REGROUPEMENT

« La science des données ne remplace pas la modélisation statistique et l'analyse des données, elle les enrichit. »

(P. Boily)

« Les données ne sont pas des renseignements, les renseignements ne sont pas des connaissances, la connaissance n'est pas la compréhension, la compréhension n'est pas la sagesse. »

(attribué à Cliff Stoll dans Keeler's *Nothing to Hide: Privacy in the 21st Century*, 2006)

OBJECTIFS D'APPRENTISSAGE

Se familiariser avec les concepts élémentaires du regroupement et certains algorithmes courants.

Se familiariser avec une variante du regroupement de partition (k-moyennes).

Se familiariser avec les critères de validation d'un regroupement.

TABLE DES MATIÈRES

1. Étude de cas : OK Cupid
2. Fondements du regroupement
3. Algorithmes de regroupement
4. Validation d'un regroupement
5. Notes
6. Exemple : Iris

ÉTUDE DE CAS : DONNÉES D'OK CUPID

REGROUPEMENT

Trouver l'amour véritable au moyen de l'analyse des regroupements.

(K. Poulsen, *How a Math Genius Hacked OK Cupid to Find True Love*, WIRED)

CONTEXTE

Chris McKinlay, étudiant de 35 ans au doctorat en mathématiques à UCLA, recherchait en ligne un partenaire romantique sans trop de succès.

- Les algorithmes d'*OK Cupid* utilisent seulement les questions auxquelles les deux partenaires potentiels décident de répondre et les questions qu'il avait choisies (plus ou moins de manière aléatoire à ce point) n'étaient pas les plus utilisées.

Entre juin 2012 et décembre 2013,

- il a utilisé un échantillonnage statistique pour trouver les questions qui auraient importées au type de partenaire qu'il recherchait;
- il a créé un nouveau profil qui répondait seulement à ces questions;
- il a établi une correspondance seulement avec des femmes de L.A. avec lesquelles il avait des affinités.

PROCESSUS

Cette histoire procure un excellent exemple du processus d'exploration des données, du début à la fin :

1. **recueillir** des données;
2. recueillir **d'autres données**, **légèrement meilleures** et **différentes**;
3. recueillir **encore d'autres** données;
4. déterminer la technique d'exploration des données qui **conviendrait** aux renseignements recherchés (regroupement);
5. **valider** les résultats de l'analyse.

PROCESSUS

Cette histoire procure un excellent exemple du processus d'exploration des données, du début à la fin (suite) :

6. **examiner** les résultats et faire ressortir les résultats véritablement intéressants;
7. analyser **davantage** les résultats intéressants et utiliser ces résultats pour résoudre le problème original;
8. utiliser les données pour **améliorer les autres aspects** de son profil;
9. attendre et récolter les bénéfices de l'exploration des données?

MÉTHODOLOGIE ET RÉSULTATS

Il a utilisé le regroupement selon le mode k pour regrouper 20 000 femmes dans sept groupes statistiquement distincts, en fonction de leurs questions et de leurs réponses.

Il a validé le regroupement au moyen de 5 000 autres profils tirés du site.

Il a analysé les regroupements pour en repérer deux qu'il trouvait intéressants :

- les femmes dans la vingtaine qui étaient indépendantes, musiciennes et artistes;
- les femmes un peu plus âgées qui occupaient des emplois professionnels de création, comme des éditrices ou des designers.

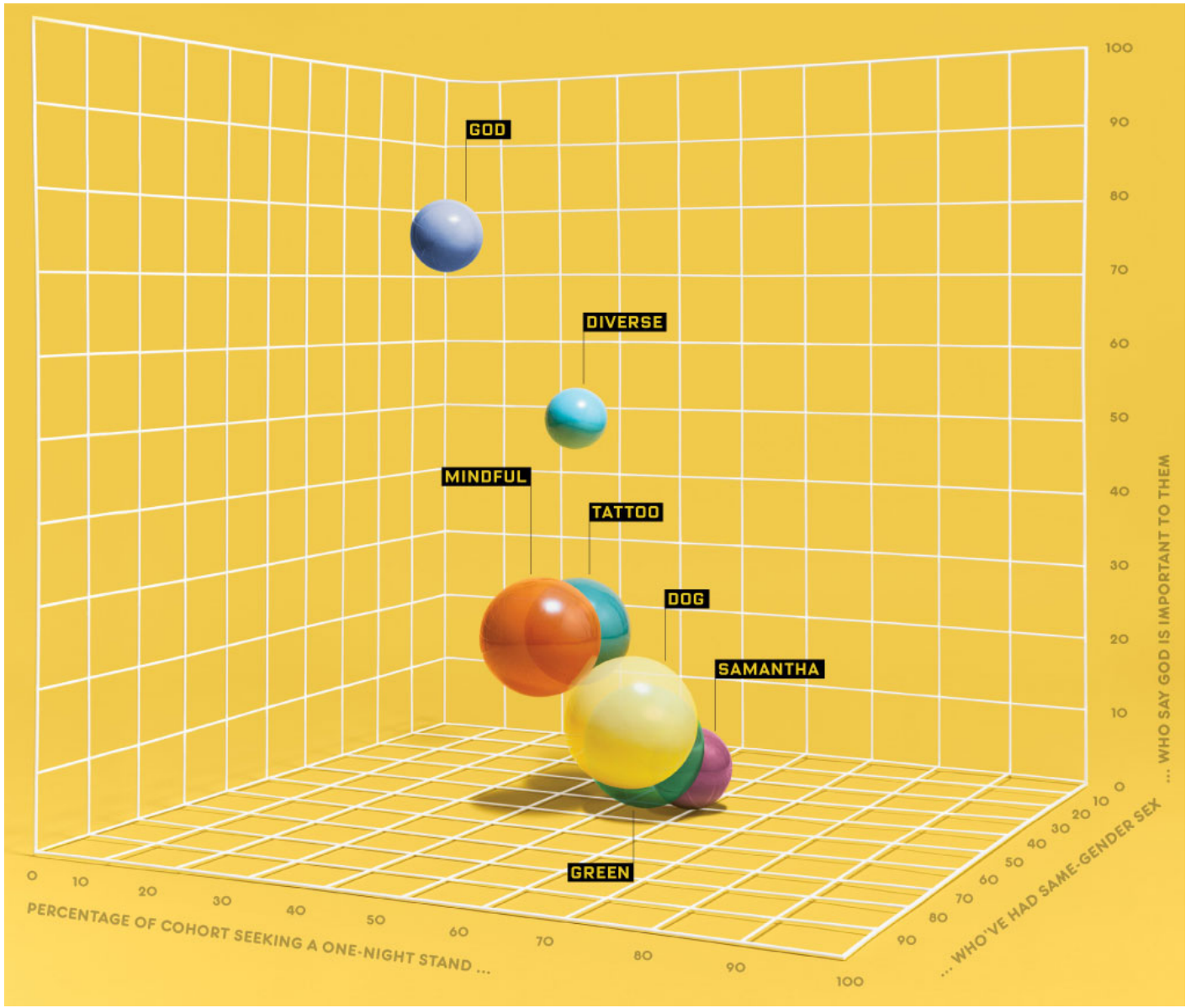
MÉTHODOLOGIE ET RÉSULTATS

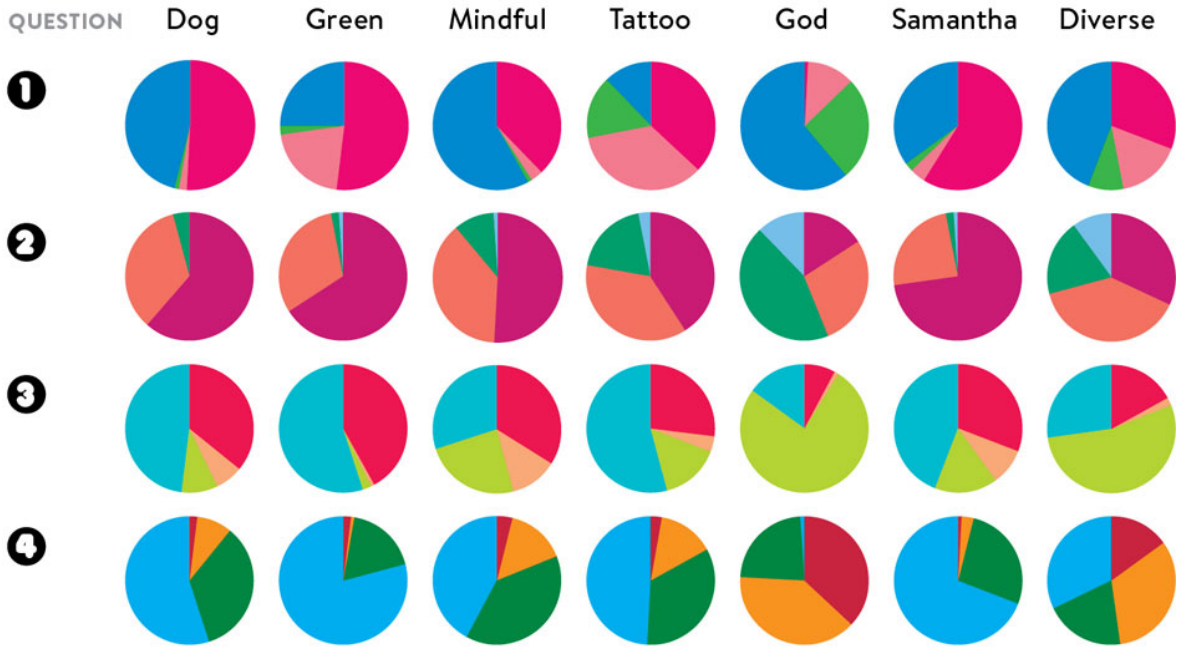
Il a utilisé le boosting adaptatif (algorithme d'apprentissage machine) pour trouver les questions auxquelles il devrait répondre dans son profil.

Un plus grand nombre de femmes correspondaient à son profil, ce qui a mené à un plus grand nombre de premières rencontres, quelques deuxièmes rencontres et une seule troisième rencontre.

À la fin, une femme a communiqué avec lui parce qu'elle était intriguée par son profil.

Elle lui a demandé de le rencontrer et ils vivaient ensemble lors de la rédaction de l'article.


GOD
DIVERSE
MINDFUL
TATTOO
DOG
SAMANTHA
GREEN
0 10 20 30 40 50 60 70 80 90 100
PERCENTAGE OF COHORT SEEKING A ONE-NIGHT STAND ...
0 10 20 30 40 50 60 70 80 90 100
... WHO'VE HAD SAME-GENDER SEX
0 10 20 30 40 50 60 70 80 90 100
... WHO SAY GOD IS IMPORTANT TO THEM



1. About how long do you want your next relationship to last?

- One night
- A few months to a year
- Several years
- The rest of my life

2. Say you've started seeing someone you really like. As far as you're concerned, how long will it take before you have sex?

- 1-2 dates
- 3-5 dates
- 6 or more dates
- Only after the wedding

3. Have you ever had a sexual encounter with someone of the same sex?

- Yes, and I enjoyed myself
- Yes, and I did not enjoy myself
- No, and I would never
- No, but I'd like to

4. How important is religion/God in your life?

- Extremely important
- Somewhat important
- Not very important
- Not important at all

DISCUSSION

Que pensez-vous de cette utilisation de l'apprentissage machine?

FONDEMENTS DU REGROUPEMENT

REGROUPEMENT

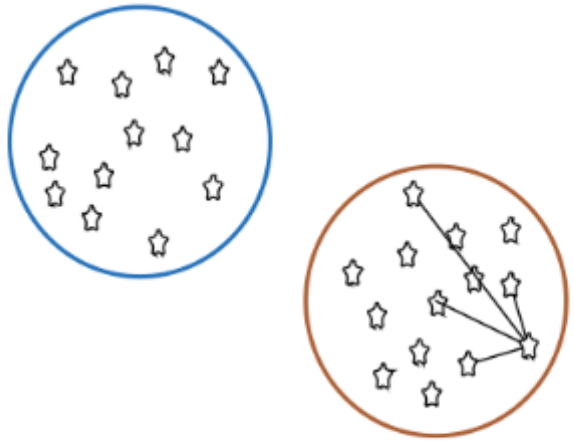
« La Voie lactée n'est qu'un amas d'innombrables étoiles. »
(Galileo Galilei, *Sidereus Nuncius*)

APERÇU DU REGROUPEMENT

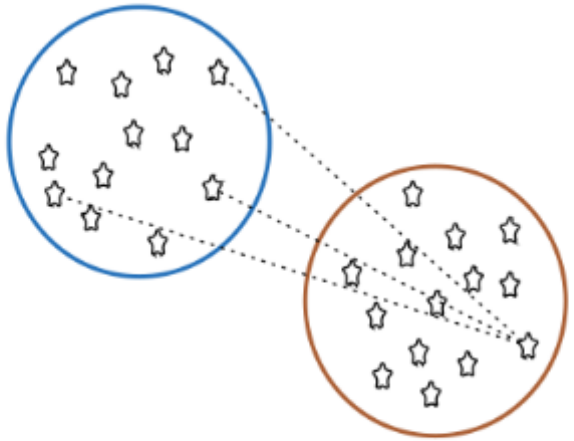
Dans un **regroupement**, les données sont réparties en **groupes formés naturellement**. Dans chaque groupe, les points de données sont **similaires**; d'un groupe à un autre, les points de données sont **distincts**.

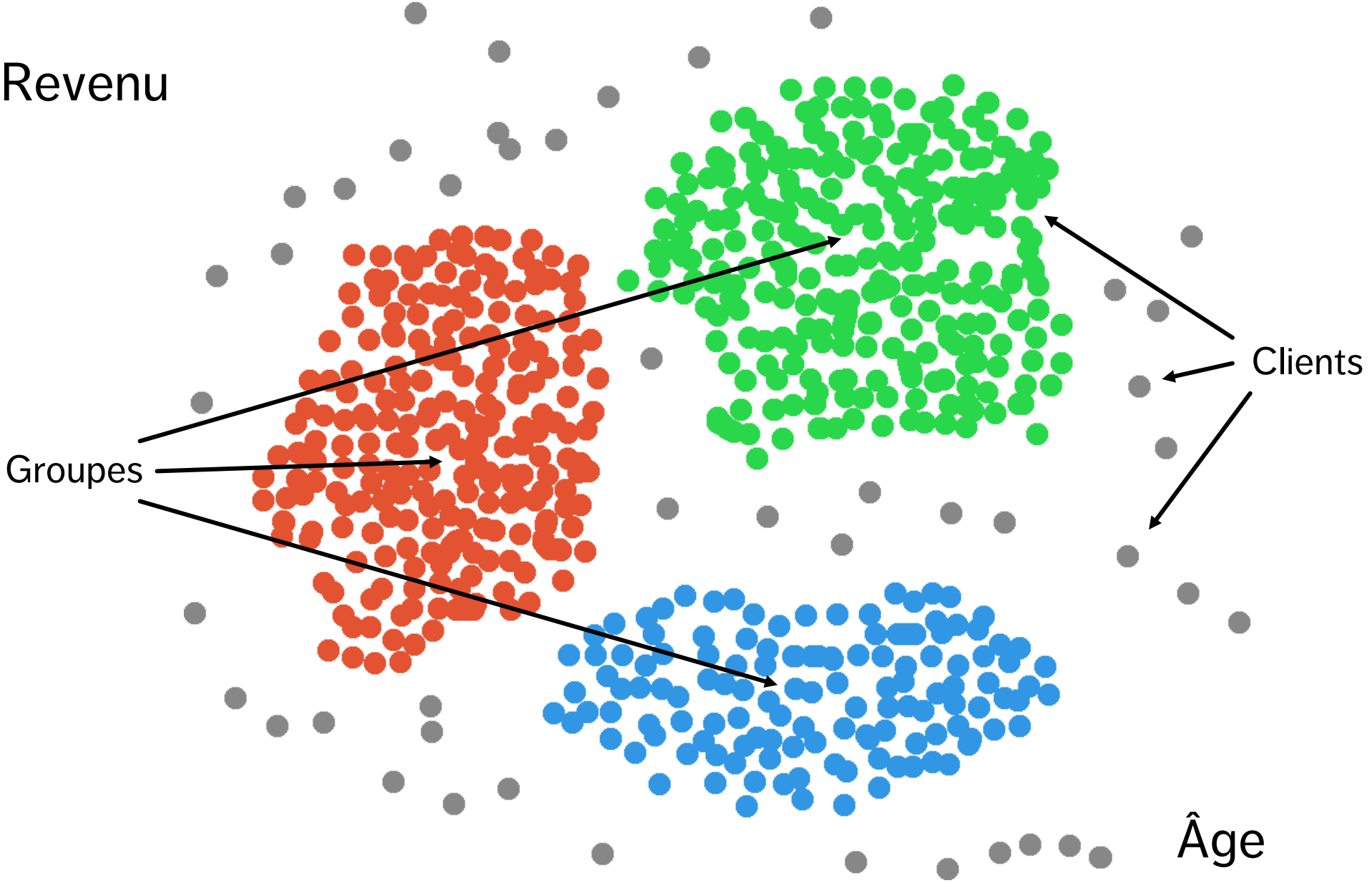
Les étiquettes des groupes ne sont pas déterminées au préalable, donc le regroupement est un exemple d'apprentissage **non supervisé**.

distance moyenne entre les points dans le même groupe (**de préférence, une courte distance**)



distance moyenne entre les points dans le groupe voisin (**de préférence, une grande distance**)




Revenu
Clients
Groupes
Âge

APERÇU DU REGROUPEMENT

Le regroupement est un concept relativement **intuitif** pour les êtres humains, car nos cerveaux le font de manière inconsciente.

- reconnaissance faciale
- recherche de modèles, etc.

En général, les gens sont très bons avec des données **désordonnées**, mais les ordinateurs et les algorithmes ont de la difficulté.

Une partie de la difficulté tient au fait qu'il n'existe **aucune définition consensuelle d'un groupe** :

- « Je peux ne pas être en mesure de définir ce que c'est, mais je le sais quant j'en vois un. »

APERÇU DU REGROUPEMENT

Les algorithmes de regroupement peuvent être **complexes** et **non intuitifs**, selon les diverses notions de similarités entre les observations.

- Malgré tout, il est **très** tentant d'expliquer les groupes *a posteriori*.

Ils sont aussi (typiquement) **non déterministes** :

- le même algorithme, exécuté deux fois (ou plus) sur le même ensemble de données, peut donner lieu à des groupes totalement différents;
- l'ordre de présentation des données peut jouer un rôle;
- tout comme la configuration de départ.

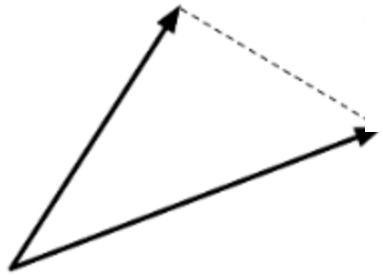
DISCUSSION

Que signifie cette non-reproductibilité (potentielle) pour la validation?

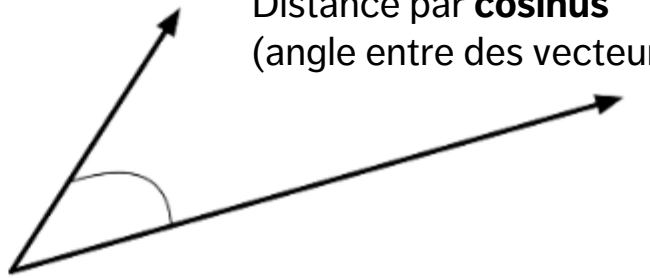
IMPORTANT: les données doivent être mises sur une même échelle avant d'effectuer le regroupement.

EXIGENCE DE REGROUPEMENT

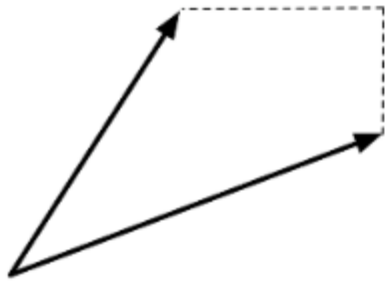
Mesure de la **similarité** w (ou d'une distance d) entre des observations.



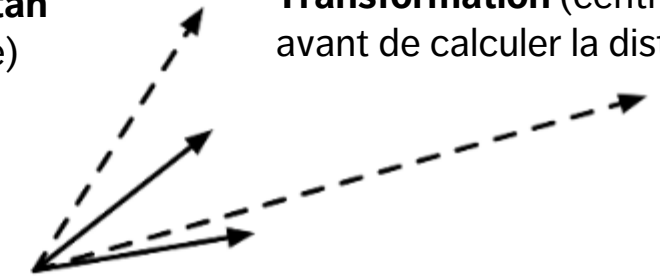
Distance **euclidienne**
(vol d'oiseau)



Distance par **cosinus**
(angle entre des vecteurs)



Distance de **Manhattan**
(si on devait conduire)



Transformation (centre normalisé)
avant de calculer la distance

En général, $w \to 1$, car $d \to 0$, et $w \to 0$, car $d \to \infty$.

* les données catégoriques doivent être dichotomisées

MESURES DE LA DISTANCE (PARAMÈTRES)

Variables de partitionnement*

- Distance de Hamming
- Indice de Russel et Rao
- Jaccard
- Coefficient de concordance
- Coefficient de Dice
- etc.

Variables numériques*

- Euclidiennes
- Manhattan
- Corrélation
- Cosinus
- Pearson
- etc.

Il n'y a aucune règle constante pour déterminer la distance à utiliser avec l'algorithme des k-moyennes.

Les schémas concurrents sont souvent produits à l'aide de différents paramètres.

APPLICATIONS

Documents de texte

- Regrouper des documents similaires en fonction de leurs sujets, de l'utilisation des mots courants ou inhabituels qu'ils contiennent.

Recommandations de produits

- Regrouper des clients en ligne en fonction des produits visualisés, achetés, aimés ou détestés.
- Regrouper des produits en fonction des commentaires des clients.

Marketing et affaires

- Regrouper des profils de clients en fonction de leurs données démographiques et de leurs préférences.

