

Christine Laney, Data Scientist
With content provided by Steve Jacobs, Janae Csavina,
Claire Lunch, Caren Scott, Bridget Hass, Tristan Goulden,
and Andrew Bigoney

NSF CI Compass Workshop: Data Management: From
Instrument to First Storage

January 24, 2025



neon
Operated by Battelle

Weaving NEON data into the cloud

National Ecological Observatory Network (NEON)

A continental-scale, long-term (30 year) Observatory, funded by NSF and operated by Battelle that enables:

- **Analysis:** Free and open data and samples on the drivers of and responses to environmental change
- **Comparison:** Standardized and reliable framework for research and experiments
- **Interoperability:** Integration with other national and international network science projects



NEON is important for ecology

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



Downed trees at The Jones Center at Ichauway, GA



Fire damage at Great Smoky Mountains National Park, TN



Ice breakup on Oksrukuyik Creek, AK

Data and Sample Collection

Standardized, colocated methods across sites



Automated
instrument
systems



Airborne
observation
platform

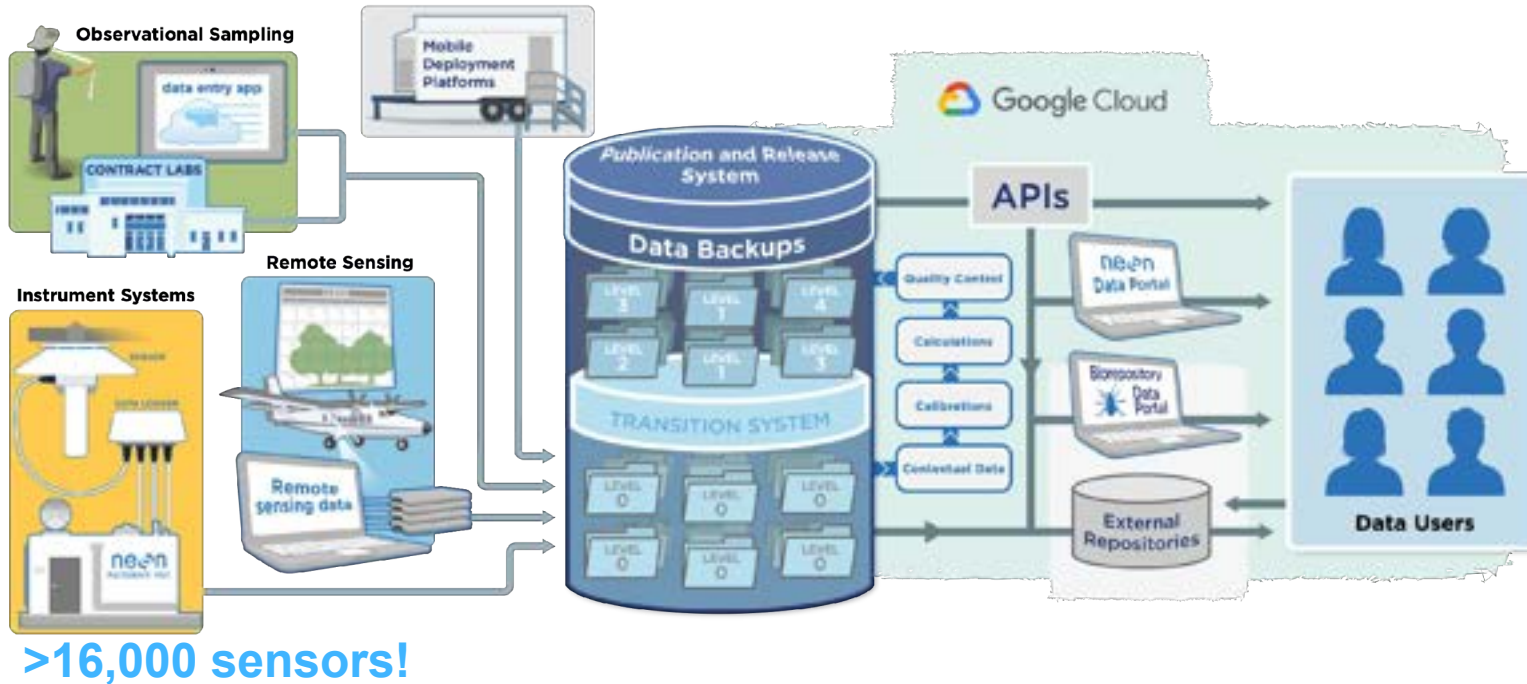


Observational
sampling



Data lifecycle

Collection, processing, publication, and management



System	Total Data Volume (raw and published)
Remote sensing (AOP)	1.2 PB
Instrumented systems (IS)	375 TB
Observational systems (OS)	71 TB

- **NEON ingests 5.6 billion data points per day** (~65K data points/second) of instrumented data
- In comparison, VISA Net handles 150 million transactions per day* (1.7K transactions per second)

* <https://usa.visa.com/run-your-business/small-business-tools/retail.html>

NEON Publishes >180 Data Products

Atmosphere



H₂O, CO₂
Heat
Micromet
Isotopes
Turbulence
Storage
Fluxes

Biogeochemistry



Soil conditions
Chemistry
Particulates
Foliar characteristics

Ecohydrology



Water quality
Precipitation
Discharge
Radiation
Geomorphology
Riparian Structure

Land Cover & Processes



Spectrometry
Hi-Res Imagery
LiDAR

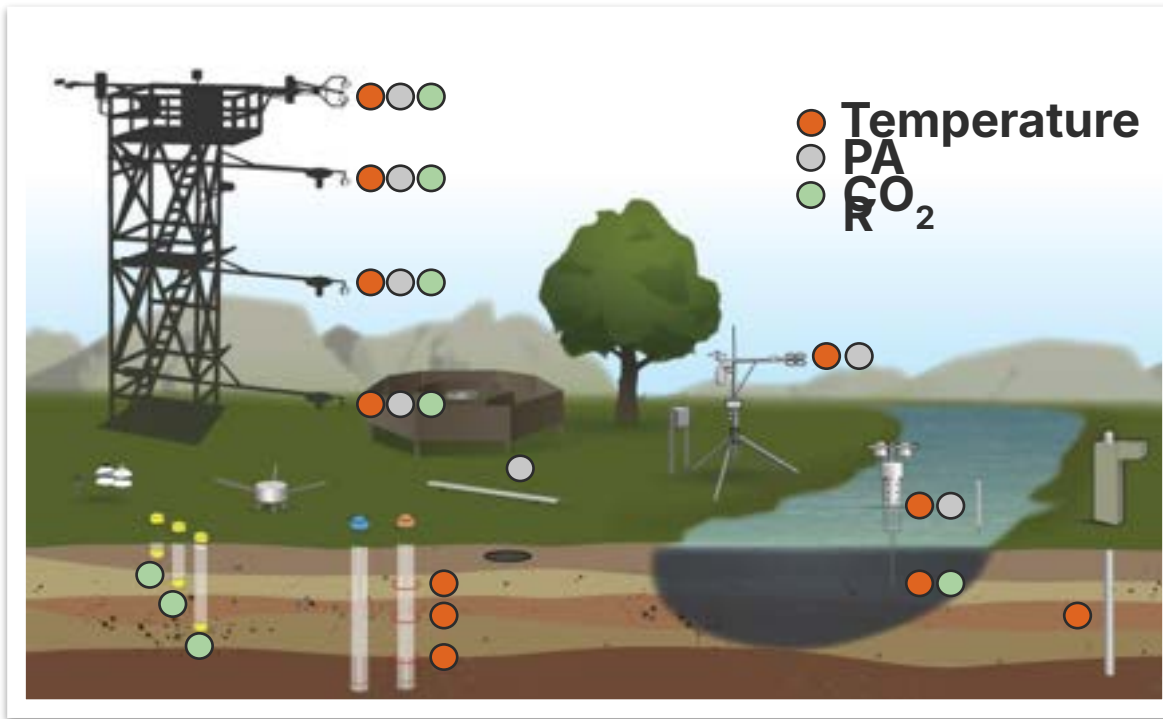
Organisms, Populations, & Communities



Abundance
Composition
Biomass
Productivity
Pathogens
Phenology
Marker Genes
Metagenomics

