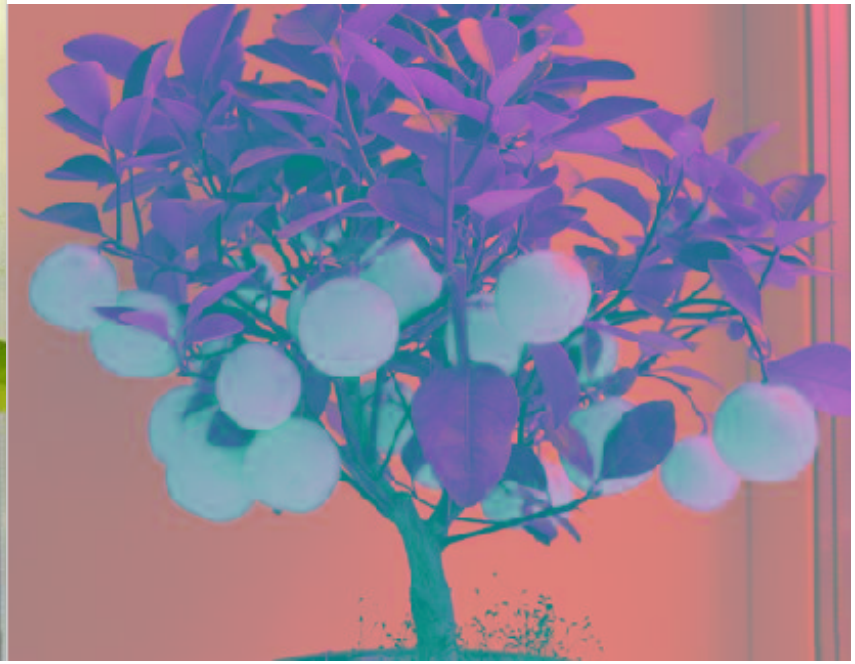
New clustering methods into a parallel code

