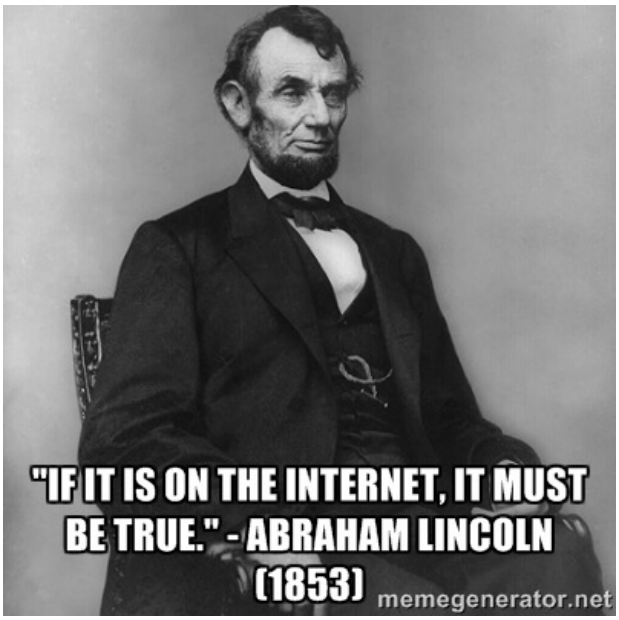
Responsabilité en analyse de données massives : équité, neutralité et transparence

Serge Abiteboul

Inria et ENS Cachan

Avec la participation de Julia Stoyanovich



Un déluge de données

Et même une explosion de données

Données personnelles

que nous produisons nous-mêmes

Que d'autres humains produisent sur nous

Produites par des capteurs divers

et par des programmes

Les données Web :

4V: Volume, véracité, vitesse, variété

Les individus et la société perdent le contrôle sur ces données

Promesses et risques des données massives

Améliorer la vie des gens : recommandations

Accélérer la découverte scientifique : médecine

Nourrir l'innovation : voitures autonomes

Transformer la société : gouvernement ouvert (open government)

Optimiser le business (publicités ciblées)

Un ressentiment croissant contre :

- Les comportements déviants : racisme, terrorisme, pédophilie, vol d'identité, cyber-harcèlement, cybercrime.
- Les entreprises : marketing agressif, personnalisation cryptique, décisions commerciales…
- Les gouvernements : NSA et ses analogues européens
- Une prise de conscience croissante de l'asymétrie entre ce que les systèmes connaissent de nous et ce que nous connaissons.

Motivation

- Beaucoup de problèmes sociaux sont liés à l'acquisition et au traitement de données
- Ce qu'on devrait faire
  - Changer la manière dont nous traitons les données personnelles ?
  - Changer le web ?

Références

Data responsibly, with Julia Stoyanovich (Drexel) & Gerome Miklau (U. Mass), **EDBT** Tutorial 2016

Data responsibly, with Julia Stoyanovich (Drexel), **Sigmod Blog** (in French, **Le Monde**), 2016

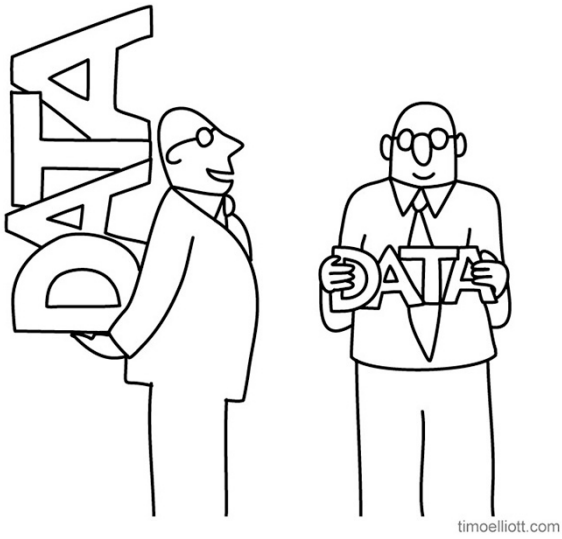
Managing your digital life with a Personal information management system, with Benjamin André (Cozy Cloud) & Daniel Kaplan (Fing), **CACM** 2015

Personal information management systems, with Amélie Marian (Rutgers), **EDBT** Tutorial 2015

Platform Neutrality, **CNNum** Report*, 2015*

Organisation

Motivation
Vie privée
Analyse de données
Evaluation de la qualité des données
Dissémination des données
Mémoire des données



"I think you'll find that mine is bigger..."

