ME 5374-ST

Machine Learning for Materials Science and Discovery

Fall 2025

Asst. Prof. Peter Schindler

Lecture 1 - Introduction

- Introductions
- Course Logistics
- Why Data-driven Materials Discovery?
- Materials Informatics in Industry

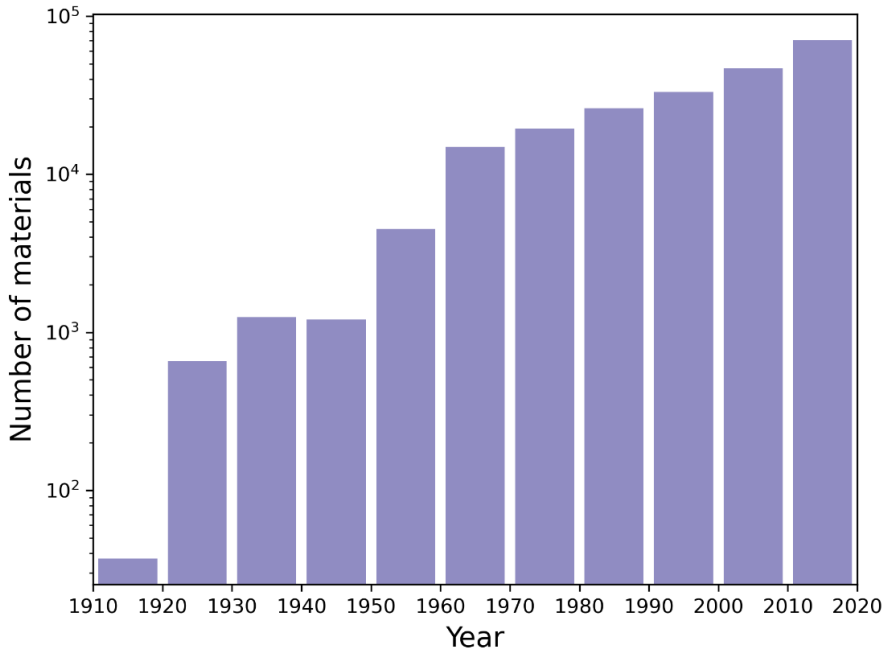
Materials-Discovery over Time

Of all (solid state) materials that we know of today,
how many were discovered in the **last 10 years**?

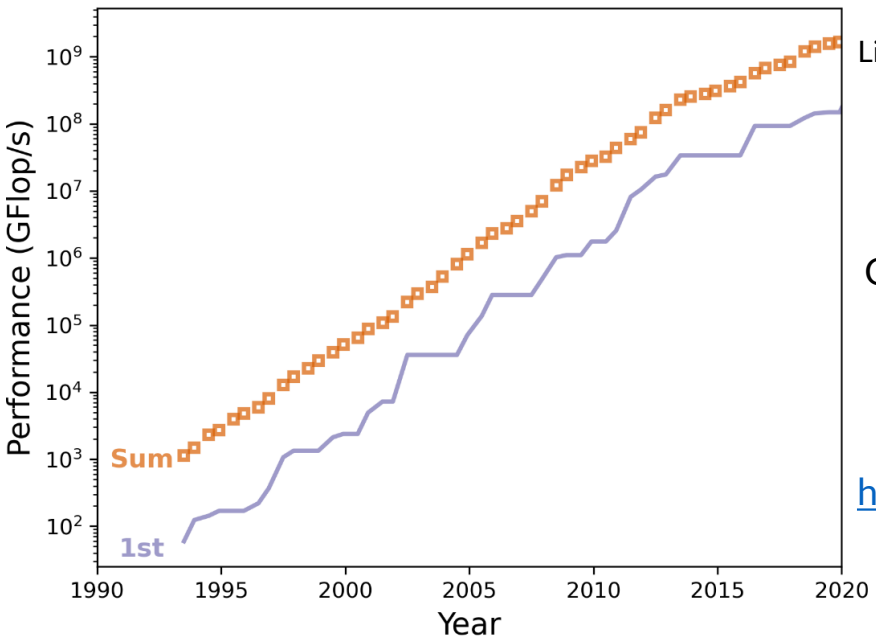
- 1 %
- 10%
- 33%
- 66%
- 99%

Materials-Discovery over Time

Materials in ICSD



Top 500 HPC



"El Capitan"
Livermore, Los Alamos, and Sandia NLs

Exascale HPC

Exaflop = Quintillion 64-bit operations per second

Check out https://top500.org

Doubles every ~22 yrs

Doubles every ~1.3 yrs

1st : Empirical Science	2nd: Model-based Science	3rd: Computational Science	4th: Data-driven Science
Experiments	**Physical Laws**	**DFT, MD**	**ML, Clustering**

