Machine Learning for Materials Science

An Introduction

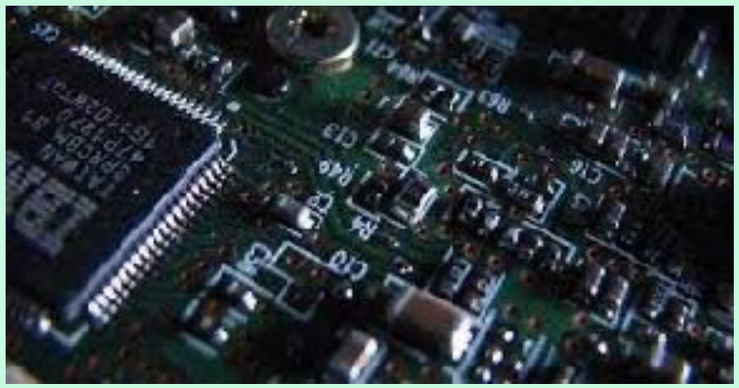
Profs. Prasenjit Sen (HRI) & Subhankar Mishra (NISER)

Materials and civilization

Stone Age

Bronze Age

Historical ages



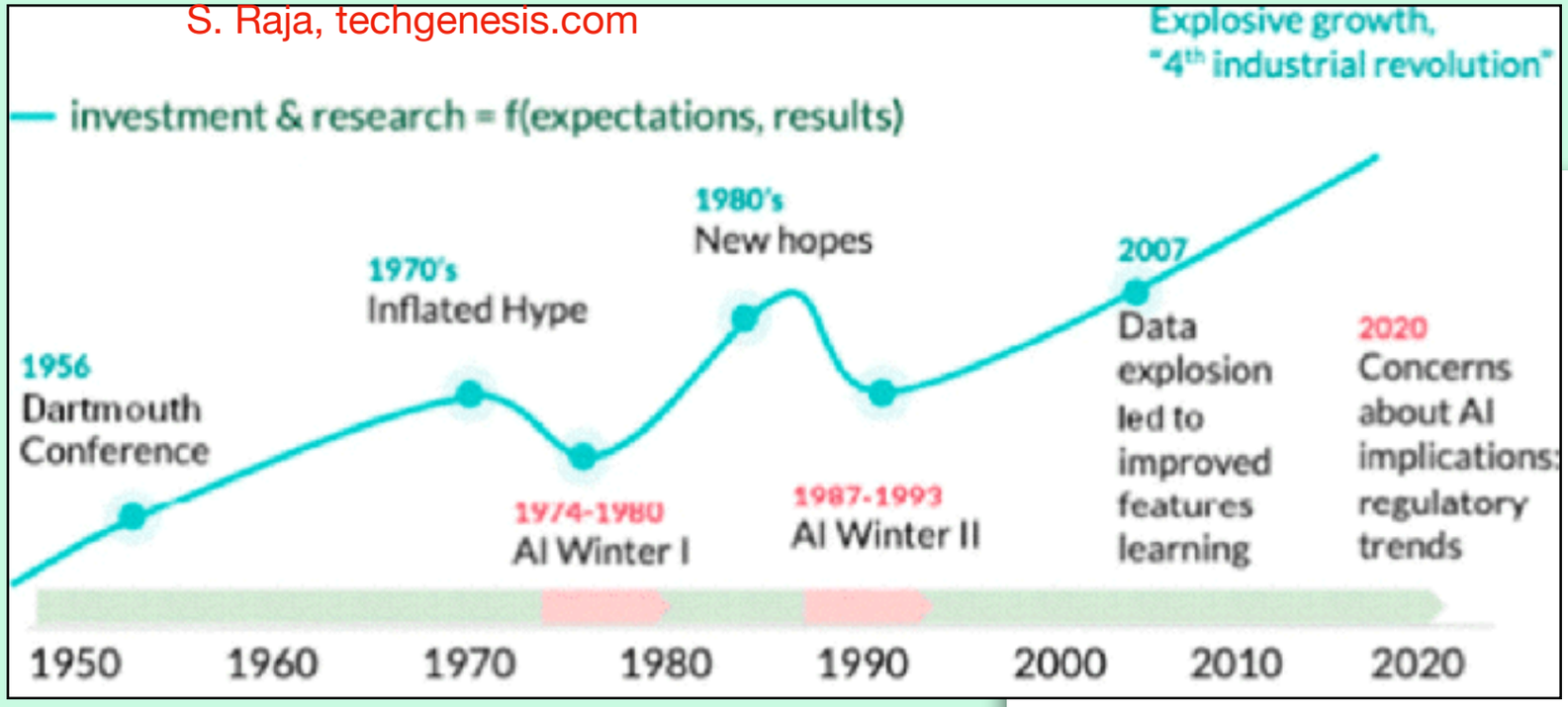
Silicon Age

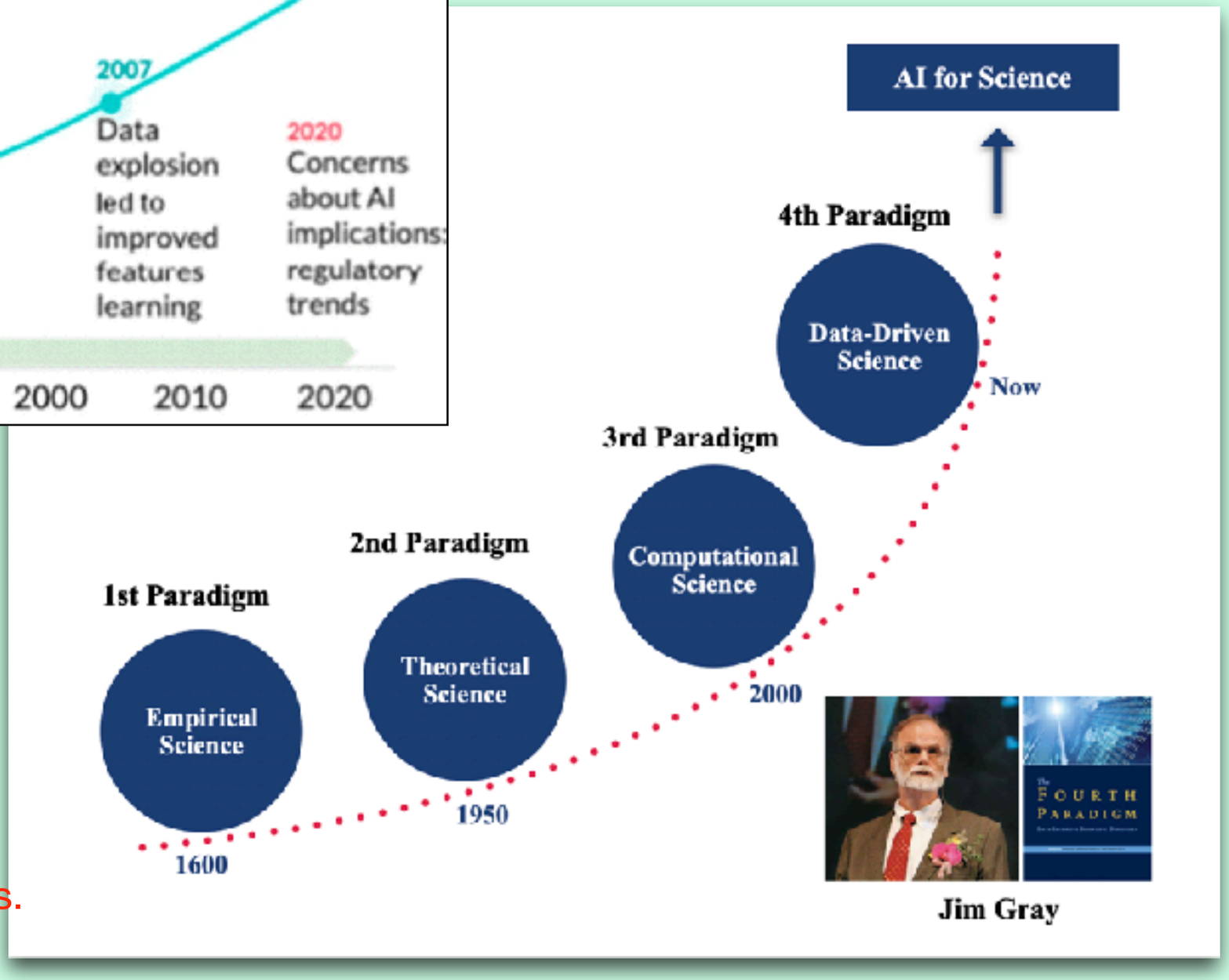
Iron Age

What next ???