Instrument System: *In-situ* sensors

- >16,000 sensors generate 5.6 billion data points per day
- The NEON instrument & spatial design allows measurements to be linked within & across sites



Tower at Niwot Ridge, CO



Aquatic sensors

Raw data collection

- **Two basic types of data:**
 - **Readings:** have periodicity – up to 40hz
 - **Events:** happen at unscheduled times
- A single sensor may generate more than one data stream, but if all streams have data at a single timestamp, this set may be saved as a single record
- Some data are only used on-site to inform system command and control processes, such as the turning on/off aspiration fans or gas valves
- Raw data are retained in as close to manufacturer specifications as possible

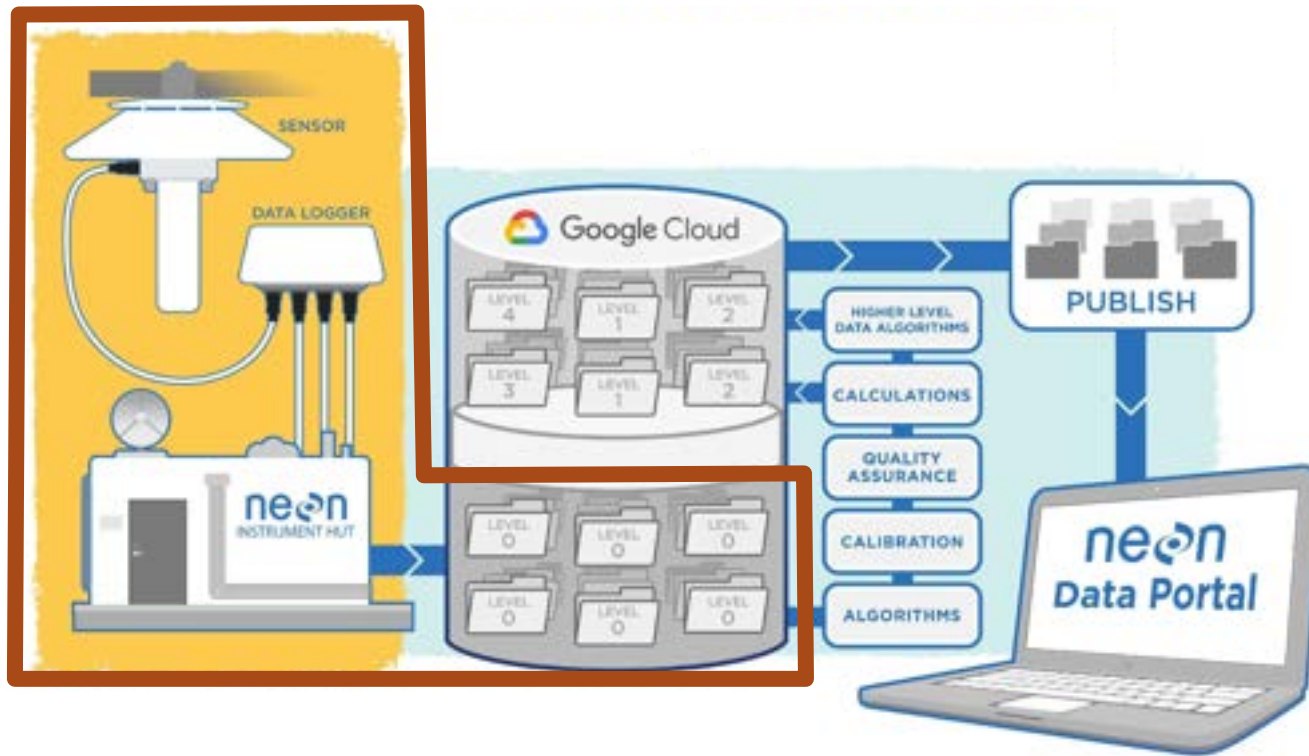


Soil array at Moab, UT



Interior of instrument hut, Yellowstone, WY

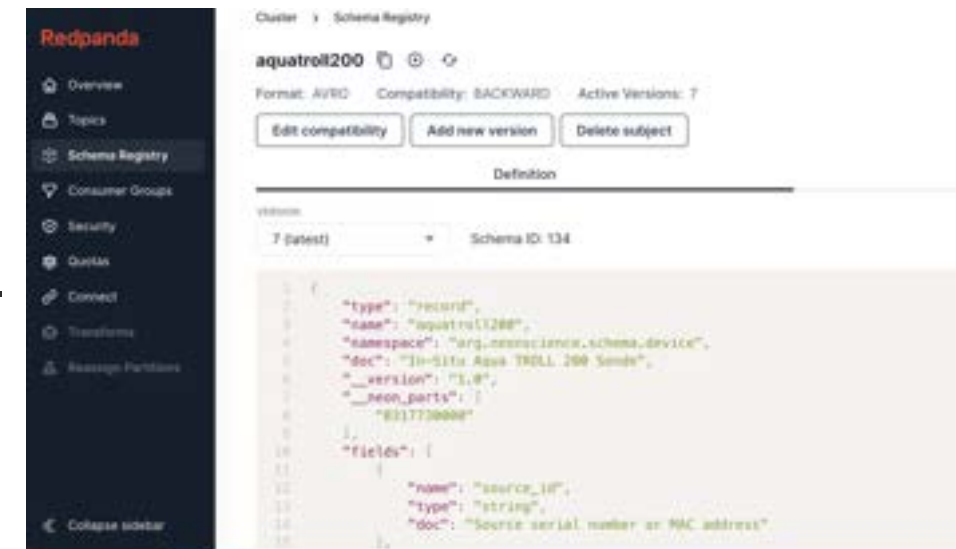
Instrumented systems pipeline



- Data are streamed within site and to headquarters via numerous technologies:
 - Custom data loggers (GRAPes), which collect analog and digital measurements
 - Campbell data loggers, via LoggerNet
 - Sensors' own built-in network capabilities, over ethernet
- Data are retained at **sites for 60 days** and in **GCP for 2 weeks post ingest** (usually only 1 day is needed) to protect against data loss if site connectivity is interrupted.

