Materials Data Facility

Streamlined and automated data sharing, discovery, access, and analysis

Ben Blaiszik[1,2] (blaiszik@uchicago.edu),
Logan Ward[1] (loganw@uchicago.edu)
Ian Foster (foster@uchicago.edu)[1,2],
Jonathon Gaff[1], Kyle Chard[1], Jim Pruyne[1],
Rachana Ananthakrishnan[1], Steven Tuecke[1]
Michael Ondrejcek[3], Kenton McHenry[3], John Towns[3]
University of Chicago[1], Argonne National Laboratory[2], University of Illinois at Urbana-Champaign[3]

materialsdatafacility.org

globus.org

Materials Genome Initiative

Team (unordered)

UC/Argonne


Ian Foster (PI)


Ben Blaiszik

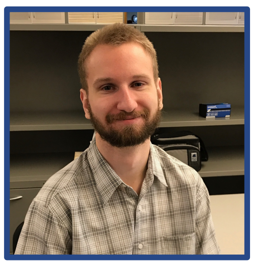

Steve Tuecke


Jim Pruyne


Rachana Ananthakrishnan


Kyle Chard


Logan Ward


Jonathon Gaff


Stephen Rosen

Illinois (Urbana-Champaign)


John Towns (PI)


Kenton McHenry


Michal Ondrejcek

Streamline and automate: Four keys

- Simplify data publication, regardless of size, type, and location
- Automate data and metadata ingest, to enable capture of many valuable materials datasets
- Enable unified search of disparate materials data sources
- Deploy APIs to foster community development, data creation, and data consumption

MATERIALS DATA FACILITY

Streamline and automate

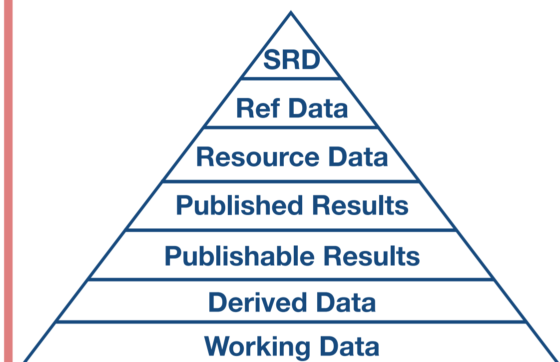
Publication

REST APIs

Discovery

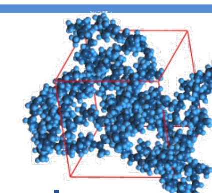
- Identify datasets with persistent identifiers (e.g., DOI or Handle)
- Describe datasets with appropriate metadata and provenance
- Handle big (and small) data: We have already ingested datasets with > 1.5 M files and > 1.5 TB in size

Materialsdatafacility.org



- Search, query, and access datasets in modern ways
- Automatically index flexible metadata and harvest file contents
- Provide simple user interfaces (c.f., Google and Amazon)

