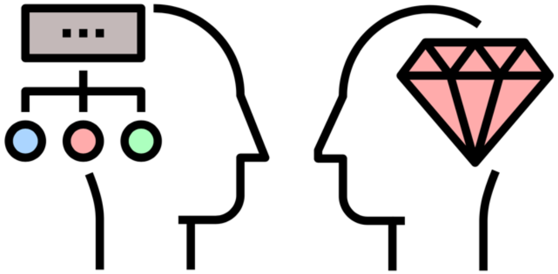
IMPERIAL
Department of Materials

Machine Learning for Materials

2. Machine Learning Basics

Aron Walsh

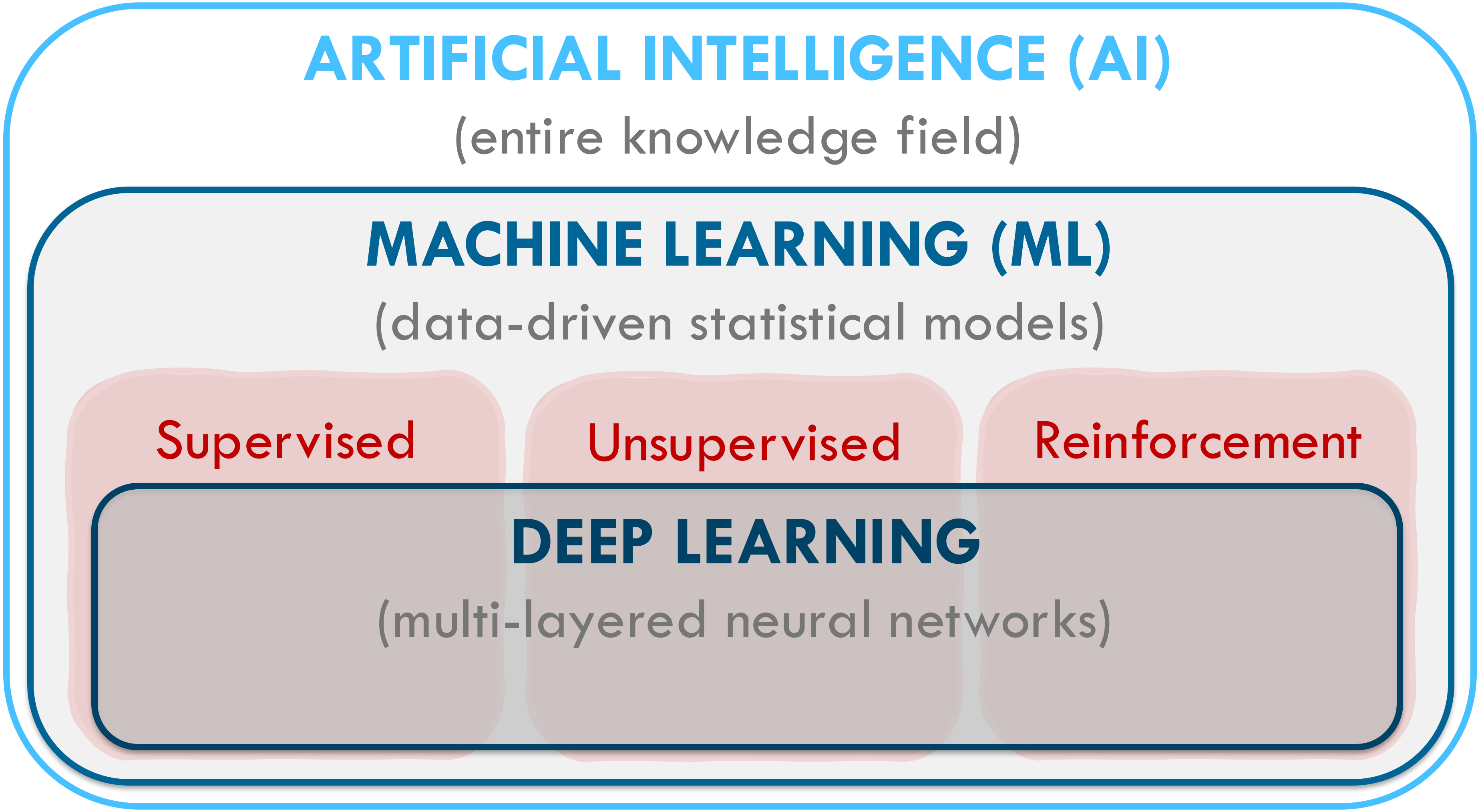
Department of Materials
Centre for Processable Electronics



Module Contents

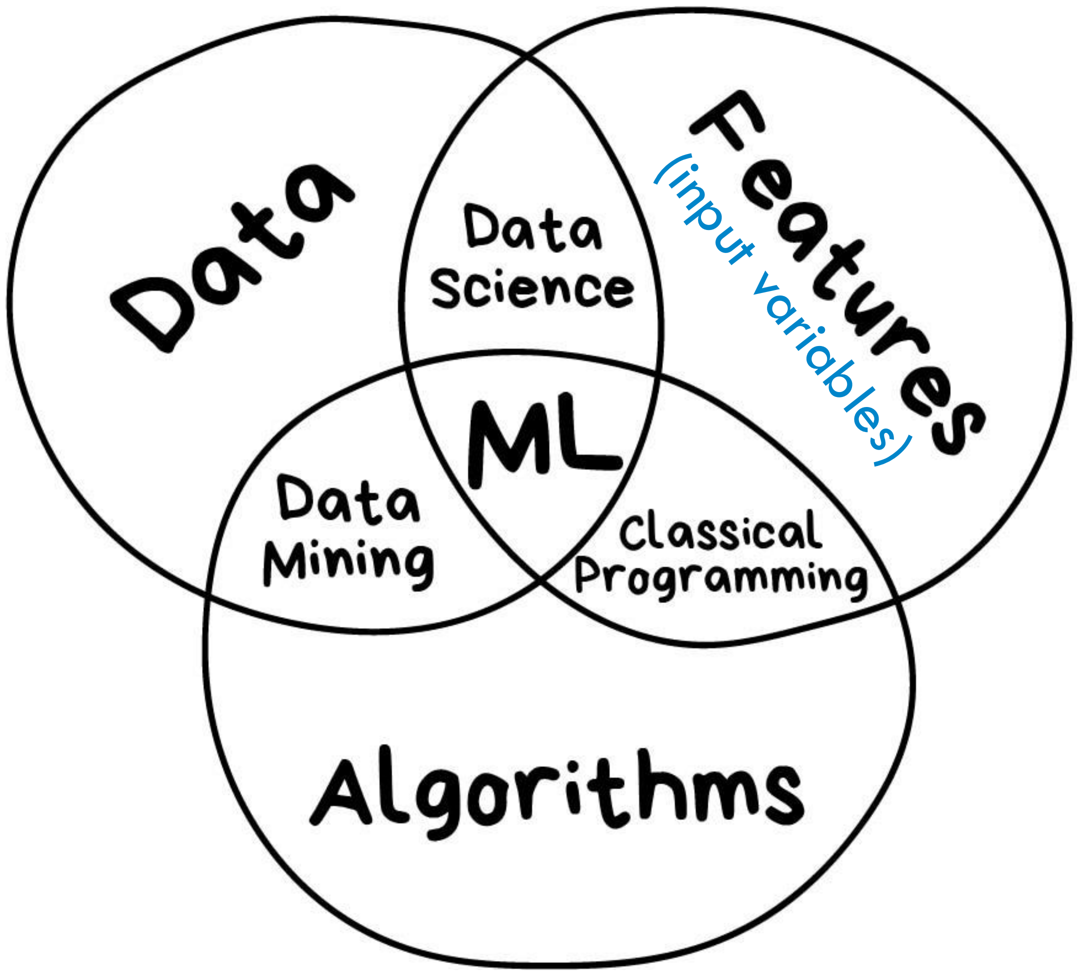
Artificial Intelligence

Computational techniques that mimic human intelligence



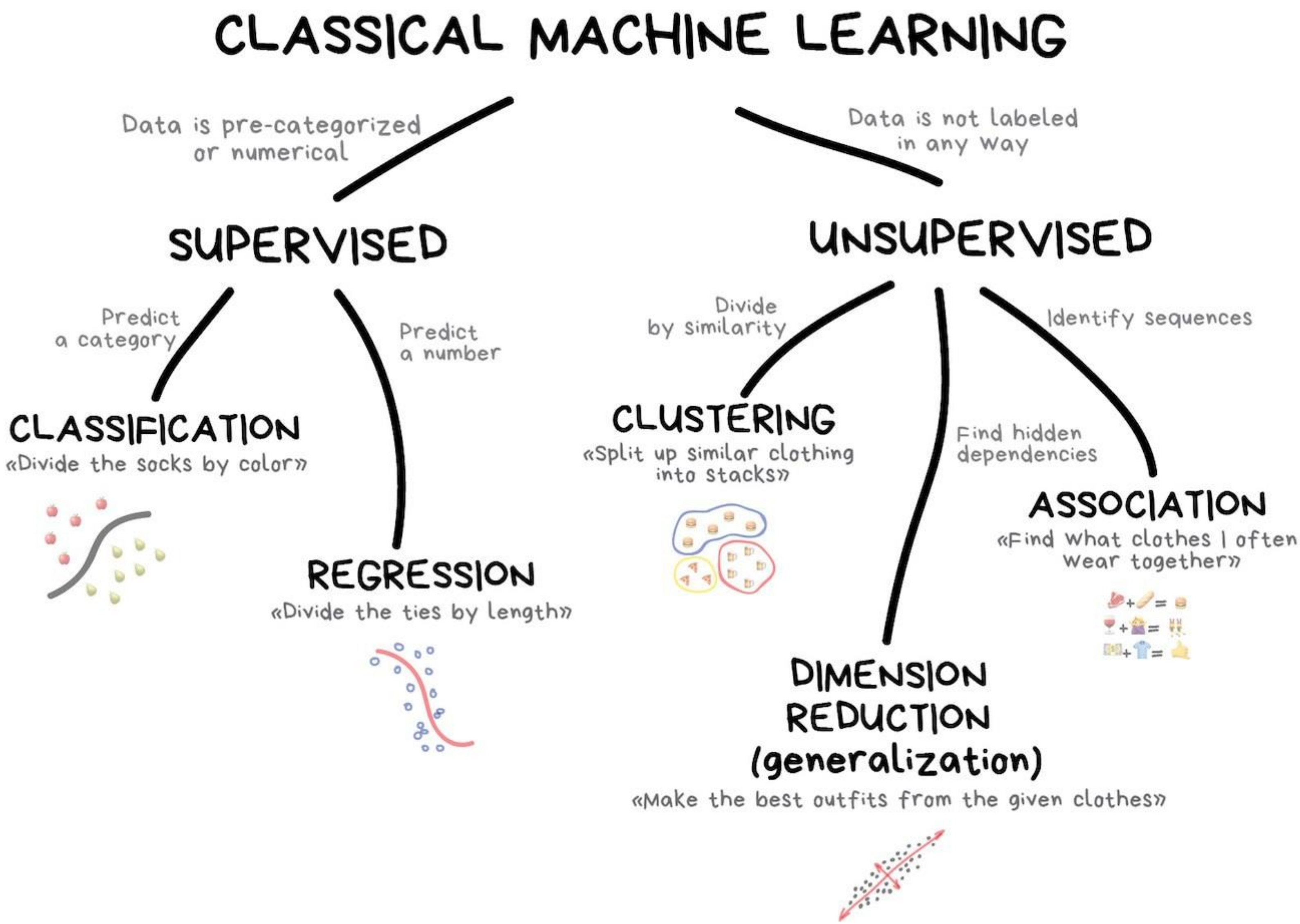
Focus on Machine Learning (ML)

Statistical techniques that improve with experience



https://vas3k.com/blog/machine_learning

Focus on Machine Learning (ML)



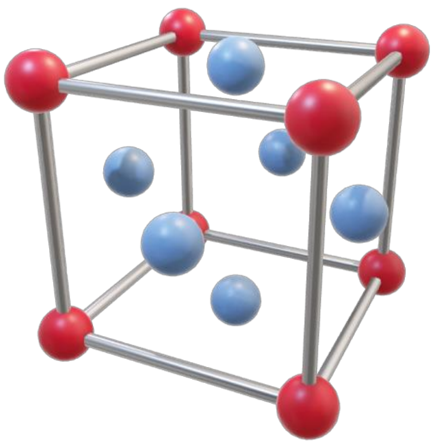
https://vas3k.com/blog/machine_learning

Quiz

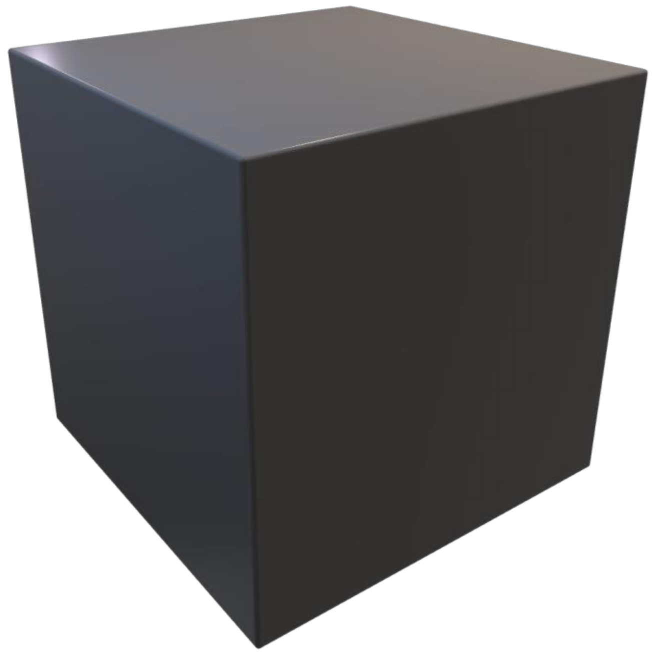
What type of learning is this?

