Integrating Big Data Analysis with Data Management in the Materials Domain: The SciServer Perspective

David Elbert[1, 2], Nick Carey[2], and Tamas Budavari [1, 3, 4]

1. Hopkins Extreme Materials Institute
2. Earth and Planetary Sciences
3. Computer Sciences
4. Applied Mathematics and Statistics

Johns Hopkins University

contact: elbert@jhu.edu

Powered by SciServer

Johns Hopkins University

idies The Institute for Data Intensive Engineering and Science

CMEDE Center for Materials in Extreme Dynamic Environments

19-Mar-2018

Materials Community has a Lot of Data

In science it is not enough to think of an important problem on which to work. It is also necessary to know the means which could be used to investigate the problem.

– *Leo Szilard*

Fundamental Questions:

- ***How does one work with all that data?***
- ***What's new about what we can accomplish if we use a data-science perspective?***

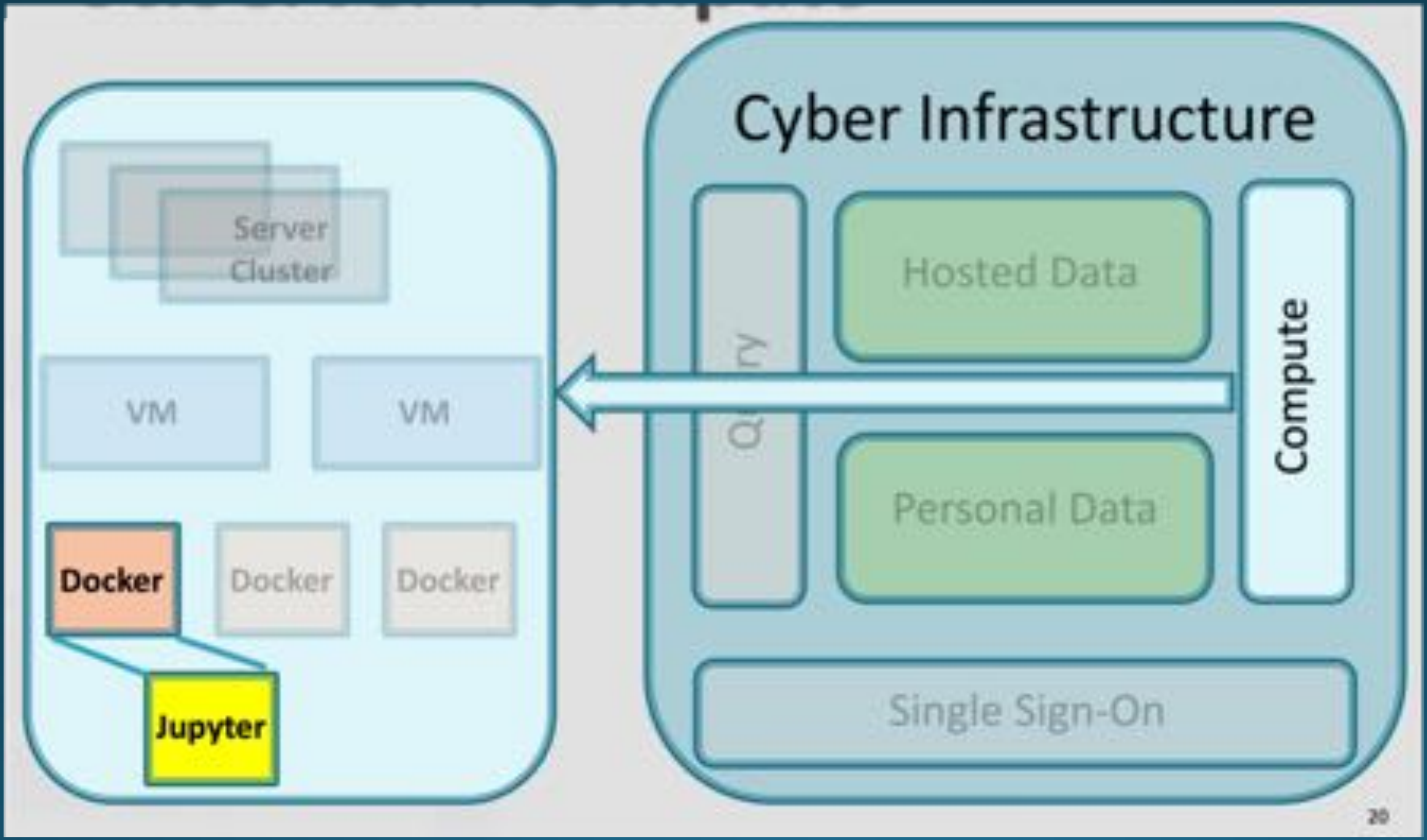
Data Management for the Future not just for the Investigator

SciServer – Data Centric

“bring the analysis to the data”

