# Introduction à la science des données

**Instructeur**: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

# APPRENTISSAGE STATISTIQUE

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics

[avec des fichiers de Jen Schellinck | Sysabee]

# TYPES D'APPRENTISSAGE

Le problème central de la science des données et de l'apprentissage machine est le suivant :

**pouvons-nous** (devrions-nous) concevoir des algorithmes capables d'apprendre ?

**Apprentissage supervisé (**apprentissage avec un enseignant**)**

- classification, régression, classements, recommandations
- utilisation de données de **formation étiquetées** (l'élève donne une réponse à chaque question d'examen en fonction de ce qu'il a appris à partir d'exemples élaborés)
- le rendement est évalué à l'aide de **données d'essai** (l'enseignant fournit les bonnes réponses)
- il existe une **cible / référence** sur laquelle on peut entraîner le modèle

# LES TYPES D'APPRENTISSAGE

**Apprentissage non supervisé** (regroupement d'exercices semblable en tant qu'outil d'aide à l'étude)

- agglomération, découverte de règles d'association, profilage de liens, détection d'anomalies
- utilisation des observations **non étiquetées** (l'enseignant n'est pas impliqué)
- l'exactitude **ne peut pas** être évaluée (les élèves pourraient ne pas se retrouver avec les mêmes regroupements)
- Le concept de cible n'est pas applicable.

**Autres:**

- **Apprentissage semi-supervisé** (l'enseignant fournit des exemples et une liste de problèmes non résolus)
- **Apprentissage de renforcement** (entreprendre un doctorat avec un conseiller)

# NOTIONS DE BASE SUR LES RÈGLES D'ASSOCIATION

La **découverte de règles d'association** est un type d'apprentissage non supervisé qui trouve des liens entre des attributs (et des combinaisons d'attributs).

**Exemples:**

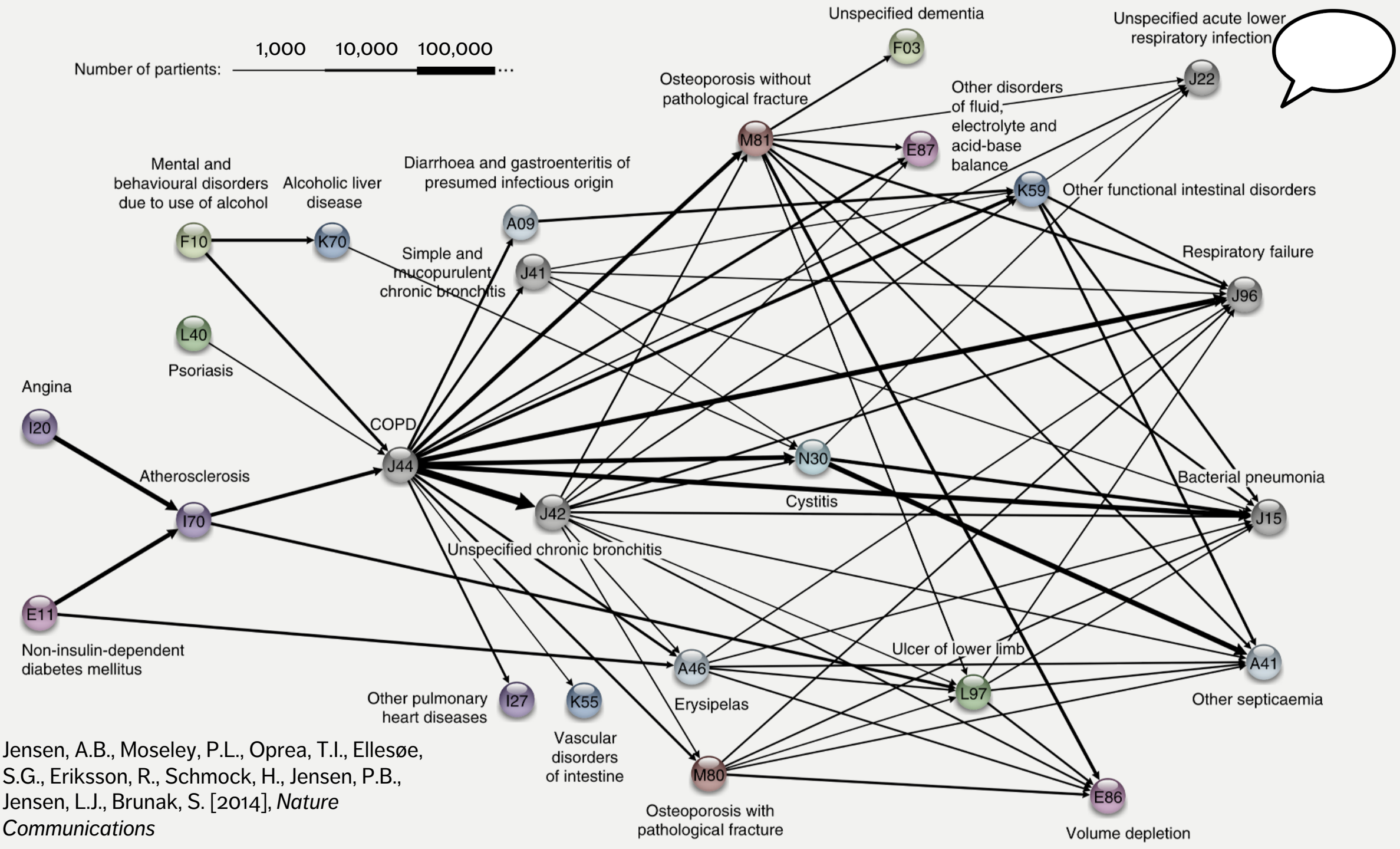
- le pain et le lait sont souvent achetés ensemble... est-ce intéressant ?
- les hot-dogs et la moutarde sont également souvent achetés par paire, mais plus rarement achetés individuellement... est-ce intéressant ?

