



Bases de Données Avancées-L3

Des statistiques au k -anonymat

Jean-François COUCHOT

Université de Franche-Comté, UFR-ST



Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat



Big data et vie privée

De l'intérêt du big data

Protéger la vie privée ?

Aspects législatifs

Publication non sure de données

Un premier modèle de PVP : le k -anonymat



Big data et vie privée

De l'intérêt du big data

Protéger la vie privée ?

Aspects législatifs

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

PVP : Exacerbé par le Big Data



Big Data & Data mining

- L'exploration de données (Data Mining) : inférence de connaissances intéressantes à partir de grandes quantités de données (Big Data)
- Tendances générale : exploration des données en croissance // Big Data
- Analyse/extraction de connaissance : techniquement réalisable aujourd'hui

Big Data : application et volume

- Domaines : veille économique, découverte scientifique, santé, profilage
- Marché du Big Data en santé : $\approx 67,82$ G\$ en 2025 (Globe News Wire)
- 90% de toutes les données : créées au cours des deux dernières années (IBM)
- 97,2% des organisations : investissent en Big Data et IA. (New Vantage)
- Offres d'emploi dans le domaine : $\approx 2,7$ M en 2020. (Forbes)





Big data et vie privée

De l'intérêt du big data

Protéger la vie privée ?

Aspects législatifs

Publication non sure de données

Un premier modèle de PVP : le k -anonymat



Historique

- Expression du “droit d’être laissé tranquille” (the right to be let alone) ¹
- “Nul ne sera l’objet d’immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, [. . .]. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.” ²
- A l’heure d’Internet et des données (personnelles) transmises par : smartphones, messageries, GPS, appareils de fitness, moteur de recherche. . .

Des exemples d’inférences problématiques

- Réfrigérateur commandant des produits consommés : \rightsquigarrow nbre. de présents/absents au domicile, risque sanitaire ? assurance ?
- Application de suivi de la santé : positions, fréquences card. partagées avec Apple/Google seulement ?

1. Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.

2. Organisation des Nations Unies, (1949), Déclaration universelle des droits de l’homme.

3. <https://openclassrooms.com/fr/courses/5280946-protégez-les-donnees-personnelles>

Scandales de non protection de la vie privée

National Security Agency (NSA) PRISM (2007-2013)

- En 2007, création par la NSA du programme PRISM de surveillance des communications échangées sur les services en ligne des GAFAM notamment
- Transmission des données brutes à des pays tiers
- Espionnage de leaders politiques internationaux (dont A.Merkel)

Cambridge Analytica (CA) (2014-2018)

- "This is your digital life" : application pour Facebook qui permet l'aspiration de données personnelles présentes du réseau
- 87M comptes utilisateurs Facebook aspirés dès 2014
- Ciblage politique pour convaincre de voter pour D. Trump en 2016
- "Sans CA, il n'y aurait pas eu de Brexit" selon C. Wylie



Big data et vie privée

De l'intérêt du big data

Protéger la vie privée ?

Aspects législatifs

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Rôles des personnes accédant à la donnée - 1

Fournisseur de contenu

- Exemple d'une photo : les personnes sur la photo + photographe ev.

Autres entités connues du fournisseur avec lesquelles il souhaite la partager

- Exemple d'une photo : amis pour partager, fournisseur de service
- Moyens utilisés : application, mécanisme de groupes et de permissions
- Règlement Général sur la Protection des Données (RGPD) :
 - A. 6 : "Données minimales collectées et traitées de manière loyale et licite"
 - A. 12 : "Transparence des informations et des communications et modalités de l'exercice des droits de la personne concernée"
 - A. 16 : "Droit à la rectification"
 - A. 17 : "Droit à l'effacement, à l'oubli"
 - A. 20 : "Droit à la portabilité" : données récupérables (format lisible)

Rôles des personnes accédant à la donnée - 2

Gestionnaire de cette donnée

- Exemple d'une photo : employés de réseau social, d'entreprises de stockage dans le nuage
- Catégorie critique :
 - Modèle économique du service : extraction d'information
 - Administrateurs de BD : souvent honnête mais à risque
 - Chiffrement des données : souvent une solution