Shifting paradigms of science
& the ‘seasons’ of AI research

AI History: The First Summer and Winter of AI
S. Raja, techgenesis.com





- Dartmouth conference (1956)
  - John McCarthy, ‘father of AI’.
- Natural Language Processing (NLP) (‘**Student**’, Bobrow 1964); **ELIZA** (Weizenbaum 1966); machine translation.
- Media hype, slow progress.
- Lighthill report (1973, British Science Res. Council): ‘In no part of the field have discoveries made so far produced The major impact that was then promised’.

Traditional approach to materials design

Approach 0:
Trial-and-error experimentation

Slow pace, uncertain, expensive, often serendipitous

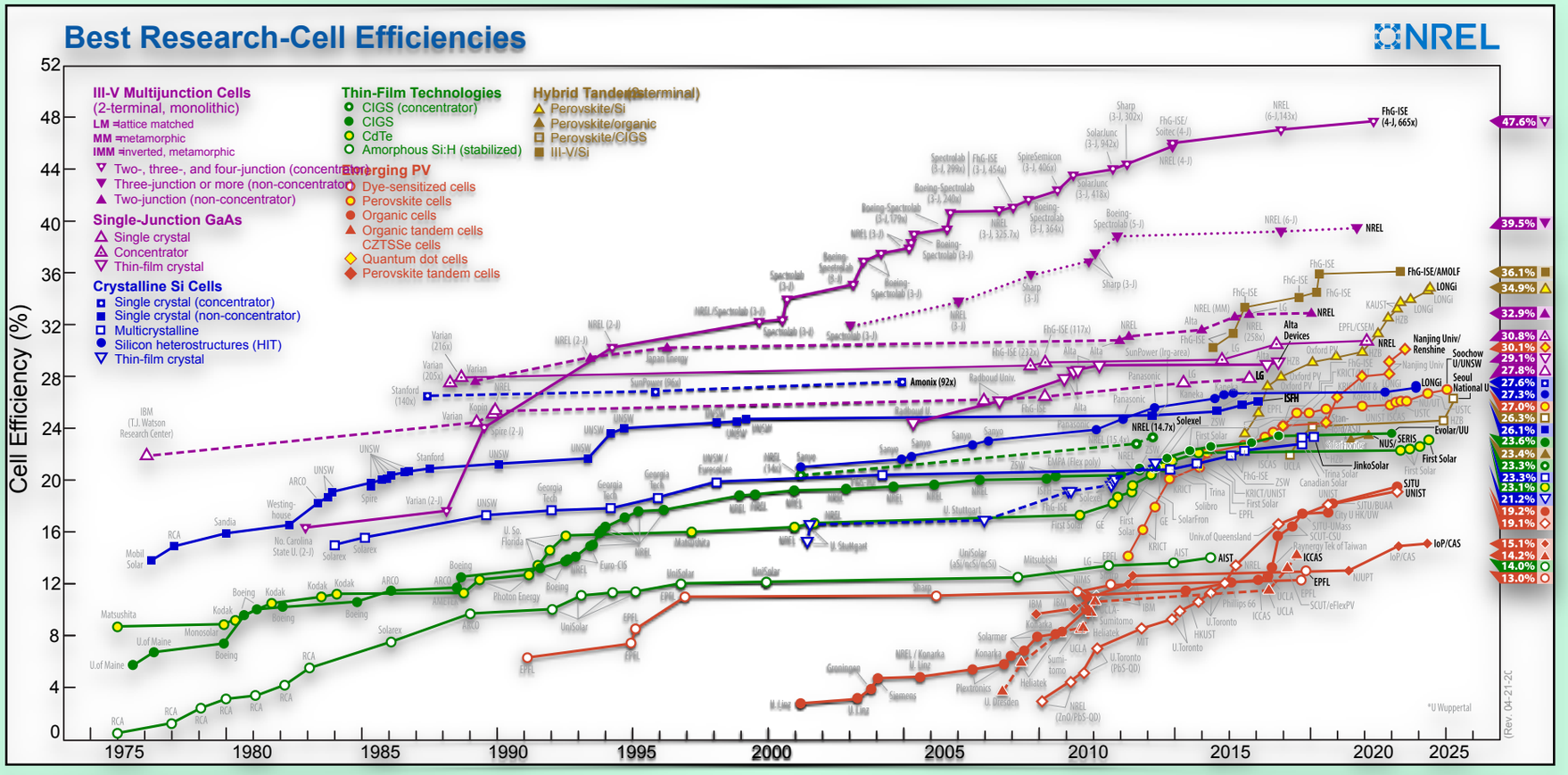
Li-ion battery tested in 1970s

Widely used only in 1990s

Still some way before widely used for mobility

Giant magnetoresistance materials (1988)

