MODELADO ESPACIO-TEMPORAL DE PROCESOS GEOAMBIENTALES EN R

Marcos Rodrigues Mimbrero

rmarcos@eagrof.udl.cat

Universitat de Lleida

What's a model?

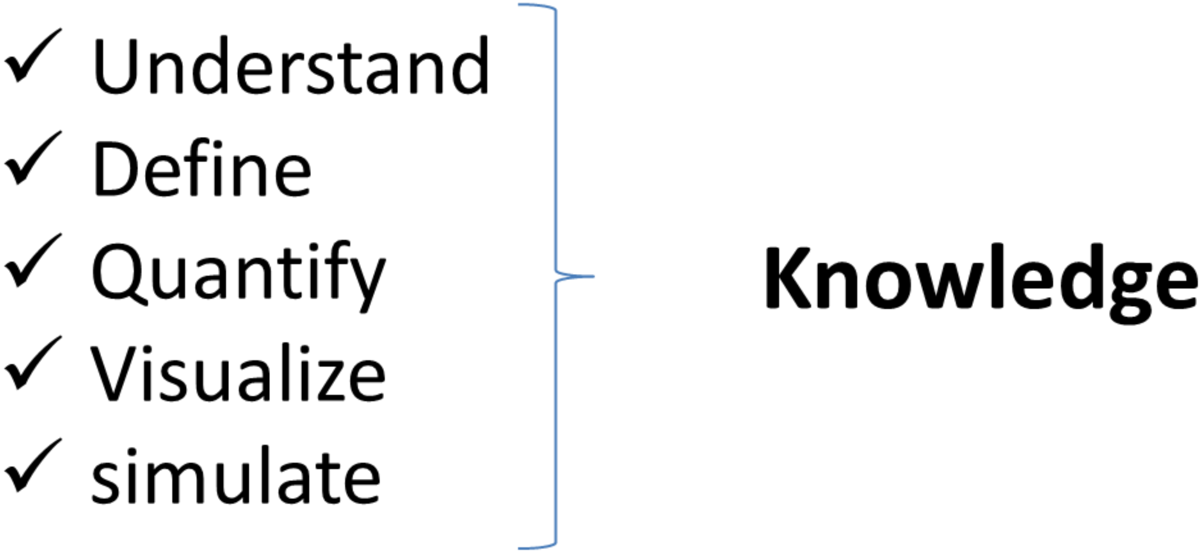
In science, a model is a **representation** of an **idea**, an object or even a **process** or a **system** that is used to **describe** and **explain** phenomena that cannot be experienced directly.

Models are central to what **scientists** do, both in their research as well as when **communicating their explanations.**

Taken from https://www.sciencelearn.org.nz/resources/575-scientific-modelling

What’s a model?

Scientific modeling

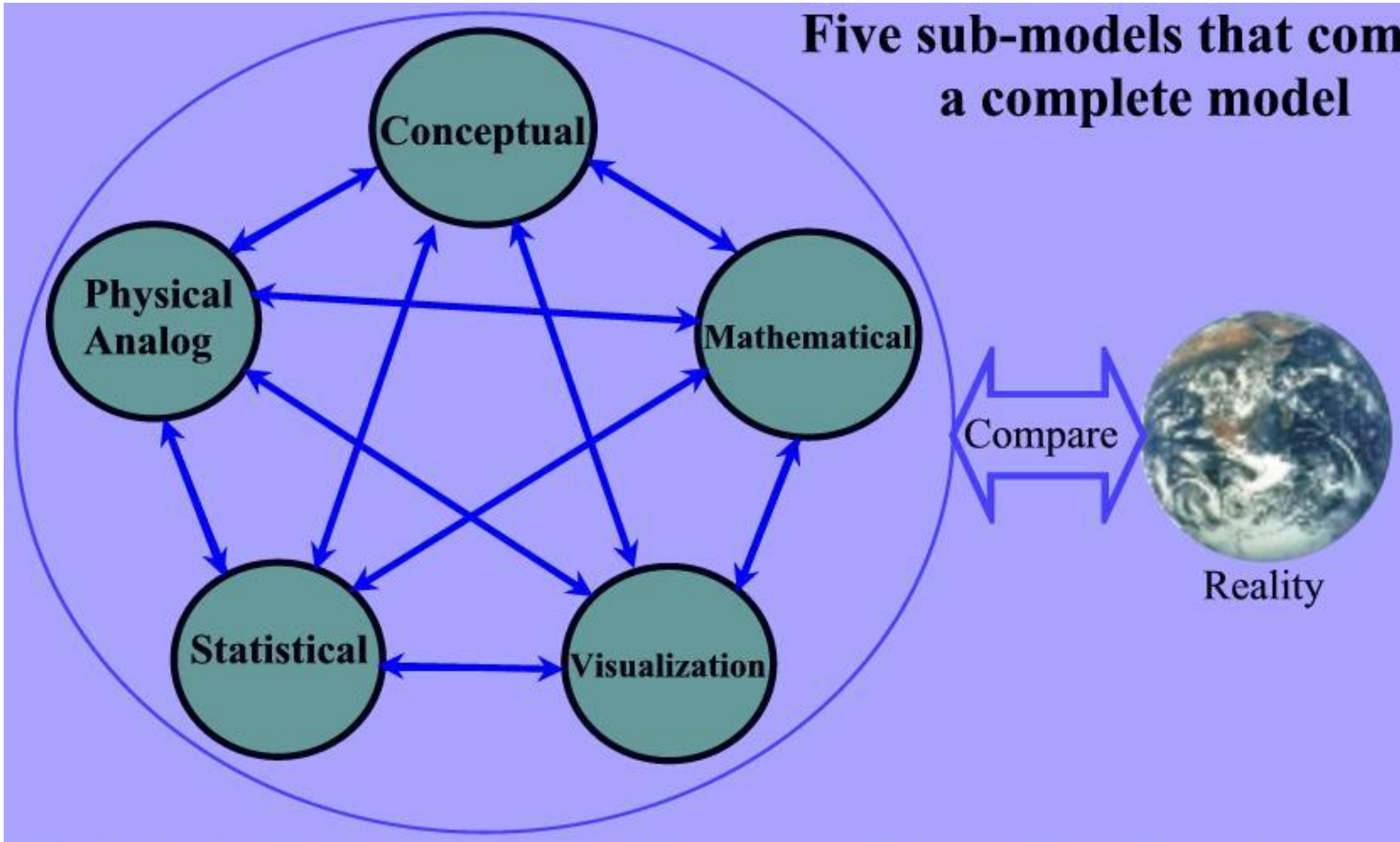


- ✓ Understand
- ✓ Define
- ✓ Quantify
- ✓ Visualize
- ✓ simulate

Knowledge

Types of models

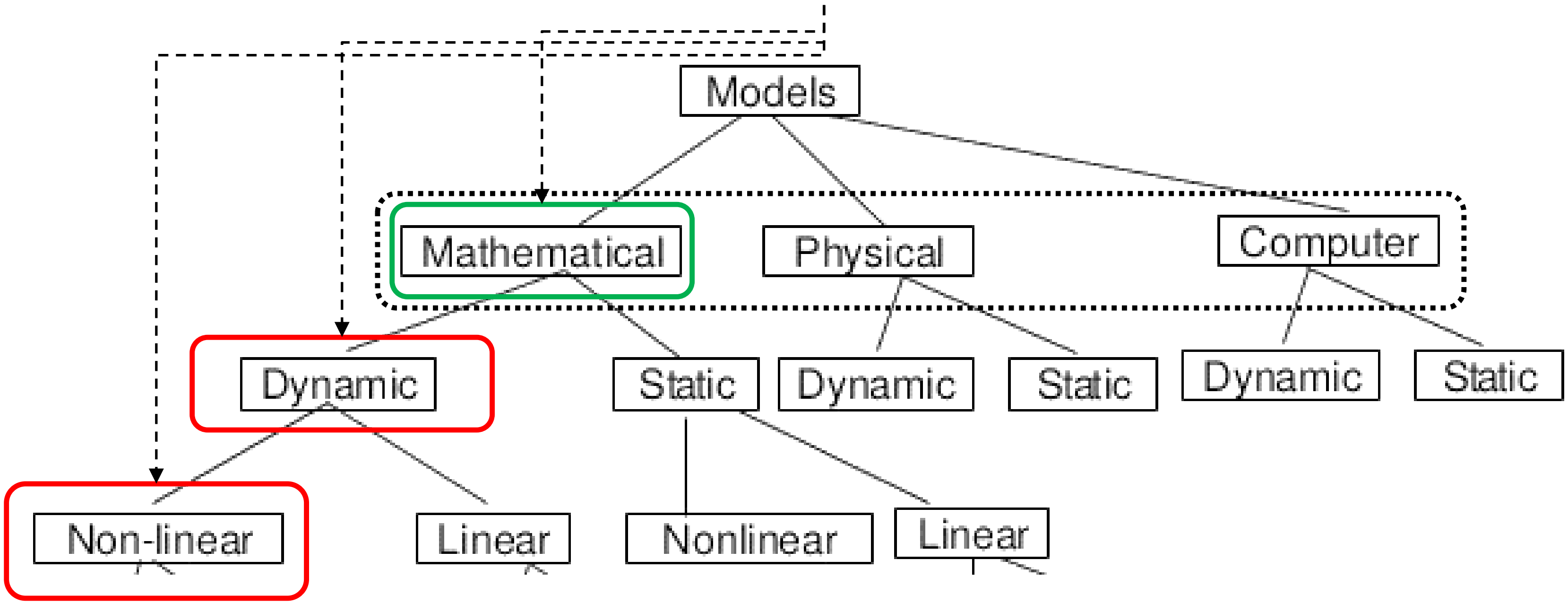
Five sub-models that comprise a complete model



During development information is continuously exchanged between the sub-models and the real system to optimize model performance.

Types of models

Procesos geo-ambientales



Types of models

- ✓ **Explanatory models:** they aim at understand which factors trigger the occurrence of an event and vice versa.

- ✓ **Predictive models:** extend explanatory models to forecast the probability of occurrence of a given type of event into space and/or time.

Types of models

What’s the difference? What does imply going from explanatory to predictive?