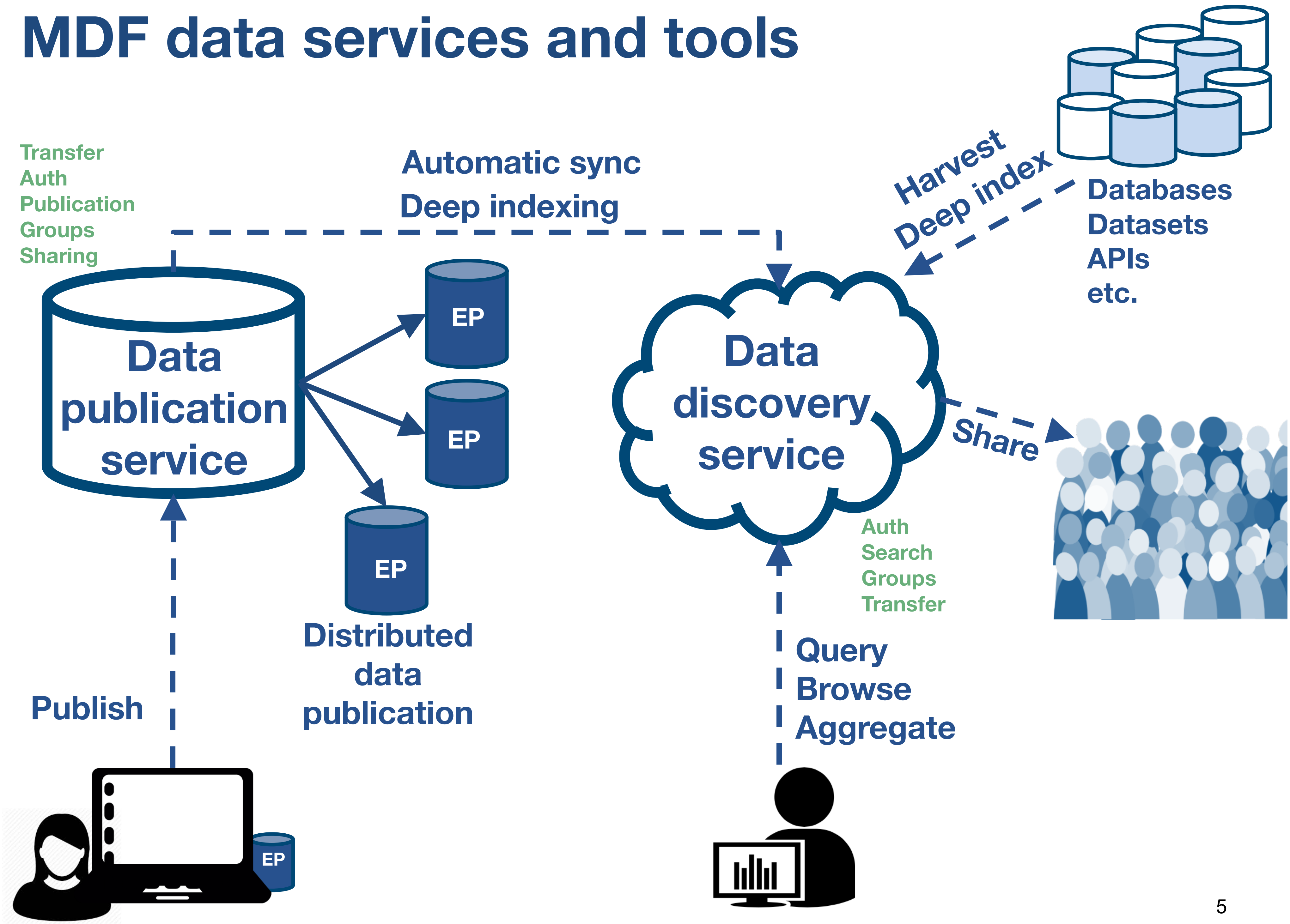
PPPDB



- Extract key polymer properties from literature via natural language processing and crowdsourcing
- Build interfaces to explore curated χ and other property values
- Includes 375 polymer-polymer values and 1,014 polymer-solvent values

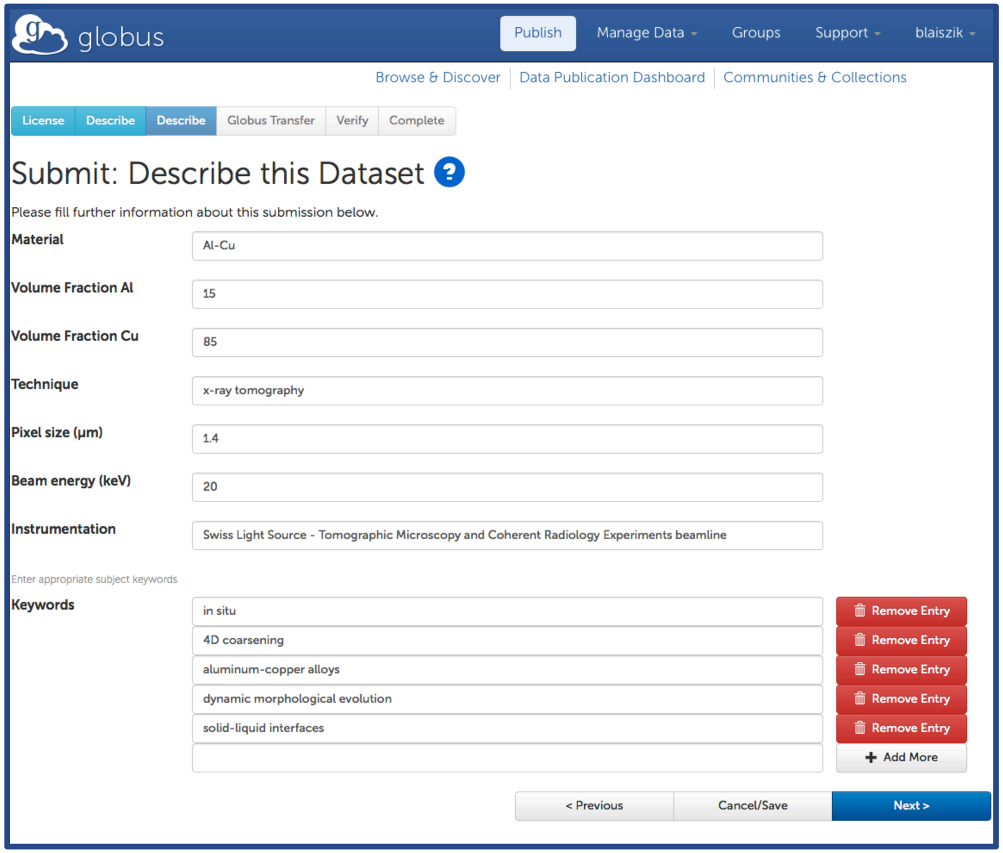
NIST National Institute of Standards and Technology U.S. Department of Commerce · CHiMaD Center for Hierarchical Materials Design · The University of Chicago · Northwestern University · Illinois University of Illinois at Urbana-Champaign · Argonne National Laboratory

