

REGROUPEMENT

PRÉPARATION DU TERRAIN

« 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)

TABLE DES MATIÈRES

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

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 :

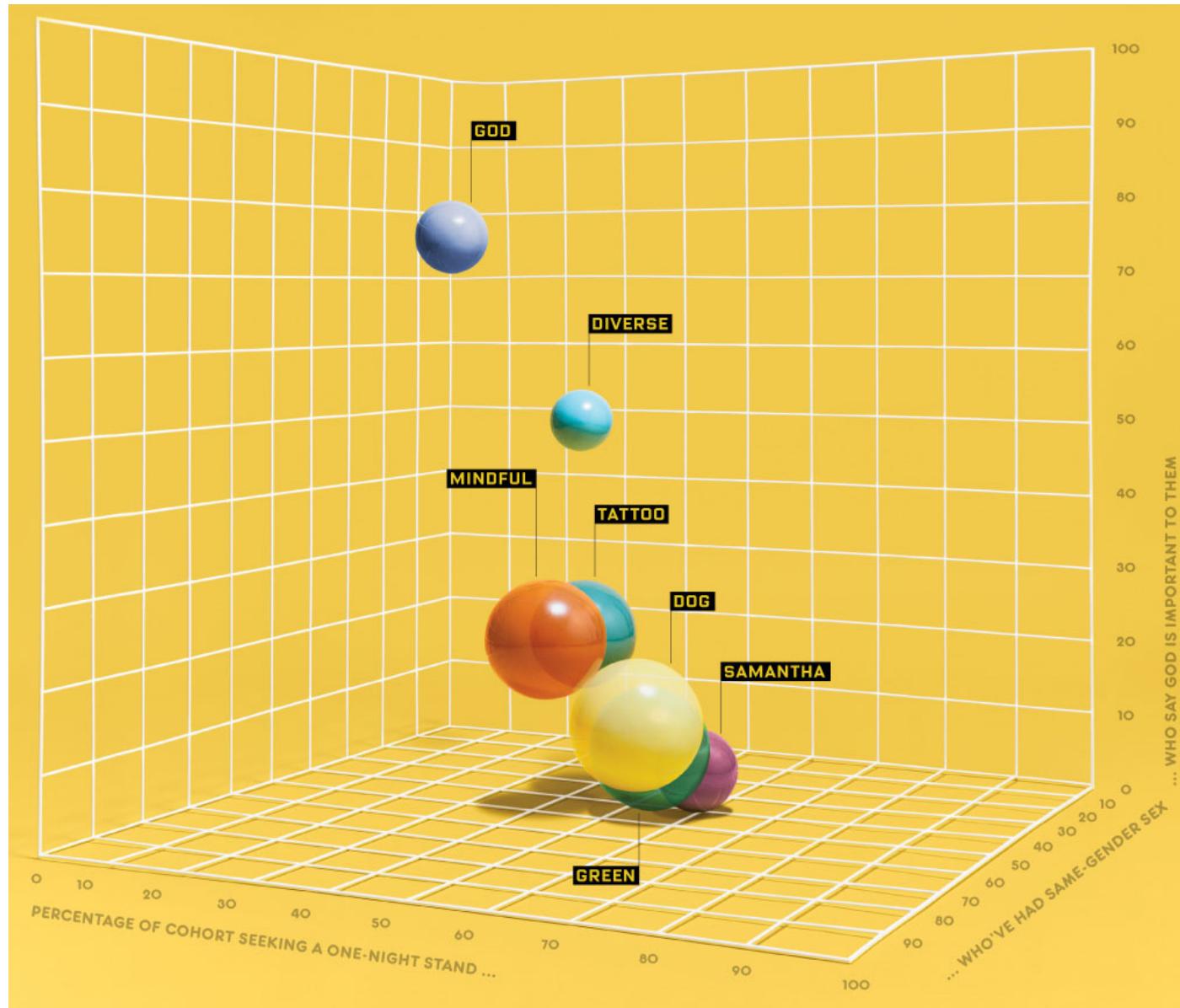
1. **recueillir** des données;
2. recueillir **d'autres données, légèrement meilleures et différentes**;
3. recueillir **encore d'autres** données;
4. déterminer la technique d'exploration des données qui **conviendrait** aux renseignements recherchés (regroupement);
5. **valider** les résultats de l'analyse.

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?

