

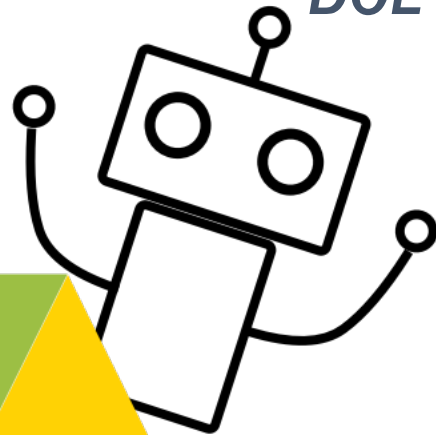
# Towards an Integrated Biological and Environmental Data Infrastructure

*Kjiersten Fagnan, CIO*

*DOE Joint Genome Institute*

*June 1, 2025*

**BERtron**



# The DOE Joint Genome Institute at a glance



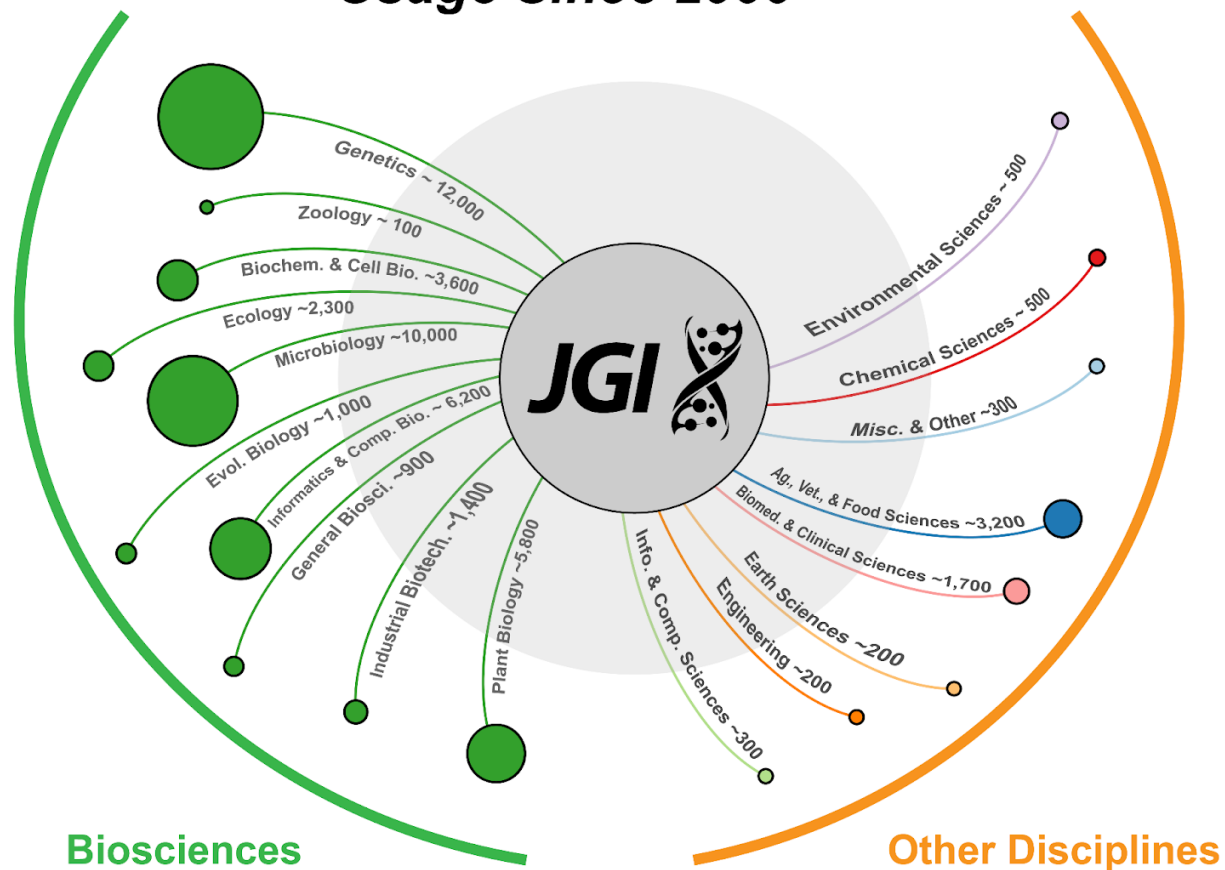
## JGI MISSION:

To provide the global research community with free access to the most advanced integrative genome science capabilities in support of the DOE energy & environmental research mission