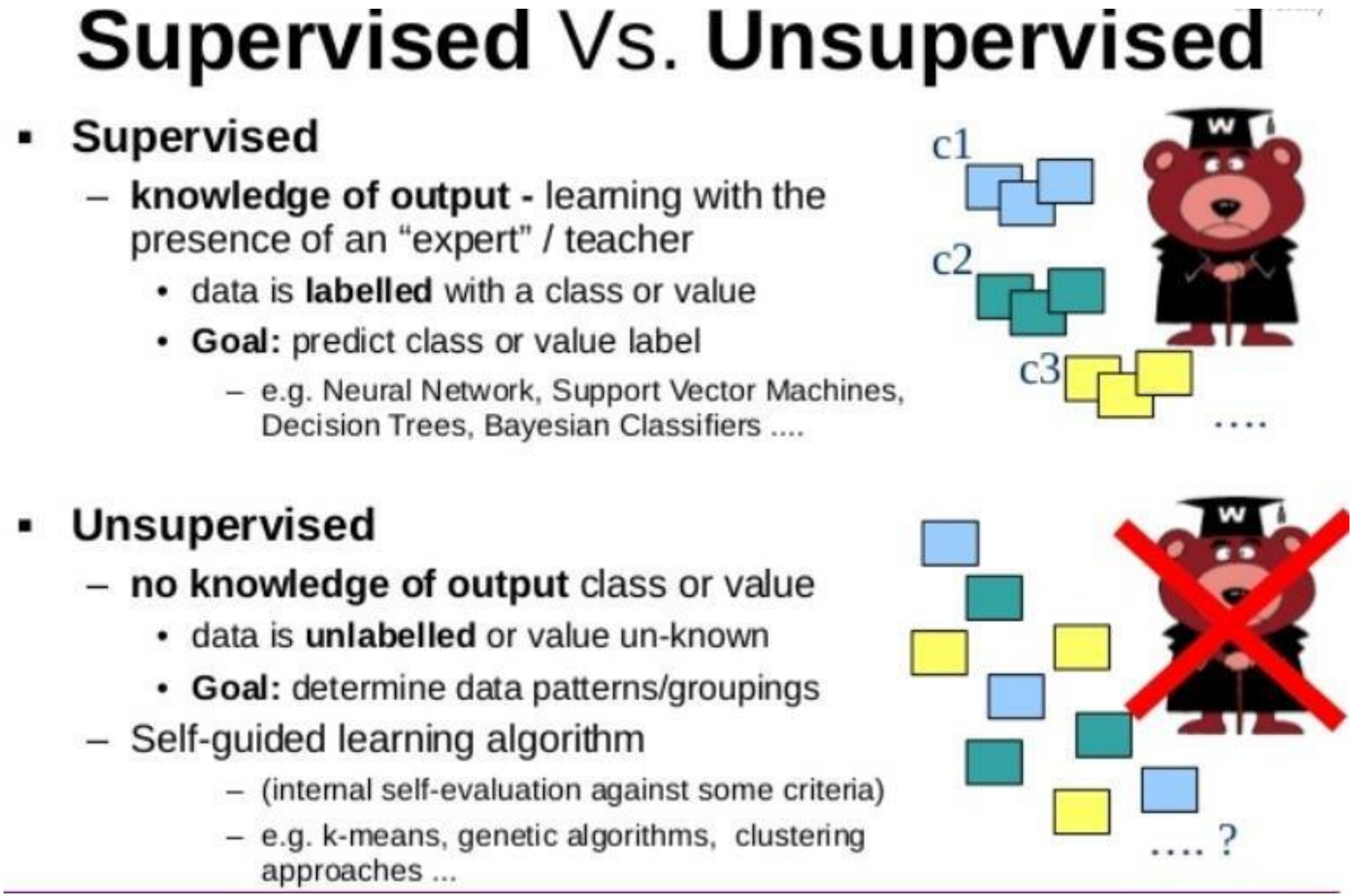
Predictive models need an independent test sample to test their predictive performance, which is not necessarily the case of the explanatory ones. That means we must implement a **validation procedure**.

Types of models

- ✓ **Structural models:** they model or provide an ‘static’ picture of the phenomenon.

- ✓ **Dynamic models:** they focus on the temporal or dynamic evolution of the phenomenon, i.e., provide a different forecast based on time or in the change over time of a certain predictor.

- ✓ **Empirical/statistical methods** leverage **historical data** or **observations.** Two main approaches:

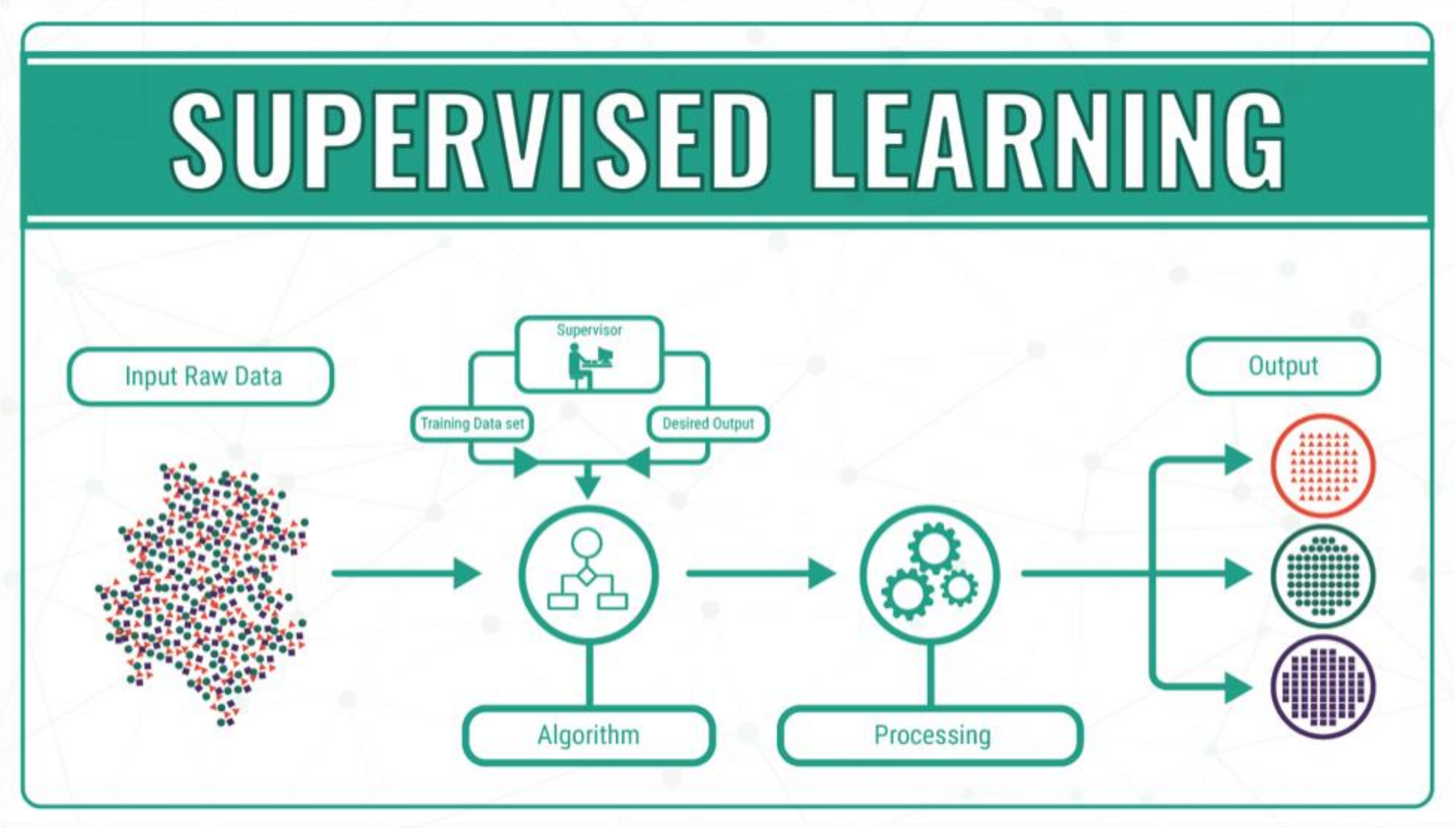


Supervised Vs. Unsupervised

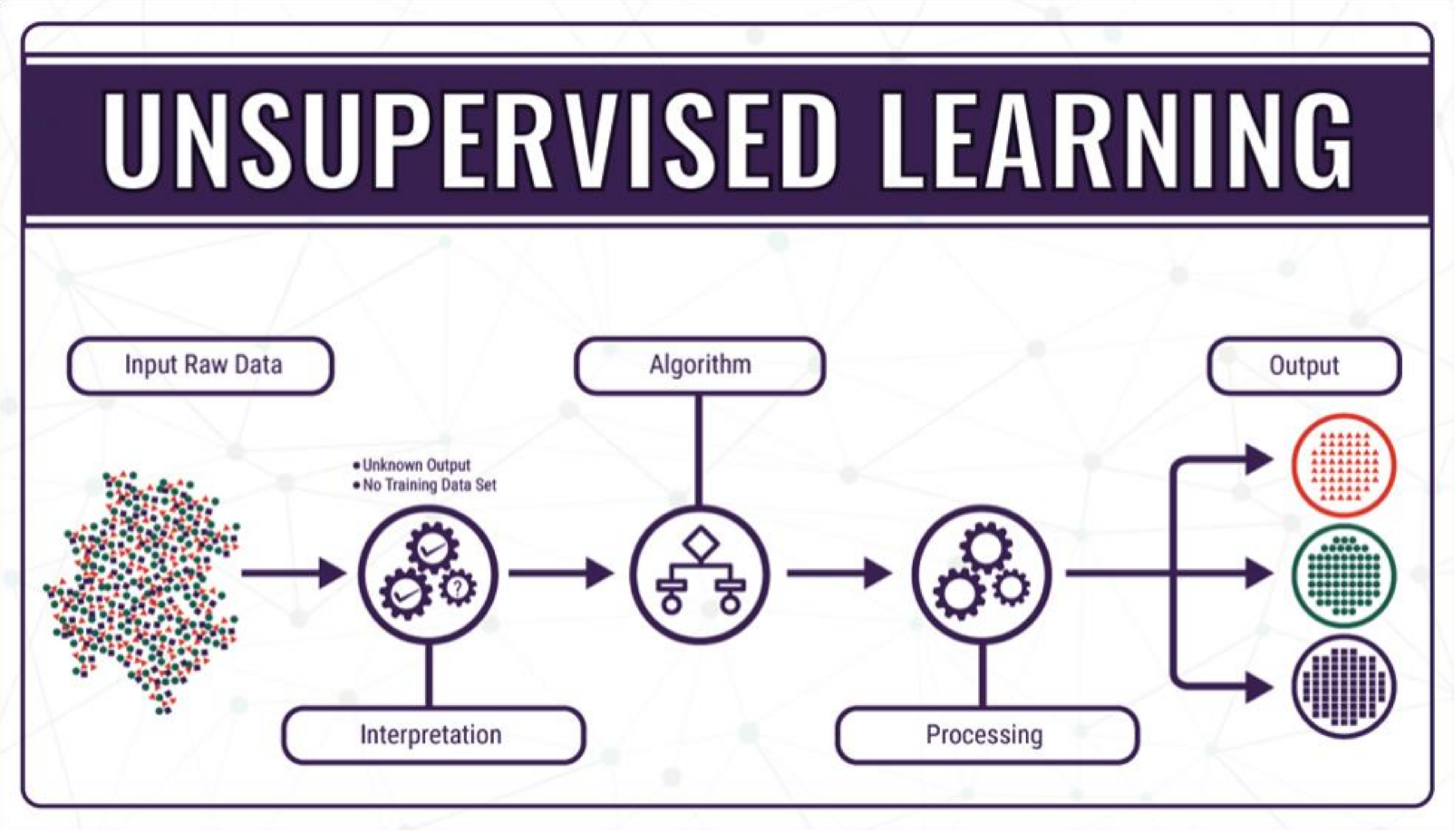
- **Supervised**
  - **knowledge of output -** learning with the presence of an "expert" / teacher
    - data is **labelled** with a class or value
    - **Goal:** predict class or value label
      - e.g. Neural Network, Support Vector Machines, Decision Trees, Bayesian Classifiers ....
- **Unsupervised**
  - **no knowledge of output** class or value
    - data is **unlabelled** or value un-known
    - **Goal:** determine data patterns/groupings
  - Self-guided learning algorithm
    - (internal self-evaluation against some criteria)
    - e.g. k-means, genetic algorithms, clustering approaches ...

