NREL
Transforming ENERGY

Data-Centric AI and the Open Energy Data Initiative (OEDI)

Nicole Taverna

SETO Workshop on Solar Applications of AI and ML

10/31/2023

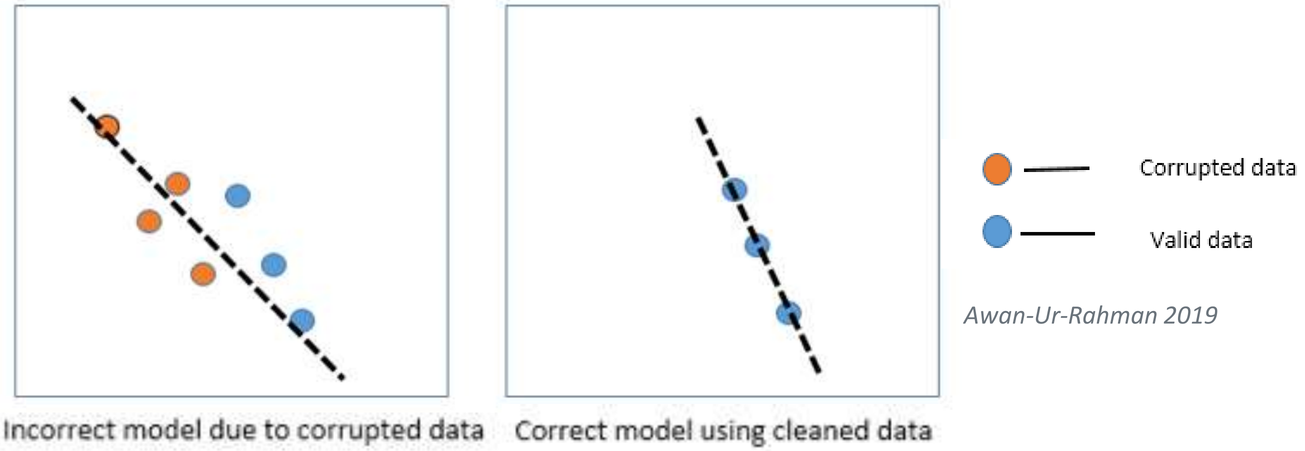
Photo by Dennis Schroeder, NREL 55200

The Importance of High-Quality Data for AI & ML

Data-Centric AI

Limitations of Model-centric AI

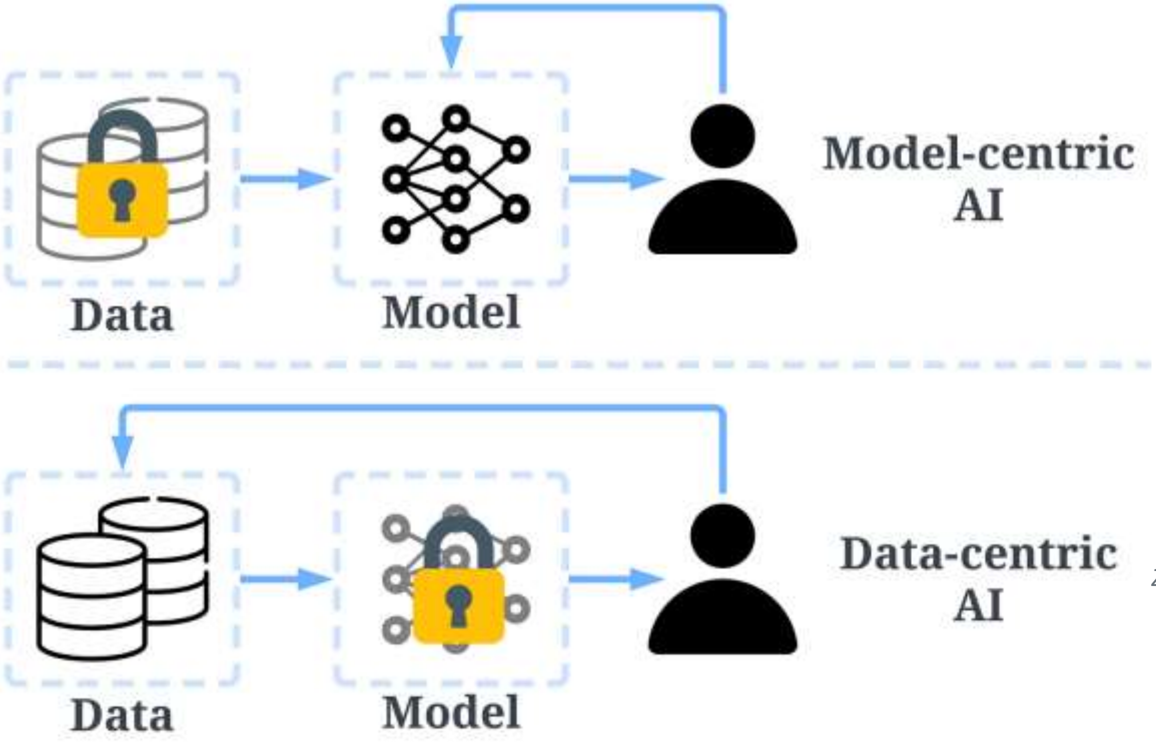
- Limitations of tuning model parameters:
  - **Training on inaccurate data leads to inaccurate results**
  - Insufficient data points → Disappointing outputs
  - Significant number of incorrectly labeled data points → Worse results than when fewer but accurate labels are used



