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

Le forage des données

Bradley A. Malin, Ph.D.

Professeur adjoint

Département d'informatique biomédicale

Université Vanderbilt

Collecte des données

Données démographiques

Séjour à la clinique

Dossiers Médicaux électroniques

Collecte ***ouverte***

Visite à l'hôpital pour recevoir un traitement

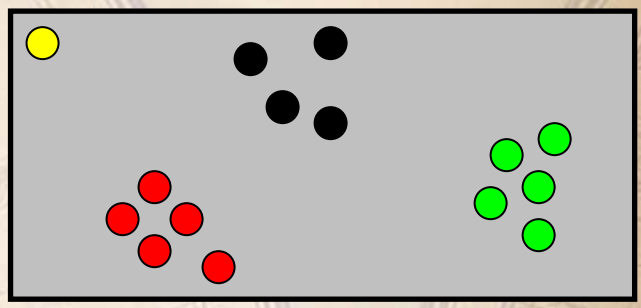
Collecte ***secrète***

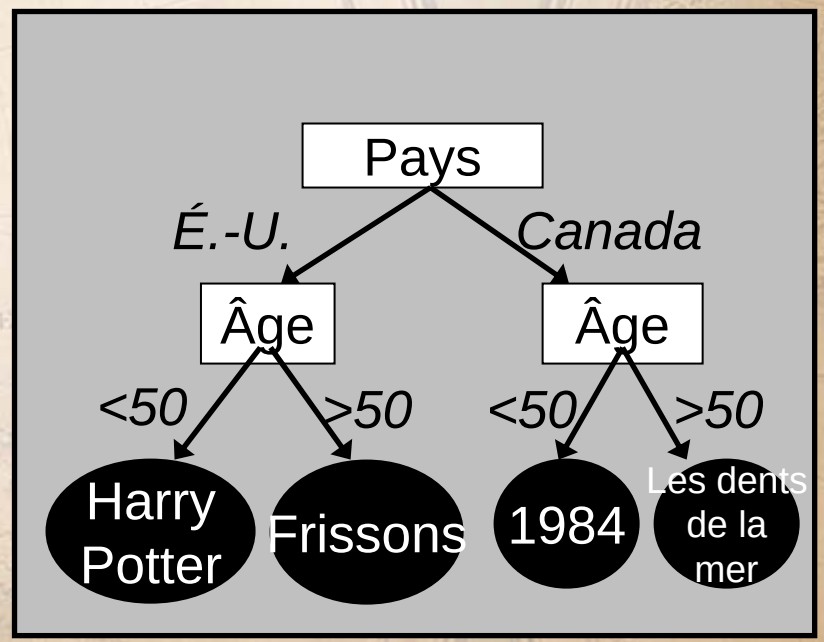
cybercaméra en marchant

- Localisée : dossiers personnels dans les bases de données à la source
- Répartie : Intégration des dossiers à partir de nombreuses sources

Forage des données

- Non supervisé
  - « Étiquettes » *inconnues* à l'avance, à la recherche de schémas de données intrinsèques
  - Grouper les personnes « similaires »
    - Achats des mêmes produits
- Supervisé
  - « Étiquettes » *connues* d'avance
  - Appliquer des modèles aux données-échantillons pour classer de nouveaux cas



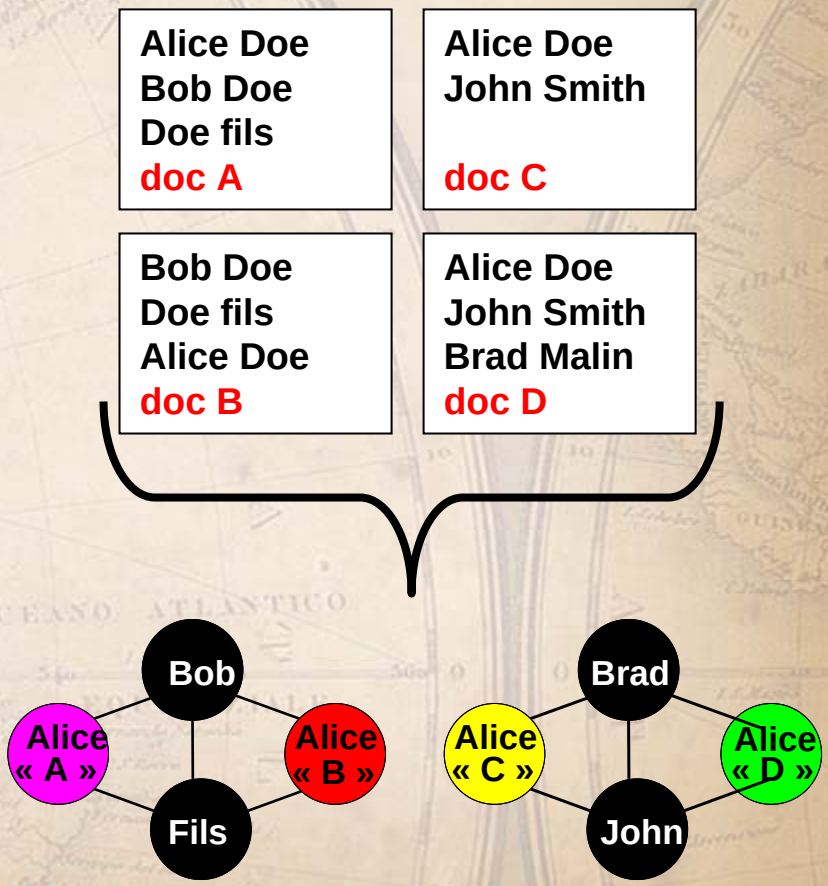


Personnalisation d'un site Web

- Est-ce que le site Web peut prédire ce que je veux voir?
  - *Intrapersonnalisation* : Quelles pages / quels sujets est-ce que j'ai consultés au cours de ma visite précédente?
  - *Interpersonnalisation* : Est-ce que mon exploration ou mes antécédents d'achat sont semblables à ceux d'autres personnes?
- Est-ce que mon comportement révèle mon identité ou des choses délicates à propos de ma vie?
  - Quelles informations ne devraient pas être révélées?