1. Vie privée et données
2. Les systèmes de gestion d'information personnelle, (PIMS – personal information management systems)

VIE PRIVÉE

Sécurité des données et vie personnelle

- De plus en plus de soucis avec la vie privée (privacy)
- Des limites sur ce que les compagnies de gestion de données peuvent faire
- Des lois forcent les compagnies à demander une autorisation pour construire des bases de données avec des informations personnelles (France)
- Des règles sur ce que les utilisateurs devraient pouvoir faire
- Des lois obligent les compagnies (banques, sociétés de crédit à laisser les usagers voir et corriger les informations qui les concernent (USA)
- Ces lois dépendent des pays et il est difficile de les faire respecter

Déconnecter ?

Confidentialité des données : est-ce qu'il y a quelque chose à faire ?

Existence de moyens pour garantir la confidentialité des données : inutilisés

- trop compliqués à utiliser ou à comprendre

Outils de cryptographie

Droits d'accès

Contrats d'utilisation illisibles

Difficulté à transférer les données d'un logiciel à l'autre : « enfermement propriétaire » (Vendor lock in)

Protection des données: les PIMS

Un système d'informations personnelles (PIM) est un système en nuage qui gère toute l'information d'une personne

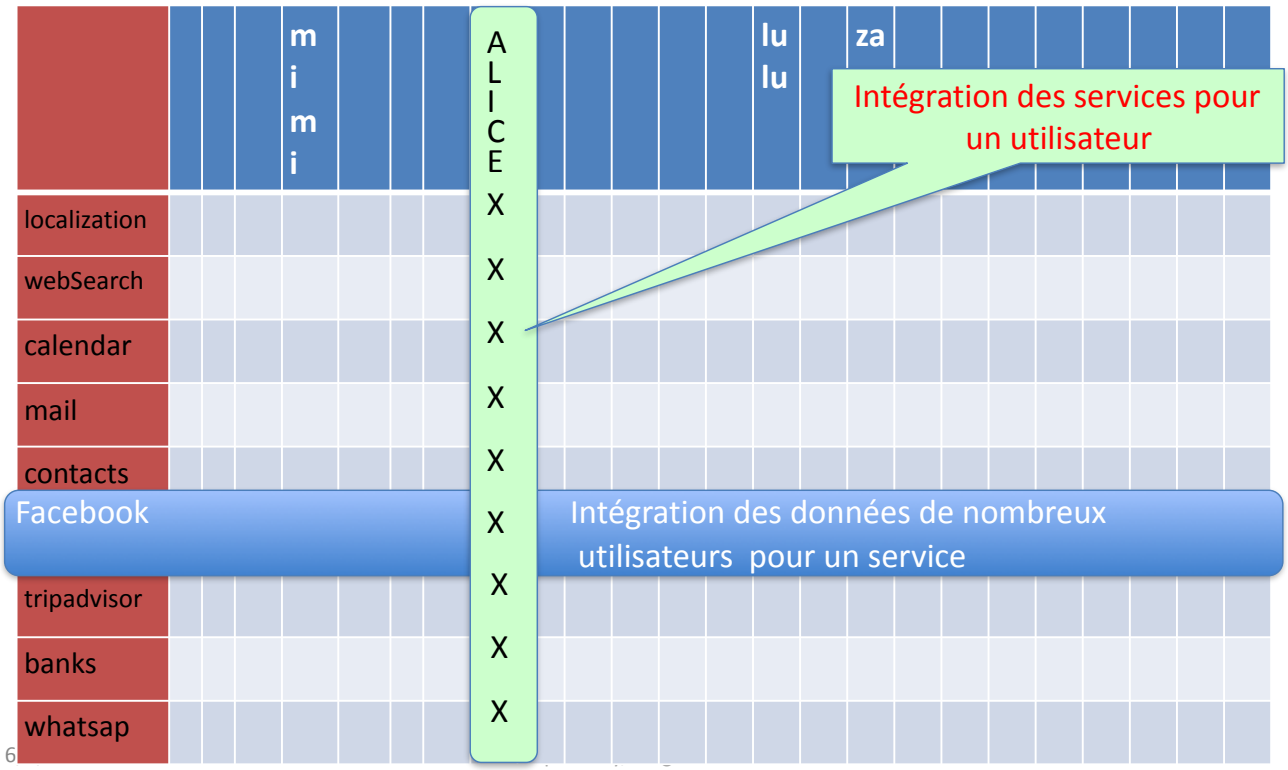
Un service du web s'exécute

- Sur une machine inconnue
- avec nos données
- traitées par un logiciel inconnu

Perspective

- **Une machine personnelle, connue**
- **Avec nos données**
  - répliquant des systèmes que nous apprécions
- Avec notre logiciel ou permettant l'utilisation de services externes