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

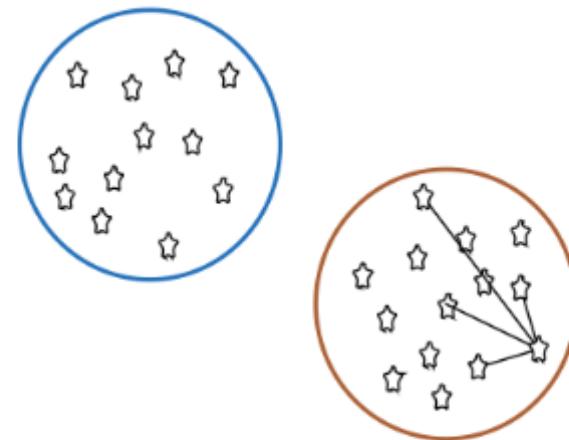


APERÇU DU REGROUPEMENT

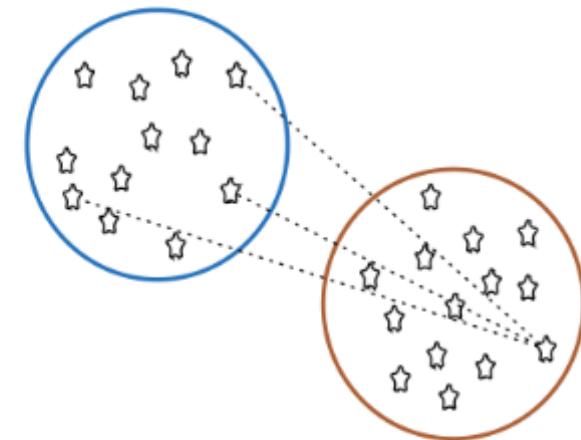
Dans un **regroupement**, les données sont réparties en **groupes formés naturellement**. Dans chaque groupe, les points de données sont **similaires**; d'un groupe à un autre, les points de données sont **distincts**.

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

Groupes

Âge

Clients

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 quand 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.

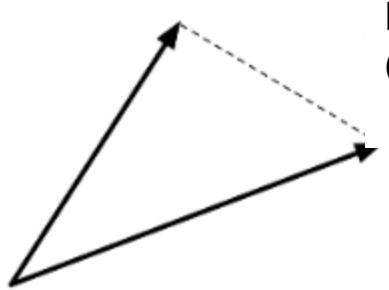
- Malgré tout, il est **très** tentant d'expliquer les groupes *a posteriori*.

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

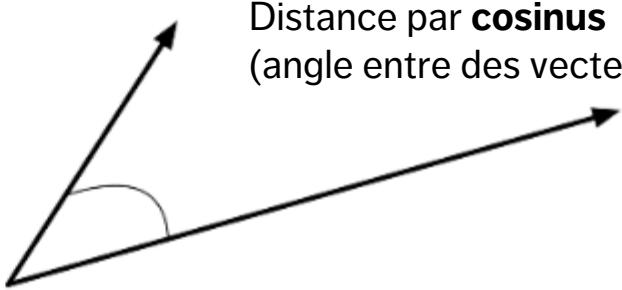
- le même algorithme, exécuté deux fois (ou plus) sur le même ensemble de données, peut donner lieu à des groupes totalement différents;
- l'ordre de présentation des données peut jouer un rôle;
- tout comme la configuration de départ.

EXIGENCE DE REGROUPEMENT

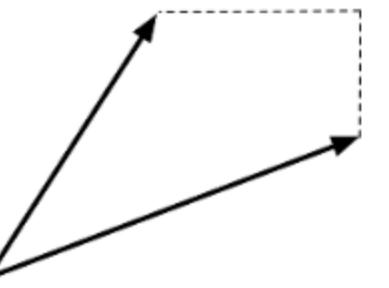
Mesure de la **similarité** w (ou d'une distance d) entre des observations.



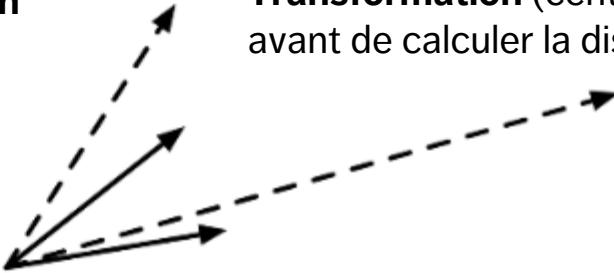
Distance **euclidienne**
(vol d'oiseau)



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



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



Transformation (centre normalisé)
avant de calculer la distance

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

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.