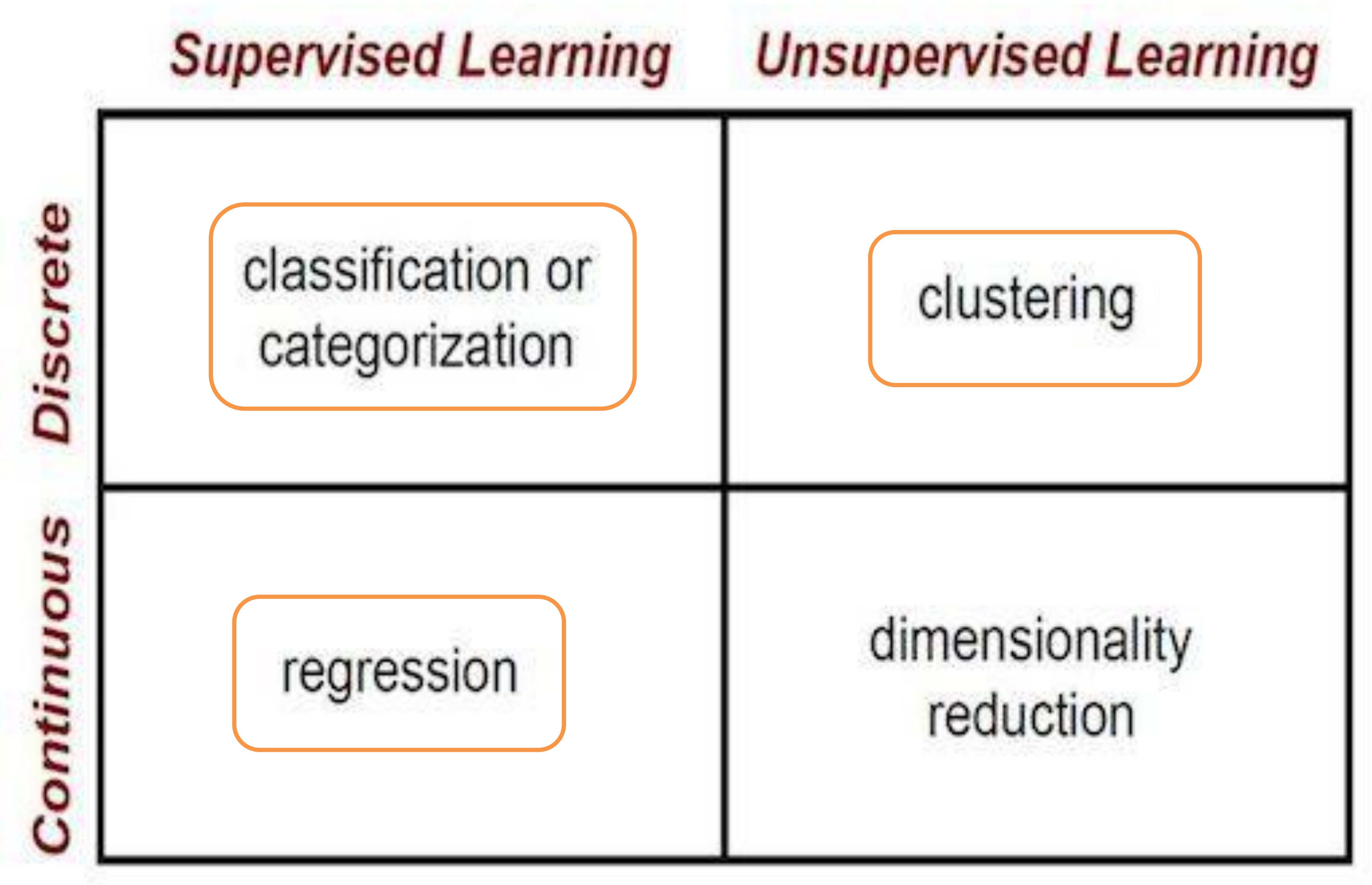
Introduction to Machine Learning by Ahmad F. Al Musawi

In Supervised learning, you train the machine using data which is well **"labeled."** It means some data is already tagged with the correct answer.

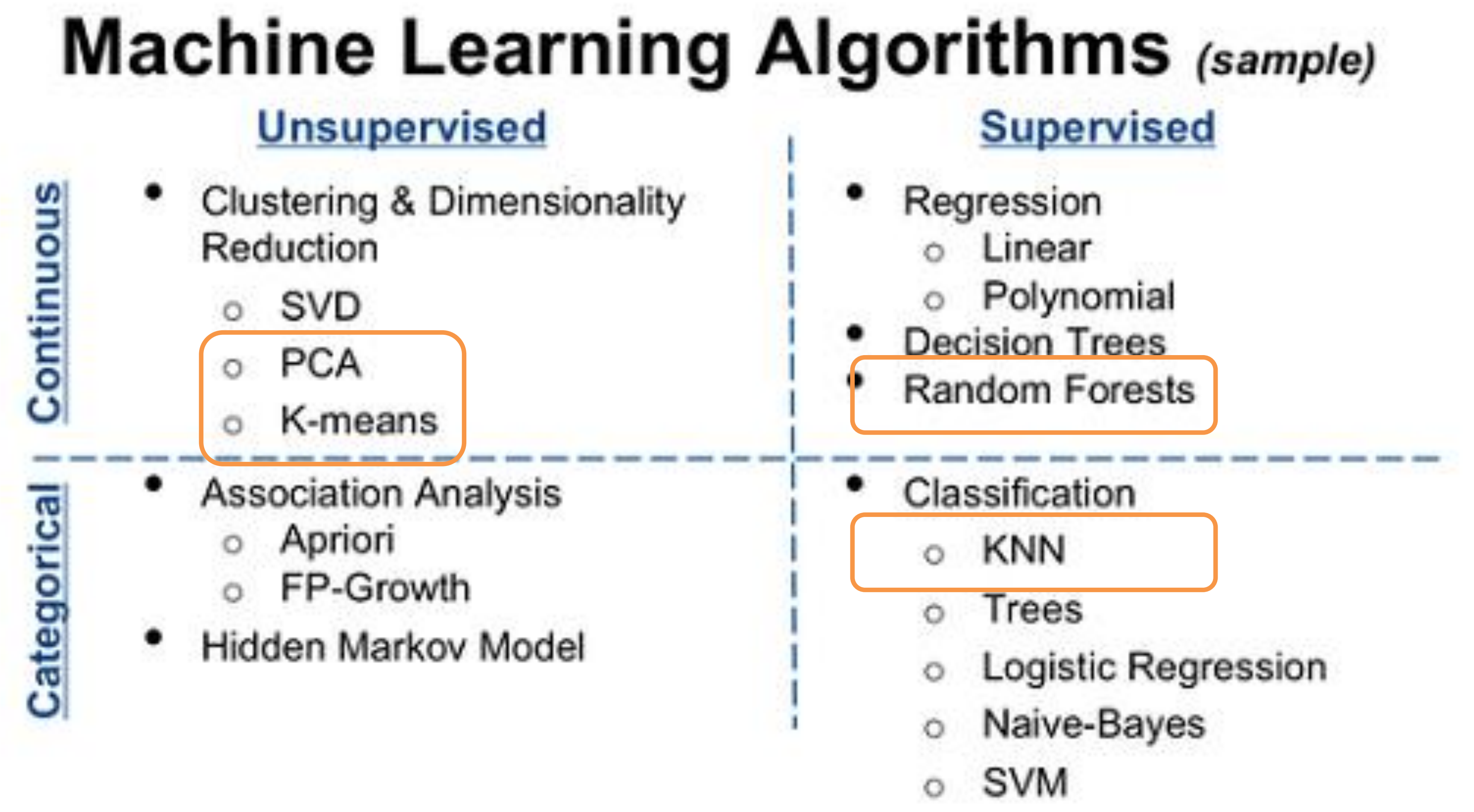


Unsupervised learning is a technique where you do not need to (can’t) supervise the model. Instead, you need to allow the model to work on its own to discover patterns.





	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction



Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Supervised learning - Regression

Regression is a **mathematical method** that models the **relationship** between a **dependent variable** and a series of **independent variables** (Draper and Smith 1998).

The regression methods allow modeling the **relationship** between a **response** variable (Yi) and a set of **explanatory variables or drivers** (Xi…Xn) that are related to the dependent variable.

$$Yi=\beta 0+\beta iXi+\epsilon i$$

Supervised learning - Regression

- ✓ **Binary/logit regression**

The primary objective of binary regression is to model how the probability of occurrence of an event, usually dichotomous, is influenced by the presence of various factors.

As noticeable this matches the presence/absence nature of hazards, so it suits the purpose of modeling the likelihood of a hazard event to occur.

