Enhancing Fusion Research with TokSearch:
Updates and Integration into the Fusion Data Platform

presented by **Brian Sammuli, General Atomics**

Brian Sammuli[1], Erik Olofsson[1], Tom Neiser[1], Mitchell Clark[1], David Orozco[1], Cihan Akcay[1], Javier Hernandez Nicolau[2], Fabio Andrijauskas[2], Annmary Justine Koomthanam[3], Aalap Tripathy[3], Rishabh Sharma[3], Matthew Waller[4], Ruqi Pei[4], Zeyu Li[1], Amitava Majumdar[2], Rose Yu[2], Sicun Gao[2], Frank Wuerthwein[2], Raffi Nazikian[1], Martin Foltin[3], Craig Michoski[4], David Schissel[1]

[1]General Atomics
[2]University of California, San Diego
[3]Hewlett Packard Enterprise
[4]Sapientai

14th IAEA Technical Meeting on Control Systems, Data Acquisition, Data Management and Remote Participation in Fusion Research

July 17, 2024

Work supported by the U.S. Department of Energy under Award No. DE-FC02-04ER54698 and Award No. DE-SC0024426

Overview

- The Fusion Data Platform project
- TokSearch
- Project Timeline

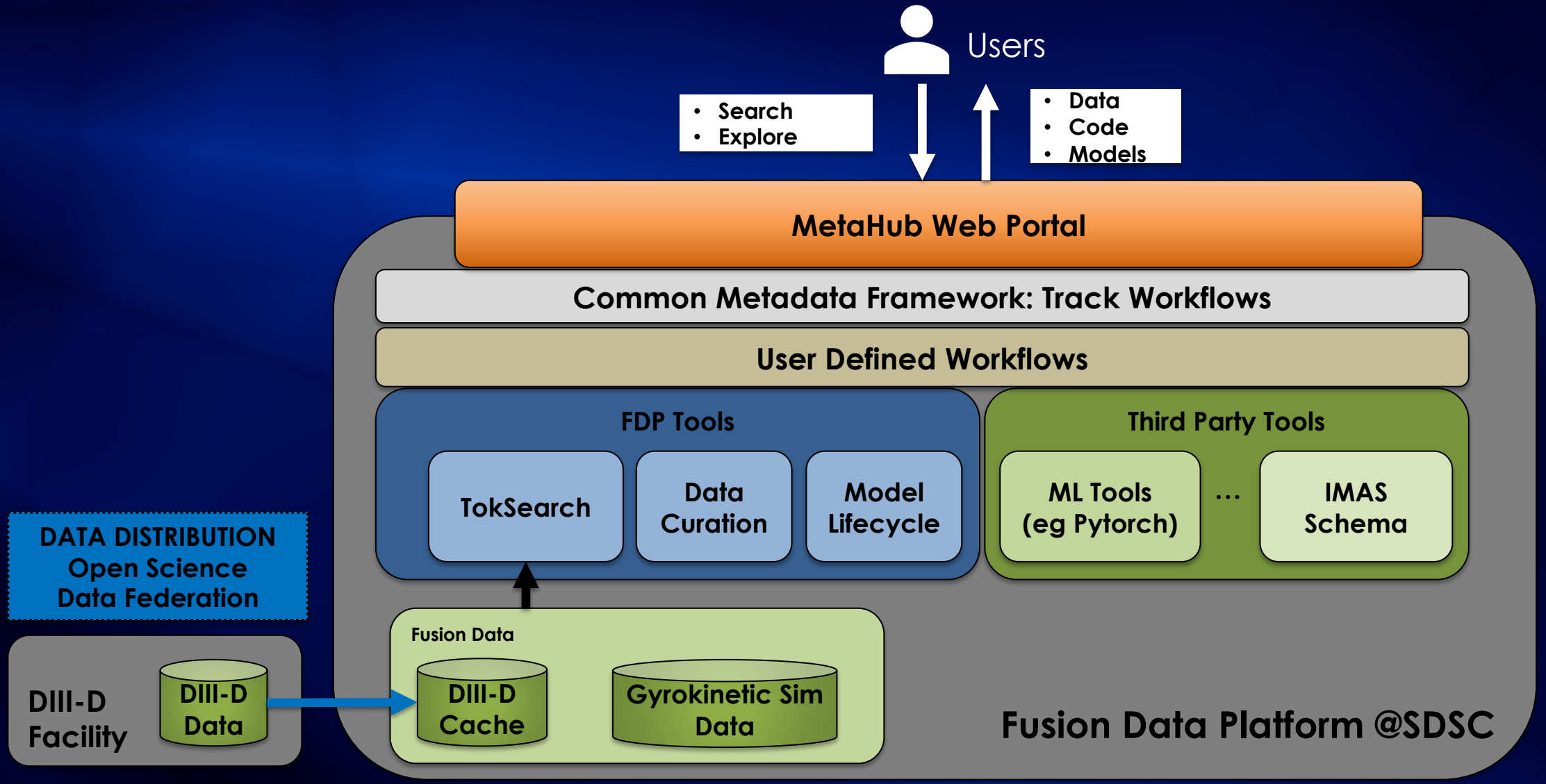
Motivation: Why do we need a Fusion Data Platform?

Need		FDP Solution
Data Distribution	No community-wide broad distribution of fusion data	Use Open Science Data Federation software stack (HEP tested)
Data Curation	No standard tools for curating fusion data	Equip FDP with fusion-specific data curation tools
Standardization	MDSplus commonly used, but no unified schema	Use IMAS schema for curated data
Reproducibility and Provenance	ML models often published with insufficient info to reproduce results	Use HPE's Common Metadata Framework for workflows, with complete provenance tracking of data, models, and metadata
Discoverability	No standard repository of curated data, no ability to search existing models/data	Provide metadata database with web portal and search interfaces

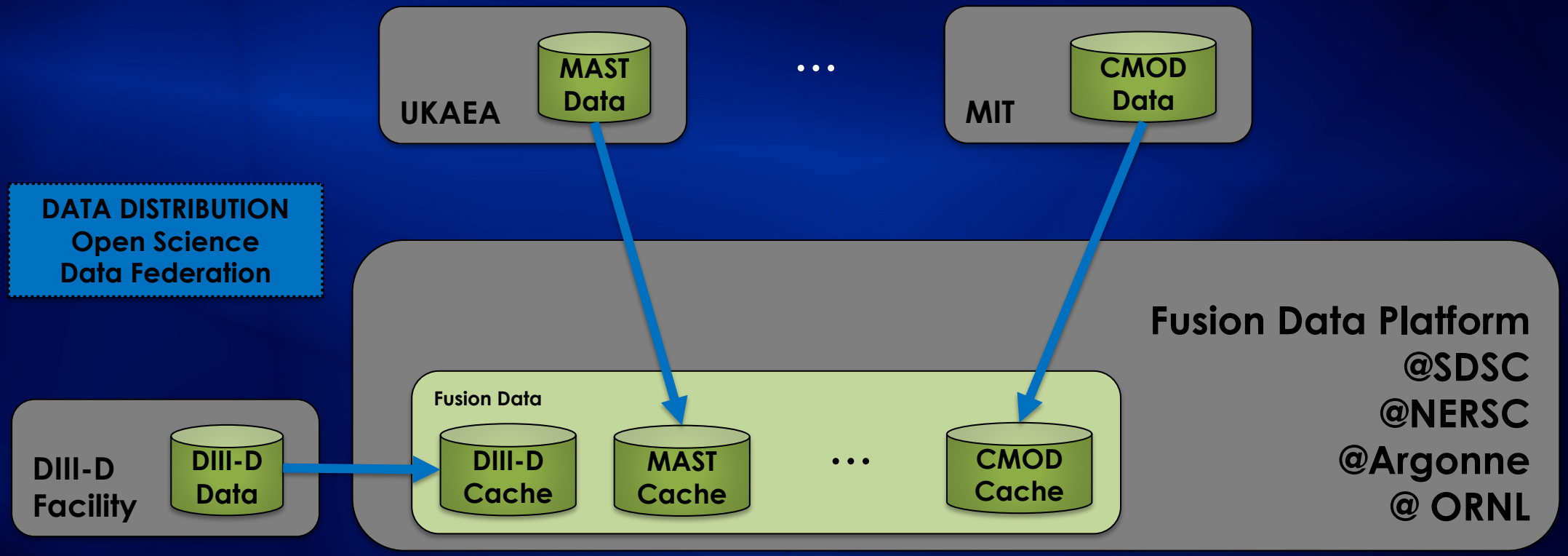
The Fusion Data Platform (FDP) project aims to address these challenges

- **Development environment for fusion data-driven applications with fusion-specific data processing and curation tools**
- **Uses distributed version control semantics for code, metadata, generated artifacts**
- **Provides federated access to data**
- **Team:**
  - General Atomics (PI: Brian Sammuli)
  - Hewlett Packard Enterprise (PI: Martin Foltin)
  - UCSD/SDSC(PI: Frank Wuerthwein)
  - Sapientai (PI: Craig Michoski)
- **3 year, $7.4M project, FES funded**

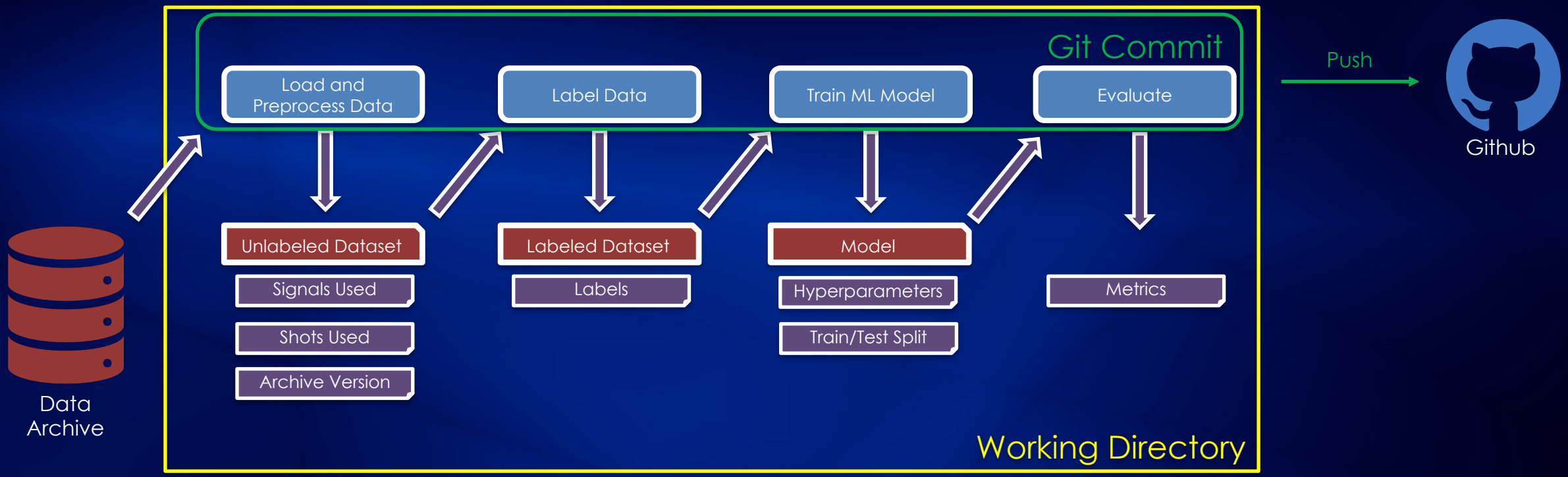
Fusion Data Platform will provide comprehensive environment for community-driven AI/ML development, supporting both experimental and simulation data