Raw data storage & processing

- Kafka is used to handle data records.
 - Certain data needed for command and control are written every second.
 - Most data are published into Kafka every 8 seconds in a compressed format to reduce storage and bandwidth needs.
 - Metrics for monitoring processes at the sites use data recorded every 15 seconds and are not run through Kafka.
- An AVRO schema is used for both raw data and processed data.
- Kafka MirrorMaker 2 is used to copy Kafka topics from the site computers to Google Cloud Platform



Recent Network Improvements

Starlink Low Earth Orbit (LEO) satellite service has been installed successfully at 20 sites to date; 5 additional Starlink satellite deployments are planned in AY25. This is especially important for very remote field sites.

- **Lower cost:** \$140/month compared to \$1000/month for cellular network sites
- **Lower latency:** ~ 60ms less than cellular sites, ~400ms less than (GEO) geostationary satellite sites
- **Increased bandwidth:** up to 100Mb/sec download and 10Mb/sec upload for many existing cellular and satellite sites
- **Improved reliability/uptime** at sites where cellular and satellite signals were degraded or experiencing regular network congestion and overload.

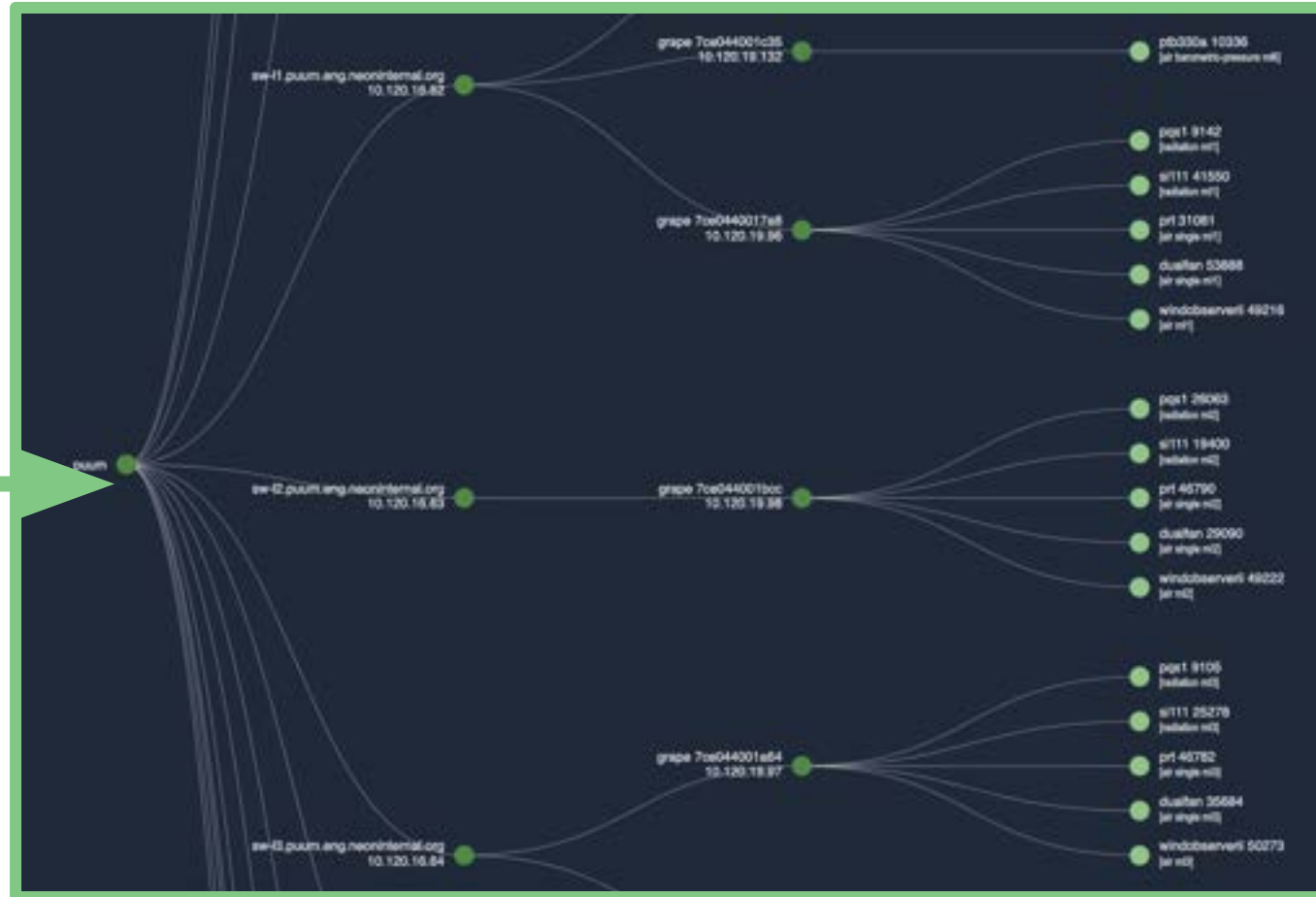
This week (of Jan 20, 2025) was the first time since February 2023 where all sites have been online during a morning status call. 🎉



