

Orange Labs
October, 2017

Khiops Enneade 8.0.1.35

Tutorial for « Predictive Clustering » functionality

Outline of this tutorial

- If you know what is a Clustering
 - Part I : How to use the tool [p4-p15]
- If you know what is a Clustering
 - But you don't know what is "Predictive Clustering"
 - Part II - Details about Predictive clustering [p16-p41]
 - Part III - Algorithms in Khiops Ennéade [p42-p44]
 - Part IV - The different Criteria (classification, clustering, Predictive Clustering) [p45-p55]
- Whether you know or not what is a Clustering or a Predictive Clustering
 - Part V - How to set the number of clusters [p56-p67]
 - Part VI - How to define a cluster profile [p69-p99]

Depending on your knowledge, you may skip parts of this tutorial...

Khiops Ennéade

- If you work for Orange  (wherever your office is in the world) and you would like to use the tool:
 - contact Vincent Lemaire [at] orange.com
- If you do not work for Orange : in that case, [Predicxis](https://khiops.predicxis.com) (khiops.predicxis.com) will be your official Khiops Enneade provider.

Remember that this tutorial is complementary to the tool's "User Guide". We also provide courses on the tool (contact 'Orange Learning' for that)

Khiops Ennéade

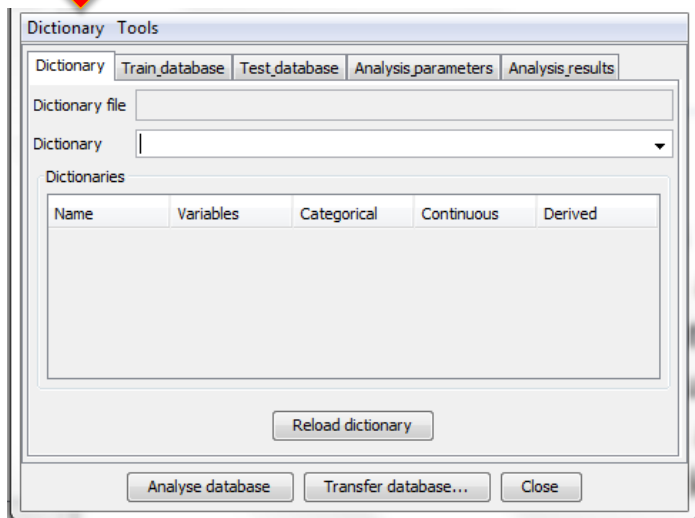
Part I

How to use the tool

Predictive Clustering

Example on the German database

- **Step 1 : Open an existing dictionary** (ex: sample Iris.kwc)
 - Description of variables to use during analysis



Available actions :

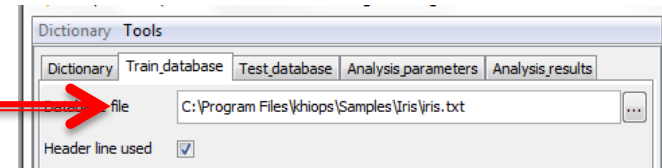
- Open, Save, Close, Reload
- Edition (*menu « Dictionaries Inspect », or NotePad*)

```
KWClass      Iris
{
    Continuous SepalLength;
    Continuous SepalWidth;
    Continuous PetalLength;
    Continuous PetalWidth;
    Symbol Class;
};
```

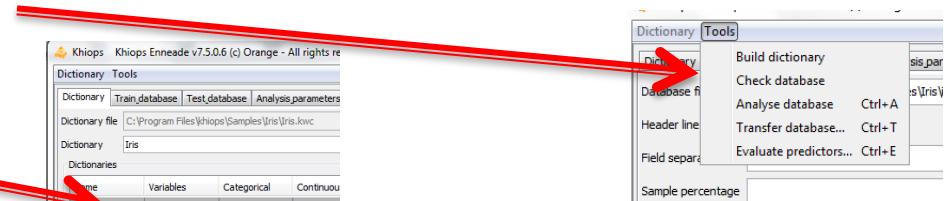
Supervised clustering

■ Step 1, bis : Build a new dictionary from a database (If no available dictionary)

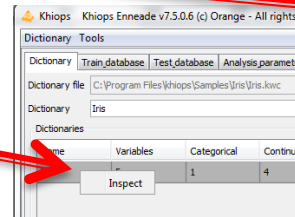
- 1. Specify database in Train database pane



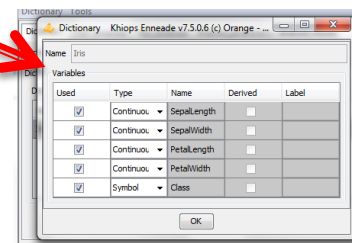
- 2. Menu Tool -> Build dictionary



- 3. Inspect the dictionary

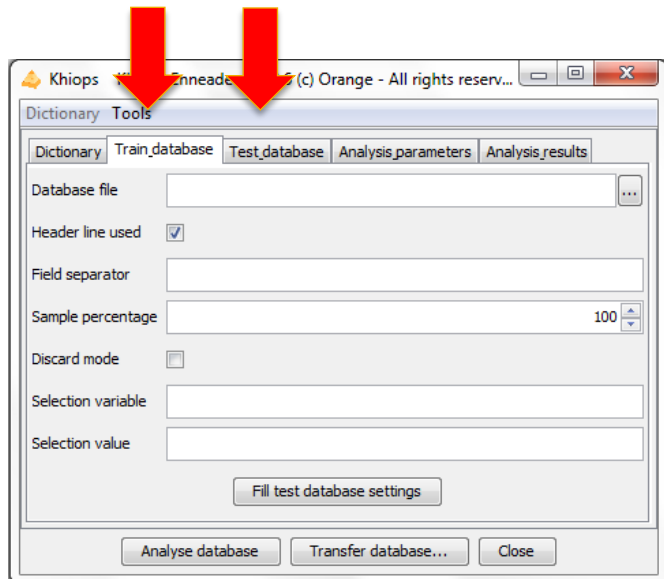


- 4. Check variables types
Select Used variables



Supervised clustering

■ Step 2 : Specify train and test databases



Three train/test split strategies :

- 1- Use two distinct files for train and test
- 2 - Discard mode : one single file, split into train and test according to a sample percentage
Use Fill test database settings to have a test database complementary of the train database
- 3 - Controlled way of splitting the instances by the means of a selection variable and selection value

Supervised clustering

■ Step 3 : Analysis parameters

Dictionary Tools

Dictionary Train_database Test_database Analysis_parameters Analysis_results

Target variable Class

Main target value Iris-setosa

Predictors System parameters

K-mean predictor ☒

KMean clusters number (K) 2

K-Means advanced parameters

Analyse database Transfer database... Close

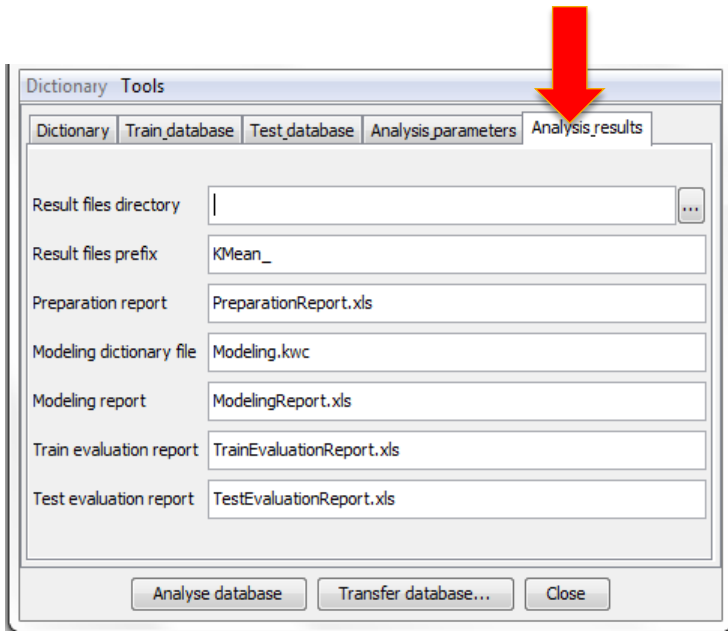
Type of target variable implies type of analysis
Categorical -> supervised clustering

One type of predictors

- K-mean :
 - Keeps only informative variables
 - Selection of variables

Supervised clustering

■ Step 4 : Analysis results



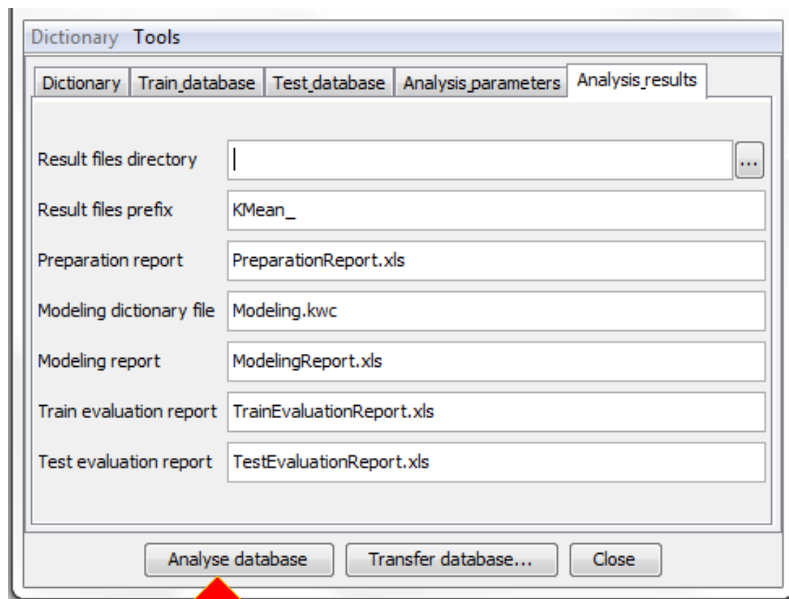
The screenshot shows a software window titled 'Dictionary Tools' with five tabs: 'Dictionary', 'Train_database', 'Test_database', 'Analysis_parameters', and 'Analysis_results'. A red arrow points to the 'Analysis_results' tab. The 'Analysis_results' tab contains several input fields and buttons. The fields are: 'Result files directory' (empty), 'Result files prefix' (KMean_), 'Preparation report' (PreparationReport.xls), 'Modeling dictionary file' (Modeling.kwc), 'Modeling report' (ModelingReport.xls), 'Train evaluation report' (TrainEvaluationReport.xls), and 'Test evaluation report' (TestEvaluationReport.xls). At the bottom are three buttons: 'Analyse database', 'Transfer database...', and 'Close'.

- Directory where all results files are written
- Prefix (*ex: in case of several experiments*)
- Description of trained univariate preparation models
- Technical description for deployment purposes
- Description of trained model with selected variables
- Evaluation on train database
- Evaluation on test database

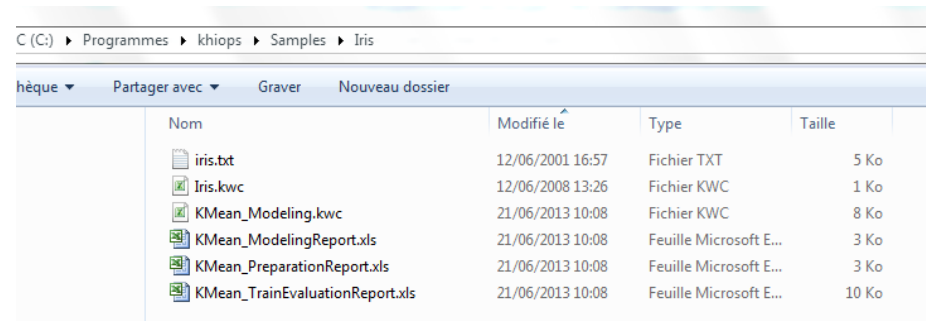


Supervised clustering

■ Step 5 : Start the analysis



1 – Start the analysis



2 - Inspect the directory and open the results files

Exploratory of results with Excel

Files produced by the tool:

- **Preparation report:** report file produced at the end of the univariate data analysis on the train database (default: KMean_PreparationReport.xls).
- **Modeling dictionary file:** name of the dictionary file produced once the predictor has been built (default: KMean_Modeling.kdic).
- **Modeling report:** name of the report file produced once the predictor has been built (default: KMean_ModelingReport.xls).
- **Train evaluation report:** name of the report file produced at the end of the evaluation of the predictor on the train database (default: KMean_TrainEvaluationReport.xls).

The content of the files and the meaning of the tables in the files are described in the document “Khiops Enneade Guide 8.0.1.34”

Exercises A and B ...

A : Perform a supervised clustering on Iris sample database

B : Perform a supervised clustering on Adult sample database

➡ Interpret the analysis results

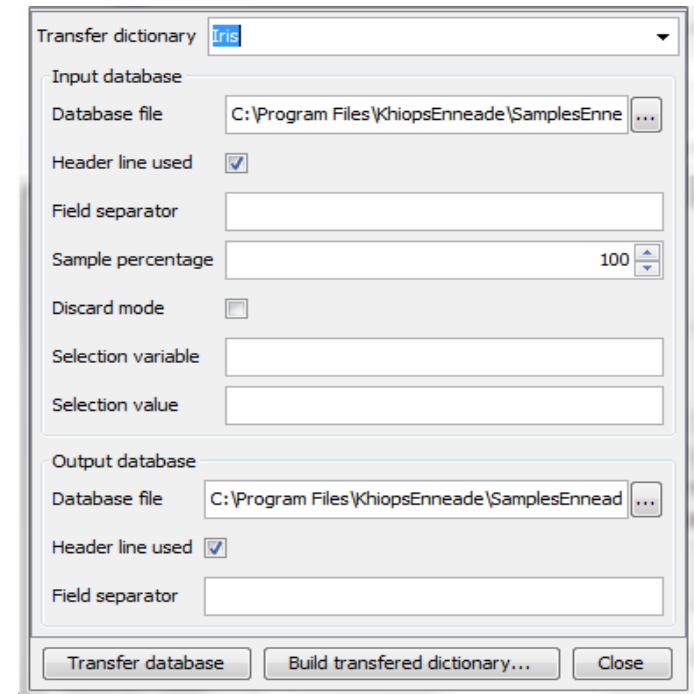
Notes...

A and B

Deploy a model

- Steps for model deployment

- 1- Open a modeling dictionary « *KMean_Modeling* », using a text editor
- 2- Remove all « *Unused* » tags from identifier variables
- 3- Open the modeling dictionary, using Enneade
- 4- Menu : « *Tools -> Transfer database* »



Exercise I ...

I: Deploy a clustering on database Iris

Khiops Ennéade

Part II

Details about Predictive clustering

Predictive clustering ?

The idea is to train a model to :

Predictive clustering ?

The idea is to train a model to :

- predict (a target variable)

Predictive clustering ?

The idea is to train a model to :

- predict (a target variable)
- and describe the data (using the explanatory variables)

Predictive clustering ?

The idea is to train a model to :


- predict (a target variable)
- and describe the data (using the explanatory variables)



simultaneously

Predictive clustering ?

The idea is to train a model to :

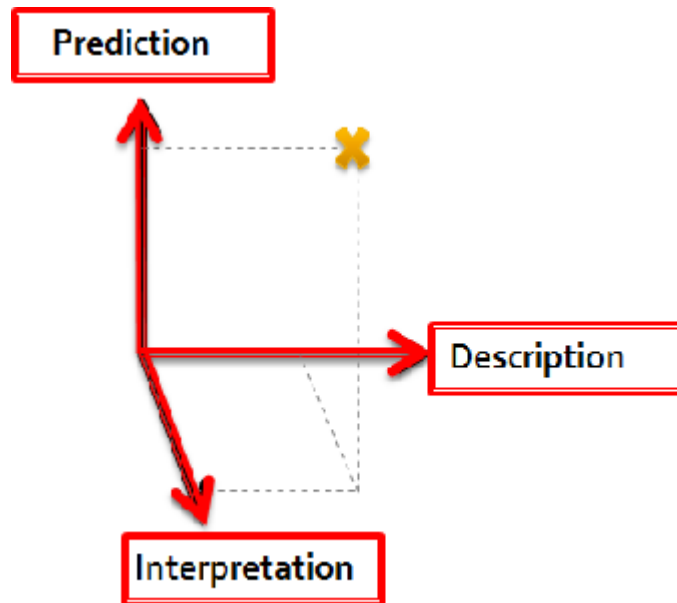
- predict (a target variable)
 - and describe the data (using the explanatory variables)
 - using a model easy to understand
- 
- simultaneously

Predictive clustering ?