Reste du monde

- Exemple d'une photo : celles/ceux avec lequel on ne veut pas la partager



Big data et vie privée

Publication non sure de données

Bases de données statistiques : attaquables

Anonymisation par pseudonymisation : attaquable

Un premier modèle de PVP : le k -anonymat



Big data et vie privée

Publication non sure de données

Bases de données statistiques : **attaquables**

Anonymisation par pseudonymisation : **attaquable**

Un premier modèle de PVP : le *k*-anonymat



Motivations pieuse

- Permettre l'accès à des statistiques concernant des groupes d'individus
- Tout en restreignant l'accès aux informations individuelles

Pb. : des vestiges d'informations individuelles contenus dans chq. stat.

- Comparaison des salaires totaux de deux groupes qui ne diffèrent que d'un individu
- Possibilité de déduire le salaire de l'individu manquant à l'un

Les M attributs d'une relation (issue d'une jointure ev.)

- Chaque attribut A_j , $1 \leq j \leq M$: $|A_j| \geq 2$ valeurs possibles
- x_{ij} : valeur de l'attribut A_j pour l'individu i
- Hypothèse simplificatrice : un seul enregistrement par individu
- N enregistrements

Caractérisation de sous-groupe par une formule C

- C : formule en logique propositionnelle
- Opérateurs : “+” (OR), “.” (AND), “—” (NOT)
- Relation issue de C : ensemble des enregistrements qui vérifient C

4. Denning, D. E. R. (1982). Cryptography and data security (Vol. 112). Reading : Addison-Wesley.



La base

| Nom | Sexe | Dpt. | Année | Test | Note |
|-------|------|------|-------|------|------|
| Allen | F | Info | 00 | 600 | 3,4 |
| Baker | F | Ing | 00 | 520 | 2,5 |
| Cook | H | Ing | 98 | 630 | 3,5 |
| Davis | F | Info | 98 | 800 | 4,0 |
| Evans | H | Bio | 99 | 500 | 2,2 |
| Frank | H | Ing | 01 | 580 | 3,0 |
| Good | H | Info | 98 | 700 | 3,8 |
| Hall | F | Psy | 99 | 580 | 2,8 |
| Ilies | H | Info | 01 | 600 | 3,2 |
| Jones | F | Bio | 99 | 750 | 3,8 |
| Kline | F | Psy | 01 | 500 | 2,5 |
| Lane | H | Ing | 98 | 600 | 3,0 |
| Moore | H | Info | 99 | 650 | 3,5 |

- Identifiant : Nom
- Champs sensibles : Note, Test
- Exemple de sous groupe C :
(Sexe = H).((Dpt = Info) + (Dpt = Bio)), en abrégé en H .(Info + Bio)
- Relation issue de C :
 $\{(Evans, H, Bio, 99, 500, 2, 2),$
 $(Good, H, Info, 98, 700, 3, 8),$
 $(Ilies, H, Info, 01, 600, 3, 2),$
 $(Moore, H, Info, 99, 650, 3, 5)\}$



Statistiques simples

- Informations uniquement agrégées par
 - $\text{count}(C) = |C|$
 - $\text{sum}(C, A_j) = \sum_{i \in C} x_{ij}$
- Exemple : $\text{count}(H.(Info + Bio)) = |H.(Info + Bio)| = 4$
- Autres requêtes identifiantes interdites : $F.Info = \{Allen, Davis\}$

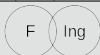
Statistiques sensibles

- Statistique déclarée comme sensible : si dévoile trop d'informations confidentielles à propos d'un individu
- Statistiques interdites : celles issues d'un singleton
- Exemple $F.Ing$ est le singleton $\{(Baker, F, Ing, 00, 520, 2.5)\}$
 $\text{sum}(F.Ing, Note)$ est sensible et sera interdite
- Idée : pour un seuil n choisi, n'autoriser que les requêtes sur des relations de cardinalité dans $[n, N - n]$,

