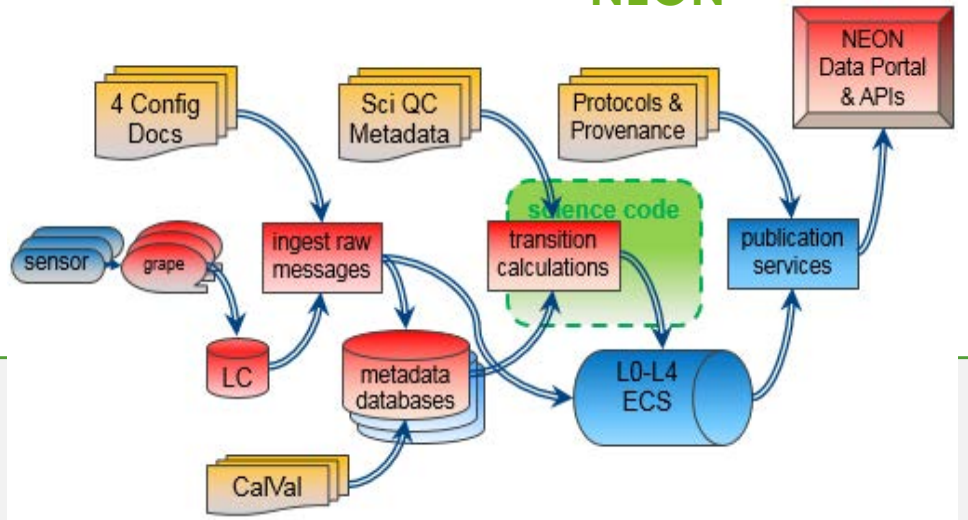
USC Viterbi School of Engineering Information Sciences Institute
University of Notre Dame
The University of Utah
renci
Trusted CI The NSF Cybersecurity Center of Excellence
Contact: Ewa Deelman, deelman@isi.edu

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
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. Discovering opportunities for CI sharing
6. Defining new working groups and discussion topics
   - Broadening the disaster recovery discussion
   - Data archiving and preservation
   - CI workforce enhancement, training

2019 LF CI Workshop Recommendations

- CI discovery and sharing of existing solutions, services, training resources, best practices
- Evaluate new technologies and provide training
- Maintain expertise in specialized areas (e.g., Internet of Things, workflows, data modeling, data archiving)
- Provide assistance in science-driven CI blueprinting
- Foster communication, collaboration, and community across LFs and CI projects
- Assist facilities in overcoming workforce challenges

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