

Facility and Instrument PIDs in the Materials Research Data Domain

Is there a roadmap?

Are we optimistic or pessimistic?

Got PID?

David Elbert: elbert@jhu.edu

CDO PARADIM Materials Innovation Platform (MIP)

ARL HTMDEC Extreme Data PI

2 NSF DMREF Co-PI

DOE Supported Catalysis Co-PI

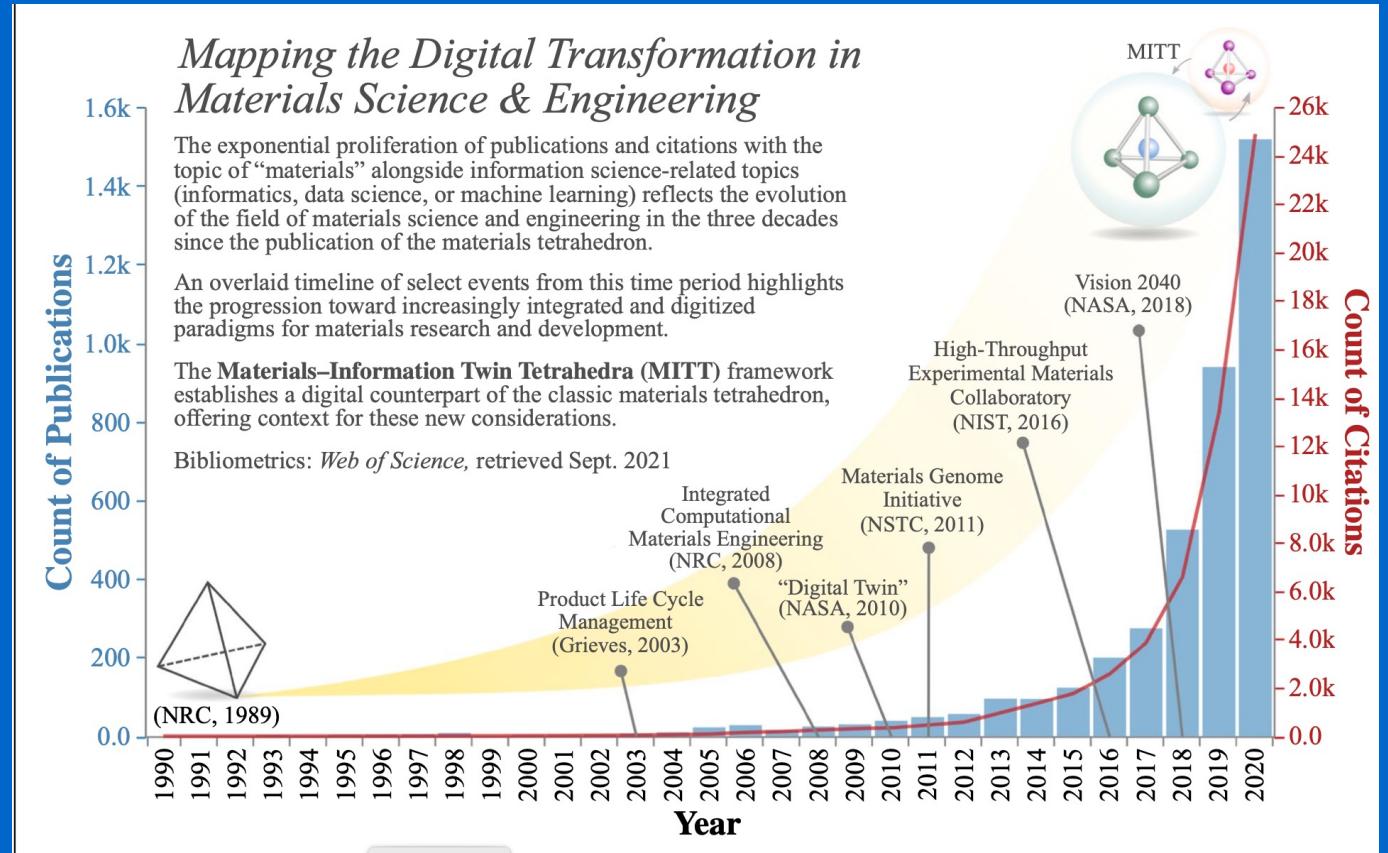
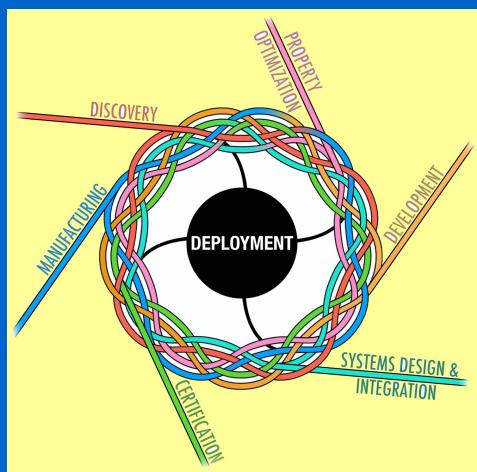
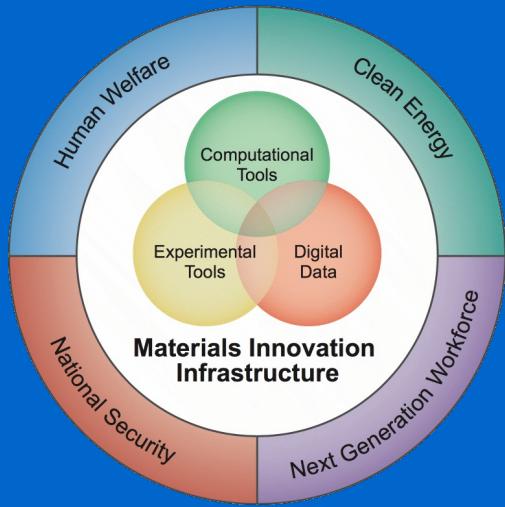
Materials Research Data Alliance (MaRDA)

MaRCN FAIROS-RCN PI



Illegitimi non carborundum

Materials Genome Initiative (MGI) Driven Directions



Deagen et al, 2022 MRS Bulletin
doi:10.1557/s43577-021-00214-0

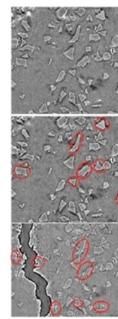
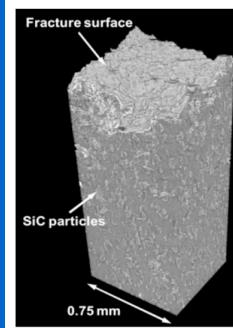
Materials Science Motivations