Welcome to ML4MSD

- **Canvas**: https://northeastern.instructure.com/courses/229749
- **Website**: https://ml4msd.d2r2group.com
- **Time**: Tuesday 11:45 – 1:25 and Thursday 2:50 – 4:30
- **Location**: Ryder Hall 153
- **Credit**: 4 semester-hours
- **Instructor**:
  Asst. Prof. Peter Schindler ("Prof. Peter")
  Office: 275 SN
  E-mail: p.schindler@northeastern.edu

My Academic Background

Originally from Austria, Vienna
M.Sc. & Ph.D. in **Physics** from University of Vienna

universität wien

Postdoc at **Stanford University** (EE/MatSci)

Senior scientist at **Aionics Inc.** (Materials AI)

AIONICS

My Group: Data-Driven Renewables Research

Using *high-throughput density functional theory* calculations and *data-driven materials property prediction* to discover new materials for renewable energy applications.

Surface science of thin-film growth, semiconductor fabrication, and heterogeneous catalysis.

Data Driven Renewables Research

www.d2r2group.com

Course Outline (Tentative)

Week	Day	Topic
1	1	Introduction and Getting Started (GitHub and VSCode)
2	1	Python Crash Course
2	2	Data in Python: Numpy, Pandas, and Matplotlib
3	1	Machine Learning Basics 1
3	2	Machine Learning Basics 2
4	1	Types of Data in Materials Science
4	2	Crystallography Crash Course and Pymatgen
5	1	Object-Oriented Programming Illustrated with Pymatgen
5	2	Working with APIs (Materials Project)
6	1	*Project Discussions*
6	2	Computational Materials Properties 1
7	1	Computational Materials Properties 2
7	2	Featurization of Materials 1
8	1	Featurization of Materials 2

Course Outline (Tentative)

8	2	Deep Learning 1
9	1	Deep Learning 2
9	2	Machine Learning Interatomic Potentials (MLIPs)
10	1	Large-Language Models (LLMs) and Generative AI in Materials Science
10	2	**Student Paper Presentations**
11	1	Veteran's Day (no class)
11	2	**Student Paper Presentations**
12	1	*Guest Speaker*
12	2	*Guest Speaker*
13	1	*Guest Speaker*
13	2	Thanksgiving recess (no class)
14	1	*Guest Speaker*
14	2	**Final project presentations (in-class)**
15	1	**Final project presentations (in-class)**

Learning Outcomes

By the end of this course, students will be able to:

- Explain the **fundamentals of ML** and its relevance to materials science.
- Analyze and manipulate **materials datasets** with *Numpy*, *Pandas*, and *Matplotlib*.
- Use **Pymatgen** to analyze and manipulate crystal structures.
- Access and use APIs like the **Materials Project** for data retrieval.
- Identify both experimental/computational **data types** in materials science.
- Perform **featurization** of materials data for use in ML models.
- Interpret the output of ML models in the context of **materials discovery**.
- **Train ML models** with packages like *scikit-learn* and *PyTorch*.
- Design and implement **projects** applying ML to real-world materials problems.
- Explore the application of **deep learning**, MLIPs, and genAI in materials science.
- Build foundational **Python** programming skills, including **OOP**.
- Use *GitHub* and *VSCode* for version control, and project development.

Coding Skill Requirements?

- **No formal prerequisite** in Python (or other) coding experience
  However, prior experience is beneficial
- Python has a **forgiving learning curve** and lightweight syntax
- *MATLAB* is somewhat similar, with a few key differences
- Python is the **de facto standard for ML** and open-source software development
- The first 1-2 weeks of this course aim to **cover** (most) **coding basics**
- *Resources* (next slide) will be critical to keep up with the course
- Course schedule tentative – can be **adjusted based on student needs**

