A Data Ecosystem to Support Machine Learning in Materials Science

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

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

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
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**

- **Submit**
- **Enrich**
- **Dispatch**

**MDF Connect**
- Extract: Crystal structure, Composition, File information, Other metadata
- Transform: File format, Representations, Vocabularies, ...

**MDF Publish**
- Support for large datasets
- Persistent storage for dataset
- DOI for referencing
- Globus endpoint for access

**MDF Discover**
- Cloud-hosted metadata index
- Advanced search capabilities
- Access-controlled searches
- Python tools and REST API

**Beta testing**

Flexible interfaces simplify submission

Data can be ingested from many sources

- Web
- Programmatic
  - REST API and Python SDK

- Connectors
  - Auth
  - Transfer
  - Groups
  - Search

Other community data services

NIST MRR

CITRINE
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

https://www.materialsdatafacility.org

Or with a Python script
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)
```

https://www.materialsdatafacility.org
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
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
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

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

**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

---

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

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

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

---

**Exascale Cancer Research**

**CANDLE**

---

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

Logan Ward1,2, Ben Blaiszik1,3, Ian Foster1,2, Rajiv S. Assary4,5, Badri Narayanan4,6, Larry Caris44

---

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

Zhengchun Liu, Tekin Becer, Rajkumar Kettimuthu, Doga Gursoy, Francesco De Carlo, Ian Foster
Bringing it Together

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

Helge S. Stein, Dan Quevarra, 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@jcaptech.edu

Get Data

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

Run Model

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
res = dl.run(name="mschwarting_anl/stein_encoder",
inputs=[images])
```
Bringing it Together

Get Data

Models and Code

Run Model

Data

Compute

- Auth
- Transfer
- Groups
- Connectors
- Search

(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

Original  AE  VAE

2018 Argonne Adv. Computing LDRD
References

Websites

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

Project Papers

• DLHub

• MDF
Thanks to our sponsors!