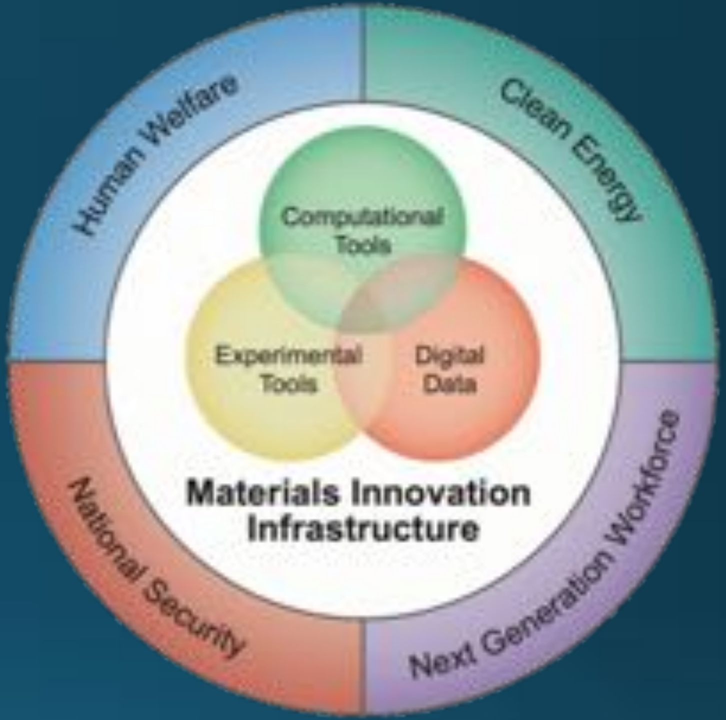
- NSF Data Infrastructure Building Block Center (DIBB)
  - ~$10M over five years
- Host and serve petabyte datasets (1024 terabytes = million gigabytes)
- Scalable compute services
- Analysis on datasets too large to load into memory or handle locally
- Core Components
  - Compute: Data Analysis in Jupyter Notebooks – Python, MatLab, R kernels
  - SciDrive: Data Storage using Binary Large OBject (BLOB) storage for unstructured data
  - CASJobs: Database storage and query – SQL/Set Theory Science

SciServer Architecture: Data and Analysis in the Cloud



Materials Genome Initiative (MGI)

-*Discover, Develop, and Deploy Twice as Fast*



Strategic Goals:

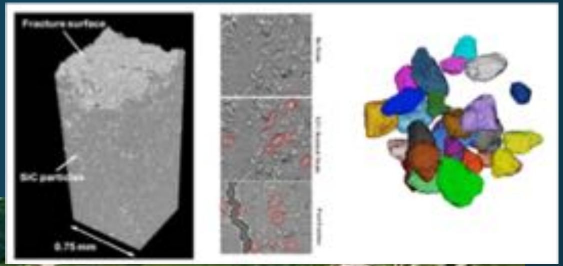
- Facilitate Access to Materials Data
- Equip the Next-Generation Materials Workforce
- Integrate Experiments, Computation, and Theory
- Enable a Paradigm Shift in Materials Development

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

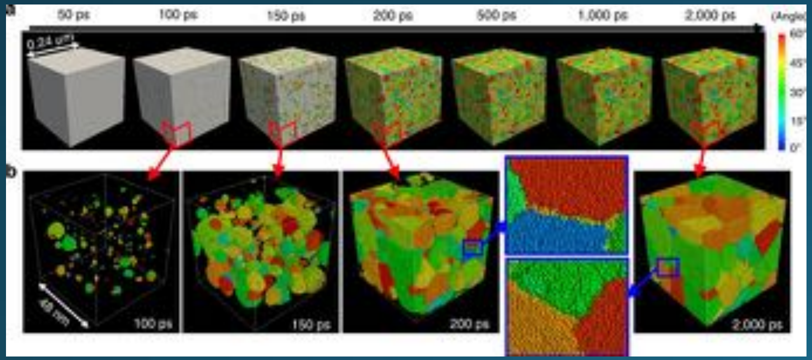
Refs: https://www.mgi.gov/content/mgi-infographic and https://www.mgi.gov/sites/default/files/documents/wadia_mgi_talk.pdf

