

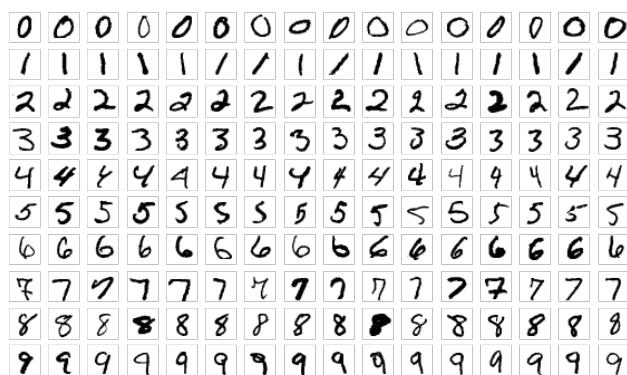
### Document 1 : algorithme des k plus proches voisins (KNN)

L'algorithme des k plus proches voisins est un algorithme d'apprentissage machine **supervisé**. Il vise à classifier une donnée, à trouver l'étiquette d'une donnée.

L'algorithme est dit supervisé car il nécessite une intervention extérieure pour réaliser sa tâche. Un ensemble de données doit en effet être étiqueté au préalable pour permettre à l'algorithme d'opérer.

Le principe de l'algorithme est de calculer d'une façon ou d'une autre la distance entre la nouvelle donnée et chacune des données d'apprentissage, il classe ensuite ces distances en ordre croissant et enfin, il attribue l'étiquette la plus donnée (le mode) parmi les k données les plus proches (d'où son nom).

Exemple d'utilisation : on donne à l'algorithme un ensemble d'images de chiffres manuscrits et on lui indique le chiffre réellement représenté (l'étiquette) sur chaque image. Charge ensuite l'algorithme de déterminer le chiffre écrit sur une nouvelle image manuscrite fournie.



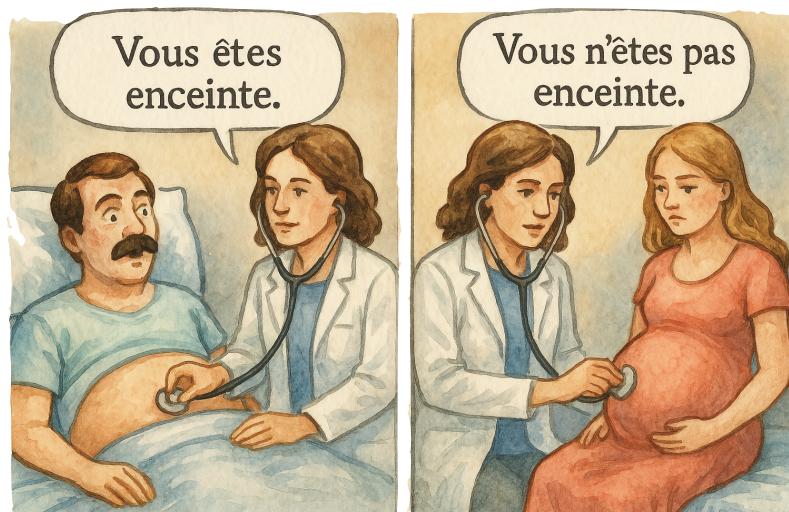
### Document 2 : précision, sensibilité et exactitude

Supposons que l'on spécialise notre algorithme à identifier les 3. Il se borne alors juste à déterminer si l'image soumise est un 3 ou non. Quatre cas de figure peuvent se présenter :

- Vrai positif (**VP**) : un 3 est reconnu comme un 3 par l'algo ;
- Faux positif (**FP**) : un autre chiffre est reconnu comme un 3 ;
- Vrai négatif (**VN**) : un autre chiffre n'est pas reconnu comme un 3 ;
- Faux négatif (**FN**) : un autre chiffre est reconnu comme un 3.

On définit alors trois grandeurs permettant d'évaluer la qualité de la prédiction :

- La **précision** : nombre de données bien prédites parmi les prédictions positives.
- La **sensibilité** : nombre de données bien prédites parmi les données positives.
- L'**exactitude** : nombre de données bien prédites parmi l'ensemble des données.



1. Quel cas de figure est illustré par chacune des images ?

Supposons que l'on connaisse les effectifs de VP, FP, VN et FN.

2. Construire la grandeur précision à partir de ces effectifs.
3. Construire la grandeur sensibilité à partir de ces effectifs.
4. Construire la grandeur exactitude à partir de ces effectifs.

Un chef de projet vous demande d'obtenir la meilleure sensibilité possible sans plus de détail.

5. Comment pourriez-vous lui fournir un algorithme très simple qui réponde à son cahier des charges trop limité.
6. Pourquoi augmenter la précision se fait souvent au détriment de la sensibilité ?

