

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

« 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)

OBJECTIFS D'APPRENTISSAGE

Se familiariser avec les différentes approches d'apprentissage statistique (supervisé, non supervisé, etc.).

Se familiariser avec les concepts fondamentaux des règles d'association et leur application aux données.

APPRENTISSAGE STATISTIQUE

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

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

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)

ANALOGIE DE L'EXPLORATION

Qu'extrayons-nous? données (**terre**)

Qu'utilise-t-on pour l'extraction? techniques d'extraction de données (**outils de fouille**)

Que cherche-t-on? recherche de modèles/connaissances (**minéraux bruts**)

Que faisons-nous de la matière première? décrire les modèles/relations (**transformer les minéraux en quelque chose d'utile**)

Quel est le résultat ou le produit? modèles (**Ge, Ga, Si pour construire des transistors**)

Que faisons-nous avec le produit? appliquer des modèles à la prise de décisions fondées sur des données probantes (**utiliser des transistors dans les systèmes électriques**)

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.

EXERCICES

Quels sont quelques exemples de tâches d'apprentissage supervisées et non supervisées dans le monde des affaires? Dans un contexte de politique publique/gouvernemental?

EXERCICES

En supposant que les techniques d'extraction de données sont utilisées dans les cas suivants, déterminez si la tâche requise relève d'un apprentissage **supervisé (S)** ou **non supervisé (NS)**.

- Décider d'accorder ou non un prêt à un demandeur sur la base de données démographiques et financières (en se référant à une base de données comportant des données semblables sur des clients antérieurs).
- Dans une librairie en ligne, faire des recommandations aux clients concernant des articles supplémentaires à acheter en fonction des habitudes d'achat des transactions précédentes.
- Identifier un paquet de données de réseau comme étant dangereux (virus, attaque de pirates informatiques) sur la base d'une comparaison avec d'autres paquets ayant un statut de menace connu.
- Identifier les segments de clients semblables.
- Prévoir si une entreprise fera faillite en comparant ses données financières à celles d'entreprises semblables en faillite et non en faillite.

EXERCICES

En supposant que les techniques d'extraction de données sont utilisées dans les cas suivants, déterminez si la tâche requise relève d'un apprentissage **supervisé (S)** ou **non supervisé (NS)**.

- Estimation du temps de réparation requis pour un avion sur la base d'un dossier de panne.
- Tri automatique du courrier par numérisation des codes postaux.
- Il est plus difficile et plus coûteux de gagner de nouveaux clients que de conserver les clients existants. Le fait d'évaluer la probabilité qu'un client parte peut aider une organisation à concevoir des interventions efficaces, comme des rabais ou des services gratuits, afin de fidéliser les clients rentables de façon rentable.
- Certains médecins effectuent des tests inutiles ou surfacturent leur gouvernement ou leurs compagnies d'assurance. En utilisant les données de vérification, il est possible d'identifier ces fournisseurs et de prendre les mesures qui s'imposent.

EXERCICES

En supposant que les techniques d'extraction de données sont utilisées dans les cas suivants, déterminez si la tâche requise relève d'un apprentissage **supervisé (S)** ou **non supervisé (NS)**.

- Une analyse du panier de consommation peut aider à élaborer des modèles de prévision pour déterminer quels produits se vendent souvent ensemble. Cette connaissance des affinités entre les produits peut aider les détaillants à créer des forfaits promotionnels pour associer les articles qui se vendent mal à un ensemble de produits qui se vendent bien.
- Diagnostiquer la cause d'un état de santé est la première étape cruciale de l'intervention médicale. Outre l'état actuel, d'autres facteurs peuvent être pris en considération, notamment les antécédents médicaux du patient, les antécédents pharmaceutiques, les antécédents familiaux et d'autres facteurs environnementaux. Un modèle de prévision peut absorber toute l'information disponible à ce jour (pour ce patient et d'autres) et établir des diagnostics probabilistes, sous la forme d'un arbre de décision, ce qui élimine une bonne partie de la conjecture.

EXERCICES

En supposant que les techniques d'extraction de données sont utilisées dans les cas suivants, déterminez si la tâche requise relève d'un apprentissage **supervisé (S)** ou **non supervisé (NS)**.

- Les écoles peuvent élaborer des modèles pour identifier les élèves qui risquent de ne pas retourner à l'école. Ces étudiants peuvent être ciblés par des mesures correctives.
- Outre les données sur les clients, les entreprises de télécommunications stockent également des enregistrements détaillés des appels, qui décrivent précisément le comportement d'appel de chaque client. Les données uniques peuvent être utilisées pour établir le profil des clients, auxquels on peut vendre des produits en fonction de la similarité de leur enregistrement détaillé des appels avec celui d'autres clients.
- Statistiquement, tout équipement est susceptible de tomber en panne à un moment donné. Le fait de prévoir quelle machine est susceptible de s'arrêter est un processus complexe. Des modèles décisionnels permettant de prévoir les défaillances de machines pourraient être établis à partir de données antérieures, ce qui permettrait de faire des économies grâce à l'entretien préventif.

ÉTUDE DE CAS : ÉTUDE MÉDICALE DANOISE

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

Trajectoires temporelles des maladies condensées à partir des données d'un registre à l'échelle de la population couvrant 6,2 millions de patients

(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*).

CONTEXTE

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

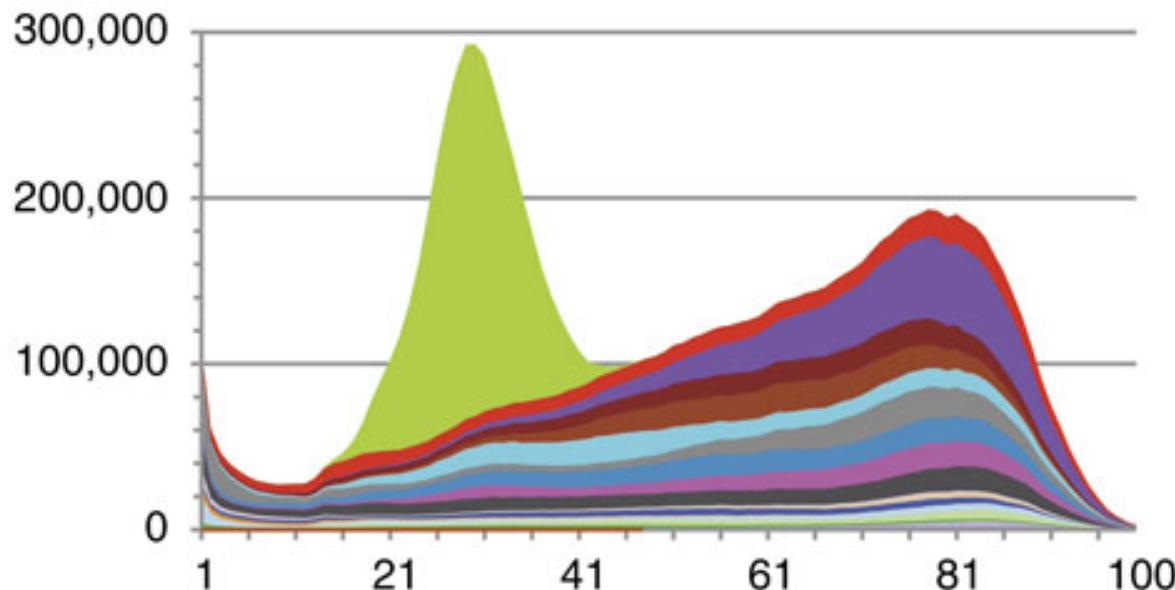
L'analyse des données a permis d'établir, entre autres :

