Accelerating Time to Science *sans* Human Interaction:

Materials Data Science Enabled by Integration of Distributed & High Performance Computing

Roger H. French
Director, Materials Data Science for Stockpile Stewardship (MDS^3) COE
Faculty Director, Applied Data Science
Case Western Reserve University, Cleveland OH

co-Directors: Laura Bruckman, Yinghui Wu

Strengthening NNSA's Capability to Modernize Manufacturing & Production

Modern materials science research produces petabyte-scale, heterogeneous datasets that span multiple modalities. Coherently integrating such data presents a significant unsolved challenge not addressed by current high performance computing approaches. CRADLE, an infrastructure and framework tackles these materials data science challenges in several ways: 1) scaling to handle large, diverse datasets through distributed computing and vertical scaling; 2) supporting the full data lifecycle from data ingestion to model deployment; 3) providing accessible tools that enable novice to experienced users to construct end-to-end machine learning pipelines.

We demonstrate "CRADLE analytics" on terabyte-scale multi-modal data at scale through four exemplar cases: 1) photovoltaic (PV) power time series imputation using generative graph neural networks given billions of power measurements, 2) integrating geospatial data to track fertilizer runoff, 3) X-ray Diffraction (XRD) analysis of in-situ movies, and 4) crack/precipitate analysis with summary graph generation on timeseries X-ray Computed Tomography (XCT) creep test datasets.

"SDLE Research Center" & Materials Data Science

Create Cross-cutting Solutions Based in Materials Data Science

SDLE's Research Centers and Funded Projects

MDS-RELY | CASFER ERC | MDS^3-COE | PV, Edifice, EERS

FAIRifcation, Domain Ontologies, Study Protocols, Global Geospatial Repository, netSEM, PVimage, ddiv,....

SDLE Core Facility: Equipment, Tools, Outdoor Testing, CRADLE Computing, Benchmark Datasets,

SDLE as a Data Science Incubator and Foundation

SDLE Research Center:
Develops Materials Data Science Solutions & Leading Edge Materials Science Research
Across a broad range of research projects and centers

Common Research Analytics & Data Lifecycle Environment
CRADLE Computing & Analytics

- Integrate Distributing Computing
  - "Scaled Out Computing"
- With High Performance Computing
  - "Scaled Up Computing"

Agile Team Science

- Agile Manifesto for Software Development
  - Slack, Jira KanBan, Confluence, Bitbucket
- Use 4 Month Long Cross-cutting Sprints

SDLE Research Center: Acknowledgements

25 UG | 26 GR | 4 Postdocs | 3 Staff | 20 Faculty

Create Cross-cutting Solutions Based in Materials Data Science

SDLE's Research Centers and Funded Projects

MDS-RELY | **CASFER ERC** | **MDS^3-COE** | **PV, Edifice, EERS**

Agile Team Science

Useful "Home Page" for ATS: https://start.atlassian.com/

Confluence Spaces
- SDLE Lab Meetings
- MDS^3 Meetings
- SDLE Wiki
- Collaborators

Jira KanBan Boards
- Admin Boards
- Research Packages Boards
- Project Boards

Bitbucket (Git) Repositories
- Project Repositories
- Research Package Repositories
- Manuscript LaTeX Repositories
- Thesis LaTeX Repositories

SLACK Team Messaging

SDLE Core Facility: Equipment, Tools, Outdoor Testing, CRADLE Computing, Benchmark Datasets

SDLE as a Data Science Incubator and Foundation

SDLE Research Center:

Develops Materials Data Science Solutions & Leading Edge Materials Science Research

Across a broad range of research projects and centers

The vision of the MDS^3 COE

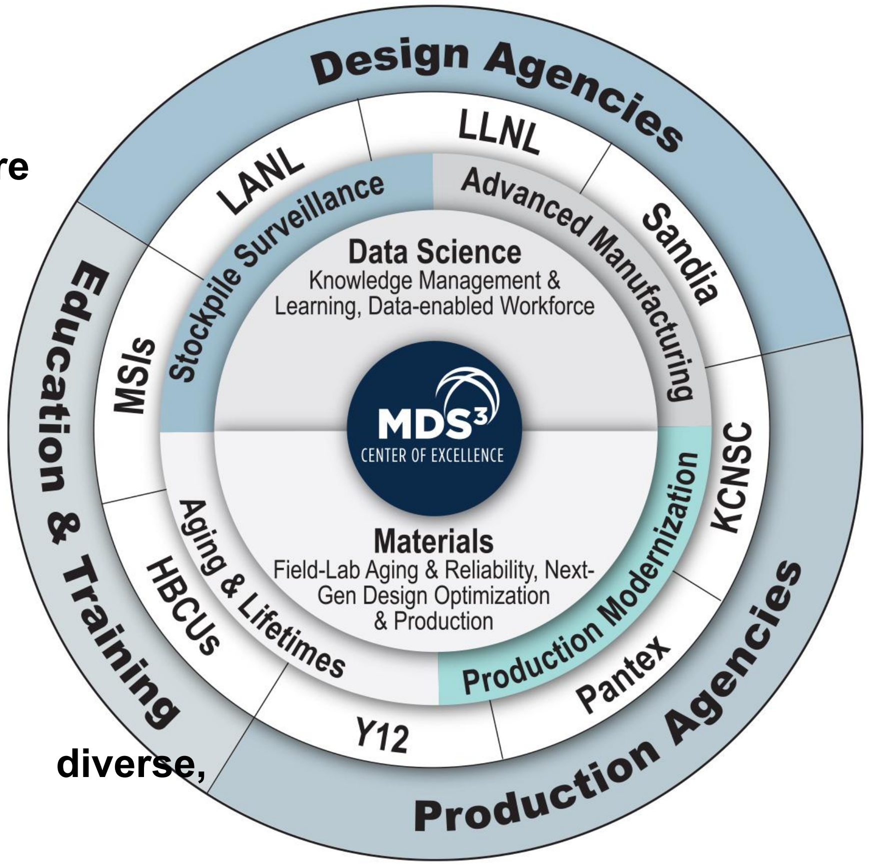
- **Develop, demonstrate, and deploy**
  - Novel Materials Data Science (MDS) tools
- **Frameworks, codes, and computing infrastructure**
  - "Research Packages"

- **To advance our understanding of**
  - Materials degradation
  - Parts Design and Optimization for Fabrication
  - Failure of materials, parts, and subsystems

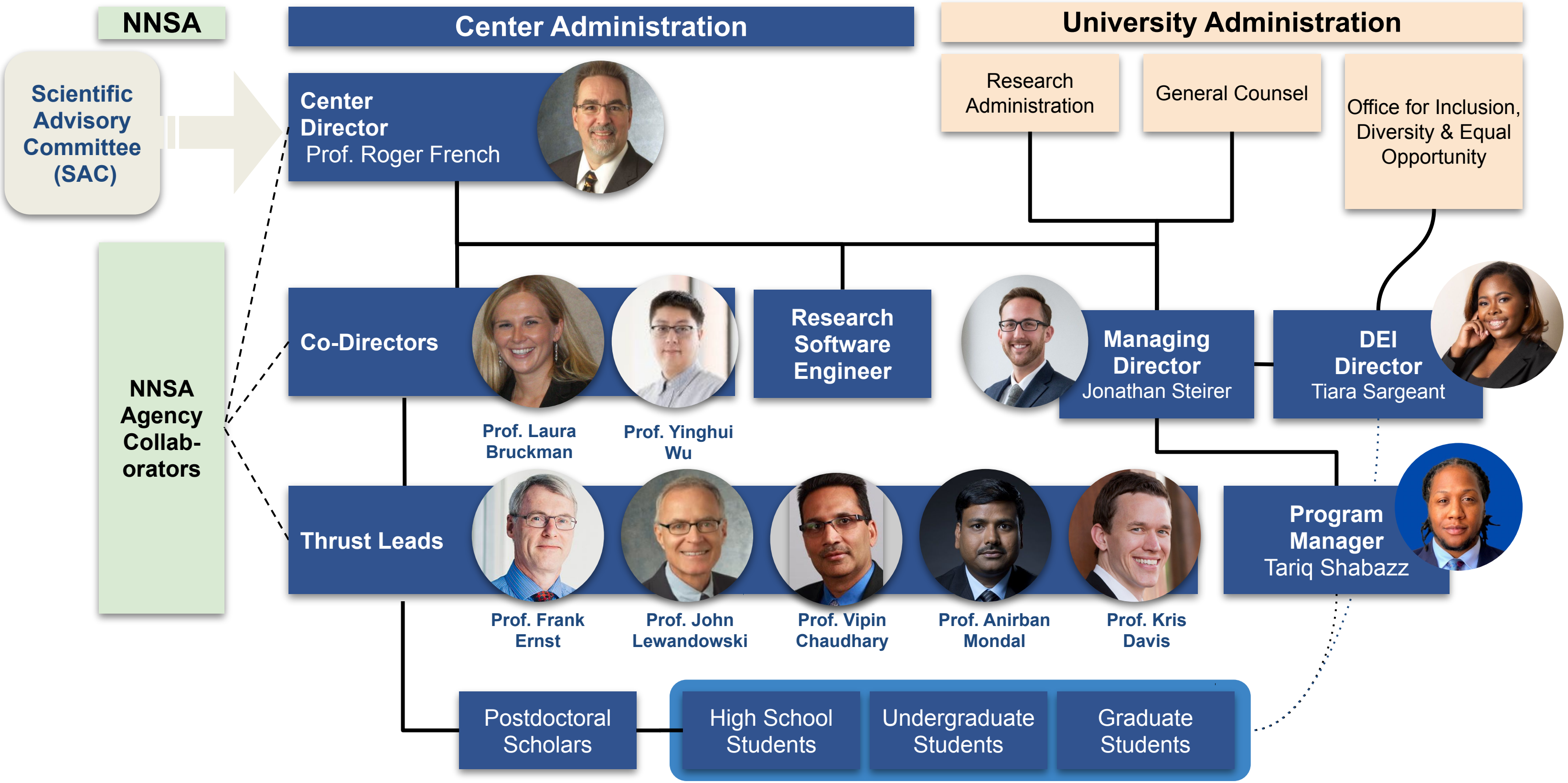
- **Using novel computer science and data science**

- **Empowering current NNSA/NSE employees**

- **Delivering a pipeline of diverse, data enabled workforce (DEW) for the future**

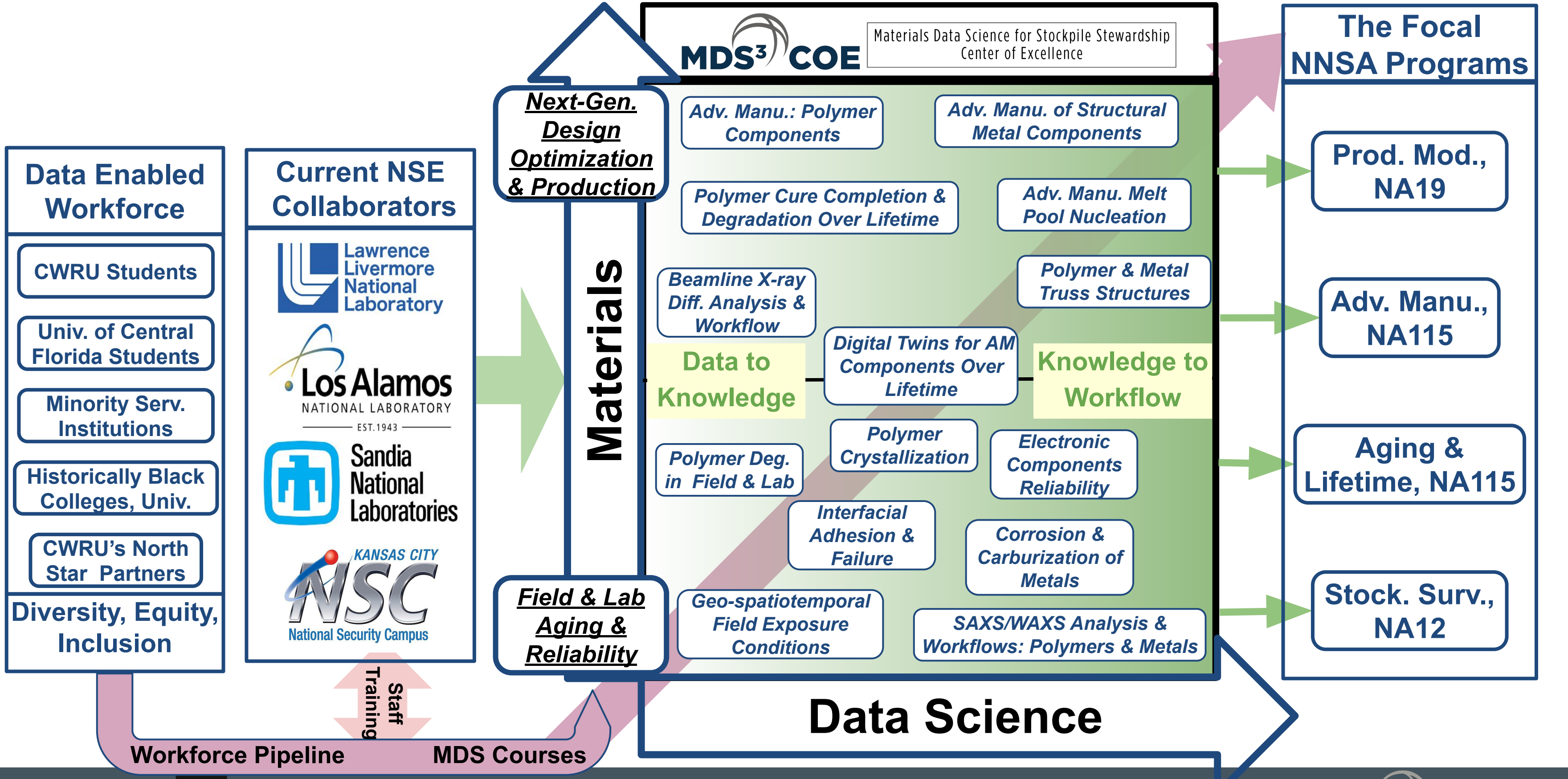


MDS^3 COE Structure



MDS^3 Center of Excellence organizational structure.

MDS^3 COE's focus: Initial NNSA Programs & Collaborations



The Components of our (MDS^3) Data-enabled Workforce Pipeline

Research Experience for Undergraduates (REUs)

- MDS^3 – 4 students
- MDS-Rely – 2 students
- CASFER – 1 student

ACS Project Seed

High Schools:
- Cleveland Metropolitan School District
- St. Martin de Porres
- Heights High

CWRU - EMSE

EMSE SDLE Research Center

- MDS^3
- MDS-Rely
- CASFER
- PV Reliability
- Geospatial

Teaching - DSCI

- Applied Data Science
- Materials Data Science

Courses:
- Undergraduate
- Graduate

CWRU NorthStar Partners

UCF (MDS^3)
- Undergrad
- Graduate

UTRGV
- Undergrad
- Graduate

Tuskegee
- Undergrad
- Graduate

PITT (MDS-Rely)
- Undergrad
- Graduate

Employment at DOE Labs

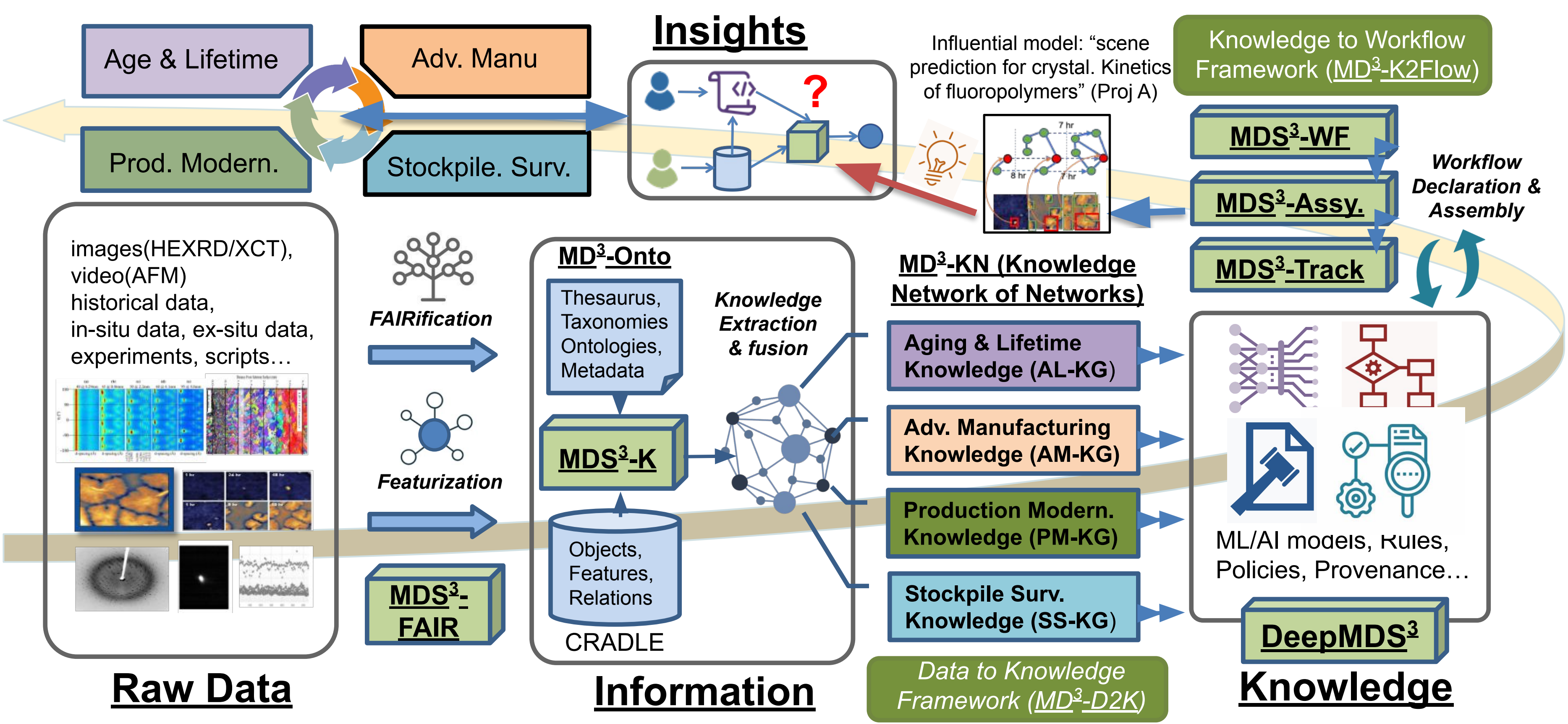
DOE Lab Internships
Summer & Year Round

How to create a new pipeline

For Data-enabled Workforce

To supply the needs of
- DOE Labs Text
- And US Industry

MDS³ Data to Knowledge, Knowledge to Workflow Framework



CASFER : A NSF ERC (Engineering Research Center)

CASFER

Center for Advancing Sustainable and Distributed Fertilizer Production

TEXAS TECH UNIVERSITY

CASE WESTERN RESERVE UNIVERSITY EST. 1826

Georgia Tech

FAMU [FLORIDA AGRICULTURAL AND MECHANICAL UNIVERSITY]

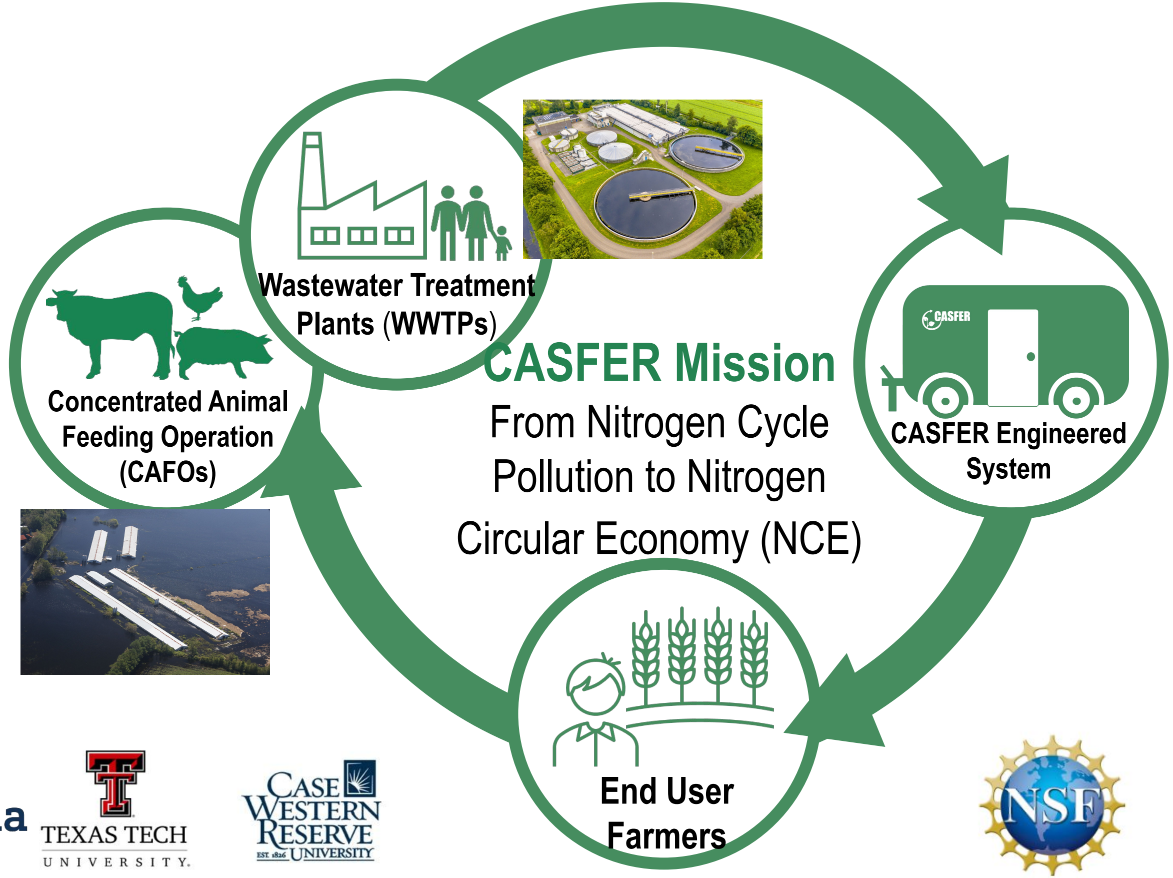
Towards a nitrogen Circular Economy

CASFER Center for Advancing Sustainable and Distributed Fertilizer Production

CASFER will enable

- **Resilient & sustainable food production**
- By developing
  - Next generation,
  - Modular,
  - Distributed, &
  - Efficient technology

To capture, recycle & produce Nitrogen Based Fertilizers (NBFs)



FAMU FLORIDA AGRICULTURAL AND MECHANICAL UNIVERSITY | Georgia Tech | TEXAS TECH UNIVERSITY | CASE WESTERN RESERVE UNIVERSITY EST 1826 | NSF

MATERIALS DATA SCIENCE
MDS-RELY

The Center on Materials Data Science for Reliability and Degradation

NSF Award# 2052776 / 2052662

Director, Laura S. Bruckman
Pitt Site Directory, Paul Leu

Case Western Reserve University EST. 1826

University of Pittsburgh

MDS-Rely NSF Ind./Univ. Collab. Research Center (IUCRC)

Lawrence Livermore National Laboratory

NAVAL NUCLEAR LABORATORY

EATON

Parker

NIST National Institute of Standards and Technology

electroninks

NETL NATIONAL ENERGY TECHNOLOGY LABORATORY

U.S.ARMY

DEVCOM ARMAMENTS CENTER

MDS-Rely 2023/24 Research Portfolio

Polymers, Elastomers & Coatings

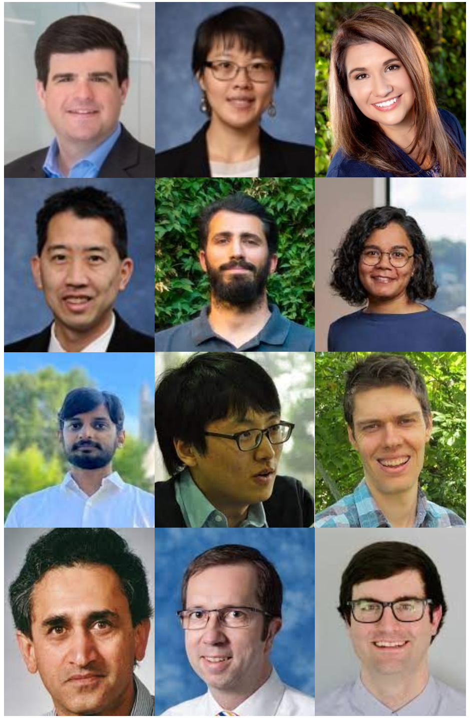
- Non-Invasive Detection of Defects during Coatings Manufacturing, Chris Wirth
- Predictive Framework to Indicate the Age of Plastics for Proper Recycling, Metin Karayilan, Divita Mathur, Sanmukh Kuppannagari
- Machine Learning Methods for Optimizing and Innovating Structural Color Paints and Coatings, Paul Leu, Oliver Hinder, Jungtaek Kim

Metals & Alloys

- Achieving Reliable Laser Powder Bed Fusion based Additive Manufacturing via Machine Learning of in-situ Optical Profilometry Monitoring Data, Xiayun Zhao
- Data-driven Analysis of Hydrogen-Degraded, Additive Manufactured Zircaloy, **Markus Chmielus & Zachary Harris**

Components, Devices & Systems

- Effects of Aerosol Jet Printing Parameters on the Lifetime Performance of Additively-Manufactured Flexible Circuits, Janet Gbur
- Enhancing Degradation Analysis and Failure Prediction through Modern Machine Learning Techniques, Satish Iyengar



A BHEF CASE STUDY

CREATING A MINOR IN

APPLIED DATA SCIENCE

Case Western Reserve University Engages Business Leaders to Produce T-Shaped Professionals

THROUGH THE COLLABORATION of its business and higher education members, the Business-Higher Education Forum (BHEF) launched the National Higher Education and Workforce Initiative (HEWI) to create new undergraduate pathways in high-skill, high-demand fields such as data science and analytics. Data science and analytics must be integrated with T-shaped skills, such as critical thinking, collaboration, and effective communication, which are critical for all graduates entering the 21st century workforce. Knowledge of data science and analytics in recent years has become as fundamental as any other skill for graduates' career readiness. BHEF's Strategic Business Engagement Model with higher education addresses this demand by moving the two sectors from transactional relationships to strategic partnerships through five strategies:

1. ENGAGE corporate leadership;
2. FOCUS corporate philanthropy on undergraduate education;
3. IDENTIFY and tap core competencies and expertise;

undergraduate education.

This case study examines how BHEF member Case Western Reserve University (Case Western Reserve) is integrating T-shaped skills into a minor in applied data science.

PROGRAM OVERVIEW

THE APPLIED DATA SCIENCE (ADS) MINOR AT CASE WESTERN RESERVE serves as a national model for undergraduate education in data science. Available to every undergraduate student across all schools at the university, this program of study requires experiential learning opportunities, embeds T-shaped skills, and allows students to master fundamental ADS concepts in their chosen domain area. From strong leadership engagement to funded undergraduate research opportunities, Case Western Reserve applied BHEF's Strategic Business Engagement Model to create a minor that responds to the fundamental need for data science in today's global business community.

AY 2014-15	AY 2015-16	AY 2016-17	AY 2017-18	AY 2018-19	AY 2019-20	AY 2020-21	AY 2021-22	AY 2022-23	Total
9	36	49	57	100	106	92	159	220	828

Medical Mutual of Ohio
Medtronic
Philips Healthcare
Sherwin-Williams Company
Siemens
Teradata Corporation
Timken Company
University Hospitals

18

http://www.bhef.com/publications/creating-minor-applied-data-science

BHEF BUSINESS HIGHER EDUCATION FORUM
Creating Solutions. Inspiring Action.

Open Source, Open Data, Reproducible Research Tools For Science

Using Open Source tools

- R & Python coding
- Git code versioning & collaboration
- Cross-Platform (Linux, Mac, Windows)
- LaTeX & Markdown

Reproducible Research

- Distribute Code & Datasets
- At time of paper publication
- Your research can be reproduced by others
- Others can build on your research and data

Use [Agile Development Tools]()

- Slack team messaging
- [Jira Cloud Issue Tracking]()
- BitBucket/GitHub/GitLab

Build Packages for Science

Use Package-based systems

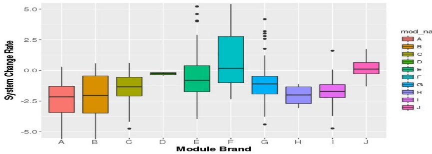
- Rely on well-vetted Open Source Codes

R Packages

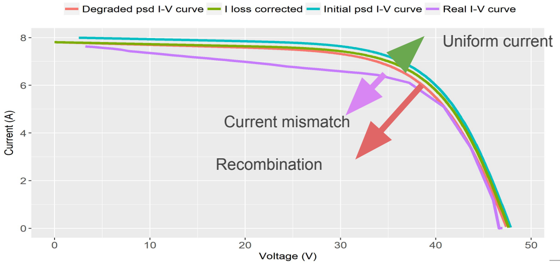
- Well vetted, with know package dependencies
- With Vignettes on Theory, & Use
- With Data Sets and Results for Validation

Performance Loss Rate Determination

IEA PVPS Task 13 PV Reliability

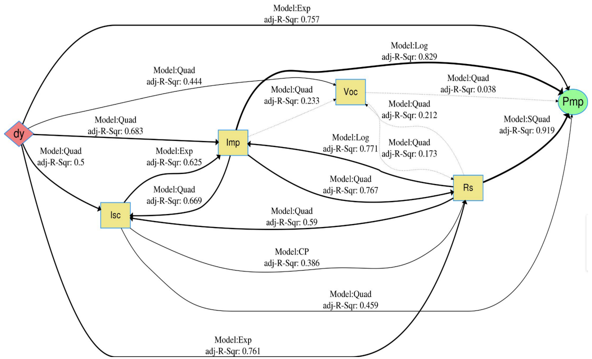


Suns-V_{oc} from Time-Series I-V



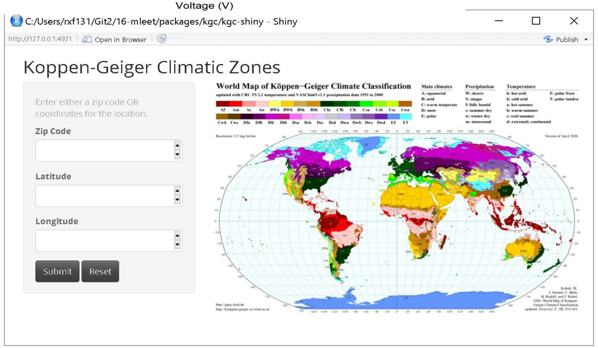
[NetSEM on CRAN]()

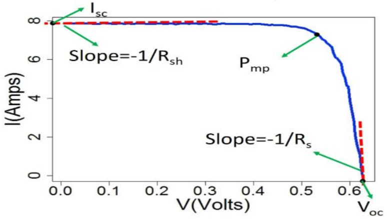
Network Structural Equation Modeling



[Kgc on CRAN]()

Köppen-Geiger Climate Zone Package





[ddiv on CRAN]()

Data-driven I-V Feature Extraction

CRADLE Analytics: Enable Sparse to Massive Data Analytics

<u>Materials Data Science</u>

Distributed/High Performance Computing
Coherent Data Lifecycle Environment

- Data & Modeling Stay Integrated
- Over years, Building on prior work

Low Barriers for novice Data Scientists

Automated Data Analysis Pipelines

<u>Enable Terabyte Dataset Analysis</u>

- Adv. Manu. Datastreams
- Beamline HEXRD
- Other Big (or Sparse) Datasets

Write-back All Models & Results

Future Analysis Builds On Priors
Datasets & AI/ML Models Get Smarter

Minimize Large Data Transfer

Prefer In-place Analytics (Hadoop/Spark)

Focus on Fast/Efficient Modeling

Such as high speed segmentation
For Autonomous Driving

The Challenges, & Opportunities, of AI/ML: Accelerating Time to Science

To develop AI/ML for Science, such as Materials Science

We have High Performance Computing (HPC)

- "Scaled Up" Computing: Works for Physics Simulation Modeling
  - But doesn't handle massive datasets

Yet Big Tech uses Distributed Computing (DC)

- "Scaled Out" Computing: e.g. used by Google, Meta, etc.

AI/ML for Science needs D/HPC Computing

- Needs the integration of "Scaled Out & Scaled Up" Computing
- CRADLEtm: Common Research Analytics & Data Lifecycle Environment[1]
  - Automated pipelines, FAIRification[2], Efficient Insights

Data Centric AI[3] presents humans with a grand opportunity

- "*Computational Inflection Point for Scientific Discovery*"[4]
  - Augmenting human reasoning; Working alongside human researchers
  - Scientific investigations restructured around the "salient human tasks"
    - With computers handling the routine and onerous tasks
    - Supplementing our human capabilities

While decreasing reductionist approaches in scientific research

In SDLE Res. Cntr.

- Dist. Compute
  - 2.5 Pb Cluster
  - 7 TB Ram
  - 1164 CPU Cores
  - 30 GPUs
    - 480 GPU VRAM
    - 384k Cuda Cores
    - 1.2k Tensor Cores
- High Perf. Compute
  - 7152 CPU Cores
- Nvidia AISC 8 DGX
  - 2.5 Tb VRAM
  - 4 Tb RAM
  - 15 Tb nvme storage

1. A. Khalilnejad, ry s;., "Automated Pipeline Framework for Processing of Large-Scale ...," PLOS ONE, 15, 12, p. e0240461, Dec. 2020.
2. W. C. Oltjen et al., "FAIRification, Quality Assessment, and Missingness Pattern ..., IEEE PVSC, Jun. 2022, pp. 0796–0801.
3. M. H. Jarrahi, et al., "The Principles of Data-Centric AI," Commun. ACM, vol. 66, no. 8, pp. 84–92, Jul. 2023,
4. T. Hope, et al., "A Computational Inflection for Scientific Discovery," Commun. ACM, vol. 66, no. 8, pp. 62–73, Jul. 2023,

AI4Science: An inflection point for Science

DOE NNSA & DOE Office of Science

- Are individually funded by congress
- And working towards $2B for AI4Science

Both have noticed our MDS^3 COE

- As a demo of what the opportunity is

DOI:10.1145/3571724

Uniting data-centric perspectives and concepts to trace the foundations of DCAI.

BY MOHAMMAD HOSSEIN JARRAHI, ALI MEMARIANI, AND SHION GUHA

The Principles of Data-Centric AI

» key insights

- DCAI is an emerging paradigm that emphasizes the importance of data quality and dynamism in AI systems, using an iterative, systematic approach.
- DCAI is redefining the role of data from being merely a preprocessing concern to a continuous improvement factor, encouraging consistent enhancement of both data and model throughout the AI life cycle by incorporating strategies such as data augmentation.
- A specific contribution of this article is its focus on the human-centered nature of data that feeds AI systems, presenting data as a sociotechnical system, embodying both technological elements and social norms, and biases.

IMAGE BY OLLOMY

DOI:10.1145/3576896

Enabling researchers to leverage systems to overcome the limits of human cognitive capacity.

BY TOM HOPE, DOUG DOWNEY, OREN ETZIONI, DANIEL S. WELD, AND ERIC HORVITZ

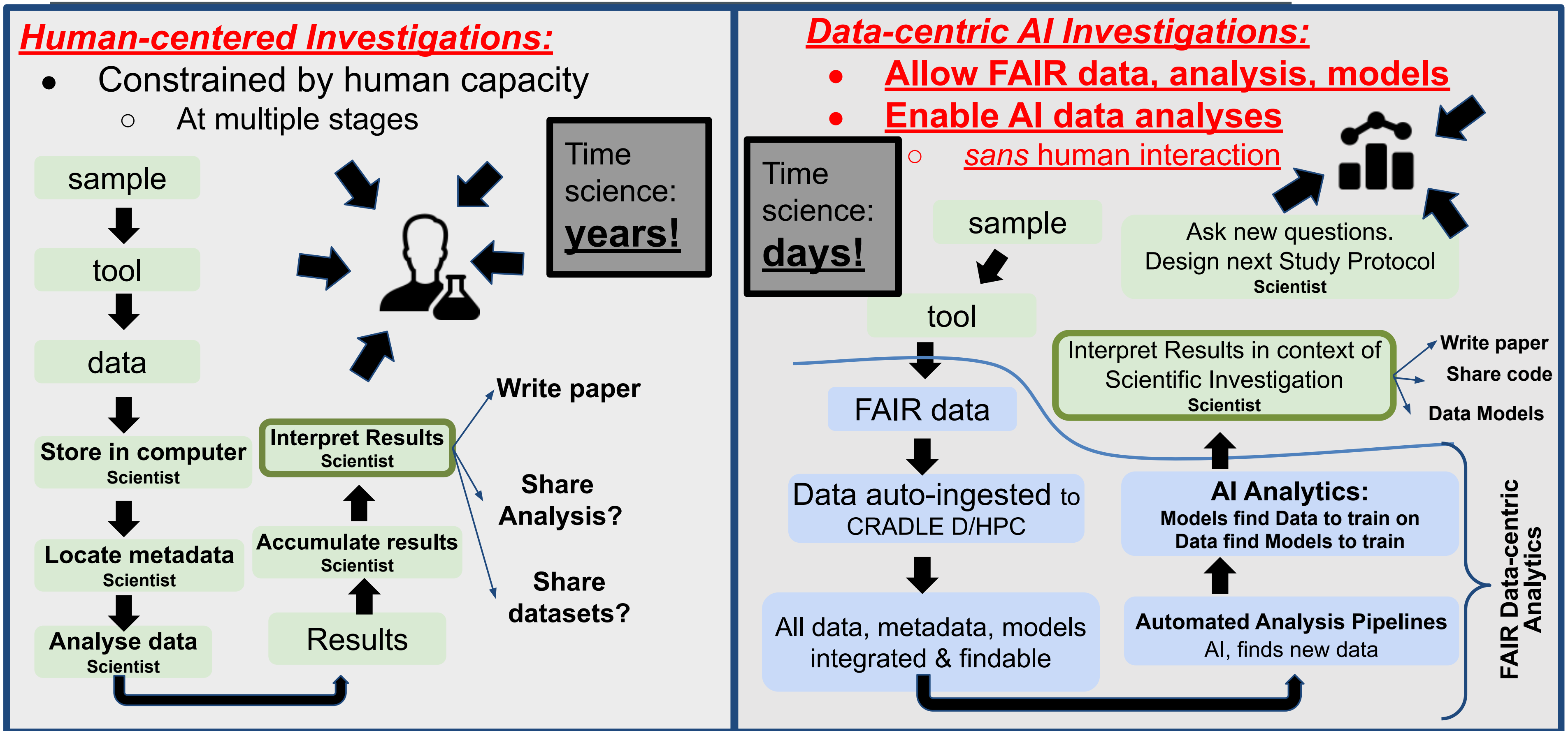
A Computational Inflection for Scientific Discovery

» key insights

- Recent advances in AI present great opportunities for augmenting human scientific reasoning. Future systems may work alongside humans throughout the scientific process: detecting and explaining relevant literature, generating hypotheses and directions, suggesting experiments and actions.
- The scientific process may be decomposed into salient human tasks. Task-guided scientific knowledge retrieval systems retrieve and synthesize external knowledge in a manner that serves a task-guided utility of a researcher, while taking into consideration the researcher's goals, state of inner knowledge, and preferences.
- Progress has recently been made in building such systems yet fundamental challenges remain: in computational representation and synthesis of scientific knowledge, and in modeling the diversity of human tasks, contexts, and cognitive processes involved in consuming and producing scientific knowledge.

