

# Facility and Instrument PIDs in the Materials Research Data Domain

Got PID?

*Is there a roadmap?*

*Are we optimistic or pessimistic?*

David Elbert: [elbert@jhu.edu](mailto:elbert@jhu.edu)

CDO PARADIM Materials Innovation Platform (MIP)

ARL HTMDEC Extreme Data PI

2 NSF DMREF Co-PI

DOE Supported Catalysis Co-PI

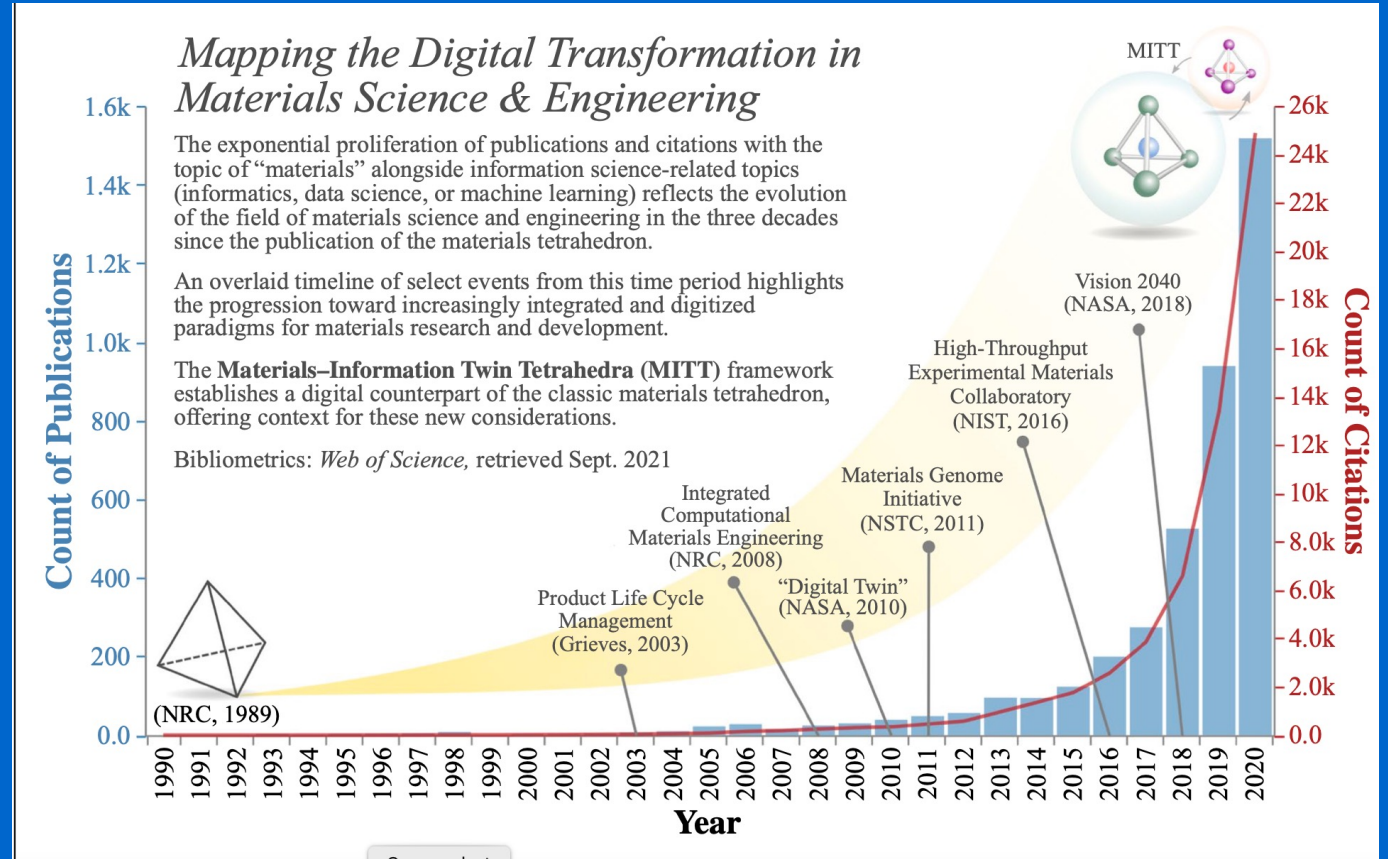
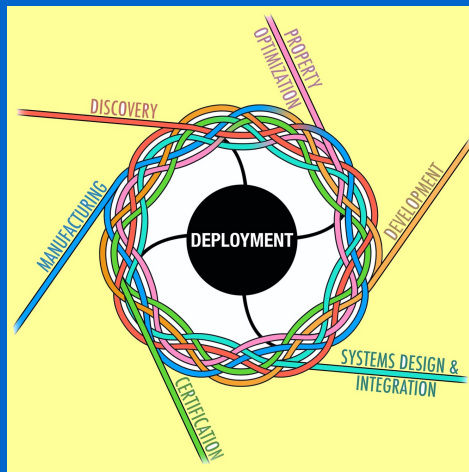
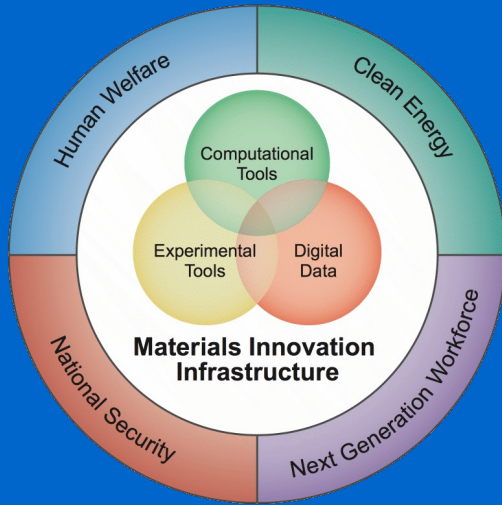
Materials Research Data Alliance (MaRDA)