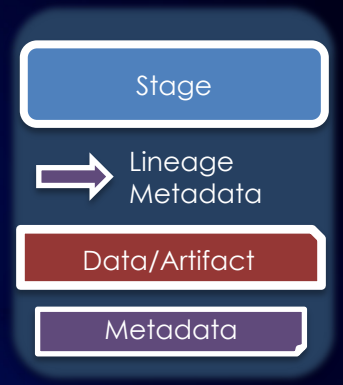
Aim is to expand platform to include additional devices, and deploy tools across multiple computing sites



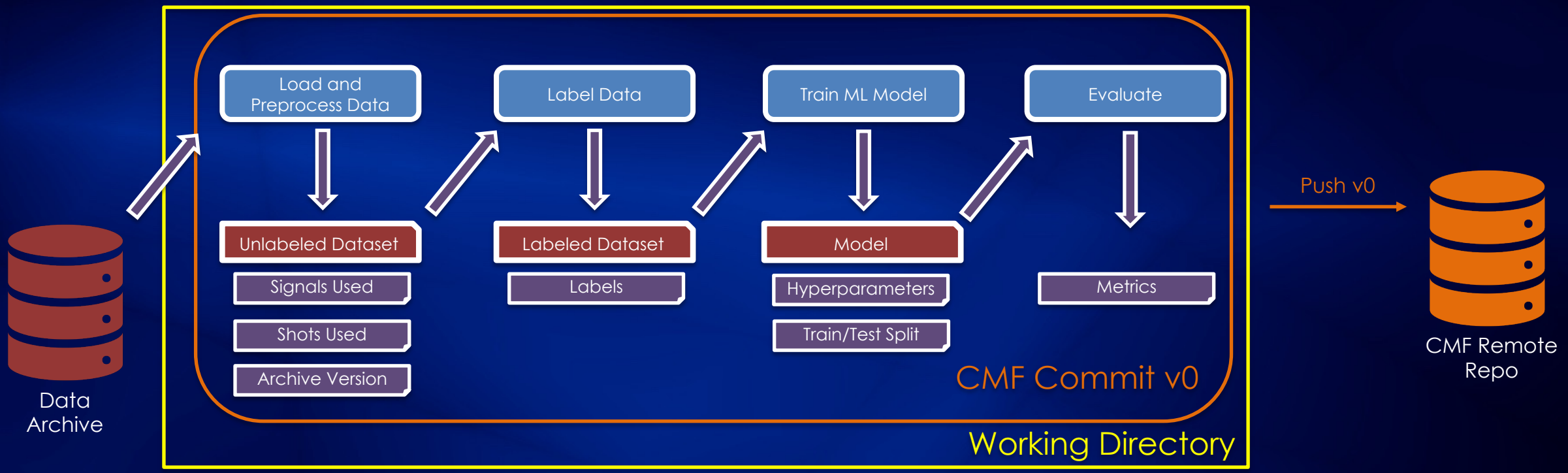
Workflows can be represented as directed graphs that capture the relationship between code, data, generated artifacts, and metadata



- **Typical Usage Pattern – Just version the code**
  - Code and model can diverge

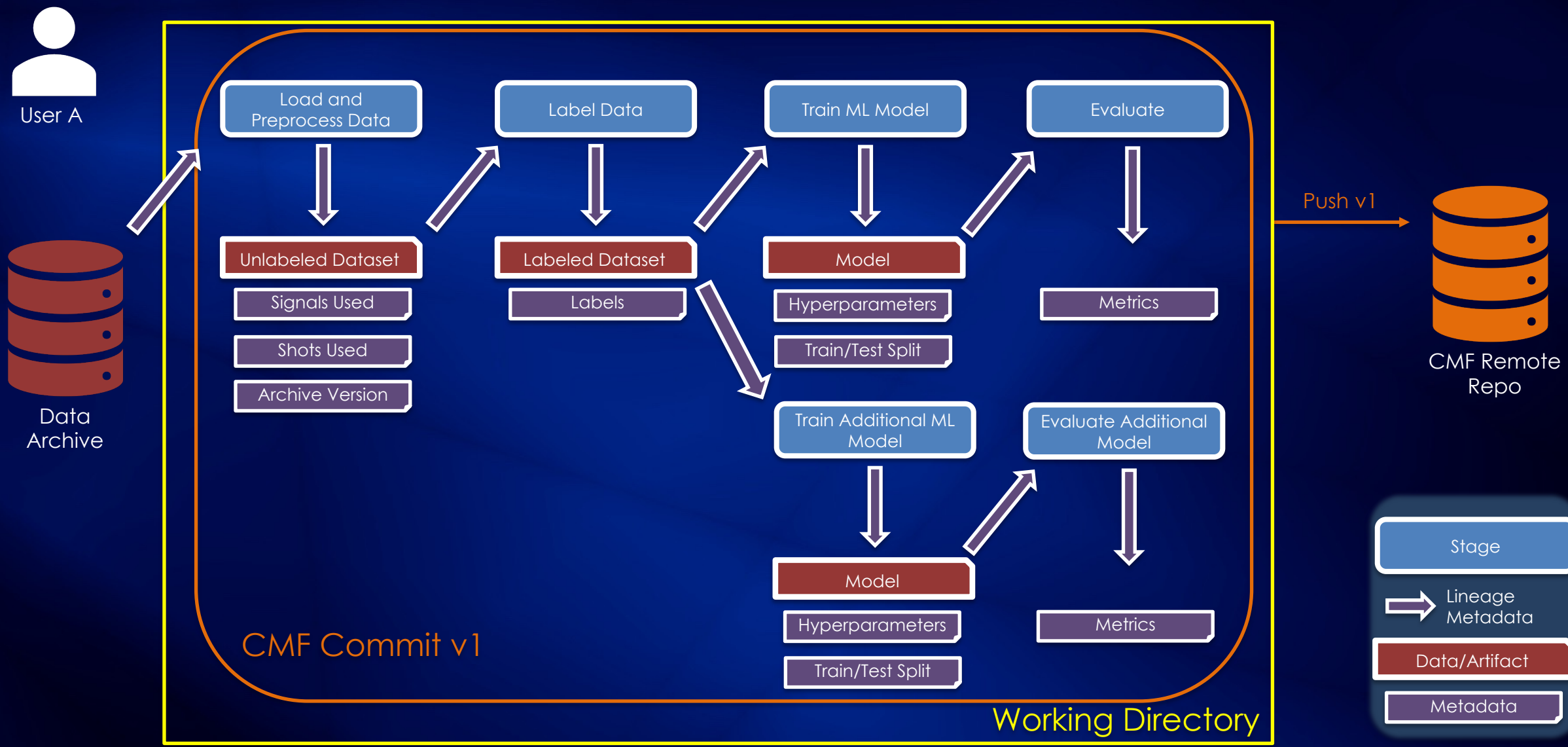


Workflows can be represented as directed graphs that capture the relationship between code, data, generated artifacts, and metadata

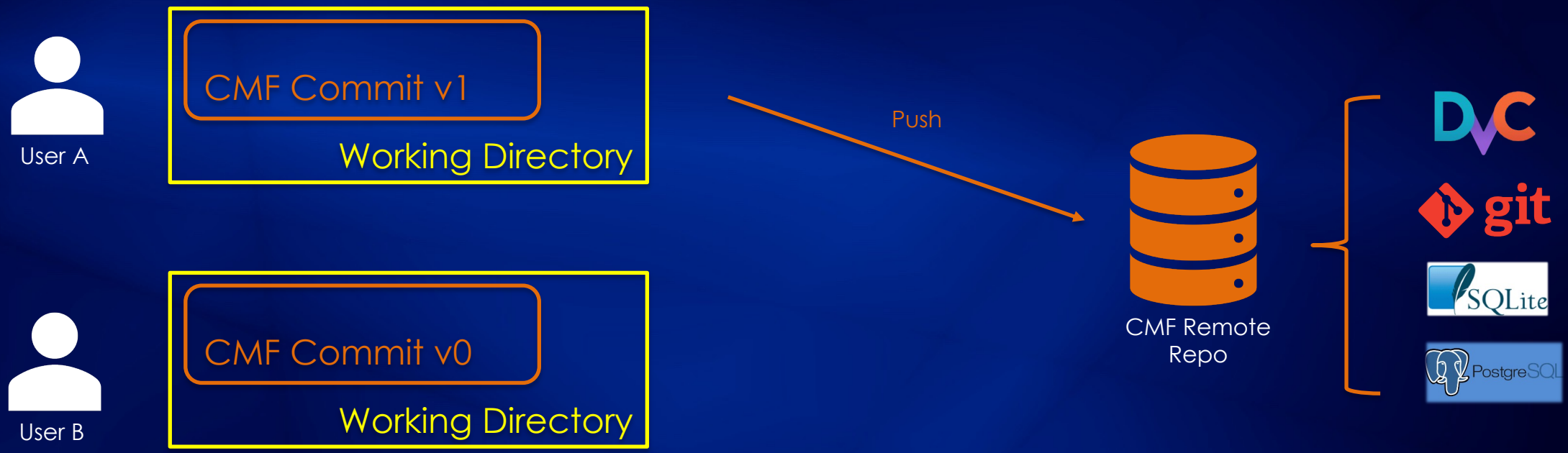


- **HPE's Common Metadata Framework (CMF) bundles entire workflow in a single atomic unit, allowing for consistent snapshots**
  - Enables provenance tracking, reproducibility, and sharing

