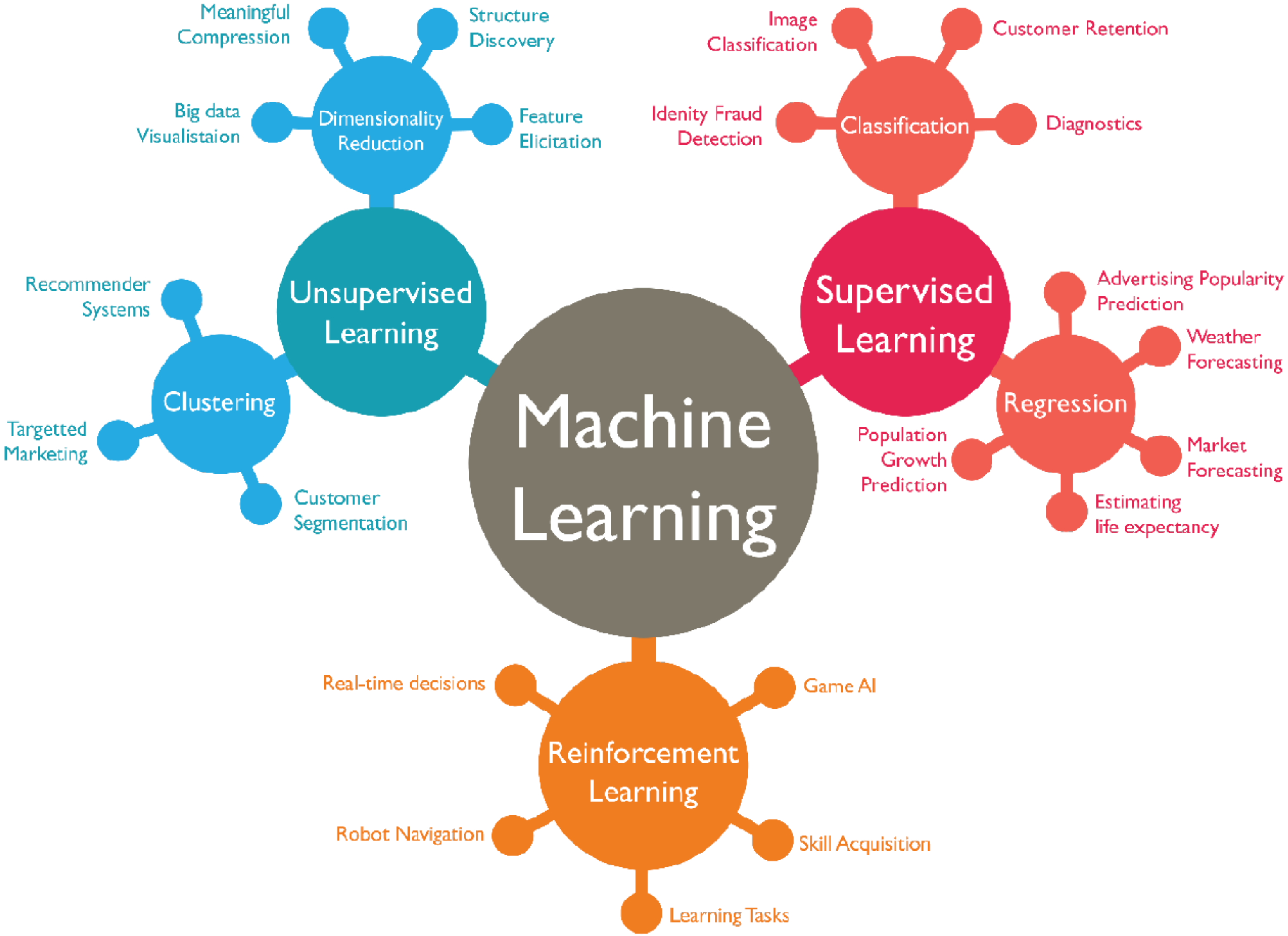
Initiation à l’apprentissage automatique en science des matériaux

2. Fundamentals in Machine Learning

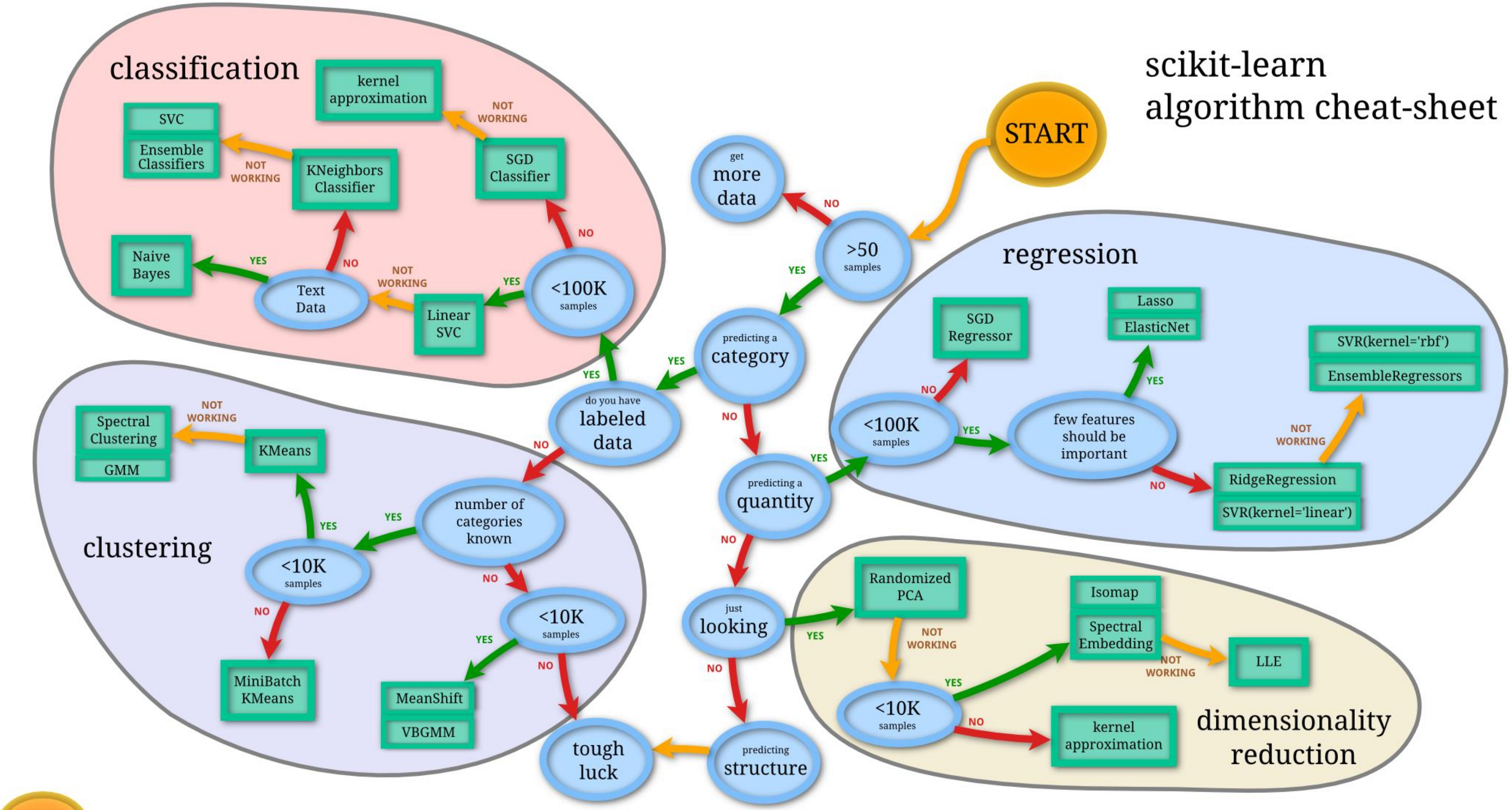
J.-C. Crivello, LINK : jean-claude.crivello@cnrs.fr
C. Barreteau, ICMPE : celine.barreteau@cnrs.fr
S. Junier, ICMPE : sebastien.junier@cnrs.fr

https://link.cnrs.fr/ML/

Machine learning



Machine learning



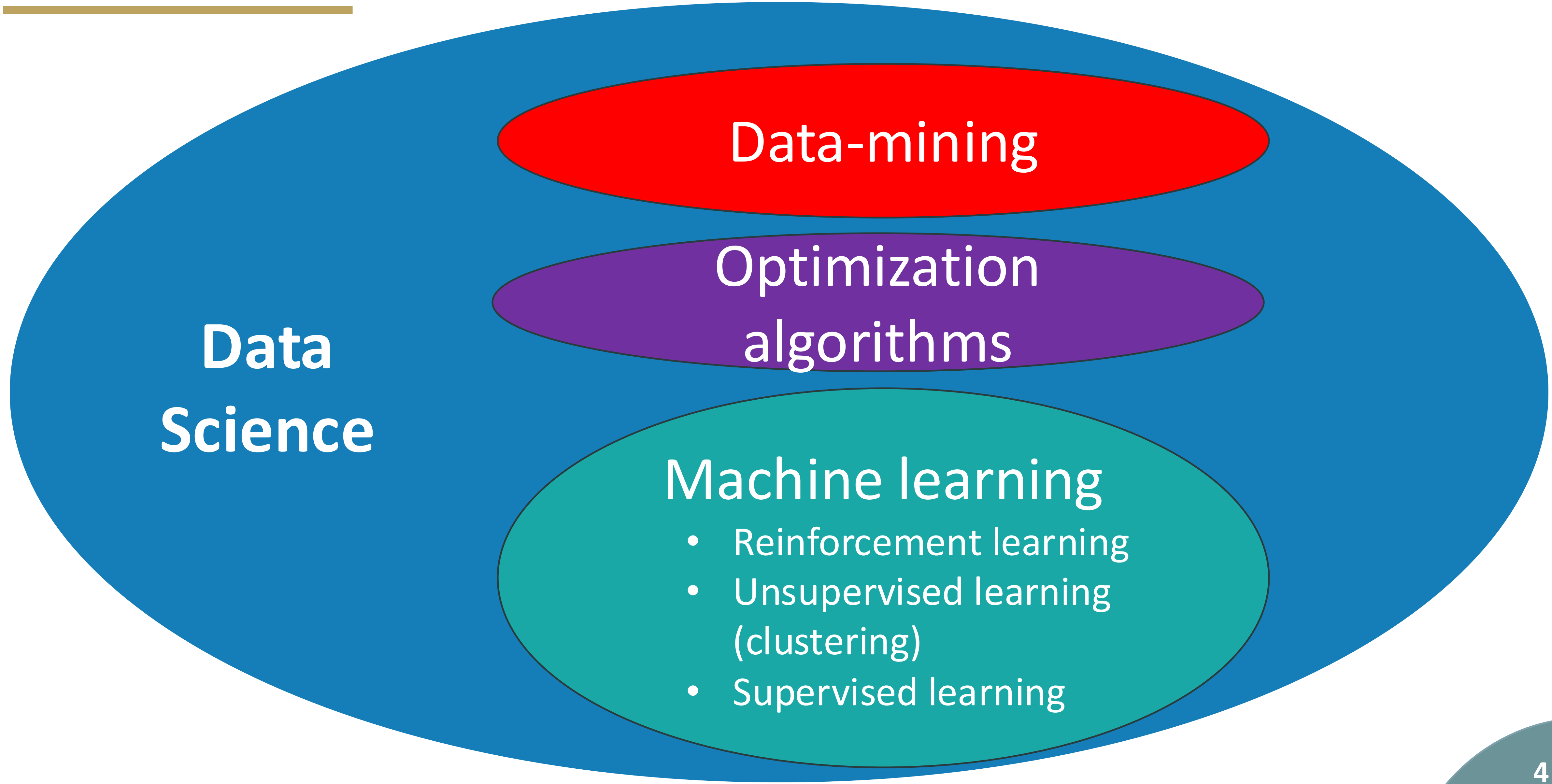
Back

scikit learn

Choosing the right estimator?

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Several approaches of the data science



Data Science

Data-mining

Optimization algorithms

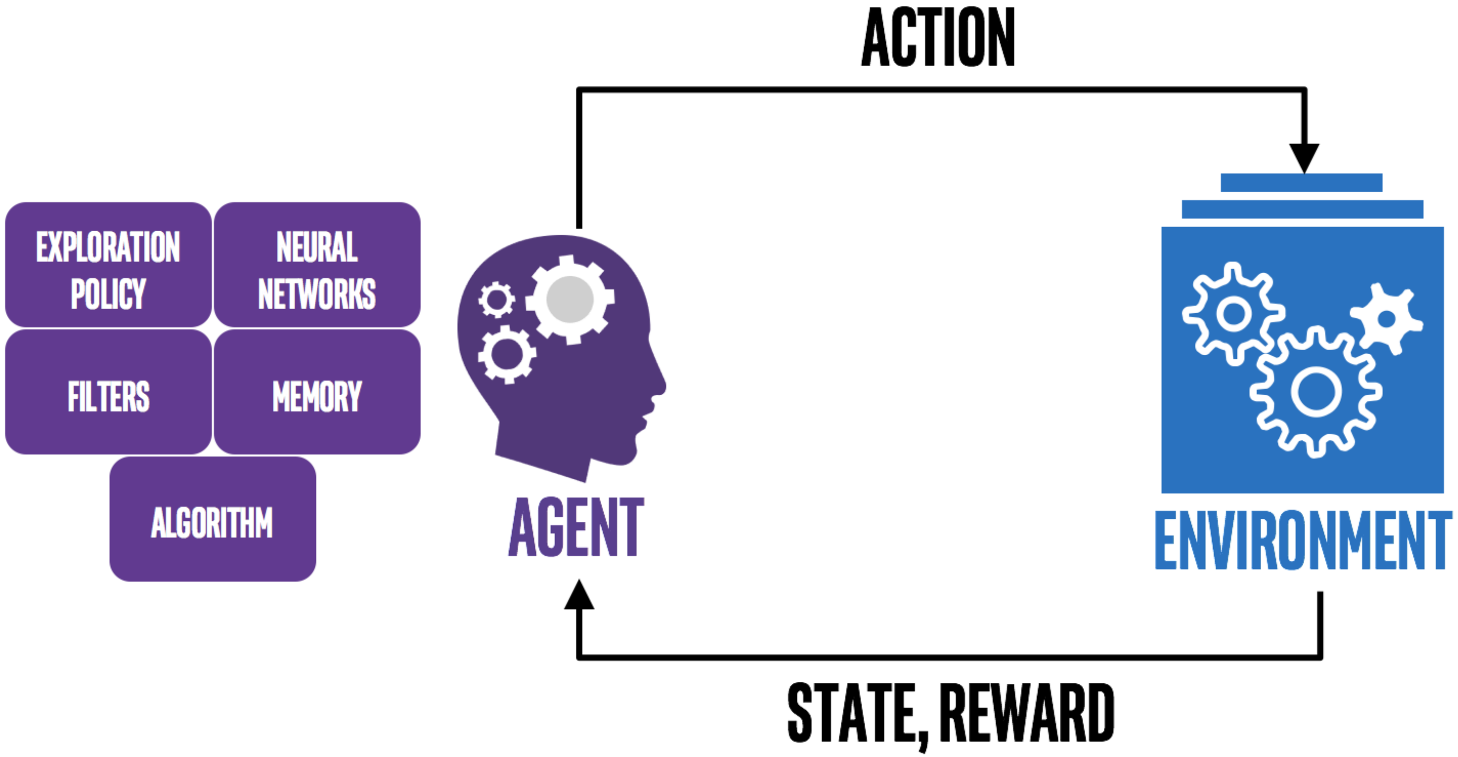
Machine learning

- Reinforcement learning
- Unsupervised learning (clustering)
- Supervised learning

1. Reinforcement learning

RL is learning from experiences.

RL teaches an agent how to choose an action from its action space, within a particular environment, in order to maximize rewards over time



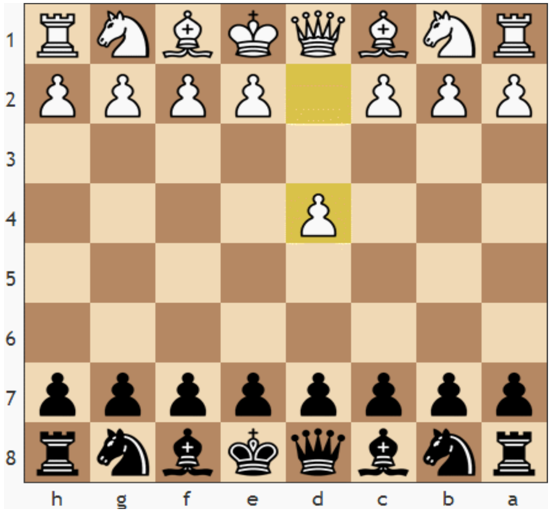
Game theory: Brute force VS RL

1997 – Deep Blue - IBM

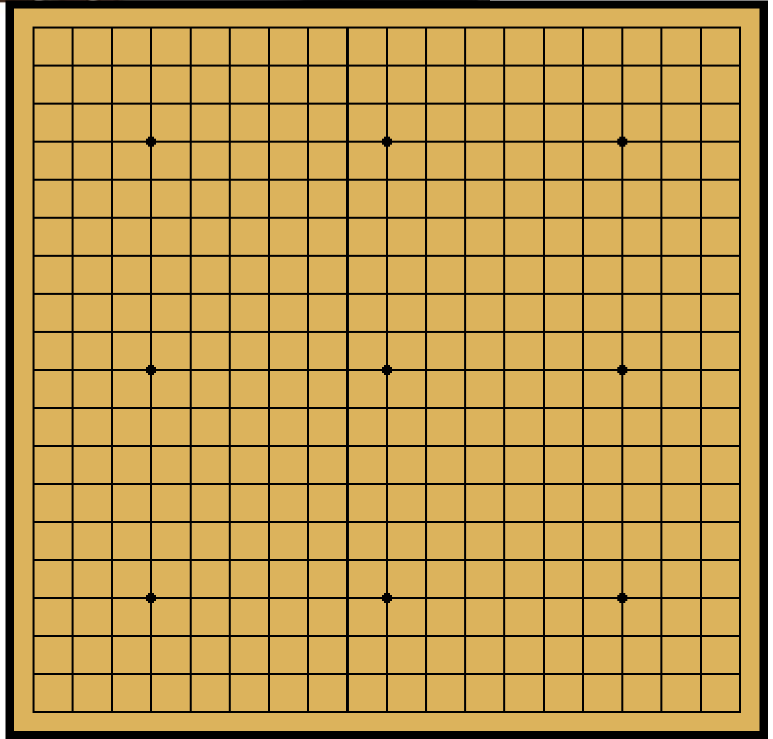
© Louie Psihoyos/CORBIS

2016
Alpha Go

Number of possible sequences:



