Application of AI/ML in Materials Science

Prof N S **Harsha** Gunda

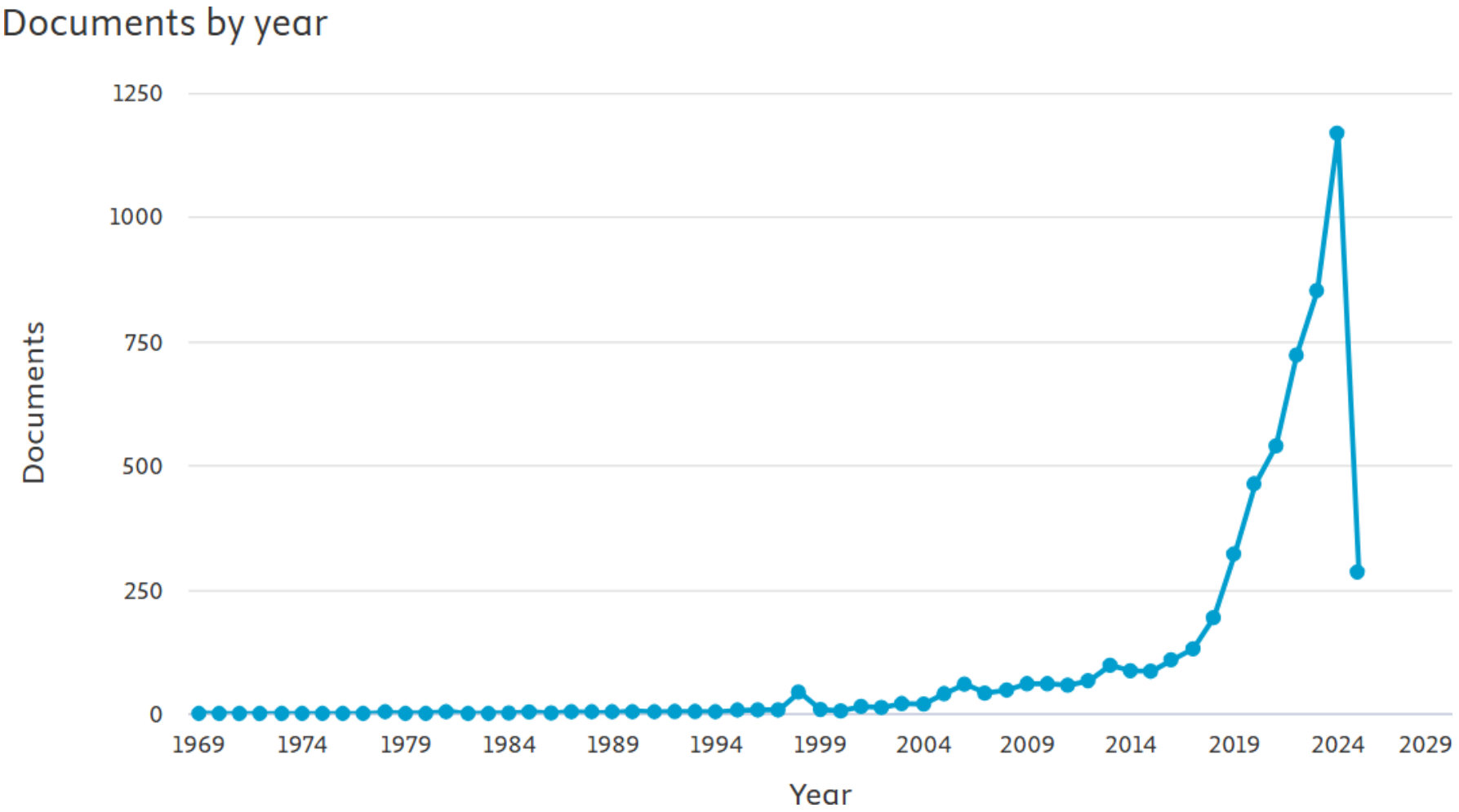
What is Machine Learning?

Using data-driven algorithms to discern patterns and make predictions on big, high-dimensional data

Clustering, classification, regression, generation

In materials science, we are most interested in processing-structure-property-performance relationships

Scopus Citation Report



Keywords:
Materials Informatics
Machine Learning & Materials Science

https://www.scopus.com/search/form.uri#basic

ARTICLE OPEN

Magnetic and superconducting phase
temperatures predicted using text min

Callum J. Court[1] and Jacqueline M. Cole[1,2,3,4]✉

ARTICLE

https://doi.org/10.1038/s42005-020-0317-3 OPEN

Artificial-intelligence-driven s
microscopy

A. Krull[1,2,7]✉, P. Hirsch[3,7], C. Rother[4], A. Schiffrin[5,

An Automated Machine Learning architectu
prediction of Metal-Organic Frameworks pe
environmental applications

Ioannis Tsamardinos[a,d], George S. Fanourgakis[b], Elissa
Konstantinos Gkagkas[e], George E. Froudakis[b,*]

IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, VOL. 30, NO. 4, JUNE 202

Critical Temperature Prediction fo
A Variational Bayesian Neural N

Thanh Dung Le, *Member, IEEE*, Rita Noumeir, *Member, IEEE*, H
Jung Ho Kim, and Ho Min Kim, *Mem*

The Materials Simulation Toolkit for Machine l
automated open source toolkit to accelerate dat

Ryan Jacobs[a,*], Tam Mayeshiba[a], Ben Afflerbach[a], Luke Miles
Raphael Finkel[b], Dane Morgan[a]

Extraction of mechanical properties of materials through deep learning from instrumented indentation

Lu Lu, Ming Dao, Punit Kumar, Upadrasta Ramamurty, George Em Karniadakis, and Subra Suresh

PNAS March 31, 2020 117 (13) 7052-7062; first published March 16, 2020 https://doi-org.proxy.lib.ohio-state.edu/10.1073/pnas.1922210117

Contributed by Subra Suresh, February 10, 2020 (sent for rev Ting Zhu)

**Stability and anion diffu
stabilized zirconia resol
global potential energy surface exploration**

J. Chem. Phys. 152, 094703 (2020); https://doi-org.proxy.lib.ohio-state.edu/10.1063/1.5142591

Shu-Hui Guan[1,2], Ke-Xiang Zhang[2], Cheng Shang[2,a)], *and* Zhi-Pan Liu[2,b)]

more...

Machine learning models for the prediction of energy, forces, and stresses for Platinum

J. Chapman, R. Batra, R. Ramprasad ✉

Published: 16 March 2020

Machine-learning-assisted prediction of the mechanical properties of Cu—Al alloy

Zheng-hua Deng, Hai-qing Yin ✉, Xue Jiang, Cong Zhang, Guo-fei Zhang, Bin Xu, Guo-qiang Yang, Tong Zhang, Mao Wu & Xuan-hui Qu

International Journal of Minerals, Metallurgy and Materials **27**, 362–373(2020) | Cite this article

15 Accesses | Metrics

ic algorithm-driven discovery of
ected thermal conductivity
cement by disorder

Hua Bao [a] ✉, Xiulin Ruan [b] ✉

International Journal of Quantum Chemistry
Volume 120, Issue 8, 15 April 2020, Article number e26151

Representations and descriptors unifying the study of molecular and bulk systems (Article)

Rossi, K.[a], Cumby, J.[b] ✉

anuary 2020

amework for brittle porous
e learning and geometric

ş ✉

ce **55**, 4734–4747(2020) | Cite this article

r 645

rmulations with carbon nanotubes and intrinsic
tation analysis with machine learning on

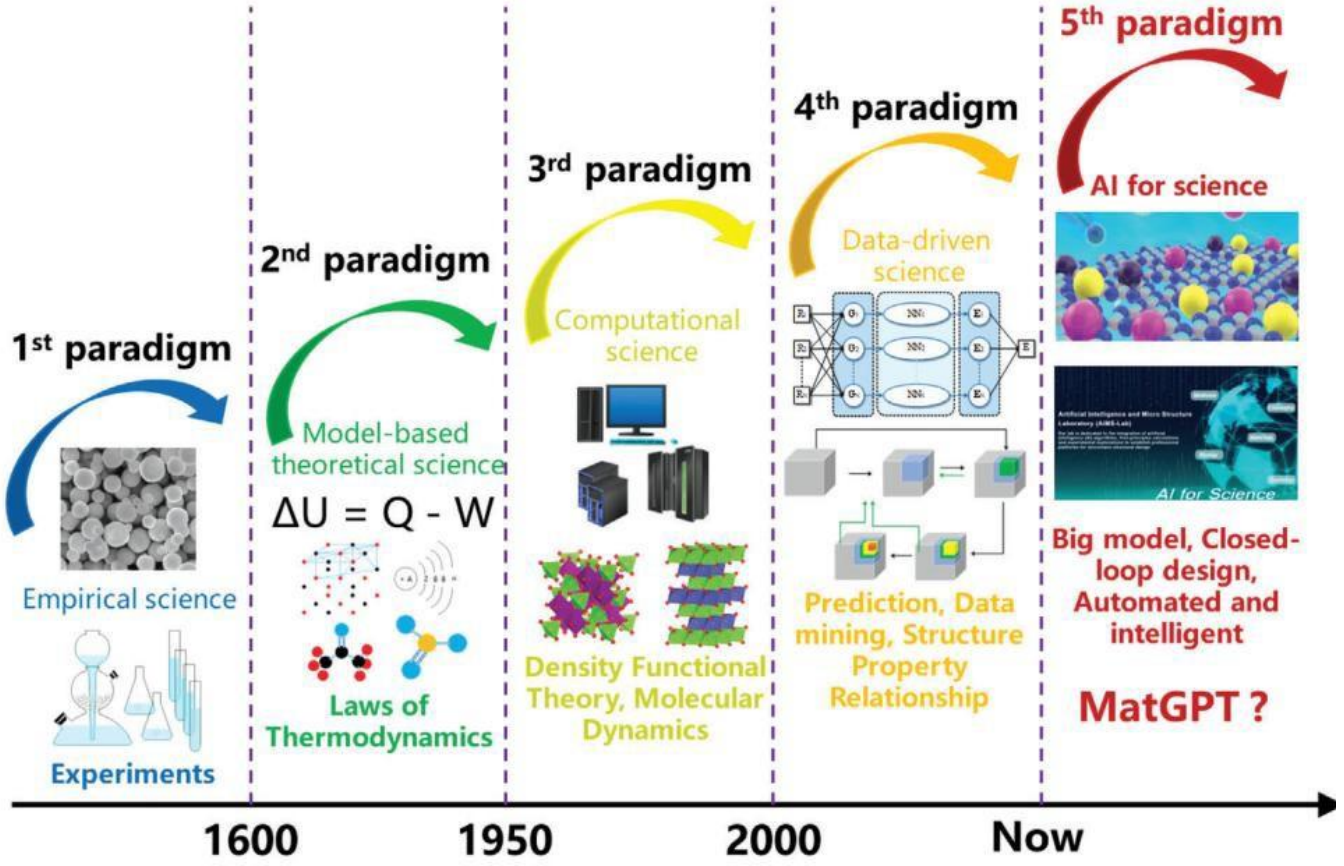
icle) (Open Access)

, Charitidis, C.A.[a] ✉

posite, Nano Materials & Nanotechnology, School of Chemical
ens, Zographos Athens, GR-15773, Greece
(IRES), Boulevard Edmond Machtens 79/22, Brussels, 1080, Belgium

Design paradigm in Computational Materials Research

The ultimate goal is **Accelerated Materials Discovery and Design**!