- 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.**

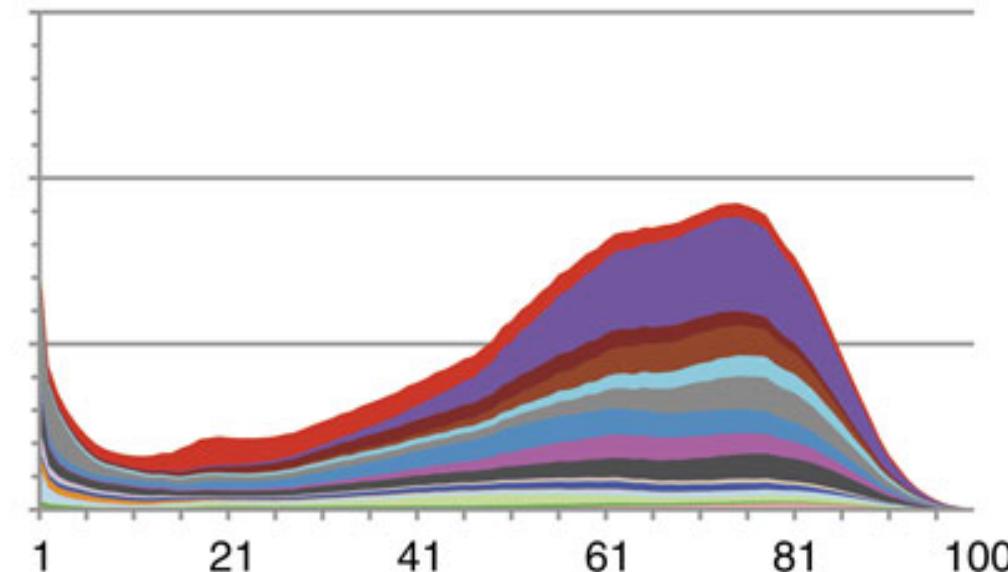


Diagnosis count
Inpatient

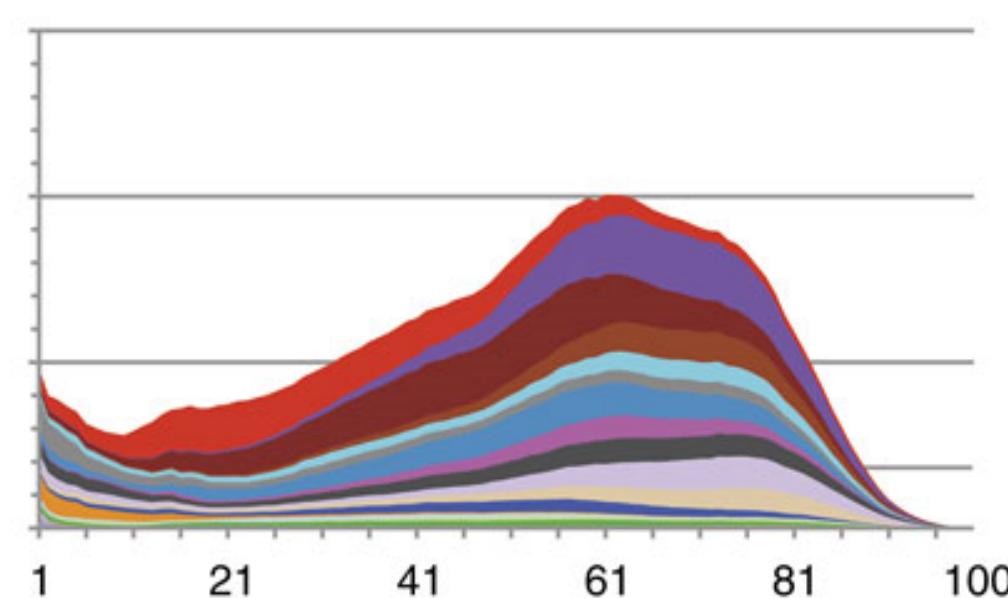
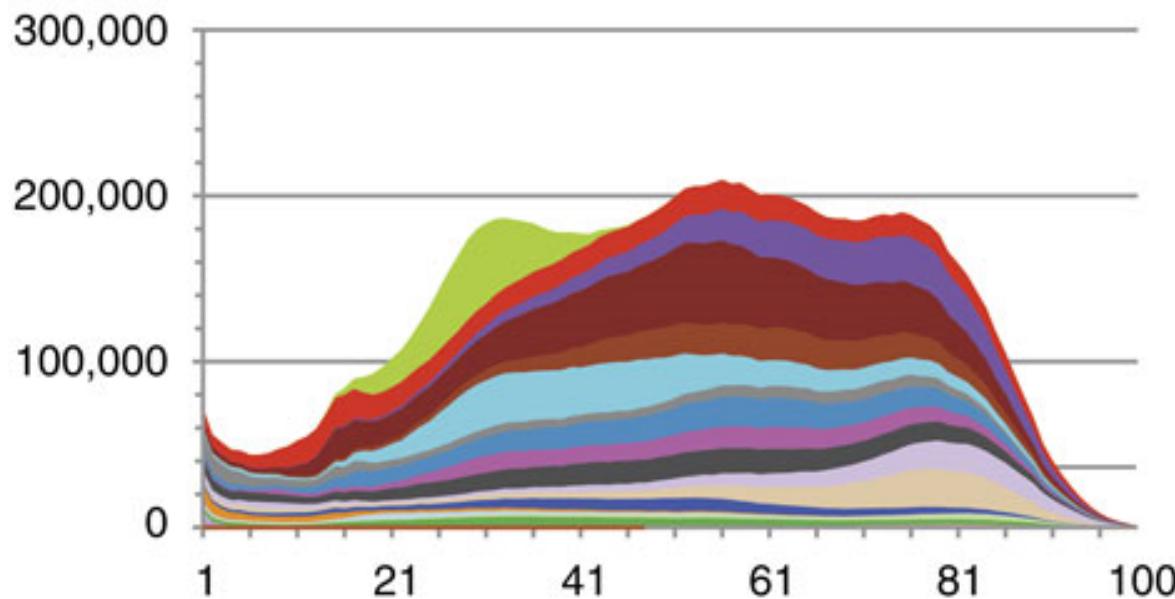
Female