Data

	Y ₁	Y ₂	...	Y _p
01	x _{01,1}	x _{01,2}	...	x _{01,p}
02	x _{02,1}	x _{02,2}	...	x _{02,p}
03	x _{03,1}	x _{03,2}	...	x _{03,p}
04	x _{04,1}	x _{04,2}	...	x _{04,p}
05	x _{05,1}	x _{05,2}	...	x _{05,p}
06	x _{06,1}	x _{06,2}	...	x _{06,p}
07	x _{07,1}	x _{07,2}	...	x _{07,p}
08	x _{08,1}	x _{08,2}	...	x _{08,p}
...			...	
%%	x _{%%,1}	x _{%%,2}	...	x _{%%,p}

Cluster Assignment

	Y ₁	Y ₂	...	Y _p	■
01	x _{01,1}	x _{01,2}	...	x _{01,p}	■
02	x _{02,1}	x _{02,2}	...	x _{02,p}	■
03	x _{03,1}	x _{03,2}	...	x _{03,p}	■
04	x _{04,1}	x _{04,2}	...	x _{04,p}	■
05	x _{05,1}	x _{05,2}	...	x _{05,p}	■
06	x _{06,1}	x _{06,2}	...	x _{06,p}	■
07	x _{07,1}	x _{07,2}	...	x _{07,p}	■
08	x _{08,1}	x _{08,2}	...	x _{08,p}	■
...			...		■
%%	x _{%%,1}	x _{%%,2}	...	x _{%%,p}	■

External Info
(if available, appropriate)

	▲
01	▲
02	▲
03	▲
04	▲
05	▲
06	▲
07	▲
08	▲
...	...
%%	▲

Clustering Algorithm

Model

Clustering Validation

Deployment

MODÈLES DE REGROUPEMENT

***k*-moyennes (*k*-means)**

Regroupement hiérarchique

Allocation de Dirichlet latente

Maximisation de l'espérance

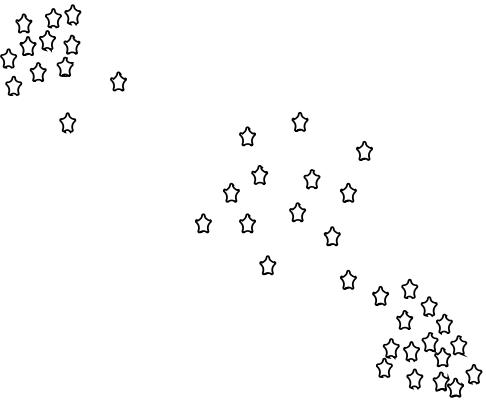
Réduction et regroupement itératifs et équilibrés au moyen de hiérarchies (BIRCH)

Regroupement par densité spatiale des applications avec bruit (DBSCAN)

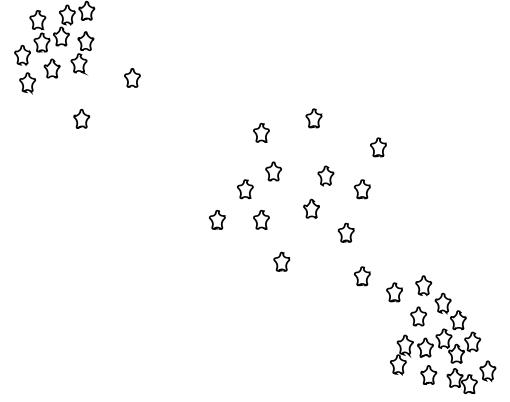
Propagation par affinités

Regroupement spectral

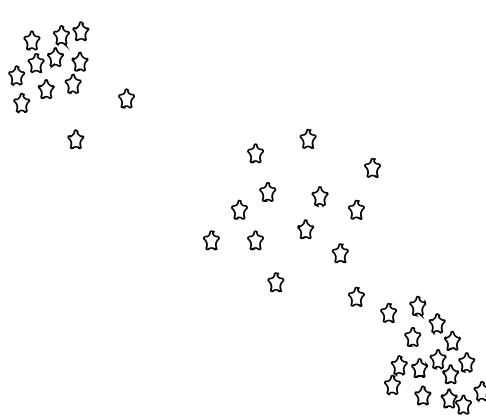
FORME GÉNÉRALE D'UN ALGORITHME DE REGROUPEMENT