Data-storage storage (1997)



Materials space

- Space of all materials structure and composition
  - Inorganic Crystal Structure Database (ICSD): 307,301 crystal structures as on 01/10/2024
  - Crystallography Open Database (COD): 526,936 entries as on 06/08/2025
  - At least 108,423 experimentally verified, unique 2D materials with up to 6 different elements as in 2018 (Mount et al. Nat. Nanotech. 2018)
  - Number of possible inorganic materials $> 10^{10}$.
- **A vast materials space to be explored**

How to do this efficiently?

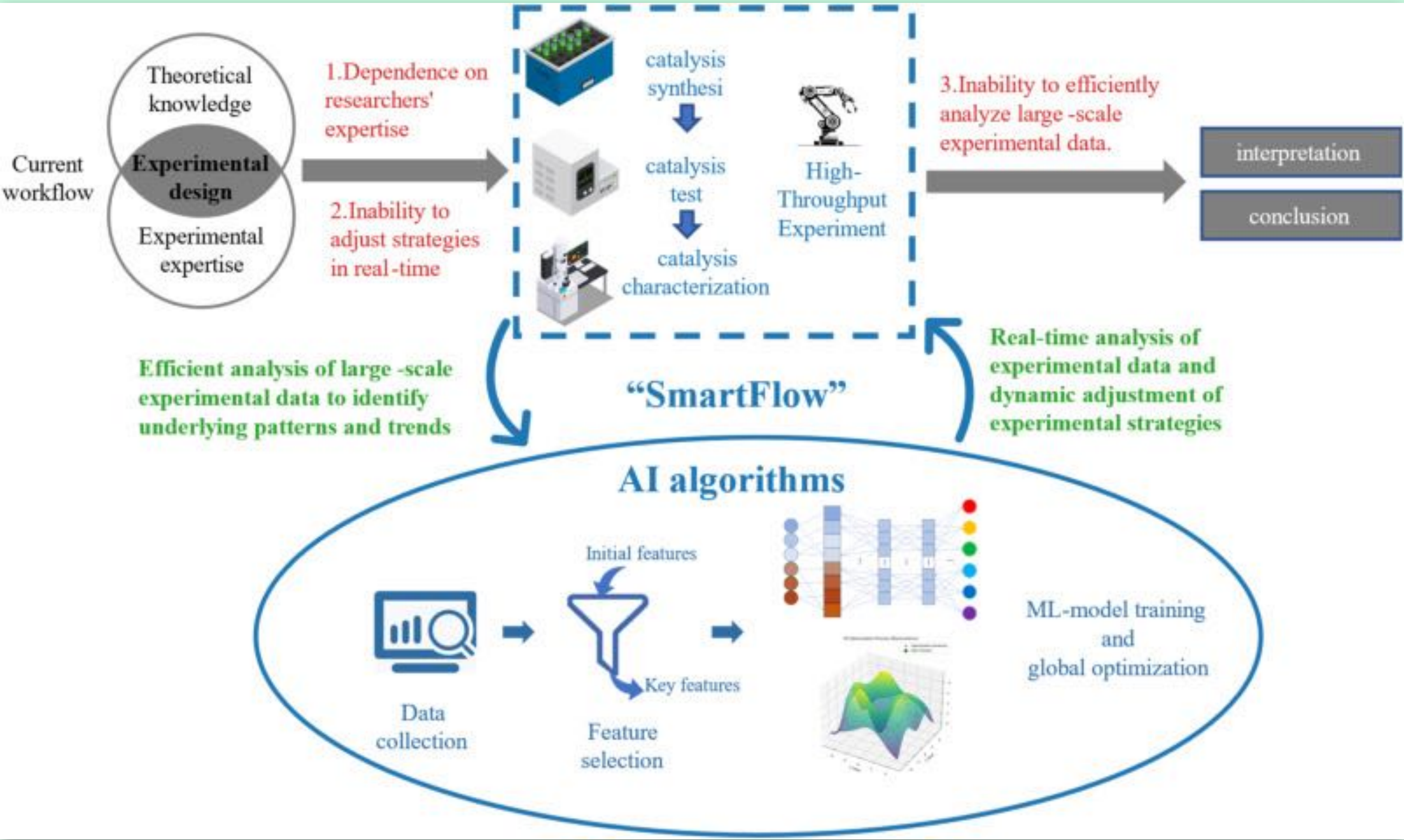
- **High throughput experiments**
  - Faster, but still expensive
  - Can explore only a limited number of materials
  - Could **autonomous experiments** be the answer?