SORIANO Tristan, AUBENEAU Simon, DUVAL Quentin,
PRIEUL Simon, SANTINA Jeremy

INP TOULOUSE ENSEEIHT

March 12, 2015

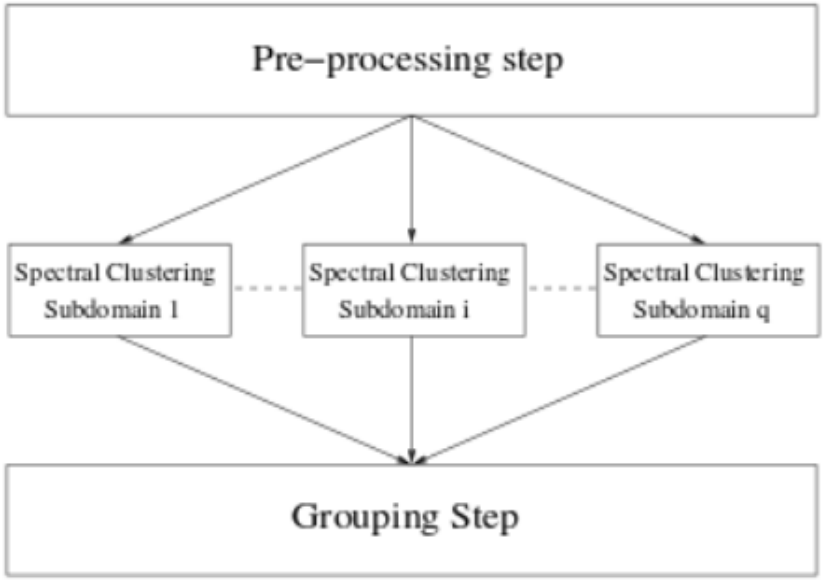
Clustering method



High computational complexity

Challenge

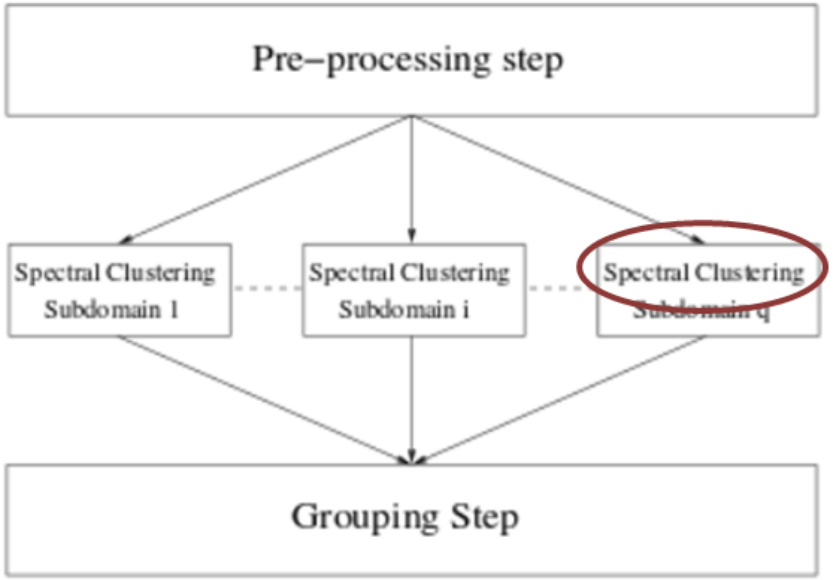
The data used leads to computations using big dense matrices.



High computational complexity

Challenge

The data used leads to computations using big dense matrices.



Sommaire

Development Plan

Step 1 Theoretical study + Matlab new methods implementation

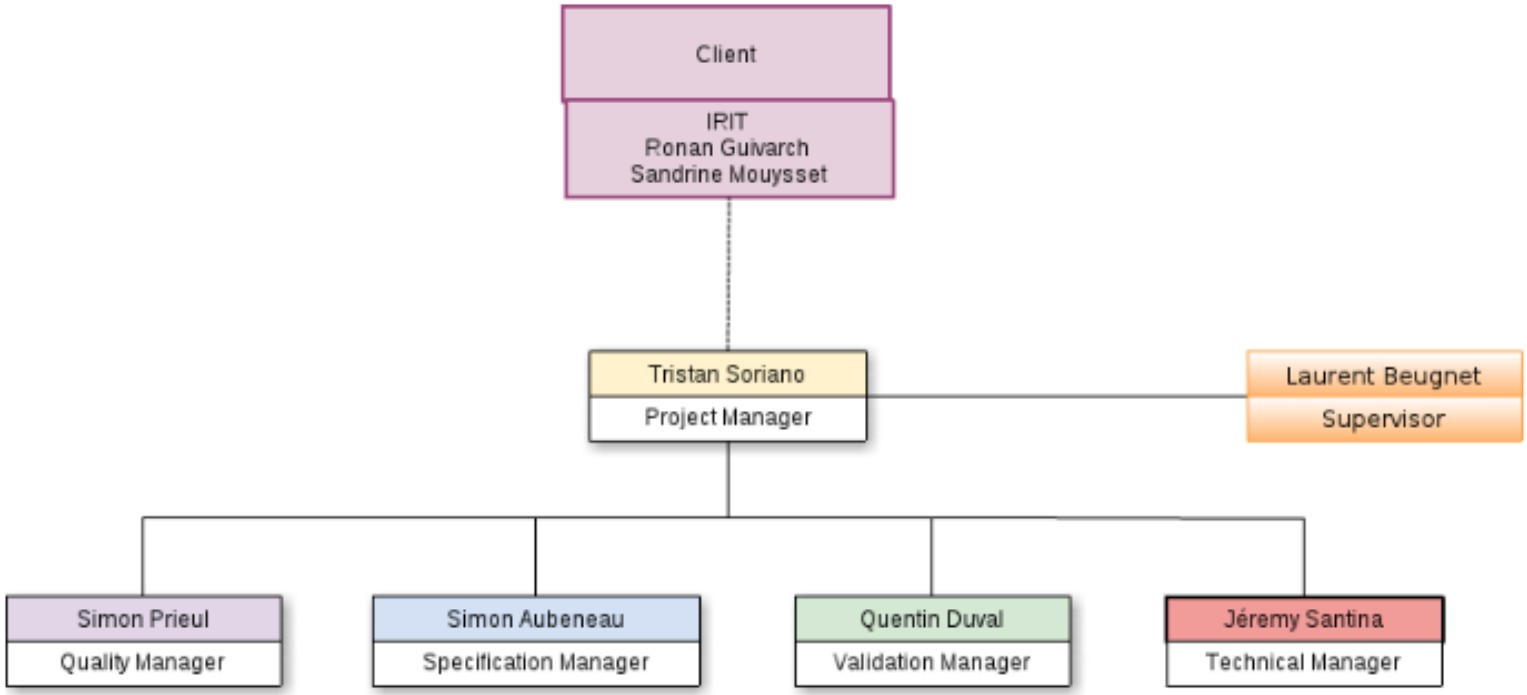
Step 2 Code documentation + Fortran interfaces specification

