Des données, à l'information, aux connaissances : Le Web de demain

Serge Abiteboul

INRIA & ENS Cachan

Conseil national du numérique & Académie des sciences

Je ne connais pas d'être vivant, de cellule, tissu, organe, individu et peut-être même espèce, dont on ne puisse pas dire qu'il stocke de l'information, qu'il traite de l'information, qu'il émet et qu'il reçoit de l'information.

Michel Serres

Introduction ←

Deux grands succès du 20e siècle

- Les systèmes relationnels
- Les moteurs de recherche de la Toile

Deux défis du 21e siècle

- Réseaux et connaissances collectives
- La Toile des connaissances

Conclusion

Gestion de données/information

Les systèmes informatiques servent à calculer

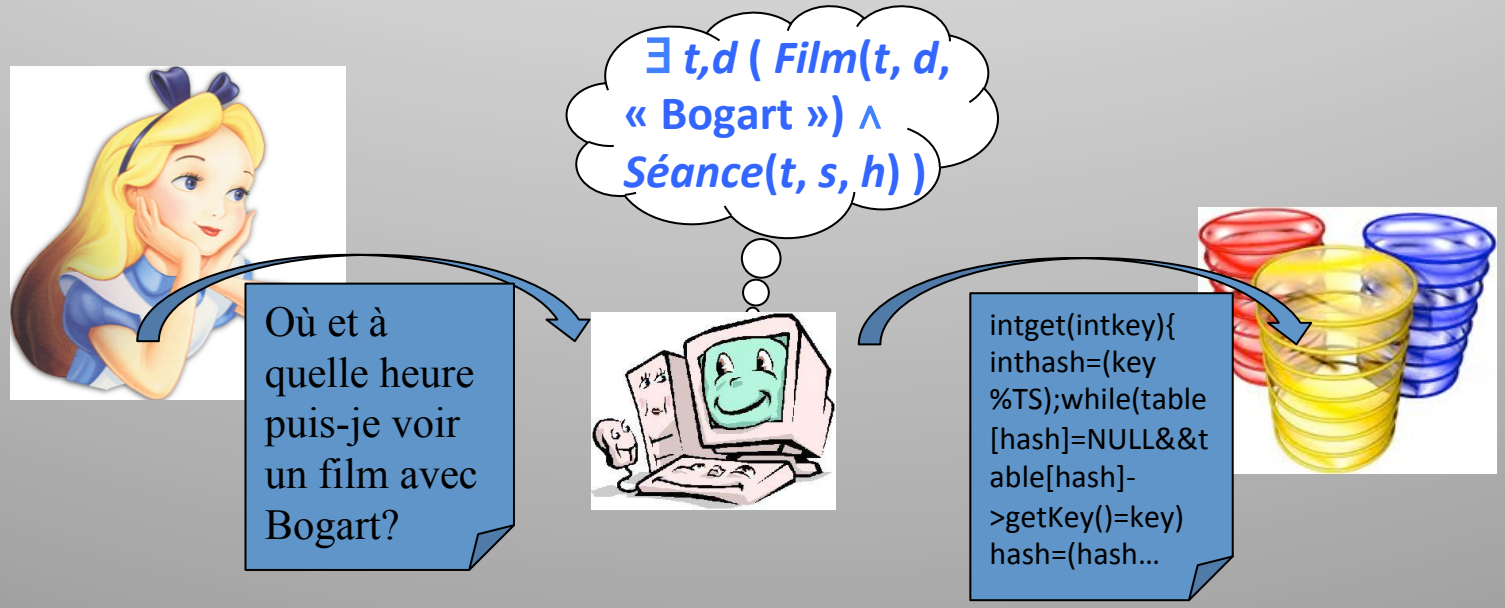
- Simulation de la météo
- Cryptographie
- Etc.

Ils servent beaucoup à stocker/gérer des **données**

- Comptabilité
- Catalogue de produits
- Inventaire
- Agenda
- Contacts
- Bibliothèque
- Médiathèque, etc.

