Université Paris 13 , Sorbonne Paris Cité,
LIPN, UMR 7030 du CNRS
99 Avenue J-B. Clément - 93430 Villetaneuse - France

CLUSTERING DE FLUX DE DONNÉES

Mustapha LEBBAH

MCF, HdR

LIPN – Univ. Paris 13

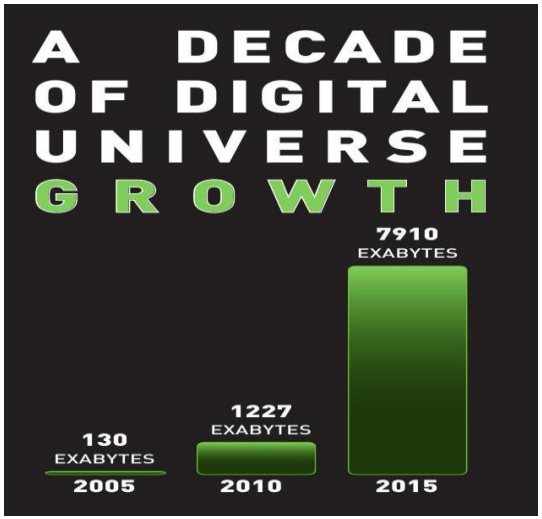
H. Azzag, T. Duong, **M. Ghesmoune (doctorant)**

Journée Clustering, Issy Les Moulineaux le 20 Oct 2015

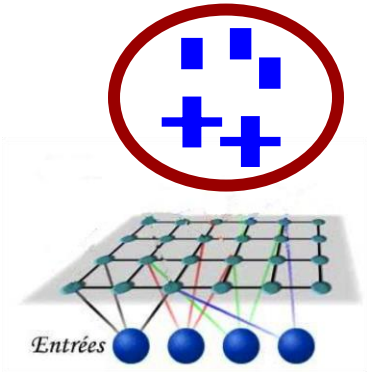
Plan

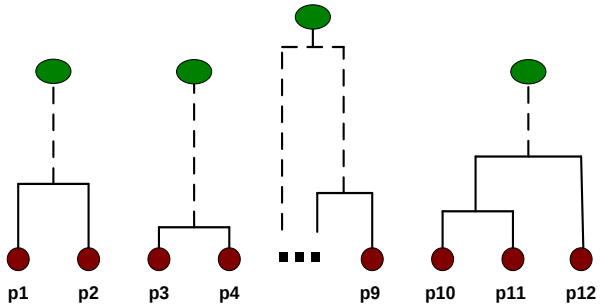
- Contexte
- Le Big Data comme du «Data Stream»
- Stratégie du clustering de flux de données
- Approches utilisées
- Conclusion & perspectives

Partitionnement «clustering» et Big Data

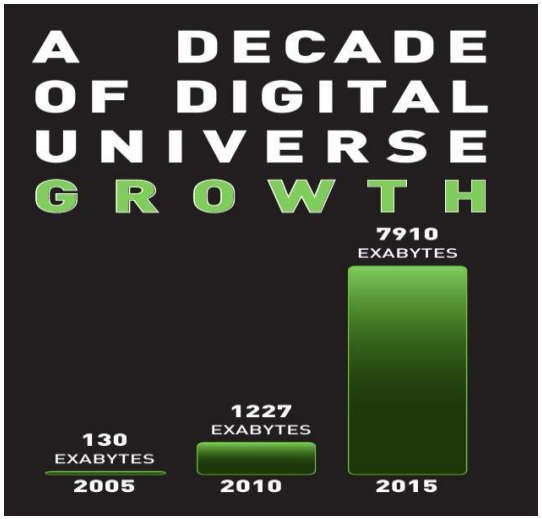


- Approche hiérarchique
- Approche par partitionnement
- Approche spectrale
- Approche à base de modèles auto-organisés
- Versions probabilistes

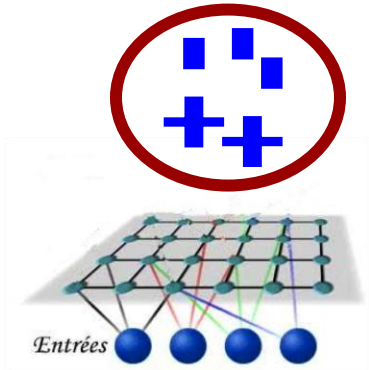


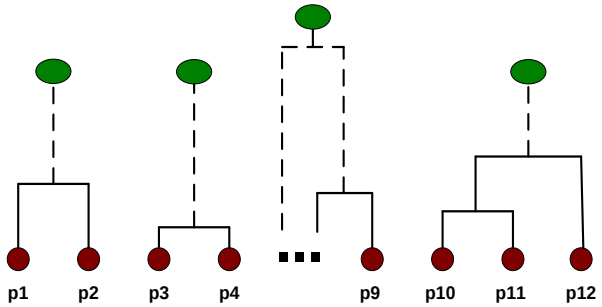


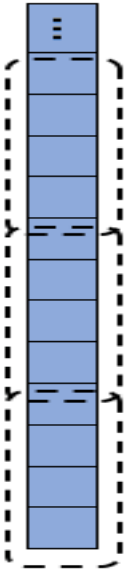
PARTITIONNEMENT «CLUSTERING» ET BIG DATA

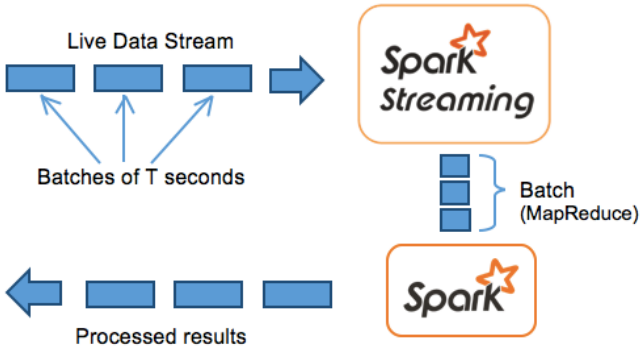


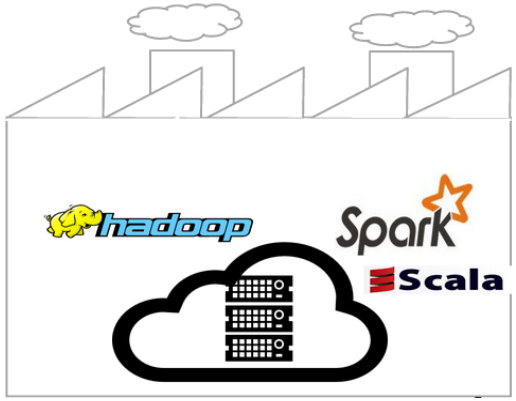
- Approche hiérarchique
- Approche par partitionnement
- Approche spectrale
- Approche à base de modèles auto-organisés
- Versions probabilistes











MLI: An API for Distributed Machine Learning. E. Sparks, et al. In ICDM, 2013.

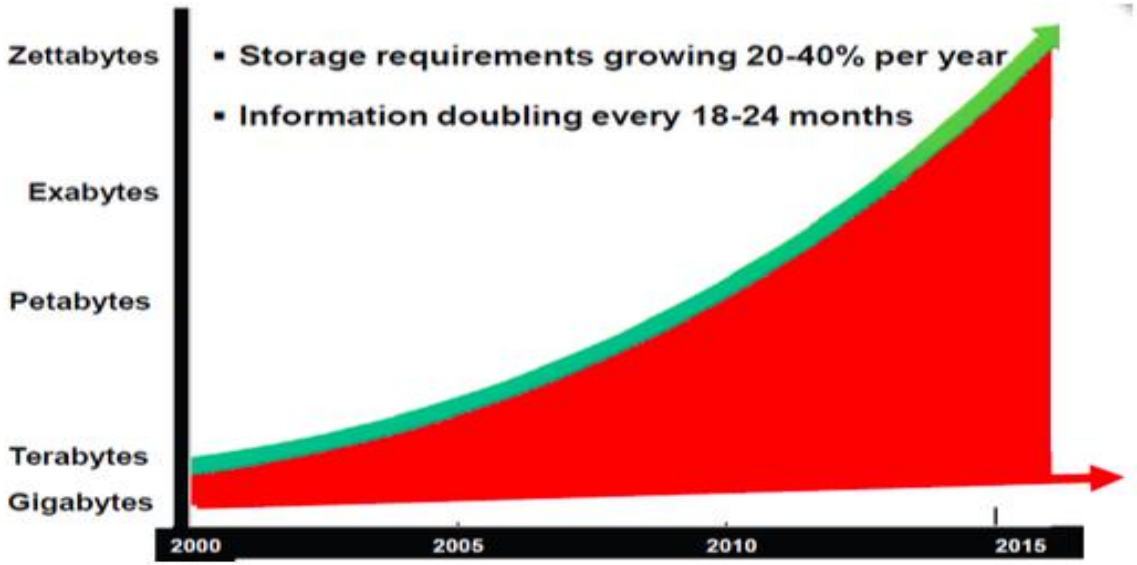
Flux de données / Data Stream

- Les données massives sont collectées à un rythme exceptionnel.
- Besoin d'une réponse rapide

- Exemples:
  - Les données de suivi de comportement pour étudier les changements dans les habitudes (achat/de circulation/ intrusion)
  - Suivi des données météorologiques,
  - La taille des bases de données et des échantillons est énorme

La collecte des données ne s'arrête jamais !

http://www.flightradar24.com

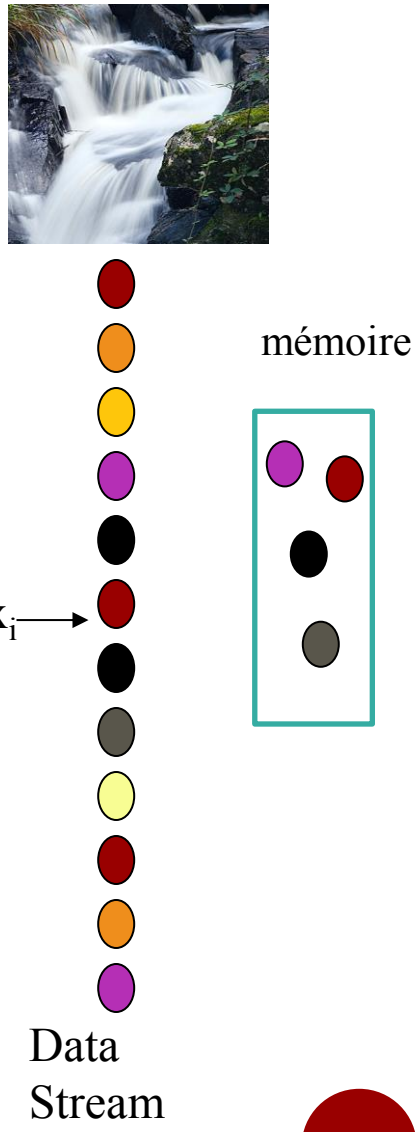


Albert Bifet et al 2015

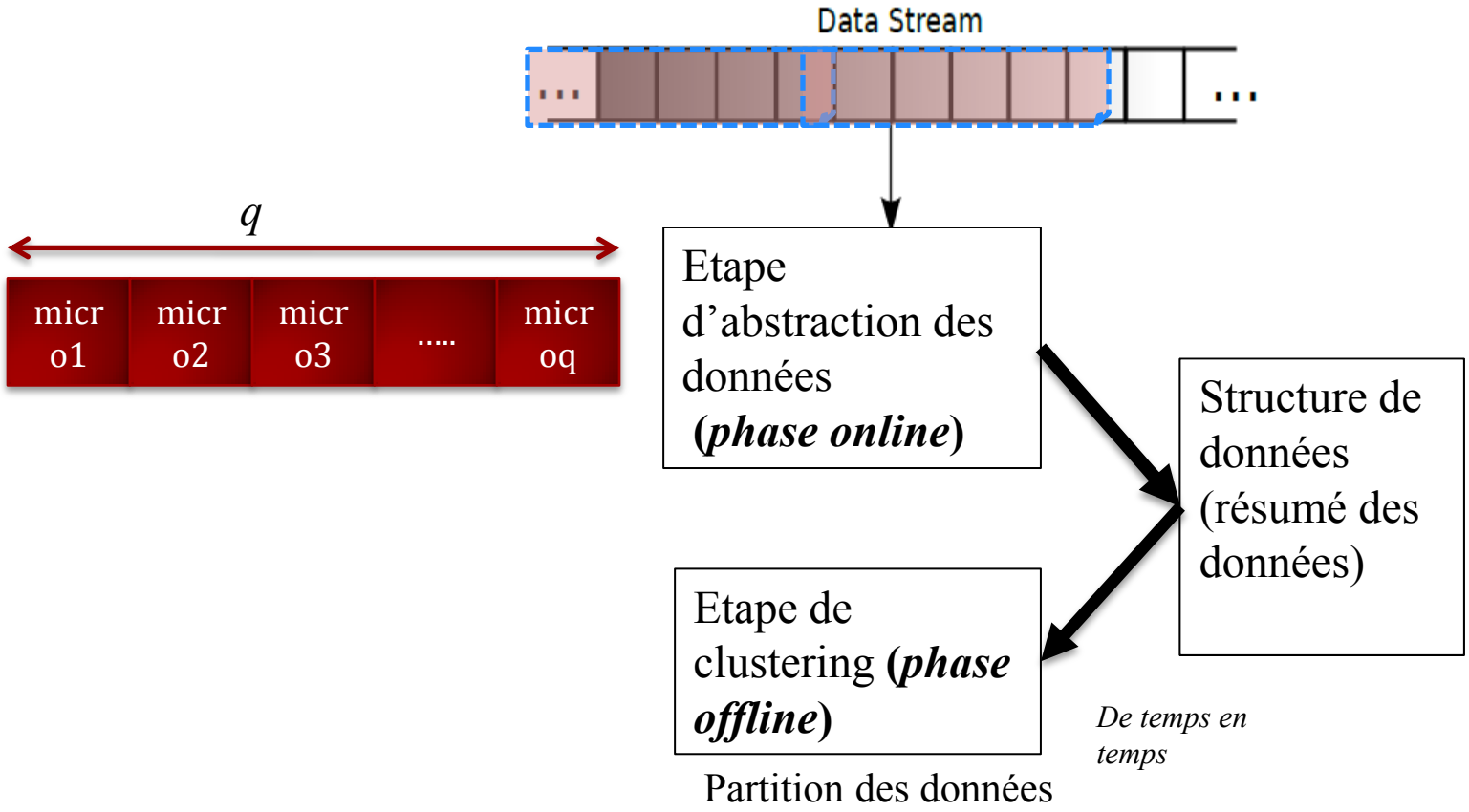
Challenges

Traiter les « Big Data » comme du Data Stream

- Ensemble x_1, …, x_n d'observations
  - Potentiellement infini
  - Non-stationnaire
- Rapide, un seul passage
- Doit être traité dans cet ordre en une passe
- Ressources CPU et mémoire utilisées par élément faibles
- L'accès aléatoire n'est pas permis



Le Big Data comme le Data Stream

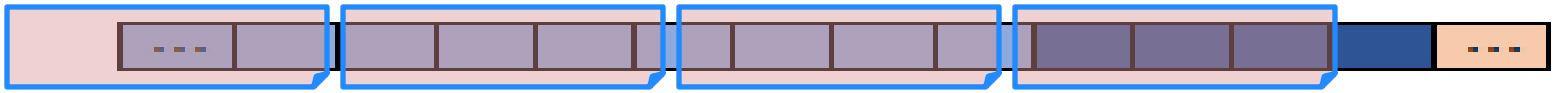


Framework de clustering de flux de données online-offline

Clustering de flux de données

Principe de base

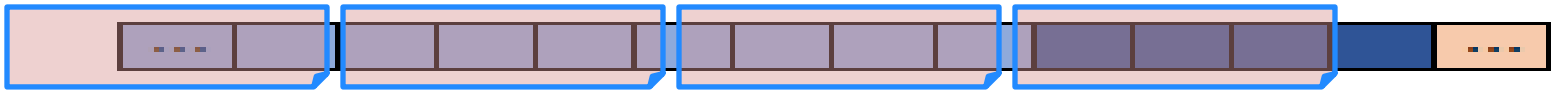
- **Stratégie**
  - **Fenêtre de temps**: diviser de manière indépendante le flux en fenêtres temporelles.