Firehose of Data

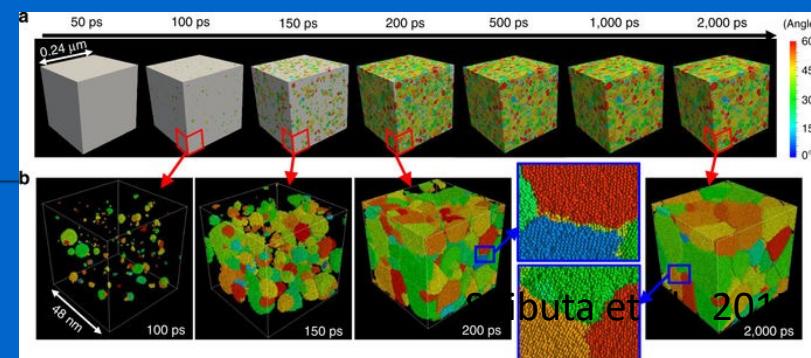
- Higher resolution
- Shorter time scales
- Higher dimensionality
- Dynamic experiments
- Larger simulations
- Tighter processing control

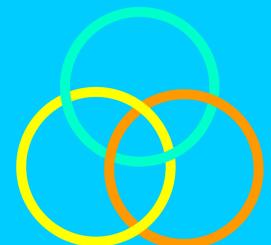
Diverse, Distributed Data

- Experimental
 - labs great and small
- Modeling
 - finite element to ab initio
 - desktops to supercomputers



De Carlo et al., 2012





MaRDA

Materials Research Data Alliance

MaRCN

<https://marda-alliance.org>

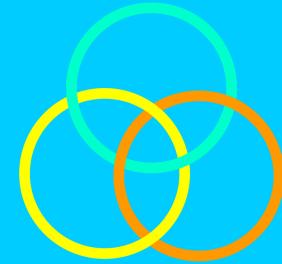
MGI 2.0: *People are infrastructure, too!*

2021 Revised Strategic Plan Goal 1, Objective 2:

Unify the Materials Innovation Infrastructure

- Bridge, Build, and Bolster Elements of the MII.
- Establish a National Materials Data Network
- Accelerate through National Grand Challenges





MaRDA

Materials Research Data Alliance

MaRDA is a **community network** focused on developing the *open, accessible, and interoperable materials data* that fuels the Materials Genome Initiative (MGI).

MaRDA is a **convergence of people and ideas working together** to connect materials data infrastructure to accelerate discovery, enable new insights into materials mechanisms, and lay a foundation for both human-centered and artificial intelligence-assisted approaches to materials design.

FAIR for Materials:

- FAIR is still poorly understood by most of our investigators
- They don't know that FAIR are principles not implementation



<https://doi.org/10.1557/s43577-023-00498-4>

The FAIR Guiding Principles	
Findable:	F1 Data and metadata are assigned a globally unique and persistent identifier F2 Data are described with rich metadata (defined by R1 below) F3 Metadata clearly and explicitly include the identifier of the data it describes F4 Data and metadata are registered or indexed in a searchable resource
Accessible:	A1 Data and metadata are retrievable by their identifier using a standardized communications protocol A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorization procedure, where necessary A2 Metadata are accessible, even when the data are no longer available
Interoperable:	I1 Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2 Data and metadata use vocabularies that follow FAIR principles I3 Data and metadata include qualified references to other (meta)data
Reusable:	R1 Data and metadata are richly described with a plurality of accurate and relevant attributes R1.1 Data and metadata are released with a clear and accessible data usage license

Adapted from Wilkinson et al., 2016
<https://fairtoolkit.pistoiaalliance.org/fair-guiding-principles/>



FAIR for Materials

FAIR means metadata, too



Community action
will fuel a revolution
in research

L. Catherine Brinson,*^{ID} Laura
Ian Foster, Alejandro Strachan

<https://doi.org/10.1557/s43577-023-00498-4>

Individual Actions



Example: Coordinated Development for Interoperability

- Metadata Extractors
Matthew Evans, UC Louvain
Peter Kraus, TU Berlin
David Elbert, Johns Hopkins

- Interoperability Layer
- LinkML
 - Translate LinkML instance data to OWL (TBoxes and ABoxes)

Screenshot of the GitHub repository github.com/marda-alliance/metadata_extractors

The repository has 1 branch and 0 tags. The main branch contains 17 commits from ml-evs. The commits include:

- .github: Add CODEOWNERS (3 months ago)
- meetings: Add link to slides (2 months ago)
- .gitignore: Initial commit (5 months ago)
- LICENSE: Initial commit (5 months ago)
- README.md: Add contributing notes (3 months ago)

The README.md file contains the following content:

MaRDA Extractors WG

This repository contains organizational info for a [MaRDA](#) working group (WG) focused on connecting and advancing interoperability of efforts on automated extraction of metadata from materials files.

Contacts:

- [Matthew Evans, UCLouvain](#) (matthew.evans[at]uclouvain.be)
- [Peter Kraus, TU Berlin](#) (peter.kraus[at]ceramics.tu-berlin.de)
- [David Elbert, Johns Hopkins University](#) (elbert[at]jhu.edu)

Contributing

This working group is completely open. If you would like to be added to the mailing list please reach out to us over email, or just turn up at a meeting!

The GitHub discussions on this repo can be used for pretty much any related chat, and specific code suggestions/feedback can be made as pull requests to the GitHub repos for each subproject:

- [marda-alliance/metadata_extractors_schema](#)
- [marda-alliance/metadata_extractors_registry](#)
- [marda-alliance/metadata_extractors](#)



- Interoperability Layer
- LinkML
 - Translate LinkML instance data to OWL (TBoxes and ABoxes)

data_extractors Public

Full requests Discussions Actions Projects Wiki Security Insights Settings

Announcements Welcome announcement, come introduce yourself! davidelbert

General Arranging informal "office hours" ml-evs

Search: is:open Clear Sort by: Latest activity Label Filter: Open New discussion

Categories

- View all discussions
- Announcements
- General
- Ideas
- Polls
- Q&A
- Show and tell

Most helpful

Be sure to mark someone's comment as an answer if it helps you resolve your question — they deserve the credit!