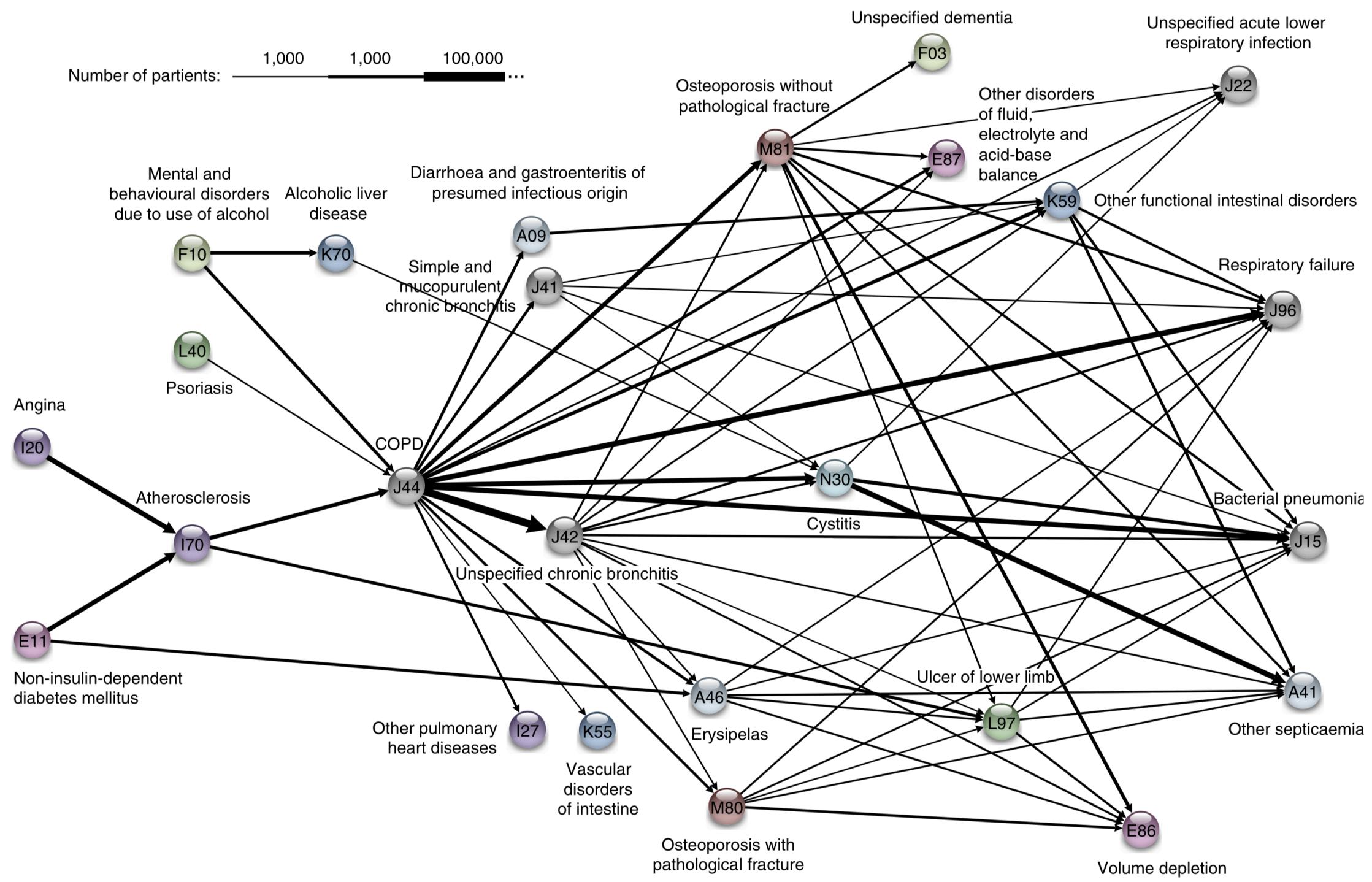


Male

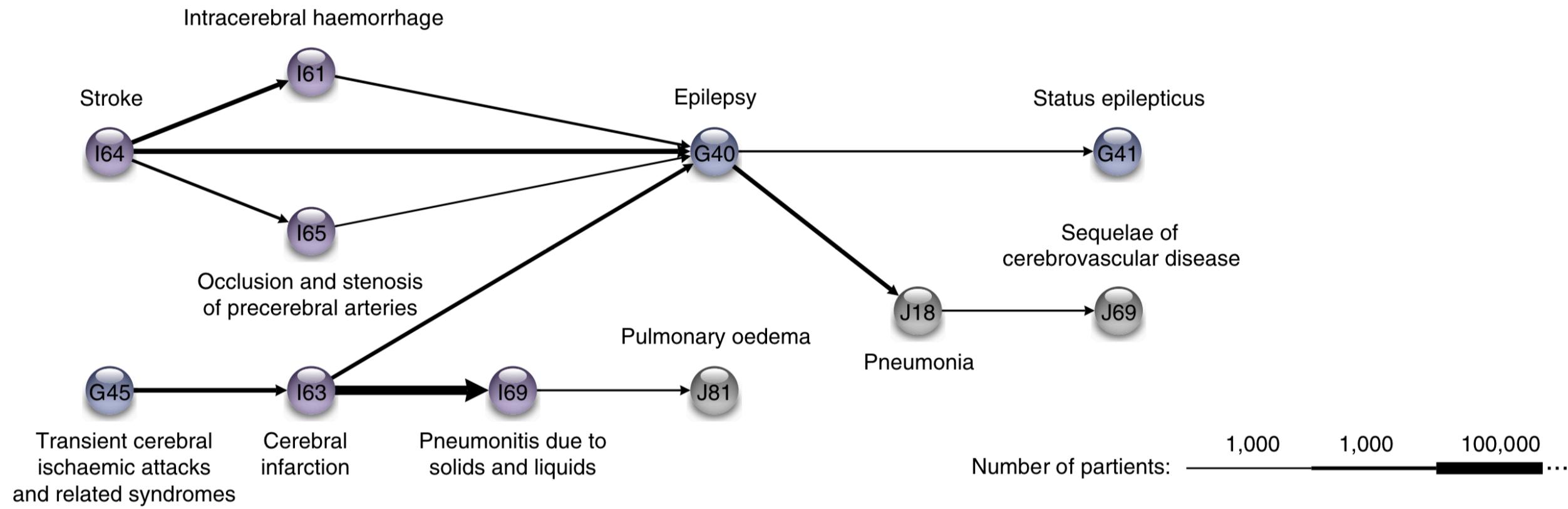


Diagnosis count
Outpatient





RÉSULTATS



À RETENIR

Les données permettent d'**étudier les maladies dans un contexte plus large**.

La recherche pourrait avoir des **effets bénéfiques tangibles sur la santé** à mesure que nous dépasserons le stade de la médecine universelle.

Le fait d'identifier tôt un modèle de risque pour la santé nous permettra de mieux **prévenir et traiter les maladies graves**.

Au lieu d'examiner chaque maladie de façon isolée, il est possible de le voir comme un système complexe avec de nombreux facteurs d'interaction différents.

L'ordre d'apparition des différentes maladies peut aider à trouver des **tendances** et des **corrélations complexes** indiquant la direction à prendre pour chaque personne.

DISCUSSION

Cette recherche est-elle applicable au contexte canadien? Au contexte chinois?

Selon vous, quels étaient certains des défis techniques?

NOTIONS DE BASE SUR LES RÈGLES D'ASSOCIATION

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

« La corrélation n'est pas la causalité. Mais c'est un gros indice. »

(E. Tufte)

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 triathloniens* (**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

AUTRES UTILISATIONS

Établissement de prévisions et prise de décisions en fonction de ces règles

Modification des circonstances ou de l'environnement pour tirer parti de ces corrélations

Utilisation des liens pour modifier la probabilité de certains résultats

Imputation des données manquantes

Remplissage automatique et correction automatique du texte

DISCUSSION

Quelles sont certaines applications des règles d'association en matière de politique publique/gouvernementale?

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

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

UN EXEMPLE SIMPLE

Les quatre mesures sont :

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

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.

EXERCICE

Évaluez les règles suivantes destinées aux candidats :

- Si un particulier possède un album de musique classique (W), il possède également un album de hip-hop (Z), étant donné que

$$\text{Freq}(W) = 2010, \text{Freq}(Z) = 6855, \text{Freq}(W \cap Z) = 132$$

- Si un individu possède à la fois un album des Beatles et un album de musique classique, il est né avant 1976, étant donné que $\text{Freq}(Y \cap W) = 1852, \text{Freq}(Y \cap W \cap X) = 1778$

Sur les trois règles qui ont été établies ($X \rightarrow Y, W \rightarrow Z, Y \& W \rightarrow X$), laquelle vous semble la plus utile? Qu'est-ce qui est le plus surprenant?

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

EXERCICE

Un magasin qui vend des accessoires pour téléphones cellulaires fait une promotion sur les écrans de protection.

Les clients qui achètent plusieurs écrans de protection parmi un choix de six couleurs différentes bénéficient d'un rabais. Les directeurs de magasin, qui aimeraient savoir quelles couleurs d'écrans protecteurs sont susceptibles d'être achetées ensemble, ont rassemblé les transactions passées dans `Transactions.csv`.