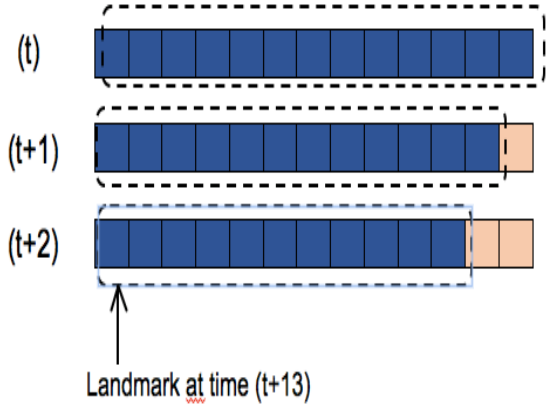


Clustering de flux de données

Principe de base

- **Stratégie**
  - **Fenêtre de temps**: diviser de manière indépendante le flux en fenêtres temporelles.
  - Pondération des fenêtres temporelles (données) -> « fading»

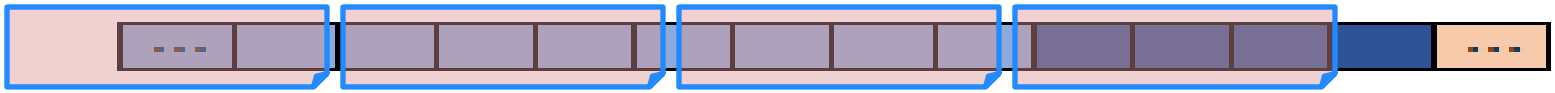


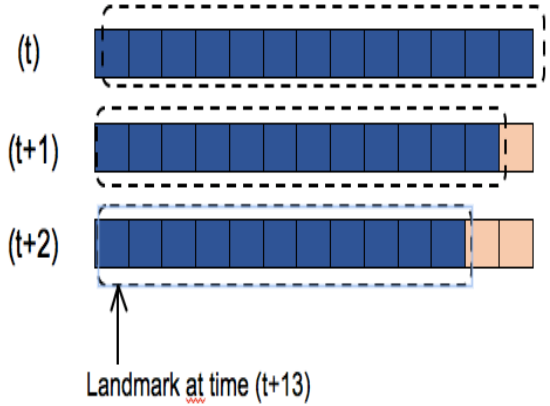


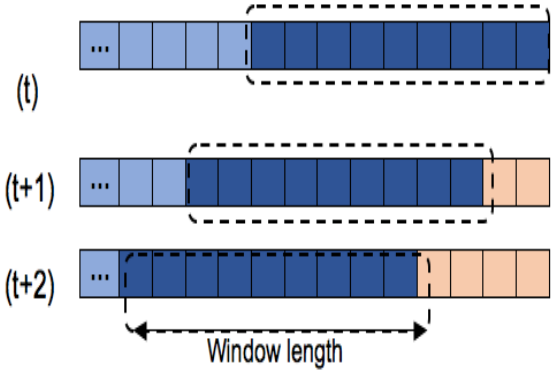
Clustering de flux de données

Principe de base

- **Stratégie**
  - **Fenêtre de temps**: diviser de manière indépendante le flux en fenêtres temporelles.
  - Pondération des fenêtres temporelles (données) -> « fading»



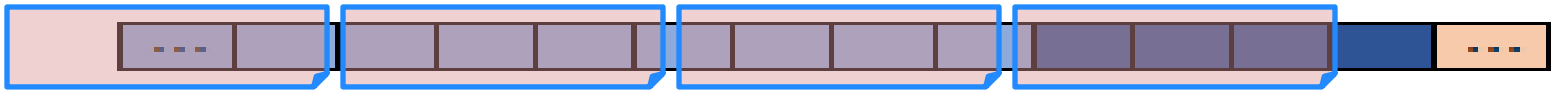


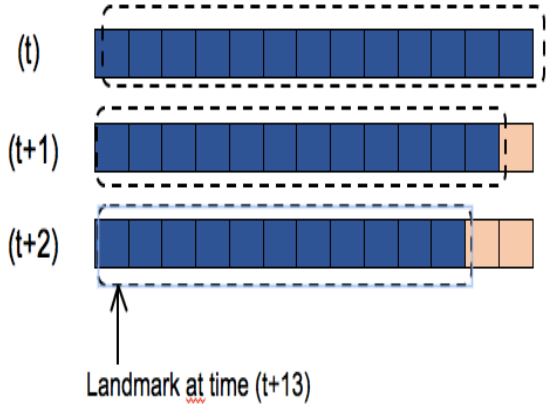


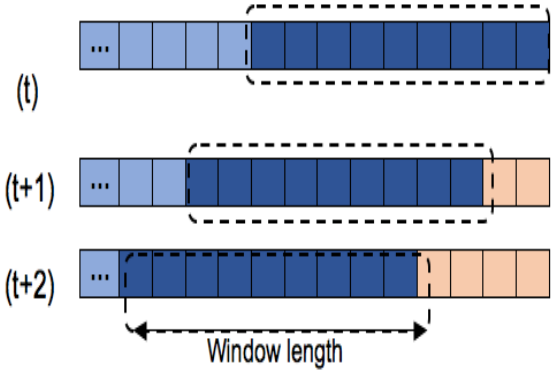
Clustering de flux de données

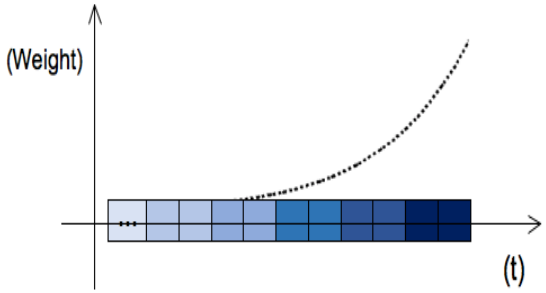
Principe de base

- **Stratégie**
  - **Fenêtre de temps**: diviser de manière indépendante le flux en fenêtres temporelles.
  - Pondération des fenêtres temporelles (données) -> « fading»







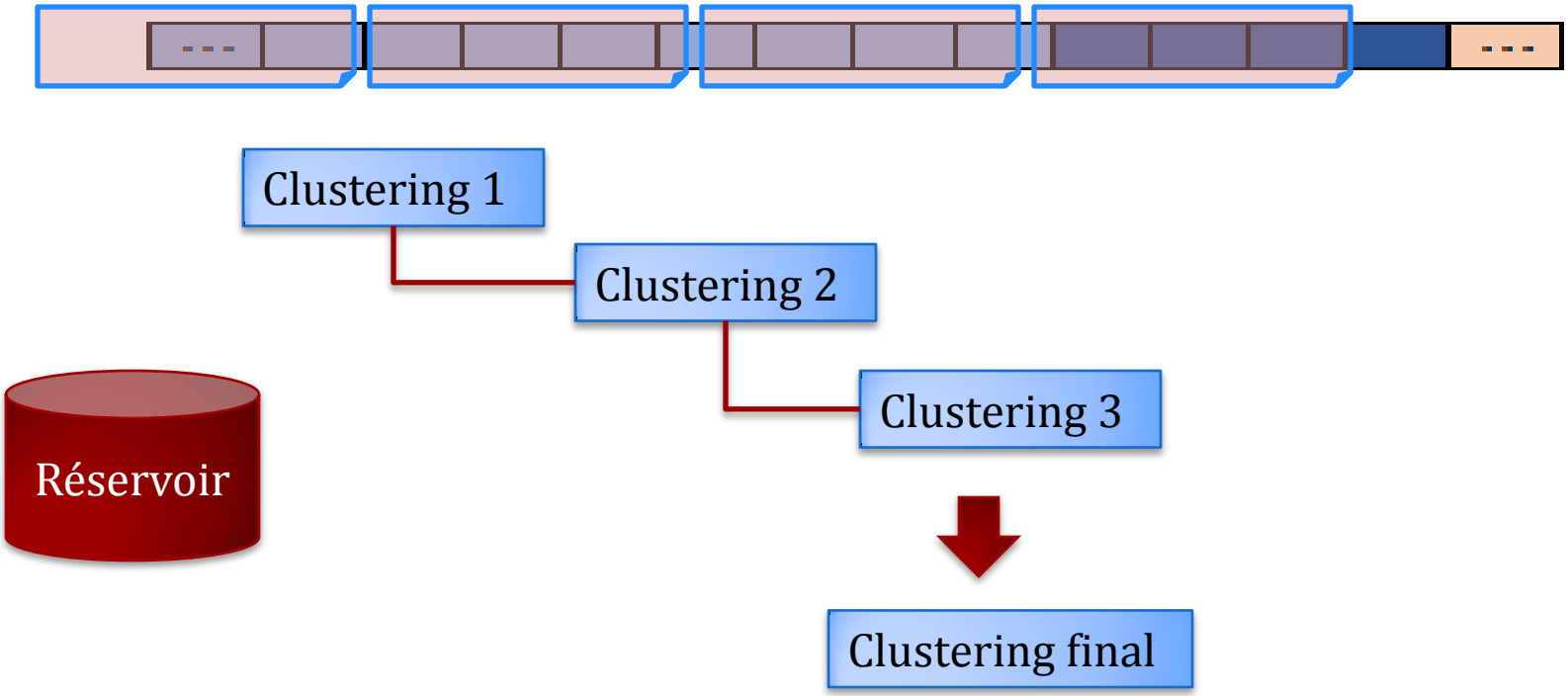


Décroissance exponentielle pour diminuer l'influence des données affectées aux micro-clusters (traiter la dérive de concept)

Clustering de flux de données

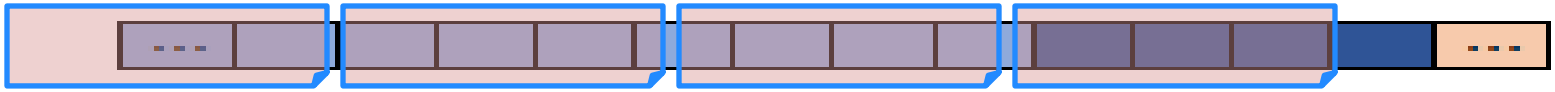
Principe de base

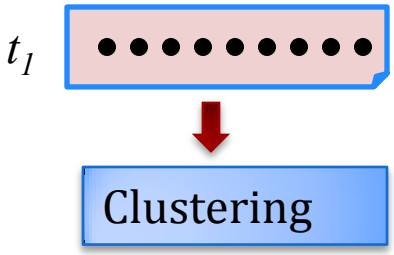
- **Stratégie**
  - **Fenêtre de temps**: diviser le flux en fenêtres temporelles

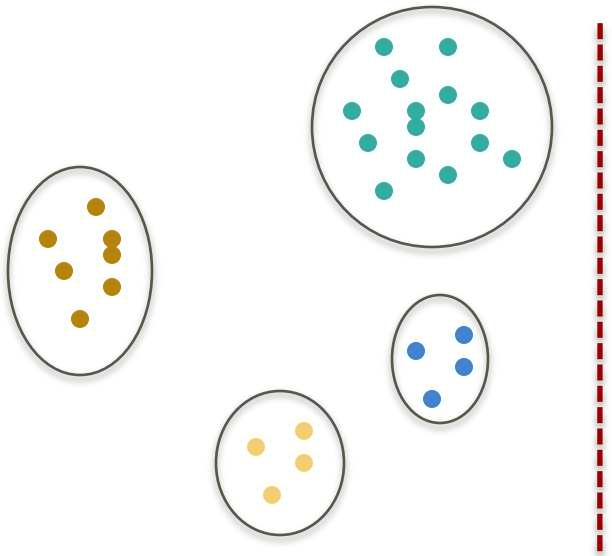


Partitionnement des micro-clusters: utilisation d'algorithme de clustering traditionnel (k-means, dbscan) en mode «off-line» de combiner.

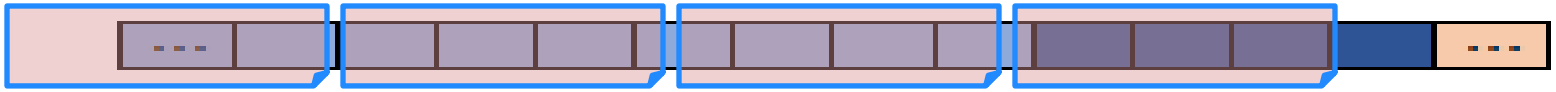
EVOLUTION D'UN CLUSTERING

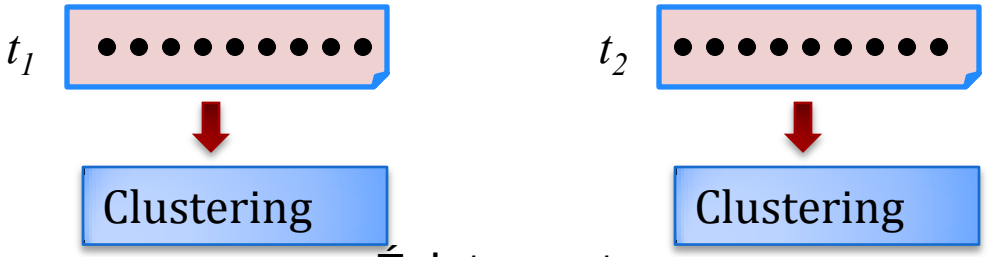


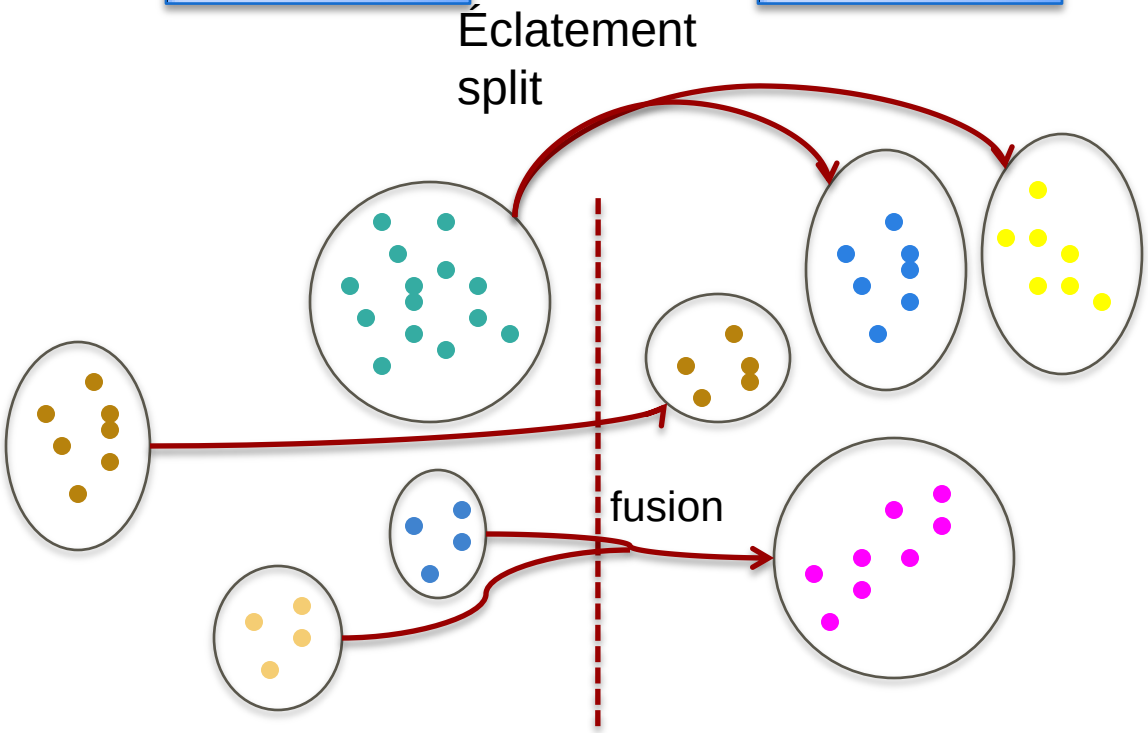




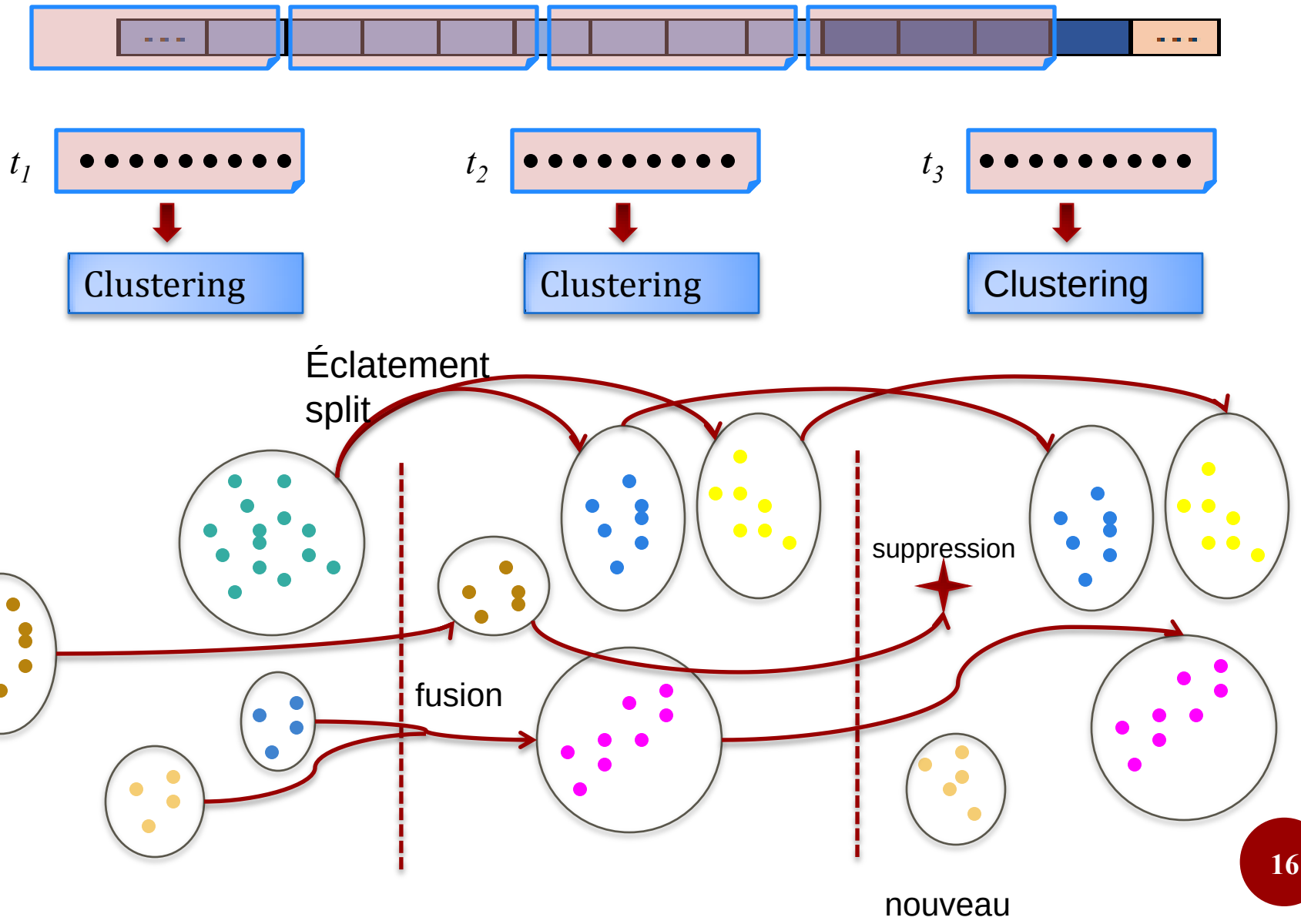
EVOLUTION D'UN CLUSTERING







EVOLUTION D'UN CLUSTERING



Les questions

- Comment traiter les nouvelles données?
  - Affecter à un ancien micro-cluster
  - Création d'un nouveau micro-cluster
- Comment traiter les anciens micro-clusters
  - Supprimer
  - Fusionner des micro-clusters (les plus proches)