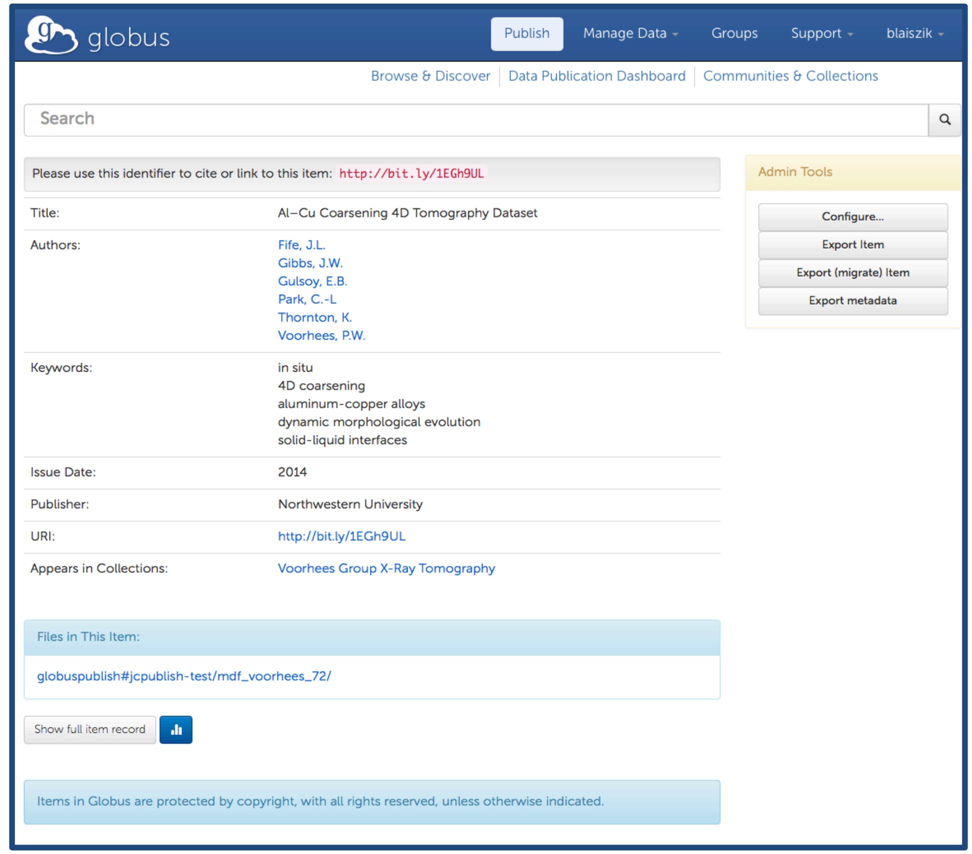
MDF data services and tools



Data publication

Data publication service





- Mechanisms to create and enforce schemas and logical collections
- Web UI to create datasets and manage curation and admin tasks
- Tools to automate publication process

- Dataset record permanent landing page for DOI link
- Record shows some metadata links to the rest
- Direct link to underlying files
- Download statistics

Tool to automate data publication

User perspective

1. Setup endpoint at data origin and create config (one time)
2. Collect data on EP
3. Describe data (JSON)
4. Publish!

1 Configure

```
import autopublish
config = {
    "source_ep" : "e38ee745-6d04-11e5-ba46-22000b92c6ec",
    "source_path" : "/MDF/testing_publication/data2/",
    "metadata_path": "/MDF/testing_publication/data2/data2_metadata.json"
}
```

2 Publish dataset

```
autopublish.publish(**config)
```

Tool to automate data publication

User perspective

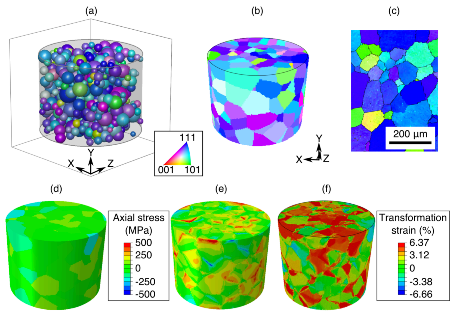
1. Setup endpoint at data origin and create config (one time)
2. Collect data on EP
3. Describe data (JSON)
4. Publish!

Behind the scenes

- Publication created in MDF Data Publication service
- Directory created on publication endpoint
- ACLs set on endpoint as defined by collection
- Submitted metadata added to the database and verified against schema
- Transfer automated between origin and publication endpoint
- (optional) Curation flow started
- DOI minted
- Metadata registered in search → metadata pushed to MRR

Published Data Highlights

Grain Structure, Grain-averaged Lattice Strains, and Macro-scale Strain Data for Superelastic Nickel-Titanium Shape Memory Alloy Polycrystal Loaded in Tension

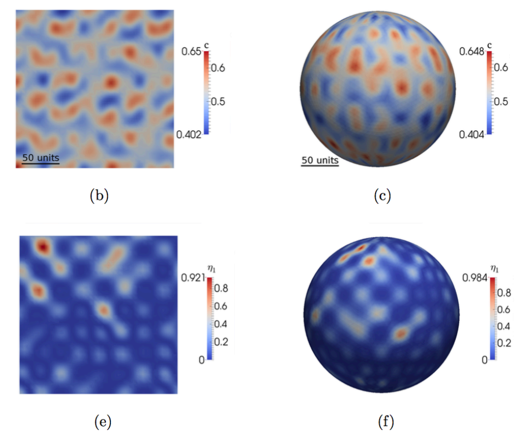


- Largest dataset to date (>1.5 TB). Showcases MDF unique capabilities and makes a unique dataset discoverable for code development, analysis, and benchmarking

Paranjape *et al.*

http://dx.doi.org/10.18126/M2NK5W

Phase Field Benchmark I Dataset



Jokisaari *et al.*

http://dx.doi.org/doi:10.18126/M2101X

Electron Backscattering and Diffraction Datasets for Ni, Mg, Fe, Si

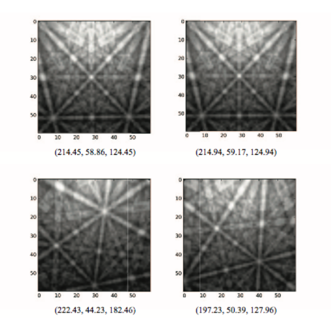
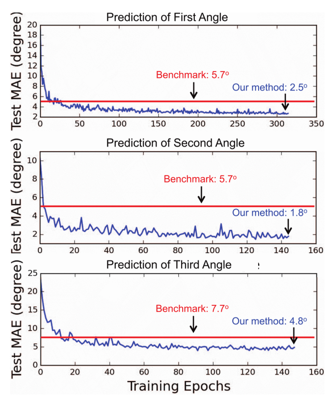


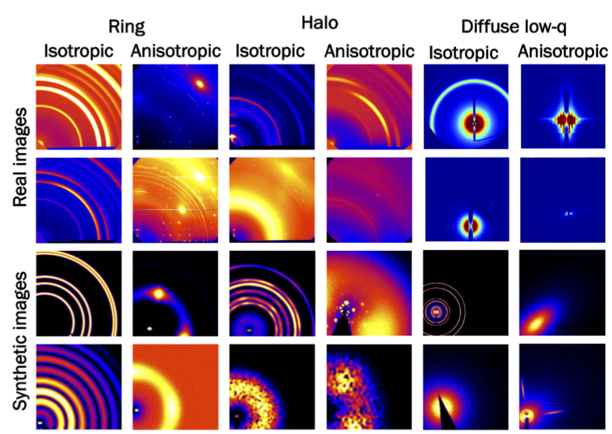
Figure 2: Examples of four EBSD patterns, each denoted with its corresponding Euler angles $(\varphi_1, \Phi, \varphi_2)$, used as regression target in deep net training. The upper left and upper right patterns are very similar, and also have a small difference in target angles.



