Regroupement et projection

Javiera CASTILLO NAVARRO
Gaston LENCZNER
Guillaume VAUDAUX RUTH
Adrien CHAN-HON-TONG
ONERA

Introduction

Apprentissage supervisé : on veut construire une fonction f tel que le signe de $f(x)$ approxime celui de $y(x)$ avec y connu uniquement sur une base d'apprentissage.

Introduction

Apprentissage supervisé : on veut construire une fonction f tel que le signe de $f(x)$ approxime celui de $y(x)$ avec y connu uniquement sur une base d'apprentissage.

Regroupement : on veut regrouper dans des points en groupe en fonction de leur distance.

Introduction

Regroupement : on veut regrouper dans des points en groupe en fonction de leur distance.

$\Rightarrow$ Si on fixe le nombre de groupes, le problème est bien posé !

$\Rightarrow$ Si plus difficile de voir à quoi ça peut servir mais le problème est bien posé.

Exemple

Pour compresser une image, on peut vouloir grouper des couleurs.

En fixant une couleur par groupe, on diminue le nombre de couleurs présentes.

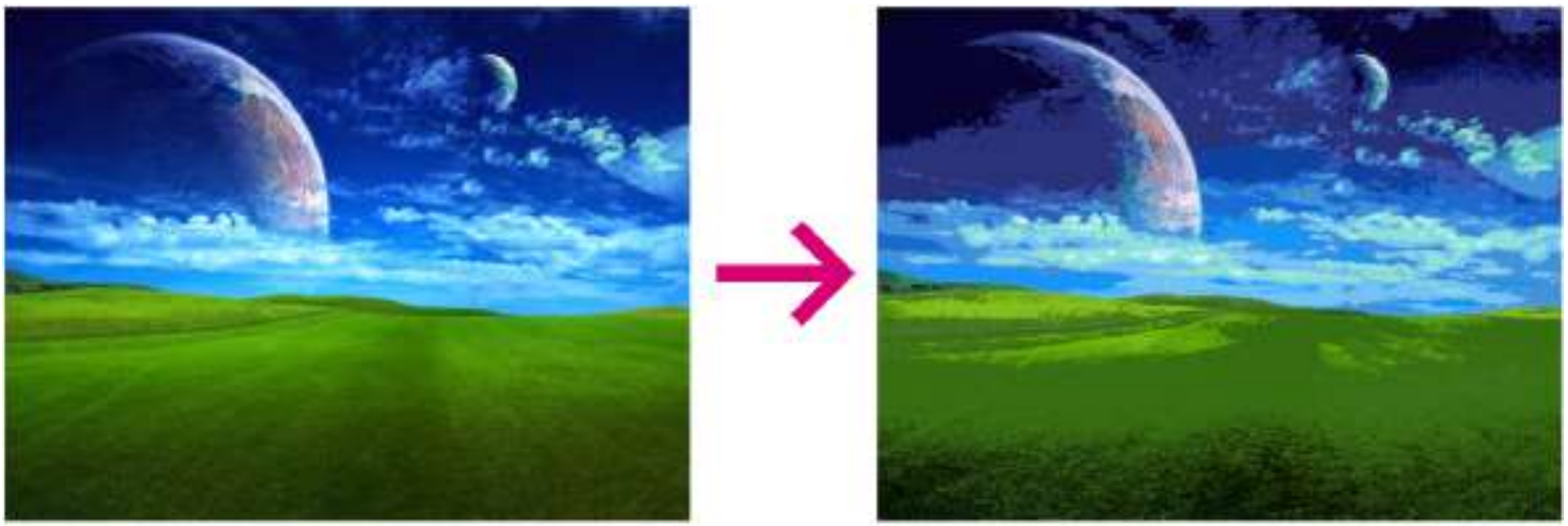


Fig.1a: The Visual Effect of 24K Colors Depth Image Segmentation

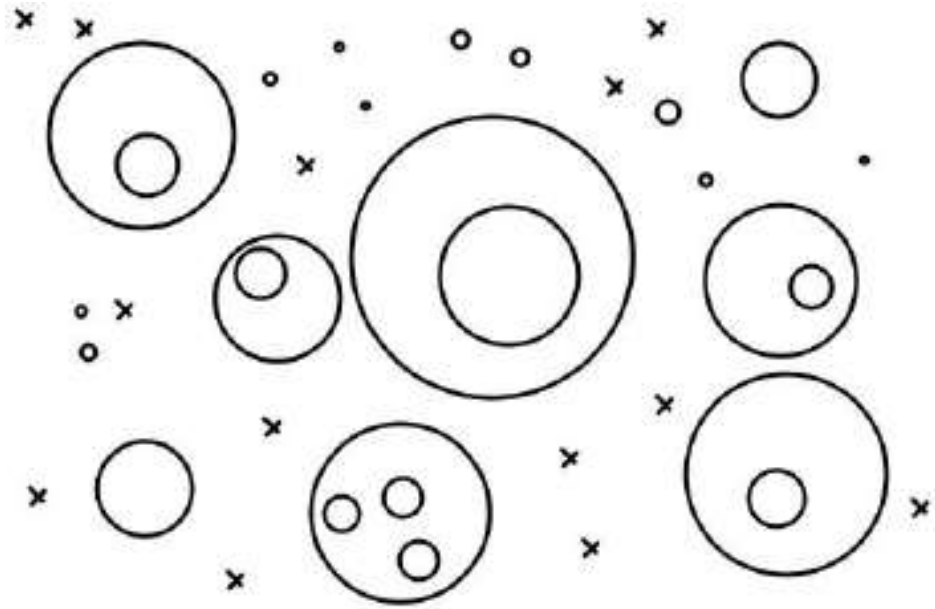
Introduction

Compression et apprentissage peuvent trouver des points communs.

si vous devez résumé une imagette en 1 nombre, choisir celui qui est représenté peut être une *bonne* solution.

Introduction

La sémantique est généralement une compression mais une compression n'est **PAS** nécessairement sémantique :



Coder la taille plutôt que de coder le fait d'être une croix est optimal vis à vis de la distance euclidienne, mais détruit toute la sémantique croix/cercle.

Plan du cours

2 méthodes de *compression* (qu'il conviendrait d'étudier pour cela - ce qu'on ne fera pas)
Puis une application à l'apprentissage.

- K moyennes
- PCA
- Approche Sac de mots

Sac de mots ce qu'on faisait de mieux avant le deep learning...

K moyennes

Soit $x_1, ..., x_N \in \mathbb{R}^D$ et K un nombre,
le problème dit des K moyennes consiste à résoudre

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} \min_{k \in \{1,...,K\}} ||c_k - x_n||_2^2$$

Il est important de ne pas confondre *problème* et solution. Ici, on se donne un problème, on ne parle pas encore de solution.

K moyennes

Soit $x_1, ..., x_N \in \mathbb{R}^D$ et K un nombre,
le problème dit des K moyennes consiste à résoudre

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} \min_{k \in \{1,...,K\}} ||c_k - x_n||_2^2$$

ou de façon équivalente

$$\min_{c \in \mathbb{R}^{K \times D}, \sigma \in \{0,1\}^{N \times K}} \sum_{n \in \{1,...,N\}, k \in \{1,...,K\}} \sigma_{n,k} ||c_k - x_n||_2^2$$

$$sc : \forall n \in \{1, ..., N\}, \sum_{k \in \{1,...,K\}} \sigma_{n,k} = 1$$

$$\forall n \in \{1, ..., N\}, i, j \in \{1, ..., K\} \ \sigma_{n,i} = 1 \Leftrightarrow ||c_i - x_n|| \leq ||c_j - x_n||$$

K moyennes

Soit $x_1, ..., x_N \in \mathbb{R}^D$ et K un nombre,
le problème dit des K moyennes consiste à résoudre

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} \min_{k \in \{1,...,K\}} ||c_k - x_n||_2^2$$

ou de façon équivalente

$$\min_{c \in \mathbb{R}^{K \times D}, \sigma \in \mathbb{R}_+^{N \times K}} \sum_{n \in \{1,...,N\}, k \in \{1,...,K\}} \sigma_{n,k} ||c_k - x_n||_2^2$$

$$sc : \forall n \in \{1, ..., N\}, \sum_{k \in \{1,...,K\}} \sigma_{n,k} \geq 1$$

K moyennes

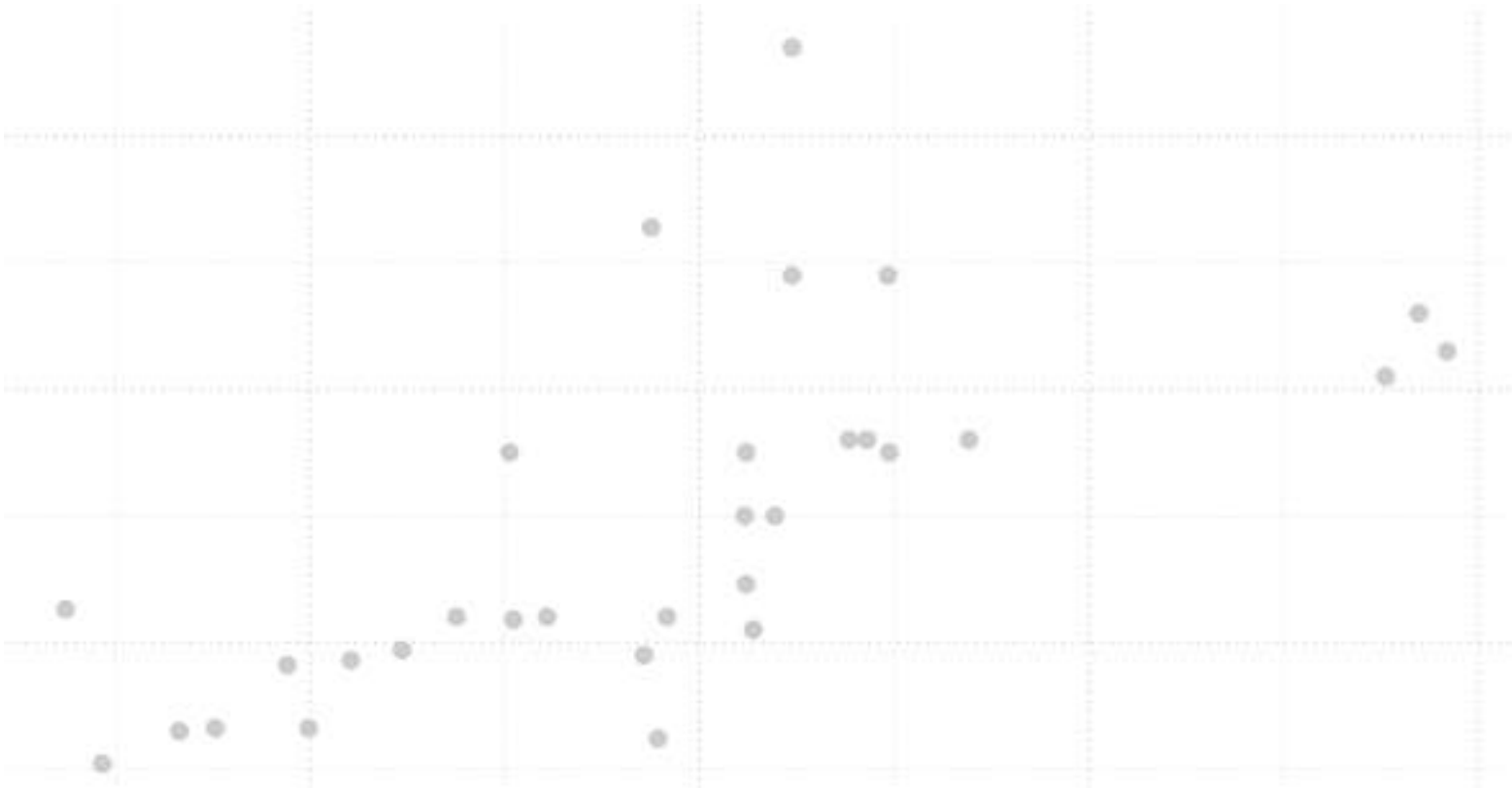
Soit $x_1, ..., x_N \in \mathbb{R}^D$ et K un nombre,
le problème dit des K moyennes consiste à résoudre

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} \min_{k \in \{1,...,K\}} ||c_k - x_n||_2^2$$