Awan-Ur-Rahman 2019

Data-Centric AI Movement

Movement focused on improving the quality of data used to train models, rather than tuning model parameters, to improve accuracy



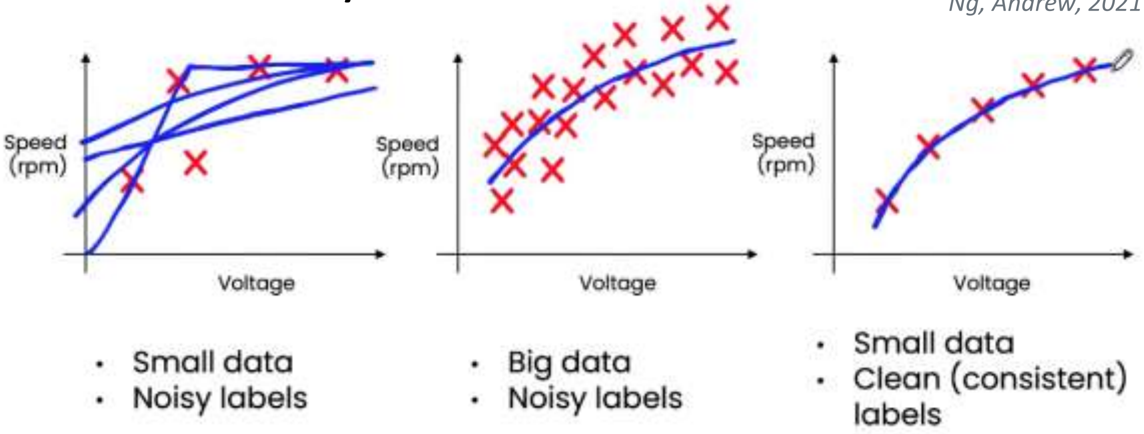
Zha, Daochen et al., 2023

Benefits of Data-Centric AI

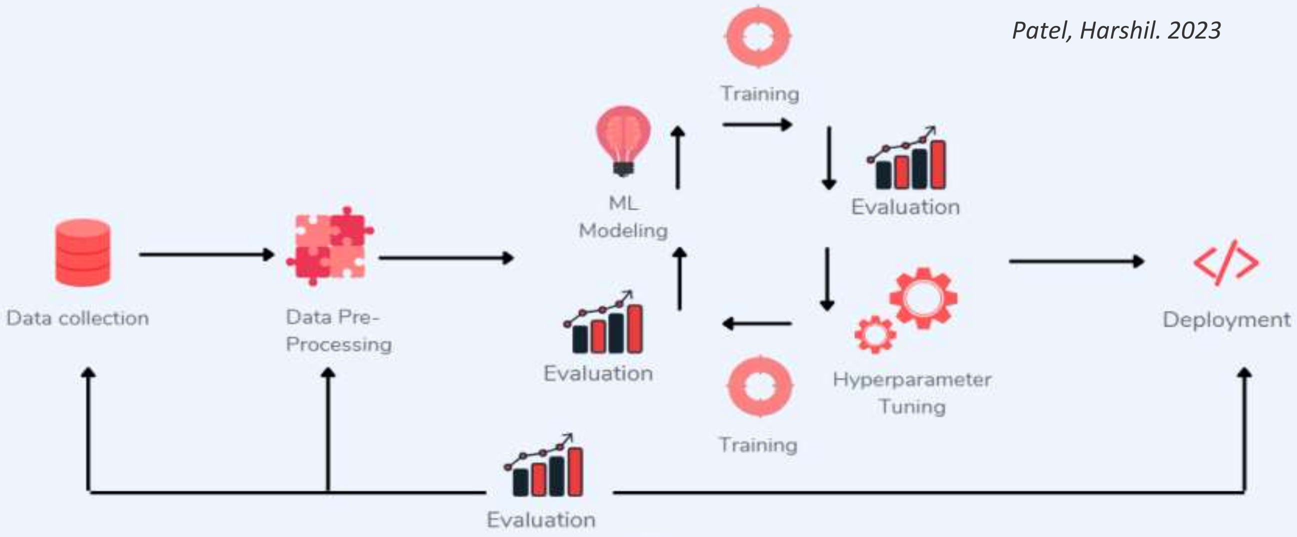
★ **We can be more accurate with less data**

- Reduced data errors and inconsistencies, and improved data reliability
- Better insight into trends helping to interpret results and make better decisions
- Lower overall cost
- Makes data more accessible to key stakeholders
- Reduced data redundancy

