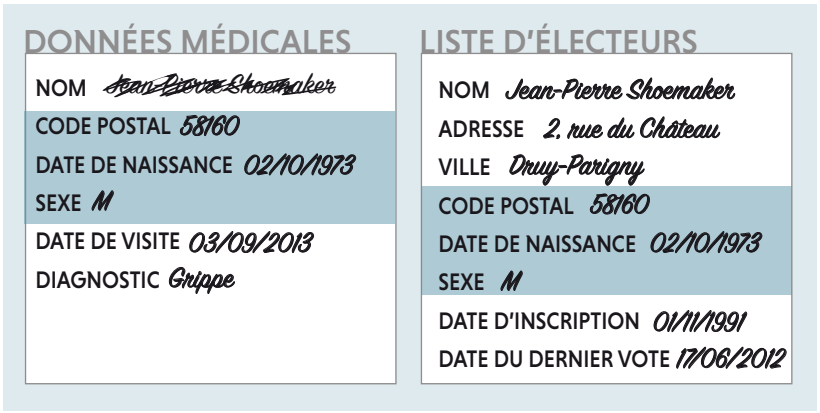
> Comment rendre anonyme un jeu de données? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de 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.

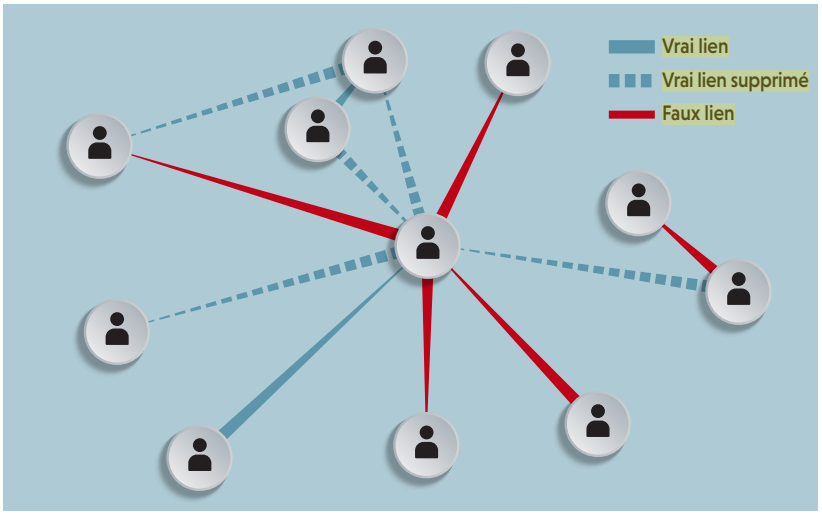
Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'université Cornell, à New

York, et ses collègues, quantifie la connaissance préalable qu'a l'attaquant de sa cible par la probabilité de trouver la valeur convoitée avant l'observation du jeu de données: si l'attaquant recherche le diagnostic médical d'une personne nommée Dupont, qu'il ne connaît de sa cible que son nom et qu'il sait que 10% des Dupont ont un cancer, il a 1 chance sur 10 de ne pas se tromper en affirmant que sa cible a un cancer. Cette connaissance est dite *a priori*. Après la découverte des données, l'attaquant a plus de chances de trouver la donnée qu'il cherche; la connaissance de sa cible qui en résulte, connaissance dite *a posteriori*, est quantifiée par la probabilité de lui associer la bonne valeur sensible.

C'est ici que la théorie bayésienne intervient, car elle permet de calculer des probabilités conditionnelles: les probabilités de déduire telle ou telle information d'un jeu de données sachant qu'on a telle ou telle connaissance préalable. La confidentialité des données publiées est caractérisée par la différence entre connaissance *a posteriori* et connaissance *a priori*, autrement dit par la connaissance supplémentaire sur la cible apportée par les données. Assurer un niveau adéquat de protection revient à limiter cette valeur. Cependant, cette technique est difficile à mettre en œuvre, car elle suppose de connaître exactement ce que l'attaquant sait de sa cible.

La $l$-diversité, méthode conçue la même année et par les mêmes chercheurs, est davantage applicable. Elle consiste à imposer la présence de $l$ valeurs sensibles distinctes dans chaque groupe (*voir l'encadré page 102*). Un attaquant qui ne connaît de sa cible que le groupe auquel elle appartient ne peut trouver sa valeur sensible avec une probabilité supérieure à $1/l$.

De nombreuses autres techniques ont été élaborées depuis, afin de s'adapter à la diversité des données, des attaques et des attaquants. La «$t$-proximité» vise à créer des groupes au sein desquels la distribution des données est à peu près la même que dans la population globale.




Les graphes de réseau social représentent les individus par des points et leurs liens par des traits. Pour qu'ils ne dévoilent pas la vie privée des individus, on peut leur appliquer un algorithme dit de confidentialité différentielle, consistant, par exemple, à supprimer de nombreux vrais liens et à les remplacer par autant de faux liens.

La «$(c, k)$-sûreté» prend en compte la capacité de l'attaquant à formuler $k$ déductions logiques du type: «Si Adrien a la grippe, alors Bettie aussi.» La «3D-confidentialité» considère que l'attaquant connaît *a priori* trois types de données: $l$ valeurs sensibles que sa cible n'a pas, $k$ valeurs sensibles que d'autres individus ont et $m$ déductions logiques entre individus. La «$m$-invariance» protège contre les comparaisons entre les publications successives d'un même jeu de données, telles certaines informations statistiques sur un hôpital, dont les variations reflètent les arrivées, les départs et les évolutions des patients.

L'application de ces techniques commence souvent par celle de l'algorithme de Mondrian. Il sert à construire des groupes $k$-anonymes, au sein desquels on vérifie la conformité de la distribution des données sensibles vis-à-vis du modèle choisi (*voir l'encadré page 102*).

Volontairement ou non, tous ces modèles sont des exemples du paradigme de non-information, formulé en 1977 par le statisticien suédois Tore Dalenius: selon ce paradigme, un jeu de données est d'autant plus confidentiel qu'il renseigne peu l'attaquant sur sa cible. En 2006, Cynthia Dwork, alors chercheuse à Microsoft, a proposé une nouvelle façon de définir la confidentialité, qualifiée de confidentialité différentielle: un jeu de données serait confidentiel s'il n'est presque pas modifié par l'ajout des données d'un individu, quelles qu'elles soient.

## NOYER LES DONNÉES

L'assainissement consiste alors à perturber les données de sorte que celles de chaque individu se trouvent noyées dans la perturbation (et non plus dans la foule, comme dans le $k$-anonymat). On introduit, par exemple, un grand nombre de fausses données (typiquement 100 fois plus que de vraies), afin que le jeu de données qui contient la contribution d'un individu particulier ressemble beaucoup à celui qui ne la contient pas. Certains algorithmes suppriment aussi de vraies données. La protection est assurée par l'impossibilité de distinguer les fausses données des vraies dans le jeu de données assaini, dont on ne peut même pas garantir qu'il contient la contribution d'un individu précis.

L'idée de la confidentialité différentielle est en plein essor. Elle est désormais applicable à des types variés de données, tels les graphes de réseaux sociaux. Ces graphes représentent les individus par des nœuds, connectés par des liens. La perturbation peut alors consister à supprimer de vrais liens et à en introduire de faux. Des informations tel le nombre de liens sont toujours extractibles du graphe, sans que l'on puisse déterminer les connexions précises d'un individu.

Le succès de la confidentialité différentielle s'explique en partie par sa simplicité: l'attaquant >