Practical Experimental Metadata and Curation Challenges

David Elbert[1], Nick Carey[2], Ali Rachidi[2] , Adam Phelan[1,2] ,Tyrel McQueen[1,2]

1. Hopkins Extreme Materials Institute
2. Department of Computer Sciences
3. Department of Physics and Astronomy
4. Department of Chemistry

contact: elbert@jhu.edu

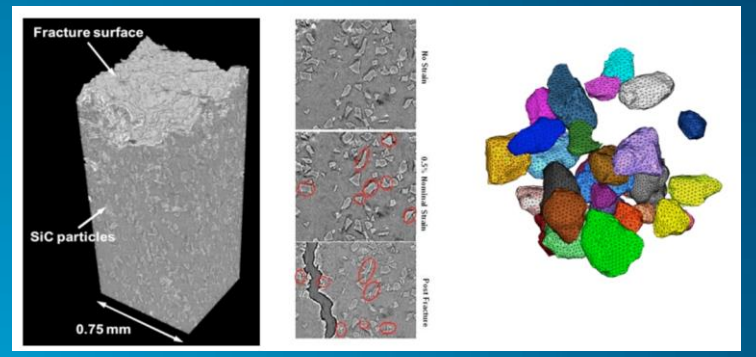
Major Challenges:

- Complex and Diverse Data
- Data Privacy (Personal, Confidential, Sensitive/Classified)
- Citation/Credit for Data
- Meeting Domain Experts Comfort Level
- FAIR Principles (Findable, Accessible, Interoperable, Reusable)
- Connectivity (API)
- Futureproofing
- Metadata Extraction
- Semantics – Connecting Data to Data and Data to Research

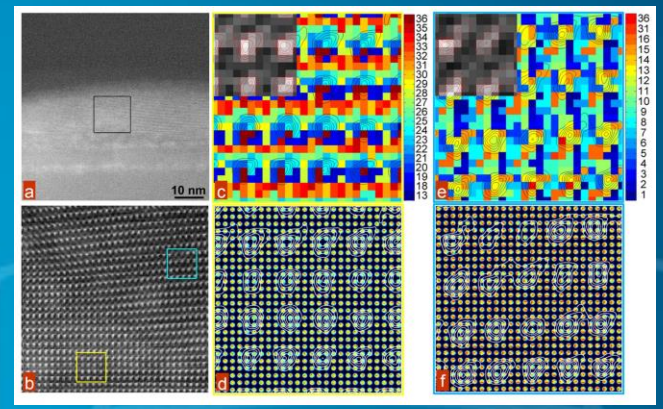
Good News! They fit on one slide…

Materials Big Data

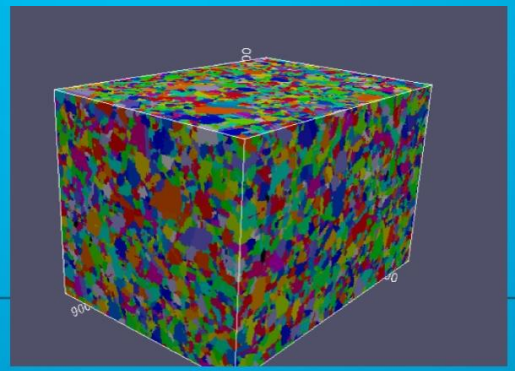
- Higher resolution
- Shorter time scales
- Higher dimensionality
- Dynamic experiments
- Larger simulations
- Tighter processing control



