Christine Laney
Data Scientist
Boulder, CO
09/16/2022

Unique Opportunities for Ecological Data Management with NEON

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

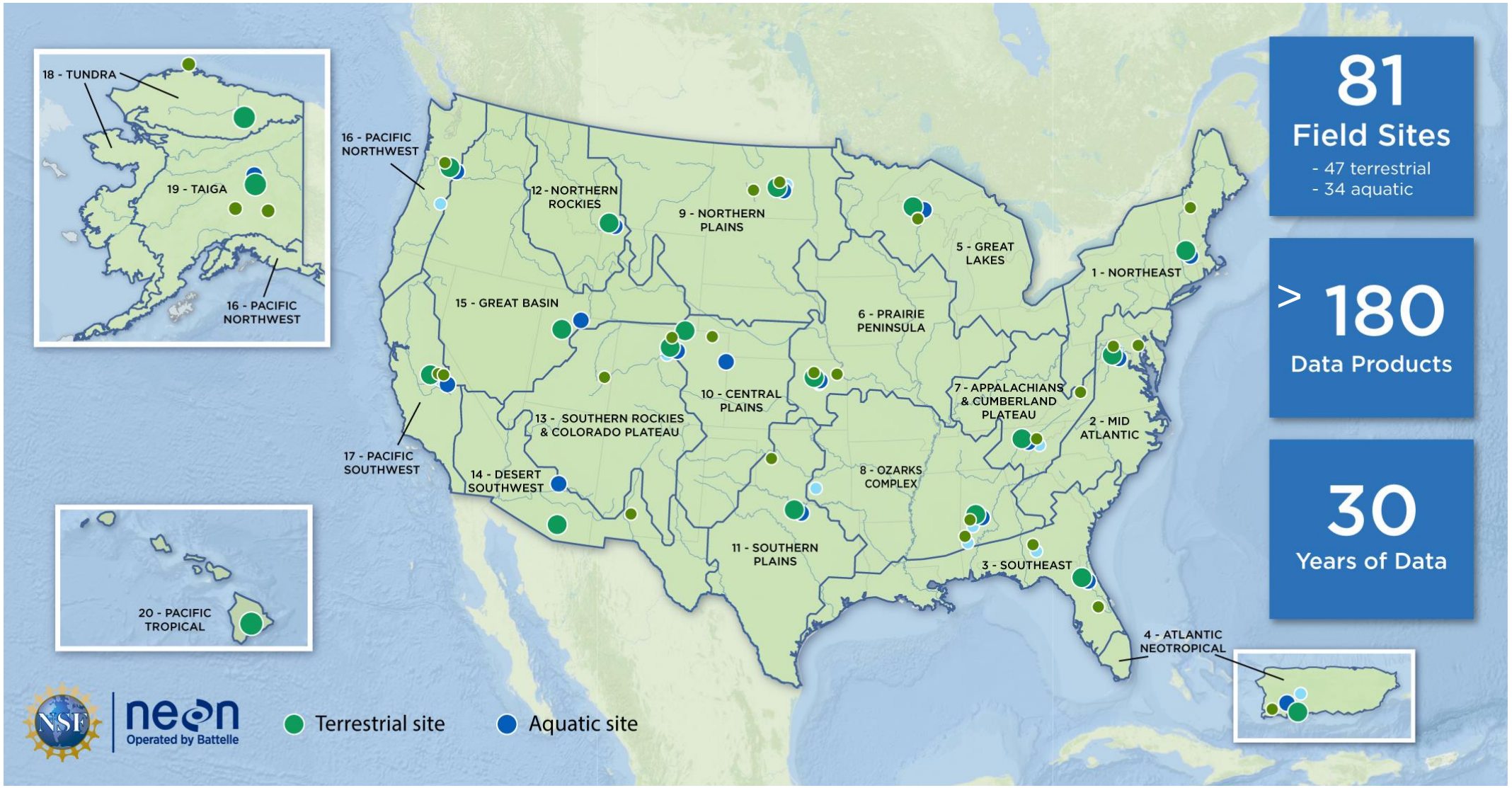
NEON Designed to Address Grand Challenges

[Credit: NASA]

"Recent changes in the climate are widespread, rapid, and intensifying, and unprecedented in thousands of years.

