Introduction à la science des données

Instructeur: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

Patrick Boily

Carrière :

Professeur [uOttawa] (~55 cours/ ~150 journées d'atelier)

Gérant ['12 – '19, CQADS/CAQAD, Carleton]

Fonctionnaire ['08 – '12, ASFC | StatCan | TC | TPSGC]

Clients :

AMC, SGDN, ACSTA, plusieurs autres (~40 projets)

Spécialités :

Visualisation des données, nettoyage des données, application d'un large éventail de méthodes quantitatives.

APPRENTISSAGE STATISTIQUE

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics

[avec des fichiers de Jen Schellinck | Sysabee]

CONTEXTE D'APPRENTISSAGE

APPRENTISSAGE STATISTIQUE

« Nous apprenons de l'échec, pas du succès ! »
(Bram Stoker, *Dracula*)

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)

RÈGLES D'ASSOCIATION

APPRENTISSAGE STATISTIQUE

M. SNIFF : Qu'est-ce que tu cherches ?
SNOOP : Un billet de cinq dollars.
SNIFF : Tu es sûr de l'avoir perdu dans cette rue ?
SNOOP : Oh non ! Je l'ai perdu dans le bloc suivant, mais je regarde ici parce que la lumière est meilleure.

(*Boys' Life Magazine*, 1932)

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.

APPLICATIONS

Concepts apparentés

- Recherche de paires (triplets, etc.) de mots qui représentent un concept commun
- {Ottawa, Sénateurs}, {Michelle, Obama}, {veni, vidi, vici}, etc.

Plagiat

- Recherche de phrases qui apparaissent dans divers documents
- Recherche de documents qui ont des phrases en commun

Biomarqueurs

- maladies fréquemment associées à un ensemble de biomarqueurs

CAUSALITÉ ET CORRÉLATION

Les règles d'association peuvent automatiser la découverte d'hypothèses, mais il faut **rester prudent en matière de corrélation**.

Si les attributs A et B sont corrélés, alors soit :

- A et B sont corrélés **entièrement par accident** dans l'ensemble de données
- A est un nouvel étiquetage de B
- A est (une) cause de B et/ou B est (une) cause de A
- une combinaison d'attributs $C_1, \ldots, C_n$ (connus ou non) est responsable de A & B

[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

DÉFINITIONS

Une règle $X \rightarrow Y$ est un énoncé prenant la forme “si X alors Y” établi à partir de n’importe quelle combinaison logique d’attributs d’un ensemble de données.

Il **n’est pas nécessaire qu’une règle soit valide pour toute les observations** de l’ensemble de données (les règles ne sont pas nécessairement exactes).

Parfois, les meilleures règles sont celles qui sont exactes 10% du temps, en opposition à celles qui ne le sont que 5% du temps, mettons.

Comme toujours, **cela dépend du contexte**.

Défi technique: trouver un **petit** ensemble de règles raisonnables.

DÉFINITIONS

On détermine la force d'une règle à l'aide de **mesures**:

- Le **support** (couverture) mesure la fréquence à laquelle une règle se produit dans un ensemble de données. Une petite couverture indique que la règle se produit rarement (qu'elle soit valide ou non).
- La **confiance** (exactitude) mesure la fiabilité de la règle: à quelle fréquence le conséquent apparaît-il lorsque l'on observe l'antécédent ? Les règles avec une confiance élevée sont « plus vraie ».
- L'**intérêt** mesure la différence entre la confiance et la fréquence relative du conséquent. Les règles ayant un intérêt absolu plus élevée sont plus intéressantes, en général.
- Le « **lift** » mesure l'augmentation de la fréquence d'apparition du conséquent attribuable à la présence de l'antécédent. Si le lift est élevé (> 1), le conséquent se produit plus fréquemment qu'il ne le ferait s'il était indépendent de l'antécédent.

