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, University of Notre Dame

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

USC Viterbi School of Engineering
Information Sciences Institute

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

Develop a model and a plan for a Cyberinfrastructure Center of Excellence

Funded by NSF

Co-funded by OAC (Bill Miller) and BIO (Roland Roberts)
10/2018- 9/2020

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/

- 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.

Workshop Key Findings

- The need for, and benefits of, **close interactions, collaborations, and sharing** among the facilities and with the CI communities: sharing of CI related **expertise, technical solutions, best practices, and innovations** across NSF large facilities as well as DOE, NIH, NASA,
- There is a need for, and a current **lack of easily accessible information** about current **CI technologies, solutions, practices, and experiences**.
- There is a critical **lack of a focused entity that could facilitate interactions** and sharing across facilities. A model such as that used by the NSF-funded Center for Trustworthy
- **Workforce development, training, retention, career paths, and diversity** are major crosscutting challenges that the community shares. They may be best addressed coherently across all facilities through a coordinated approach.
- **Scientific Cyberinfrastructure (CTSC)** was explicitly and repeatedly noted as an effective model that should be explored to address this gap.

Workshop Key Recommendations

- **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**.
- 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.

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
Paul Ruth
Erik Scott

Resource Management, Networking, Clouds

USC Viterbi School of Engineering Information Sciences Institute
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
CENTER FOR APPLIED CYBERSECURITY RESEARCH
renci

Project Collaborators

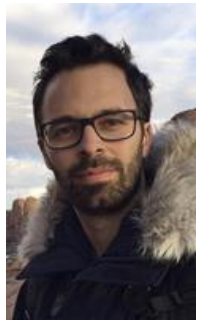
University of Notre Dame

Jarek Nabrzyski
Jane Wyngaard
Charles Vardeman

Workforce development, Sensors, Semantic technologies

University of Utah

Valerio Pascucci, Rob Ricci,
Timo Bremer, Attila Gyulassy,
Steve Petruzza



Data management, visualization, clouds, large-scale CI deployment

Indiana University

Susan Sons (co-funded by Trusted CI)
Von Welch (unfunded collaborator)
Mary Conley

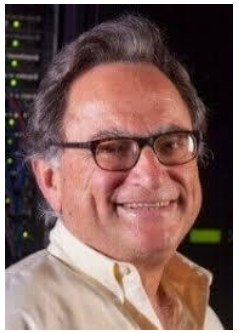
Cybersecurity

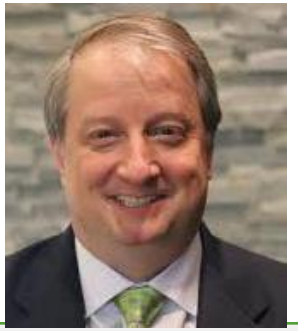
USC Viterbi School of Engineering Information Sciences Institute
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
CENTER FOR APPLIED CYBERSECURITY RESEARCH
renci

CI CoE PILOT

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**, Indiana University
- **Michael Zentner**, Purdue University





