Réconcilier exploitation de données massives et anonymat


WARDEN
AUTONOMOUS SECURITY SOLUTION

Le modèle de Warden

Mutualiser les données pour le bénéfice commun

- ➔ Récolte de données sensibles
- ➔ Partage de connaissances
- ➔ Apprentissage machine sécuritaire

Introduction

Cette présentation n’est pas

- ➔ Sur les dérives des entreprises qui génèrent des profits monstre en vendant votre identité.

Cette présentation n’est pas

- ➔ Sur les dérives des entreprises qui génèrent des profits monstre en vendant votre identité.
- ➔ Sur les data brokers et autres Cambridge Analytica

Cette présentation n'est pas

- ➔ Sur les dérives des entreprises qui génèrent des profits monstre en vendant votre identité.
- ➔ Sur les data brokers et autres Cambridge Analytica
- ➔ Sur le *blockchain*

Cette présentation n’est pas

- ➔ Sur les dérives des entreprises qui génèrent des profits monstre en vendant votre identité.
- ➔ Sur les data brokers et autres Cambridge Analytica
- ➔ Sur le *blockchain*
- ➔ Sur comment se protéger d’une fuite de données

Cette présentation n’est pas

- ➔ Sur les dérives des entreprises qui génèrent des profits monstre en vendant votre identité.
- ➔ Sur les data brokers et autres Cambridge Analytica
- ➔ Sur le *blockchain*
- ➔ Sur comment se protéger d’une fuite de données
- ➔ Sur l’intelligence artificielle

Mon objectif

➔ Faire comprendre la pertinence du partage de données

Mon objectif

- ➔ Faire comprendre la pertinence du partage de données
- ➔ Faire comprendre le risque qu'un partage mal opéré peut générer pour l'anonymat

Mon objectif

- ➔ Faire comprendre la pertinence du partage de données
- ➔ Faire comprendre le risque qu'un partage mal opéré peut générer pour l'anonymat
- ➔ Donner un aperçu des techniques et pistes de réflexions pour protéger l'anonymat

Exploitation de données massives (mégadonnées)

➔ Qu’est-ce que sont les données massives?

Exploitation de données massives (mégadonnées)

- ➔ Qu’est-ce que sont les données massives?
- ➔ Qu’est-ce que l’exploitation de données massive?

Exploitation de données massives (mégadonnées)

- ➔ Qu’est-ce que sont les données massives?
- ➔ Qu’est-ce que l’exploitation de données massive?
- ➔ Quel est le travail du “scientifique de données”

Exploitation de données massives

Id	Code postal	Age	Nationalité	Maladie
1	13053	28	Russe	Cardiaque
2	13068	29	Américain	Cardiaque
3	13068	21	Indien	Virus
4	13053	23	Russe	Virus
5	14853	50	Américain	Cancer
6	14850	55	Américain	Cardiaque
7	14850	47	Américain	Virus
8	14823	49	Mexicain	Cardiaque
9	13023	31	Indien	Infection
10	13068	37	Japonais	Cancer
11	13068	36	Américain	Cancer
12	13023	35	Américain	Cancer