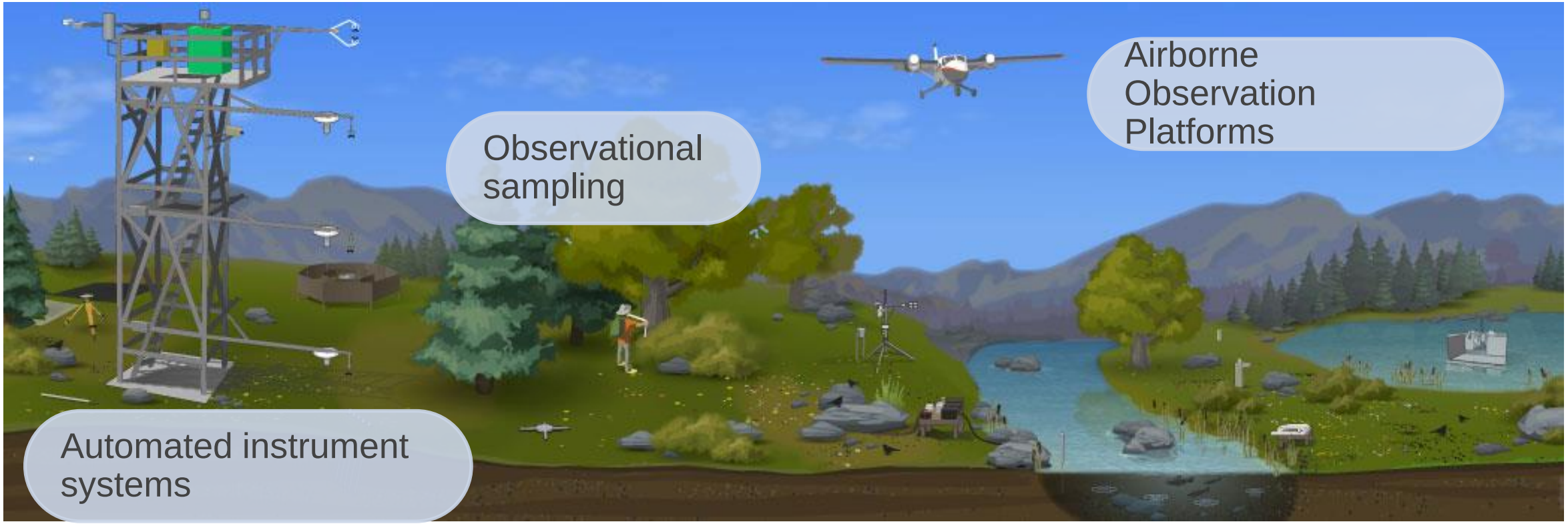
ipcc
INTERGOVERNMENTAL PANEL ON climate change
WMO UNEP