Un supermarché pourrait alors faire des soldes sur les hot-dogs pour attirer les clients, tout en augmentant le prix des condiments pour maintenir les marges bénéficiaires.

[E. Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die]

# CAUSALITÉ ET CORRÉLATION

| Observations | Organisation |
|---|---|
| Achats de Pop-Tarts avant un ouragan | Walmart |
| Plus le taux de crime est élevé, plus les gens prennent des Uber | Uber |
| Le fait d'utiliser correctement les majuscules est corrélé à la solvabilité | Jeune entreprise de services financiers |
| Les utilisateurs des navigateurs Chrome et Firefox font de meilleurs employés | Cabinet de services professionnels en ressources humaines se fiant aux données sur les employés de Xerox et d'autres entreprises |
| Les hommes qui sautent le petit-déjeuner ont plus de maladies coronariennes | Chercheurs en médecine de l'Université Harvard |
| Les employés les plus motivés ont moins d'accidents | Shell |
| Les gens intelligents aiment les frites ondulées | Chercheurs à l'Université de Cambridge et à Microsoft Research |
| Les ouragans portant des noms féminins sont plus meurtriers | Chercheurs universitaires |
| Plus leur statut est élevé, moins les gens sont polis | Des chercheurs examinant les comportements sur Wikipédia |




Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesøe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J., Brunak, S. [2014], *Nature Communications*

# APERÇU DE LA CLASSIFICATION

Dans la **classification**, un échantillon de données (**l'ensemble d'apprentissage**) est utilisé pour déterminer les règles et les modèles qui divisent les données en groupes prédéterminés, ou classes (apprentissage supervisé ; analyse prédictive).

Les données d'apprentissage sont généralement constituées d'un sous-ensemble de données **étiquetées** (cibles) sélectionné de manière **aléatoire**.

**L'estimation de la valeur** (régression) s'apparente à la classification lorsque la variable cible est numérique.

# APERÇU DE LA CLASSIFICATION

Dans la phase de **test**, le modèle est utilisé pour attribuer une classe aux observations pour lesquelles l'étiquette est cachée, mais finalement connue (ensemble de test).

