PRIVACY HORIZONS: TERRA INCOGNITA

29th International Conference of Data Protection and Privacy Commissioners

September 25 to 28, 2007
Montreal, Canada

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE : TERRA INCOGNITA

29e Conférence internationale des commissaires à la protection des données et de la vie privée

du 25 au 28 septembre 2007
Montréal, Canada

M^me^ Ann Cavoukian, Ph. D.

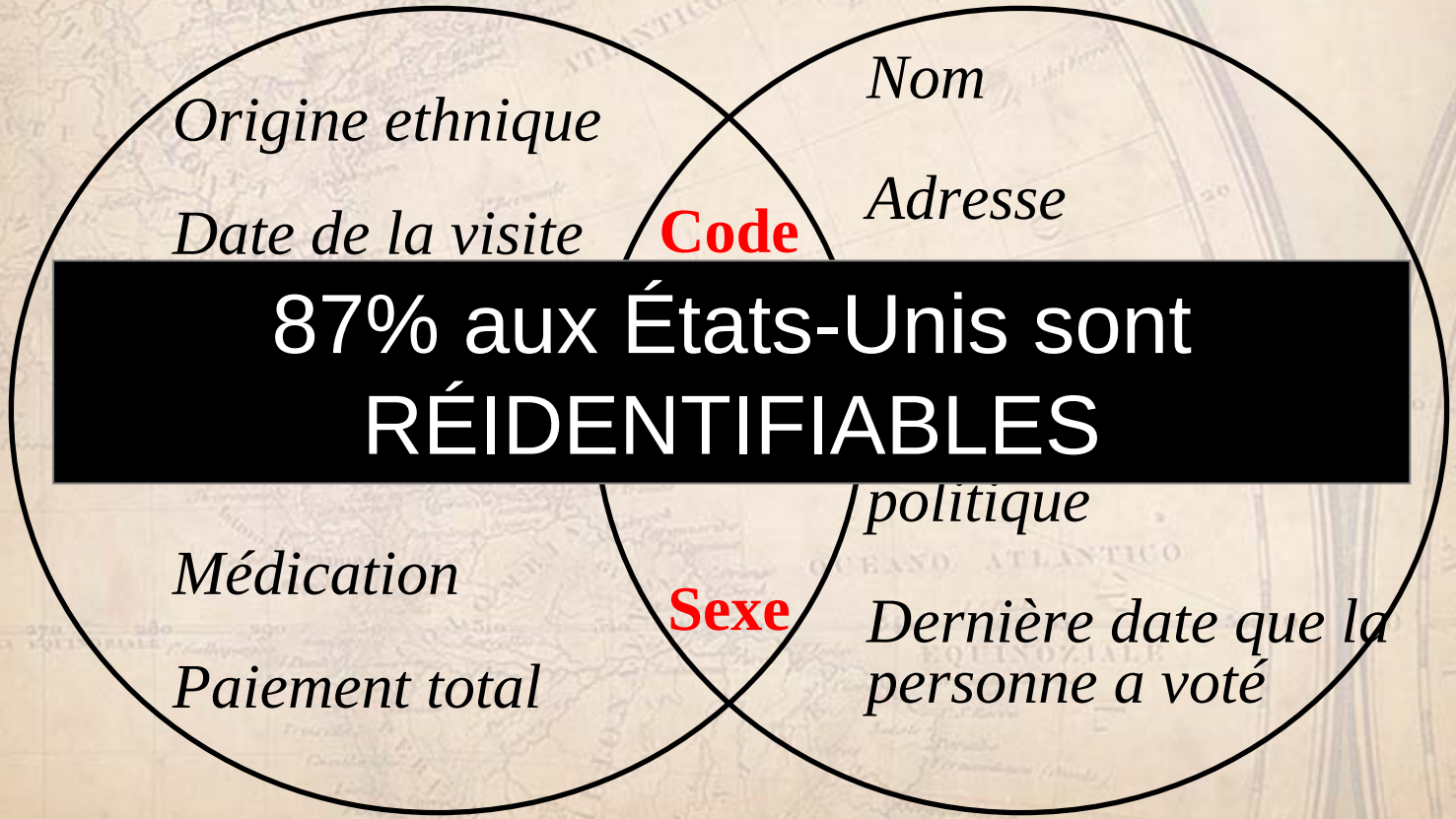
Commissaire à l'information et à la protection de la vie privée de l'Ontario

Désidentification des données, risques et resolution

Bradley Malin, Ph.D.
Professeur adjoint
Vanderbilt University

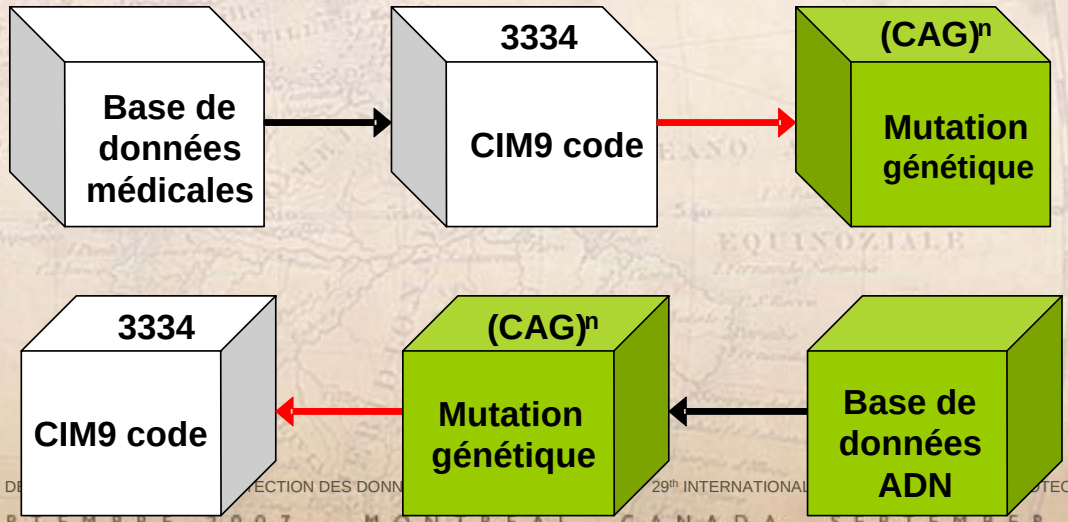
Désidentifié ne veut pas dire anonyme

(*Sweeney 1998, 2000*)