Médiation

Les systèmes informatiques jouent le rôle de **médiateurs** entre des utilisateurs intelligents et des objets qui stockent l'information



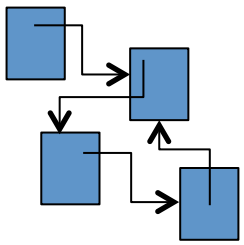
La Toile

Aujourd'hui, on trouve l'information sur la Toile

- « World Wide Web », littéralement la « toile d'araignée mondiale »

La Toile est un système hypertexte (*) public fonctionnant sur Internet (**) qui permet de consulter, avec un navigateur, des pages accessibles notamment via des moteurs de recherche

(*) Hypertext



(**) Internet

Un réseau qui permet de transférer des flux d'information entre des machines connectées au réseau

Success stories sur la Toile

Google : gestion des pages du Web

Facebook : informations personnelles et co

Wikipedia : encyclopédie

Amazon, eBay : catal

YouTube, P

Twitt ion, news

Flickr, sa : photos

iTunes, Kazaa, Emule, Batanga, BearShare : musique en ligne

Myspace : pages personnelles

Meetic : fiches individuelles

Wikileaks : secrets d'Etats

C'est de la gestion de données/d'information

Quel est leur point commun ?

Le quantitatif : le monde numérique

Des milliards d'objets communicants

Des centaines de millions de sites de la Toile

1000 milliards de pages (Septembre 2008)

Plus de 10 milliards de recherches sur le Web/mois (Avril

Nous baignons dans un monde numérique véritablement gigantesque

Le quantitatif : le volume de données

8 bits = **1** octet

1 téraoctet = **10^{12}** octets

- 200 téraoctets = tous les livres é

1 pétaoctet = **10^{15}**

- 100 [illegible] données produites par le collisionneur de [illegible] une minute

1 [illegible] = **10^{18}** octets

- 5 exaoctets = le volume des mots prononcés depuis que l'homme parle

1 zettaoctet = **10^{21}** octets

- **½ zetta = le trafic Internet en 2012** – $0.5.10^{21}$
- 66 zetta : l'information visuelle envoyée au cerveau en une année

Le monde numérique double tous les 18 mois

Source : Cisco Visual Networking Index – Forecast, 2007-201 - Via Michael Brodie

Le qualitatif : données, informations et connaissances

Données	Description élémentaire d'une réalité	*Mesures de températures dans une station météo*
Informations	Données avec un sens (pour construire une représentation de la réalité)	*Une courbe donnant l'évolution des minima & maxima moyens en un lieu suivant le mois de l'année*
Connaissances	Informations avec une vérité, plus généralement une loi qui est considérée comme vraie	*Le fait que la température sur terre augmente du fait de l'activité humaine*

Introduction

Deux grands succès du 20e siècle

- **Les systèmes relationnels** ←
- Les moteurs de recherche de la Toile

Deux défis du 21e siècle

- Réseaux et connaissances collectives
- La Toile des connaissances

Conclusion

La gestion de données « classique »

Un grand succès de l'informatique du 20e siècle

- Recherche industrielle et académique
- Fondements théoriques
- Systèmes commerciaux comme Oracle, DB2, SQL Server
- Logiciels libres comme mySQL

Modèle relationnel, Tedd Codd-1970

Fortement inspiré par la *Logique du premier ordre*

- Développée à la fin du 19e par des mathématiciens
- Pour formaliser le langage des mathématiques

Logic is the beginning of wisdom, not the end. Mr. Spock, *Star Trek*

Les données sont organisées en relations

Film		
Titre	Réalisateur	Acteur
Casablanca	M. Curtiz	H. Bogart
Casablanca	M. Curtiz	P. Lore
Les 400 coups	F. Truffaut	J.-P. Léaud
Star Wars	G. Lucas	H. Ford

Séance		
Titre	Salle	Heure
Casablanca	Le Grand Rex	19:00
Casablanca	Max Linder Panorama	20:00
Star Wars	Sèvres Espace Loisirs	20:30
Star Wars	Sèvres Espace Loisirs	20:45

Les requêtes sont exprimées en calcul relationnel

q_{HB} = { *salle, heure* | ∃ *réalisateur, titre*

(Film(*titre, réalisateur*, « Humphrey Bogart ») ∧

Séance(*titre, salle, heure*) }

En pratique les systèmes relationnels utilisent une syntaxe encore plus simple à comprendre :

SQL :

select *salle, heure*

from Film, Séance

where Film.*titre* = Séance.*titre* **and** *acteur*= «Humphrey Bogart»