NEON: Data at a Continental Scale



NEON data collected via 3 data collection systems

Standardized, colocated methods across sites; 4 major data processing pipelines



5.6 B Records daily
400K Observations per year

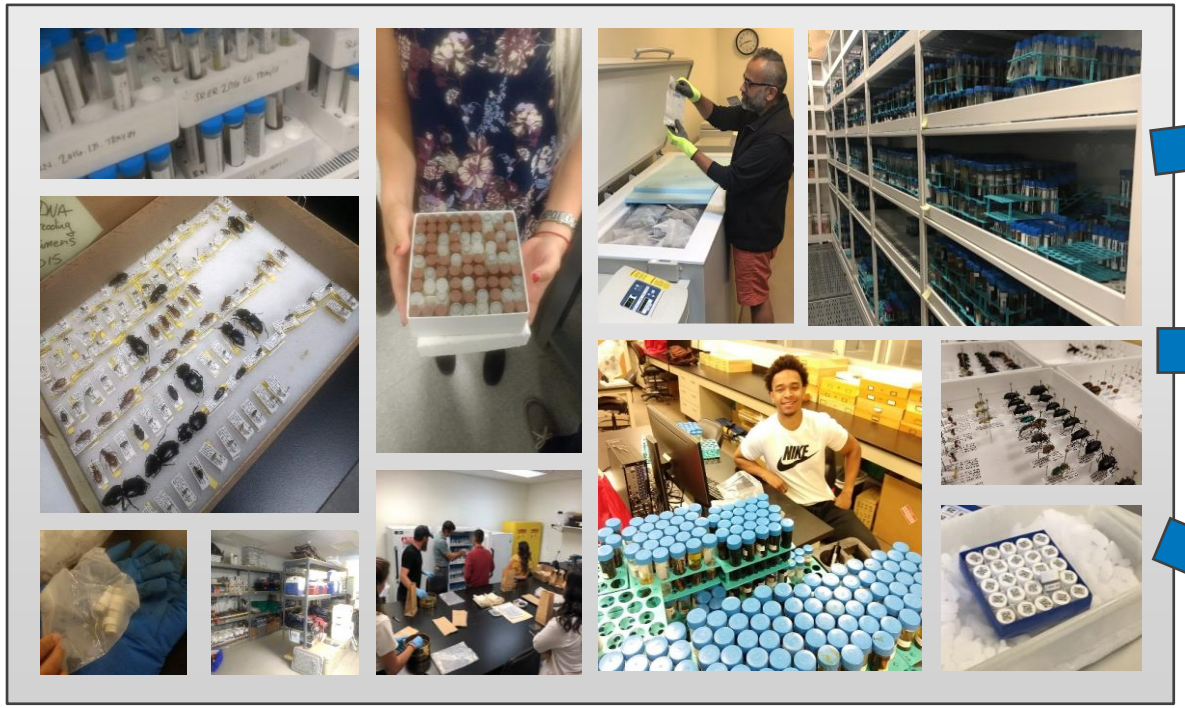
Physical Specimens and Samples Sent to NEON Biorepository for Loan and Research

65 sample types

- Small mammals
- Fishes
- Ground beetles
- Mosquitos
- Ticks
- Zooplankton
- Benthic macroinvertebrates
- Vascular plants, algae, bryophytes and lichens
- Soil microbes
- Soil
- Dust
- Wet deposition

