PROTECTION STATISTIQUE DE LA CONFIDENTIALITÉ DES DONNÉES

Anne-Sophie Charest
Professeure adjointe, Département de mathématiques et statistique
Université Laval

Big Data, le défi du traitement des données

Université Laval, 29 octobre 2015

LE PROBLÈME DE LA CONFIDENTIALITÉ

DEUX OBJECTIFS CONTRADICTOIRES

DEUX OBJECTIFS CONTRADICTOIRES

Promesse de confidentialité

Vos réponses seront conservées confidentielles et ne seront utilisées qu'à des fins statistiques.

DEUX OBJECTIFS CONTRADICTOIRES

Promesse de confidentialité

Vos réponses resteront confidentielles et ne seront utilisées qu'à des fins statistiques.

Utilité des données

- Des rapports sont publiés à partir des données récoltées.
- Parfois, certaines données sont accessibles aux chercheurs ou à la population en général.

DEUX OBJECTIFS CONTRADICTOIRES

Promesse de confidentialité

Vos réponses resteront confidentielles et ne seront utilisées qu'à des fins statistiques.

Utilité des données

- Des rapports sont publiés à partir des données récoltées.
- Parfois, certaines données sont accessibles aux chercheurs ou à la population en général.

Comment peut-on s'assurer de respecter notre promesse de confidentialité tout en utilisant les données collectées?

SOLUTION INTUITIVE:

Anonymisation des données

i.e. enlever toutes les variables qui pourraient permettre d'identifier directement le répondant (nom, NAS, adresse, …)

SOLUTION INTUITIVE:

Anonymisation des données

i.e. enlever toutes les variables qui pourraient permettre d'identifier directement le répondant (nom, NAS, adresse, …)

Ce n'est malheureusement pas suffisant…

PREMIER EXEMPLE

- Informations médicales sur 135 000 employés de l'état du Massachusetts.
- Version anonyme partagée pour la recherche.
- Aucune information personnelle, mais certaines caractéristiques individuelles.

PREMIER EXEMPLE

- Informations médicales sur 135 000 employés de l'état du Massachusetts.
- Version anonyme partagée pour la recherche.
- Aucune information personnelle, mais certaines caractéristiques individuelles.

À l'aide d'une liste des voteurs, Dr. Latyana Sweeney identifie WilliamWeld, alors gouverneur de l'état, et obtient donc accès à son historique medical.

PREMIER EXEMPLE (SWEENEY,2000)

- Informations médicales sur 135 000 employés de l'état du Massachusetts.
- Version anonyme partagée pour la recherche.
- Aucune information personnelle, mais certaines caractéristiques individuelles.

À l'aide d'une liste des voteurs, Dr. Latyana Sweeney identifie WilliamWeld, alors gouverneur de l'état, et obtient donc accès à son historique medical.

« According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code." »

SUITES

- Recensement américain de 1990 :
  87% des gens pourraient être identifiés ainsi

SUITES

- Recensement américain de 1990 :
  87% des gens pourraient être identifiés ainsi
- Celui de 2000 :
  l'estimation est de 63 %

SUITES

- Recensement américain de 1990 :