ML Model

Input



Crystal structure



Black box of trained parameters

Output

{Cubic,
Tetragonal,
Orthorhombic,
Hexagonal,
Trigonal,
Monoclinic,
Triclinic}

Crystal system

"Black box" is a common criticism of ML – with the right tools you can open it up!

Class Outline

Materials Learning Basics

A. *Terminology*

B. *Evaluation metrics*

C. *Learning by example*

Function Approximation

Output $\quad$ Input

$$y = f(x)$$

Property

Features

1. Composition
2. Structure

Linear Regression

Property $\quad$ Constant $\quad$ Composition $\quad$ Structure

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Learned weights

Function Approximation

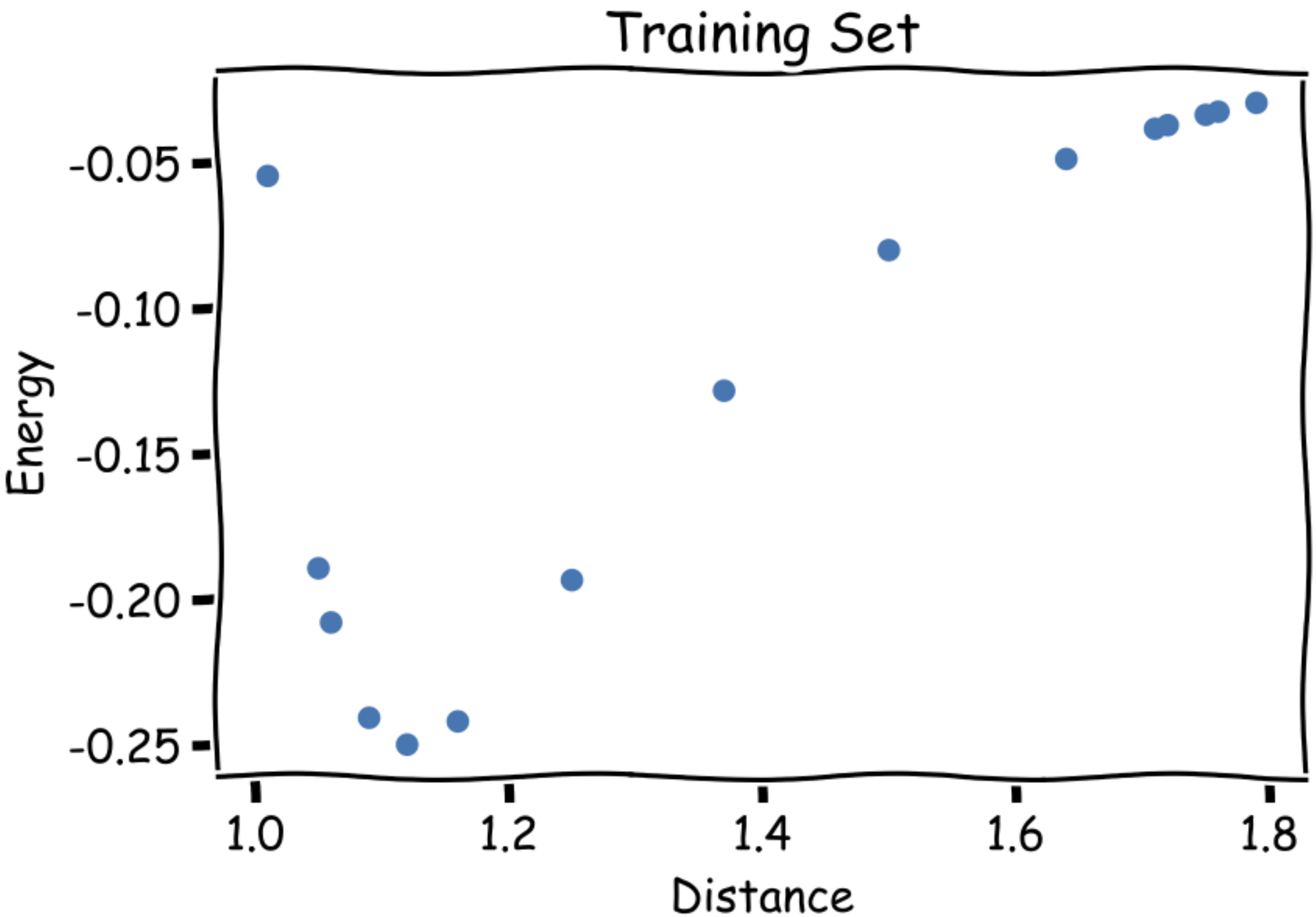
Generalised Linear Models

$$\overset{\text{Property}}{y} = \overset{\text{Constant}}{\beta_0} + \overset{\text{Composition}}{f_1(x_1)} + \overset{\text{Structure}}{f_2(x_2)}$$

Non-Linear Interactions

$$\overset{\text{Property}}{y} = \overset{\text{Constant}}{\beta_0} + \overset{\text{Composition}}{\beta_1 x_1} + \overset{\text{Structure}}{\beta_2 x_2} + \underset{\text{Coupling}}{\beta_3 x_1 x_2}$$

Function Approximation



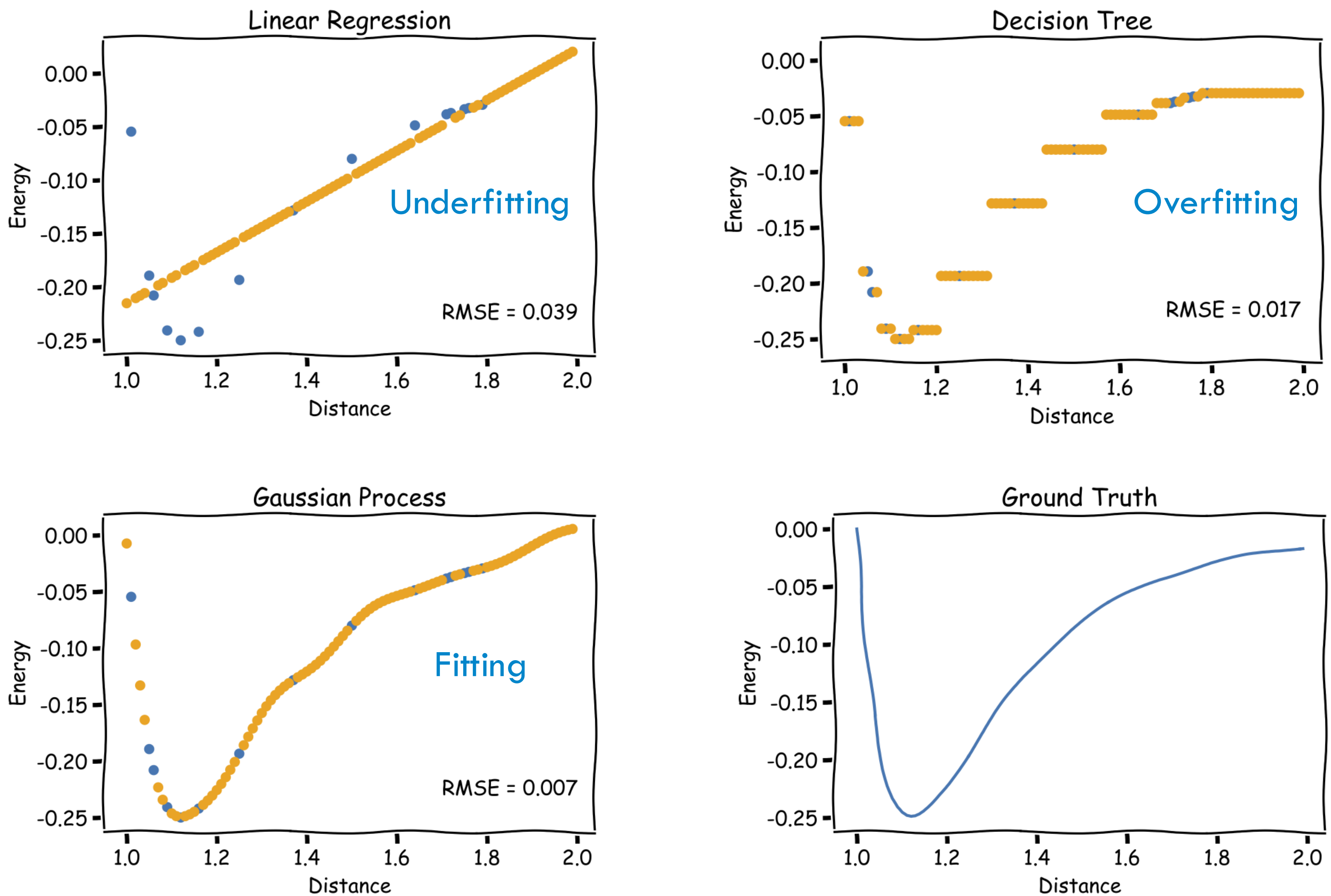
You should recognise the underlying function your undergraduate classes

Function Approximation

```
def truth(r):
    ε = 1
    v = (ε/r)**12 - (ε/r)**6
    return v

xvals = np.arange(1, 2, 0.01)
yvals = truth(xvals)
```

My reference function to generate data for model training and testing

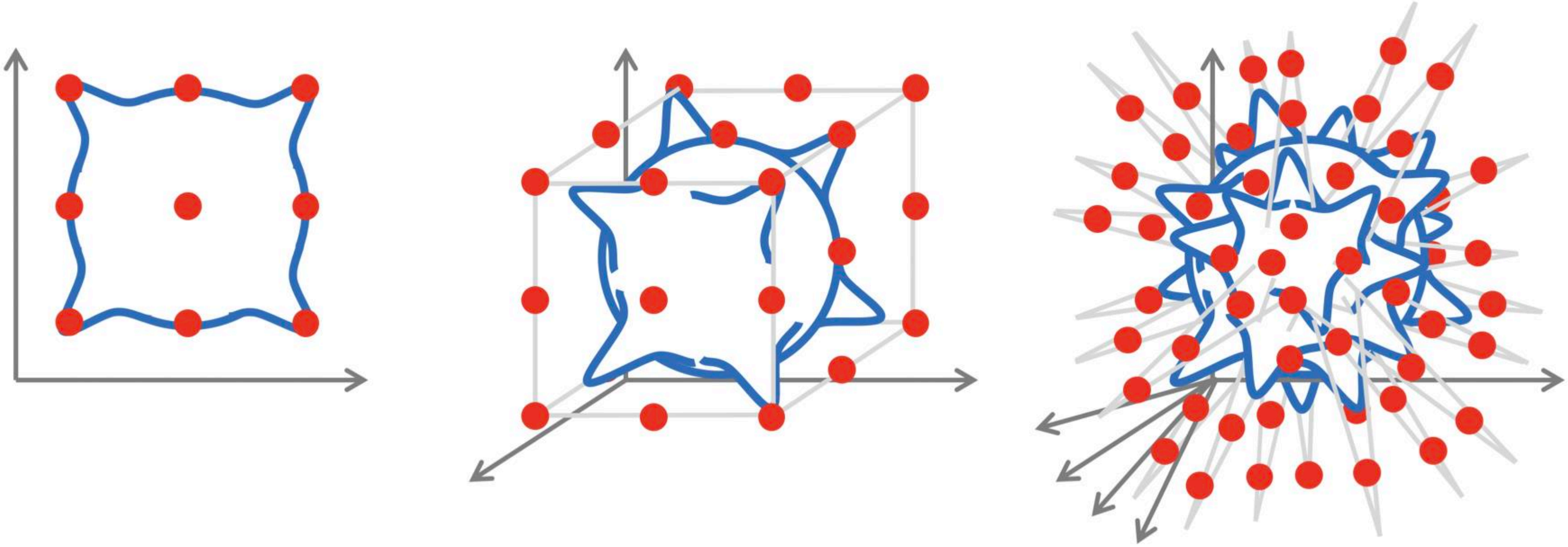
Function Approximation



Default parameters with the scikit-learn Python package

Function Approximation

Standard expansions work in low dimensions (D).
Real problems face the **"curse of dimensionality"**