De Carlo et al., 2012

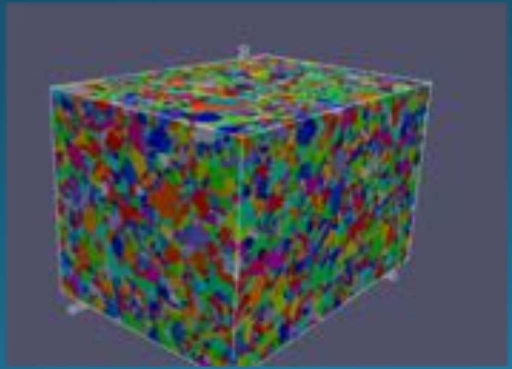


It Is More than Beamlines

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



Shibuta et al., 2017



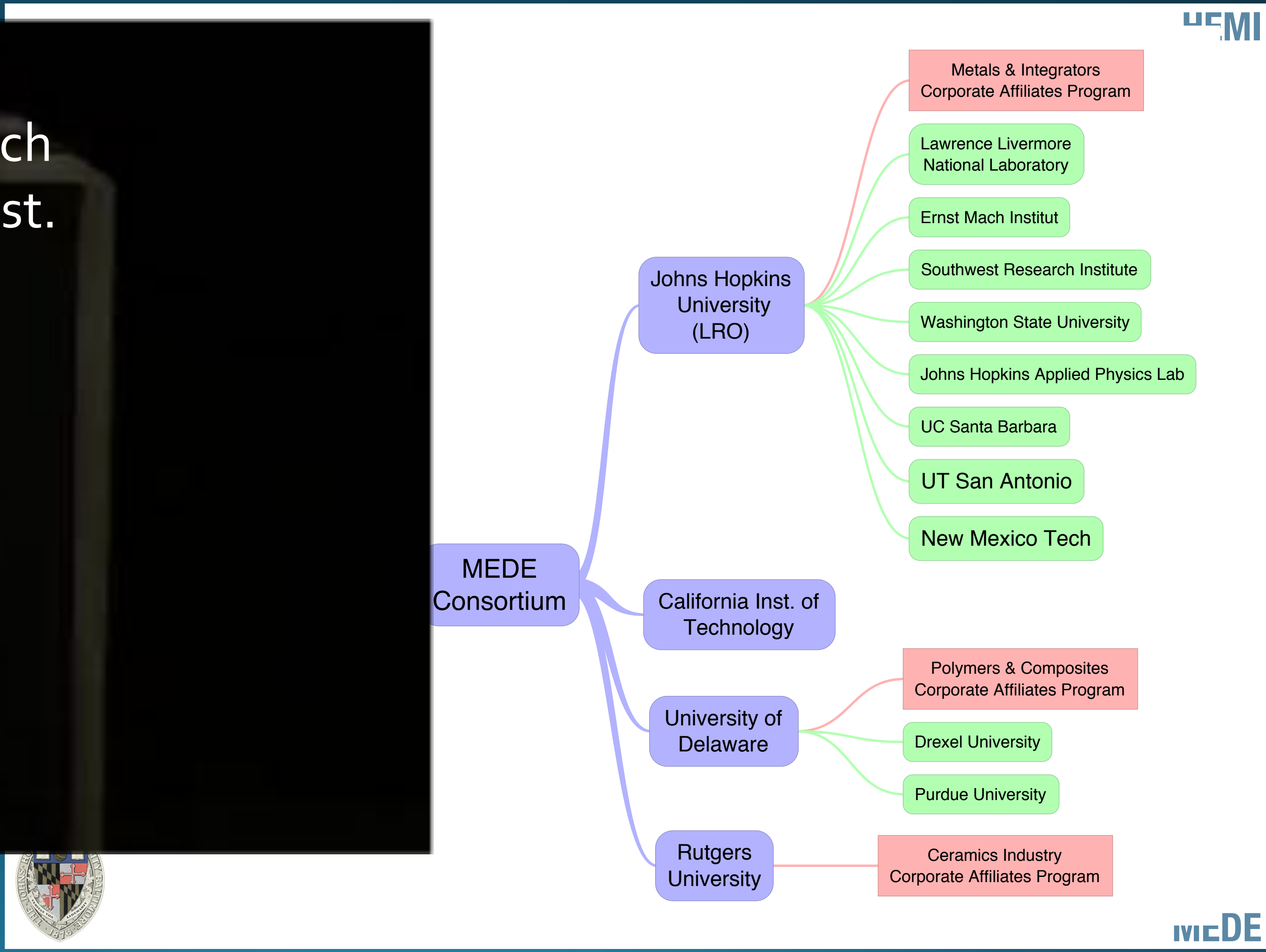
Courtesy Dream3D software

Materials in Extreme Dynamic Environments

Collaborative Research Alliance with Army Research Lab

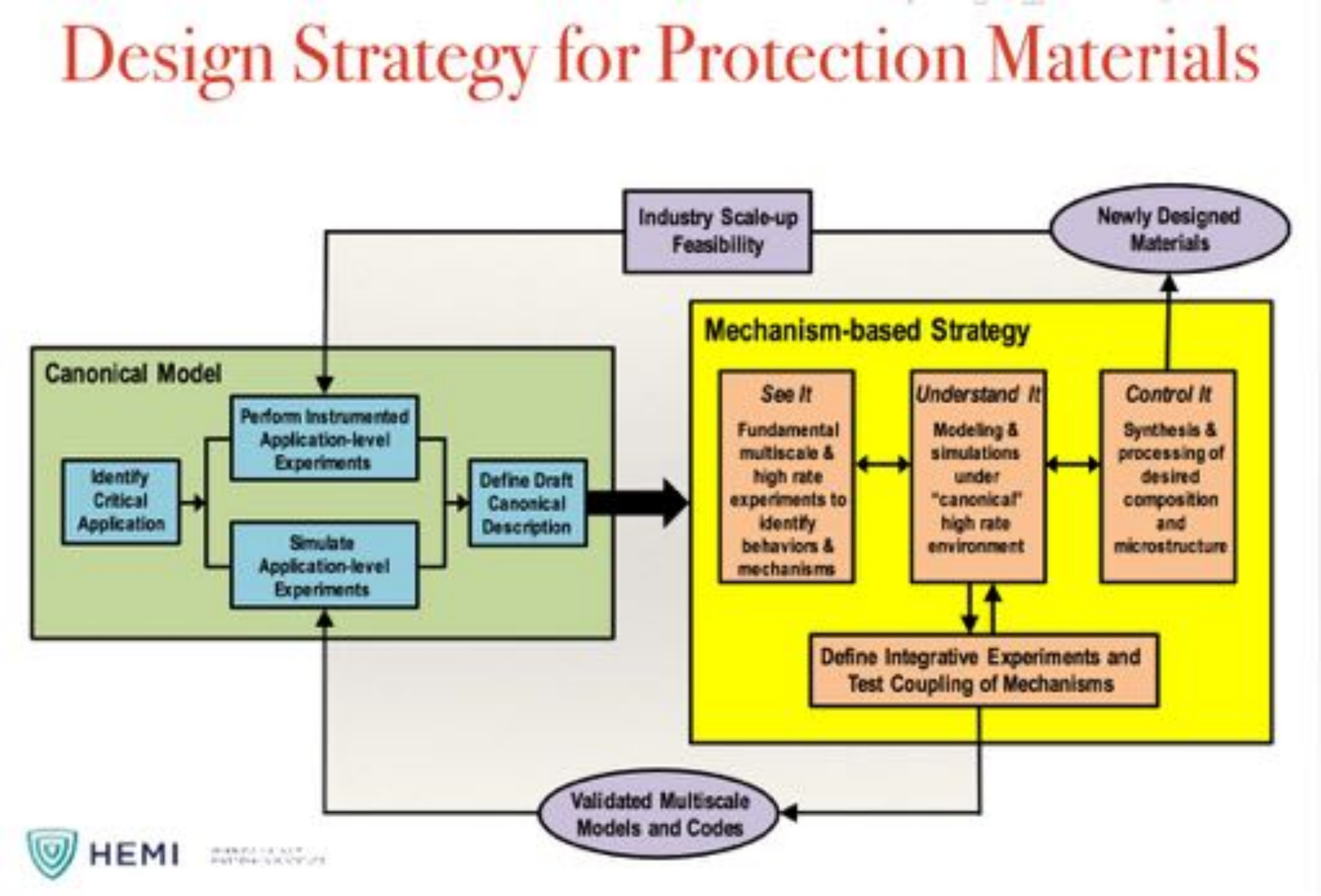
- Multi-institution collaborative research
- Part of Hopkins Extreme Materials Inst. (HEMI)
- Academia, industry, and ARL

"As the local energy density increases, the energy dissipation in the system must explore smaller and smaller length scales."