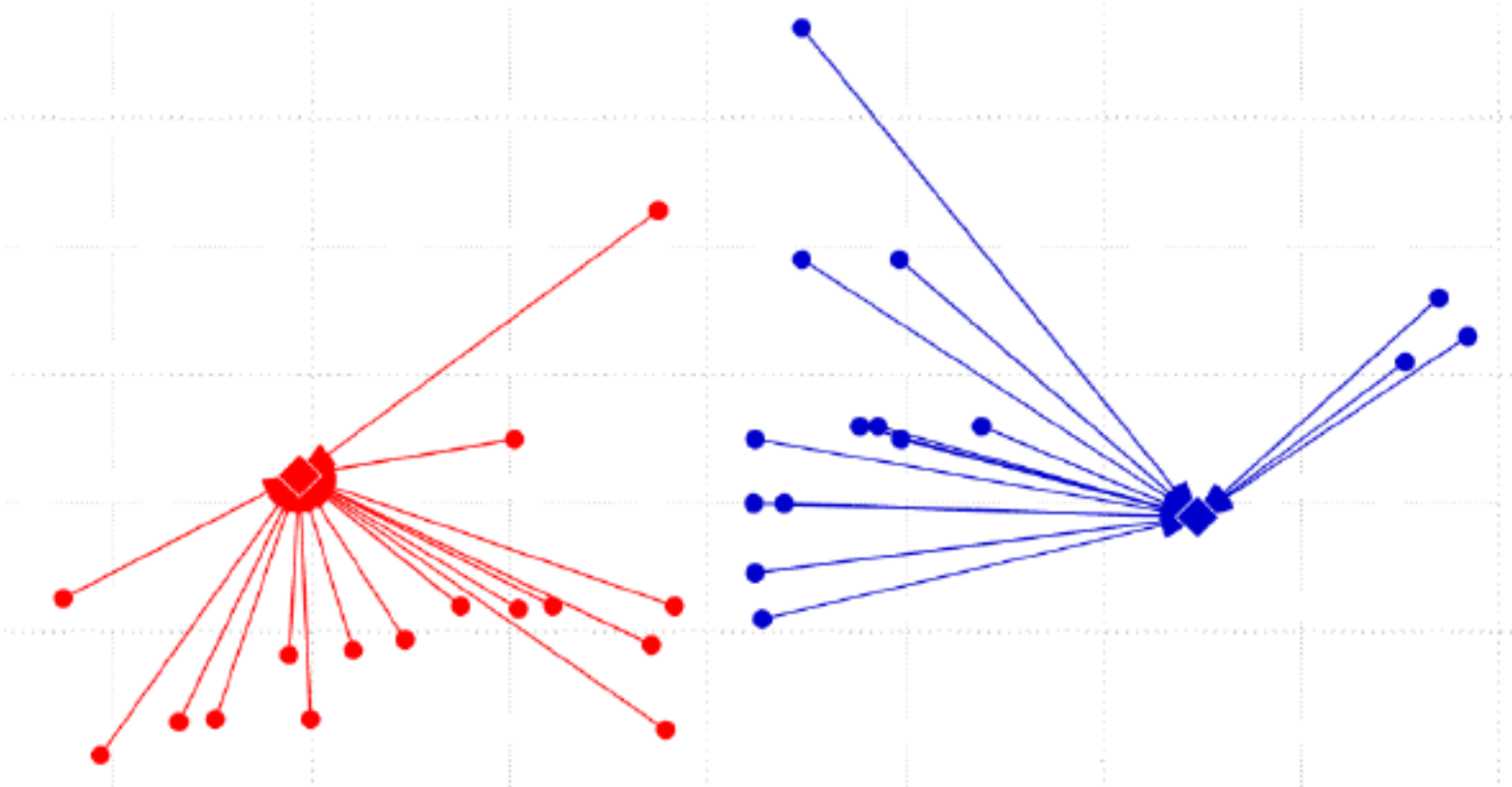
ou de façon équivalente

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,...,K\}^N} \sum_{n \in \{1,...,N\}} ||c_{\rho_n} - x_n||_2^2$$

K moyennes



K moyennes



K moyennes

Soit $x_1, ..., x_N \in \mathbb{R}^D$ et K un nombre,
le problème dit des K moyennes consiste à résoudre

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} \min_{k \in \{1,...,K\}} ||c_k - x_n||_2^2$$

C'est un problème NP complet dès $D = 2$ quand $N, K \to \infty$ et dès $K = 2$ quand $N, D \to \infty$.
Donc, en pratique on ne sait pas résoudre ce problème **exactement**.

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Si c est fixé ?

$$\min_{\rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

$$\Leftrightarrow \sum_{n \in \{1,\dots,N\}} \min_{\rho_n \in \{1,\dots,K\}} ||c_{\rho_n} - x_n||_2^2$$

$$\Leftrightarrow \forall n \in \{1, \dots, N\},\ \rho_n = \arg\min_{k \in \{1,\dots,K\}} ||c_k - x_n||_2^2$$

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Si c est fixé, la solution optimale est d'associer chaque point x_n à son plus proche voisin dans $\{c_1, \dots, c_K\}$

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,...,K\}^N} \sum_{n \in \{1,...,N\}} ||c_{\rho_n} - x_n||_2^2$$

Si ρ est fixé ?

$$\min_{c \in \mathbb{R}^{K \times D}} \sum_{n \in \{1,...,N\}} ||c_{\rho_n} - x_n||_2^2$$

$$\Leftrightarrow \min_{c \in \mathbb{R}^{K \times D}} \sum_{k \in \{1,...,K\}} \sum_{n \in \{1,...,N\} \ / \ \rho_n = k} ||c_k - x_n||_2^2$$

$$\Leftrightarrow \forall k \in \{1, ..., K\}, \ c_k = \arg\min_{c_k \in \mathbb{R}^D} \sum_{n \in \{1,...,N\} \ / \ \rho_n = k} ||c_k - x_n||_2^2$$

Algorithme des K moyennes

$$\min_{\psi \in \mathbb{R}^D} \sum_{r \in \{1,\dots,R\}} ||\psi - \phi_r||_2^2$$

$$\Leftrightarrow \min_{\psi \in \mathbb{R}^D} \; R\psi^T\psi - 2\psi^T\Big(\sum_{r \in \{1,\dots,R\}} \phi_r\Big) + \sum_{r \in \{1,\dots,R\}} \phi_r^T\phi_r$$

$$\Leftrightarrow \min_{\psi \in \mathbb{R}^D} \; R\psi^T\psi - 2\psi^T\Phi$$

$$\Leftrightarrow \psi = \frac{1}{R}\Phi$$

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Si ρ est fixé, la solution optimale est de choisir c_k comme le centre (la moyenne) de $\{x_n \ / \ \rho_n = k\}$

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Algorithme des K moyennes (approximation) ?

1. initialiser c
2. calculer ρ optimal à c fixé
3. $\rho_c = \rho$
4. faire en boucle
   - 4.1 calculer c optimal à ρ fixé
   - 4.2 calculer ρ optimal à c fixé
   - 4.3 si $\rho = \rho_c$ arrêter sinon, $\rho_c = \rho$

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Algorithme des K moyennes ?

Optimum local possible !

$D = K = 2$, $N = 4$, imaginez un rectangle non carré, l'algorithme s'arrête si on a 2 groupes de 2 points adjacents – sauf que c'est optimal que s'il s'agit des petits cotés et non des grands !

Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

Algorithme des K moyennes (approximation) ?
Convergence garantie car le critère décroît :
Quand on actualise partiellement le critère $\sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$ vis à vis de c ou ρ (l'autre étant fixé), on pourrait laisser la variable inchangée – si on la change, c'est pour faire diminuer le critère.
Donc le critère décroît strictement (s'il y a du changement).
Et il y a un nombre fini de ρ !

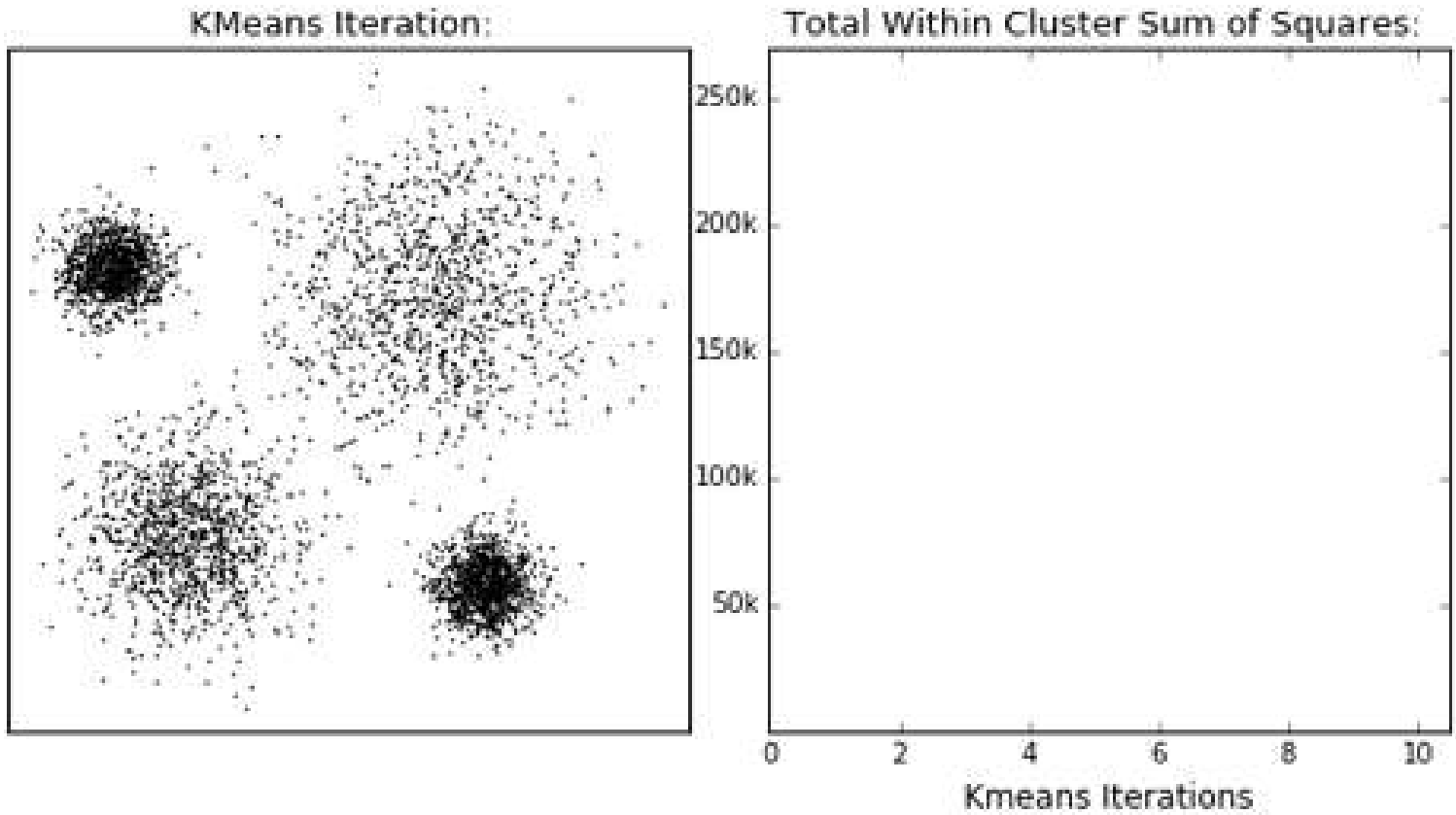
Algorithme des K moyennes

$$\min_{c \in \mathbb{R}^{K \times D}, \rho \in \{1,\dots,K\}^N} \sum_{n \in \{1,\dots,N\}} ||c_{\rho_n} - x_n||_2^2$$

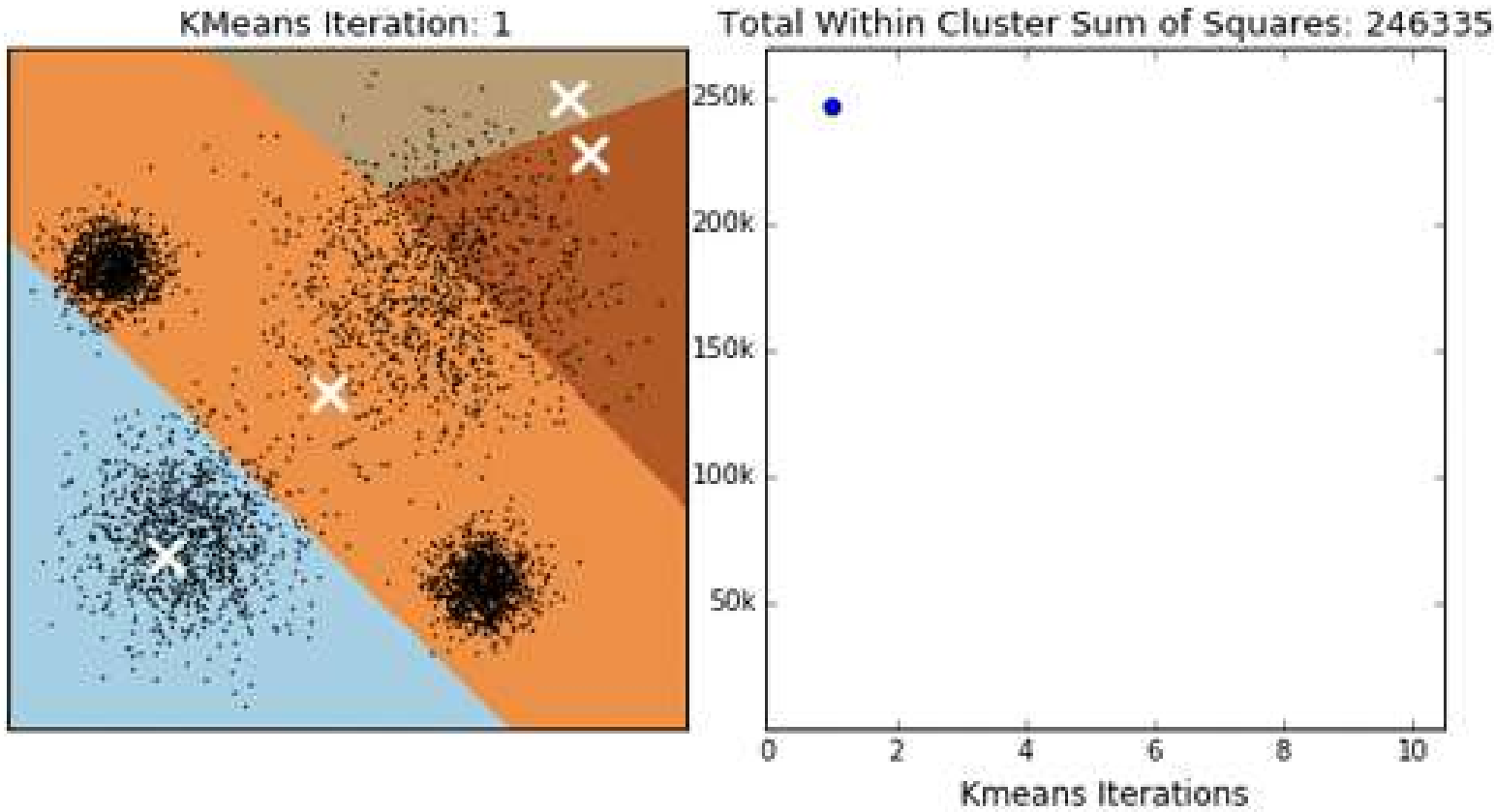
Algorithme des K moyennes (approximation) :

