Module: **Machine Learning and Data-Driven Materials**

Machine Learning Intro

Institute for Materials Discovery (IMD)

University College London (UCL)

Dr. Zied Hosni, Dr. Miguel Rodriguez Pineda, Prof. Adham Hashibon
Email: ucqshos@ucl.ac.uk

Objectives and learning outcome

- Data splitting and cross-validation
- Regression techniques and QSPR
- Features selection in machine learning
- Descriptors and fingerprints
- Difference between overfitting and underfitting
- Practical
  - python to pre-process a messy iris dataset
  - multiple linear regression
  - Regression alternatives in Knime

Data splitting

- Data splitting is a technique used in machine learning to divide a dataset into two or more subsets, typically called training set and testing set.
- The purpose of data splitting is to provide a way to evaluate the performance of a machine learning model on unseen data, and to prevent overfitting.

Full Dataset

No	X1	X2	X3	X4	Y
0	0	1	0	1	0.23
1	0	0	1	0	0.55
2	0	1	1	1	0.74
3	1	0	0	0	0.44
4	1	1	0	1	0.11

Training set

No	X1	X2	X3	X4	Y
0	0	1	0	1	0.23
1	0	0	1	0	0.55
2	0	1	1	1	0.74
3	1	0	0	0	0.44

Test set

No	X1	X2	X3	X4	Y
4	1	1	0	1	0.11

Holdout

- randomly splits the data into two subsets, a training set and a testing set. The model is trained on the training set and evaluated on the testing set.

K-fold Cross-Validation

- This technique divides the data into k subsets, called "folds". The model is trained on k-1 folds, and evaluated on the remaining fold. This process is repeated k times, with a different fold used as the testing set each time. The final evaluation is the average of the k evaluations.

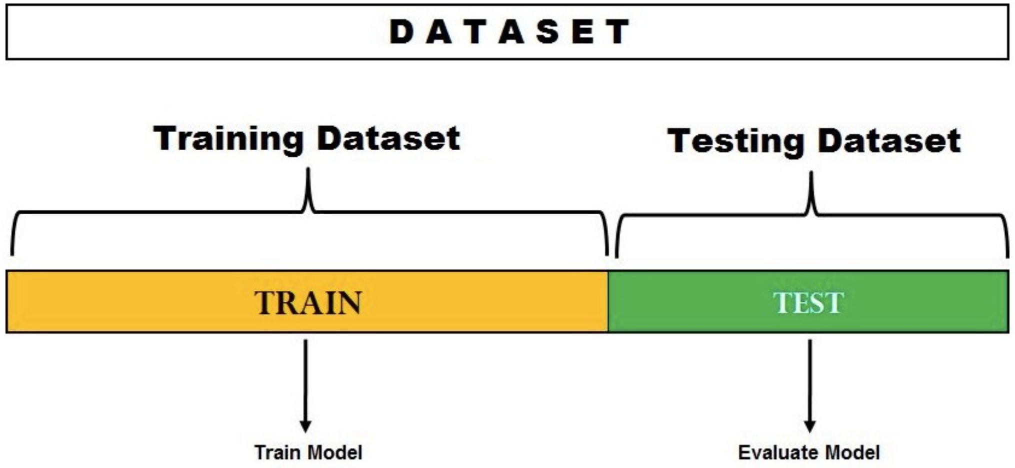
Stratified K-fold Cross-Validation

- This technique is similar to K-fold cross-validation, but it ensures that the class distribution is roughly the same in each fold.

Leave-One-Out Cross-Validation

- This technique is a special case of K-fold cross-validation where k is set to the number of samples in the dataset. The model is trained on all but one sample, and evaluated on the left-out sample.

Holdout splitting



DATASET

Training Dataset

Testing Dataset

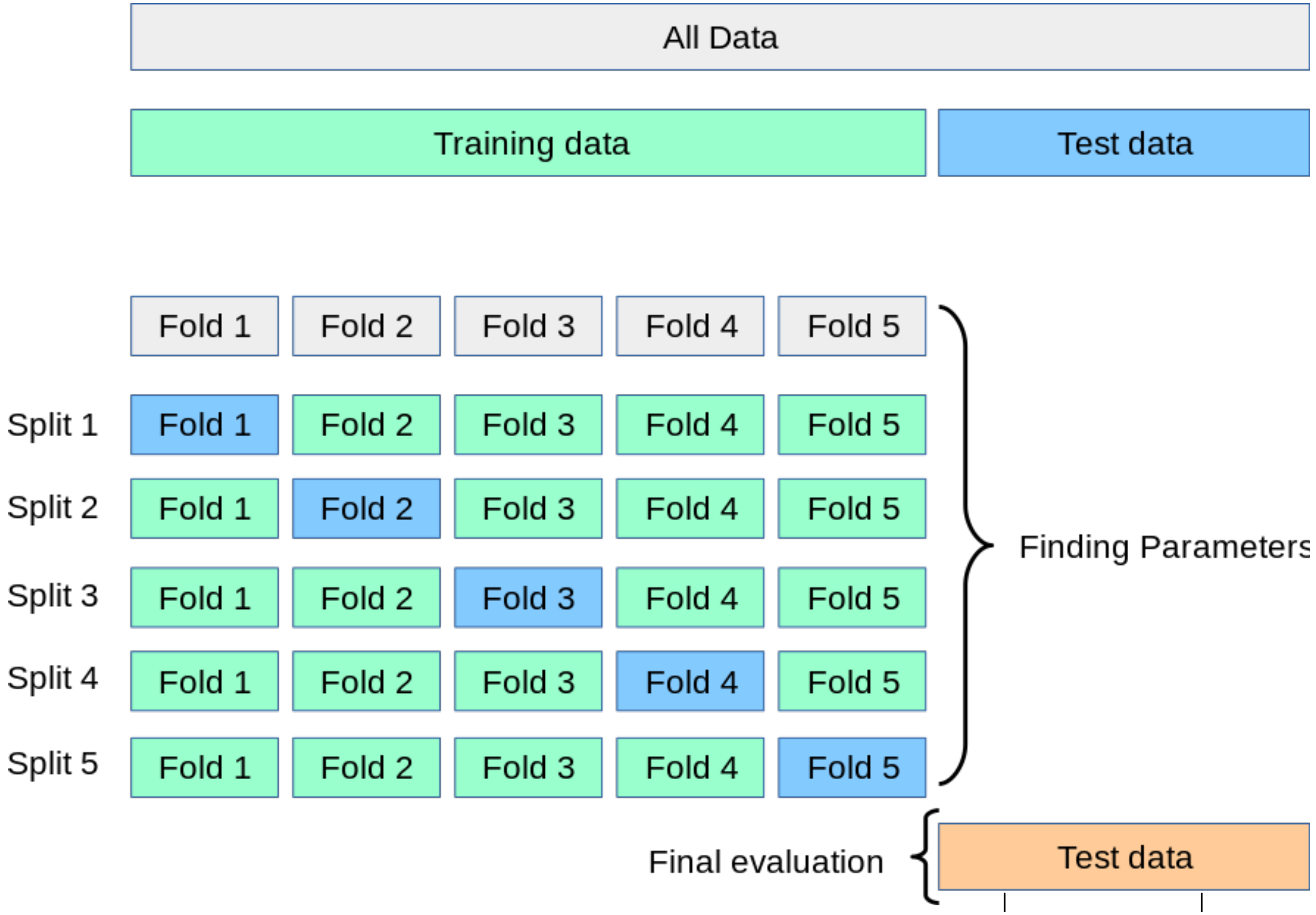
TRAIN

TEST

Train Model

Evaluate Model

K-fold cross-validation splitting



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

Some of the pros of using data splitting include:

- It provides a way to evaluate the performance of a machine learning model on unseen data.
- It helps to prevent overfitting.

Some of the cons of using data splitting include:

- It may not be suitable for very small datasets.
- It may not provide an accurate estimate of the model's performance on new data if the data is highly imbalanced.

A practical example of data splitting using scikit-learn library is:

```
from sklearn.model_selection import train_test_split

# Split the data into a training set and a testing
set

X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2)
```

X1	X2	X3	X4	Y
0	1	0	1	0.23
0	0	1	0	0.55
0	1	1	1	0.74
1	0	0	0	0.44
1	1	0	1	0.11

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Alpaydin, E. (2010). Introduction to Machine Learning. Cambridge, MA: MIT Press.

G. Montavon, G. B. G. Stukowski, "Machine learning for the identification of crystal structures from diffraction patterns" Acta Materialia, vol. 150, pp. 278-292, 2018.

D.J. Li, Y. Li, "A machine learning approach to predict mechanical properties of metallic glasses" Acta Materialia, vol. 160, pp. 45-54, 2018.

