Clustering Big Data

Violaine Antoine

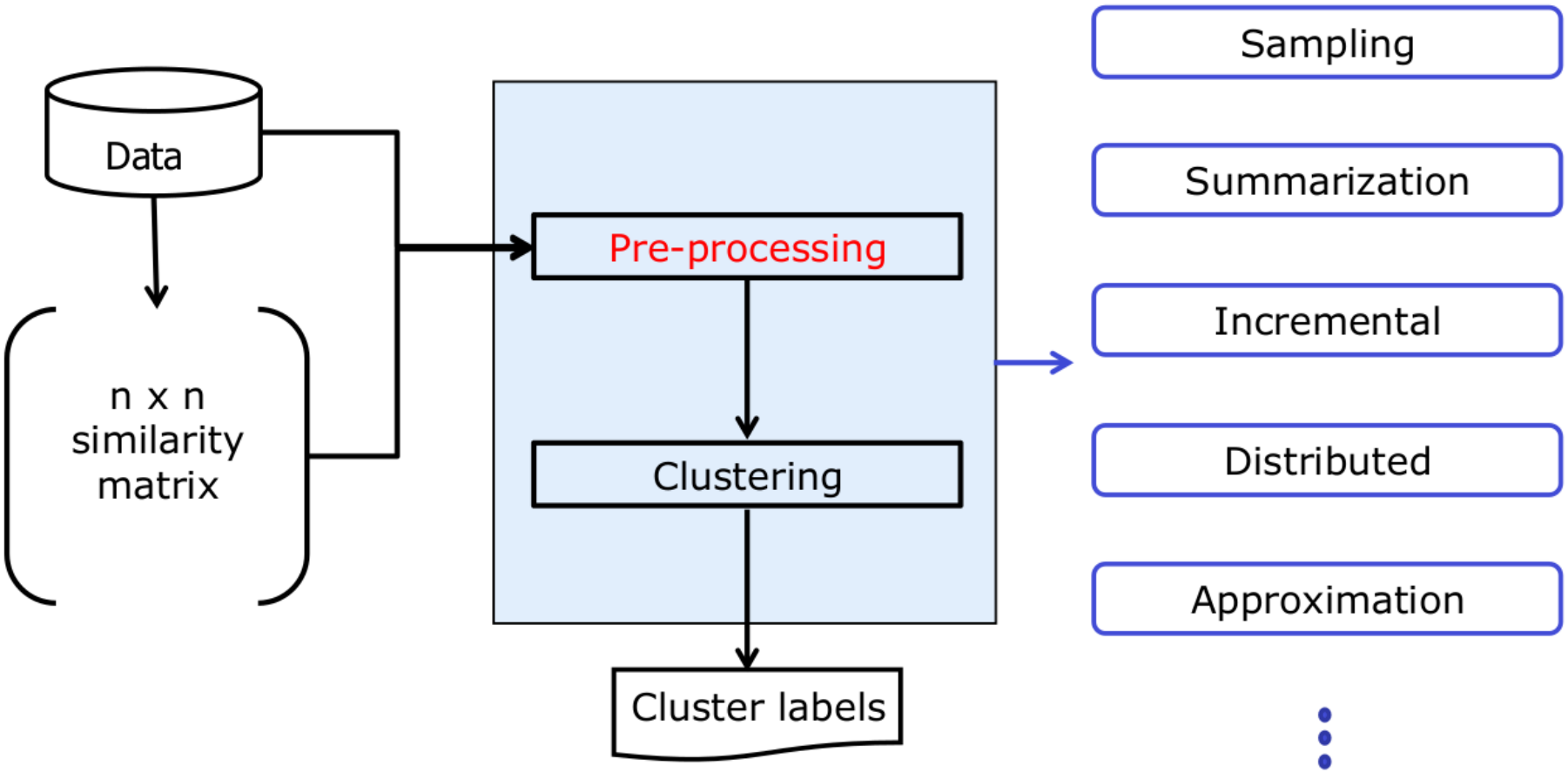
ISIMA / LIMOS

January, 2019

Algorithms complexity

categories	algorithm	complexity
hierarchical	single-link	$O(n^2)$
	complete-ling	$O(n^2 \log n)$
	average-link	$O(n^2 \log n)$
partitional	dbscan	$O(n^2)$
	dbscan with spatial index	$O(n \log n)$
	k-means	$O(ncp)$
	FCM	$O(n)$
	Kernel k-means	$O(n^2 c)$
grid-based	SOM	$O(n^2 m)$
	Ant	/