35 ^ 80



250 ^ 150

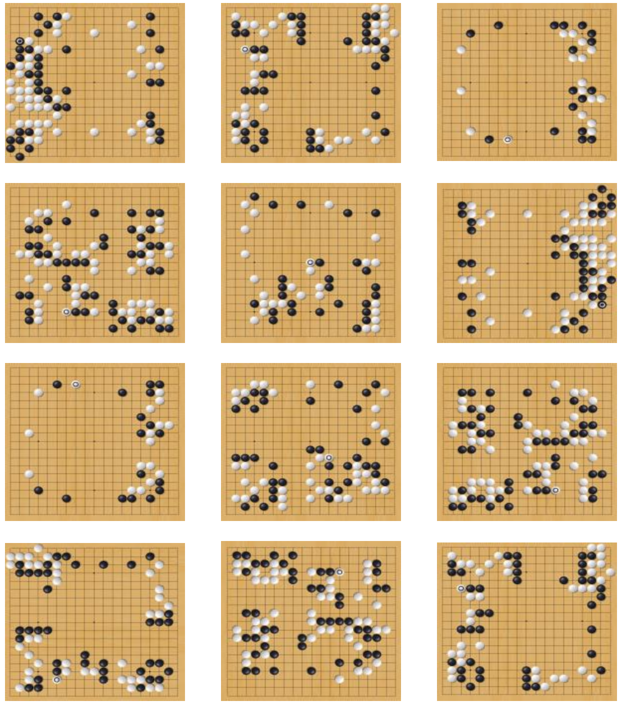
Mastering the game of Go

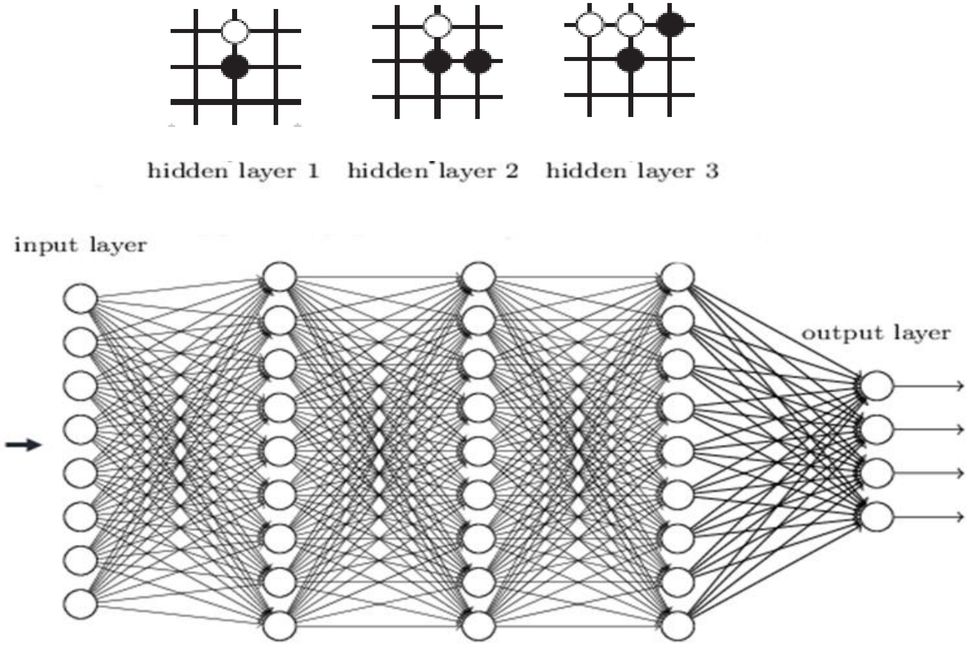
How to find the best local move for wining the whole game?

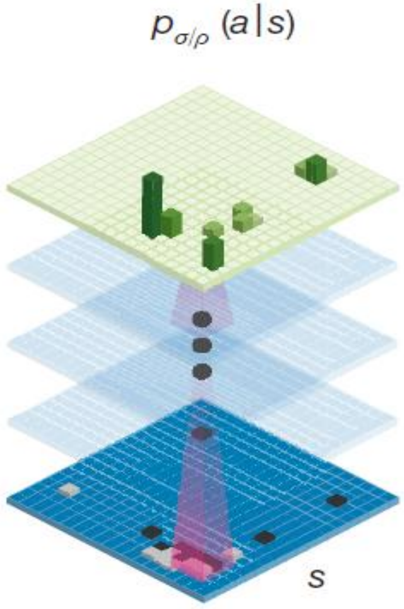
(1) Deep neural networks supervised training

(2) Monte Carlo tree search programs

(3) Alpha-go Zero (2017) : only reinforcement from scratch

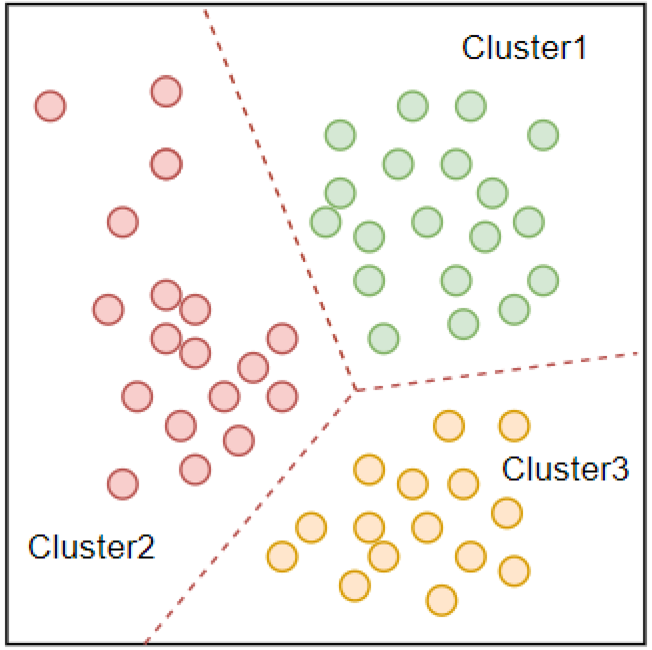


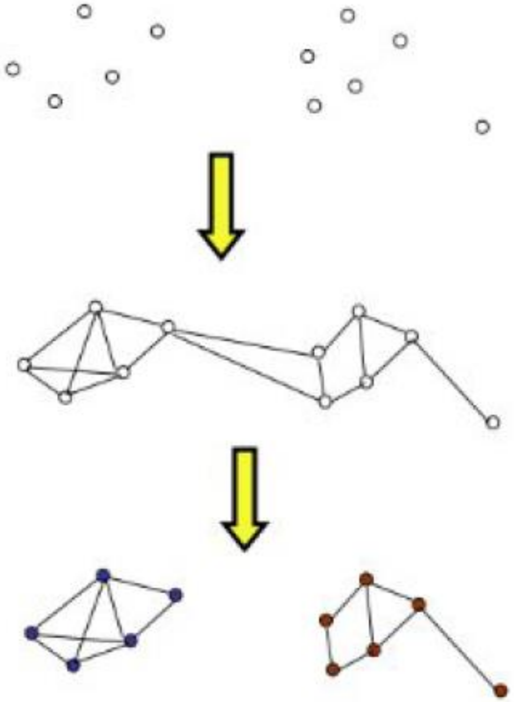




2. Unsupervised learning (clustering)

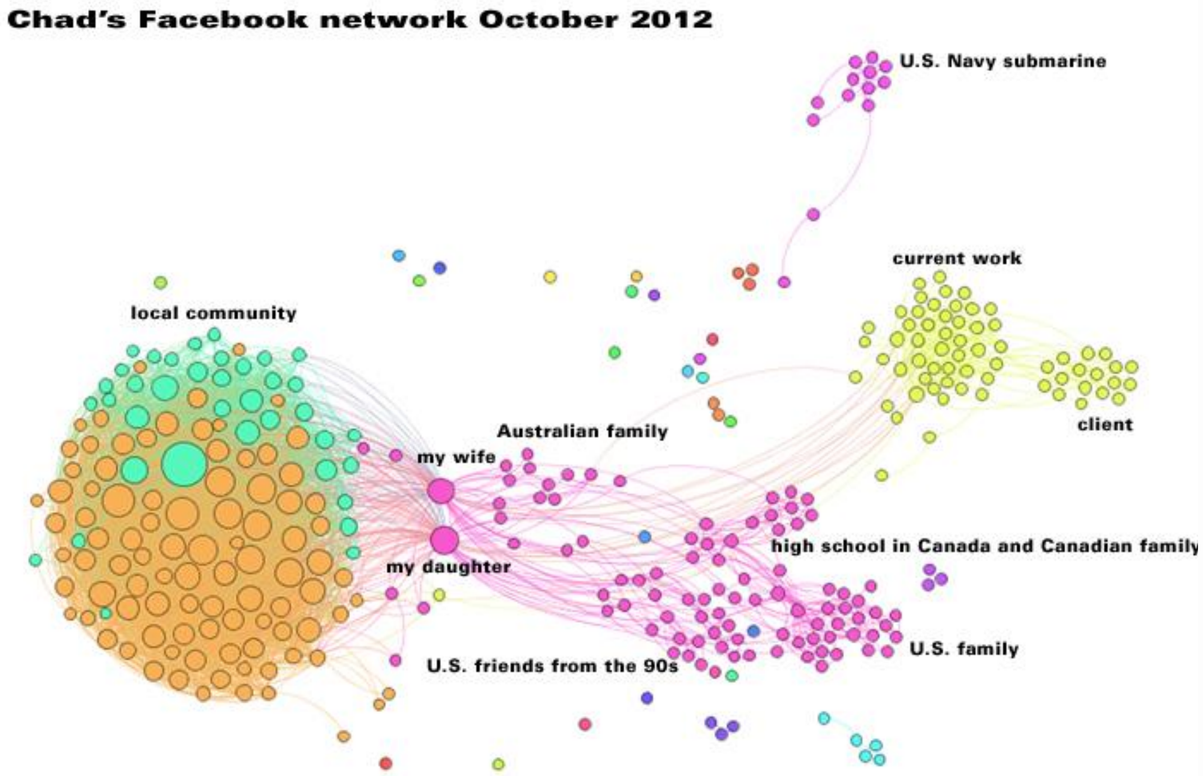
Unsupervised ML learns from a dataset without any labels. The algorithm can automatically classify or categorize the input data.



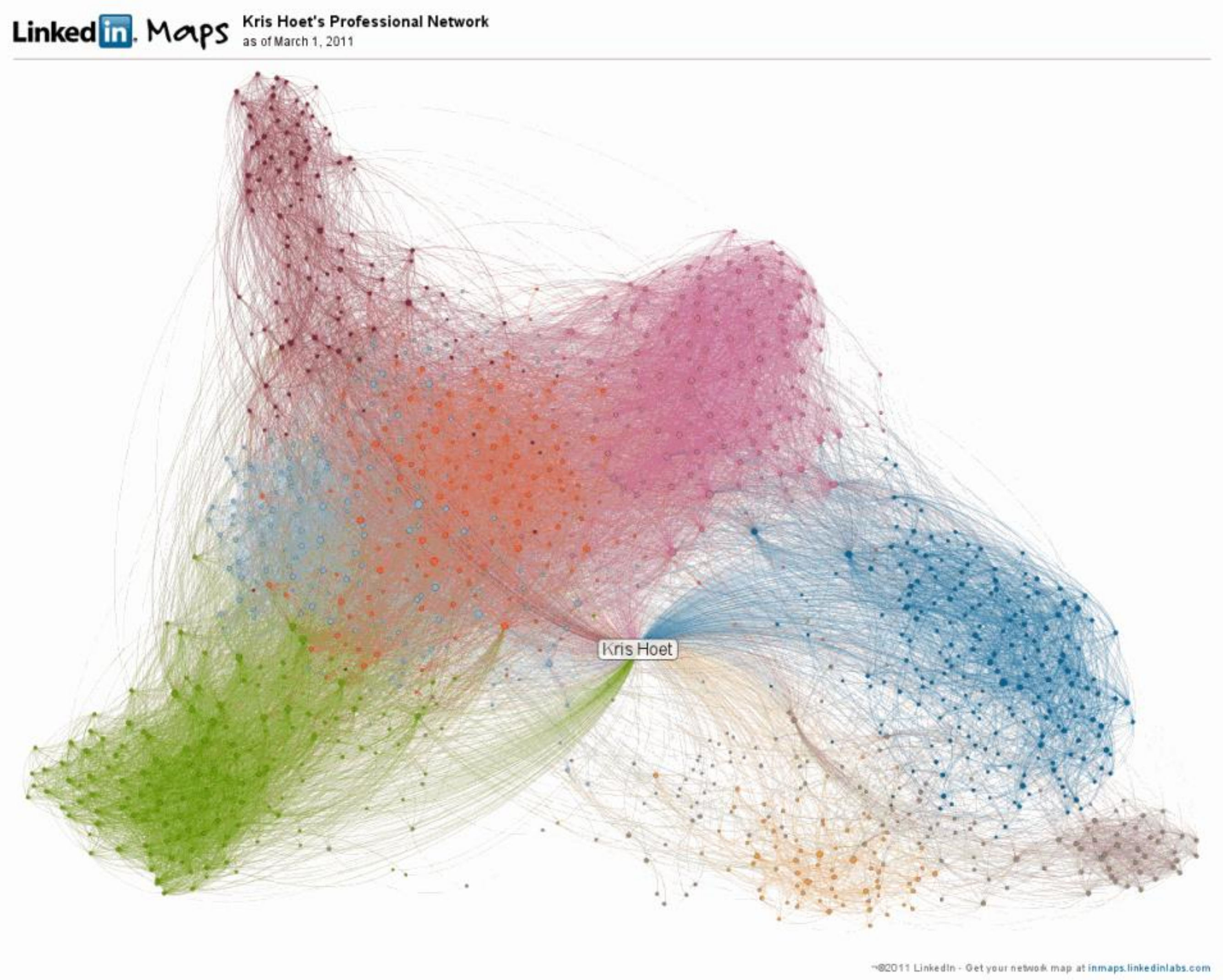


The application of unsupervised learning mainly includes cluster analysis, association rule or dimensionality reduce.

Social Network Analysis



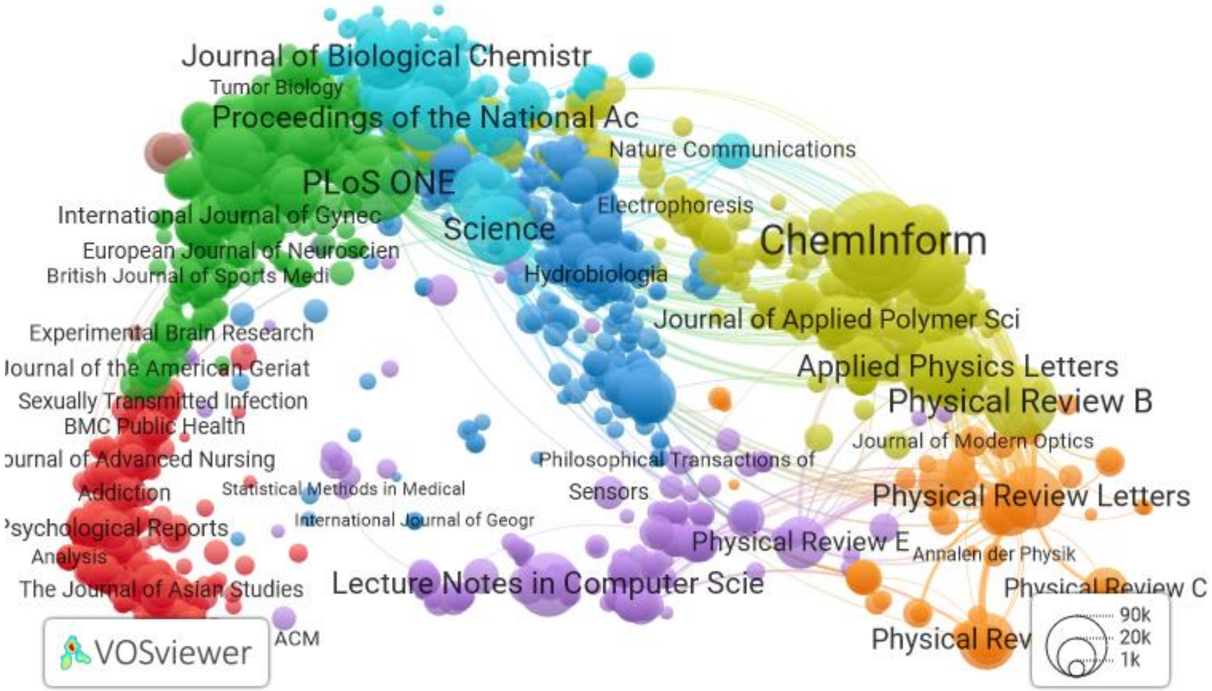
Chad's Facebook network October 2012



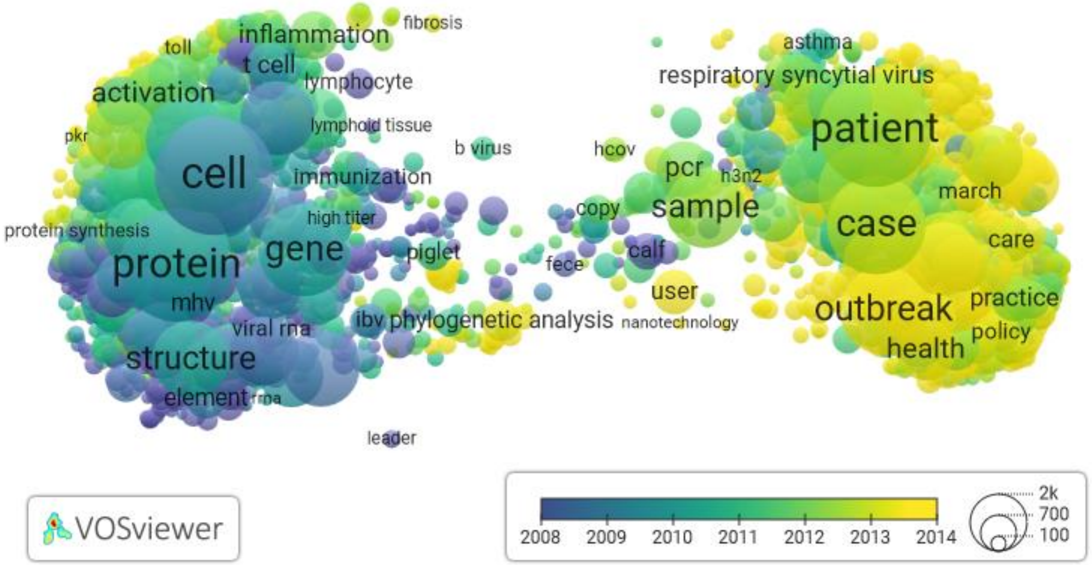
Kris Hoet's Professional Network
as of March 1, 2011

©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

Social Network Analysis



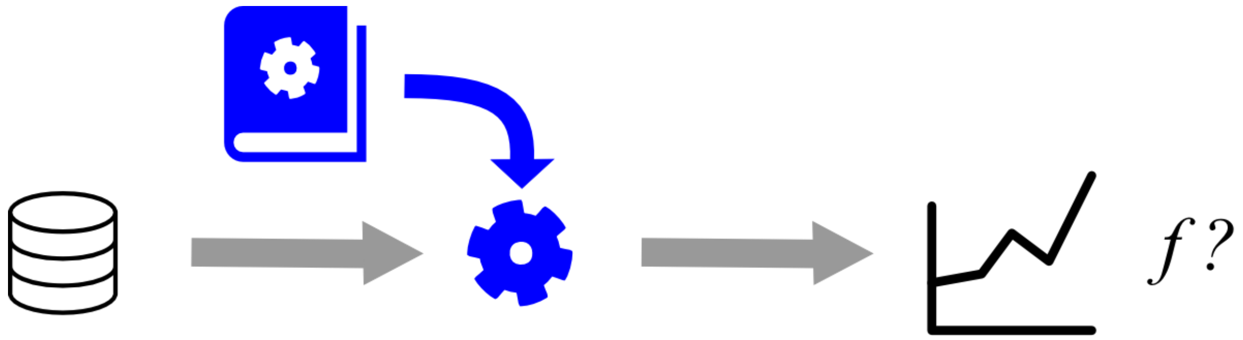
Citation network of journals



Co-occurrence network of terms in COVID-19 articles

3. Supervised learning

Supervised ML learns from a trained tagged dataset, builds a function, predicts the output based on the function.