An exponential increase in the data requirements
needed to cover the parameter space effectively, $O(e^D)$

M. M. Bronstein et al, arXiv:2104.13478 (2021)

Three Components of ML Models

1. Representation

Type of data and model architecture

2. Evaluation

Objective (or scoring) function to distinguish good from bad models

3. Optimisation

Update of model parameters to improve performance

ML Vocabulary

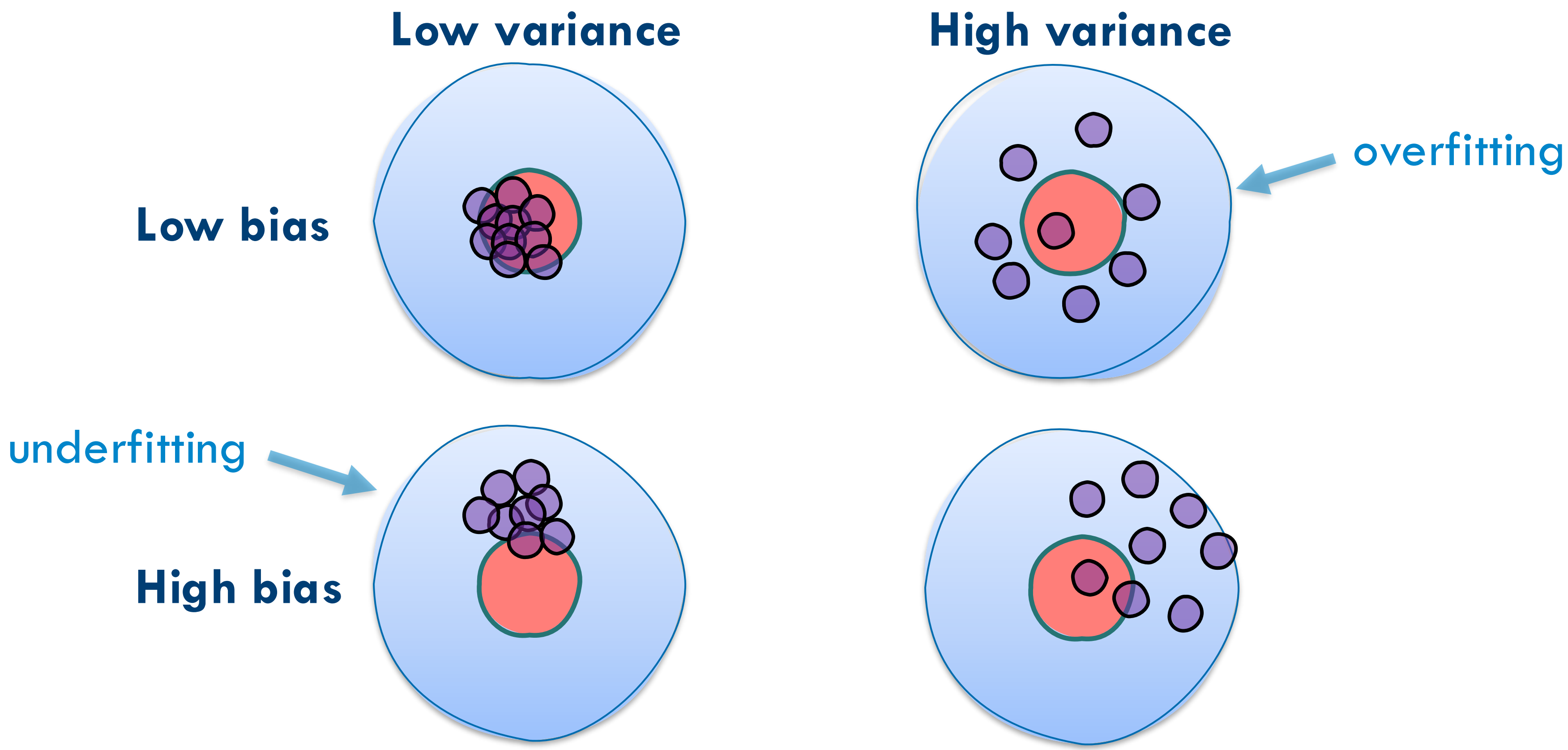
- **Classification model** – input a tensor of *feature values* and output a single discrete value (the *class*)
- **Regression model** – input a tensor of *feature values* and output a continuous (predicted) value

- Feature – an input variable
- Labelled example – a feature with its corresponding label (the "*answer*" or "*result*")
- Ground truth – reliable reference value(s)
- Hyperparameter – model variables that can be tuned to optimise performance, e.g. learning rate

See https://developers.google.com/machine-learning/glossary

ML Vocabulary

- **Bias** – systematic error in the average prediction
- **Variance** – variability around the average prediction



Predicted values (purple circles) compared to the ground truth (red centre)

ML Vocabulary

- **Underfitting** – model too simple to describe patterns
- **Overfitting** – model too complex and fits noise

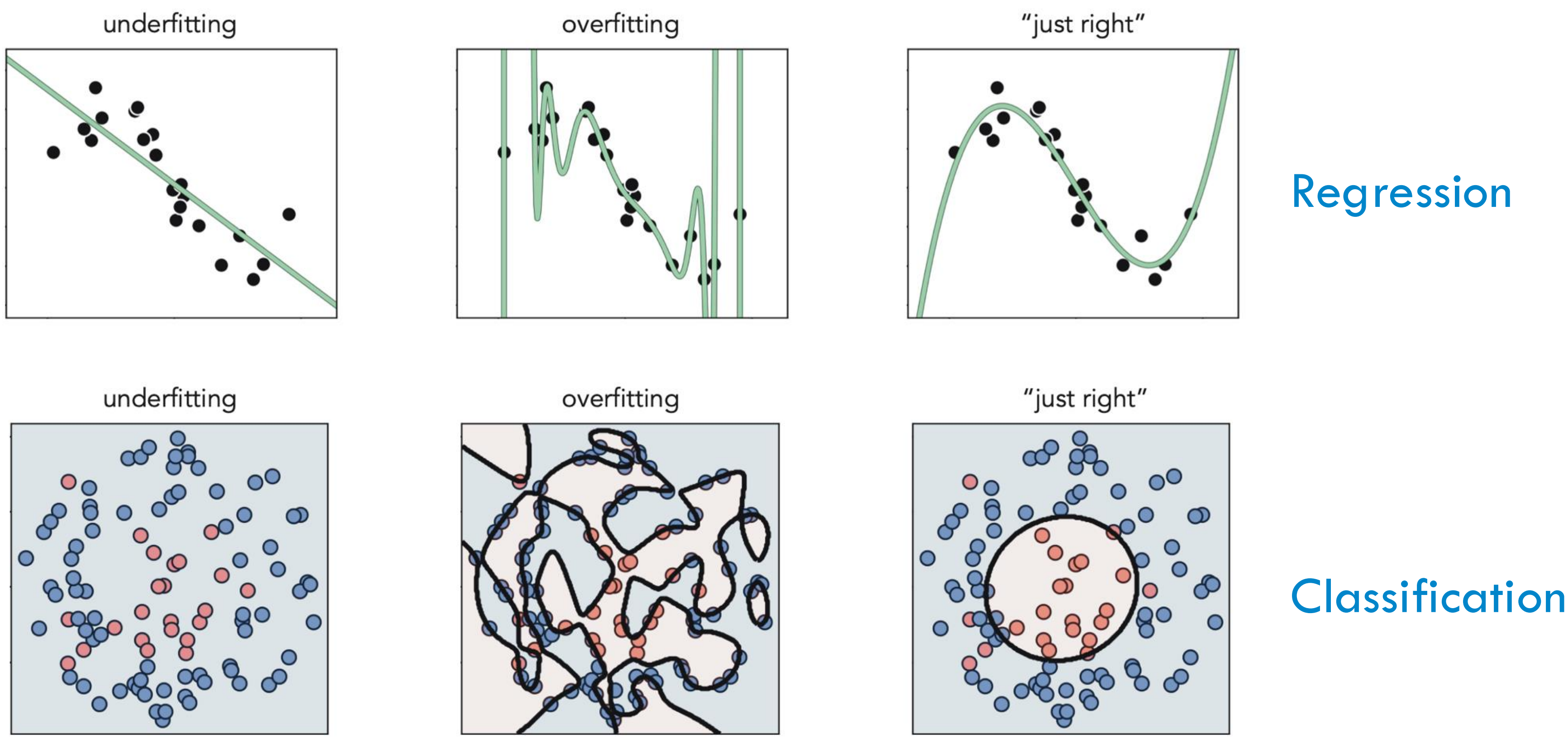
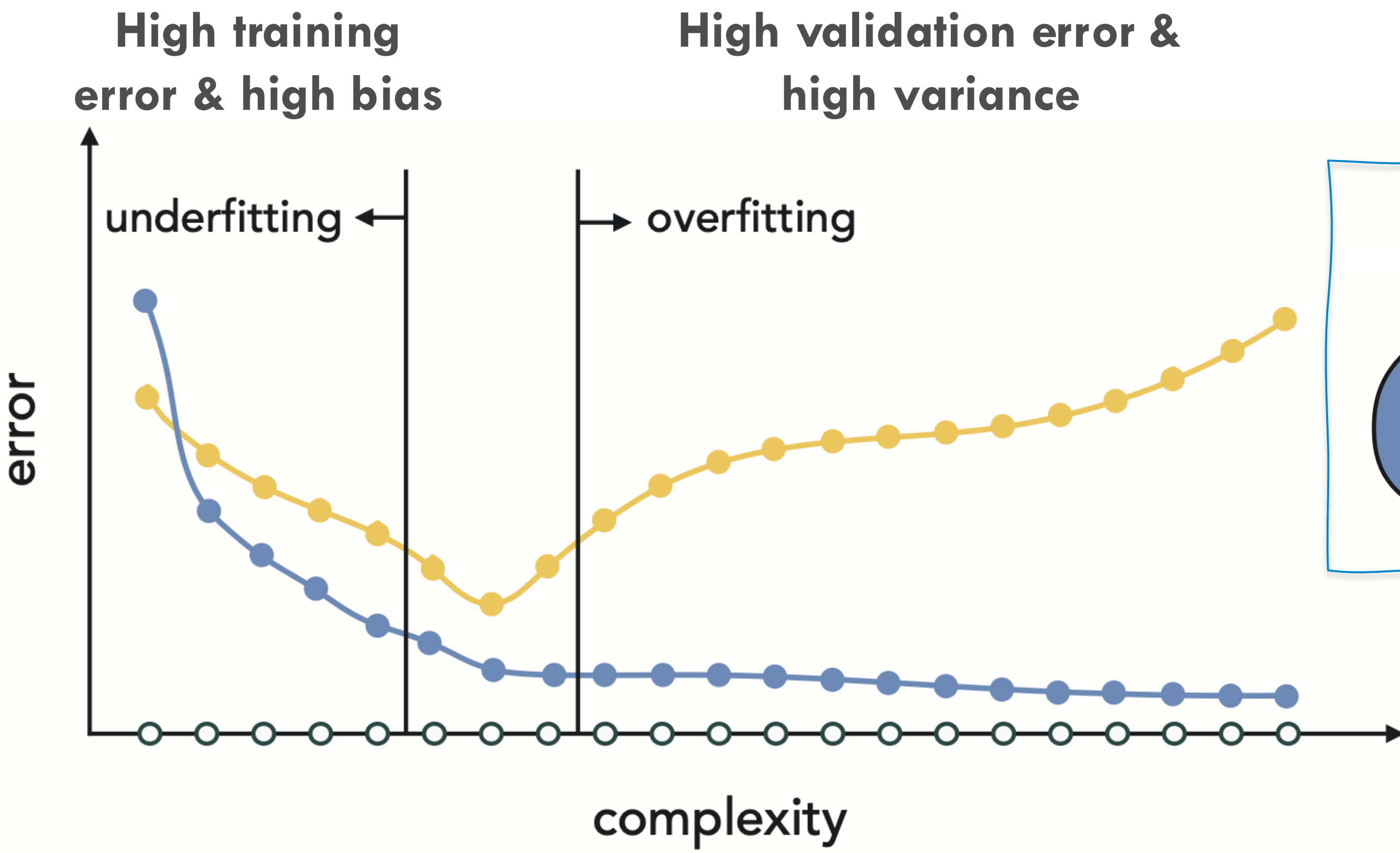


