



OCEAN
OBSERVATORIES
INITIATIVE

OCEANOBSERVATORIES.ORG



Preparing an Ocean of Data for AI

Jeffrey Glatstein

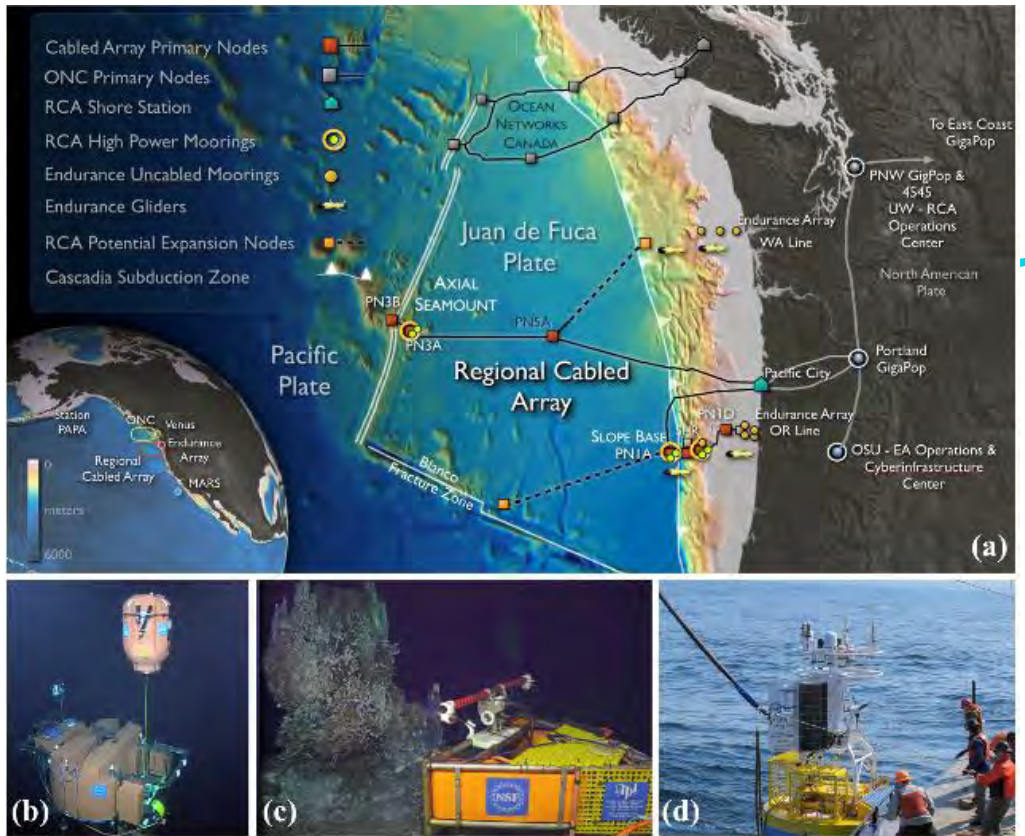
Senior Manager of Cyberinfrastructure

January 12, 2026

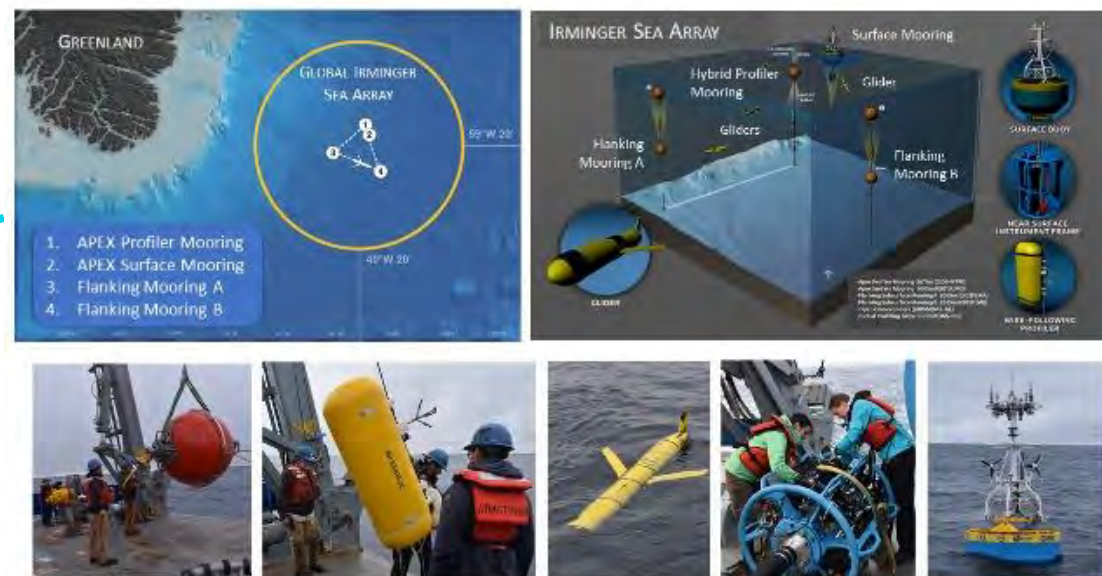


