

Les données personnelles

Techniques d'anonymisation



Guillaume Raschia @univ-nantes.fr



UNIVERSITÉ DE NANTES

Vie privée

- Droit fondamental, affirmé en 1948 dans l'article 12 de la Déclaration universelle des droits de l'homme des Nations unies
- En France, l'article 9 du Code civil protège ce droit depuis la loi du 17 juillet 1970

La sphère privée

- protection du domicile
- secret professionnel et médical
- protection de l'image
- protection de l'intimité
- écoutes téléphoniques réglementées

Quel périmètre dans la société numérique ?

Usages de l'anonymat

Secret du vote

Alcoolique anonyme

Revendication d'anonymat pour les 500 signatures de l'élection présidentielles

Accouchement sous X

Anonymisation des requêtes sur les moteurs de recherche

Argent liquide

Télécarte, téléphone mobile sans abonnement

Anonymat des gagnants du Loto

Anonymat des contributeurs de Wikipedia

Lettre de dénonciation anonyme

Anonymat des sources

Protections des témoins

Centre d'appels info Sida

Secret bancaire



Numéro d'appel maltraitance

L'identité à la Légion Étrangère

Anonymat du don d'organes

Anonymat des agents en fonction (GIGN, RAID)

Anonymat des copies d'examens

Anonymat de l'offre dans les marchés publics

Anonymat des enquêtes, questionnaires, sondages

Bons au porteur

CV anonyme

Source : travail réalisé par des élèves de l'EHESP (École des Hautes Études en Santé Publique) de Rennes lors d'une formation dispensée le 6 mars 2008 par A. Belleil et Y Le Hegarat.

Valeur morale

hors-propos



De quoi parle-t-on ?

- **Anonymat** : “—possibilité de suivre une personne unique dans la durée avec—impossibilité de connaître sa véritable identité” (Lexique AFCDP)
- **Anonymisation** : “suppression des données à caractère personnel” (id.)

Mais aussi...

- **Pseudonymat** : anonymat *réversible* avec une responsabilité juridique
e-Commerce : pseudos pour particuliers uniquement (LCEN, 2004)
- **Hétéronymat** : plusieurs identités autonomes (70 écrivains = F. Pessoa)
- **Homonymat (?)** : banalisation de l'identité

Mais encore

- Norme ISO 15408
- Non-chaînabilité : impossibilité—pour un tiers—d'établir un lien entre différentes opérations faites par un même individu
- Non-observabilité : impossibilité—pour un tiers—de déterminer si une opération est en cours
- *pseudo < ano < non-chaîn < non-obs*

Au menu



- **Chapitre 1 Questions de vie privée**
- **Chapitre 2 Cadre juridique & enjeu de société**
- **Chapitre 3 Les PET's**
- **Chapitre 4 Anonymisation**

Vie privée

Faits et actualités

- Fichage et traçage
- Exposition de soi
- « Sousveillance »



Profil individuel

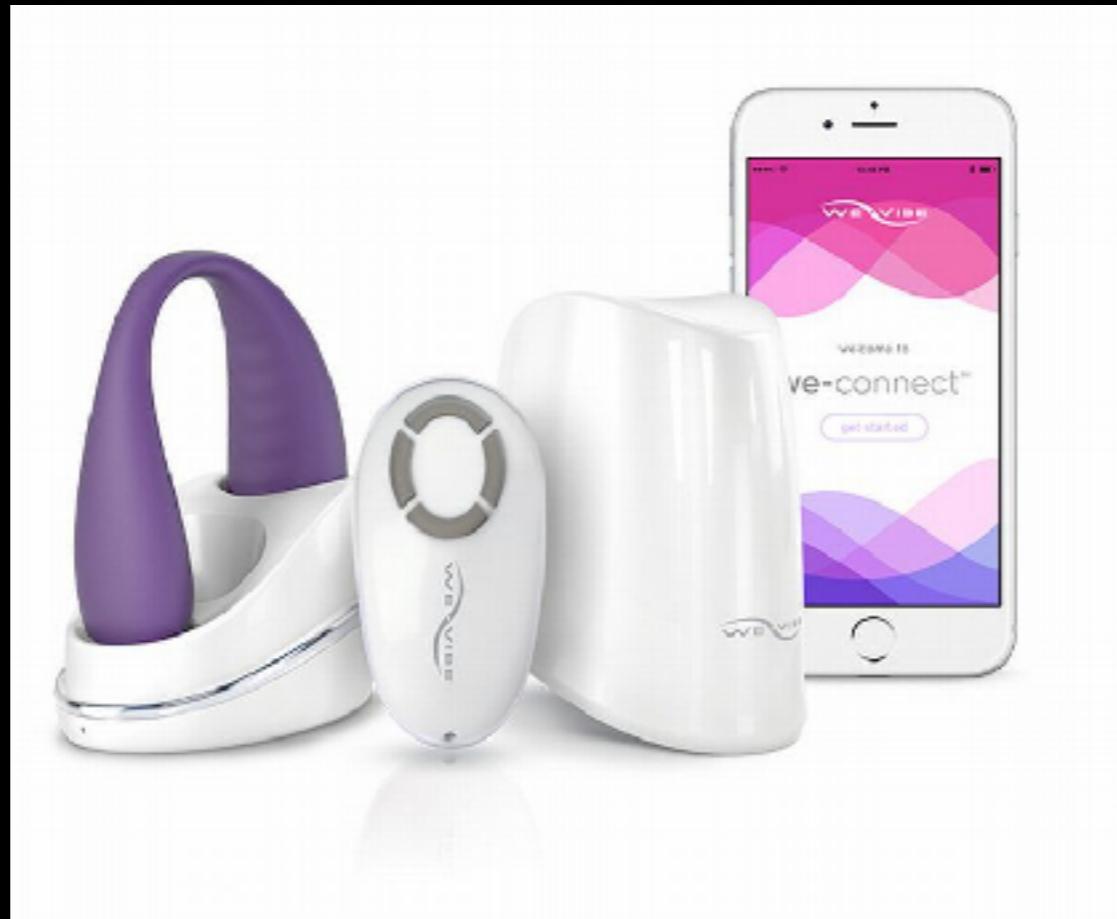
- *Trace d'activité* : préférence, consommation, évaluation, déplacement, communication...
- *Auxiliaires de collecte* : sites marchands, services en ligne, apps, titres de transport, cartes bancaires, cartes de fidélité, objets connectés, sondes et capteurs
- *Technologies de transmission* : Ethernet, wi-fi, puces sans contact (RFID), réseaux mobiles, GPS

Vizio



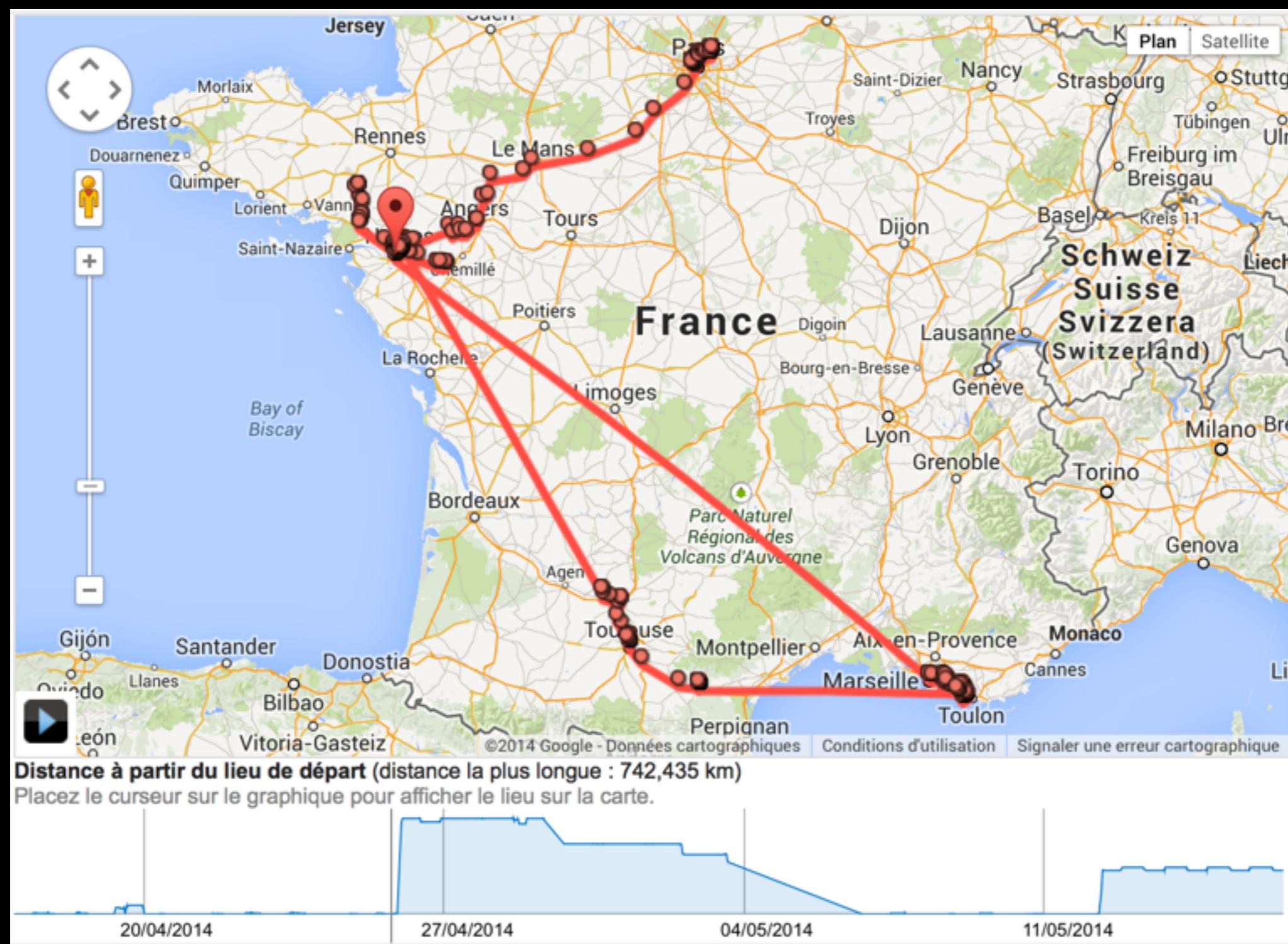
Février 2017 : collecte de données à partir
de la Smart TV de 11 millions de clients
amende = \$2,2 millions

Standard Innovation



mars 2017 : collecte à partir de We-Vibe
amende = \$3 millions

<https://maps.google.com/locationhistory/>



Paranoïa ?



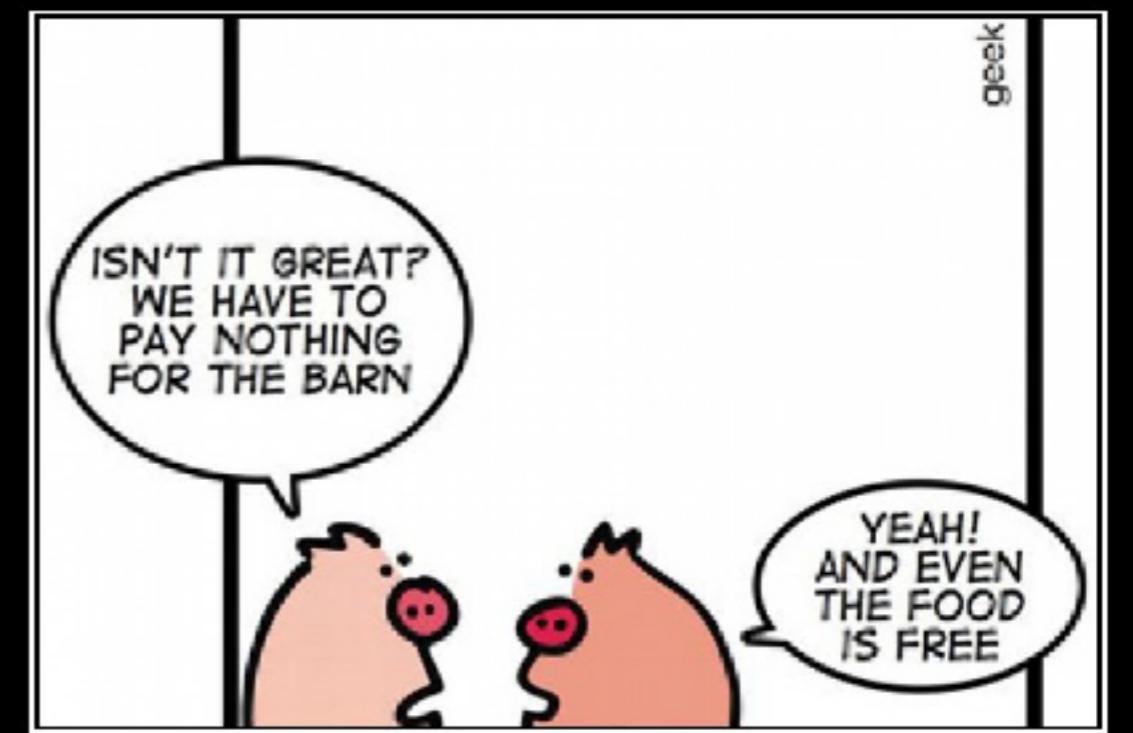
Echelle mondiale



- Cloud : données et services externalisés



The Big Four: GAFA



FACEBOOK AND YOU

If you're not paying for it, you're not the customer. You're the product being sold.