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

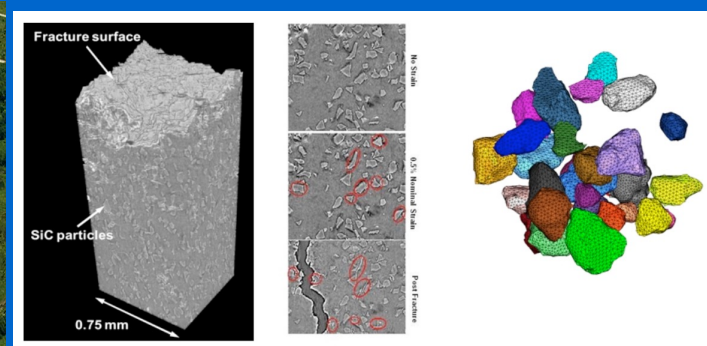
# Materials Science Motivations

## Firehose of Data

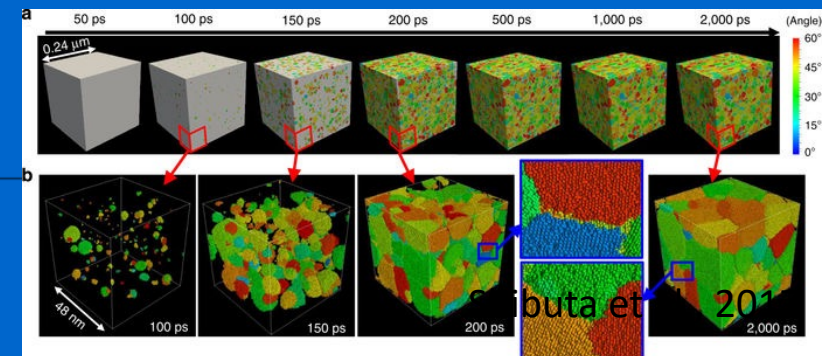
- Higher resolution
- Shorter time scales
- Higher dimensionality
- Dynamic experiments
- Larger simulations
- Tighter processing control

## 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

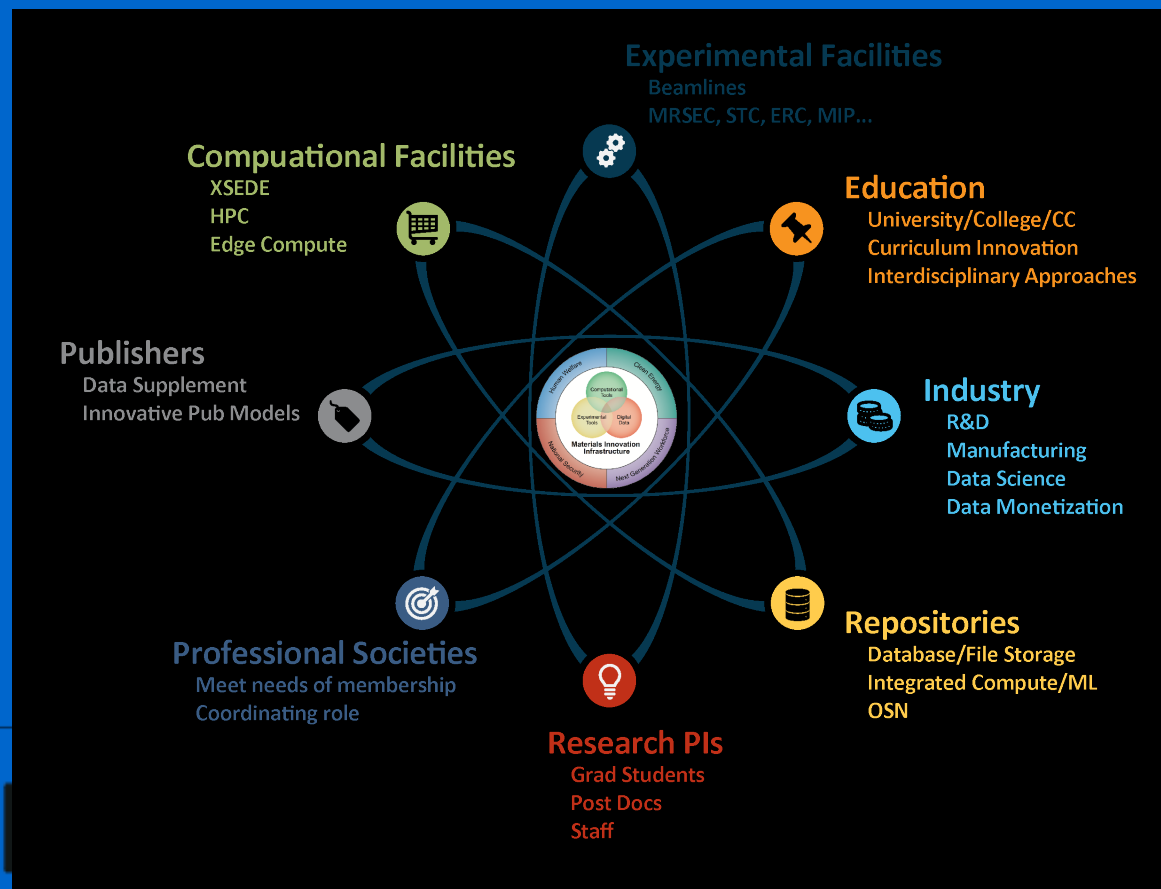
<https://marda-alliance.org>

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

The FAIR Guiding Principles	
<b>Findable:</b>	
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
<b>Accessible:</b>	
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
<b>Interoperable:</b>	
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
<b>Reusable:</b>	
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/>

Impact Opinion & Perspective



## Community action on FAIR data will fuel a revolution in materials research

L. Catherine Brinson,\* Laura M. Bartolo, Ben Blaiszik, David Elbert, Ian Foster, Alejandro Strachan, and Peter W. Voorhees



