MATERIALS DATA INFRASTRUCTURE

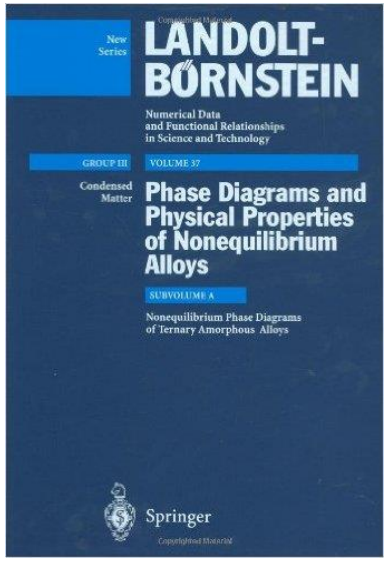
Logan Ward
Asst. Computational Scientist
Argonne National Laboratory

6 March 2021

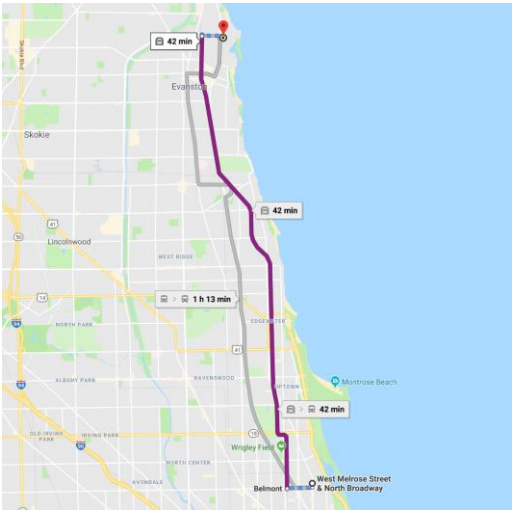
Good Data is Hard to Find

Even well-curated data can require effort to use with AI

Good Data



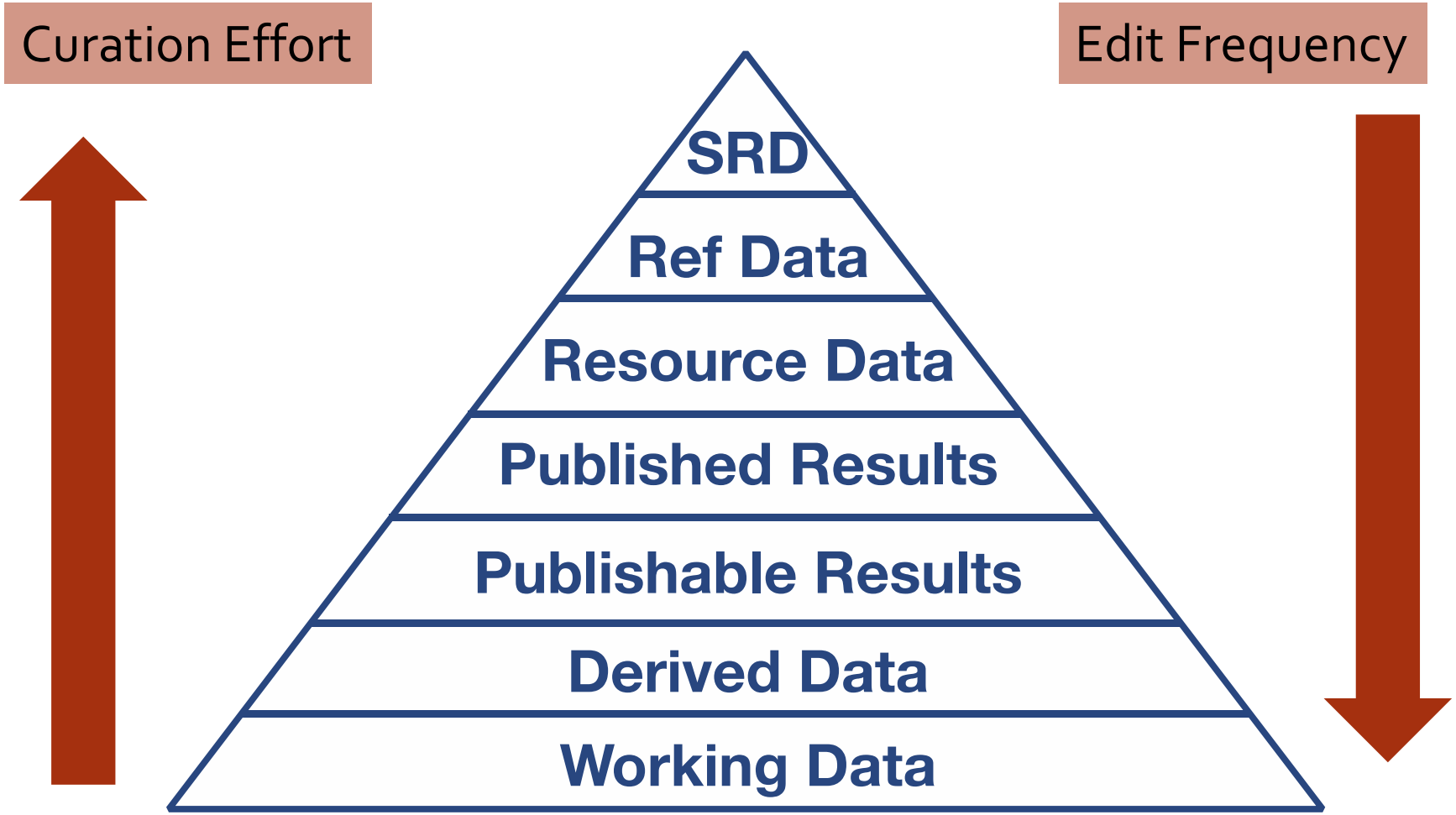
Long Commute



Usable Data

Entry	Diagram	Comp_1	Frac_1	Comp_2	Frac_2	Comp_3	Frac_3	Class	Check_Fra
1	1	Ag	20	Al	25	La	55	AM	Check
2	2	Ag	15	Al	10	Mg	75	AM	Check
3	2	Ag	25	Al	10	Mg	65	AM	Check
4	2	Ag	25	Al	20	Mg	55	AM	Check
5	2	Ag	35	Al	10	Mg	55	AM	Check
6	2	Ag	35	Al	20	Mg	45	AM	Check
7	2	Ag	45	Al	20	Mg	35	AM	Check
8	3	Ag	10	Ce	6	Cu	84	AM	Check
9	3	Ag	10	Ce	10	Cu	80	AM	Check
10	3	Ag	15	Ce	6	Cu	79	AM	Check
11	3	Ag	20	Ce	6	Cu	74	AM	Check
12	3	Ag	20	Ce	10	Cu	70	AM	Check
13	3	Ag	25	Ce	6	Cu	69	AM	Check
14	3	Ag	30	Ce	6	Cu	64	AM	Check
15	3	Ag	30	Ce	10	Cu	60	AM	Check
16	3	Ag	35	Ce	4	Cu	61	AM	Check
17	3	Ag	35	Ce	5	Cu	60	AM	Check
18	3	Ag	35	Ce	6	Cu	59	AM	Check
19	3	Ag	40	Ce	3	Cu	57	AM	Check
20	3	Ag	40	Ce	4	Cu	56	AM	Check
21	3	Ag	40	Ce	5	Cu	55	AM	Check
22	3	Ag	40	Ce	8	Cu	52	AM	Check
23	3	Ag	40	Ce	10	Cu	50	AM	Check
24	3	Ag	45	Ce	3	Cu	52	AM	Check
25	3	Ag	45	Ce	4	Cu	51	AM	Check

No single solution to data management



Different Types of Data == Different Database Requirements!

