Time series prediction, from preprocessing to basic algorithms

RainsMore – 24/10/2022

Kelly Grassi

Weather FORCE

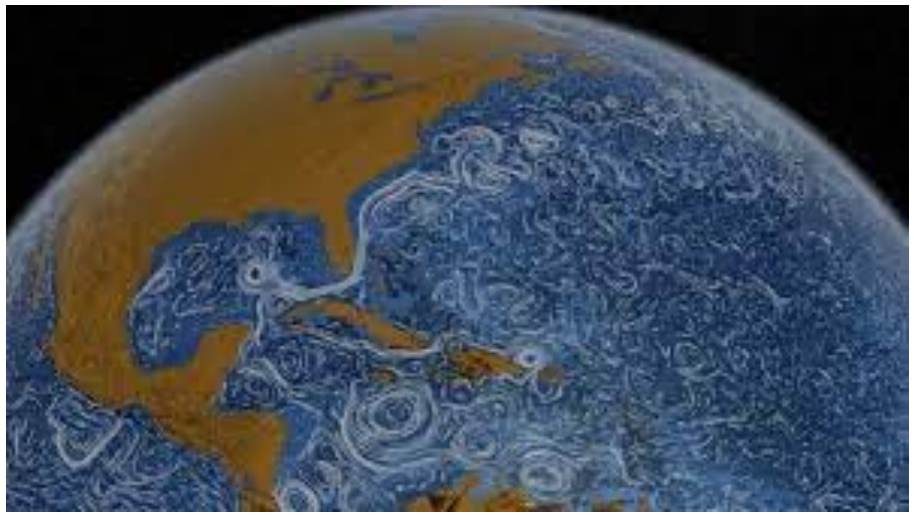
Introduction

1. Overview to preprocessing
2. Resume to basic IA methods and concept

Weather FORCE

Common goal

To improve our knowledge of ecosystem functioning and physical and biological processes



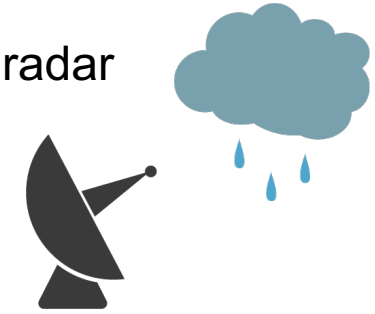
Context

→ Multi-source database

station

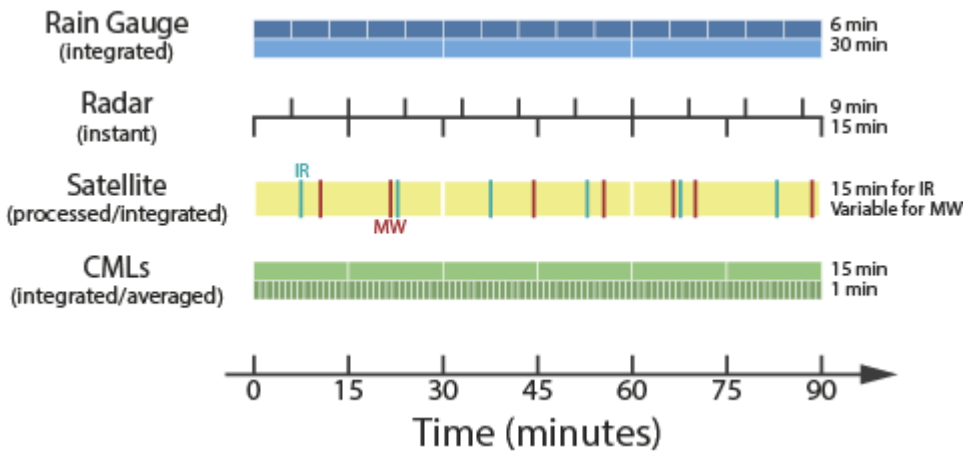
satellite

radar

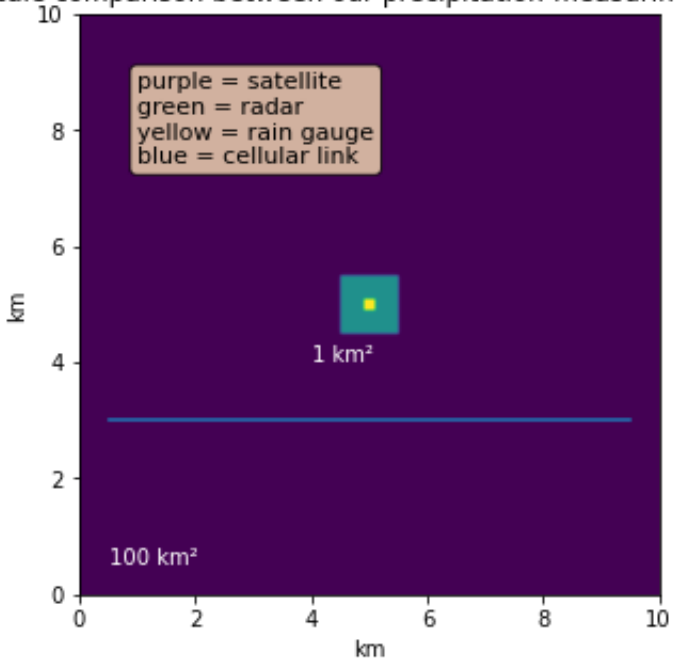


Opportunity data :CLM

Time scale differences between sources



Scale comparison between our precipitation-measuring tools

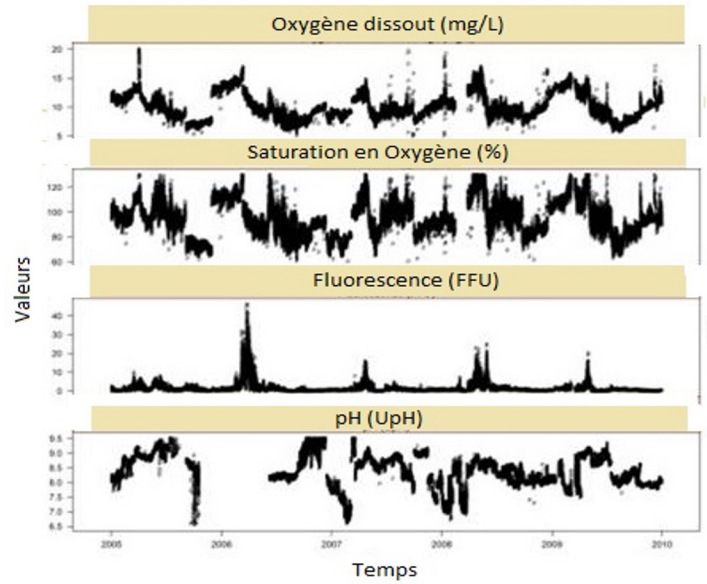


Weather FORCE

Context

→ Multi-varied database

Exemple de base de données



Essential variables

GCOS GLOBAL CLIMATE OBSERVING SYSTEM

The Global Ocean Observing System

→ 54 Essential Climate Variables (ECV)
Ex: air temperature, precipitation

→ Over 100 Essential Ocean Variables (EOV)
Ex: Fluorescences, Taxonomy, [Nutrients].]

Weather FORCE

Why is data preprocessing important?

→ Operating constraints

- **Sampling difference**
  - Acquisition frequencies
  - Resolution
  - Measurement units
- **Difference of the data**
  - Missing data
  - Noise
  - Nature of the information : estimation, measurement station ...

Weather FORCE

Preprocecing Data

Different types :

1. Data integration
2. Data cleaning
3. Data tranformation
4. Data reduction

Weather FORCE

Data integration - MetaData

→ Homogenized Metadata

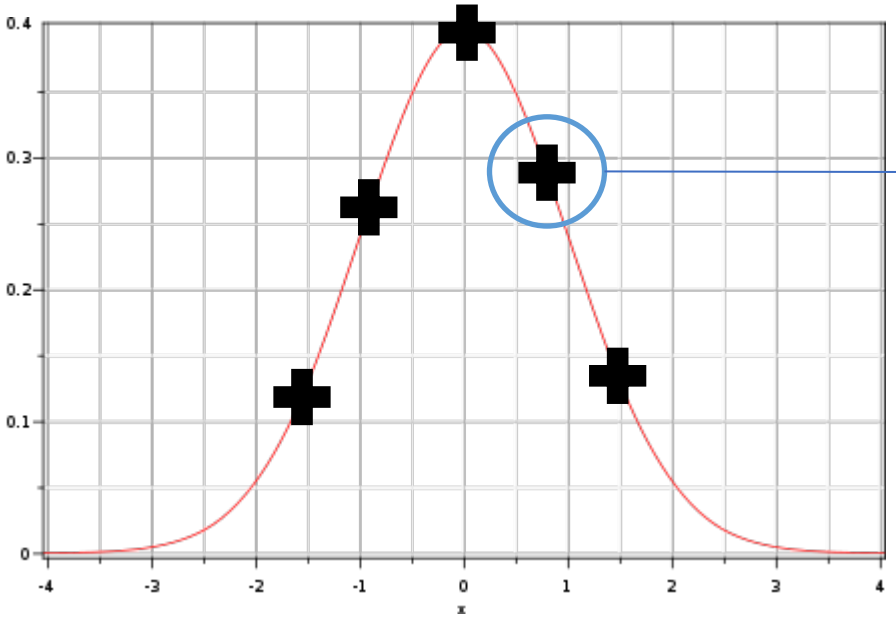
- Check Entity identification problem
  - ID Station or name
- Detect and resolve data value concepts
  - Date format "MM/DD/YYYY" O "DD/MM/YYYY".