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



# FAIR for Materials

FAIR means metadata, too



**Community action will fuel a revolution in research**

L. Catherine Brinson,\*<sup>ORCID</sup> Laura  
Ian Foster, Alejandro Strachan

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



# 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)

The screenshot shows a GitHub repository page for 'marda-alliance/metadata\_extractors'. At the top, there are navigation links for Pull requests, Discussions, Actions, Projects, Wiki, Security, Insights, and Settings. Below the repository name, it shows the current branch 'main', 1 branch, and 0 tags. A pull request is visible, merged by 'ml-evs' from 'marda-alliance/PeterKraus-patch-1' on Jan 20, with 17 commits. A table lists recent files and their commit dates:

File	Commit Description	Time Ago
.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 content is displayed below, titled 'MaRDA Extractors WG'. It describes the repository's purpose: '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.' It lists contacts: Matthew Evans (UCLouvain), Peter Kraus (TU Berlin), and David Elbert (Johns Hopkins University). It also includes a 'Contributing' section stating the group is open and provides instructions on how to reach out via email or GitHub discussions. At the bottom, there are links to related repositories: 'marda-alliance/metadata\_extractors\_schema', 'marda-alliance/metadata\_extractors\_registry', and 'marda-alliance/metadata\_extractors\_schema'.