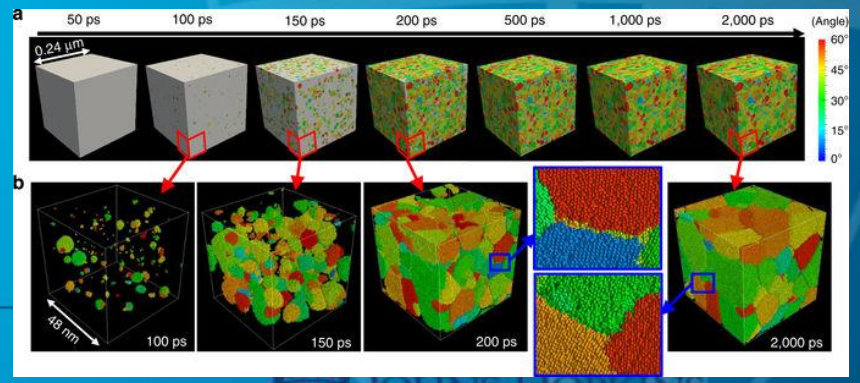
De Carlo et al., 2012



Jesse et al., 2016 3.5 GB/sec



Courtesy Dream3D software



Shibuta et al., 2017

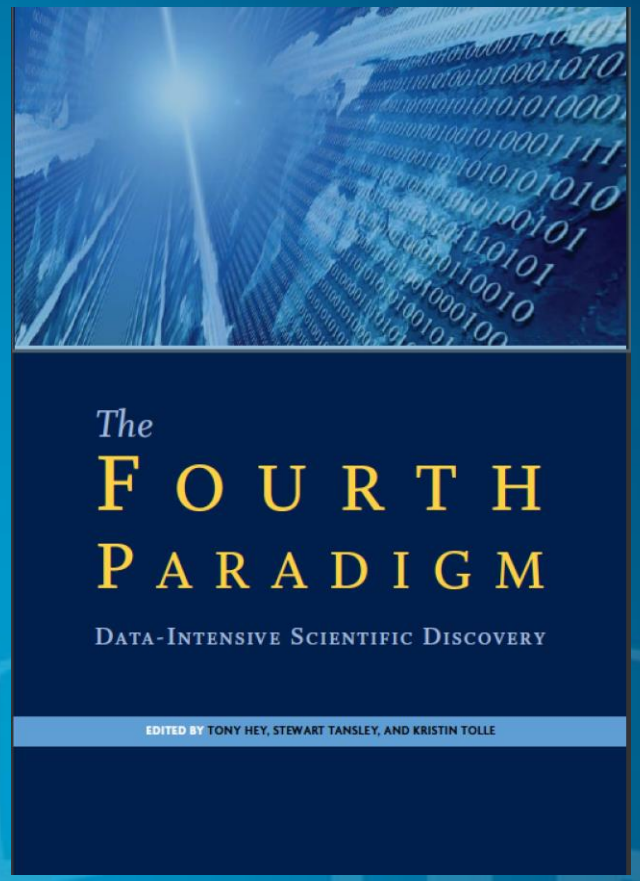
It's all Big Data

"Big data from little acorns grow"

Big Data Changes Everything

When we connect data we change the way we think in three principal ways:

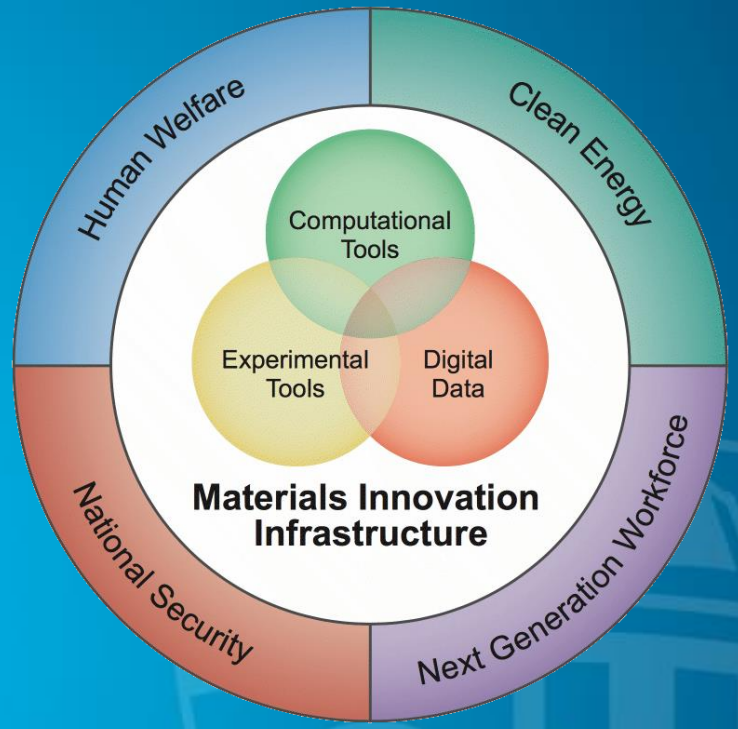
1. Think differently – Experiment Differently
2. Understand differently – Data Driven Science
3. Collaborate differently



Materials Genome Initiative (MGI)

Strategic Goals:

- Facilitate Access to Materials Data
- Equip the Next-Generation Materials Workforce
- **Integrate Experiments, Computation, and Theory**
- Enable a Paradigm Shift in Materials Development



Cross Cutting Themes:

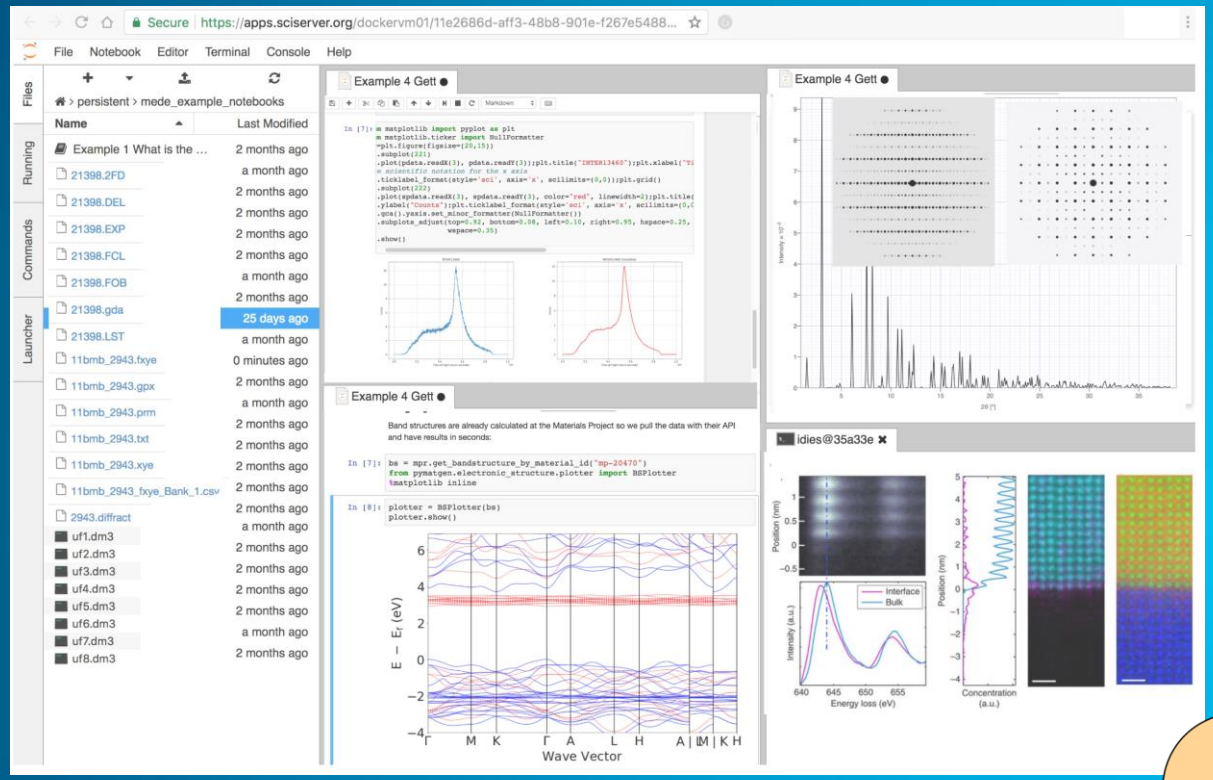
- **Incentivizing open data and access of tools**
- Structuring public-private partnerships
- **Driving innovation across computation, data