1. initialiser c
2. calculer ρ optimal à c fixé
3. $\rho_c = \rho$
4. faire en boucle
   - 4.1 calculer c optimal à ρ fixé
   - 4.2 calculer ρ optimal à c fixé
   - 4.3 si $\rho = \rho_c$ arrêter sinon, $\rho_c = \rho$

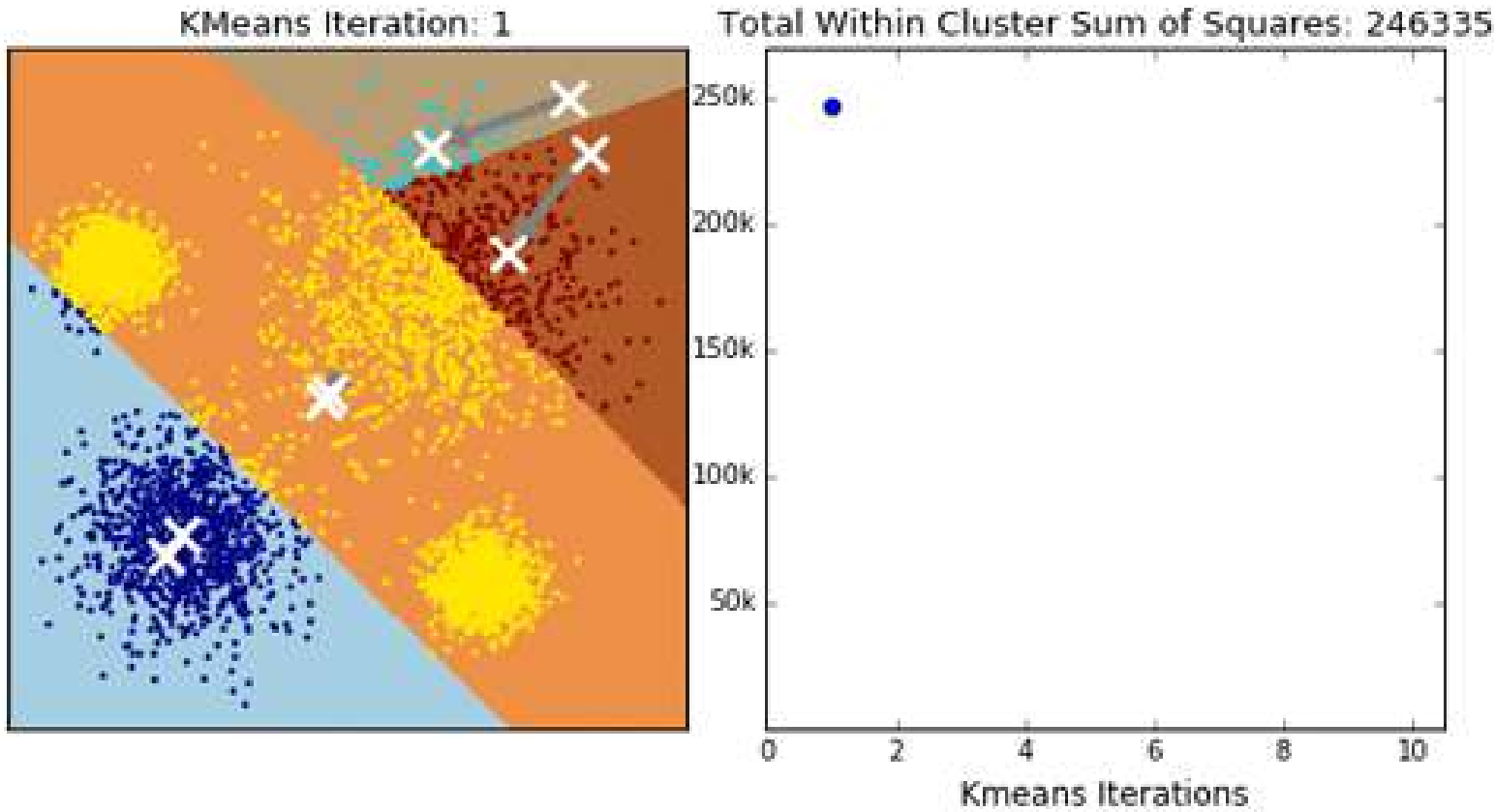
Algorithme des K moyennes



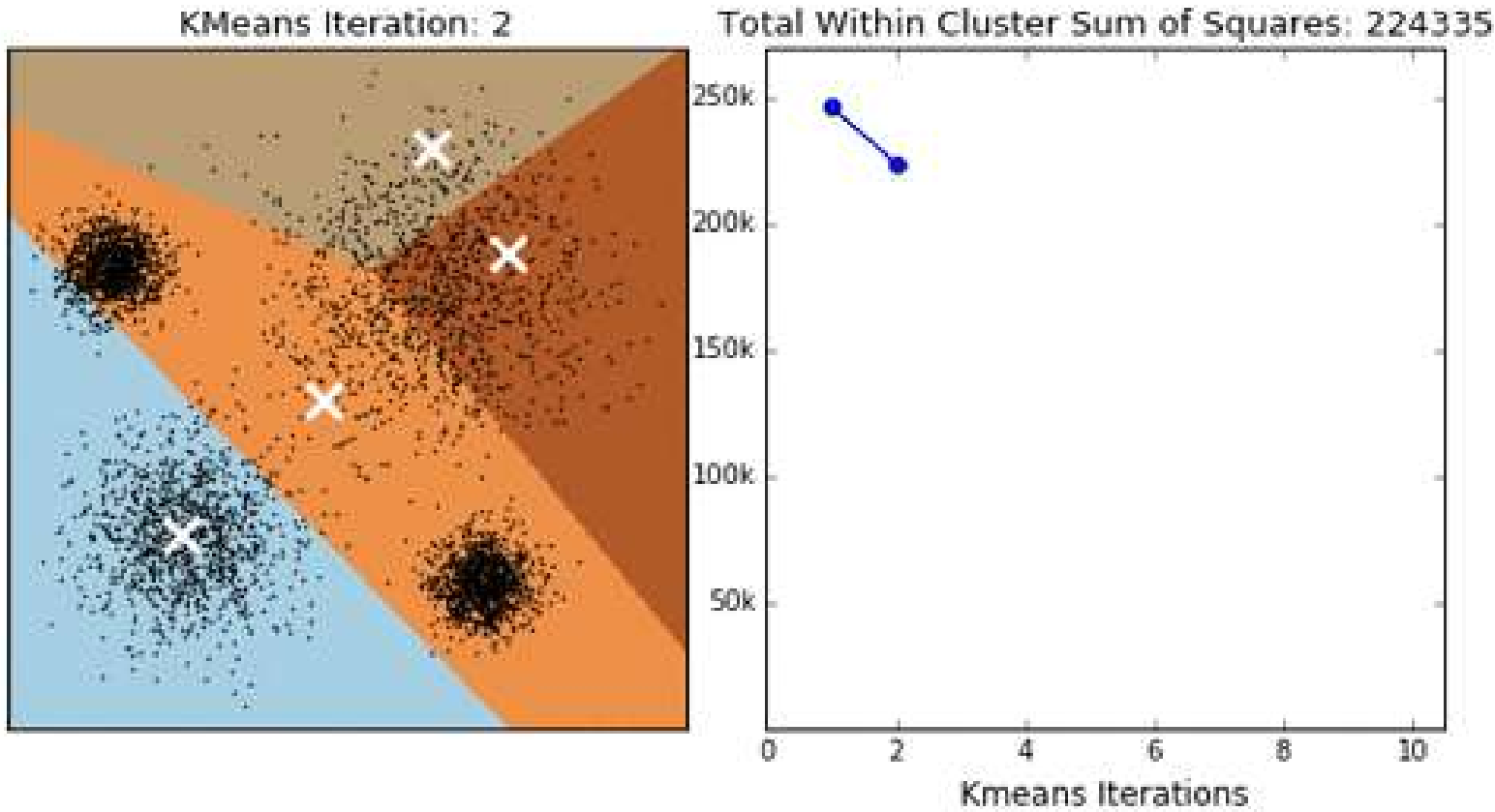
Algorithme des K moyennes



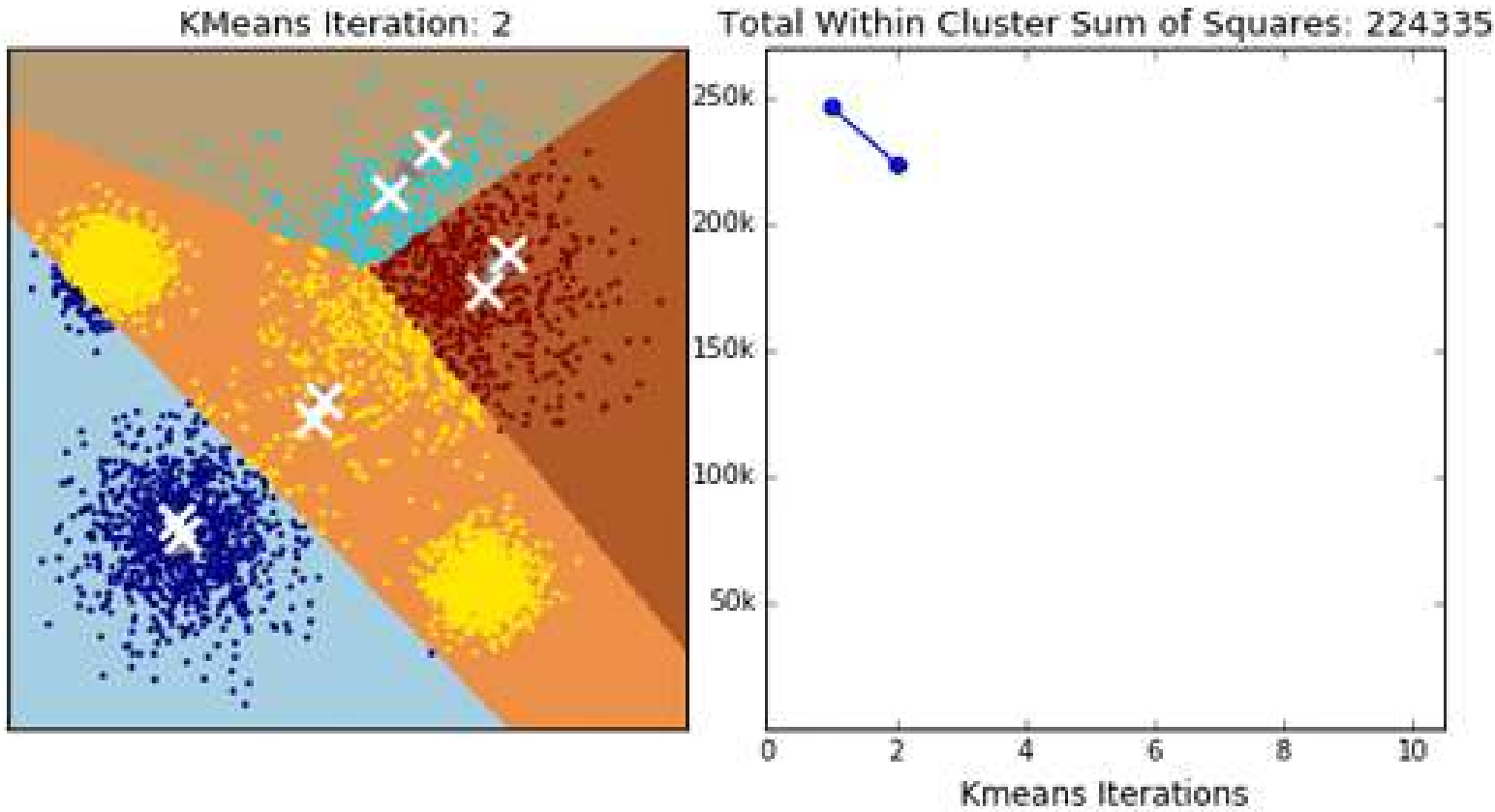
Algorithme des K moyennes



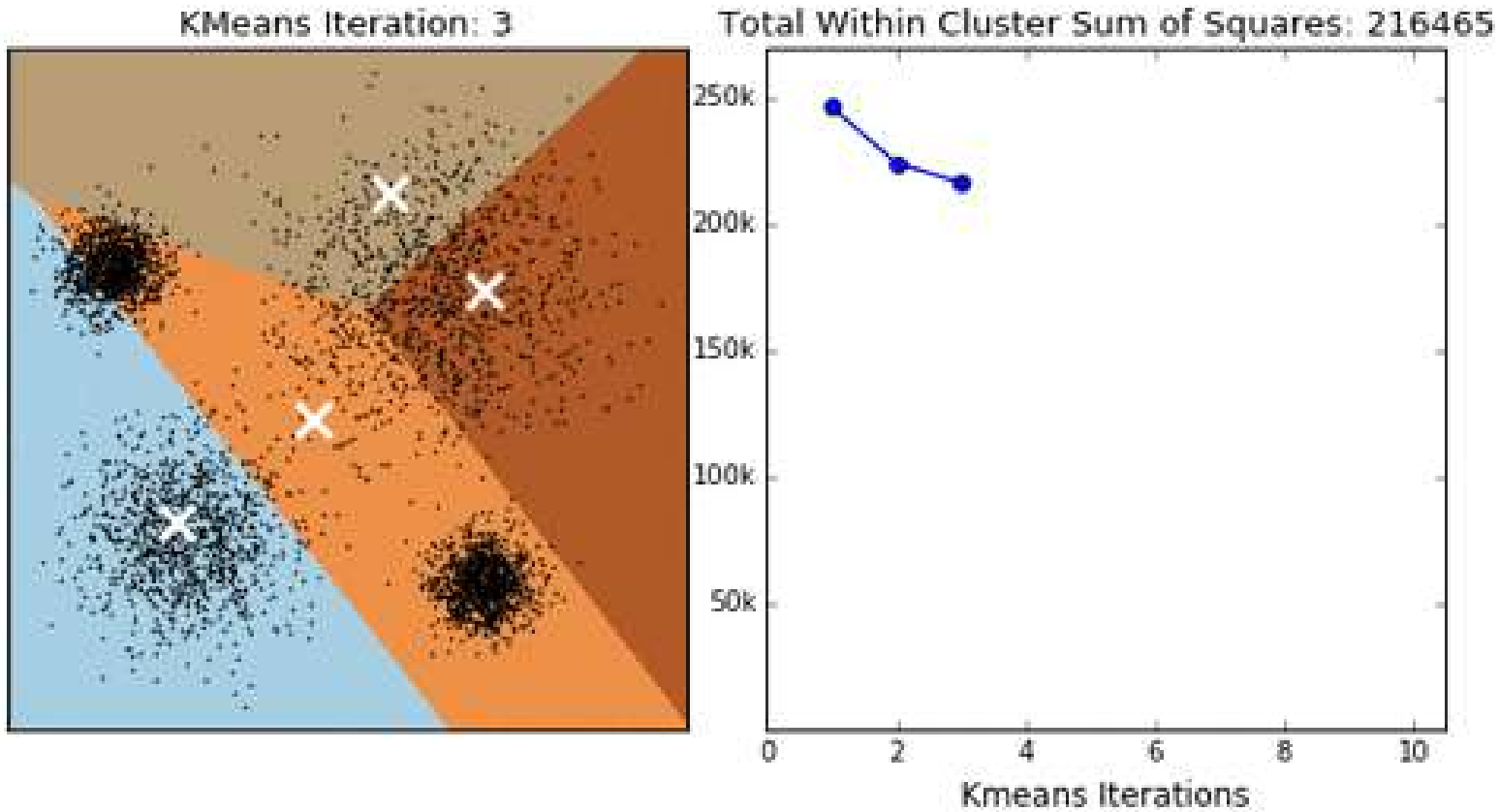