L'idée principale : l'intégration de données



Des problèmes de recherche et développement

Des problèmes anciens, revisités

Intégration des informations personnelles

Connaissance des contextes

Analyse des données personnelles

Synchronisation / sauvegarde et séquencement des tâches

Contrôle des accès et de l'échange d'informations

Contrôle des accès connectés

1. Equité
2. Transparence
3. Diversité
4. Vie privée (Privacy)

ANALYSE DE DONNÉES

Créer du savoir à partir des données

Trouver des corrélations statistiques

Publier des statistiques agrégées

Détecter

- Les points « aberrants »
- Les tendances

Techniques disponibles : fouille de données, données massives, apprentissage machine

Analyse de données : équité

Origine des biais

- Collecte des données
  - p.ex, données non représentatives
- Analyse des données
  - p. ex moteur de recherche qui favorise certains sites pour des raisons commerciales

Ce biais peut être illégal

- Faire des offres financières moins avantageuses aux membres de certaines minorités (« sterling »).

Exemple: analyse des données scientifiques

- Devrait expliquer comment les données ont été obtenues
- Quelles analyses ont été menées avec ces données
- Les expérimentations doivent être reproductibles

Domaine très exploré ; beaucoup de problèmes de recherche

Effets sur des sous populations

Paradoxe de Simpson
Une inégalité au niveau de la population disparaît ou s'inverse quand on considère des sous populations

		Admissions dans les écoles visées	
		Admis	**Refusés**
Genre	**F**	1512	2809
	M	3715	4727

résultats positifs

35% de femmes

44% d'hommes

UC Berkeley 1973: les femmes candidatent à des départements plus sélectifs, avec de faibles taux d'admission.

Equité de groupe ou individuelle

Au niveau du groupe : l'allocation « moyenne » aux individus ne dépend pas de la sous population

		Score	
		Bon	**Mauvais**
race	**noirs**	⊕	⊖ ⊖ ⊕ ⊖
	blancs	⊕ ⊖ ⊕	⊖ ⊖

résultats positifs **crédit obtenu**

40% de noirs

40% de blancs

Au niveau individuel
Deux personnes similaires à l'égard d'un facteur particulier devraient avoir des évaluations semblables

Analyse de donnée : diversité

Pertinence du classement (pour des recommandations)

- se base généralement sur la popularité
  - Les informations moins populaires deviennent de moins en moins populaires
  - un manque de diversité peut engendrer un risque de discrimination et d'exclusion

Exemples

- site de rencontres en ligne (match.com)
- marché de financement collaboratif comme Amazon Mechanical Turk
- ou une plate-forme de financement comme Kickstarter

Le riche s'enrichit alors que le pauvre s'appauvrit …

Analyse de donnée: Transparence

Exemple : manque de transparence dans le traitement des données par Facebook

- En général, contrat de licence d'utilisation illisible
- Les utilisateurs veulent contrôler ce qui est enregistré les concernant et comment ces informations sont utilisées

La transparence aide à vérifier que le service fonctionne comme il devrait le faire, comme c'est annoncé

Permet aussi au fournisseur de données de vérifier que ses données sont utilisées comme cela a été spécifié

Vie privée et analyse de données

Publication de statistiques :

protéger les personnes

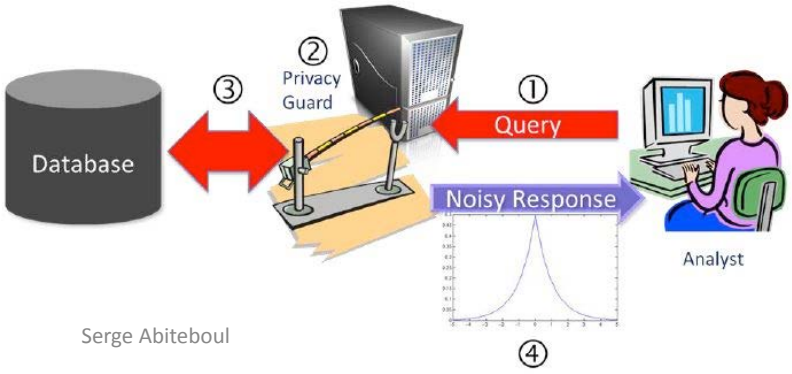
Anonymisation

« intimité différentielles »