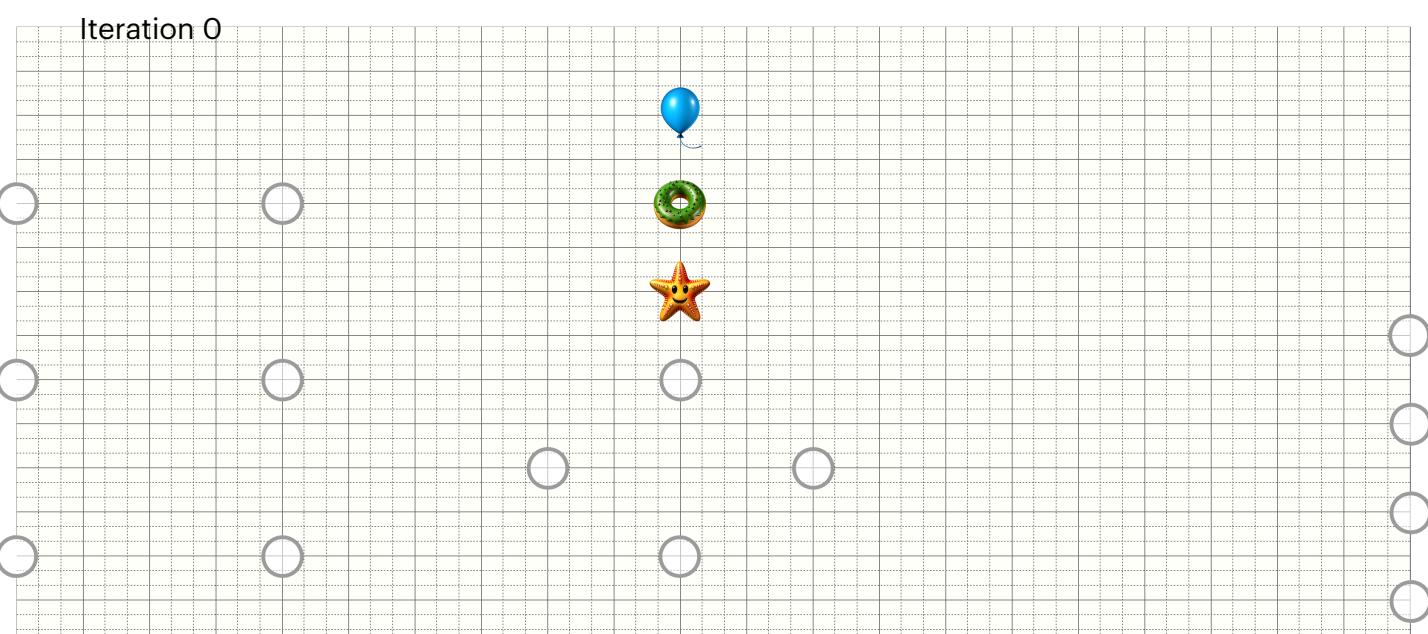
### Document 3 : l'algorithme des k-moyennes (k-means)

L'algorithme des k-moyennes est un algorithme d'apprentissage machine **non supervisé**, c'est-à-dire qu'il réalise sa tâche sans avoir besoin qu'on le guide. Et sa tâche consiste à partitionner un groupe de données, c'est-à-dire à les regrouper en k différents groupes.

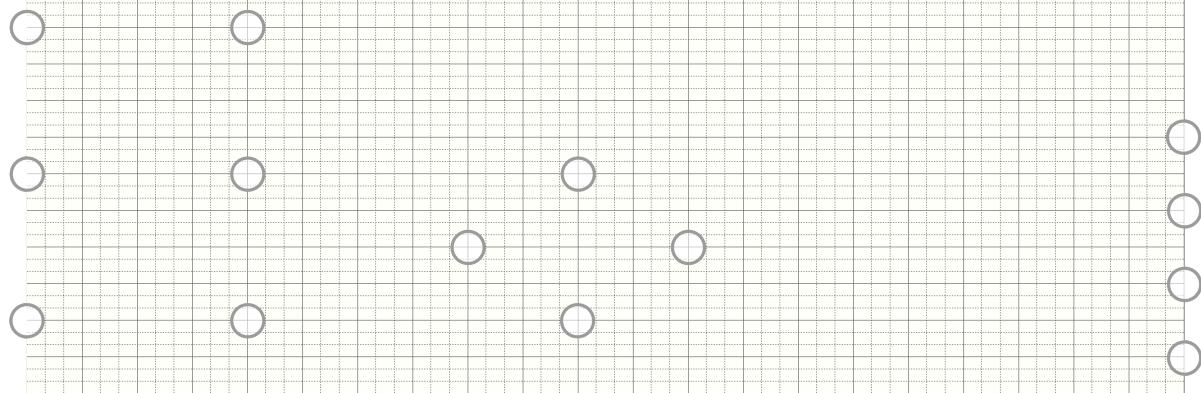
Cet algorithme est par exemple utilisé pour comprimer une image en réduisant son nombre de couleurs ou encore en marketing pour segmenter un marché en groupes de consommateurs aux habitudes similaires.