Resources

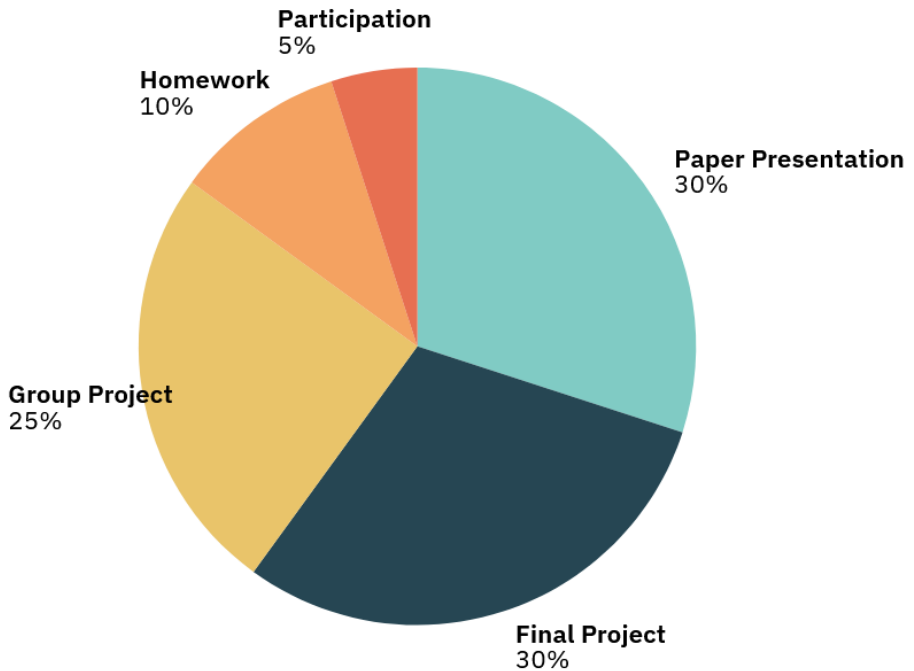
Materials Informatics is a new and rapidly evolving field, and hence, conventional textbooks are often not an ideal place to start.
This course **will not follow a specific textbook** but will rely on prior efforts of various aspects of this field (ML, Python, data science, materials science, etc.) and **free** online resources:

- *Machine Learning in Materials Science*, Keith T. Butler, Felipe Oviedo, and Pieremanuele Canepa, ACS (2022). (free with NEU login)
- *Machine Learning Refined*, Jeremy Watt, Reza Borhani, and Aggelos K. Katsaggelos, 2nd Edition (2024). (free)
- *Understanding Deep Learning*, Simon J.D. Prince, The MIT Press (2023). (free)
- *Deep Learning for Molecules and Materials*, Andrew D. White, Living Journal of Computational Molecular Science (2021). (free)

Further detailed resources and readings will be shared alongside the course.

Course Grade

- 5% in-class participation (in person)
- 10% homework assignments
- 30% paper presentation
- 25% group project
- 30% final project (individual)



Homework Assignments

Designed to *complement* and deepen skills taught during class

Typically, short coding assignments and/or additional reading assignments

Graded for **effort** (not correctness):

- Not submitted/minimal effort: 0%
- Student tried: 50%
- Effort is evident: 100%

Homework Due:
1 week after it was posted at 11:59 pm

Late Homework Policy:

All assignments will be accepted up to 2 calendar days late (cutoff 11:59 pm).
- Assignments turned in *one day late* will be **penalized by 10%**.
- Assignments turned in *two days late* will be **penalized by 20%.**
- After two calendar days, assignments will not be accepted, and you will receive a 0 on the assignment.

The lowest homework score will be dropped.

Projects

Group Project

- Analysis of generalization performance of materials science ML models on standard datasets
- Topic is clearly defined - students only choose a specific model and/or dataset
- All students carry out a very similar workflow
- Results will be gathered at the end and published (ideally until summer 2026)
- *Deliverables*: Results in an agreed-upon standard format (likely JSON); 2-3 page summary

Final Project (Individual)

- Each student chooses their own project topic (I'll help with suggestions)
- Ideally, the topic is related to their current/past research
- *Deliverables*: Project report (~3 pages) and In-class presentation

Office Hours

- Once a week in my office: SN 275 (2nd floor)
- Format: In-person with *optional hybrid option*
- Individual virtual meetings upon email request

- Office hours day/time: **Thursday 4:30 – 5:30 pm**

Academic Integrity

Examples of academic integrity violations:

- Copying another student's homework solutions (a reasonable amount of collaboration among students is permitted, however, not direct copying) or from online solutions
- Not citing sources/figures appropriately when writing the project paper
- Read the *Utilization of Generative AI* section of the syllabus

Preparations for Next Week

The majority of classes will have:
A lecture component *and* an **interactive coding component**.

