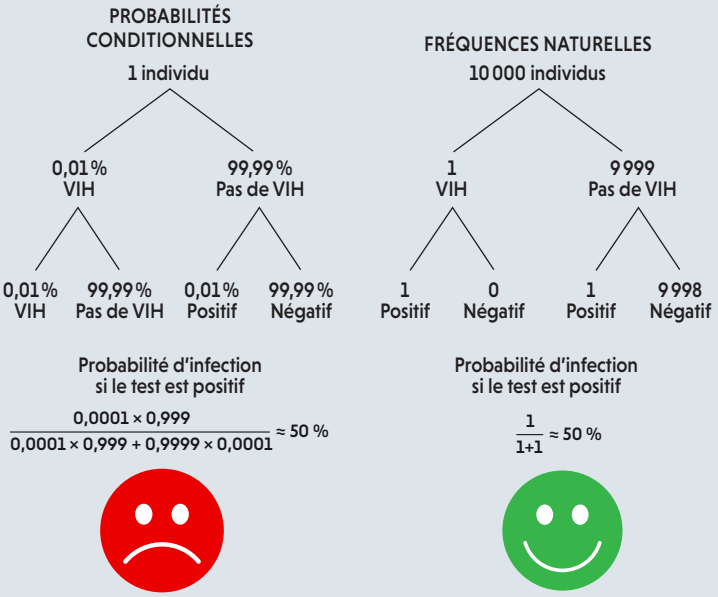
PRÉVOIR UNE INFECTION

Si votre test de dépistage du VIH est positif et que vous êtes un homme à faible risque d'infection (hétérosexuel, non drogué…), quel est le risque que vous soyez effectivement porteur du virus ? Les probabilités conditionnelles *(à gauche)* proposent un calcul compliqué. En revanche, en se fondant sur les fréquences naturelles *(à droite)*, on obtient facilement la réponse : sur 10 000 hommes, on s'attend à ce qu'un seul soit infecté et que son test soit donc positif ; parmi les 9 999 qui ne sont pas infectés, 1 devrait aussi avoir un résultat positif. En conséquence, on a deux tests positifs, mais un seul individu infecté. Donc la probabilité d'infection donnée par un test positif n'est pas de 100 %, mais de 50 %.



Les taux de mortalité sont des indicateurs plus fiables de la valeur des programmes de dépistage que les taux de survie à cinq ans. Si l'on se fie à ces taux, un homme doit-il faire systématiquement un dosage de l'antigène PSA ? Un fumeur doit-il passer systématiquement un scanner des poumons ? Il est vrai que ces deux examens détectent plus de cancers à un stade précoce ; mais aucun des deux ne permet de réduire la mortalité.

LES ÉPIDÉMIES DE DIAGNOSTICS

Les gens considèrent souvent les tests de dépistage comme des garants de leur santé. Toutefois, des examens supplémentaires peuvent conduire à des interventions médicales inutiles dont les effets sont parfois délétères. Et pour les nombreux patients inutilement diagnostiqués, le traitement a forcément des conséquences indésirables. Une épidémie de diagnostics peut être aussi dangereuse pour la santé que la maladie.

Les erreurs d'interprétation des statistiques seraient moins fréquentes si les chercheurs, les médecins et les médias utilisaient des données chiffrées directes au lieu de nombres qui prêtent à confusion : le risque absolu au lieu du risque relatif, les fréquences naturelles à la place des probabilités conditionnelles, et les taux de mortalité plutôt que les taux de survie à cinq ans. En outre, nous devons mieux éduquer les jeunes à la science du risque et de l'incertitude.

Comme le suggérait H. G. Wells, les statistiques devraient être enseignées en même temps que la lecture et l'écriture. De fait, aux États-Unis, l'Association nationale des enseignants de mathématiques insiste pour que l'enseignement des statistiques et des probabilités commence à l'école primaire. Si les enfants apprenaient que le monde n'est pas fait de certitudes et ce d'une façon ludique, les statistiques seraient mieux comprises.

Au lieu d'apprendre aux étudiants comment appliquer des formules de probabilité pour résoudre des problèmes virtuels, les professeurs devraient leur montrer comment utiliser les statistiques pour résoudre des problèmes concrets. Par exemple, ils pourraient leur enseigner l'usage des statistiques et des probabilités quand il s'agit de décider comment se comporter face aux drogues, à la consommation d'alcool, à la conduite automobile, aux biotechnologies et à d'autres questions importantes pour la vie quotidienne.