REVIEW

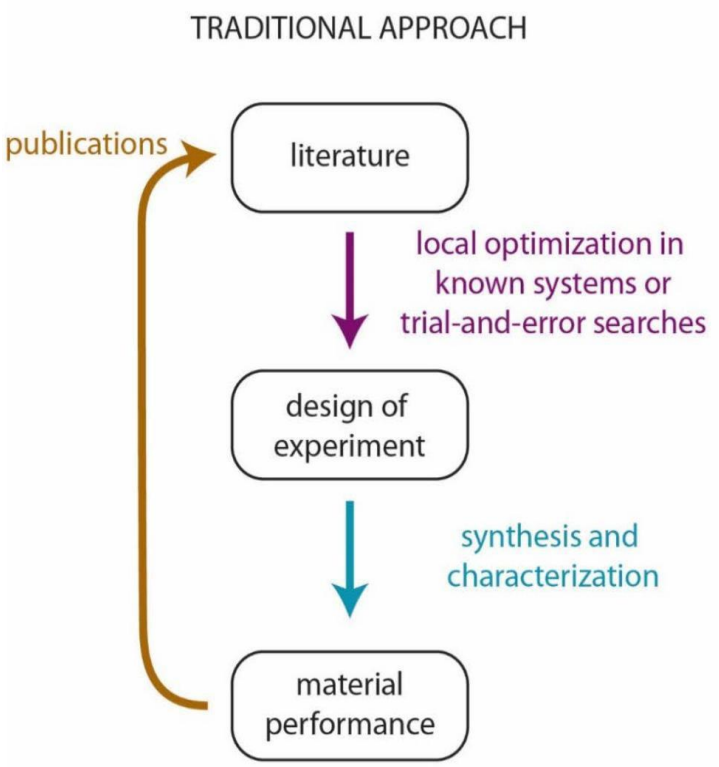
ADVANCED MATERIALS
www.advmat.de

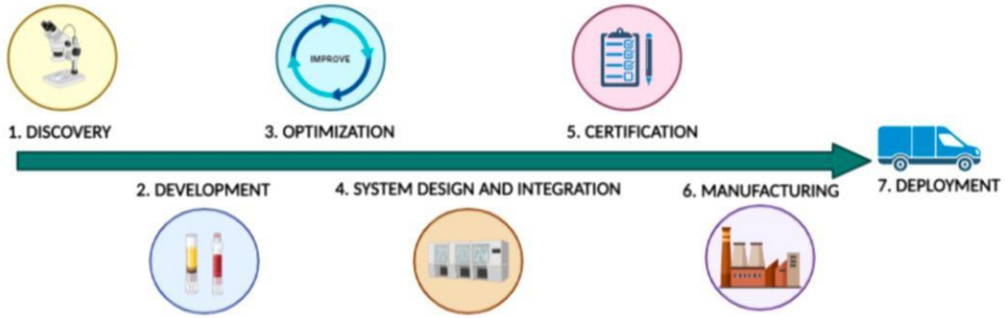
MatGPT: A Vane of Materials Informatics from Past, Present, to Future

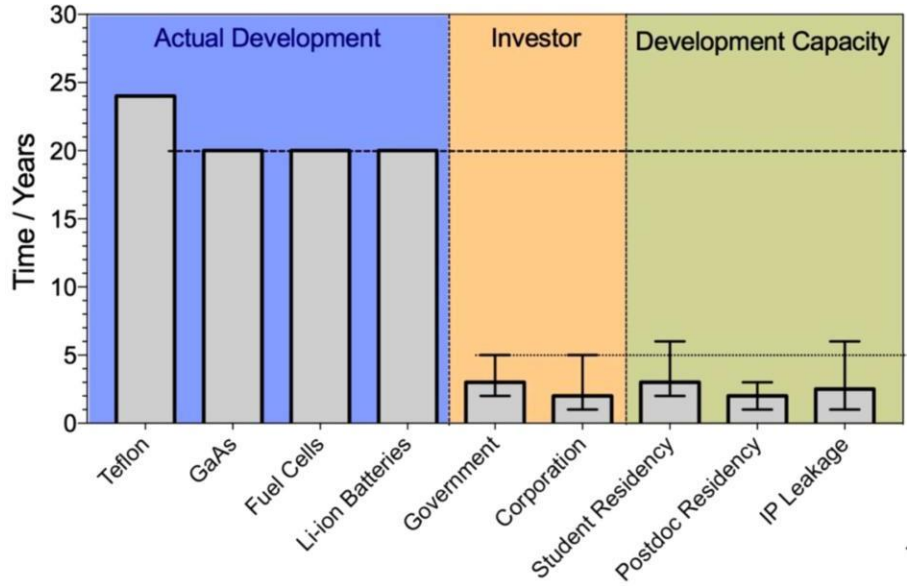
*Zhilong Wang, An Chen, Kehao Tao, Yanqiang Han, and Jinjin Li**

Materials Discovery

"Experience"
"Intuition"
"Common Sense"

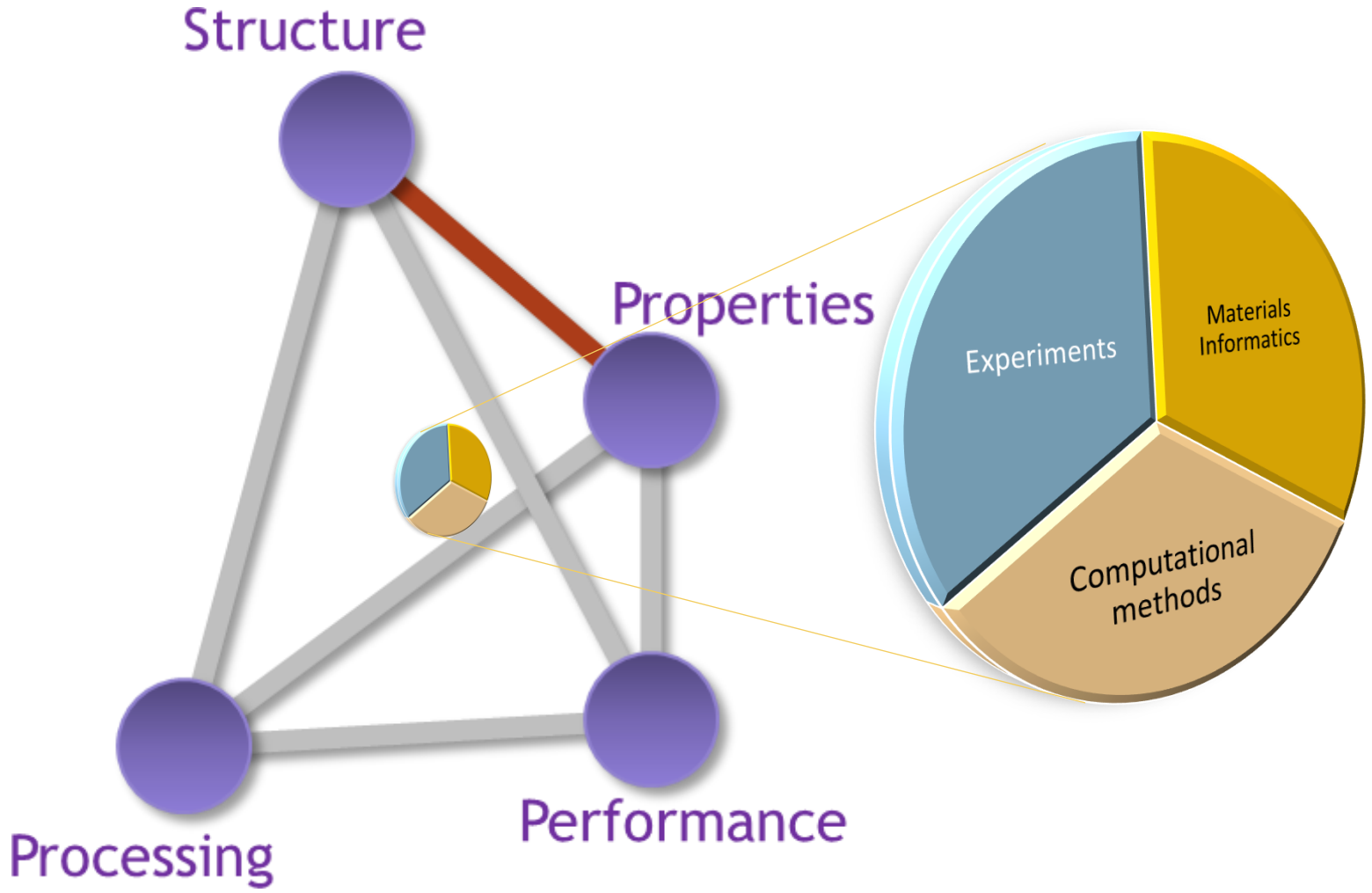




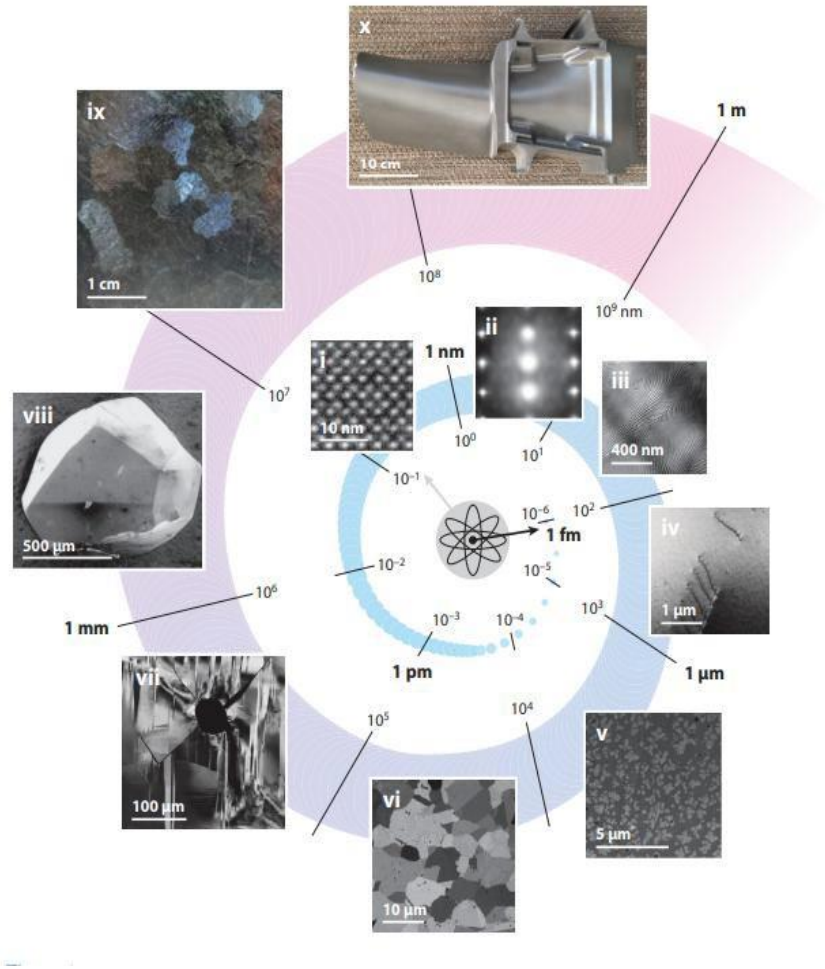


Dhruv Menon and Raghavan Ranganathan, ACS Omega, (2022)