- Real-time data from more than 900 instruments on 80 platforms at 5 arrays.
- The arrays are designed to conduct science at global, coastal & regional scales.

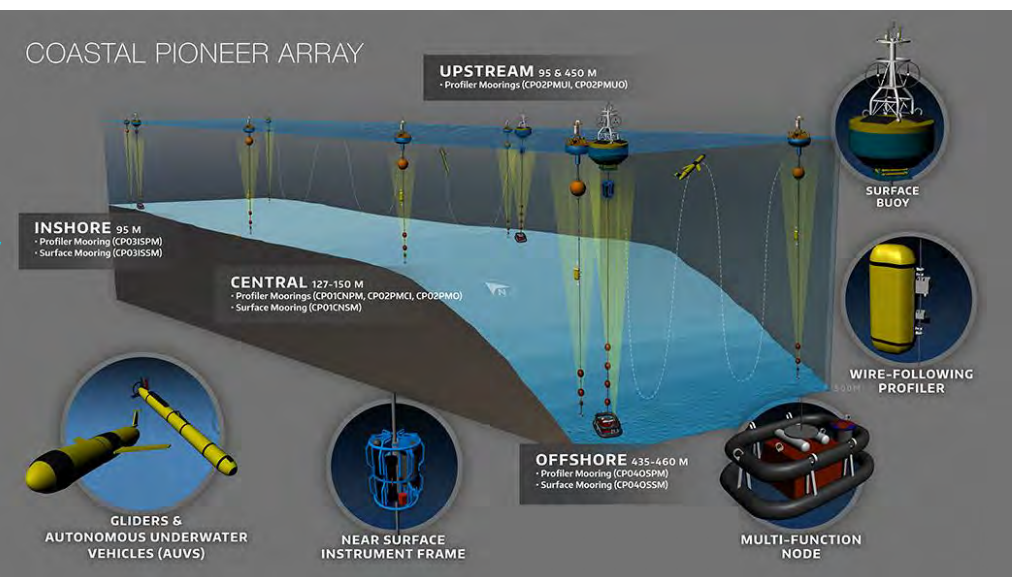
Regional Cabled Array & Coastal Endurance Array