ALGORITHME SIMPLE

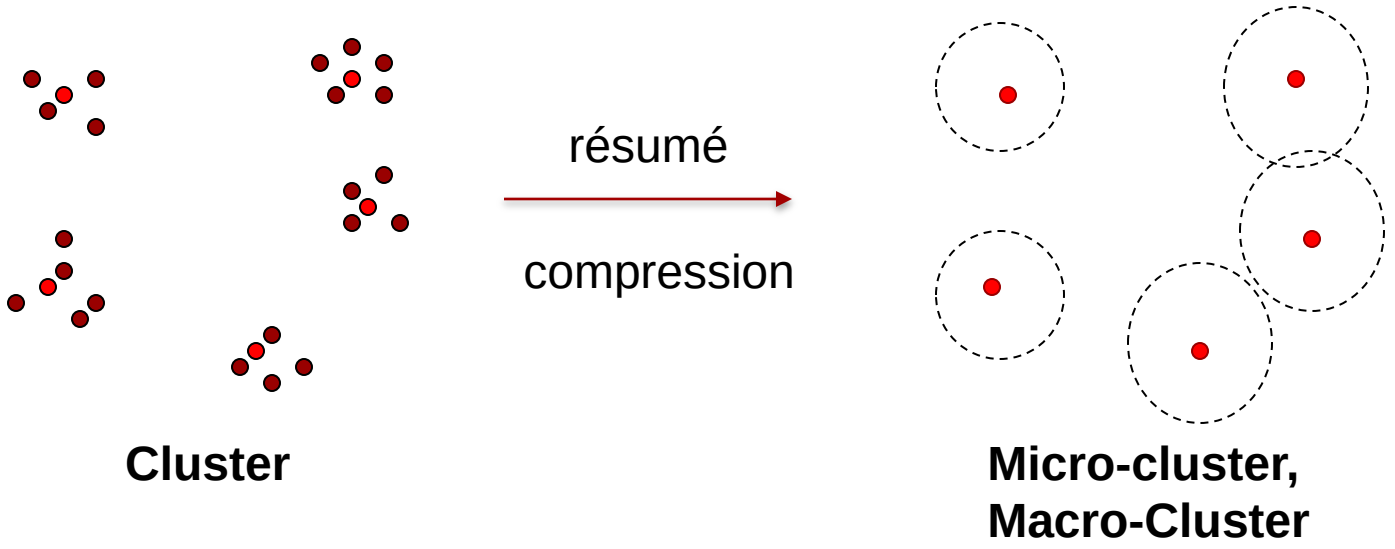
- Initialiser des **micro-cluster** vides
- Pour chaque nouvelle observation
  - Trouver le **micro-cluster** le plus proche
  - Affecter *x* au **micro-cluster** ou créer un autre

- Mise à jour de la structure sous-jacente des **micro-clusters** (une topologie, une hiérarchie ….)

CLUSTERING DE FLUX DE DONNÉES

PRINCIPE DE BASE

- **Stratégie** : **Micro-Clusters**
  - Un micro-cluster est une quantification « **résumé**» d'un ensemble de données qui sont proches les uns aux autres et seront traitées comme une seule unité.

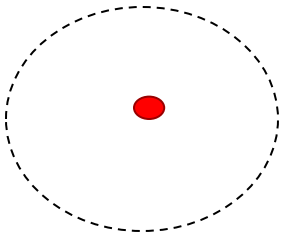


- Utiliser les prototypes : centre de gravité, vecteur caractéristique, variance...etc
- Utiliser le temps d'arrivée du cluster

VECTEUR CARACTÉRISTIQUE : BIRCH

- Utiliser la structure ou un vecteur caractéristique pour sauvegarder les données

- $CF = (N, \sum \mathbf{x}_i, \sum \mathbf{x}_i^2)$



$$CF = (N, LS, SS)$$

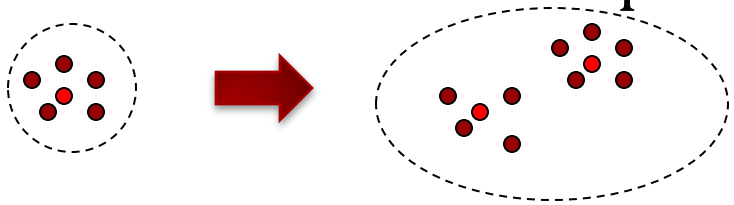
  - Permet de déduire : le centre de gravité, rayon, diamètre
  - Toutes les informations nécessaires pour le clustering

Birch: Balanced Iterative Reducing and Clustering using Hierarchies, by Zhang, Ramakrishnan, Livny 1996

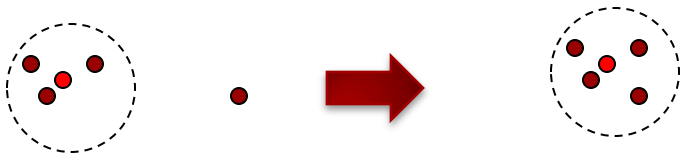
VECTEUR CARACTÉRISTIQUE : BIRCH

- Permet de déduire : le centre de gravité, rayon, diamètre

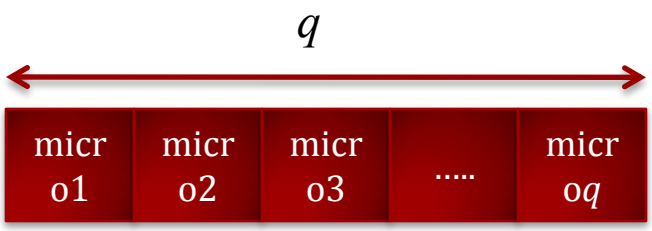
- **Incrémentalité** : à la présentation d'un nouveau point
  - *LS=LS+X*
  - *SS=SS+ $(X)^2$*
  - *N=N+1*



- **Additivité** : fusion de micro-cluster *(CF1, CF2)*
  - *N3=N1+N2*
  - *LS3=LS1+LS2*
  - *SS3=SS1+SS3*



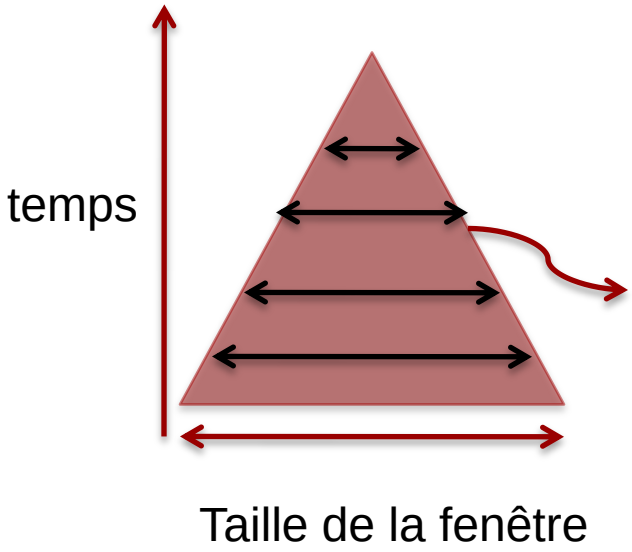
MICROCLUSTERS



$$CF = (N, \sum \mathbf{x}_i, \sum \mathbf{x}_i^2, LST, SST)$$

CluStream

- Sauvegarder la cardinalité, la somme linéaire et la somme au carré de des données du micro-cluster
- Fusion/suppression des clusters proches
- Pour chaque période on génère les macro-clusters
- le snapshot contient *q* micro-clusters, *q* dépend de la disponibilité de la mémoire

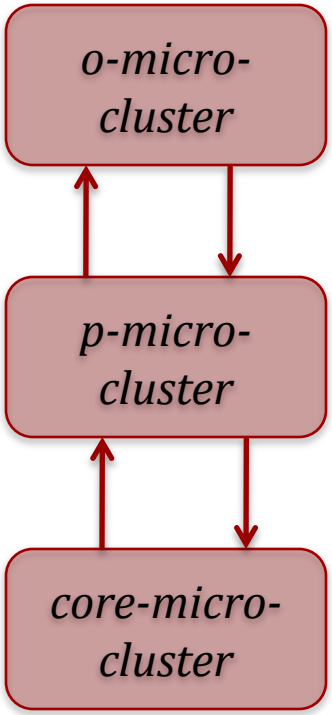


Génération des Macro-clusters

CluStream: A Framework for Clustering Evolving Data Streams, Aggarwal, J. Han, J. Wang, P. Yu (VLDB03)

MICROCLUSTERS

$$CF = (N, \sum \mathbf{x}_i, \sum \mathbf{x}_i^2, LST, SST)$$



DenStream

- Introduction des clusters “outliers”

Les Micro-Clusters sont classés par rapport à leur poids *w*:

Si $w \geq \mu$, *MC* est un *core-micro-cluster*

Durant la partie online, on distingue:

1. **“Potential core-micro-clusters”** (p-micro-clusters), avec $w \geq \beta.\mu$
2. **“Outlier micro-clusters”** (o-micro-clusters), avec $w < \beta.\mu$

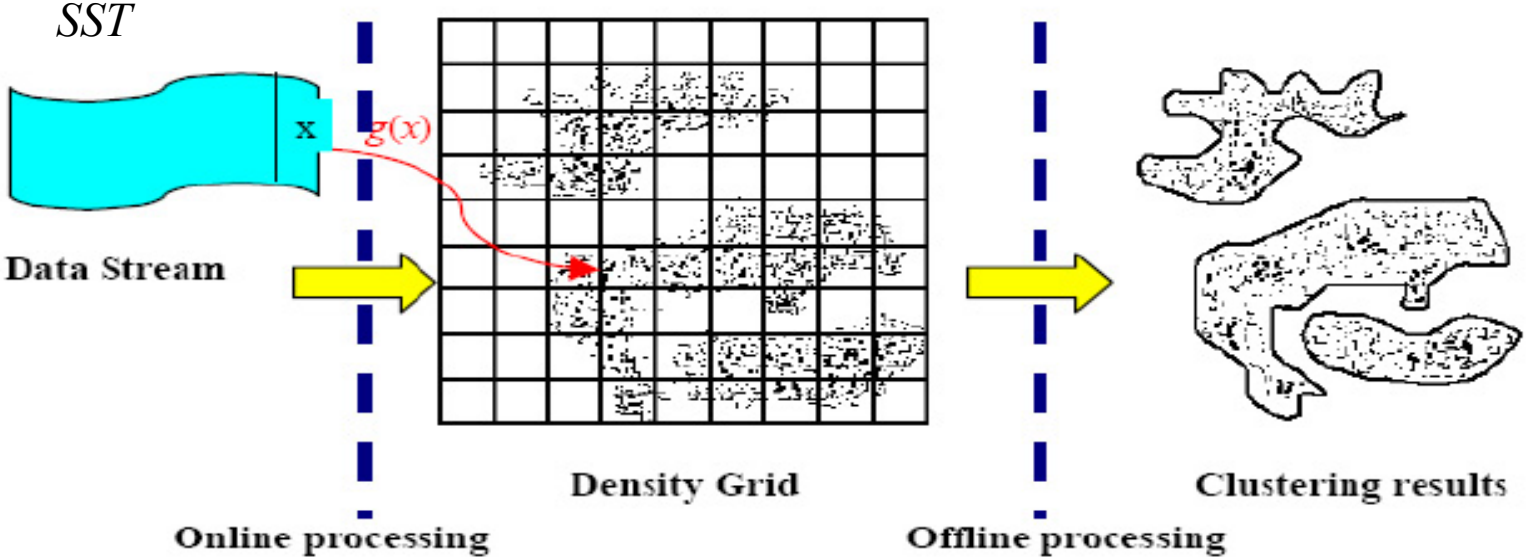
Fenêtre temporelle: poids des données diminue avec le temps: $f(t) = 2^{-\alpha t}$, $\alpha > 0$

MICROCLUSTERS

D-Stream

- Affecte la donnée à la grille
- Grilles pondérés par les récents points

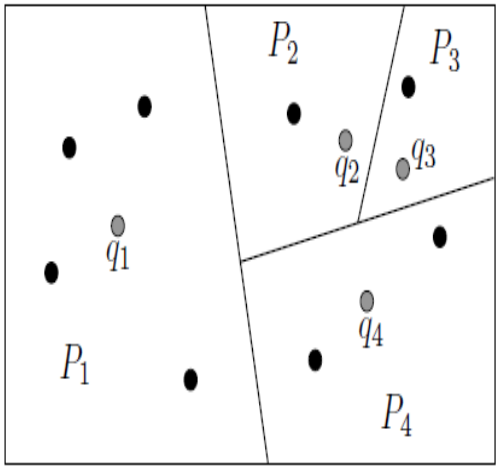
N
LS
SS
LST
SST

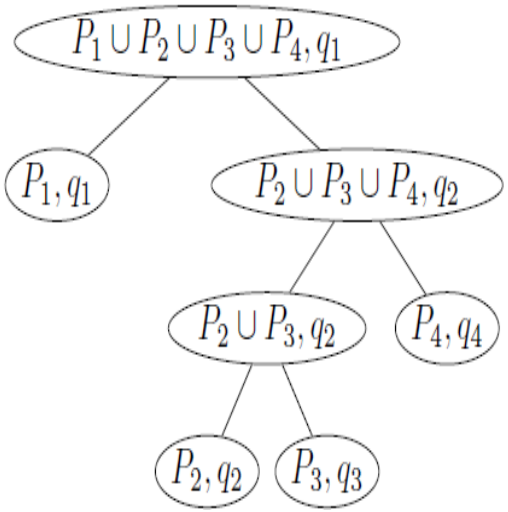


StreamKM++ (Coresets)

- Un ensemble *S* pondéré est un (k,)-corset pour un ensemble de données *X* si le partitionnement de *S* se rapproche du partitionnement de *X* avec une marge d'erreur de

$$(1-\epsilon)dist(X,C) \leq dist_w(S,C) \leq (1+\epsilon)dist(X,C)$$

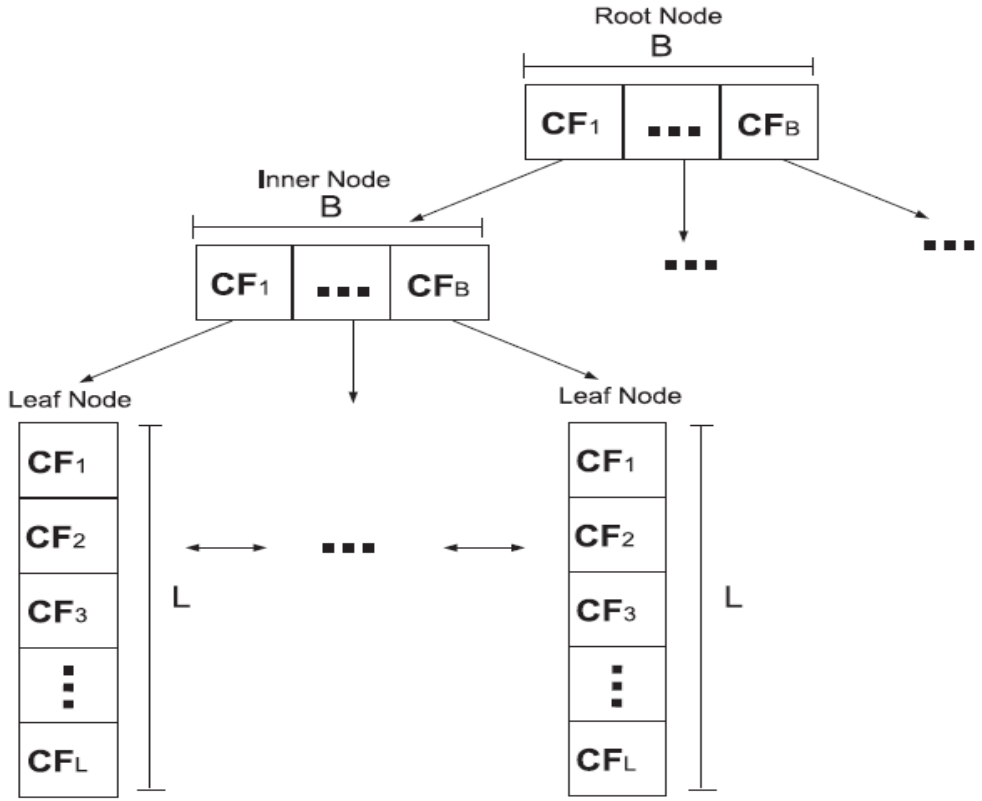




StreamKM++: A Clustering Algorithm for Data Streams. Ackermann et al 2010
LBG: Linde, Y.;, A.; Gray, R. (1980). "An Algorithm for Vector Quantizer Design". IEEE Tran .Buzo et al 1980.