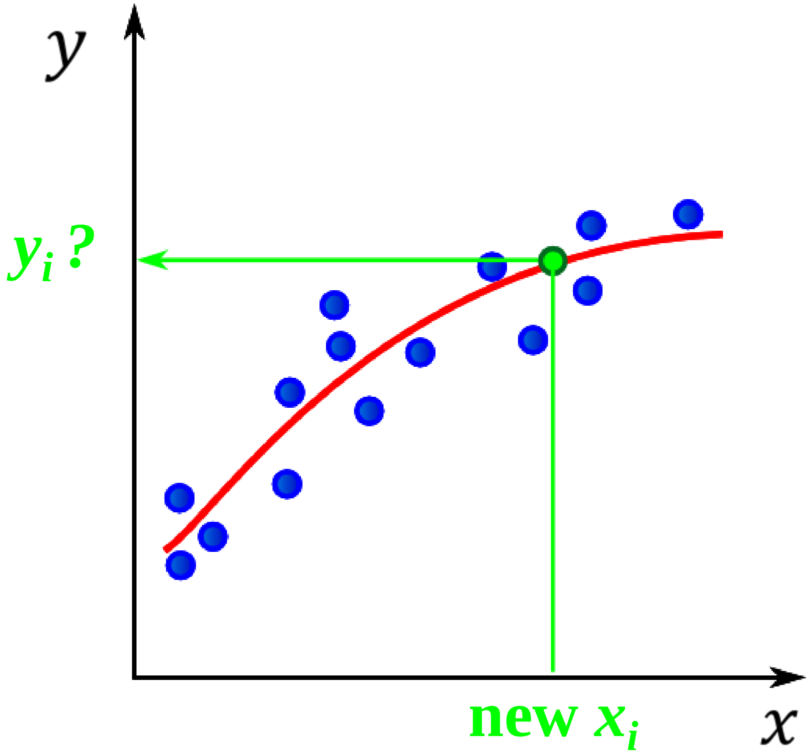
***f*(x) ?**

(i) quantitative variables:
→ Regression

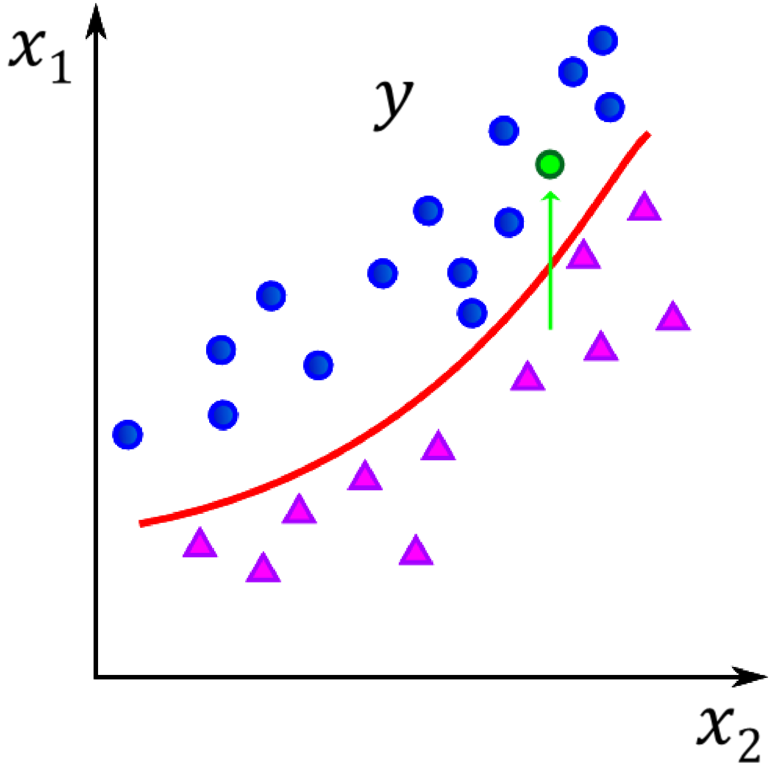
(ii) qualitative variables:
→ Classification

3. Supervised learning

The training dataset often consists of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), the output of the function can be **regression** or **classification**.

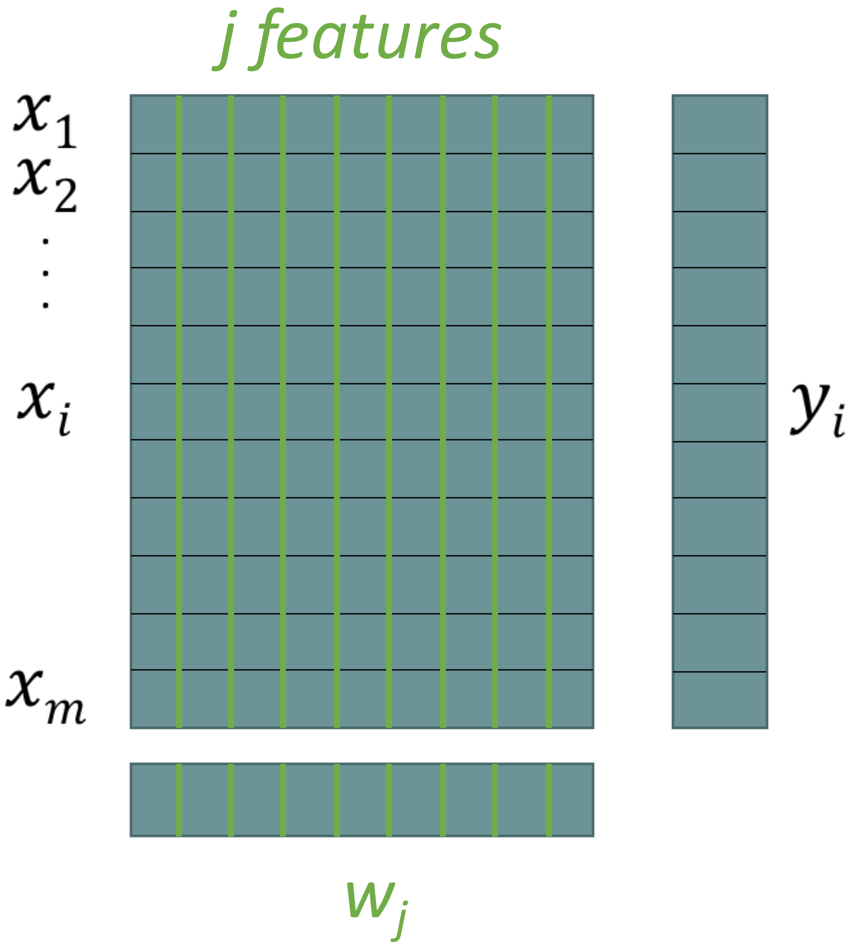


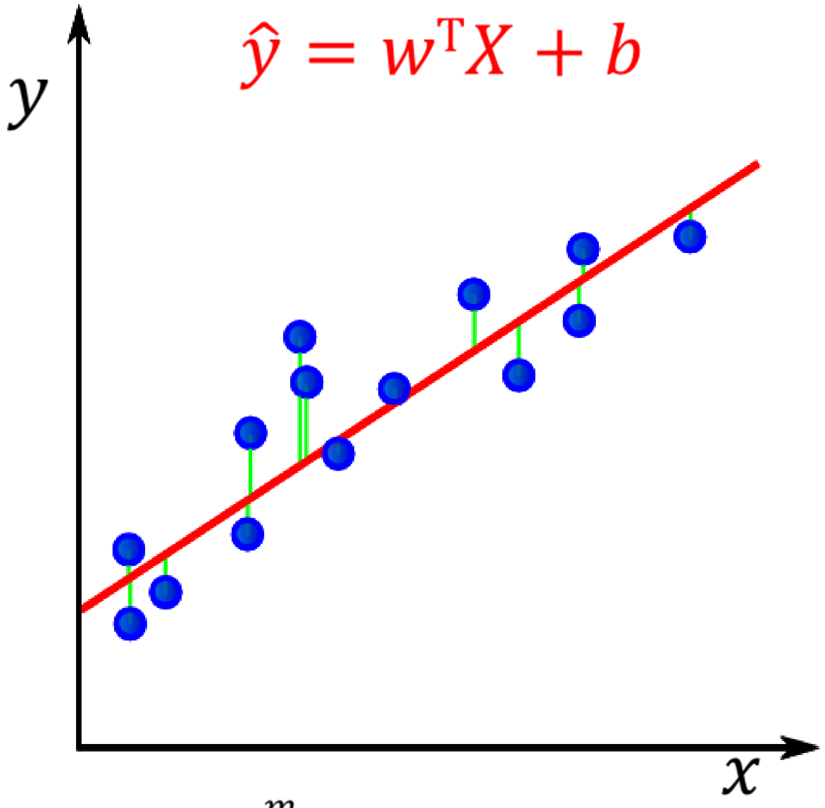
Regression



Classification

Optimization of the learning

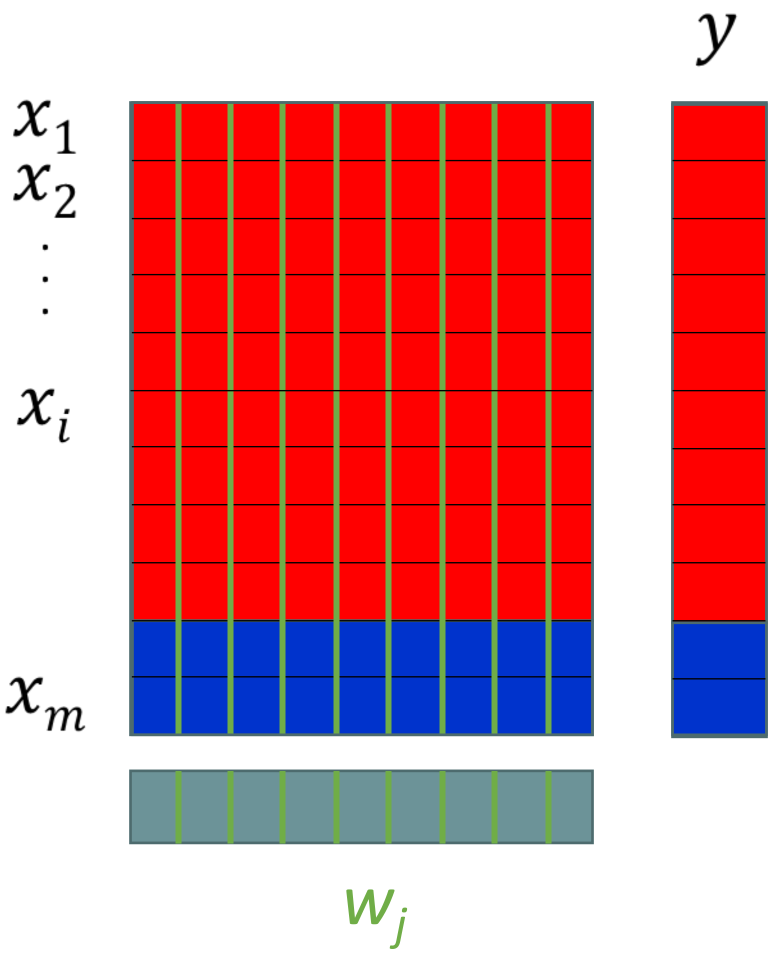




$$\mathrm{SSR} = \sum_{i}^{m} (\hat{y}_i - yi)^2$$

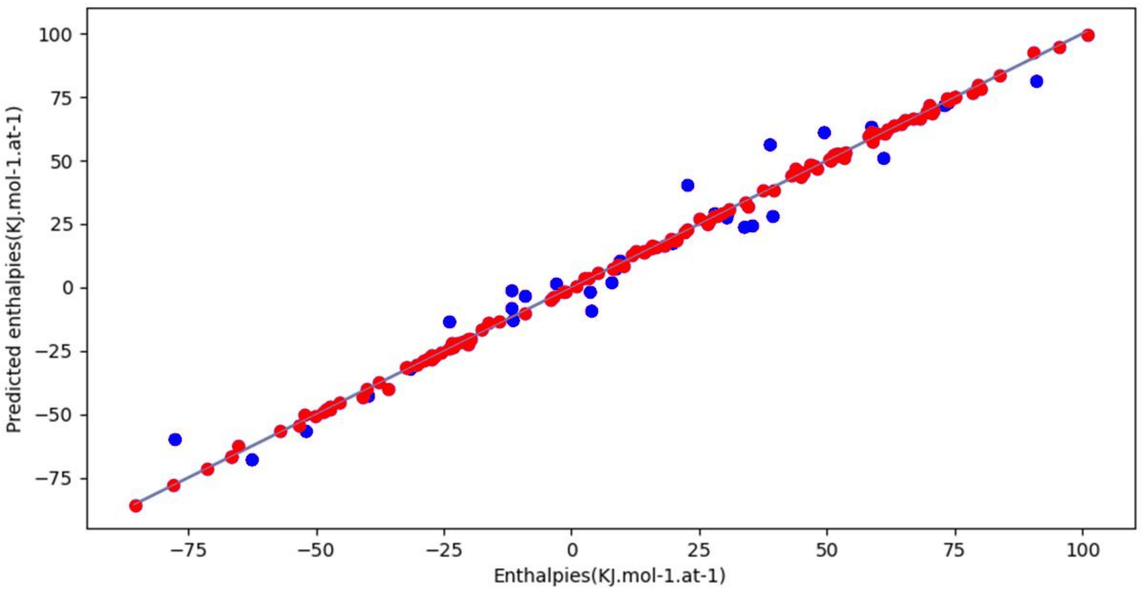
$$R^2 = 1 - \sum \frac{(\hat{y}_i - yi)^2}{(\bar{y}_i - yi)^2}$$

