ACTUAIRES 100% DATA SCIENCE
21 NOVEMBRE 2023

Catégorisation d'anomalies à l'aide d'apprentissage non-supervisé

Irchad MAMODE VALJEE
Actuaire manager
GALEA

Fabrice DIAKHATE
Responsable Service Actuariat Inventaire
Crédit Agricole Assurances

Stéphanie BRUGIRARD
Responsable Equipe Data & Risks Monitoring
Crédit Agricole Assurances

//galea

CRÉDIT AGRICOLE ASSURANCES

Qu'est-ce qu'un contrat d'assurance-vie ?

Utilité :

- Moyen d'épargne à moyen et long terme.
- Fiscalité avantageuse.

Types :

- Contrat **EURO** : Placement sur des produits sans risque comme des obligations d'Etat.
- Contrat **UC** : Placement sur des produits à risque, notamment liés à la bourse.

Attributs : Définissent la situation du contrat de manière mensuelle.

(ex : versements effectués, frais de gestion prélevés, etc.)

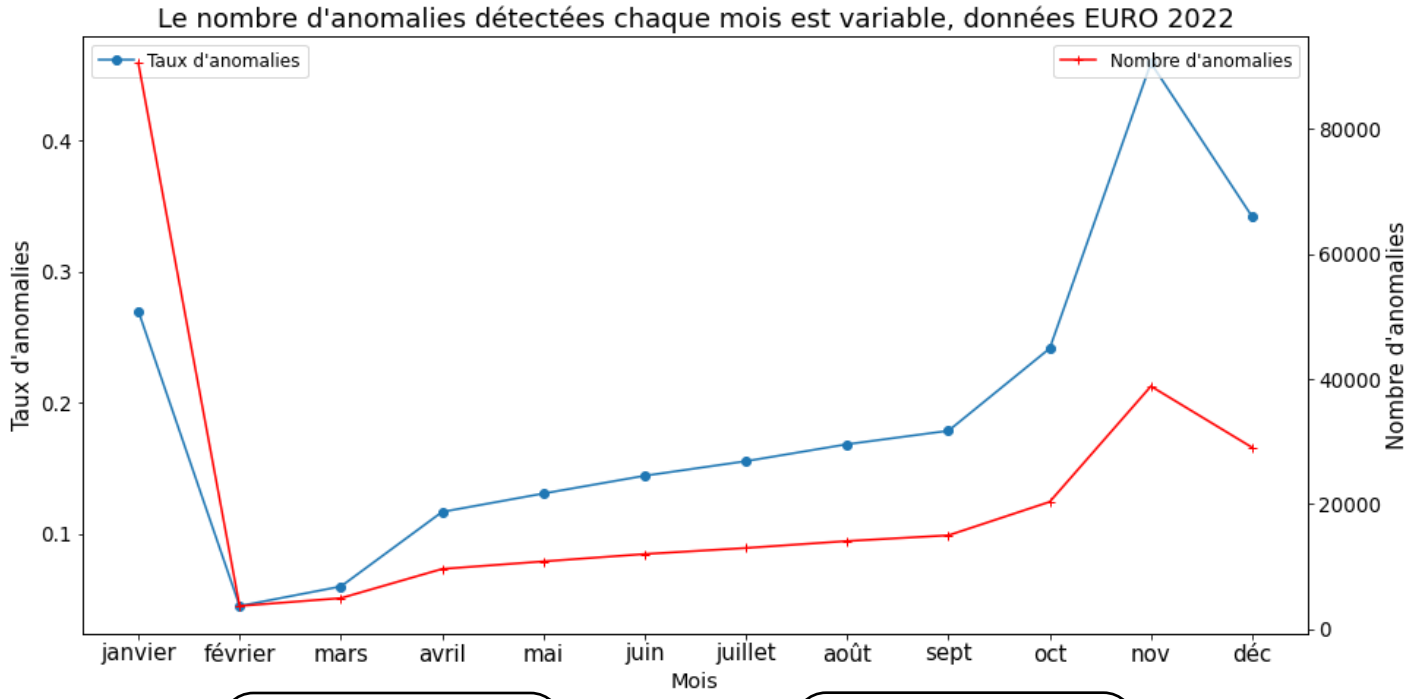
Problématique

Le traitement des anomalies de contrats d'assurance-vie est complexe :

→ De nouvelles anomalies **chaque mois**.

→ Requiert du **temps** de travail.

→ Corrections à réaliser dans des **délais courts**.



125 M
de contrats à traiter par an

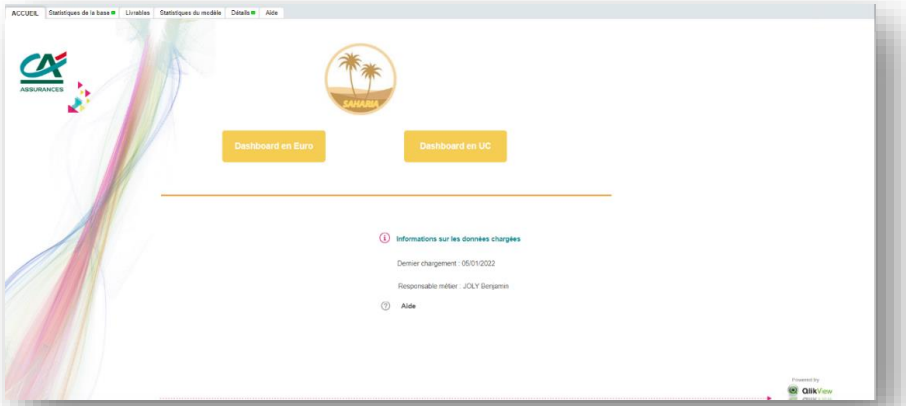
0,19%
d'anomalies par mois en moyenne

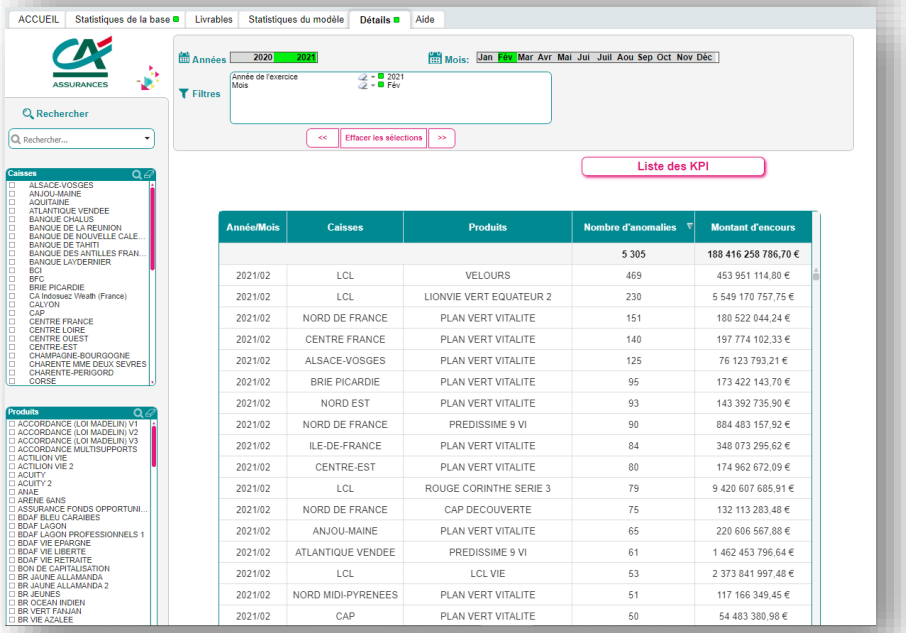
Développement d'une méthode de réduction du temps de traitement des anomalies

//galea

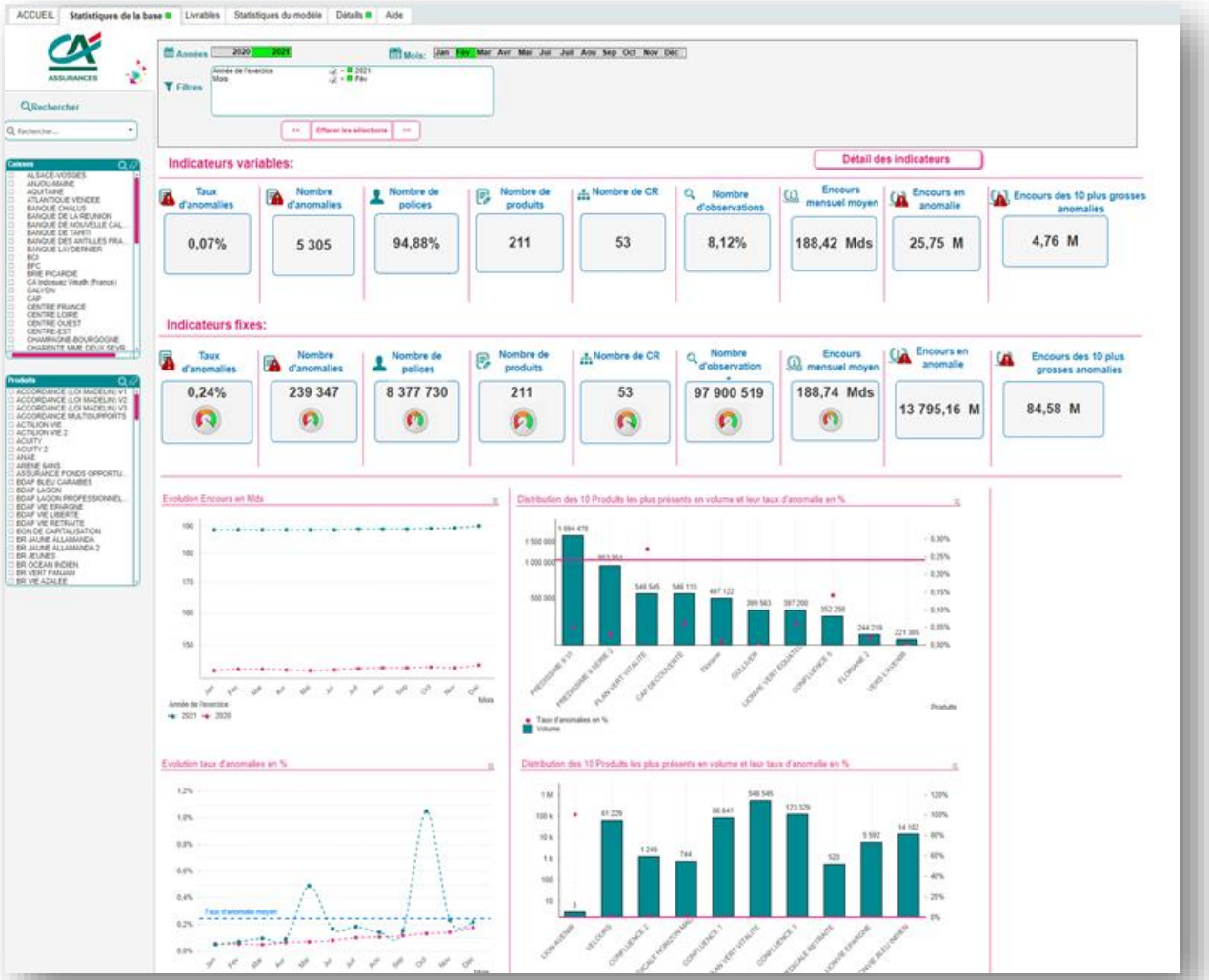
CRÉDIT AGRICOLE ASSURANCES

Monitoring Métier

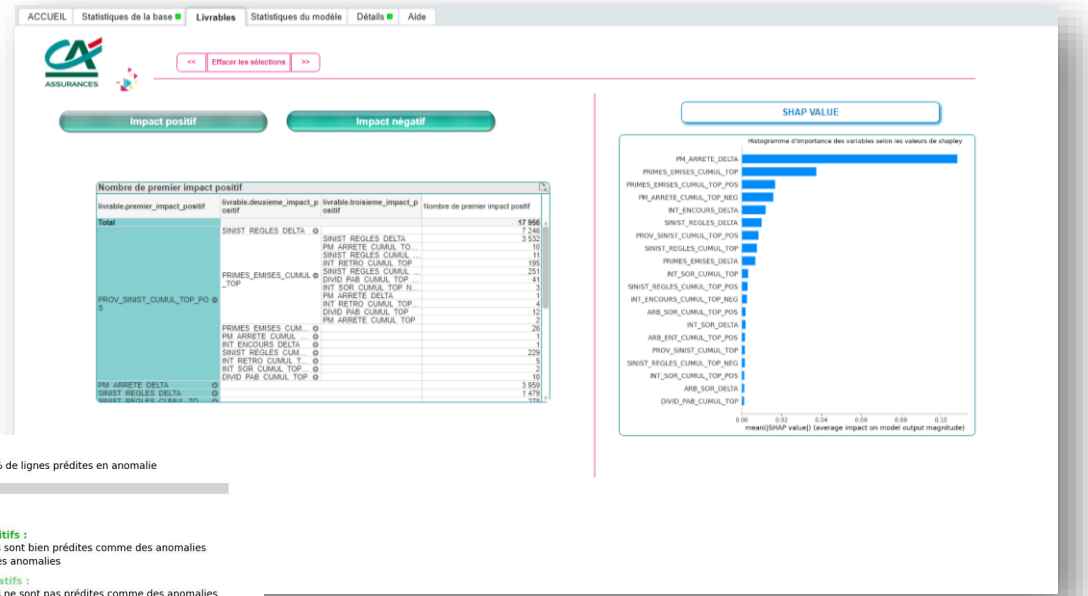


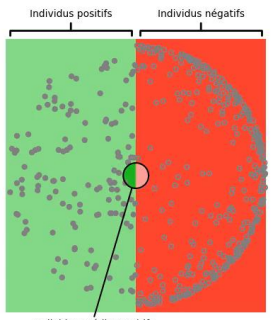


Année/Mois	Caisses	Produits	Nombre d'anomalies	Montant d'encours
			5 305	188 416 258 786,70 €
2021/02	LCL	VELOURS	469	453 951 114,80 €
2021/02	LCL	LIONVIE VERT EQUATEUR 2	230	5 549 170 757,75 €
2021/02	NORD DE FRANCE	PLAN VERT VITALITE	151	180 522 044,24 €
2021/02	CENTRE FRANCE	PLAN VERT VITALITE	140	197 774 102,33 €
2021/02	ALSACE-VOSGES	PLAN VERT VITALITE	125	76 123 793,21 €
2021/02	BRIE PICARDIE	PLAN VERT VITALITE	95	173 422 143,70 €
2021/02	NORD EST	PLAN VERT VITALITE	93	143 392 735,90 €
2021/02	NORD DE FRANCE	PREDISSIME 9 VI	90	884 483 157,92 €
2021/02	ILE-DE-FRANCE	PLAN VERT VITALITE	84	348 073 295,62 €
2021/02	CENTRE-EST	PLAN VERT VITALITE	80	174 962 672,09 €
2021/02	LCL	ROUGE CORINTHE SERIE 3	79	9 420 607 685,91 €
2021/02	NORD DE FRANCE	CAP DECOUVERTE	75	132 113 283,48 €
2021/02	ANJOU-MAINE	PLAN VERT VITALITE	65	220 606 567,88 €
2021/02	ATLANTIQUE VENDEE	PREDISSIME 9 VI	61	1 462 453 796,64 €
2021/02	LCL	LCL VIE	53	2 373 841 997,48 €
2021/02	NORD MIDI-PYRENEES	PLAN VERT VITALITE	51	117 166 349,45 €
2021/02	CAP	PLAN VERT VITALITE	50	54 483 380,98 €



Monitoring IA : Suivi de la performance de l'IA





0.0% de lignes prédites en anomalie

Vrais positifs :
0 anomalies sont bien prédites comme des anomalies
soit 0% des anomalies

Faux négatifs :
383 anomalies ne sont pas prédites comme des anomalies
soit 100% des anomalies

Vrais négatifs :
1026 lignes régulières sont bien prédites comme régulières
soit 100% des lignes régulières

Faux positifs :
0 lignes régulières sont prédites comme des anomalies
soit 0% des lignes régulières

$$\text{Precision} = \frac{}{} = 100.0\%$$

$$\text{Recall} = \frac{}{} = 0.0\%$$

Vue d'ensemble de la méthode de traitement des anomalies

- **Détecter** les contrats présentant une anomalie de récurrence un mois donné.
- **Partitionner** les contrats en groupes basés sur la similarité de leur situation.
- **Extraire** les variables discriminantes de chaque partition.
- **Visualiser** les relations entre partitions.

Comment repérer les anomalies ?

Méthode de l'équation de récurrence :

- **Egalité** liant tous les attributs d'un contrat d'assurance-vie devant être respectée pour chaque contrat.
- Détecter si les évènements d'entrée et de sortie d'argent arrivant sur un contrat sont **déséquilibrés**.
- **Détecter** la présence d'une anomalie, mais pas sa cause.

Equation de récurrence (simplifiée) :

$$Encours_{initial} + Entrées - Sorties - Frais - Encours_{final} = 0$$

Prétraitement des données

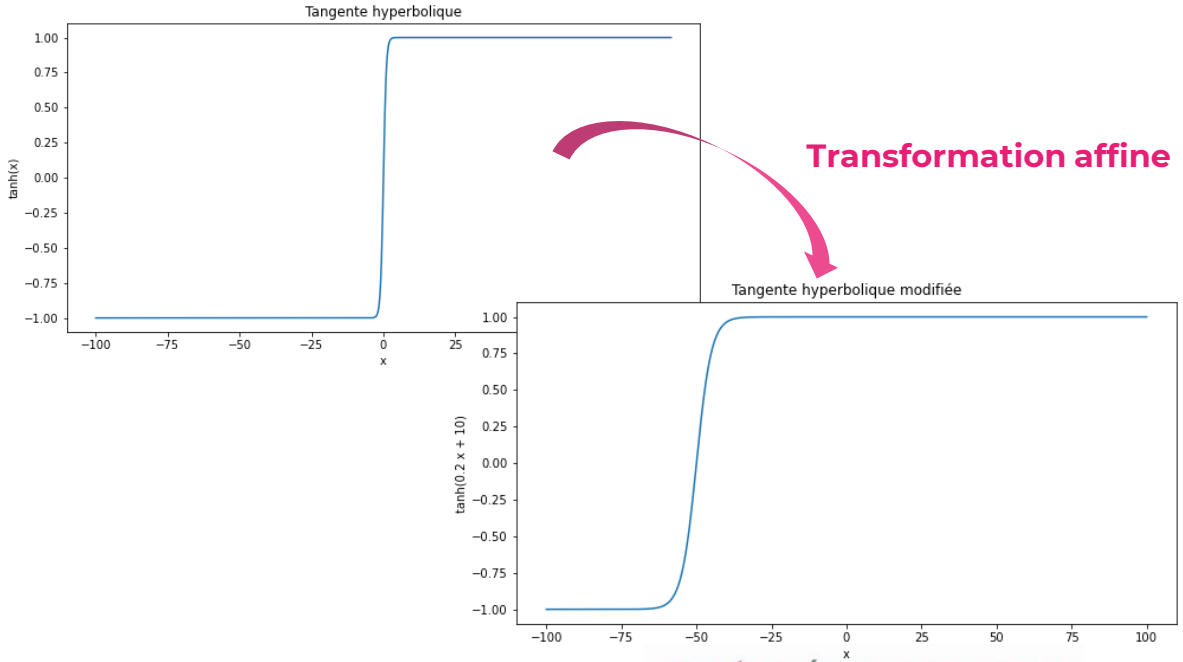
Ecarts d'échelle importants entre les montants gérés sur différents contrats.

→ Fonction de transformation de données non-linéaire pour réduire l'écart entre valeurs extrêmement fortes et faibles : **Tangente hyperbolique**.

