ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

| Cours 6 |

Collection et préparation des données

AVD | Cours 6 | AMD

- SOURCES ET ACQUISITION DES DONNÉES
- NETTOYAGE ET PRÉTRAITEMENT DES DONNÉES
- TYPES DE DONNÉES
- LA MESURE

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

Clustering

LE CLUSTERING EST UNE TECHNIQUE D'APPRENTISSAGE AUTOMATIQUE ET D'ANALYSE DE DONNÉES UTILISÉE POUR REGROUPER DES POINTS DE DONNÉES SIMILAIRES EN CLUSTERS OU EN CATÉGORIES SUR LA BASE DE CERTAINES CARACTÉRISTIQUES.

L'objectif : D'identifier des modèles ou des structures au sein d'un ensemble de données, ce qui facilite la compréhension et l'analyse de données complexes.

Le clustering est une *méthode d'apprentissage non supervisée*, ce qui signifie qu'elle ne nécessite pas d'étiquettes prédéfinies pour les points de données, mais qu'elle découvre plutôt des modèles ou des similitudes inhérents aux données.

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

POINTS FORTS ESSENTIELS

- On ne connaît pas le nombre de classes (clusters). Il faut donc déterminer ce nombre.
- Chaque cluster possède certaines caractéristiques telles que le centre ou le nombre de points dans le cluster. Toutes ces caractéristiques ne seront disponibles qu'à la fin du processus.
- Il n'y a pas d'exemples ou d'autres connaissances de la structure interne des données pour aider à la construction des clusters proprement dits.
- L'objectif est de découvrir la structure interne de l'ensemble de données actuel.
- Il n'y a pas d'attribut ciblé : les points de données ne sont pas étiquetés à la fin du processus, mais les clusters obtenus peuvent être utilisés ultérieurement comme entrée d'un algorithme d'apprentissage supervisé.

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

- L'évaluation des clusters se fait généralement à l'aide de caractéristiques calculées des clusters résultants. Ces caractéristiques calculées, souvent appelées mesures d'évaluation interne, permettent d'évaluer la qualité et la structure des clusters générés par un algorithme de clustering.

- Ces métriques sont utiles lorsque vous n'avez pas accès à des étiquettes de vérité terrain et que vous devez évaluer le regroupement uniquement sur la base des données elles-mêmes.

Silhouette Score

Indice Davies-Bouldin

Inertie (somme des carrés à l'intérieur d'une classe)
« Within-Cluster Sum of Squares / WCSS »

Indice de Calinski-Harabasz (critère du rapport de variance)

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Inertie: L'inertie mesure la compacité des clusters en calculant la somme des carrés des distances entre les points de données et les centres de leurs clusters respectifs.

Remarque 1: On peut la considérer comme une fonction objective implicite qui aide à déterminer le bon nombre de centroïdes ou de clusters à inclure dans l'ensemble de données.

Remarque 2: Une inertie plus faible indique un meilleur regroupement, car elle suggère que les points de données sont plus proches de leurs centres de regroupement.

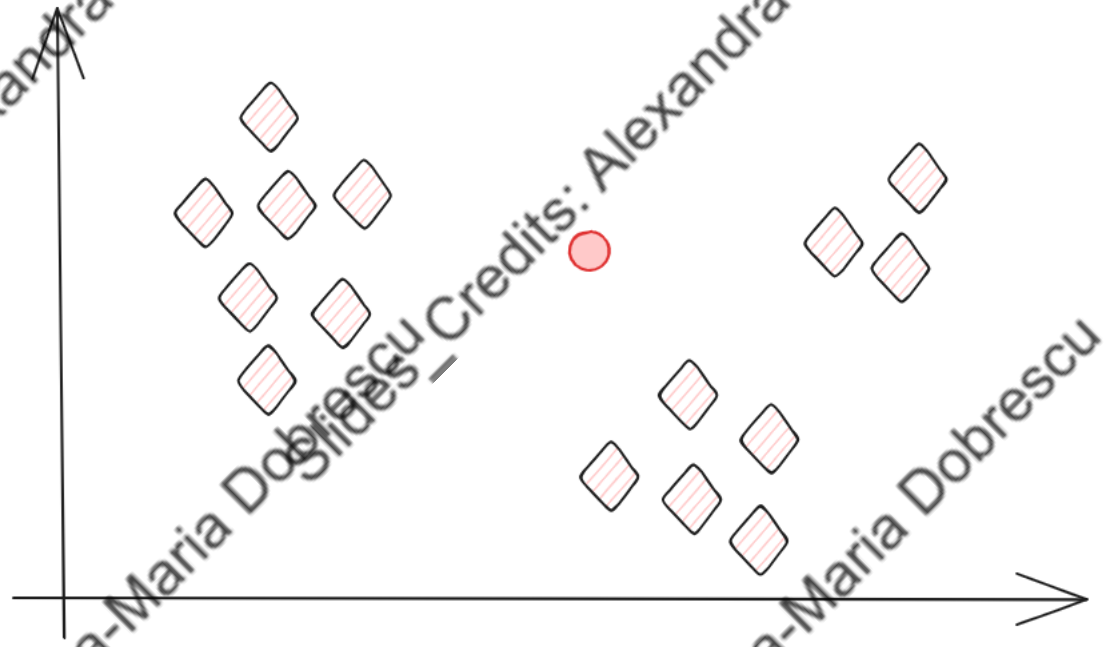
$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$

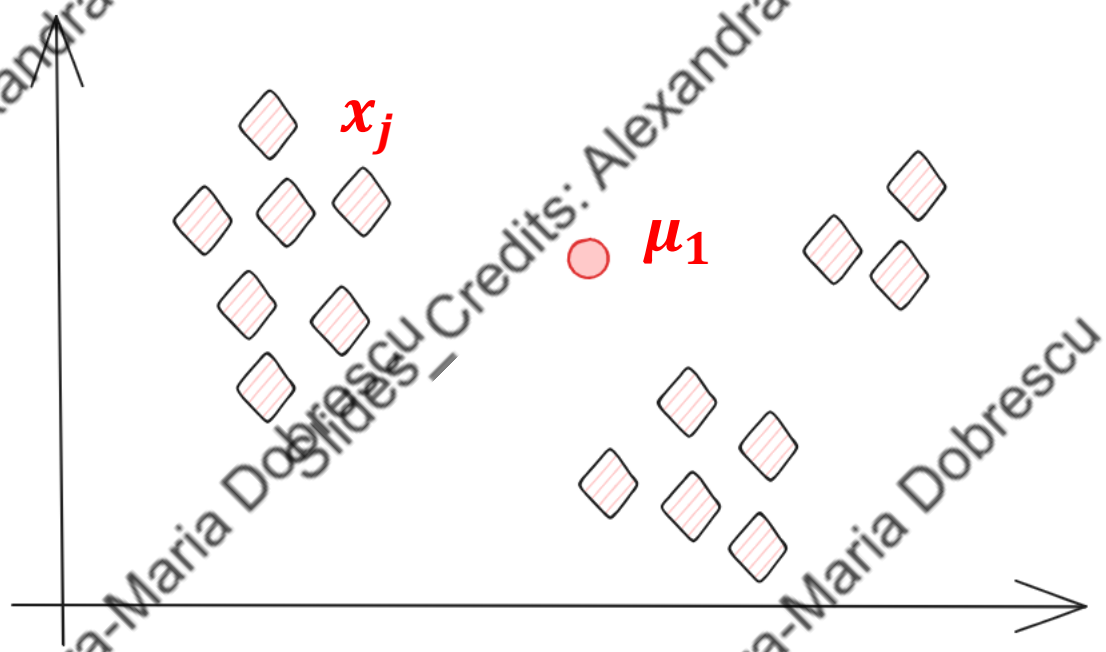


Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$



Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$


```
data_points= [ (2, 3), (3, 5), (4, 6), (1, 2), (5, 4),
               (10, 12), (11, 14), (13, 15), (7, 8), (8, 10),
               (6, 9), (9, 11)
]
```