  87% des gens pourraient être identifiés ainsi

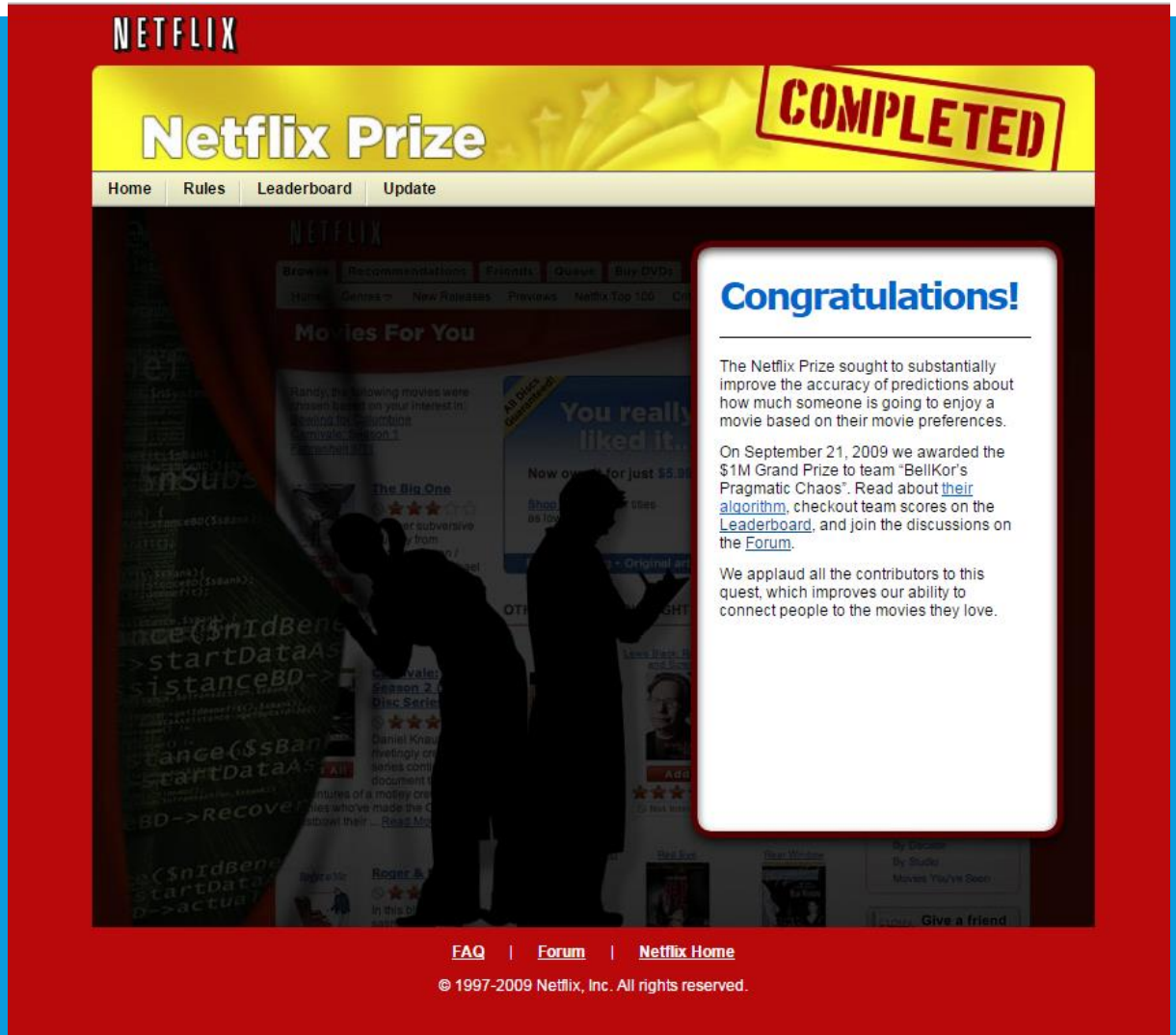
- Celui de 2000 :

  l'estimation est de 63 %

- Identification de participants à un projet sur le génome humain.

  Sweeney L, Abu A, and Winn J. Identifying Participants in the Personal Genome Project by Name. Harvard University. Data Privacy Lab. White Paper 1021-1. April 24, 2013.

AUTRE EXAMPLE – CONCOURS NETFLIX



LE PROBLÈME

Lesbian Sues Netflix Amid Privacy Concerns

1:31 PM - December 18, 2009 - By Jane McEntegart - Source : Tom's Guide US

Like 0 Send Twitter 0 0 StumbleUpon 31 Share 31

A lesbian is suing Netflix amid concerns that the company did not do enough to ensure user data would remain anonymous once released and made available to the public.

Wired reports that the mother of two is suing the movie rental company alleging Netflix made it possible for her to be "outed" by disclosing insufficiently anonymous information about nearly half-a-million customers. The information was disclosed as part of the company's bid to find a more reliable recommendation system for customers.

Zoom

When Netflix released the 100 million movie ratings, along with the date of the rating the company assigned a unique ID number to the subscriber, and the movie information. However, according to Wired, two Texas University students quickly identified a number of Netflix subscribers by comparing their “anonymous” reviews in the data to ones posted on IMDb.

CONSÉQUENCE

3/12/2010 @ 12:35PM | 1,417 views

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley, Contributor

+ Comment now

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

The lawsuit called attention to academic research that suggests that Netflix indirectly exposed the movie preferences of its users by publishing anonymized customer data. In the suit, plaintiff Paul Navarro and others sought an injunction preventing Netflix from going through the so-called "Netflix Prize II," a follow-up challenge that Netflix promised would offer up

DEUXIÈME ESSAI:

Ne publier que des données agrégées

DEUXIÈME ESSAI:

Ne publier que des données agrégées

Encore une fois, ce n’est pas suffisant.

DEUXIÈME ESSAI:

Ne publier que des données agrégées

Encore une fois, ce n'est pas suffisant.

Pensez par exemple à un tableau avec une catégorie complètement vide, ou à une cellule avec un seul individu.

DEUXIÈME ESSAI:

Ne publier que des données agrégées

Encore une fois, ce n'est pas suffisant.