FORMULES

Si N est le nombre d'observations dans l'ensemble de données:

- $\text{Support}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{N} \in [0,1]$ ← Proportion de cas où l'antécédent et le conséquent se produisent ensemble
- $\text{Confiance}(X \to Y) = P(Y|X) = \frac{\text{Freq}(X \cap Y)}{\text{Freq}(X)} \in [0,1]$ ← Proportion de cas où le conséquent survient lorsque l'antécédent est observé
- $\text{Intérêt}(X \to Y) = \text{Confiance}(X \to Y) - \frac{\text{Freq}(Y)}{N} \in [-1,1]$
- $\text{Lift}(X \to Y) = \frac{N^2 \cdot \text{Support}\,(X \to Y)}{\text{Freq}(X) \cdot \text{Freq}(Y)} \in (0, N^2]$

← … ?!?

EXAMPLE

Ensemble de données musicales hypothétique contenant des données pour $N = 15{,}356$ mélomanes.

Règle musicale (RM): si une personne est née avant 1976 (X), elle possède alors une copie d'au moins un album des Beatles (Y).

Supposons que

- Freq$(X) = 3888$ personnes sont nées avant 1976
- Freq$(Y) = 9092$ personnes possède une copie d'au moins un album des Beatles
- Freq$(X \cap Y) = 2720$ personnes sont nées avant 1976 et possède une copie d'au moins un album des Beatles

