User-defined Cyberinfrastructure
Company Background

- Founded in 2006
- 20 employees
- Offices in AK, OR and RI
- Mission driven: Build capabilities which accelerate the synthesis and re-use of earth science data
- Federal Partners: NSF, NOAA, USGS, BOEM, NASA, FWS, DARPA, ONR
- Other Partners: MBARI, Shell, UAF, UAA, Carnival, Smartfin, Leidos
The National Science Foundation defines cyberinfrastructure as:

In scientific usage, cyberinfrastructure is a technological and sociological solution to the problem of efficiently connecting laboratories, data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.

- Community developed software, standards and protocols
- Scalable compute and storage infrastructure (HPC)
- Human capacity - data scientists, data librarians, data coordinators, software engineers….
Data Management Lifecycle

Supporting the entire data lifecycle

DATA COLLECTION & QUALITY CONTROL
Scientists or Ingestion

REUSE & TRANSFORMATION
Jupyter Notebook & data analyses

ACCESS & DISCOVERY
Data portals, search catalogs, and services

ARCHIVE & PRESERVATION
Repository submission pathway

STORAGE
Databases, Messages, Files, etc.

DESCRIPTION
Metadata, Apply standards
Supporting the entire data lifecycle

DATA COLLECTION & QUALITY CONTROL
Scientists or Ingestion

REUSE & TRANSFORMATION
Jupyter Notebook & data analyses

ACCESS & DISCOVERY
Data portals, search catalogs, and services

DESCRIPTION
Metadata, Apply standards

STORAGE
Databases, Messages, Files, etc.

ARCHIVE & PRESERVATION
Repository submission pathway

DATA COLLECTION & QUALITY CONTROL
Scientists or Ingestion

REUSE & TRANSFORMATION
Jupyter Notebook & data analyses

ACCESS & DISCOVERY
Data portals, search catalogs, and services

DESCRIPTION
Metadata, Apply standards

STORAGE
Databases, Messages, Files, etc.

ARCHIVE & PRESERVATION
Repository submission pathway
## Data Categories

### Biodiversity
- count, richness, diversity indices

### Platforms
- moorings, shore stations

### Products
- skill assessment, shoreline change, etc.

### Grids
- models, satellite, radar

### GIS
- Habitat types, bathymetry, fishing zones, etc.

### Moving Platforms
- auv, wave glider, argo, particle trajectory models

### Unstructured Data
- PDF, DOC, WAV, PNG, MPG

---

**Axiom**

DATA SCIENCE
Data Pipelines

- Develop a set of **data pipelines**
- Keep a **streaming first** mentality
- Support multiple **integration points** (entry and exit)
- **Targets** can be anything (stream, file, database entry, etc.)

```
Raw
  Binary Stream
       ↓
  Parsed
      ASCII Stream
          ↓
Converted
         Database
      ↑
Instrument Stream
  Recovered Analysis
       ↑
Quality Control

```

7
Data Pipelines

- Connect pipelines together with a messaging system
Data Pipelines - Enabling Discovery

- Catalog pipelines together with metadata updates

```
New raw data

Raw
Binary Stream

New parsed data

Parsed
ASCII Stream

Catalog

New database data

Converted
Database
```
Data Pipelines

- Build in data **update and invalidation** at integration points
Data Pipelines

- Design pipelines and targets to meet **user needs**
- Develop targets around **community standards for interoperability**

![Data Pipelines Diagram]

- **Raw**
  - Binary Stream
- **Parsed**
  - ASCII Stream
- **Converted**
  - Database

Flow:
- **netCDF Export** to **File**
- **File** to **Changed file**
- **Changed file** to **New database data**
- **New database data** to **Update ERDDAP**
- **Update ERDDAP** to **Converted**
- **Converted** to **File**
- **File** to **netCDF Export**
Data Pipelines

- Well defined integration points helps migrate and transition systems
Example Pipelines - AIS

AIS Data

Raw AIS Messages → Parsed and cleaned Messages → Daily Vessel Voyages → Filtered Voyages by time frame, region, and ship type → Voyages GeoJSON / Shapefile

Shapefiles

Shapefile → PostGIS Table → GeoServer → OGC Services

NetCDF Grids

netCDF Grid → THREDDS DAP → ncWMS2 WMS

Catalog

Vessel Traffic Heatmaps

Heatmaps GeoTIFF / NetCDF
Example Pipelines - Sensors

metadata data
Eye Candy

Catalog
- General filtering, tagging, full-text search: https://portal.secoora.org/#search
- Regional grouping by parameter: https://portal.secoora.org/#metadata/7/sensor/list
- Global grouping by parameter: https://sensors.ioos.us/#metadata/14/sensor/list
- Source grouping (CO-OPS): https://portal.aoos.org/?sensor_version=v2#metadata/9/sensor_source/list
- Filter by type (gliders): https://data.cencoos.org/#search?type_group=all&tag%7Ctag=gliders&page=1

Data Inventories
- Provider: https://portal.secoora.org/?sensor_version=v2#metadata/144/sensor_source/inventory
- Station: https://portal.secoora.org/?sensor_version=v2#metadata/60387/station/inventory

Visualization
- Station page: https://sensors.ioos.us/#metadata/57301/station
- ~18 year profiling time-series: https://sensors.ioos.us/#metadata/75365/station
- Sensor QC (with config) https://portal.aoos.org/?&sensor_version=v2#metadata/75367/station/25/sensor
- Provider supplied QC: https://sensors.ioos.us/?sensor_version=v2#metadata/63965/station/36/sensor
- West-coast Glider deployment: https://data.cencoos.org/#platform/4e3d20cf-d3fe-4646-b1bb-b89f38d9597d
- Research Cruise (ECOA1): https://portal.secoora.org/#platform/81ad3445-245c-4b9d-b758-0c4595f1f5d0
- Hurricane Tracker: https://hurricane.portal.secoora.org/#map-view/hurricane-florence-2018/data/2018-09-14T17:00:00Z

Sites
- AIS Pipeline - https://ais.axds.co
- IOOS Sensor Map - https://sensors.ioos.us
- Northern Gulf of Alaska NSF LTER Site - https://nga.lternet.edu
Remote Compute

- Not all data can be analyzed from a web browser
- Data volumes are growing, especially at DACs
  - NOAA Big Data Project, GOES-R, etc.
  - Availability of audio/video
- Downloading all data no longer viable

- What are the options?
  - Produce analyses and products users request
  - Give users fast access to the data without downloading
Remote Compute - Examples

WebCAT - https://secoora.org/webcat/

- Private/public partnership between SECOORA and Surfline
- 8 streaming video cameras
- 1 year of raw video data hosted at SECOORA
- Access to raw video streams was “too much”

AIS Histograms - Notebook

- Analyze the Daily Vessel Voyages (CSV) for a specific region
- Access to a Dask cluster

Passive Acoustic Spectrograms - Notebook

- Hydrophones in Bering Sea from 2009-2013
- Raw access to .wav files
Extra Slides
Results
Partners
High Performance Computing (HPC)

- Ocean data is BIG DATA (models, sensors, hydroacoustics, satellite imagery, etc...)
- Requires scalable compute and storage infrastructure

- ~5,000 processor cores
- ~1.5 petabytes of functional storage/5 petabytes of actual storage (~1,500 hard drives)
- Level 2 Fat Tree Infiniband Network, 40 Gb/Sec node to node). 240 Gb/Sec cluster to cluster