informatics, and experimentation**
- Moving the community to a different cultural norm

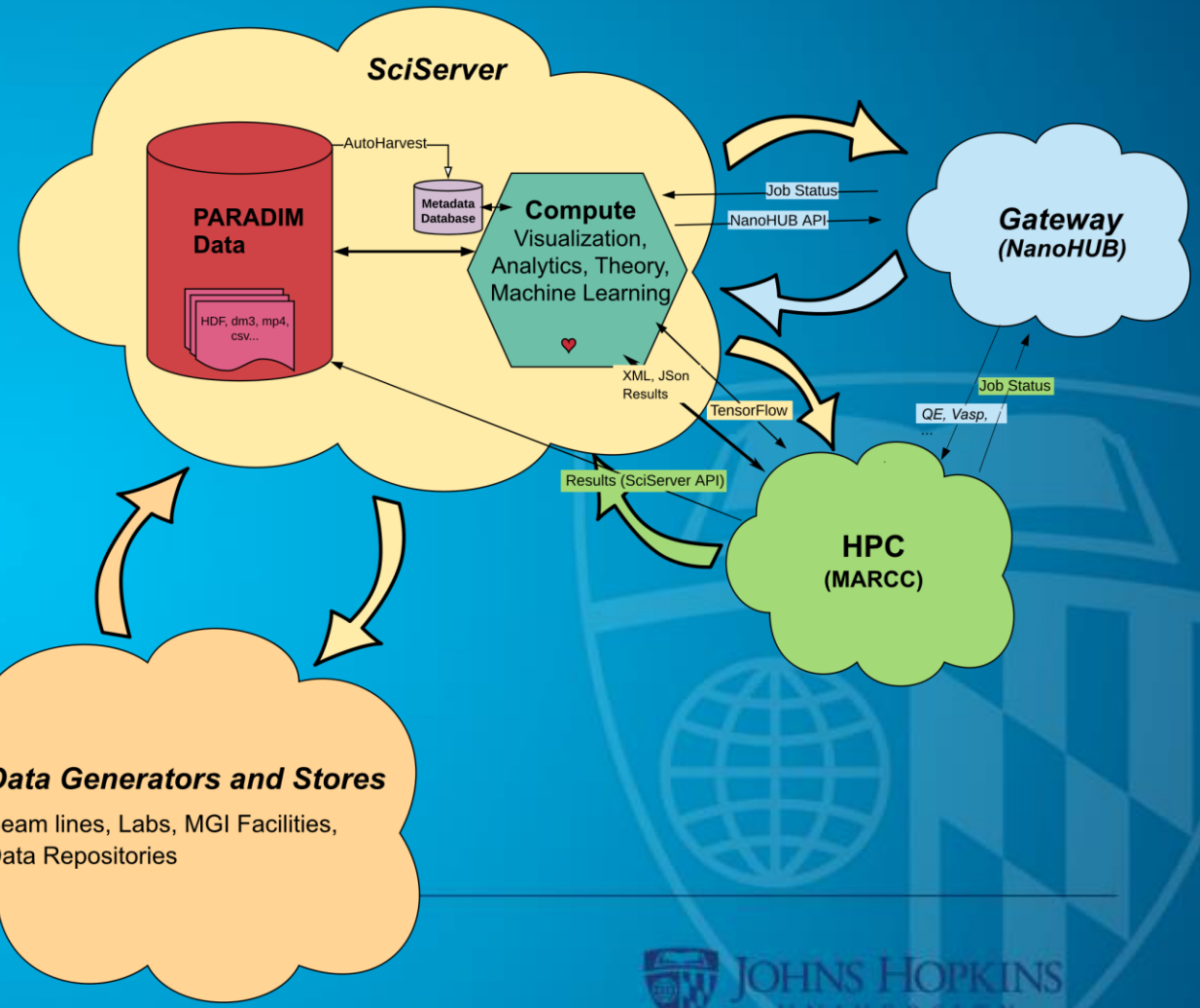
HEMI | Hopkins Extreme Materials Institute