Niwot Ridge, CO, instrument hut with Starlink dish

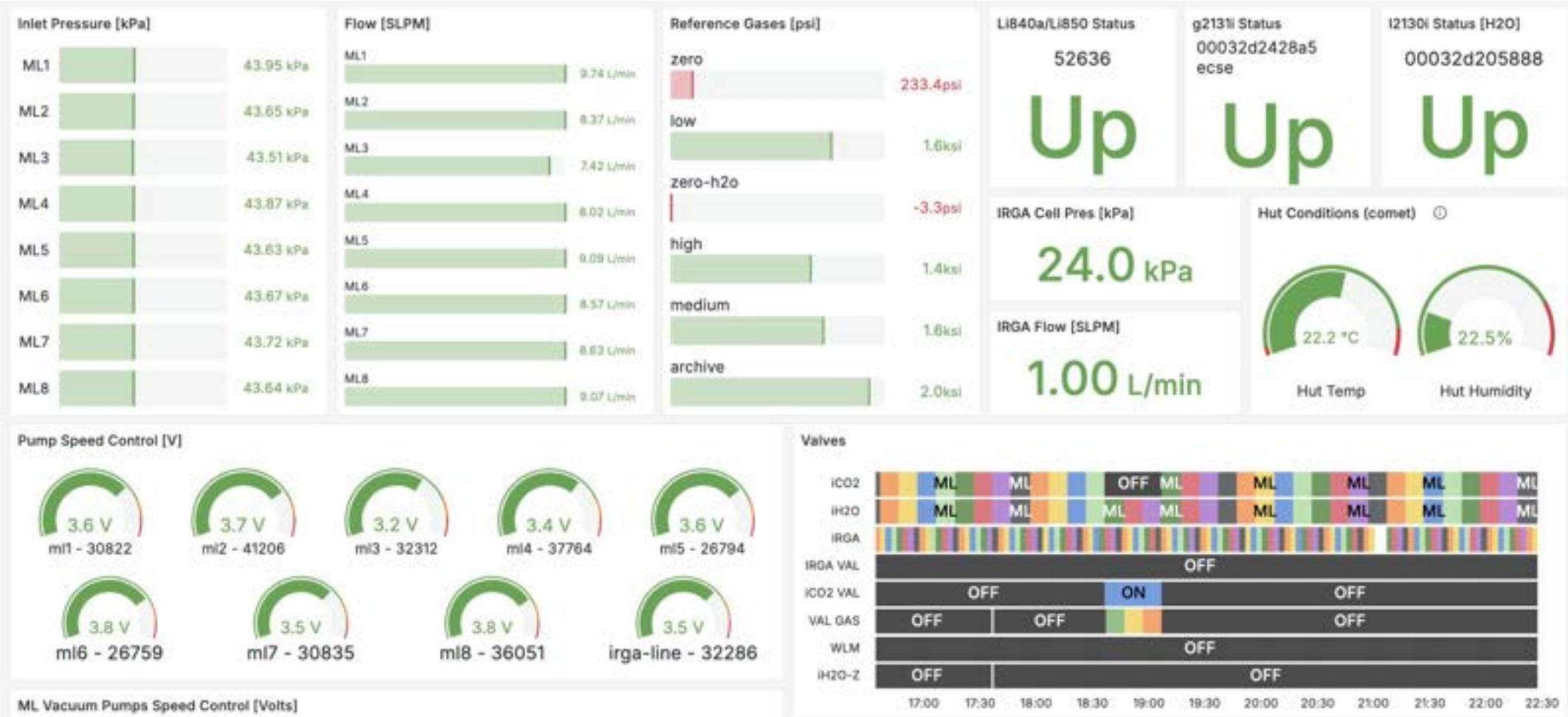


The screenshot shows a mobile application interface with a dark background. At the top, there is a navigation bar with a hamburger menu icon, a home icon, and the text "PUUM". Below the navigation bar, there is a list of items. A red rectangular box highlights a specific item in the list. The item appears to be a book or document with a cover image and some text. The text "PUUM" is visible in the navigation bar.



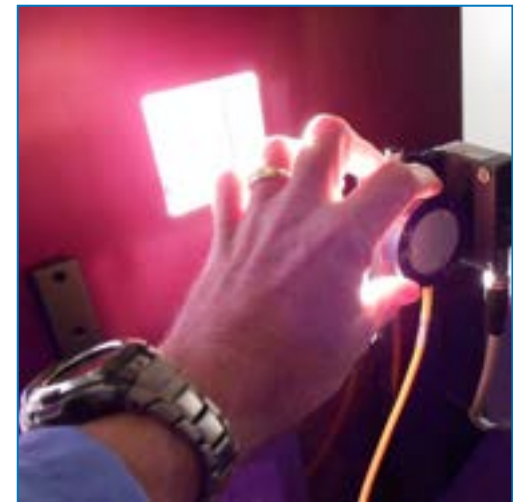
Monitoring instrumentation

ECSE - Eddy Covariance Storage Exchange System at Wind River Experimental Forest (WREF), WA,
2025-01-23 15:20



Calibration, Validation, and Audit Laboratory (CVAL)

- Annually calibrates/validates sensors and data acquisition systems (DAS).
 - Rotates 10,000 sensors and 3,000 DAS annually to meet client requirements
 - Controls quality by ensuring performance of sensors and that DASs meet project requirements
 - Provides internal assessment of measurement uncertainty
 - Informs drift in measurements by comparing years of calibrations to understand field use impacts, drift corrections, calibration cycle and sensor life



NEON's in-house metrology lab

- Indoor and outdoor facilities with metrology infrastructure
- Internal expertise developed metrology fixtures
 - Designed, engineered and verified **53 different calibration and validation fixtures** with traceability to national and international standards for NEON measurements
 - Developed **LabView GUIs** to provide automated calibrations with customized controls, operator instruction, data transformations, and outputs
 - Developed Laboratory Information Management System (**LIMS**) to provide adaptability of data management and quality controls
 - Developed **Mobile applications** to calibrate non-rotating sensors



More CVAL!

- Creates aquatic standards, isotopic carbon dioxide gas standards, and isotopic water standards
 - Annually create 100 aquatic standards, 500 gas standards and 300 isotopic water standards.
 - Used in the field for traceable calibrations and validations.
- Obsolescence and Technology Upgrades
 - Evaluate obsolete or previous sensors to new sensors with the same standards to understand changes in the 30 years of the project.



Airborne Observation Platform: Remote Sensing

- 40 sites planned for imaging in 2025 via 2 payloads mounted on rented airplanes
- A 3rd payload is reserved for filling in times when other payloads are offline or for PI-driven research.

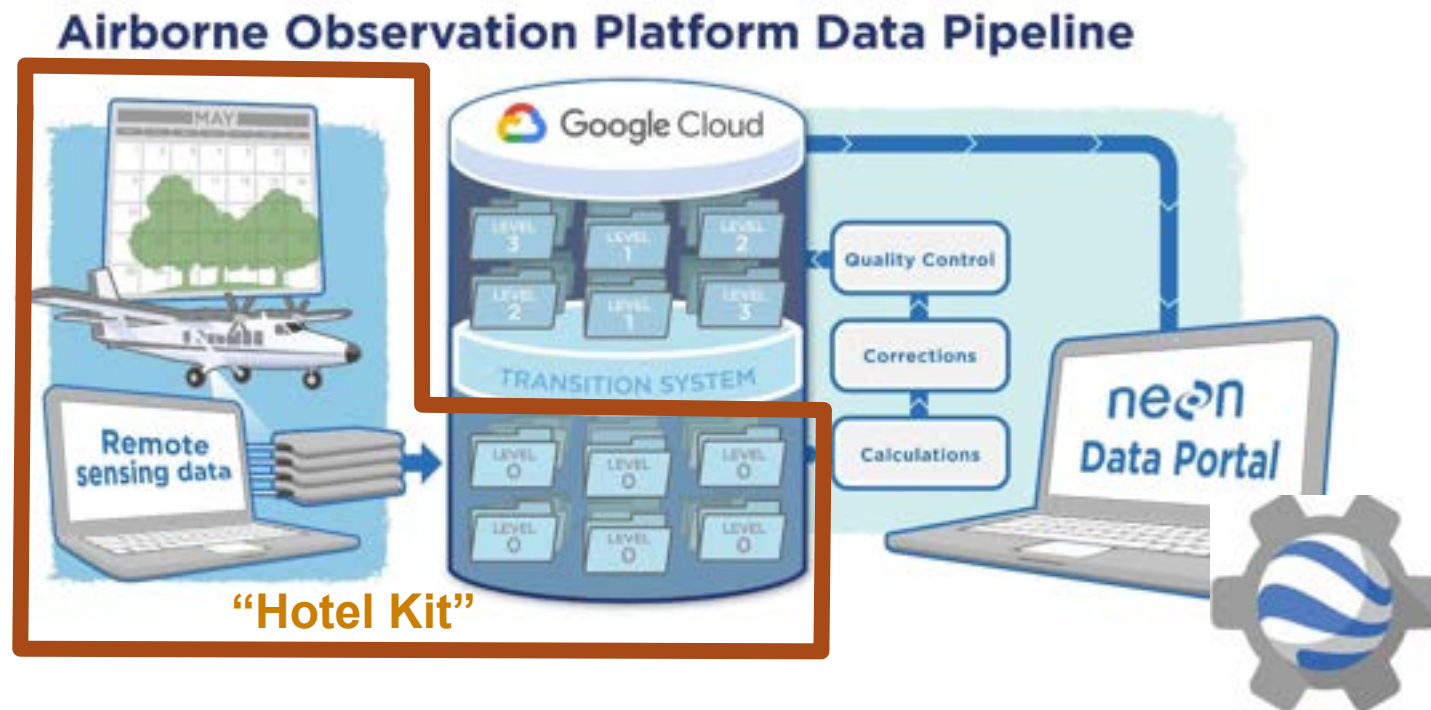


Camera, imaging hyperspectrometer, and lidar peering from the underside of the plane