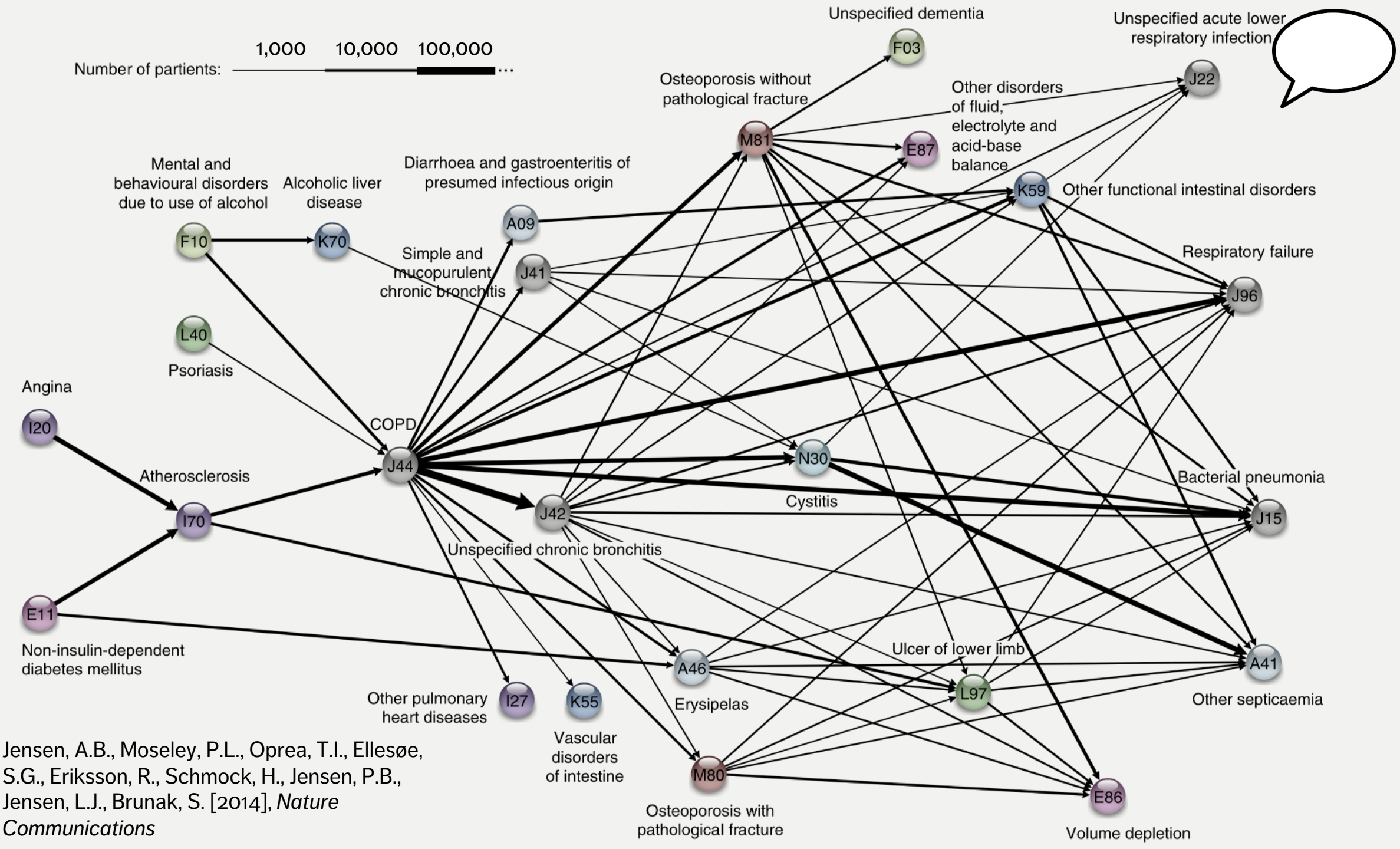
EXEMPLE

$$1.2 \approx \frac{0.70}{0.56}$$

Les 4 mesures sont:

- $\text{Support}(RM) = \frac{2720}{15{,}356} \approx 18\%$ (RM se produit dans 18% des observations)
- $\text{Confiance}(RM) = \frac{2720}{3888} \approx 70\%$ (RM est valide pour 70% des pré- 1976)
- $\text{Intérêt}(RM) = \frac{2720}{3888} - \frac{9092}{15356} \approx 0.11$ (RM n'est pas très intéressante)
- $\text{Lift}(RM) = \frac{15{,}356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$ (faible corrélation entre le fait d'être né avant 1976 et le fait de posséder une copie d'un album des Beatles)

Interprétation du lift: 70% de ceux nés avant 1976 possède une copie, tandis que 56% de ceux nés après 1976 possède une copie.



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*

CLASSIFICATION

APPRENTISSAGE STATISTIQUE

« La science des données ne remplace pas la modélisation statistique et l'analyse des données ; elle les augmente. »
(P. Boily)

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.

APPLICATIONS

Médecine et sciences de la santé

- prédire quel patient risque de subir une crise cardiaque mortelle dans les 30 jours suivant une première crise cardiaque, sur la base de facteurs de santé (pression artérielle, âge, problèmes de sinus, etc.)

Politiques sociales

- prédire la probabilité d'avoir besoin d'un logement d'assistance à la vieillesse sur la base d'informations démographiques/de réponses à une enquête

Marketing et affaires

- prédire quels clients sont susceptibles de passer à un autre opérateur de téléphonie mobile en fonction de leurs caractéristiques démographiques et de leur utilisation.