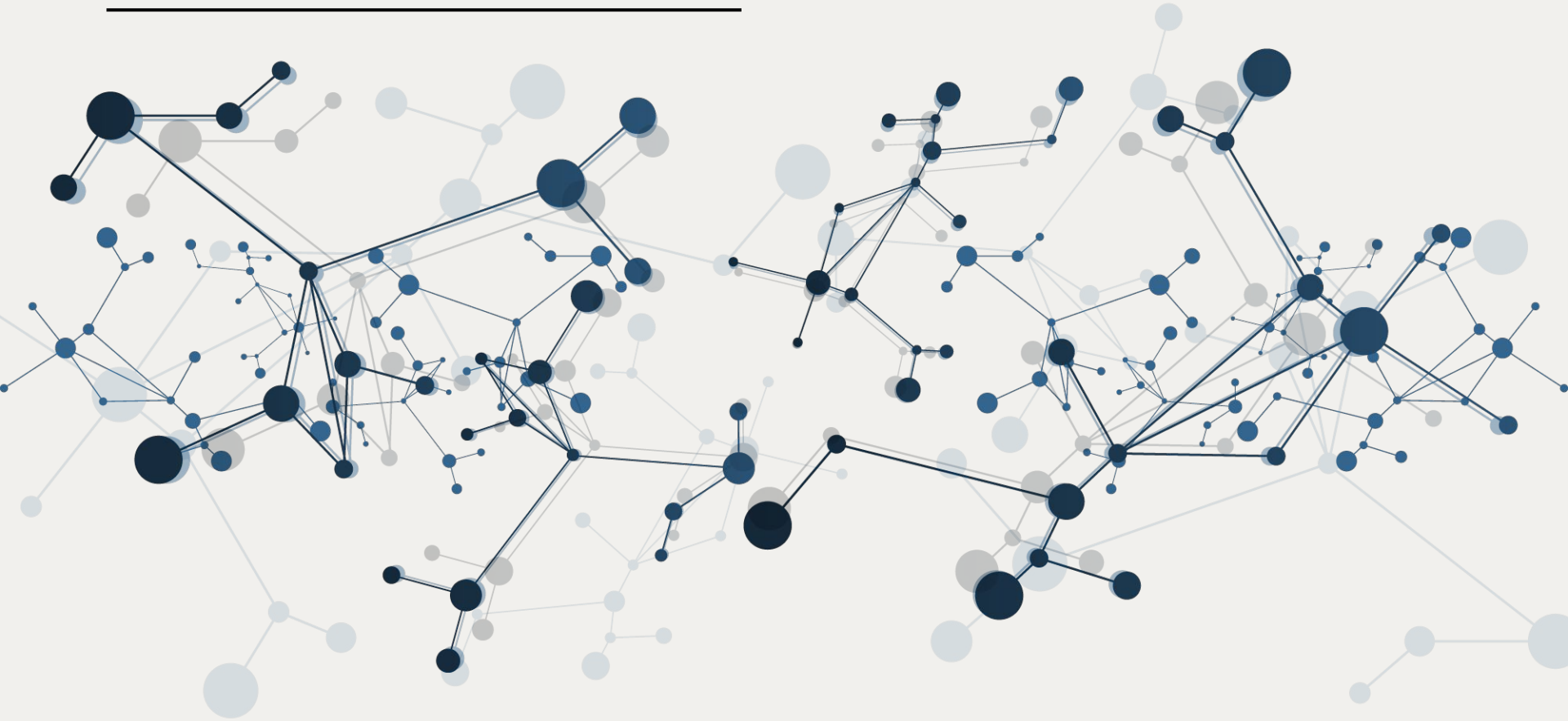
Exploitation de données massives

Id	Code postal	Age	Nationalité	Maladie
1	13053	28	Russe	Cardiaque
2	13068	29	Américain	Cardiaque
3	13068	21	Indien	Virus
...				
10	13068	37	Japonais	Cancer
11	13068	36	Américain	Cancer
12	13023	35	Américain	Cancer

Id	Dernier pays visité
1	Russie
2	États-Unis
3	Cuba
...	
10	Cuba
11	Canada
12	États-Unis

Id	Profession	Salaire annuel
1	Avocat	120,000$
2	Plombier	69,000$
3	Maire	125,000$
...		
10	Chef cuisinier	45,000$
11	Technicien	40,000$
12	Chômeur	25,000$

Exploitation de données massives



Exploitation de données massives

Partage de données: Bénéfices

- Données ouvertes
  - Ville intelligente
  - Données de géolocalisation
  - Recherche
- Partage et mutualisations de données
  - Potentiel gain en intelligence d’affaire
  - Exemple: Quartier des spectacles
- Concours
  - Netflix
  - Kaggle

Partage de données: Bénéfices

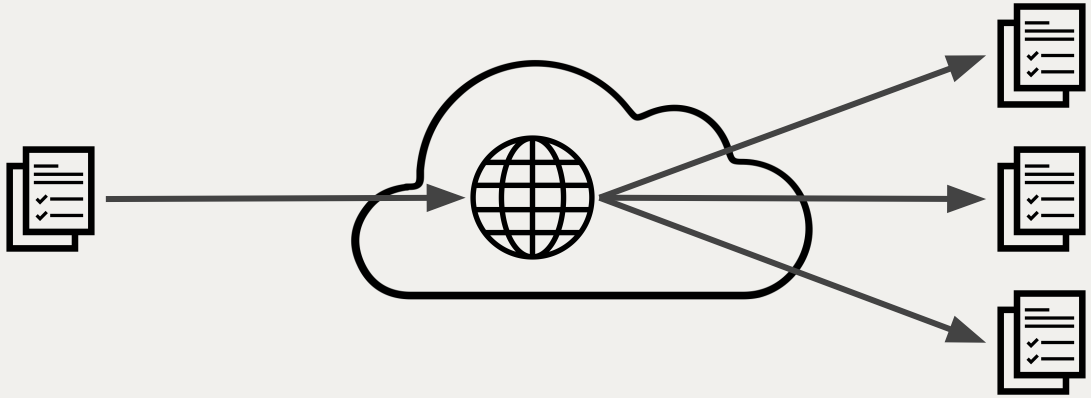
- Consultation, tiers partis et confiance partielle
  - Élections
- Données de test
  - Environnement de développement et production
- Respect de l’anonymat, valeur d’entreprise?
  - De bonnes pratiques pour un public sensibilisé

Problème 1

Les frontières de l’anonymat ne sont pas bien définies

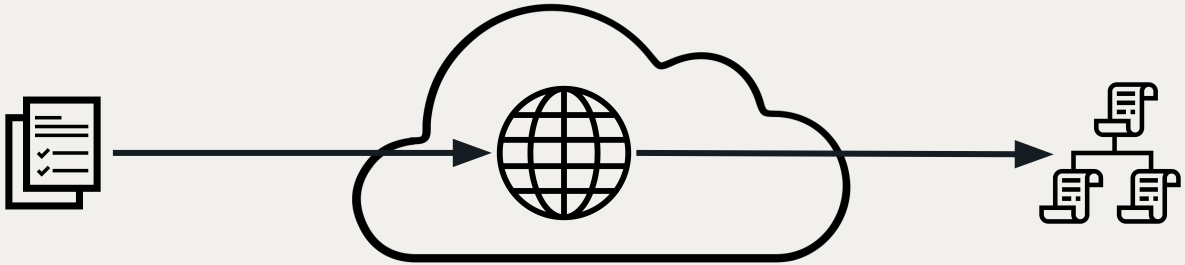
Problème 2

Une fois publiée, il est impossible de rapatrier l’information



Problème 3

Une fois publiée, une information ne peut que se composer



Contexte

Parenthèse: 33 bits

➔ On peut identifier n’importe quel humain avec seulement 33 bits d’information.

Parenthèse: 33 bits

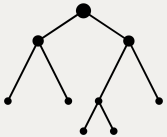
➔ On peut identifier n’importe quel humain avec seulement 33 bits d’information.

➔ 1 bit = Une question vrai/faux bien posée

Parenthèse: 33 bits

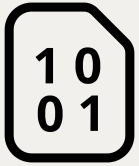
➔ On peut identifier n’importe quel humain avec seulement 33 bits d’information.

➔ 1 bit = Une question vrai/faux bien posée



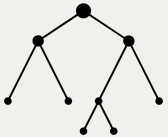
➔ 2 bits = Un choix de 4 réponses

Parenthèse: 33 bits



- ➔ On peut identifier n’importe quel humain avec seulement 33 bits d’information.