Centroïde = (3, 4)

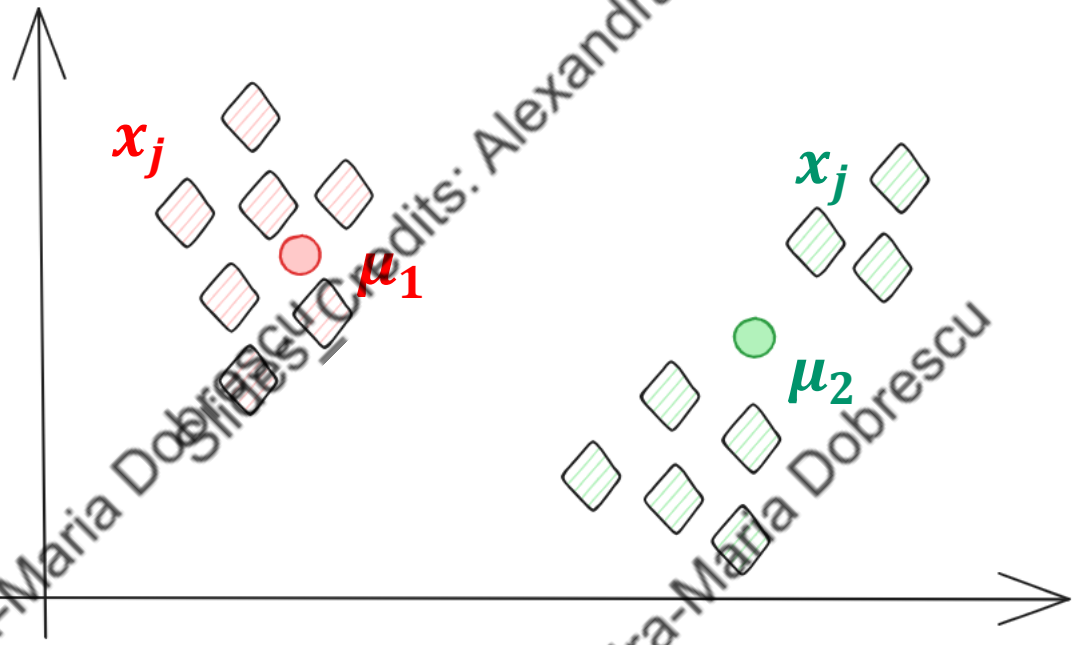
Inertie = ?

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$



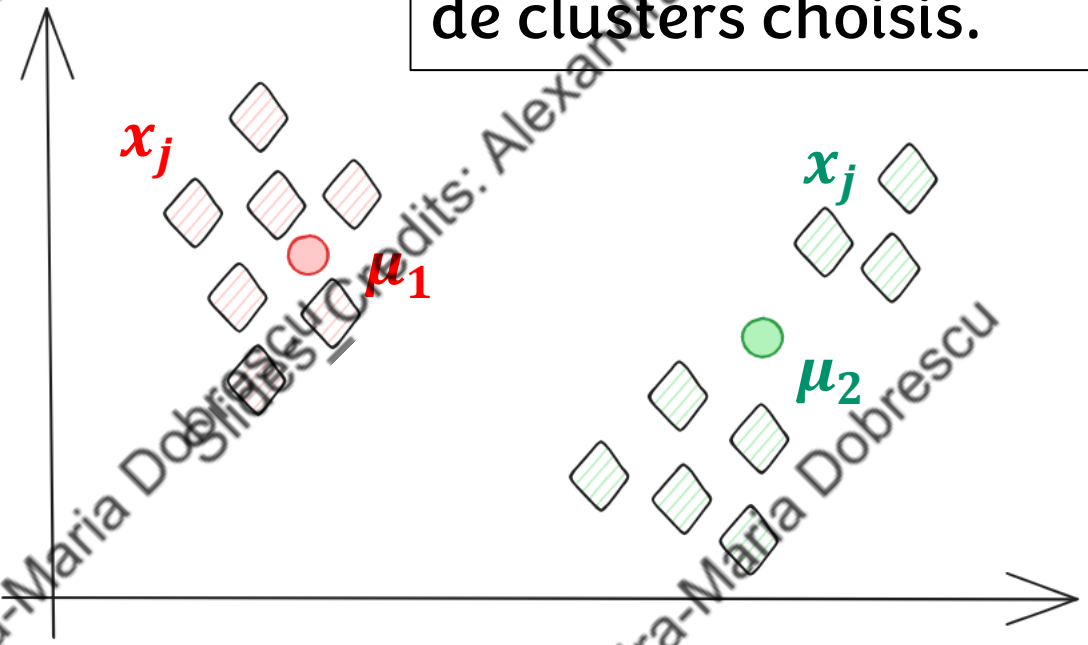
Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$

Les centroïdes varient en fonction du nombre de clusters choisis.



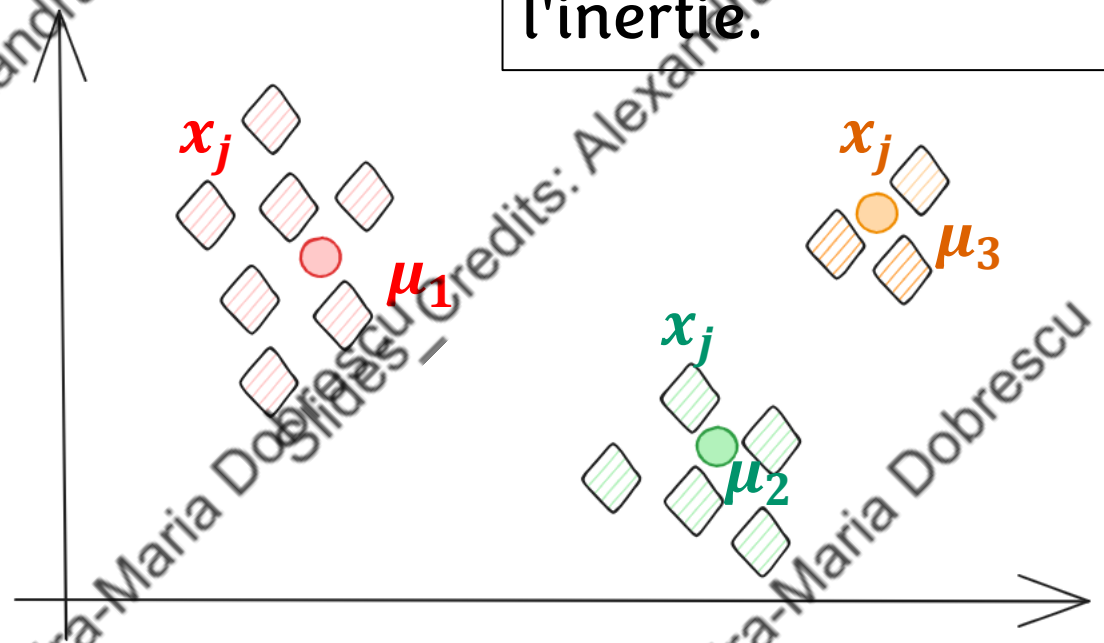
Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$

Exécutez un algorithme de clustering et calculez l'inertie.



Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

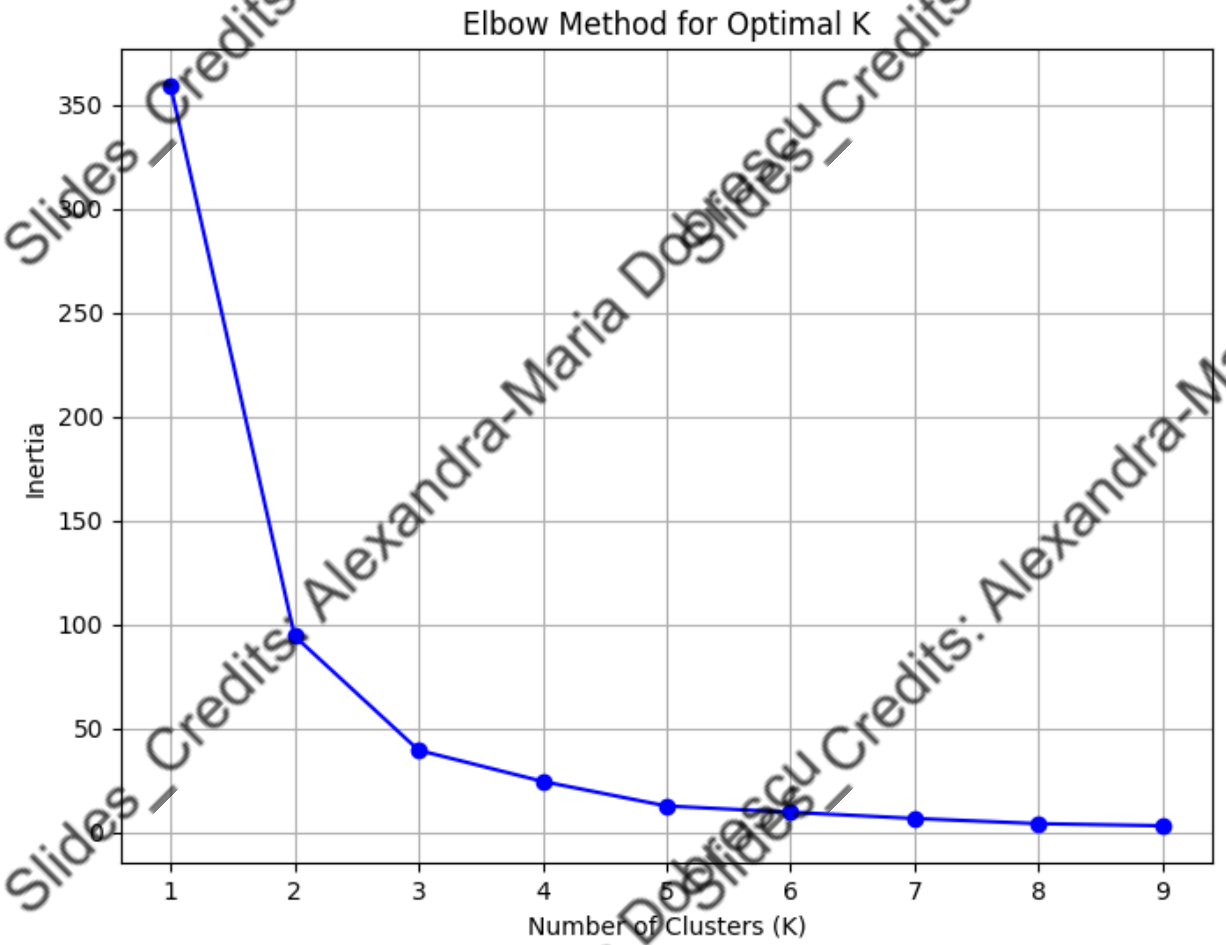
L'ÉVALUATION

$$Inertie = \sum_{i}^{k} \sum_{x_j \in c_i} \|x_j - \mu_i\|^2$$

Vous pouvez répéter les calculs ci-dessus pour chaque valeur des clusters K afin d'obtenir l'inertie correspondante.

Mais jusqu'à quand ?

Slides_Credits: Alexandra-Maria Dobrescu



La tendance montrera qu'une augmentation de K entraîne une diminution de l'inertie, mais le taux de diminution peut commencer à ralentir après un certain point, ce qui vous aidera à identifier le point d'inflexion dans la méthode du coude.

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Silhouette Score

Pour chaque point *D* d'un cluster, une valeur de « silhouette » peut être calculée, et cette valeur est en même temps :

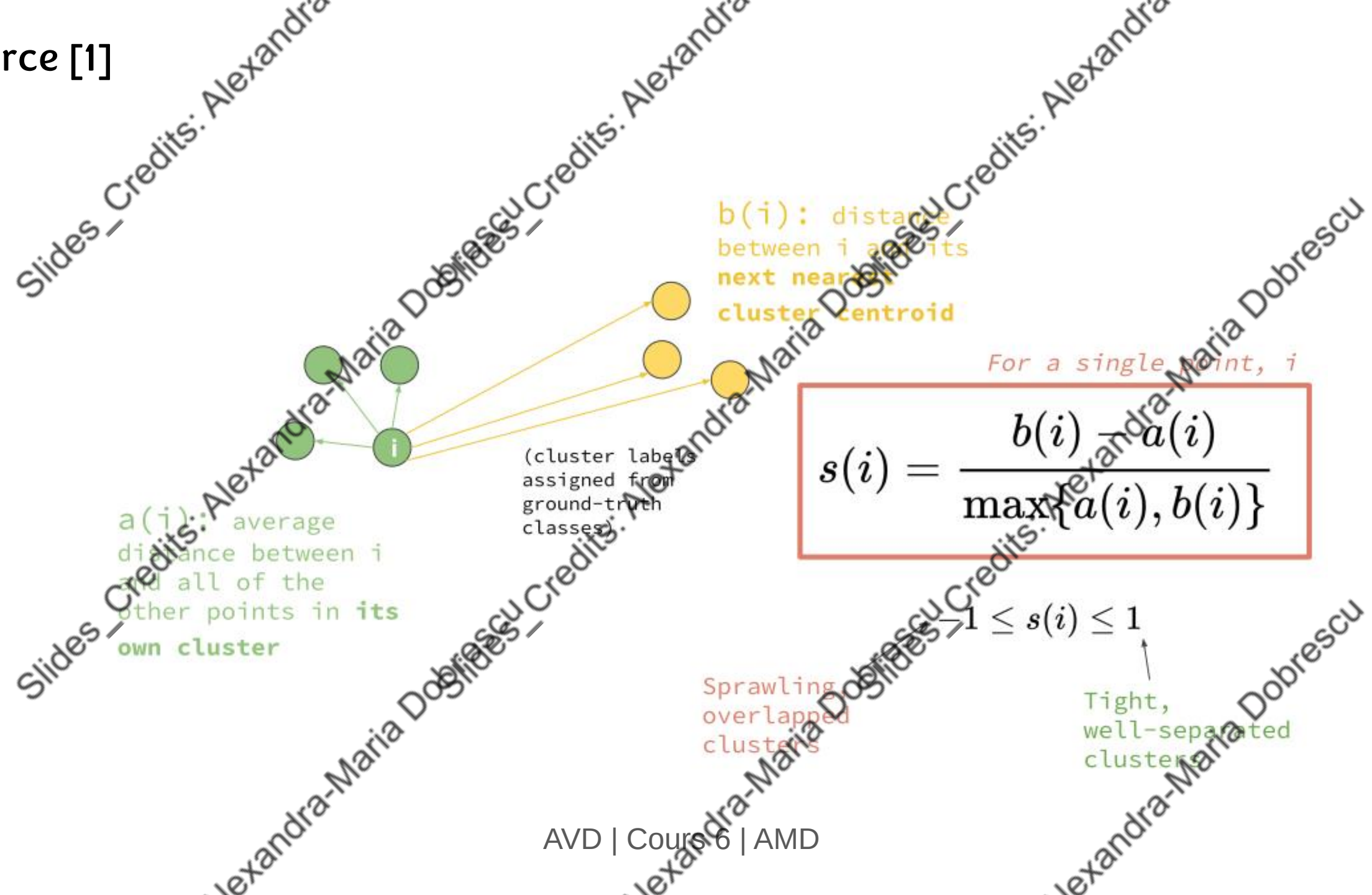
- une mesure de la similarité de *D* avec les points de son cluster,
- une mesure de la dissimilarité de *D* avec les points des autres clusters.