Algorithme des K moyennes



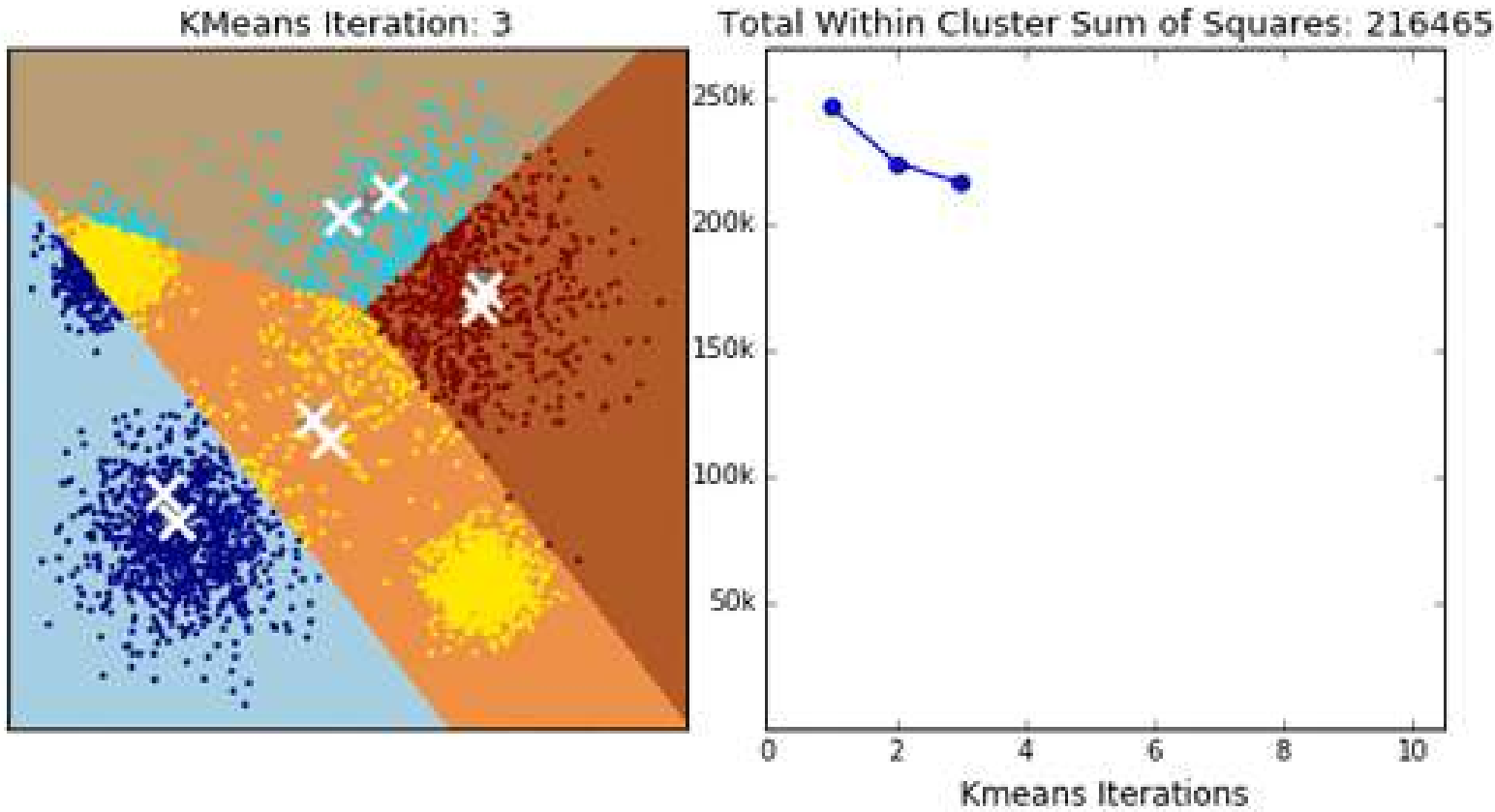
Algorithme des K moyennes



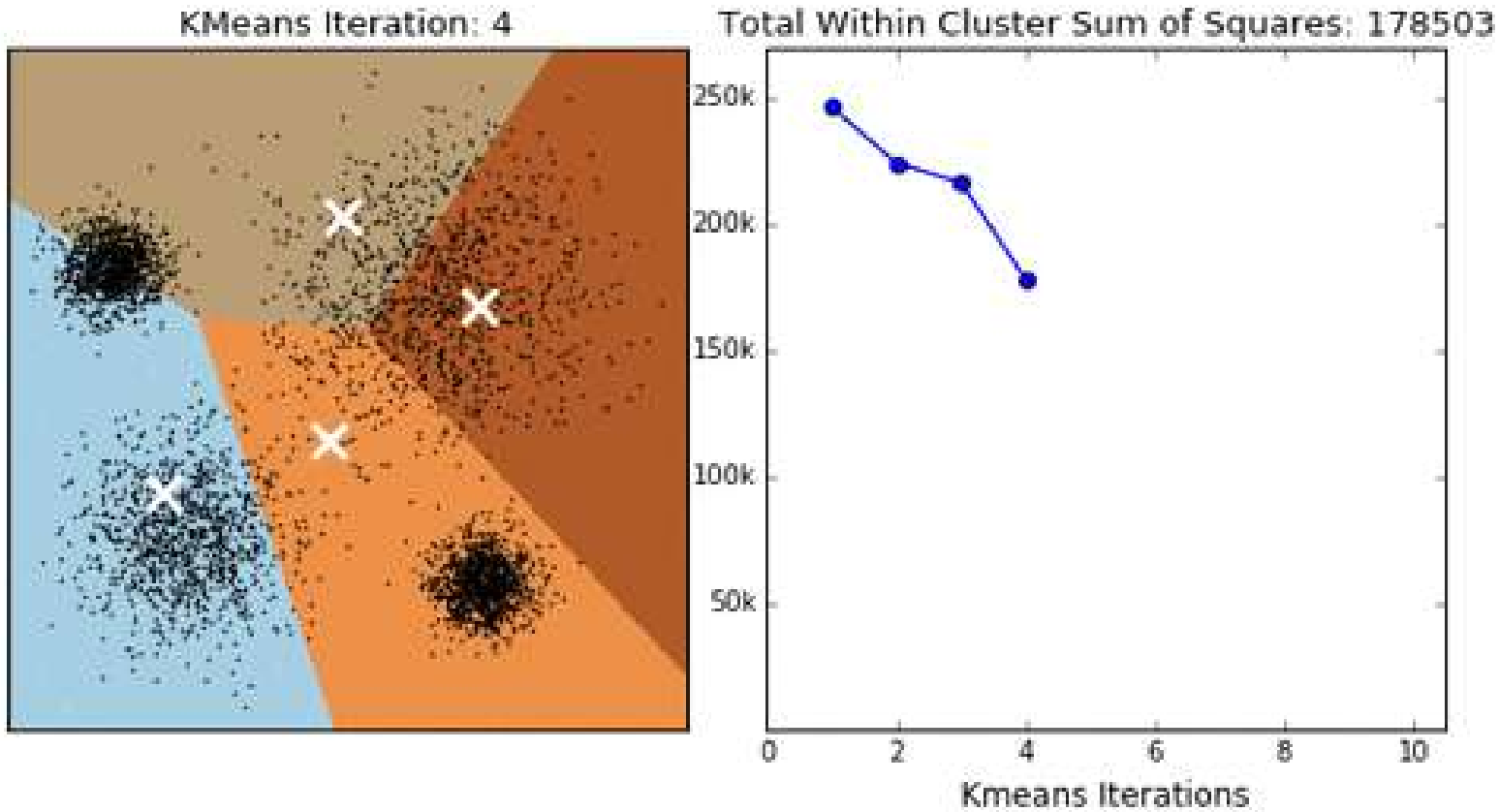
Algorithme des K moyennes



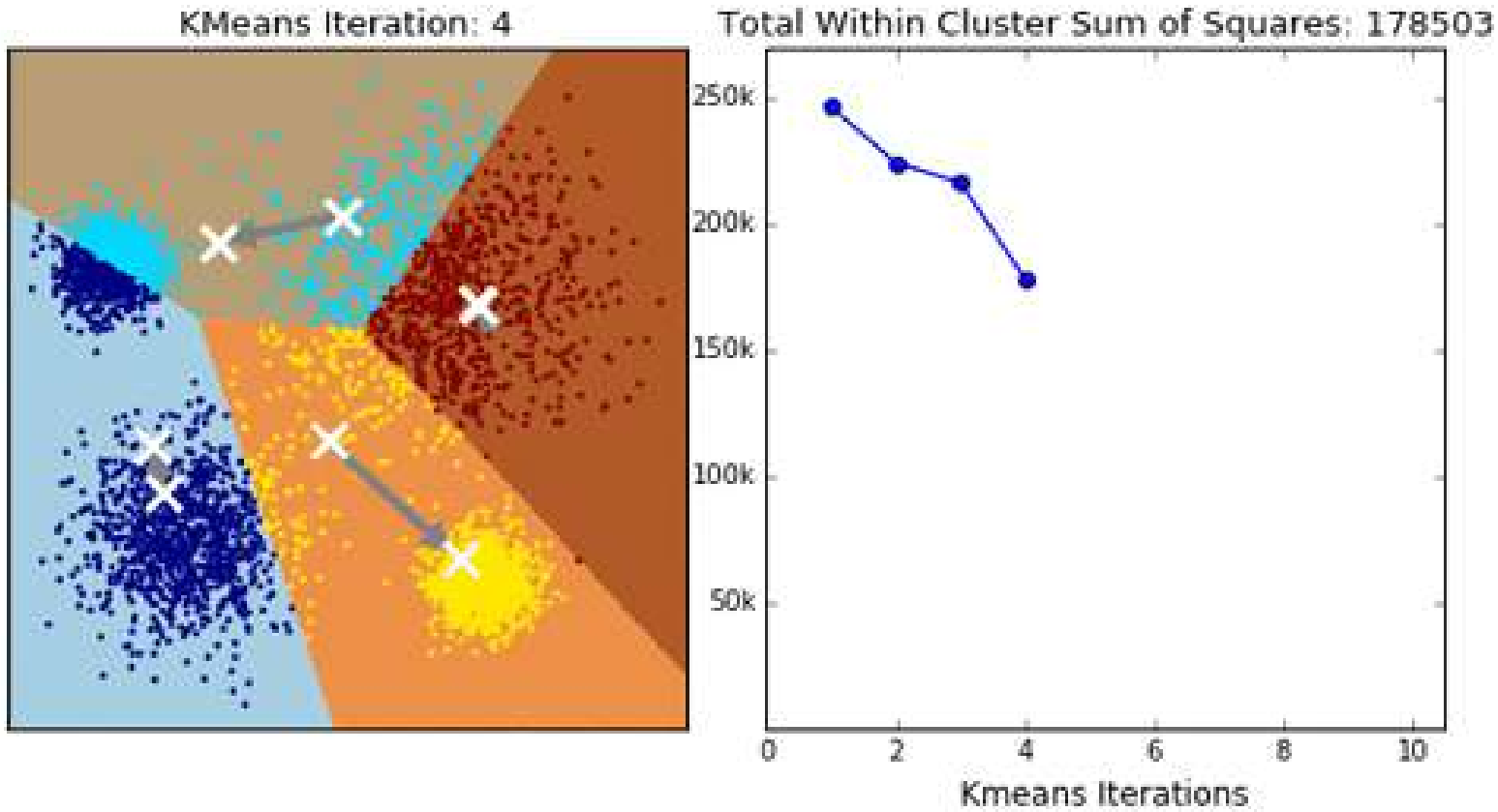
Algorithme des K moyennes



Algorithme des K moyennes



Algorithme des K moyennes



Algorithme des K moyennes

L'initialisation est primordiale.
L'initialisation de référence Kmeans++ : tirer les centres via leur distance aux centres précédents