ILLUSTRATION BY PETER CROWTHER ASSOCIATES

An inflection point for science



Outline: Common Research Analytics & Data Lifecycle Environment

CRADLE Computing & Analytics

- Hardware, Frameworks, Middleware & Automated Pipelines
- FAIRification: Making Data & Models FAIR

CRADLE Data Lifecycle

- Scientific Investigations, Study Protocols & Materials Data Science

Spatiotemporal-Graph (st-Graph) Learning

- Timeseries Imputation & Trend Estimation

Geospatial Data Science

- Eutrophication: Motion of Nitrogen Through Watersheds

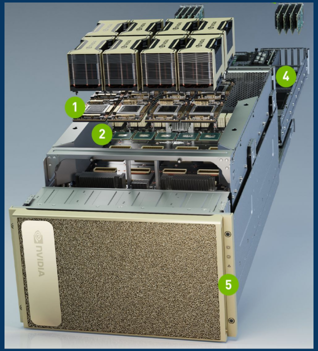
Synchrotron 2D X-ray Diffraction HEXRD: Automated NN Analysis Pipelines

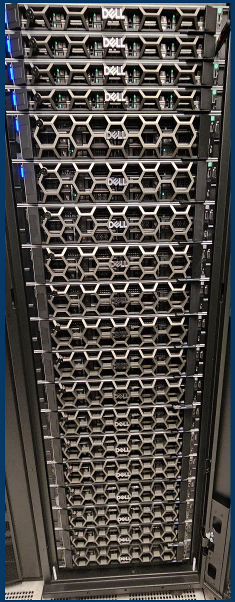
- “Scientist Ground Truth” Learning Approach
- Kinematic Diffraction Forward Model Learning

3.5D X-ray Computed Tomography: Pipelines & Spatiotemporal Feature Extraction

- Observing Pitting Corrosion of Aluminum Wires
- Al:Mg Alloy: Stress Corrosion Cracking

Conclusions





CRADLE Computing & Analytics: Hardware

GS: Arafath Nihar[1], Olatunde Akanbi[1], Tommy Ciardi[1], Tian Wang[1]
UG: Rachel Yamamoto[1], Rounak Chawla[1], Hayden Caldwell[1],
Faculty: Yinghui Wu[1], Vipin Chaudhary[1], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

Horizontal Scaling vs Vertical Scaling

Horizontal Scaling:

- Add more machines
- To increase capacity

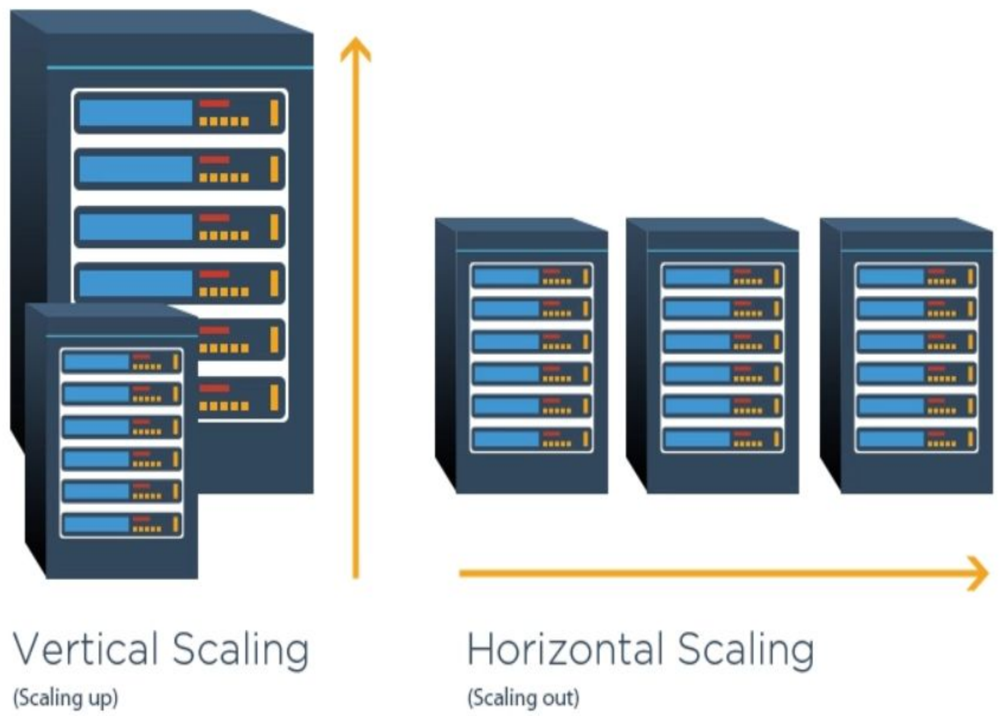
Distributes workloads

- **Across multiple machines**

Increases redundancy

- **And fault tolerance**

Generally more cost-effective



CRADLE Compute Environment: Distributed & High Performance Computing

Running in CWRU's HPC

- Pioneer (RHEL8 OS)
- Markov (DSCI Teaching Cluster)

Dist. Comp. Frameworks

- Apache Hadoop, Hbase, Spark
- Apache Ozone, Impala, Ranger, etc
- JanusGraph, GraphX

Cloudera Data Platform

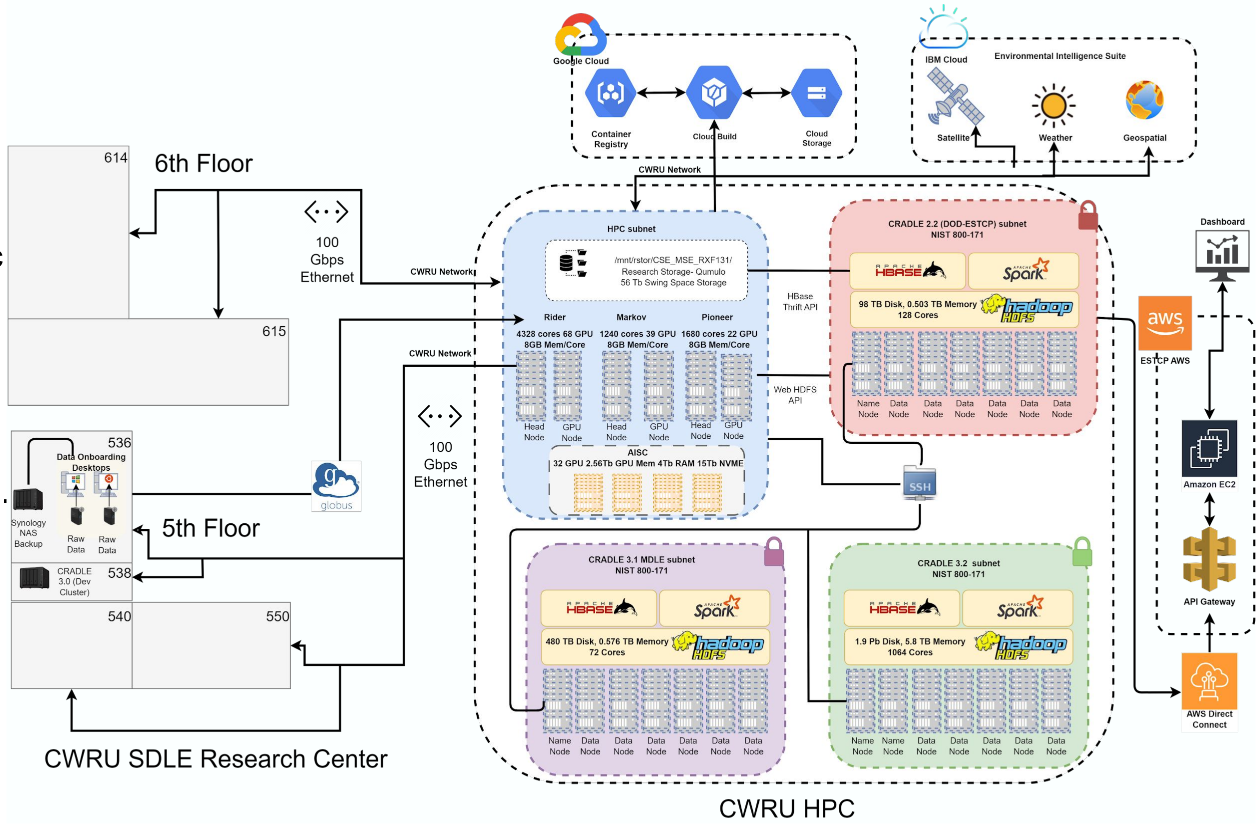
- Commercial supported distribution
- Of Apache Hadoop/Hbase/Spark/….

OnDemand Containerized Apps

- Using Ubuntu 20.04 OS

Able to train 100s of Deep Learning Models

Common Research Analytics & Data Processing Environment



CRADLE Hardware: HPC Scaling up

Pioneer HPC: 5912 cores

- 32 gpu nodes

Markov HPC: 1240 cores

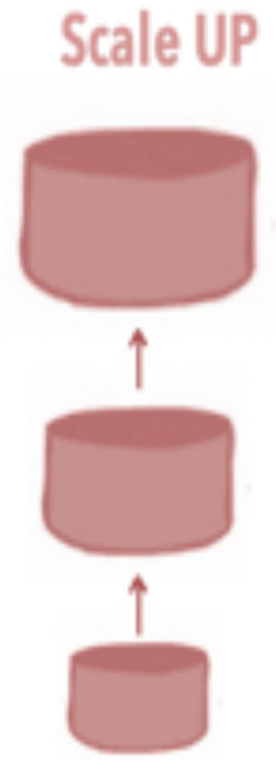
- 16 gpu nodes

One Compute Node

- Up to 40 cores
- Up to 1Tb RAM memory
- Nvidia v100
- Up to 32 GB of GPU VRAM

HPC Compute Model

- Lots of FLOPS
- But Limited, Expensive Data Storage



CRADLE Hardware: HPC Scaling up

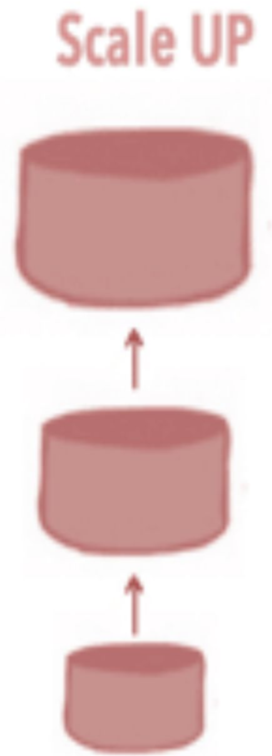
Nvidia AISC: 32 integrated GPU nodes
- 4 Nvidia DGX Pods, of 8 A100 GPUs
- 2.56 Tb GPU VRAM
- 4 Tb of RAM memory
- 15 Tb NVME storage

Pioneer HPC: 5912 cores
- 32 gpu nodes

Markov HPC: 1240 cores
- 16 gpu nodes

One Compute Node
- Up to 40 cores
- Up to 1Tb RAM memory
- Nvidia v100
- Up to 32 GB of GPU VRAM

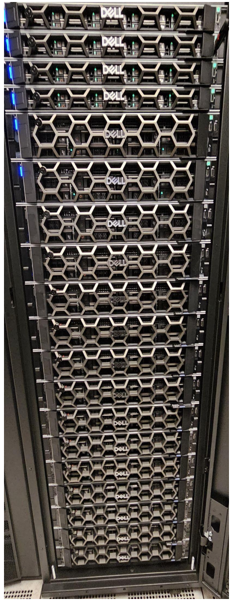
HPC Compute Model
- Lots of FLOPS
- But <u>Limited, Expensive Data Storage</u>

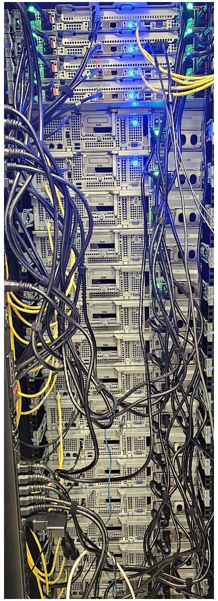


CRADLE Hardware: Distributed Hadoop Scaling Out, for CRADLE 3.2

4 Name Nodes

- 224 Cores
- 2 Tb of RAM memory
- 21.6 Tb Storage

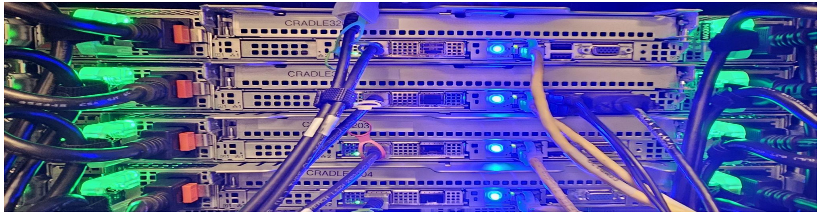




15 Data Nodes

- 840 Cores
- 3.84 Tb of RAM memory
- 1.92 Pb of Storage TB
- 30 NVIDIA Ampere A2 GPU

= 1.95 Pb of storage



CRADLE D/HPC

- Dist. Compute
  - 2.5 Pb Cluster
  - 7 TB Ram
  - 1164 CPU Cores
  - 30 GPUs
    - 480 GPU VRAM
    - 384k Cuda Cores
    - 1.2k Tensor Cores
- High Perf. Compute
  - 7152 CPU Cores
- Nvidia AISC 8-DGX
  - 2.5 Tb VRAM
  - 4 Tb RAM
  - 15 Tb nvme storage

CRADLE Hardware: Distributed Hadoop Scaling Out, for CRADLE 3.2

4 Name Nodes

- 224 Cores
- 2 Tb of RAM memory
- 21.6 Tb Storage

15 Data Nodes

- 840 Cores
- 3.84 Tb of RAM memory
- 1.92 Pb of Storage TB
- 30 NVIDIA Ampere A2 GPU

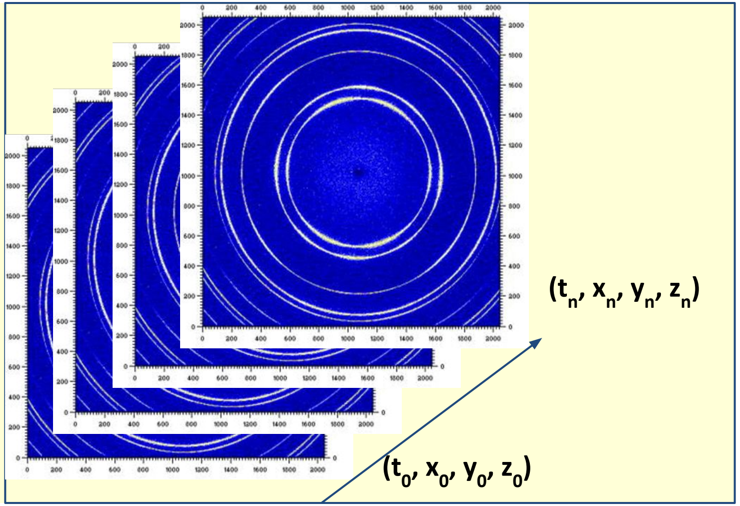
CRADLE D/HPC

- Dist. Compute
  - 2.5 Pb Cluster
  - 7 TB Ram
  - 1164 CPU Cores
  - 30 GPUs
    - 480 GPU VRAM
    - 384k Cuda Cores
    - 1.2k Tensor Cores
- High Perf. Compute
  - 7152 CPU Cores
- Nvidia AISC 8-DGX
  - 2.5 Tb VRAM
  - 4 Tb RAM
  - 15 Tb nvme storage

Current 2D-HEXRD Datasets

from Don Brown @ LANL

- ~ 21 Tb
- ~ 4.5 million HEXRD images
- In-situ heating, texture, strain analysis of Ti-6Al-4V at CHESS,
- Wire arc additive manufacturing of stainless steel etc.



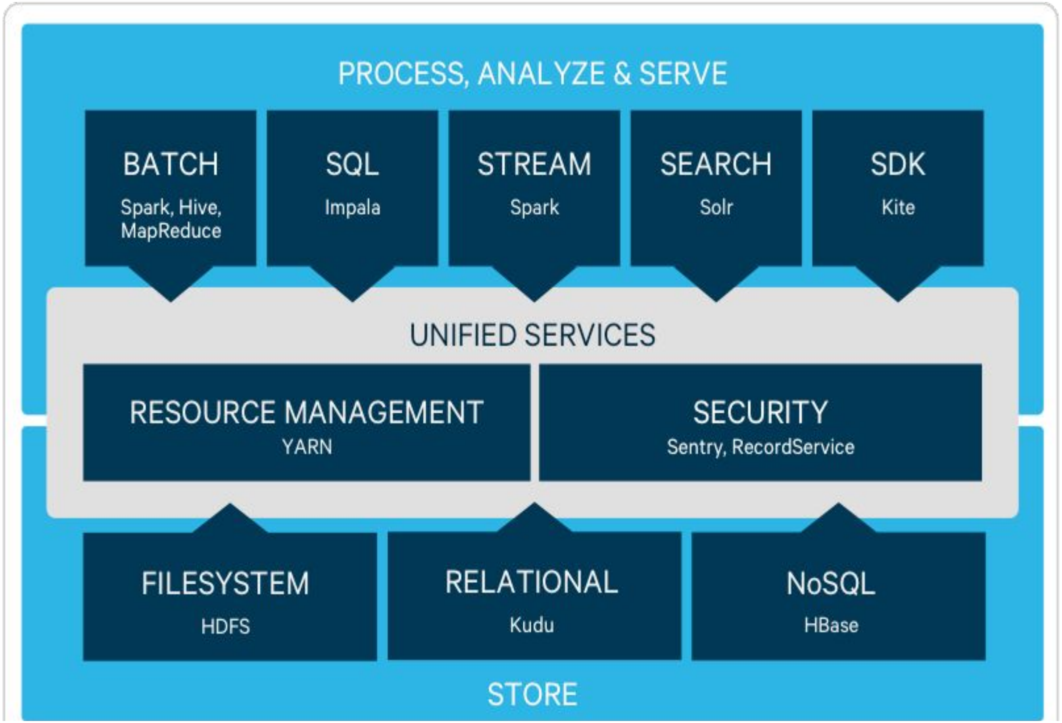
Distributed Computing: Cloudera Data Platform Distribution

Hadoop Distributed File System
- HDFS Storage

Apache Spark:
- Unified analytics engine for large-scale data processing

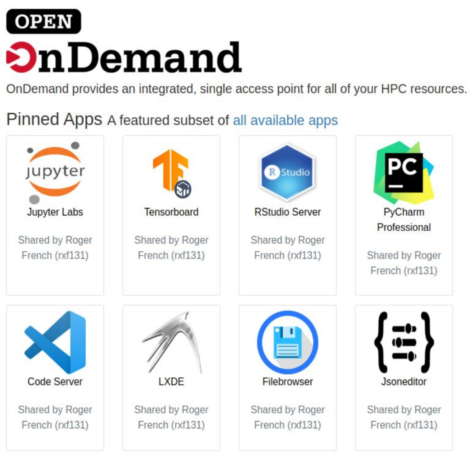
Apache Impala:
- Massively parallel processing SQL query engine

Kerberos:
- User authentication protocol

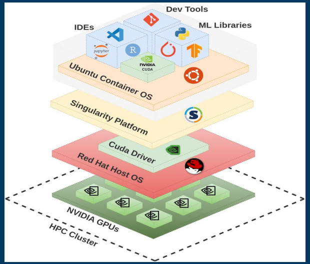


CRADLE Computing:

Frameworks, Middleware & Automated Pipelines



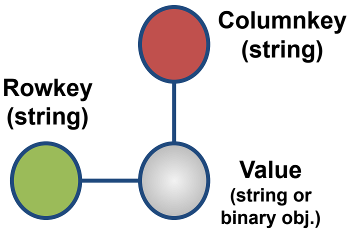
Web-based Access to Cloud & Softwares



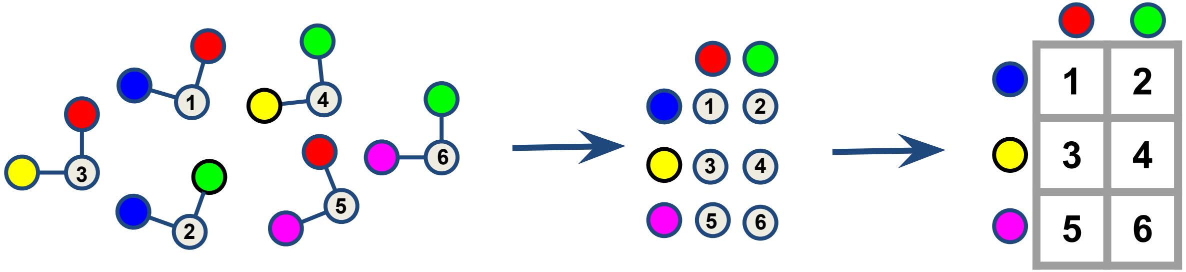
GS: Arafath Nihar[1], Olatunde Akanbi[1], Tommy Ciardi[1], Tian Wang[1]
UG: Rachel Yamamoto[1], Rounak Chawla[1], Hayden Caldwell[1],
Faculty: Yinghui Wu[1], Vipin Chaudhary[1], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

The "NoSQL" Database Abstraction of Hadoop/Hbase: RDF Triples



HBase RDF 'Triple'



Combines Lab data (Spectra, Images, Videos etc.)

With Geospatiotemporal Data (PV Power Plant Data)

Distributed & High Performance Computing:

Petabyte Data Lake In A Petaflop HPC Environment

- In-place Analytics: Distributed Spark Analytics in Hadoop/HDFS/Hbase
- In-memory Data Extraction: To Separate HPC Compute Nodes

A non-relational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites

Yang Hu, *Member, IEEE*, Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE*, Timothy J. Peshek, *Member, IEEE*, Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE*, and Roger H. French, *Member, IEEE*

Automated pipeline framework for processing of large-scale building energy time series data

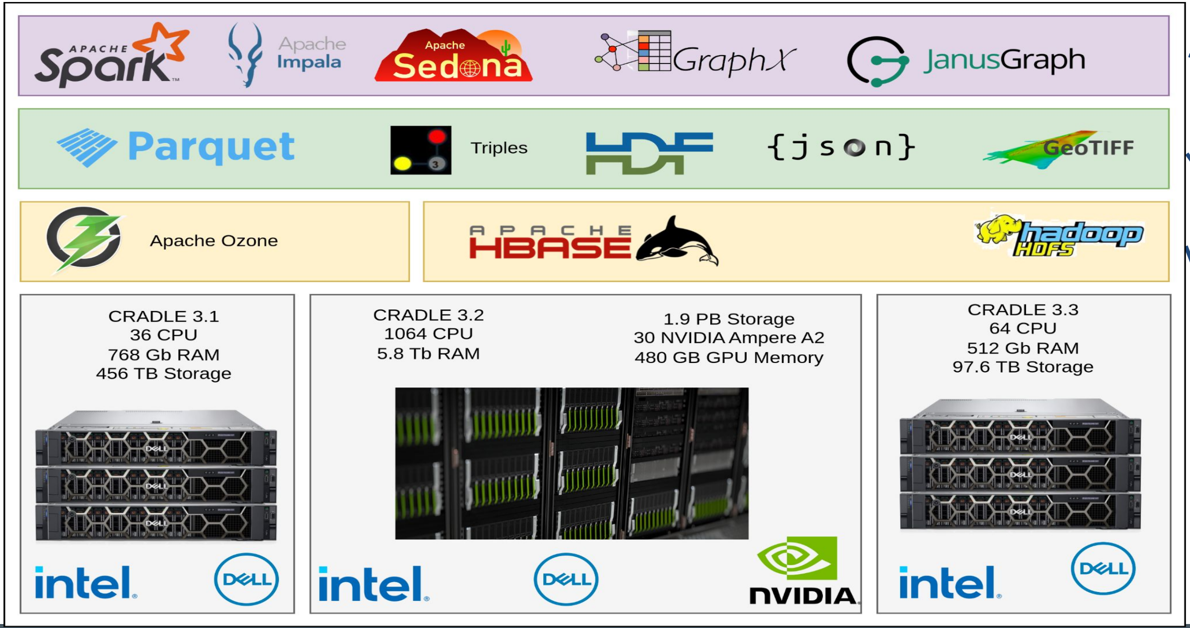
Arash Khalilnejad[1,5], Ahmad M. Karimi[2,5], Shreyas Kamath[1,5], Rojiar Haddadian[2,5], Roger H. French[2,4,5]*, Alexis R. Abramson[3,6¤]

Hu, Y., et al., "A Nonrelational Data Warehouse for the Analysis of Field & Lab Data From Multiple Heterogeneous Photovoltaic Test Sites." IEEE JPV, 7, 1, 2017, 230–36.
A. Khalilnejad, et al., Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data, PLOS ONE. 15 (2020) e0240461.

Multimodal Hadoop Cluster for Heterogeneous Data

CRADLE's Hadoop cluster prioritizes the scientific data workflow

- Leverages a unique combination of **open source technologies**
- To manage **heterogeneous data at scale** (Petabytes)
- Prioritizing **multi-modality, reproducibility, and security**



Multi-modal Processing

Lightning fast Petabyte scale data pipelines

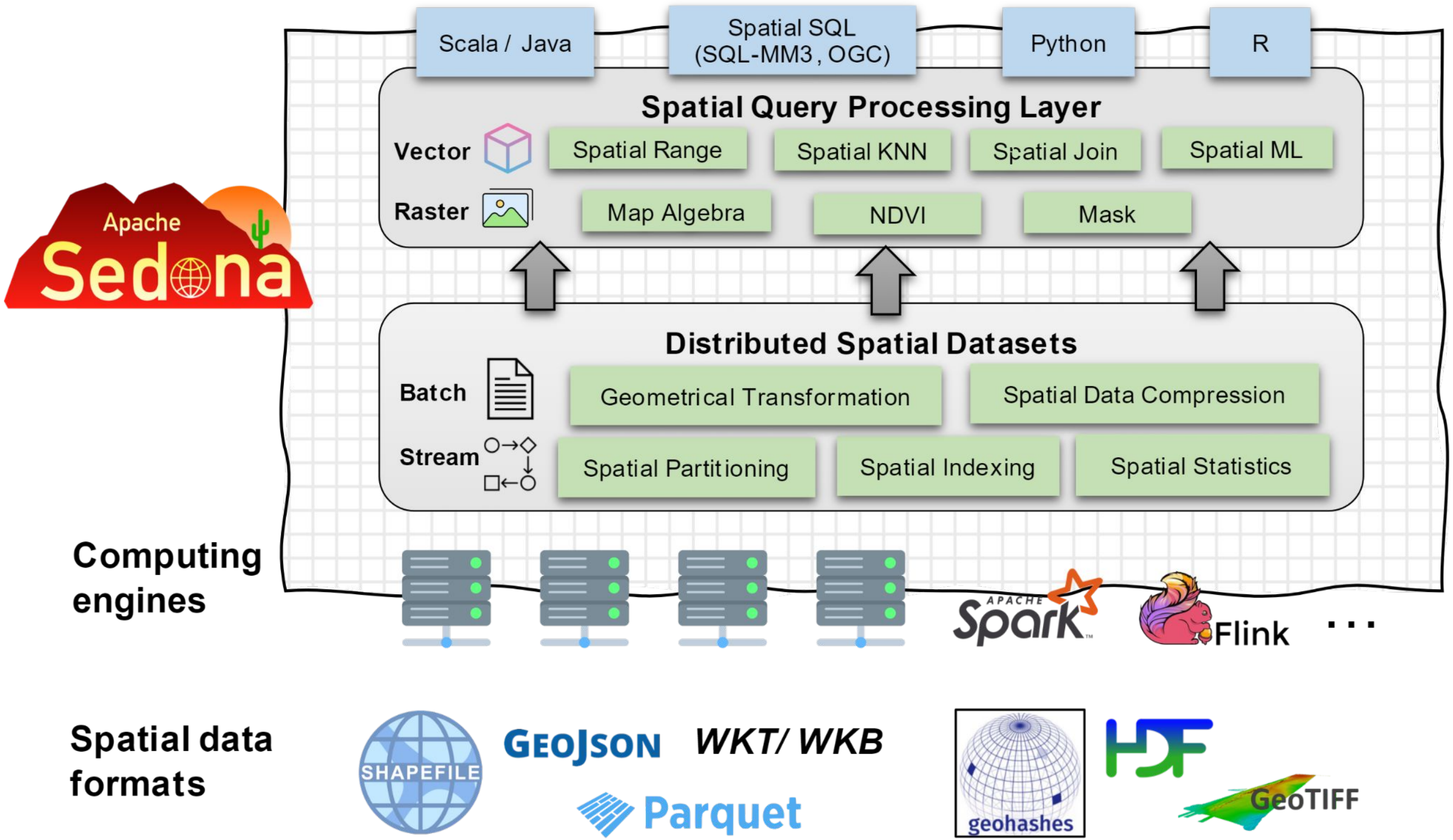
Multi-modal Storage

Low latency queries for heterogeneous data sets

HDFS Base Storage

raw data lake for provenance and reproducibility

Example: Apache Sedona for Handling Geospatial Raster Data



CRADLE Middleware

Complex computational tools made easily accessible through simple Python & R interfaces

CRADLEtools

Apache tool wrapper for simple interactive queries

desired metadata

DataFrame

row

column

Example: return all power plants with weather data 2018-2020

CRADLEsgis

Input coord and time

API call

SOLARGIS

{JSON}

clean / format

DataFrame

row

column

SDLEfleets

Submit 1000's of compute jobs through a single function call

Parquet

SDLEfleets

slurm jobs created

slurm workload manager

LOG

DataFrame

row

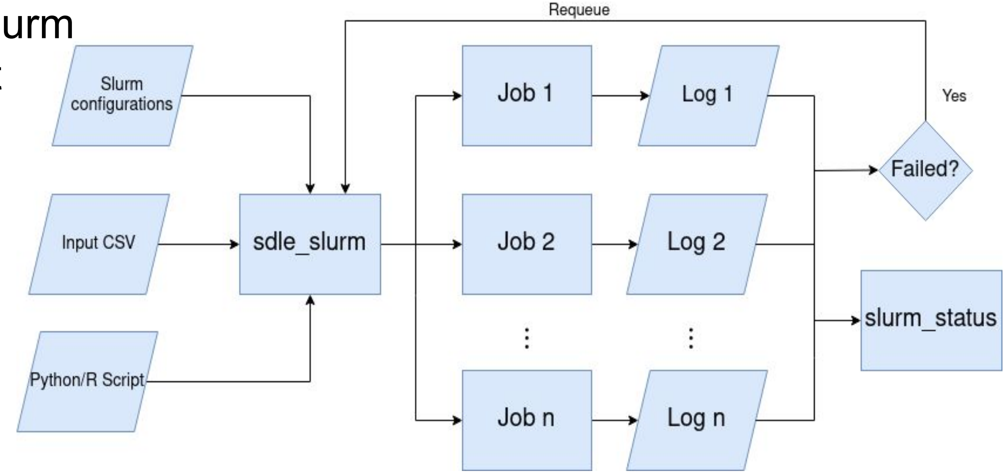
column

SDLEfleets Package: Fleets of ML Jobs

SLURM (Simple Linux Utility for Resource Management)
- allocates and releases computational resources
- when available to jobs in its queue

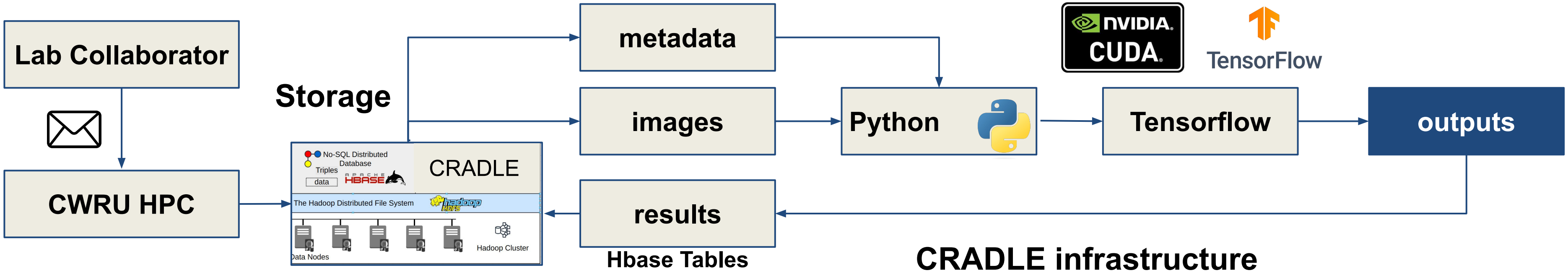
Drawbacks:
- User unfriendly for data scientists (requires proficiency with shell scripting)
- Difficult to scale
- No aggregate job status checking/error reporting

SDLEfleets Package
- A scalable Python and R interface over Slurm
  - for job fleet submission & management
- Key features:
  - Integrated with other HPC tools
    - (pyCRADLEtools3/rCRADLEtools3)
  - Simple workflow
  - Containerized
  - Improved and aggregated logging (json)
  - Job requeue



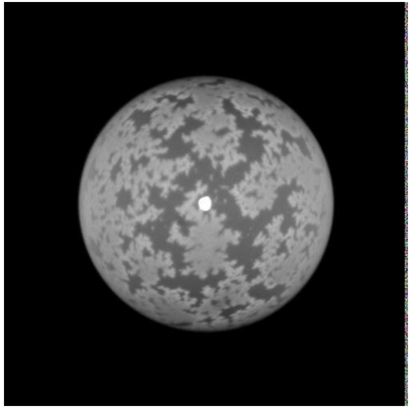
[1] https://slurm.schedmd.com/

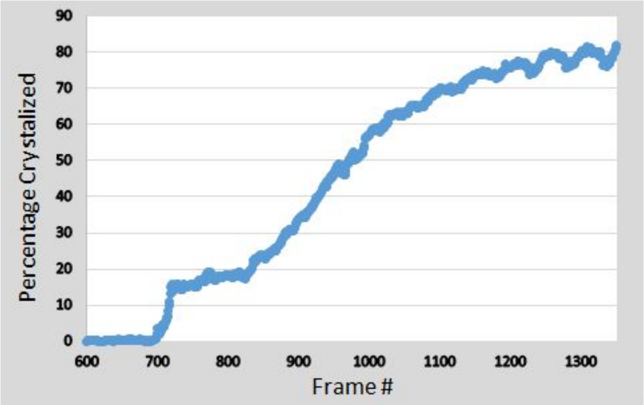
Data Processing Infrastructure: A Data Analysis Pipeline (Python or R)



Nucleation & Growth of AlN Crystals

- 1 million images of Al/Ni Melt





CRADLE infrastructure

NoSQL database

- Apache HBase

Object storage

- Apache Ozone

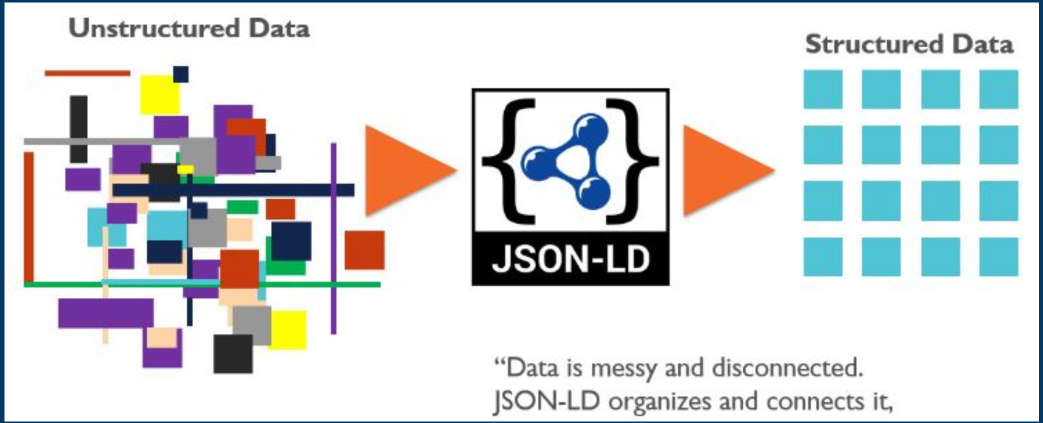
HPC environment

- Nvidia GPU acceleration for deep learning

Python/TensorFlow

[1] M. Adachi, S. Hamaya, D. Morikawa, B. G. Pierce, A. M. Karimi, Y. Yamagata, K. Tsuda, R. H. French, H Fukuyama, Temperature dependence of crystal growth behavior of AlN on Ni–Al using electromagnetic levitation and computer vision technique", Mat. Sci. in Semicon. Proc., 153, 2023, 107167, ISSN 1369-8001, https://doi.org/10.1016/j.mssp.2022.107167 .
[2] A. Khalilnejad, et al., "Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data," PLOS ONE, p. e0240461, Dec. 2020, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240461 .
[3] Masayoshi Adachi et al., "In-situ observation of AlN formation from Ni-Al solution using an electromagnetic levitation technique," J Am Ceram Soc, p. jace.16960, Jan. 2020, https://onlinelibrary.wiley.com/doi/abs/10.1111/jace.16960 .

FAIRification: Making {Meta}Data & Models FAIR



GS: Alexander Harding Bradley[1], Priyan Rajamohan[1]
UG: Jiana Kambo[1], Hyangmok Baek[1]
Postdoc: Erika I. Barcelos[2]
Faculty: Yinghui Wu[1], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

Traditional Scientific Investigations versus FAIR Reproducible Science

Today's Research…

Research Question	Each user defines their own "naming" convention
Experiments/ Simulations	Metadata stored in log books
Data stored locally	Data not findable by computers
Data shared by email	No metadata description

Data gets lost or become non-reusable!

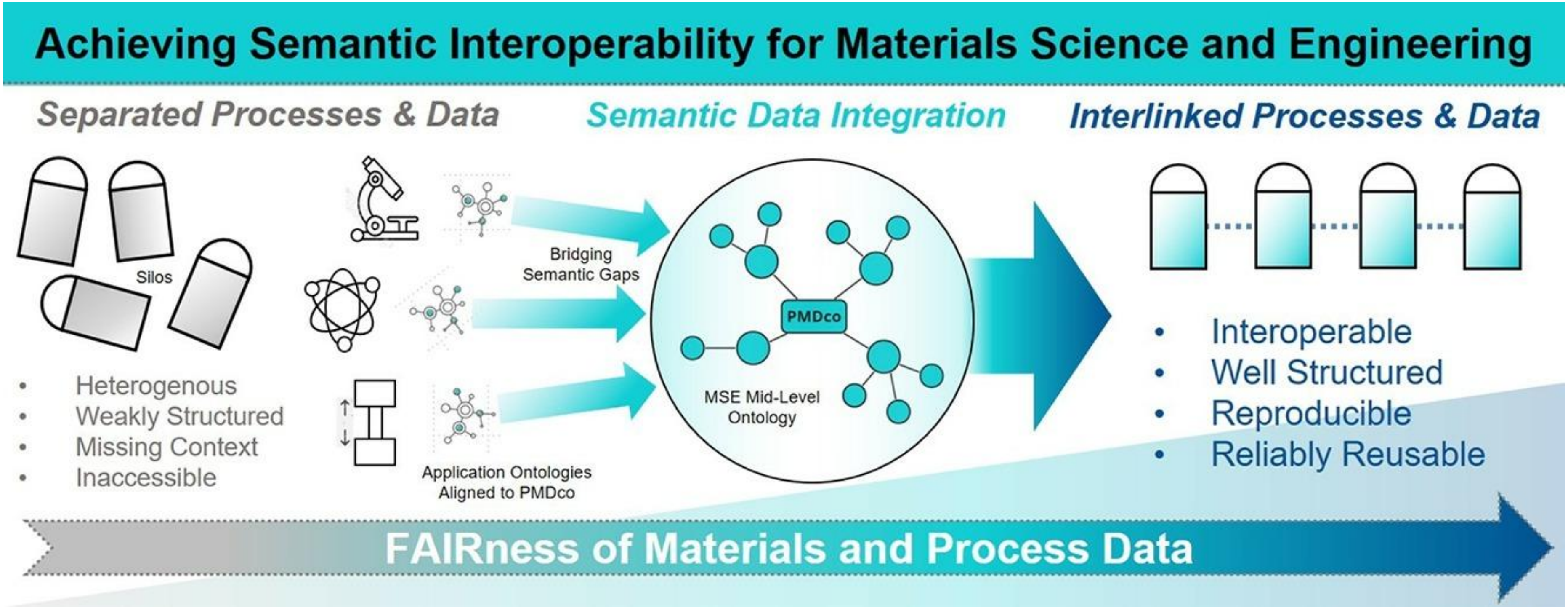
Findable
- Should be findable by humans and computers
- Detailed descriptive metadata
- (Meta)data assigned to a globally unique and persistent identifier

Accessible
- (Meta)data accessible even when data no longer available
- (Meta)data retrievable by their identifier using standardized protocol

Interoperable
- (Meta)data use formal, accessible, shared, knowledge representation
- (Meta)data follows FAIR domain ontology & references other metadata

Reusable
- (Meta)data are released with a clear & accessible data license
- (Meta)data meet domain-relevant community standards

PDMco Mid-level Ontology



B. Bayerlein et al., "PMD Core Ontology: Achieving semantic interoperability in materials science," Materials & Design, vol. 237, p. 112603, Jan. 2024, doi: [10.1016/j.matdes.2023.112603](https://doi.org/10.1016/j.matdes.2023.112603)

“Bi-lingual” R & Python Package: With Common JSON-LD Domain Templates

FAIRmaterials Package website

- https://cwrusdle.bitbucket.io/

~ 30 Scientific Domain Ontologies

- Defined by OWL Files
- And 1 Combined OWL file

48 json-ld templates

- For these domains

~ 30 domain documentation vignettes

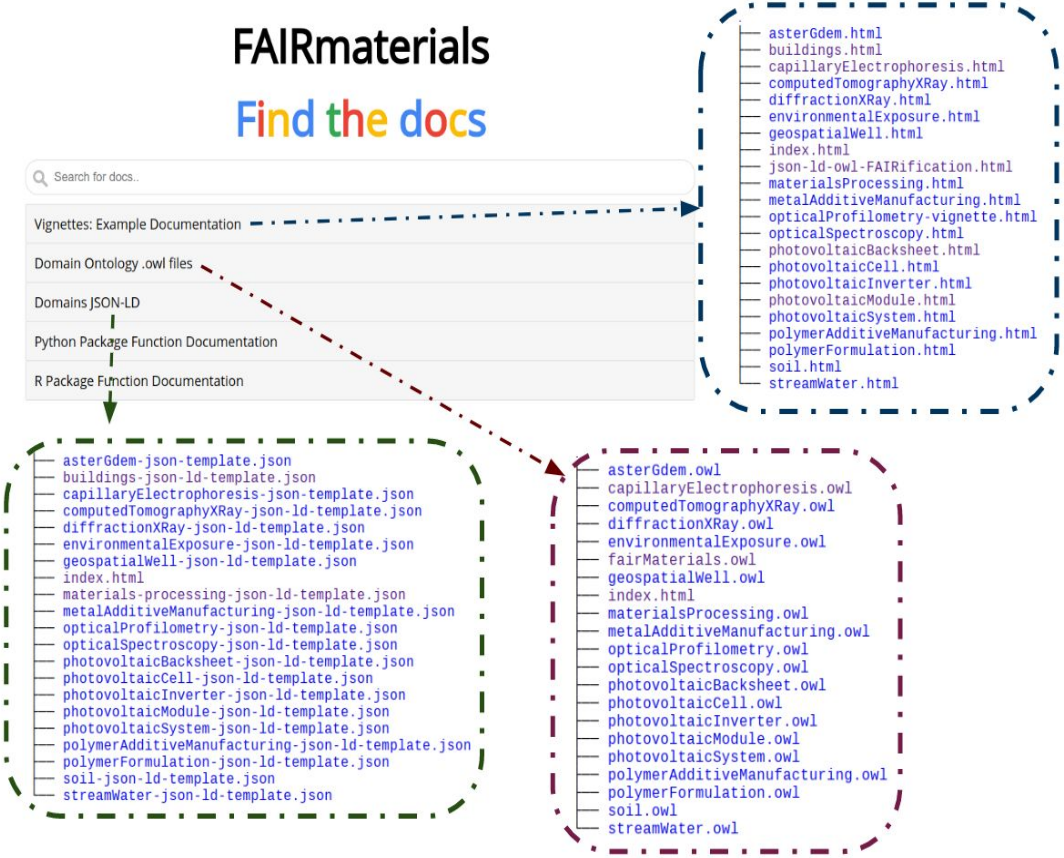
- How to FAIRify for that domain

Towards automation of

JSON-LD & Ontology

Creation and validation

- No existing tools for this purpose
  - Manual work
- Now automating with RDFLib & PyLODE



CRADLE FAIRification & Data Science Workflow Pipeline

Datasets shared on osf.io with the community

OSFHOME ▾ Search Support Donate Sign Up Sign In

CWRU SDLE Research Center | Metadata | Files | Wiki | Analytics | Registrations

CWRU SDLE Research Center

0.0B | Public | 0

Contributors: Roger H. French, Arafath Nihar, Erika Barcelos, Raymond Wieser, Alan Curran, Ahmad Maroof Karimi, Jennifer L Braid, Devin Gordon, JiQi Liu, Menghong Wang, Laura Bruckman, Arash Khalilnejad, Kunal Rath, Sascha Lindig, Xuanji Yu, Sameera Nalin Venkat, Will Oltjen, Olatunde Akanbi

Affiliated institutions: Case Western Reserve University

Date created: 2019-03-25 10:34 AM | Last Updated: 2023-06-04 06:41 PM

Identifier: DOI 10.17605/OSF.IO/WN35J

Category: Project

Description: *Open Data and Open Codes shared from CWRU SDLE Research Center*

License: *CC-By Attribution 4.0 International*

Lawrence Livermore National Laboratory

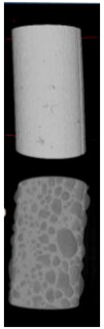
Los Alamos National Laboratory EST. 1943

Sandia National Laboratories

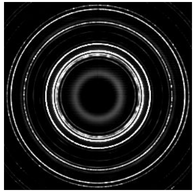
FAMU Florida Agricultural and Mechanical University

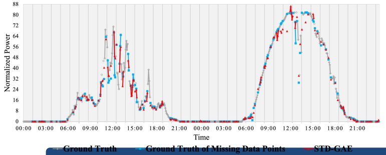
solaredge

Brookfield Renewable

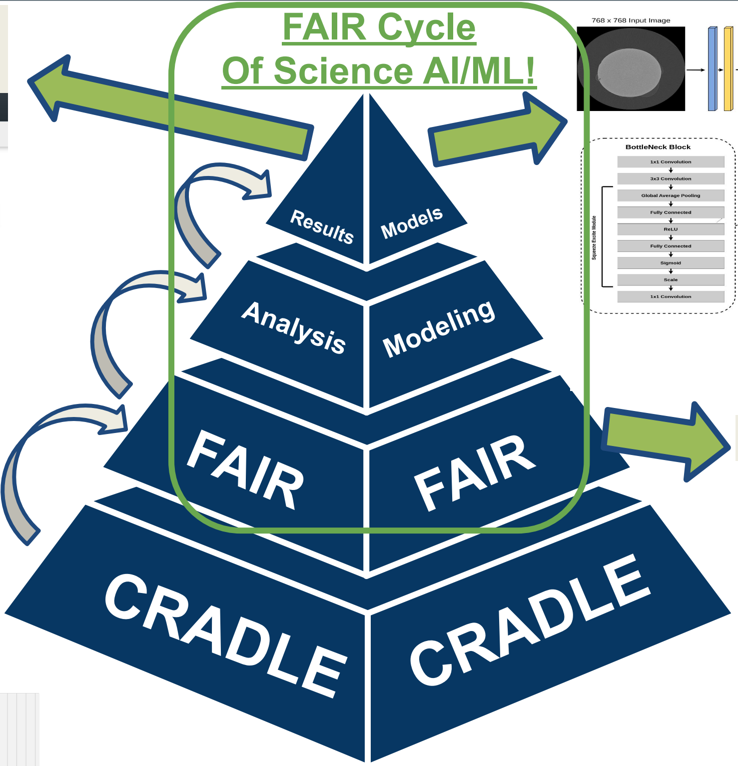


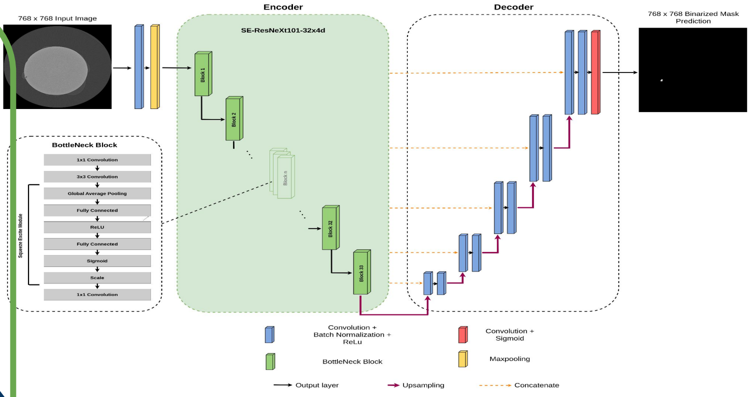
OrderDate	Region	Rep	Item	Units	Unit Cost	Total
9/1/14	Central	Smith	Desk	2	125.00	
6/17/15	Central	Kivell	Desk	5	125.00	
9/10/15	Central	Gill	Pencil	7	1.29	
11/17/15	Central	Jardine	Binder	11	4.99	
10/31/15	Central	Andrews	Pencil	14	1.29	
2/26/14	Central	Gill	Pen	27	19.99	
10/5/14	Central	Morgan	Binder	28	8.99	
12/21/15	Central	Andrews	Binder	28	4.99	
2/9/14	Central	Jardine	Pencil	36	4.99	
8/7/15	Central	Kivell	Pen Set	42	23.95	
1/15/15	Central	Gill	Binder	46	8.99	
1/23/14	Central	Kivell	Binder	50	19.99	
3/24/15	Central	Jardine	Pen Set	50	4.99	
5/14/15	Central	Gill	Pencil	53	1.29	
7/21/15	Central	Morgan	Pen Set	55	12.49	





DATA





Open source tools to community

Search projects

fairmaterials 0.4.0

```
pip install fairmaterials
```

PVplr: Performance Loss Rate Analysis Pipeline

The pipeline contained in this package provides tools used in the Solar I feature correction, power predictive modeling, PLR determination, and usage. This material is based upon work supported by the U.S Departme of the High Performance Computing Resource in the Core Facility for Ad

FAIRmaterials: Make Materials Data FAIR

We provide here tools used by the Solar Durability and Lifetime Extension Center (SDLE) for FAIRifying data from materials science projects. Functions have been created for numerous tools common in the field in order to make the metadata more Findable, Accessible, Interoperable, and Reproducible.