Les valeurs sont comprises entre -1 et 1:

- les valeurs positives indiquent que *D* est similaire aux points de son cluster,
- les valeurs négatives que *D* n'est pas bien assigné (il ferait mieux d'être assigné à un autre cluster).

Slides_Credits: Alexandra-Maria Dobrescu

Source [1]



Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Silhouette Score

La valeur moyenne de *s(i)* pour les points d'un cluster est une mesure de la cohésion des points du cluster.

Remarque 1: La valeur moyenne de s pour tous les points de l'ensemble de données est une mesure de la performance du processus de clustering.

Remarque 2: L'examen des silhouettes des clusters permet de déterminer la meilleure valeur pour ***k***. Par exemple, Pour le *K-Means*, si ***k*** est trop grand ou trop petit, certaines clusters ont des silhouettes plus étroites que les autres.

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Indice Davies-Bouldin (DBI): Une mesure permettant d'évaluer la séparation et la compacité des clusters.

Remarque: Il est basé sur l'idée que les bonnes grappes sont celles qui présentent une faible variation à l'intérieur du cluster et une forte séparation entre les clusters.

Le DBI est calculé comme la moyenne du rapport maximal entre la distance intra-groupe et la distance inter-groupes pour chaque groupe. Plus le DBI est faible, meilleure est la qualité du regroupement.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(x_i) + \Delta(x_j)}{\delta(x_i, x_j)}\right)$$

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Indice Davies-Bouldin (DBI): Une mesure permettant d'évaluer la séparation et la compacité des clusters.

Remarque: Il est basé sur l'idée que les bonnes grappes sont celles qui présentent une faible variation à l'intérieur du cluster et une forte séparation entre les clusters.

Le DBI est calculé comme la moyenne du rapport maximal entre la distance **intra-groupe ou inter-cluster** et la distance inter-groupes pour chaque groupe. Plus le DBI est faible, meilleure est la qualité du regroupement.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(x_i) + \Delta(x_j)}{\delta(x_i, x_j)}\right)$$

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

Indice Davies-Bouldin (DBI): Une mesure permettant d'évaluer la séparation et la compacité des clusters.

Remarque: Il est basé sur l'idée que les bonnes grappes sont celles qui présentent une faible variation à l'intérieur du cluster et une forte séparation entre les clusters.

Le DBI est calculé comme la moyenne du rapport maximal entre la distance intra-groupe et la distance **inter-groupes ou inter-clusters** pour chaque groupe. Plus le DBI est faible, meilleure est la qualité du regroupement.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(x_i) + \Delta(x_j)}{\boldsymbol{\delta(x_i, x_j)}}\right)$$

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

L'indice de Davies-Bouldin est très efficace par rapport à d'autres mesures d'évaluation des clusters pour les raisons suivantes :

- Il est flexible et fonctionne pour n'importe quel nombre de clusters.
- Il ne fait aucune hypothèse sur la forme des clusters, contrairement à la mesure d'évaluation du score de Silhouette.
- Il est facile à utiliser et intuitif.

L'indice de Davies-Bouldin (IDB) n'est cependant pas sans inconvénients:

- Il peut être sensible aux « outliers » et au « noise », ce qui donne une fausse indication d'un mauvais clustering.
- Il suppose une forme sphérique avec des tailles et des densités similaires pour chaque cluster, ce qui peut ne pas être vrai dans de nombreux cas réels.
- Il ne tient pas compte de la structure ou de la distribution des données, telles que les clusters à l'intérieur des clusters ou les relations non linéaires.

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

L'indice de Calinski-Harabasz (CH): Il évalue le rapport entre la variance entre les clusters et la variance à l'intérieur des clusters.

Remarque: Un CH élevé signifie un meilleur regroupement puisque les observations dans chaque cluster sont plus proches les unes des autres (plus denses), tandis que les clusters eux-mêmes sont plus éloignés les uns des autres (bien séparés).

CH ou Critère du Rapport de Variance est calculé comme un rapport entre la somme de la dispersion inter-groupes et la somme de la dispersion intra-groupes pour tous les groupes (où la dispersion est la somme des distances au carré).

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

CH ou Critère du Rapport de Variance est calculé comme un rapport entre la somme de la dispersion **inter-groupes** et la somme de la dispersion intra-groupes pour tous les groupes (où la dispersion est la somme des distances au carré).

$$BGSS = \sum_{k=1}^{K} n_k \times ||C_k - C||^2$$

nombre d'observations dans le cluster k → n_k
le centroïde du cluster k → C_k
le centroïde de l'ensemble de données → C
nombre des clusters → k

Slides_Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

CH ou Critère du Rapport de Variance est calculé comme un rapport entre la somme de la dispersion inter-groupes et la somme de la dispersion **intra-groupes** pour tous les groupes (où la dispersion est la somme des distances au carré).
Pour chaque cluster on a:

$$WGSS_k = \sum_{i=1}^{n_k} ||X_{ik} - C_k||^2$$

nombre d'observations dans le cluster *k* → n_k
le centroïde du cluster *k* → C_k
la *i*-ième observation du cluster *k* → X_{ik}
nombre des clusters → *k*

Slides Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

CH ou Critère du Rapport de Variance est calculé comme un rapport entre la somme de la dispersion inter-groupes et **la somme de la dispersion intra-groupes** pour tous les groupes (où la dispersion est la somme des distances au carré).
Pour chaque cluster on a:

$$WGSS_k = \sum_{i=1}^{n_k} ||X_{ik} - C_k||^2$$

nombre d'observations dans le cluster *k* → n_k
le centroïde du cluster *k* → C_k
la *i*-ième observation du cluster *k* → X_{ik}
nombre des clusters → *k*

Donc, la somme finale devient:

$$WGSS = \sum_{k=1}^{K} WGSS_k$$

Slides Credits: Alexandra-Maria Dobrescu

L'apprentissage non supervisée

L'ÉVALUATION

CH ou Critère du Rapport de Variance est calculé comme un rapport entre la somme de la dispersion inter-groupes et la somme de la dispersion intra-groupes pour tous les groupes (où la dispersion est la somme des distances au carré).
La formule finale devient:

$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1}$$

dispersion entre les groupes → *BGSS*
dispersion à l'intérieur d'un groupe → *WGSS*
nombre total d'observations → *N*
nombre total des clusters → *k*

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

Clustering

Tout algorithme de clustering a la structure générique suivante:

Entrée:

a) Un ensemble de *n* objets (généralement des points: $D = \{d_1, d_2, \ldots d_n\}$). Les objets ne sont pas étiquetés et aucun ensemble d'étiquettes de classe n'est pas défini.
b) Une fonction de distance qui peut être utilisée pour calculer la distance entre deux points. Example: La mesure de dissimilarité, pour laquelle une distance de faible valeur signifie « proche », une distance de taille grande signifie « éloigné ».
c) Normalement, les points devraient également être accompagnés de leurs coordonnées dans l'espace dimensionnel où ils sont définis.
d) Une valeur prédéfinie pour le nombre de clusters dont on abesoin à la fin.