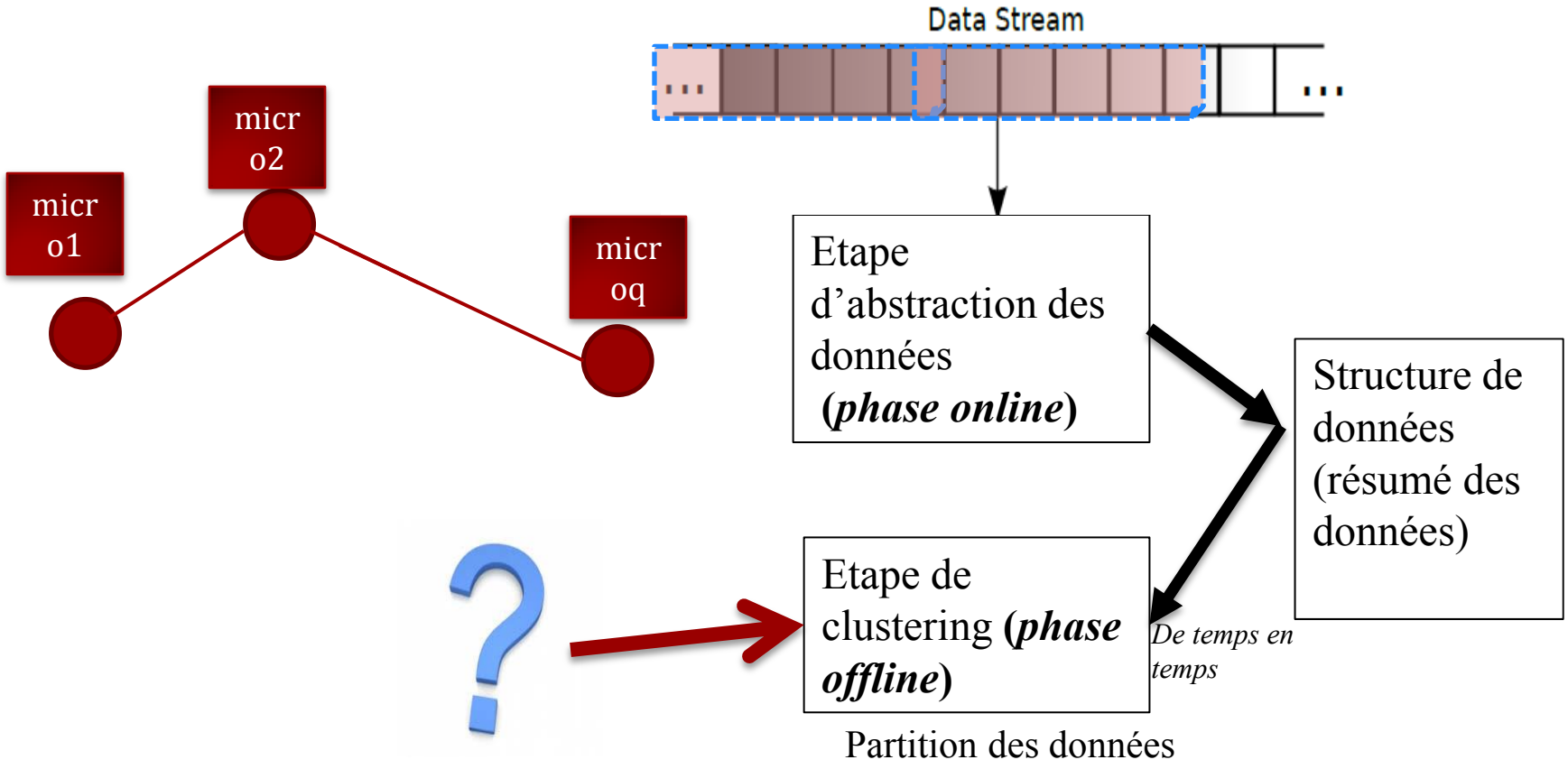
Construction d'un arbre de CF

- B-arbre, seuil de diamètre T et L le nombre de micro-clusters maximum par feuille



- Les "petits" micro-clusters sont considérés comme des "outliers"

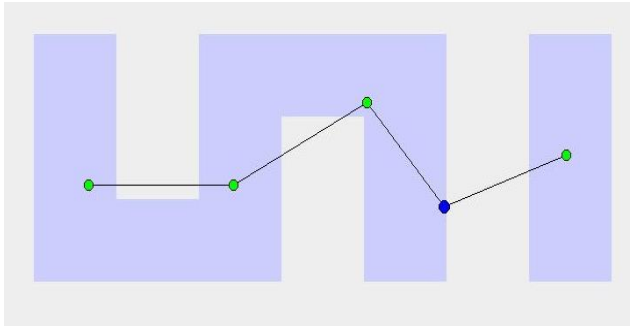
Le Big Data comme le Data Stream

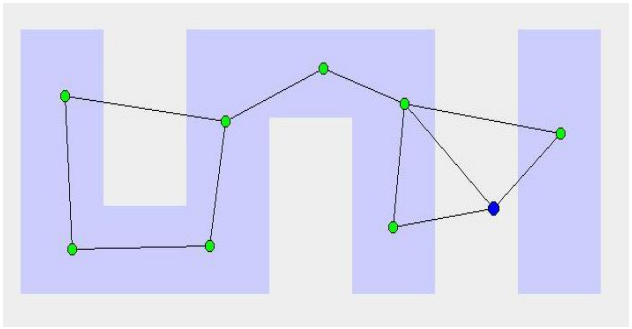


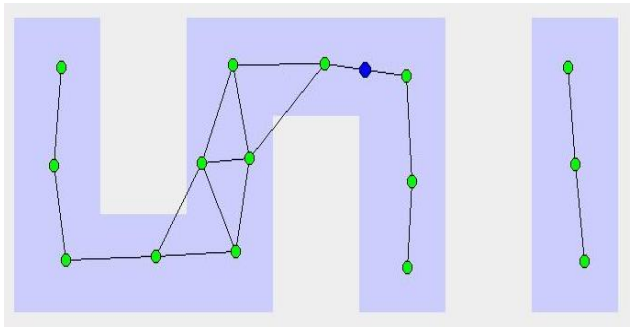
Framework de clustering de flux de données online-offline

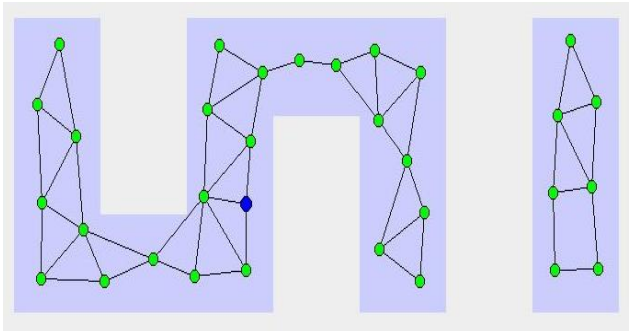
GNG

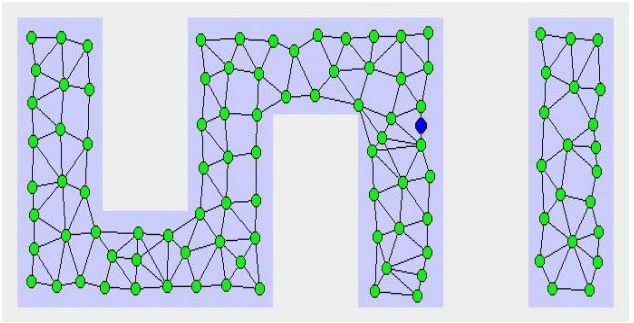
- Topologie évolutive
- Nombre de cellules non fixé











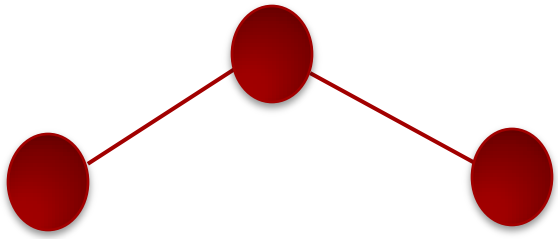
GNG: A Growing Neural Gas Network Learns Topologies. Fritzke 1995

G-STREAM ET LE RESTE

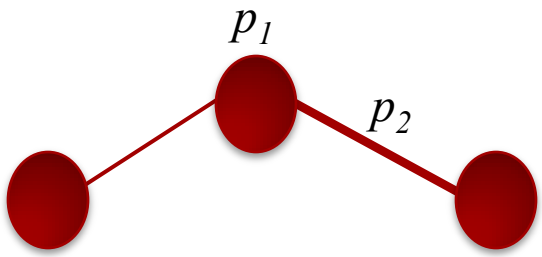
Algorithms	based on	topology	WL	phases	remove	merge	split	fade
G-Stream	NG	✓	✓	online	✓	✗	✗	✓
AING	NG	✓	✗	online	✗	✓	✗	✗
CluStream	*k*-means	✗	✗	2 phases	✓	offline	✗	✗
DenStream	DbScan	✗	✗	2 phases	✓	offline	✗	✓
SOStream	DbScan, SOM	✗	✗	online	✓	✓	✗	✓
E-Stream	*k*-means	✗	✗	2 phases	✓	✓	✓	✓
StreamKM++	*k*-means	✗	✗	2 phases	✓	✓	✓	✓
StrAP	AP	✗	✗	2 phases	✓	✗	✗	✓
SVStream	SVC, SVDD	✗	✗	online	✓	✓	✓	✓

CARACTÉRISTIQUES

- Un graphe représentant la structure topologique,

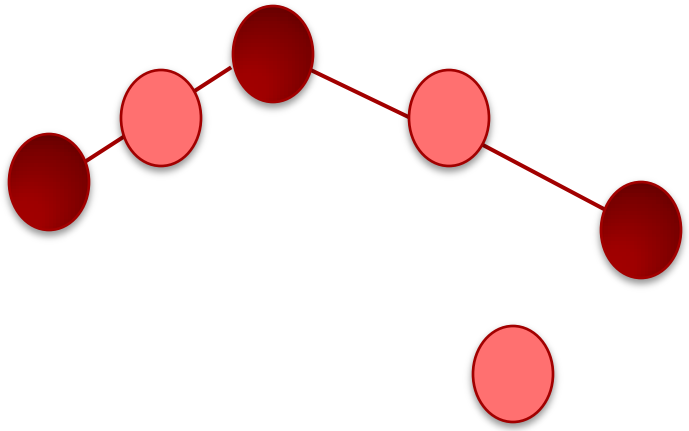


- Pondération des liens et des nœuds



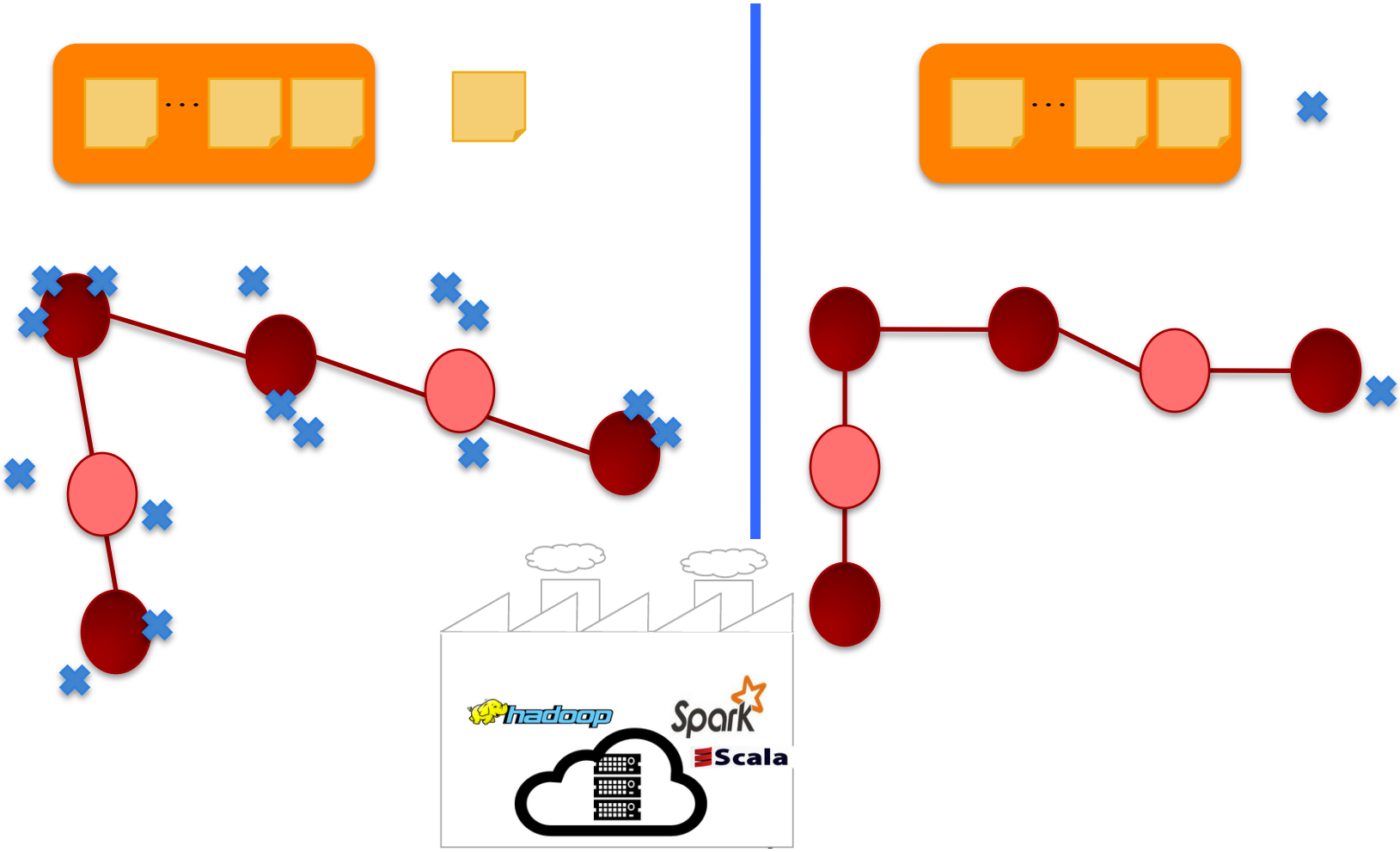
G-STREAM: CARACTÉRISTIQUES ET AVANTAGES

- Pas de phase d'initialisation du modèle,
- Création de plusieurs nœuds (potentiel-micro-cluster) à la fois,