…many more

100,000 specimens & samples/year



GBIF
Global Biodiversity Information Facility

iDigBio
Integrated Digitized Biocollections

= includes associated genomic data

Arizona State University

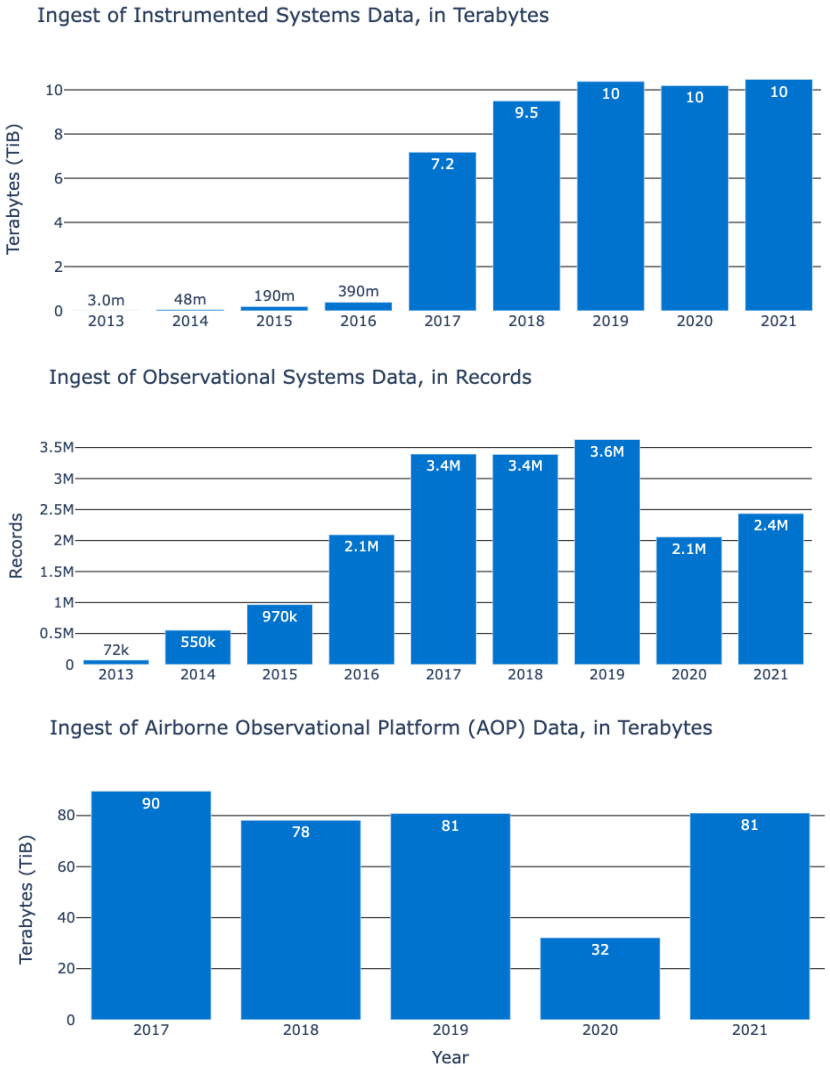
biorepo.neonscience.org

neon Operated by Battelle

Challenges of Data Collection, Storage and Processing

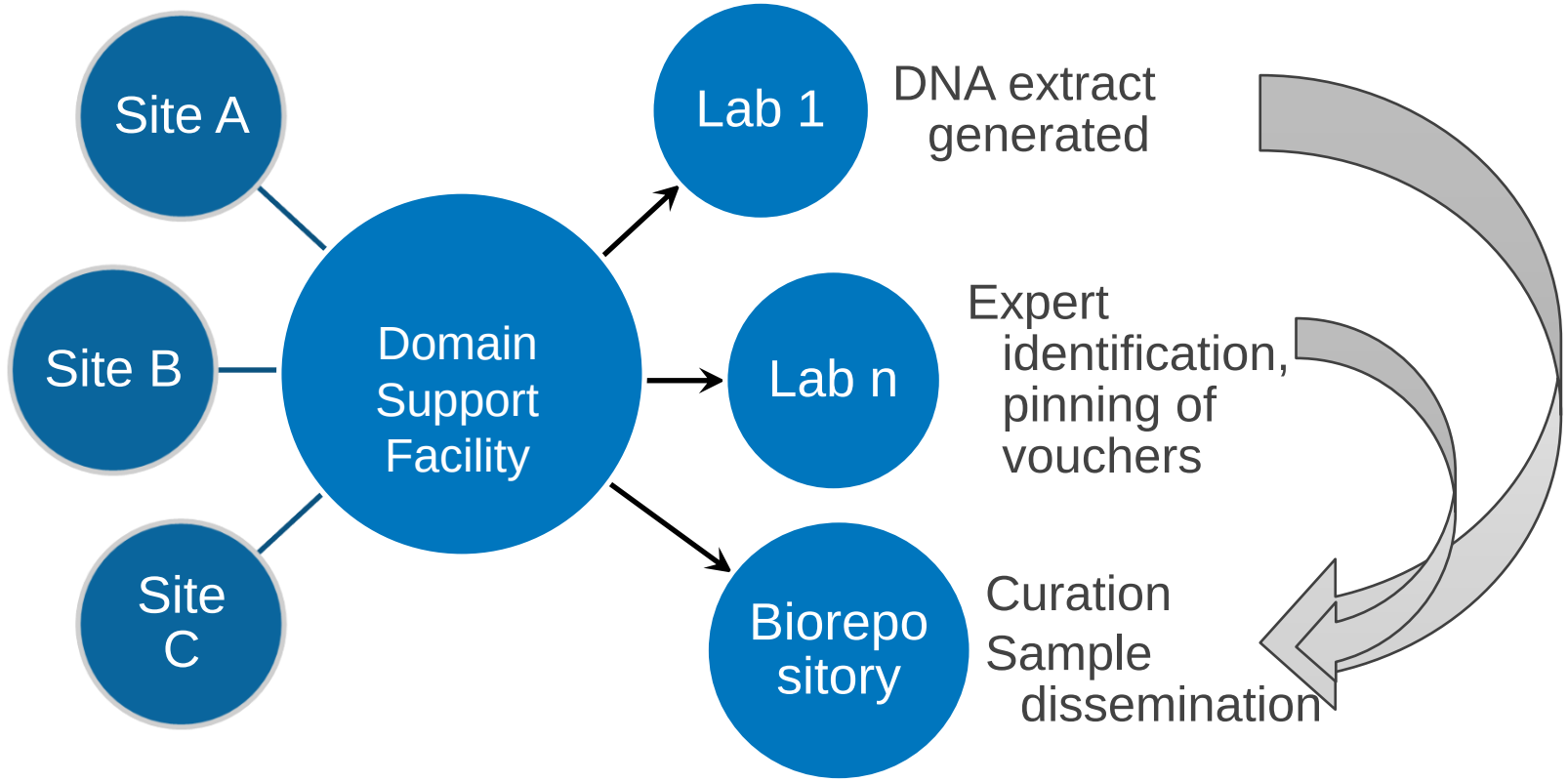
- High-volume remote sensing data - $$$$
- Data from automated instruments - >8,000 sensors of 64 types
- Data from the field
  - >150 permanent field staff, >200 seasonal across 81 sites
  - Coordination of training across dozens of standardized protocols → get standardized data
  - Biological and environmental samples – at 500K by 2023

