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

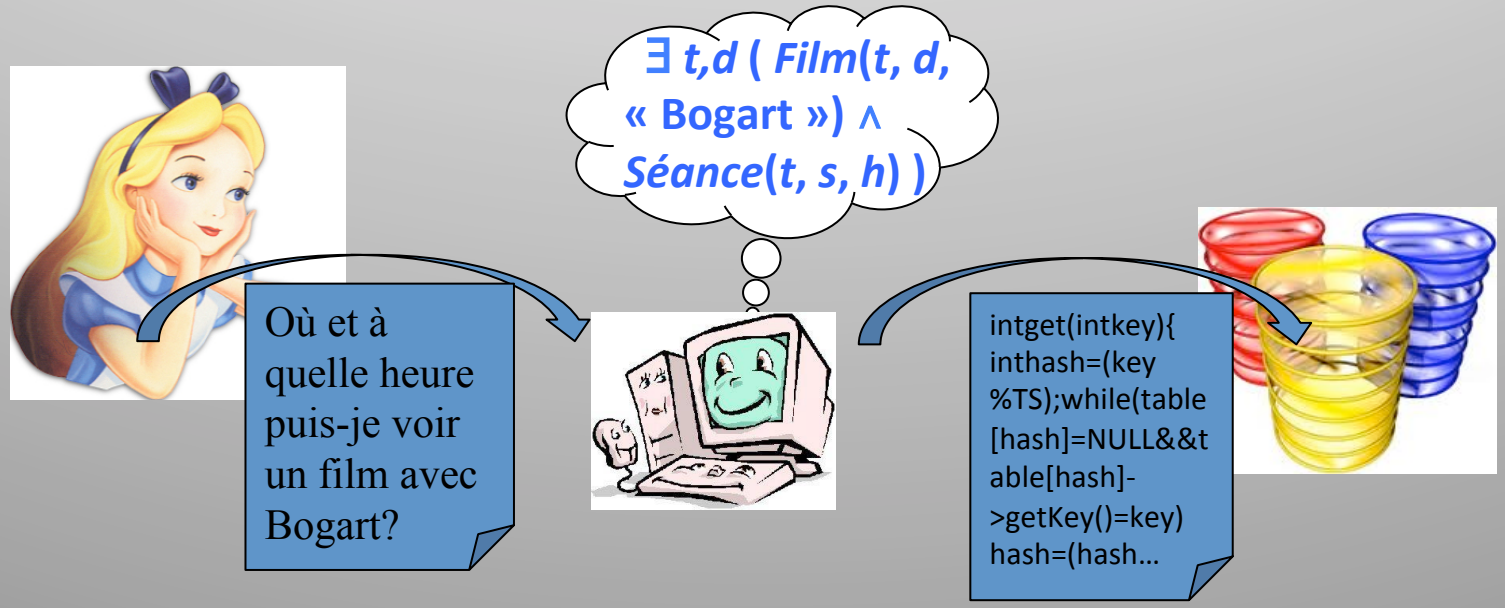
- Simulation de la météo
- Cryptographie
- Etc.

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



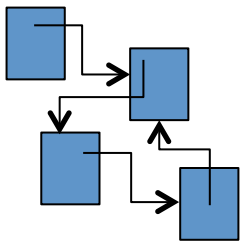
La Toile

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 communautés
Wikipedia : encyclopédie
Amazon, eBay : catalogues de vente sur le Web
YouTube, Dailymotion : vidéos
Twitter : communication, news
Flickr, Picasa : photos
iTunes, Kazaa, Emule, Batanga,
Myspace : pages personnelles
Meetic : fiches indiv
Wikileaks :

Quel est leur point commun ?

