Regroupement

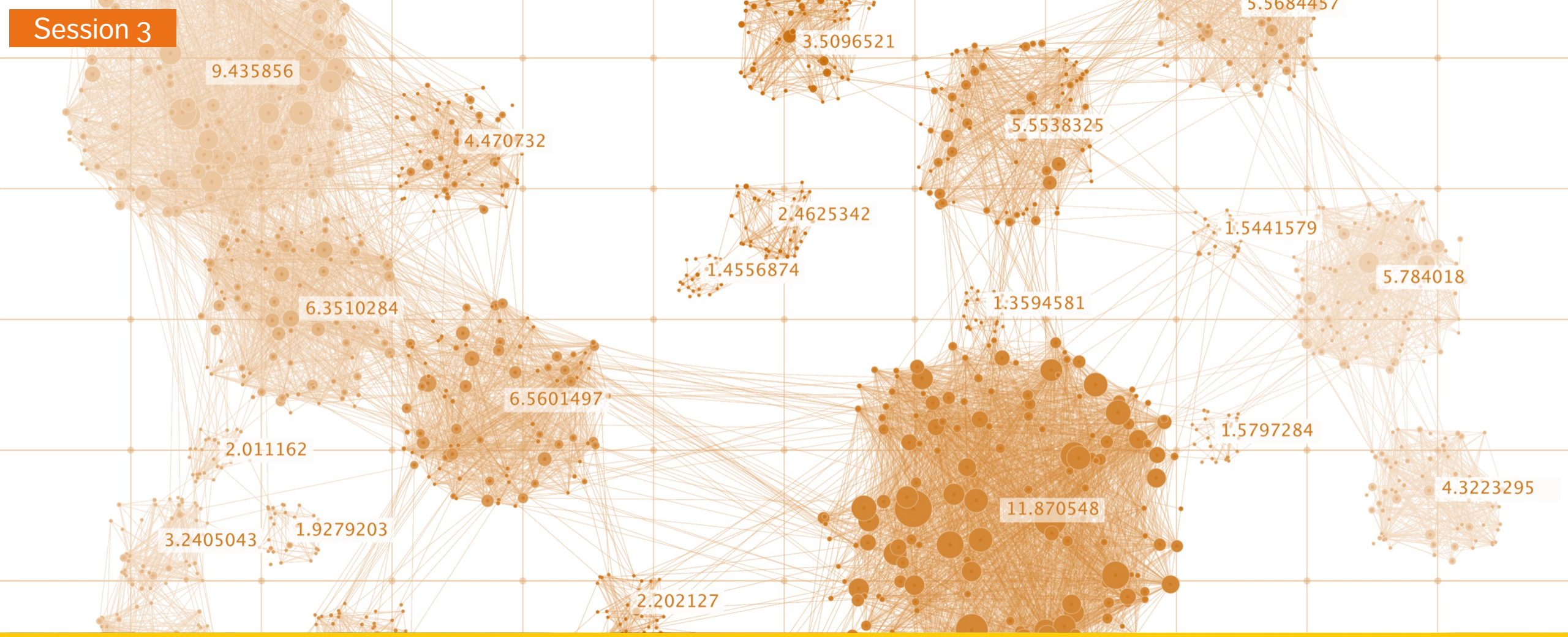
INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

Le regroupement est dans la perception de ceux qui jette un coup d'oeil. Les chercheurs ont proposé de nombreux principes et modèles d'induction dont les problèmes d'optimisation correspondant ne peuvent être résolus qu'approximativement par un nombre encore plus grand d'algorithmes.

[V. Estivill-Castro, *Why So Many Clustering Algorithms* ?]

Les malheurs se regroupent. Rares sont les malheurs solitaires ; ils aiment le train, ils se marchent sur les pieds.

[E. Young]



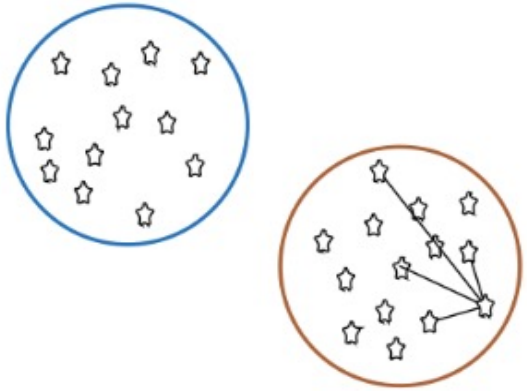
7. Aperçu du regroupement

Vue d'ensemble

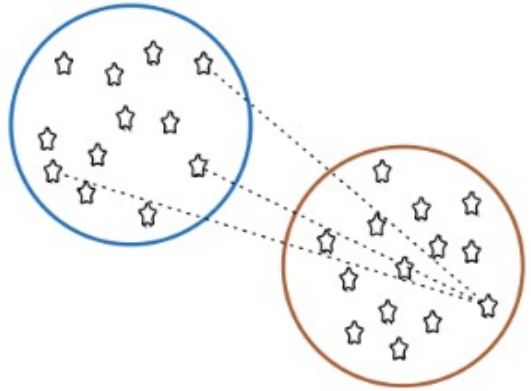
Dans le cas du **regroupement**, les données sont divisées en **groupes naturels** (ou **grappes**). Au sein de chaque groupe, les observations sont **similaires** ; d'un groupe à l'autre, elles sont **différentes**.

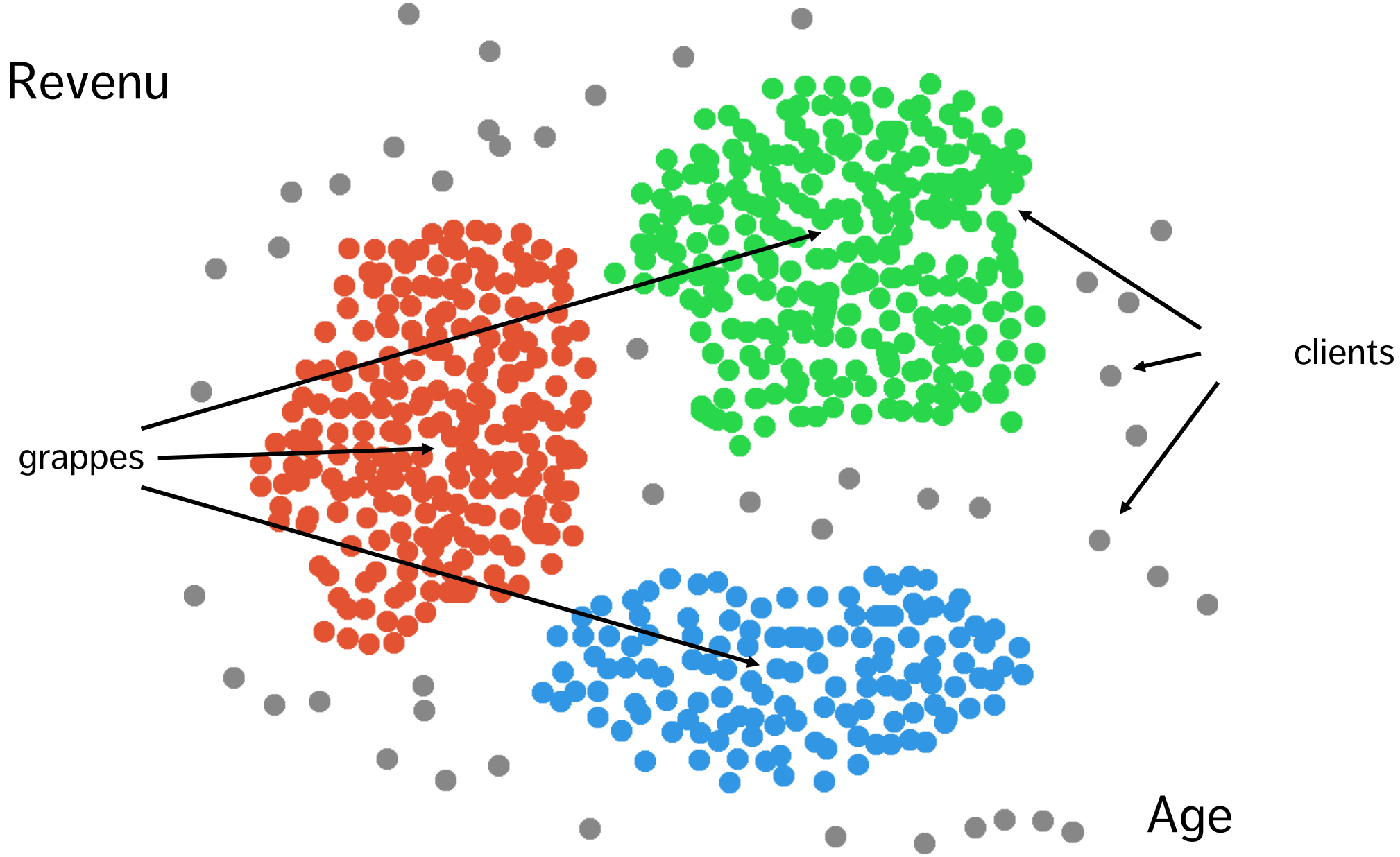
Les étiquettes et le nombre de groups ne sont pas déterminées à l'avance, c'est donc un exemple d'apprentissage **non supervisé.**

distance moyenne entre les points de son propre groupe (une **distance faible est préférable**)



distance moyenne entre les points de la grappe voisine (une **distance élevée est préférable**)




Revenu
clients
grappes
Age

Vue d'ensemble

Les algorithmes de regroupement peuvent être **complexes** et **non intuitifs**, car ils reposent sur différentes notions de similitudes entre les observations.

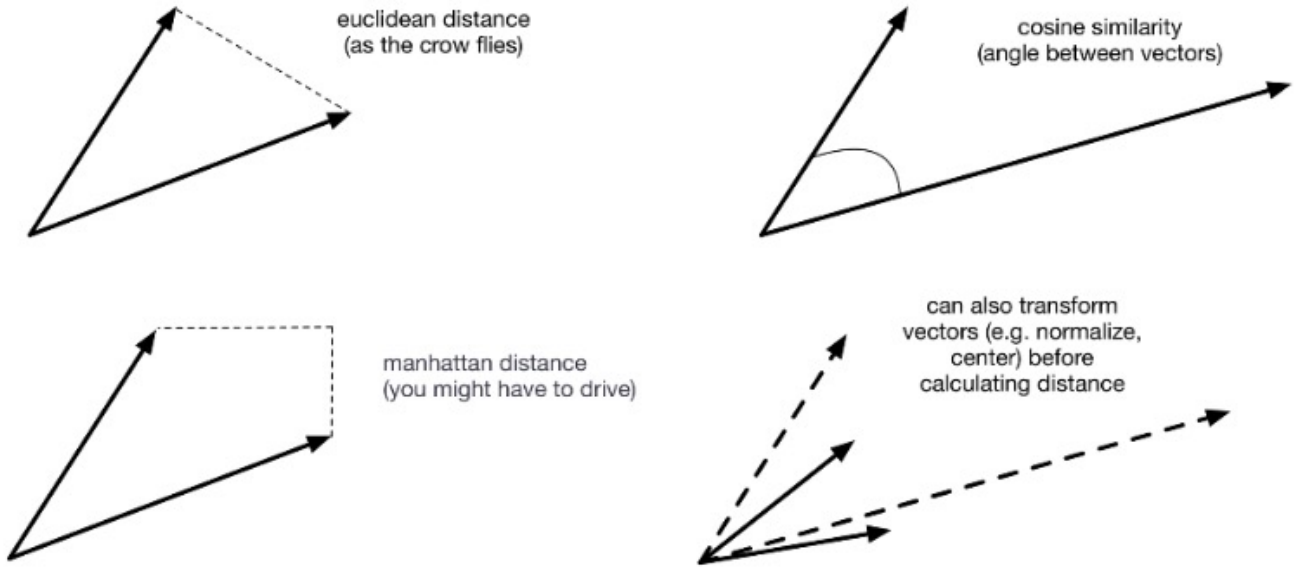
- malgré cela, la tentation d'expliquer les grappes *a posteriori* est **forte**

Ils sont également (généralement) **non déterministes** :

- le même algorithme, appliqué deux fois (ou plus) au même ensemble de données, peut découvrir des grappes complètement différentes
- l'ordre de présentation des données peut jouer un rôle
- Il en va de même pour les configurations de départ

Exigences en matière de regroupement

Une mesure de **similarité** (ou une distance) entre les observations :



IMPORTANT : les données doivent être mises à l'échelle avant d'être utilisées dans les algorithmes de regroupement

En règle générale, , $w \to 1$ correspond à $d \to 0$, et $w \to 0$ à $d \to \infty$.