The idea is to train a model to :

- predict (a target variable)
- and describe the data (using the explanatory variables)
- using a model easy to understand

} simultaneously

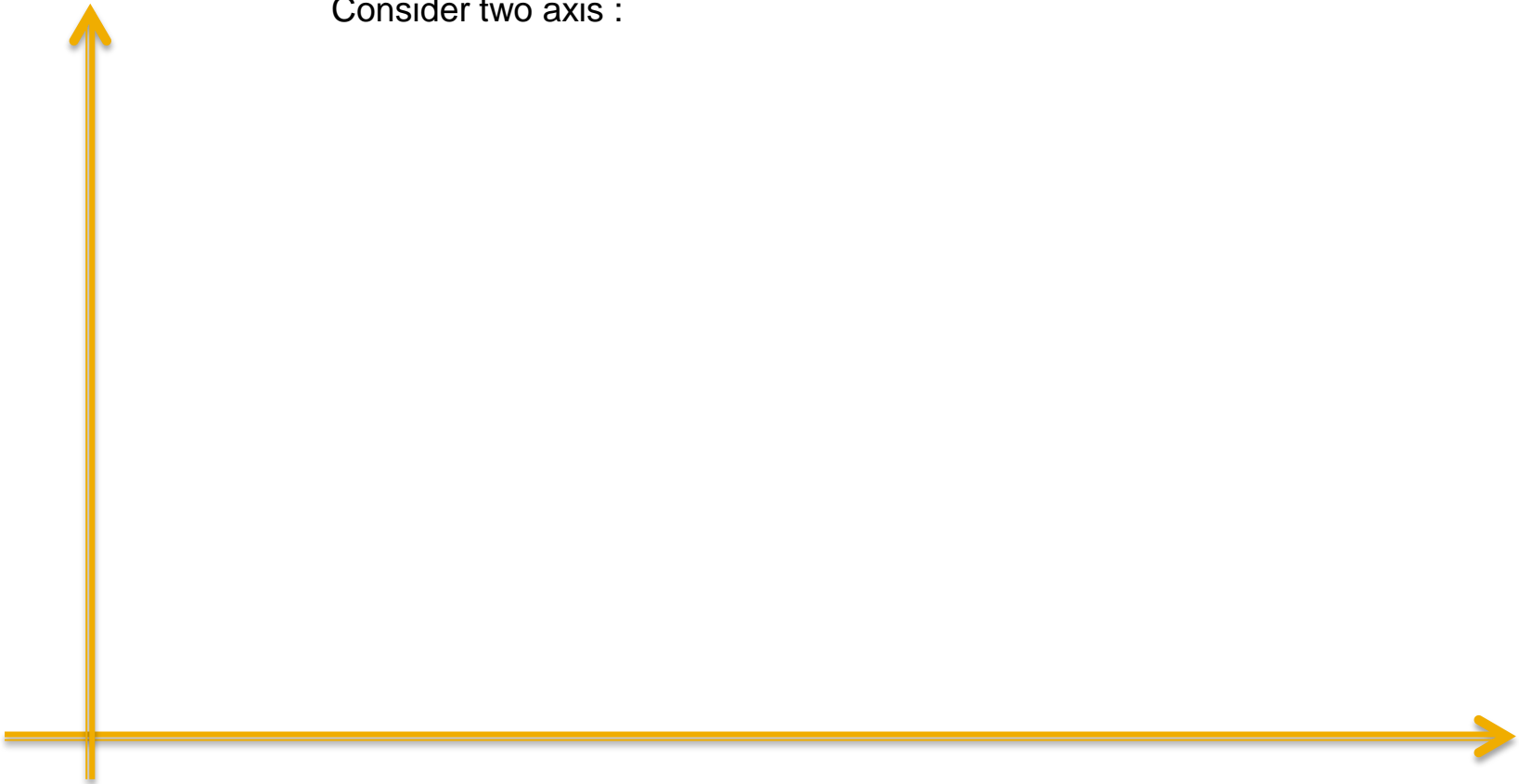


From explanatory analysis to supervised classification through “Predictive Clustering”

Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”

Consider two axis :



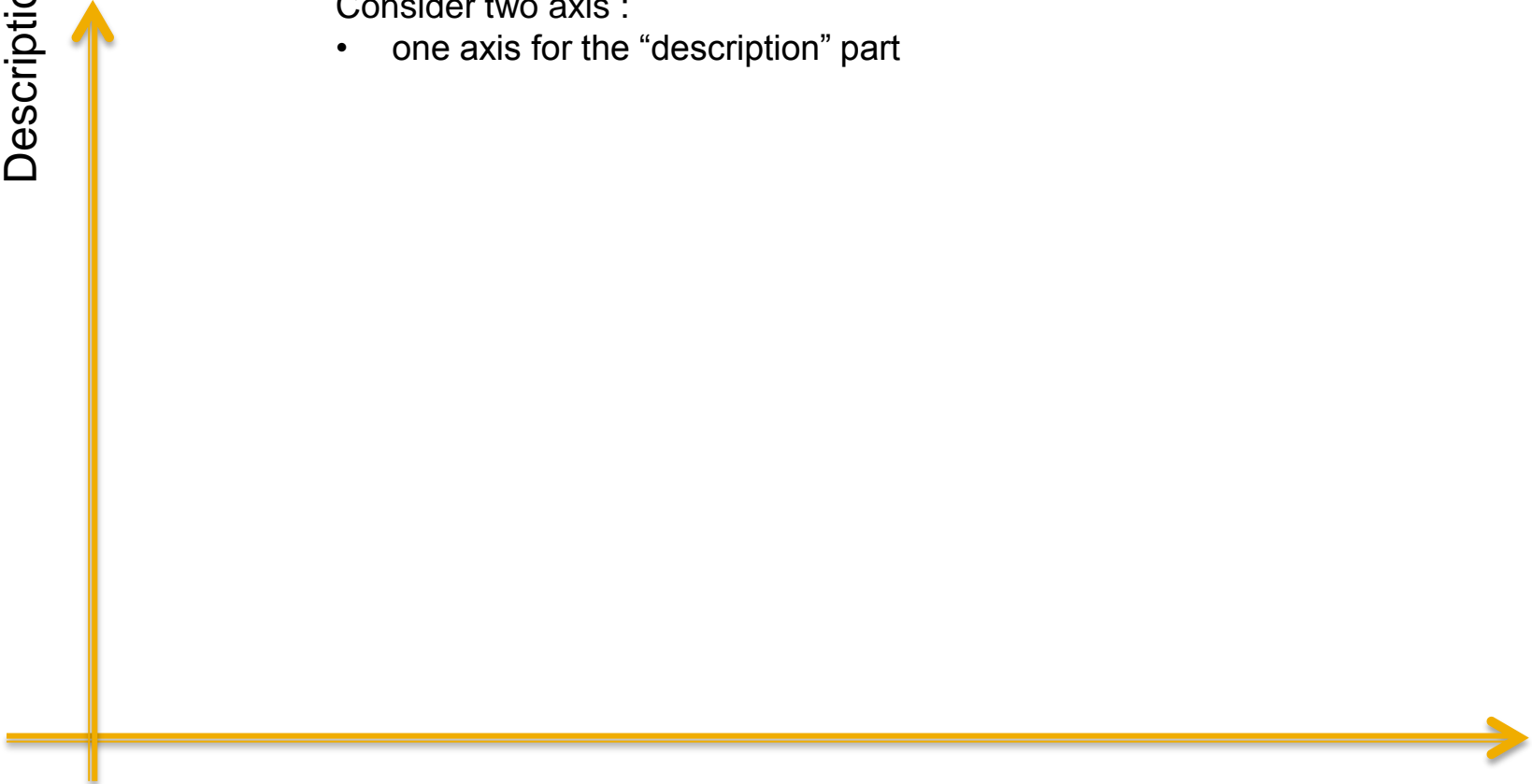
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”

Description

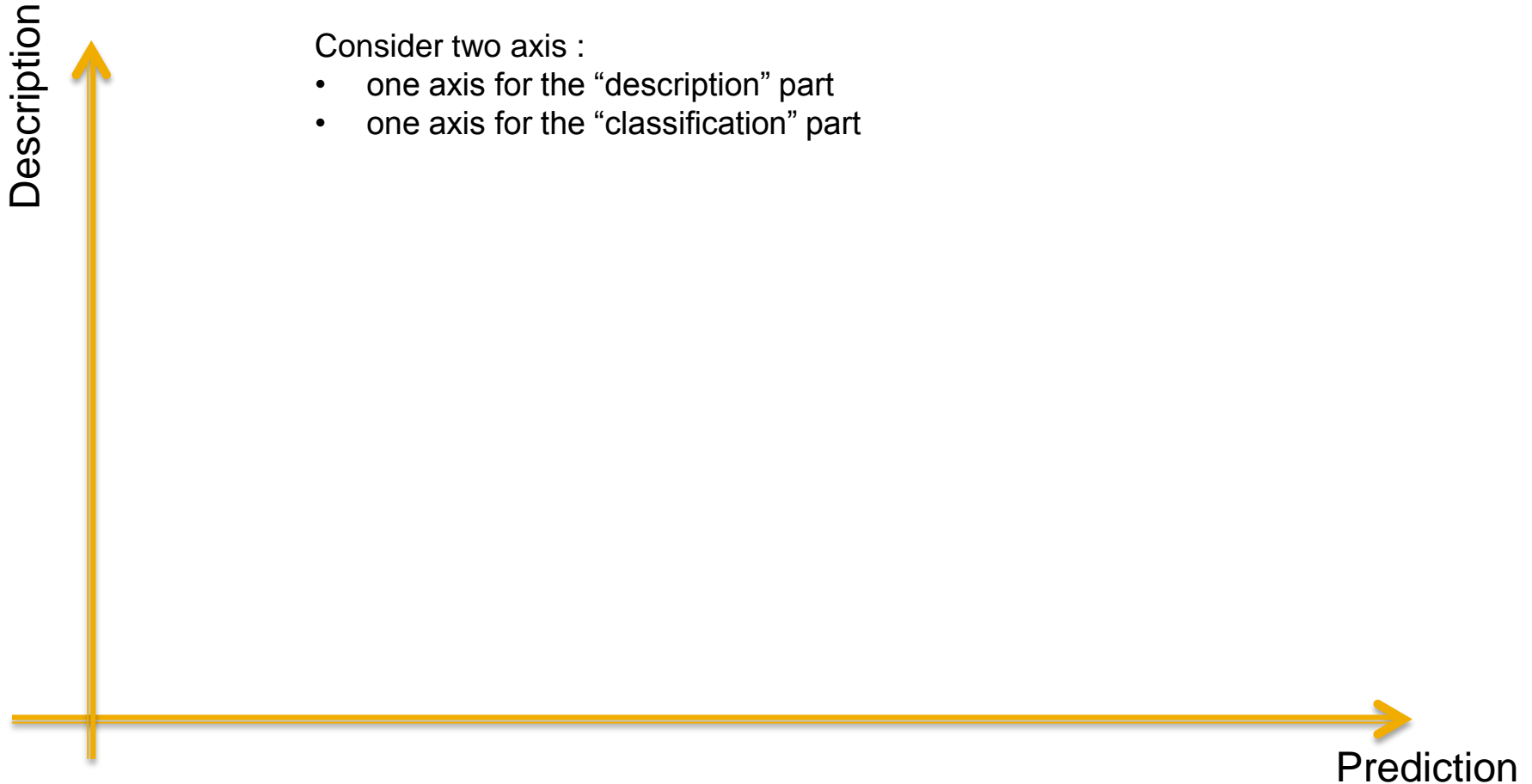
Consider two axis :

- one axis for the “description” part



Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



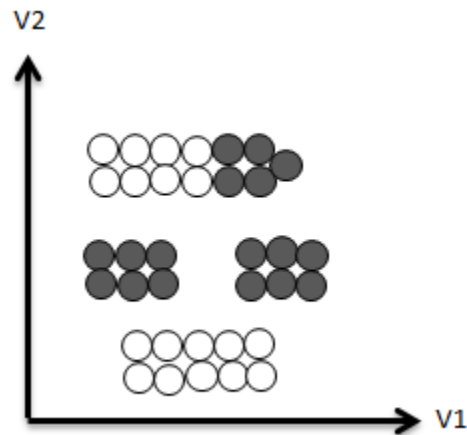
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”

Description

Consider two axis :

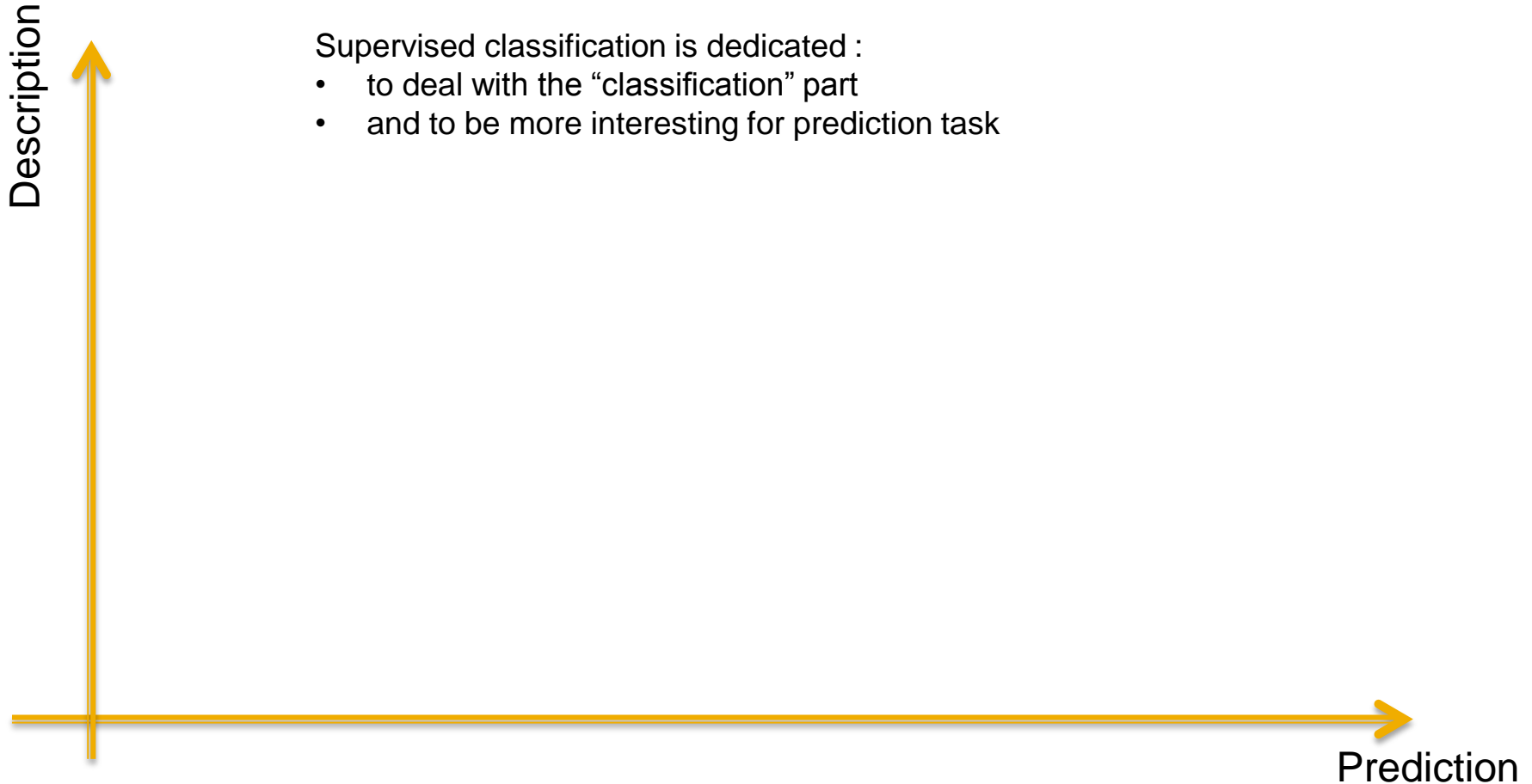
- one axis for the “description” part
- one axis for the “classification” part
- and a toy example (below) of a dataset constituted by:
 - examples described by two explanatory variables ($V1$, $V2$)
 - and a target variable (examples are black or white)



Prediction

Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



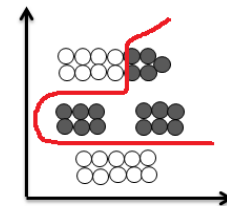
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”

Description

Supervised classification is dedicated :

- to deal with the “classification” part
- and to be more interesting for prediction task



supervised classification
(using a target variable)

Prediction

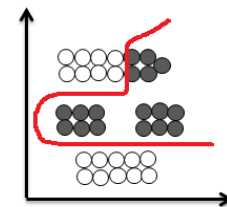
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”

Description

Clustering methods are more dedicated :

- to deal with the “description” part
- and to be more interesting for explanatory analysis