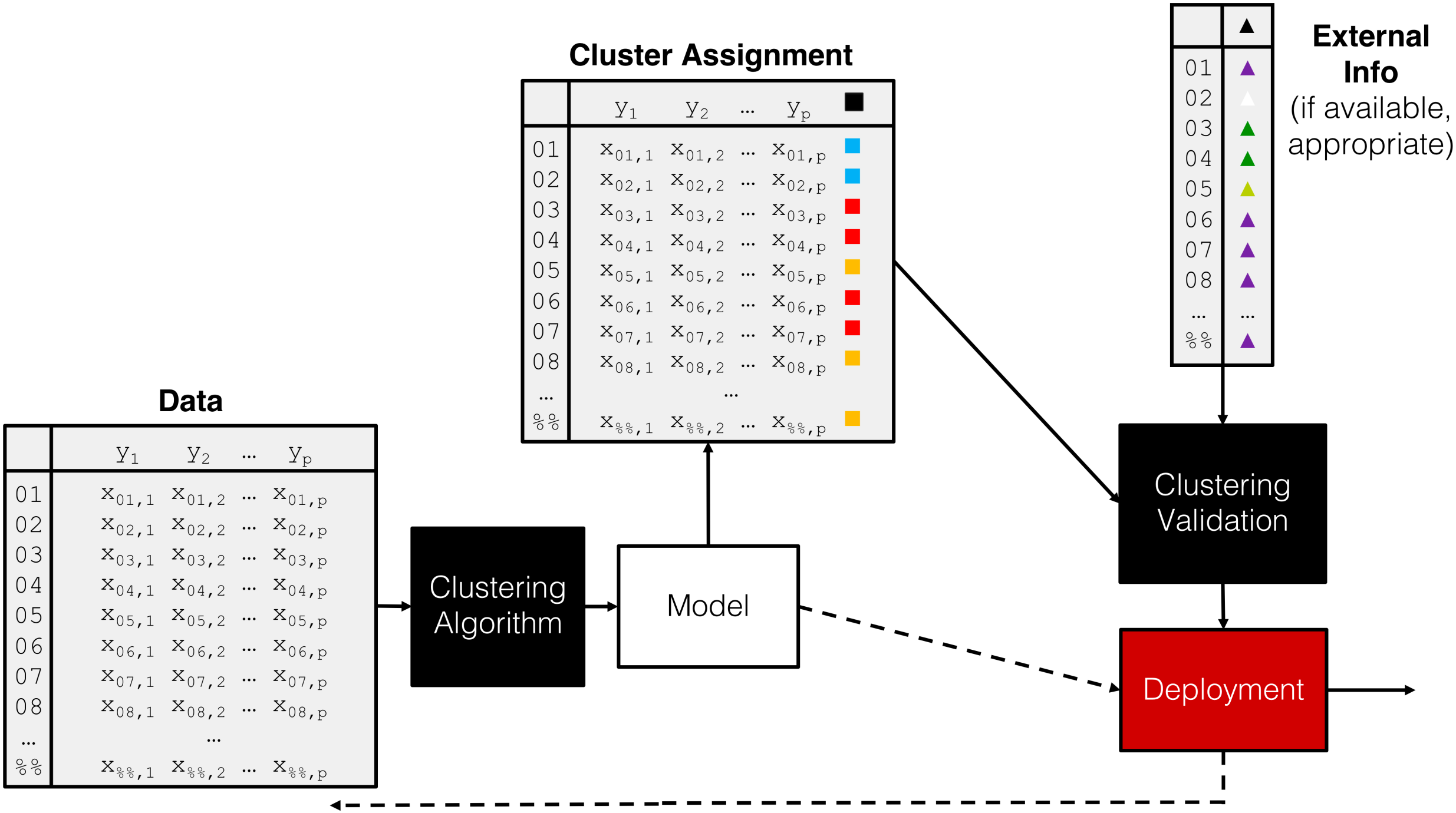
AUTRES UTILISATIONS

Diviser un grand groupe (ou une superficie ou une catégorie) en groupes **plus petits**, dans lesquels les membres ont certaines similarités.

- Les tâches peuvent être accomplies séparément par chaque plus petit groupe.
- Ce processus peut donner lieu à une précision accrue après l'agrégation des résultats distincts.

Créer de (nouvelles) taxonomies **à mesure que l'on progresse**, et que de nouveaux éléments sont ajoutés à un groupe.

- Facilite la navigation dans les produits affichés sur un site web, comme Netflix, par exemple.


Cluster Assignment
External Info
(if available, appropriate)
Data
Clustering Algorithm
Model
Clustering Validation
Deployment

ALGORITHMES DE REGROUPEMENT

REGROUPEMENT

« Le regroupement dépend du créateur et, en ce sens, les chercheurs ont proposé de nombreux principes et modèles d'intégration dont le problème d'optimisation correspondant peut seulement être résolu approximativement par un plus grand nombre d'algorithmes. »

(V. Estivill-Castro, *Why So Many Clustering Algorithms?*)

MODÈLES DE REGROUPEMENT

k-moyennes

- modèle classique (et surutilisé)
- hypothèses visant la forme des groupes

Regroupement hiérarchique

- facile à interpréter, déterministe

Allocation de Dirichlet latente

- utilisée dans la modélisation des sujets

Maximisation de l'espérance

MODÈLES DE REGROUPEMENT

Réduction et regroupement itératifs et équilibrés au moyen de hiérarchies

- c.-à-d. BIRCH

Regroupement par densité spatiale des applications avec bruit

- repose sur des graphiques

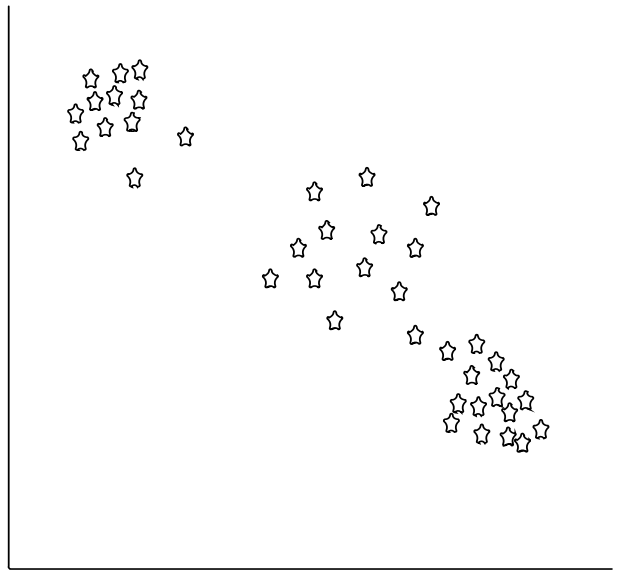
Propagation par affinités

- sélectionne automatiquement le nombre optimal de groupes

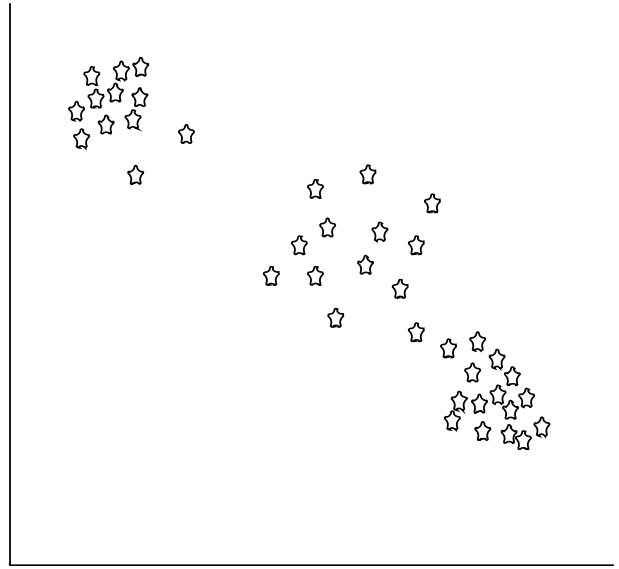
Regroupement spectral

- reconnaît les groupes non globulaires

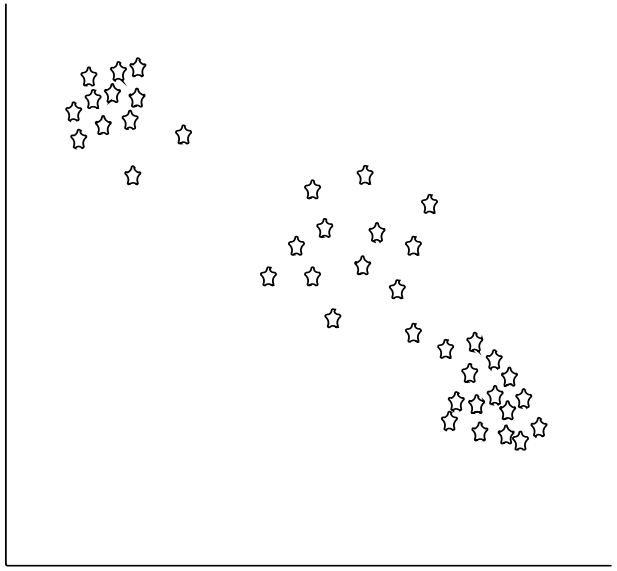
FORME GÉNÉRALE D'UN ALGORITHME DE REGROUPEMENT



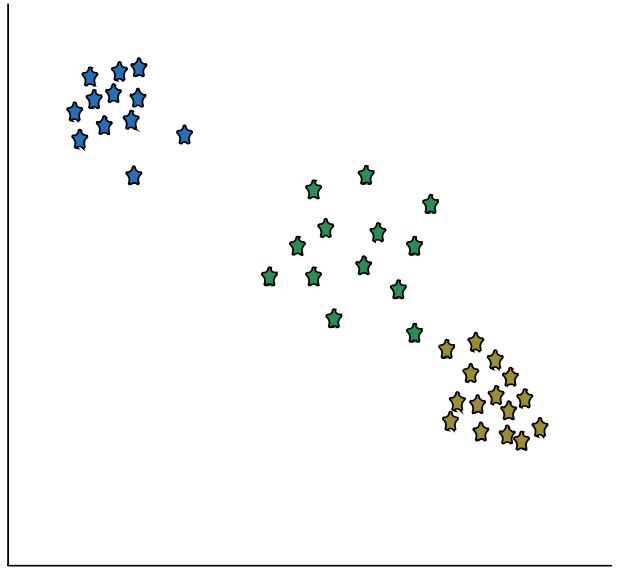
Initialization



Clustering Step A (Usually Repeated, Possibly in Conjunction with Next Step)



Clustering Step B (Usually Repeated, Possibly in Conjunction with Previous Step)

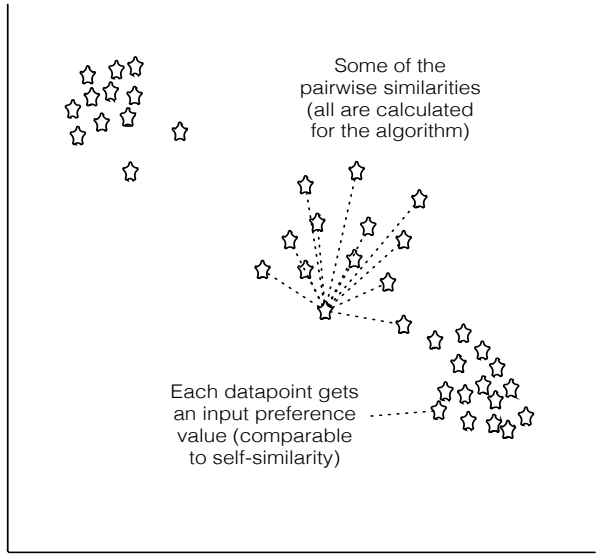


End Condition (Usually When Iterations of Steps A and B Produce Stable Results)

COMPARAISON DES ÉTATS DE DÉPART DE CERTAINS REGROUPEMENTS COURANTS

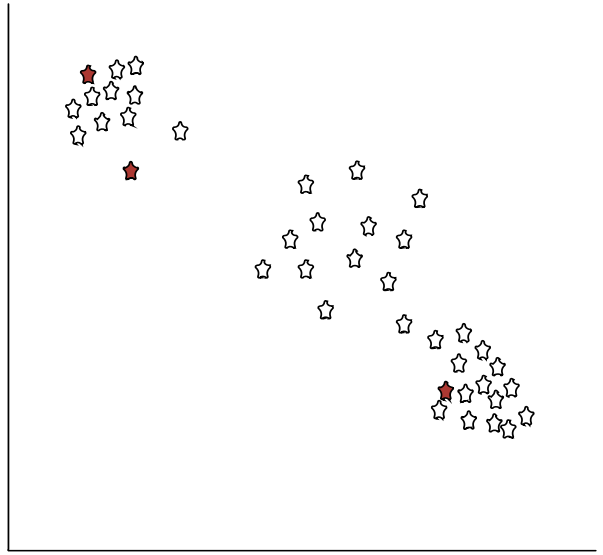
Donne une idée de la très grande différence dans les résultats d'un regroupement créé au moyen d'algorithmes différents.

Affinity Propagation Initialization



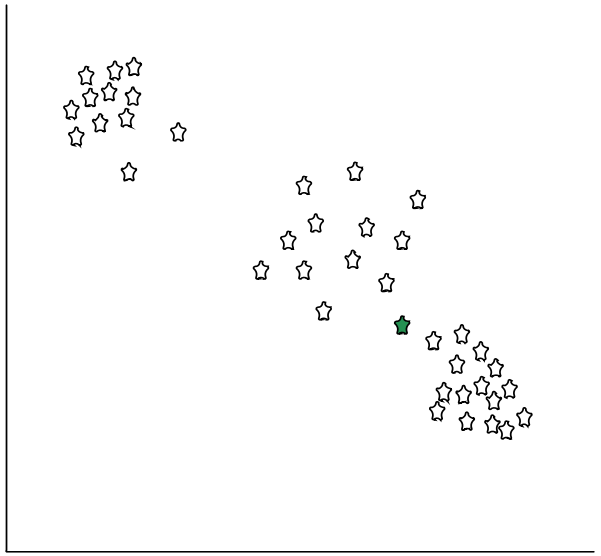
Calculate All Pairwise Similarities, Set Input Preference Values

K-Means Initialization



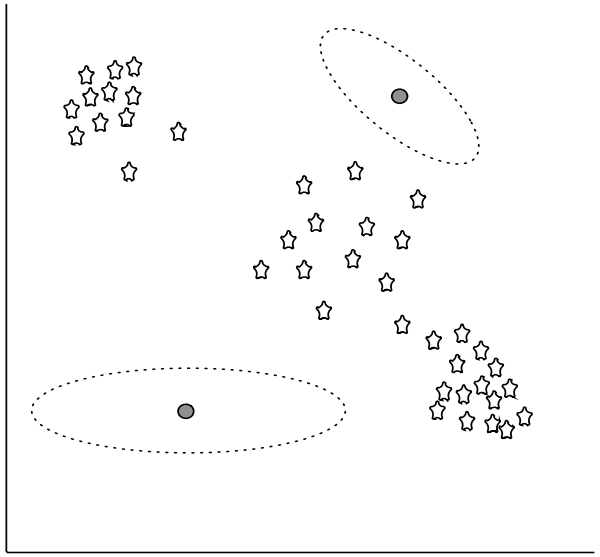
Randomly Pick k Centers

DBSCAN Initialization



Randomly Pick a DataPoint

Expectation Maximization Initialization



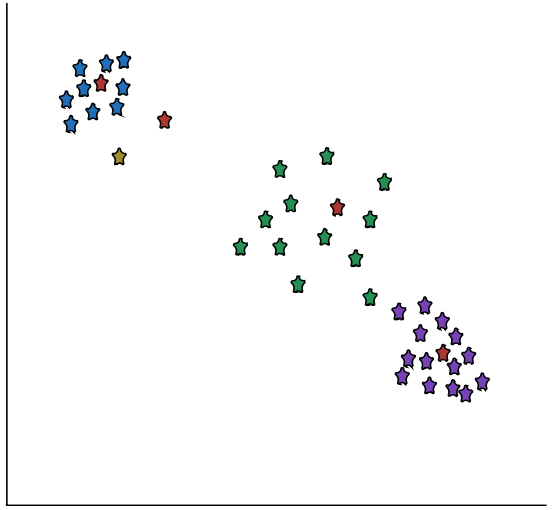
Set Initial Statistical Models

IDLEWYLD Sysabee DAVHILL u

COMPARAISON DES ÉTATS INTERMÉDIAIRES DE CERTAINS ALGORITHMES COURANTS

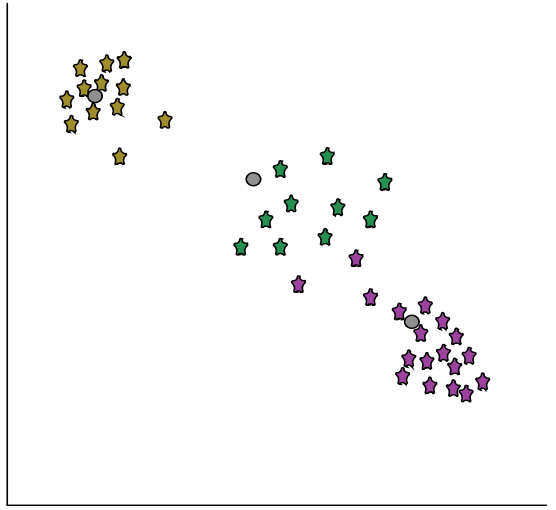
Donne une idée de la très grande différence dans les résultats d'un regroupement créé au moyen d'algorithmes différents.

Affinity Propagation



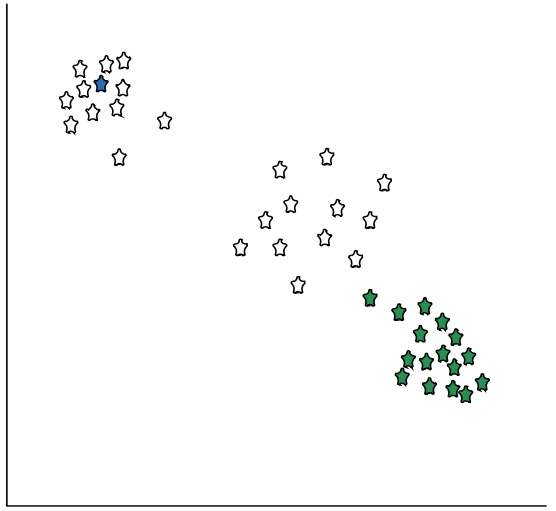
Select Good Over-All Exemplars (Based on Responsibility and Availability Scores). Assign Points to Clusters Based on Which Exemplars Are Most Suitable for Each Point

K-Means



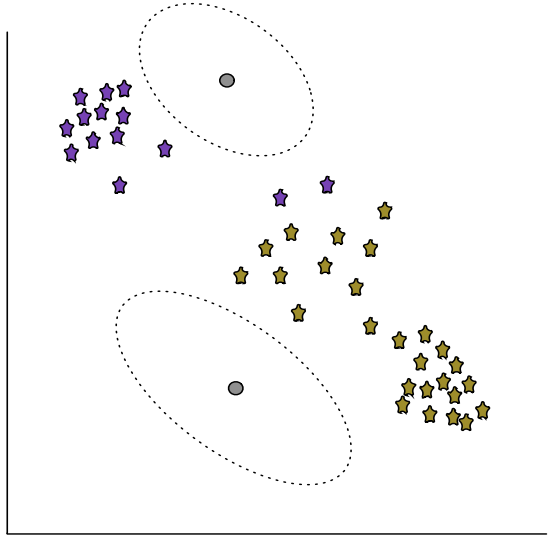
Re-assign Points Based on Centroids. Repeat from Previous Step (Calculate New Centroids) Until Stable

DBSCAN



Randomly Pick a New Unclustered Point and Try to Grow Another Cluster.

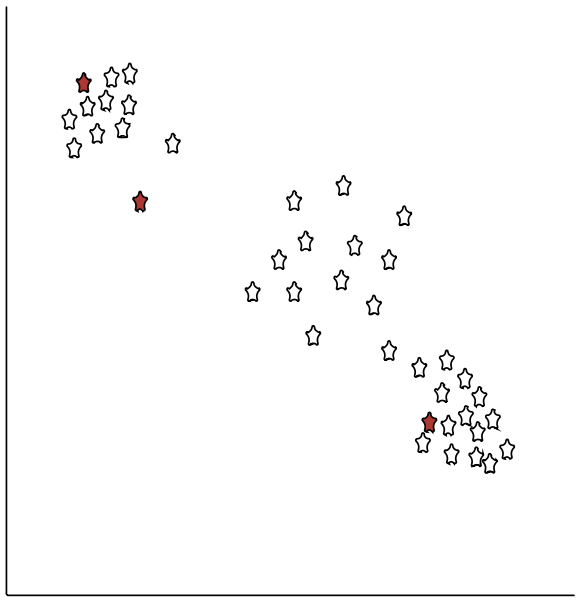
Expectation Maximization



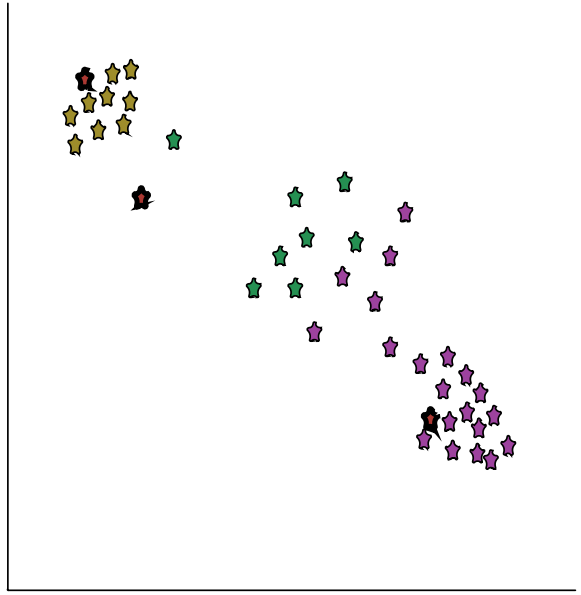
Adjust Clustering Assignment. Repeat from Previous Step (Adjust Models) Until Stable

IDLEWYLD Sysabee DAVHILL uOttaw

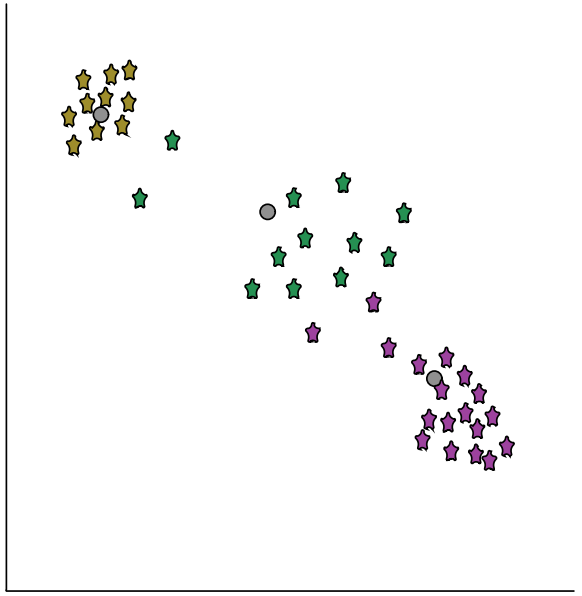
ÉCLAT DE L'ALGORITHME DES K-MOYENNES



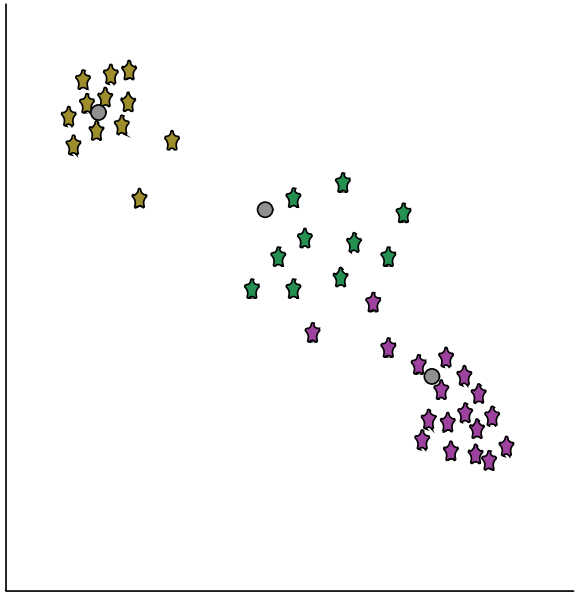
Initialization (e.g. Randomly Pick k Centers)



Assign Initial Clusters (Based on Distance to Centers)



Calculate Centroids of Clusters



Re-assign Points Based on Centroids. Repeat from Previous Step Until Stable

ALGORITHME DES k-MOYENNES

1. Sélectionnez le **nombre désiré de groupes**, disons k.
2. Sélectionnez au hasard k instances à titre de **centres initiaux du groupe**.
3. Calculez la **distance** de chaque observation par rapport au centre.
4. Placez chaque instance dans le groupe en fonction du centre **le plus proche**.
5. Calculez le **centre de masse** de chaque groupe.
6. Répétez les étapes 3 à 5 avec les nouveaux centres de masse.
7. Répétez l'étape 6 jusqu'à ce que les groupes soient **stables**.