Weather FORCE

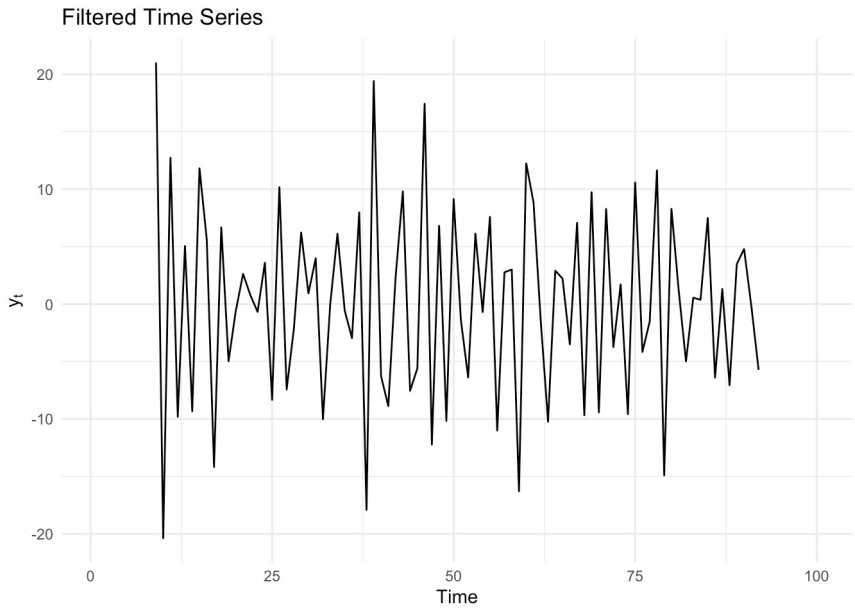
Data cleaning – Missing Data

- ignoring missing data
- Data completion

Low frequency



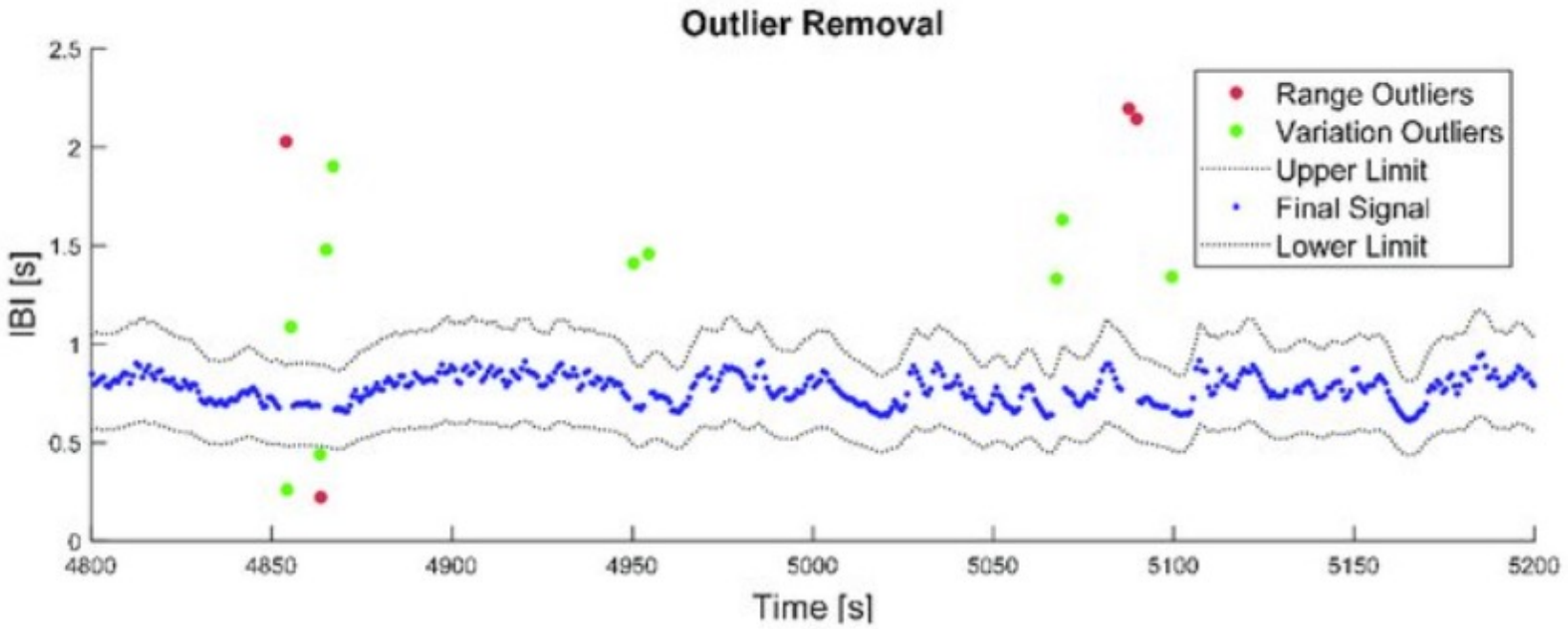
High frequency



Weather FORCE

Data cleaning – data correction

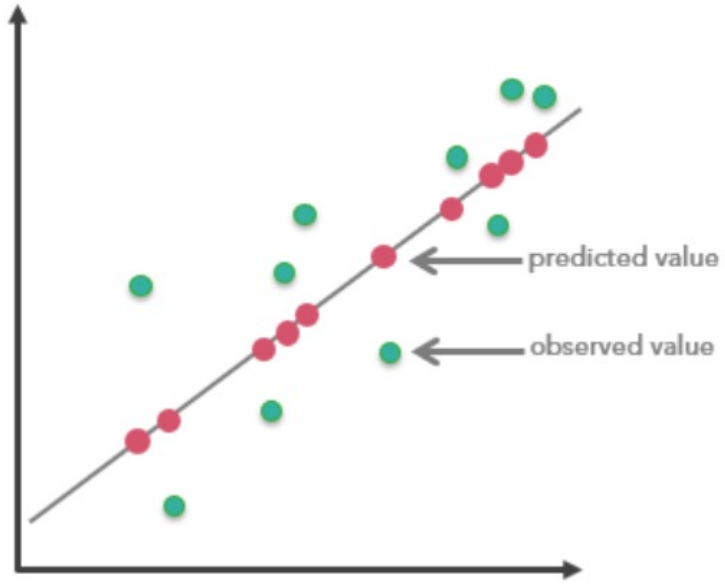
- Removing outlier data

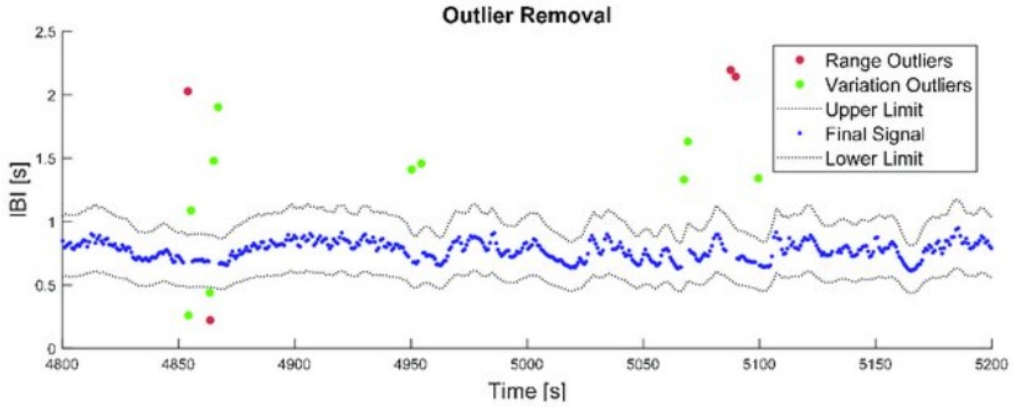


Data cleaning – Missing Data

- Removing outlier data
- Regression

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

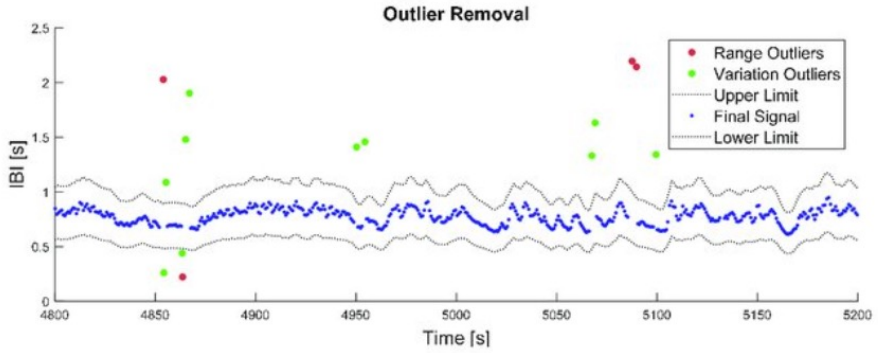


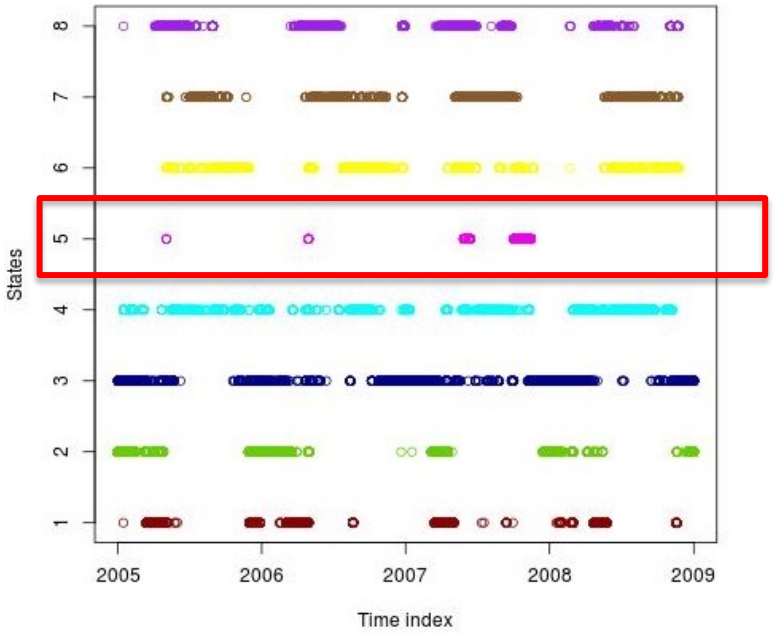