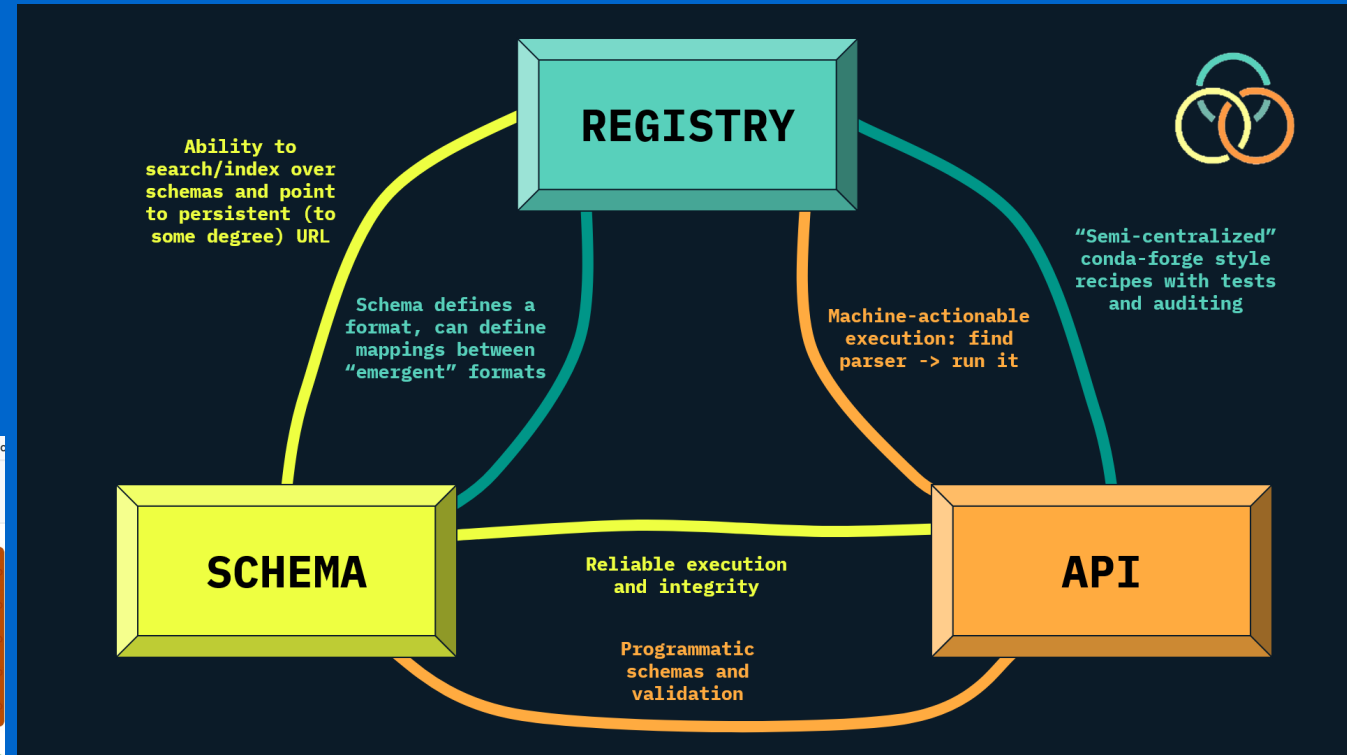


# • Interoperability Layer

## • LinkML

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

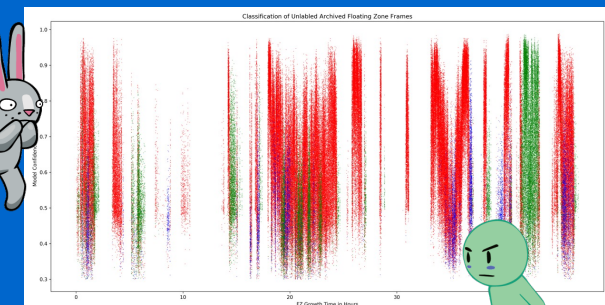
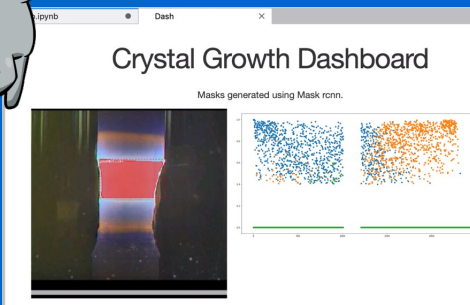
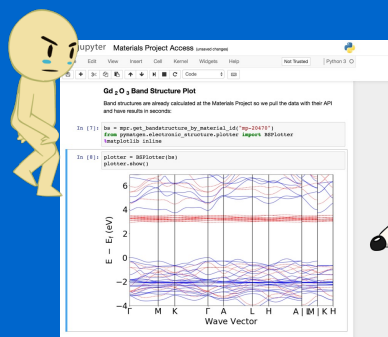
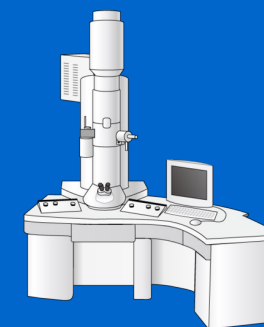
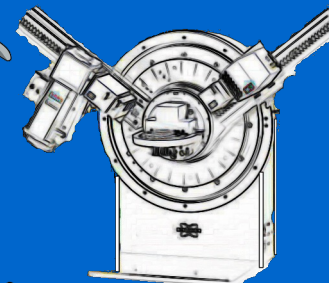
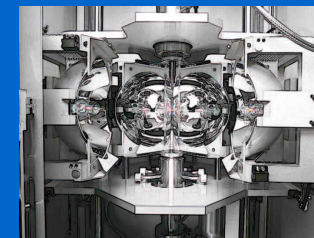
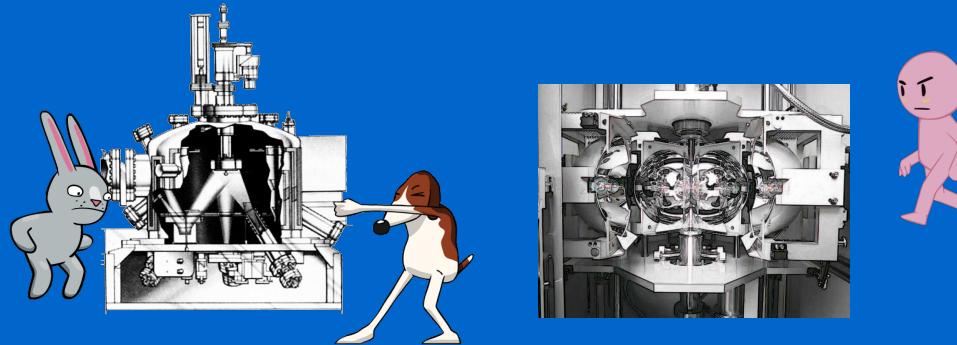
The screenshot shows a GitHub repository page for 'data\_extractors'. The top navigation bar includes 'Discussions', 'Actions', 'Projects', 'Wiki', 'Security', 'Insights', and 'Settings'. Below the navigation, there are two announcement cards: a purple one from 'davidelbert' saying 'Welcome announcement, come introduce yourself!' and a red one from 'ml-evs' titled 'Arranging informal "office hours"'. A search bar shows 'is:open' with filters for 'Sort by: Latest activity', 'Label', and 'Filter: Open'. The 'Discussions' section lists several topics, including 'Arranging informal "office hours"', 'Meeting 3: March 21st 15:00 UTC', 'Meeting 2: January 19th 2023 @ 15:00 UTC', 'I. Lightweight metadata schema for parsers', 'Definitions...', 'synergy with "advanced data analytics for multiple microscopies" DOE SBIR FY23 topic', and 'II. A common API specification for executing parser code'.



# 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: 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](#)

Review Article | [Published: 09 March 2023](#)

## Combinatorial synthesis for AI-driven materials discovery

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

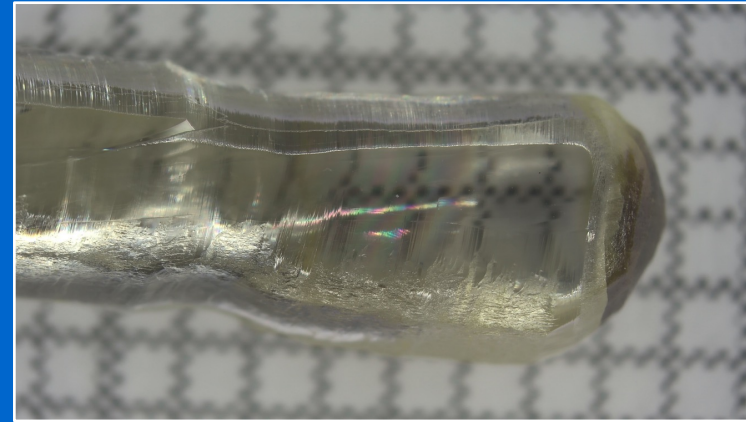
*Nature Synthesis* **2**, 493–504 (2023) | [Cite this article](#)