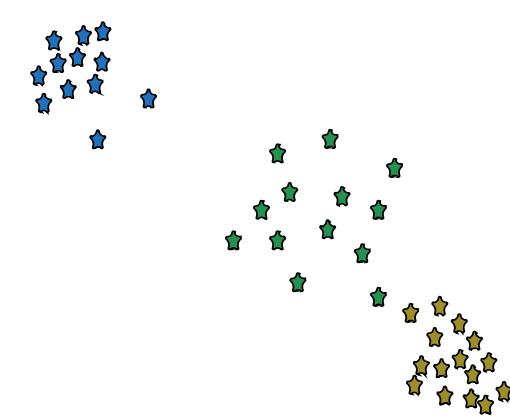
Initialization



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

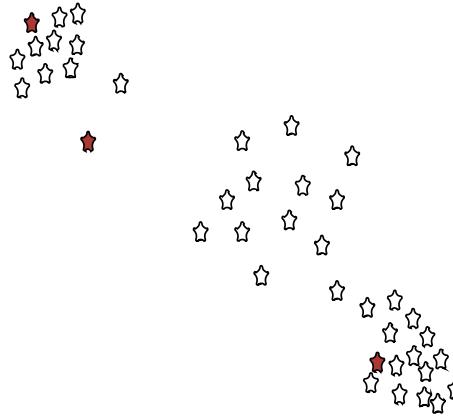


Clustering Step B (Usually Repeated,
Possibly in Conjunction with Previous Step)

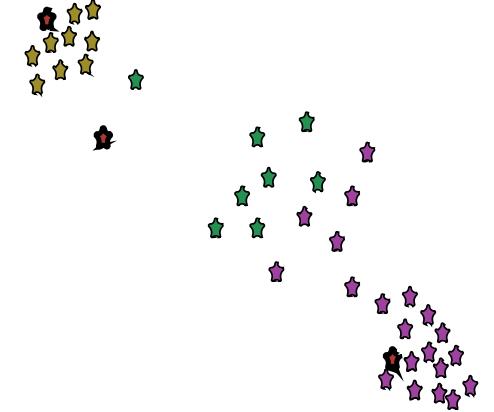


End Condition (Usually When Iterations of
Steps A and B Produce Stable Results)

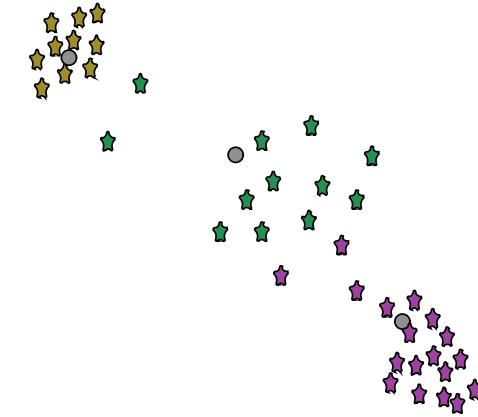
SURVOL DE L'ALGORITHME DES K-MOYENNES



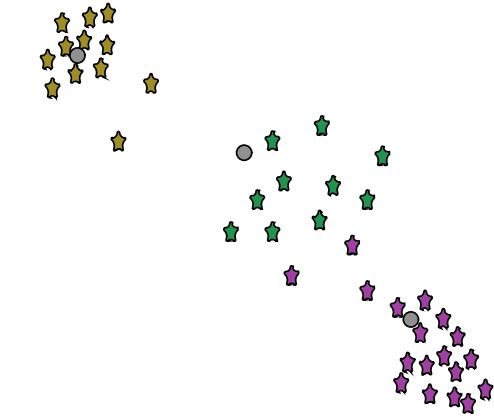
Initialization (e.g. Randomly Pick k Centers)



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 ET POINTS FAIBLES DES k -MOYENNES

Facile à créer

Souvent une méthode **naturelle** d'examiner les regroupements

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

Vous pouvez assigner les points de données à un seul groupe

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

Vous devez présumer que les groupes sont distincts (discrets)

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?

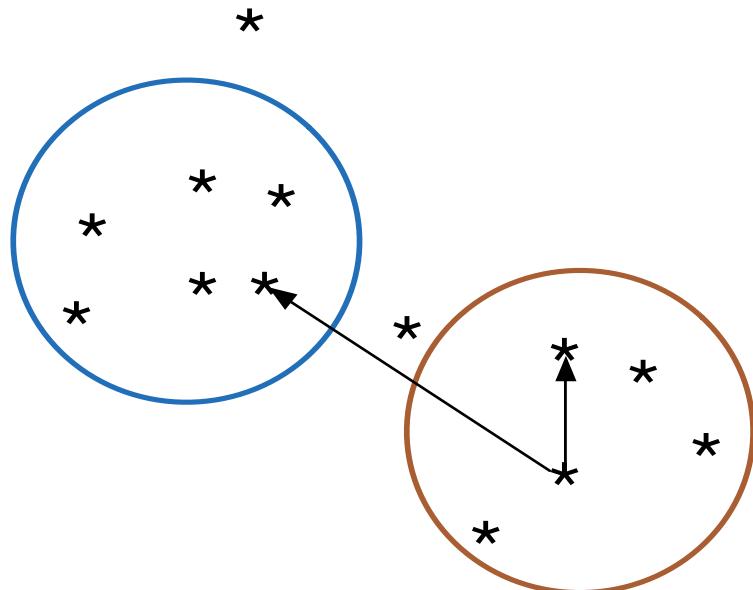
Y a-t-il plusieurs regroupements « naturels » possibles?

