Analyse de Données – ID
Centres Mobiles

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – FRE 2729
Site de l'Ecole Polytechnique de l'université de Nantes

Plan

Points abordés

- Généralités
- Clustering hiérarchique
- **Clustering par partitionnement**
  - **K-Means, Nuées dynamiques**
  - CLARA (Clustering LARge Applications), Fuzzy C-Means, ...

K-Means

Centres mobiles – K-means – K-moyennes

Principe

- (Forgy 1965, MacQueen 1967)
- répartir les N points en K ensembles disjoints
- regrouper les points proches
- problème de minimisation :

$$J = \sum_{g=1}^{K} \sum_{i \in C_g} d^2(x_i, \mu_g)$$

- $\Rightarrow$ NP difficile
- on peut juste trouver un minimum local

K-Means

Algorithme

- Initialiser $\mu_1, \cdots \mu_K$
- Répeter
  - affectation de chaque point à son cluster le plus proche

$$C(x_i) = \min_g d(x_i, \mu_g)$$

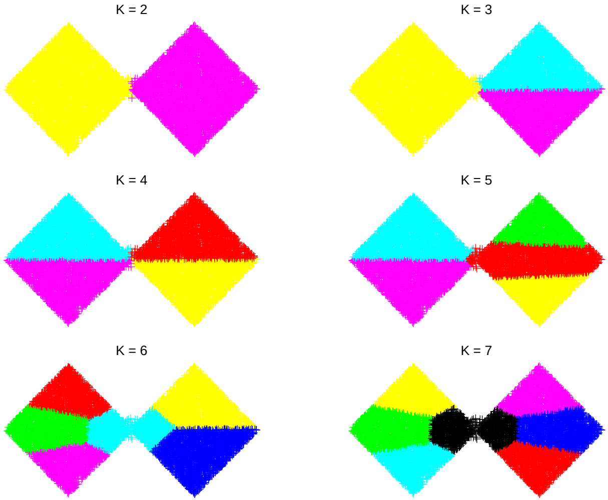
  - recalculer le centre μ_i de chaque cluster

$$\mu_g = \frac{1}{N_g} \sum_{i \in C_g} x_i$$

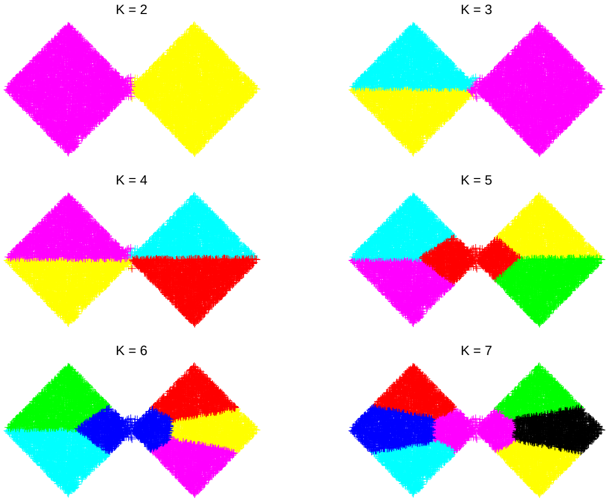
- Tant que $\| \Delta\mu \| > \epsilon$

Complexité $= O(KnI)$ (I: itérations)