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

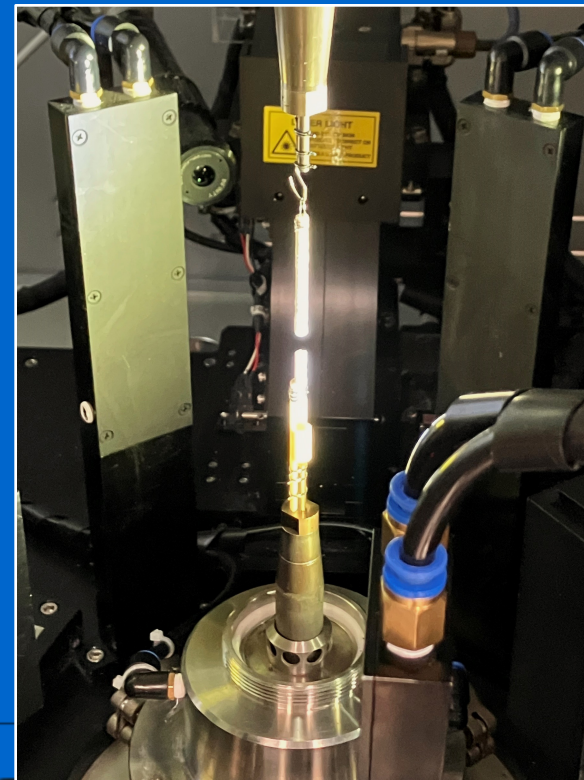


# Linking Data GEMD Graphical Model

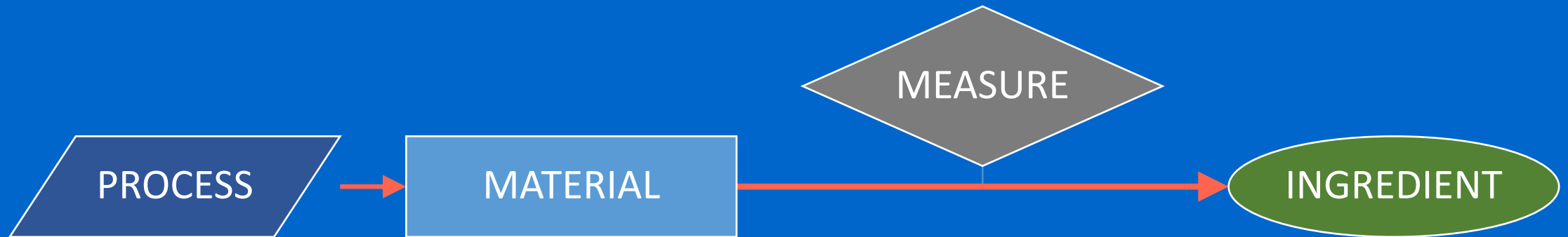
Yttrium Orthovanadate  $YVO_4$



Gannon Murray PARADIM  
REU from Earlham College

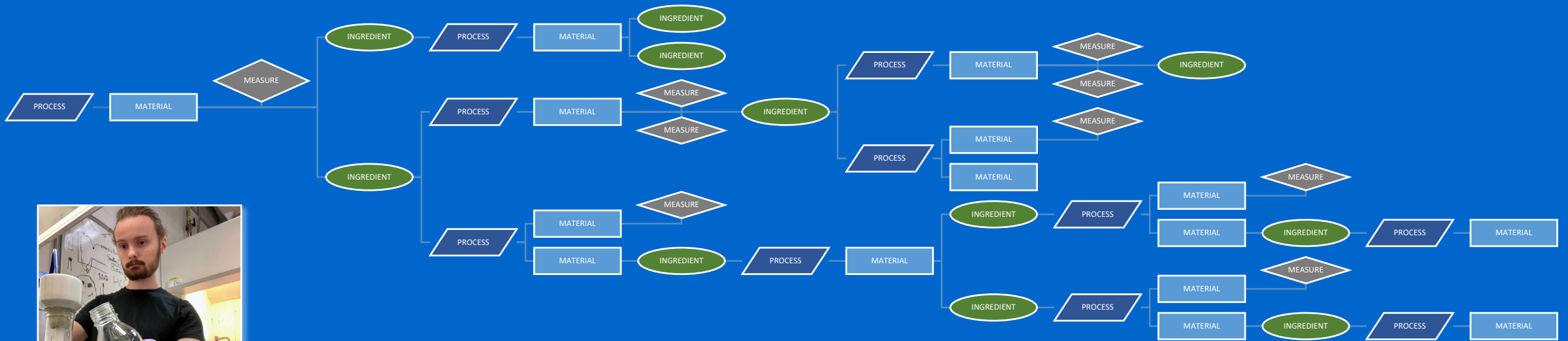


# Graphical Expression of Materials Data (GEMD)



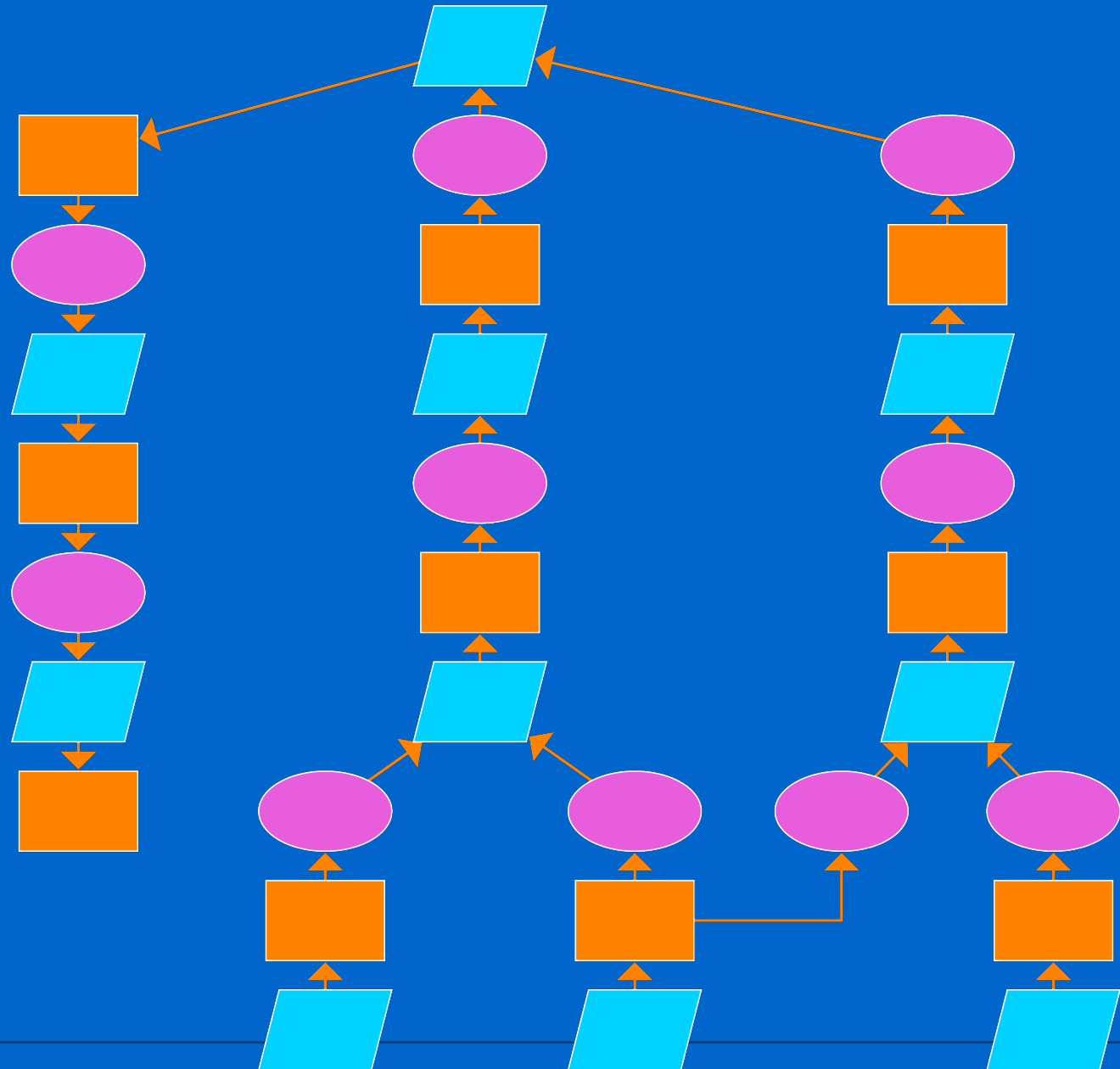
Data driven approaches are transformative  
*but don't underestimate that:*  
*Connected data is unknown territory*