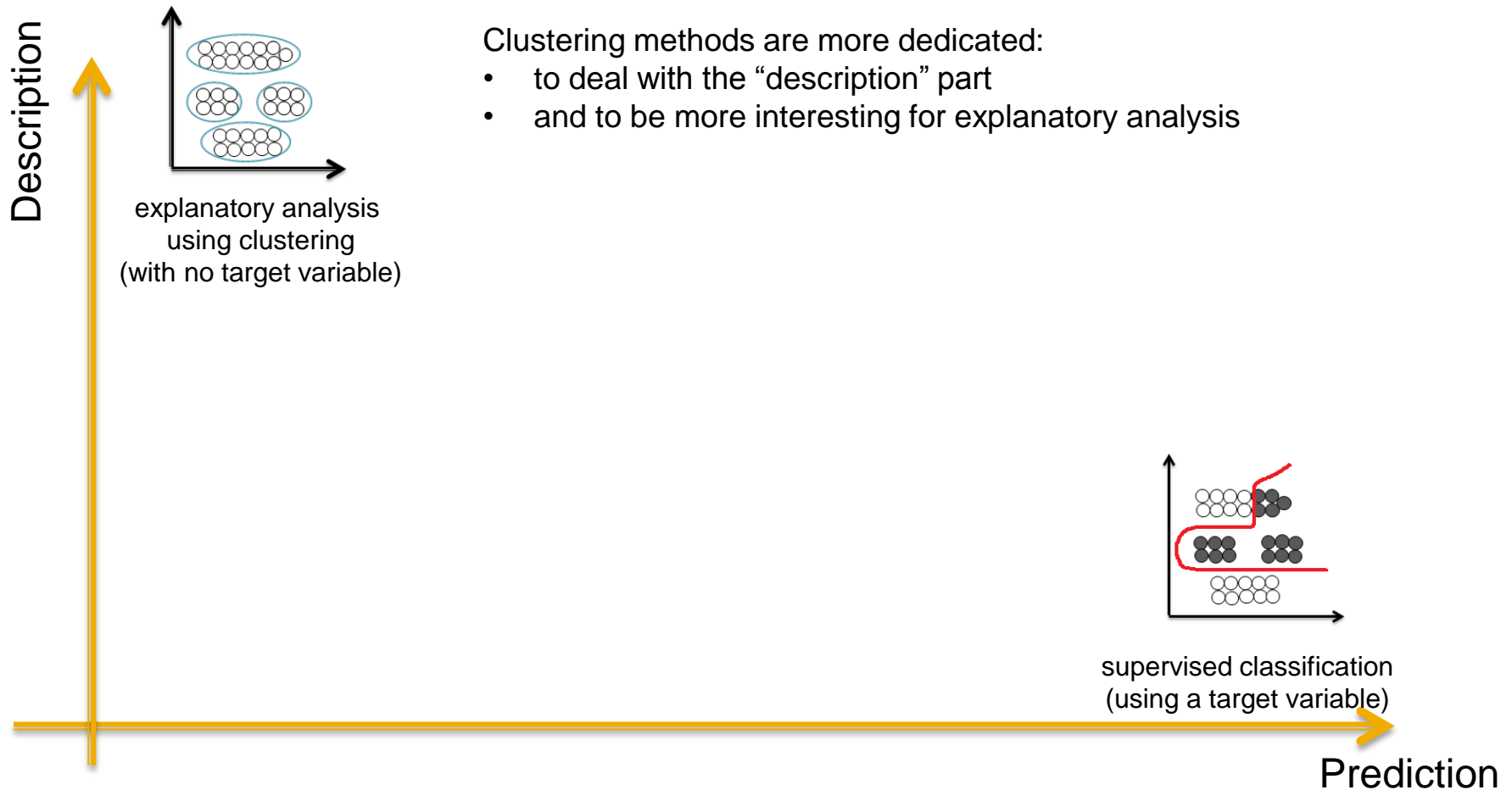


supervised classification
(using a target variable)

Prediction

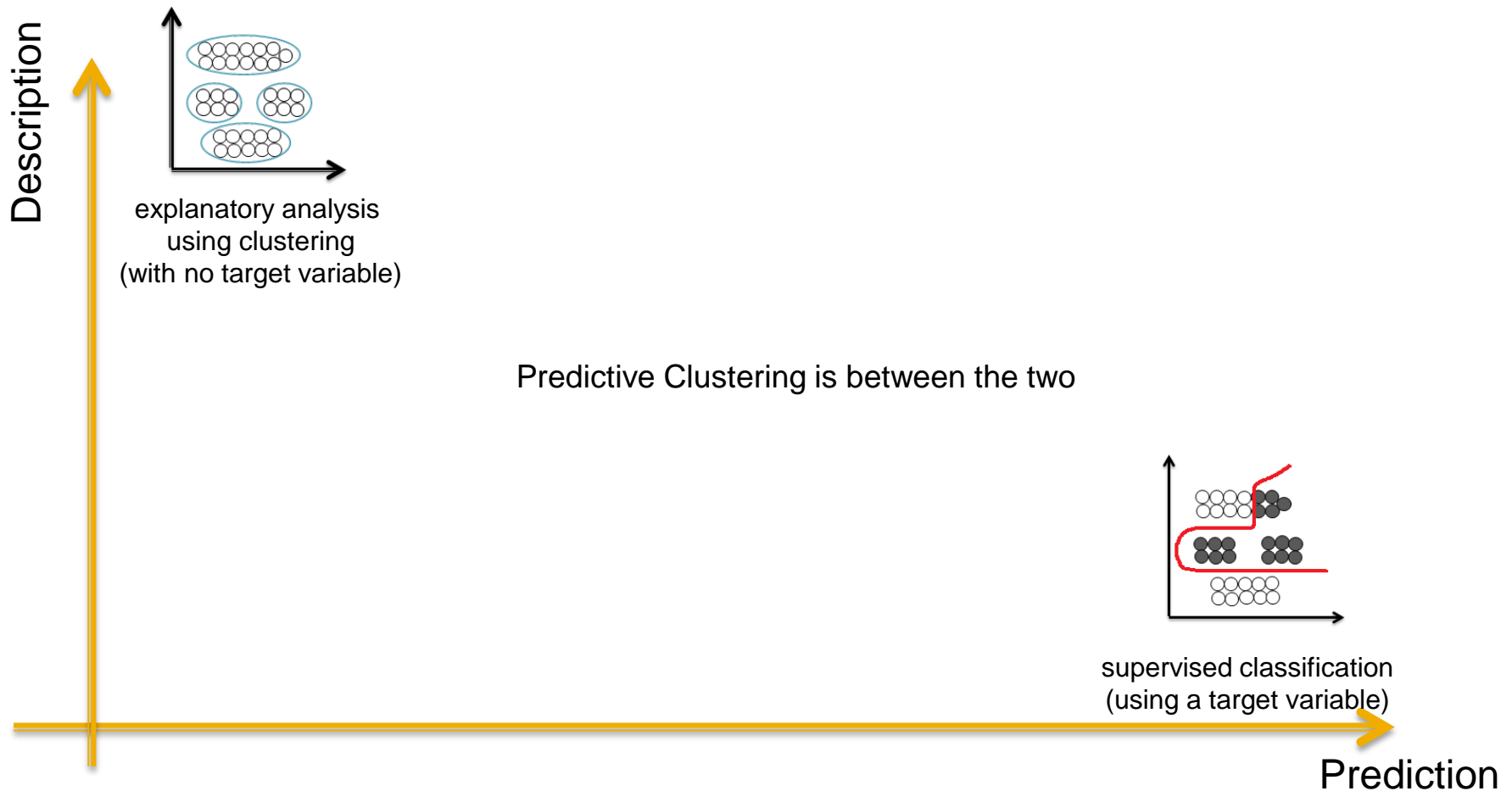
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



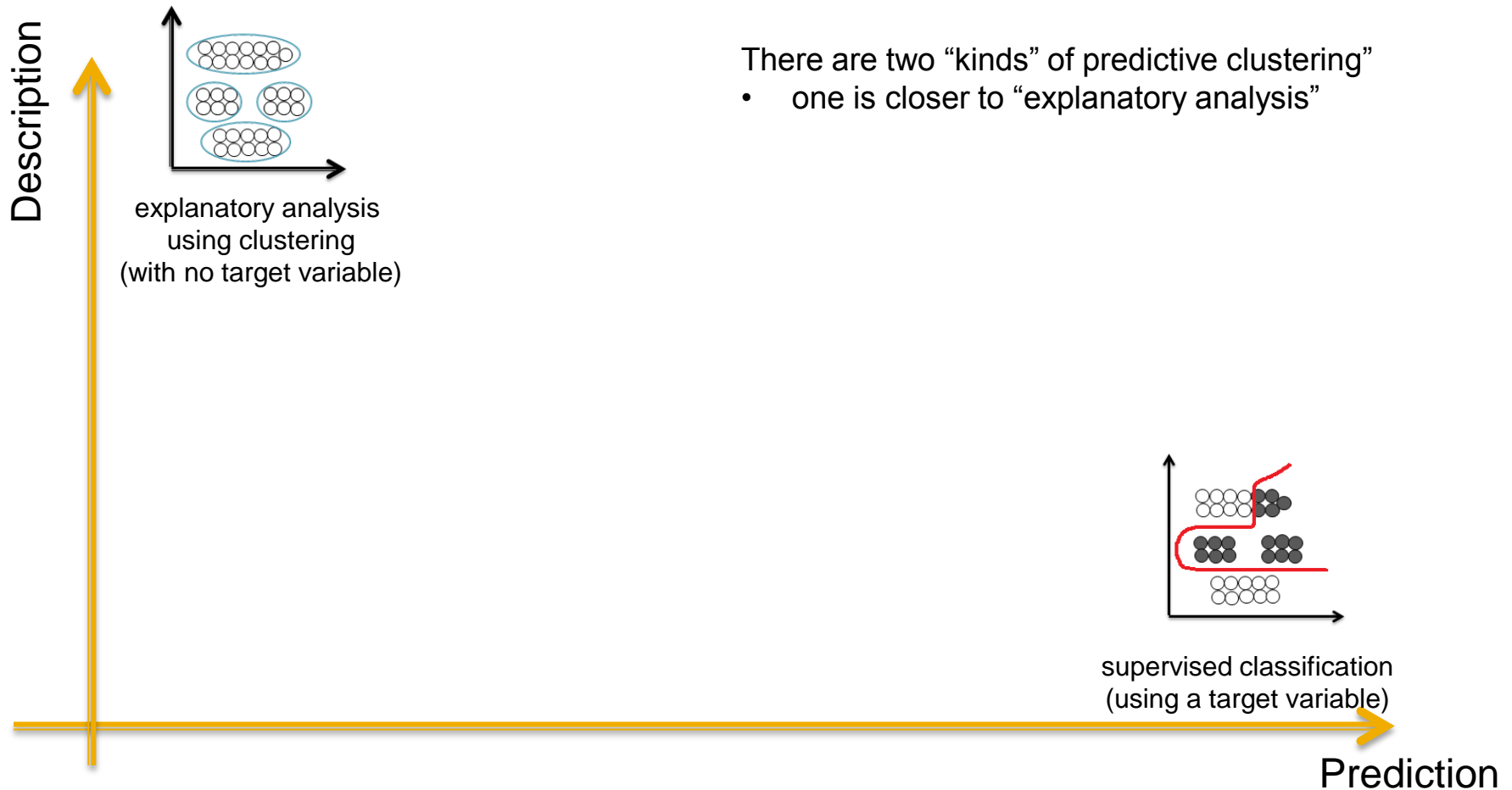
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



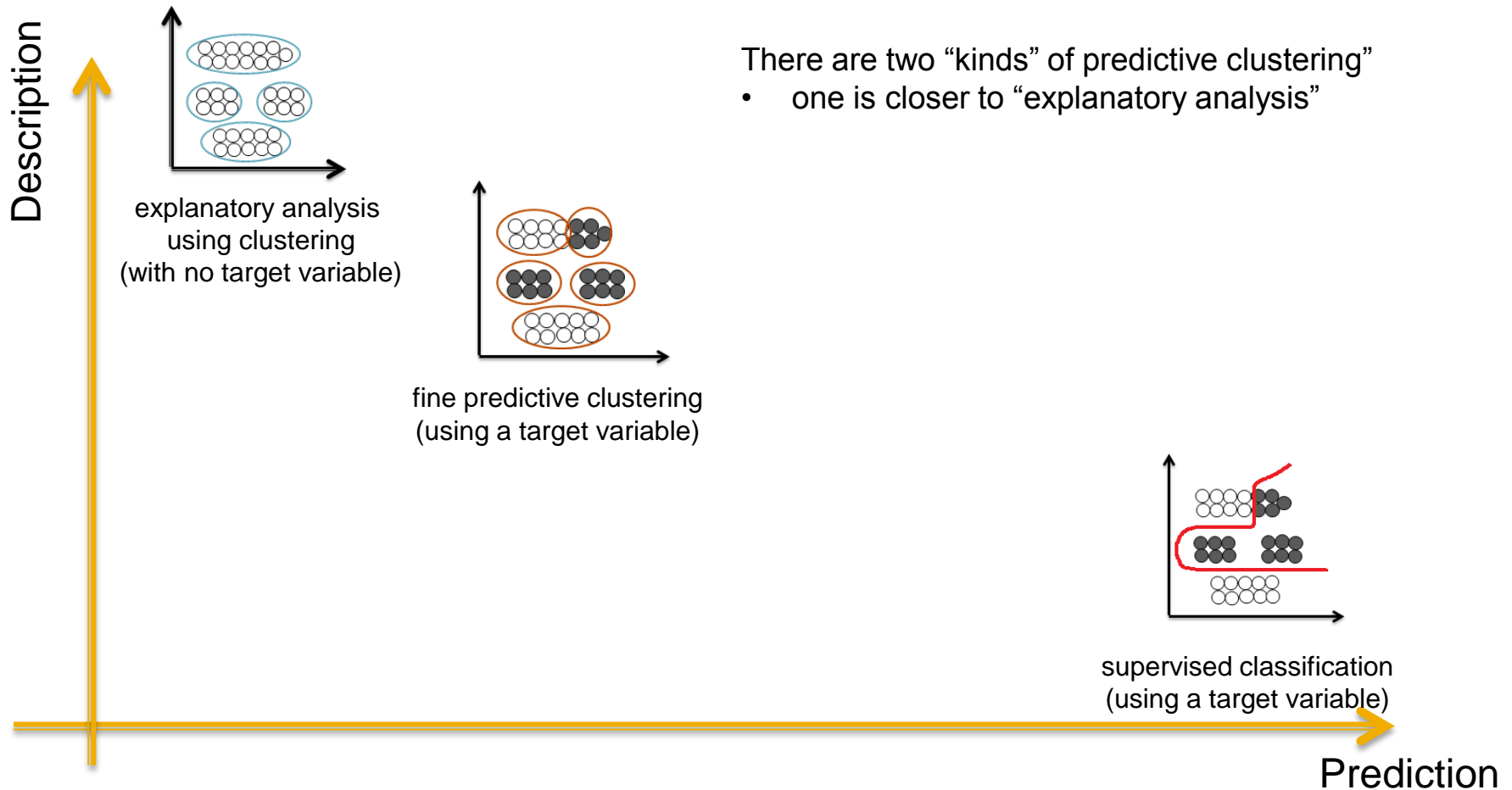
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



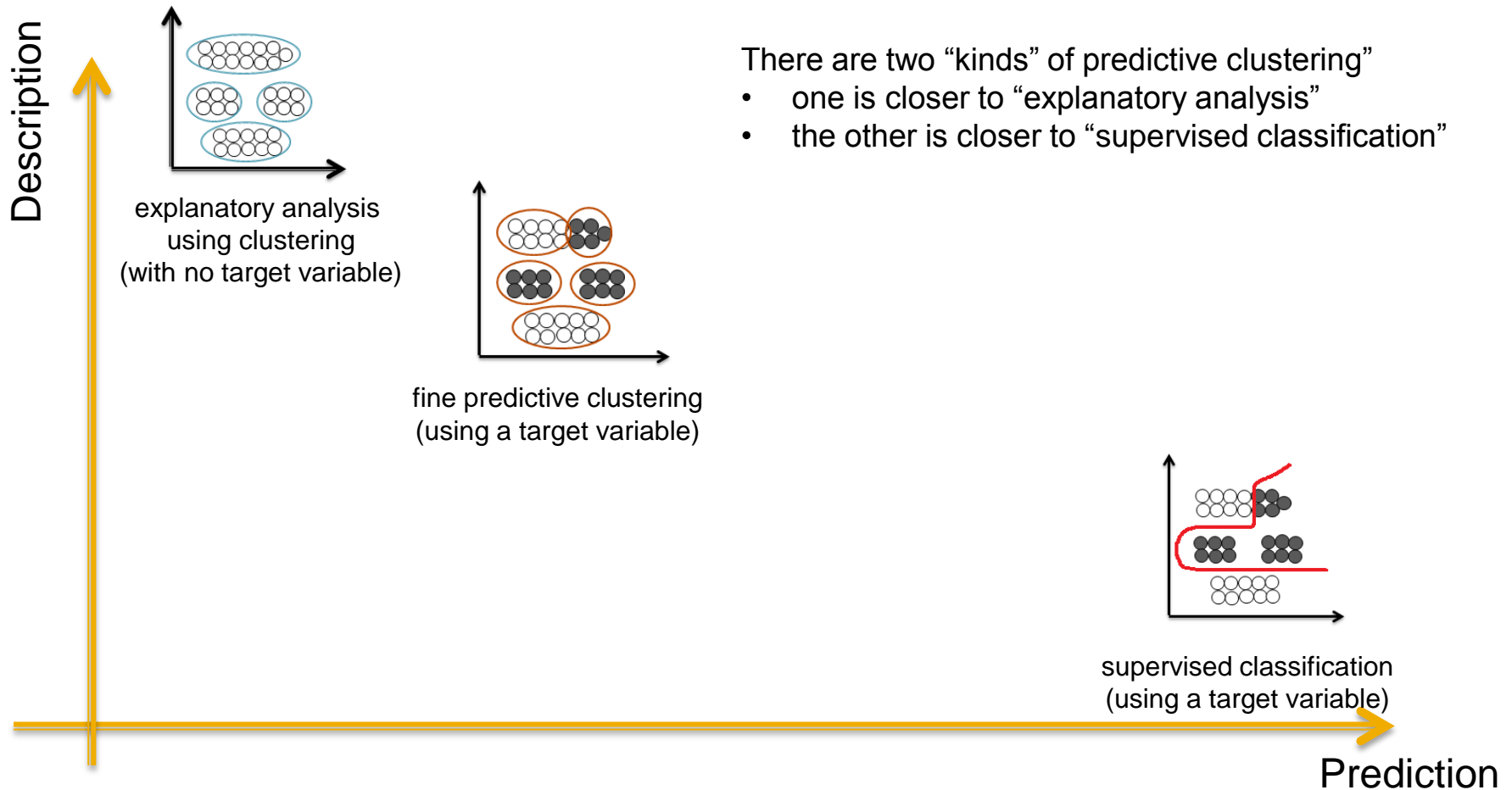
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