Step 3 Code refactoring + Fortran implementation

Step 4 Validation

Risk Description	**Probability**	**Impact**	**Action**
Unsatisfied client	Light	Heavy	Specification with the client
Unsuitable hardware resources	Medium	Heavy	Lighter test creation, early access request
Insufficient knowledge	Heavy	Medium	Increase the time dedicated to each risky task

Team organization



Technologies

Sommaire

Environment

Source code

A package containing source code file have been provided: functional but with a poor design.

Configuration and setup

- Libraries: MPI, Lapack and Arpack
- F90 Compiler: GFortran
- Configuration: Makefile

Sommaire

Changing source code

Purpose

Improve software quality and portability

1. Changing the types "REAL" and "REAL(KIND=*)" into "DOUBLE PRECISION" and "INTEGER(KIND=*)" into "INTEGER"
2. Changing the types "CHARACTER*x" into "CHARACTER (LEN=x)"
3. Changing the types of some "INTEGER" variables used as boolean into "LOGICAL"

Organising source code

Purpose

Improve readability and maintainability of the code

1. Fortran keywords transformed in UPPERCASE
2. Removing of the commented code
3. Declarations of parameters and variables at the beginning of a subroutine or program : one declaration line, ordered by type and name
4. Removing the semicolons : only one instruction per line
5. Removing the unused variables

Renaming and translating

Purpose

Improve readability and internationalization

1. Translation of the comments helping the understanding of the code
2. Rename methods : standardization (underscore to separate the different words), translation in english, better naming, blank space after each comma in the signature, blank space after each comma in the signature
3. Rename parameters of methods and variables : standardization (underscore to separate the different words), translation in english, better naming
4. Translation of all the “PRINT” messages (displayed in the console during execution)

Sommaire

Autogeneration with Doxygen

Produced output formats

- *HTML*: Documentation readable by a web browser
- *PDF*: Documentation written in LaTex to generate a PDF file

Targets

Description for each module, program, routine and parameters.

Example

```
1 !> Here is a brief description of the routine.
2 !! @details Here is a detailed description.
3 !! @param[dir] param Here is the parameter description
4 SUBROUTINE my_routine(parameter)
```

How to proceed

Main issues

1. Renaming and reordering parameters before documenting
2. Repetitive work : 62 routines and 249 parameters

Automated comments writing

A small software have been programmed to automatically write headers.