Version:	0.4.1
Depends:	R (≥ 2.10)
Imports:	stringr, jsonlite, svDialogs, tidyjson, RColorBrewer, igraph, dplyr, graphics, jsonld, utils
Suggests:	knitr, rmarkdown, testthat (≥ 3.0.0)
Published:	2023-08-09
Author:	Will Oltjen [aut], Mingjian Lu [aut], Xuanji Yu [aut], Liangyi Huang [aut], Arafath Nihar [aut], Tommy Ciardi [aut], Erika Barcelos [aut], Pawan Tripathi [aut], Abhishek Daundkar [aut], Deepa Bhuvanagiri [aut], Hope Omodolor [aut], Hein Htet-Aung [aut], Kristen Hernandez [aut], Mirra Rasmussen [aut], Olatunde Akanbi [aut], Raymond Wieser [aut], Sameera Nalin-Venkat [aut], Tian Wang [aut], Weiqi Yue [aut], Yangxin Fan [aut], Rounak Chawla [aut], Leean Jo [aut], Jarod Kaltenbaugh [aut], Zelin Li [aut], Jonathan Gordon [aut], Jeffrey Yarus [aut], Mengjie Li [aut], Kristopher O Davis [aut], Laura Bruckman [aut], Yinghui Wu [aut], Roger French [aut, cre], Solar Durability and Lifetime Extension Center [cph, fnd]
Maintainer:	Roger French <roger.french at case.edu>
License:	BSD_3_clause + file LICENSE

Data ingested at CRADLE/HPC

Linking Data in a Domain for Efficient Pipelining & Modeling

Metadata and data are linked by unique ids

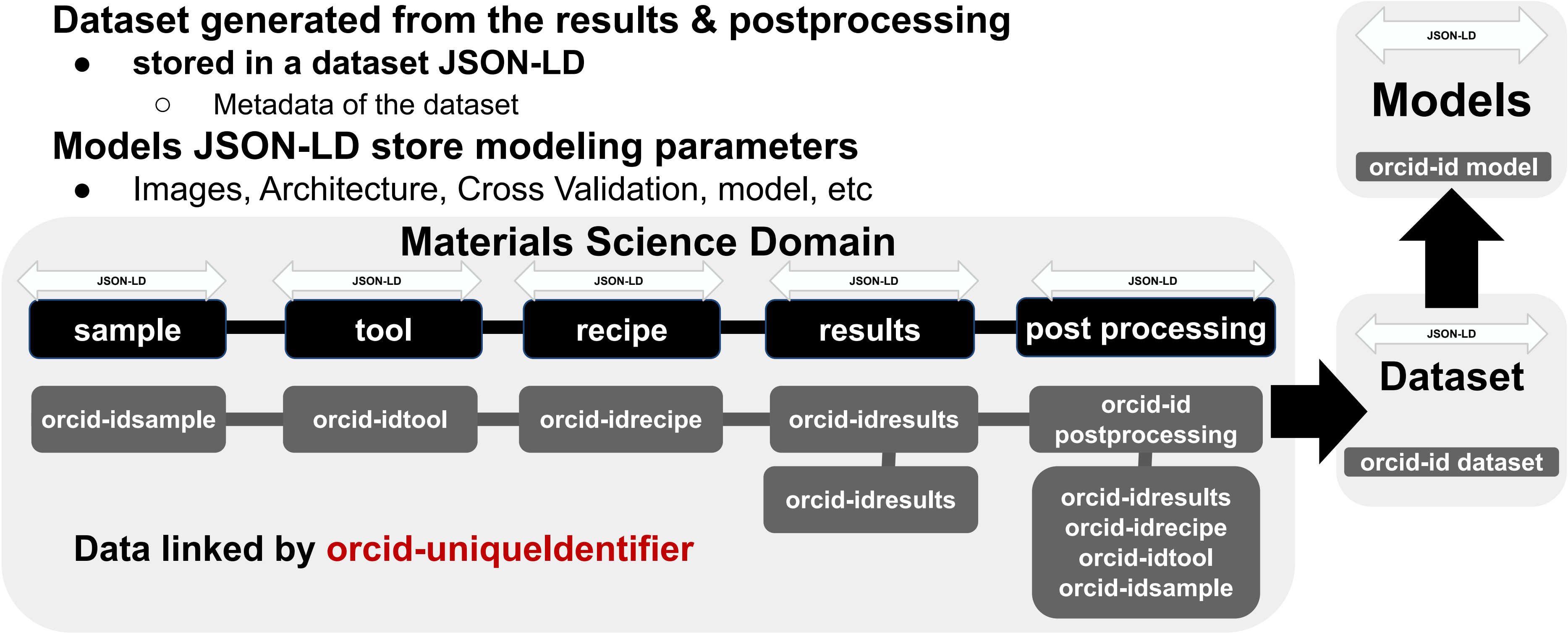
- associated to the user's ORCID

Dataset generated from the results & postprocessing

- **stored in a dataset JSON-LD**
  - Metadata of the dataset

Models JSON-LD store modeling parameters

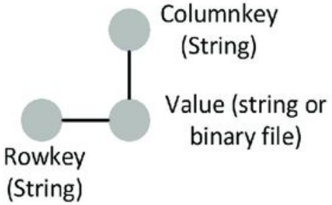
- Images, Architecture, Cross Validation, model, etc



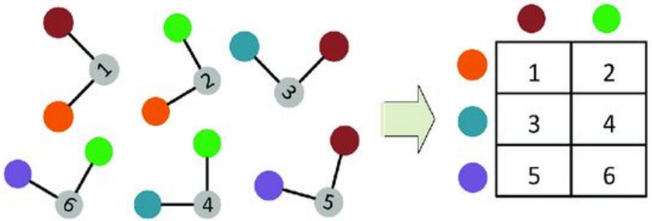
Development of Domain Ontologies: Knowledge Graphs

Apache HBase:
- Data Storage and
- Represented **in RDF Triples**

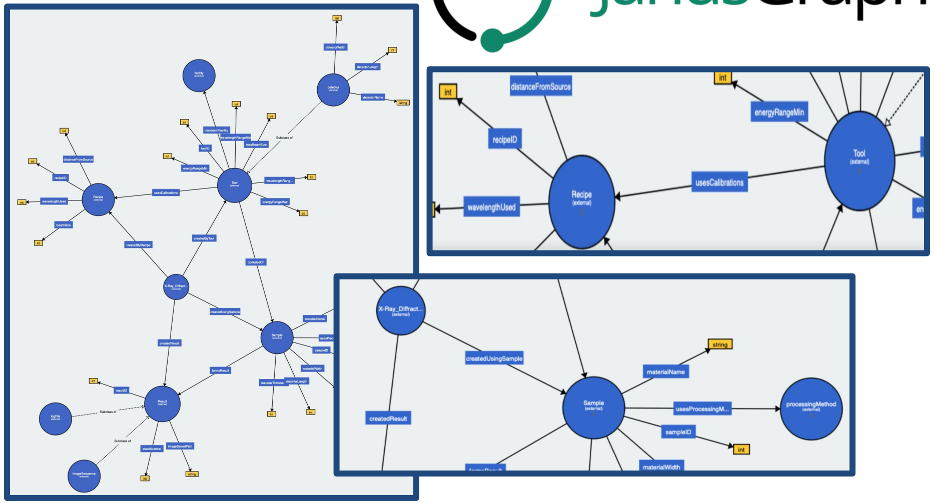
Onotologies created in OWL language
- Builds on top of RDF
- Extends RDF for complex knowledge & reasoning
- Provides a more expressive language
  - And larger vocabulary



Hbase Triple



Creation of Ontology-driven Knowledge Graphs
- **JanusGraph Distributed Database**
  - Scalable graph database optimized for
    - storing and querying graphs
    - containing hundreds of billions
    - of vertices and edges
    - distributed across D/HPC CRADLE



Khalilnejad A, et al., Automated pipeline framework for processing of large-scale building energy time series data. PLoS ONE 2020 m15(12): e0240461.
W. C. Oltjen, et al., "FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data," IEEE PVSC, 2022.

Development of Domain Ontologies: Knowledge Graphs

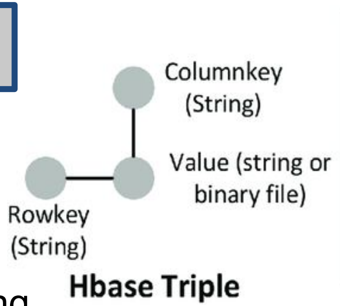
Apache HBase:

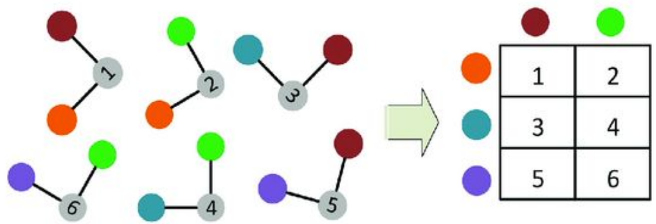
- Data Storage and
- Represented **in RDF Triples**

RDFs is a cold watery coffee, while **OWL** is a hot espresso

Onotologies created in OWL language

- Builds on top of RDF
- Extends RDF for complex knowledge & reasoning
- Provides a more expressive language
  - And larger vocabulary

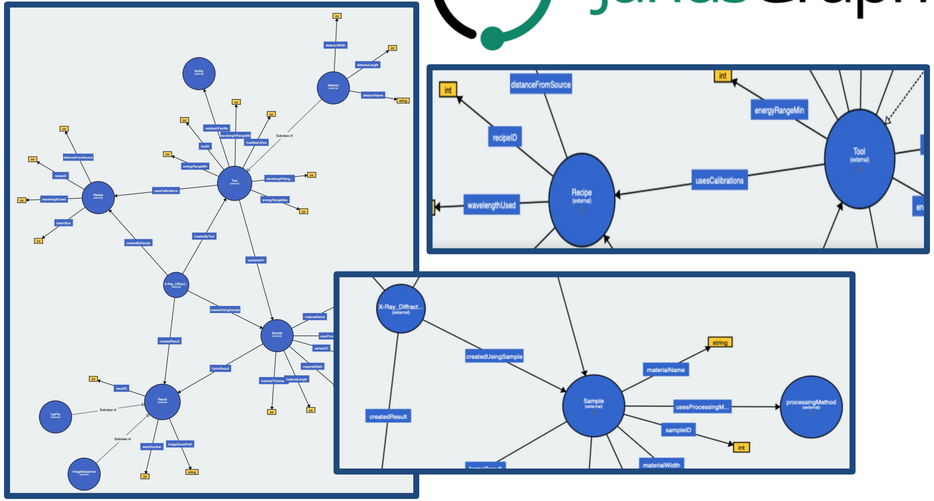




JanusGraph

Creation of Ontology-driven Knowledge Graphs

- **JanusGraph Distributed Database**
  - Scalable graph database optimized for
    - storing and querying graphs
    - containing hundreds of billions
    - of vertices and edges
    - distributed across D/HPC CRADLE



Khalilnejad A, et al., Automated pipeline framework for processing of large-scale building energy time series data. PLoS ONE 2020 m15(12): e0240461.
W. C. Oltjen, et al., "FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data," IEEE PVSC, 2022.

CRADLE Data Lifecycle:

Scientific Investigations, Study Protocols & Materials Data Science

GS: Kristen Hernandez[1], Hein Htet Aung[1], Ayorinde Olatundei[2], Arafath Nihar[3], Olatunde Akanbi[1], Tommy Ciardi[3], Tian Wang[3],
UG: Rachel Yamamoto[3], Rounak Chawla[1], Hayden Caldwell3,
Faculty: Anirban Mondal[1], Laura S. Bruckman[1] , Yinghui Wu[2], Vipin Chaudhary[3], Roger H. French[1,2]

1. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA
2. Department of Computer and Data Sciences, CWRU, Cleveland, OH
3. Department of Mathematics, Applied Mathematics, and Statistics, CWRU, Cleveland, OH

Scientific Investigations: Study Protocol Pipeline Schema

Systematic Review

Identify Research Question
Uncalibrated measurement
Abnormal to standard
Falsifiable

Scientific Investigations

Experimentation

Design

Production Optimization

Aging & Lifetime

Surveillance

Materials Data Science

Data FAIRification

Data Curation & Ingestion

Exploratory Data Analysis

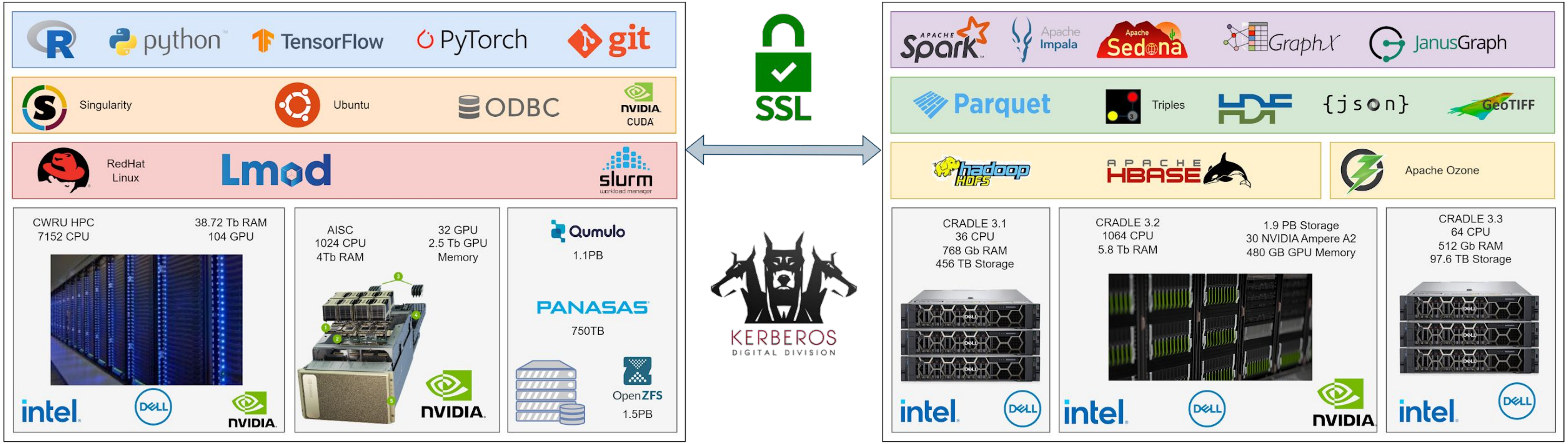
Knowledge Management & Learning

Statistical Learning

Machine Learning

Deep Learning

CRADLE Frameworks: Enabling Materials Data Science

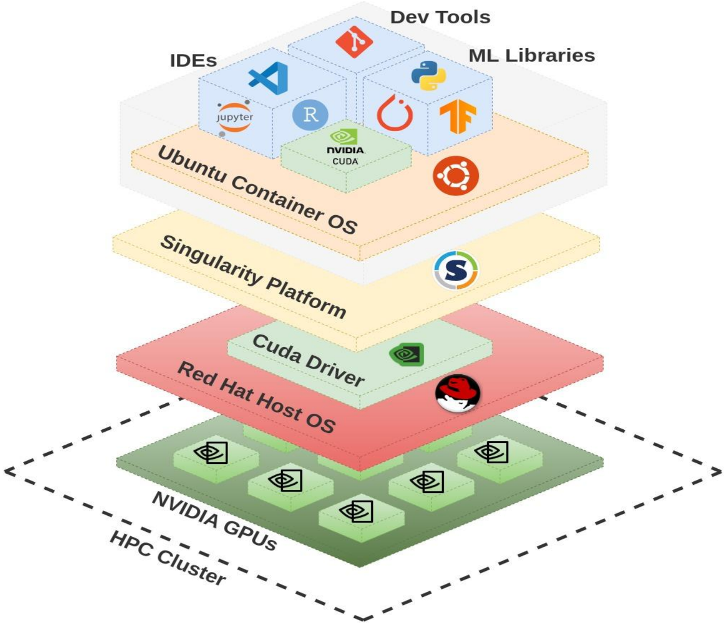


- **Combined Distributed & High Perf. Computing**
- **Distributed Computing to reduce Data Motion**
- **Integrated AI Engines such as AISC**
- **Permanent Data Storage: To enable data curation**
- **“Low Barriers To Entry” Accessible Data Science Tools**

A Containerized Environment for Researcher Ease of Use

Containerized environments enable:

- Researchers: To use CRADLE
  - without extensive compute training
- Group: consistent tools/code packages
  - for an entire team

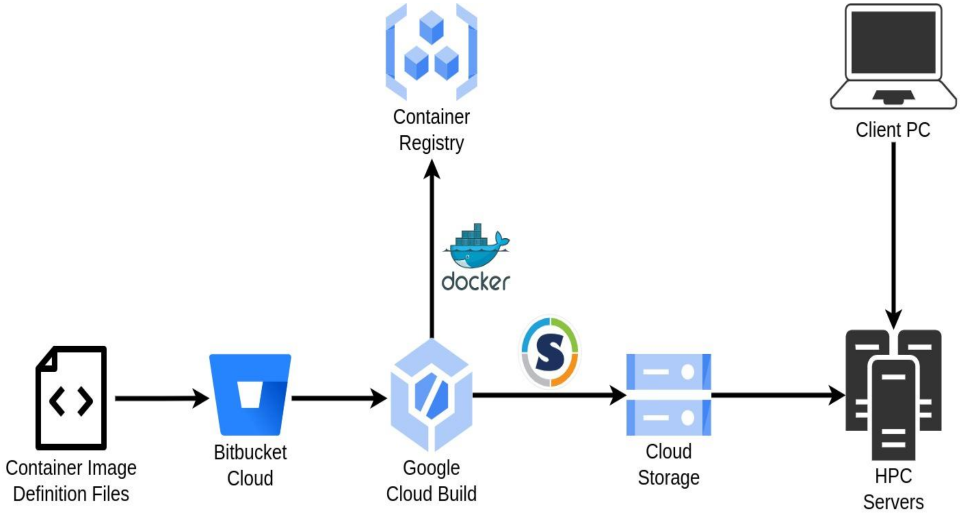


Cloud based container building pipelines

- Ensures features and fixes
- Are released to production
  - For the entire research group
- Users don't need to manage dependencies

From a single source

- Using our Container Registry



OnDemand Apps: Using Containerized OS & Applications

Containerized environments enable:

- Researchers: Use CRADLE
  - Without extensive compute training
- Group: consistent tools/code packages
  - For an entire team

Browser access to CRADLE D/HPC

- Pre-configured data science environment

Easy Access to CRADLE D/HPC

- Storage, CPUs & GPUs

Providing

- Integrated Development Environments: R/Python
- ***CRADLE Data Explorer***
- ***SDLE Diagnostics***
  - Web app to detect & fix infrastructure issues
- **WebVOWL** & **JSON-LD** Servers: FAIRmaterials

CRADLE Data Explorer: PV Systems, {meta}data, Quality

Ingest 100,470 Photovoltaic Systems

- To CRADLE3
  - Into HDFS
  - As Parquet Files
- Using **Apache Spark3**

Distributed Across

- 1000 CPUs
- 100 HDDs

Apache Impala

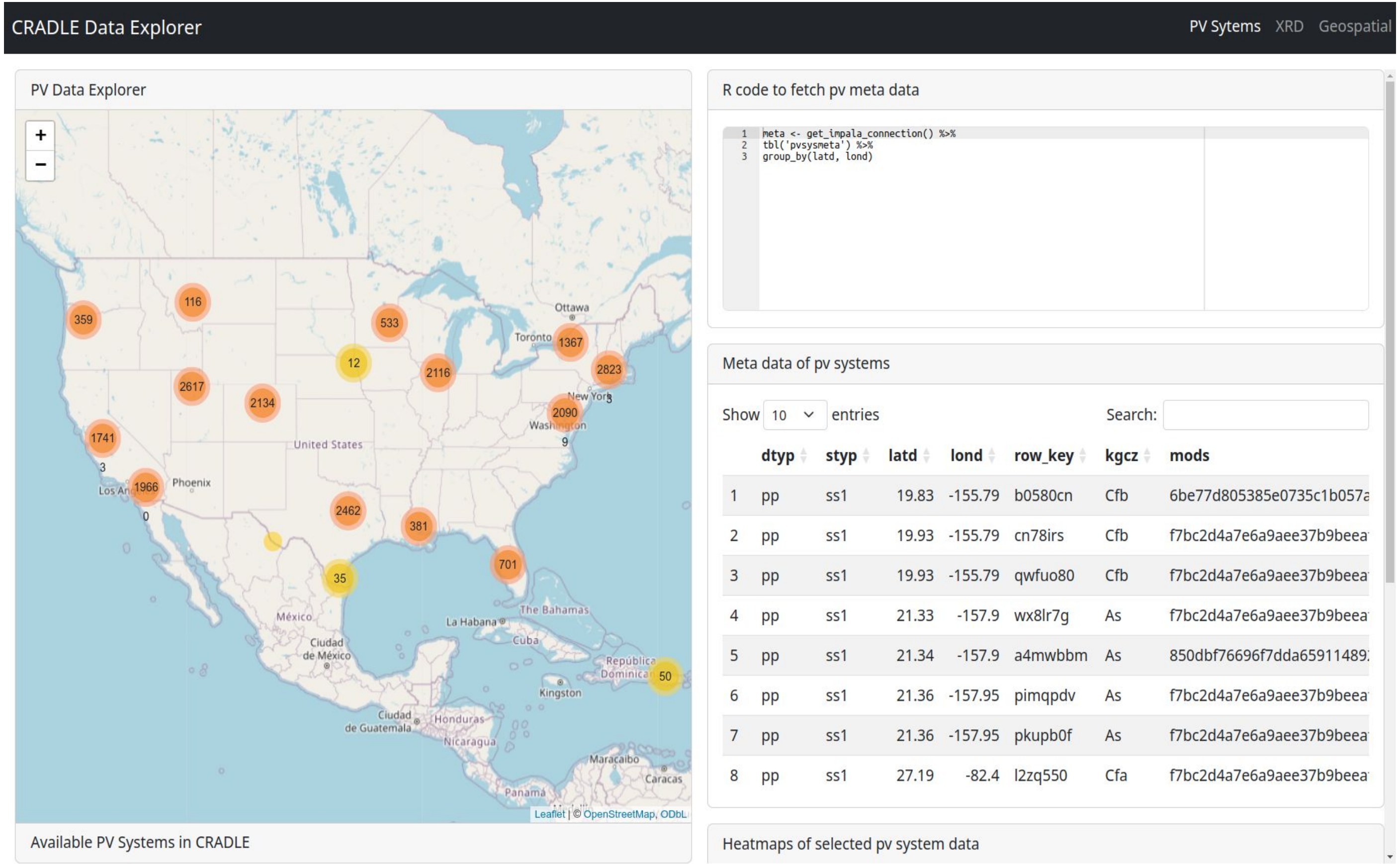
- For SQL Queries

Provide Codebox

- For Customized Queries

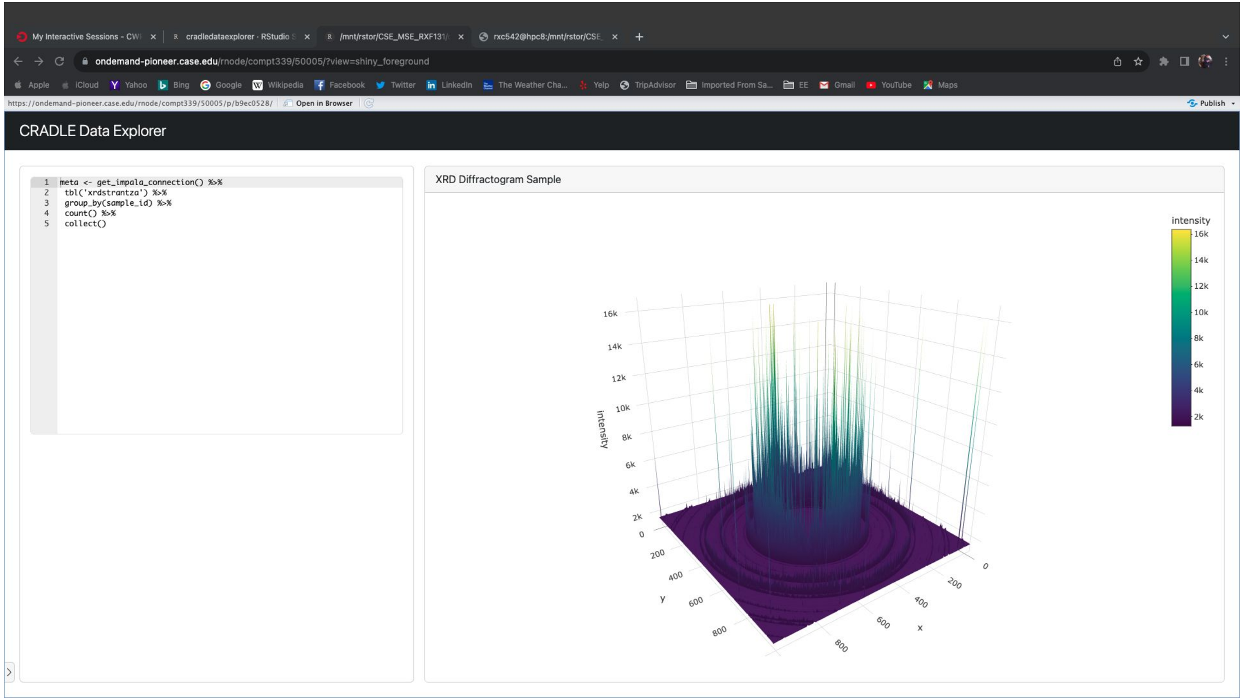
Retrieve All Metadata

- Data Quality Heatmap



Interactive 3D Plots of XRD Diffractograms

- Securely query data from CRADLE
- And interact with it in your browser



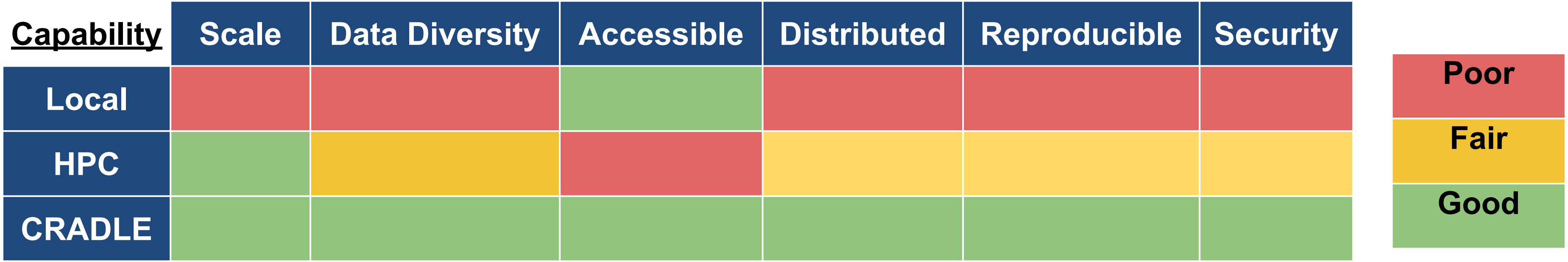
CRADLE's D/HPC Architecture Offers Next Generation Capabilities

HPC and Hadoop hybrid infrastructure enables the ability to **handle next generation datasets**

Raw compute and distributed

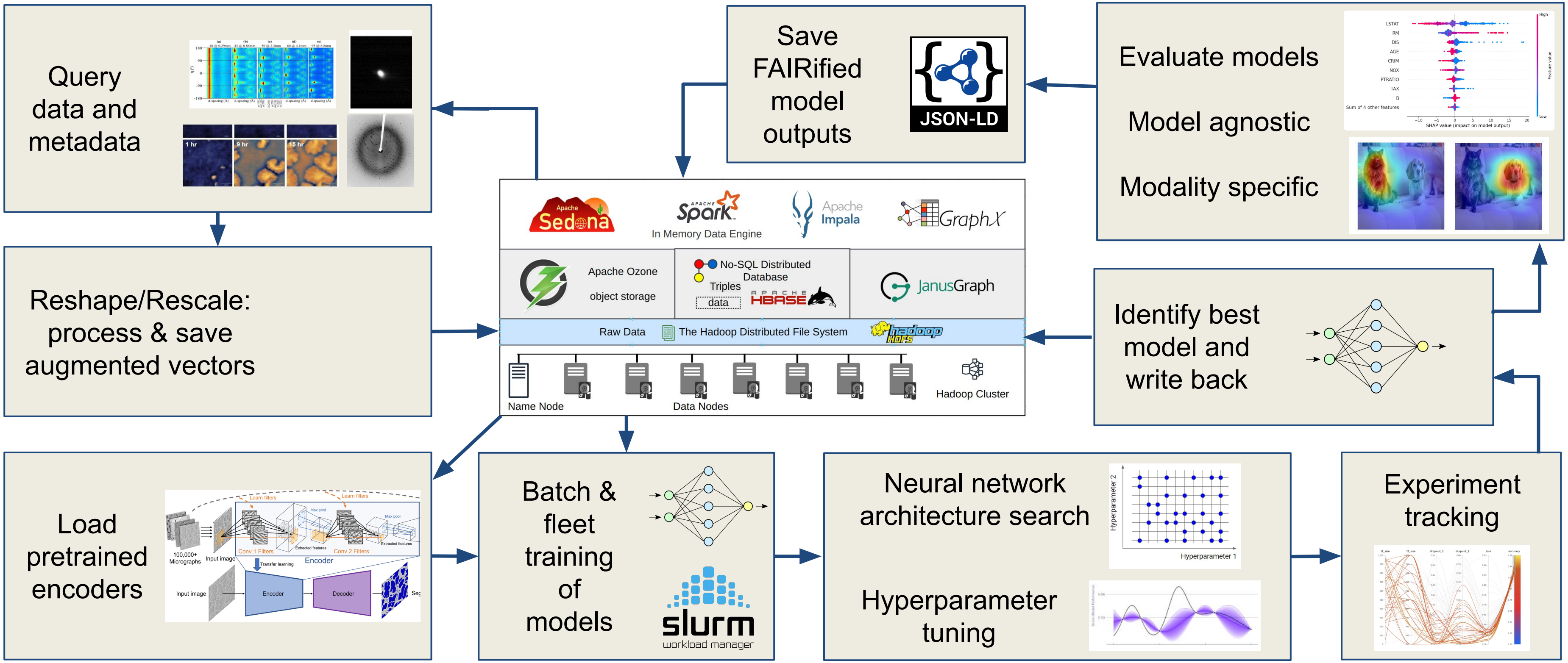
In one night a scientist can train on…

- **laptop** → **1-2 runs** → **1-6 GB datasets**
- **HPC** → **1000 runs** → **2 TB dataset**
- **CRADLE (HPC/ Hadoop)** → **10000+ runs** → **10+ TB datasets**



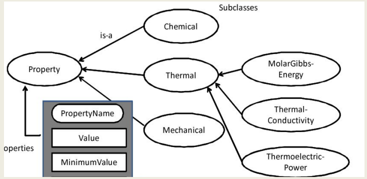
CRADLE Data Science Modeling & Learning Framework

How do we find the best possible model and make our efforts reproducible?



MDS^3-COE's: Knowledge Graph Learning Framework

Metadata, Ontologies
RDF triples, JSON
(abstraction/semantics/constraints)



... to data standardization & knowledge sharing

1

validate

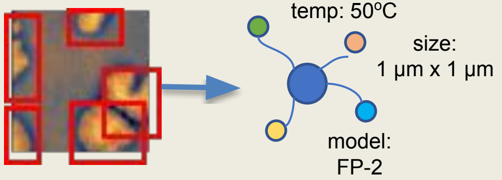
curate

FAIRification

Objects, Observations & Properties
(instances)

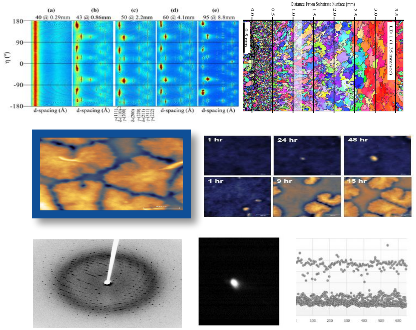
2

Featurization



... to create AI/ML ready data resources

Raw datasets
Image/videos, design data, in-situ/ex-situ data..



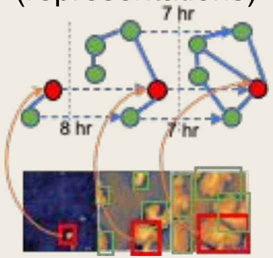
ML/AI models

physical con-straints

linked entities

enriched features

Scenes or st-Graphs
(representations)



Deep st-graph representation learning
... to inferential & predictive models

Deep Learning

4

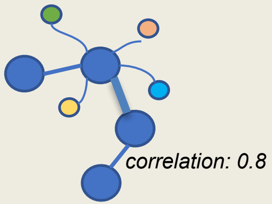
cost-effective learning

efficient access & interpretation

Summarization

3

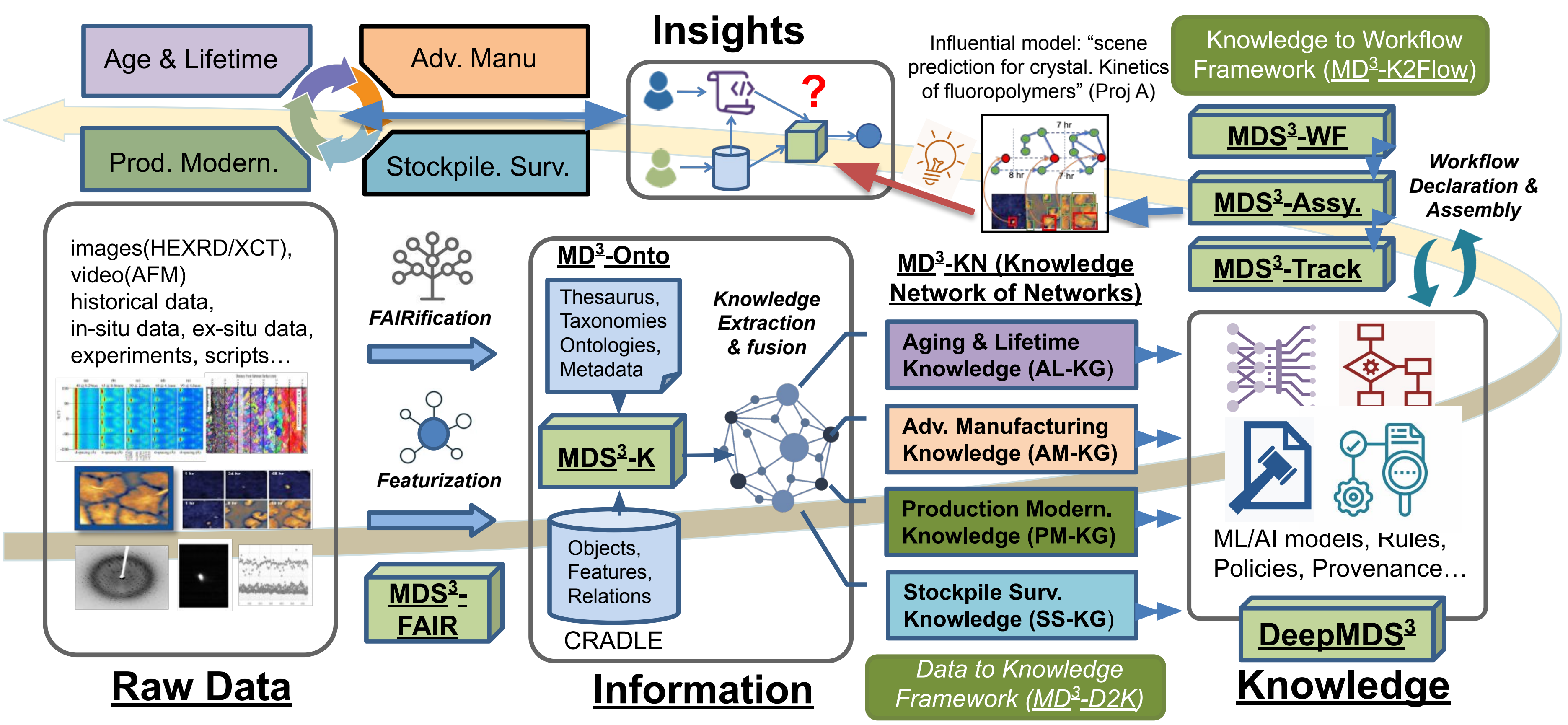
Summary Graphs
(patterns)



... to cost-effective data access, analysis & interactive exploration

[1] Roger H. French et al., "Fairmaterials." The Python Package Index (PyPI), Oct. 08, 2021, https://pypi.org/project/fairmaterials/ .
[2] Willam C. Oltjen, et al., "FAIRmaterials: Make Materials Data FAIR." CRAN, Sep. 14, 2021, https://CRAN.R-project.org/package=FAIRmaterials .
[3] A. M. Karimi, et al., "Spatiotemporal Graph Neural Network for Performance Prediction of Photovoltaic Power Systems," in Proc. of AAAI,, 2021, https://ojs.aaai.org/index.php/AAAI/article/view/17799 .