Réidentification par empreintes génétiques

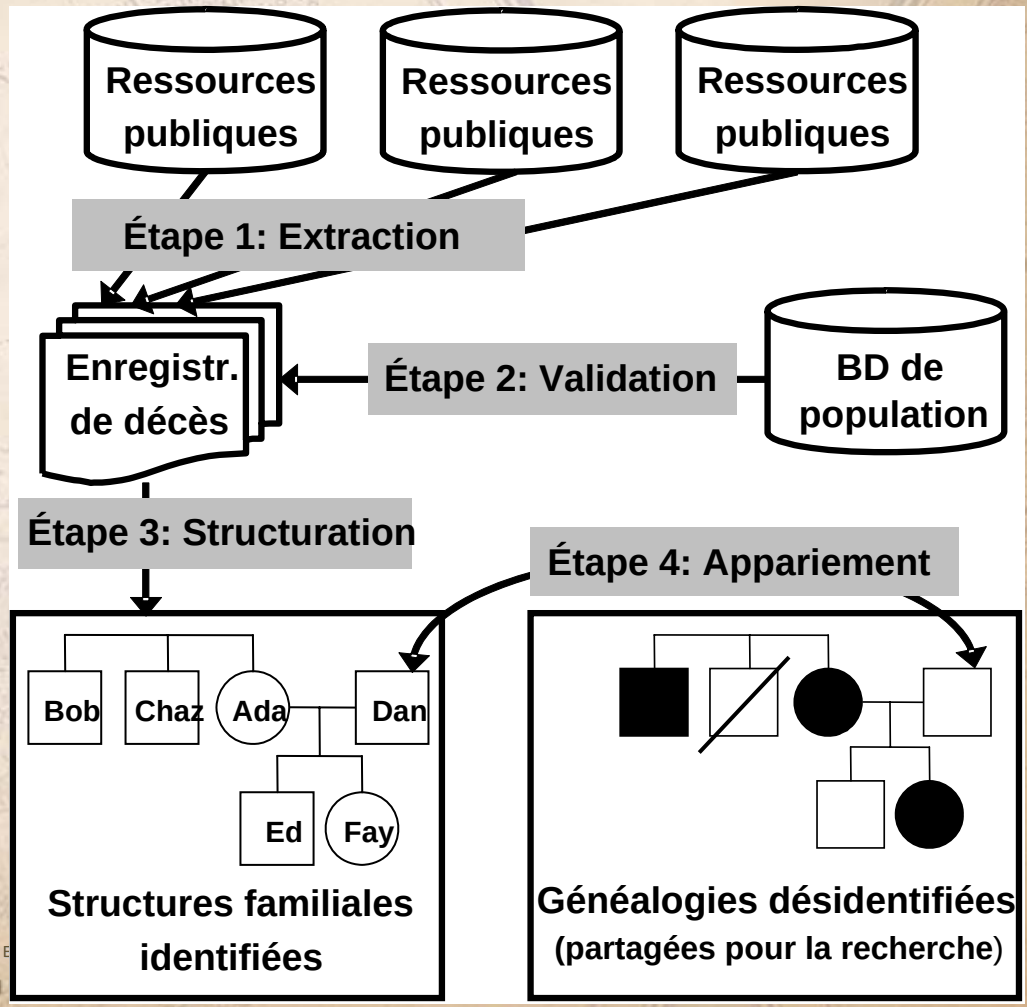
- Plusieurs des technologies de protection de la vie privée visant le génome permettent la réidentification par l'ADN (*Malin 2005*)
- L'ADN est réidentifié par des méthodes automatiques, par exemple :
  - Génotype – Inférence phénotype (*Malin & Sweeney, 2000, 2002*)



Réidentification généalogique

(*Malin 2006*)

- *IdentiFamily*:
  - logiciel qui apparie des généalogies désidentifiées à des personnes désignées
  - Se sert d'informations accessibles au public, p.ex., nécrologies, enregistrements de décès, la base de données de la *Social Security Death Index* pour établir des généalogies



Réidentification généalogique

(*Malin 2006*)

WyomingNews.com
Wyoming Tribune-Eagle

OBITUARIES

Richard R. Mann

1924-2007

Richard R. Mann, 82, of Cheyenne died Jan. 12 at Cheyenne Regional Medical Center.

He was born June 29, 1924, in Allentown, Pa., and had lived here since 1956.

Mr. Mann served in the Army Air Corp during World War II in South Africa and Italy.

He retired as a flight engineer for the Wyoming Air National Guard.

Mr. Mann was a member of St. Mary's Catholic Church, Elks, Moose and the Knights of Columbus, where he had been a past grand knight and state deputy.

He is survived by two sons, Gerald Mann and Thomas Mann, both of Cheyenne; seven daughters, Teresa Johnson, Kathryn Schroll, Judith Oldenburg, Cheryl Thibault, and Jon Cameron, all of Cheyenne, Lou Ann Golden of Sidney, Neb., and Kimberly Byron of Littleton, Colo.; his companion, Katie Heaton of Cheyenne; 25 grandchildren and two great-grandchildren.

He was preceded in death by his wife of more than 50 years, Patricia A. Mann; two daughters, Mary Constance Grant and Jeanane Rhodes; his parents, Russell and Viola Mann; two brothers, Roland Mann and Robert Mann; and a sister, Rochelle Behrandt.

Vulnérabilité du système

(Malin, JAMIA 2005)

	Systèmes de protection de la vie privée			
Quoi	Tiers de confiance	Tiers de semi-confiance	Dénominalisation	Désidentification
Où	deCode Genetics Inc.	University of Gent, Custodix	Université de Montreal	University of Utah, University of Sydney, Australian National University
Vulnérabilité à une attaque				
Structures de famille	■		■	■
Sillage		■		■
Génotype-Phénotype	■	■	■	■
Dictionnaire	■	■		