28 mars 2017



Le Congrès américain autorise les fournisseurs d'accès à vendre les données d...

Le Monde · Il y a 1 jour

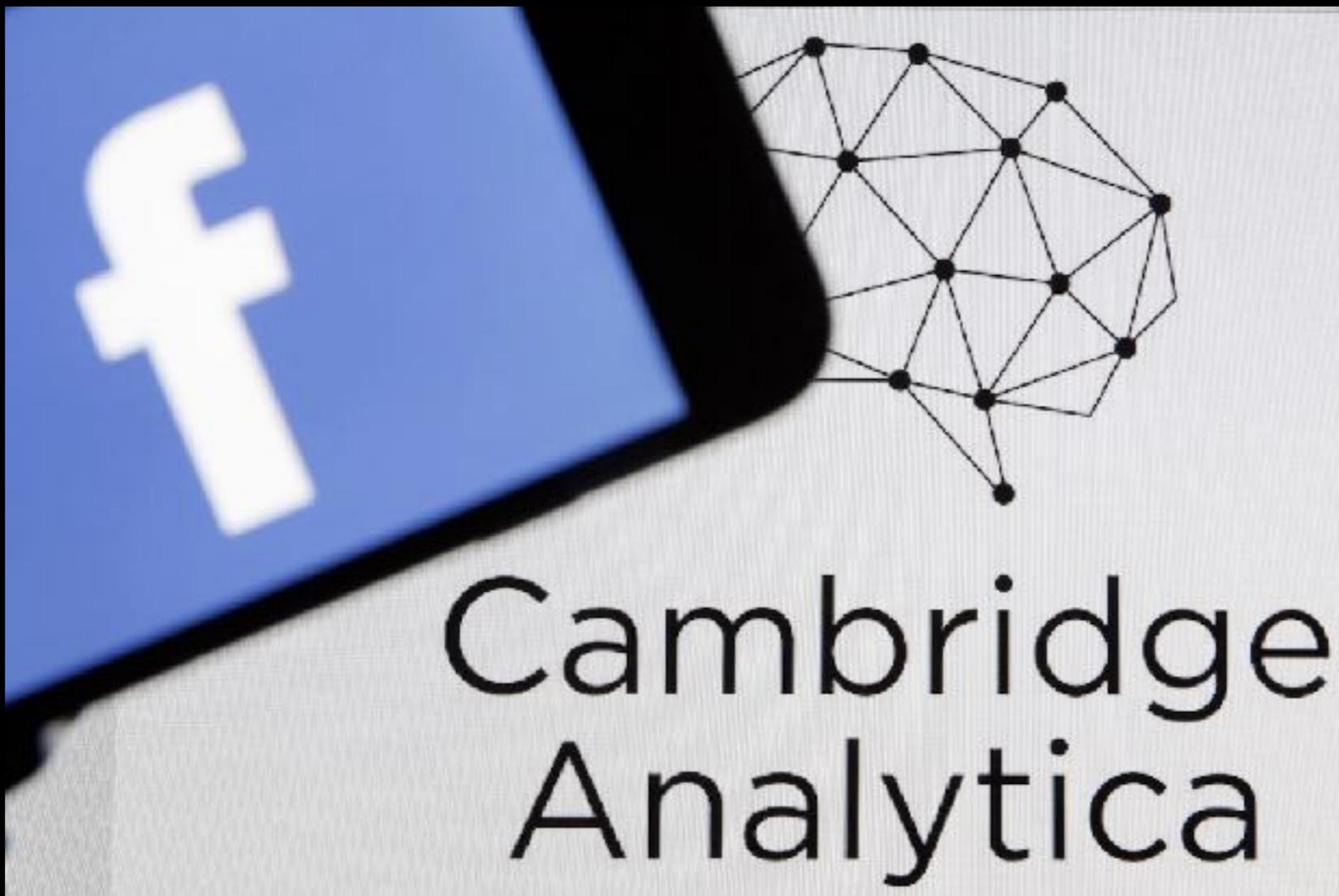
Quand le Congrès américain torpille la vie privée en ligne

Libération · Il y a 18 heures

Protection des données personnelles des internautes : recul du Congrès américain |...

ICI.Radio-Canada.ca · Il y a 1...

17 mars 2018



Cambridge Analytica

- 87 millions de profils Facebook siphonnés

Surveillance de masse

- Pratique des Etats : renseignements, police, *anti-terrorisme*
 - Aux USA : programme PRISM de la NSA
 - révélations de Edward Snowden (juin 2013—?)
 - En Europe : directive 2006/24/CE
 - conservation des données de connexion
 - invalidée par la Cour de justice de l'UE le 08 avril 2014
 - En France : article 20 de la Loi de Prog. Militaire

Fichiers de police



- **58 fichiers en 2009**
 - Edvige/Edvirsp : RG et DST :: données sensibles à partir de 13 ans
 - FNAEG : 1.276.769 fiches (2%) en Avr. 2010 vs. 2.100 en 2002 (+1000/j)
 - Traitement des Procédures Judiciaires TPJ : auteurs, victimes, témoins
 - STIC : 44,5M pers., 79% erreurs en 2010
 - JUDEX : 2,15M pers., 48% erreurs en 2009
 - NS2I : Nouveau Système d'Information dédié à l'Investigation

Les informations contenues dans cette fiche ont SIMPLE VALEUR DE RENSEIGNEMENT
susceptible D'ORIENTER L'ENQUETE. Il ne pourra en être fait état
que sous réserve de VERIFICATION

SMET, JEAN-PHILIPPE Né le 15/06/1943 à Paris 09^e
Père : SMET Sexe : MASCULIN Situation maritale : Niveau d'études :
Nationalité : INDETERMINEE
Resident : Séjour :
Validité état-civil : IDENTITE DECLARÉE Alias : HALLIDAY JOHNNY Né le 15/06/1943 à PARIS 09^e
Nationalité : INDETERMINEE
Stat :

Emmêlages

Profession : Non enregistrée ou inconnue Photo : Non enregistrée ou inconnue

Cette personne a été citée dans cette procédure pour le ou les faits suivants mais en aucun cas il ne peut être déduit de ce document qu'elle a été reconnue comme responsable des faits.

Procédure : - DIV STAT ET DOC CRIM DRPJ PARIS, N° 1992/009260

Archivages : - DIV STAT ET DOC CRIM DRPJ PARIS, N°1992/0199852/DOS COLLECTIF
Situation du mis en cause : DEFERE Suites judiciaires : inconnues

Cité comme AUTEUR : ESCROQUERIE

Cité comme AUTEUR : ABUS DE BIENS SOCIAUX

Faits commis le 14/01/1992 à PARIS

Procédure : - DIV STAT ET DOC CRIM DRPJ PARIS, N° 1973/005339

Archivages : - DIV STAT ET DOC CRIM DRPJ PARIS, N° 1973/0646945/DOS INDIVIDUEL
Situation du mis en cause : DEFERE Suites judiciaires : Inconnue

Cité comme AUTEUR : VIOLENCES VOLONTAIRES

Cité comme AUTEUR : OUTRAGE AUX BONNES MOEURS

Faits commis le 22/01/1973 à PARIS

Cité comme AUTEUR : VIOLENCES VOLONTAIRES

Faits commis le 05/09/1973 à PARIS 08^e

Cité comme AUTEUR : INFRACTION À LA LEGISLATION SUR LES ARMES

Faits commis le 26/02/1973 à PARIS 08^e

Procédure : - DIV STAT ET DOC CRIM DRPJ PARIS, N°1963/001115

Archivages : - DIV STAT ET DOC CRIM DRPJ PARIS, N°1963/0775231/DOS INDIVIDUEL
Situation du mis en cause : DEFERE Suites judiciaires : Inconnue

Cité comme AUTEUR : OUTRAGE A AGENT DE LA FORCE PUBLIQUE

Cité comme AUTEUR : REBELLION

Faits commis le 27/03/1972 à PARIS 08^e

STIC de J. H.

- Profession : *non enregistrée ou inconnue*
- Nationalité : *indéterminée*
- 26.10.67, Paris XVle : auteur de *violences volontaires* [Déféré]
- [suites judiciaires inconnues] pour
 - *violence volontaire (1973)*
 - *outrage aux bonnes moeurs (1973)*
 - *infraction à la législation sur les armes (1975)*
 - *abus de biens sociaux (1992)*
 - *escroquerie (1992)*
- **absence de mises à jour**
- **conservation illégale des données (15/40 ans)**
- **revente d'infos ("tricoche")**

TES

- Titre Electronique Sécurisé

- déjà 29 millions de français (demandes de passeport depuis 2008)
 - données d'état civil, avec filiation
 - couleur des yeux, taille, adresse
 - photo, empreinte digitale, signature
- projet de généralisation pour délivrer des cartes d'identité
- 30/10/2016 : expérimentation et audit (rapports critiques)
- 28/03/2017 : application sur tout le territoire national

1974 SAFARI * 2012 ~~CNI~~ * 2017 TES !



Log de requêtes

- Durée de rétention
 - Juin 2007 : Google à 18 mois (vs. 24 mois)
 - Avr. 2008 : G29 recommande <6 mois
 - Juil. 2008 : Ixquick : label européen (48h!)
 - Sept. 2008 : Google à 9 mois
 - Déc. 2008 : Microsoft à 6 mois... avec conc.
 - Déc. 2008 : Yahoo! à 3 mois... en 2010
 - Jan. 2010 : MS Bing à 6 mois... en 2011

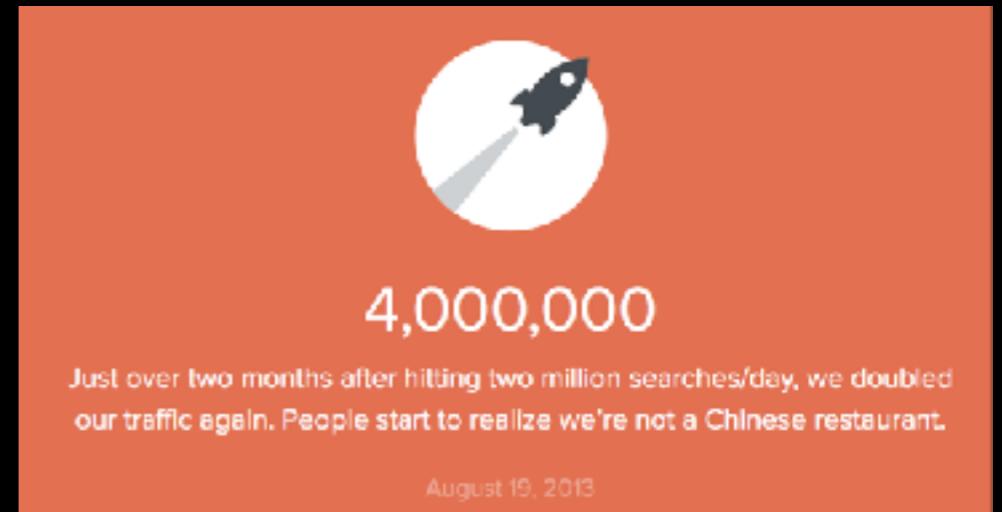


Réversibilité



DuckDuckGo

- Robustesse de la méthode d'anonymisation
@IP et sessionIDs (cookies)
 - Microsoft : @IP seulement
 - Google : anonymat réversible...
- <https://duckduckgo.com>
 - 9.000.000 requêtes/jour