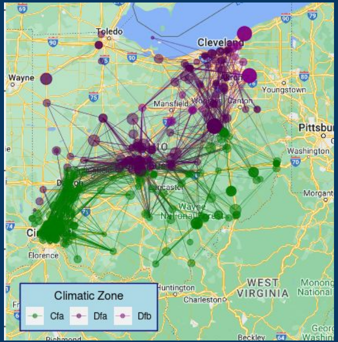
MDS^3 Data to Knowledge, Knowledge to Workflow Framework



Spatiotemporal-Graph (st-Graph) Learning:

Timeseries Imputation

& Trend Estimation

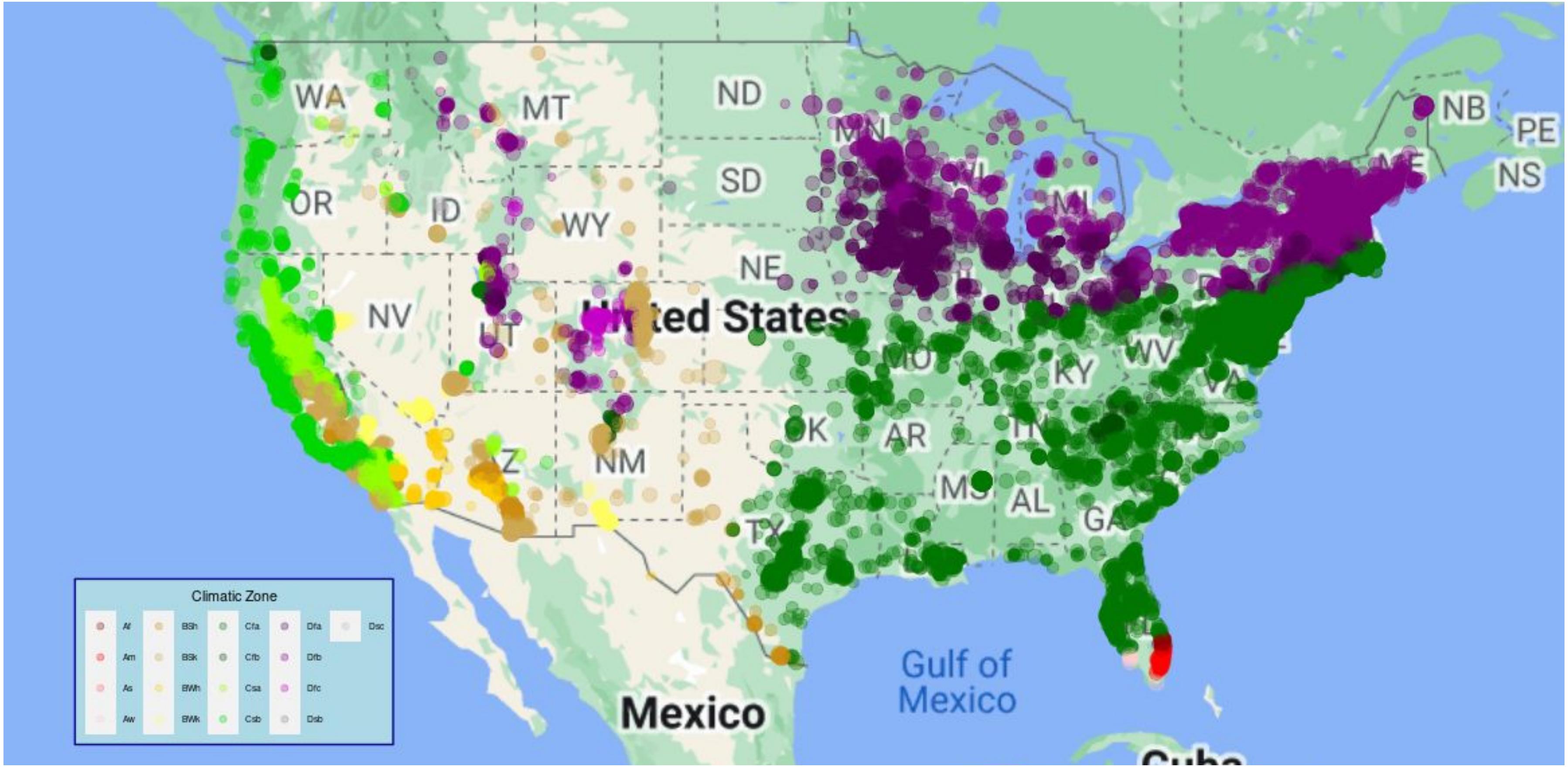


GS: Yangxin Fan[1], Raymond Wieser[1]
UG: Jiana Kambo[1], Hyangmok Baek[1]
Faculty: Yinghui Wu[1], Laura Bruckman[2], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

Award No: DE-EE0009353

Large Scale Photovoltaic Fleet Monitoring: 104,700 PV Systems



PV Network Representation

Inverters

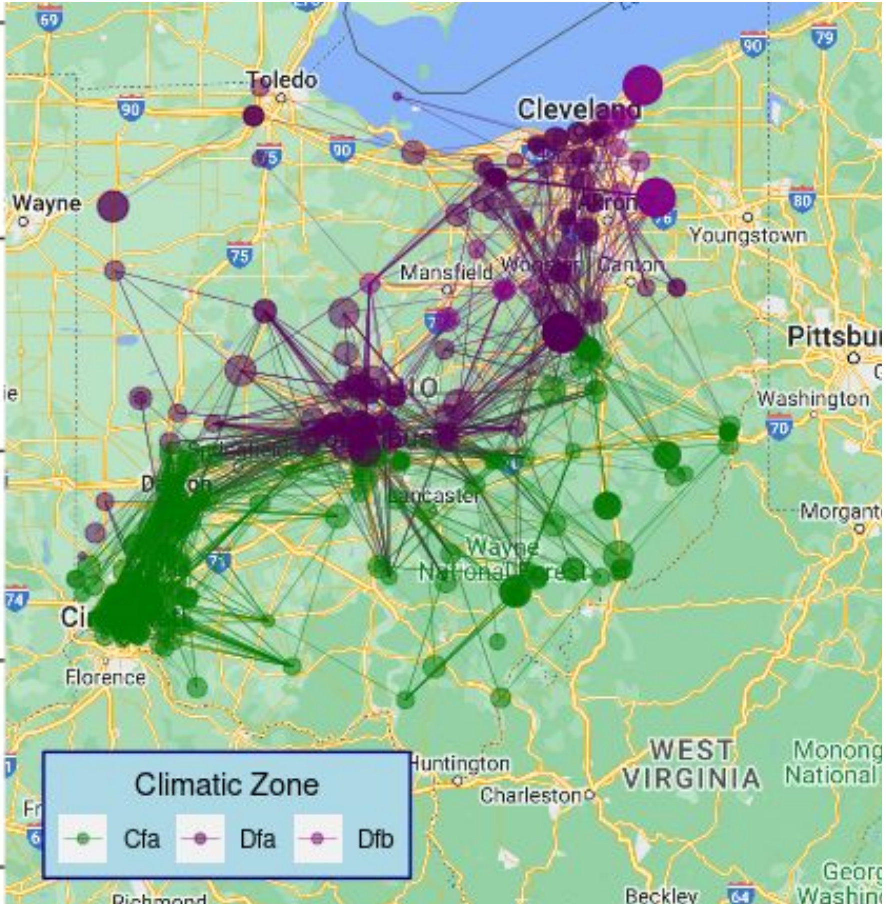
- "Nodes"
- Individual Timeseries

Site "Similarity"

- "Edges"
  - How much information
  - Should connections "share"

Evaluating "Similarity"

- Distance (Spatial Coherence)
- Cell Type
- Nameplate Power
- Benefits from "FAIRified" datastreams



Network Representation of 295 Inverters
(edges sparsified for visualization)

Why Spatiotemporal Graphs (st-graphs)?

Graphs are enhanced data structures with

- **Nodes:**
  - information about a particular object
- That provide:
  - information about other objects through
- **Edges:**
  - through their relationships ("Edges")

st-Graphs have distance-based

Spatial coherence threshold **epsilon**

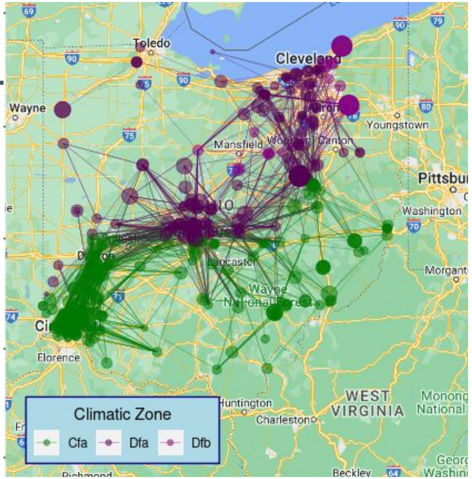
- Values between 0 and 1
- epsilon = 0.75 (st-GAE)
- epsilon = 0.25 (Decomposition)

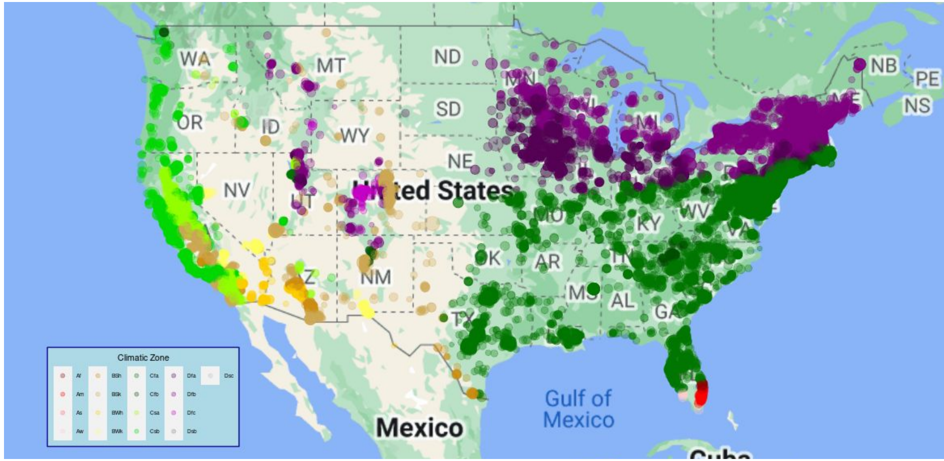
We have over 100,000 Nodes

- Of Photovoltaic Power Plants
  - Timeseries Power data (5 min. interval)
- As st-Nodes ingested
  - Into CRADLE infrastructure!

PV systems

- Local Weather
- System Age
- Technology
- Module Brand





Graph Neural Network Computing at Scale

In a GNN model

- Computing the embeddings of a node
  - depends on the embeddings of its neighbors
- This leads to **exponential growth**
  - of the number of nodes
  - involved with number of GNN layers

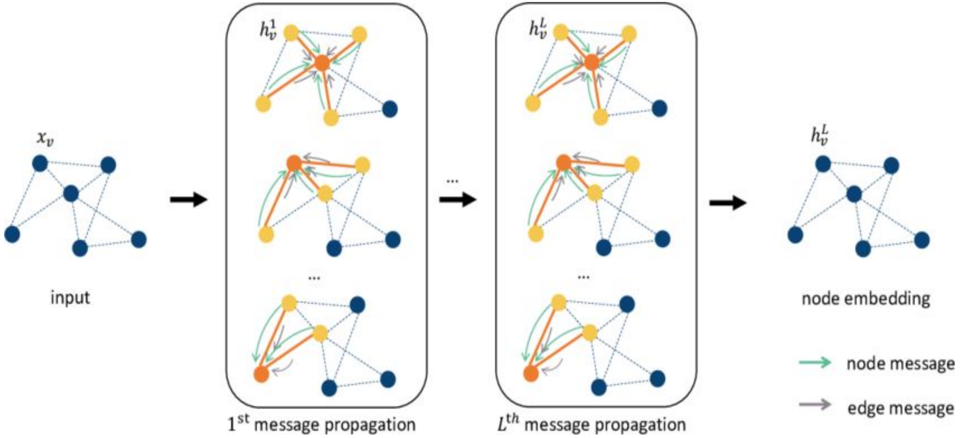


Illustration of message passing in GNN[1]

Hence, large-scale graph learning is very challenging

- Vanilla GNNs fails to scale up,
  - Limited by the GPU memory space

Most large-scale graph learning leverage sampling-based methods

- Such as
  - neighbor sampling,
  - layer sampling, and
  - random-walk sampling
- But may **sacrifice** prediction accuracy

[1] Liu W et al. Item relationship graph neural networks for e-commerce. IEEE Trans. Neu. Net. & Learn. Sys., 2021, 33(9): 4785-4799.

Large st-Graph Calculation Benchmarks

Benchmark tests using CRADLE's

- State-of-the-art CPUs & GPUs

Compute 100K^2 adjacency matrix

- Using multi-processing per compute node
- And fleet out jobs across compute nodes
  - Using SDLEfleets package:
    - ~19 Days ➞ ~ 2 hrs

Use 1 NVIDIA A100 GPU, 80GB VRAM

- For large-scale graph learning
  - Without subgraph sampling

AISC is 32 Integrated A100 GPUs!

- With integrated RAM & NVME Storage
- A different, but critical form of
  - "Converged Computing"

Benchmark Results

- **Model: st-GAE-Impute**
- **Large Scale st-Graph AutoEncoder**
  - 10k nodes, ~1 million edges
  - 1-year timeseries for each node
  - 5-minutes interval
- **Training time**
  - Using 1 year of timeseries data
    - 5 min. Interval
  - Run time: 1 hour 55 minutes
    - Epsilon = 0.25
- **Inference time: For Data Imputation**
  - On two-months data
  - Run time: 56 seconds

CRADLE Benchmarks Table

Benchmarking compute task performance on HPC versus CRADLE (Hadoop3) distributed computing, with Spark3 and Nvidia Rapids distributed GPU computing.		
Task \ Compute Infrastructure	**HPC**	**CRADLE**
PV: 100K adjacency matrix construction (**10 billion distance calculations**)	19 days (24 CPUs)	19.1 minutes (Spark3)
PV: 100K PV system graph community detection	8.47 hours (40 CPUs)	5.1 minutes (RAPIDS)
PV: Query 5000 PV systems power data, from **2.6 billion rows**	N/A (overloads HPC RAM >250 GB)	~1 hour (Impala, Spark3)
Image Conversions: 100k `.ibw` image files to `.tiff`	11.1 hours	0.62 hours (Spark3)
Image Deep Learning: Hyperparameter tuning: Training 240 deep learning segmentation models	5.1 days	5.2 hours (SDLEFleets)
Image: Nearest neighbor crystallite calculations	3.8 minutes	1.8 minutes (RAPIDS)

Timeseries Data Reconstruction and Generative Data Imputation

For PV: Performance Loss Rate (PLR)

- Critical to profitability of asset

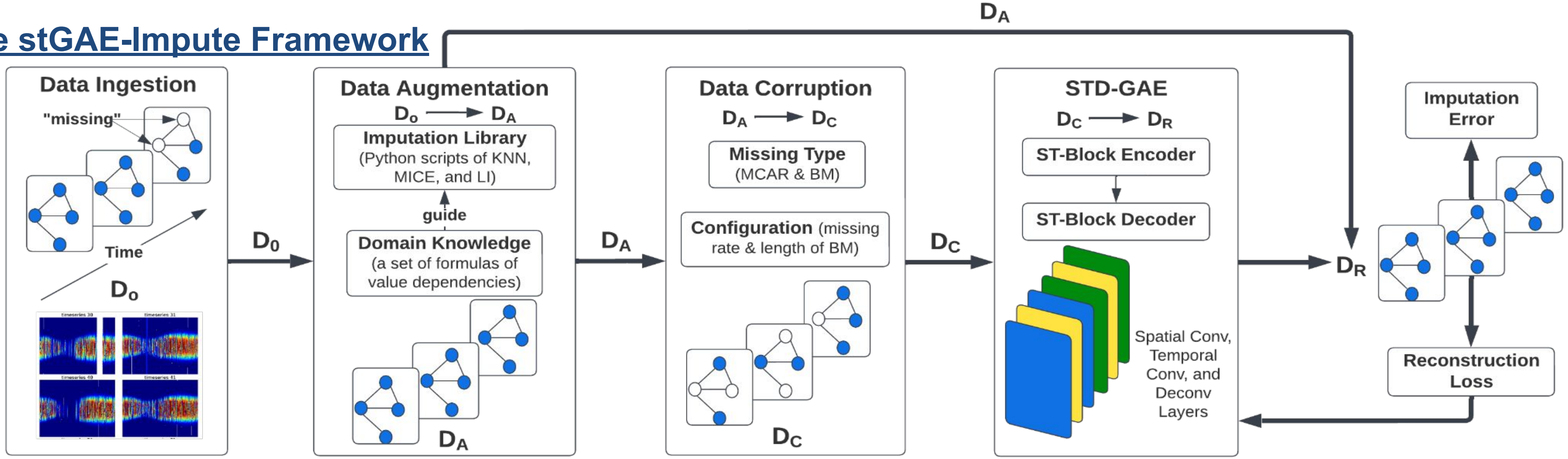
Data Quality Impacts PLR estimates

- Low Quality Data
  - Low Quality PLR estimation
  - High uncertainty, Low accuracy

Data Imputation improves low quality data

- Physical Models
- Predictive Mean Matching
- Gradient Boosting Regression
- Traditional Imputation Methods

The stGAE-Impute Framework



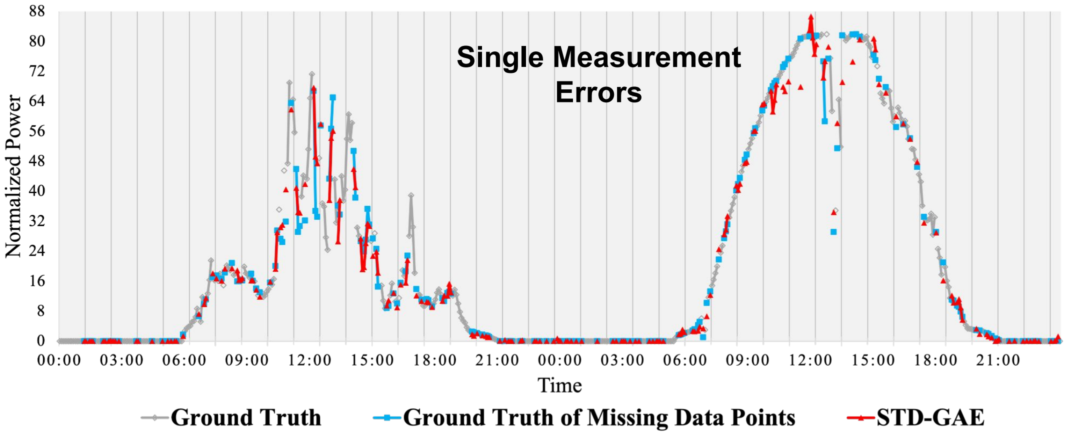
D_O: Observed PV Data; D_A: Augmented PV Data; D_C: Corrupted PV Data; D_R: Recovered PV Data.

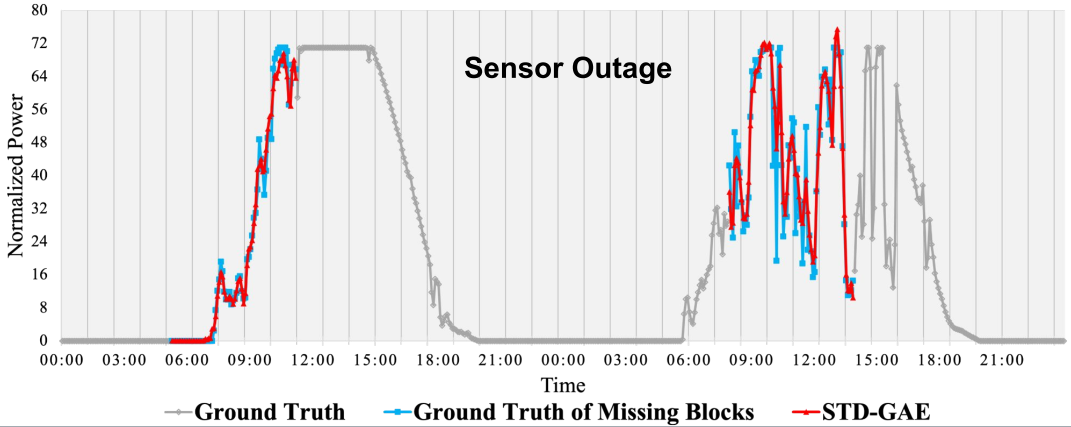
Data Imputation Accuracy

st-GAE

- Missingness Types
  - Single Value Corruption
  - Measurement Outage

- Missingness Severity
  - 10% - 60% Measurements Missing
  - 2hrs - 6hrs Inverter Outage



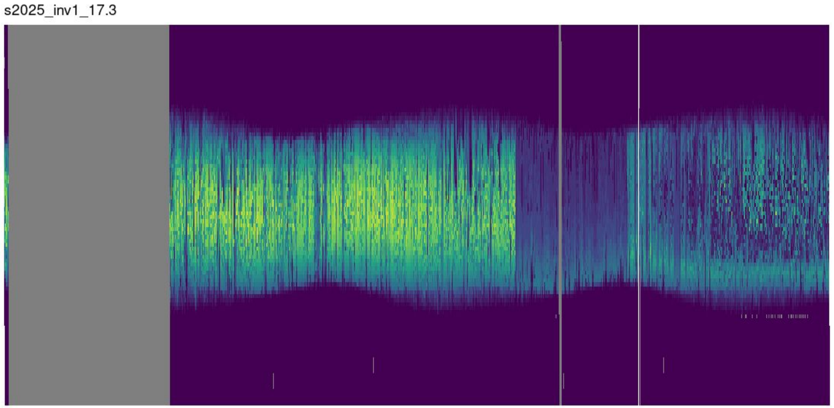


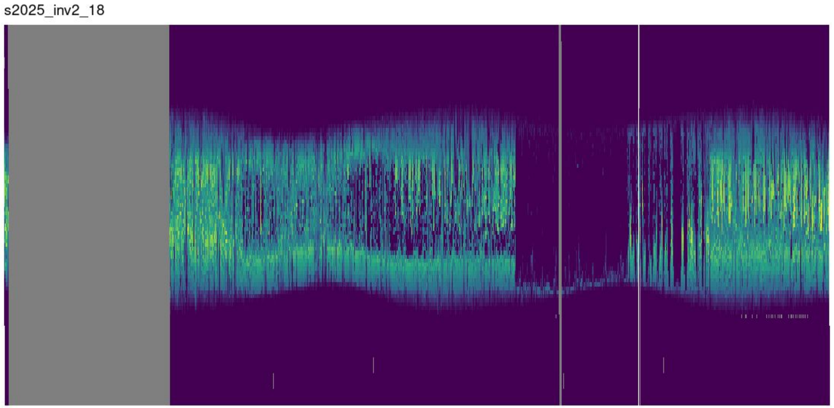
Model Accuracy

- Insensitive
  - Missingness Types
  - Missingness Severity
- st-GAE Outperforms
  - Traditional
  - Deep Learning

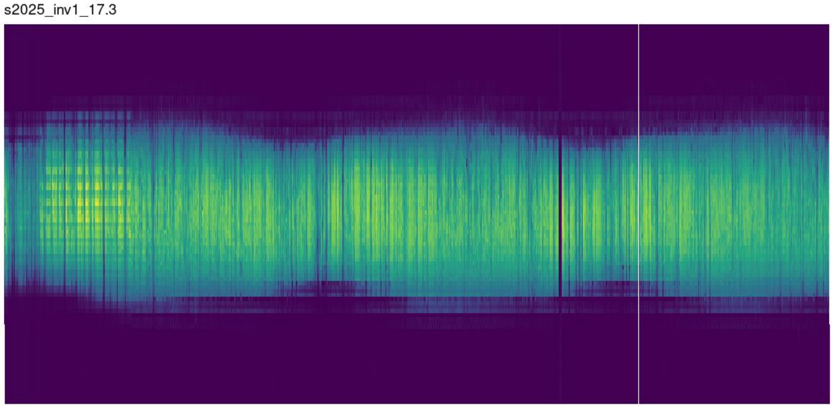
Data Reconstruction: Block Outages & Anomalous Measurements

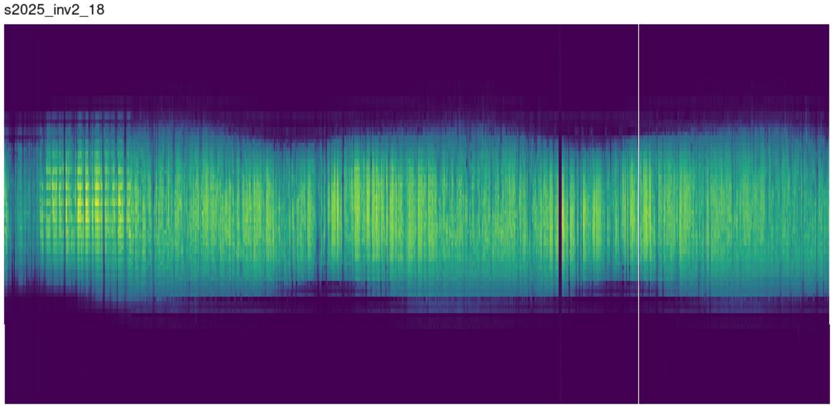
RAW



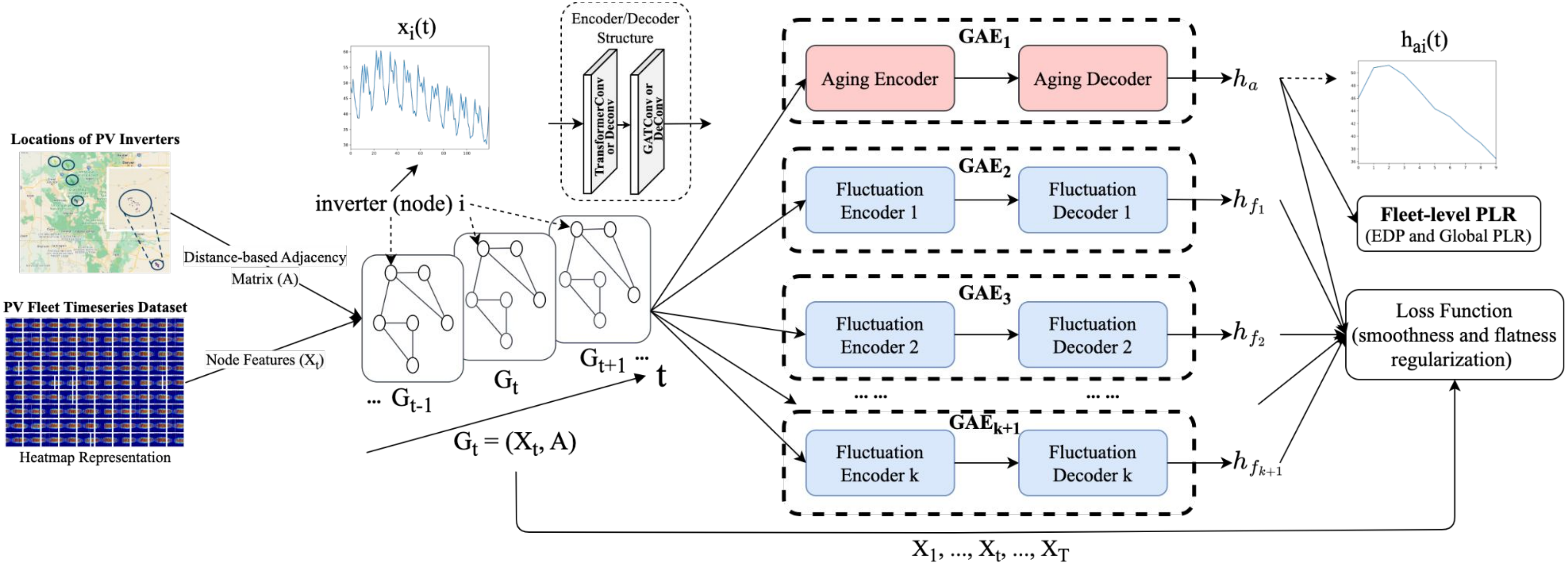


Reconstruction



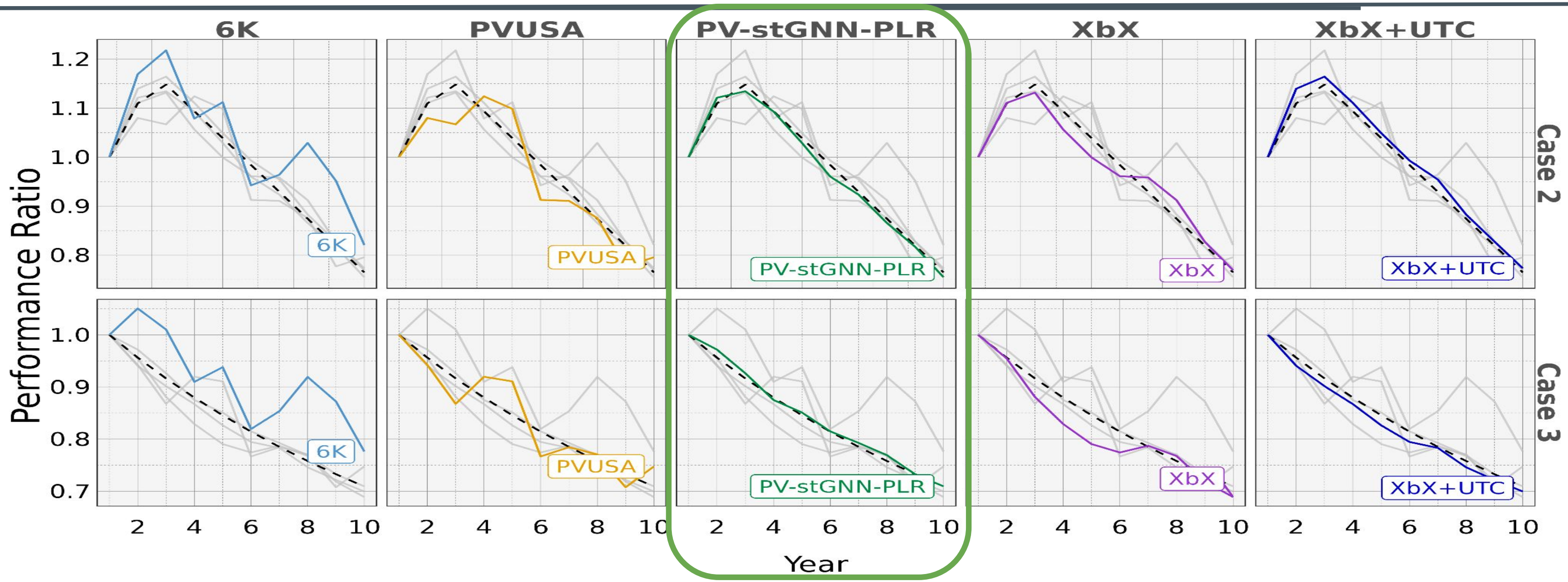


Timeseries Decomposition Framework: For PLR Determination



- **"Parallel-friendly" K+1 GAE (graph autoencoder) blocks**
- **One aging-term**
  - Extracts the long-term degradation pattern for PLR analysis
- **K different fluctuation terms**
  - Captures seasonalities and noises at different temporal resolutions

Trend Decomposition and Extraction



We compare **Estimated Degradation Pattern (EDP)** extracted by st-DynGNN

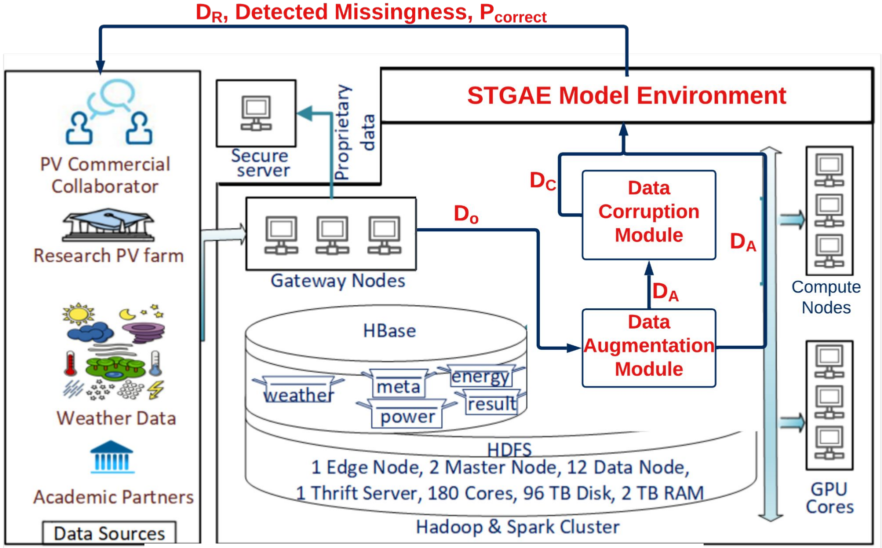
With top six best-performed baselines with **Real Degradation Pattern (RDP).**

- st-DynGNN can better recover real degradation pattern
- EDP extracted by st-DynGNN is the closest to RDP
  - in both case 2 and case 3 figures
  - followed by XbX+UTC and STGAE2

Spatiotemporal Graph AutoEncoder Takeaways

st-GAE exploits:

- Temporal Coherence
- Spatial Coherence
- Value Dependencies



Proposed Workflow in CRADLE

st-GAE:

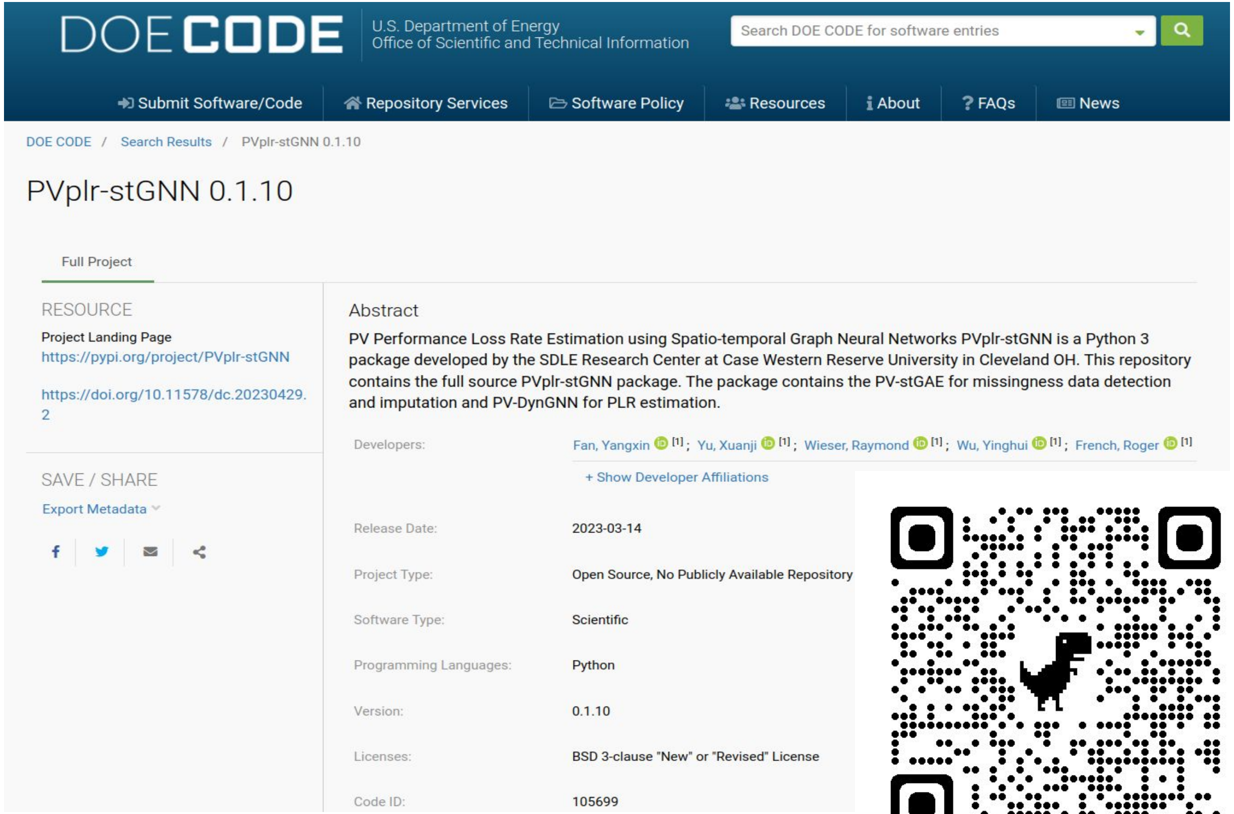
- Obtains **superior imputation accuracy**
- **Retains Raw Data** properties
  - ➔ Seasonality
  - ➔ Magnitude
- **Maintains robust performance** gains
  - ➔ % Missingness
  - ➔ Seasonality
- Graph-based Outlier Detection
  - ➔ "**Learned**" from **Fleet**
  - ➔ **Physics Informed Loss**
  - ➔ Data Similarity

ALL at TERABYTE SCALE tabular data!

Pre-trained Model: Availability

PVplr-stGNN

- pypi
  - https://pypi.org/project/PVplr-stGNN/
- DOE CODE
  - OSTI 105699
  - https://www.osti.gov/doecode/biblio/105699

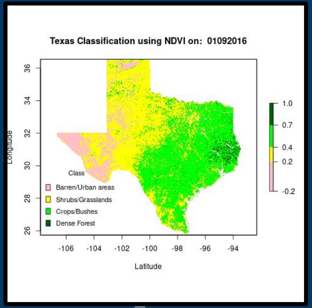


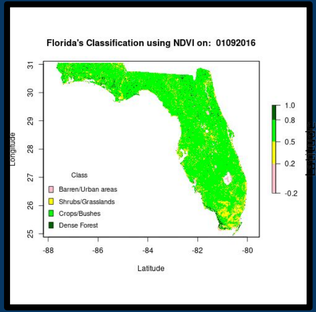
CASFER
Center for Advancing Sustainable and Distributed Fertilizer Production

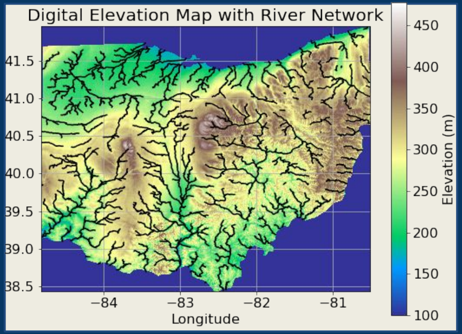
Geospatial Data Science

Eutrophication:

Motion of Nitrogen Through Watersheds







GS: Deepa Bhuvanagiri[1], Olatunde Akanbi[1]
UG: Vibha Mandayam[1], Lam Nguyen[1]
Postdoc: Erika Barcelos
Faculty: Yinghui Wu[1], Roger H. French[1,2], Jeffrey Yarus[2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

Water Contamination

Agriculture runoff

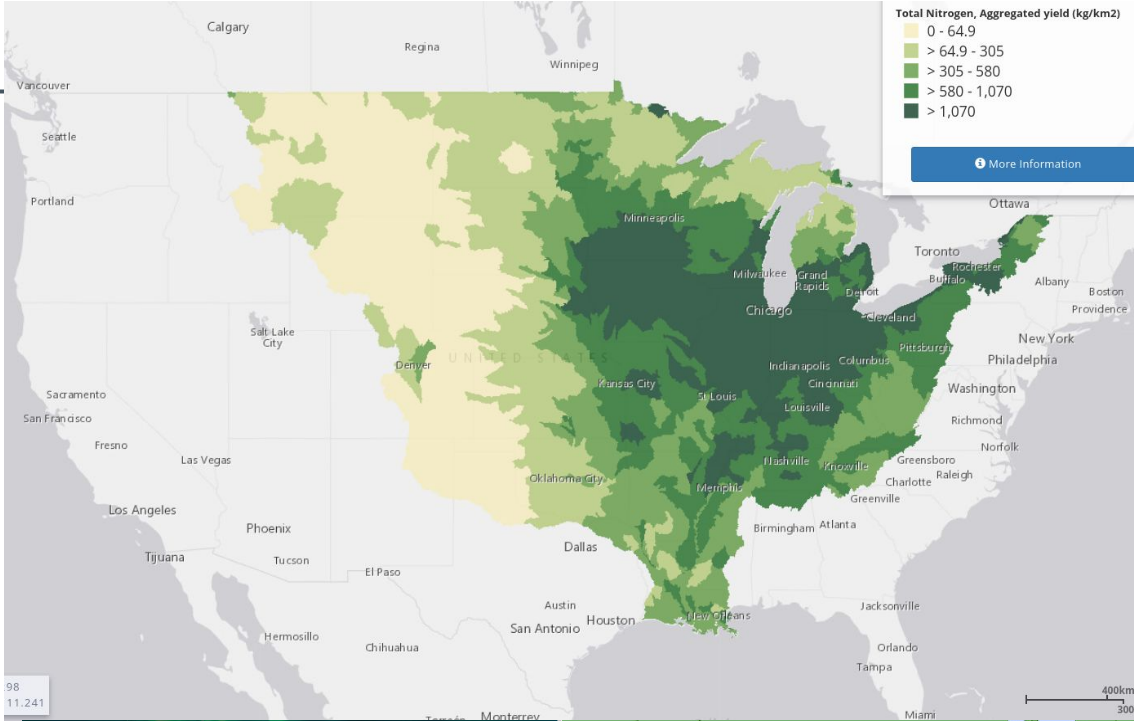
Manure

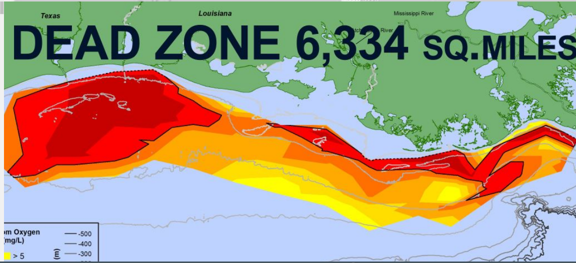
CAFOS lagoons
Rain, runoff, or faulty storage

Excess of Oxygen in Water

Algae Bloom Proliferation
Block sunlight
Release toxine
Consume Oxygen

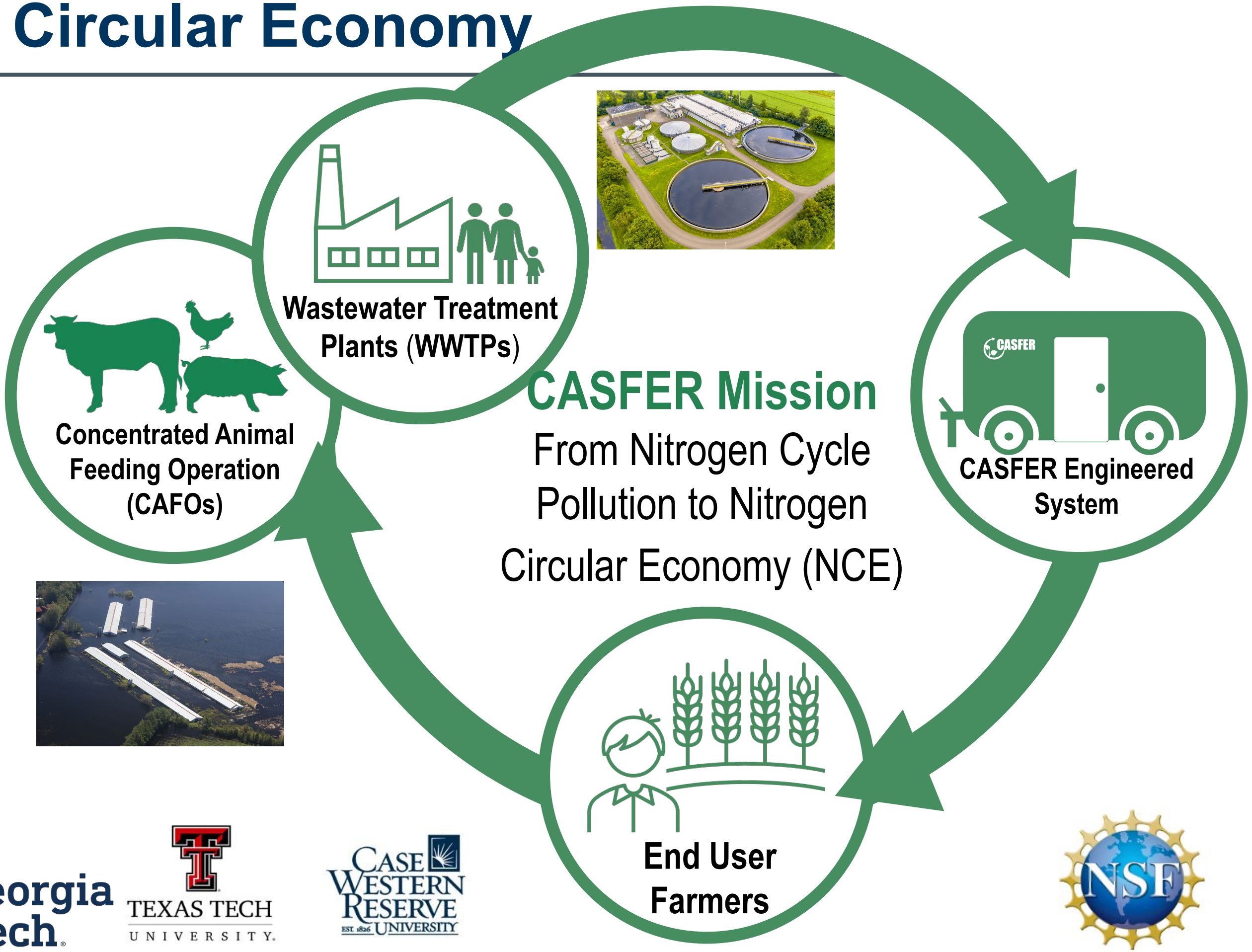
Dead fish & animals
Economic, Social and Environment