Image from https://github.com/jermwatt/machine_learning_refined

ML Vocabulary

- **Underfitting** – model too simple to describe patterns
- **Overfitting** – model too complex and fits noise



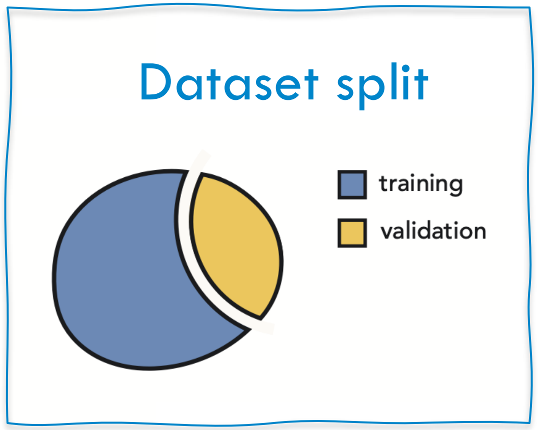
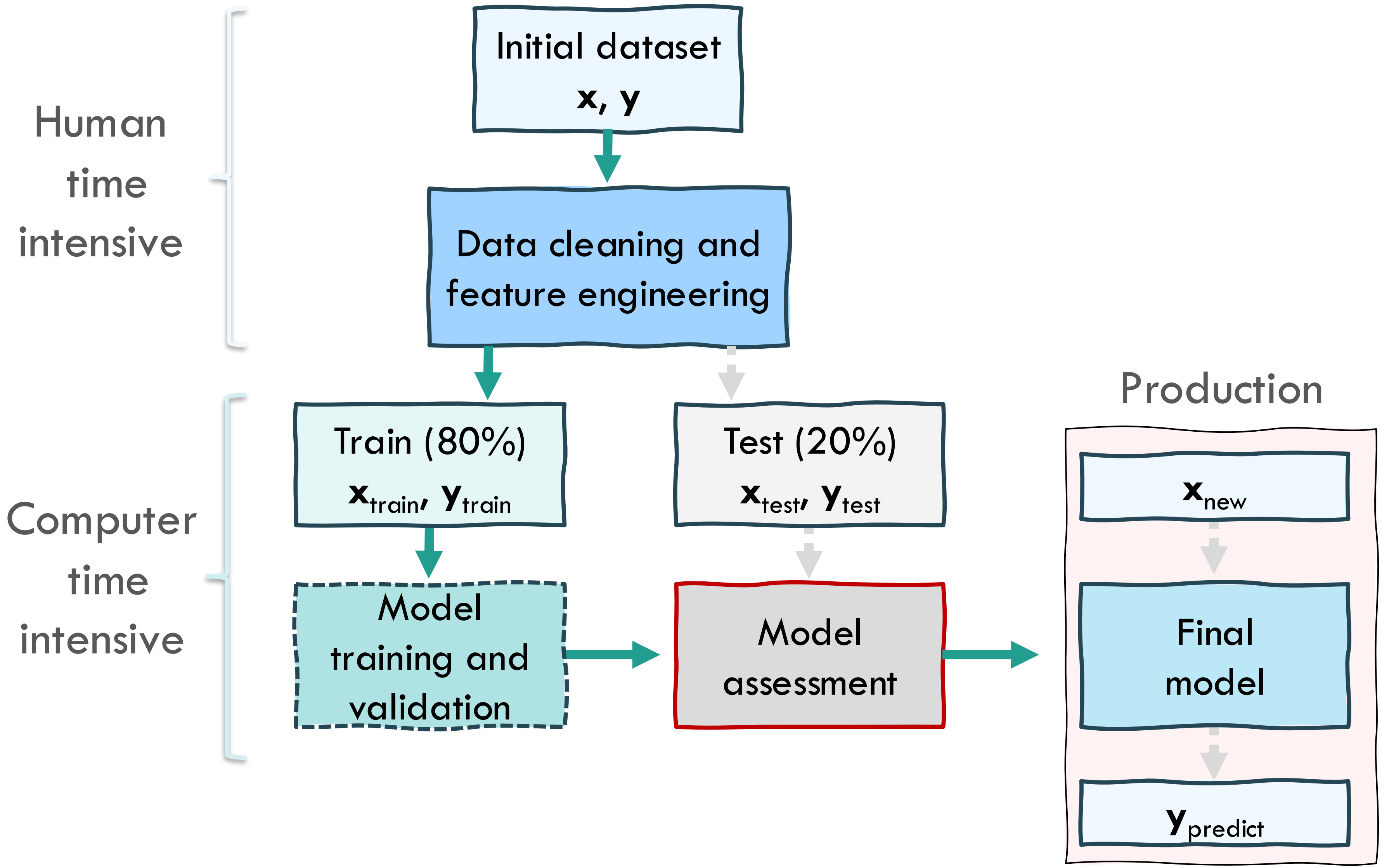


Image from https://github.com/jermwatt/machine_learning_refined

Typical Supervised ML Workflow



The exact workflow depends on the type of problem and available data

Class Outline

Materials Learning Basics

A. *Terminology*

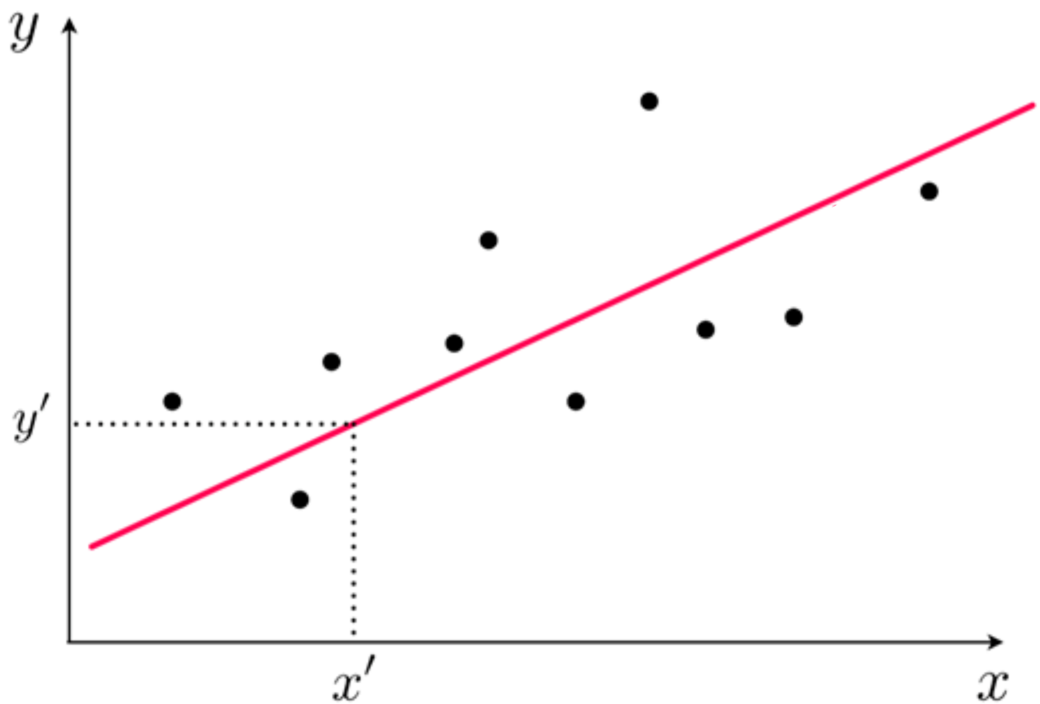
B. *Evaluation metrics*

C. *Learning by example*

Model Assessment

Consider a linear model with optimal weights $\boldsymbol{w}$

$$y^{predicted} = \text{model}(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_i$$



Mean squared error (MSE) $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{model}(x_i, \boldsymbol{w}))^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$

Squaring the error ensures non-negativity and penalises larger deviations

Model Assessment

- Residual – a measure of prediction error $e_i = y_i - y_i^{predicted}$
- MAE – Mean Absolute Error = $\frac{\sum_{i=1}^{n}|e_i|}{n}$
- RMSE – Root Mean Square Error = $\sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}}$
- Standard Deviation – a measure of the amount of dispersion in a set of values. Small = close to the mean. Expressed in the same units as the data, e.g.

lattice parameters *a* = 4 Å, 5 Å, 6 Å
mean = (4+5+6)/3 = 5 Å
deviation = -1, 0, 1; deviation squared = 1, 0, 1
sample variance σ^2 = (1+0+1)/2 = 1
standard deviation σ = 1 Å

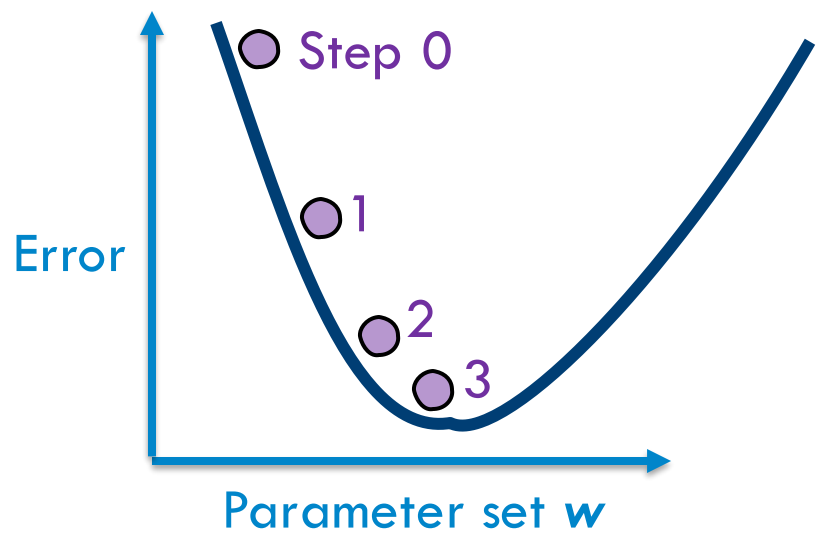
Model Training → Minimising Error

Model weights w are adjusted until a cost function (e.g. RMSE) is minimised

Gradient descent is a popular choice:

$$w_i \rightarrow w_i - \alpha \frac{d\,\text{Error}}{dw_i}$$

Learning rate



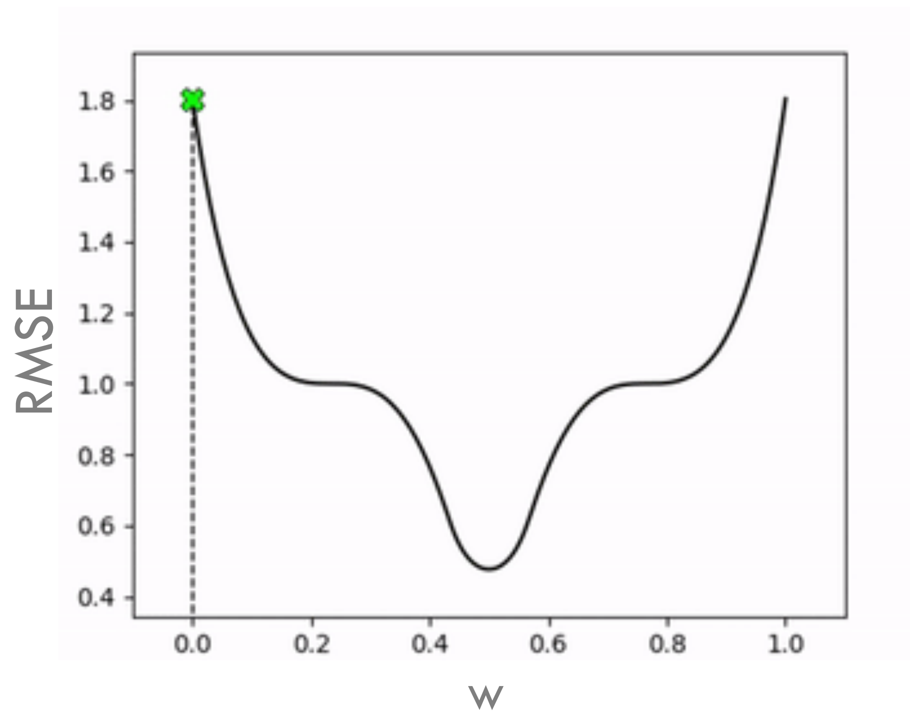
Warning: local optimisation algorithms often miss global minima

