Data Clustering

Introduction

- Data clustering : finding structures in unlabeled data

- Possible Application : Marketing, biology, insurance, earthquake studies ...

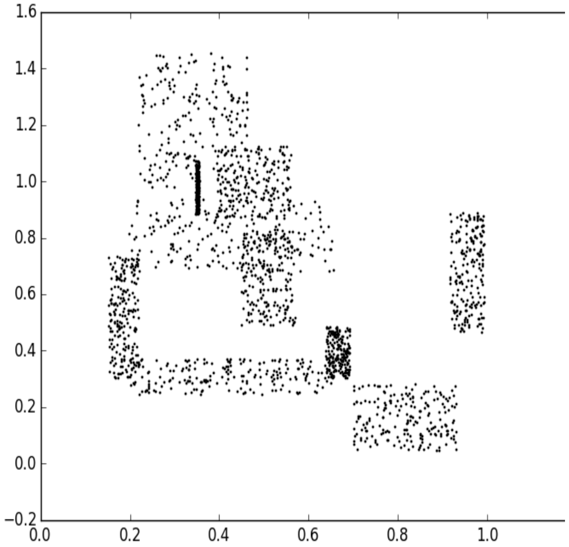
- Problems complexity : number of dimensions, distance definition, number of points

Content

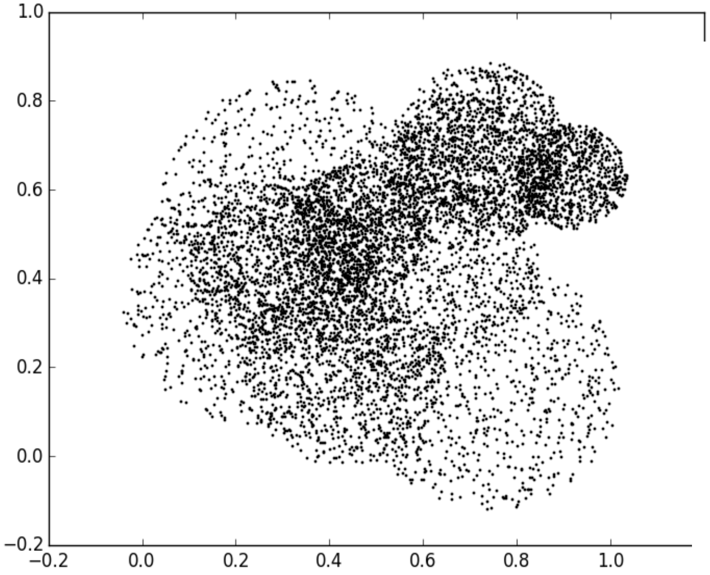
- Context generation
- Resolution
- Results

Context generation

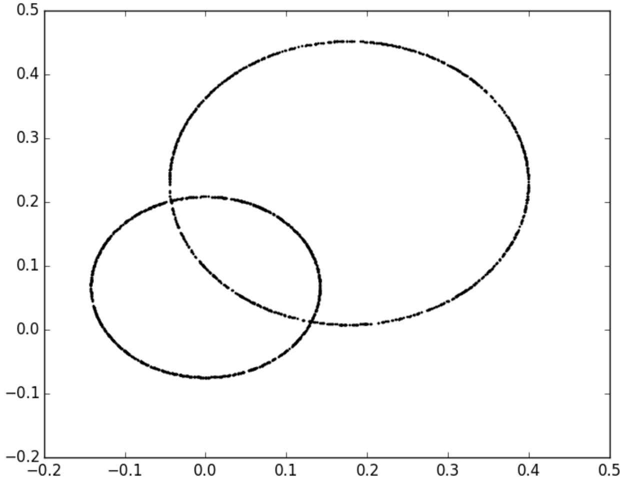
- Generates cluster and creates points in