Clustering big data



A. Jain & al, Clustering Big Data, 2012.

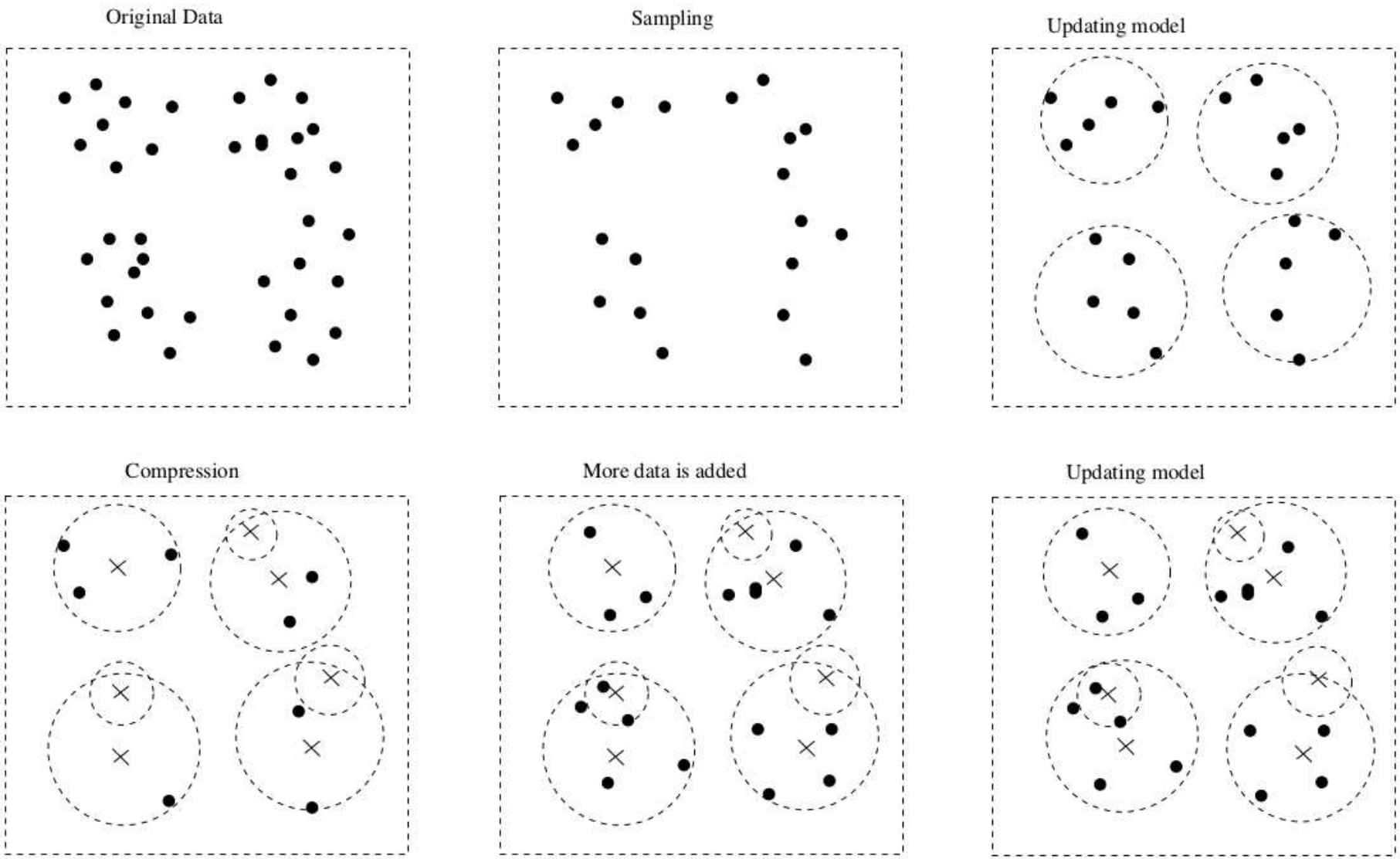
Strategies can be mixed.