Supervised learning - Regression

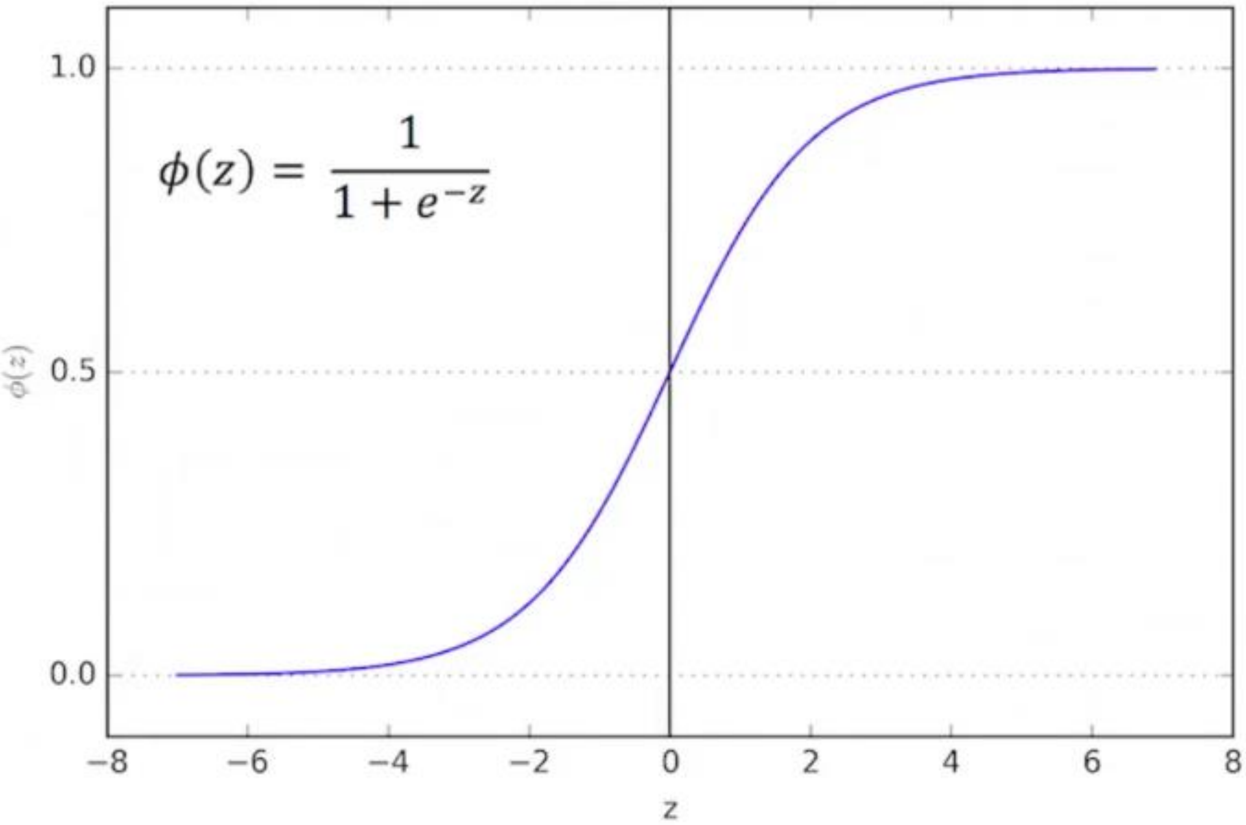
✓ **Logistic regression or logit models**

This kind of regression method belongs to the so-called Generalized Linear Models.

log(P)=β0+βnXn

P=1/1+exp−z

z=β0+βnXn



Supervised learning - Regression

Logistic regression or logit models actually fall within the category of binary modeling techniques.

All of them take a binary response variable (1-presence of a phenomena and 0-absence of that phenomena).

Other algorithms such as Machine Learning (such as - Random Forest)are able to reproduce this kind of models.

Supervised learning - Regression

What is the outcome of these models?

The **probability** (0 to 1) of that event to occur given the modeled relationship with their drivers.

This might be conducted using spatial data so that we can not only understand this relationship but map the likelihood of occurrence.

Supervised learning - Classification

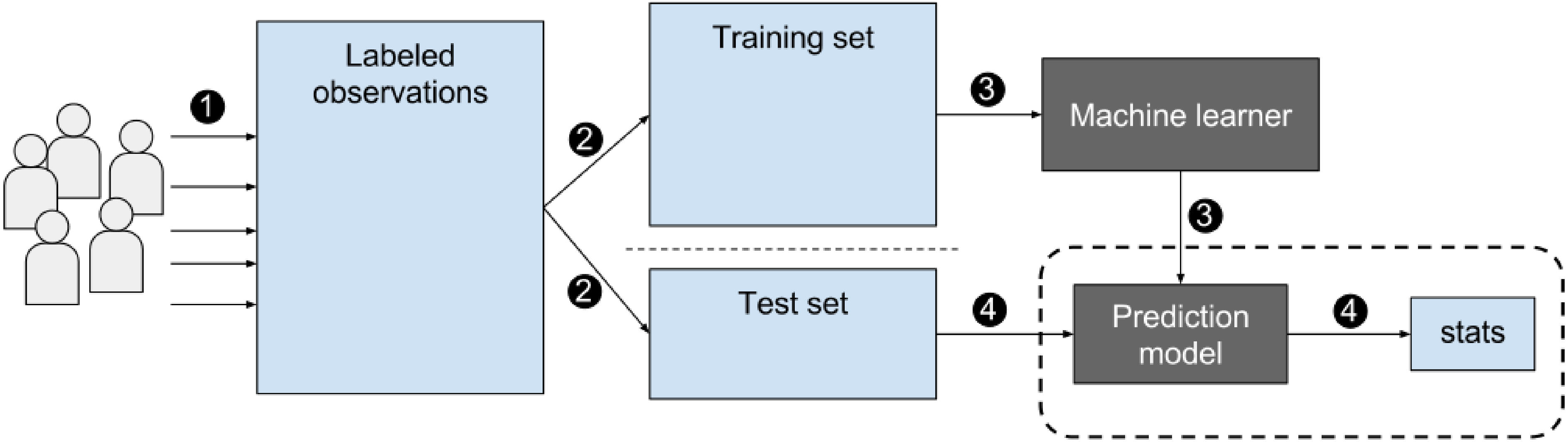
Classification is a subcategory of supervised learning where the goal is to predict the **categorical class labels** (discrete, unoredered values, group membership) of new instances based on past observations.

There are two main types of classification problems:

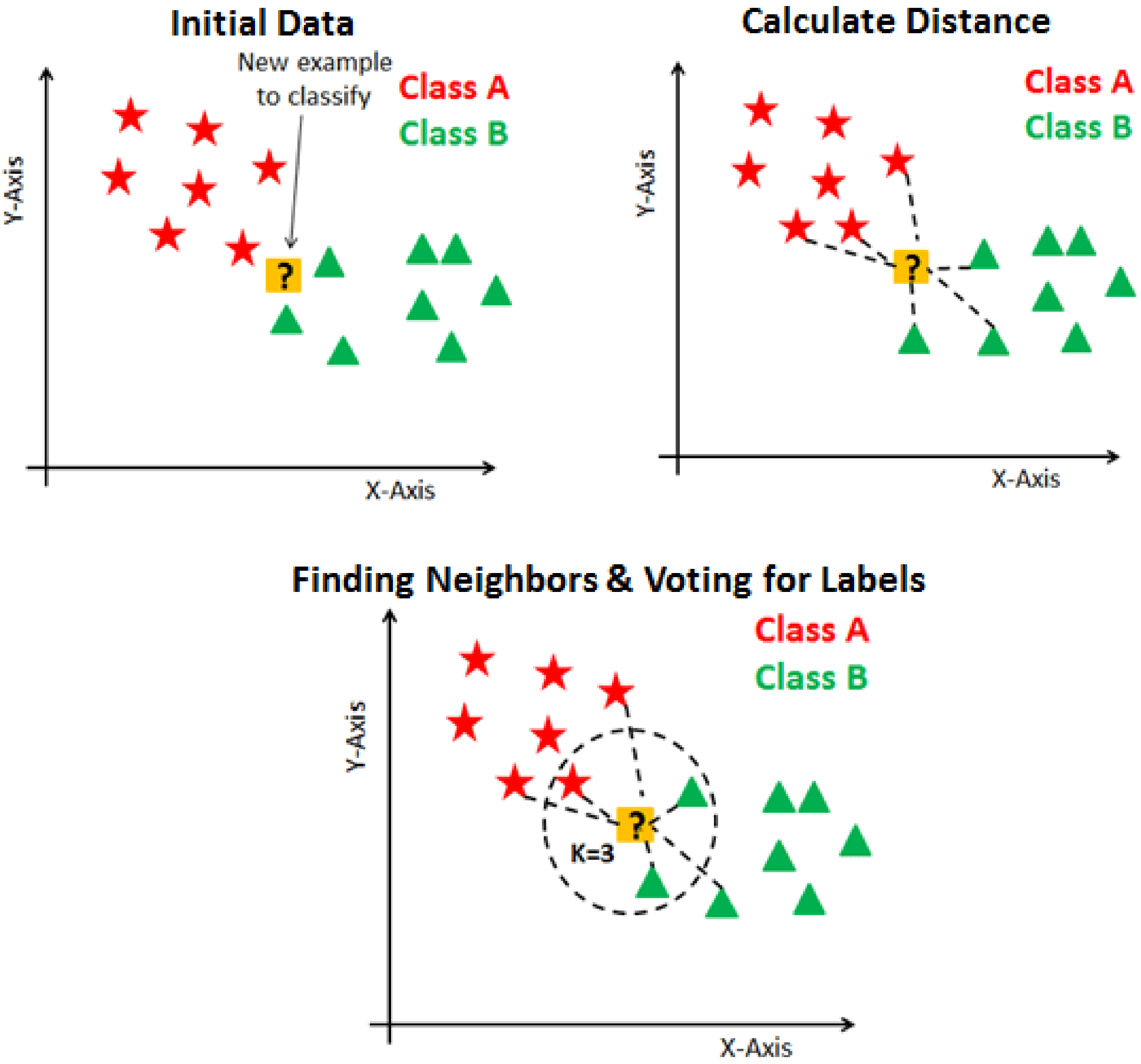
- Binary classification
- Multi-class classification

Supervised learning - Classification

Standard procedure in supervised learning:



Supervised learning - KNN



Initial Data

Calculate Distance

Finding Neighbors & Voting for Labels

Supervised learning – validation/testing

Model calibration

Optimizing the model to use the best set of model parameters using a calibration sample:

- Training sample: set of observations used to fit the model.
- Validation sample: independent sample used to retrieve the optimal parameters of the model

Model testing

Test sample: independent sample used to test the performance of the model

https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets

Supervised learning – validation/testing

What's an independent sample? A data subset or record that has not been used to construct the model. It can be build or obtained following several procedures:

✓ **Extract a random sample from a dataset**

This is valid for almost any kind of validation, though independence might be questioned.

✓ **Preserve a subset of data meeting some criteria**

Keep the last year of a time series to test the model. This is particularly interesting when we build a dynamic model.

Supervised learning – validation/testing

How many samples do we have to build? Well, at least one but the more the better because are able to account for uncertainty or dispersion in model performance.