- Quantum mechanical calculations (**density functional theory**) replacing experiments
  - Still (computationally) expensive, slow & labor intensive
  - High throughput computations, faster but computationally equally expensive

- This is where data-driven approach becomes crucial

High-throughput & autonomous experiments

Ma et a., Chinese J Chem. Engg. (2025)

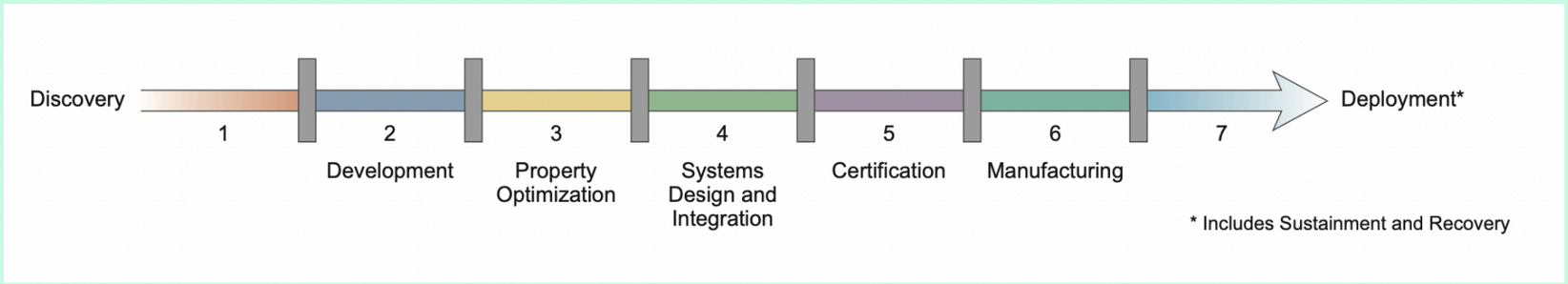


High throughput experimentation

Materials Genome Initiative

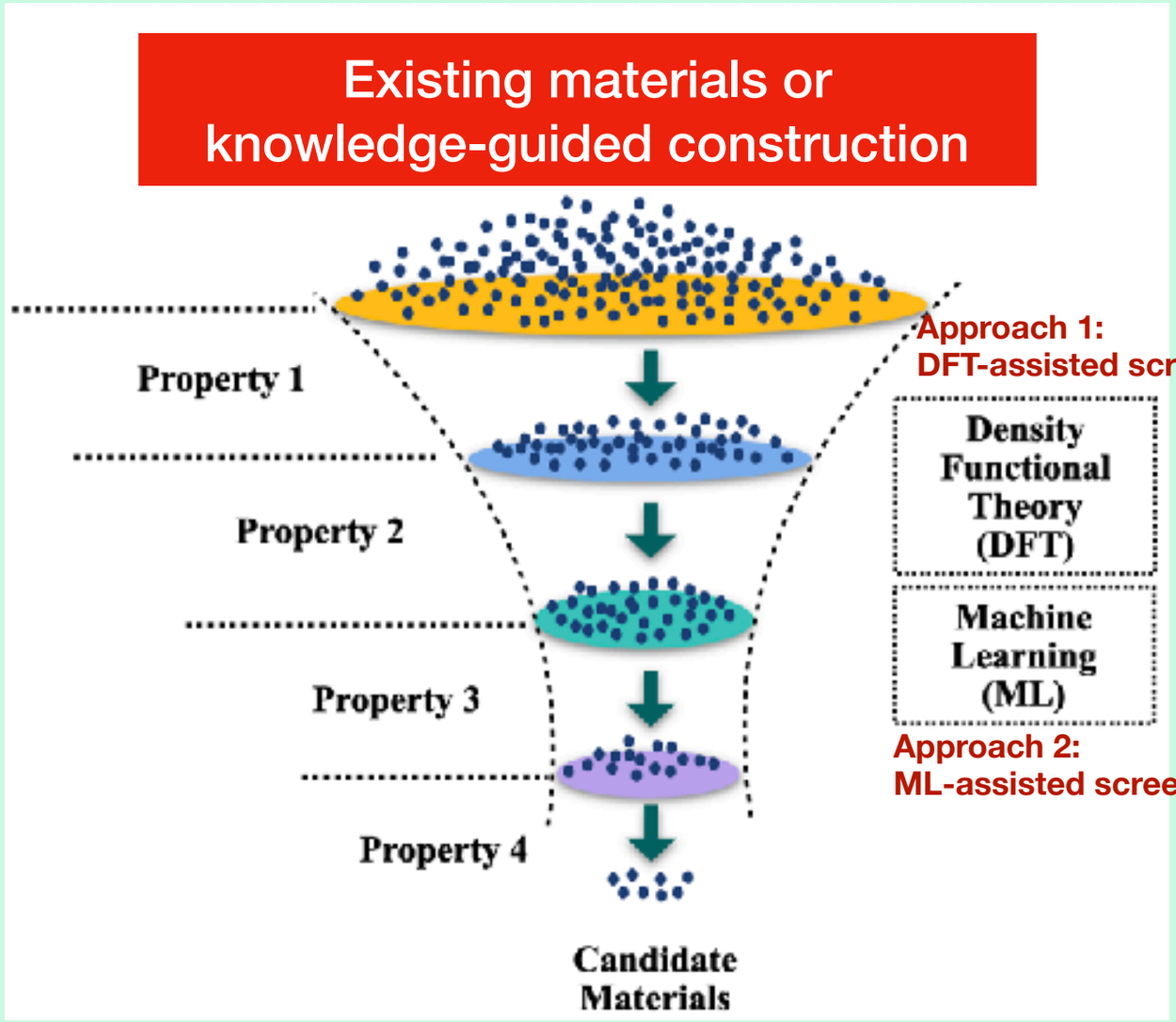
www.mgi.gov

- Multi-agency US government initiative
  - To reduce the time to develop and deploy advanced materials
  - Materials development & deployment typically one to two decades
  - MGI aimed to reduce it to half, at a much reduced cost
  - Bottleneck: Seven-stage development continuum (figure), little feedback between stages
- Need for better integration of and feedback between stages
- Encouraged open innovation ecosystem for accelerated materials discovery through
  - Large accessible databases
  - Advancing computational tools, including AI-driven methods