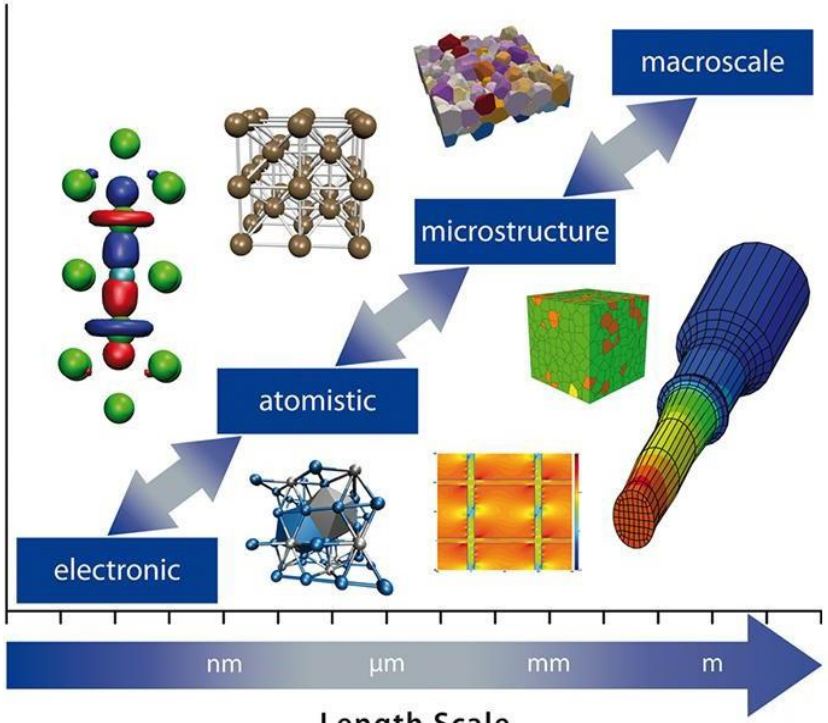
Materials Discovery



Materials Discovery



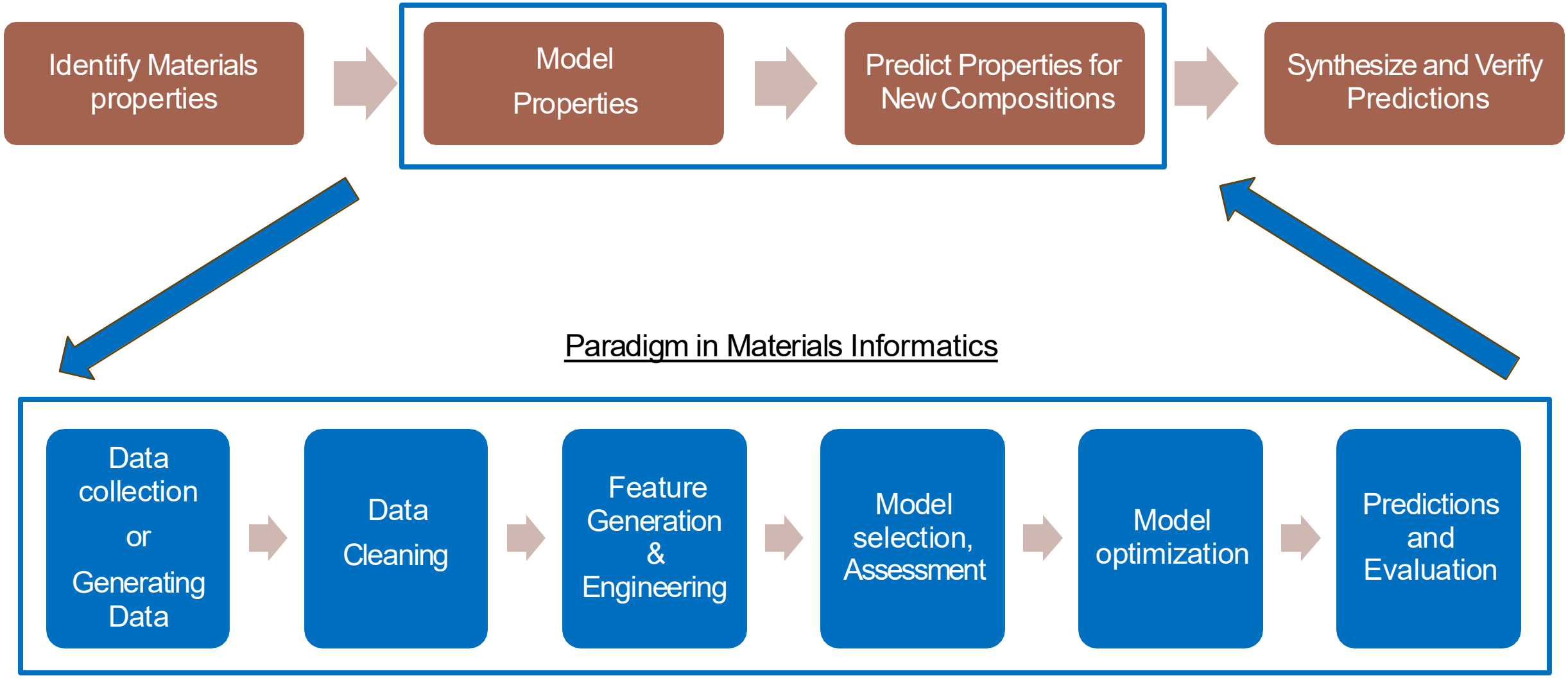
Heterogenous data



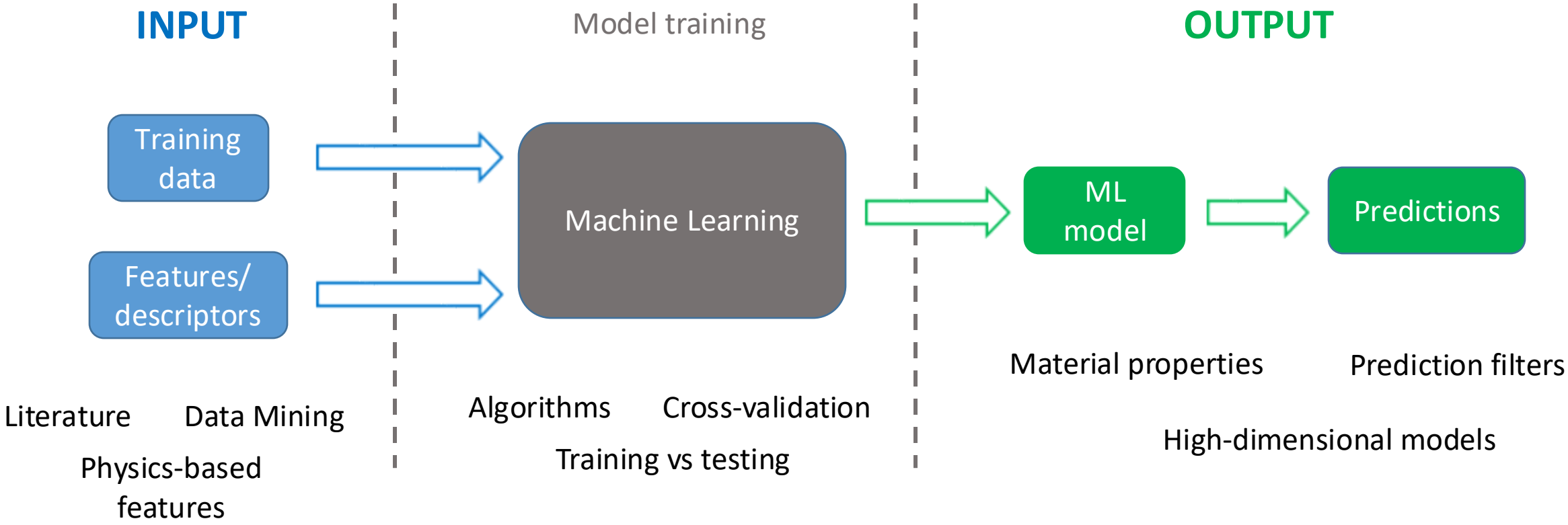
Materials Data Science: Current Status and Future Outlook

Annual Review of Materials Research
Vol. 45:171-193 (Volume publication date July 2015)
https://doi.org/10.1146/annurev-matsci-070214-020844

Typical workflow in Materials Discovery



Identifying features and robust training data are necessary for ML materials



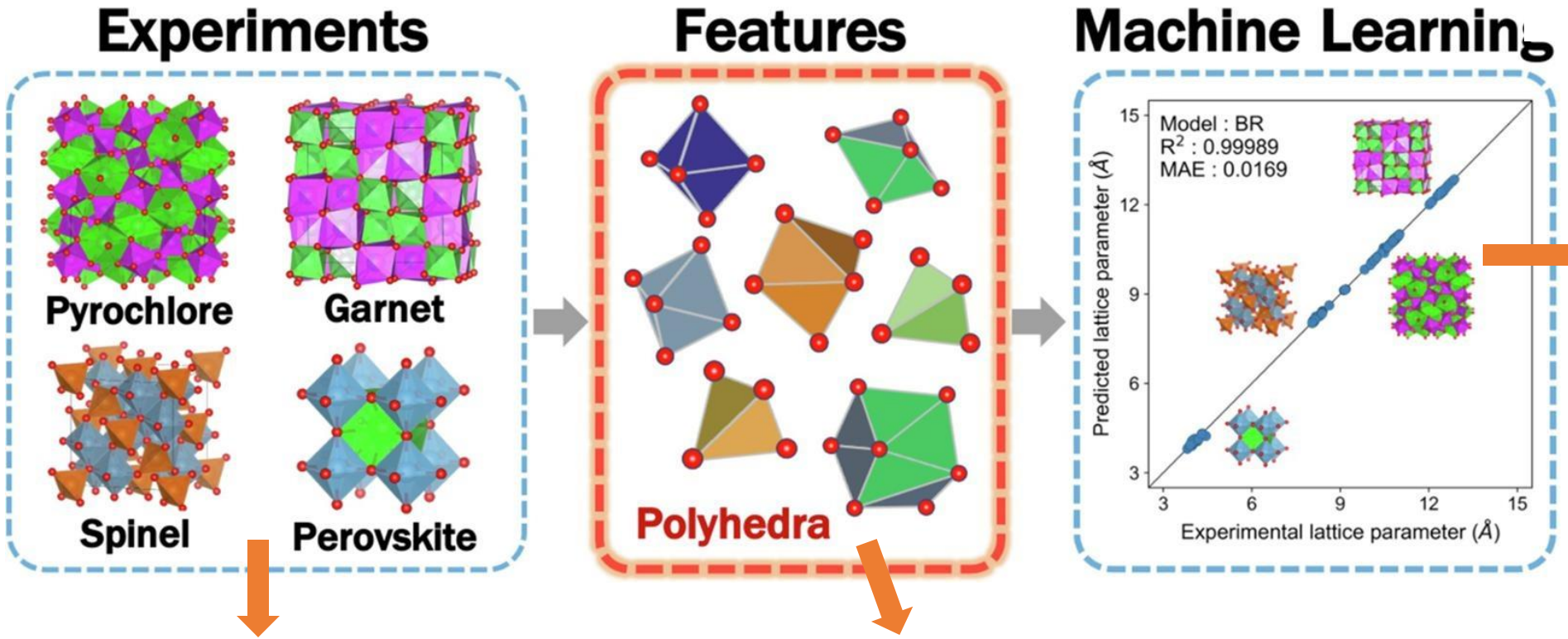
Lets look at two case studies from my research to demonstrate the power for ML fitting

ML to predict high temperature properties of oxides

- Training accurate ML models using simple but effective polyhedral features of oxides
- Target to predict accurate thermal expansion in oxides

Thermal Expansion Coefficient

$$TEC(K^{-1}) = \frac{L_1 - L_0}{T_1 - T_0} * \frac{1}{L_0}$$



Models trained on high-temperature lattice parameters of oxides

Database of experimental observation in binary and ternary oxides

Number of polyhedra (bonds) as ML features

Jian Peng & N. S. Harsha Gunda et al.*, Computational Materials Science, Volume 210, July 2022, 111034. (*Equally contributing first-author)

Setting up the data for training a ML model

77 oxides

Polyhedral count

72 descriptors

700 data points

Name	Al^{3+} (oct)	Al^{3+} (tet)	Cu^{2+}(tet)	…	…	Temperature, K	Lattice constants (Å)
$CuAl_2O_4$	16	0	8	…	..	293	8.081
$CuAl_2O_4$	16	0	8	..	..	473	8.092
$Y_3Al_5O_{12}$	0	16	0	..	..	296	12.008
$Y_3Al_5O_{12}$	0	16	0	..	..	573	12.034
…							
…							

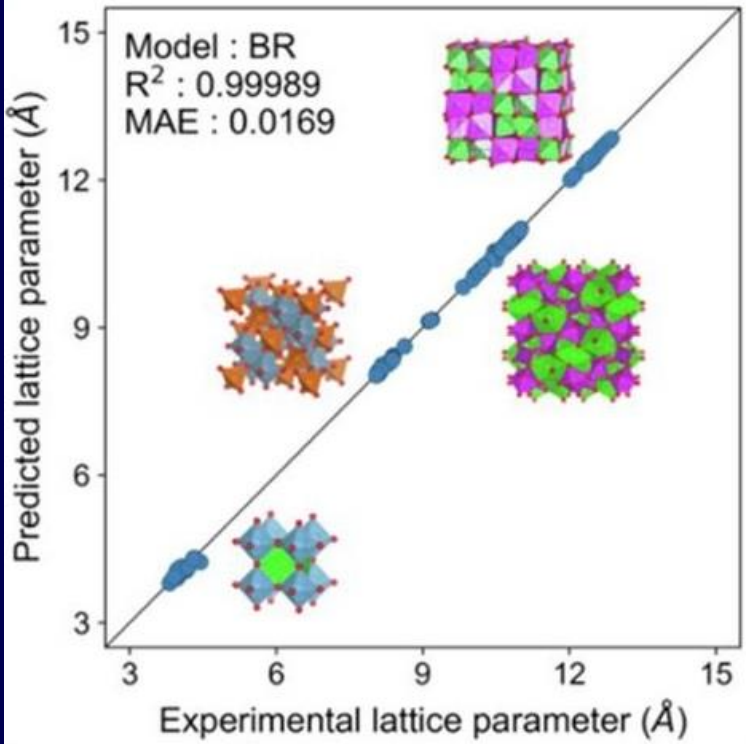
Input

Features

Training data

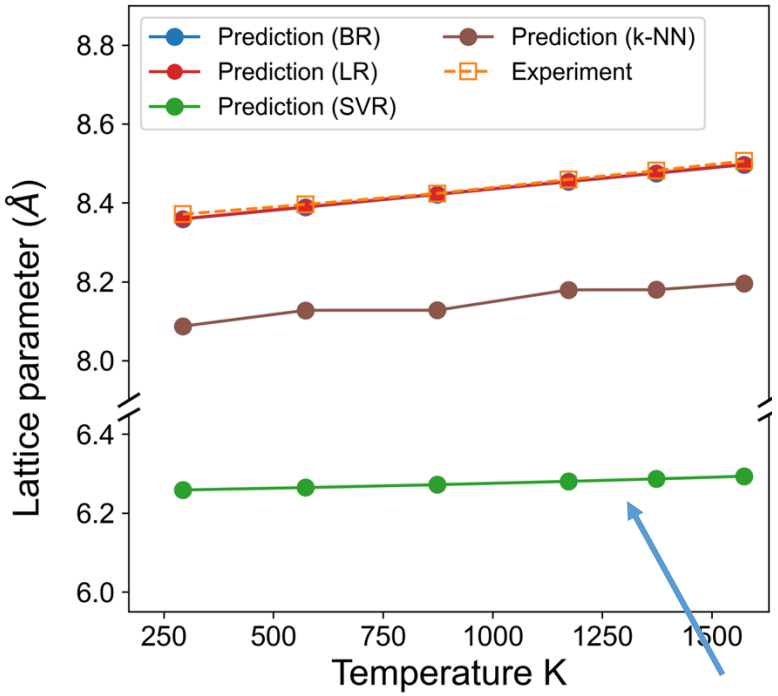
ML with adequate data and physical descriptors can predict materials properties

