Sciences des données : De la Logique du premier ordre à la Toile

Serge Abiteboul

Chaire informatique et sciences numériques

informatiques mathématiques
Inria

COLLÈGE DE FRANCE
1530

À l'étudiante
en informatique,
en mathématiques
ou en 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 réalisations 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

Sciences des données : de la Logique du premier ordre à la Toile

Les systèmes informatiques servent à calculer

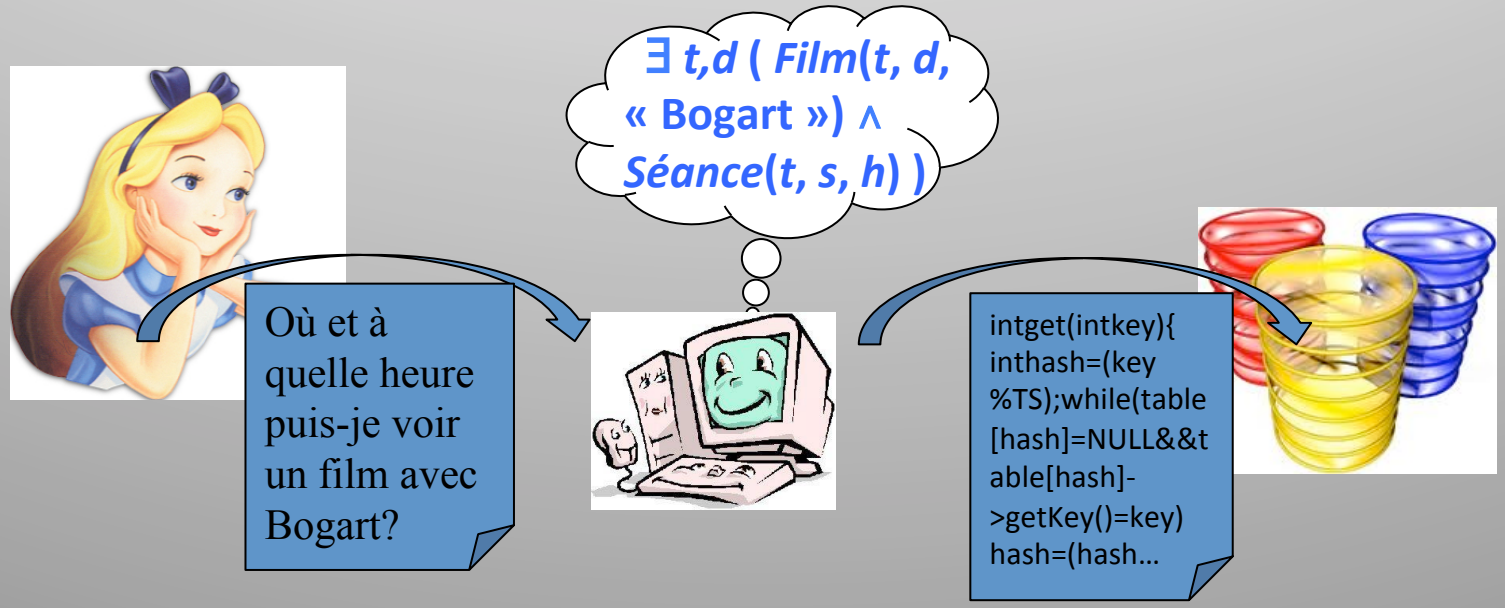
- Simulation de la météo
- Cryptographie
- Etc.

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

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

Sciences des données : de la Logique du premier ordre à la Toile

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



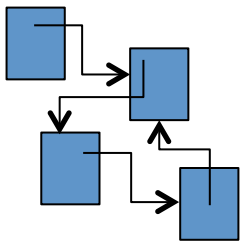
Sciences des données : de la Logique du premier ordre à la Toile

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[illegible]munautés

Wikipedia : encyclopédie

Amazon, eBay : catal[illegible]eb

YouTube, [illegible]

Twitt[illegible]on, news

Flickr, [illegible]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 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 2008)

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 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 prop… …omme parle

1 zettaoctet = **10^{21}** octe…

- **½ zetta = le …**
- …e envoyée au cerveau en une année

L'univers digital double tous les 18 mois

Source : C… …rking 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*

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

- Introduction
- Deux réalisations 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

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. Leaud
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 :