USC Viterbi School of Engineering Information Sciences Institute

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

Project Goals

Award #1842042

Develop a model and a plan for a Cyberinfrastructure Center of Excellence

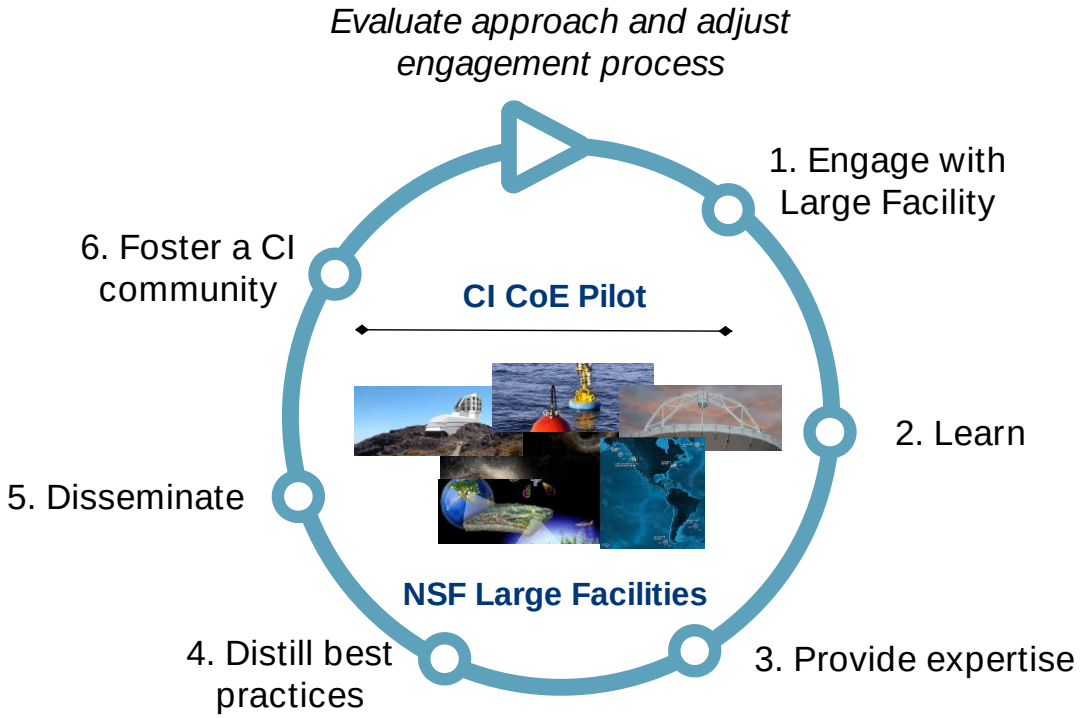
- Platform for knowledge sharing and community building
- Key partner for the establishment and improvement of Large Facilities with advanced CI architecture designs
- Grounded in re-use of dependable CI tools and solutions
- Forum for discussions about CI sustainability and workforce development and 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

Engagement Methodology

Developing and improving Engagement Processes



Engagement with a Facility

- Engage at the management level, potentially seek introductions from NSF PO, participate in meeting (LF 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

NEON Engagement

- 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, Boulder Nov 2018
    - Understanding and prototyping sensor data pipelines
    - Exploring user-facing data presentation options
    - Learning about data processing tools (workflow systems)
    - Collecting information about data lifecycle and disaster recovery approaches

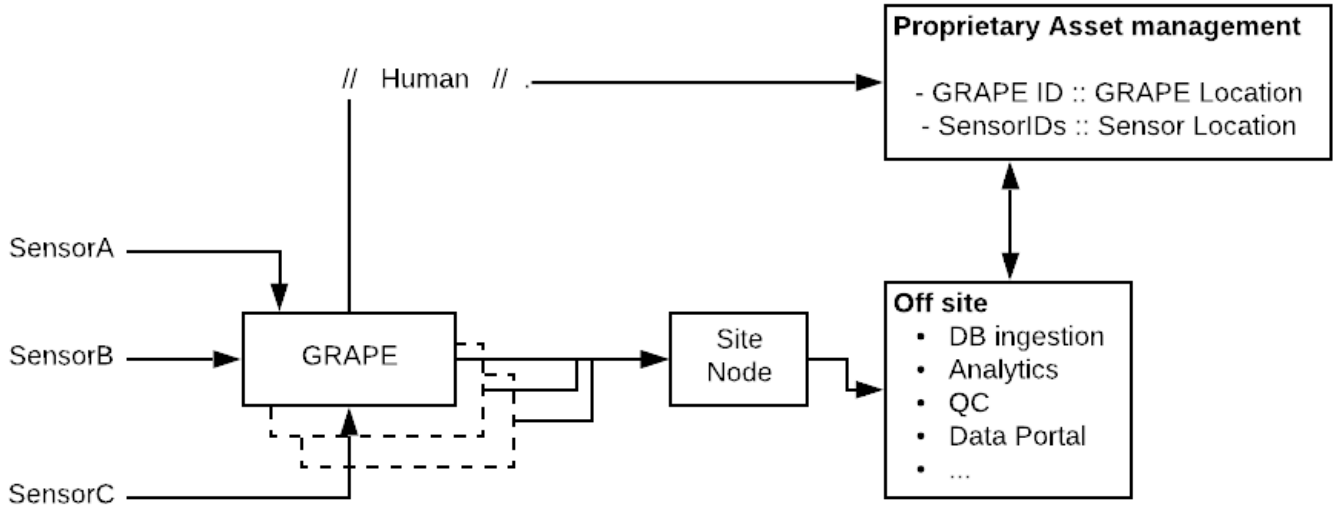
Working Groups

- Data capture
- Data processing
- Data storage/curation/preservation
- Data access/visualization/dissemination
- Disaster recovery
- Identity management
- Engagement with Large Facilities