Marc De Graef *et al.*

http://dx.doi.org/doi:10.18126/M2D593

X-ray Scattering Image Classification Using Deep Learning



layer name	output size	kernels	
conv1	112×112	7×7, 64, stride 2	
		3×3 max pool, stride 2	
conv2_x	56×56	1 × 1, 64 3 × 3, 64 1 × 1, 256	× 3
conv3_x	28×28	1 × 1, 128 3 × 3, 128 1 × 1, 512	× 4
conv4_x	14×14	1 × 1, 256 3 × 3, 256 1 × 1, 1024	× 6
conv5_x	7×7	1 × 1, 512 3 × 3, 512 1 × 1, 2048	× 3
pooling	1×1	average pooling	
fc	1×1	2048×num of attributes	

Yager *et al.*

http://dx.doi.org/10.18126/M2Z30Z

Data discovery and demos

MDF data search: ingest and indexing

Start your search here

Data index	17 Data sources indexed	~1M Records

Metadata index	6 Repositories harvested ~200 Datasets ~260 TB Made discoverable	• MDF • NIST MML Repo • MATIN • Materials Commons • CXIDB • NDS Materials Resource Registry

>2.75 PB Identified for future indexing

MDF data ingest and indexing

Datasets/Databases

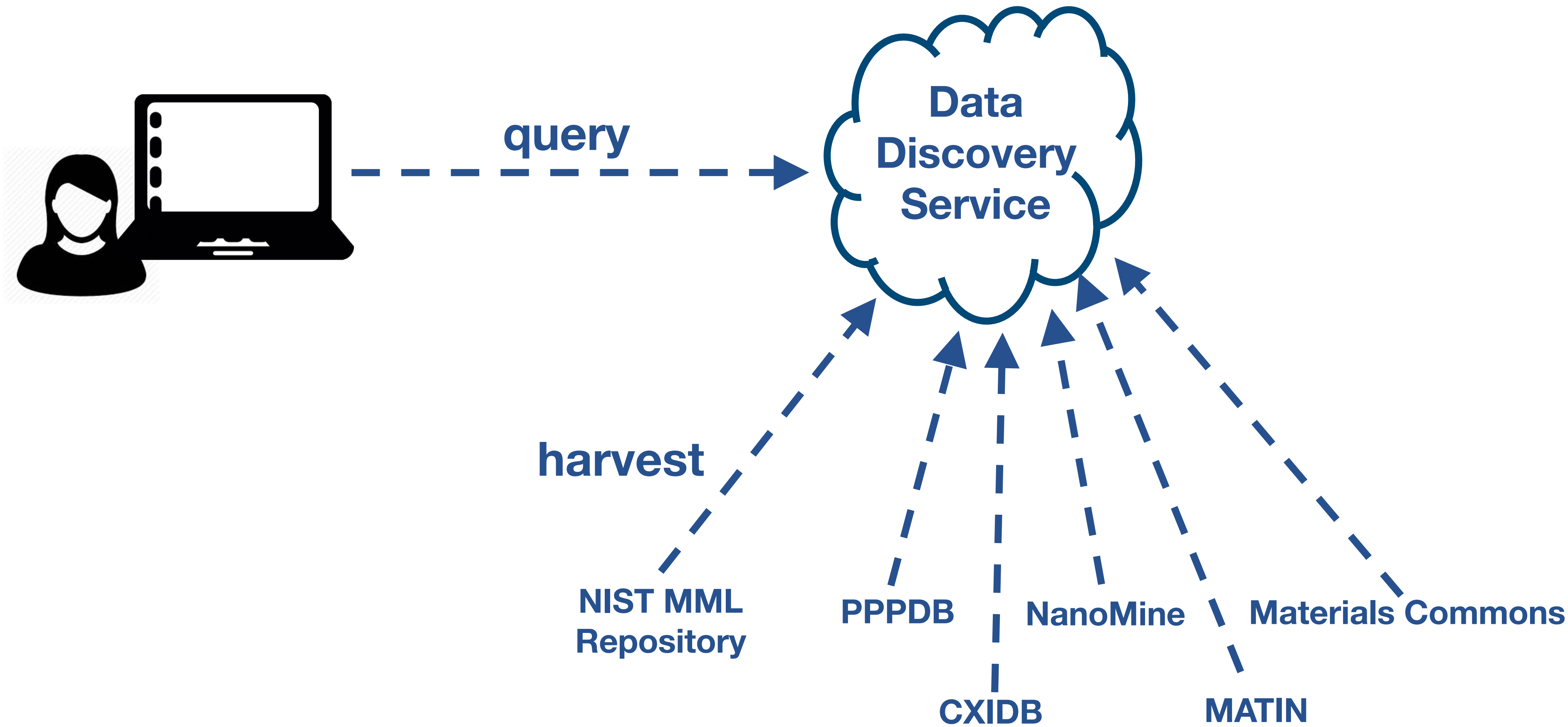
- NanoMine (CHiMaD)
- PPPDB (CHiMaD)
- Khazana Polymers
- Khazana VASP
- Ab Initio Solute Solvent Diffusion Dataset
- JANAF (NIST)
- Harvard Organic Photovoltaic Database
- SLUCHI (VASP)
- Crystallography Open Database
- Classical Interatomic Potentials (NIST)
- Interatomic Potentials Repository (NIST)
- XAFS Data Library
- Lytle XAFS Dataset
- OQMD (CHiMaD)
- NREL Organic Electronic DB
- CoRE Metal-organic Frameworks
- MD Trajectories of $C_7O_2H_{10}$

Repositories

- MDF (CHiMaD)
- MATIN
- Materials Commons
- CXIDB
- MML Repository (NIST)

Unified search across repositories, datasets, and DBs

Problem: Many materials data repositories, databases, and datasets exist but interfaces and access to them are fragmented