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

Clustering

Tout algorithme de clustering a la structure générique suivante:

Entrée:

Obs: Les coordonnées d'un point peuvent être considérées comme des valeurs d'attribut. Pourquoi ? Chaque dimension détermine un attribut pour l'ensemble des points.

Les fonctions de distance

dans l'espace 2D

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

Clustering

Tout algorithme de clustering a la structure générique suivante:

Sortie:

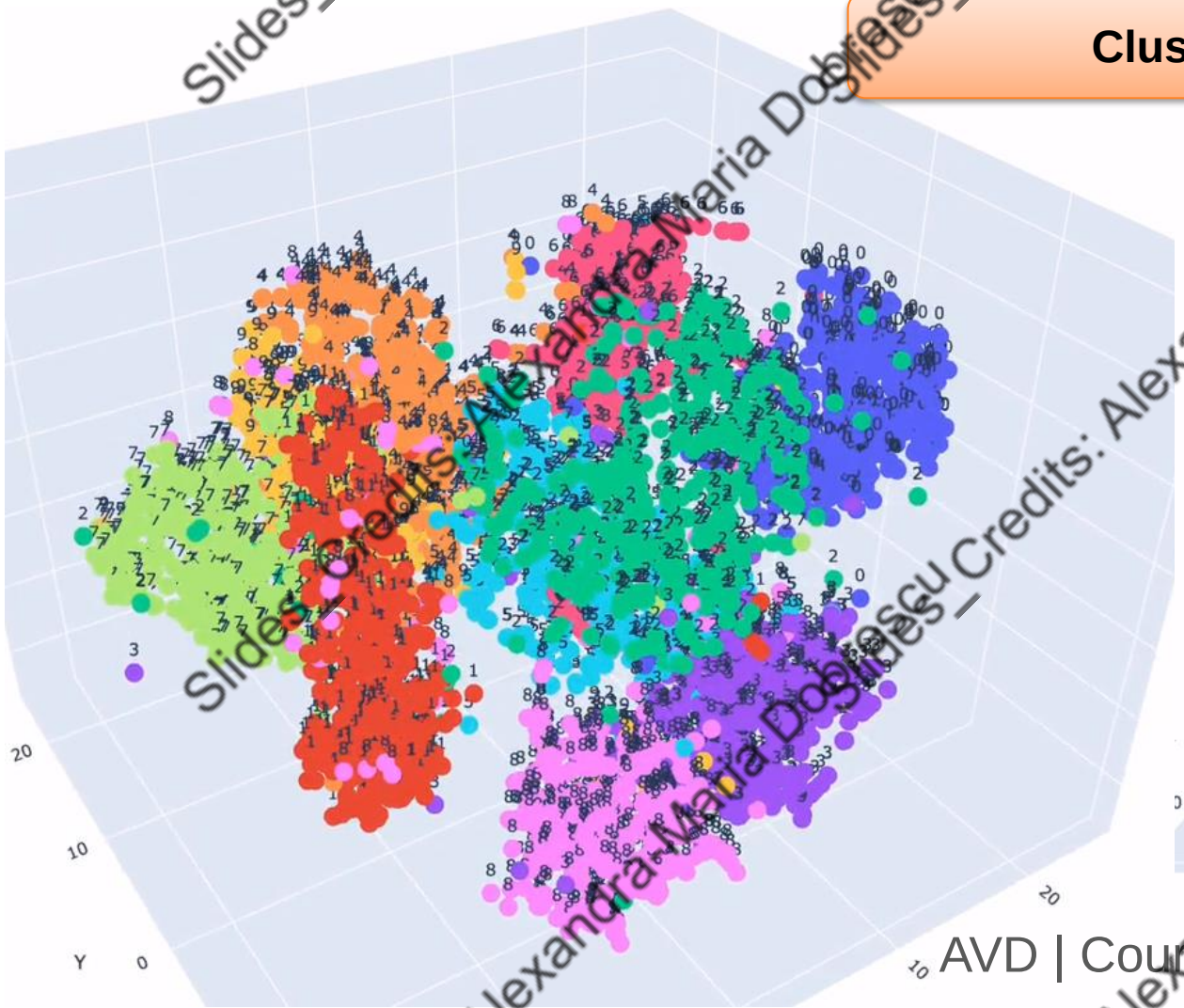
Un ensemble de groupes d'objets (points) appelés **clusters** où les points d'un même cluster sont proches les uns des autres. Les points de clusters différents sont éloignés les uns des autres, en tenant compte de la fonction de distance.

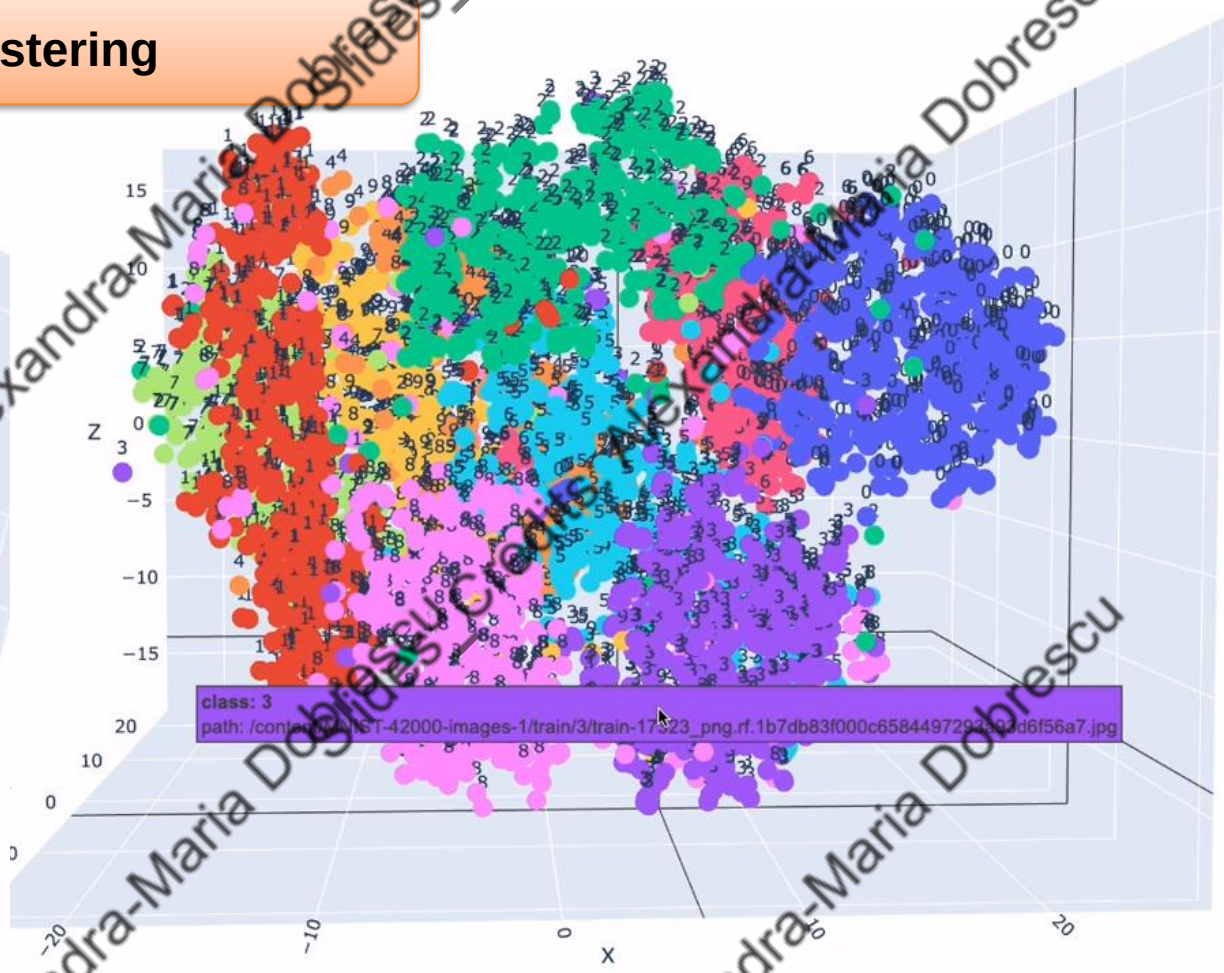
Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