Mesures de distance

Variables catégorielles*

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

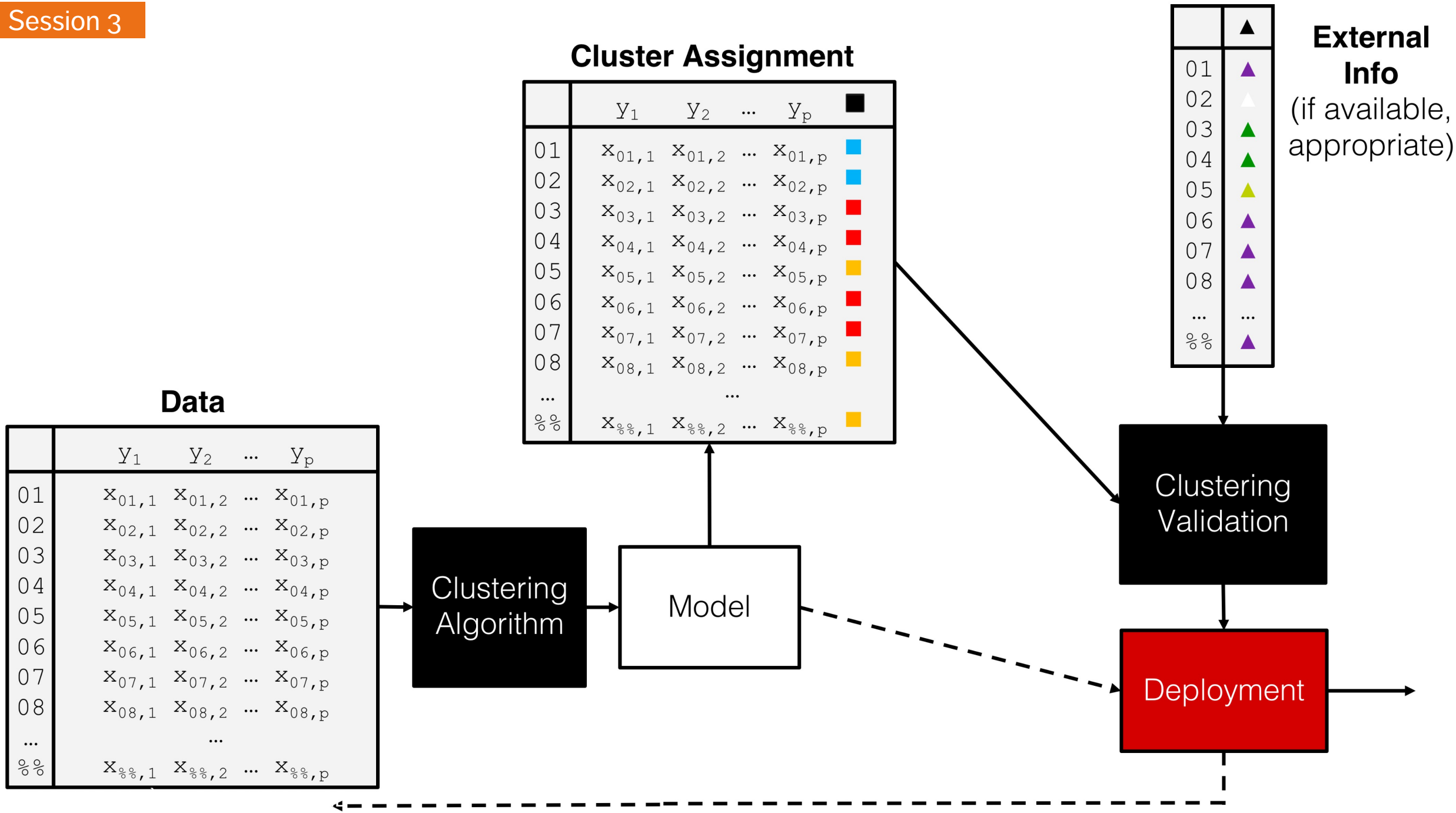
Il n'y a pas de règle fixe pour déterminer la distance à utiliser ; des systèmes concurrents sont souvent produits avec des distances différentes.

Variables numériques

- Euclidienne
- Manhattan
- corrélation
- cosinus
- etc.

Il peut être nécessaire de créer des métriques hybrides pour les ensembles de données comportant à la fois des variables catégorielles et numériques.

Session 3


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

Applications

Documents textuels

- regrouper des documents similaires en fonction de leurs thèmes, sur la base des modèles de mots communs et inhabituels

Recommandations sur les produits

- regrouper les acheteurs en ligne en fonction des produits qu'ils ont consultés, achetés, aimés, ou détestés
- regrouper les produits en fonction des avis des clients

Marketing et affaires

- regrouper les profils des clients en fonction de leurs caractéristiques démographiques et de leurs préférences

Applications

Division d'un grand groupe (ou d'une zone, ou d'une catégorie) en groupes **plus petits**, les membres des groupes plus petits ayant des similitudes d'un certain type

- les tâches peuvent ensuite être résolues séparément pour chacun des petits groupes
- cela peut conduire à une plus grande précision une fois que les résultats distincts sont agrégés

Création de taxonomies **à la volée**, au fur et à mesure que de nouveaux éléments sont ajoutés à un groupe d'éléments.

- permettant de faciliter la navigation des produits sur un site comme Netflix, par exemple

Étude de cas

"Livehoods"

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Objectif

Lorsque nous pensons à la similitude au niveau urbain, nous envisageons généralement des voisinages ou des quartiers. Existe-t-il un autre moyen d'identifier les parties similaires d'une ville ?

Les chercheurs ont pour objectif de tracer les limites des **"livehoods"**, des zones de **caractère semblable** au sein d'une ville, à l'aide de modèles de regroupement. Contrairement aux quartiers administratifs **statiques**, les "livehoods" sont définis en fonction des **habitudes** de leurs habitants.

Étude de cas

"Livehoods"

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Méthodologie

Les auteurs utilisent le **regroupement spectral** pour découvrir des **zones géographiques distinctes** de la ville sur la base des **schémas de déplacement** collectif.

Les grappes de "livehoods" s'obtiennent comme suit :

1. une **distance géographique** est calculée sur la base des paires de coordonnées de **lieux** ;
2. une **similarité sociale** est calculée entre chaque paire de **lieux à** l'aide de mesures de cosinus ;
3. le regroupement spectral produit des **candidats** de **"livehoods"** ;
4. des entretiens sont menés avec les résidents afin d'**explorer**, d'**étiqueter,** et de **valider** les grappes découvertes par l'algorithme.

Étude de cas

"Livehoods"

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Données

Les données proviennent de deux sources, combinant environ 11 millions "check-ins" issus de l'ensemble de données de Chen et al. (un site de recommandation de lieux basé sur les expériences des utilisateurs) et un nouvel ensemble de données de 7 millions de "check-ins" Twitter téléchargés entre juin et décembre 2011.

Pour chaque "check-in", les données comprennent l'**identifiant de l'utilisateur**, l'**heure**, la **latitude et la longitude**, le **nom du lieu** et sa **catégorie**.

Dans cette étude de cas, les données de la ville de Pittsburgh, en Pennsylvanie, sont examinées *via* 42787 "check-ins" de 3840 utilisateurs dans 5349 lieux.

Étude de cas

"Livehoods"

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Points forts et limites de l'approche

- La technique utilisée dans cette étude est **agnostique** par rapport à la source particulière des données : elle ne dépend pas d'une méta-connaissance des données.
- L'algorithme peut être sujet à des préjugés "majoritaires", ce qui peut entraîner une représentation erronée ou une dissimulation des comportements minoritaires.
- L'ensemble de données est construit à partir d'un échantillon **limité** de "check-ins" partagés sur Twitter et sont donc biaisés vers les types de visites/lieux que les gens veulent généralement partager **publiquement.**
- L'ajustement des grappes n'est pas trivial : le biais de l'expérimentateur peut se combiner au "biais de confirmation" des personnes interrogées lors de l'étape de validation – si les chercheurs sont des résidents de Pittsburgh, verront-ils des grappes alors qu'il n'y en avait pas ?

Étude de cas

"Livehoods"

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

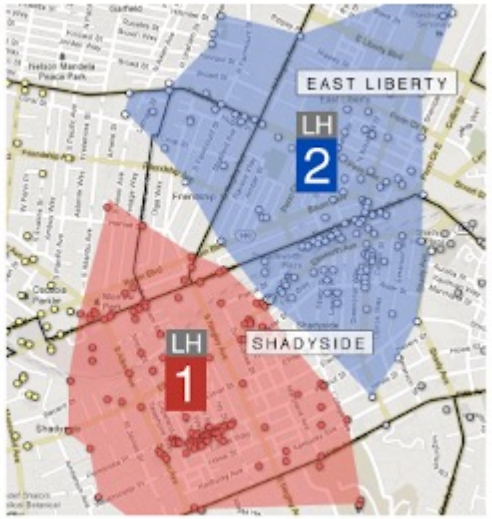
Résultats, évaluation, et validation

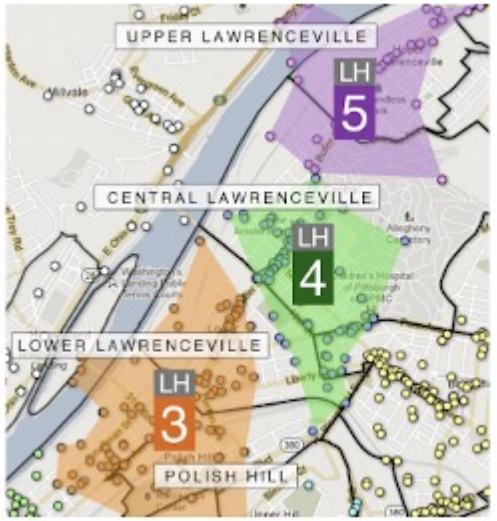
Dans trois quartiers de la ville, neuf modes de vie ont été identifiés et validés par 27 habitants de Pittsburgh.

- **Frontières :** les "livehoods" sont dynamiques et évoluent en fonction des comportements des gens, contrairement aux quartiers fixes établis par le gouvernement de la ville.
- **Démographie :** les entretiens ont montré que la démographie des résidents et des visiteurs d'une région joue un rôle important dans l'explication des "livehoods".
- **Développement et ressources :** le développement économique peut affecter le caractère d'une zone. De même, les ressources fournies par une région ont une forte influence sur les personnes qui la visitent, et donc sur le caractère qui en résulte.

Étude de cas

"Livehoods"

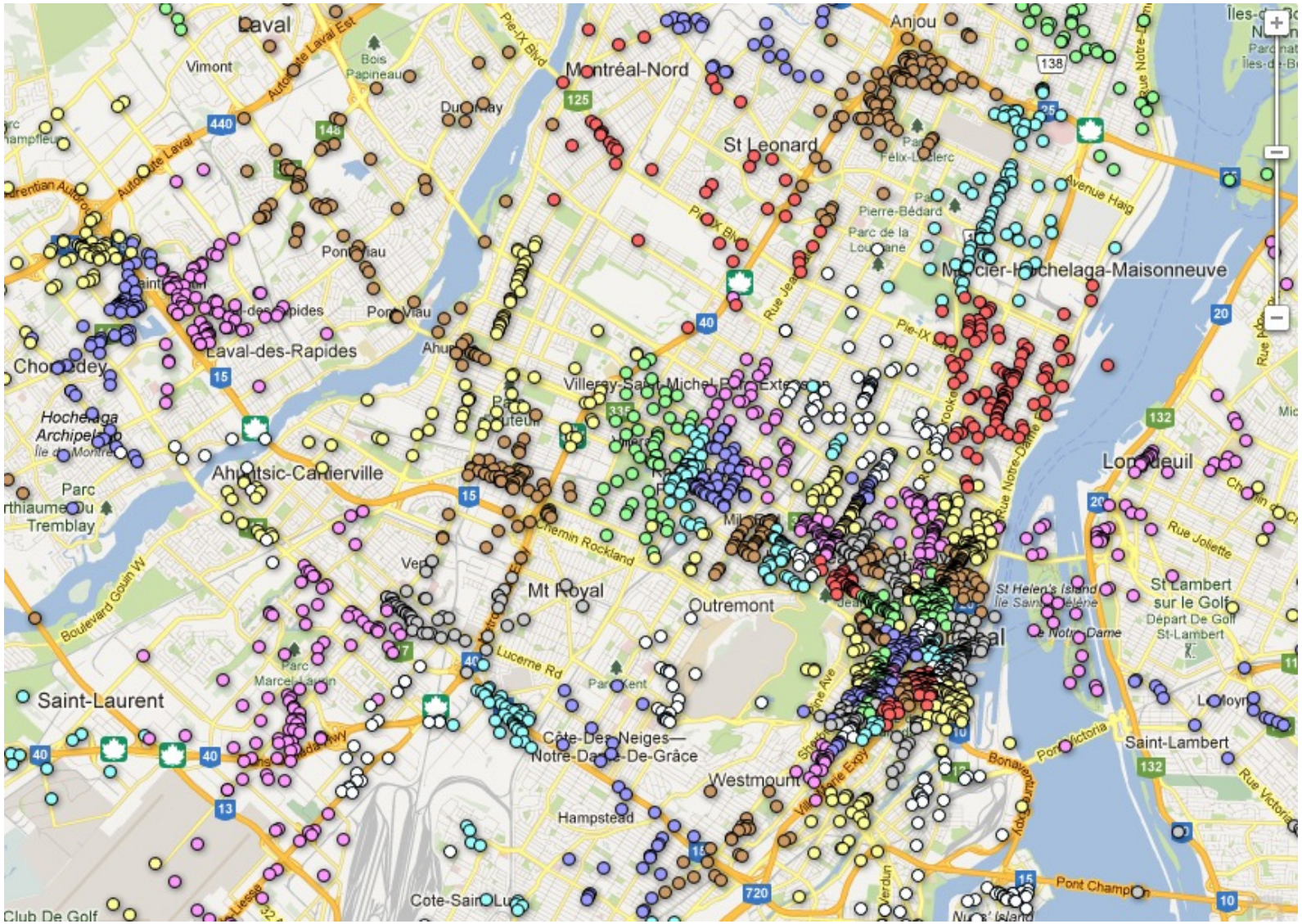




Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Étude de cas

"Livehoods"



Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Remarques générales

Le regroupement est un concept relativement **intuitif** pour les êtres humains, car notre cerveau le fait inconsciemment :

- reconnaissance faciale
- la recherche de modèles, etc.