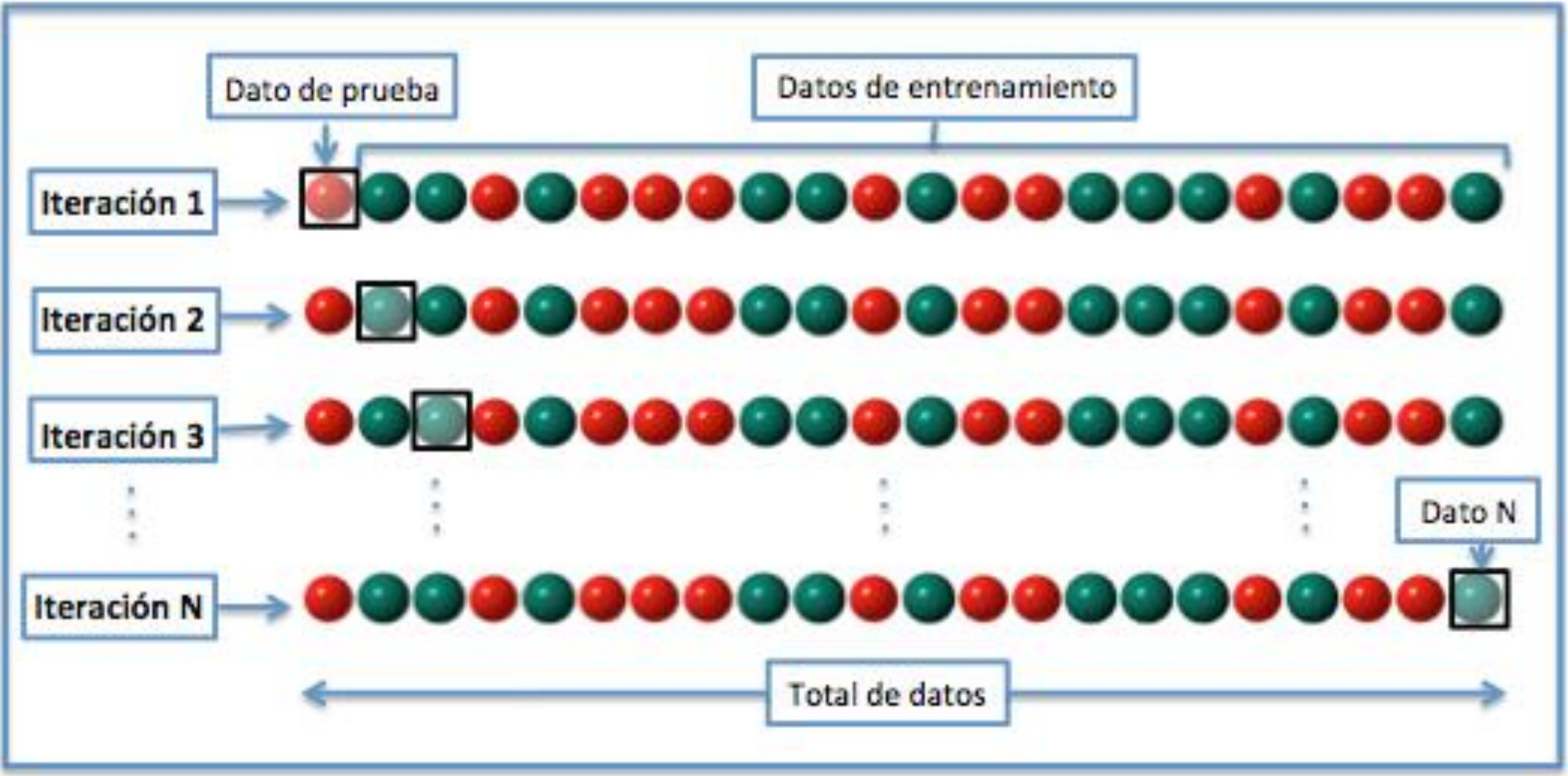
Which fraction of data may we keep for validation purposes? The truth is that there is no single answer to this question, although there is some consensus that the larger the sample of data, the greater the proportion of data that we can allocate for validation.

Supervised learning – validation/testing

Are there any specific techniques to be used? Fortunately for us, there are several well-known techniques and approaches available:

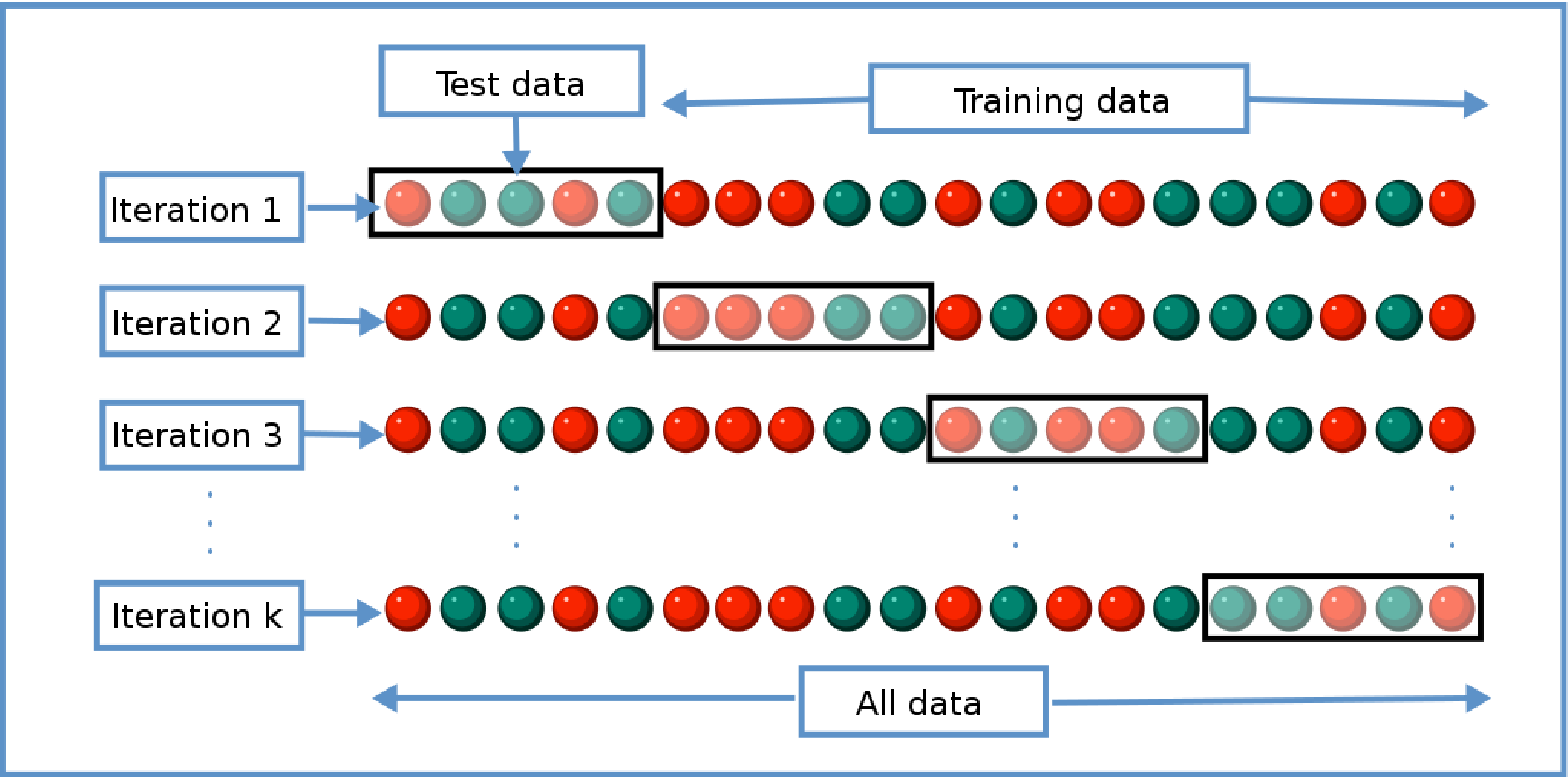
- ✓ Single random split: just extract one validation subset.
- ✓ Cross-validation: iterative resampling data
  - Leave-one-out: Fit the model with all records but one and check the prediction for that single record
  - k-fold: Split data into *k* groups and fit as many models as required to evaluate all groups.

Leave-one-out cross-validation



https://upload.wikimedia.org/wikipedia/commons/2/2d/Leave-one-out.jpg

K-fold cross-validation



Source: https://upload.wikimedia.org/wikipedia/commons/thumb/b/b5/K-fold_cross_validation_EN.svg/1024px-K-fold_cross_validation_EN.svg.png

How do we evaluate a model?

We know how we to organize and treat the data to conduct a validation but, **how do we calculate the efficiency of a model?**

✓ **Regression error estimate**

Methods to calculate the difference between observed and predicted values: residual, MSE, RMSE

✓ **Classification performance**

Methods to measure the accuracy of a classification. **Binary regression falls within this category since we work with integer 1/0 values.**

How do we evaluate a model?

Residual: raw difference between observed value and predicted value.

Mean squared error (MSE):

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

Root mean squared error (MSE):

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

How do we evaluate a model?

Accuracy

Fraction of correctly classified observations

Kappa Cohen’s

Statistic index that is used to measure inter-rater reliability qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation (wikipedia.org)

How do we evaluate a model?

$$k \equiv \frac{p_o - p_e}{1 - p_e}$$

		B	
		Yes	No
A	Yes	a	b
	No	c	d

		B	
		Yes	No
A	Yes	10	15
	No	20	05

The observed proportionate agreement is:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{10+5}{50} = 0.3$$

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.

How do we evaluate a model?

In order to calculate any of this indicators of goodness of fit we have to transform probability values (remember that it is the output of any binary model) to 1/0 categories.

The traditional way is to use a **threshold** to split probabilities:

- P>0.5 is assumed as 1
- P<0.5 is assumed as 0

How do we evaluate a model?

However, the **0.5 threshold** is rather **arbitrary** and not necessarily the most representative.

For instance, hazard events are rare phenomenon that occur few times so assuming 0.5 might be unrealistic.