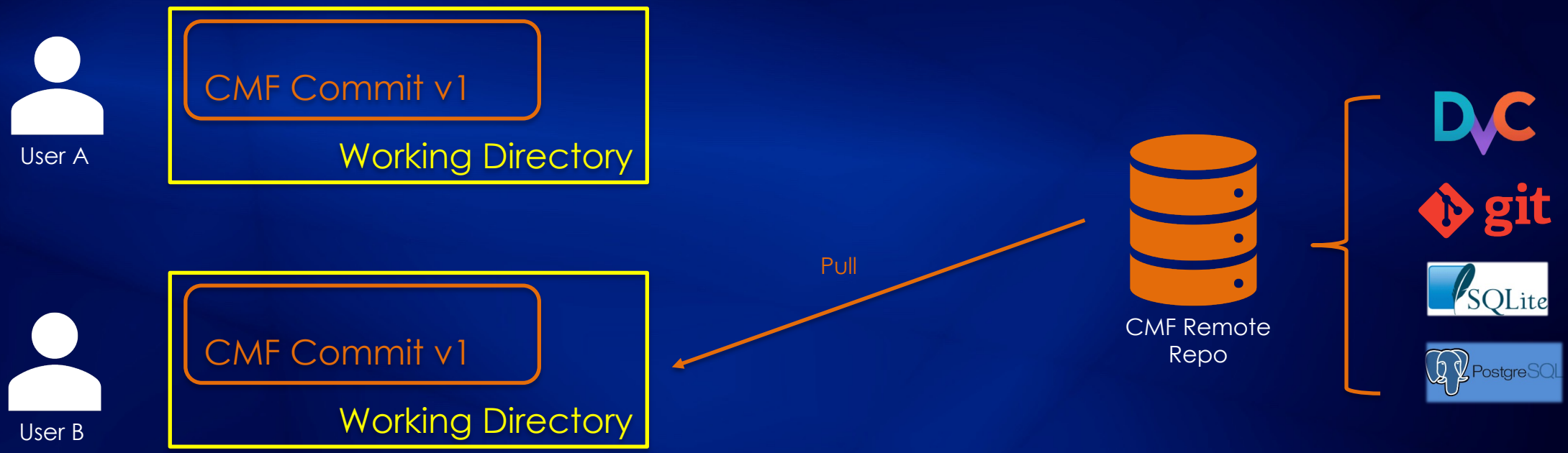
The FDP uses distributed version control semantics, allowing for collaborative development and workflow sharing



The FDP uses distributed version control semantics, allowing for collaborative development and workflow sharing



The FDP uses distributed version control semantics, allowing for collaborative development and workflow sharing



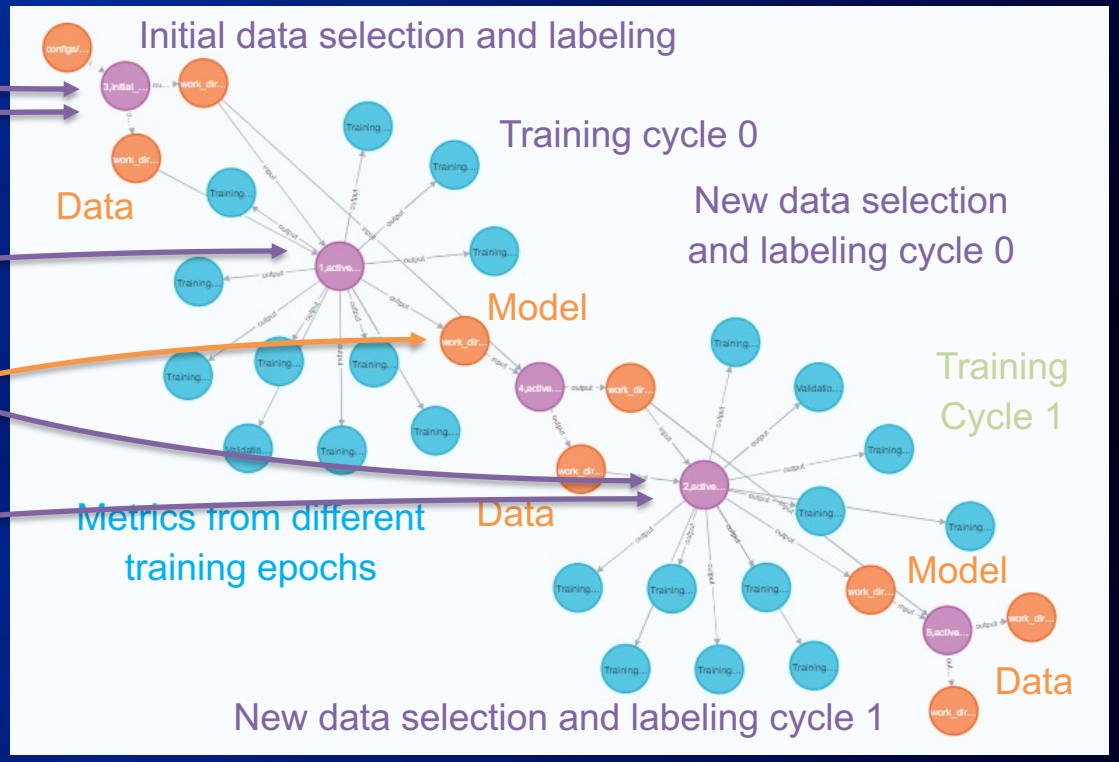
Hewlett Packard Enterprise's Common Metadata Framework (CMF) orchestrates coordinated version control of code, metadata, generated artifacts

- **CMF supports definition of arbitrarily complex workflows and integrates with popular AI/ML tools**

Simplified active learning workflow definition:

```
initial_data_selection:
    cmd:  python initial_select.py input_data selected_data_0 …
    deps: - input_data
    outs: - selected_data_0
labeling:
    foreach: - cycle: 0 - cycle: 1
    do:
      cmd: python label.py selected_data_${item.cycle} labeled_data_${item.cycle} …
      deps: - selected_data_${item.cycle}
      outs: - labeled_data_${item.cycle}
training:
    foreach: - cycle: 0 - cycle: 1
    do:
      cmd: python train.py labeled_data_${item.cycle} ai_model_${item.cycle} …
      deps: - labeled_data_${item.cycle}
      outs: - ai_model_${item.cycle}
new_data_selection:
    foreach: - cycle: 0 next_cycle: 1 - cycle: 1 next_cycle: 2
    do:
      cmd:  python select.py input_data selected_data_${item.cycle}
                             ai_model_${item.cycle} selected_data_${item.next_cycle} …
      deps: - input_data, selected_data_${item.cycle}, ai_model_${item.cycle}
      outs: - selected_data_${item.next_cycle}
```

Corresponding CMF lineage
(Labeling & Selection merged for simplicity):

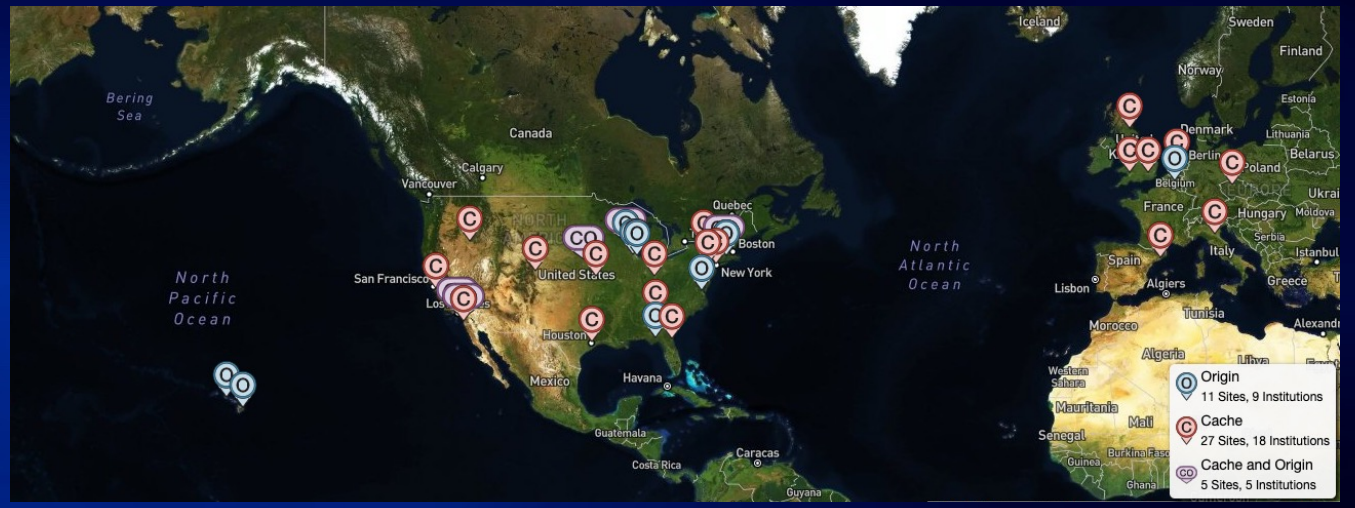


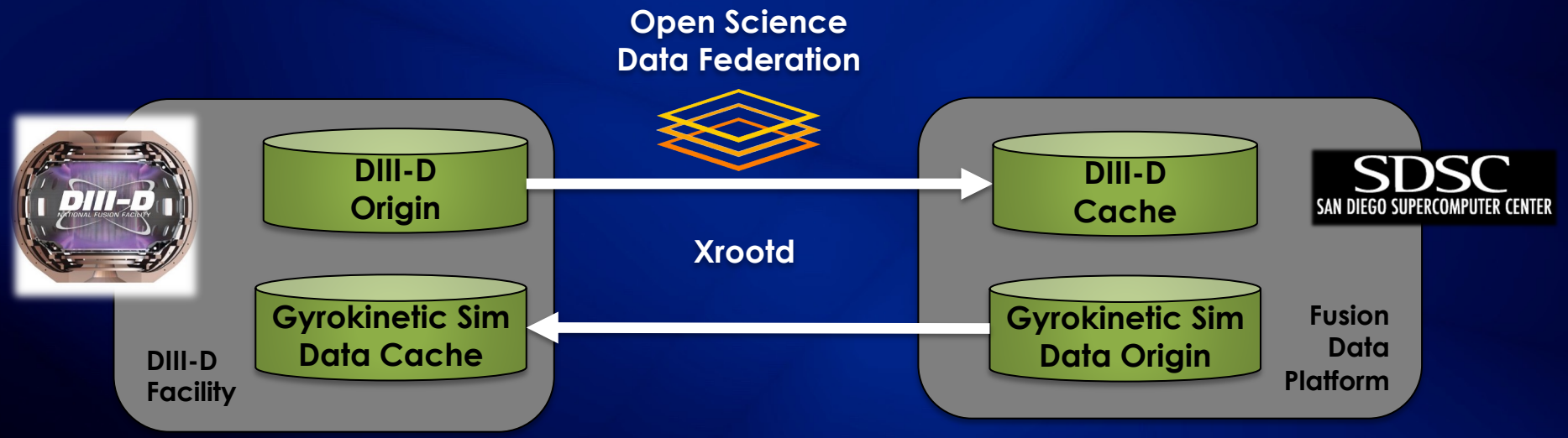
GENERAL ATOMICS

Federated access to data achieved using Open Science Data Federation (OSDF)