Tenez compte des règles suivantes :

- {rouge, blanc} \Rightarrow {vert}
- {vert} \Rightarrow {blanc}
- {rouge, vert} \Rightarrow {blanc}
- [vert] \Rightarrow {rouge}
- {orange} \Rightarrow {rouge}
- [blanc, noir] \Rightarrow {jaune}
- [noir] \Rightarrow {vert}

EXERCICE

Pour chaque règle, calculer le **support**, la **confiance**, l'**intérêt** et le **lift**.

Parmi les règles pour lesquelles le support est positif (> 0), laquelle a le lift le plus élevé? La confiance? L'intérêt?

Déterminer cinq à dix autres règles et les évaluer. Lesquelles de ces 12 à 17 règles seraient les plus utiles pour les directeurs de magasin, selon vous?

Comment déterminer des valeurs seuil raisonnables pour le support, la couverture, l'intérêt et le lift des règles établies à partir d'un certain ensemble de données?

NOTES ET VALIDATION

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

« Rappelez-vous que tous les modèles sont faux; la question pratique est de savoir dans quelle mesure ils doivent être faux avant de ne plus être utiles. »

(Box, G.E.P., et Draper, N. R., *Empirical Model Building and Response Surfaces*)

NOMBRE DE RÈGLES

Considérons 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

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

M. SNIFF : Qu'est-ce que tu cherches?

M. SNOOP : Un billet de cinq dollars.

M. SNIFF : Tu es sûr que tu l'as perdu dans cette rue?

M. SNOOP : Oh non! Je l'ai perdu dans le bloc d'à côté, mais je cherche ici parce que la lumière est meilleure.

(Boys' Life Magazine, 1932)

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

Élage 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.

EXEMPLE : TITANIC

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

« Oh, ils ont construit le bateau Titanic pour naviguer sur l'océan bleu;
Et ils pensaient qu'ils avaient un bateau que l'eau ne traverserait pas;
Mais la toute-puissante main du Destin savait que le vaisseau ne tiendrait pas.
C'était triste quand ce grand vaisseau a coulé. »

(The Titanic Disaster, chanson traditionnelle)

ENSEMBLE DE DONNÉES SUR LE TITANIC

Compilé par Robert Dawson en 1995; il se compose de quatre attributs catégoriques pour chacune des 2 201 personnes à bord du Titanic lors de son naufrage en 1912.

Les attributs sont :

- **classe** (première classe, deuxième classe, troisième classe, membre d'équipage)
- **âge** (adulte, enfant)
- **sexé** (homme, femme)
- **survie** (oui, non)

ENSEMBLE DE DONNÉES SUR LE TITANIC

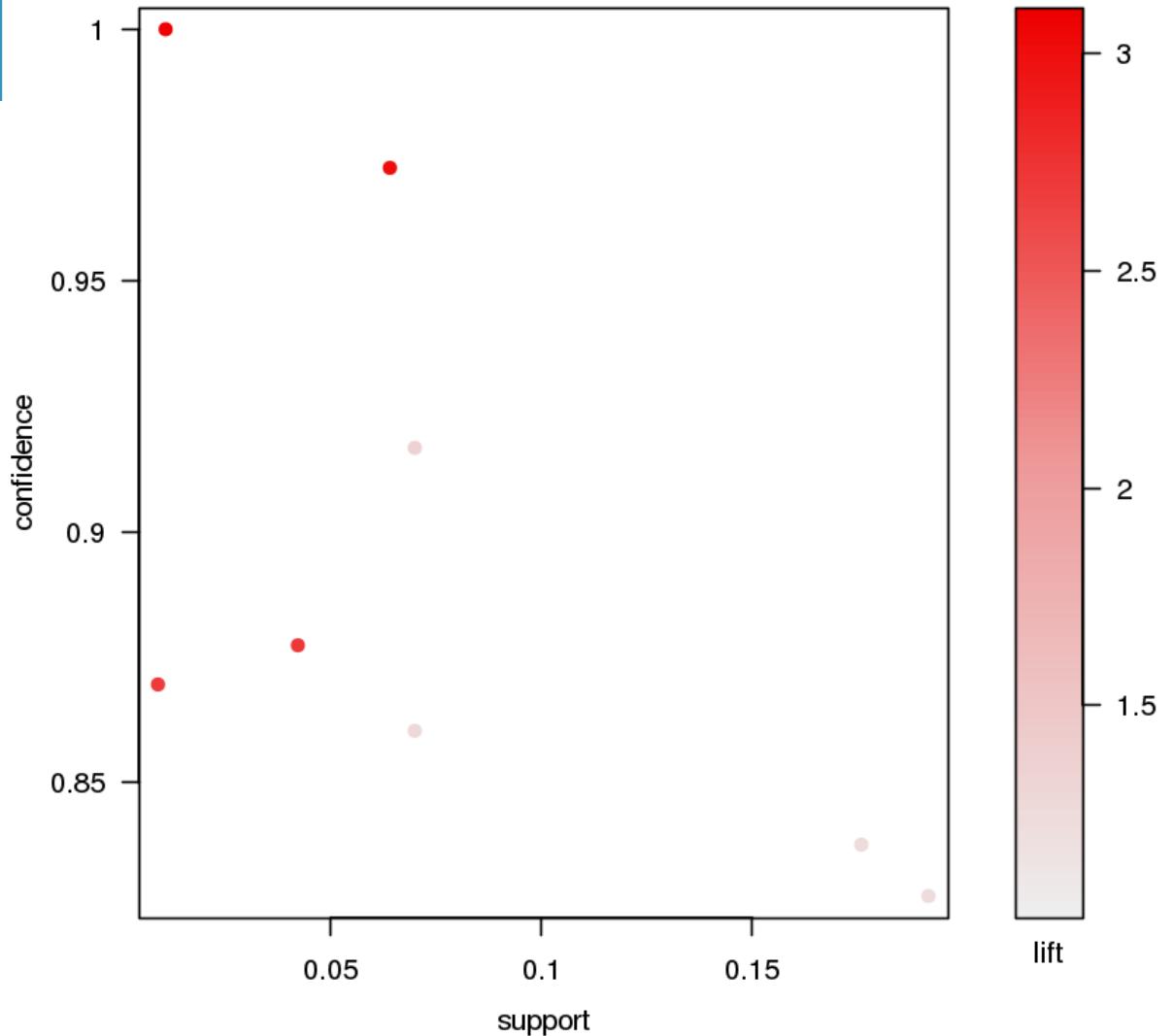
La question naturelle pour cet ensemble de données est de savoir comment la survie est liée aux autres attributs.

Nous utilisons la mise en œuvre des règles d'*Apriori* dans R pour produire et personnaliser les règles des candidats, pour finalement obtenir **huit règles**.

S'agit-il d'une tâche supervisée ou non supervisée?