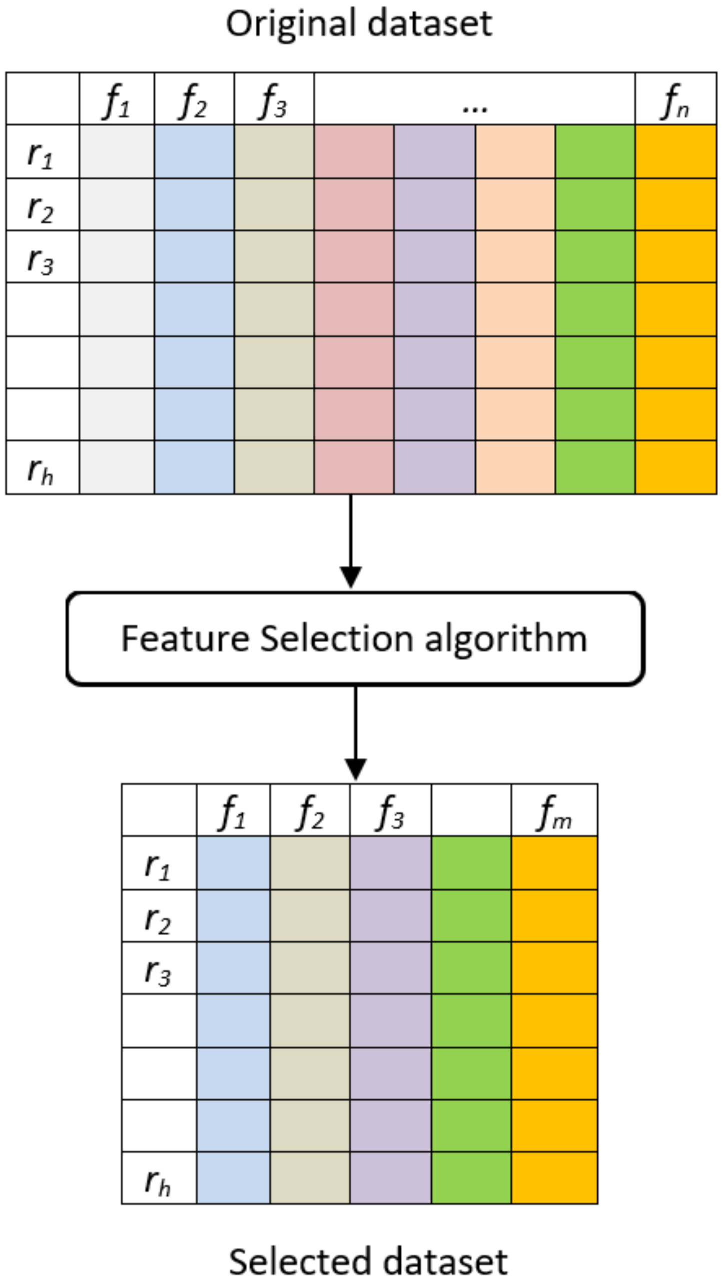
Features selection in machine learning

A Challenge in ML..

..datasets often contain a large number of features, which can make the training and analysis of models more complex and computationally expensive.

Features selection in machine learning

- Feature selection is a technique used to identify and select a subset of the most relevant features from a larger set of features.

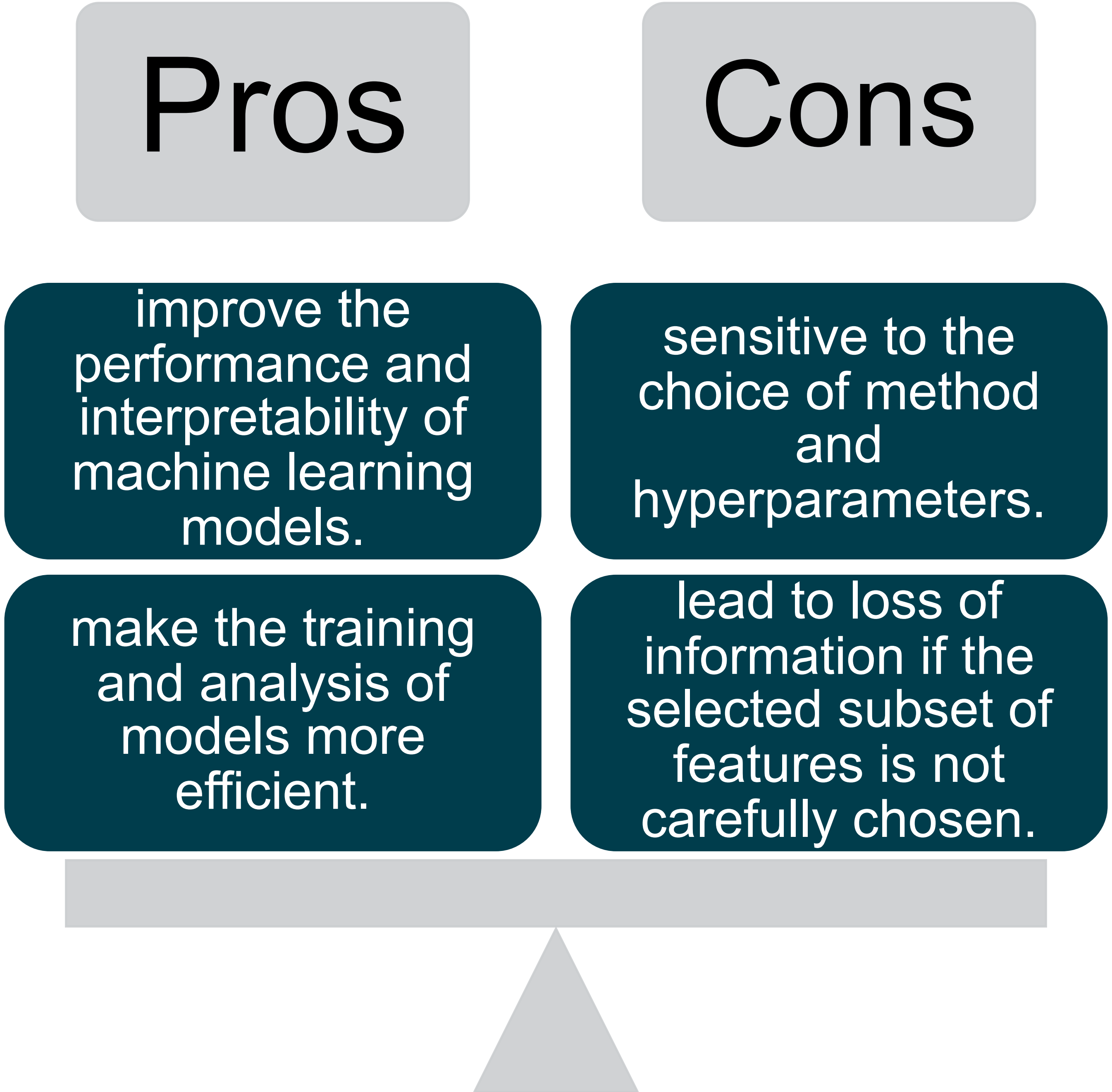


The need to feature selection

- ✓ reduce the dimensionality of the data
- ✓ improve the interpretability of the model
- ✓ improve the performance of the model

Some popular feature selection techniques

- ❖ filter methods
- ❖ wrapper methods
- ❖ embedded methods



Pros

improve the performance and interpretability of machine learning models.

make the training and analysis of models more efficient.

Cons

sensitive to the choice of method and hyperparameters.

lead to loss of information if the selected subset of features is not carefully chosen.

Python code:

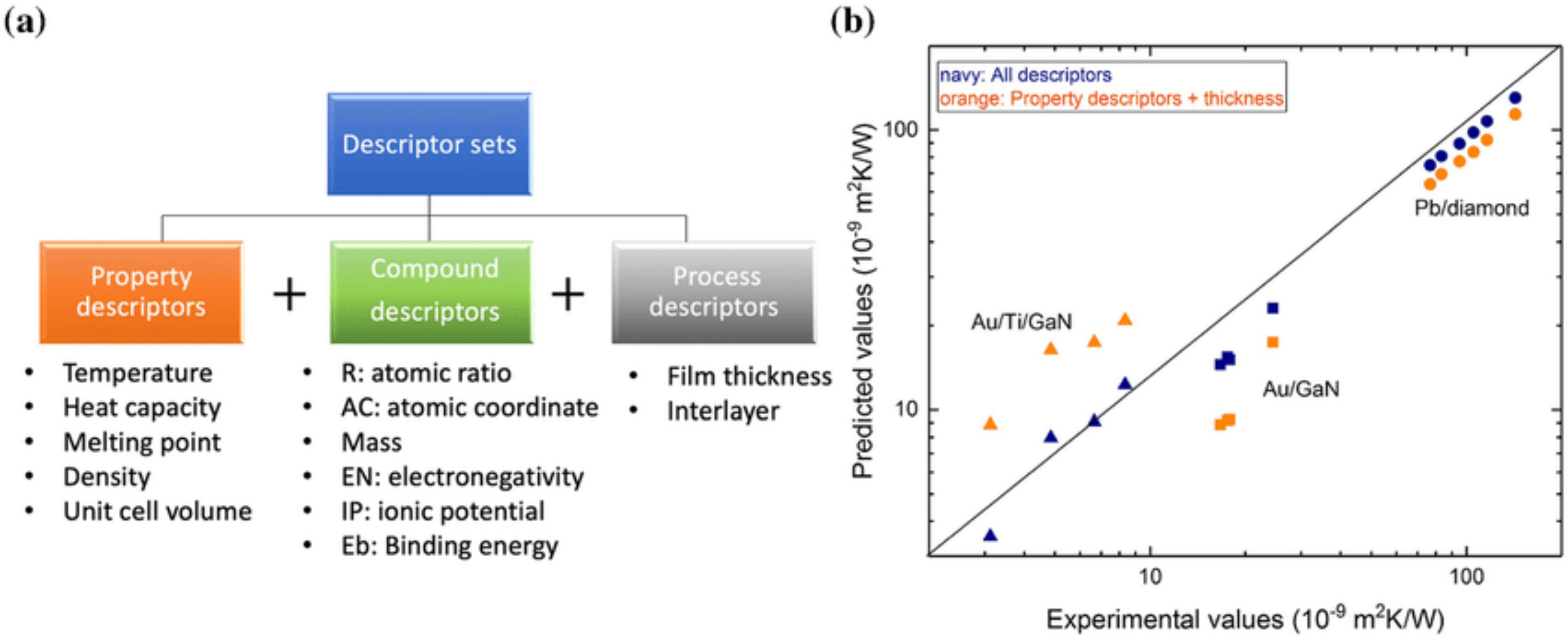
```
from sklearn.feature_selection import SelectKBest, mutual_info_regression, RFE
from sklearn.linear_model import LassoCV

# SelectKBest
X_new = SelectKBest(mutual_info_regression, k=5).fit_transform(X, y)

# RFE
estimator = LassoCV()
selector = RFE(estimator, 5, step=1)
selector = selector.fit(X, y)
X_new = X[:, selector.support_]
```

Descriptors and fingerprints

Descriptors and fingerprints are terms used in materials science to refer to representations of the properties of materials that can be used as inputs to machine learning models.