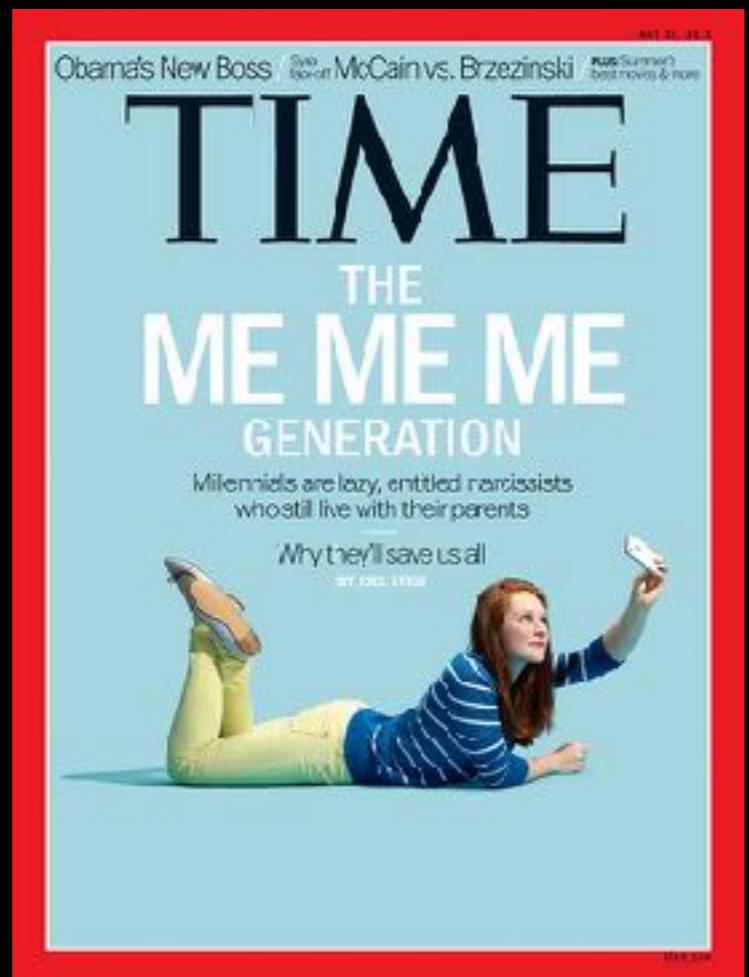
Réseaux sociaux

les conditions sont toujours aussi inaccessibles, illisibles, remplies de liens hypertextes – entre 40 et 100 liens hypertextes - renvoyant parfois à des pages en langue anglaise. Pire, les réseaux persistent à s'autoriser très largement la collecte, la modification, la conservation et l'exploitation des données des utilisateurs et même de leur entourage (« amis », « followers », « +1 », etc.)... Ils s'octroient toujours, sans l'accord particulier des utilisateurs, une licence mondiale, illimitée et sans rémunération, d'exploitation et de communication des données à des partenaires économiques.

Source : UFC Que Choisir, communiqué de presse, 15 mars 2014



Réseaux sociaux



- Droit européen vs. droit U.S.
personnalité vs. données commerciales
- Longue dépression de *Nathalie Blanchard*

Réseaux sociaux

12 Juin 2009 : avis G29 sur réseaux sociaux

définir des paramètres par défaut limitant la diffusion des données des internautes

mettre en place des mesures pour protéger les mineurs

supprimer les comptes qui sont restés inactifs pendant une longue période permettre aux personnes, même si elles ne sont pas membres des réseaux sociaux, de bénéficier d'un droit de suppression des données qui les concernent

proposer aux internautes d'utiliser un pseudonyme, plutôt que leur identité réelle

mettre en place un outil accessible aux membres et aux non membres, sur la page d'accueil des réseaux sociaux, permettant de déposer des plaintes relatives à la vie privée.

Révélations

- *Portrait Google de Marc L****
Revue Le Tigre, vol.28 (Nov. 2008)

• Flickr : “voyages” 17.000 photos + Facebook + Youtube + ...
“Elle a habité successivement Angers puis Metz, son chat s’appelle Lula, et, physiquement, elle a un peu le même genre que Claudia. À l’été 2006, vous êtes partis dans un camping à Pornic, dans une Golf blanche. La côte Atlantique, puis la Bretagne intérieure. Tu avais les cheveux courts, à l’époque, ça t’allait moins bien.”
• Graphe relationnel :

- Gaydar (2007), MIT
- Prévisions diverses : opinion politique, genre, localisation, race du chien



« Sousveillance » ambivalente

- Lanceurs d'alerte
 - Affaire NSA et Edward Snowden
 - Wikileaks
 - @ délinquants sexuels : familywatchdog.us
- Pilori numérique (« *bashing* »)
 - Ghyslain Raza, avr. 2003 : *Star Wars Kid*
 - Affaire de la *Dog Poop Girl*, 2005



Où en sommes-nous ?

- Chapitre 1 Autour des données personnelles
- **Chapitre 2 Cadre juridique & enjeu de société**
- Chapitre 3 Les PETs
- Chapitre 4 Anonymisation



La vie privée

- Droit fondamental consacré (D.U.D.H. 1948)
- Une liberté vs. la sécurité “*rien à me reprocher*”
- Une liberté vs. des avantages économiques
- Progrès technologiques
- Exposition et mesure de soi
- Application difficile du cadre juridique



La législation

- Loi informatique et liberté du 6 jan. 1978
 - Création de la CNIL
- Directive européenne 95/46/CE de oct. 1995
- Transposition en droit Fr : loi du 6 août 2004 décret n° 2005-1309 du 20 oct. 2005
- Règlement Eu 2016/679 : **25 mai 2018 !**
- Données à caractère personnel | traitement | fichiers

“à caractère personnel”

- Toute information relative à une personne physique identifiée ou identifiable, directement ou indirectement, telle que :
 - nom, prénom, photographie, biométrie (empreinte digitale, etc.)
 - adresse, numéro de S.S.
 - numéro de tél./carte d'identité/compte bancaire/matricule/dossier/etc.
 - plaque d'immatriculation automobile
 - @mail, cookie (sessionId), etc.
- Cas épineux de l’@IP

Les données sensibles

- Des données « très » personnelles
 - origines ethniques
 - opinions politiques, philosophiques, religieuses
 - appartenance syndicale
 - orientation et pratiques sexuelles
 - données de santé
 - données génétiques
 - données biométriques

Cas pratique

A votre avis, la collecte des données suivantes doit-elle être considérée comme un traitement de données personnelles ?

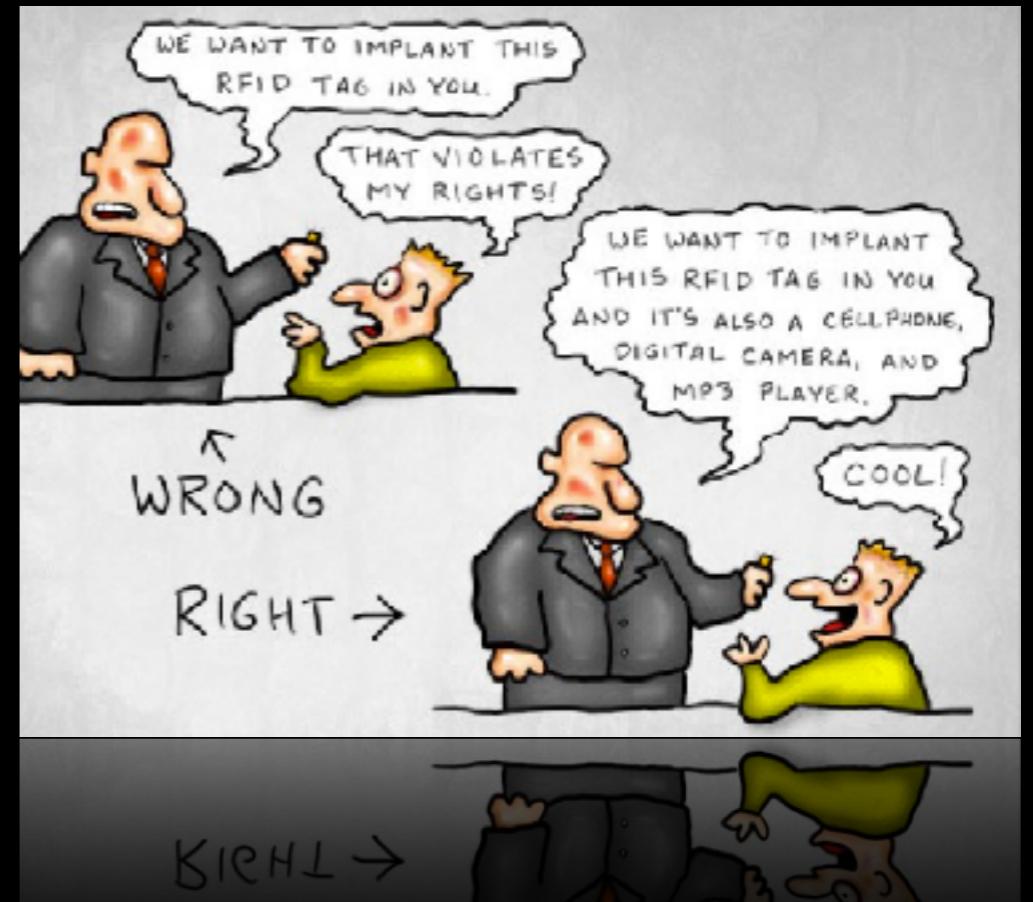
Année de naissance	Lieu de naissance	Profession	Nationalité	Date de l'interruption volontaire de grossesse
1978	Paris	Enseignante	FR	10/10/2000
1976	Strasbourg	Agricultrice	FR	11/08/1999
1975	Ceintrey	Avocate	IT	12/01/1995
1974	Lyon	Commerçante	DK	23/03/1996

Réponse de la CNIL : effectivement, il s'agit bien d'un traitement de données indirectement nominatives (la question était posée avant 2004). Pourquoi ? Simplement en raison du risque d'atteinte à la vie privée qui résulterait de l'identification des personnes. Risque suffisant pour dicter à l'autorité administrative la plus grande prudence dans sa décision.

Source : post de blog <http://www.donneespersonnelles.fr/adresse-ip-est-elle-donnee-personnelle> par T. Devergranne le 18/11/2011

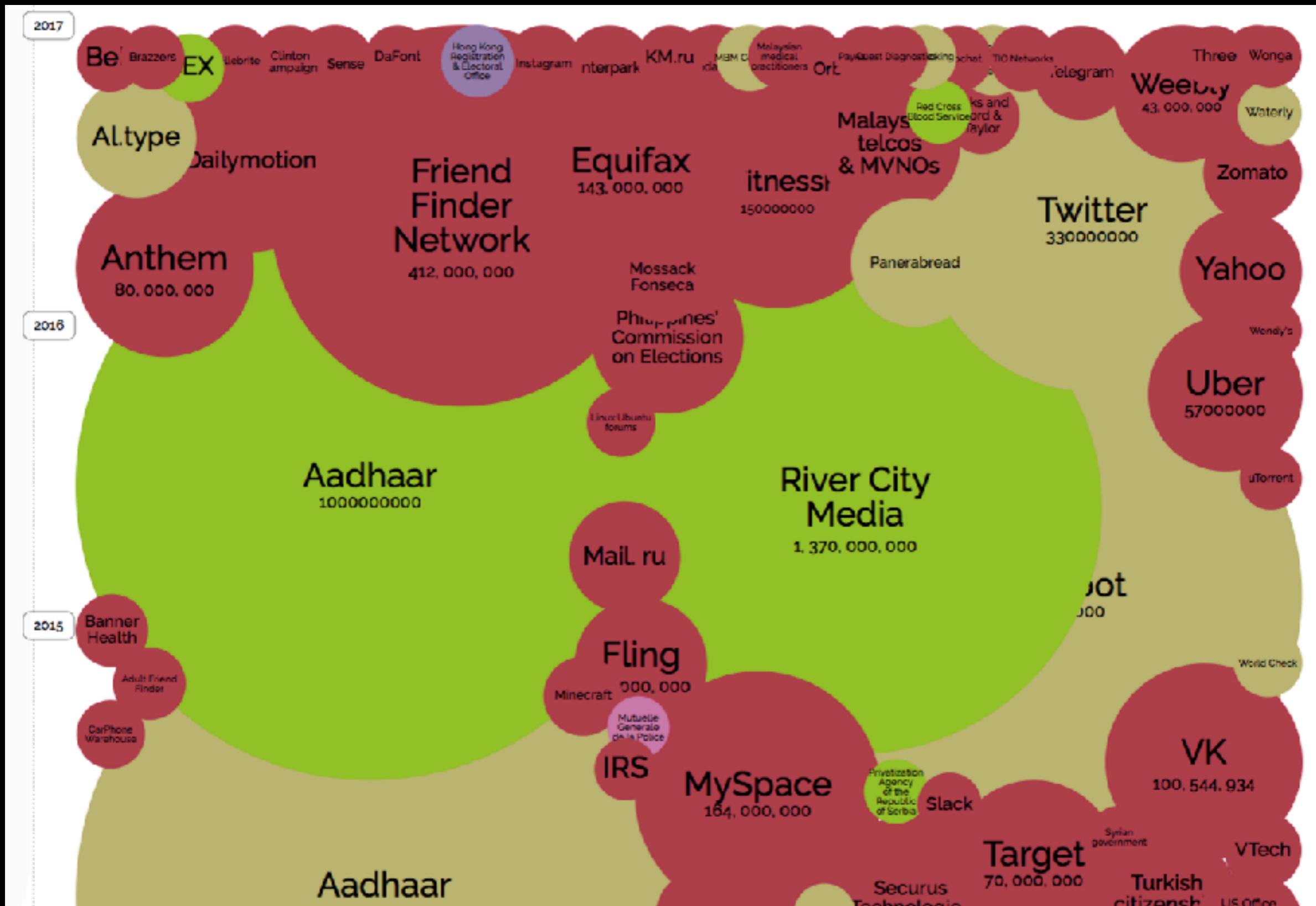
Les raisons de la collecte

- **Obligation** légale et réglementaire
 - état civil, dossier médical, impôts, recensement, etc.
- **Contribution active** à des études
 - épidémie de grippe HINI
- **Exposition de soi**
 - Facebook, Flickr
- **Avantage économique**