Serait-il possible d'utiliser des regroupements différents?

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



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**.

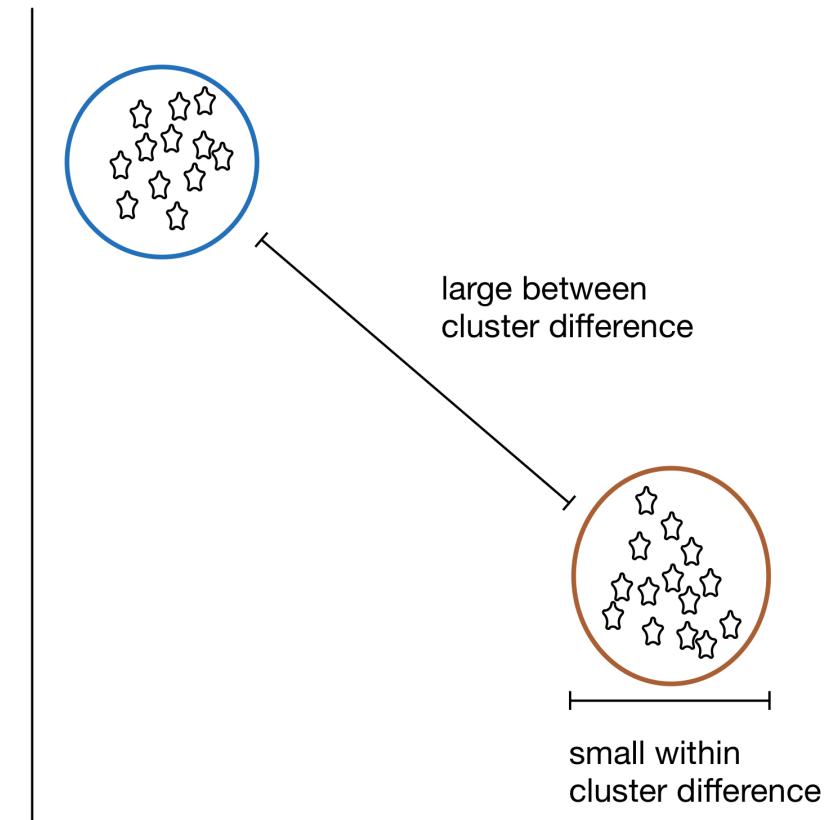
OBJECTIFS TRÈS GÉNÉRAUX

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.

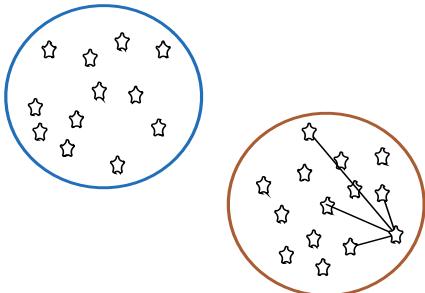
Dans certains cas, comment pouvons-nous pondérer les aspects positifs (p. ex., une note élevée pour la similarité à l'intérieur d'un groupe) et les aspects négatifs (p. ex., une note faible pour la séparation des groupes).

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

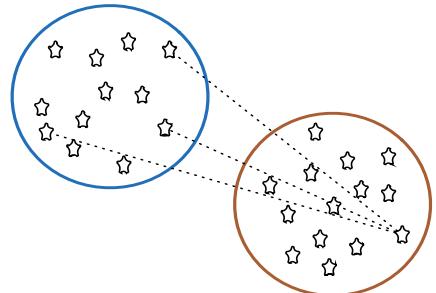


INDICE DE SILHOUETTE

average distance to points in own cluster (low is good)



