

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

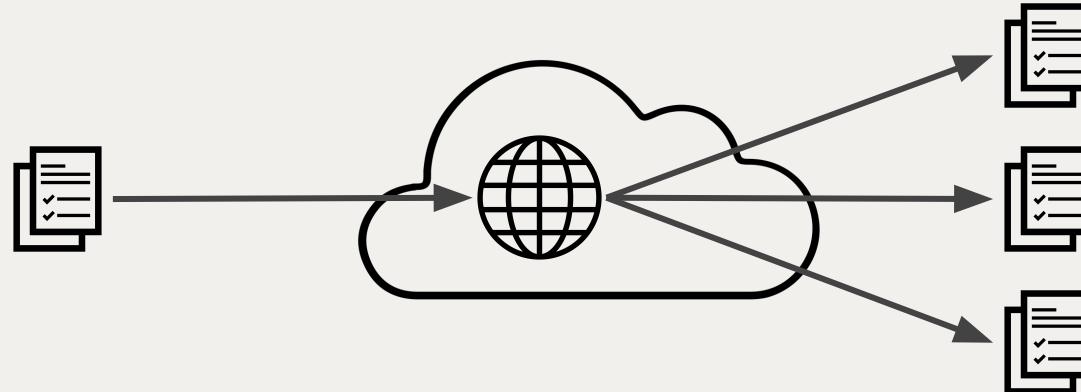
Problème 1

Les frontières de l'anonymat ne
sont pas bien définies



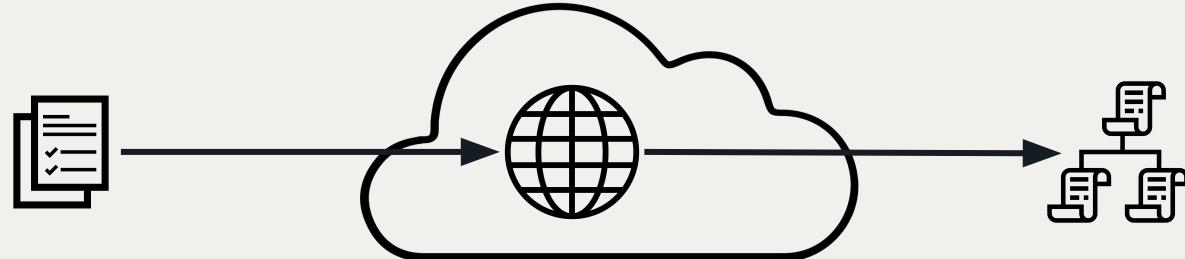
Problème 2

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

1	0
0	1

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

Parenthèse: 33 bits

1	0
0	1

- On peut identifier n'importe quel humain avec seulement 33 bits d'information.
- 1 bit = Une question vrai/faux bien posée



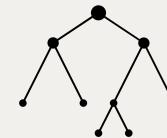
Parenthèse: 33 bits

1	0
0	1

- 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

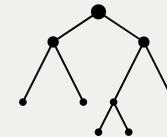
Parenthèse: 33 bits

1	0
0	1

- 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

1	0
0	1

Exemple:

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



Parenthèse: Contexte légal

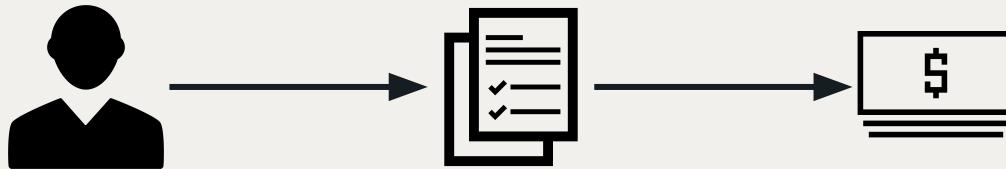


- Droits de l'homme
 - ◆ Déclaration universelle des droits de l'homme à l'assemblée des nations unies (article 12), 1948
- Contexte canadien
 - ◆ Personal Information Protection and Electronic Documents Act (PIPEDA)
 - ◆ Loi sur la protection des renseignements personnels et les documents électroniques
 - ◆ 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