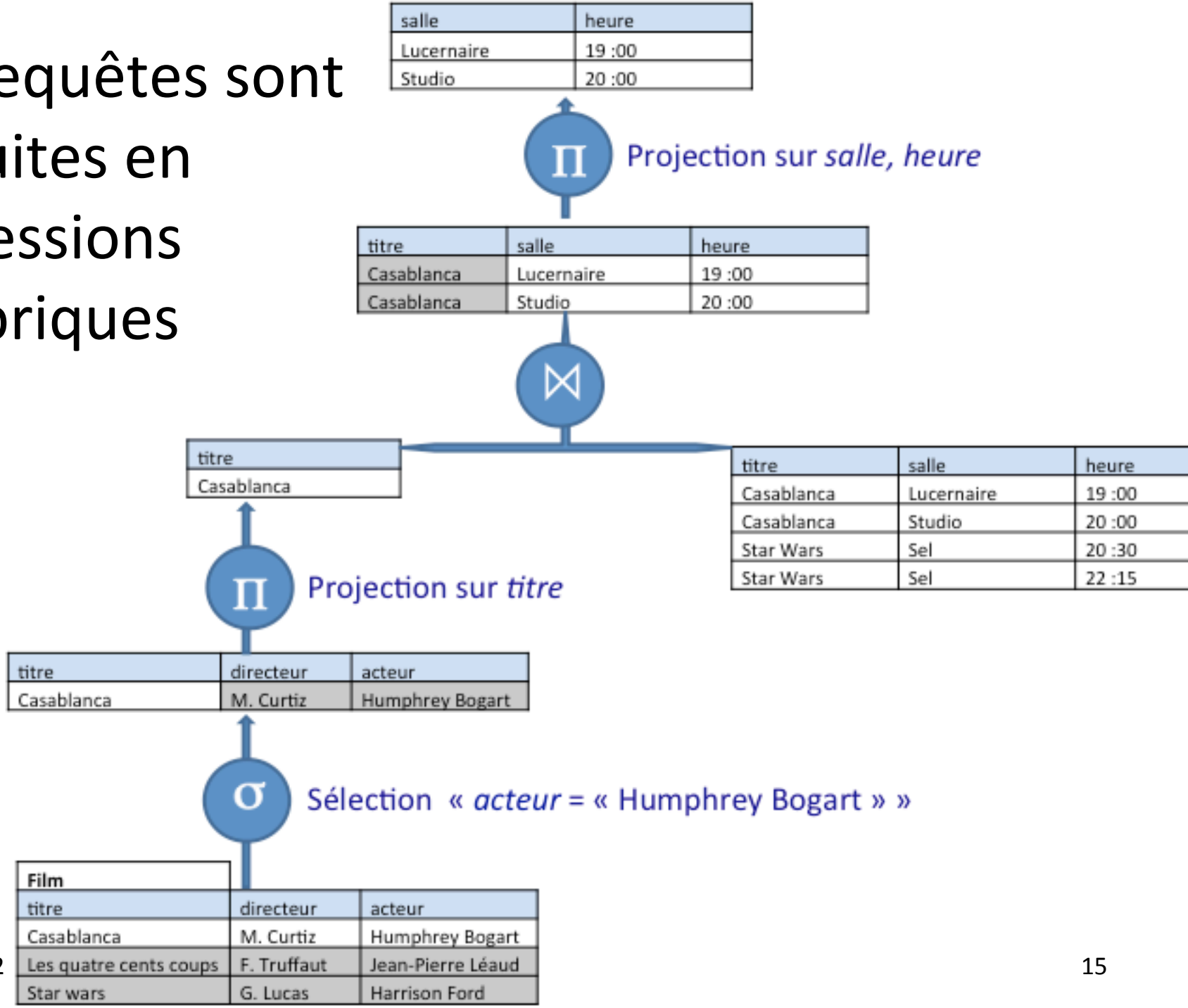
SQL :

select *salle, heure*

from Film, Séance

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

Les requêtes sont traduites en expressions algébriques



Optimisation de requêtes

C'est-à-dire : choisir le plan d'exécution le moins coûteux possible (typiquement en temps) pour calculer la requête

Problème :

- L' « espace de recherche », c'est-à-dire l'espace dans lequel nous voulons trouver le plan d'exécution, est potentiellement gigantesque. On utilise des « heuristiques » pour éviter de le parcourir
- On doit être capable d'estimer très rapidement le coût de chaque plan candidat (pour trouver le moins coûteux)

Les optimiseurs de systèmes relationnels comme Oracle ou DB2 font des merveilles sur des requêtes simples

- En pratique, la plupart des requêtes posées sont simples

De la complexité des requêtes

Certaines propositions ne peuvent être ni démontrées ni réfutées & certains problèmes ne peuvent être résolus [Church-Turing]

Certaines tâches sont trop coûteuses à réaliser

- Par exemple, factoriser un très grand entier en nombres premiers

Complexité d'une tâche

- Pour nous : en fonction de la taille des données
- En temps : quel temps est nécessaire pour la réaliser ?
- En espace : quel espace disque (ou quelle mémoire) est nécessaire ?