How to proceed

Automated comments writing

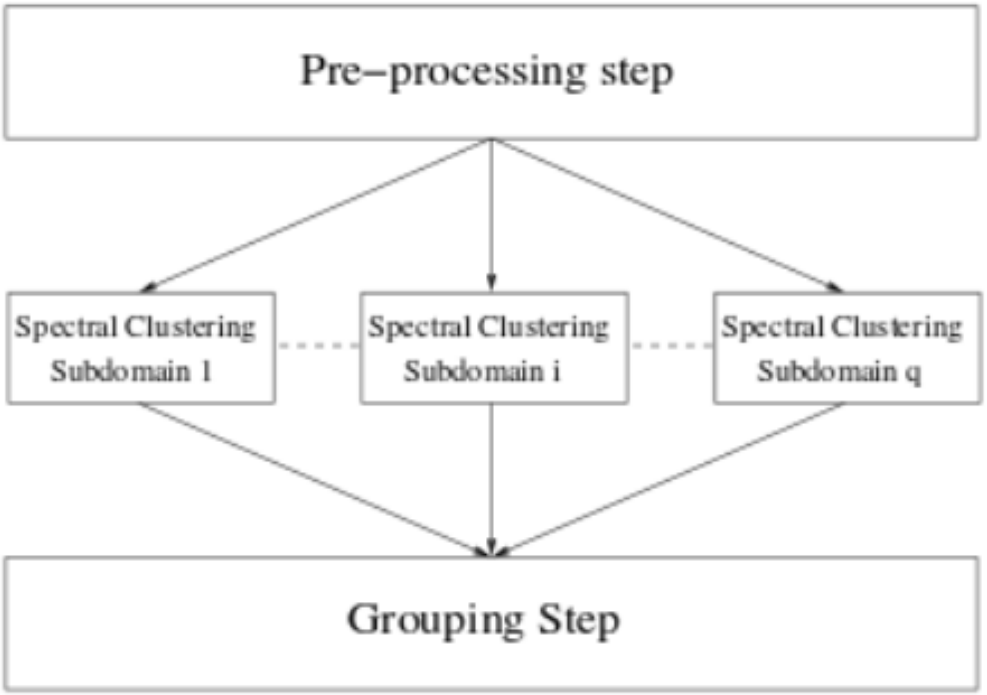
A small software have been programmed to automatically write headers.

1. Classify modules, methods and parameters
2. Write descriptions in a database
3. Extract information from the database
4. Write header into the source file