Renseignement

- Les listes d'entités sont de plus en plus courantes
  - Comptes rendus de renseignement, tableaux de service, réseaux
- Combien d'Alice retrouve-t-on? Laquelle est laquelle?
- Quelle est la relation entre Alice et Bob?



Surveillance

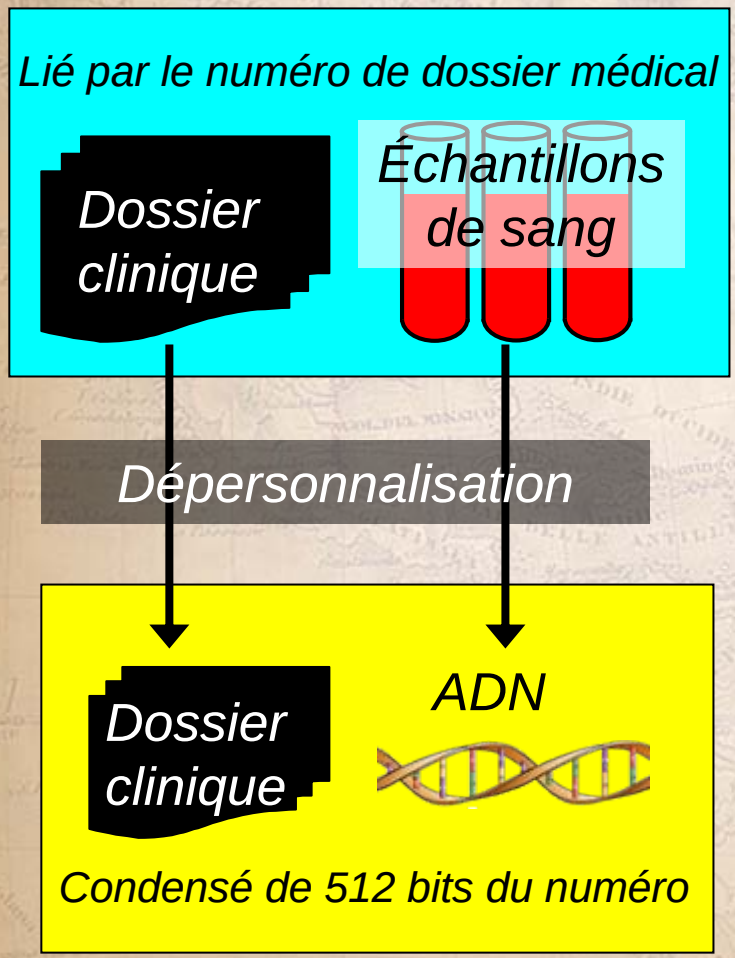
- Surveillance de localisation : est-ce que quelqu'un qui figure sur la liste de surveillance d'Interpol s'est rendu à l'hôtel *X*? A utilisé les services du transporteur aérien Y?

- **Défi :** Les détenteurs de données veulent collaborer, mais ils craignent la connaissance stratégique et les contraintes juridiques