Borrowed with appreciation from NIST

Working Data: Close to the Scientist

[Data] scientists need…

1. Unrestricted access to data
2. Portability
3. Easy use from other tools
4. Ability to share with collaborators

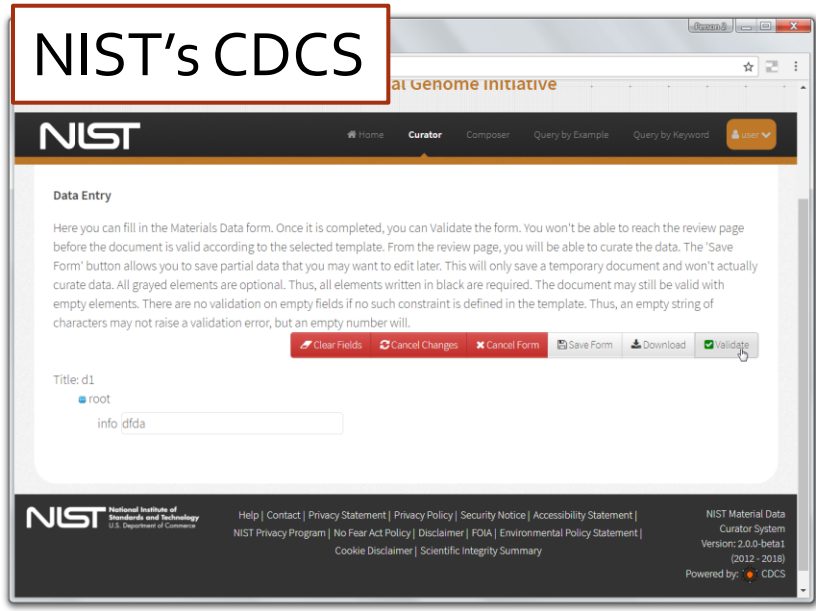
Everyone has their own workflow

<u>Key challenge is achieving flexibility while also limiting chaos</u>

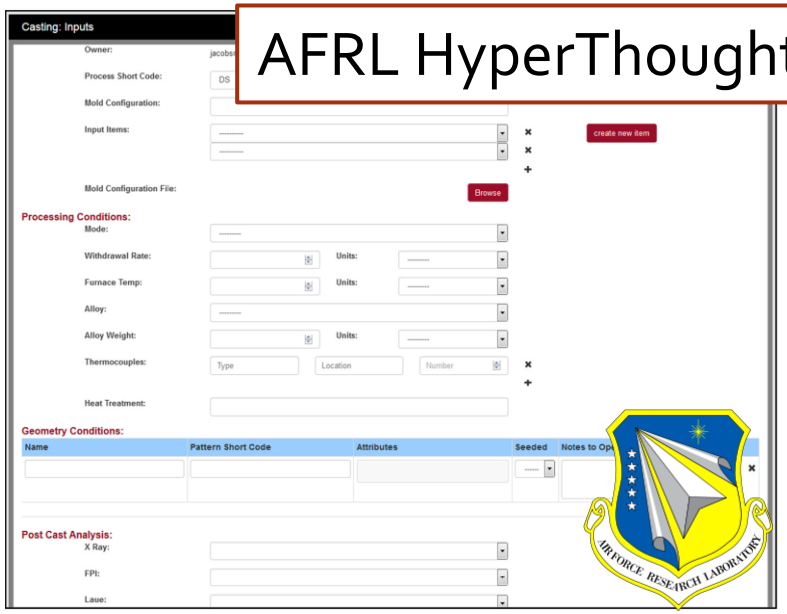
Creating usable data management systems is a huge, and well-studied problem

Laboratory Inventory Management (US)

NIST's CDCS

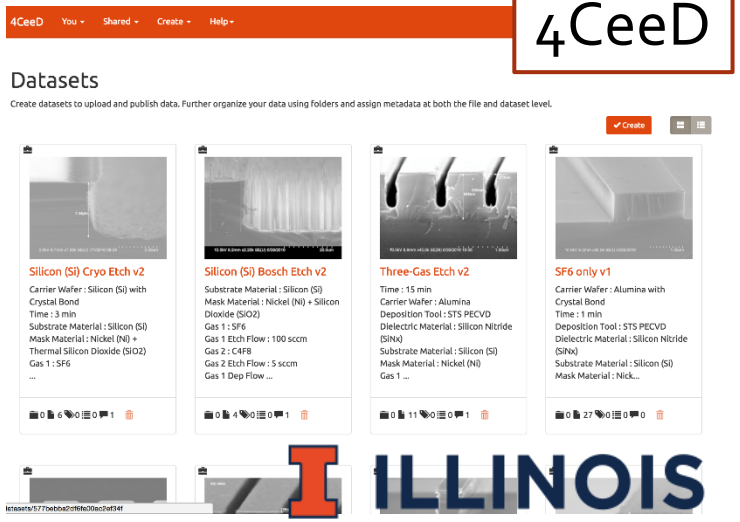


AFRL HyperThought



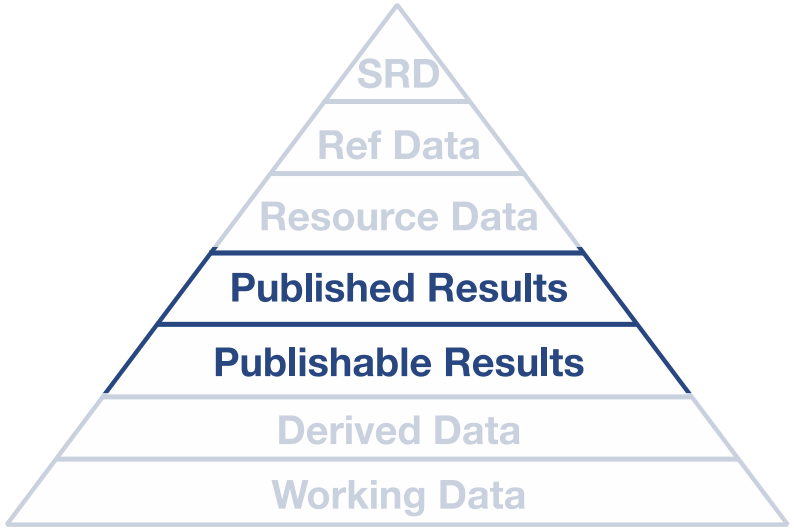
Materials Commons

4CeeD



“Sharable” Data and Publication

Need: “Publish and Forget”



Requirements:

1. Provenance Information
2. Archival Storage
3. Detailed Descriptions
4. Rewards for Data Publication

Common Features of All Services

There are Plenty of publication services

MATERIALS DATA FACILITY

zenodo

Many services, some better for different kinds of data

DRYAD

MATERIALSCLOUD

NOMAD

What Does Published Data Look Like?

Basic Provenance Information

Data from: Charting the complete elastic properties of inorganic crystalline compounds

de Jong M, Chen W, Angsten T, Jain A, Notestine R, Gamst A, Sluiter M, Ande CK, van der Zwaag S, Plata JJ, Toher C, Curtarolo S, Ceder G, Persson KA, Asta M

Date Published: March 16, 2015

DOI: https://doi.org/10.5061/dryad.h505v

Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.