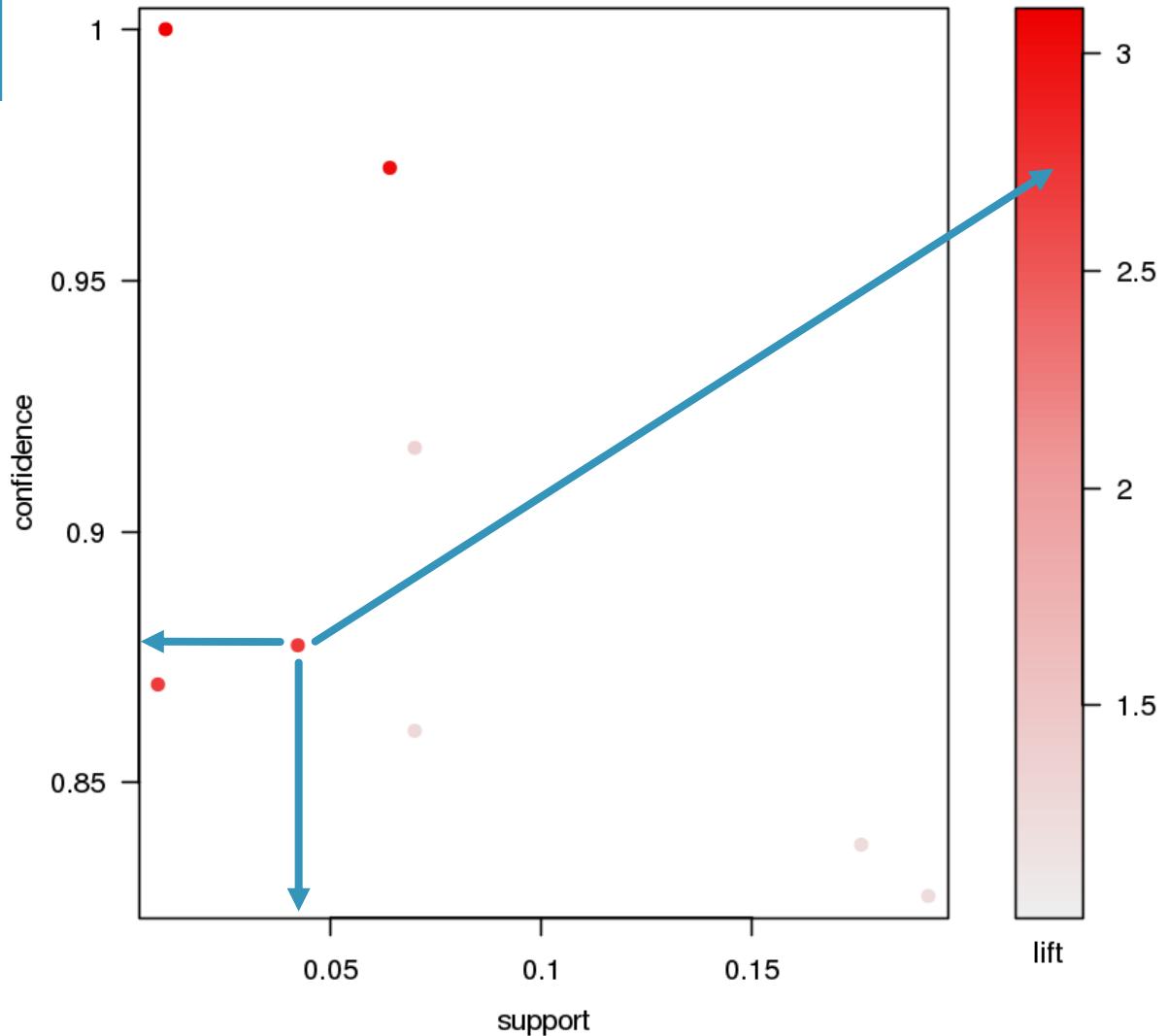
ENSEMBLE DE DONNÉES SUR LE TITANIC

Règle	Support	Confiance	Lift
SI classe = 2e ET âge = Enfant ALORS survécu = Oui	0,01	1	3,10
Si classe = 1er ET sexe = Femme ALORS survécu = Oui	0,06	0,97	3,01
Si classe = 2e ET sexe = Femme ALORS survécu = Oui	0,04	0,88	2,72
Si classe = Équipage ET sexe = Femme ALORS survécu = Oui	0,00	0,87	2,70
SI classe = 2e ET sexe = Homme ET âge = Adulte ALORS survécu = Non	0,07	0,92	1,35
SI classe = 2e ET sexe = Homme ALORS survécu = Non	0,07	0,86	1,27
SI classe = 3e ET sexe = Homme ET âge = Adulte ALORS survécu = Non	0,18	0,84	1,24
SI classe = 3e ET sexe = Homme ALORS survécu = Non	0,19	0,83	1,22

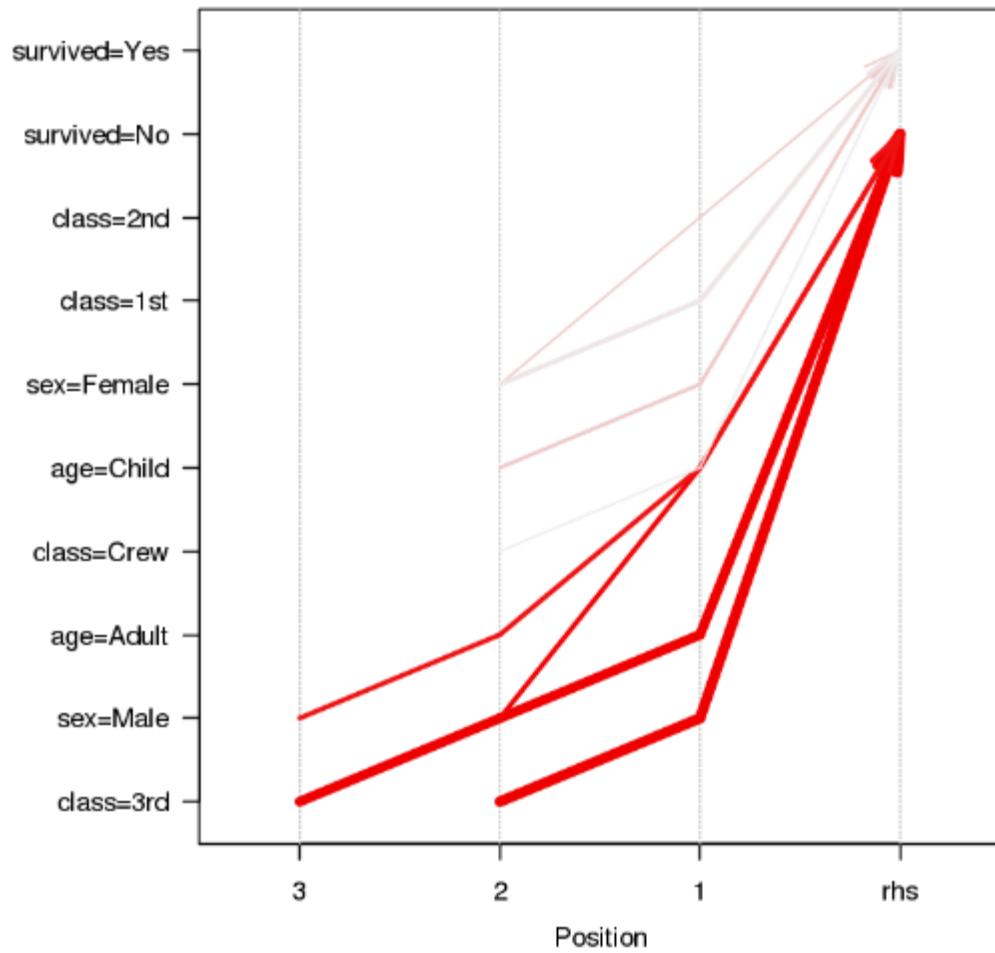
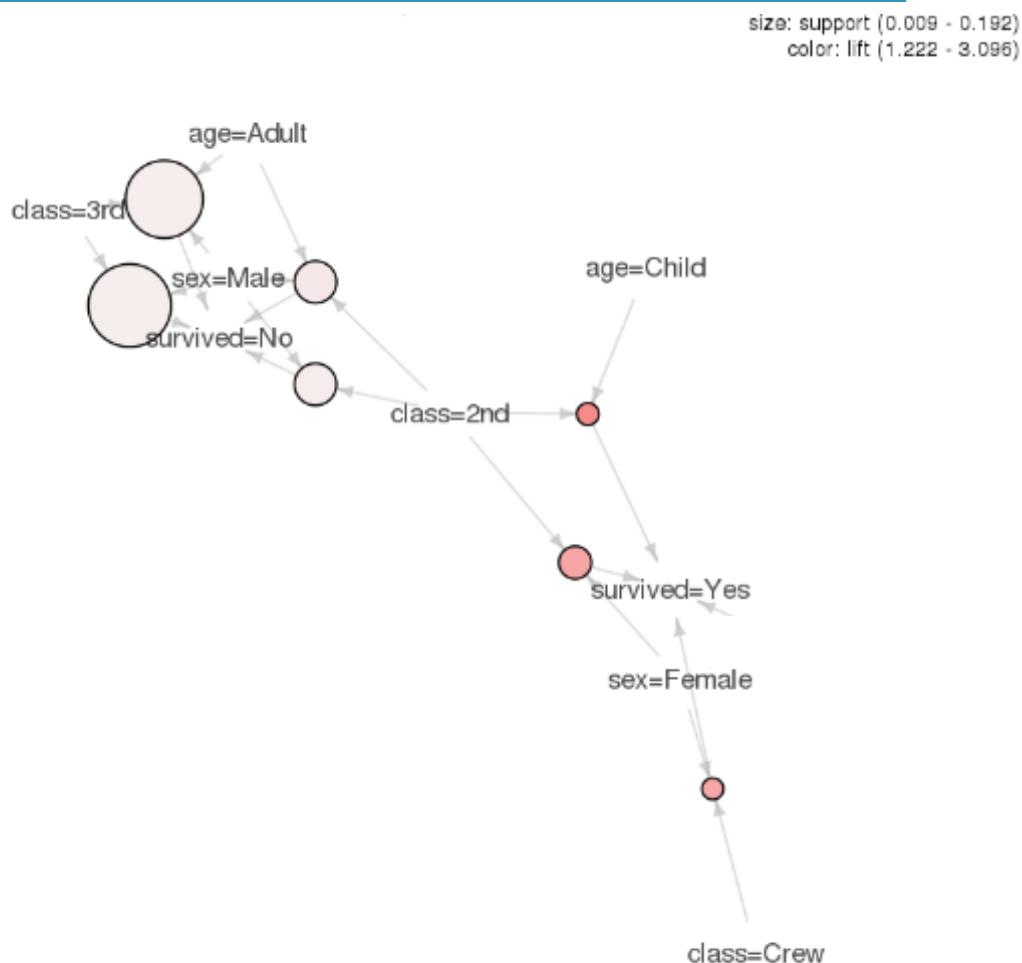