Solution

Use threshold-independent measure such as the Area Under the Receiver Operating Characteristic Curve (AUC)

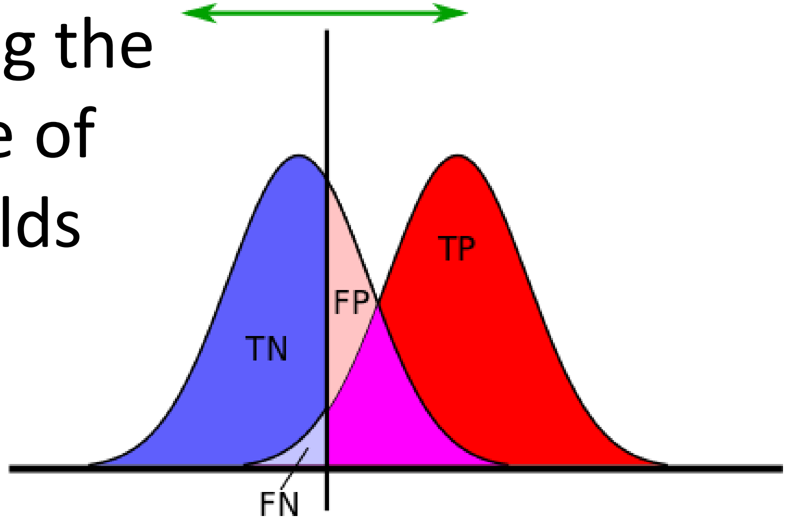
How do we evaluate a model?

In statistics, a receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The **AUC** is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

How do we evaluate a model?

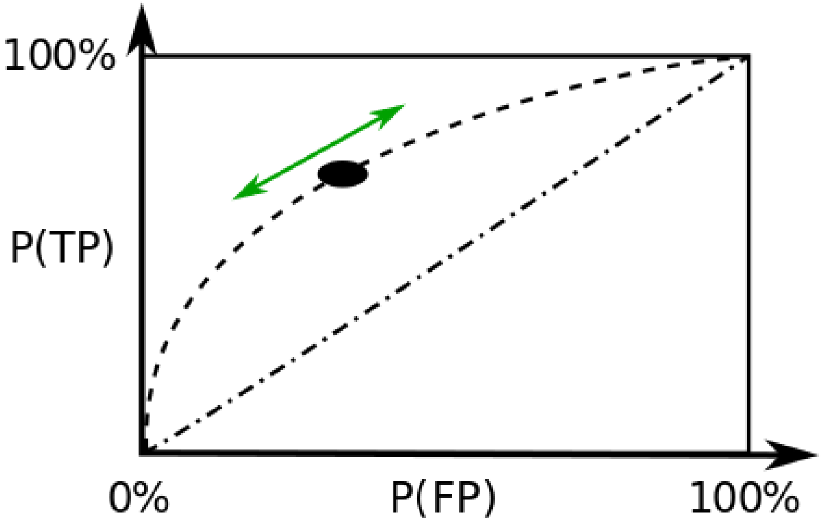
We plot TP/FP along the complete sequence of separation thresholds between 0 to 1.



TP	FP
FN	TN

The value of AUC ranges from:

- 0.5 random prediction
- 1 perfect fit



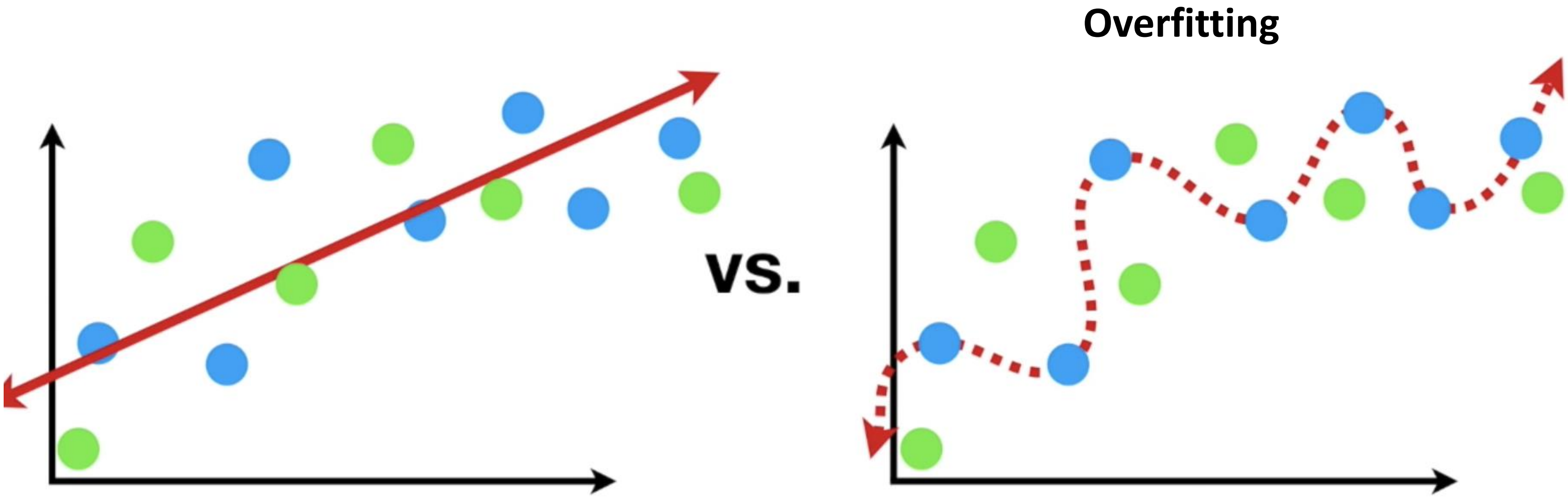
How do we evaluate a model?

Checking for overfitting

Any successful model must be able to model a series of relationships and generalize them when applied to an independent dataset.

Overfitting occurs when our model fits the training data but fails to reproduce the same phenomena when applied to other sample of data.

Overfitting



https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=2ahUKEwidt9_Wo4jnAhVBQhoKHQc0D18QjRx6BAgBEAQ&url=https%3A%2F%2Fmedium.com%2F%40carlos.cutillasdefrutos%2Fdo-we-care-enough-about-bias-60a211d072be&psig=AOvVaw1h0eYn_lUVXnqI5ZKhwlMj&ust=1579269520126819

Assessing variables’ performance

A key step towards an effective model is **selecting meaningful drivers** or, in a regression framework, independent variables, covariates or predictors.

The initial selection is often based on **literature review** and/or experience.

However, we must investigate what’s the contribution of each variable to the model, and **optimize** it selecting only **variables that** actually **contribute**.

Assessing variables' performance

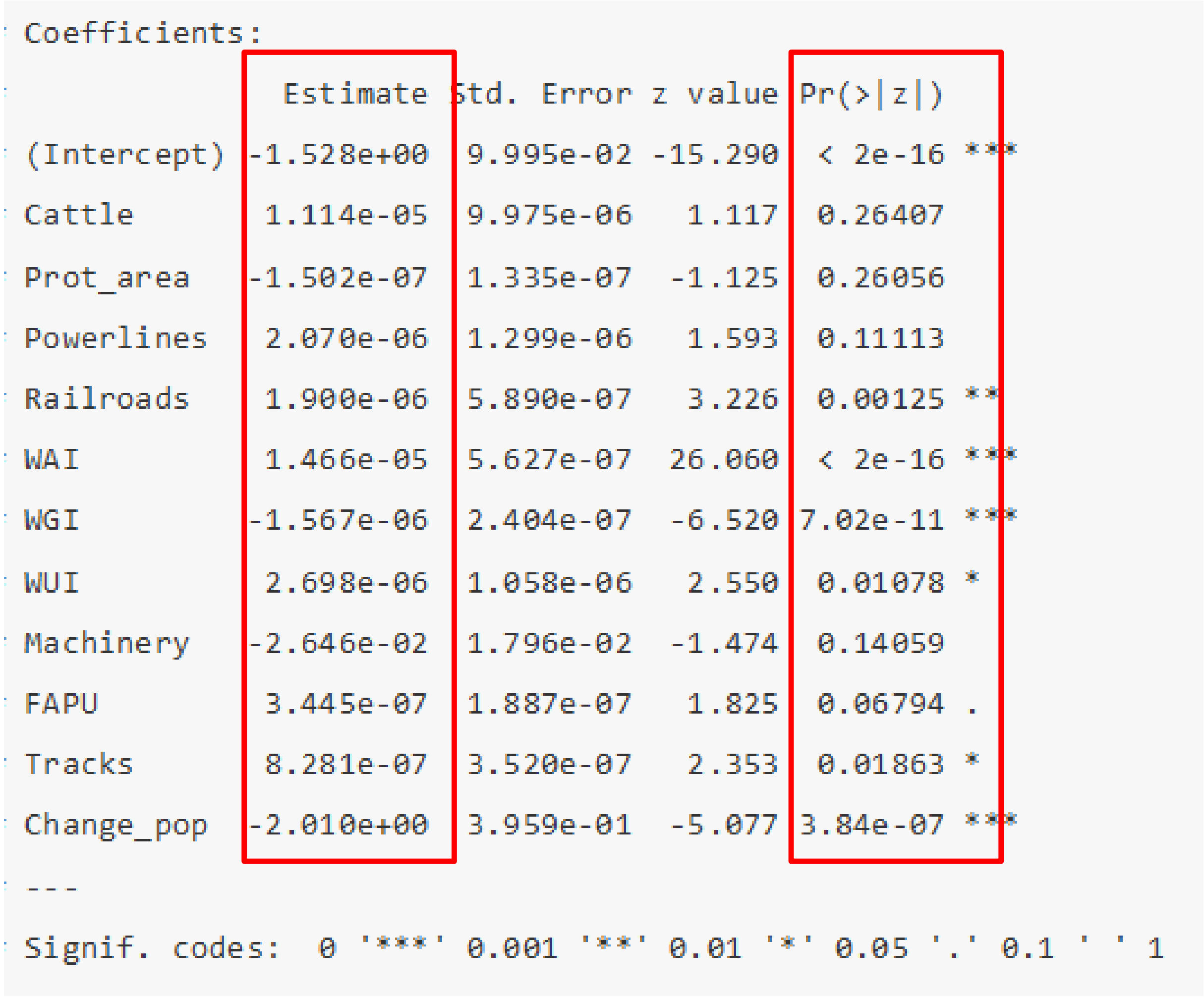
Likewise, the explanatory sense, that is, the way a variable interplays with the response, must be investigated. In its most simple expression we find:

- ✓ **Direct** (or positive) relationships, i.e., the value of the response increases as the value of the predictor does.
- ✓ **Inverse** (or negative) relationships. Vice versa.

Again, traditional regression methods express this information in the fashion of regression coefficients

Assessing variables’ performance

Coefficients **Significance**



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.528e+00  9.995e-02 -15.290  < 2e-16 ***
Cattle       1.114e-05  9.975e-06   1.117  0.26407    
Prot_area   -1.502e-07  1.335e-07  -1.125  0.26056    
Powerlines   2.070e-06  1.299e-06   1.593  0.11113    
Railroads    1.900e-06  5.890e-07   3.226  0.00125 ** 
WAI          1.466e-05  5.627e-07  26.060  < 2e-16 ***
WGI         -1.567e-06  2.404e-07  -6.520 7.02e-11 ***
WUI          2.698e-06  1.058e-06   2.550  0.01078 *  
Machinery   -2.646e-02  1.796e-02  -1.474  0.14059    
FAPU         3.445e-07  1.887e-07   1.825  0.06794 .  
Tracks       8.281e-07  3.520e-07   2.353  0.01863 *  
Change_pop  -2.010e+00  3.959e-01  -5.077 3.84e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients

Positive -> direct

Negative -> inverse

Significance

P-value < 0.05

Assessing variables’ performance

Unfortunately, machine learning algorithms do not offer a clear way to explore variable performance.