Please,

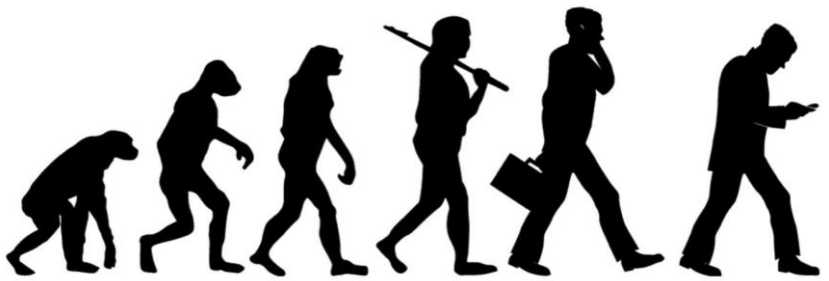
- bring your laptop!
- install VSCode: https://code.visualstudio.com/
- install Python 3.11.9: https://www.python.org/downloads/release/python-3119/
- create a GitHub account (and enable 2-factor authentication) – it's free and **email me your GitHub username**
- install GitHub Desktop: https://desktop.github.com/download/

Canvas: Course Announcements, Grades, and Homework
GitHub: All Other Files (Code, Slides, Readings)

Materials Discovery Precursor to Progress in Society

Materials *Changed* Societies and *Enabled new Technology*:

Stone → Bronze → Iron → … → Silicon Age



wikimedia.org, Google images

14 Grand Challenges for Humanity in the 21st Century

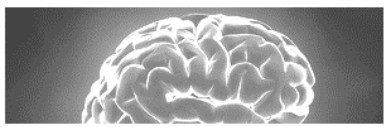
ENGINEER THE TOOLS OF SCIENTIFIC DISCOVERY

MAKE SOLAR ENERGY ECONOMICAL

ENHANCE VIRTUAL REALITY

ADVANCE HEALTH INFORMATICS

REVERSE-ENGINEER THE BRAIN



NAE GRAND CHALLENGES FOR ENGINEERING
NATIONAL ACADEMY OF ENGINEERING

SECURE CYBERSPACE

DEVELOP CARBON SEQUESTRATION METHODS

ENGINEER BETTER MEDICINES

PROVIDE ACCESS TO CLEAN WATER

ADVANCE PERSONALIZED LEARNING

RESTORE AND IMPROVE URBAN INFRASTRUCTURE

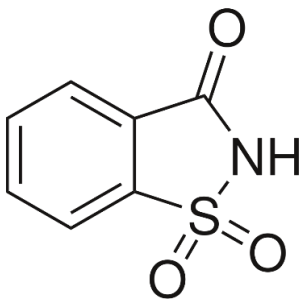
PROVIDE ENERGY FROM FUSION

PREVENT NUCLEAR TERROR

MANAGE THE NITROGEN CYCLE

How are Materials Discovered?

By Luck / Accident: Stainless steel, vulcanized rubber (car tires), Teflon, Play-doh, Saccharin, Super Glue,...