POINTS FORTS DE L'ALGORITHME DES k-MOYENNES

Facile à créer (sans avoir à réellement calculer les distances entre paires).

- très courant en conséquence
- élégant et simple

Dans de nombreux contextes, l'algorithme des k-moyennes est une méthode **naturelle** d'examiner les regroupements.

Aide à fournir une **compréhension élémentaire de la structure des données** au premier examen.

LIMITES DE L'ALGORITHME DES k-MOYENNES

Vous ne pouvez assigner les points de données qu'à un seul groupe.

- Peut donner lieu à un surapprentissage.
- Solution rigoureuse : examinez la **probabilité** de faire partie de chaque groupe.

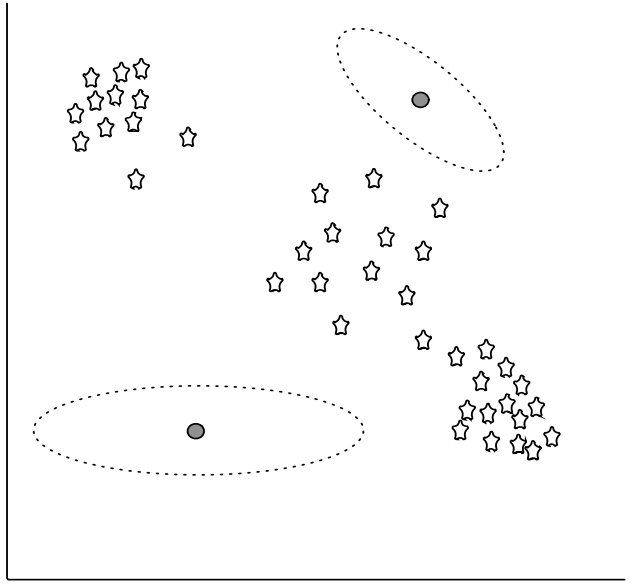
Vous devez présumer que les groupes sous-jacents sont de **forme globulaire**.

- L'algorithme des k-moyennes échoue à produire des groupes utiles si l'exercice pratique ne peut pas confirmer l'hypothèse.

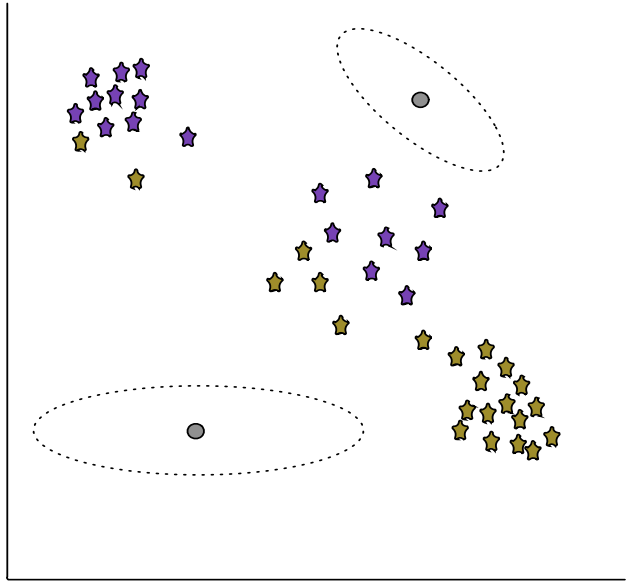
Vous devez présumer que les groupes sont distincts (discrets).

- L'algorithme des k-moyennes ne permet pas la présence de groupes **chevauchants** ou **hiérarchiques**.

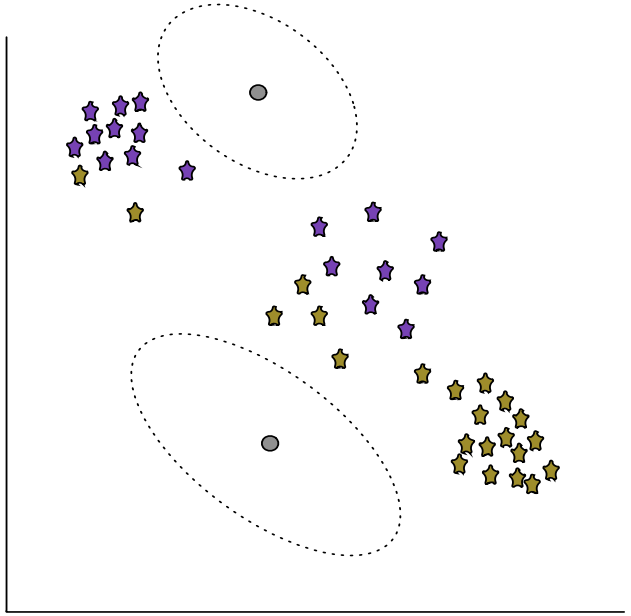
ÉCLAT DE L'ALGORITHME DE LA MAXIMISATION DE L'ESPÉRANCE



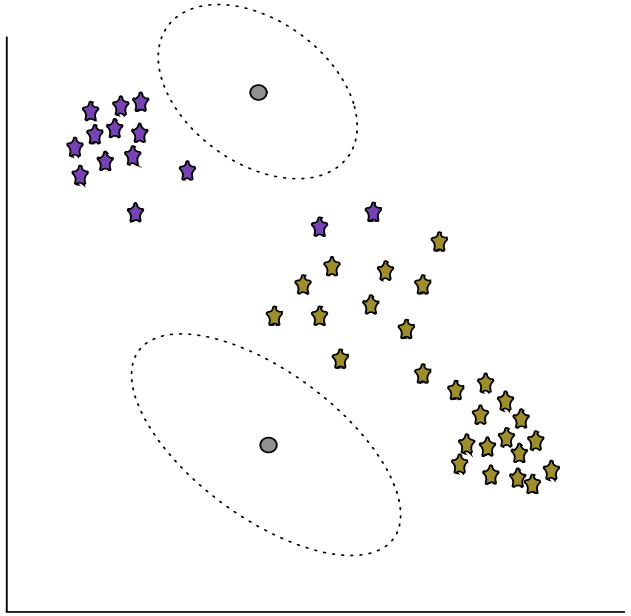
Initialization (Set Initial Statistical Models)



Assign Clusters Based on Models

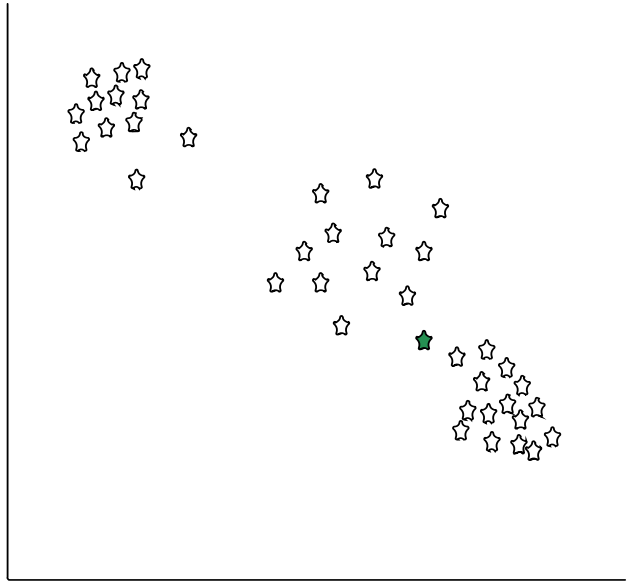


Adjust Statistical Models

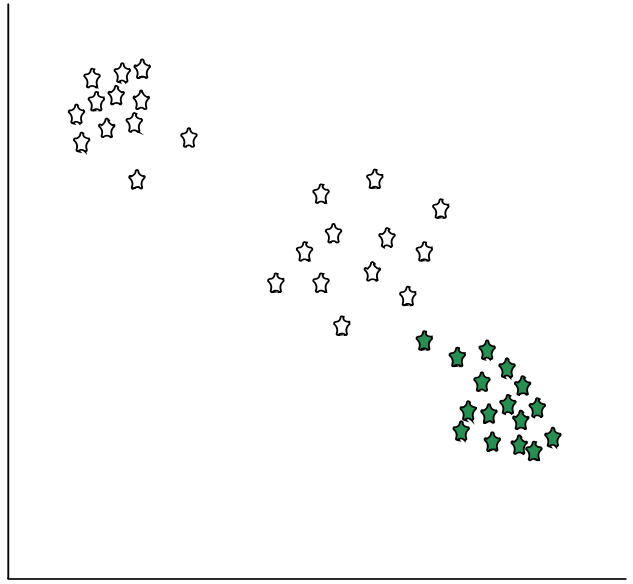


Adjust Clustering Assignment.
Repeat from Previous Step Until Stable

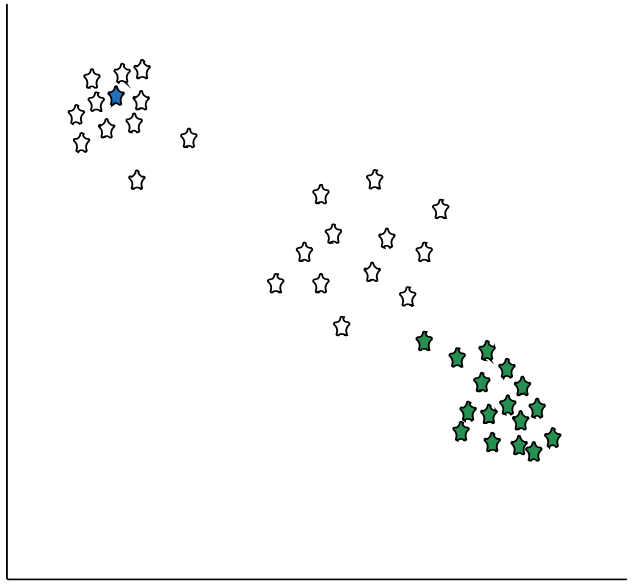
ÉCLAT DE L'ALGORITHME DBSCAN



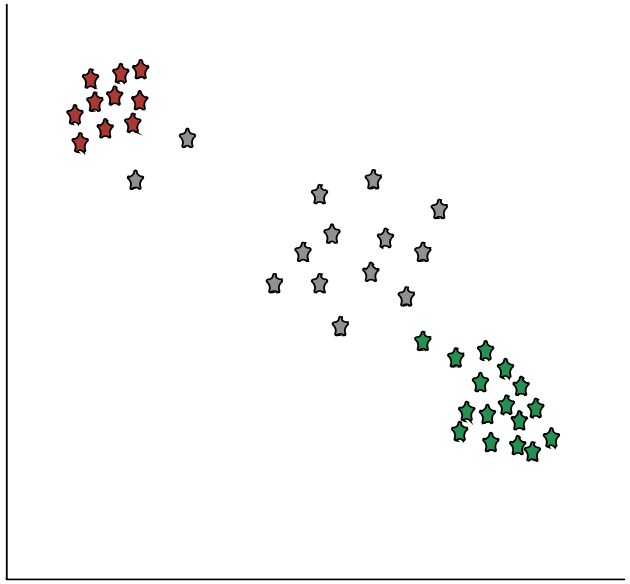
Initialization (Randomly Pick a DataPoint)



(Try to) Grow A Cluster. Stop When There Are No More Close Points

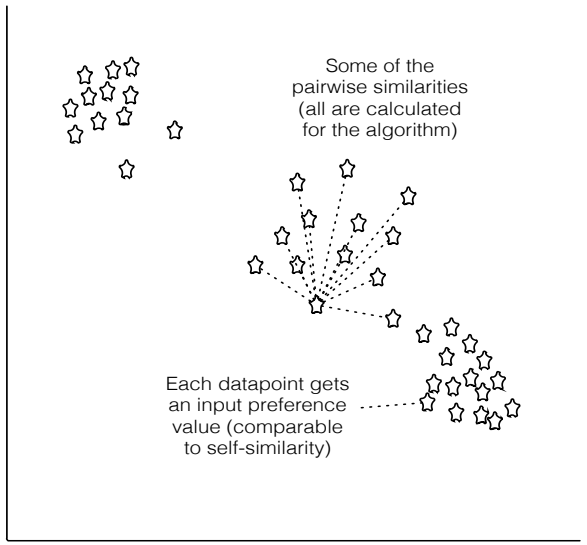


Randomly Pick a New Unclustered Point and Try to Grow Another Cluster.

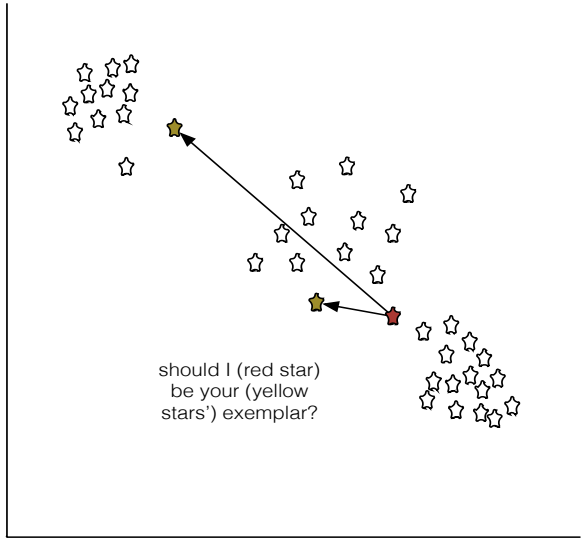


Stop Clustering When All Points Are Clustered (or Marked as Anomalies)

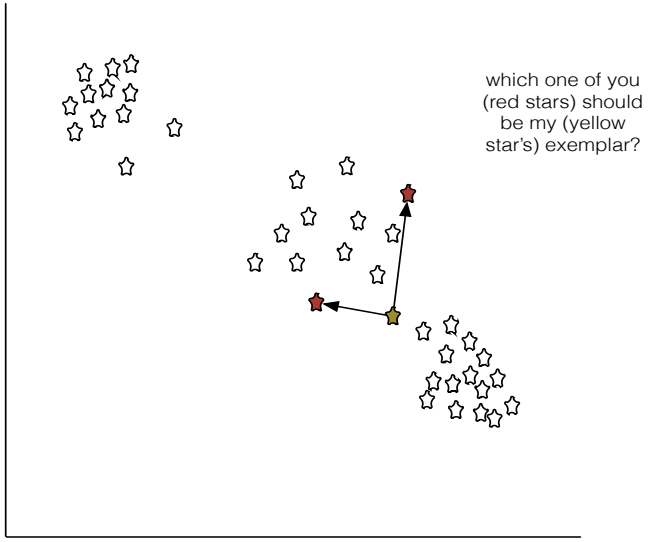
ÉCLAT DE LA PROPAGATION PAR AFFINITÉS



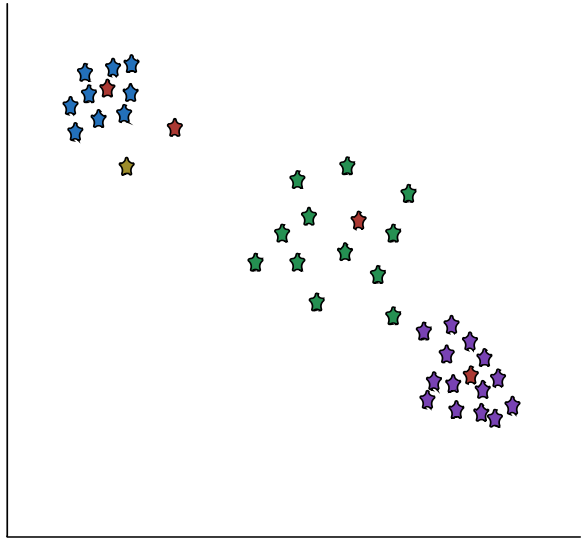
Initialization (Calculate All Pairwise Similarities, Set Input Preference Values)



Consider, for Each Point, its Suitability to Be an Exemplar For Each Other Point (Responsibility Score)



Consider, For Each Point, The Suitability of Every Other Point To Be an Exemplar for that Point (Availability Score - Based Also on Responsibility). Iterate Between This And Previous Step Until Stable



Select Good Over-All Exemplars (Based on Responsibility and Availability Scores). Assign Points to Clusters Based on Which Exemplars Are Most Suitable for Each Point

DISCUSSION

Jusqu'à quel point le choix de l'algorithme (et des paramètres de distance et autres paramètres) dépend des données et des types de données offerts?

VALIDATION D'UN REGROUPEMENT

REGROUPEMENT

VALIDATION D'UN REGROUPEMENT

Qu'est-ce que ça signifie qu'un modèle de regroupement soit **mieux** qu'un autre?

Qu'est-ce que ça signifie qu'un modèle de regroupement soit **valide**?

Qu'est-ce que ça signifie qu'un groupe soit **bon**?

Combien de groupes y a-t-il réellement dans les données?

La notion de bon ou de mauvais ne veut rien dire : il faut rechercher les regroupements **optimaux** plutôt que les regroupements **sous-optimaux**.

VALIDATION D'UN REGROUPEMENT

Modèle de regroupement **optimal** :

- séparation maximale entre les groupes
- similarité maximale dans les groupes
- réussit le test de l'œil humain
- est utile pour atteindre les objectifs

Types de validation

- externe (utilise des renseignements supplémentaires)
- interne (utilise seulement les résultats du regroupement)
- relative (établit des comparaisons entre diverses tentatives de regroupement)

DISCUSSION

Le principal défi du regroupement tient au fait que nous ne savons pas **avec quoi** nous comparons les résultats du modèle de regroupement (des versions de ce problème minent les tâches non supervisées).

Alors, pourquoi faire des regroupements?

ENSEMBLE DE DONNÉES DES IMAGES DE FRUITS

20 images de fruits

Est-ce que cet ensemble de données permet de bons et de mauvais regroupements?

Y a-t-il plusieurs regroupements « naturels » possibles?

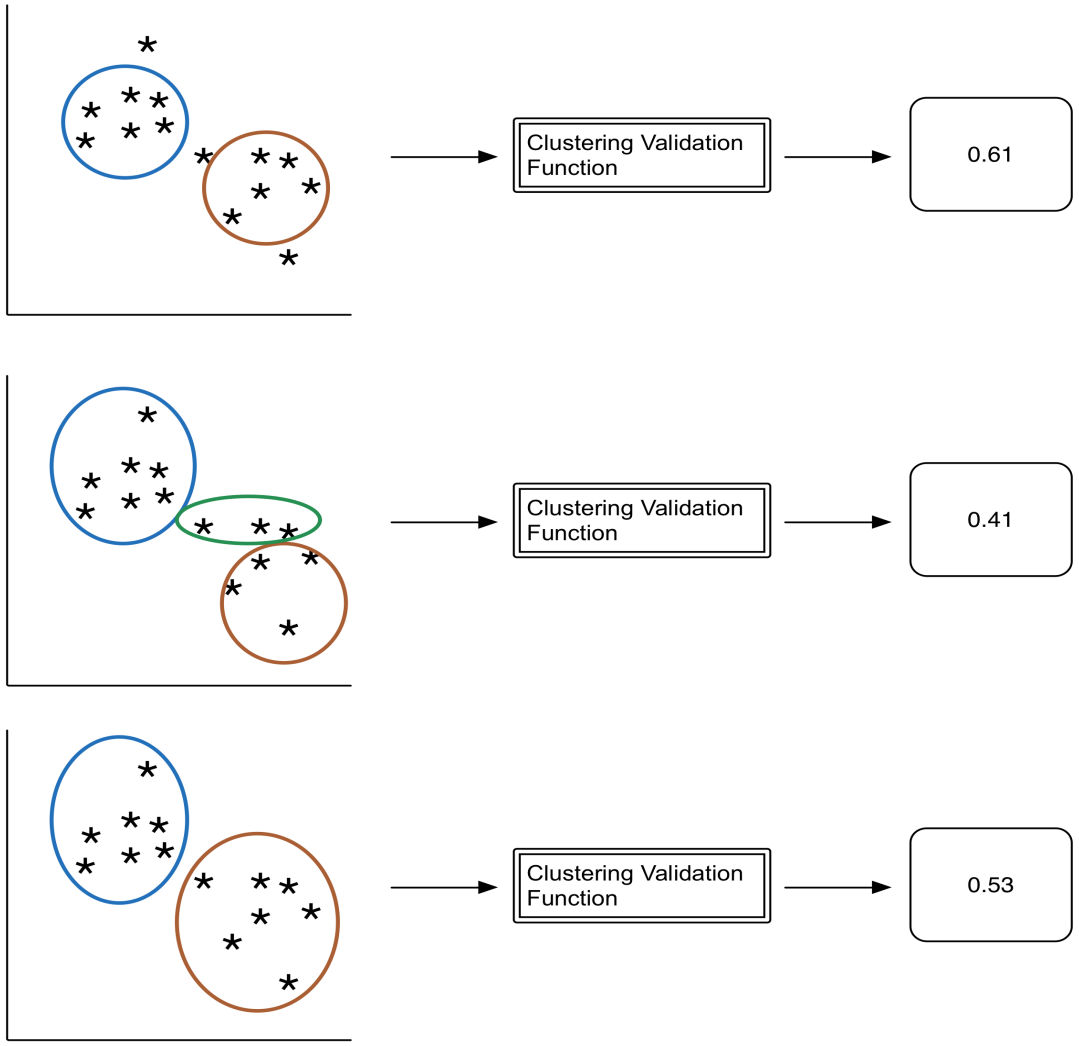
Serait-il possible d'utiliser des regroupements différents?