square



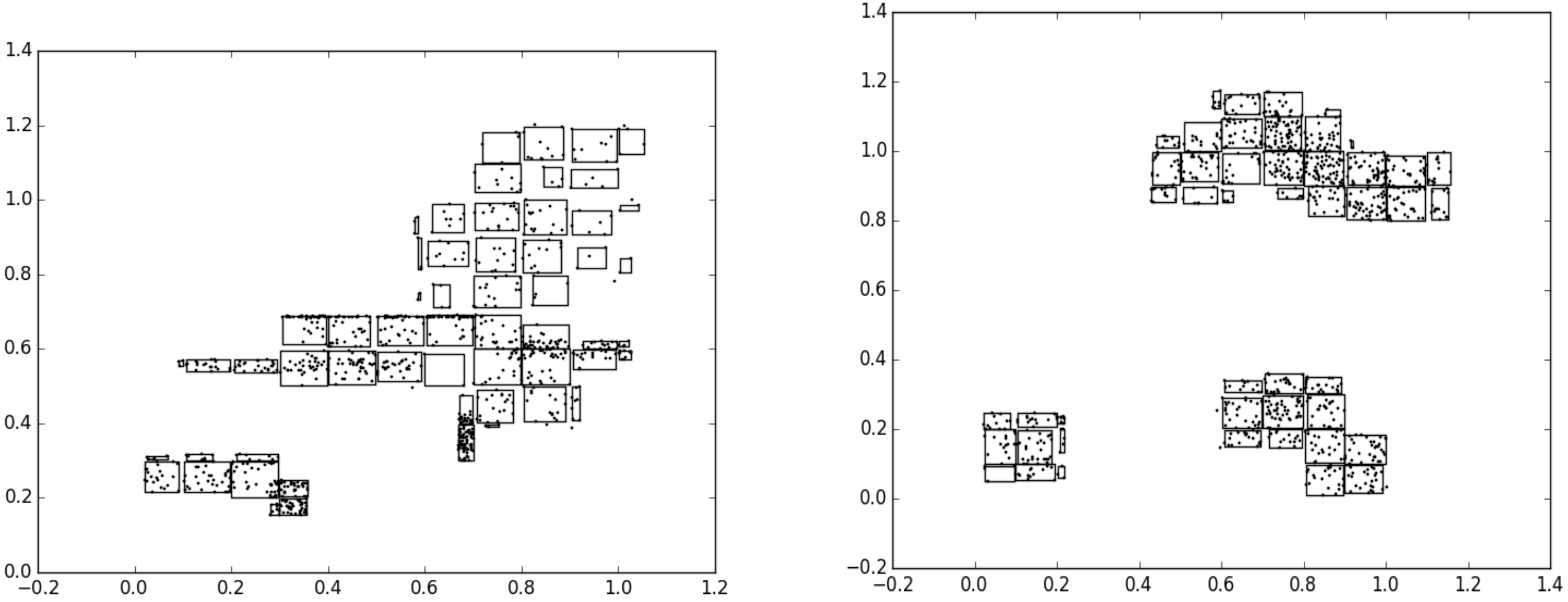
sphere



circle

Resolution

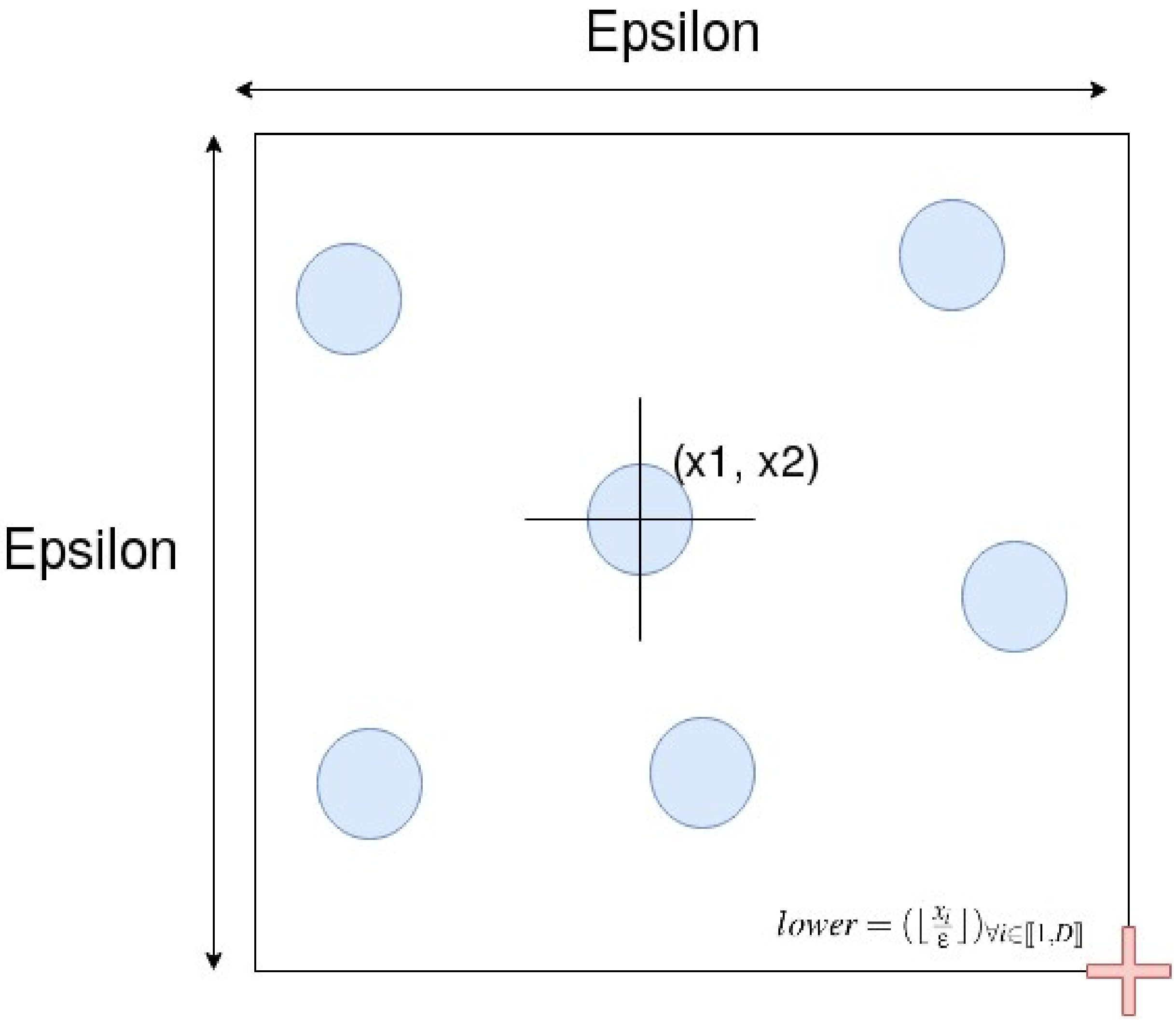
- Create an hash table : local sensitive parsing