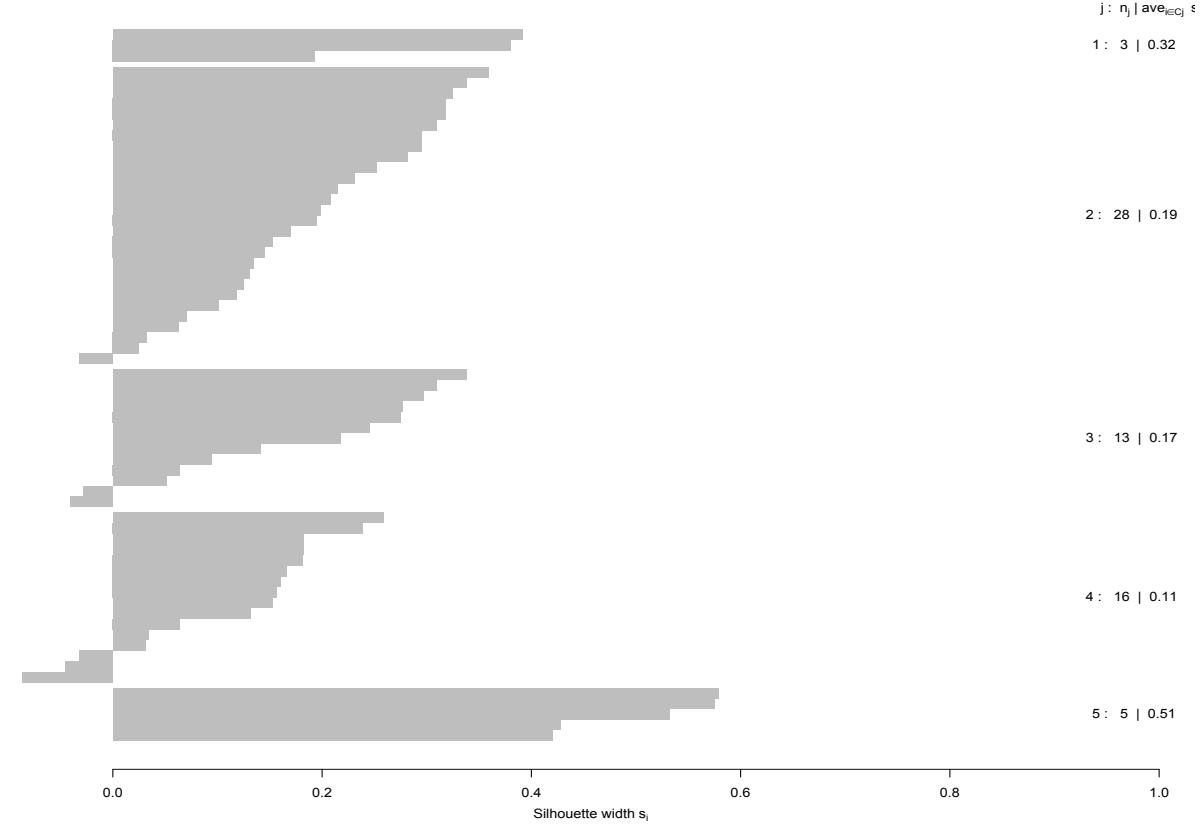
average distance to points in neighbouring cluster (high is good)



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

Silhouette plot of pam(x = ndf, k = 5)

n = 65



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

(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)