- ➔ 1 bit = Une question vrai/faux bien posée



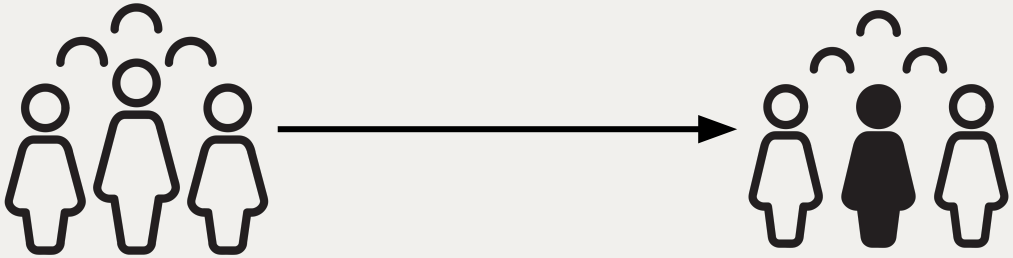
- ➔ 2 bits = Un choix de 4 réponses

- ➔ Et ainsi de suite...

Parenthèse: 33 bits

Exemple:

- ➔ Homme ou femme? 1 bit
- ➔ Code postal US: ~9-15 bits
- ➔ Date d’anniversaire: ~15 bits



Parenthèse: Contexte légal

- Droits de l'homme
  - Déclaration universelle des droits de l'homme à l'assemblée des nations unies (article 12), 1948
- Contexte canadien
  - Personal Information Protection and Electronic Documents Act (PIPEDA)
  - Loi sur la protection des renseignements personnels et les documents électroniques
  - Commissariat à la protection de la vie privée du Canada
- Contexte américain
  - Health Insurance Portability and Accountability Act of 1996 (HIPAA)
- Contexte européen
  - Réglementation Générale sur la Protection des Données (RGPD)

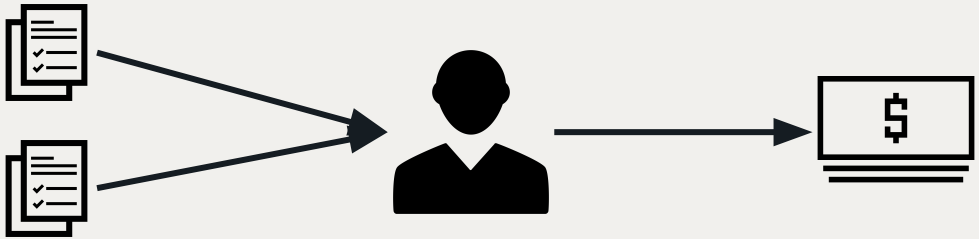
Parenthèse: Contexte légal

➔ Vendre des données clients

Parenthèse: Contexte légal

➔ Vendre des données clients

➔ Ré-identifier des individus à partir de données anonymes pour vendre les données



Concepts

Identifiants

Information nominative

Information d'identification personnelle directe

- ➔ Nom et prénom
- ➔ Adresse
- ➔ NAS
- ➔ Données biométriques

Quasi-identifiant

Information composable pour identifier un individu

- ➔ Données sociales
- ➔ Données de géolocalisation
- ➔ Données de préférences

Cycle de vie de l'information

1) Récolte d'information

Cycle de vie de l'information

1) Récolte d'information

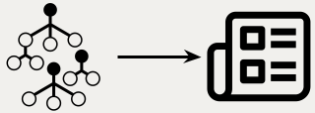
2) Analyse de l'information

Cycle de vie de l'information

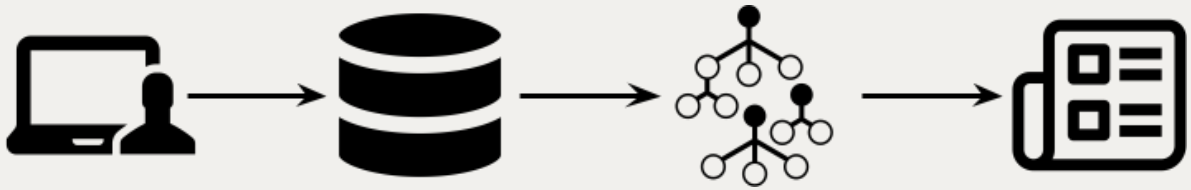
1) Récolte d'information

2) Analyse de l'information

3) Publication de résultats



Cycle de vie de l'information



Qu’est-ce que l’anonymat?

La revendication d'individus, de groupes ou d'institutions à déterminer par eux-mêmes quand, comment et dans quelle mesure les informations les concernant sont communiquées aux autres.

Qu’est-ce qu’une donnée anonyme?

Donnée sur un individu ne permettant pas d’inférer directement l’identité de l’individu.

Qu’est-ce qu’une donnée anonyme?

Donnée sur un individu ne permettant pas d’inférer directement (**ou indirectement**) l’identité de l’individu.

Qu’est-ce qu’une donnée anonyme?

Donnée sur un individu ne permettant pas d’inférer directement (**ou indirectement**) l’identité de l’individu (**avec des moyens raisonnables**).