Community guidelines Community insights Screenshot

Upvote Downvote Comment

Discussion details

Arranging informal "office hours" ml-evs started on Dec 1, 2022 in General

Meeting 3: March 21st 15:00 UTC ml-evs announced on Feb 3 in Announcements

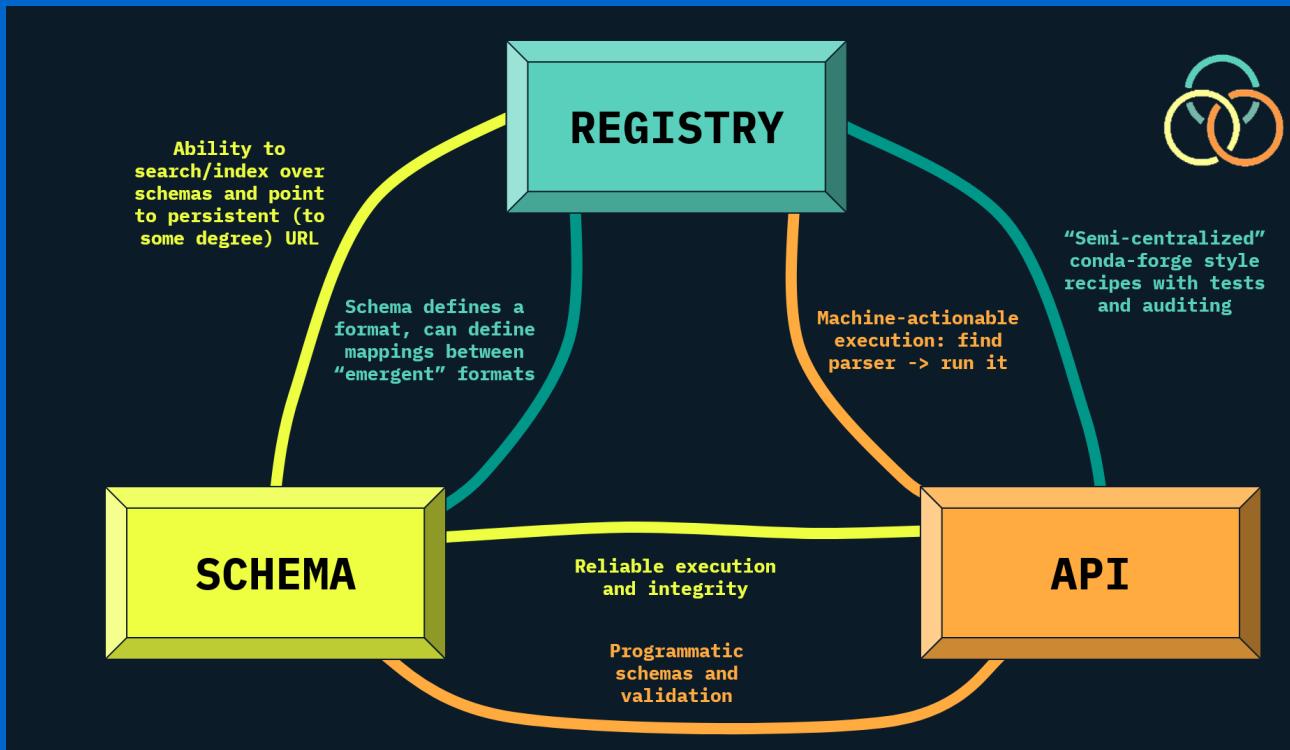
Meeting 2: January 19th 2023 @ 15:00 UTC ml-evs announced on Dec 16, 2022 in Announcements

I. Lightweight metadata schema for parsers ml-evs started on Dec 1, 2022 in Ideas

Definitions... SteffenBrinckmann started on Jan 20 in General

synergy with "advanced data analytics for multiple microscopies" DOE SBIR FY23 topic dwinston started on Dec 19, 2022 in Ideas

II. A common API specification for executing parser code ml-evs started on Dec 1, 2022 in Ideas



MaRCN

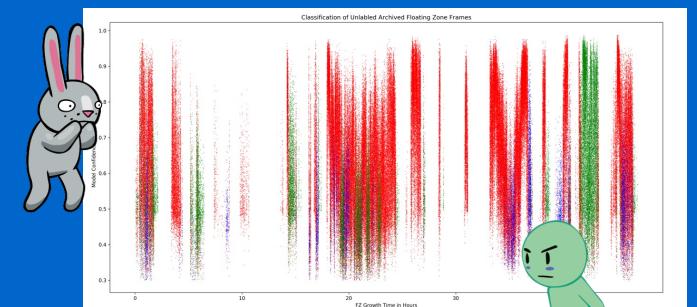
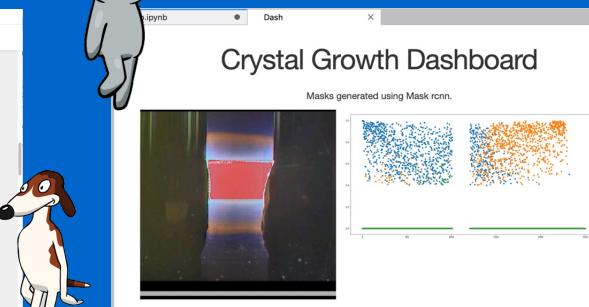
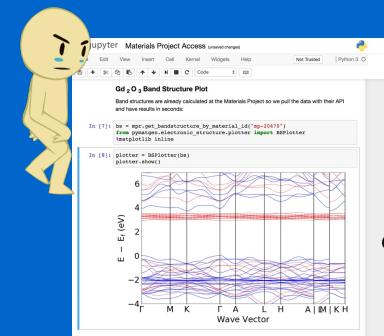
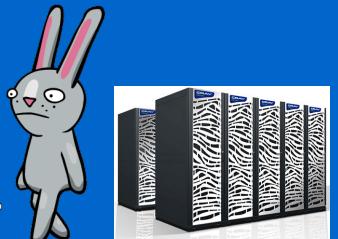
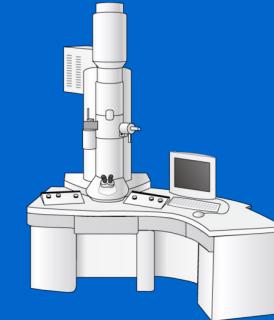
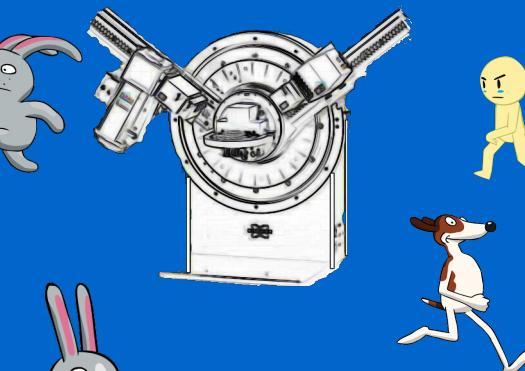
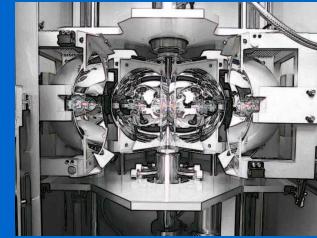
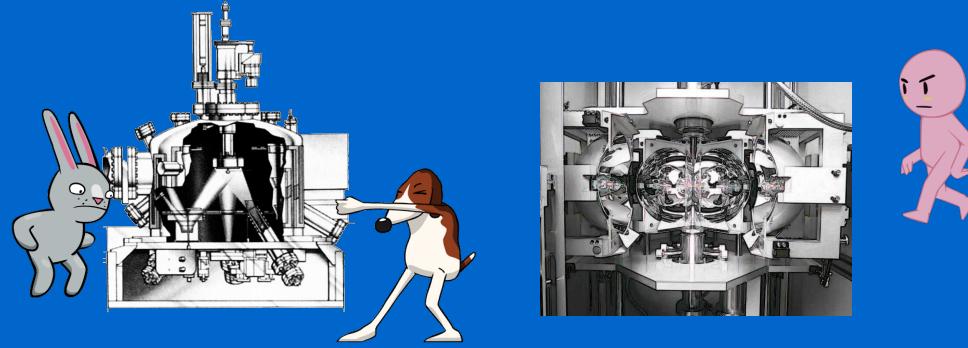