→ Application d'une **transformation affine** pour modifier l'intensité et le biais de la pente selon la variable.

```
count    3057.000000
mean       46.946524
std       219.592058
min         0.000000
25%         0.000000
50%         0.231600
75%        16.186300
max      7351.377900
Name: NB_PM_OUVERTURE
```

Ordre de magnitude 70 entre les quantiles 50 % et 75 % de la variable « PM d'ouverture ».



galea

CRÉDIT AGRICOLE ASSURANCES

Prétraitement des données

Typographies différentes pour désigner les mêmes anomalies au sein de la base de données.

→ Création de labels communs.

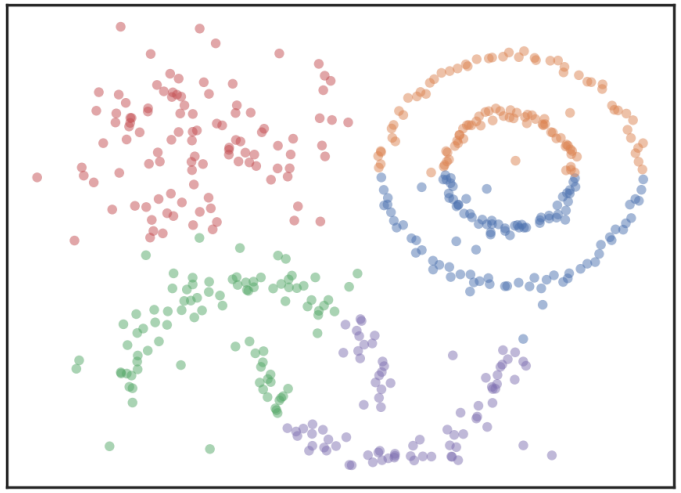
→ Mise en place de scripts de correction automatique.

suite fiche 936384 RPP non pris en compte PM fin mai OK	65
Pm 31/12/2021 ok dans ARPEGE - PM fin OK	35
ev non pris en compte ou annulation non restituée	15
PM 31/12/2021 erronée	13
à surveiller en cours XFRI en juin	13

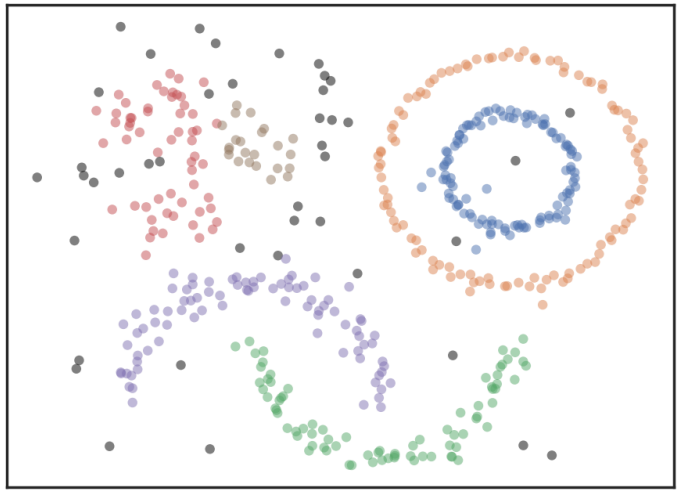
Les deux labels encadrés en rouge représentent la même anomalie concernant l'attribut « provision mathématique ».

Méthodes de Clustering

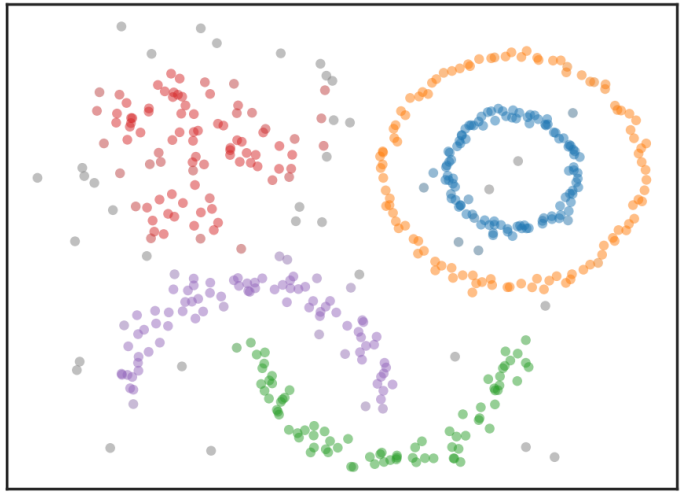
K-Means



DBSCAN



HDBSCAN

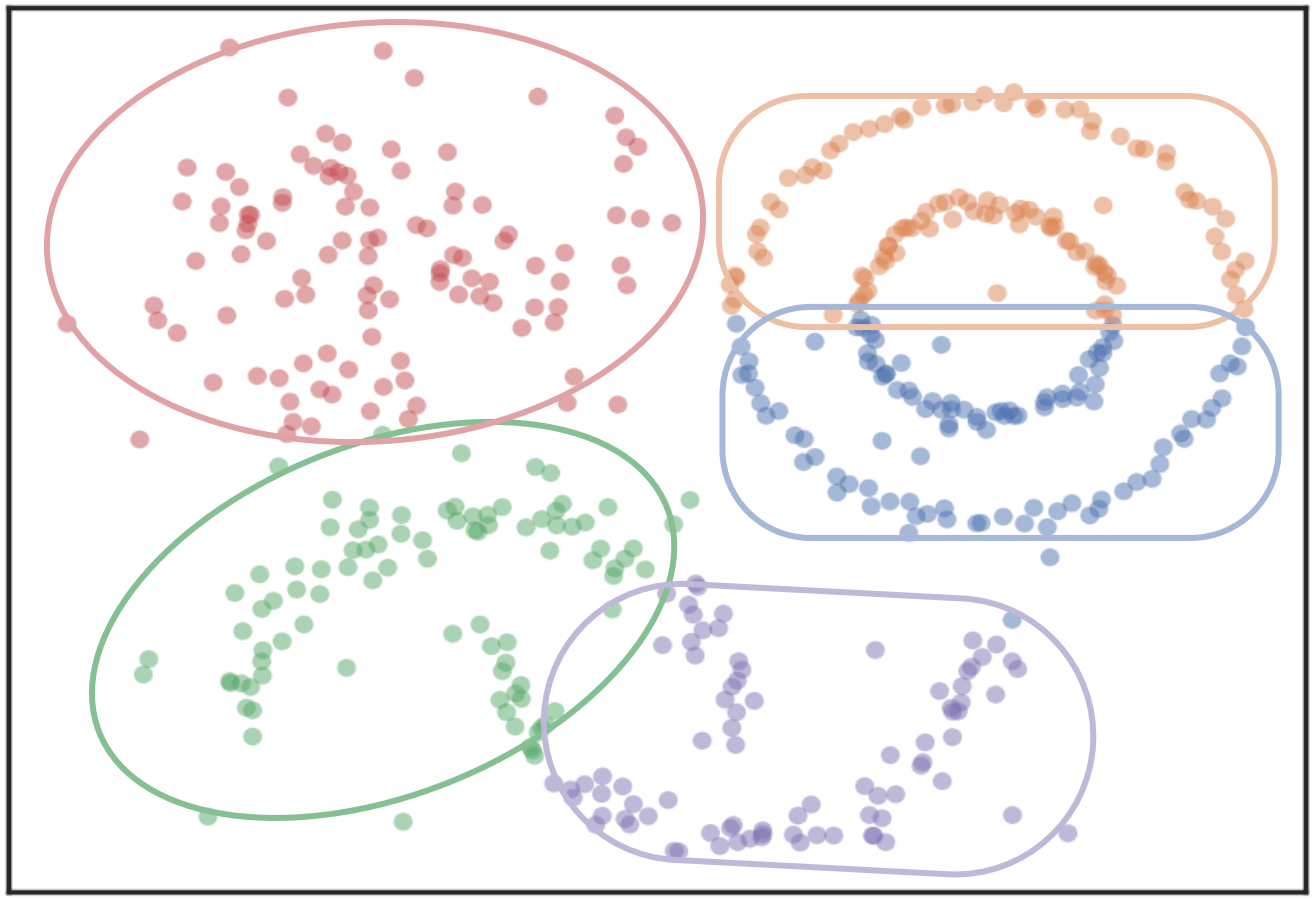


Méthodes de Clustering

K-Means

Partition de l'espace :

- Hypothèse forte sur la forme des clusters.
- Inadapté pour la détection d'outliers.
- Connaissance du nombre de clusters.

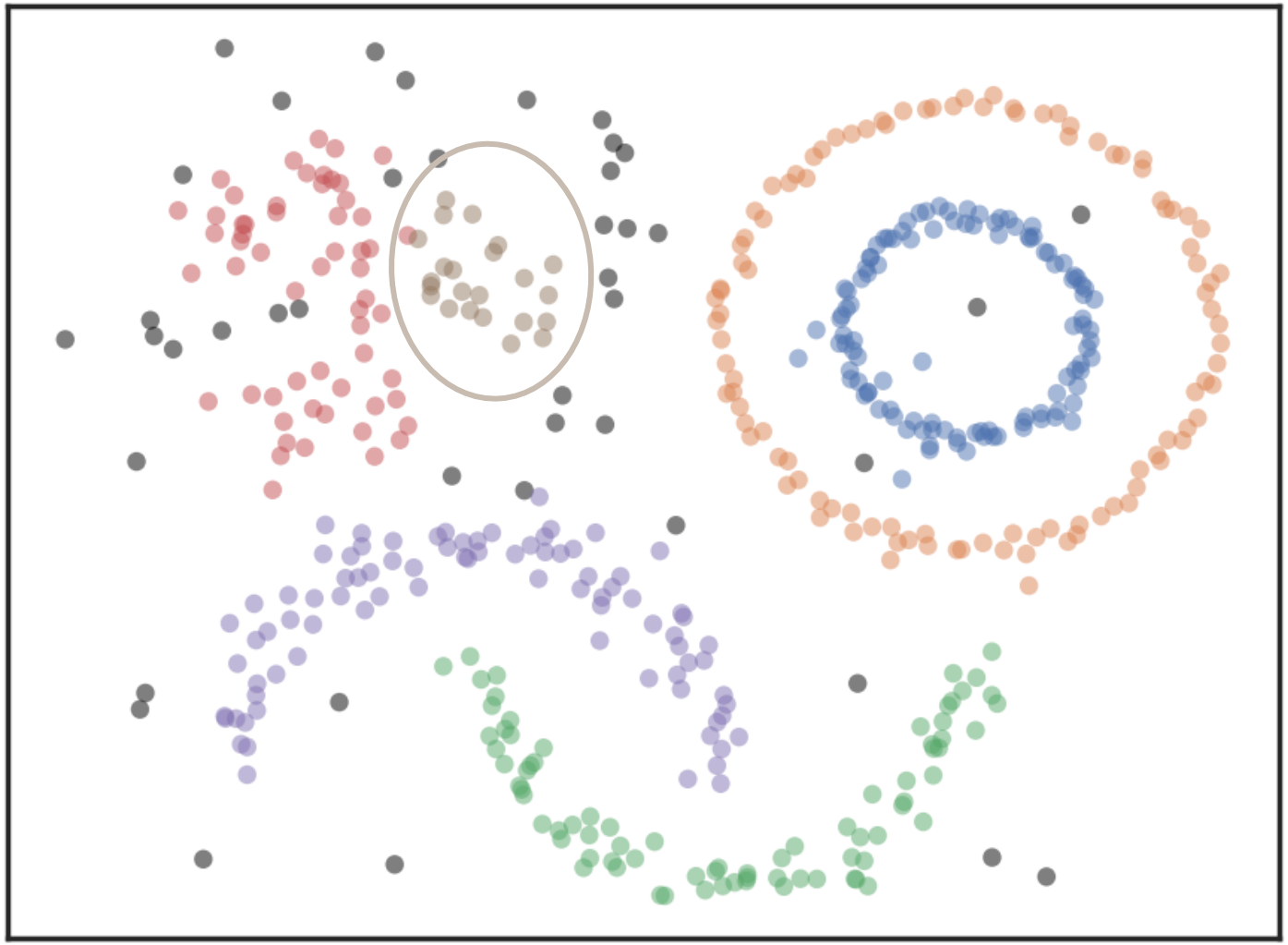


Méthodes de Clustering

DBSCAN (1/13)

Partition de l'espace :

- Aucune hypothèse sur la forme des clusters.
- Robuste au bruit.
- Aucune connaissance sur nombre de clusters.
- Peu robuste à une densité non uniforme
- Paramétrage non intuitif

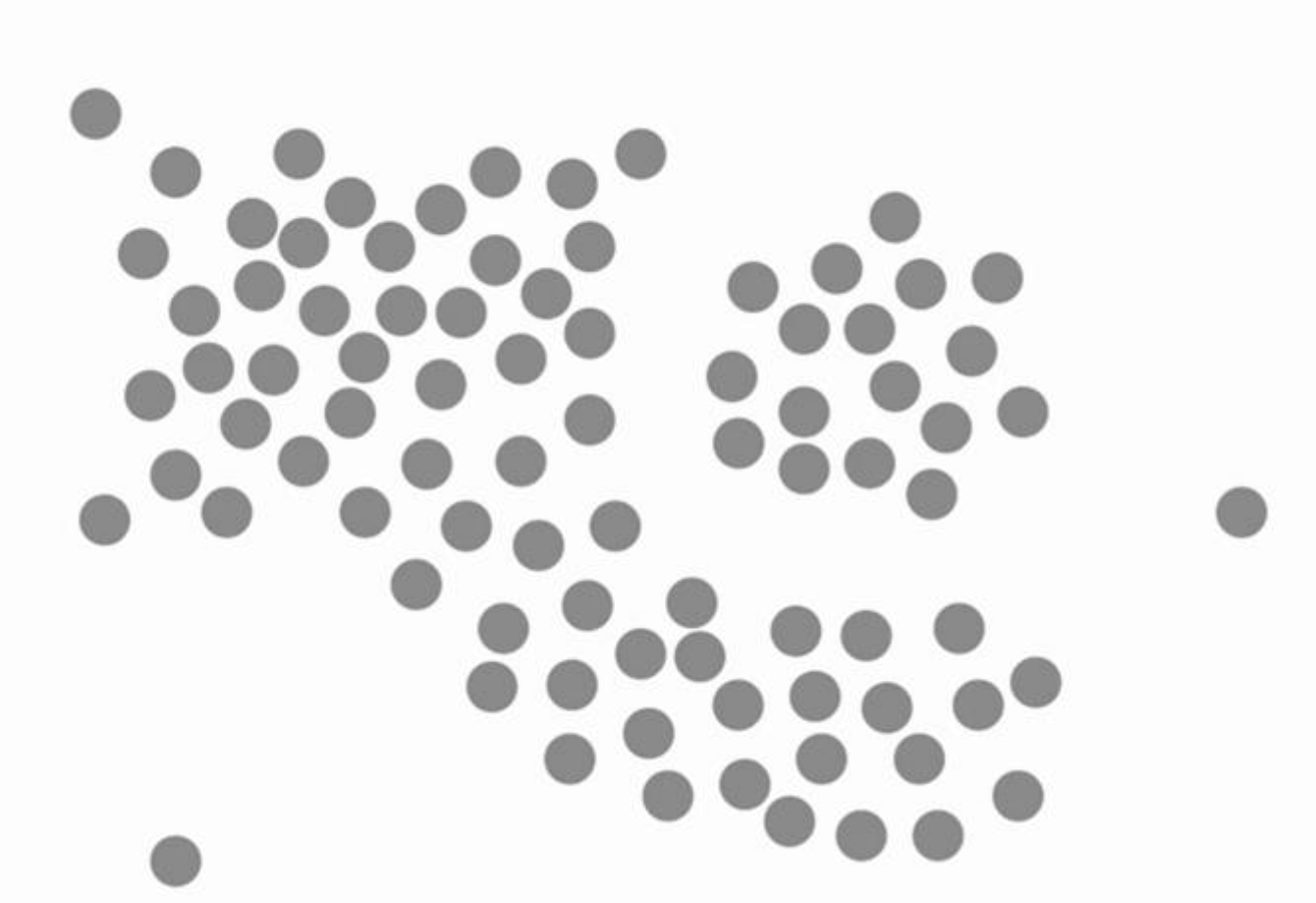


Méthodes de Clustering

DBSCAN (2/13)

Objectif : Obtenir des **clusters non sphériques en fonction de la densité** des points.

Comment : Définir la notion de **point dense** à travers deux paramètres ε et min_sample.



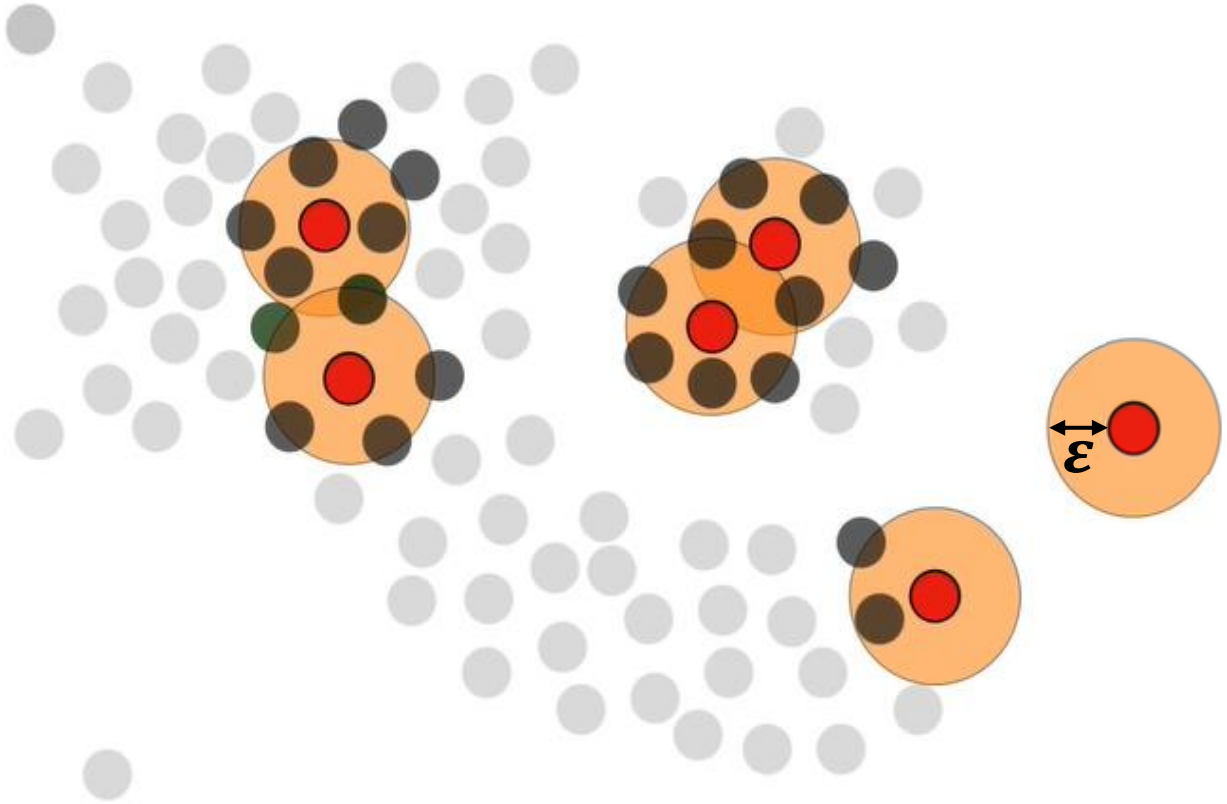
Méthodes de Clustering

DBSCAN (3/13)

1 – Notion de ε voisinage

Objectif : Définir une **métrique** pour mesurer la distance entre les observations.

Comment : Définir un **seuil de distance ε** pour détecter les observations considérées comme proches.