Challenges of Data Volume and Processing

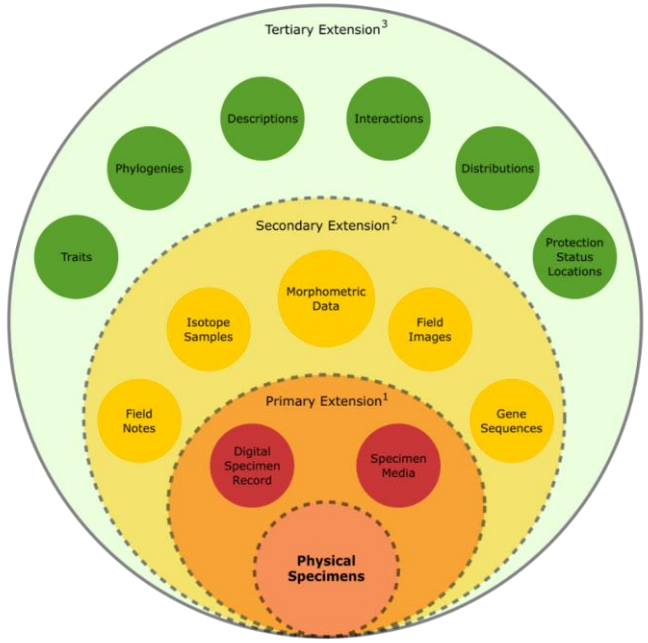


- Total ingested data volumes (approx.) through 2021:
  - Remote sensing: 362 TiB
  - Instrumented: 48 TiB
  - Observational: 18.6 M records
- Processed into 182 Data Products (more storage)
  - Often complex, multi-table, multi-parameter
  - Quality controlled snapshots are delivered annually as Releases
  - Documentation is necessary in multiple formats and types

