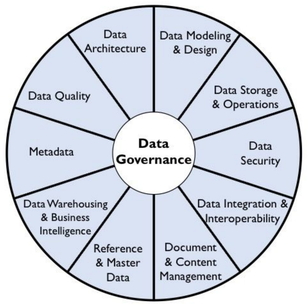
Data Governance and Stewardship for AI:
An Effective Data Lifecycle Using Distributed Computing, Ontologies, and Workflows for NeuroSymbolic AI

Roger H. French[1,2,3], Arafath Nihar[1,3], Tian Wang[1,3], B. Priyan Rajamohan[1,3], Thomas G. Ciardi[1,3], Olatunde Akanbi[1,2], Benjamin Pierce[1,2], Mohommad Redad Medhi[1,2], Hayden W. Caldwell[1,3], Ozan Dernek[1,2], Quynh D. Tran[1,2], Pawan K. Tripathi[1,2], Laura S. Bruckman[1,2], Vipin Chaudhary[1,3], Yinghui Wu[1,3], Erika I. Barcelos[1,3]

1. Materials Data Science for Stockpile Stewardship Center of Excellence, Case Western Reserve University, Cleveland OH 44106
2. Dept. of Materials Science & Engineering, Case Western Reserve University, Cleveland OH 44106
3. Dept of Computer and Data Sciences, Case Western Reserve University, Cleveland OH 44106

DMBOK[1]: Data Governance



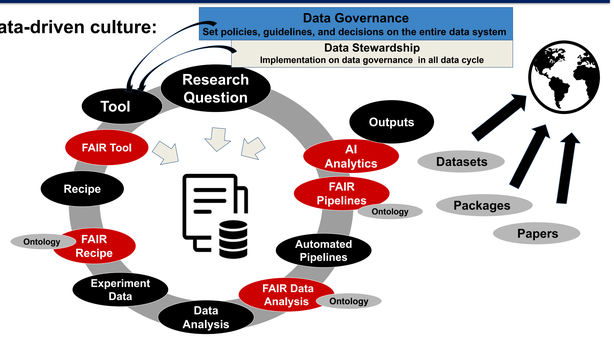
Copyright© 2017 DAMA International

We need to shift our data flows to a data-driven culture:

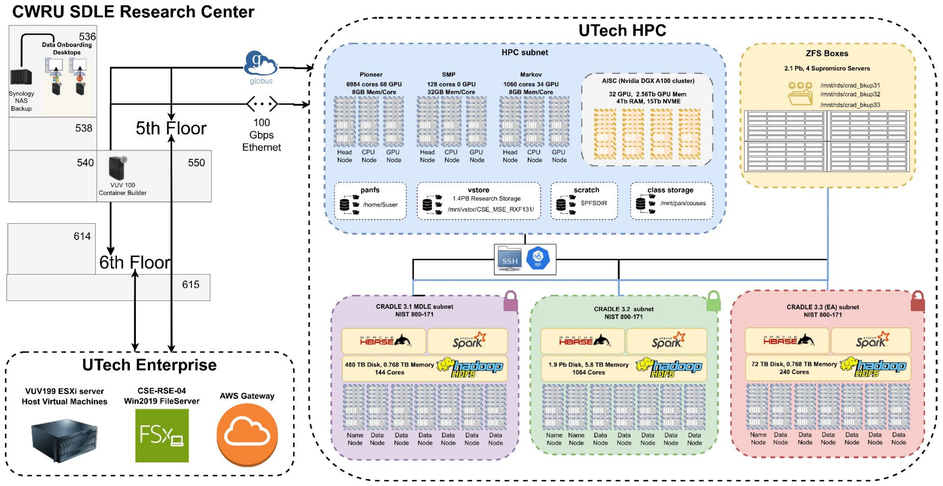
- work on data quality
- complements structured frameworks

All streams of data :

- **Ontologies** created
  - to unify terms in domains
- Facilitates data integration,
  - exchange & access
- **Automated frameworks**
  - created & fully scalable
- Across domains & applications



Distributed Computing: CRADLE[3] with Data Governance



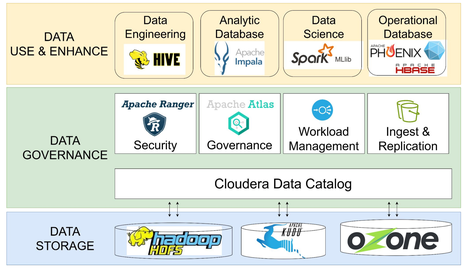
In SDLE Res. Cntr.

- 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

Benchmarking compute task performance on HPC versus CRADLE (Hadoop3) distributed computing, with Spark3 and Nvidia Rapids distributed GPU computing [2]

Domain	Task	HPC	CRADLE: Hadoop Private Cloud
PV:	100K adjacency matrix construction (10 billion distance calculations)	**19 days**, *On 24 CPUs*	**19.1 minutes**, *Spark3, Hadoop*
PV:	100K PV system graph community detection	**8.47 hours**, *On 40 CPUs*	**5.1 minutes**, *Nvidia RAPIDS*
PV:	Query 5,000 PV systems power, from 2.6 billion rows	**N/A**, *Out-of-Memory for 250 GB RAM*	**~1 hour**, *Impala, Spark3, Hadoop*
Image Conversion:	10,000 `.ibw` image files to `.tiff`	**11.1 hours**	**0.62 hours**, *Spark3*
Image Deep Learning:	Hyperparameter tuning: Train 240 deep learning Unet segmentation models	**5.1 days**	**5.2 hours**, *SDLEFleets*
Image Sequences:	Nearest neighbor crystallite calculations	**3.8 minutes**	**1.8 minutes**, *Nvidia RAPIDS*