Source: MGI

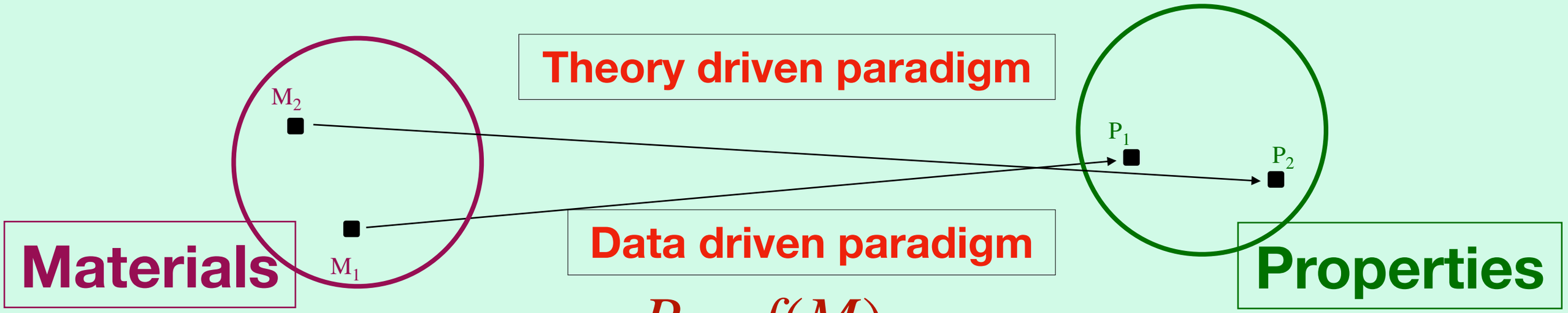
More rational approaches



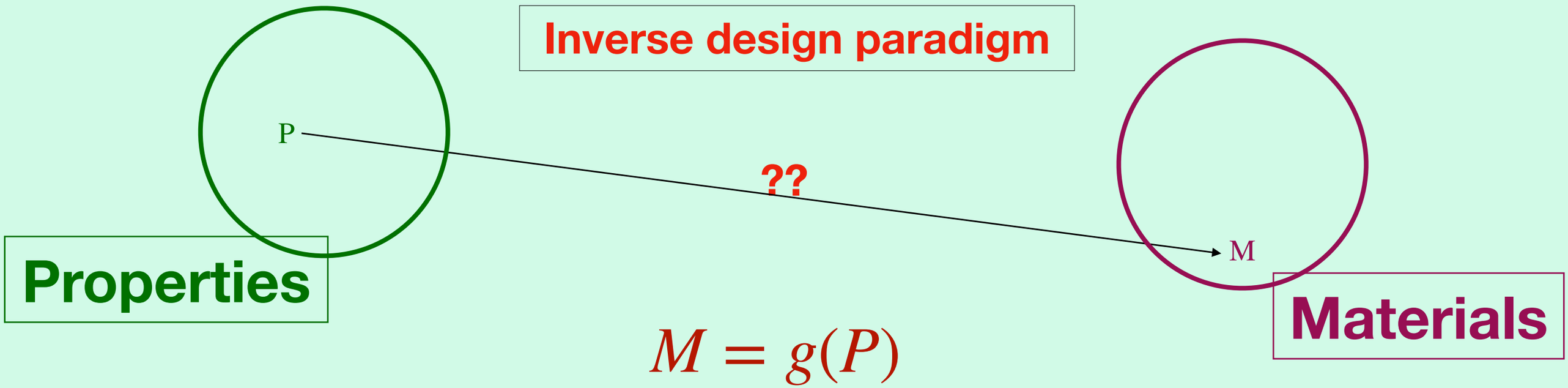
Materials of interest
Experiments

Approach 2:
Inverse design of materials

Inverting the question



$$P = f(M)$$



$$M = g(P)$$

Materials property prediction (& screening)

Examples

- Is a material stable? — thermodynamic, dynamical and mechanical stability.
- Calculate formation energy, distance from hull, phonon spectrum, usually DFT.

- Band gap of a semiconductor — DFT underestimates band gaps, more advanced calculations more expensive.

- Ferromagnet or anti-ferromagnet? What is the saturation magnetization? Coercivity?