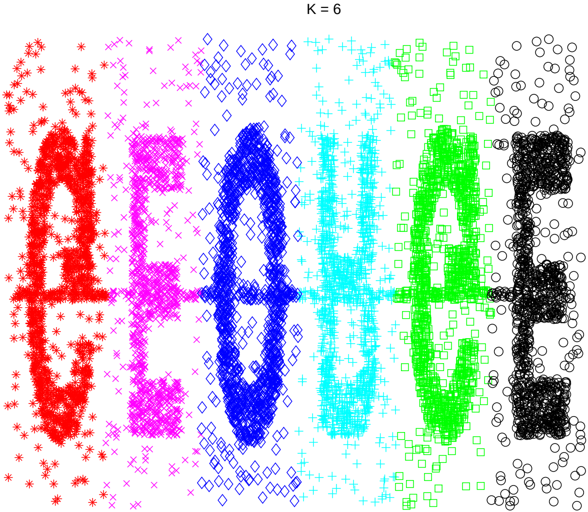
K-Means : exemple 1



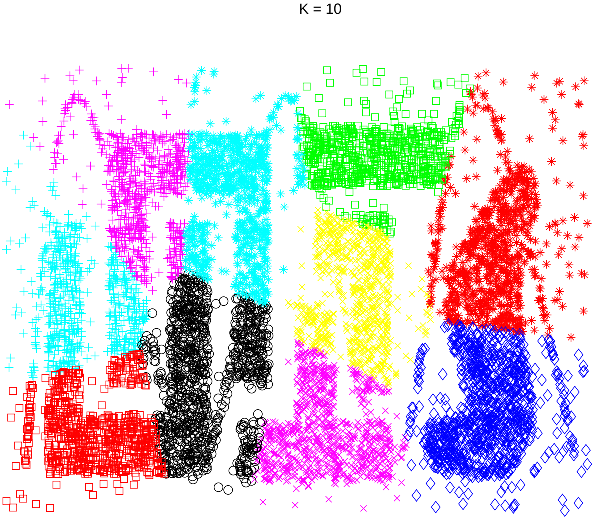
K-Means : exemple 1



K-Means : exemple 2

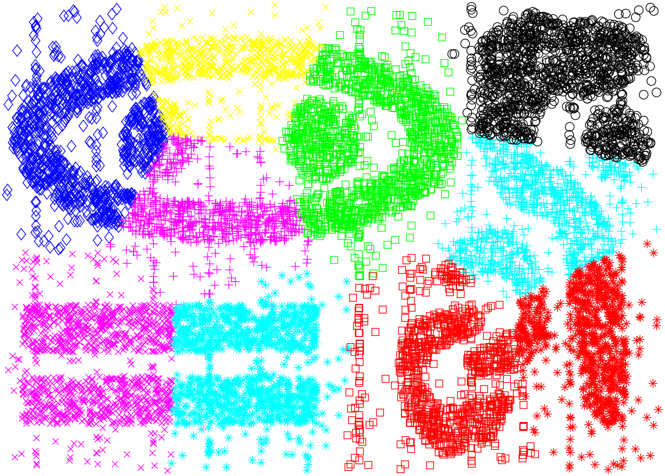


K-Means : exemple 3



K-Means : exemple 4

K = 10



K-Means

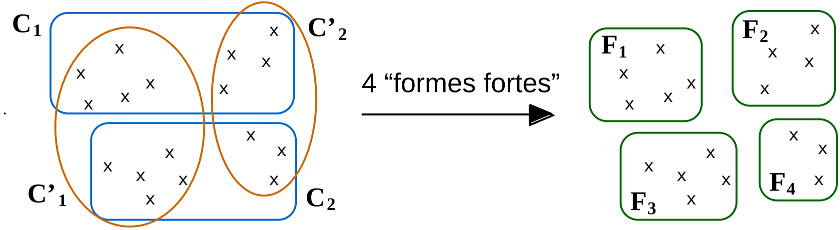
Initialisation

- Initialisation des μ_i :
  - aléatoirement dans l'intervalle de définition des x_i
  - aléatoirement dans l'ensemble des x_i
- Des initialisations différentes de peuvent mener à des clusters différents (problème de minima locaux)

Formes fortes

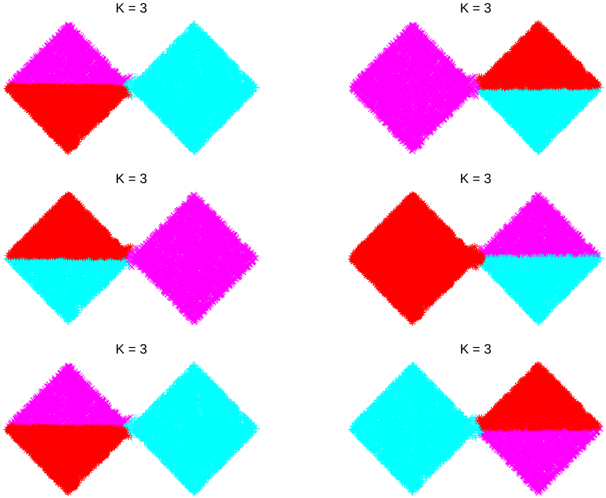
- méthode **générale** pour obtenir des clusters "stables"
  - on répète l'algo des K-Means r fois
  - on regroupe ensemble les x_i qui se retrouvent toujours dans les mêmes clusters.
  - on supprime les regroupements "faibles"

K-Means : formes fortes (Bisson 2001)



K-Means : formes fortes

K-Means répété 6 fois



K-Means : formes fortes

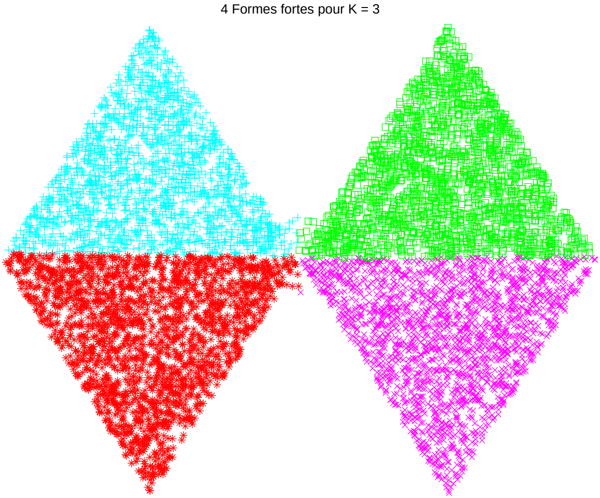
- On trouve 5 regroupements de points différents :

	F_1	F_2	F_3	F_4	F_5
N_i	2040	1940	49	2042	1929

- F_3 n'est pas représentatif
- F_1, F_1, F_4 et F_5 sont les formes fortes
- on peut recalculer les clusters à partir des centres des formes fortes

K-Means : formes fortes

K-Means répété 6 fois



K-Means séquentiels

Adaptation des k-Means lorsque les exemples arrivent au fur et à mesure

Algorithme

- Initialiser $\mu_1, \cdots \mu_K$
- Initialiser $n_1, \cdots n_K$ à 0
- Répeter
  - acquérir x
  - affectation de chaque point à son cluster le plus proche

$$\mu_i = argmin_g d(x, \mu_g)$$

  - incrémenter n_i
  - recalculer le centre μ_i de ce cluster

$$\mu_i = \mu_i + \frac{1}{n_i}(x - \mu_i)$$

Nuées Dynamiques (Diday 1972, 1974)

Principe

- Généralisation des K-Means
- Utilisation de noyaux = représentation d'un cluster
  - barycentre ($= \mu$ pour les K-means)
  - *n* points représentatifs
  - ...