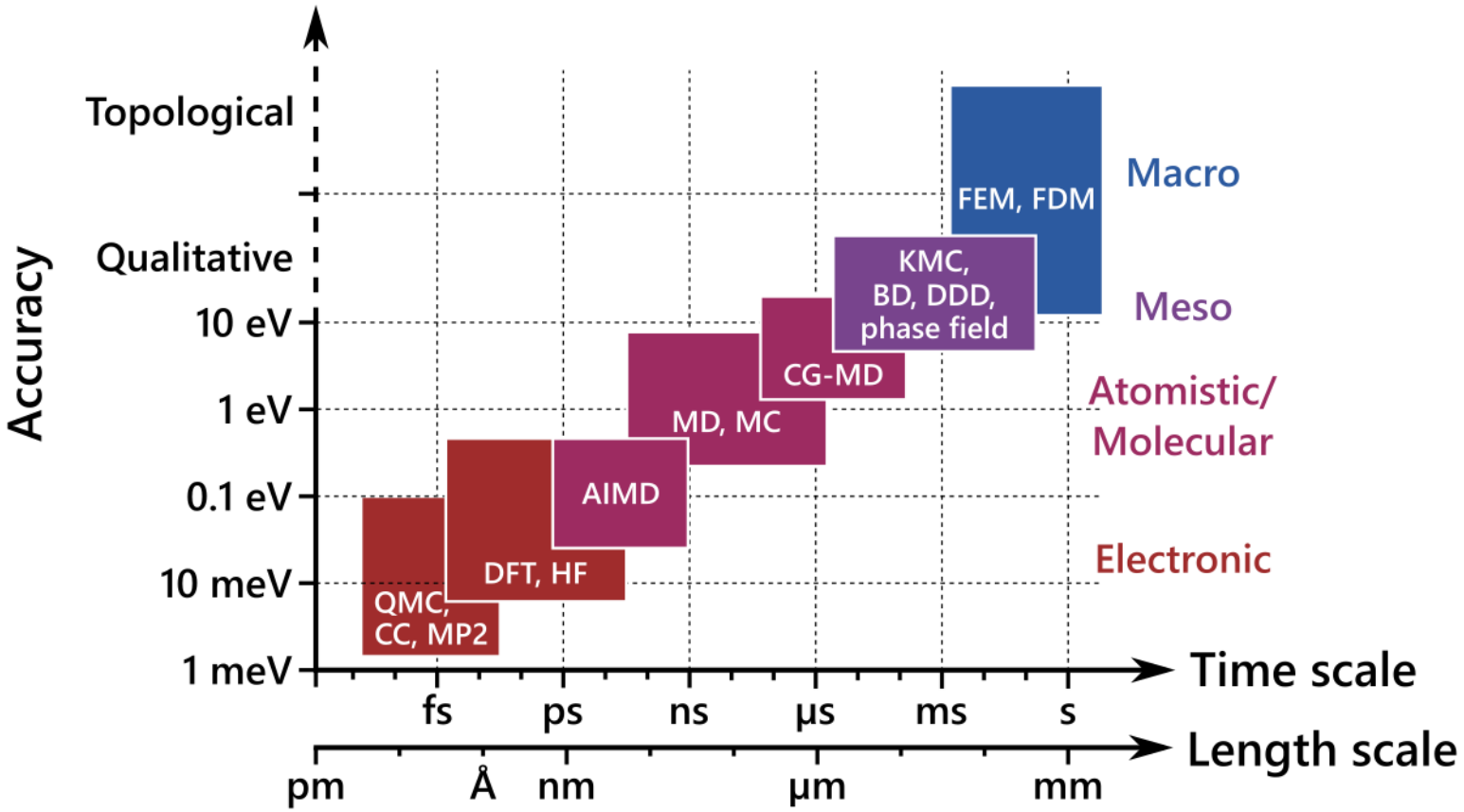
Edisonian (Trial and Error) Approach:

He tested **over 6,000 plant materials** to discover the final light bulb filament

1st : Empirical Science	2nd: Model-based Science	3rd: Computational Science	4th: Data-driven Science
Experiments	**Physical Laws**	**DFT, MD**	**ML, Clustering**

wikimedia.org, Google images

Computational Materials Science: Accuracy / Scale Tradeoff



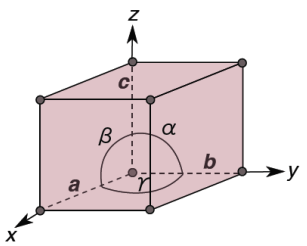
The 3rd Paradigm: Computational Discovery

$$\left[-\frac{\hbar\nabla^2}{2m}+V_{\text{nuc}}(\vec{r})+\sum_j\int \mathrm{d}\vec{r}'\,\frac{\rho(\vec{r})}{|\vec{r}-\vec{r}'|}+\frac{\delta E_{\text{XC}}[\rho(\vec{r})]}{\delta\rho(\vec{r})}\right]\psi_i(\vec{r})=E_i\cdot\psi_i(\vec{r})$$

"ab-inito"