Towards a Nitrogen Circular Economy

- CASFER will enable **resilient** and **sustainable** food production by
  - Developing **next generation, modular, distributed,** and **efficient** technology
    - For capturing, recycling, and producing NBF



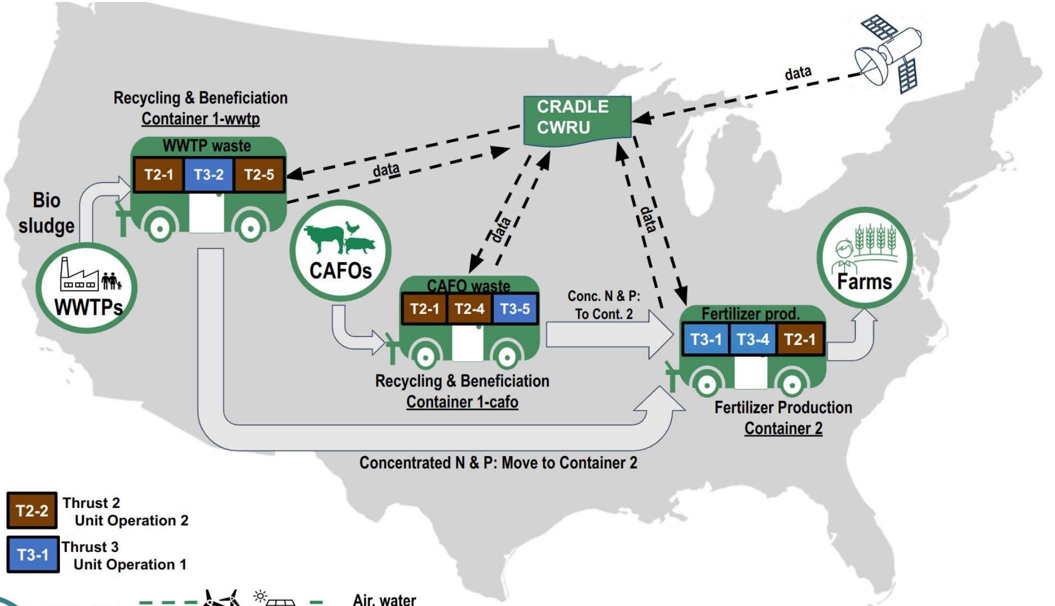
FAMU FLORIDA AGRICULTURAL AND MECHANICAL UNIVERSITY | Georgia Tech | TEXAS TECH UNIVERSITY | CASE WESTERN RESERVE UNIVERSITY EST. 1826 | NSF

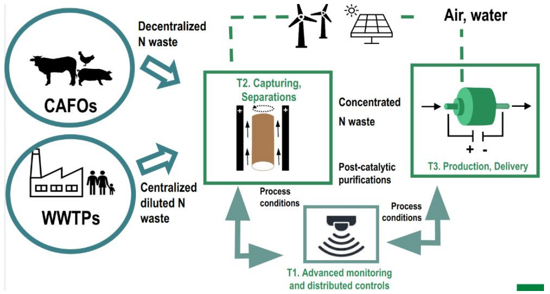
Spatiotemporal Predictive Modeling : Goal

Develop spatiotemporal models to **predict nutrients distribution in watershed**

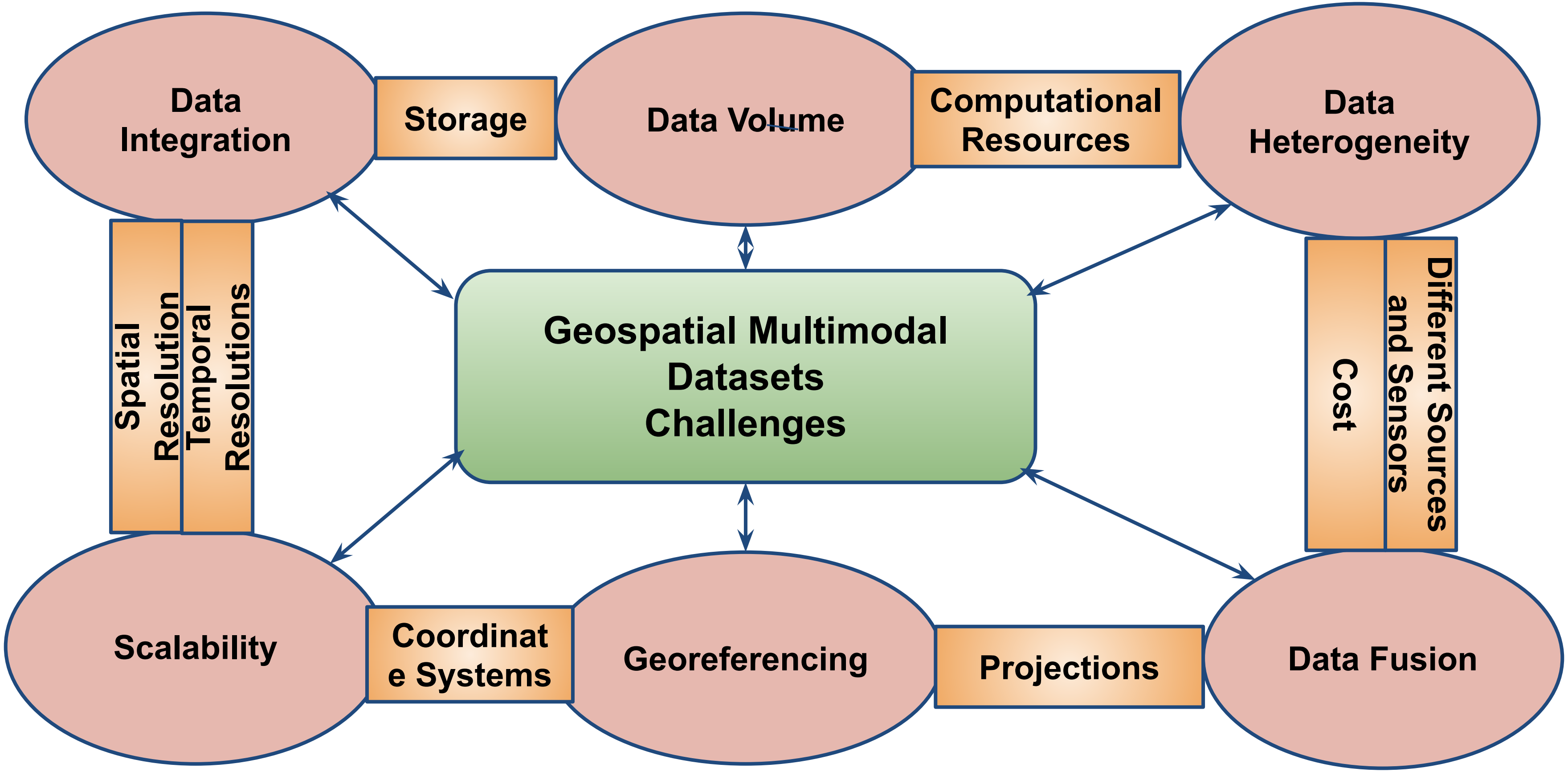
Understand and rank factors controlling flow of N and P:

- Rain, wind, crops, soil type, type of fertilizer, elevation, CAFOS, practices of applications
- Type of crops, type of animals, etc





Challenges of Geospatial Multimodal data

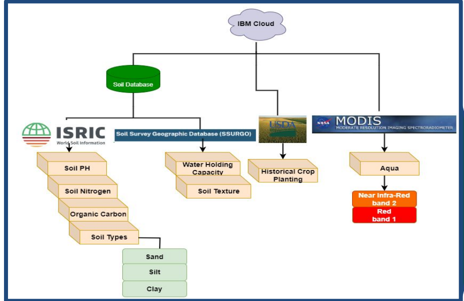


Geospatiotemporal Integration for Multimodal Datasets

The Challenge

Integrating Spatio Temporal Multimodal Big Data

IBM EIS dataset



4 datasets, different resolutions

Over 1 billion data points (Texas only) x 365 days x 2 bands

Featured Regions

Ohio, Texas and Florida (2019)

Pre-processing

Stack, Mask, Resample (Temporal)

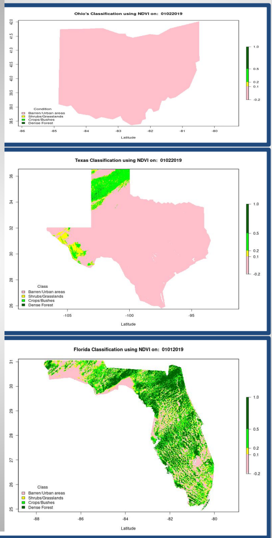
Land Use with NDVI

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

Ohio

Texas

Florida

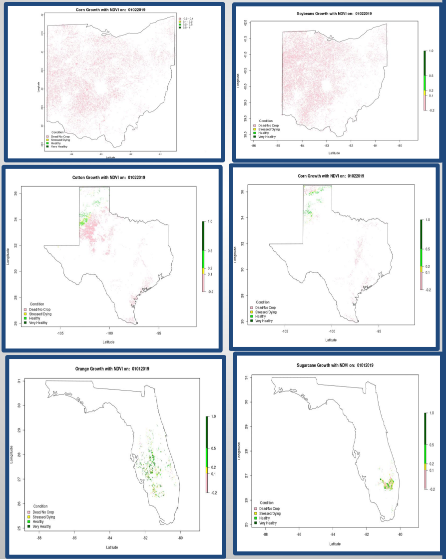


Daily Land Use Classification with Satellite Images (365 maps per state)

Data Integration

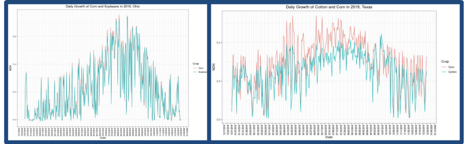
NDVI + Historical Crop Data (365 maps per crop)

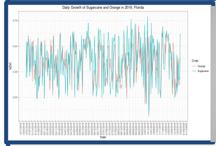
Crop Health and Growth



Major Two crops per state

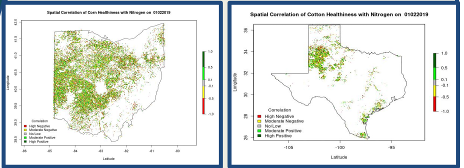
Average Daily Crop Growth

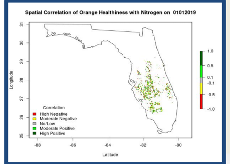




Ohio: Corn, Soybeans
Texas: Corn, Cotton
Florida: Orange, Sugarcane

Integration and Correlation



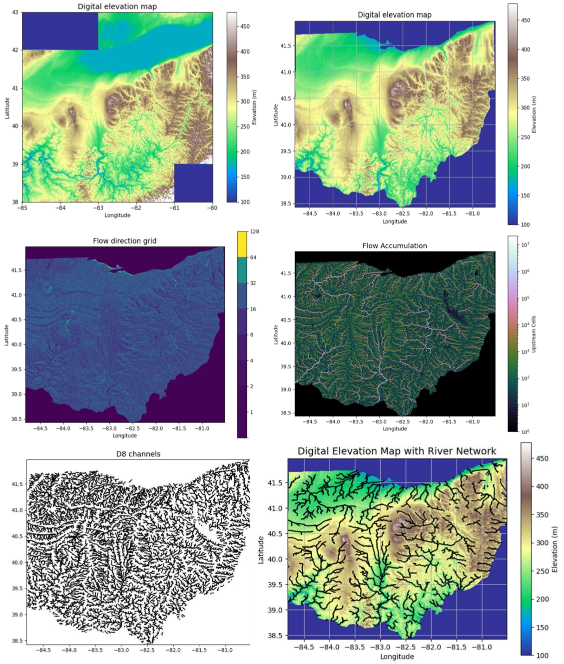


<u>With Nitrogen</u>
Ohio: Corn
Texas: Cotton
Florida: Orange

NDVI + Historical Crop + Soil data=Spatiotemporal correlation

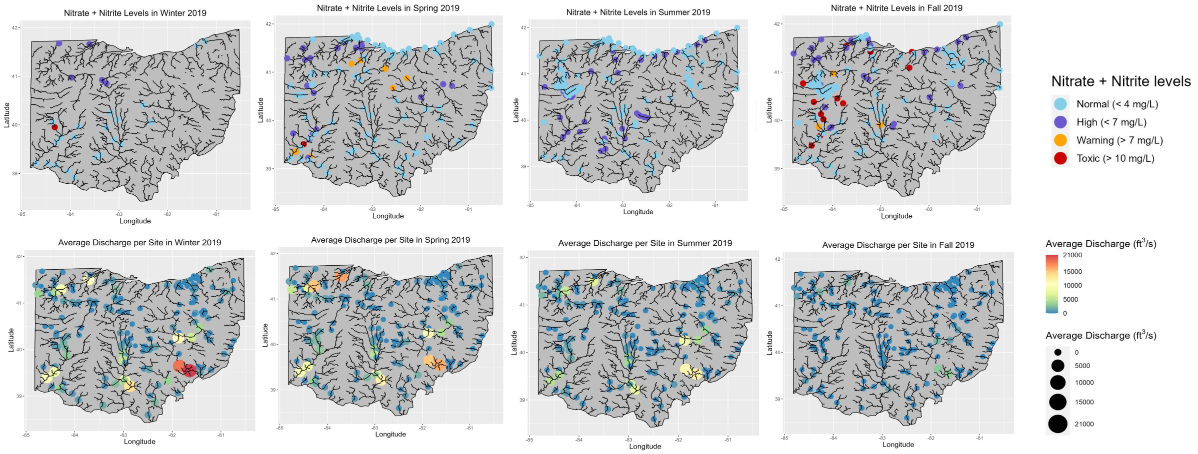
Analysis of Hydrologic Features

Extraction of of River Networks from Global Digital Elevation Models

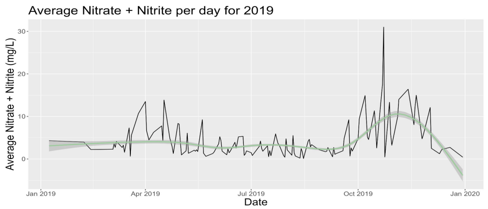


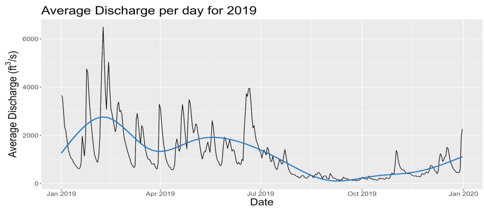
Datasets used: USGS, WQP and GDEMs

Behavior of Discharge and Nitrate + Nitrate Seasonally and Temporally



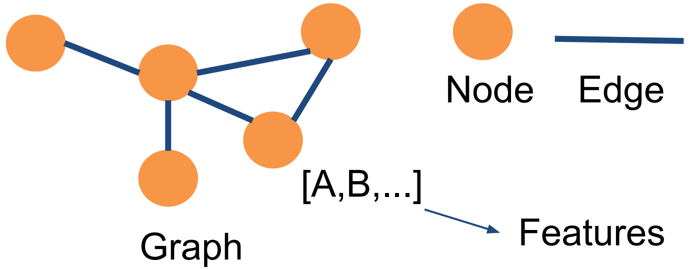
Behavior of Discharge and Nitrate + Nitrate Temporally



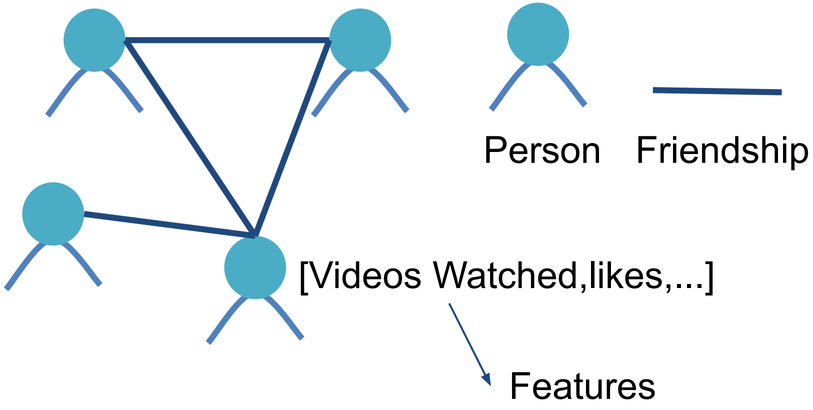


Overview of Graphical Neural Nets (GNNs)

What are GNNs



Friend Circle Example



Examples of GNN Applications

Stream Networks PV st-GNNs

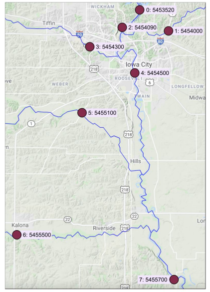
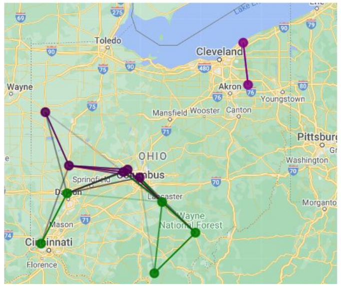
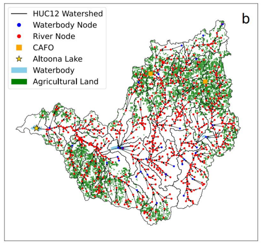


Figure 1. Visualization of study area and USGS sensor locations on Google Maps with their *id_in_graph: usgs_id*

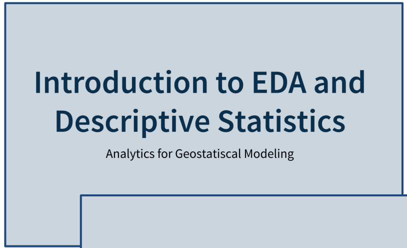


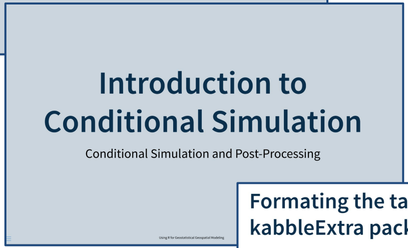
Purpose of GNNs in Watershed Modeling

- Predicting what nutrient concentrations would be at specific location
- Model structure of watersheds
- Could give us sources/quantification of nutrient contamination



Specialization in Geospatial Modeling



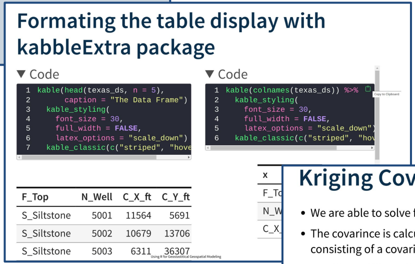


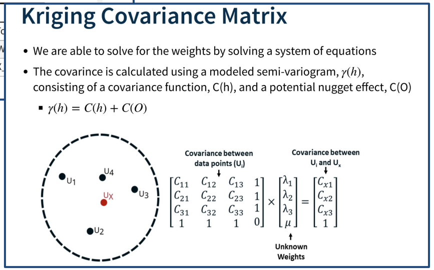
Geostatistical Geospatial Modeling

Course 1 : Principals

Course 2: Methods and Coding

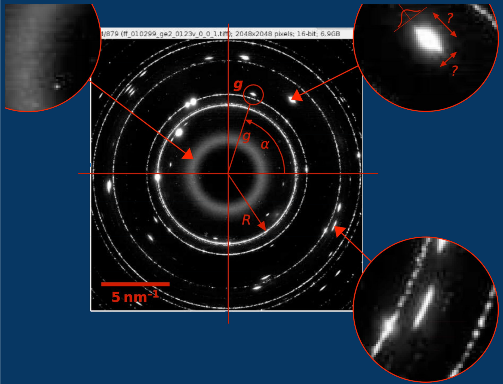
Course 3: Case Studies and Practical Examples





Automated Analysis Pipelines for 2D HEXRD

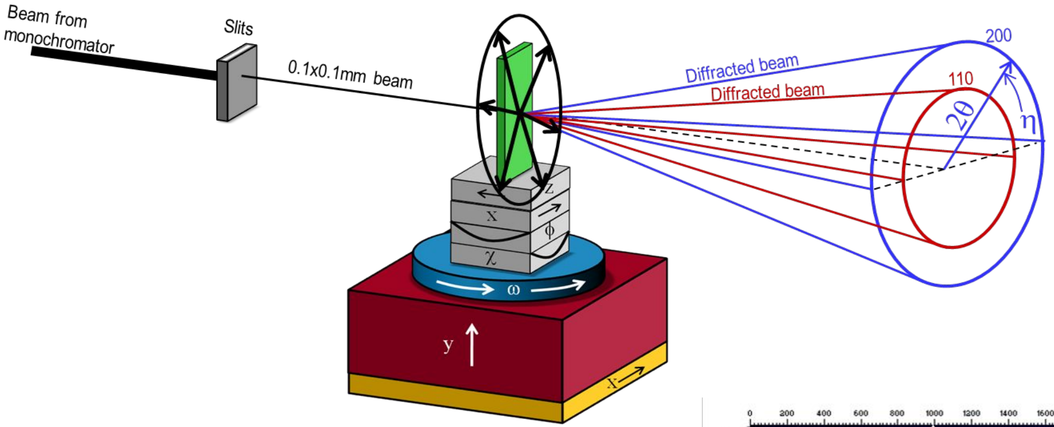
Diffraction Analysis Framework & "Scientist Ground Truth" Deep Learning Approach

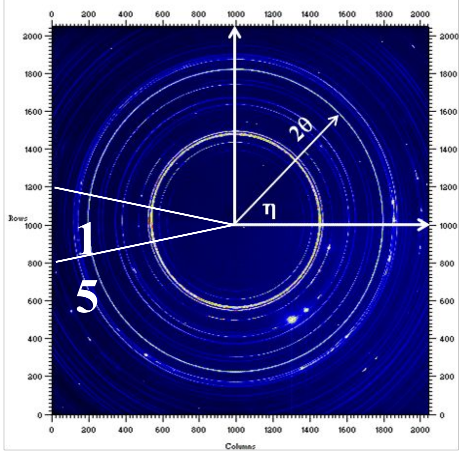


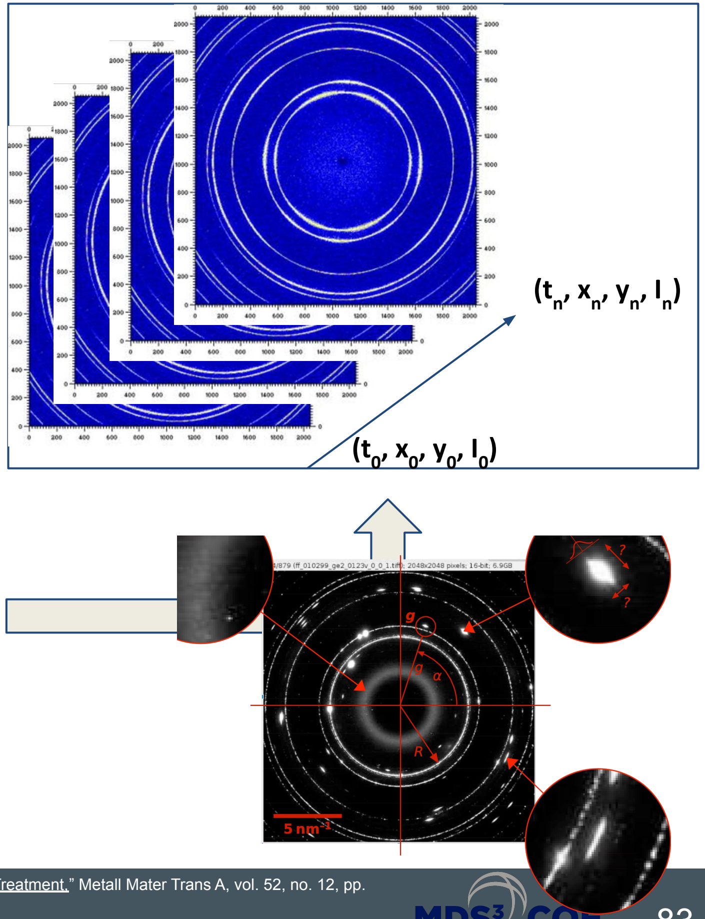
GS: Weiqi Yue[1], Redad Mehdi[1], Finley Holt[2]
UG: Gabriel Ponon[1], Ethan Fang[1]
Postdoc: Pawan K. Tripathi[2]
Faculty: Vipin Chaudhary[1], Frank Ernst[2], Matthew Willard[2], Bjourn Clausen[3]
Donald W. Brown[3], Daniel Savage[3], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA
3. Los Alamos National Laboratory, New Mexico, USA

2D-HEXRD Data Analysis Challenge: Extract All Information





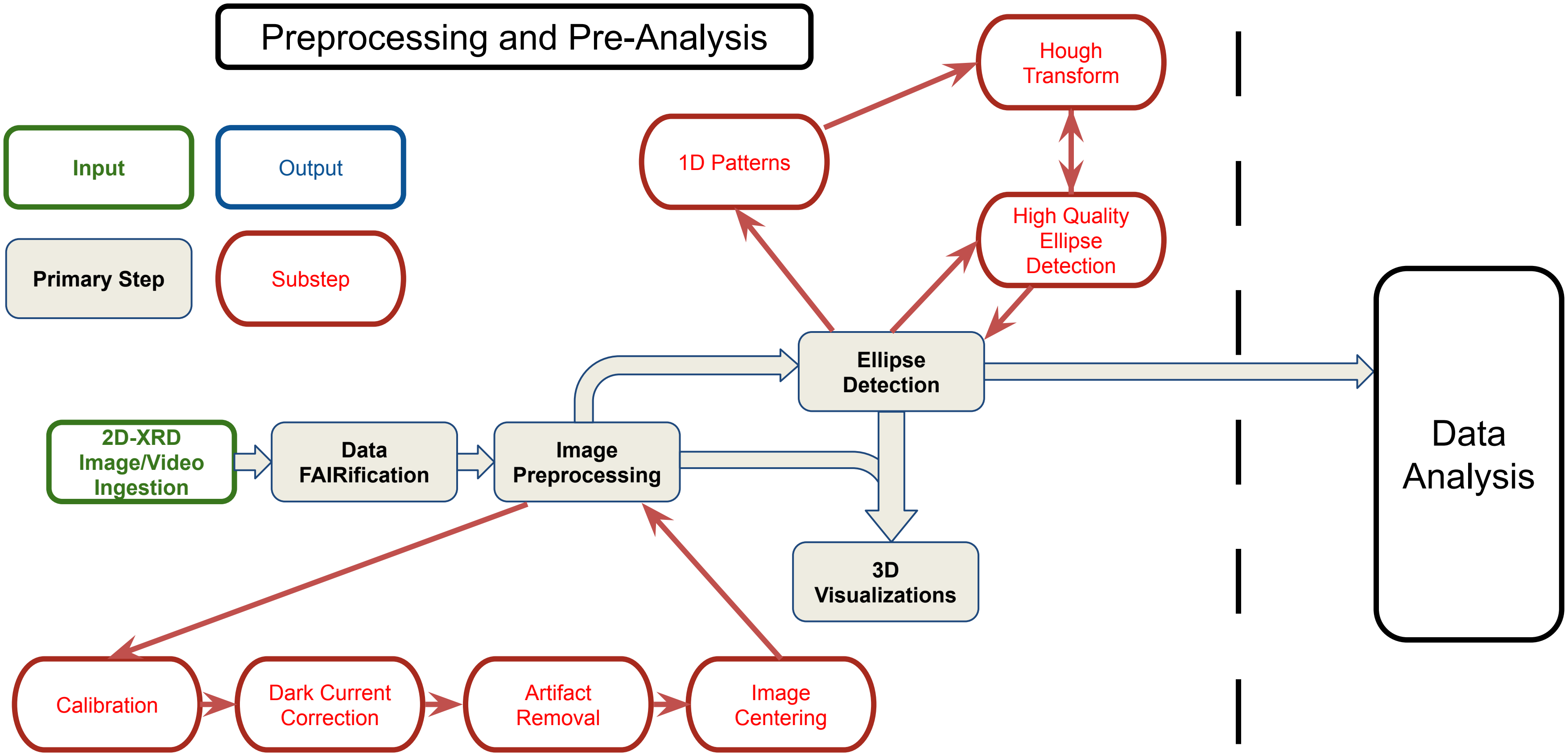


<u>Current 2D-HEXRD Datasets</u>
from Don Brown @ LANL

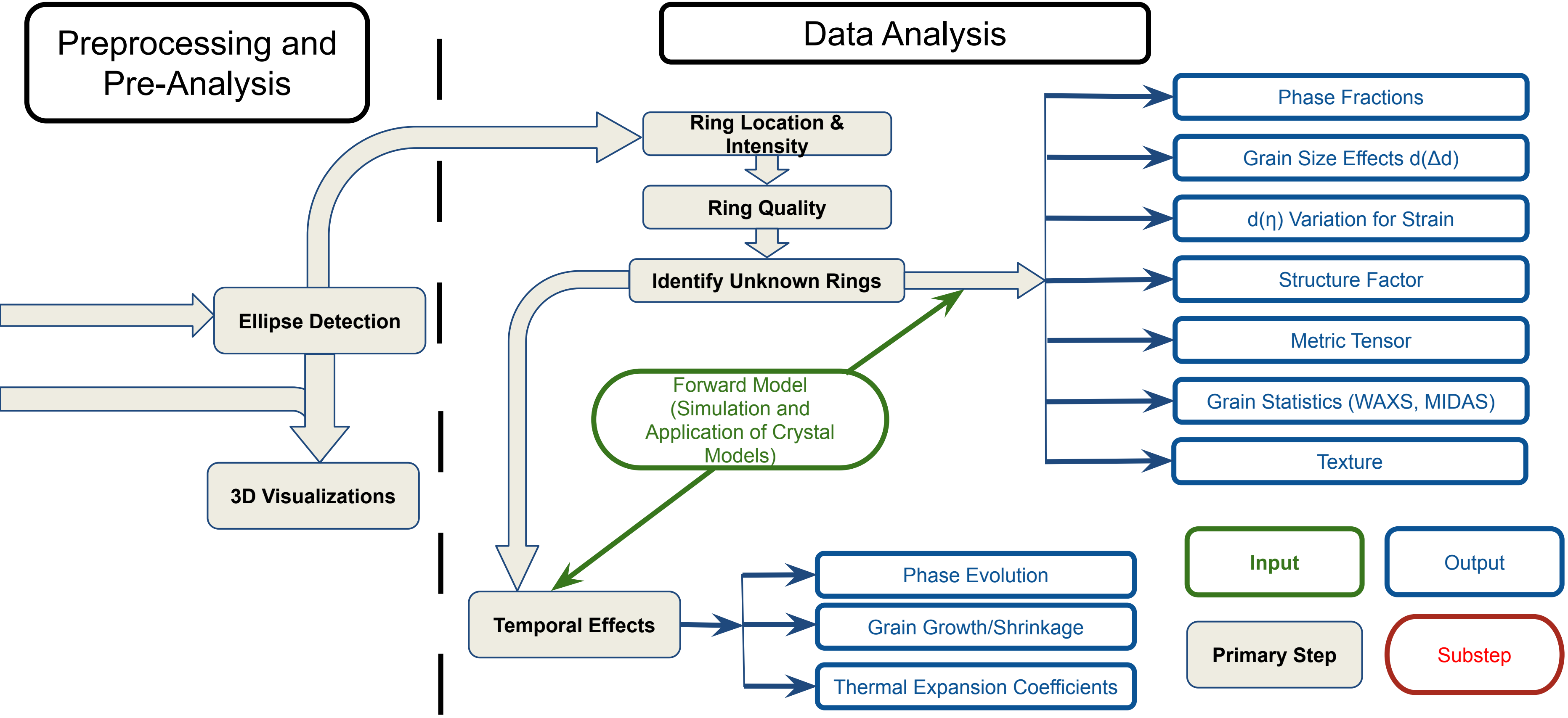
- ~ 22.1 TB
  - ~ 3.5 million 2D HEXRD images/movies
- Ti-6Al-4V:
  - In-situ heat treatment, texture, strain
- Stainless Steel
  - Wire arc Additive Manufacturing
- In-situ casting of Ti-Nb

D. W. Brown et al., "Evolution of the Microstructure of Laser Powder Bed Fusion Ti-6Al-4V During Post-Build Heat Treatment," Metall Mater Trans A, vol. 52, no. 12, pp. 5165–5181, Dec. 2021.

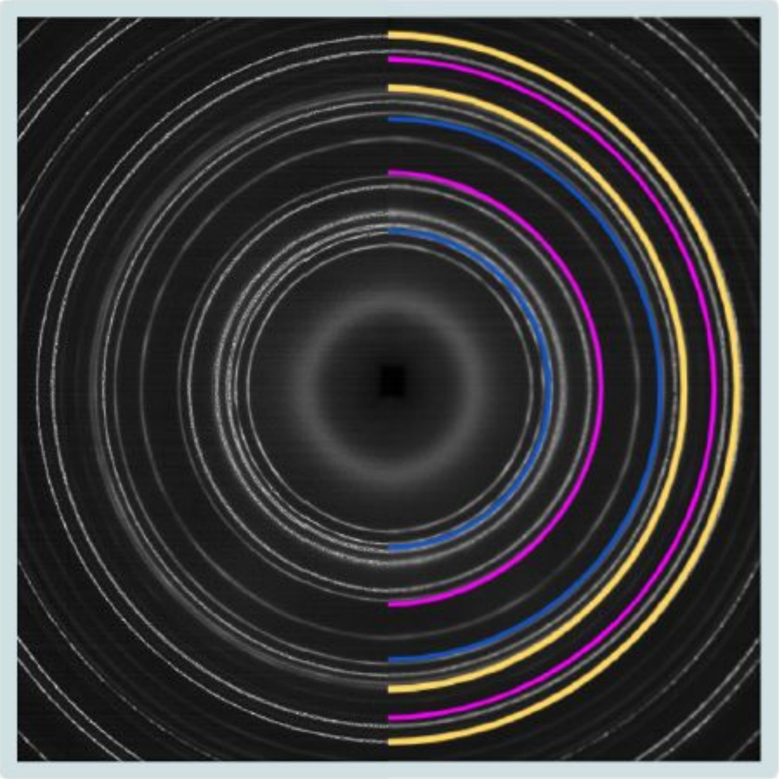
2D-HEXRD Analysis Pipeline: Preprocessing & Pre-Analysis



2D-HEXRD Analysis Pipeline Cont.



XRD Analysis Example: β-phase Phase Detection in Ti-64



Using deep learning framework

- Aim to identify β phase volume fraction

During in-situ heat-treatment of Ti-64 alloy

- At APS synchrotron 1ID XRD beamline

Ti-64 exhibit two phases

  - α-phase (HCP)
  - β-phase (BCC)
- Ti-64 sample contained
  - in stainless steel container

Phase Detection (from set of rings) =>

- The ring color indicates the crystalline phase
- The blue rings are about **α** phase,
- The pink rings are about **β** phase, and
- The yellow rings are about the **stainless steel**.

Automated XRD Phase Detection Analysis Pipeline

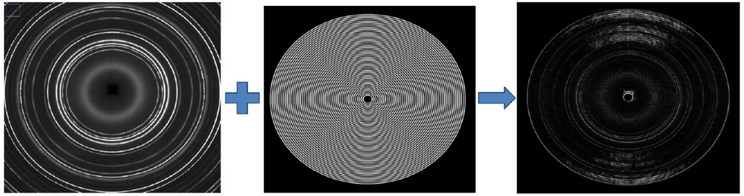
Image Datasets

- 10 datasets- 4 labelled & 6 Unlabelled.

	Sample Names	Number of Images	HT Temp.	Sample Description
4*Labeled Datasets	PB-HT1	1078	1043K	LPBF:LLNL
	PB-HT2	1102	1113K	LPBF:LLNL
	PB-HT3	863	1281K	LPBF:LLNL
	WR-HT2	1103	1113K	Wrought:Israel
6*Unlabeled Datasets	WR-HT1	1079	1043K	Wrought:Israel
	WR-HT3	1057	1281K	Wrought:Israel
	LENS-HT1	963	1043K	AM:PenState(LENS)
	EBM-HT1	1082	1043K	AM:Israel(EBM)
	EBM-HT2	1080	1113K	AM:Israel(EBM)
	EBM-HT3	880	1281K	AM:Israel(EBM)

Image Pre-processing

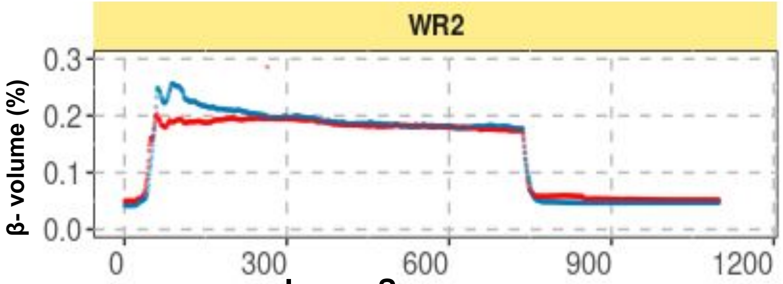
- Dark correction (subtraction)
- Image Centering
  - Multiple rings mask
  - Image registration



- Room Temperature Translation
  - Resize SS rings
  - Pixel-wise correlation

Deep Learning: using CNNs

- CNN model predicts **β phase volume fraction**
  - in external 2D XRD



Ti-6Al-4V Samples & Datasets

10 Ti-64 samples

- Processed
  - Using different methods
  - And at different facilities
- XRD movies acquired at CHESS

4 labeled datasets

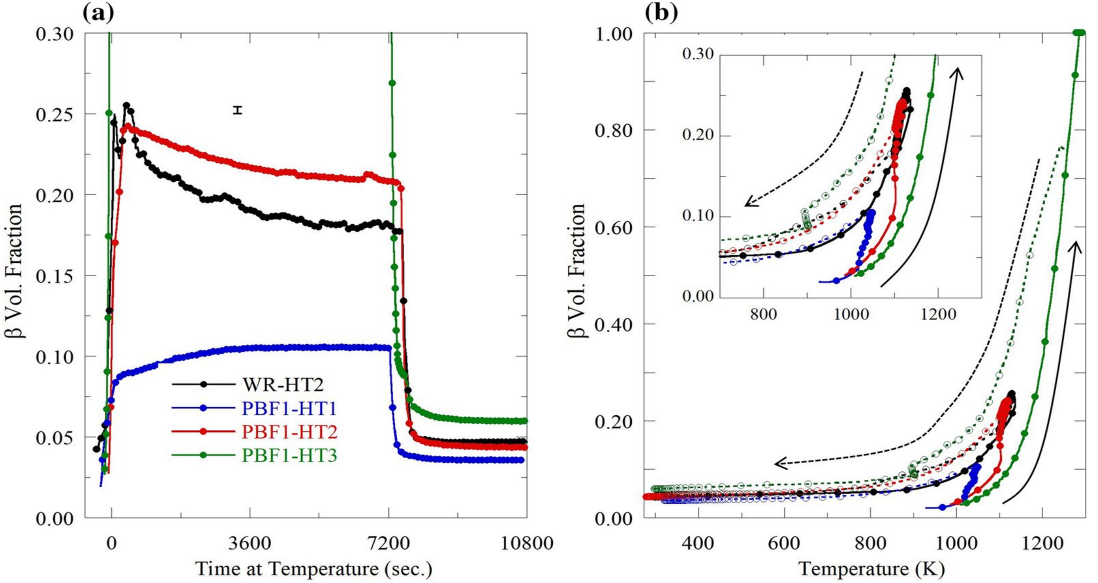
- β volume is known
  - From Don Brown publications
  - A form of "ground truth"

6 Unlabeled (un-analyzed) datasets

During the heat treatment, samples

- Heated from room temperature
- Held at maximum temp. for 2 hours
- Cooled back down to room temp.
- Samples HT1, HT2, and HT3
- Different max. heat treatment temp.

	Sample Names	Number of Images	HT Temp.	Sample Description
4*Labeled Datasets	PB-HT1	1078	1043K	LPBF:LLNL
	PB-HT2	1102	1113K	LPBF:LLNL
	PB-HT3	863	1281K	LPBF:LLNL
	WR-HT2	1103	1113K	Wrought:Israel
6*Unlabeled Datasets	WR-HT1	1079	1043K	Wrought:Israel
	WR-HT3	1057	1281K	Wrought:Israel
	LENS-HT1	963	1043K	AM:PenState(LENS)
	EBM-HT1	1082	1043K	AM:Israel(EBM)
	EBM-HT2	1080	1113K	AM:Israel(EBM)
	EBM-HT3	880	1281K	AM:Israel(EBM)



Deep Learning Approach & Hyperparameter Tuning

"Neural Network Architecture Search"[1]

- Critical topic in deep learning performance
  - Major topic in Data Science today
  - Which is the best Neural Network architecture to learn from a specific dataset?