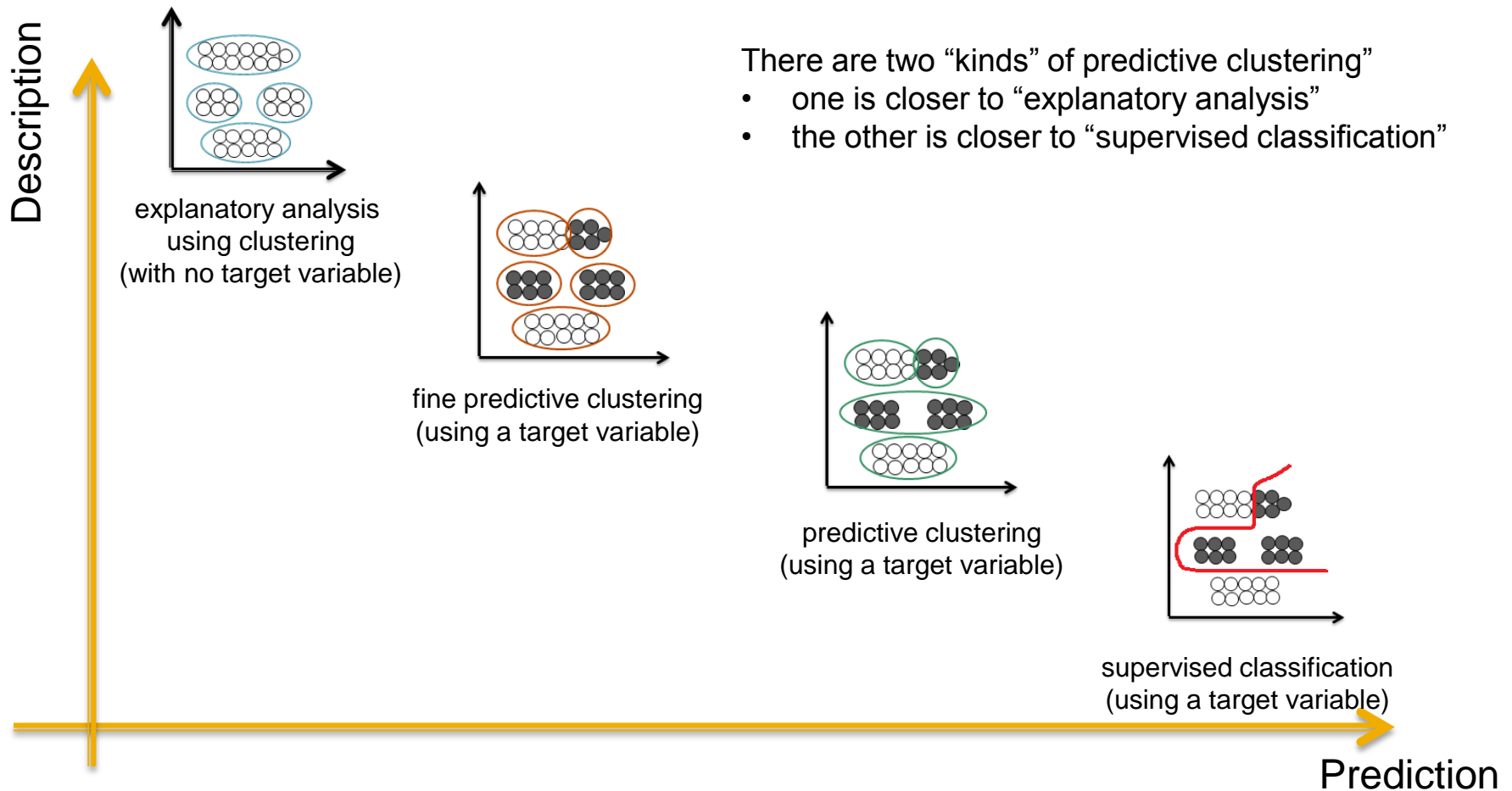
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



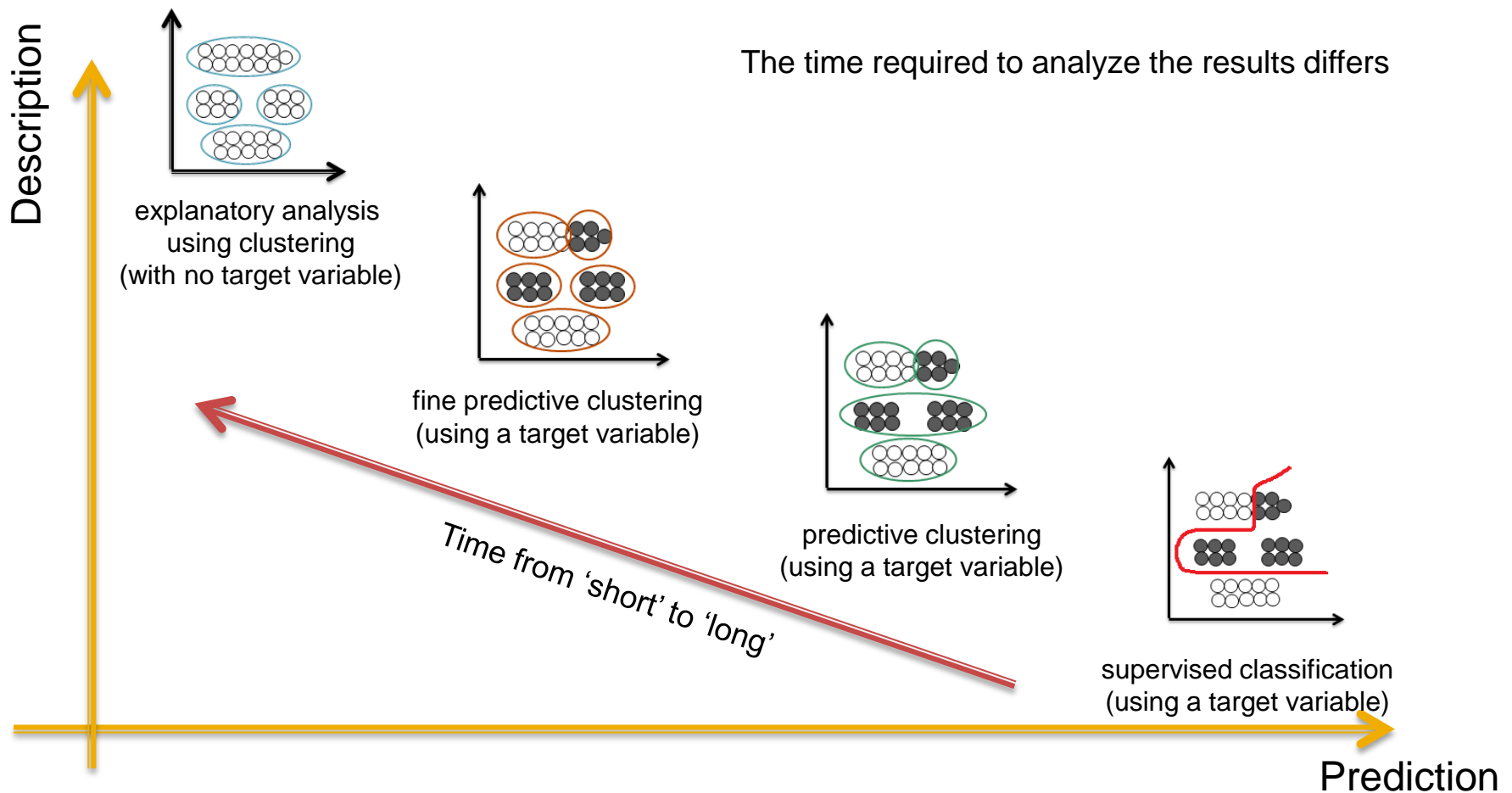
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



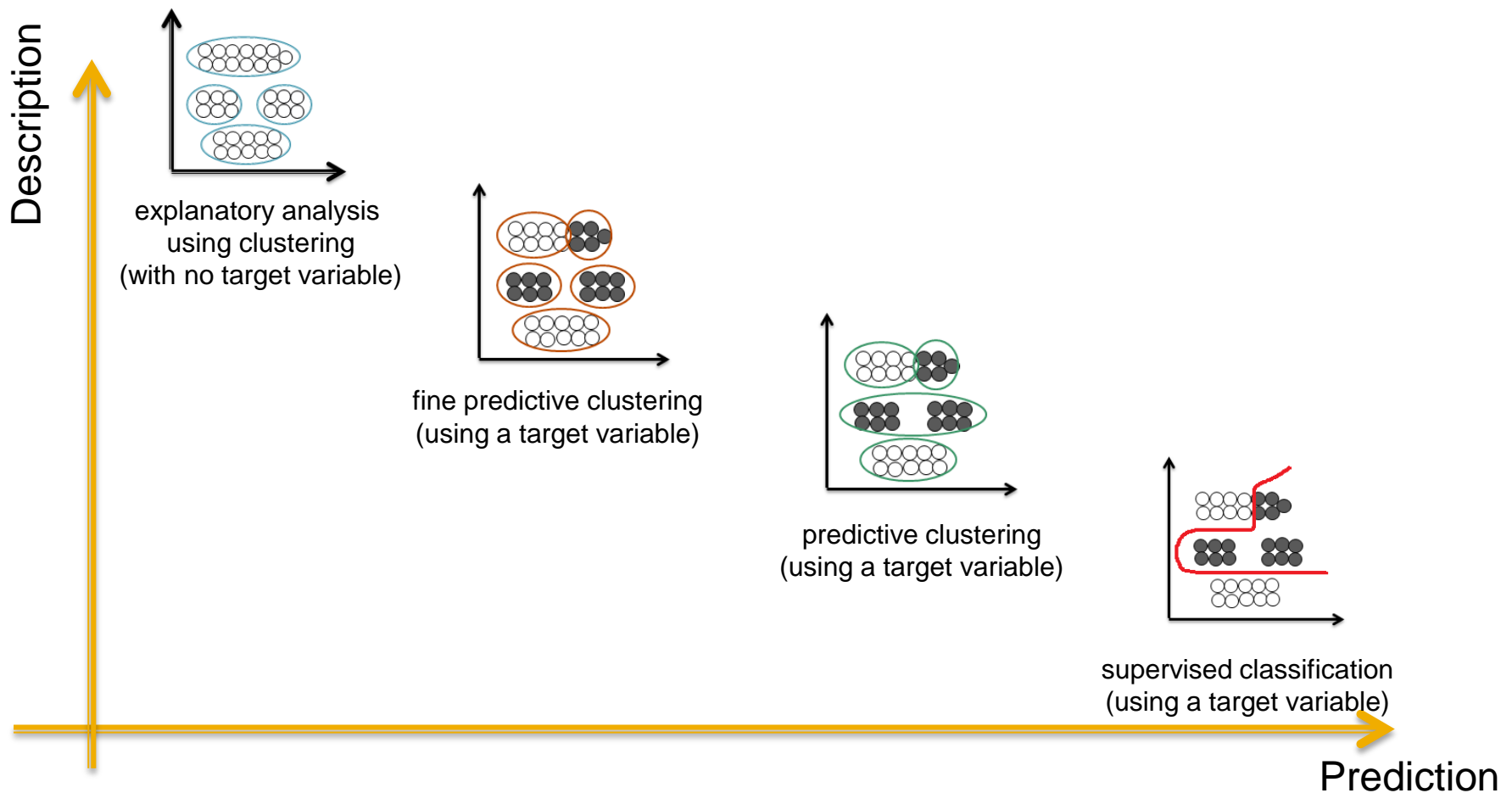
Predictive Clustering

From explanatory analysis to supervised classification through “Predictive Clustering”



Which tool to use...

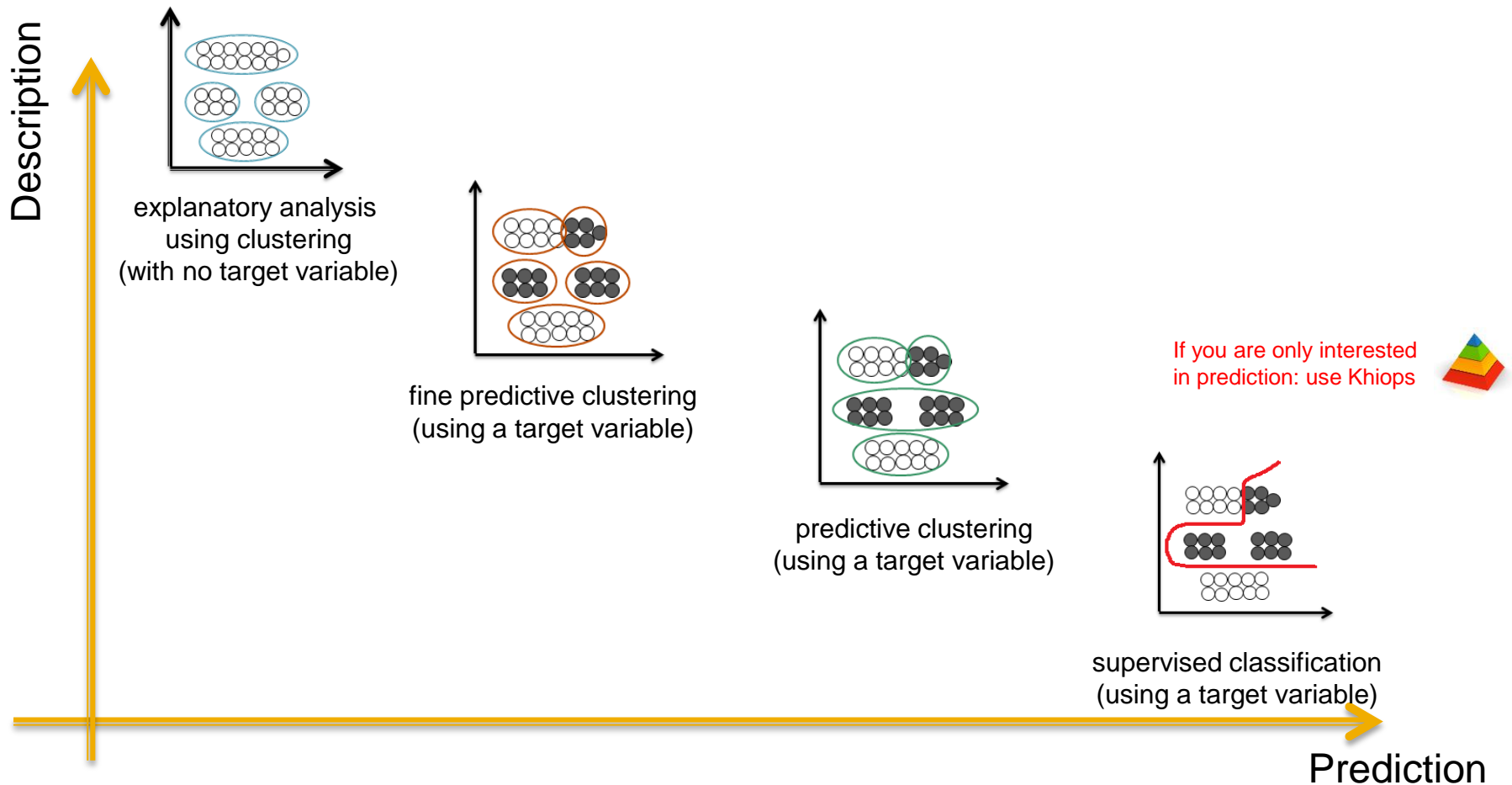
in the Khiops Family ☺



Predictive Clustering

Which tool to use...

in the Khiops Family ☺



Predictive Clustering

Which tool to use...

