A Data Ecosystem to Support Machine Learning in Materials Science

Ben Blaiszik (blaiszik@uchicago.edu), Jonathon Gaff, Logan Ward, Ryan Chard, Kyle Chard, Steven Tuecke, Ian Foster

MATERIALS DATA FACILITY

DLHub
Data and Learning Hub for Science
https://www.dlhub.org

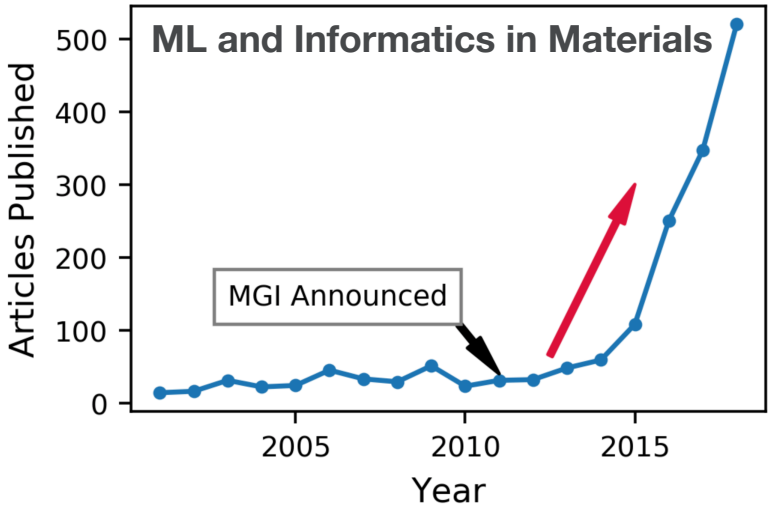
https://www.materialsdatafacility.org

https://www.dlhub.org

A Growing Opportunity in Machine Learning in the Sciences

ML in Science

- Number of publications across domains is growing rapidly
- Access to datasets is improving, but still a challenge
- Access to models and codes is a particular challenge
- Benchmarks are lagging



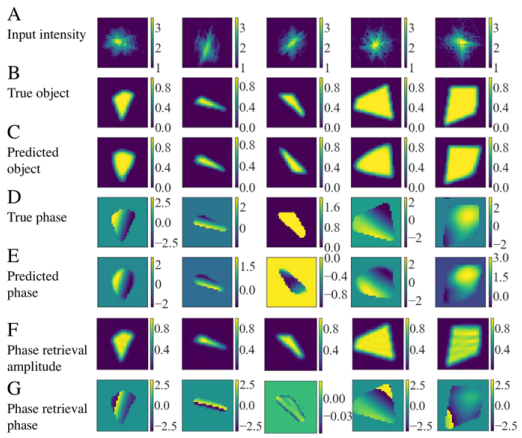
A Motivation

Article | OPEN | Published: 08 November 2018

Real-time coherent diffraction inversion using deep generative networks

Mathew J. Cherukara, Youssef S. G. Nashed & Ross J. Harder

Scientific Reports **8**, Article number: 16520 (2018) | Download Citation



Model ??

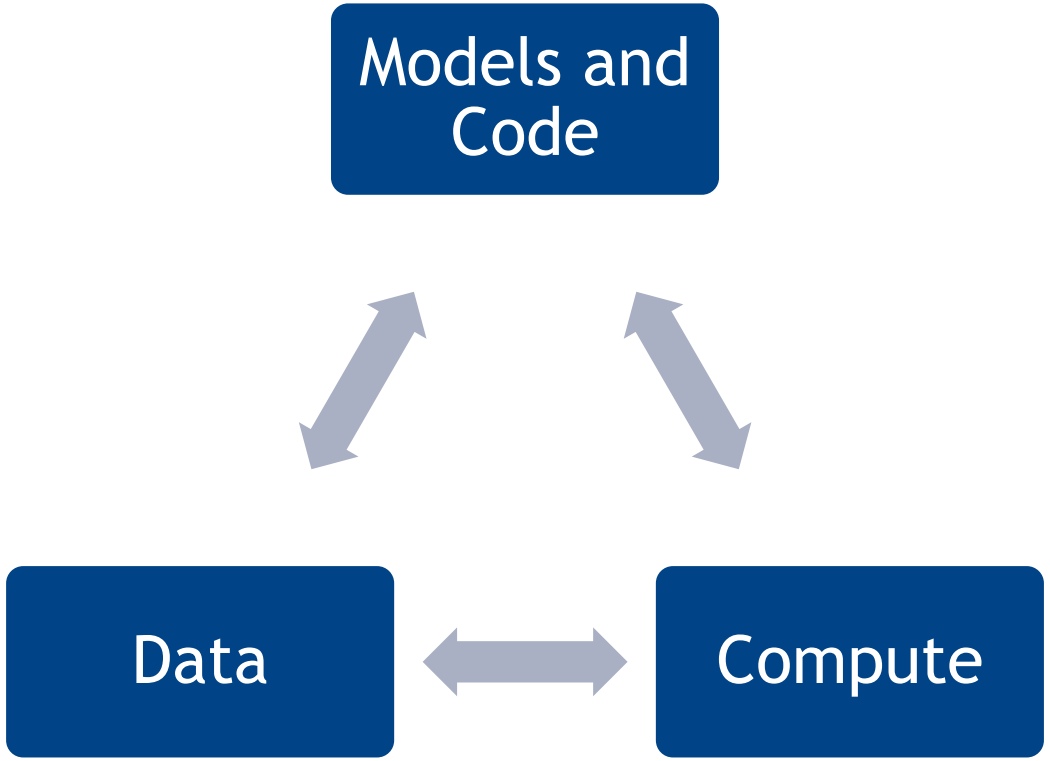
Code ???

Data ????