Title	**Elastic constant**
Downloaded	385 times
Download	ec.json (8.098 Mb)
Details	View File Details

When using this data, please cite the original publication:

de Jong M, Chen W, Angsten T, Jain A, Notestine R, Gamst A, Sluiter M, Ande CK, van der Zwaag S, Plata JJ, Toher C, Curtarolo S, Ceder G, Persson KA, Asta M (2015) Charting the complete elastic properties of inorganic crystalline compounds. Scientific Data 2: 150009. https://doi.org/10.1038/sdata.2015.9

Additionally, please cite the Dryad data package:

de Jong M, Chen W, Angsten T, Jain A, Notestine R, Gamst A, Sluiter M, Ande CK, van der Zwaag S, Plata JJ, Toher C, Curtarolo S, Ceder G, Persson KA, Asta M (2015) Data from: Charting the complete elastic properties of inorganic crystalline compounds. Dryad Digital Repository. https://doi.org/10.5061/dryad.h505v

Cite | Share

Ward, Logan; O'Keeffe, Stephanie C.; Stevick, Joesph; Jelbert, Glenton R.; Aykol, Muratahan; Wolverton, Chris, "A Machine Learning Approach for Engineering Bulk Metallic Glass Alloys, 2018, http://dx.doi.org/doi:10.18126/M2662X

Download Citation

Title:	A Machine Learning Approach for Engineering Bulk Metallic Glass Alloys
Authors:	Ward, Logan O'Keeffe, Stephanie C. Stevick, Joesph Jelbert, Glenton R. Aykol, Muratahan Wolverton, Chris
Keywords:	Machine learning Bulk Metallic Glasses
Appears in Collections:	MDF Open

Links to Files

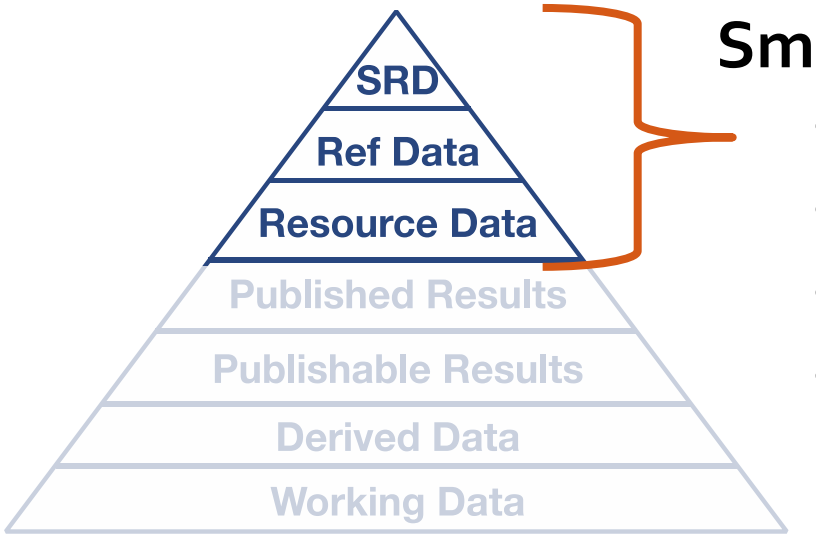
Endpoint and path to dataset

82f1b5c6-6e9b-11e5-ba47-22000b92c6ec/published/publication_1106/

Show full record | Return to data publication dashboard

Data is Available (!), But Only Usable by Humans

Reference Data: What People Want

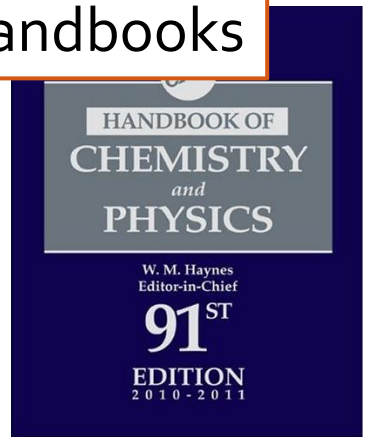


Smallest fraction of data. Typically…

- Extensively curated
- Composed of many experiments
- Specific goal of collection
- Consistent format (schema)

Accordingly, reference data are most widely used and usable

Handbooks



Web Databases

OQMD

Web APIs

```
a.get_in_chemsys(
    ['Ca', 'O']
)
```

What is special about "reference data"?

Al-B-Fe 612

Table 12.

1. Radio frequency melting
2. Melt-spinning
3. Ar-gas
4. Ribbon width × thickness:
1-2mm × 0.02-0.03mm
5. TEM; XRD

81INOU2

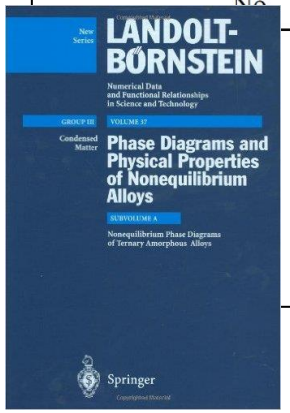
cont.