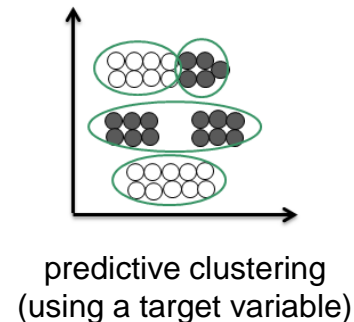
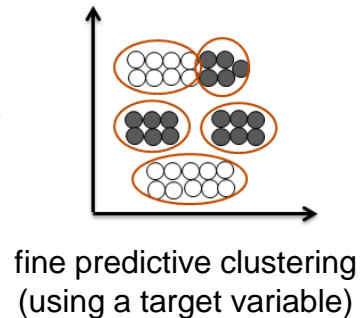
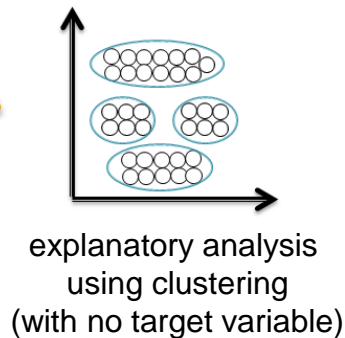
If you are only interested in description:

- if you may pose your question as CoClustering problem: use Khiops CoClustering
- use an appropriate tool (data dependent)
- or try Khiops Ennéade

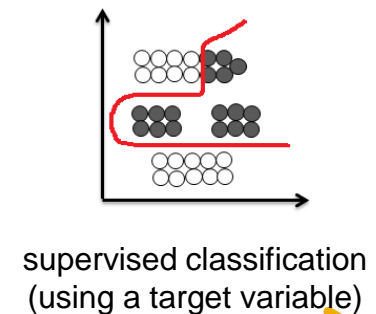


in the Khiops Family ☺

Description



If you are only interested
in prediction: use Khiops



Prediction

Predictive Clustering

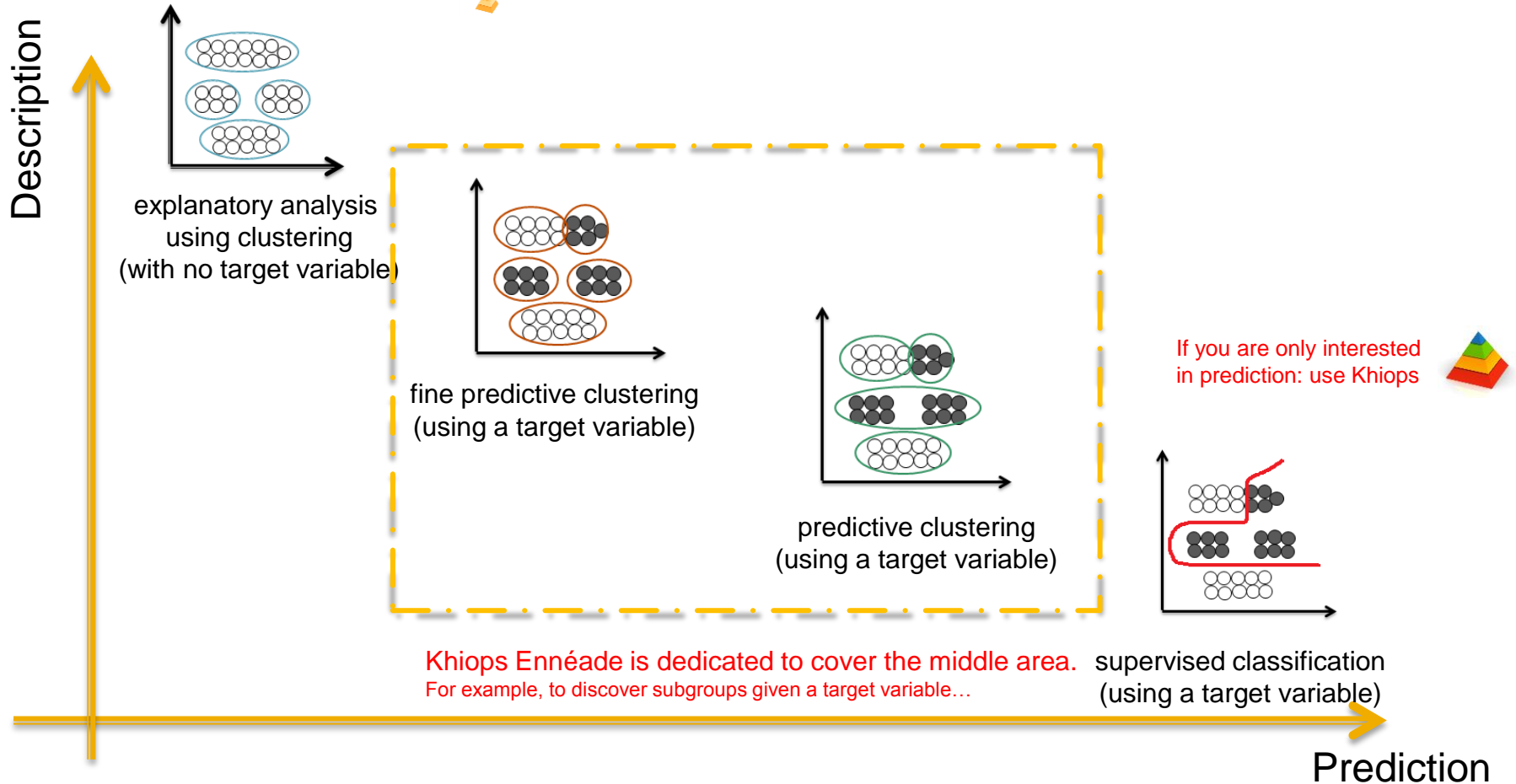
Which tool to use...

If you are only interested in description:

- if you may pose your question as CoClustering problem: use Khiops CoClustering
- use an appropriate tool (data dependent)
- or try Khiops Ennéade



in the Khiops Family ☺



Khiops Ennéade

Part III

Algorithms in Khiops Ennéade

Predictive K-Means

The tools run a modified version of the K-means :

- Different initializations has been suggested to address 'fine predictive clustering' or 'predictive clustering'
 - "Une méthode supervisée pour initialiser les centres des K-moyennes", Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols, in Extraction et Gestion des Connaissances (EGC), Reims, 2016
 - "Initialisation des k-moyennes à l'aide d'une décomposition supervisée des classes" Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols, Congrès de la Société Française de Classification (SFC)
 - "An Initialization Scheme for Supervized K-means" Vincent Lemaire, Oumaima Alaoui Ismaili, Antoine Cornuéjols, in International Joint Conference on Neural Networks (IJCNN), IEEE, Ireland, 2015
- The preprocessing has been modified
 - "Supervised pre-processings are useful for supervised clustering", Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols, in the Springer Series "Studies in Classification, Data Analysis, and Knowledge Organization" which will be released in 2015
- The criteria used to find K has been modified
 - "Evaluation of predictive clustering quality", Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols, in Model-Based Clustering and Classification (MBC2), 2016.
 - "Un critère d'évaluation pour les K-moyennes prédictives" Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols, in Extraction et Gestion des Connaissances (EGC), Grenoble, 2017.
- and some others aspects not yet published...

You may find more details in the PhD : «Clustering prédictif - Décrire et Prédire simultanément», Ouaima Alaoui Ismaili, PhD université Paris Sacaly (AgroParisTech)

Predictive K-Means

That means that the tools

- has **an auto adaptation capacity** to choose :
 - the right initialization
 - the right preprocessing and
 - the right criterion

...depending on if you ask :

- a “clustering” (in this case a classical K-means)
- or a “predictive clustering” (in this case our modified K-means)

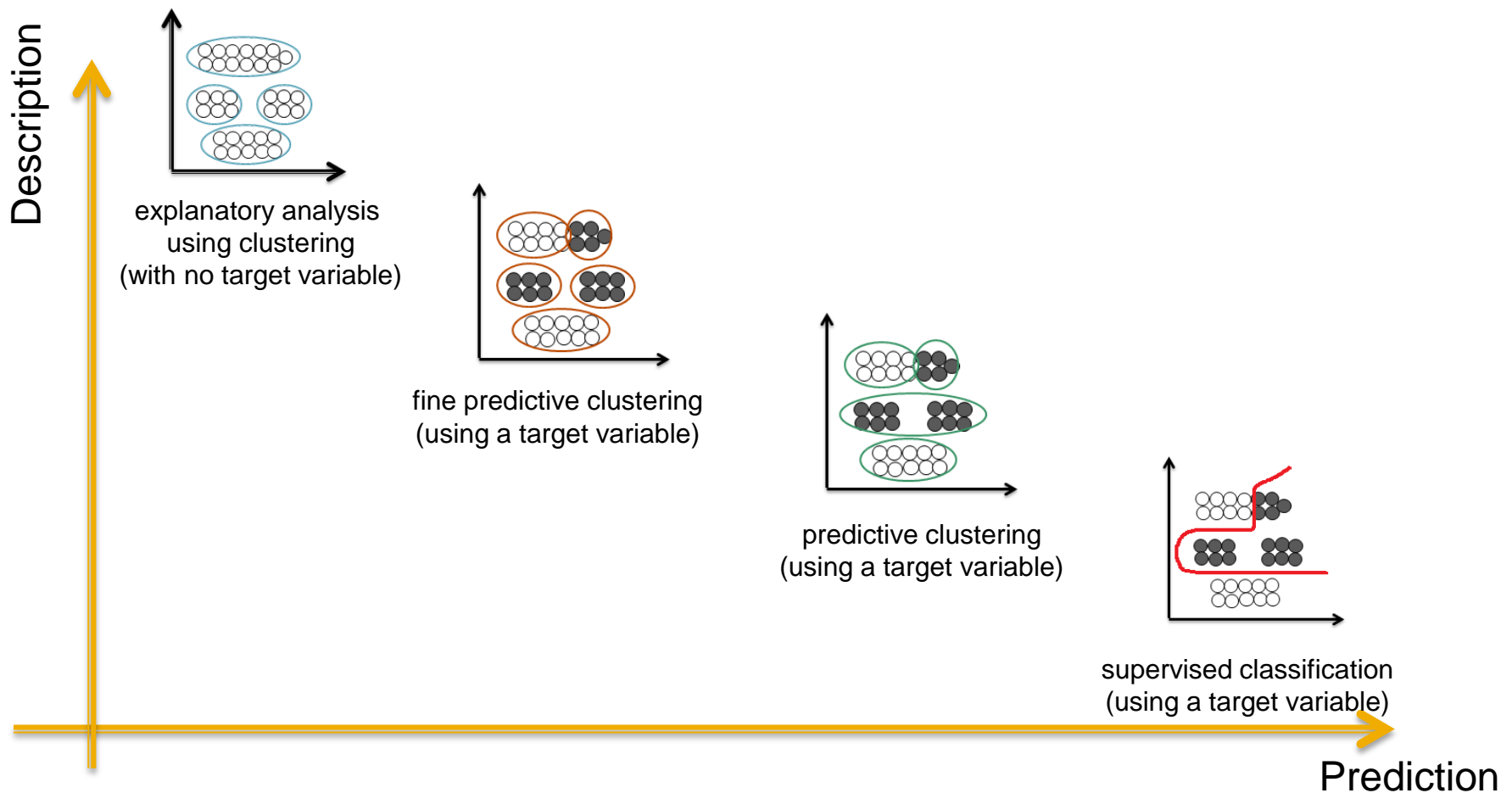
by indicating or not the presence of a target variable (in the Analysis parameters tab when using the tool)

Khios Ennéade

Part IV

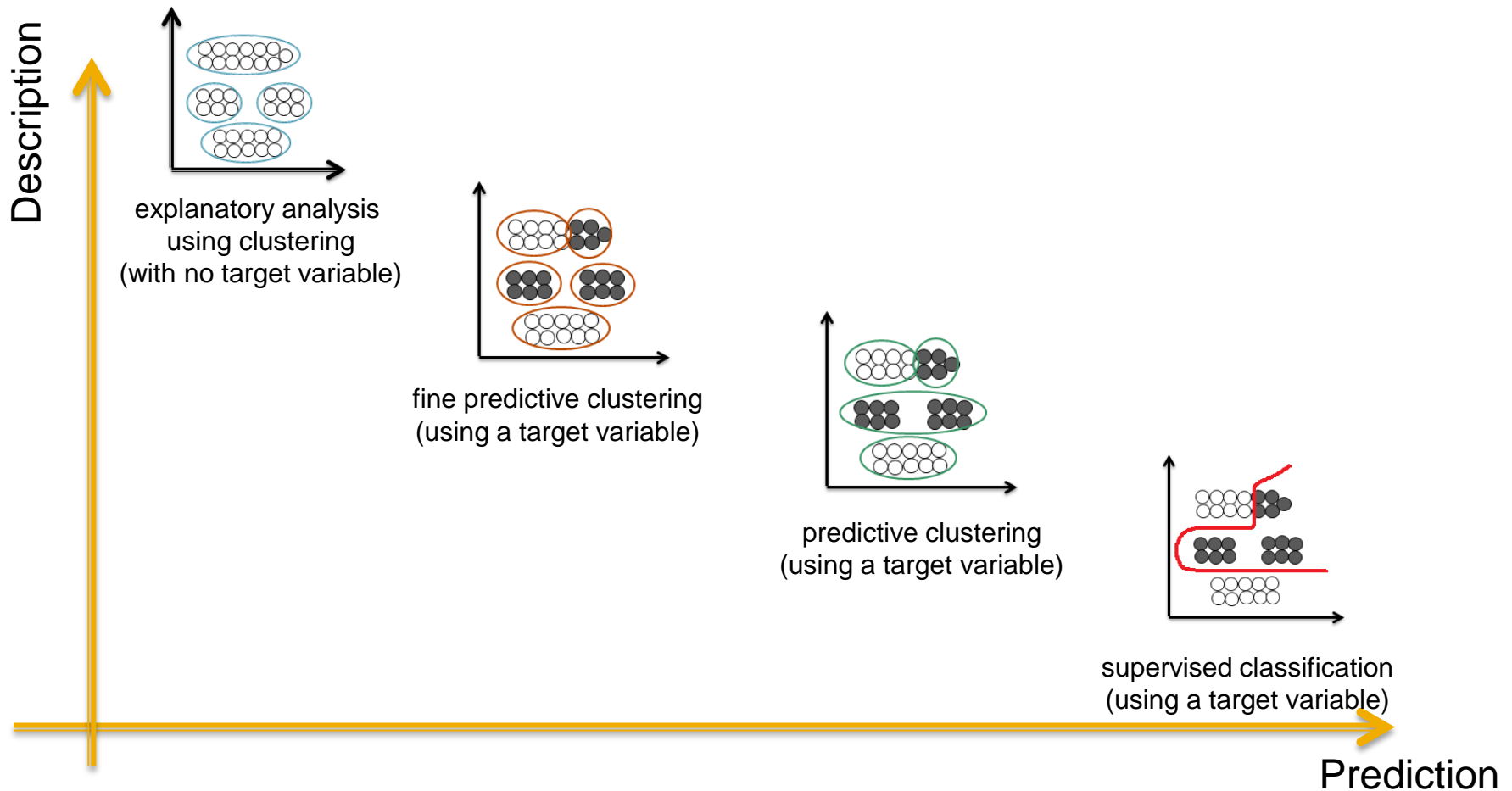
The different Criteria

“Predictive Clustering” vs “Criteria”



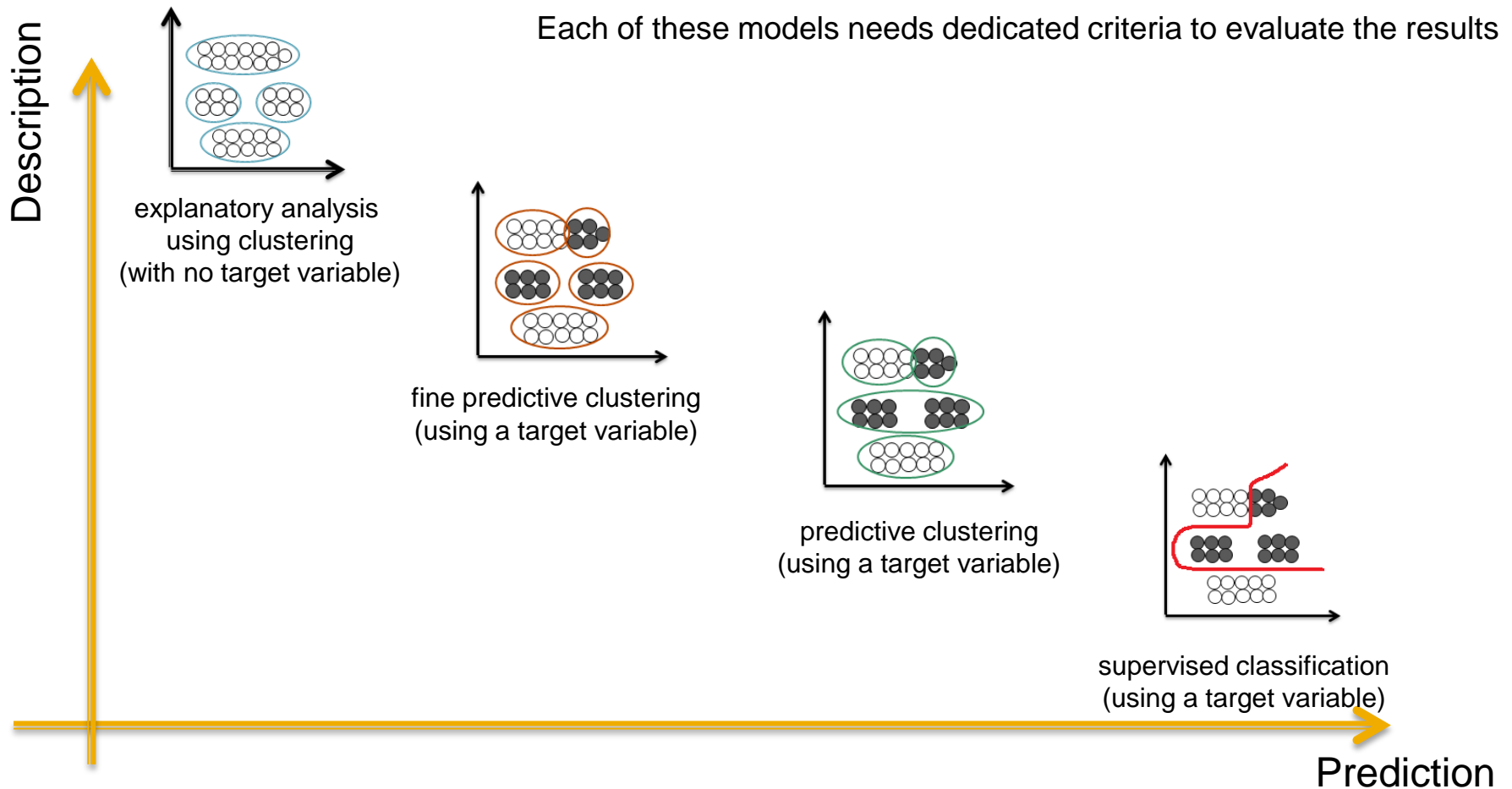
Predictive Clustering

"Predictive Clustering" vs "Criteria"



