

# APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

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 *Nothing to Hide: Privacy in the 21st Century* de Keeler, 2006)

# QU'EST-CE QUE LA SCIENCE DES DONNÉES? (REPRISE)

La science des données est l'ensemble des processus par lesquels nous extrayons **des renseignements utiles et exploitables** à partir des données.

(paraphrasé d'après T. Kwartler)

La science des données constitue l'**intersection fonctionnelle** de la statistique, de l'ingénierie, de l'informatique, de l'expertise du domaine et du « hacking ». Elle s'articule autour de deux axes principaux : l'**analyse** (compter les choses) et l'**invention de nouvelles techniques** pour tirer des enseignements des données.

(Paraphrasé d'après H. Mason)

# APPRENTISSAGE EN GÉNÉRAL

Au-delà d'un « simple coup d'œil rapide », les personnes apprennent par l'intermédiaire de ce qui suit :

- en répondant à des questions
- en testant des hypothèses
- en créant des concepts
- en faisant des prévisions
- en créant des catégories et en classant des objets
- en regroupant des objets

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

**peut-on concevoir des algorithmes qui peuvent apprendre?**

# TYPES D'APPRENTISSAGE

## 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)

## 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)

# TYPES D'APPRENTISSAGE

**Apprentissage semi-supervisé** (l'enseignant fournit des exemples **et** une liste de problèmes non résolus)

**Apprentissage de renforcement** (entreprendre un doctorat avec un conseiller)

---

Dans **l'apprentissage supervisé**, il existe une cible par rapport à laquelle il faut former le modèle. Dans **l'apprentissage non supervisé**, nous ne savons pas quelle est la cible, ni même s'il y en a.

La distinction est **cruciale**. Assurez-vous de la comprendre.

## ÉTUDE DE CAS: ÉTUDE MÉDICALE DANOISE

Le *Danish National Patient Registry* contient **68 millions** d'observations médicales sur **6,2 millions de patients** sur une période de 15 ans (janvier 1996 – novembre 2010).

### Objectifs :

- trouver des liens entre les différents diagnostics
- déterminer comment un diagnostic à un moment donné permettrait de prévoir un autre diagnostic à un moment ultérieur

# MÉTHODOLOGIE

1. Calcul du **degré de corrélation** pour des paires de diagnostics sur une période de cinq ans sur un sous-ensemble représentatif des données.
2. Tester la **directionnalité** des paires de diagnostics (un diagnostic survenant de façon répétée avant l'autre).
3. Déterminer des trajectoires de diagnostic raisonnables (**voies de communication**) en combinant de plus petites trajectoires fréquentes avec des diagnostics qui se chevauchent.
4. Valider les trajectoires par comparaison avec des données **non danoises**
5. Regrouper les voies de communication pour identifier les conditions médicales centrales (**principaux diagnostics**) autour desquelles s'organise la progression de la maladie.