Trained 168 CNN architectures, with different hyperparameters

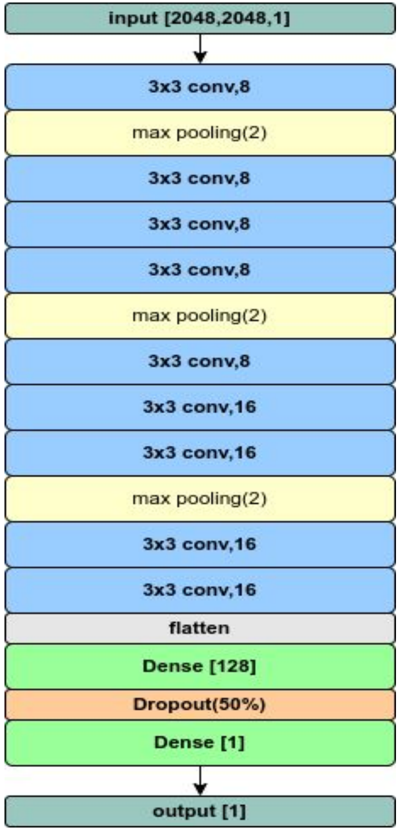
- Tuning CNN models' hyper-parameters
  - Using our Distributed & HPC system CRADLE[2]
  - And SDLEfleets to train on GPUs in HPC Compute Nodes
    - Not using the Nvidia AISC

Models used for 2D HEXRD learning:

- Regression Convolutional Neural Networks (CNN)

Training & Testing Details

- Trained on 2D XRD datasets from three different heat treatment runs
  - Total 2451 XRD diffractogram images
  - i.e. PB1, PB2, PB3,
  - Train: 81% (1955 images), Validation 19% (451 images)
- Utilizing the trained CNN model to predict on a test dataset, WR-HT2
  - 1103 diffractogram images



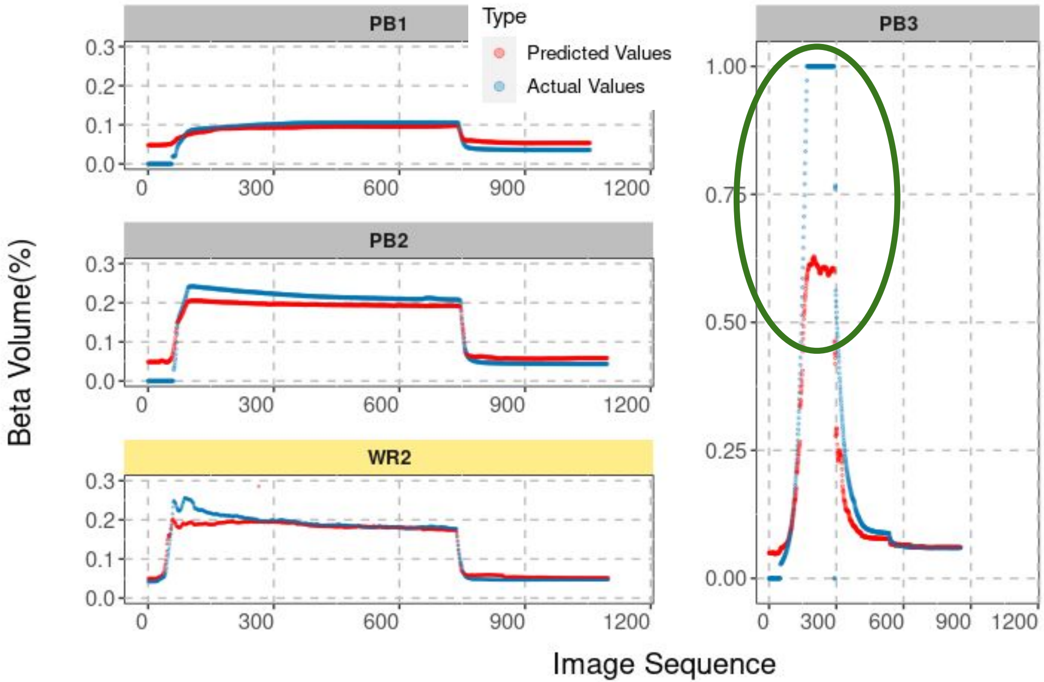
Architecture of CNN Model #80

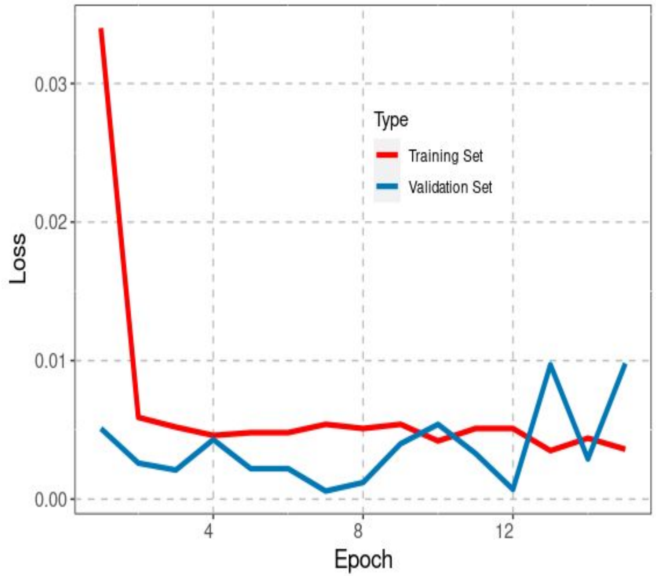
[1] C. Ying, et al., "NAS-Bench-101: Towards Reproducible Neural Architecture Search," in Proc.36th Intl. Conf. Mach. Learning, PMLR, May 2019, pp. 7105–7114. Available: [Accessed: Oct. 15, 2023]
[2] A. Khalilnejad, A. M. Karimi, S. Kamath, R. Haddadian, R. H. French, and A. R. Abramson, "Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data," PLOS ONE, vol. 15, no. 12, p. e0240461, Dec. 2020.

General Analysis of CNN #80's Performance

While CNN #80 has a low MSE on WR-HT2 (0.02%)

- But it performs poorly on the training set
- Particularly on PB3 dataset.





Takeaway:

No 1 "best" CNN Architecture

For 2D XRD Analysis

Datasets & Models have biases

Robust Models on Testing and WR-HT2 Datasets

Regr. CNNs Models that perform well

- On both of the train dataset
- And WR-HT2 datasets

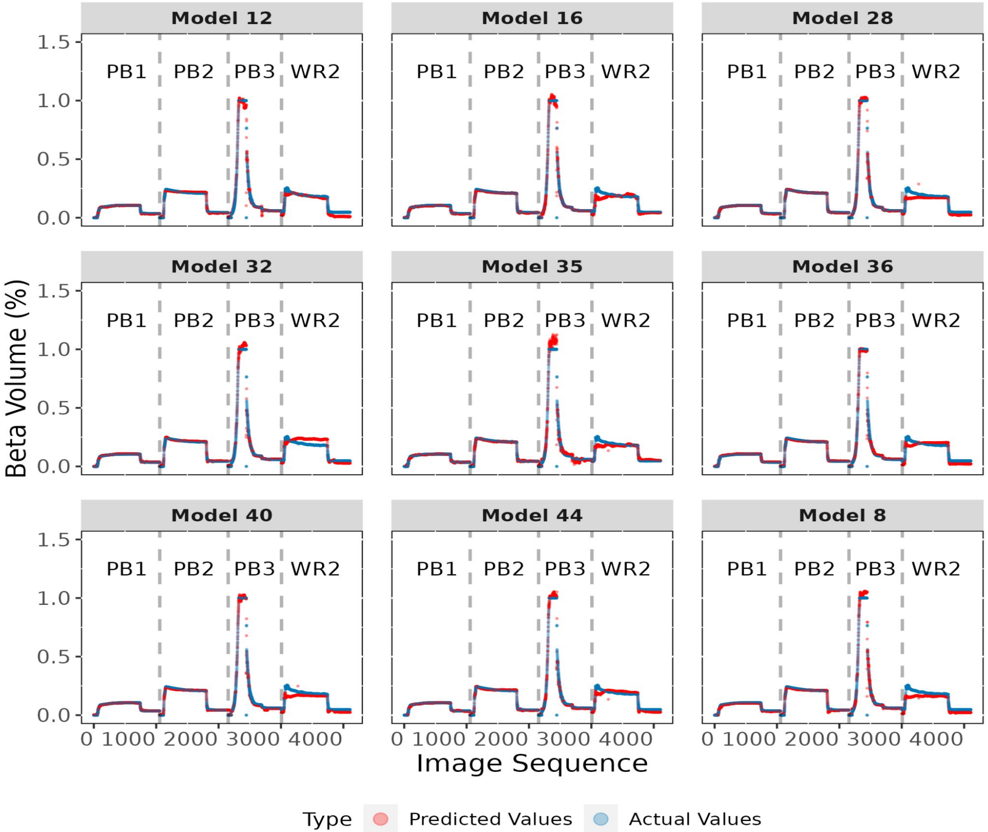
Selected top 15 models on the WR-HT2 datase

- And the top 15 models on the train dataset.

Nine of these 15 best performing CNN models

<u>Were the same CNN Architectures</u>

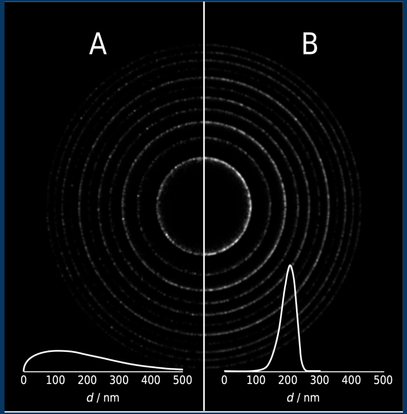
<u>Indicating their robustness.</u>



Deep Learning for 2D HEXRD

Use a Kinematic Diffraction Forward Model

For Regression CNN Training



GS: Weiqi Yue[1], Redad Mehdi[1] , Finley Holt[2]
UG: Gabriel Ponon[1], Ethan Fang[1]
Postdoc: Pawan K. Tripathi[2]
Faculty: Vipin Chaudhary[1], Frank Ernst[2], Matthew Willard[2], Donald W. Brown[3] , Daniel Savage[3], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA
3. Los Alamos National Laboratory, New Mexico, USA

A Forward Model to Simulate X-Ray Diffractograms

Challenges training neural networks to predict {micro}structure from diffractograms:

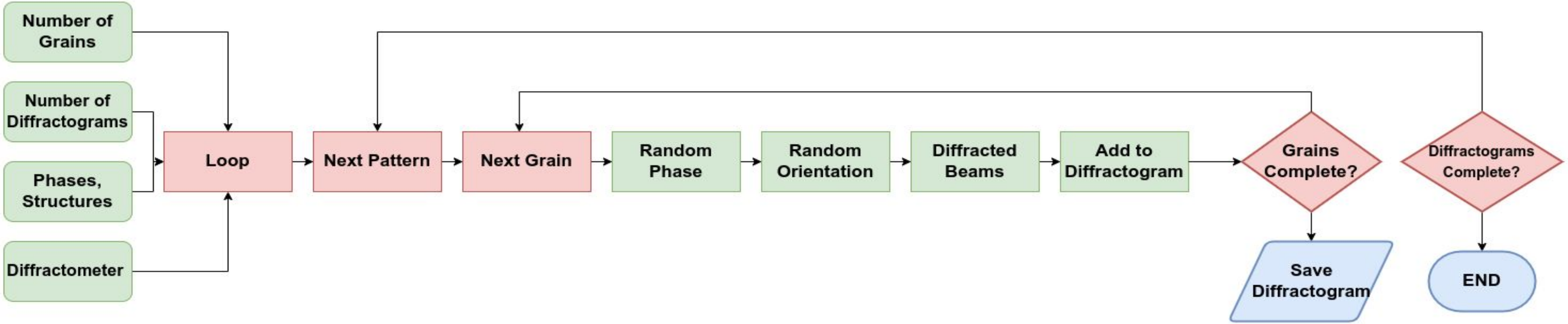
- No large experimental datasets for which microstructure is known
- Nor are features varied over a large range
  - That are independent from other parameters
- Even “human labelled data” is not the absolute “ground-truth”

Simulation XRD data can incorporate all parameters of the

- X-ray diffractometer and the sample’s crystal structure and grain microstructure

Provides granular control over parameters we want the neural networks to learn

- Reduces need for “scientist ground truth” experimental data
- Needed for NN training, Which may not exist



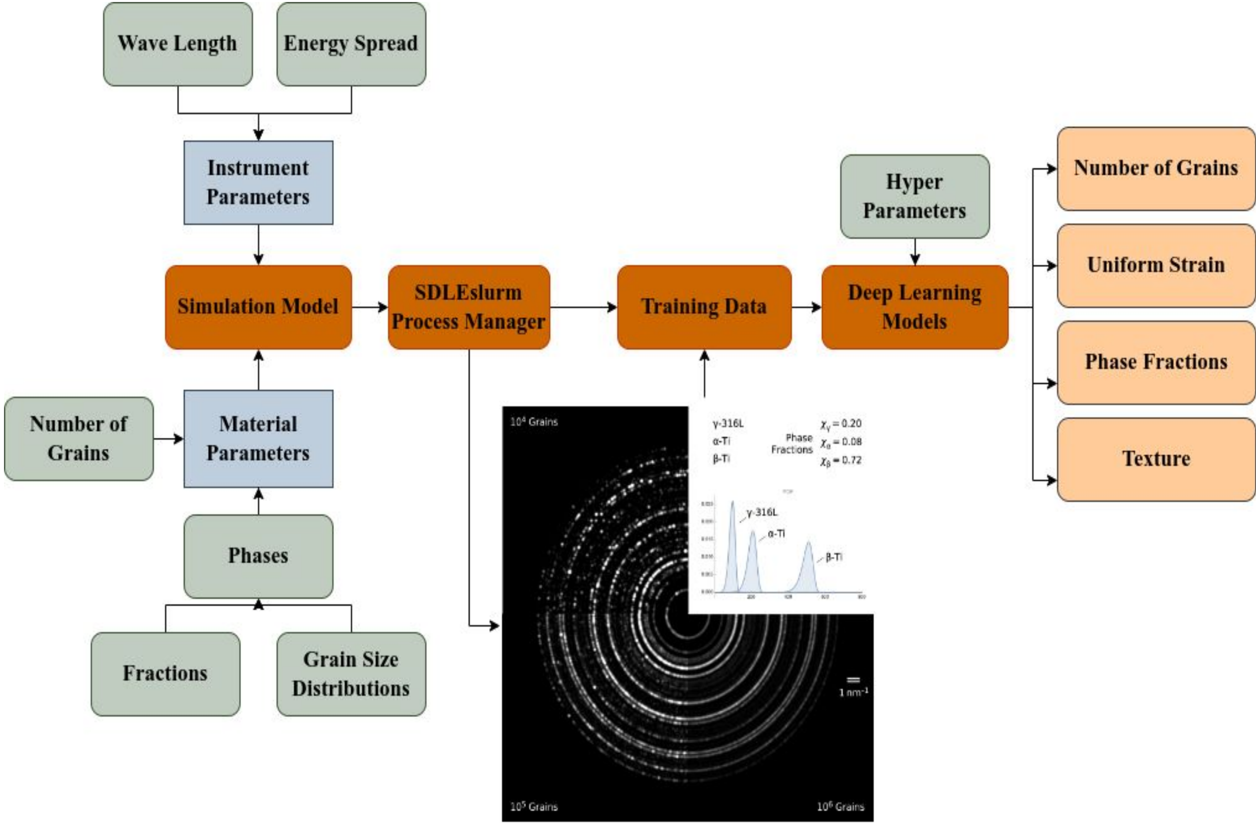
Kinematic Diffraction Forward Model Pipeline for 2D HEXRD

Goal

- Retrieve info from diffractograms
- Replace human experimentalists
- By Neural Networks (NN)
- Quantify *microstructure*

Approach

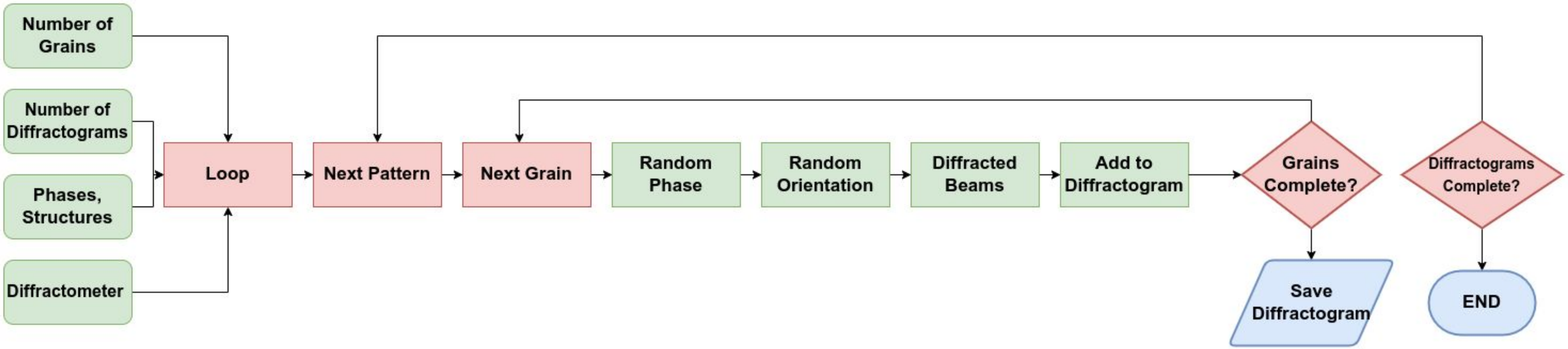
- Train NNs to learn information
- Need: Training data
- *Ab-initio simulations* for data
- NN training
  - Varying hyperparameters
- Simulations verified by
  - *Labeled* experimental data
- The trained NN is then
  - Applied to experimental data



Kinematic Diffraction Simulation Package

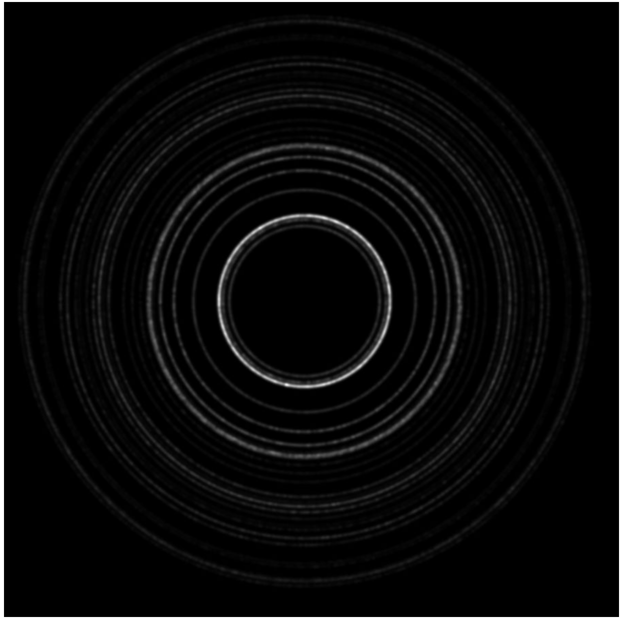
Kinematic Diffraction Simulation: (Diff-Sim)

- Mathematica paclet
  - Written in Wolfram Language
- X-ray Diffractometer parameters
  - Wavelength of the primary beam
  - Beam divergence
- Sample parameters
  - Any crystal structure
  - Any number of phases
  - Grain size distribution per phase
  - Any texture per phase

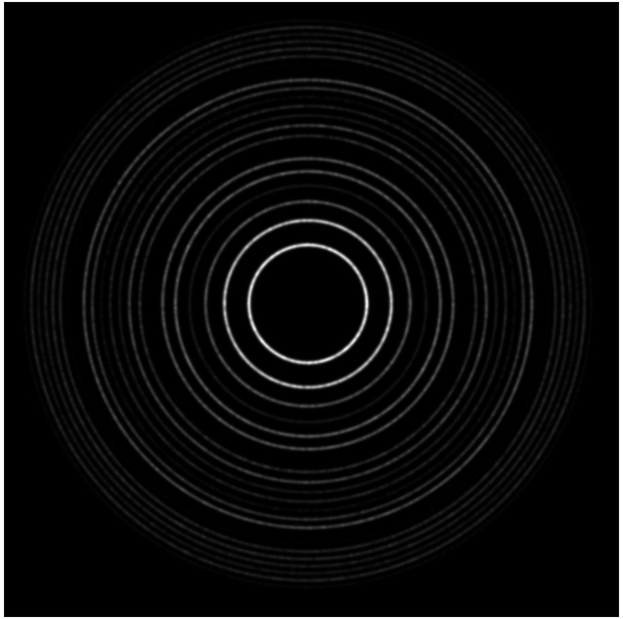


Simulated Diffractograms of Ti-6Al-4V: 0% & 100% β phase

For 100,000 grains in irradiated volume



100%α- 0%β Ti



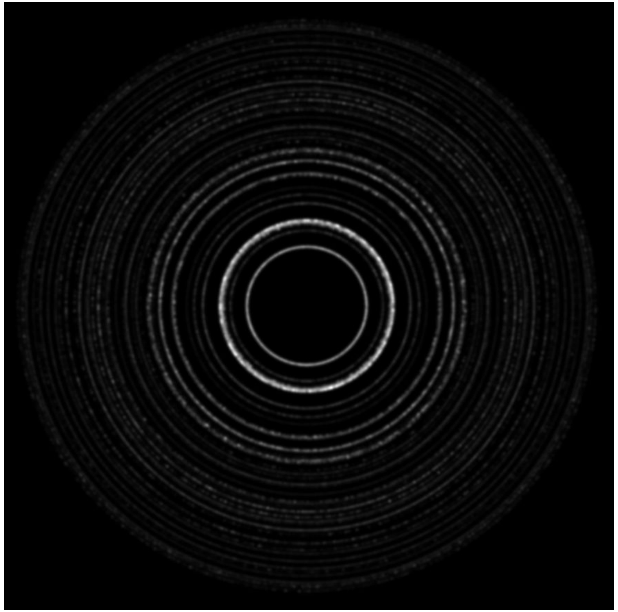
0%α- 100%β Ti

The intensity of the rings associated with the β-phase

- Increases as it’s mole fraction increases

Simulated Diffractograms of Ti-6Al-4V: 80% & 100% β phase

For 100,000 grains in irradiated volume



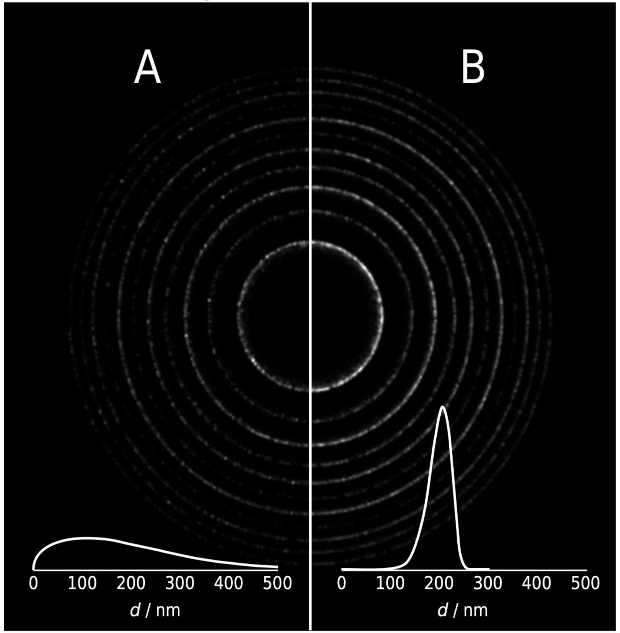
20%α- 80%β Ti

As we get to pure β-Ti,

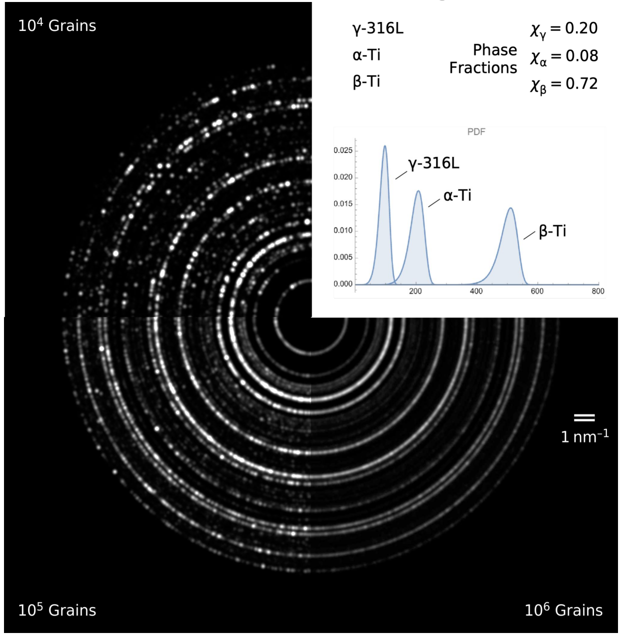
- The rings associated with the α-phase disappear and
- The entire intensity is from the β- phase

Effect of Various Parameters on The Simulated Diffractograms

Effect of grain-size distribution



Effect of number of grains

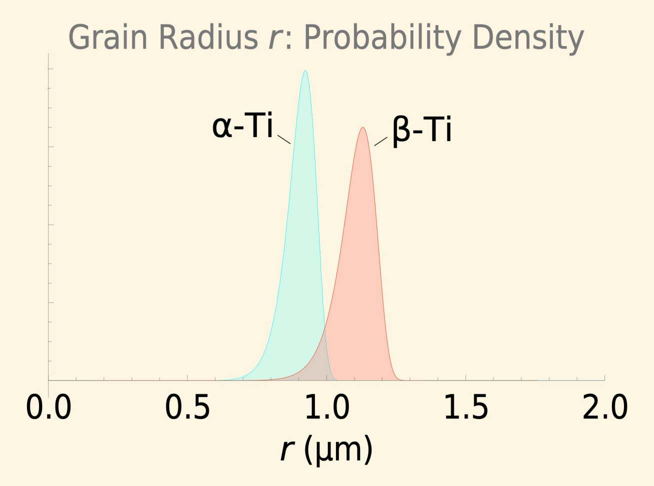


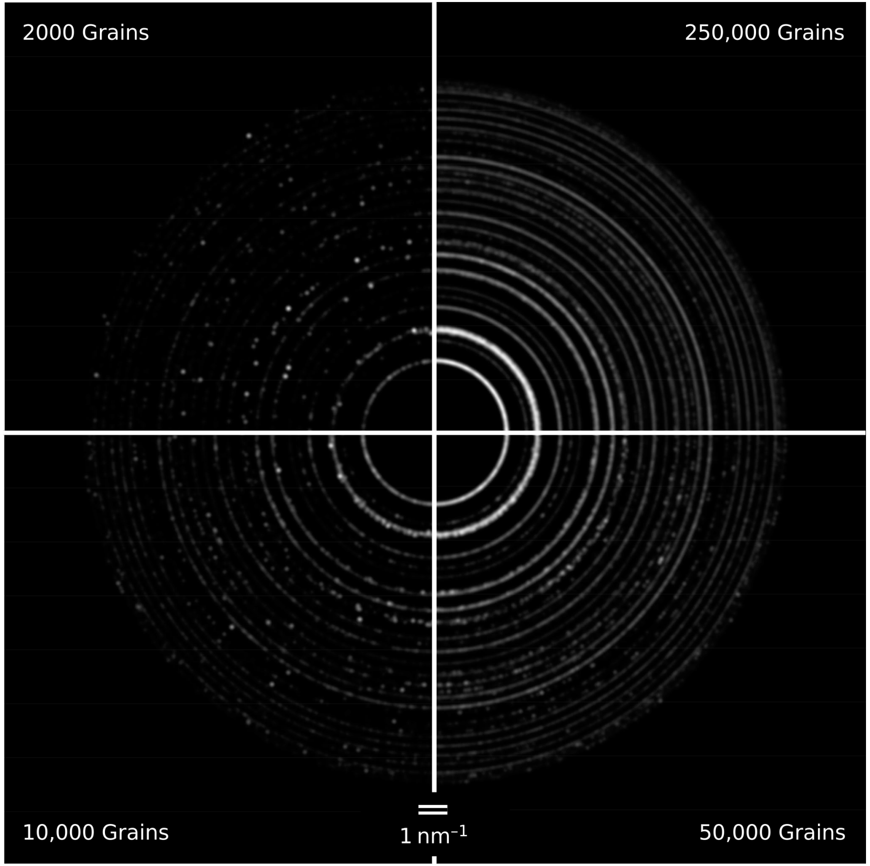
Ring continuity (spotiness) depends on

- The grain-size distribution
- The number of grains

Effect of Microstructure Parameters on Simulated Diffractograms

Phase Fractions $\chi_\alpha = 0.10$ $\chi_\beta = 0.90$

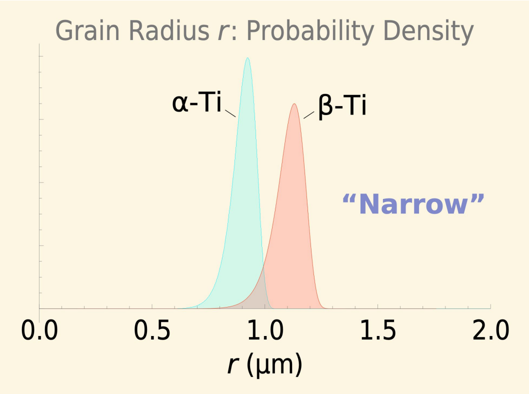


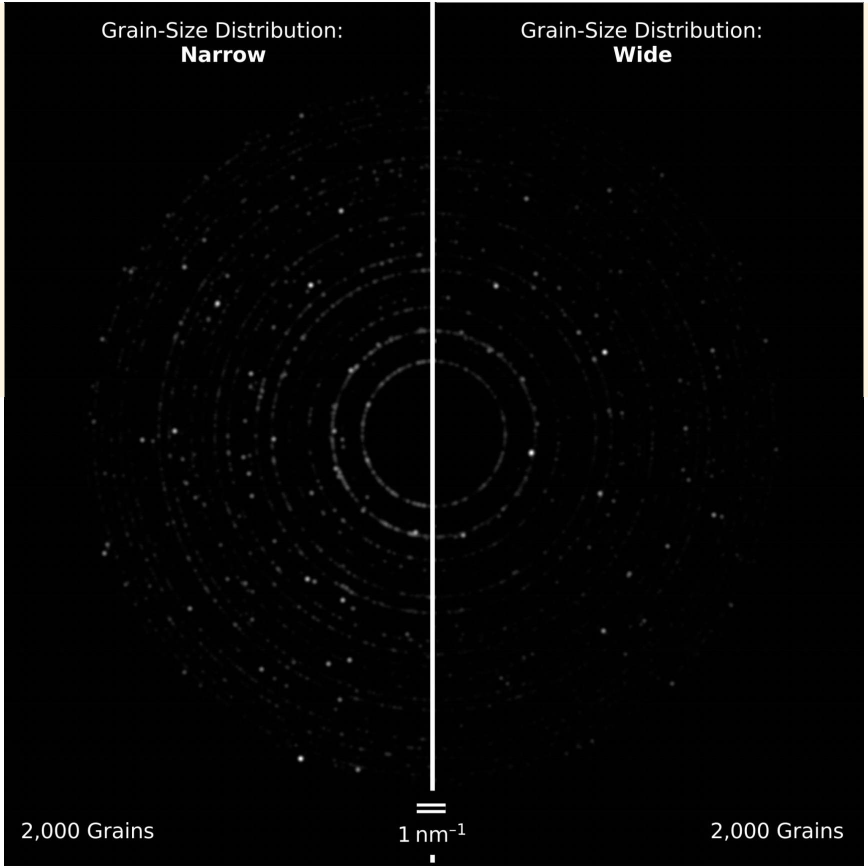


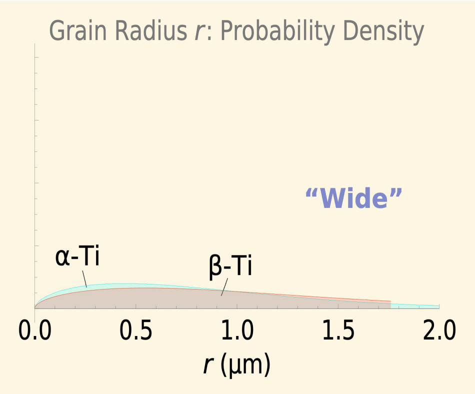
Ring continuity (spotiness) depends on:

- Number of grains.

Effect of Microstructure Parameters on Simulated Diffractograms







Ring continuity (spotiness) depends on

- Grain-size distribution.

Phase Fractions $\chi_\alpha = 0.10$ $\chi_\beta = 0.90$

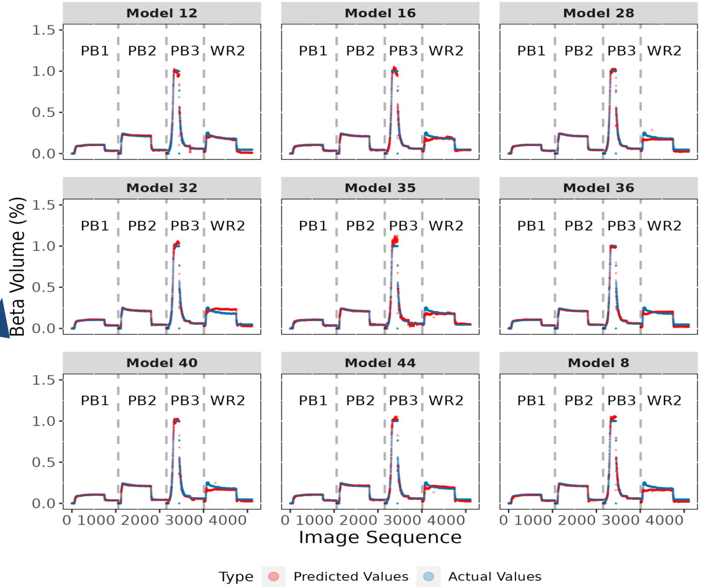
β-phase Detection: Train CNN Models on Forward Model Simulation Data

Regression CNN Training & Testing Details

- **~5500 Diffractograms for Training**
  - An equal split of ~2250 each
  - The diffractograms are either
    - Pure α-Ti (0% β), or
    - Pure β-Ti (100% β)

- **~1000 Diffractograms for Testing**
  - Contains data at every 5% β-phase fraction
  - So, around 50 diffractograms at every 5% β-Ti

- **Construct Models With an <u>Identical Architecture</u> to the**
  - Top-performing model
  - From our prior experimental datasets



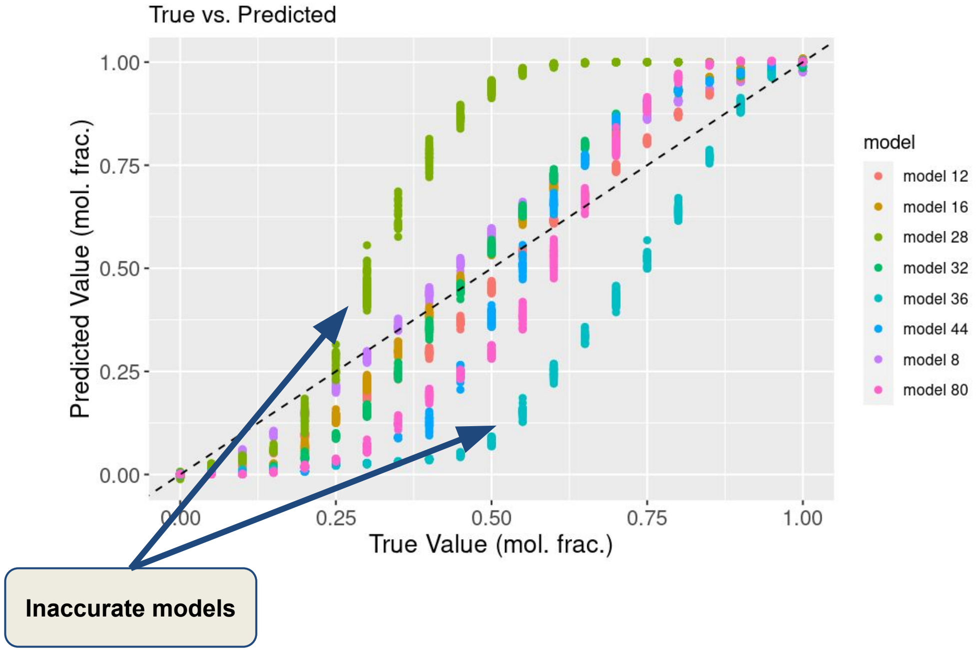
Performance of models with different architectures on testing set

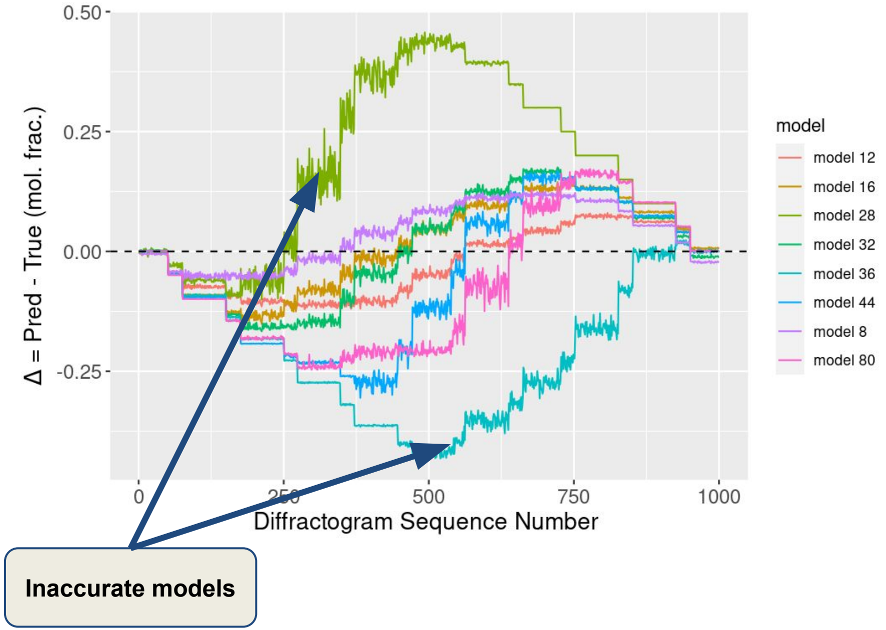
Diffractogram in testing set is

- Sorted from 0% to 100% β mole fraction

Visualize models' performance based on:

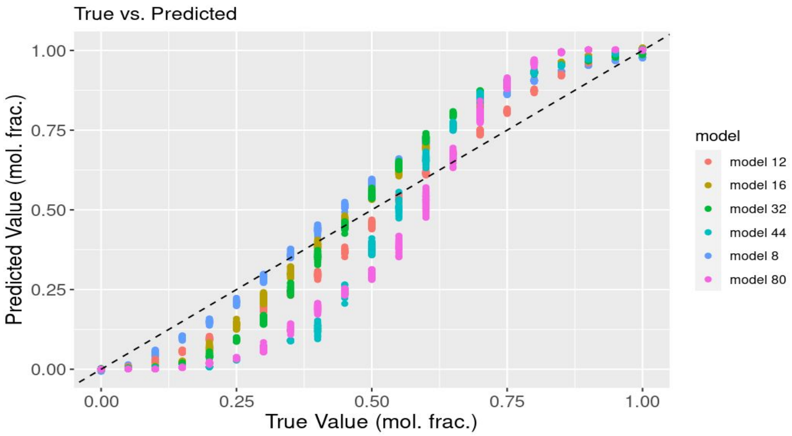
- Difference between predicted and true values
- Predicted values vs. true values

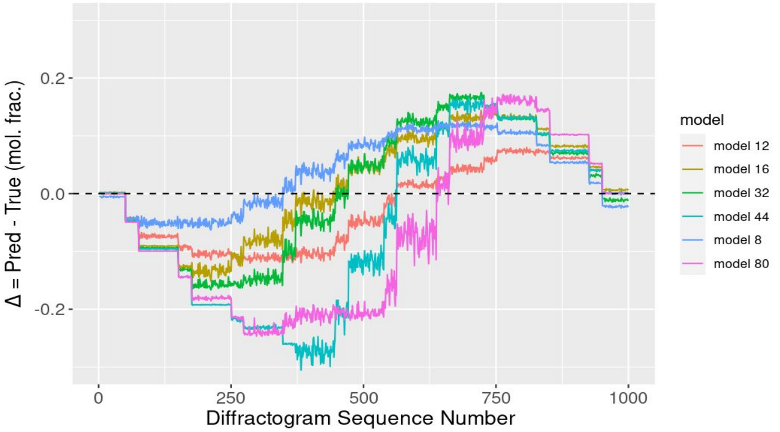




Neural Network Architectures & Performance of the Trained CNN Models

Model Index	Convolutional layers	Dense layers	Parameter number	Metric (MSE)
8	4	128-64	260 M	0.00527
12	4	128-64-32	260 M	0.00522
16	5	128	126 M	0.00833
32	6	128	520 M	0.012
44	7	128-64	129 M	0.0231
80	9	128	125 M	0.0230





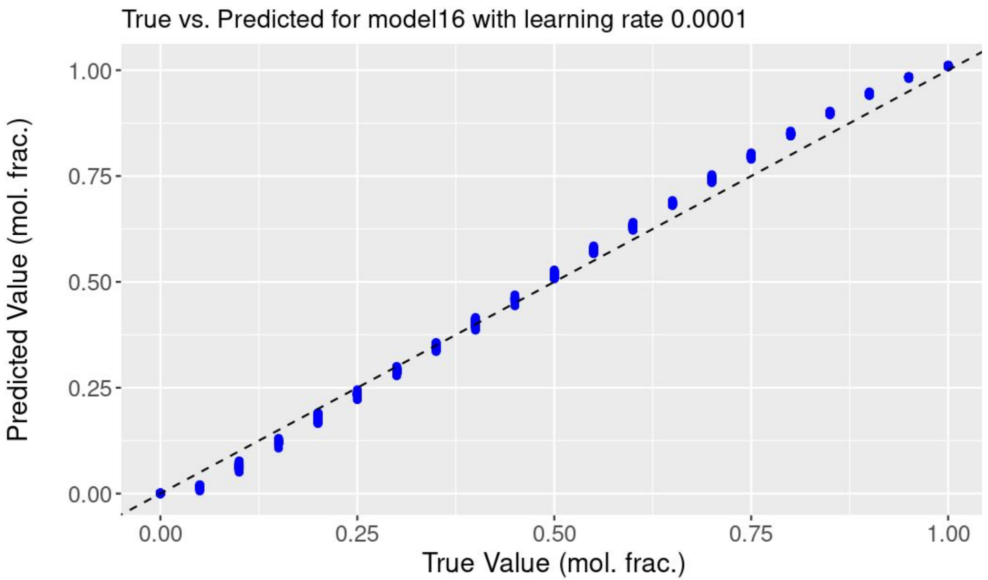
Visualization for architectures of these CNN models

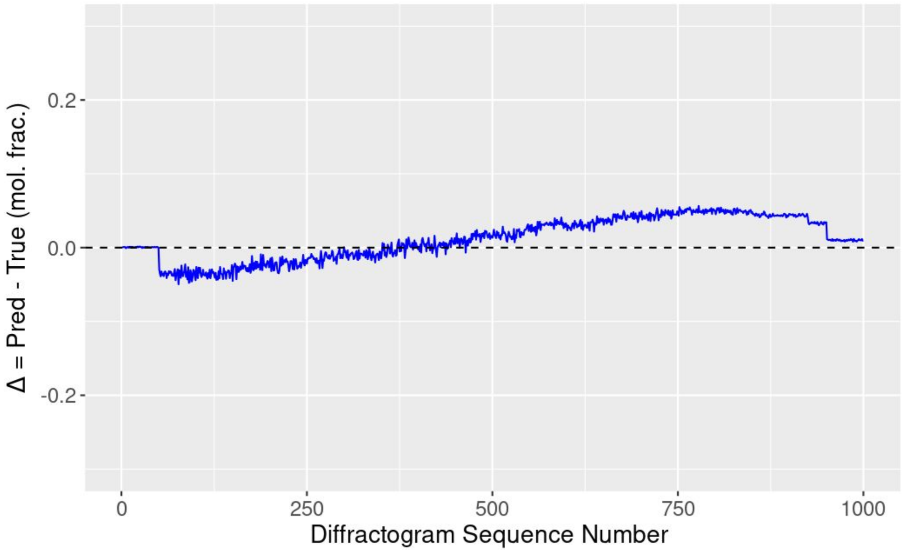
- Ignored two low performance models (model 28 and 44)

‘Best’ Regression CNN model so far

After fine-tuning the learning rates,

- Determined that a learning rate of 1 x 10^{-4} resulted in
- <u>Model 16 achieving its best performance</u>





Further Hyperparameters Tuning

Model Index	MSE	Batch size	LR
16	0.0345	16	0.00005
16	0.00094	16	0.0001
16	0.00833	16	0.0005
16	0.0833	16	0.001
16	0.0342	16	0.005

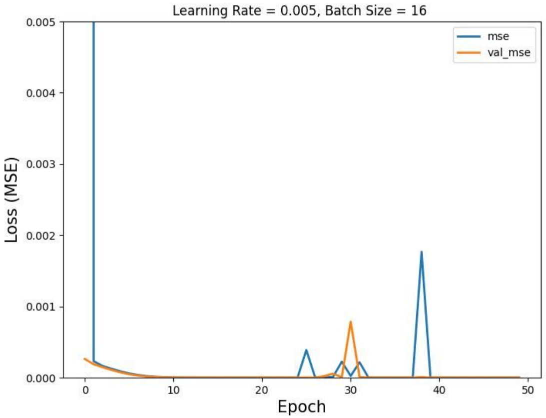
Even for the same Neural Network architecture models,

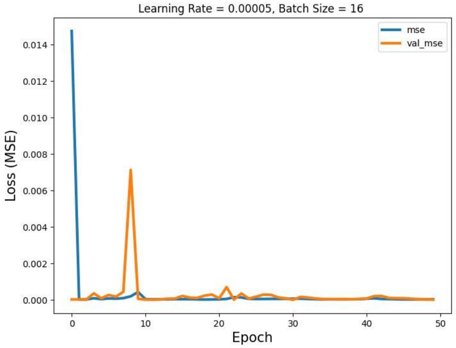
- Different Hyperparameter settings during training
  - Learning rate
  - Batch size
- Affect the model's learning and final performance

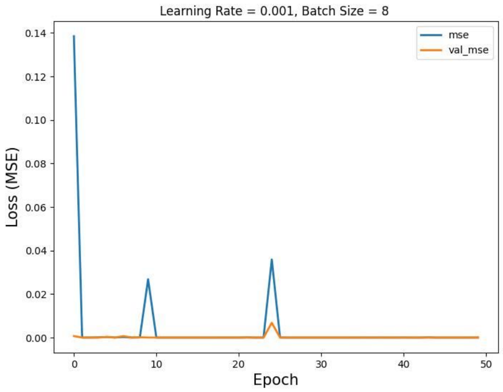
Numerous hyperparameters can be varied during the training.