ENSEMBLE DE DONNÉES SUR LE TITANIC

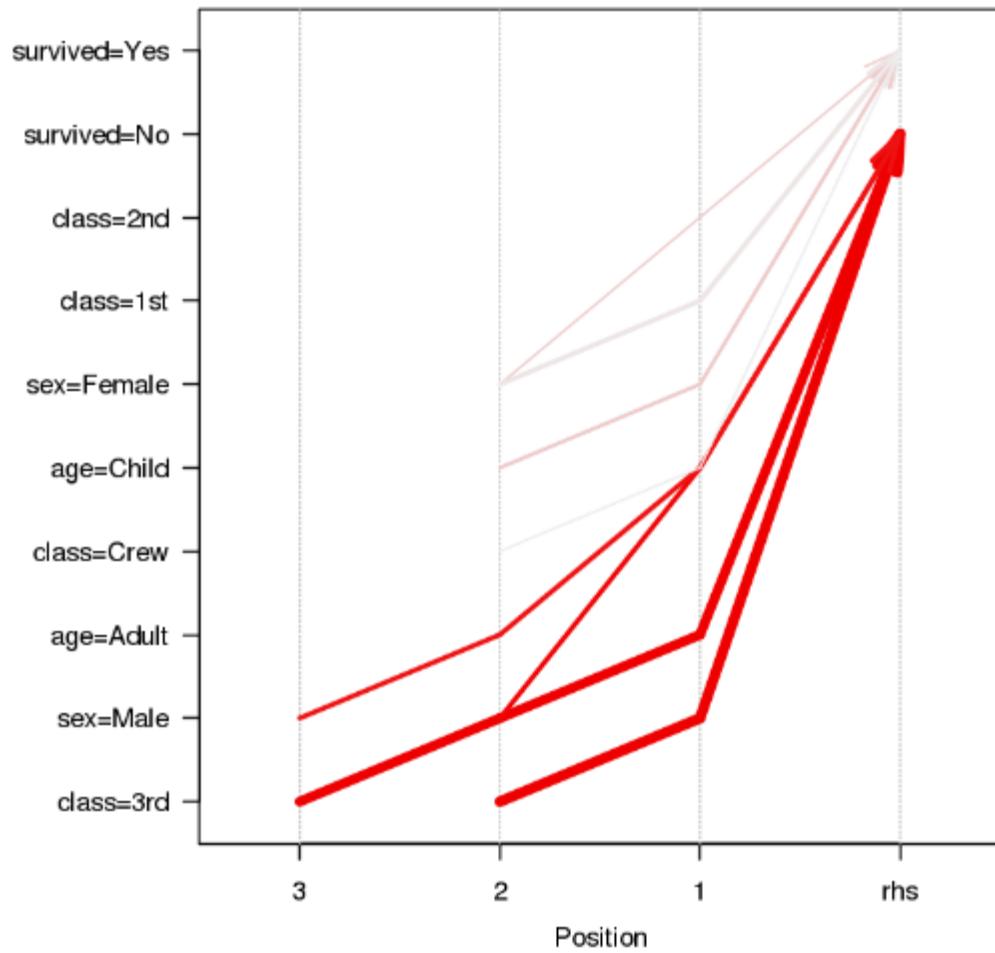
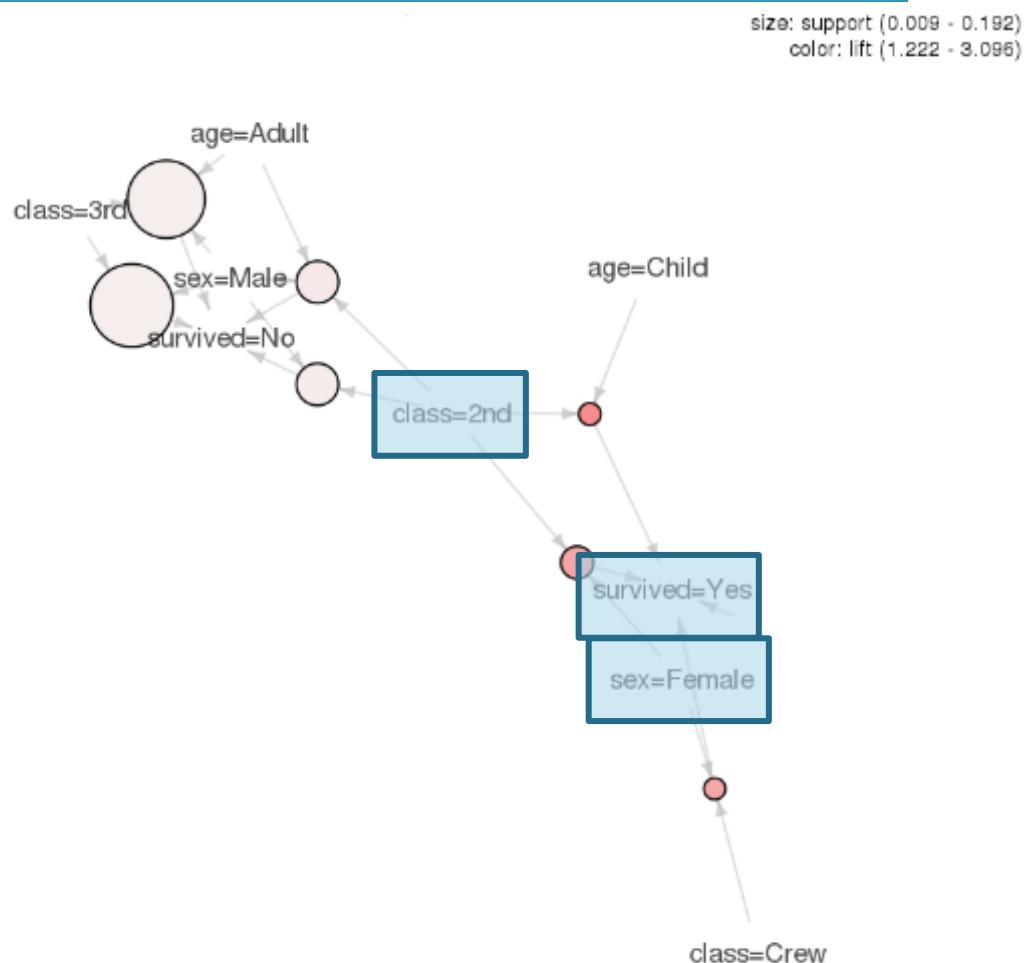
Règle	Confiance		
	Support	Confiance	Lift
SI classe = 2e ET âge = Enfant ALORS survécu = Oui	0,01	1	3,10
Si classe = 1er ET sexe = Femme ALORS survécu = Oui	0,06	0,97	3,01
Si classe = 2e ET sexe = Femme ALORS survécu = Oui	0,04	0,88	2,72
Si classe = Équipage ET sexe = Femme ALORS survécu = Oui	0,00	0,87	2,70
SI classe = 2e ET sexe = Homme ET âge = Adulte ALORS survécu = Non	0,07	0,92	1,35
SI classe = 2e ET sexe = Homme ALORS survécu = Non	0,07	0,86	1,27
SI classe = 3e ET sexe = Homme ET âge = Adulte ALORS survécu = Non	0,18	0,84	1,24
SI classe = 3e ET sexe = Homme ALORS survécu = Non	0,19	0,83	1,22