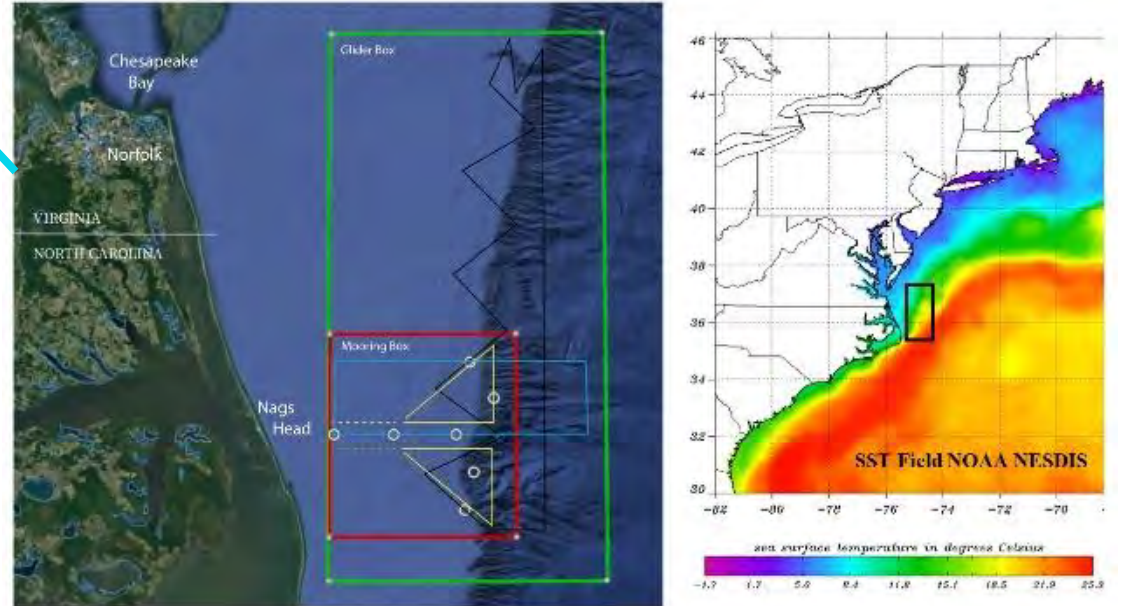
Global Irminger Sea



Coastal Pioneer NES



Coastal Pioneer MAB



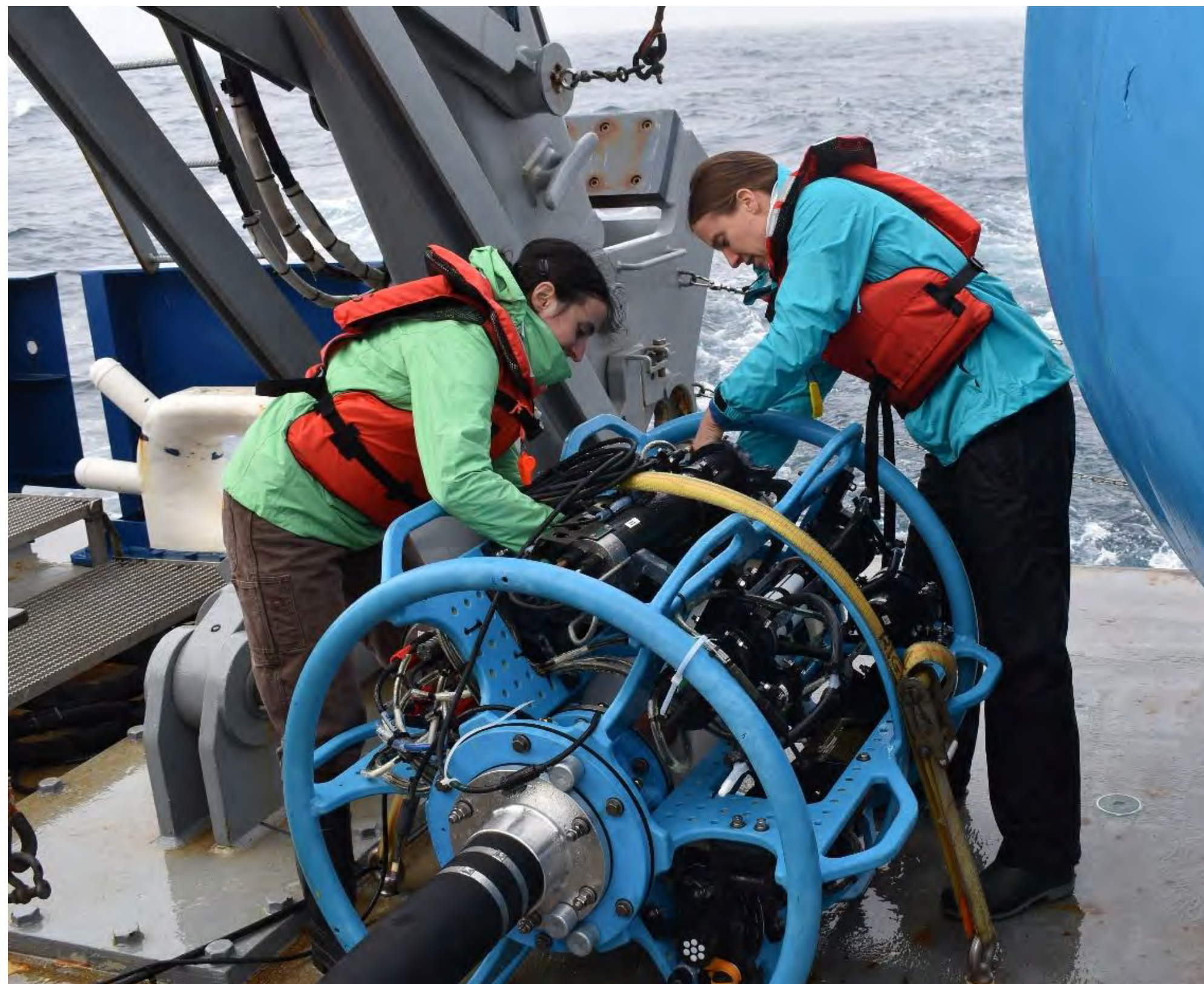
- Served by a united Cyberinfrastructure
- Data freely available online
- 30-year lifetime starting from 2016.
- OOI is a UN Decade Action attached to OASIS program
- OOI is endorsed by SOLAS