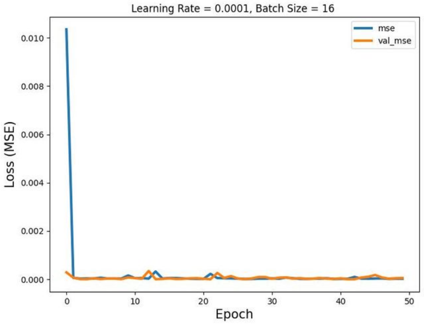
- It's always a tradeoff between compute resource and models' performance

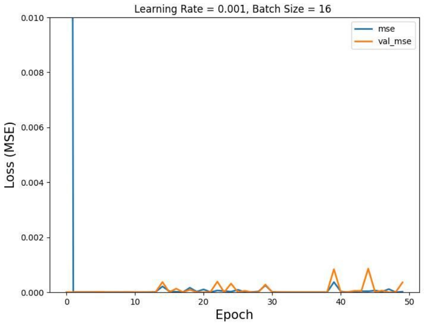
Training history curve for different learning rates

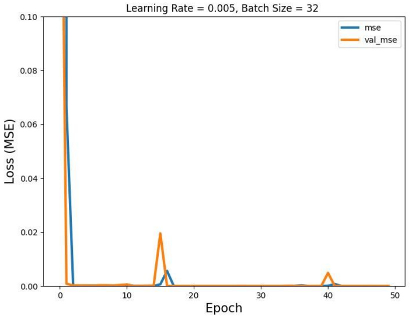












HEXRD Analysis Takeaway

Comprehensive deep learning 2D HEXRD Diffractogram analysis pipeline

- For automated phase fraction detection
- Complex feature analysis in 2D XRD
- Can handle terabyte-level XRD datasets

Forward model simulates kinematic diffraction data

- Details for microstructure for materials (ground truth)

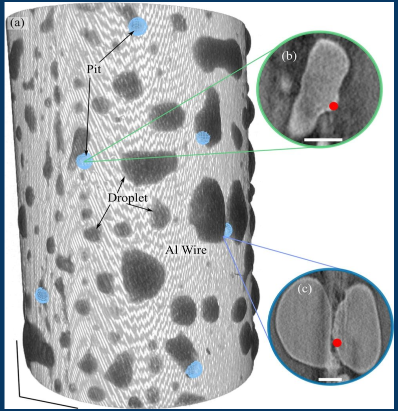
Hyperparameter tuning pipeline for Deep Learning Models

- Avoid invalid models
- Achieves high accuracy on external datasets
- Generates robust model architectures

Not all NN Models learn the same information from a particular dataset!

Automated Pipeline for X-ray Computed Tomography

Observing Pitting Corrosion of Aluminum Wires



GS: Tommy Ciardi[1], Maliesha Sumudmalie[2]
Postdoc: Pawan K. Tripathi[2]
Faculty: Alp Sehirlioglu[2], Philip Noell[3] , Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA
3. Sandia National Laboratory, New Mexico, USA

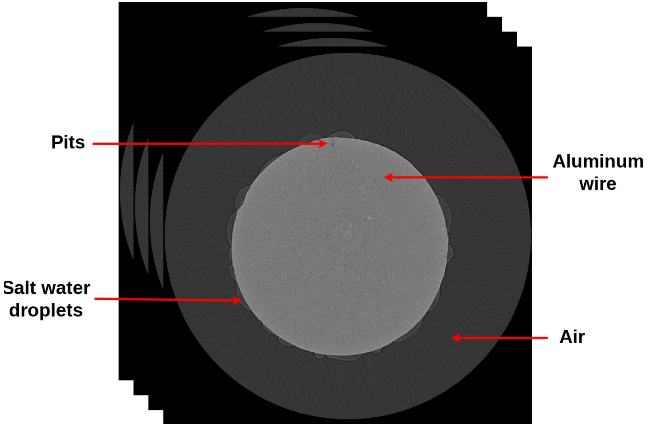
Al Wire Sample & XCT Scan Details

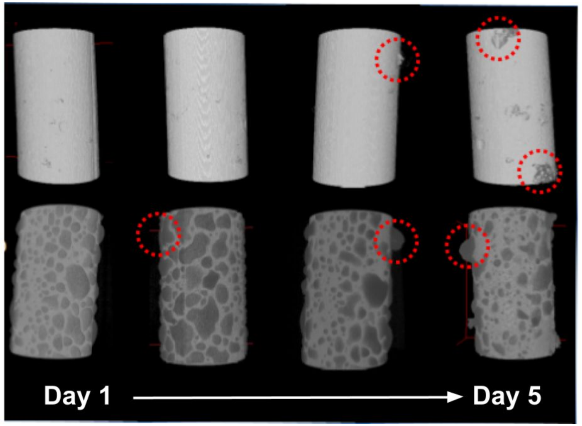
In-situ XCT observations of Aluminum Wire

- 0.813 mm diameter 1100 Al wire
  - Commercial-purity Al
- NaCl picoliter-sized droplets
- Exposed to 98% RH at ~25° C
  - for 122.33 hours
- 1.25 mm length of the wire imaged by XCT
  - Over the course of the exposure
  - 996 slices
- Voxel size of 1.25 μm
- Spatial resolution of 15.6 μm^3
  - (2 × 2 × 2 voxels)

A total of 88 XCT datasets were collected

- Over 122.33 h (~5 days)
- At a 83 min temporal resolution
- Total number of images = 996 x 88
  - **= 87,648 (~100GB)**





Characterizing Pitting Corrosion in Al-1100 Bond-Wires

Features of interest

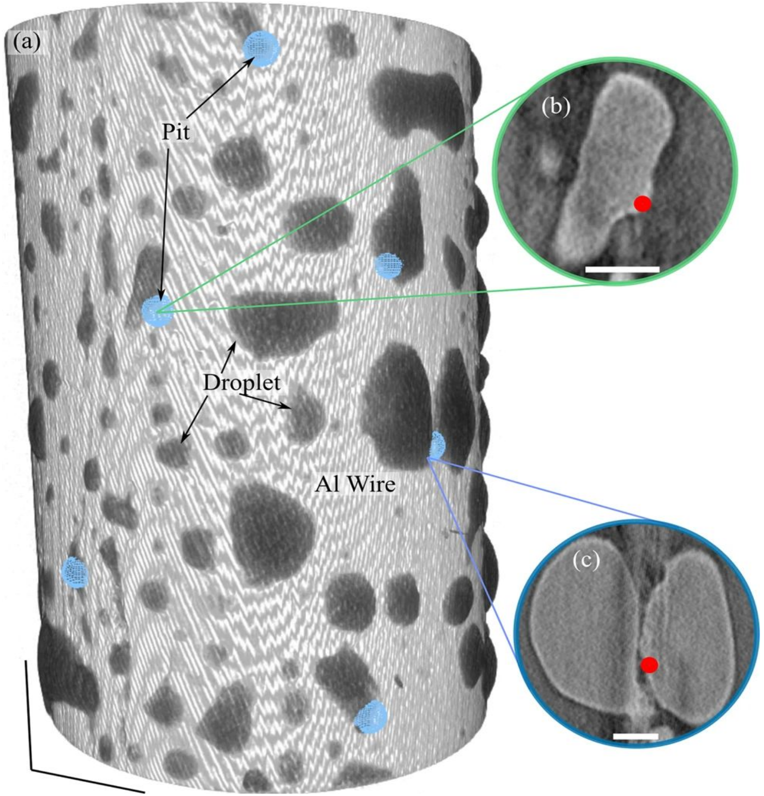
- Growth kinetics of cumulative pits
- Growth kinetics of individual pits
- Evolution of pit morphology

Current Approach (at Sandia)

- Manual segmentation of the pits using
  - Commercial software: *Dragonfly* 3D
  - Based on grayscale values and location
  - Evaluate pit volume and surface area

Goal:

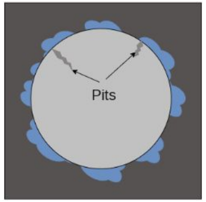
Build a pipeline to study pitting corrosion behavior through a large scale XCT dataset



P. J. Noell et al., "The evolution of pit morphology and growth kinetics in aluminum during atmospheric corrosion," npj Mater Degrad, 7, 1, 1, Feb. 2023.

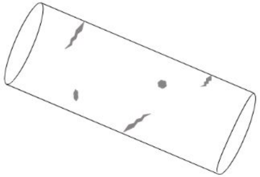
Pipelining Process is Great for Communicating Code

Cross-sectioned XCT image

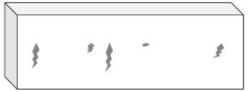


Pit Segmentation Maps

3D Cylindrical Pit Volume



3D Unwrapped Pit Volume



2D XCT Image →

Pit Segmentation on 2D XCT images
Semantic Segmentation with U-Net
→

Pit Reconstruction in 3D
Stack 2D pit Segmentation Maps
→

Unwrap 3D Cylindrical Pit Volume
Python Package *pointextract*

→ **Visualization of All the Pits Within the Wire**
3D Visualization Tool *MayaVi*

→ **Statistical Quantification of Pitting Corrosion**
Pit Morphology Evolution

From Pit Reconstruction in 3D:

→ **Visualization of Individual Pits**
3D Visualization Tool *MayaVi*

→ **Statistical Quantification of Pitting Corrosion**
Growth Kinetics of Individual Pits
Growth Kinetics of Cumulative Pits

U-Net Architecture for Image Segmentation

Train an U-Net model on 2D XCT images

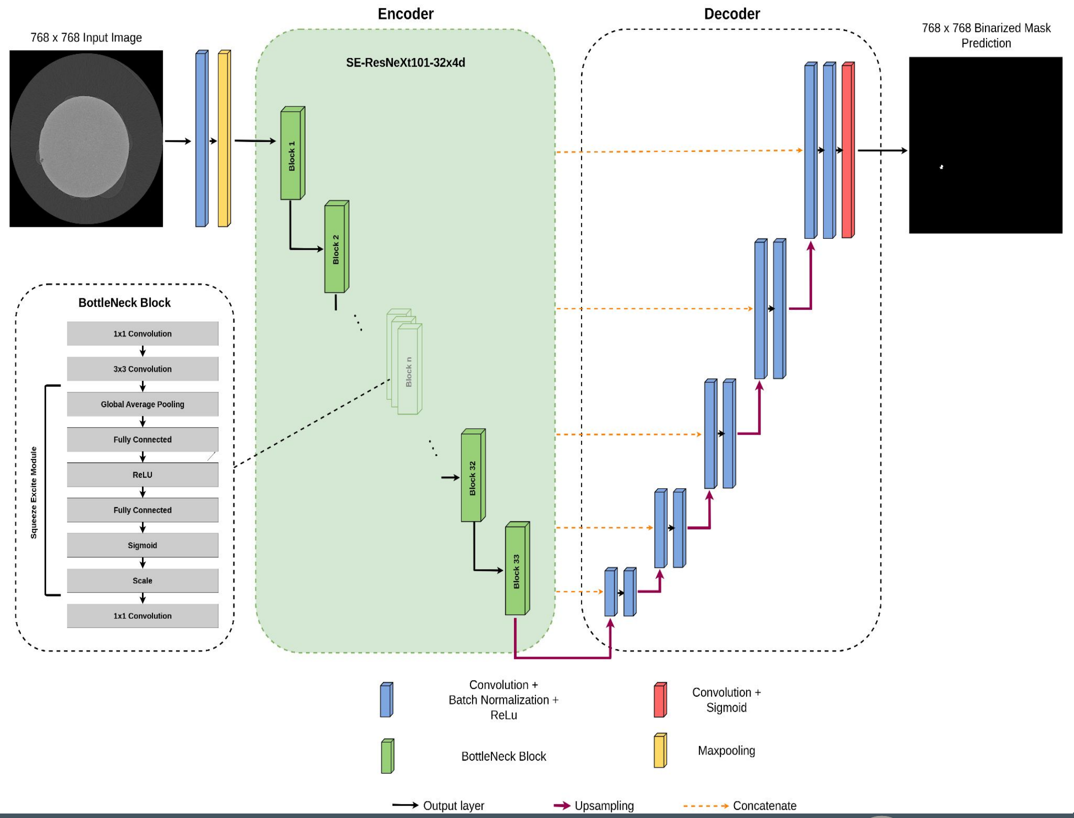
- # training images = 293
- # epochs = 100
- Batch size = 4

SE-ResNeXt101 encoder

- Provides sufficient depth as standard four block encoder failed

Hybrid focal & Jaccard loss function:

- Focal: class imbalance
- Jaccard: IoU focus
  - Intersection over Union (IoU)



U-Net Model Results

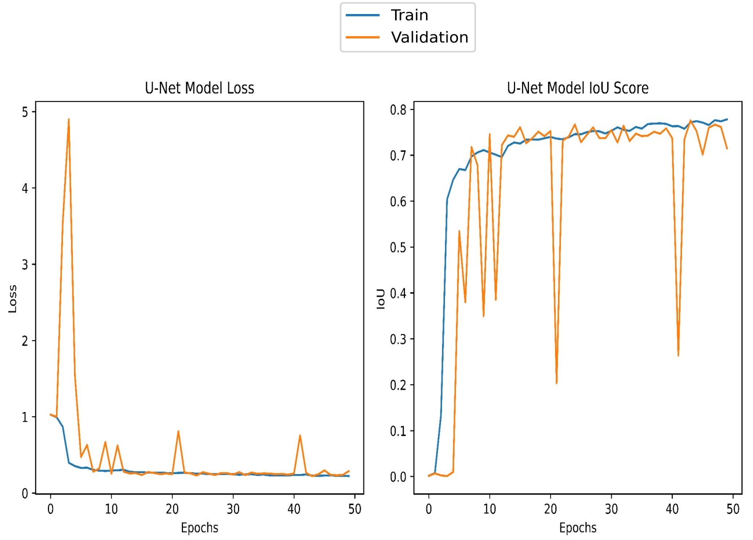
Model Performance

Accuracy = 99.9 %

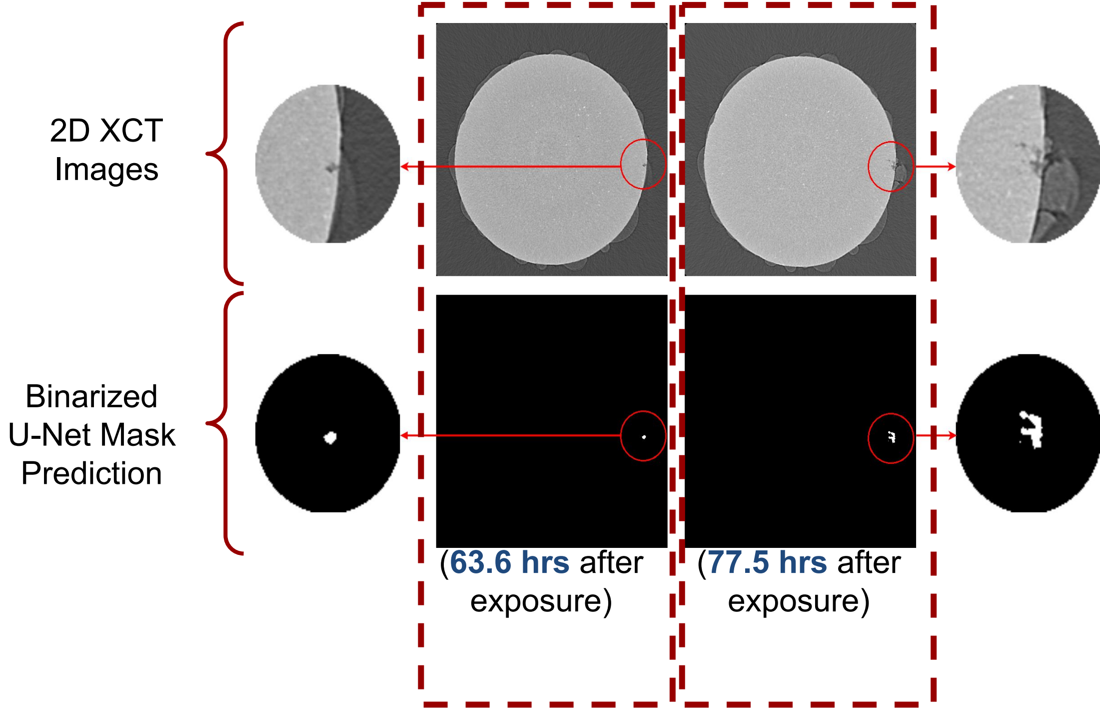
Precision = 88.2 %

Recall = 90.4 %

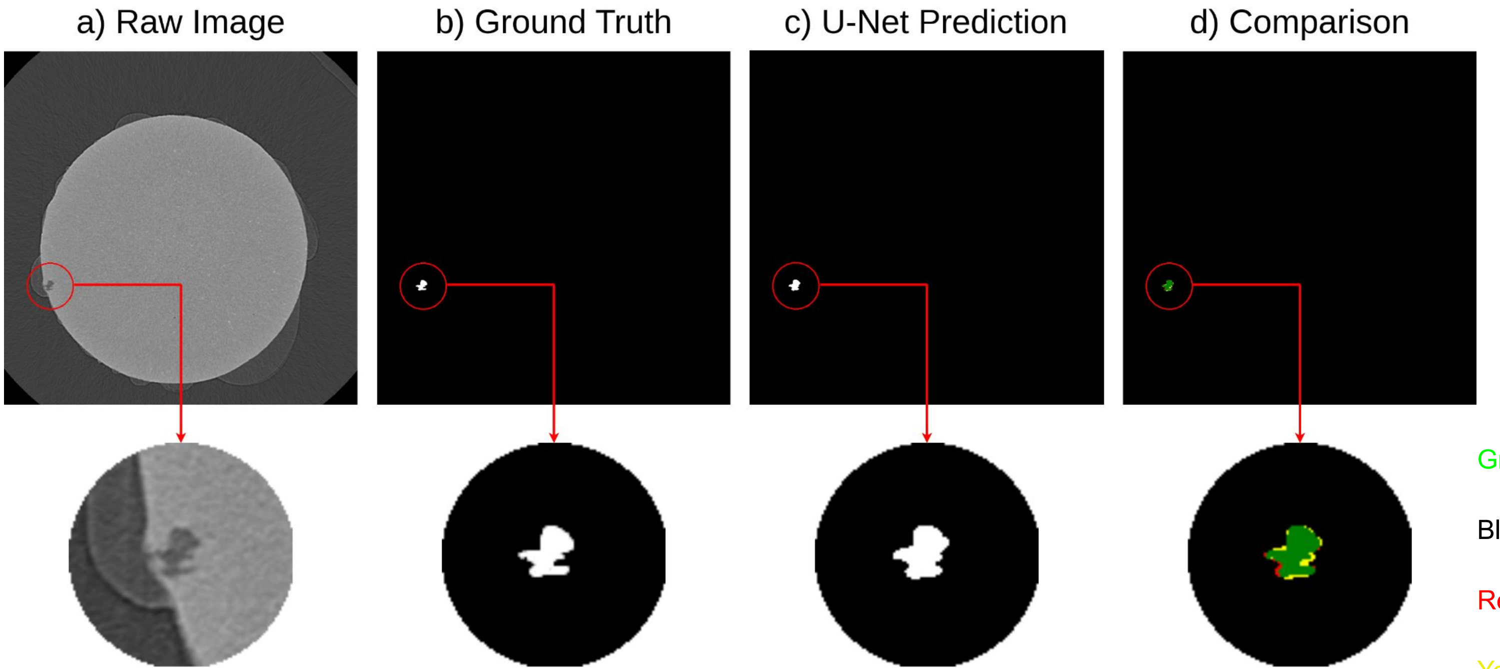
Binary IoU = 79.2 %



Segmentation Prediction Example



Segmentation Comparison



Green	True Positives
Black	True Negatives
Red	False Negatives
Yellow	False Positives

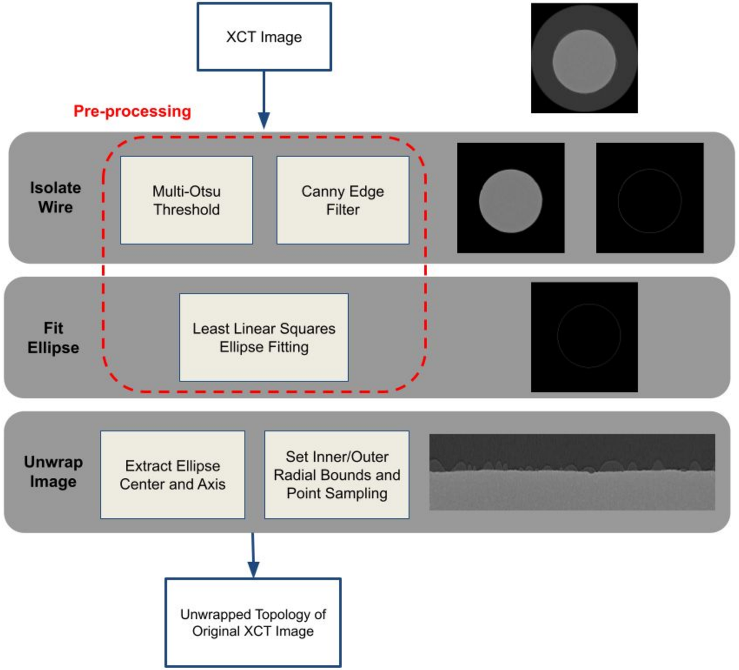
Unwrapping the Cylindrical Wire, for Pit Visibility

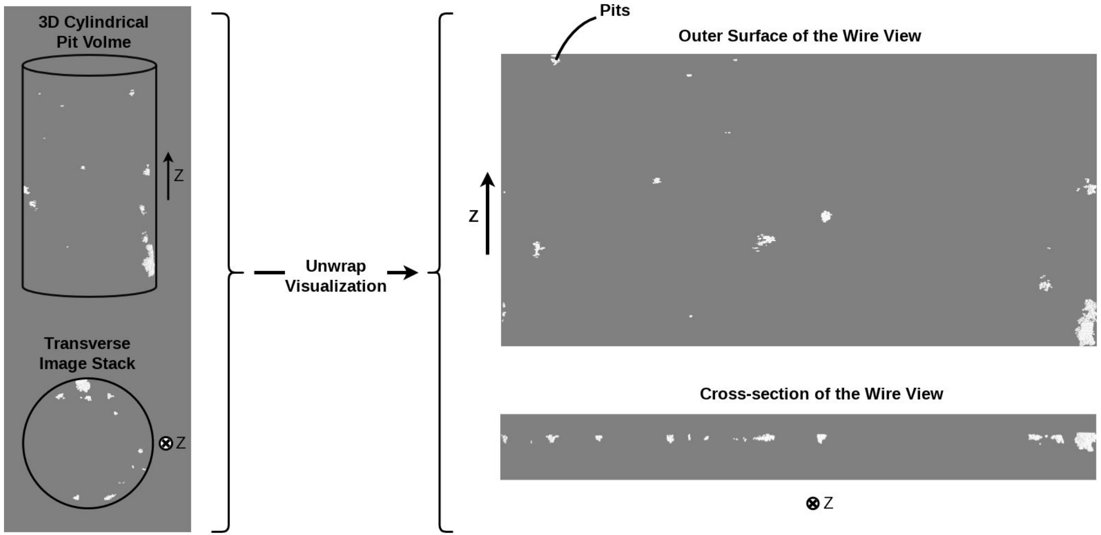
pointextract.py

- Translates a 2D wire cross section to a rectangular version

Applied this transformation to

- the entire 3D volume of pit segmentation maps
    - generated by U-Net



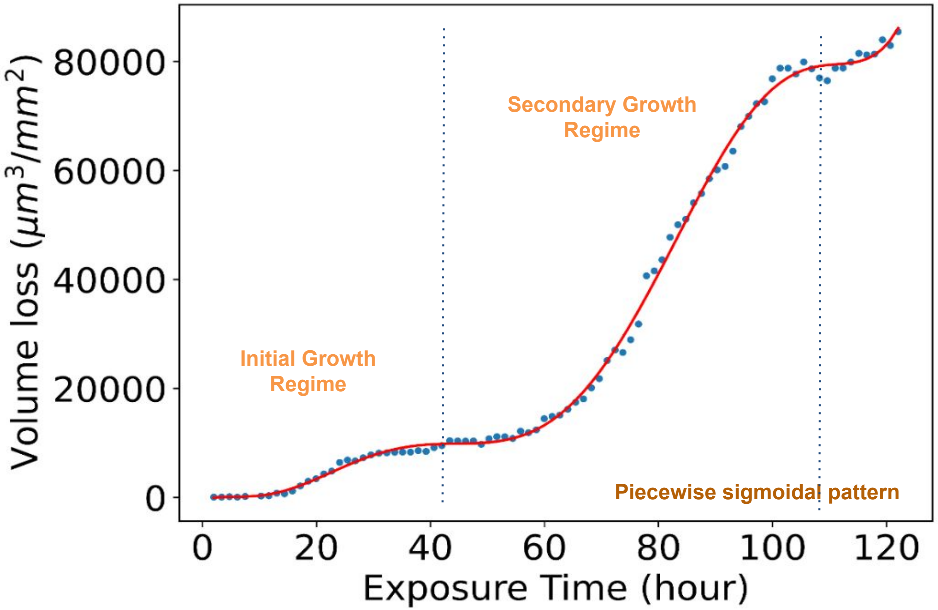


Z - Along the wire's axis

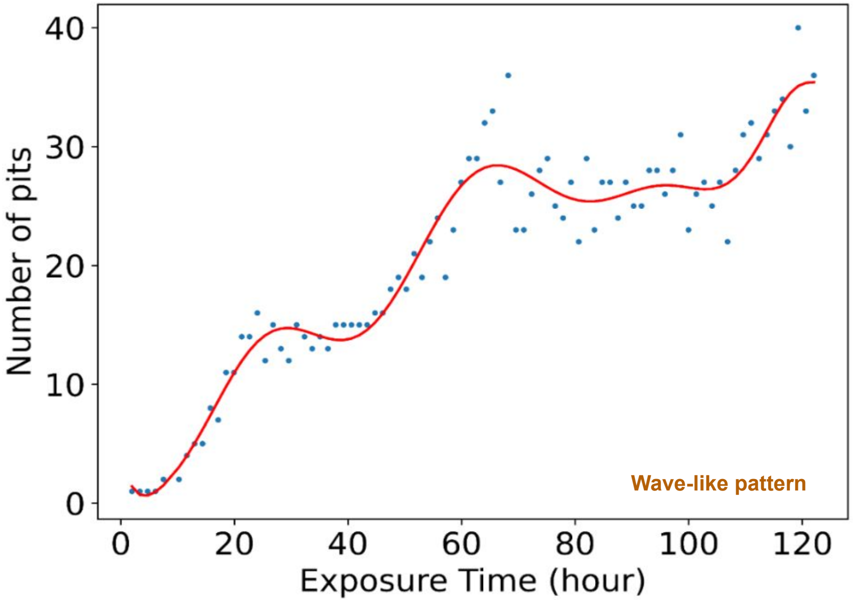
[1] Liangyi Huang and Roger H. French, "pointextract." 11-May-2022 [Online]. https://pypi.org/project/pointextract/

Temporal Variations of Cumulative Pits

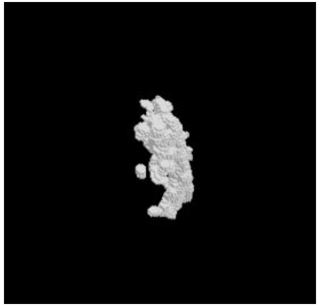
Cumulative Volume Loss Over Time (Pitting Volume)

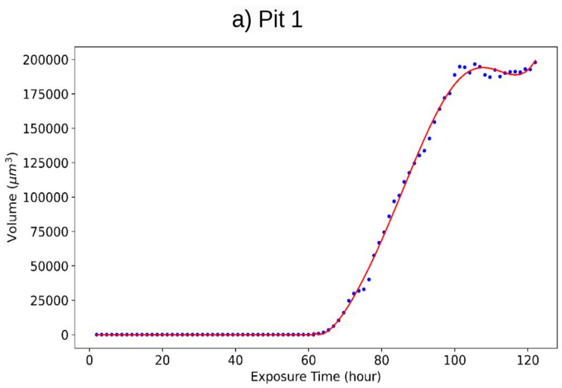


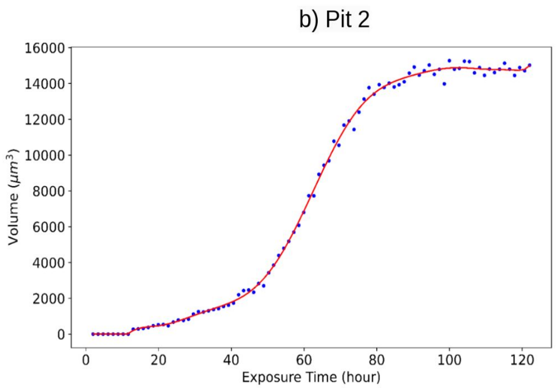
Number of Pits Over Time

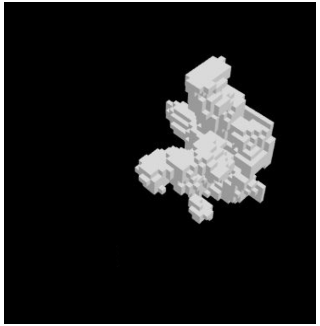


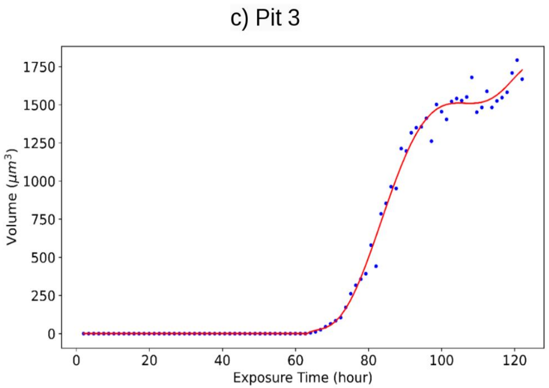
Growth Kinetics of Individual Pits

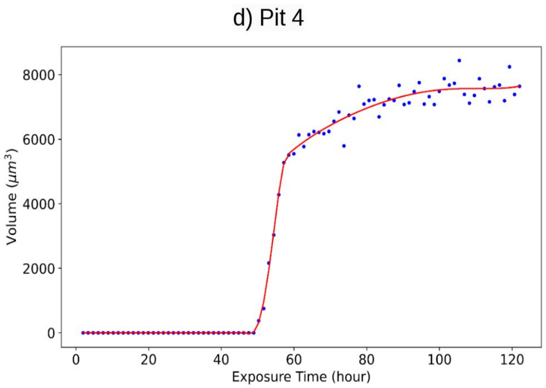


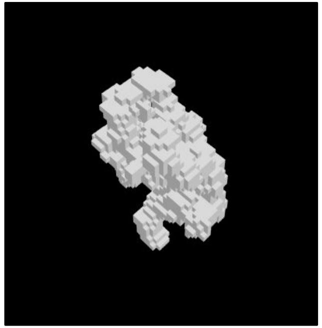








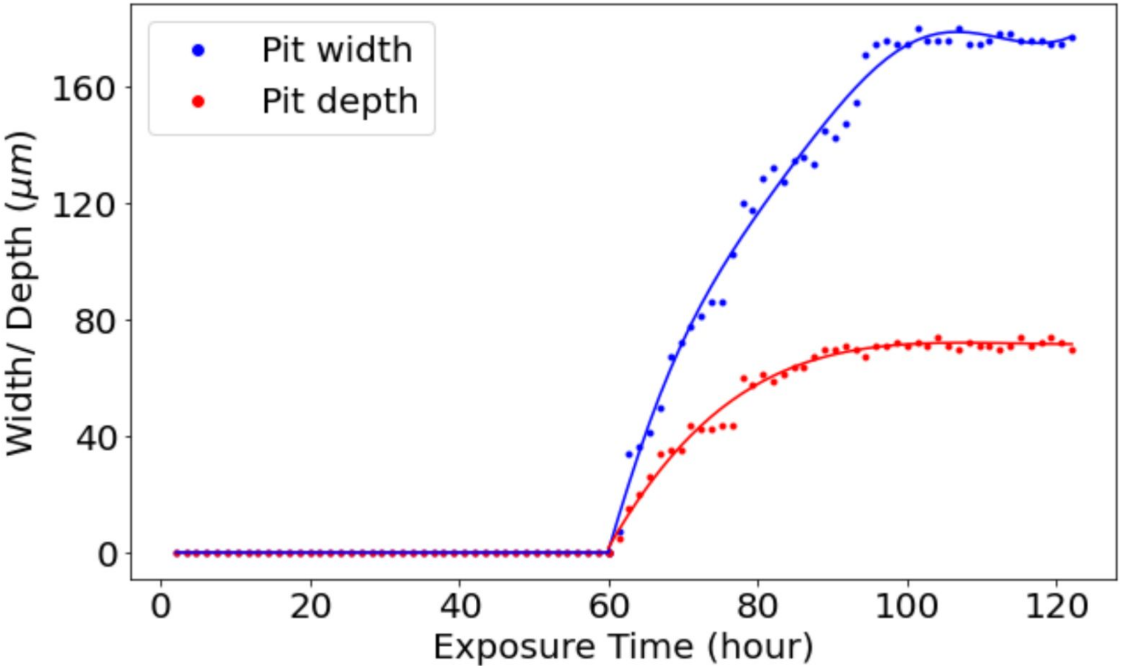




All 4 pits exhibit sigmoidal growth kinetics.

Pit Morphology Evolution

Pit width and pit depth evolution over time for the largest pit



Both depth & width start expanding

- from the point of nucleation

The width of the pit

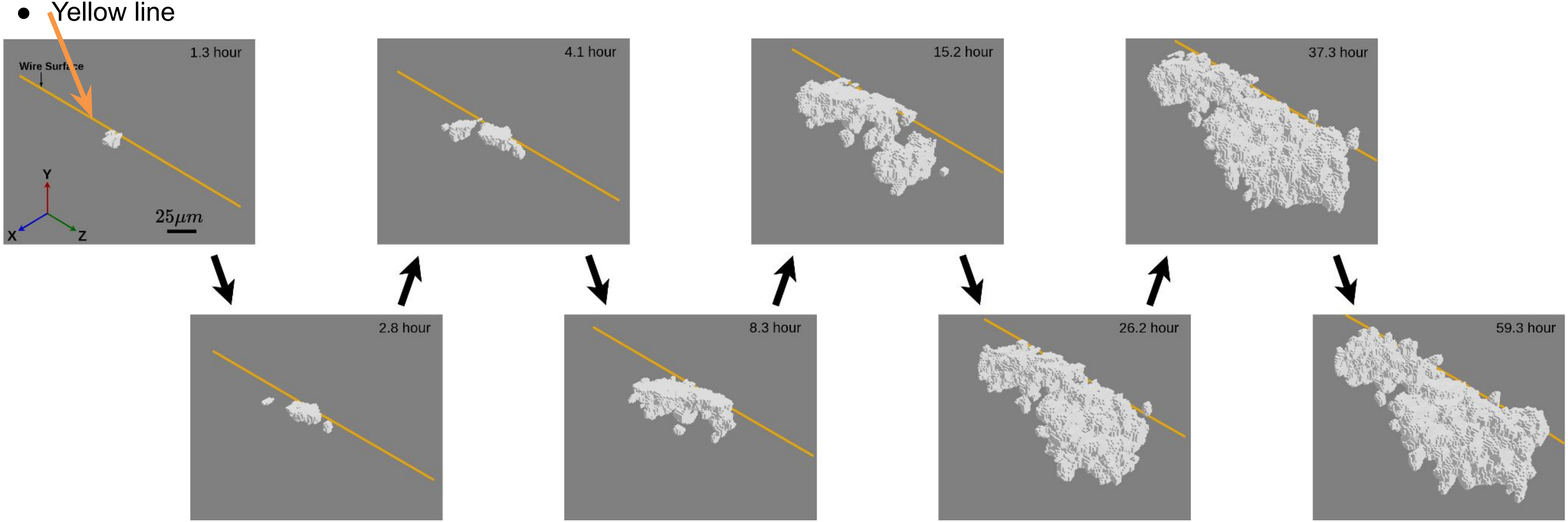
- Is growing at a faster rate
- than its depth.

MDS³ COE, SDLE Research Center, Roger H. French © 2023 https://mds3-coe.com http://sdle.case.edu DE-NA0004104

Pit Morphology Evolution Over Time: Impact of Texture?

Plane of Wire's Surface

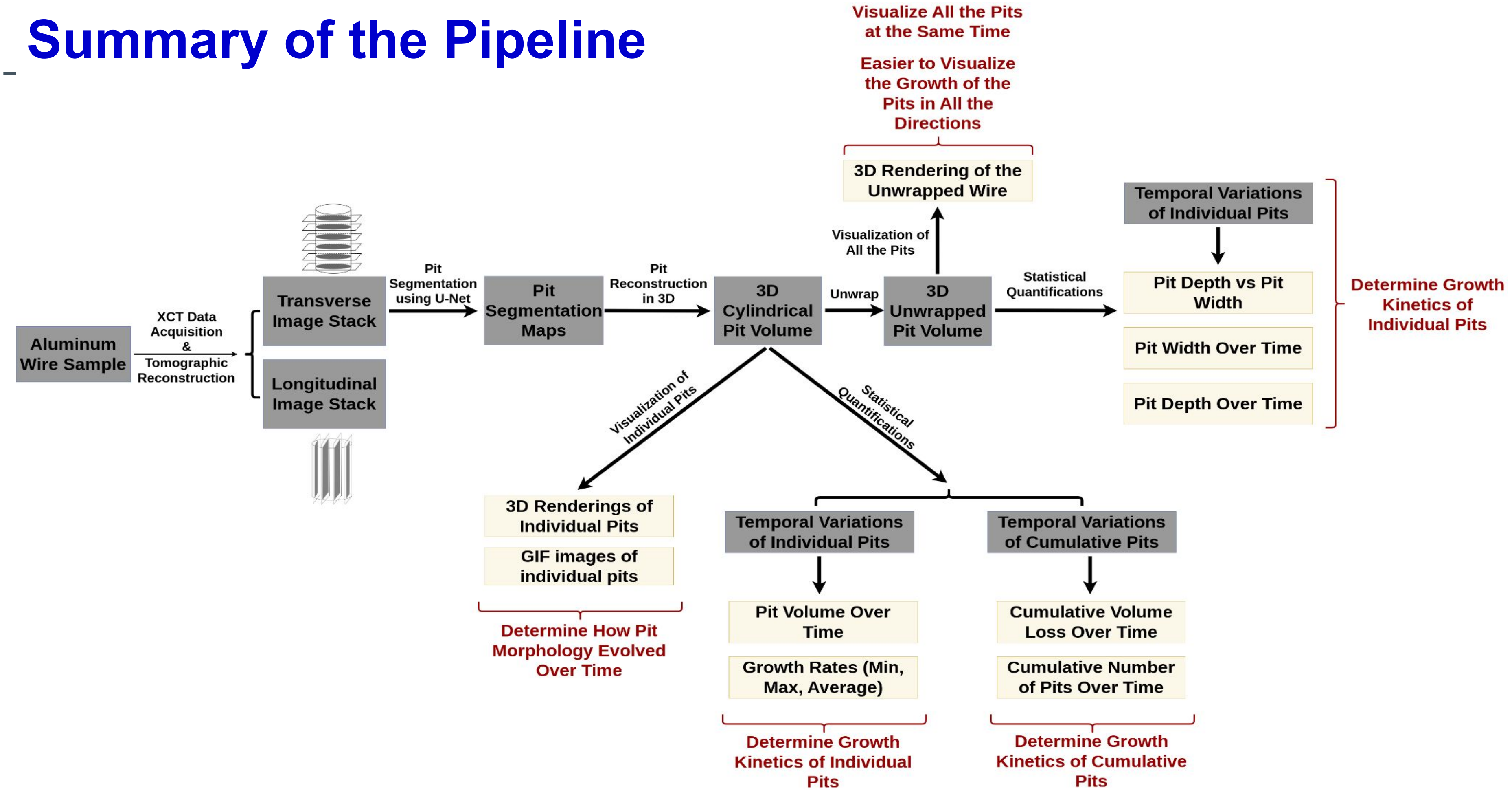
- Yellow line



For this large pit

- Growth progresses into, and along the wire axis
- Possibly arising from the textured microstructure of the wire

Summary of the Pipeline

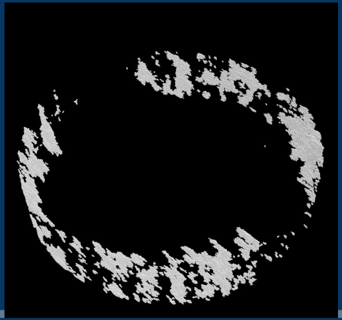


Pitting Corrosion Takeaways

Our automated pipeline has a global impact by:

- **Enhanced efficiency in corrosion detection and assessment**
  - Reduce time and resources vs. manual inspection.
- **Ability to assess the lifetimes, enhance reliability, and**
  - Ensure long-lasting durability of passive alloys.
- **Improved maintenance and safety in infrastructure**
  - Allows for timely maintenance and replacement of affected components
  - Reducing the risk of failures, outages, and accidents.
- **Environmental impact and sustainability**
  - Reduce waste and the environmental footprint
  - Associated with alloy components production and disposal
  - By extending the lifespan of them.