Requêtes statistiques dans $[n, N - n]$: pas sûres

Tracker individuel, ordre 2

| Nom | Sexe | Dpt. | Année | Test | Note |
|-------|------|------|-------|------|------|
| Baker | F | Ing | 00 | 520 | 2,5 |
| Cook | H | Ing | 98 | 630 | 3,5 |
| Frank | H | Ing | 01 | 580 | 3,0 |
| Lane | H | Ing | 98 | 600 | 3,0 |



- $|F.Ing| = 1$: sensible !
- $|Ing| = 4$, $|\overline{F}.Ing| = 3$: OK
- $\text{sum}(F.Ing, Note) = \text{sum}(Ing, Note) - \text{sum}(\overline{F}.Ing, Note) = 12,0 - 9,5 = 2,5$

Tracker individuel, ordre 3

| Nom | Sexe | Dpt. | Année | Test | Note |
|-------|------|------|-------|------|------|
| Evans | H | Bio | 99 | 500 | 2,2 |
| Hall | F | Psy | 99 | 580 | 2,8 |
| Jones | F | Bio | 99 | 750 | 3,8 |
| Moore | H | Info | 99 | 650 | 3,5 |

| | H.Bio | 00 | 01 | 11 | 10 |
|----|-------|----|----|----|----|
| 99 | | | | | |
| 0 | | | | | |
| 1 | | | | X | |

- $q(H.Bio.99) + q((\overline{H} + \overline{Bio}).99) = q(99)$
- $q(H.Bio.99) = q(99) - q((\overline{H} + \overline{Bio}).99)$
- $\text{sum}(H.Bio.99, Note) = \text{sum}(99, Note) - \text{sum}((\overline{H} + \overline{Bio}).99, Note) = 12,3 - 10,1 = 2,2$

Inconvénient : nécessite de caractériser l'individu

Présentation agrégée : pas toujours sure



Macrostatistiques

- Collections de statistiques agrégées, usuellement présentées sous la forme de tableaux à double entrées
- Inconvénient : réduction extrême de l'utilité des données, difficulté d'établir des corrélations entre attributs. . .

Exemple de macrostatistiques

Effectifs par année/sexe

| Sexe | 98 | 99 | 00 | 01 | Total |
|-------|----|----|----|----|-------|
| F | 1 | 2 | 2 | 1 | 6 |
| H | 3 | 2 | 0 | 2 | 7 |
| Total | 4 | 4 | 2 | 3 | 13 |

Tests par année/sexe

| Sexe | 98 | 99 | 00 | 01 | Total |
|-------|------|------|------|------|-------|
| F | 800 | 1330 | 1120 | 500 | 3750 |
| H | 30 | 1150 | 0 | 1120 | 4260 |
| Total | 2730 | 2480 | 0 | 1180 | 8010 |

Suppressions

Tests par année/sexe

| Sexe | 98 | 99 | 00 | 01 | Total |
|-------|------|------|------|------|-------|
| F | * | 1330 | 1120 | * | 3750 |
| H | 1930 | 1150 | 0 | 1120 | 4260 |
| Total | 2730 | 2480 | 0 | 1180 | 8010 |

Tests par année/sexe

| Sexe | 98 | 99 | 00 | 01 | Total |
|-------|------|------|------|------|-------|
| F | * | 1330 | 1120 | * | 3750 |
| H | * | 1150 | 0 | * | 4260 |
| Total | 2730 | 2480 | 0 | 1180 | 8010 |

Pas de garantie de sûreté.

