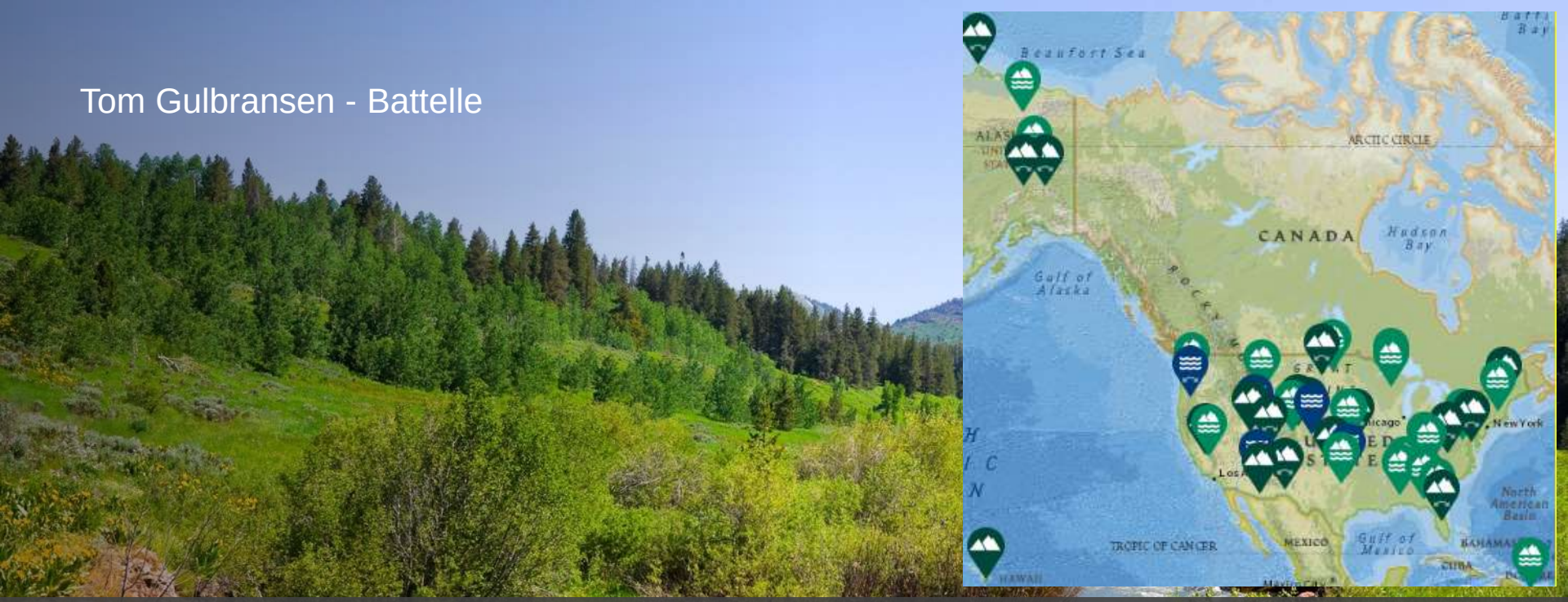
Tom Gulbransen - Battelle



NEON Data Delivery & Cyberinfrastructure

(mission, data products, data delivery, processing, architecture, interoperability, improvement plans)

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Why is NEON important?

NEON provides a coordinated national system for monitoring a number of critical ecological and environmental properties at multiple spatial and temporal scales.

…transformative science …workforce development








This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON's 176 Data Products Overlap Multiple Themes

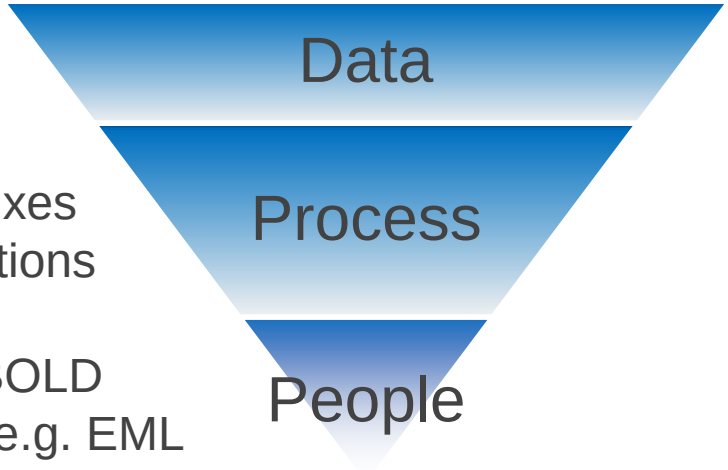
Atmospheric 58	Organismal 51	Ecohydrology 47	Biogeochemistry 85	Land Related 47
				
+11 from AmeriFlux	+1 from PhenoCam +3 from MG-RAST	+2 from AeroNet		
H_2O, CO_2 Heat Isotopes Turbulence Storage Fluxes	Abundance Composition Pathogens Phenology DNA Barcodes Marker Genes Metagenomics	Water quality Precipitation Levels Discharge Radiation Geomorphology	Soil conditions Chemistry Particulates Foliar characteristics	Spectrometry Hi-Res imagery LiDAR

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON CI & Data Science Offerings

National
Ecological
Observatory
Network

Cyberinfrastructure = resources, tools & services, observation to delivery



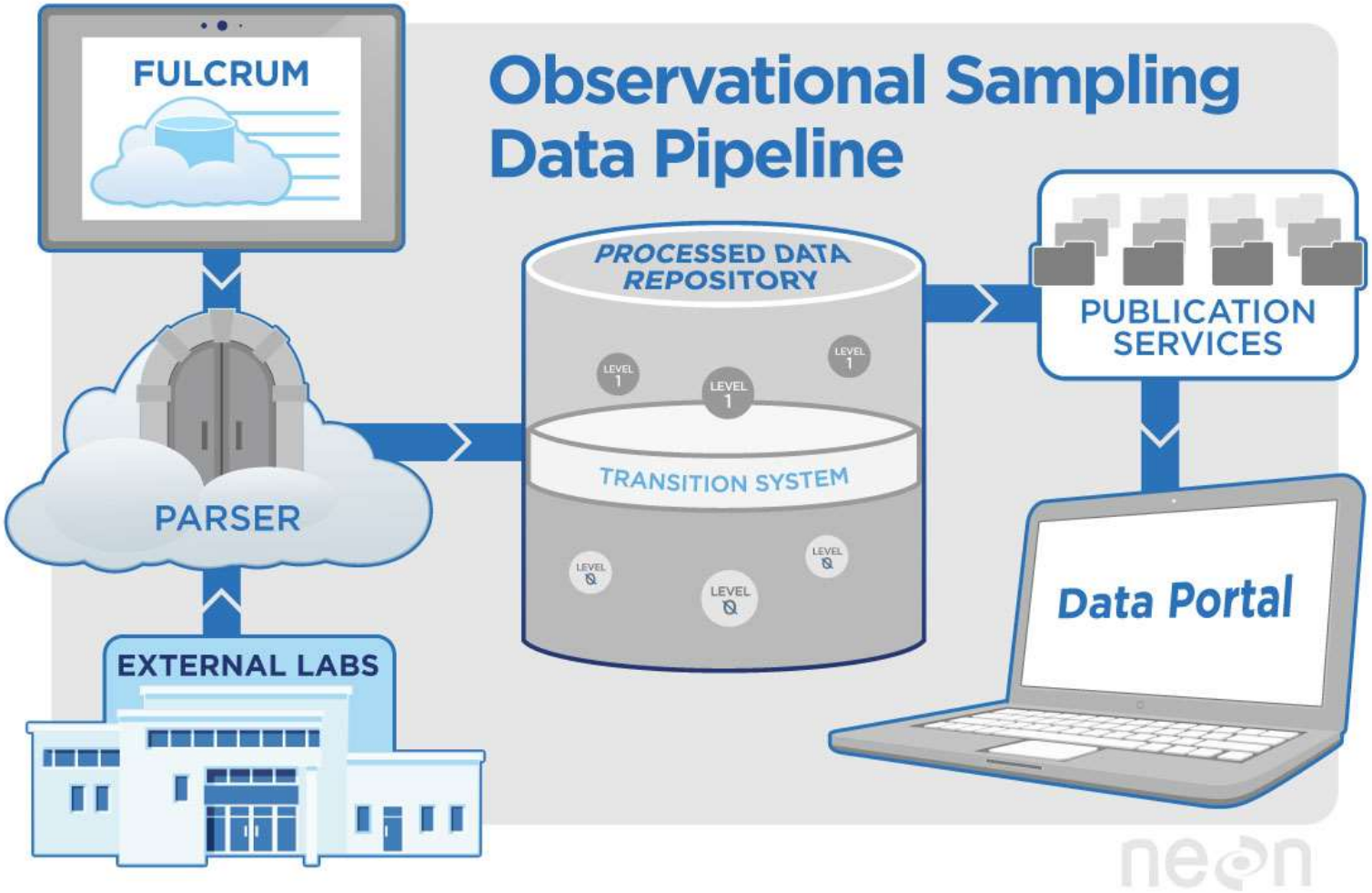
- of **field observational results**
- of **sampling protocols**
- of **algorithms**
  - surface atmospheric exchange/fluxes
  - meteorological temporal interpolations
- of **data science methods**
  - APIs @ PhenoCam, MG-RAST, BOLD
  - Ecological semantic conventions e.g. EML
  - Dictionary-driven generic observational pipeline
  - Data integration via workflow models
- of **practitioners**
  - ***21 Working Groups; ~1,500 unique users & 75k API hits/month***
  - Carpentry workshops & interns
  - Digital identifiers for data use/provenance
  - Large data facilities coalitions, e.g. RDA, CDF, ESIP