Ng, Andrew, 2021



Data-Centric AI Takeaways



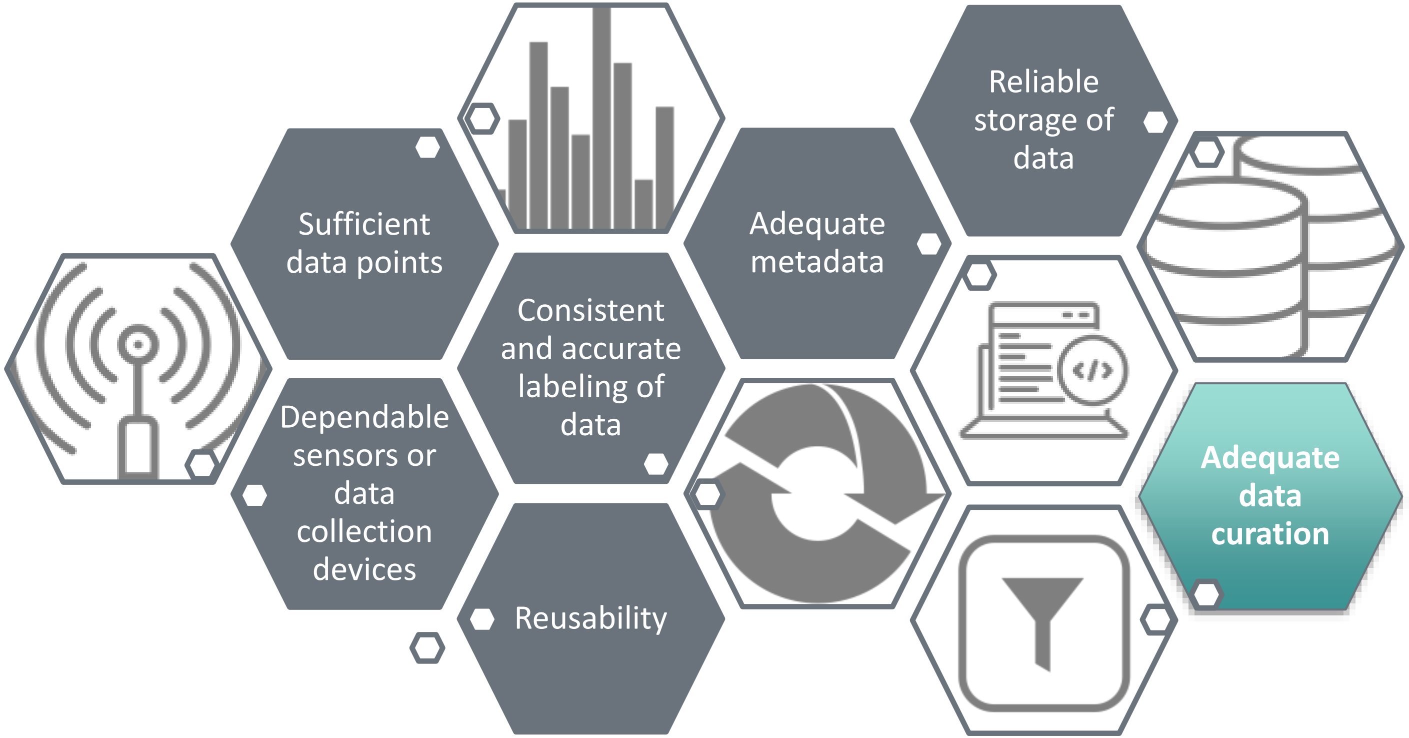
Patel, Harshil. 2023

- The accuracy of your model depends on the quality of your data
  - **Accurate information is needed to make good decisions**
- Best to adopt a **hybrid approach**
  - Considering both the data and the model
- You must have enough data to solve your problem, but **having a large amount of data is a benefit, not a must**

Best Practices for Data Curation

In AI & ML Projects

What Constitutes High Quality Data?

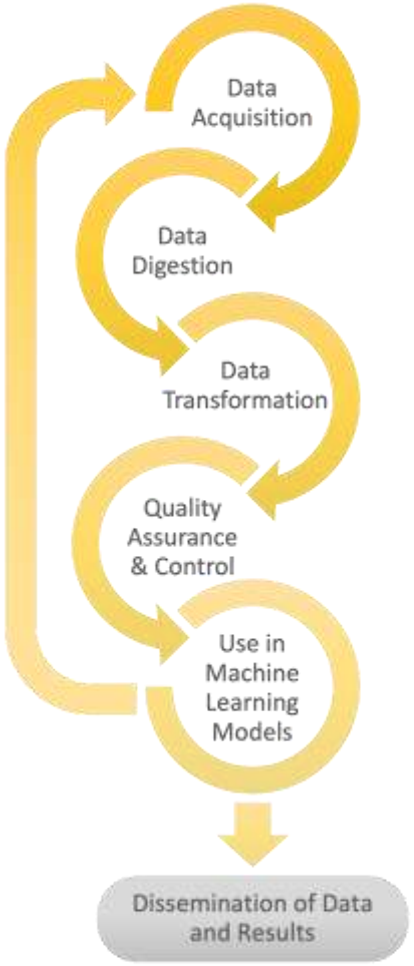


Data quality can be enhanced through adhering to data curation best practices

Data Curation Best Practices

1. **Acquisition** of data from data owners
2. **Digestion** of data to gain an understanding of what is included
3. **Transformation** of data into a machine-readable format
4. **Quality assurance & quality control**
5. **Use in machine learning algorithms**
6. **Repetition** of previous steps until all data needs are met
7. **Dissemination** of curated dataset and ML outputs

Supports a data-centric philosophy with goal of improving real-world applicability of results



GOOML data curation process as outlined in Taverna et al., 2022

Where can you find high quality data?