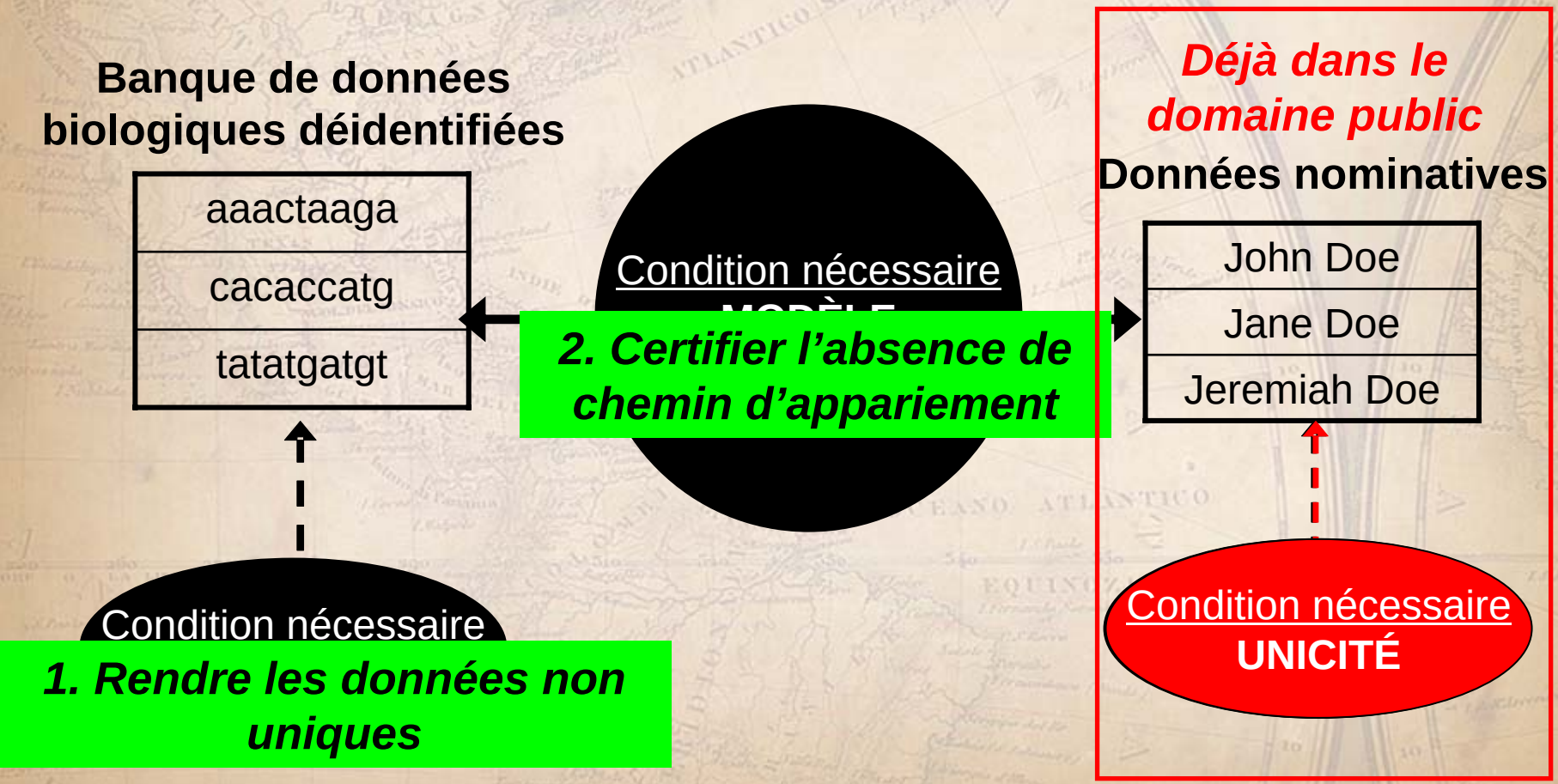
■ ***Vulnérable*** ***Pas vulnérable***

La modification des données n'assure pas la protection

- Science Magazine (*Lin et al, 2004*)
  - < 100 SNPs rendent l'ADN unique
  - –
  - –
  - –
    - De nombreuses perturbations sont requises pour empêcher l'appariement
    - Garder les enregistrements sous scellés

AVERTISSEMENT :
L'unicité ne garantit pas que la vie privée sera compromisée

Modèle formel de réidentification

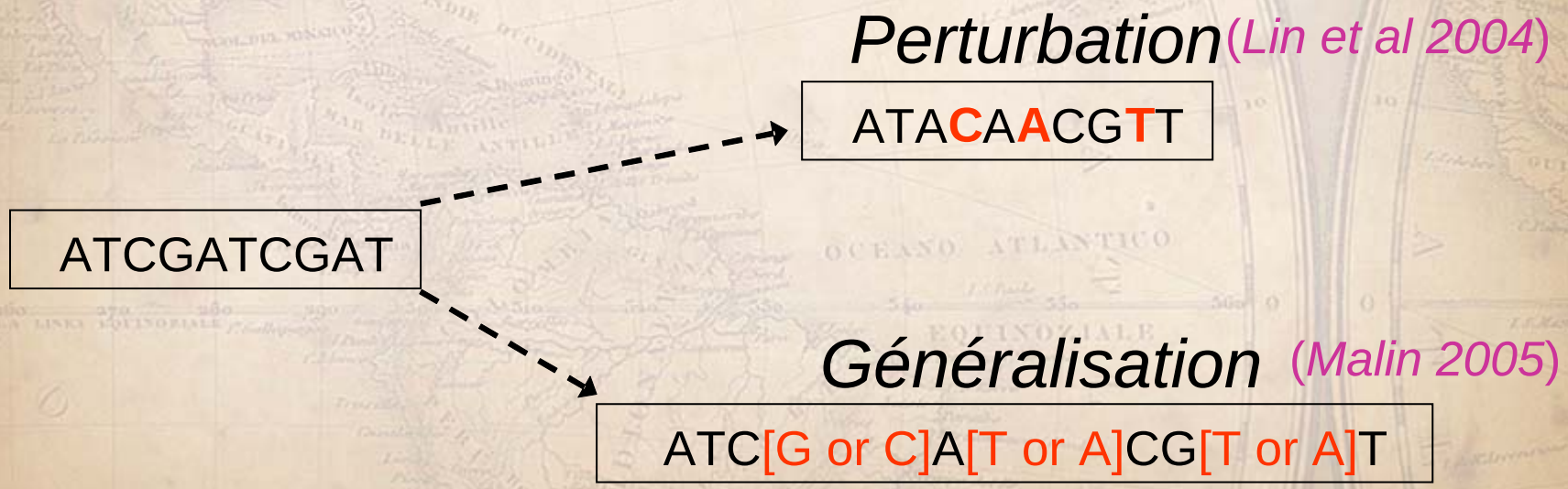


Protection formelle

- *k*-mappage (Sweeney, 2002)
  - Chaque enregistrement partagé désigne au moins *k* unités dans la population
- *k*-anonymat (Sweeney, 2002)
  - Chaque enregistrement partagé est semblable à au moins *k-1* autres enregistrements
- *k*-non appariement (Malin 2006)
  - Chaque enregistrement partagé s'apparie à au moins *k* identités à travers de son sillage
  - Satisfait le modèle de protection par *k*-mappage