Before and after I

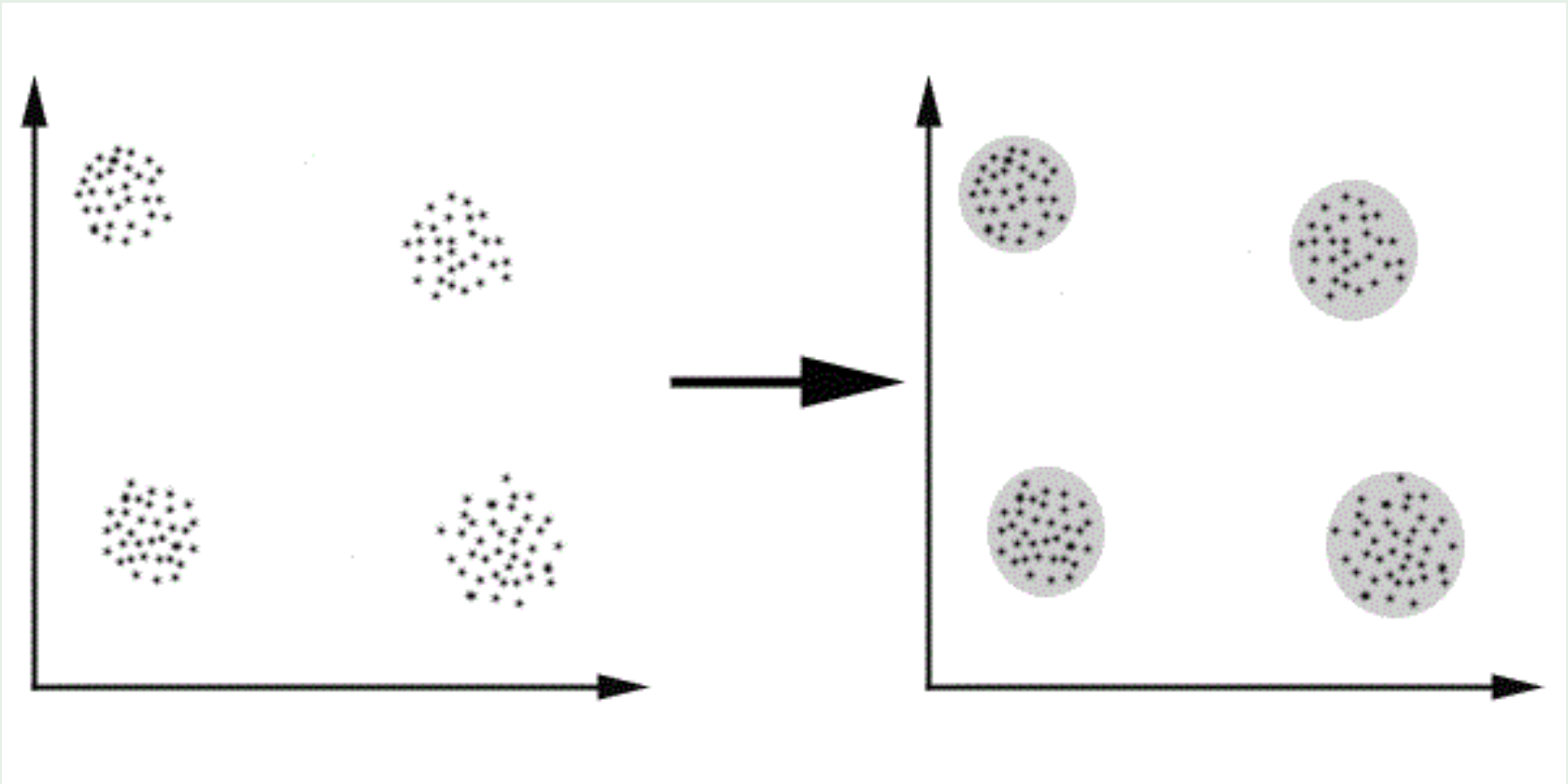
Explanations

- Spectral Clustering
- Kernel K-Means
- Mean Shift



Explanations

Example



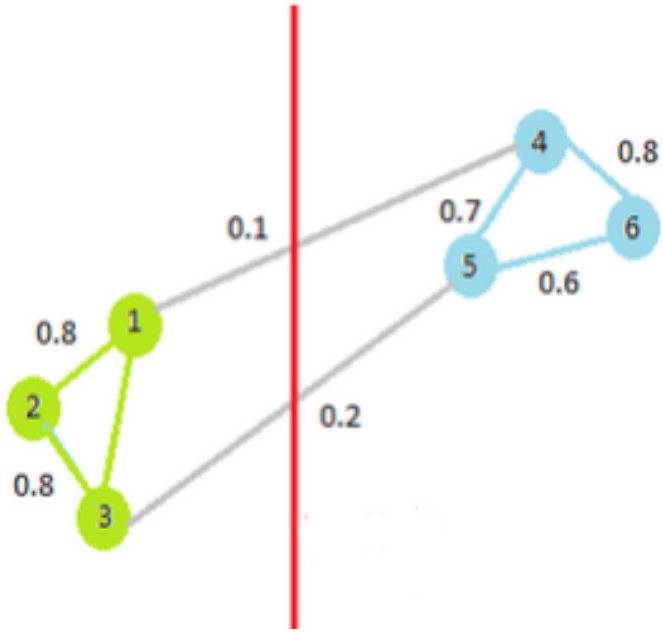
Sommaire

Description

Main idea

Create a weighted graph between points

- Each vertice weight represents the similarity between points
- The final graph is cut on the lowest similarity vertices



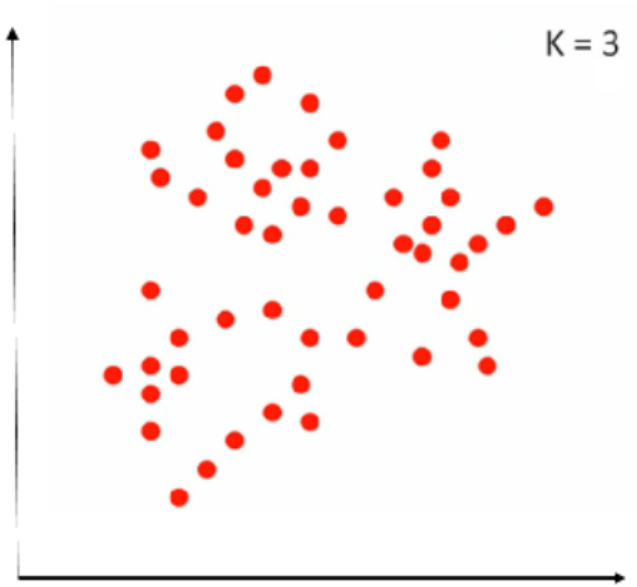
Sommaire

K-Means

Main idea

Select random centers for cluster and move them until they reach centers of density.