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

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 2008)

Nous baignons dans un monde numérique véritablement gigantesque

Le quantitatif : le volume de données

Le monde numérique double tous les 18 mois

8 bits

1 téraoctet = 10^{12} octets

- 200 téraoctets = tous les livres écrits à ce jour

1 pétaoctet = 10^{15} octets

- 100 pétaoctet = la quantité de données produites par le collisionneur de particules du CERN en une minute

1 exaoctet = 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

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 minimas & maximas 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*

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

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

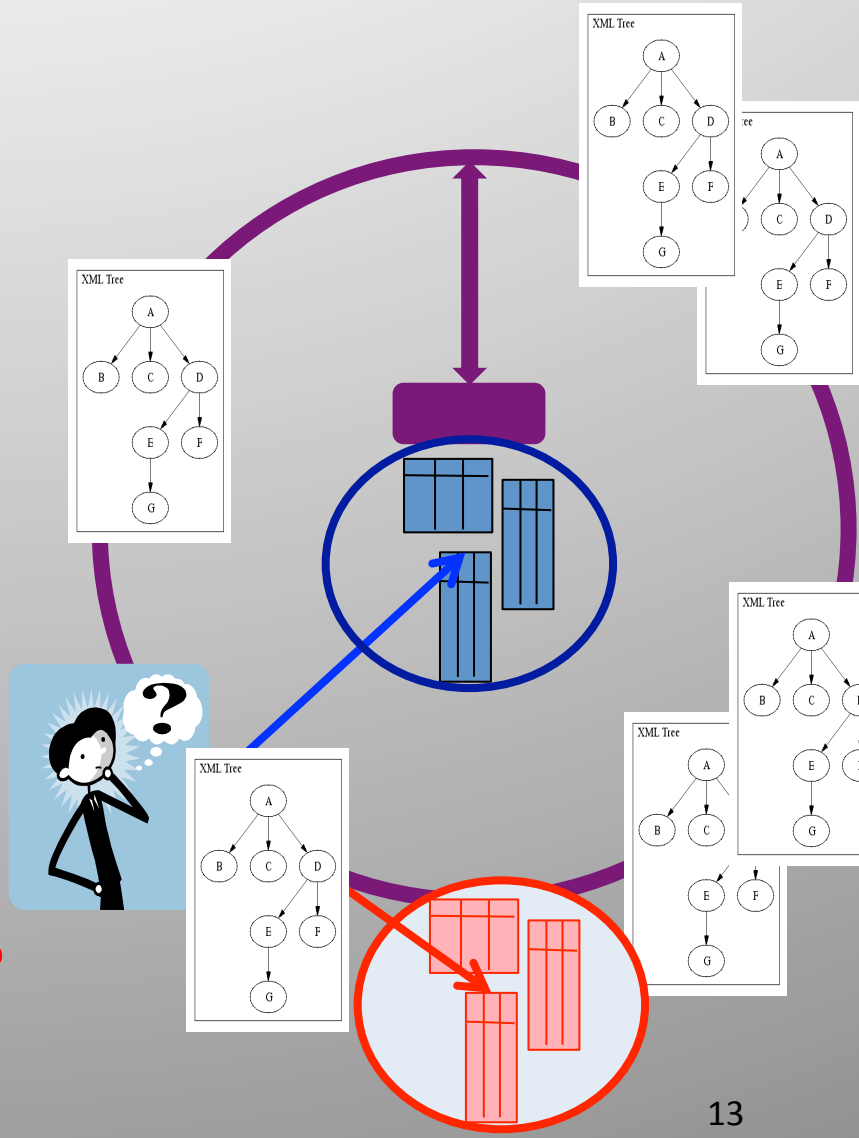
Ce qui a changé avec la Toile

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
	…

Moteur de recherche du Web

Problème de 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

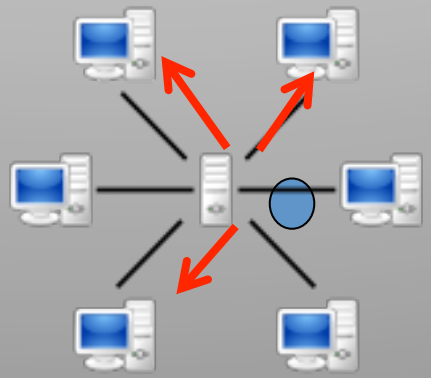
- Meilleure disponibilité, performance, etc.