An Emerging Data Infrastructure Ecosystem

We need and can build a **cohesive infrastructure** for AI-driven science

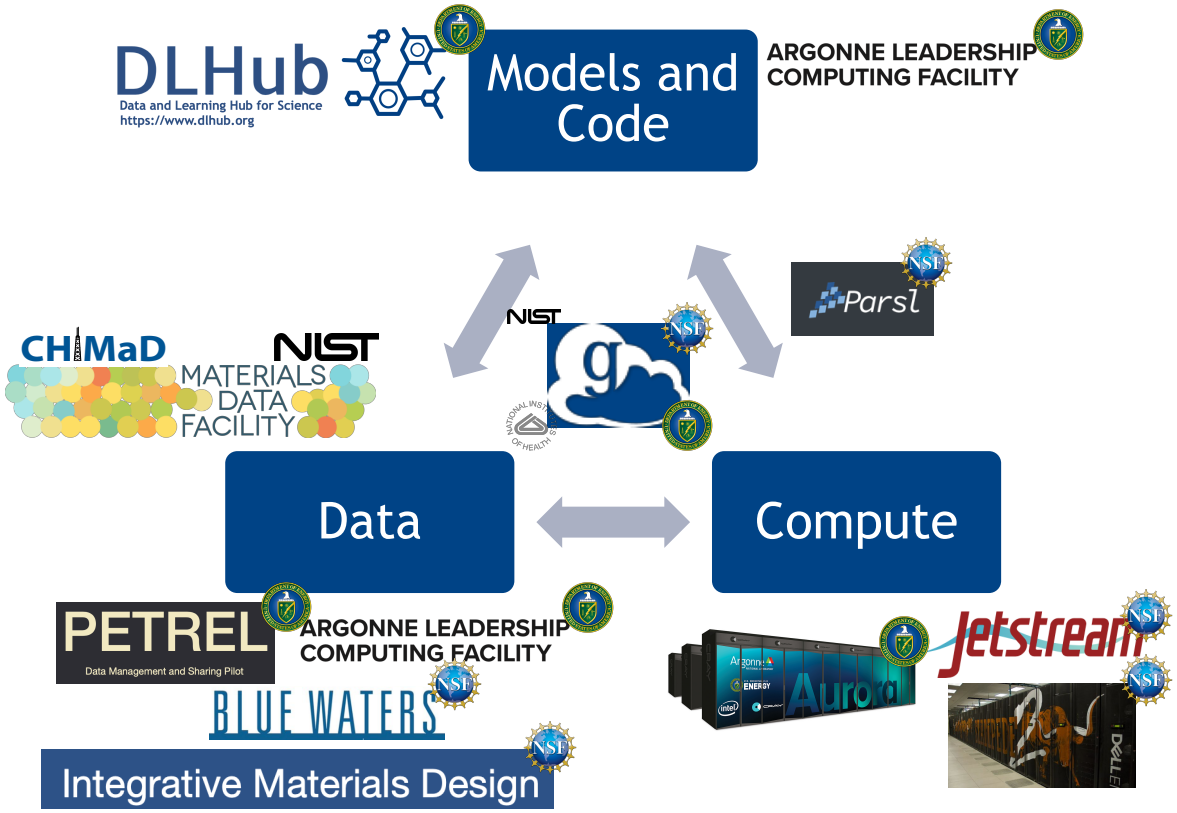
- Data: capture, organize, analyze, share move and deliver data
- Compute: accessible at many scales, types, and locations
- Models: discoverable, described, and portable to wherever data and/or computer are located
- Workflows: easily discovered, adapted, composed, scaled, and reused



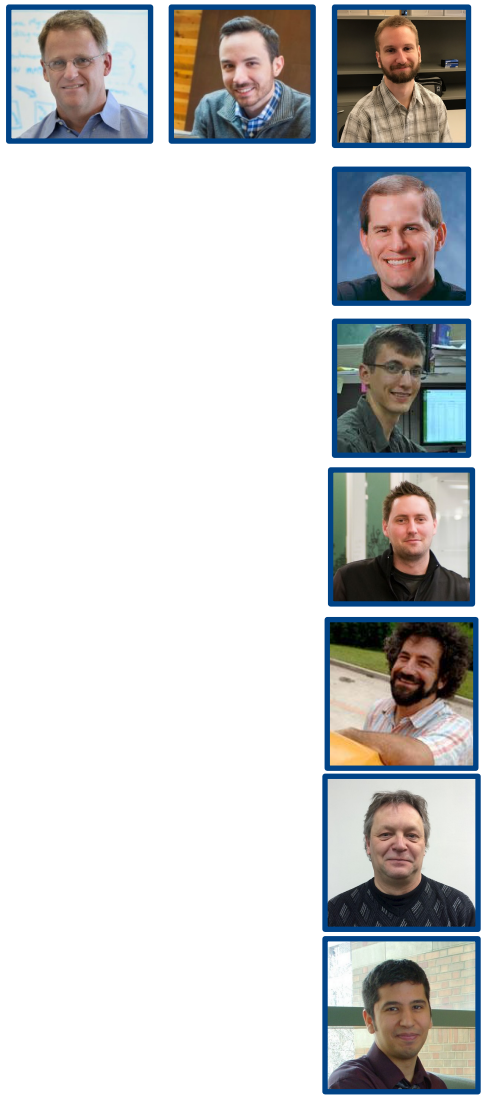
An Emerging Data Infrastructure Ecosystem

We need and can build a **cohesive infrastructure** for AI-driven science

- Data: capture, organize, analyze, share move and deliver data
- Compute: accessible at many scales, types, and locations
- Models: discoverable, described, and portable to wherever data and/or computer are located
- Workflows: easily discovered, adapted, composed, scaled, and reused