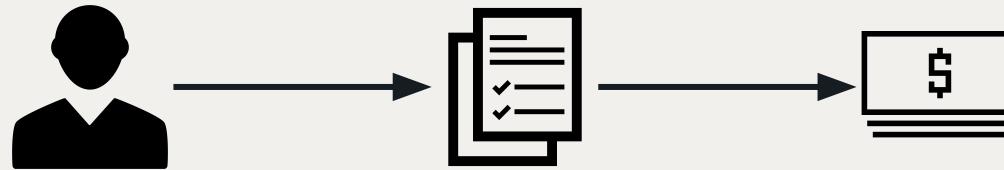
- Vendre des données clients



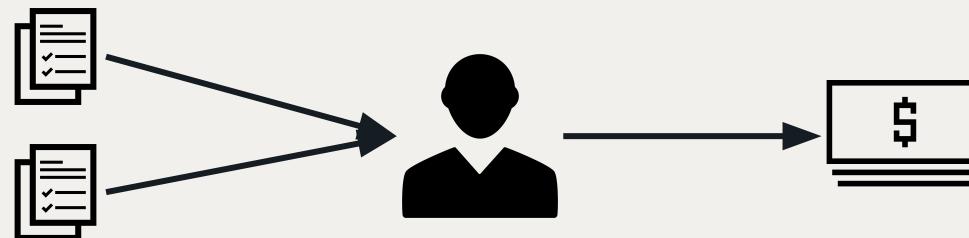
Parenthèse: Contexte légal



- 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

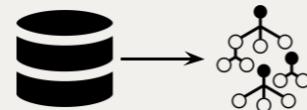


Cycle de vie de l'information

1) Récolte d'information



2) Analyse de l'information



Cycle de vie de l'information

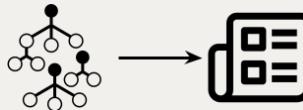
1) Récolte d'information



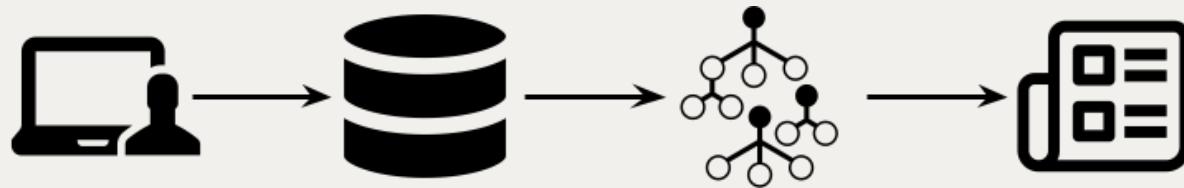
2) Analyse de l'information



3) Publication de résultats



Cycle de vie de l'information



Anonymization

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.

Anonymization

Qu'est-ce qu'une donnée anonyme?

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



Anonymization

Qu'est-ce qu'une donnée anonyme?

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



Anonymization

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

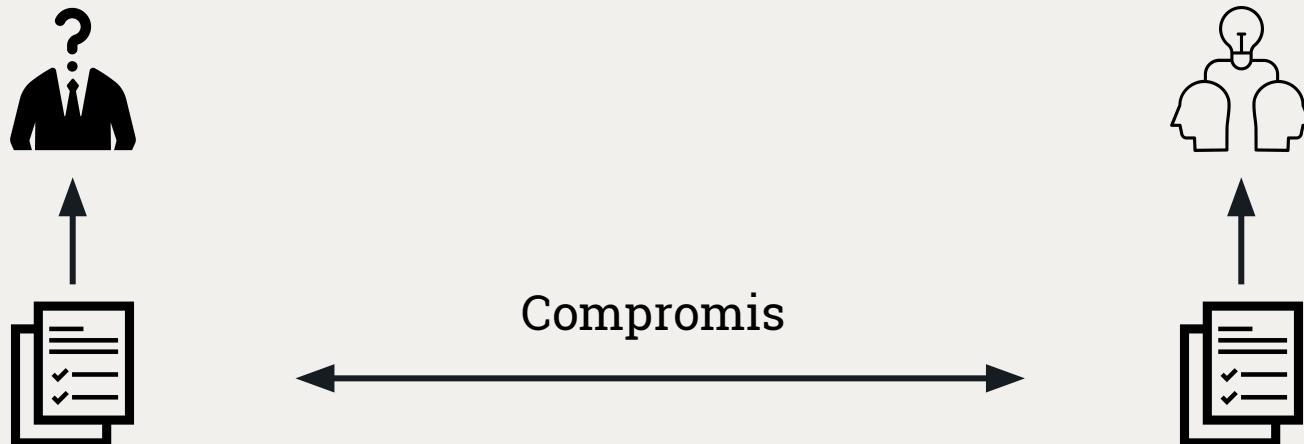
Nom	Age	Nationalité	Maladie
*	>25	Russe	Cardiaque
*	>25	Américain	Cardiaque
*	<25	Indien	Virus