AOP Data Pipeline

- **~1.7 TB of data are generated per flight**; ~40 sites flown per year plus calibration and PI-driven research support flights -> **157 TB in 2024**
- Initial extraction of waveform lidar ~doubles volume
- Basic QA is performed in the Hotel Kit
- Checksums generated on the database during the extraction are verified as part of the ingestion, ensuring data integrity on GCS



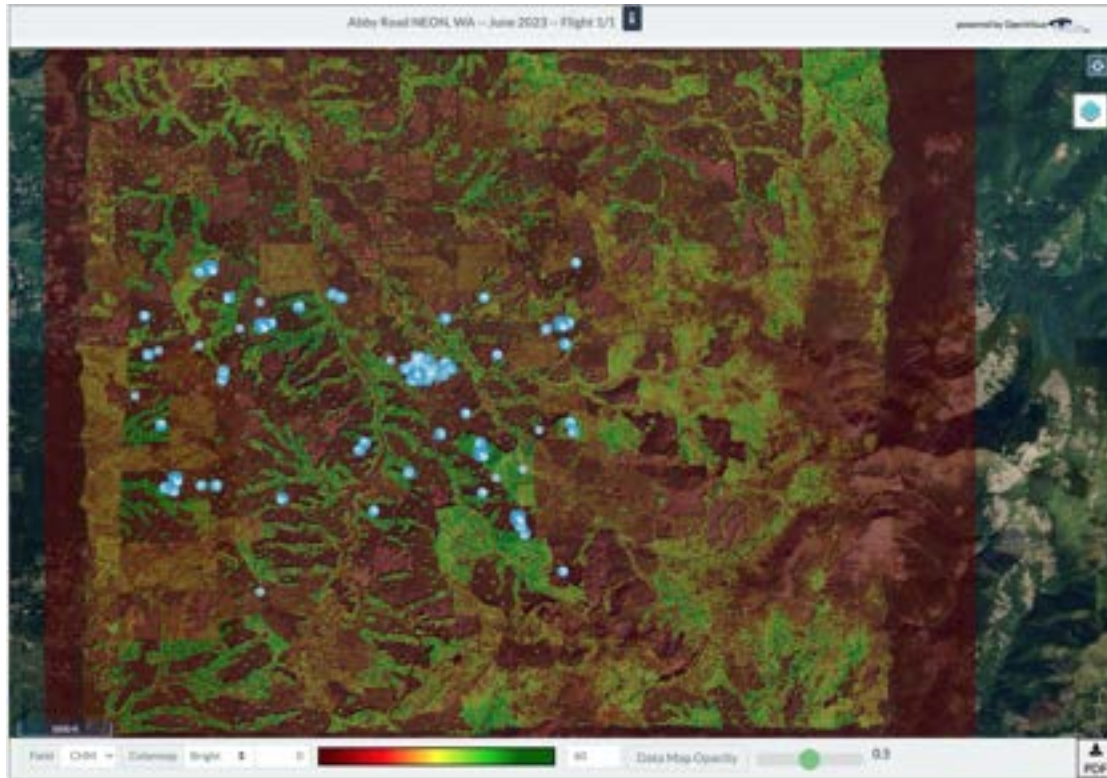
Data Volumes Increasing

- Last year, NEON produced more than twice as much raw sensing data than previous years, due to an improved waveform lidar that can operate at higher pulse frequencies (i.e., much denser data).

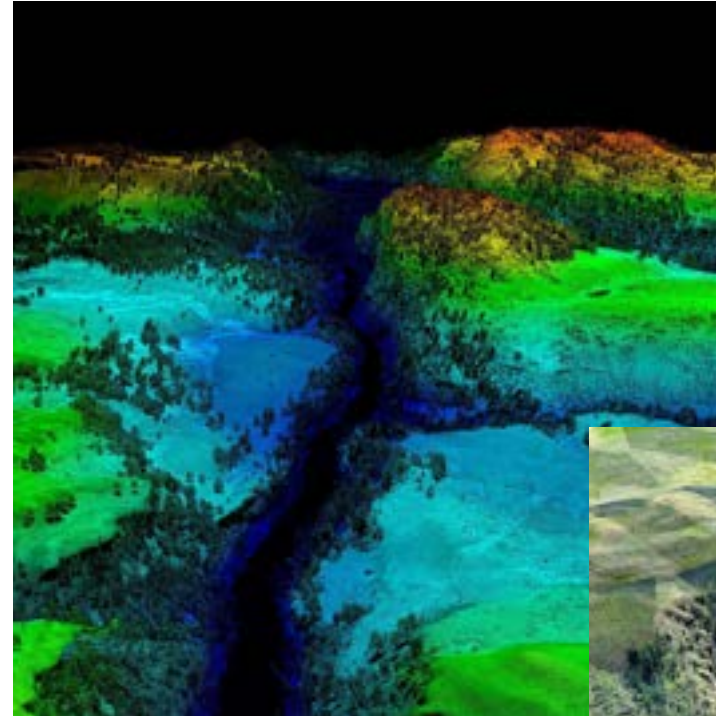
Year	# of missions	Lidar (TB)	NIS (TB)	RGB (TB)	Total (TB)
2019	141	9.22	43.65	12.59	65.46
2020	69	4.19	18.53	5.3	28.01
2021	98	9.82	35.29	14.25	59.36
2022	91	7.73	33.91	13.36	55.00
2023	80	6.67	28.14	12.17	46.98
2024	86	82.87	28.19	12.35	123.41

- This is likely to happen anywhere across the suite of instruments NEON deploys, so we need to evolve the cyberinfrastructure to meet science needs.