Mesures de protection en matière de vie privée

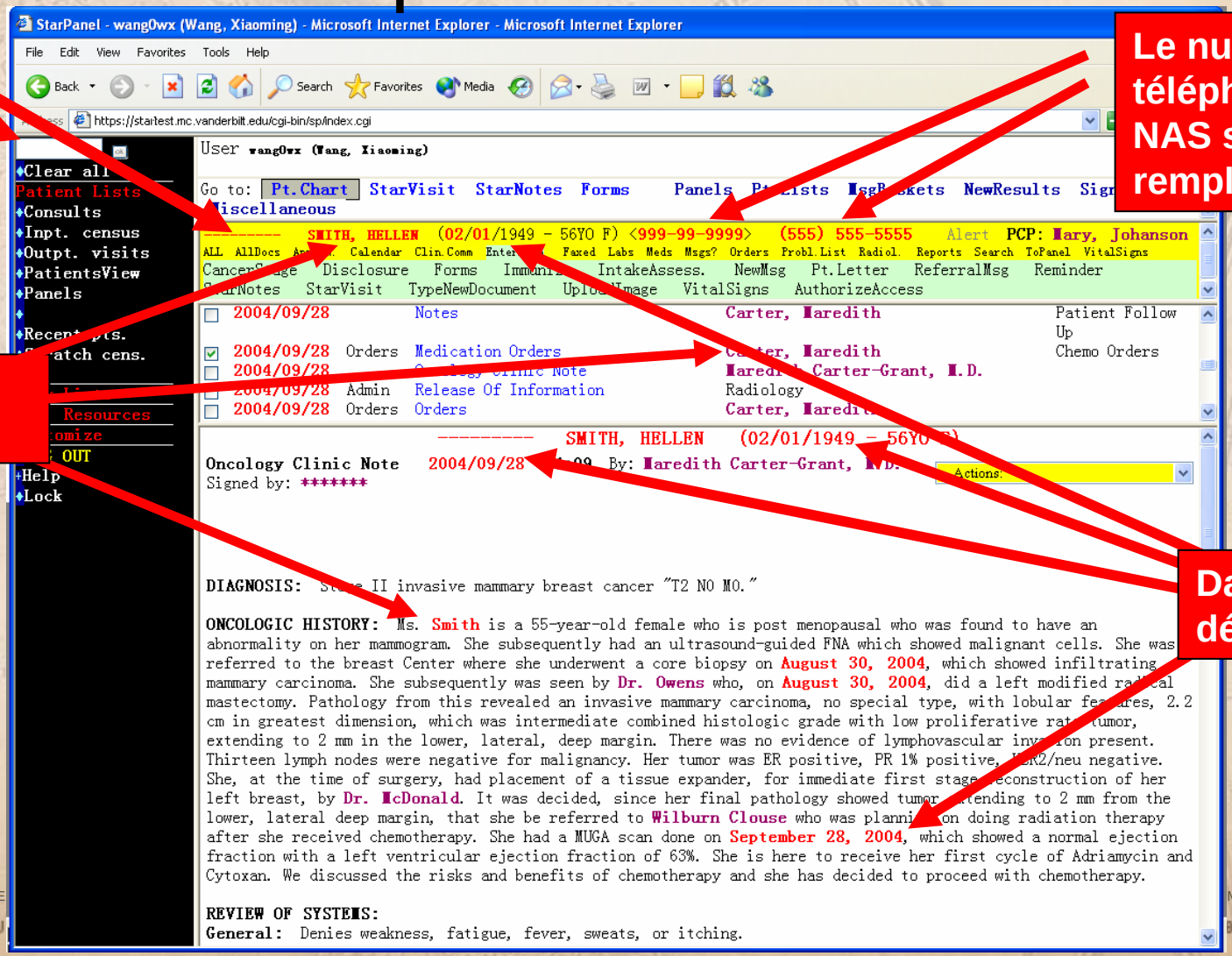
- Protégé l'anonymat
  - Retirer les renseignements signalétiques codés
  - Supprimer les inférences susceptibles de révéler l'identité
- Protégé la confidentialité
  - Cacher les règles de nature délicate
  - Perturbation et généralisation
- Garantir le calcul multipartite
  - E(a) + E(b) = E(a + b) [homomorphisme]
  - E(E(John),x),y) = E(E(John),y),x) [commuter]

Génomique clinique



- Banque de données génétiques à l'Université Vanderbilt
- ADN de sang « non utilisé »
  - *25-75 K par année, 250 K en 5 ans*
- Joint à des dossiers médicaux électroniques dépersonnalisés
  - *600 Go sur 1,4 million de patients*
- « Élaboration d'hypothèses » pour explorer les corrélations entre les caractéristiques cliniques et l'ADN

Exemple de dossier médical dépersonnalisé



Le numéro de DM est enlevé

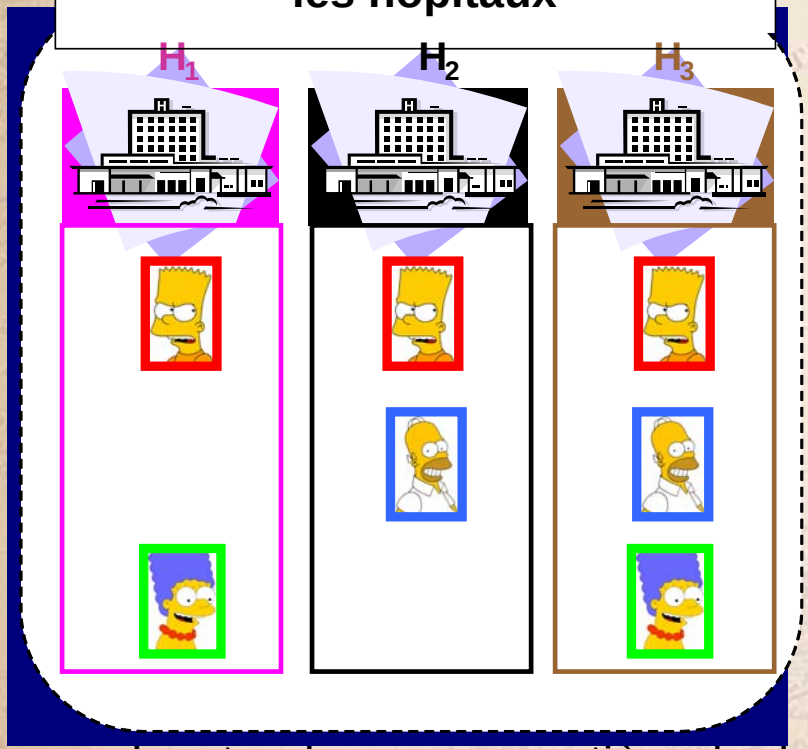
Le numéro de téléphone et le NAS sont remplacés