MÉTHODES DE CLASSIFICATION

Régression logistique

Réseaux neuronaux

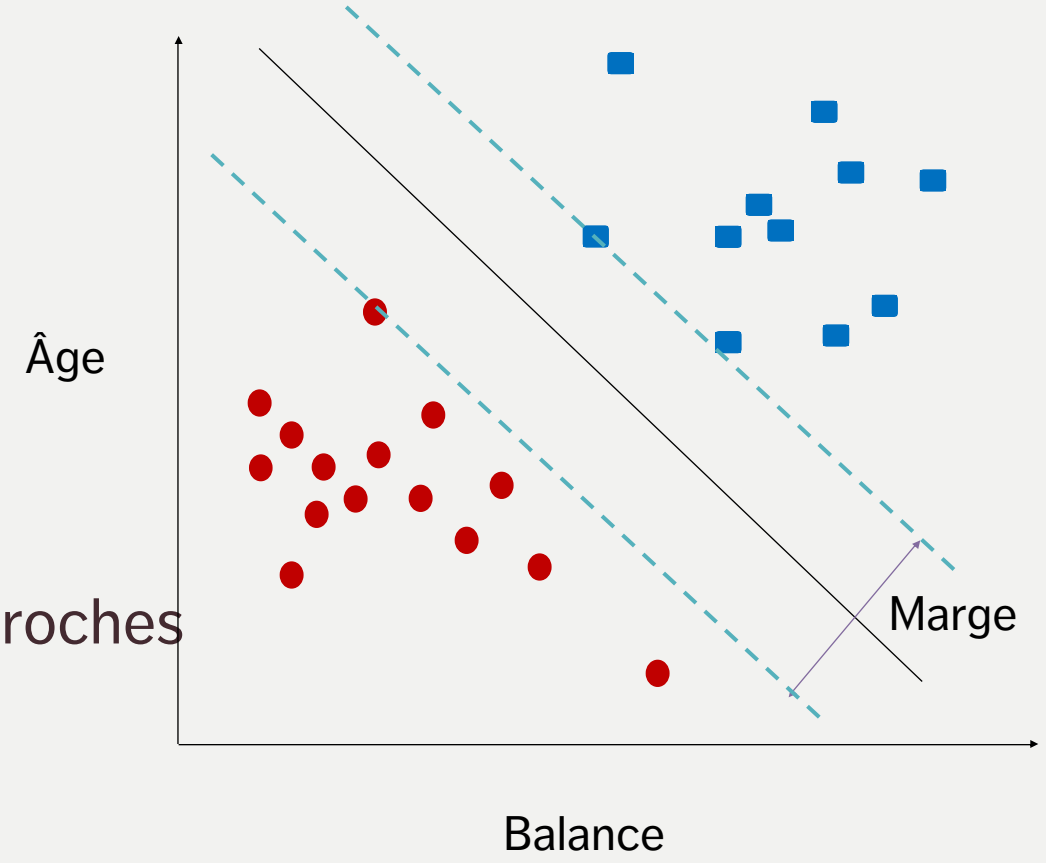
Arbres de décision

Classificateurs Naïve Bayes

Machines à vecteurs de soutien

Classificateurs à base de voisins les plus proches

etc.



[adapted from A. Ng, K. Soo, *Numsense!*, Surviving a Disaster]

ARBRES DE DÉCISION

Les arbres de décision sont peut-être la plus **intuitive** de ces méthodes.

La classification est réalisée en suivant un chemin le long de l'arbre, depuis sa **racine**, en passant par ses **branches**, jusqu'à ses **feuilles**.

D'AUTRES POINTS DE RÉFLEXION

La classification est liée à **l'estimation des probabilités**

- les approches basées sur des modèles de régression pourraient s'avérer utiles

Les occurrences rares (souvent plus intéressantes/importantes) continuent de perturber les tentatives de classification

- les données historiques du réacteur nucléaire de Fukushima avant la fusion n'ont pas pu être utilisées pour tirer des conclusions sur les fusions.

« Rien n'est gratuit » : Aucun classificateur ne donne les meilleurs résultats pour toutes les données.

Avec les mégadonnées, on doit également tenir compte de l'**efficacité**.

ÉVALUATION DE LA PERFORMANCE

		Predicted A	Predicted B		Total
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted A	Predicted B		Total
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
Total		70	11	81	
		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

REGROUPEMENT

APPRENTISSAGE STATISTIQUE

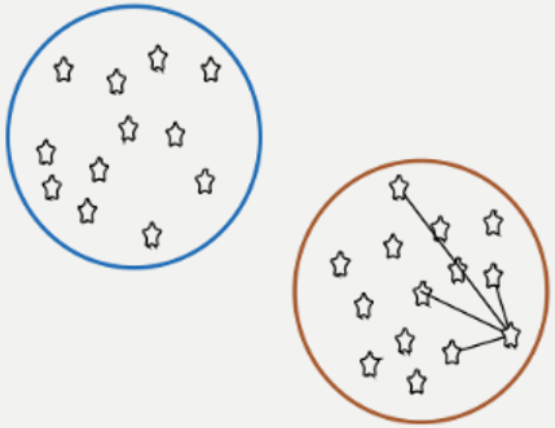
« Les données ne sont pas des informations, les informations ne sont pas des connaissances, les connaissances ne sont pas la compréhension, la compréhension n'est pas la sagesse. »
(C. Stoll)

APERÇU DU REGROUPEMENT

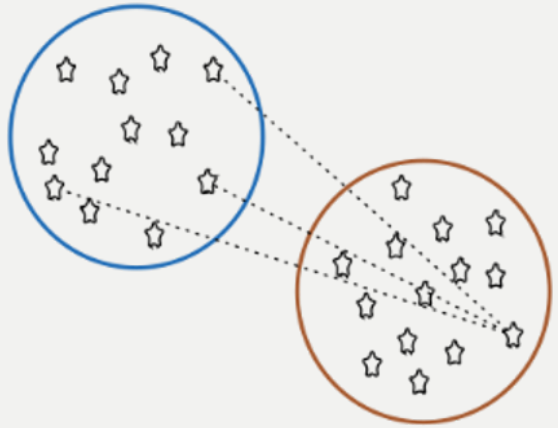
Dans un **regroupement**, les données sont réparties en groupes formés naturellement. Dans chaque groupe, les points de données sont similaires; d'un groupe à un autre, les points de données sont distincts.