Arrondi systématique de valeurs : pas sûr



Arrondi systématique de q dans une base b

- Soit $b' = \lfloor \frac{b+1}{2} \rfloor$ et $d \equiv q \pmod{b}$
- $$r(q) = \begin{cases} q & \text{si } d = 0 \\ q - d & \text{si } d < b' \\ q + (b - d) & \text{si } d \geq b' \end{cases}$$
- $r(q) \in [r(q) - b' + 1, r(q) + b' - 1]$

| | $r(q_i)$ | l_i | u_i |
|--------|----------|----------|----------|
| q_1 | 15 | 13 | 17 |
| q_2 | 10 | 8 | 12 |
| q_3 | 15 | 13 | 17 |
| q_4 | 20 | 18 | 22 |
| \sum | | 52 | 68 |
| Q | 70 | $L = 68$ | $U = 72$ |

Cas d'arrondis réversibles pour un attribut A

- Ensembles disjoints C_1, \dots, C_m et $q_i = \text{sum}(C_i, A)$, $1 \leq i \leq m$,
- $Q = q_1 + \dots + q_m$ le total de toutes les sommes
- $[l_i, u_i]$ l'intervalle estimant $r(q_i)$, $[L, U]$ l'intervalle estimant $r(Q)$
- Si $L = \sum_i u_i \rightarrow \forall i, 1 \leq i \leq m, q_i = u_i$
- Si $U = \sum_i l_i \rightarrow \forall i, 1 \leq i \leq m, q_i = l_i$



Big data et vie privée

Publication non sure de données

Bases de données statistiques : attaquables

Anonymisation par pseudonymisation : attaquable

Un premier modèle de PVP : le k -anonymat

Pseudonymisation



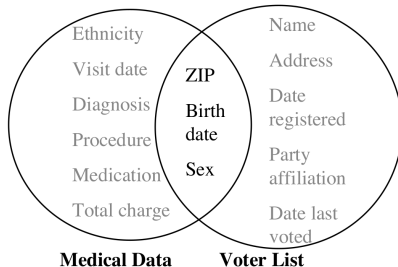
- Champs identifiants : supprimés et remplacés par un id (H(NSS)).

| | Non-sensibles | | | | Sensibles |
|----|---------------|-----|-------|-------------|-------------------|
| H | CP | Age | Genre | Nationalité | Pathologie |
| 1 | 13053 | 28 | H | russe | trouble cardiaque |
| 2 | 13068 | 29 | H | américaine | trouble cardiaque |
| 3 | 13068 | 21 | F | japonaise | infection virale |
| 4 | 13053 | 23 | H | américaine | infection virale |
| 5 | 14853 | 49 | H | indienne | cancer |
| 6 | 14853 | 48 | F | russe | trouble cardiaque |
| 7 | 14850 | 47 | H | américaine | infection virale |
| 8 | 14850 | 49 | F | américaine | infection virale |
| 9 | 13053 | 31 | H | américaine | cancer |
| 10 | 13053 | 37 | H | indienne | cancer |
| 11 | 13068 | 36 | F | japonaise | cancer |
| 12 | 13068 | 35 | F | américaine | cancer |

- Avantage : calculs identiques à ceux sur la base de données initiale (Age moyen/cancer=37,8)

Pseudonymisation : attaque par quasi-identifiants⁵

- Base de donnée médicale pseudonymisée et publique



- Liste d'électeurs publique, recensement USA, 1990 : "87% of the population in the US had characteristics that likely made them unique based only on 5-digit Zip, gender, date of birth"
- CP, genre, date de naissance : quasi-identifiants
- Identification de données médicales du gouverneur Weld

5. Sweeney, L. (2002). k-anonymity : A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Pseudonymisation : attaque par intersection ⁶

2006, diffusion par AOL de 20M requêtes, 658K utilisateur sans nom

| AnonID | Query | QueryTime |
|--------|---|------------|
| 1326 | "holiday mansion houseboat" | 2006-03-29 |
| 1326 | "back to the future" | 2006-04-01 |
| 591476 | "english spanish translator" | 2006-03-20 |
| 591476 | "panama vacations" | 2006-03-20 |
| 591476 | "breast reduction" | 2006-03-23 |
| 591476 | "volunteer work at hospitals in brooklyn" | 2006-05-24 |
| 591476 | ... | ... |
| 591476 | "how to secretly poison your ex" | 2006-03-12 |

Thelma Arnold, 62 ans, veuve vivant à Lilburn, Ga., réidentifiée en 3 j.

| AnonID | Query |
|---------|-----------------------------------|
| 4417749 | "people with last name 'Arnold'" |
| 4417749 | "landscapers in Lilburn, Ga" |
| 4417749 | "60 single men" |
| 4417749 | "dog that urinates on everything" |
| 4417749 | dog-related queries |



↪ Suppression hâtive des données sur le site d'AOL.