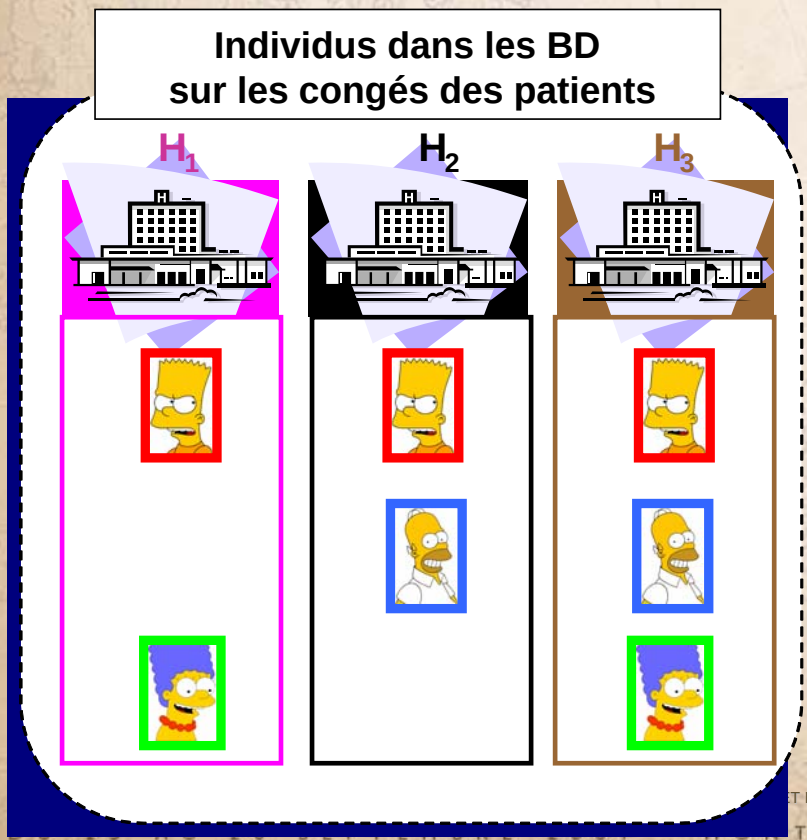
Au delà des protections *ad hoc*

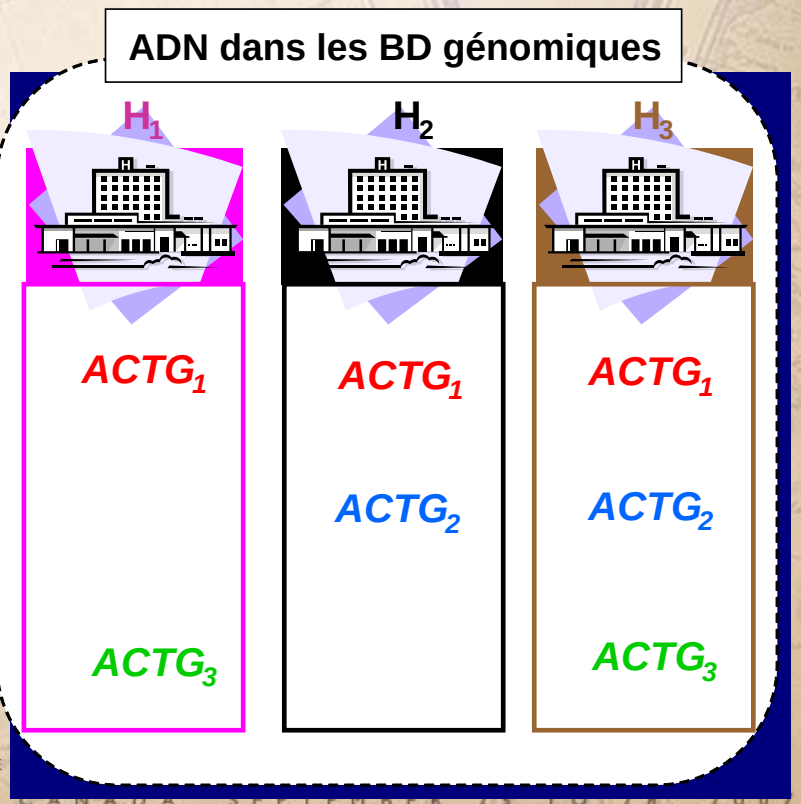
- La perturbation ne garantit pas la protection de la vie privée
- Alternative : Généralisation des données



Savoir qui vous êtes à partir d'où vous avez été (« sillage »)

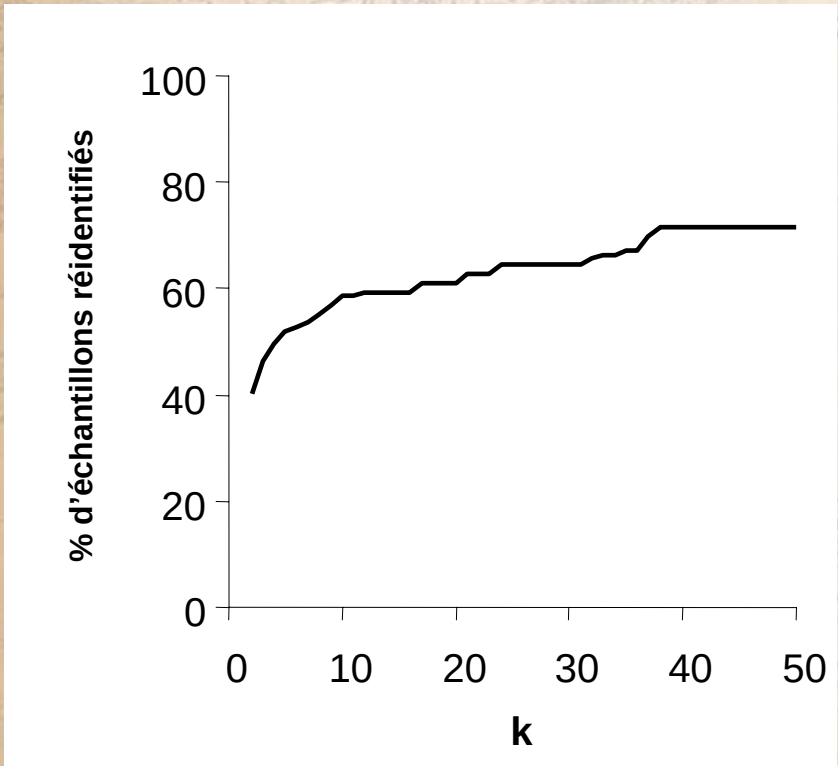
(Malin & Sweeney, 2001; 2004, Malin & Airoldi 2006)

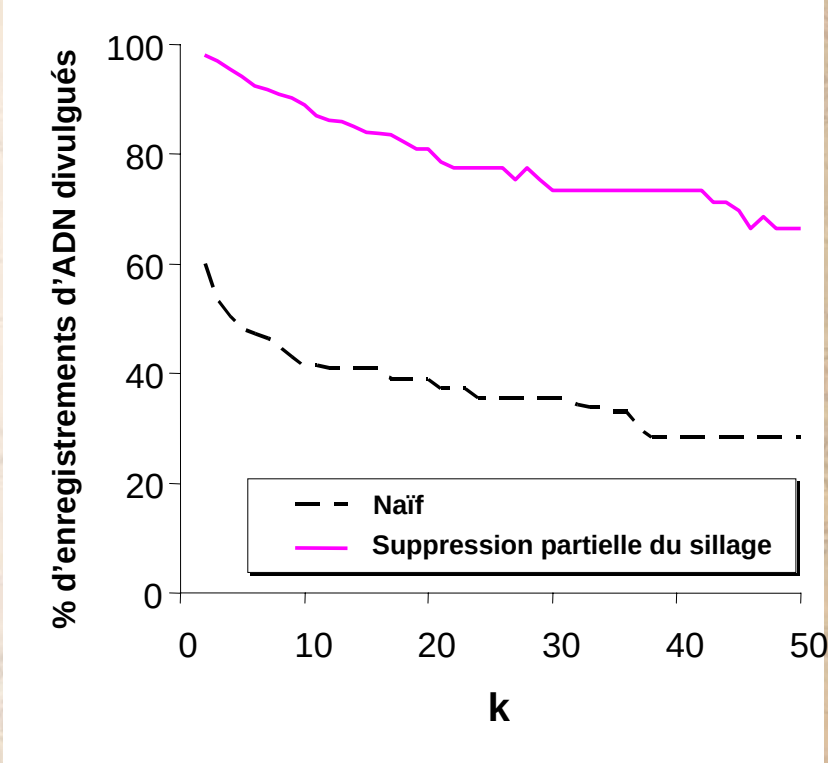




Empêcher le sillage : population avec la fibrose kystique

(1149 échantillons)