Training data

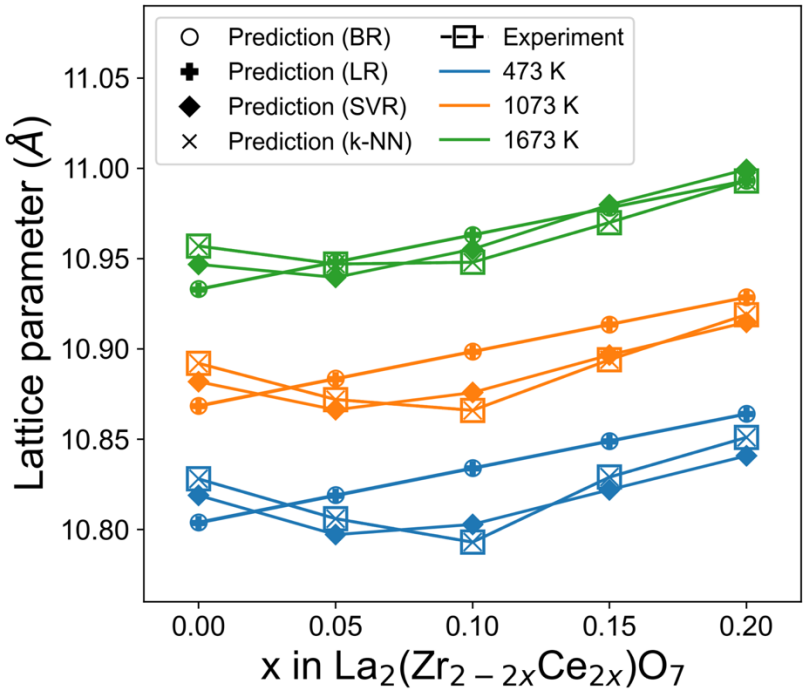


Predictions (not included in training)

$MgFe_2O_4$



$La_2(Zr_{2-2x}Ce_{2x})O_7$



Not every ML algorithm works for modeling

It is crucial to build a "good" dataset and find meaningful physical descriptors

Jian Peng & N. S. Harsha Gunda et al.*, Computational Materials Science, Volume 210, July 2022, 111034. (*Equally contributing first-author)

Accelerated development of ReaxFF-MD potentials

$$E_{system} = E_{bond} + E_{over} + E_{angle} + E_{tors} + E_{vdWaals} + E_{Coulomb} + E_{Specific}$$

Test system

bcc-Cr

14 ReaxFF parameters

P_1	P_2	P_3	..	..	P_{13}	P_{14}

Material property

High dimensional models

Modelling

MD Model

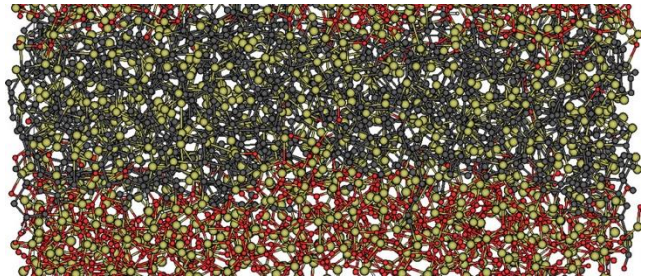
Manual

Automated

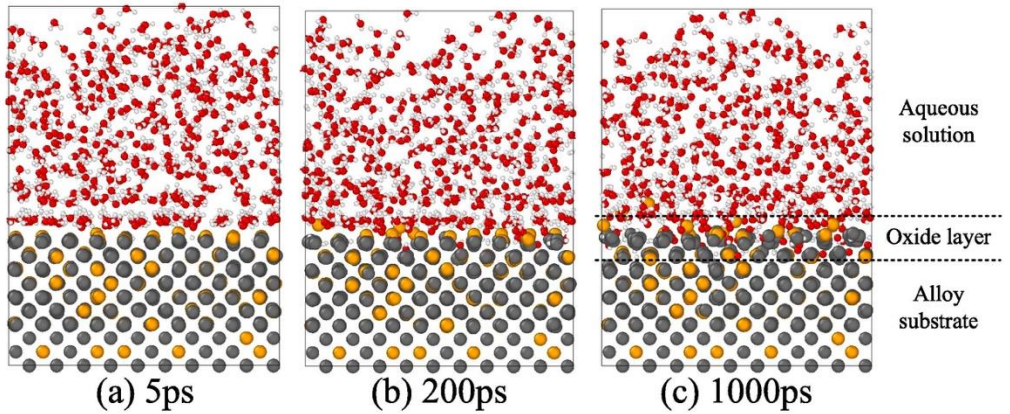
Months

Weeks

Reactive force field (ReaxFF) potentials uses bond order



Example: Simulation of oxidation in Fe-Cr alloys

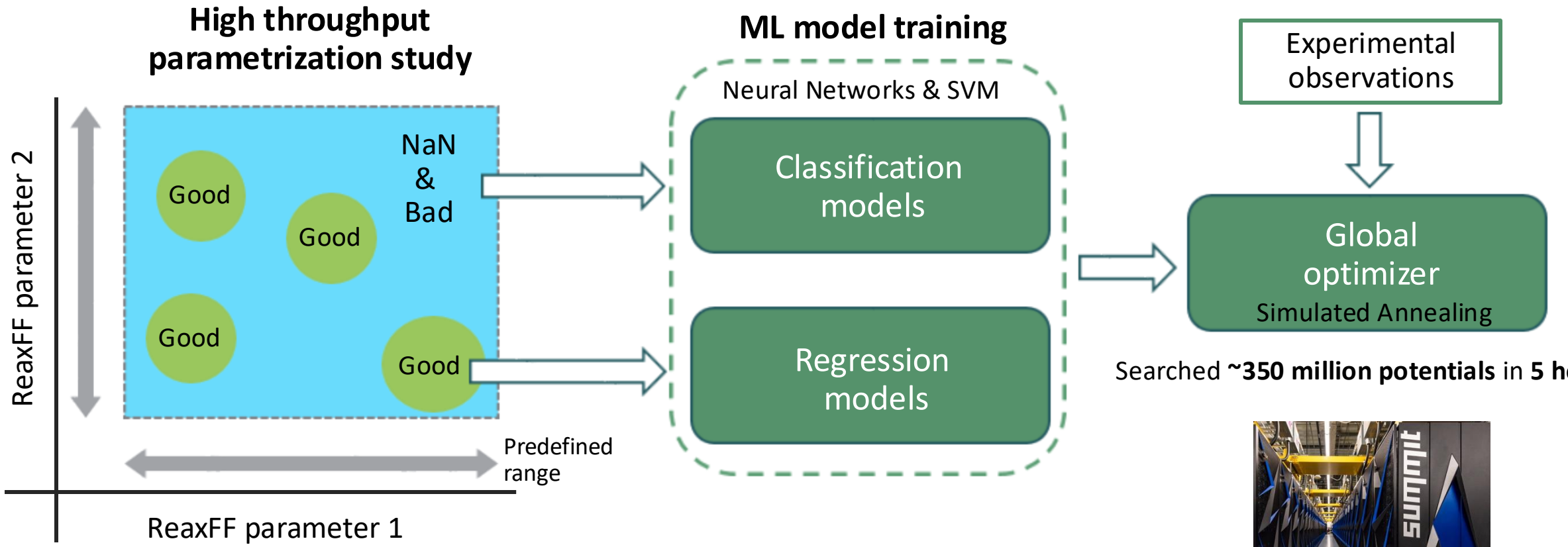


Jiang et al. Applied Surface Science 548 (2021) 149159

Ideal development scheme

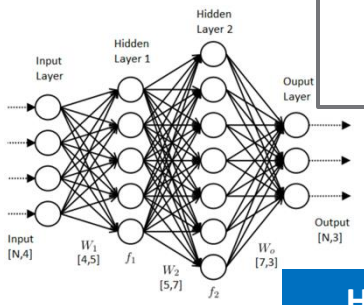
ReaxFF-MD calculations are extensively used to simulate bulk phenomenon in metals alloys

ML models can identify “good” ReaxFF parameters and predict MD response accurately



How good are the optimized potentials? Encouraging

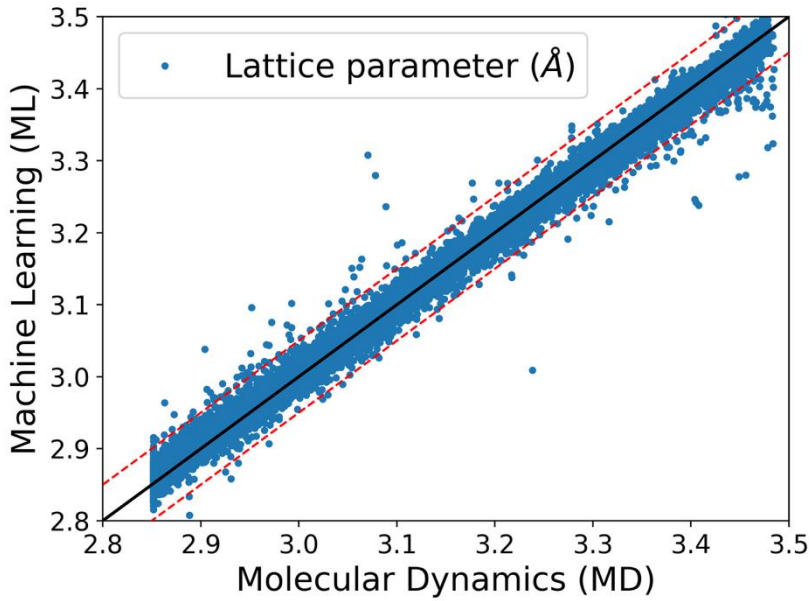
Neural networks



ML Models → Simulated Annealing → Optimized Potential → ReaxFF-MD Validation

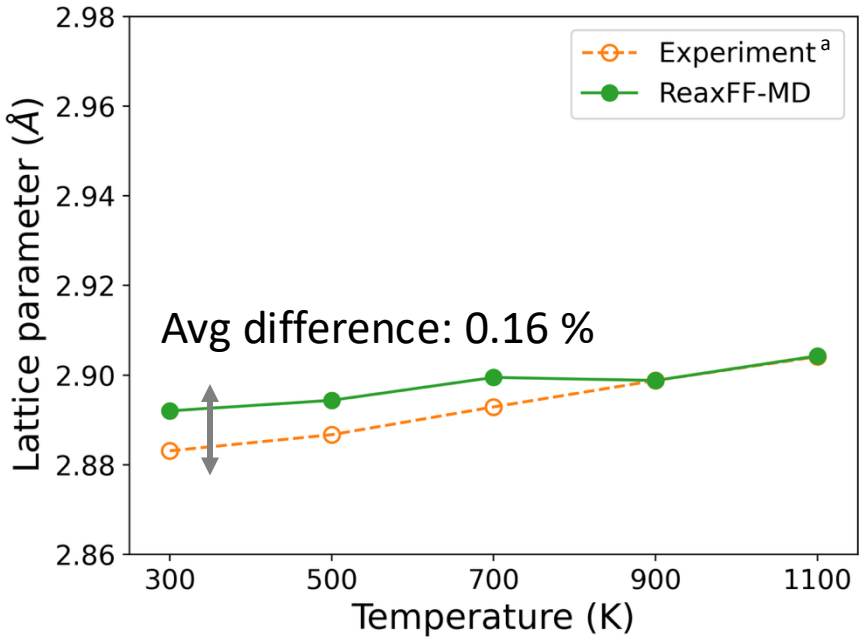
Searched **~350 million potentials** in **5 hours**