Les performances d'un modèle de classification sont évaluées sur l'ensemble de test, **jamais** sur l'ensemble de formation.

Les questions techniques comprennent :

- la sélection des caractéristiques à inclure dans le modèle
- le choix de l'algorithme
- etc.

# MÉTHODES DE CLASSIFICATION

Régression logistique

Réseaux neuronaux

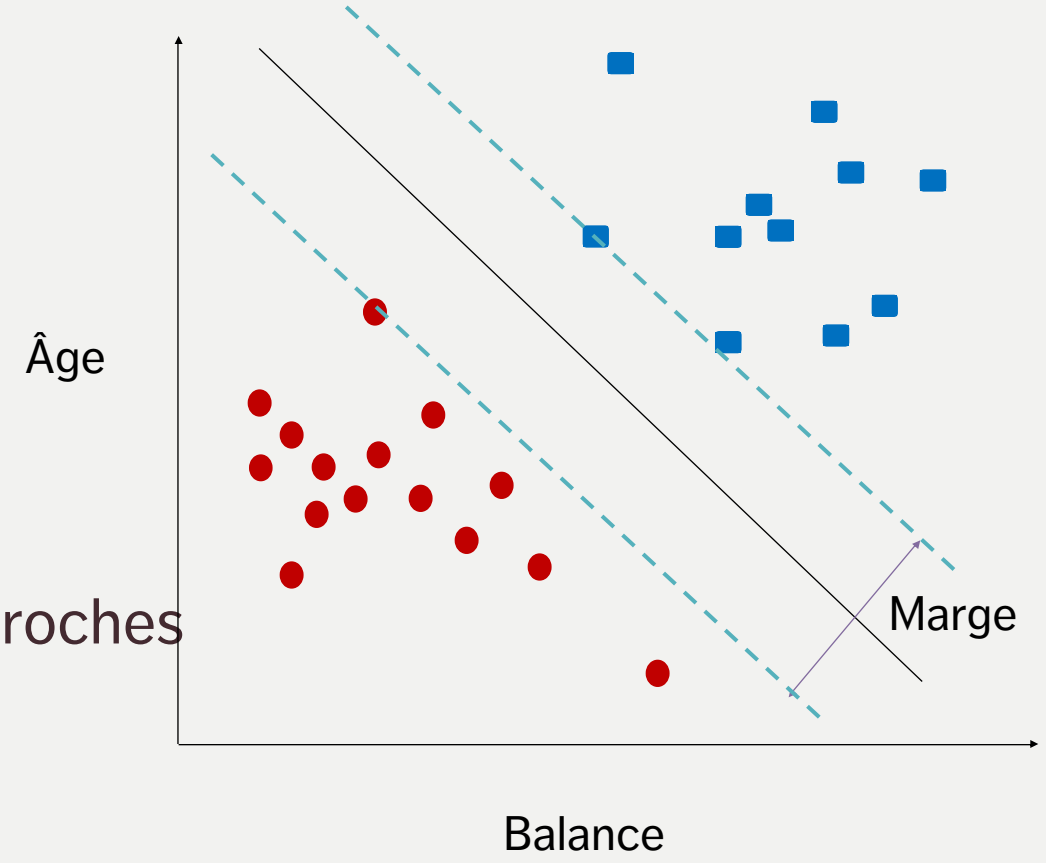
Arbres de décision

Classificateurs Naïve Bayes

Machines à vecteurs de soutien

Classificateurs à base de voisins les plus proches

etc.