AVANT STRANON
100% échantillons en entrepôt

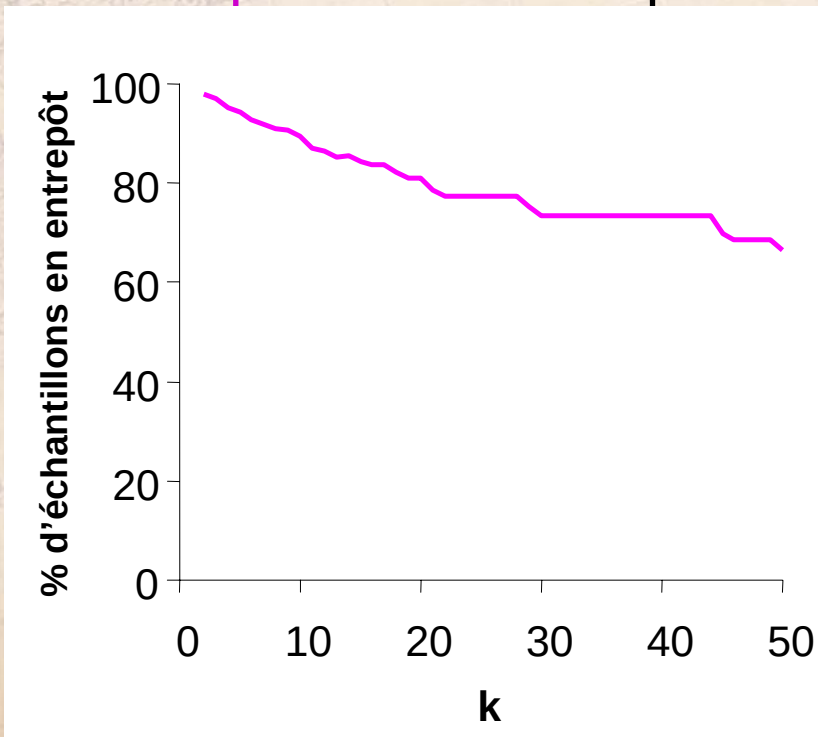
APRÈS STRANON
0% échantillons k-réidentifiés

Utilité : Risque quantifié

Réglage forcé

Réglage initial

Quantité demandée



- Modification au risque de réidentification
- Déplace le fardeau de l'accroissement du risque vers l'analyste requérant
- Lie les modèles légal et informatique

Évaluer et gérer les risques liés à la repersonnalisation

par

Khaled El Emam

Université d'Ottawa

Gérer les risques liés à la repersonnalisation – I

- Avant la collecte de données
  - Scénarios
    - Au moment de la préparation d'un protocole
    - Aux fins d'examen par un comité d'éthique
    - Dans le cadre de la formulation de nouvelles politiques et procédures
    - Au cours de la rédaction d'ententes sur l'échange de données
  - Outils
    - Heuristiques
    - Simulations

Gérer les risques liés à la repersonnalisation – II

- Après la collecte de données
  - Scénarios
    - Fournir des données aux administrateurs, aux chercheurs ou aux ministères gouvernementaux
    - Répondre à une demande d'accès à l'information
  - Outils
    - Masquage
    - Dépersonnalisation en fonction des risques

Outils heuristiques, masquage, dépersonnalisation

- La règle des 20 000, des 70 000, des 100 000…
- Outils de décision tirés d'expériences similaires
- Il y a environ 18 outils de masquage disponibles sur le marché
- Détermination d’un seuil de risque pour la dépersonnalisation

Risque acceptable lié à une nouvelle personnalisation

- Quelles sont les bases de données auxquelles un utilisateur malveillant a accès pour coupler des dossiers?
- Que sait l'utilisateur malveillant avant de passer à l'action?
- Quels sont les coûts liés à la vérification?
- Comment rendre compte des compromis liés à la protection de la vie privée pour le public ?
- Quelles sont les répercussions de l'utilisation de modèles de consentement?

Bases de données

- Renseignements et registres publics
- Bases de données commerciales, mais accessibles au public
- Bases de données confidentielles et de propriété exclusive

Coûts liés à la vérification

- À un moment donné, les coûts liés à la vérification deviennent trop élevés comparativement aux avantages que tire l'utilisateur malveillant.
- Il est important d'évaluer la proportion des données propres à une population.
- Il est également important d'évaluer l'étendue des couplages réussis dans l'ensemble.
- Il est possible de contrôler ces deux variables au moyen de la dépersonnalisation.

Compromis

- Le public est disposé à sacrifier sa vie privée en échange d'avantages.
- Le public ne se comportera pas nécessairement comme il le prétend.
- Jusqu'à quel point le public est-il disposé à sacrifier sa vie privée pour obtenir des avantages?

Modèles de consentement

- Les répercussions sur les taux de recrutement et d'erreur sont-elles fonction du modèle de consentement choisi ou de sa mise en œuvre?
- De nombreux facteurs influencent le consentement — ces facteurs ont-ils tous été contrôlés lors de la comparaison des modèles de consentement?

Atelier 4

Protéger la vie privée au moyen de la dépersonnalisation: réalité ou illusion?

1re partie: Discussion

M^me Debra Grant, Ph. D.

Spécialiste principale de la protection de la vie privée en matière de santé

Bureau du commissaire à l'information et à la protection de la vie privée de l'Ontario

Les problèmes de dépersonnalisation posés par les données génétiques et génomiques

William W. Lowrance, Ph.D.
(lowrance@iprolink.ch)

26 septembre 2007

Fondement physique des problèmes

Le génome humain :

- est élaboré et extrêmement détaillé
- influence de nombreuses caractéristiques personnelles
- est intrinsèque au corps humain
- reste le même durant toute la durée d'une vie
- est unique à une personne

Le génome complet est porté par l'ADN dans chaque cellule du corps, à l'exception des globules rouges.

Les données génomiques ressemblent à

...tttccgtatgcgtagccagacttaccctcctagtag...

- à raison de 3 000 000 000 « cellules de données » qui portent chacune les bases a, t, g, c.

La modification ou l'insertion de quelques a,t,g,c peut faire une grande différence, quelle que soit la manière dont on envisage le génome :

- comme un ruban de programmation dynamique
- comme un « code à barres » intrinsèque.

Les données génétiques ressemblent à ce qui suit :

- à l'échelle d'une séquence : |ctag…ctccca|
- à l'échelle d'un gène : « Gène porteur du diabète SLC308A »
- à l'échelle du corps : « cheveux roux », « dysplasie rénale héréditaire »
- à l'échelle de la famille : ascendance familiale, antécédents familiaux en matière de santé, autres indicateurs.