HOPKINS EXTREME
MATERIALS INSTITUTE

Scientists can be myopic

Many Moving Parts

- Equipment
- People
- Ideas



Somas “Data Landscape is Changing!”

- Automating Everything
- High Throughput
- Automation + Decisions = Autonomy
- Linked Data Is Foundational

Scientists are the worst judges
of the reuse of their own data

Review Article | Published: 09 March 2023

Combinatorial synthesis for AI-driven materials discovery

John M. Gregoire [✉](#), Lan Zhou & Joel A. Haber

Nature Synthesis 2, 493–504 (2023) | [Cite this article](#)

1258 Accesses | 5 Altmetric | [Metrics](#)

Review Article | Published: 30 January 2023

The rise of self-driving labs in chemical and materials sciences

Milad Abolhasani [✉](#) & Eugenia Kumacheva

Nature Synthesis 2, 483–492 (2023) | [Cite this article](#)

The high-throughput highway to computational materials design

Stefano Curtarolo [✉](#), Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito & Ohad Levy

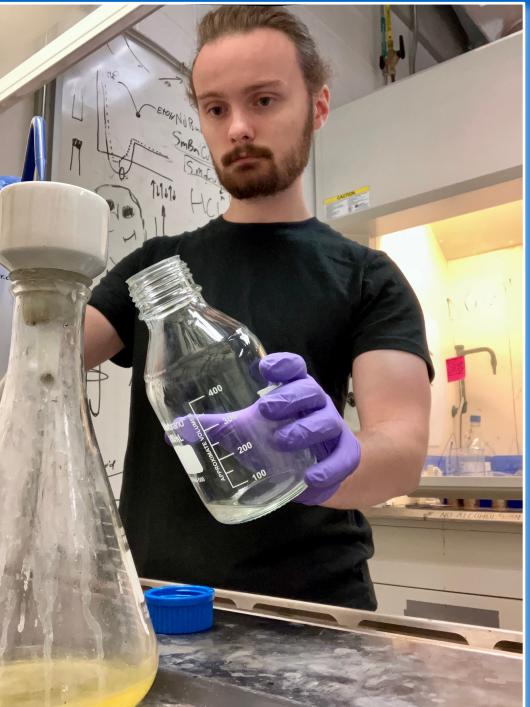
Nature Materials 12, 191–201 (2013) | [Cite this article](#)

43k Accesses | 1357 Citations | 34 Altmetric | [Metrics](#)