Les raisons du succès de ces systèmes

Les requêtes sont exprimées dans **le calcul relationnel**

- un langage logique, simple et compréhensible surtout dans des variantes comme SQL

Une requête du calcul est traduite en une requête de **l'algèbre**

- facile à évaluer; Théorème de Codd

On peut **optimiser** l'évaluation d'expressions de l'algèbre

- parce que c'est un modèle de calcul limité (qui ne permet pas de calculer n'importe quelle fonction)

Le **parallélisme** permet de passer à l'échelle de très grandes bases de données

- les requêtes du calcul relationnel sont dans la classe de complexité AC0
- très parallélisables

Playboy : Is your company motto really "Don't be evil"?
Brin : Yes, it's real.
Playboy : Is it a written code?
Brin : Yes. We have other rules, too.
Page : We allow dogs, for example.

Sergey Brin et Larry Page, fondateurs de Google.
Interview dans le magasine *Playboy*, 2004

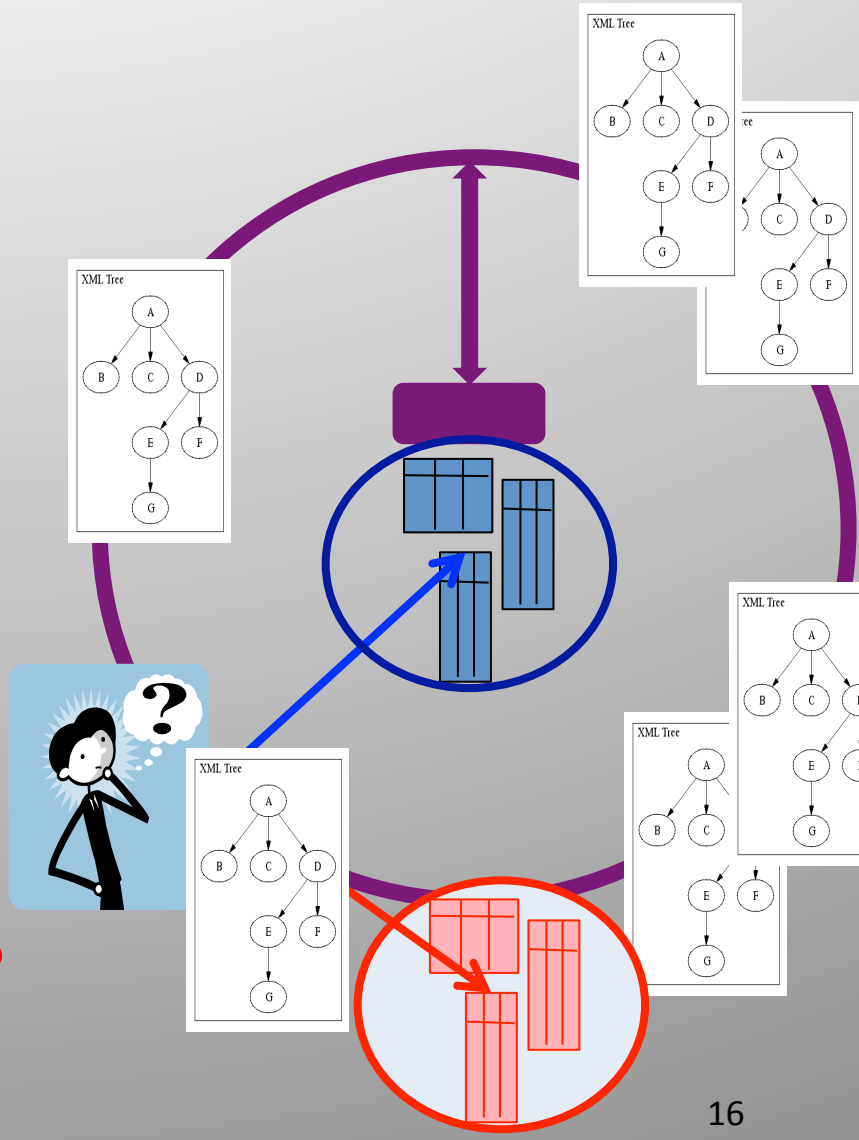
Ce qui a changé avec la Toile

L'information résidait sur des iles avec des formats, des langages de programmation, des applications, des systèmes d'exploitations différents

Grâce à des **standards universels** pour échanger de l'information, nous avons maintenant :

1. Un accès uniforme et universel à l'information
2. L'accès à des volumes gigantesques d'information

Comment trouver de l'information ?