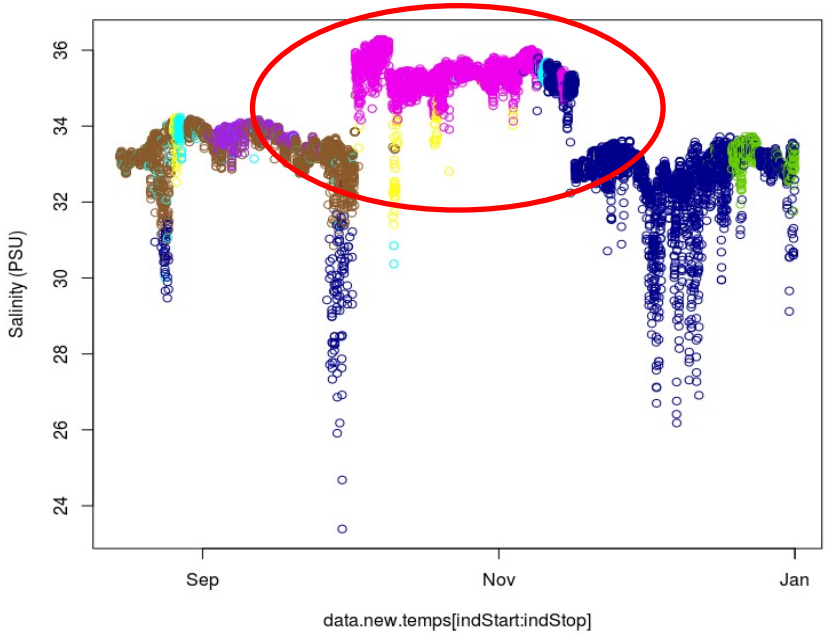
Weather FORCE

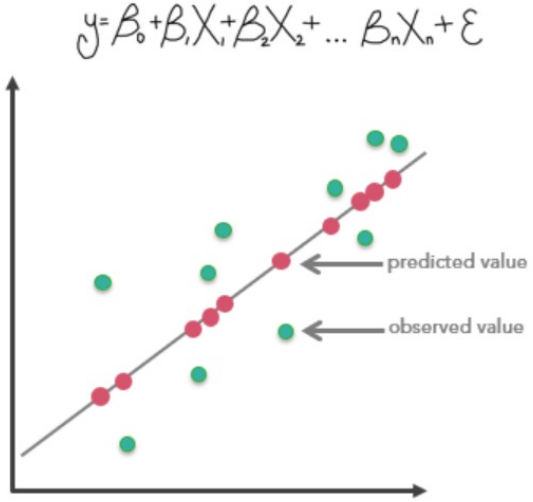
Data cleaning – Missing Data

- Removing outlier data
- Regression
- Clustering





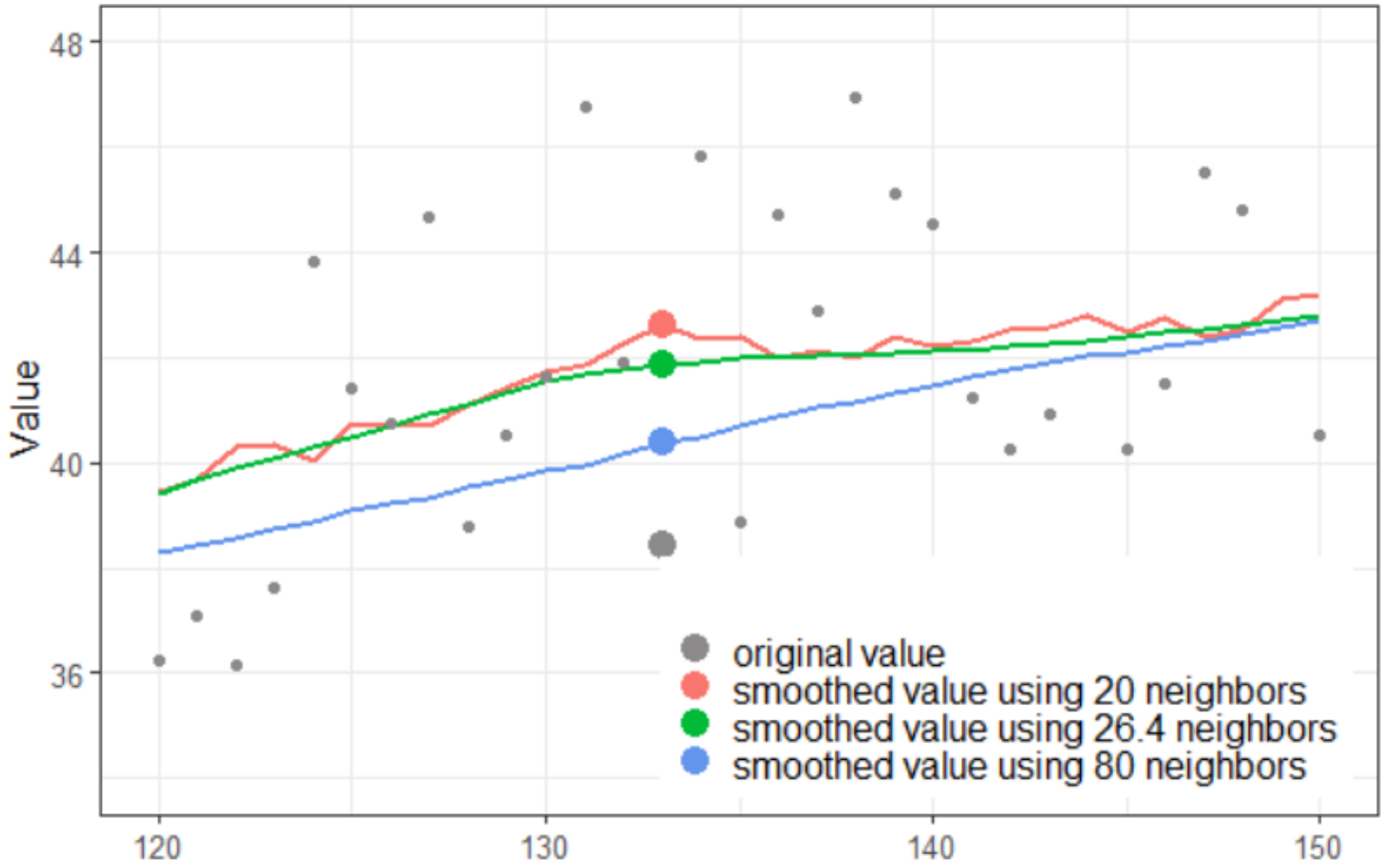




Data transformation

- Smoothing
- Remove the noise
- Help to knows

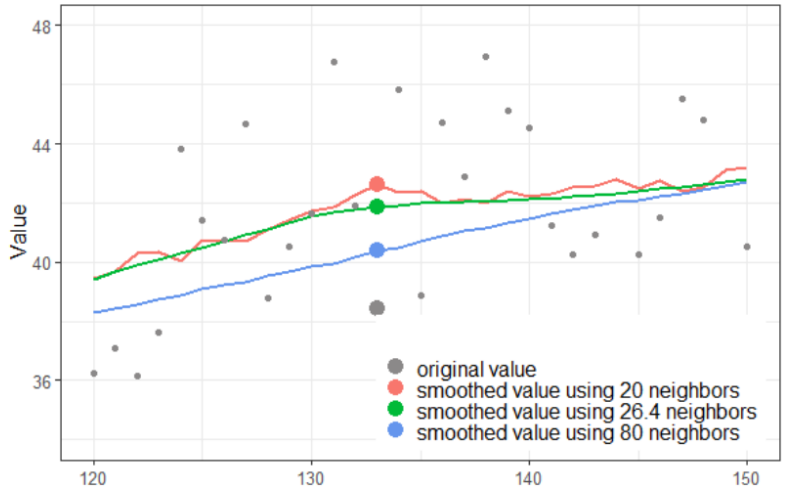
global problems

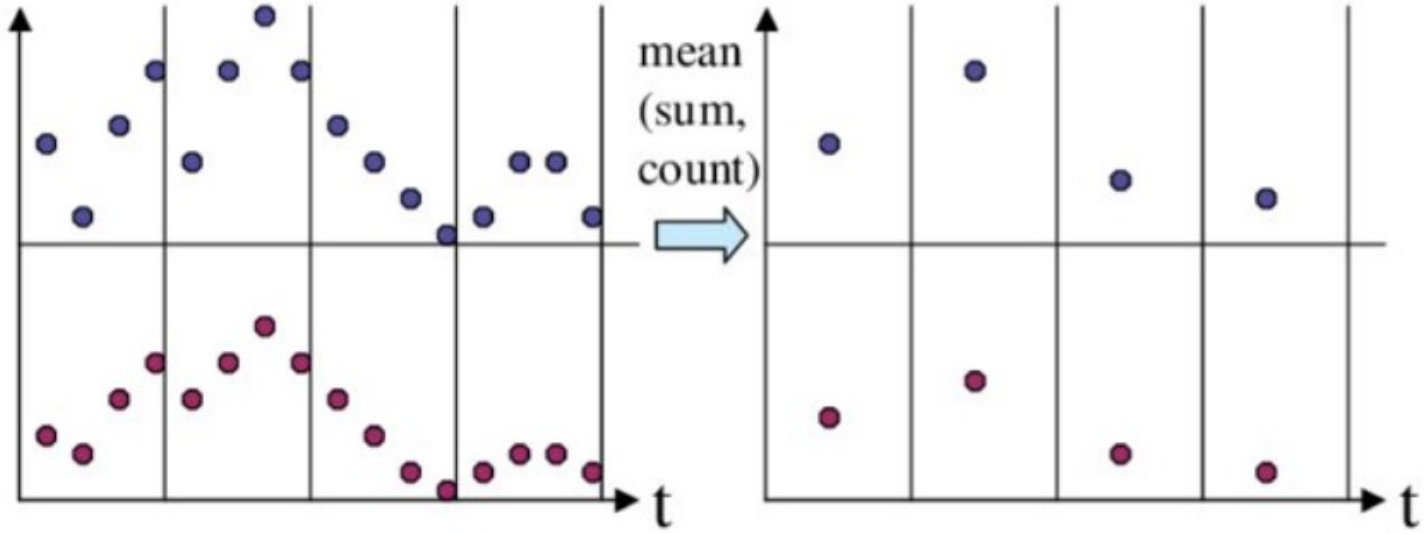


Weather FORCE

Data transformation

- Smoothing
- Aggregation
  - correlate and reduce dataset

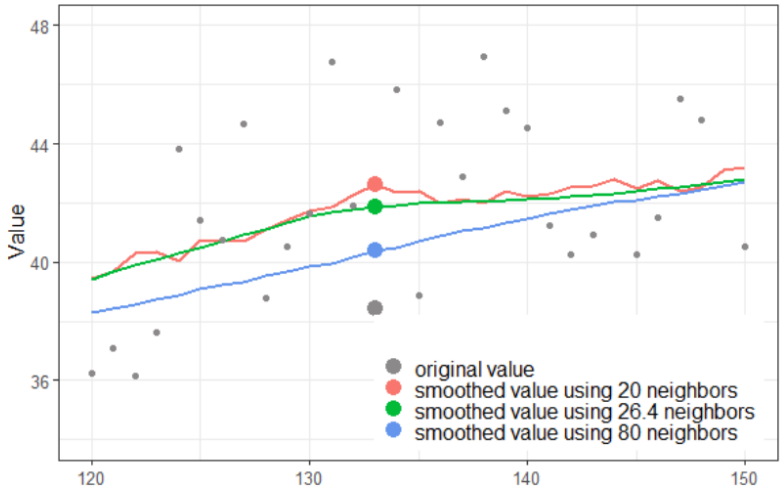


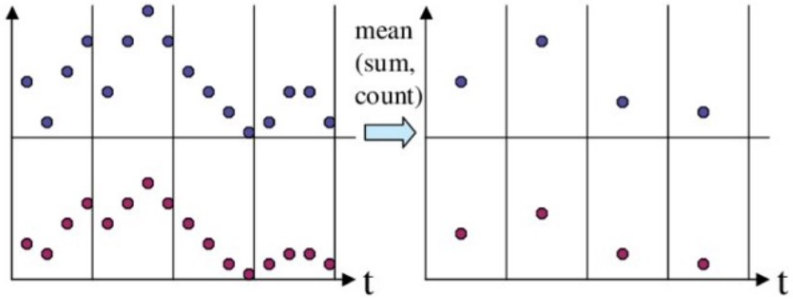