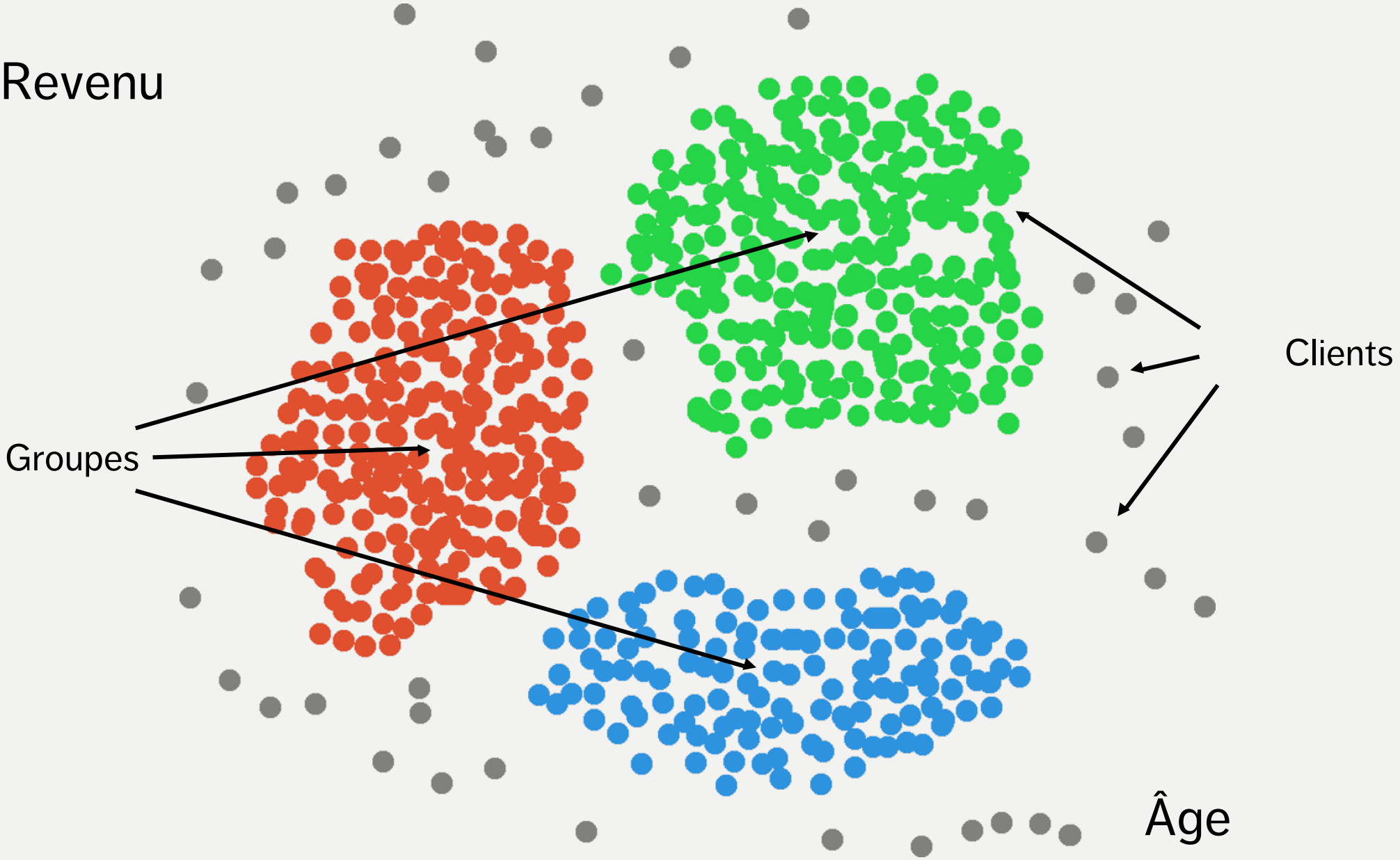
Les étiquettes des groupes ne sont **pas déterminées** au préalable (apprentissage non supervisé).