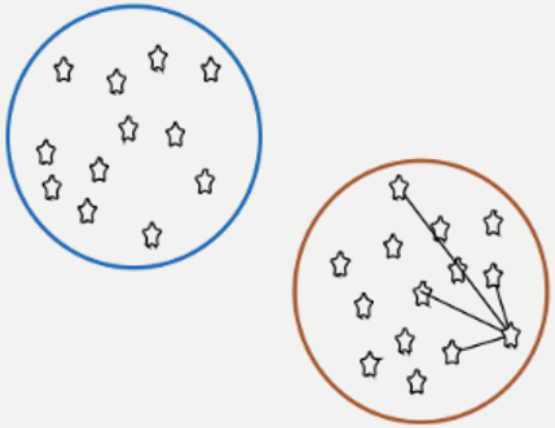


# APERÇU DU REGROUPEMENT

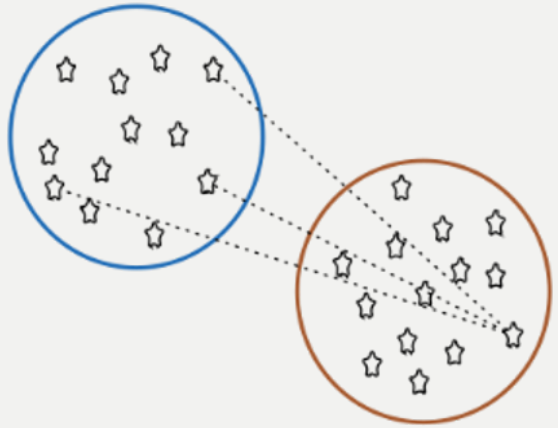
Dans un **regroupement**, les données sont réparties en groupes formés naturellement. Dans chaque groupe, les points de données sont similaires; d'un groupe à un autre, les points de données sont distincts.

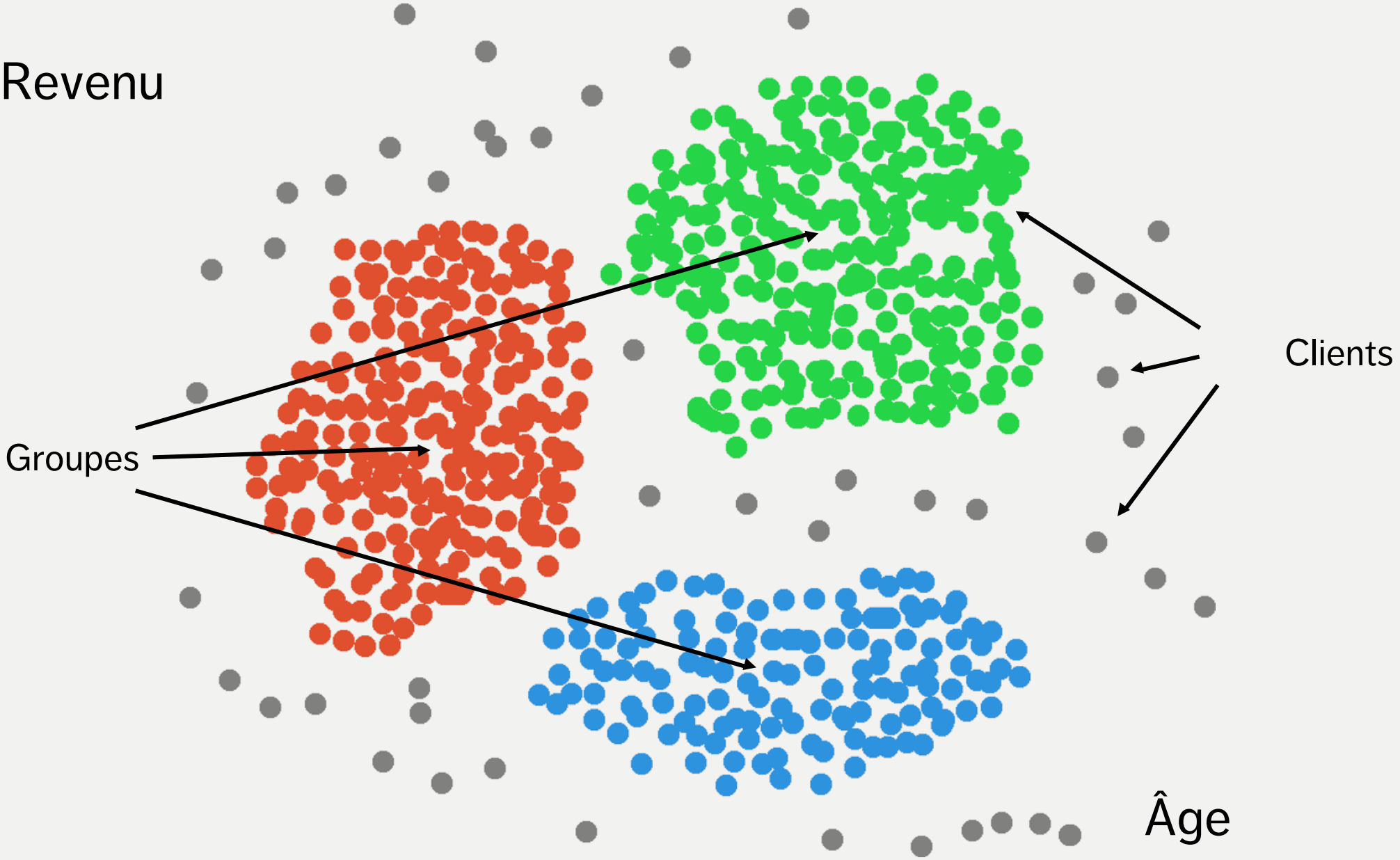
Les étiquettes des groupes ne sont **pas déterminées** au préalable (apprentissage non supervisé).