Model Training → Minimising Error

Model weights w are adjusted until a cost function (e.g. RMSE) is minimised

Gradient descent is a popular choice:

$$w_i \rightarrow w_i - \underbrace{\alpha}_{\text{Learning rate}} \frac{d\,\text{Error}}{dw_i}$$

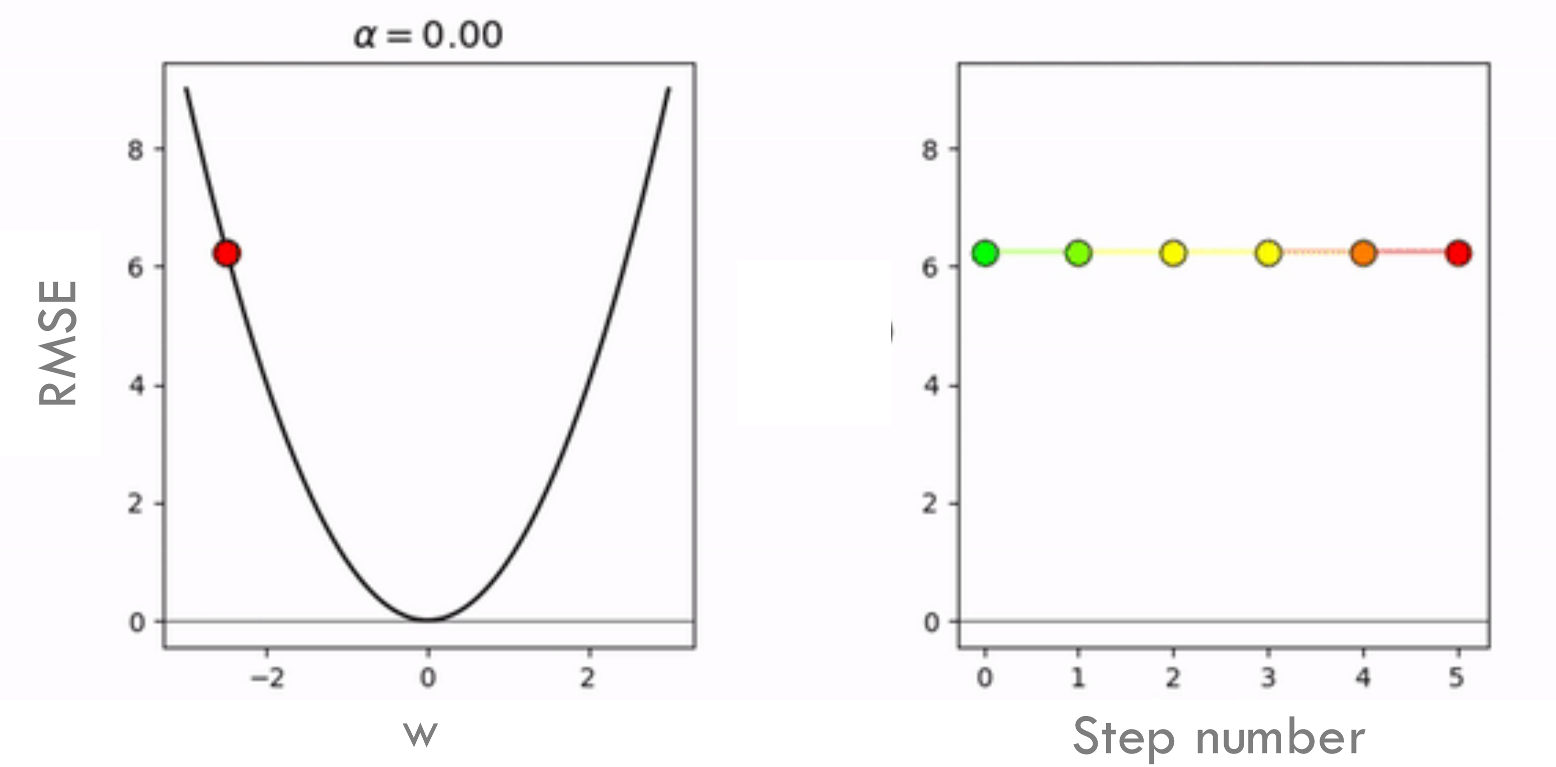


Animation from https://github.com/jermwatt/machine_learning_refined

Model Training → Minimising Error

Optimisation algorithms have their own parameters, e.g. step size and no. of iterations

Learning rate (step size)



Animation from https://github.com/jermwatt/machine_learning_refined

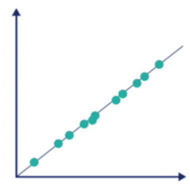
```python
# Define a function and its gradient
def f(x):
    return x**2

def df(x):
    return 2*x

# Initialise starting point and learning rate
x = 5
learning_rate = 0.1

# Perform gradient descent for 10 iterations
for i in range(10):

    # Compute the gradient of the function at x
    grad = df(x)

    # Update x using gradient descent
    x -= learning_rate * grad
```

Correlation Coefficient (r)

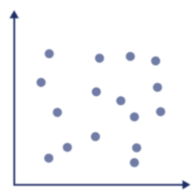
Describes the strength of the relationship between two variables (e.g. “ground truth” vs predicted values)

$r \in [-1,1]$

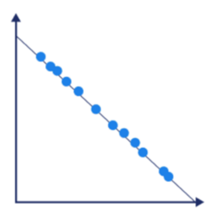
Positive: variables change in the same direction



Zero: no relationship between the variables



Negative: variables change in opposite directions



Pearson correlation*

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{x})^2}}$$

Reminder: correlation does not imply causation

*Outlined by Auguste Bravais (1844); https://bit.ly/3Kv75GJ

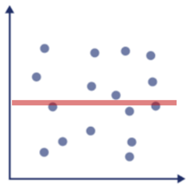
Coefficient of Determination (r^2)

Measure of the goodness of fit for a model.

Describes how well that known data is approximated

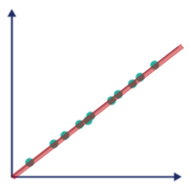
$r^2 \in [0,1]$

Zero: baseline model with no variability that predicts $\bar{y}$



0.5: 50% of the variability in y is accounted for

One: model matches observed values of y exactly



Three equivalent definitions

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

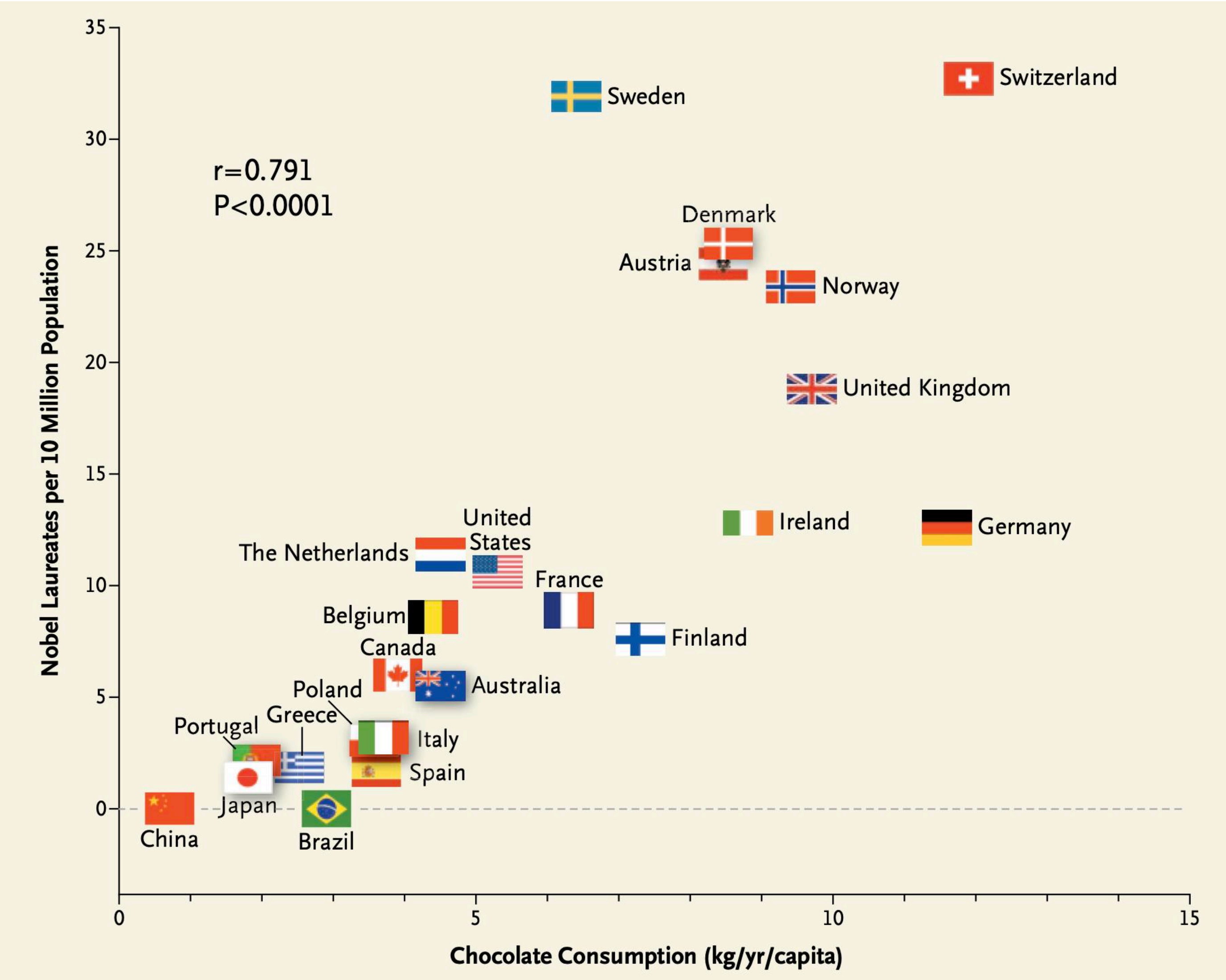
$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^{predicted}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

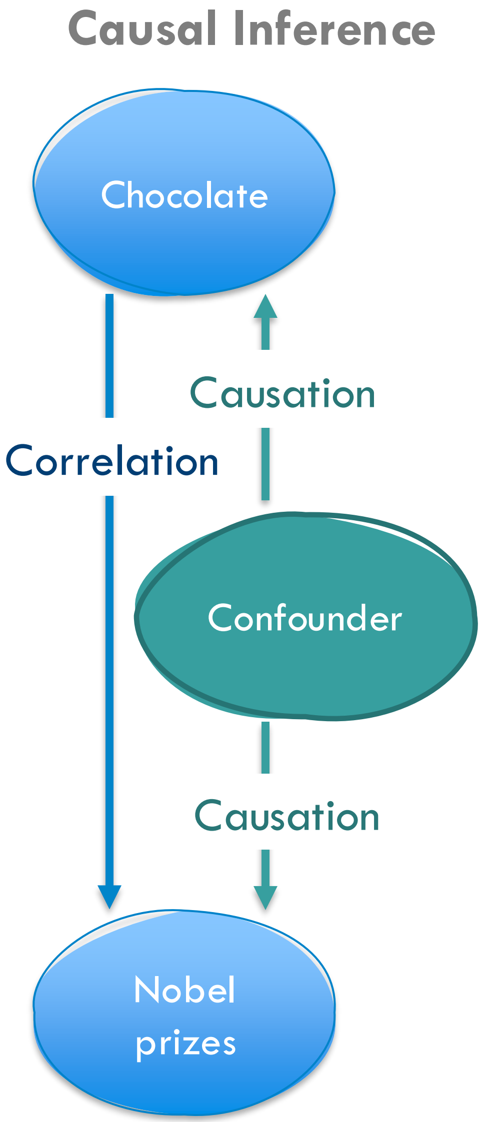
$$r^2 = 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Note: a unitless metric. Alternative definitions are sometimes used