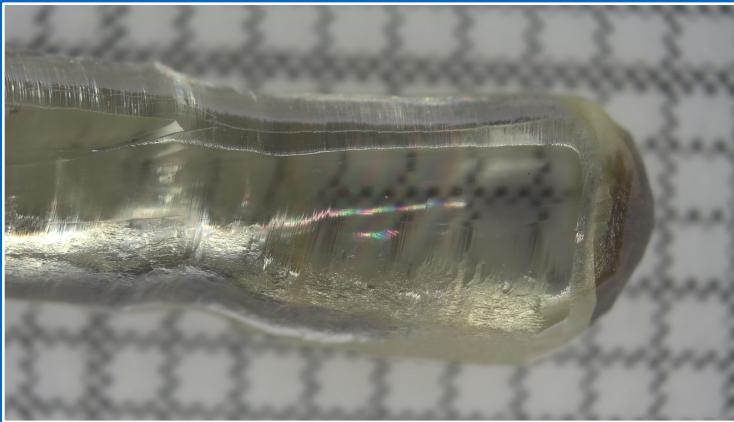


Linking Data GEMD Graphical Model

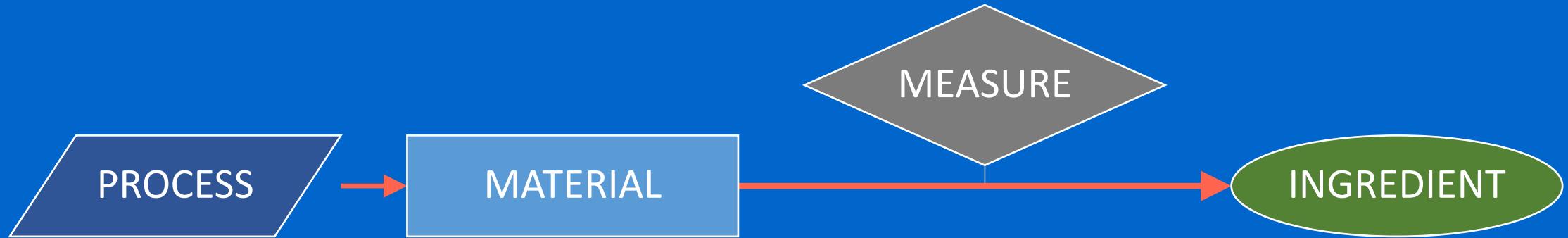
Yttrium Orthovanadate YVO_4



Gannon Murray PARADIM
REU from Earlham College

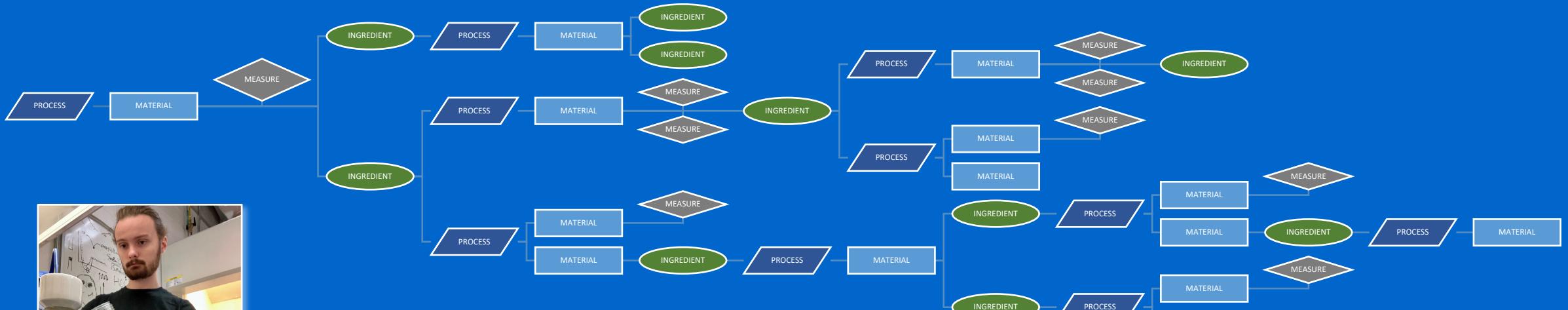


Graphical Expression of Materials Data (GEMD)

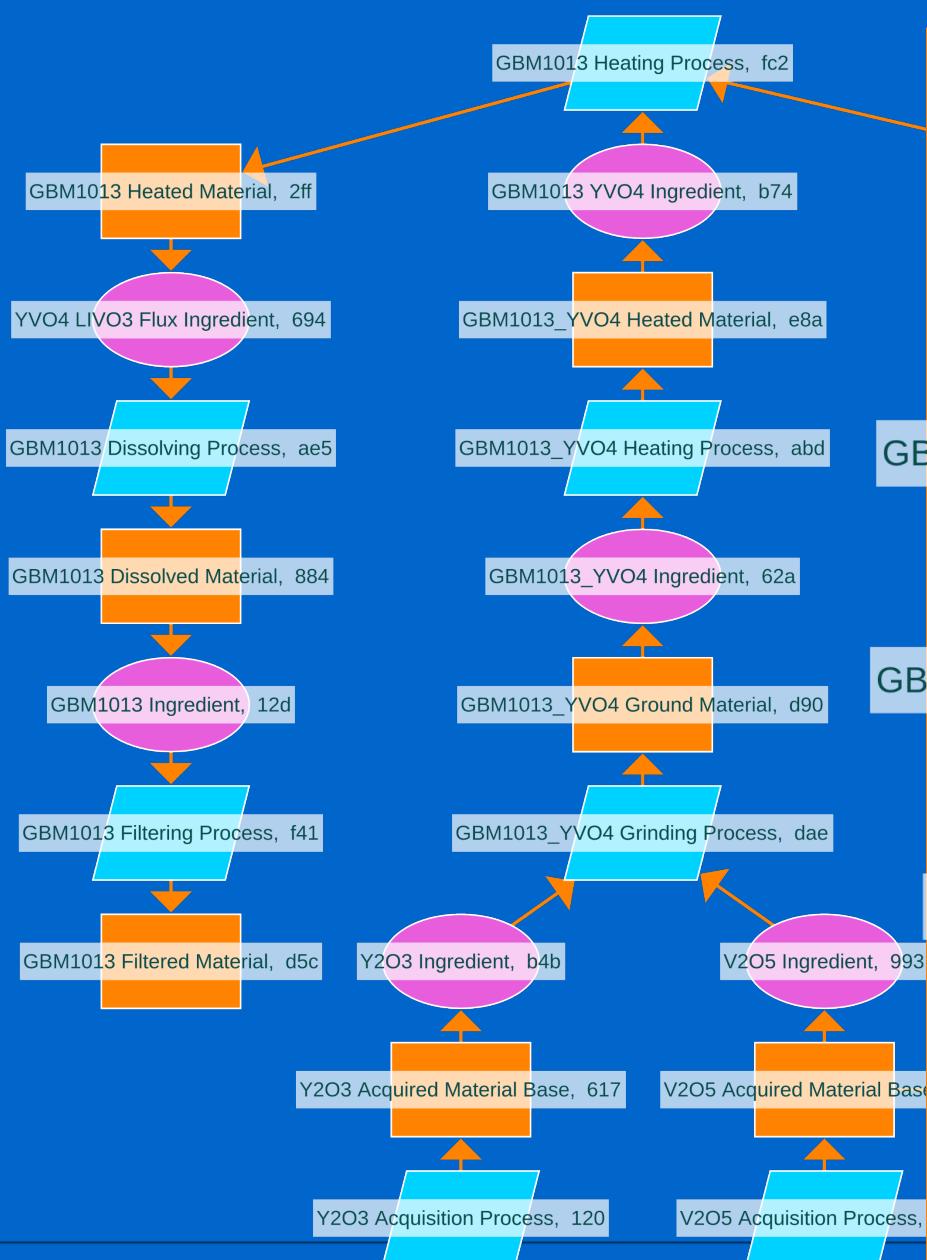
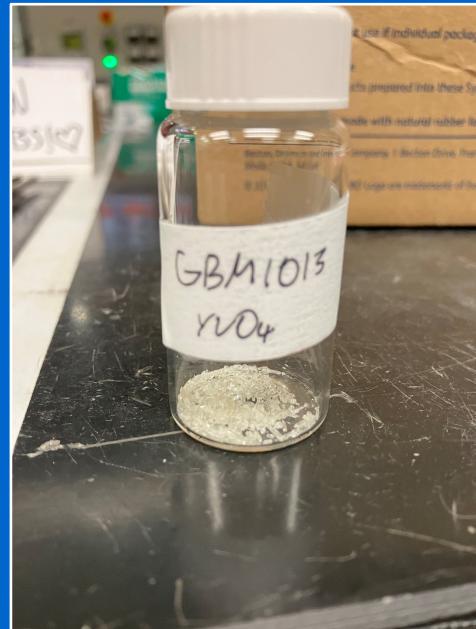


Data driven approaches are transformative
but don't underestimate that:
Connected data is unknown territory