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/>

user-ct-test-collection-06.txt					
1998497	anthony burger	2006-03-05 13:01:36	2	http://www.anthonyburger.com	
1998497	gaither	2006-03-05 13:02:22	4	http://www.bill.gaither.com-music.homepages.org	
1998497	allegiant air	2006-03-05 15:27:59	1	http://www.allegiantair.com	
1998497	goithe	2006-03-05 17:07:32			
1998497	goithe	2006-03-05 17:07:44	7	http://www.goither.com	
1998497	goithe	2006-03-05 17:09:53			
1998497	goithe	2006-03-05 17:10:03	7	http://www.goither.com	
1998497	allegiant air	2006-03-05 18:22:26	1	http://www.allegiantair.com	
1998497	disney coronado springs resort orlando fl	2006-03-07 14:09:08	5	http://hotels.about.com	
1998497	www.hlt.com	2006-03-18 09:05:39			
1998497	heritage lottery international	2006-03-18 09:06:56	1	http://blog.supersurge.com	
1998497	googlemaps.com	2006-03-11 08:12:28	1	http://www.googlemaps.com	
1998497	any grant	2006-03-11 19:29:34	7	http://www.windspring.com	
1998497	any grant	2006-03-11 19:29:34	2	http://www.anygrant.com	
1998497	any grant	2006-03-11 19:29:34	5	http://en.wikipedia.org	
1998497	davidphelpes	2006-03-12 19:33:55	1	http://www.davidphelpes.com	
1998497	lancer.com social security	2006-03-12 13:50:19			
1998497	lancer.com social security	2006-03-12 13:58:39			
1998497	www.shc.com	2006-03-12 15:07:03	1	http://www.shc.com	
1998497	www.setlife.com	2006-03-12 15:31:06	2	http://www.setlife.com	
1998497	www.vsp.com	2006-03-12 15:36:37	1	http://www.vsp.com	
1998497	www.birdsandblooms.com	2006-03-15 20:06:15			
1998497	www.birdsandblooms.com	2006-03-15 20:06:27	2	http://www.birdsandblooms.com	
1998497	yahoo.com	2006-03-18 13:32:15	1	http://www.yahoo.com	
1998497	google.com	2006-03-18 13:51:35	1	http://www.google.com	
1998497	google.com	2006-03-18 14:13:57			
1998497	google.com	2006-03-18 14:14:25			
1998497	google.com	2006-03-18 14:14:52			
1998497	google.com	2006-03-18 14:15:17			
1998497	google.com	2006-03-18 14:15:54			
1998497	google.com people	2006-03-18 14:16:17			
1998497	www.bostonmarket.com	2006-03-20 19:48:38	1	http://www.bostonmarket.com	
1998497	american heart association	2006-03-24 16:58:34	1	http://www.americanheart.org	
1998497	american cancer society	2006-03-24 19:45:55	5	http://www.acs-tx.org	
1998497	american cancer society	2006-03-24 19:49:13			
1998497	american cancer society	2006-03-24 19:49:23			
1998497	american cancer society	2006-03-24 19:50:08			
1998497	american cancer society	2006-03-24 19:51:33			
1998497	american cancer society	2006-03-24 19:51:54			
1998497	american cancer society	2006-03-24 19:52:02			