S. Wright "Correlation and Causation", J. Agri. Res. 20, 557 (1921)

Correlation, Causation…





F. Messereli, New England Journal of Medicine 367, 1562 (2012)

Classification Metrics

Confusion (or error) matrix provides a summary of classification model performance

Predicted class

$$\begin{array}{cc} & \begin{array}{cc} + & - \end{array} \\ \begin{array}{c} \textbf{Actual class} \\ \begin{array}{c} + \\ - \end{array} \end{array} & \begin{bmatrix} \text{True positive (TP)} & \text{False negative (FN)} \\ \text{False positive (FP)} & \text{True negative (TN)} \end{bmatrix} \end{array}$$

$$\begin{bmatrix} 70 & 0 \\ 0 & 30 \end{bmatrix}$$

Perfect model to classify metals and insulators (N = 100)

$$\begin{bmatrix} 66 & 4 \\ 8 & 22 \end{bmatrix}$$

My best model

Accuracy = Correct/Total

(70+30)/100 = 100 %

(66+22)/100 = 88 %

K. Pearson "Mathematical Contributions to the Theory of Evolution" (1904)

Classification Metrics

Confusion (or error) matrix provides
a summary of classification model performance

$$\begin{array}{ccc} & \multicolumn{2}{c}{\textbf{Predicted class}} \\ & + & - \\ \begin{array}{c}\textbf{Actual}\\\textbf{class}\end{array}\begin{array}{c}+\\ \\-\end{array} & \multicolumn{2}{c}{\begin{bmatrix} \text{True positive (TP)} & \text{False negative (FN)} \\ \text{False positive (FP)} & \text{True negative (TN)} \end{bmatrix}} \end{array}$$

$$\begin{bmatrix} 70 & 0 \\ 0 & 30 \end{bmatrix}$$

Perfect model to classify metals and insulators ($N = 100$)

$$\begin{bmatrix} 66 & 4 \\ 8 & 22 \end{bmatrix}$$

My best model

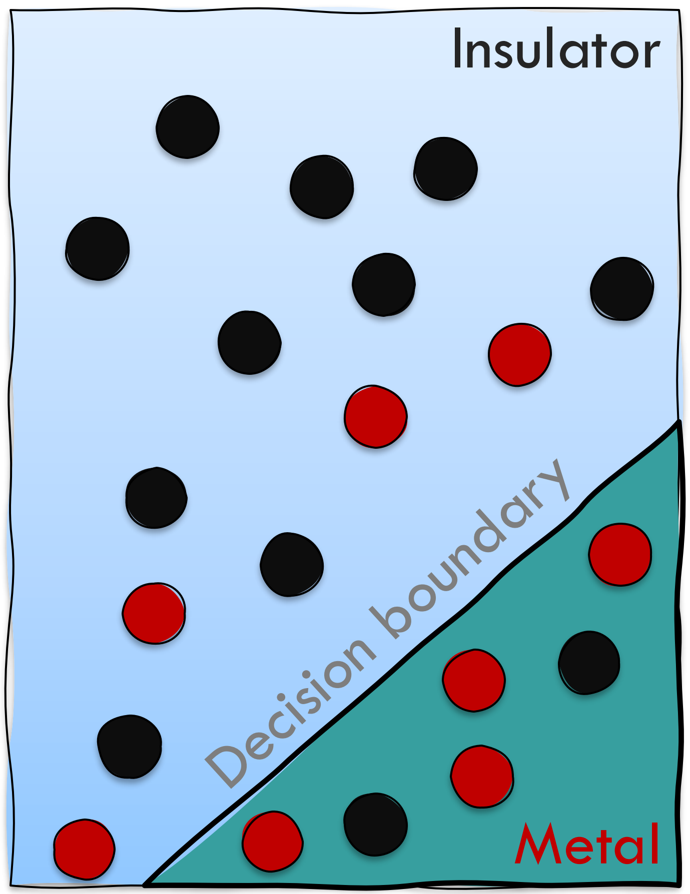
Sensitivity = TP/(TP+FN)

70/(70+0) = 100 %

66/(66+4) = 94 %

Quiz

Fill in "?" for this confusion matrix



		Predicted class	
		Insulator	**Metal**
Actual class	**Insulator**	10	2
	Metal	?	4

There are 20 data points in total

Class Outline

Materials Learning Basics

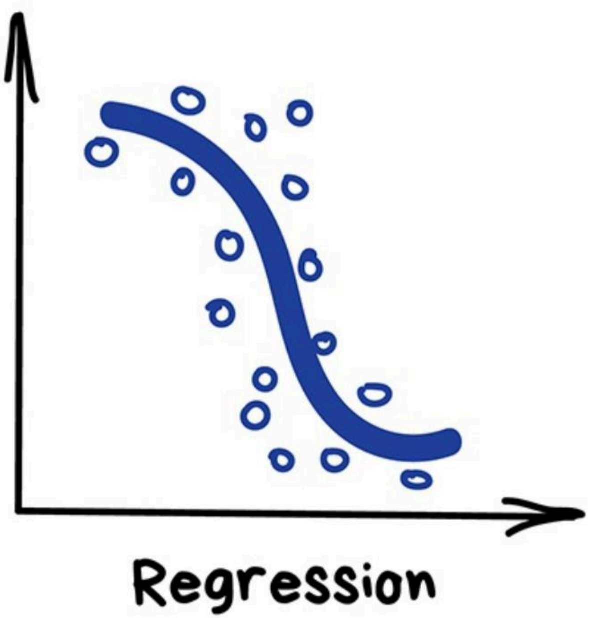
A. *Terminology*

B. *Evaluation metrics*

C. ***Learning by example***

Supervised Regression

Model that maps an input to an output based on example input-output pairs (labelled data)



Regression predicts a continuous value, e.g. to extract a reaction rate

Target variable

Error

$$\boxed{\boldsymbol{y} = f(\mathbf{x}) + \varepsilon}$$

Learned function

Regression Example

Predict the dielectric constant of a crystal

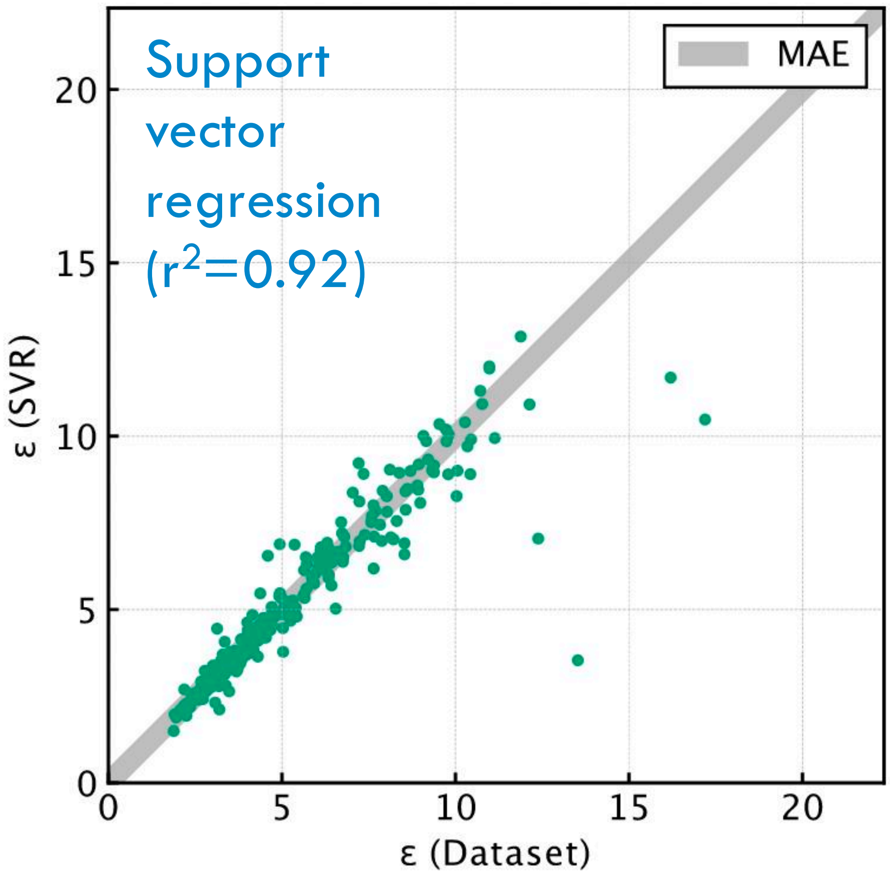


TABLE I. Features used to train the support vector regression model.

Feature	Dimensions
Bandgap[a]	1
Δ Pauling energy[b]	1
Material density[a]	1
Formation energy (per atom)[a]	1
Oxidation state (minimum, variation)[a]	2
Madelung energy (minimum, maximum)[c]	2
Ionic species (one hot encoded)[c]	85

TABLE II. Performance metrics of support vector regression (SVR) and deep neural network (DNN) for training and test data. Metrics are mean Pearson's correlation coefficient (r^2), average error (MAE), mean squared error (MSE), and root mean square error (RMSE).

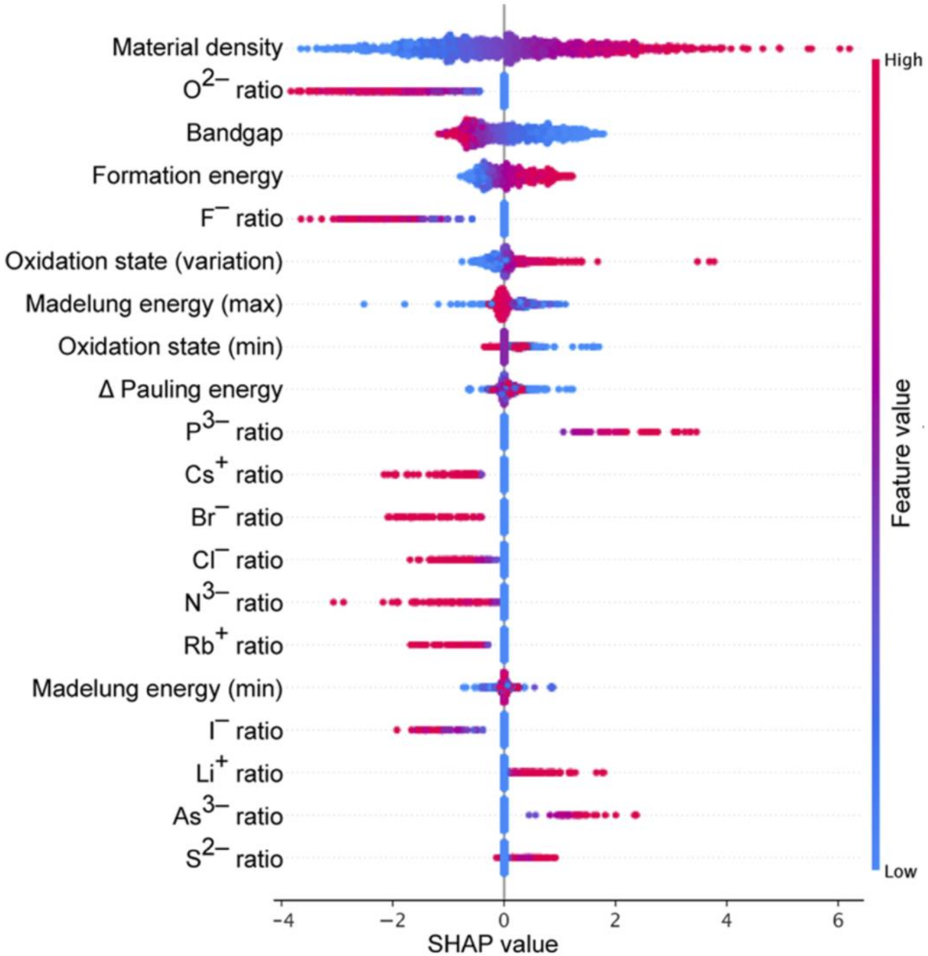
	SVR		DNN	
Metric	Training	Test	Training	Test
r^2	0.92	0.86	0.95	0.84
MAE	0.24	0.44	0.20	0.55
MSE	0.69	0.99	0.38	1.17
RMSE	0.83	0.99	0.62	1.08