Pensez par exemple à un tableau avec une catégorie complètement vide, ou à une cellule avec un seul individu.

Ou encore à des données beaucoup plus compliquées…

EXAMPLE - ÉTUDE D'ASSOCIATION PANGÉNOMIQUE (GWAS)

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1*]

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, **2** University of California Los Angeles, Los Angeles, California, United States of America

Abstract

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

EXAMPLE - ÉTUDE D'ASSOCIATION PANGÉNOMIQUE (GWAS)

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1*]

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, **2** University of California Los Angeles, Los Angeles, California, United States of America

Abstract

We use h…nonstrate the ability…omic DNA mixture. …within a mixture, t…y demonstr…individuals within a s… contributo…utility of SNPs for …uture research …sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies.

EXAMPLE - ÉTUDE D'ASSOCIATION PANGÉNOMIQUE (GWAS)

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1*]

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, **2** University of California Los Angeles, Los Angeles, California, United States of America

Abstract

We use h… …onstrate the ability… …omic DNA mixture. … …ithin a mixture, t… …y demonstr… …individuals within a s… contribut… …utility of SNPs for … …uture research … … sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies.

Suite à la parution de cet article en 2008, le National Institute of Health (NIH) a enlevé toutes les données génomiques agrégées de son site internet.

QU'EST-CE QU'ON FAIT ALORS?

PLUSIEURS APPROCHES

- Limiter l'accès aux données.
  - Autoriser seulement les chercheurs d'une institution reconnue avec un projet de recherche sérieux à utiliser les données.
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus.

PLUSIEURS APPROCHES

- Limiter l'accès aux données.
  - Autoriser seulement les chercheurs d'une institution reconnue avec un projet de recherche sérieux à utiliser les données.
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus.

- *Confidentialiser* les jeux de données, données agrégées (tableaux) ou sorties statistiques
  - Avant la publication
  - À l'aide d'un logiciel statistique qui donne accès aux données

PLUSIEURS APPROCHES

- Limiter l'accès aux données.
  - Autoriser seulement les chercheurs d'une institution reconnue avec un projet de recherche sérieux à utiliser les données.
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus.

- *Confidentialiser* les jeux de données, données agrégées (tableaux) ou sorties statistiques
  - Avant la publication
  - À l'aide d'un logiciel statistique qui donne accès aux données

Objectif : Maximiser l'utilité tout en minimisant le risque de divulgation.

CONFIDENTIALISER?

- Méthodes de réduction (non-perturbatives)

- Méthodes perturbatives

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  -
  -
  -
- **Méthodes perturbatives**

Table 3

Business structure of FFNHP establishments by sub-sector, 2011

Symbols | Next table | Previous table

	Functional food establishments [1]	Natural health product establishments [2]	Establishments active in both fields [3]	Service only establishments	All establishments
	number				
Business structure					
Private corporation	95	387	121	58	660
Publicly traded corporation	x	9	12	x	38
Sole proprietorship	0	x	x	x	19
Unincorporated partnership	x	x	0	0	x
Cooperative	10	6 E	7	0	23
Other	0	x	x	0	x
Establishments with head offices located in Canada	108	419	140	65	732
Establishments who are a subsidiary of a multi-national enterprise	41	50	x	x	109

1. Functional Food establishments: Includes Functional Food establishments with or without a service-providing component.
2. Natural Health Product establishments: Includes Natural Health Product establishments with or without a service-providing component.
3. Functional Food and Natural Health Product establishments: Includes establishments involved in Functional Food and Natural Health Products with or without a service-providing component.

Note(s): Totals may not add up due to rounding.

Source(s): Statistics Canada, The Functional Foods and Natural Health Products Survey, 2011.

x	suppressed to meet the confidentiality requirements of the *Statistics Act*
E	use with caution

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  -
  -
- **Méthodes perturbatives**

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  -
- **Méthodes perturbatives**

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique

- **Méthodes perturbatives**

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique
- **Méthodes perturbatives**
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  -
  -
  -
  -

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique
- **Méthodes perturbatives**
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  - Échanger les valeurs de certaines variables entre des répondants
  -
  -
  -

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique
- **Méthodes perturbatives**
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  - Échanger les valeurs de certaines variables entre des répondants
  - Ajouter du bruit aléatoire aux données
  -
  -

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique

- **Méthodes perturbatives**
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  - Échanger les valeurs de certaines variables entre des répondants
  - Ajouter du bruit aléatoire aux données
  - Arrondir les fréquences dans un tableau
  -

CONFIDENTIALISER?