Plan du cours

- *K moyennes*
- PCA
- Approche Sac de mots

PCA

$x_1, ..., x_N \in \mathbb{R}^D$ peut-on projeter les x dans un espace plus petit $\mathbb{R}^{\mathcal{D}}$ ($\mathcal{D} \ll D$) en perdant un minimum d'information ?

PCA

$x_1, ..., x_N \in \mathbb{R}^D$ peut-on projeter les x dans un espace plus petit $\mathbb{R}^{\mathcal{D}}$ ($\mathcal{D} \ll D$) en perdant un minimum d'information ?

$\Rightarrow$ information ? ? ?

encore une fois garder l'information essentiel c'est compresser mais compresser peut détruire l'information...

PCA

$x_1, ..., x_N \in \mathbb{R}^D$ / $\sum_n x_n = 0$ peut-on **linéairement** projeter les x dans un espace plus petit $\mathbb{R}^{\mathcal{D}}$ ($\mathcal{D} \ll D$) en conservant un maximum de variance ?
(variance == information ? ?)

$$\max_{A \in \mathrm{R}^{D \times D} \ / \ AA^T = I} \sum_n ||A_{\{1,...,\mathcal{D}\}} x||^2$$

Cas simple PCA

$x_1, ..., x_N \in \mathbb{R}^D$ / $\sum_n x_n = 0$ peut-on linéairement projeter les x dans un espace plus petit $\mathbb{R}^{\mathcal{D}}$ ($\mathcal{D} \ll D$) en conservant un maximum de variance ?

$$D = 2 \text{ et } \mathcal{D} = 1$$

on cherche donc $u \in \mathbb{R}^2$ tel que

$$u = \max_u \sum_n \frac{(u^T x_n)^2}{u^T u}$$

Cas simple PCA

on cherche $u \in \mathbb{R}^2$ tel que

$$u = \max_u \sum_n \frac{(u^T x_n)^2}{u^T u}$$

Notons, $\alpha = \sum_n x_{n,1}^2$, $\beta = \sum_n x_{n,2}^2$ et $\gamma = \sum_n x_{n,1} x_{n,2}$ et paramétrons $u = (1 \quad t)$, on obtient l'équation :

$$\max_t \frac{1}{1+t^2} \sum_n x_{n,1}^2 + x_{n,2}^2 t^2 + 2 x_{n,1} x_{n,2} t = \frac{1}{1+t^2}(\alpha + \beta t^2 + 2\gamma t)$$

Cas simple PCA

$$\max_t \frac{1}{1+t^2}(\alpha + \beta t^2 + 2\gamma t)$$

Posons $f(t) = \frac{1}{1+t^2}(\alpha + \beta t^2 + 2\gamma t)$
$f'(t) = \frac{1}{1+t^2}(2\beta t + 2\gamma) - \frac{2t}{(1+t^2)^2}(\alpha + \beta t^2 + 2\gamma t)$
$f'(t) = 0 \Leftrightarrow (1+t^2)(\beta t + \gamma) - t(\alpha + \beta t^2 + 2\gamma t) = 0$
$\Leftrightarrow \beta t + C + \beta t^3 + Ct^2 - \alpha t - \beta t^3 - 2\gamma t^2 = 0$
$\Leftrightarrow -\gamma t^2 + (\beta - \alpha)t + \gamma = 0$
$\Rightarrow$ si $\gamma = 0$, alors $t = 0$
sinon, $\Rightarrow t^2 - \frac{\beta-\alpha}{\gamma} t - 1 = 0$
$\Leftrightarrow t = \frac{\beta-\alpha}{2\gamma} \pm \sqrt{1 + (\frac{\beta-\alpha}{2\gamma})^2}$

Cas simple PCA

on cherche $u \in \mathbb{R}^2$ tel que

$$u = \max_u \sum_n \frac{(u^T x_n)^2}{u^T u}$$

en pratique si on projette les x dans la base orthonormale composée de u et de son complémentaire, on a $\sum_n x_{n,1} x_{n,2} = 0$

Cas général PCA

Soit $x_1, ..., x_N \in \mathbb{R}^D$ tel que $\forall i \neq j \ \sum_n x_{n,i} x_{n,j} = 0$ alors $\forall u$

$$\sum_n (u^T x)^2 = \sum_n \sum_{i,j} u_i u_j x_{n,i} x_{n,j}$$

$$= \sum_{i,j} u_i u_j \sum_n x_{n,i} x_{n,j}$$

$$= \sum_d u_d^2 \sum_n x_{n,d}^2$$

dans ce cas, le meilleur vecteurs u (orthogonal) est tel que $u_d = 0$ sauf pour d tel que $\sum_n x_{n,d}^2$ est maximal (1 pour ce d).

PCA

$$\max_{A \in \mathrm{R}^{D \times D} \ / \ AA^T = I} \sum_n ||A_{\{1,\dots,d\}} x||^2$$

revient à diagonaliser la matrice de covariance $C_{i,j} = \sum_n x_{n,i} x_{n,j}$ en base orthonormale (possible car c'est une matrice symétrique) !

(Ce problème est lui même lié à la notion de décomposition en valeur singulière.)

PCA

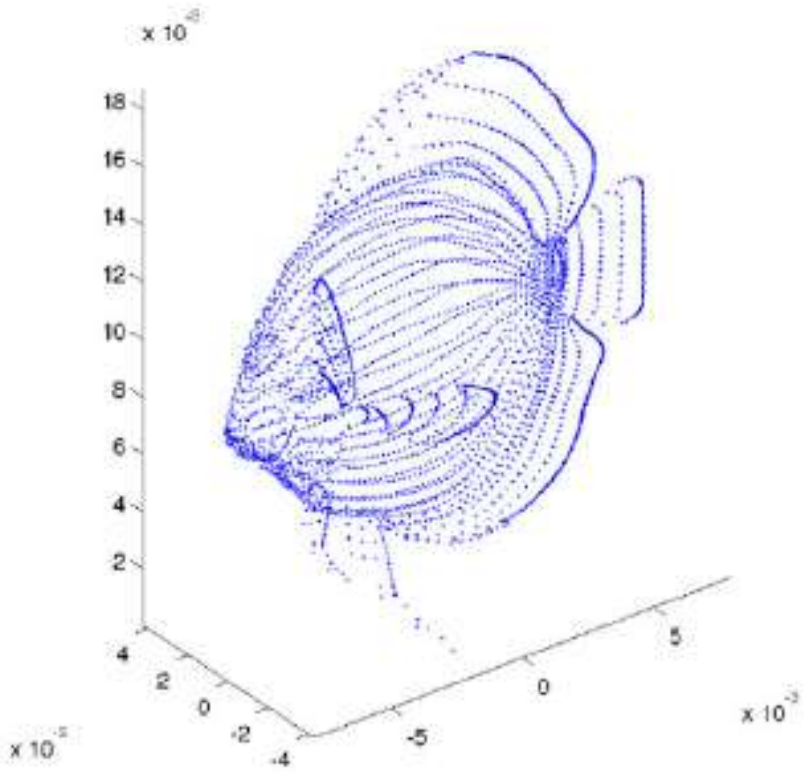
Soit $x_1, ..., x_N \in \mathbb{R}^D$ avec $\sum_n x_n = 0$,

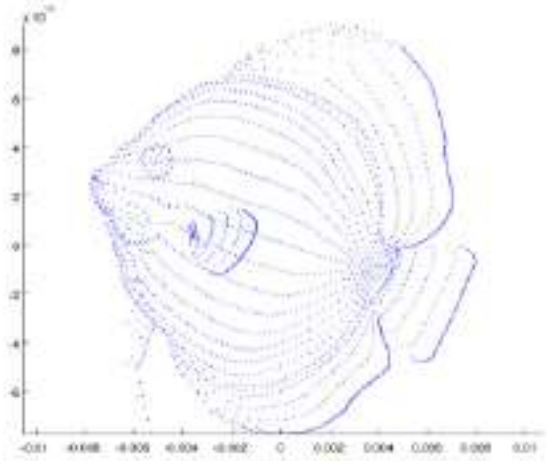
$C_{i,j} = \sum_n x_{n,i} x_{n,j}$ la matrice de covariance,

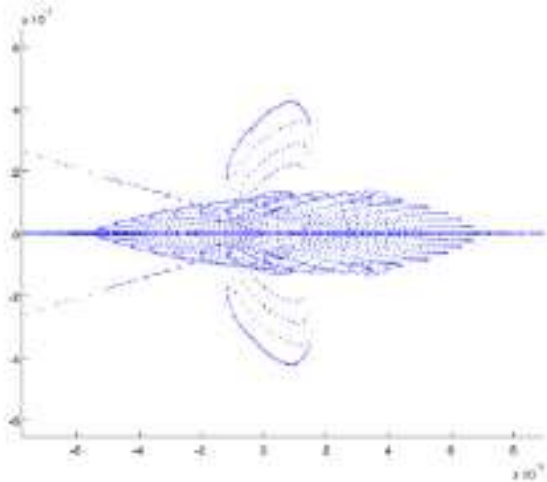
$$C = U^T \begin{pmatrix} \lambda_1 & 0 & 0 & ... & 0 \\ 0 & \lambda_2 & 0 & ... & 0 \\ ... & & & & \\ 0 & ... & 0 & 0 & \lambda_D \end{pmatrix} U \text{ avec } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$$

Les vecteurs $U_{\{1,...,\mathcal{D}\}}$ résolve le problème de la projection linéaire vers $\mathbb{R}^{\mathcal{D}}$ maximisant la variance !

