

:dvs ny

ο**ι** (=x^p

f h o c[\ i i K l

R:g D

/ | ~ X I

## **The Data Pyramid**

Prediction Decision support

Everyone wants to jump to the top of the pyramid...

**Optimize** 

AI

ML

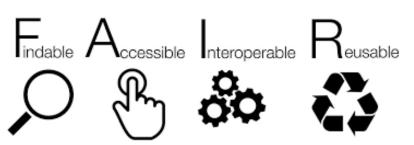
Test and validate

Data analytics tools and best practices must be shared.

Researchers spend ~80% of their time acquiring, cleaning, & organizing data.

**Curate and Label** 

Clean, transform, label, visualize, identify features



## **Store and Share Data**

Clean and organize data, metadata, workflow for cloud computing

## **Collect Data**

Instruments, sensors, simulations, lab reports & literature.

Automate data collection (Data Hub).

... but these steps at the bottom must come first.



Ram Devanathan, ram.devanathan@pnnl.gov



dvs ny

R:g D

/ | ~ X I

## Challenges when Al meets domain science

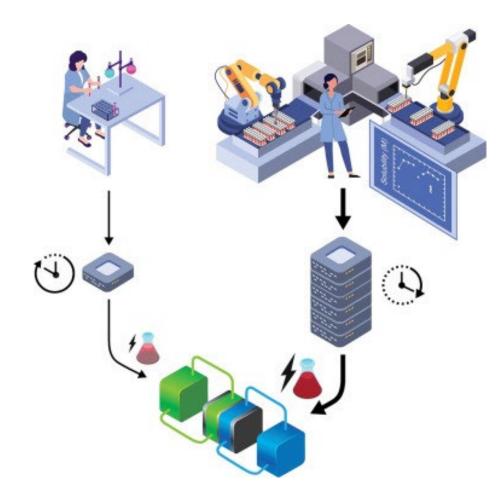
- 1. Small & expensive data.

  Limited data relevant to energy storage.
- 2. Data quality and uncertainty.
- 3. Complex processing-structure-propertyperformance relationships.
- 4. Use of black box models.
- 5. Sloppy use of ML + hype.

Solutions: Data curation, data sharing, few shot ML, emphasize statistical learning and physical basis of results. Domain-specific foundation models.

Dimensions of data quality: Completeness, accuracy, usability, and standardization.

M. Wenzlick, J. R. Bauer, K. Rose, J. Hawk, and R. Devanathan. Integrating Materials and Manufacturing Innovation 9 (2020) 89.



Y. Liang et al, High-throughput solubility determination for datadriven materials design and discovery in redox flow battery research. Cell Reports Physical Science 4 (2023) 10.