Est-ce que certains regroupements seront (objectivement) de meilleure **qualité** que les autres?

CONCRÉTISATION DES CONCEPTS

Pour évaluer la validation d'un regroupement, il est utile de lier les concepts à un aspect concret.

Sur les prochaines diapositives, prenez le temps de penser à la manière de lier les concepts présentés aux images du présent petit ensemble de données.



VALIDATION D'UN REGROUPEMENT

Le regroupement fait appel à deux activités principales :

- la création de groupes;
- **l'évaluation de la qualité des groupes**.

Nous créons des fonctions pour exercer ces deux activités.

Fonctions de regroupement

- Intrant : instances (vecteurs)
- Extrant : assignation des groupes à chaque instance

Évaluation de la qualité des groupes

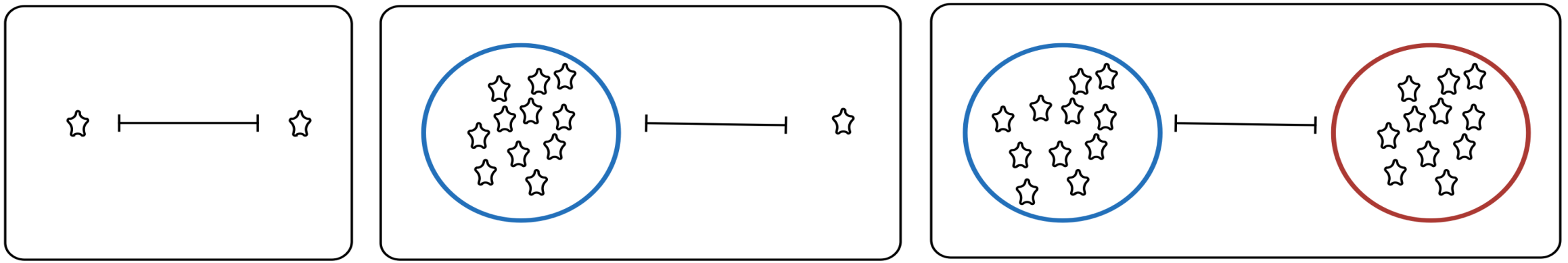
- Intrant : instances + assignation des groupes (+ matrice de similarité, en général)
- Extrant : valeur numérique

ÉLÉMENTS D'UNE FONCTION

Il existe un grand nombre de fonctions de regroupement et de validation des groupes.

Toutefois, toutes ces fonctions reposent sur les mesures de base liées aux propriétés de l'instance ou du groupe que nous avons déjà examinées :

- **Propriétés de l'instance**
- **Propriétés du groupe**
- **Propriétés des liens entre deux instances**
- **Propriétés des liens entre un groupe et une instance**
- **Propriétés des liens entre deux groupes**

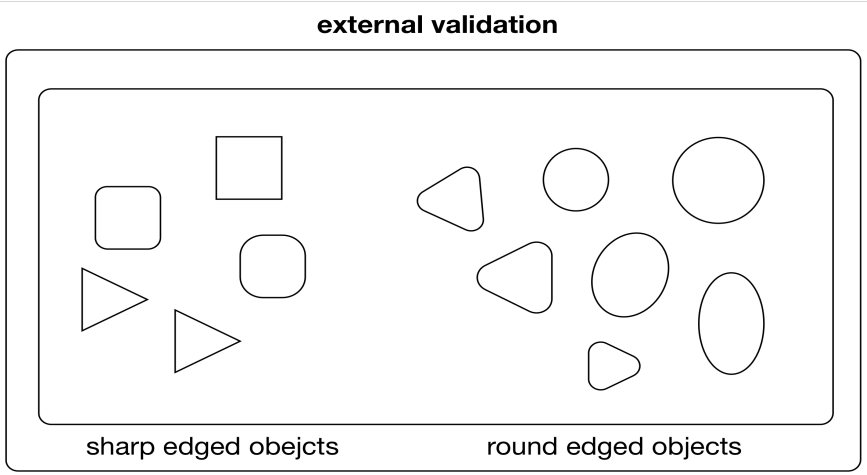


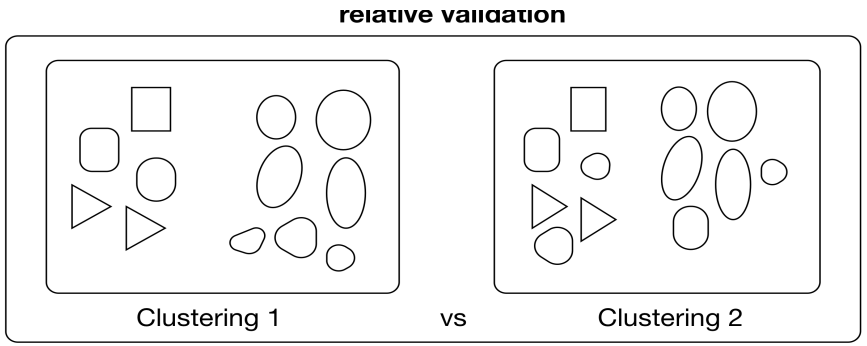
TROIS TYPES DE VALIDATION

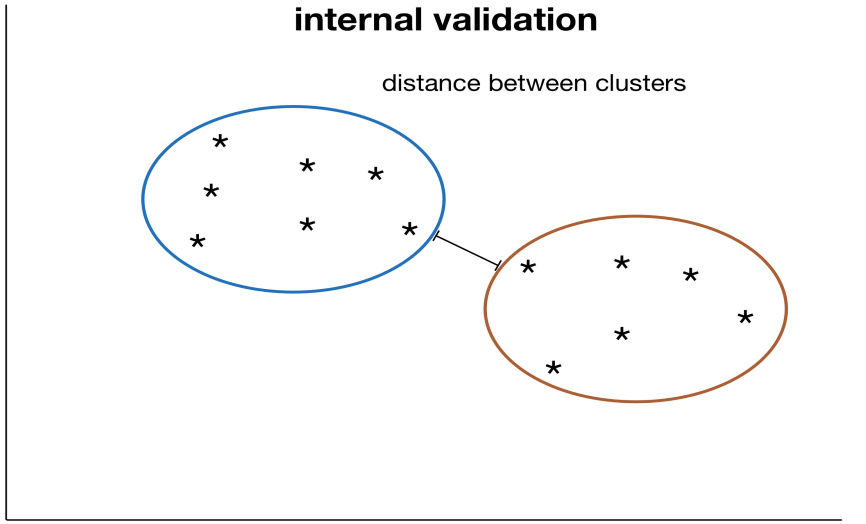
Validation interne : Repose seulement sur les propriétés incluses dans les résultats d'un seul groupe (il faut noter que cela comprend plusieurs groupes).

Validation relative : Comparaison des résultats d'un groupe avec les résultats d'un autre groupe.

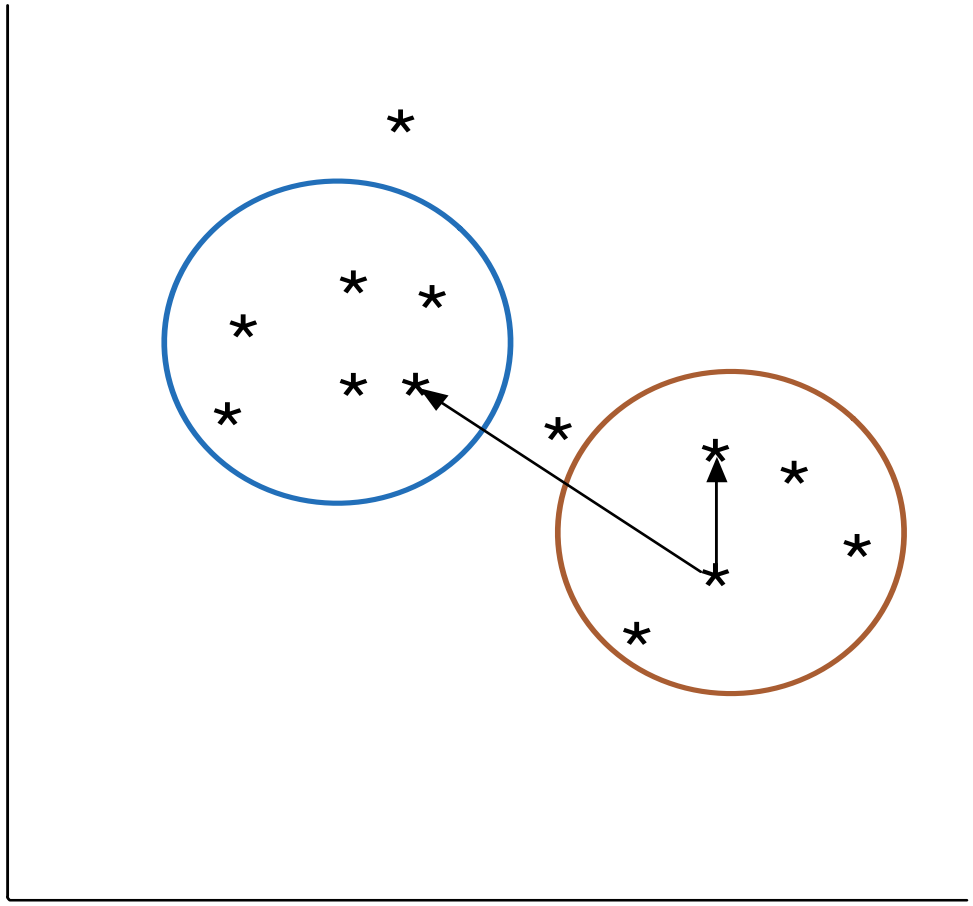
Validation externe : Comparaison des résultats d'un groupe avec une norme externe.







VALIDITÉ ET QUALITÉ



Le contexte est très important pour la qualité d'un regroupement, mais que se passe-t-il s'il n'y a aucun contexte?

Y a-t-il une manière de mesurer objectivement la qualité d'un groupe sans tenir compte d'un contexte particulier?

Le terme « validité » laisse entendre qu'il y a un regroupement **correct**, et tout ce que nous devons faire, c'est de vérifier comment proche nous y parvenons.

Par ailleurs, Lewis, Ackerman et de Sa (2012) utilisent plutôt le terme **mesures de la qualité d'un groupe**.

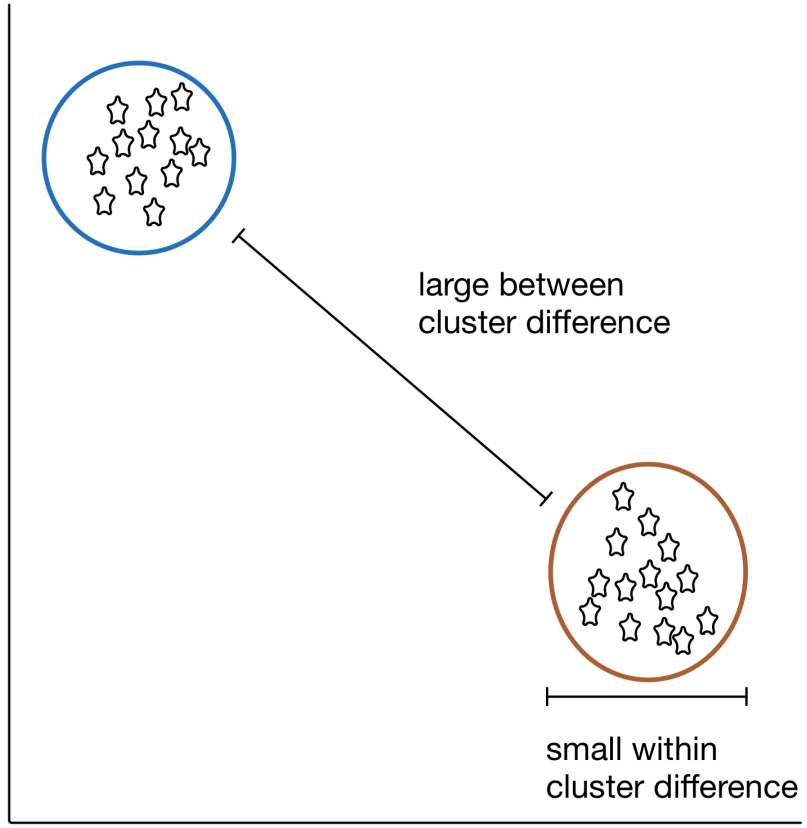
OBJECTIFS TRÈS GÉNÉRAUX

Dans un groupe, tout est très similaire. D'un groupe à un autre, la différence est énorme.

Le problème tient au fait que les groupes ont un grand nombre de manières de s'éloigner de cet idéal.

Dans certains cas, comment pouvons-nous pondérer les aspects positifs (p. ex., une note élevée pour la similarité à l'intérieur d'un groupe) et les aspects négatifs (p. ex., une note faible pour la séparation des groupes).

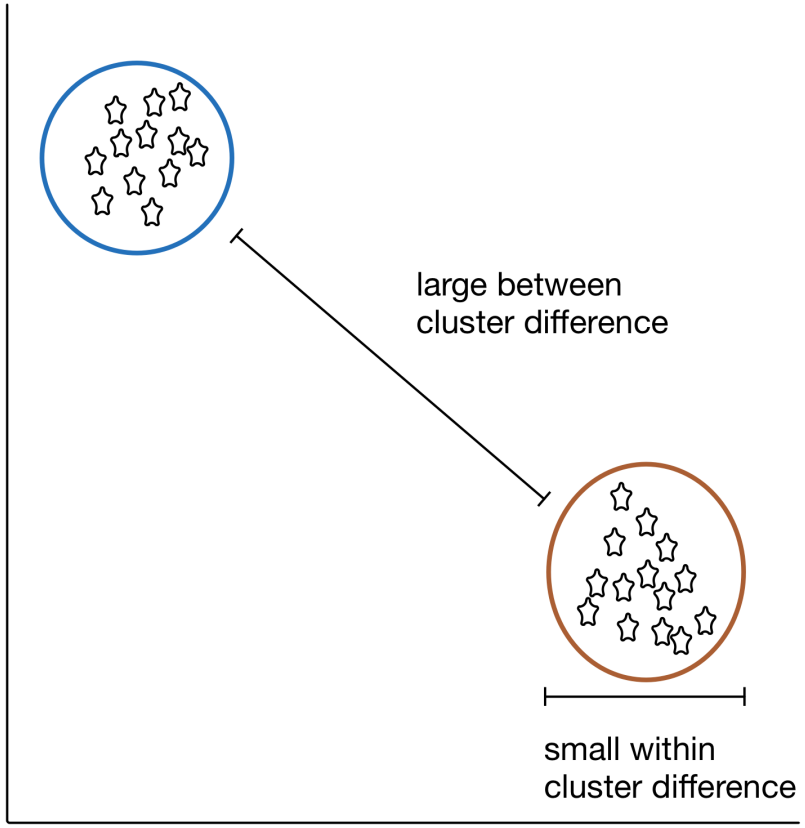
C'est la raison pour laquelle il y a un grand nombre de mesures de la qualité d'un groupe.



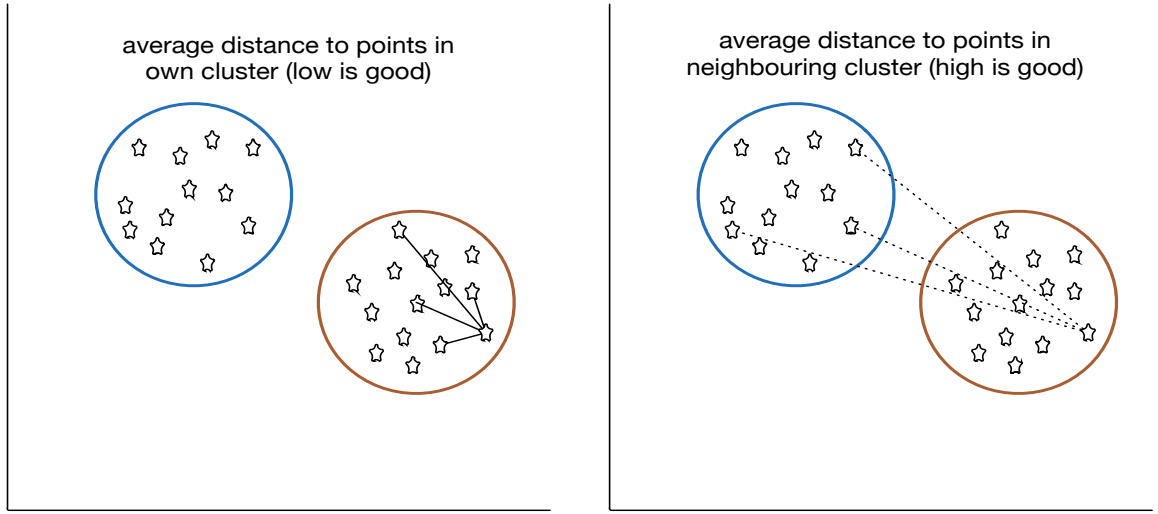
OBJECTIFS TRÈS GÉNÉRAUX

Question : est-ce que ce compromis (et les mesures de la qualité qui en résultent) n'a vraiment aucun lien avec le contexte?

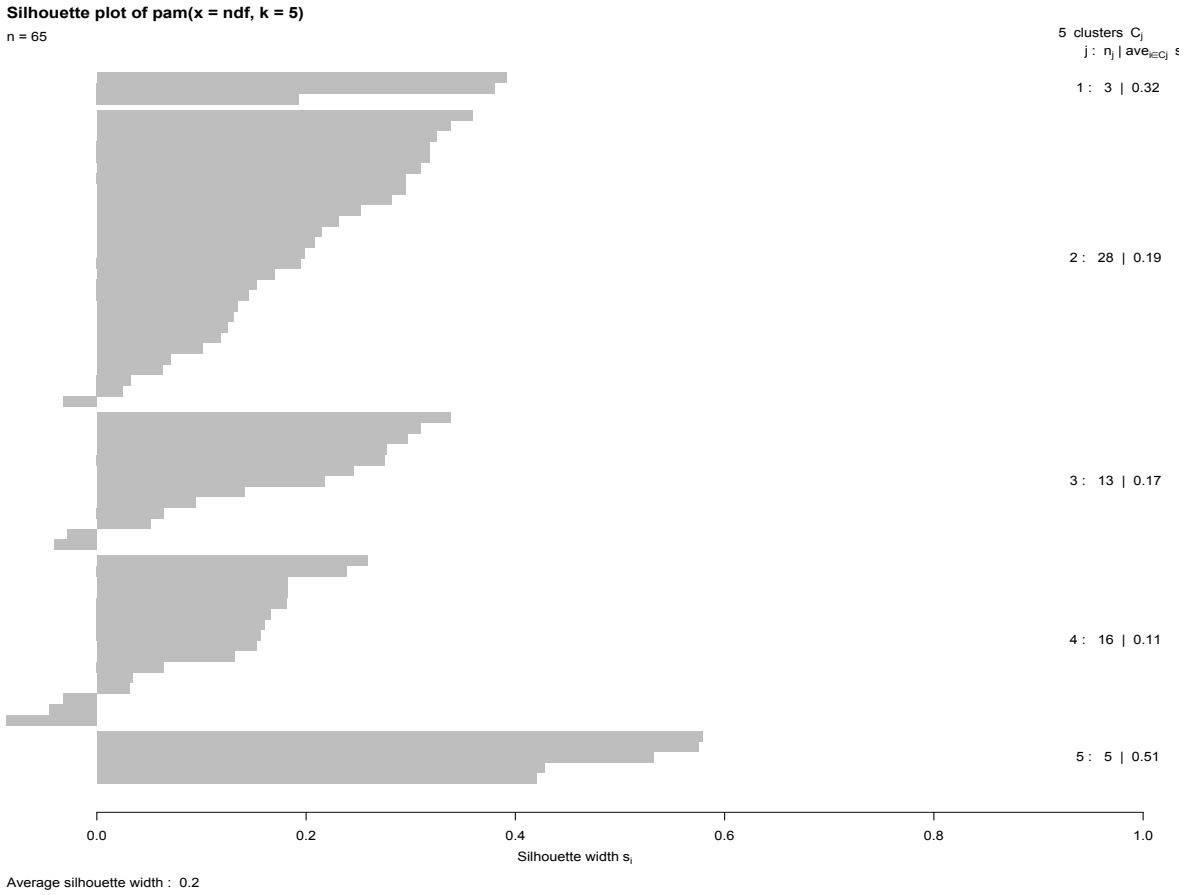
Il se peut que des pondérations différentes soient plus pertinentes dans des contextes différents?



INDICE DE SILHOUETTE



$$\text{silhouette metric for a point} = \frac{\text{(average dissimiliarity with neighbouring cluster - average dissimilarity with own cluster)}}{\text{maximum dissimilarity value (own or neighbour)}}$$



Excellente mesure de la validation interne qui repose sur plusieurs mesures.

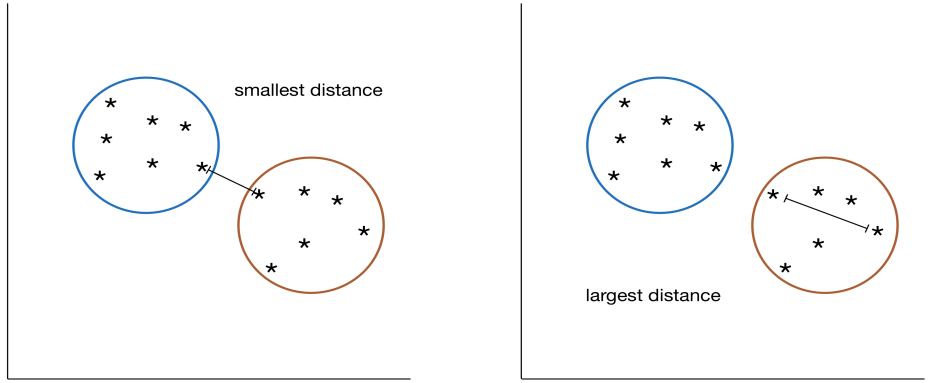
INDICE DE DUNN

Dans un groupe, taille du groupe (p. ex., la distance maximale entre les points).

Entre deux groupes, distance entre ces groupes (p. ex., la distance minimale entre les points).