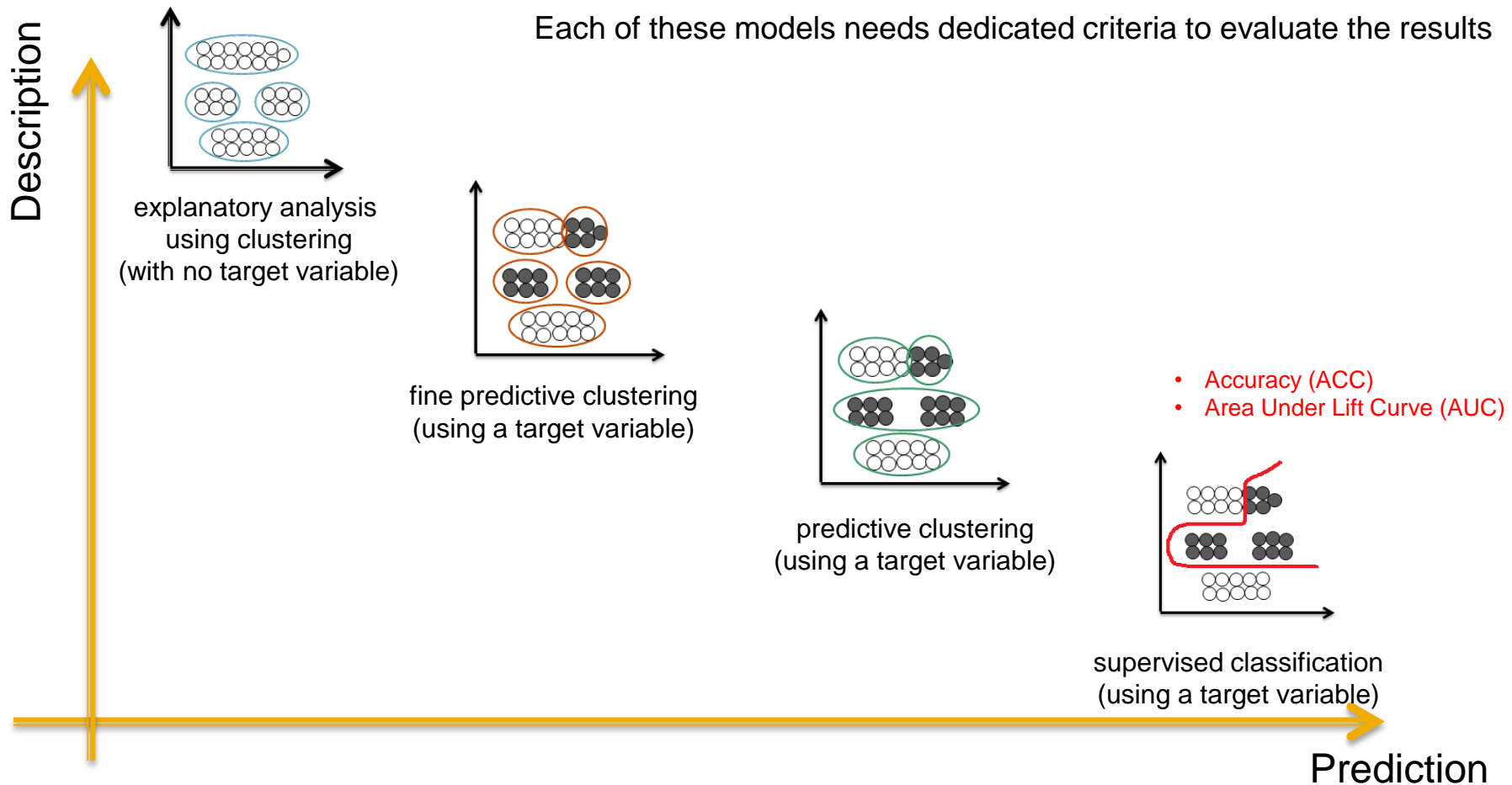
Predictive Clustering

"Predictive Clustering" vs "Criteria"



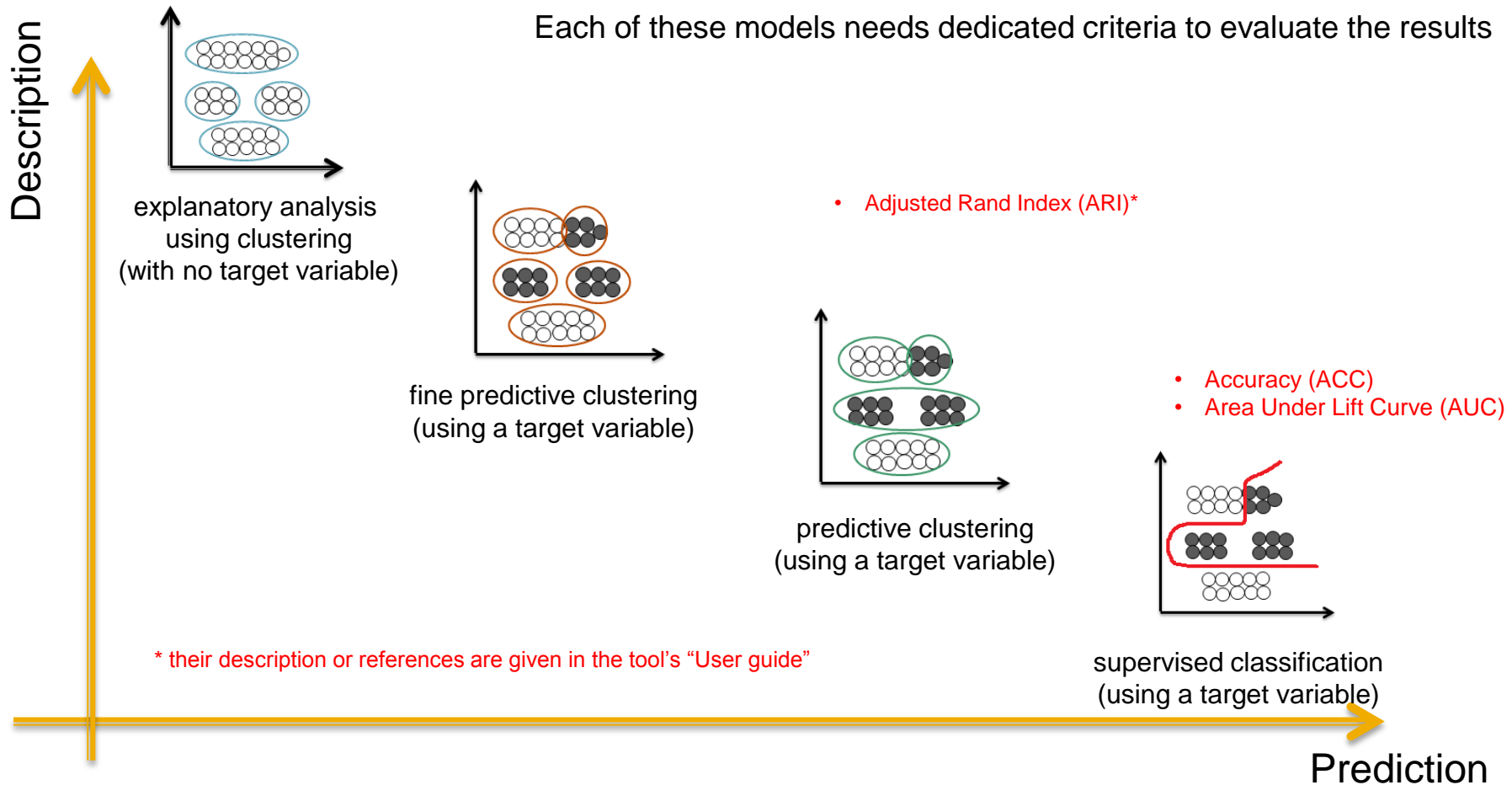
Predictive Clustering

“Predictive Clustering” vs “Criteria”



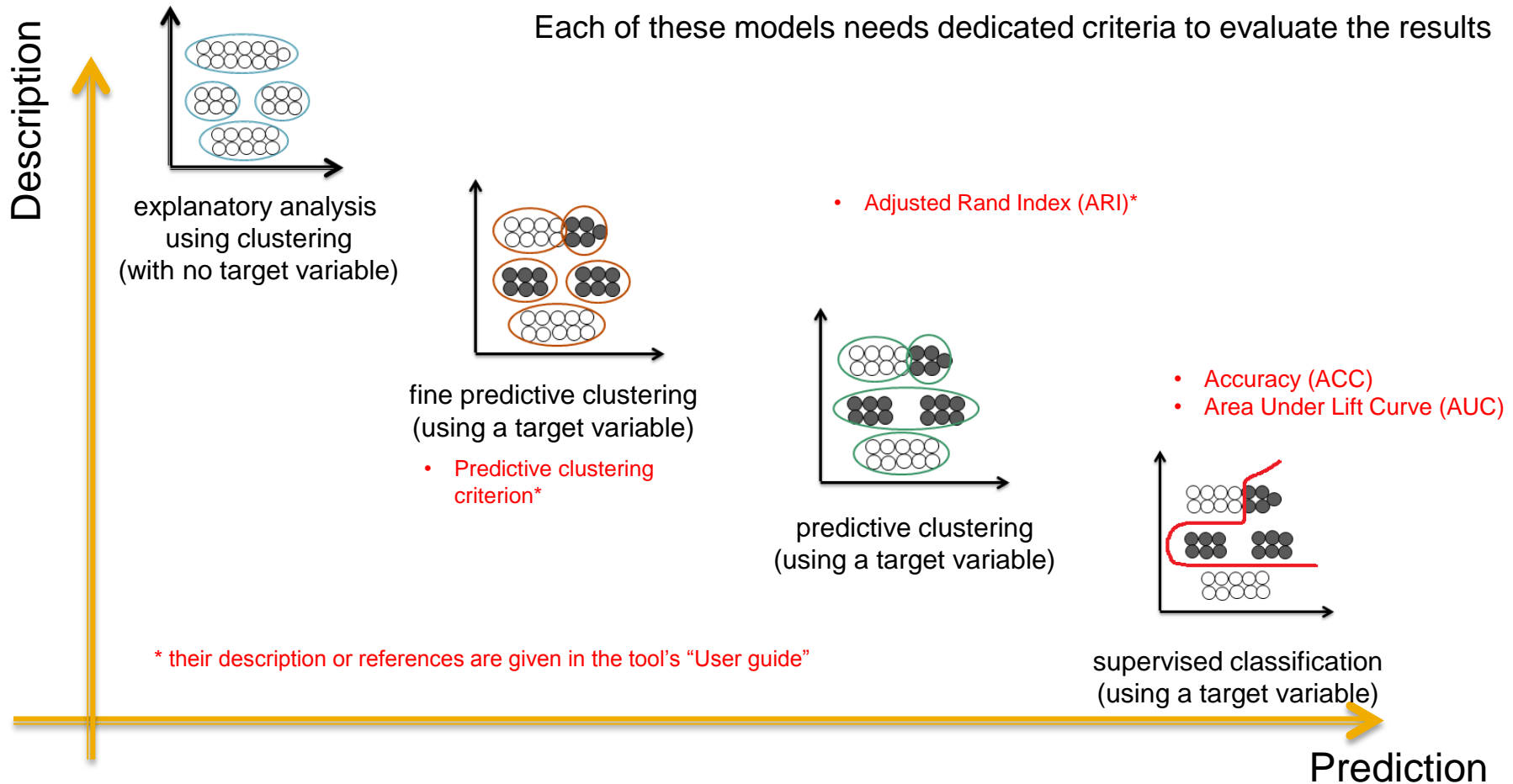
Predictive Clustering

“Predictive Clustering” vs “Criteria”



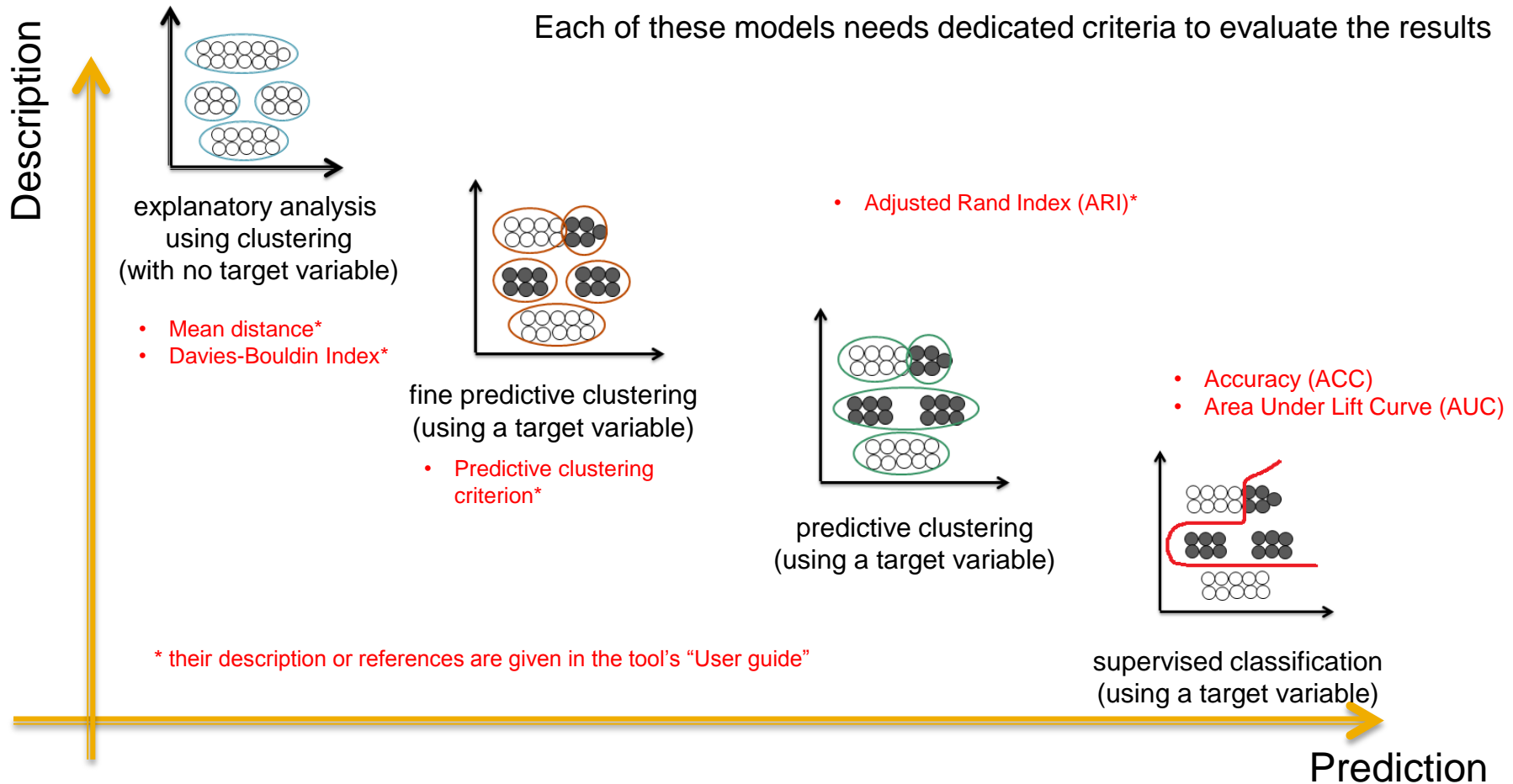
Predictive Clustering

“Predictive Clustering” vs “Criteria”



Predictive Clustering

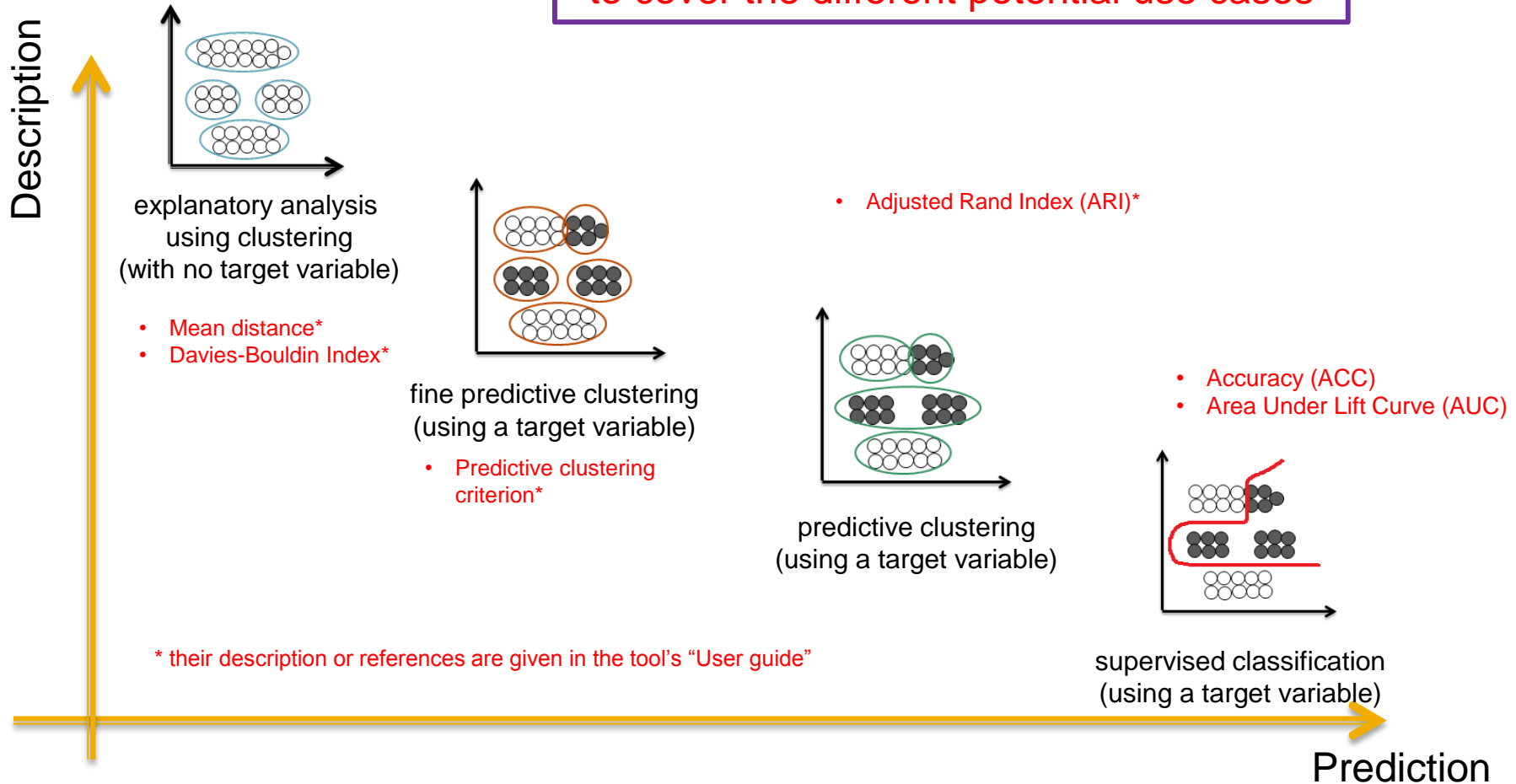
“Predictive Clustering” vs “Criteria”



Predictive Clustering

“Predictive Clustering” vs “Criteria”

Khiops Ennéade contains all these criteria to cover the different potential use cases



Predictive Clustering

Range of values of these criteria

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+ ∞]	minimized
Davies-Bouldin index	[0.0,+ ∞]	minimized
Mean distance	[min**,max**]	minimized

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+ ∞]	minimized
Davies-Bouldin index	[0.0,+ ∞]	minimized
Mean distance	[min**,max**]	minimized

Khiops Ennéade

Part V

Determining the number of clusters

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized

You may decide on your own the number of clusters because :

- this number corresponds to the time you have to analyze the results
- this number corresponds to the number of sub profiles you will want to address
-

Otherwise ...