Data Governance Tools in the CRADLE Hadoop Platform



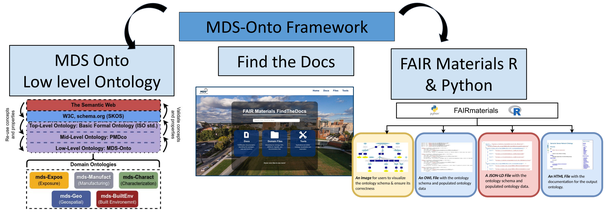
FAIR-Ontology[4]

To ensure interoperability and efficiency in data analysis and modeling pipelines we need Ontologies

- **Ontologies are created for unification of terms in specific domains, which facilitates**
  - Data integration, exchange & access
- Research data Stewardship is increasingly conceptualized in terms of **FAIR Principles**

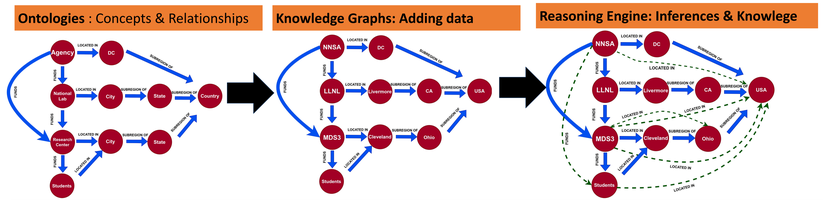
MDS-Onto : A low Level Ontology in the domains of Materials & Data Science

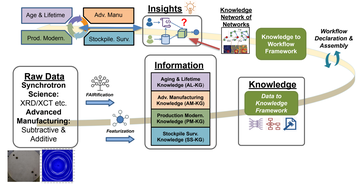
FAIR Enables Semantic Reasoning and New Knowledge



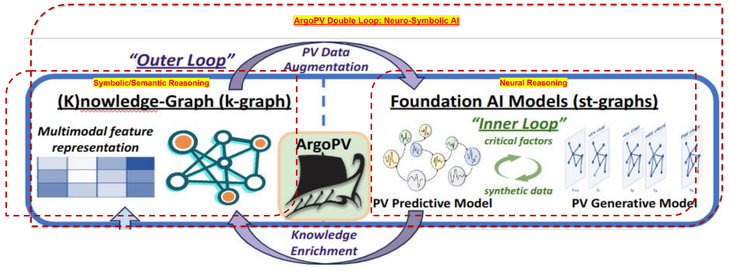
FAIR & Ontologies enable scalable and automated data pipelines

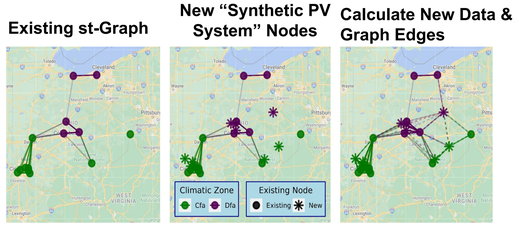
1. Data from multiple sources are ingested ino CRADLE
2. Metadata and Data are FAIRified & Ontologies created
3. FAIR data enable efficient, transparent and reliable data analysis and Modeling
4. Results enable new knowledge and insights





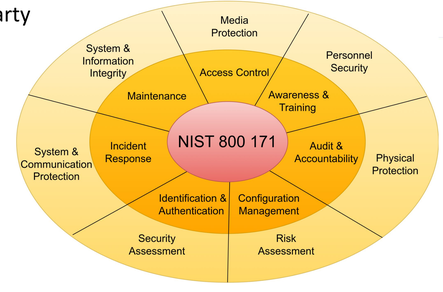
Neuro-Symbolic AI: Semantic Learning from Historical Data

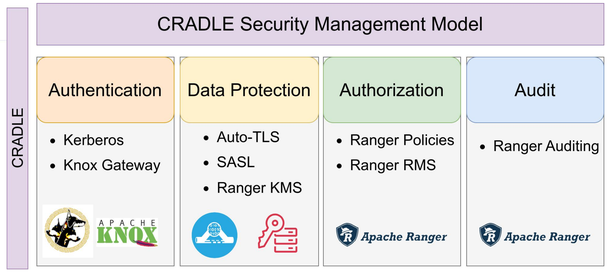




Cybersecurity & CMMC Level 2

- Ongoing third-party assessments
- Ongoing self-assessment
- 110 practices aligned with NIST SP 800-171





References

[1] International, Dama., "DAMA-DMBOK: Data management body of knowledge," Technics Publications, LLC, 2017.
[2] Arafath Nihar *et al.*, "Accelerating Time to Science using CRADLE: A Framework for Materials Data Science," in *2023 IEEE 30th HiPC*, 2023, 234–245.
[3] T. G. Ciardi *et al.*, "Materials data science using CRADLE: A distributed, data-centric approach ," *MRS Commun.*, 2024.
[4] Kristen J. Hernandez *et al.*, "A Data Integration Framework of Additive Manufacturing Based on Fair Principles," *MRS Adv.*, 2024.
[5] W. C. Oltjen *et al.*, "FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data," in *2022 IEEE 49th PVSC*, 2022, 0796–0801.

Acknowledgement

This material is based upon research in the **Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE)**, and supported by the **U.S. Department of Energy's National Nuclear Security Administration** under Award Number(s) **DE-NA0004104**.
This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at **Case Western Reserve University.**

This material is based upon work supported by the National Science Foundation under Grant No. 2133576.