Rapport : Distance minimale à l'intérieur d'un groupe entre toutes les paires de groupes ou distance maximale entre deux groupes.

Il existe diverses manières de définir la distance à l'intérieur d'un groupe et la taille d'un groupe.



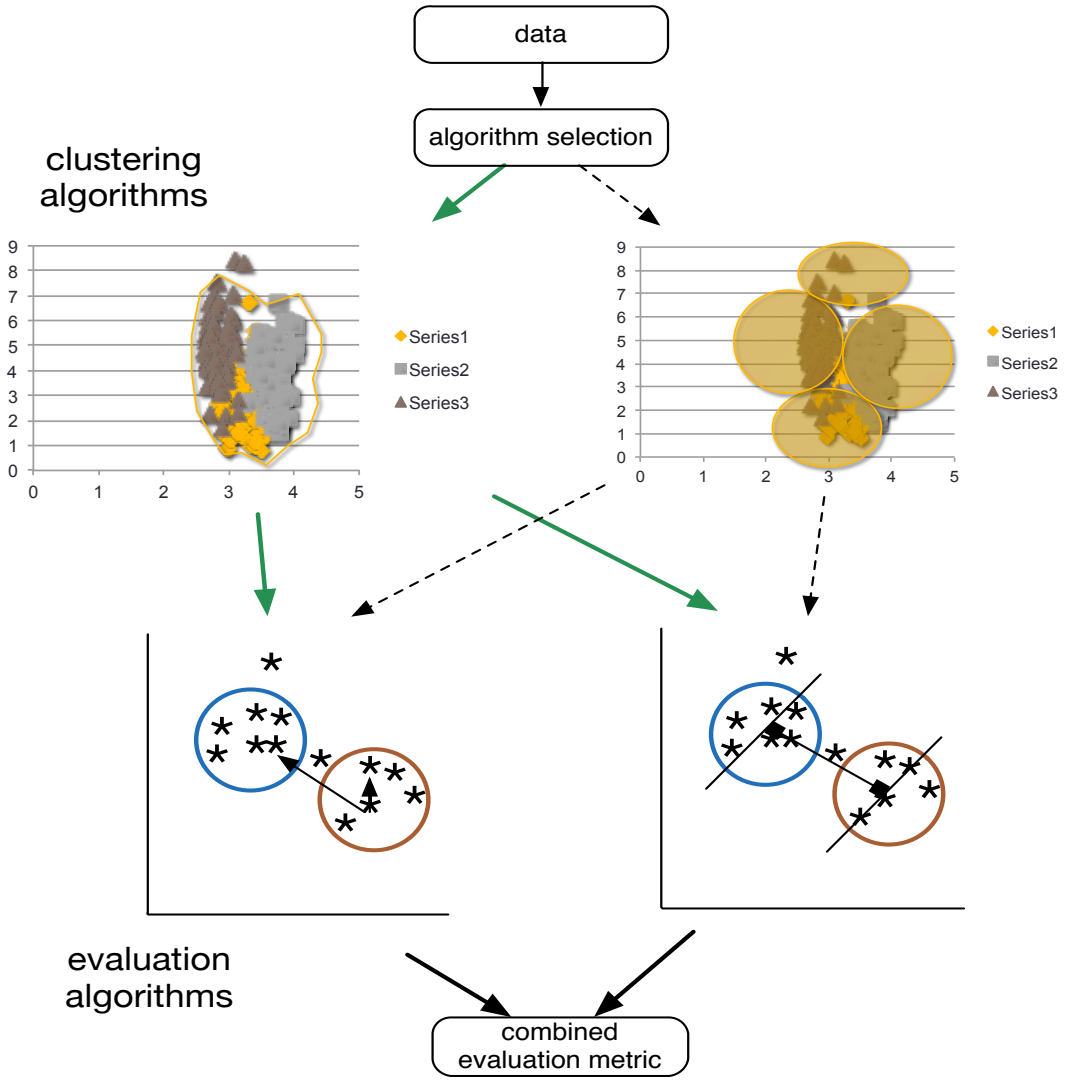
Comparaison faite au moyen de l'indice de silhouette : selon un certain point de vue, il s'agit d'une mesure plus simple. Il s'agit davantage d'une mesure de la totalité du groupe, plutôt que d'une mesure d'un point avec un autre point. Cet indice permet une évaluation en fonction des valeurs extrêmes (maximum, minimum).

L'ABONDANCE EST PRÉFÉRABLE (VALIDATION RELATIVE)?

Il n'est pas très utile d'obtenir une seule mesure de validation pour un seul regroupement. Pouvons-nous améliorer les résultats? Est-ce que c'est le mieux que nous pouvons espérer?

Et si nous comparions les résultats de diverses exécutions ou des paramètres choisis?

Le principal objectif de la validation relative consiste à trouver une manière de comparer les résultats de chaque exécution.



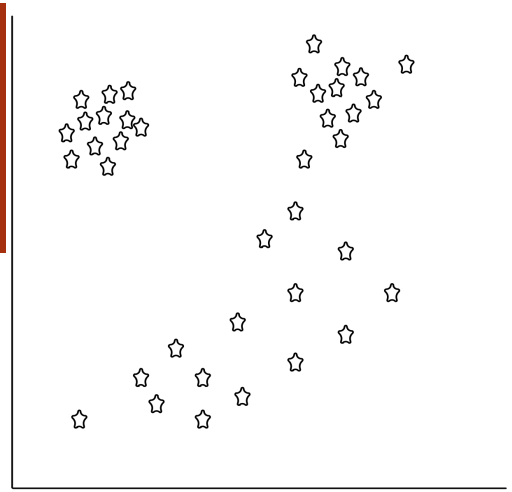
STABILITÉ

Quelques options :

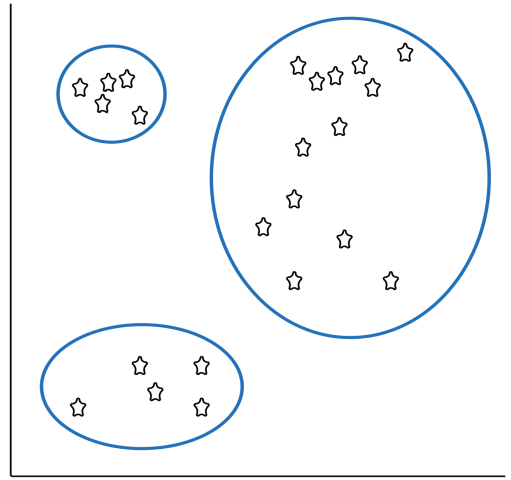
- plusieurs ensembles de données extraits de la même source
- utilisation de différentes colonnes pour créer les groupes (c.-à-d. sélectionner une colonne différente lors de chaque exécution)

Vous mesurez la similarité des résultats.

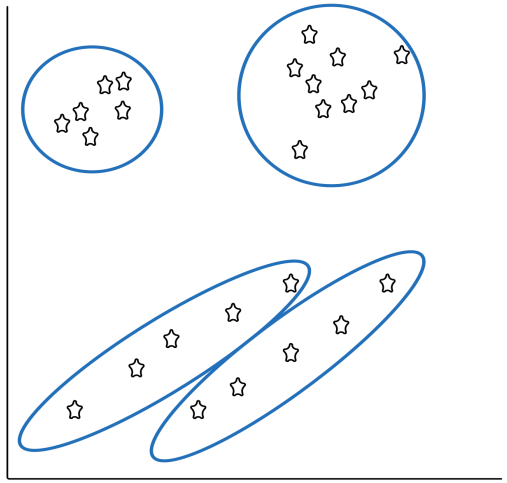
Si vous n'obtenez pas des résultats stables d'un regroupement à un autre, vous devez poursuivre les recherches.



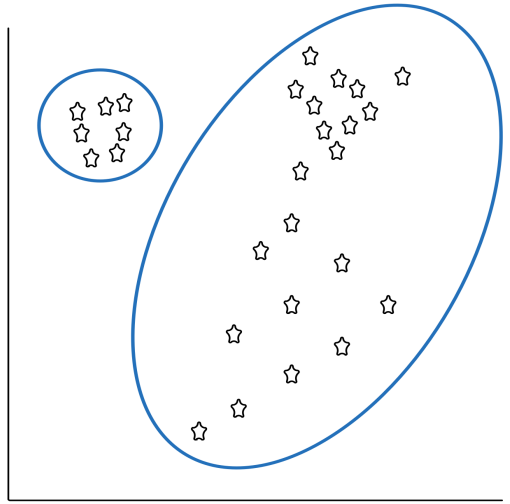
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

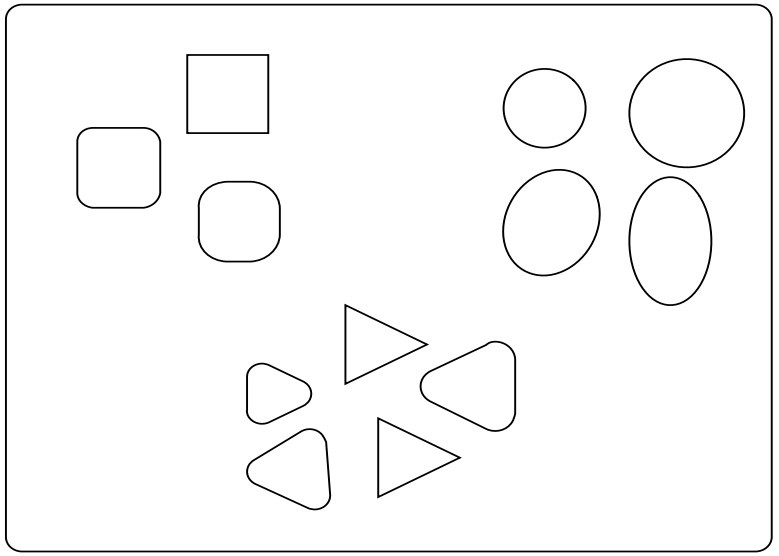
DE RETOUR AU CONTEXTE (VALIDATION EXTERNE)

Utiliser des renseignements externes pour évaluer les groupes.

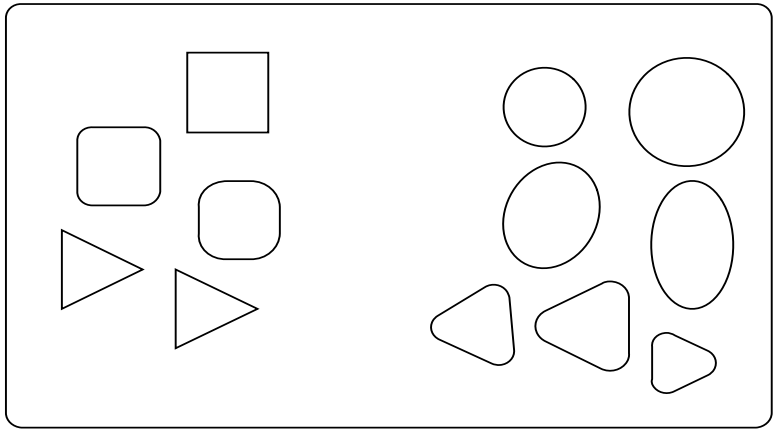
Ces renseignements externes représentent en général la « bonne » catégorie.

En quoi cette méthode est-elle différente de la classification?

Elle sert souvent à renforcer la confiance dans l'approche générale, selon les résultats préliminaires ou les résultats de l'échantillonnage.



Natural Groupings



Clustering Results

PURETÉ (VALIDATION EXTERNE)

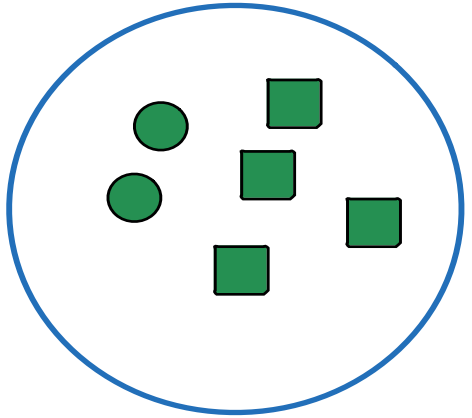
Avec cette mesure, chaque groupe est assigné à la catégorie la plus fréquente.

Pour calculer la pureté, on divise le nombre de points correctement assignés par le nombre de points dans le groupe.

Quelques options supplémentaires : précision, rappel.

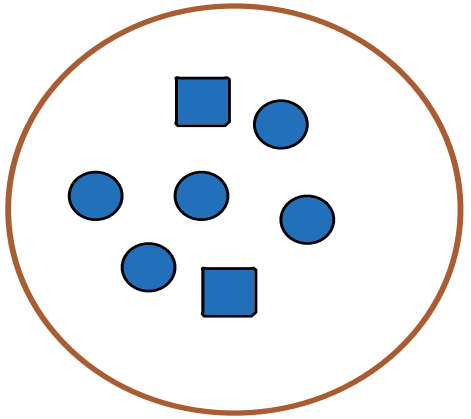
Assuming we are interested in shape…

SQUARE CLUSTER



purity = 66%

CIRCLE CLUSTER



purity = 71%

FAIRE PLUSIEURS ESSAIS

Diversité des techniques de validation des regroupements.

Soyez au fait des types de validation, et des variations offertes par chaque type.

Recherchez des concordances d'une technique à une autre.

Il existe plusieurs manières d'obtenir le bon regroupement – vous devez décider des aspects importants et des aspects dont vous n'avez pas à tenir compte.

Tout dépend en grande partie du **contexte**.

NOTES

REGROUPEMENT

« Des grappes de malheurs. Les malheurs solitaires sont rares; ils aiment se regrouper, ils se talonnent l'un l'autre. »

(Edward Young)

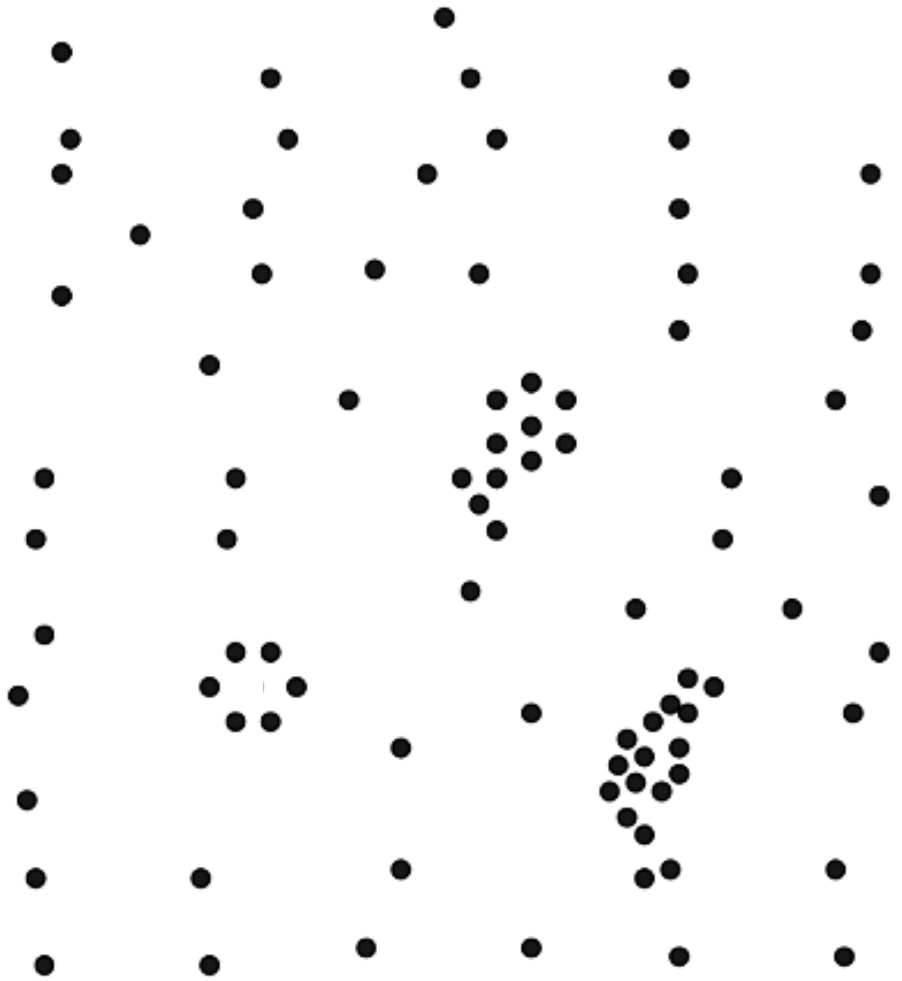
DÉFIS DU REGROUPEMENT

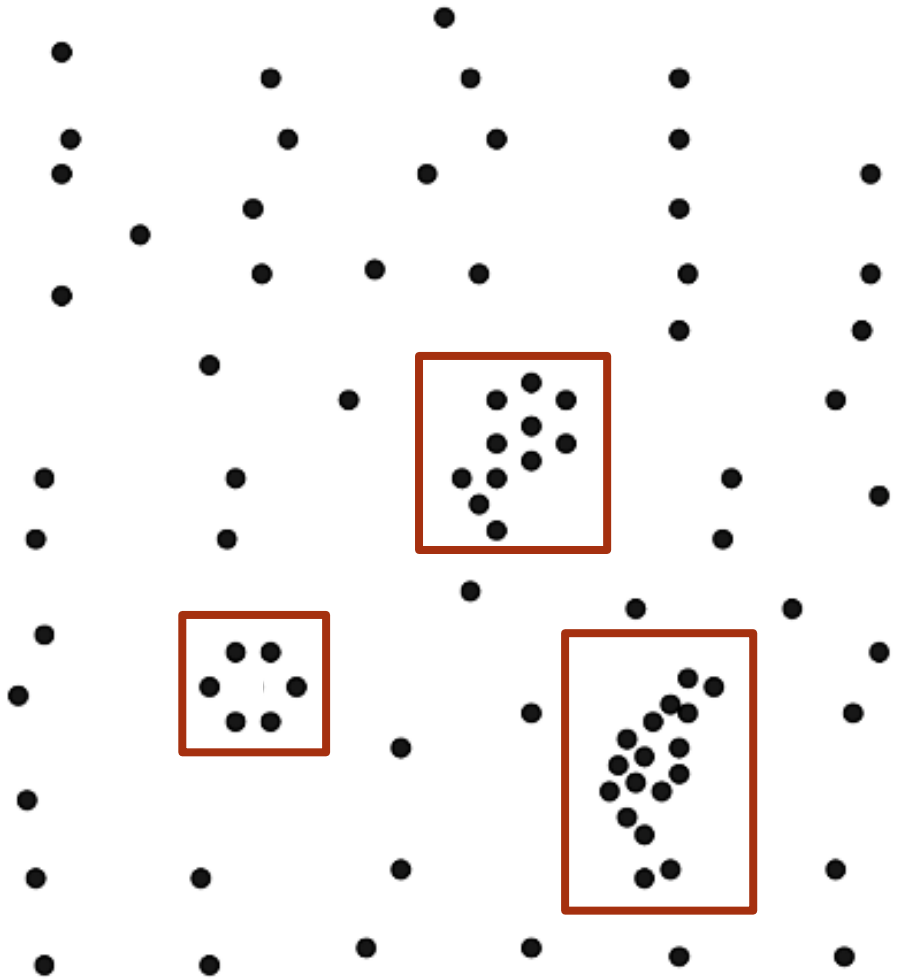
Automatisation
Relativement intuitif pour les humains, mais difficile à automatiser.

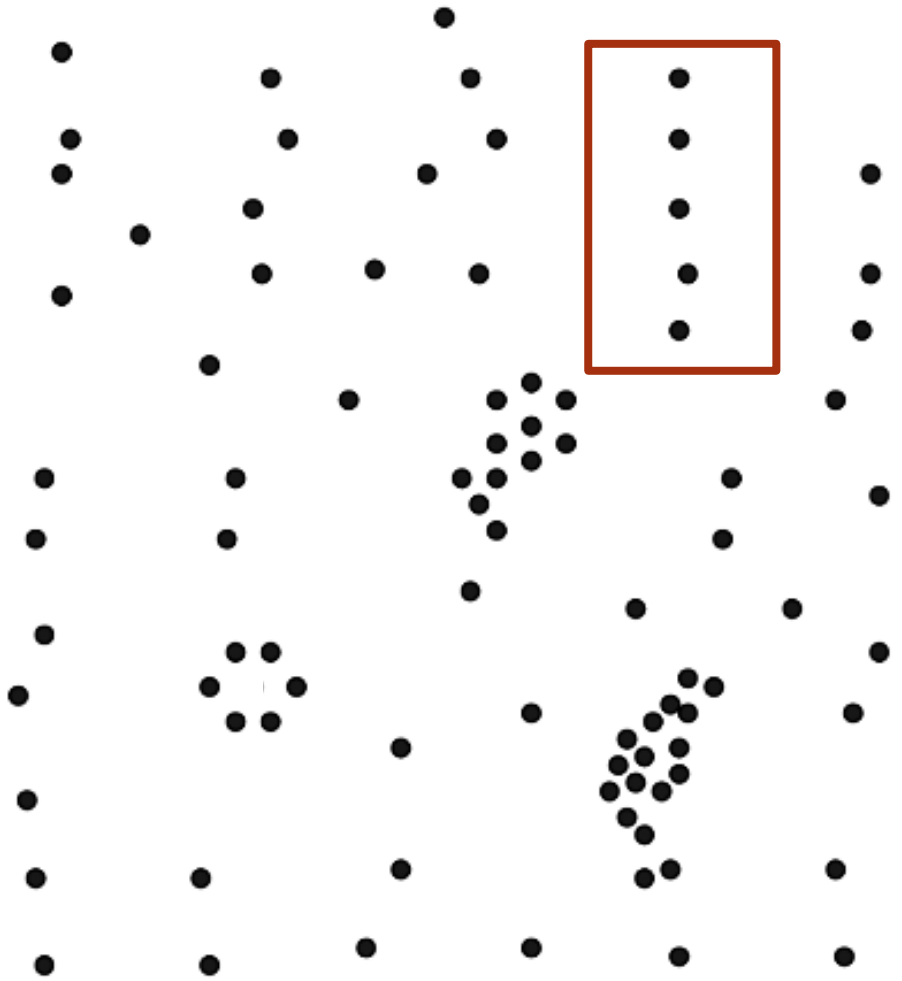
Absence d'une définition précise
Aucun consensus universel de la définition d'un groupe.