Our Mission

The OOI does not conduct research, the OOI provides research quality data to the user community (scientists, educators, and the general public).

<https://oceanobservatories.org/how-to-access-data/>



Data Collected

Surface Variables

- Pressure
- Radiative Fluxes
- Temperature
- Humidity
- Precipitation
- Moisture/Evaporation
- Vector Wind
- DC Stress & Buoyancy Flux

Oceanographic

- Net Surface Heat Flux
- Temperature Profiles
- Salinity Profiles
- Currents Profiles
- Sea Level

Biogeochemical Sensors (at multiple depths)

- pH
- pCO₂
- Oxygen
- Plankton & zooplankton
- Nitrate
- Chlorophyll a

and we monitor an active volcano!



OOI Data By The Numbers

- Data Collected (900+ instruments)
 - 200 billion rows of numerical data
 - 2.8 PB of raw data
 - 15,000 hours of HD video
 - 3.00 million digital still pictures
 - 400K hours of Hydrophone acoustic sampling
- Data Delivered Last Quarter
 - 35TB of raw data for 12 million requests
 - 16TB of calculated data for 62 million requests



Popular AI Bot – How can I make scientific data AI ready?



Making scientific data AI-ready involves a systematic process of **collecting, cleaning, structuring, and enriching** data with high-quality metadata, focusing on a specific use case to ensure relevance, and establishing robust governance for ongoing quality and accessibility. Key steps include standardizing formats, handling missing values, adding context (like [labels](#) and [tags](#)), ensuring data lineage, and implementing security, all within a scalable architecture that supports discovery and access.

List of Six categories with fifteen specific actions.