Noms modifiés

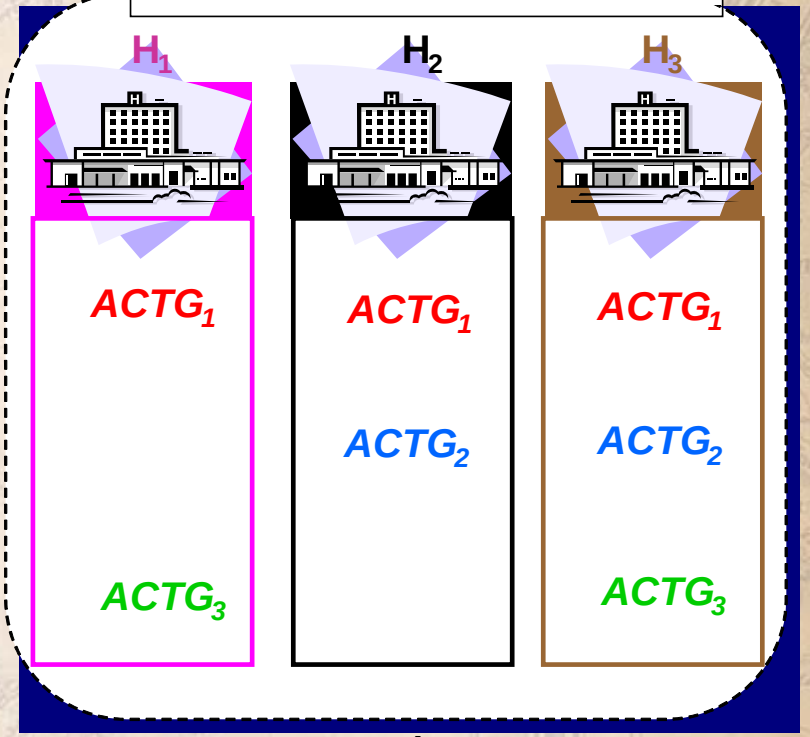
Dates déplacées

Protection naïve

Identités et banques de données sur les départs dans les hôpitaux



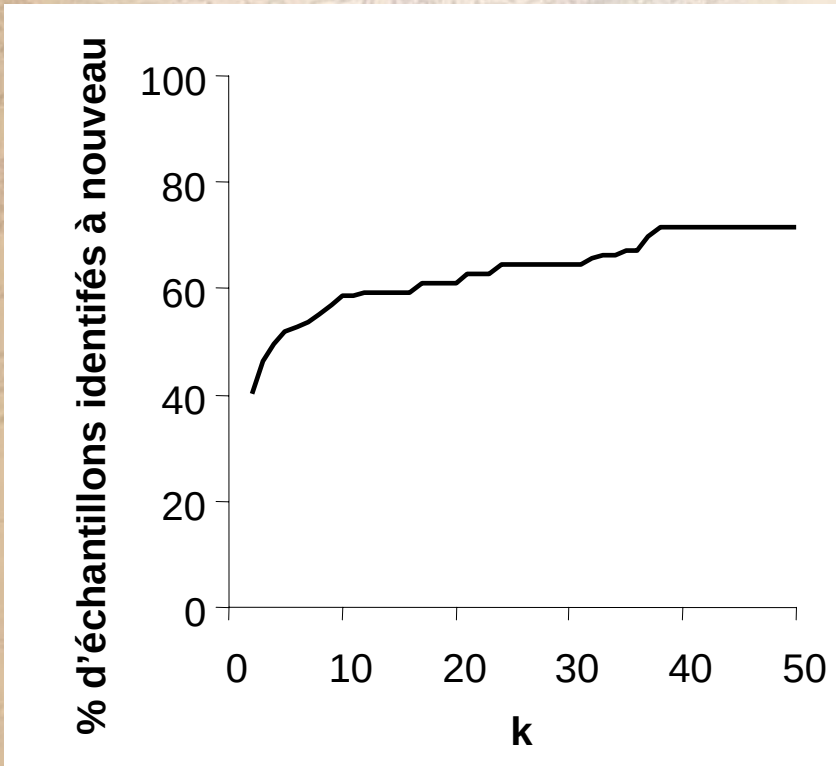
L'ADN dans les bases de données en génomique



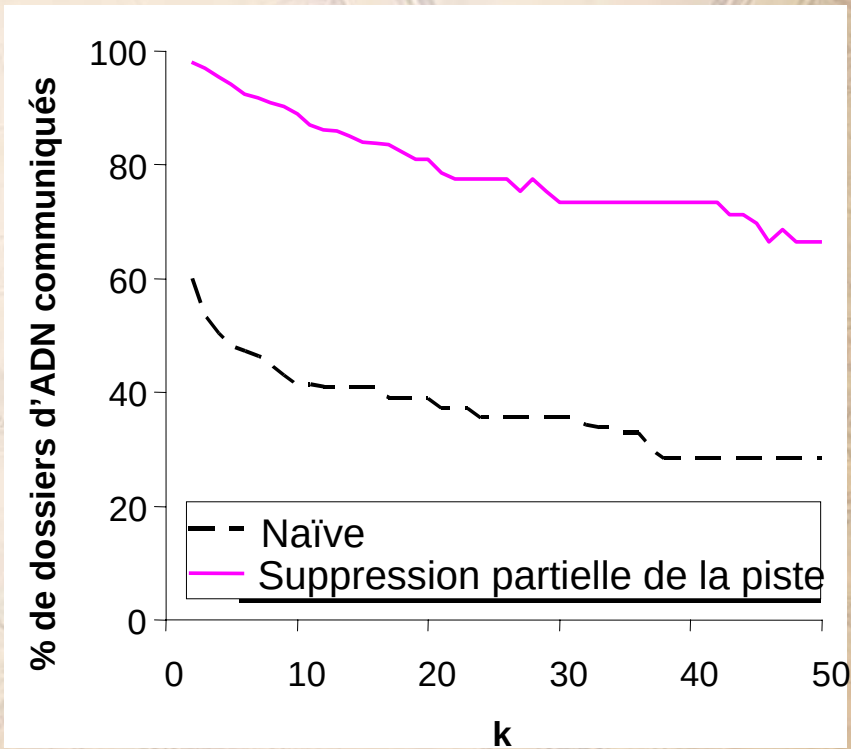
- Les tendances en matière de données peuvent conduire à des compromis sur le plan de la protection de la vie privée
- Dissimuler les tendances « intelligemment » pour aller dans le sens des objectifs visés