Sampling and Summarization

Preprocessing step to reduce the information (data dependent)

- create several samples for
  - parallel clustering
  - online clustering
- summarize data without loss of information

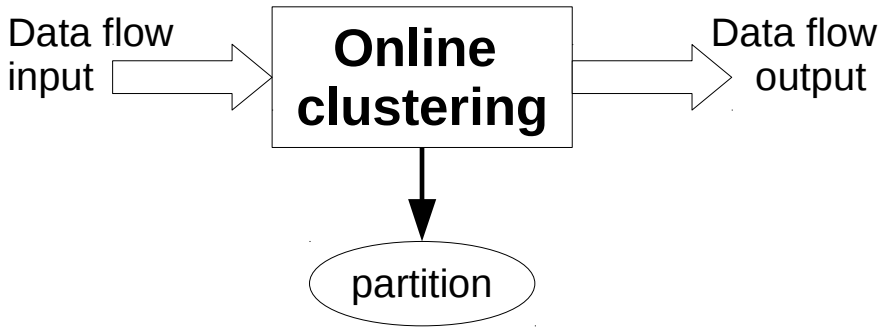
Sampling and Summarization : Scalable k-means



J. Béjar, Strategies and Algorithms for Clustering Large Datasets : A Review, 2013

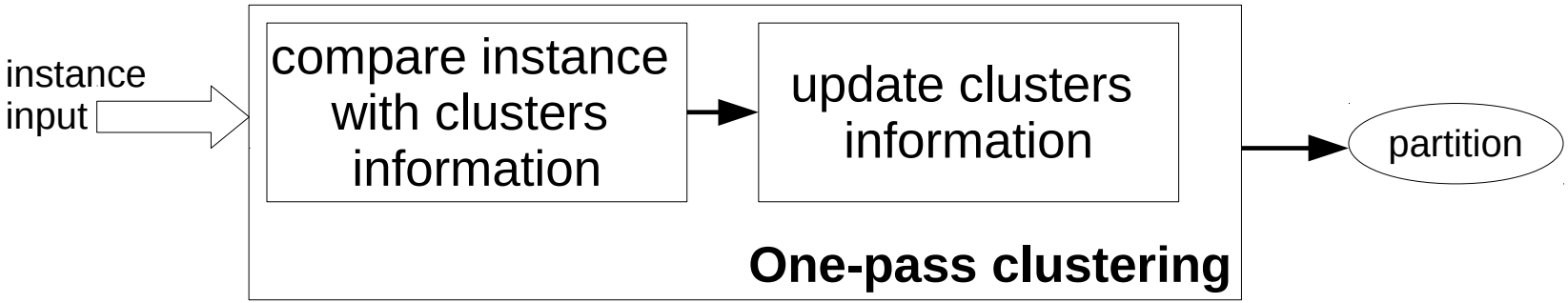
Incremental

- online clustering



- one-pass strategies
  - reduce the number of scans of the data to only one sequential process
  + enable to fit in memory
  - low accuracy $\Rightarrow$ used as a first stage to obtain general information

One-pass strategy : the leader algorithm



- Assign first object $\mathbf{x}_1$ to ω_1,
- For each objects $\mathbf{x}_i$ from 2 to n
  1. compute $d^2(\mathbf{x}_i, \omega_j)\ \forall \omega_j \in \Omega$
  2. if $d^2(\mathbf{x}_i, \omega_j) > \theta\ \forall \omega_j \in \Omega$ then create a new cluster
     otherwise, put $\mathbf{x}_i$ in the closest cluster and recompute the centroid.

Use of One-pass strategy : Scalable hierarchical clustering

Method l-Single Link (l-SL) :

- Apply a one-pass algorithm (ex : the leader algorithm with θ) : $O(nc)$
- Apply modified hierarchical clustering with centroids obtained : $O(c^2)$
- Replace centroids by its instances to obtain clusters.

Method al-SL : variant of l-SL to obtain the same dendrogram as SL.