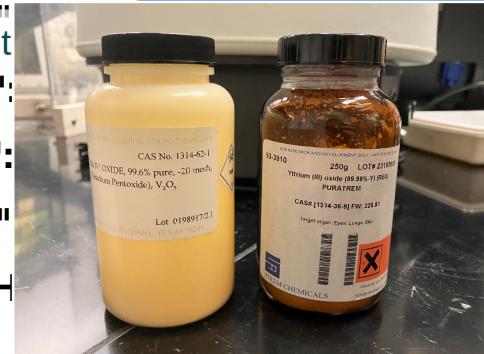
Creating a Material History

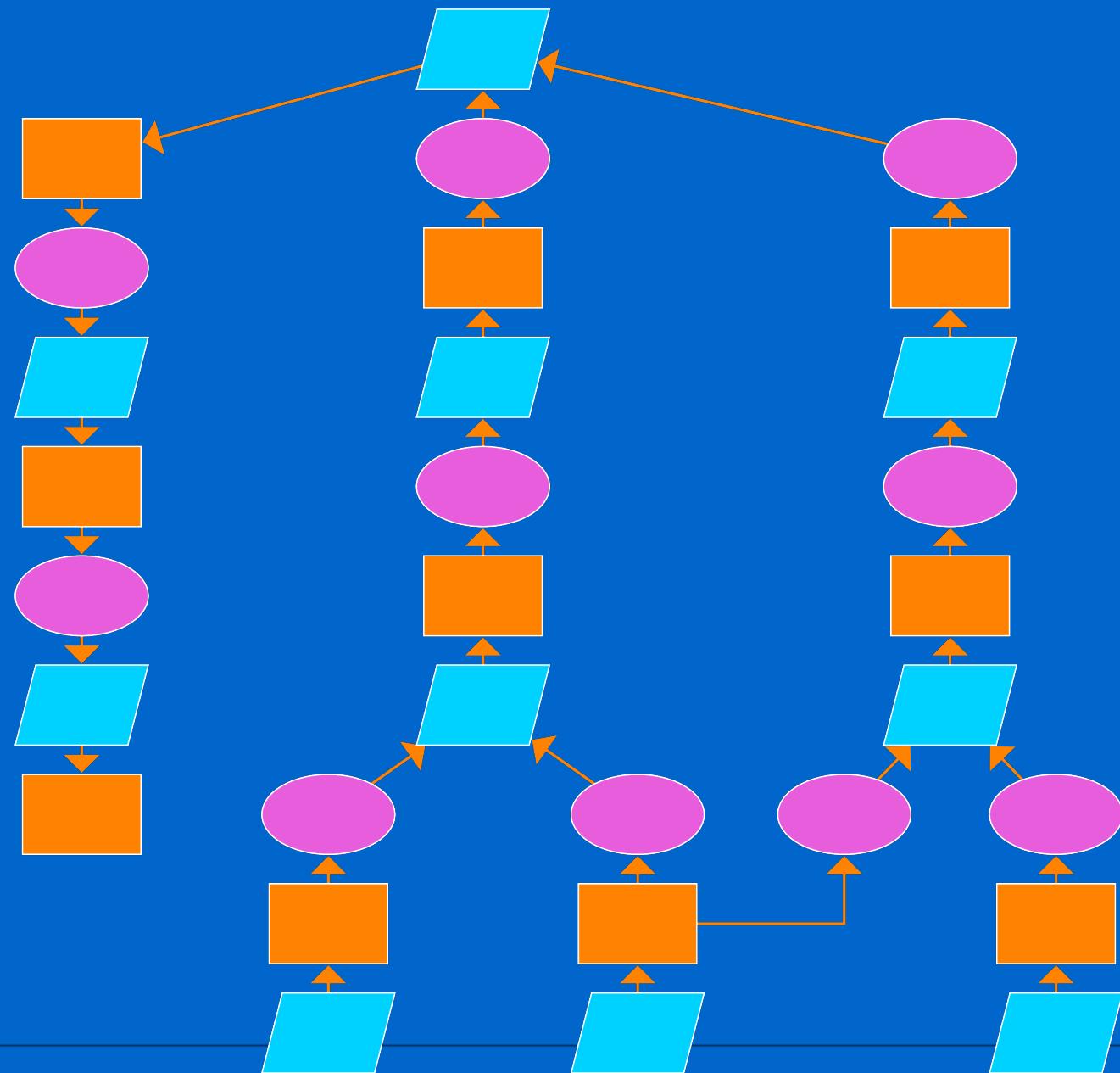


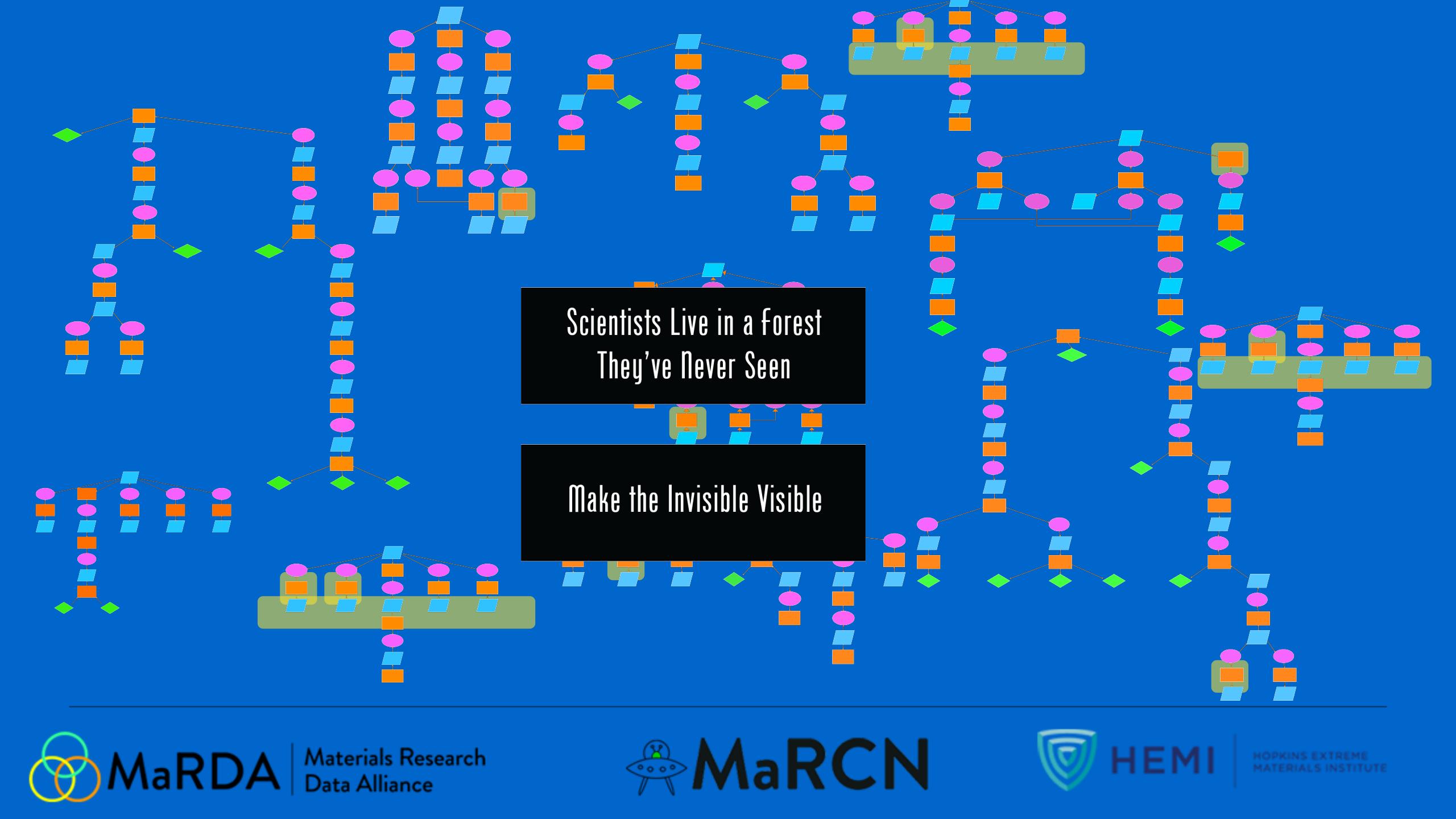
Gannon Murray PARADIM
REU from Earlham College