Outputs



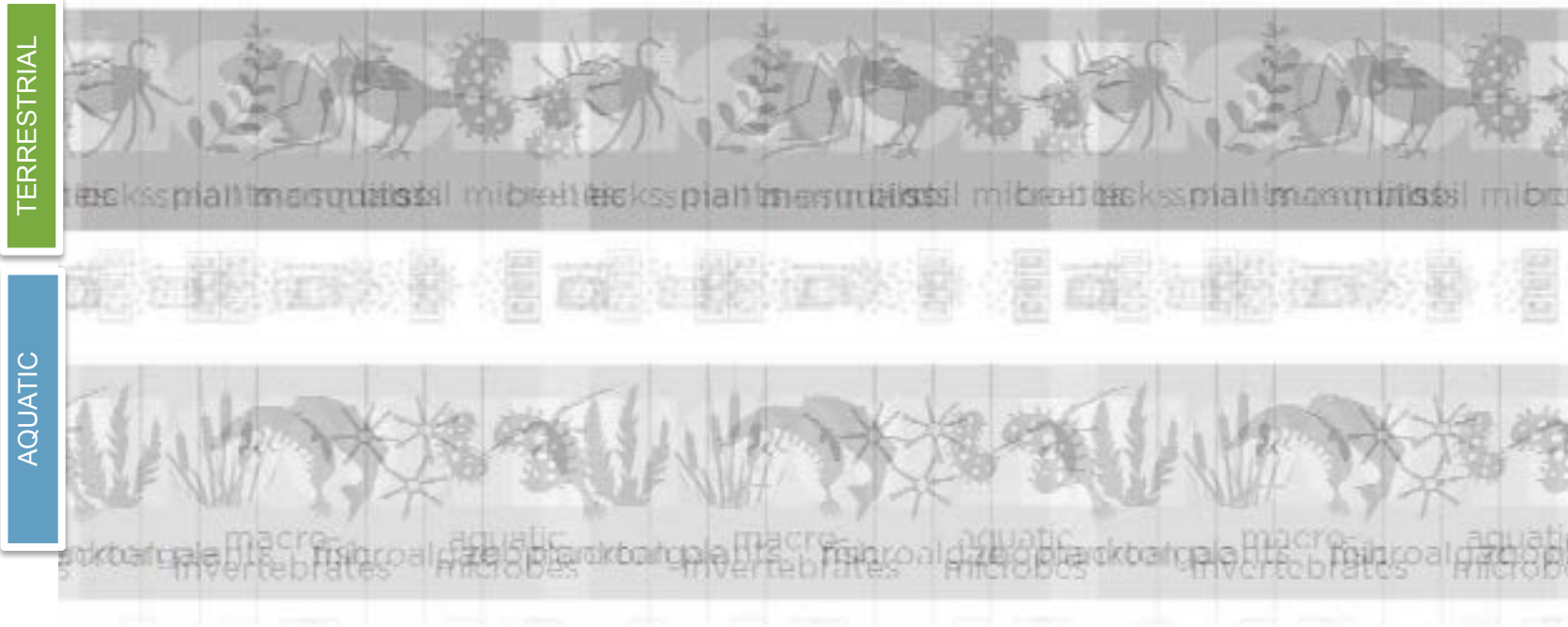
Abby Road, WA – canopy height



Lidar at Yellowstone, WY
– false and true color



Observation System



- 250-400 field science staff (seasonal flux) across 18 domain offices
- Make observations via > 50 standardized protocols & maintain infrastructure

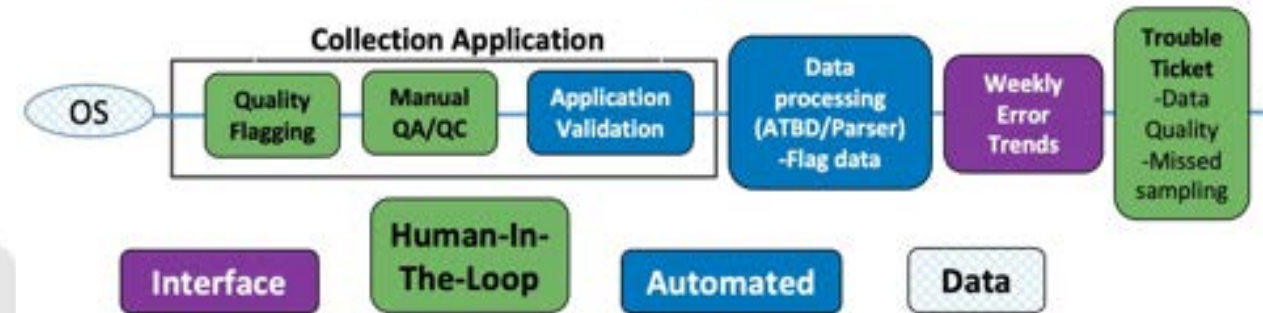
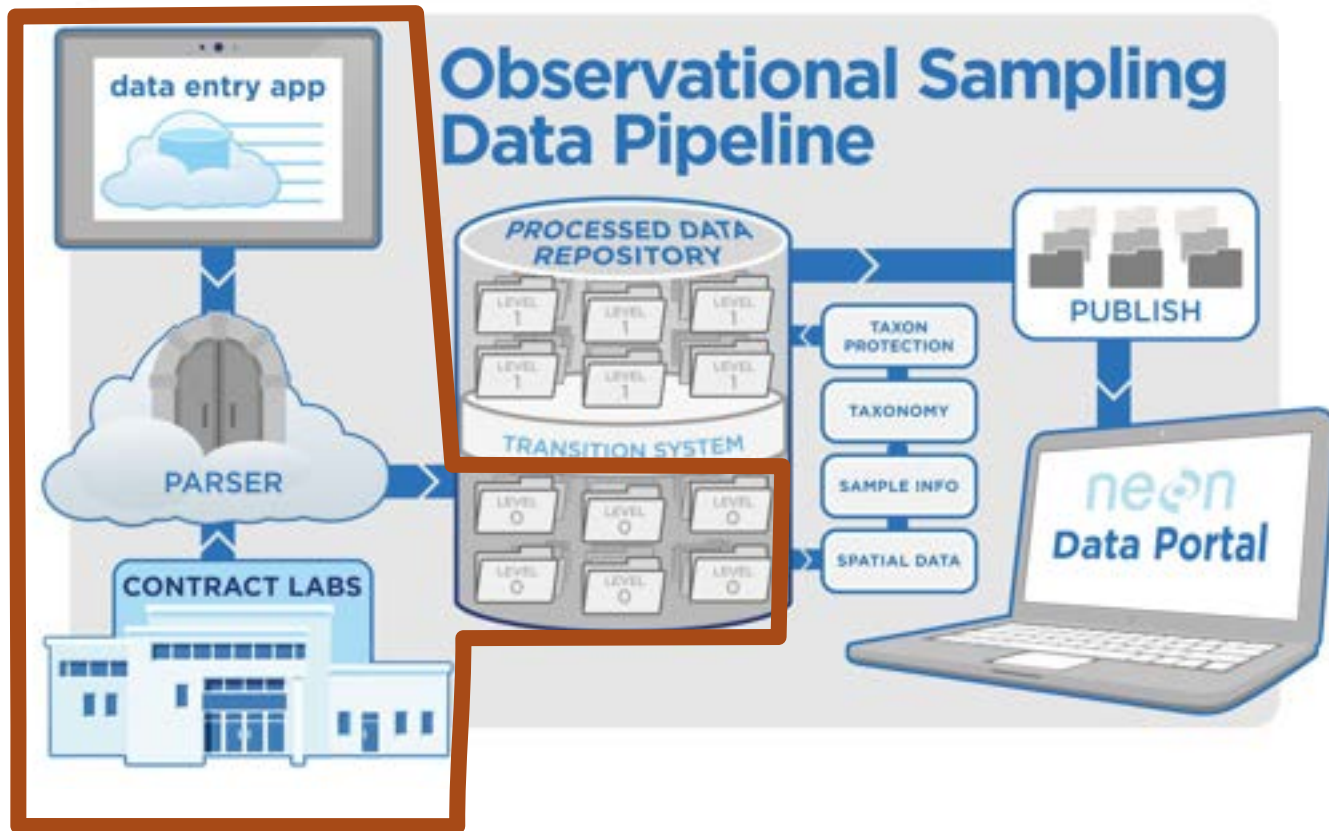