Techniques de base

Nom	Age	Nationalité	Maladie
Igor	28	Russe	Cardiaque
Jack	29	Américain	Cardiaque
Pratesh	21	Indien	Virus

Suppression

Age	Nationalité	Maladie
28	Russe	Cardiaque
29	Américain	Cardiaque
21	Indien	Virus

Masque

Nom	Age	Nationalité	Maladie
I****	28	Russe	Cardiaque
J****	29	Américain	Cardiaque
P****	21	Indien	Virus

Techniques de base

Nom	Age	Nationalité	Maladie
Igor	28	Russe	Cardiaque
Jack	29	Américain	Cardiaque
Pratesh	21	Indien	Virus

Permutation

Nom	Age	Nationalité	Maladie
Pratesh	28	Russe	Cardiaque
Igor	29	Américain	Cardiaque
Jack	21	Indien	Virus

Pseudonymisation

Id	Age	Nationalité	Maladie
1	28	Russe	Cardiaque
2	29	Américain	Cardiaque
3	21	Indien	Virus

Techniques de base

Nom	Age	Nationalité	Maladie
Igor	28	Russe	Cardiaque
Jack	29	Américain	Cardiaque
Pratesh	21	Indien	Virus

Généralisation

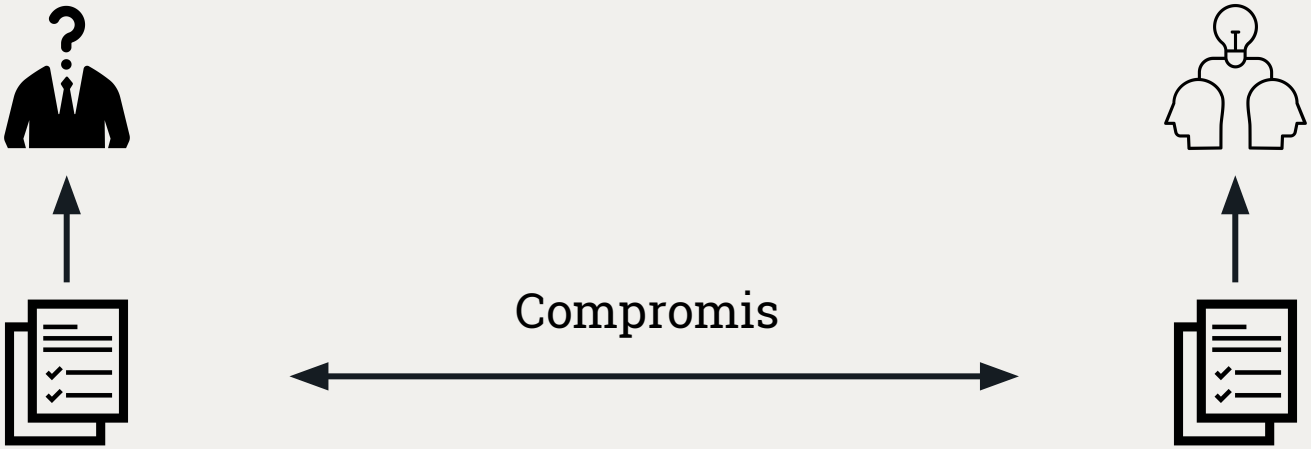
Nom	Age	Nationalité	Maladie
*	>25	Russe	Cardiaque
*	>25	Américain	Cardiaque
*	<25	Indien	Virus

Randomization

Nom	Age	Nationalité	Maladie
*	28+-2	Russe	Cardiaque
*	29+-2	Américain	Cardiaque
*	21+-2	Indien	Virus

Problème fondamental de l’anonymisation

Tout processus d’anonymisation vient nécessairement dégrader le potentiel d’extraction d’information d’un jeu de données.



Exemples		Type d'attaques
AOL	⇨	Inférence
Sweeny vs Governor Weld	⇨	Intersection
Structure dans un réseau social	⇨	Active
Prédire le déplacement	⇨	Modèle prédictif

Techniques d’assainissement

Techniques naïves

- ➔ Masquer les informations nominatives
- ➔ Pseudonymisation
- ➔ Permutation

Techniques naïves

Particulièrement inefficace contre des données fortement corrélées

- Données de géolocalisation
- Données de “graphe social”
- Recherches sur le web
- Données éparses

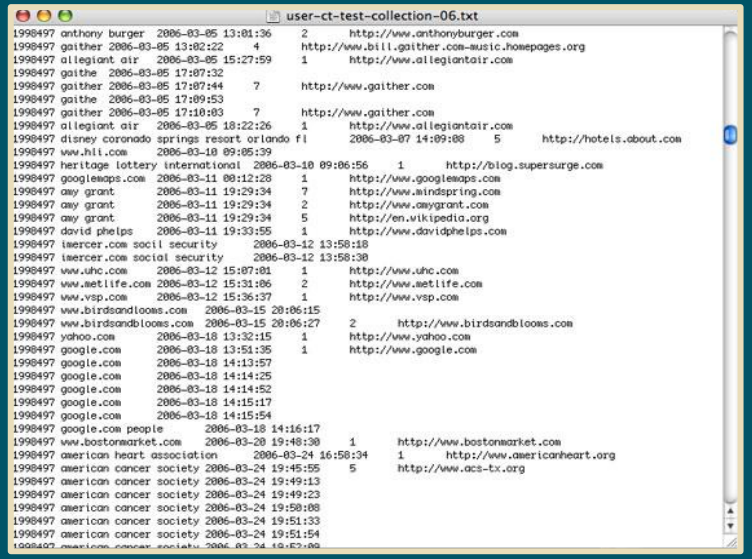
Incident AOL

AOL “pseudonymise” un ensemble de 20 millions de recherches sur 650 000 utilisateurs.

Incident d'AOL

98280	prayers to break curses	2006-04-09 5
98280	prayers for cleansing	2006-04-09 2
98280	prayers for defeating enemy	2006-04-09 1
98280	bible scriptures for defeating the enemy	2006-04-09 4
98280	prayers to plead the blood of jesus against problems	2006-04-09 2
98280	prayers to plead the blood of jesus against problems	2006-04-09 1
98280	prayers to plead the blood of jesus against problems	2006-04-09 3
98280	how does a male's cocaine use affect a fetus	2006-04-10 1
98280	how does a male's cocaine use affect a fetus	2006-04-10 5
98280	birth defects caused by father's cocaine use	2006-04-10 1
98280	birth defects caused by father's cocaine use	2006-04-10 4
98280	are chainletter scams ever successful	2006-04-10 0

https://www.somethingawful.com/weekend-web/aol-search-log/



https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Incident d'AOL

- ➔ Beaucoup de requêtes de vanités

Incident d'AOL

- ➔ Beaucoup de requêtes de vanités
- ➔ “A Face Is Exposed for AOL Searcher No. 4417749”