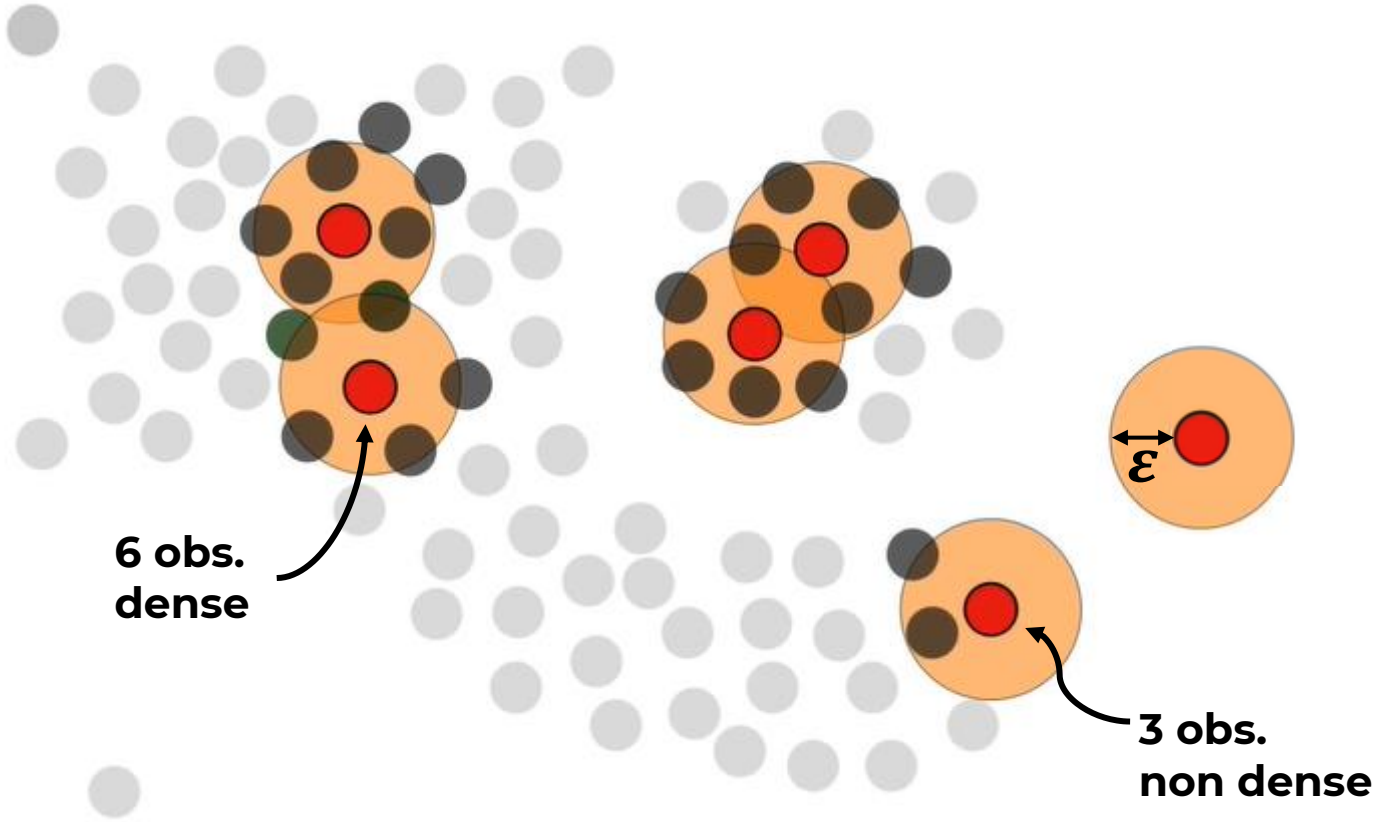
Méthodes de Clustering

DBSCAN (4/13)

2 – Notion de point dense

Objectif : Sélectionner les observations formant des **zones denses**.

Comment : Définir le nombre minimal d'observations min_sample dans le voisinage proche (ε voisinage) pour **identifier les points denses**.



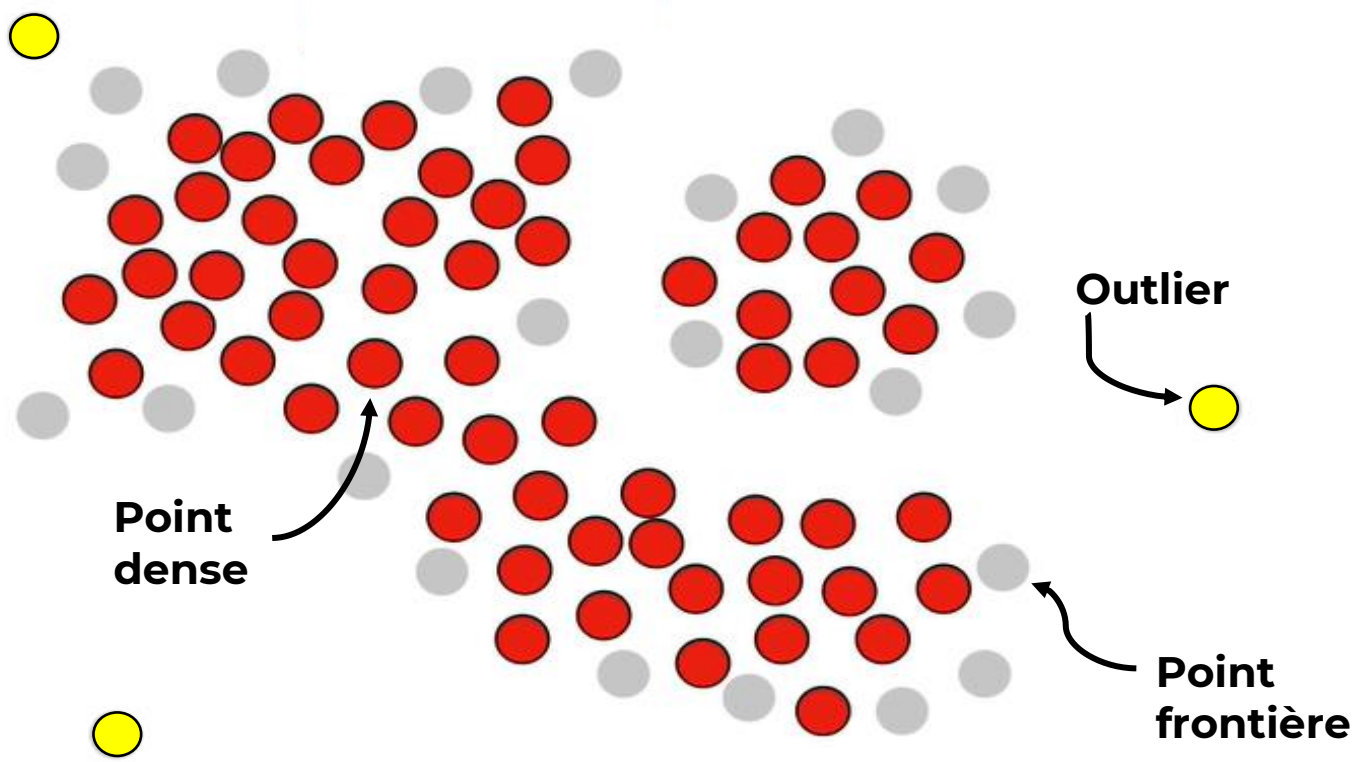
Méthodes de Clustering

DBSCAN (5/13)

3 – Notion de point frontière et d'outlier

Objectif : **Etiqueter** chaque point selon son niveau de densité :

- Point dense
- Point frontière
- Outlier



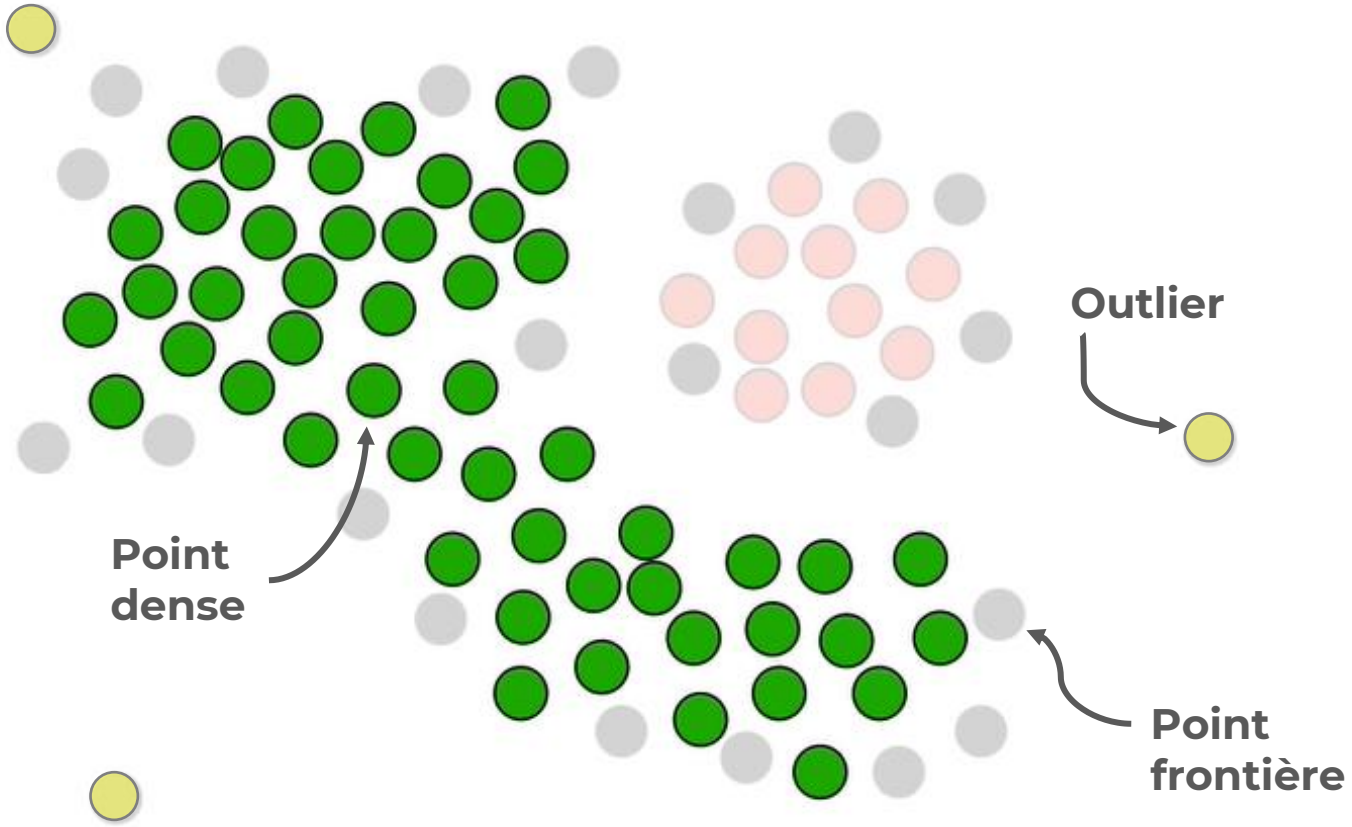
Méthodes de Clustering

DBSCAN (6/13)

4 – Création des clusters : Etape 1

Objectif : Utiliser les points denses pour créer des clusters.

Comment : **Agréger les points denses** à distance ε afin de former un cluster



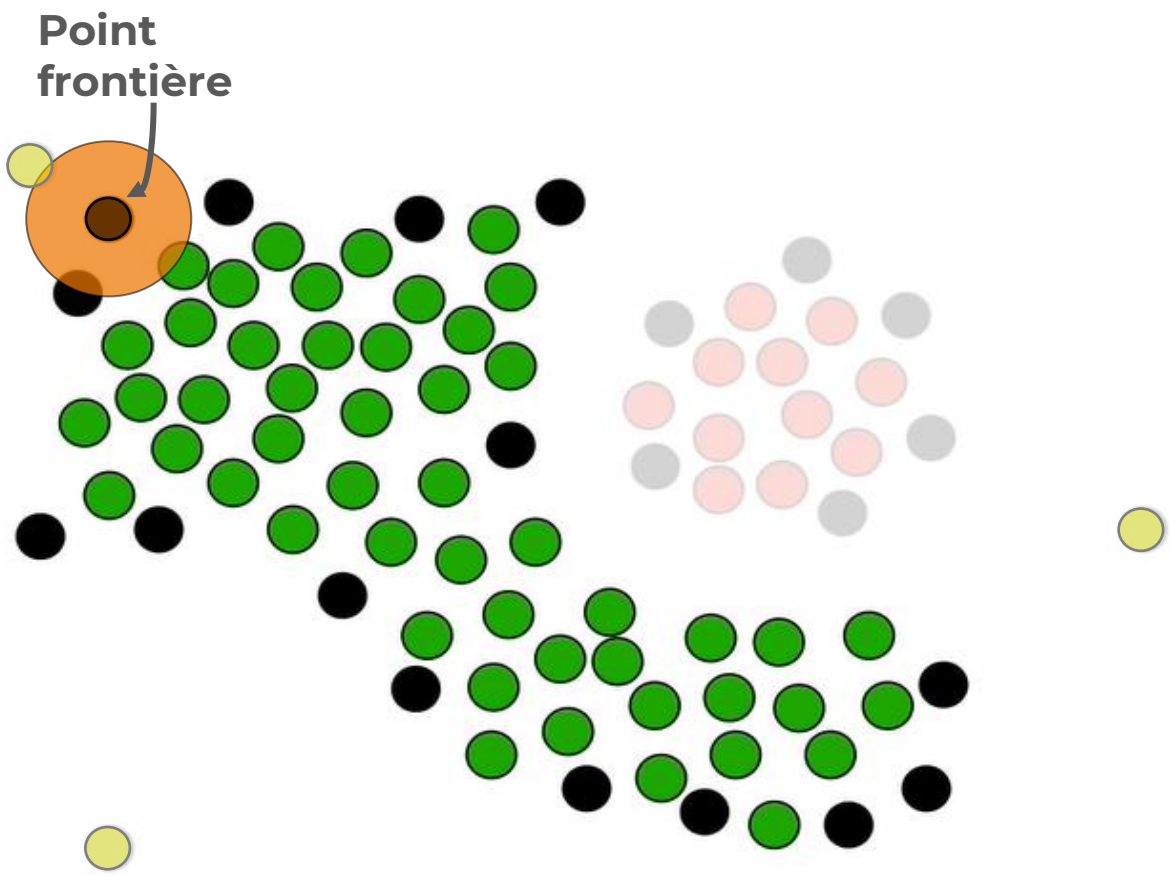
Méthodes de Clustering

DBSCAN (7/13)

4 – Création des clusters : Etape 2

Objectif : Utiliser les points denses pour créer des clusters.

Comment : **Agrandir le cluster avec les points frontières** à distance ε d'un point du cluster.



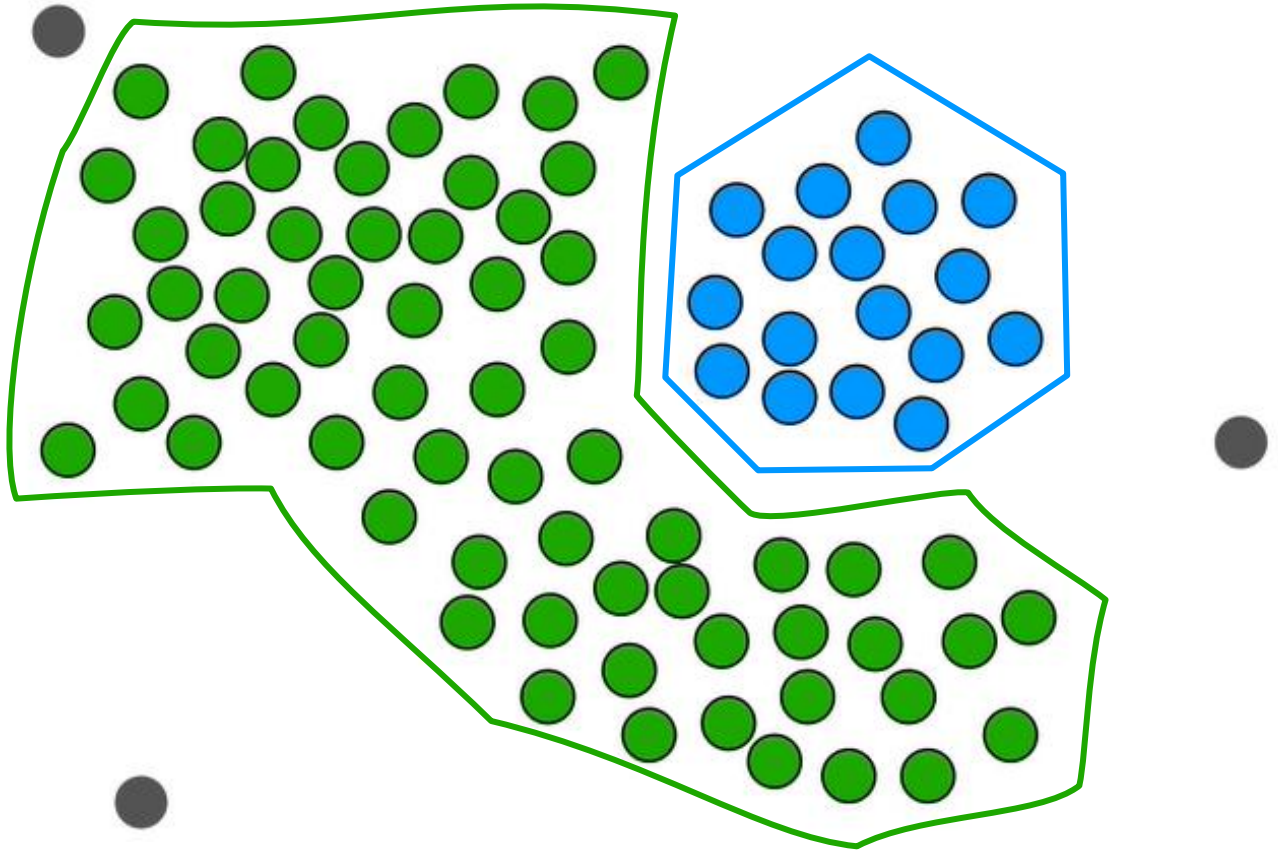
Méthodes de Clustering

DBSCAN (8/13)

5 – Résultat de l'algorithme

→ Les clusters sont de **formes différentes** (non sphérique pour le cluster 1)

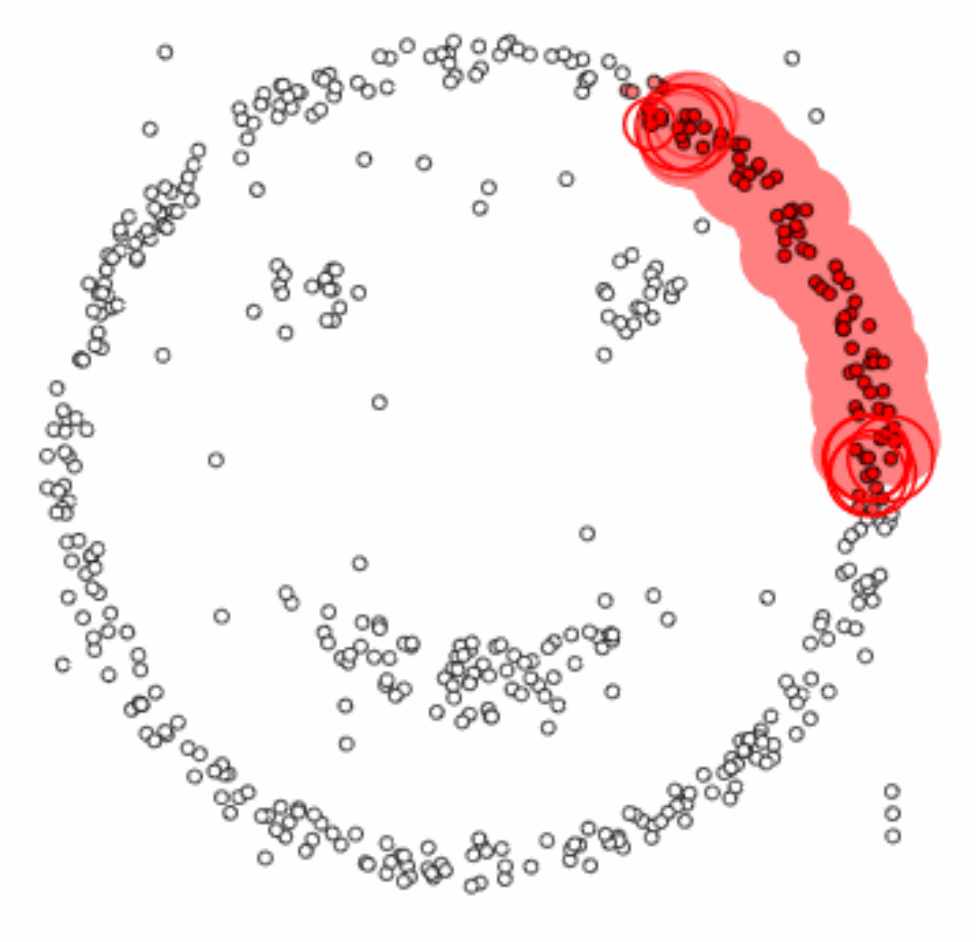
→ Des points éloignés peuvent appartenir à un même cluster via une suite de points connectés (**phénomène de propagation**)