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON CI Data Delivery

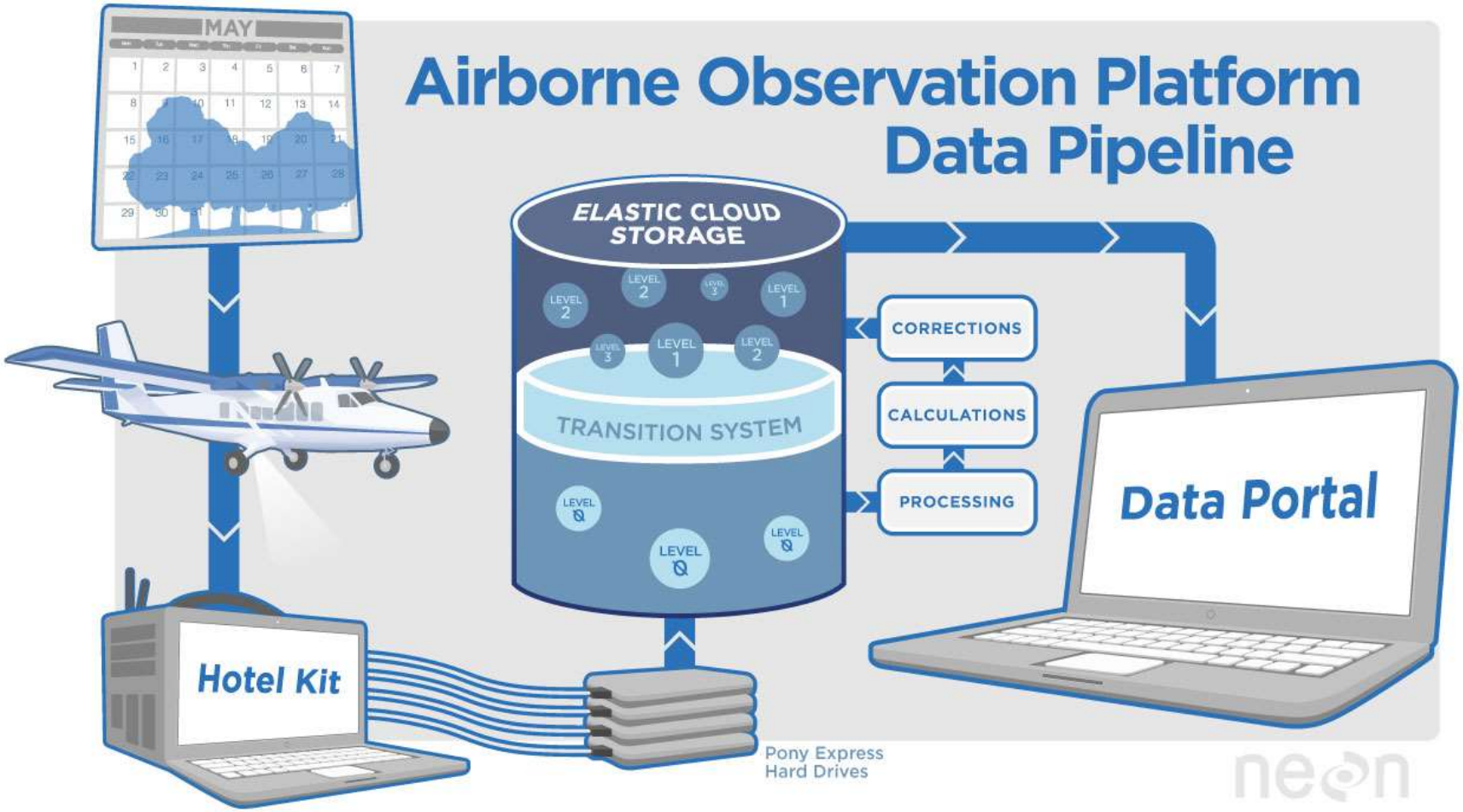
Data Product Type	General Latency from Sampling to Portal	Primary Influence on General Latency	Existing Backlog	Estimated Complete Publication of Backlog
AOP	2 months	Staff during flights	2018 ongoing	Nov2018
AOP legacy		Reprocessing with new algorithms	2013-2016	Dec2018
Land, Water, Soil, Meteorology	15th of next month	5 day communications completion buffer	2018 ongoing	Monthly
Land, Water, Soil, Meteorology legacy		Volume of prior site-months	2013-2017	Nov2018
Eddy Covariance	15th of next month	5 day communications completion buffer	2018 ongoing	Monthly
Eddy Covariance legacy		Volume of prior site-months & configurations	2013-2017	Nov2018
OS – field observations	1-4 months (~9-13 for fish & veg structure)	Reviews as season progresses. HQ staff.		Oct2018
OS – domain lab results	1-3 months (6 & 13 for morphospp. & bathy)	Reviews as season progresses. HQ staff.		Sep2018
OS – external lab rolling batches	1-9 months	Reviews as season progresses. Lab deliveries.		Dec2018
OS – one time samples (megapits, soil)		external lab processing - MBL		Dec2018

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.



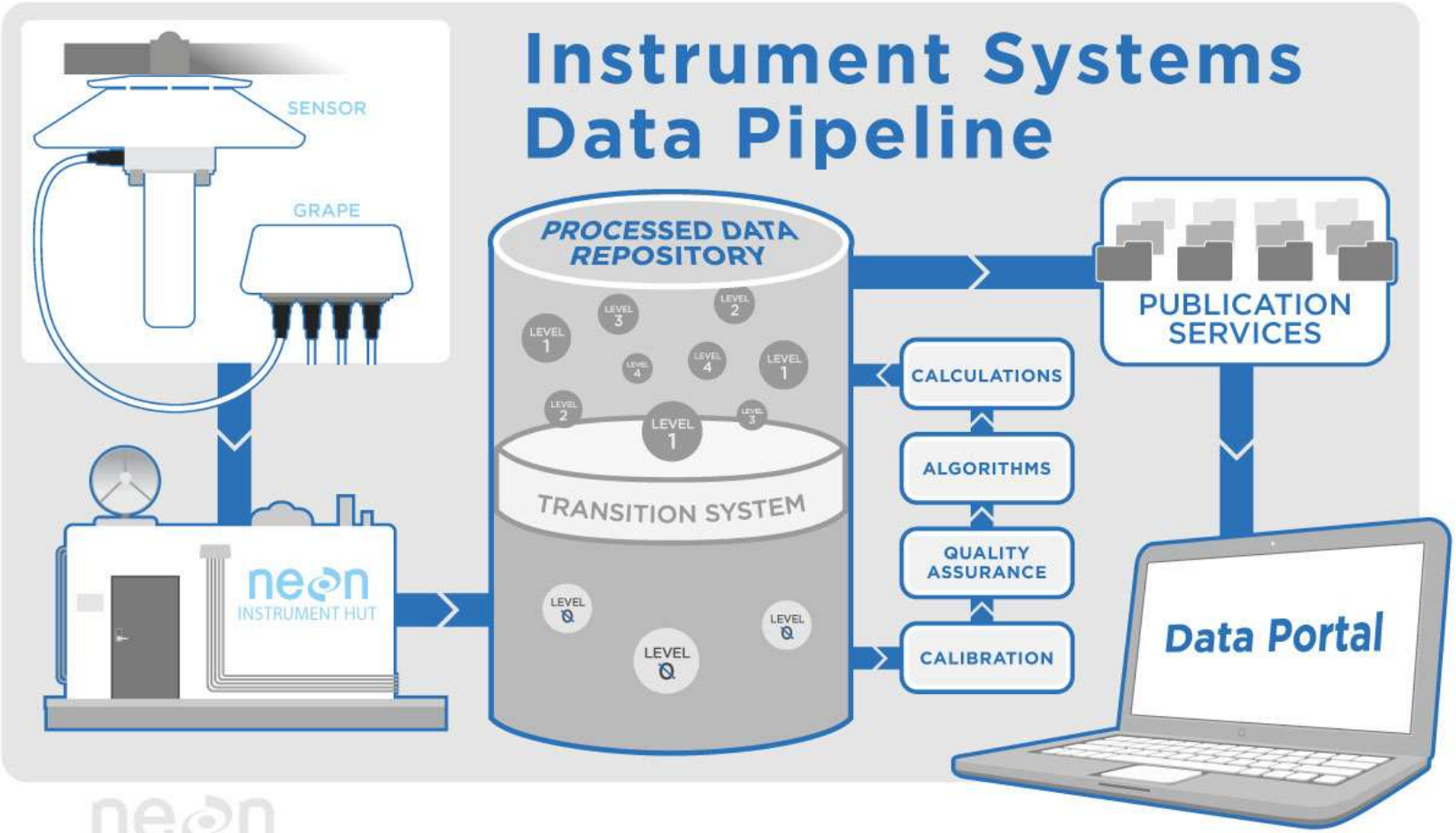
- 47 terrestrial & 34 aquatic sites, 100s protocols, dozens of labs
- Latencies from 5 to 365+ days

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.



- 19 domains, 100-300km^2 flown annually at peak greenness
- Images, Spectroscopy (380 to 2500 nm), LiDAR
- 1-500GB/data product, ~2PB/yr

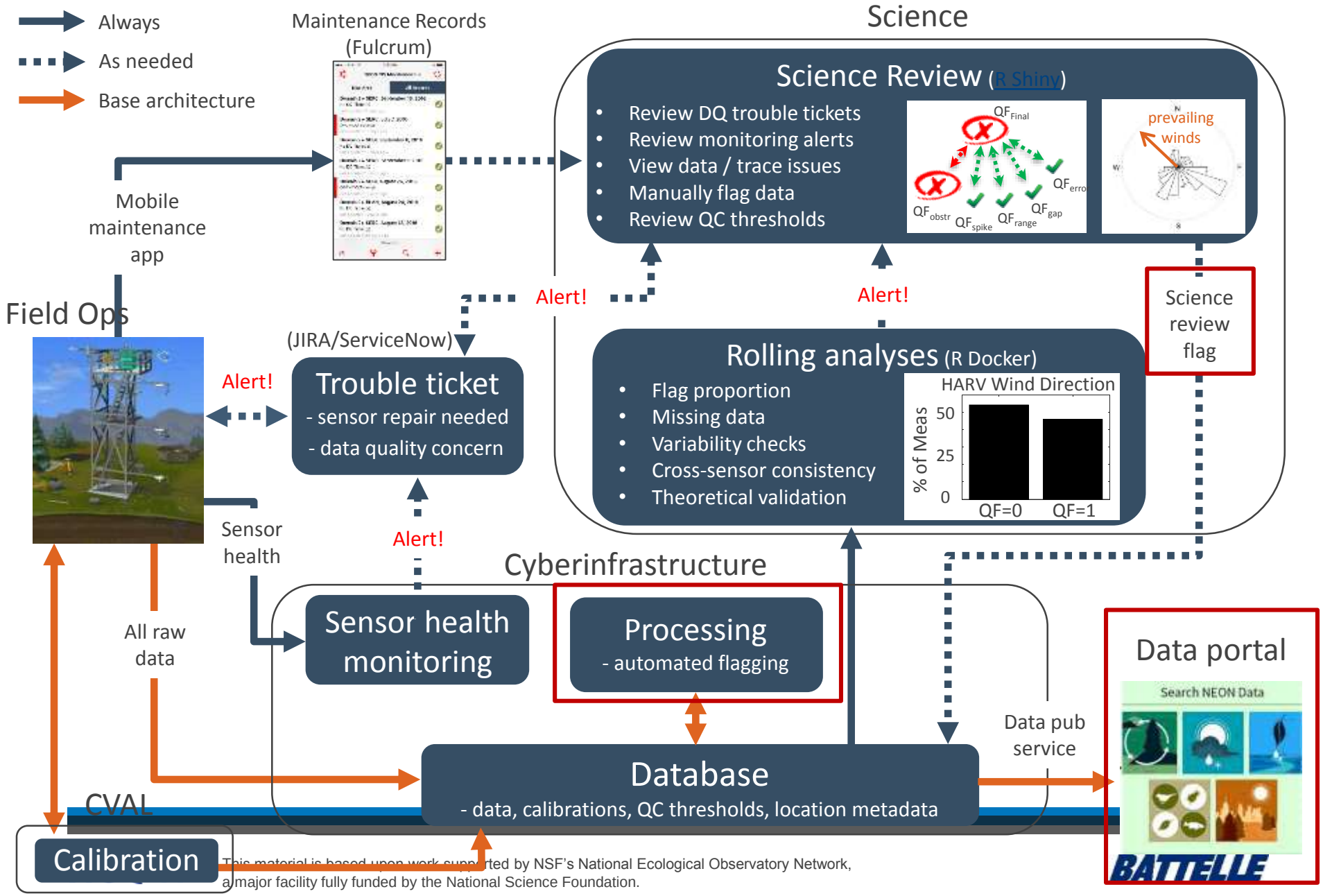
This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.