Johns Hopkins University

Refs: https://www.mgi.gov/content/mgi-infographic and https://www.mgi.gov/sites/default/files/documents/wadia_mgi_talk.pdf

PARADIM Data Collective

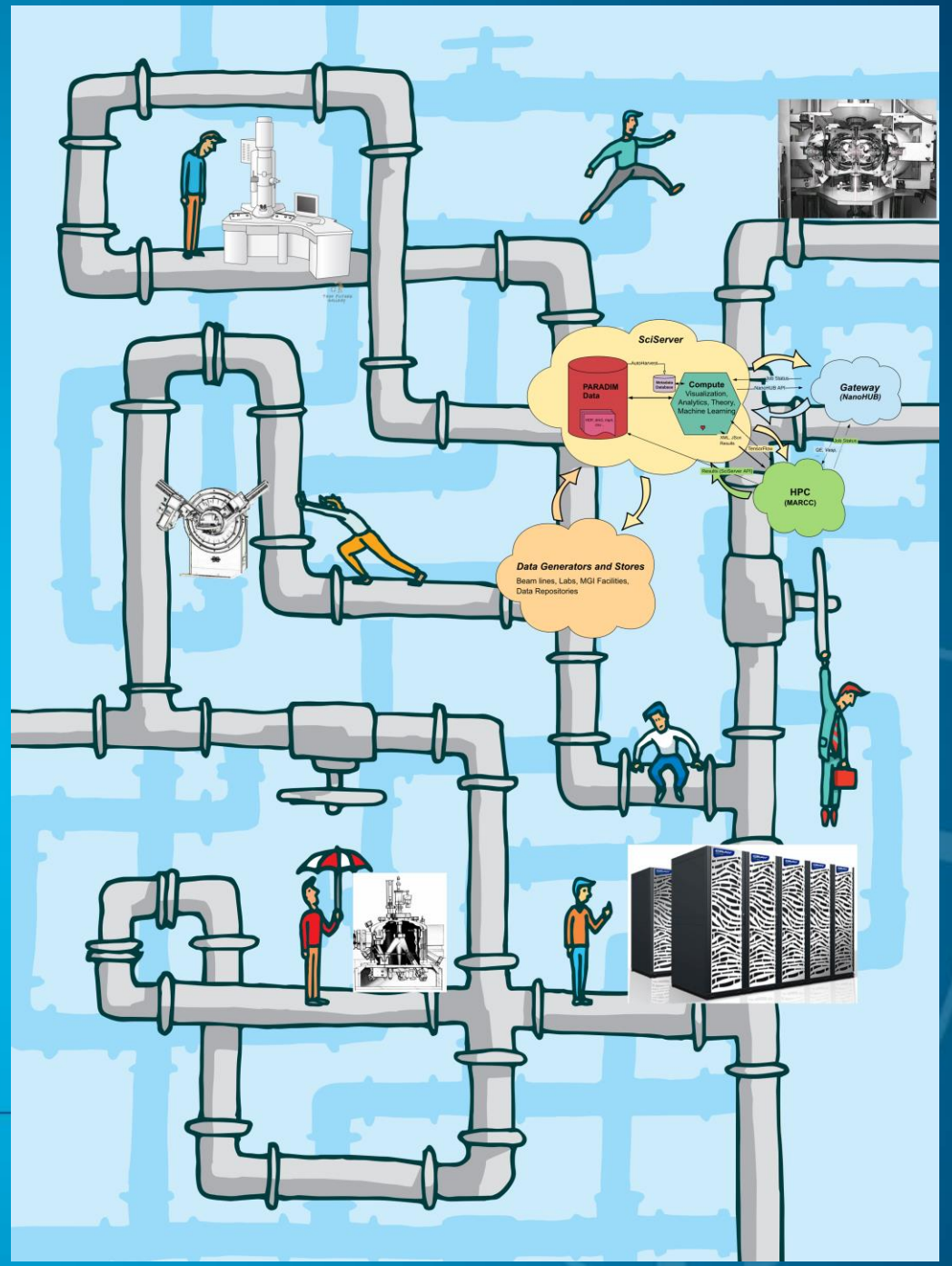




PARADIM Data Pipeline:

How Flow Can We Go??

- Instruments
  - Synthesis
  - Characterization
- Compute
  - Wrangling/Visualization
  - HPC
- Repositories
  - File Stores
  - Databases
- Restful APIs
- Secure and Citable



Repositories: Database vs Files

File Based:

- data are small enough to fit in computer memory
- data are static
- only one person at a time accesses the data
- security is a minor concern

Database:

- Terabytes of data
- Data being updated or added to frequently
- Serving a community of users
- Sensitive

NSF 2D Data Framework Student Workshop

- ~4.5 Days
- 26 students
- Instructors from NIST and Hopkins
- 5 Participant Lightning Talks
- Topics:
  - Terminal Shell
  - Git/GitHub
  - Python and Jupyter Notebooks
  - Databases (SQL/NoSQL)
  - Basic Data Wrangling in Python
  - Materials APIs
  - Atomistics from Notebook