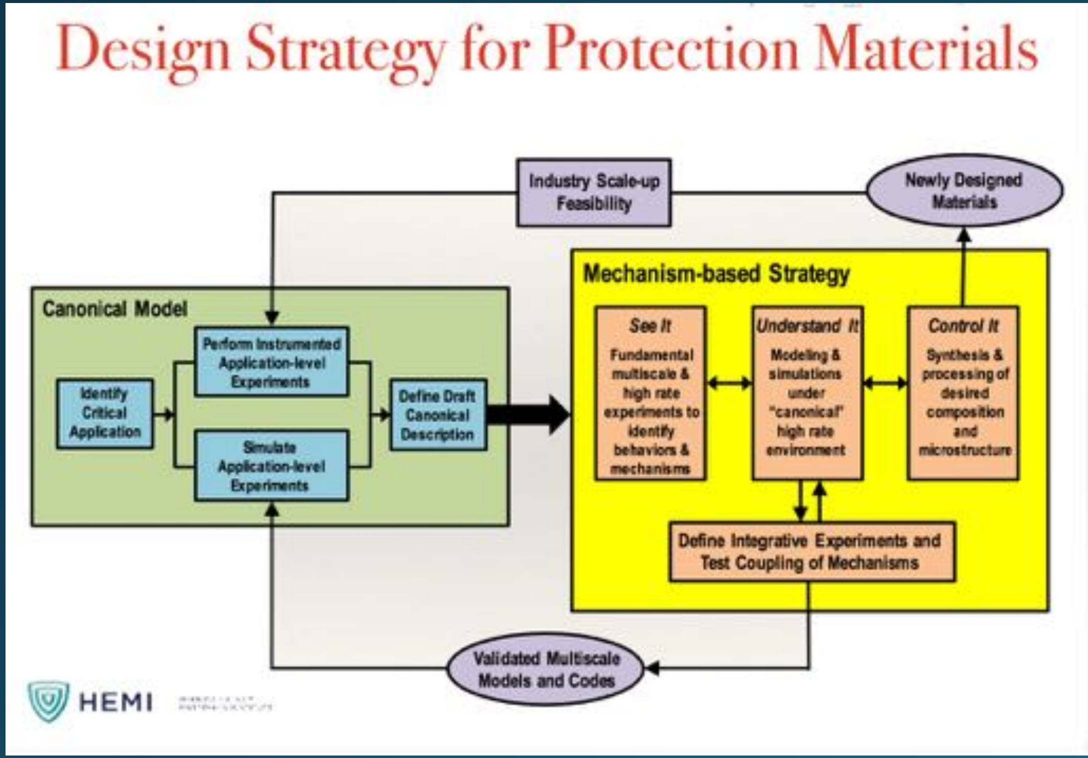


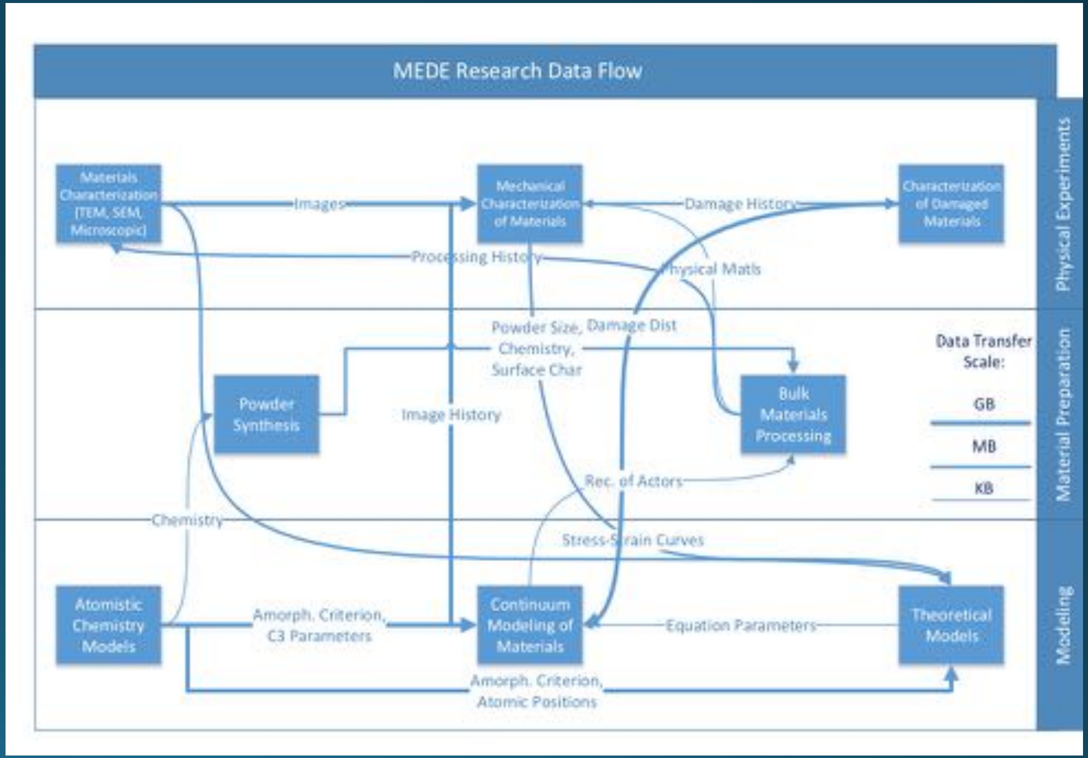
Friday, April 12, 13

Materials By Design



Bring the Data Science to MEDE





MEDE–DSC (Data Science Cloud)