Materials Data Facility (MDF) Overview

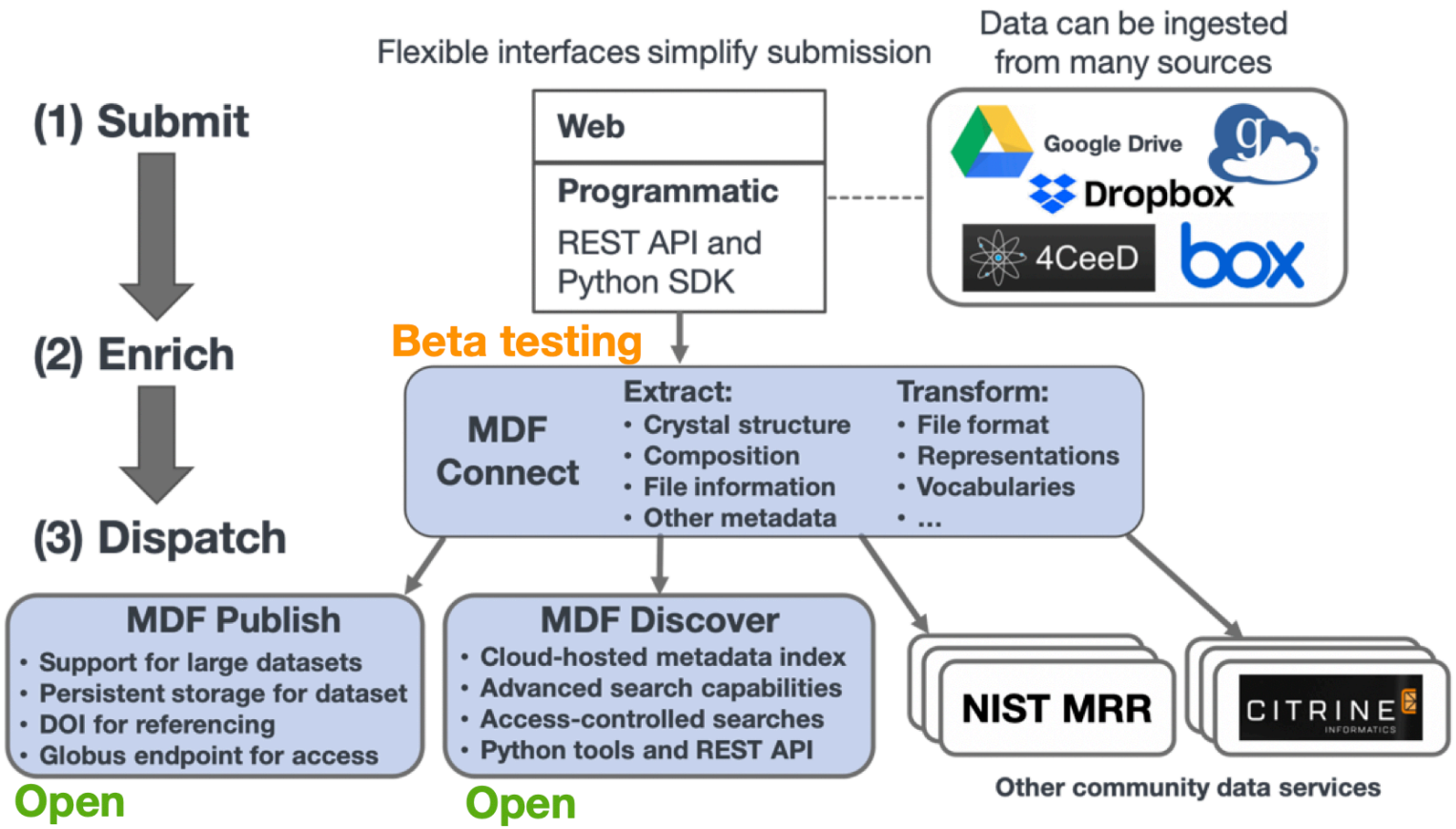


Build data services to

- Empower researchers to publish data, regardless of size, type, and location
- Automate data and metadata extraction and ingest
- Enable unified search and discovery across disparate materials data sources

Deploy with APIs to simplify connection to other data efforts and to enable automation

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
- Currently holds ~30TB of data from over 150 authors, millions of individual results

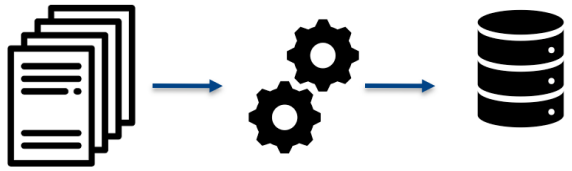
MDF - Automating Metadata Extraction

Key features

- Programmatic access to discover, access, and aggregate indexed data
- Data indexed from many sources
- Automated indexing of materials-specific information
- Goals:
  - apply extractions in an intelligent way
  - Build individual parsers on best-in-class parsers for each file type
  - provide reliable testing and benchmarking of parsers
  - Provide general framework for other domains
    - DFT
    - Thermo
    - CIF
    - STM
    - images
    - SEM/TEM
    - JSON/CSV/YAML
    - …

MATERIALS DATA FACILITY

MaterialsIO CITRINE INFORMATICS



build passing | docs passing | coverage 93%

MaterialsIO is a library of tools that generate summaries of the data contained in scientific data files. The goal of MaterialsIO is to provide a shared resources of these tool to avoid duplication of effort between the many emerging materials databases. Each parser is designed to generate the sum of all data needed by each of these databases with a uniform API so that specific projects can write simple adaptors for their needs.

Crystal Structure

```
from crystal_structure import ParseCrystalStructure
p = ParseCrystalStructure()
groups = p.group('./data/cif')
[p.parse(group) for group in groups]
```

```
[{'crystal_structure': {'number_of_atoms': 5070.0,
   'space_group_number': 146,
   'stoichiometry': 'A45B54C154D280E312',
   'volume': 101836.44086588411},
  'material': {'composition': 'Co270H1680C1872N3240924'}}]
```

```
from crystal_structure import ParseCrystalStructure
p = ParseCrystalStructure()
groups = p.group('./data/outcar')
[p.parse(group) for group in groups]
```

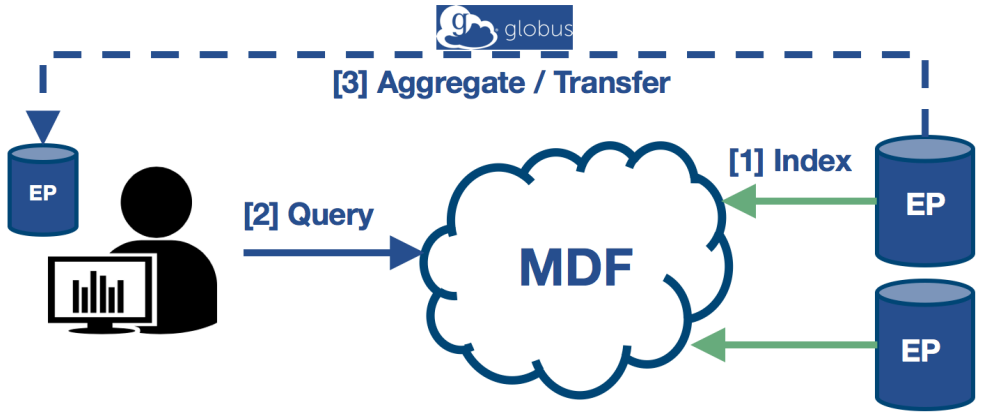
```
[{'crystal_structure': {'number_of_atoms': 53.0,
   'space_group_number': 160,
   'stoichiometry': 'AB52',
   'volume': 611.90620674445},
  'material': {'composition': 'Al1Fe52'}}]
```

https://www.materialsdatafaclity.org

CHiMaD NIST

MDF – Simplifying Data Discovery and Usage

Assemble Training Set



Compile records for a larger, mixed source, result set

```
elements = ["Al"]
sources = ["khazana_vasp", "sluschi", "ab_initio_solute_database"]
my_ep = "c8ee7e5c-6d04-11e5-ba46-22000b92c6ec"
my_path = "/Users/ben/Desktop/blaiszik-macbookpro/dft_training_set"

mdf = Forge()
res = mdf.search_by_elements(elements=elements, sources=sources, limit=9999)
mdf.get_globus(res, dest=my_path,
                local_ep=my_ep, preserve_dir=True)
```

```
Processing records: 100%|■■■■■■■■■■| 10/10 [00:00<00:00, 19.05it/s]
Submitting transfers: 100%|■■■■■■■■■■| 1/1 [00:00<00:00,  3.60it/s]

All transfers submitted
Submission IDs: 3fbfc637-7181-11e7-a9fd-22000bf2d287
```

Dataset Mixing

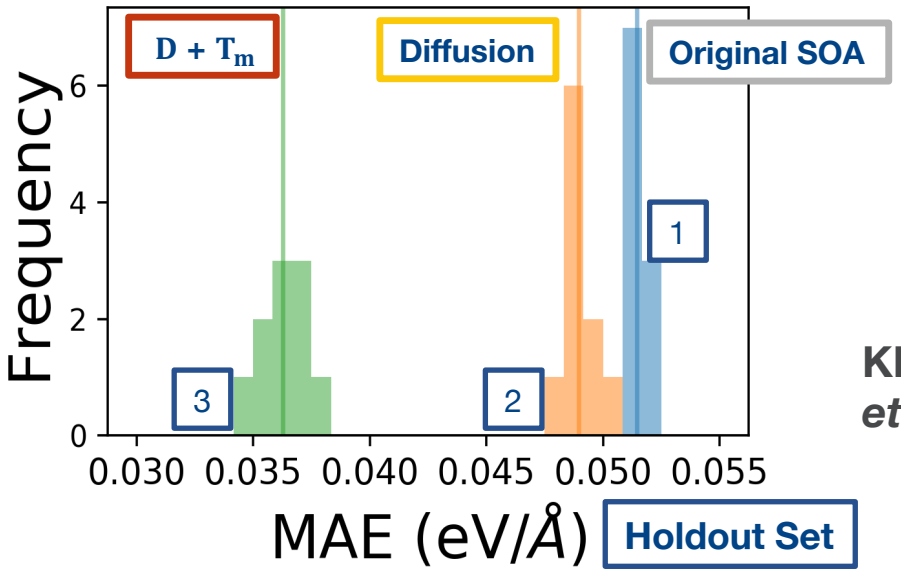
Building force-field potentials from different datasets