Predictive Clustering

How to determine the value of K ?

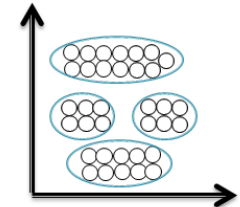
Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized

If you want to perform an explanatory analysis:

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



explanatory analysis
using clustering
(with no target variable)

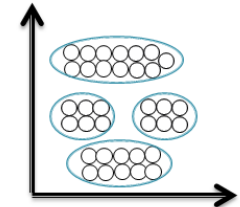
If you want to perform an explanatory analysis:

- That means that you do not have (or do not want to use) a 'target variable'

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



explanatory analysis
using clustering
(with no target variable)

If you want to perform an explanatory analysis:

- That means that you do not have (or do not want to use) a 'target variable'
- In this case run the tool for different values of K
- And chose the value which minimizes the 'Davies-Bouldin index' in train

Predictive Clustering

How to determine the value of K ?

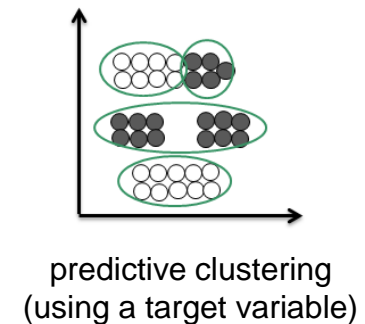
Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized

If you want to perform a predictive clustering:

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



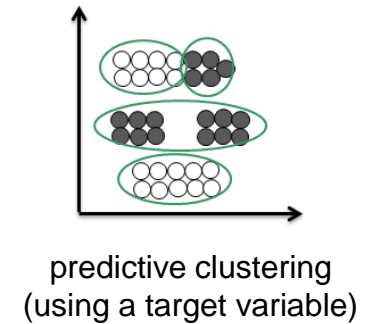
If you want to perform a predictive clustering:

- but you are **mainly** interested by the **classification** performance

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



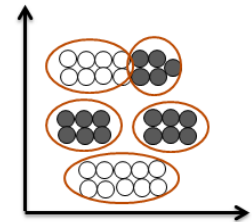
If you want to perform a predictive clustering:

- but you are **mainly** interested by the **classification** performance
 - In this case run the tool, and set it to try different values of K
 - then, choose the value which maximizes the 'Adjusted Rand Index (ARI)' in train

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



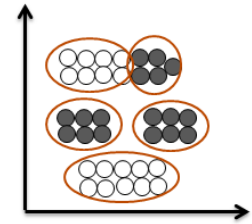
fine predictive clustering
(using a target variable)

If you want to perform a predictive clustering:

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



fine predictive clustering
(using a target variable)

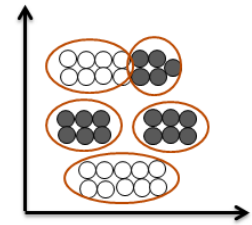
If you want to perform a predictive clustering:

- but you are mainly interested in a **tradeoff** between classification performance and description

Predictive Clustering

How to determine the value of K ?

Criterion	range of values	Has to be ...
Accuracy	[0.0,1.0]	maximized
Area Under Lift Curve (AUC)	[0.0,1.0]	maximized
Adjusted Rand Index (ARI)	[0.0,1.0]	maximized
Predictive Clustering (index)	[0.0,+∞]	minimized
Davies-Bouldin index	[0.0,+∞]	minimized
Mean distance	[min**,max**]	minimized



fine predictive clustering
(using a target variable)

If you want to perform a predictive clustering:

- but you are mainly interested in a **tradeoff** between classification performance and description
 - In this case, run the tool, and set it to try different values of K
 - then, choose the value which minimizes the 'Predictive Clustering (index)' in train

Khiops Ennéade

Part VI

How to define a profile?

Predictive Clustering

How to define a profile

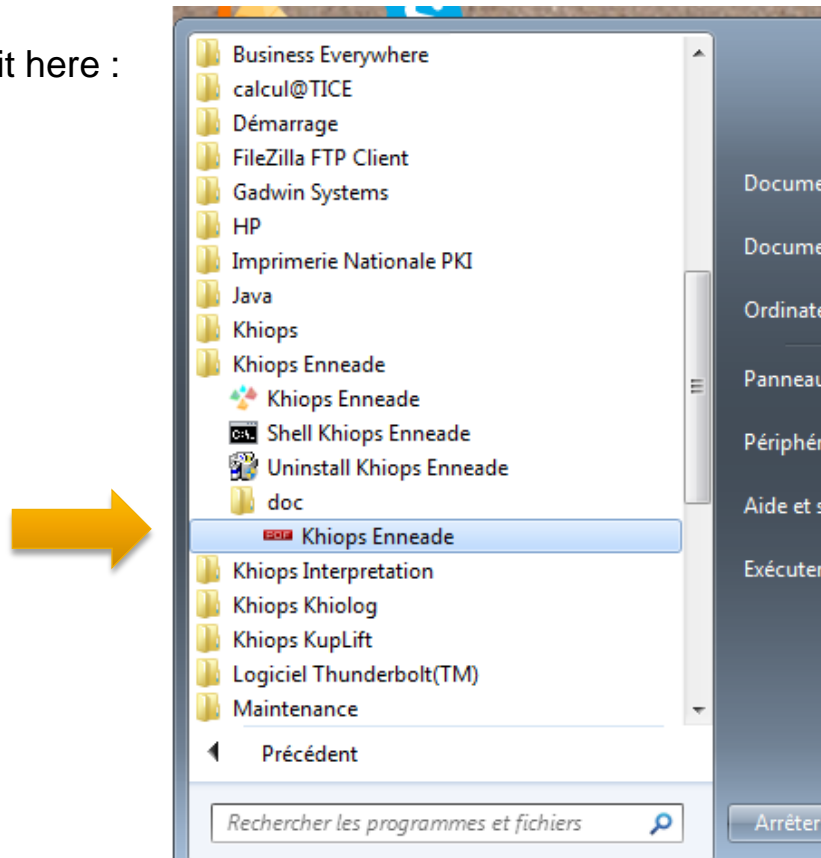
Preamble

Predictive Clustering

How to define a profile

Note :

- Remember that all the descriptions of the tables are given in the “Khiops Enneade Guide” which has been put on your computer during the installation of the tool.
- You may find it here :



In next slides we are going to use a part of them to show how to define the profile of a cluster

Predictive Clustering

How to define a profile

First at all :

- Khiops Ennéade performs a modified version of “K-means Algorithm”.

Predictive Clustering

How to define a profile

First at all :

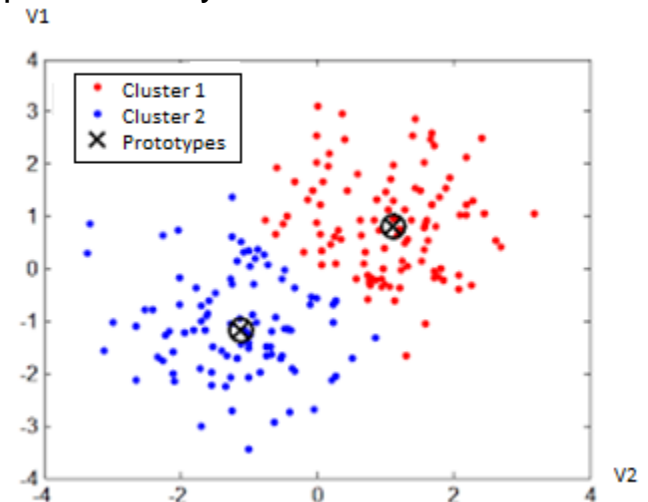
- Khiops Ennéade performs a modified version of “K-means Algorithm”.
- The produced model, after the training step, is a set of K clusters:
 - a cluster is a set of examples in which each example is closer (more similar) to the prototype that defines the cluster than to the prototype of any other cluster
 - a prototype is the mean vector (profile) of examples belonging to the cluster.

Predictive Clustering

How to define a profile

First at all :

- Khiops Ennéade performs a modified version of “K-means Algorithm”.
- The produced model, after the training step, is a set of K clusters:
 - a cluster is a set of examples in which each example is closer (more similar) to the prototype that defines the cluster than to the prototype of any other cluster
 - a prototype is the mean vector (profile) of examples belonging to the cluster.
- For example in the easy case here, where examples are represented by two explanatory variables (V1, V2), we see:
 - K=2
 - examples belonging to Cluster 1 are represented in blue*
 - examples belonging to Cluster 1 are represented in red*
 - Black crows represent the two prototypes



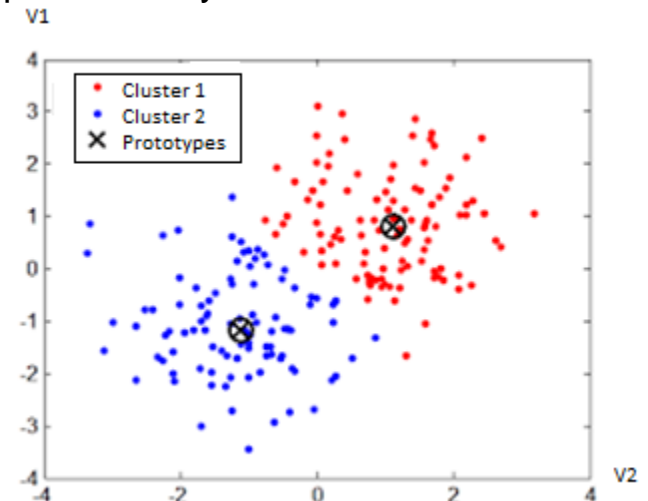
* (whatever is their value for the target variable)

Predictive Clustering

How to define a profile

First at all :

- Khiops Ennéade performs a modified version of “K-means Algorithm”.
- The produced model, after the training step, is a set of K clusters:
 - a cluster is a set of examples in which each example is closer (more similar) to the prototype that defines the cluster than to the prototype of any other cluster
 - a prototype is the mean vector (profile) of examples belonging to the cluster.
- For example in the easy case here, where examples are represented by two explanatory variables (V1, V2), we see:
 - K=2
 - examples belonging to Cluster 1 are represented in blue*
 - examples belonging to Cluster 1 are represented in red*
 - Black crows represent the two prototypes
- Then Khiops Ennéade gives detailed results on the K prototypes allowing the user to understand the mean “profile” of each cluster.
 - This results are presented in many tables in the Train (or Test) Evaluation Report



* (whatever is their value for the target variable)

Predictive Clustering

How to define a profile

Several tables may be viewed as histograms (which may be considered as profiles).

- Histograms which describe each cluster individually, along the different explanatory variables
- Histograms which describe a given value* of an explanatory variable along the different clusters

We suggest to define a profile as a comparison of given histogram compared to a reference.

For example :

- the “histogram” of a given cluster, compared to the “histogram” of the global population
- the “histogram” of a given cluster, compared to the “histogram” of another cluster

* interval for numerical attributes or group of values for categorical attributes when asking a predictive clustering to the tool

Predictive Clustering

How to define a profile

At first, we use the table “gravity center”

Cluster	Frequency	Coverage	Iris-versicolor	Iris-setosa	Iris-virginica
Cluster 1	48	0.32	0.9375	0	0.0625
Cluster 3	52	0.346667	0.0961538	0	0.903846
Cluster 2	50	0.333333	0	1	0
Total	150	1	0.333333	0.333333	0.333333

This table contains for every cluster, from the left column to the right:

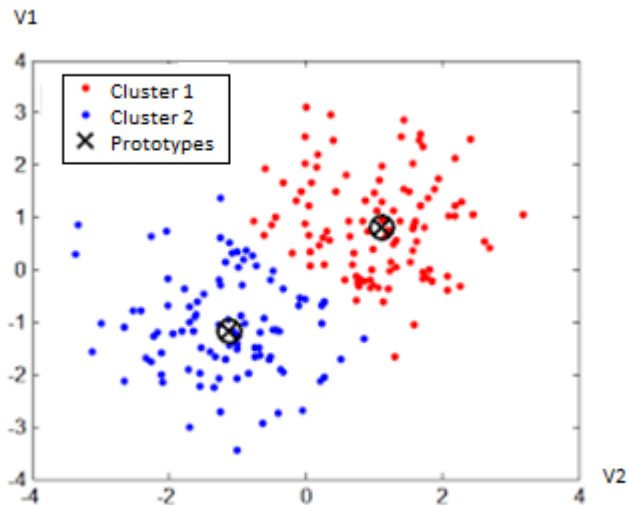
- Cluster: id of the cluster
- Frequency: Number of examples belonging to the cluster
- Coverage: Percentage of examples belonging to the cluster.
- C columns (one column per Class if a Target Variable has been specified):
 - Percentage of the examples belonging to the cluster ‘id’ and of the Class where in the name is the header of the table

Predictive Clustering

How to define a profile

Table “Gravity centers” in the Evaluation Report (Train or Test) :

For each cluster:



note: the draws on the figure are illustrative

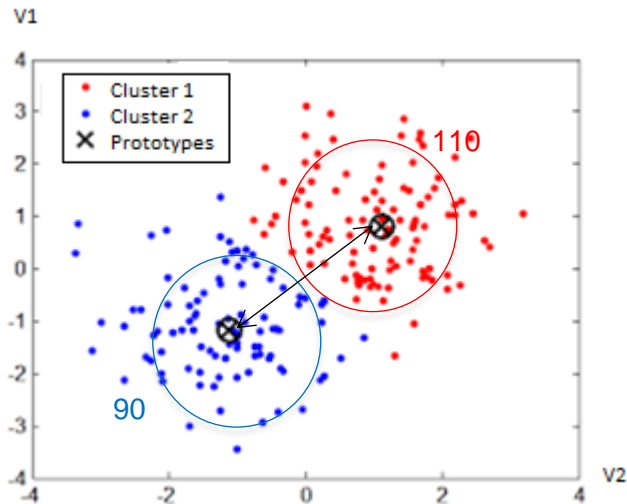
Predictive Clustering

How to define a profile

Table “Gravity centers” in the Evaluation Report (Train or Test) :

For each cluster:

- Frequency: Number of examples belonging to the cluster
- Coverage: percentage of examples belonging to the cluster



note: the draws on the figure are illustrative

Predictive Clustering

How to define a profile

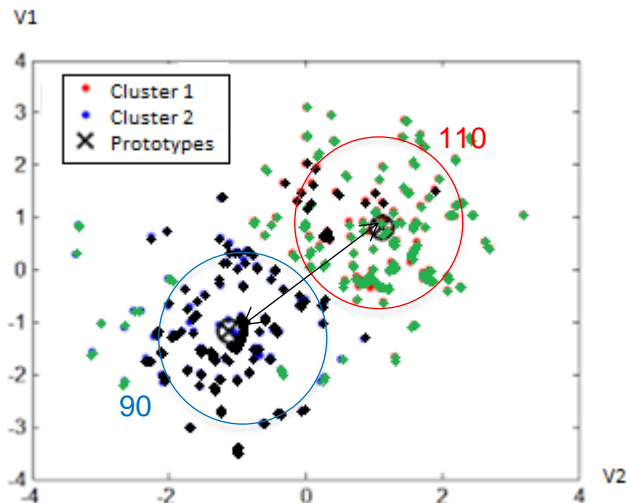


Table “Gravity centers” in the Evaluation Report (Train or Test) :

For each cluster:

- Frequency: Number of examples belonging to the cluster
- Coverage: percentage of examples belonging to the cluster
- For all the different values of the target variable:
 - Percentage of the examples having this value

For example (here the target value has two values, green or black):

- Cluster 1 contains 90% of examples with a target value as “green” and 10% as “black”
- Cluster 2 contains 20% of examples with a target value as “green” and 80% as “black”

note: the draws on the figure are illustrative

Predictive Clustering

How to define a profile

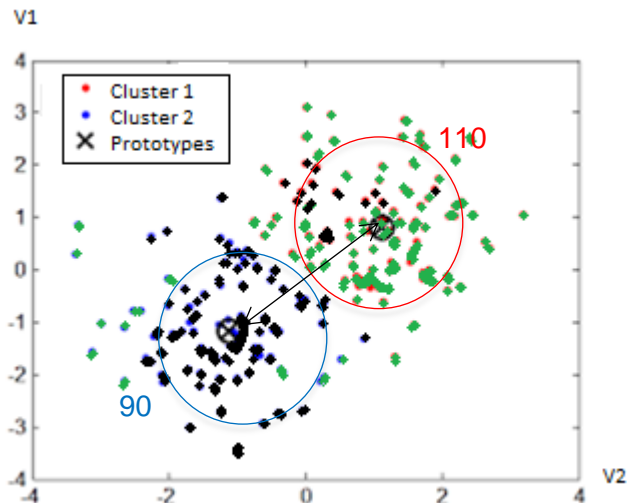


Table “Gravity centers” in the Evaluation Report (Train or Test) :

For each cluster:

- Frequency: Number of examples belonging to the cluster
- Coverage: percentage of examples belonging to the cluster
- For all the different values of the target variable:
 - Percentage of the examples having this value

Note : Using this information, we are able to predict the “target value” (of the target variable) of an example.