NEON sensor upgrade goals:

- Focus: "Instrumented systems"
- Primary goal: Move to a COTS GRAPE
- Considerations:
  - Partial migration
  - Cost (HW, SW, and process)
  - Long term sustainability

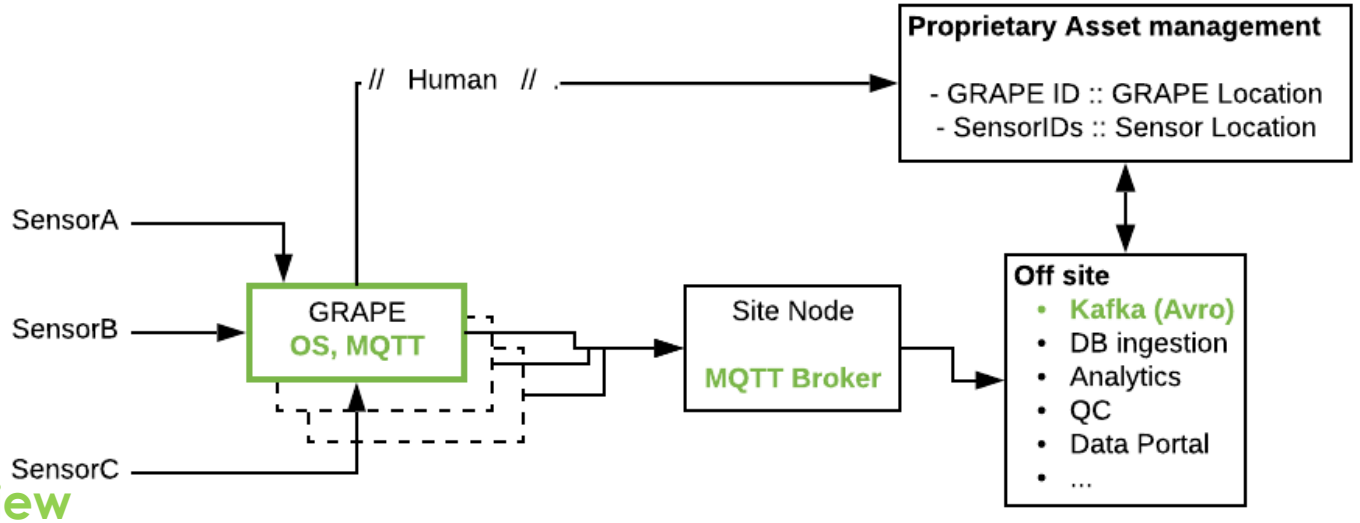
Jane Wyngaard, lead



NEON sensor upgrade goals:

- Focus: "Instrumented systems"
- Primary goal: Move to a OTS GRAPE
- Considerations:
  - Partial migration
  - Cost (HW, SW, and process)
  - Long term sustainability