Challenges of High Quality Samples and Data Generated Through 30+ Partner Labs

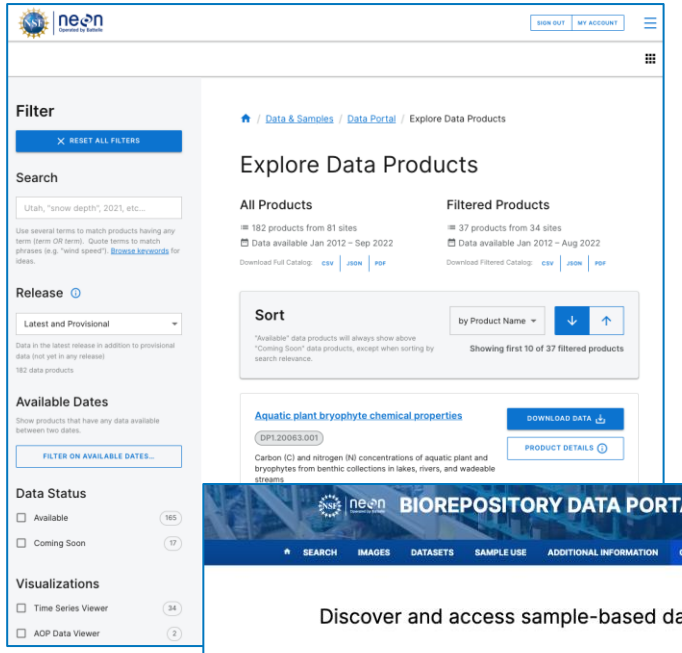


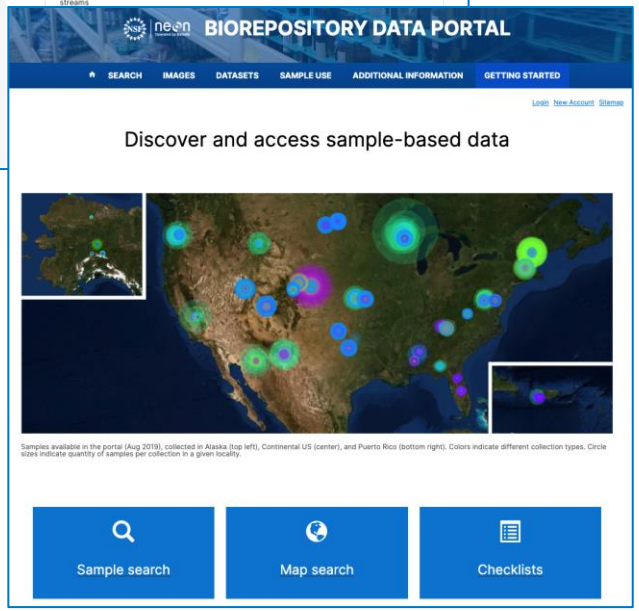
Extended Specimen



Challenges of Data Discovery and Dissemination

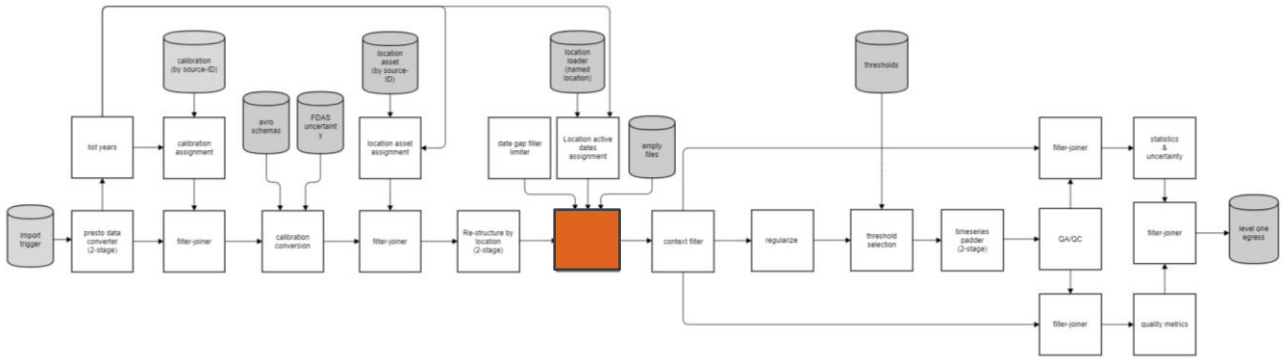
- Numerous points of discovery for NEON products & derivatives
  - NEON's Data Portal, API, Biorepository Portal
  - Third-party **aggregators** – e.g., GBIF (specimens, observations), DataONE
  - Third-party **repositories** – e.g., AmeriFlux, AERONET, BOLD, EDI
  - Third-party, short-term research projects – e.g., StreamPulse
- Researchers may have high expectations for accessibility and usability of highly heterogeneous data