DÉFIS DU REGROUPEMENT

Automatisation

Absence d'une définition précise

Absence de reproductibilité

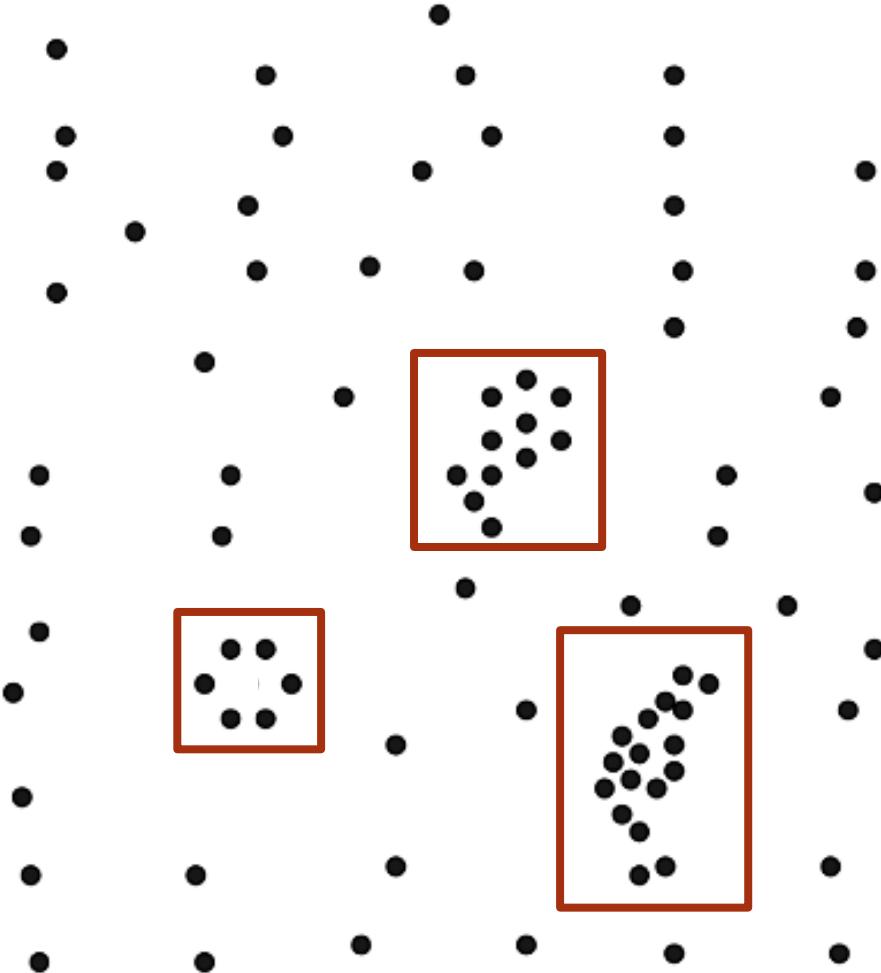
Nombre de groupes

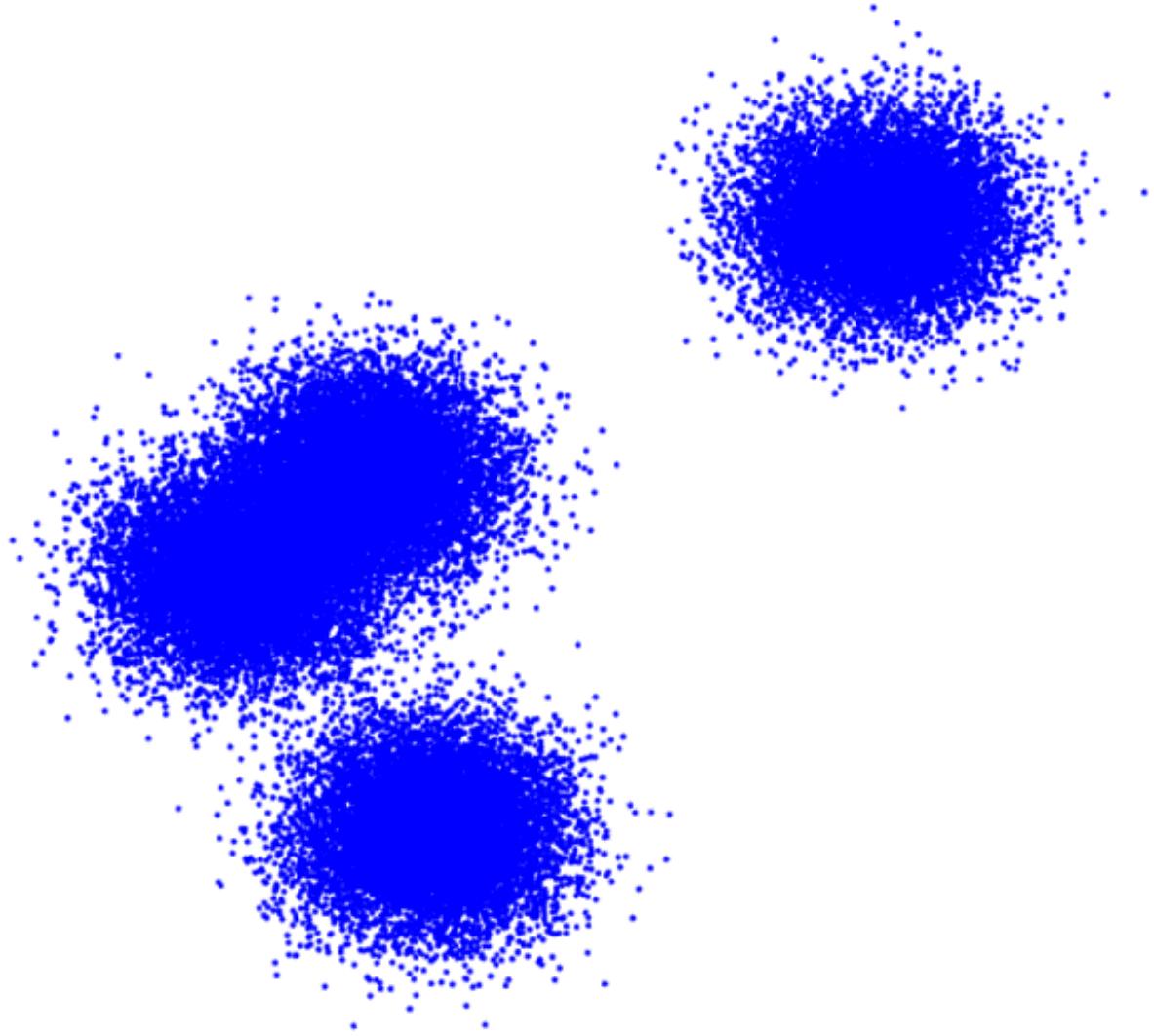