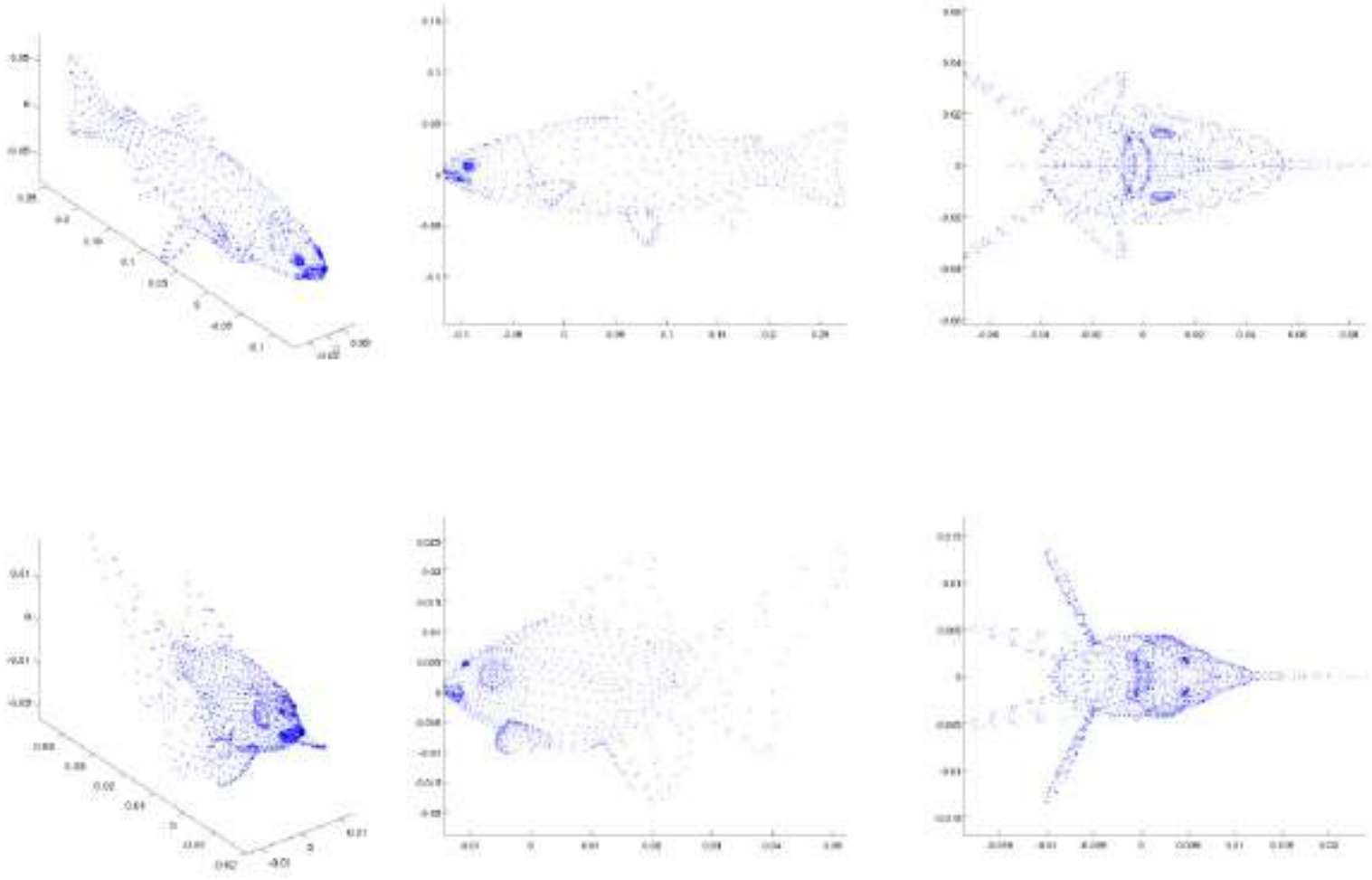
PCA







PCA

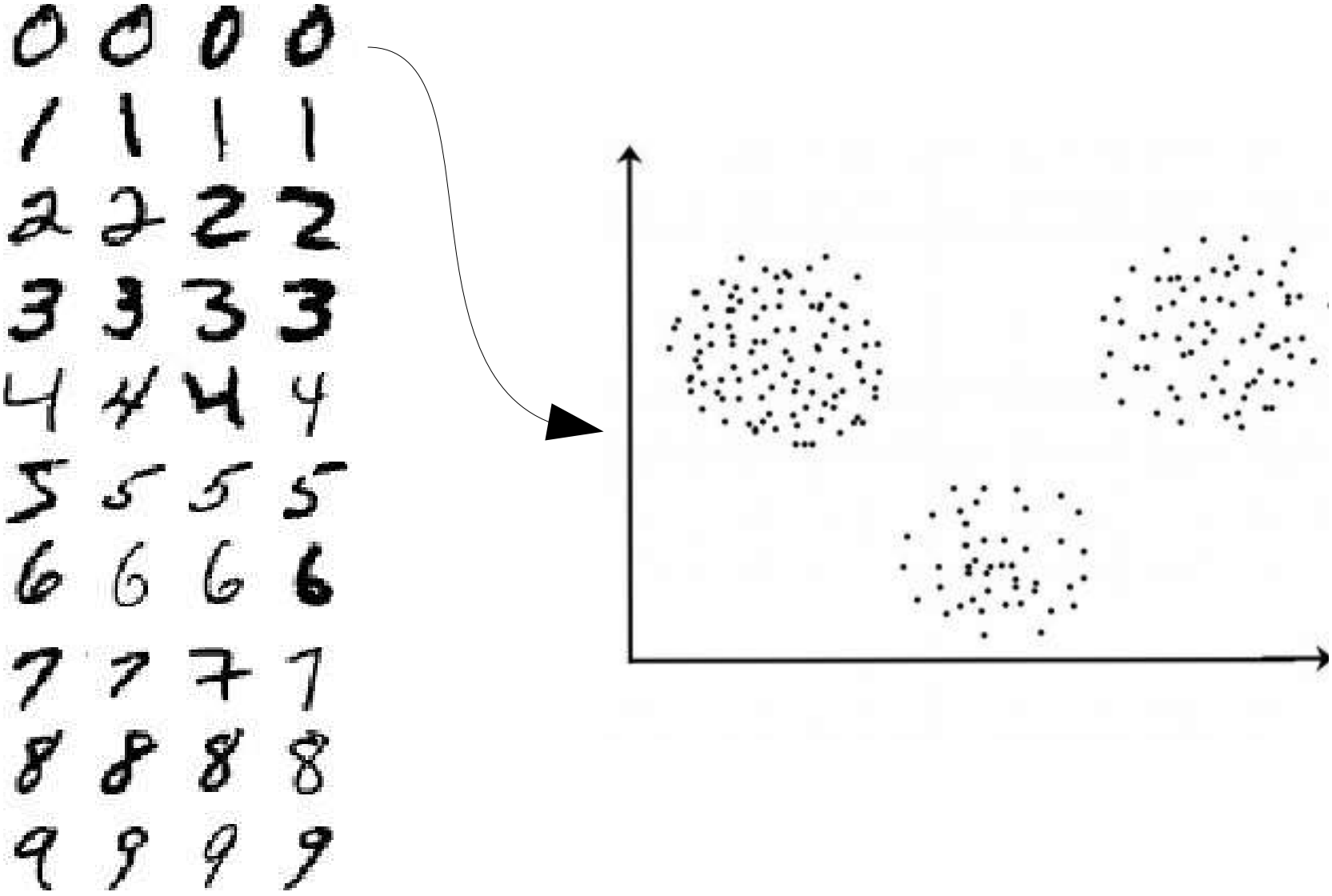


Plan du cours

- *K moyennes*
- *PCA*
- Approche Sac de mots

On a vu 2 méthodes de *compression* - sachant que la sémantique c'est une compression mais qu'une compression n'est pas toujours sémantique !
Maintenant une application à une approche d'apprentissage.

Classification brute



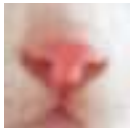
problème

En encodant la valeur brute plutôt
que les formes relative
deux images équivalentes sont différentes !

Comment obtenir des représentations
invariantes
et
discriminantes

Une idée

Un bout d’un objet permet parfois de savoir ce qu’est l’objet



Or les bouts sont plus facilement normalisables
(coder les invariances et etc)

Sac de mots

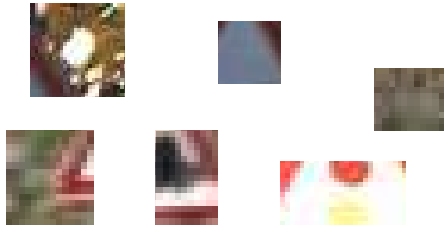
Sac de mots : phase 1 création d’un dictionnaire

On prend toutes les images disponibles (indépendamment de leur classe)

Sac de mots : phase 1 création d'un dictionnaire

On prend toutes les images disponibles (indépendamment de leur classe)

On les coupe en plein de petits bouts (de façon dense généralement)

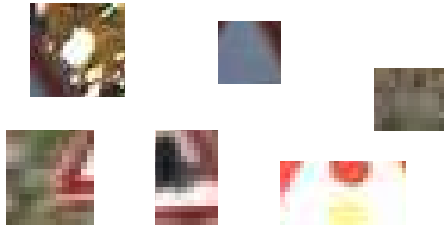


N images → K x N bouts

Sac de mots : phase 1 création d'un dictionnaire

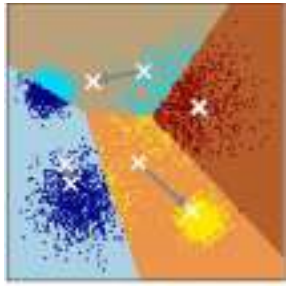
On prend toutes les images disponibles (indépendamment de leur classe)

On les coupe en plein de petits bouts (de façon dense généralement)



N images → R x N bouts

On applique les K moyennes à l'ensemble de bouts → K groupes



Avec par exemple K = 100000

Sac de mots : phase 2 encodage des images

Pour chaque image

On extrait ses R bouts

Qu’on associe au K centres

On obtient R valeurs dans {1,…,K}

Sac de mots : phase 2 encodage des images

Pour chaque image

On extrait ses R bouts

Qu'on associe au K centres

→ On obtient R valeurs dans {1,…,K}

Qu'on peut encoder dans un histogramme x de taille K
(x_k est le nombre de valeurs égales à k ; la somme des x_k fait R)

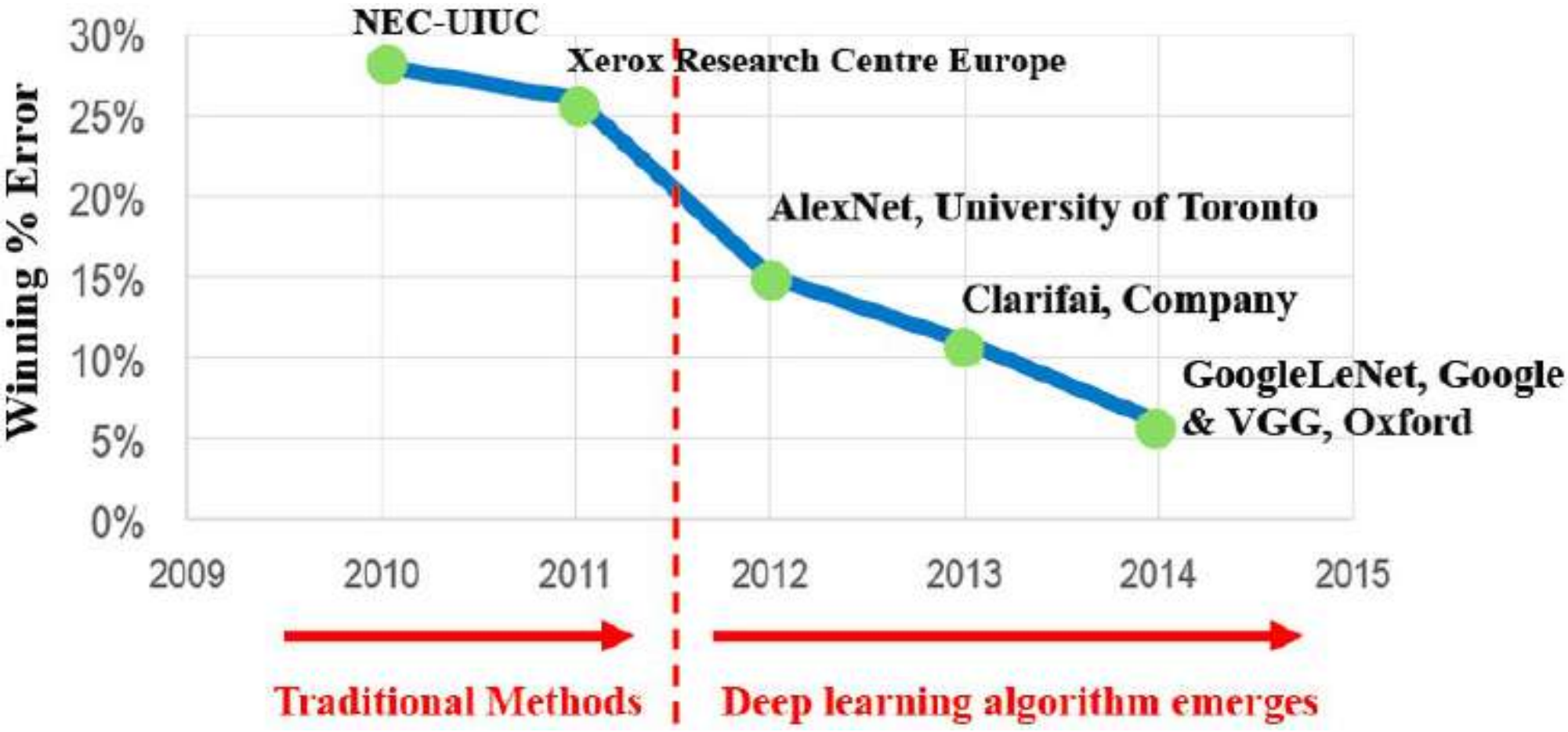
→ **On sait transformer une image en un vecteur de taille K**
relativement invariant à des translations
et relativement discriminant !
(sans utiliser les labels)