Méthodes de Clustering

DBSCAN (9/13)

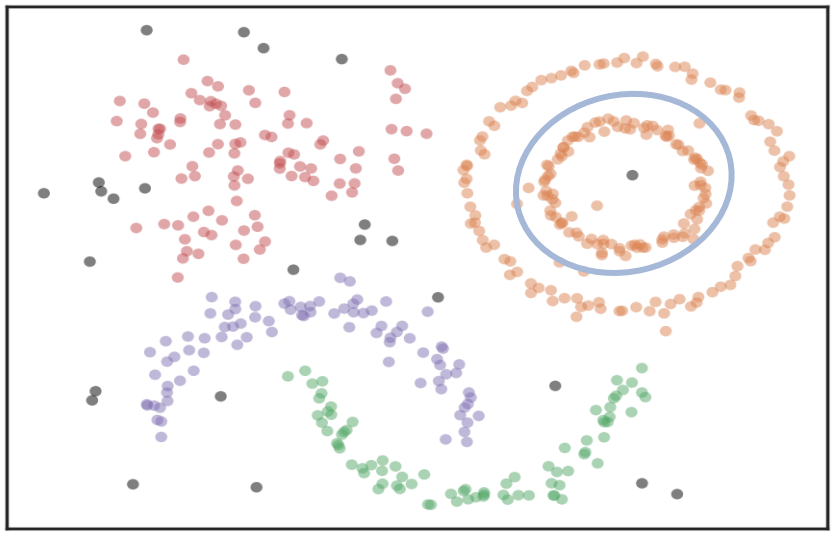
Exemple

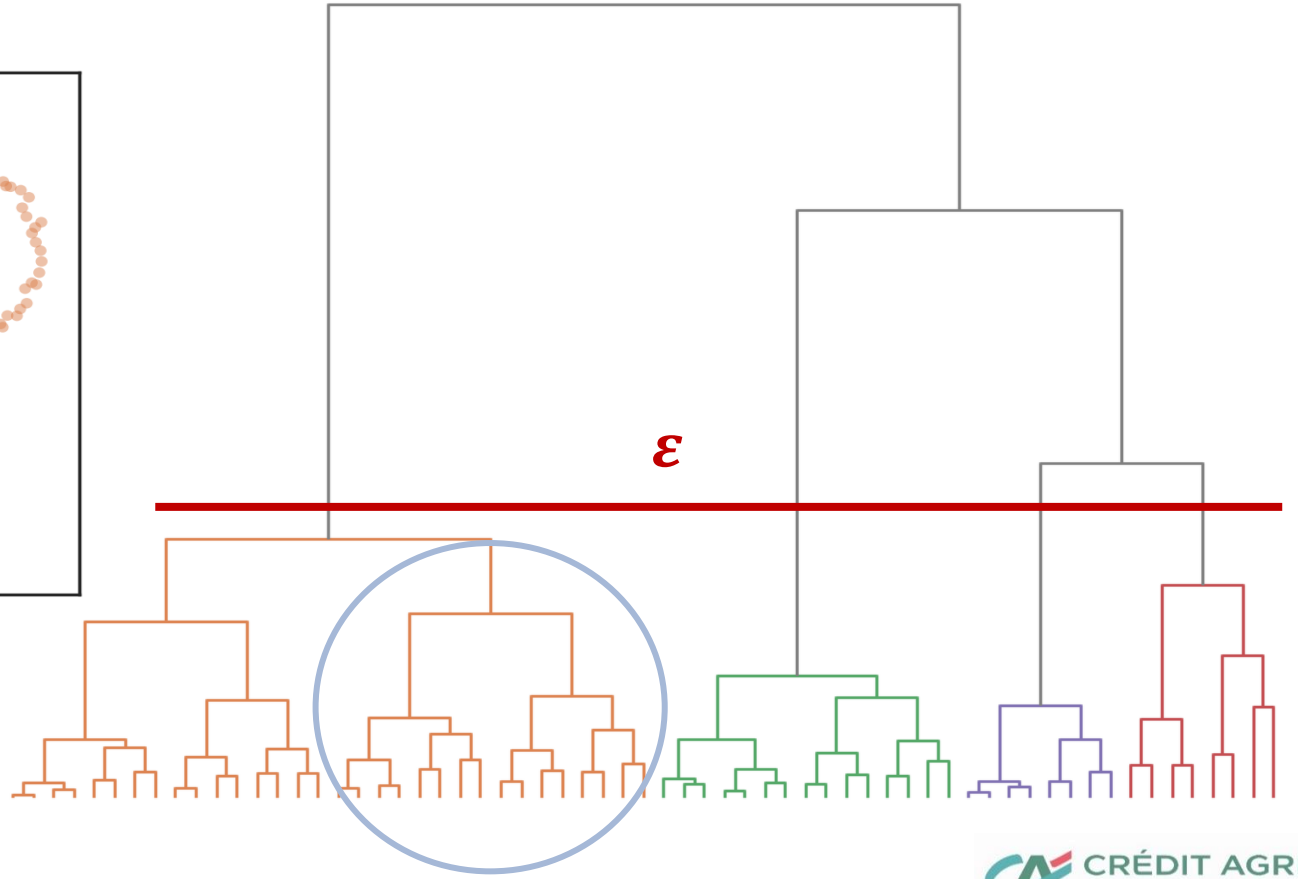


Méthodes de Clustering

DBSCAN (10/13)

Hiérarchisation du clustering : Approche par dendrogramme

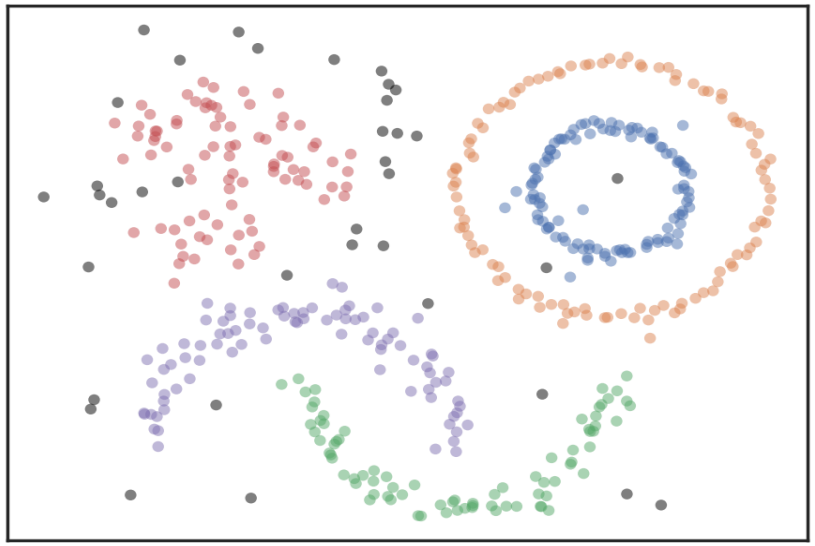


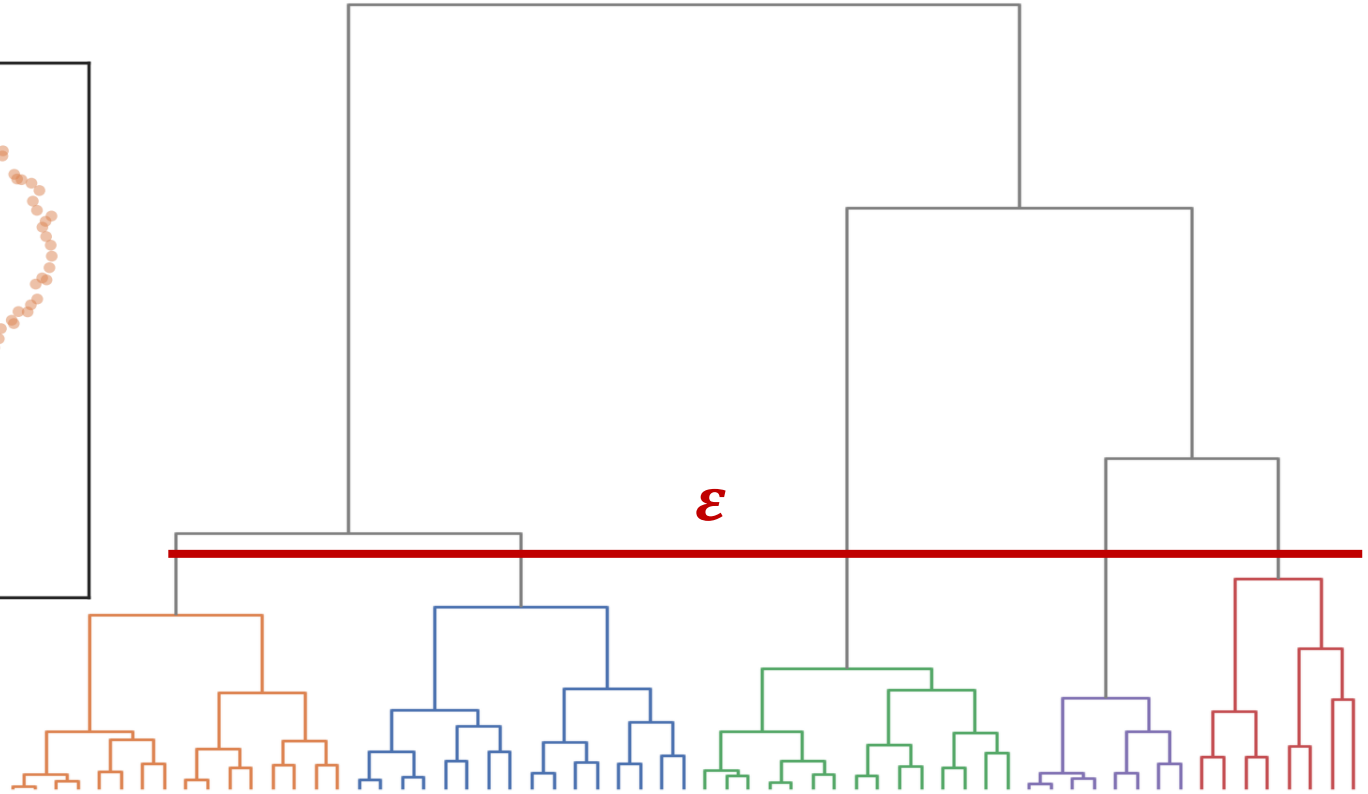


Méthodes de Clustering

DBSCAN (11/13)

Hiérarchisation du clustering : Approche par dendrogramme

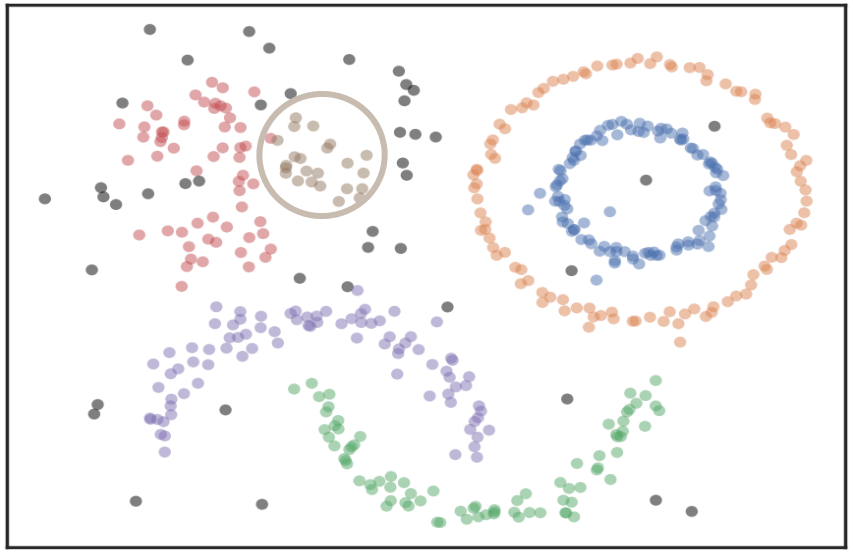


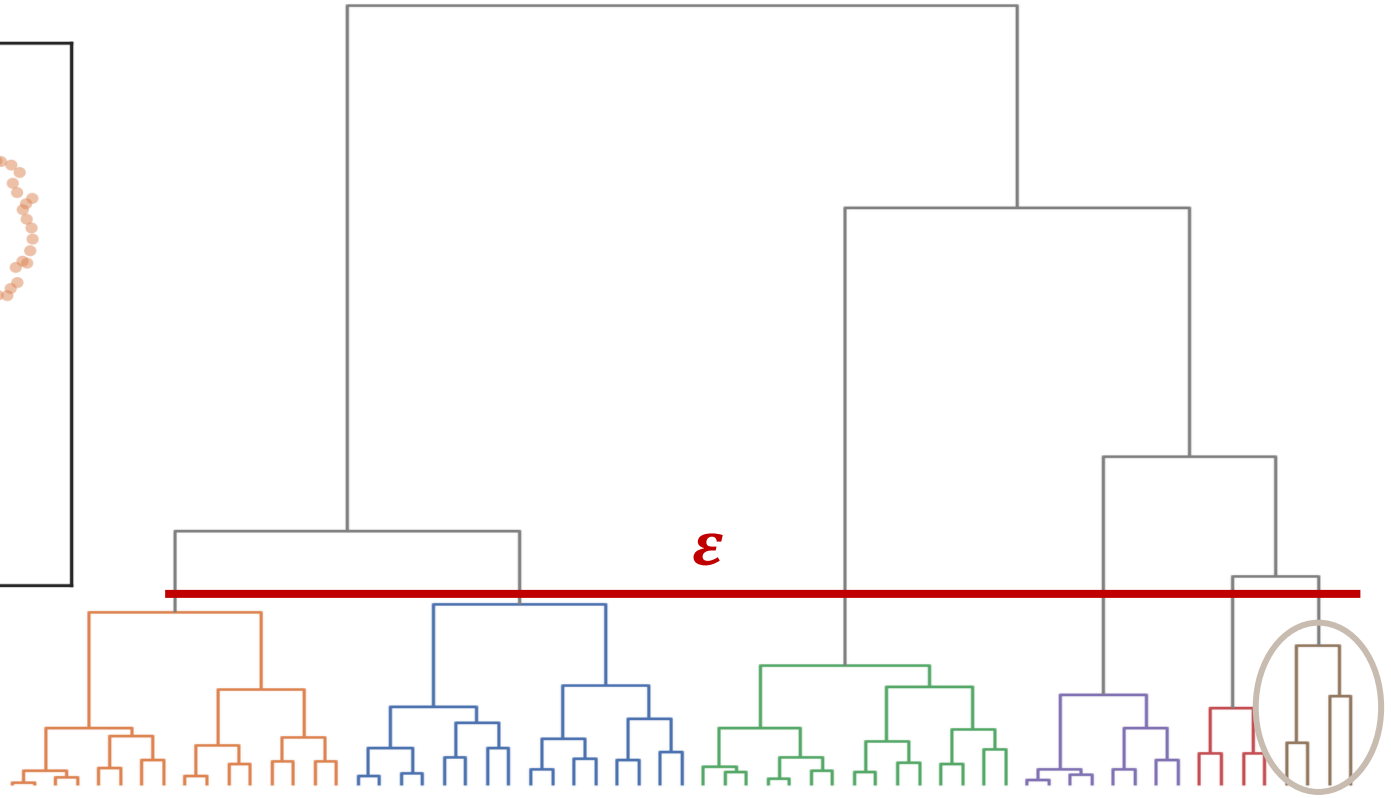


Méthodes de Clustering

DBSCAN (12/13)

Hiérarchisation du clustering : Approche par dendrogramme





Méthodes de Clustering

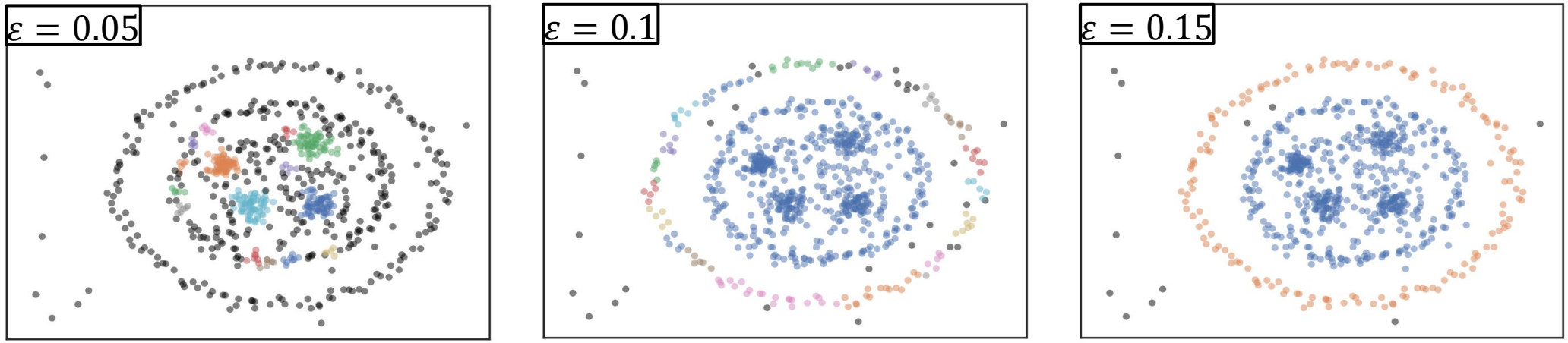
DBSCAN (13/13)

6 – Limites du modèle

Le choix du paramètre ε est très **peu intuitif** :

→ Pour une zone avec de **nombreux points rapprochés**, un niveau de **seuil faible** est à privilégier sinon tous les points appartiendront à un **unique cluster**.

→ Pour une zone avec des **points espacés**, un **seuil élevé** est à privilégier sinon de nombreux points sont identifiés en **outlier**.