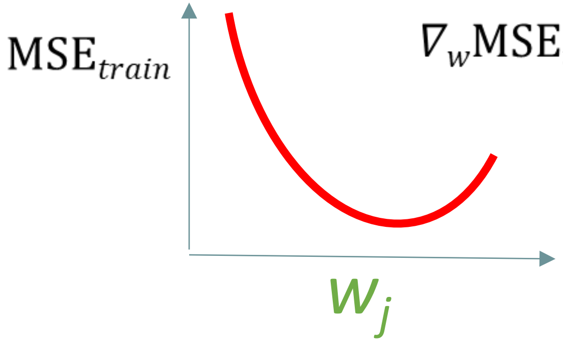
Estimation of the learning



training set

testing set



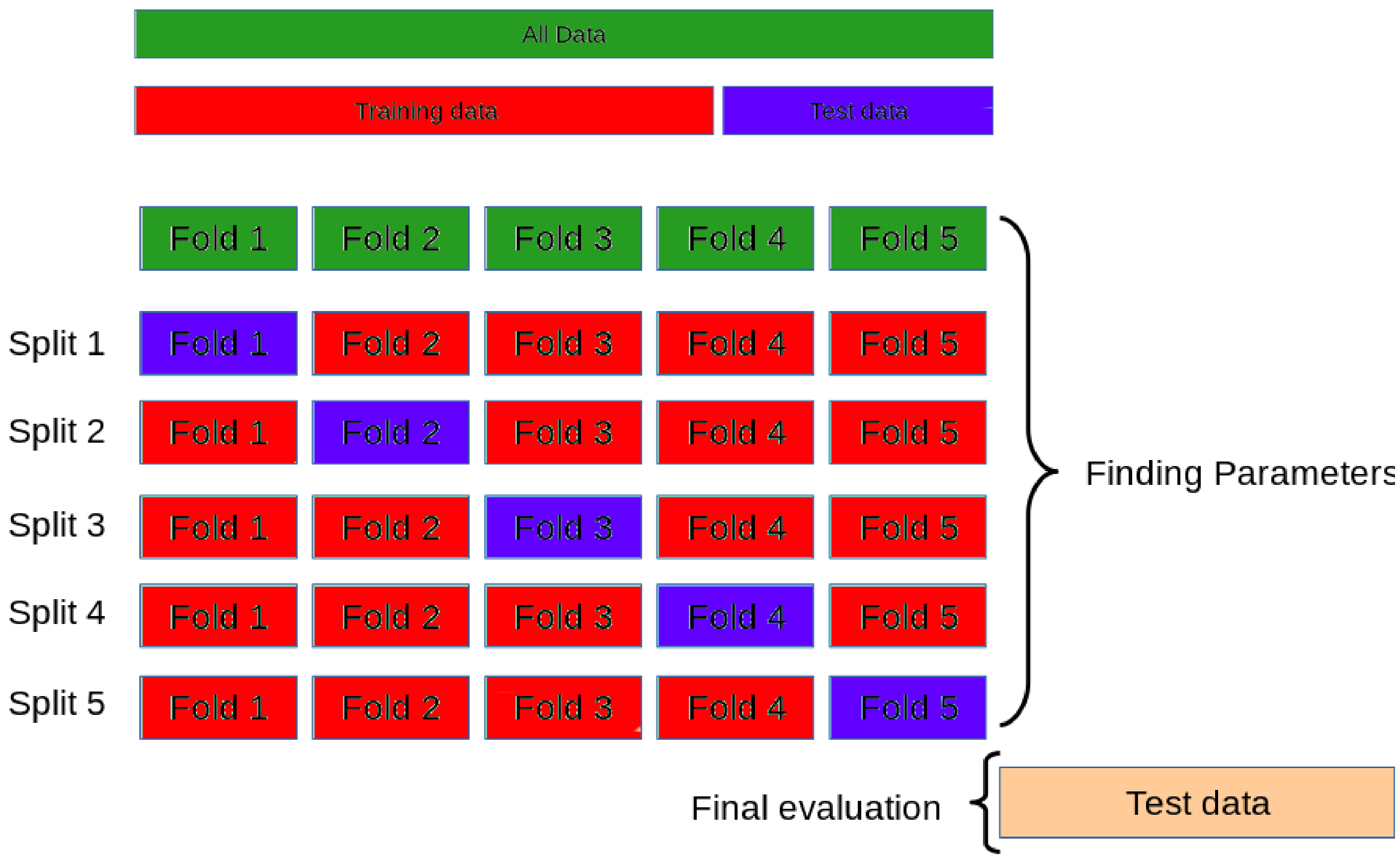


$$\nabla_w \mathrm{MSE}_{train} = 0$$

$$\mathrm{MSE}_{test} = \frac{1}{m}\sum_{i}^{m}(\widehat{y_{test}} - yte_{st})^2$$

$$\mathrm{RMSE}_{test} = \sqrt{\mathrm{MSE}_{test}}$$

Cross Validation



All Data

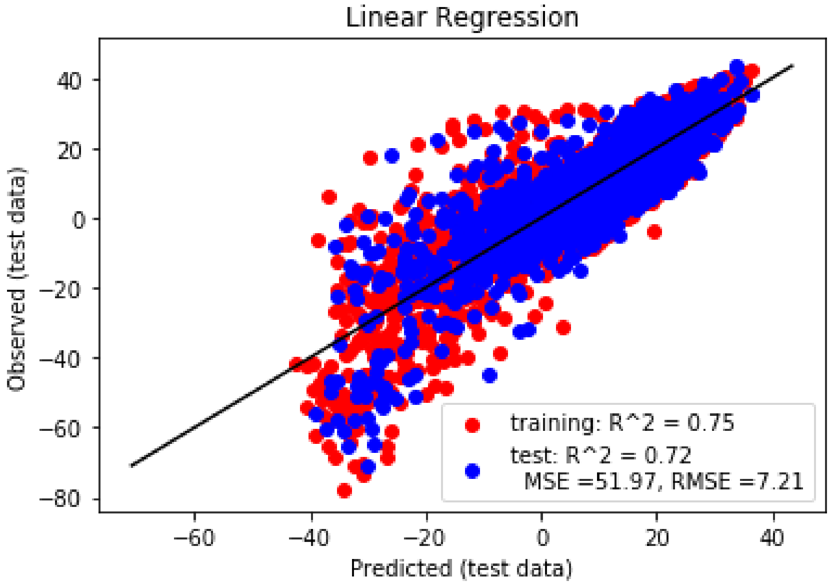
Training data | Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

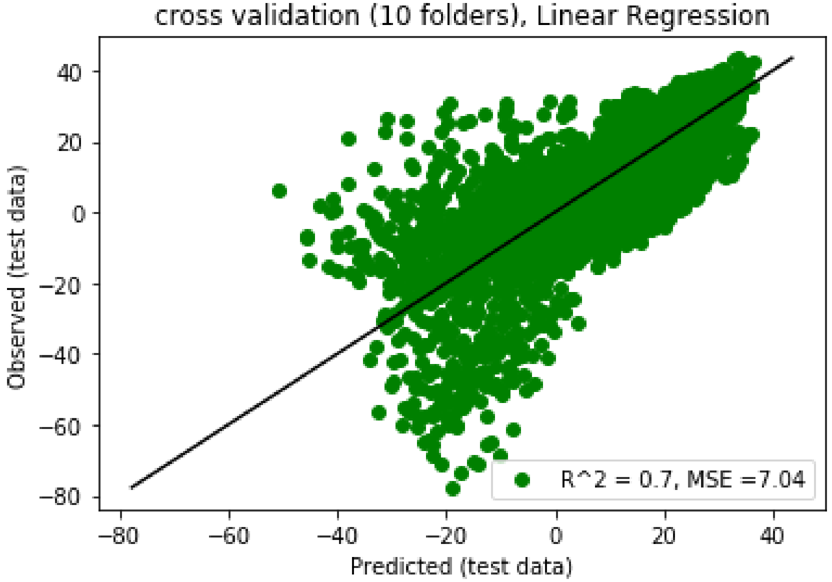
Finding Parameters

Final evaluation

Test data



Linear Regression

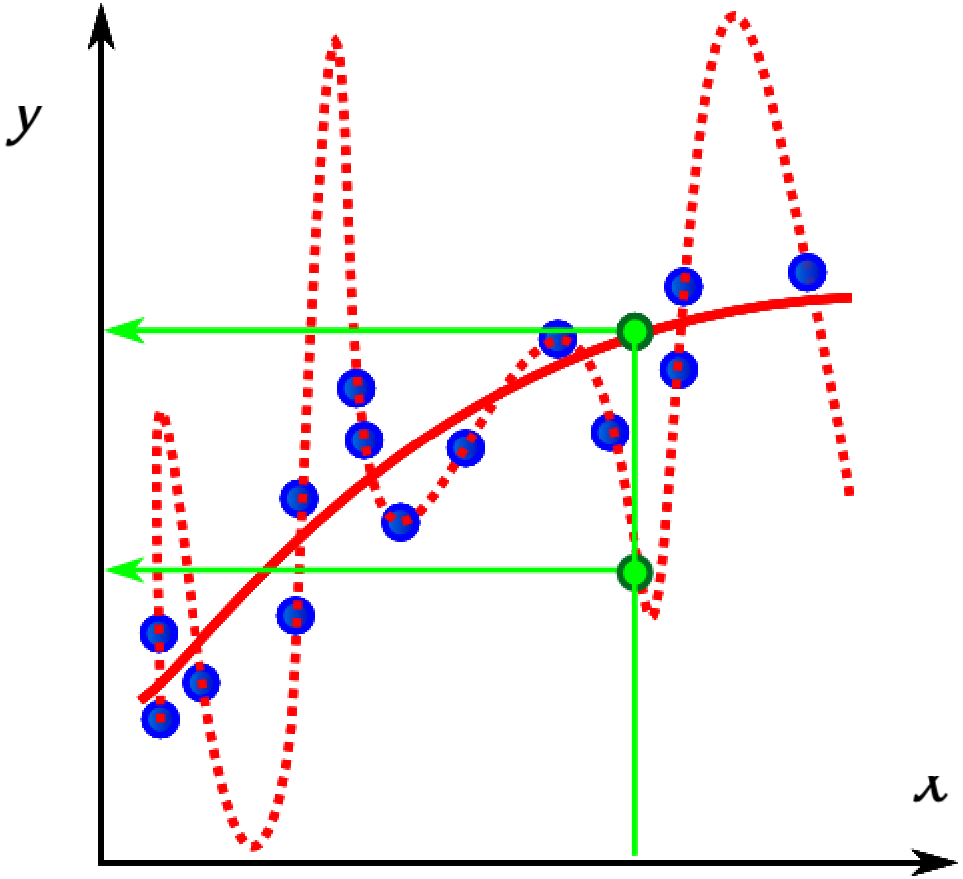


cross validation (10 folders), Linear Regression

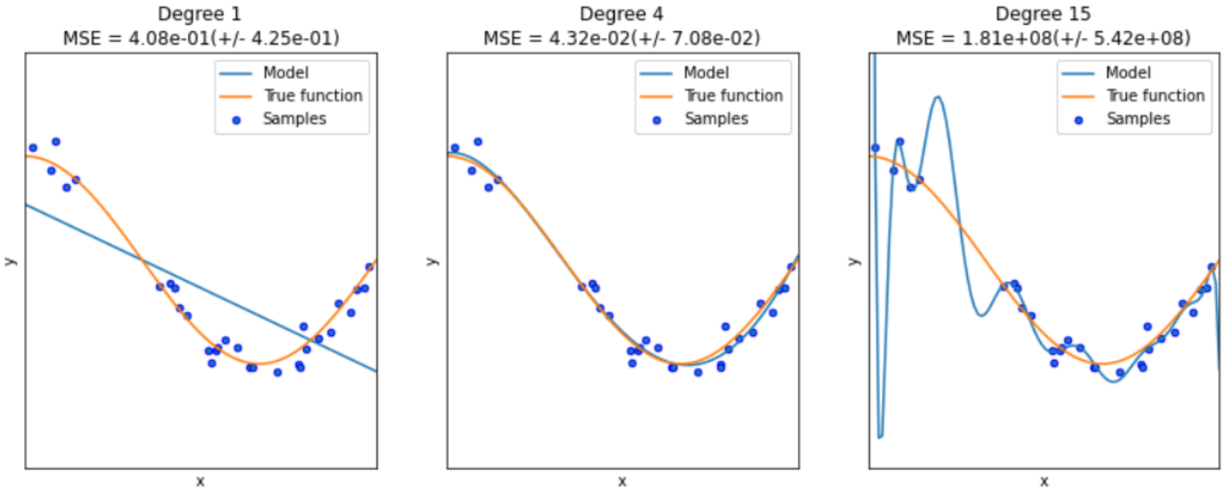
If large dataset, CV is need by repeating training/testing procedure under *k*-folder: partitions formed by splitting into *k* non-overlapping subsets.

Overfitting

= the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.