If $n >> c$, reduce the complexity $O(n^2)$ to $O(nc)$

B. Patra & al. A distance based clustering method for arbitrary shaped clusters in large datasets, 2011.

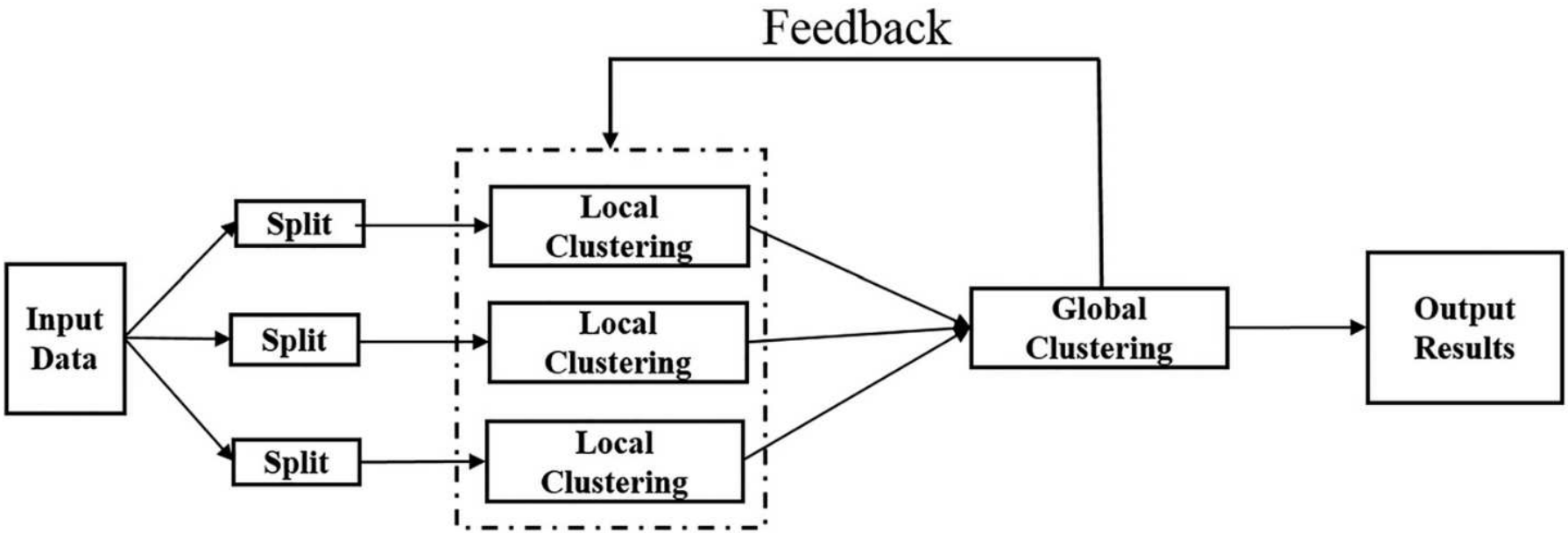
Use of One-pass strategy : Scalable hierarchical clustering

Dataset	Distance (h)	Method	Time (s)	Rand index (RI)
Pendigits	90.0	l-SL	0.46	0.999
	90.0	al-SL	0.79	1.000
	90.0	SL	392.59	–
	70.0	l-SL	1.38	0.993
	70.0	al-SL	2.14	1.000
	70.0	SL	430.46	–
Shuttle	0.02	l-SL	9.13	0.999
	0.02	al-SL	19.38	–
	0.02	SL (40,000)	6929.77	
	0.01	l-SL	9.32	0.999
	0.01	al-SL	20.27	–
	0.01	SL (40,000)	6929.77	
GDS10	700	l-SL	0.50	0.999
	700	al-SL	1.15	1.000
	700	SL	4105.35	–
	900	l-SL	0.26	0.999
	900	al-SL	0.62	1.000
	900	SL	4105.35	–

	n	p
Pendigit	7494	16
Shuttle	58 000	9
GDS10	23 709	28

Distributed

- parallel clustering
- map reduce



M. Chen & al, Clustering in Big Data.

Distributed : Divide and conquer strategy : canopy

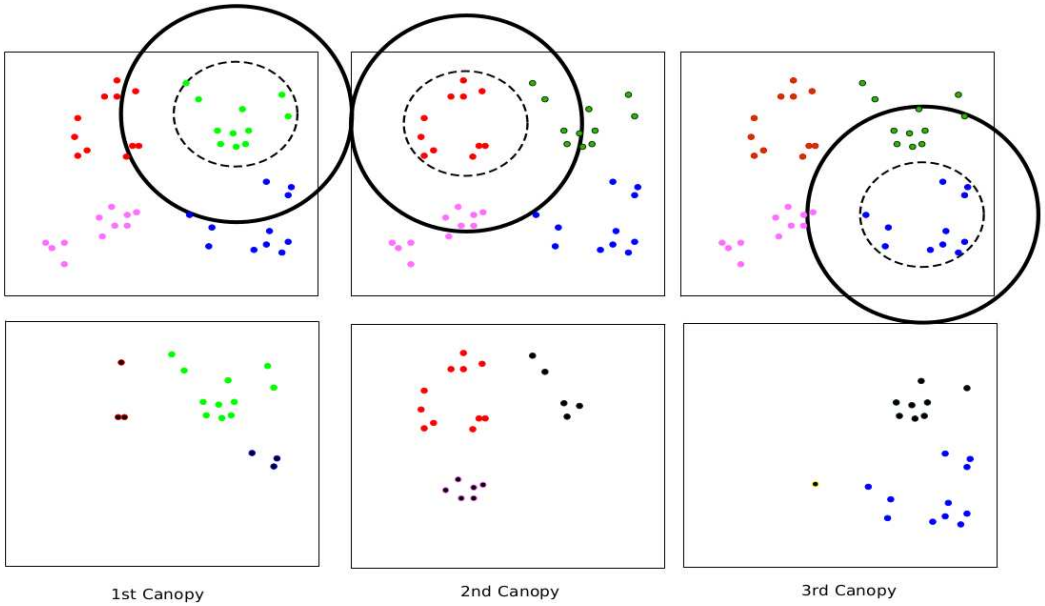
Method :