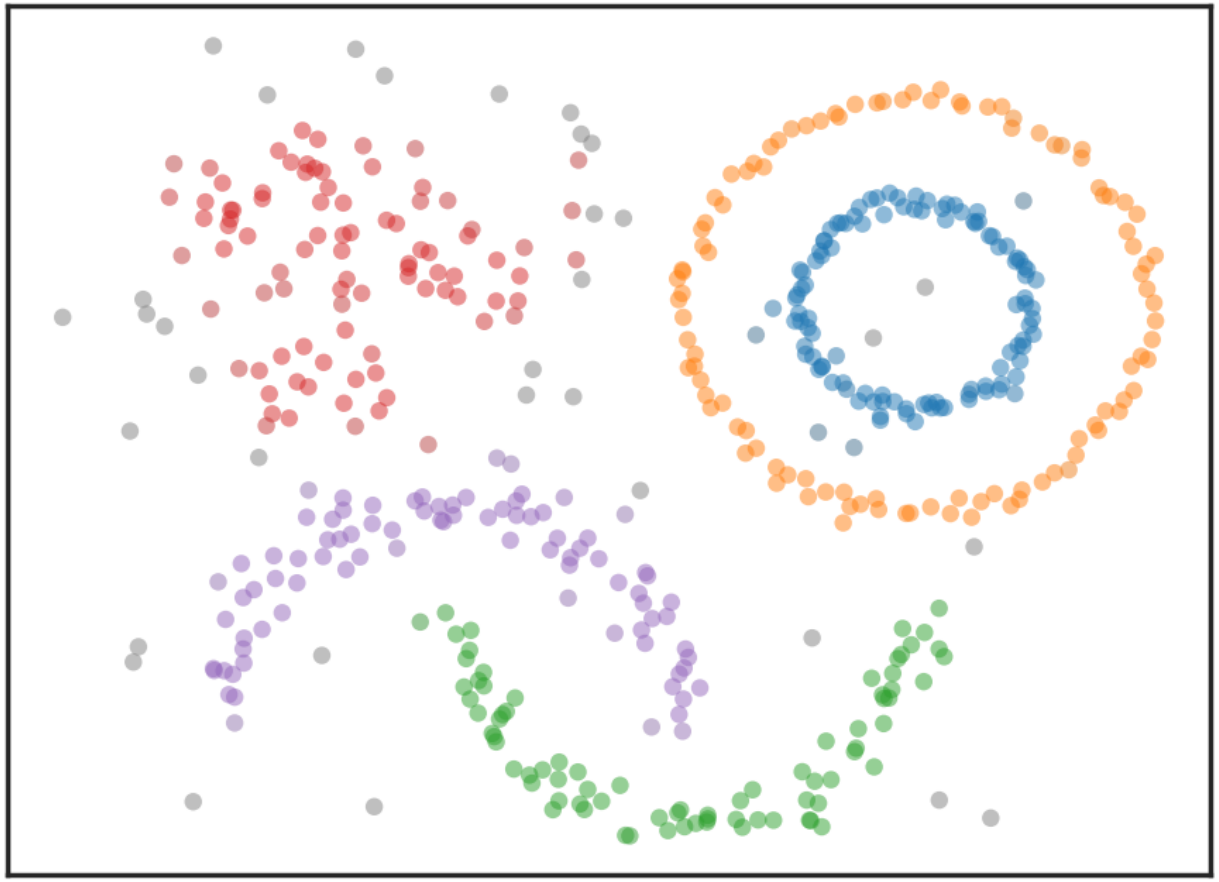
→ Modèle **inadapté** en présence de **plusieurs densités** différentes.

Méthodes de Clustering

HDBSCAN (1/8)

Partition de l'espace :

- Aucune hypothèse sur la forme des clusters.
- Robuste au bruit.
- Aucune connaissance sur nombre de clusters.
- Robuste à une densité non uniforme
- Paramétrage intuitif



Méthodes de Clustering

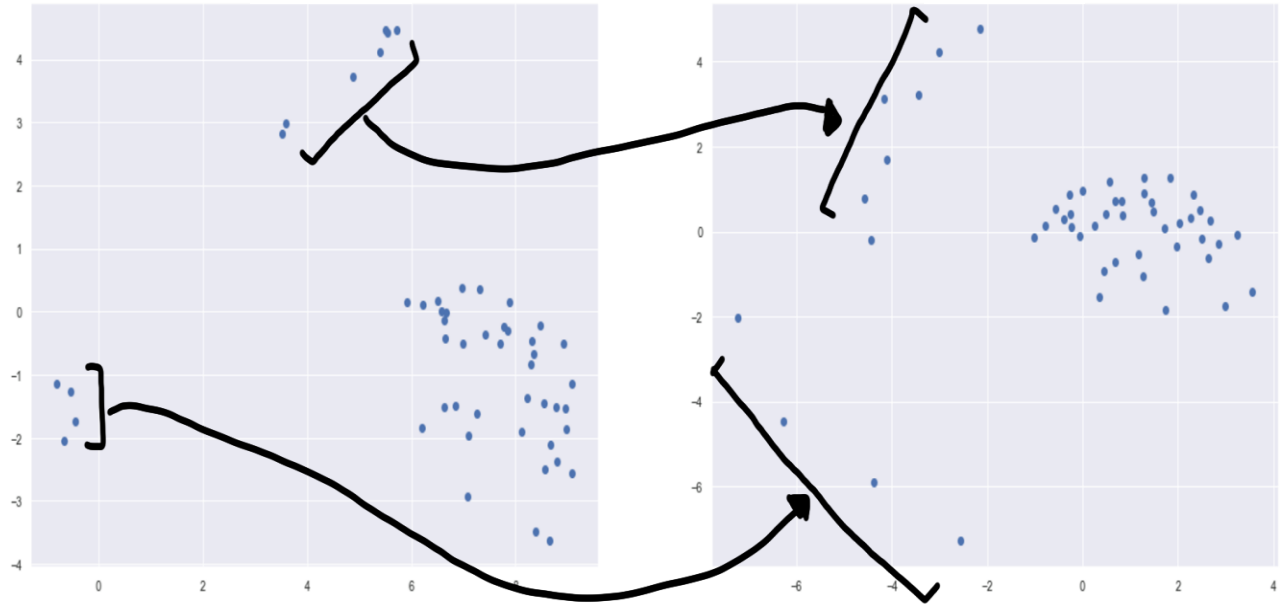
HDBSCAN (2/8)

1 – Choix d'une nouvelle métrique

Objectif : Obtenir une meilleure **mesure de la proximité** des points.

Idée : « **Eloigner** » les observations dans les espaces clairsemés sans modifier les zones denses.

Comment : Créer la **distance d'accessibilité mutuelle** $d_{mutual\ reach}$.



$$d_{mutual\ reach}(a, b) = \max(d_{core}(a), d_{core}(b), d_{euclid}(a, b))$$

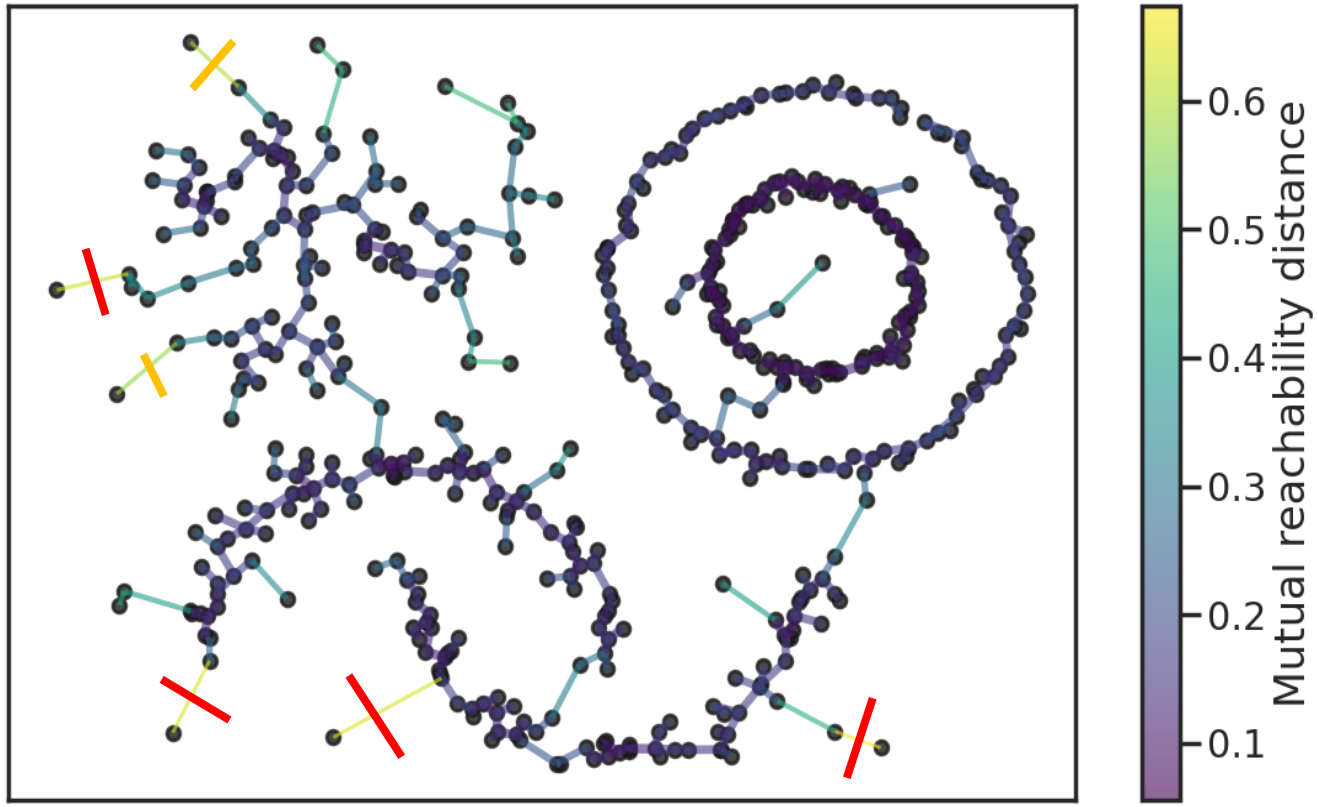
Méthodes de Clustering

HDBSCAN (3/8)

2 – Mise en place d'un arbre couvrant minimal (Minimum Spanning Tree)

Objectif : **Connecter l'ensemble des points** en minimisant la somme des distances d'accessibilité mutuelles.

Comment : Chaque nœud du graphe est un point d'observation et **chaque arête est égale à la distance** d'accessibilité mutuelle entre les points.

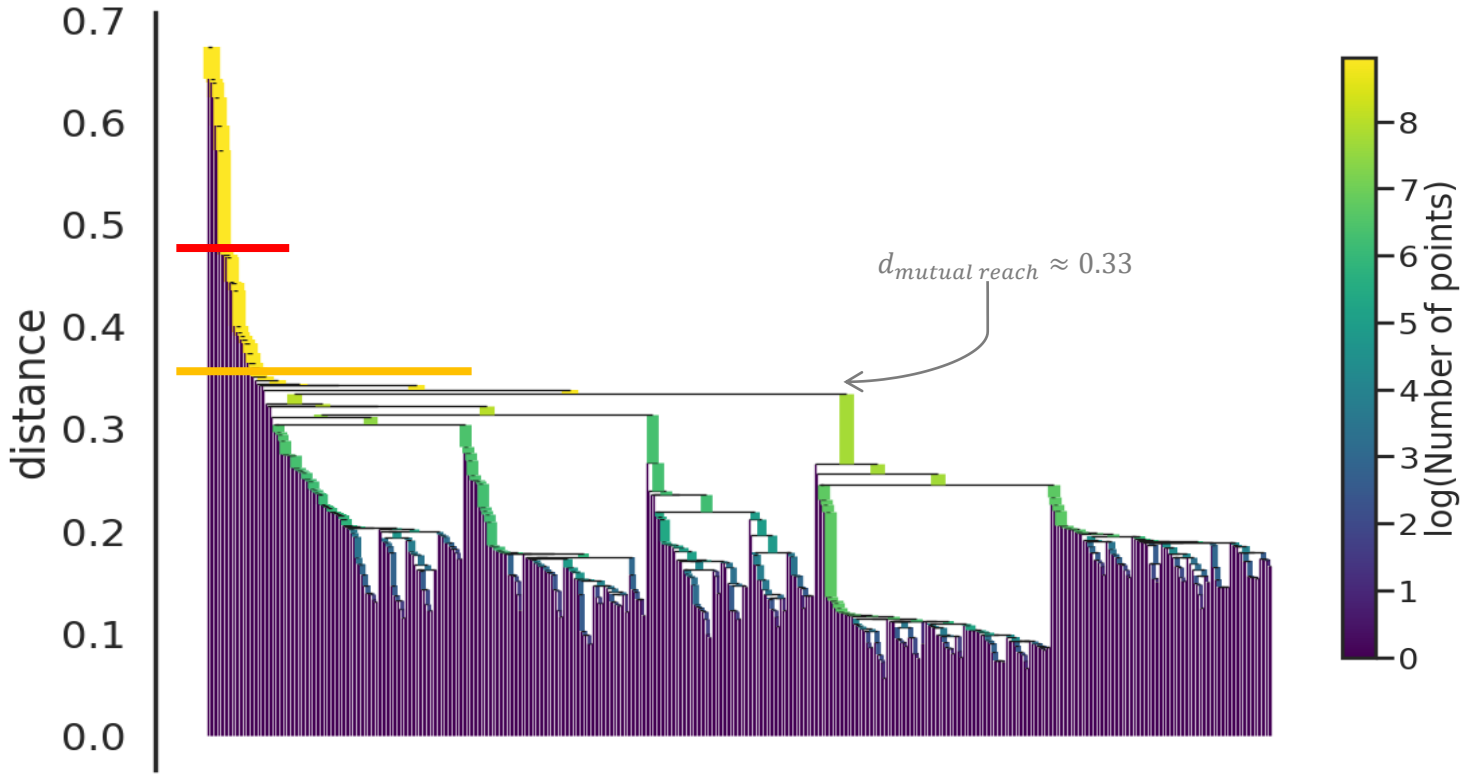


Méthodes de Clustering

HDBSCAN (4/8)

3 – Visualisation des divisions de l'arbre couvrant minimal

Obtenu en « élaguant » l'arbre minimum couvrant en coupant les arêtes par ordre décroissant de $d_{mutual\ reach}$.



Méthodes de Clustering

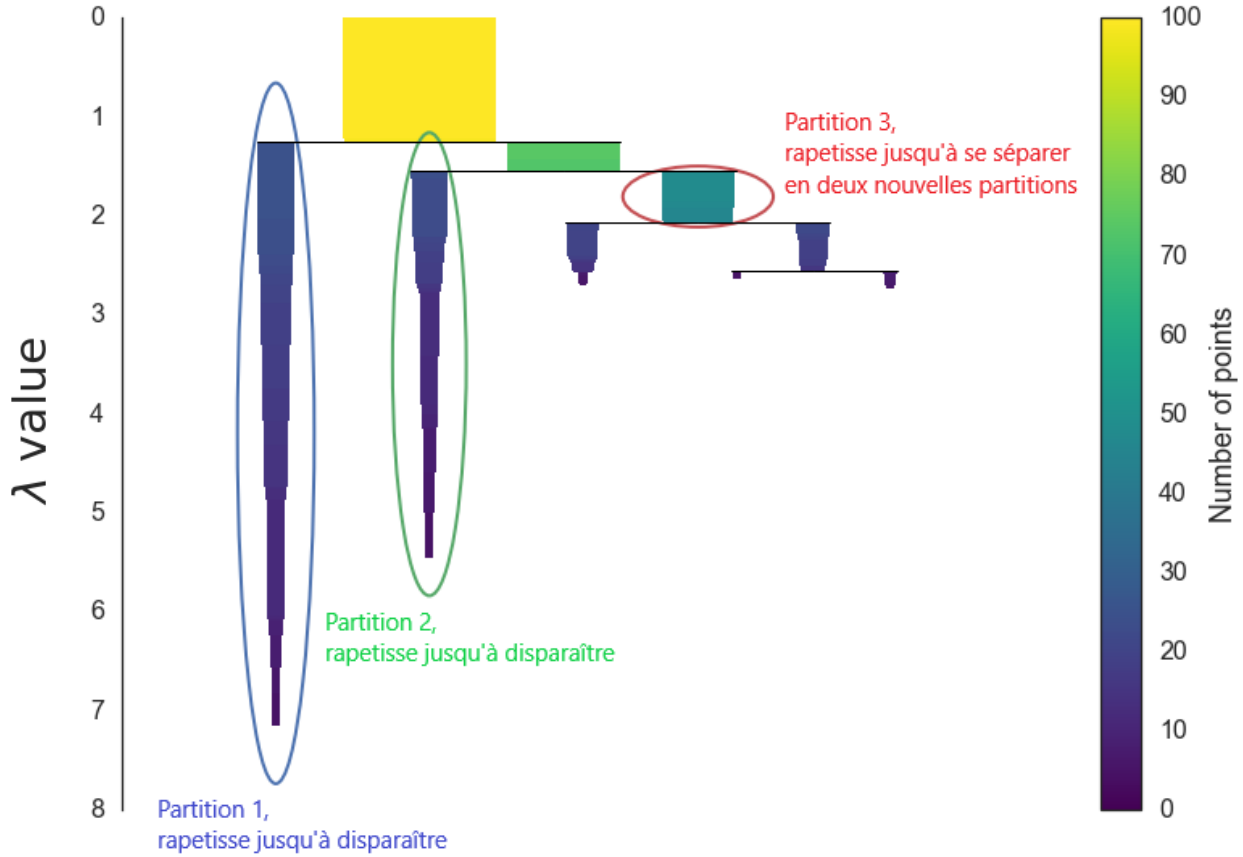
HDBSCAN (5/8)

4 – Création des clusters

Objectif : **Simplifier le dendrogramme** en « lissant » les clusters.

Comment :

- Considérer uniquement les clusters avec **au moins $min\_cluster\_size$ observations**.
- « Retourner » le dendrogramme en calculant $\lambda = \frac{1}{d_{mutual\ reach}}$



Méthodes de Clustering

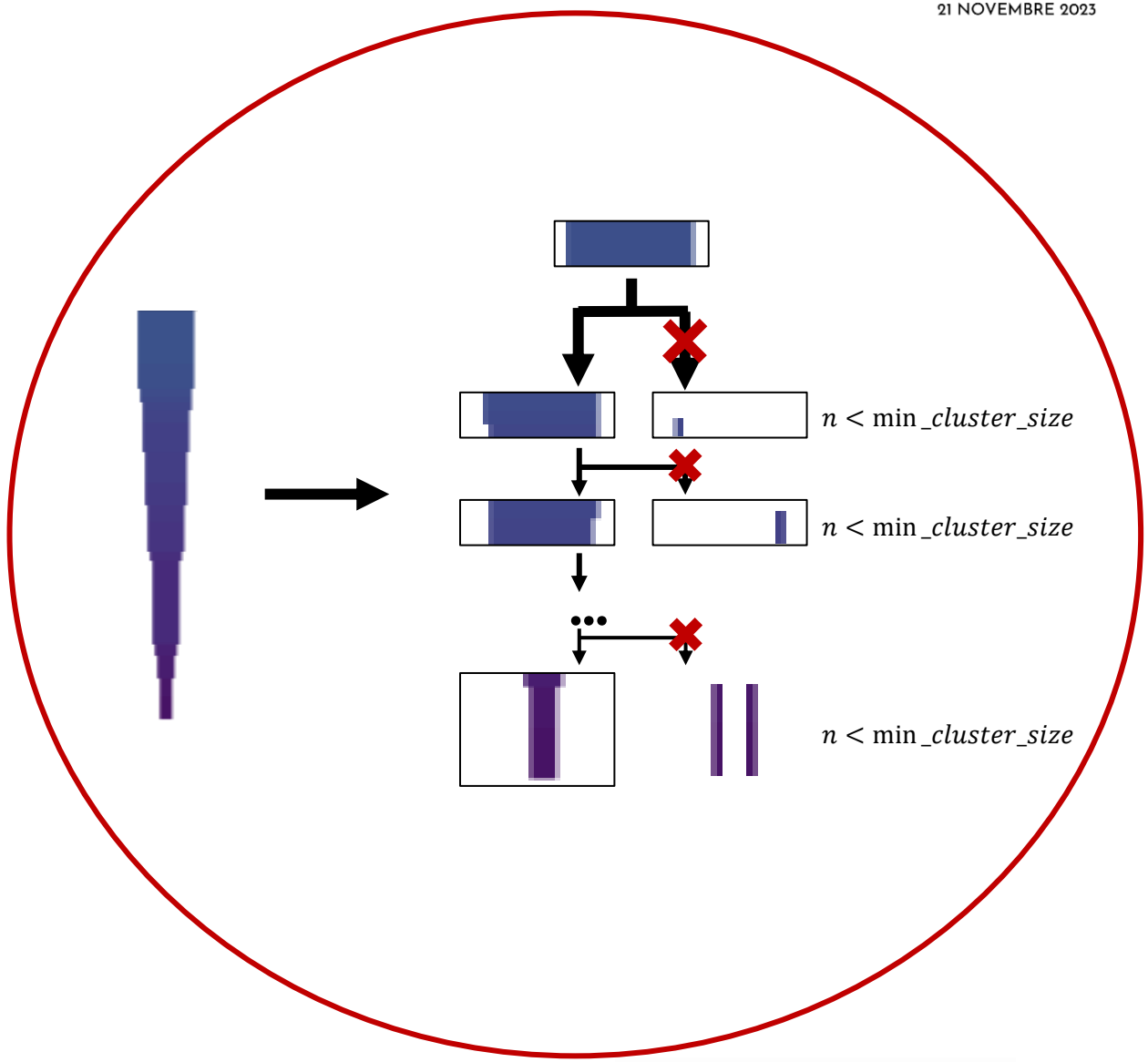
HDBSCAN (6/8)

4 – Création des clusters

Objectif : **Simplifier le dendrogramme** en « lissant » les clusters.