En détail : Fibrose kystique

(1 149 patients, 174 hôpitaux)



AVANT la protection
100 % des échantillons archivés



APRÈS la protection
0 % des échantillons identifiés à nouveau

Les conséquences du forage des données sur la protection de la vie privée dans les secteurs public et privé

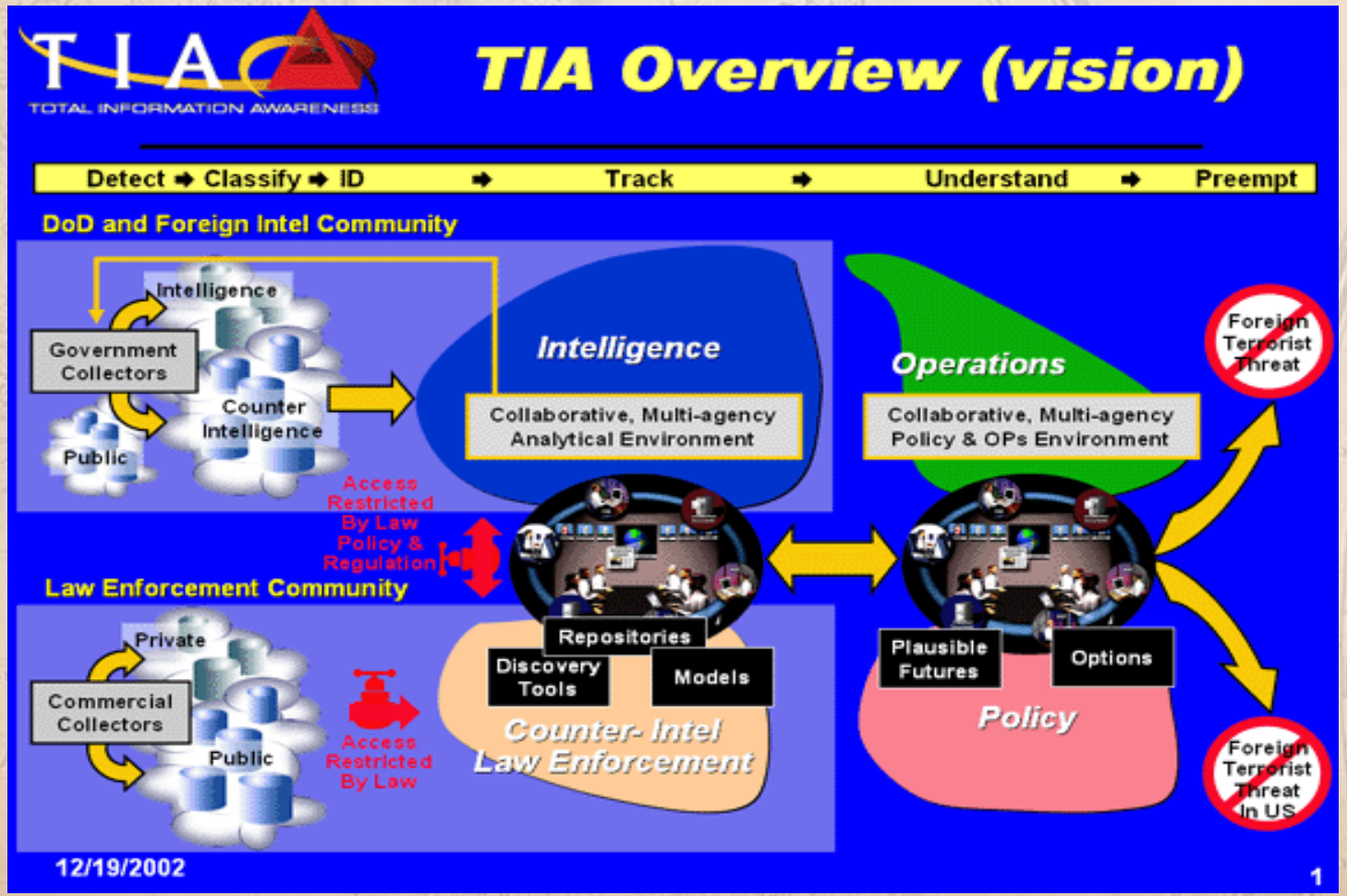
Richard S. Rosenberg

Professeur émérite, Département de l'informatique, Université de la Colombie-Britannique et Président de la BC Freedom of Information and Privacy Association