- 81 sites, 100s sensors per site, up to 40Hz
- Sensor→GRAPE→LC→HQ 5day lag, ~5TB/month

This material is based upon work supported by NSF's National Ecological Observatory Network,
a major facility fully funded by the National Science Foundation.

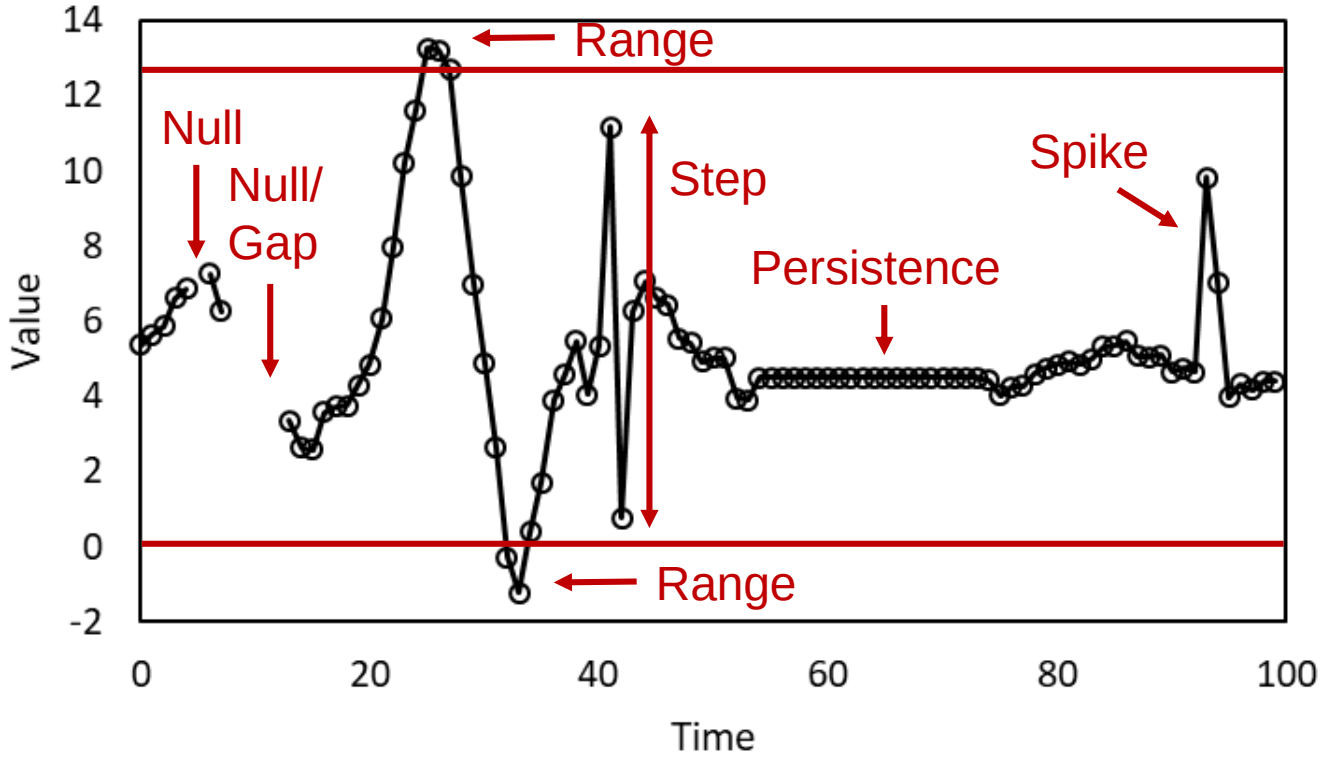
Instrumented Systems Data QA/QC Framework



This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

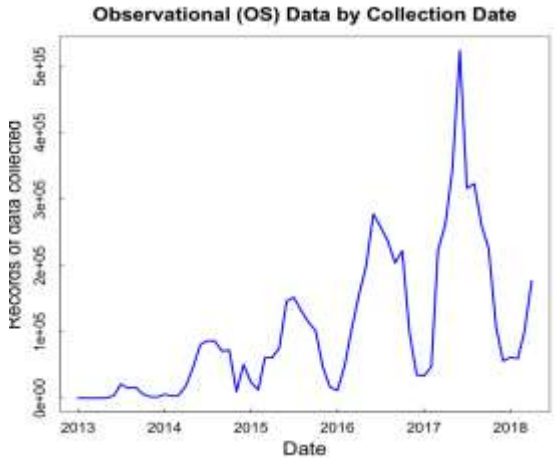
Automated flagging

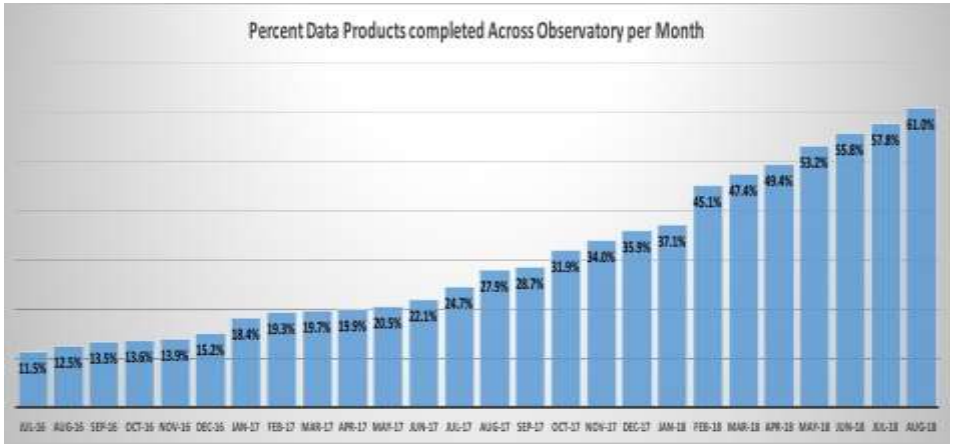
- Applied on each calibrated data value (native resolution)
- Basic tests: Null, Gap, Range, Step, Spike, Persistence
- Sensor-specific tests: e.g. sensor diagnostic
- If chosen, can remove any data point that fails a test



What's in a data package?

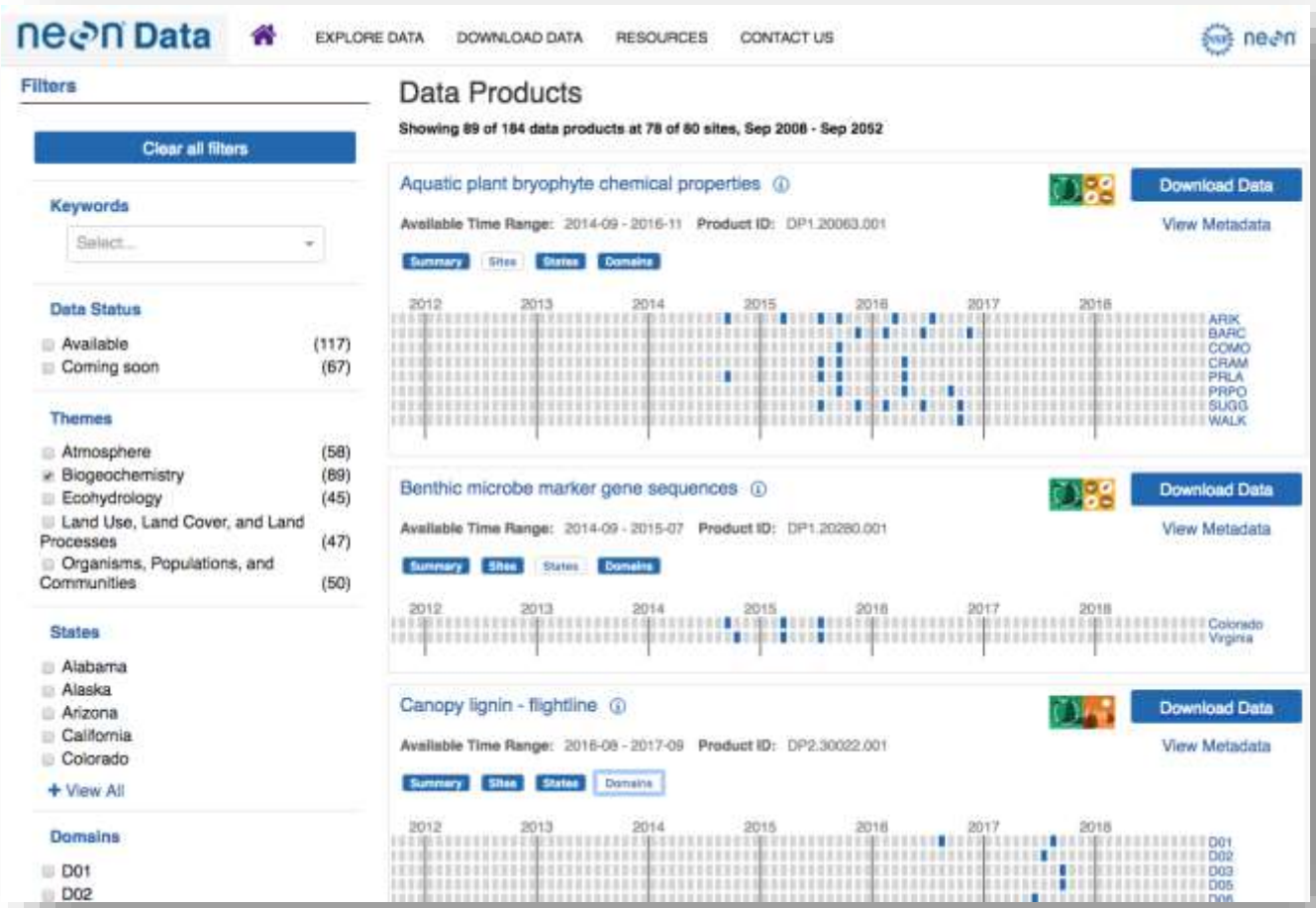
- Monthly or Annual data files with all basic data and quality flags
- CSV (IS, OS), HDF5 (EC, AOP), TIF (AOP), LAS (AOP), FASTA (metagenomics)
- Readme text file – general info about the data product
- Additional quality metrics, external lab data
- User guides & protocols; algorithm & sensor configuration documents
- Variable definitions, validation rules, and sensor positions
- Machine-readable metadata (EML)