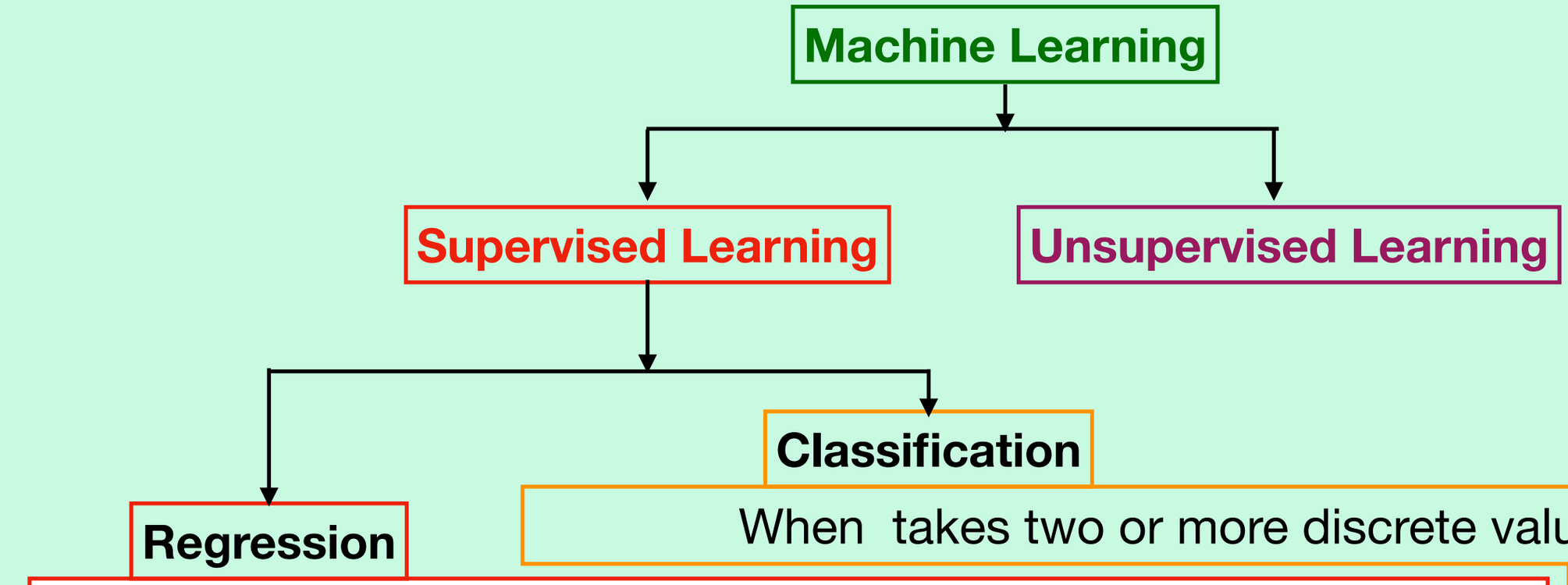
Training models to predict properties

Using existing data

- **Materials genome initiative**
  - https://www.mgi.gov
- **Materials Project**
  - https://materialsproject.org
- NOMAD
  - https://nomad-coe.eu
- **Open quantum materials database**
  - http://oqmd.org
- **AFLOW**
  - http://www.aflow.org
- **Computational materials repository**
  - https://cmr.fysik.dtu.dk
- **novomag**
  - https://www.novomag.physics.iastate.edu/structure-database
- **Novamag**
  - https://zenodo.org/records/3241267
- **Alexandria**
  - https://figshare.com/articles/dataset/Alexandria_DB/27174897?file=49622718
- JARVIS-DFT
  - https://jarvis.nist.gov/

Machine Learning

In one slide



Classification

- Stable or unstable?
- Magnetic or NM?
- FM or AFM?

- **Regression**
  - $Y = f(X) + \epsilon$; $X = (X_1, X_2, \ldots, X_p)$ called the predictors, descriptors, features
  - ϵ is error independent of X, $\bar{\epsilon} = 0$
  - Estimate f (say $\hat{f}$) from observed points, predict $\hat{Y} = \hat{f}(X)$ for a new X
  - Linear regression: Predicted $\hat{Y}^i = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$.
  - β_j's obtained minimizing $RSS = \sum_{i=1}^{N} \left(Y^i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j^i\right)^2$
  - Ridge regression: Minimize $\sum_{i=1}^{N} \left(Y^i - \beta_0 - \sum_{i=1}^{p} \beta_j X_j^i\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

parameters

$\lambda > 0$

hyper-parameter

Details we will learn

- How to access/download from (free) databases
- Materials representation for machine learning
  - Properties that materials features should ideally satisfy
  - What features to use, and how to create them
- Training models, measuring their performance
- Few examples where ML is used to accelerate materials characterization

Topics we won't learn

- **Generative models** (used for inverse design), most elegant way of exploring the vast materials space
  - Generative adversarial network (GAN)