Descriptor sets

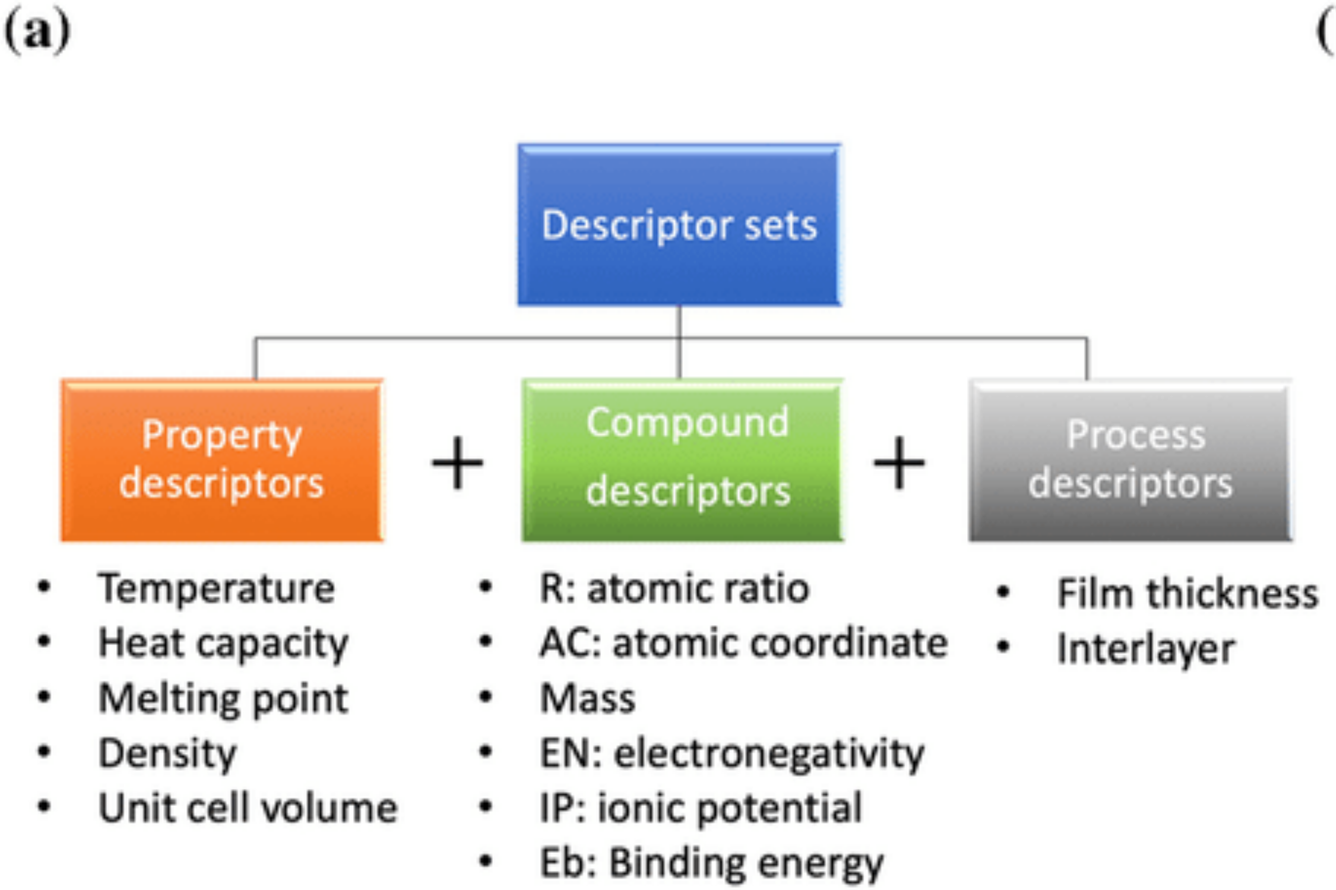
Property descriptors + Compound descriptors + Process descriptors

- Temperature
- Heat capacity
- Melting point
- Density
- Unit cell volume

- R: atomic ratio
- AC: atomic coordinate
- Mass
- EN: electronegativity
- IP: ionic potential
- Eb: Binding energy

- Film thickness
- Interlayer

(a)



Descriptor sets

Property descriptors + Compound descriptors + Process descriptors

Property descriptors:

- Temperature
- Heat capacity
- Melting point
- Density
- Unit cell volume

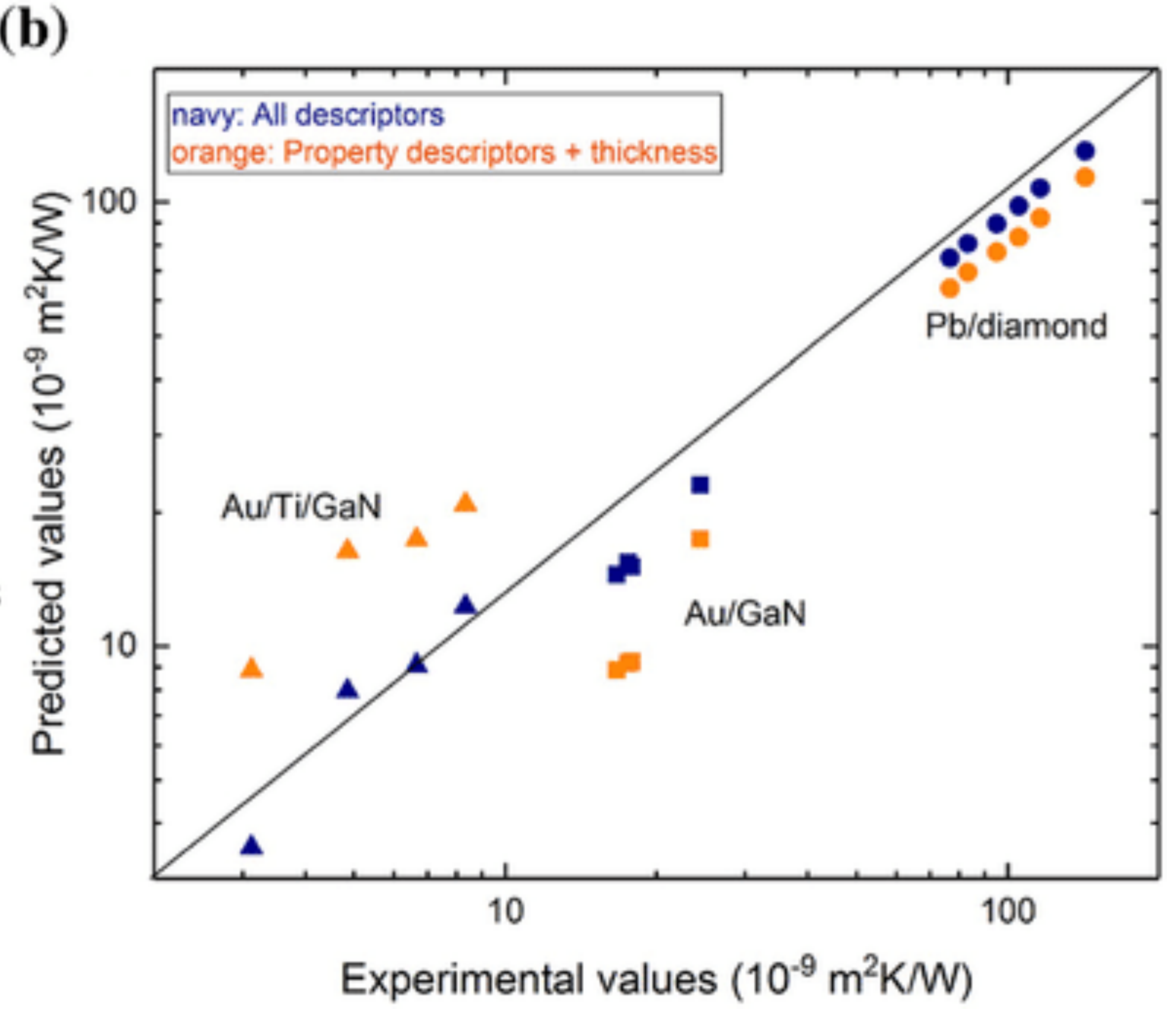
Compound descriptors:

- R: atomic ratio
- AC: atomic coordinate
- Mass
- EN: electronegativity
- IP: ionic potential
- Eb: Binding energy

Process descriptors:

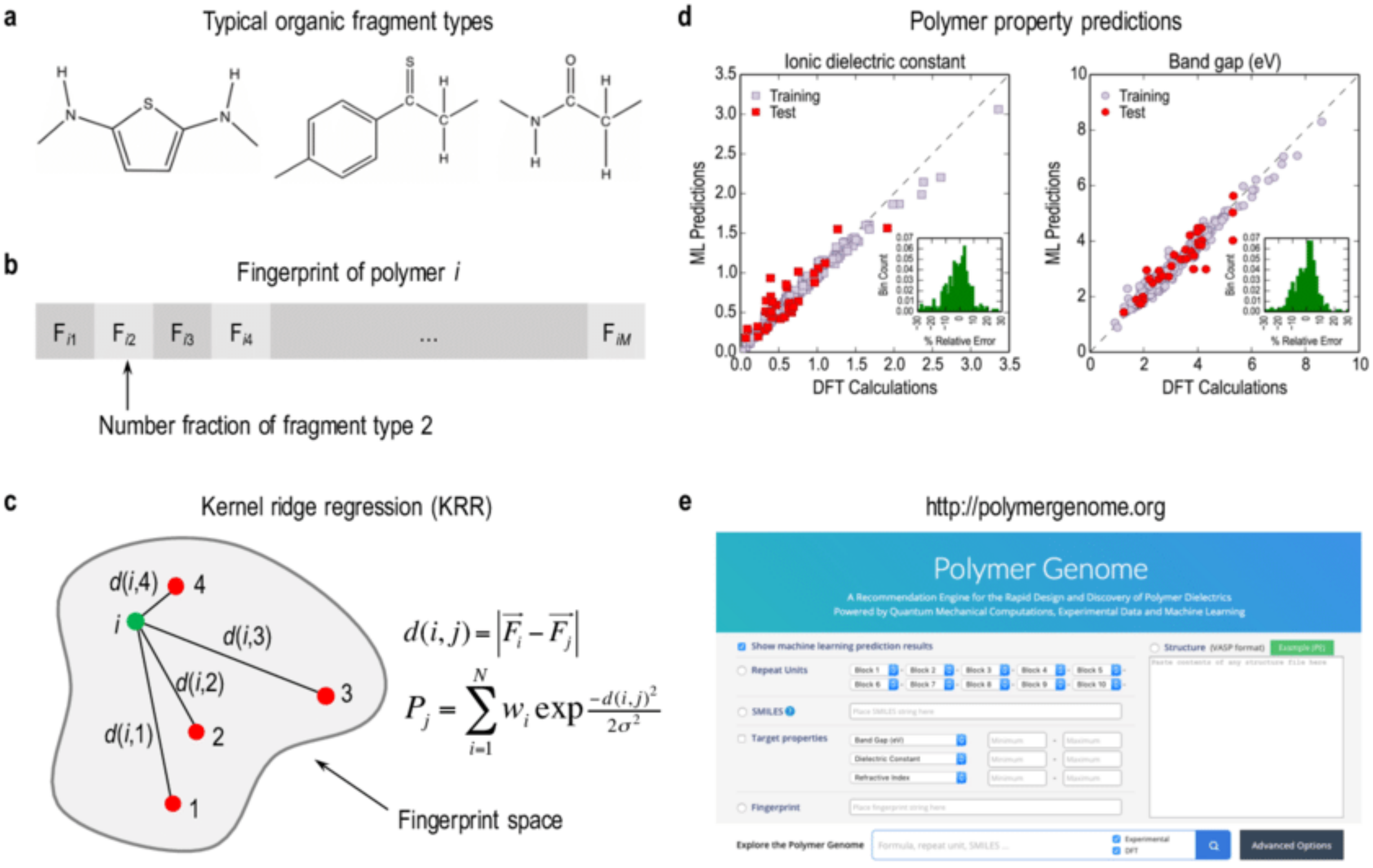
- Film thickness
- Interlayer

(b)



- Descriptors are numerical or categorical properties that describe the materials, such as composition, crystal structure, or electronic properties.

- Fingerprints are a set of binary or numerical features that capture the local or global features of the materials, such as local coordination or chemical environment.



$$d(i,j)=\left|\vec{F}_i-\vec{F}_j\right|$$

$$P_j=\sum_{i=1}^{N} w_i \exp\frac{-d(i,j)^2}{2\sigma^2}$$

Polymer Genome

A Recommendation Engine for the Rapid Design and Discovery of Polymer Dielectrics
Powered by Quantum Mechanical Computations, Experimental Data and Machine Learning

☑ Show machine learning prediction results

○ Repeat Units: Block 1 - Block 2 - Block 3 - Block 4 - Block 5 - Block 6 - Block 7 - Block 8 - Block 9 - Block 10 -

○ SMILES: Place SMILES string here

□ Target properties: Band Gap (eV) Minimum - Maximum; Dielectric Constant Minimum - Maximum; Refractive Index Minimum - Maximum

○ Fingerprint: Place fingerprint string here

○ Structure (VASP format) Example (PI)

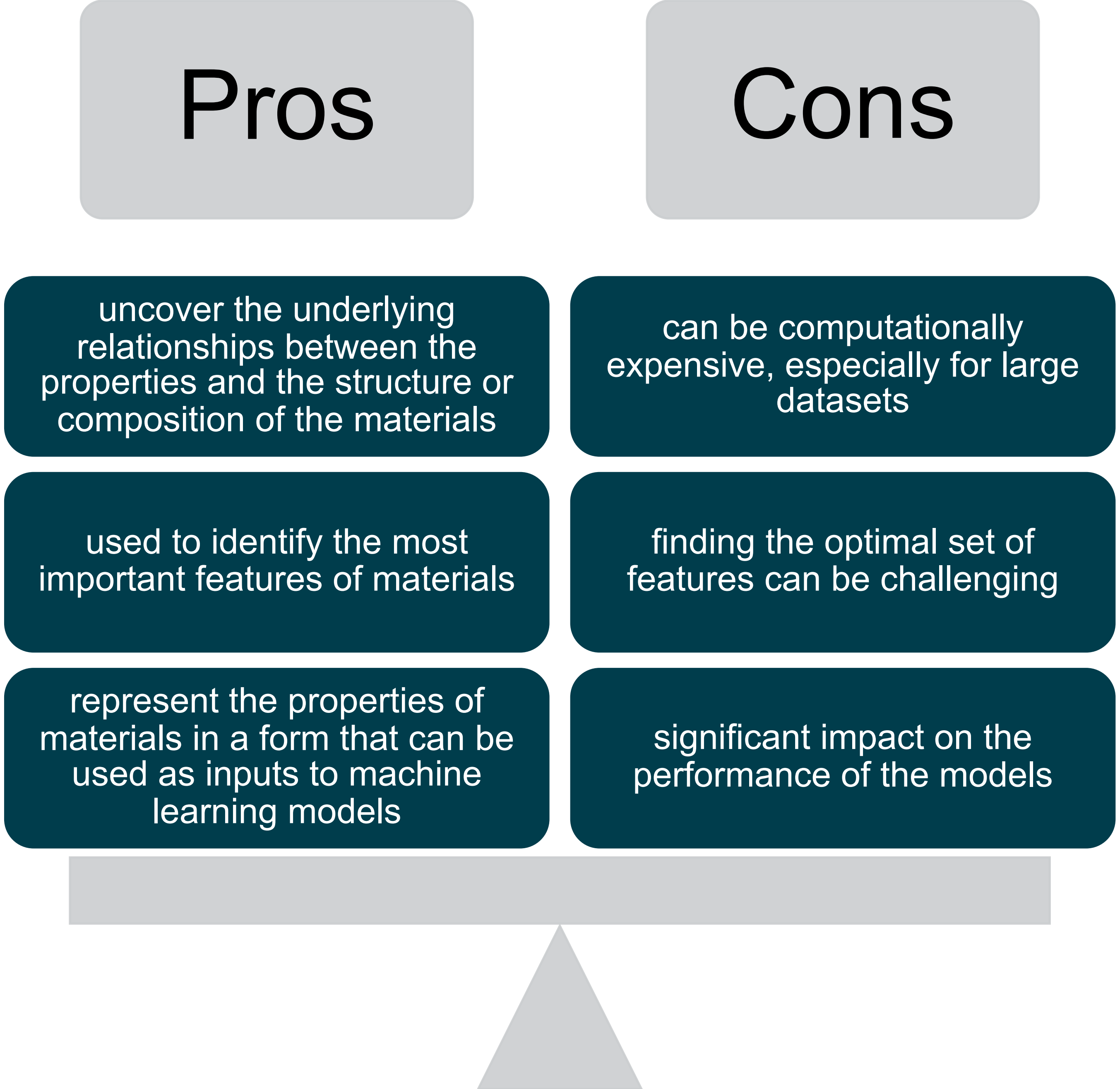
Paste contents of any structure file here

Explore the Polymer Genome: Formula, repeat unit, SMILES ... ☑ Experimental ☑ DFT

Advanced Options

References:

- Ward, L., Agrawal, A., Chollet, A., & Persson, K. A. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials, 2(1), 1-13.
- Rupp, M., Tkatchenko, A., Muller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters, 108(5), 058301.



Pros	Cons
uncover the underlying relationships between the properties and the structure or composition of the materials	can be computationally expensive, especially for large datasets
used to identify the most important features of materials	finding the optimal set of features can be challenging
represent the properties of materials in a form that can be used as inputs to machine learning models	significant impact on the performance of the models

Descriptors

the concentration of each element in the material

the crystal structure

the lattice parameters (calculated from experimental measurements)

Fingerprints

the local coordination environment of each atom in the material (calculated from molecular dynamics simulations or quantum chemical calculations)

Difference between overfitting and underfitting

❖ Overfitting and underfitting are common problems in machine learning that occur when a model is not properly trained or is too complex.

❖ Overfitting and underfitting occur when a model is trained on a specific dataset and then evaluated on a different dataset.