Incident d'AOL

- Beaucoup de requêtes de vanités
- "A Face Is Exposed for AOL Searcher No. 4417749"
- Exemples de mot clé cherchés
  - Care packages
  - Movies for dogs
  - Best dog for older owner
  - Rescue of older dogs
  - Gwinnett county yellow pages
  - Retirement in australia

The New York Times

Latanya Sweeney
vs
Gouverneur Weld

K-anonymat

Données médicales

https://en.wikipedia.org/wiki/Latanya_Sweeney

Ethnicité

Date de visite

Diagnostique

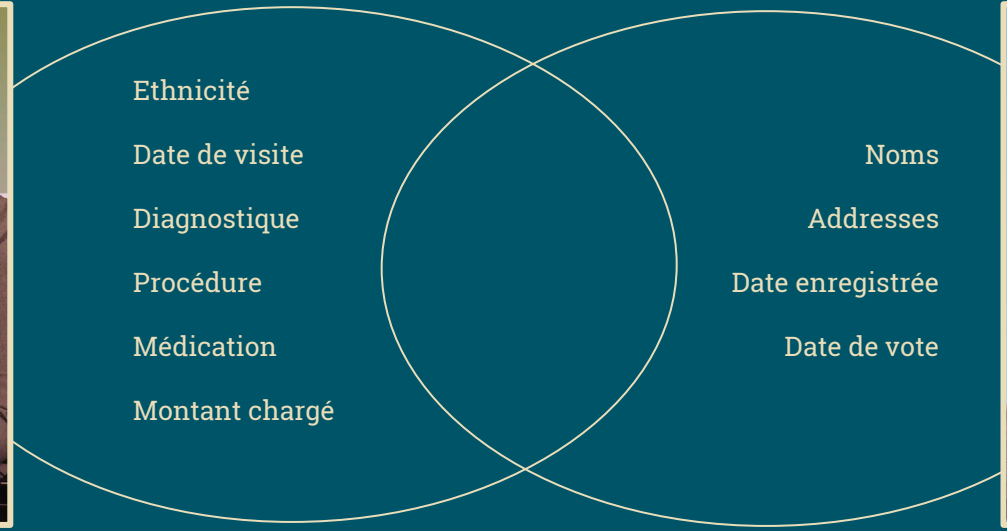
Procédure

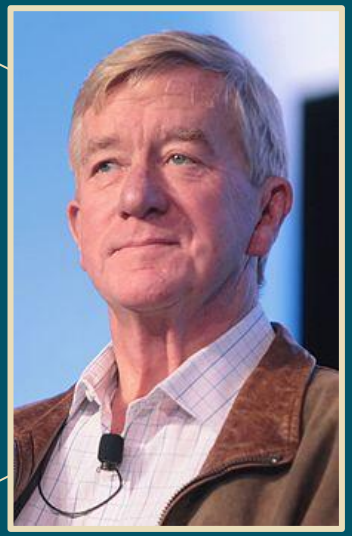
Médication

Montant chargé

K-anonymat

https://en.wikipedia.org/wiki/Latanya_Sweeney

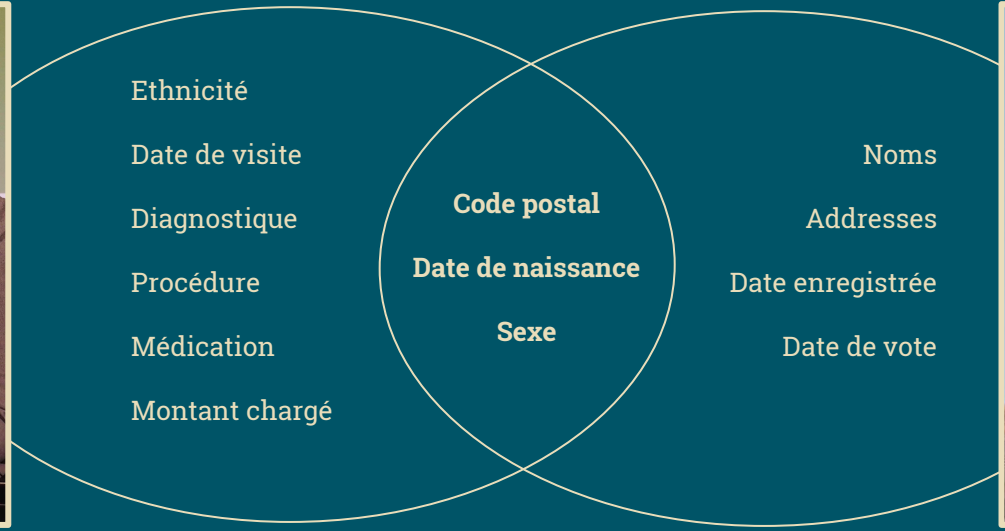


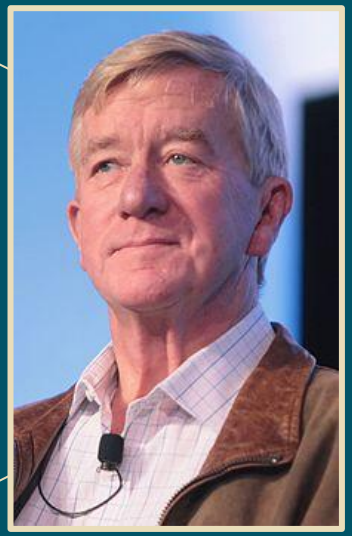


https://www.revolvy.com/main/index.php?s=William+Weld&item_type=topic

K-anonymat

https://en.wikipedia.org/wiki/Latanya_Sweeney





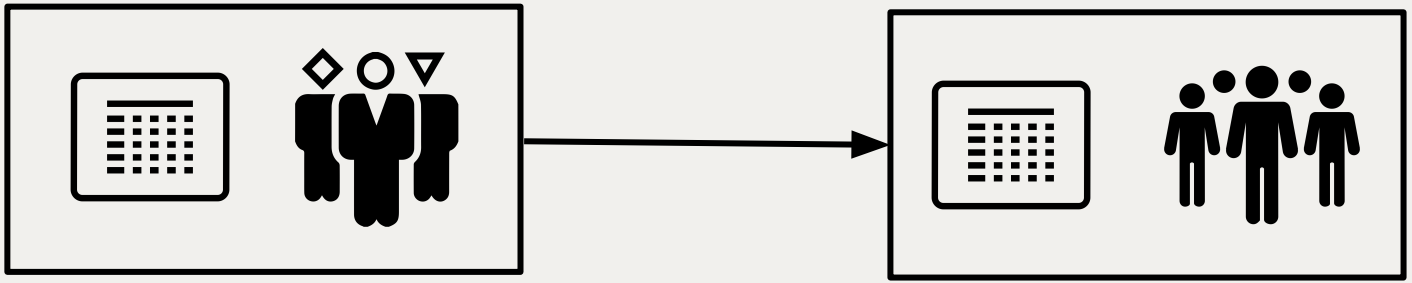
https://www.revolvy.com/main/index.php?s=William+Weld&item_type=topic

Solution

K-anonymat

K-anonymat: Solution

Pour tout groupe de données, chaque individu est indistinguable d’au moins k-1 autres individus.

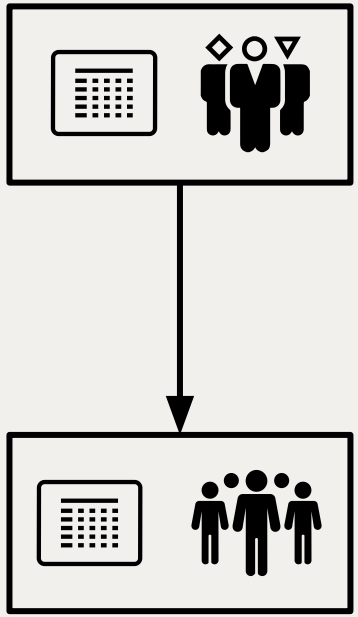


K-anonymat: Exemple

Id	Quasi-identifiant			Donnée sensible
	Code postal	Age	Nationalité	Maladie
1	13053	28	Russe	Cardiaque
2	13068	29	Américain	Cardiaque
3	13068	21	Indien	Virus
4	13053	23	Russe	Virus
5	14853	50	Américain	Cancer
6	14850	55	Américain	Cardiaque
7	14850	47	Américain	Virus
8	14823	49	Mexicain	Cardiaque
9	13023	31	Indien	Infection
10	13068	37	Japonais	Cancer
11	13068	36	Américain	Cancer
12	13023	35	Américain	Cancer

K-anonymat: Exemple

Id	Quasi-identifiant			Donnée sensible
	Code postal	Age	Nationalité	Maladie
1	13053	28	Russe	Cardiaque
2	13068	29	Américain	Cardiaque