Comment :

- Considérer uniquement les clusters avec **au moins $min\_cluster\_size$ observations**.
- « Retourner » le dendrogramme en calculant $\lambda = \frac{1}{d_{mutual\ reach}}$



Méthodes de Clustering

HDBSCAN (7/8)

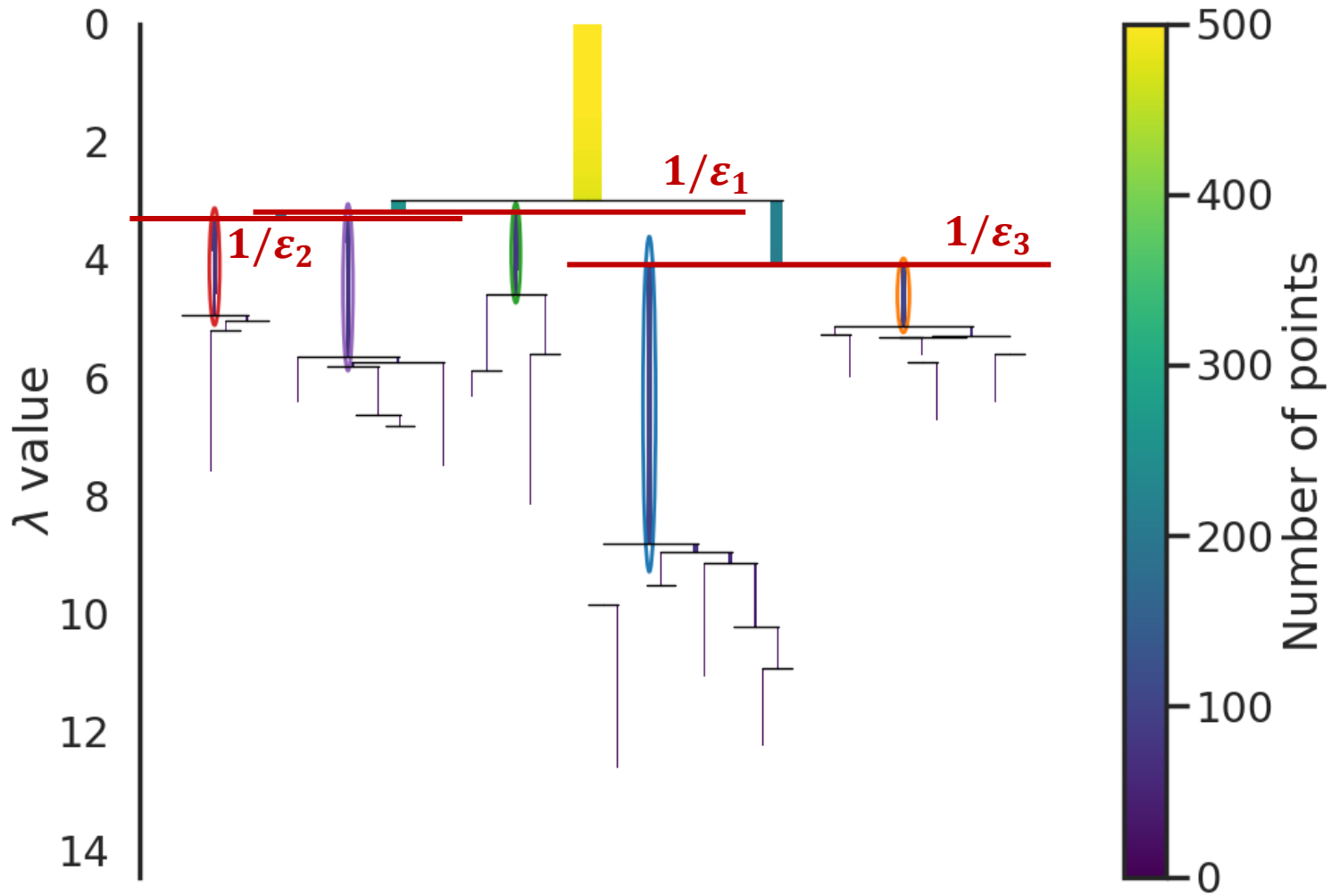
5 – Sélection des clusters

Comment :

- Définir la **stabilité** d'un cluster :

$$S(C) = \sum_{p \in C} (\lambda_p - \lambda_C)$$

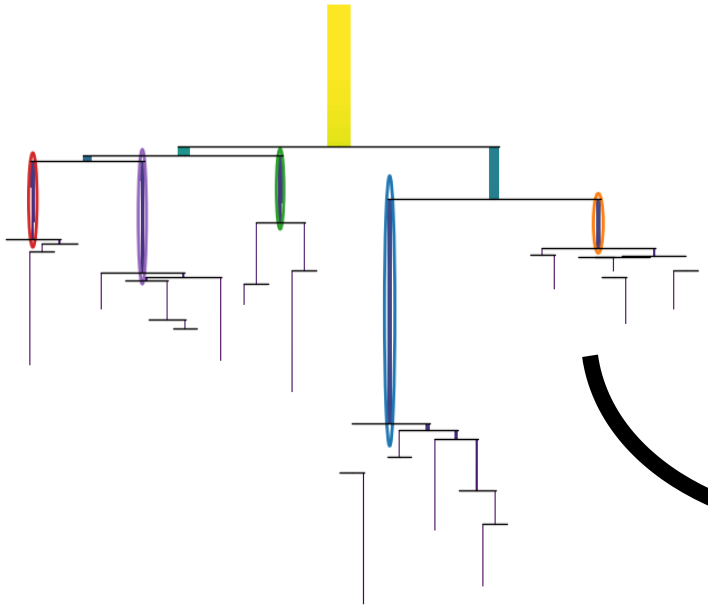
- Sélectionner les clusters **maximisant** $\sum_C S(C)$.

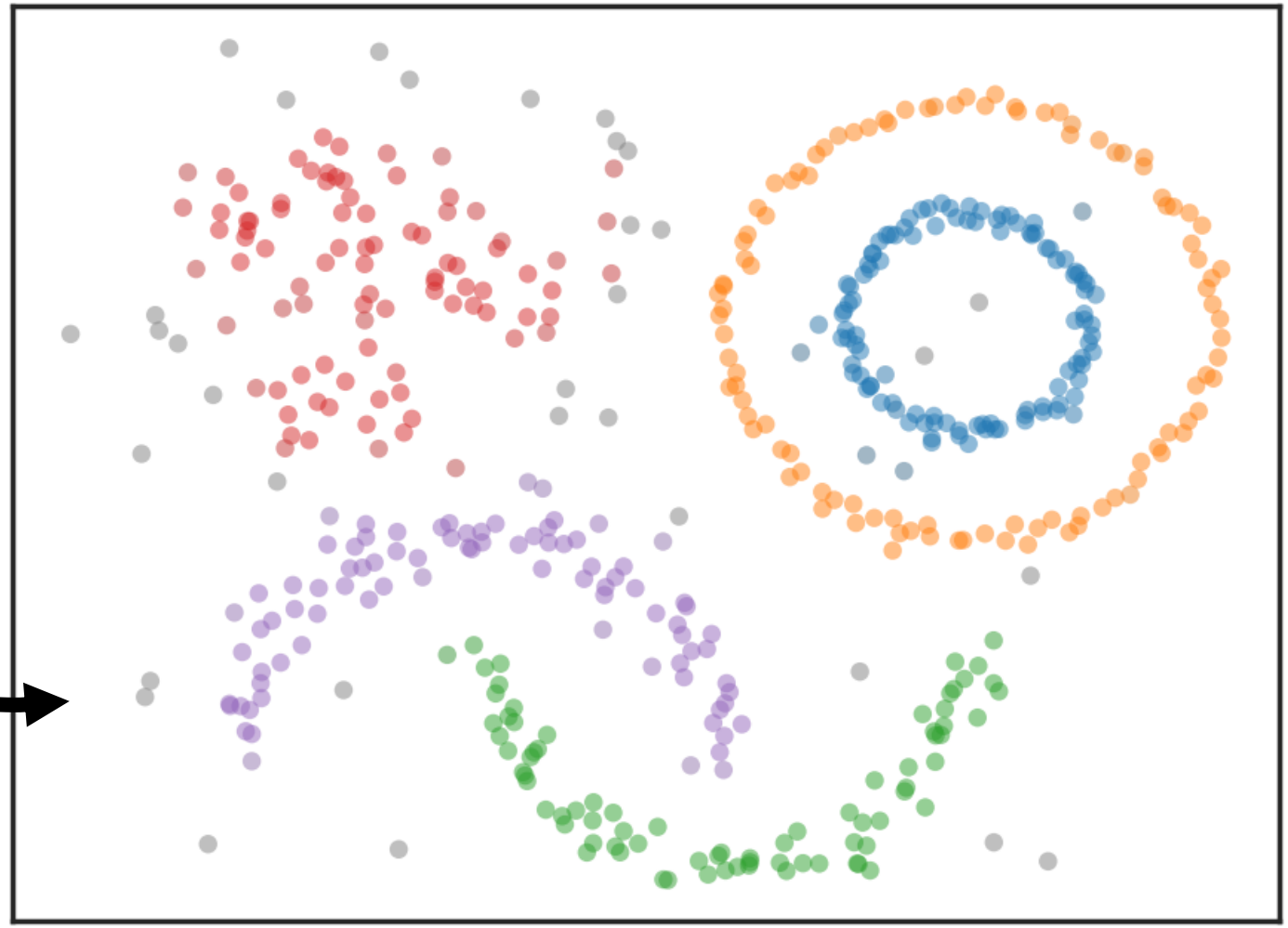


Méthodes de Clustering

HDBSCAN (8/8)

Résultats

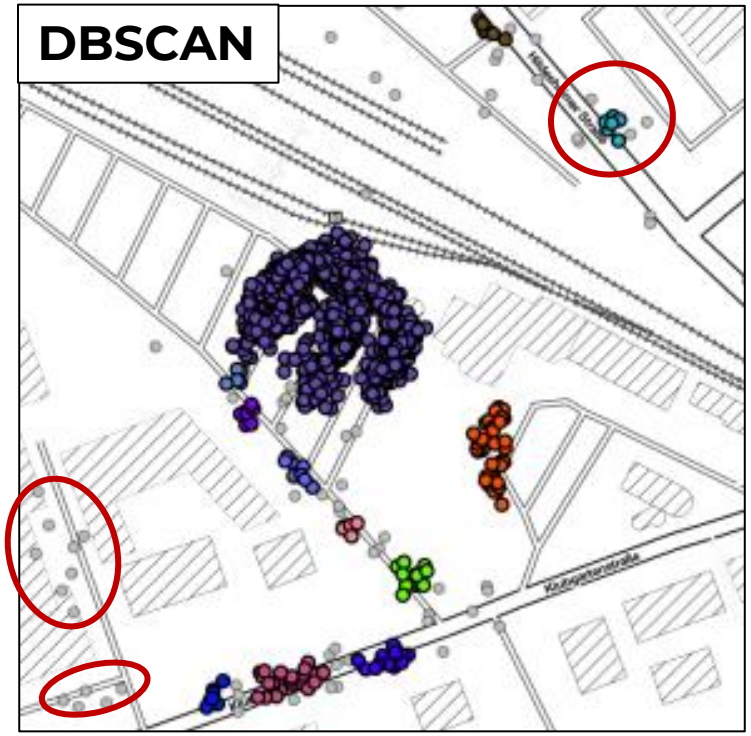


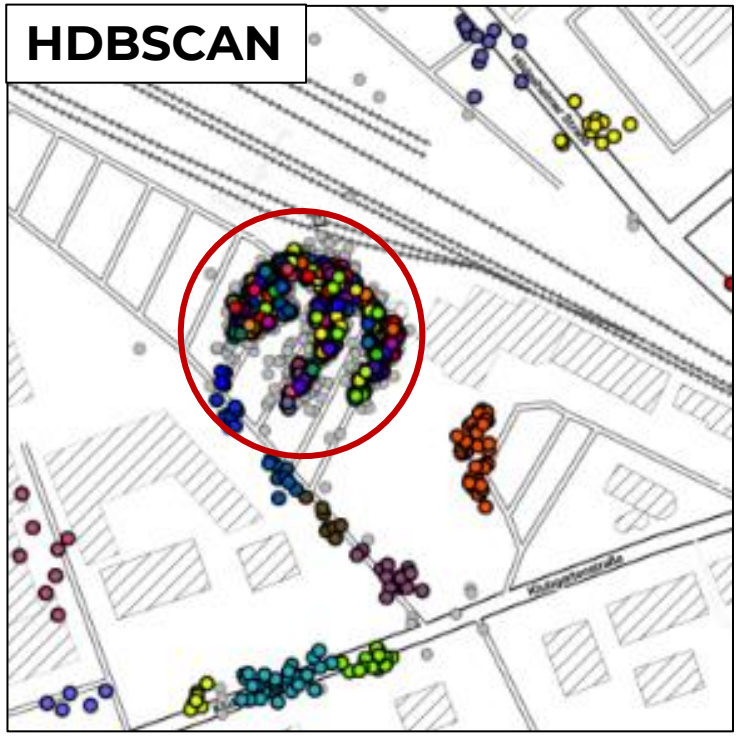


HDBSCAN & DBSCAN

Combinaison (1/3)

Pourquoi : Si la stabilité des petits clusters est trop importante, HDBSCAN en sélectionne trop.



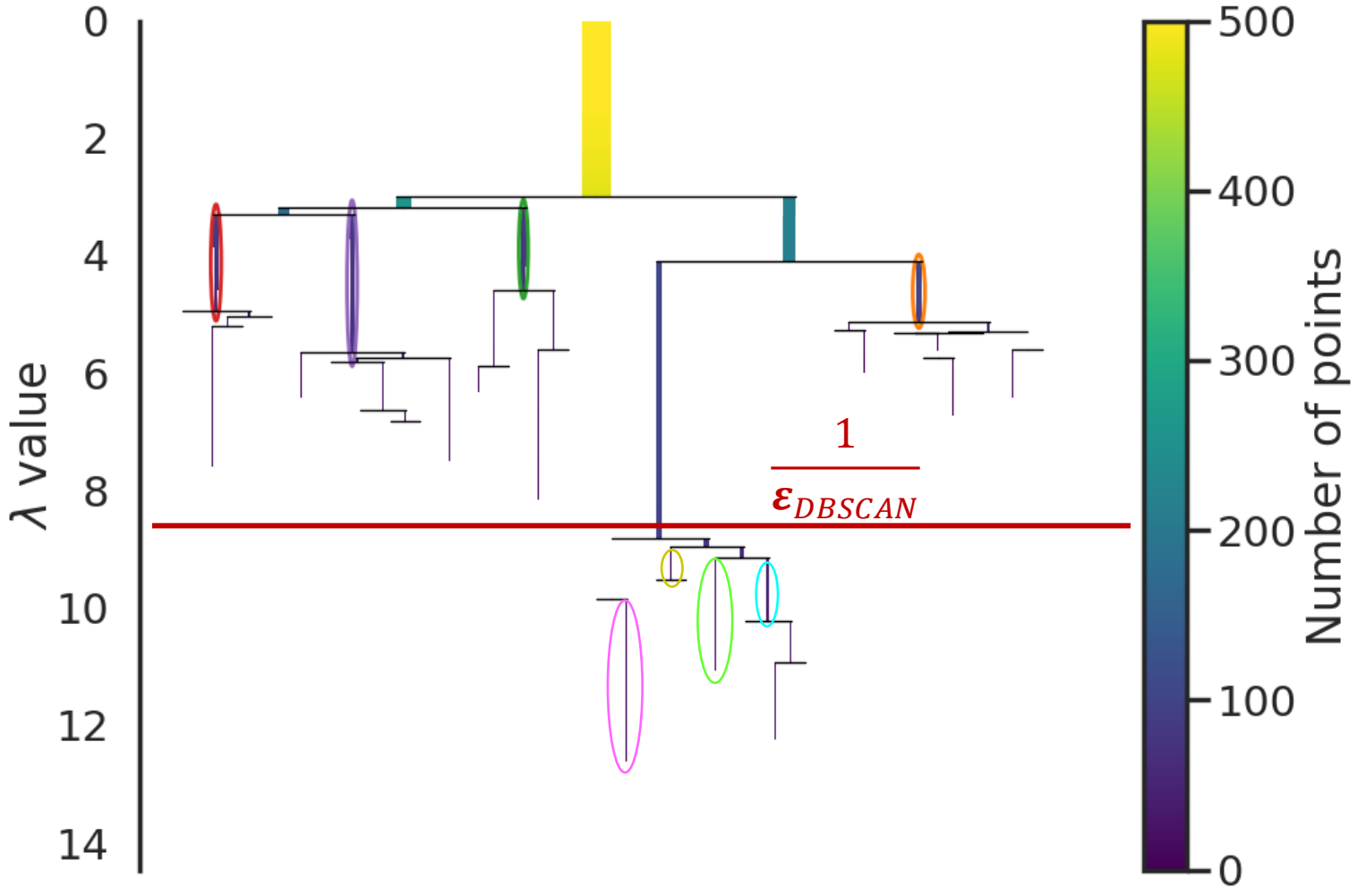


HDBSCAN & DBSCAN

Combinaison (2/3)

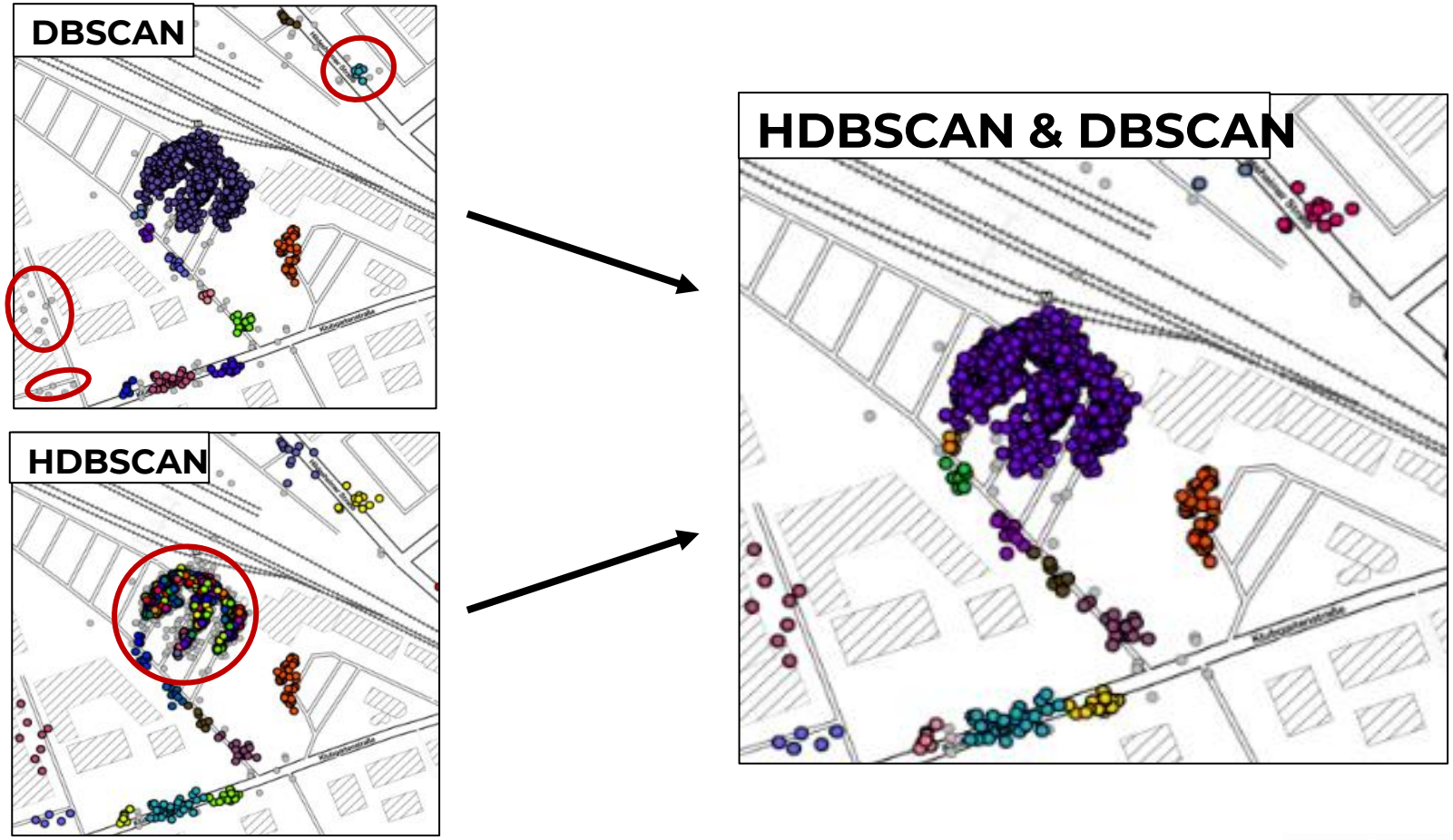
Pourquoi : Si la stabilité des petits clusters est trop importante, HDBSCAN en sélectionne trop.