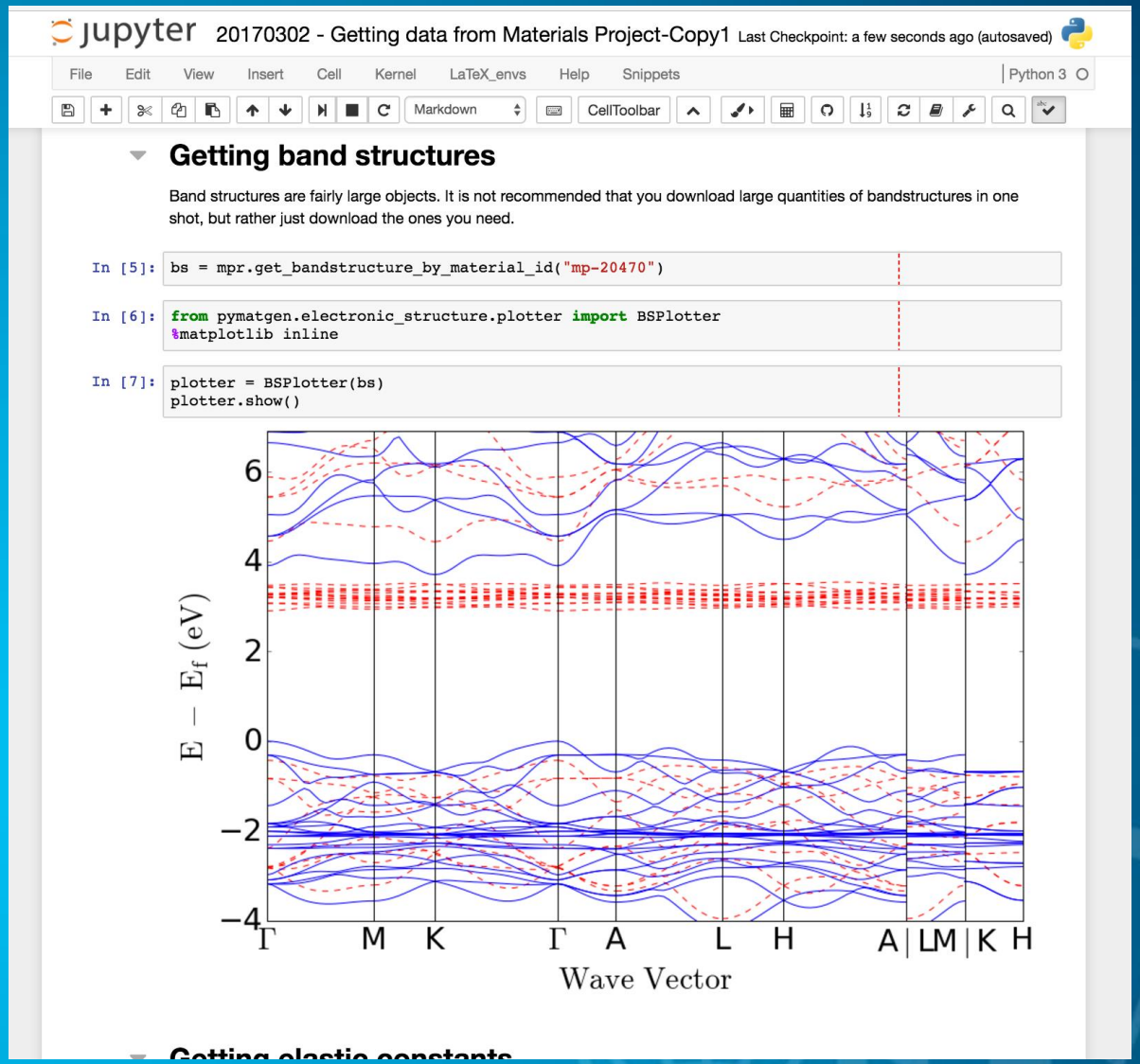
NSF 2D Data Framework Student Workshop

Goals for the week:

- New skills to work with materials data
- Better appreciation of MGI-related, materials data resources
- Motivation to expand your data science skill set
- New friends

Jupyter Notebooks ….the future is Python

- Interactive Computing Environment
  - Live code, narrative text, mathematics,
  - plots and rich media in one document
- "Narrates Computation and Analysis"
- Flexible
  - Diverse tools match diversity of work
  - Facilitates interactive collaboration
- "Reproducible" Analysis
- Extensible
  - > 30,000 related GitHub repositories
- Increase efficiency
  - Combine Analysis with Visualization
  - Implicit sharing of workflow
- "BFT" – Better, Faster, Traceable Science



HEMI HOPKINS EXTREME MATERIALS INSTITUTE

JOHNS HOPKINS UNIVERSITY

NSF 2D Data Framework Student Workshop

Key Take Aways:

- Most students need data skills
- It is important to them
- We need to better bridge the computational community
- Students ***want*** a community

open as possible in providing us helpful feedback.

At the beginning of the week we told you our goals for the workshop were for students to acquire:

- New skills to work with materials data
- A better appreciation of MGI-related, materials data resources
- Motivation to expand your data-science skill set
- New friends

In that context, please answer the following:

Questions (Do you know why the list starts with zero?):	**1 = disagree** **5 = agree strongly** n/a means not applicable due to your prior expertise
0. Before the workshop I was very skilled at related topics	1 2 3 4 5 n/a
1. I have improved my data-related skills	1 2 3 4 5 n/a
2. I learned about data related tools or resources new to me	1 2 3 4 5 n/a
3. I learned new things about data related tools or resources I was already aware of or using	1 2 3 4 5 n/a
4. The expertise of the instructors was appropriate	1 2 3 4 5 n/a
5. I have a better appreciation of the breadth of materials tools available to or being developed in the community	1 2 3 4 5 n/a
6. My appreciation for data-centric approaches to science has expanded	1 2 3 4 5 n/a
7. Learning data-related skills and methodologies has the potential to be important to my career path	1 2 3 4 5 n/a
8. My motivation to continue learning data-science methods has increased	1 2 3 4 5 n/a
9. I hope to share data-centric skills and ideas with my home research group	1 2 3 4 5 n/a
10. Incorporating data science methods in my research will expand my career opportunities	1 2 3 4 5 n/a
11. I enjoyed the methods and approaches of the workshop	1 2 3 4 5 n/a
12. I found the interaction with fellow attendees valuable	1 2 3 4 5 n/a
13. I would like to maintain connections with the people I met at the workshop	1 2 3 4 5 n/a