À mon avis, l'interprétation la plus utile de la notion d'identifiabilité pour les données génomiques est la suivante :

> *« L'identifiabilité » est l'associativité potentielle des données précises à une personne.*

Moyens par lesquels les données génomiques peuvent être personnalisées

a) par association : en associant un génotype à des données génotypiques identifiables (p. ex. données policières, militaires ou par lien familial)

b) par couplage : en couplant des données génomiques et d'autres données connexes (sur la santé, sociales, etc.) avec d'autres données

c) par profilage : en décrivant de manière « probabilistique » l'apparence, les facteurs liés à la santé ou d'autres caractéristiques possibles.

Tactiques permettant de dépersonnaliser les données génomiques

a) en limitant la proportion de renseignements relatifs à un génome qui sera communiquée
b) en faussant de manière statistique les données avant de les communiquer
c) en dépersonnalisant les données de manière irréversible
d) en séparant les identifiants et en effectuant un codage par clé.

Tactique a) en limitant la proportion de renseignements relatifs à un génome qui sera communiquée

- s'effectue déjà et peut protéger les données;
- toutefois, elle limite souvent l'utilité des données, puisqu'on ignore le plus souvent quelles seront les portions du génome qui seront pertinentes;
- il est difficile de déterminer la « bonne quantité » de renseignements à communiquer.

Tactique b) en faussant de manière statistique les données avant de les communiquer

- peut se faire, par exemple en substituant de manière aléatoire des a/t/g/c;
- fausse presque toujours l'utilité des données, puisque la plupart des analyses s'effectuent sur le plan de détails précis.

Tactique c) en dépersonnalisant les données de manière irréversible

- s'effectue parfois, par exemple lorsque l'objectif consiste à sonder l'occurrence d'un phénomène particulier ou à fournir des données à des fins éducatives.

Tactique d) en séparant les identifiants et en effectuant un codage par clé

- fonctionne bien lorsqu'elle est effectuée de manière adéquate, que la clé est bien protégée et que l'utilisation de la clé pour reconstituer les données est strictement supervisée;
- est de plus en plus utilisée dans le cadre d'activités comme la recherche en santé.

Dépersonnaliser ou pas?

Les motifs justifiant la dépersonnalisation des données et les manières de le faire sont fonction :

- de la nature des données;
- du consentement;
- des usages prévus;
- du potentiel de couplage avec des données génotypiques de référence ou d'autres données;
- des protections.

Dépersonnalisation : autres solutions et compléments

- Fournir l'accès aux données au moyen de communications supervisées (régies par contrat, supervisées par un comité de gérance, etc.).
- Sanctionner l’utilisation malveillante des données (comme la repersonnalisation inappropriée) ou l’utilisation abusive des données (comme la discrimination).

Mot de la fin

La dépersonnalisation est une forme de protection pratique et essentielle — pour les données génomiques comme pour les autres formes de données — et il faut fortement encourager son utilisation!

Référence générale : LOWRANCE et COLLINS, « Identifiability in genomic research », *Science* 317, pages 600 à 602, 3 août 2007.

L'accès à des renseignements personnels à des fins de recherche en santé et le consentement à leur utilisation – *le point de vue du public*

Don Willison, D. Sc.

Centre for Evaluation of Medicines, St. Joseph's Healthcare,
Département d'épidémiologie clinique et de biostatistique,
Université McMaster,
willison@mcmaster.ca

- **Équipe de recherche :**
  - **Université McMaster**
    - Don Willison (chercheur principal – protection de la vie privée, politiques, méthodes de recherche)
    - Lisa Schwartz (philosophie, bioéthique)
    - Julia Abelson (mobilisation du public)
    - Cathy Charles (mobilisation du public, méthodes qualitatives)
    - Lehana Thabane (statisticien, méthodes quantitatives)
    - Marilyn Swinton (coordonnatrice de la recherche, méthodes qualitatives)
  - **Université York**
    - David Northrup (méthodes d'enquête)
  - **Réseaux canadiens de recherche en politiques publiques**
    - Mary Pat MacKinnon, Judy Watling (dialogue)
- **Financement :** Instituts de recherche en santé du Canada
- **Publication :** JAMIA – Novembre 2007

Contexte : Accroissement de l'utilisation des renseignements personnels pour la recherche en santé

- Accroissement de la portée et de la complexité de l'utilisation des données
  - Couplage de données
    - données administratives et cliniques
    - données d'enquête et génétiques
  - Études uniques à délai fixe → registres et biobanques
  - Dossier de santé électronique (DSE) : accès élargi aux renseignements sur la santé pour :
    - la recherche en santé publique / de la population
    - des essais pragmatiques
- Les chercheurs ont besoin de données au niveau des particuliers
  - Le défi : le camouflage de l'identité
  - Le débat : traiter les données comme identifiables?

Questions entourant le consentement

- Le point de vue du patient/du public :
  - Comment obtenir un consentement utile et valable?
- Le point de vue du chercheur :
  - Faisabilité d'obtenir le consentement
    - éventuels biais de sélection dans un système fondé sur le consentement
  - Limites en cas de renonciation au consentement :
    - Impossibilité de joindre le patient / Qui peut faire l'examen des fiches médicales?
- Généralités :
  - Devons-nous nous limiter à l'option binaire du consentement et de l'absence de consentement?

Notre enquête :

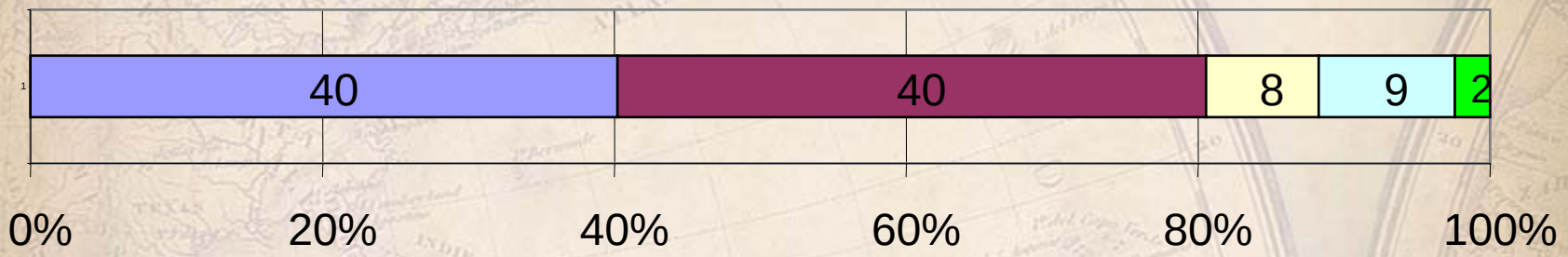
- Enquête téléphonique à composition aléatoire à l'étendue du Canada
  - Mars-avril 2005
  - n=1230 (taux de réponse de 58 %)
- Structure :
  - **Questions générales**
    - Caractéristiques démographiques, altruisme
    - La santé et la protection de la vie privée dans le contexte d'autres priorités
  - **Questions particulières**
    - attitudes vis-à-vis de la recherche et de la protection de la vie privée
    - confiance faite aux établissements
    - utilisation des dossiers médicaux pour différents types de recherche
  - **Scénarios particuliers. Le rôle du consentement dans :**
    - la recherche fondée sur les dossiers médicaux
    - le dossier de santé électronique
    - le couplage d'enregistrements