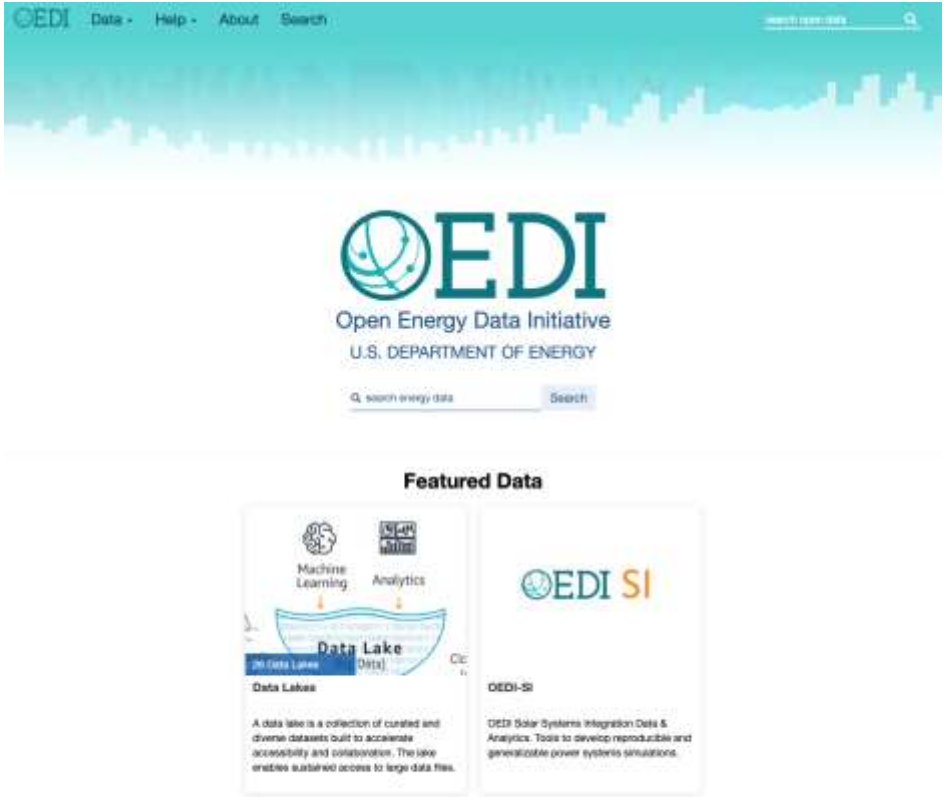
Open Energy Data Initiative (OEDI)

The Open Energy Data Initiative (OEDI)

→ Home to data generated by projects funded by the DOE Solar Energy Technologies Office along with other curated energy data