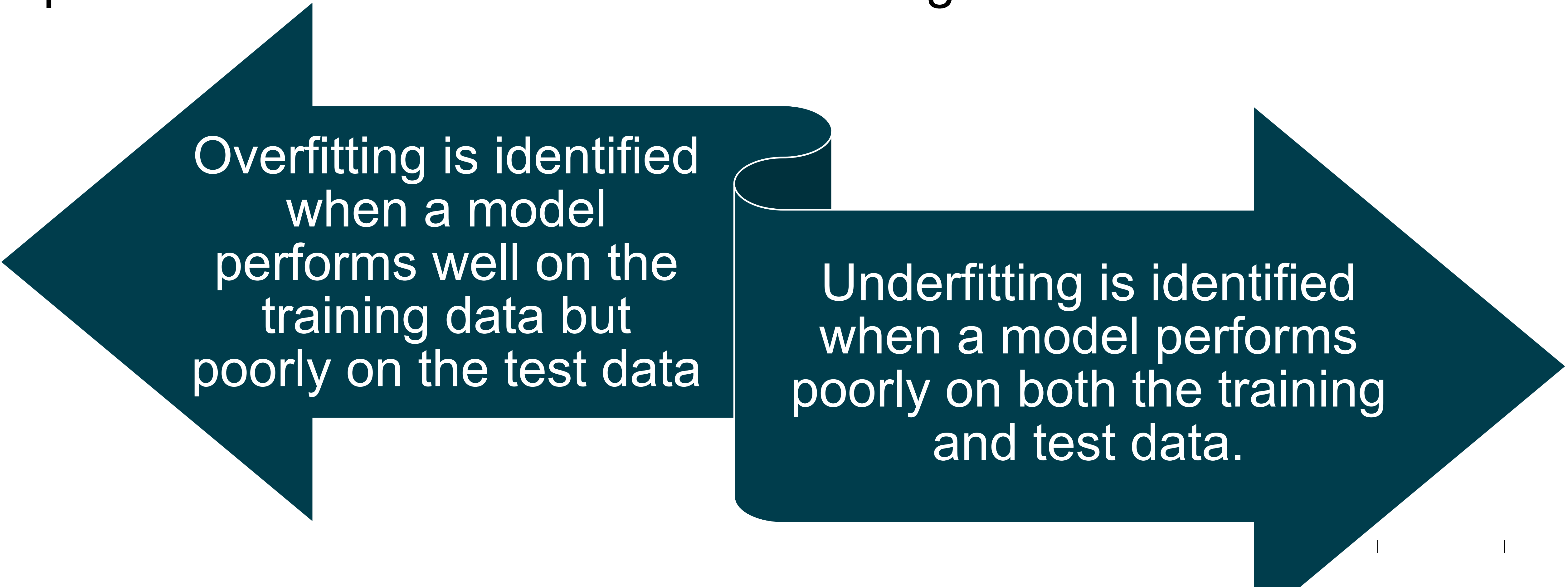
Difference between overfitting and underfitting

- ❑ Overfitting occurs when a model is too complex and is able to fit the noise in the training data.

- ❑ Underfitting occurs when a model is not complex enough and is unable to capture the underlying pattern in the data.

Difference between overfitting and underfitting

Overfitting and underfitting can be identified by comparing the performance of a model on the training data and the test data.



Overfitting is identified when a model performs well on the training data but poorly on the test data

Underfitting is identified when a model performs poorly on both the training and test data.

✓Understanding overfitting and underfitting can help you to **improve the performance** of your machine learning model.

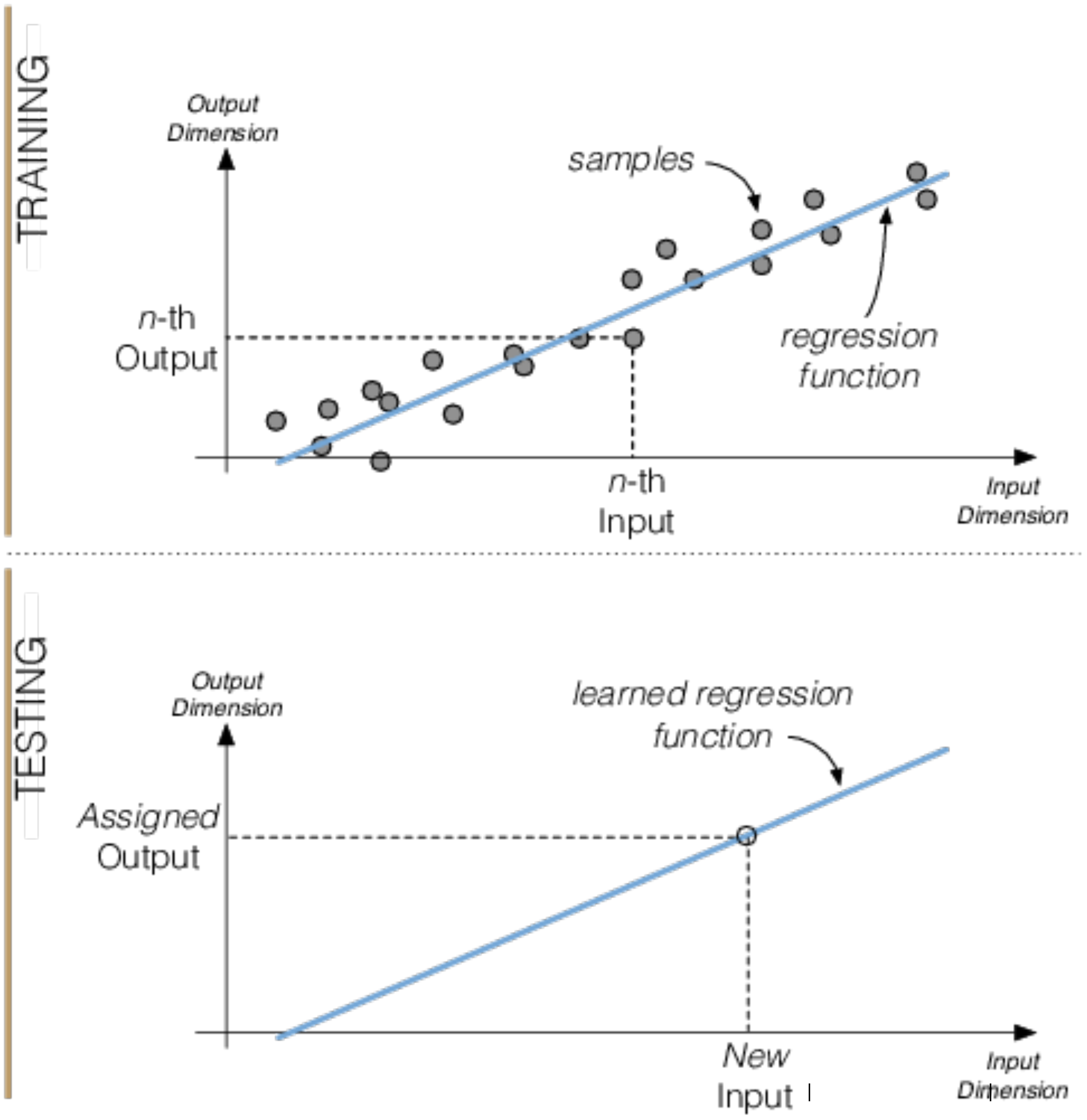
✓Overfitting and underfitting can lead to **poor performance** of a model and can be hard to detect.

References

- Overfitting and Underfitting
  - https://towardsdatascience.com/underfitting-and-overfitting-in-machine-learning-d4fa802aabae
- Bias-Variance Trade-off
  - https://towardsdatascience.com/bias-variance-trade-off-f2e03821f40b
- Regularization techniques in machine learning
  - https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

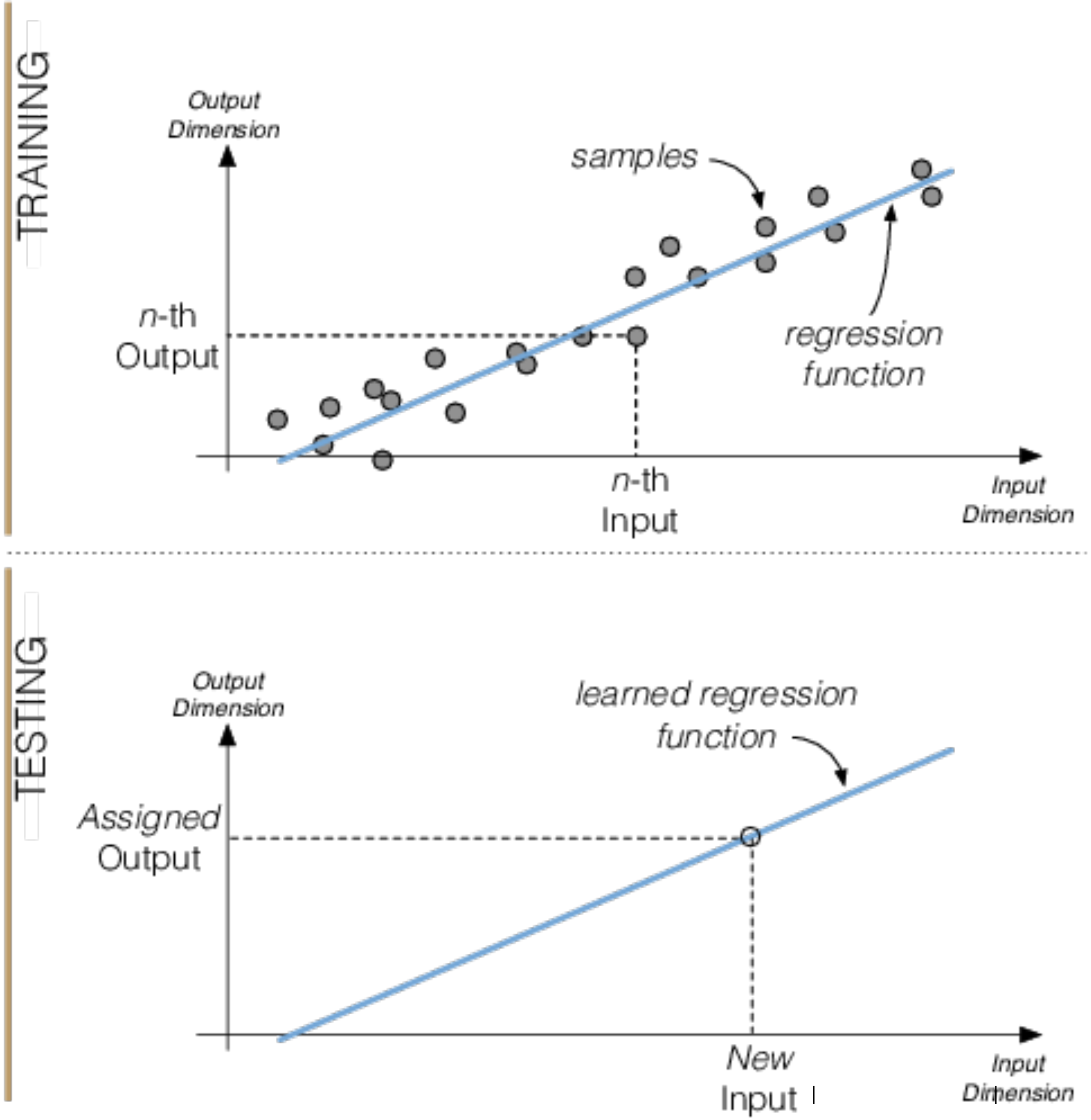
1.Definition of regression

Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.