Jane Wyngaard lead

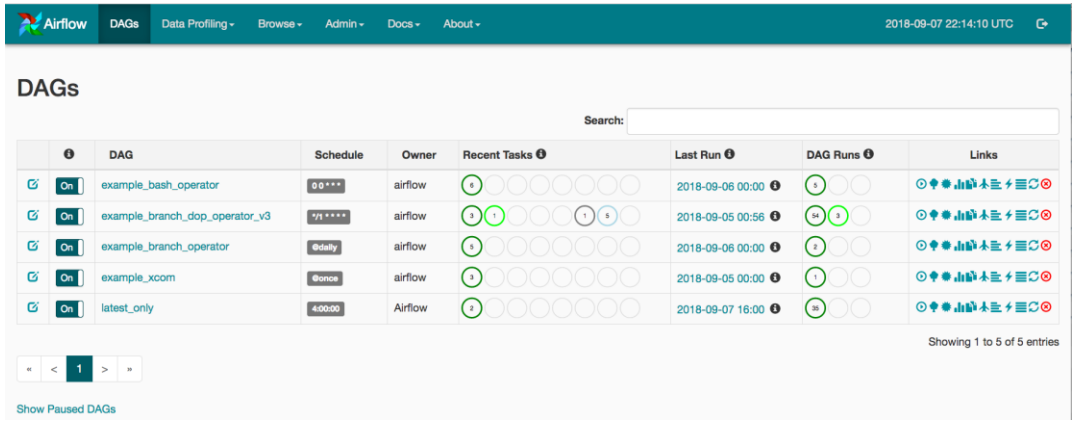


- GRAPES requirements review
- GRAPES to CI architecture requirements review
- Potential pipeline prototype

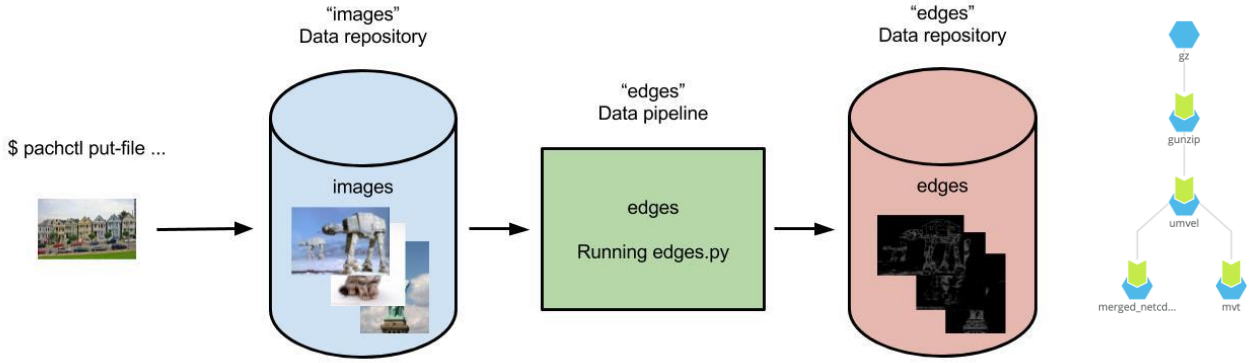
Data semantics discussion

Data Processing

- NEON is using Airflow for datacenter data processing, exploring Pachyderm
- CI CoE Pilot is comparing the ability to model complex data analysis using various workflow systems

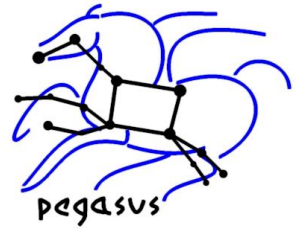


Airflow is time-driven, provides extensive UI, ability to manage scheduled data processing



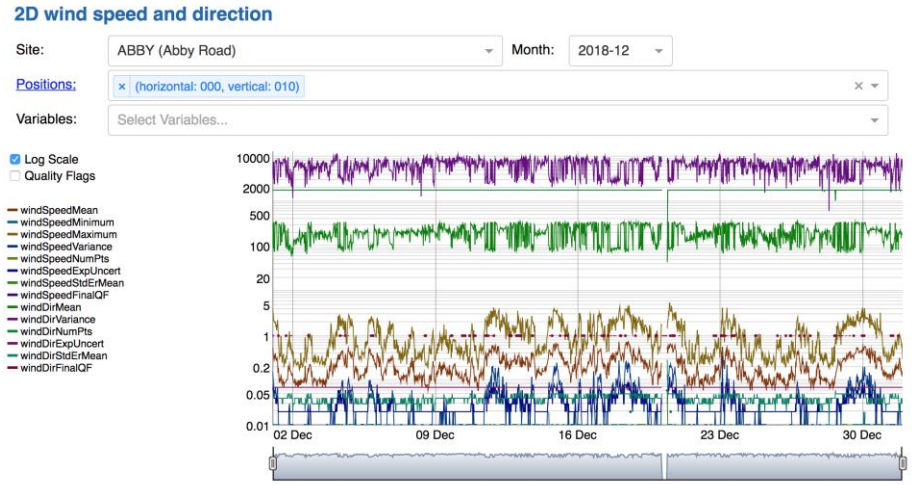
Pachyderm is data-driven, operates on data repository triggers, supports containers

USC's Pegasus is task-driven, focuses on portability across CI, scalability, robust execution.