Table 12. (cont.)

Comment:

See Al-B-Co system.

No.	Al	B	Fe	Phase
	0	13	87	AM
	0	18	82	AM
	0	22	78	AM
	4	14	82	AM
	4	18	78	AM
	4	22	74	AM
	8			
	8			



Large amount of data

Curated metadata

Link to original source

Data in Tabular Form

Such data is rare, yet a requirement for ML

What defines “structured” data and a web APIs?

Structured Data

What makes data “structured?”

A defined “data model”

```
{
 “number”: 5735,  “direction”: “S”
 “street”: “Ellis Ave.”, “zip”: 60637
}
```

Why is this good?

- Predictability: Write simpler code
- Documentation: Understand what you write

Web APIs

Common Approach: REST API

Key features:

- Access a website via HTTP requests
- Send/receive structured data

Illustrative Example: Materials API (MAPI)

```
GET …/rest/v1/materials/24972/vasp
Response: {"valid_response": true,
"response": [{
  "formation_energy_per_atom": -1.833,
  "elements": ["O", "Fe“],
```

Reference Data: A Bright Future

Commercial/Industrial

Academic/National Laboratory

Bulk Metallic Glasses

ID: 156839 - Version 1 - Create new version - Edit - Delete

Description:

Dataset associated with "A Machine Learning Approach for Engineering Bulk Metallic Glass Alloys." Contains the glass-forming ability (either bulk, ribbon, or none), critical casting diameter, supercooled liquid range, and glass transition temperature for many metallic alloys.

Show Less

This dataset has not been used in any views.

Search this dataset

Material Name or Chemical Formula

Property Name

Units

Advanced Search Options

Search

Showing results 1 to 24 of 7093

$B_{12}Fe_{78}Mn_{10}$

Chemical formula: $B_{12}Fe_{78}Mn_{10}$

Glass forming ability: None

CITRINE INFORMATICS

Polymer Genome

An informatics platform for accurate property estimates of existing and hypothetical polymers using machine learning models trained on a benchmark dataset

Home Guide Dataset Summary ML Performance Sign-in

Draw Polymer

Polymer name, repeat unit, SMILES ...

Predict Properties

Predict Solvent

repeat units or SMILES strings.

Reference Databases are Proliferating Rapidly!

OQMD

ChemAxon

ChemSpider

The free chemical database

GRANTA

MATERIAL INTELLIGENCE

Materials Platform for Data Science

MatNavi

NIMS Materials Databa

NanoMine Data Resource

Integrated Web Interface & Data Exchange

Database | Analysis Tools | Simulations

Curation | Exploration | Visualization | Dissemination

Microstructure Characterization & Reconstruction | Interphase tools | Image Processing | Data Mining & Analytics

Dispersion State & Microstructure | Interphase Model | Physical Property Simulations

Processing | Structure | Properties

UHCSDB microstructure explorer

What Are the Trends?

- Data is Getting Published!
- Repositories are Digital
- Efforts are Community Driven

What Are the Major Trends?

1. ~~Data is Getting Published~~ **Deluge of Data**
   - Data Management Systems seldom used
   - Publication repositories lack metadata

2. ~~Repositories are Digital~~ **APIs are Uncommon**
   - Tools Do Not Work with Databases