# RÉSULTATS

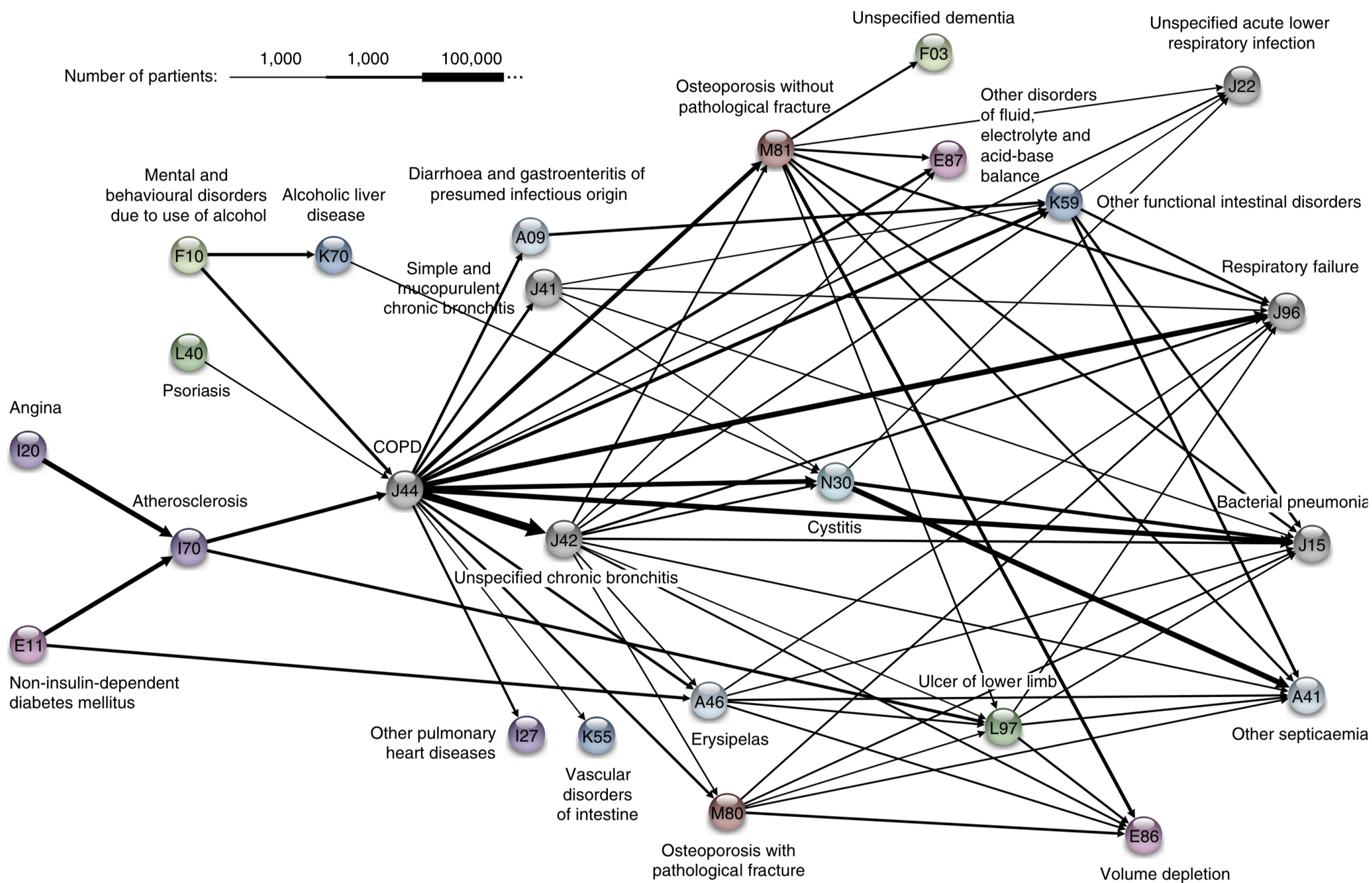
Les données ont été réduites à 1 171 voies de communication visant :

- le diabète
- la maladie pulmonaire obstructive chronique (MPOC)
- le cancer
- l'arthrite
- les maladies cardiovasculaires

L'analyse des données a permis d'établir, entre autres :

- que des diagnostics d'anémie sont ultérieurement suivis de la découverte d'un cancer du côlon
- que la goutte est un précurseur de maladies cardiovasculaires
- que la MPOC est **sous-diagnostiquée** et **sous-traitée**.





# 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).

**Exemple :** nous pourrions analyser un ensemble de données sur les activités physiques et les habitudes d'achat de la population nord-américaine et découvrir que

- *les coureurs qui sont aussi des triathlonsiens (l'**antécédent**) ont tendance à conduire des Subarus, à boire des bières de microbrasserie et à utiliser des téléphones intelligents (le **conséquent**);*
- les personnes qui ont acheté de l'équipement de gymnastique à domicile sont peu susceptibles de l'utiliser un an plus tard (pour ne nommer que quelques possibilités fictives).

# APPLICATION ORIGINALE

Les supermarchés enregistrent le contenu des paniers aux caisses pour déterminer les articles qui sont souvent achetés ensemble.

## Exemples

- Le pain et le lait sont souvent achetés ensemble, mais ce n'est pas très intéressant étant donné la fréquence à laquelle ils sont achetés individuellement.
- Les « hot dogs » et la moutarde sont aussi souvent achetés ensemble, mais plus rarement à l'unité.

Ainsi, un supermarché pourrait offrir une réduction sur les hot dogs tout en augmentant le prix des condiments.

# AUTRES 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** (ce qui est moins répandu chez les scientifiques des données qu'on ne l'espère...).