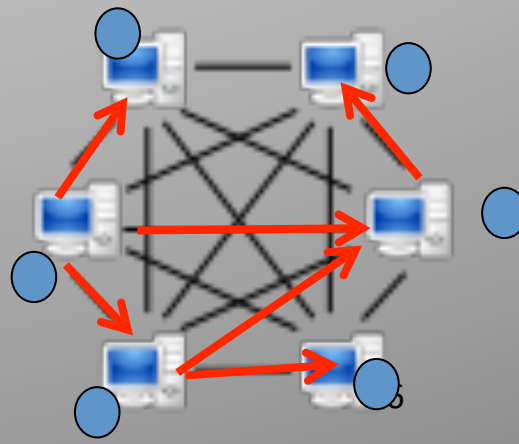
Quel parallélisme?

- 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

La magie : les classer avec PageRank

Surfeur aléatoire de la Toile

- **Popularité = probabilité de se trouver sur la page**
- Probabilité est plus forte pour lemonde.fr que pour la page personnelle de Monsieur Michu

Mise en équation : $pop = \Theta \times pop$

Et comment on calcule cela ?

- pop_0 défini par $pop_0[i] = 1/N$
  - toutes les pages sont supposées aussi populaires
- $pop_1 = \Theta \times pop_0$
- $pop_2 = \Theta \times pop_1$
- $pop_3 = \Theta \times pop_2 \ldots$

Le point fixe donne la popularité

Des problèmes ouverts

Simplisme des requêtes actuelles

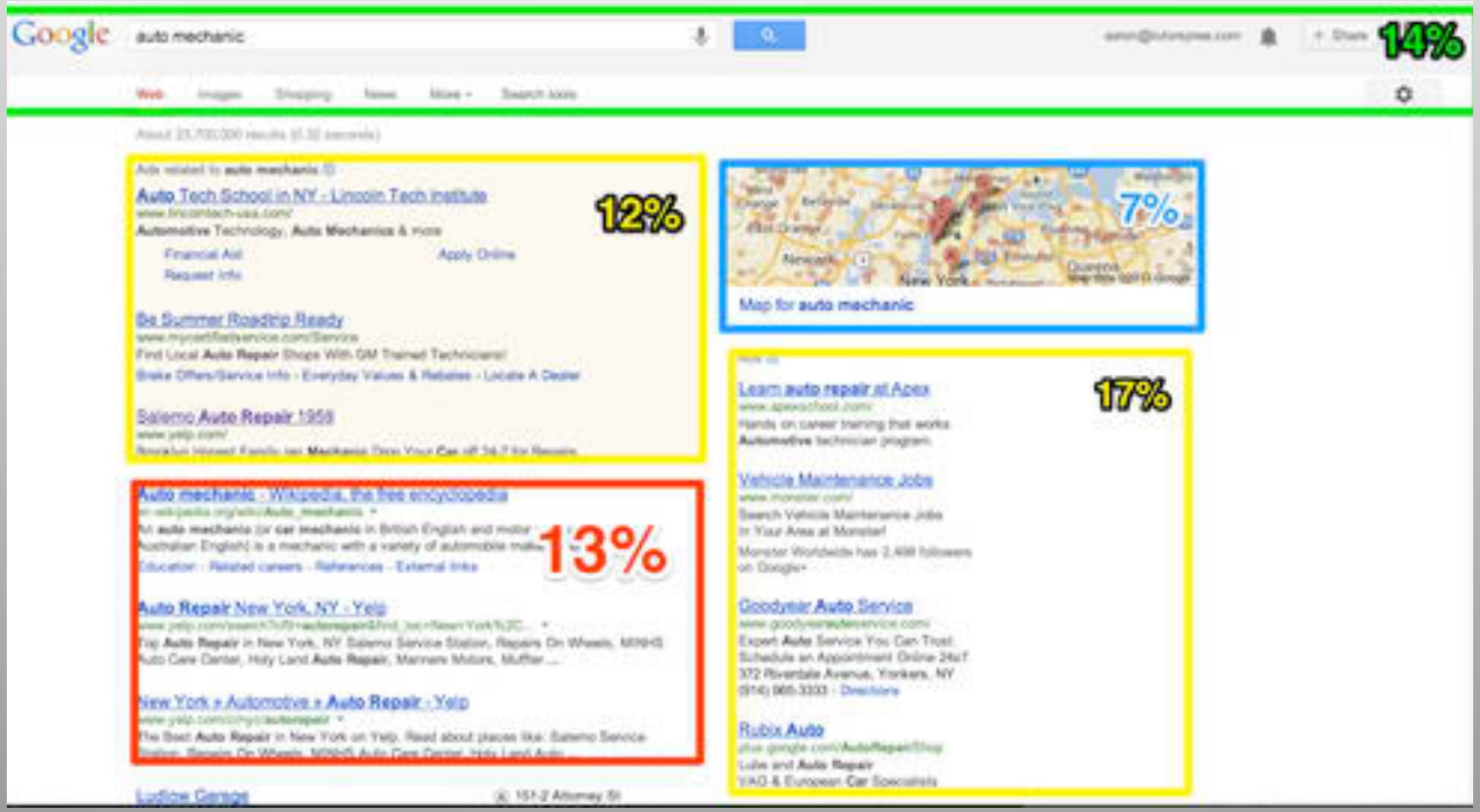
- Langue primitive quasiment sans grammaire : Liste de mots-clés
- Résultat imprécis : liste de pages
- Il est possible de faire mieux