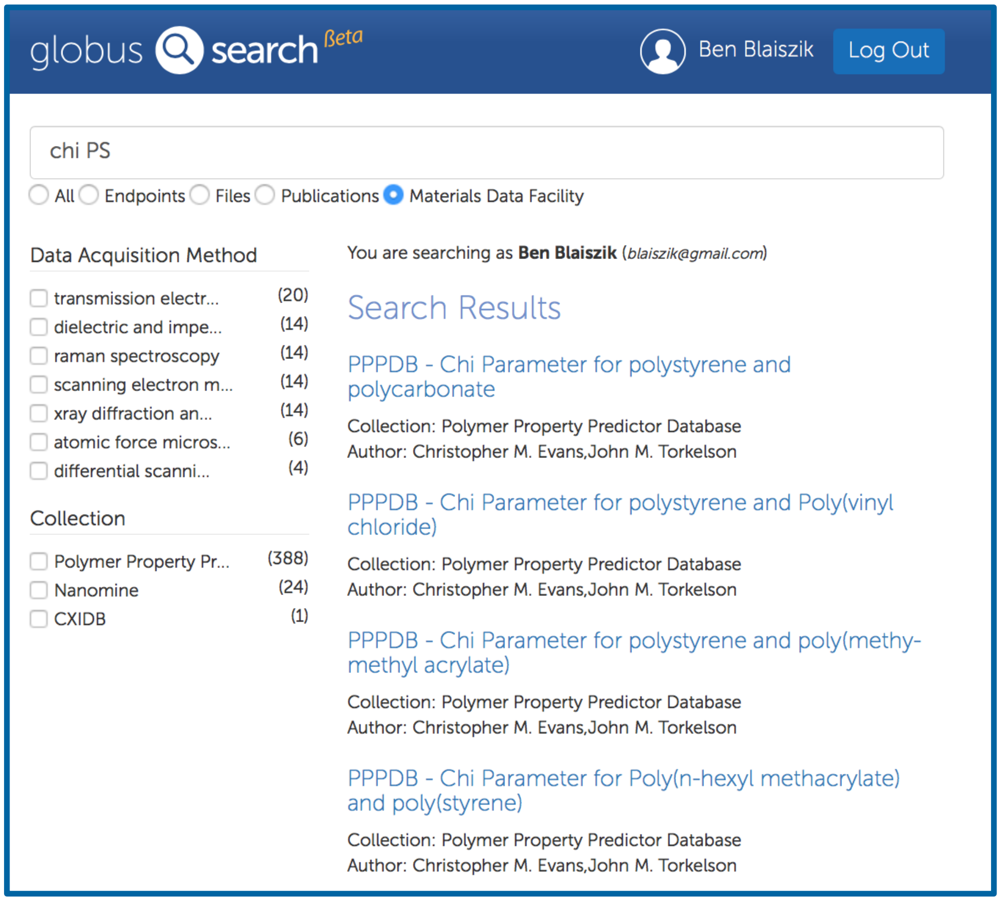


GitHub [Contribute a harvester (coming soon)](https://github.com/materials-data-facility/)
https://github.com/materials-data-facility/

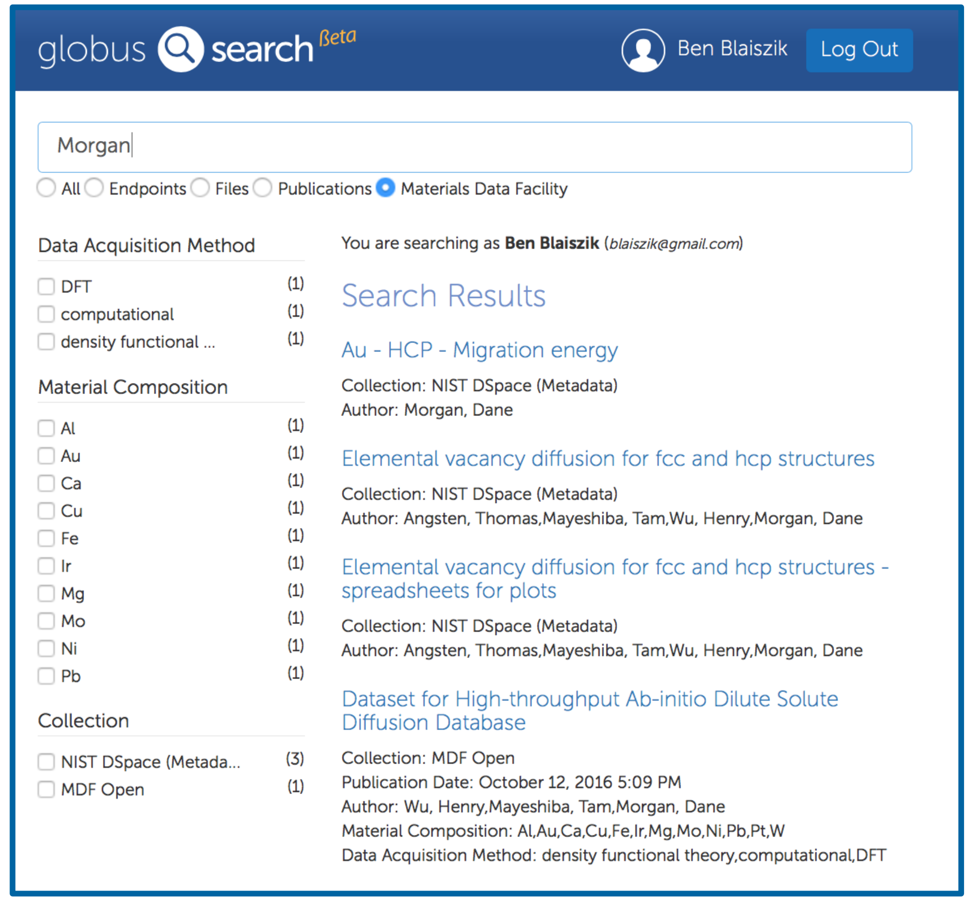
In this example: NanoMine, PPPDB, MML Repository (NIST), MATIN, Materials Commons (publications), CXIDB, linking with MRR…