→ Provides public access to energy-related data sets

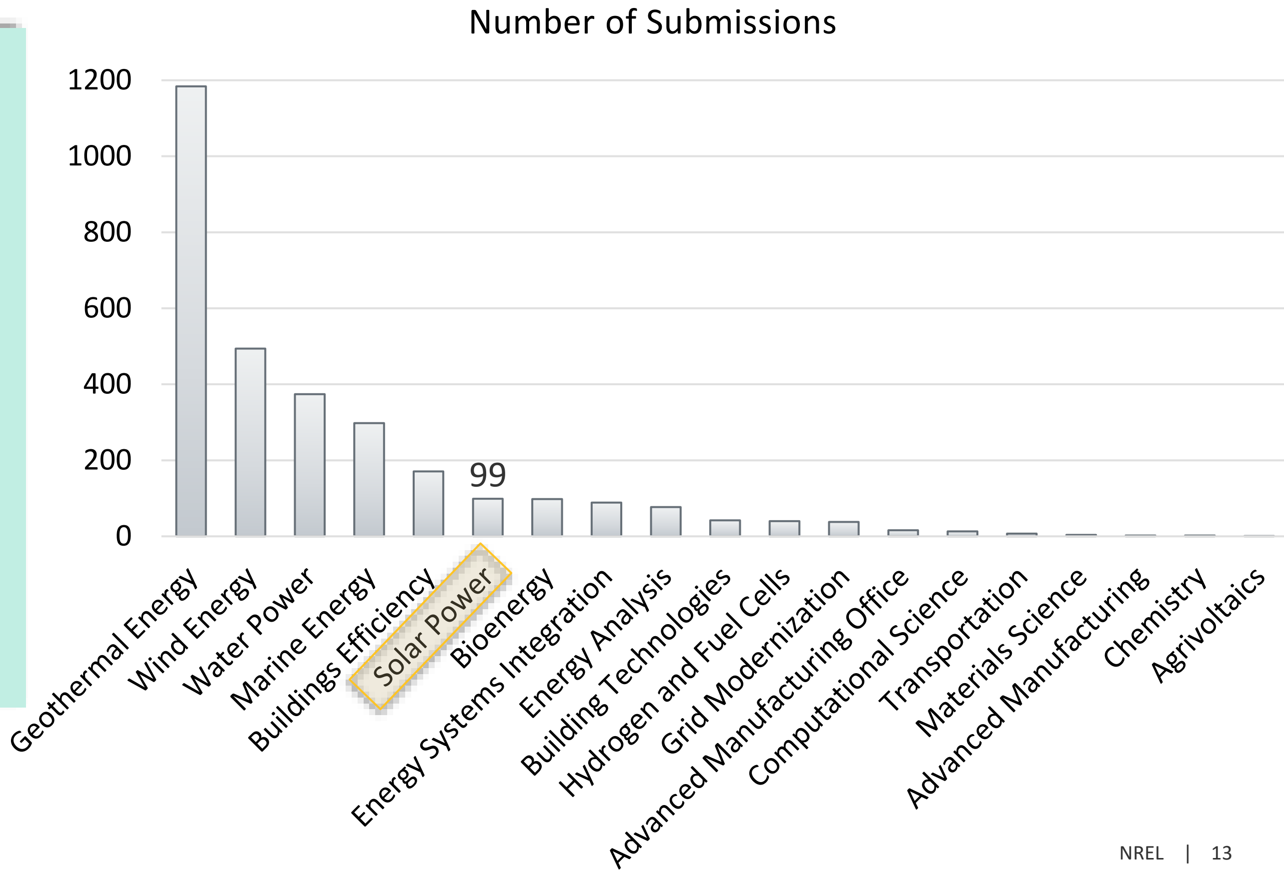
→ Consistently working to improve the convenience and efficiency of using its datasets in ML projects



Read more about the GDR in Weers et al., 2022

What kinds of Data are in OEDI?