- **OSDF uses origin/cache scheme**
  - Origin: Original source of data
  - Cache: Access data at remote site on fast storage

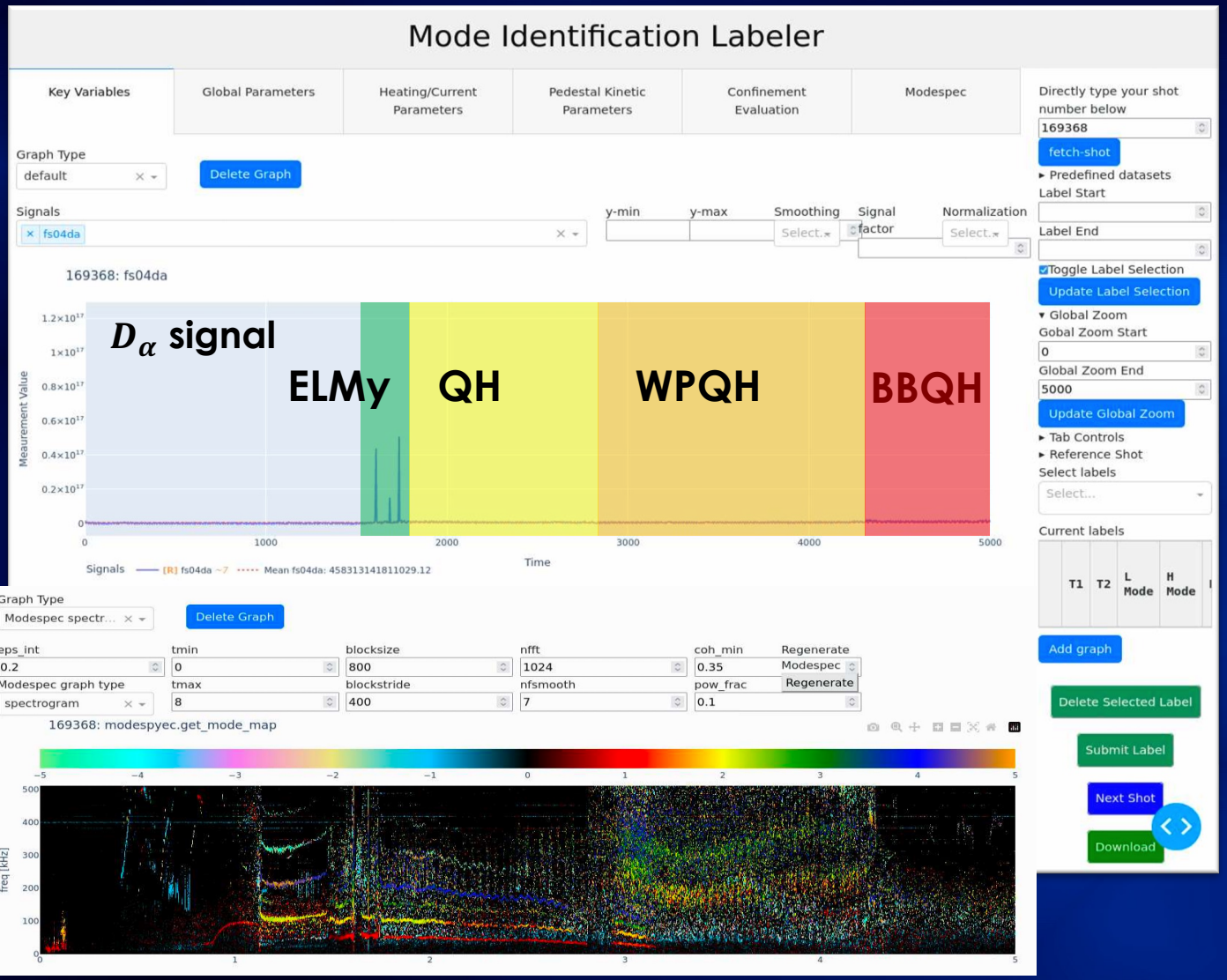
- **DIII-D origin/cache operational**
- **Gyrokinetic sim. origin/cache in progress**





GENERAL ATOMICS

Sapientai has created a new visual labeling tool: First applied to edge plasma regimes at DIII-D



Labels:

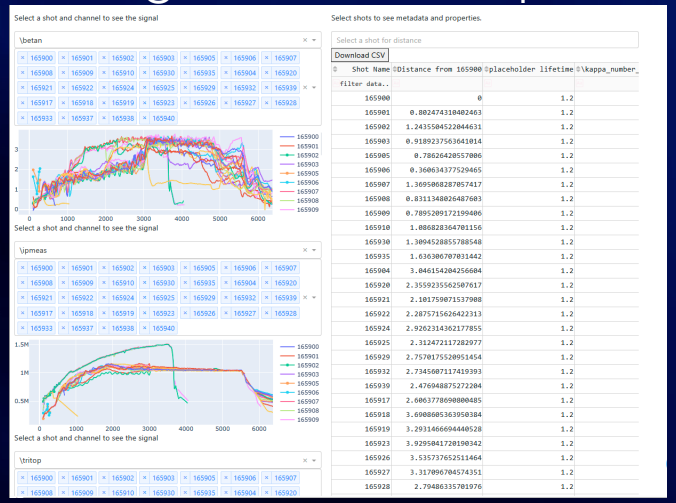
- Edge-localized-mode (ELMy)
- Standard quiescent-H mode (QH)
- Wide pedestal quiescent-H mode (WPQH)
- Broadband quiescent-H mode (BBQH)

	shots	Time Spent	Speed up
Previous database	145	~2-3 weeks	-
New Database	400	1 week	**~5-8x**

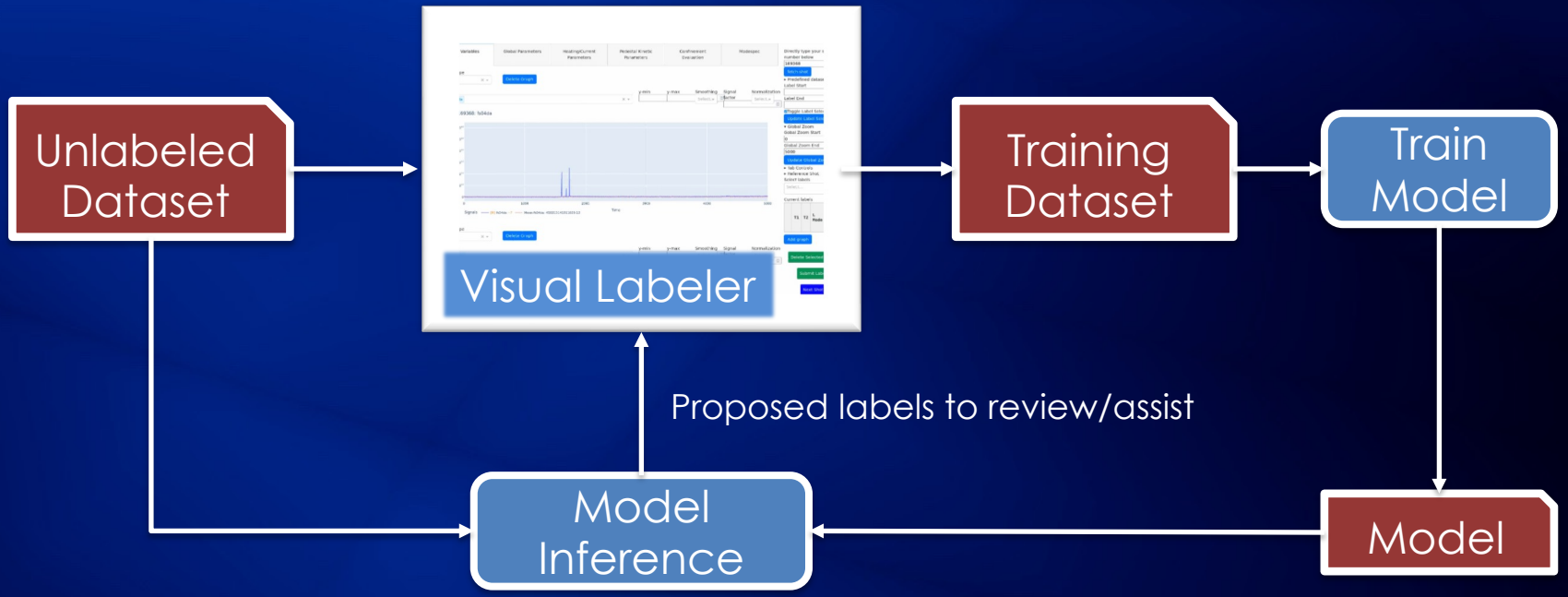
Sapientai is developing fusion-specific data exploration tools and ML-assisted labeling that will be integrated with visual labeling capability