-> Differential privacy

Déjà très étudié

Sujet non clos



Problèmes : vérifier ces propriétés

instruments pour collecter des données et les analyser de manière responsable

instruments pour vérifier qu'une analyse a été réalisée de manière responsable

plus facile si la responsabilité est prise en compte très tôt,

conception des instruments en rapport avec des utilisations responsables > *responsibility by design*

Pour vérifier le comportement d'un programme, on peut :

en analyser le code ≈ preuve par les théorèmes mathématiques

analyser ses effets ≈ études de phénomènes (tels le climat ou le coeur humain)

Vérification : analyse des effets

Analyse statistique

- Détecter les biais
- Détecter les utilisations illégales d'attributs protégés

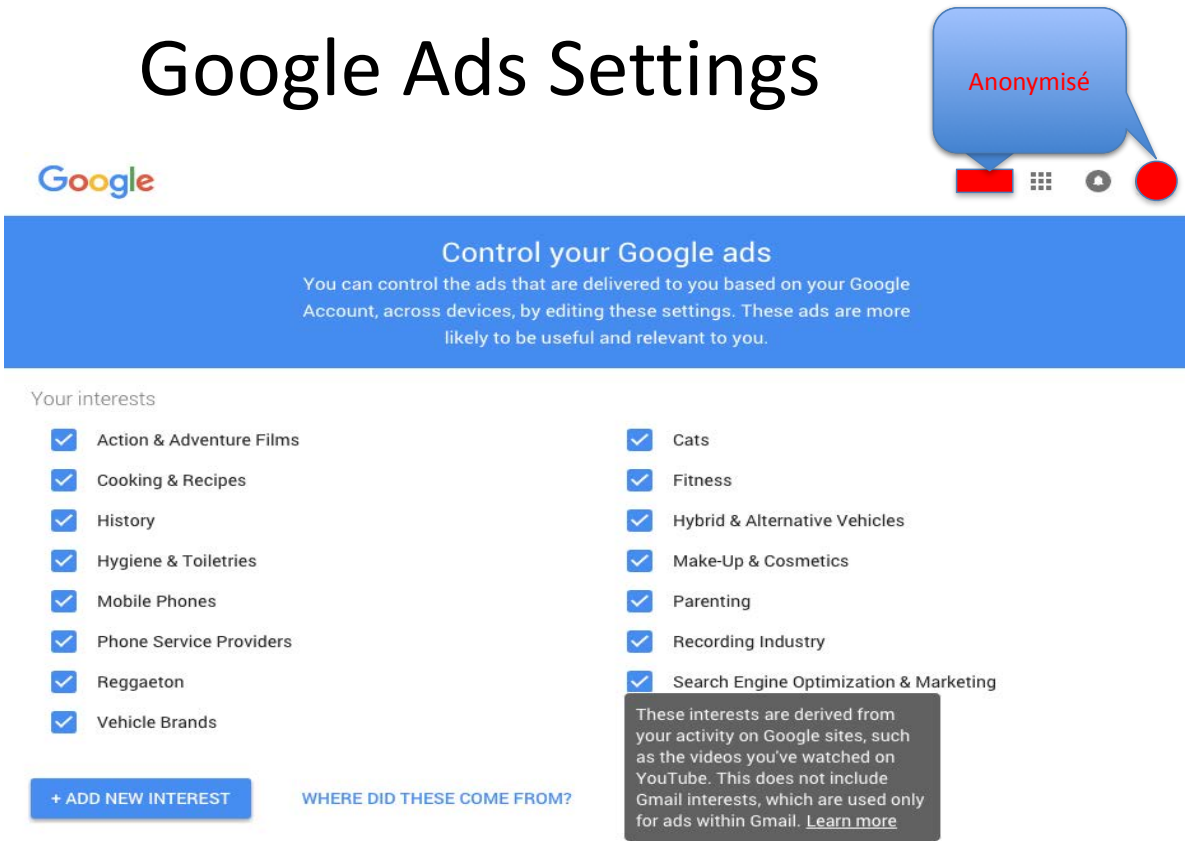
Vérifier la transparence

Vérifier la "loyauté"