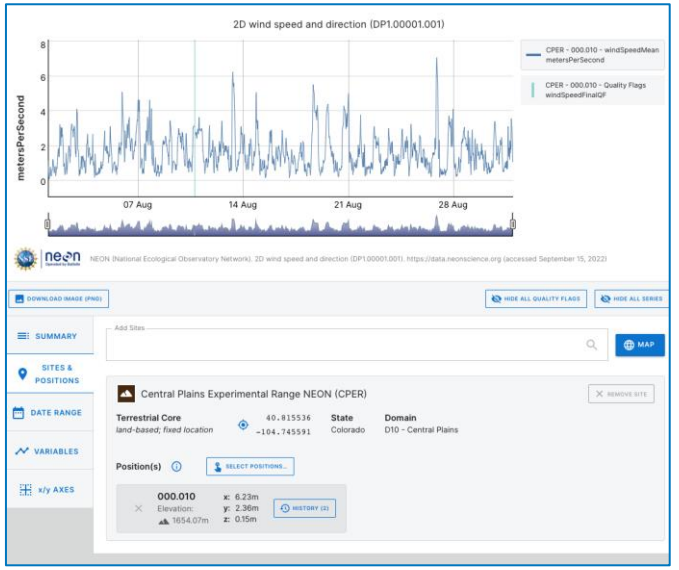
Practices to Support Data Collection Through Dissemination

- ***Data transmission & initial QA***
  - Streaming data from thousands of sensors to site hub to central servers
  - Mobile applications for field crews with built in data QA and streaming to cloud (Fulcrum)
- ***Storage***
  - Store all raw data but only provide instant access to processed data
  - Raw data in cold storage ($)
- ***Data monitoring & processing***
  - Open source, modular software e.g., Airflow, Kubernetes, Pachyderm
  - Foster tight community between software developers and scientific programmers



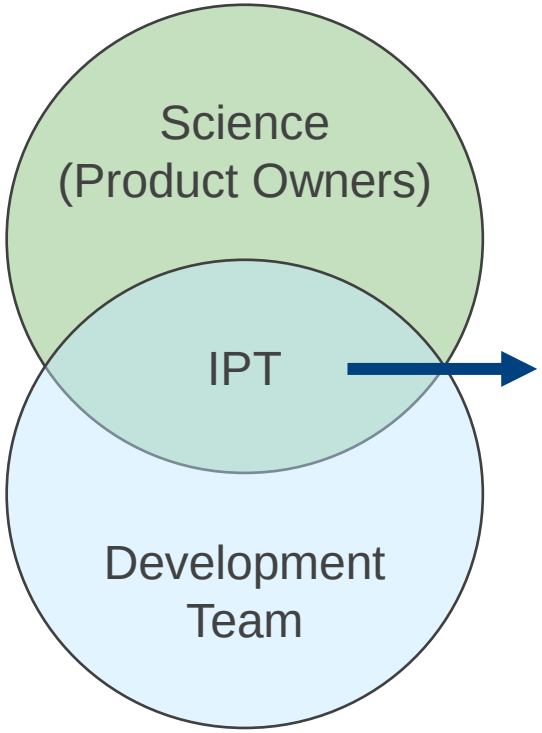
Practices to Support Data Collection Through Dissemination

- ***Sample storage & record annotation*** – Use and support improvements to software & machine-readable metadata standards, e.g., Symbiota, DarwinCore, EML, DataCite
- ***Linking data*** - Globally persistent unique identifiers (e.g., DOIs, IGSNs)
- ***Open, accessible data*** – CC0 license, no logins required
- ***Data exploration*** – Leveraging and improving open-source frameworks (e.g., React), libraries for websites (e.g., Material UI), and packages for data visualization (e.g., Dygraphs)