Description d'un groupe

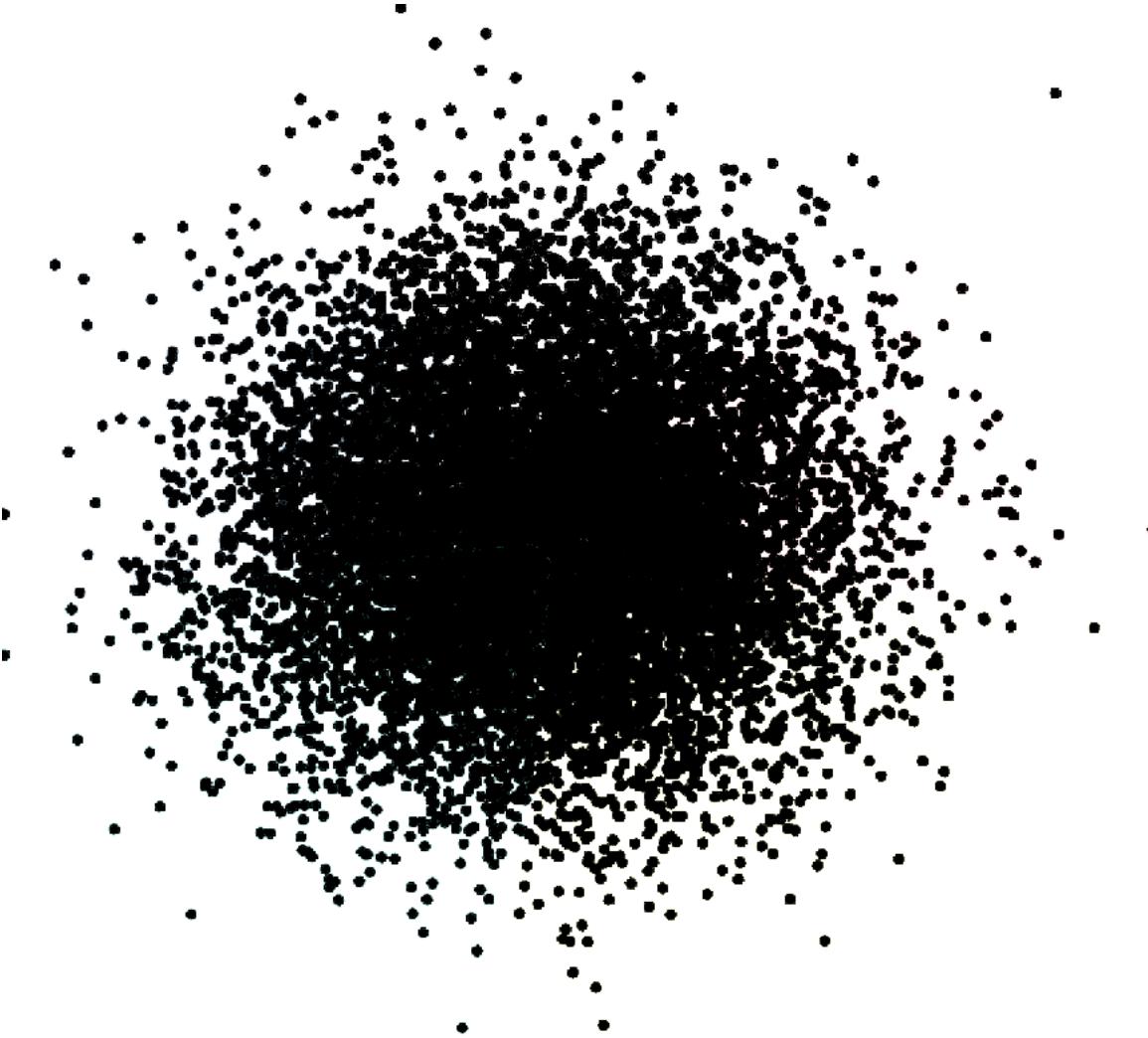
Validation modèle

Regroupement fantôme

Rationalisation *a posteriori*







IDLEWYLD Sysabee DAVHILL

data-action-lab.com