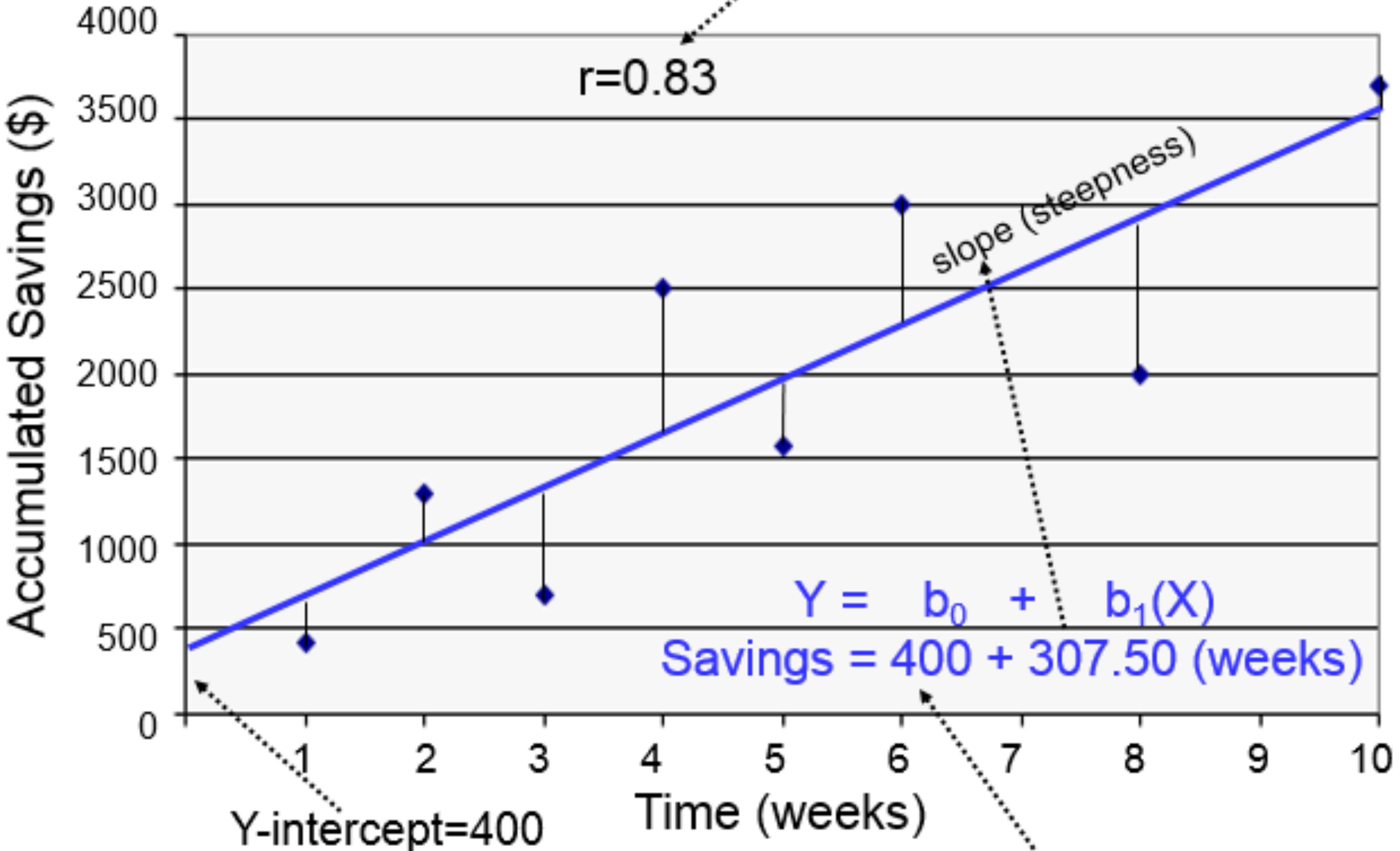


The goal of regression is to find the mathematical equation that best fits the data, in order to make predictions about future values.

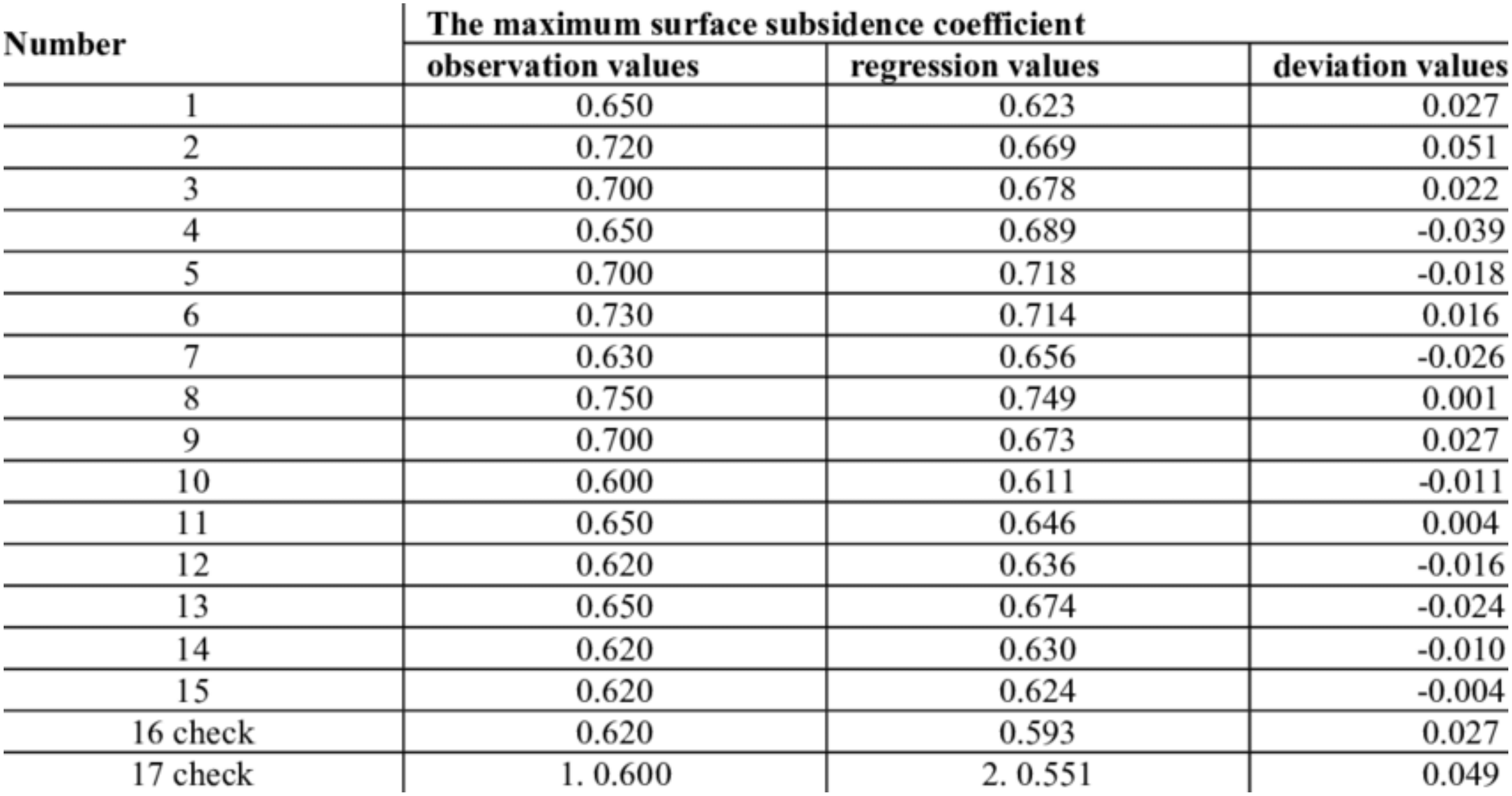
$$Y = a X1 + b X2 +..+ n Xn$$



Correlation coefficient (r): a measure of how closely observation points conform to the regression line



Regression: equation describing the relationship



Number	The maximum surface subsidence coefficient		
	observation values	regression values	deviation values
1	0.650	0.623	0.027
2	0.720	0.669	0.051
3	0.700	0.678	0.022
4	0.650	0.689	-0.039
5	0.700	0.718	-0.018
6	0.730	0.714	0.016
7	0.630	0.656	-0.026
8	0.750	0.749	0.001
9	0.700	0.673	0.027
10	0.600	0.611	-0.011
11	0.650	0.646	0.004
12	0.620	0.636	-0.016
13	0.650	0.674	-0.024
14	0.620	0.630	-0.010
15	0.620	0.624	-0.004
16 check	0.620	0.593	0.027
17 check	1. 0.600	2. 0.551	0.049

1.Types of Regression

There are many different types of regression techniques, including linear regression, polynomial regression, multiple regression, and non-linear regression.