NOS CONSTATATIONS
Les attitudes vis-à-vis de la protection de la vie privée

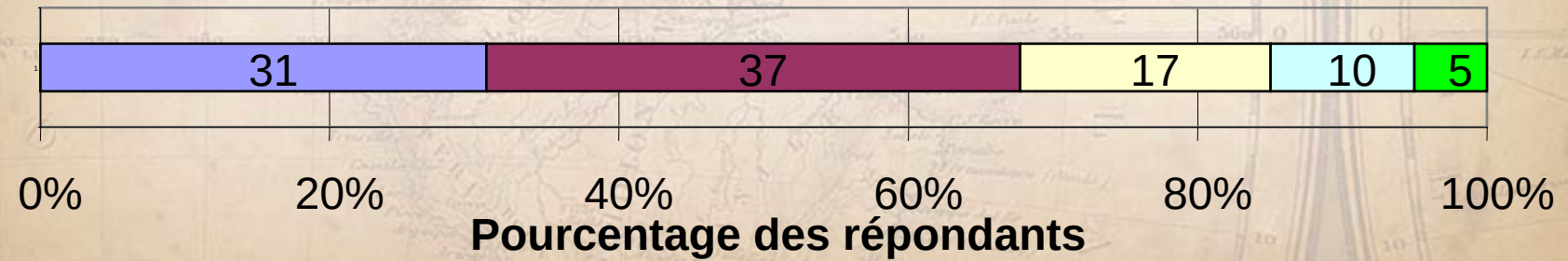
- Les participants attachent une grande importance à la protection de la vie privée en principe :
  - 97 % considèrent que la protection de leurs renseignements personnels est importante
    - très importante pour 74 %; plutôt importante pour 23 %
  - 91 % sont d'accord pour dire qu'il faut s'efforcer davantage de protéger la vie privée
    - 59 % sont tout à fait d'accord / 32 % sont plutôt d'accord
  - 92 % sont d'accord pour dire que tout le monde y gagne si la vie privée des gens est respectée
    - 66 % sont tout à fait d'accord / 26 % sont plutôt d'accord

La protection de la vie privée et la recherche

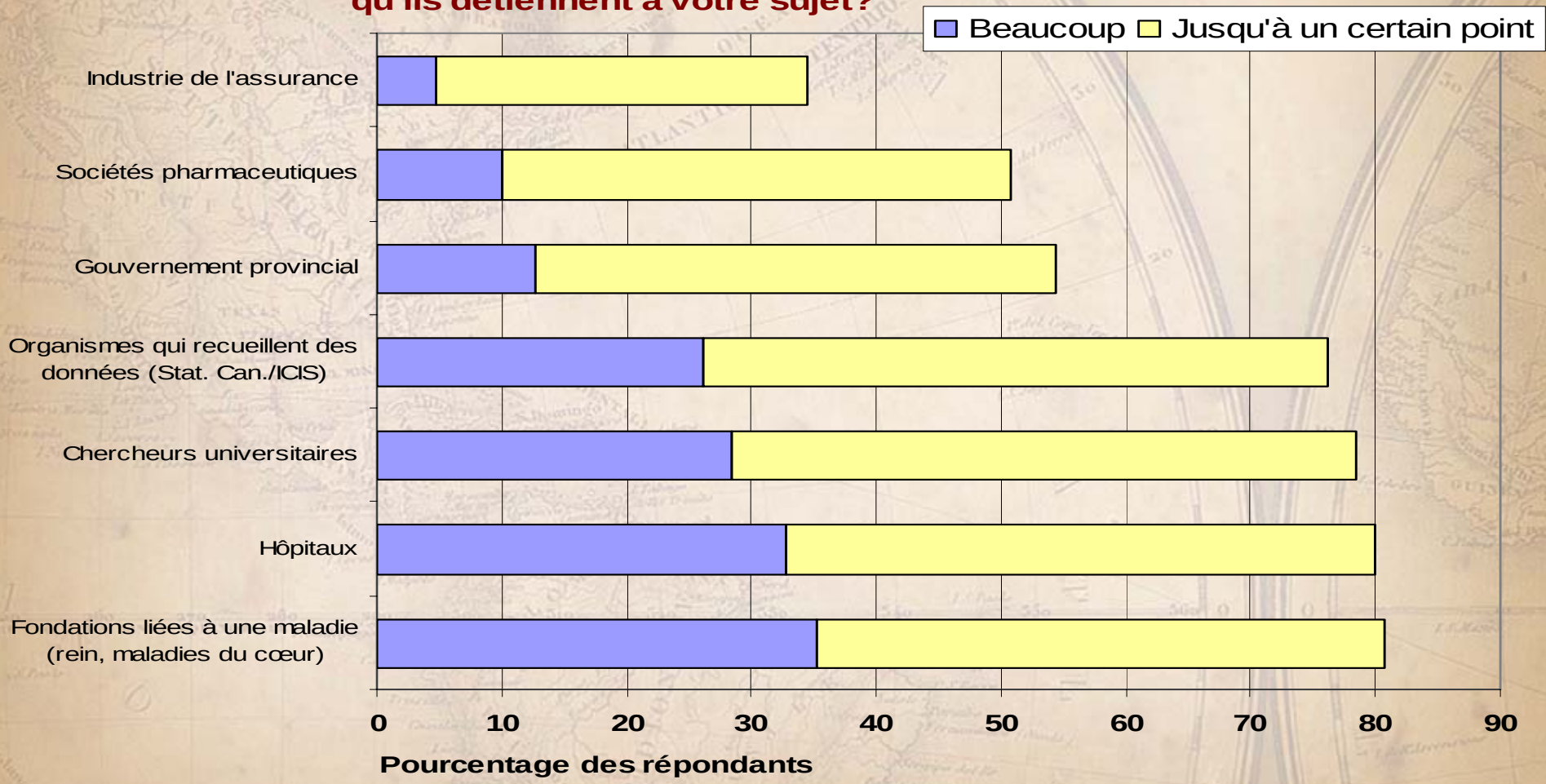
Les gens devraient permettre l'utilisation de leurs renseignements au profit de la société si cela ne cause pas de préjudice aux particuliers.



La recherche qui pourrait être bénéfique pour la santé des gens est plus importante que ne l'est la protection de la vie privée des gens



Dans quelle mesure faites-vous confiance aux établissements suivants pour qu'ils assurent la confidentialité des renseignements sur la santé qu'ils détiennent à votre sujet?