- Utilisation d'un réservoir

G-STREAM

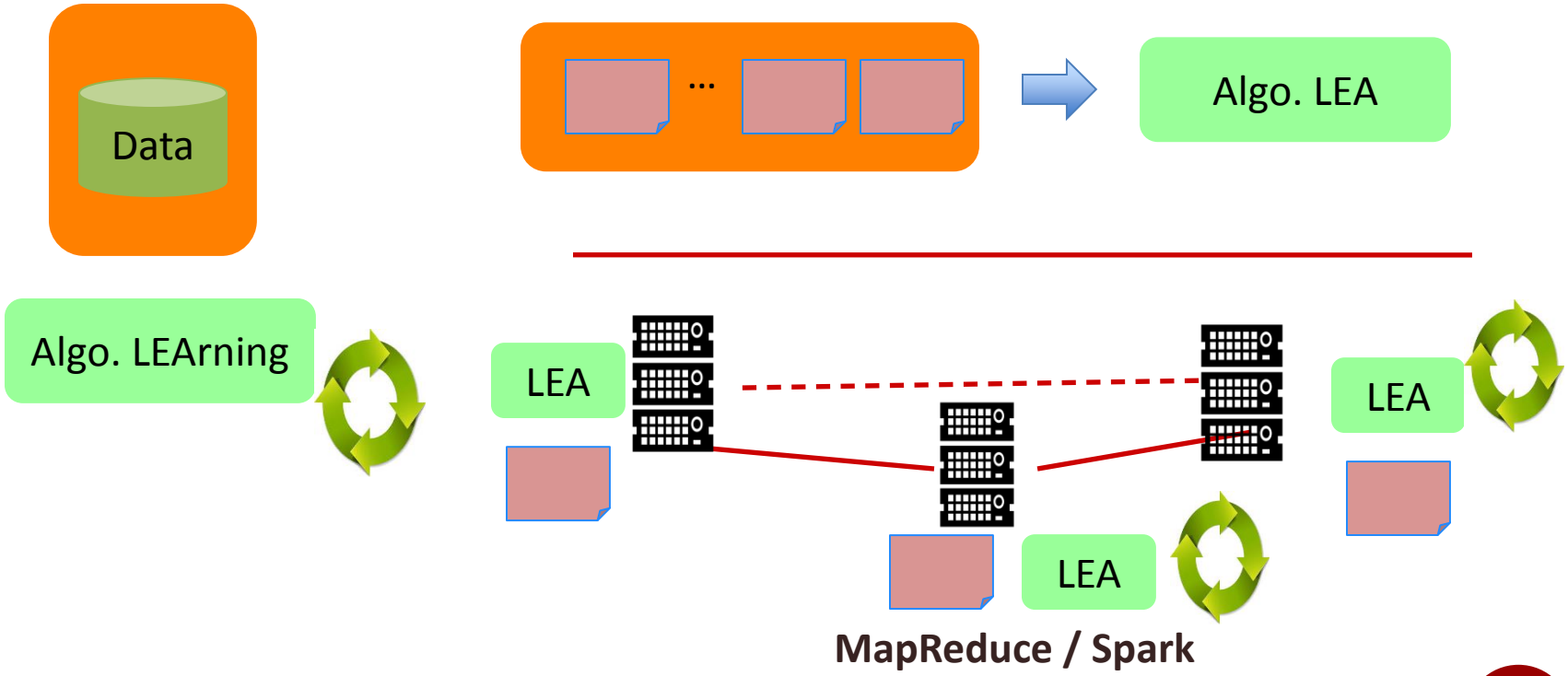


Mise à jour d'une manière «Batch»

Mise à jour d'une manière stochastique

Stream + Batch + MapReduce

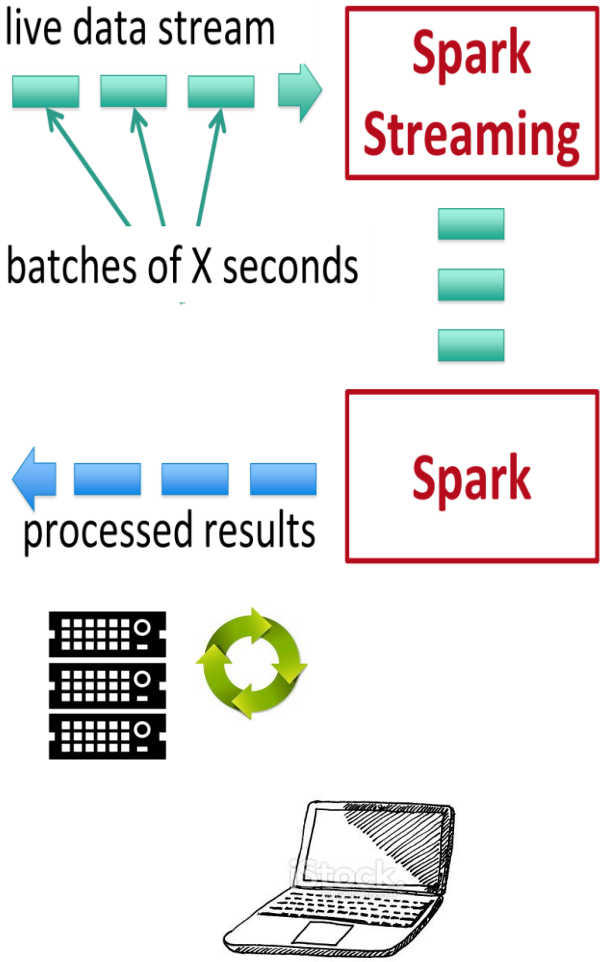
Aucune maitrise de la distribution des données dans l'échosystème.



SPARK STREAMING

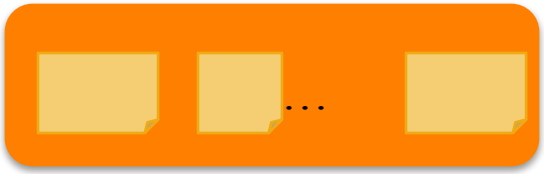
Combinaison du traitement du flux et le traitement en batch

```
for (i <- 1 until 10) {
 val closest = data.map(p =>
   (closestCentroid(p, centroids),
(p, 1))
 )
 val pointStats=closest.reduceByKey{
   case ((p1, sum1), (p2, sum2)) =>
(p1 + p2, sum1 + sum2)
 }
 pointStats.foreach{case(id, value)
=>
 centroids(id) = value._1 / value._2
 }
}
```



G-STREAM : PRINCIPE DE LA VERSION BATCH

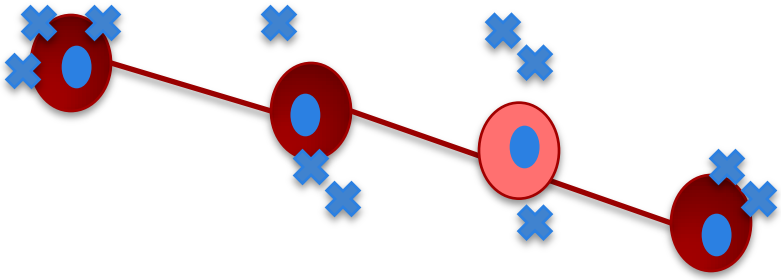
$$\mathcal{DS} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L\} \qquad \mathbf{X}_i = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$$

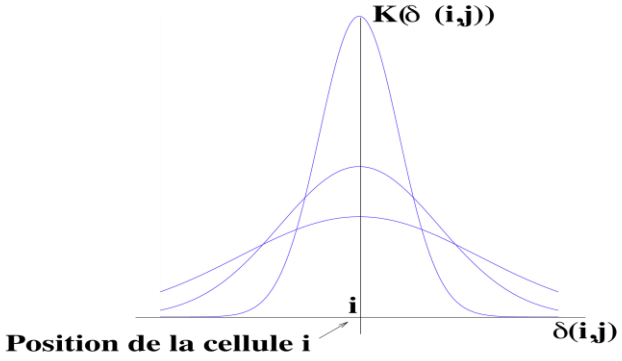


Version batch de SOM

$\chi(\mathbf{x}_i)$ Affecter au plus proche

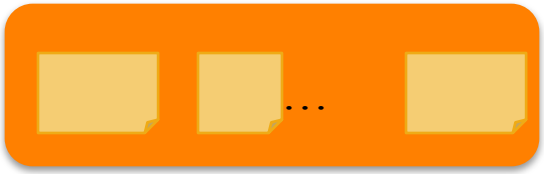
K Fonction de voisinage





G-STREAM : PRINCIPE DE LA VERSION BATCH

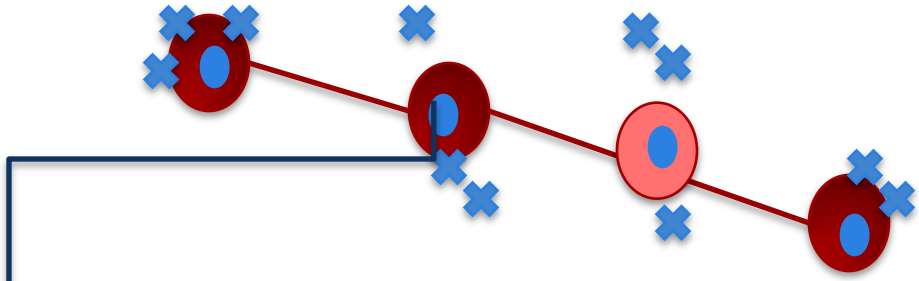
$$\mathcal{DS} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L\} \qquad \mathbf{X}_i = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$$



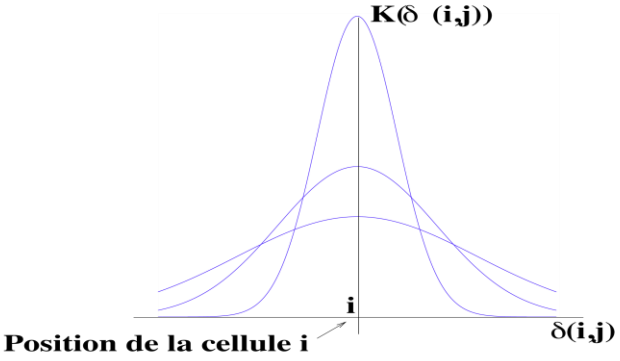
Version batch de SOM

$\chi(\mathbf{x}_i)$ Affecter au plus proche

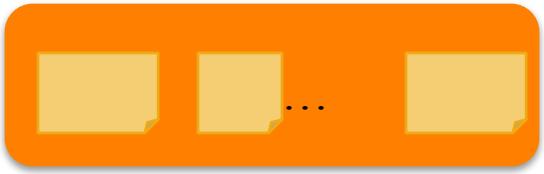
K Fonction de voisinage

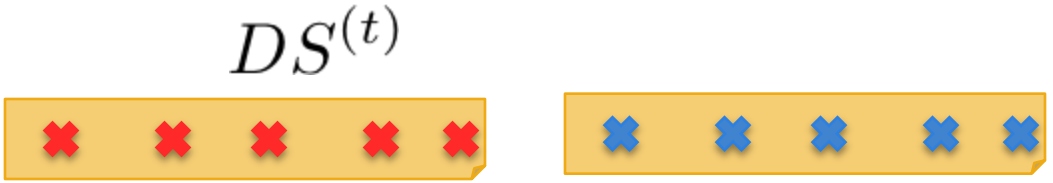


$$\mathbf{w}_c = \frac{\sum_{\mathbf{x}_i} K(c, \chi(\mathbf{x}_i))\mathbf{x}_i}{\sum_{\mathbf{x}_i} K(c, \chi(\mathbf{x}_i))},$$



G-STREAM : LE PROTOTYPE

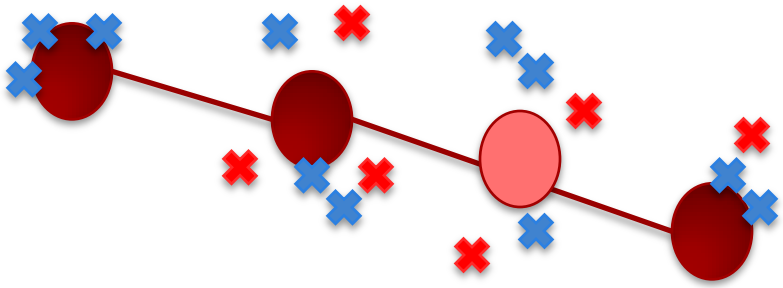




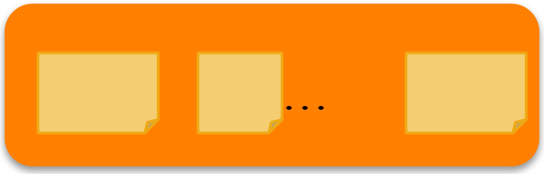
Nouvelles données

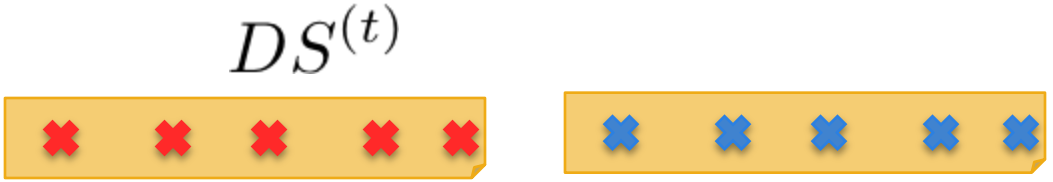
Anciennes données

$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))\mathbf{x}_i}{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))},$$



G-STREAM : LE PROTOTYPE

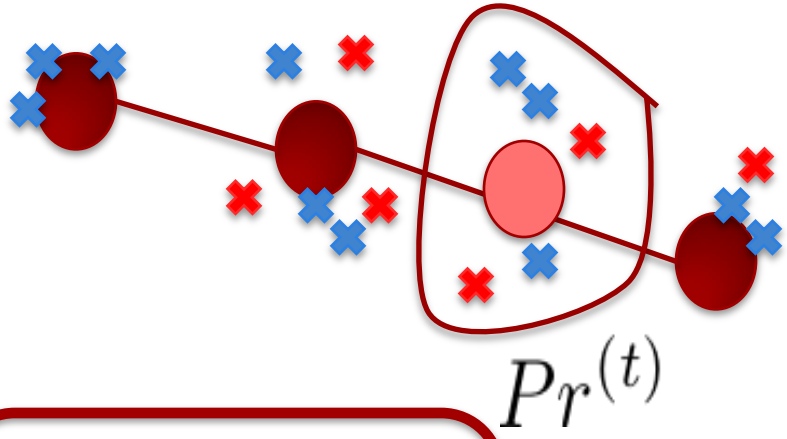




Nouvelles données

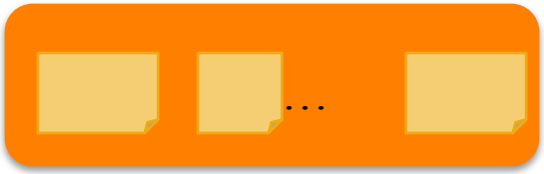
Anciennes données

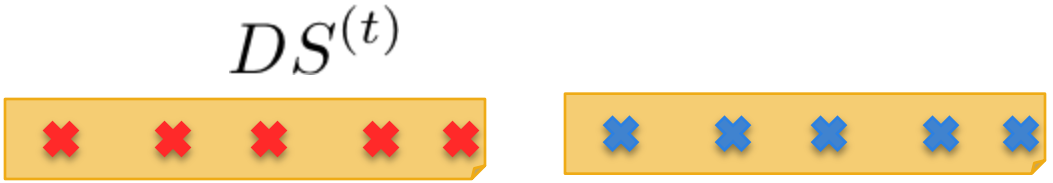
$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))\mathbf{x}_i}{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))},$$



$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{r \in C} \sum_{\mathbf{x}_i \in Pr^{(t-1)}} K(c, r)\mathbf{x}_i + \sum_{r \in C} \sum_{\mathbf{x}_i \in Pr^{(t)}} K(c, r)\mathbf{x}_i}{\sum_{r \in C} \sum_{\mathbf{x}_i \in Pr^{(t-1)}} K(c, r) + \sum_{r \in C} \sum_{\mathbf{x}_i \in Pr^{(t)}} K(c, r)},$$

G-STREAM : LE PROTOTYPE

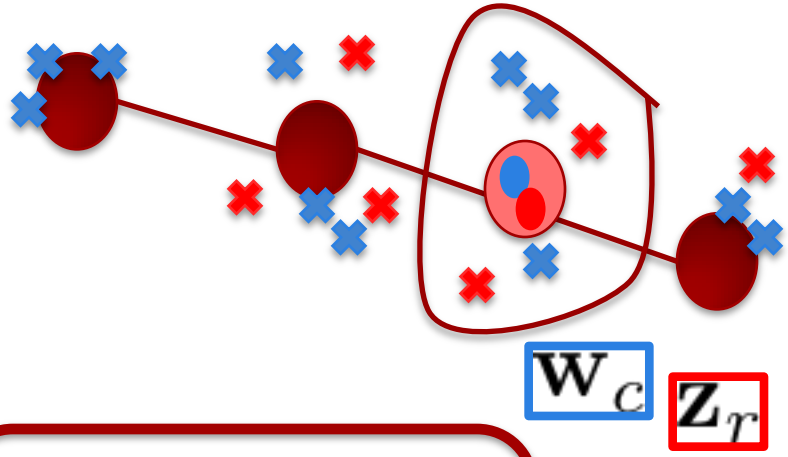




Nouvelles données

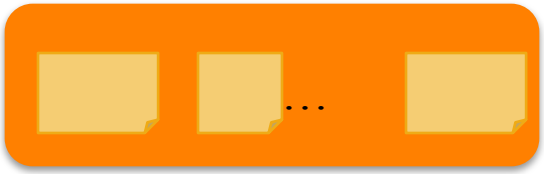
Anciennes données

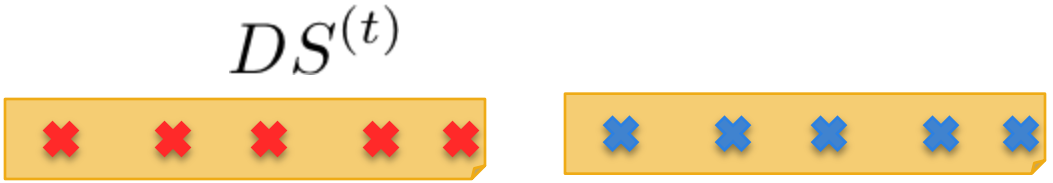
$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))\mathbf{x}_i}{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))},$$



$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{r \in C} K(c,r)\mathbf{w}_c^{(t-1)} n_r^{(t-1)} + \sum_{r \in C} K(c,r)\mathbf{z}_r^{(t)} m_r^{(t)}}{\sum_{r \in C} K(c,r) n_r^{(t-1)} + \sum_{r \in C} K(c,r) m_r^{(t)}}$$