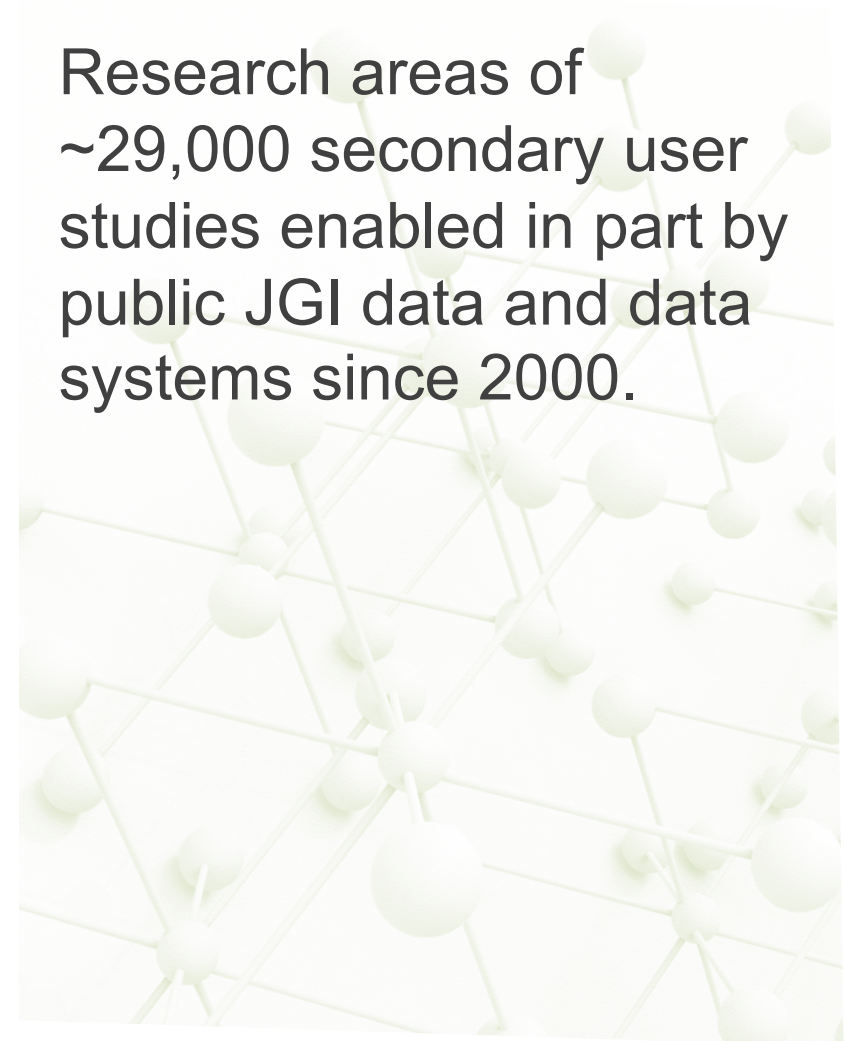
## U.S. Department of Energy Office of Science User Facility

- JGI established in 1997, User facility from 2004
- Located at Lawrence Berkeley National Laboratory
- ~285 staff; ~\$80M annual funding
- 2,038 Global Primary Users in FY20; >10,000 Data Users

## Known Data & System Usage Since 2000



Research areas of ~29,000 secondary user studies enabled in part by public JGI data and data systems since 2000.



## AlphaFold Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search

Examples:

[See search help](#) [Go to online course](#) [See our updates – March 2025](#)

## Evo 2: DNA Foundation Model

Evo 2 is a genomic foundation model capable of generalist prediction and design tasks across DNA, RNA, and proteins. It uses a frontier deep learning architecture to enable modeling of biological sequences at single-nucleotide resolution with near-linear scaling of compute and memory relative to context length. Evo 2 is trained with 40 billion parameters and 1 megabase context length on over 9 trillion nucleotides of diverse eukaryotic and prokaryotic genomes.

[Evo 2 Preprint](#)

[Evo 1 Published November 2024, \*Science\*](#)

[Evo Designer](#)

[Evo Mechanistic Interpretability Visualizer](#)

[Github](#)

[Evo 2 on Github](#)

[Evo 1 on Github](#)

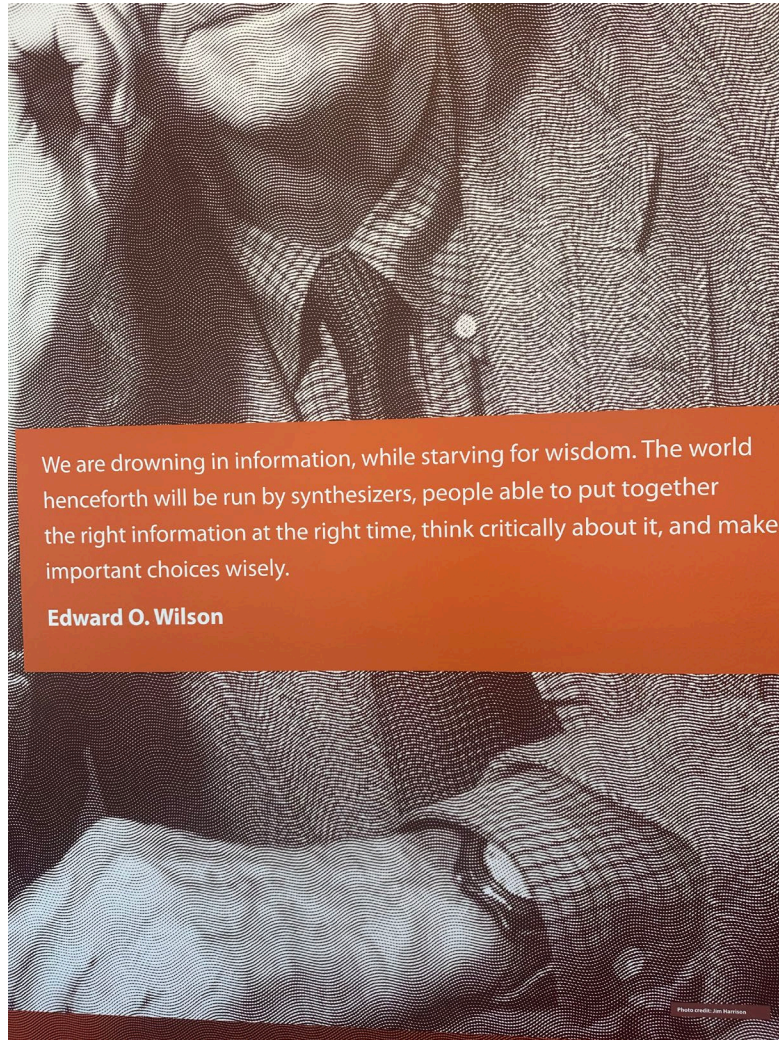
February 27, 2025 Release Product

## Introducing GPT-4.5

A research preview of our strongest GPT model.  
Available to Pro users and developers worldwide.

[Try in ChatGPT ↗](#)

# The State of our Data Systems



We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely.