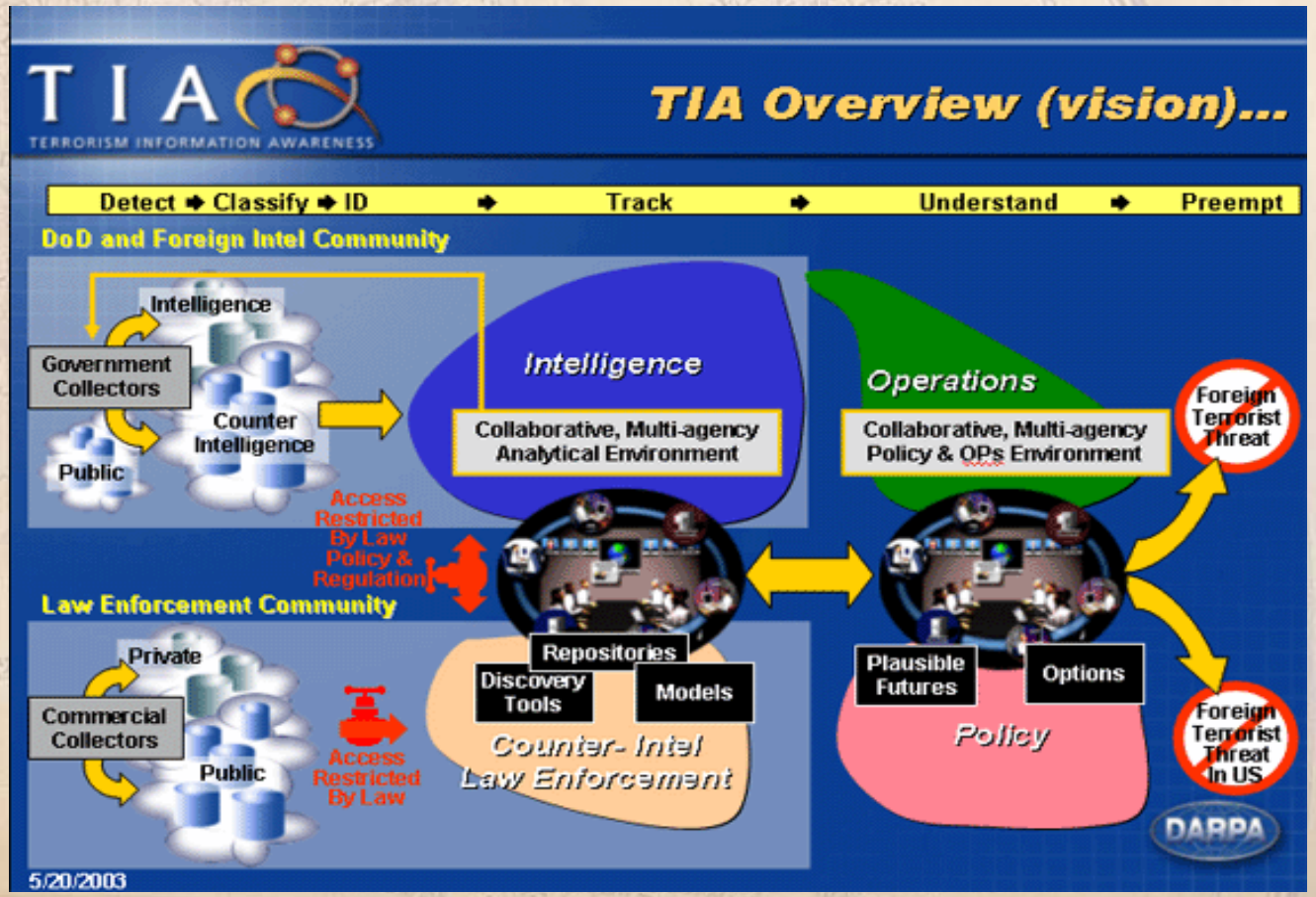
Vancouver (C.-B.)