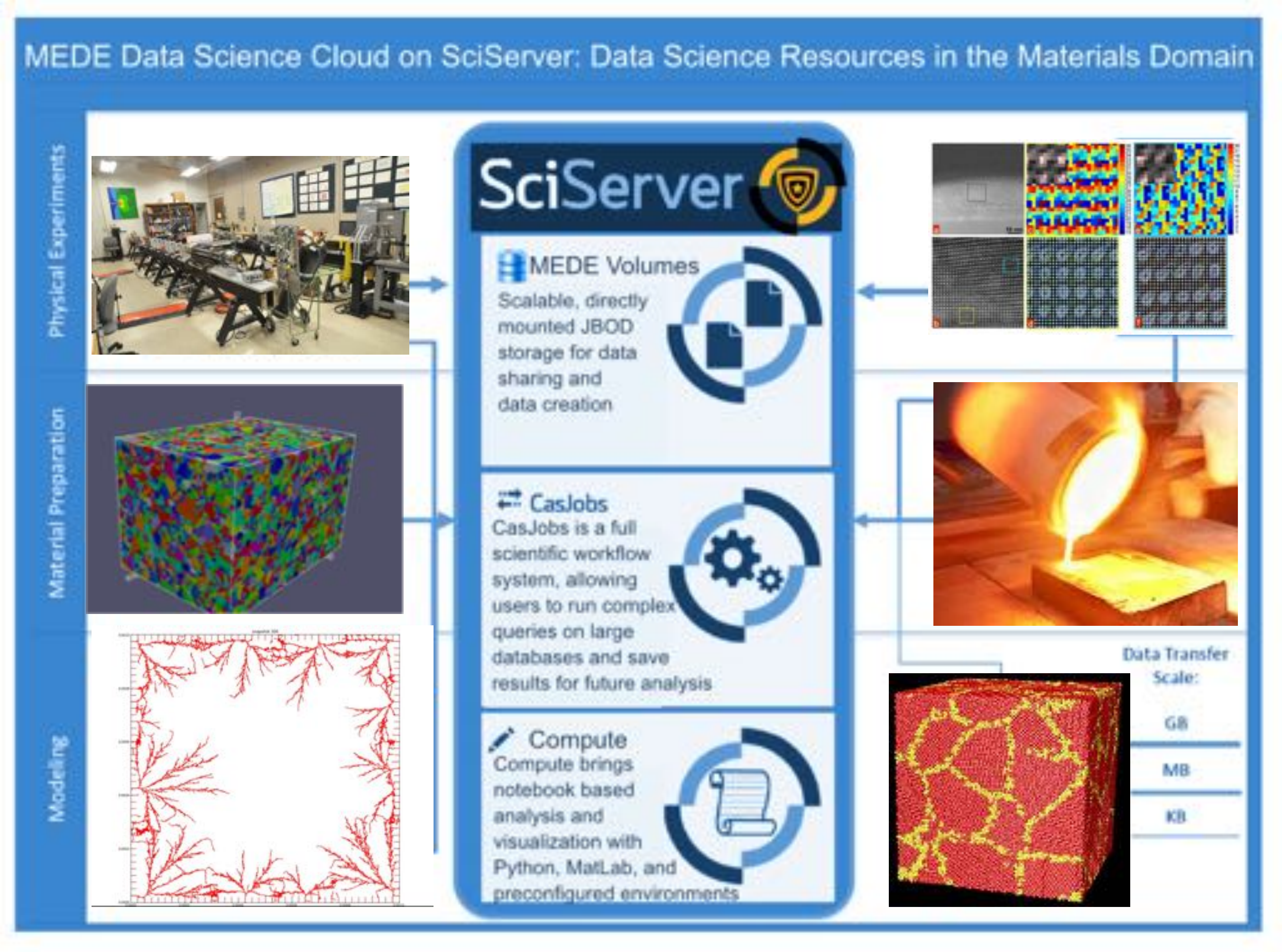
Linking MEDE Experiment, Computation and Theory

Architecture:

- Built on NSF DIBB SciServer
- Scalable, Shared Data Volumes
- Data Ingress Tools
- Materials Computation Environments
- Training and Workshops
- API Integration

Highlights:

- Foundation of Shared Analysis
- Big Data Analytics
- Tools to Advance Collaboration
- Repeatable/Reproducible Science