En général, les gens sont très doués pour les données **désordonnées**, mais les ordinateurs et les algorithmes ont plus de mal.

La difficulté réside en partie dans le fait qu'il **n'existe pas de définition commune de ce qui constitue un cluster :**

- "Je ne suis peut-être pas en mesure de définir ce que c'est, mais j'en reconnais un quand j'en vois un".

Lectures conseillées

Vue d'ensemble du regroupement

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.5 Clustering
    - Overview
    - Case Study: Livehoods

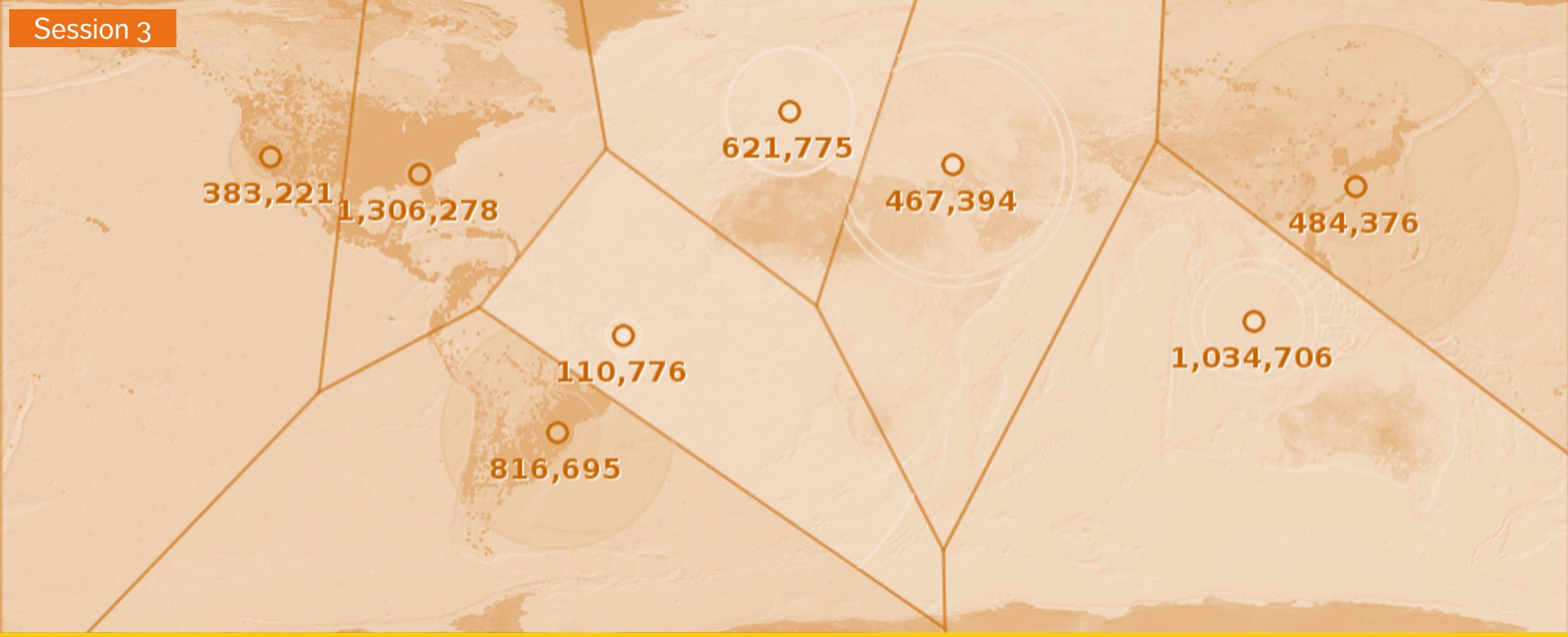
22. Focus on Clustering
    22.1 Overview

Exercices

Vue d'ensemble du regroupement

1. Qu'est-ce que la non-reproductibilité (potentielle) du regroupement implique pour la validation ? Pour le “buy-in” du client et/ou des parties prenantes ?

2. Identifier des scénarios et des questions qui pourraient faire appel au regroupement dans vos activités professionnelles quotidiennes.



8. k-moyennes et autres algorithmes

Algorithmes de regroupement

k-moyennes

- modèle classique (et surutilisé)
- les hypothèses faites sur la forme des grappes

Regroupement hiérarchique

- facile à interpréter, déterministe

Ensembles de grappes

Allocation latente de Dirichlet

- utilisé pour la modélisation des sujets

Maximisation de l'espérance

Algorithmes de regroupement

Réduction itérative équilibrée et regroupement à l'aide de hiérarchies (BIRCH)

Regroupement spatial basé sur la densité des applications avec bruit (DBSCAN)

- basé sur un graphique

Propagation de l'affinité

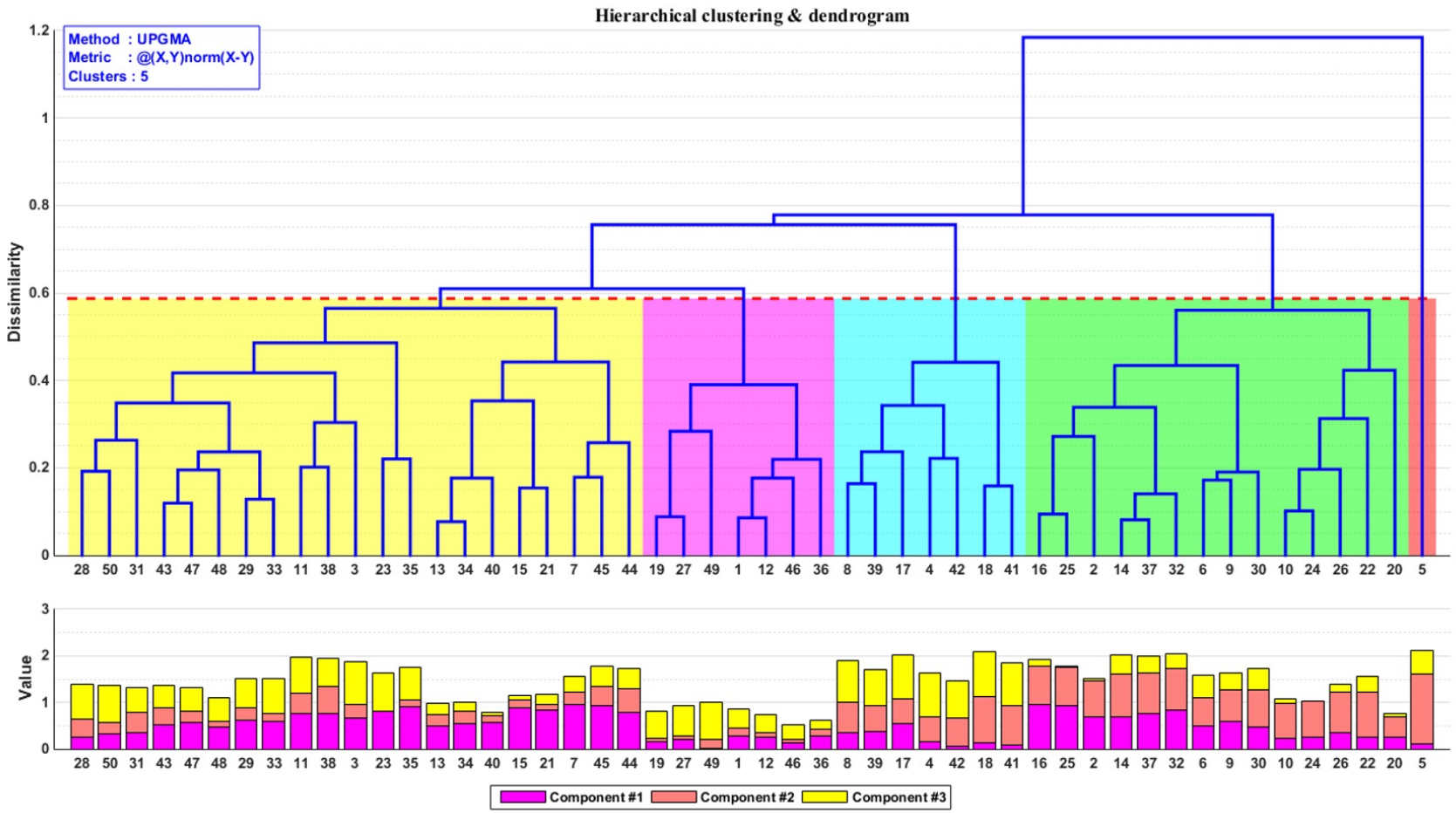
- sélectionne automatiquement le nombre optimal de grappes

Regroupement spectral

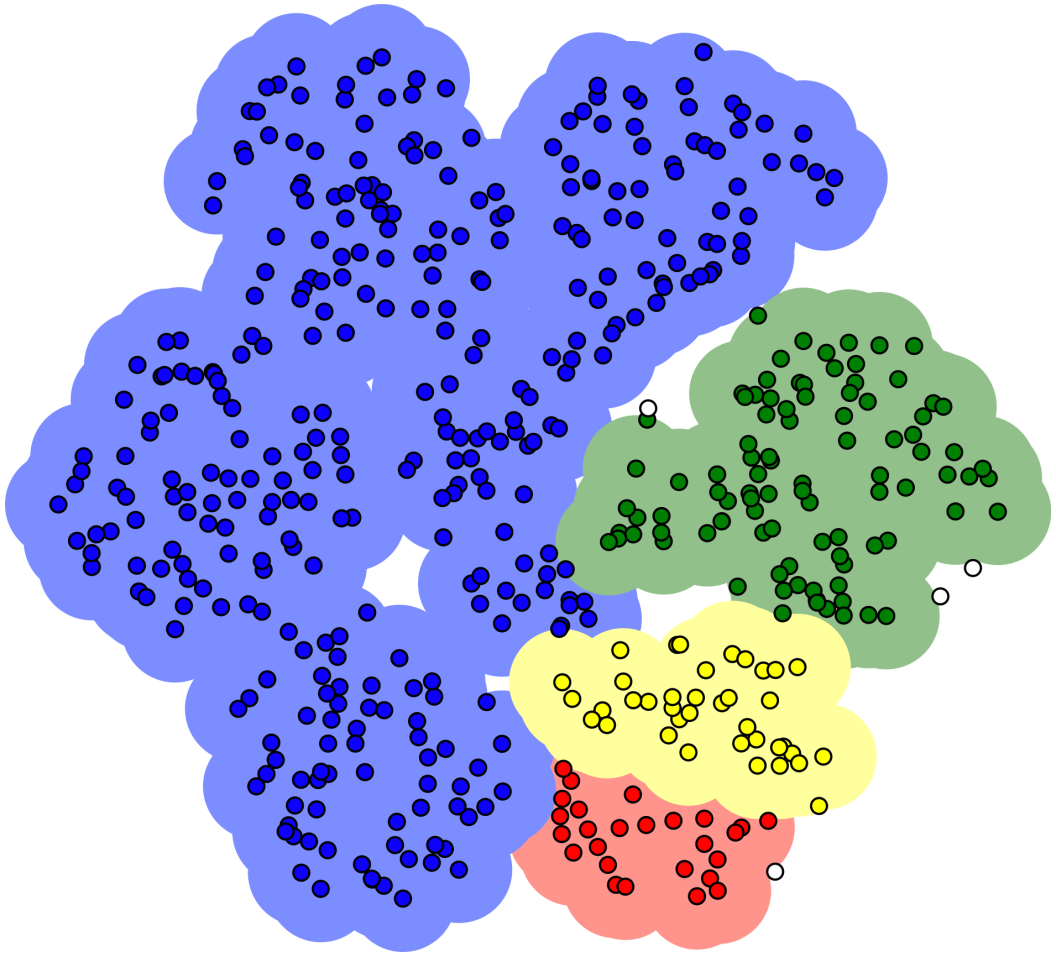
- reconnaît les clusters non-blob

Regroupement flou (Fuzzy)

Regroupement hiérarchique

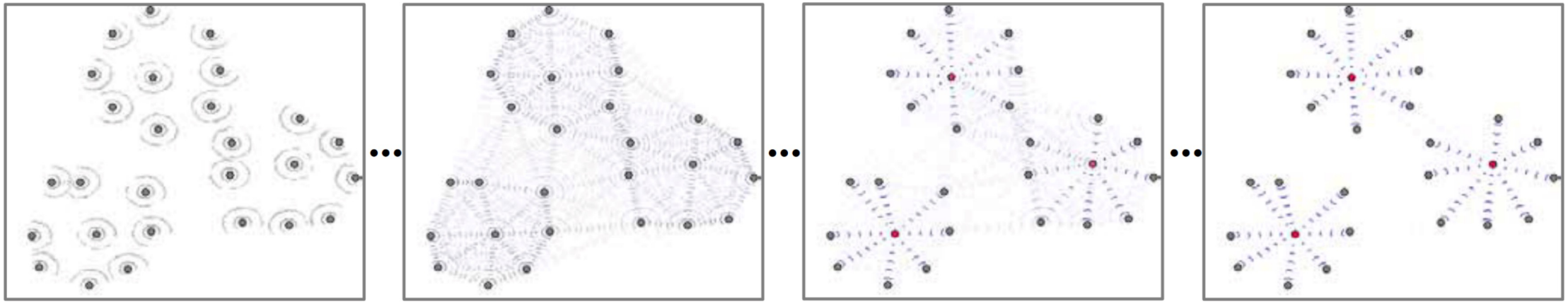


DBSCAN



[D. Dueck]