Simplisme de PageRank

- Privilégie la popularité
- Encourage l'uniformité
- Opinions négatives

Et pourquoi le secret sur les critères de classement des pages ?

La recherche « neutre » : 13% et moins



Digression : neutralité

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 perte de neutralité nuit à notre liberté d'accès à l'information et à notre liberté d'expression

Les systèmes relationnels
comment on en est arrivé là

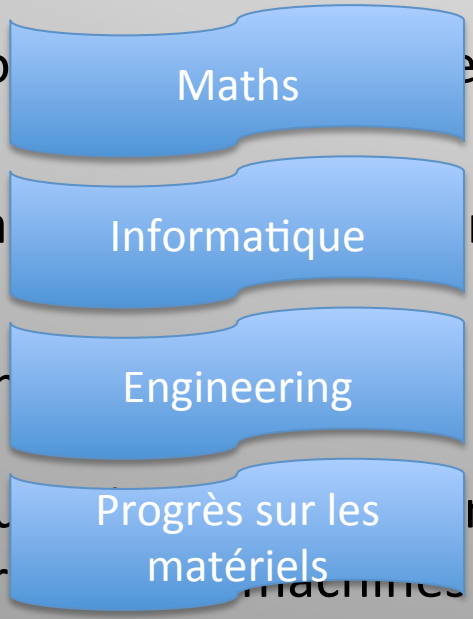
L'amélioration d'une fonction existante ou une nouvelle fonctionnalité