Un livre scolaire américain du secondaire raconte l'histoire vraie d'une mère célibataire de 26 ans. À la suite d'un test positif de dépistage du VIH, elle perd son travail, déménage dans un foyer hébergeant d'autres personnes séropositives, a des relations sexuelles non protégées avec l'un d'entre eux, et attrape une bronchite. Son médecin lui prescrit alors un nouveau test de dépistage. Le résultat est négatif, tout comme celui de son échantillon sanguin précédent, qui a été réanalysé. Cette femme a vécu un cauchemar parce que ses médecins n'ont pas compris qu'un résultat positif à ce test n'était pas un verdict définitif, mais qu'il signifiait que cette femme avait une probabilité d'être infectée de 50 %, étant donné son appartenance à un groupe à faible risque.

L'éducation statistique peut changer des vies, aider les gens à prendre de meilleures décisions personnelles, à reconnaître les messages trompeurs et à développer une attitude plus sereine envers leur santé. Comme le recommandait le philosophe Emmanuel Kant : « Osez savoir ! » ■

BIBLIOGRAPHIE

WOLOSHIN *ET AL.*, *Know your chances : Understanding health statistics*, University of California Press, 2008.

A. PLEASANT, Communiquer sur les statistiques et le risque, *SciDev Net*, 15 décembre 2008.

A. FAGERLIN *ET AL.*, Making numbers matter : present and future research in risk communication, *Am. J. of Health and Behav.*, vol. 31, pp. S47-S51, 2007.

G. GIGERENZER, *Calculated risk : how to know when numbers deceive you*, Simon & Schuster, 2003.

B. FALISSARD, *Comprendre et utiliser les statistiques dans les sciences de la vie*, Masson, Abrégés, 3e éd., 2005.

L'ESSENTIEL

- Cachés derrière un pseudonyme, nous laissons de nombreuses données personnelles sur Internet ou à divers organismes.
- C'est insuffisant pour garantir l'anonymat de ces données !
- Diverses techniques sont développées pour protéger les données sensibles sans empêcher leur exploitation.

LES AUTEURS

TRISTAN ALLARD est postdoctorant à l'Inria (Institut national de recherche en informatique et en automatique).

BENJAMIN NGUYEN est chercheur à l'Inria et maître de conférences à l'Université de Versailles-Saint-Quentin-en-Yvelines.

PHILIPPE PUCHERAL est chercheur à l'Inria et professeur à l'Université de Versailles-Saint-Quentin-en-Yvelines.

L'art de préserver L'ANONYMAT

Presque tous nos comportements laissent des empreintes numériques. Les informaticiens conçoivent et développent des techniques pour que l'analyse de ces gisements de données ne compromette pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information: toutes ces facettes de nos vies laissent des empreintes numériques qui sont stockées et classées dans des bases de données gigantesques, telles celles des moteurs de recherche, qui gardent l'historique des requêtes associées à une adresse IP (la carte d'identité d'un appareil connecté) pendant des années. Ces masses d'informations constituent une manne sans précédent pour les sociétés humaines. Dans le domaine de la santé, par exemple, on peut analyser les caractéristiques individuelles de chaque patient, afin de mieux le prendre en charge, ou de mener des études de santé publique. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements, la quantité et la diversité des informations aujourd'hui disponibles en multiplient l'intérêt potentiel… et les dangers.

En effet, ces myriades d'informations font planer une menace, elle aussi sans précédent, sur la vie privée des individus. La plupart du temps, les données ne sont pas publiques, mais elles circulent néanmoins entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre une partie des données personnelles de leurs utilisateurs… Certaines données sont même accessibles à tous sur Internet : par exemple, le site Netflix a publié des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand celui-ci ne contient ni son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec des préjudices multiples: discrimination au crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique... Protéger les données personnelles sans empêcher leur exploitation est devenu un enjeu majeur à l'ère du tout-numérique. La diffusion de ces données est un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus.

UN BESOIN D'ÉQUILIBRE

Le souci de la protection des données s'est accru pendant la seconde moitié du XX^{e} siècle, à mesure que l'informatique se généralisait. Il a fait naître un domaine de recherche spécifique dans les années 1970, visant à garantir un anonymat plus ou moins complet aux titulaires des données. Le phénomène s'est encore accéléré depuis le début des années 2000, avec l'essor d'Internet, qui permet de consulter et de croiser instantanément ou presque de multiples sources d'information (réseaux sociaux, annuaires...). Lors de la publication de données, il est essentiel de prendre en compte les connaissances annexes accessibles à un individu mal intentionné. En outre, les analystes se contentent de moins en moins de statistiques sur les données personnelles et demandent d'accéder directement à celles-ci, afin d'augmenter la précision de leurs études.

La conjonction de ces deux facteurs a exacerbé la nécessité d'une protection robuste. Au cours des dix dernières années, de nombreux chercheurs ont étudié la façon de publier des données tout en préservant la vie privée des individus concernés. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée, afin de réduire la probabilité que l'on puisse réidentifier la personne correspondante ou accéder à des informations sensibles la concernant. >

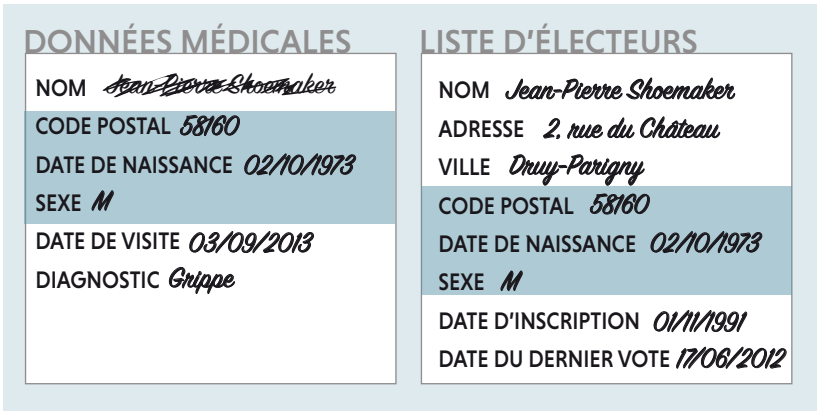
> Comment rendre anonyme un jeu de données? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650000 utilisateurs sur une période de trois mois (pour les rendre disponibles aux chercheurs). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots-clés des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont réussi à identifier la personne portant le pseudonyme 4417749 en croisant les mots-clés de ses requêtes avec des données disponibles sur le Web.

ASSAINIR LES DONNÉES

De nombreux cas similaires illustrent l'échec de la pseudonymisation à rendre les données anonymes. Remplacer les seuls champs directement identifiants, tels que le nom et le prénom ou le numéro de sécurité sociale ne prémunit pas contre une réidentification ultérieure (*voir la figure ci-dessus*). Pourtant, la pseudonymisation reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

Au début des années 2000, Latanya Sweeney, de l'université Carnegie-Mellon, aux États-Unis, a proposé une méthode, nommée *k*-anonymat, pour prévenir les réidentifications *via* des croisements de données. Elle a d'abord montré qu'il était parfois possible de retrouver le titulaire d'un jeu de données médicales pseudonymisé à partir des champs {Sexe, Date de naissance, Code postal}. En effet, ces champs sont aussi présents dans d'autres jeux de données comprenant le nom du titulaire. En outre, la combinaison des renseignements correspondants est unique pour la plupart des gens, selon plusieurs études récentes.

Pour empêcher les réidentifications, Latanya Sweeney a proposé de scinder les champs du jeu de données en deux catégories: les champs quasi identifiants (notés QID), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD), qui doivent rester privés (par exemple, un diagnostic médical). Les valeurs des quasi-identifiants d'un individu sont ensuite rendues identiques à celles d'au moins $(k - 1)$ autres individus dans le jeu de données, pour un entier k convenablement choisi. En d'autres termes, le *k*-anonymat dissimule chaque individu dans une foule d'au moins k personnes impossibles à distinguer par leurs quasi-identifiants. Tout croisement avec une autre source de données aura ainsi une précision inférieure à $1/k$.



Le titulaire d'un dossier médical dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données. En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(en rouge)* est unique. Il suffit alors de trouver un autre jeu de données, par exemple une liste électorale *(à droite)*, qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière.

Les algorithmes du *k*-anonymat se fondent sur le principe de généralisation, qui consiste à remplacer les valeurs précises des quasi-identifiants par des ensembles de valeurs. Ceux-ci peuvent être des intervalles numériques ou des catégories. Il serait simple d'élaborer un algorithme qui parcoure les quasi-identifiants et forme un ensemble les incluant tous, tel {âge de 0 à 150 ans, habitant sur la Terre}. Mais les données seraient trop dégradées (c'est-à-dire trop imprécises) pour être utiles aux analystes. On a alors développé des algorithmes dits par partitionnement: ils forment des ensembles d'au moins k individus en regroupant les quasi-identifiants voisins, qu'ils remplacent ensuite par un intervalle ou une catégorie les incluant tous. Le plus utilisé est l'algorithme dit de Mondrian (*voir l'encadré page 102*).

Cependant, le *k*-anonymat ne résout pas tous les problèmes. Considérons la situation suivante: un attaquant dispose d'un jeu de données *k*-anonymes et recherche la valeur sensible (tel le diagnostic médical) d'un individu cible dont il connaît le quasi-identifiant. Il peut retrouver le groupe auquel appartient sa cible, et donc l'ensemble des valeurs sensibles de ce groupe. La protection apportée par le *k*-anonymat est d'autant plus faible que le nombre de ces valeurs est petit. Dans le cas extrême, supposons que les k individus du groupe partagent la même valeur sensible, par exemple un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer: la valeur sensible n'est pas protégée.

BROUILLER LES PISTES

De multiples techniques ont succédé au *k*-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre en compte des attaques potentielles réalistes.

Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'université Cornell, à New