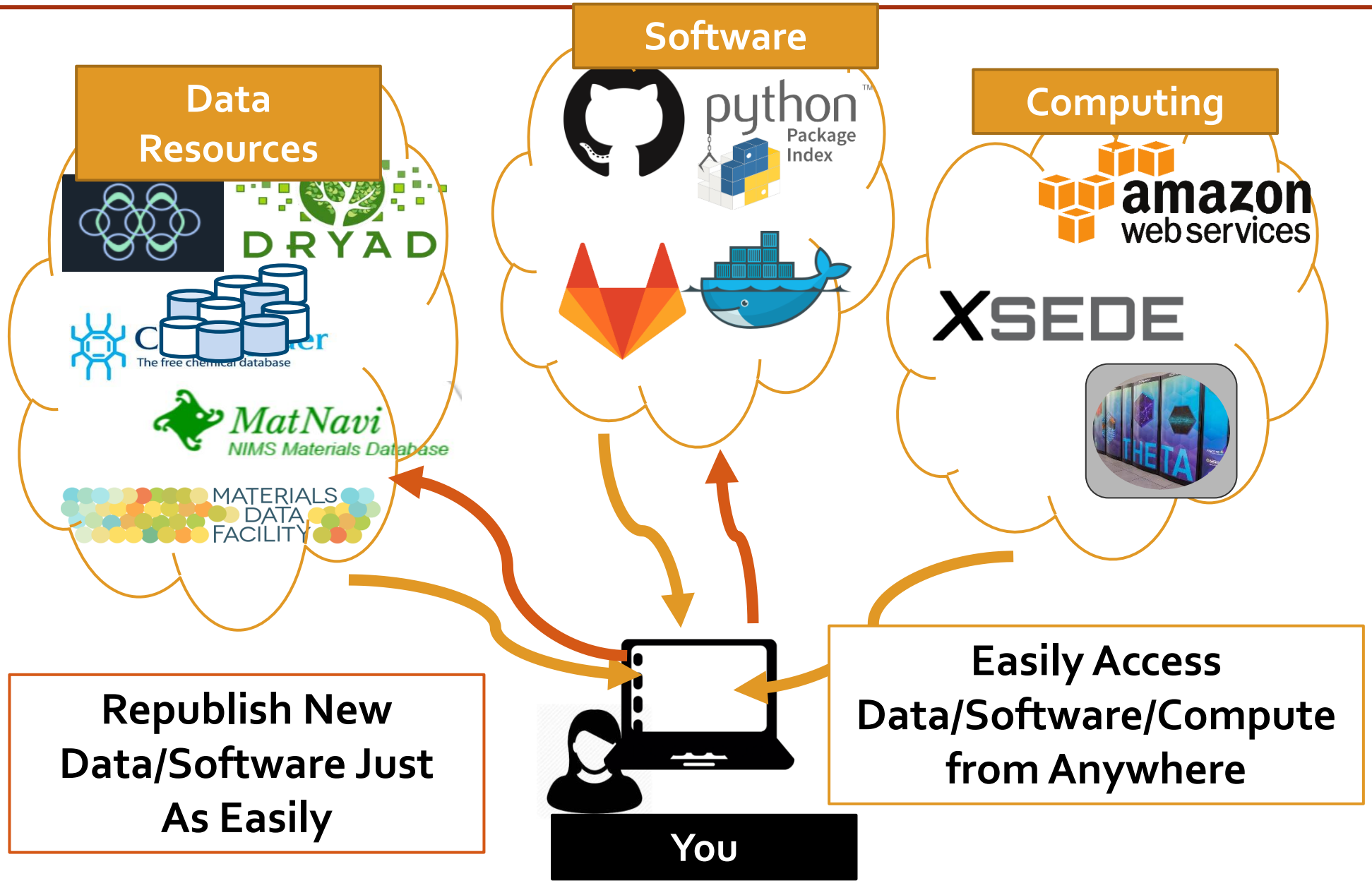
3. ~~Efforts are Community Driven~~ **Many Silos**
   - Finding Best Dataset Difficult