Structural

Lattice constants
Bond length

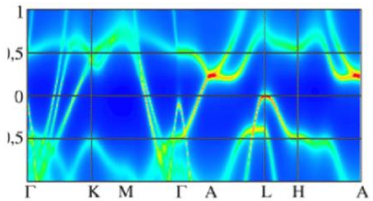


Mechanical

Bulk modulus
Stress tensor σ_{ij}

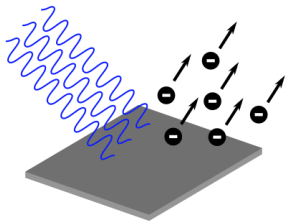
Optical / Electrical

Dielectric constant
Absorption spectra
Density of states
Band structure



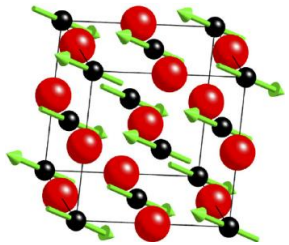
Surface

Work function
Surface/cleavage energy
Adsorption energy



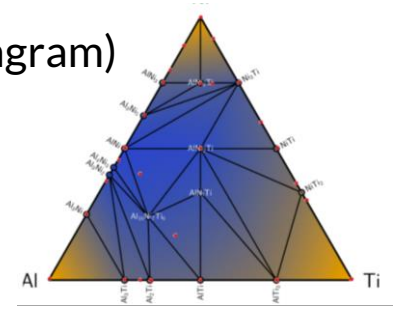
Magnetic

Magnetic ordering
Magnetic moment



Thermodynamic

Vibrational entropy
Phase stability (Hull diagram)



Time Required for Experiment vs. Computation

Experiments (Synthesis) ~ **weeks to months** (to a Ph.D.)

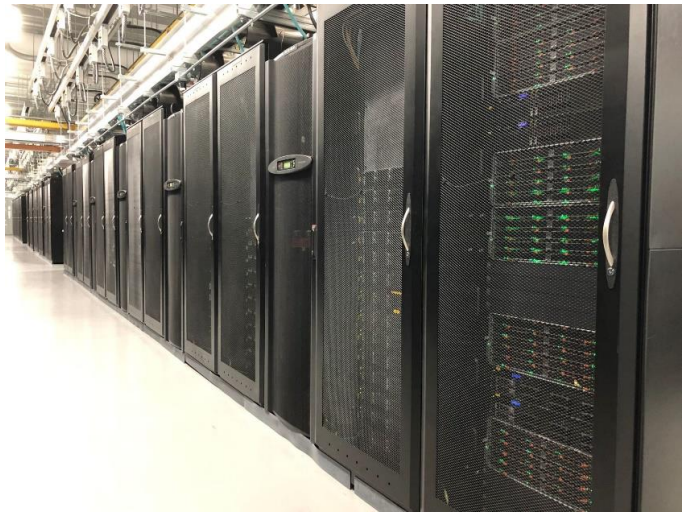
First principles calculations ~ **hours to days** (to weeks)