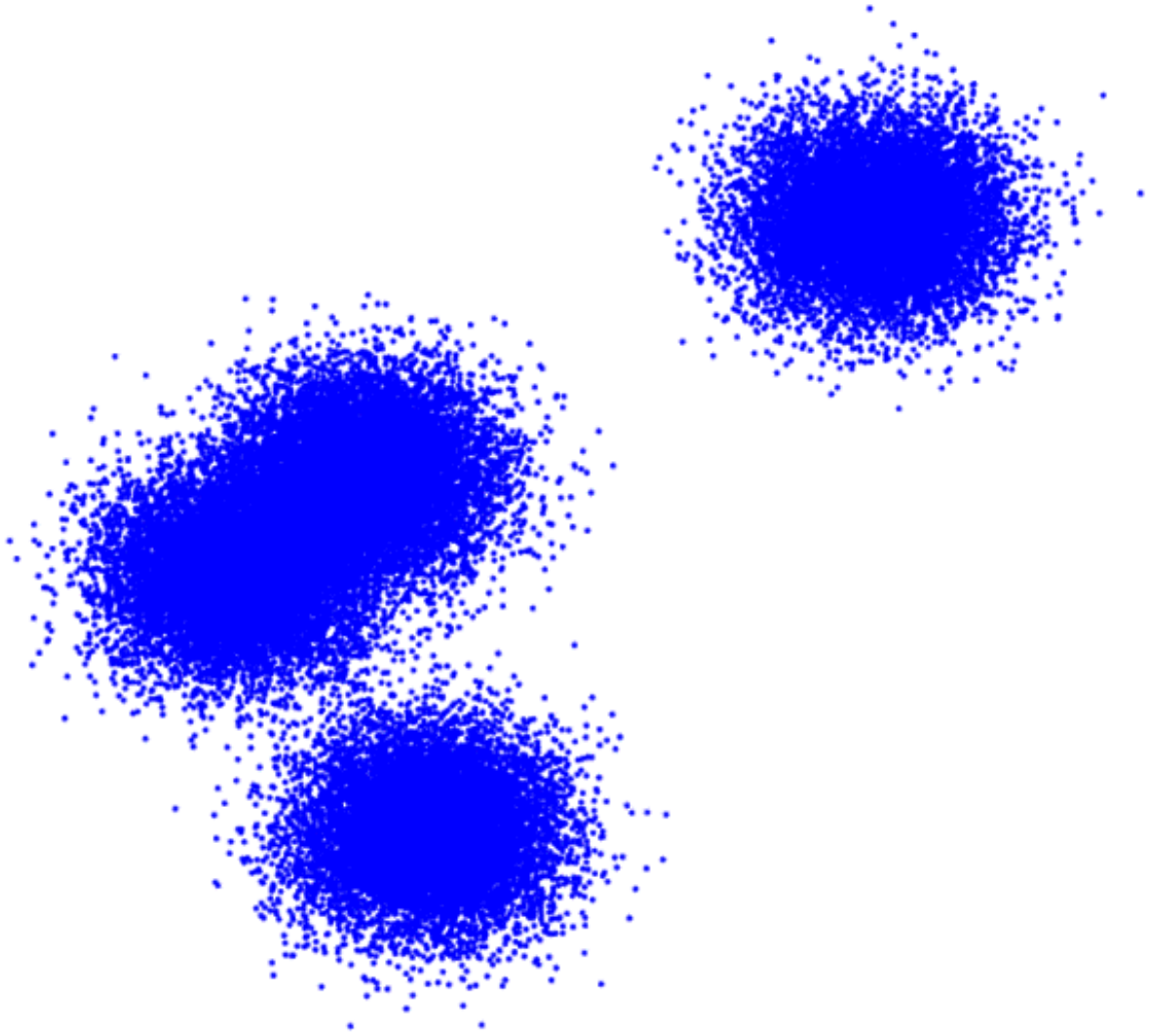
Absence de reproductibilité
Non déterministe : le même algorithme, exécuté deux fois sur le même ensemble de données, peut donner lieu à des groupes totalement différents.

Nombre de groupes
Il est difficile de déterminer le nombre optimal de groupes.









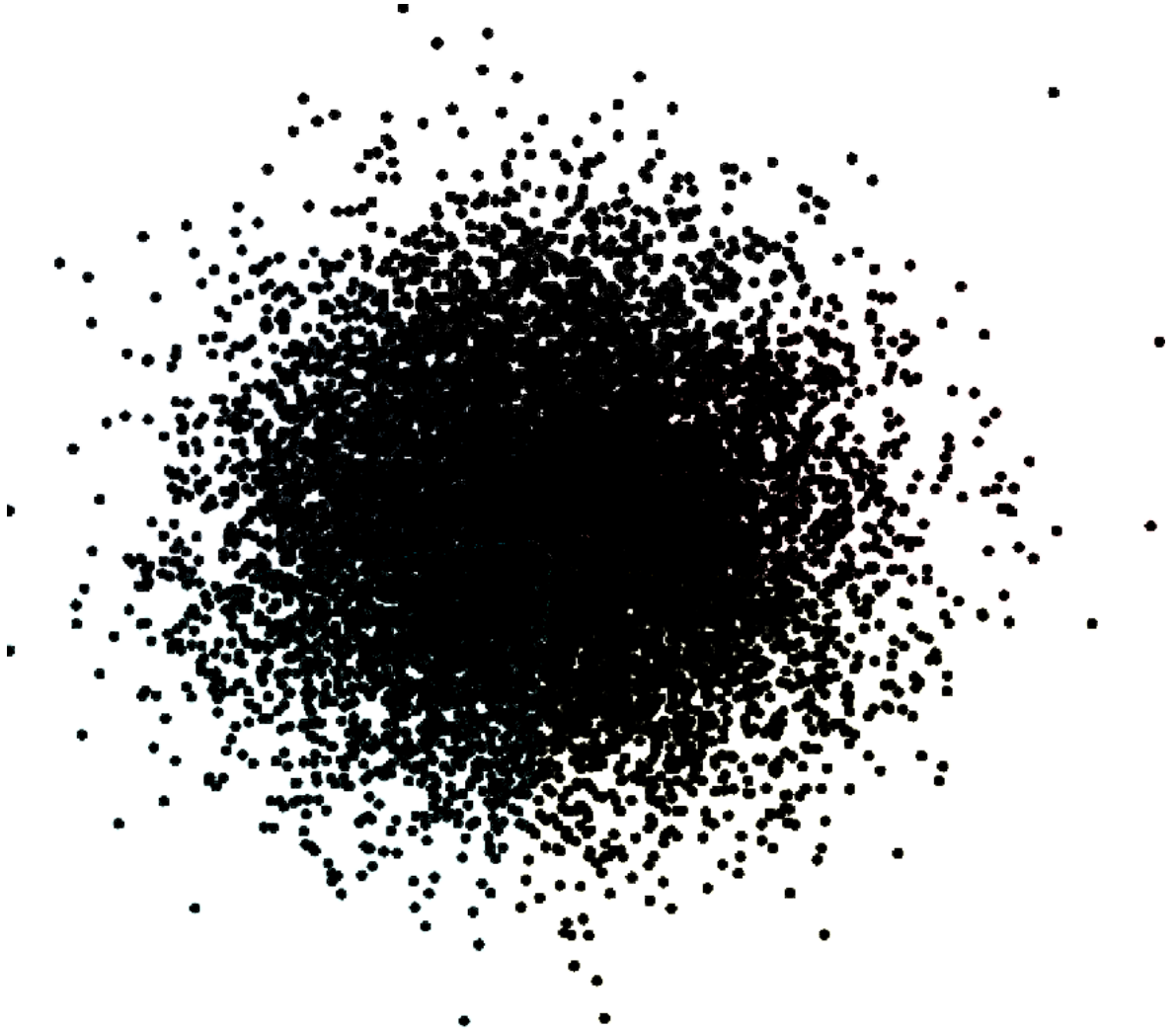
DÉFIS DU REGROUPEMENT

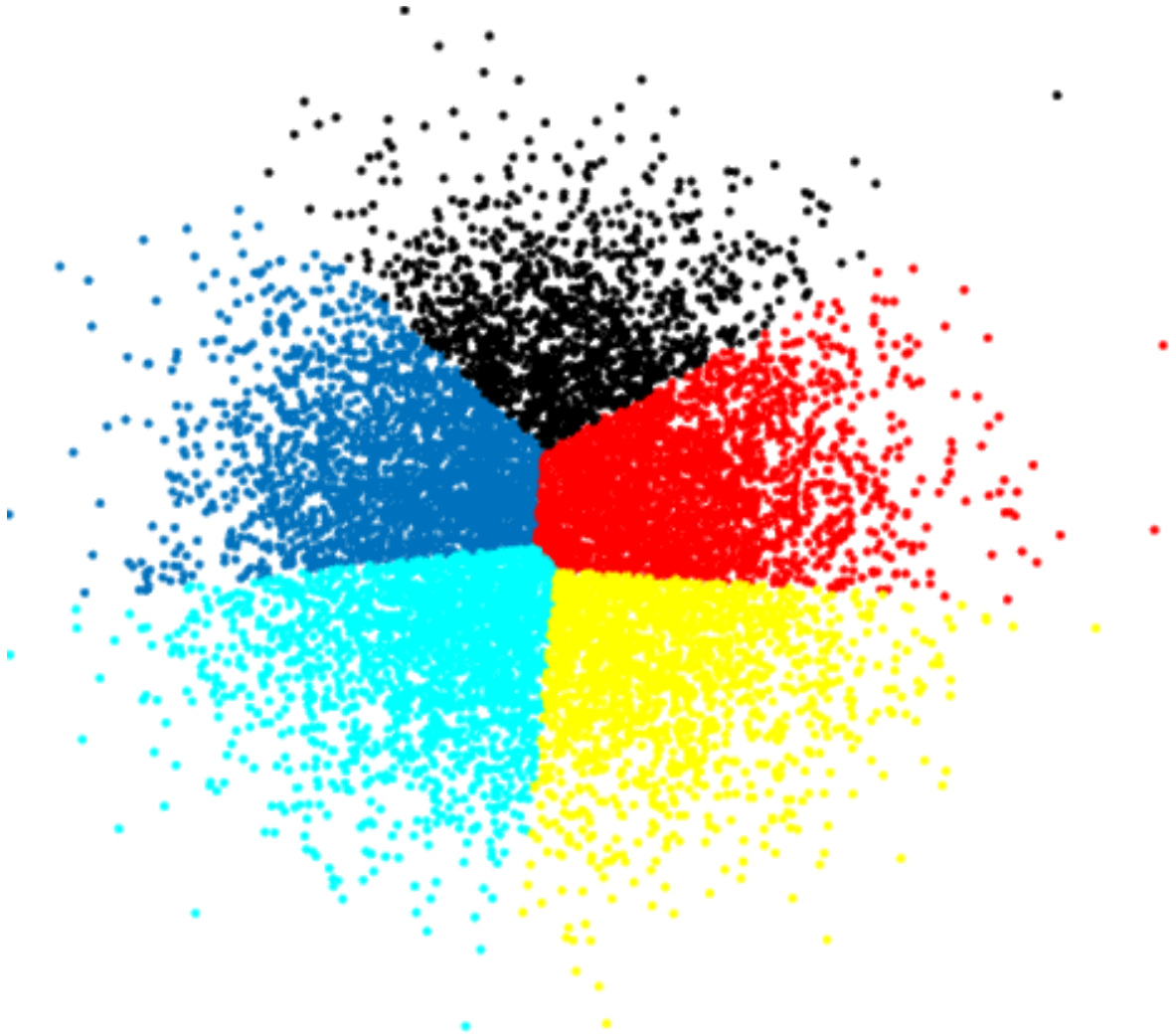
Description d'un groupe
Devons-nous décrire les groupes au moyen d'instances représentatives ou de valeurs moyennes?

Validation modèle
Il n'existe aucun renseignement sur le véritable regroupement par rapport auquel nous pouvons comparer le modèle de regroupement. Donc, comment pouvons-nous déterminer la pertinence d'un regroupement?

Regroupement fantôme
La plupart des méthodes crée des groupes, même s'il n'y en a aucun dans les données.

Rationalisation *a posteriori*
Après la création des groupes, il est tentant d'essayer de les « expliquer »...





EXEMPLE : IRIS

REGROUPEMENT

« Il n'est pas nécessaire que les étudiants en science des données soient jardiniers, mais ça aide. »
(anonyme)

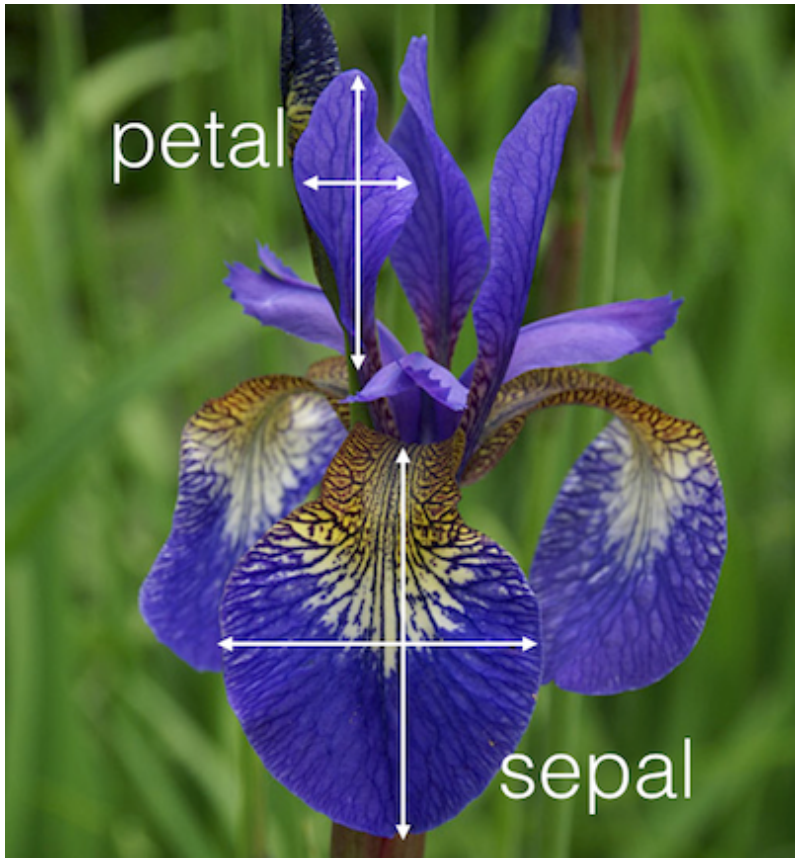
[http://blog.kaggle.com/wp-content/uploads/2015/04/iris_petal_sepal.png]

EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS

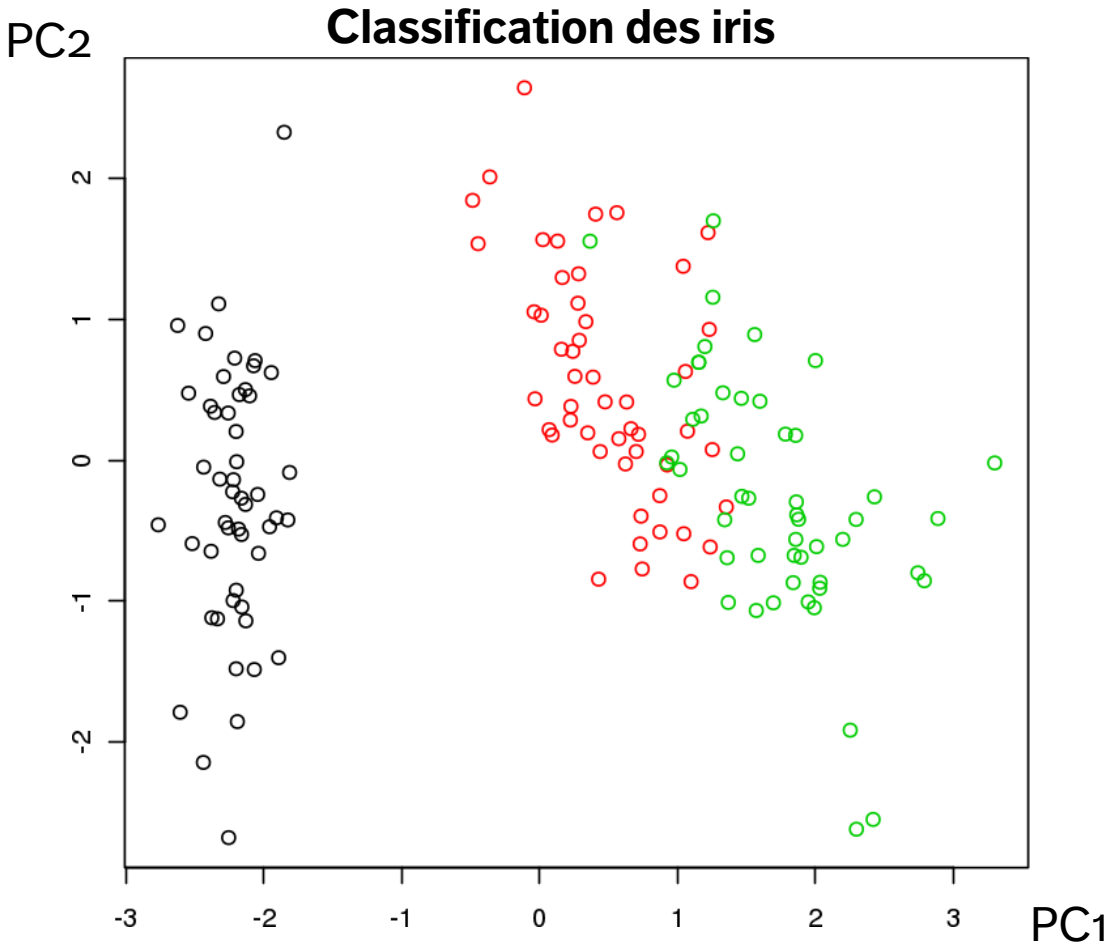
L'iris est un genre de plante à fleur voyante.

L'ensemble de données de l'iris de Fisher contient 150 observations regroupées sous cinq attributs pour les spécimens recueillis par Anderson, la plupart dans un pâturage de la Gaspésie dans les années 1930 :

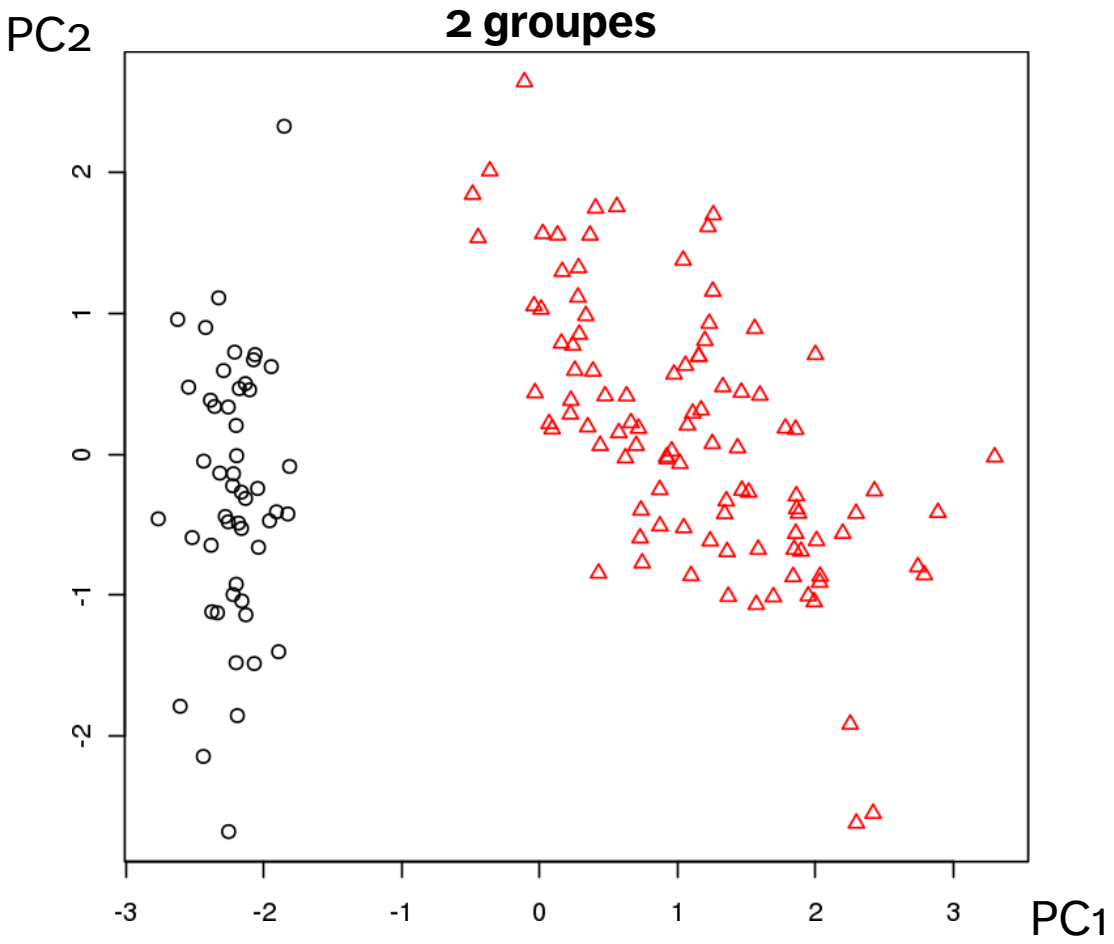
- **largeur des pétales**
- **longueur des pétales**
- **largeur du sépale**
- **longueur du sépale**
- **espèces**