- **Méthodes de réduction (non-perturbatives)**
  - Masquer la valeur de certaines cellules dans un tableau de valeurs
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique

- **Méthodes perturbatives**
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  - Échanger les valeurs de certaines variables entre des répondants
  - Ajouter du bruit aléatoire aux données
  - Arrondir les fréquences dans un tableau
  - Créer des jeux de données complètement synthétiques

LA PROMESSE DE CONFIDENTIALITÉ

QU'EST-CE QU'ON PROMET EXACTEMENT?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

QU'EST-CE QU'ON PROMET EXACTEMENT?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

- Dalenius 1977: Soit D_K la valeur de la caractéristique D pour l'individu K.

  « Si la publication d'une statistique S permet de déterminer la valeur D_K plus précisément que sans accès à S, alors une divulgation a eu lieu »

QU'EST-CE QU'ON PROMET EXACTEMENT?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

- Dalenius 1977: Soit D_K la valeur de la caractéristique D pour l'individu K.

  « Si la publication d'une statistique S permet de déterminer la valeur D_K plus précisément que sans accès à S, alors une divulgation a eu lieu »

- Les conclusions statistiques seraient en elles-mêmes une divulgation!

UN COMPROMIS INTÉRESSANT: LA CONFIDENTIALITÉ DIFFÉRENTIELLE

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**

UN COMPROMIS INTÉRESSANT: LA CONFIDENTIALITÉ DIFFÉRENTIELLE

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**

- Protection **rigoureusement mesurable** de la confidentialité des données.

UN COMPROMIS INTÉRESSANT: LA CONFIDENTIALITÉ DIFFÉRENTIELLE

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**
- Protection **rigoureusement mesurable** de la confidentialité des données.
- Le mécanisme de protection est complètement public.

DÉFINITION

(DWORK ET AL., 2006)

Une fonction randomisée κ garantie la confidentialité différentielle de niveau ϵ si et seulement si pour tous jeux de données voisins D_1 et D_2 et pour tout $S \in Image(\kappa)$,

$$e^{-\epsilon} \leq \frac{Pr(\kappa(D_1) \in S)}{Pr(\kappa(D_2) \in S)} \leq e^{\epsilon}$$

AUTRES APPROCHES

- Estimer la probabilité de pouvoir ré-identifier un individu dans la base de données.

- Estimer la precision avec laquelle un adversaire peut reconstituer votre information à l'aide d'un jeu de données.

(nécessitent certaines hypothèse sur l'adversaire, le jeu de données, les autres informations disponibles…)

LE TRAVAIL DU STATISTICIEN

- Développer des méthodes pour la publication de statistiques et de résultats obtenus à partir de données confidentielles

LE TRAVAIL DU STATISTICIEN

- Développer des méthodes pour la publication de statistiques et de résultats obtenus à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité

LE TRAVAIL DU STATISTICIEN

- Développer des méthodes pour la publication de statistiques et de résultats obtenus à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité
- Évaluer l'effet de la protection des données sur la qualité des analyses statistiques

LE TRAVAIL DU STATISTICIEN

- Développer des méthodes pour la publication de statistiques et de résultats obtenus à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité
- Évaluer l'effet de la protection des données sur la qualité des analyses statistiques
- Développer des méthodes pour analyser les données publiées si elles ont été modifiées pour protéger la confidentialité

EN BREF

QUOI RETENIR?

- Le problème de la confidentialité des données est important pour chacun
  - en tant que citoyens partageant ses données,
  - et chercheur utilisant des jeux de données confidentiels.
-
-
-

QUOI RETENIR?

- Le problème de la confidentialité des données est important pour chacun
  - en tant que citoyens partageant ses données,
  - et chercheur utilisant des jeux de données confidentiels.
- Les solutions intuitives pour la protection de la confidentialité ne sont pas suffisantes.
-
-

QUOI RETENIR?

- Le problème de la confidentialité des données est important pour chacun
  - en tant que citoyens partageant ses données,
  - et chercheur utilisant des jeux de données confidentiels.
- Les solutions intuitives pour la protection de la confidentialité ne sont pas suffisantes.
- Définir/mesurer la protection de la confidentialité n'est pas simple.
-

QUOI RETENIR?

- Le problème de la confidentialité des données est important pour chacun
  - en tant que citoyens partageant ses données,
  - et chercheur utilisant des jeux de données confidentiels.
- Les solutions intuitives pour la protection de la confidentialité ne sont pas suffisantes.
- Définir/mesurer la protection de la confidentialité n'est pas simple.
- La statistique est essentielle pour approcher ce problème.

EN SAVOIR PLUS

anne-sophie.charest@mat.ulaval.ca