High-fidelity Machine Learning Models



*Predicting MD data not included in ML training

High-temperature lattice parameters

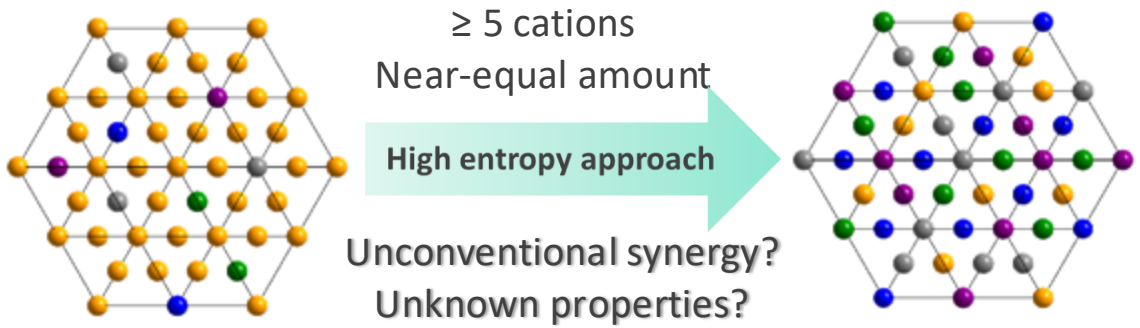


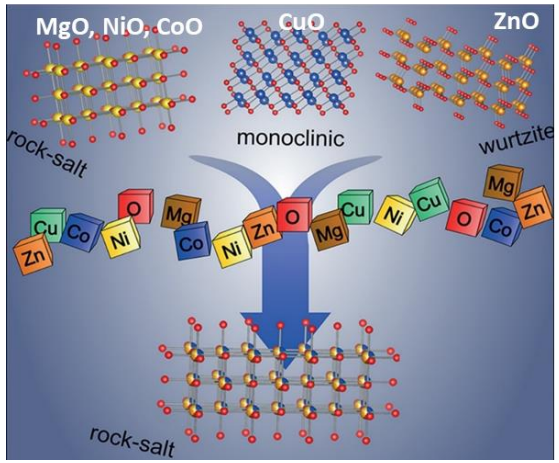
[a] Dubrovinskaia N. et al. CALPHAD **21**, 497-508 (1997)

*Manuscript under preparation

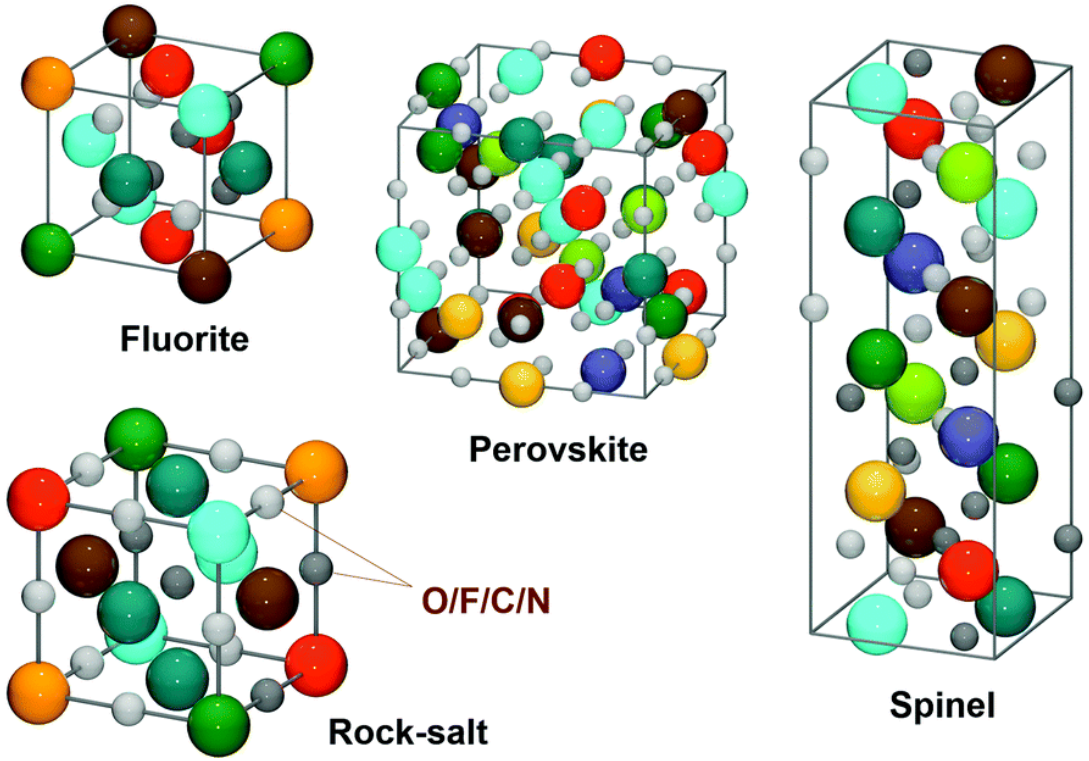
Design and discovery of new high-entropy ceramics (HEC)

- High-entropy ceramics have disordered cation lattices with ordered anions such as O, C, N, B etc.
- Discovered in 2015 (in bulk phases) and are stable in extreme conditions with good mechanical properties.
- Favorable properties such as oxidation resistance, wear resistance, and high melting point.





Ma et al. Energy Environ. Sci., 2021, 14, 2883-2905

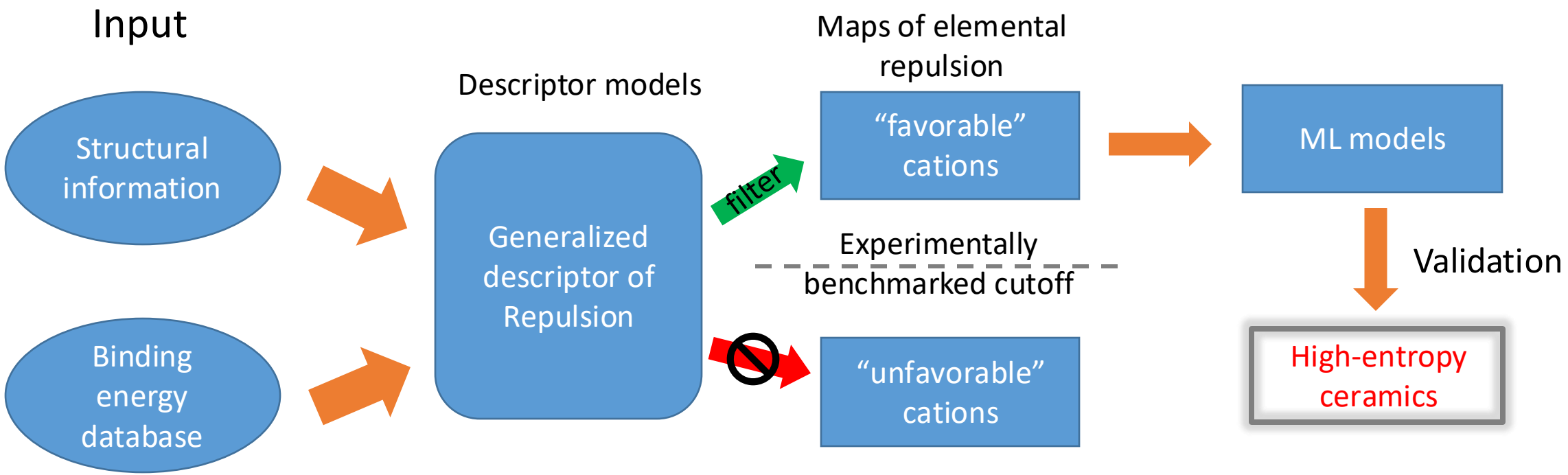


Models to predict the single-phase stability in HEC's are not well established

Sarkar et al., *Nature Communications* (2018)
Sarkar et. al., Advanced Materials (2019)

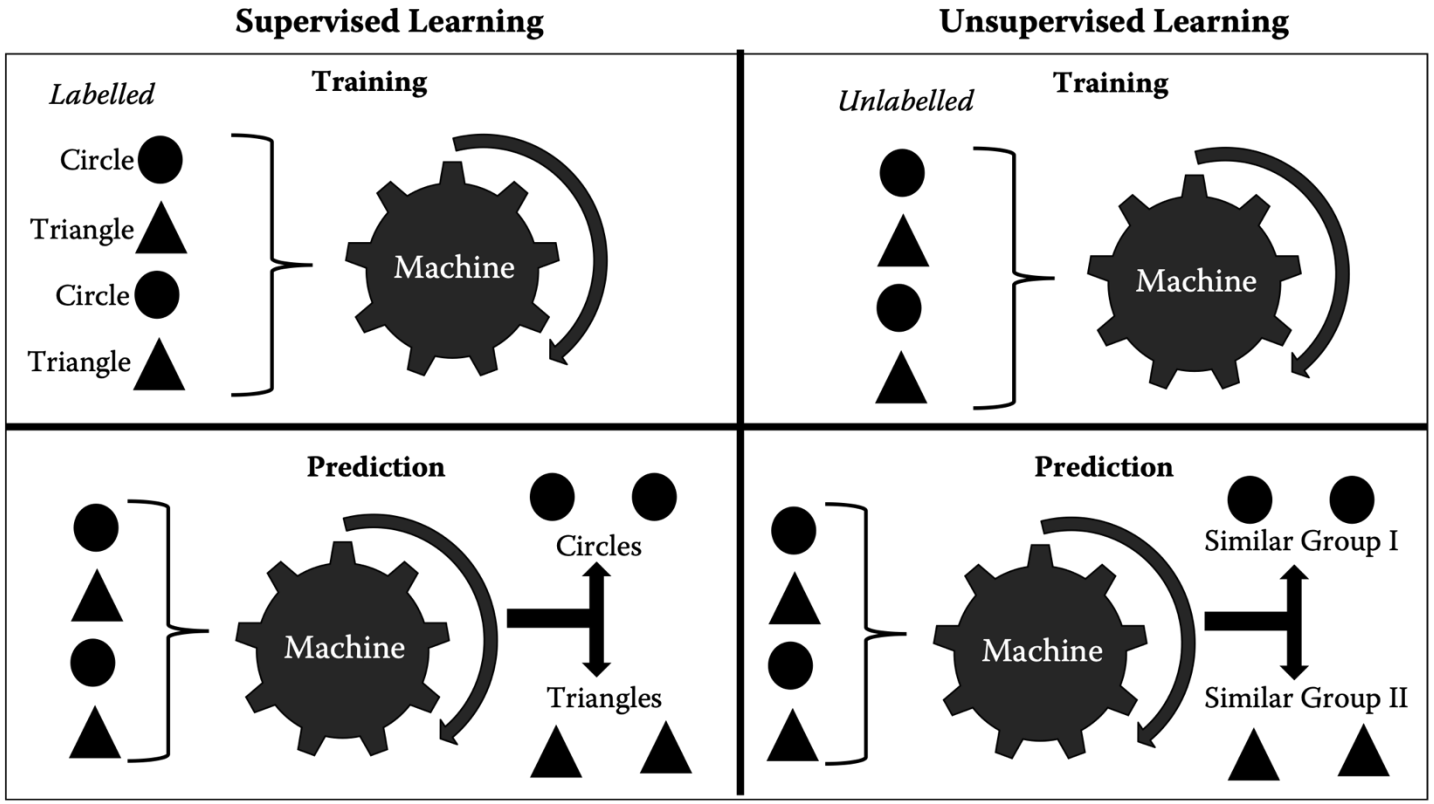
Generating new indicators to predict phase stability in high entropy ceramics