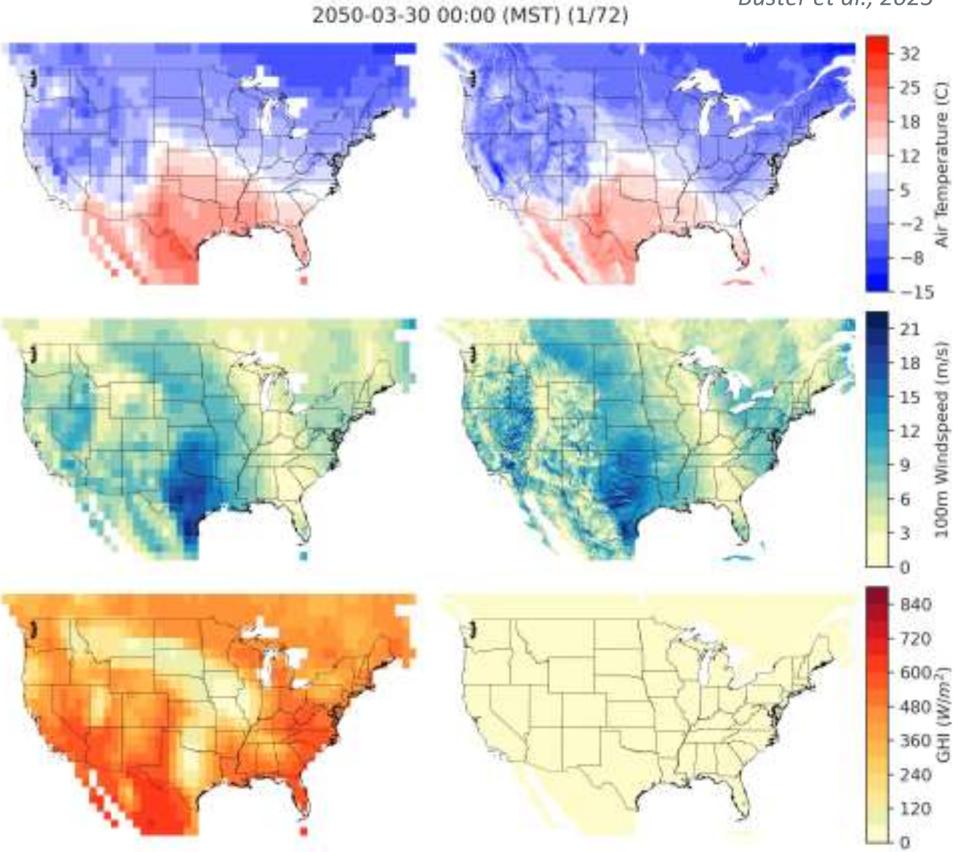
- **1,951 publicly accessible datasets**
- 11,725 total resources
  - includes files, links and APIs
- 2.72 PB of submitted data



Featured Datasets

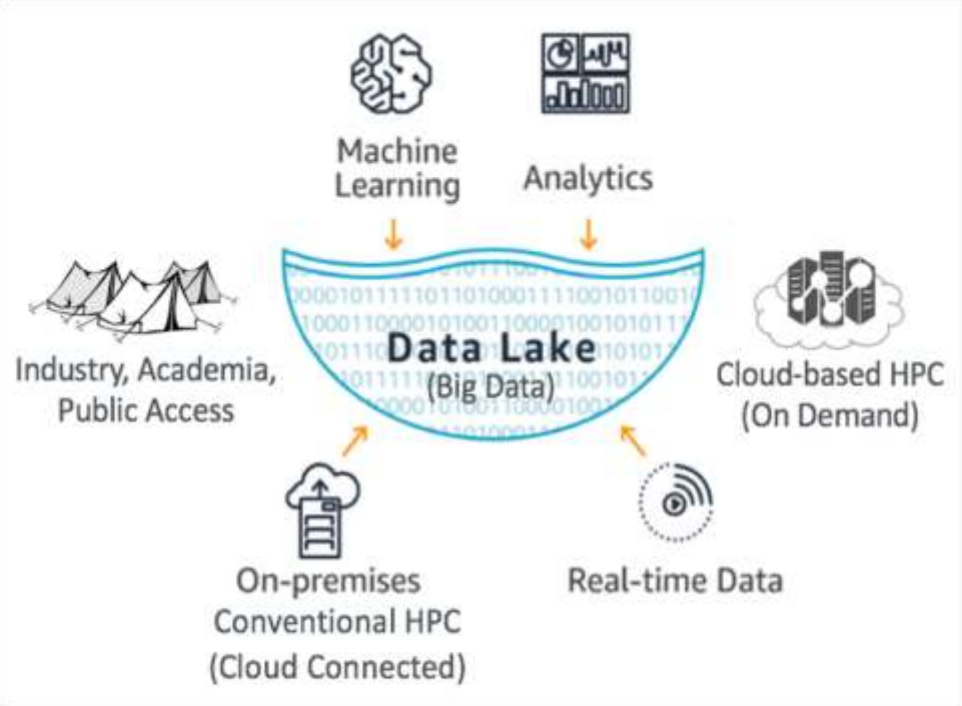
- **Sup3rCC**: 4km hourly wind, solar, temperature, humidity, and pressure fields for the U.S. under climate change scenarios. https://data.openei.org/submissions/5839
- **PVDAQ**: Large-scale time-series db of system metadata and performance data from public PV sites. https://data.openei.org/submissions/4568
- **NSRDB**: Serially complete collection of meteorological and solar irradiance data sets for the U.S. and international locations. https://data.openei.org/submissions/1