<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



Ethnicité
Date de visite
Diagnostique
Procédure
Médication
Montant chargé

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

K-anonymat

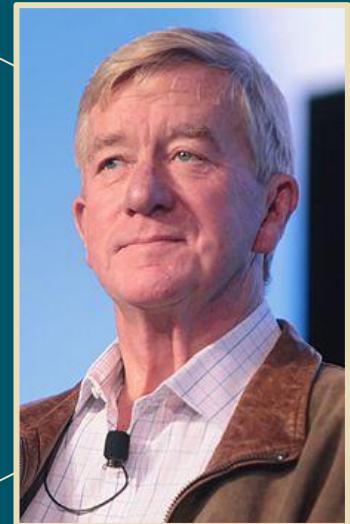
Données médicales



Ethnicité
Date de visite
Diagnostique
Procédure
Médication
Montant chargé

Données de vote

Noms
Adresses
Date enregistrée
Date de vote



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

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

K-anonymat

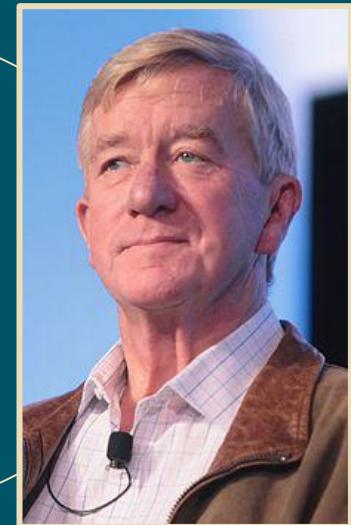
Données médicales



Ethnicité
Date de visite
Diagnostique
Procédure
Médication
Montant chargé

Données de vote

Noms
Adresses
Date enregistrée
Date de vote



Code postal
Date de naissance
Sexe

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

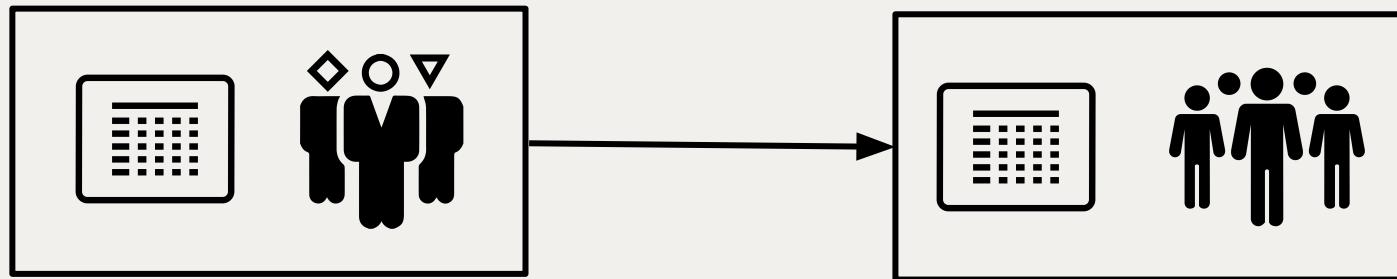
https://www.revovy.com/main/index.php?s=William+Wel&item_type=topic

Solution

K-anonymat

K-anonymat: Solution

Pour tout groupe de données, chaque individu est indistinguorable d'au moins $k-1$ autres individus.

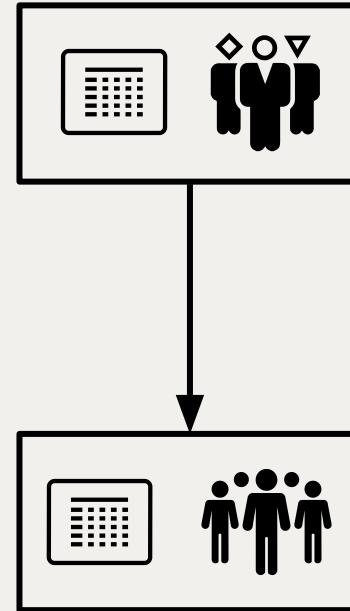


K-anonymat: Exemple

Id	Quasi-identifiant			Donnée sensible
	Code postal	Age	Nationalité	
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	
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	
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



Acquired by **imshealth™**
INTELLIGENCE. ANALYSE.

K-anonymat: Problèmes

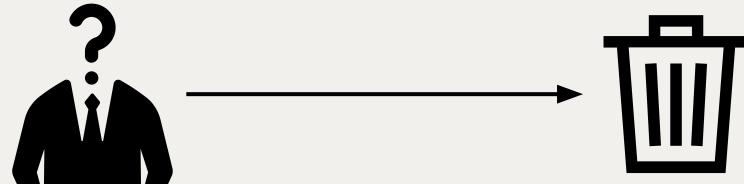
- N'est pas robuste à un changement de statut de l'information.
 - ◆ Exemple: Publication de nouvelles données
- Extraire des connaissances utiles d'un tel ensemble de données est souvent difficile
 - ◆ Compromis utilité - anonymat
- Devient impraticable lorsque les informations sont complexes ou éparses.