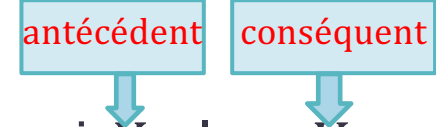
Si les attributs  $A$  et  $B$  sont corrélés, il y a (au moins) cinq possibilités :

- $A$  et  $B$  sont **entièrement corrélés par hasard** dans cet ensemble de données particulier
- $A$  est un nouvel étiquetage de  $B$
- $A$  donne  $B$
- $B$  donne  $A$
- les combinaisons d'autres attributs  $C_1, \dots, C_n$  (connus ou non) donnent  $A$  et  $B$

# 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 de « 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 vraie pour toutes les observations** de l'ensemble de données (c.-à-d. que les règles ne sont pas nécessairement exactes à 100 %).

En fait, parfois, les « meilleures » règles pourraient être celles qui ne sont exactes que 10 % du temps, par opposition aux règles qui ne sont exactes que 5 % du temps, par exemple.

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

# DÉFINITIONS


Pour déterminer la force d'une règle, nous évaluons certains paramètres :


- **Le support** (couverture) mesure la fréquence à laquelle une règle se produit dans un ensemble de données. Une valeur de couverture faible indique que la règle se produit rarement (qu'elle soit vraie ou non).
- **La confiance** (exactitude) mesure la fiabilité de la règle : à quelle fréquence le conséquent se vérifie-t-il lorsque l'antécédent est observé? Les règles avec une grande confiance sont « plus vraies ».
- **L'intérêt** mesure la différence entre la confiance et la fréquence relative du conséquent. Les règles ayant un intérêt absolu élevé sont plus intéressantes.
- **Le « lift »** mesure l'augmentation de la fréquence d'apparition du conséquent attribuable à l'antécédent. Dans le cas d'une règle avec un lift élevé ( $> 1$ ), le conséquent se produit plus fréquemment qu'il ne le ferait s'il était indépendant de l'antécédent.




# FORMULES

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

- $\text{Support}(X \rightarrow Y) = \frac{\text{Fréq}(X \cap Y)}{N} \in [0,1]$  

Proportion de cas où l'antécédent et le conséquent se produisent ensemble
- $\text{Confiance}(X \rightarrow Y) = P(Y|X) = \frac{\text{Fréq}(X \cap Y)}{\text{Fréq}(X)} \in [0,1]$  

Proportion de cas où le conséquent survient lorsque l'antécédent est observé
- $\text{Intérêt}(X \rightarrow Y) = \text{Confiance}(X \rightarrow Y) - \frac{\text{Fréq}(Y)}{N} \in [-1,1]$
- $\text{Lift}(X \rightarrow Y) = \frac{N^2 \cdot \text{Support}(X \rightarrow Y)}{\text{Fréq}(X) \cdot \text{Fréq}(Y)} \in (0, N^2]$  

...?!?

## UN EXEMPLE SIMPLE

Ensemble de données musicales hypothétiques contenant des données pour  $N = 15,356$  mélomanes.

**Règle destinée aux candidats ( $RM$ ) :** « Si une personne est née avant 1976 ( $X$ ), elle possède alors une copie d'au moins un album des Beatles, dans un format quelconque ( $Y$ ) ».

Supposons que

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

## UN EXEMPLE SIMPLE

$$1,2 \approx \frac{0,70}{0,56}$$

Les quatre 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 vrai pour 70 % des personnes nées avant 1976)
- $\text{Intérêt}(RM) = \frac{2720}{3888} - \frac{9092}{15356} \approx 0.11$  ( $RM$  n'est pas très intéressant)
- $\text{Lift}(RM) = \frac{15,356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$  (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 % des personnes nées avant 1976 en possèdent une copie, alors que 56 % de celles nées après 1976 en possèdent une copie.

# ALGORITHME DE FORCE BRUTE

1. Générer des ensembles d'éléments (de taille 1, 2, 3, 4, etc.)
  - p. ex. {achat = Typique, adhésion = Faux, coupon = Oui}.
2. Créer des règles à partir de chaque ensemble d'éléments.
  - p. ex. **SI** (achat = Typique ET adhésion = Faux) **ALORS** coupon = Oui
3. Calculer le support, la confiance, l'intérêt, le lift pour chaque règle.
4. Ne conserver que les règles avec une couverture, une précision, un intérêt ou un lift (ou d'autres paramètres) « assez élevés ».
5. Ces règles sont considérées comme étant **vraies** pour l'ensemble de données – il s'agit de **nouvelles connaissances établies à partir des données**.

# PRODUCTION DE RÈGLES

Un **ensemble d'éléments** (ou cas) est une liste d'attributs et de valeurs.