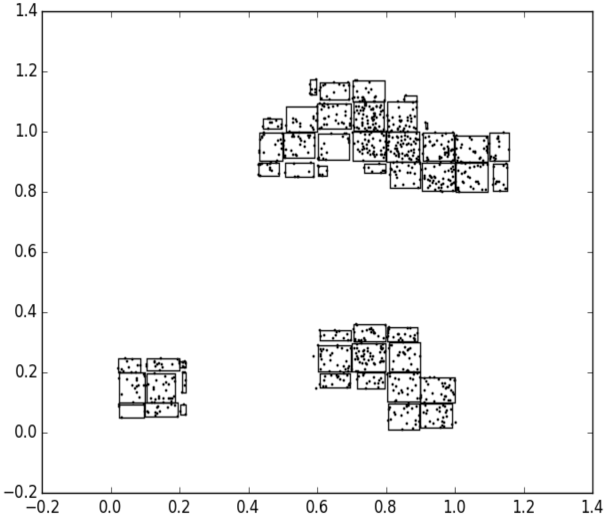
Exemple of hashing

Resolution

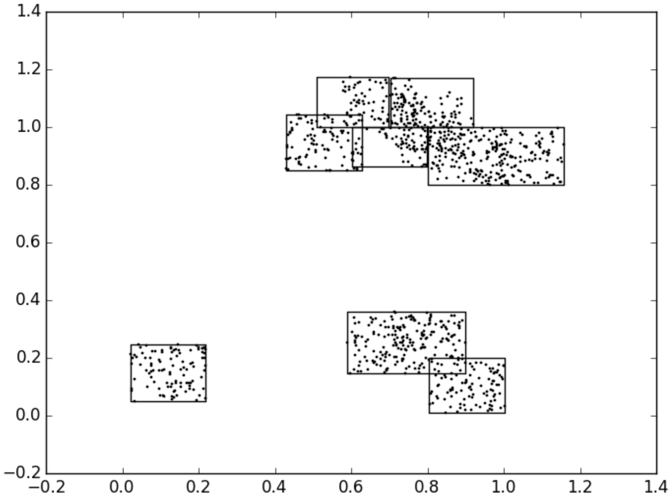
- Hashing principle:


Epsilon
Epsilon
(x1, x2)
$lower = (\lfloor \frac{x_i}{\varepsilon} \rfloor)_{\forall i \in [\![1, n]\!]}$

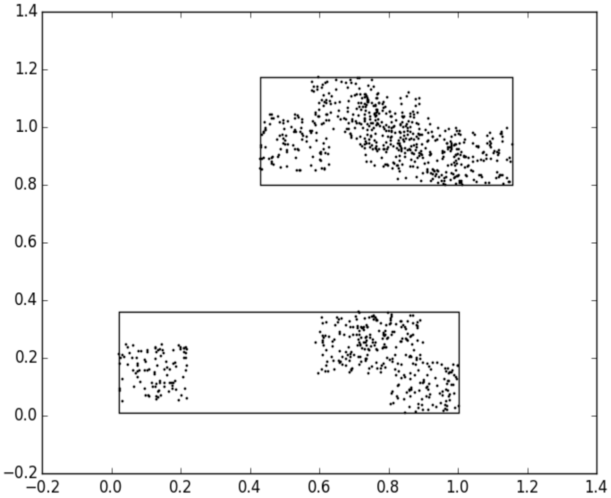
Nearest Neighboor



1st step clustering



N/2 step clustering



N step clustering

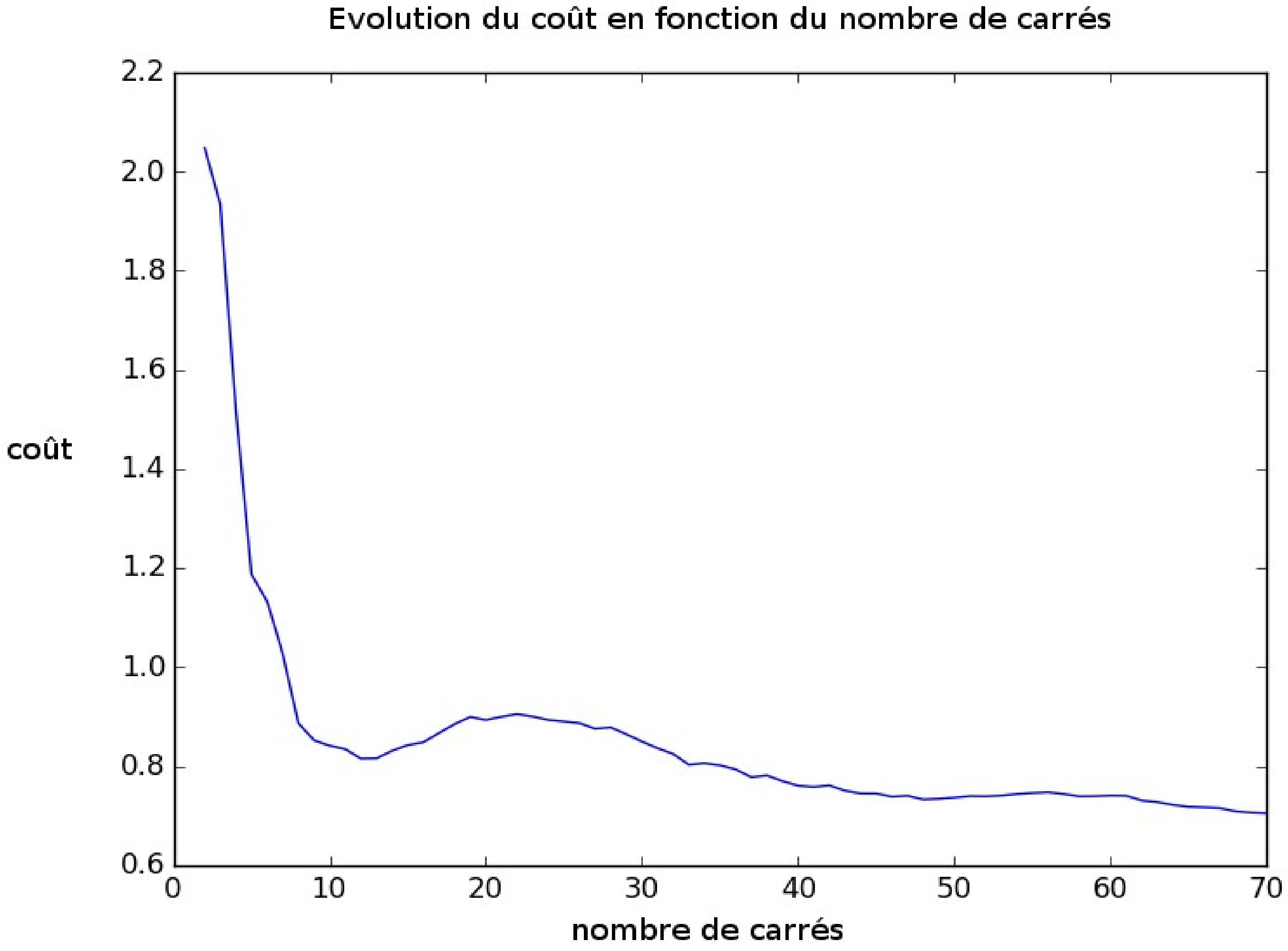
Animation

Data structure used for NN

- First method : O(n^2)
- Second method : Sorted list O(nlogn)