  - Variational auto-encoder (VAE)
  - Diffusion-based models

- **Machine learning force fields (MLFF)**
  - Molecular dynamics is a powerful simulation technique in physics, chemistry (i.e., materials) and biology
  - Parametrized potentials
    - Example: Lennard-Jones $e_{ij} = -\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^6 - \left(\frac{\sigma}{r_{ij}}\right)^{12}\right]$. Such potentials for C, Si, metals, e.g. Fe, Co. Fitted to DFT or experiments. Inexpensive calculations, but constrained to represent specific environments.
  - Ab initio or DFT-generated potentials. Accurate but expensive.
  - ML models trained on DFT results to predict energies and forces for a collection of atoms; universal force fields. DFT-level accurate calculations at the cost of classical MD.

Possibilities for the future

- Future of materials science, particularly materials design—
  - Efficient exploration of materials space via
    - Generative models
    - Accurate property predictions via surrogate models, materials screening
    - Efficient materials simulations using MLFFs
    - High-throughput and autonomous materials synthesis, characterization and property measurements

- Benefits —
  - Accelerated materials development and deployment in technologies
  - High fidelity; precise, large-scale synthesis of materials, enhanced reproducibility, safety, cost-effectiveness
  - Tasks taking months can be completed in days (Burger et al. Nature (2020))

- Integrated materials design, synthesis and characterization platforms that are
  - Connected, Autonomous, Shared and High-throughput (CASH) (Shimizu et al. APL Mater. (2020))
  - A dream for the future: A single command, or even a voice command
    - Explores materials space & generates new materials
    - Checks for stability & and desired properties
    - If promising, hands over to the autonomous experimental agent who
    - Synthesizes, characterizes, and measures properties
    - Different components may be distributed geographically