Unified search across repositories, datasets, and DBs

Problem: Many materials data repositories, databases, and datasets exist but interfaces and access to them are fragmented



Results from PPPDB

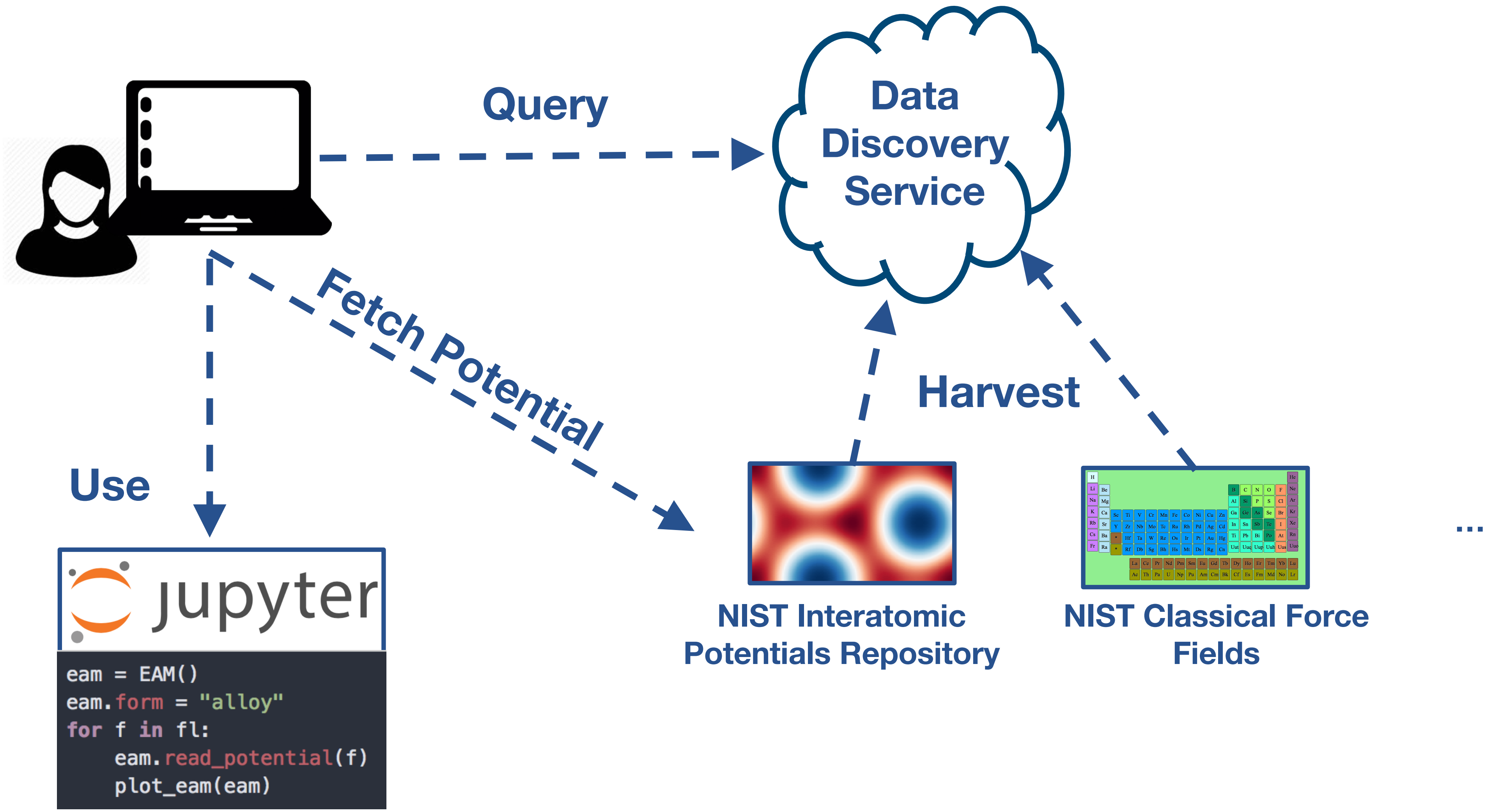


Results from MML Repository and MDF

In this example: NanoMine, PPPDB, MML Repository (NIST), MATIN, Materials Commons (publications), CXIDB, linking with MRR…

Finding and using interatomic potentials

Problem: Materials data exist in disparate locations, but finding and importing them into analysis scripts is time consuming



In this example: NIST Interatomic Potentials Repository, NIST Classic Potentials

Finding and using interatomic potentials

Problem: Materials data exist in disparate locations, but finding and importing them into analysis scripts is tricky

Search for potentials in NIST Interatomic Potentials Repo

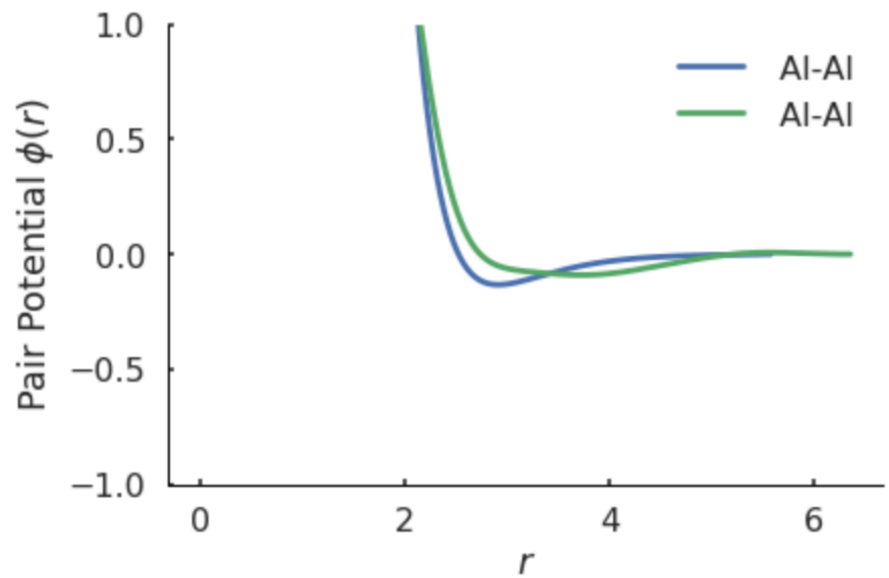
```
search_domain = "globus_search"
client = globus_auth.login("https://datasearch.api.demo.globus.org/", search_domain)

r = client.search("+eam aluminum^2")
```