$$\hat{y} = \sum_{i}^{n} w_i X^i + b$$

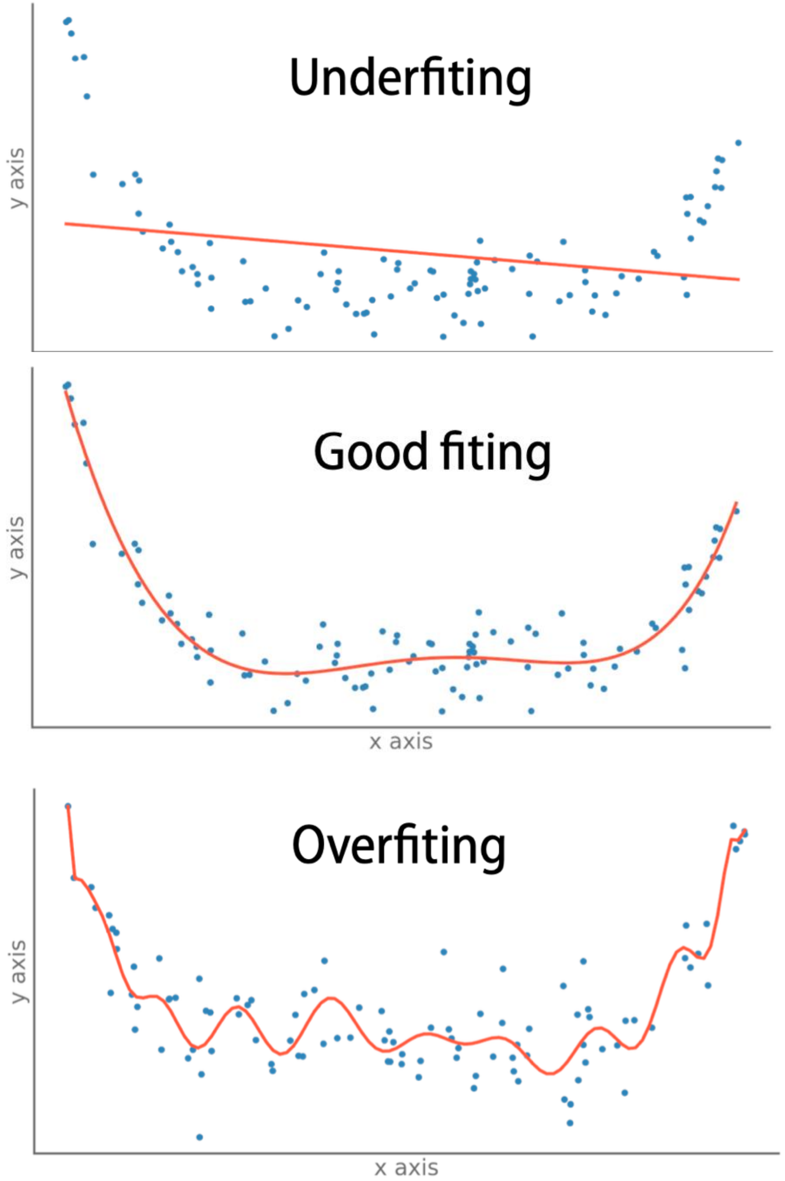


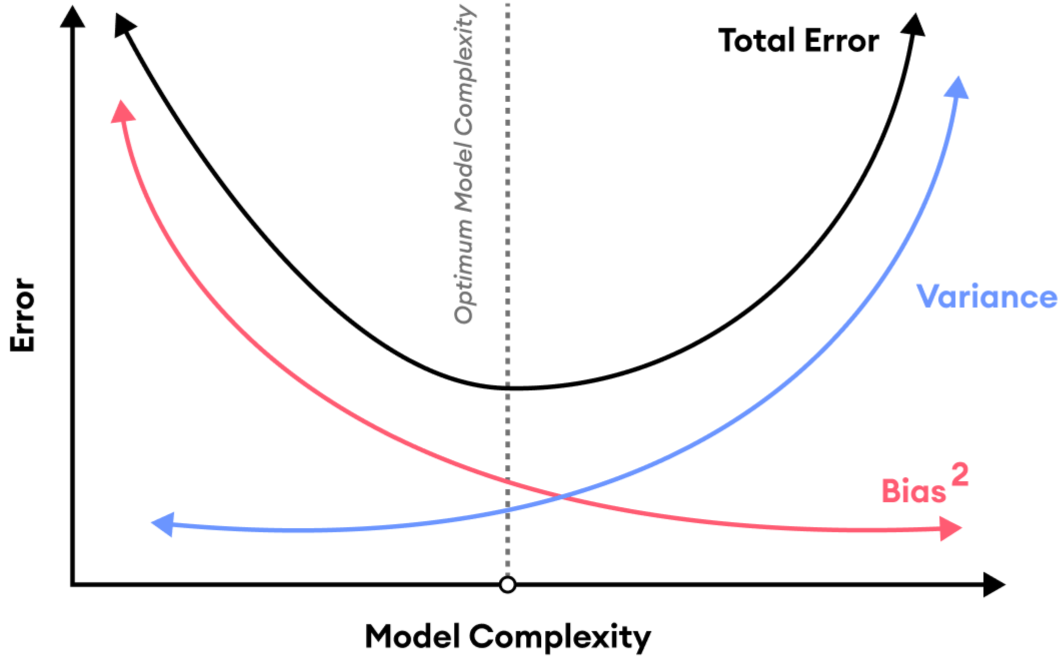
high biais
underfitting

just right

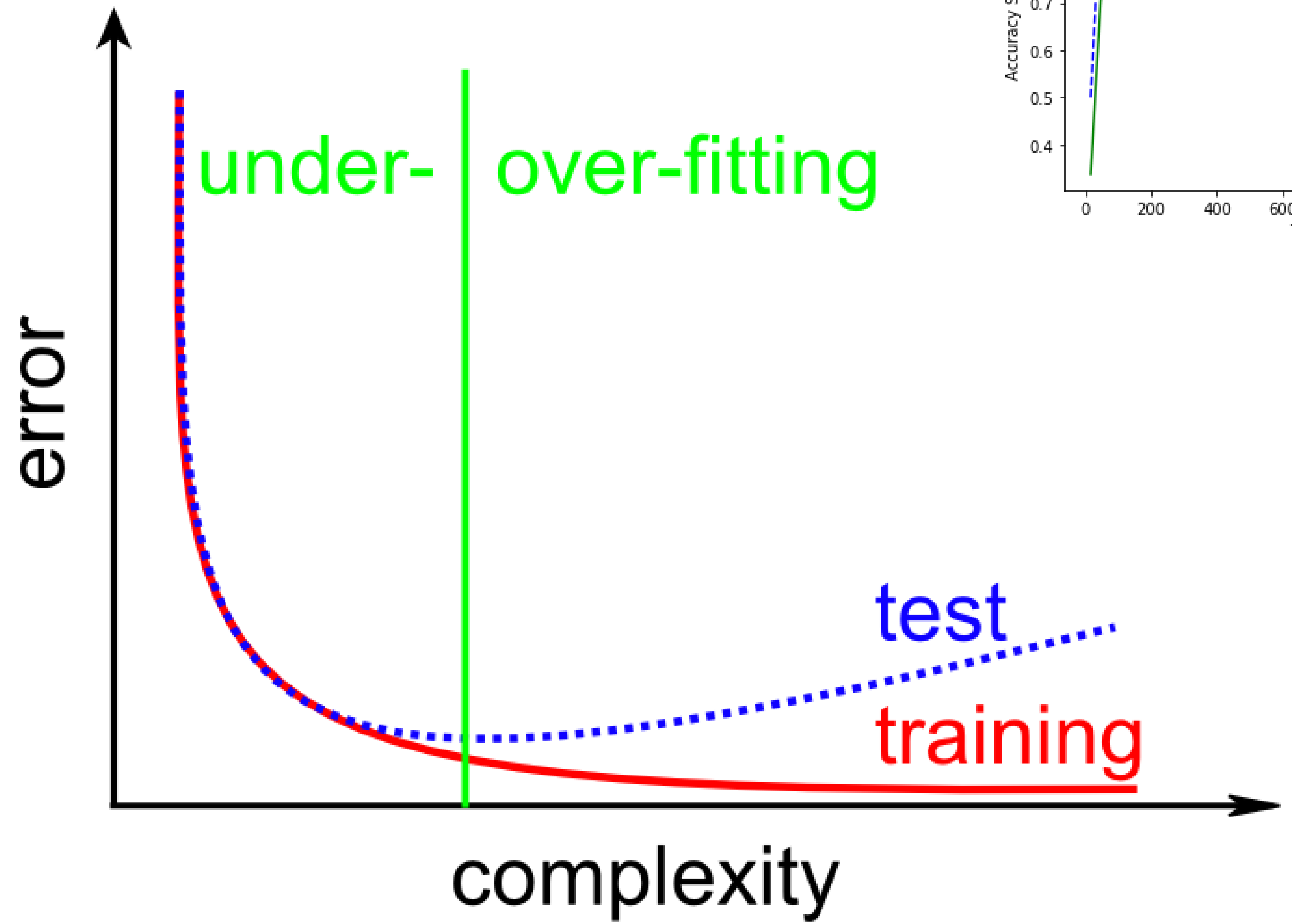
low biais
overfitting

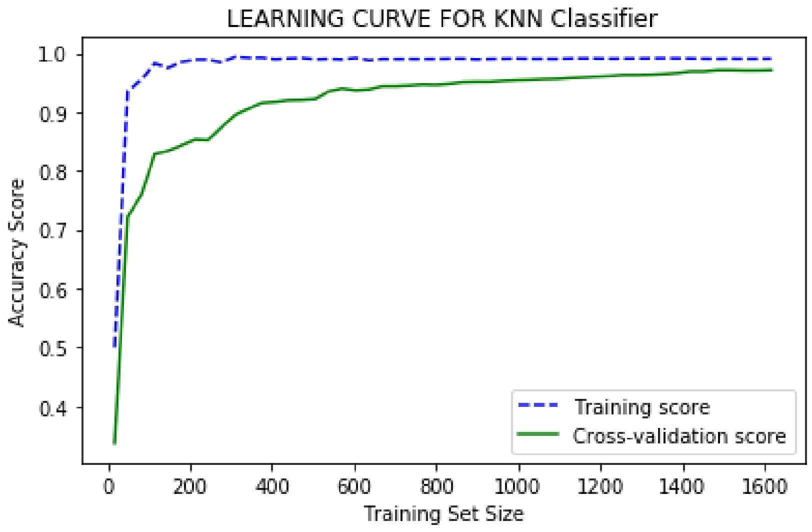
Bias-variance tradeoff



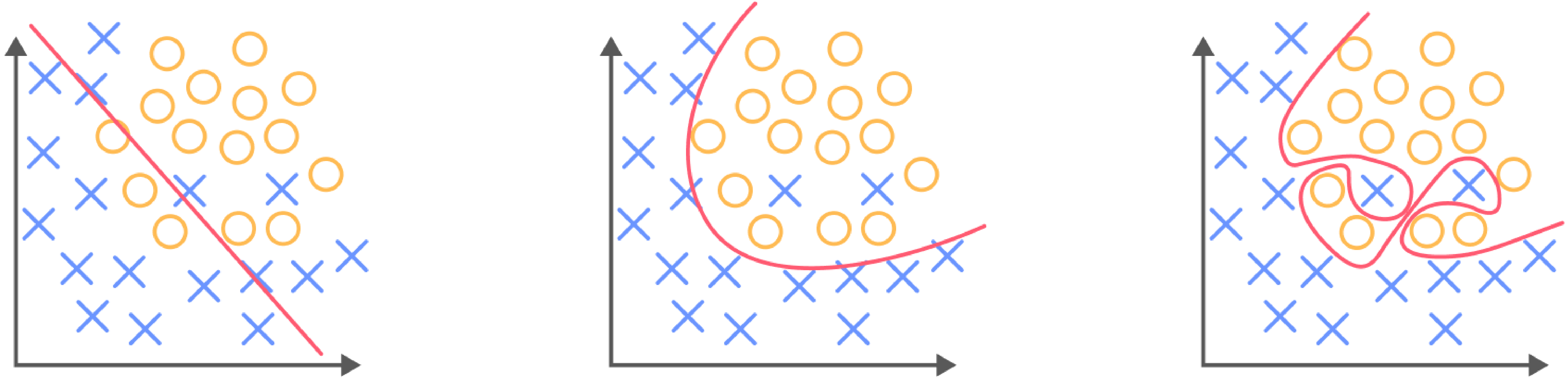


Overfitting





Overfitting in classification

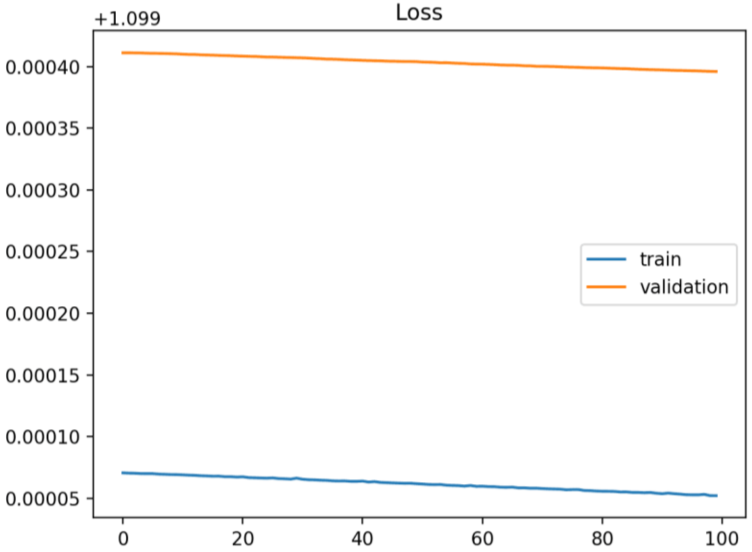


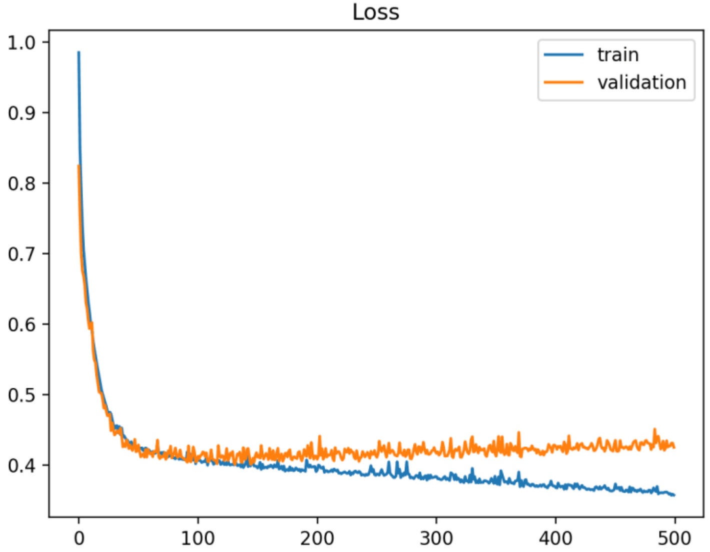
Underfitting

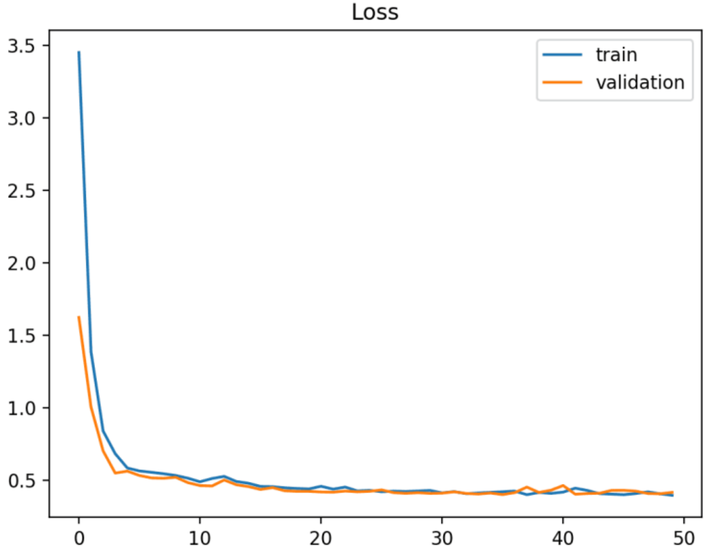
Appropriate fitting

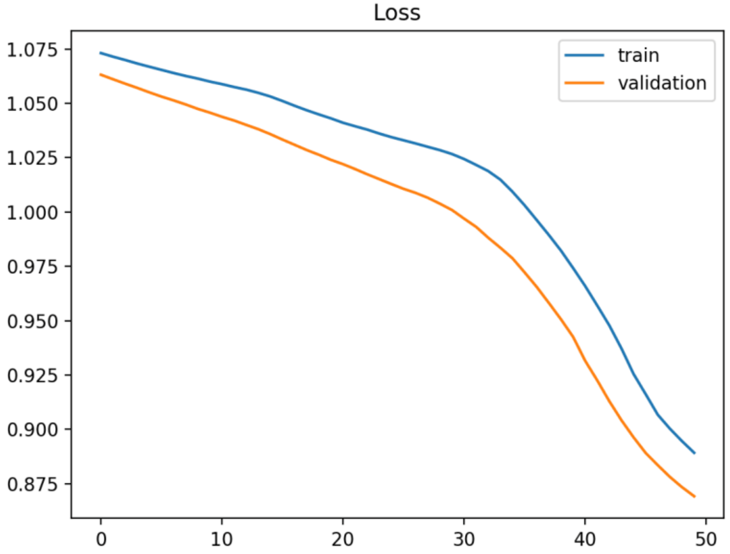
Overfitting

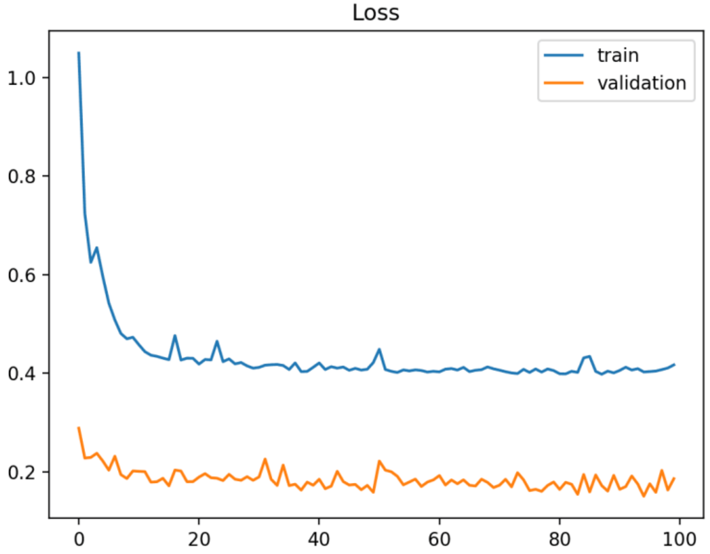
Learning curve: performance VS time

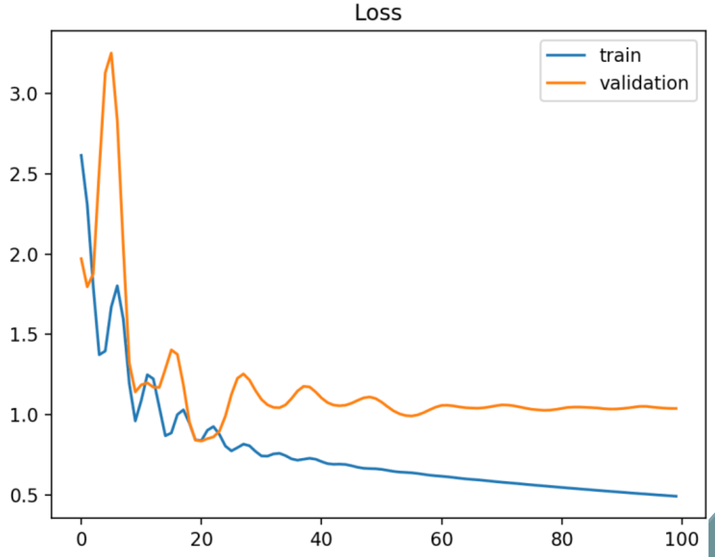








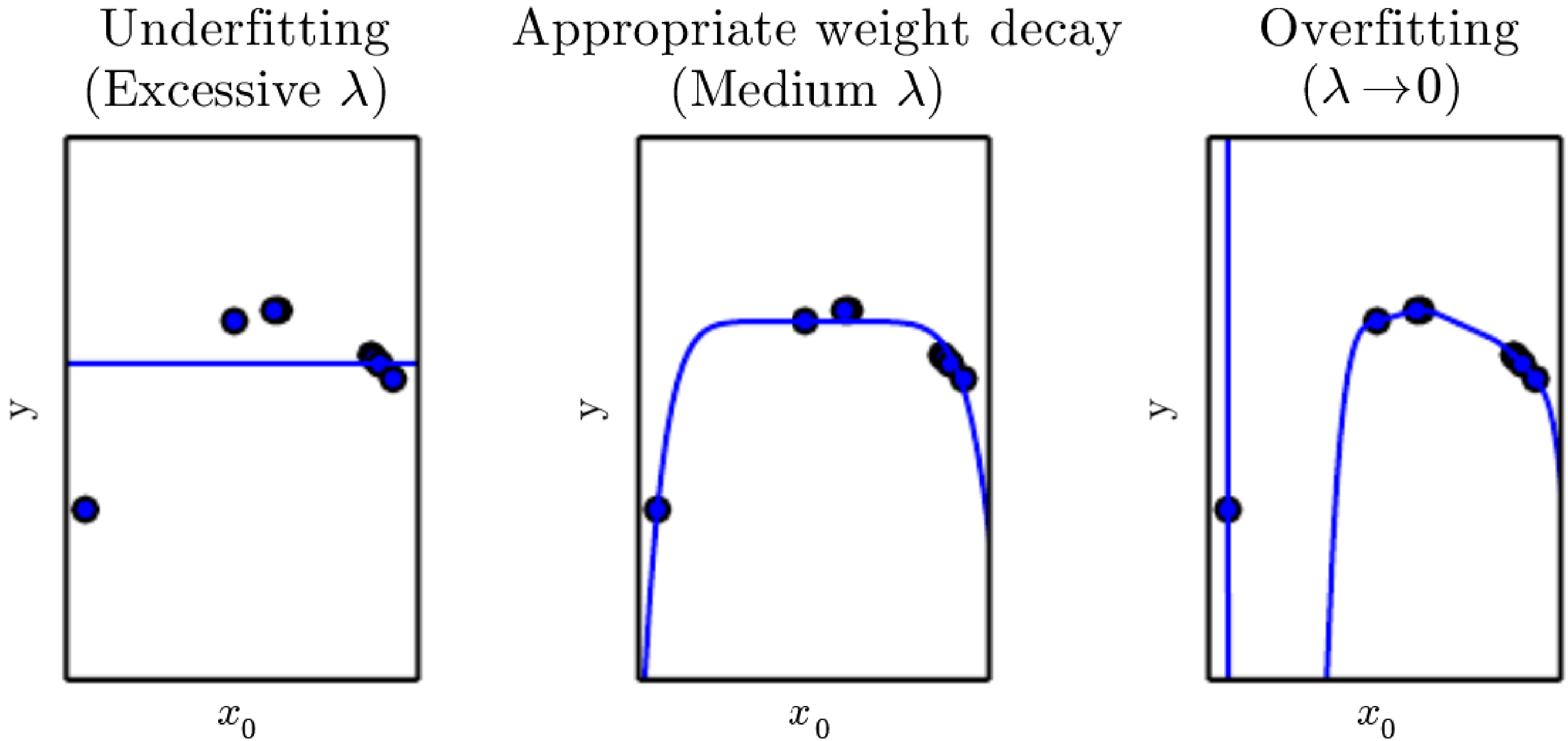




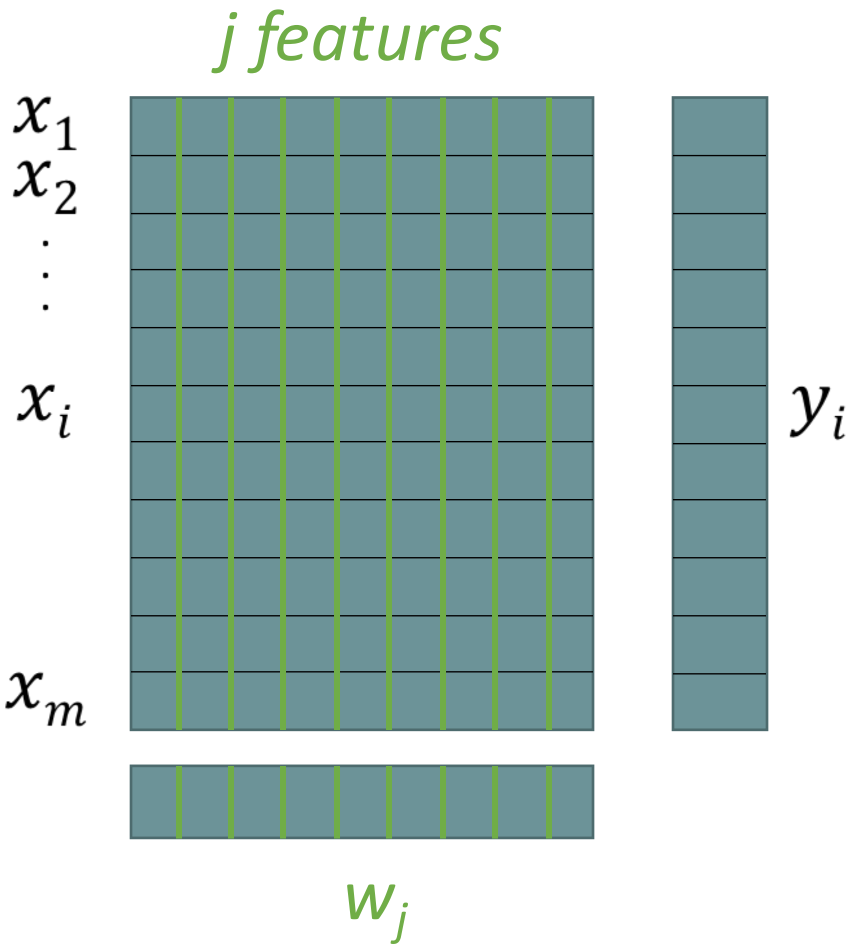
Regularization by penalties

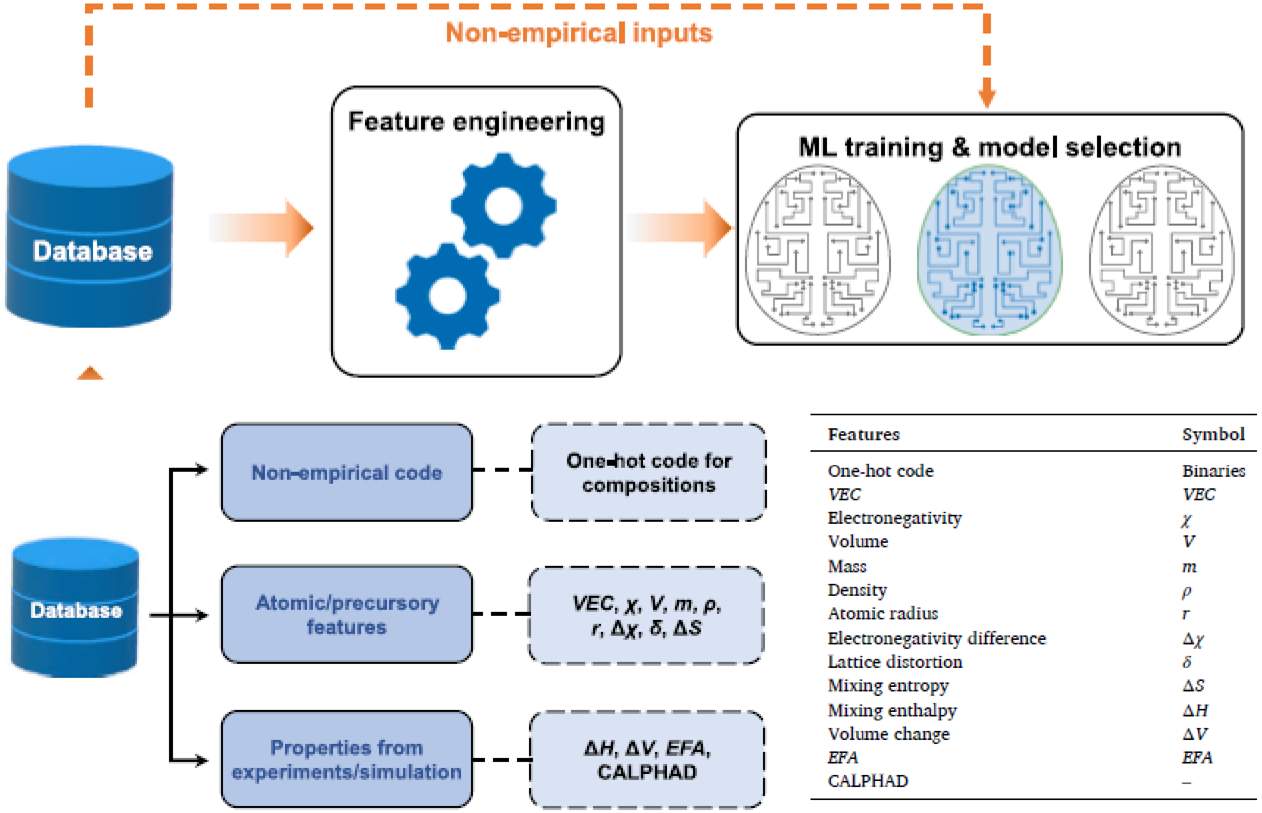
We can introduce a weight decay to degrade the learning.

$$\mathrm{error}(w) = \mathrm{MSE}_{train} + \lambda w^{\mathrm{T}} w$$

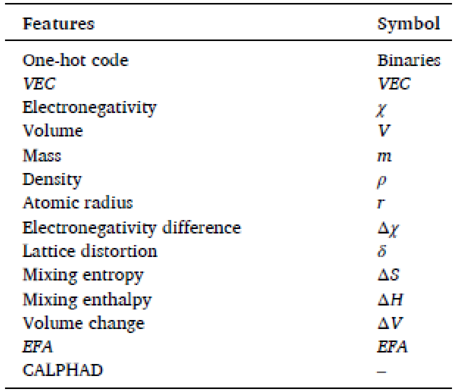


Choice of descriptors / features





Non-empirical inputs
Database
Feature engineering
ML training & model selection
Database
Non-empirical code