6. BARBARO, Michael, ZELLER, Tom, et HANSELL, Saul. A face is exposed for AOL searcher no. 4417749. New York Times, 2006, vol. 9, no 2008, p. 8.



Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Un algorithme pour atteindre le k -anonymat

Mesures d'utilité

Attaques du k -anonymat



Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Un algorithme pour atteindre le k -anonymat

Mesures d'utilité

Attaques du k -anonymat

Le k -anonymat⁵ et les quasi-identifiants -1

Quasi-identifiants : QID

- QID, intuition⁷ : “des éléments qui ne sont pas en eux-mêmes des identificateurs uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants afin de créer un identificateur unique”
- QID, définition⁸ : Les attributs de $Q \subseteq \{A_1, \dots, A_M\}$ sont quasi-identifiants de la relation T si la requête suivante retourne au moins un résultat

```
SELECT Q FROM T
GROUP BY Q
HAVING COUNT(*)=1
```

- (CP, genre, date de naissance) : triplets uniques dans 87% des cas \rightsquigarrow quasi-identifiants

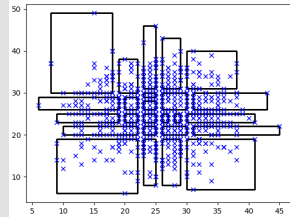
7. <https://en.wikipedia.org/wiki/Quasi-identifiant>

8. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Le k -anonymat⁵ et les quasi-identifiants -2

Intuition : regrouper les QID pour casser l'unicité

- Niveau de détail des valeurs des QID : à réduire pour qu'il y ait au moins k individus différents dont les QIDs sont égaux
- Individus avec mêmes QIDs : font partie de la même classe d'équivalence





Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Un algorithme pour atteindre le k -anonymat

Mesures d'utilité

Attaques du k -anonymat

k-anonymat par généralisation



Réduction des niveaux de détail

Sur l'exemple :

- CP : laisser, **regroupements par 2**, **suppr.**
- Age : laisser, par intervalles d'amplitudes 10, 20, **suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**
- $\rightsquigarrow 3 \times 4 \times 2 \times 3 = 72$ combinaisons de généralisation !

Regroupement par classes d'équivalence de card. \geq à 4

| CP | Age | Genre | Nationalité | Pathologie | |
|------------------|----------|-------|-------------|-------------------|---------------|
| {13053 13058} | [21; 31[| * | * | trouble cardiaque | } 4 individus |
| | [21; 31[| * | * | trouble cardiaque | |
| | [21; 31[| * | * | infection virale | |
| | [21; 31[| * | * | infection virale | |
| {14850 14853} | [41; 50[| * | * | cancer | } 4 individus |
| | [41; 50[| * | * | trouble cardiaque | |
| | [41; 50[| * | * | infection virale | |
| | [41; 50[| * | * | infection virale | |
| {13053 13058} | [31; 41[| * | * | cancer | } 4 individus |
| | [31; 41[| * | * | cancer | |
| | [31; 41[| * | * | cancer | |
| | [31; 41[| * | * | cancer | |



Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Un algorithme pour atteindre le k -anonymat

Mesures d'utilité

Attaques du k -anonymat

C_{AVG} : nbre. moy. d'éléments par classe normalisé

Définition pour $|T|$ enregistrements et propriétés

$$C_{AVG} = \frac{|T|}{|EQs|} \times \frac{1}{k}$$

- et $|EQs|$: nbre. de classes d'équiv.
- $\frac{|T|}{|EQs|}$ nbre. moy. d'éléments par classe ($\geq k$) $\rightsquigarrow C_{AVG} \geq 1$
- Utilité décroissante à mesure que C_{AVG} croît.