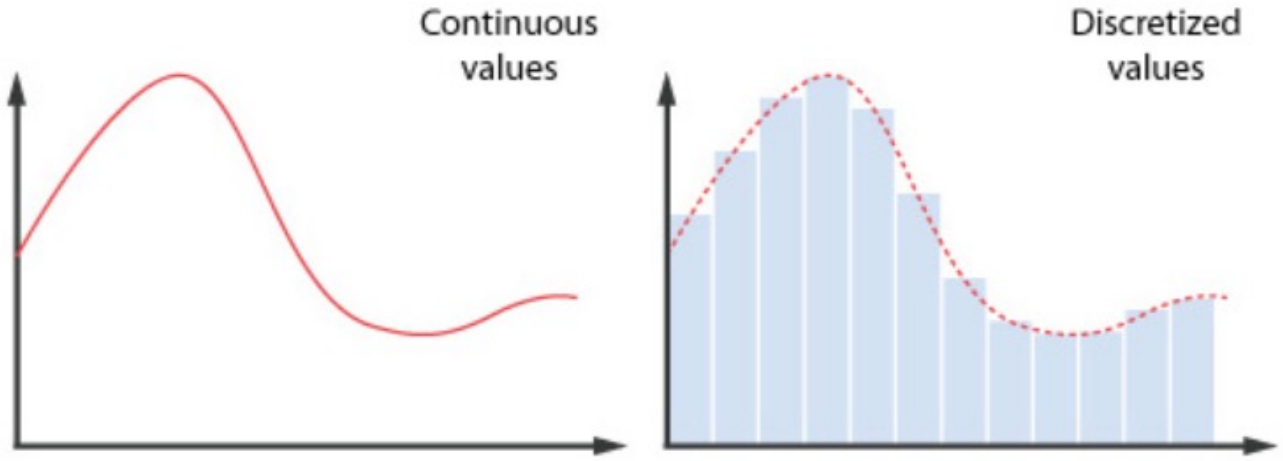
Weather FORCE

Data transformation

- Smoothing
- Aggregation
- Discretization
  - Reduce the size of data



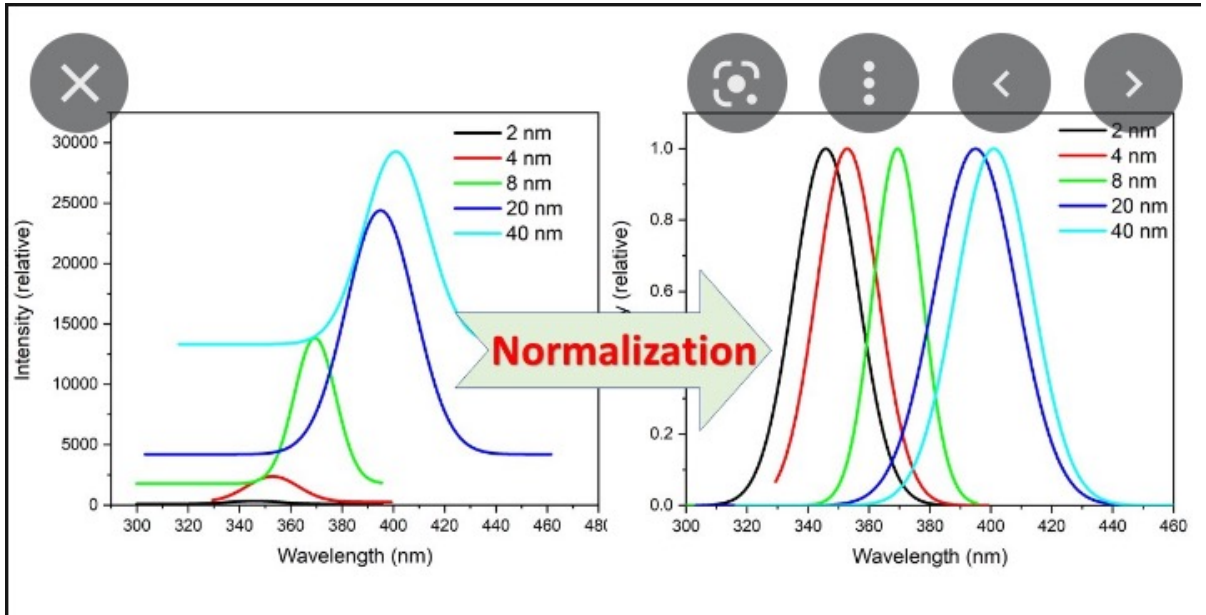


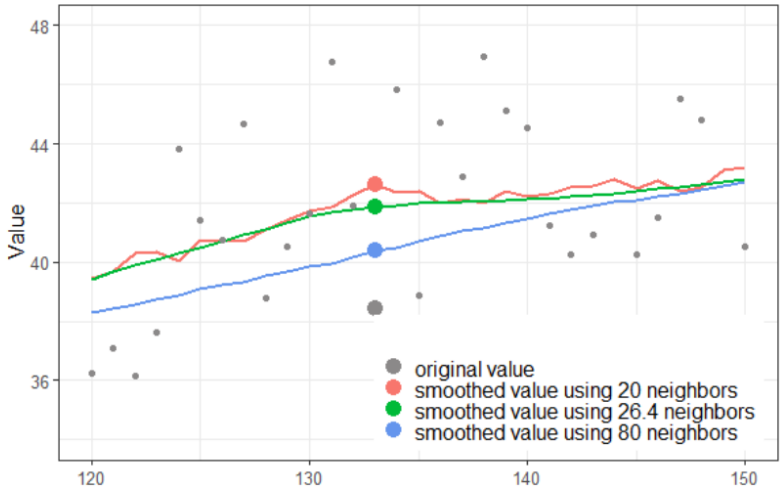


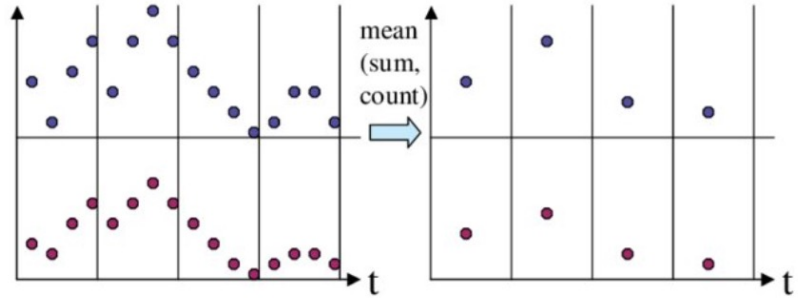
Data transformation

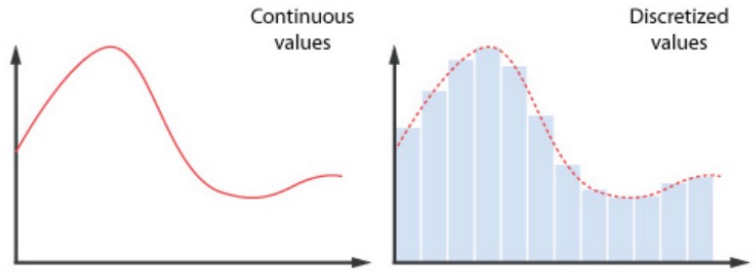
- Smoothing
- Aggregation
- Discretization
- Normalization

– Smaller range and homogenous data









Weather FORCE

Data reduction

- Dimensionality reduction

– wavelet transforms and PCA (principal component analysis).