3	13068	21	Indien	Virus
4	13053	23	Russe	Virus
5	14853	50	Américain	Cancer
6	14850	55	Américain	Cardiaque
7	14850	47	Américain	Virus
8	14823	49	Mexicain	Cardiaque
9	13023	31	Indien	Infection
10	13068	37	Japonais	Cancer
11	13068	36	Américain	Cancer
12	13023	35	Américain	Cancer



K-anonymat: Exemple

Id	Quasi-identifiant			Donnée sensible
	Code postal	Age	Nationalité	Maladie
1	13053	28	Russe	Cardiaque
2	13068	29	Américain	Cardiaque
3	13068	21	Indien	Virus
4	13053	23	Russe	Virus
5	14853	50	Américain	Cancer
6	14850	55	Américain	Cardiaque
7	14850	47	Américain	Virus
8	14823	49	Mexicain	Cardiaque
9	13023	31	Indien	Infection
10	13068	37	Japonais	Cancer
11	13068	36	Américain	Cancer
12	13023	35	Américain	Cancer

Id	Quasi-identifiant			Donnée sensible
	Code postal	Age	Nationalité	Maladie
1	130**	< 30	*	Cardiaque
2	130**	< 30	*	Cardiaque
3	130**	< 30	*	Virus
4	130**	< 30	*	Virus
5	148**	>= 45	*	Cancer
6	148**	>= 45	*	Cardiaque
7	148**	>= 45	*	Virus
8	148**	>= 45	*	Cardiaque
9	130**	3*	*	Infection
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

K-anonymat: Privacy Analytics

- ➔ Entreprise canadienne
- ➔ Fondée en 2007
- ➔ Méthodologie basée sur une analyse de risque

K-anonymat: Problèmes

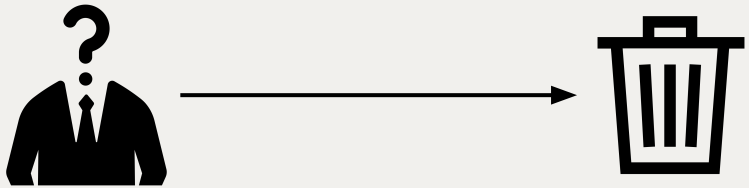
- ➔ N’est pas robuste à un changement de statut de l’information.
  - ◆ Exemple: Publication de nouvelles données
- ➔ Extraire des connaissances utiles d’un tel ensemble de données est souvent difficile
  - ◆ Compromis utilité - anonymat
- ➔ Devient impraticable lorsque les informations sont complexes ou éparses.

Article 17 du RGPD

Droit à l'oubli

Confidentialité différentielle

Étant donnée le caractère non inversible de la publication et du processus de réidentification, comment peut-on garantir un tel requis?