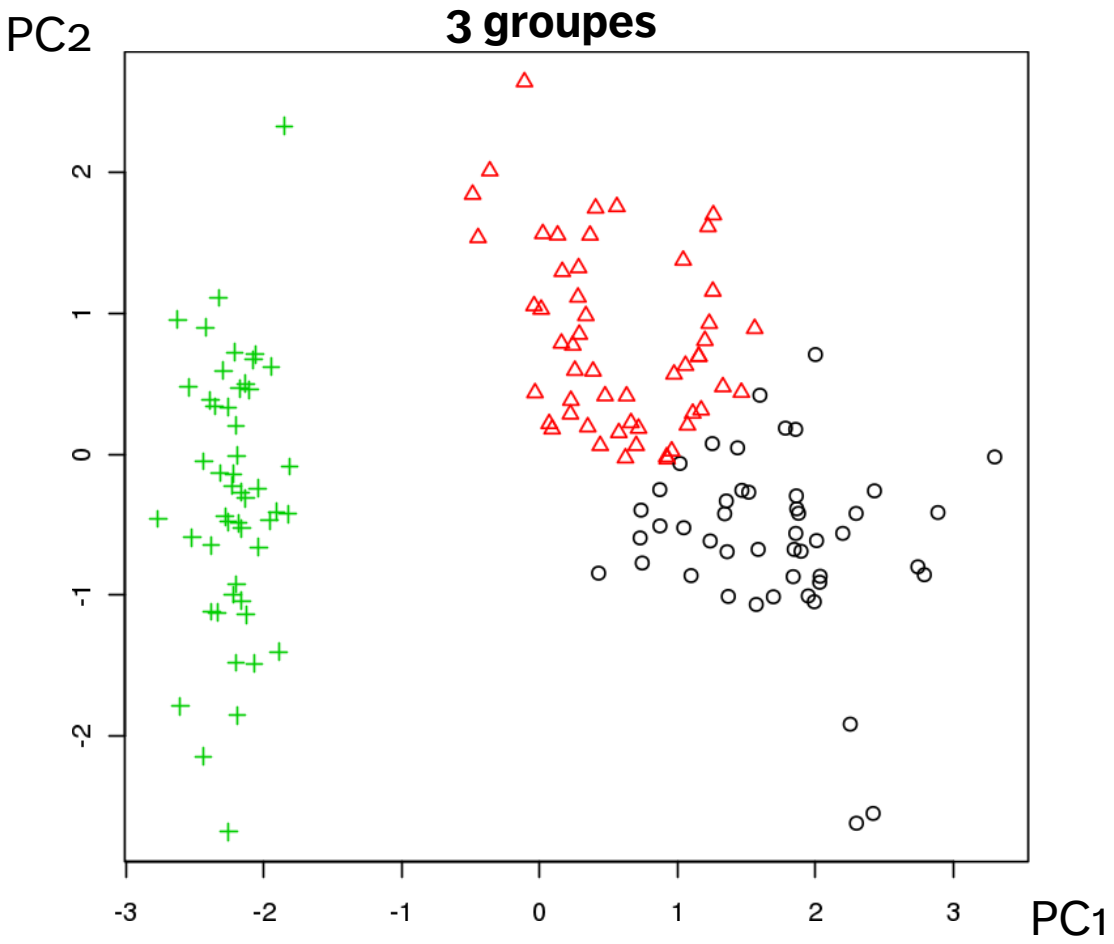
EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



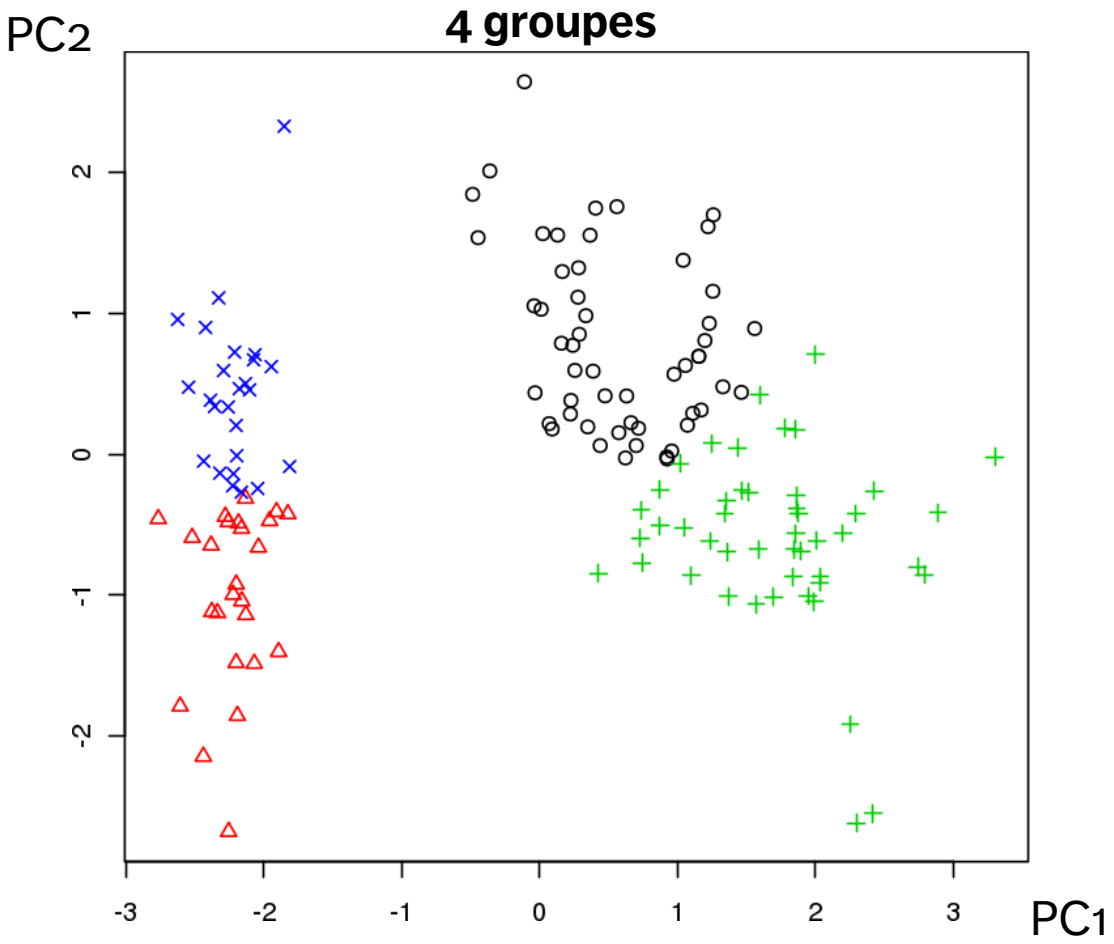
EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



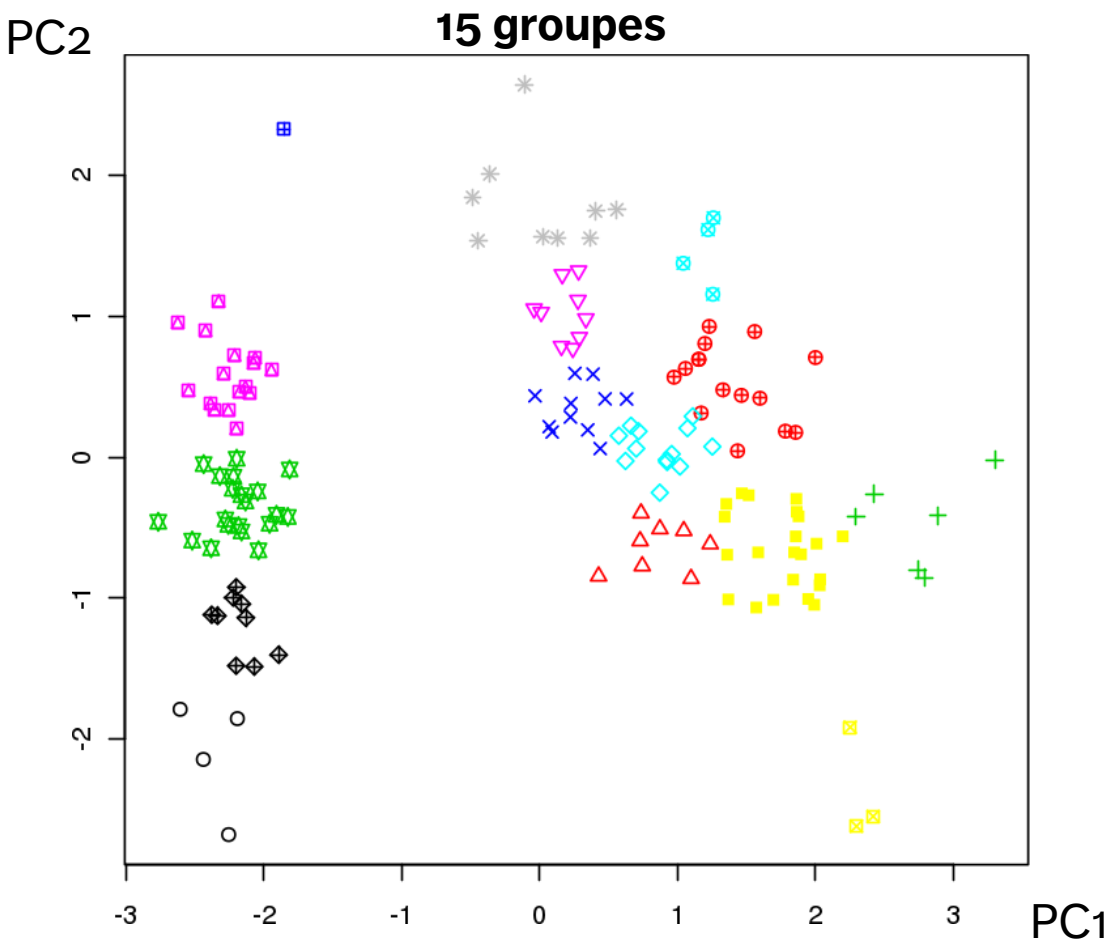
EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



(PETIT) ÉCHANTILLON DES MESURES DE LA QUALITÉ INTERNE

Ball-Hall	Gplus	Scott-Symons
Banfeld-Raftery	KsqDetW	SD
C	LogDetRatio	SDbw
Calinski-Harabasz	LogSSRatio	Silhouette
Davies-Bouldin	McClain-Rao	Tau
Det Ratio	PBM	Trace
Dunn	Point-Biserial	TraceWiB
Baker-Hubert Gamma	Ratkowsky-Lance	Wemmert-Gancarski
GDI	Ray-Turi	Xie-Beni

Que devons-nous faire de toutes ces différentes mesures, supposées sans contexte, de la qualité du regroupement?

(offertes dans le langage R au moyen de la fonction `clusterCrit()`)

VALIDATION INTERNE D'UN REGROUPEMENT

Indice de Davies-Bouldin

$$DB = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} \frac{s_i + s_j}{d(c_i, c_j)},$$

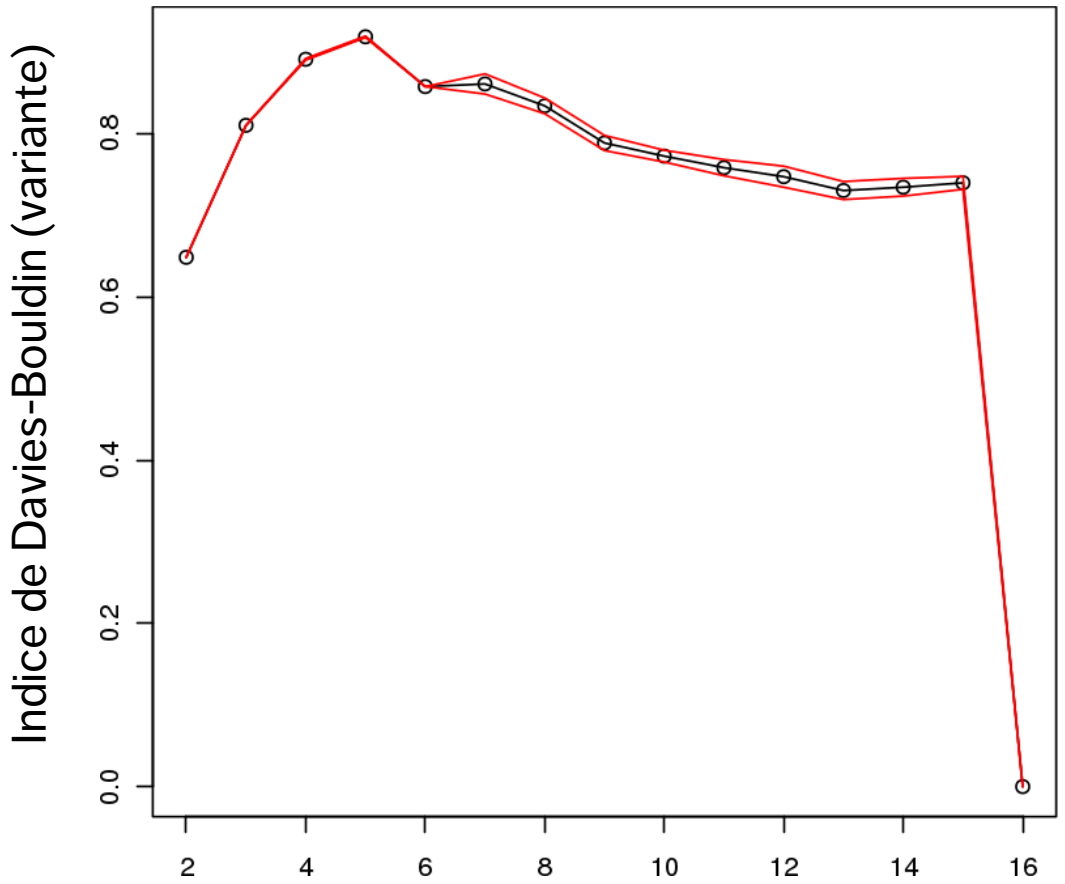
où N représente le nombre de groupes, c_m représente le centre de masse du m^{e} groupe, et s_m représente la distance moyenne des points dans le groupe m^{e} par rapport à c_m;

peut servir à déterminer le nombre de groupes dans un algorithme des k-moyennes.

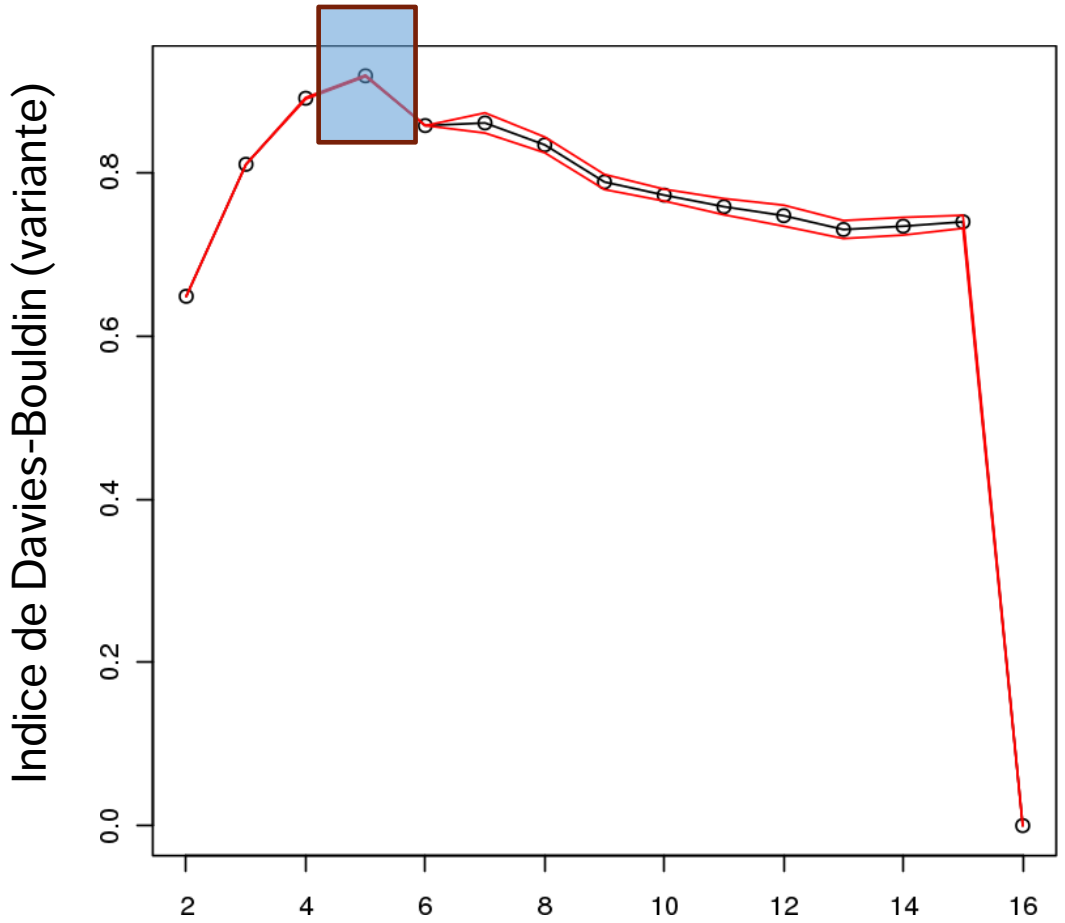
Autres méthodes

- somme des erreurs proportionnelles au carré, indice de Dunn, indice de silhouette, etc.

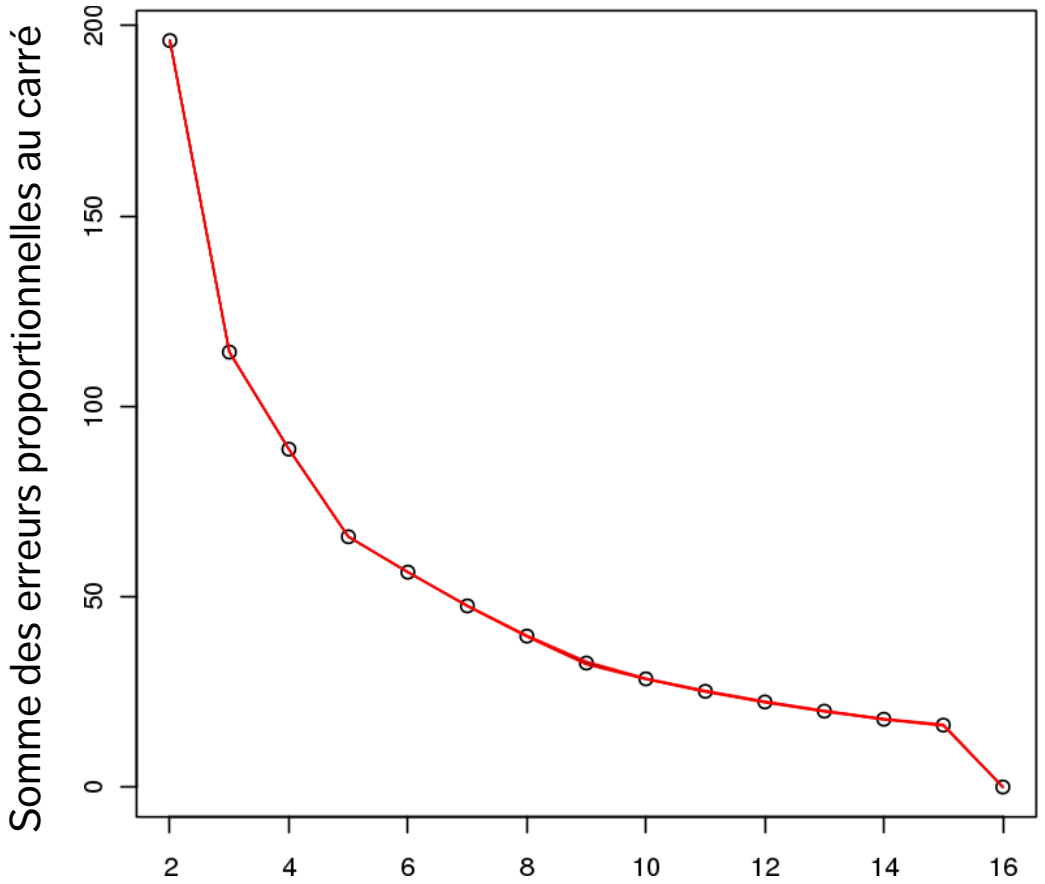
EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



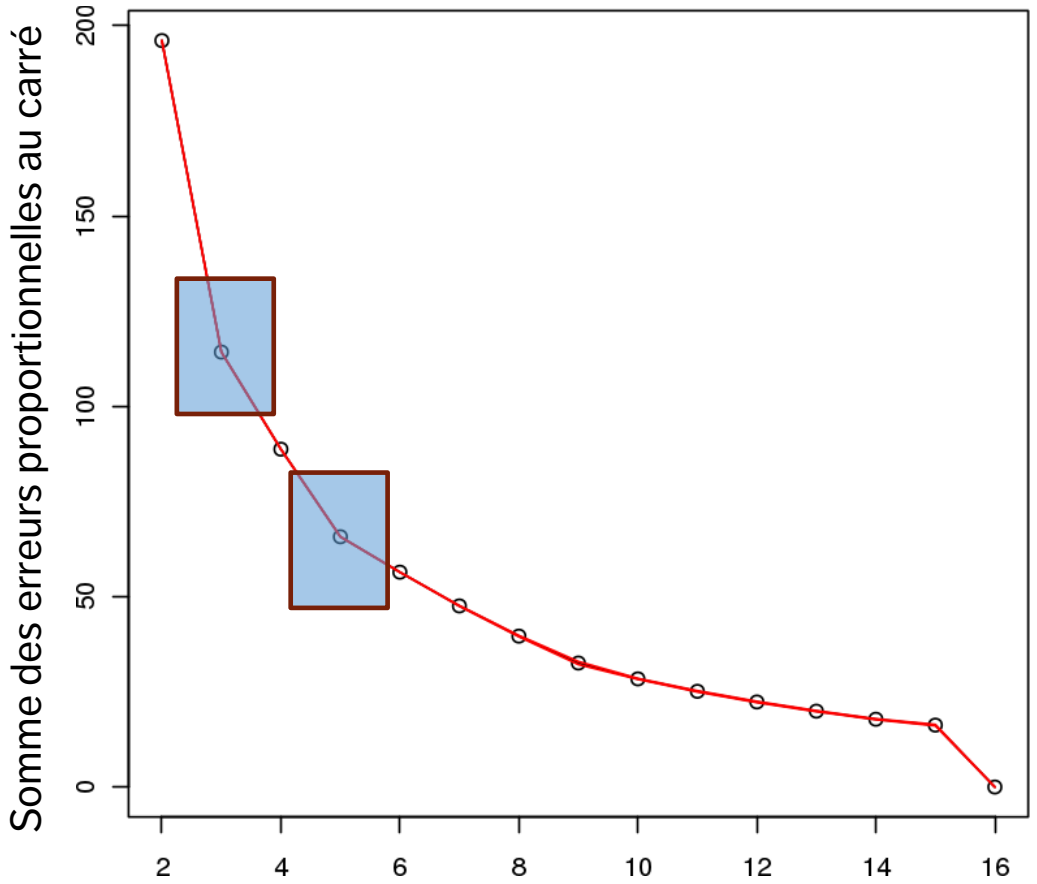
EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS

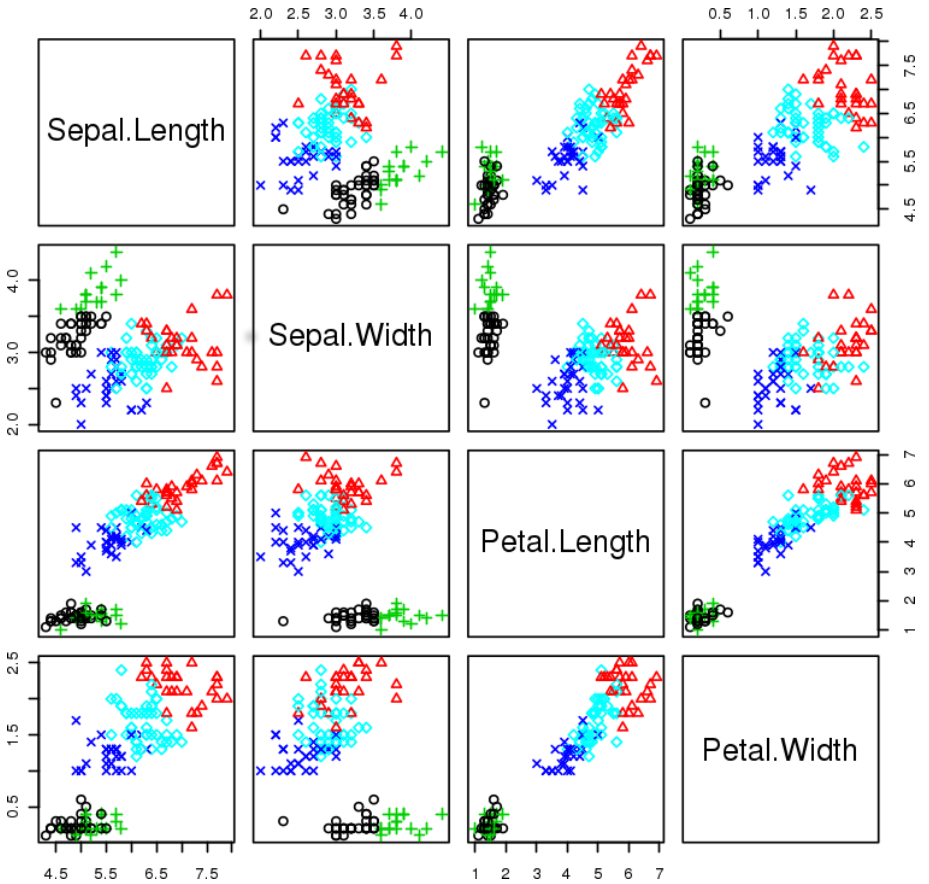


EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS



EXEMPLE – ENSEMBLE DE DONNÉES DE L'IRIS

5 groupes (1 doublon)



DISCUSSION

Est-ce que c'est un « bon » modèle de regroupement?