- Number Reduction

– Volume is reduced

- Attribute subset selection

– most relevant attributes are selected

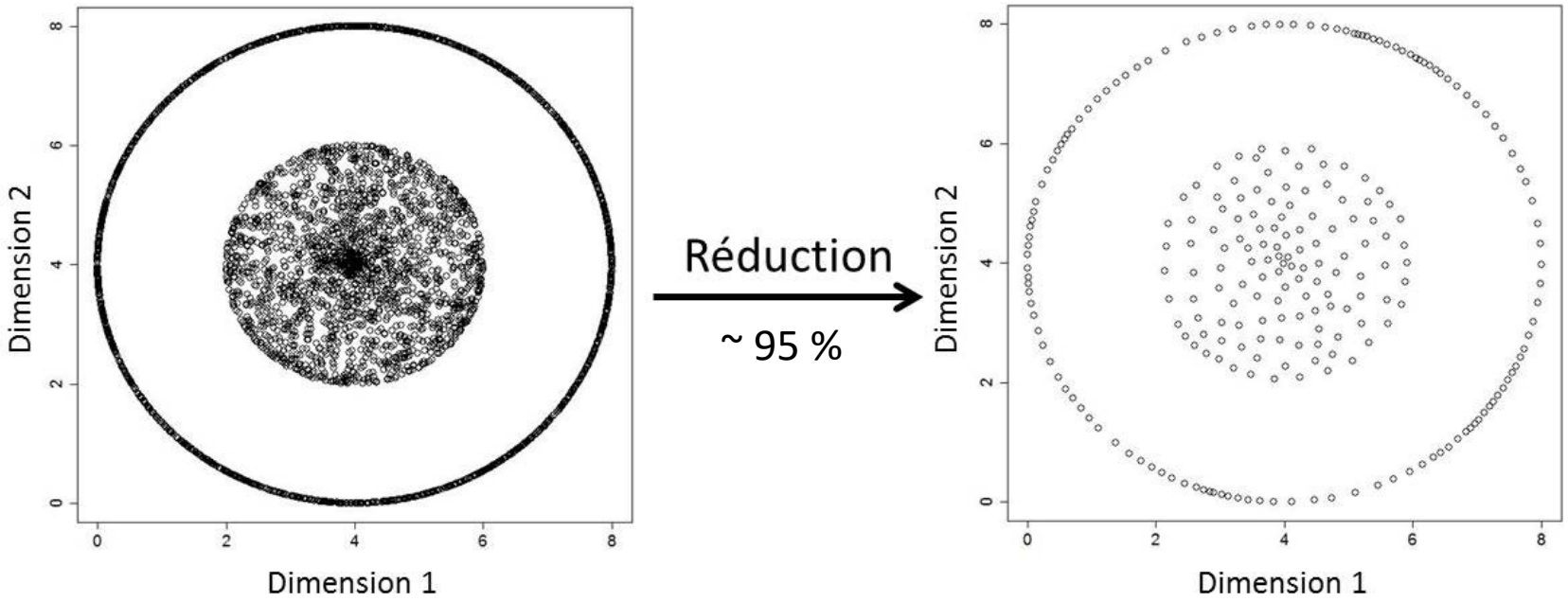
Data compression

–optimization of storage capacity

Weather FORCE

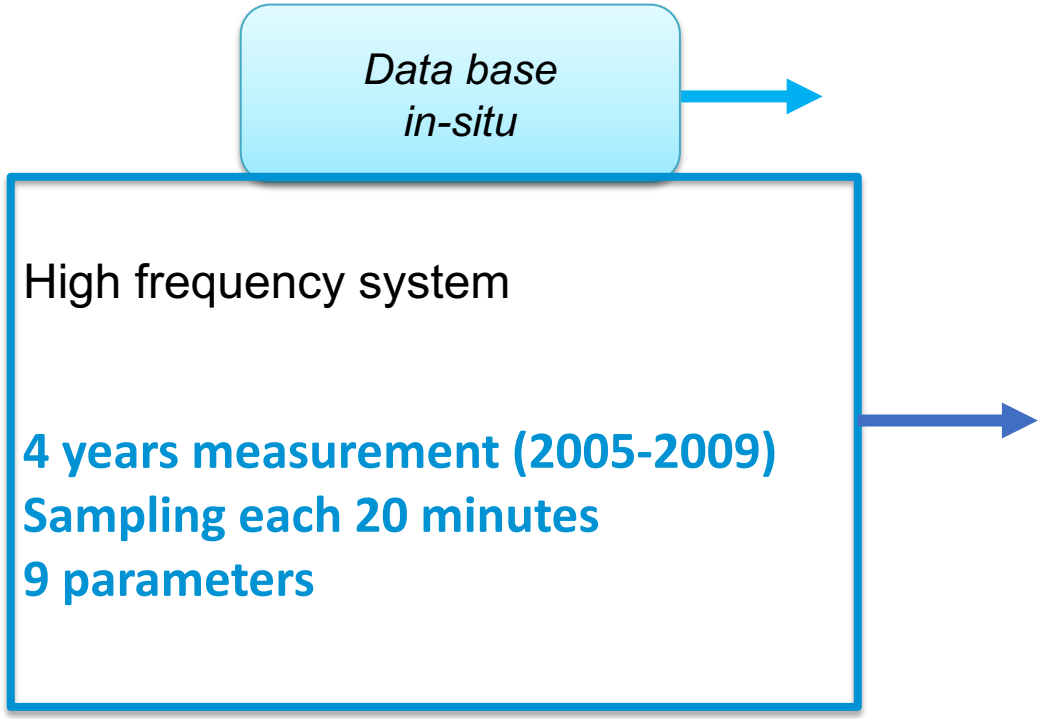
reduction step :
selected representative dataset by a vectors quantification of space : use K-means and Elbow criterion

- Reduction 4 000 to 260 points
  - Elbow criterion:
    - 95 % of variance explaned

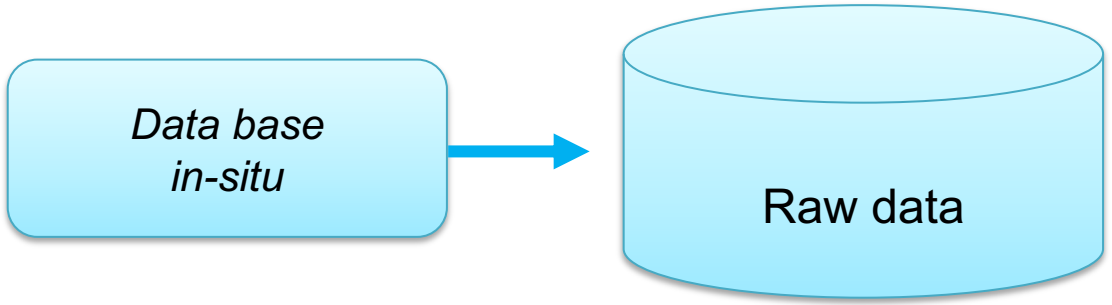


Preprocessing exemple

Weather FORCE



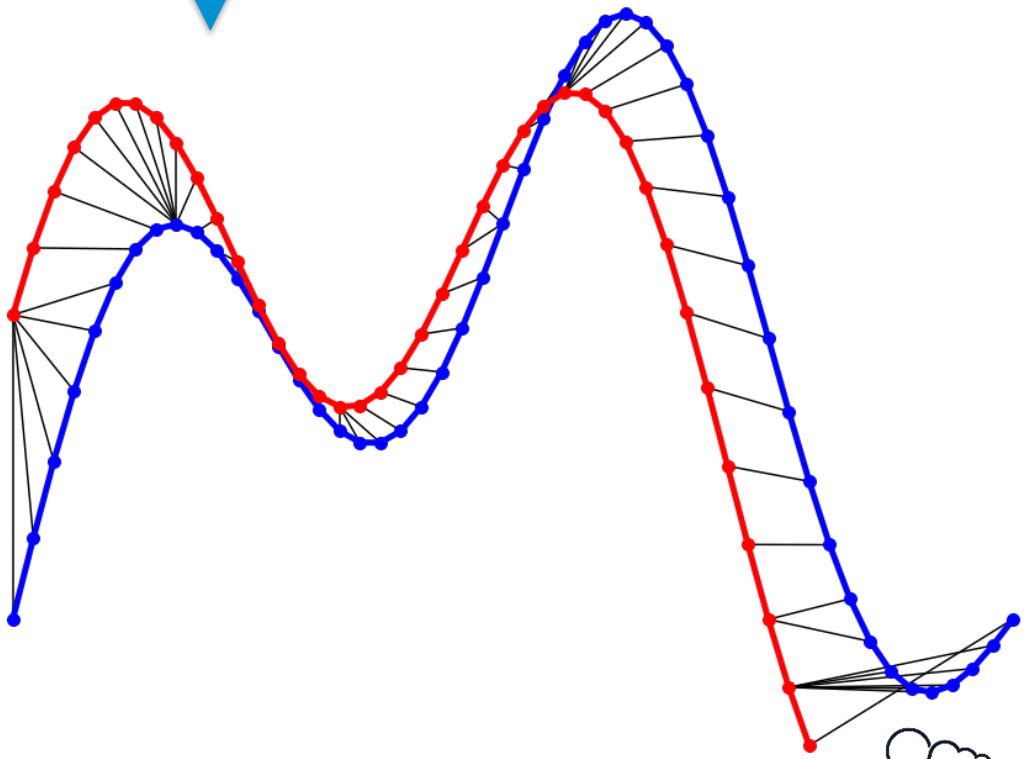
- Acquisition frequencies
- Measurements units
- Sensor Failure
- Missing data



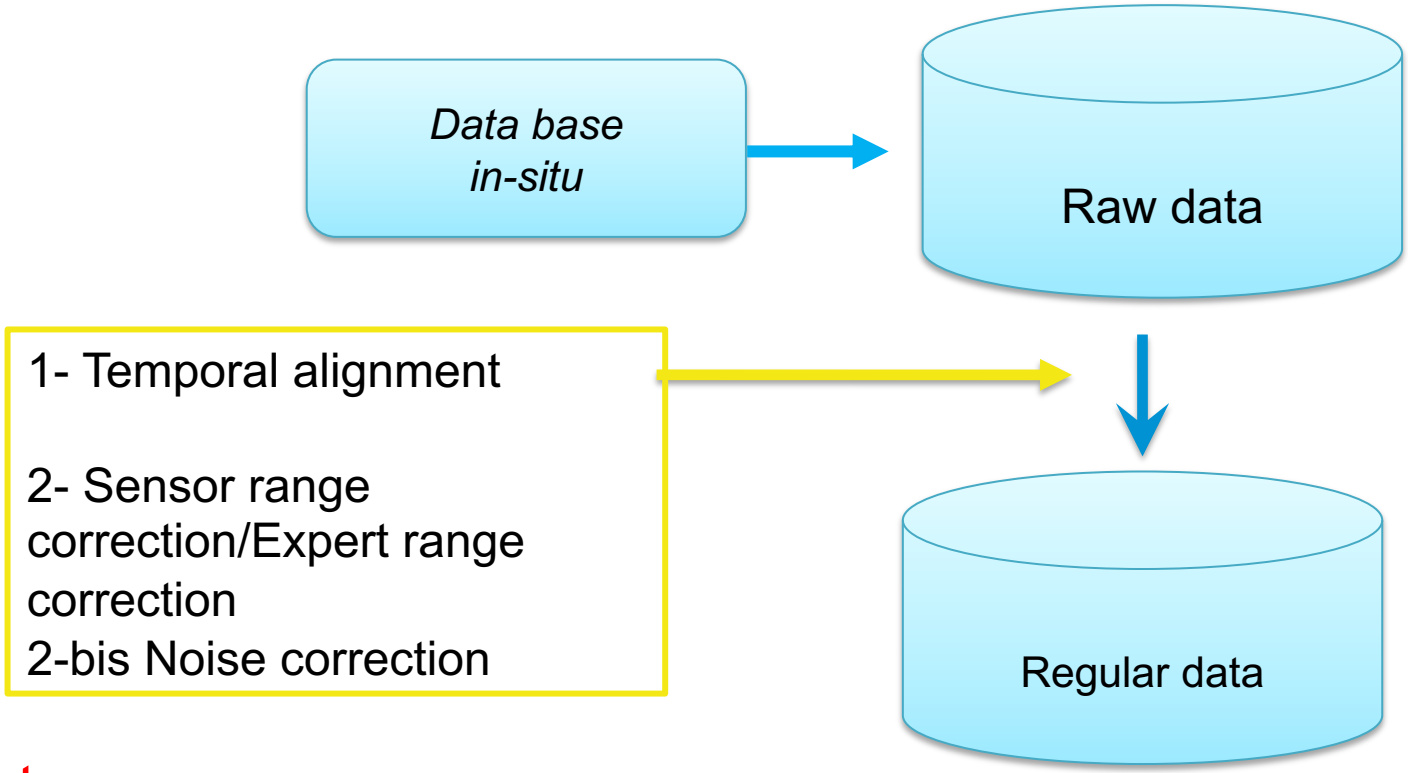
1- Temporel alignment

Aggregation

- Computation of the average time step of the time series
- Definition of an ideal time variable