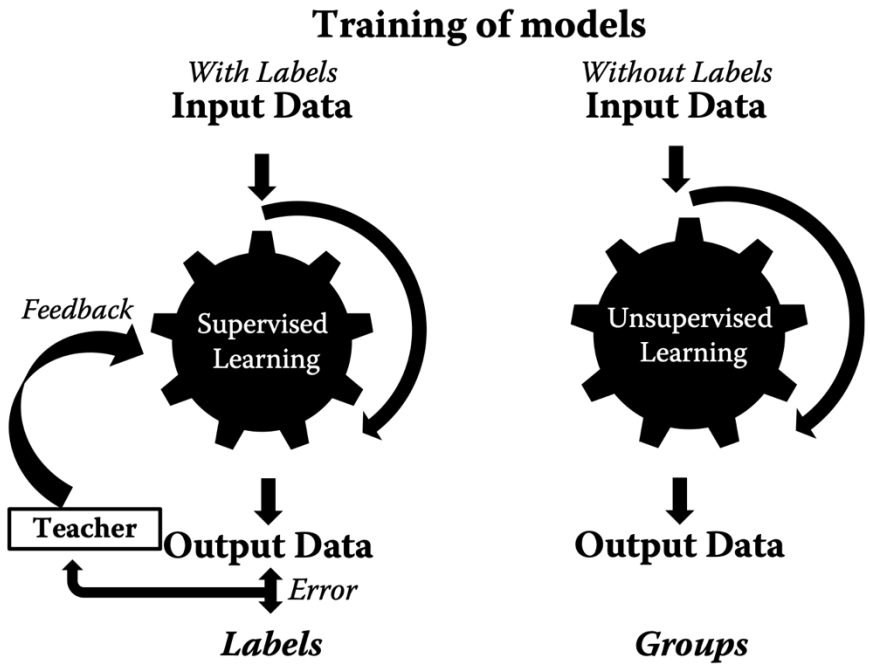
- Developing a generalized robust descriptor of repulsion across the elements
- Benchmarking with experiments to identify cutoffs in repulsion to form single phases



Let’s get to know some basics of ML

Two types of ML methods: Supervised and Unsupervised





There is also Semi-supervised Learning and Reinforcement Learning

Unsupervised Machine Learning

Training

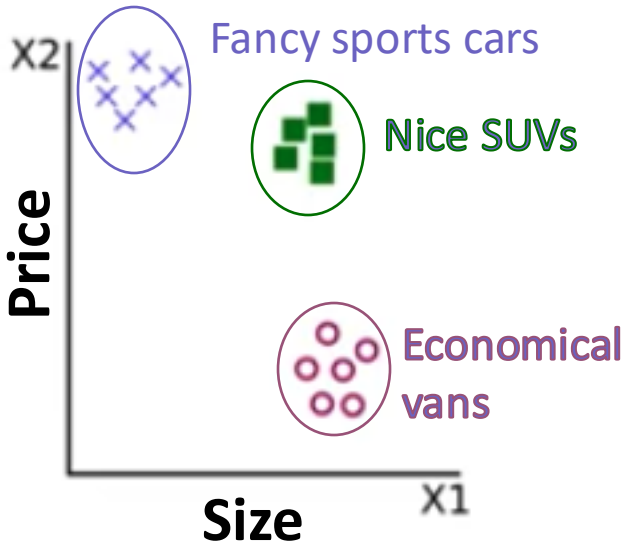
Inputs → **ML Model**

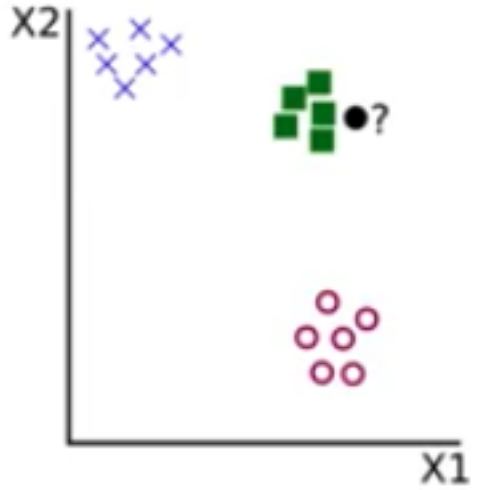
Can tell you that you have three clusters in your data (but not necessarily what those clusters correspond to)

Prediction

Inputs → **ML Model** → Outputs

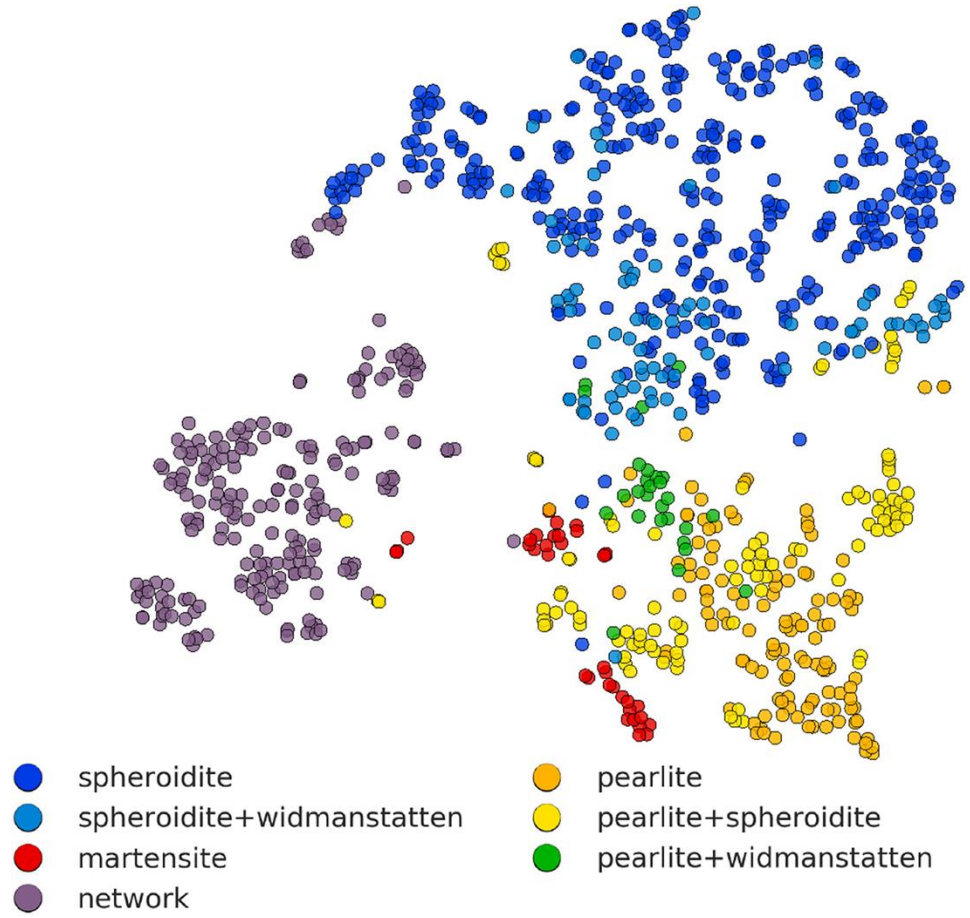
Ex.: Cars





Unsupervised Machine Learning: Examples

- How do you visualize high-dimensional data sets?
- Dimension reduction techniques
  - Principal Component Analysis (PCA)
    - Project data onto principle eigenvectors
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
    - Optimization to preserve local distances
    - Keep nearby points near each other



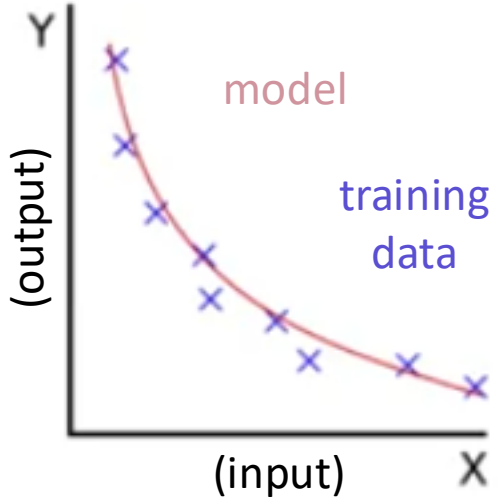
BL DeCost, T Francis, EA Holm, Acta Mater. 133 (2017) 30-40.

Supervised Machine Learning

Training

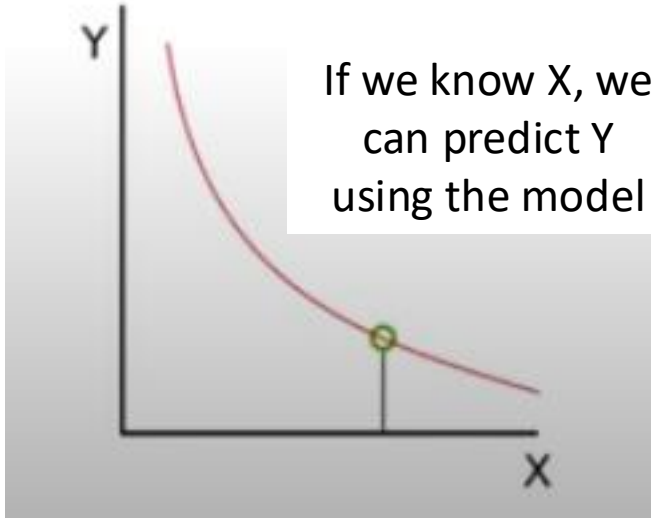
(Inputs, Outputs) → **ML Model**

Can fit model using inputs and outputs



Prediction

Inputs → **ML Model** → Outputs

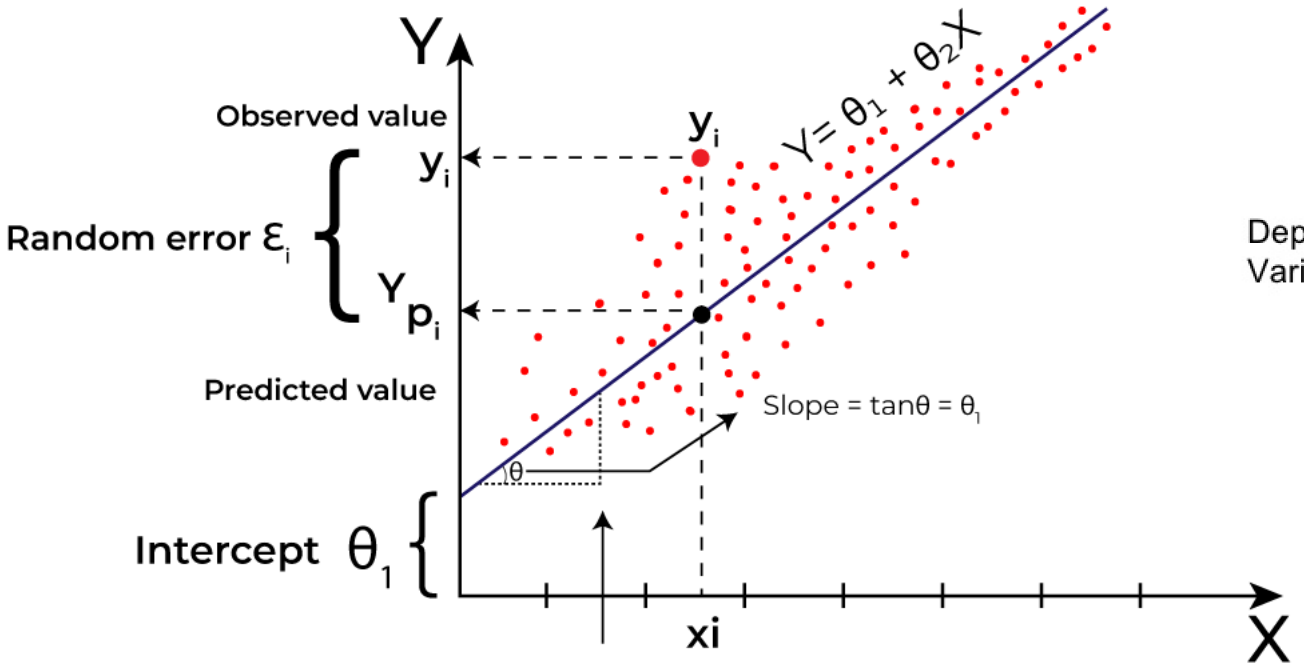


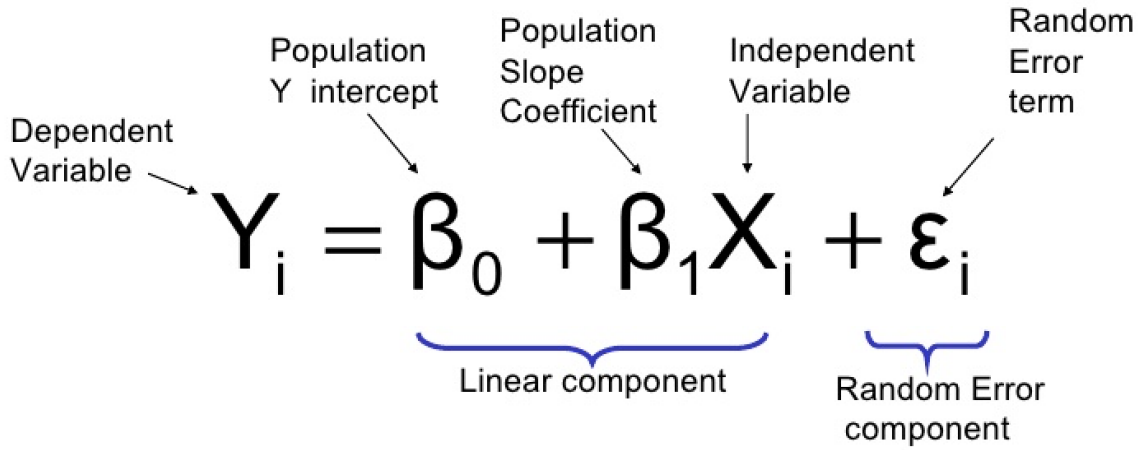
If we know X, we can predict Y using the model