Des o…es

Des a…nts

Un er…

Et tou…mpte des progr…



Notamment, des modèle plus abstraits pour gérer des données

Notamment, la logique et l'algèbre relationnelles

Notamment, pour l'optimisation de requête

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

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

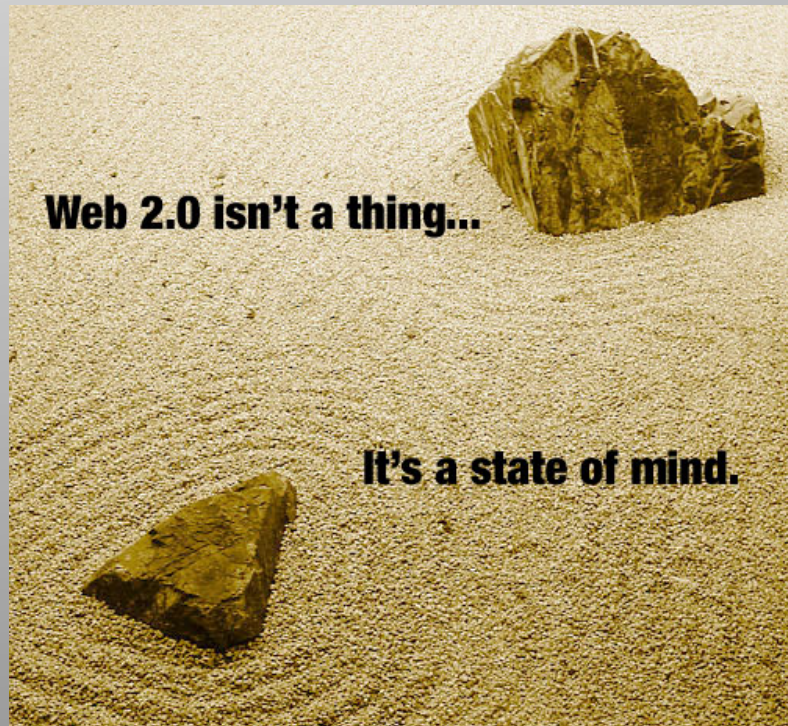
- Introduction
- Deux grands succès du 20e siècle
  - Les systèmes relationnels
  - Les moteurs de recherche de la Toile
- Deux défis du 21e siècle
  - **Réseaux et connaissances collectives** ←
  - La Toile des connaissances
- Conclusion

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

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

Problèmes

Difficulté de l'analyse statistique sur de gros volumes de données & de gros volumes d'utilisateurs

- Vérifier l'information, évaluer la qualité, résoudre contradictions

Manque d'explication

Manque de sérendipité

Manque de « privacy »

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 = Conf invitée à PFIQ ; type = Powerpoint ;

date = Juillet 2013 ; lieu = Lille ; 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

Par les systèmes qui nous entourent

- Senseurs, téléphones, appareils photo, e-commerce, réseaux sociaux…

Par extraction de connaissances du texte (web, emails…)

- Recherche de formes syntaxiques comme

Napoléon *est mort à* Sainte-Hélène

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

Problème : le raisonnement distribué

En utilisant des faits comme

Psychose est un film d'Hitchcock et Alice ne l'a pas vu

Et des règles comme

SouhaiteVoir(Alice, *t*) ← Film(*t*, Hitchcock, *a*), *not* Vu(Alice, *t*)

On peut **déduire** des faits « intentionnels » comme

Alice souhaiterait voir le film Psychose

Répondre à une requête est devenu plus compliqué

- Inférence de nouveaux faits en évitant de les inférer tous
- Collaboration entre des systèmes qui ont et infèrent des faits

Changement de contexte

- Immersion dans un monde de systèmes qui ont/échangent/ infèrent des connaissances
- Modification de notre manière de savoir et de penser

Illustration : Webdamlog

- Echange de connaissances en pair à pair
- Chaque pair Webdamlog
  - A sa propre base de données
  - A son propre moteur d'inférence (datalog)
  - Echange des données et des règles avec les autres
  - Délégation : mécanisme pour installer des règles chez un autre pair
- Raisonnement distribué

Webdamlog - exemple

Alice : je veux des photos où je suis avec Bob disponibles dans les téléphones de mes amis

```
result@alice($X, $U, $Meta) :-
    friends@facebook(alice,$X),
    smartphone@SNdirectory($X,$P),
    photos@$P($U,$Meta),
    contains@$P($Meta, “Alice”) ,
     contains@$P($Meta, “Bob”)
```

Plusieurs sortes de règles

- Local : tous les prédicats du corps sont locaux
- Extensionnel (stocké) vs. Intensionnel (dérivé)