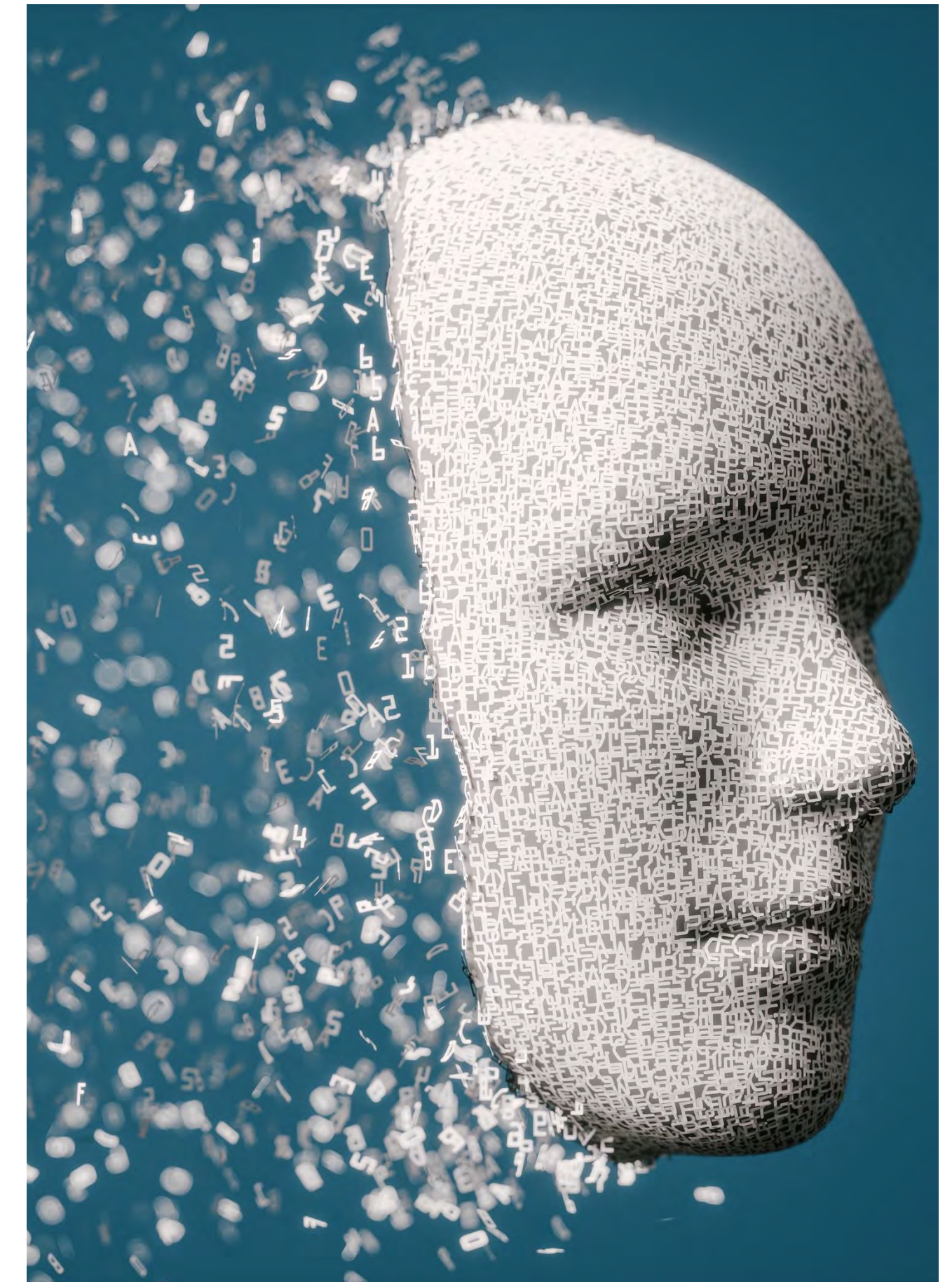
- **Define the use case**: Start with a specific science question.
 1. **Data Collection**: Audit Sources, Standardize formats, Use high quality sensors
 2. **Data Cleaning**: Handle missing values, Remove duplicates, Detect anomalies
 3. **Address Batch Effects**: Correct variations (e.g. calibration or parameter calc changes)
 4. **Metadata**: Add context, Catalog and tag, data lineage
 5. **Governance and Ethics**: Establish policies, Ensure security, Monitor and version
 6. **Accessibility and architecture**: Centralize (yes plus GPU access), Enable APIs (JupyterHub)

Key: **Operational** **More required** **Partially implemented** **Not met**



AI Implementation Status

- Machine Learning (Supervised and Unsupervised)
 - Data quality checks of both photos and videos – successful POC
 - Content categorization – successful POC
 - Object identification – successful POC and soon to be production implementation
- Large Language Model (Conversational)
 - Translate code into usable documentation – early stage development
 - Translate documentation into chat bot – early stage development
- Observational data-centric AI (Natural language/Agentic)
 - Conversion of NetCDF file farm to data lake – early stage POC
 - Analysis of data structure for AI consumption using analysis tool (AIDRIN)





OCEAN
OBSERVATORIES
INITIATIVE



OCEANOBSERVATORIES.ORG



Thank you



OCEAN
OBSERVATORIES
INITIATIVE