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

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.

MODÈLES DE REGROUPEMENT

k −moyennes

Regroupement hiérarchique

Allocation de Dirichlet latente

Maximisation de l'espérance

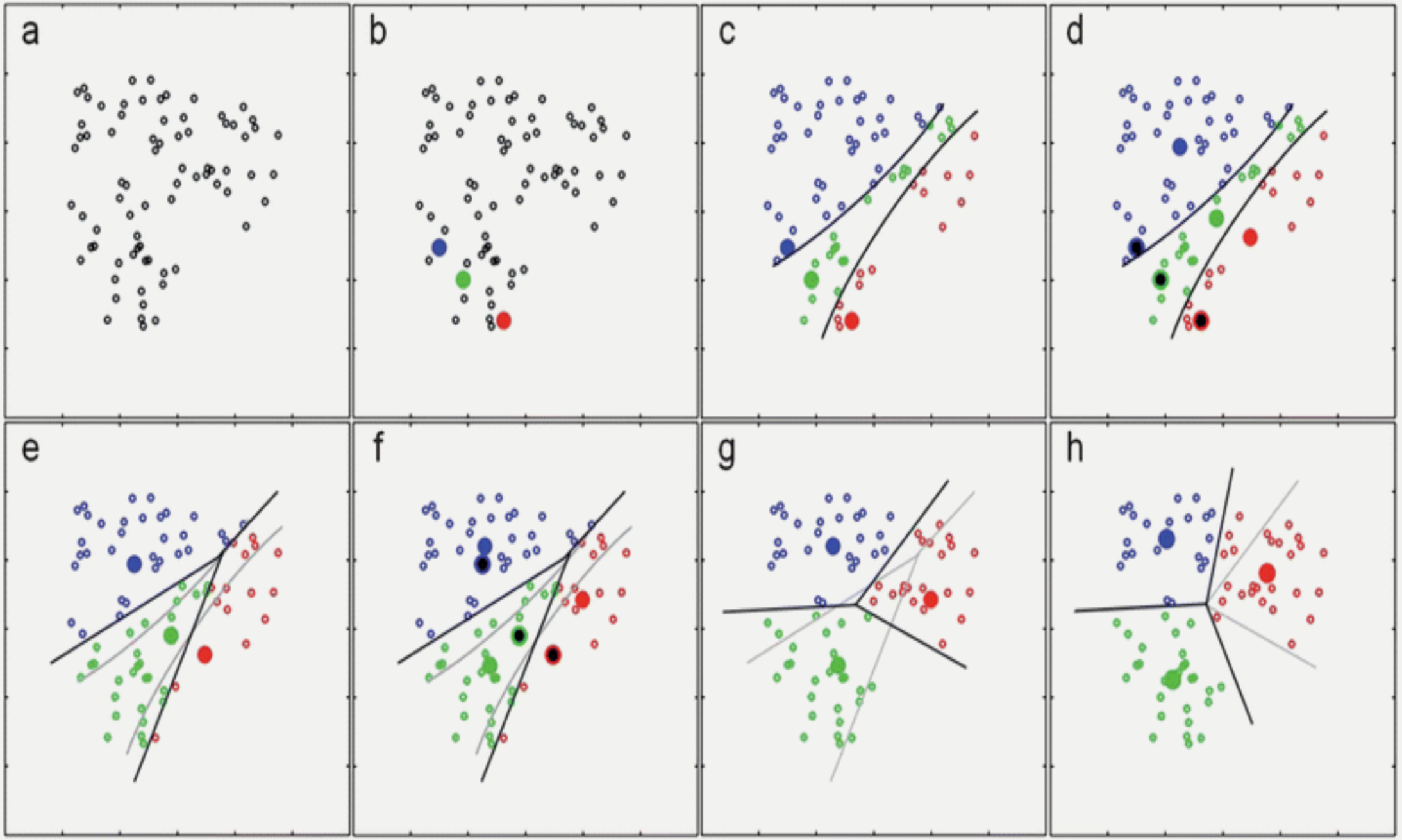
Réduction et regroupement itératifs et équilibrés au moyende hiérarchie (BIRCH)

Regroupement par densité spatiale des applications avec bruit (DBSCAN)

Propagation par affinité

etc.