Moteur de recherche de la Toile

L'index donne, pour chaque mot, la liste des pages qui contiennent ce mot

Mot	Numéro de page
…	
collège	34,56,223,9900,111111…
…	
france	56,778,6560,9900,9999…
…	
informatique	9890,11122290…
…	

num	url
1	www.inria.fr
2	www.bnf.com
3	www.inria.fr/~bhe
4	www.inria.fr/a/b
	…

Passage à l'échelle

Plus le moteur indexe de pages, plus l'index grandit

- Des milliards de pages
- L'index est du même ordre de grandeur que les pages indexées
- Chaque requête devient de plus en plus coûteuse à évaluer

Plus le moteur a d'utilisateurs, plus il reçoit de requêtes

- Des dizaines de milliards de requêtes de recherche par mois

Solution : le **parallélisme**

Digression: Le parallélisme

Essentiel pour gérer de gros volumes de données

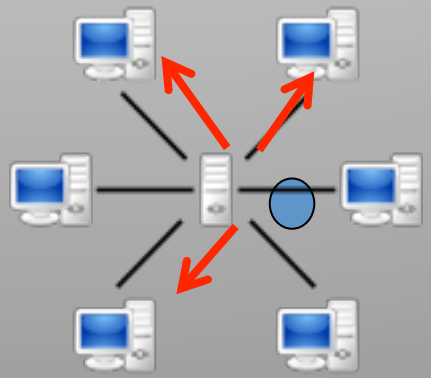
- Meilleure disponibilité, performance, etc.

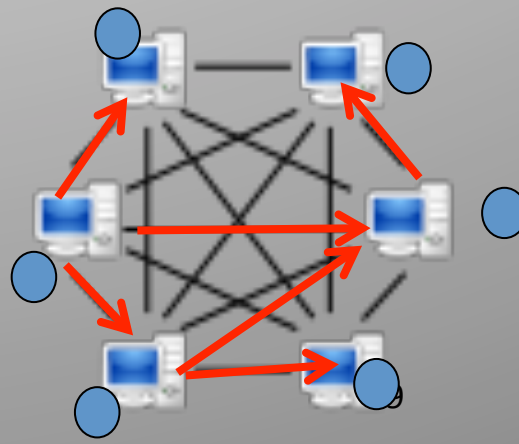
Quel parallélisme?

- Les machines sont de plus en plus multi-processeurs
- Collaboration entre les serveurs des différents sites d'une entreprise
- Centaines voire milliers de serveurs d'une « grappe »
- Millions de serveurs de la Toile

Illustration : deux types d'organisations sont possibles pour la diffusion de films

- Chaque film sur un serveur unique vite saturé
- Architecture *pair-à-pair*, chaque machine est à la fois serveur et client





Prouesse et magie

On vous a dit

- Les moteurs de recherche de la toile sont extraordinaires par la quantité d'informations qu'ils indexent – des milliards de pages

Non

- Ils sont merveilleux parce qu'ils savent comment choisir dans le résultat de l'index qui peut faire des centaines de millions de pages

La prouesse : indexer des milliards de pages

- En utilisant des techniques comme le hachage

La magie : trouver ce que vous voulez (en général)

- En utilisant des « mesures » pour classer les pages comme PageRank et TFIDF

Digression

La **neutralité du réseau** garantit l'égalité de traitement de tous les flux de données sur Internet. Ce principe exclut ainsi toute discrimination à l'égard de la source, de la destination ou du contenu de l'information transmise sur le réseau

- Si un opérateur d'internet sur mobile bloque les services de Skype – c'est pas neutre
- Si un opérateur de télécom et télévision bloque YouTube – c'est pas neutre
- Si un moteur de recherche déclasse un site pour plaire à un de ses clients – c'est pas neutre