Current State: Data and Tools are Available
What would make things better?

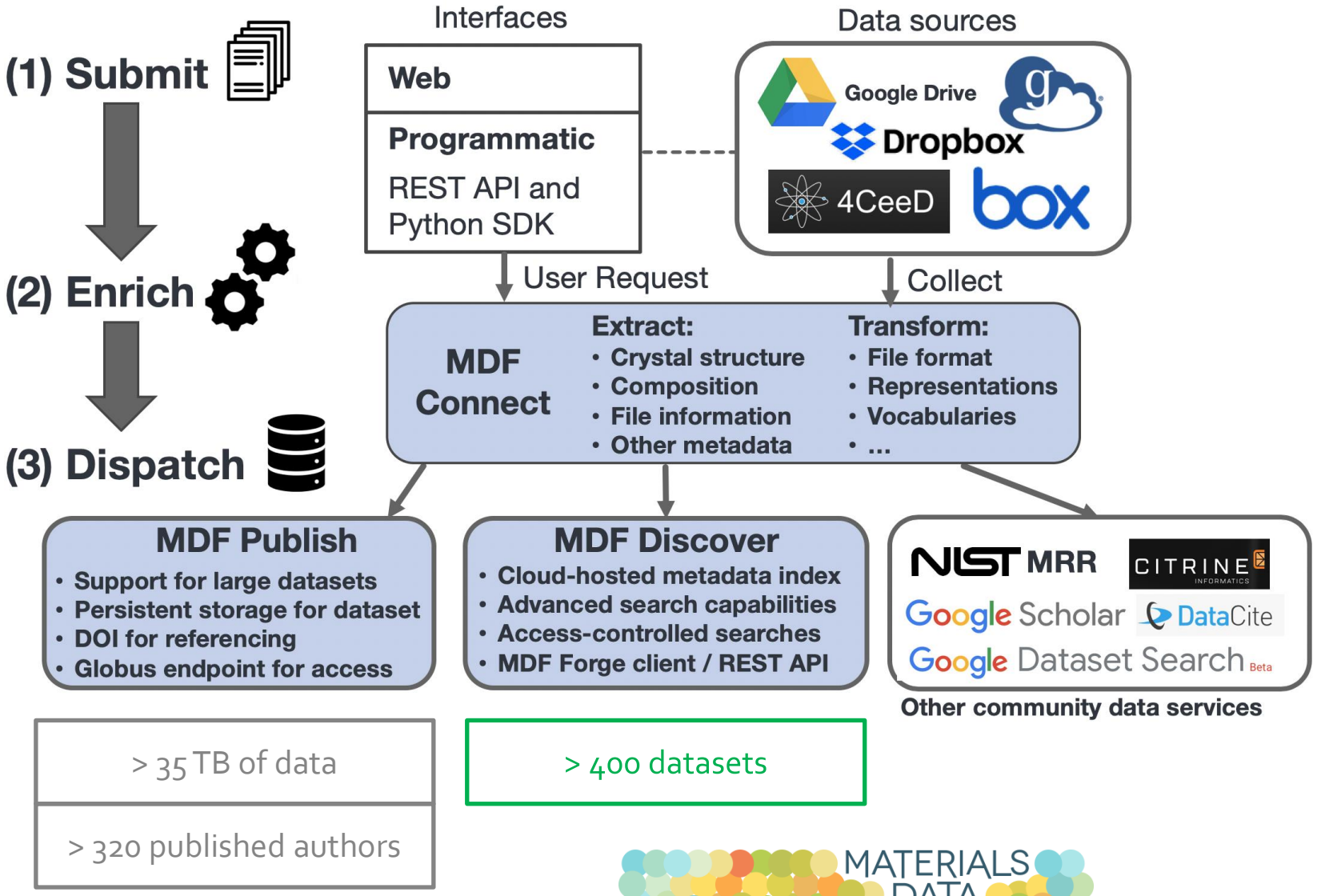
WHAT IS THE PATH FORWARD?