Les risques

- Usage détourné
- Revente
 - marché noir des données
 - Inccohérence des fichiers
- Fuite de données
 - Exploitation d'une faille de sécurité
 - Malveillance interne
 - Maladresse



informationisbeautiful.net / données: databreaches.net

Les conséquences

- **Usurpation d'identité**
 - 212.762 vols d'identité / an en France (CREDOC)
- **Réputation, crédibilité et vie privée**
 - 35% recruteurs ont déjà écarté un candidat ayant une mauvaise e-réputation
- **Intégrité physique et morale**
 - opposition politique
 - cambriolage, vendetta, etc.

RGPD

Règlement Général sur la Protection des Données

- Uniformiser les règles communautaires
- Sanctuariser le droit des personnes
- Renforcer les sanctions
- Clarifier les responsabilités
- Cadrer la conformité

Les acteurs

- La **personne**
- Responsable de traitement (**Data Controller**): entité qui détermine les **finalités** et les moyens (pas de décharge en cas de sous-traitance)
- Sous-traitant (**Data Processor**)
- Destinataire : récipiendaire des données
- Tiers

Les droits CNIL

- Droit d'information
- Consentement (*Opt-in*)
- Droit d'opposition légitimée
- Droit d'accès direct ou indirect
- Droit de rectification
- Durée de conservation limitée au service

Complément RGPD

- Droit à l'oubli (ou effacement)
- Portabilité
- Réclamation, recours et réparation
- Consentement pour les mineurs

Les obligations de conformité

- Contrats de sous-traitance
- Délégué à la protection des données (*Data Privacy Officer*)
- Politique de sécurité/intégrité
- Registre des activités de traitement
- Notification des violations de données
- Analyse d'impact

Les obligations de conformité

- Documentation
- Pas de déclaration CNIL préalable
- Analyse d'impact (*Data Protection Impact Assessment*)
 - des traitements présentant un risque élevé pour les droits et libertés
 - à soumettre à l'autorité de contrôle (CNIL)

Les sanctions

- Enquêtes, contrôles, audits
- Notification des violations de données
- Avertissement, mise en demeure
- Suspension de flux et/ou certification
- Amende administrative
 - 20 millions € ou 4% du CA annuel mondial
 - Principe de proportionnalité et de dissuasion

Le DPO

- ex-Correspondant Informatique et Libertés
- référent pour une collectivité territoriale, une entreprise publique ou privée, une association
- « chef d'orchestre » de la conformité
- rôle d'information, contrôle, conseil
- Délégué à la protection des données (*Data Privacy Officer*) à partir de mai 2018

Nouveaux paradigmes

- Privacy by Design
- Privacy by Default

Moyens techniques

- Chiffrage
- Pseudonymisation
- Anonymisation

Périmètre d'application

- les entreprises privées ou publiques qui :
- proposent des biens et services dans le marché de l'UE
- collectent et traitent des données personnelles sur les résidents de l'UE
- sont concernées les entreprises hors-UE !

« Sphère de sécurité »

- RGPD et transferts trans-frontaliers
- *Safe Harbor*
- Passerelle droit Eu/droit US pour la protection des données personnelles des ressortissants Eu aux USA

Max Schrems



- Citoyen autrichien, étudiant en droit
- Fondateur de <http://europe-v-facebook.org/>
 - 2011 : 1200 pages de données Facebook pour 22 plaintes
 - 2013 : révélations de E. Snowden sur la surveillance de masse de la NSA, nouvelle plainte contre Facebook (à Dublin, siège Eu)
 - 06/10/2015 : invalidation du Safe Harbor par la Cour européenne de justice
 - Depuis le 01/08/2016 : *Privacy Shield...*

<https://www.privacyshield.gov>

Où en sommes-nous ?

- Chapitre 1 Autour des données personnelles
- Chapitre 2 Cadre juridique & enjeu de société
- **Chapitre 3 Les PETs**
- Chapitre 4 Anonymisation



Les PETs

Privacy-Enhancing Technologies

<http://cyberlaw.stanford.edu/wiki/index.php/PET>

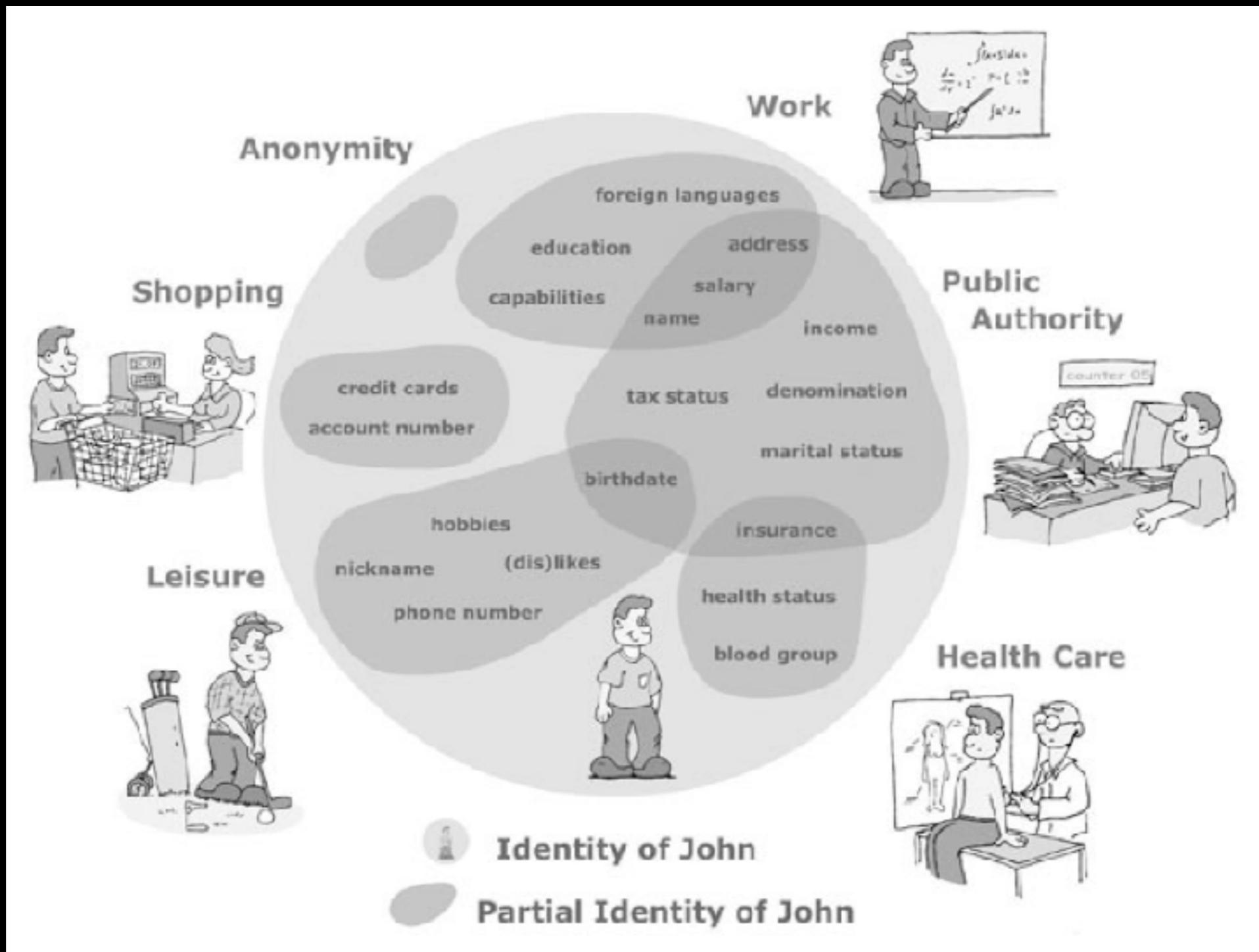
- Technologies pour la protection de la vie privée
 - Gestion d'identités multiples
 - Autorisation respectant la vie privée
 - Communication et accès anonyme
 - Gestion des données personnelles



xIDs

- Réduire/contrôler le lien personne-donnée
- Accès personnalisé et privilégié : hétéronymat
 - préférences : météo :: cookies
 - rôles : voyageur / électeur :: pseudos
 - niveaux de sécurité : Twitter / banque :: mdp|certificat|biométrie|RFID
 - durée de vie : CAF / DL ressource :: pseudos “jetables”
- Identités virtuelles multiples vs. Single-Sign-On
 - I-card  vs. OpenID 

identités partielles



Source : PRIME Primer (rev. 2006)

@IP : identification

Received: from localhost (debian [127.0.0.1])
by smtp-tls.univ-nantes.fr (Postfix) with ESMTP id 214E812810D
for ; Tue, 13 Oct 2009 12:02:55 +0200 (CEST)
X-Virus-Scanned: Debian amavisd-new at univ-nantes.fr
Received: from smtp-tls.univ-nantes.fr ([127.0.0.1])
by localhost (SMTP-TLS.univ-nantes.fr [127.0.0.1]) (amavisd-new, port 10024)
with LMTP id YQzhP7Vh6Imy for ;
Tue, 13 Oct 2009 12:02:55 +0200 (CEST)
Received: from [192.168.248.203] (nantes.wifi.univ-nantes.fr [193.52.107.31])
(using TLSv1 with cipher AES128-SHA (128/128 bits))
(No client certificate requested)
by smtp-tls.univ-nantes.fr (Postfix) with ESMTPSA id 0C1221280E4
for ; Tue, 13 Oct 2009 12:02:55 +0200 (CEST)
Message-Id:
From: Guillaume Raschia
To: test@yopmail.com
Content-Type: text/plain
Content-Transfer-Encoding: 7bit
Mime-Version: 1.0 (Apple Message framework v936)
Subject: test : @IP?
Date: Tue, 13 Oct 2009 12:02:38 +0200
X-Mailer: Apple Mail (2.936)

@IP : contenu sensible



@IP : géo-localisation



@IPs dans le Futur

- IPV6, réseaux ad hoc : informatique ubiquitaire, intelligence ambiante, internet des objets, réseaux de capteurs, convergence 4G, etc.
- Une @IP unique et fixe / “machin”
- Plusieurs “machins” / personne
- Une communication et des échanges permanents