RÉFÉRENCES

REGROUPEMENT

MATIÈRE SUPPLÉMENTAIRE

Regroupement hiérarchique
https://www.data-action-lab.com/wp-content/uploads/2019/03/Hierarchical-Clustering.pdf

DBSCAN
https://www.data-action-lab.com/wp-content/uploads/2019/03/Density-Based-Clustering.pdf

Regroupement spectral
https://www.data-action-lab.com/wp-content/uploads/2019/03/Spectral-Clustering.pdf

Cahiers sur les regroupements
https://www.data-action-lab.com/wp-content/uploads/2019/03/ClusteringNotebooks.zip

RÉFÉRENCES

https://en.wikipedia.org/wiki/Davies–Bouldin_index

https://algobeans.com/2015/11/30/k-means-clustering-laymans-tutorial/

http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

Aggarwal, C.C. et Reddy, C.K. (éditeurs), *Data Custering: Algorithms and Applications*, CRC Press, 2014.

Torgo, L., *Data Mining with R: Learning with Case Studies,* 2e édition, CRC Press, 2017.

Aggarwal, C.C., *Data Mining: the Textbook*, Springer, 2015.

Maheshwari, A.K., *Business Intelligence and Data Mining*, Business Expert Press, 2015.

Leskovec, J., Rajamaran, A. et Ullman, J.D., *Mining of Massive Datasets*, Cambridge Press, 2014.

RÉFÉRENCES

Provost, F. et Fawcett, T., *Data Science for Business*, O'Reilly, 2013.

Hastie, T., Tibshirani, R. et J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2e édition, Springer, 2008.

Frank, E. et Witten, I.H., *Data Mining: Practical Machine Learning Tools and Techniques*, 2e édition, Elsevier, 2005.

https://en.wikipedia.org/wiki/Cluster_analysis

RÉFÉRENCES

Équipe principale de R et collaborateurs du monde entier, *hclust {stats}, R Documentation: Hierarchical Clustering*, version 3.3.0. des statistiques de la trousse. Récupéré le 11 octobre 2016.

Wikipédia, *Hierarchical clustering*. Dernière modification : 16 septembre 2016. Récupéré le 11 octobre 2016.

Manning, C.D., P. Raghavan et H. Schütze, *Hierarchical agglomerative clustering*, Introduction to Information Retrieval. Publié (en ligne) le 7 avril 2009.

Vogler, R., *Hierarchical Clustering with R (feat. D3.js and Shiny)*. Publié le 14 décembre 2014. Récupéré le 11 octobre 2016.

Mazza, R., D. Brodbeck, M. Lanza et R. Wettel, *Introduction to Visualizing Hierarchies*. Récupéré le 11 octobre 2016.

Greenacre, M. et R. Primicerio, *Hierarchical cluster analysis, in Multivariate Analysis of Ecological Data,* Fundación BBVA, 2013, Plaza de San Nicolás, 4 48005 Bilbao, 2013.

TIBCO Spotfire Documentation: What is a Treemap? Dernière modification : 12 février 2015. Récupéré le 11 octobre 2016.

RÉFÉRENCES

Wikipédia, *Silhouette (clustering)*. Dernière modification : 15 juillet 2016. Récupéré le 11 octobre 2016.

Maechler, M., P. Rousseeuw, A. Struyf et M. Hubert, *silhouette {cluster} Compute or Extract Silhouette Information from Clustering*, version 2.0.3 du groupe de la trousse. Publié le 8 octobre 2016.

Vendramin, L., R.J.G.B. Campello and E.R. Hruschka, *Relative Clustering Validity Criteria: A Comparative Overview*, Wiley InterScience. Publié en ligne le 30 juin 2010.

Gower, J. C., *A General Coefficient of Similarity and Some of Its Properties*, Biometrics, 1971.

Buttigieg, P.L., et A. Ramette, *Dissimilarity Measures*, A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses.

O'Connor, B., *Cosine similarity, Pearson correlation, and OLS coefficients*.

Benoît, G., *Data Mining Portfolio Similarity and Dissimilarity Measures*.

Greenacre, M., et R. Primicerio, *Measures of distance and correlation between variables*, Multivariate Analysis of Ecological Data, 2013.

RÉFÉRENCES

d'Huy, J., *Scientists Trace Society's Myths to Primordial Origins*, Scientific American (en ligne). Publié le 29 septembre 2016. Récupéré le 11 octobre 2016.

Habib, U., K. Hayat et G. Zucker, *Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering*, Complex Adaptive Systems Modeling, 4:8, 2016.

Orłowska, M., K. Pytlakowskae, A. Mrozek-Wilczkiewicz, R. Musioł, M. Waksmundzka-Hajnos, M. Sajewicz et T. Kowalska, *A Comparison of Antioxidant, Antibacterial, and Anticancer Activity of the Selected Thyme Species by Means of Hierarchical Clustering and Principal Component Analysis*, Acta Chromatographica, 28, 2016.

Codd, M., J. Mehegan, C. Kelleher et A. Drummond, *Use of hierarchical cluster analysis to classify prisons in Ireland into mutually exclusive drug-use risk categories*.

Patnai, A.K., P.K. Bhuyan et K.V.K. Rao, *Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets*.

RÉFÉRENCES

Johnson, Jesse, *Clusters and DBScan*. Publié le 20 août 2013. Récupéré le 11 octobre 2016.

Développeurs de scikit-learn, *Comparing different clustering algorithms on toy datasets*, scikit-learn. Récupéré le 11 octobre 2016.

Bäcklund, Henrik, Anders Hedblom et Niklas Neijman, *Data Mining TNM033 Notes: DBSCAN A Density-Based Spatial Clustering of Application with Noise*. Publié le 30 novembre 2011. Récupéré le 11 octobre 2016.

Ligges, Uwe, et Martin Mächler, *Scatterplot3d: an R package for Visualizing Multivariate Data,* 2003. Récupéré le 11 octobre 2016.

Wikipédia, *DBSCAN*.

Schubert, Erich, Sander, Jörg, Ester, Martin, Kriegel, Hans Peter et Xu, Xiaowei, *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN*, ACM Trans. Database Syst. **42** (3): 19:1–19:21, doi:10.1145/3068335, ISSN 0362-5915, juillet 2017.

RÉFÉRENCES

Plant, C., S.J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A.W. Bokde, H. Hampel et M. Ewers, *Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease*.

Panchami, V.U., N. Radhika et A.V. Vidyapeetham, *A Novel Approach for Predicting the Length of Hospital Stay With DBSCAN and Supervised Classification Algorithms*.

Francis, Z., C. Villagrasa et I. Clairand, *Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm*.

Jawad, A., K. Kersting et N.V. Andrienko, *Where traffic meets DNA: mobility mining using biological sequence analysis revisited*.

Schoier, G., et G. Borruso, *Individual movements and geographical data mining. clustering algorithms for highlighting hotspots in personal navigation routes*.

RÉFÉRENCES

von Luxburg, Ulrike, *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics.

Singh, Aarti, *Spectral Clustering*.

Ng, Andrew Y., Michael I. Jordan et Yair Weiss, *On Spectral Clustering: Analysis and an Algorithm.*

Wang, Jing, *An Introduction to Support Vector Machine and Spectral Clustering.*

Zelnik-Manor, Lihi, et Pietro Perona, *Self-Tuning Spectral Clustering.*

Hamad, Denis, et Philippe Biela, *Introduction to spectral clustering.*

Meila, Marina, *Classic and Modern Data Clustering*.

Kung, H.T., et Dario Vlah, *A Spectral Clustering Approach to Validating Sensors via Their Peers in Distributed Sensor Networks*.

Chehreghani, Morteza, Alberto Busetto et Joachim Buhmann, *Information Theoretic Model Validation for Spectral Clustering*.

Kamvar, Sepandar D., Dan Klein et Christopher Manning, *Spectral Clustering*.

RÉFÉRENCES

Vendramin, L., Campello, R. J. G. B. et Hruschka, E., *Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining*, 3. 209-235. 10.1002/sam.10080, 2010.

Amigó, E., Gonzalo, J., Artiles, J. et Verdejo, M., *Comparison of extrinsic clustering evaluation metrics based on formal constraints*, Information Retrieval, 12, 461-486, 2009.

Lewis, J. M., Ackerman, M. et de Sa, V., *Human Cluster Evaluation and Formal Quality Measures: A Comparative Study*, compte-rendu de la 34e conférence de la Cognitive Science Society, 2012.

Desgraupes, Bernard, *Clustering Indices*, Lab Modal'X, Université Paris-Ouest, 2013.

Cranshaw, Justin, Raz Schwartz, Jason I. Hong et Norman Sadeh, *The Livehoods Project: Utilizing Social Media to Understand the Dynamics of A City*, compte-rendu de la 6e Conférence internationale de l'AAAI sur les blogues web et les réseaux sociaux (ICWSM–12), Dublin, Irlande, pp. 1–8, 2012.

Kung, H. T. et Vlah, D.A., *Spectral clustering approach to validating sensors via their peers in distributed sensor networks*, compte-rendu de la 18e Conférence internationale de l'IEEE sur les communications informatisées et les réseaux (ICCCN '09), 2009.

MATIÈRE SUPPLÉMENTAIRE

REGROUPEMENT

COMPARAISON DES MESURES ENTRE DES ENSEMBLES DE DONNÉES

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Point-Biserial	A	0.000	0.046	0.226	0.247	0.262	0.289	0.306	0.373	0.390	0.408	0.488	0.555	0.566	0.571	0.584	0.636	0.642	0.645	0.694	0.705	0.729	0.736	0.768	0.822	0.837	1.107
Tau	B	-0.046	0.000	0.180	0.201	0.216	0.243	0.260	0.327	0.344	0.362	0.442	0.509	0.520	0.525	0.538	0.590	0.597	0.599	0.649	0.659	0.683	0.690	0.722	0.776	0.791	1.061
C/k$^{1/2}$	C	-0.226	-0.180	0.000	0.021	0.036	0.063	0.080	0.147	0.164	0.182	0.263	0.329	0.340	0.345	0.358	0.410	0.417	0.419	0.469	0.479	0.504	0.510	0.542	0.596	0.611	0.881
ASWC	D	-0.247	-0.201	-0.021	0.000	0.015	0.042	0.060	0.127	0.143	0.161	0.242	0.308	0.319	0.324	0.338	0.390	0.396	0.398	0.448	0.458	0.483	0.489	0.521	0.575	0.590	0.860
ASSWC	E	-0.262	-0.216	-0.036	-0.015	0.000	0.027	0.045	0.112	0.128	0.146	0.227	0.293	0.304	0.309	0.323	0.375	0.381	0.384	0.433	0.443	0.468	0.474	0.506	0.560	0.575	0.846
PBM	F	-0.289	-0.243	-0.063	-0.042	-0.027	0.000	0.017	0.084	0.101	0.119	0.199	0.266	0.277	0.282	0.295	0.347	0.353	0.356	0.406	0.416	0.440	0.447	0.479	0.533	0.548	0.818
SWC	G	-0.306	-0.260	-0.080	-0.060	-0.045	-0.017	0.000	0.067	0.083	0.102	0.182	0.249	0.260	0.265	0.278	0.330	0.336	0.339	0.388	0.399	0.423	0.430	0.462	0.516	0.530	0.801
SSWC	H	-0.373	-0.327	-0.147	-0.127	-0.112	-0.084	-0.067	0.000	0.016	0.035	0.115	0.181	0.193	0.198	0.211	0.263	0.269	0.272	0.321	0.332	0.356	0.363	0.395	0.449	0.463	0.734
Dunn12	I	-0.390	-0.344	-0.164	-0.143	-0.128	-0.101	-0.083	-0.016	0.000	0.018	0.099	0.165	0.176	0.181	0.195	0.247	0.253	0.255	0.305	0.315	0.340	0.346	0.378	0.432	0.447	0.717
Dunn62	J	-0.408	-0.362	-0.182	-0.161	-0.146	-0.119	-0.102	-0.035	-0.018	0.000	0.080	0.147	0.158	0.163	0.176	0.228	0.234	0.237	0.287	0.297	0.321	0.328	0.360	0.414	0.429	0.699
Dunn13	K	-0.488	-0.442	-0.263	-0.242	-0.227	-0.199	-0.182	-0.115	-0.099	-0.080	0.000	0.066	0.078	0.082	0.096	0.148	0.154	0.157	0.206	0.217	0.241	0.248	0.280	0.334	0.348	0.619
VRC	L	-0.555	-0.509	-0.329	-0.308	-0.293	-0.266	-0.249	-0.181	-0.165	-0.147	-0.066	0.000	0.011	0.016	0.030	0.082	0.088	0.090	0.140	0.150	0.175	0.181	0.213	0.267	0.282	0.552
Ball and Hall	M	-0.566	-0.520	-0.340	-0.319	-0.304	-0.277	-0.260	-0.193	-0.176	-0.158	-0.078	-0.011	0.000	0.005	0.018	0.070	0.076	0.079	0.129	0.139	0.163	0.170	0.202	0.256	0.271	0.541
Trace(W)	N	-0.571	-0.525	-0.345	-0.324	-0.309	-0.282	-0.265	-0.198	-0.181	-0.163	-0.082	-0.016	-0.005	0.000	0.013	0.065	0.072	0.074	0.124	0.134	0.159	0.165	0.197	0.251	0.266	0.536
DB	O	-0.584	-0.538	-0.358	-0.338	-0.323	-0.295	-0.278	-0.211	-0.195	-0.176	-0.096	-0.030	-0.018	-0.013	0.000	0.052	0.058	0.061	0.110	0.121	0.145	0.152	0.184	0.238	0.252	0.523
Nlog(|T|/|W|)	P	-0.636	-0.590	-0.410	-0.390	-0.375	-0.347	-0.330	-0.263	-0.247	-0.228	-0.148	-0.082	-0.070	-0.065	-0.052	0.000	0.006	0.009	0.058	0.069	0.093	0.100	0.132	0.186	0.200	0.471
Trace(CovW)	Q	-0.642	-0.597	-0.417	-0.396	-0.381	-0.353	-0.336	-0.269	-0.253	-0.234	-0.154	-0.088	-0.076	-0.072	-0.058	-0.006	0.000	0.003	0.052	0.063	0.087	0.094	0.126	0.180	0.194	0.465
k^2|W|	R	-0.645	-0.599	-0.419	-0.398	-0.384	-0.356	-0.339	-0.272	-0.255	-0.237	-0.157	-0.090	-0.079	-0.074	-0.061	-0.009	-0.003	0.000	0.049	0.060	0.084	0.091	0.123	0.177	0.192	0.462
log(SSB/SSW)	S	-0.694	-0.649	-0.469	-0.448	-0.433	-0.406	-0.388	-0.321	-0.305	-0.287	-0.206	-0.140	-0.129	-0.124	-0.110	-0.058	-0.052	-0.049	0.000	0.010	0.035	0.041	0.074	0.128	0.142	0.413
Dunn11	T	-0.705	-0.659	-0.479	-0.458	-0.443	-0.416	-0.399	-0.332	-0.315	-0.297	-0.217	-0.150	-0.139	-0.134	-0.121	-0.069	-0.063	-0.060	-0.010	0.000	0.024	0.031	0.063	0.117	0.132	0.402
Gamma	U	-0.729	-0.683	-0.504	-0.483	-0.468	-0.440	-0.423	-0.356	-0.340	-0.321	-0.241	-0.175	-0.163	-0.159	-0.145	-0.093	-0.087	-0.084	-0.035	-0.024	0.000	0.007	0.039	0.093	0.107	0.378
McClain and Rao	V	-0.736	-0.690	-0.510	-0.489	-0.474	-0.447	-0.430	-0.363	-0.346	-0.328	-0.248	-0.181	-0.170	-0.165	-0.152	-0.100	-0.094	-0.091	-0.041	-0.031	-0.007	0.000	0.032	0.086	0.101	0.371
C-Index	W	-0.768	-0.722	-0.542	-0.521	-0.506	-0.479	-0.462	-0.395	-0.378	-0.360	-0.280	-0.213	-0.202	-0.197	-0.184	-0.132	-0.126	-0.123	-0.074	-0.063	-0.039	-0.032	0.000	0.054	0.069	0.339
|T|/|W|	X	-0.822	-0.776	-0.596	-0.575	-0.560	-0.533	-0.516	-0.449	-0.432	-0.414	-0.334	-0.267	-0.256	-0.251	-0.238	-0.186	-0.180	-0.177	-0.128	-0.117	-0.093	-0.086	-0.054	0.000	0.015	0.285
Trace(W^{-1}B)	Y	-0.837	-0.791	-0.611	-0.590	-0.575	-0.548	-0.530	-0.463	-0.447	-0.429	-0.348	-0.282	-0.271	-0.266	-0.252	-0.200	-0.194	-0.192	-0.142	-0.132	-0.107	-0.101	-0.069	-0.015	0.000	0.270
G(+)	Z	-1.107	-1.061	-0.881	-0.860	-0.846	-0.818	-0.801	-0.734	-0.717	-0.699	-0.619	-0.552	-0.541	-0.536	-0.523	-0.471	-0.465	-0.462	-0.413	-0.402	-0.378	-0.371	-0.339	-0.285	-0.270	0.000
	Mean	0.959	0.913	0.733	0.712	0.697	0.670	0.653	0.586	0.569	0.551	0.471	0.404	0.393	0.388	0.375	0.323	0.316	0.314	0.264	0.254	0.230	0.223	0.191	0.137	0.122	-0.148

Fig. 10 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 25$.

En 2010, Vendramin et ses collègues ont utilisé un certain nombre de tests de référence pour comparer un grand nombre de mesures de validation intrinsèques.

Conclusion générale : des variantes de l'indice de silhouette ont donné de bons résultats d'un test à un autre.

APPRENTISSAGE MACHINE ET APPRENTISSAGE HUMAIN

- Lewis et ses collègues ont comparé six mesures courantes de la qualité des regroupements et ont procédé à une évaluation humaine des résultats des regroupements.
- Principale constatation : l'évaluation humaine des regroupements ressemble davantage à l'indice de silhouette et à l'indice Calinski-Harabasz.
- Il se peut que la validation interne et les mesures de la qualité des regroupements nous en disent plus sur les regroupements dans tous les contextes.
- Il est peut-être plus facile de repérer un groupe manifestement mauvais que toutes les variations d'un bon groupe?

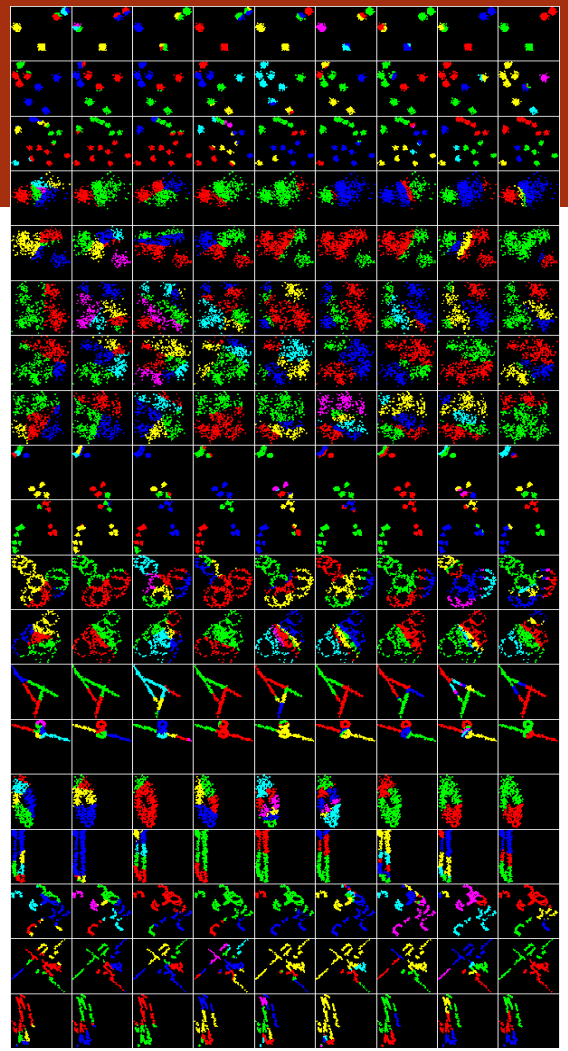


Figure 1: All stimuli. Datasets are in rows; partitions are in columns.

MESURES DE LA CORRÉLATION

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	0	0	0	0	1

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	1	0	0	1	1

Deux résultats très similaires de regroupements (mais le nombre de groupes est différent).

Examinez la corrélation entre les regroupements.

Rand, Jaccard, Gamma

Une corrélation parfaite donne une valeur maximale à la mesure.