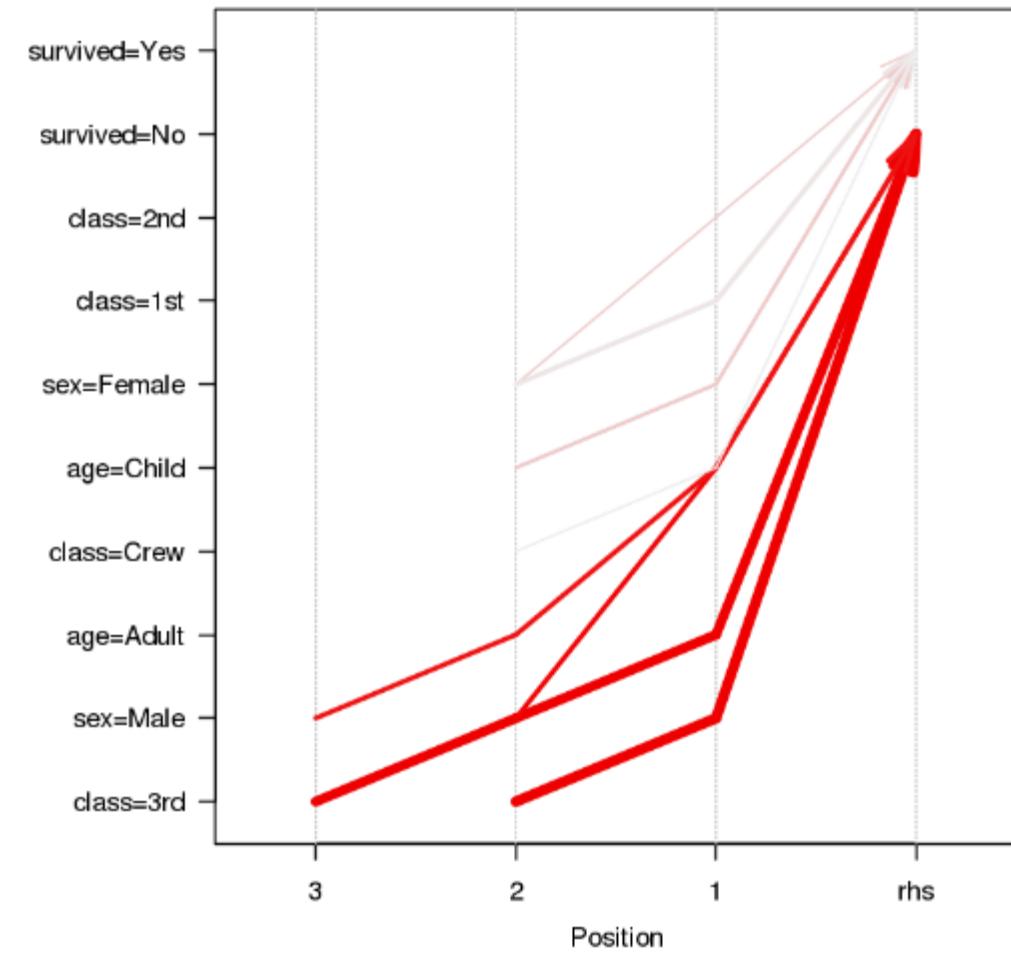
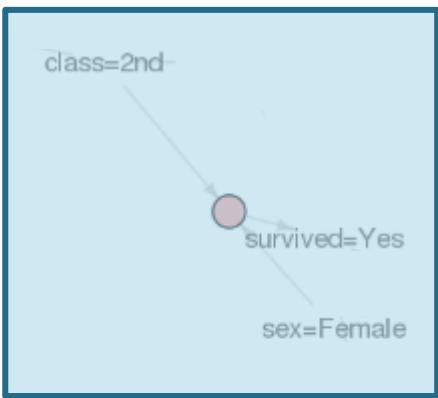
ENSEMBLE DE DONNÉES SUR LE TITANIC