Propagation de l'affinité



initialization

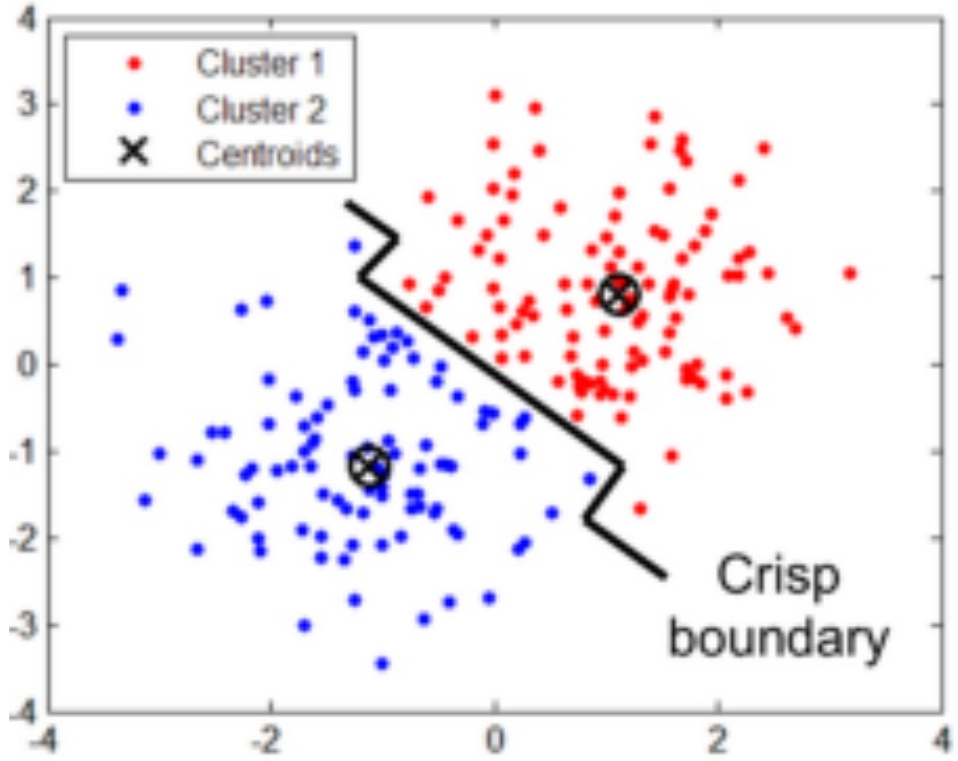
iteration X
(uncertainty)

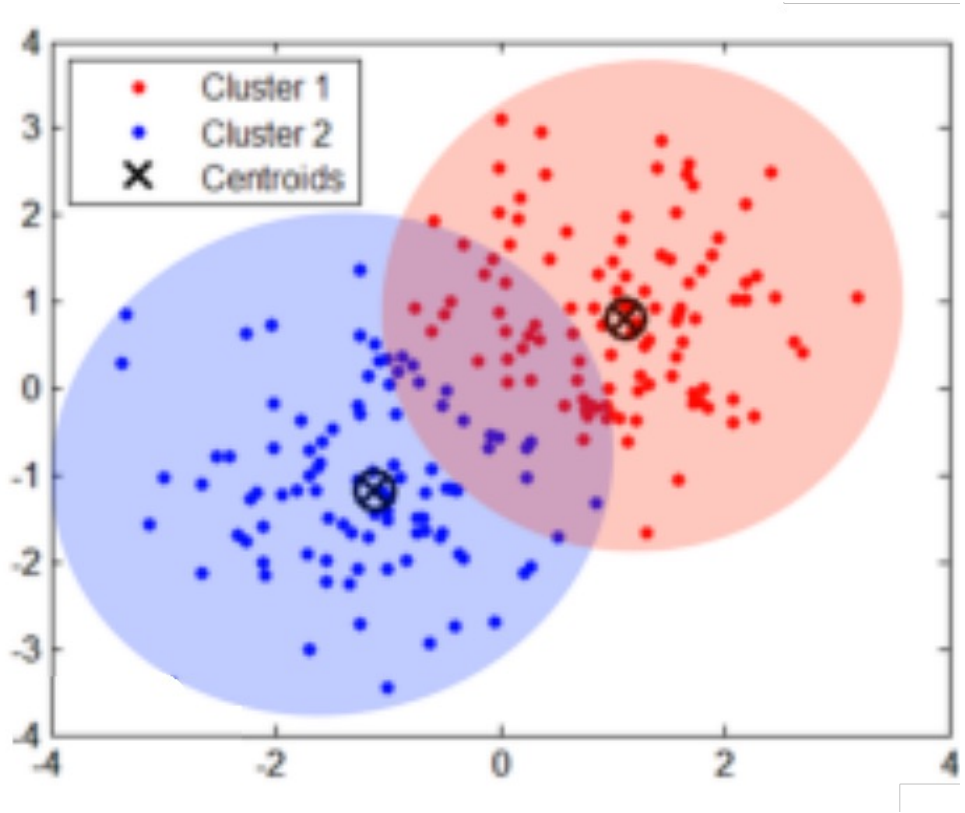
iteration Y
(emergence)

final iteration
(convergence)

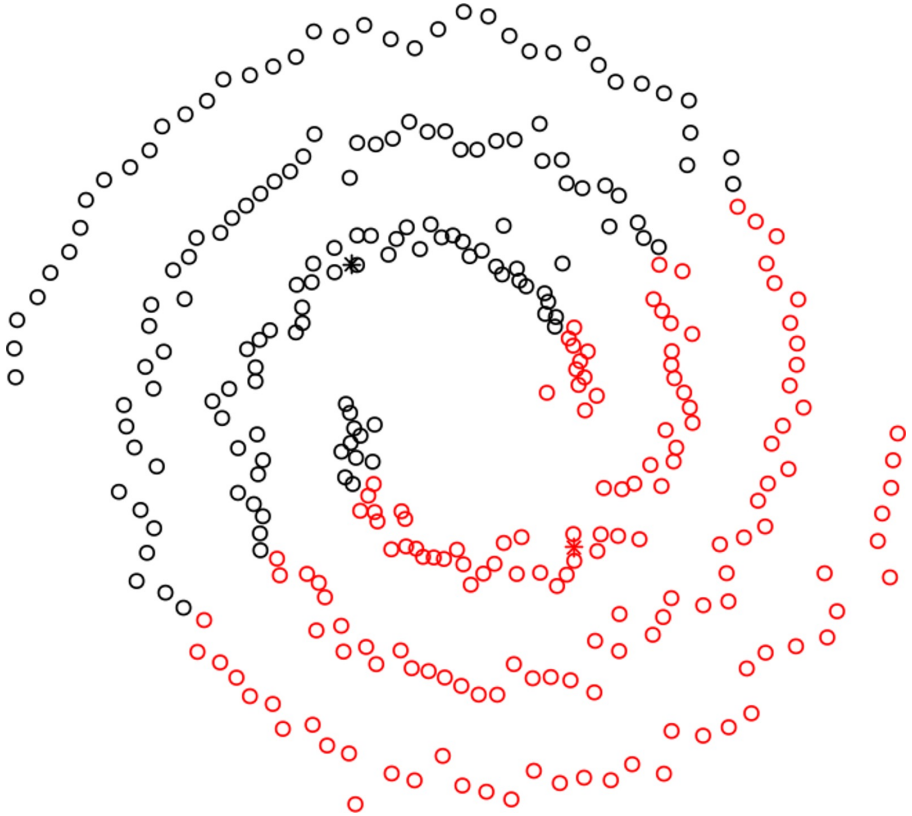
[A. Bagnaro, F. Baltar et G. Brownstein]

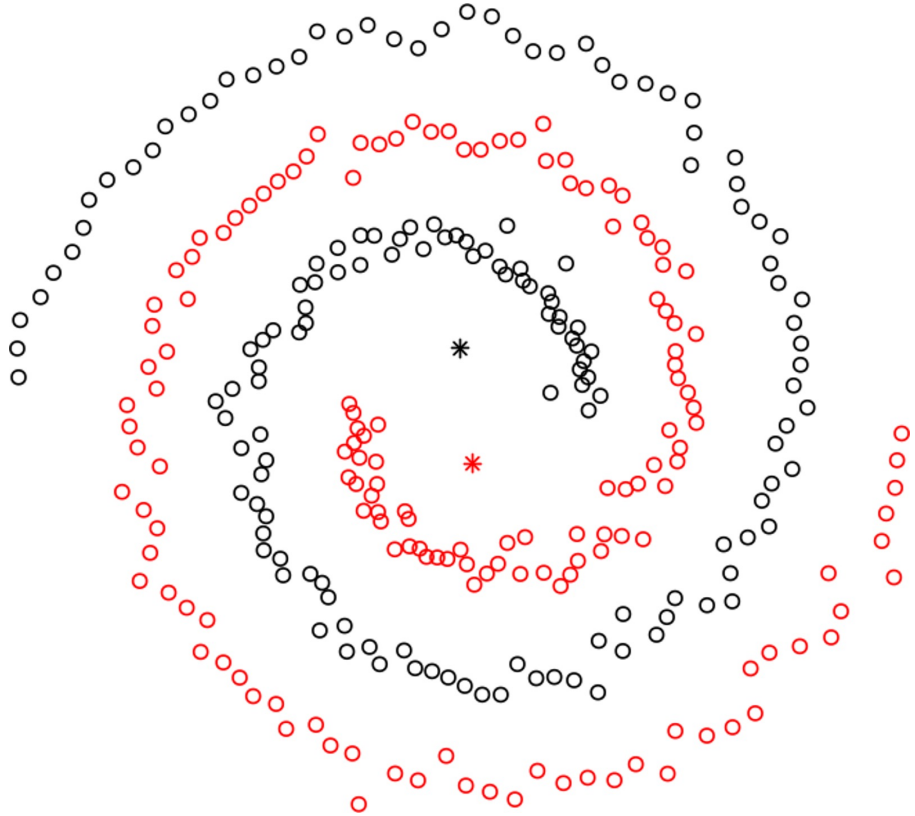
k –moyennes floues





Regroupement spectral

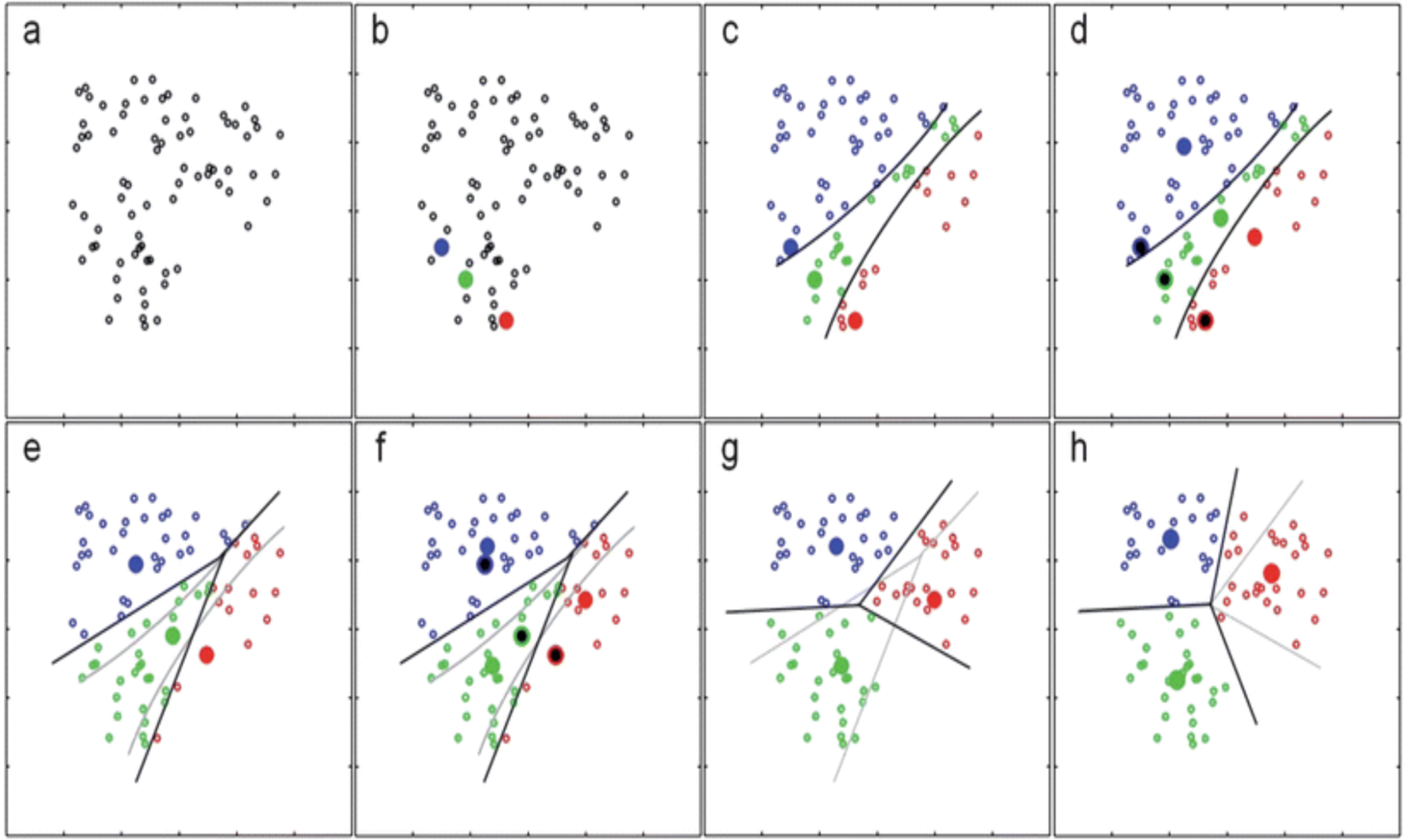




Algorithme des k-moyennes

1. Sélectionnez le **nombre de grappes** souhaité : k
2. Choisir de manière aléatoire k instances comme **centres de grappe**
3. Calculer la **distance** entre chaque observation et chaque centre
4. Placer chaque instance dans la grappe dont le centre est le **plus** proche
5. Calculer le **centroïde** pour chaque grappe
6. Répéter les étapes 3 à 5 avec les nouveaux centroïdes
7. Répéter l'étape 6 jusqu'à ce que les grappes soient stables

[A.H. Du]



Points forts

Facile à mettre en œuvre (sans avoir à calculer les distances paire par paire).