# Creating a Material History

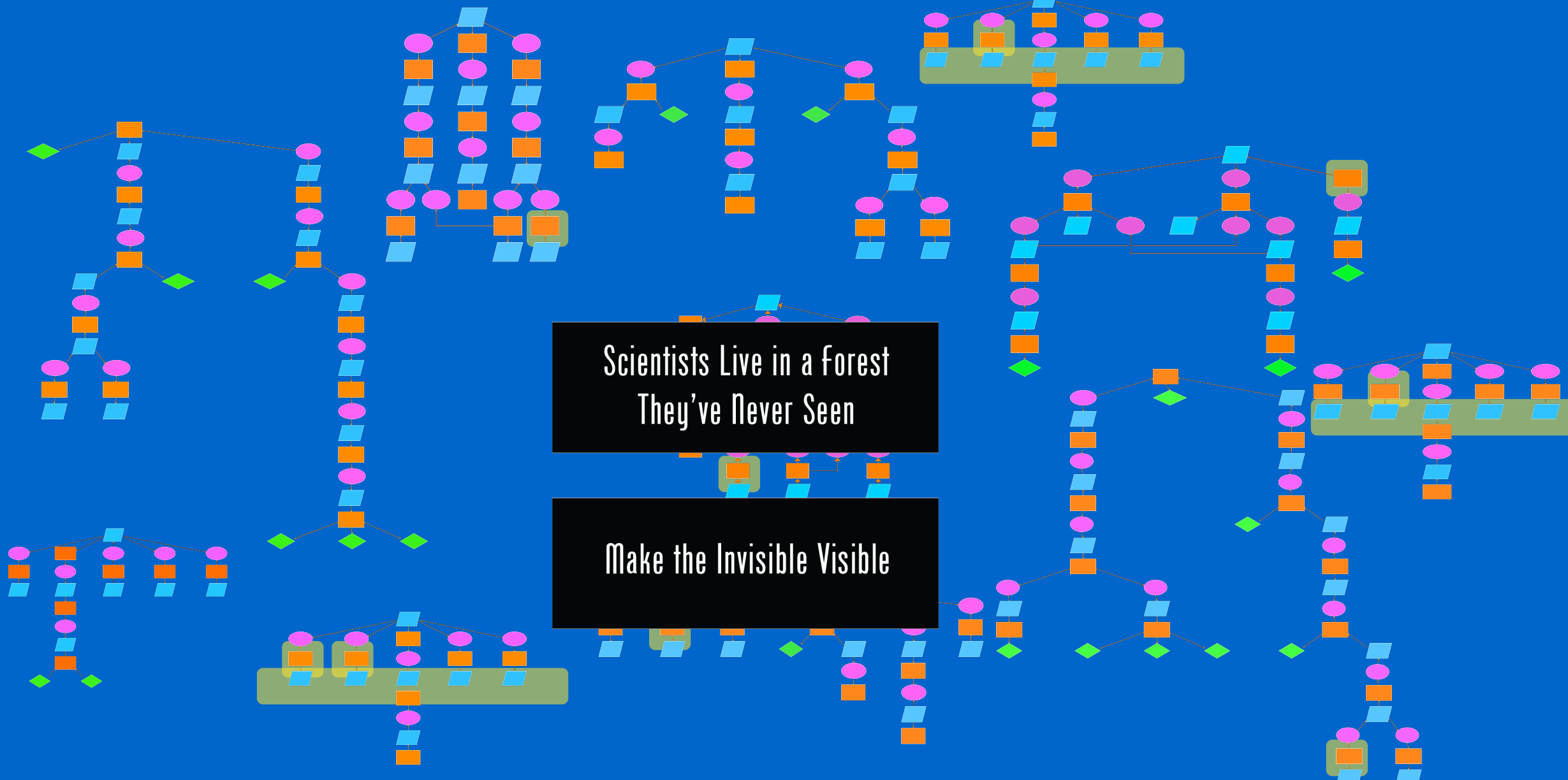


Gannon Murray PARADIM  
REU from Earlham College









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

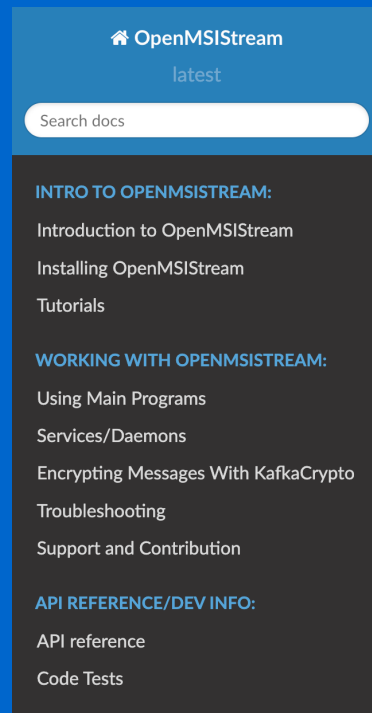
Active Learning

Analysis Pipelines

Path to Deployment

Maintainable

Flexible (Future Foundation)