[https://notebook.community/alanhdu/AccessibleML/4. K-means clustering and unsupervised learning]



DÉFIS DU REGROUPEMENT

Automatisation

Absence d'une définition precise

Absence de reproductibilité

Nombre de groups

Description d'un groupe

Validation modèle

Regroupement fantôme

Rationalisation *a posteriori*

ENJEUX ET DÉFIS

APPRENTISSAGE STATISTIQUE

« Nous disons tous que nous aimons les données, mais ce n'est pas vrai. Ce que nous aimons, c'est obtenir des perspectives grâce aux données. Cela n'équivaut pas tout à fait à aimer les données. En fait, j'ose dire que je ne me soucie pas vraiment des données, et il semblerait que je ne suis pas le seul. »

(Q.E. McCallum, *Bad Data Handbook*

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)

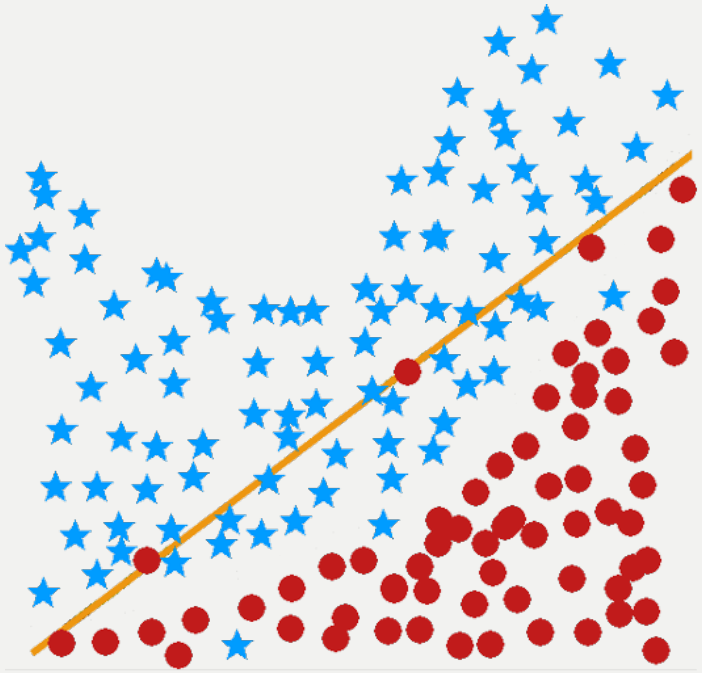
Est-ce que l'approximation est suffisante ?

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

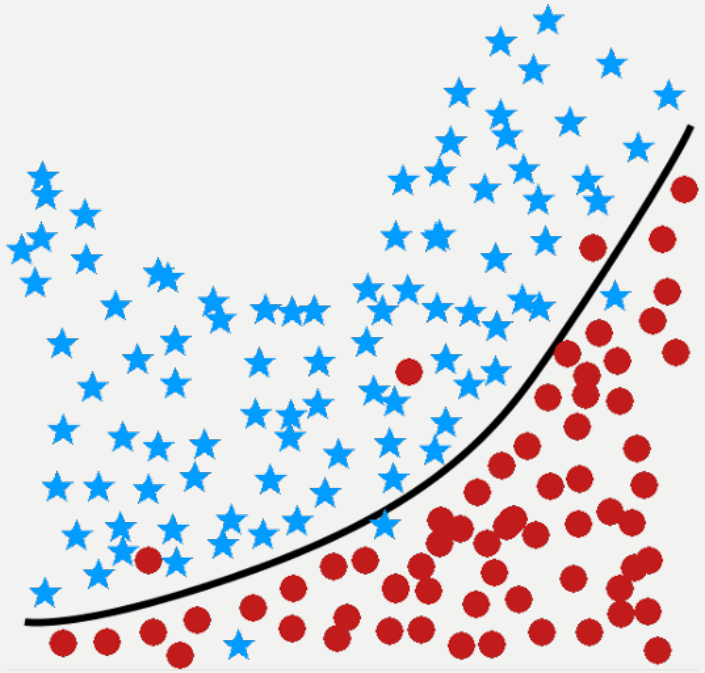
Recherche de la **perfection** (données universitaires, professionnelles, gouvernementales, relatives au service)

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 ?