- utilisation extrêmement fréquente en conséquence
- élégant et simple

Dans de nombreux contextes, les k-moyennes sont une façon **naturelle** d'envisager le regroupement d'observations.

Permet d'acquérir une **compréhension de base de la structure des données** lors d'une première passe.

Limites

Les points de données ne peuvent être affectés qu'à **une seule** grappe

- cela peut conduire au sur-ajustement
- solution robuste : considérer la probabilité d'appartenance à chaque groupe

Les grappes sous-jacentes sont supposées avoir la **forme d'un blob**

- k-moyennes ne produiront pas de grappes utiles si cette hypothèse n'est pas respectée dans la pratique.

Les grappes sont supposées être distinctes (discrètes)

- k-moyennes ne permettent pas les **chevauchements** ou regroupements **hiérarchiques**

Limites

Il existe de nombreuses façons de choisir le **nombre optimal de** grappes .

L'algorithme est **stochastique :** différentes configurations initiales peuvent donner des **résultats différents**, ce qui peut donner un nombre optimal différent.

Les résultats peuvent également dépendre de la **taille** des données, du choix de la mesure de **distance**, du choix de la **métrique de qualité**, etc.

Lectures conseillées

k-moyennes et autres algorithmes

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.5 Clustering
    - Clustering Algorithms
    - k-Means
    - Toy Example: Iris Dataset

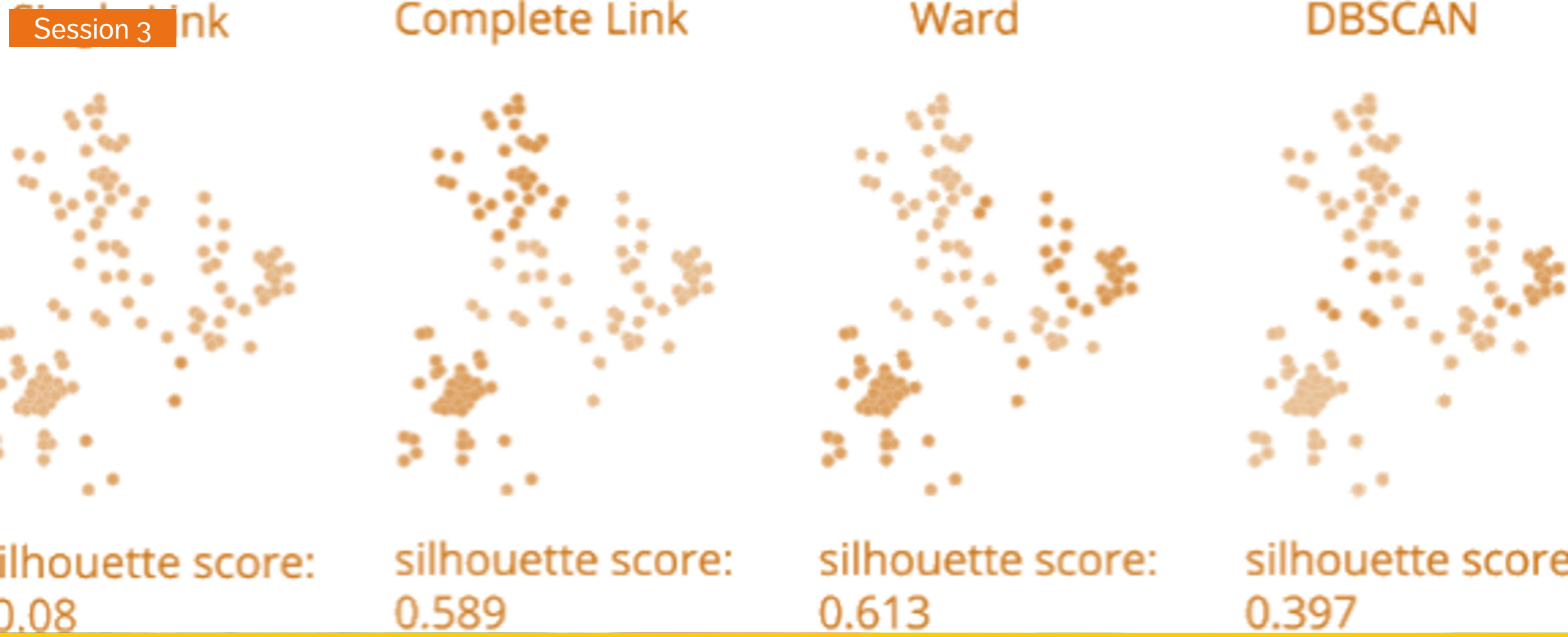
    19.7 R Examples
    - Clustering: Iris Dataset

22. Focus on Clustering
    22.2 Simple Clustering Methods
    22.4 Advanced Clustering Approaches

Exercices

k-moyennes et autres algorithmes

1. Passez en revue l'exemple de regroupement d'iris trouvé dans DUDADS (voir lecture conseillée). Répétez le processus avec l'ensemble de données UniversalBank (vous pouvez d'abord visualiser l'ensemble de données) afin de construire un schéma de regroupement. Déterminez le nombre optimal de grappes à l'aide de l'indice de Davies-Bouldin.



9. Validation et commentaires

Validation de regroupement

Qu'est-ce qui fait qu'un schéma de regroupement est **meilleur** qu'un autre ?

Qu’entend-on par un schéma de regroupement valide ?

Qu'est-ce que cela signifie pour une grappe spécifique d'être **bonne** ?

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

Bon vs. mauvais n'a pas de sens : il s'agit plutôt d’**optimal** vs **sous-optimal**.

Validation de regroupement

Schéma de regroupement **optimal :**

- séparation maximale entre les grappes
- similarité maximale au sein des groupes
- est en accord avec le test de l'œil humain
- utile pour atteindre ses objectifs

Types de validation

- **externe** (utilise des informations supplémentaires)
- **interne** (utilise uniquement les résultats du regroupement)
- **relative** (comparaison entre les tentatives de regroupement)

Validation

Le regroupement implique 2 activités :

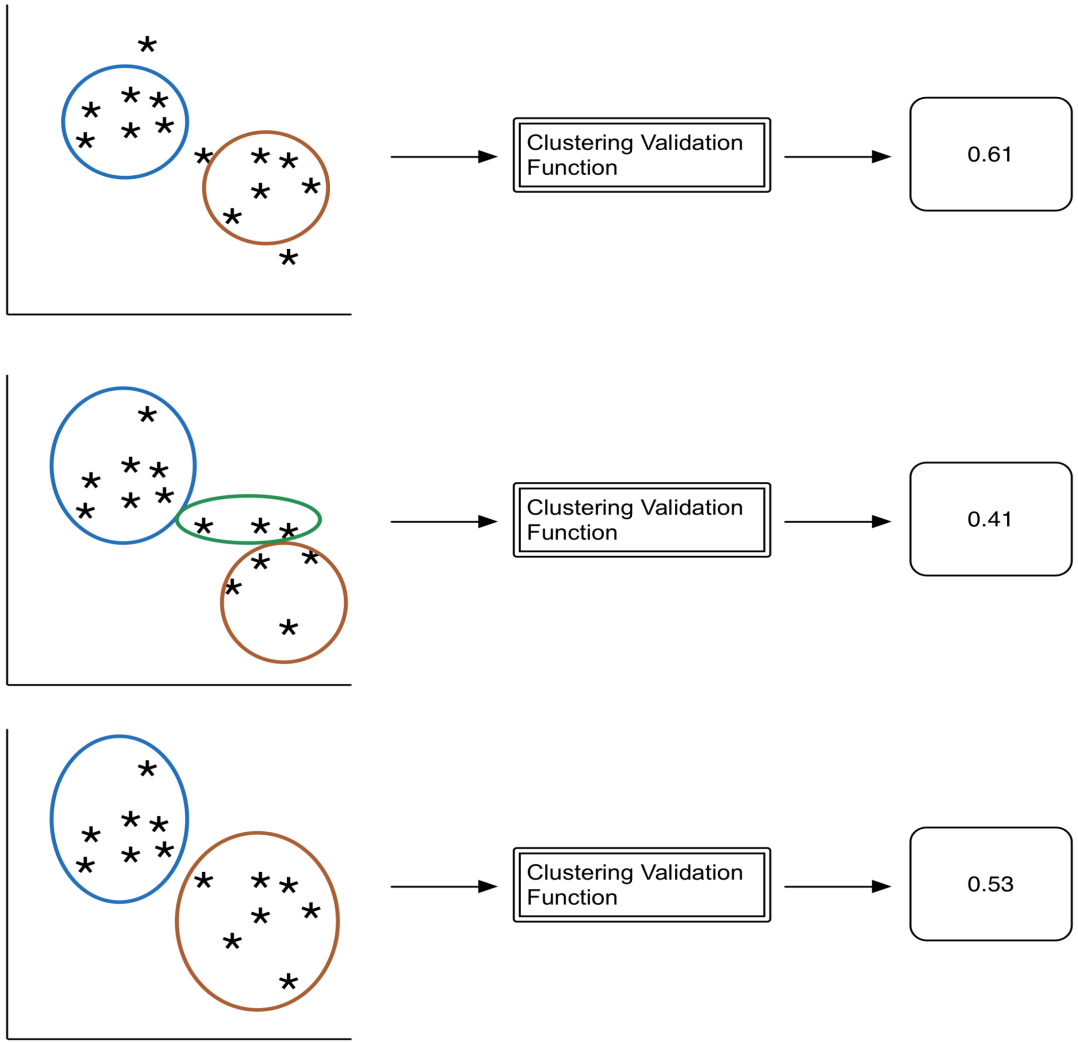
- la création de grappes
- l'**évaluation de la qualité des grappes**

Fonctions de regroupement

- entrée : instances (vecteurs)
- résultat : affectation à une grappe

Évaluer la qualité des grappes

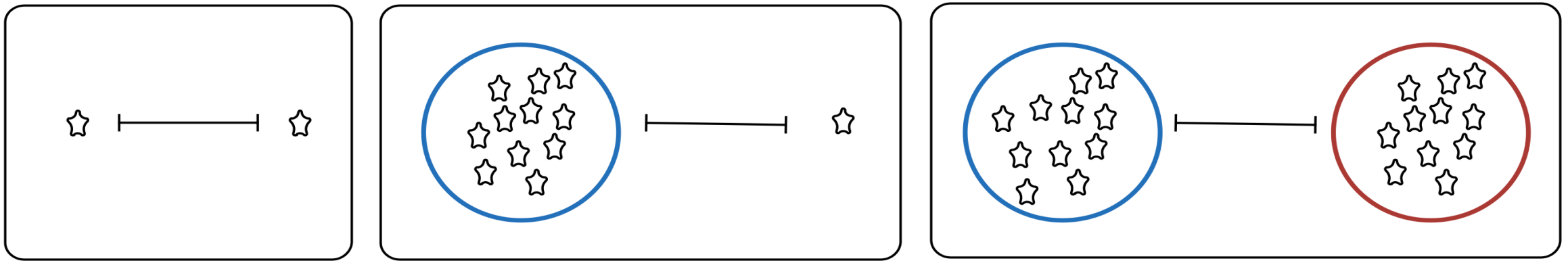
- données d'entrée : instances + affectations (+ matrice de similarité, en général)
- résultat : une valeur numérique



Composantes de la fonction

Il y a plusieurs fonctions de regroupement et de validation des grappes, mais elles sont toutes construites à partir de mesures de base:

- **propriétés de l'instance**
- **propriétés de la grappe**
- **propriétés de la relation instance - instance**
- **propriétés de la relation grappe – instance**
- **propriétés de la relation grappe – grappe**



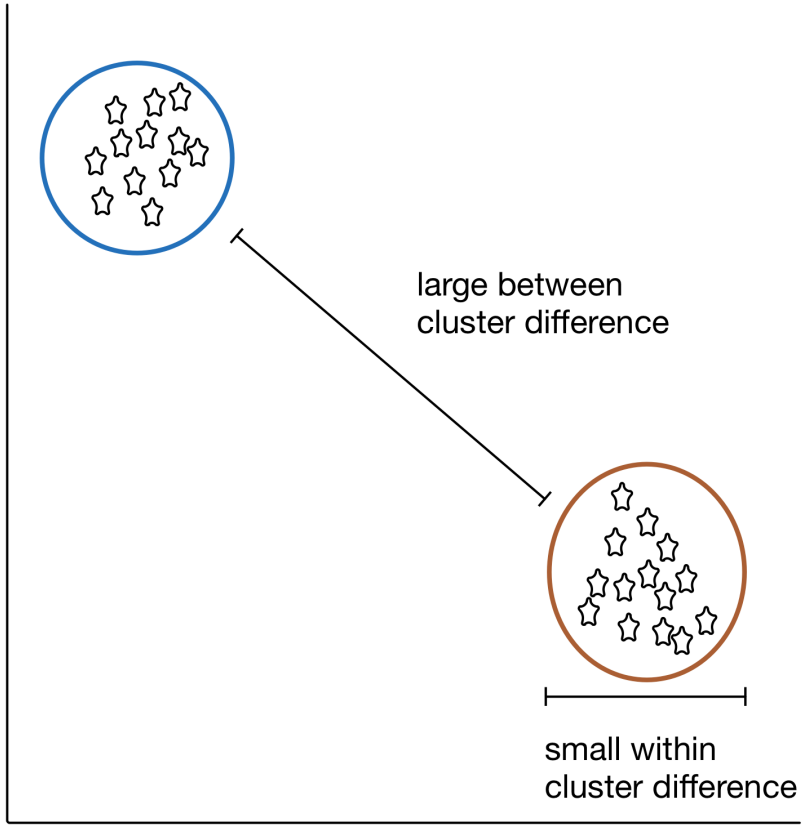
Objectifs de la validation interne

Au sein des grappes, tout est très similaire. Entre les grappes, il y a beaucoup de différences.