Sac de mots : phase 3 classification

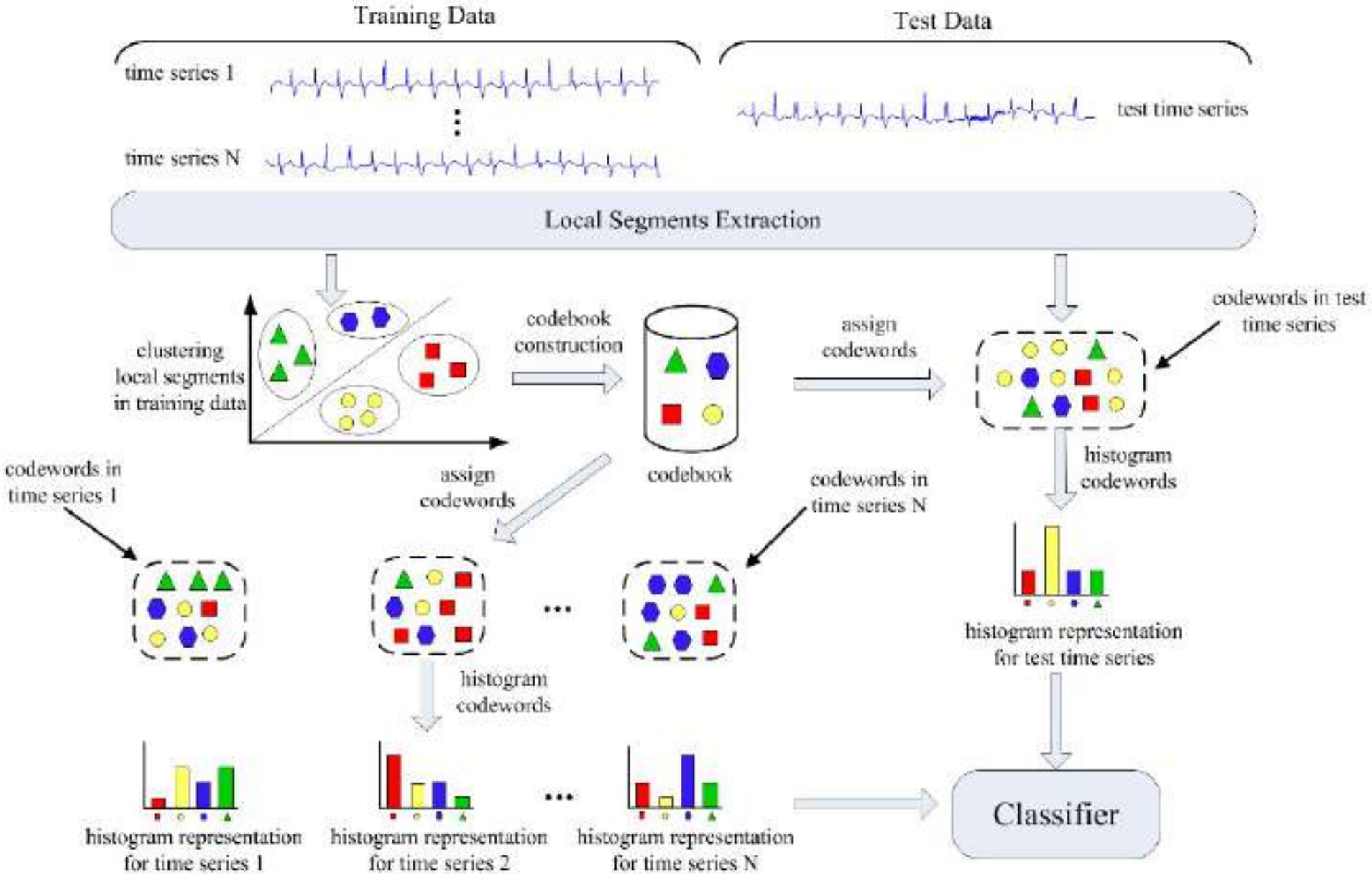
On peut alors apprendre un classifier sur ces vecteurs de dimension K

(Par exemple un MLP ou un SVM ou un arbre et non pas un CNN)

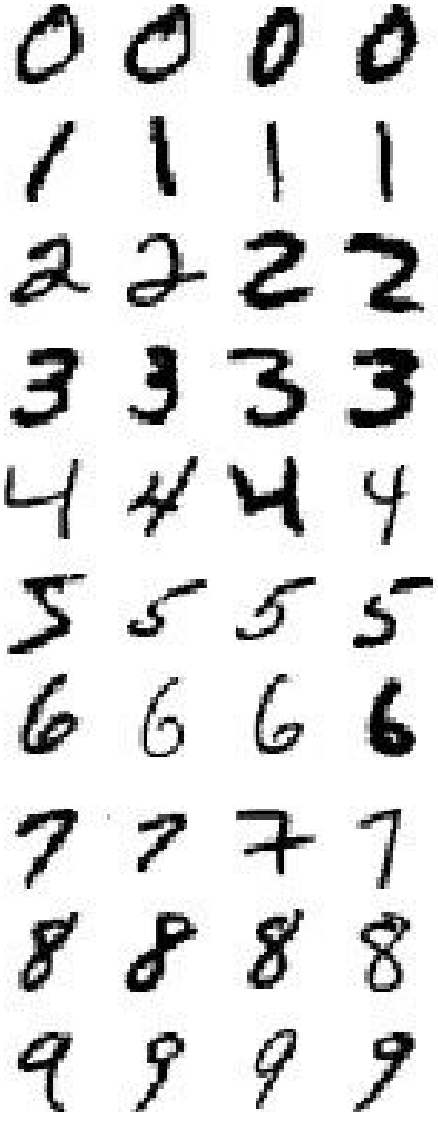
Sac de mots + MLP vs CNN



Sac de mots + MLP vs CNN

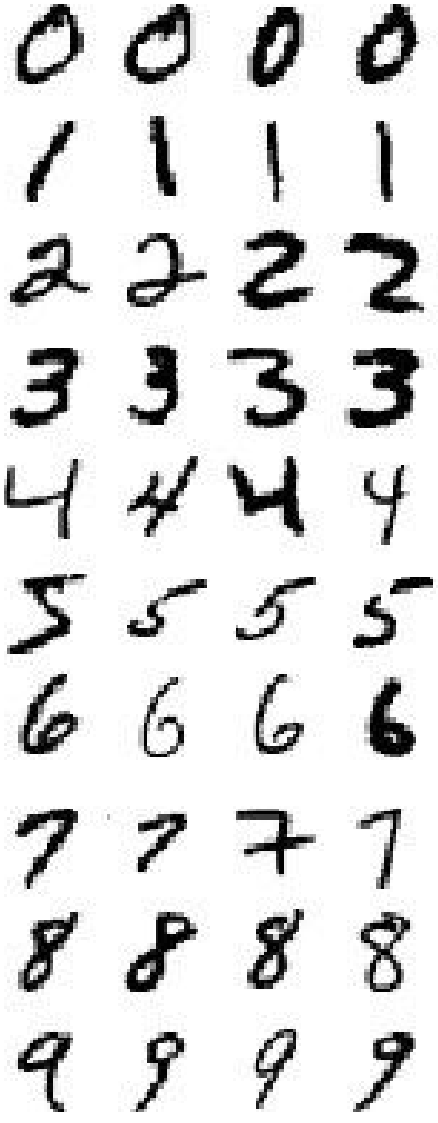


Auto-encodage, auto-supervisé : la rencontre du non supervisé et des modèles modernes

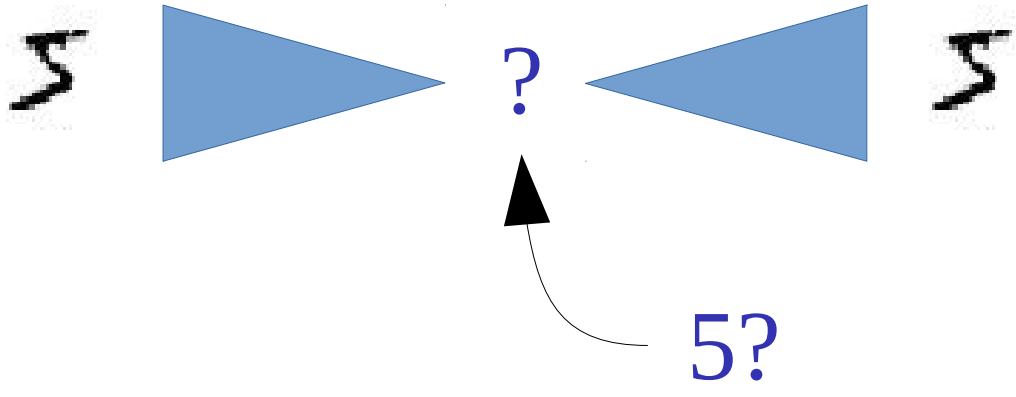


Le point de départ :
Si on devait compresser
chacune de ces images en 1 nombre
ce serait le nombre représenté

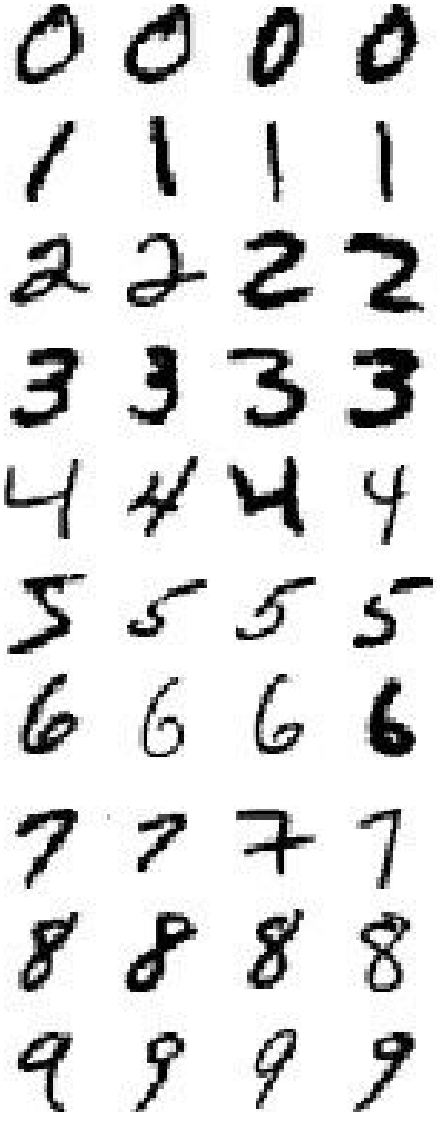
Auto-encodage, auto-supervisé :
la rencontre du non supervisé et des modèles modernes



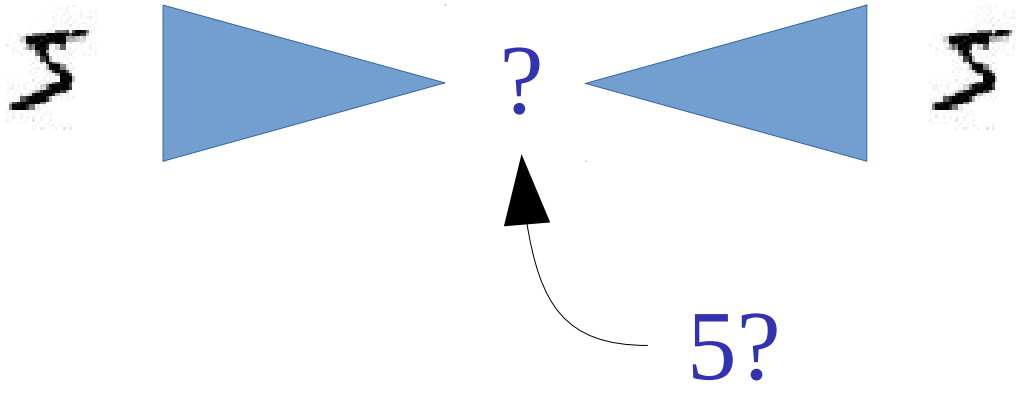
Le point de départ :
Si on devait compresser
chacune de ces images en 1 nombre
ce serait le nombre représenté



Auto-encodage, auto-supervisé : la rencontre du non supervisé et des modèles modernes



Le point de départ :
Si on devait compresser
chacune de ces images en 1 nombre
ce serait le nombre représenté



Ce type d'architecture est facile à réaliser avec des réseaux de neurones (cf segmentation)

Auto-encodage, auto-supervisé : la rencontre du non supervisé et des modèles modernes

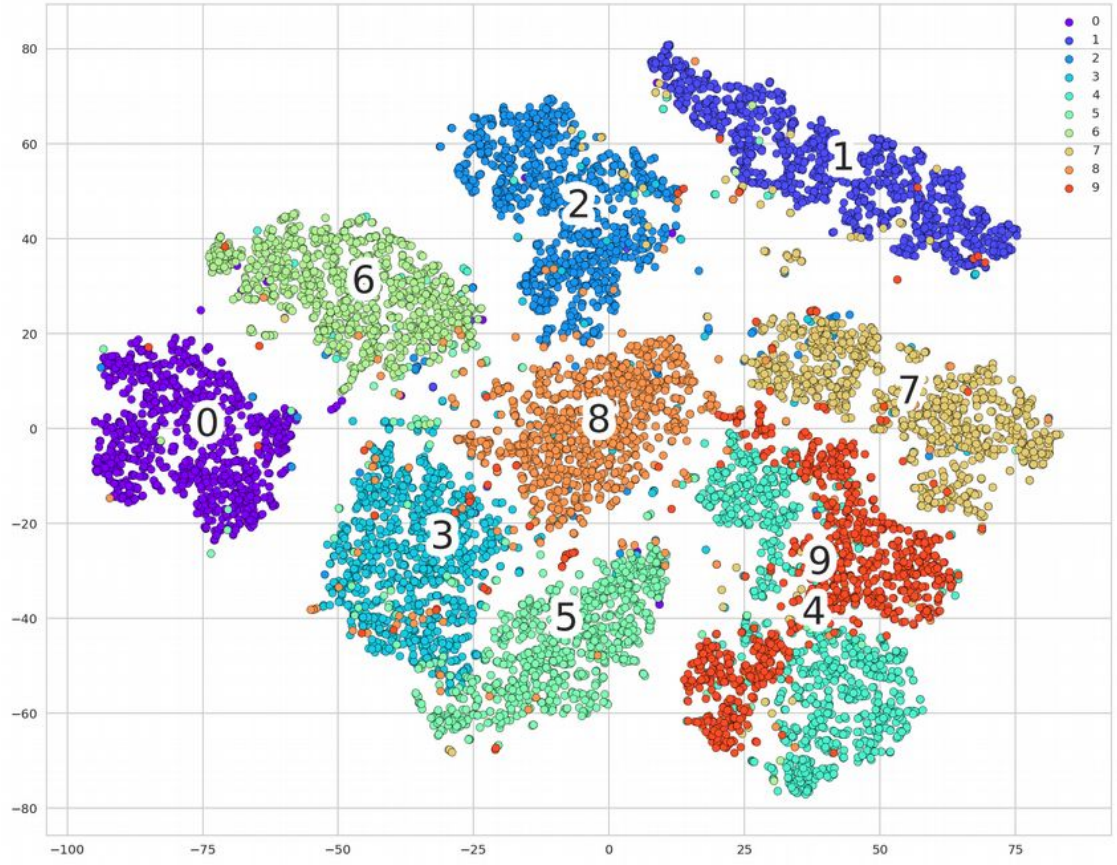
→ utiliser des réseaux de neurones pour compresser le signal