### Document 4 : déroulé de l'algorithme des k-moyennes

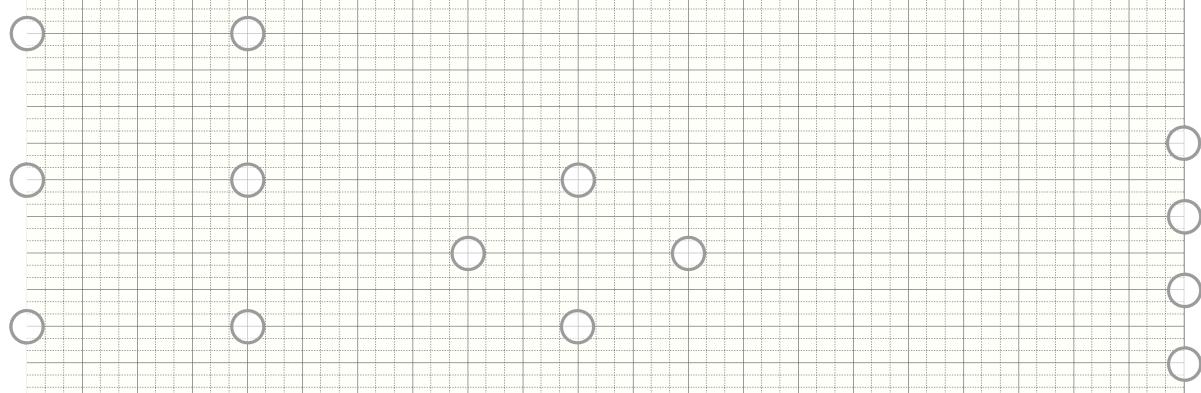
1. on choisit k centroïdes de k couleurs différentes disposés parmi les données ;
  2. on colorie les données de la couleur du centroïde le plus proche ;
  3. on déplace chaque centroïde au barycentre (centre géométrique) des données de sa couleur ;
  4. on recommence les étapes 2 et 3 tant qu'il y a du changement.
7. Vous allez appliquer l'algorithme des k-moyennes sur l'exemple ci-dessous où les 14 données sont les ronds blancs et où les 3 centroïdes sont le ballon qui colore les données en bleu, le donut qui colore les données en vert et l'étoile de mer qui les colore en jaune (ou rouge).  
L'algo est censé ici converger au bout de 4 itérations.  
Sur chacune des itérations 1, 2, 3 et 4, commencer par colorier les données de la couleur du centroïde le plus proche (en considérant les positions des centroïdes de l'itération précédente), puis déplacer les centroïdes sur leur nouvelle position.



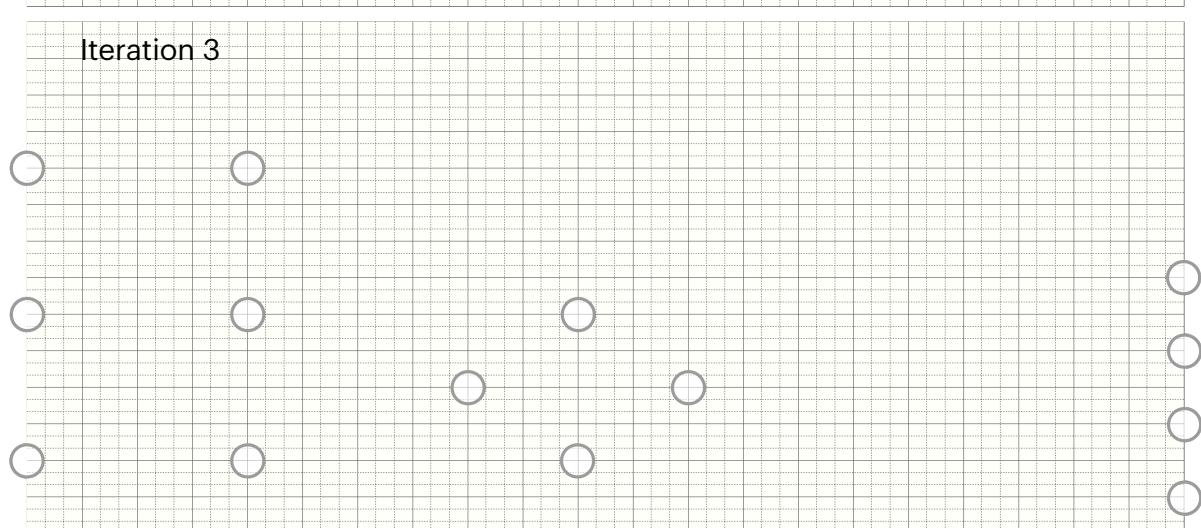
Iteration 1



Iteration 2



Iteration 3



Iteration 4