Exemple avec 4-anonymat

| CP | Age | Genre | Nationalité | Pathologie |
|------------------|----------|-------|-------------|-------------------|
| {13053 13058} | [21; 31[| * | * | trouble cardiaque |
| | [21; 31[| * | * | trouble cardiaque |
| | [21; 31[| * | * | infection virale |
| | [21; 31[| * | * | infection virale |
| {14850 14853} | [41; 50[| * | * | cancer |
| | [41; 50[| * | * | trouble cardiaque |
| | [41; 50[| * | * | infection virale |
| | [41; 50[| * | * | infection virale |
| {13053 13058} | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |

- $C_{AVG} = \frac{12}{3} \times \frac{1}{4} = 1$
- Optimal pour cette métrique

Loss : Perte due à la généralisation



Définition pour $|T|$ enregistrements, n QID et propriétés

$$Loss = \frac{1}{n|T|} \sum_{j=1}^{|T|} \sum_{i=1}^n \frac{R_{ij}}{R_i}$$

- R_{ij}/R_i : rapport acquis/acquérables
 - discret : $(|généralisation\ i_j| - 1) / (|Qid\ i| - 1)$
 - continu : $(U_{ij} - L_{ij}) / (U_i - L_i)$
- Moyenne des acquis/acquérables, utilité décroissante // Loss croît

Exemple avec 4-anonymat

| CP | Age | Genre | Nationalité | Pathologie |
|------------------|----------|-------|-------------|-------------------|
| {13053 13058} | [21; 31[| * | * | trouble cardiaque |
| | [21; 31[| * | * | trouble cardiaque |
| | [21; 31[| * | * | infection virale |
| | [21; 31[| * | * | infection virale |
| {14850 14853} | [41; 50[| * | * | cancer |
| | [41; 50[| * | * | trouble cardiaque |
| | [41; 50[| * | * | infection virale |
| | [41; 50[| * | * | infection virale |
| {13053 13058} | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |
| | [31; 41[| * | * | cancer |

- $Nat. = \{\text{russe, am., jap., ind.}\} \rightsquigarrow$
 $R_{Nat} = 3, R_{CP} = 3, R_{Genre} = 1,$
- $R_{Age} = 50 - 21 = 29$
- $Loss = \frac{1}{4 \times 12} (8 \times (\frac{1}{3} + \frac{10}{29} + \frac{1}{1} + \frac{3}{3})) + 4 \times (\frac{1}{3} + \frac{9}{29} + \frac{1}{1} + \frac{3}{3}) = \frac{2}{3}$



Big data et vie privée

Publication non sure de données

Un premier modèle de PVP : le k -anonymat

Introduction au k -anonymat

Un algorithme pour atteindre le k -anonymat

Mesures d'utilité

Attaques du k -anonymat

k-anonymat : attaques



Exemple

| CP | Age | Genre | Nationalité | Pathologie | |
|------------------|----------|-------|-------------|-------------------|---------------|
| {13053 13058} | [21; 31[| * | * | trouble cardiaque | } 4 individus |
| | [21; 31[| * | * | trouble cardiaque | |
| | [21; 31[| * | * | infection virale | |
| | [21; 31[| * | * | infection virale | |
| {14850 14853} | [41; 50[| * | * | cancer | } 4 individus |
| | [41; 50[| * | * | trouble cardiaque | |
| | [41; 50[| * | * | infection virale | |
| | [41; 50[| * | * | infection virale | |
| {13053 13058} | [31; 41[| * | * | cancer | } 4 individus |
| | [31; 41[| * | * | cancer | |
| | [31; 41[| * | * | cancer | |
| | [31; 41[| * | * | cancer | |

Attaques

- Homogénéité :
 - \oplus Patient de 35 ans connu \rightsquigarrow cancer.
 - \ominus Patient de 29 ans connu \rightsquigarrow ~~cancer~~.
- Connaissance supplémentaire : un japonais de 21 ans, $P(\text{trouble cardiaque}|\text{japonais})=\text{faible} \rightsquigarrow$ infection virale.