Le **problème :** les grappes peuvent s'écarter de cet idéal de bien des façons.

Comment pondérer les bons aspects (par exemple, une forte **similarité à l'intérieur d'une grappe**) par rapport aux mauvais (par exemple, une faible **séparation entre les grappes**) ?

En conséquence, il y a un très grand nombre de **mesures de la qualité des grappes** (CQM).



CQM de validation interne

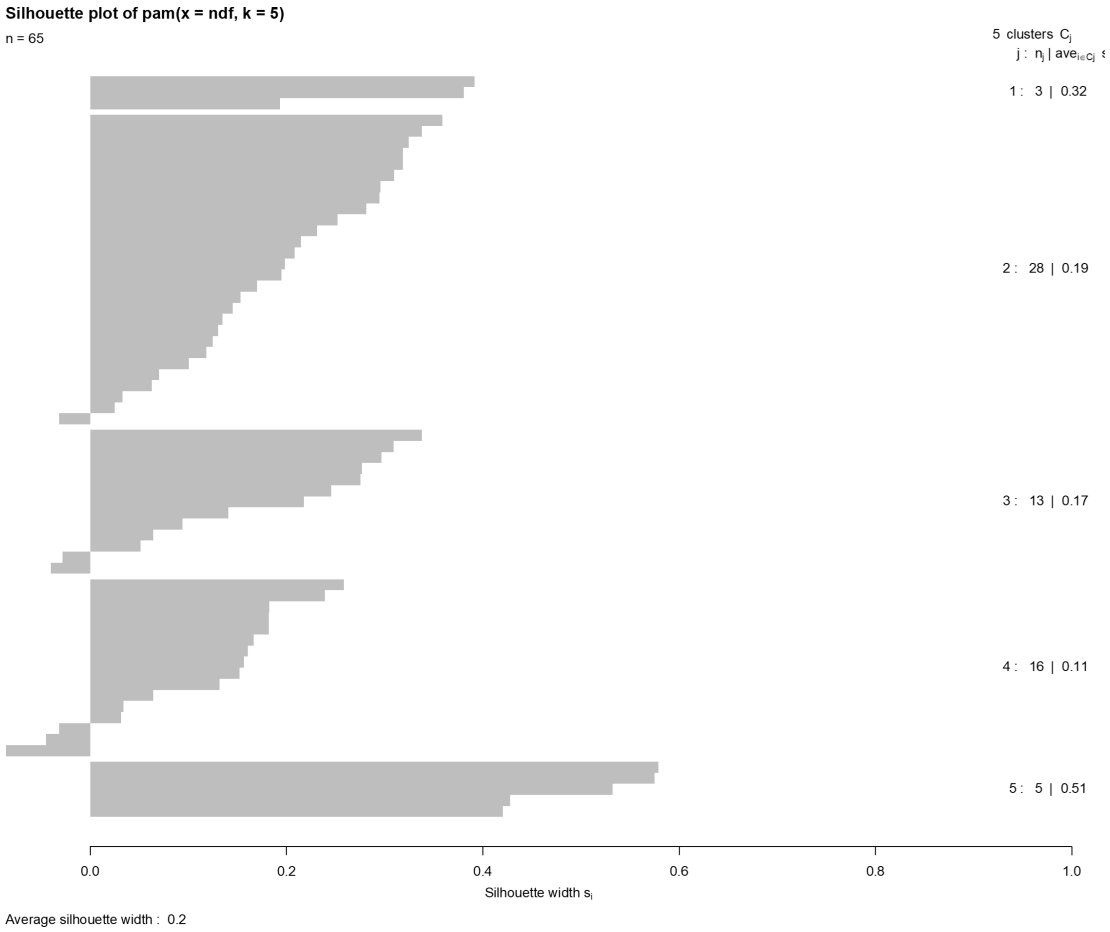
Indice Davies-Bouldin

Indice de Dunn

Silhouette

Somme des carrés à même la grappe

etc. (il y en a des tonnes !)

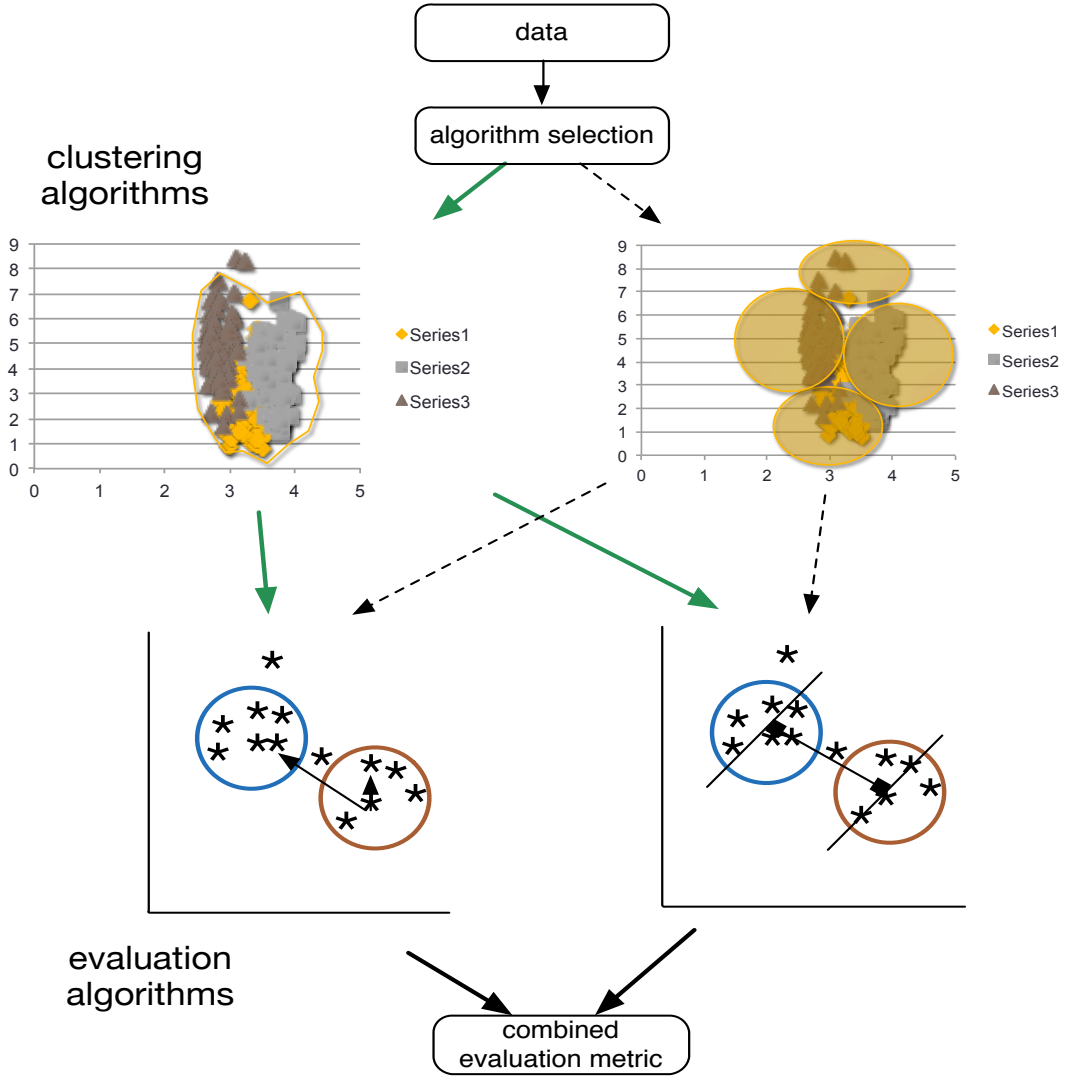


Validation relative

Obtenir une seule mesure de validation pour un seul regroupement n'est pas très utile – les résultats pourraient-ils être meilleurs ? Est-ce le mieux que nous puissions espérer ?

Nous pourrions **comparer les résultats** d'une série à l'autre ou d'un paramètre à l'autre.

La principale difficulté consiste à déterminer comment comparer les résultats des **différentes séries**.



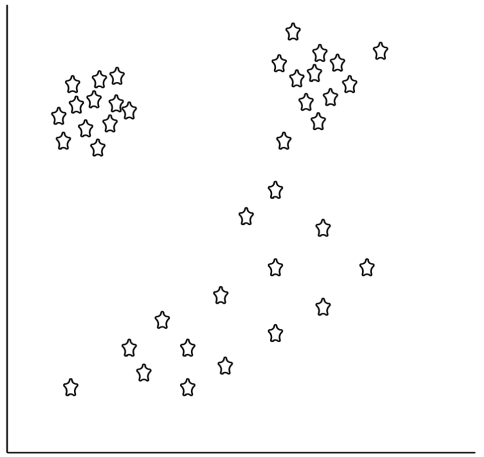
Ensembles

Quelques options :

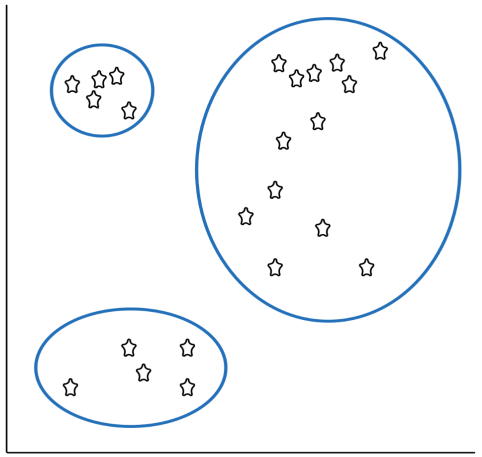
- échantillons multiples provenant de la même source
- différents sous-ensembles de colonnes sont utilisés
- différents algorithmes sont utilisés

La **similarité** des résultats du regroupement est mesurée.

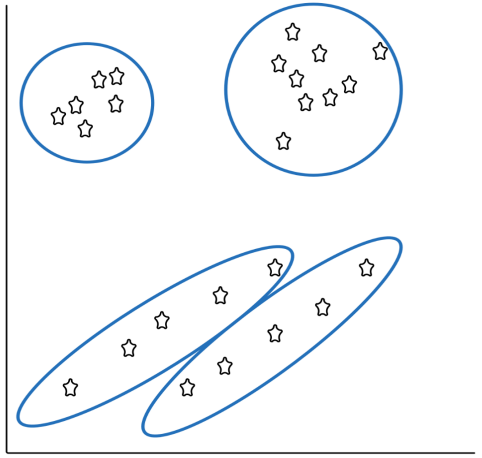
Si les résultats **ne sont pas stables** d'une approche à l'autre, des recherches plus approfondies sont nécessaires.