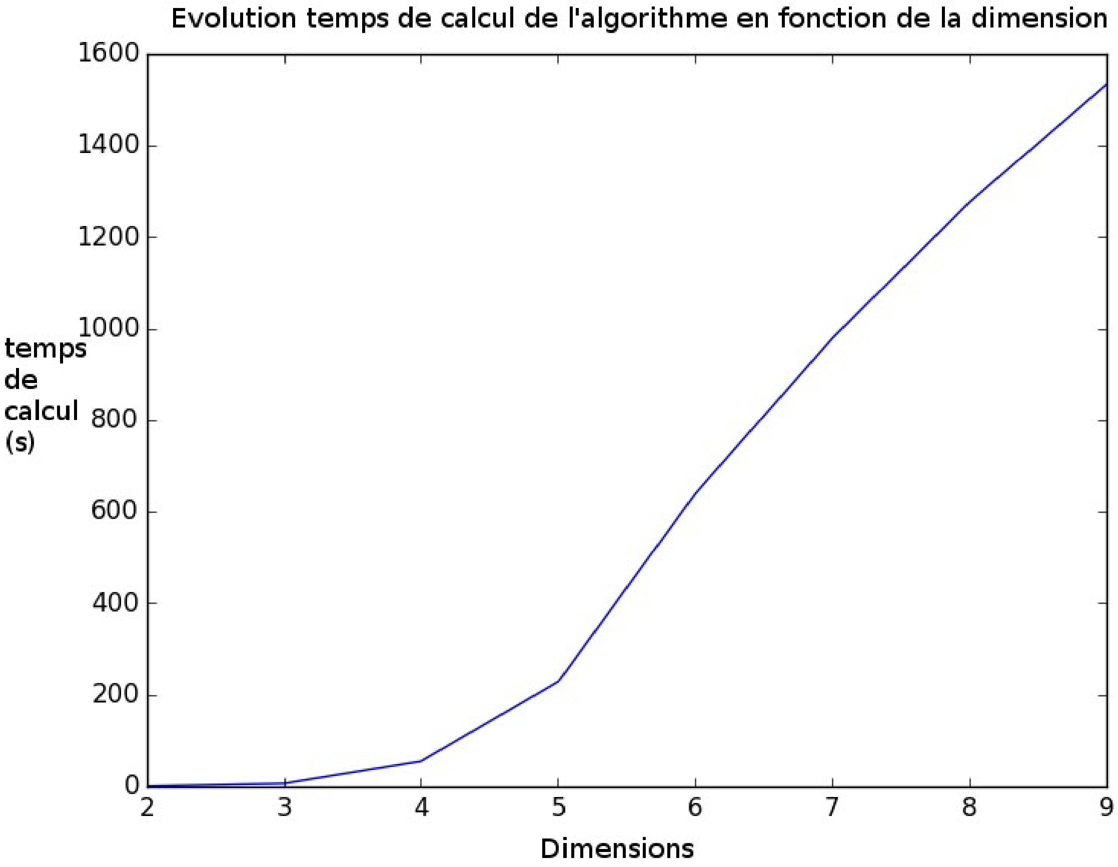
Results

- Cost evolution depending number of cluster