Note: outliers are often interesting cases (poor or exceptional data)

K. Morita et al, J. Chem. Phys. 153, 024503 (2020)

Regression Example

Predict the dielectric constant of a crystal



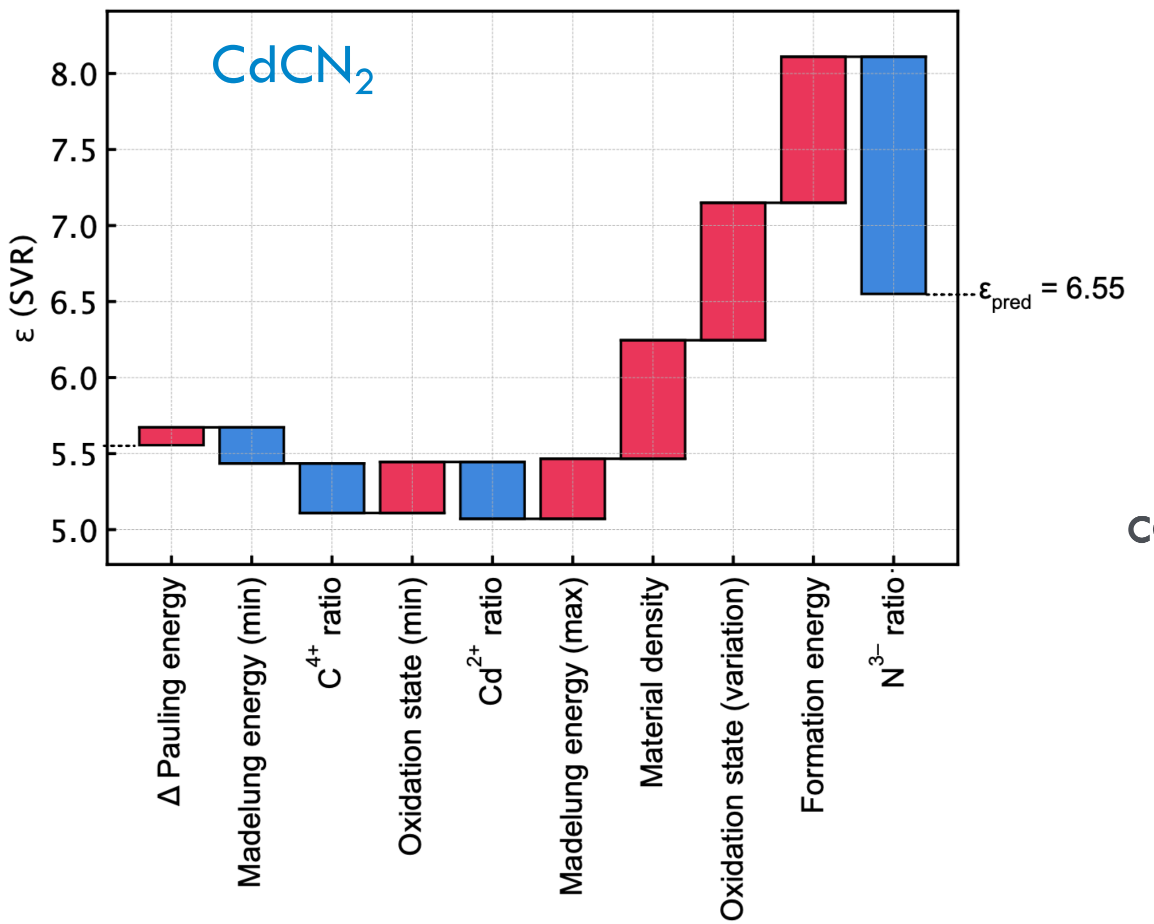
SHAP (SHapley Additive exPlanations) analysis is a method for interpreting ML models

Relative importance of input features in making predictions

A positive SHAP indicates a feature contributes to an increase in the prediction

Regression Example

Predict the dielectric constant of a crystal



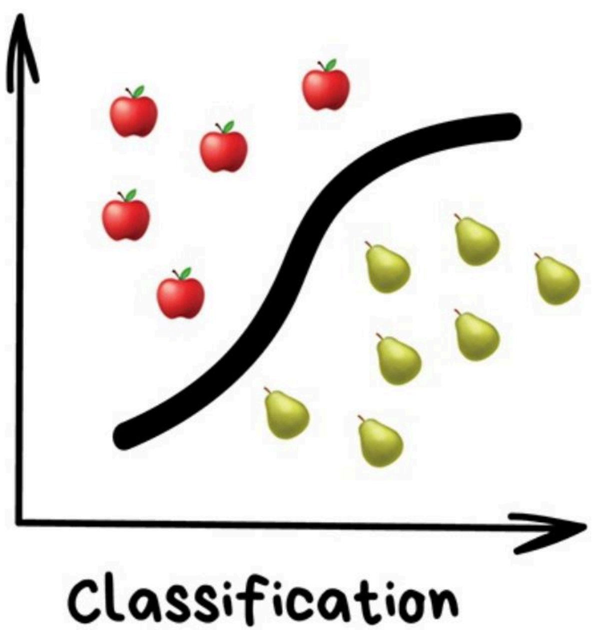
Transparent breakdown of a predicted value

Note: a physical connection is not implied (only a correlation)

K. Morita et al, J. Chem. Phys. 153, 024503 (2020); https://github.com/slundberg/shap

Supervised Classification

Model that maps an input to an output based on example input-output pairs (labelled data)



Classification predicts a category, e.g. decision trees for reaction outcomes

Class label

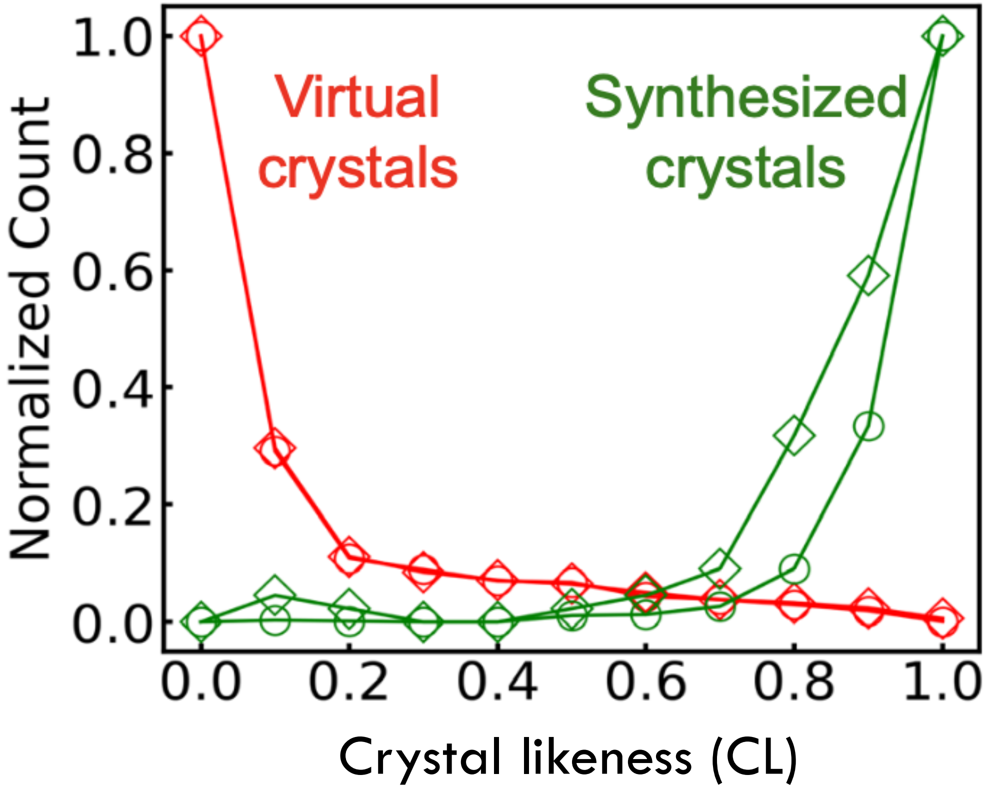
$$\boldsymbol{y} = f(\mathbf{x})$$

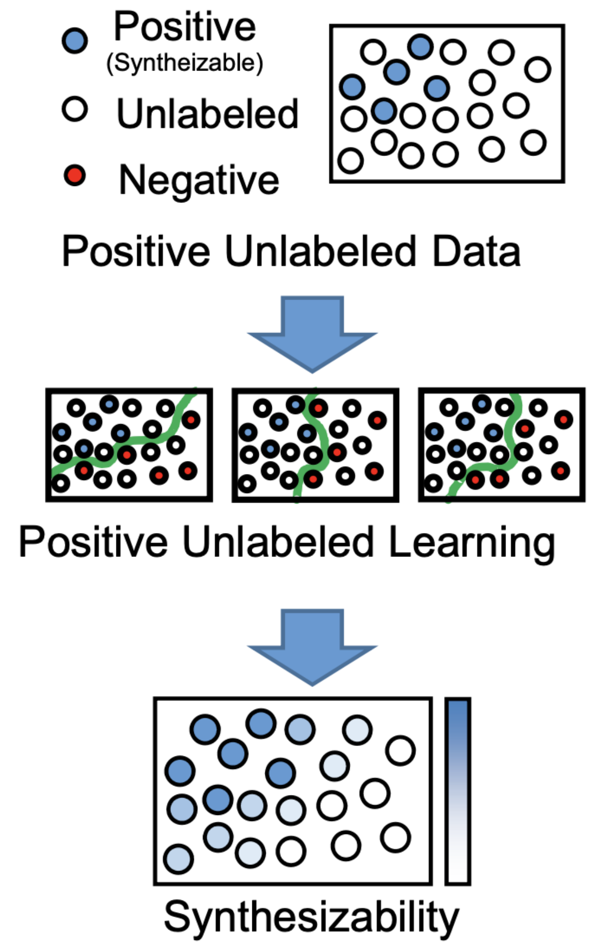
Classifier

Assignment can be absolute or probabilistic (e.g. 90% apple, 10% pear)