Un ensemble de **règles** peut être créé en ajoutant « **SI ... ALORS** » à chacun des cas. À titre d'exemple, à partir du cas défini

{adhésion = Vrai, âge = Jeune, achat = Typique}

nous pouvons créer les règles

- **SI** (adhésion = Vrai ET âge = Jeune) **ALORS** achat = Typique
- **SI** adhésion = Vrai **ALORS** (âge = Jeune ET achat = Typique)
- **SI** ∅ **ALORS** (adhésion = Vrai ET âge = Jeune ET achat = Typique)
- etc.

# NOMBRE DE RÈGLES

Considéons un ensemble d'éléments  $C$  avec  $n$  membres.

Dans une règle établie à partir de  $C$ , chacun des  $n$  membres apparaît soit dans l'**antécédent**, soit dans le **conséquent**, donc il y a  $2^n$  de ces règles.

La règle selon laquelle chaque membre fait partie de l'antécédent (et le conséquent est nul) n'est pas permise; on peut donc établir  $2^n - 1$  règles à partir de  $C$ .

Le nombre de règles augmente de façon exponentielle lorsque le nombre de fonctions augmente linéairement.

Ce n'est pas une bonne chose.

# VALIDATION

L'algorithme de force brute fonctionne relativement bien pour de **petits ensembles de données** (petit nombre de caractéristiques).

Pour les **ensembles de données plus importants**, il peut être coûteux de produire des règles de cette façon (surtout lorsque le nombre d'attributs augmente). Comment produire des **règles** généralement **porteuses**?

Quelle est la **fiabilité** des règles d'association? Quelle est la probabilité qu'elles se produisent par **hasard**? Quelle est leur **pertinence**? Peut-on les généraliser en **dehors** de l'ensemble de données ou par rapport à de **nouvelles** données?

## REMARQUES

Comme les règles fréquentes correspondent à des occurrences répétées dans l'ensemble de données, les algorithmes qui produisent des ensembles d'éléments essaient souvent de **maximiser la couverture**.

Lorsque des **événements rares** sont plus significatifs (comme la détection d'une maladie rare), nous avons besoin d'algorithmes qui peuvent produire des ensembles d'éléments rares. **Il ne s'agit pas là d'un problème banal.**

Un rappel, malgré la réplique de Tufte : **il ne faut pas confondre corrélation et causalité.**



## AUTRES ALGORITHMES

**Données continues** ou **nominales** : les données continues doivent être regroupées en catégories pour que les règles d'association soient pertinentes. Il y a plus d'une façon de s'y prendre.

Les ensembles d'éléments sont parfois appelés **paniers de consommation**.

Autres algorithmes :

AIS, SETM, Apriori, AprioriTid, AprioriHybrid, Eclat, PCY, Multistage, Multihash, etc.

# ALGORITHME APRIORI

Élaboré au départ pour les données de transaction

- chaque ensemble de données raisonnable peut être transformé en un ensemble de données de transaction à l'aide de variables fictives

Trouve des **ensembles d'éléments fréquents** à partir desquels proposer des règles

- au lieu d'établir des règles à partir de tous les ensembles d'éléments possibles

Commence par identifier les éléments individuels fréquents dans la base de données et les étend à des ensembles d'éléments de plus en plus grands, en supposant qu'ils sont encore trouvés **assez fréquemment** dans l'ensemble de données.

- approche **ascendante**, utilise la propriété de fermeture décroissante du support

# ALGORITHME APRIORI

Élague les candidats qui présentent des **sous-tendances fréquentes**.

- exige un seuil de support
- ce seuil doit être suffisamment élevé pour réduire au minimum le nombre d'ensembles d'éléments fréquents

Par exemple, si un ensemble comportant un élément n'est pas fréquent, tout ensemble de deux éléments le contenant est également peu fréquent.

L'algorithme se termine lorsqu'aucune autre bonne extension n'est trouvée.

# FORCES ET LIMITES

Facile à mettre en œuvre, facile à paralléliser.

L'algorithme Apriori est **lent** et nécessite des balayages fréquents des ensembles de données.

- solutions possibles : **échantillonnage** et **séparation**

Pas idéal pour trouver des règles pour les ensembles d'éléments **peu fréquents** ou **rares**.

D'autres algorithmes l'ont supplanté depuis (valeur historique) :

- **Max-Miner** essaie d'identifier les ensembles d'éléments fréquents sans les énumérer; effectue des sauts dans l'espace au lieu d'utiliser une approche ascendante.
- **Eclat** est plus rapide et utilise la recherche en profondeur d'abord, mais nécessite une capacité de mémoire importante.