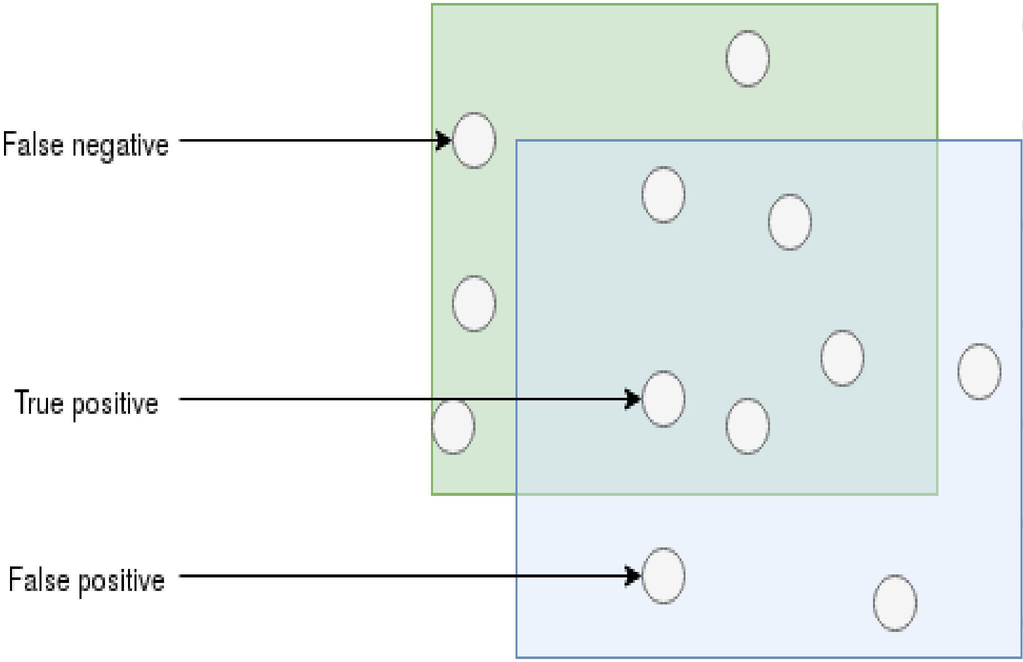
Results

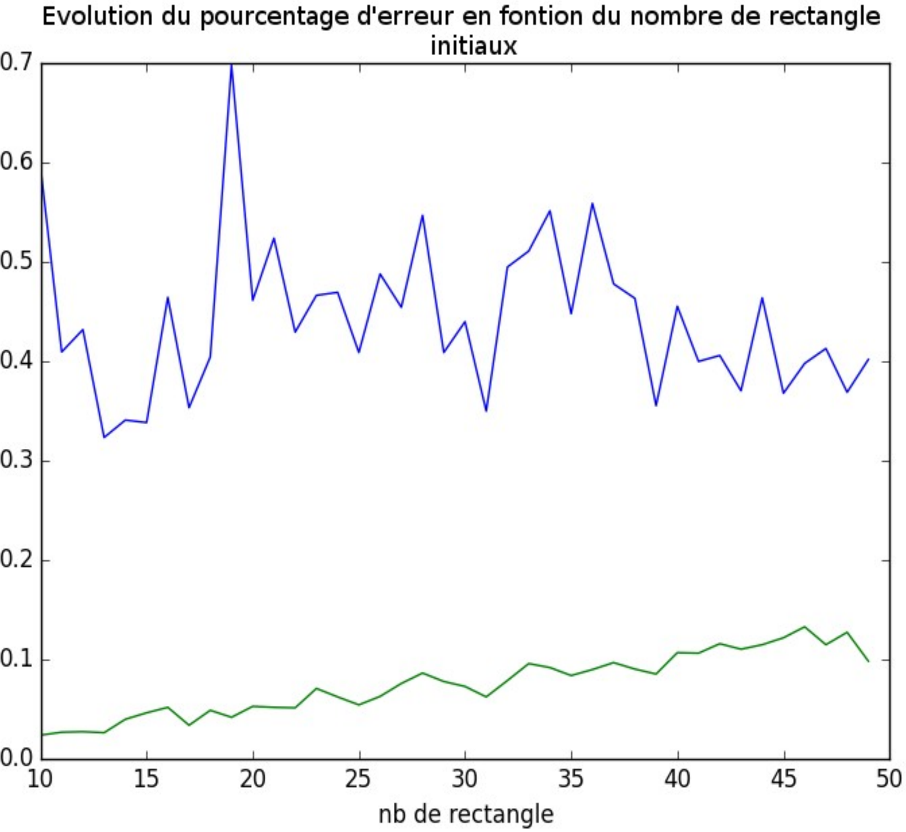
- Curse of dimensionality (with method 1)