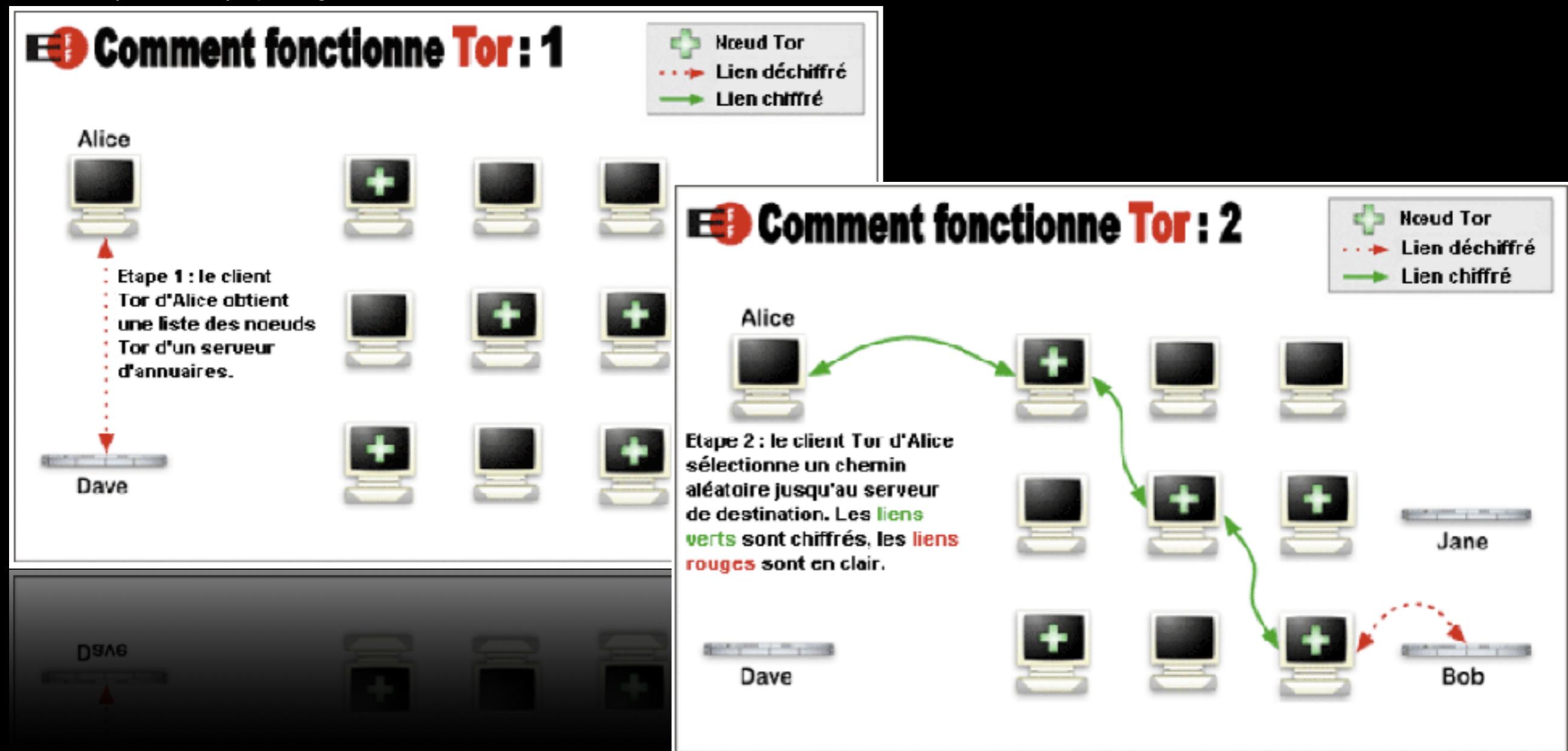
Communications anonymes

- Protection des @IP
- Affectation dynamique : DHCP, PPP, NAT, etc.
- Routeurs d'anonymat :
 - MIX [Chaum, 1981]
 - DCnet [Chaum, 1988]
 - Blind message service [Cooper & Birman, 1995]
 - Onion Routing [1996] / TOR [2004]
 - Crowds [1997]



Onion Routing

Source : <http://www.torproject.org/>



Accès anonyme

- Relais (*proxy*) d'anonymat
 - E-mail, news (Usenet) :
 - `nym servers :: anon.penet.fi` (700.000 utilisateurs en 1996!)
 - Cypherpunk / Mixmaster / Mixminion
 - Web :: proxify.com
- Serveurs de pseudonymes
 - @E-mail jetables
 - Identités multiples fournies par des F.A.I.

Preuves

- *Credential* : certificat, garantie, accréditation
- Multiplication des certificats :: PKI
 - carte d'abonnement, de membres d'association, etc.
 - permis de conduire, carte d'électeur, etc.
- P3P (*Platform for Privacy Preferences*) du W3C

Propriétés déterminantes

Confidentialité : communication secrète

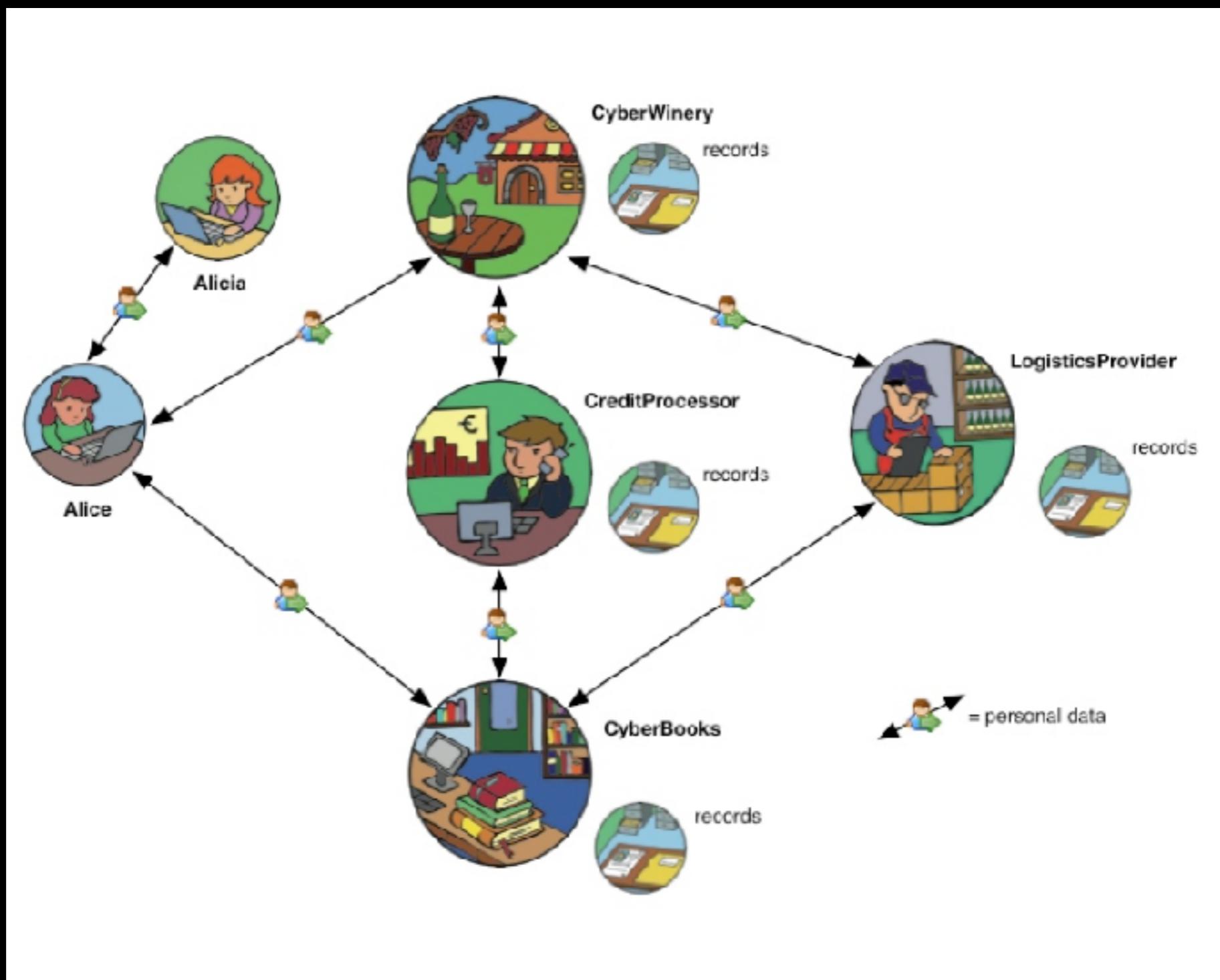
Authentification : auteur identifié

Intégrité : message intact

Non-répudiation : responsabilité

Primitives cryptographiques...

Qui détient mes données ?



Données personnelles

- **Minimisation** (*need-to-know*)
 - proportionnalité et finalités légitimes
 - fragmentation
 - **anonymisation** (*neutralisation*)
- **Souveraineté**
 - péremption, notification de transfert/usage

Où en sommes-nous ?

- Chapitre 1 Autour des données personnelles
- Chapitre 2 Cadre juridique & enjeu de société
- Chapitre 3 Les PETs
- Chapitre 4 Anonymisation

Pourquoi anonymiser ?

- Donnée « dé-personnalisée » = donnée neutralisée vis-à-vis de la loi
- Conformité à la loi :
 - durée de conservation : logs de requêtes
 - usage secondaire :
 - statistiques, prédiction, marché des données
 - réglementation : publication des décisions de justice
 - Open data Open data et protection de la vie privée, Rapport d'information de MM. Gaëtan GORCE et François PILLET, fait au nom de la commission des lois du Sénat, n° 469 (2013-2014) - 16 avril 2014
 - dé-responsabilisation : sous-traitance et transfert trans-frontalier

Données maltraitées

- Enquêtes Compuware et Institut Ponemon
- *Uncertainty of Data Breach Detection, 2008*
 - Violation de données : 62% professionnels IT (Fr) dont 52% interne
- *Test data and security : the unseen crisis, 2008*
 - Données personnelles/sensibles : 43% en pré-production (Fr)
 - 59% externalisation des tests (Fr)
 - 81% partage des données de production avec sous-traitants (Fr)

Révélation d'identité

- AOL (2006)

Source : TNYT. *A Face Is Exposed for AOL Searcher No. 4417749*, 2006



- 20M requêtes de 658.000 internautes, @IP hachée
- “chien qui fait pipi partout”, “taxe foncière de Harrisburg, Virginie”, etc.
- #4.417.749 = Thelma Arnold, veuve de 62 ans
- Exposition du fichier : <http://search-id.com>

- Concours Netflix (2006)

Source : A. Narayanan, V. Shmatikov.
How To Break Anonymity of the Netflix Prize Dataset, 2006

- 100M reco. 1999-2005 de 500.000 clients sur 10M titres
- Croisement avec IMDB : films+notes et #id
- avec 8 reco.+date (+/- 3 j.) : 96% abonnés identifiés

Le gouverneur malade

Source : L. Sweeney, *k-Anonymity: a Model for Protecting Privacy*, 2002

- GIC : dossiers médicaux des employés du MA
 - zip, sexe, dn, ethnie, l/O date, diag, medic, prix, ... (x100)
- Pour US20\$ registre d'électeurs Cambridge, MA
 - zip, sexe, dn, nom, prénom, @, affiliation, date_vote
- Croisement de fichiers !
- Gouverneur William Weld :
 - 6 dn > 3 hommes > 1 zip ! Données médicales confidentielles...