Nonetheless, there are ways to inquiry the models in order to understand the relative influence of the variables.

Partial plots and variable importance are perhaps the most widespread method to achieve this purpose.

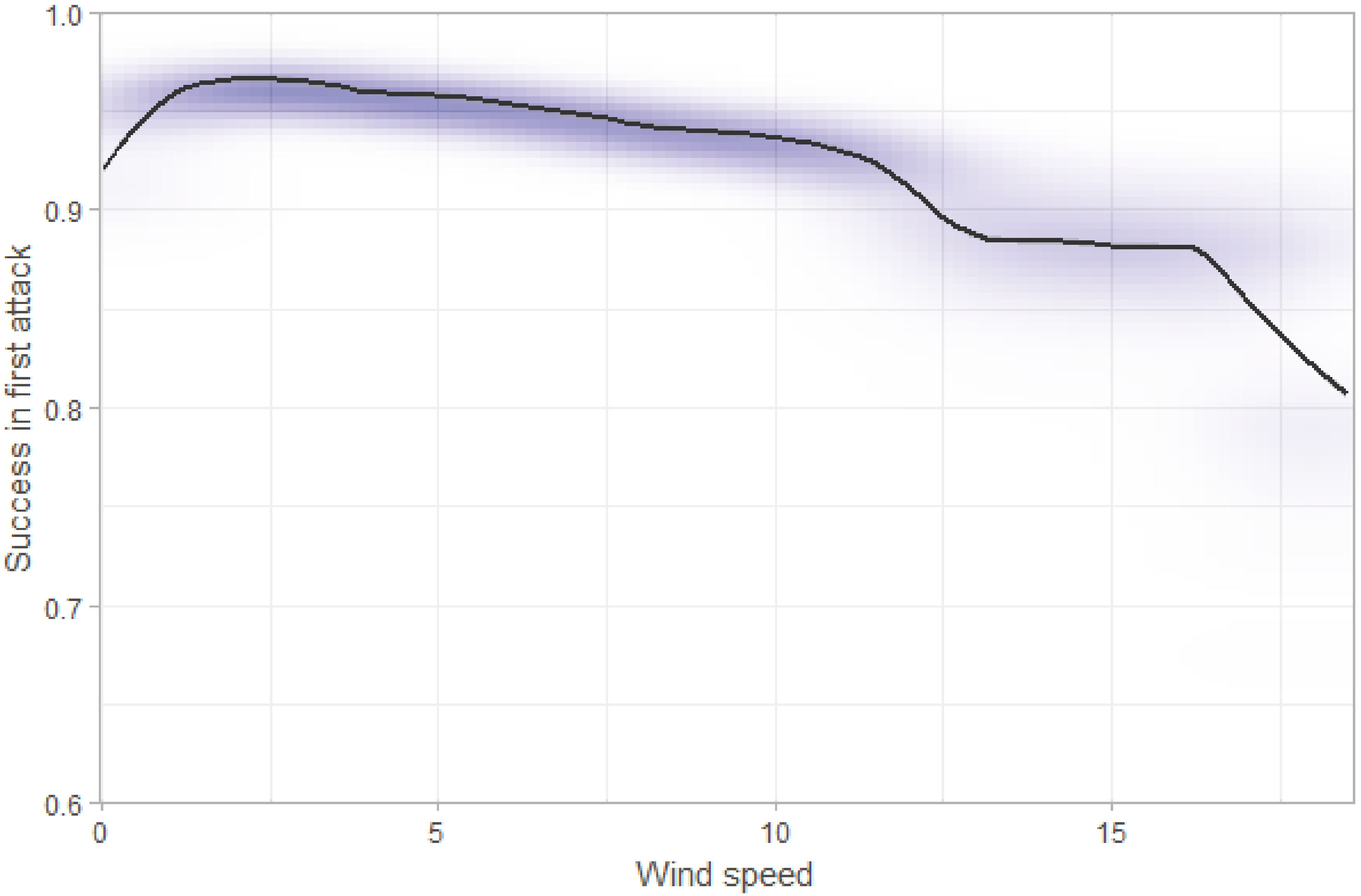
Assessing variables' performance

Partial dependence plots (PDP) are a graphical representation of the influence of a given covariate on the predicted response.

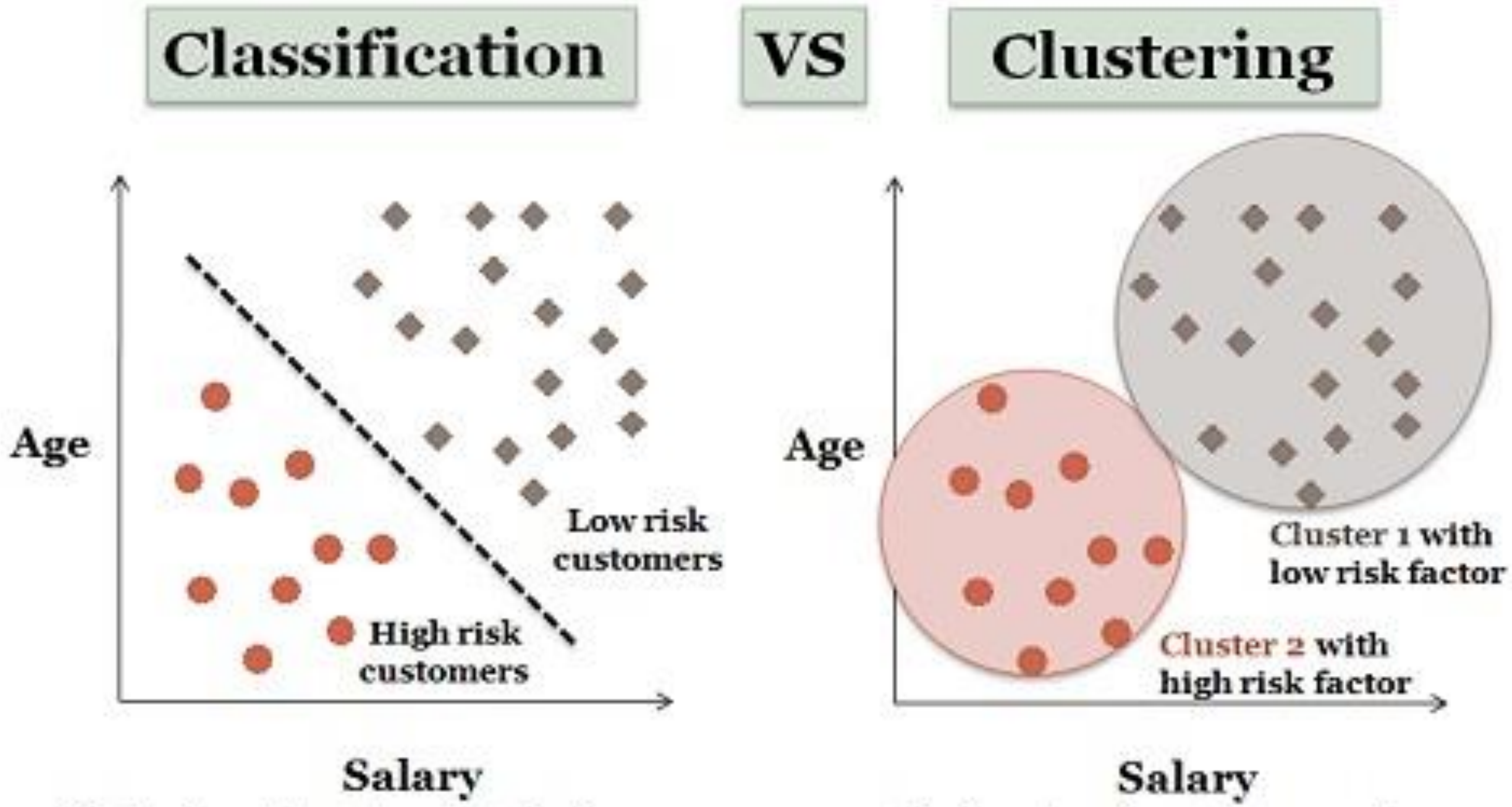
The x-axis in a PDP represents the value of the covariate, whereas the y-axis displays the associated predicted response.

If more than one variable entered the model they are set to their median value.