Data resources: 3 Aluminum DFT datasets

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

Result: Improved performance by integrating data sources

Logan Ward



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

Better performance in original application: No new DFT calculations

CHiMaD NIST

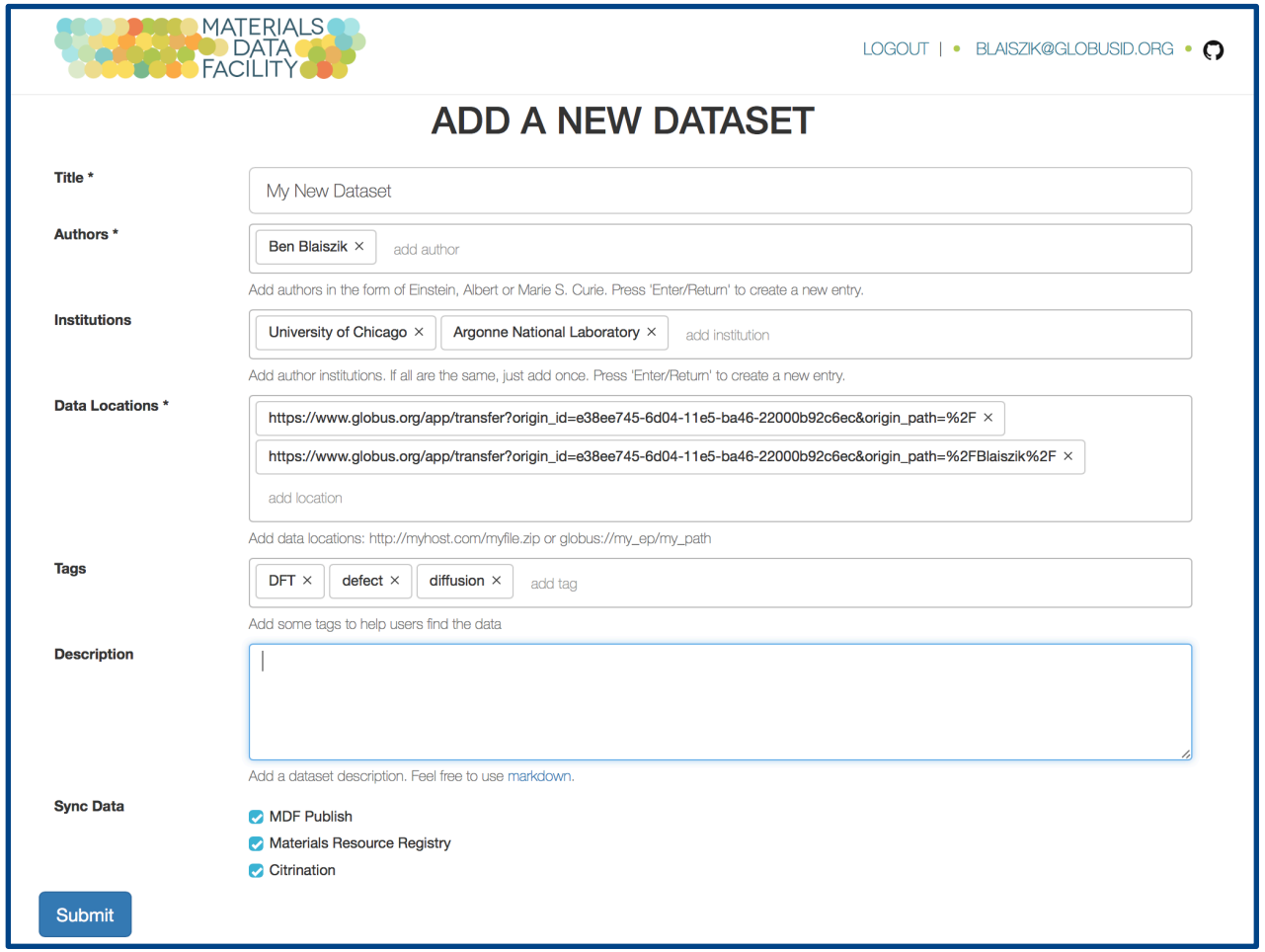
MDF – Enabling Flexible Data Publication

Key features

- Receive a citable DOI for your dataset
- Native support for large datasets
  - millions of files or many TB of data
- Distributed storage enabled
- Host data on high availability, reliable, performant storage (ALCF, UIUC) or your own storage cluster

- Currently hold over 30 TB of data from over 150 authors
- Free of charge for researchers

Publish via Web Interface



Or with a Python script

https://www.materialsdatafacility.org

CHiMaD NIST

MDF – Enabling Flexible Data Publication

Key features

- Receive a citable DOI for your dataset
- Native support for large datasets
  - millions of files or many TB of data
- Distributed storage enabled
- Host data on high availability, reliable, performant storage (ALCF, UIUC) or your own storage cluster

- Currently hold over 30 TB of data from over 150 authors
- Free of charge for researchers

Publish via Python Script

```python
from mdf_connect_client import MDFConnectClient

mdfcc = MDFConnectClient()
mdfcc.set_title("My Test Google Dataset")
mdfcc.set_authors(["Ben Blaiszik","Logan Ward"])
mdfcc.set_data(["google:///mdf_connect_test_dataset"])
mdfcc.set_services(["publish","citrine","mrr"])

sub_id = mdfcc.submit_dataset()
mdfcc.get_status(sub_id)
```