G-STREAM : LE PROTOTYPE

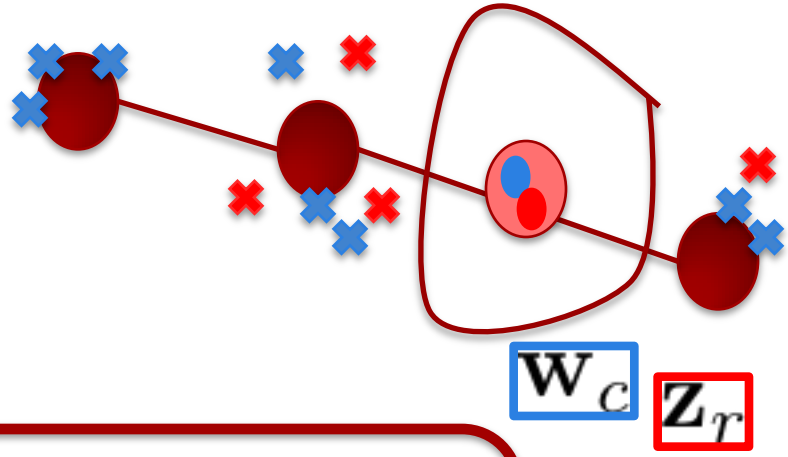




Nouvelles données

Anciennes données

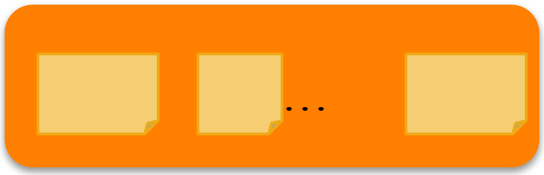
$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))\mathbf{x}_i}{\sum_{\mathbf{x}_i \in DS^{(t)}} K(c, \chi(\mathbf{x}_i))},$$



$$\mathbf{w}_c^{(t+1)} = \frac{\sum_{r \in C} K(c,r)\mathbf{w}_c^{(t-1)} n_r^{(t-1)} + \sum_{r \in C} K(c,r)\mathbf{z}_r^{(t)} m_r^{(t)}}{\sum_{r \in C} K(c,r) n_r^{(t-1)} + \sum_{r \in C} K(c,r) m_r^{(t)}}$$

G-STREAM: AFFECTATION, MAP

$\mathcal{DS} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L\}$



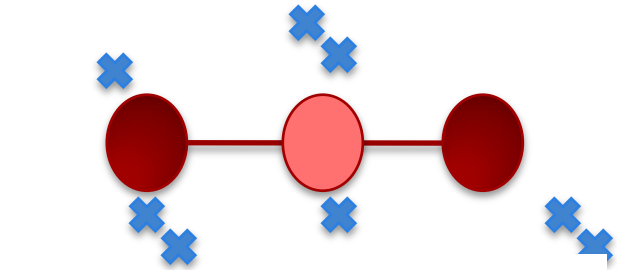
$\mathbf{X}_i = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$

$\mathbf{x}_i$

Map

$(bmu_1, bmu_2, \mathbf{x}_i)$

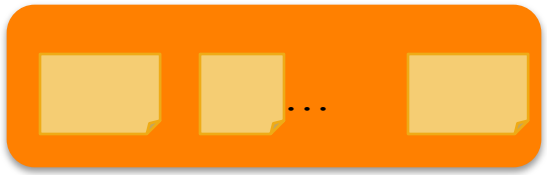
Numéro de cluster

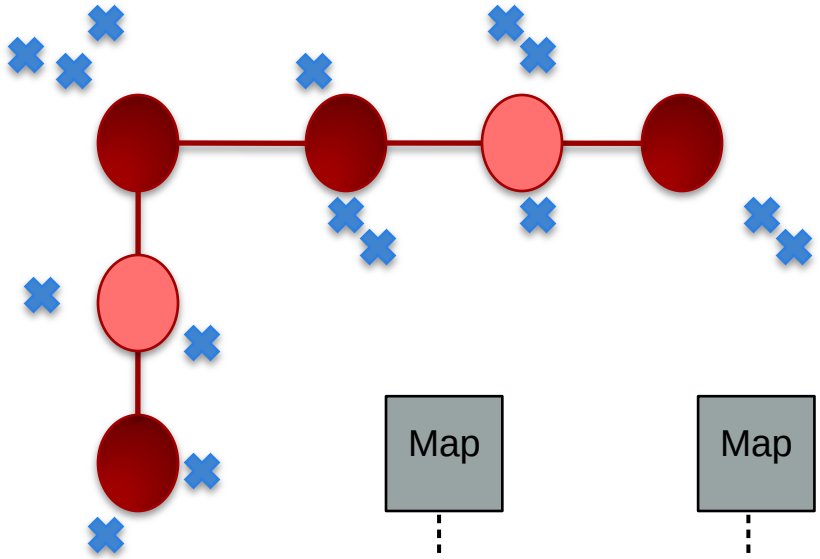


bmu_1 bmu_2

G-STREAM

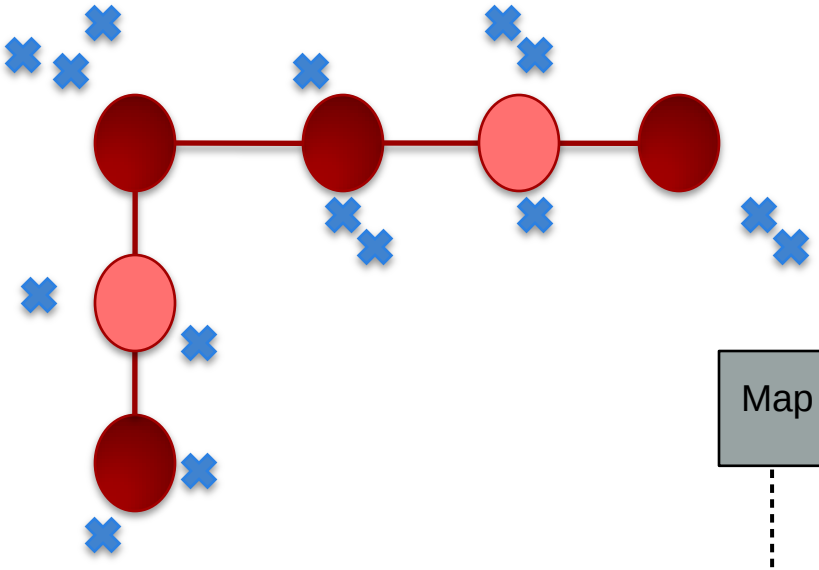
$$\mathcal{DS} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L\} \qquad \mathbf{X}_i = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$$

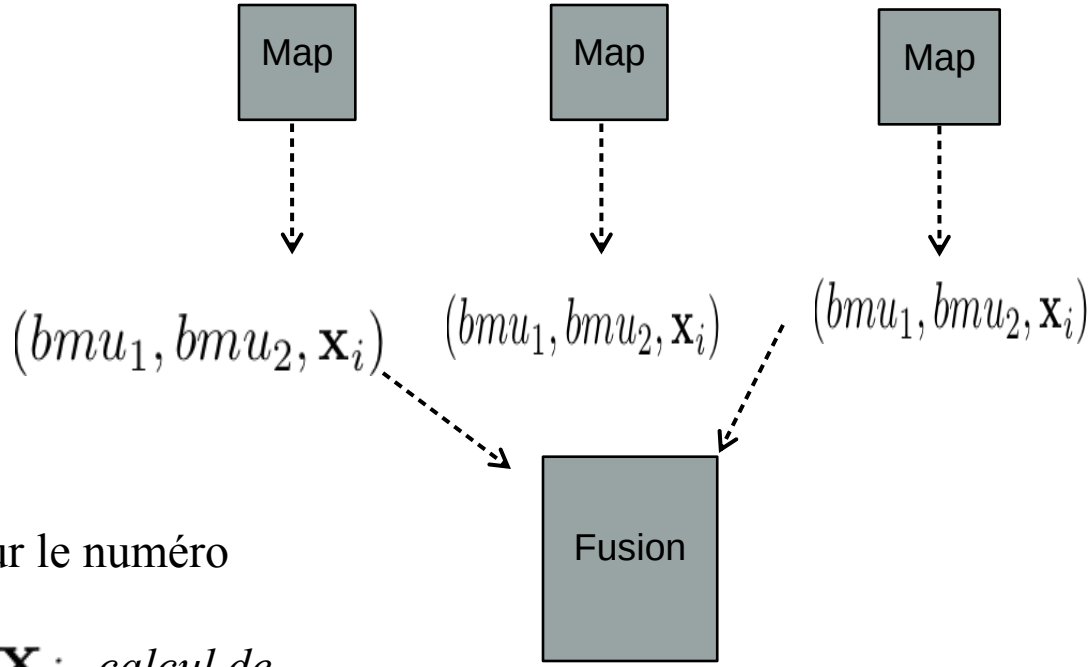




$(bmu_1, bmu_2, \mathbf{x}_i)$ $(bmu_1, bmu_2, \mathbf{x}_i)$ $(bmu_1, bmu_2, \mathbf{x}_i)$

G-STREAM



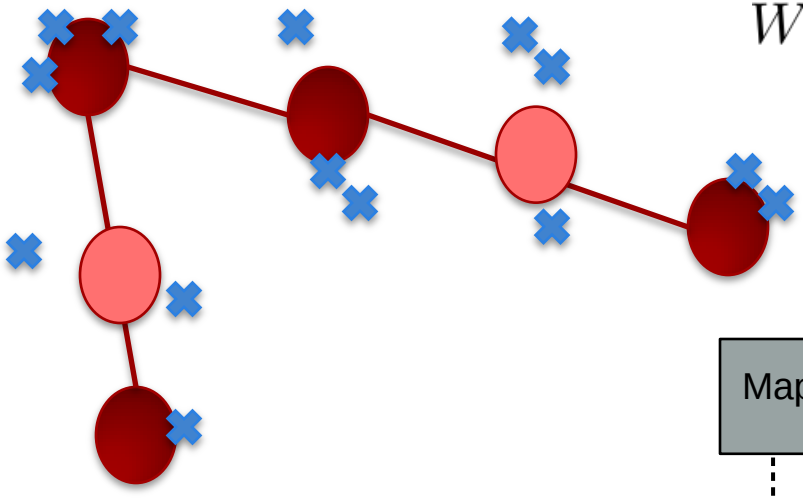


Fusion :

•Contrainte must-link sur le numéro de *bmu (ou autre)*

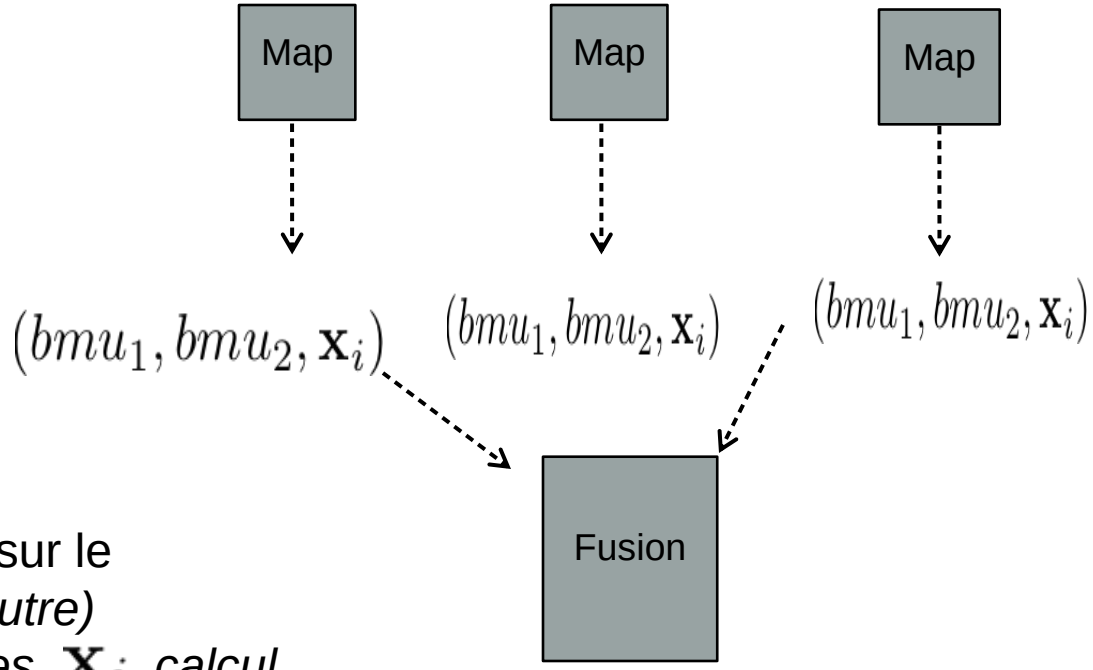
•*action: sommation des* $\mathbf{x}_i$*, calcul de la fonction de voisinage,...etc*

G-STREAM



$$Weight(c) = \sum 2^{-\lambda(t-t_0)}$$

$$f(t) = 2^{-\lambda(t-t_0)}$$

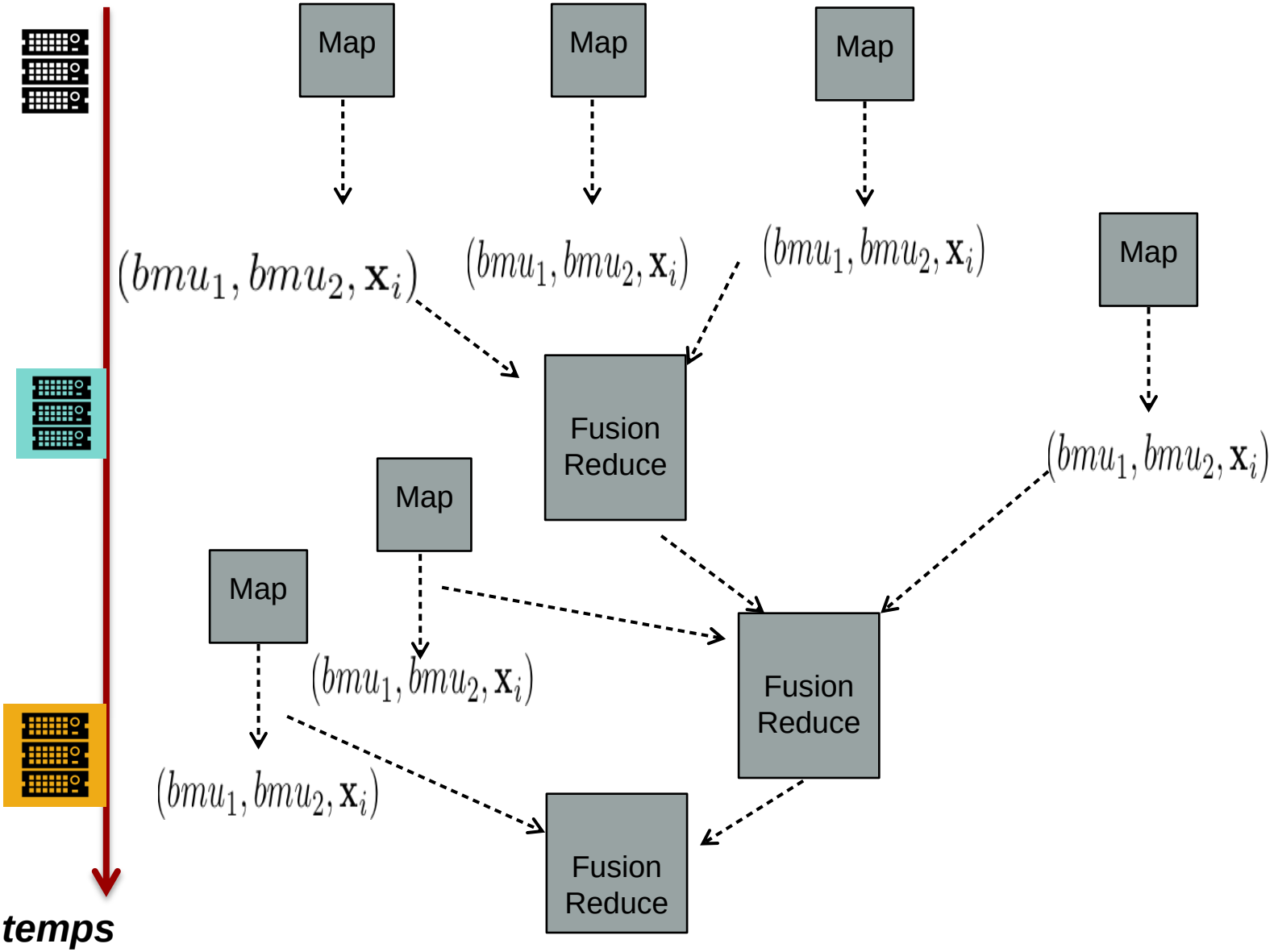


Fusion :

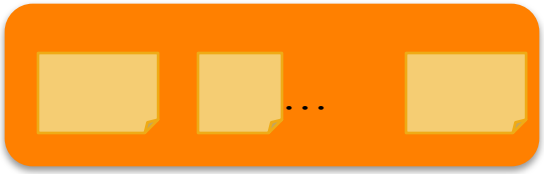
•Contrainte must-link sur le numéro de *bmu (ou autre)*