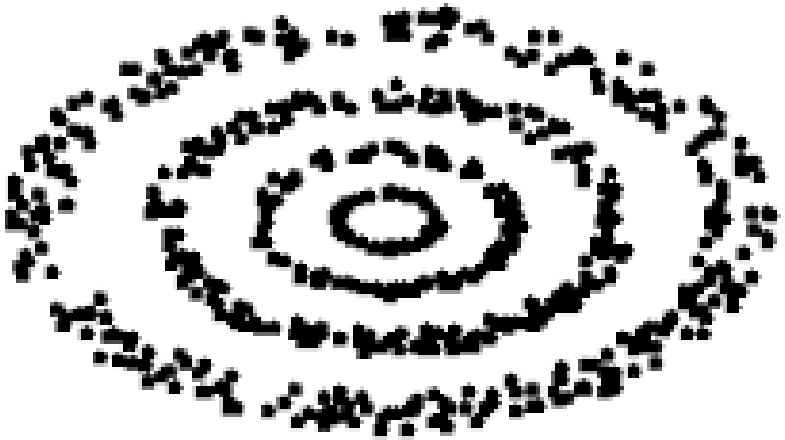
1. Find minimum distances to cluster centers
2. Compute density centers
3. Select new cluster centers
4. Continue until cluster centers and density centers are similar

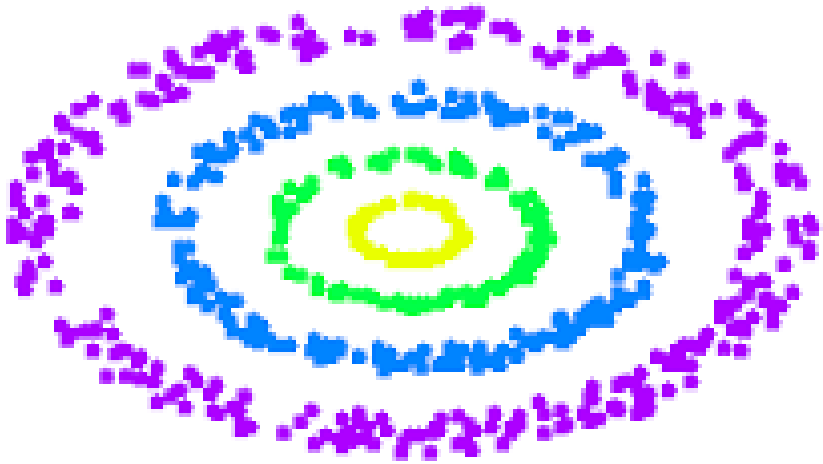


K-Means

Limitation

K-Means does not work for non-linear separations.