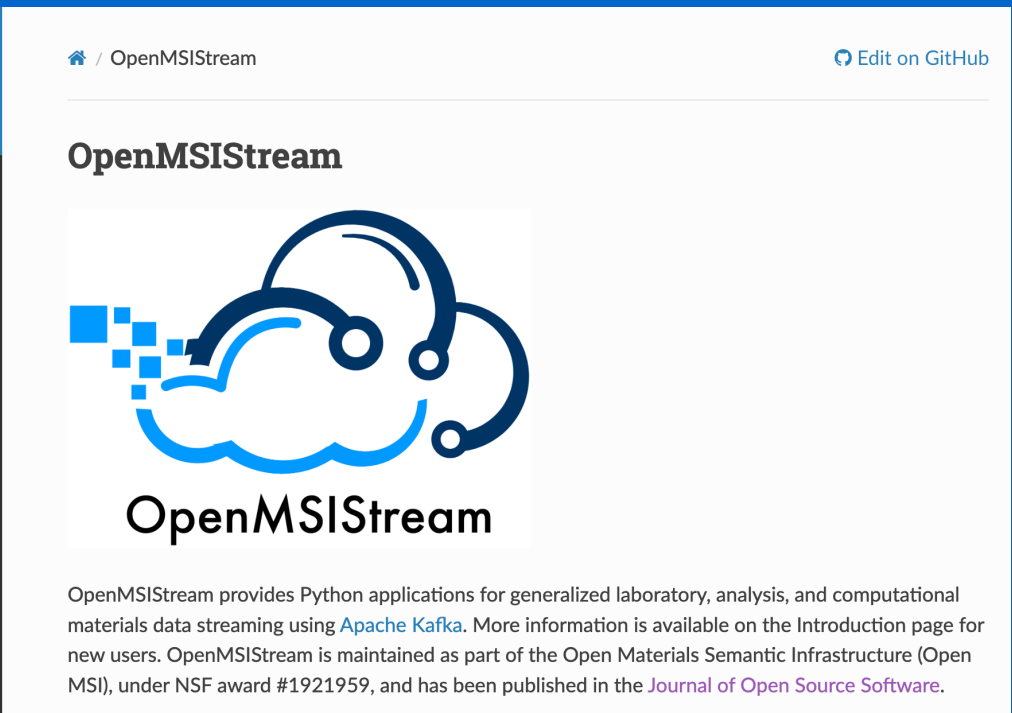
OpenMSIStream  
latest

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<sup>1¶</sup>, Sam Tabrisky<sup>2,3,4</sup>, Amir Sharifzadeh<sup>1</sup>, Christopher DiMarco<sup>4</sup>, Jacob M. Diamond<sup>4,6</sup>, K. T. Ramesh<sup>4</sup>, Todd C. Hufnagel<sup>4,5,6</sup>, Tyrel M. McQueen<sup>4,5,7,8</sup>, and David Elbert<sup>1,4</sup>

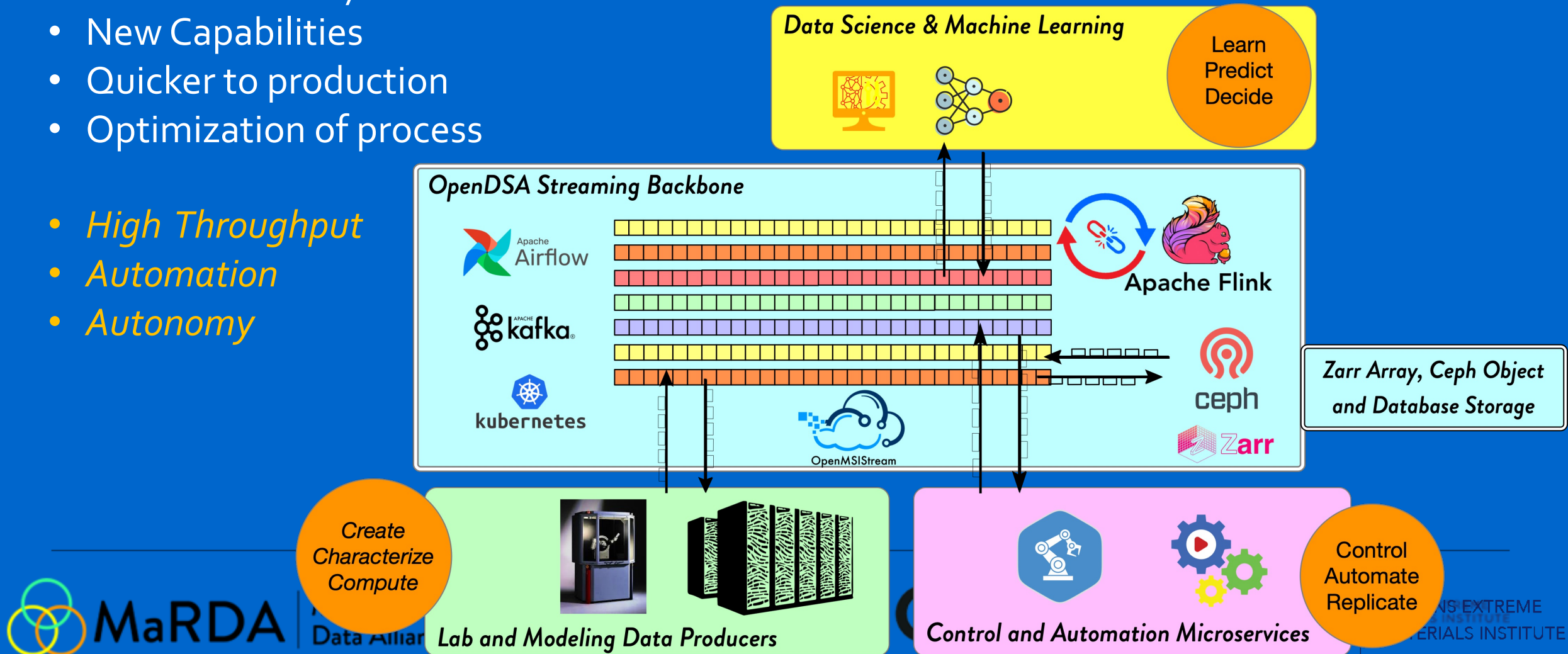
1 Institute for Data Intensive Engineering and Science (IDIES), The Johns Hopkins University, USA 2 Department of Biology, Dartmouth College, USA 3 Department of Computer Science, Dartmouth College, USA 4 Hopkins Extreme Materials Institute (HEMI), The Johns Hopkins University, USA 5 Department of Materials 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