Règles locales avec tête extensionnelle

```
ext-s@loc(x,y) :- r@loc(x,y)        % insertion dans une BD locale
ext-rn-loc(4,4) :- r@loc(3,3)       % envoie de message à un autre pair
```

Règles locales avec tête locale et intentionnelle

```
int-t@loc(x,y) :- r@loc(x,y)        % déduction locale à la datalog
```

Règles locales avec tête non-local et intentionnelle

```
int-t@n-loc(x,y) :- r@loc(x,y)      % délégation de vue
```

Règles non-locales

```
int-t@n-loc2(x,y) :- r@n-loc1(x,y)  % délégation générale
```

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 ←

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. …

La Toile est multiforme

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

Incontournable

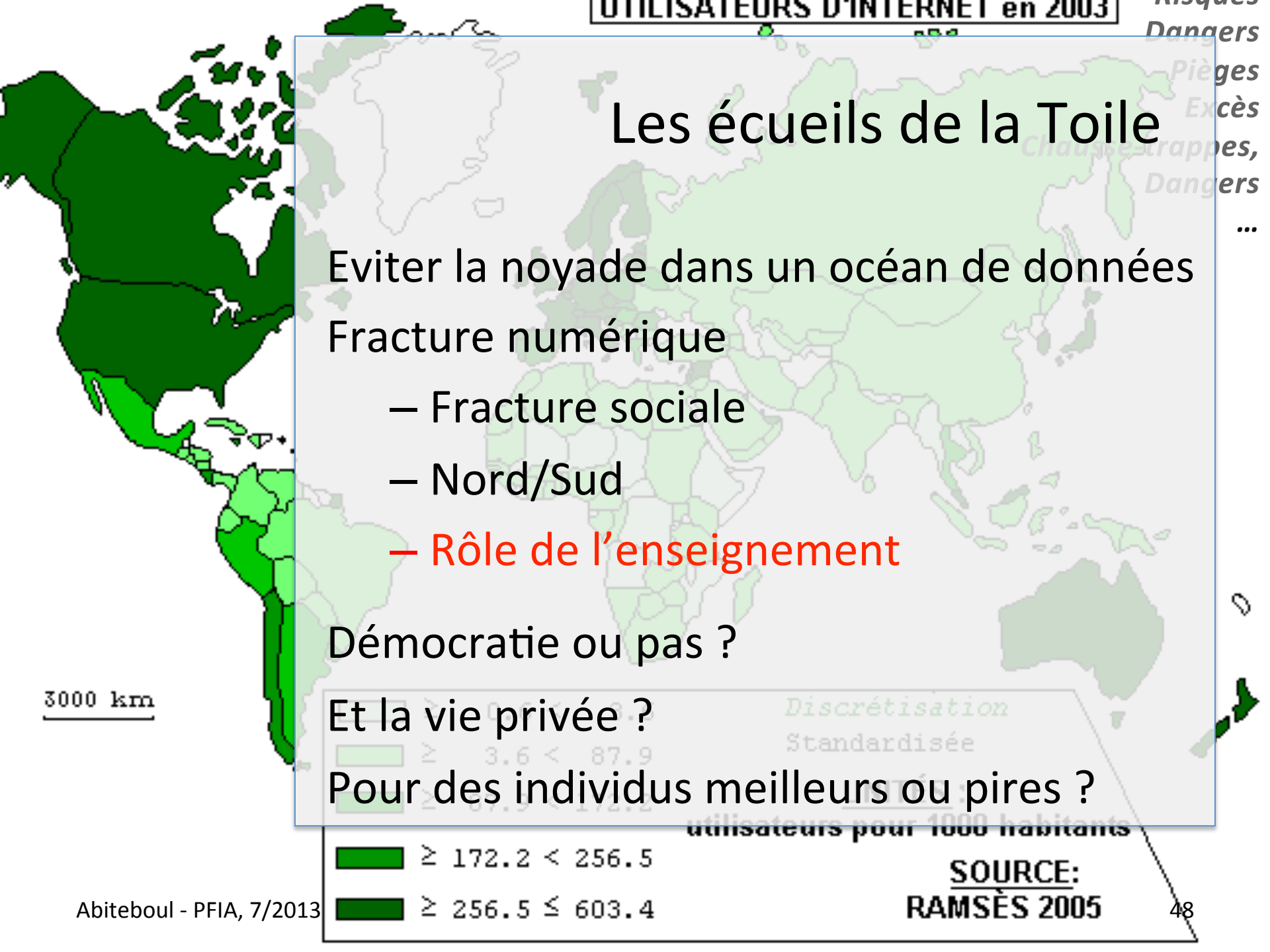
- Trouver du travail, travailler, se loger, gérer ses comptes bancaires, faire partie d'une association…