MDF – Curating Datasets

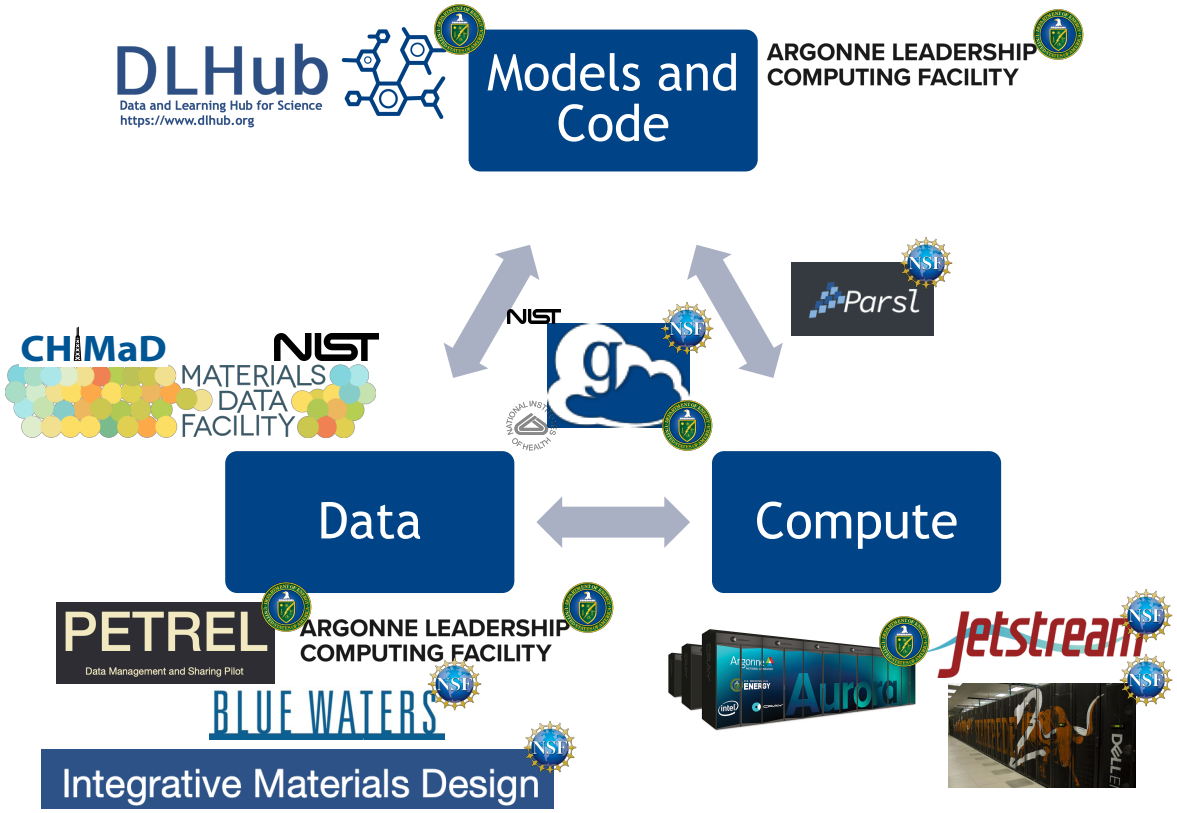
Key features

- Asynchronous process, temporarily halts the indexing flow if curation is required
- Curation requirements can be defined by each organization
- Curation groups can be defined to specify which users can approve submissions
- Specific metadata fields may be required
- Metadata fields may be reused
- Types can be enforced on metadata fields
- Prescribed licenses may be required

An Emerging Data Infrastructure Ecosystem

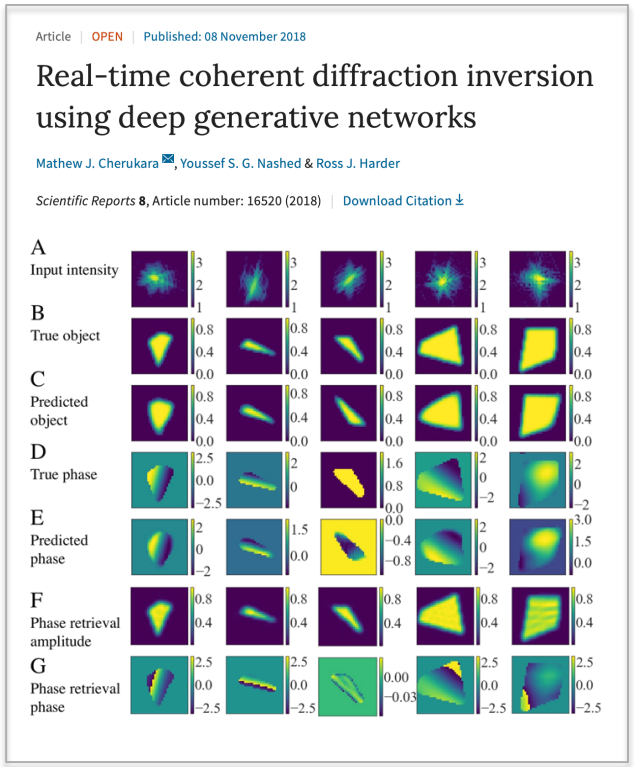
We need and can build a **cohesive infrastructure** for AI-driven science

- Data: capture, organize, analyze, share move and deliver data
- Compute: accessible at many scales, types, and locations
- Models: discoverable, described, and portable to wherever data and/or computer are located
- Workflows: easily discovered, adapted, composed, scaled, and reused



DLHub – A Data and Learning Hub for Science

- Collect, publish, categorize models and processing logic from many disciplines (materials science, physics, chemistry, genomics, etc.)
- Serve models via DLHub operated service to simplify sharing, consumption, and access
- Mint persistent identifiers for all artifacts
- Enable new science through reuse, real-time integration, and synthesis of existing models