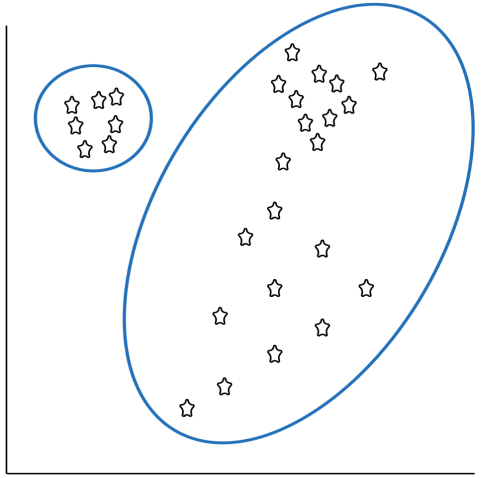
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

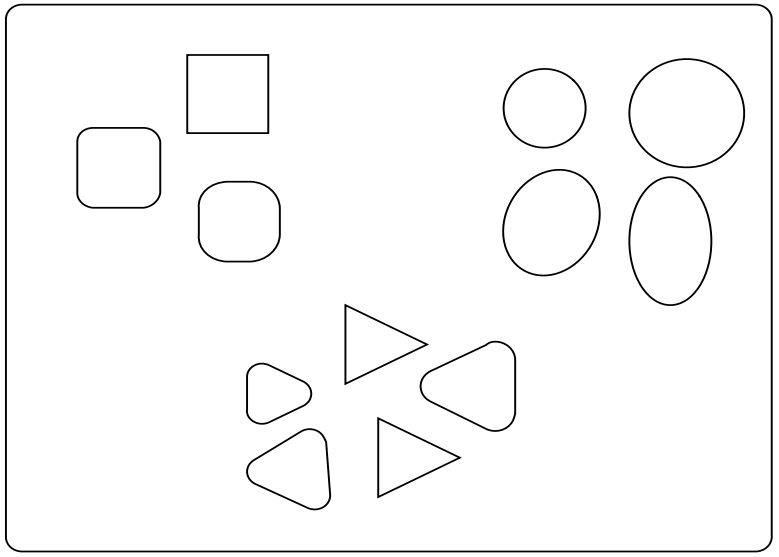
Validation externe

Nous pouvons aussi apporter de l'information de l'extérieur afin d'**évaluer** les grappes.

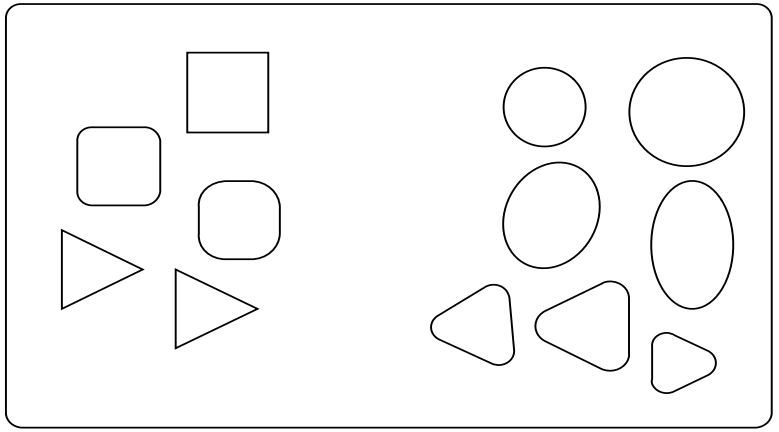
Ces informations extérieures constituent une classification "correcte".

En quoi cela diffère-t-il de la classification ?

Souvent utilisé pour renforcer la confiance dans l'approche globale, sur la base de résultats préliminaires ou d'échantillons.



Natural Groupings



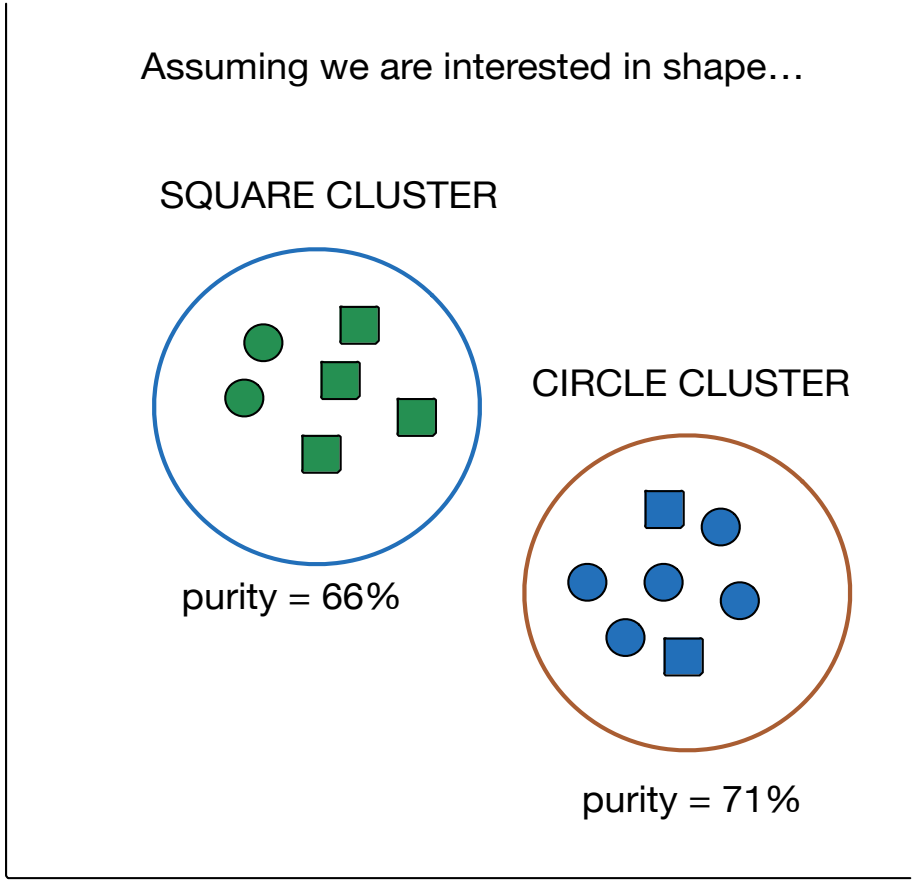
Clustering Results

La pureté

Pour cette mesure de validation externe, chaque grappe est attribuée à la classe la **plus fréquente** dans la grappe.

Nous calculons la **pureté** comme suit : nombre de points correctement attribués divisé par le nombre de points dans la grappe.

Autres options : **précision**, **rappel**.



Défis en matière de regroupement

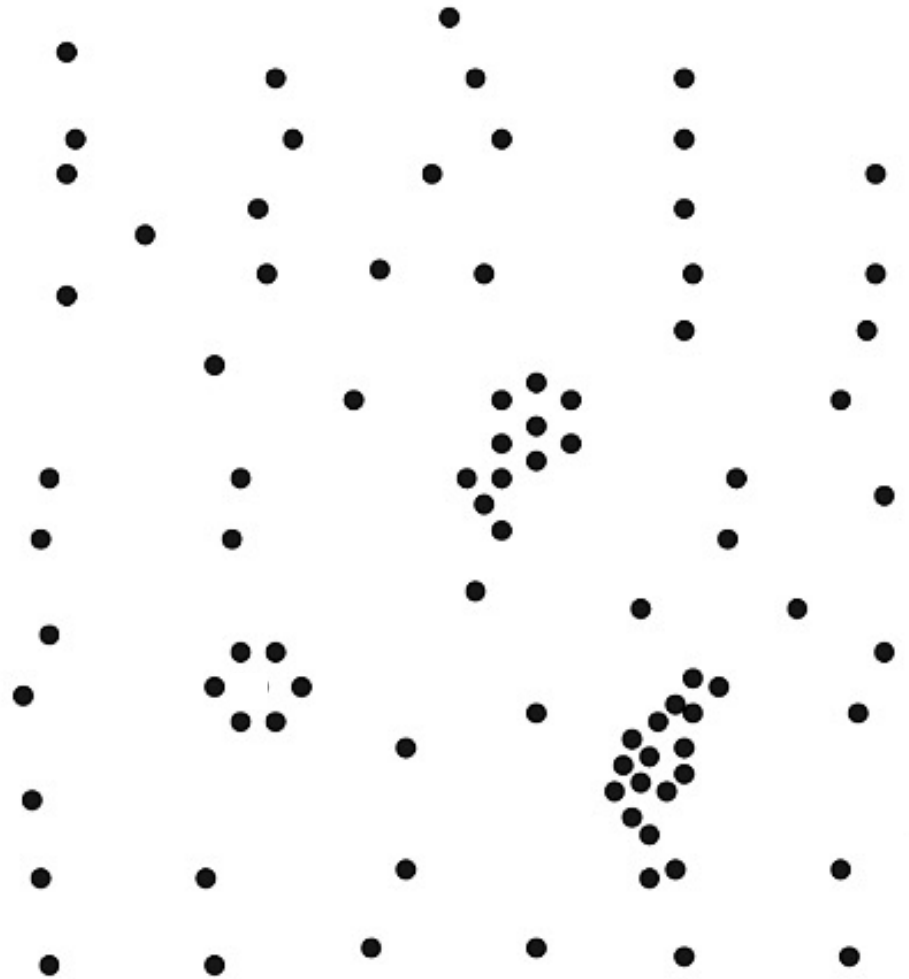
Automatisation
relativement intuitive pour les humains, mais plus difficile pour les machines

Absence de définition claire
pas d'accord universel sur ce qui constitue une grappe

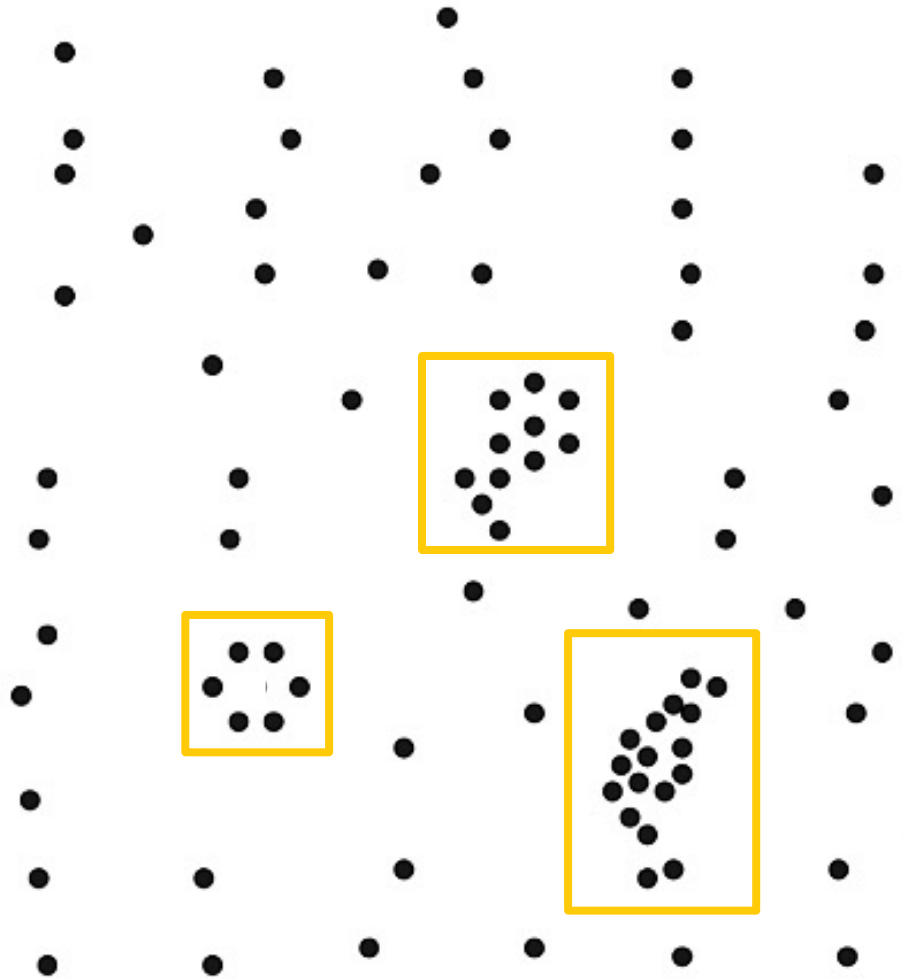
Non-répétabilité
non déterministe : le même algorithme, appliqué deux fois au même ensemble de données, peut découvrir des grappes complètement différentes