Deep Learning Framework for Spatiotemporal Feature Extraction and Statistical Characterization of Terabyte-Scale XCT Datasets



GS: Tommy Ciardi[1],

Faculty: John Lewandowski[2], Roger H. French[1,2]

1. Department of Computer and Data Sciences, CWRU, Cleveland, OH
2. Department of Materials Science & Engineering, CWRU, Cleveland OH, USA

Strengthening NNSA's Capability to Modernize Manufacturing & Production

The Big Picture (and Challenge)

A Materials Science Problem

How do **inclusions** influence **stress corrosion cracking** in **Al-Mg alloys** in **different environments**?

Enabled through X-ray Computed Tomography

Challenge: scale of the data

- Terabytes per sample
- Outpaces the current analysis software

A Materials Science Domain Challenge

How do **microstructural features** influence **temporal changes** in **materials** under **certain conditions**?

Advances in instrumentation and computational power

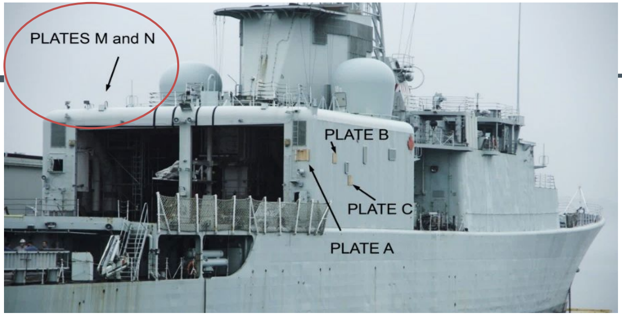
Challenge: scale of the data

- Order of Terabytes
- Outpaces software and infrastructure

Challenge: experimental philosophy

- Reduction of data

Experimental Background



AlMg plates from HMCS Iroquois:

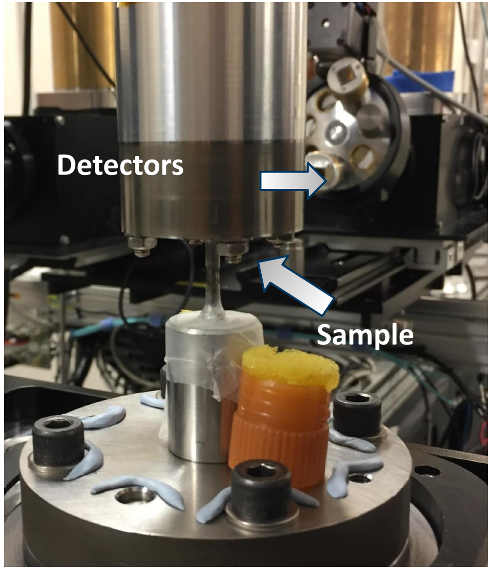
- Decommissioned Navy destroyer
- 1972 to 2014 in Gulf theatre, Domalia, and Caribbean Sea
- Aluminum: 5XXX rolled plates, ~H116 temper, ~4.7-5.5wt% Mg

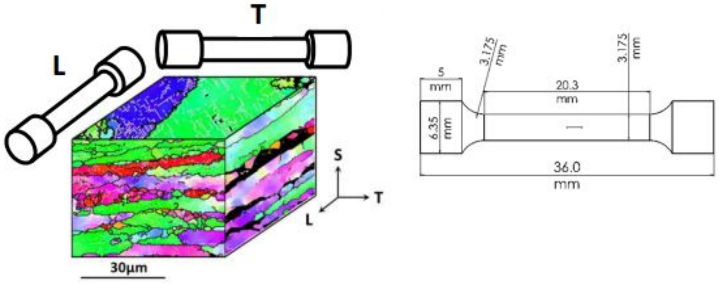
Sample Processing:

- Plane N (roughly 6mm thick)
- Exposure to sun = higher degradation
- T orientation

Slow strain-rate tension test

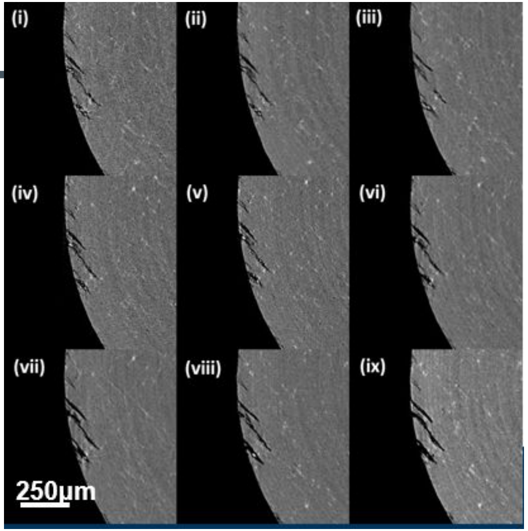
- Synchrotron at Diamond Light Source
    - (Didcot, UK)
- Intermittent holds on load to scan XCT





[1] Burnett, T.L., Holroyd, N.J.H., Lewandowski, J.J., Ogurreck, M., Rau, C., Kelly, R., Pickering, E.J., Daly, M., Sherry, A.H., Pawar, S., Slater, T.J.A., and Withers, P.J. (2017). "Degradation of Metallic Materials Studied by Correlative Tomography", in 38 th Riso International Symposium on Materials Science – IOP Conf. Series: Materials Science and Engineering, 219(1), 012001.
[2] Gudla, V.C., Garner, A., Storm, M., Gajjar, P., Carr, J., Palmer, B.C., Lewandowski, P.J., Holroyd, N.J.H., and Burnett, T.L. (2019). "Initiation and Short Crack Growth Behavior of Environmentally Induced Cracks in AA5083-H131 Investigated Across Time and Length Scales", Corrosion Reviews, 37(5), pp. 469-481.

Experimental Data from 3D XCT Scans



Series of 3D volumetric scans as 2D images in a movie through time

Each scan (one 3D image) sliced into 2110 2D .tiff images

- 12MB per image at 2510x2510 resolution = 24.7 GB per scan

	< 1% RH (dry)	50% RH
Number of scans	36	77
Size	**929.5 GB total (~1TB)**	**1949.64 GB total (~2TB)**

Scale of Data:

~3TB of image data from two samples (3.4 TB of total img/non-img data)

231 subdirectories

238,430 images (239,879 files)

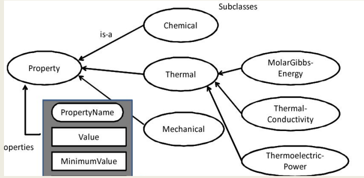
Previous analysis has been limited to hand selected subsets of the dataset[1]

- **Data reduction problem**

[1] Gudla, V.C., Storm, M., Palmer, B., Lewandowski J.J., Withers, P.J., Holroyd, N.J.H, and Burnett, T. (2020). “Environmentally Induced Crack (EIC) Initiation, Propagation, and Failure: A 3D In-situ Time-lapse Study of AA5083-H131”, Corrosion Science, 174, 10.1016/j.corsci.2020.108834

Image to Scene Knowledge Learning Framework

Metadata, Ontologies
RDF triples, JSON
(abstraction/semantics/constraints)



... to data standardization
& knowledge sharing

1

FAIRification

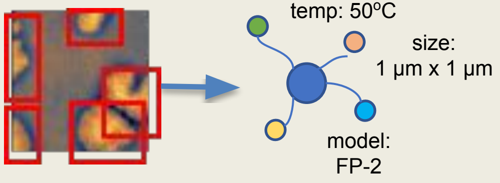
validate

curate

2

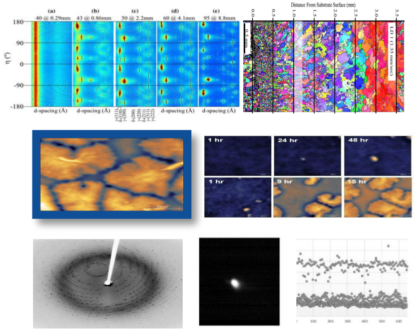
Featurization

Objects, Observations & Properties
(instances)



... to create AI/ML ready
data resources

Raw datasets
Image/videos, design data,
in-situ/ex-situ data..



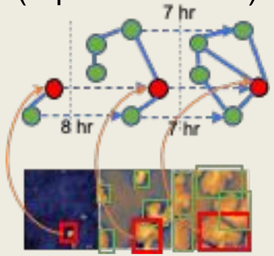
ML/AI models

constraints/ validation

linked entities

enriched features

Scenes or st-Graphs
(representations)



Deep (graph) representation learning
... to inferential & predictive models

Deep Learning

4

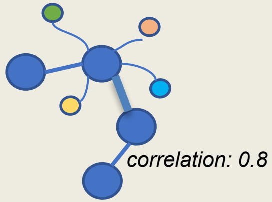
cost-effective learning

efficient access & interpretation

3

Summarization

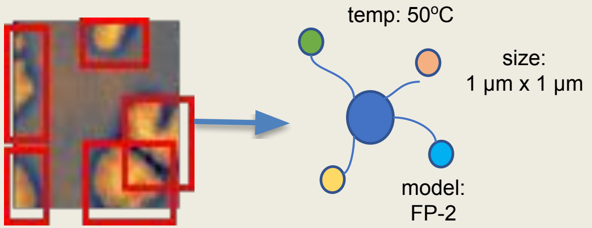
Summaries & Patterns
(data regularities)



... to cost-effective data access,
analysis & interactive exploration

Featurization

Objects, Observations & Properties
(instances)

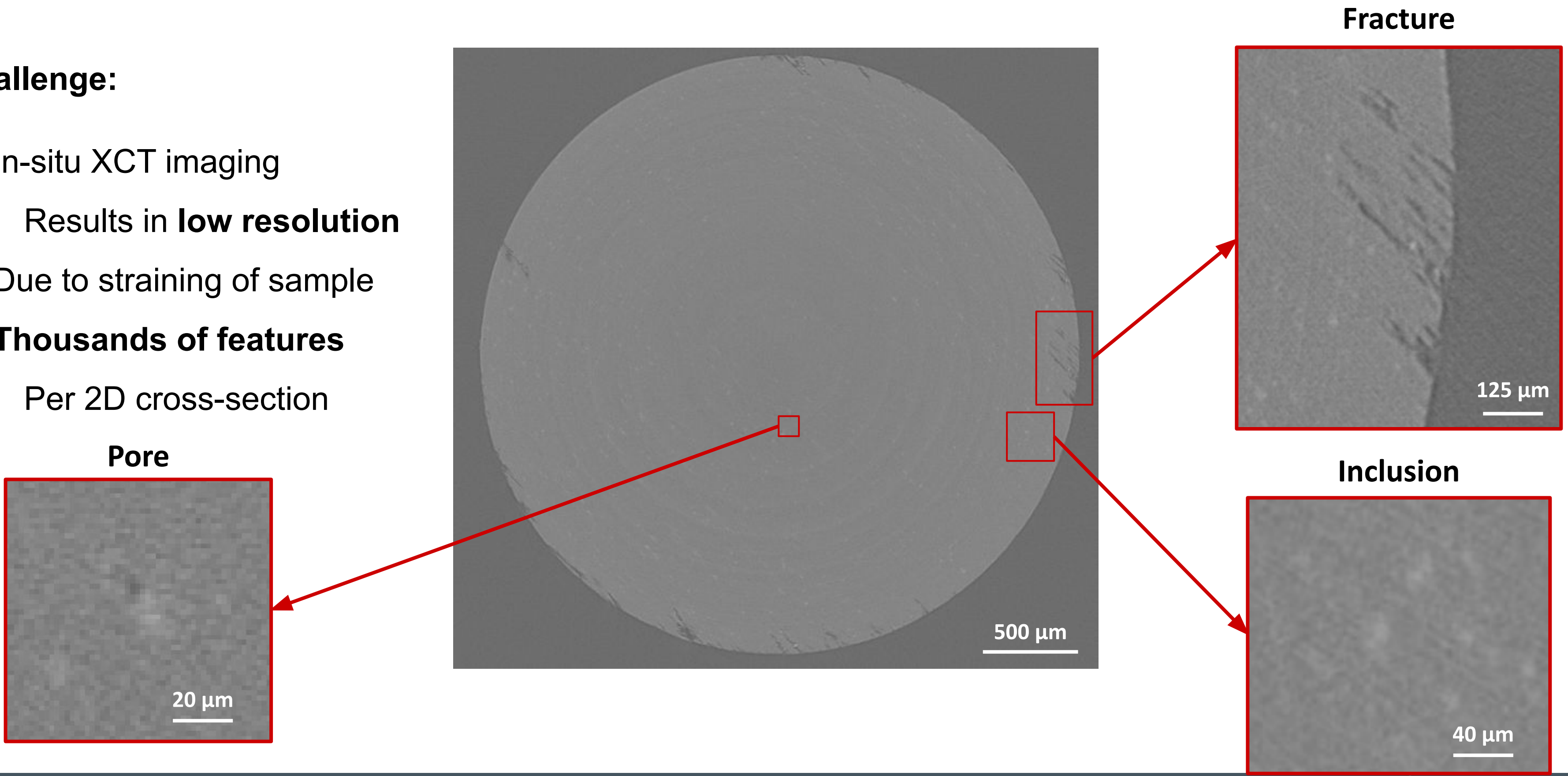


... to create AI/ML ready data resources

Features of Interest: Overview

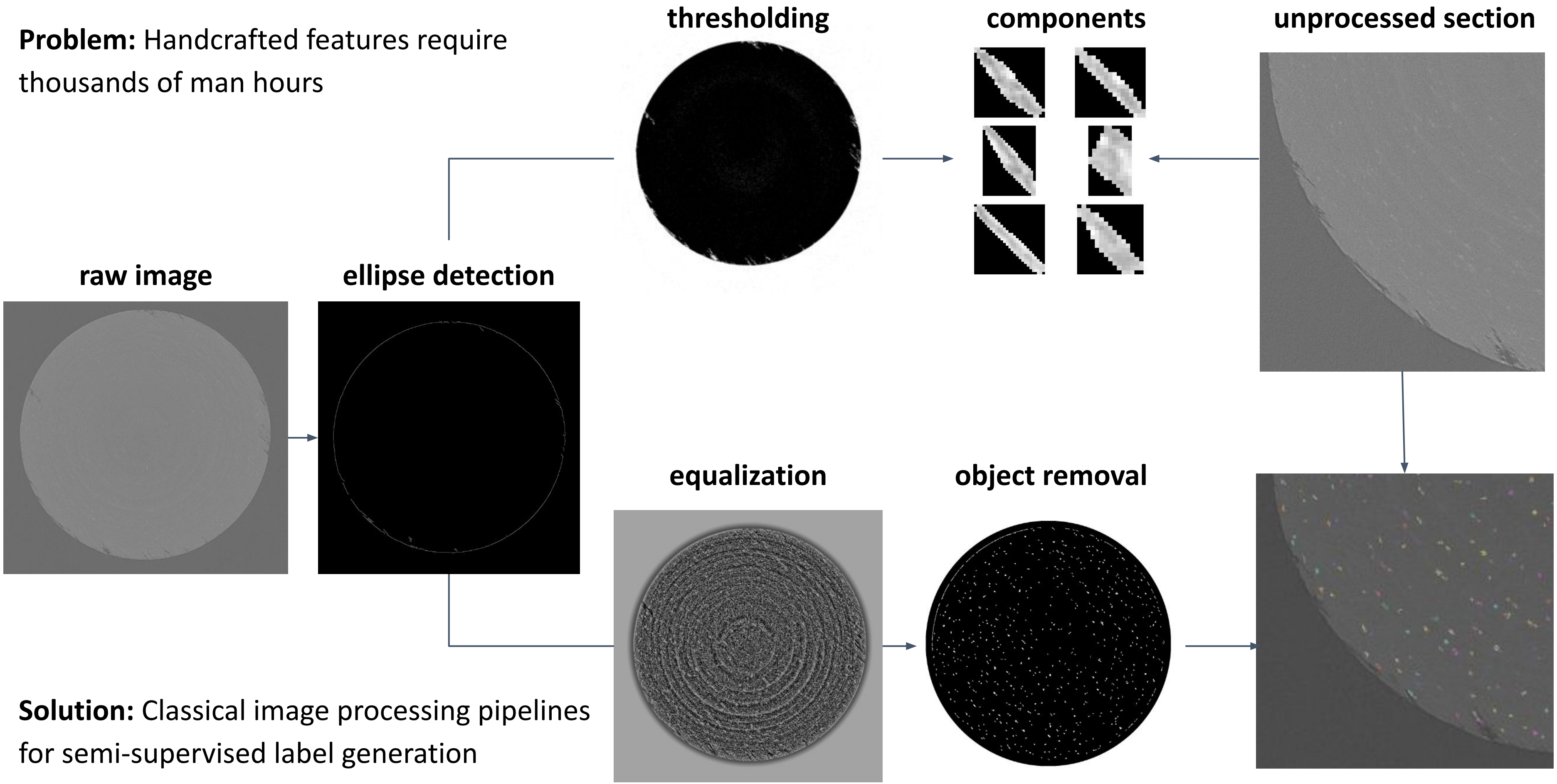
Challenge:

- In-situ XCT imaging
  - Results in **low resolution**
- Due to straining of sample
- **Thousands of features**
  - Per 2D cross-section



Feature Extraction: Classical Image Processing

Problem: Handcrafted features require thousands of man hours

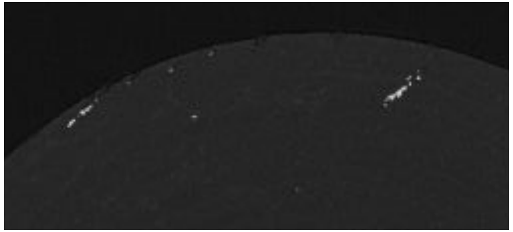
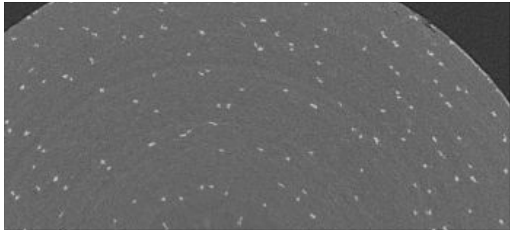


Solution: Classical image processing pipelines for semi-supervised label generation

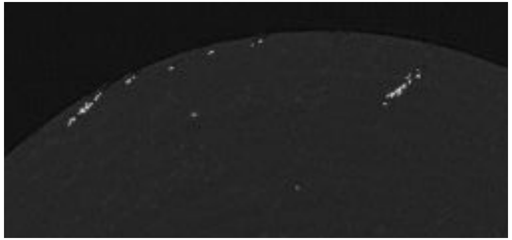
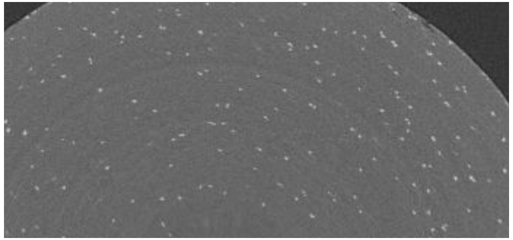
Feature Extraction: Deep Learning

- Image processing is parameter dependent and computationally heavy
- Deep learning networks offer robust, transferable segmentation models

Image Processing

Deep Learning

	Fracture	Inclusion
Precision	0.94	0.91
Recall	0.89	0.88
Binary IoU	0.92	0.89

* Comparison of UNet segmentation to image processing label.

Example failure case that becomes solved

Raw

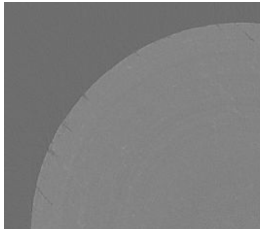
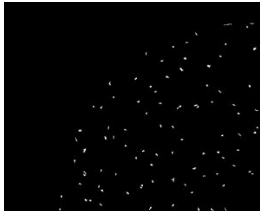
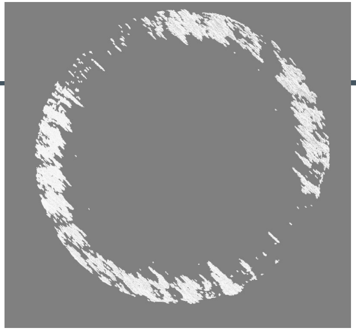


Image Process

Deep Learning

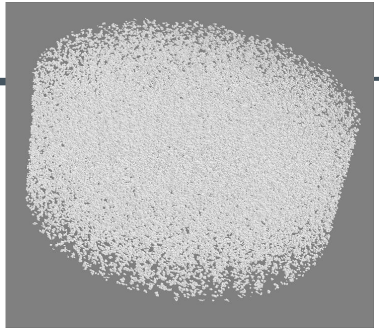


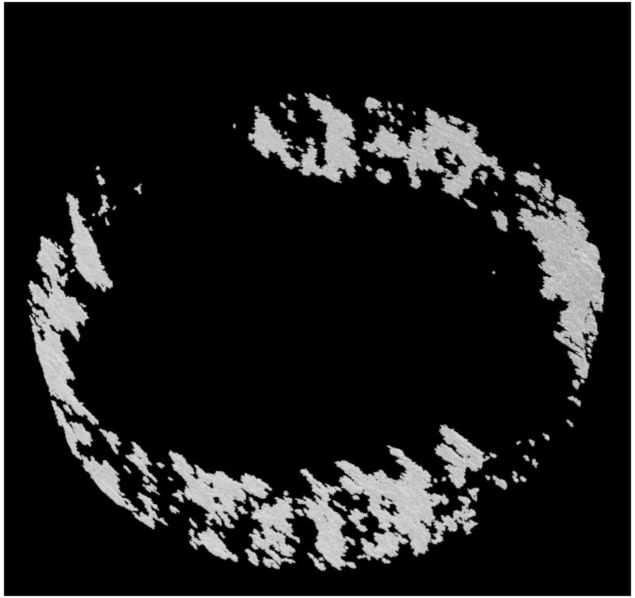
3D Reconstruction of Features



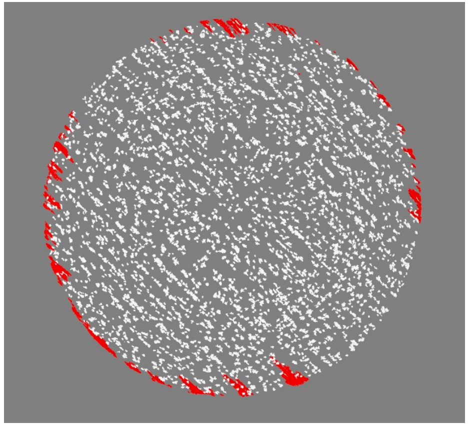
We can predict features on 2D cross sections

Then stack the segmentation masks to reconstruct our features in a 3D space

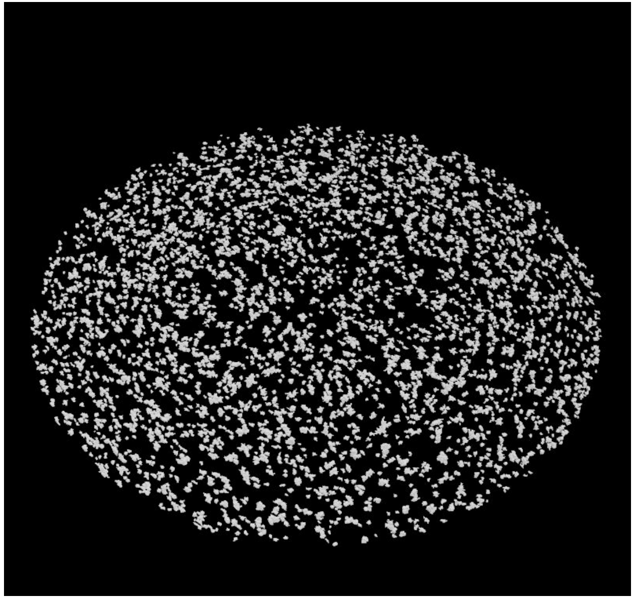




Fracture reconstruction: 250 slices



Top down view of 50 slices with labeled fracture and inclusion

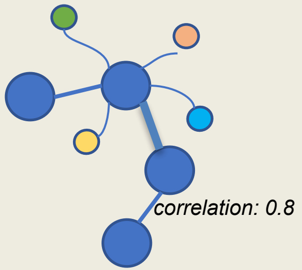


Inclusion reconstruction: 50 slices

Summarization

Summaries & Patterns

(data regularities)



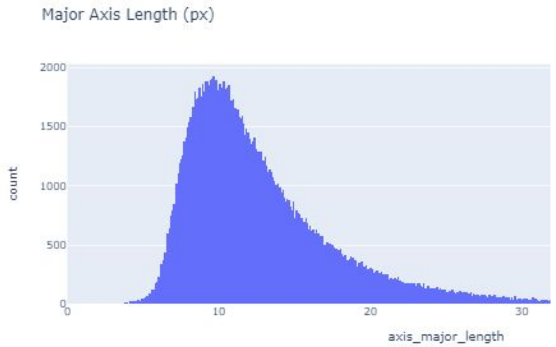
… to cost-effective data access, analysis & interactive exploration

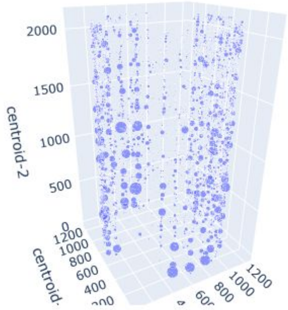
Statistical Quantification - Summary

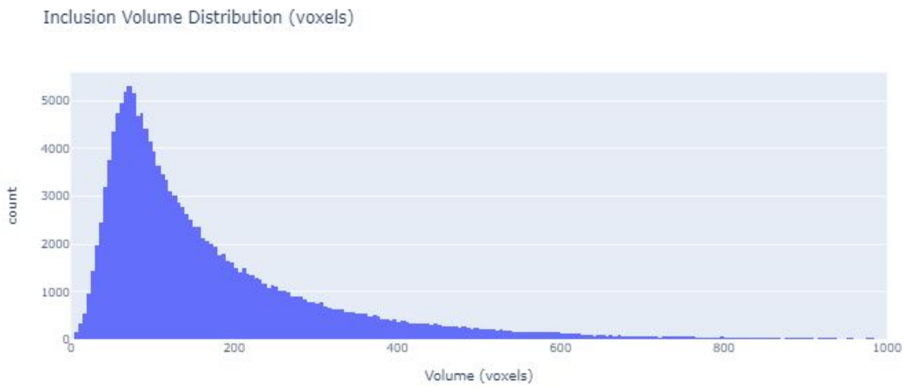
Quantification of defects for a full 3D XCT volume enables us to build a **complete microstructural defect profile**

We can query and understand our complete dataset:

Spatial: How many inclusions exist in one mm^3?
Temporal: What is the average fracture length over time?





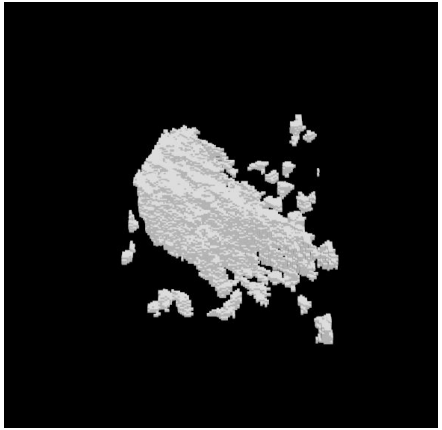


Inclusion Feature (single 3D volume)	Value
Count	**161574**
Average major axis (px)	10.978
Average volume (voxel)	180.904
Volume fraction	~0.9%

Statistical Quantification - Granular

Quantification of every individual feature enables us to **investigate a single defect** of interest

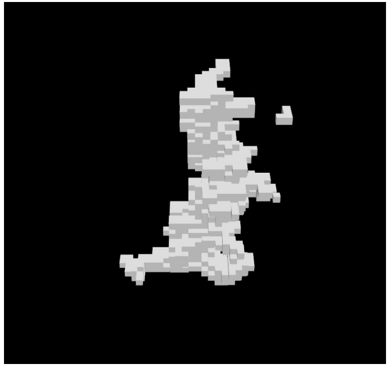
Query x feature for attributes of the any 100,000+ features detected



Largest detected fracture at timestep 20

Query profile → Retrieve attributes → Gain Insight

Largest inclusion at timestep 25 and attributes



Inclusion Feature	Value
Major axis (px)	43.01
Volume (voxels)	1340

Automated extraction of 13,000,000+ total features

Next Steps: Spatiotemporal Scene Graphs

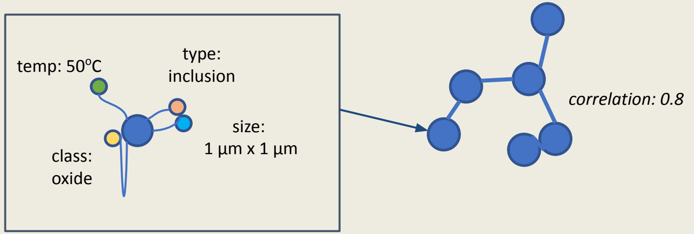
How can we ask more complex questions

- (E.g. do fractures tend to extend towards regions of higher defect density?)

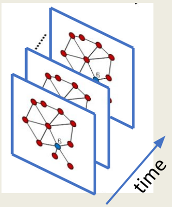
Generate scene graphs[1] for an interpretable full-scale microstructural and degradation analysis

Summary Graph Generation

Labeled features can be turned into nodes in a graph and then edges created between corresponding nodes

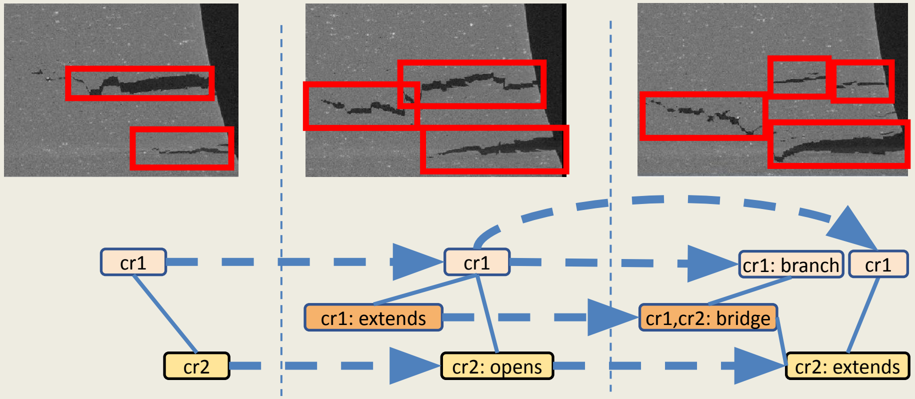


A single graph represents one point in time, multiple graphs can be stacked for temporal analysis



Spatiotemporal Scene Graph Generation

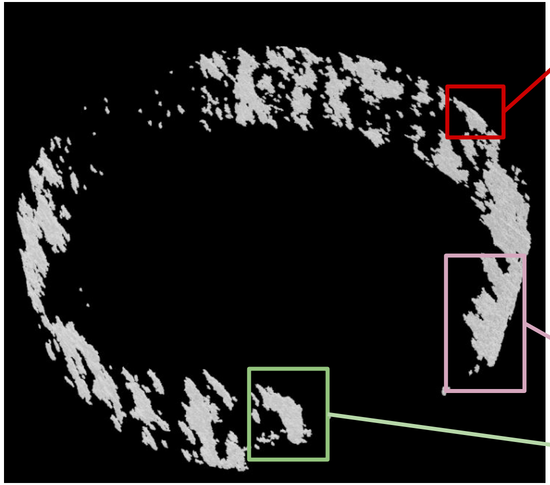
Scene graphs will be generated to label actions and relationships to identify what is occurring both spatially and temporally



[1] Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10236-10247).

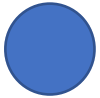
Summary Graph Generation

Translating 3D feature stacks and attributes into graphs

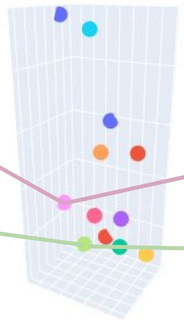


Fracture Features	Value
Major axis (px)	43.01
Volume (voxels)	1340
Orientation	intergranular

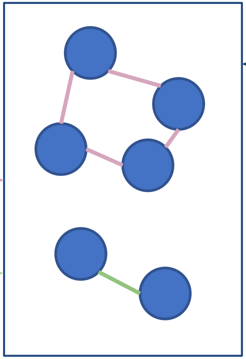
fracture embedded as node



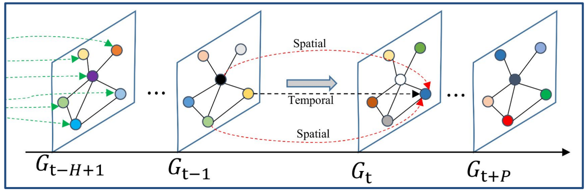
[43.01, 1340, 1]



density based clustering



graph for timestep 32

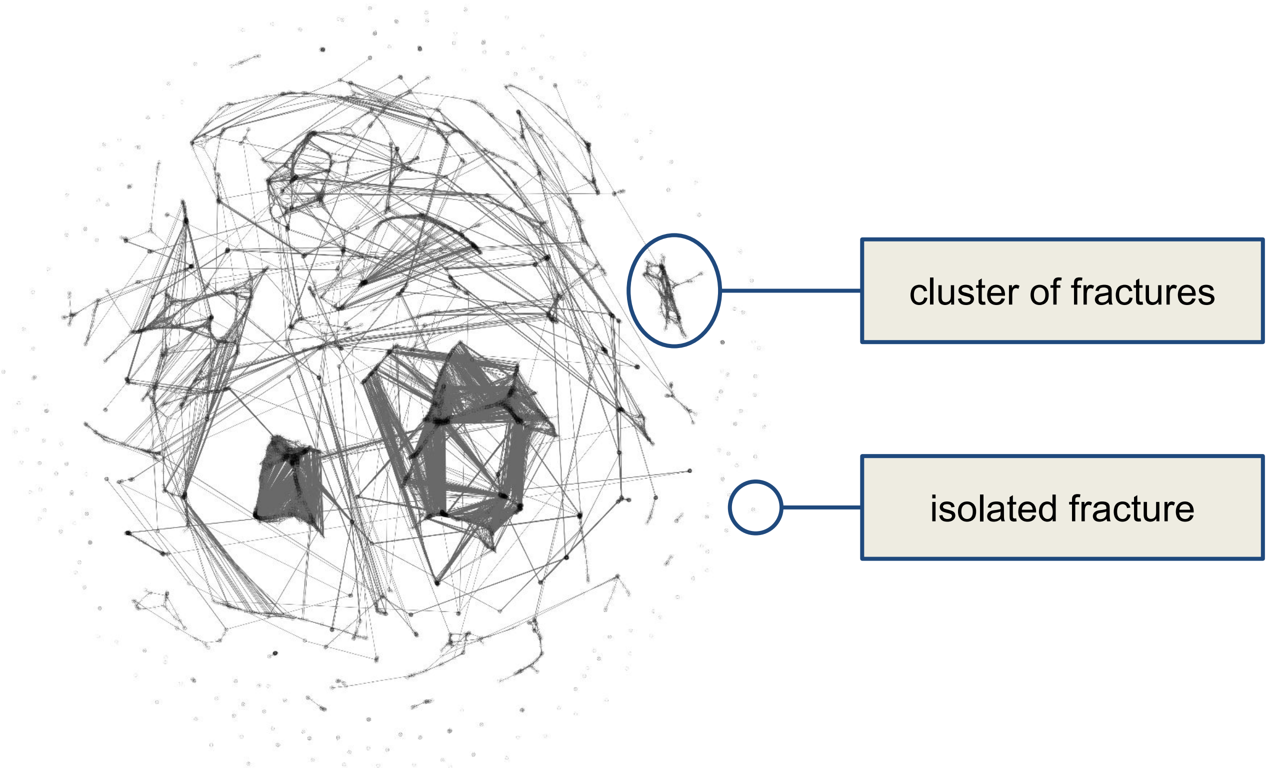


input for graph neural networks

Summary Graph Example: Fractures for a Single Timestep

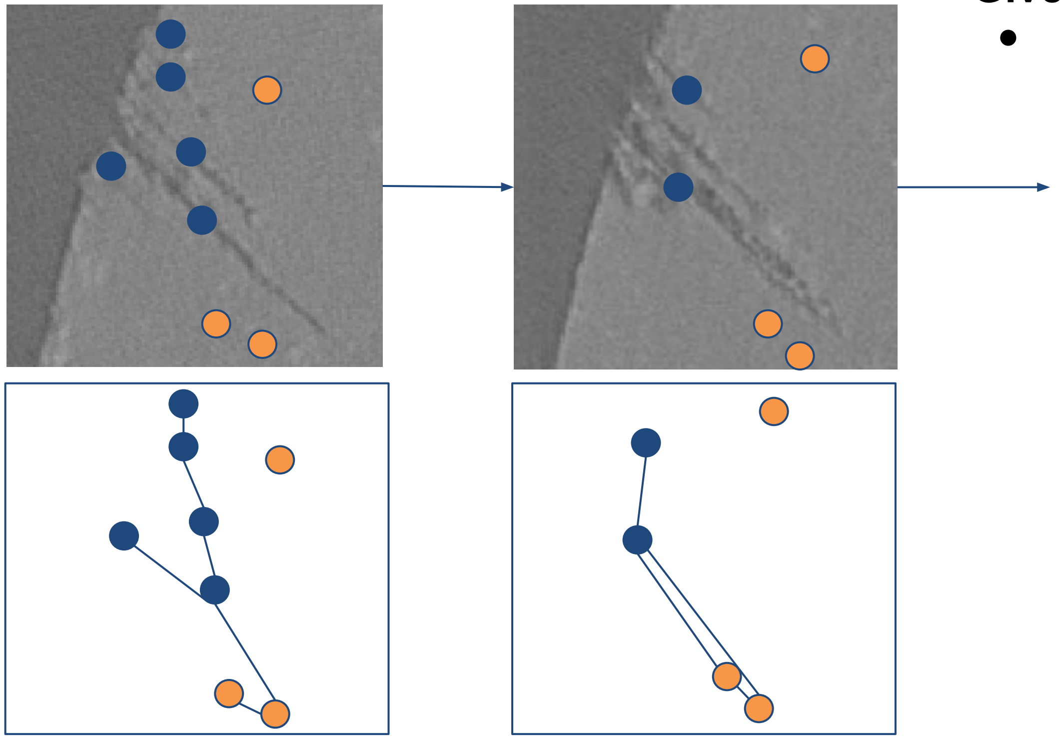
Graph Representation

- Full 3D volume
- Timestep 54
- Fractures only
- 15602 nodes (fractures)
- 198151 edges



Generative Graph Representations

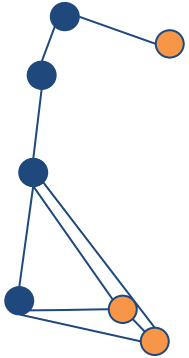
timestep 1 **timestep 2**



Given t_1 … t_i

- predict graph representation for t_{i+1}

…



inclusion

fracture

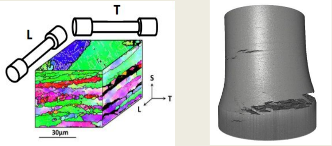
Spatiotemporal Feature Extraction Framework for Large-Scale XCT Datasets

Grand Challenge

Framework to analyze full Tera/Petabyte scale datasets

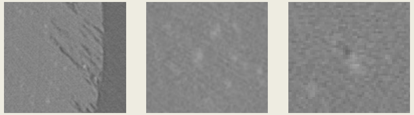
exemplar dataset

stress corrosion cracking in Al:Mg alloy creep test in synchrotron



2 samples = 3 TB data
240,000 2D cross sections

features of interest

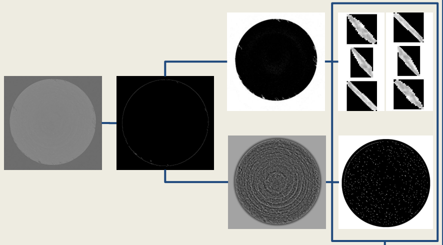


fracture - inclusion - pore

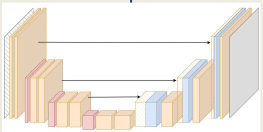
Semi-supervised Deep Learning

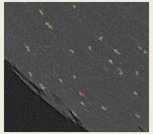
image processing and zero-shot learning annotations

semi-supervised label pipeline



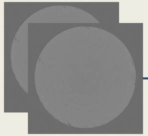
deep learning model





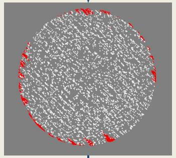
robust feature extraction

3D Defect Reconstruction



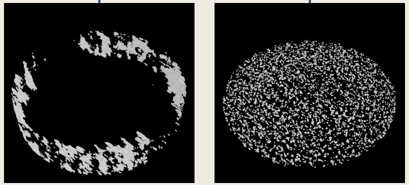
trained deep learning model

feature segmentation mapping



stack feature maps

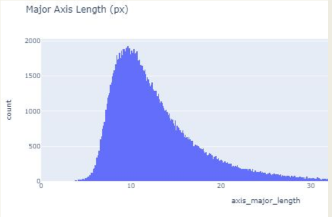
filter spatial inconsistencies



defect volumes

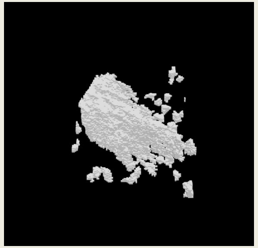
Spatiotemporal Feature Profile

Sample-level Insights



volume fraction	0.903%

Defect-level Granularity



spatiotemporal quantification of over **13,000,000 defects**

Conclusions: Lets Collaborate on Materials Data Science !

AI/ML for Materials Data Science needs D/HPC Computing
- Needs the integration of “Scaled Out & Scaled Up” Computing

CRADLE: Common Research Analytics & Data Lifecycle Environment
- Automated pipelines, FAIRification, Efficient Insights
- Broadly Applicable

CRADLE represents a different mind-set on how to do Materials Science
- Don’t initially simplify, and constrain variables
- Collect all the data
- Analyze ALL the data
- Then summarize it, using Graphs

Data Centric AI presents humans with a grand opportunity
- Augmenting human reasoning; Working alongside human researchers
- Scientific investigations restructured around the “salient human tasks”
- With computers handling the routine and onerous tasks
- Supplementing our human capabilities

While reducing use of reductionist approaches in scientific research

Conclusions: AI Represents an Inflection Point for Science!

AI/ML for Materials Data Science needs D/HPC Computing
- Needs the integration of "Scaled Out & Scaled Up" Computing

CRADLE: Common Research Analytics & Data Lifecycle Environment
- Low barriers to entry for scientists
- Broadly Applicable: Automated pipelines, FAIRification, Efficient Insights
- While Introducing State-of-the-art Data Management, AI/ML, and Scientific Workflows

CRADLE represents a different mind-set on how to do Materials Science
- FAIRified Datasets and FAIRfied Models enable automated AI Materials Science
- Don't initially simplify, and constrain variables
- Analyze ALL the data
- Then summarize it, using Graphs

Data Centric AI presents humans with a grand opportunity
- Augmenting human reasoning; Working alongside human researchers
- Scientific investigations restructured around the "salient human tasks"
- With computers handling the routine and onerous tasks
- Supplementing our human capabilities

MDS3

CENTER OF EXCELLENCE

CASE SCHOOL OF ENGINEERING

CASE WESTERN RESERVE UNIVERSITY