Buster et al., 2023



Data Lakes

- Large or complex datasets are stored in the OEDI data lake
- Allows users to query or work with the data without downloading the full dataset
- Integration with cloud-based tools

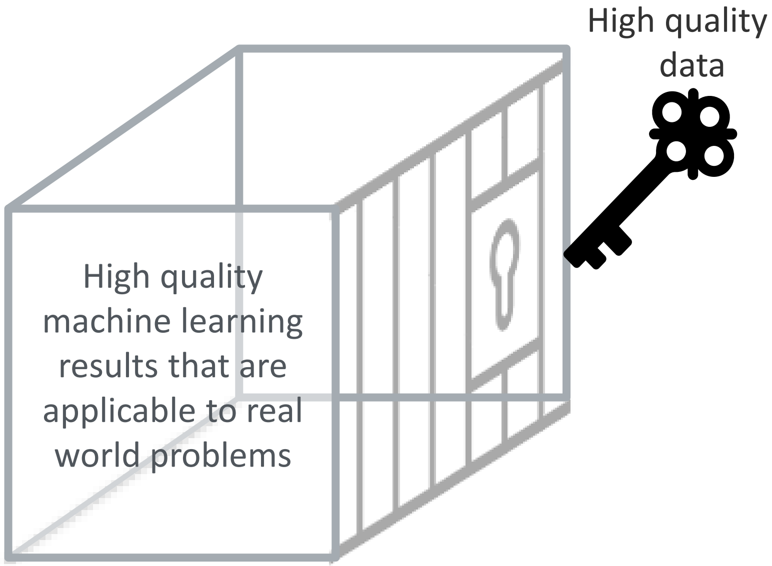


Weers et al., 2021

Big Picture and Conclusions

Conclusions

- Data-centric AI allows you to obtain **better results with less data**
- Adhering to best practices for **data curation** can help improve the quality of your data
- ML is exploratory in nature, meaning that **data curation is often iterative**
- **OEDI** is a great starting place for obtaining **high quality data**



Taverna et al., 2023

Q&A

www.nrel.gov

NREL/PR-6A20-87936

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office and Geothermal Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Photo from iStock-627281636

NREL Transforming ENERGY

References

- Patel, H., 2023. “Data-Centric Approach vs Model-Centric Approach in Machine Learning.” MLOps Blog. https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning
- Pres, G. “Andrew Ng Launches A Campaign For Data-Centric AI.” Forbes. https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=690c4dd674f5
- Taverna, N., Buster, G., Huggins, J., Rossol, M., Siratovich, P., Weers, J., Blair, A., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., and Akerley, J. “Data Curation for Machine Learning Applied to Geothermal Power Plant Operational Data for GOOML: Geothermal Operational Optimization with Machine Learning.” Proceedings of the 47th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA.
- Taverna, N., Weers, J., Huugins, J., Porse, S., Anderson, A., Frone, F., Scavo, R.J. 2023. “Improving the Quality of Geothermal Data Through Data Standards and Pipelines Within the Geothermal Data Repository.” Proceedings of the 48th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA.
- Taverna, N., Weers, J., Porse, S., Anderson, A., Frone, Z., Holt, E. 2023. “An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data.” Proceedings of the 49th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA.
- Weers, J., Porse, S., Huggins, J., Rossol, M., and Taverna, N. “Improving the Accessibility and Usability of Geothermal Information with Data Lakes and Data Pipelines on the Geothermal Data Repository.” GRC Transactions, Vol. 45 (2021).