unique pour la plupart des gens, selon plusieurs études récentes – qui portaient sur les États-Unis, mais c'est probablement aussi le cas ailleurs.

Pour empêcher les réidentifications, L. Sweeney a proposé de scinder les champs du jeu de données en deux catégories : les champs quasi identifiants (QID, pour *Quasi identifiers data*), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD, pour *Sensitive data*), 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 $1/k$.

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. Par exemple, l'intervalle d'âges [15, 20] (c'est-à-dire de 15 à 20 ans) englobe plus d'individus qu'un âge précis. De même, la catégorie « France » englobe plus d'individus que la catégorie « Bourgogne ».

Il serait simple d'élaborer un algorithme qui parcourre 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 66)*.

Cependant, le k-anonymat est loin de résoudre tous les problèmes. Considérons la situation suivante : un attaquant dispose d'un jeu de données k-anonyme 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 qu'ils aient tous un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer : la valeur sensible n'est pas protégée.

Brouiller les données confidentielles

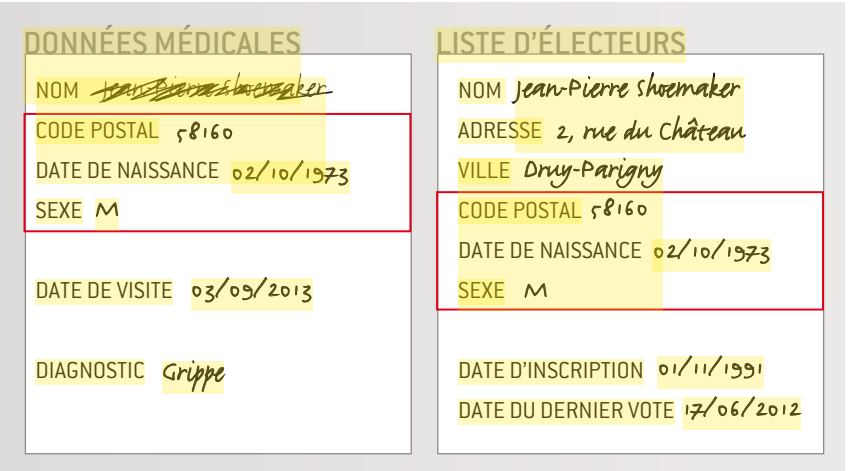
De multiples techniques ont succédé au k-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre en 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 dix pour cent des Dupont ont un cancer, il a une chance sur dix 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. C'est d'autant plus difficile qu'on ignore souvent qui sera le ou les attaquants – ces derniers pouvant même avoir des connaissances *a priori* différentes.

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 66)*. 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 $1/l$.



2. LE TITULAIRE D'UN DOSSIER MÉDICAL dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données (de tels dossiers circulent parfois entre les hôpitaux et divers instituts de recherche ou de statistique). En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(encadrée en rouge)* est unique. Il suffit alors de trouver un autre jeu de données qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière. C'est le cas, par exemple, d'une liste électorale *(à droite)*, disponible sur simple demande à la mairie.

De nombreuses autres techniques ont été élaborées au cours des années suivantes, 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. 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 ci-dessous)*.

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.

L'assainissement consiste alors à perturber les données de telle 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.

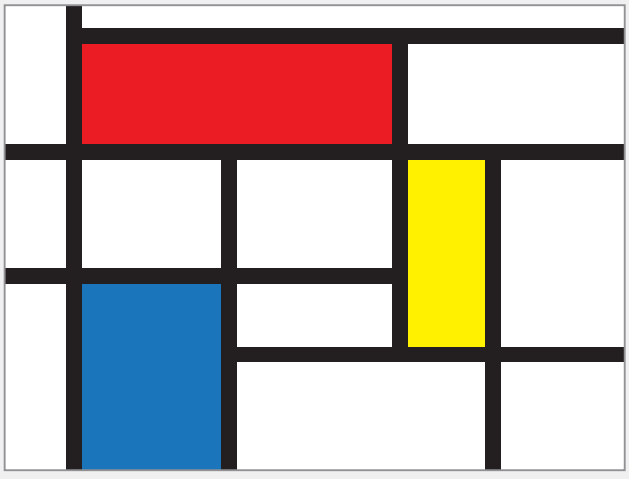
Une vision alternative de la confidentialité