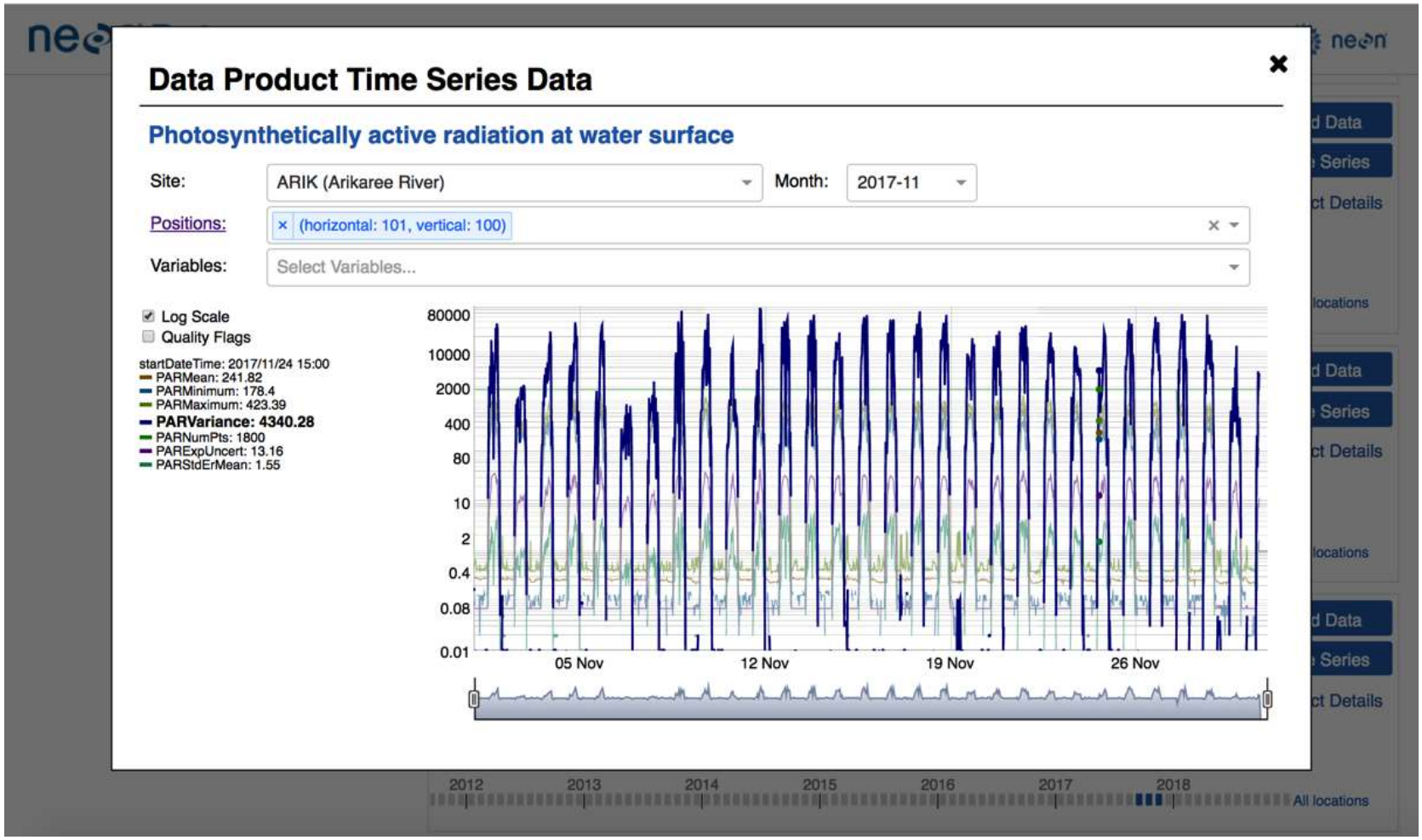
This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Interactive Browse

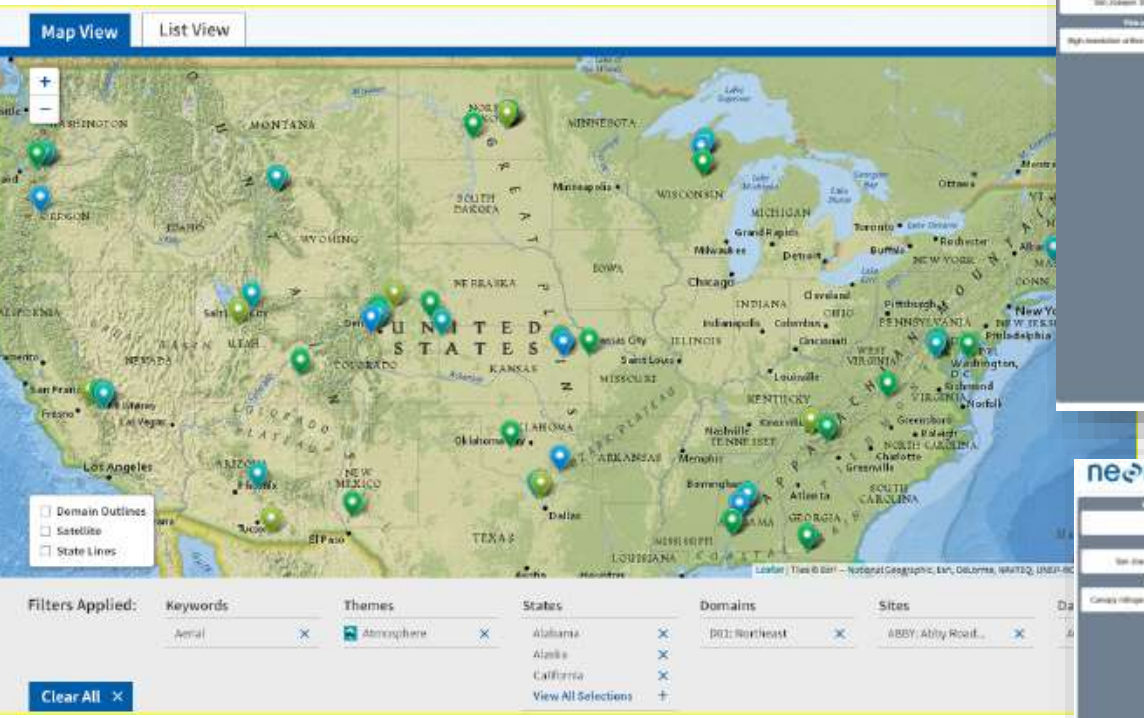


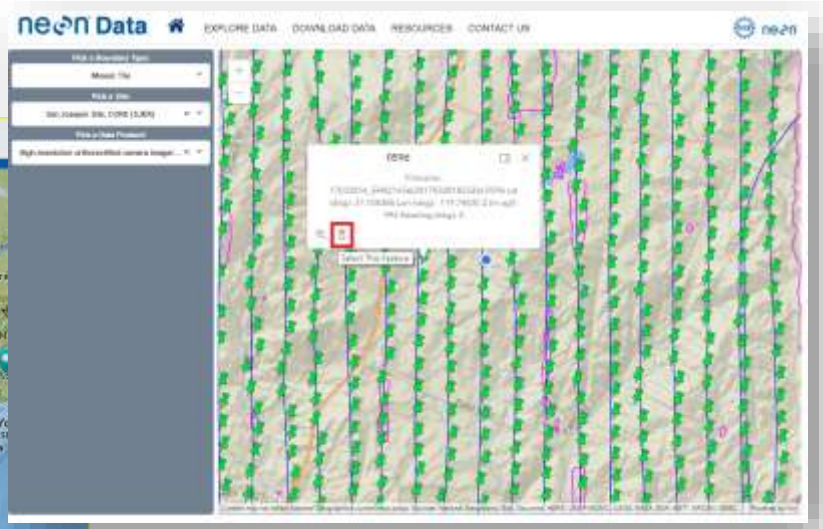
This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Interactive Time Series



Spatial Viewer (prototypes)





This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Partner Repositories

- Work with specialized repositories for domain-specific data
- Use their APIs to keep NEON caches in sync
- Field data on portal; specialized data hosted