Format the discovered potentials

```
fl = get_potentials(r, 2)
```

Load potentials into ASE and plot

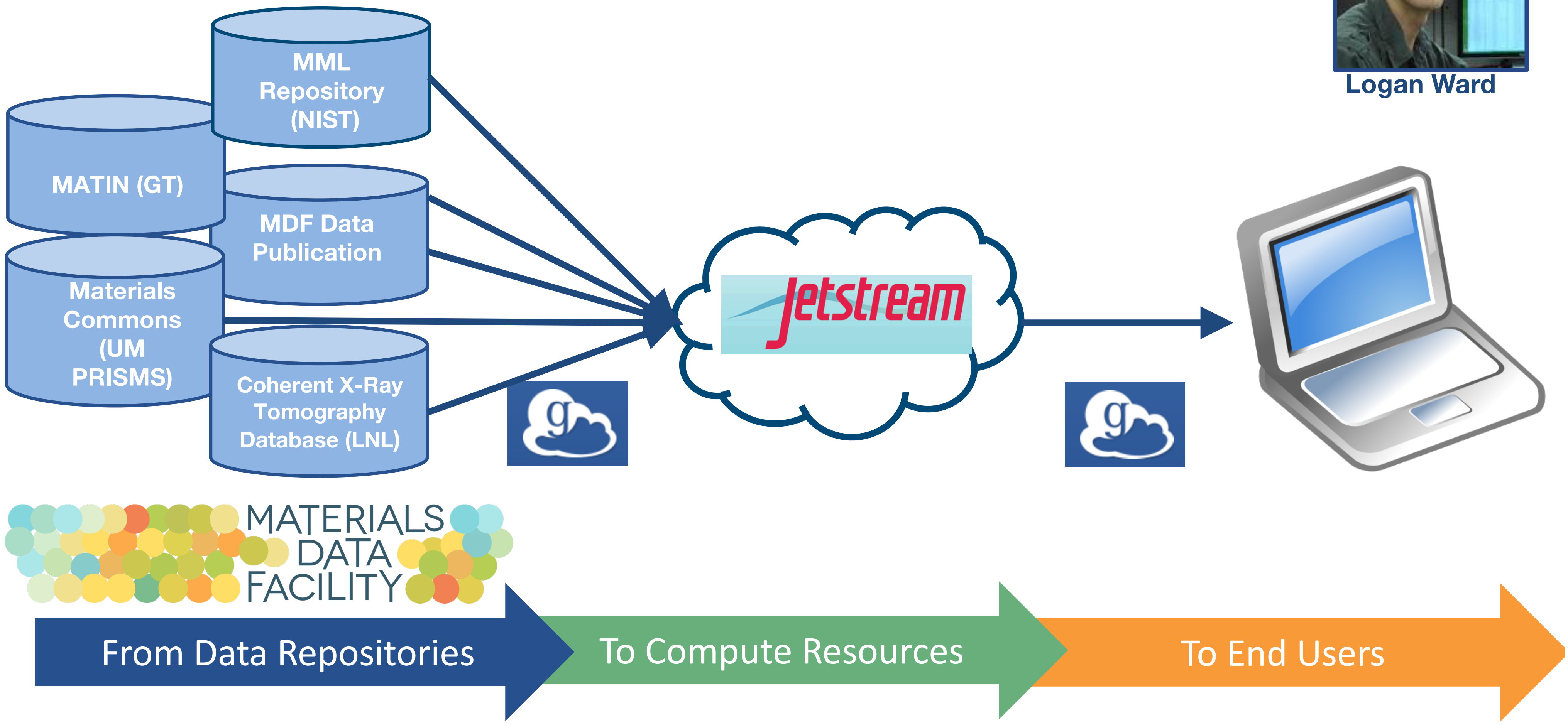
```
eam = EAM()
eam.form = "alloy"
for f in fl:
    eam.read_potential(f)
    plot_eam(eam)
```



In this example: NIST Interatomic Potentials Repository, NIST Classic Potentials

Integrating analytics tools with MDF

Logan Ward



<u>Result</u>: Scientists connected with data, analytics tools, and compute capability

Jetstream is a self-provisioned, scalable science and engineering cloud environment operated by Indiana University for the National Science Foundation: jetstream-cloud.org

Building a machine learning model using MDF

A simple web service to simplify AGNI model building

agni.demo.materialsdatafacility.org

demo | Autofit

MATERIALS DATA FACILITY

Welcome to Autofit!

This webpage allows you to build potentials using the AGNI method using data accessible through the MDF.