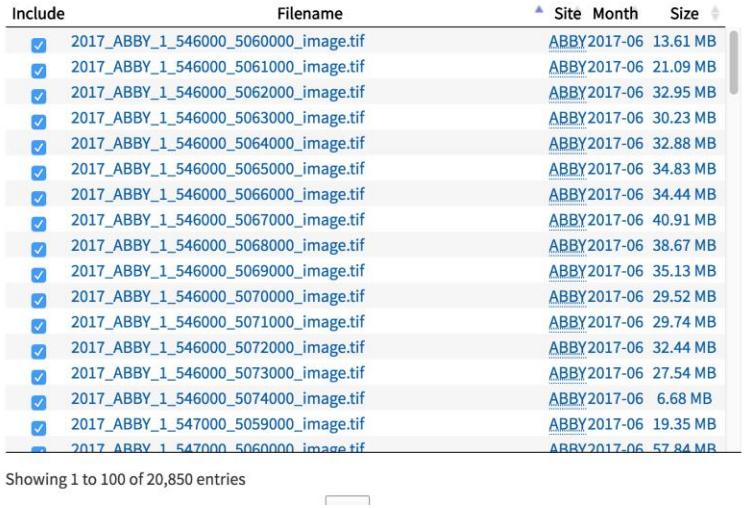


Data Visualization and Presentation

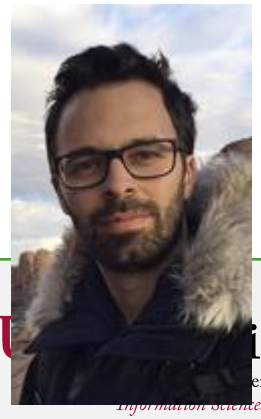
- NEON has a large amount of data that is shared with the community through their **data portal**
- There exist **APIs** to download those data in bulk (per site, per year, per data product)
- For some data, such as sensor measurements, the portal provide an **interactive** navigation system
- For others, like **Airborne Observation Platforms data**, there is only 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



Atmospheric data



AOP data



Steve Petruzza, lead

Information Sciences Institute

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

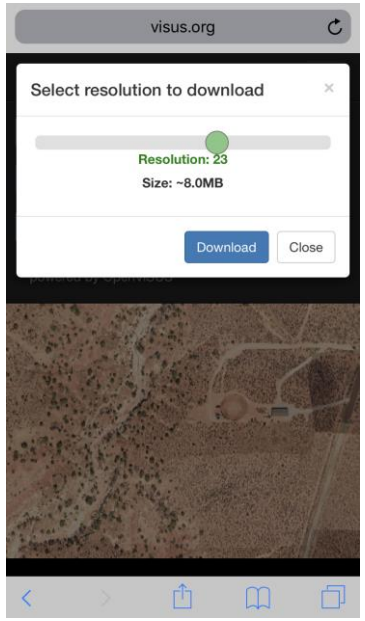
CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

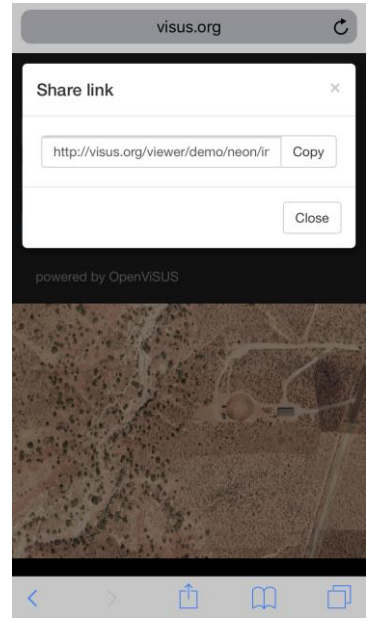
Data Presentation – New interactive data access