Unsupervised clustering helps inform labeling

Historical context of many shots guides curation process



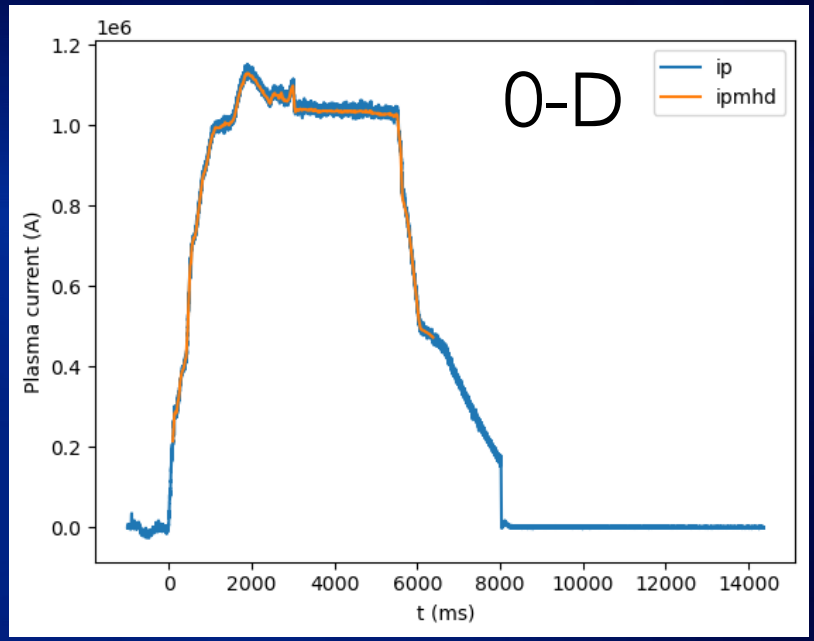
ML-assisted labeling shows proposed labels and allows user to not start from scratch

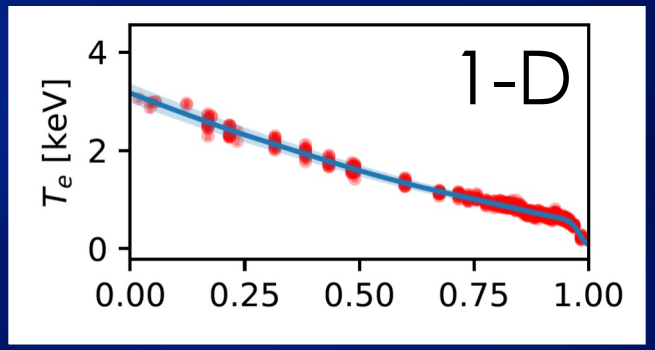


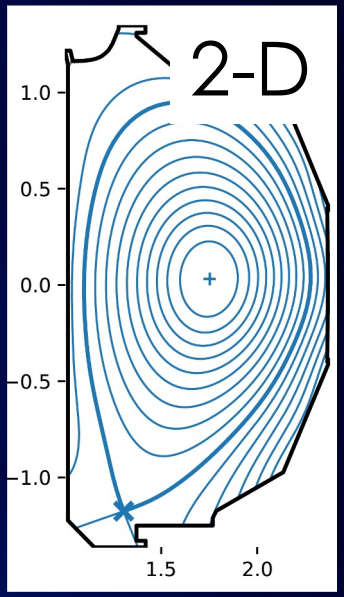
GENERAL ATOMICS

Constructing AI/ML datasets from fusion data is a challenge!

- Large volume of data
- Large variability of data types
  - Sample rates span orders of magnitude: O(1 Hz) – O(100 Mz)
- Variety of dimensionalities
  - 0-D scalar time series
    - e.g. magnetics, currents
  - 1-D profile time series
    - e.g. temperature profile
  - 2-D grid data time series
    - e.g. equilibrium reconstructions
  - Image time series
    - e.g. infrared camera data

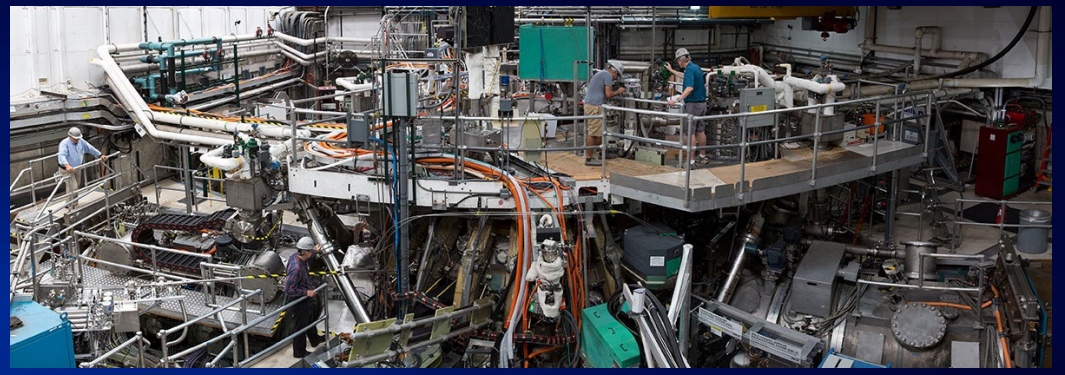






GENERAL ATOMICS

DIII-D data is a stored as in multiple formats, adding to data curation difficulty



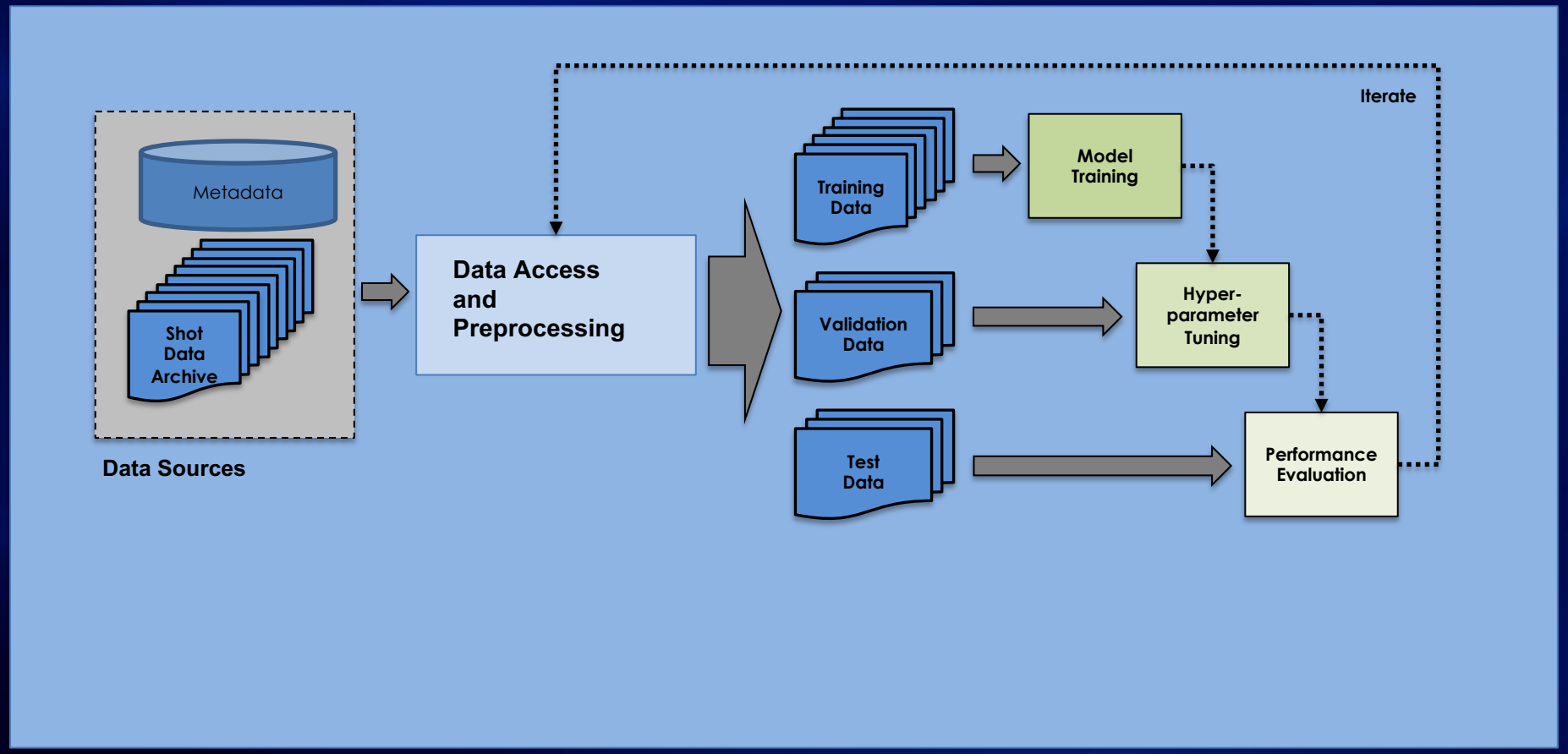
Total data volume: 700 TB

Data Archive Types:

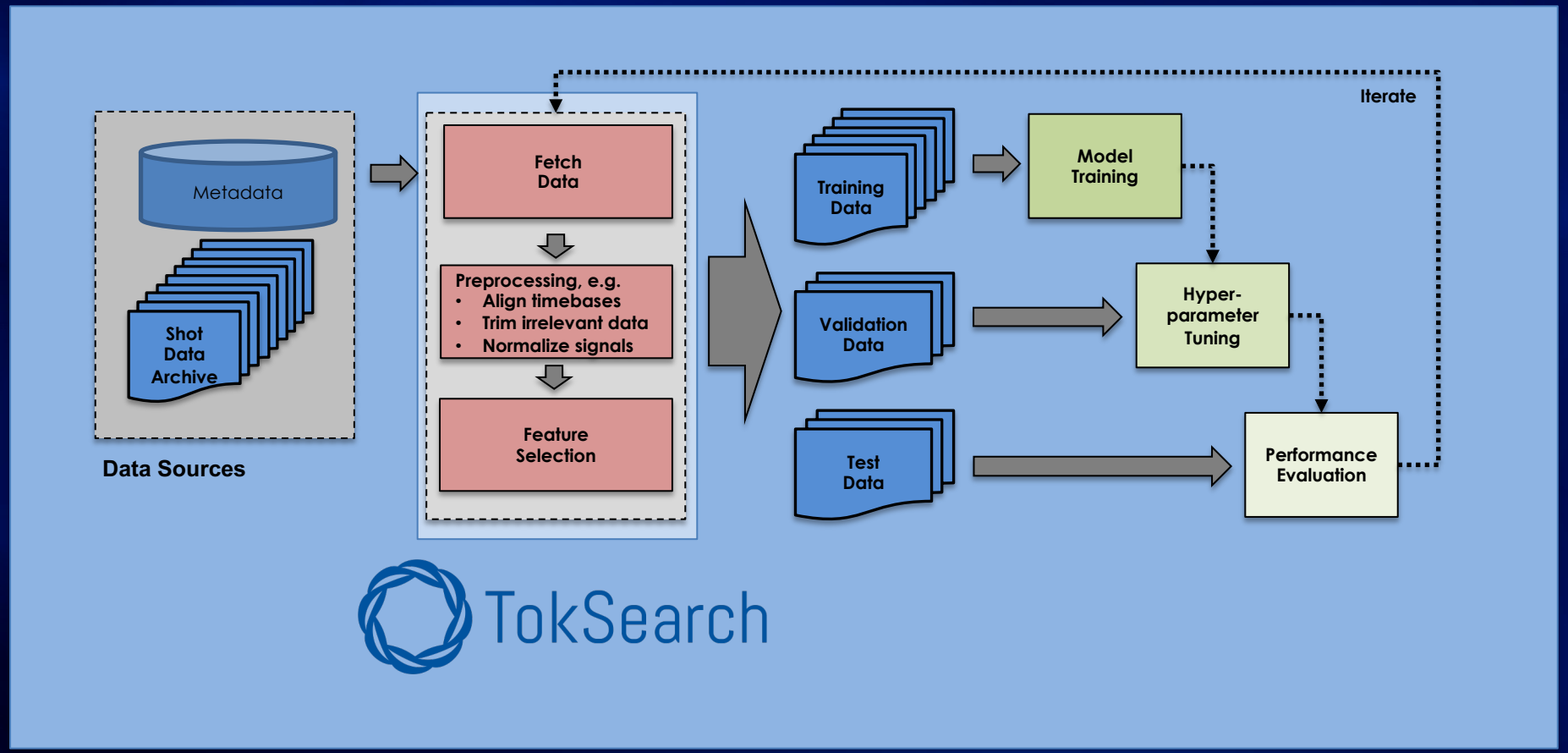
- **PTDATA**
  - Raw data from data acquisition systems (ie digitizers)
  - ~90% of total data
- **MDSplus**
  - Store output of analysis code (e.g. plasma equilibria)
  - ~10% of total data
- **Relational Database**
  - Shot metadata (e.g. date, operator notes)
  - << 1% of total data

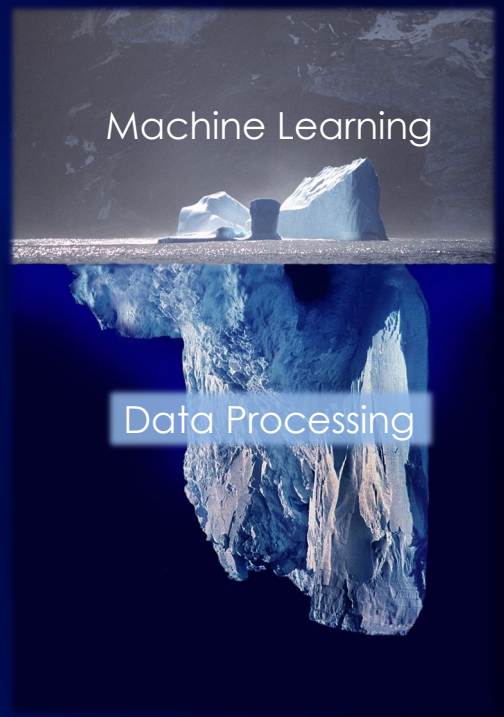
→ Typically use all three when doing AI/ML

Creating real world datasets involves significant data retrieval and preprocessing



Creating real world datasets involves significant data retrieval and preprocessing





Quick poll: How many of you have written something like this?

```
results = []
for shot in shots:
    data = read_data(shot)
    processed_data = process_data(data)
    if meets_some_criteria(processed_data):
        results.append(processed_data)
```

TokSearch

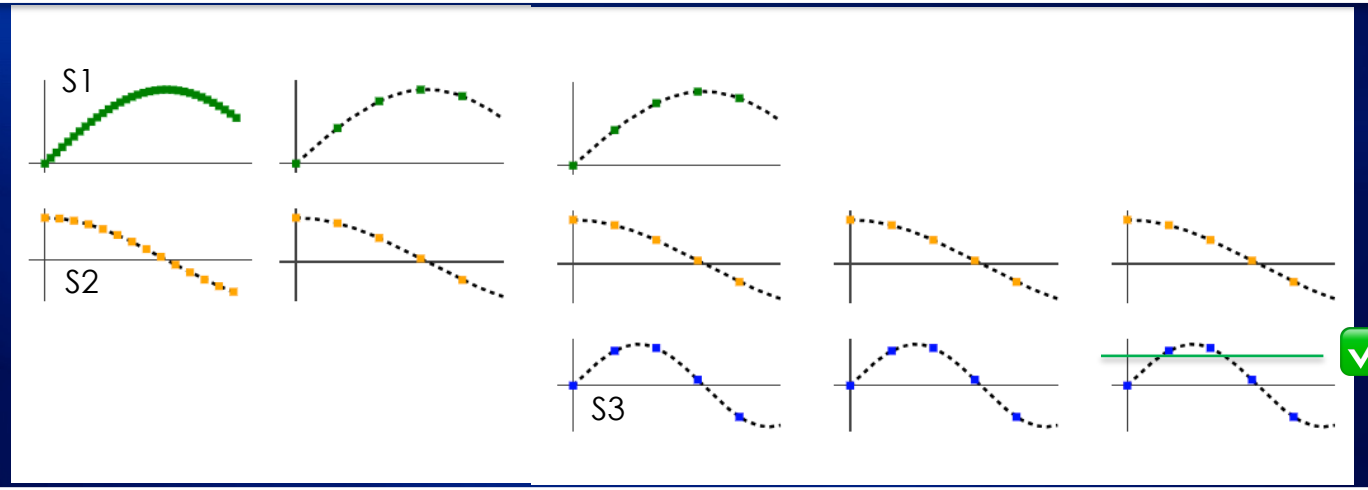
TokSearch accelerates this type of processing pipeline!

TokSearch composes built-in and user-defined functions into a pipeline applied to each shot

Align:
Put signals onto a common timebase

Keep:
Only preserve selected data

Shot No.



Output

PIPELINE

Fetch:
Pull signals from archives (0D, 1D, 2D, ...)

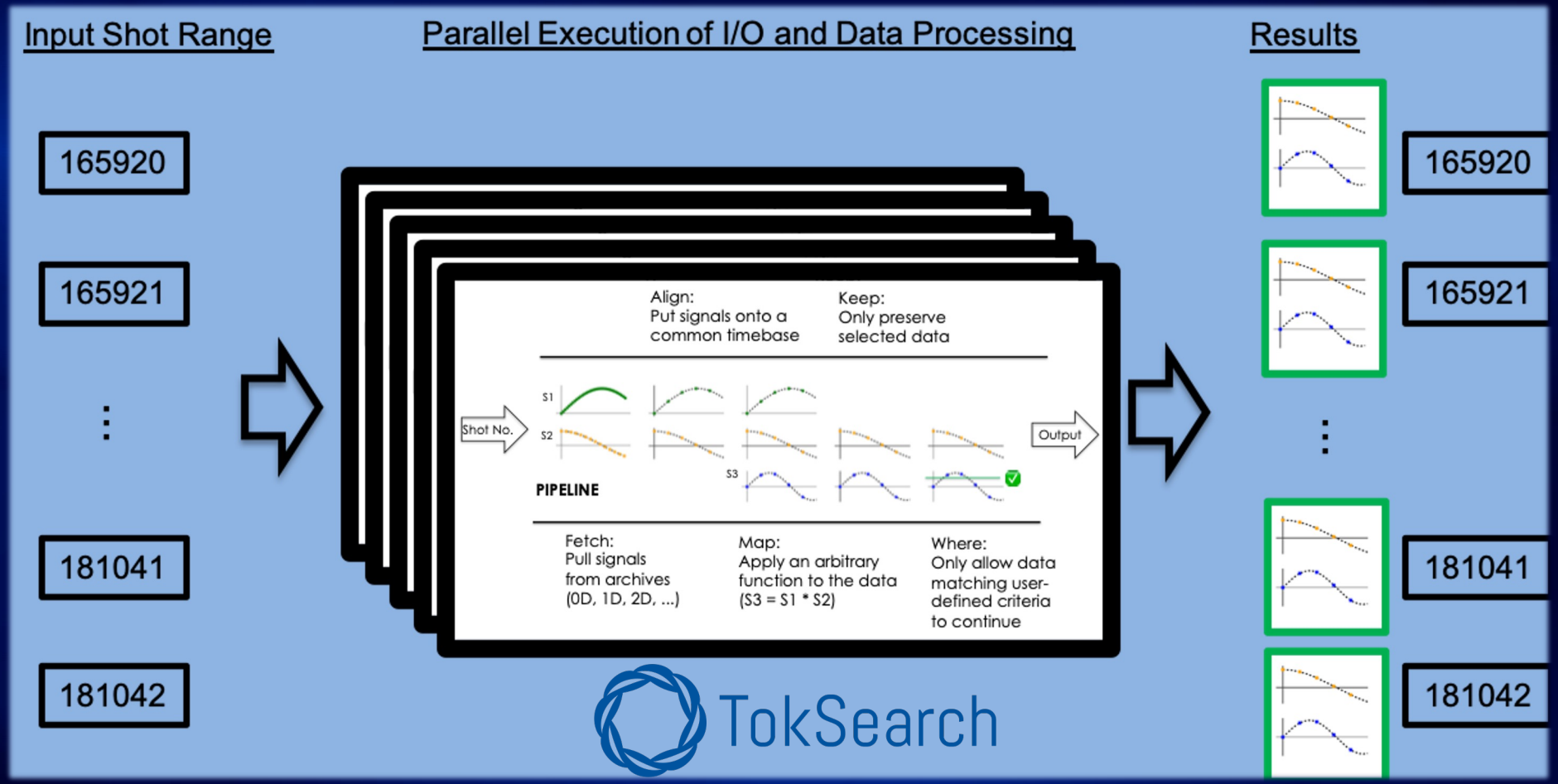
Map:
Apply an arbitrary function to the data (S3 = S1 * S2)