Mécanisme d'inférence

AdN	Sortie	Diagnostic
1990	12-12-2006	schizophrénie

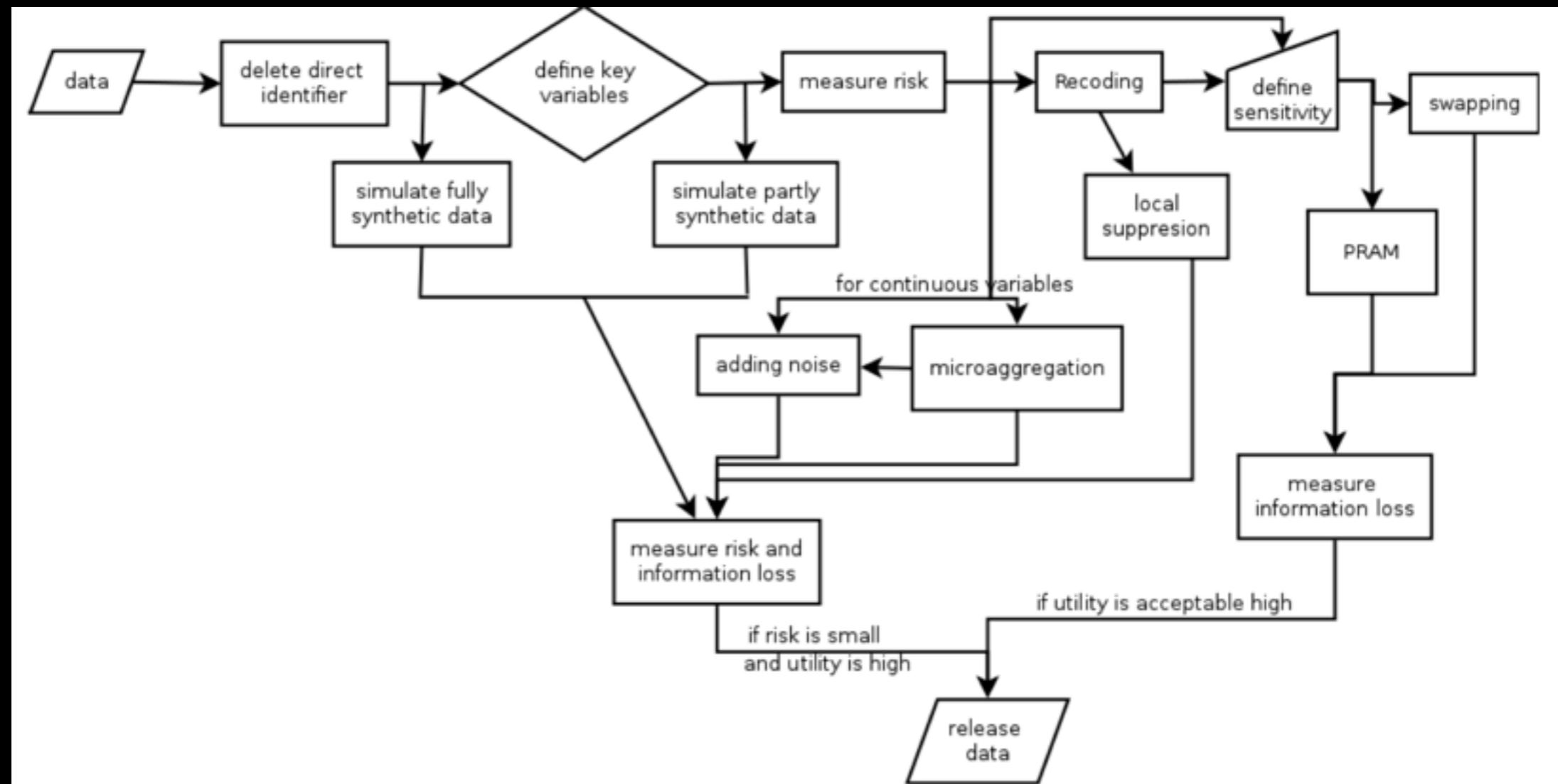
Nom	AdN	Sortie
Alice	1990	12-12-2006

- **Croisement de fichiers**

Source : présentation CNIL

- 762.407 naissances en France en 1990
- 45 sorties de l'hôpital le 12-12-2006
- 1990 et 12-12-2006 = 1 seule personne : elle est schizophrène.

La démarche



Source : Guidelines for the Anonymization of Microdata Using R-package sdcMicro (Version 1.0) B. Meindl, M. Templ and A. Kowarik, April 22, 2013

Quelques outils

- [déprécié] *Datafly / k-Similar*, Carnegie Mellon U.
- *sdcMicro* package R, Matthias Templ
- *ARX*, Université Technique de Munich
- *Projet Lamane*, Télécom Bretagne
- *u-Argus*, CASC project <http://neon.vb.cbs.nl/casc/>
- *DataMasker / DataBee*, Network 2000 Ltd.
- Compuware *Data Privacy*
- IBM *Optim Data Privacy* techno Princeton Softech

Le risque

- Modèle d'attaque : croisement de fichiers
- Quel est le nombre d'individus en danger ?
- Evaluation et quantification du risque

Le risque

- Identifiants indirects (quasi-id) Qid
 - Ex. : {Date-de-naissance, Genre, CP}
- Taille du domaine de définition : $|q_i|$
 - Ex. : $|CP| = 100.000$, de 00-000 à 99-999
- Nombre maximum de valeurs de Qid :

$$m = \prod_{Qid} |q_i|$$

Le risque (bis)

- Principe des tiroirs (*Pigeon-hole principle*)
 - Population de taille n , domaine des Qid de taille m
 - Enoncé :

il existe une valeur de Qid représentée au moins chez $\lceil n/m \rceil$ individus

Exemple

- Population de 100.000 individus
 - Âge $\in [0..99]$, $|CP|=500$, Genre $\in \{M, F\}$
- Existe-t-il un fragment de population qui ne peut être uniquement représenté sur les Qid's ?
 - $|Qid| = 100 \times 500 \times 2 = 100.000$
 - $100.000/100.000=1 \rightarrow \text{NON}$

Exemple

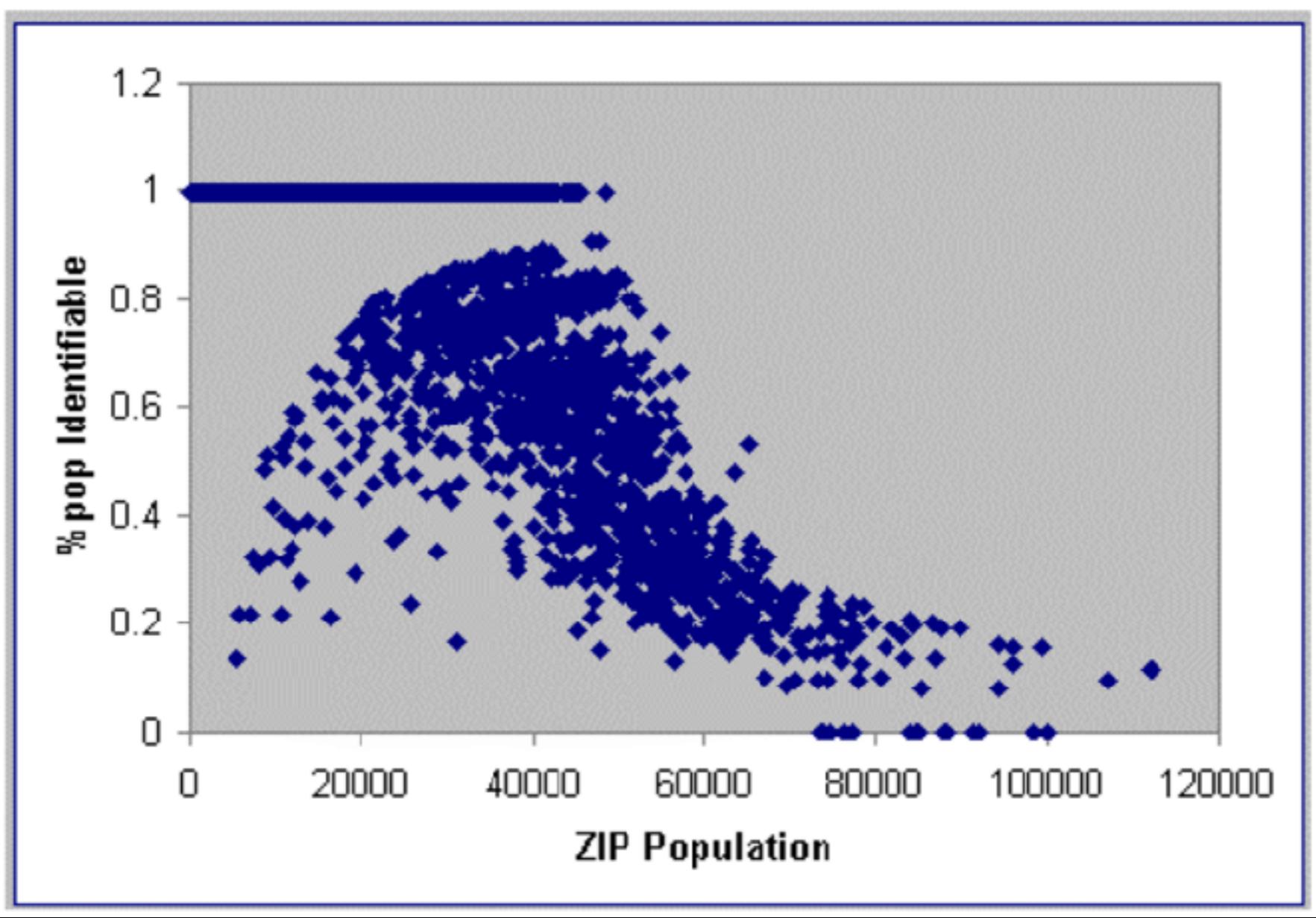
- Population de 100.000 individus
 - Âge $\in [0..50]$, $|CP|=50$, Genre $\in \{M, F\}$
- Existe-t-il un fragment de population qui ne peut être uniquement représenté sur les Qid's ?
 - $|Qid| = 50 \times 50 \times 2 = 5.000$
 - $100.000/5.000=20$ individus identiques

Les bornes

- #max d'individus pouvant être identifiés ?
 - $m-l$ si $m < n$, n sinon
- #min d'individus pouvant être identifiés ?
 - l si $n=l$, 0 sinon
- Avec au moins l individu par valeur de Qid's
 - $\#min=0$ si $n \geq 2m$, $2m-n$ sinon ($n \geq m$)

Quantification du risque

- Tableau de contingence à $\#Qid's$ dimensions
 - $\#cellules$ de valeur l
- Estimation possible des Qid's
 - Ex. : Année-de-naissance vers Date-de-naissance ?
 - Hypothèse de répartition uniforme (?)
 - La contingence de chaque année est répartie sur 365 jours



- ZIP+Genre+DdN. = 87% population U.S.
Sources : L. Sweeney, *Uniqueness of Simple Demographics in the U.S. Population*, 2000
P. Golle. *Revisiting the uniqueness of simple demographics in the US population*, WPES, pp. 77-80, 2006
- DdN 2 enfants = identité de la mère (Fr)
Source : G. Trouessin, JSSI 2008
- 4 relevés GSM+temps = 95% ré-identification
Source : Y.-A. de Montjoye et al., *Unique in the Crowd: The privacy bounds of human mobility*, Nature, 2013

L'anonymisation

- Données $D \subset Id \times Qid \times S$
- Mécanisme $\mathcal{M}(D)$ d'anonymisation produisant D'
- Une ou plusieurs fonctions f qui modifient les valeurs d'Id et Qid
- « Pseudonymisation » ou suppression des Id
- Perturbation des Qid

Les transformations I/2

- Suppression : $f(x) = \emptyset$
- Mélange : $f(x) = y$
- Vieillissement ou décalage : $f(x) = x + a$
- Masquage : $f(x) = .suff(x) \mid pre(x)$.
- Appauvrissement ou généralisation : $x \in f(x)$

Les transformations 2/2

- Micro-agrégation : $f(x) = \text{agg}([x])$
- Obfuscation ou substitution : $f(\emptyset|x) = y \{y \text{ fictif}\}$
- Chiffrement : $f(x) = k \{f^{-1}(k) = x\}$
- Hachage : $f(x) = k \{f^{-1}(k) = \text{n/a}\}$
- Randomisation : $f(x) = \text{rand}(x)$

Mise en oeuvre

Source : L'anonymisation de données en masse chez Bouygues Telecom, J.-L. Lambert et P. Chambet, JSSI 2011

Les grands principes (3/3)

Attribut	Exemple de transformation
AdresseDeclaree.boitePostale	Préfixe à 'BP' + renumérotation sur 7 digits
AdresseDeclaree.complementAdr1	Préfixe à 'APT' + renumérotation sur 7 digits
AdresseDeclaree.complementAdr2	Préfixe à 'APT' + renumérotation sur 7 digits
AdresseDeclaree.cp	Sans anonymisation
AdresseDeclaree.num	renumérotation
AdresseDeclaree.rue	Préfixe à 'RUE ANONYME' + renumérotation sur 7 digits
AdresseDeclaree.ville	Sans anonymisation
AdresseElectronique.eMail	Préfixe "email" + renumérotation sur 6 digits + suffixe "@bouyguestelecom.fr"
AdresseNormalisee.ligne1	Deduit des translations sur les champs métiers élémentaires
AdresseNormalisee.ligne2	Deduit des translations sur les champs métiers élémentaires
CarteBancaire.noCarte	renumérotation sur 16 digits. Pas de recherche de compatibilité bancaire
ClientPayeur.identifiantFonctionnelClient	Conservation du préfixe [123456NK]. + renumérotation respectueuse des règles métiers (8 digits pour le prefixe '1', 2 digits pour le prefixe '2', ...)
CoordonneesTelephoniques.noTel	Conservation du préfixe (0033 +33 33 0 aucun), des 5 digits suivants + renumérotation des 4 derniers digits. Inchangé pour les numéros courts.
Entreprise.noSiren	renumérotation sur 9 digits
Entreprise.raisonSociale	Préfixe à 'raisonSociale' + renumérotation sur 6 digits
Individu.civilité	Sans anonymisation
Individu.nom	Préfixe à 'NOM' + renumérotation sur 6 digits
Individu.prenom	Préfixe à 'PRE' + renumérotation sur 6 digits
JusticatifIdentite.numero	Préfixe à 'PID' + renumérotation sur 7 digits
LigneGsm.msisdn	Conservation du préfixe (0033 +33 33 0 aucun), des 5 digits suivants + renumérotation des 4 derniers digits. Inchangé pour les numéros courts.
SimLogique.imsi	Conservation des 11 premiers digits + renumérotation des 4 derniers digits
SimPhysique.iccid	Sur forme '893320[0-9]{13}', conservation des 15 premiers digits + renumérotation des 4 derniers digits
TerminalMobile.imei	Conservation des 10 premiers digits, renumérotation des 4 digits suivants, recalcul du digit de contrôle pour compatibilité avec l'algorithme de Luhn

k-Anonymat

Source : P. Samarati and L. Sweeney. *Generalizing data to provide anonymity when disclosing information.* PODS, p. 188, Seattle, WA, USA (1998)