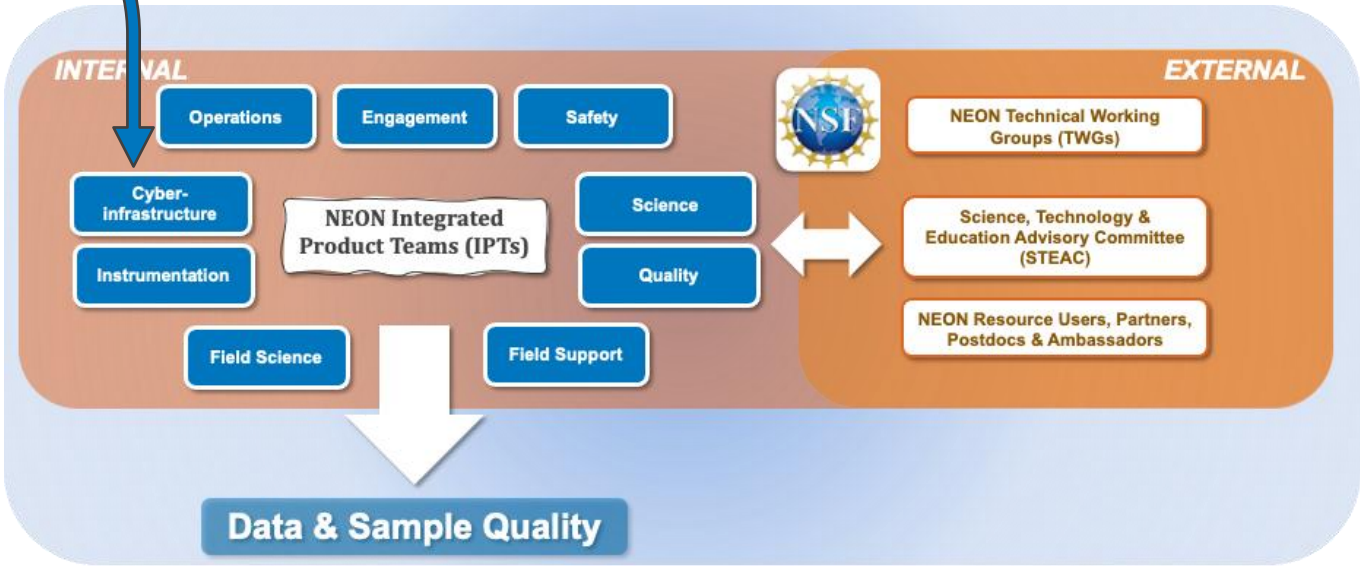
Prioritizing Data Management Best Practices Within a Limited Budget

Data Services Prioritization Integrated Product Team (IPT)



Development Roadmap

Feeds into the larger mechanism of inputs and feedbacks across the Observatory



Future Data Management and Data Dissemination Challenges

- NEON is managing increasingly large data and sample volumes by
  - Migrating data to Google Cloud with backups in Amazon Cloud
    - Cost-savings
    - Remote sensing data in Google Earth Engine
  - Expanding Biorepository space and Symbiota capabilities, as well as APIs that connect NEON & Biorepository
- Continue to decrease the time from collection to dissemination while keeping quality high by continuing to
  - Work with the research community to find where these cases are needed
  - Research and add to new scientific workflow applications
  - Build trust between traditional software developers and scientists who program

Future Data Management and Data Dissemination Challenges

- Retaining data quality and citability as data are disseminated, used, and their derivatives are posted
  - DOIs, metadata on annual data Releases
  - Education on data use and citation (via NEON's Data Skills program)
- Improving data discovery and dissemination through NEON's outlets as well as through third parties.
  - Improve metadata so that it is more easily harvestable by aggregators and repositories
  - Continue work with communities on what they need to do research with NEON data
  - Keep up to date with and help improve standards (especially in domains where few exist)

A Few Continuing Challenges

- Data volumes
- Data granularity and dynamic metadata
- Linkages between samples, their subsamples, and their data records over time
- Community contributions to major processing workflows
- Data citation/attribution with derivatives
- Data harmonization with other networks

Operated by Battelle

720.746.4844 | neonscience@battelleecology.org | neonscience.org