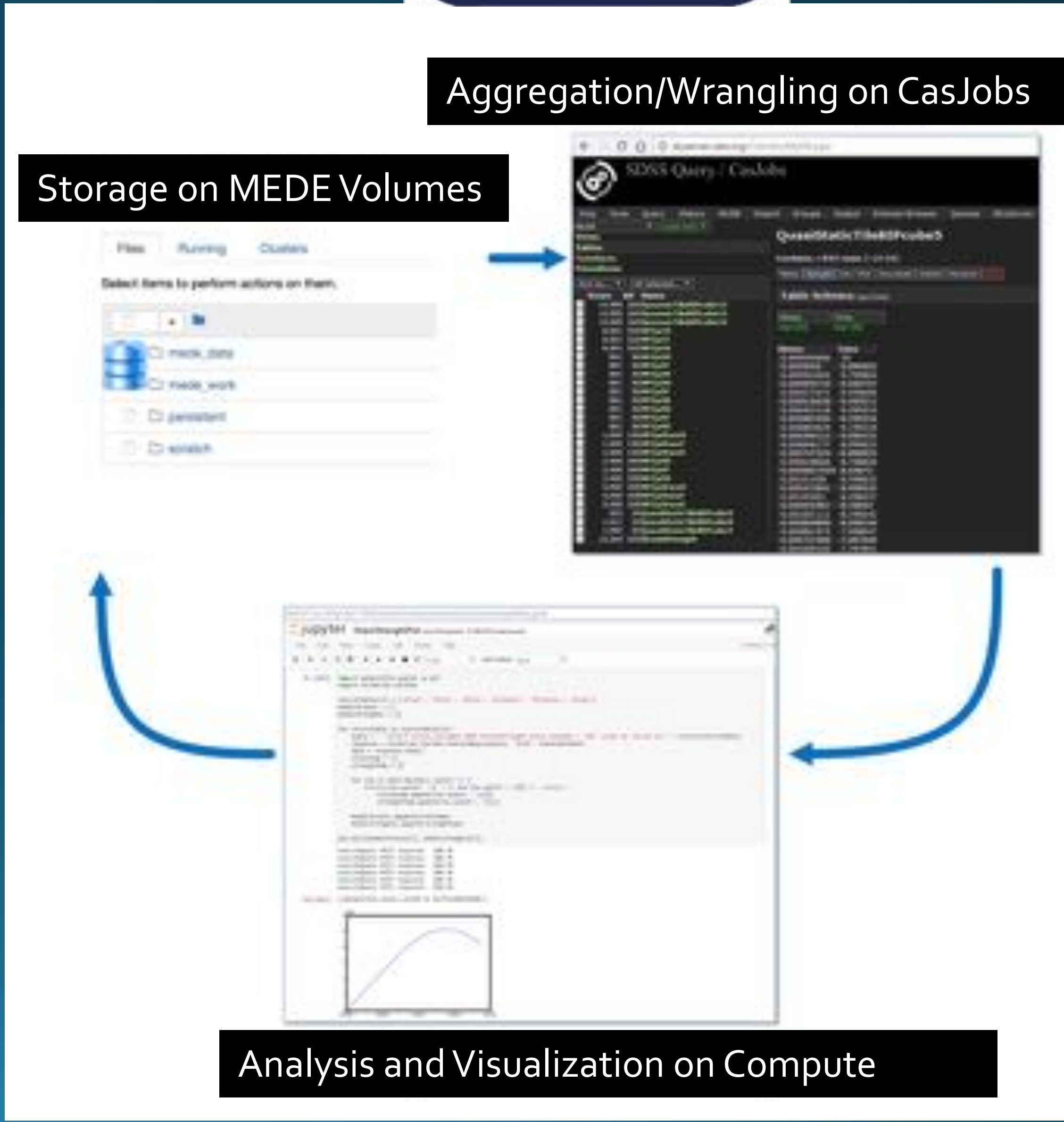


MEDE – DSC

Linking MEDE Experiment, Computation and Theory

Bring the analysis to the data

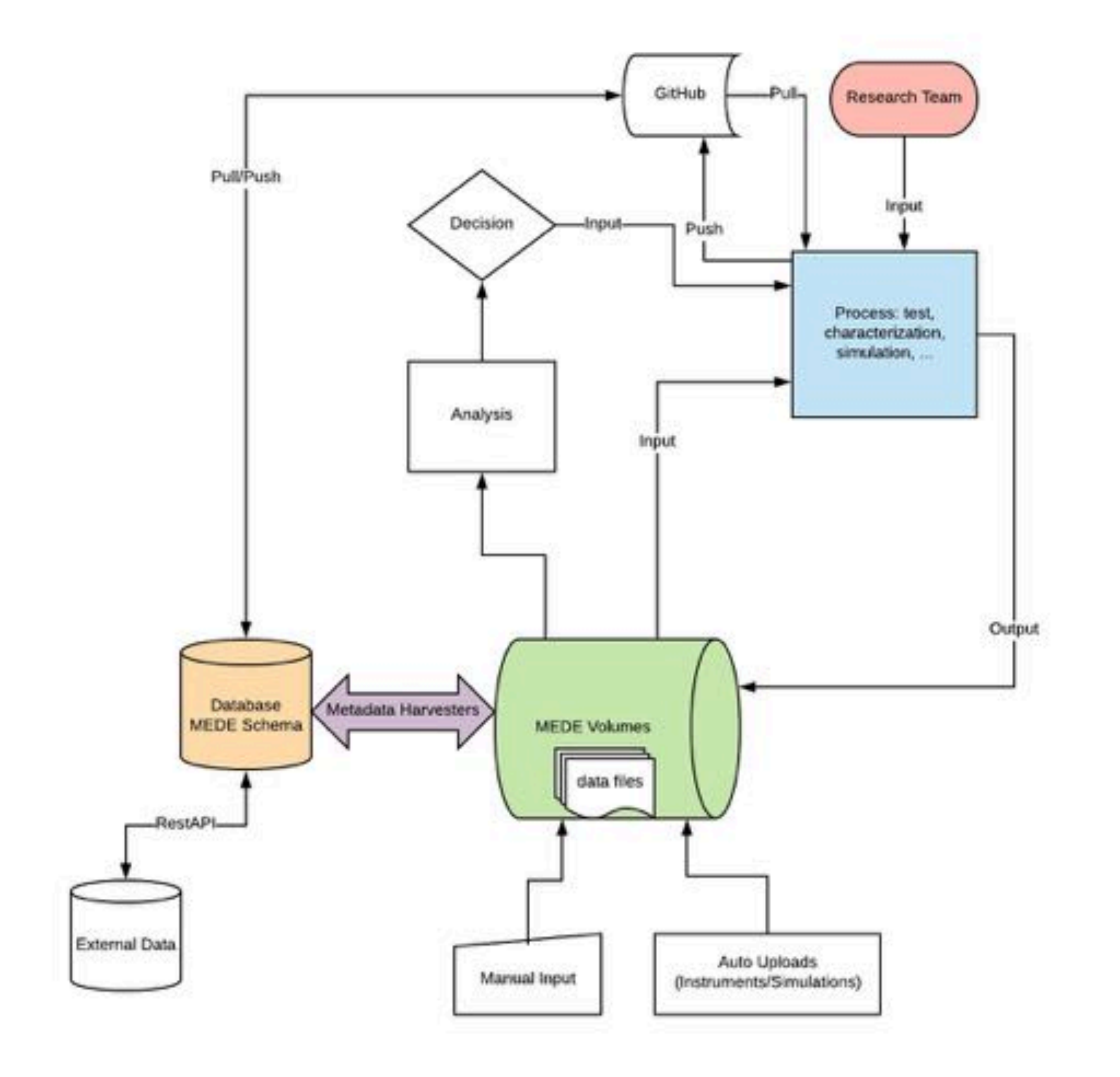
- Visualization and analysis in materials tailored Compute containers
  - Python 3/2
  - MatLab,
  - R, Julia, Ruby if requested
  - Materials packages
- Scalable, virtual machine architecture
- Analysis in the Database