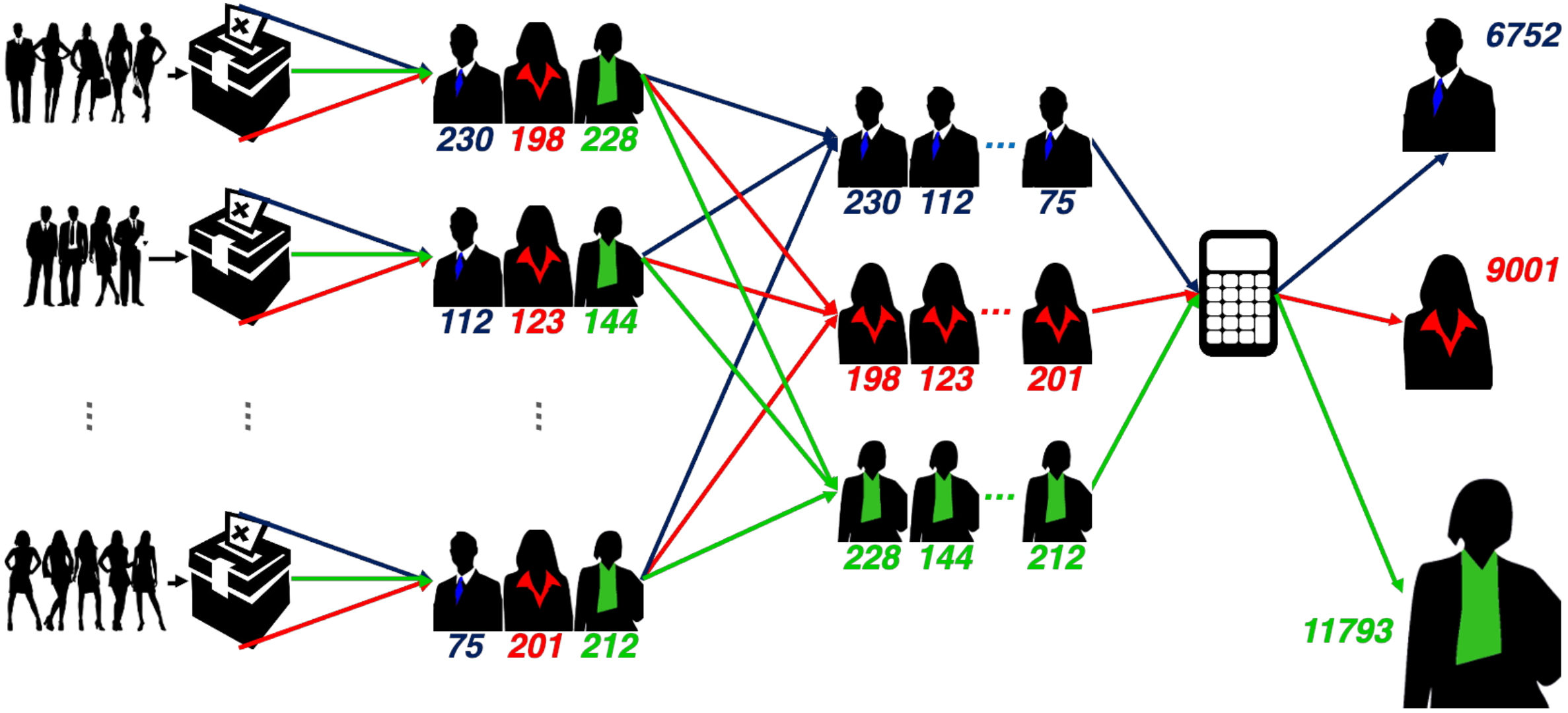
- les ensembles de données sont **VOLUMINEUX**
- problèmes : collecte, saisie, accès, stockage, analyse, visualisation

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


230 198 228
112 123 144
75 201 212
230 112 ... 75
198 123 ... 201
228 144 ... 212
6752
9001
11793

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?