La **neutralité des plateformes**

Les systèmes relationnels
comment on en est arrivé là

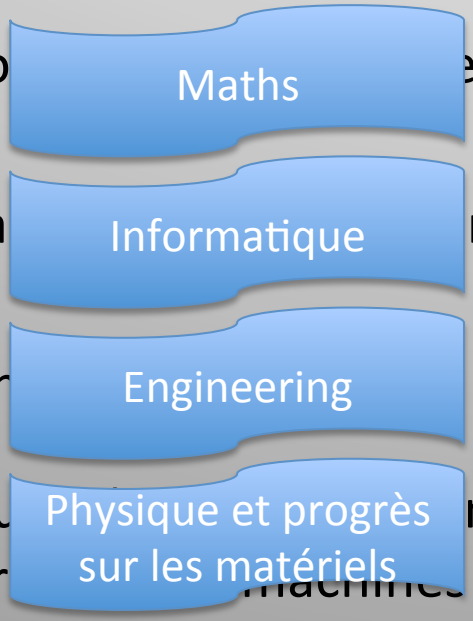
L'amélioration d'une fonction existante ou une nouvelle fonctionnalité

Des o…es

Des a…nts

Un er…

Et tou…mpte des progr…



Notamment, des modèles plus abstraits pour gérer des données

Notamment, la logique et l'algèbre relationnelles

Notamment, pour l'optimisation de requêtes

Notamment la reprise sur pannes et la gestion de la concurrence

Amélioration de la capacité des disques

Moteurs de recherche de la Toile

comment on en est arrivé là

L'amélioration d'une fonction existante ou une nouvelle fonctionnalité

Meilleur classement des pages

Des o…es

Maths

Notamment, les techniques de point fixe

Des a…nts

Informatique

Notamment, l'utilisation du parallélisme massif

Un er…

Engineering

Notamment, faire fonctionner des fermes de machines

Et tou…mpte des progr…machines

Physique et progrès sur les matériels

Baisse du prix des mémoires

21e siècle

- Masses de données disponibles
- Masses d'information disponible

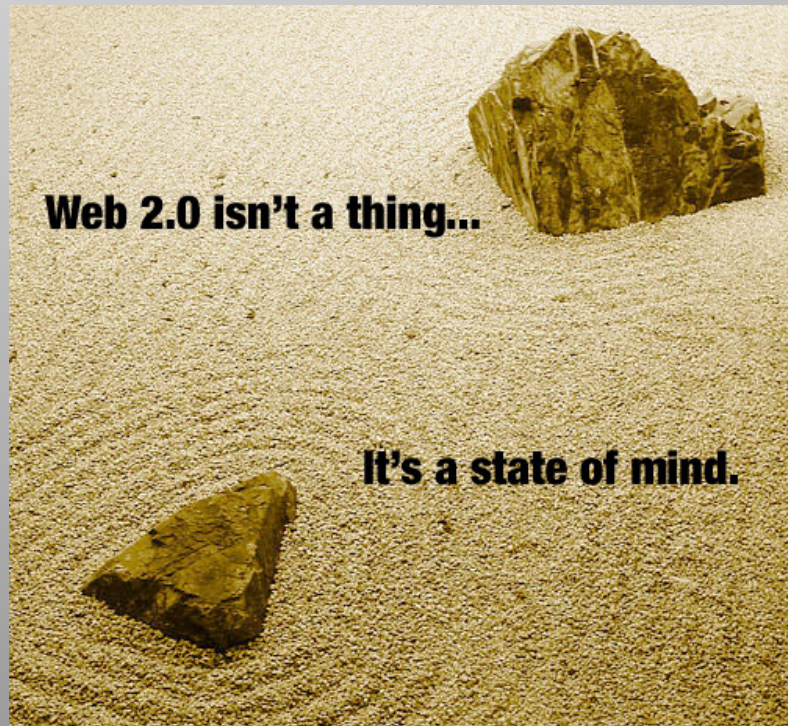
- Construire des bases de connaissances collectives

Après les réseaux de machines, puis de contenus, les **réseaux d'utilisateurs**

La Toile n'est pas juste faite pour obtenir des données

Tout le monde peut participer : tweets, Wikipédia, mashups

Mots-clés : interaction, communauté, communication, réseaux



Comment découvrir collectivement des connaissances

- La notation par l'internaute
  - PageRank
  - eBay
- L'évaluation de l'expertise des internautes
- La recommandation pour l'internaute
- La collaboration des internautes
- Le crowdsourcing des internautes

- L'analyse de données et les big data

La notation

Connaître l'avis de l'internaute

- quantitatif (notes)
- qualitatif (restaurant d'ambiance)

eBay : les clients notent les vendeurs

De plus en plus répandu

- Cinéma comme Allociné
- Restaurant comme ViaMichelin
- Pages de la Toile : annotations dans Delicious

L'évaluation de l'expertise

Evaluer

la qualité de l'information

la qualité des sources d'information

Illustration : travail récent sur la corroboration

Comment se construit l'expertise sur la Toile ?