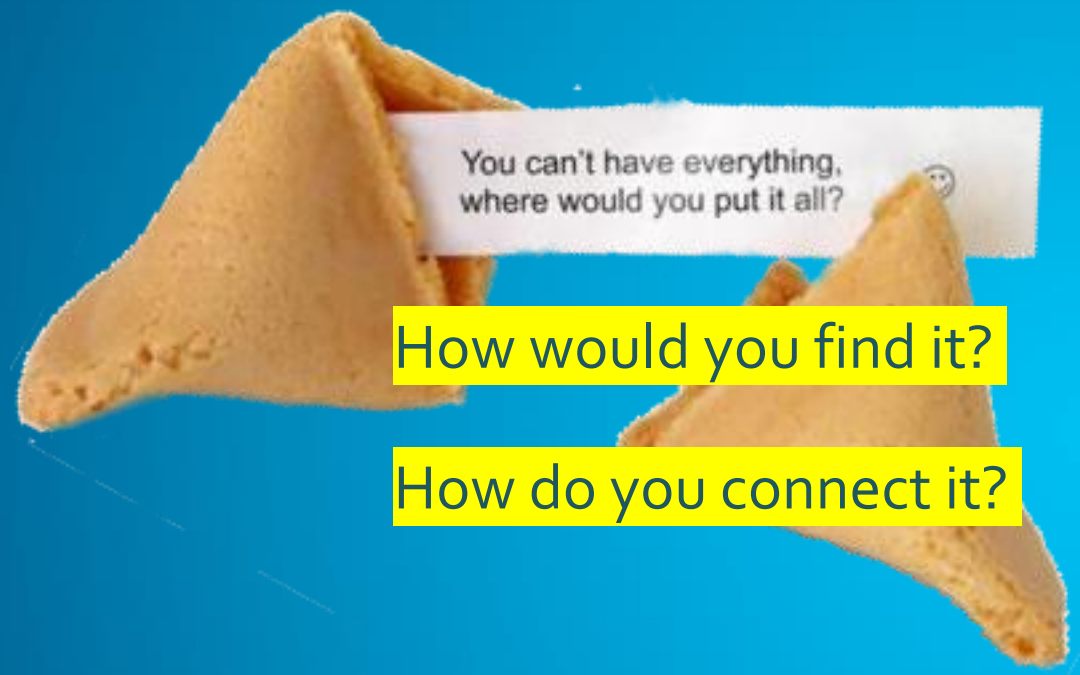
HEMI | JOHNS HOPKINS EXTREME MATERIALS INSTITUTE

Major Challenges:

- Complex and Diverse Data
- Data Privacy (Personal, Confidential, Sensitive/Classified)
- Citation/Credit for Data
- Meeting Domain Experts Comfort Level
- FAIR Principles (Findable, Accessible, Interoperable, Reusable)
- Connectivity (APIs)
- Futureproofing
- Metadata Extraction
- Semantics – Connecting Data to Data and Data to Research

Metadata Are Data About Data

Metadata Connect Data



How would you find it?

How do you connect it?

Obvious Metadata:

1. Investigator
2. Material
3. Date
4. Funding
5. …

Less Obvious:

1. History of the material
2. History of the data

"Metadata is a love note to the future"

-Jason Scott (2011)

Data

Metadata

What do we want to know?

1. What's in the jar? When was it made? Who made it?
2. How many jars were made each year?
3. Did production drift towards different flavors over the last 20 years? If so, why?

Fundamental Challenge: What metadata do we need?

Answer: Depends on what you'll ask of the data

Question: How do you winnow metadata to make it manageable? (Create a useful data model?)

Answer: Define 20 queries.

"Most selections involving human choices follow a "long tail," or so-called 1/f distribution, it is clear that the relative information in the queries ranked by importance is logarithmic, so the gain realized by going from approximately 20 (24.5) to 100 (26.5) is quite modest." -Jim Gray

Example: Microscopy Metadata (Lehigh/NIST Workgroup)

Resource metadata (who, when, where)
Modality:
    Probe: Light, SEM, TEM, SPM, X-ray Microscope…
    Contrast: BSE, BF, DF, HR, STEM (HAADF, BF, DF), Diffraction (CBED, SAED), EFTEM, Lorentz…
Length scale (pixels to lengths, etc))
Spatial and data dimension:
Features(polycrystalline, phases, texture, twins, etc.)
    Interpreter
        Submitter
        Other User
        Agent
    Interpretations
        Polycrystal, Dendrite, Twinned, Precipitate…
    Confidence (need serious discussion here)) (
sample identity (nominal chemistry, name), unique identifiers
image identity (Not sure what this means?)
file format
    Lossless (Y/N)
    PNG, Tiff, PDF, gif, other suffix
    Size in bytes
errors?
related hyperspectral data
        Orientation Map
pixel range of raw data
crystal symmetry
        Space group
        Point group
        Ordered/Disordered
reference frame (Euler angles and conventions)
what hyperspectral information is recorded
        OIM
        EDS
        WDS
        EELS
        HAADF
        XANES, etc…
DOI to published paper or some other related resource
instrumentation (use emerging RDA spec)
experimenters (ORCID, funding)
processing conditions (primary, secondary, tertiary)
simulation code (name, version, repo, DOI)

Lather, Rinse, Repeat!