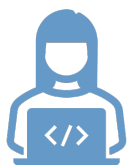
Edward O. Wilson

“We are drowning in information, while starving for wisdom.



The world henceforth will be run by synthesizers, **people able to put together the right information, at the right time, think critically about it and make important choices wisely.**”

*E. O. Wilson (2014). "Consilience: The Unity of Knowledge", p.399, Vintage*

# 5 BER Resources Collaborating to Prototype a Unified Data Access Layer




Unifying Access Layer Components  
BERtron - Global Search supported by common APIs - **find and reuse** data  
Data Transfer Service - **maintain provenance, propagate credit**



Sample metadata,  
standardized data  
products



Biogeochemical  
measurements,  
sensor data



Proteomics,  
metabolomics, imaging

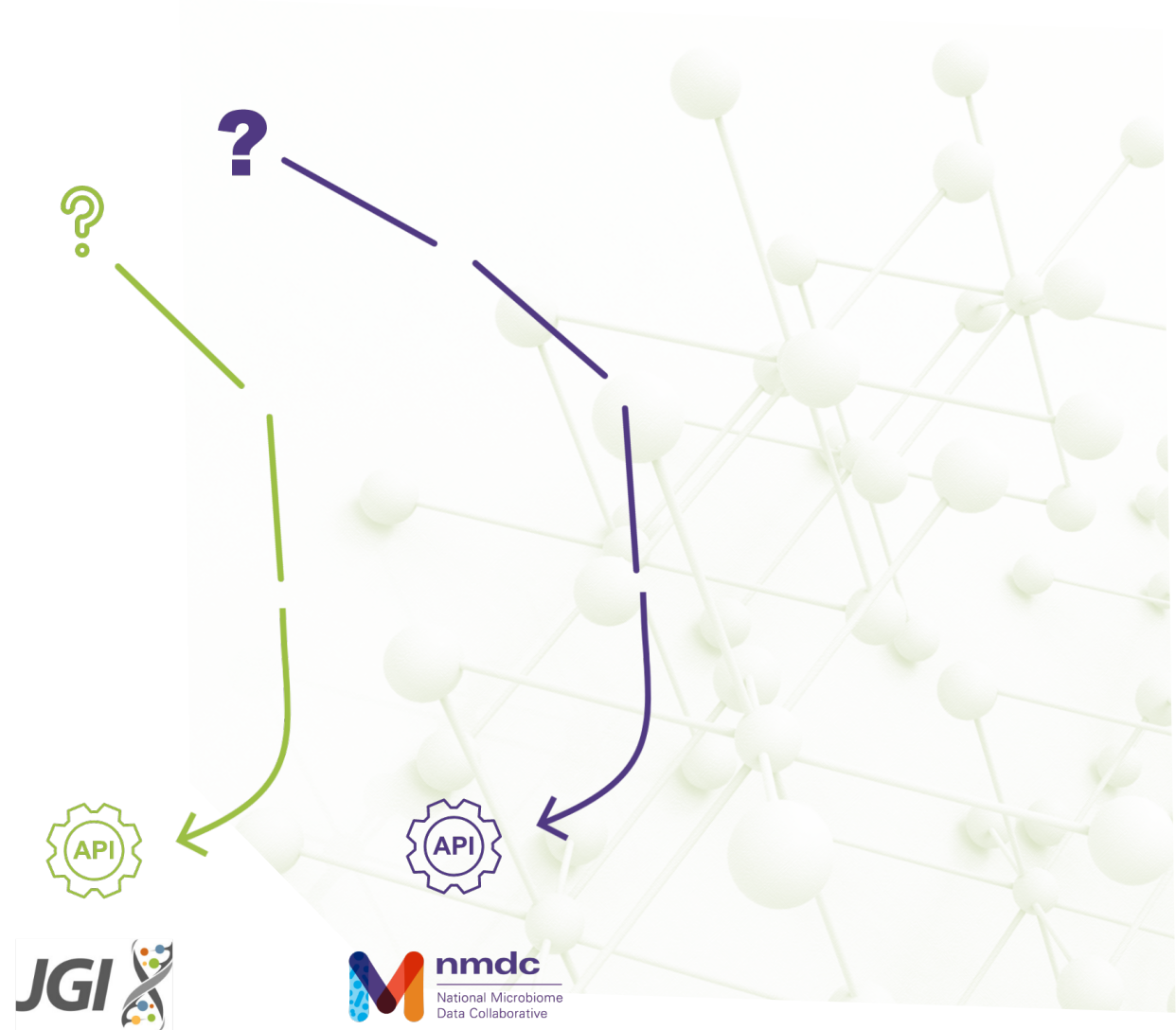
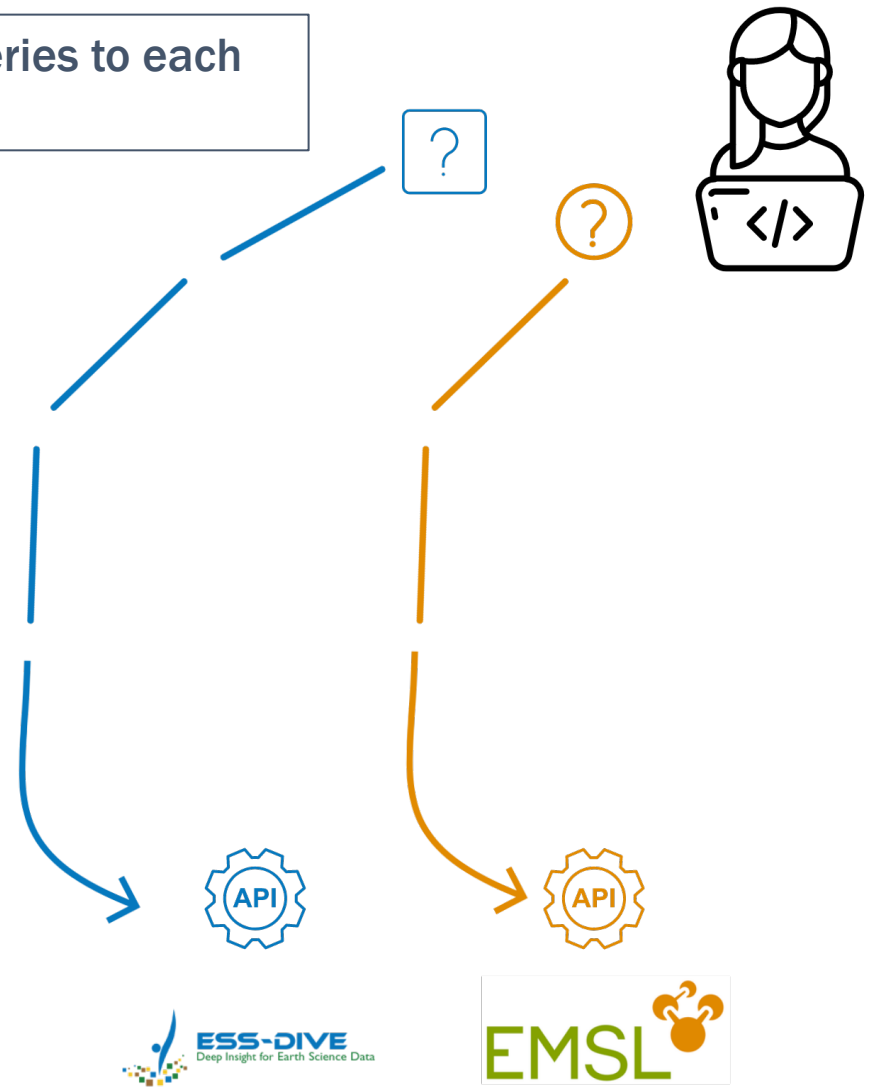