- Prototype an interactive web portal for image data visualization/exploration
- We performed some manual conversions (e.g., using ad-hoc scripts) of some of the NEON image data (e.g., 65 GB from the entire MOAB site) to a hierarchical multiresolution data format
- For each test site, **thousands of images** have been stitched into a single large multiresolution image
- Setup a streaming server on a small machine, to allow the data streaming of different resolutions of the data
- Responsive web portal that allows
  - Interactively explore entire site image data
  - Download a subsampled resolution of the site data
  - Share the current using an auto-generated link
  - Available at: https://visus.org/viewer/demo/neon

Interactive exploration

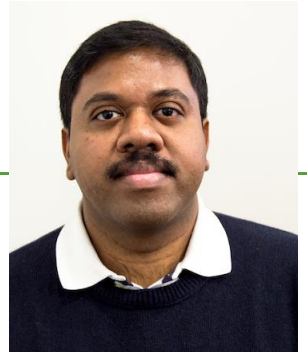


Multiresolution download

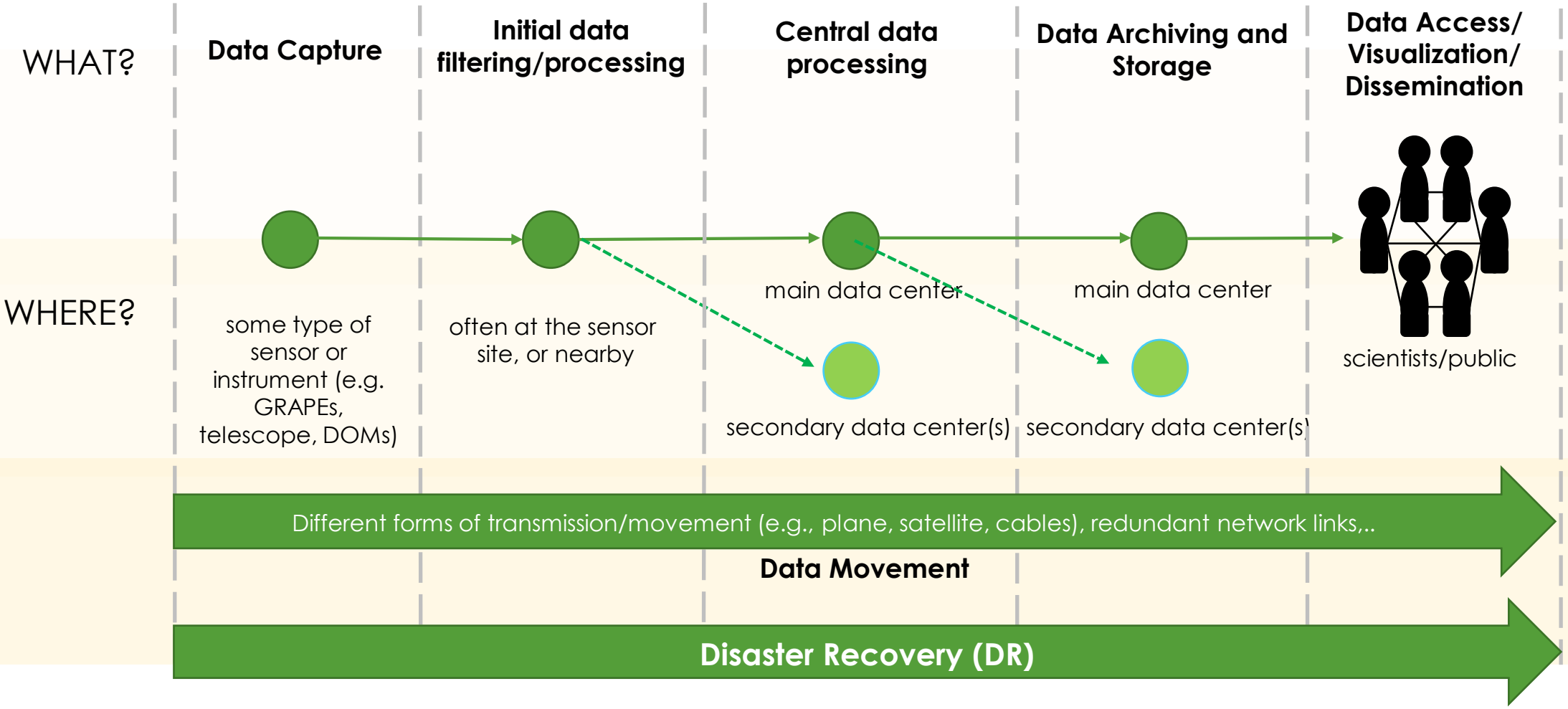


View sharing

Data Lifecycle for LFs



Anirban Mandal, lead



What services correspond to the data lifecycle stages?

Disaster Recovery Effective Processes

- **Cross-cutting finding:** Although some DR strategies exist across some stages of the data life cycle for some LFs, DR hasn't been taken into account to the fullest extent it warrants when designing the CI architecture for LFs.
- There is a need for some careful consideration of **requirement analysis and planning for DR** as an effective process to be followed **before and after** a possible disaster.
- Developing a working draft for a possible **DR Planning Phase template** that Large Facilities can follow for planning for Disaster Recovery
  - Based on the NIST guidance for developing an *Information System Contingency Plan (ISCP)* and/or *Disaster Recovery Plan (DRP)* by doing a thorough *Business Impact Assessment (BIA)* – **NIST 800-34r1** (https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-34r1.pdf)

NEON Engagement

Lessons learned so far:

- Importance of f2f discussions
- The need to formalize the engagement: expectation and timelines
- Importance of LF priorities and challenges
- Identified how to organize ourselves (working groups and work products)

- Complexity of the end-to-end data lifecycle
- Need to formalize the questions to ask the facilities (templates)
- Understanding of the CI decision making processes: for example the level of support of a software
- Learned about new CI solutions (distributed query processing, workflow management)
- Co-existence of old and new systems, making for a heterogeneous CI landscape

