Protection de la vie privée
2 - Principes de conception et de gestion

Guillaume Piolle
guillaume.piolle@centralesupelec.fr
http://guillaume.piolle.fr/

CentraleSupélec – majeure SIS

10 janvier 2017

Principes de conception et de gestion

Principes de conception

La vie privée selon l'ISO

Common Criteria for Information Technology Security Evaluation

Norme ISO/IEC 15408, successeur de l'*Orange Book* du DoD.
Section 7 : protection de la vie privée.

Exigences techniques pour assurer la vie privée

- **Anonymat** (*anonymity*) : incapacité des observateurs à déterminer l'identité d'un utilisateur ;
- **Pseudonymat** (*pseudonymity*) : idem, mais en imposant à l'utilisateur de répondre de ses actions ;
- **Non-chaînabilité** (*unlinkability*) : incapacité des observateurs à déterminer si deux actions ont été réalisées par le même utilisateur ;
- **Non-observabilité** (*unobservability*) : incapacité des observateurs à déterminer si une action est en cours.

Souveraineté des données

Faire en sorte que **l'utilisateur conserve le contrôle** sur les données personnelles le concernant :

- Stocker en priorité données et/ou clés sur ses terminaux personnels ;
- Contrôler étroitement usage et diffusion, en imposant des obligations (obligations de sécurité, notifications, suppression...).

Souveraineté au sens juridique : attention toute particulière à porter aux données externalisées, en particulier lorsque le prestataire est de droit US, et ce **même si les données sont stockées dans l'union européenne** : le *Stored Communication Act*, le *PATRIOT ACT* et *FISAA* permettent à l'administration US d'accéder aux données, silencieusement, et pour des motifs pouvant parfois relever de simples intérêts économiques (cf. analyses et recommandations SGDSN, délégation interministérielle à l'intelligence économique, Parlement européen).

Minimisation des données

Minimisation des données

cf. principe de proportionnalité

- Ne collecter que les données absolument nécessaires à la finalité ;
- Ne les transmettre/conserver que si c'est absolument nécessaire ;
- Détruire dès que possible les données non absolument nécessaires ;

Le tout dans les limites des obligations d'auditabilité des systèmes.

Le problème, c'est que dans certains cas d'usage il n'est pas facile, voire impossible, de savoir à l'avance quelles données vont être utiles/nécessaires. . . Ce n'est pas quelque chose que la CNIL ou un CIL aime forcément entendre !

Le *Privacy by Design*

Principe

La protection de la vie privée, comme la sécurité, ne peut être efficace que si elle est pensée dès la conception du système. Les ajouts postérieurs ne peuvent pas espérer colmater des brèches de conception.

Le principe, de plus en plus mentionné dans les textes, doit concerner à la fois les intervenants techniques et non techniques, conjointement.

Exemples de mise en œuvre :

- Travail de spécification incluant experts techniques, juristes et décideurs ;
- Application de méthodes formelles de conception ;
- *Privacy impact assessments* ;
- Systèmes contraints par les politiques ;
- ...

Le *Privacy by Design*

Les sept principes du PbD

- Proactif plutôt que réactif, préventif plutôt que correctif ;
- Considérer la protection de la vie privée comme le réglage par défaut (*Privacy by default*) ;
- Intégrer la protection de la vie privée dans la conception du système ;
- Assurer des fonctionnalités complètes : viser une somme positive, pas une somme nulle ;
- Assurer la sécurité de bout en bout, avec une protection tout au long du cycle de vie ;
- Visibilité et transparence – viser l'ouverture ;
- Montrer du respect pour la vie privée des utilisateurs – centrer les systèmes sur les utilisateurs.

Sécurité, vie privée, auditabilité

- La protection de la vie privée (ou des données personnelles) peut être considérée **du point de vue de la sécurité informatique** ;
- Certaines exigences de la vie privée **peuvent être remplies** grâce aux outils classiques de la sécurité informatique ;
- Certaines exigences de la vie privée **ne peuvent pas être remplies** grâce aux outils classiques de la sécurité informatique ;
- Certaines exigences de vie privée sont **incompatibles** avec certaines exigences de la sécurité informatique.

Parfois présentée comme une **sous-discipline**, parfois comme une discipline **connexe** ou **transverse**, parfois comme une discipline **concurrente**.

Sécurité, vie privée, auditabilité

Besoin d'auditabilité

Un impératif de la sécurité informatique : se donner les moyens de détecter les comportements malveillants ou erronés et de désigner des responsables.
Principal outil : conservation de **journaux** retraçant l'activité d'un système (logiciel, serveur web, etc.).

auditabilité vs vie privée ?

- Certes, la conservation des journaux, très riches en informations, est un risque potentiel pour la vie privée des usagers ;
- La protection de la vie privée et des données personnelles nécessite aussi de conserver beaucoup d'informations : charge de la preuve, effectivité du droit d'accès, du droit à l'oubli, contrôles, notification des brèches, communication...

→ axes de recherche sur l'auditabilité préservant la vie privée

Sécurité, vie privée, auditabilité

Obligations de journalisation

- 2001, Loi sur la Sécurité Quotidienne (LSQ) : les opérateurs télécom doivent conserver les données de connexion pendant un an (mesure temporaire, prolongée *ad vitam*) ;
- 2004, Loi sur la Confiance dans l'Économie Numérique (LCEN) : conservation des informations identifiant les personnes déposant des contenus en ligne (étendu à tous les fournisseurs d'accès) ;
- 2011, décret d'application de la LCEN : conservation des identifiants, peudonymes, mots de passe, données de paiement, coordonnées (étendu aux hébergeurs et éditeurs de sites web).

En cas de journalisation insuffisante ?

Jusqu'à 375 k€ d'amende pour une société, 75 k€ et un an d'emprisonnement pour son dirigeant.

Sécurité, vie privée, auditabilité

Qui peut accéder aux journaux ?

- La justice (commission rogatoire, décision en référé ou en instance) ;
- La police, sur réquisition simple (sans autorisation judiciaire), depuis la loi du 23 janvier 2006 sur la lutte contre le terrorisme ;
- L'administrateur système/réseau, qui « est tenu d'une **obligation de confidentialité** » (même vis-à-vis de l'employeur, en tout cas en ce qui concerne les e-mails) et peut accéder aux données « dans le cadre de sa mission de sécurité du réseau informatique » (Cour de Cassation, 17 juin 2009).

Un risque opérationnel aggravé ?

À des fins de sécurité (lutte contre le terrorisme), on augmente le risque de dommages en cas d'intrusion et on fournit une incitation aux attaquants éventuels. *[Déportez la journalisation !]*

Bases de données anonymes ou anonymisées

- Absence de données permettant d'identifier une personne de manière unique :
  - Retrait des nom et prénom ;
  - Remplacement par un numéro aléatoire ;
  - Remplacement par des psudonymes arbitraires. . .
- Sondages anonymes, officiels ou non ;
- Sondages et questionnaires dont la partie identifiante est ensuite désolidarisée du reste.

Cadre juridique des bases de données anonymes

Si l'on considère qu'il n'y a pas de « données à caractère personnel » parce qu'il n'y a pas d'éléments identifiants, alors la loi Informatique et Libertés **ne s'applique pas** !

Conséquence :

- **Aucun droit** pour les personnes concernées ;
- **Aucune obligation** pour les responsables de traitements ;
- **Aucune restriction** à la conservation, la publication, l'exploitation, le rapprochement avec d'autres bases de données.

Mais. . . aucun problème puisque tout est anonyme ?

L'anonymisation parfaite est impossible

Dans la majorité des cas, une « anonymisation » des données ne suffit pas à empêcher l'identification des individus.

On considère généralement que l'anonymisation est une opération impossible dans le cas général et que le terme est donc impropre. On préfère souvent parler de « pseudo-anonymisation », parfois de « désidentification » ou « d'assainissement » suivant les cas (ou encore de « pseudonymisation », mais cela peut faire référence à d'autres notions).

Désanonymisation : l'affaire Netflix

Une démarche inoffensive

Netflix : plate-forme de vidéo à la demande permettant l'évaluation des films visionnés et la recommandation personnalisée.

2010 : Netflix publie des données d'évaluation anonymes, dans le cadre d'un concours (*Netflix prize*, 1M$) visant à améliorer son algorithme de recommandation.

Un chercheur recoupe les données anonymes avec celles du site IMDb et « désanonymise » la base. Les goûts cinématographiques des utilisateurs deviennent des données identifiantes !
La connaissance de deux notes suffit à identifier 68 % des utilisateurs.

Plainte fédérale, Netflix se rétracte et met fin au concours pour cause de risque pour la vie privée.

Désanonymisation : les aventures de Latanya Sweeney, épisode 1

L'affaire du GIC

Milieu des années 90 : le *Group Insurance Commission* du Massachusetts décide de rendre publiques des données « anonymisées » concernant les hospitalisations des employés de l'état.

L. Sweeney, étudiante à Carnegie Mellon, recoupe ces données avec les listes électorales et envoie le détail de son dossier médical au gouverneur.

Le gouverneur fait faire marche arrière au GIC...

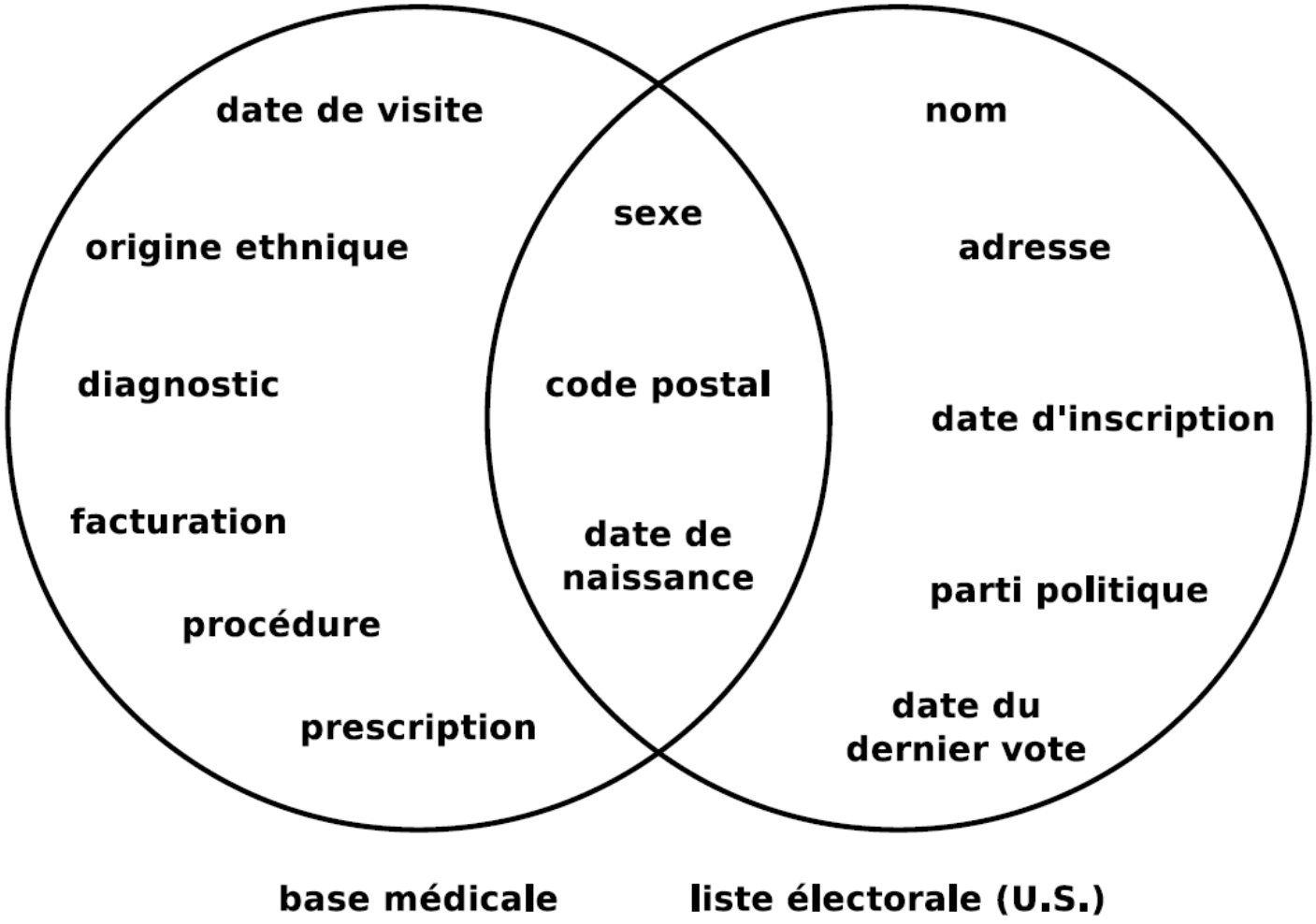
Désanonymisation : les aventures de Latanya Sweeney, épisode 2

Naissance du k-anonymat

2000 : L. Sweeney montre que 87 % des citoyens U.S. peuvent être identifiés de manière unique par leur sexe, leur date de naissance et leur code postal (recoupements faciles avec les registres publics).

Publication en 2002 : introduction du concept de *k-Anonymity*, qui mesure de manière mathématique le degré d'anonymat d'une base de données « anonyme ».

Problème de l'interconnexion de bases de données



Exemple fictif : une base de données « anonyme »

Sondage anonyme (fictif) sur les étudiants d'un campus

Sexe	Taille	Orientation sexuelle
...	...	...
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	bisexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	180-190	hétérosexuel
M	180-190	homosexuel
M	180-190	hétérosexuel
M	180-190	autre
M	180-190	hétérosexuel
M	190-200	hétérosexuel
M	200-210	homosexuel

Orientation sexuelle : **sensible** au sens de l'article 8 de la loi « Informatique et Libertés ».

MAIS : sondage complètement anonyme... donc hors du champ de la loi !

- Réalité de cet « anonymat » ?
- Les étudiants sont-ils tous égaux devant cet « anonymat » ?
- À quelles questions répondez-vous lors de sondages « anonymes » ?

Principes du k-anonymat

Quasi-identifiant

Ensemble d'attributs d'une base de données pouvant permettre, dans au moins un cas, d'identifier un tuple à l'aide d'informations externes.

N'importe quel attribut peut appartenir à un quasi-identifiant !

Nouvel éclairage sur la notion de « donnée à caractère personnel » (art. 2 de la loi « Informatique et Libertés »).

k-Anonymat

Une base de données est dite **k-anonyme** si tout tuple est indistingable d'au minimum $k - 1$ autres tuples de la base projetée sur tout quasi-identifiant.

Retour sur la base d'exemple

Sondage anonyme (fictif) sur les étudiants d'un campus

Sexe	Taille	Orientation sexuelle
...	...	...
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	bisexuel
M	**170-180**	**hétérosexuel**
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	170-180	hétérosexuel
M	180-190	hétérosexuel
M	180-190	homosexuel
M	180-190	hétérosexuel
M	180-190	autre
M	180-190	hétérosexuel
M	190-200	hétérosexuel
M	200-210	homosexuel

La base est **1-anonyme** : c'est le pire cas !

Au moins une personne (deux ici) est ré-identifiable à l'aide d'une base de données externe facile à concevoir.

On peut dire que l'individu en gras est 8-anonyme dans la base.

k-anonymisation

Principe

À partir d'une base 1-anonyme, on cherche à obtenir d'une base « publiable » k-anonyme, avec k suffisamment grand.

Attention !

- En k-anonymisant, on limite l'intérêt (utilité) de la base ;
- La taille du quasi-identifiant dépend de l'ensemble des bases dans le monde extérieur.

k-anonymisation

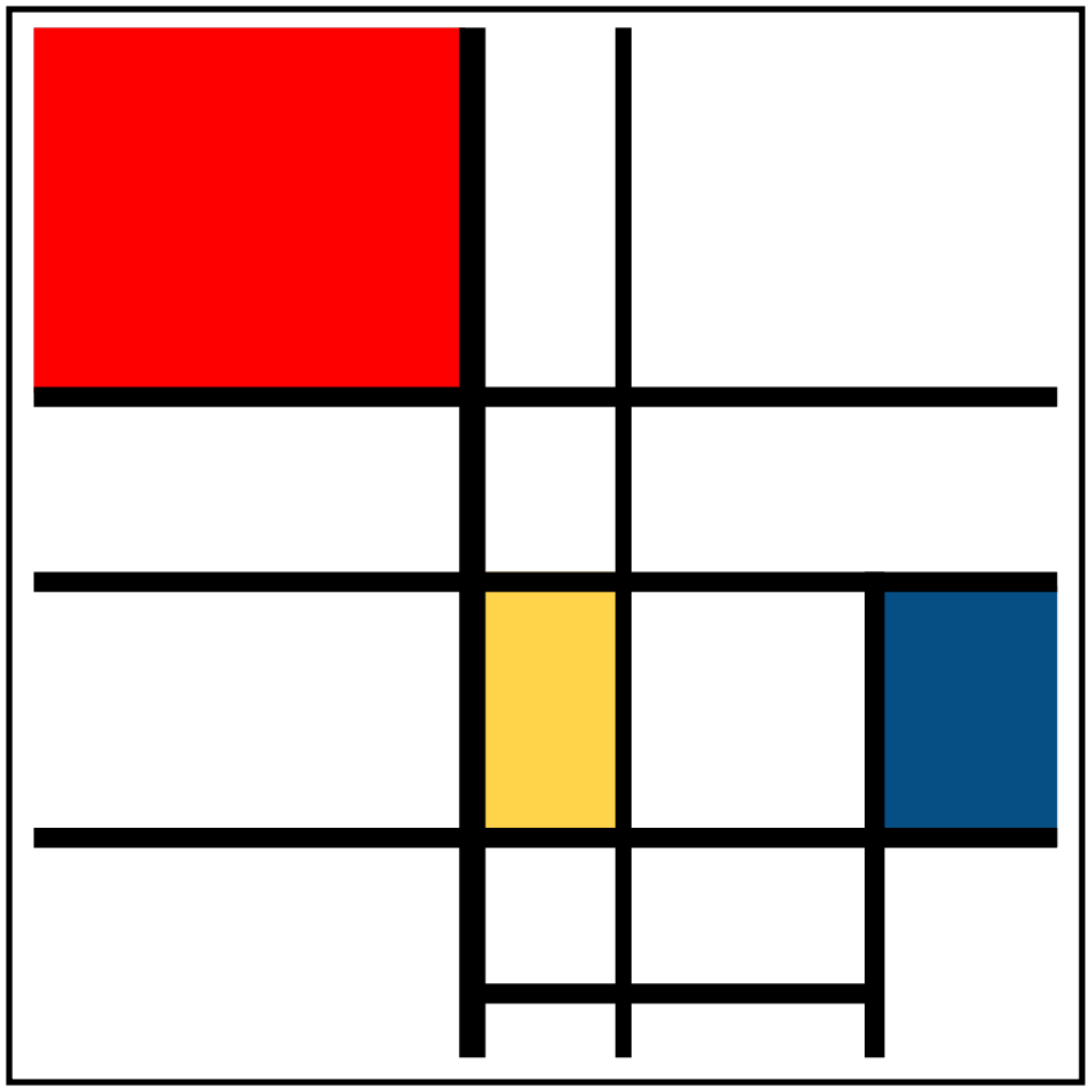
Exemple de k-anonymisation : algorithmes de type Mondrian

- **Hypothèse :** un attribut de la base est soit partie d'un QI, soit une donnée sensible ;
- **Objectif :** rendre le plus équivoque possible le lien entre une valeur de QI et les données sensibles correspondantes ;
- **Mécanisme :** partitionner l'espace des QI de manière à former des groupes d'au moins k éléments, puis remplacer dans la BD les QI par l'identifiant de la partition.

- On part de l'ensemble de tous les QI ;
- On partitionne au fur et à mesure (par dichotomie par exemple) jusqu'à obtenir un maximum de zones de k éléments au minimum.

Ces algorithmes sont paramétrables par les caractéristiques des distributions dans chaque zone, par exemple.

k-anonymisation



Attaques contre les bases k-anonymes

Complementary release attack

Corrélation de deux extraits de la même base, k-anonymisés de manière différente.

Unsorted matching attack

Valide uniquement pour des bases où l'ordre des tuples a un sens.
Principe : on effectue des correspondances entre deux bases de données k-anonymisées, mais dont les tuples sont publiés dans le même ordre.

Attaques temporelles

Comparaison du résultat de requêtes faites à des moments différents, corrélées avec des informations extérieures.

Homogénéité et diversité des données

Homogeneity attack

Age	CP	Diagnostic
...	...	...
20-25	35000	Colite
20-25	35000	Liposarcome
20-25	35000	Rhume des foins
20-25	35000	Entorse
25-30	35000	Grippe aviaire
25-30	35000	Angine virale
25-30	35000	Coqueluche
25-30	35000	Pneumonie
20-25	35510	Syphilis
20-25	35510	Syphilis
20-25	35510	Syphilis
20-25	35510	Syphilis

La base est 4-anonyme, mais pourtant on peut apprendre avec certitude des informations sensibles sur certains individus.

La cause en est une trop grande homogénéité dans les résultats de certaines classes (pas assez de **diversité**).

Homogénéité et diversité des données

l-diversité (*l-diversity*)

Une classe (dans une base k-anonyme, par exemple), est dite **l-diverse** s'il y a au moins *l* valeurs *bien représentées* pour l'attribut sensible.

« *l* valeurs *bien représentées* » peut signifier :

- qu'il y a au moins *l* valeurs distinctes ;
- que l'entropie de la classe (par référence à l'espace des valeurs pour l'attribut sensible) est supérieure à $log_2(l)$;
- que, suivant d'autres métriques ((c-l)-diversité), la valeur la plus courante n'apparaît pas *trop fréquemment* et que la valeur la moins courante n'est pas *trop rare* ;
- ...

Représentativité des données

Homogeneity attack

Age	CP	Diagnostic
...	...	...
20-25	35510	Syphilis
20-25	35510	Gonorrhée
20-25	35510	Chlamydia
20-25	35510	Herpès génital

La base est 4-anonyme et 4-diverse, mais on apprend encore des informations significatives sur les individus concernés, *notamment* parce que la distribution des attributs sensibles ne correspond pas à la distribution globale dans la population.

La *t*-**closeness** mesure la proximité de la distribution des attributs avec la distribution connue ou supposée dans la population d'où est tirée la base.

Primitives de pseudo-anonymisation

Des outils à combiner et à adapter à chaque usage

- **Pseudonymisation** (remplacement d'attributs par des identifiants aléatoires ou issus d'un hachage / chiffrement) : n'améliore pas les métriques mentionnées ;
- **Agrégation** ;
- **Projection** (impact fort sur l'utilité) ;
- **Sélection** (nécessite un modèle fiable de la distribution, risque de biais) ;
- Insertion d'**enregistrements artificiels** (idem) ;
- **Ajout de bruit** (à régler avec soin) ;
- **Permutation d'attributs** (plutôt du « brouillage », risques divers) ;
- ...

Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques*, Commission européenne, 2014.

Métriques de l'indistingabilité

- k-anonymity (Sweeney 2002) ;
- l-diversity (Machanavajjhala 2006) ;
- **Differential privacy** (Dwork 2006) ;
- t-closeness (Li 2007) ;
- (c, k)-Safety (Martin 2007) ;
- 3D-Privacy (Chen 2007) ;
- (d, γ)-Privacy (Rastogi 2007) ;
- ϵ-Privacy (Machanavajjhala 2009) ;
- Towards a unified theory of privacy and utility (Kifer 2010) ;
- ...

Differential privacy

Principe « gros grain »

Lorsqu'un attaquant interroge une base de données, il ne doit pas apprendre plus d'information sur moi si j'y figure que si je n'y figure pas.

Cynthia Dwork *et al.*, *Calibrating noise to sensitivity in private data analysis*, 2006

Une méthode d'interrogation $\mathcal{A}$ est ϵ-*differentially private* si et seulement si (pour tout sous-ensemble S de l'image de $\mathcal{A}$) :

$$\frac{P(\mathcal{A}(D) \in S)}{P(\mathcal{A}(D') \in S)} < e^{\epsilon}$$

où D' diffère de D par un seul tuple.

$\mathcal{A}$ est généralement une méthode qui injecte du bruit (laplacien) soit dans la base, soit dans les résultats.

Crédits iconographiques

-  – Hay Kranen *Mondrian lookalike.svg*, 2007 (CC-BY 2.5, Wikimedia Commons).