DOI: 10.21105/joss.04896

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:

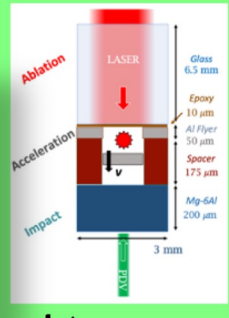
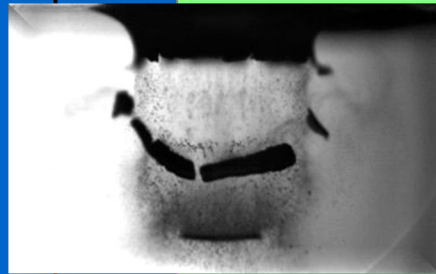
- Faster Discovery
- New Capabilities
- Quicker to production
- Optimization of process
- *High Throughput*
- *Automation*
- *Autonomy*



# Example: DMREF Spall Resistant Aluminum OpenMSIStream

Transform Data to Propel the MGI Future

Lab Experiment



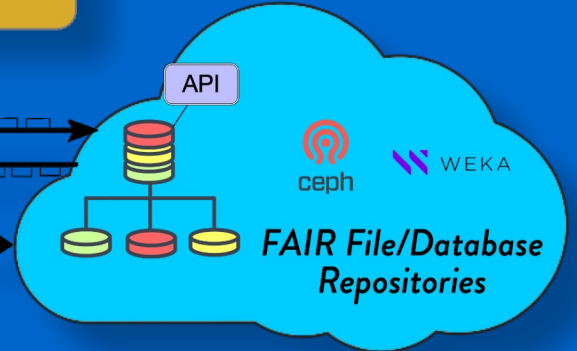
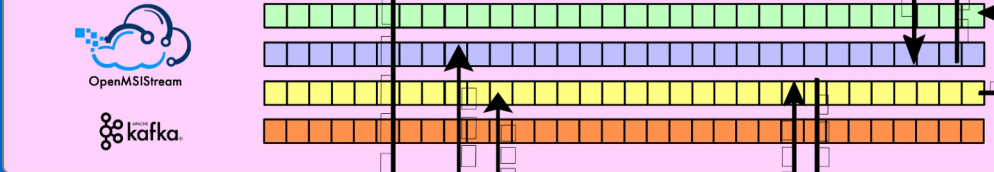
Metadata Extraction

Name	Date modified	Type
2021-03-26	2/26/2021 8:17 PM	File folder
2021-03-26 - Evening Shots	2/26/2021 10:00 PM	File folder
F2-20200917-00000	9/17/2020 8:03 PM	Text Document
F2-20200918-00000	9/18/2020 11:06...	Text Document
F2-20200919-00000	9/19/2021 6:19 PM	Text Document
F2-20210000-00000	9/19/2021 6:23 PM	Text Document
F2-20210001-00000	9/16/2021 5:01 PM	Text Document
F2-20210006-00001	9/6/2021 5:18 PM	Text Document
F2-20210006-00002	9/6/2021 5:29 PM	Text Document
F2-20210009-00000	9/6/2021 5:49 PM	Text Document
F2-20210009-00006	9/6/2021 5:49 PM	Text Document

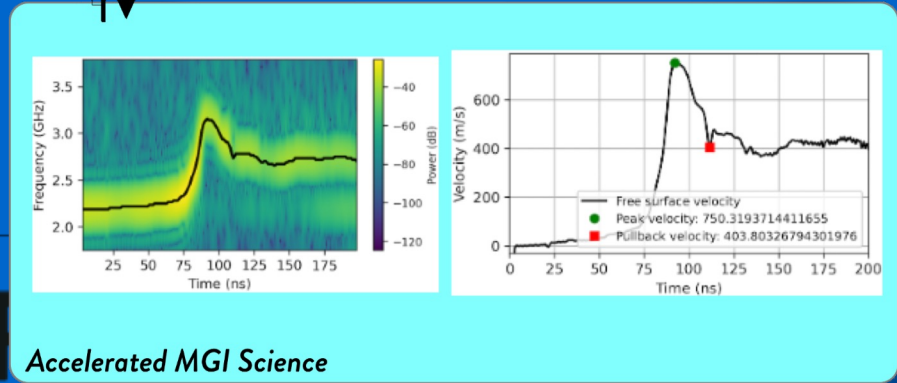
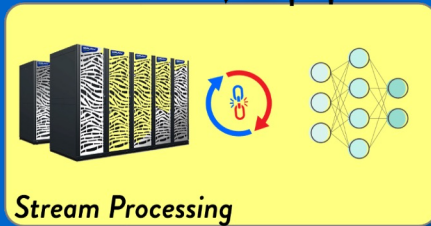
  

Time, Ampl
0.05995000000353, 0.0674841
0.05995000001603, 0.074
0.05995000002853, 0.0777666
0.05995000004103, 0.0774638
0.05995000005353, 0.0747145
0.05995000006603, 0.0699427
0.05995000007853, 0.062688
0.05995000009103, 0.0513761
0.05995000010353, 0.0379689
0.05995000011603, 0.0264511
0.05995000012853, 0.01698
0.05995000014103, 0.0063342
0.05995000015353, -0.00698821
0.05995000016603, -0.0200805

Streaming Layer



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

2030

2035...

Do you know

where your data are?

## Got PID?

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