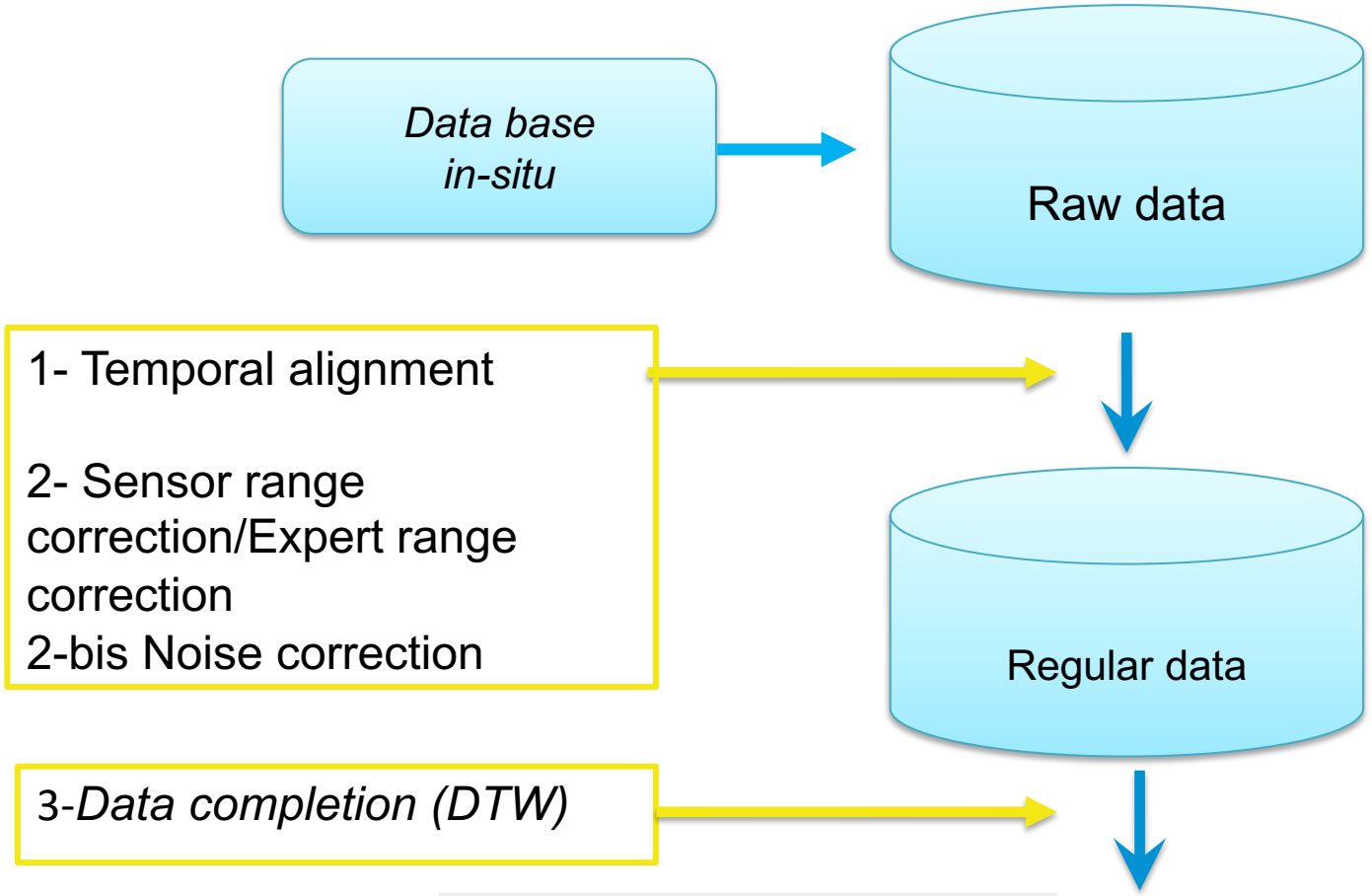


Weather FORCE



Cleaning step :

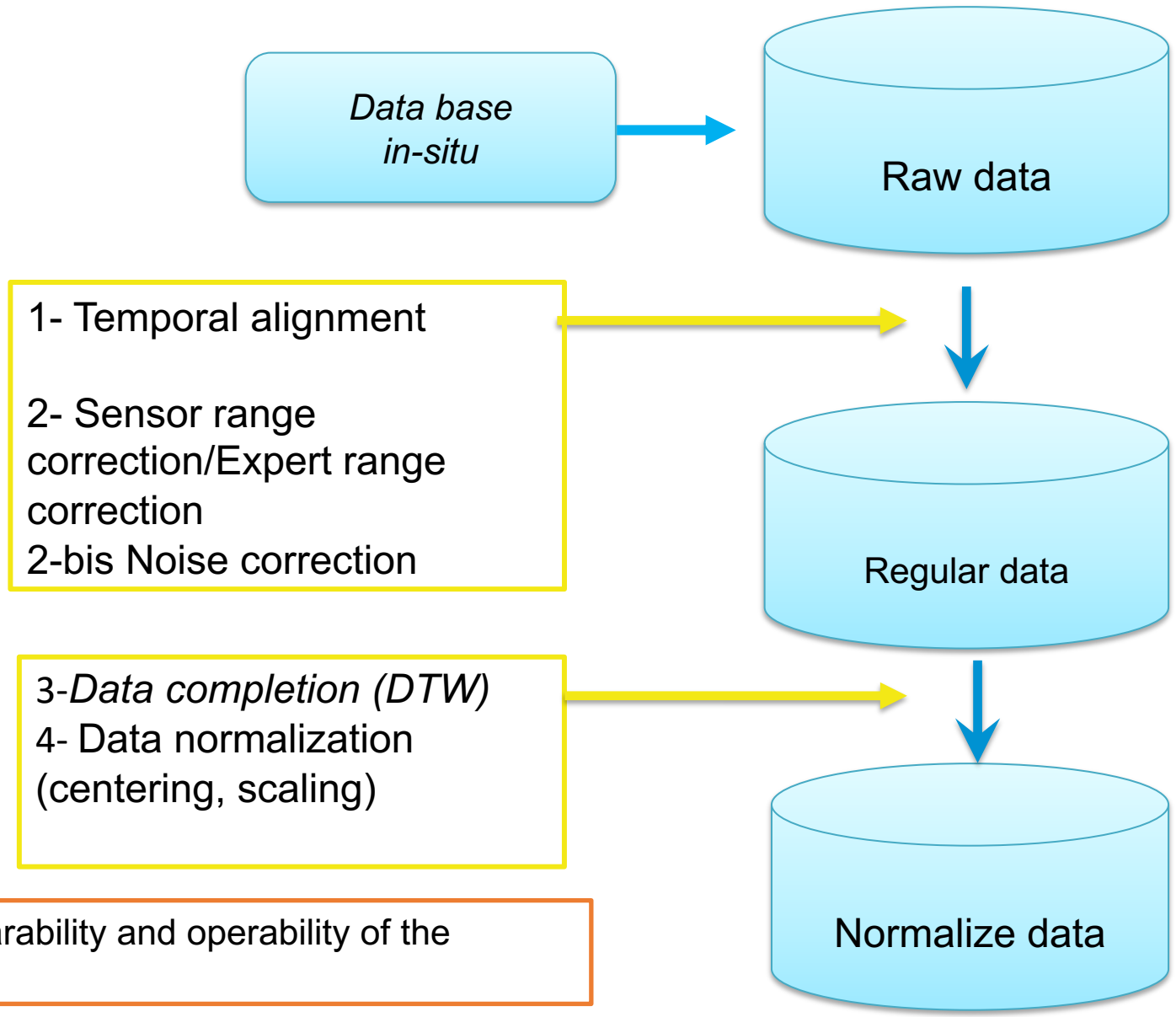
- Setting up of "sensors" and "experts" ranges
- Removal of out-of-range values
- Replacement by out of range values (NA)



DTW : *Dynamic Time Warping*

Cleaning step

- Fills in large gaps of missing data
- Preserves the dynamics of complex signals (preserves seasonality and accepts temporal variabilities)



→ Facilitate inter-comparability and operability of the different variables

IA basic algorithms

Weather FORCE

Machine Learning Base

Data: input/output pairs X, Y
Objectives: associate values of Y to X

X Observations = one or N characteristics (e.g. height, age)
Y Labels = what we want to know (e.g. male or female)

X = Environmental variables (Precipitation, cloud cover)
Y = Environmental states (thunderstorm)