One-hot code for compositions
Atomic/precursory features
VEC, χ, V, m, ρ, r, Δχ, δ, ΔS
Properties from experiments/simulation
ΔH, ΔV, EFA, CALPHAD

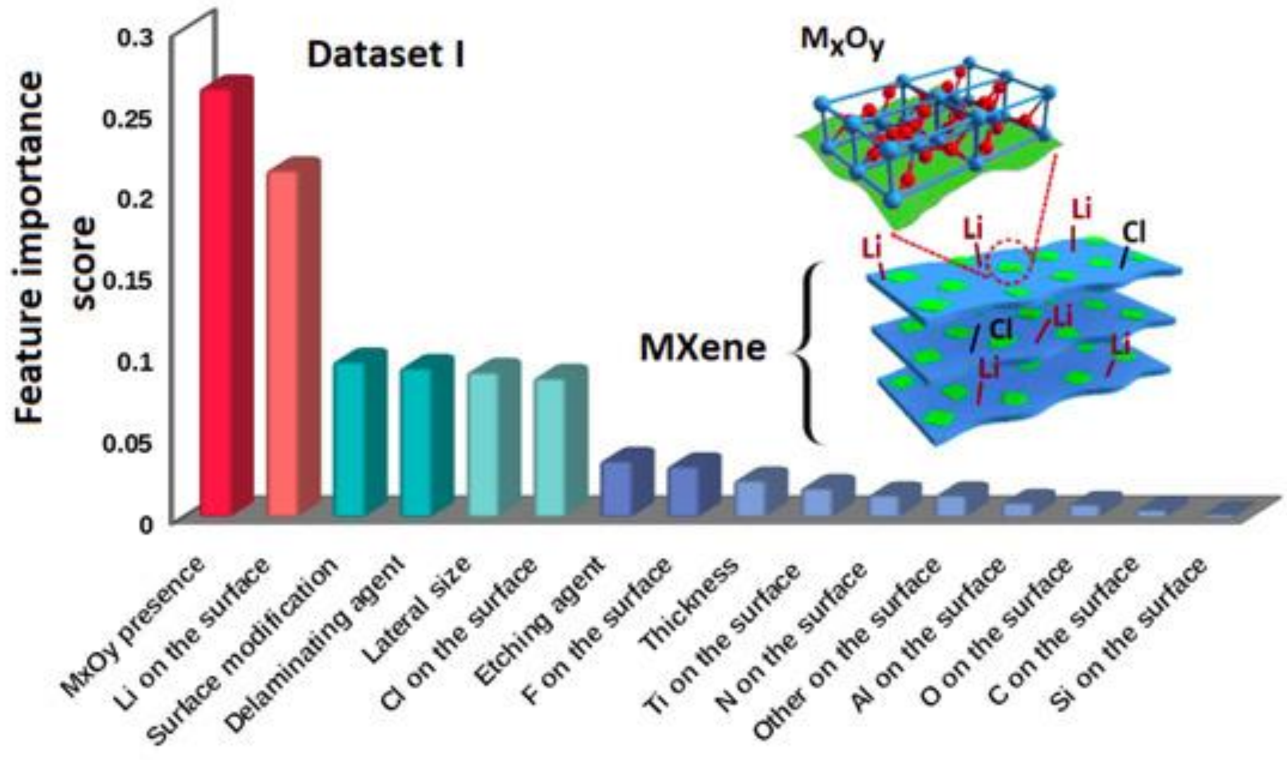


Features	Symbol
One-hot code	Binaries
VEC	*VEC*
Electronegativity	χ
Volume	V
Mass	m
Density	ρ
Atomic radius	r
Electronegativity difference	$\Delta\chi$
Lattice distortion	δ
Mixing entropy	ΔS
Mixing enthalpy	ΔH
Volume change	ΔV
EFA	*EFA*
CALPHAD	–

Zhang *et al*. Curr. Opin. Solid State Mater. Sci. (2020)

Rational design of high-entropy ceramics based on machine learning – A critical review

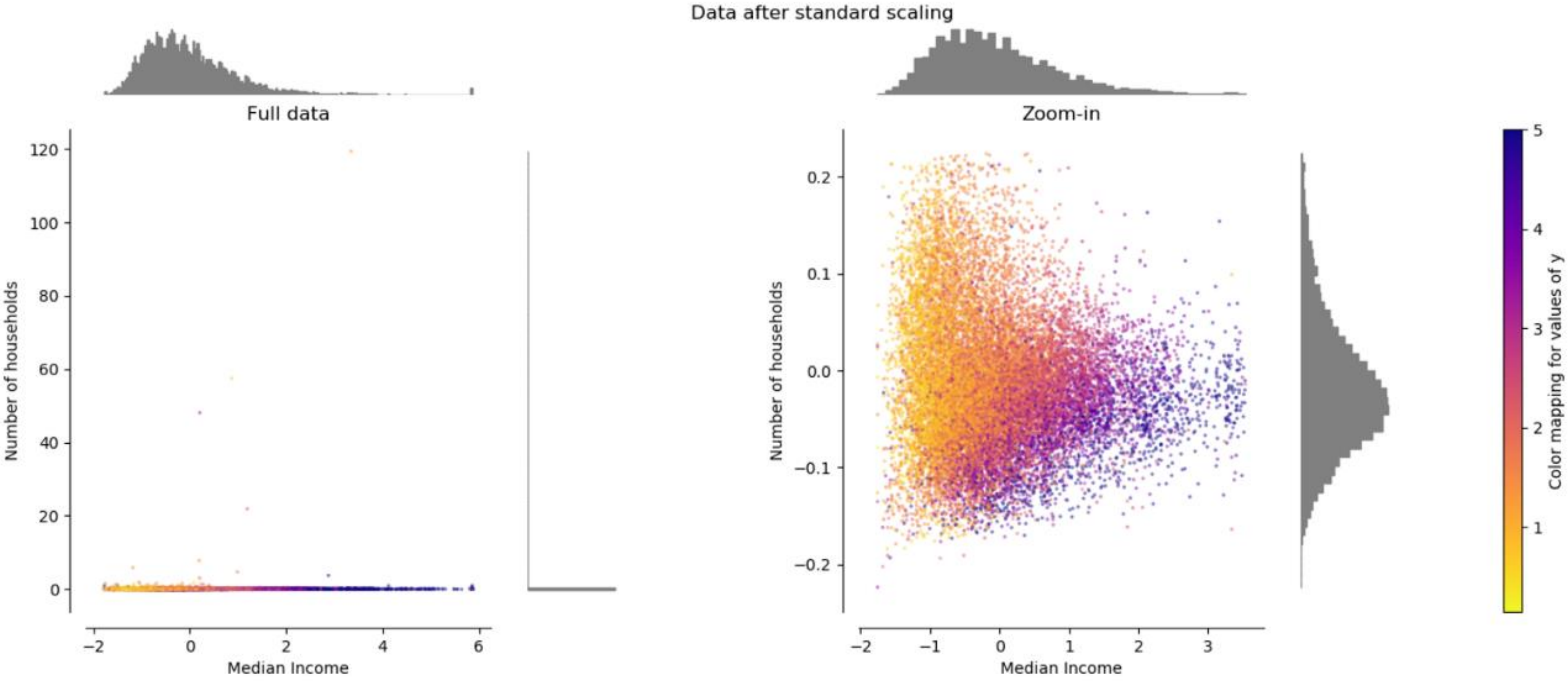
Feature importance



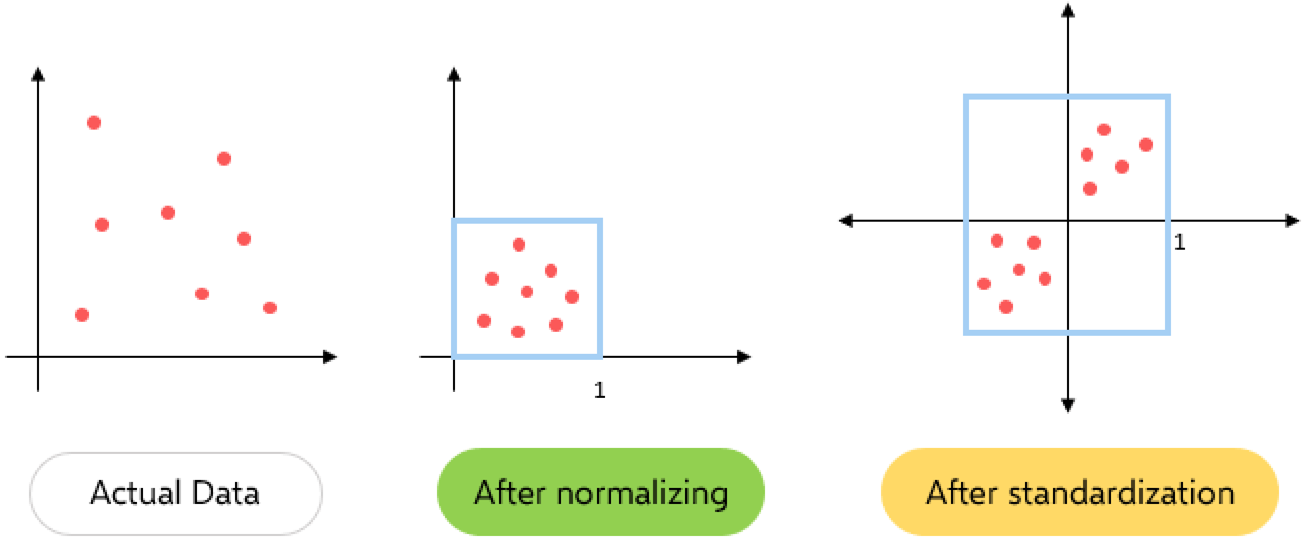
Marchwiany *et al*. Materials (2020)

Surface-Related Features Responsible for Cytotoxic Behavior of MXenes Layered Materials Predicted with Machine Learning Approach

Standardization



Standardization



Actual Data

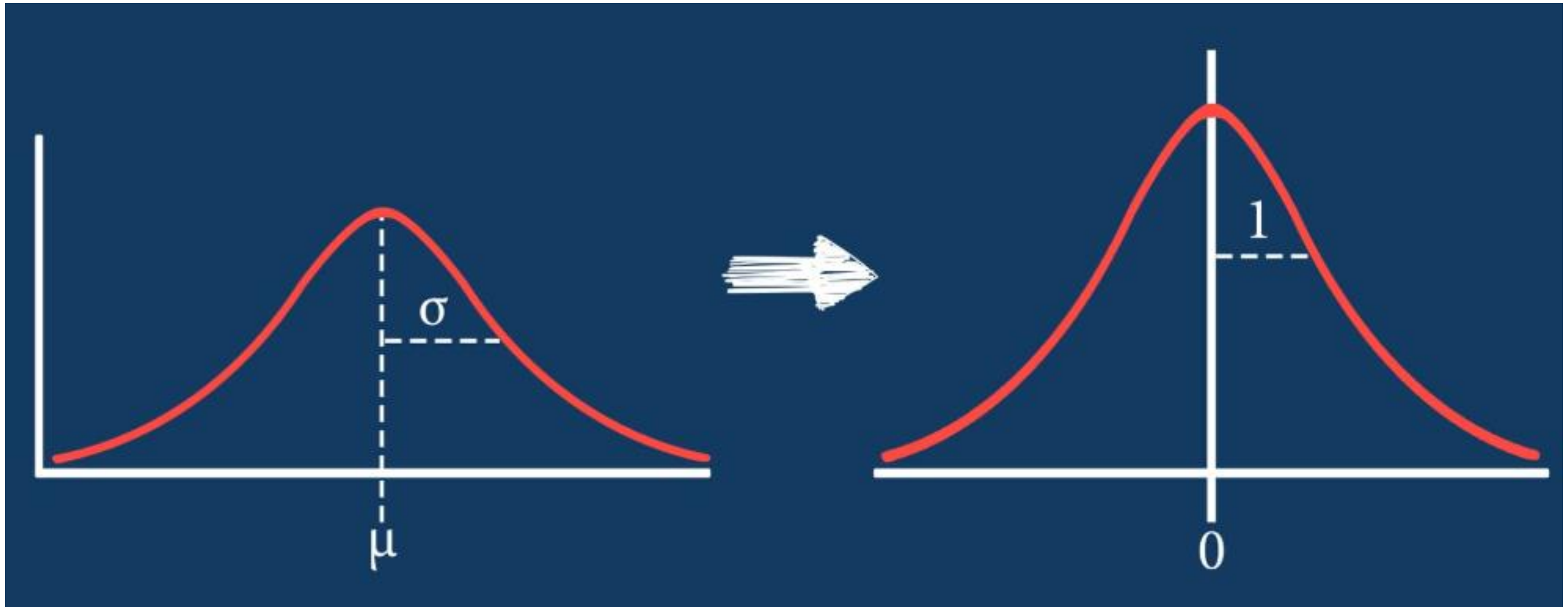
After normalizing

After standardization

Moyenne: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

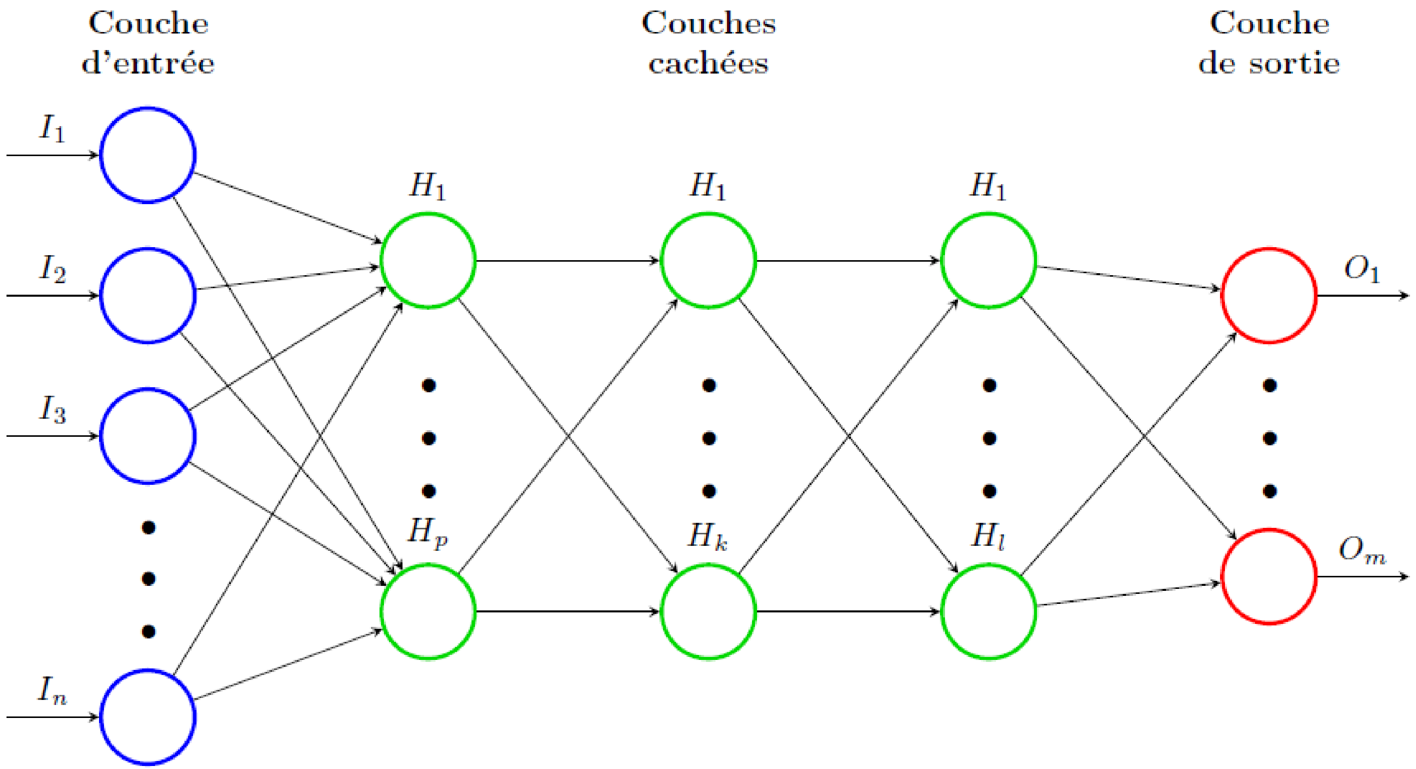
Variance: $V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Ecart type: $\sigma = \sqrt{V}$