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
Clients
Âge

# MODÈLES DE REGROUPEMENT

$k$ −moyennes

Regroupement hiérarchique

Allocation de Dirichlet latente

Maximisation de l'espérance

Réduction et regroupement itératifs et équilibrés au moyende hiérarchie (BIRCH)

Regroupement par densité spatiale des applications avec bruit (DBSCAN)

Propagation par affinité

etc.

# MAUVAISES DONNÉES

L'ensemble de données semble-t-il fiable ? (entrées non valides, etc.)

Détection des **mensonges** et des **erreurs** (erreurs de déclaration, langage polarisant)

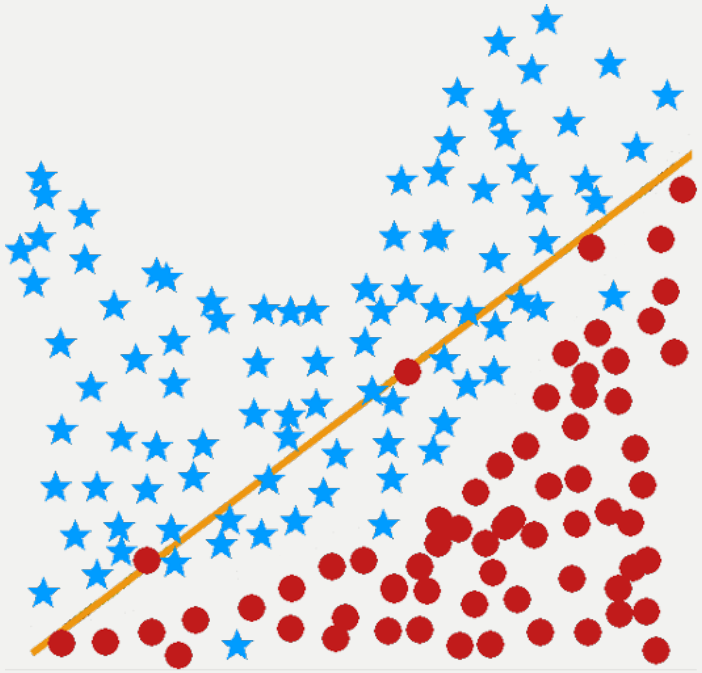
**Est-ce que l'approximation est suffisante ?**