- Des blogs, comme celui de Maitre Eolas pour les affaires juridiques
- Blogs de simples citoyens en Tunisie ou en Syrie

Elle sera un jour déterminée par des programmes ?

La recommandation

Utiliser les données du Web pour « recommander »

- Meetic organise des rencontres
- Netflix suggère des films
- Amazon des livres

Analyses statistiques pour mettre en évidence des « proximités »

- Entre clients dans Meetic
- Entre clients et produits dans Netflix et Amazon

amazon

La collaboration

Des internautes réalisent collectivement une tâche qui les dépasse individuellement

Wikipédia : encyclopédie

- 281 éditions ; 3 millions d'articles pour la version anglaise
- Place considérable dans la diffusion des connaissances
- Couverture bien plus large qu'une encyclopédie traditionnelle
- Qualité très controversée

Linux : operating system en logiciel libre

Web des données (linked data) : corpus de données ouvertes

Le crowdsourcing

Publication de questions ☛ réponses des internautes

Mechanical Turk d'Amazon

- Référence au *Turc mécanique*, un automate joueur d'échecs de la fin du 18e siècle

Foldit : décodage de la structure d'une enzyme proche de celle du virus du sida

- Comprendre comment cette enzyme se replie dans un espace en trois dimensions pour construire sa structure
- Jeu

L'analyse de données et les big data

- Croiser
  - Des données structurées/propres d'une entreprise
  - Avec des informations moins structurées/plus sales
    - Des données personnelles (comme des emails)
    - Des données de réseaux sociaux
    - Et des flux de données (générées par ex. par des senseurs)…
- Valoriser ces données
- Découvrir de nouvelles connaissances
- Offrir de nouveaux services

Analyse statistique de gros volumes de données

Technologue sous-jacente : machine learning

Très efficace

Bases scientifiques peu claires

Manque d'explication

Manque de sérendipité

Problèmes de « privacy »

Ref : article dans *Le Monde* du mardi 20 octobre 2015 avec Julia Stoyanovich

Mais de l'arbre de la connaissance du bien et du mal, tu n'en mangeras pas; car, au jour que tu en mangeras, tu mourras certainement.

Genèse 2:17

- Introduction
- Deux grands succès du 20e siècle
  - Les systèmes relationnels
  - Les moteurs de recherche de la Toile
- Deux défis du 21e siècle
  - Réseaux et connaissances collectives
  - **La Toile des connaissances** ←
- Conclusion

Du texte aux connaissances

La Toile des documents est basée sur le fait que les gens aiment écrire, lire, dire, écouter du texte

Les machines comprennent mieux des **connaissances** plus formatées

Texte	Connaissance
Je suis presque certain que Bob est amoureux d'Alice	Aime(Bob, Alice, 95%)

Le Web sémantique

Ajouter des indications sémantiques pour expliquer le sens des documents de la Toile

Sur cette présentation

auteur = Serge Abiteboul ; titre = Des données, à l'information…

nature = Forums régionaux du savoir 2013 ; type = « Powerpoint ;

date = Avril 2013 ; lieu = Rouen; langue = français

A l'intérieur d'un document

Woody Allen <*dbpedia:Woody_Allen*> était à Cannes <*geo:ville_France*> pour la première de …

Les bases de connaissances comme dbpedia sont appelées des **ontologies**

Ontologies

Des phrases logiques comme :

- ***classes*** *sa:Personne, sa:Réalisateur, sa:Cinéaste*
- *sa:Réalisateur* ***sous classe de*** *sa:Personne*
- *sa:Réalisateur* ***synonyme de*** *sa:Cinéaste*
- *sa:Woody_Allen* ***est un*** *sa:Réalisateur*
- ***relation*** *sa:a_réalisé*
- *sa:Woody-_Allen sa:a_réalisé sa:movie_Manhattan*

A quoi ça sert ?