•*action: sommation des* $\mathbf{x}_i$*, calcul de la fonction de voisinage,…etc*

La chaine Map (affectation) et Réduce (fusion)



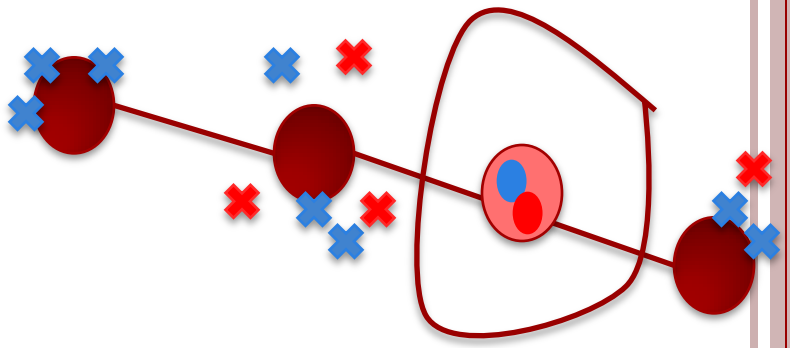
G-STREAM : LE PROTOTYPE



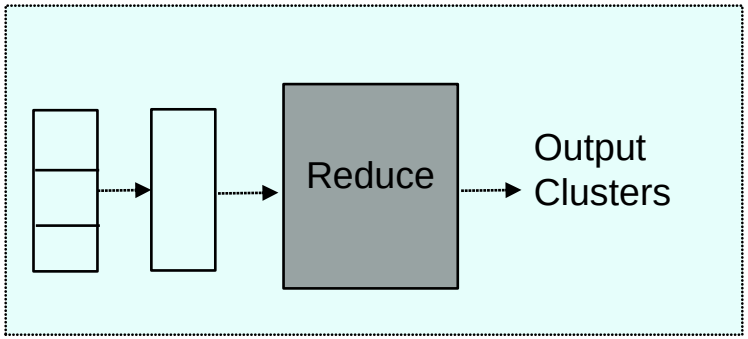
Nouvelles données

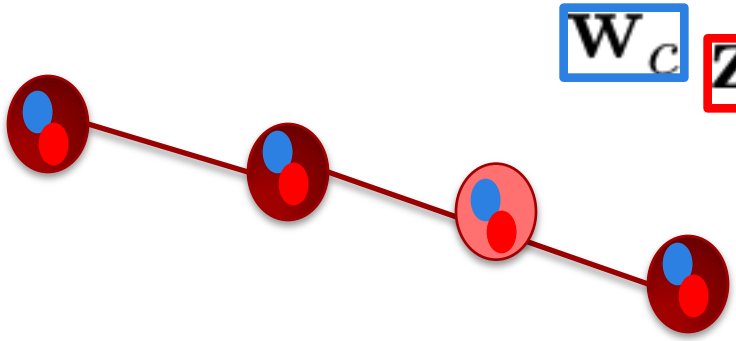
Anciennes données

$$\mathbf{w}_c^{(t+1)} = \frac{\mathbf{w}_c^{(t)} n_c^{(t)} \alpha + \sum_{r \in \mathcal{C}} \mathcal{K}(r,c) \mathbf{z}_r^{(t)} m_r^{(t)}}{n_c^{(t)} \alpha + \sum_{r \in \mathcal{C}} \mathcal{K}(r,c) m_r^{(t)}}$$



(key, [Value0, Value1, Value2,…])

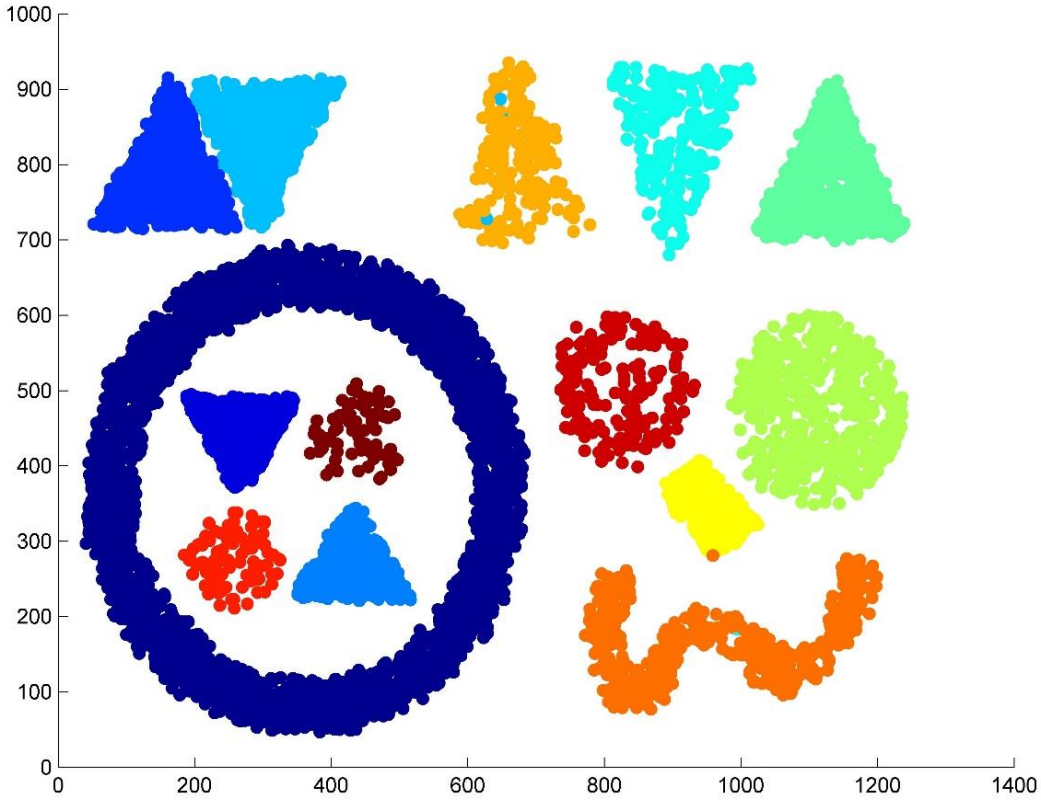


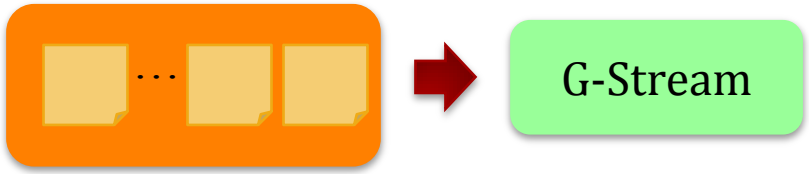


BASE DE DONNÉES

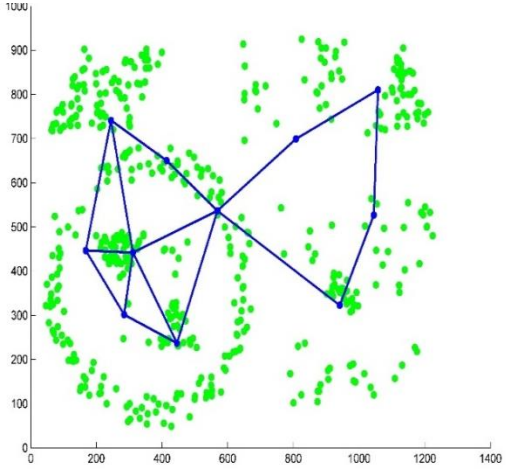
Datasets	#records	#features	#classes
DS1	9,153	2	14
DS2	5,458	2	13
letter4	9,344	2	7
Sea	60,000	3	2
HyperPlan	100,000	10	5
KddCup99	494,021	41	23
CoverType	581,012	54	7
Sensor	2,219,803	5	54