Le cas Netflix

Confidentialité différentielle

- ➔ Compétition d’apprentissage machine en 2006
- ➔ Objectif: Améliorer le système de recommandation
- ➔ Données anonymisées sur 480 189 utilisateurs et 17 770 films sur la période de 98 à 2005

Confidentialité différentielle

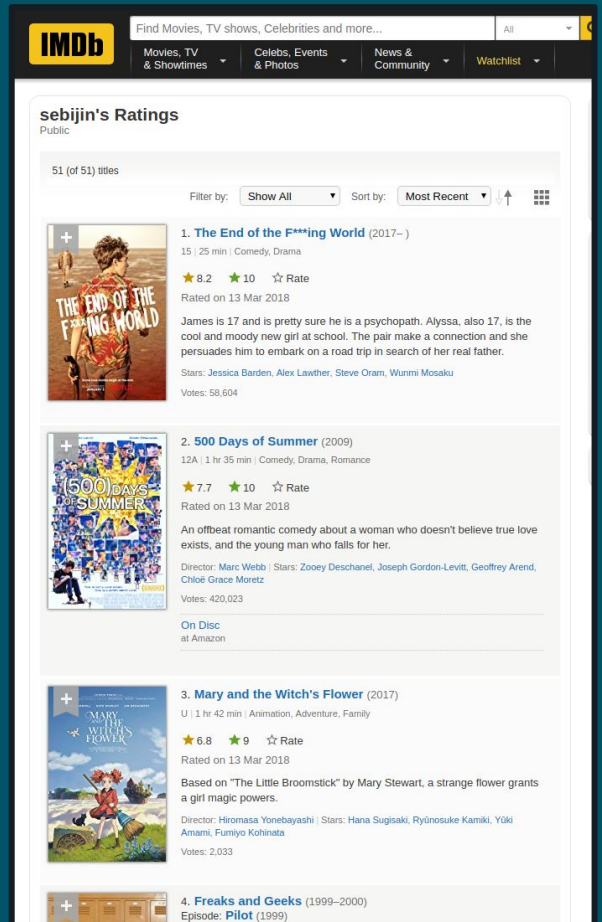
- ➔ Compétition d'apprentissage machine en 2006
- ➔ Objectif: Améliorer le système de recommandation
- ➔ Données anonymisées sur 480 189 utilisateurs et 17 770 films sur la période de 98 à 2005

	movie 1	movie 2	movie 3	...	movie 4 654	...	movie 12 820	...	movie 17 770
User 1	9				7				
User 2	3		8						
user 3					5				7
...									
User 480 189							9		

Confidentialité différentielle

Le cas Netflix

- ➔ Utilisation de données auxiliaire: l'IMDB
- ➔ 99% des utilisateurs peuvent être ré identifié avec 8 notes et les dates associées
- ➔ 68% avec 2 notes



Confidentialité différentielle: Idée

➔ Inspiré par le concept de déni plausible et la robustesse de la cryptographie.

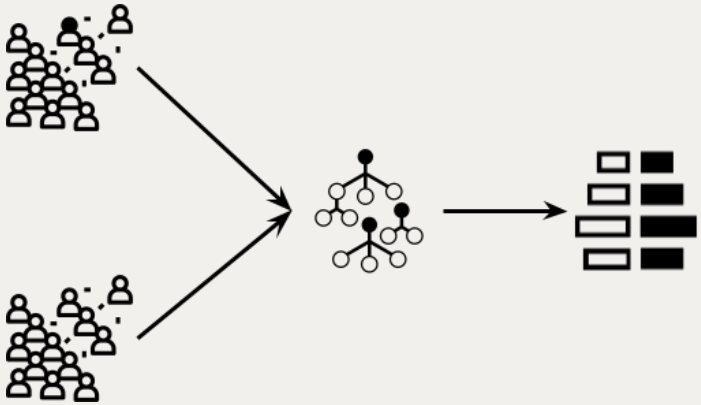
Confidentialité différentielle: Idée

- ➔ Inspiré par le concept de déni plausible et la robustesse de la cryptographie.

- ➔ **Objectif**: Garantir que la présence ou l'absence d’un individu dans un jeu de donnée n'influence pas le résultat de manière perceptible.

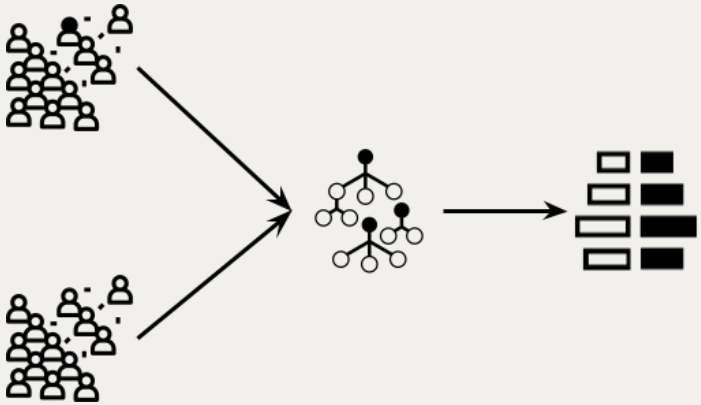
Confidentialité différentielle: Solution

Soit **A** un algorithme de requête. **D1** et **D2** n'importe quels ensembles de données qui diffèrent par un seul élément.