Article | OPEN | Published: 08 November 2018

Real-time coherent diffraction inversion using deep generative networks

Mathew J. Cherukara, Youssef S. G. Nashed & Ross J. Harder

Scientific Reports **8**, Article number: 16520 (2018) | Download Citation

Funding: 2018 Argonne Adv. Computing LDRD

Argonne NATIONAL LABORATORY

ARGONNE LEADERSHIP COMPUTING FACILITY

PETREL

parsl

DLHub – A Data and Learning Hub for Science

Describe → Publish → Discover

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

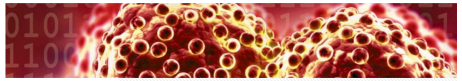
Run

- 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

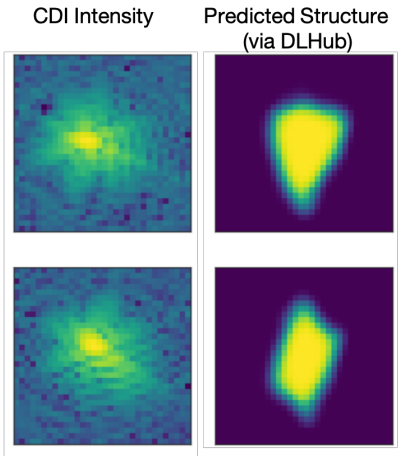
CANDLE



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)
```



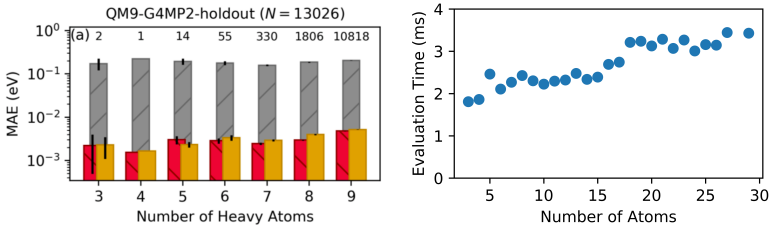
Cherukara et al., 2018

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]

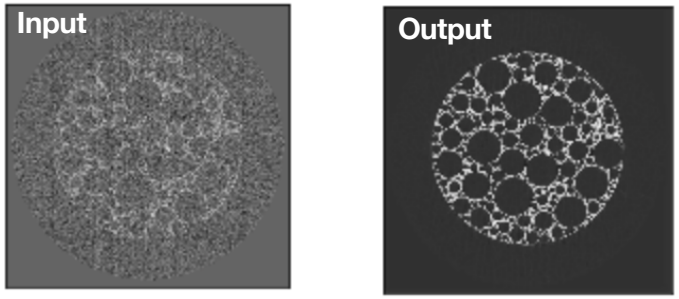


Tomography

- 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



Funding: 2018 Argonne Adv. Computing LDRD

Argonne NATIONAL LABORATORY

ARGONNE LEADERSHIP COMPUTING FACILITY

PETREL

parsl

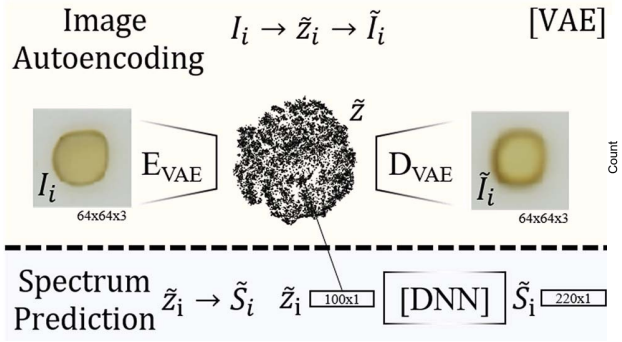
Bringing it Together

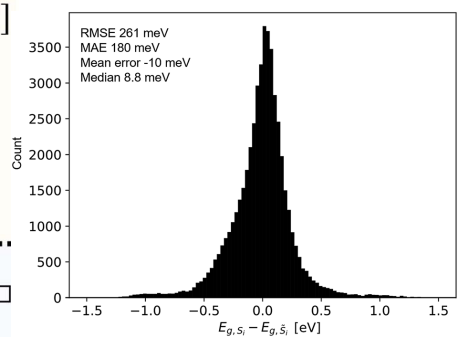
DOI: 10.1039/C8SC03077D (Edge Article) *Chem. Sci.*, 2019, **10**, 47-55

Machine learning of optical properties of materials – predicting spectra from images and images from spectra‡

Helge S. Stein *, Dan Guevarra , Paul F. Newhouse , Edwin Soedarmadji and John M. Gregoire *

Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California 91125, USA. E-mail: stein@caltech.edu; gregoire@caltech.edu





Get Data

Run Model

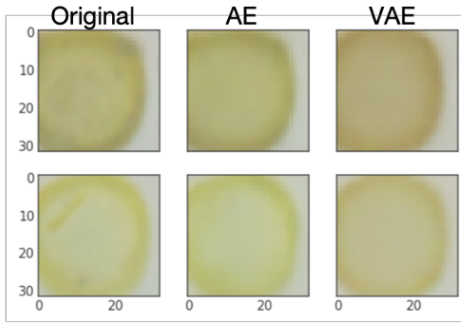
(a) MDF Data Retrieval

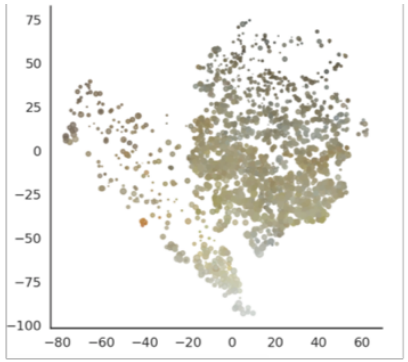
```
from mdf_forge.forge import Forge

mdf = Forge()
q = "mdf.source_id:jcap_optical_spectroscopy"
res = mdf.search(q, advanced=True)
```

(b) DLHub to Encode Images

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

res = dl.run(name="mschwarting_anl/stein_encoder",
             inputs=[images])
```