Classification Example

Predict if a material will be stable or unstable





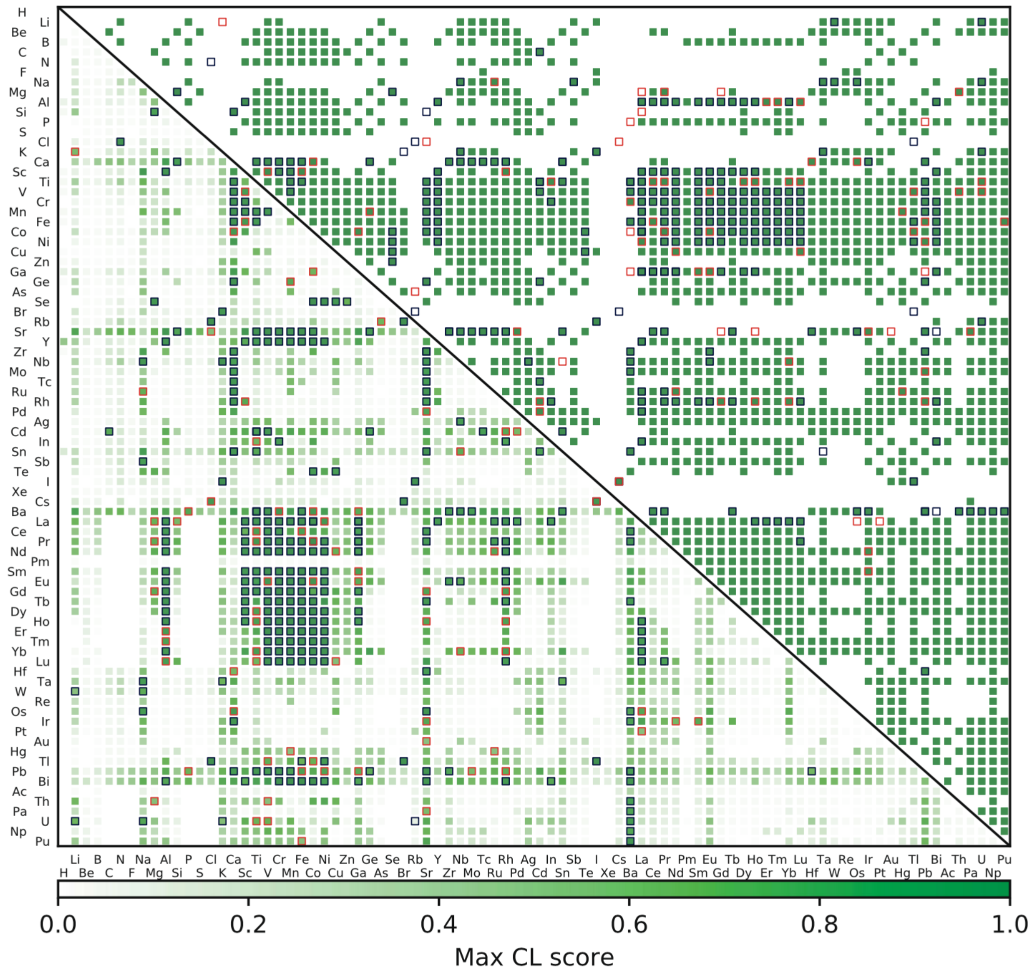
Probabilistic score for the class label

Classification Example

ABX_3 perovskite crystals

Neural network model

Improved selectivity for promising compositions



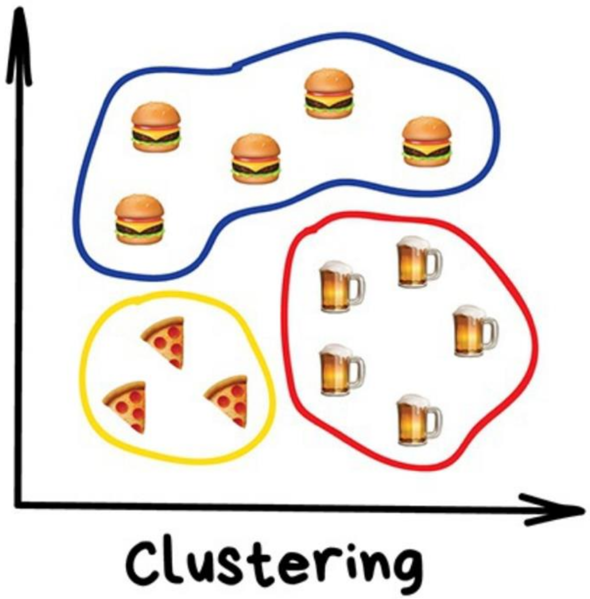
Radius ratio rules

Likelihood of formation

G. H. Gu et al, npj Computational Materials 8, 71 (2022)

Unsupervised Learning

Model that can identify trends or correlations within a dataset (unlabeled data)



Clustering groups data by similarity, e.g. high-throughput crystallography

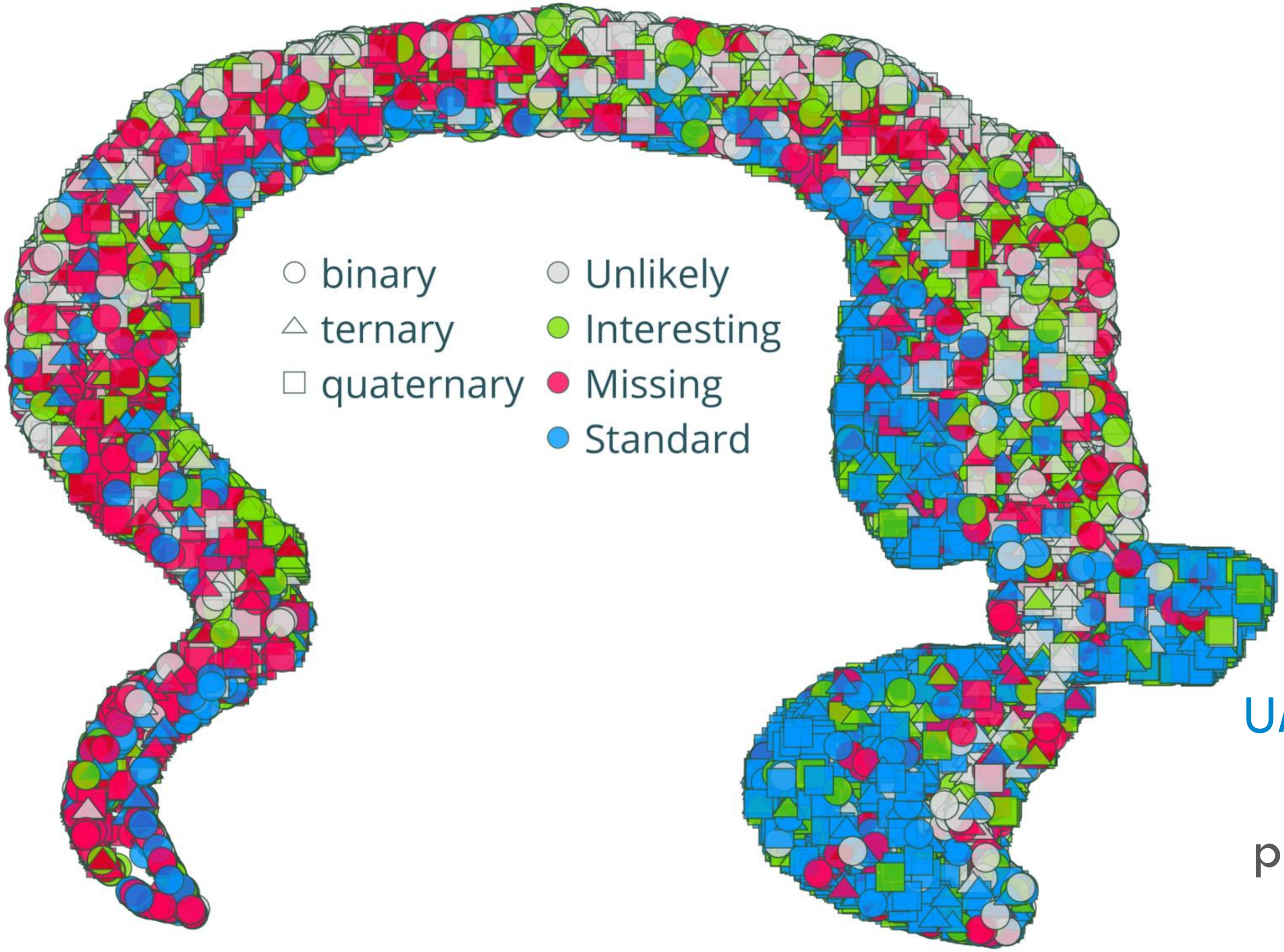
Input data

$$\mathbf{x} \rightarrow f(\mathbf{x})$$

Transformation to new representation

Unsupervised Example

Map materials space according to their features



Dimensionality reduction techniques:

PCA: Principal component analysis

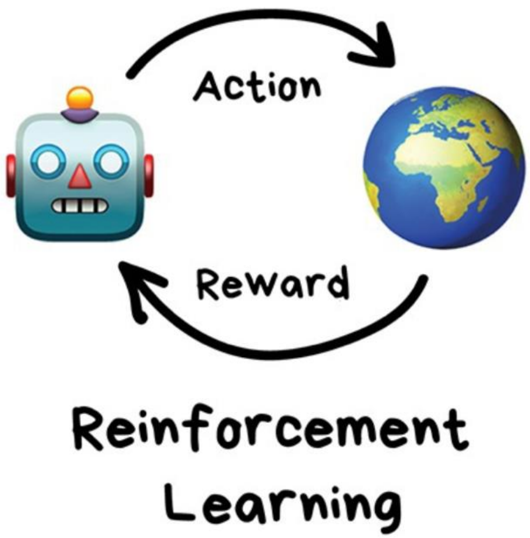
t-SNE: t-distributed stochastic neighbour embedding

UMAP: Uniform manifold approximation and projection for dimension reduction

Hyunsoo Park et al, Faraday Discussions 256, 601 (2025)

Reinforcement Learning

Model that performs a series of actions by trial and error to achieve an objective



Maximise reward, e.g. reaction conditions to optimise yield

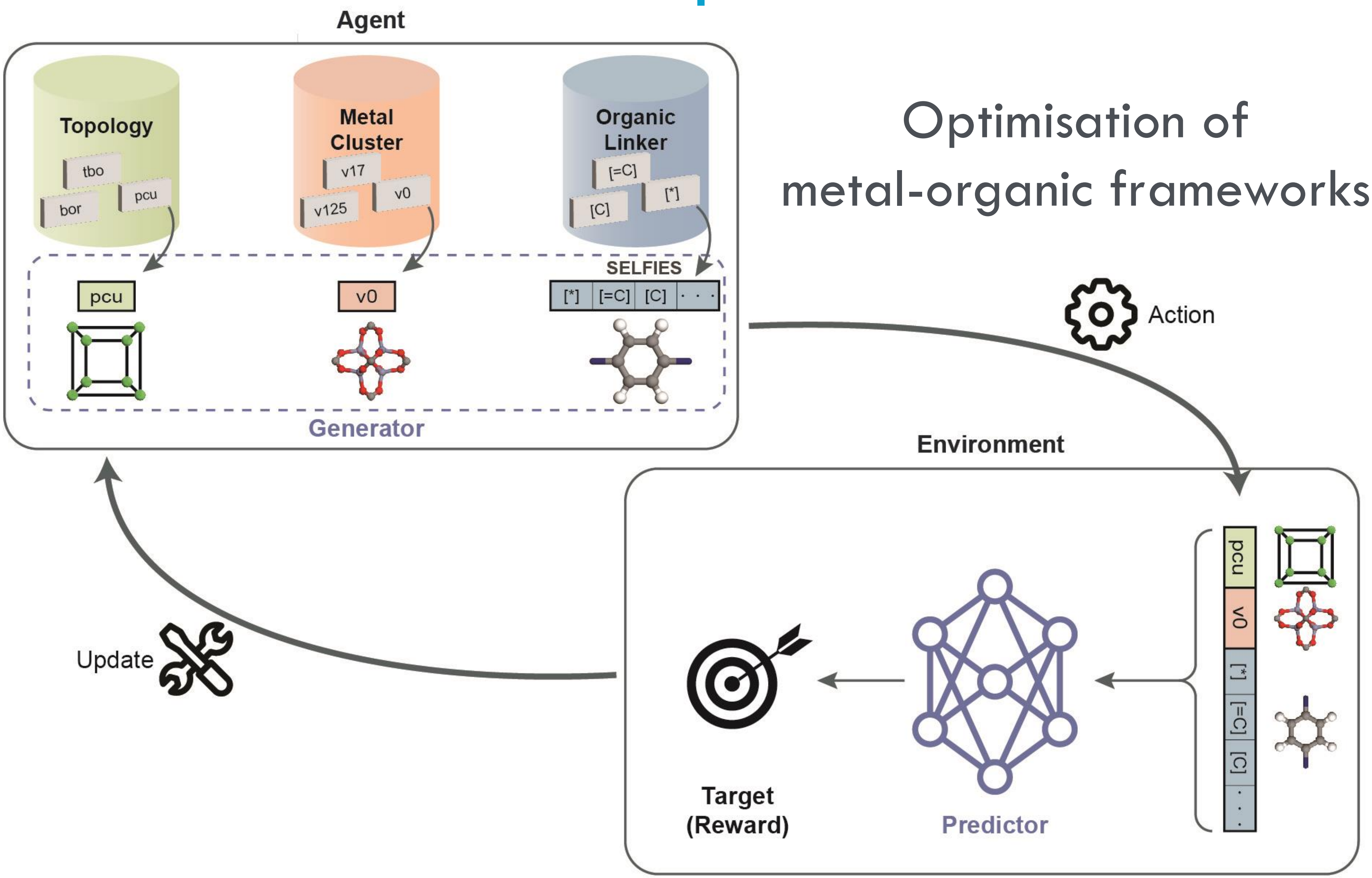
Automated experiments

Samples, conditions, etc.

Agent $\rightleftharpoons$ Environment

Actions

Reinforcement Example



Optimisation of metal-organic frameworks

Hyunsoo Park et al, Digital Discovery 3, 728 (2024)

Class Outcomes

1. Define machine learning
2. Describe the three components of machine learning with examples
3. Explain the statistical metrics used to assess model performance

Activity:

Crystal hardness