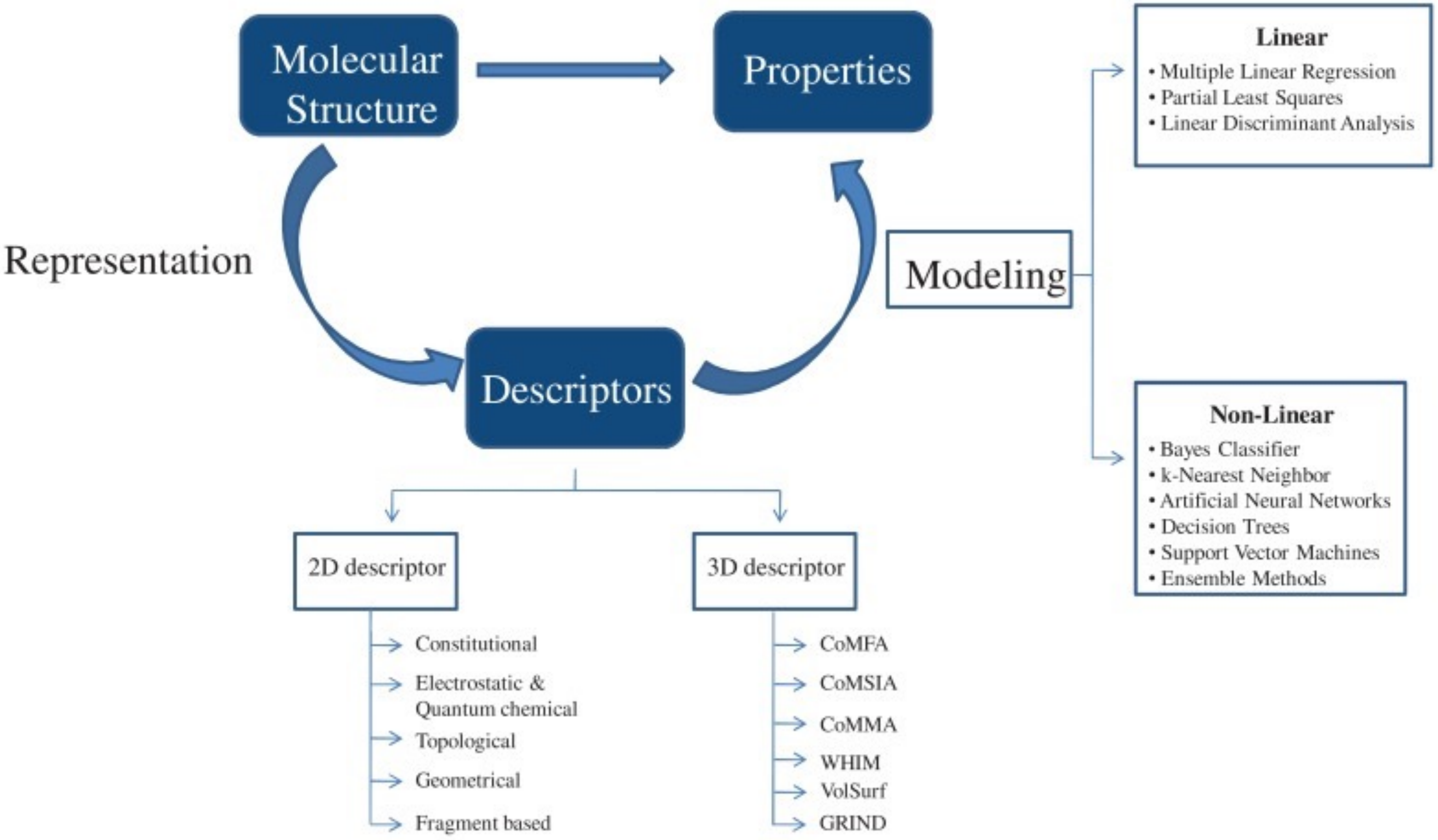
Each type of regression is designed to model different types of relationships between the dependent and independent variables.

Applications of Regression Techniques in Materials Science

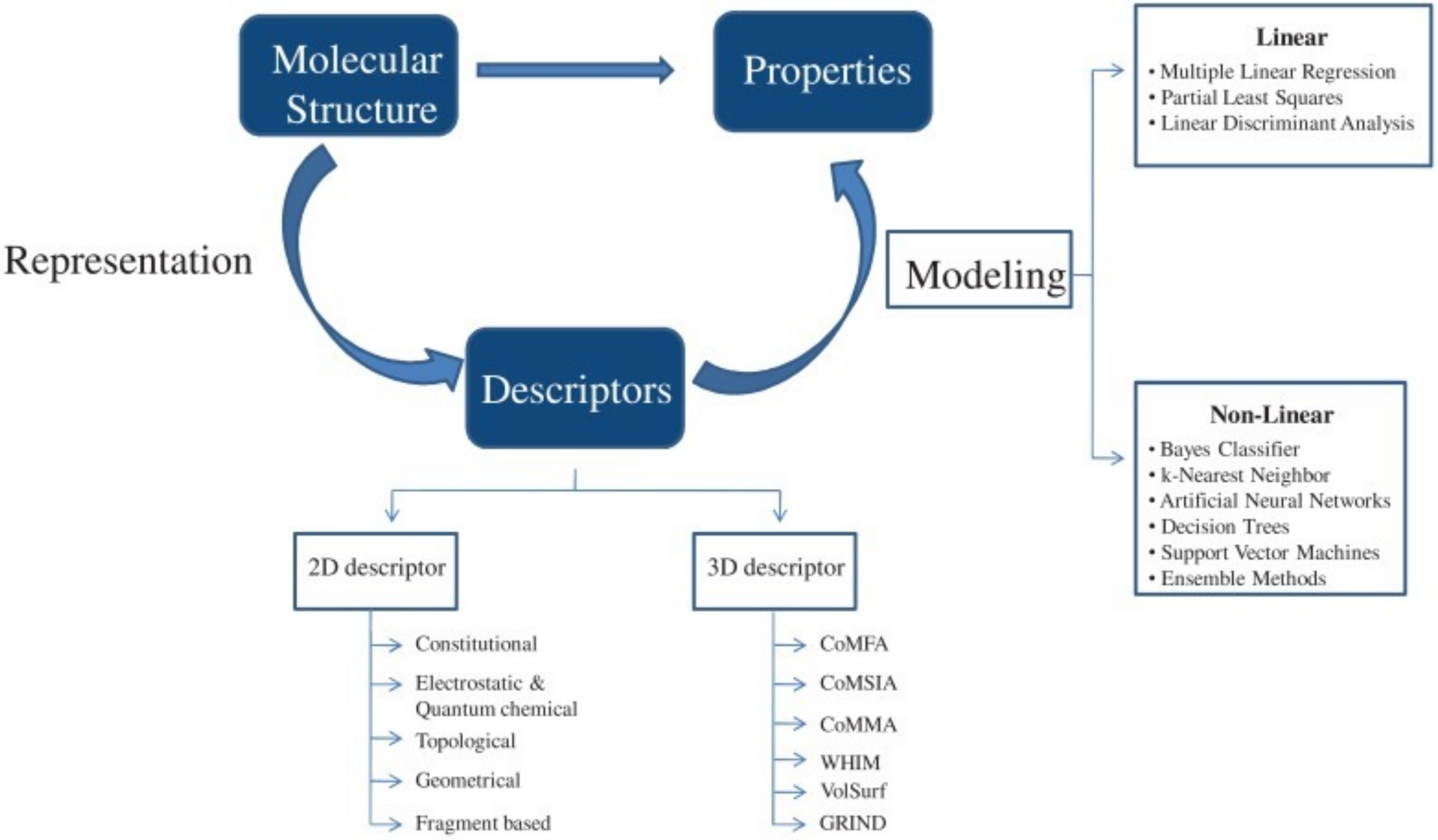
- predict the properties of materials, such as their thermal conductivity, electrical conductivity, and mechanical strength.
- make predictions about the properties of new materials or materials under different conditions or the conditions that lead to the best materials for a given application.
- screen large numbers of potential materials for specific properties.
- quickly identify the materials with the best potential for a given application, without the need for expensive and time-consuming experiments.

QSPR (Quantitative Structure-Property Relationships)

QSPR is a subfield of computational materials science that uses statistical and mathematical techniques to model the relationship between the molecular structure of a material and its properties.



- In QSPR, the molecular structure is represented by a set of descriptors, which are numerical representations of the molecular structure.
- These descriptors are then used as independent variables in a regression model to predict the properties of the material.



Useful metrics in regression

- Evaluating the performance of a model is an important step in machine learning as it allows to have an idea of how well the model is able to generalise to new data.

- Some of the most commonly used metrics are accuracy, precision, recall, F1 score, and R-squared.