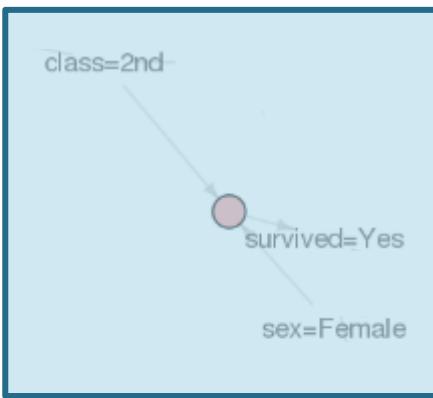
ENSEMBLE DE DONNÉES SUR LE TITANIC



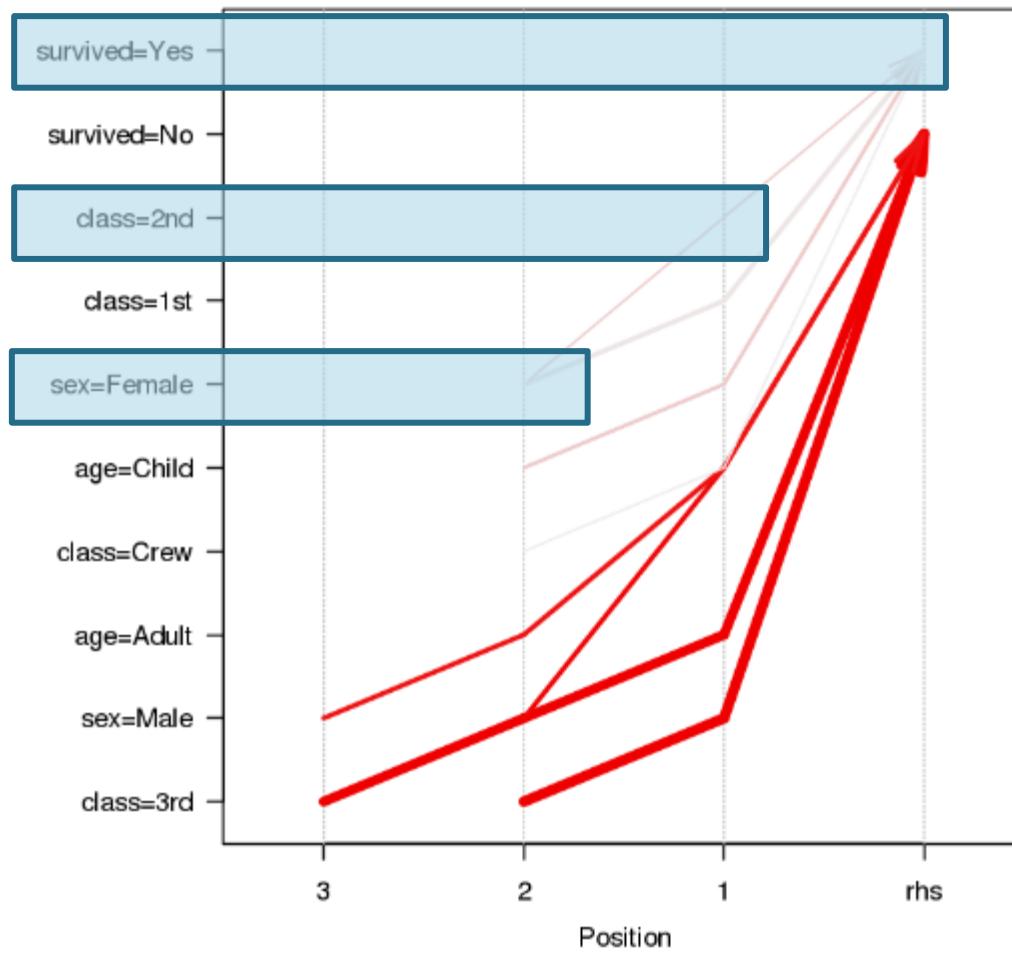
ENSEMBLE DE DONNÉES SUR LE TITANIC



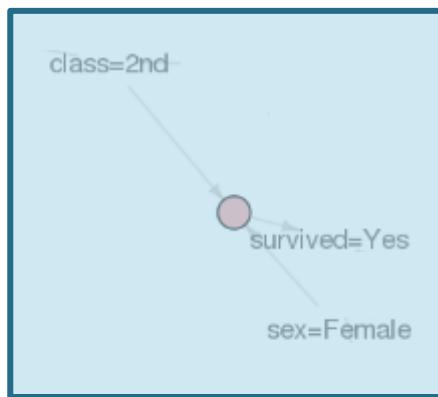
ENSEMBLE DE DONNÉES SUR LE TITANIC



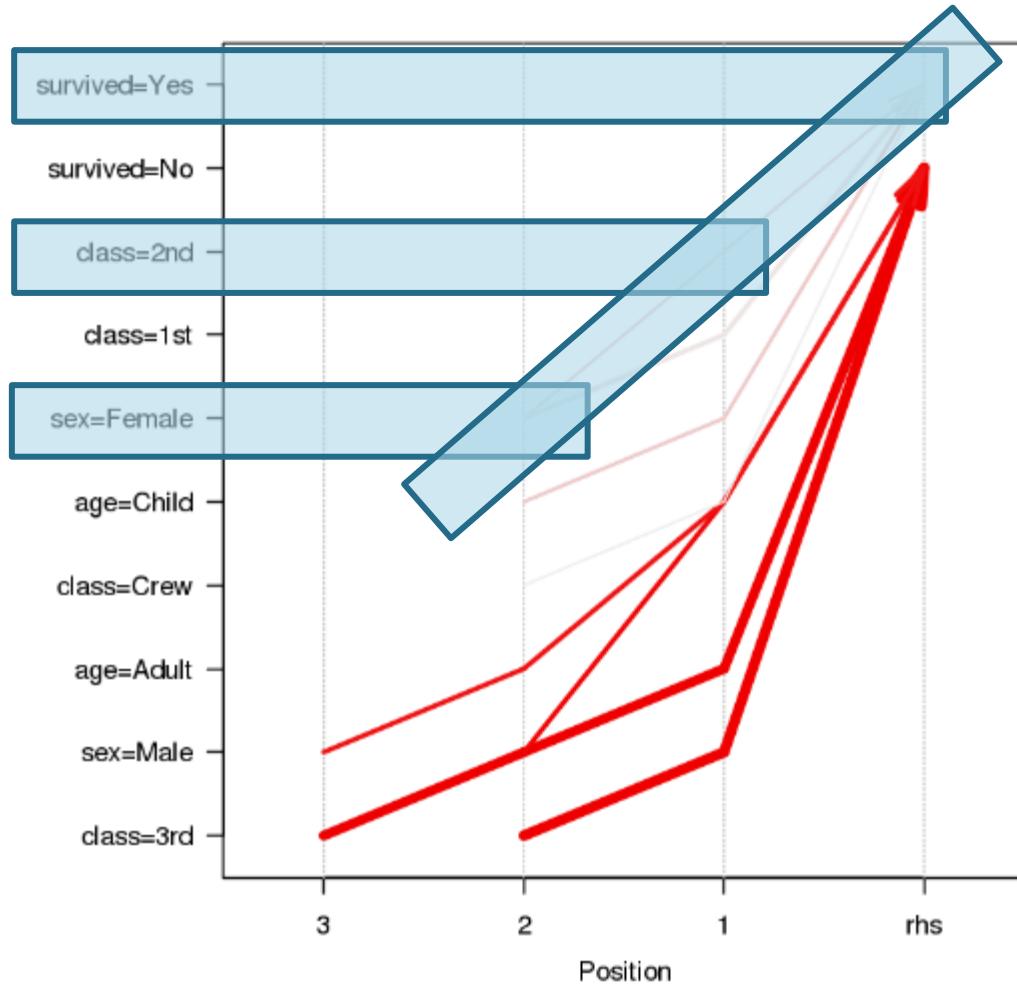
size: support (0.009 - 0.192)
color: lift (1.222 - 3.096)



ENSEMBLE DE DONNÉES SUR LE TITANIC



size: support (0.009 - 0.192)
color: lift (1.222 - 3.096)



EXERCICE

Effectuer une analyse semblable pour obtenir les règles d'association relatives aux ensembles de données *Life in L.A.* et *Transactions*.

RÉFÉRENCES

APPRENTISSAGE STATISTIQUE ET EXPLORATION DES RÈGLES D'ASSOCIATION

RÉFÉRENCES

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., Moser, S.A. (1998), Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance, Journal of American Medical Informatics Association, vol. 5, no 4, p.373-381

Garcia, E., Romero, C., Ventura, S., Calders, T. (2007), Drawbacks and solutions of applying association rule mining in learning management systems, Proceedings of the International Workshop on Applying Data Mining in e-Learning, 2007.

Boily, P., Schellinck, J., Hagiwara, S. [2019], *Introduction to Quantitative Consulting*, Data Action Lab.

Aggarwal, C.C., [2015], *Data Mining: The Textbook*, Springer.

Aggarwal, C.C., Han, J. (eds.) [2014], *Frequent Pattern Mining*, Springer.

RÉFÉRENCES

<http://www.rdatamining.com/examples/association-rules>

<https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>

<https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf>

<https://www.lynda.com/R-tutorials/Up-Running-R/120612-2.html>

http://michael.hahsler.net/research/arules_RUG_2015/demo/