- **Répondre** plus finement aux requêtes
- Permettre d' « **intégrer** » plusieurs sources d'information et, à terme, intégrer toutes les connaissances de la Toile

Problème : l'acquisition de connaissances

Les internautes

- aiment publier sur la Toile dans leur langue naturelle
- n'apprécient pas les contraintes d'un éditeur de connaissances
- veulent garder leur visibilité

Les connaissances vont être générées automatiquement

- Recherche de formes syntaxiques comme

Napoléon *est mort à* Sainte-Hélène

Construction de grosses bases de connaissances

Tâche complexe

- Compréhension de la langue
- La Toile fourmille d'imprécisions, d'erreurs, de faits controversés

Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information? T.S. Eliot

- Introduction
- Deux grands succès du 20e siècle
  - Les systèmes relationnels
  - Les moteurs de recherche de la Toile
- Deux défis du 21e siècle
  - Réseaux et connaissances collectives
  - La Toile des connaissances
- **Conclusion** ←

Des données,

à l'information,

aux connaissances…

La Toile est multiforme

Industrie, santé, culture, gouvernement, sciences, écologie…

Incontournable

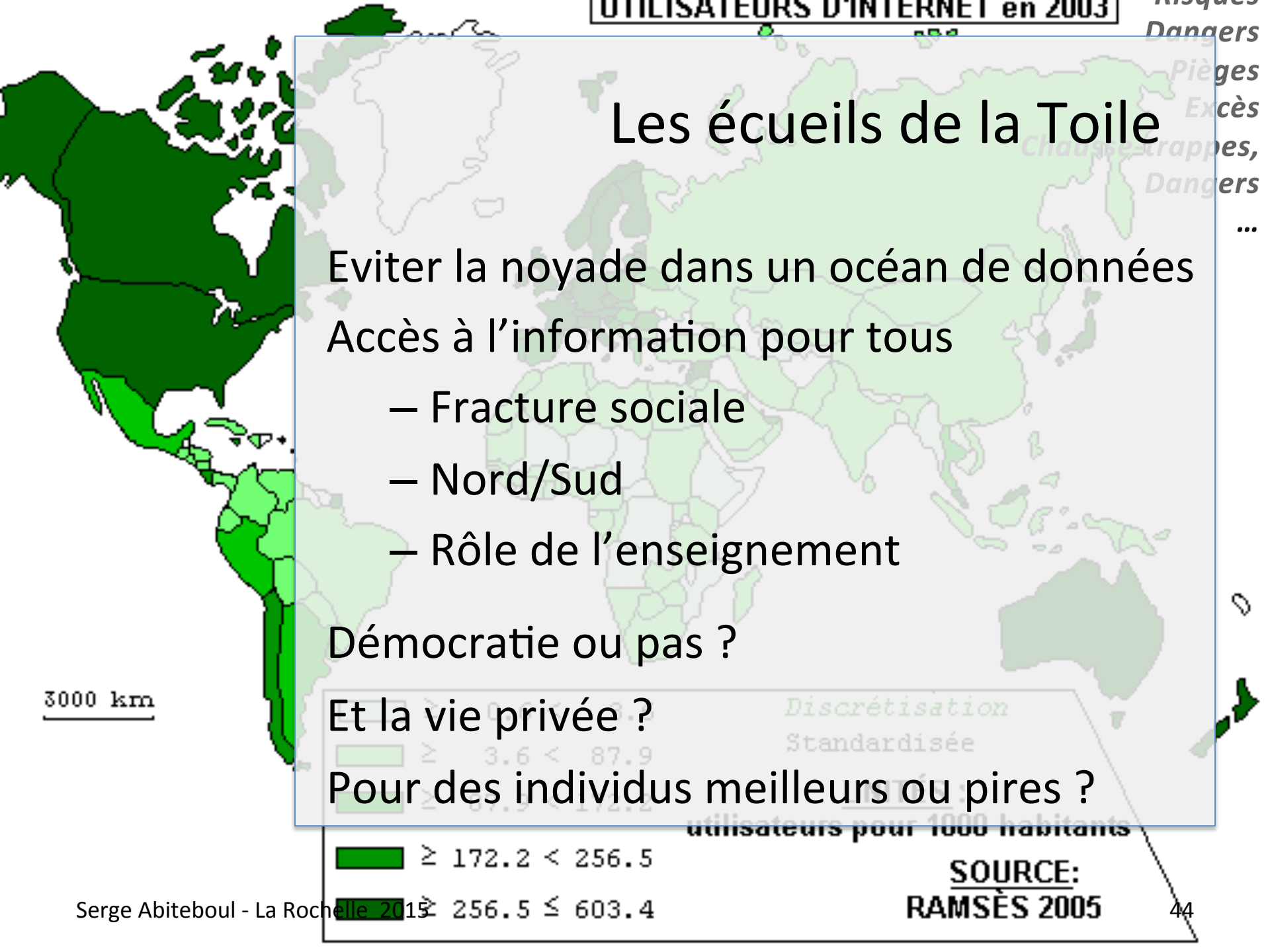
- Trouver du travail, travailler, se loger, gérer ses comptes bancaires, faire partie d'une association…