Still too long to screen
>100,000 candidates

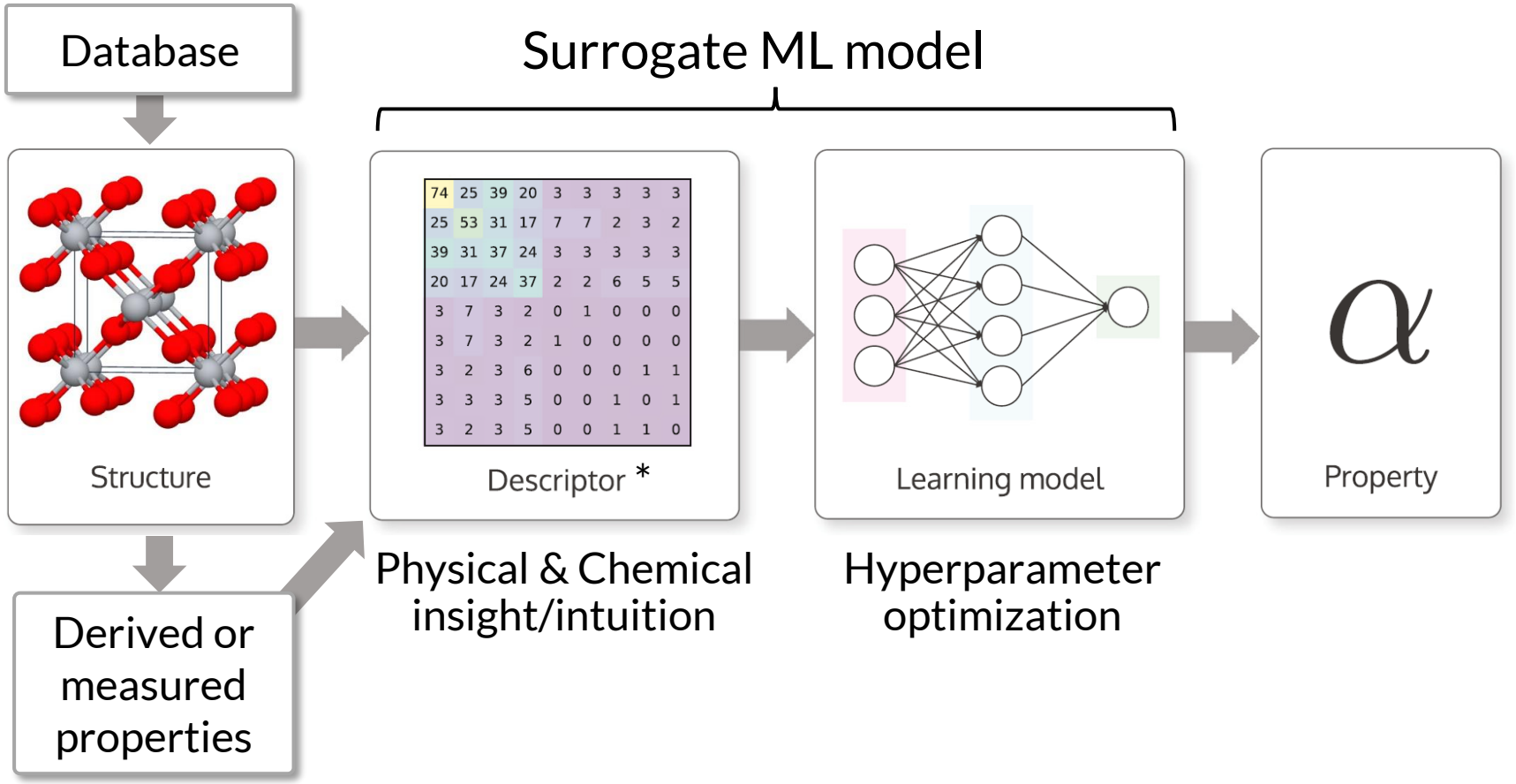
Discovery Cluster
Over 20,000 CPU Cores and Over 200 GPUs

Hosted at MGHPCC (90,000-square-foot facility)

MGHPCC
MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER



The 4th Paradigm: Data-Driven Discovery



* = fingerprint = feature (vector)

Adapted from L. Himanen, et al. *Comp. Phys. Comm.*, *247*, (2020).

High-Quality Data?