"Location, Hot Lab": "False"
"Step 1, {'quantities': {'Duration': 0.0,
'Temp': 0.0, 'Type': 'Init'}, 'type':
'nominal_composition'}": "False"
"Step 2, {'quantities': {'Duration': 9.0,
'Temp': 900.0, 'Type': 'Ramp'}, 'type':
"GBM1013¹³Flux_Medium Heated Material, 08a
'nominal_composition'}": "False"
"Step 3, {'quantities': {'Duration': 1.0,
'Temp': 900.0, 'Type': 'End'}, 'type':
"GBM1013_Flux_Medium Heating Process, 5e1
'nominal_composition'}": "False"
"Step 4, {'quantities': {'Duration': 9.0,
'Temp': 0.0, 'Type': 'Ramp'}, 'type':
'nominal_composition'}": "False"
"GBM1013_Flux_Medium Ingredient
"Step 5, {'quantities': {'Duration': 0.0,
'Temp': 0.0, 'Type': 'End'}, 'type':
'nominal_composition'}": "False"
label:"GBM1013_Flux_Medium H
Process, 5e1"







Scientists Live in a forest
They've Never Seen

Make the Invisible Visible

Event Driven Design

Seamless
Dropbox-like entry point
Automate Anything Repetitive
Curation
Reduction
Active Learning
Analysis Pipelines
Path to Deployment
Maintainable
Flexible (Future Foundation)

OpenMSIStream
latest

Search docs

INTRO TO OPENMSISTREAM:
Introduction to OpenMSIStream
Installing OpenMSIStream
Tutorials

WORKING WITH OPENMSISTREAM:
Using Main Programs
Services/Daemons
Encrypting Messages With KafkaCrypto
Troubleshooting
Support and Contribution

API REFERENCE/DEV INFO:
API reference
Code Tests

/ OpenMSIStream


OpenMSIStream

OpenMSIStream provides Python applications for generalized laboratory, analysis, and computational materials data streaming using Apache Kafka. More information is available on the Introduction page for new users. OpenMSIStream is maintained as part of the Open Materials Semantic Infrastructure (Open MSI), under NSF award #1921959, and has been published in the *Journal of Open Source Software*.

<https://openmsistream.readthedocs.io/>



OpenMSIStream: A Python package for facilitating integration of streaming data in diverse laboratory environments

Margaret Eminizer  ¹¶, Sam Tabrisky ^{2,3,4}, Amir Sharifzadeh  ¹, Christopher DiMarco  ⁴, Jacob M. Diamond  ^{4,6}, K. T. Ramesh  ⁴, Todd C. Hufnagel  ^{4,5,6}, Tyrel M. McQueen  ^{4,5,7,8}, and David Elbert  ^{1,4}

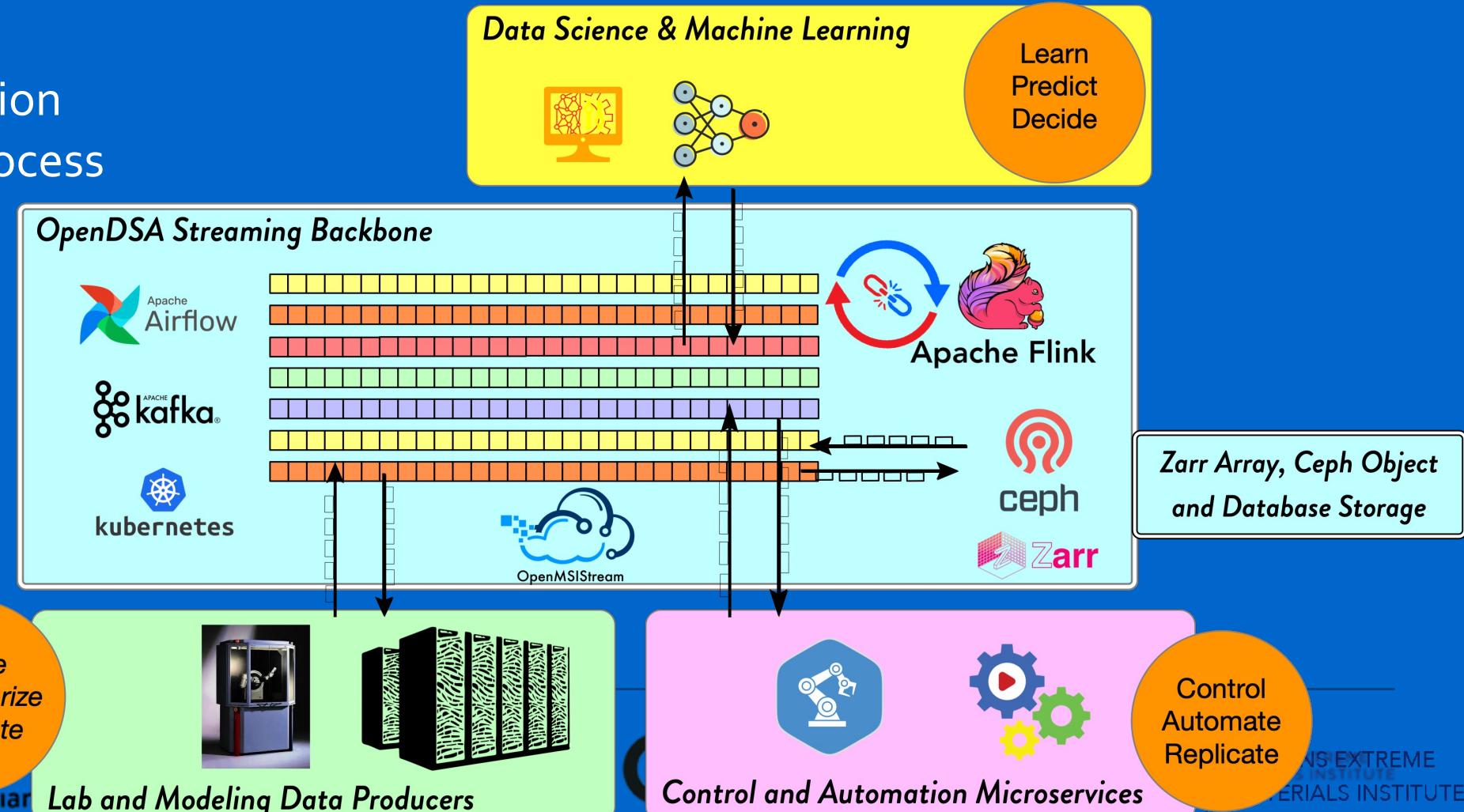
¹ Institute for Data Intensive Engineering and Science (IDIES), The Johns Hopkins University, USA ² Department of Biology, Dartmouth College, USA ³ Department of Computer Science, Dartmouth College, USA ⁴ Hopkins Extreme Materials Institute (HEMI), The Johns Hopkins University, USA ⁵ Department of Materials Science and Engineering, The Johns Hopkins University, USA ⁶ Department of Mechanical Engineering, The Johns Hopkins University, USA ⁷ Department of Chemistry, The Johns Hopkins University, USA ⁸ Institute for Quantum Matter (IQM), William H. Miller III Department of Physics and Astronomy, The Johns Hopkins University, USA ¶ Corresponding author