Data Object Outline

- Whole workflow data centric
- Harvester links to database

Open Metadata is Critical



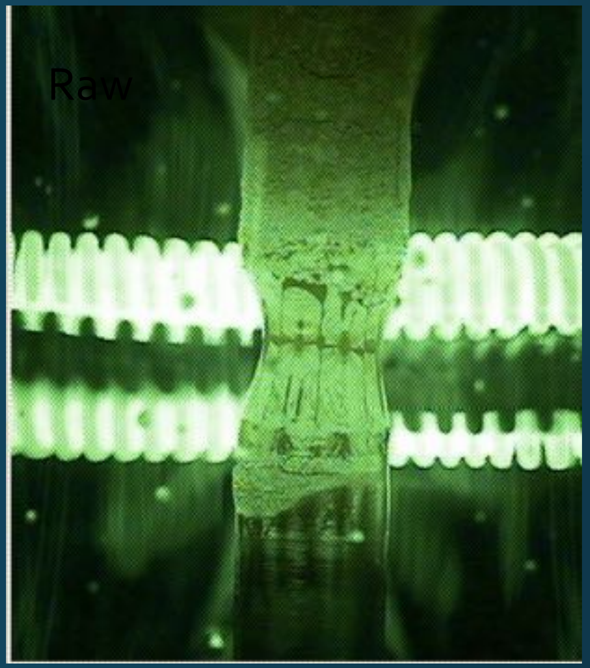
MEDE – DSC

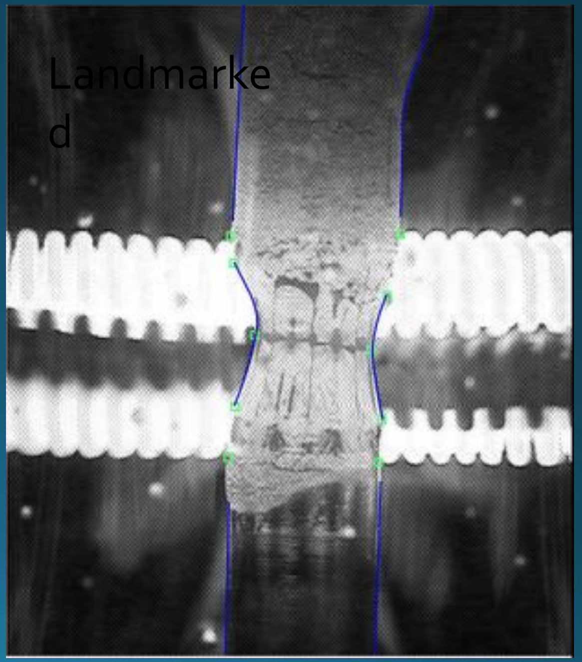
Enabling Novel Approaches:

Floating Zone Furnace Assisted Synthesis:

Big Data Facilitate Deep Learning

- Preprocessing
- Develop Parametric Deformable Object Model Training Stack:
- Supervised Learning:
  - Machine landmark identification
- Analytics
- Create Real-time Processing