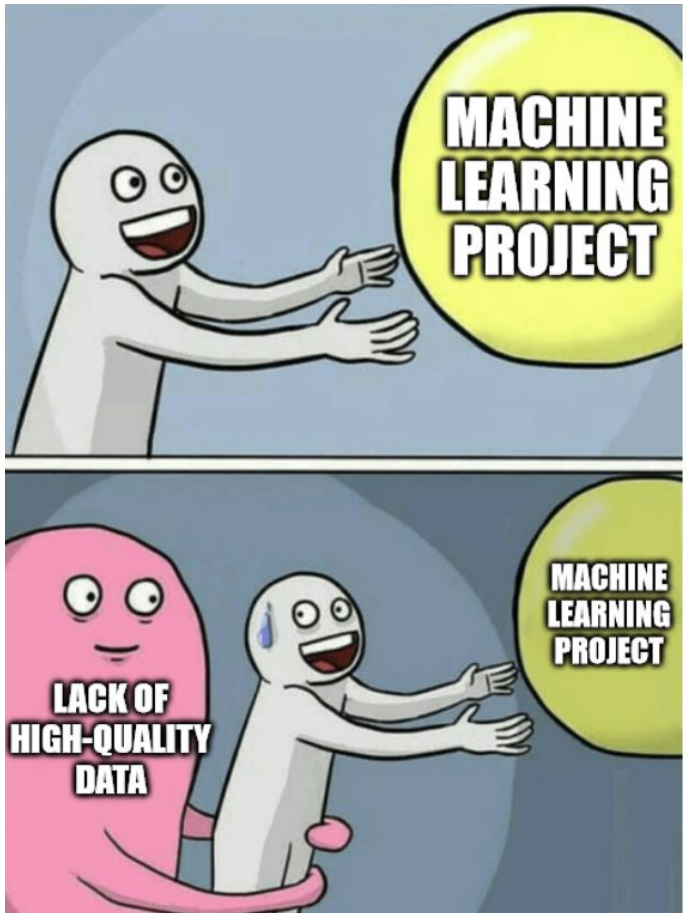
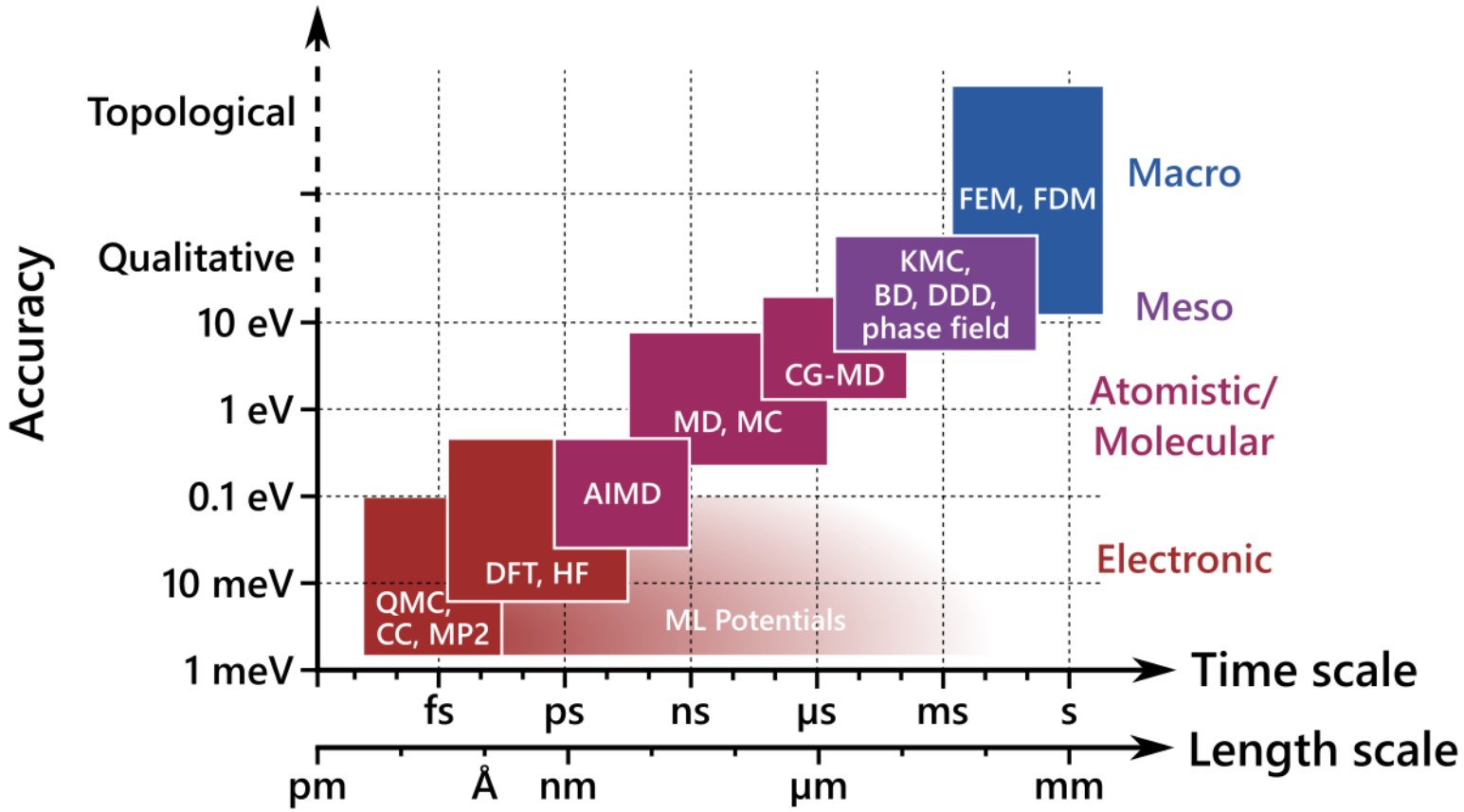


Image Credit: The Internet…

“Affordable Accuracy”



"*Materials Informatics*" in Industry?

Google Research

TOYOTA RESEARCH INSTITUTE

amazon | science

RIVIAN

TESLA

Meta

NVIDIA

"*Materials Informatics*" in Industry?

MateriaChem

Lecture Feedback

Please, scan the QR code and take a minute to let me know how the lecture was and mention any **feedback/questions**

This form is **anonymous!**