Confidentialité différentielle: Solution

A est différentiellement privé si toute requête de **A** sur **D1** ou **D2** ne permet pas de les distinguer avec une probabilité raisonnable



Confidentialité différentielle

➔ Ajout de bruit statistique au moment de la récolte de données ou au moment de la publication d’analyse

Confidentialité différentielle

- ➔ Ajout de bruit statistique au moment de la récolte de données ou au moment de la publication d'analyse
- ➔ Donne une garantie forte de pérennité de l'anonymat

Confidentialité différentielle

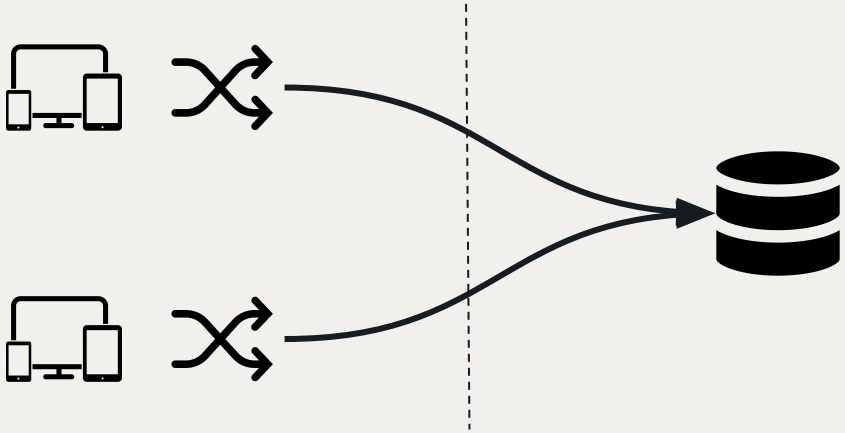
- ➔ Ajout de bruit statistique au moment de la récolte de données ou au moment de la publication d'analyse
- ➔ Donne une garantie forte de pérennité de l'anonymat
- ➔ Adoption majeure par la communauté technologique
  - ◆ Apple
  - ◆ Google
  - ◆ Uber

Exemples de confidentialité différentielle

Google

Google RAPPOR

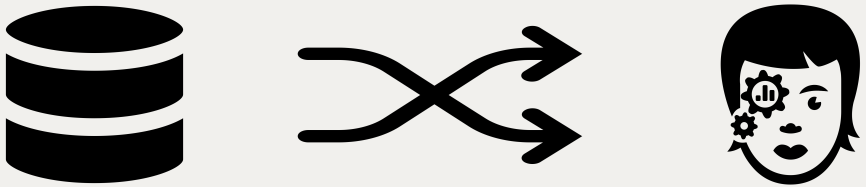
github.com/google/rappor



UBER

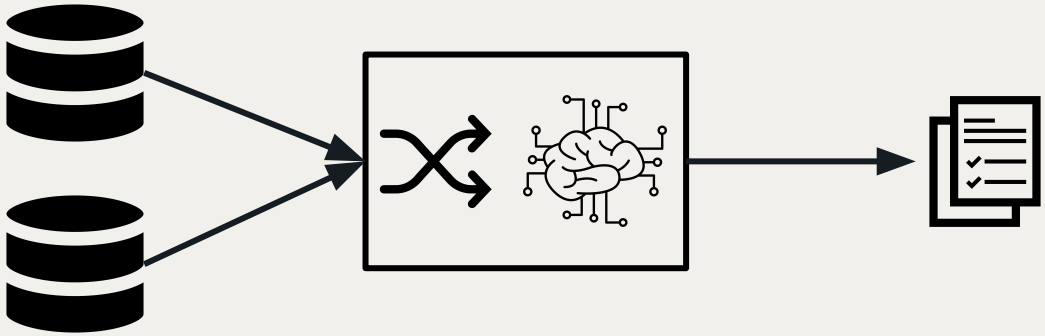
SQL Differential Privacy

github.com/uber/sql-differential-privacy



Application en intelligence artificielle

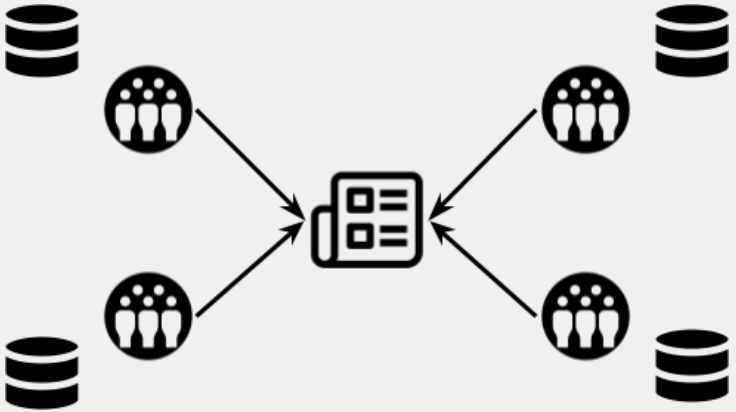
À partir d’un modèle entrainé, il peut être possible de retrouver de l’information sur les exemple d’entrainement.



Mutualisation de données

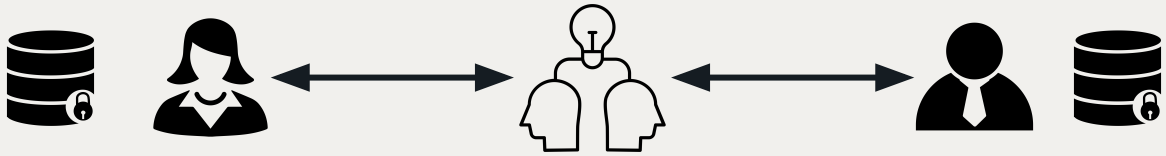
Mutualisation de données

Peut-on se partager des données de manière à conserver l'anonymat si l'on ne se fait pas totalement confiance?



Calcul multipartie

Obtenir une connaissance commune à partir de données privées



Chiffrement homomorphique

Comment utiliser les services d’un consultant honnête mais curieux.

Exemple de succès d'affaire

La suite des choses

La suite des choses

Continuer le partage de données, les bénéfices sont grands.

La suite des choses

User d’une approche prudente et informée sur la manière d’assainir les données.

La suite des choses

Besoin d’un meilleur cadre législatif.

Serge-Olivier Paquette

Chercheur et conseiller en cybersécurité

serge-olivier.paquette@delvelabs.ca