Sources de **biais** et **d'erreurs**

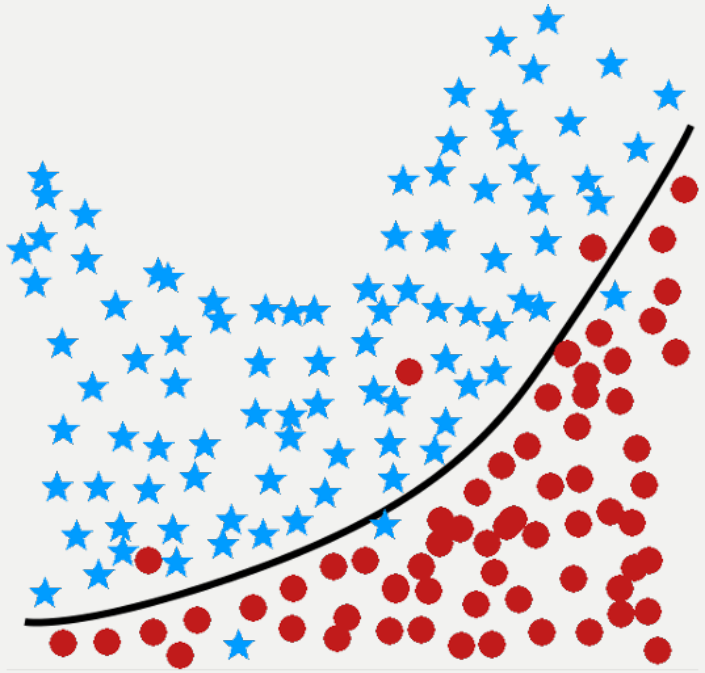
Recherche de la **perfection** (données universitaires, professionnelles, gouvernementales, relatives au service)

Les **pièges** de la science des données : analyse sans compréhension, utilisation d'un seul outil (par choix/décret), analyse pour l'analyse, attentes irréalistes à l’égard de la science des données, selon le besoin de savoir et vous n’avez pas besoin de savoir.

# SURAPPRENTISSAGE



Sous-apprentissage



Bonne représentation

Surapprentissage

# COMPARAISON ENTRE LES MÉGADONNÉES (*BIG DATA*) ET LES PETITES DONNÉES

## Quelle est la principale différence ?

- les ensembles de données sont **VOLUMINEUX**
- problèmes : collecte, saisie, accès, stockage, analyse, visualisation

## D'où viennent les données ?

- les progrès technologiques permettent de dépasser les limites de vitesse de traitement des données
- détection de l'information, appareils mobiles, appareils photo et réseaux sans fil

## Quels sont les défis ?

- la plupart des techniques ont été élaborées pour de très petits ensembles de données
- les méthodes directes peuvent pendant des années

# PERTINENCE ET PORTABILITÉ

Les méthodes de la science des données ne sont **pas** appropriées :

- si l'on doit absolument utiliser des ensembles de données existant (hérité) au lieu d'un jeu de données idéal (« ce sont les meilleures données dont nous disposons! »)
- si l'ensemble de données possède des attributs qui permettent de prédire utilement une valeur d'intérêt, mais qui ne sont pas disponibles lorsqu'une prédiction est requise
- si l'on va tenter de prédire l'appartenance à une classe en utilisant un algorithme d'apprentissage non supervisé

Si les données sont utilisées dans d'autres contextes ou pour effectuer des prédictions en fonction d'attributs sans données, on ne peut valider les résultats.

**Exemple :** Pouvons-nous utiliser un modèle qui prédit les emprunteurs hypothécaires en défaut pour prévoir également les détenteurs d'un prêt auto en défaut?

# BIAIS, SOPHISMES ET INTERPRÉTATION

La corrélation n'est pas un lien de causalité.

Les tendances extrêmes peuvent induire en erreur.

Il faut rester dans les limites d'une étude.

Gardez le taux de base à l'esprit.

Des résultats étranges se produisent parfois (paradoxe de Simpson).

Toute activité analytique comporte une composante humaine.

De petits effets peuvent quand même être (statistiquement) significatifs.

Méfiez-vous des statistiques sacro-saintes (valeur $p$, etc.)

La présence d'un biais invalide-t-elle nécessairement les résultats?