Comment : Ajout d'un paramètre de seuil ε_{DBSCAN}.



HDBSCAN & DBSCAN

Combinaison (3/3)



Performance

1 – Score de performance (1/2)

Objectif : Trouver une métrique pour **évaluer nos clusters**.

Idée : Utiliser les **classifications d'experts**.

Comment : Calculer un score de performance entre les classifications d'experts et de HDBSCAN.

	Cluster 1	Cluster 2	Cluster 3	Total
Experte 1	20	5	0	25
Experte 2	3	2	50	55
Experte 3	0	15	2	17
Total	23	22	52	**97**

Matrice de croisée entre les classes de l'expert et les clusters

→ Score = 0.58

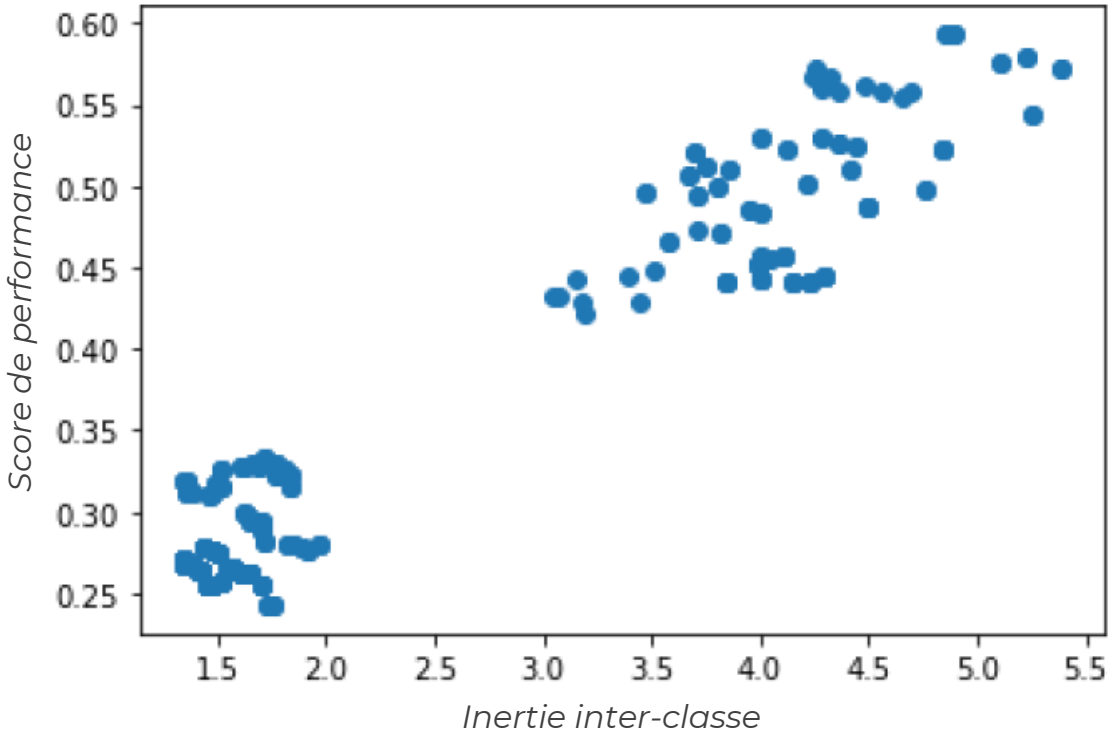
→ **Problème** : métrique coûteuse dans le cadre de travaux automatisés.

Performance

1 – Score de performance (2/2)

Objectif : Trouver une métrique pour **évaluer nos clusters**.

→ **Corrélation** entre le score de performance et l'inertie inter-classe.



Performance

2 – Inertie

Inertie Totale :

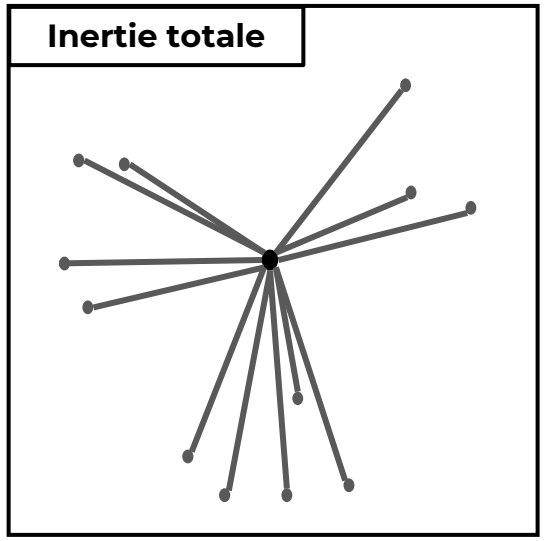
$$I_{tot} = \frac{1}{n}\sum_{i=1}^{n} d^2(e_i, g)$$

Inertie intra :

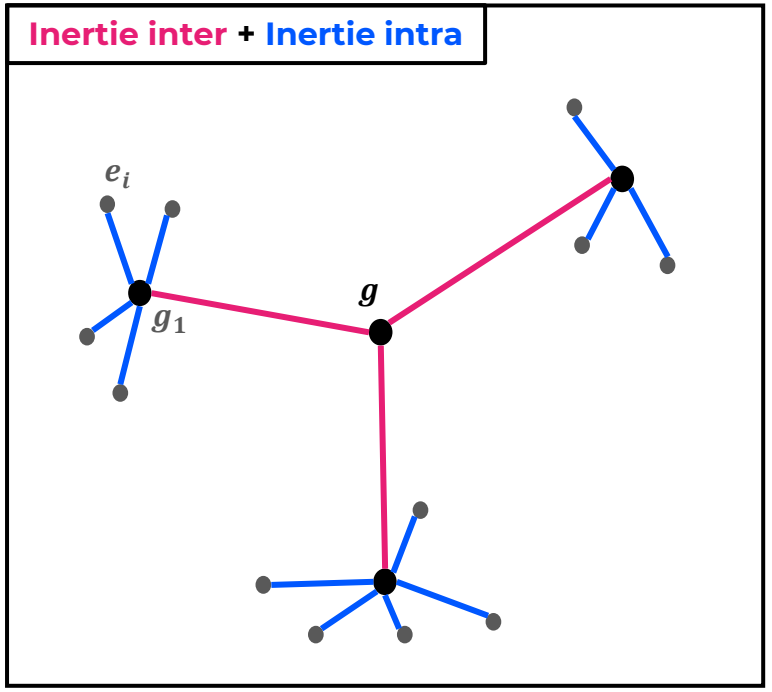
$$I_{intra} = \frac{1}{n}\sum_{i=1}^{k}\sum_{e \in g_i} d^2(e, g_i)$$

Inertie inter :

$$I_{inter} = \frac{1}{n}\sum_{i=1}^{k} n_i d^2(g_i, g)$$



$$I_{tot} = I_{intra} + I_{inter}$$



Un bon clustering maximise I_{inter} ou minimise I_{intra}.

→ **Objectif** : Maximiser I_{inter}, car corrélée avec le score de performance.

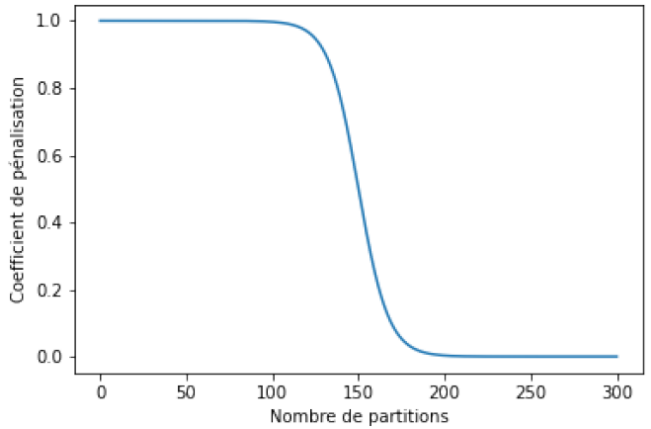
Performance

3 – Critères de pénalisation

$$I_{tot} = I_{intra} + I_{inter}$$

→ **Problème** : I_{inter} est maximisé lorsque $I_{intra} = 0$: chaque cluster contient 1 point.

→ Ajouter un critère de pénalisation basé sur le **nombre de clusters** : $p_{cluster}(n_{cluster})$.

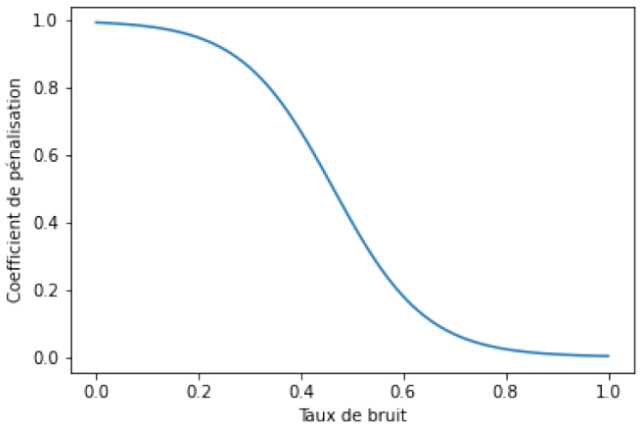


Evolution du facteur de pénalisation en fonction du nombre de clusters, Contrats UC

$$I_{tot} = \frac{1}{n}\sum_{i=1}^{n} d^2(e_i, g)$$

→ **Problème** : I_{tot} ne prend pas en compte les outliers.

→ Ajouter un critère de pénalisation basé sur le **taux d'outliers** : $p_{outlier}(t_{outlier})$.



Evolution du facteur de pénalisation en fonction du taux d'outliers, Contrats UC

Performance

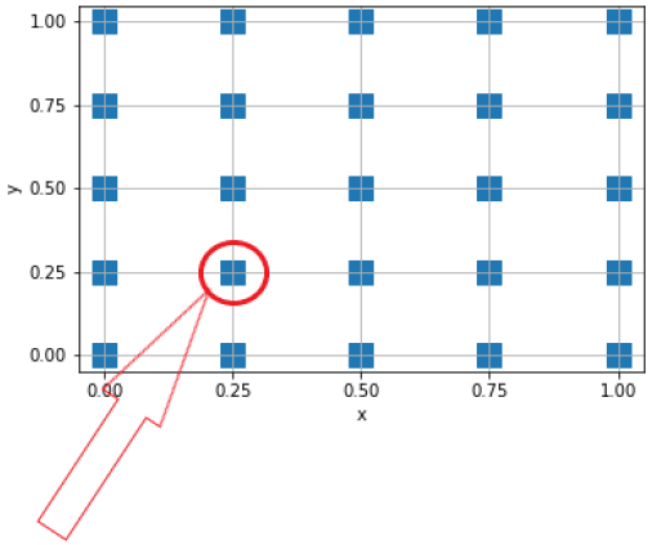
4 – Optimisation

Objectif : Maximiser le **critère d'optimisation** : $c = p_{cluster}(n_{cluster}) * p_{outlier}(t_{outlier}) * I_{inter}$

Comment : Rechercher les hyper-paramètres optimaux de notre modèle exhaustivement par **grid-search**.

Hyper-paramètres : HDBSCAN possède 3 hyper-paramètres déterminants :

- **Min samples** : Détermine la prédisposition à classer des points comme outlier en jouant sur la core_distance.
- **Min cluster size** : Détermine la taille minimale d'un cluster.
- **Epsilon** : Détermine la prédisposition à fusionner les clusters de faible cardinalité ensemble.

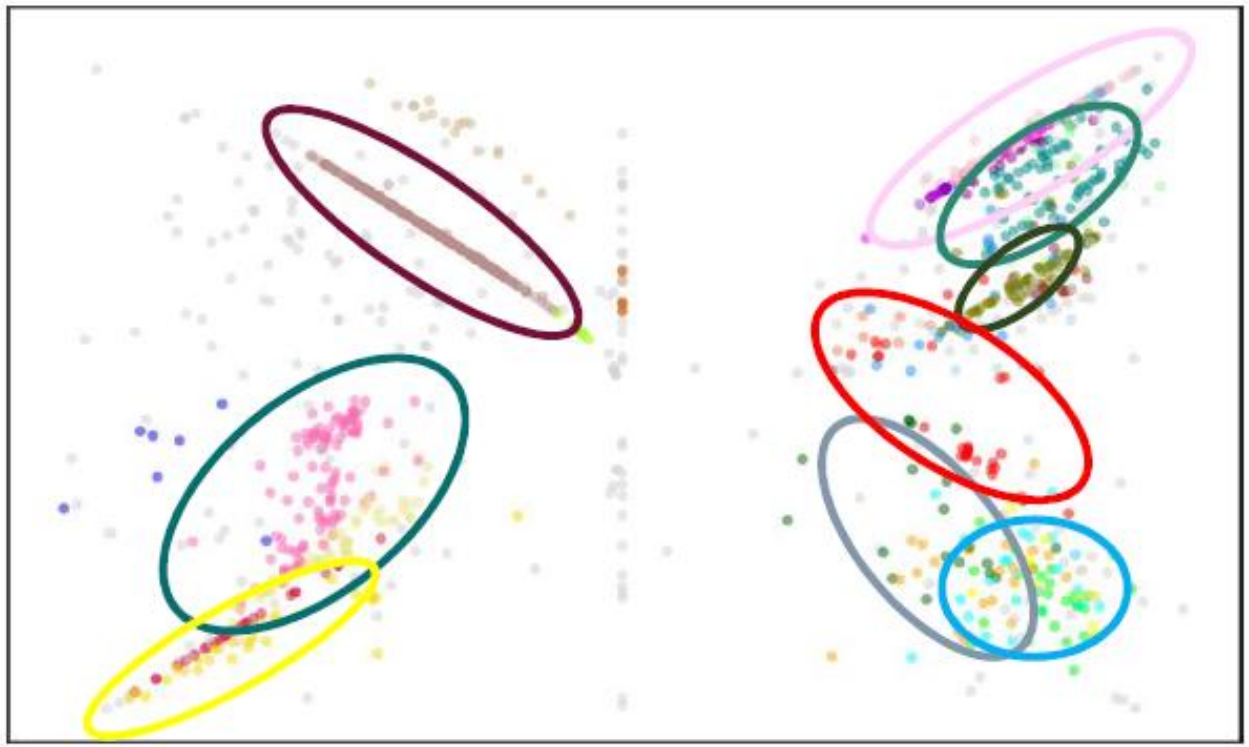


Traitement des anomalies

5 – Résultats

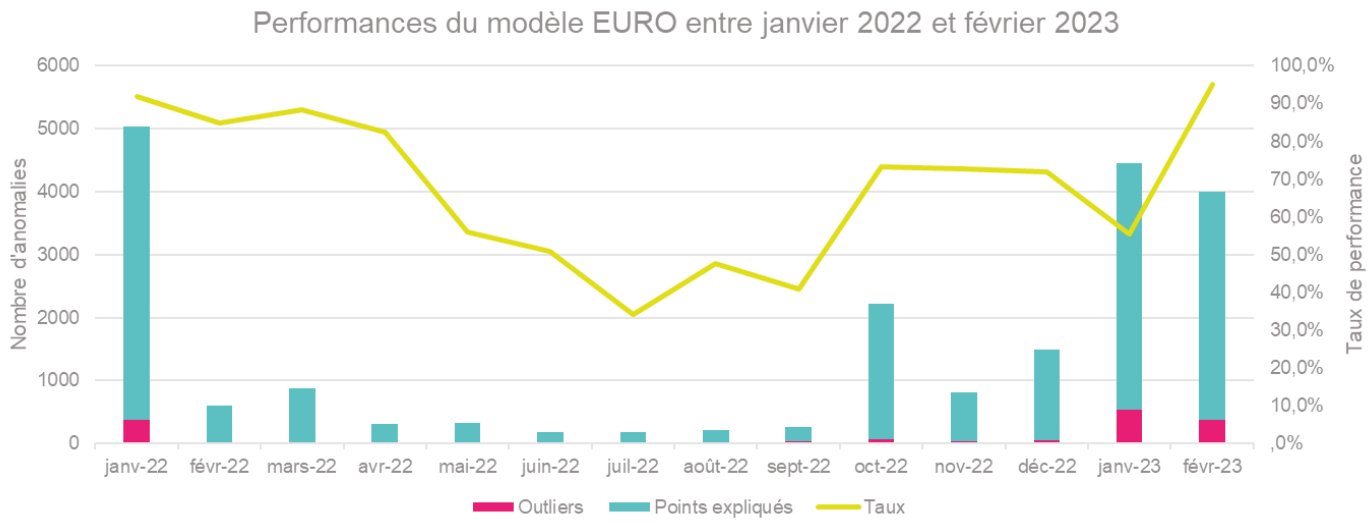
Obtention de **groupes d'anomalies identifiables**.

Les outliers sont des **anomalies non-catégorisées**.