- Partition des attributs : Id's|Qid's|Sensibles
- Définition : Soit $T(Id, Qid, S)$ un tableau de données ; $T'(Qid, S)$ est une version k -anonyme de T ssi tout enregistrement de T' est **indiscernable** sur Qid de $k-1$ autres enregistrements.
- Probabilité $1/k$ de révélation d'identité

Les modalités pratiques

Age	Sexe	Pays	Maladie
19	M	Canada	Grippe
18	M	USA	Tétanos
27	F	USA	Grippe
25	M	Brésil	Cancer

Age	Sexe	Pays	Maladie
10--20	*	America	Grippe
10--20	*	America	Tétanos
20--30	*	America	Grippe
20--30	*	America	Cancer

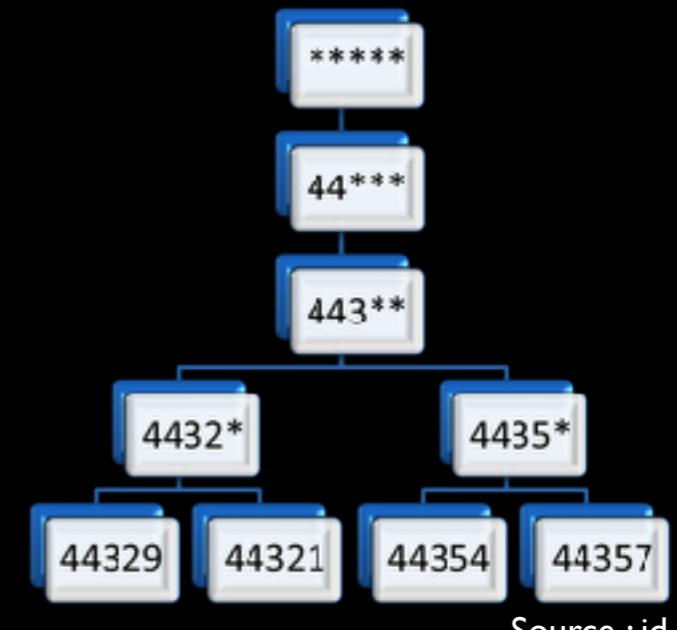
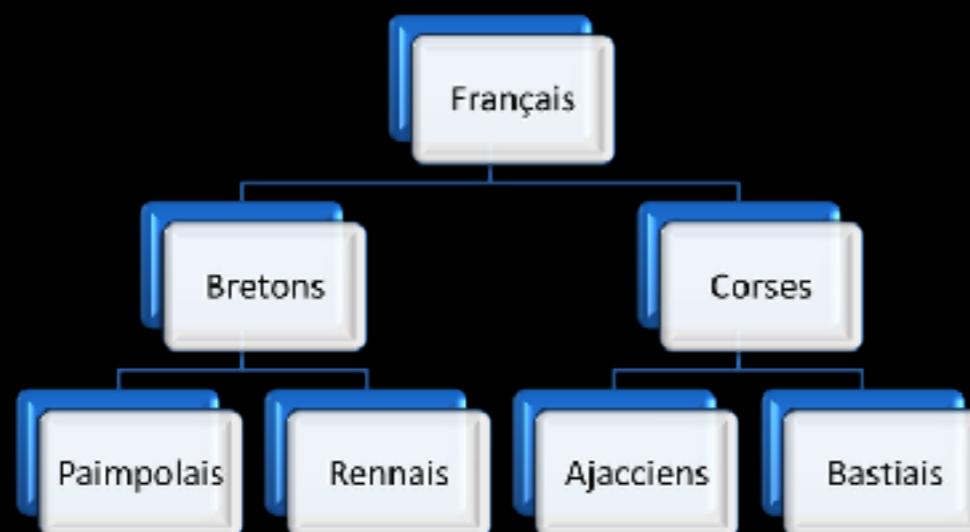
Age	Sexe	Pays	Maladie
10--20	M	N.America	Grippe
10--20	M	N.America	Tétanos
20--30	*	America	Grippe
20--30	*	America	Cancer

Source : S. Marjot et A. Piffeteau (2009). *k-Anonymat*. Rapport de Projet Polytech'Nante encadré par G. Raschia

- Tables 2-anonymes
- Suppression (*) vs. Généralisation+ (10-20)
- Encodage global (Table 2) vs. local+ (Table 3)
- ID vs. nD <A,S,P>+

Connaissances de domaine

- Emboîtement d'intervalles|valeurs par domaine
 - niveau de base : valeurs du domaine [0,100]
 - niveau 1 : [0,25]]25,50]]50,75]]75,100]
 - niveau 2 : [0,50]]50,100]
 - niveau 3 : [0,100]



Source : id.

Algorithme Datafly

- I. Calculer chaque #individus / valeur de Qid
2. Tant qu'il existe plus de k individus couverts par des Qid's de fréquence $< k$:
 - Généraliser le Qid avec la plus grande taille de domaine actif
3. Si nécessaire, supprimer (*) au moins k valeurs, dont les Qid's ont une fréquence $< k$

Exemple

Source : Lecture 14 Achieving Formal Protection, Bradley Malin, 2013

Race	Birthdate	Sex	Zip	Count	Records
Black	9/20/65	M	37203	1	r ₁
Black	2/14/65	M	37203	1	r ₂
Black	10/23/65	F	37215	1	r ₃
Black	8/24/65	F	37215	1	r ₄
Black	11/7/64	F	37215	1	r ₅
Black	12/1/64	F	37215	1	r ₆
White	10/23/64	M	37215	1	r ₇
White	3/15/64	F	37217	1	r ₈
White	8/13/64	M	37217	1	r ₉
White	5/5/64	M	37217	1	r ₁₀
White	2/13/67	M	37215	1	r ₁₁
White	3/21/67	M	37215	1	r ₁₂

Exemple

Source : Lecture 14 Achieving Formal Protection, Bradley Malin, 2013

Race	Birthdate	Sex	Zip	Count	Records
Black	1965	M	37203	2	r ₁ , r ₂
Black	1965	F	37215	2	r ₃ , r ₄
Black	1964	F	37215	2	r ₅ , r ₆
White	1964	M	37215	1	r ₇
White	1964	F	37217	1	r ₈
White	1964	M	37217	2	r ₉ , r ₁₀
White	1967	M	37215	2	r ₁₁ , r ₁₂
# of Values	2	3	2	3	2

Exemple

Source : Lecture 14 Achieving Formal Protection, Bradley Malin, 2013

Race	Birthdate	Sex	Zip	Count	Records
Black	1965	M	37203	2	r ₁ , r ₂
Black	1965	F	37215	2	r ₃ , r ₄
Black	1964	F	37215	2	r ₅ , r ₆
White	1964	M	37217	2	r ₉ , r ₁₀
White	1967	M	37215	2	r ₁₁ , r ₁₂
# of Values	2	3	2	2	# Suppressed

Exemple

Source : Lecture 14 Achieving Formal Protection, Bradley Malin, 2013

Record	Race	Birthdate	Sex	Zip
r ₁	Black	9/20/65	M	37203
r ₂	Black	2/14/65	M	37203
r ₃	Black	10/23/65	F	37215
r ₄	Black	8/24/65	F	37215
r ₅	Black	11/7/65	F	37215
r ₆	Black	12/1/64	F	37215
r ₇	White	10/23/64	M	37215
r ₈	White	3/15/64	F	37217
r ₉	White	8/13/64	M	37217
r ₁₀	White	5/5/64	M	37217
r ₁₁	White	2/13/67	M	37215
r ₁₂	White	3/21/67	M	37215

Race	Birthdate	Sex	Zip
Black	1965	M	37203
Black	1965	M	37203
Black	1965	F	37215
Black	1965	F	37215
Black	1964	F	37215
Black	1964	F	37215
*	*	*	*
*	*	*	*
White	1964	M	37217
White	1964	M	37217
White	1967	M	37215
White	1967	M	37215

Utilité des données

- Métrique de précision
 $| - \sum_{Qid's} \sum_n (\#gén./hauteur)/n \times \#Qid's$
- Précision nulle : généralisation systématique à la racine
- Précision égale 1 : valeurs originales

Précision

Source : Lecture 14 Achieving Formal Protection, Bradley Malin, 2013

Race	Birthdate	Sex	Zip	Record	Race	Birthdate	Sex	Zip	Race	Birthdate	Sex	Zip
Black	9/20/65	M	37203	r ₁	Black	1965	M	37203	0/1	2/5	0/1	0/3
Black	2/14/65	M	37203	r ₂	Black	1965	M	37203	0/1	2/5	0/1	0/3
Black	10/23/65	F	37215	r ₃	Black	1965	F	37215	0/1	2/5	0/1	0/3
Black	8/24/65	F	37215	r ₄	Black	1965	F	37215	0/1	2/5	0/1	0/3
Black	11/7/64	F	37215	r ₅	Black	1964	F	37215	0/1	2/5	0/1	0/3
Black	12/1/64	F	37215	r ₆	Black	1964	F	37215	0/1	2/5	0/1	0/3
White	10/23/64	M	37215	r ₇	*	*	*	*	1/1	5/5	1/1	3/3
White	3/15/64	F	37217	r ₈	*	*	*	*	1/1	5/5	1/1	3/3
White	8/13/64	M	37217	r ₉	White	1964	M	37217	0/1	2/5	0/1	0/3
White	5/5/64	M	37217	r ₁₀	White	1964	M	37217	0/1	2/5	0/1	0/3
White	2/13/67	M	37215	r ₁₁	White	1967	M	37215	0/1	2/5	0/1	0/3
White	3/21/67	M	37215	r ₁₂	White	1967	M	37215	0/1	2/5	0/1	0/3

Précision (bis)

Race	Birthdate	Sex	Zip	
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
1/1	5/5	1/1	3/3	4
1/1	5/5	1/1	3/3	4
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
2	30/5	2	2	12

$$I - I_2 / (I_2 \times 4) = 0,75$$

Précision (ter)

Datafly Solution

Race	Birthdate	Sex	Zip
Black	1965	M	37203
Black	1965	M	37203
Black	1965	F	37215
Black	1965	F	37215
Black	1964	F	37215
Black	1964	F	37215
*	*	*	*
*	*	*	*
White	1964	M	37217
White	1964	M	37217
White	1967	M	37215
White	1967	M	37215

More “Precise” Solution

Race	Birthdate	Sex	Zip
Black	1965	M	37203
Black	1965	M	37203
*	1965	F	3721*
*	1965	F	3721*
Black	1964	F	37215
Black	1964	F	37215
White	196*	M	37215
*	1965	F	3721*
White	1964	M	37217
White	1964	M	37217
White	196*	M	37215
White	196*	M	37215

Précision (quater)

Race	Birthdate	Sex	Zip	
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
1/1	2/5	0/1	1/3	26/15
1/1	2/5	0/1	1/3	26/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	4/5	0/1	0/3	12/15
1/1	2/5	0/1	1/3	26/15
0/1	2/5	0/1	0/3	6/15
0/1	2/5	0/1	0/3	6/15
0/1	4/5	0/1	0/3	12/15
0/1	4/5	0/1	0/3	12/15
3	6	0	1	10

$$1 - \frac{10}{12 * 4}$$

0.79

Version optimale ?