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			5			
user 3									7
...									
User 480 189							9		

Confidentialité différentielle

Le cas Netflix

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

The screenshot shows a user's ratings page on IMDb. The top navigation bar includes the IMDb logo, search bar, and links for 'Movies, TV & Showtimes', 'Celebs, Events & Photos', 'News & Community', and 'Watchlist'. The main content area is titled 'sebjin's Ratings' and is set to 'Public'. It displays 51 of 51 titles, filtered by 'Show All' and sorted by 'Most Recent'. The first three results are listed:

- 1. The End of the F***ing World** (2017-) 15 | 25 min | Comedy, Drama
★ 8.2 ★ 10 ★ Rate
Rated on 13 Mar 2018
James is 17 and is pretty sure he is a psychopath. Alyssa, also 17, is the cool and moody new girl at school. The pair make a connection and she persuades him to embark on a road trip in search of her real father.
Stars: Jessica Barden, Alex Lawther, Steve Oram, Wunmi Mosaku
Votes: 58,694
- 2. 500 Days of Summer** (2009) 12A | 1 hr 35 min | Comedy, Drama, Romance
★ 7.7 ★ 10 ★ Rate
Rated on 13 Mar 2018
An offbeat romantic comedy about a woman who doesn't believe true love exists, and the young man who falls for her.
Director: Marc Webb | Stars: Zooey Deschanel, Joseph Gordon-Levitt, Geoffrey Arend, Chloe Grace Moretz
Votes: 420,023
On Disc at Amazon
- 3. Mary and the Witch's Flower** (2017) U | 1 hr 42 min | Animation, Adventure, Family
★ 6.8 ★ 9 ★ Rate
Rated on 13 Mar 2018
Based on "The Little Broomstick" by Mary Stewart, a strange flower grants a girl magic powers.
Director: Hiromasa Yonebayashi | Stars: Hana Sugisaki, Ryūnosuke Kamiki, Yuki Amami, Fumiyo Kohinata
Votes: 2,033
- 4. Freaks and Geeks** (1999–2000)
Episode: Pilot (1999)

Confidentialité différentielle: Idée

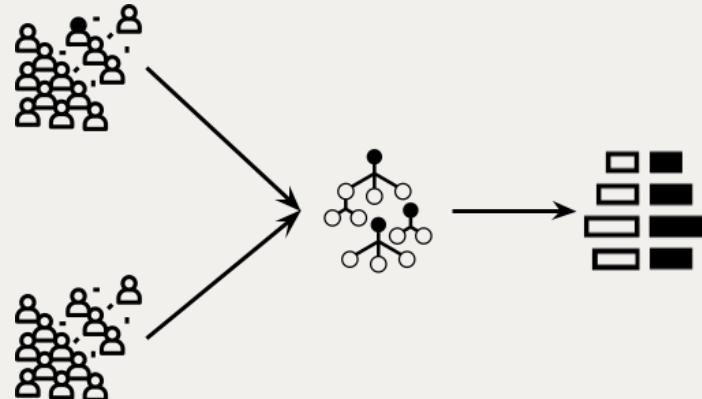
- Inspiré par le concept de déni plausible et la robustesse de la cryptographie.

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 données n'influence pas le résultat de manière perceptible.

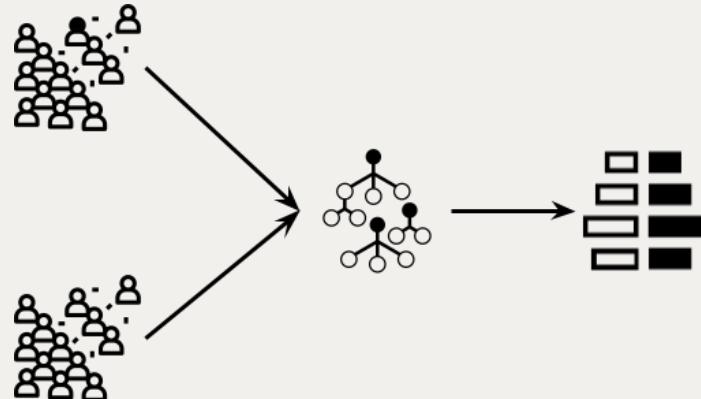
Confidentialité différentielle: Solution