CI CoE PILOT

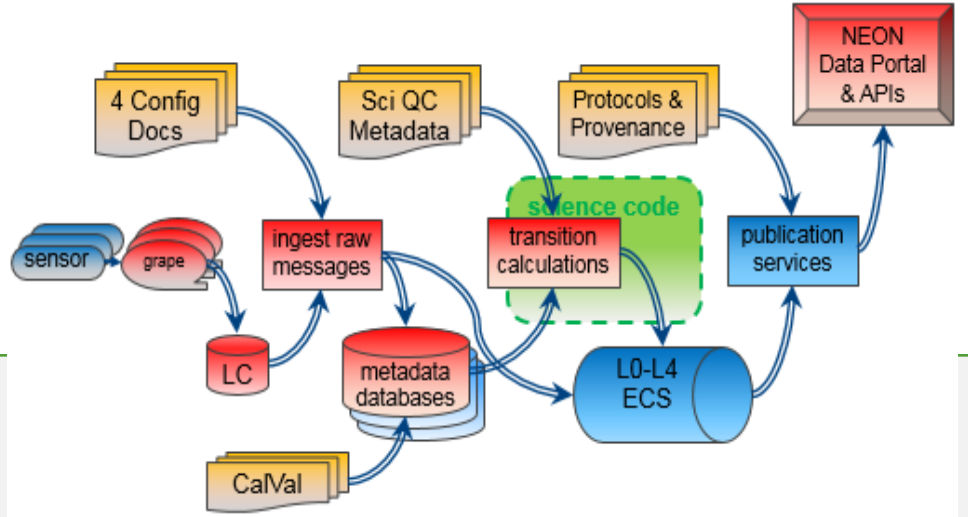
Tom Gulbransen

NEON's perspective

Project Manager for Cyber Infrastructure and Data Products Development

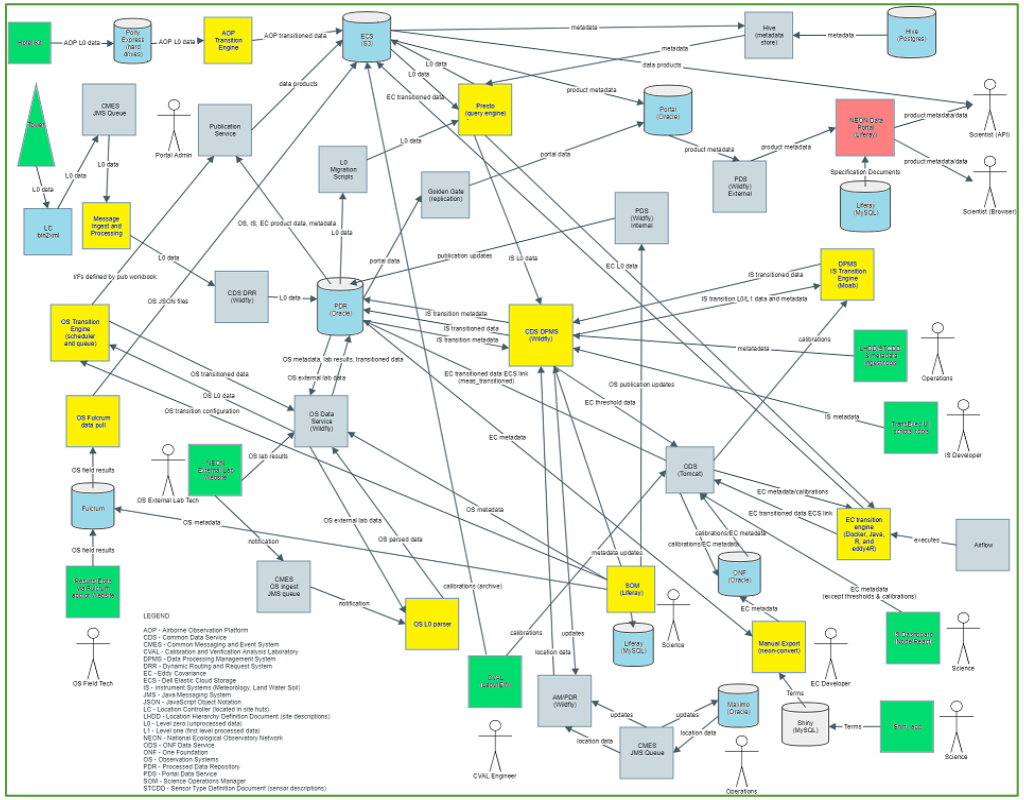
CoE Pilot Benefits to NEON Thus Far

- 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



Possible CoE Scope Amendments

- Methods for CI performance self-assessments
- Advice on CI documentation
- Consultation with CI development investors
- Inter-facility collaboration
- Workforce development?



Initial Engagement Beyond NEON

- Facilitating identity management discussions, will have a document describing the options and which facilities are using which solutions
- Initial engagement
  - IceCube-- learning about data management
  - SCIMMA (IceCube, LIGO, LSST)– contributing to the discussions around a Software Institute for multi messenger astrophysics
- Initial discussions with LSST re CI project management and data management

- **We are developing an engagement template**

NSF Cross-center Collaboration

TRUSTED CI
THE NSF CYBERSECURITY CENTER OF EXCELLENCE

trustedci.org

- Operational services and related training for NSF CI
- Community of Practice and Threat Intelligence Network
- Enabling Cybersecurity Research
- Outreach to Higher Ed Infosec regarding research CI

CI CoE PILOT

ResearchSOC

researchsoc.iu.edu

- Creating comprehensive cybersecurity programs
- Community building and leadership
- Training and best practices
- Tackling specific challenges of cybersecurity, software assurance, privacy, etc.

Jim Marsteller (Trusted CI) and Susan Sons (ResearchSOC) will be at Friday workshop.

Identity Management

- Important issue to a number of large facilities
- Facilities have different requirements (open data, embargoed data, access to data processing)
- Management of different groups (SCIMMA)
- There are a number of solutions with various costs (monetary and deployment efforts)
- Facilitating discussions across facilities
  - Presentation by **Jim Basney**
- Developing a document that outlines solutions

Future Activities

- Publish work products from the engagement with NEON
- Identify new facilities to engage
- Explore new avenues of engagement areas
- Explore issues of workforce development
- Design a blue print for a Cyberinfrastructure Center of Excellence

http://cicoe-pilot.org/