Clustering





Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

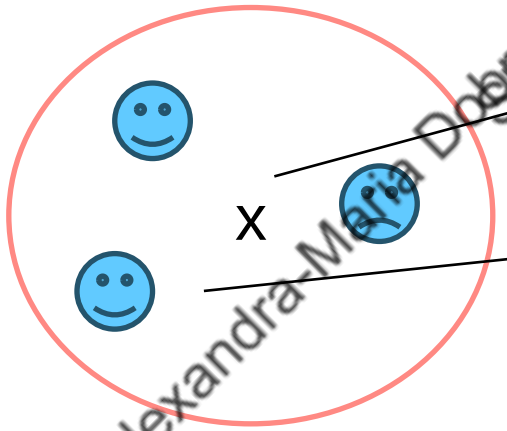
Clustering

Slides_Credits: Alexandra-Maria Dobrescu

Chaque cluster dispose des:

Caractéristiques:

a) Centroïde: le centre Euclidien, calculé comme un centre de masse des points (également pondérés). Si le cluster ne se trouve pas dans un espace euclidien, on utilise comme centre d'un groupe, le clustroïde (ou médioïde).



Le centroïde est la moyenne de tous les points de données du cluster, c-à-d, un *Point Artificiel*

Le clustroïde est un point existant qui est le plus proche de tous les autres points du cluster.

Discrétisation des données

ALGORITHMES

Clustering

Chaque cluster dispose des:

Caractéristiques:

b) Rayon: la distance maximale entre le centroïde et les points du cluster.
c) Diamètre: distance maximale entre deux points d'un cluster. Ne le considérez pas comme le double du rayon.

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

K-Means

Pseudocode K-Means

Spécifier le nombre de clusters, ***k***
Initialiser aléatoirement ***k*** centroïdes
DO
    Assigner chaque point au centroïde le plus proche
    Mettre à jour le centroïde en calculant la moyenne dans chaque cluster
WHILE
    La position du centroïde change

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

K-Means

Pseudocode K-Means

```
Spécifier le nombre de clusters, k
Initialiser aléatoirement k centroïdes
DO
    Assigner chaque point au centroïde le plus proche
    Mettre à jour le centroïde en calculant la moyenne dans chaque cluster
WHILE
    La position du centroïde change
```

à l'aide de la fonction de distance

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

K-Means

Pseudocode K-Means

```
Spécifier le nombre de clusters, k
Initialiser aléatoirement k centroïdes
DO
    Assigner chaque point au centroïde le plus proche
    Mettre à jour le centroïde en calculant la moyenne dans chaque cluster
WHILE
    La position du centroïde change
```

Critères d'arrêt:

- *Les changements du cluster sont inférieurs à un seuil donné,*
- *Le mouvement des centroïdes du cluster est inférieur à un seuil donné.*

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

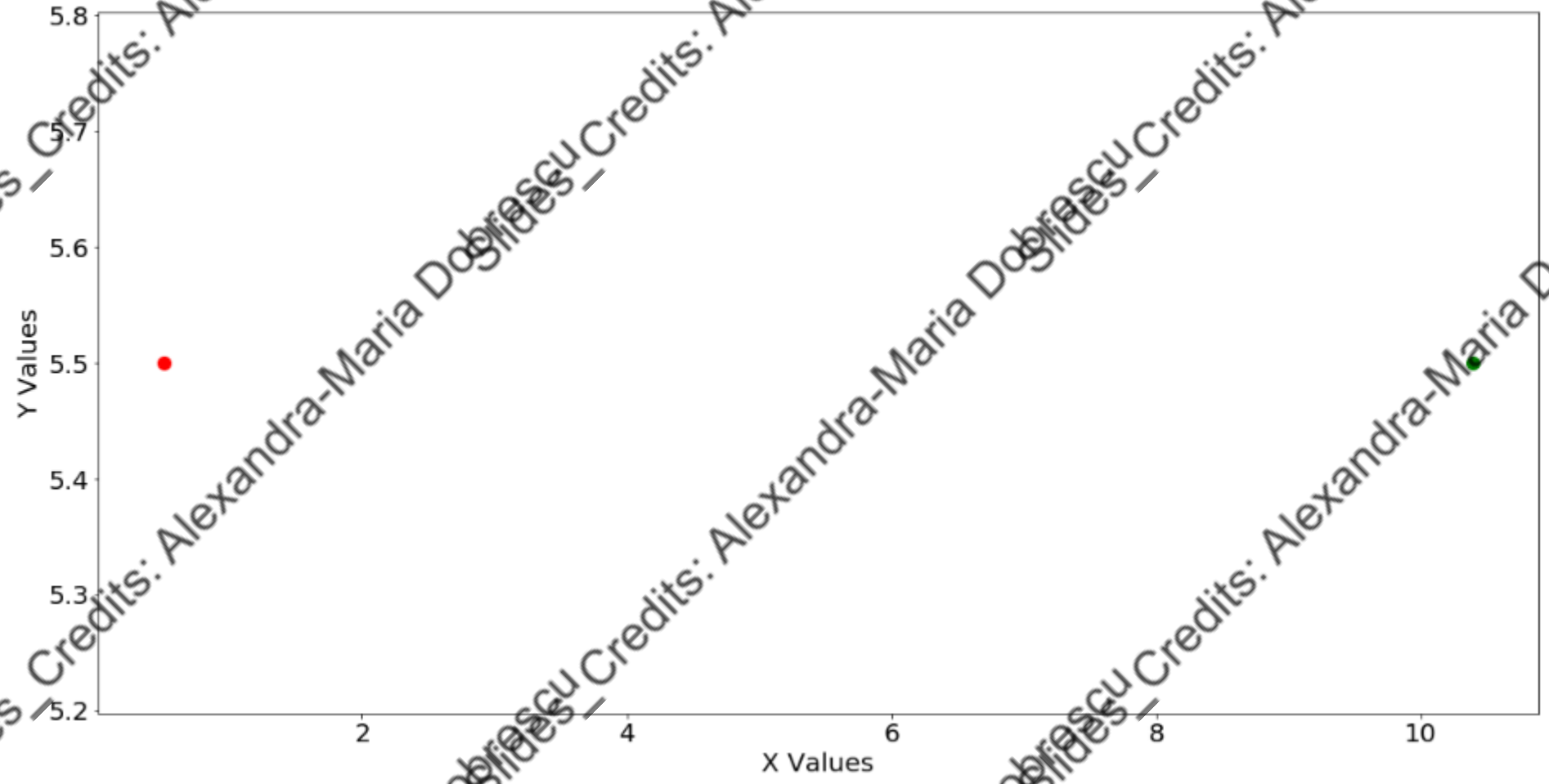
K-Means

Points faibles:

- L'algorithme est sensible aux valeurs aberrantes. Il s'agit dans de nombreux cas d'erreurs causées par des points placés loin les uns des autres. En présence des « outliers », l'algorithme tente de les inclure dans certains groupes et les nouveaux centroïdes, calculés à chaque itération, sont éloignés de leur position naturelle (sans valeurs aberrantes).

Slides_Credits: Alexandra-Maria Dobrescu

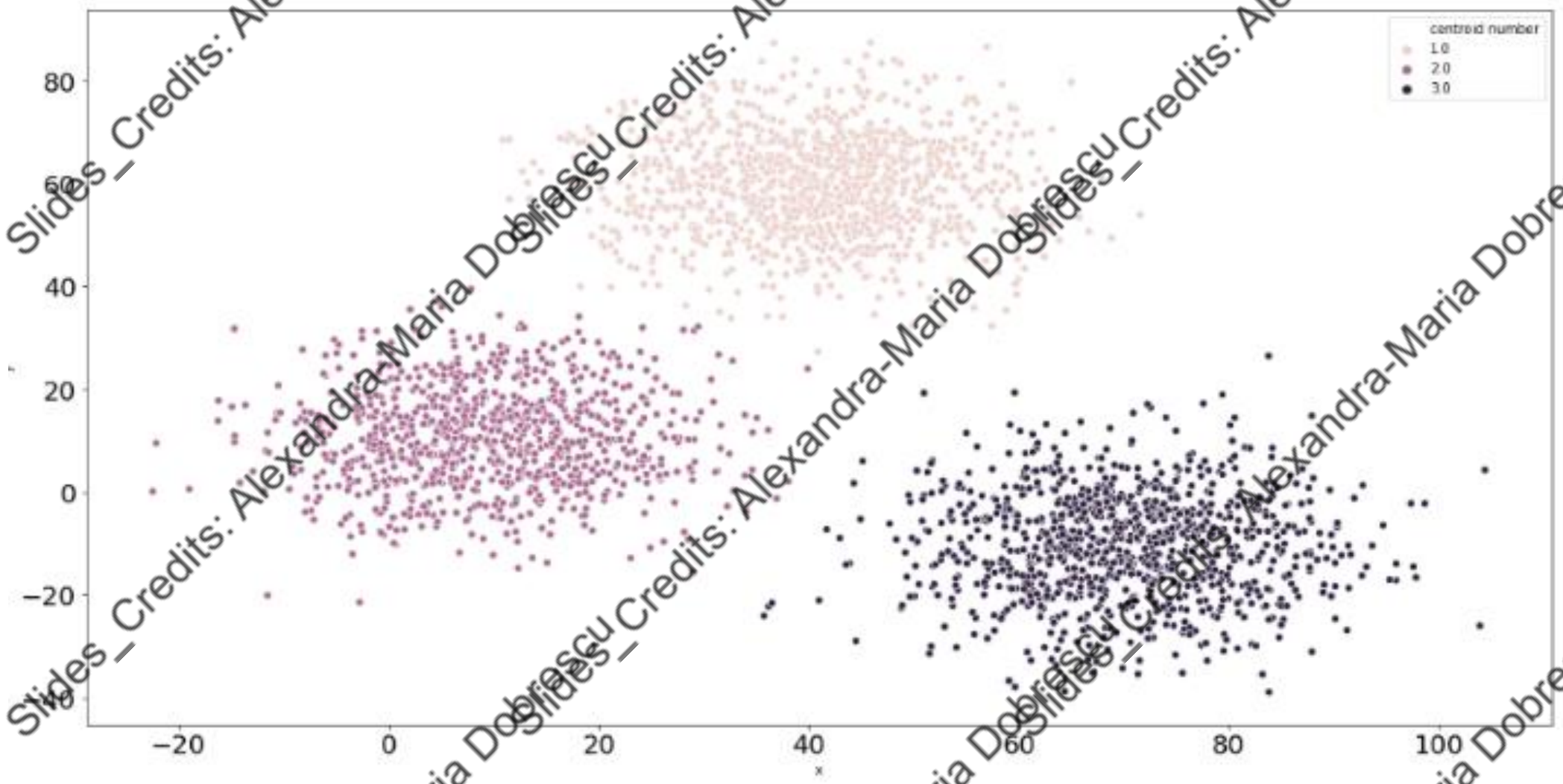
Source [3]



Slides_Credits: Alexandra-Maria Dobrescu

Le point rouge est la moyenne des données sans la valeur aberrante.
Le point vert est la moyenne des données incluant la valeur aberrante.
La valeur aberrante augmente la moyenne des données d'environ 10 unités pour 100 points.