Where:
Only allow data matching user-defined criteria to continue

TokSearch

GENERAL ATOMICS

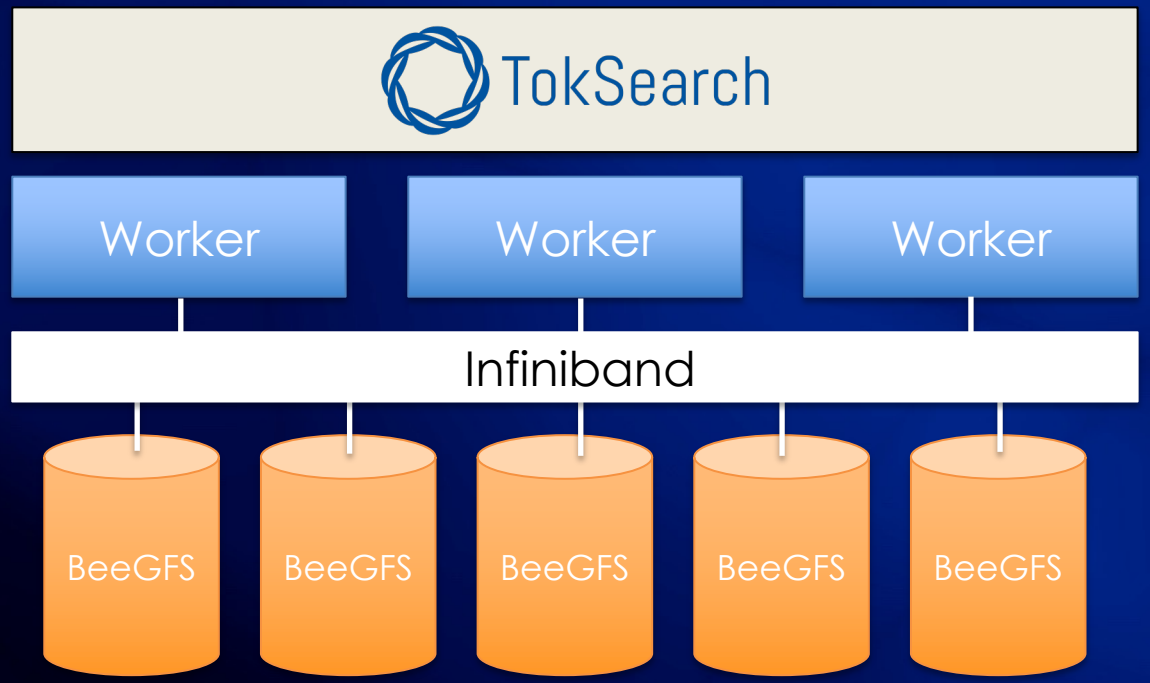
TokSearch pipelines can be executed in parallel for high throughput data processing



- Access to ~1PB of DIII-D data at high throughput

TokSearch can exploit distributed computing resources

- **Supported methods of parallelization: Ray, Apache Spark, Python Multiprocessing**
  - Others possible (MPI, Dask,...). Very easy to extend.
  - Runnable on HPC systems via SLURM
- **Full copy of DIID-D archives (0.7 PB) available on BeeGFS fast file system**
  - Currently available on Saga cluster hosted at GA
  - Will be available via OSDF on FDP



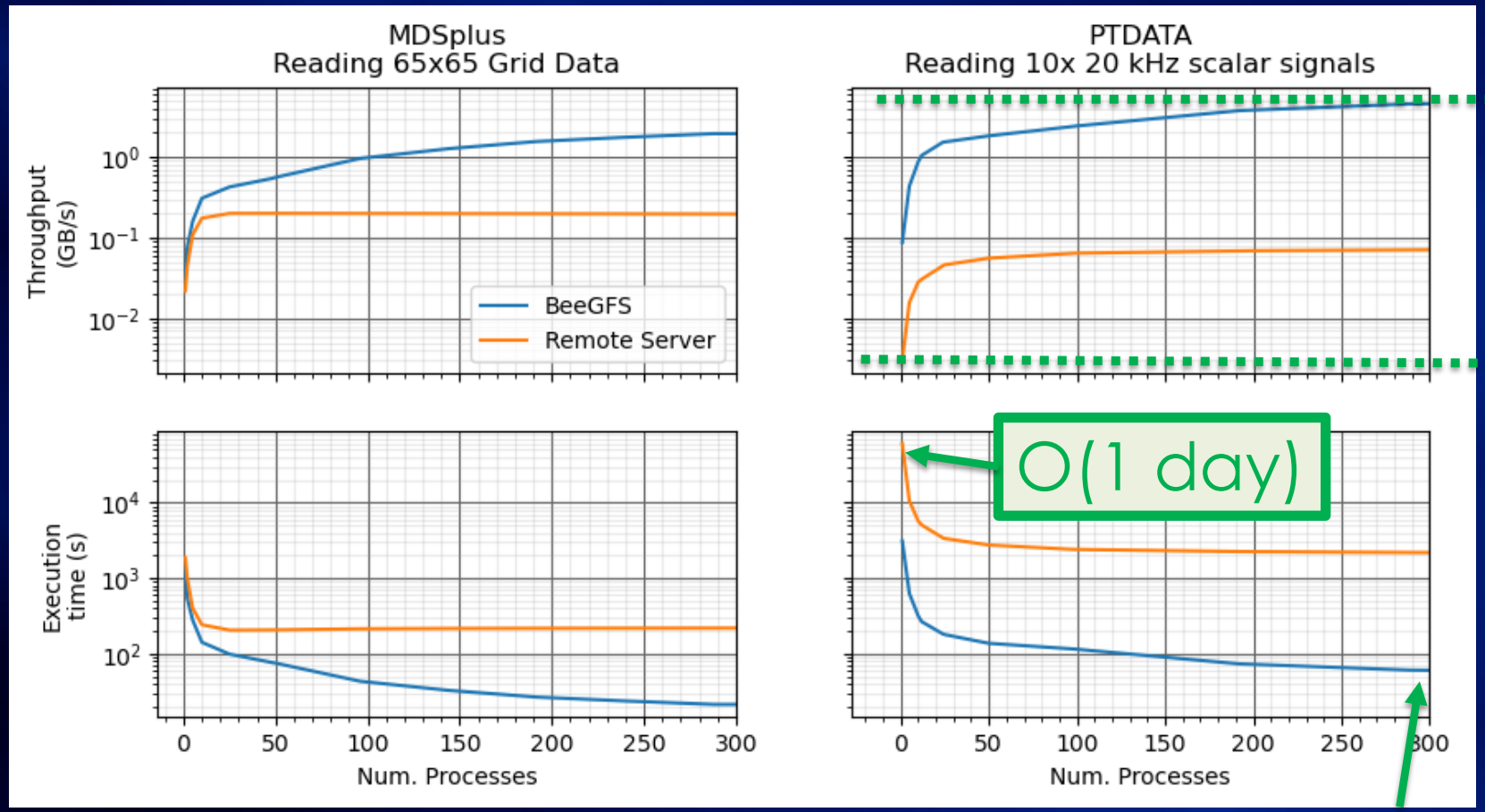
slurm workload manager

RAY

APACHE Spark™

GENERAL ATOMICS

TokSearch Performance Benchmarks:
Orders of magnitude speedup via parallelization