Types of supervised learning

- Regression: Predict a dependent variable with a continuous value
- Classification: Predict a dependent variable with discrete value

Linear Regression





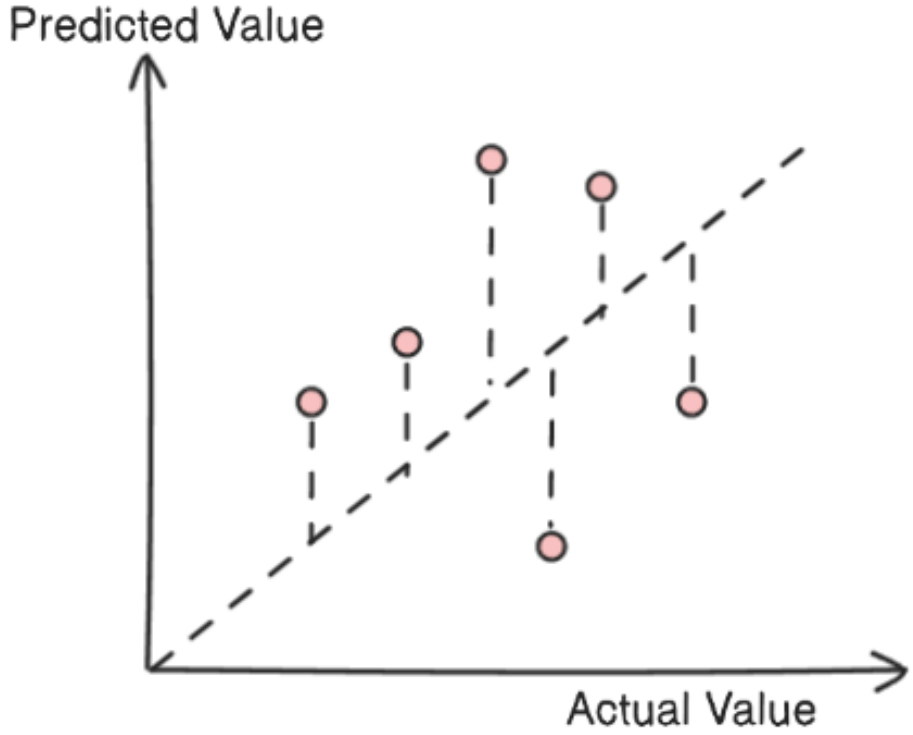
Regression Model Metrics

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$



Training vs testing data

Training

(Inputs, Outputs) → **ML Model**

Prediction

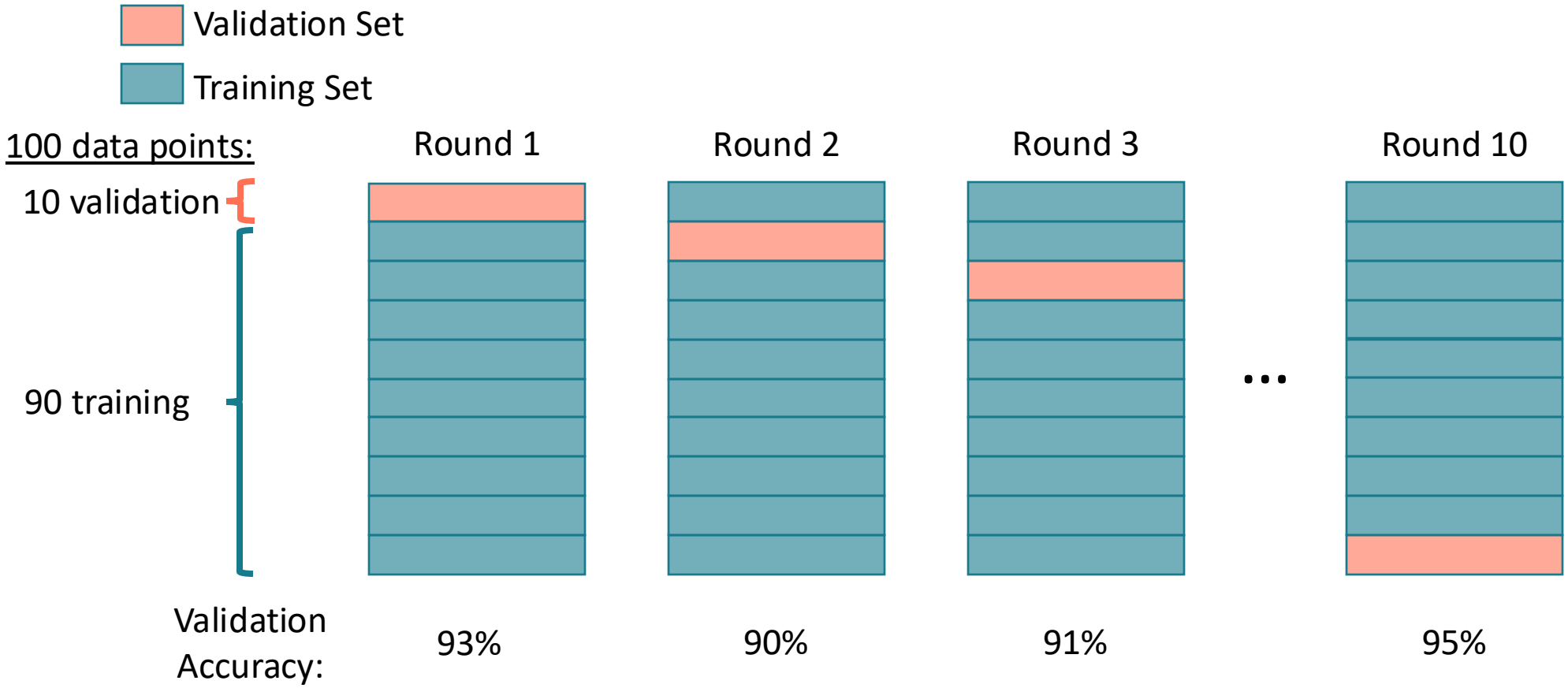
Inputs → **ML Model** → Outputs

Evaluation

Do not use the same data set that you trained on!

(Predicted Outputs, True Outputs) → Model Accuracy

Cross-Validation



Final Accuracy = Average(Round 1, Round 2,…)

Importance of scaling

- IMPORTANT: Scale Train data and use that transformation on test data. Do not scale the whole data.

Types:

Standardization:

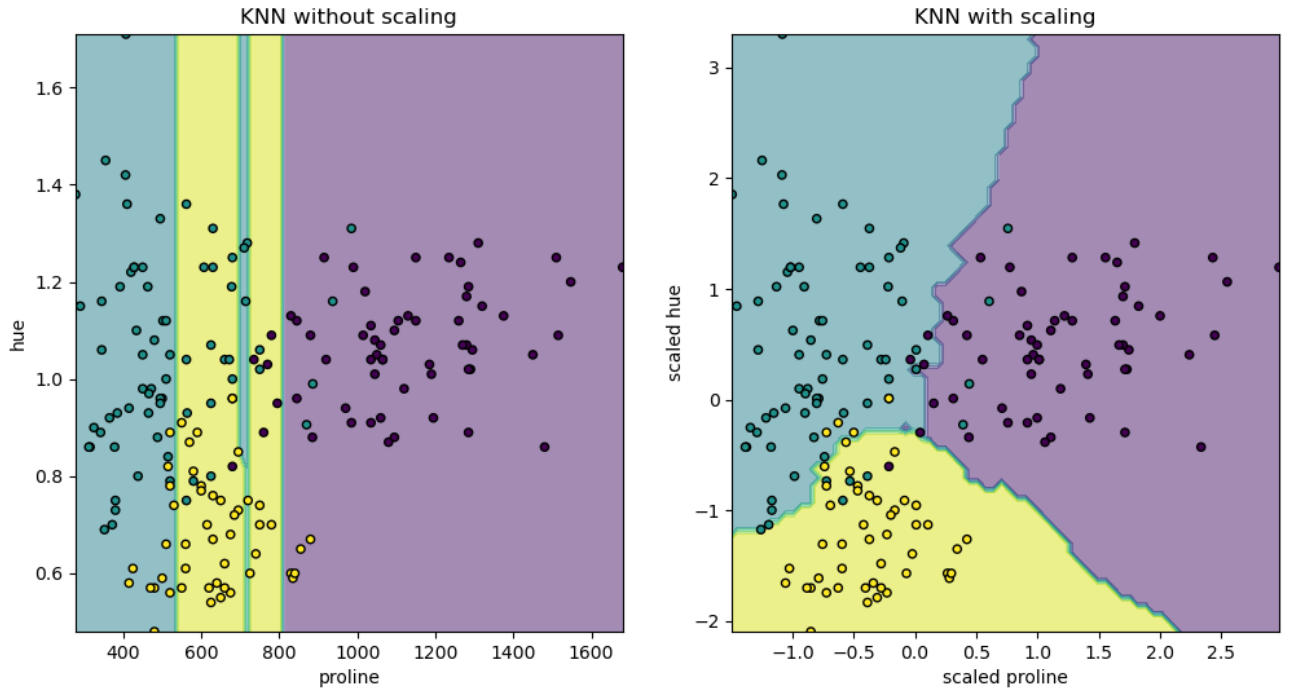
$$x' = \frac{x - \bar{x}}{\sigma}$$

Mean Normalization:

$$x' = \frac{x - \bar{x}}{max(x) - min(x)}$$

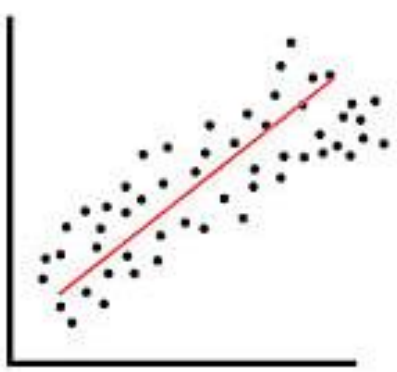
Min-Max Scaling:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$



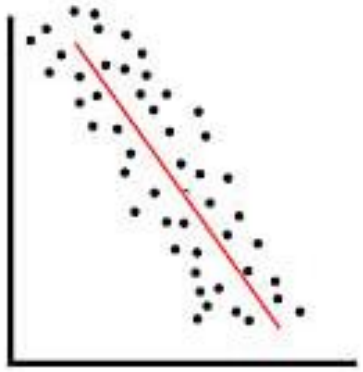
Pearson Correlation: Feature elimination