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

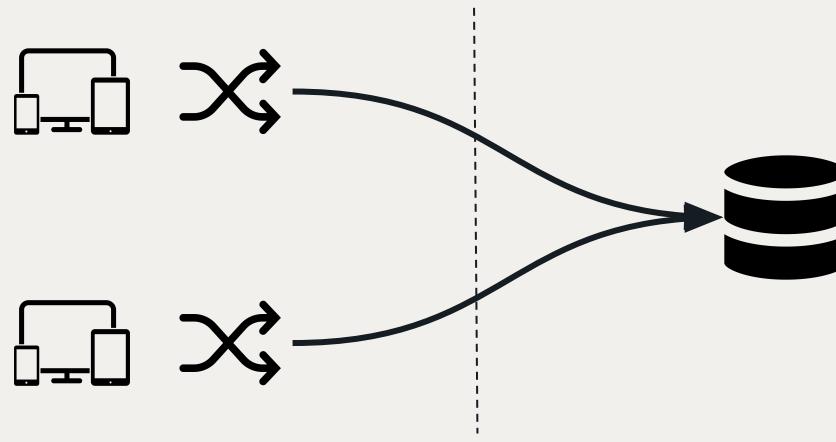
Confidentialité différentielle

- 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 RAPPOR

github.com/google/rappor



UBER

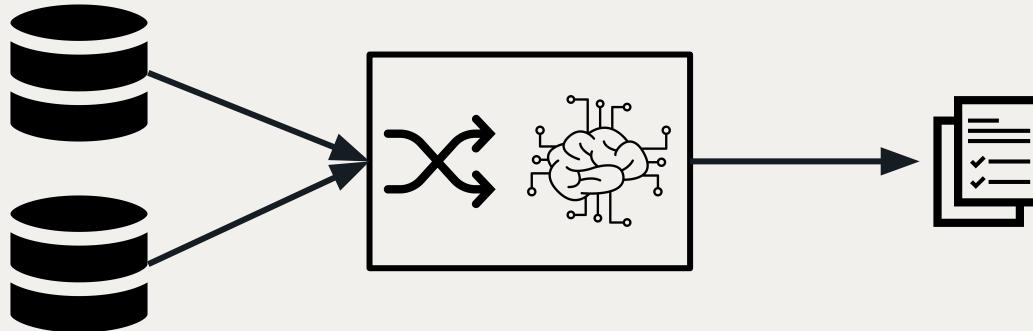
SQL Differential Privacy

github.com/uber/sql-differential-privacy



Application en intelligence artificielle

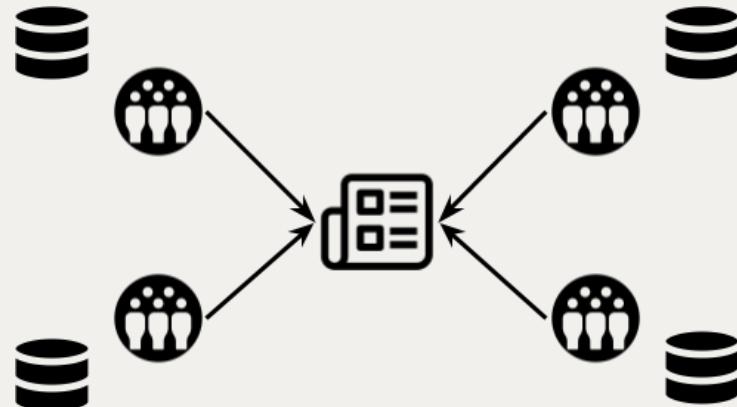
À partir d'un modèle entraîné, il peut être possible de retrouver de l'information sur les exemples d'entraînement.



Mutualisation de données

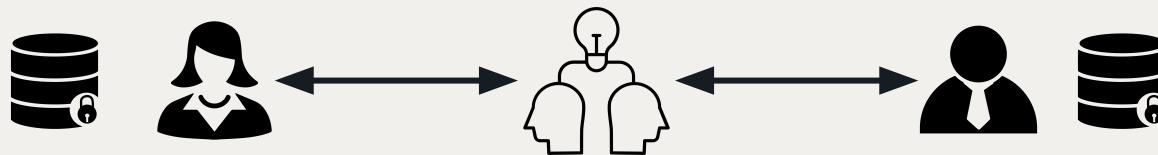
Mutualisation de données

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 homomorphe

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