(meta)genomics,  
-transcriptomics,  
metadata

# Current State of Data Access and Exploration



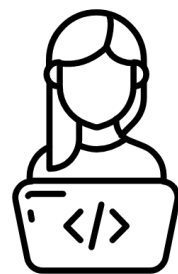
Bespoke queries to each resource



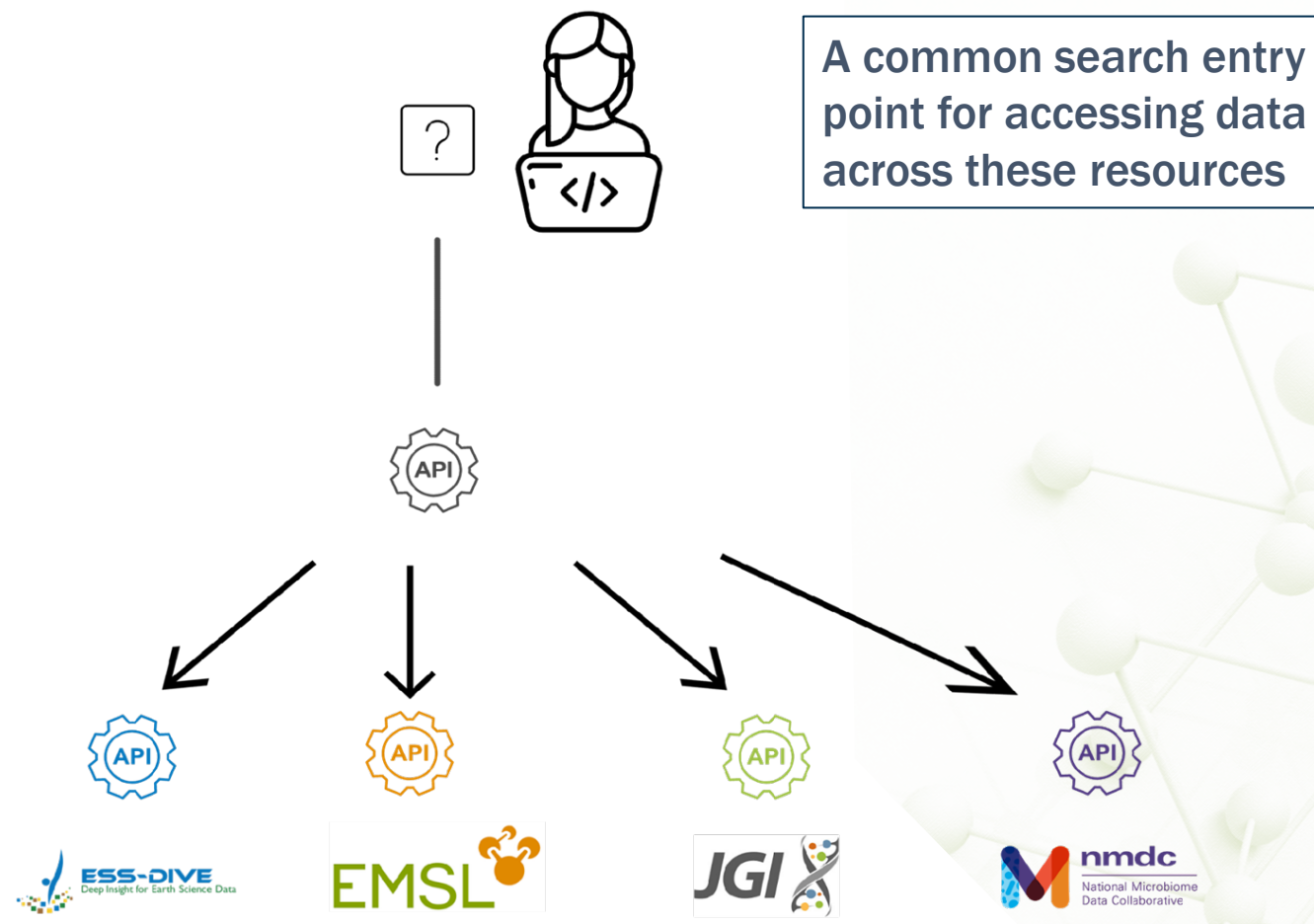
# Yielding bespoke responses that need to be harmonized



# combines the data and is able to move on



# Initial Project Goal – Global Search



# User Research – 2 rounds



## Emphasis on Search, Data Management, and Sharing Practices

- Data users (Scientists, facility users, research community) are downloading files from various locations, how do they search?
- How do they manage what they download?
- How are results shared?