rosen@cs.ubc.ca

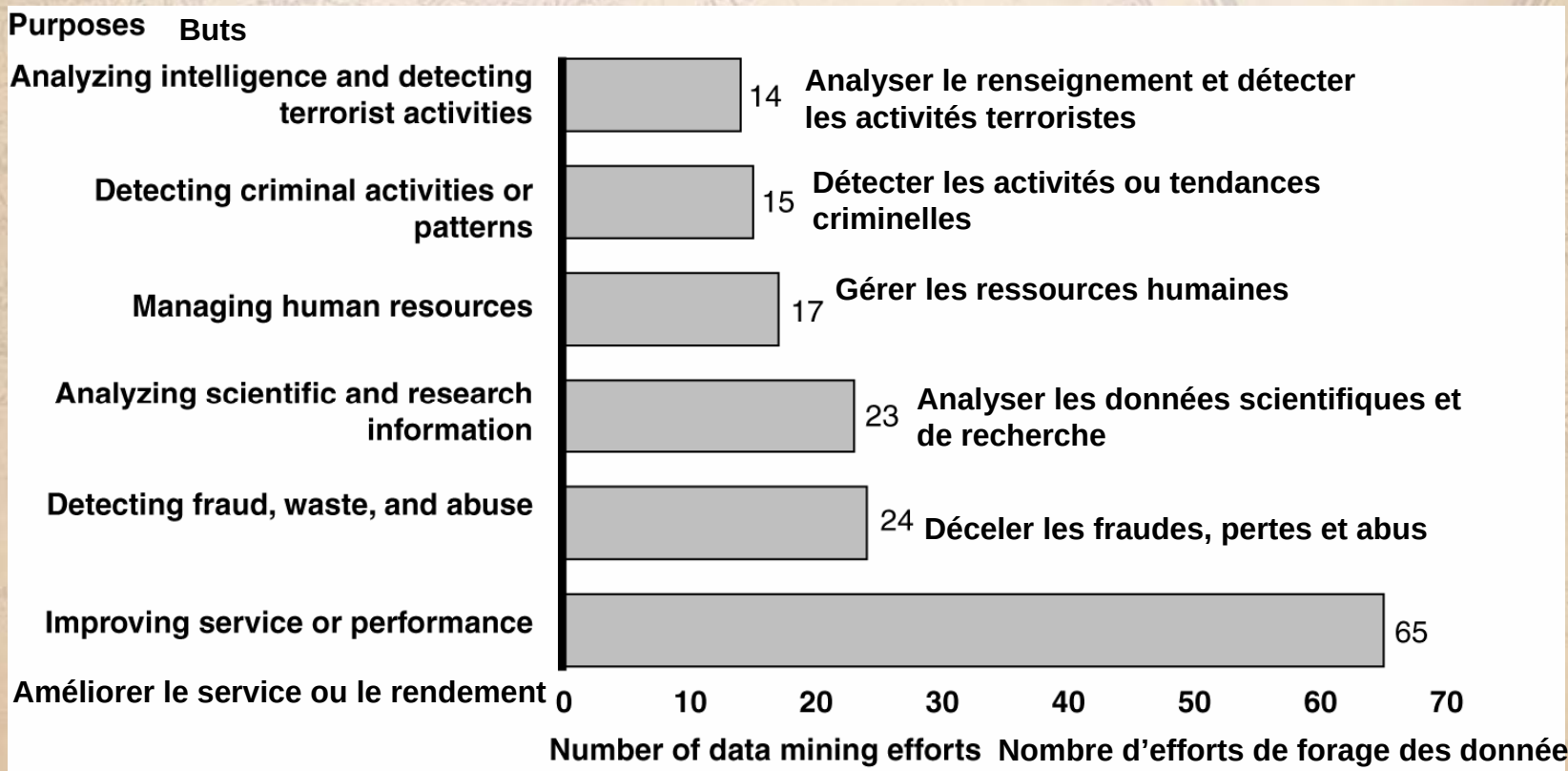
Le gouvernement américain



Une révision



Les six principaux buts que poursuivent les départements et organismes dans leurs efforts de forage des données



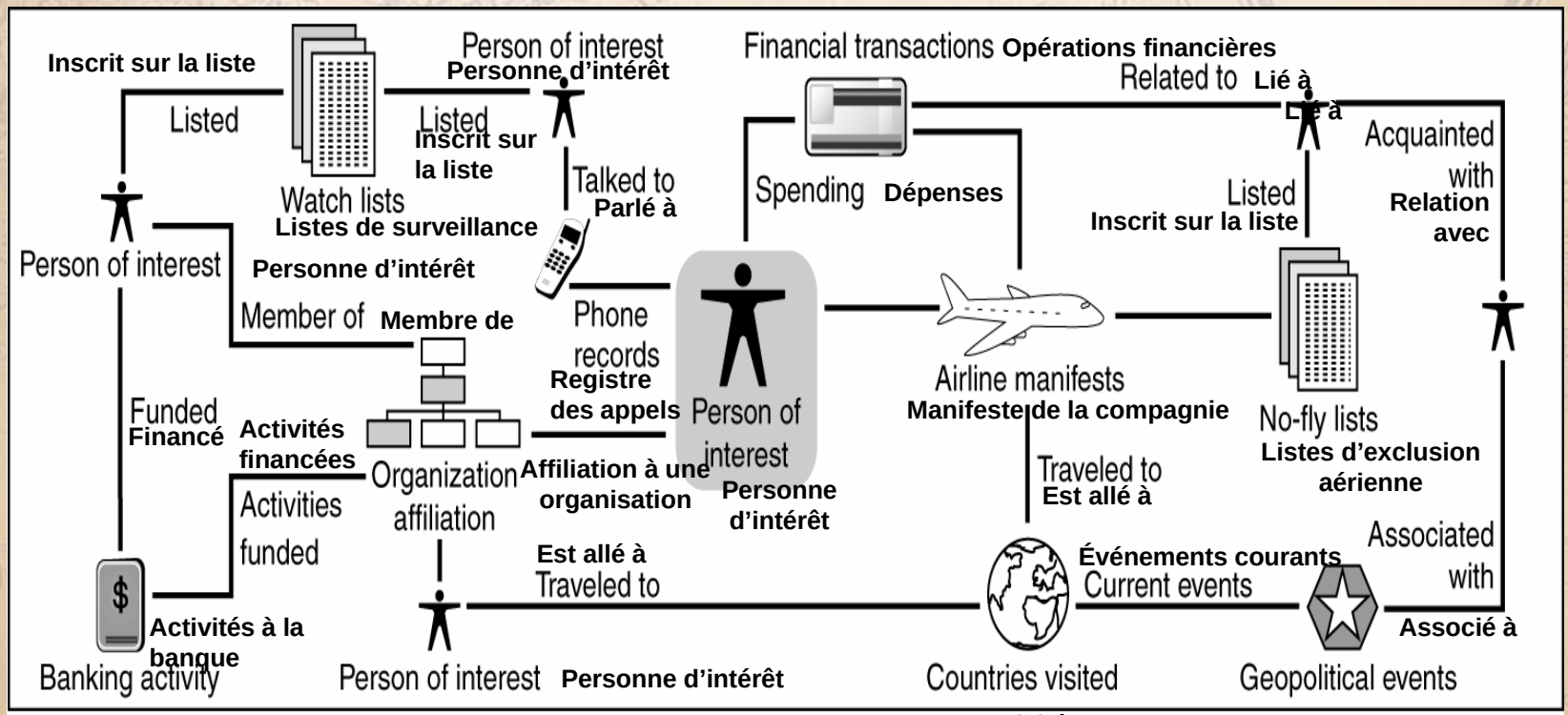
Source: GAO analysis of agency data. **Source :** Analyse des données d'organismes faite par le GAO

Tableau 1 : Principales mesures que doivent prendre les organismes afin de protéger la vie privée, et exemples de procédures et sources connexes