- The predicted class will be the majority class present in a cluster at the end of the training step
- The probability of this predicted class will be the percentage of this majority class
- For example when deploying the clustering model, example closer to prototype “1” will be predicted as “green” with a probability of 0.9 (and therefore “black” with a probability of 0.1).

This information allow to compute the Accuracy, AUC, ... given in the Table “Predictors detailed performance” and the Table “Confusion Matrix” (for the axis prediction).

Predictive Clustering

How to define a profile

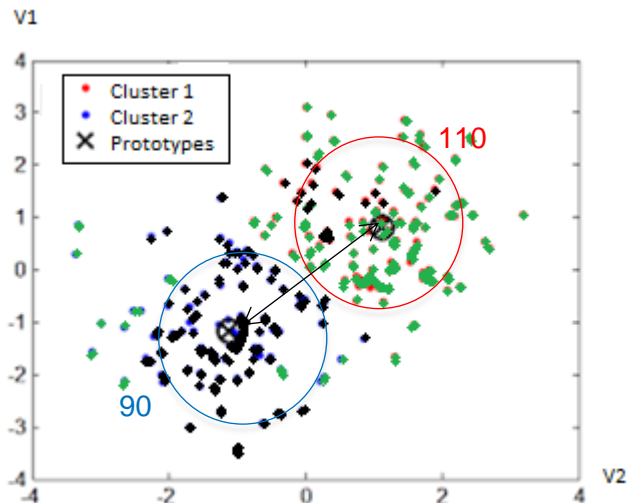


Table “Gravity centers” in the Evaluation Report (Train or Test) :

For each cluster:

- Frequency: Number of examples belonging to the cluster
- Coverage: percentage of examples belonging to the cluster
- For all the different values of the target variable:
 - Percentage of the examples having this value

Frequency and coverage give indication about the “size” of the obtained clusters

The last information gives the repartition of the values of the target variable in each cluster. This allows you to see if the tool has been able to create a cluster with a high percentage of a given value of the target variable.

Predictive Clustering

How to define a profile

Example on Iris Database ...

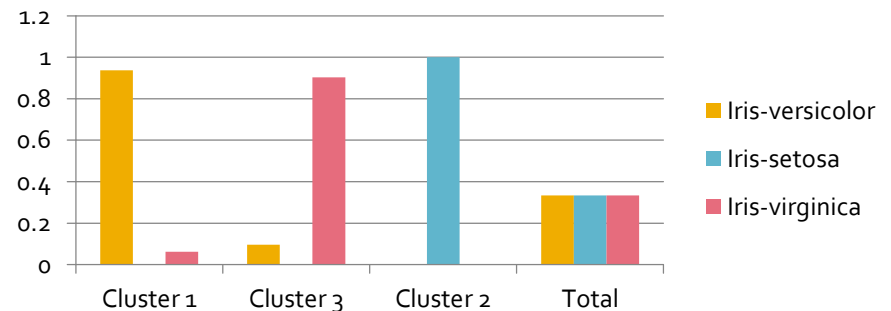
We have run the tool for 3 clusters and the “Class” as target variable, 100% of the examples for the Train and we kept all the parameters by default.

1) Firstly at all we have a look on the table “gravity center”

Cluster	Frequency	Coverage	Iris-versicolor	Iris-setosa	Iris-virginica
Cluster 1	48	0.32	0.9375	0	0.0625
Cluster 3	52	0.346667	0.0961538	0	0.903846
Cluster 2	50	0.333333	0	1	0
Total	150	1	0.333333	0.333333	0.333333

We may define at first the profiles of the clusters regarding the target variable “Class”. Using your favorite tool (excel, ..) you may easily create the corresponding histogram. With this view we understand that the clusters 1, 2, 3 are respectively more related to Iris-versicolor, Iris-virginica and Iris-setosa. Each cluster could be compared to the global population (Total).

Knowing this repartition we may continue our analysis...



Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

Reminder (see more details in the User Guide of the tool) : in this table we have for every explanatory variable used after the preprocessing step, from the left column to the right

- Column 1 - Var name: Name of the variable
- Column 2 - Modality/Interval: Intervals or Group name
- Column 2+1 to 2+k (k is the number of clusters): Percentage of instances belonging to "Cluster k" for which the variable of the Column 1 ("Var name") takes its values in the Intervals or Group name of the column 2 ("Modality/Interval") – the percentage is computed over all Intervals or Group name of the variable of the Column 1 ("Var name").
- Column "global" : Percentage of instances (whatever the cluster) for which the variable of the Column 1 ("Var name") takes its value in the Interval or Group name of the column 2 ("Modality/Interval")

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

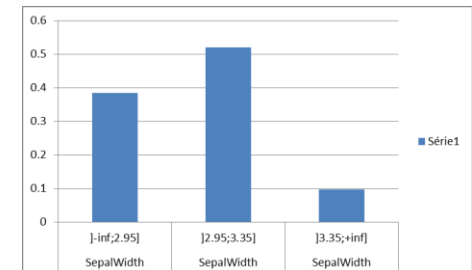
Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

Here, we can see that :

- 93.75% of the individuals belonging to cluster 1 have a PetalLength in the range]2.45;4.75] compare to 30% for the the global population
- The repartition of the values of SepalWidth in cluster 3 is

SepalWidth]-inf;2.95]	0.384615
SepalWidth]2.95;3.35]	0.519231
SepalWidth]3.35;+inf]	0.0961538

- which can be viewed (using your favorite tool) as



Predictive Clustering

How to define a profile

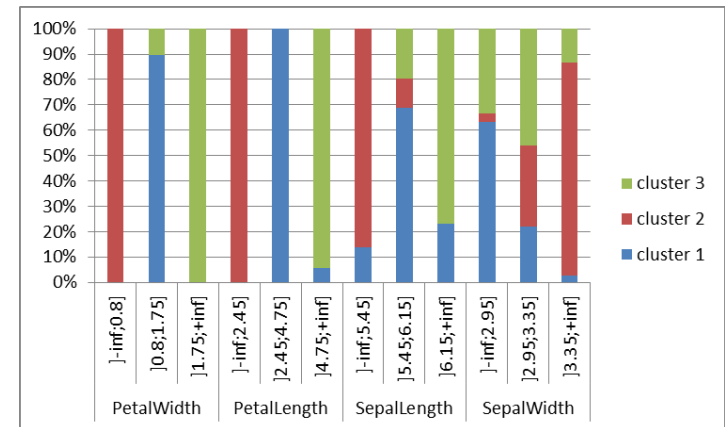
Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

Using this table you may also view multivariate histogram as :



Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

for

- the values of the PetalWidth are low

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

for

- the values of the PetalWidth are low (100% of the individuals belonging to cluster 2 have a PetalWidth in the range $]-\infty;0.8]$ compared to 33% for the the global population)

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

for

- the values of the PetalWidth are low (100% of the individuals belonging to cluster 2 have a PetalWidth in the range $]-\infty;0.8]$ compared to 33%)
- the values of the PetalLength are low (100% of the individuals belonging to cluster 2 have a PetalLength in the range $]-\infty;2.45]$ compared to 33%)

Note : Here it's a pure hazard to have 3 intervals for each explanatory variable, the number of intervals depends on the information contained in the database

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

for

- the values of the PetalWidth are low (100% of the individuals belonging to cluster 2 have a PetalWidth in the range $]-\infty;0.8]$ compare to 33%)
- the values of the PetalLength are low (100% of the individuals belonging to cluster 2 have a PetalLength in the range $]-\infty;2.45]$ compare to 33%)
- the values of the SepalLength are mainly low (90% of the individuals belonging to cluster 2 have a SepalLength in the range $]-\infty;5.45]$ compare to 34.6%)

Note : Here it's a pure hazard to have 3 intervals for each explanatory variable, the number of intervals depends on the information contained in the database

Predictive Clustering

How to define a profile

Example on Iris Database (continued)...

2) Secondly we analyze Cluster 2 (containing only Iris-setosa) to define its profile

- We use the Table 'Native attributes probas'

Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth]-inf;0.8]	0	1	0	0.333333
PetalWidth]0.8;1.75]	1	0	0.115385	0.36
PetalWidth]1.75;+inf]	0	0	0.884615	0.306667
PetalLength]-inf;2.45]	0	1	0	0.333333
PetalLength]2.45;4.75]	0.9375	0	0	0.3
PetalLength]4.75;+inf]	0.0625	0	1	0.366667
SepalLength]-inf;5.45]	0.145833	0.9	0	0.346667
SepalLength]5.45;6.15]	0.604167	0.1	0.173077	0.286667
SepalLength]6.15;+inf]	0.25	0	0.826923	0.366667
SepalWidth]-inf;2.95]	0.729167	0.04	0.384615	0.38
SepalWidth]2.95;3.35]	0.25	0.36	0.519231	0.38
SepalWidth]3.35;+inf]	0.0208333	0.6	0.0961538	0.24

by comparing the column "Cluster 2" with the column "global", we find what are the specificities of this cluster, compared to the global population...

for

- the values of the PetalWidth are low (100% of the individuals belonging to cluster 2 have a PetalWidth in the range $]-\infty;0.8]$ compared to 33%)
- the values of the PetalLength are low (100% of the individuals belonging to cluster 2 have a PetalLength in the range $]-\infty;2.45]$ compared to 33%)
- the values of the SepalLength are mainly low (90% of the individuals belonging to cluster 2 have a SepalLength in the range $]-\infty;5.45]$ compared to 34.6%)
- the values of the SepalWidth is less discriminative, we see that the individuals belonging to cluster 2 do not have low value (only 4% compared to 38%)

Note : Here, it's a pure coincidence to have 3 intervals for each explanatory variable, the number of intervals depends on the information contained in the database

Predictive Clustering

How to define a profile

Conclusion on Cluster 2...

This is a pure Cluster of Iris-Setosa

- 100% of the individuals belonging to cluster 2 have a PetalWidth in the range $[-\infty; 0.8]$ compare to 33%
- 100% of the individuals belonging to cluster 2 have a PetalLength in the range $[-\infty; 2.45]$ compare to 33%
- 90% of the individuals belonging to cluster 2 have a SepalLength in the range $[-\infty; 4.45]$ compare to 34.6%
- the values of the SepalWidth is less discriminative, we see that the individuals belonging to cluster 2 do not have low value (only 4% compare to 38%)

We may also use the Tables “Mean Values” (or the “Median Values”) to add information or to confirm the profile

Mean values for Numerical attributes :					
Var name	cluster 1	cluster 2	cluster 3	global	Missing values
PetalWidth	1.3125	0.244	2.01154	1.19867	0
PetalLength	4.24375	1.464	5.51731	3.75867	0
SepalLength	5.85833	5.006	6.63462	5.84333	0
SepalWidth	2.73333	3.418	3	3.054	0

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

For example the Table “Percentage per line - Native attributes proba” is dedicated to answer to the question :
“Where are individuals which have a given property?”

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

For example the Table “Percentage per line - Native attributes proba” is dedicated to answer to the question :
“Where are individuals which have a given property?”

Percentage per line - Native attributes proba :					
Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth] -inf;0.8]	0	1	0	1
PetalWidth]0.8;1.75]	0.888889	0	0.111111	1
PetalWidth]1.75;+inf]	0	0	1	1
PetalLength] -inf;2.45]	0	1	0	1
PetalLength]2.45;4.75]	1	0	0	1
PetalLength]4.75;+inf]	0.0545455	0	0.945455	1
SepalLength] -inf;5.45]	0.134615	0.865385	0	1
SepalLength]5.45;6.15]	0.674419	0.116279	0.209302	1
SepalLength]6.15;+inf]	0.218182	0	0.781818	1
SepalWidth] -inf;2.95]	0.614035	0.0350877	0.350877	1
SepalWidth]2.95;3.35]	0.210526	0.315789	0.473684	1
SepalWidth]3.35;+inf]	0.0277778	0.833333	0.138889	1

Note : This table presents the percentage of examples belonging to “Cluster k” for which the variable of the Column 1 (“Var name”) takes its values in the Intervals or Group name of the column 2 (“Modality/Interval”) - the percentage is computed over all the clusters and therefore sums always to 1 (100%). See the User Guide of the tool.

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

For example the Table “Percentage per line - Native attributes proba” is dedicated to answer to the question : “Where are individuals which have a given property?”

Percentage per line - Native attributes proba :					
Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth] -inf;0.8]	0	1	0	1
PetalWidth]0.8;1.75]	0.888889	0	0.111111	1
PetalWidth]1.75;+inf]	0	0	1	1
PetalLength] -inf;2.45]	0	1	0	1
PetalLength]2.45;4.75]	1	0	0	1
PetalLength]4.75;+inf]	0.0545455	0	0.945455	1
SepalLength] -inf;5.45]	0.134615	0.865385	0	1
SepalLength]5.45;6.15]	0.674419	0.116279	0.209302	1
SepalLength]6.15;+inf]	0.218182	0	0.781818	1
SepalWidth] -inf;2.95]	0.614035	0.0350877	0.350877	1
SepalWidth]2.95;3.35]	0.210526	0.315789	0.473684	1
SepalWidth]3.35;+inf]	0.0277778	0.833333	0.138889	1

For example where are the examples which have a SepalLength in the range]6.15;+inf] ?

Note : This table presents the percentage of examples belonging to “Cluster k” for which the variable of the Column 1 (“Var name”) takes its values in the Intervals or Group name of the column 2 (“Modality/Interval”) - the percentage is computed over all the clusters and therefore sums always to 1 (100%). See the User Guide of the tool.

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

For example the Table “Percentage per line - Native attributes proba” is dedicated to answer to the question :
“Where are individuals which have a given property?”

Percentage per line - Native attributes proba :					
Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth] -inf;0.8]	0	1	0	1
PetalWidth]0.8;1.75]	0.888889	0	0.111111	1
PetalWidth]1.75;+inf]	0	0	1	1
PetalLength] -inf;2.45]	0	1	0	1
PetalLength]2.45;4.75]	1	0	0	1
PetalLength]4.75;+inf]	0.0545455	0	0.945455	1
SepalLength] -inf;5.45]	0.134615	0.865385	0	1
SepalLength]5.45;6.15]	0.674419	0.116279	0.209302	1
SepalLength]6.15;+inf]	0.218182	0	0.781818	1
SepalWidth] -inf;2.95]	0.614035	0.0350877	0.350877	1
SepalWidth]2.95;3.35]	0.210526	0.315789	0.473684	1
SepalWidth]3.35;+inf]	0.0277778	0.833333	0.138889	1

For example where are the examples which have a SepalLength in the range]6.15;+inf] ?

Answer:

- 21.82% in cluster 1
- 0% in cluster 2
- 78.18% in cluster 3

Note : This table presents the percentage of examples belonging to “Cluster k” for which the variable of the Column 1 (“Var name”) takes its values in the Intervals or Group name of the column 2 (“Modality/Interval”) - the percentage is computed over all the clusters and therefore sums always to 1 (100%). See the User Guide of the tool.

Predictive Clustering

How to define a profile

Note : all the Tables in the Train (or Test) EvaluationReport have a purpose*.

For example the Table “Percentage per line - Native attributes proba” is dedicated to answer to the question : “Where are individuals which have a given property?”

Percentage per line - Native attributes proba :					
Var name	Modality/Interval	cluster 1	cluster 2	cluster 3	global
PetalWidth] -inf;0.8]	0	1	0	1
PetalWidth]0.8;1.75]	0.888889	0	0.111111	1
PetalWidth]1.75;+inf]	0	0	1	1
PetalLength] -inf;2.45]	0	1	0	1
PetalLength]2.45;4.75]	1	0	0	1
PetalLength]4.75;+inf]	0.0545455	0	0.945455	1
SepalLength] -inf;5.45]	0.134615	0.865385	0	1
SepalLength]5.45;6.15]	0.674419	0.116279	0.209302	1
SepalLength]6.15;+inf]	0.218182	0	0.781818	1
SepalWidth] -inf;2.95]	0.614035	0.0350877	0.350877	1
SepalWidth]2.95;3.35]	0.210526	0.315789	0.473684	1
SepalWidth]3.35;+inf]	0.0277778	0.833333	0.138889	1

For example where are the examples which have a SepalLength in the range]6.15;+inf] ?

Answer:

- 21.82% in cluster 1
- 0% in cluster 2
- 78.18% in cluster 3
- This indication is useful when you would like to adress a part of the global population having a given characteristic.

Note : This table presents the percentage of examples belonging to “Cluster k” for which the variable of the Column 1 (“Var name”) takes its values in the Intervals or Group name of the column 2 (“Modality/Interval”) - the percentage is computed over all the clusters and therefore sums always to 1 (100%). See the User Guide of the tool.