Observational data and sample collection volume

- ~400,000 samples per year, counting every database-tracked sample
- ~100,000 physical samples archived
- ~3 million data records per year
- Smallest collection of data at NEON, but highly valuable



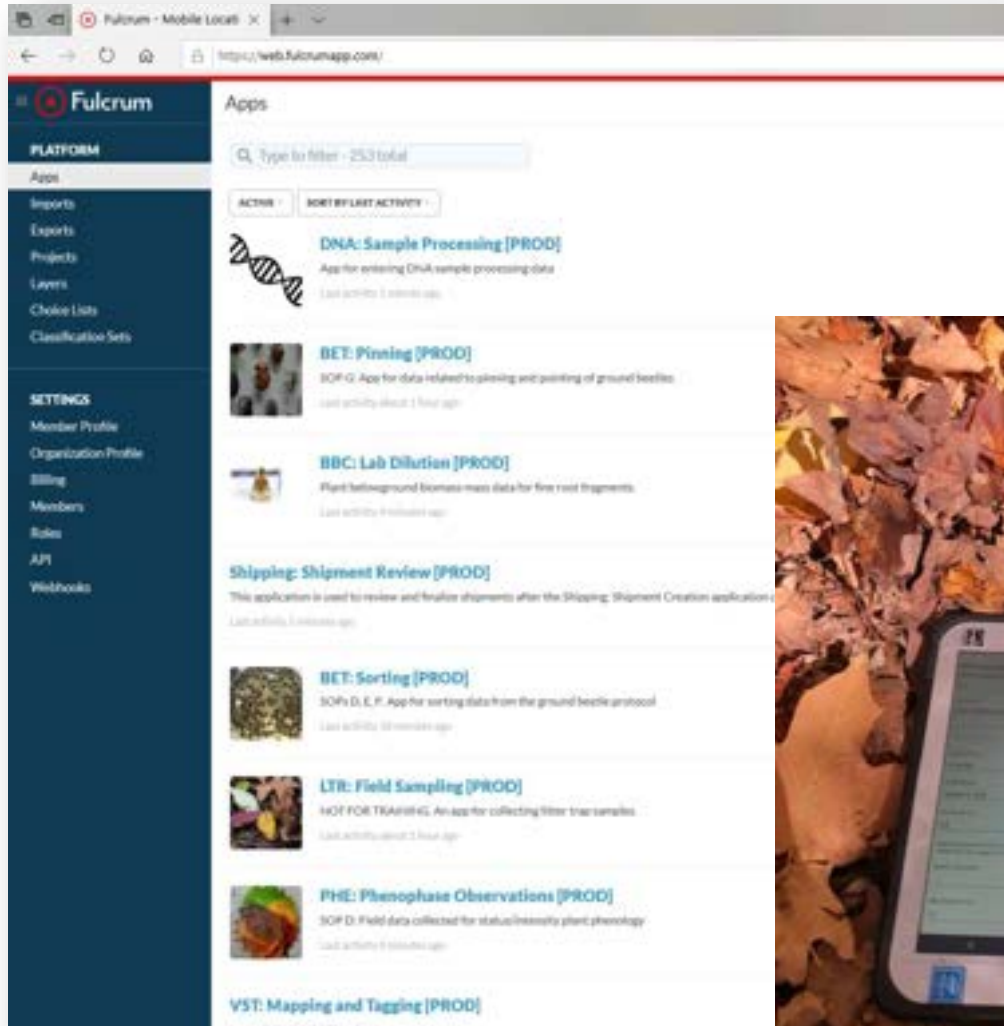
Data & Sample Pipeline



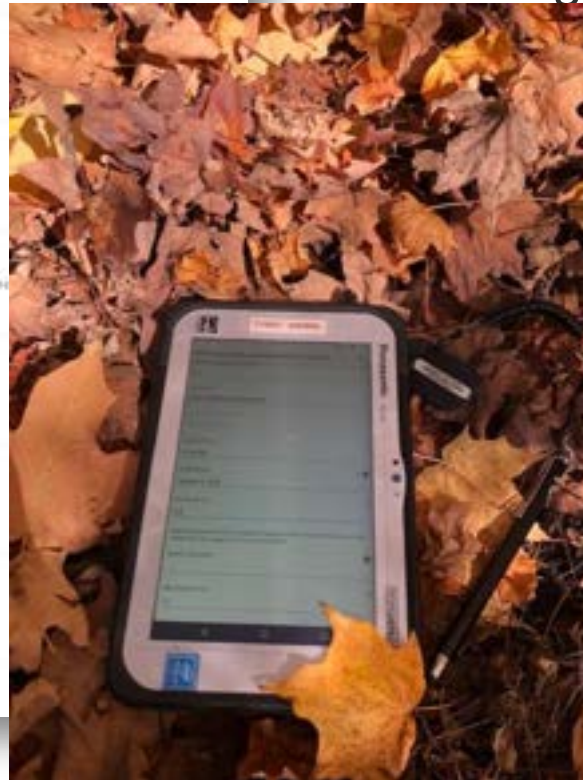
Primary principles:

- **Quality control is front-weighted:** Data entry is controlled, and data ingest is prevented if data don't meet criteria
- **Pipeline is built on standardized frameworks** and platforms that are **flexible** enough to be adapted to different data structures