- Le système se comporte comme il l'a déclaré

Exemple : Google Ads Settings & AdFisher

Google Ads Settings



Transparence et responsabilité

Analyse par AdFisher

Ne se comporte pas comme il est déclaré

Choix des publicités basés sur davantage de données, par exemple

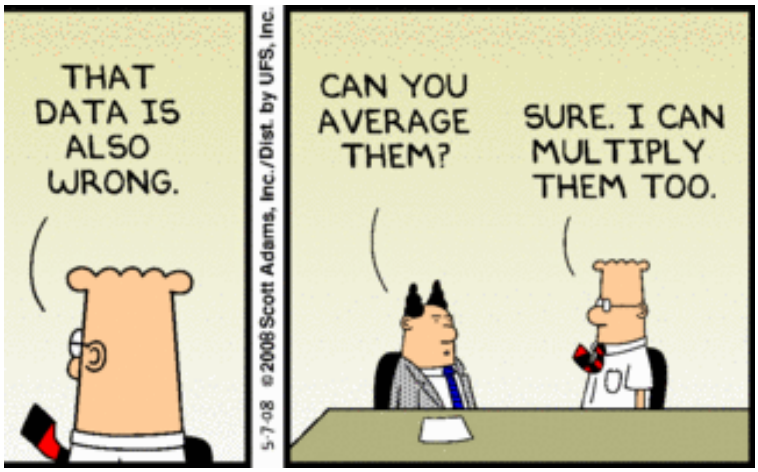
attribut protégés

les hommes reçoivent de manière significative davantage de publicités pour des postes à haut salaire que les femmes

Peu de contrôle sur les publicités

Enlever un intérêt diminue le nombre de publicités relatives à cet intérêt

par exemple, les chats



ÉVALUATION DE LA QUALITÉ DES DONNÉES

Ce qu'on ne veut pas voir sur le web

Sites Nazi

Sites terroristes

Contenu pédophile

Fausses informations sur la santé

Théorie du complot

Cybercrime

Harcèlement en ligne …

Problèmes : que peut-on faire ?

Détecter les contenus illégaux sur le Web

Évaluation automatique

- de la qualité des contenus
- de la légalité des contenus
- basé sur la transparence du classement

Analyse et classement collaboratif des pages web

Nombreux sujets de recherche

1. Protection des données
2. Accès ouverts aux données
3. Neutralité

DISSÉMINATION DES DONNÉES

Protection des données

Pour chacune de nos données en ligne, nous aimerions contrôler :

- Qui peut les lire
- Comment elles sont transmises
- Comment sont-elles modifiées
- Comment sont / seront-elles utilisées ?

Nous aimerions garder un peu de contrôle sur les paramètres de diffusion

- Contrôle de l'accès sur le web

Beaucoup de problèmes ouverts

Neutralité

Neutralité du net et des plate-formes (rapport CNNum)

Le réseau transporte des données sans biais par rapport aux sources, destinations, contenus …

Plateformes en ligne : discrimination en faveur de leurs services ?

Liens avec les problématiques de l'équité et de la diversité

Le riche s'enrichit alors que le pauvre s'appauvrit …

theguardian

European commission announces antitrust charges against Google

Inquiry will focus on accusations that internet search and tech multinational has unfairly used its products to oust competitors

Ruth Porat replaces Patrick Pichette as Google's chief finance officer. Photograph: Georges Gobet/AFP/Getty Images

Sam Thielman in New York
@samthielman
Wednesday 15 April 2015 07.27 EDT

The European Union accused Google on Wednesday of cheating competitors by distorting Internet search results in favour of its Google Shopping service and also launched an antitrust probe into its Android mobile operating system.

Problèmes

Tests de neutralité

Surveillance de la neutralité

1. Données personnelles
2. Archiver
3. Archives du web

MÉMOIRE DES DONNÉES

Archivage des données

Problèmes : décider

- Ce qu'il faut archiver
- Ce qu'il faut oublier
  - Oublier est un moyen de produire des abstractions
- Classer, résumer …

Par exemple, projet européen ForgetIT

Conclusion

De nombreux conflits politiques et sociaux sont aujourd'hui liés aux données

Les problèmes sont très clairement pas uniquement techniques

Il serait temps de changer la façon dont on utilise les données personnelles ? De changer le web ?

Des organismes y travaillent

- CNNum
- différents gouvernements (USA, UE…)

Par exemple, pour le web

- Internet Government Forum (UN)
- Global Internet Policy Observatory (UE?)
- W3C Technology Policy Internet Group

http://abiteboul.com
http://binaire.blog.lemonde.fr

informatiques mathématiques
Inria
ENS
CACHAN