1. create overlapped subsets called canopies using distances d_{min} and d_{max}
2. perform for each canopy a clustering algorithm
3. merge subclustering results

Distributed : Divide and conquer strategy : canopy

Create canopies :

1. Set $\mathcal{S}_{min}$, $\mathcal{S}_{max}$ the sets containing the data points
2. While $\mathcal{S}_{min}$ is not empty
   - Select at random a point in $\mathcal{S}_{min}$ considered as the center of the canopy c_i
   - Add in c_i all points closer than d_{max}
   - Remove from $\mathcal{S}_{max}$ all points closer than d_{min}
   - Remove from $\mathcal{S}_{min}$ all points closer than d_{max}



J. Béjar, Strategies and Algorithms for Clustering Large Datasets : A Review, 2013

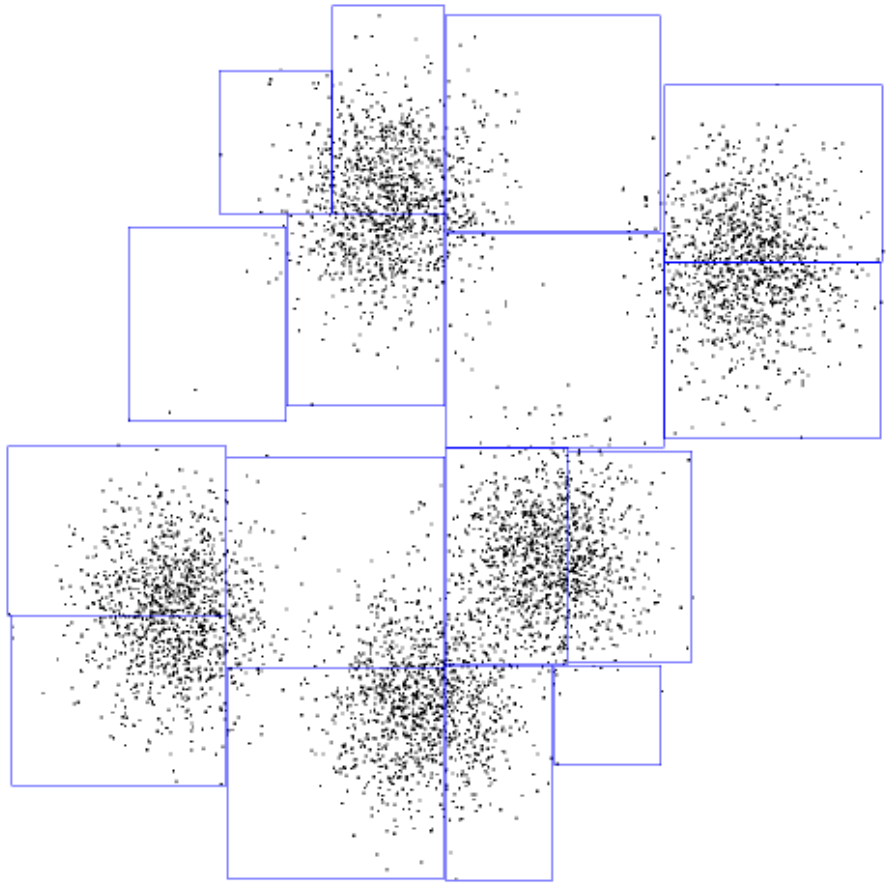
Approximation (algorithm dependent)

Some computations are saved/approximated with reduced impact

- mainly related with the distances calculation
- ex : hierarchical clustering

Approximation : the Indexed k-means method

Store points cloud an centroids in a kd-tree to avoid distances computation



Computational cost for an iteration is $\log(2^p klog(n))$
$\Rightarrow$ high dimension does not save time !