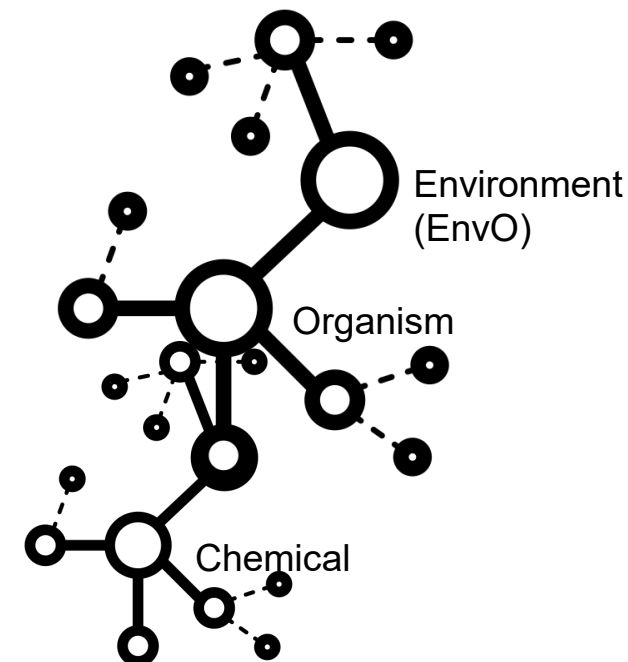
- **Data Users/Scientists generating more data and spending increasing resources on data management and building their own infrastructures**
  - # of disconnected data resources is growing, not slowing down
- **Data Users/Scientists are resourcing data management**
  - Paying for software
  - Allocating 5-10% of grants to cover staff
- **Strong desire for automation of data management activities**
  - Spend time on science, not data management
  - Applying data models is tedious
- **There are gaps in the data repository infrastructure**
  - No place for analysis-derived data products (e.g., updated genome annotations, assemblies)



# Search != Data Integration



- Search is how we learn one another's systems
- Search yields insights about data structures
- Search will improve the documentation for our respective systems
- Search is necessary (helpful), but not sufficient



**Unifying Model**  
Connect processes,  
ontologies, observations,  
etc across BER Data

# Search != Data Integration



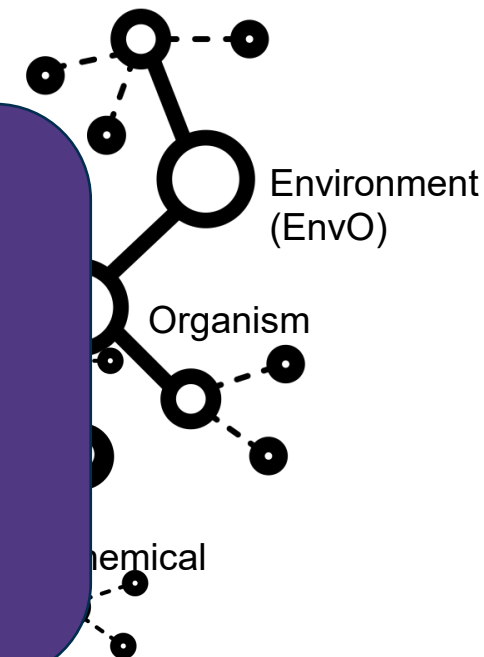
- Search is how we learn one another's systems

- Search

- Search  
respec

- Search

Metadata is a huge issue and **data harmonization** is required to support effective data integration