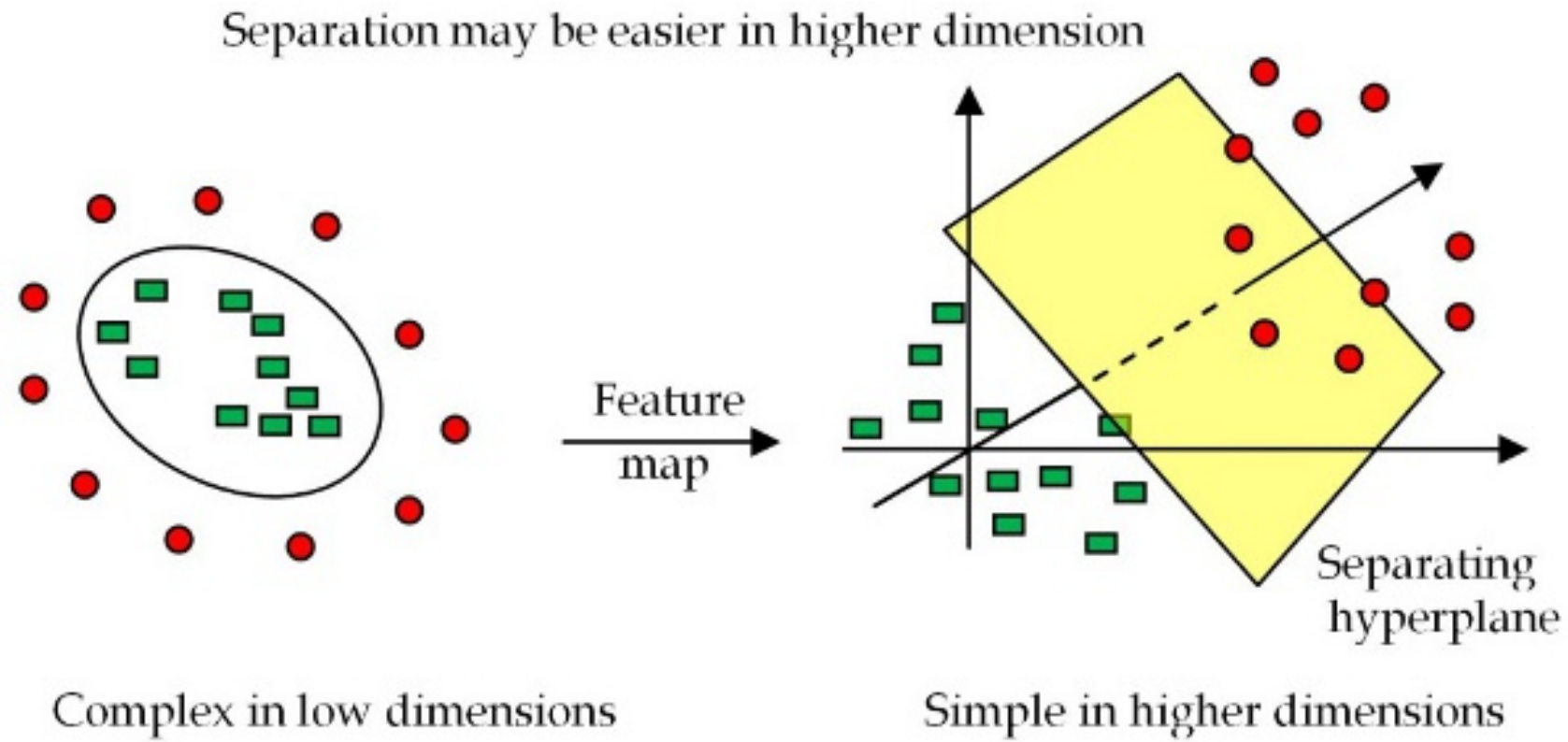




Improving K-Means

Main idea

Using K-Means on a space with higher dimensionality to highlight separations.



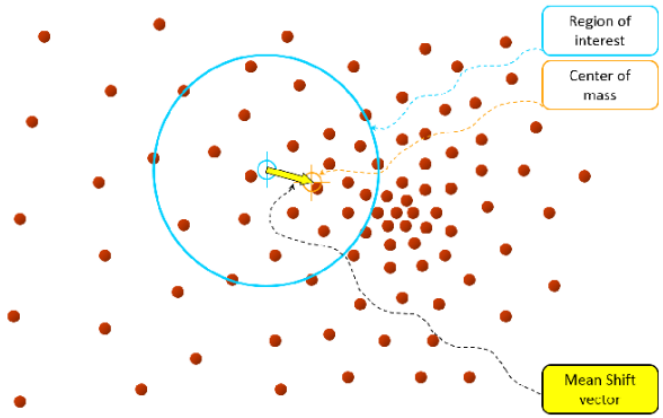
Sommaire

Description

Main idea

Move a window over the data set following the density gradient.

1. Compute the mean of density in the specified window
2. Compute difference between center of the window and computed mean
3. Select the computed mean as the new center of the window
4. Restart until difference close to 0



Sommaire