

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 1000 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 qua-

si-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 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 Mon-

drian (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.

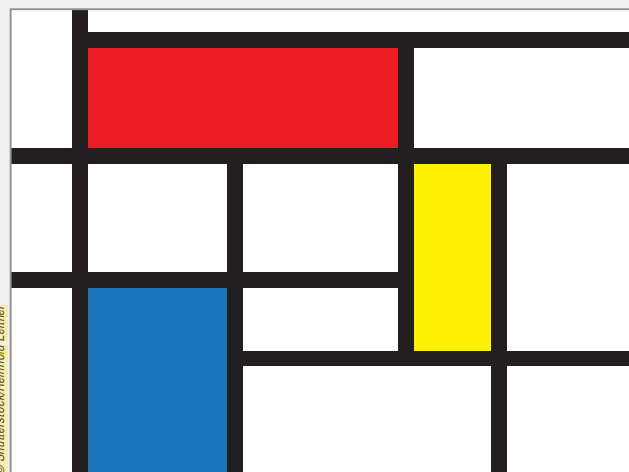
L'algorithme commence par choisir un quasi-identifiant, selon lequel il divise les individus en deux groupes, par exemple ceux dont le code postal est inférieur ou égal à 75011, et les autres (*ci-contre en haut à droite*). Si *k* est supérieur à 2, les groupes résultants sont de nouveau divisés en deux (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 atta-

quant 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*).



© Shutterstock/Reinhold Leiber

On sait alors qu'on a environ 250 valeurs fausses pour chacune de ces maladies. En comptant le nombre de grippés dans le jeu de données assaini et en lui soustrayant 250, on obtient le nombre réel de cas de grippe avec une bonne précision. Plus les vraies données sont nombreuses par rapport aux fausses, plus les informations extraites sont précises, mais moins les données sont protégées. Il existe de nombreuses façons de créer des fausses données, et les algorithmes qui s'en chargent font l'objet de multiples études.

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 (voir la figure 3). 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 telles que 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 n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et

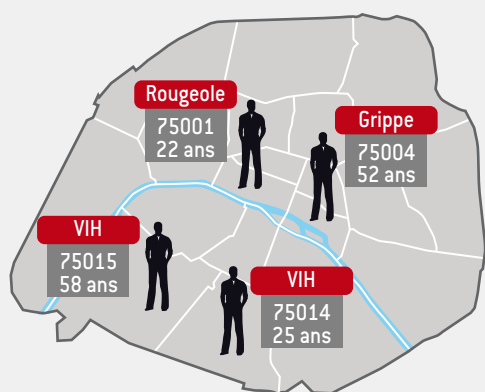
UN MODÈLE D'ASSAINISSEMENT universel des données reste donc chimérique, chaque modèle étant plus ou moins efficace selon la situation.

données sensibles, parfois peu évidente, n'est pas nécessaire. Mais cette simplicité est à double tranchant : dans des travaux publiés en 2011, A. Machanavajjhala et ses collègues ont montré que la non-prise en compte des connaissances annexes de l'attaquant et des relations potentielles entre les individus d'un jeu de données ouvre des failles dans la protection.

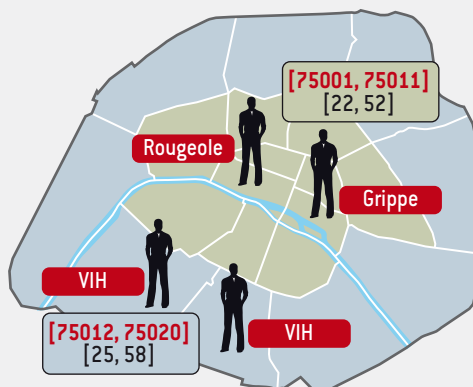
Un modèle d'assainissement universel des données reste donc chimérique. Chaque modèle a ses défenseurs et ses détracteurs, et est plus ou moins efficace selon la situation.

Outre le modèle d'assainissement, l'architecture de gestion des données a une importance. Les données nominatives sont souvent extraites du système informatique assurant leur usage quotidien (tel le serveur d'un centre de soin), puis copiées sous une forme pseudonymisée dans un entrepôt de données. C'est à partir de cet entrepôt que seront produits à la demande des jeux de données assainis pour différents destinataires. En France et en Angleterre, le système de dossier médical personnel national fonctionne ainsi. Un tel processus facilite la production de multiples versions assainies d'un même jeu de données, chacune étant adaptée aux besoins du destinataire et limitée aux usages prévus par la législation. Ainsi,

Partitionnement selon le code postal



Partitionnement selon l'âge



Les individus ayant un code postal voisin sont réunis au sein de groupes dits 2-anonymes (le code postal et l'âge ne permettent pas de distinguer les membres d'un groupe de deux personnes). Or tous les individus du groupe bleu sont infectés par le VIH. Un attaquant qui sait que sa cible appartient à ce groupe pourra donc déduire une donnée sensible de sa cible.

[75001, 75014]
[22, 25]

VIH

Rougeole

[75004, 75015]
[52, 58]

VIH

Grippe



Avec un partitionnement selon l'âge, les groupes 2-anonymes sont aussi 2-divers : chacun contient deux valeurs sensibles distinctes (deux diagnostics différents). Dans cette situation, on ne peut pas déduire le diagnostic d'un individu cible dont on connaît l'âge et le code postal.