- Datafly calcule une approximation
- Procédure *minGen* de L. Sweeney, optimale :
 1. Enumérer toutes les généralisations de T ;
 2. Filtrer les généralisations non conformes à k ;
 3. Trier selon la précision et conserver la (les) meilleure(s).
- Efficacité vs. Utilité
- Elagage de l'espace de recherche possible

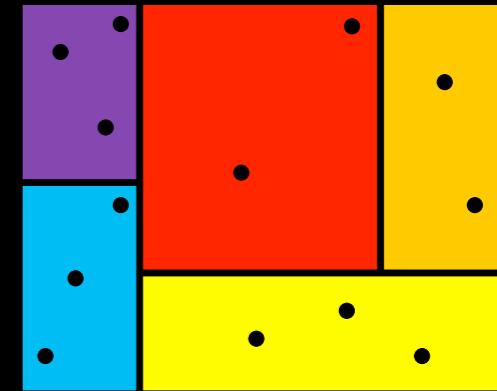
Casse-tête

- Compromis : **anonymat vs. utilité**
- Mesurer l'utilité du jeu de données
 - perte de précision, distortion, distance
- Calcul **NP-difficile** : solution optimale non garantie
- Méthodes : Datafly Incognito Mondrian
- Polémique : utilité négligeable !

Sources : J. Brickell, V. Shmatikov. *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*. KDD 2008.
Tiancheng Li and Ninghui Li. *On the Tradeoff Between Privacy and Utility in Data Publishing*. KDD 2009.

Mondrian

Source : LeFevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional k-Anonymity. ICDE 2006

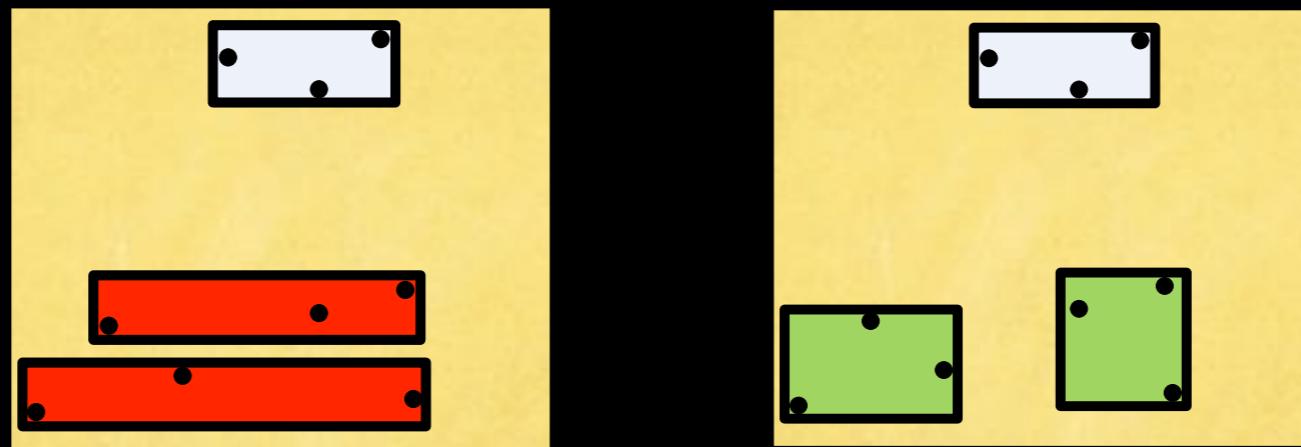


- Principe :
 - découpage successif selon chaque dimension
 - division par hyperplan médian (K-d-tree)
- Métrique de *discernabilité* : #points / classe
- Situation idéale : k individus / classe

Qualité des classes

Source : Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-chee Fu. *Utility-based anonymization using local recoding*. KDD 2006.

- Pénalité normalisée de certitude (NCP)
- Évaluation du périmètre de chaque classe
- Privilégie les classes homogènes



Les attaques

- par identification indirecte (ou jointure) +++croisement des Qid's
- par l'ordre des enregistrements ++arrangement identique

Problème n°1

	Attributs non sensibles			Attribut sensible
ID	Code postal	Age	Nationalité	pathologie
1	441**	< 30	*	infarctus
2	441**	< 30	*	infarctus
3	441**	< 30	*	infection virale
4	441**	< 30	*	infection virale
5	857**	≥ 40	*	cancer
6	857**	≥ 40	*	infarctus
7	857**	≥ 40	*	infection virale
8	857**	≥ 40	*	infection virale
9	441**	3*	*	cancer
10	441**	3*	*	cancer
11	441**	3*	*	cancer
12	441**	3*	*	cancer

	Attributs non sensibles			Attribut sensible
ID	Code postal	Age	Nationalité	pathologie
1	1305*	≤ 40	*	infarctus
4	1305*	≤ 40	*	infection virale
9	1305*	≤ 40	*	cancer
10	1305*	≤ 40	*	cancer
5	1485*	> 40	*	cancer
6	1485*	> 40	*	infarctus
7	1485*	> 40	*	infection virale
8	1485*	> 40	*	infection virale
2	1306*	≤ 40	*	infarctus
3	1306*	≤ 40	*	infection virale
11	1306*	≤ 40	*	cancer
12	1306*	≤ 40	*	cancer

Source : id.

- Individu dans la classe orange, table I
- Pathologie ?

Problème n°2

	Attributs non sensibles			Attribut sensible
ID	Code postal	Age	Nationalité	pathologie
1	441**	< 30	*	infarctus
2	441**	< 30	*	infarctus
3	441**	< 30	*	infection virale
4	441**	< 30	*	infection virale
5	857**	≥ 40	*	cancer
6	857**	≥ 40	*	infarctus
7	857**	≥ 40	*	infection virale
8	857**	≥ 40	*	infection virale
9	441**	3*	*	cancer
10	441**	3*	*	cancer
11	441**	3*	*	cancer
12	441**	3*	*	cancer

Source : id.

- Individu dans la classe bleue
- Connaissance : “population avec peu de problèmes cardiaques”
- Pathologie ?

Problème n°3

ID	Code postal	Age	Salaire	Pathologie
1	476**	2*	3k€	ulcère gastrique
2	476**	2*	4k€	gastrite
3	476**	2*	5k€	cancer de l'estomac
4	4790*	≥ 40	6k€	gastrite
5	4790*	≥ 40	11k€	grippe
6	4790*	≥ 40	8k€	bronchite
7	476**	3*	7k€	bronchite
8	476**	3*	9k€	pneumonie
9	476**	3*	10k€	cancer de l'estomac

Source : id.

- Individu dans la classe bleue
- Pathologie ? Salaire ?

Les attaques

- par identification indirecte (ou jointure) +++croisement des QIDs
- par l'ordre des enregistrements ++arrangement identique
- par homogénéité +valeurs sensibles identiques dans une classe
- par similitude -valeurs sensibles similaires dans une classe
- par dissymétrie --mauvaise répartition d'un attribut binaire
- par observabilité --existence d'un individu dans la table
- par connaissances antérieures ---informations générales

Quelques extensions

- Parades contre la révélation de caractère
 - a -désassociation : seuil de rareté ($>a\%$ v.)
 - p -sensibilité : au moins p valeurs distinctes
 - l -diversité : dispersion suffisante
 - m -unicité : valeurs uniques

Variations sur le thème

- t -proximité : distribution [c] = dist. table
- approche personnalisée : généralisation indiv.
- k -anonymat faible : volume de données
- d -présence : $[d1, d2]\%$ de révéler la présence
- r -ensemble : méthode par classification
- anonymat e-différentiel : risque pour l'rec.
- anonymat de distribution : dist. uniquement

Problème n°4

The diagram illustrates a data integration or comparison problem involving three tables. The middle table contains a single row for 'Bob' with attributes: Nom (Bob), Age (21), and Code postal (12000). Two red arrows point from this row to specific rows in the outer tables.

Left Table:

G.ID	Age	Code postal	Maladie
1	[21,22]	[12000,14000]	dyspepsie
1	[21,22]	[12000,14000]	bronchite
2	[23,24]	[18000,25000]	grippe
2	[23,24]	[18000,25000]	gastrite
3	[36,41]	[20000,27000]	grippe
3	[36,41]	[20000,27000]	gastrite
4	[37,43]	[26000,35000]	dyspepsie
4	[37,43]	[26000,35000]	grippe
4	[37,43]	[26000,35000]	gastrite
5	[52,56]	[33000,34000]	dyspepsie
5	[52,56]	[33000,34000]	gastrite

Right Table:

G.ID	Age	Code postal	Maladie
1	[21,23]	[12000,25000]	dyspepsie
1	[21,23]	[12000,25000]	gastrite
2	[25,43]	[21000,33000]	grippe
2	[25,43]	[21000,33000]	dysepsie
3	[41,46]	[20000,30000]	gastrite
3	[41,46]	[20000,30000]	grippe
4	[54,56]	[31000,34000]	gastrite
4	[54,56]	[31000,34000]	dysepsie
4	[54,56]	[31000,34000]	gastrite
5	[60,65]	[36000,44000]	gastrite
5	[60,65]	[36000,44000]	grippr

Source : id.

Les attaques

- par identification indirecte (ou jointure) +++croisement des QIDs
- par l'ordre des enregistrements ++arrangement identique
- par tables complémentaires ++2 versions publiques différentes
- par homogénéité +valeurs sensibles identiques dans une classe
- par similitude -valeurs sensibles similaires dans une classe
- par dissymétrie --mauvaise répartition d'un attribut binaire
- par observabilité --existence d'un individu dans la table
- par connaissances antérieures --informations générales
- par tables complémentaires --2 versions publiques différentes
- par chaînabilité ---suivi temporel par ID banalisé
- ...

Opinion 05/2014 du G29

- Techniques d'anonymisation

	Is Singling out still a risk?	Is Linkability still a risk?	Is Inference still a risk?
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

Table 6. Strengths and Weaknesses of the Techniques Considered

Bibliographie parti[a/e]le

Histoire

Sweeney, L. k-Anonymity:A Model for Protecting Privacy. IJFKS, 2002.

Sweeney, L. k-Anonymity:Achieving k-Anonymity Privacy Protection using Generalization and Suppression. IJFKS, 2002.

Algorithmes

LeFevre, K., DeWitt, D.J., Ramakrishnan, R. Incognito: Efficient Full-domain k-Anonymity. SIGMOD, 2005.

LeFevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional k-Anonymity. ICDE, 2006.

G. Aggarwal, T. Feder, and K. Kenthapadi. Achieving anonymity via clustering. PODS, 2006.

T. Iwuchukwu and J. Naughton. K-anonymization as spatial indexing :Toward scalable and incremental anonymization. VLDB, 2007.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A. Utility-Based Anonymization Using Local Recoding. SIGKDD, 2006.

Wong, R. C., Fu, A.W., Wang, K., Pei, J. Minimality Attack in Privacy Preserving Data Publishing. VLDB, 2007.

Études de complexité

Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. PODS, 2004.

M. Ercan Nergiz and Christopher W. Clifton. Thoughts on k-anonymization. PDM (ICDE Workshop), 2006.

Modèles d'appauvrissement

T. Marius Truta and V. Bindu. Privacy protection: p-sensitive k-anonymity property. PDM (ICDE Workshop), 2006.

X. Xiao and Y. Tao. Personalized privacy preservation. SIGMOD, 2006.

X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. SIGMOD, 2007.

M. Ercan Nergiz, M. Atzori, and C.W. Clifton. Hiding the presence of individuals from shared databases. SIGMOD, 2007.

Ashwin Machanavajjhala, Johannes Gehrke, and Daniel Kifer. l-diversity: Privacy beyond k-anonymity. ICDE, 2006.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. ICDE, 2007.

M. Atzori. Weak k-anonymity: A low-distortion model for protecting privacy. In ISC, 2006.

R. Chi-Wing Wong, J. Li, A. Wai-Chee Fu, and K. Wang. (α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. SIGKDD, 2006.

Xiao, X, Tao, Y. m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. SIGMOD, 2007.

Etat de l'art

B. C. M. Fung, K. Wang, R. Chen and P. S. Yu. Privacy-Preserving Data Publishing: A Survey on Recent Developments. ACM Computing Surveys (sous presse) 2010.

Institutions

- **CNIL** *Commission Nationale de l’Informatique et des Libertés*
- **AFCDP** *Association Française des Correspondants à la Protection des données à Caractère Personnel*
- **G29** *Groupe de travail “Article 29” de la directive européenne 95/46/CE : 27 CNILs*
- **EPIC** *Electronic Privacy Information Center*
- **NIST** *National Institute of Standards and Technology*, U.S. Dpt. of Commerce
- **EFF** *Electronic Frontier Foundation*

Merci