Useful metrics in regression

Mean Squared Error (MSE): MSE is calculated by summing up the squares of the differences between the predicted and actual values for each data point, dividing by the number of data points, and taking the square root.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE	= mean squared error
n	= number of data points
Y_i	= observed values
$\hat{Y}_i$	= predicted values

Useful metrics in regression

Mean Absolute Error (MAE): MAE is calculated by summing up the absolute differences between the predicted and actual values for each data point, dividing by the number of data points.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

MAE = mean absolute error

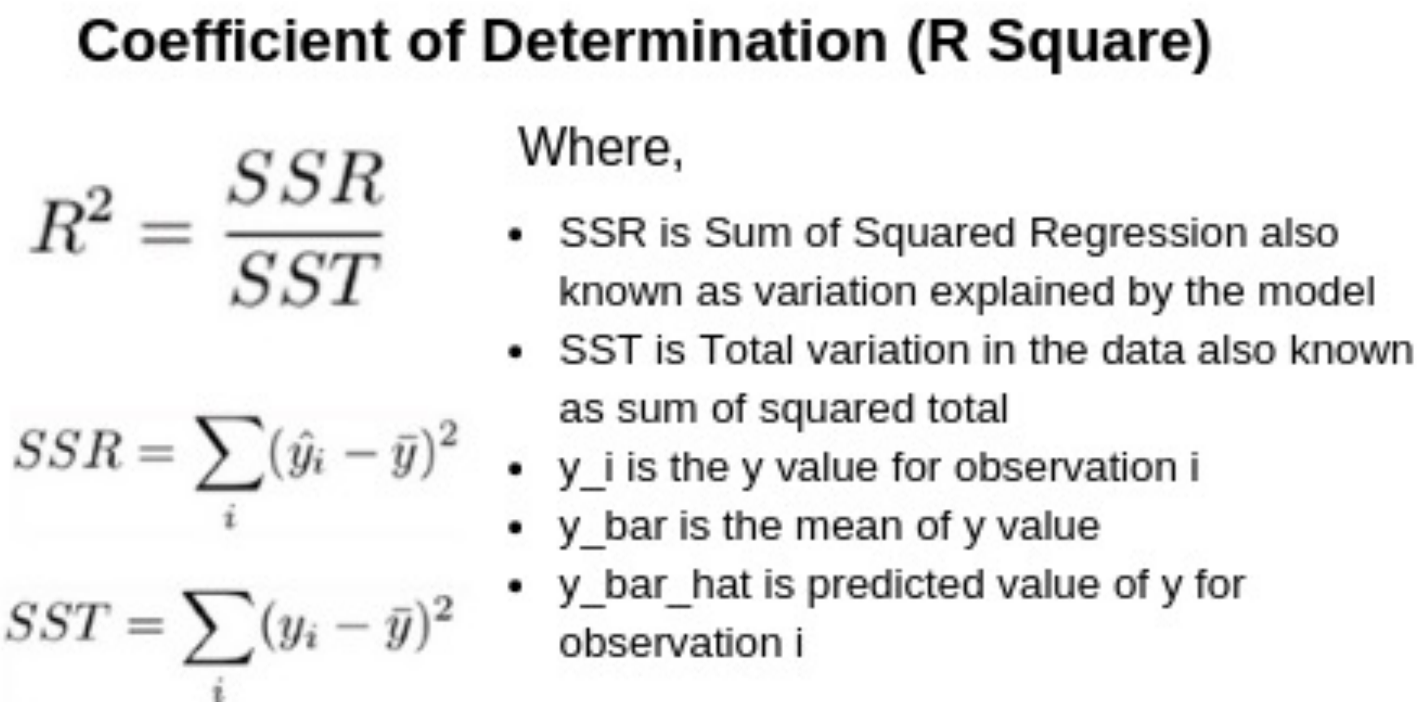
y_i = prediction

x_i = true value

n = total number of data points

Useful metrics in regression

R-squared: R-squared is calculated by dividing the explained variance of the regression model by the total variance in the dependent variable.



Coefficient of Determination (R Square)

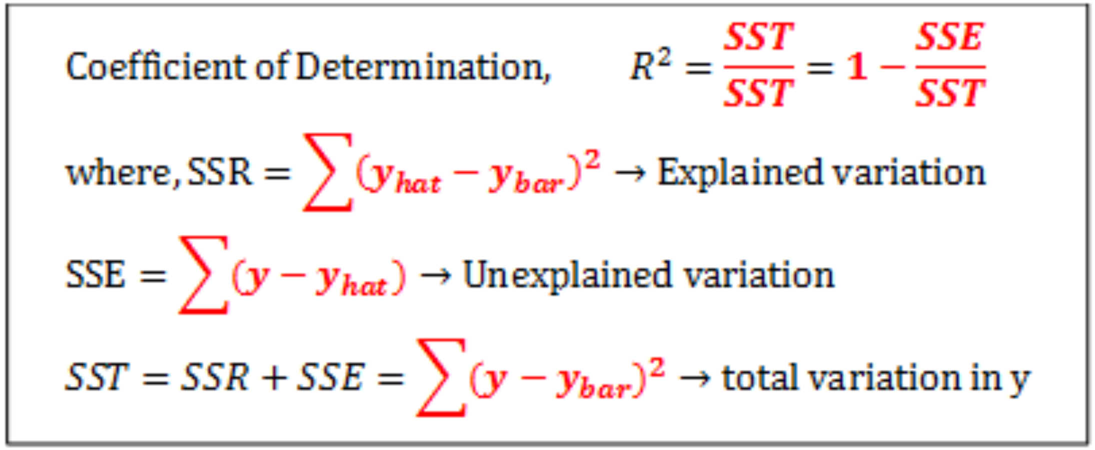
$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

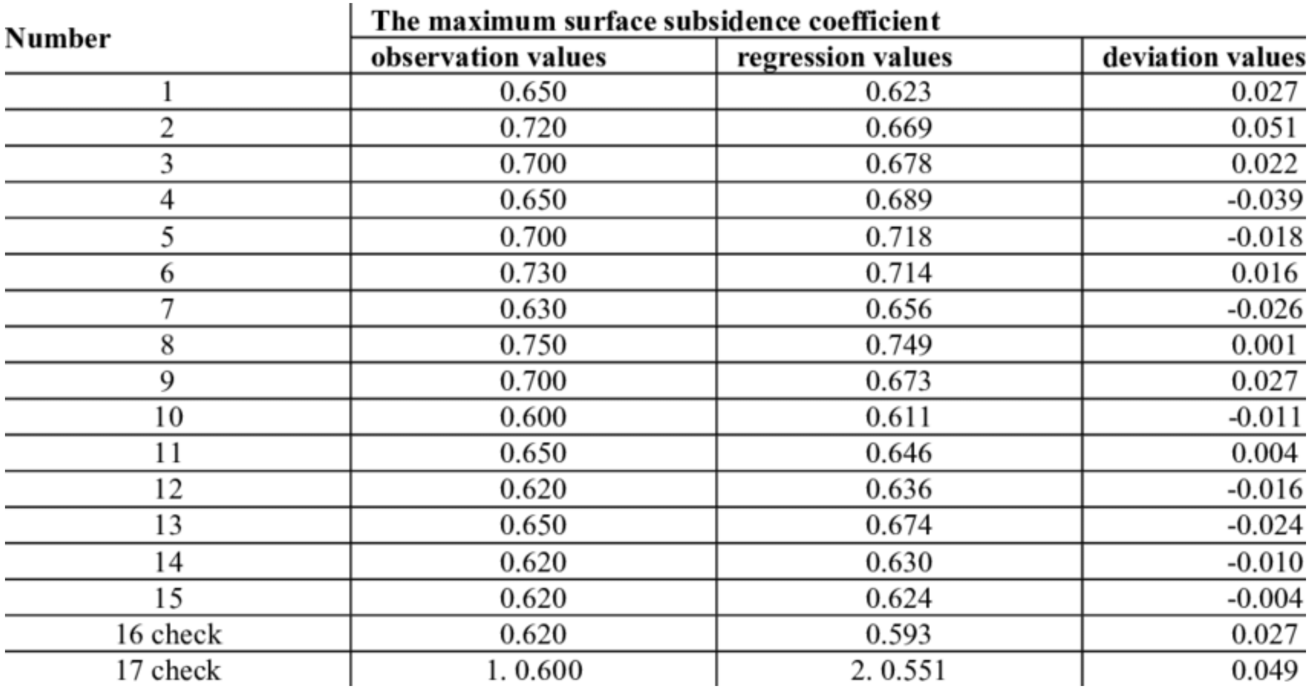


Coefficient of Determination, $R^2 = \frac{SST}{SST} = 1 - \frac{SSE}{SST}$

where, SSR $= \sum (y_{hat} - y_{bar})^2$ → Explained variation

SSE $= \sum (y - y_{hat})$ → Unexplained variation

$SST = SSR + SSE = \sum (y - y_{bar})^2$ → total variation in y



Number	**The maximum surface subsidence coefficient**		
	observation values	**regression values**	**deviation values**
1	0.650	0.623	0.027
2	0.720	0.669	0.051
3	0.700	0.678	0.022
4	0.650	0.689	-0.039
5	0.700	0.718	-0.018
6	0.730	0.714	0.016
7	0.630	0.656	-0.026
8	0.750	0.749	0.001
9	0.700	0.673	0.027
10	0.600	0.611	-0.011
11	0.650	0.646	0.004
12	0.620	0.636	-0.016
13	0.650	0.674	-0.024
14	0.620	0.630	-0.010
15	0.620	0.624	-0.004
16 check	0.620	0.593	0.027
17 check	1. 0.600	2. 0.551	0.049

Pros:

- are simple and easy to understand metrics.
- provide a comprehensive evaluation of the performance of a model, with different aspects of the model's performance.

Cons:

- may have different interpretations depending on the specific problem and the domain.

Python Code:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

from sklearn.metrics import r2_score

# calculate accuracy

accuracy = accuracy_score(y_test, predictions)

# calculate precision

precision = precision_score(y_test, predictions)

# calculate recall

recall = recall_score(y_test, predictions)

# calculate f1 score

f1 = f1_score(y_test, predictions)

# calculate R-squared (for regression)

r2 = r2_score(y_test, predictions)
```

References:

- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
- Bishop, C. M. (2006). Pattern recognition and machine learning (pp. 104-107). Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning (pp. 89-92). Springer.

Practical

Preprocessing

Features selection

Overfitting vs Underfitting

Linear Regression

Knime demonstration

Logistic regression