PSA: I work with the Materials Data Facility

A Seamless Data Infrastructure



The Materials Data Facility (MDF)



- **Connect:** Extract domain-relevant metadata / transform the data
- **Publish:** Built to handle big data (many TB, millions of files), provides persistent identifier for data, distributed storage enabled
- **Discover:** Programmatic search index to aggregate and retrieve data across hundreds of indexed data sources

MATERIALS DATA FACILITY

https://www.materialsdatafacility.org

DLHub – A Data and Learning Hub for Science

Describe

- Specify the model files
- Mark up the model with information to make it discoverable and usable

```
from dlhub_sdk.models.servables.keras import KerasModel
m = KerasModel.create_model("p1b1-example.h5")

m.set_title("CANDLE Pilot 1 - Benchmark 1")
m.set_name("candle_p1b1")
m.set_domains("genomics","biology","HPC")
```

Publish

- Register with DLHub for containerization as a servable
- DLHub service creates unique endpoint for servable

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
dl.publish(m)
```

Discover

- Discover servables with advanced search capabilities through Python SDK or web UI

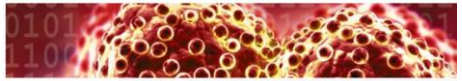
Run

- Make predictions by sending data to DLHub and specifying the servable to use

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
pred = dl.run("candle_p1b1", data)
```

Exascale Cancer Research

CANDLE

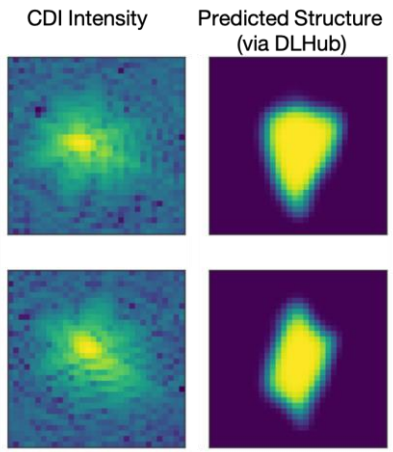


X-Ray Science

- Predict structure and phase of a material given coherent diffraction intensity
- Data available from Github

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()