Current

MG-RAST metagenomics analysis server

AERONET AEROSOL ROBOTIC NETWORK

BOLDSYSTEMS

PhenoCam Gallery

DataONE Data Observation Network for Earth

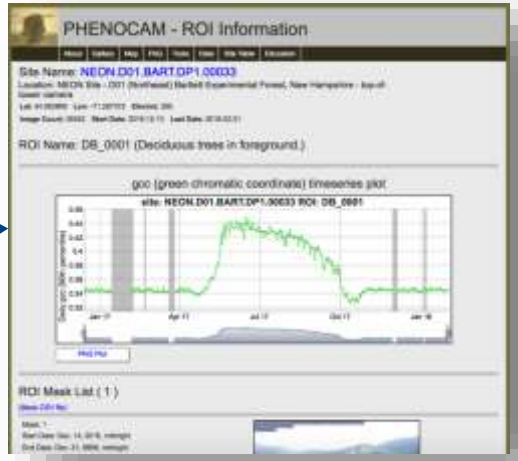
Future

AMERIFLUX

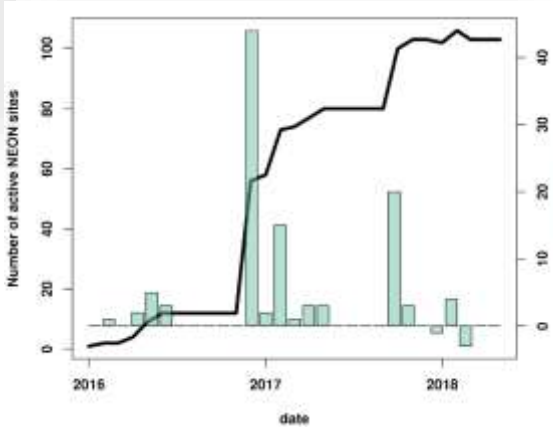
OpenTopography High-Resolution Topography Data and Tools

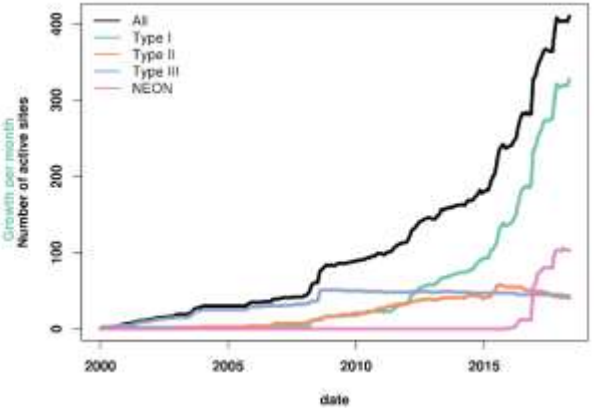
This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Data Portal & Phenocam Gallery



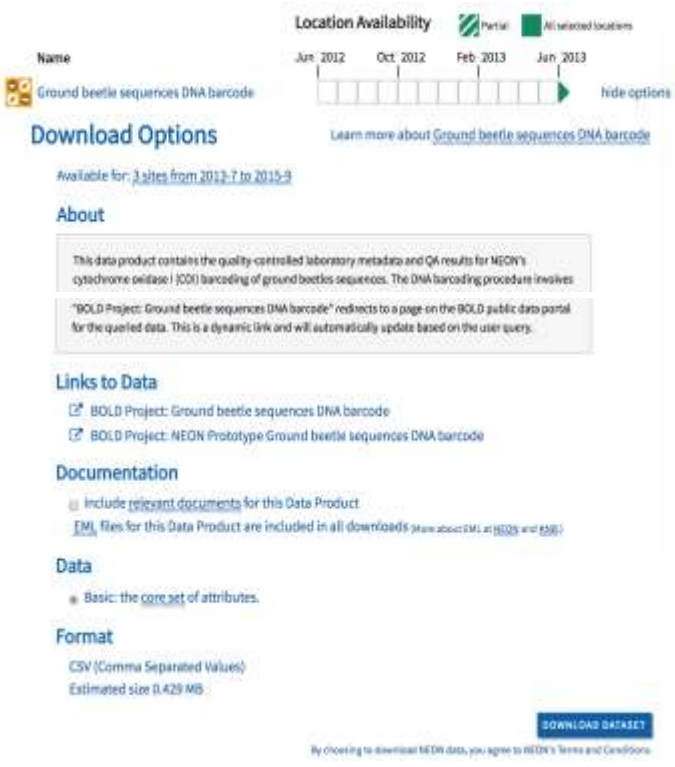
Collaborative work between 2 NSF programs – co-located sites and API modeled on NEON's.

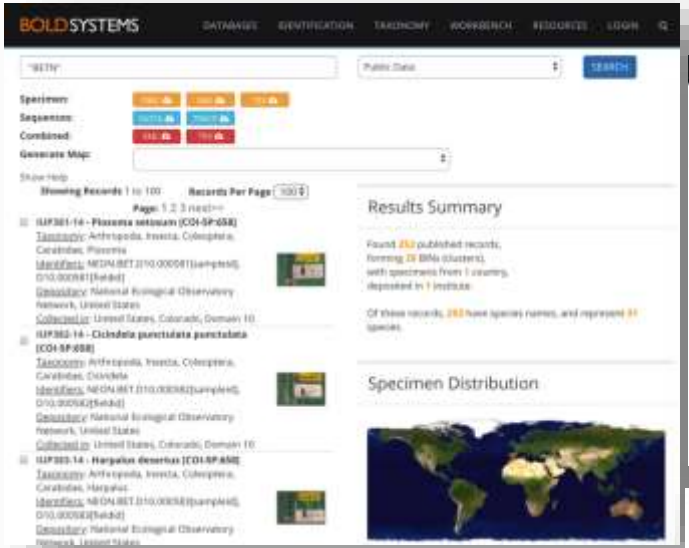




This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Data Portal & BOLD





MG-RAST is similar in functionality

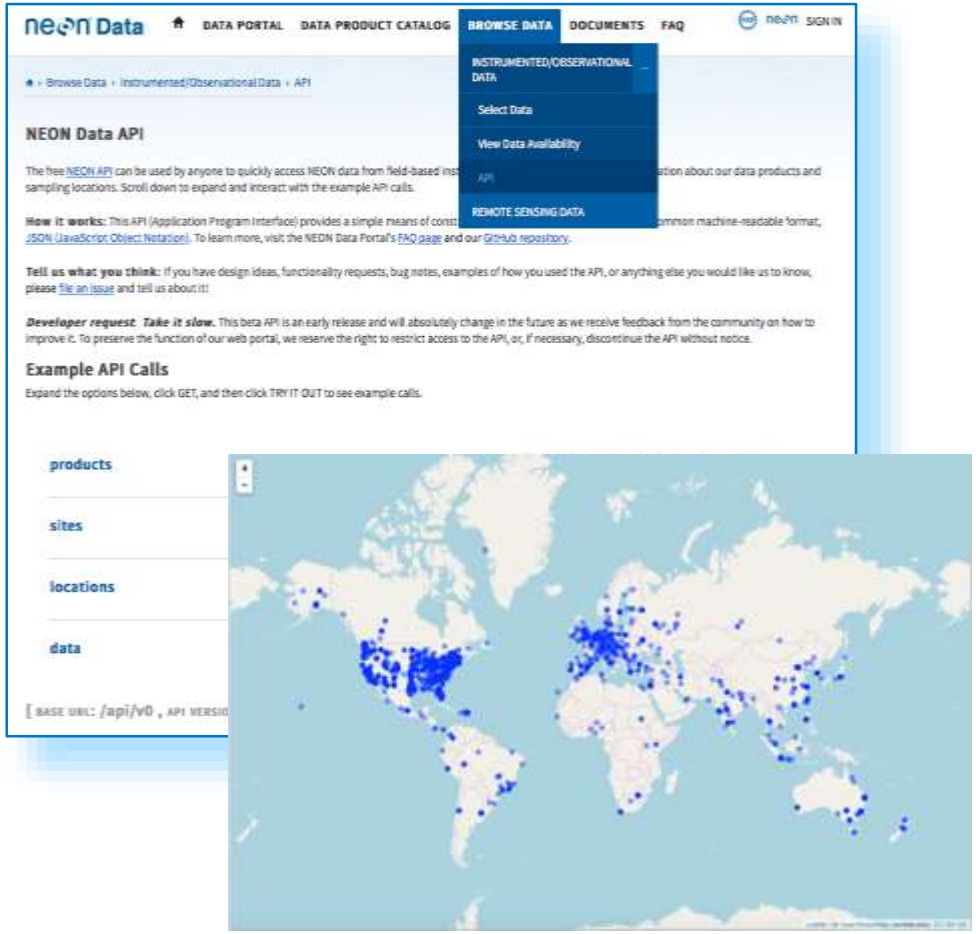
Access to thousands of beetle, mosquito, small mammal, and fish records, from current protocols as well as prototype protocols and sites.

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

REST API: Programmatic Access

http://data.neonscience.org/data-api

- Open access
- Provides product information, site and within-site information, documents, and data
- Returns information in JSON format
- Returns both zipped data packages and individual data files
- Recently added endpoints for access to taxonomic lists and sample custody histories