Event Driven Architecture makes Data the Unifying Thread that automates FAIR; accelerates analysis and AI/ML deployment; and empowers novel science and autonomy

Better Science and Production:

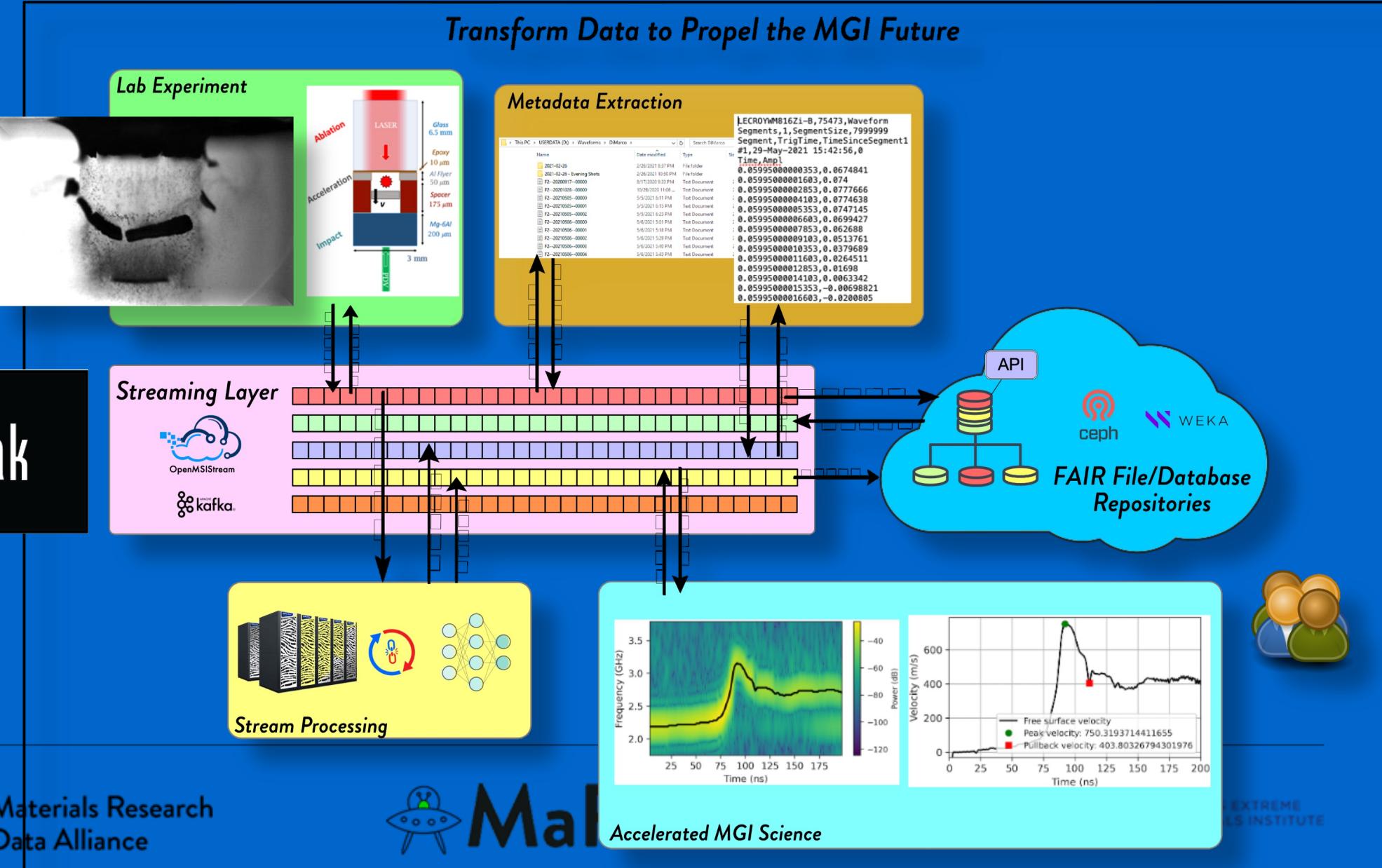
- Faster Discovery
- New Capabilities
- Quicker to production
- Optimization of process

- *High Throughput*
- *Automation*
- *Autonomy*



Example: DMREF Spall Resistant Aluminum OpenMSIStream

Seek the Squeak



A Possible Path?

- PID Interoperability
 - empowers progress
 - perfect is the enemy of the good
- Outcomes Oriented PID as a Service
- Value Added Tools
- Seek the Squeak
- Lean Into Stakeholder Connections
 - Who controls the first mile?
 - Who controls the last mile?
 - Who gets the impact?
- MaRDA/MaRCN Connect Stakeholders

2023

2025

Do you know

2030

where your data are?

2035...

Got PID?

Ask not what
your PID can
do for you...