struct = dl.run("cherukara_structure", X)
```



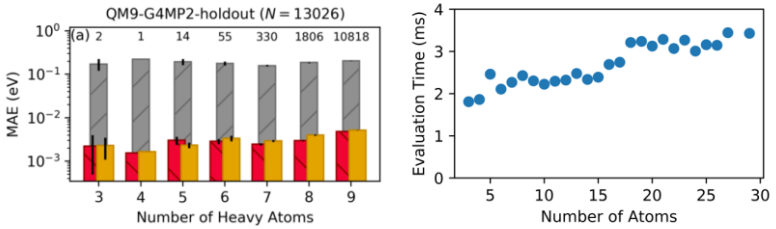
Cherukara et al., 2018

Energy Storage

- Predict molecular energies with G4MP2 accuracy at B3LYP cost
- Data available in MDF

Machine Learning Prediction of Accurate Atomization Energies of Organic Molecules from Low-Fidelity Quantum Chemical Calculations

Logan Ward[1,2], Ben Blaiszik[1,3], Ian Foster[1,2,3], Rajeev S. Assary[4,5], Badri Narayanan[5,6], Larry Curtiss[4,5]

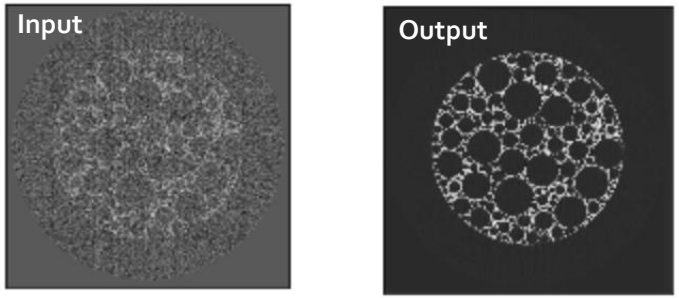


Tomography

- Enhance tomographic scans and remove noise using generative adversarial model
- Example data available on Petrel

TomoGAN: Low-Dose X-Ray Tomography with Generative Adversarial Networks

Zhengchun Liu, Tekin Bicer, Rajkumar Kettimuthu, Doga Gursoy, Francesco De Carlo, Ian Foster



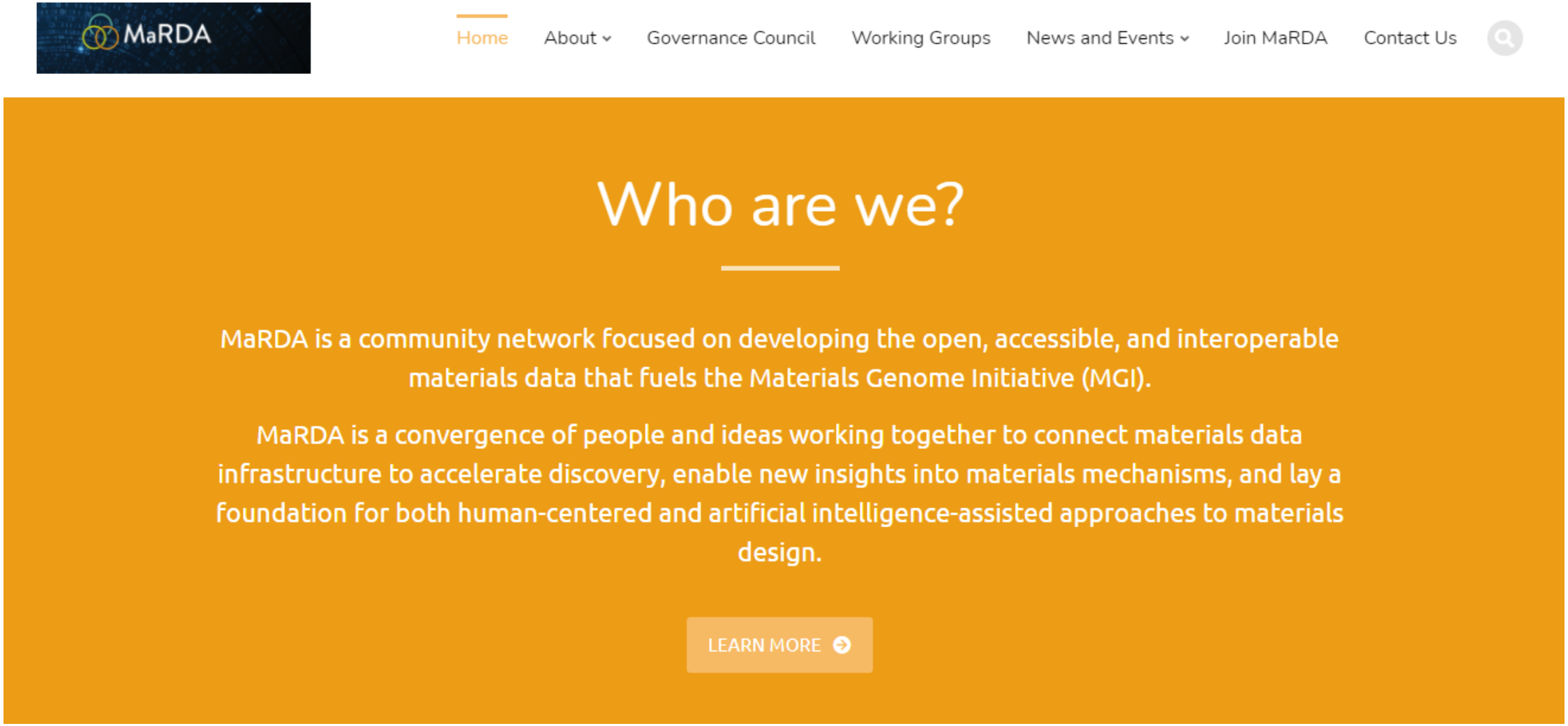
Argonne NATIONAL LABORATORY

ARGONNE LEADERSHIP COMPUTING FACILITY

PETREL

parsl

Growing communities to link resources



https://www.marda-alliance.org/

Conclusions

What should I know about materials databases?

- What are the challenges? **Depends on the type of data**
  - Working data: Laboratory Inventory Management Systems (LIMS)
  - Published data: "Publish and Forget" data systems
  - Reference data: Structured data and web APIs
- How are people trying to solve them? **Community of database software**
- What is still broken? **Linking databases together and to compute**