NEON CI Software Architecture - Layers

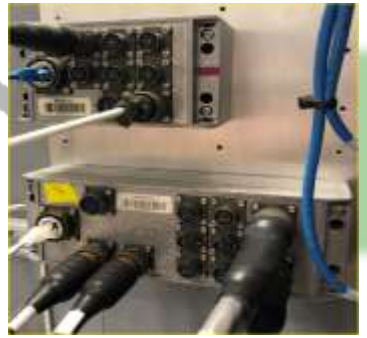
Publish

Process

Store

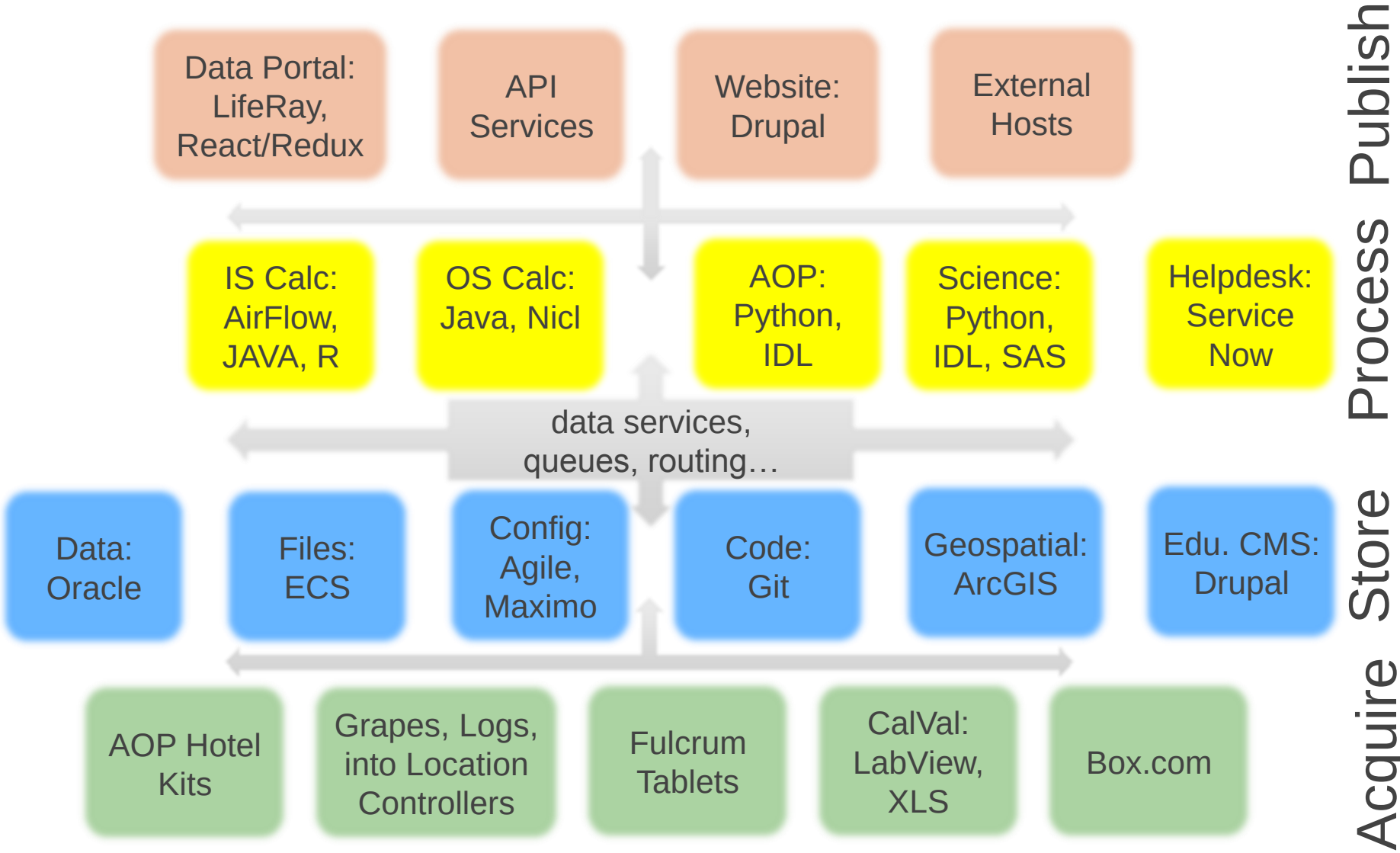
Acquire

data services, transport, queuing



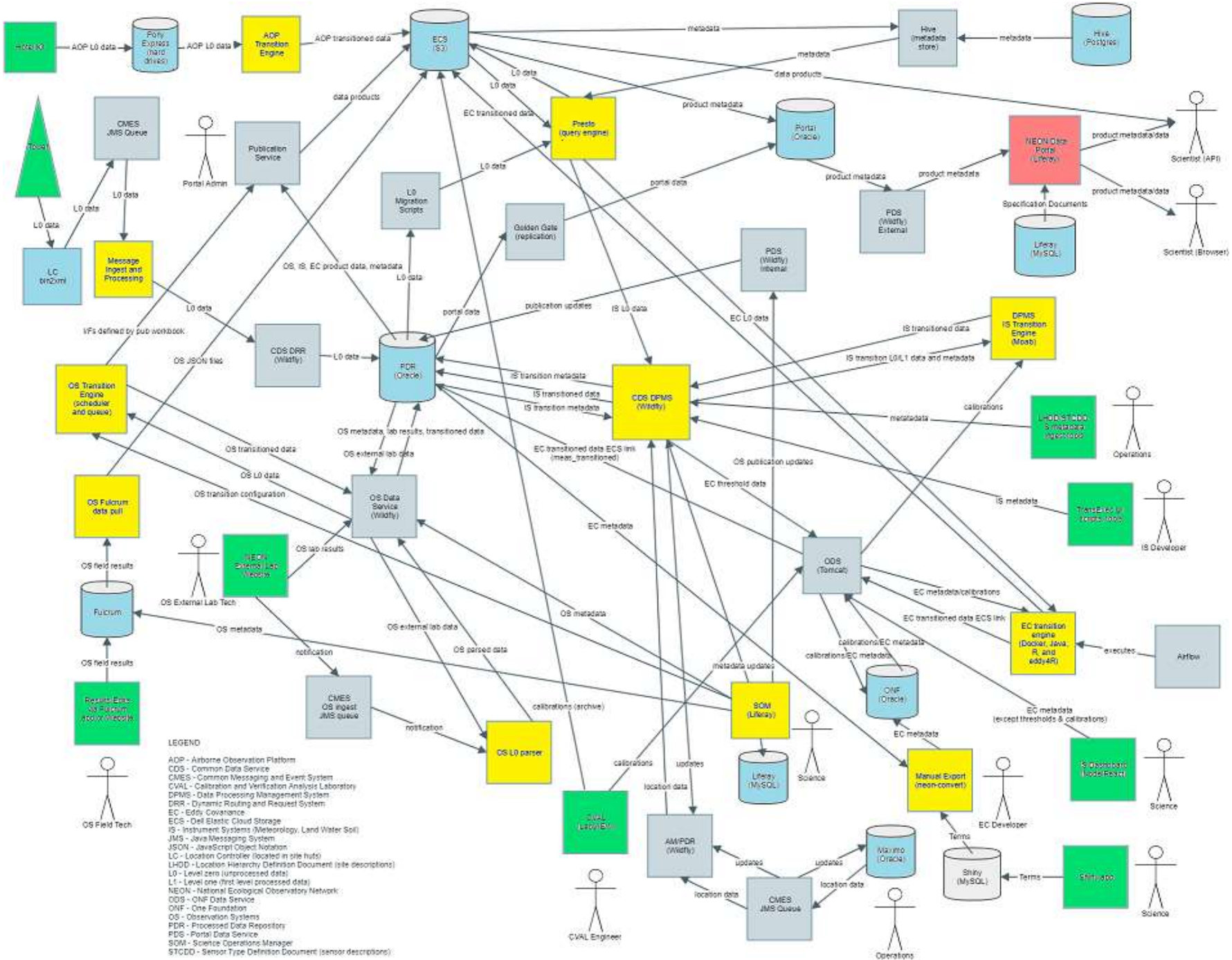
This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON CI Software Architecture - Elements

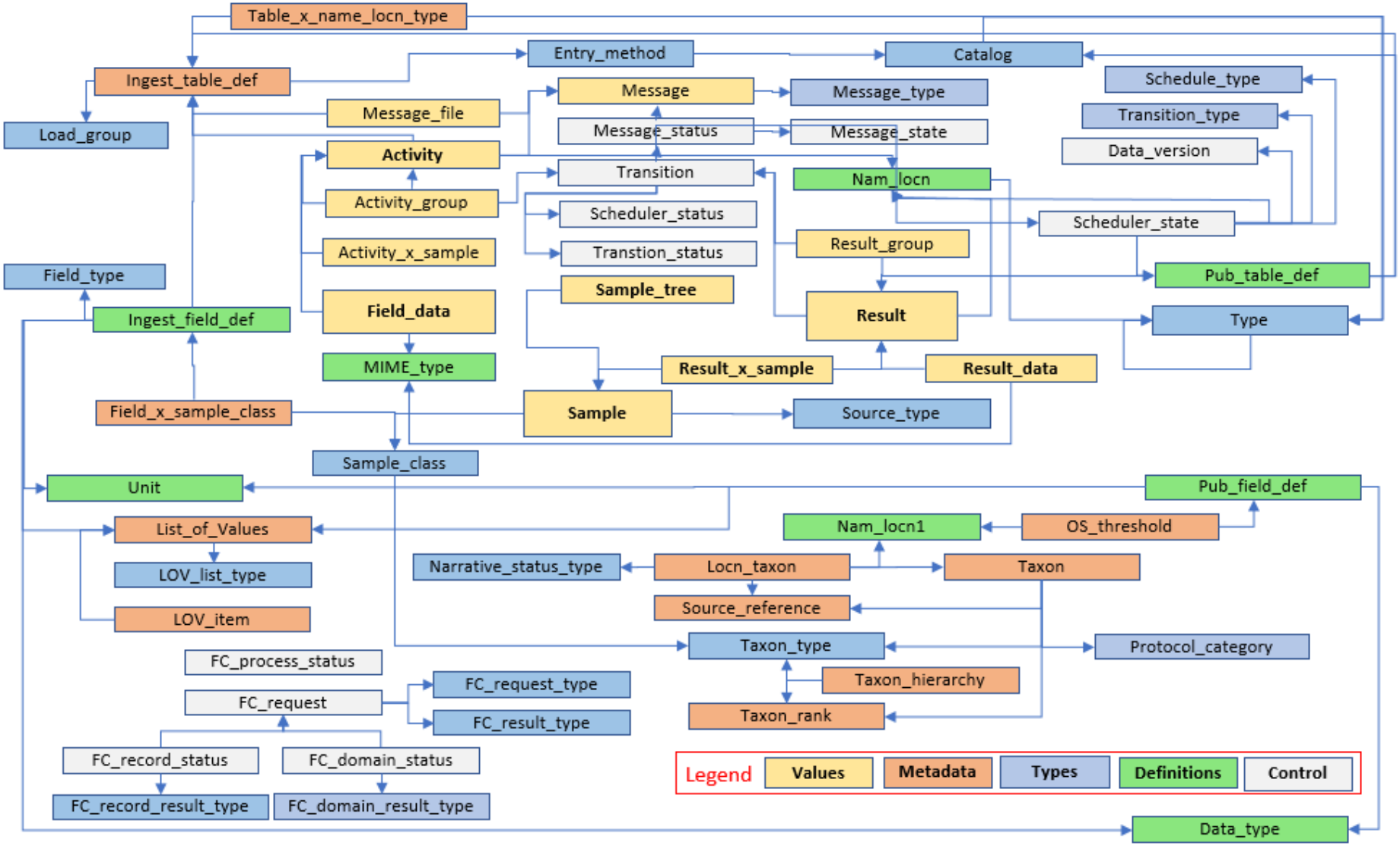


This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON Data Flow in CI Architecture