L'hébergeur de toutes les connaissances de l'humanité ?

- Des plus horribles fantasmes, de toutes les violences
- De toutes les imprécisions, les erreurs
- Un fantastique gisement de connaissances

La Toile est multiforme

1. Hypertexte
2. Bibliothèque universelle de documents
3. Les réseaux sociaux
4. Toile des connaissances
5. Téléphones « intelligents »
6. Objets communicants et intelligence ambiante
7. Mondes virtuels (jeux 3D)
8. Télé en OTT
9. …



Les écueils de la Toile

Eviter la noyade dans un océan de données

Accès à l'information pour tous

- Fracture sociale
- Nord/Sud
- Rôle de l'enseignement

Démocratie ou pas ?

Et la vie privée ?

Pour des individus meilleurs ou pires ?


LOUQUE! EN UN CLIC J'AI PÉCHO LES ZEUVRES COMPLÉTES DE VICTOR HUGO
TROP PUISSANT!
C CON QU'ON C PAS LIRE!
DELIGNE

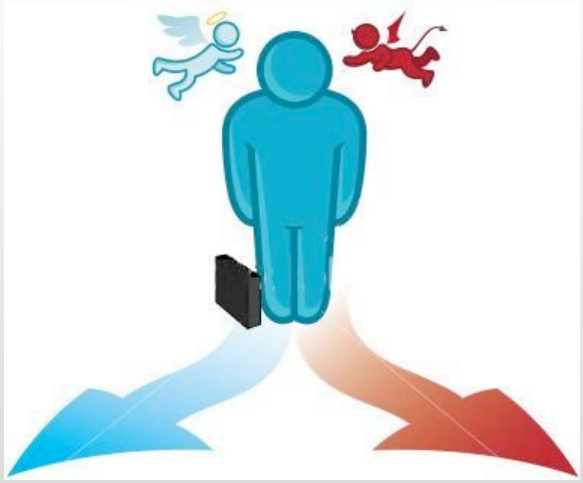


Illustration:
Big data &
La santé

Les soins personnalisés

- Toutes les données médicales de la personne
  - Son génome
- Toutes ses données sociales
- Soins personnalisés
- Mesures prédictives

Les polices personnalisées

- Plus chères pour les personnes à risque
- Personnes « trop » à risque non assurées
- Mutualisation des risques de plus en plus limitée

C'est la même science qui rend ça possible

Prediction is very difficult, especially about the future. Niels Bohr

Et demain… Déjà aujourd'hui

Nous vivons dans un monde numérique, entourés de systèmes qui traiteront l'information pour nous :

- Analysant cette information ; extrayant des connaissances ; échangeant des connaissances ; inférant collectivement des connaissances

Nous sommes passés d'un monde fermé et précis… à un monde ouvert et imprécis, parfois incohérent dans lequel les machines et les algorithmes tiendront un rôle considérable

Comment se préparer à cela ?

Par l'enseignement de l'informatique

- Pour comprendre le monde dans lequel nous vivons
- Pour choisir notre vie dans ce monde
- Pour contribuer à ce monde numérique

Rapport de l'Académie
des sciences 2013
Rapport du CNNUM
sur l'inclusion numérique 2014

Informatique et numérique : une confusion

- L'informatique a conduit à des transformations majeures de la société qui est devenue « numérique »
- Il faut préparer nos élèves à ce monde numérique
  - C'est la tâche de *tous* les enseignants
  - Il faut les former pour cela
- Il faut asseoir ces connaissances sur la science et la technique informatiques
  - C'est le rôle d'enseignants compétents en informatique
  - Il faut former des profs pour qu'ils puissent enseigner l'informatique
  - Il faut embaucher des profs d'informatique


Merci !