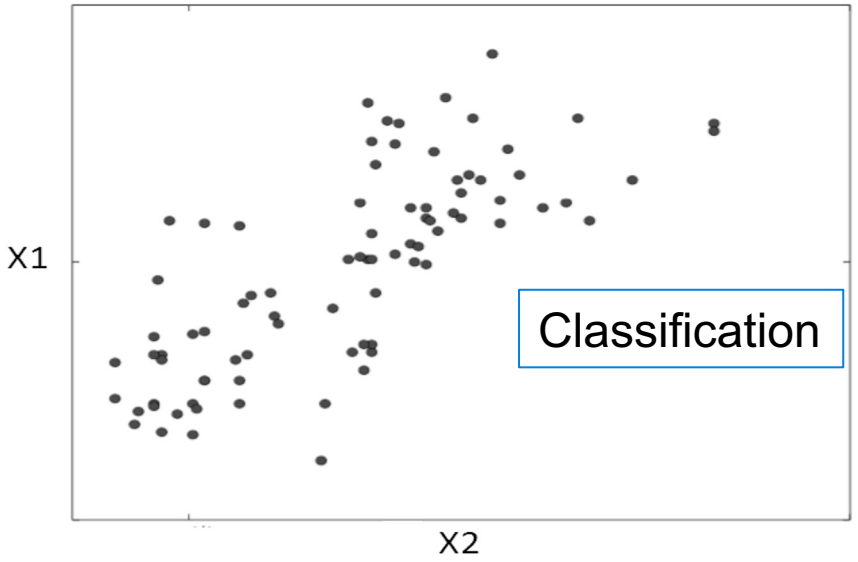
Weather FORCE

Machine Learning Base

Unsupervised

Learning sample = Obs. X
No label Y
Identification of natural structures

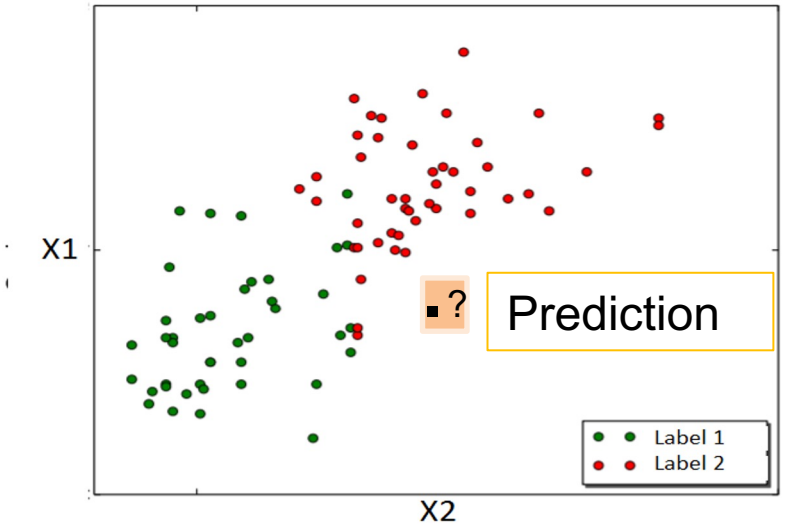
Input: observations X
Output: classes Y



Supervised

Learning sample = Obs. X, labels Y
Identification of boundaries or models allowing to establish a rule between Y and X

Input: an observation x
Output: label y associated with the new x



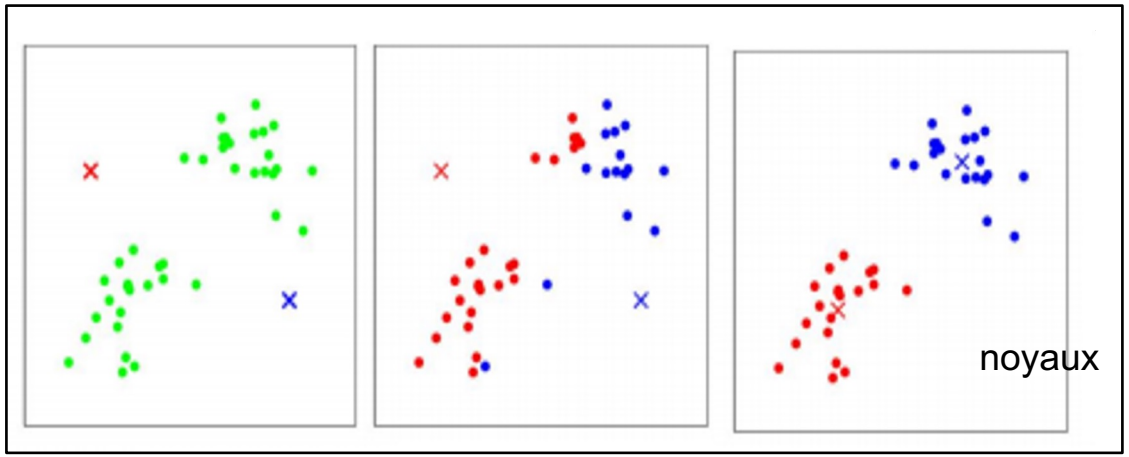
Weather FORCE

Machine Learning Base

Unsupervised

Identification of natural structures

→ Core/Convexity :
K-Means - EM



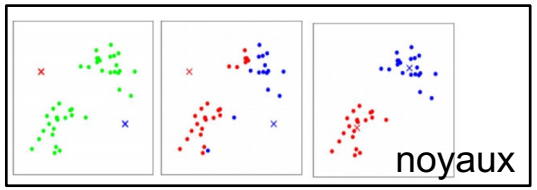
Weather FORCE

Machine Learning Base

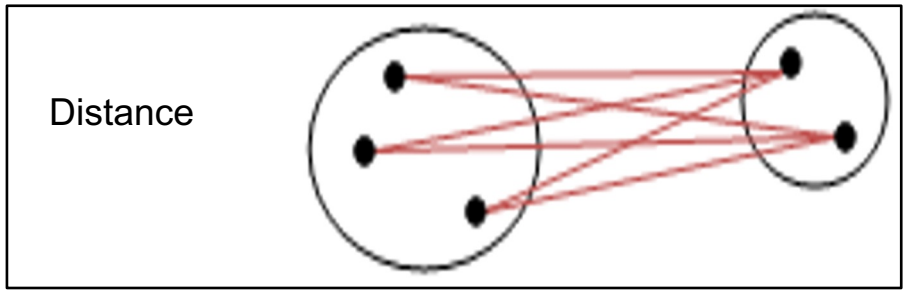
Unsupervised

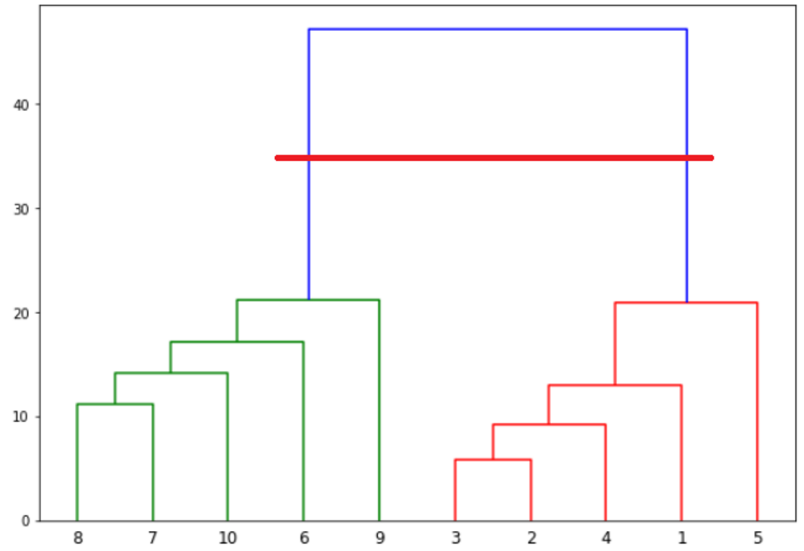
Identification of natural structures

→ Core/Convexity :
K-Means - EM



→ **Range : *Hierarchical Clustering (HC)***





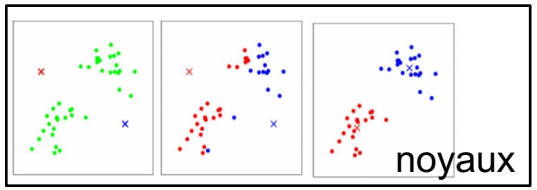
Weather FORCE

Machine Learning Base

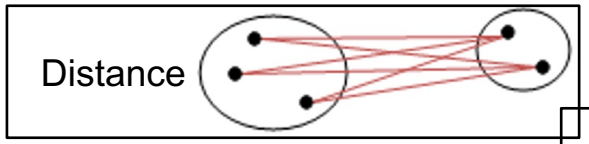
Unsupervised

Identification of natural structures

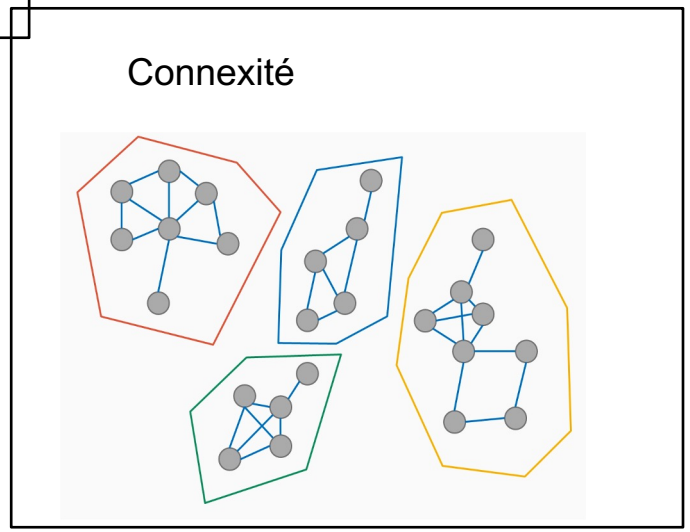
→ Core/Convexity :
K-Means - EM



→ Range : *Hierarchical Clustering (HC)*



→ **Connexity/Graph:**
Spectral Clustering (SC)

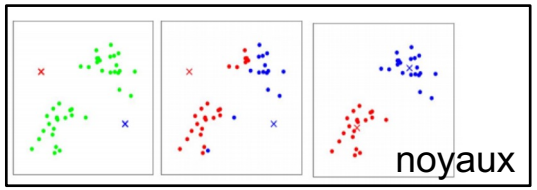


Machine Learning Base

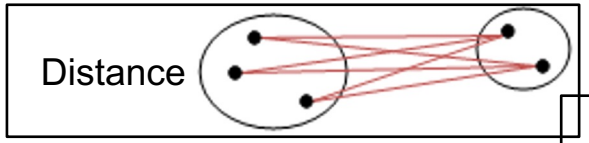
Unsupervised

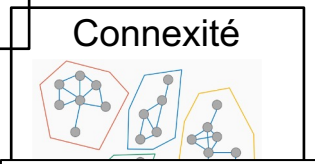
Identification of natural structures

→ Core/Convexity :
K-Means - EM



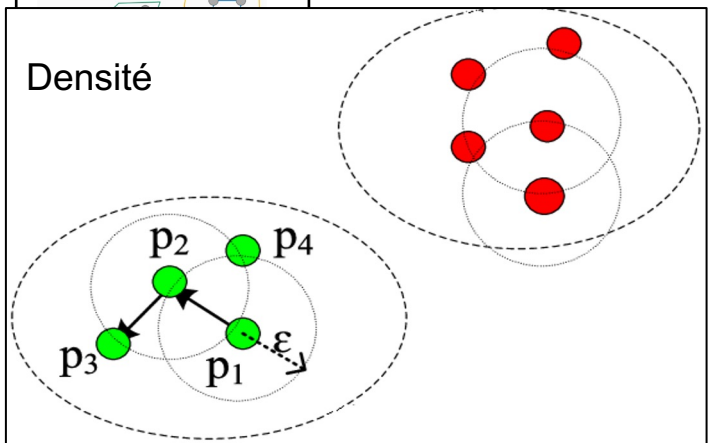
→ Range : *Hierarchical Clustering (HC)*