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

Les merveilles de la Toile

- Accès à toutes les connaissances du monde
- Accès à toute les cultures du monde
- Travail collaboratif – Vie et réseaux sociaux
- Suivi médical...



Les écueils de la Toile

Eviter la noyade dans un océan de données

Fracture numérique

- 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 ŒUVRES COMPLÈTES DE VICTOR HUGO
TROP PUISSANT!
C CON QU'ON C PAS LIRE!
DELIGNE

L'enseignement de l'informatique

Il est urgent de ne plus attendre

Extrait : La décision essentielle à prendre est de mettre en place un enseignement de science informatique depuis le primaire jusqu'au lycée, orienté vers la compréhension et la maîtrise de l'informatique, et dépassant donc largement les seuls usages des matériels et logiciels. Cette mise en place ne doit plus être différée.

Le numérique

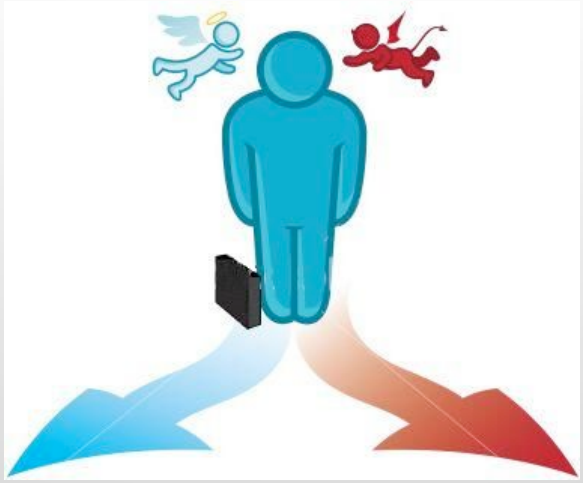


Illustration:
Big data &
La santé

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
Quel monde souhaitons-nous?