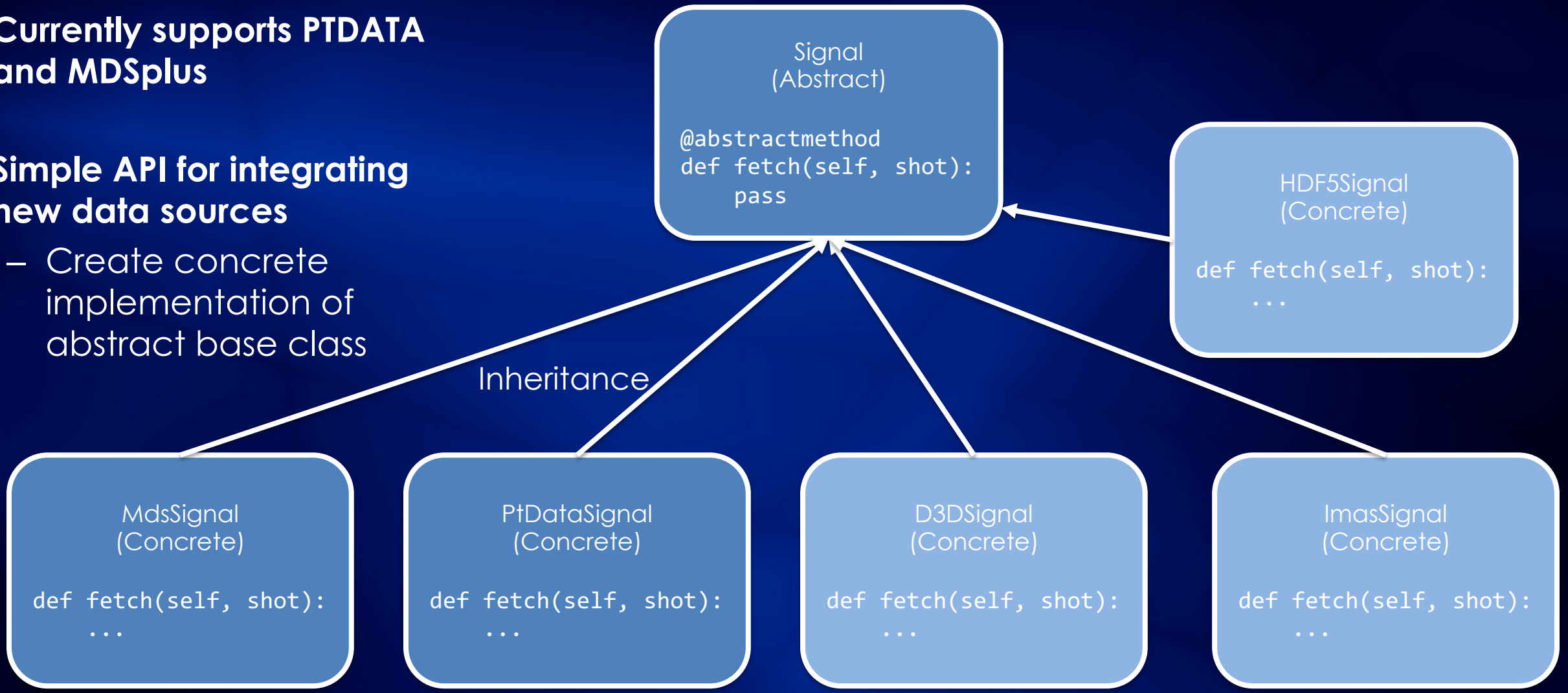


- Load and process 10,000 shots in minutes instead of days

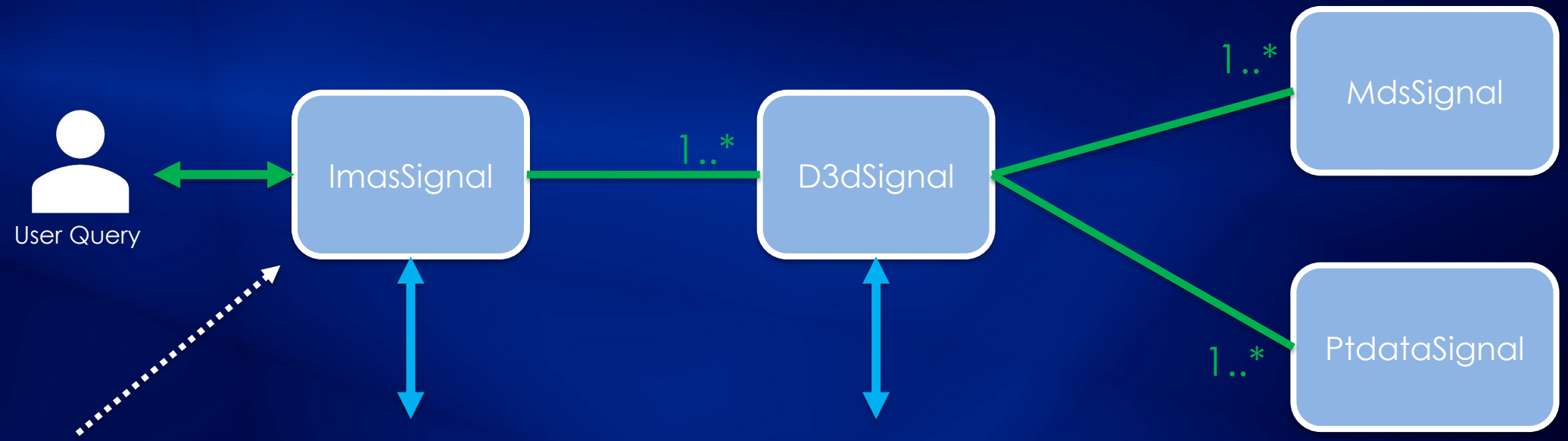
GENERAL ATOMICS

TokSearch data access is designed to be extensible

- **Currently supports PTDATA and MDSplus**
- **Simple API for integrating new data sources**
  - Create concrete implementation of abstract base class



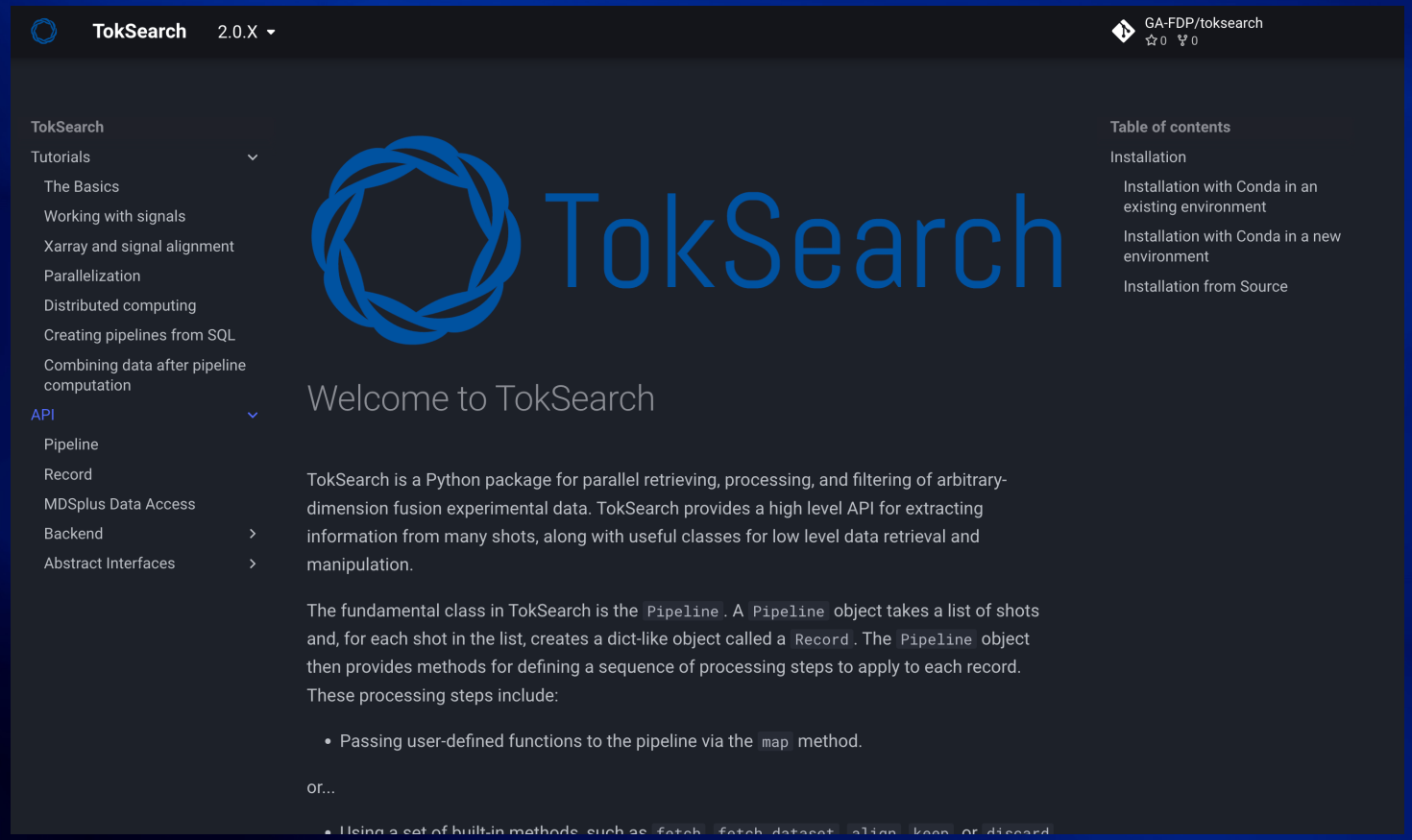
Mapping of TokSearch to IMAS Schema is Underway



- Get metadata needed to resolve IMAS entry
- Assemble compound entries
- Perform COCOS/UNIT conversions

TokSearch recently open-sourced under Apache 2.0 license

- Source available on GitHub (https://github.com/GA-FDP/toksearch)
- Docs site has tutorials, extensive API documentation

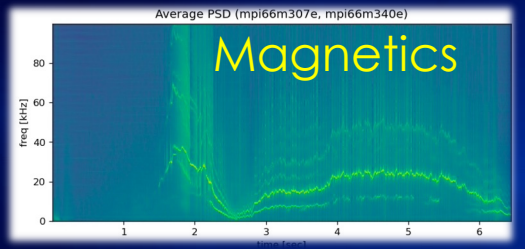


Source

Docs

GENERAL ATOMICS

A workflow that applies hazard analysis to neoclassical tearing modes has been completed using magnetic + MSE equilibria, will be extended to kinetic equilibria



TokSearch Labeling

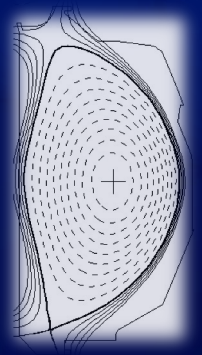
Curated Labels

Real Time Control

Train (XGBoost)

Model

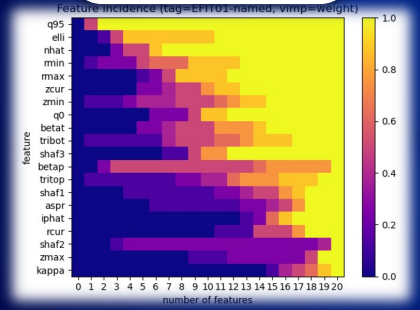
Equilibria



TokSearch Feature Extraction

Dataset

Control Room GUI (between shot)



DEMO: Identification of safe operating regimes

- **Goal: Identify high confinement regimes with low MHD activity**
- **Labels generated by heuristics analyzing locked and rotating modes in TokSearch**

Hazard Analysis using FDP tools near ready for control room deployment

- Control room tool allows experimentalists to examine causes of tearing mode precursors from previous shot, adjust subsequent control setup
- Developed collaboratively with LLNL colleagues (Holcomb, Victor)

GENERAL ATOMICS

Project Timeline

	Year 1 (FY 2024)	Year 2 (FY 2025)	Year 3 (FY 2026)
Platform Infrastructure and Tools	Initial FDP deployed and ready for use: • Core infrastructure • Initial curation tools (Visual labeling, TokSearch) • Documentation • DIII-D + gyrokinetic data	MetaHub portal deployed Model lifecycle management tools • Continual learning • Drift detection TokSearch LLM integration	Metadata indexing capability added to MetaHub Data discovery and workflow recommendation tools
Community engagement	**Announcement of alpha release at APS-DPP**	**Broad beta release** **Web portal** **DIII-D team engagement** **Student engagement** **Training, outreach**	**Full public release** **Online tutorials** **User community board**
Demonstrations	Demos initiated, documented Alpha users engaged	Migrate development effort to platform at SDSC Publish Demos and use cases on platform	Integrate model lifecycle tools, incorporate in CI/CD pipelines Publish workflows at project completion

Wrapping up

- **FDP project is underway, and is being demonstrated with DIII-D data**
- **Provides flexible environment for developing data-driven workflows**
  - Modeling
  - Data curation (labeling, tagging, cleaning)
  - Simulation
- **Distributed version control semantics and federated data access allows easy deployment at multiple computing centers (e.g. DOE, cloud)**
- **TokSearch is a key element of the FDP, allowing for accelerated data processing**

Thank you!

- **Let me know if you are interested in helping using/developing these tools**
  - Use cases welcome!
- **Also, I'm happy to provide an invite for the initial release (Oct. 2024)**
- **Contact info:**
  - Brian Sammuli
  - sammuli@fusion.gat.com