→ Connexity/Graph:
Spectral Clustering (SC)

→ **Density : DBSCAN**



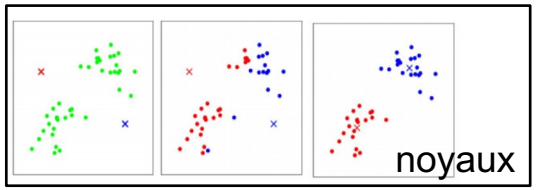
Weather FORCE

Machine Learning Base

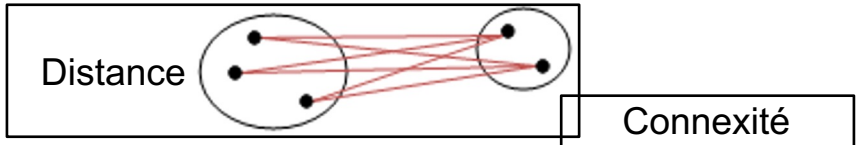
Unsupervised

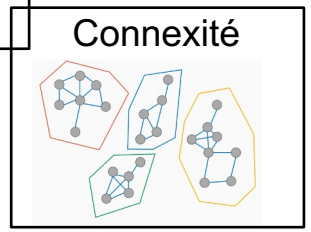
Identification of natural structures

→ Core/Convexity :
K-Means - EM

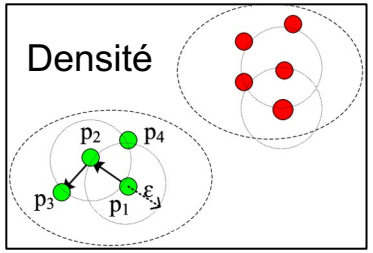


→ Range : *Hierarchical Clustering (HC)*





→ Connexity/Graph:
Spectral Clustering (SC)

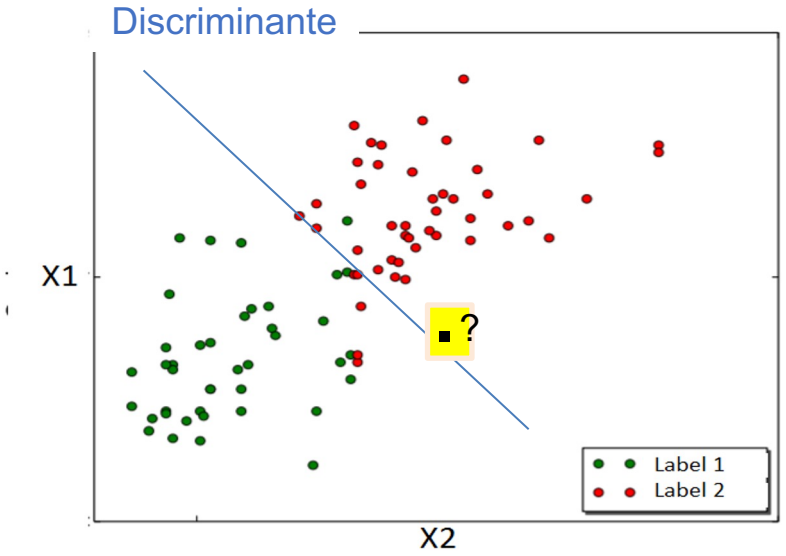


→ Density : DBSCAN

Supervised

Identification of boundaries or patterns

→Discriminant : RF, SVM, MLP, ...



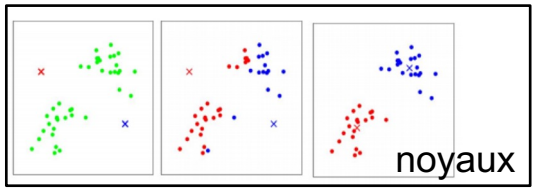
Weather FORCE

Machine Learning Base

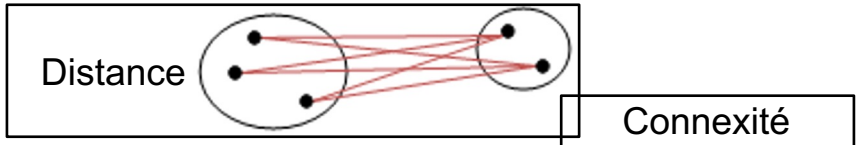
Unsupervised

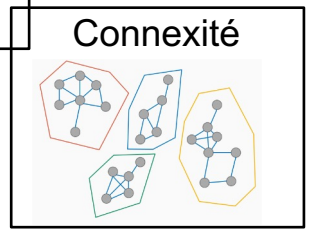
Identification of natural structures

→ Core/Convexity :
K-Means - EM

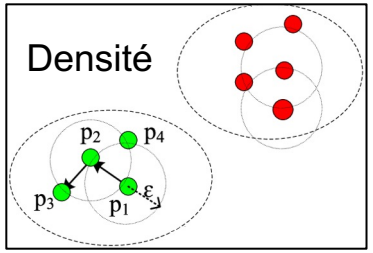


→ Range : *Hierarchical Clustering (HC)*





→ Connexity/Graph:
Spectral Clustering (SC)

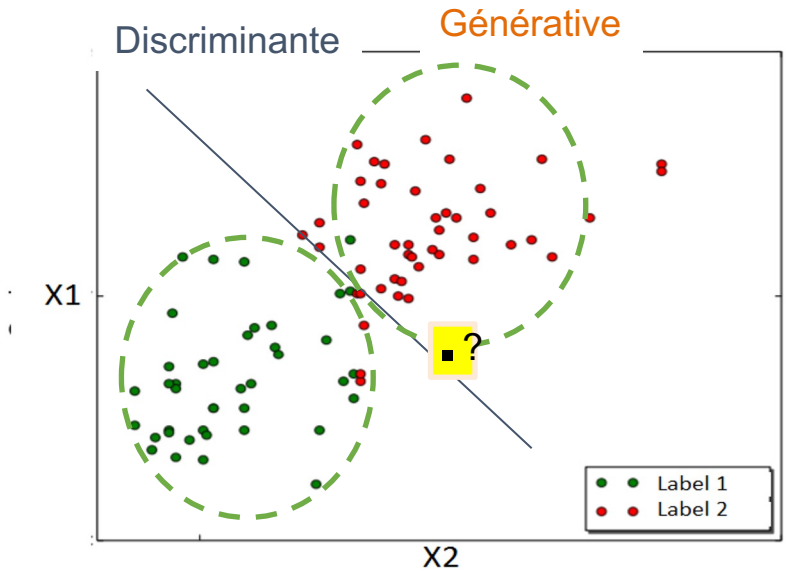


→ **Density : DBSCAN**

Supervised

Identification of boundaries or patterns

→Discriminant : RF, SVM, MLP, ...
→ Generative : HMM, ...



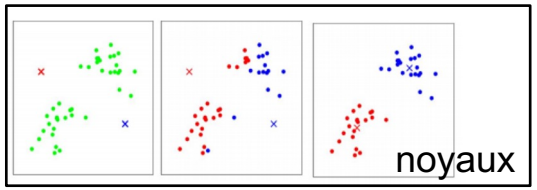
Weather FORCE

Machine Learning Base

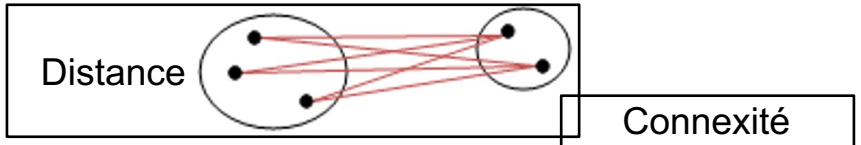
Unsupervised

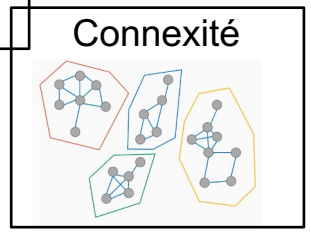
Identification of natural structures

→ Core/Convexity :
K-Means - EM

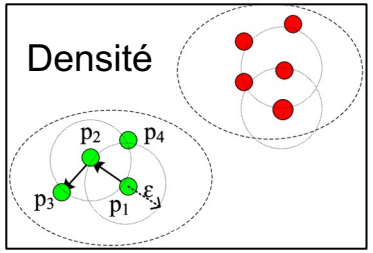


→ Range : *Hierarchical Clustering (HC)*





→ Connexity/Graph:
Spectral Clustering (SC)

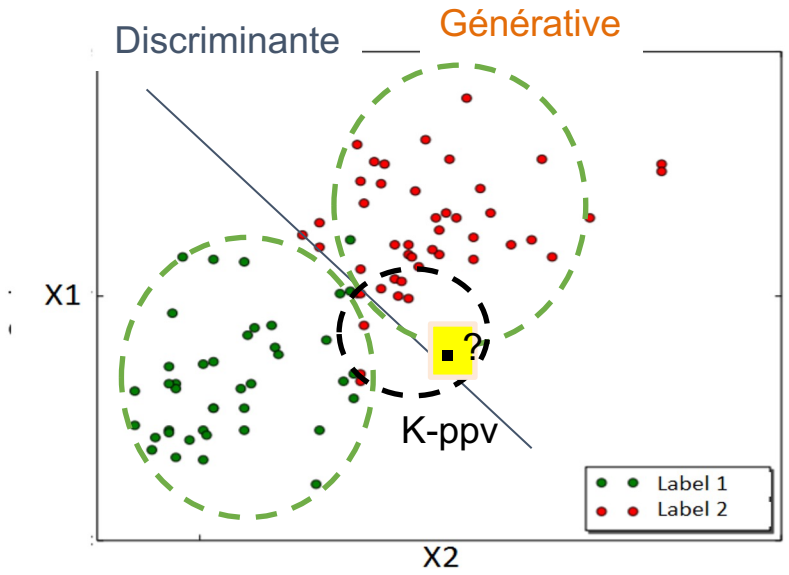


→ **Density : DBSCAN**

Supervised

Identification of boundaries or patterns

→Discriminant : RF, SVM, MLP, ...
→ Generative : HMM, ...
→ **Other(s) : K-ppv**



Weather FORCE

Thank for ours attention

Weather
FORCE