Traitement des anomalies

Performances du partitionnement (EURO)



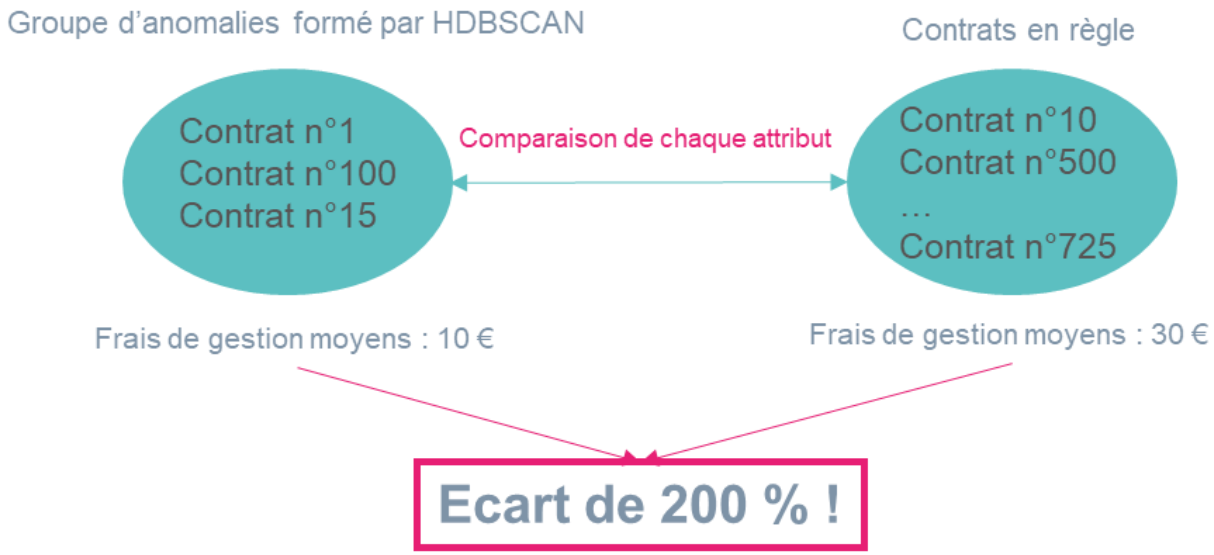
- Les **performances** sont **élevées** quand le nombre d'anomalies est conséquent.
- Faible part de points considérés comme **bruit**.
- **Faibles performances** entre mai et septembre 2022.

Traitement des anomalies

Comment déterminer les variables explicatives ?

- On calcule **l'écart entre plusieurs statistiques** des partitions par rapport à celles de la base de données des contrats en règle.

- Les variables correspondant aux **écarts les plus importants** sont considérées comme étant **explicatives** selon la méthode utilisée.

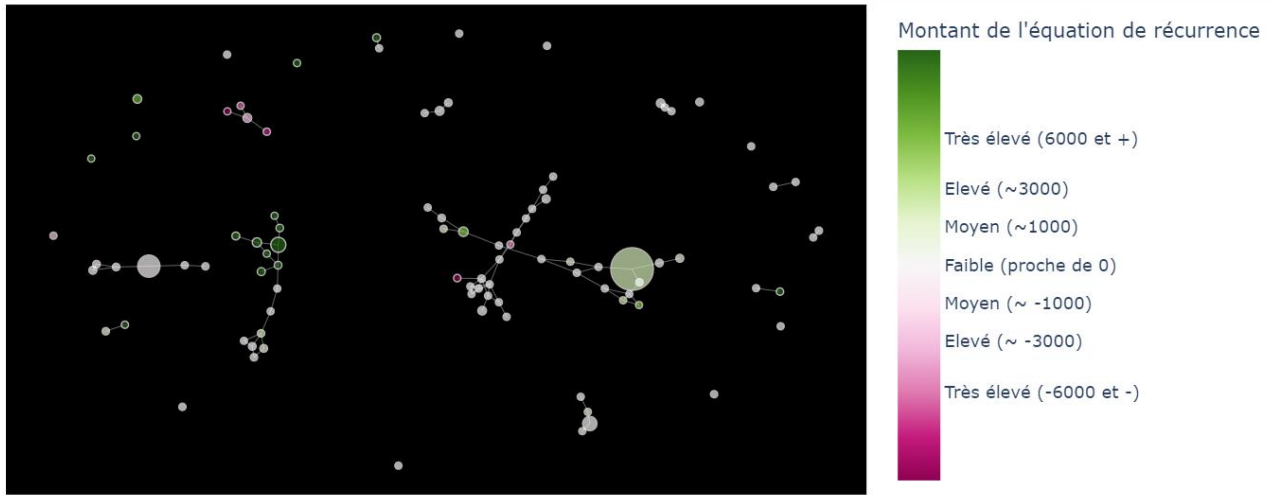


Traitement des anomalies

Visualisation du partitionnement

Beaucoup de clusters sont **liées** dans le graphe :

→ Leurs anomalies ont de fortes chances d'être **similaires**.



Remarque :
Les variables explicatives ne sont pas tout le temps **fiables**.
→ Variables retournées ne possèdent parfois pas d'**écart significatif** apparent.

Obtenir des clusters pertinents nécessite pour le moment **beaucoup de ressources**.
→ **96 cœurs** de calcul utilisés en parallèle durant la phase d'optimisation.

Conclusion

- **Validation** de l'algorithme d'optimisation du partitionnement.
- Mise en place du modèle en **production**.
- Perspective d'utilisation d'une **base de données plus détaillée** dite enrichie.

Annexes

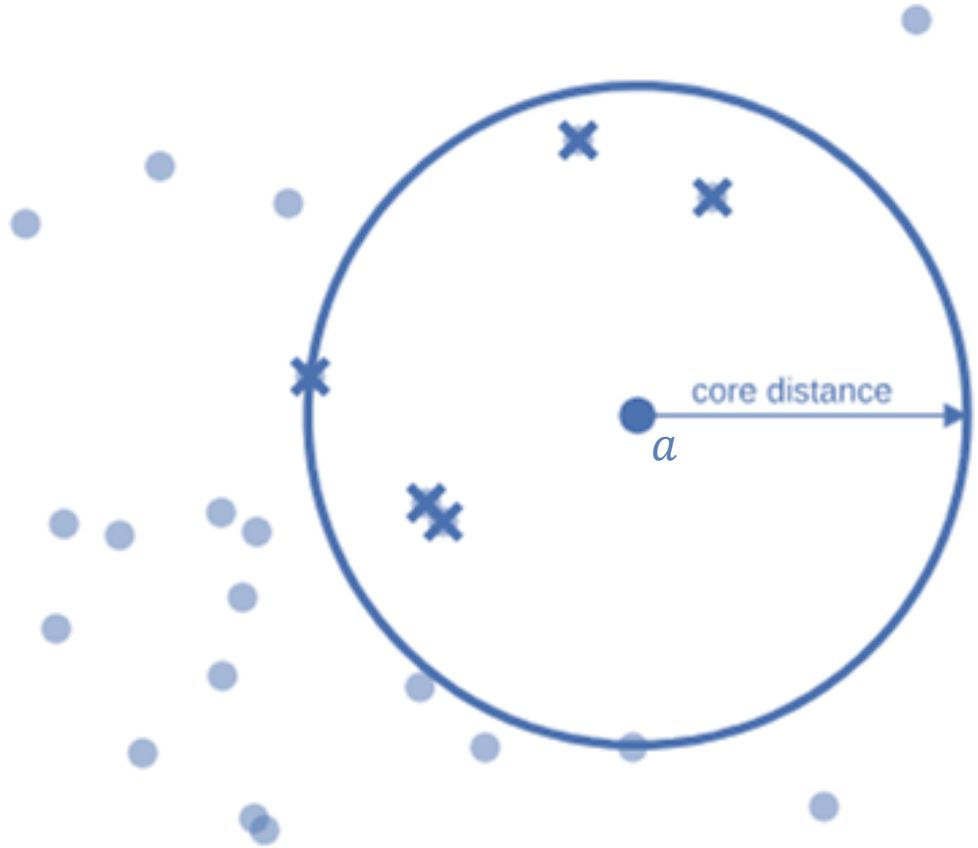
HDBSCAN

Choix d'une nouvelle métrique (1/4)

Objectif : Obtenir une meilleure **mesure de la proximité** des points.

Idée : « **Eloigner** » les observations dans les espaces clairsemés sans modifier les zones denses.

Comment : Créer la **distance d'accessibilité mutuelle** $d_{mutual\ reach}$.



La $core\_distance_a$ est le rayon d'un cercle contenant le point a et min_sample points (y compris a).

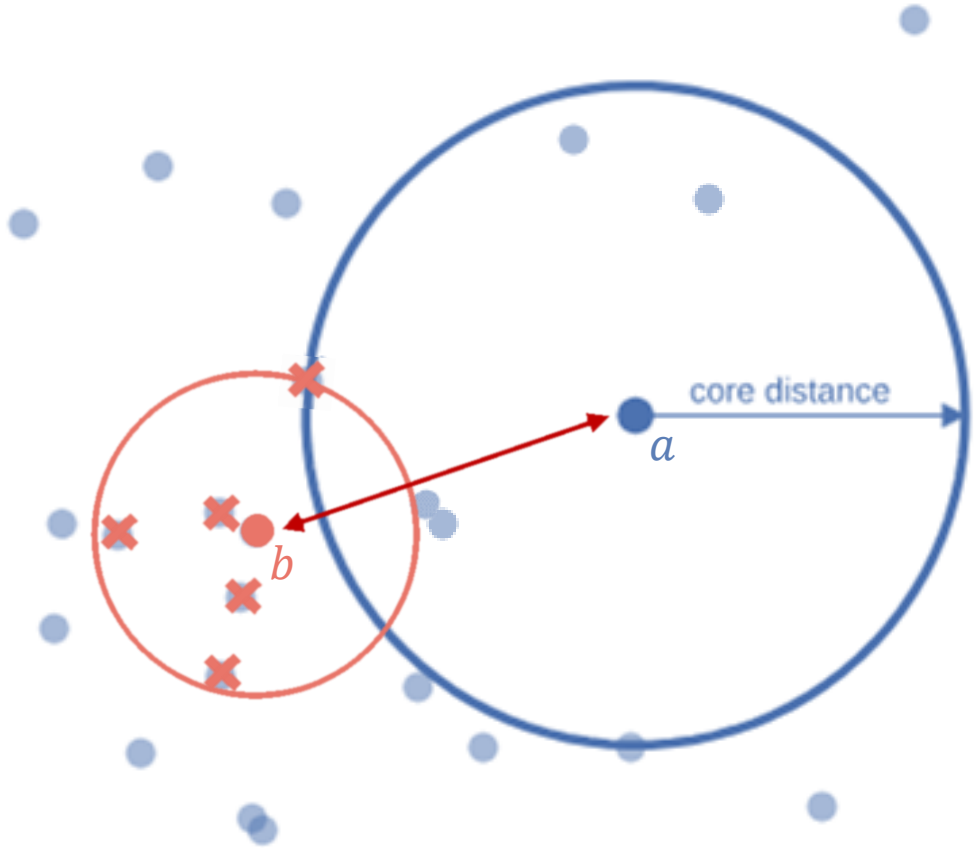
HDBSCAN

Choix d'une nouvelle métrique (2/4)

Objectif : Obtenir une meilleure **mesure de la proximité** des points.

Idée : « **Eloigner** » les observations dans les espaces clairsemés sans modifier les zones denses.

Comment : Créer la **distance d'accessibilité mutuelle** $d_{mutual\ reach}$.



$$\boldsymbol{d_{mutual\ reach}}(a, b) = \max(d_{core}(a), d_{core}(b), d_{euclid}(a, b))$$

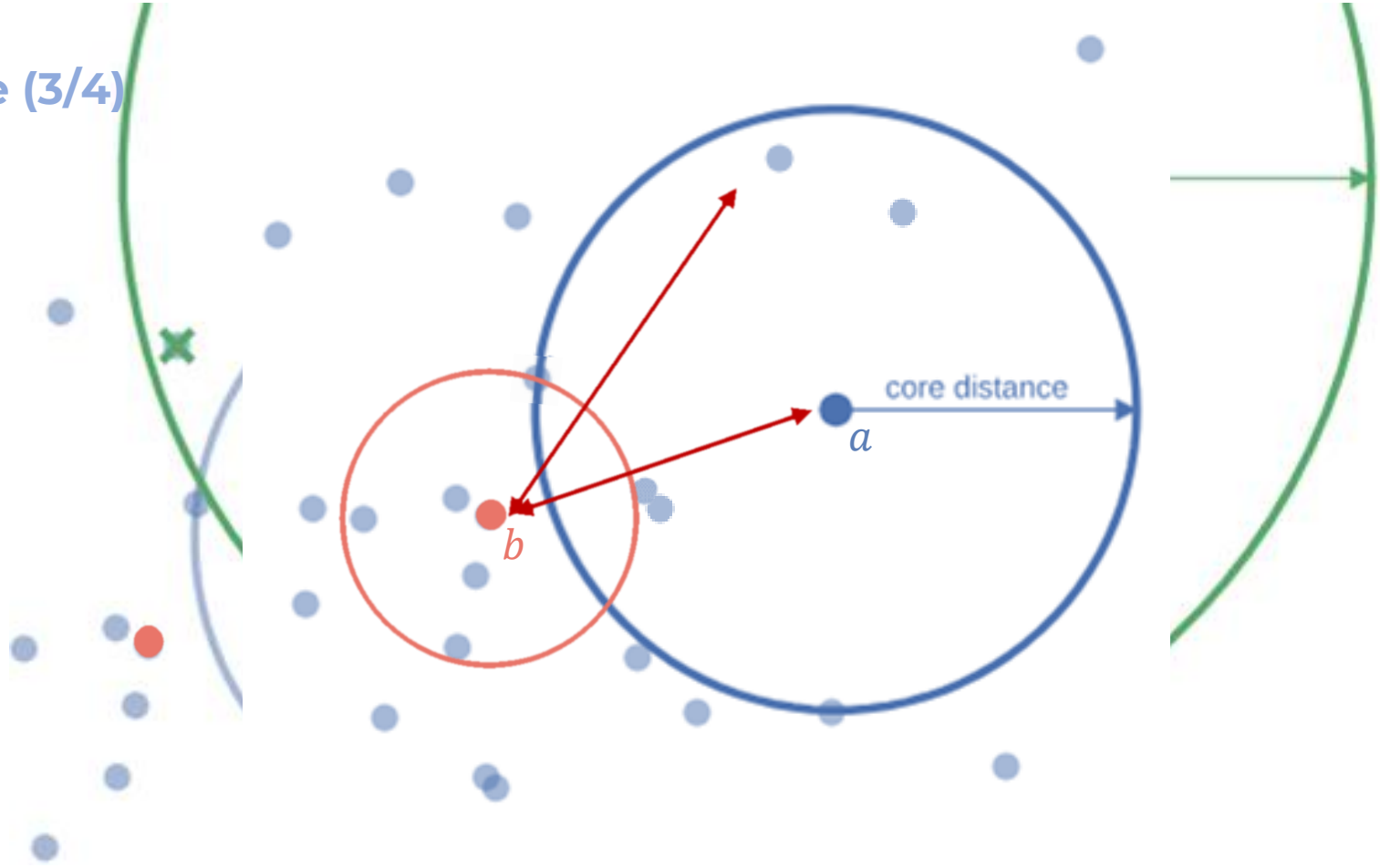
HDBSCAN

Choix d'une nouvelle métrique (3/4)

Objectif : Obtenir une meilleure **mesure de la proximité** des points.

Idée : « **Eloigner** » les observations dans les espaces clairsemés sans modifier les zones denses.

Comment : Créer la **distance d'accessibilité mutuelle** $d_{mutual\ reach}$.



$$d_{mutual\ reach}(a,b) = \max(d_{core}(a), d_{core}(b), d_{euclid}(a,b))$$

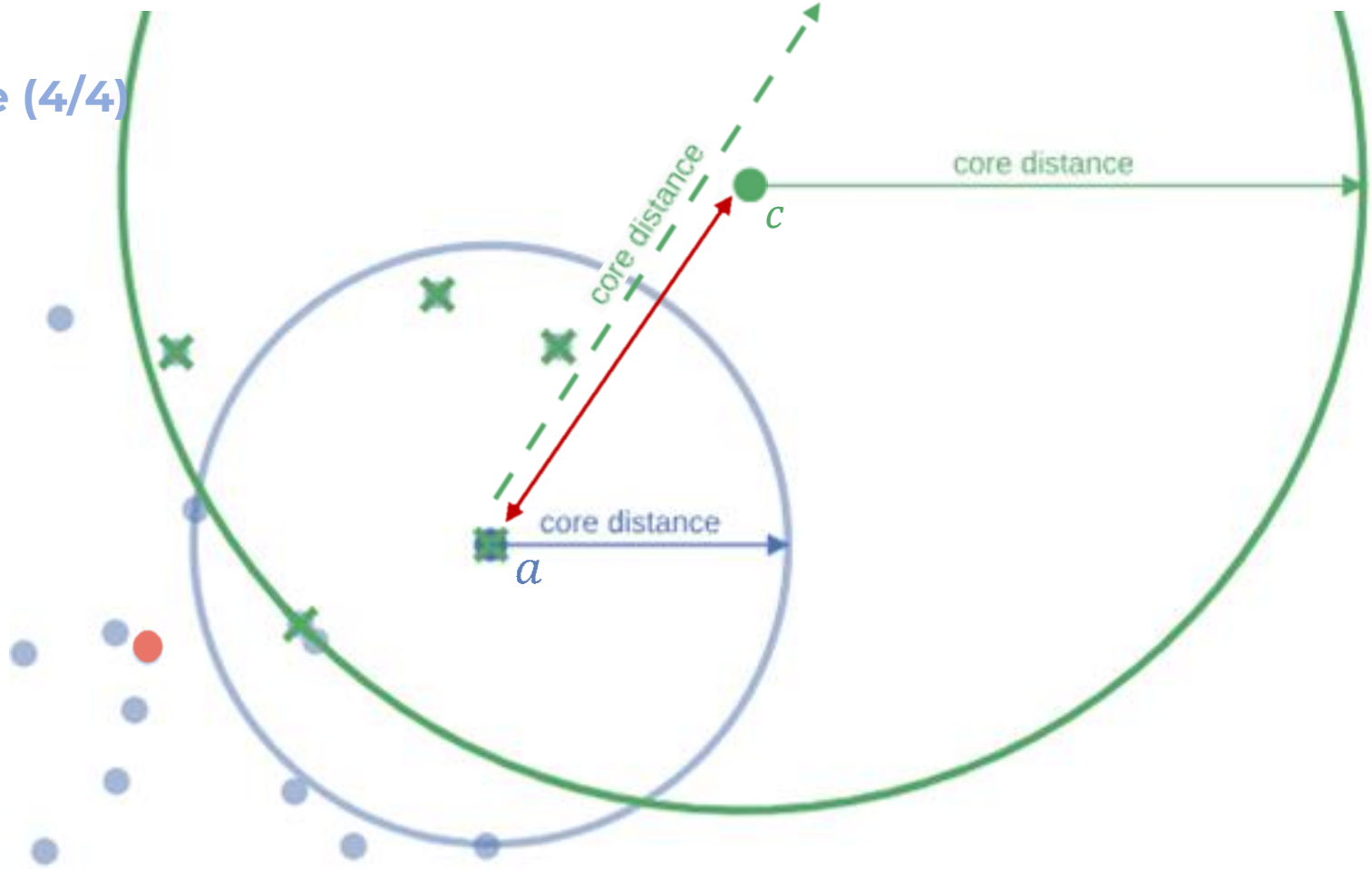
HDBSCAN

Choix d'une nouvelle métrique (4/4)

Objectif : Obtenir une meilleure **mesure de la proximité** des points.

Idée : « **Eloigner** » les observations dans les espaces clairsemés sans modifier les zones denses.

Comment : Créer la **distance d'accessibilité mutuelle** $d_{mutual\ reach}$.



$$d_{mutual\ reach}(a, c) = \max(d_{core}(a), d_{core}(c), d_{euclid}(a, c))$$