HEMI | Hopkins Extreme Materials Institute

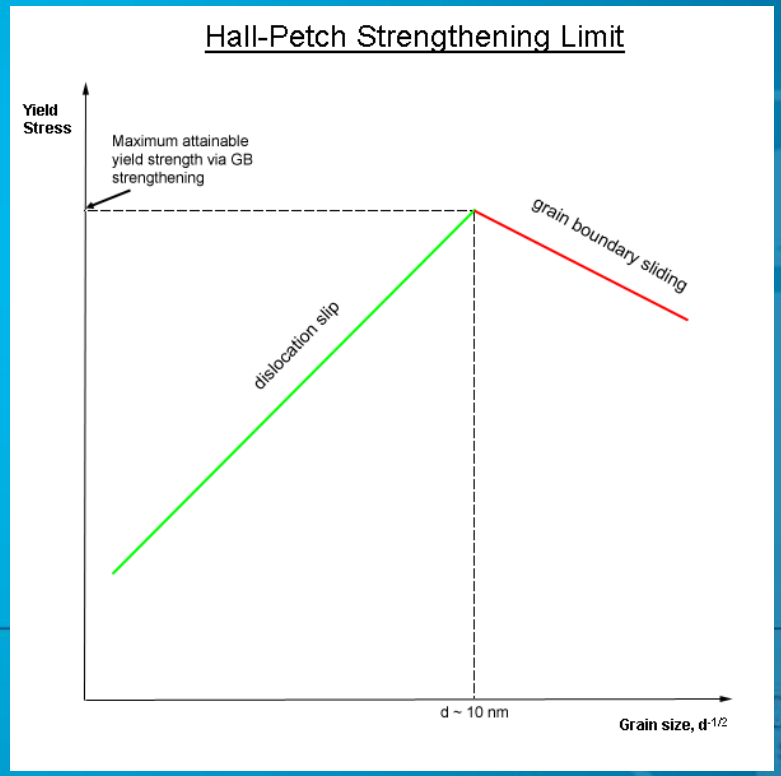
Johns Hopkins University

Example: Microscopy Metadata

Query: What's the relationship between grain size and material strength?

Can we probe microscopy data to see if the Hall-Petch Relationship is the whole story?

$$\sigma_y = \sigma_0 + \frac{k_y}{\sqrt{d}}$$



Data-Centric Materials Science and Engineering

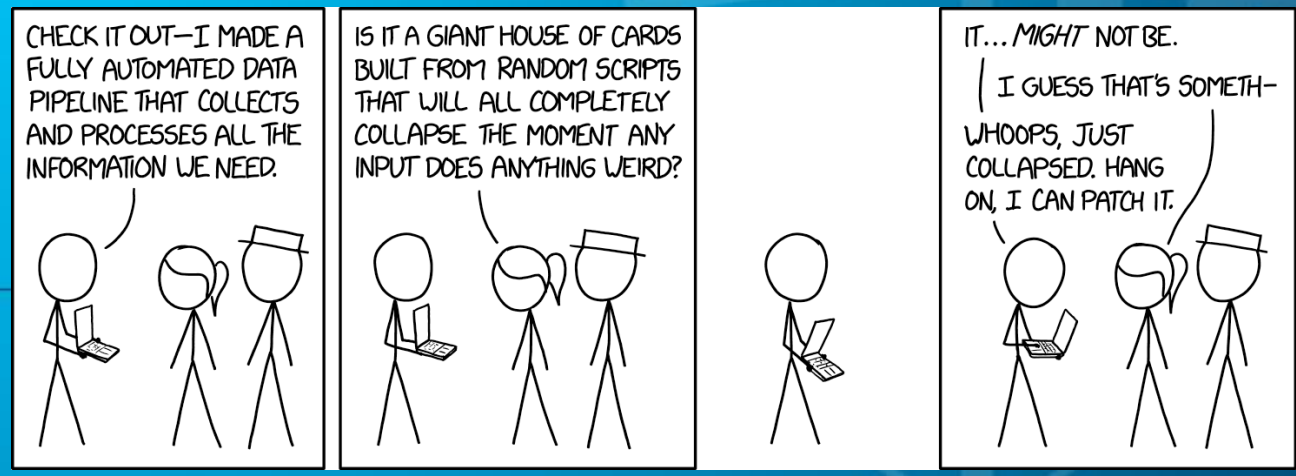
"What do we need?"

- Community structure to steer
  - Leadership not Dictators
  - Academia *and* Industry
    - Applications that meet real needs
    - Data monetization and paywalls?
  - "modern" Data Management Plan
    - Really publish your data and have it count
  - Funded projects to create data rather than only traditional answers
  - New vision for publication
- Vision for Persistent Support
  - We build it, "they will come", it better still be there!
- Curricular Modernization (Code not GUIs (Python not Excel); M.L., etc.)
- Community Valuation of Data

Data-Centric Investigations are *Services (Beyond SaaS, PaaS…)*

"Data is the contract between services"

- Complex, Diverse Data with Privacy and Citation Needs
- Educating our Domain Experts and Future Domain Experts
- Empowering Connection of our Data (RESTful APIs)
- Futureproofing – Infrastructure Persistence
- Semantics – Connecting Data to Data and Data to Research
- **Group Buy-In and Community Work**
  - **This work is fun and exciting!**
- **Plenty of concrete needs**



https://xkcd.com/2054/