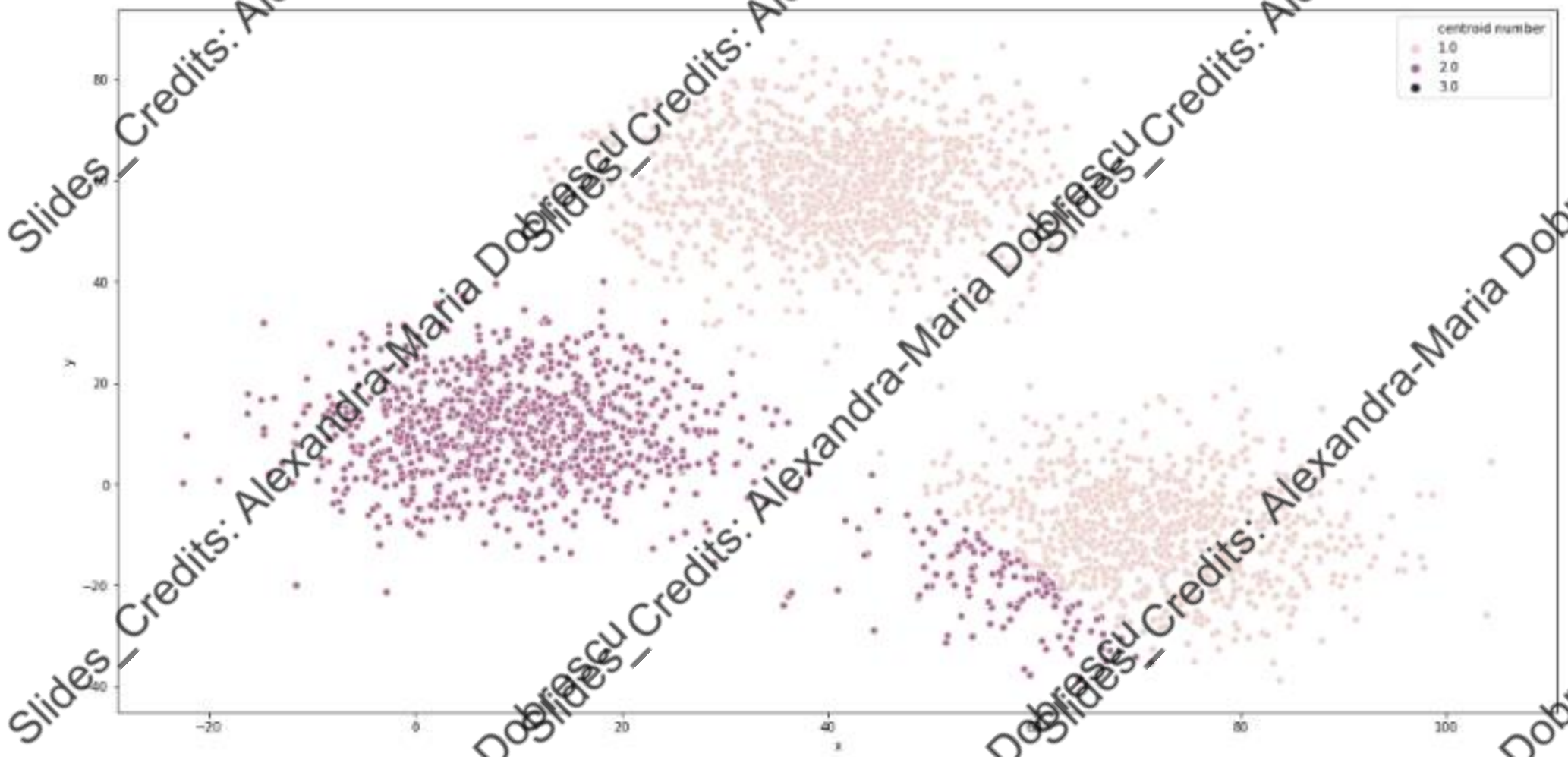
Source [3]



Slides_Credits: Alexandra-Maria Dobrescu

3000 points sans valeurs aberrantes

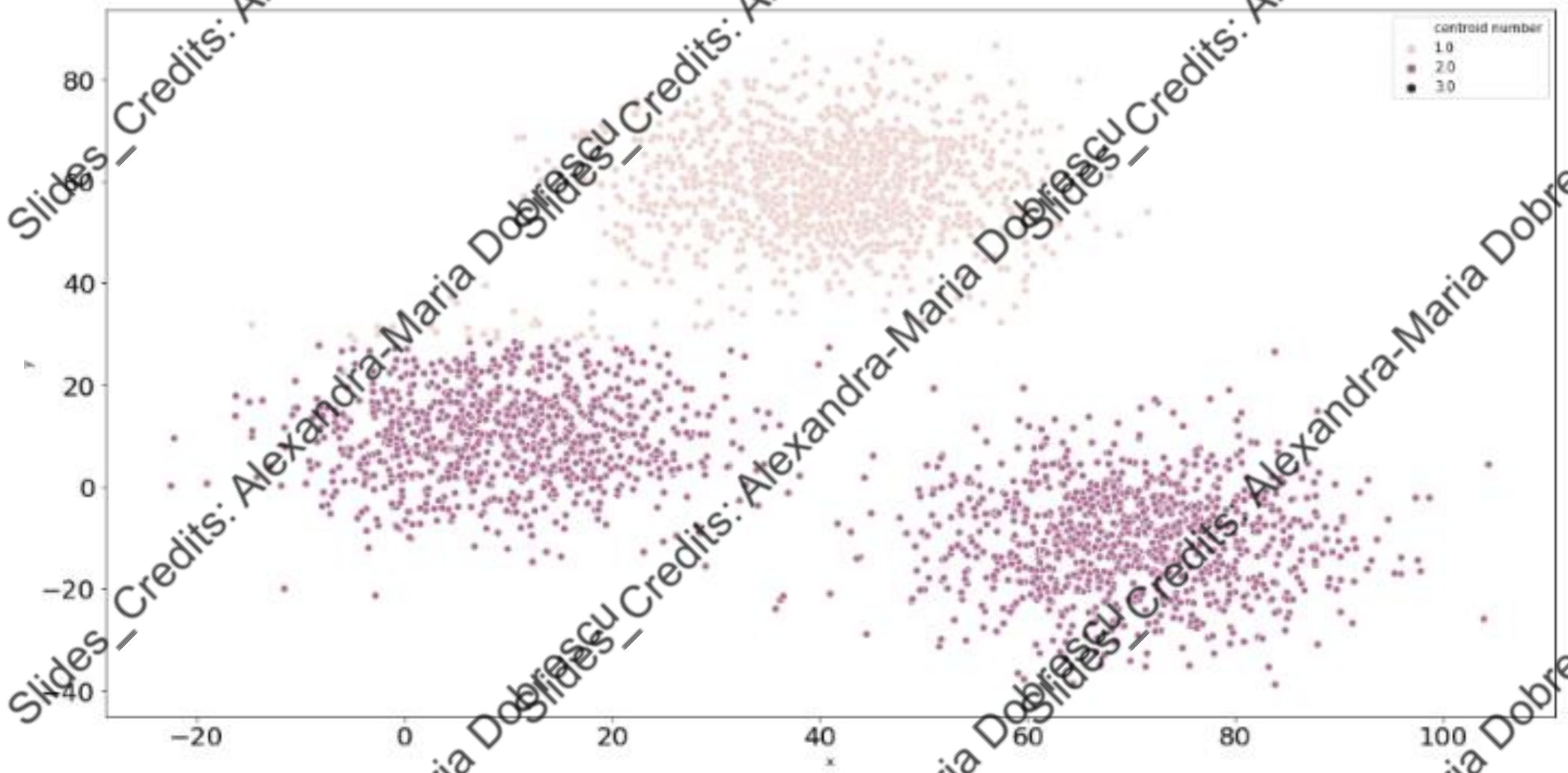
Source [3]



3000 points et 60 points comme « outliers »

Slides_Credits: Alexandra-Maria Dobrescu

Source [3]



3000 points et 60 points comme « outliers »

Slides_Credits: Alexandra-Maria Dobrescu

Discrétisation des données

ALGORITHMES

K-Means

Slides_Credits: Alexandra-Maria Dobrescu

Points faibles:

- Les semences initiales (initial seeds) ont un impact important sur les résultats finaux.

Exemple: la modification des centroïdes initiaux peut entraîner d'autres groupes résultants.

- L'ordre des données a un impact sur les résultats finaux.

Exercice: Considérons *a* et *c* comme centroïdes initiaux, puis *a* et *b* comme centroïdes initiaux.

a b

c d

Discrétisation des données

ALGORITHMES

K-Means

Points forts:

- Facile à mettre en œuvre.
- Une instance peut changer de groupe (passer à un autre groupe) lorsque les centroïdes sont recalculés.
- C'est un algorithme efficace: K-means peut être considéré comme un algorithme linéaire si le nombre d'itérations et le nombre de clusters sont réduits.

Slides_Credits:Alexandra-Maria Dobrescu

Bibliographie

[1] https://www.platform.ai/post/the-silhouette-loss-function-metric-learning-with-a-cluster-validity-index
[2] Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.
[3] https://medium.com/analytics-vidhya/effect-of-outliers-on-k-means-algorithm-using-python-7ba85821ea23

Slides_Credits: Alexandra-Maria Dobrescu