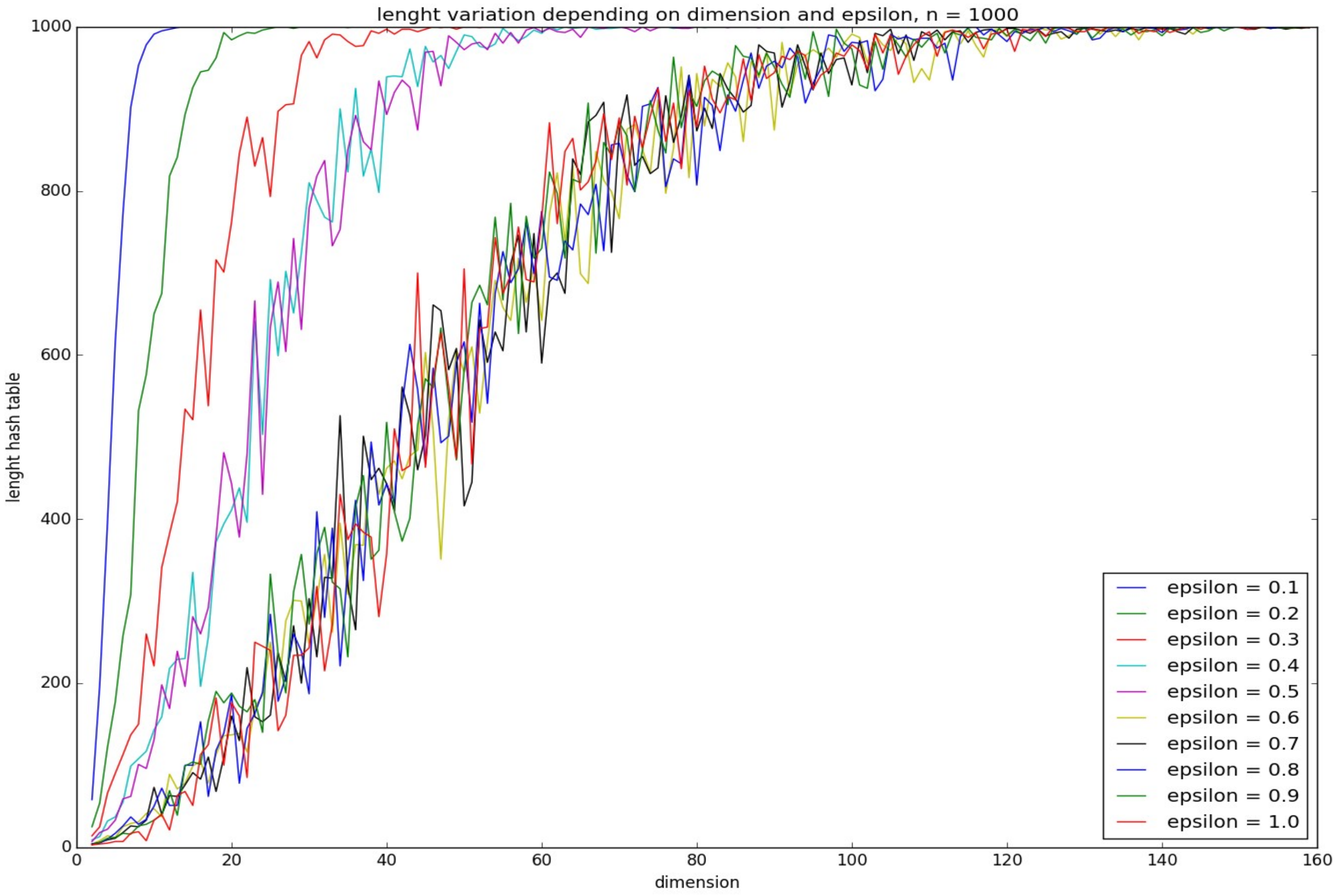
Results

- Error in clustering : evalutate percentage of false positve(green) and false negative(blue).





Results



- Remarque : high dimension points isolated

results

- Optimization of NN :

Dimension	Naive NN (s)	Sorted list NN (s)
2	0.16	0.1
3	11	4
4	211(3min)	16
5	2 151(35min)	81
6	16 082(4h)	175
7		198
8		233