Principales étapes à suivre pour protéger les renseignements personnels	Exemples de procédures	Principales sources réglementaires
Publier des avis dans le *Federal Register* au moment de créer ou de modifier des systèmes de dossiers	• Préciser les utilisations courantes du système • Donner le nom du responsable du système • Décrire brièvement les procédures que peuvent suivre les personnes pour avoir accès à leurs dossiers	• Privacy Act
Assurer aux personnes l'accès à leurs dossiers	• Permettre aux personnes d'examiner les dossiers qui les concernent • Permettre aux personnes de demander d'apporter des corrections à leurs dossiers	• Privacy Act
Notifier les personnes du motif de l'autorisation obtenue pour recueillir les renseignements	• Aviser les personnes du responsable qui a autorisé l'organisme à recueillir des renseignements • Aviser les personnes des principaux motifs pour lesquels les renseignements sont utilisés	• Privacy Act
Appliquer des directives sur le système	• Effectuer une évaluation du risque afin de déterminer les vulnérabilités du système d'information, de cerner les menaces et d'élaborer des contre-mesures pour faire face à ces menaces • Faire en sorte que la direction certifie et accrédite le système • Veiller à l'exactitude, à la pertinence, à l'actualité et au caractère complet de l'information	• FISMA • Privacy Act
Évaluer les facteurs relatifs à la vie privée	• Décrire et analyser la manière dont l'information est protégée • Décrire et analyser l'utilisation prévue de l'information • Faire examiner l'évaluation par le dirigeant principal de l'information ou quelqu'un de même niveau • Rendre l'évaluation disponible, dans la mesure du possible	• E-Government Act

L'outil de forage des données ADVISE

(Analyse, diffusion, visualisation, intuition et amélioration sémantique)



Cato Institute: Forage des données et terrorisme

- Essayer d'utiliser le forage de données de prédiction afin de découvrir des terroristes avant qu'ils passent à l'action serait faire un usage erroné, quoique subtile, des ressources de sécurité nationales.
- Vu qu'il y a un nombre relativement faible de tentatives chaque année et seulement un ou deux incidents terroristes d'importance tous les deux ans – chacun d'eux étant distinct en termes de planification et d'exécution –, il n'existe pas de tendances significatives montrant le comportement qui indique la planification et la préparation d'actes terroristes.

Le forage des données dans le secteur privé

- Nous produisons une quantité phénoménale de données issues de nos transactions quotidiennes (achat de marchandises, inscription à des cours, etc.), consultations de sites Web et interactions avec le gouvernement (impôt, recensement, immatriculation de la voiture, inscription électorale, etc.). Non seulement le nombre de dossiers que nous créons s'accroît, mais la quantité de données recueillies pour chaque type de dossier augmente également.
- En tant qu'explorateurs de données, nos tâches se heurtent à ces préoccupations. En gestion analytique des relations avec le client, nous analysons souvent les données de ce dernier en cherchant précisément à saisir le comportement des personnes et à créer des campagnes de vente fondées sur cette compréhension. Les chercheurs dans les domaines de l'économie, de la démographie, de la médecine et des sciences sociales tentent de mieux cerner la relation entre les comportements et les résultats.
- Comment pouvons-nous faire coïncider les besoins légitimes des entreprises et des chercheurs avec le désir tout aussi légitime des gens de protéger leur vie privée?

Utilisation de l'anonymat

- Tout de même, les technologies de préservation de l'anonymat ont été endossées à maintes reprises par des groupes d'experts nommés pour examiner les conséquences du forage des données. Des progrès intéressants semblent avoir été faits dans la conception de systèmes de recherche de l'information qui intègrent l'anonymisation, les listes de contrôle des utilisateurs — capables de confirmer que personne n'a examiné les dossiers au-delà de ce qui était autorisé aux fins de l'enquête — et d'autres mécanismes de protection de la vie privée.
- L'astuce consiste à aller plus loin que de simplement prendre des noms dans les dossiers. Latanya Sweeney de l'Université Carnegie Mellon — technologue de premier plan dans le domaine de la protection de la vie privée qui a déjà exécuté un projet financé par la TIA — a démontré que 87 % des Américains pouvaient être identifiés au moyen de dossiers où figuraient uniquement leur date de naissance, leur sexe et leur code postal.
- Cette idée est venue à l'esprit de Sweeney lorsqu'elle mettait au point pour le compte du Ddépartement américain du logement et du développement urbain une méthode visant à repérer de manière anonyme les sans-abri.

Un exemple dans le secteur privé

- Tesco dresse discrètement un profil sur vous tout comme sur chaque personne au pays — un portrait de votre personnalité, de vos habitudes de voyages, de vos préférences d'achat, qui dit même si vous êtes charitable et respectueux de l'environnement. Une filiale de la chaîne de supermarché a établi une base de données, appelée Crucible, qui recueille des informations détaillées sur chaque ménage du Royaume-Uni, qu'ils décident de faire leurs courses chez ce détaillant ou non.
- La compagnie refuse de révéler l'information qu'elle détient, pourtant Tesco vend l'accès à cette base de données à d'autres importants groupes de consommateurs comme Sky, Orange et Gillette. « Elle renferme des données sur chaque consommateur du Royaume-Uni, y compris l'adresse domiciliaire, et toute une gamme de données démographiques, socioéconomiques ou liées à leur mode de vie », d'après l'annonce publicitaire de Dunnhumby, la filiale de Tesco en question. Elle a ajouté « ciblage et un profilage intelligent » à ses données au moyen d'un système logiciel appelé Zodiac. Ce profilage peut catégoriser votre enthousiasme pour les promotions, votre fidélité à une marque, vos habitudes et à quel moment vous préférez faire des courses. Comme le dit la notice : « La liste est sans fin si vous connaissez bien ce que vous cherchez. »

La vue qu'on a à 9 144 m