(c) Model Outputs

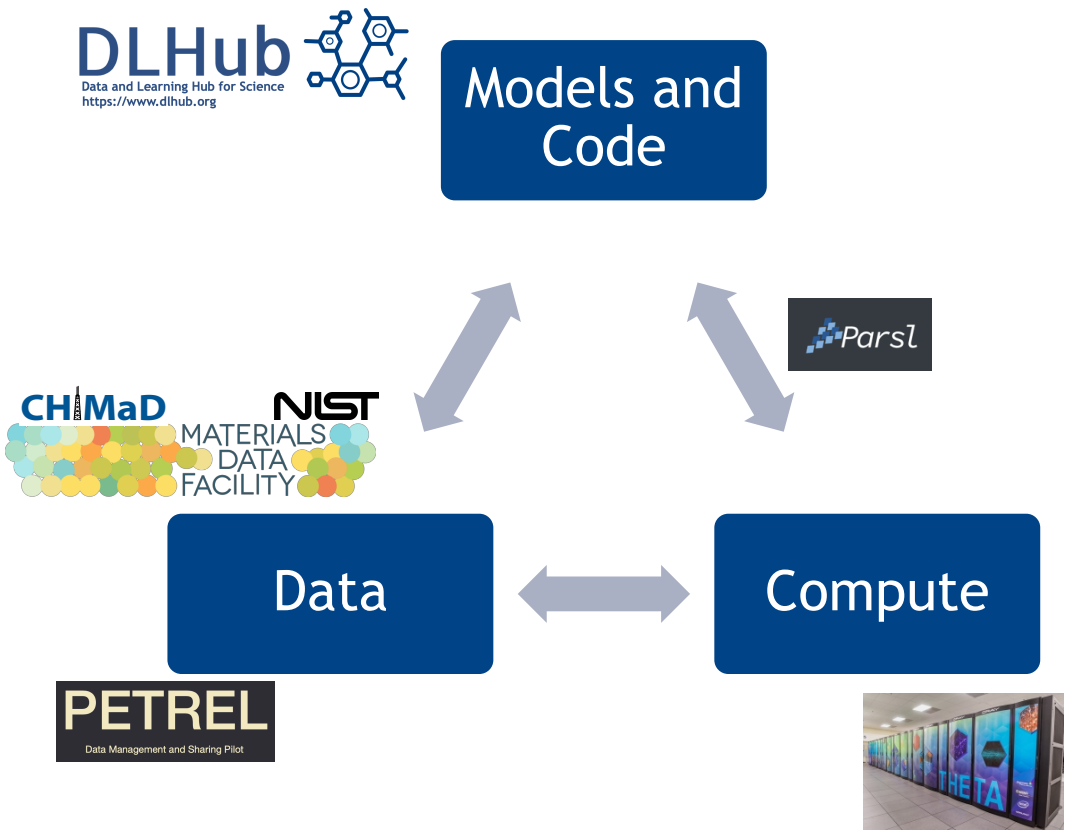




PETREL Data Management and Sharing Pilot NIST globus

2018 Argonne Adv. Computing LDRD

Bringing it Together



Get Data

Run Model

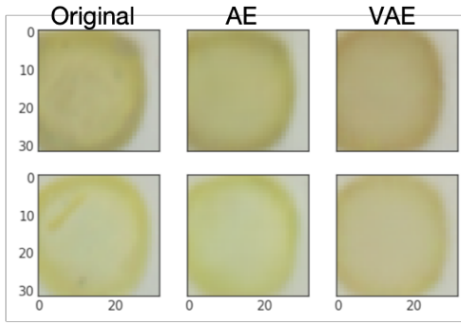
(a) MDF Data Retrieval

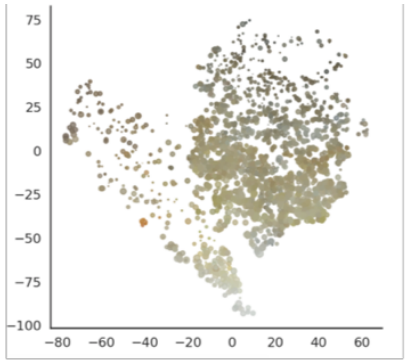
```
from mdf_forge.forge import Forge

mdf = Forge()
q = "mdf.source_id:jcap_optical_spectroscopy"
res = mdf.search(q, advanced=True)
```