→ avec une fonction de perte associée à l'erreur de reconstruction

Auto-encodage, auto-supervisé : la rencontre du non supervisé et des modèles modernes

→ utiliser des réseaux de neurones pour compresser le signal

→ avec une fonction de perte associée à l'erreur de reconstruction



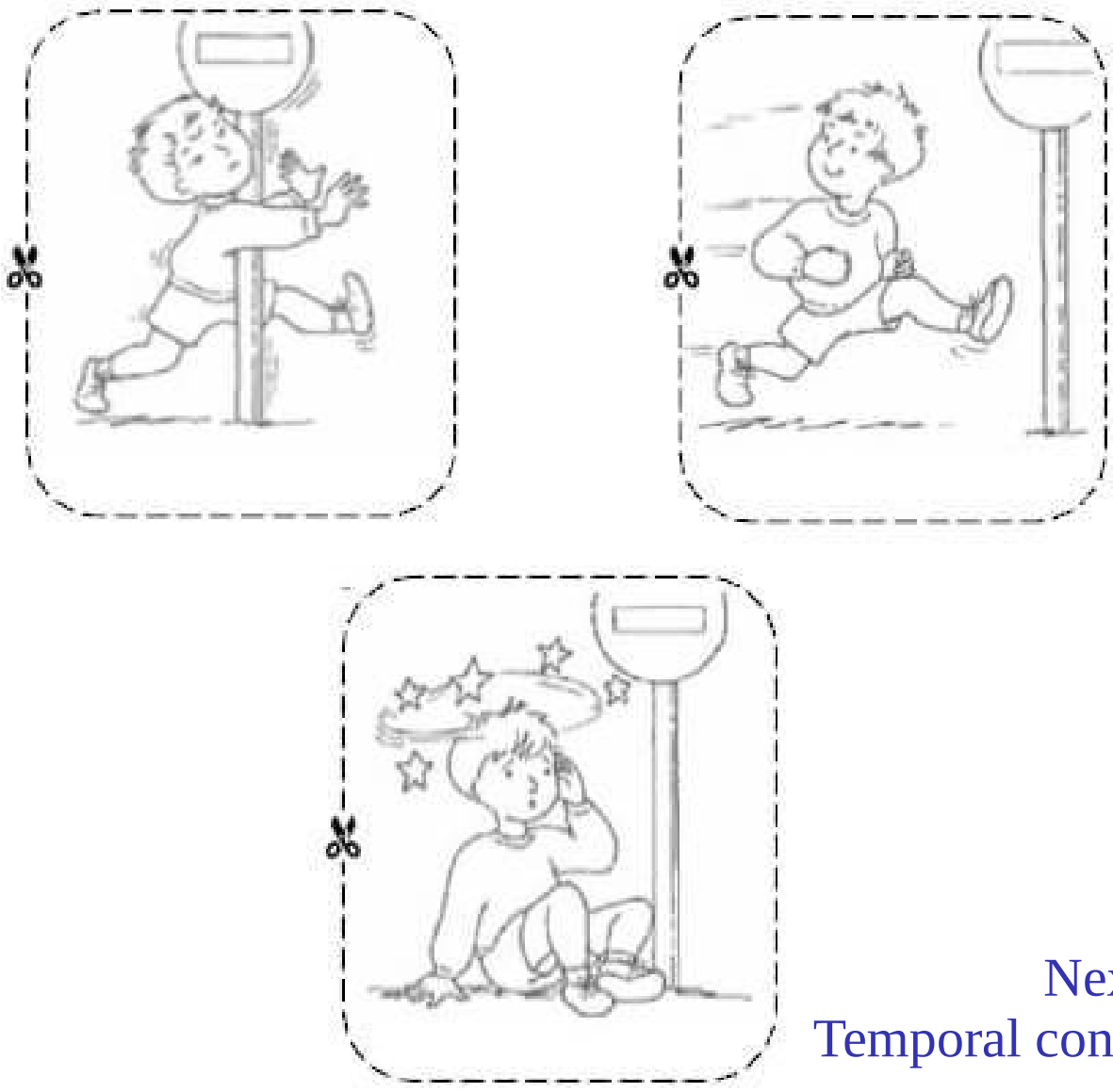
auto-supervisé :
la rencontre du non supervisé et des modèles modernes

L’auto encodage permet de pré apprendre des bouts de réseau
SANS AVEC BESOIN D’ANNOTATIONS !

→ Mais les réseaux que l’on peut apprendre ainsi ne peuvent QUE compresser

Ils ne sont pas forcément pertinent pour d’autres taches...

auto-supervisé :
la rencontre du non supervisé et des modèles modernes



Next frame prediction
Temporal consistency prediction

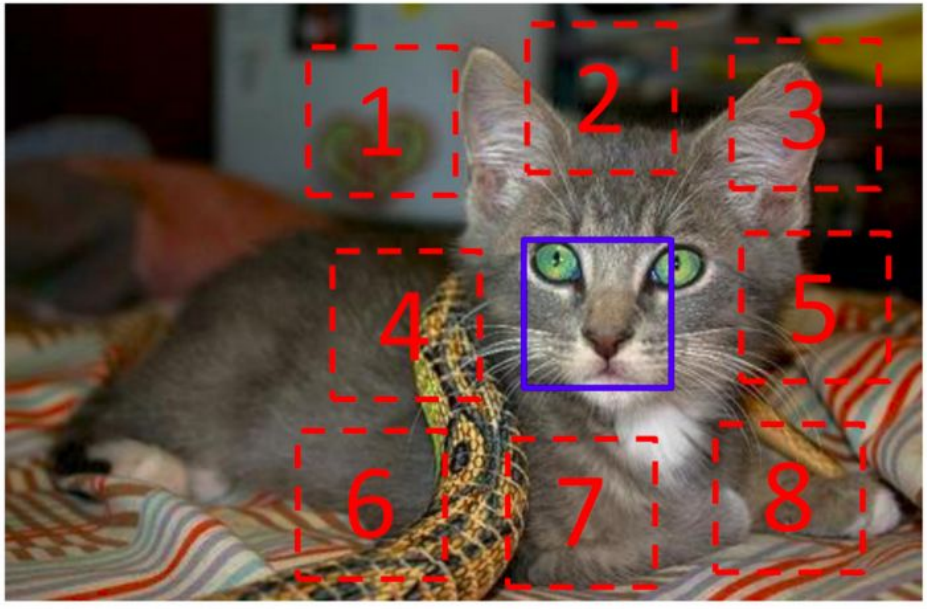
auto-supervisé :
la rencontre du non supervisé et des modèles modernes

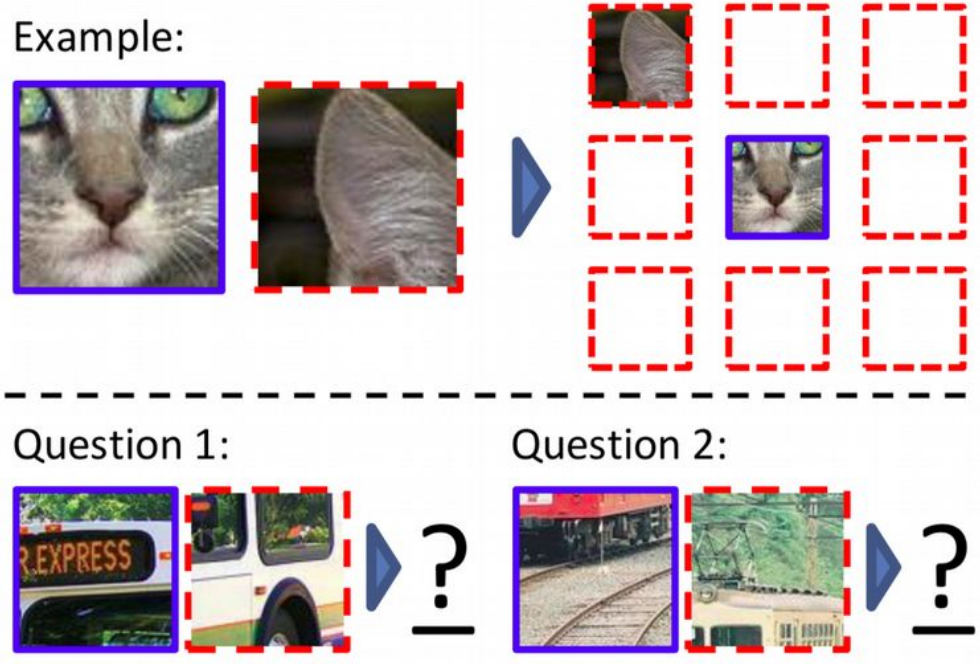
Là où l’auto encodage ne peut apprendre que des réseaux « comprimeur »

l’auto supervision peut pré-apprendre n’importe quelle type de réseau

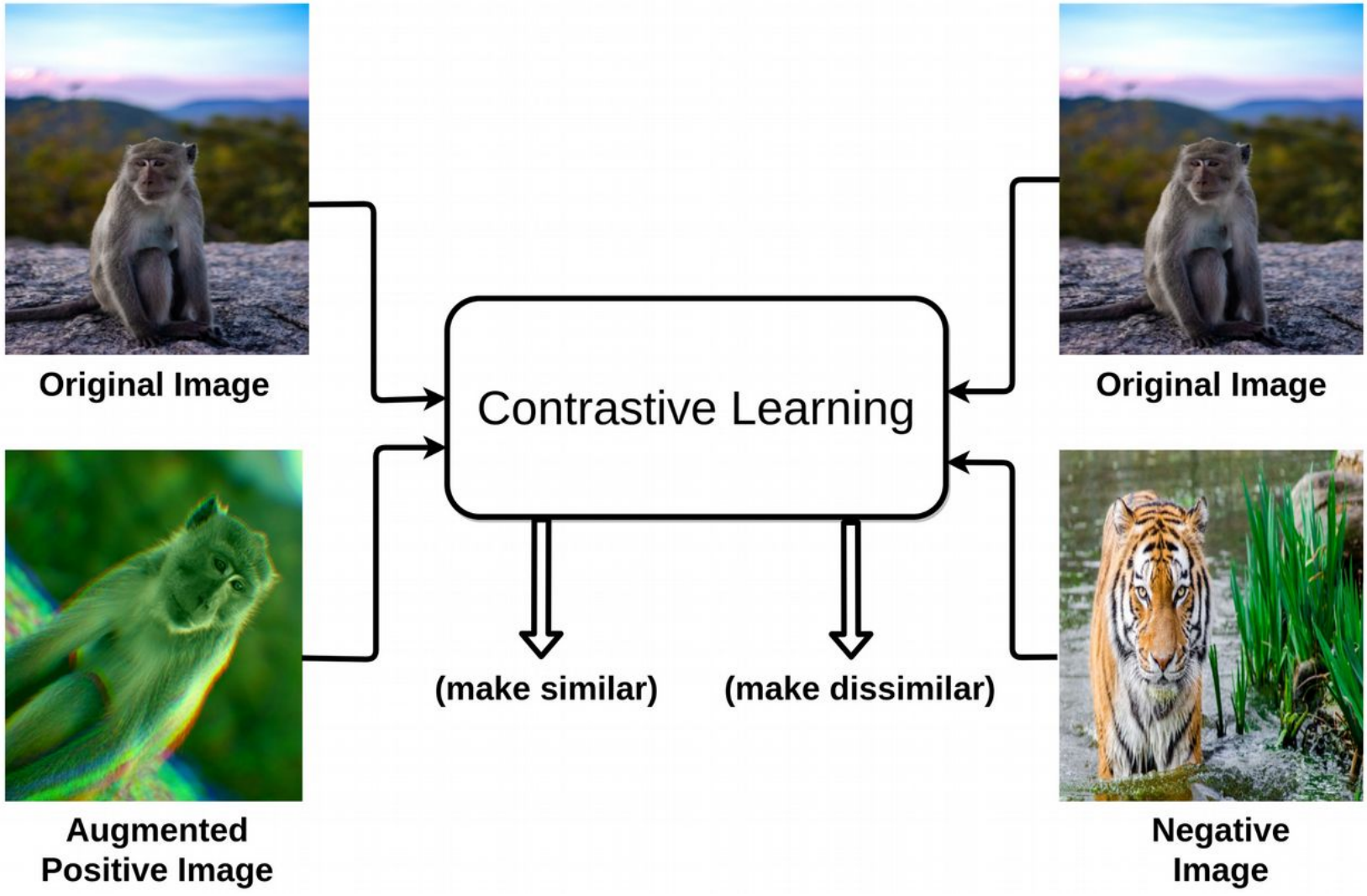
toujours sans aucun label !

auto-supervisé :
la rencontre du non supervisé et des modèles modernes





auto-supervisé :
la rencontre du non supervisé et des modèles modernes



auto-supervisé :
la rencontre du non supervisé et des modèles modernes

Le constrative learning paraît une des voies les plus prometteuse d'augmentation des performances