Observed data ingest & QA via Fulcrum Apps



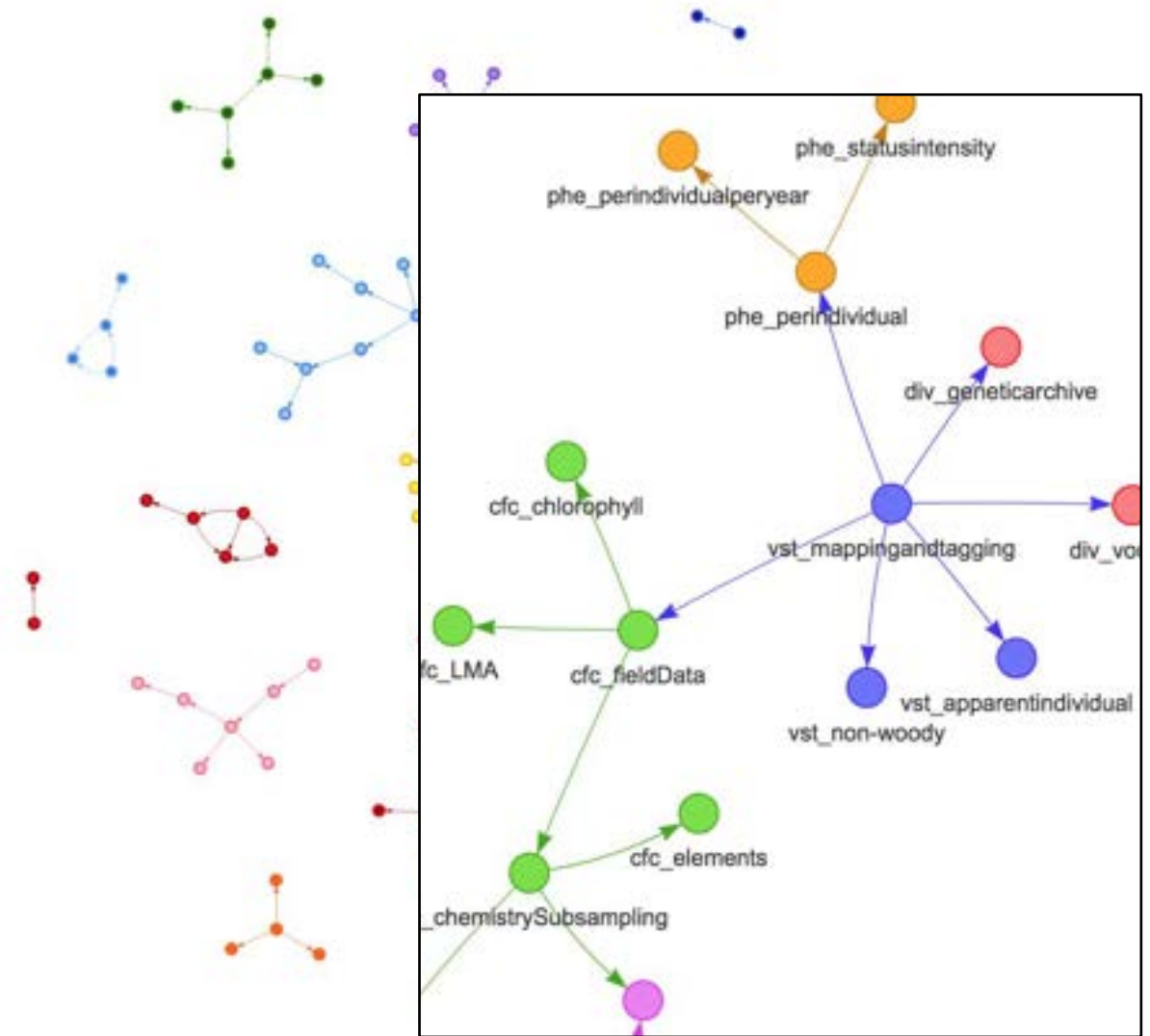
- Required fields
- Auto-generated fields
- Conditionally available fields
- Choice lists
- Range limits



Plot ID	* CPER_007_basePlot
yearBoutBegan	* 2018
eventID	HBP2018.CPER.01
SubPlot ID	* 31
Collect Date (YYYY-MM-DD)	* 07/06/2018
Collect Time (HH:MM)	* 01:54 PM
JustcollectDate	July 6, 2018
Clip Cell Range	CLIP CELL NUMBER MUST BE BETWEEN 1-350
Clip Cell Number	* 3
Clip Dimensions (m)	* 2.0m x 0.1m
clipID	CPER_007_003
sampleID	hbp.CPER007003.20180706
Target Taxa Present	* <input checked="" type="button" value="Yes"/> No
Sample Fate	* active
Sampling Protocol Version	* NEON.DOC.014037vF
hbp Type	* Non-agricultural
Fresh Sub Samples Created?	Yes No

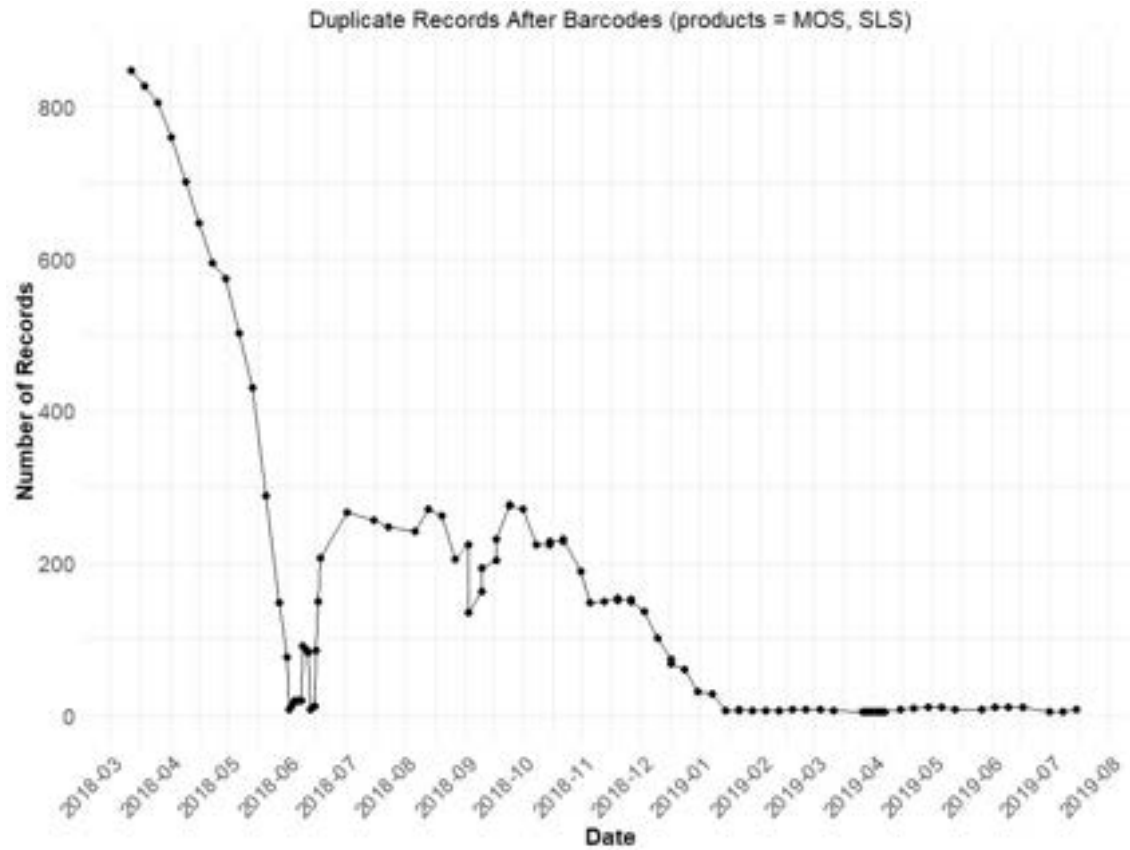
Standardized framework

- The NEON database tracks ~400,000 samples annually; manual verification of identifiers is impossible
- Reliant on barcodes and robust identifiers for both samples and staff (ORCiDs).



Data relationships as linked by samples, Observatory-wide

Reduction of errors



NEON specimens & samples: NEON Biorepository

>100,000 specimens & samples per year



ASU Arizona State University

BARCODE OF LIFE DATA SYSTEM v4

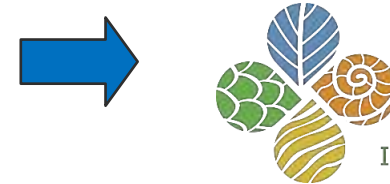
Advancing biodiversity science through DNA-based species identification.



GBIF

Global Biodiversity
Information Facility

>1 M
occurrences



iDigBio

Integrated Digitized Biocollections

65 sample types

**> 500,000
samples**

> 3,300 taxa

- Small mammals
- Fishes
- Ground beetles
- Mosquitos
- Ticks

- Soil
- Dust

many more



neon
Operated by Battelle

720.746.4844 | neonscience@battelleecology.org | neonscience.org