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$



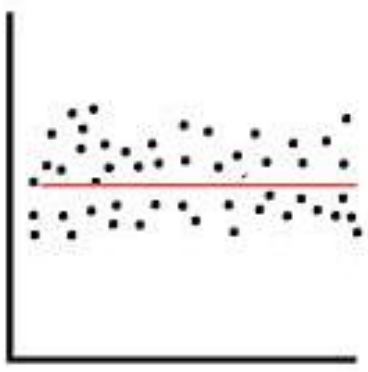
Positive Correlation

r = 1
Perfect Positive Correlation



Negative Correlation

r = -1
Perfect Negative Correlation



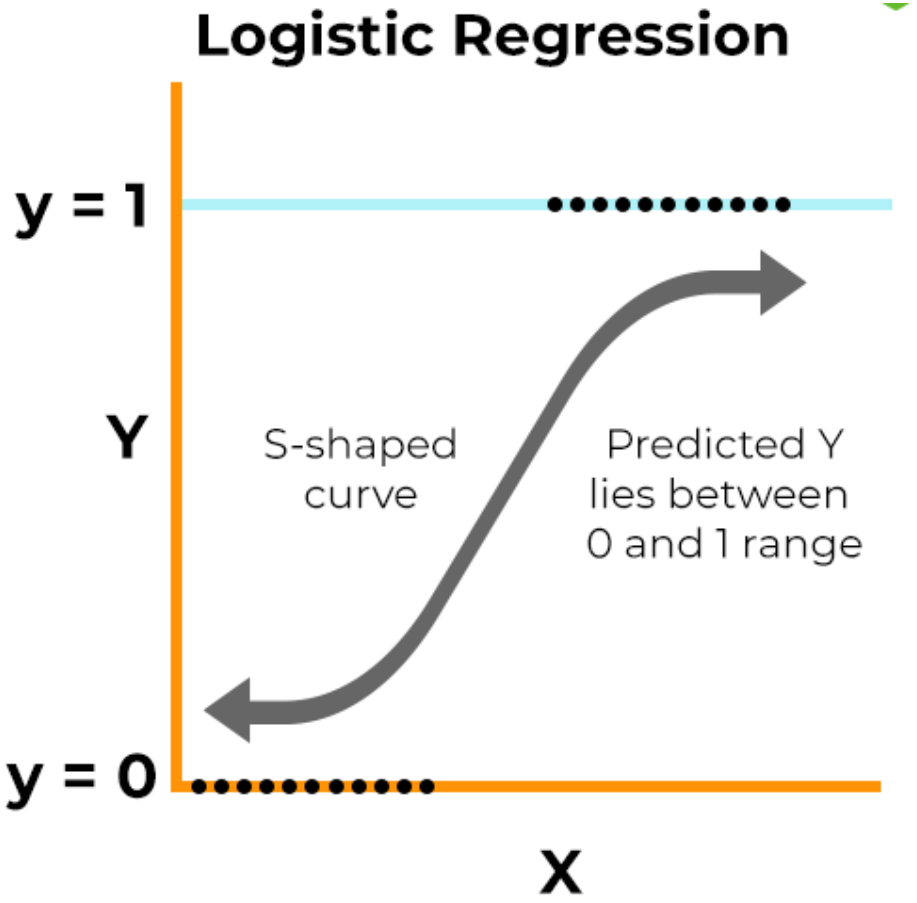
No Correlation

r = 0

- You can consider all the features that correlate > 0.5 with the target
- Eliminate features that have high correlation among each other

Logistic Regression: A classification method

- This method uses the concept of odds to predict a binary value for the dependent variable



Sigmoid curve Equation

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

A threshold will determine if the value is 0 or 1

Classification Model Metrics

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Accuracy $A = \{ TP + TN \} / Total$

Precision $P = TP/\{TP + FP\}$

Precision is the ratio of true positive and the total number of instances predicted as positive by ML model

Recall $R = TP/\{TP + FN\}$

Recall defines the sensitivity of the model i.e., the ratio of true positive and the total number of positive instances in the test data

f1 score $f1 = 2*P*R/(P+R)$

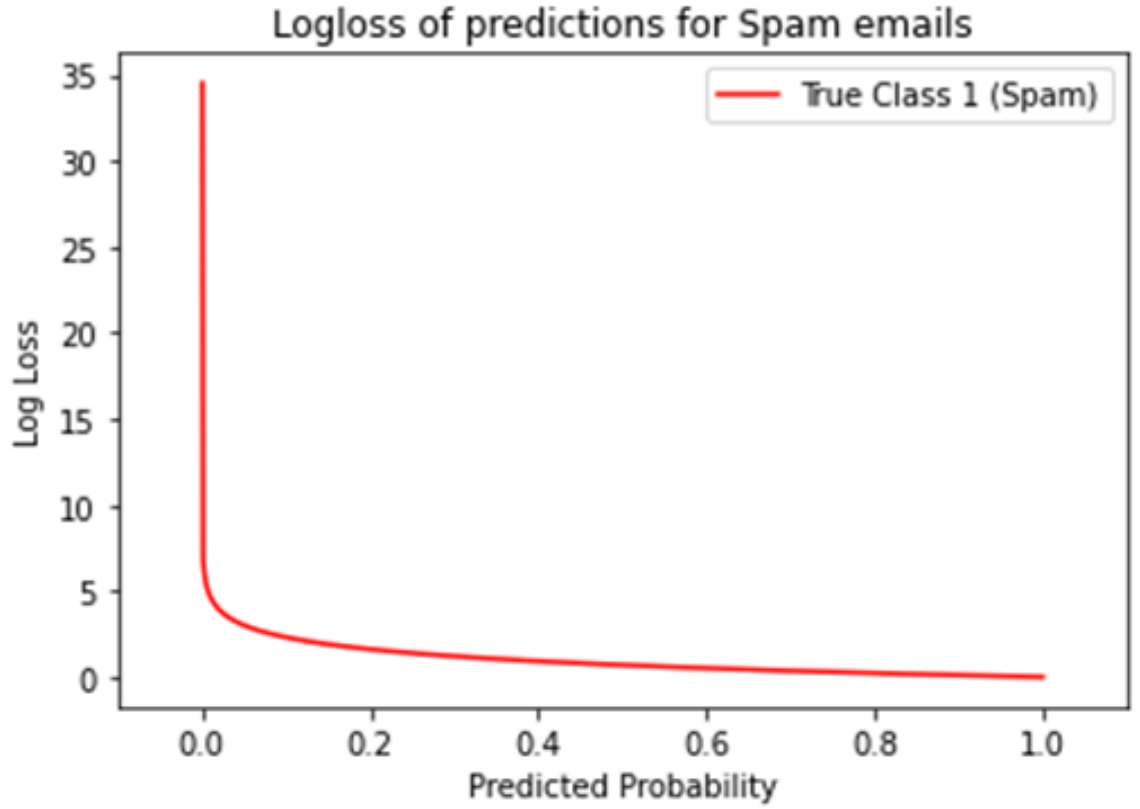
F1 score is harmonic mean of precision and recall

Classification Model Metrics: LogLoss

$$Logloss = \frac{1}{N}\sum_{i=1}^{N} logloss_i$$

$$Logloss_i = -[y_i \, ln \, p_i + (1 - y_i) \, ln(1 - p_i)]$$

The lower the Log loss the better is the fit



Other ML algorithms typically used for modeling

Regression Algorithms

Minimizes Complexity

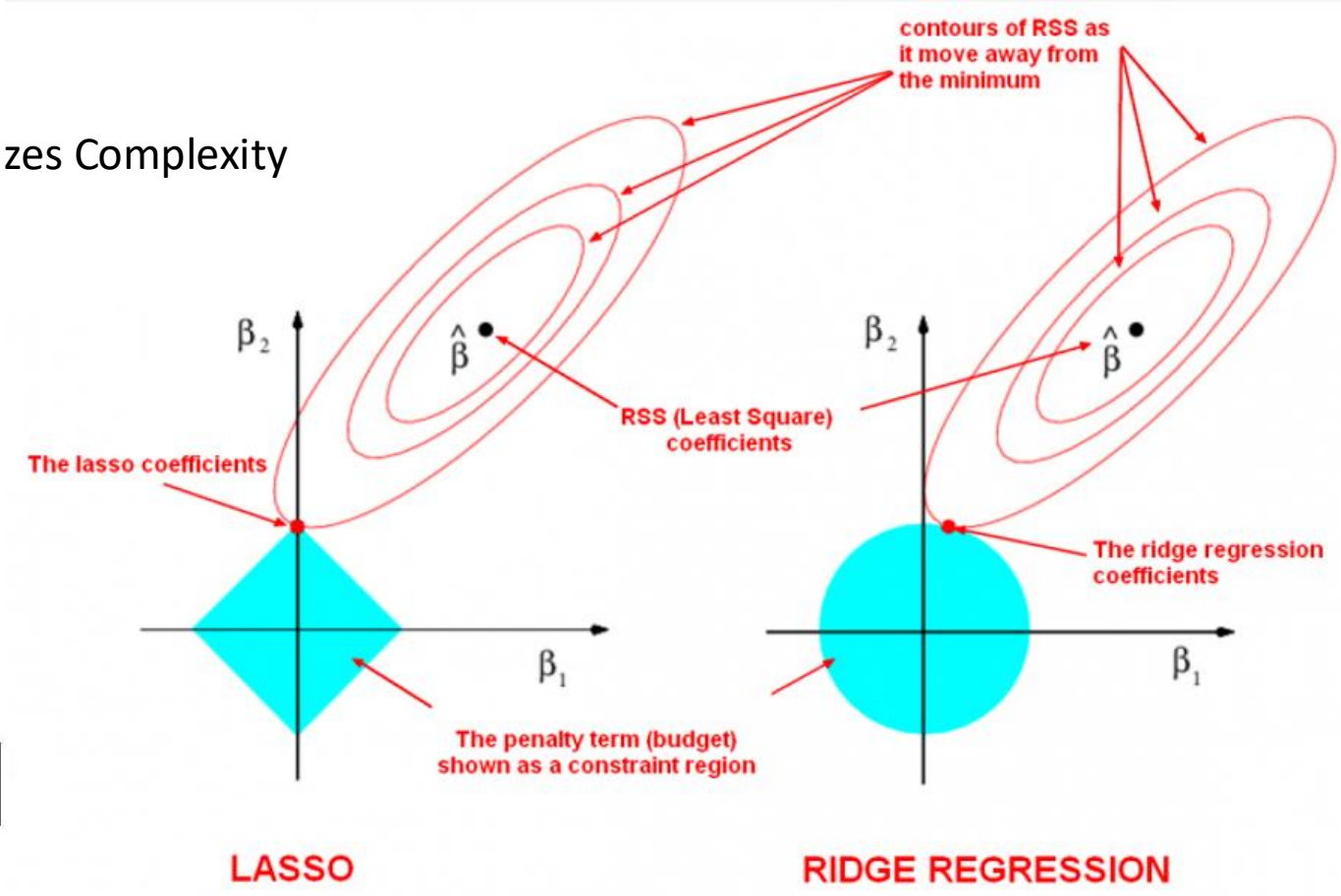
- Ridge regression

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso regression

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Automatic feature selection

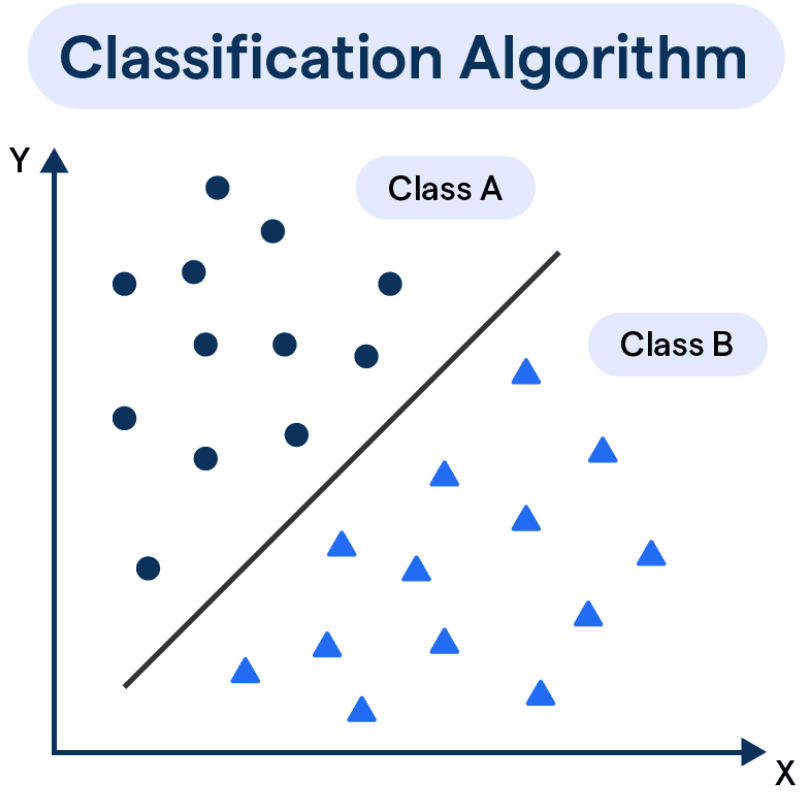


Other ML algorithms typically used for modeling

Classification Algorithms :

- Support Vector Machines
- kNN
- Random Forest

- Artificial Neural Networks



They can be used for Regression as well