Et demain…

Des données,

à l'information,

aux connaissances…

Et demain…

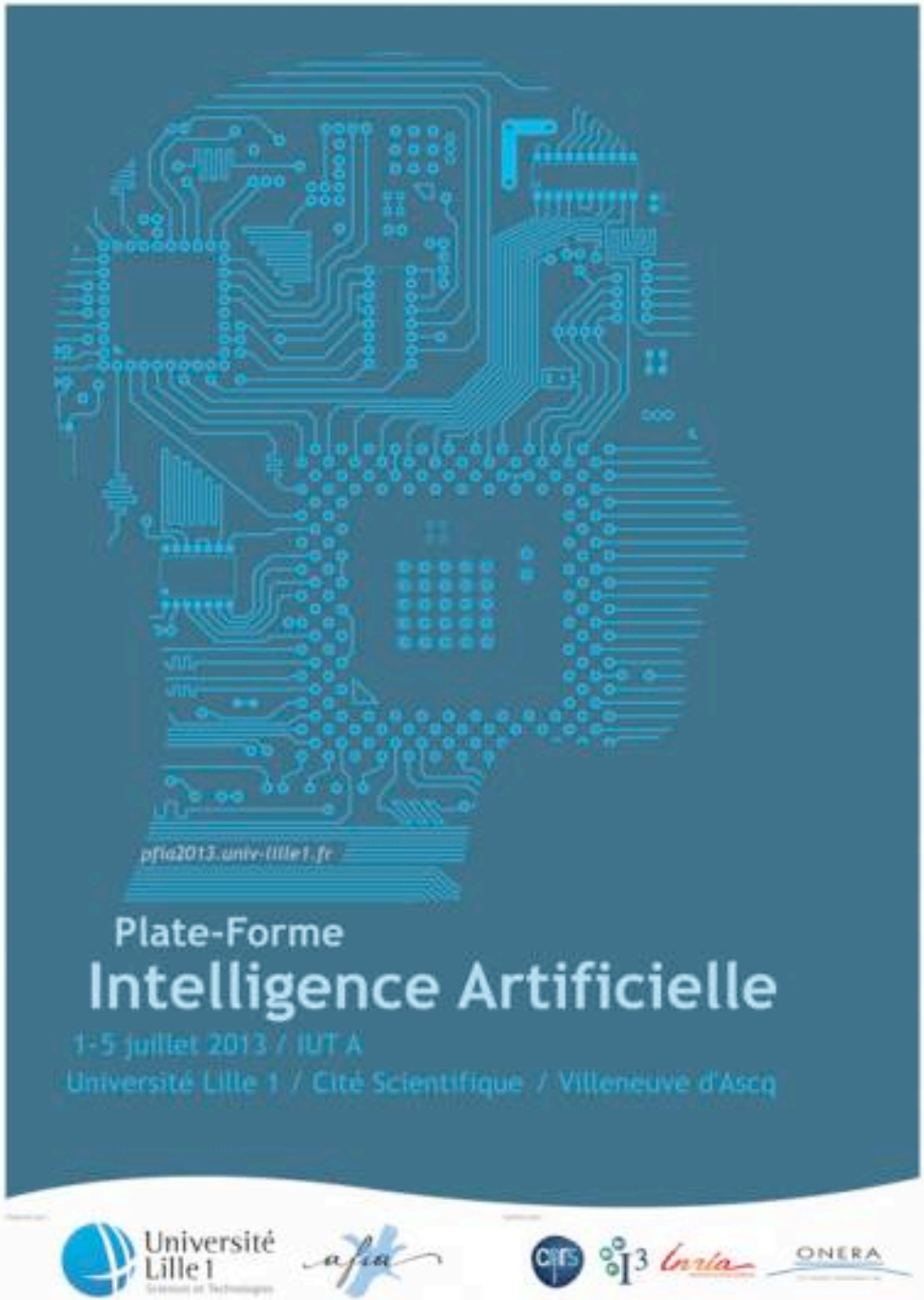
Nous vivrons dans un monde numérique entourés de systèmes qui traiteront l'information pour nous :

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

Nous passerons d'un monde fermé et précis… à un monde ouvert et imprécis, parfois incohérent

Et quelques thèmes de recherche…

- Comment extraire des connaissances de toutes les informations disponibles
- Comment décider ce qui est vrai/faux ?
- Comment décider ce qui est intéressant ?
- Comment gérer des flux massifs de données spatio-temporelles
- Comment garder le contrôle sur nos propres données & protéger notre vie privé
- Comment faire collaborer des milliards de systèmes de connaissances


pfia2013.univ-lille1.fr
Plate-Forme
Intelligence Artificielle
1-5 juillet 2013 / IUT A
Université Lille 1 / Cité Scientifique / Villeneuve d'Ascq
Université Lille 1
afia
cnrs
I3
inria
ONERA


Merci !

Pour en savoir plus

- ERC Webdam sur « Web data management »
  - Distributed reasoning, collaborative workflows, probabilistic data…
  - Corroboration : [wsdm10]
  - Webdamlog : [pods12/sigmod13]
- Cours au Collège de France
- Rapport sur l'enseignement de l'informatique [Académie des sciences]
- Avis sur la neutralité du net [Conseil national du numérique]