**Unifying Model**  
Connect processes, ontologies, observations, etc across BER Data

# Codeathon – Data Harmonization

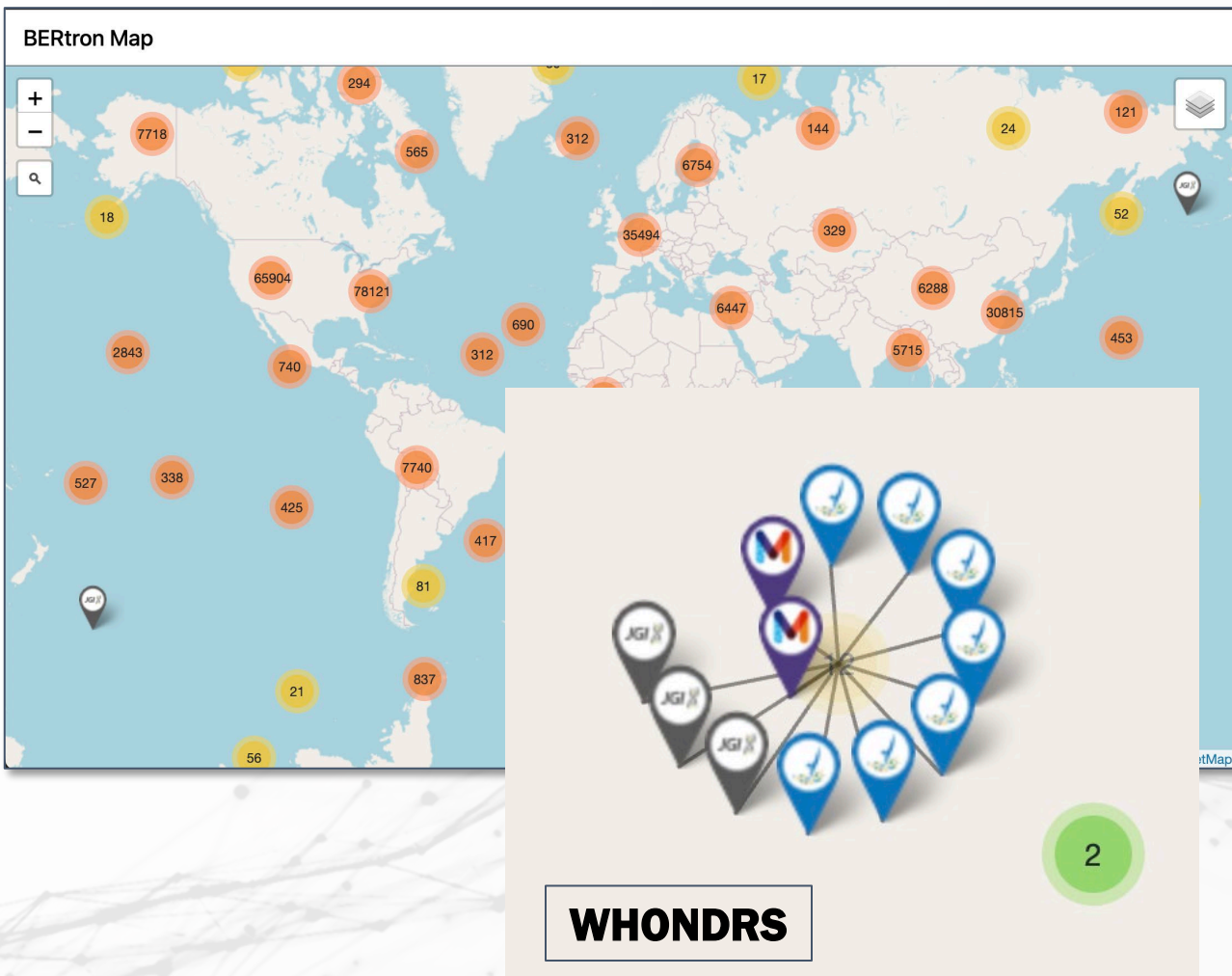


- Created the foundation for a shared data model
  - <https://ber-data.github.io/bertron-schema/>

A1	Class	BERtron Field	BERtron Field Description	Programmatic data type
1	Class	BERtron Field	BERtron Field Description	Programmatic data type
2	Dataset	Data source		Enum
3	Dataset & GeoData	Tags	from controlled vocab. max 10 tags per dataset. pick the most useful (that you think). This would give flexibility to the model.	Enum
4	Dataset & GeoData	URL	page for the entity at the BER data source	<a href="#">Uri</a>
5	Dataset & GeoData	CURIE	Compact URI (if available)	<a href="#">Curie</a>
6	Dataset & GeoData	IDs	resolvable permanent identifiers, including the same entity at other data sources	String
7	Dataset & GeoData	Name(s)	identifiers that cannot be resolved or colloquial names/synonyms	String
8	Dataset & GeoData	Description		String
9	Dataset & GeoData	Date added/updated	When the data was added or updated	<a href="#">Datetime</a>
10	GeoData	lat	latitude	<a href="#">DecimalDegree</a>
11		long	longitude	<a href="#">DecimalDegree</a>
12	GeoData	descriptor	Describes the type of data being stored	String



# Codeathon – Virtual March 6-7

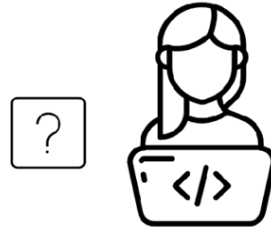
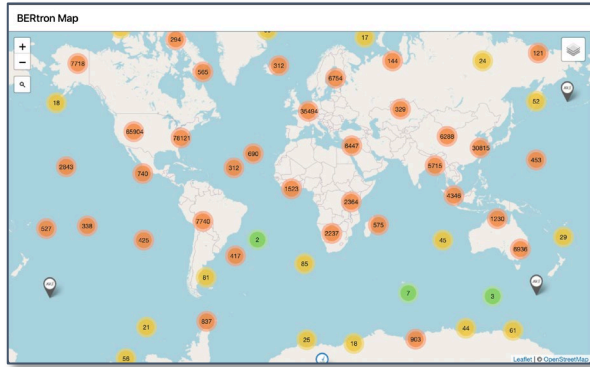


- The team developed a shared view of data based on lat/long coordinates
- Each region allows a user to zoom in to see the data from the different contributors
- Main Repository:  
<https://github.com/ber-data>
- Map:  
<https://ber-data.github.io/bertron>