G-STREAM : EXEMPLE

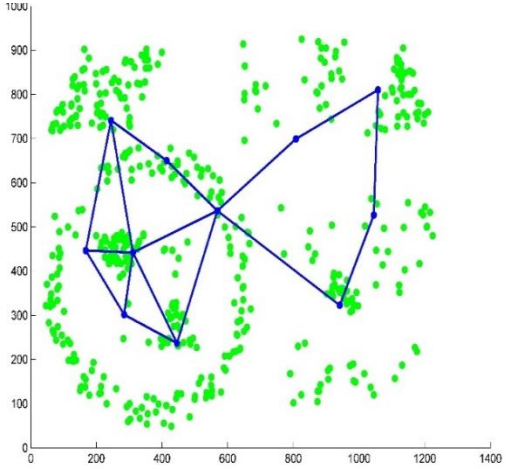


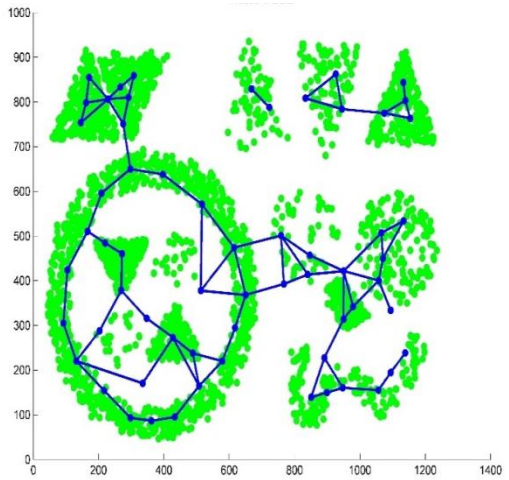


G-STREAM SUR DS1

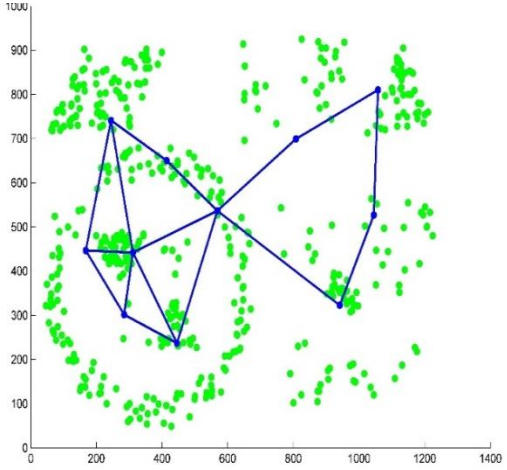


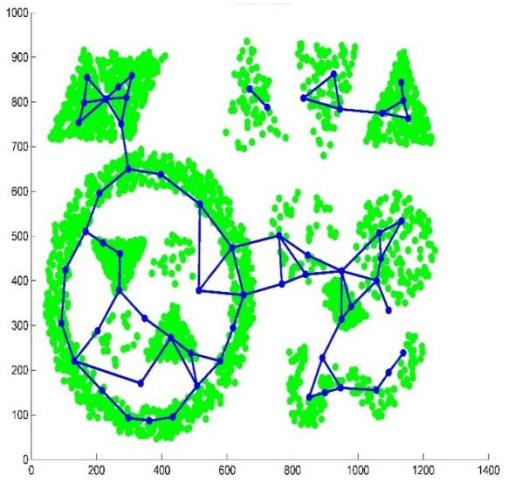
G-STREAM SUR DS1

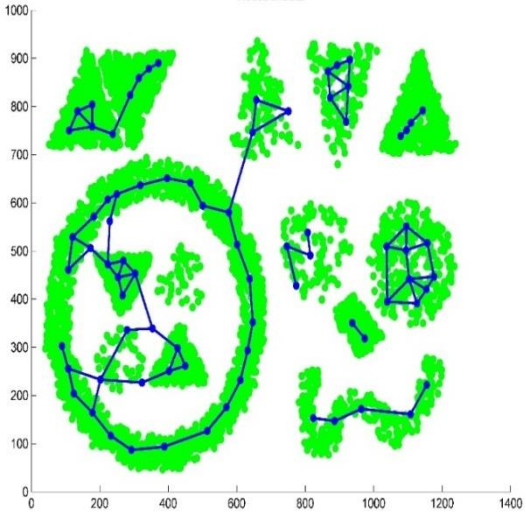




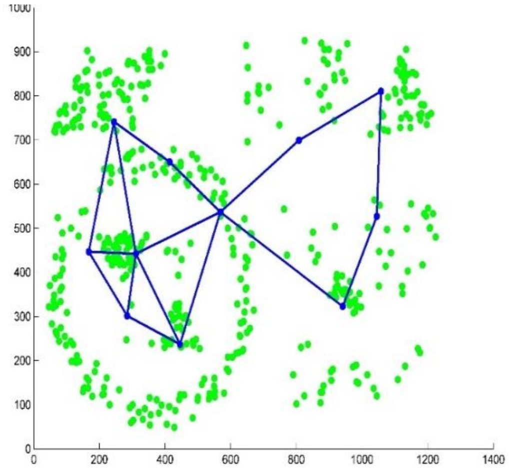
G-STREAM SUR DS1

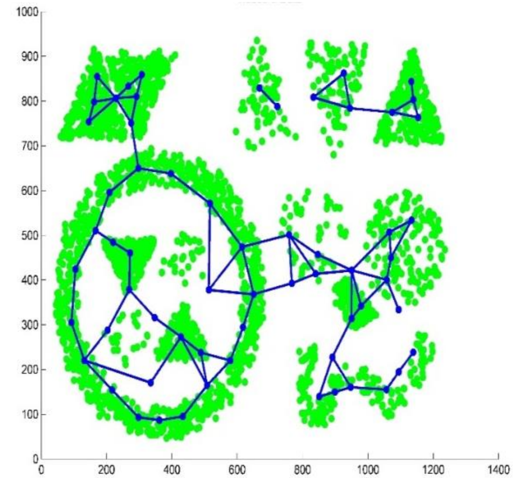


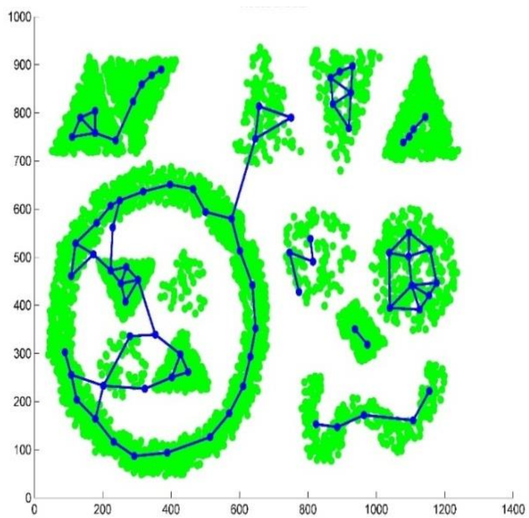


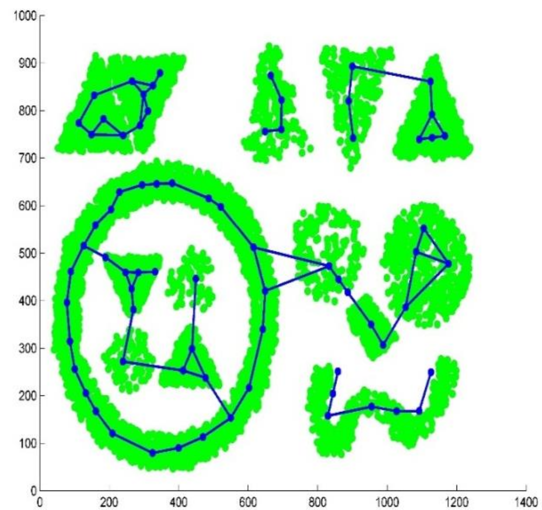


G-Stream sur DS1

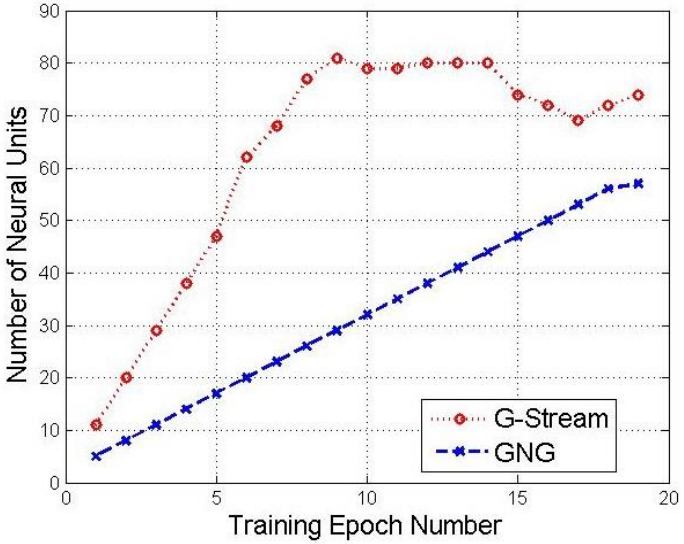




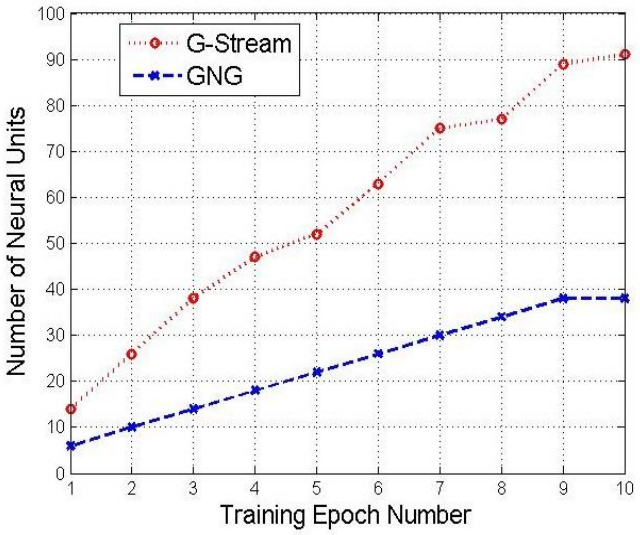




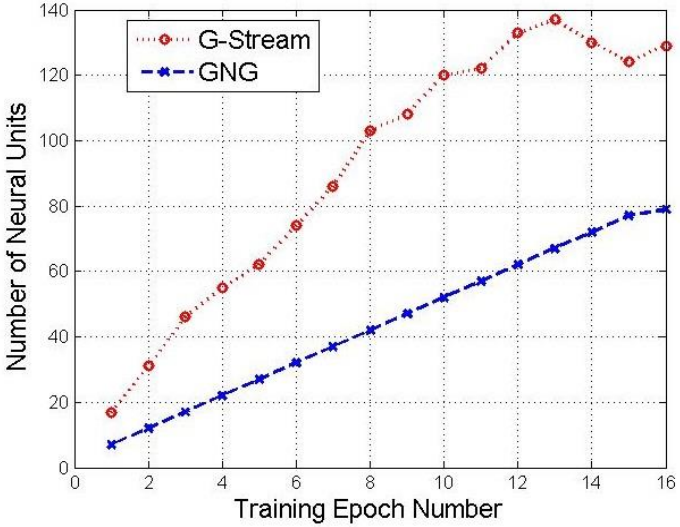
G-STREAM VS GNG-ONLINE: #NŒUDS



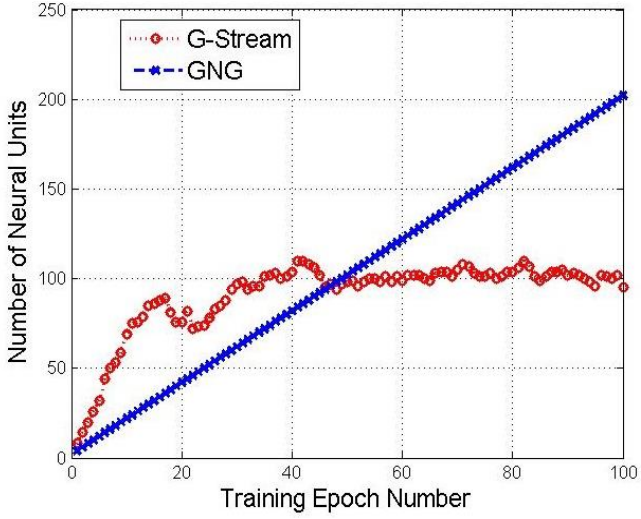
DS1



DS2

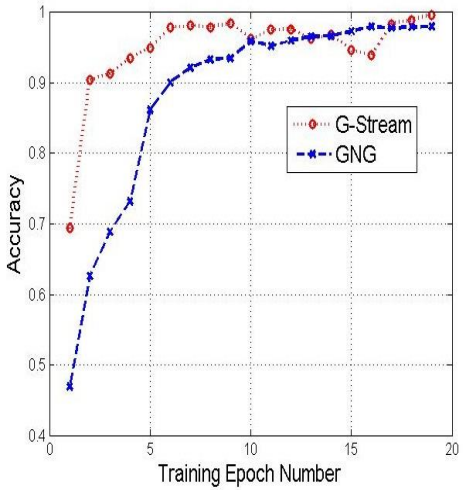


letter4

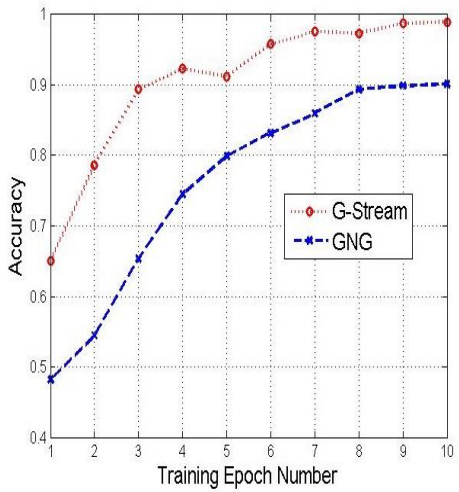


Sea

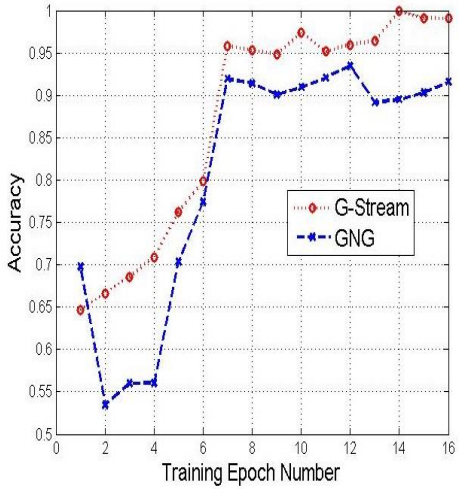
G-STREAM VS GNG ONLINE: ACCURACY



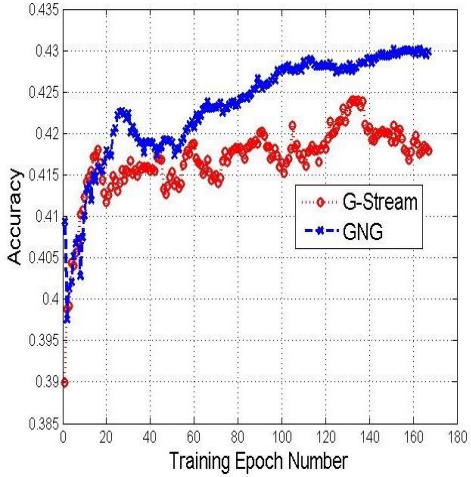
DS1



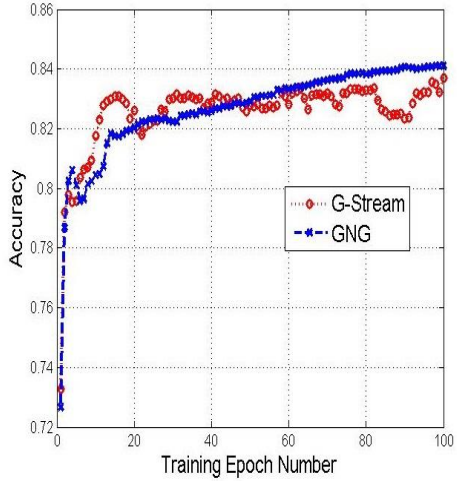
DS2



Letter4



HyperPlan

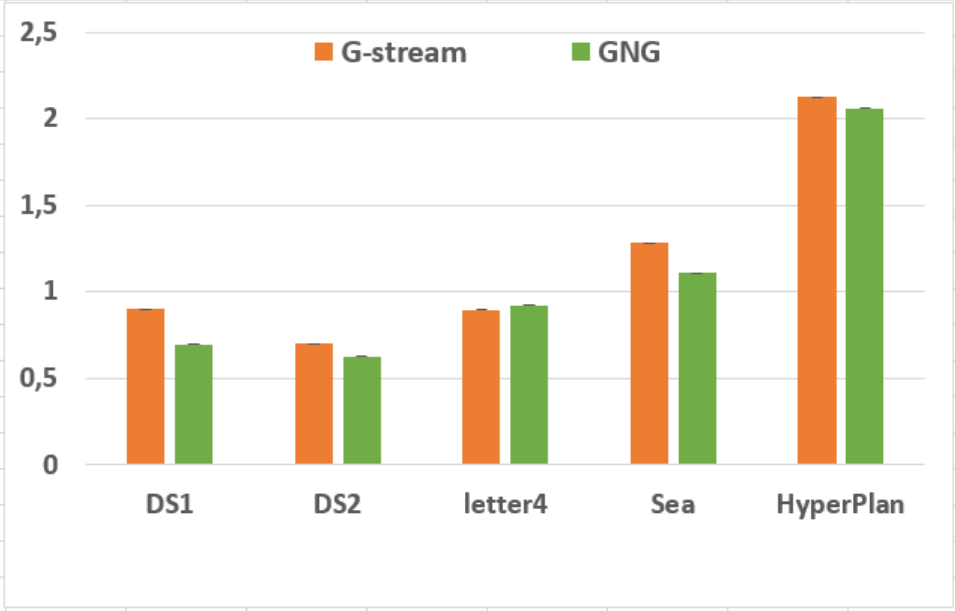


Sea

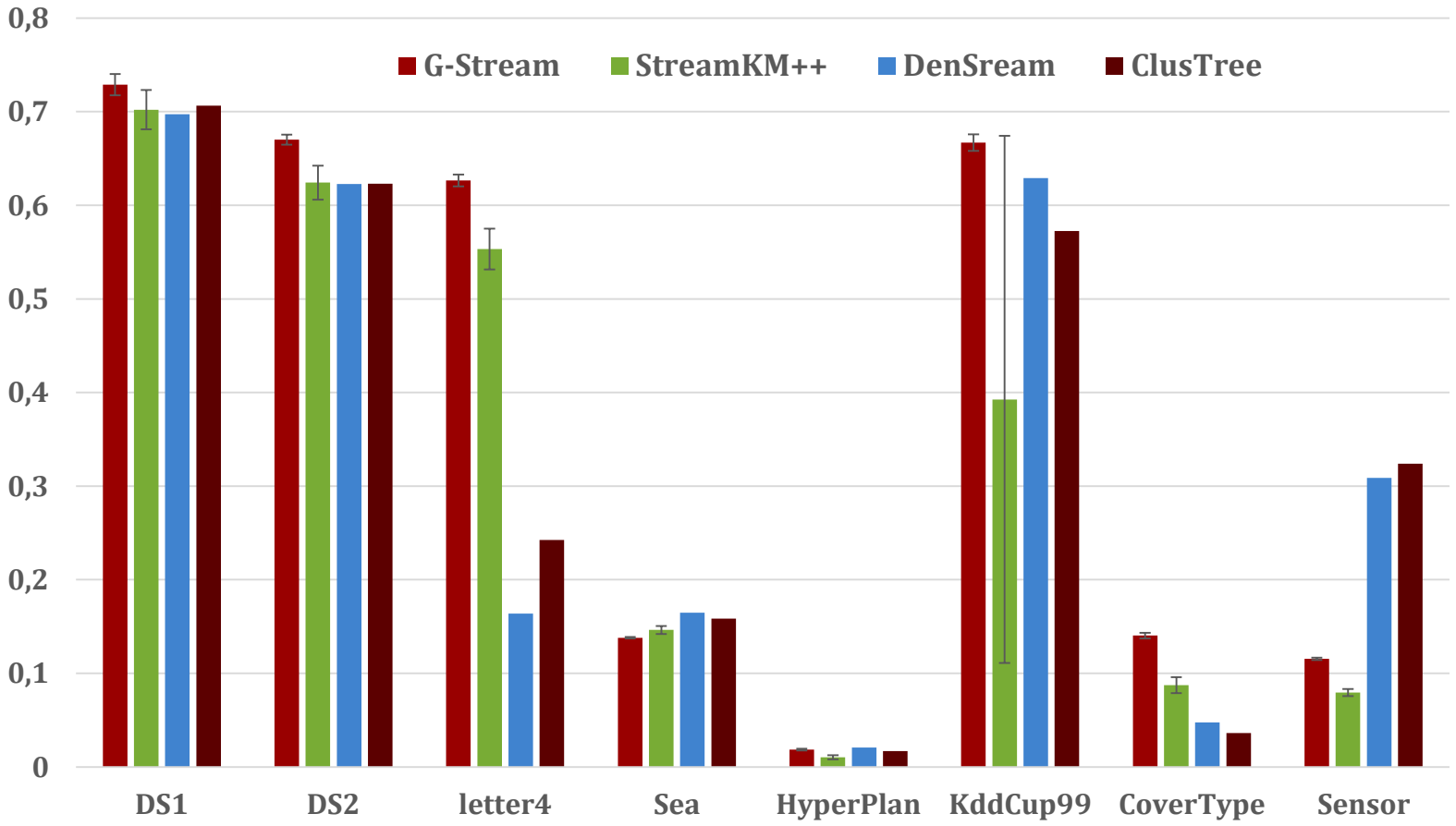
EVALUATIONS

NMI

DB



NMI


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
G-Stream
StreamKM++
DenSream
ClusTree
DS1
DS2
letter4
Sea
HyperPlan
KddCup99
CoverType
Sensor

CONCLUSION & PERSPECTIVES

- Une diversité d'approches
  - Approches probabilistes: EM online, OVEM-DyMix (Online Variationnal Expectation Maximisation for Dynamic Mixture model)

- L'apprentissage en flux de données n'est pas généralement réalisé en un seul passage. C'est un processus évolutif ..!

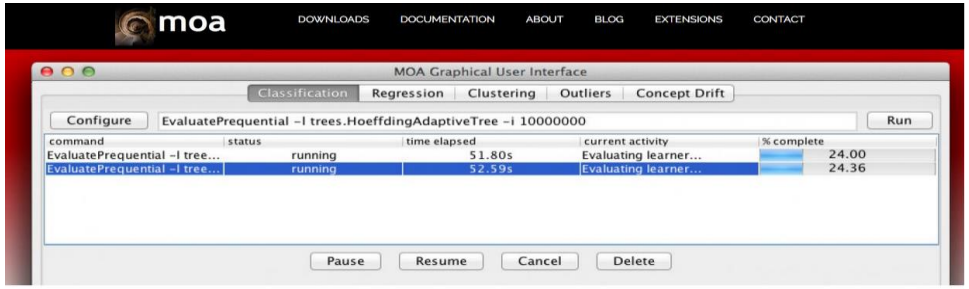
- L'intérêt de combiner du «batch» et du «streaming»

- Différents outils (dont spark)

OUTILS DISPONIBLES

- {M}assive {O}nline {A}nalysis MOA (Bifet et al. 2010)
  - Connecté à WEKA

http://moa.cms.waikato.ac.nz/details/stream-clustering/



- StreamDM

http://streamdm.noahlab.com.hk/

Spark Streaming

http://spark.apache.org/

Outils disponibles

- Apache Mahout

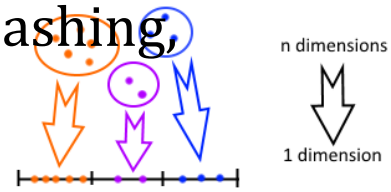
- SAMOA
  - G. De Francisci Morales, et al . "SAMOA: Scalable Advanced Massive Online Analysis." Journal of Machine Learning Research, 16(Jan):149–153, 2015.

  http://samoa.incubator.apache.org/

https://flink.apache.org/

QUELQUES CHALLENGES

- S'adapter constamment à l'évolution de l'environnement (auto-configurable)
- La prise en compte de l'évolution des données et du recouvrement
- La prise en compte des différents types de données
- Accélération des programmes: l'échantillonnage, hashing, projection aléatoire (LSH)



- La visualisation
- Facile à utiliser « **ease to use** »

Références

- G. De Francisci Morales, A. Bifet. "SAMOA: Scalable Advanced Massive Online Analysis." Journal of Machine Learning Research, 16(Jan):149–153, 2015.
- E. Sparks, A. Talwalkar, V. Smith, J. Kottalam, X. Pan, J. Gonzalez, M. Franklin, M. I. Jordan, T. Kraska. *MLI: An API for Distributed Machine Learning*. International Conference on Data Mining (ICDM), 2013
- Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. A framework for clustering evolving data streams. In Proceedings of the International Conference on Very Large Data Bases (VLDB '03), pages 81–92, 2003.
- Aggarwal C C, Han J, Wang J, Yu P S. A framework for pro- jected clustering of high dimensional data streams. In Proc. the 30th International Conference on Very Large Data Bases, Volume 30, Aug. 29-Sept. 3, 2004, pp.852-863.
- Amini A, Wah TY, Saboohi H. On density-based data streams clustering algorithms: A survey. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 116–141 Jan. 2014. DOI 10.1007/s11390-013-1416-3
- Jonathan A. Silva, Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, André C. P. L. F. de Carvalho, and João Gama. 2013. Data stream clustering: A survey. *ACM Comput. Surv.* 46, 1, Article 13 (July 2013), 31 pages
- Gianmarco De Francisci Morales, Joao Gama, Albert Bifet, Wei Fan. Big Data Stream Mining Tutorial at IEEE BigData 201
- Albert Bifet, André C P L F de Carvalho, João Gama. BigData Stream Mining. ECML-PKDD Summer school 2015
- Hani El Assaad, Allou Samé, Gérard Govaert, Patrice Aknin: Model-Based Clustering of Temporal Data. ICANN 2013: 9-16
- Mohammed Ghesmoune, Mustapha Lebbah, Hanene Azzag. Micro-Batching Growing Neural Gas for Clustering Data Streams using Spark Streaming. Procedia Computer Science journal (2015) pp. 158-166. Doi 10.1016/j.procs.2015.07.290.
- Mohammed Ghesmoune, Mustapha Lebbah, and Hanene Azzag. Clustering over data streams based on growing neural gas. PAKDD (2) 2015: 134-145.