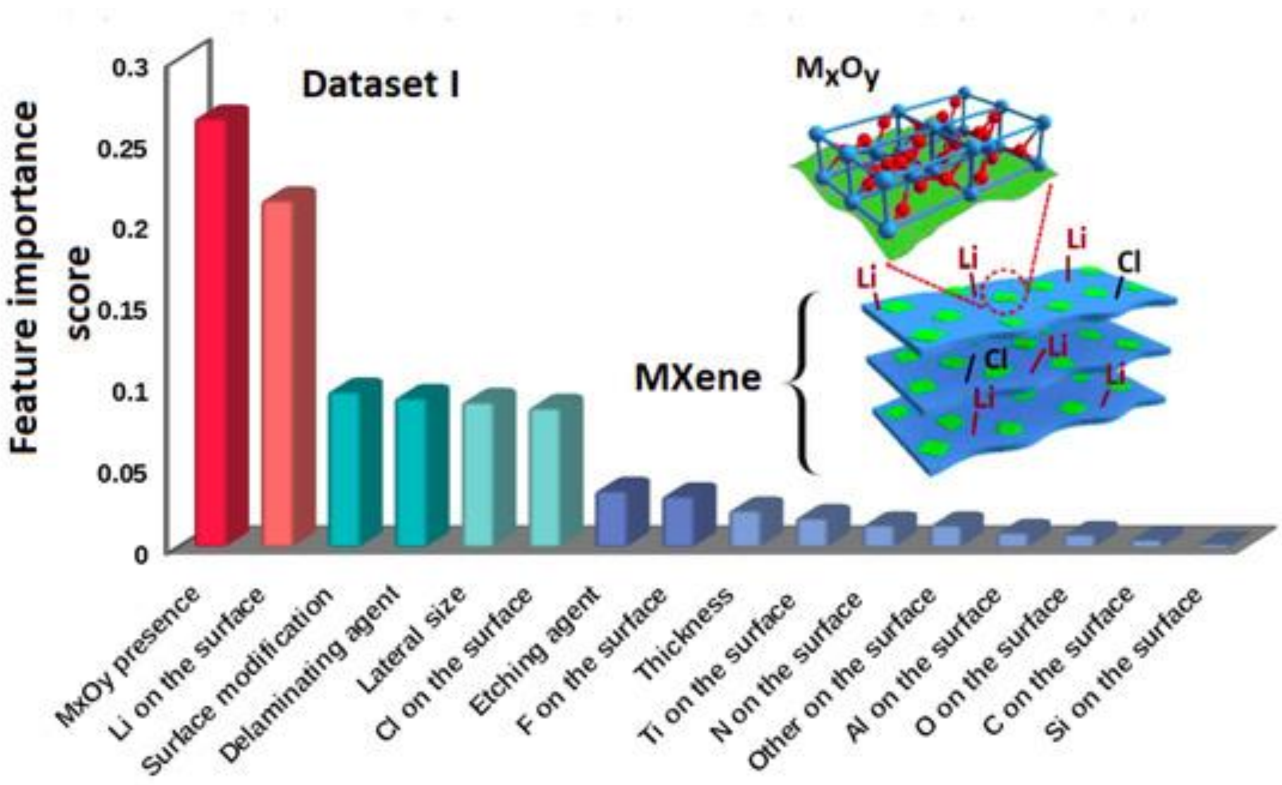


$$z_i = \frac{x_i - \mu}{\sigma}$$

Why is standardisation important?



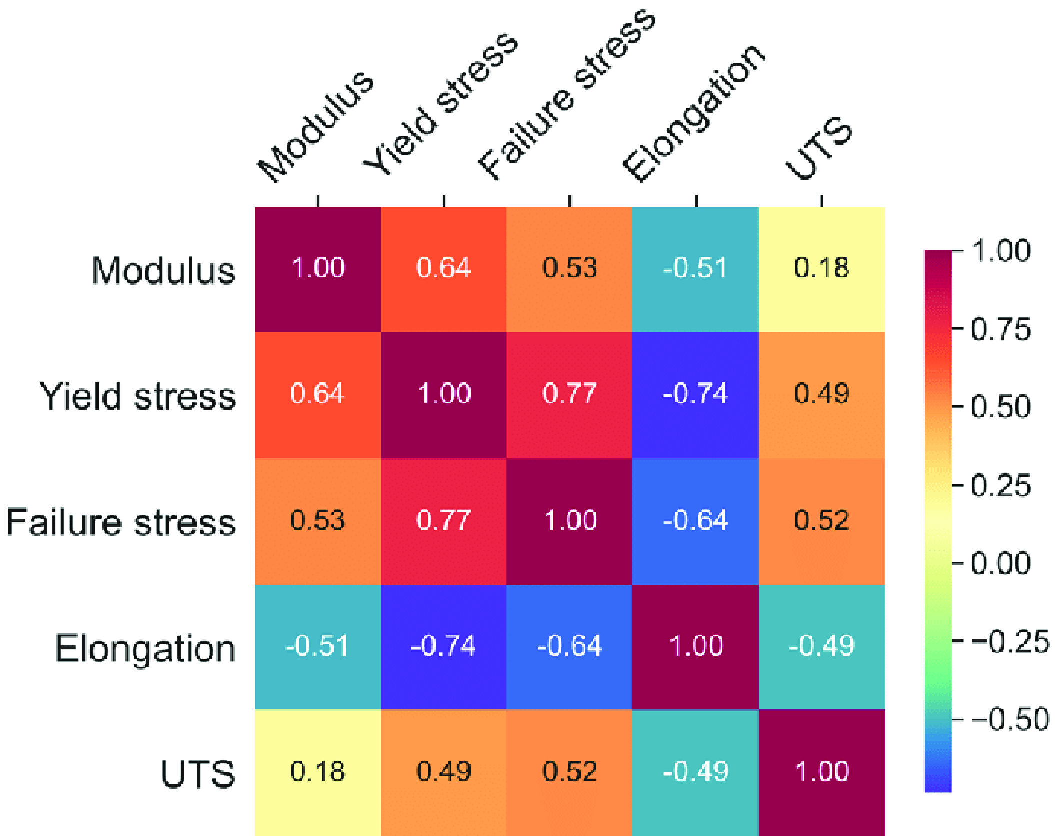
Feature importance



Marchwiany *et al*. Materials (2020)

Surface-Related Features Responsible for Cytotoxic Behavior of MXenes Layered Materials Predicted with Machine Learning Approach

Pair correlation matrix



	Modulus	Yield stress	Failure stress	Elongation	UTS
Modulus	1.00	0.64	0.53	-0.51	0.18
Yield stress	0.64	1.00	0.77	-0.74	0.49
Failure stress	0.53	0.77	1.00	-0.64	0.52
Elongation	-0.51	-0.74	-0.64	1.00	-0.49
UTS	0.18	0.49	0.52	-0.49	1.00

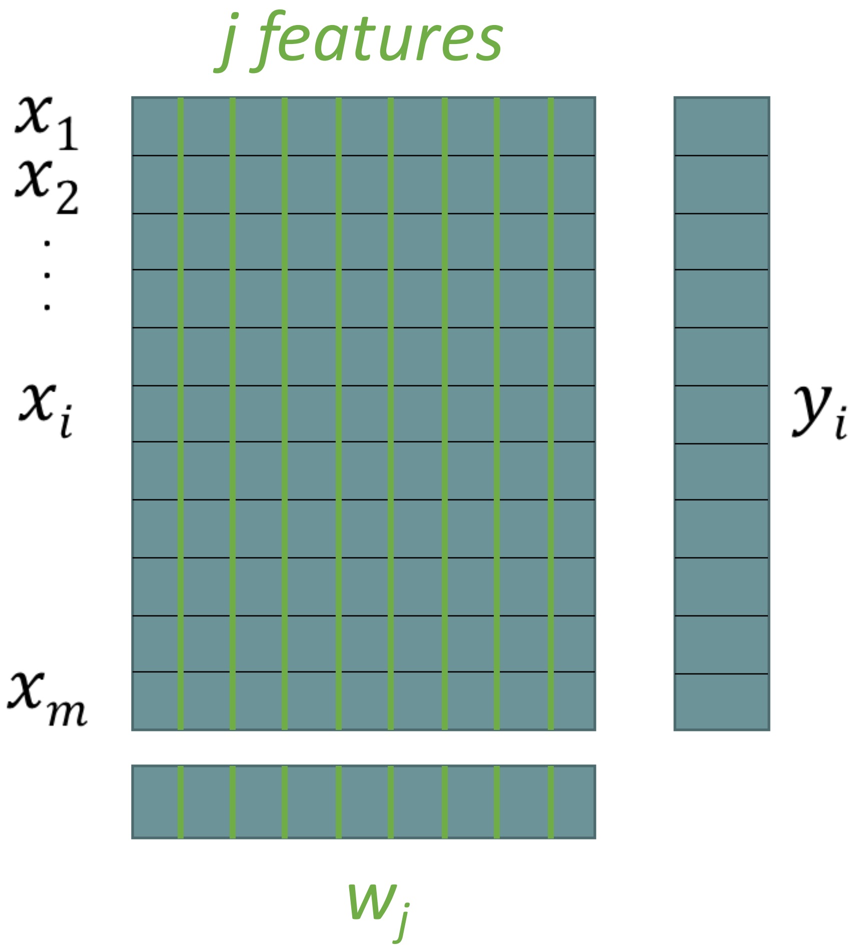
Esperance: $E[X] = \sum_{i=1}^{\infty} x_i \, p_i$

Covariance: $\mathrm{Cov}(X, Y) =$

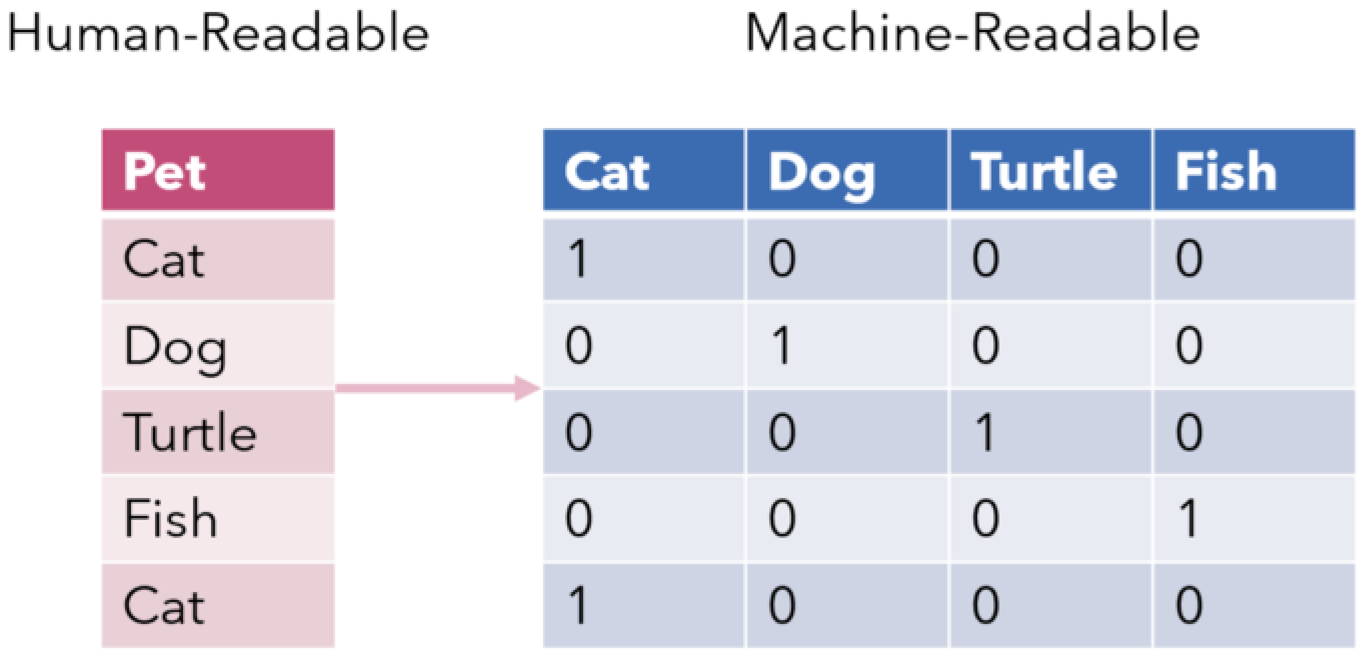
$$\sum_i \sum_j x_i y_j \, \mathrm{P}(X = xi \text{ et } Y = yj) - E[X]E[Y]$$

Xie *et al*.   Npj Comp Mater (2021)
Mechanistic data-driven prediction of as-built mechanical properties in metal additive manufacturing

One-Hot Encoding the categorical variables



otherwise known as dummy variables, is a method of converting categorical variables into several binary columns