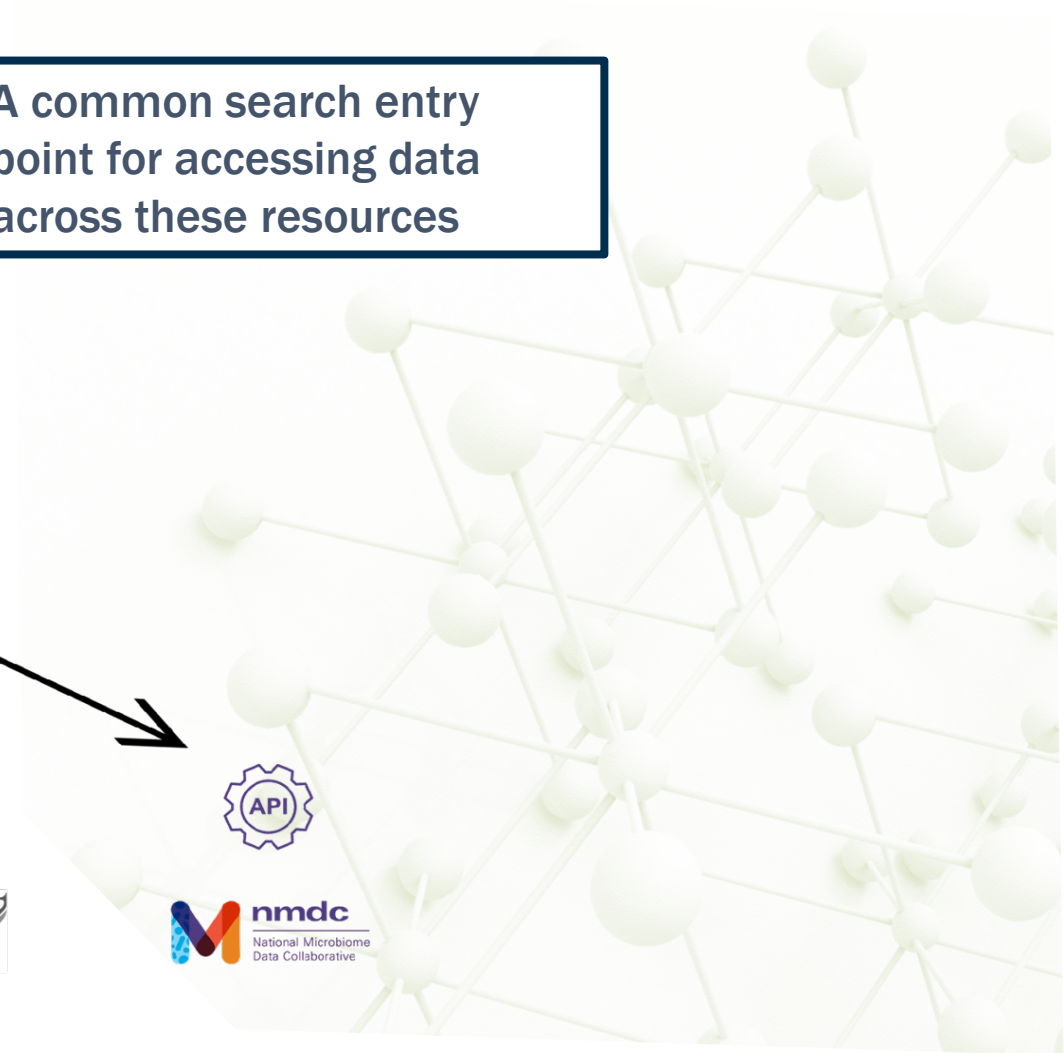
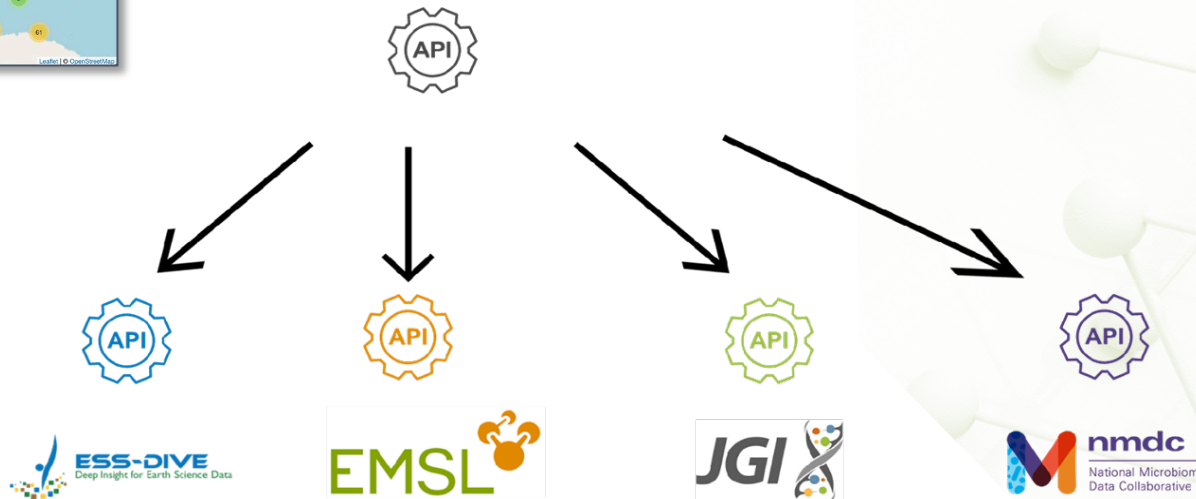
# Initial Project Goal – Global Search



Progress:



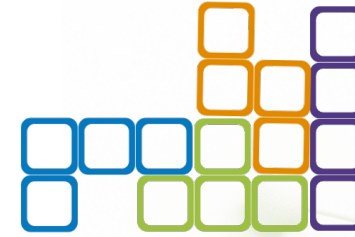
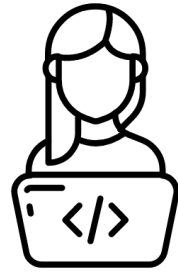
A common search entry point for accessing data across these resources



# This will still yield complex responses...

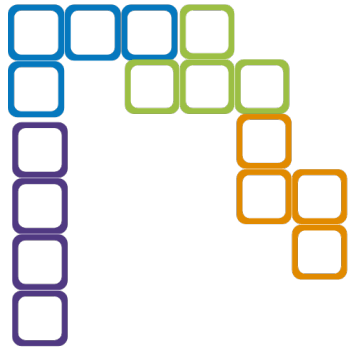
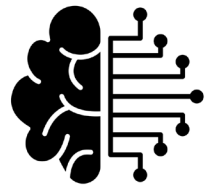