Assessing variables' performance



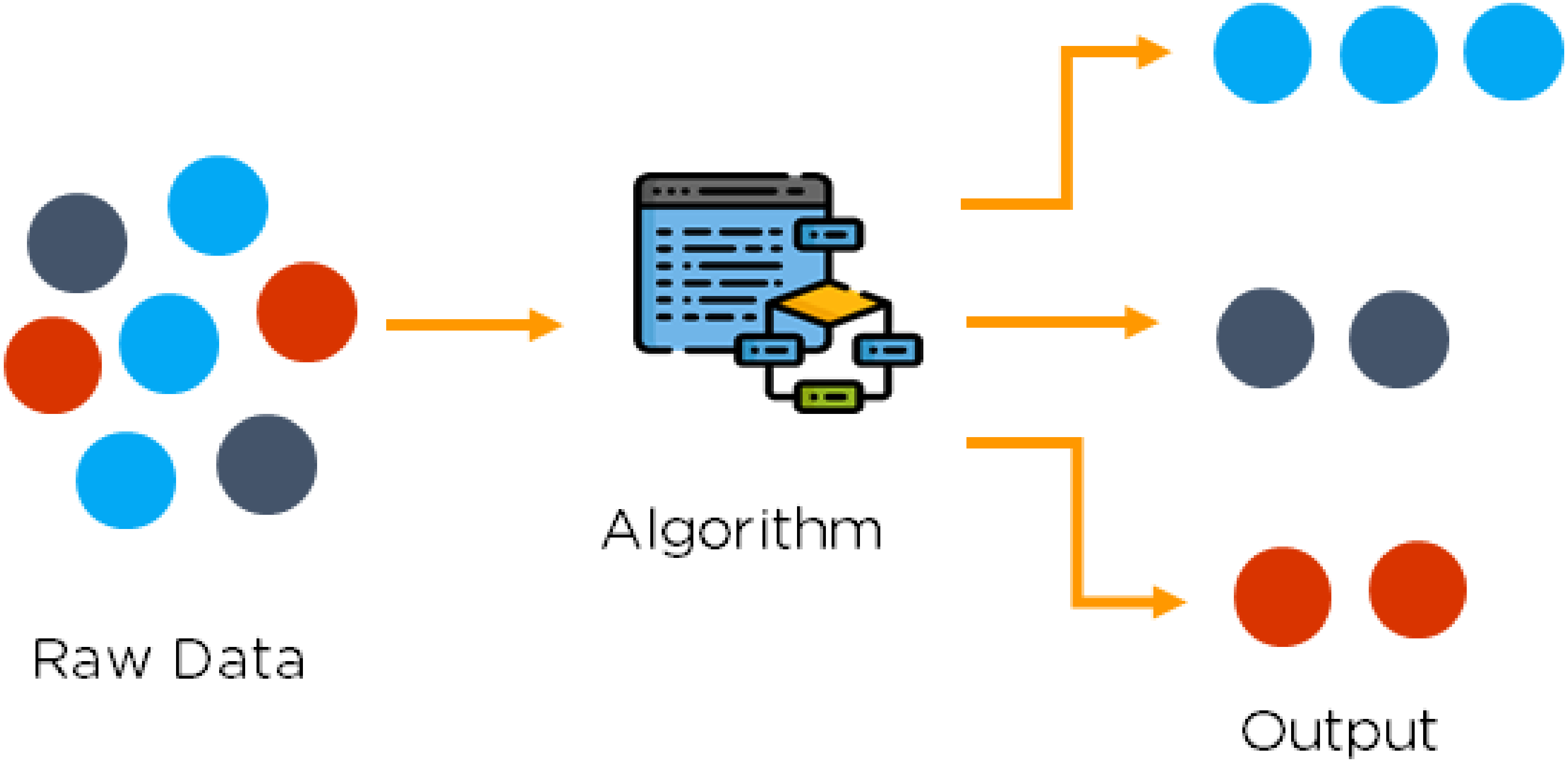
Unsupervised learning



Risk classification for the loan payees on the basis of customer salary

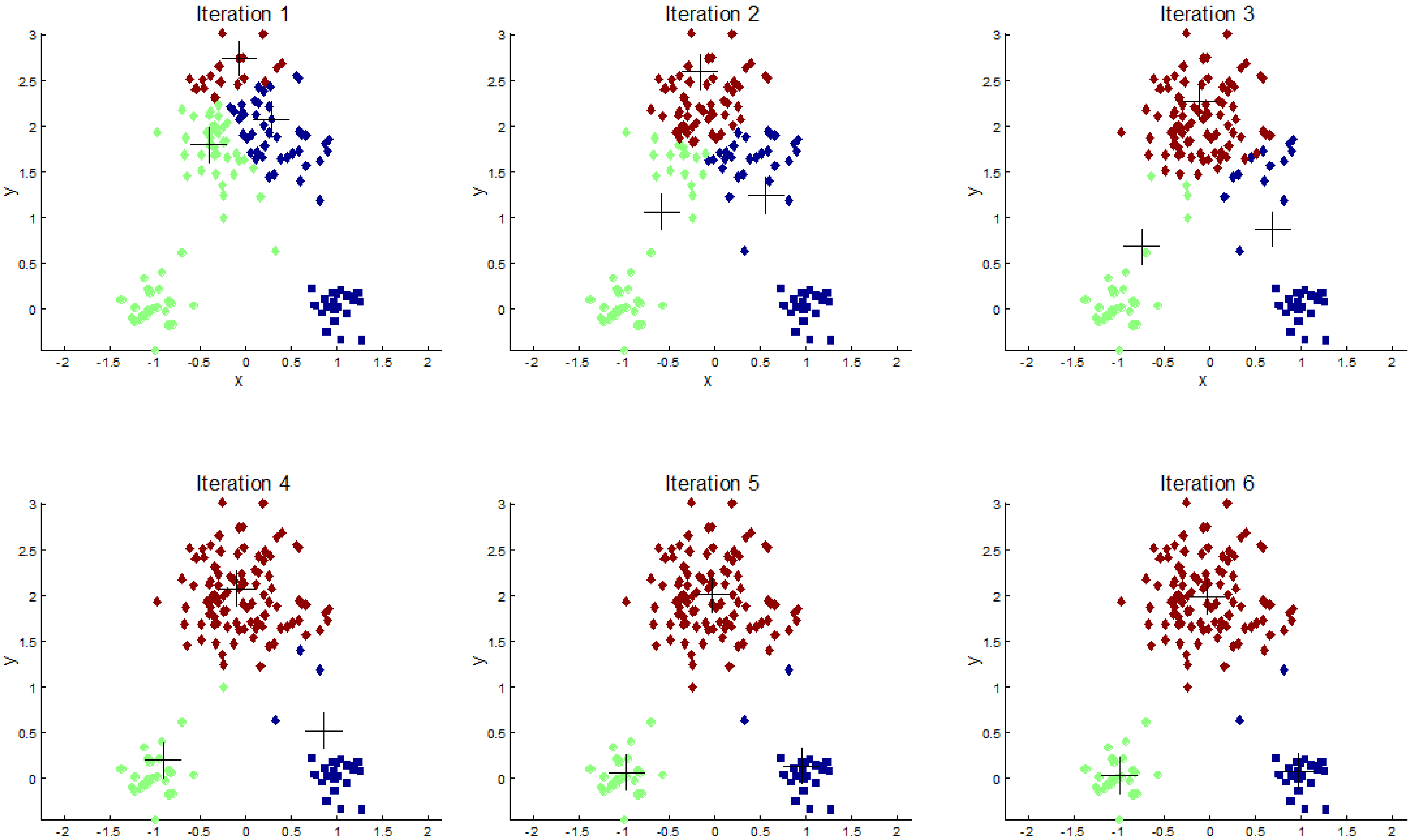
https://techdifferences.com/difference-between-classification-and-clustering.html

Unsupervised learning - clustering



https://www.quora.com/What-is-the-difference-between-k-means-and-hierarchical-clustering

Unsupervised learning – k-means



Iteration 1

Iteration 2

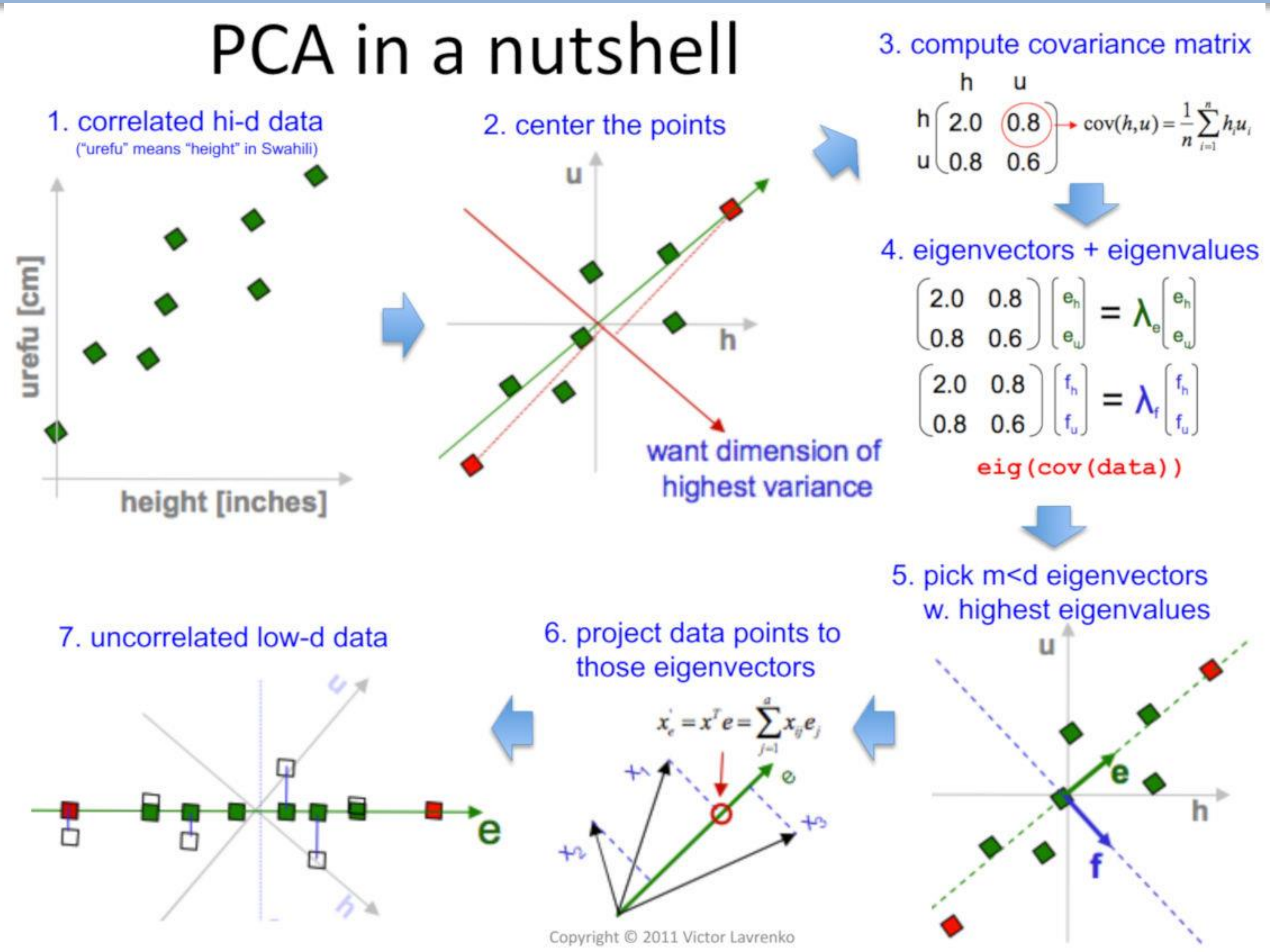
Iteration 3

Iteration 4

Iteration 5

Iteration 6

Unsupervised learning – PCA



PCA in a nutshell

1. correlated hi-d data

("urefu" means "height" in Swahili)

2. center the points

want dimension of highest variance

3. compute covariance matrix

$$\begin{array}{c|cc} & h & u \\ \hline h & 2.0 & 0.8 \\ u & 0.8 & 0.6 \end{array} \rightarrow \mathrm{cov}(h,u) = \frac{1}{n}\sum_{i=1}^{n} h_i u_i$$

4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

`eig(cov(data))`

5. pick m<d eigenvectors w. highest eigenvalues

6. project data points to those eigenvectors

$$x'_e = x^T e = \sum_{j=1}^{d} x_{ij} e_j$$

7. uncorrelated low-d data

Source: Lavrenko and Sutton 2011, slide 13. Lavrenko, Victor and Charles Sutton. 2011. "IAML: Dimensionality Reduction." School of Informatics, University of Edinburgh.