K-Means

Le premier **unsupervised algorithm** (il s'entraîne sur un jeu de données non labélisé).

- un cluster est un ensemble de points $\left\{x_{i_1^k}, \ldots, x_{i_{n^k}^k}\right\}$
- μ_k est le centre de masse du cluster k

Nous voulons que les distances entre les points au sein d'un cluster soient petites et que les distances entre les clusters soient larges.

$$\text{on veut minimiser : } \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(x_{i_j}^k - \mu_k\right)^2$$

Comment trouver les centres de masse ?

- on initialise les centres de masse à une position aléatoire
- jusqu'à ce que ça ne change plus
  - on assigne chaque point au centre de masse le plus proche (en calculant la distance euclidienne -- un point ne peut être associé qu'à un centre de masse)
  - on met à jour chaque μ_k en fonction de la moyenne des points associés

$\rightarrow$ ça ne marche pas toujours! on doit essayer avec plusieurs seeds (plusieurs positions aléatoires au début) et prendre celle qui à le meilleur résultat en termes de distance au carré