# ...that the user will need to harmonize.

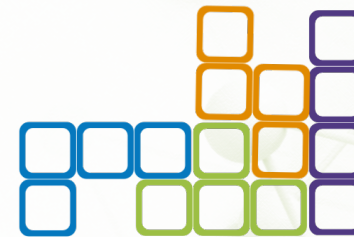
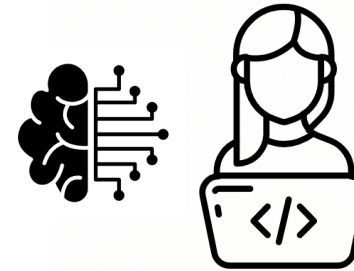




Systems where researchers can interact with AI to ease data discovery are emerging and benefit from maintaining a human in the loop



Many ways to integrate... LLM needs help to find the right one



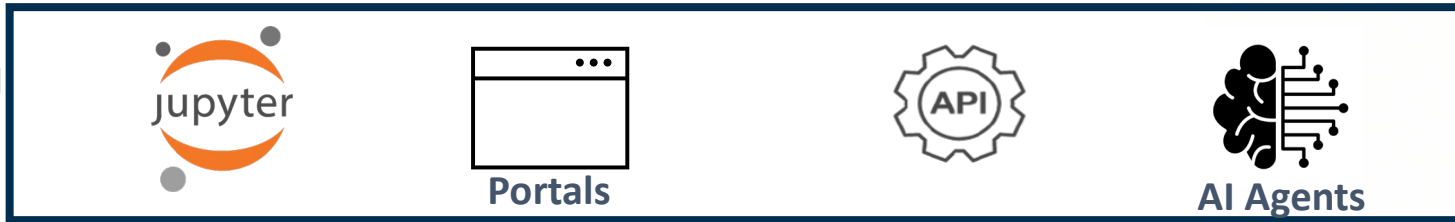
Desired result in less time?

# The BER Data Lakehouse Architecture to Enable Humans and AI



## User Interfaces for Data Science and Analytics

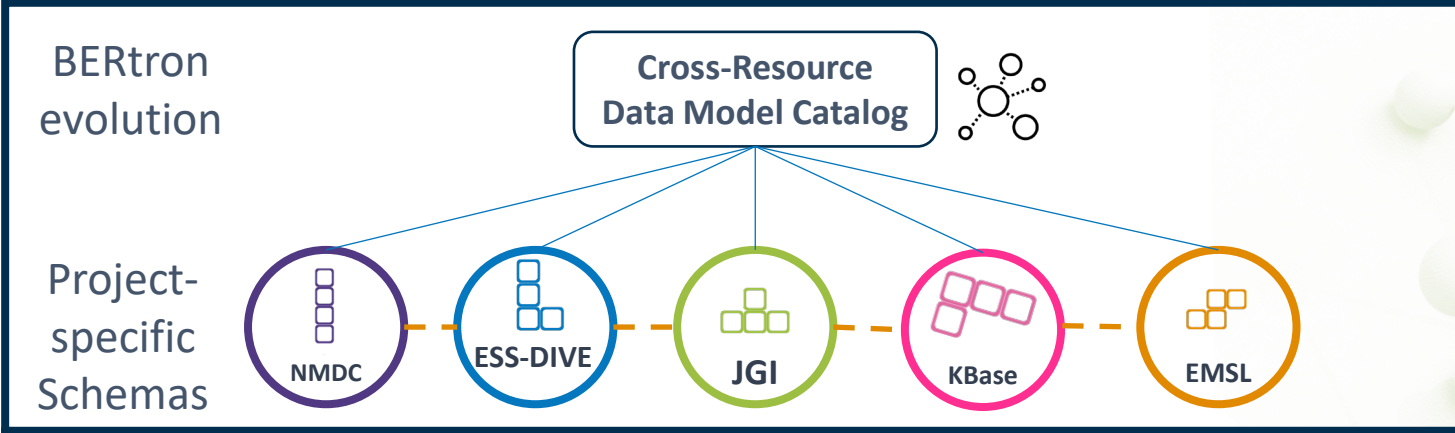
New and existing tools can be used to access BER data



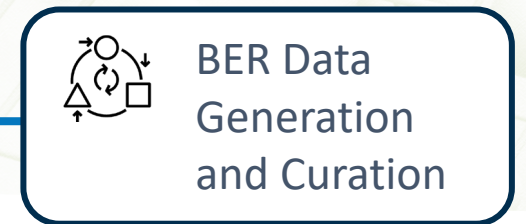
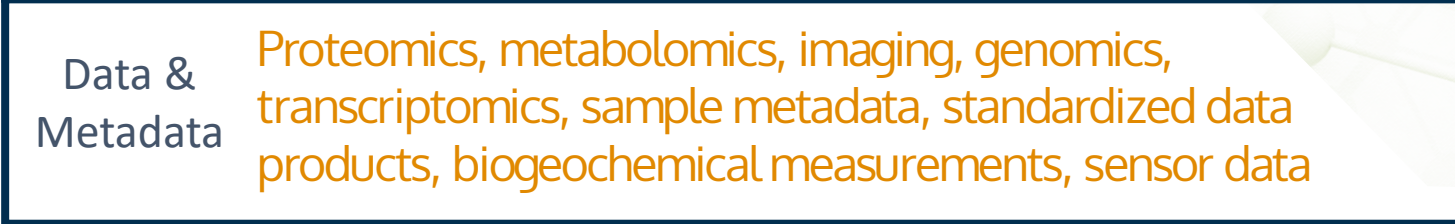
## Metadata and Governance Layer



Users can access data using either the unified catalog or project-specific schemas



Queryable information resides in an object store



# Summary -- No one can climb the mountain alone



- Getting to 80% is easy. That final 20% cannot be done alone.
- Data harmonization is necessary for integrating data from different sources
- Harmonizing sample metadata is difficult
- AI agents and LLMs are great tools, making them powerful for science requires collective effort