Step 1: Select Data

Use DataSearch to find datasets to use as a training set

- Elemental vacancy diffusion for fcc and hcp structures
- A user guide for SLUSCHI: Solid and Liquid in Ultra Small Coexistence with Hovering Interfaces
- Machine Learning Force Fields: Construction, Validation, and Outlook

Step 2: Define Potential Settings

Set both the simple and advanced setting of the AGNI method

Simple Settings

These are settings that are most likely to be changed

Training Points

500

Cutoff Distance

8

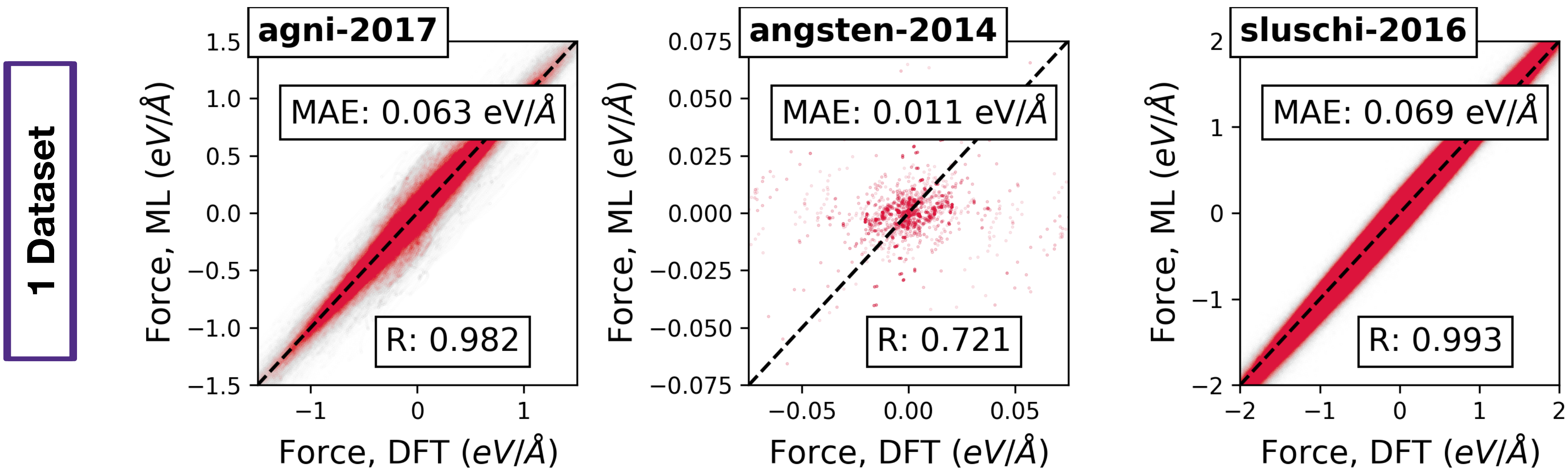
Building a machine learning model using MDF

Example: Building force-field potentials from different datasets

Data resources: 3 DFT datasets with Al data

1 dataset from khazana.uconn.edu, 2 datasets from materialsdata.nist.gov

Result: Improved performance by integrating data sources



Method: Botu *et al.* JPCC. (2017)

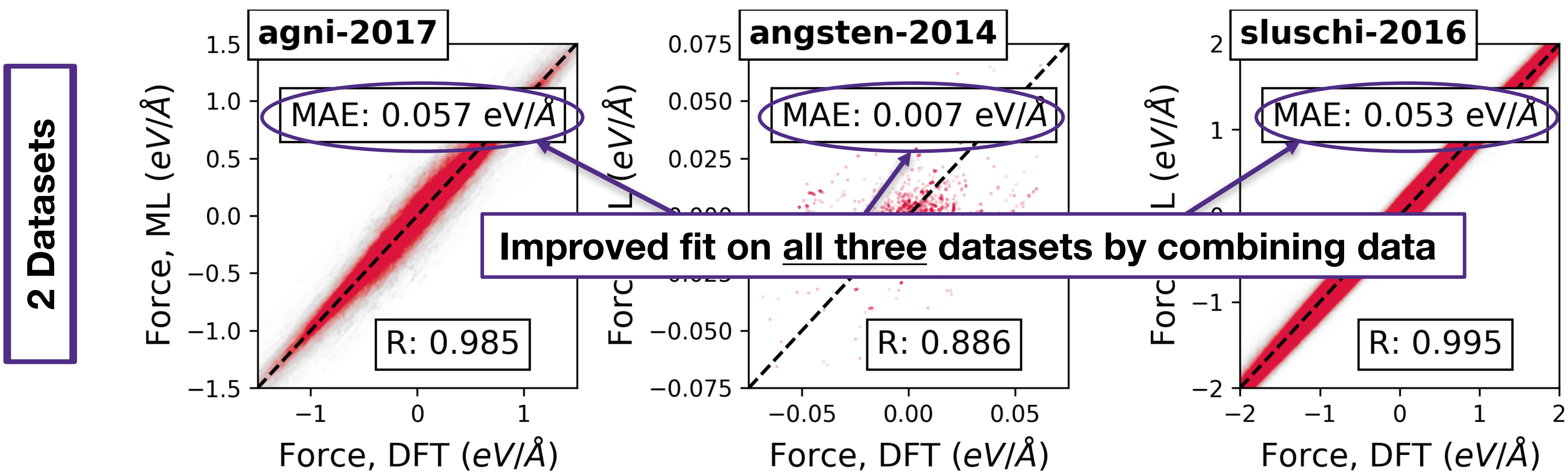
Building a machine learning model using MDF

Example: Building force-field potentials from different datasets

Data resources: 3 DFT datasets with Al data

1 dataset from khazana.uconn.edu, 2 datasets from materialsdata.nist.gov

Result: Improved performance by integrating data sources



Method: Botu *et al.* JPCC. (2017)

Recap and future work

To streamline materials data sharing, discovery, access, and analysis, we have:

- **Simplified data publication**, enabling ingest of 7 TB in 31 datasets from 11 institutions and 94 authors
- **Automated data and metadata capture** for deep indexing of large data collections, reaching 1M records from 16 sources
- **Unified search** across all of these sites and collections
- **Deployed APIs** enabling programmatic access

- We have demonstrated automated end-to-end discovery, access, and analysis pipelines

Next steps: More data, more indexing, richer search, more analyses

Thanks to our sponsors!

NIST

CHIMaD
Center for Hierarchical Materials Design

U.S. DEPARTMENT OF
ENERGY

ElectroCat
Electrocatalysis Consortium

NSF

NATIONAL INSTITUTES OF HEALTH

powered by
amazon
web services

Argonne
NATIONAL LABORATORY

THE UNIVERSITY OF
CHICAGO