Les fausses données sont élaborées de façon à ce qu'on puisse toujours extraire certaines informations. Prenons l'exemple d'un jeu de 100 dossiers médicaux, à partir duquel on souhaite savoir combien de personnes ont la grippe. On perturbe ce jeu de données en y introduisant 1 000 faux dossiers, chacun se voyant attribué de façon équiprobable une maladie parmi quatre (dont la grippe).

COMMENT RENDRE ANONYME UN JEU DE DONNÉES ?

Pour rendre le titulaire d'un jeu de données impossible à identifier, on se contente souvent de remplacer les champs tels que le nom ou le numéro de sécurité sociale par un pseudonyme. Cela n'empêche pas toujours les réidentifications ultérieures *via* des croisements de données. D'autres techniques ont alors été développées. Celle du *k*-anonymat est la plus employée aujourd'hui. Elle consiste à brouiller certains champs, dits quasi-identifiants parce que leur combinaison est parfois unique : l'âge, le sexe, le code postal… On remplace leurs valeurs précises par des intervalles de valeurs ou des catégories, de façon à ce que que la combinaison résultante soit partagée par au moins *k* personnes.

Pour appliquer cette technique, on utilise souvent l'algorithme de Mondrian, élaboré en 2006. Son nom est une référence à Piet Mondrian (1872-1944), peintre hollandais dont les œuvres partitionnent l'espace *(ci-contre, un dessin inspiré de Mondrian)* ; de même, l'algorithme partitionne l'espace des quasi-identifiants, où ces derniers sont indiqués sur des axes et où les individus sont repérés par des points.



© Shutterstock/Reinhold Leitner

L'algorithme commence par choisir un quasi-identifiant, selon lequel il divise les individus en deux groupes, par exemple ceux dont le code postal est inférieur ou égal à 75011, et les autres *(ci-contre en haut à droite)*. Si *k* est supérieur à 2, les groupes résultants sont de nouveau divisés en deux (en utilisant le code postal ou un autre quasi-identifiant, tel l'âge) pour former quatre groupes. Ce découpage continue jusqu'à ce qu'aucun groupe ne puisse plus être divisé sans englober moins de *k* individus. La dernière étape consiste à remplacer les quasi-identifiants de chacun par l'ensemble de valeurs couvert par son groupe.

Le problème est que les données personnelles à protéger, tel le diagnostic médical, peuvent être identiques au sein d'un groupe. Un attaquant qui sait que sa cible appartient à ce groupe a alors accès à sa valeur sensible (les valeurs sensibles étant les données à protéger). Plusieurs techniques visent à y remédier.

La *l*-diversité, par exemple, consiste à construire des groupes ayant au moins *l* valeurs sensibles. On choisit la valeur de *l* en fonction de ce que l'attaquant sait de sa cible, c'est-à-dire du nombre maximal de valeurs sensibles qu'on l'estime capable d'éliminer. Plus il sait de choses, plus *l* sera élevé, plus il faut inclure d'individus dans chaque groupe afin d'avoir une diversité suffisante, et moins les statistiques calculées à partir du jeu de données assaini sont fines. Le compromis entre utilité et protection est ici réalisé à travers le choix de *l*.

En pratique, on construit souvent les groupes grâce à l'algorithme de Mondrian. À chaque division d'un groupe en deux, on vérifie que les nouveaux groupes contiennent au moins *l* valeurs sensibles distinctes. Dans le cas contraire, on revient en arrière et on partitionne le groupe selon un autre quasi-identifiant *(en bas à droite)*.