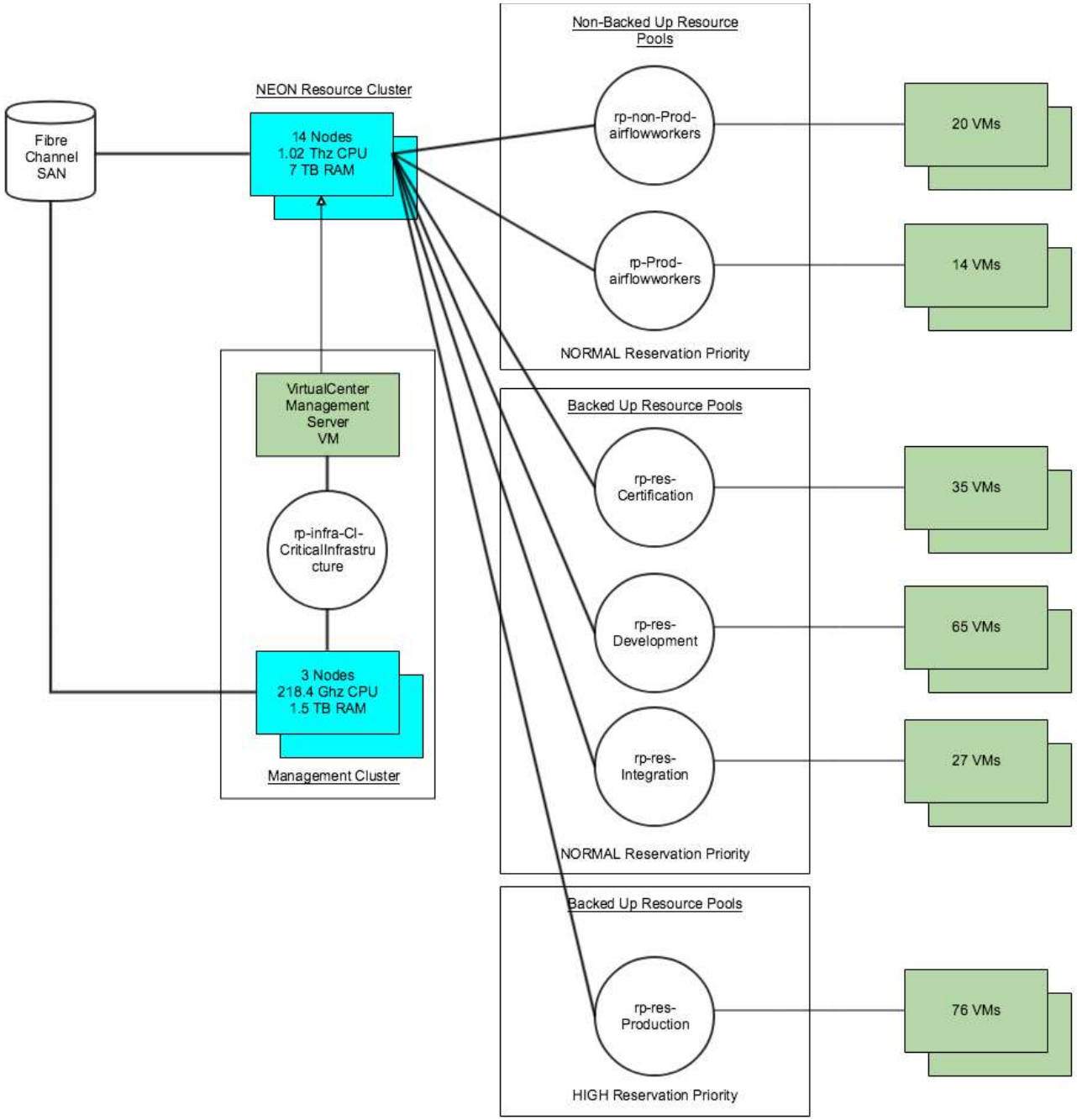


PDR Database – Observations Logical Model

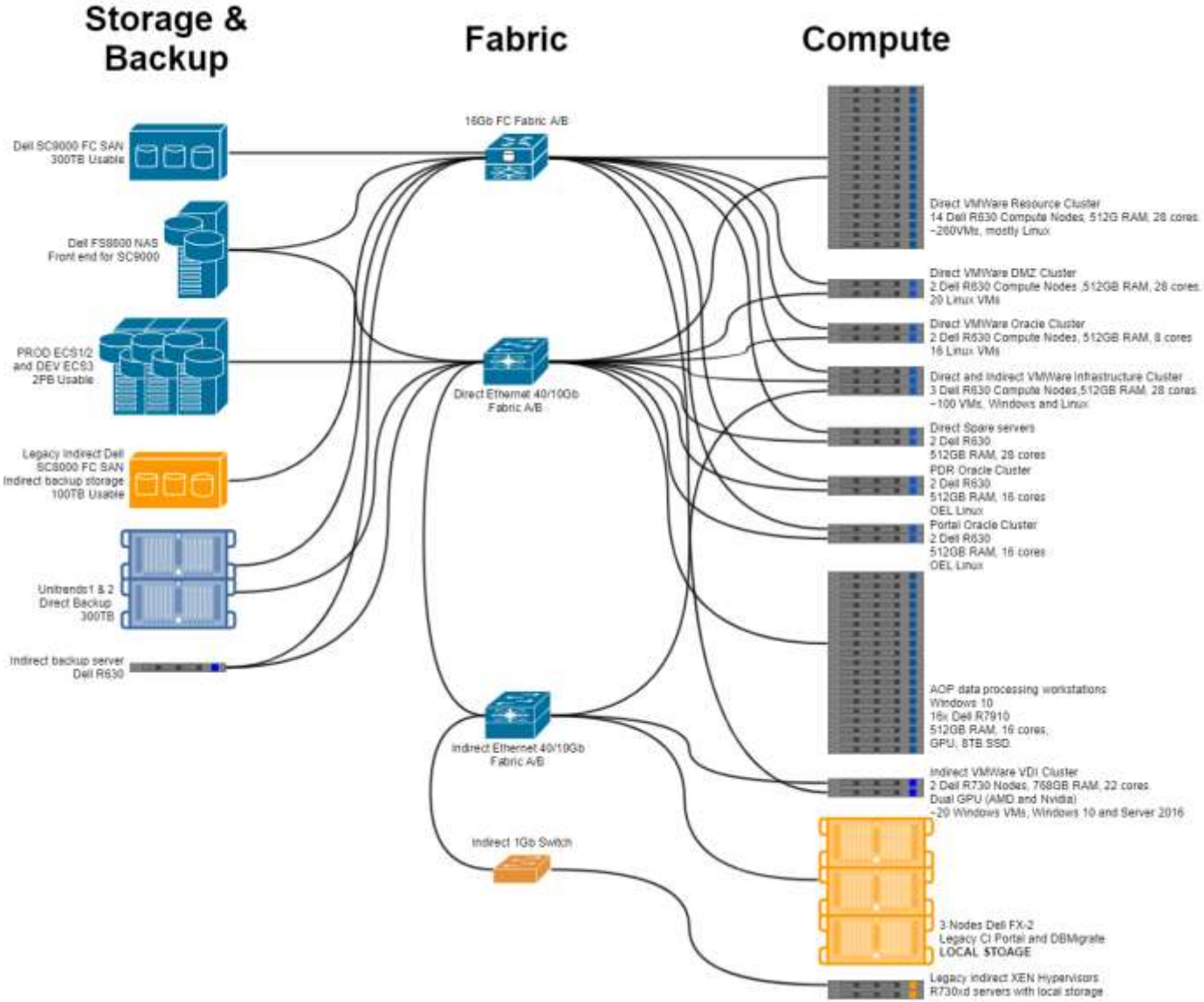


This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON CI Virtual Machine Resource Pools



Denver Datacenter – Configuration Schema



This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Cyber Security Overview

1. Security Strategy – NIST Cyber Security Framework: Identify, Protect, Detect, Respond, Recover
2. NIST CSF Timeframe – Program in progress – Goal is to address all categories by mid 2019
3. Perimeter Security – All Ingress / Egress Points are Firewall protected
4. Cloud Security – Encrypted Connections (TLS 1.0+) to Cloud Apps
5. Endpoint Security – Anti-malware / Endpoint Controls
6. Vulnerability Remediation – Constant patching schedule
7. Email Security – Microsoft Advanced Threat Protection – Phishing, Malware, Safelink protection
8. User Awareness – All employees are required to complete annual cyber security training

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

NEON CI & Data Interoperability

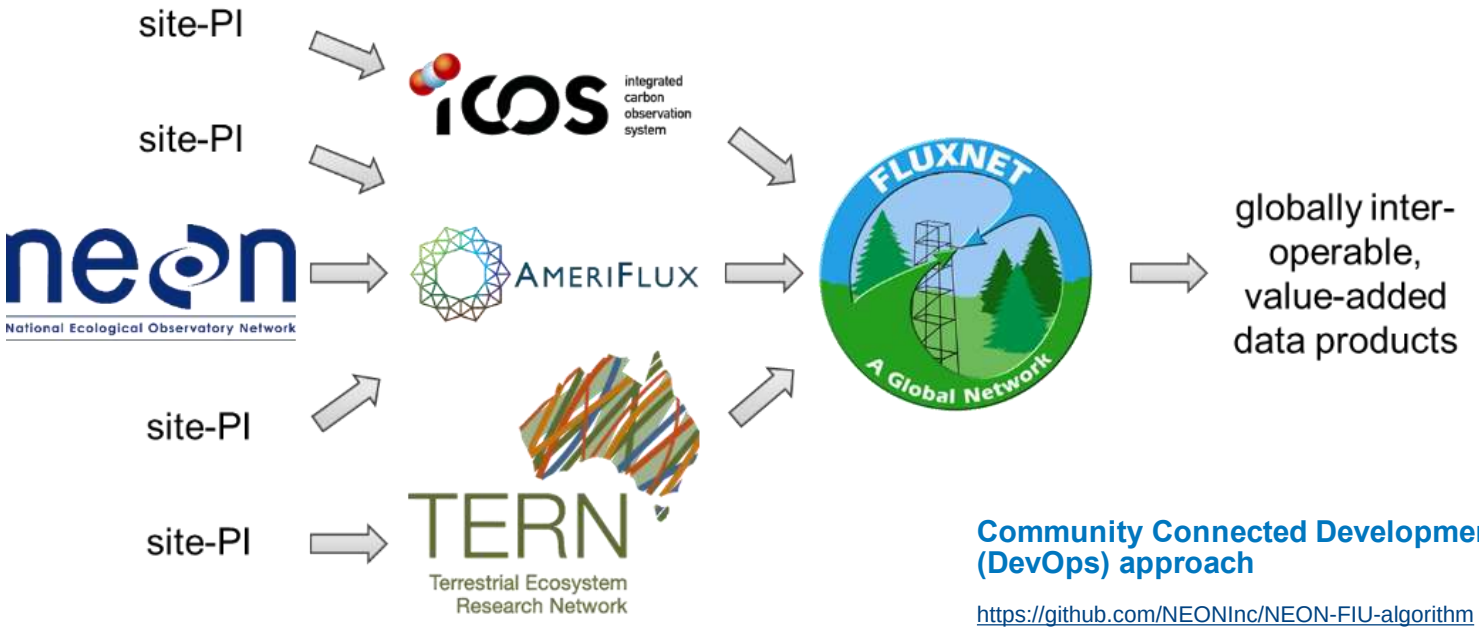
1. Across the community of researchers
   - Data collection, QA, standardized encoding and uncertainty handling protocols
   - Data exchange I/O formats & metadata
   - Data processing algorithms, code libraries