Nombre de grappes
le nombre optimal de grappes est difficile à déterminer

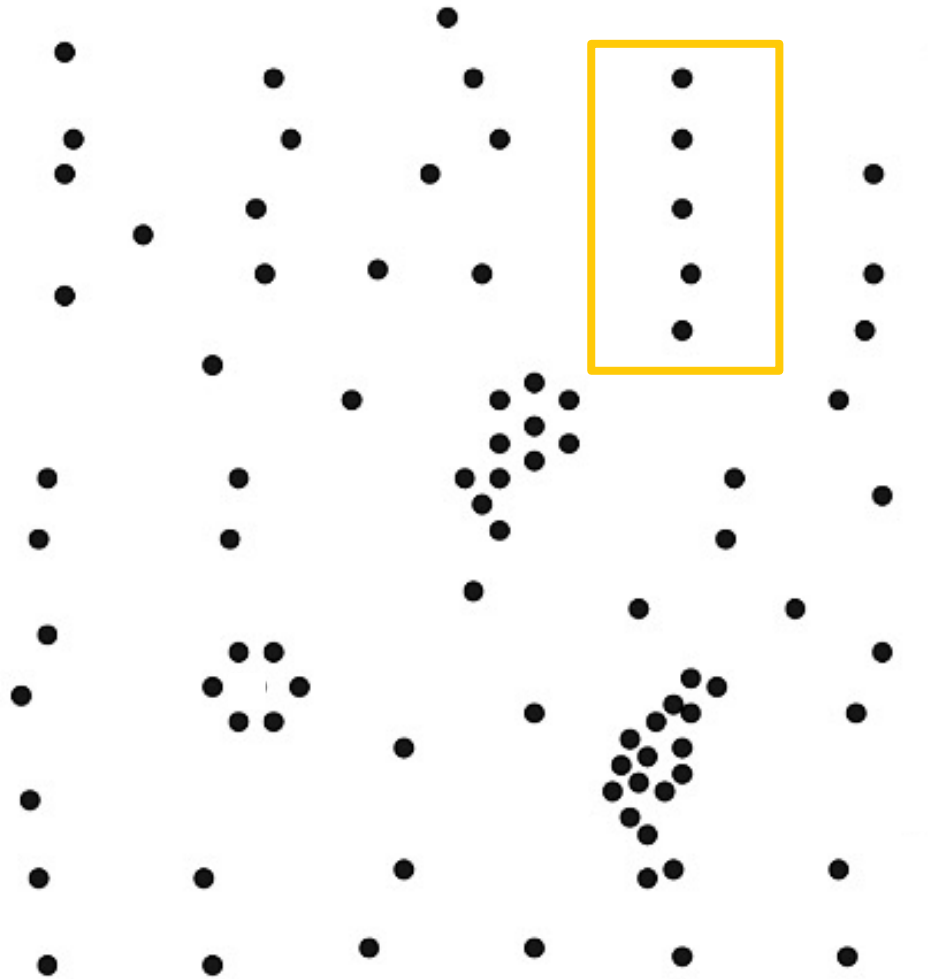
Session 3



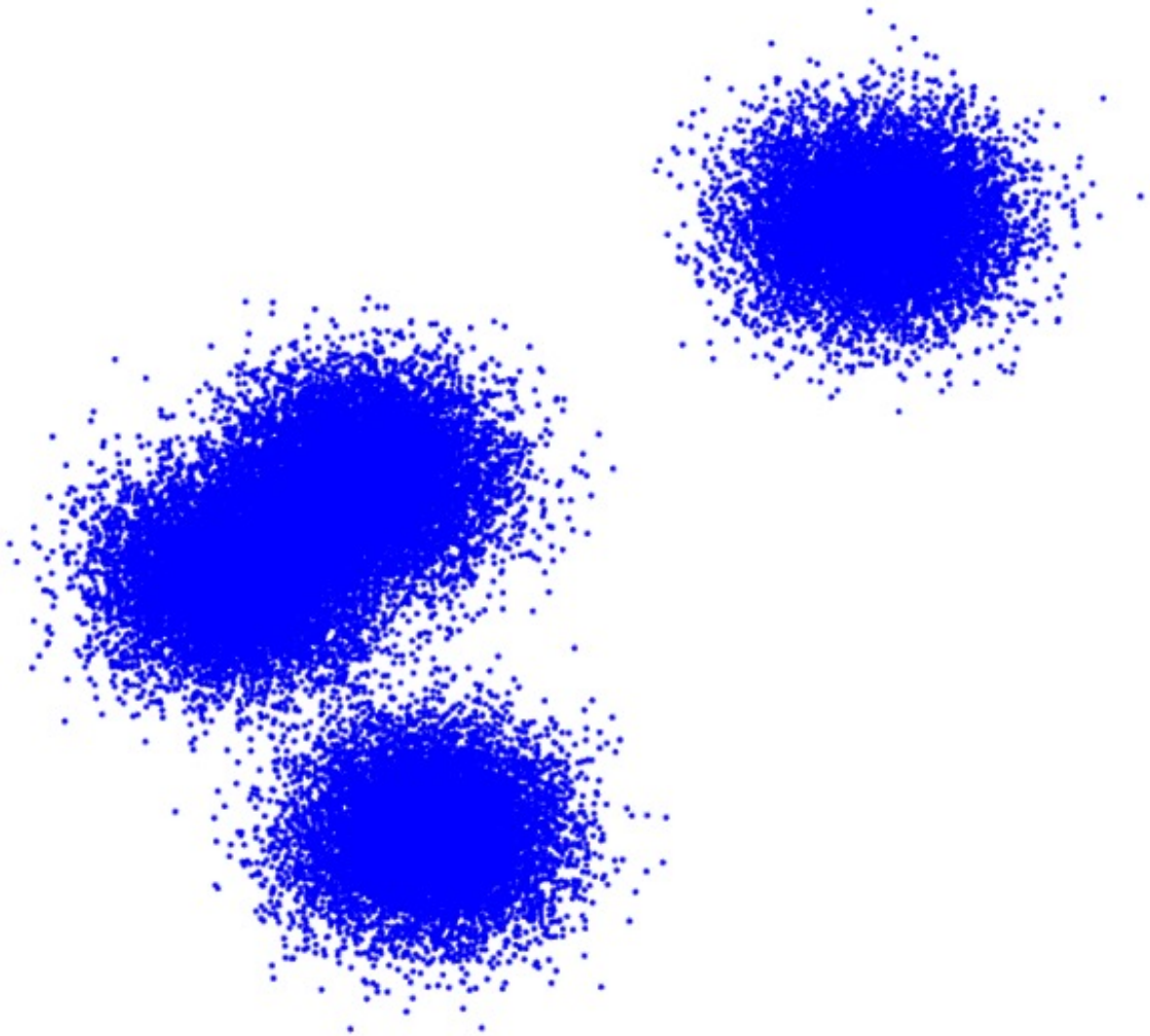
Session 3



Session 3



Session 3



Défis en matière de regroupement

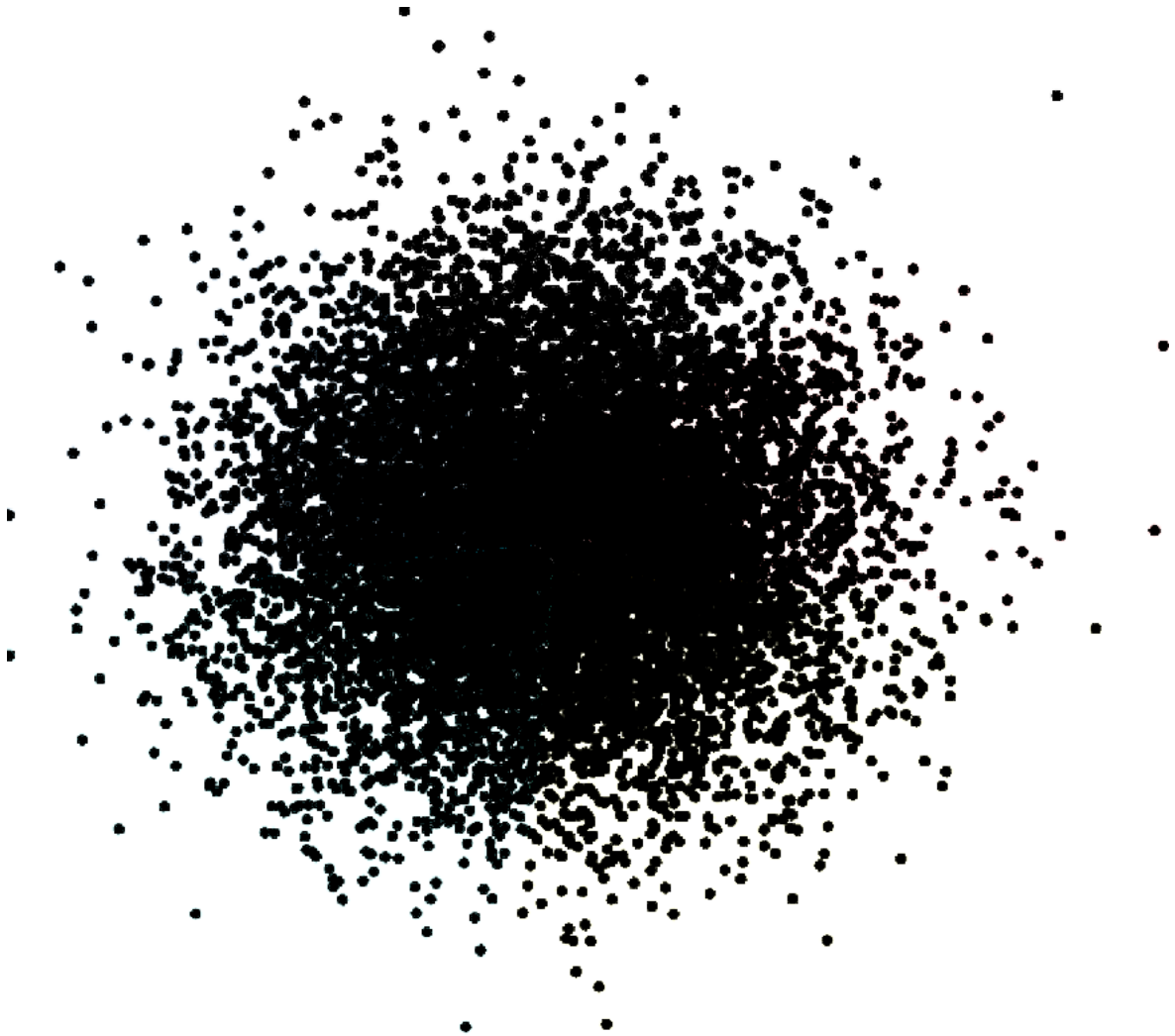
Description des grappes
Les décrit-on à l'aide d'instances représentatives ou de valeurs moyennes ?

Validation du modèle
pas d'informations sur les regroupements réels qui permettraient de comparer le schéma de regroupement, alors comment déterminer s'il est approprié ?

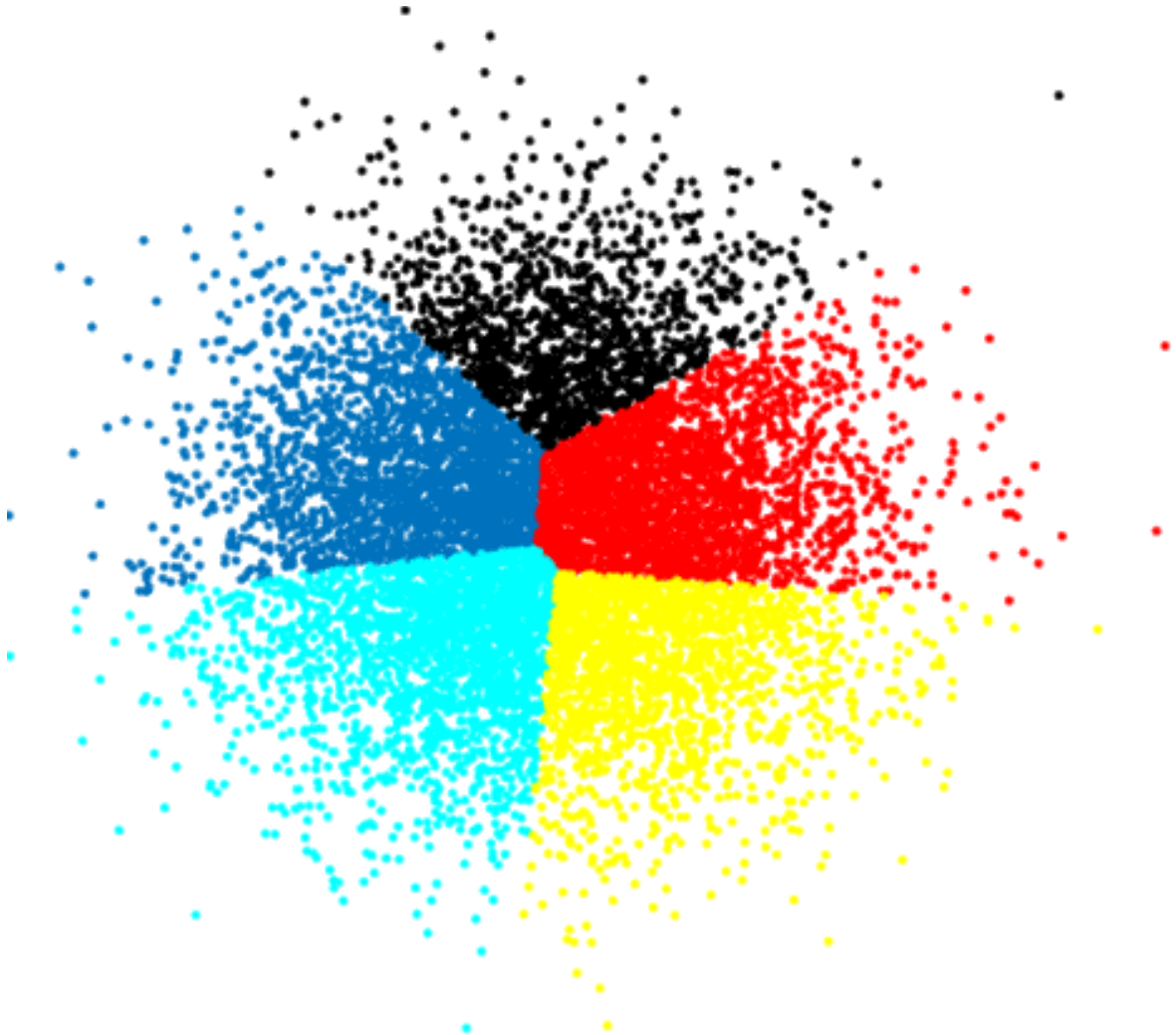
Regroupement fantôme
la plupart des méthodes trouveront des grappes même s'il n'y en a pas réellement

Rationalisation *a posteriori*
une fois les grappes trouvées, il est tentant d'essayer de les "expliquer" ...

Session 3



Session 3



Lectures conseillées

Validation et commentaires

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.5 Clustering
    - Validation

22. Focus on Clustering
    22.3 Clustering Evaluation

Exercices

Validation et commentaires

Considérez l'ensemble d'images de fruits ci-dessous.

Fournissez quelques schémas de regroupement pour les données et expliquez comment vous les valideriez.