(b) DLHub to Encode Images

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

res = dl.run(name="mschwarting_anl/stein_encoder",
             inputs=[images])
```

(c) Model Outputs





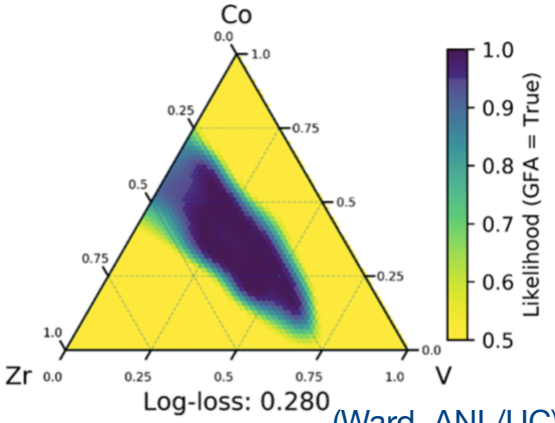
PETREL Data Management and Sharing Pilot NIST globus

2018 Argonne Adv. Computing LDRD

Select DLHub Use Cases

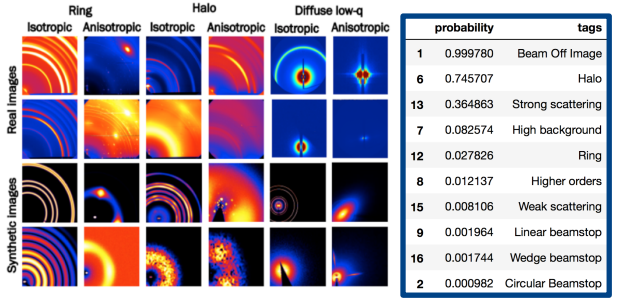
Model-in-the-Loop Science

- Metallic glass discovery [active learning]



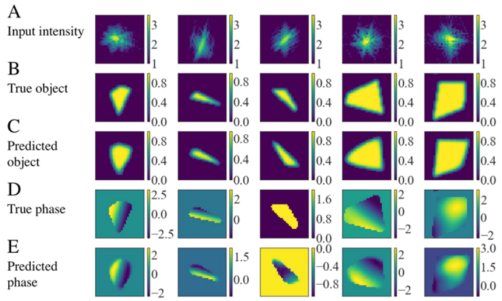
(Ward, ANL/UC)

- Models driving science at light sources



	probability	tags
1	0.999780	Beam Off Image
6	0.745707	Halo
13	0.364863	Strong scattering
7	0.082574	High background
12	0.027826	Ring
8	0.012137	Higher orders
15	0.008106	Weak scattering
9	0.001964	Linear beamstop
16	0.001744	Wedge beamstop
2	0.000982	Circular Beamstop

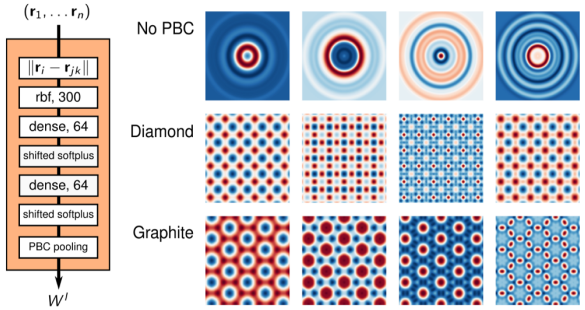
XRD image tagging
(Yager, BNL)



XRD intensity → structure/phase
(Cherukara Argonne)

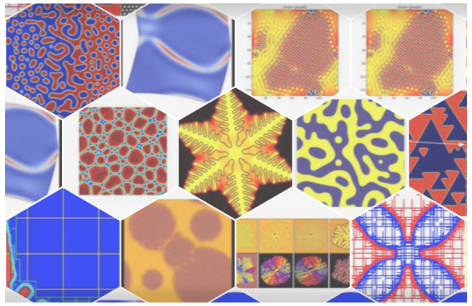
Community Model Benchmarking

- Crystal structure



(Ward, ANL/UC)

- NIST PFHub



(Wheeler, Warren, Heinonen
NIST/UC/Argonne/NU)

Automated Model Retraining with New Data

- Models linked to dynamic data sources

MATERIALS DATA FACILITY
CHiMaD
NIST

(Center for Hierarchical Materials
Design NIST/UC/Argonne/NU)

Funding: 2018 Argonne Adv. Computing LDRD

Argonne NATIONAL LABORATORY

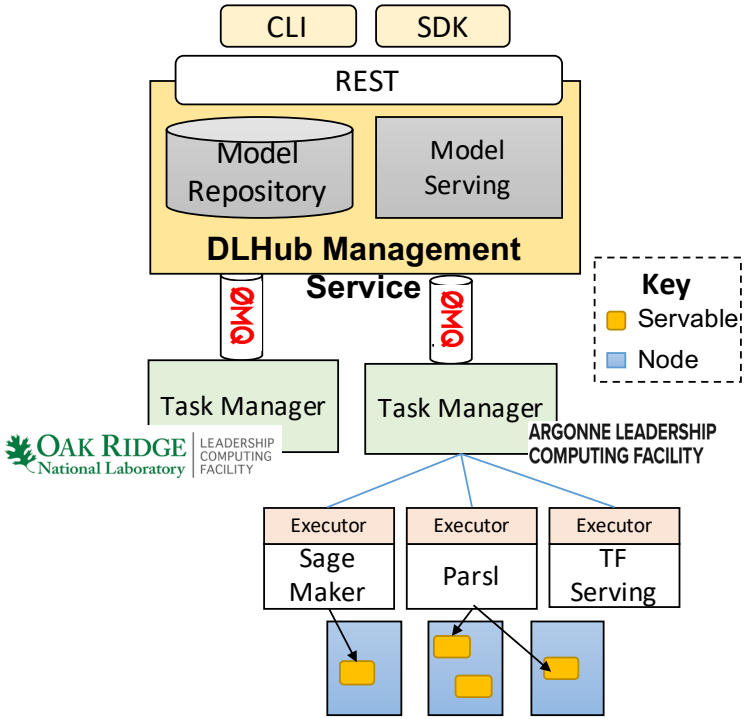
ARGONNE LEADERSHIP COMPUTING FACILITY
PETREL
parsl

DLHub Architecture and Performance

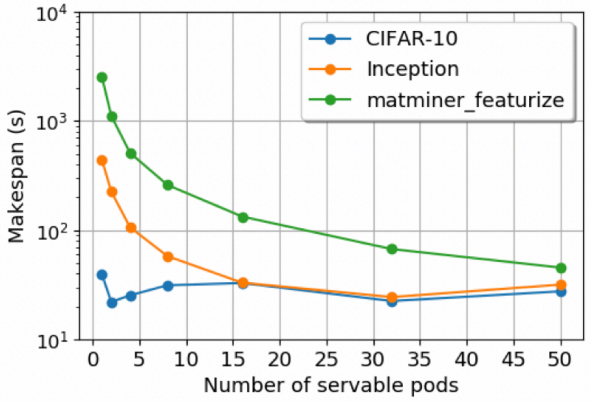
- Task Managers (TM) to support execution on various compute resources
- Executors chosen by TM to invoke a given servable'
- Caching at TM
- Data staging with Globus
- Batch submissions
- Scalability through deployment of model replicas

https://arxiv.org/abs/1811.11213

(Paper accepted to IPDPS 2019)

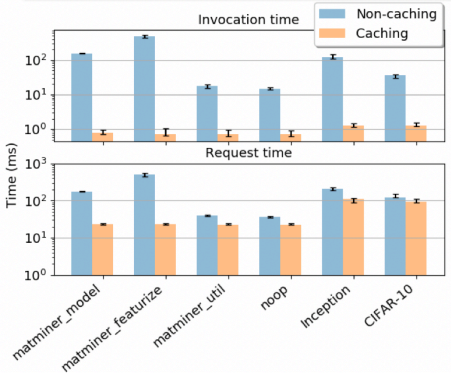


Scale Testing



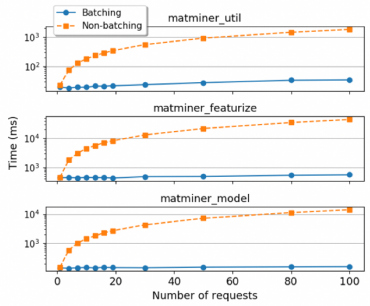
The time required for the Inception, CIFAR10, and Matminer-featurize models to process 5000 inferences with varying numbers of replicas.

Caching



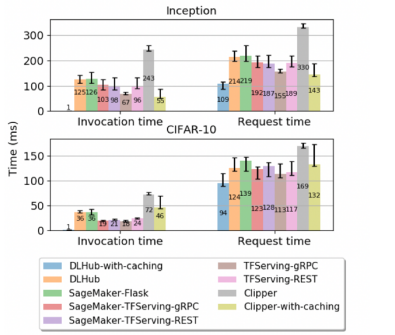
Performance impact of caching in DLHub. Bars and error bars show median and 5th/95th percentiles

Batching



Servable invocation time, with and without batching.

Serving General Models



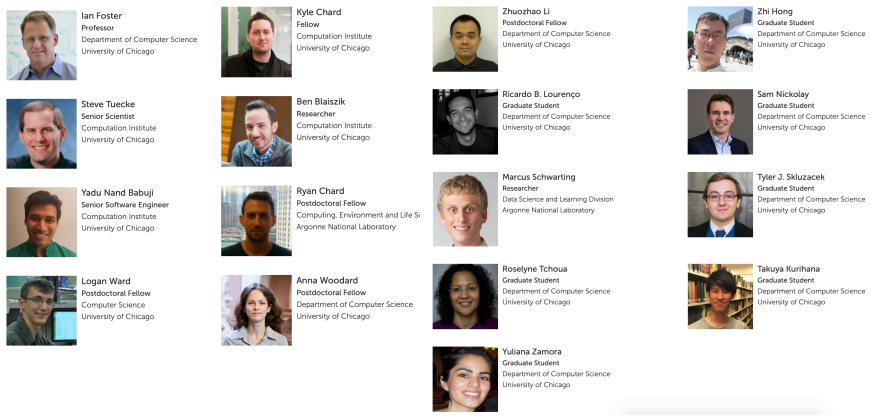
Performance of different serving systems on the Inception and CIFAR-10 problems.

Team

https://www.globus.org

Globus

- Platform as a service serving a community of >100k users, and have moved > 580 PB of data.
- Services include Auth, Transfer, Groups, Search
- All with REST APIs and web interfaces to simplify usage



https://labs.globus.org

Globus Labs

- Research group at UChicago with 17 Ph.D., M.S., Prof., scientists
- Projects span: high-performance computing; machine learning for materials science, chemistry, environmental modeling; automation; data extraction and more

MATERIALS DATA FACILITY

DLHub
Data and Learning Hub for Science
https://www.dlhub.org

References

Websites

- https://www.dlhub.org
- https://www.materialsdatafacility.org

Project Papers

- **DLHub**
  - Blaiszik, Ben, Logan Ward, Marcus Schwarting, Jonathon Gaff, Ryan Chard, Daniel Pike, Kyle Chard, and Ian Foster. **"A Data Ecosystem to Support Machine Learning in Materials Science."** arXiv preprint arXiv:1904.10423 (2019).
  - Chard, Ryan, Zhuozhao Li, Kyle Chard, Logan Ward, Yadu Babuji, Anna Woodard, Steve Tuecke, Ben Blaiszik, Michael J. Franklin, and Ian Foster. **"DLHub: Model and Data Serving for Science."** *arXiv preprint arXiv:1811.11213* (2018).
  - Blaiszik, Ben, Kyle Chard, Ryan Chard, Ian Foster, and Logan Ward. **"Data automation at light sources."** In AIP Conference Proceedings, vol. 2054, no. 1, p. 020003. AIP Publishing, 2019.
- **MDF**
  - Blaiszik, Ben, Logan Ward, Marcus Schwarting, Jonathon Gaff, Ryan Chard, Daniel Pike, Kyle Chard, and Ian Foster. **"A Data Ecosystem to Support Machine Learning in Materials Science."** arXiv preprint arXiv:1904.10423 (2019).
  - Blaiszik, B., K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster. **"The materials data facility: data services to advance materials science research."** JOM 68, no. 8 (2016): 2045-2052.

Thanks to our sponsors!

Parsl

NSF

Globus

NIST

U.S. DEPARTMENT OF ENERGY

NATIONAL INSTITUTES OF HEALTH

NSF

MATERIALS DATA FACILITY

IMaD NSF

NIST

CHiMaD
Center for Hierarchical Materials Design

ARGONNE LEADERSHIP COMPUTING FACILITY

ALCF DF

PETREL
Data Management and Sharing Pilot

DLHub

U.S. DEPARTMENT OF ENERGY

Argonne LDRD

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

THE UNIVERSITY OF CHICAGO

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov