unisanté

Anonymiser ou désidentifier ses données de recherche

Comment protéger ses participant-e-s

Céline Racine – Spécialiste en gestion des données – Unisanté
Udd.data@unisante.ch

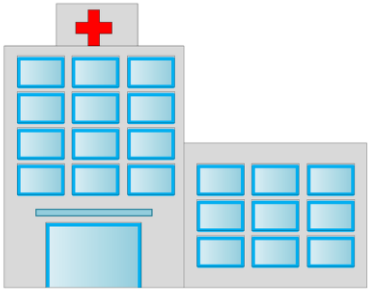
Le plan

- Un peu de théorie
  - Anonymisation
  - Risques
- Pratique : comment faire ?
  - Evaluation
  - Traitement
  - Vérification
- Conclusion
- Questions

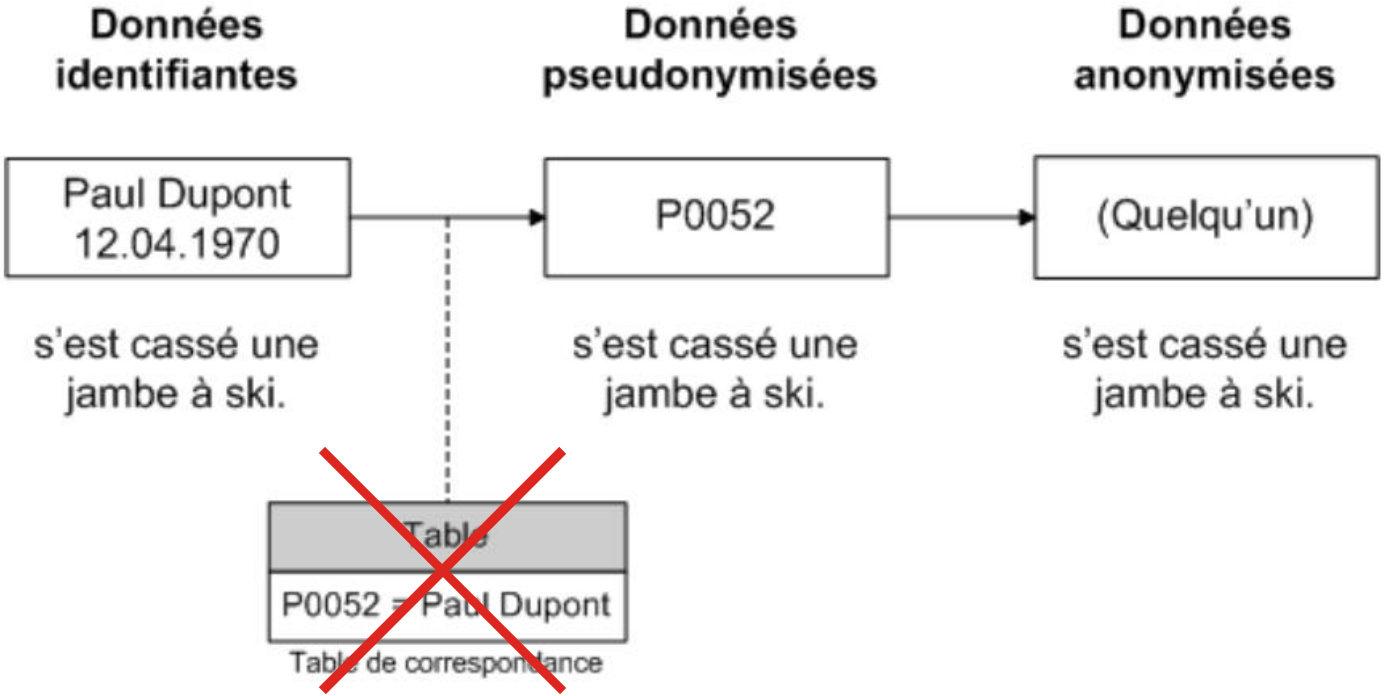
La protection des données

Cambridge Analytica

UNDERGROUND



Que dit la loi suisse ?



unisanté

Source : https://www.edoeb.admin.ch/edoeb/fr/home/protection-des-donnees/dokumentation/guides/mesures-techniques-et-organisationnelles-de-la-protection-des-do.html

En réalité

Anonymisation stricte
(= Anonymisation)

- Irréversible
- IMPOSSIBLE de réidentifier un individu
- Partage en Open Access

Anonymisation relative
(= Désidentification)

- Risques de réidentification
- Niveau de protection
- Partage en accès restreint

Terminologie

	Identifiées / identifiables	**Codées**	**Désidentifiées (faible)**	**Désidentifiées (fort)**	**Anonymes**
Identifiants directs (HIPAA)	Intact	Transformés ou supprimés	Transformés ou supprimés	Transformés ou supprimés	Supprimés
Identifiants indirects	Intact	Intact	Intact	Transformés ou supprimés	Fortement transformés ou supprimés
Table de correspondance	-	Intacte	Supprimée	Supprimée	Supprimée

La désidentification est réalisée au cas par cas. Son objectif est de trouver le bon équilibre entre utilité du set de données et protection des participant-e-s.

Les données anonymes ne peuvent plus être reliées à un être humain et sortent du cadre légal de la LRH et LPD. L'impossibilité de ré-identification doit être garantie (expl : differential privacy) ou les données doivent être fortement agrégées.

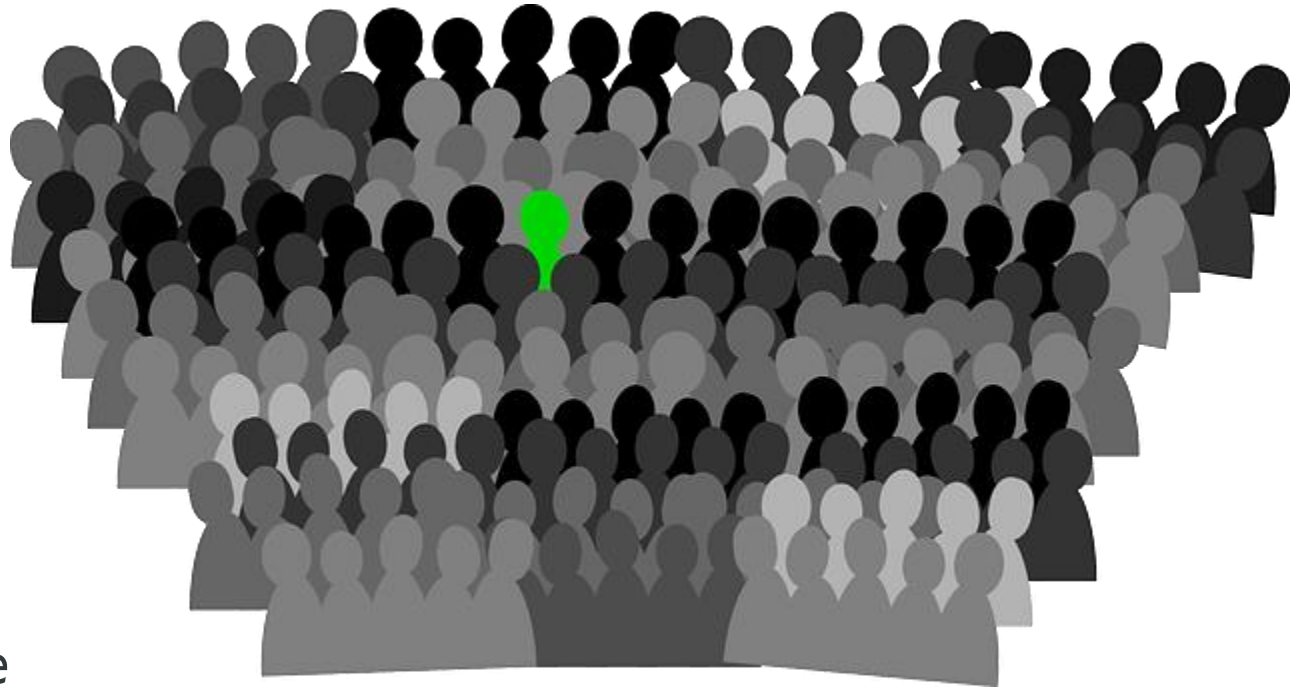
Risques de réidentification

unisanté

Image by Gordon Johnson from Pixabay

Exercice

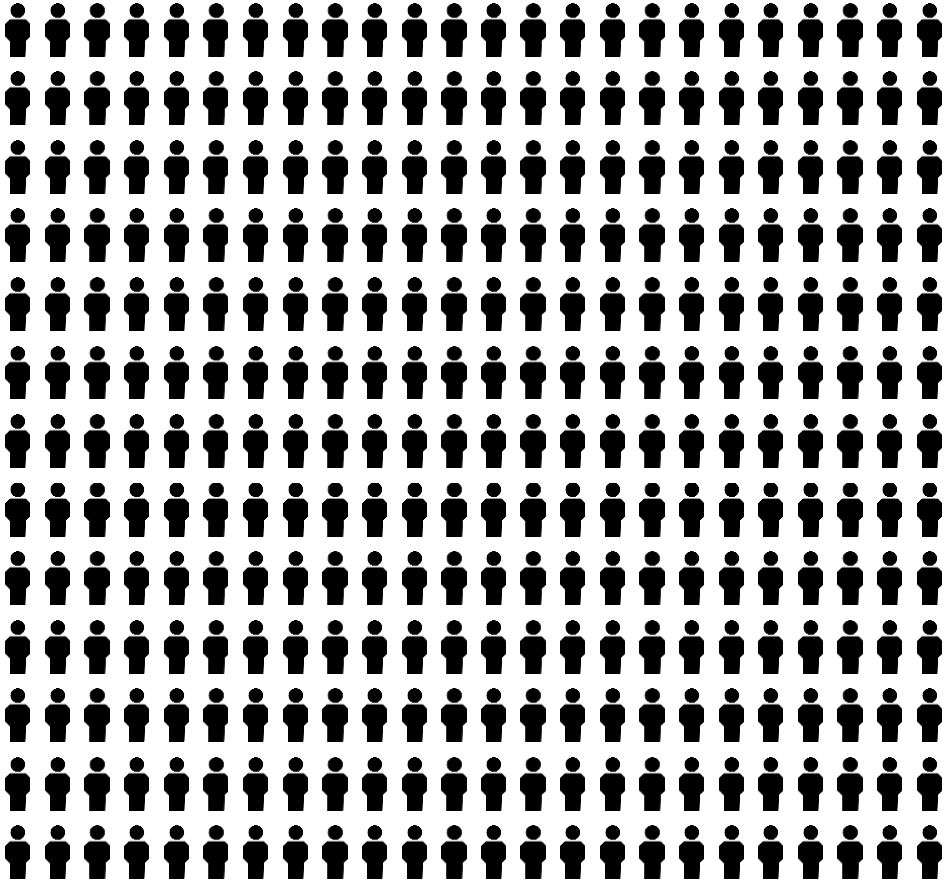
- Avec 3 indicateurs indirects, je trouve un outlier



Risque 1: Individualisation

Possibilité d'isoler 1 individu dans le set de données

- Attention particulière:
  - Petite population
  - Caractéristiques originales



unisanté

Risque 2: Corrélation

Possibilité de ré-identifier un individu en croisant des données

- Attention particulière:
  - Identifiants forts et faibles
  - Données codées/pseudonymisées

ID	0000515
Tél.	4.89.33
Institution	Unisanté

ID	9745831
Tél.	4.89.33
Personne	Céline Racine

ID 0000515 = Céline Racine

Exemples de sources de données

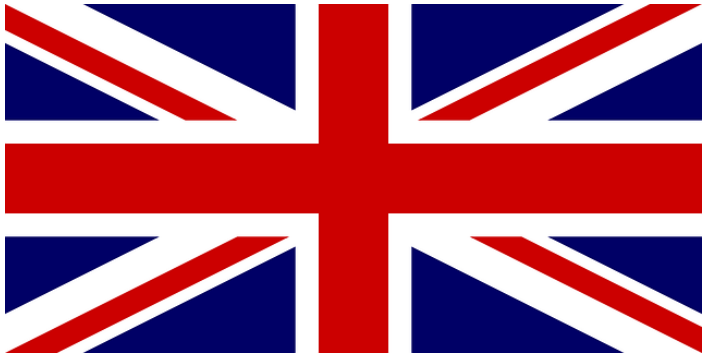
- List broker
  - https://shop.kbdata.ch/consumer/selection
- Dépôts de données
- Réseaux sociaux
- Entreprises
- Darknet

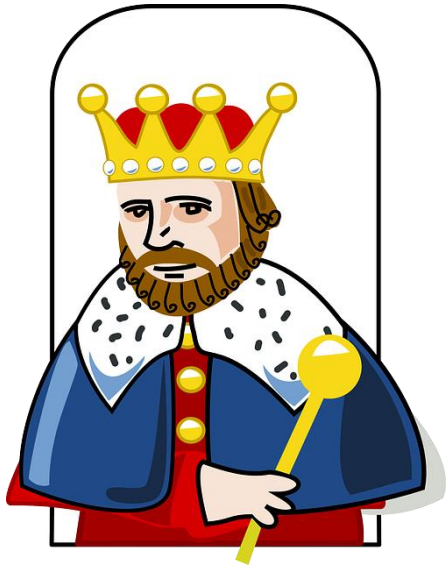
Risque 3:

Inférence

Possibilité de deviner l'identité d'un individu sans information supplémentaire

- Attention particulière:
  - Contexte
  - Informations publiques





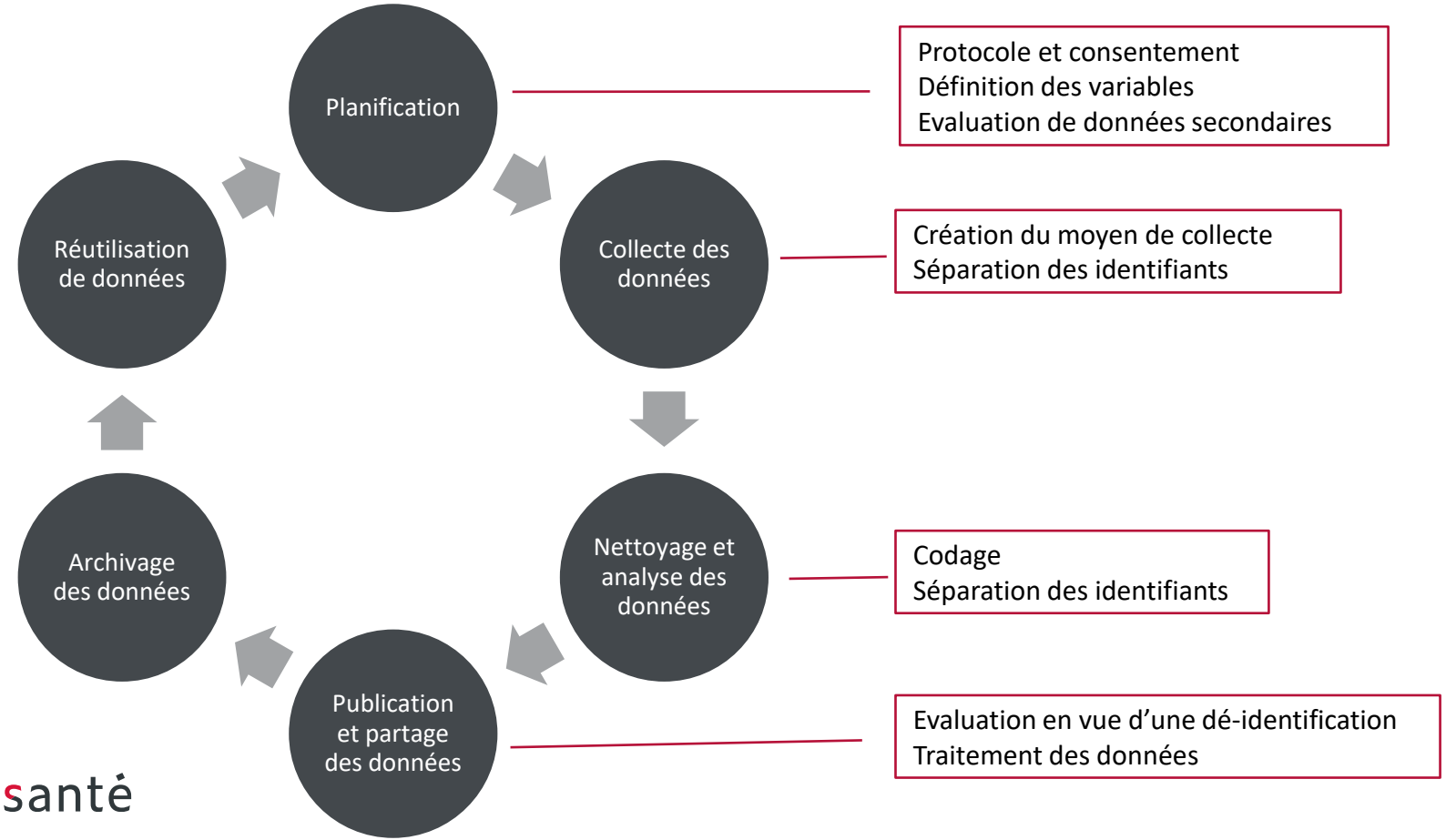
unisanté

Risques de réidentification

- S’il y a un risque, ce n’est pas anonyme

- Attention particulièrement à
  - Données sensibles
  - Petite population
  - Caractéristiques originales / données socio-démographiques
  - Destinataire des données

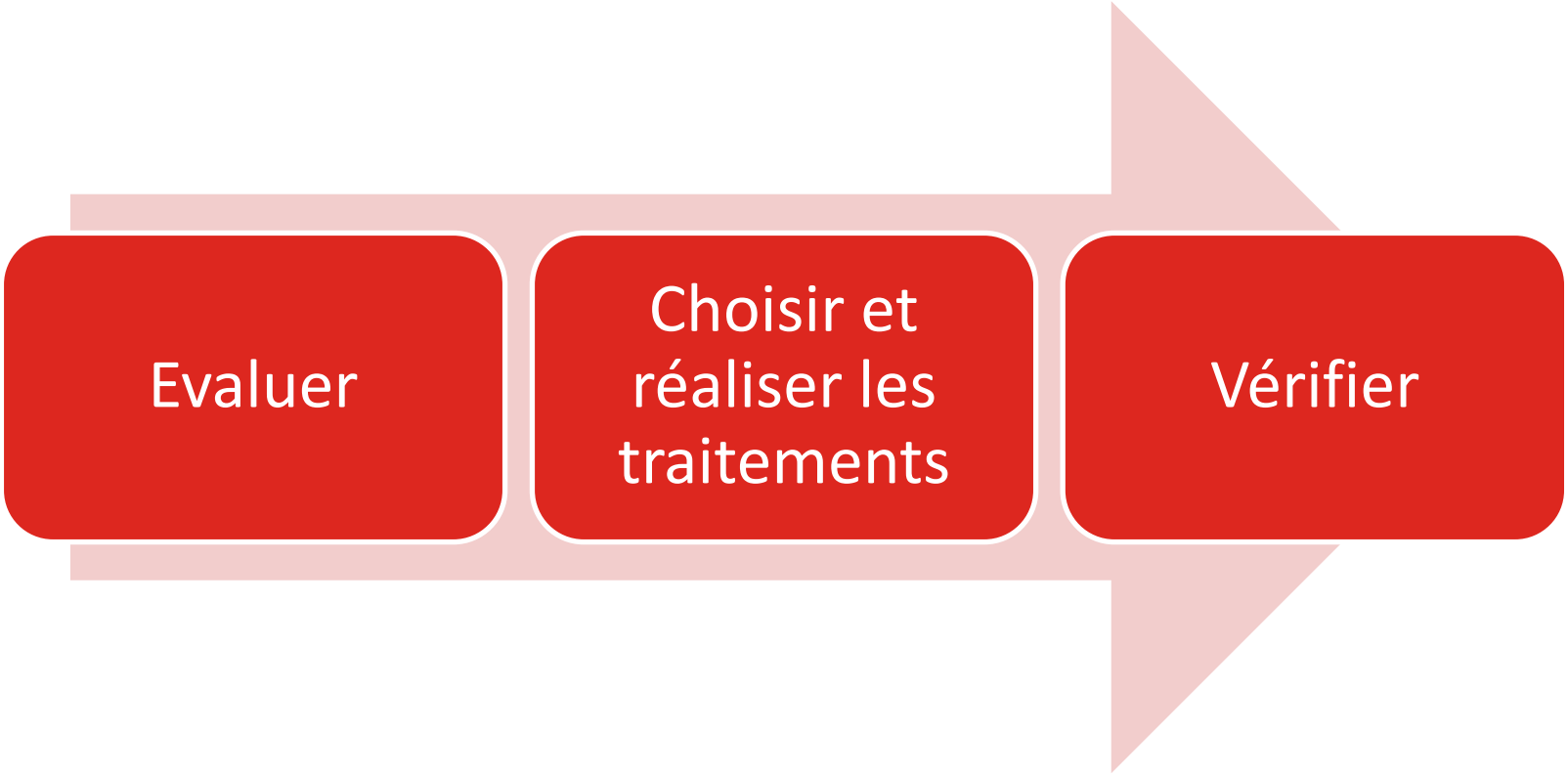
A quel moment ?



unisanté

Techniques de désidentification

En pratique




Obligation ?
Consentement ?

Evaluer le niveau de sensibilité

- Comment faire ?
  - Connaissance du domaine
  - Identifiants
  - Fusion de données

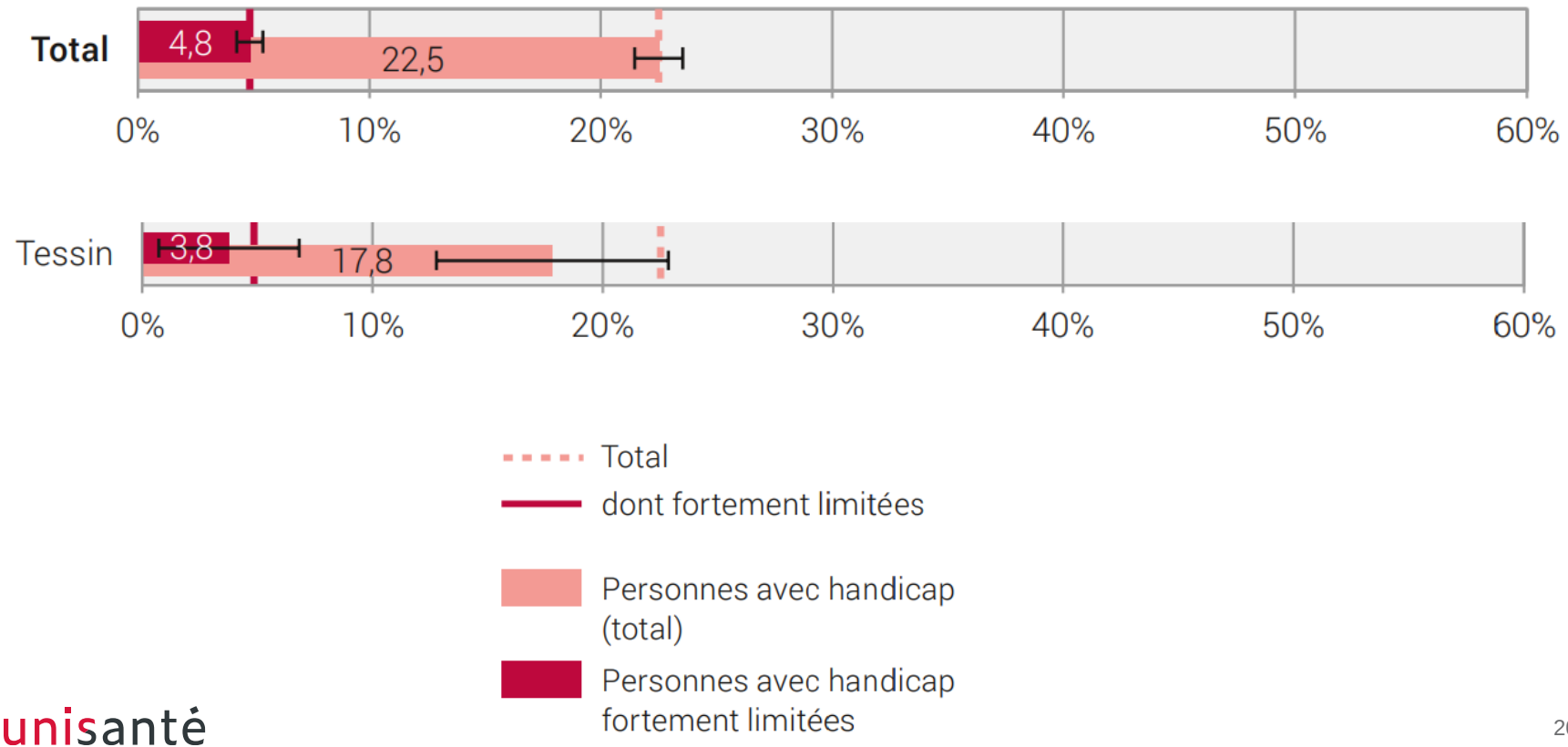
  - Utiliser les risques

Exercice

- Projet : est-ce que le handicap limite la consommation de produits BIO au Tessin ?

- Méthode de collecte : questionnaire en ligne
- Répondant-e-s : 3000 personnes

Connaissance du domaine



unisanté

Catégorisation des identifiants

- Directs
  - Réidentification certaine (nom, numéro AVS)
- Forts
  - Réidentification très probable en croisant des données (adresse IP, maladie rare)
- Faible
  - Réidentification probable en croisant beaucoup de données (genre, date de naissance, lieu de travail)

Norme HIPAA

1. Nom
2. Adresse (y compris les subdivisions plus petites que les États, telles qu'une adresse postale, une ville, un comté ou un code postal)
3. Toutes les dates (à l'exclusion des années) qui sont directement liées à une personne, y compris l'anniversaire, la date d'admission ou de sortie, la date de décès ou l'âge exact des personnes âgées de plus de 89 ans
4. Numéro de fax
5. Numéro de téléphone
6. Adresse email
7. Numéro de dossier médical
8. Numéro de sécurité sociale
9. Numéro de bénéficiaire d’assurance
10. Numéro de compte
11. Numéro de certificat / permis de conduire
12. Identifiants de véhicule, numéros de série ou numéros de plaque d'immatriculation
13. Identifiants d'appareils ou numéros de série
14. URL Web
15. Adresse IP
16. Identifiants biométriques tels que les empreintes digitales ou les empreintes vocales
17. Photos identifiantes (visage, tatouages…)
18. Tout autre numéro d'identification, caractéristique ou code unique

Catégorisation des identifiants

Variable	Type
Date de naissance	Numérique (DD-MM-AAAA)
Genre	Liste à choix
Code postal	Numérique
Situation de handicap	Booléen (Oui/Non)
Type de handicap	Liste à choix
Proportion d'aliment bio	Numérique (%)
Lieu d'achat	Liste à choix
Accessibilité du magasin	Liste à choix
Commentaires	Texte libre

Aller sur
www.menti.com

Entrer le code
3238 3038

unisanté

Catégorisation des identifiants

Variable	Type	Direct	Fort	Faible	Non
Date de naissance	Numérique (DD-MM-AAAA)			X	
Genre	Liste à choix			X	
Code postal	Numérique			X	
Situation de handicap	Booléen (Oui/Non)			X	
Type de handicap	Liste à choix			X	
Proportion d'aliment bio	Numérique (%)				X
Lieu d'achat	Liste à choix				X
Accessibilité du magasin	Liste à choix				X

Catégorisation des identifiants

Variable	Type	Direct	Fort	Faible	Non
Date de naissance	Numérique (DD-MM-AAAA)			X	
Genre	Liste à choix			X	
Code postal	Numérique		X		
Situation de handicap	Booléen (Oui/Non)			X	
Type de handicap	Liste à choix		X		
Proportion d'aliment bio	Numérique (%)				X
Lieu d'achat	Liste à choix				X
Accessibilité du magasin	Liste à choix				X

Results

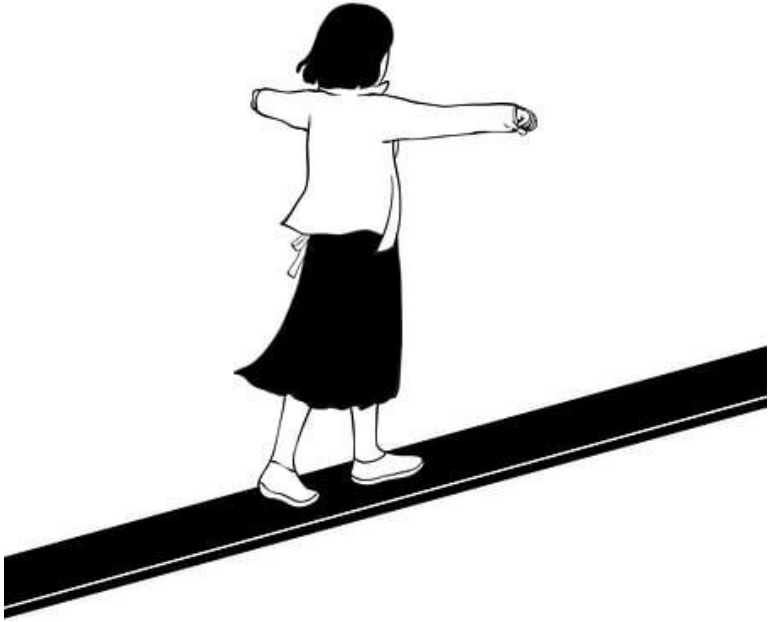
- Identifiers
- «Small» population
- Re-identification risks :
  - Singling out
  - Data linking

We will treat the data

Comment faire ?

Choisir les traitements selon…

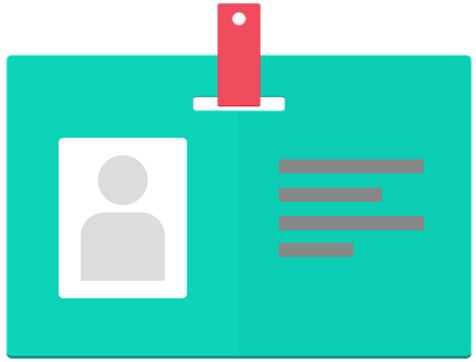
- Type d’identifiant
- Sensibilité
- Outliers
- Utilité



KEEP CALM AND DE-IDENTIFY DATA

Suppression

- Les identifiants directs
- Les commentaires
- Les données inutiles / inutilisables



Utilité du set pour de futures recherches

Modifications

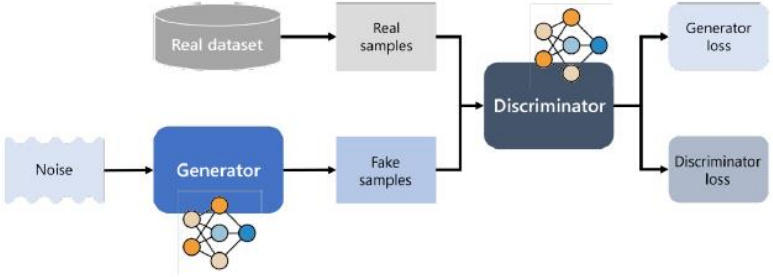
- Transformer
- Crypter
- Généraliser – réduire la granularité
- Agréger
  - Les valeurs
  - Les individus -> pour Open Access
- Randomiser (= altérer les données)

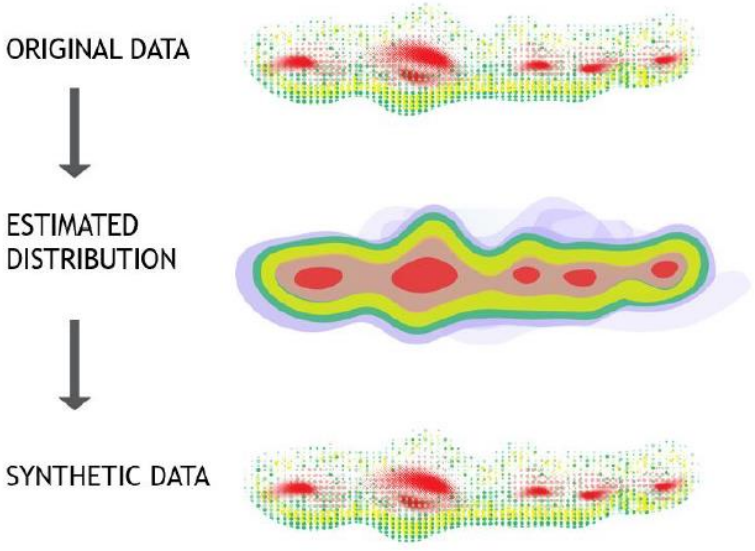
Bruit et données fictives

- C’est un type de randomisation
- Permet un niveau de protection plus élevé

- Méthodes statistiques
- Algorithmes de machine learning

Synthetic data: a promising solution to alleviate the concerns on the privacy-utility trade-off





- Can address privacy concerns associated with real data
- Can address bias in real data with synthetic data diversification
- Can be a cost-effective approach for creating large datasets

Choisir les traitements - résumé

A supprimer

- Noms et initiales
- Adresse postale, code postal *
- Email, numéro de téléphone, de fax
- Numéros identifiants (AVS, prothèses, plaques d'immatriculation, compte bancaire, adresse IP, dossier médical…)
- Identifiants biométriques (empreintes digitales et vocales, images identifiantes (visage, tatouage…), vidéos identifiantes…)
- Commentaires *
  - Extraire les informations et les coder dans de nouvelles variables puis supprimer

A transformer

- Indications géographiques -> devient le canton ou plus grand
- Dates
  - Garder uniquement l'année
  - Décaler la date
- Date de naissance / âge
  - Garder uniquement l'année
  - Faire des catégories (fortement recommandé pour les plus de 89 ans)
- Toute caractéristique qui permet d'identifier un individu unique (Exemple : prince héritier d'Angleterre)

* Si cette information est nécessaire pour de futures recherches, transformation possible

Rappel : mes données sont…

	Identifiées / identifiables	**Codées**	**Désidentifiées (faible)**	**Désidentifiées (fort)**	**Anonymes**
Identifiants directs (HIPAA)	Intact	Transformés ou supprimés	Transformés ou supprimés	Transformés ou supprimés	Supprimés
Identifiants indirects	Intact	Intact	Intact	Transformés ou supprimés	Fortement transformés ou supprimés
Table de correspondance	-	Intacte	Supprimée	Supprimée	Supprimée

La désidentification est réalisée au cas par cas. Son objectif est de trouver le bon équilibre entre utilité du set de données et protection des participant-e-s.

Les données anonymes ne peuvent plus être reliées à un être humain et sortent du cadre légal de la LRH et LPD. L'impossibilité de ré-identification doit être garantie (expl : differential privacy) ou les données doivent être fortement agrégées.

Vérifier l’anonymisation

- Reproductibilité
- Utilité

- Risques de réidentification
- K-anonymity
- Differential privacy
- SPHN template

Conclusion

- Désidentifier : au cas par cas
- Garder une balance protection / utilité
- Attention au cadre légal et consentement
- A plusieurs moments du projet
- Utiliser les soutiens à disposition

unisanté

Merci pour votre attention

Références

- BOUTET, Antoine, 2023. Données personnelles : rien à cacher, mais beaucoup à perdre. *The Conversation*. [en ligne]. 29 mars 2023. [Consulté le 2 mai 2023]. Disponible à l'adresse: http://theconversation.com/donnees-personnelles-rien-a-cacher-mais-beaucoup-a-perdre-201494
- CNIL, Commission nationale de l'informatique et des libertés, 2020. L'anonymisation de données personnelles. *cnil.fr*. [en ligne]. 19 mai 2020. [Consulté le 1 août 2022]. Disponible à l'adresse: https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles
- DATA PROTECTION COMMISSION, 2019. *Guidance Note: Guidance on Anonymisation and Pseudonymisation*. [en ligne]. Dublin: Data protection Commission. [Consulté le 2 février 2022]. Disponible à l'adresse: https://www.dataprotection.ie/sites/default/files/uploads/2020-09/190614%20Anonymisation%20and%20Pseudonymisation.pdf
- FINCH, Kelsey, 2016. A Visual Guide to Practical Data De-Identification. *https://fpf.org/*. [en ligne]. 25 avril 2016. [Consulté le 2 mai 2023]. Disponible à l'adresse: https://fpf.org/blog/a-visual-guide-to-practical-data-de-identification/
- How to GO FAIR, [sans date]. *GO FAIR*. [en ligne]. [Consulté le 18 juin 2022]. Disponible à l'adresse: https://www.go-fair.org/how-to-go-fair/
- JOHNSTON, Lisa (éd.), 2017. *Curating research data. Volume two: A handbook of current practices*. [en ligne]. Chicago, Illinois: Association of College and Research Libraries, a division of the American Library Association. ISBN 978-0-8389-8862-6. Disponible à l'adresse: https://renouvaud1.primo.exlibrisgroup.com/permalink/41BCULAUSA_LIB/1vikse1/alma991021107671202852
- JOTTERAND, Alexandre, 2022. Personal Data or Anonymous Data: where to draw the lines (and why)? *Jusletter*. [en ligne]. 15 août 2022. No. 1119. [Consulté le 24 août 2022]. Disponible à l'adresse: https://jusletter.weblaw.ch/juslissues/2022/1119/personal-data-or-ano_173939252d.html
- KLEINER, B, HEERS, M. (soon to be publish). Quantitative data anonymisation: practical guidance for anonymising sensitive social science data. FORS Guide No. 23, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS.

Références

- NGUYEN, Benjamin et CASTELLUCCIA, Claude, 2020. Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. *1024 : Bulletin de la Société Informatique de France*. 2020. No. 15, pp. 23. Disponible à l'adresse : https://hal.science/hal-02570847
- ROCHER, Luc, 2019. Données anonymes… bien trop faciles à identifier. *The Conversation*. [en ligne]. 17 septembre 2019. [Consulté le 13 décembre 2021]. Disponible à l'adresse: http://theconversation.com/donnees-anonymes-bien-trop-faciles-a-identifier-123157
- ROCHER, Luc, HENDRICKX, Julien M. et DE MONTJOYE, Yves-Alexandre, 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. 23 juillet 2019. Vol. 10, no. 1, pp. 3069. DOI 10.1038/s41467-019-10933-3.
- SPRUMONT, Dominique, 2019. Protection des données, anonymisation et recherche. [en ligne]. Lunch LRH, Lausanne. 3 octobre 2019. [Consulté le 2 août 2022]. Disponible à l'adresse: https://static1.squarespace.com/static/60b94bed393f8064950b2821/t/616e9f1dce9f7f7a05927376/1634639647288/Presentation_Protection_des_donnees_anonymisation_et_recherche_191003.pdf
- STAM, A, DIAZ, P. (2023). Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts. FORS Guide No. 20, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2023-00020