2. Between research facilities
   - NEON, AmeriFlux and Fluxnet common methods
   - Fully compliant metadata provisioning of NEON results to 6 partner host systems
   - Participation in meta-analyses of international CI methods, e.g. workflow objects

3. In active collaboration with aggregators
   - DataOne partner and metadata practitioner
   - EarthCube, CDF, ESIP, RDA coalition participants

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Data Product Collaboration (example)



Community Connected Development and Operations (DevOps) approach

https://github.com/NEONInc/NEON-FIU-algorithm

https://hub.docker.com/u/stefanmet/

Metzger, et.al. 2017. eddy4R 0.2.0: a DevOps model for community-extensible processing and analysis of eddy-covariance data based on R, Git, Docker, and HDF5. Geosci. Model Dev., 10, 3189–3206.

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Data Science Tools and Training: NEONScience on GitHub

- Open-source code packages to facilitate NEON data access and use
- Access to code used to generate NEON data products

```
Console   Terminal
~/
downloaded 29 KB

85 files downloaded to /Users/clunch/Desktop/filesToStack10003
> stackByTable("/Users/clunch/Desktop/filesToStack10003/", folder=T)
Unpacking zip files
  |=================================================================================================| 100%
Stacking table brd_countdata
  |=================================================================================================| 100%
Stacking table brd_perpoint
  |=================================================================================================| 100%
Finished: All of the data are stacked into  2  tables!
Copied the first available variable definition file to /stackedFiles and renamed as variables.csv
Copied the first available validation file to /stackedFiles and renamed as validation.csv
Stacked brd_countdata which has 74604 out of the expected 74604 rows (100%).
Stacked brd_perpoint which has 5506 out of the expected 5506 rows (100%).
Stacking took 4.126218 secs
All unzipped monthly data folders have been removed.
>
```

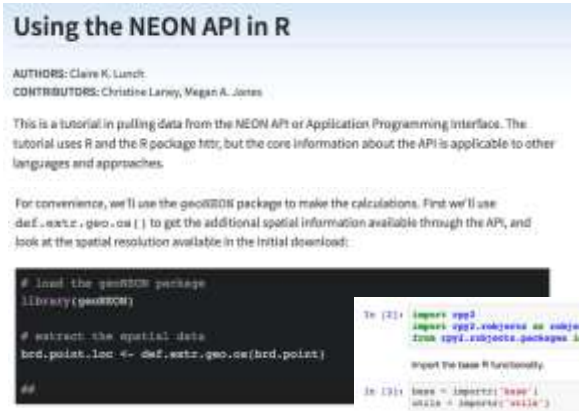
neonUtilities R package:

- Download data from NEON API
- Merge data files from Portal or API
- Convert file format for interoperability

This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

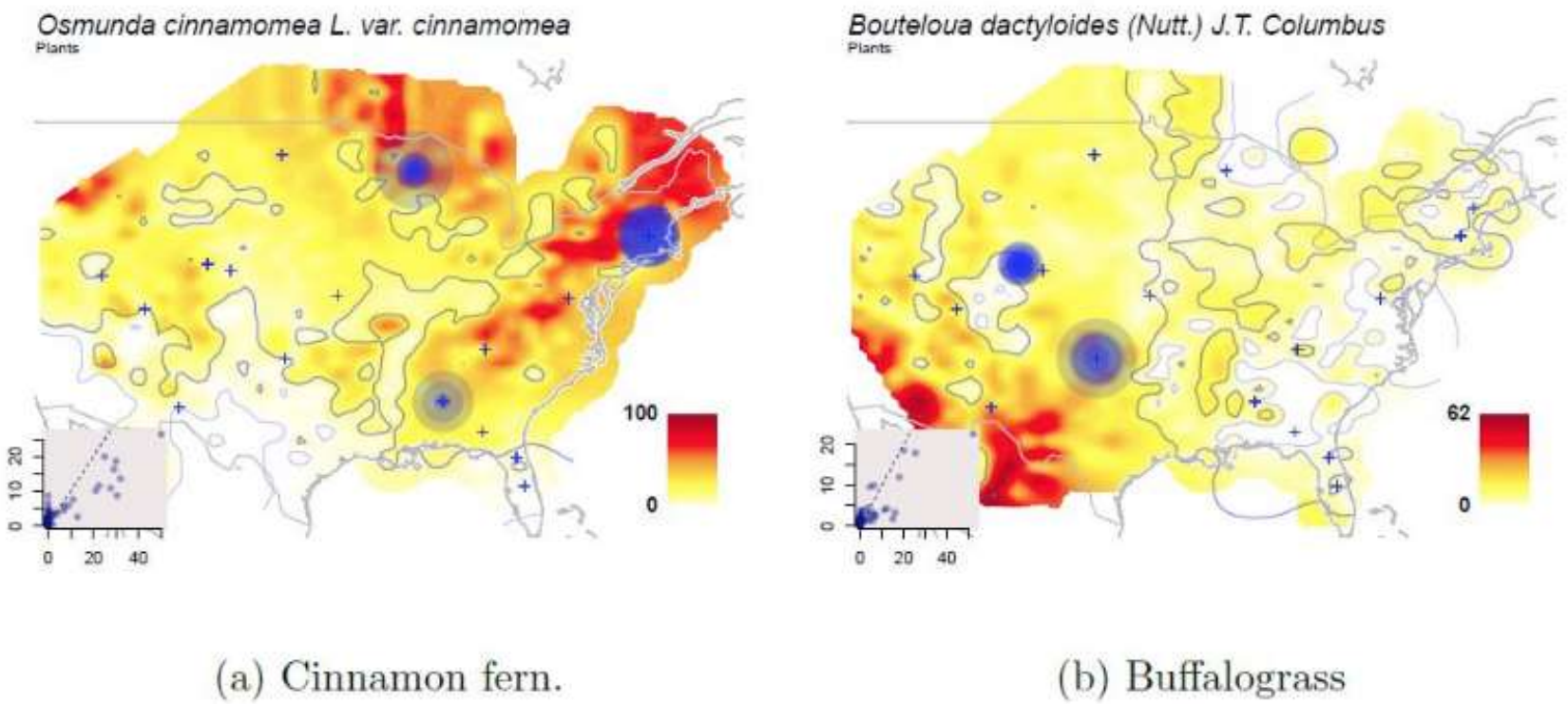
Data Science Tools and Training: Tutorials & Workshops

- Advance users' data analysis skill levels
- Online tutorials for self-directed training
- Data Institute on-site at NEON HQ
- Workshops at meetings and conferences
- Includes training in using the GitHub.com/NEONScience packages



This material is based upon work supported by NSF's National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

User Community Projects: Predictive Distributions



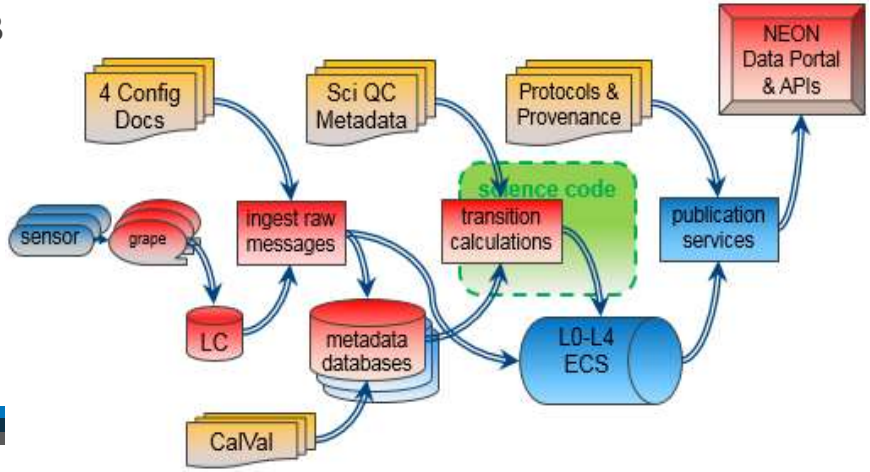
(a) Cinnamon fern. (b) Buffalograss

Figure 10: Predictive distributions for plant cover. Symbolism follows Figure 4.

Clark lab: sites.duke.edu/neon/prediction

NEON CI Improvement Plans

1. Ongoing cyberengineering support
   - e.g. better interface to manage sensor metadata; unexpected data gap checks…
2. Prioritized queue of CI or Data Product enhancements
   - e.g. improve minimum viable products; address tech debt; better algorithms…
3. Service Management continued expansion/integration
   - e.g. more workflow delegation; closer linkage with SOM & monitoring…
4. Asset management tools & methods assessment
   - e.g. mitigate risks of vulnerable sensor installation metadata addressing…
5. Proposed research & development towards VI version2
   - Redesign data acquisition hw&sw (i.e. “GRAPE” data streams)
   - Pilot optimized transition engine & DB
   - Upgrade Data Portal platform & tools
   - Evaluate CI capacity, connectivity & offsite disaster recovery options



This material is based upon work supported by NSF’s National Ecological Observatory Network, a major facility fully funded by the National Science Foundation.

Proudly operated by BATTELLE

720.746.4844 | neonscience@battelleecology.org | www.neonscience.org