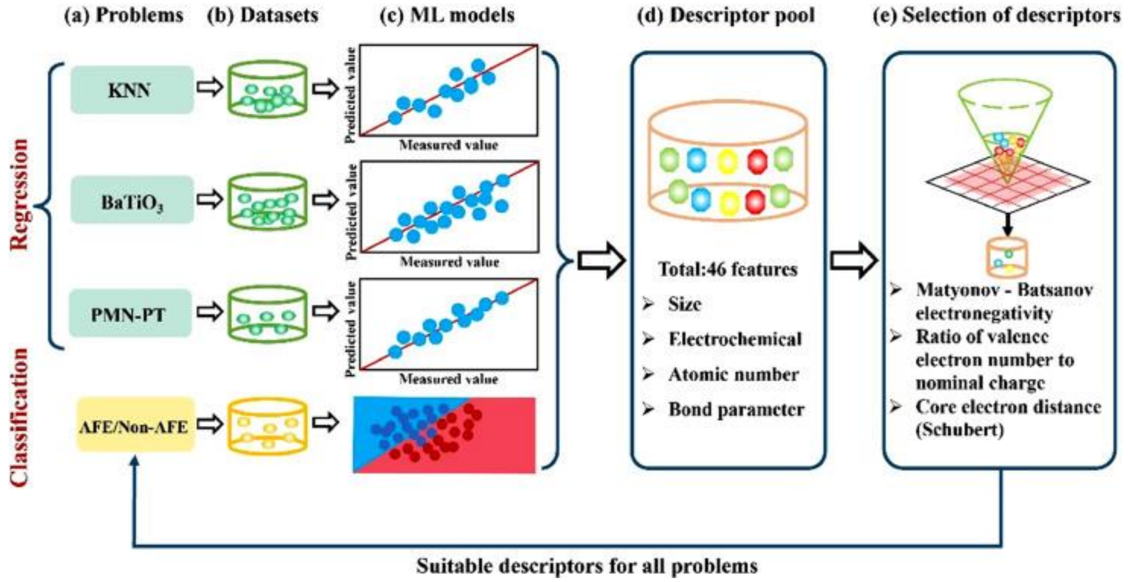
Human-Readable

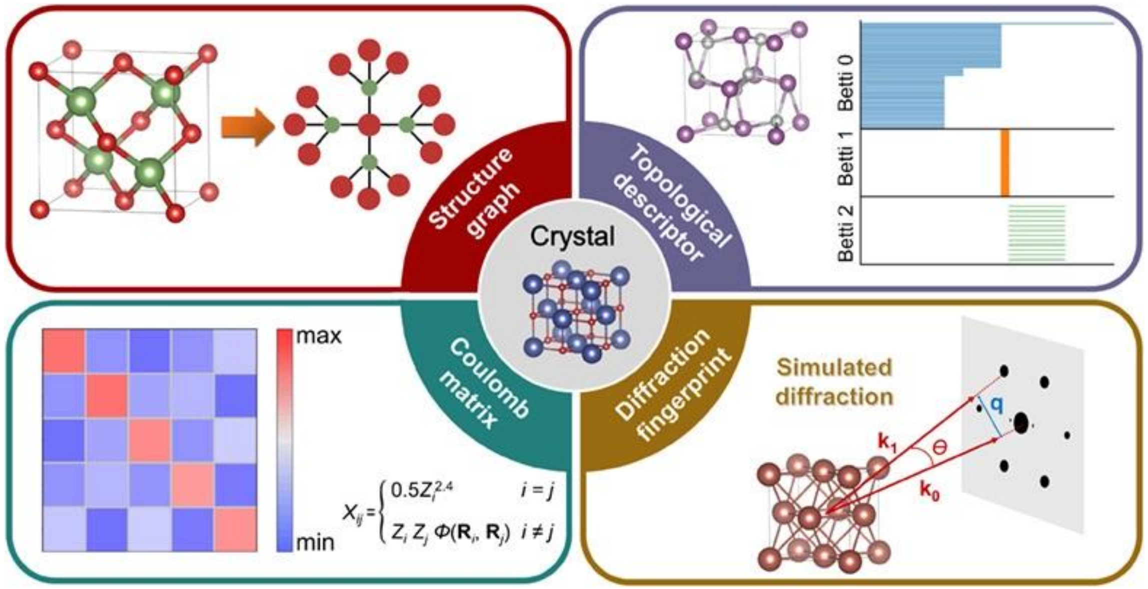
Pet
Cat
Dog
Turtle
Fish
Cat

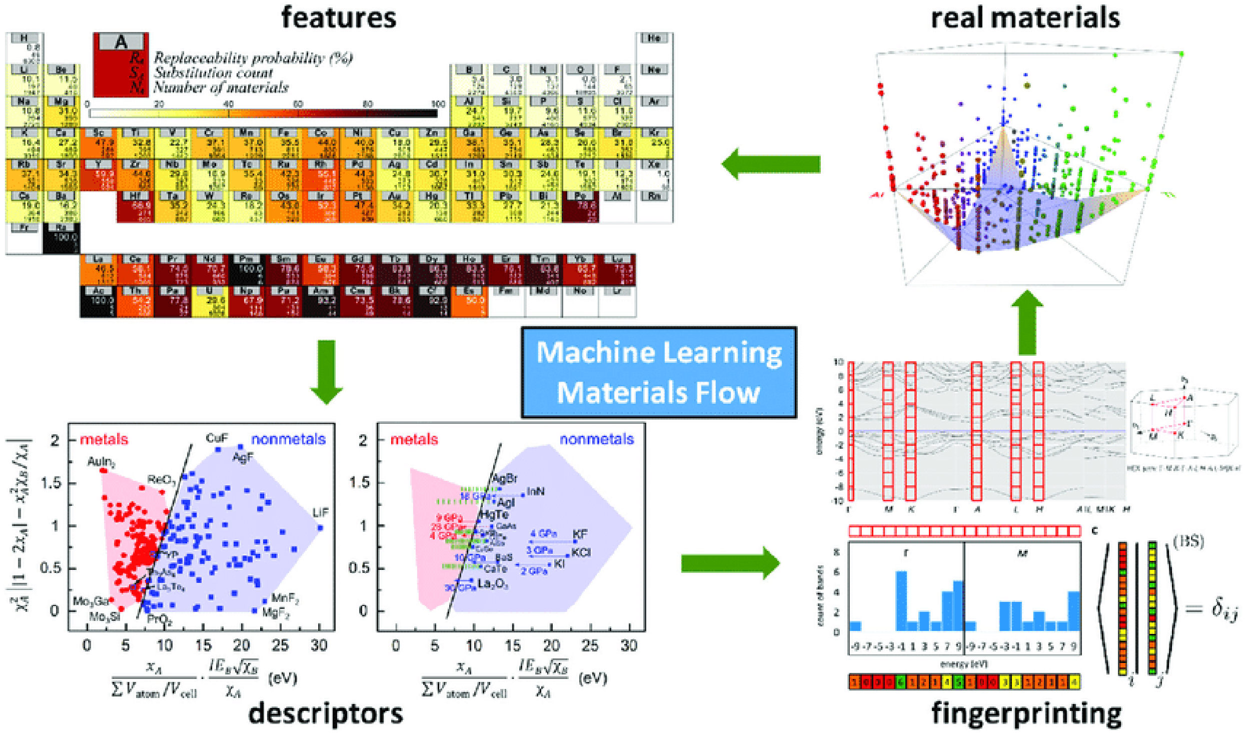
Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

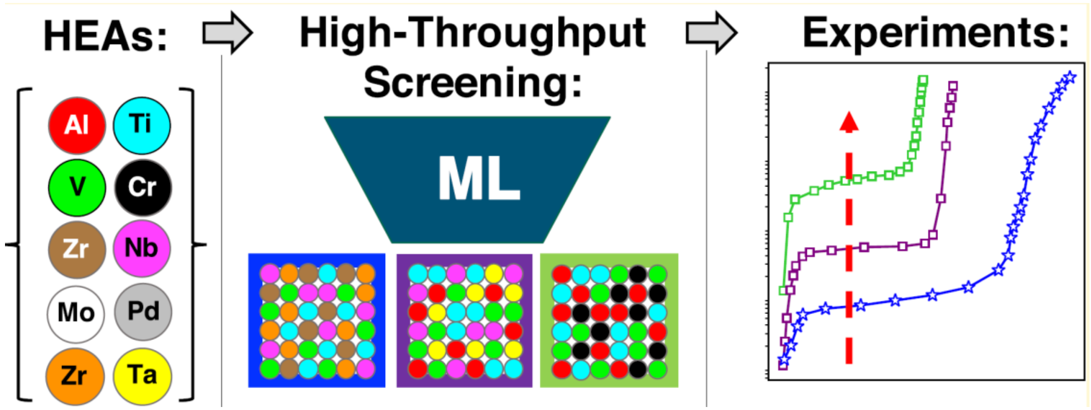
Choice of descriptors

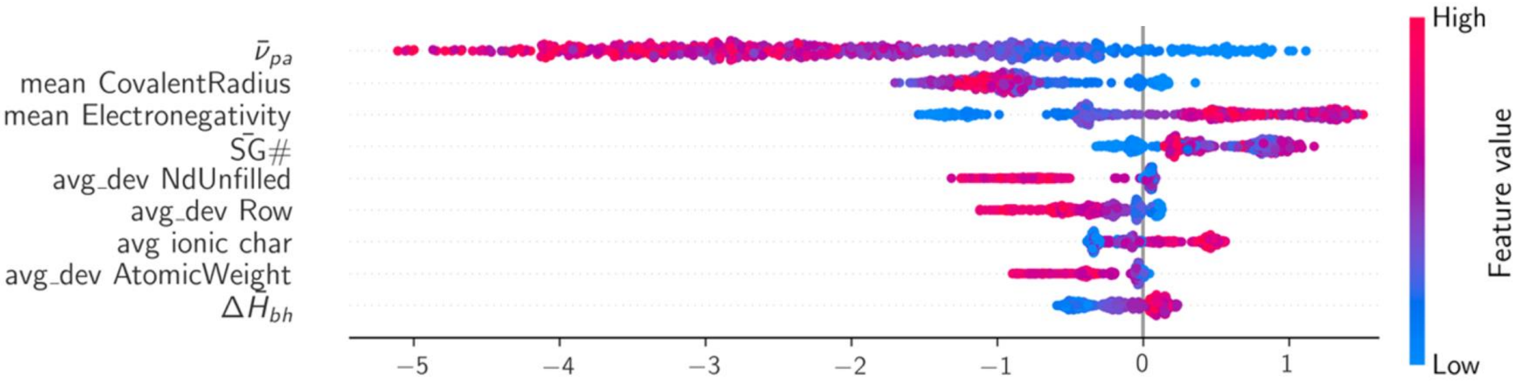






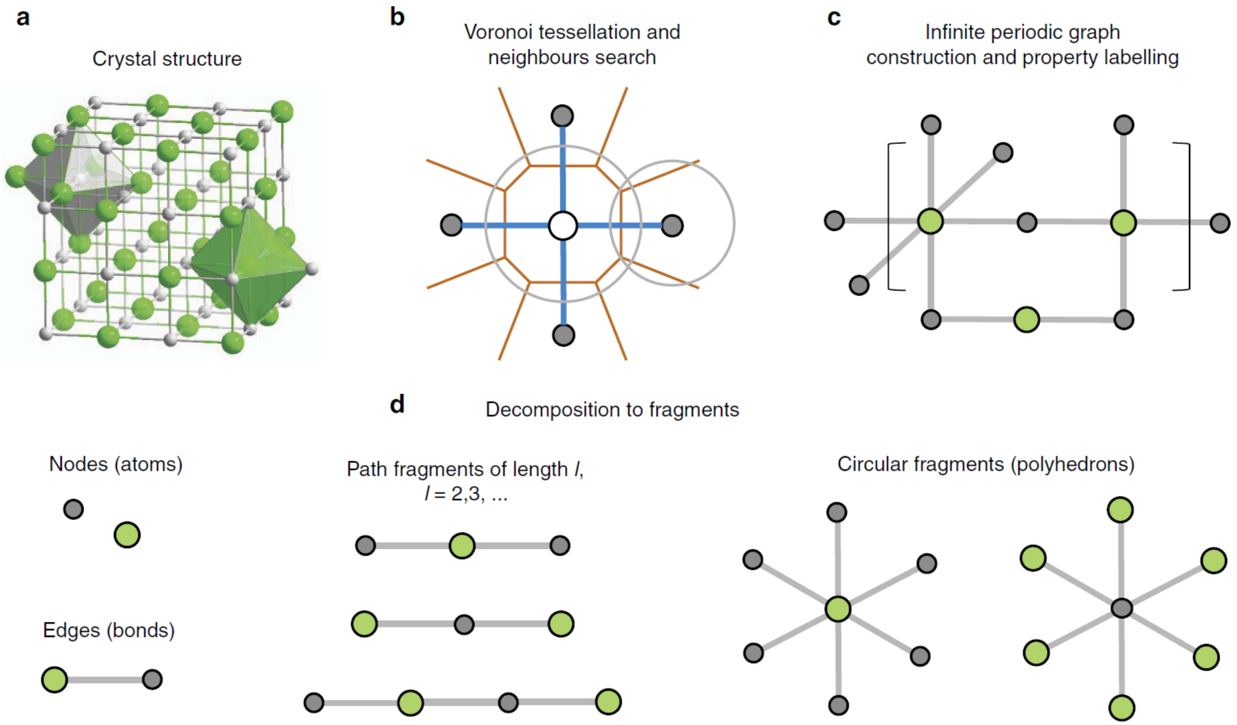
Improved performance with embedded physics



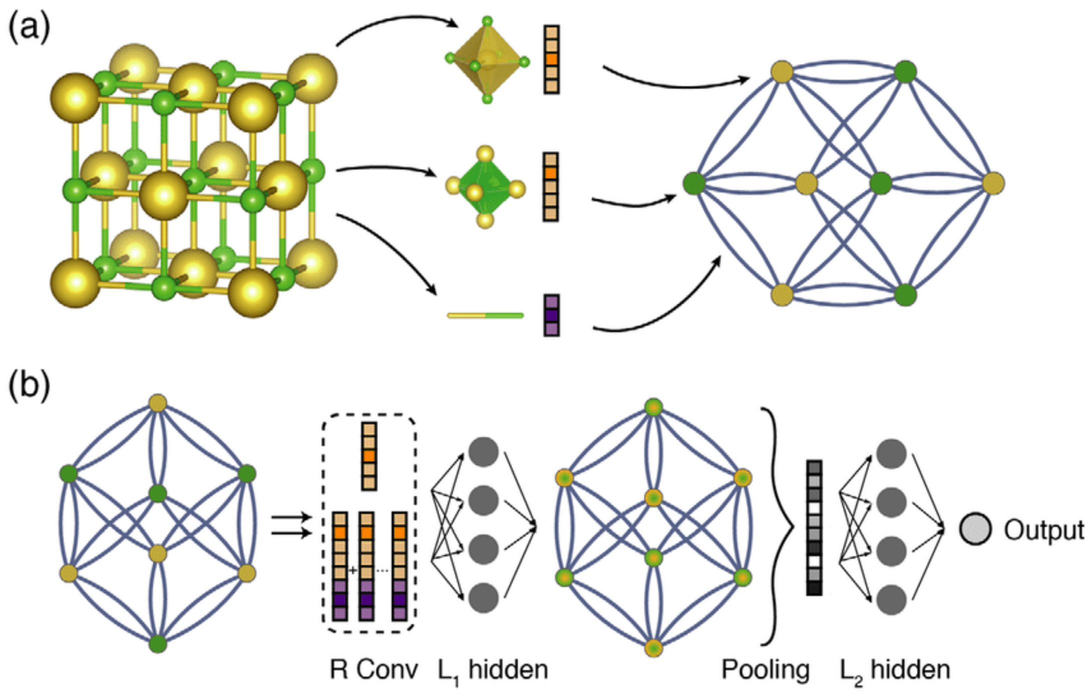


Witman *et al*.
Chem Mater (2021)
Data-Driven Discovery and Synthesis of High Entropy Alloy Hydrides with Targeted Thermodynamic Stability

Graphs description


a
Crystal structure
b
Voronoi tessellation and neighbours search
c
Infinite periodic graph construction and property labelling
d
Decomposition to fragments
Nodes (atoms)
Edges (bonds)
Path fragments of length *l*, *l* = 2,3, ...
Circular fragments (polyhedrons)

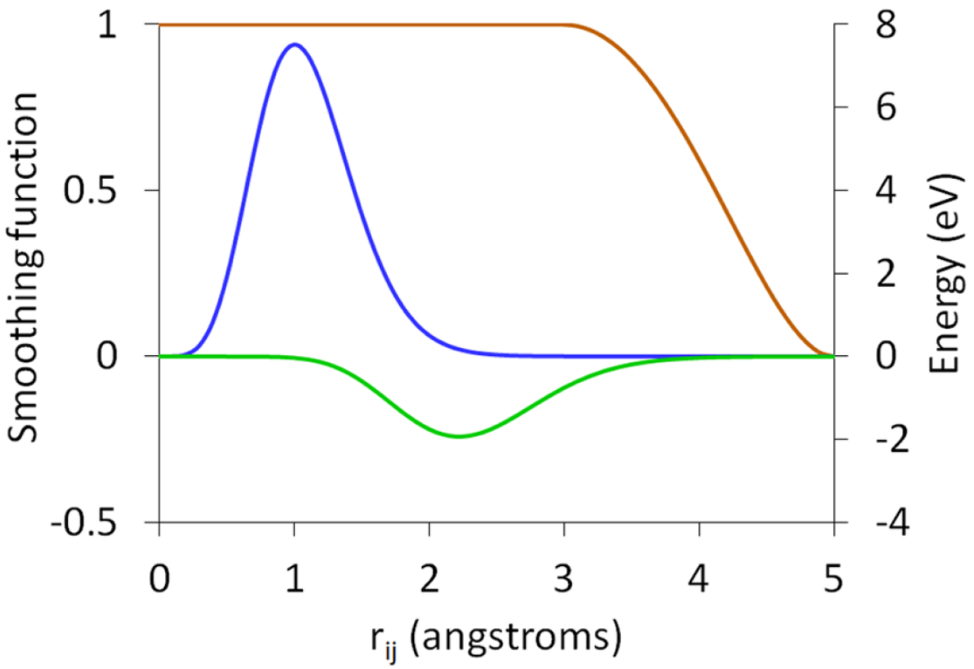
Isayev *et al*.
Nature Com (2017)
Universal fragment descriptors for predicting properties of inorganic crystals



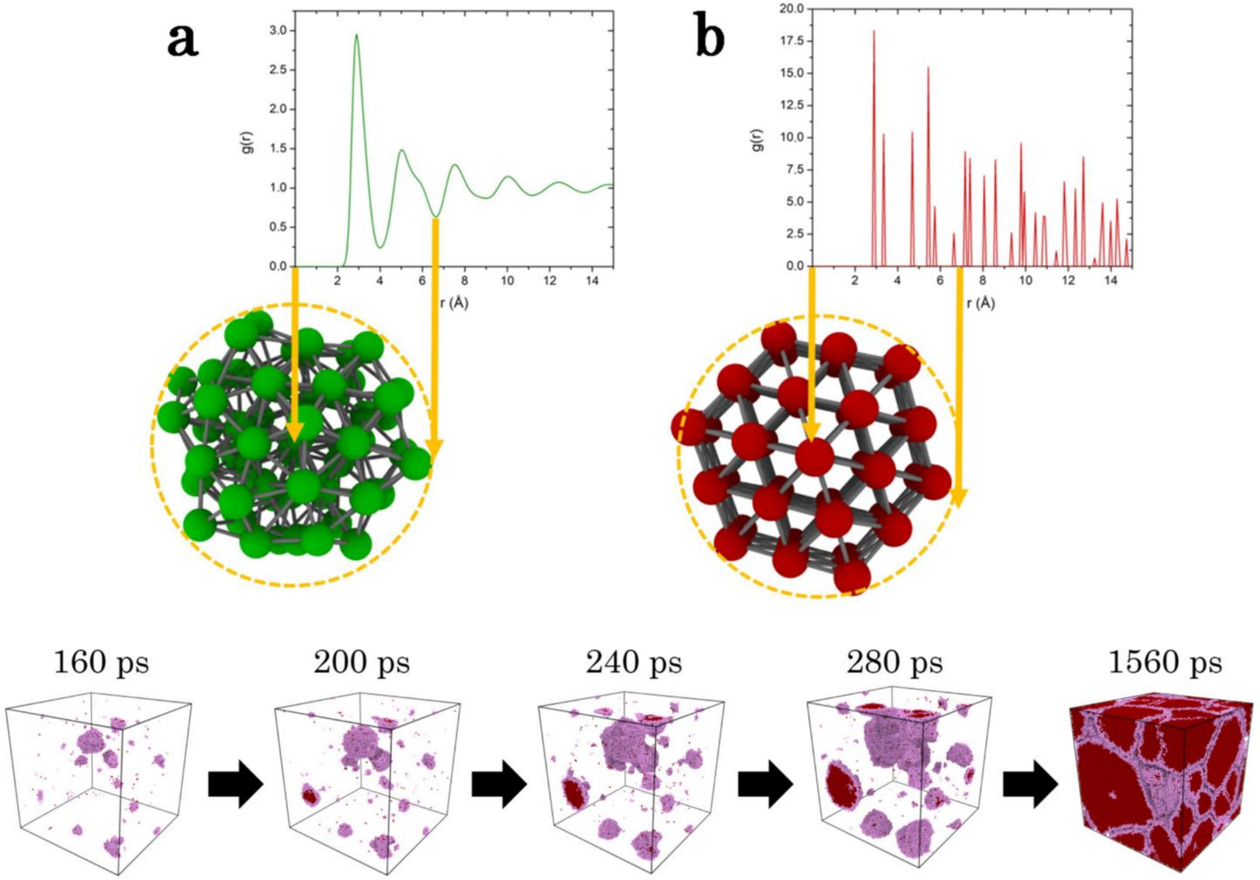
Xie *et al*.
Phys Rev Lett (2018)
3-D Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties

Interatomic potential models

$$\sum 7.51 r^{3.98-3.93r} f(r) + \left(28.01 + \sum -0.03 r^{11.73-2.93r} f(r)\right)\left(\sum f(r)\right)^{-1}$$



Mueller *et al*.
J Chem Phys (2020)
3-D Machine learning for interatomic potential models



Becker *et al*.
Sci report (2022)
Unsupervised topological learning approach of crystal nucleation