Scénarios de recherche

- 4 scénarios :
  - Extraction des renseignements du dossier de santé à des fins de recherche
  - Utilisation du dossier de santé électronique (DSE) à des fins de recherche
  - Couplage des données sur le niveau de scolarité à celles du DSE
  - Couplage des données sur le revenu à celles du DSE
- Identificateurs directs supprimés des données
  - Rend difficile, mais pas impossible, une nouvelle identification

Opinion relative au consentement et aux différentes options selon le scénario

Scénario	n	Choix concernant le consentement					
		Ne pas utiliser	Demander d'abord la permission			Avis / refus	Utiliser tout simplement
			Chaque fois	Générale, renouvelable	Générale, une seule fois		
Extraction manuelle des données du dossier médical	1 207	4 %	32 %	23 %	5 %	24 %	12 %
			60 %				
Extraction automatisée des données du DSE	941	9 %	36 %			28 %	27 %
Couplage des données sur le niveau de scolarité à celles du DSE	858	10 %	41 %			26 %	23 %
Couplage des données sur le revenu à celles du DSE	853	27 %	40 %			16 %	17 %

Dialogues publics

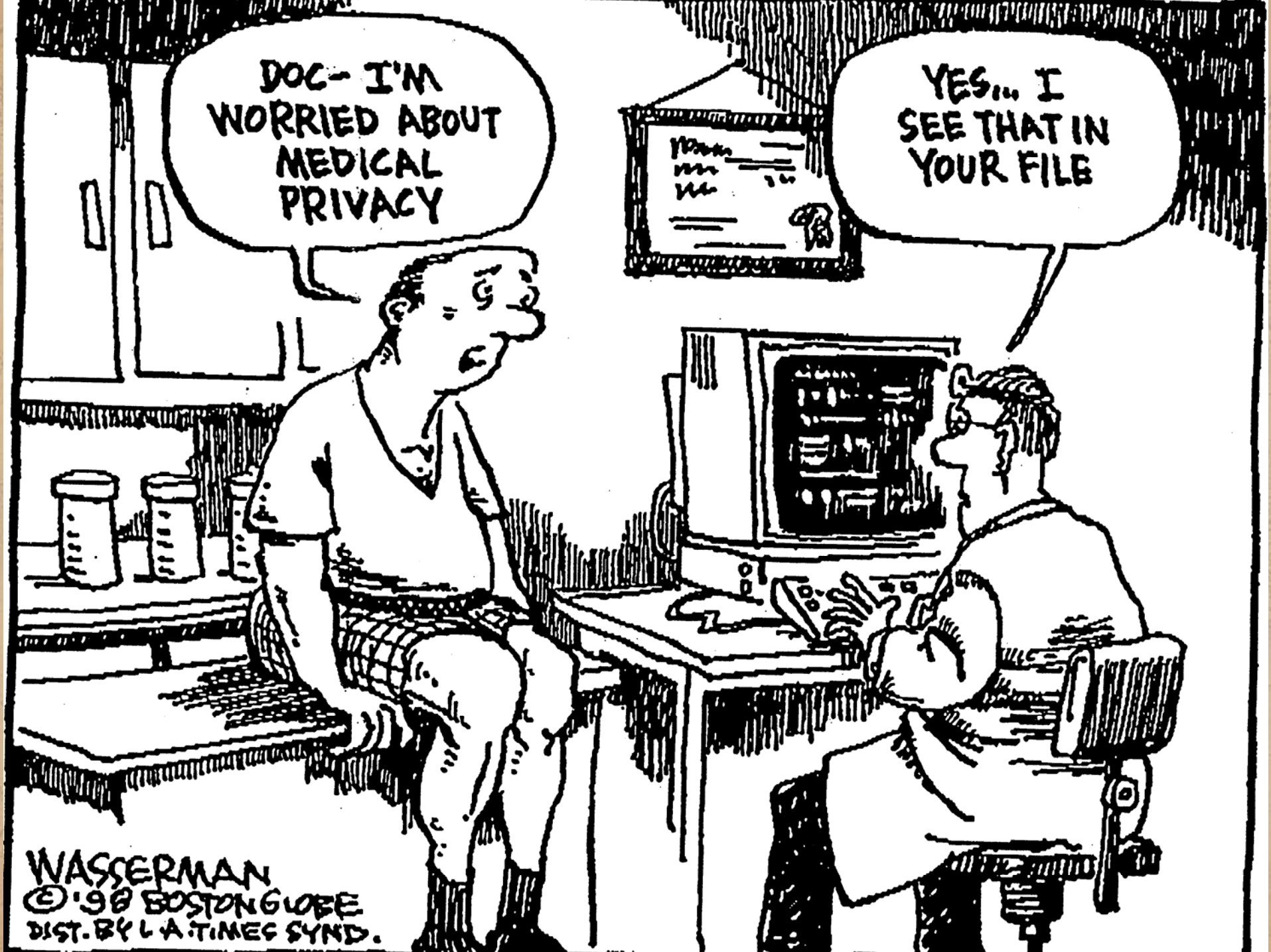
- Messages clés :
  - Vif sentiment d'altruisme, dans la mesure où le public en bénéficie
  - Désir de contrôle accru en présence d'un aspect commercial
  - Importance de la confiance dans le chercheur
    - bienfaisance / absence de malfaisance
  - Le choix concernant le consentement diffère peu selon que les renseignements sont identifiables ou pas
    - Une question de respect de la personne

Conclusions

- Le public attache de l'importance tant à la recherche en santé qu'à la protection de la vie privée
  - Avec un peu d'insistance, la recherche tend à l'emporter sur la protection de la vie privée
  - Les gens sont en faveur de l'utilisation des renseignements personnels à des fins de recherche
    - Dans une large mesure, cet appui est conditionnel
    - Les chercheurs doivent prendre soin d'entretenir la confiance du public
    - Importance de faire attention aux mesures de protection
- Le degré de contrôle que les gens souhaitent exercer sur l'utilisation de leurs renseignements personnels varie
  - La majorité (~65 %) sont réceptifs à différentes façons d'exprimer leur consentement d'une étude à l'autre
  - Seulement 12 %-27 % sont disposés à permettre l'utilisation de leurs renseignements à leur insu ou sans leur consentement

Incidences sur le plan des politiques :

- Appui insuffisant du public à l'égard d'un consentement présumé ou réputé généralisé à l'utilisation des renseignements personnels à des fins de recherche en santé
- Consignation des choix des particuliers concernant le consentement aux utilisations secondaires des renseignements personnels – modèle d'autorisation
  - Embrasser l'éventail des options relatives au consentement
  - Quelle est la meilleure façon de procéder?
    - Suivi des choix au moyen d'un DSE interopérable commun (Inforoute Santé du Canada)
    - Besoin d'infrastructures pour la détermination et la gestion des choix concernant le consentement
    - Mesures de protection et structures de gouvernance


DOC- I'M WORRIED ABOUT MEDICAL PRIVACY
YES... I SEE THAT IN YOUR FILE
WASSERMAN
©'98 BOSTON GLOBE
DIST. BY L.A. TIMES SYND.