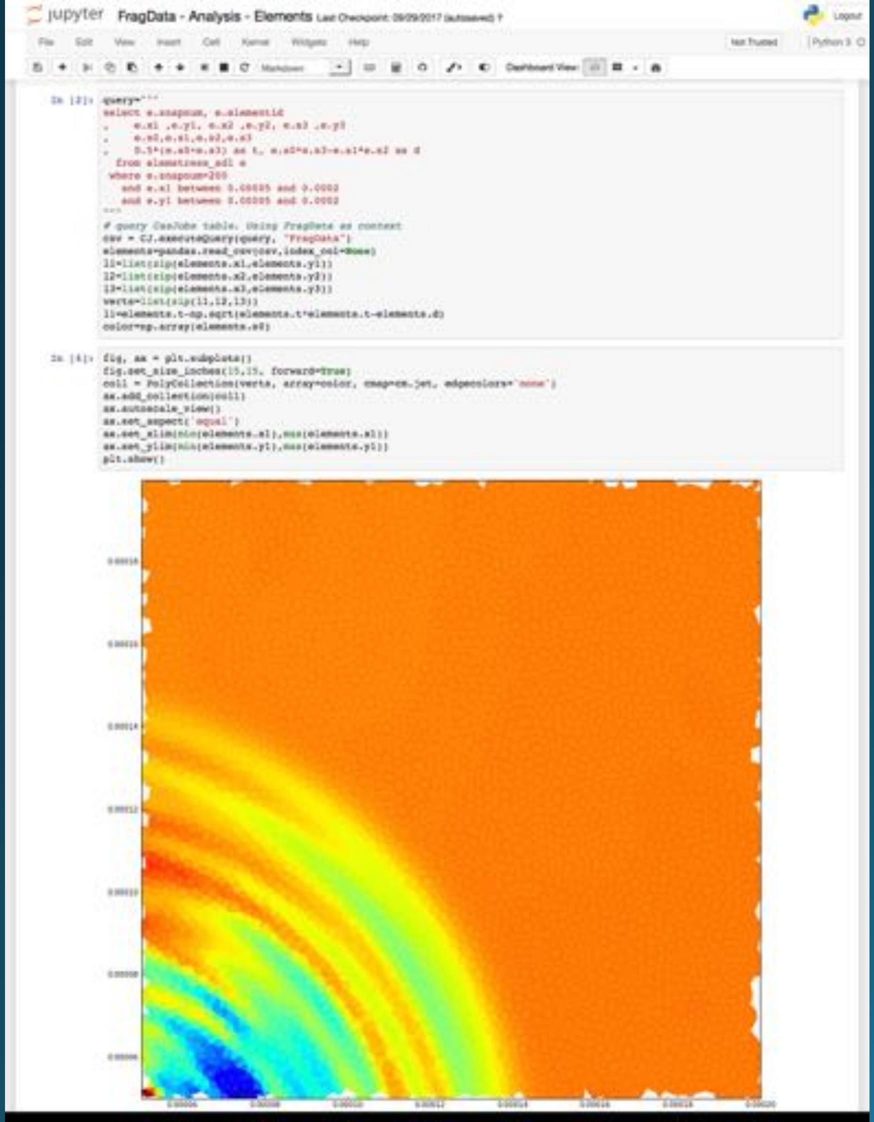


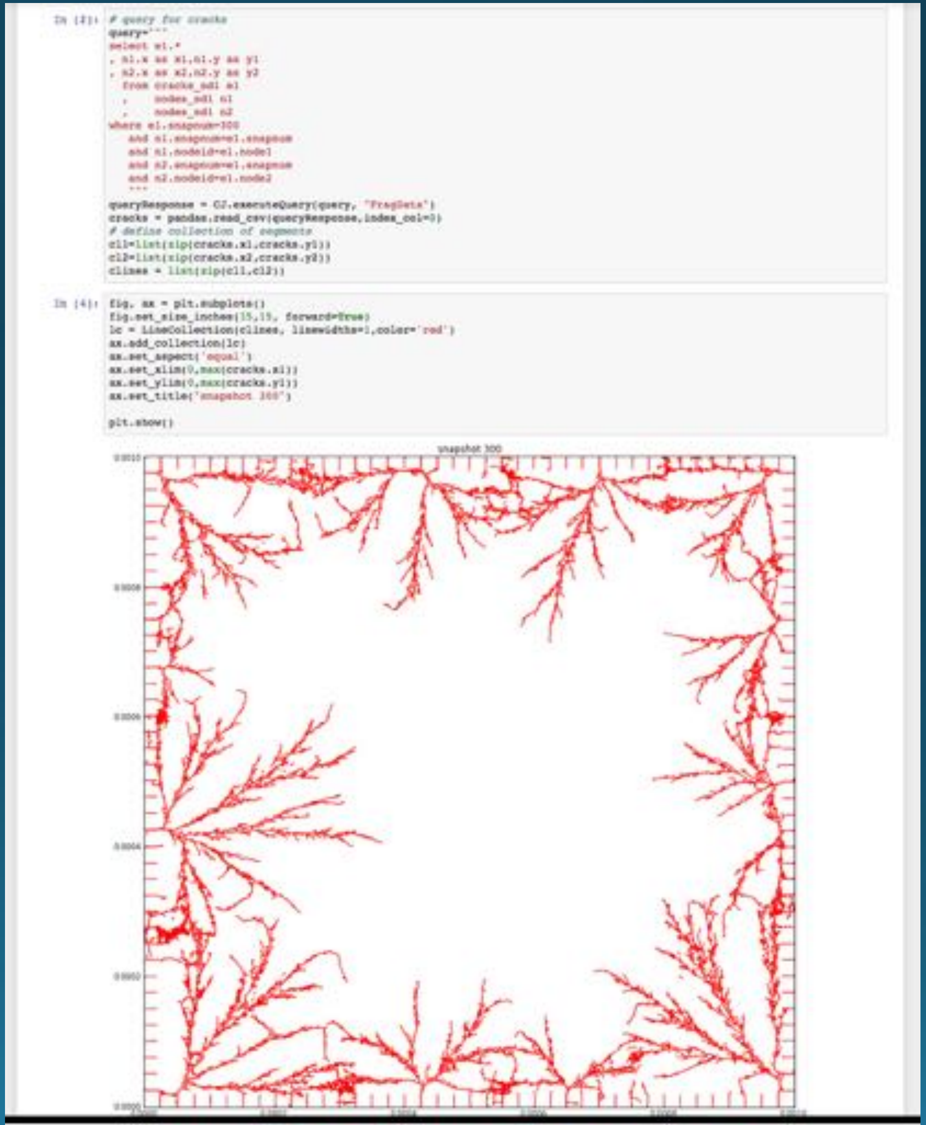


PARADIM MIPS collaboration. Images courtesy of Tyrel McQueen

MEDE – DSC

Enabling Novel Approaches:





Daphalapurkar and Lemson, in prep

Live Demo

SciServer – Data Centric

"bring the analysis to the data"

- Infrastructure and Tools
- Can be tailored to the users
  - "Materials tools on a silver platter"
- Integrate with the larger Materials Community
- Attention to data ingress and training

In science it is not enough to think of an important problem on which to work. It is also necessary to know the means which could be used to investigate the problem.

– *Leo Szilard*