Exemple

- Temps linéaire : si je double la taille des données, je double le temps
- Temp *P* ; en n^k où *n* est la taille des données
- Temps *exptime* : en k^n

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; Th. 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

Un problème ouvert

Toute requête du calcul relationnel peut être évaluée en *P*

Réciproquement, peut-on exprimer en calcul toutes les requêtes calculables en *P?* Non

- Étant donné un graphe *G*, et deux points *a, f* de ce graphe, est-ce qu'il existe un chemin de *a* à *f* ?
- On peut demander s'il existe un chemin de longueur 3 ou même *k*, pour un *k* fixé

Trouver un langage logique qui permettrait d'exprimer **toutes les** requêtes calculables en temps polynomial mais qui ne permettrait d'exprimer **que des** requêtes en temps polynomial

Un pont entre le logiquement exprimable et le facilement calculable

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.

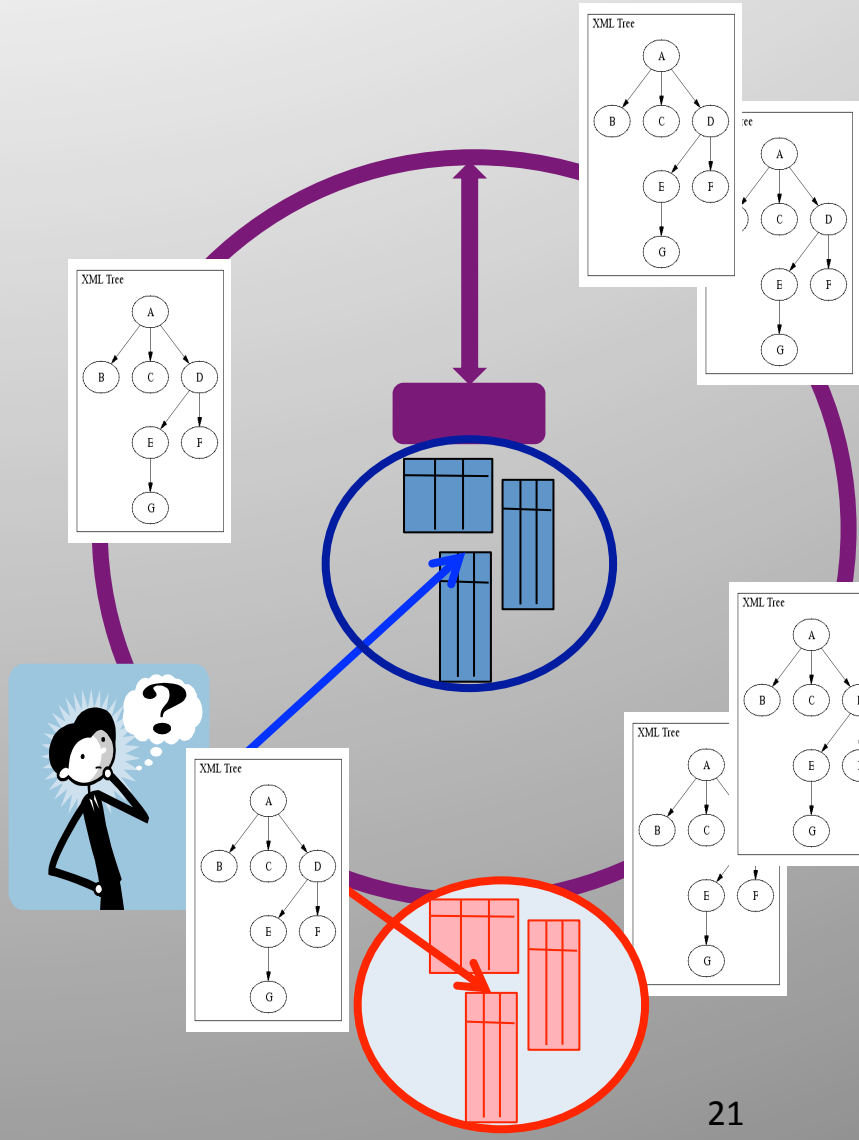
Sergey Brin et Larry Page,
fondateurs de Google.
Interview dans le magasine *Playboy*, 2004

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



Parallélisme

Essentiel pour gérer de gros volumes de données

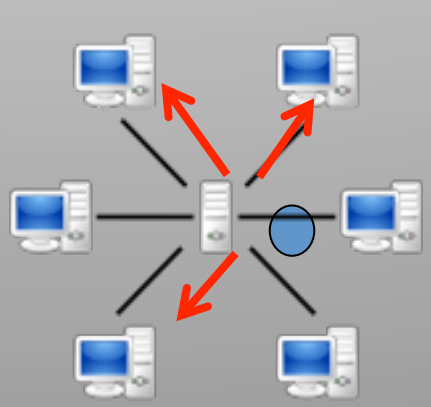
- Meilleure disponibilité, performance, etc.

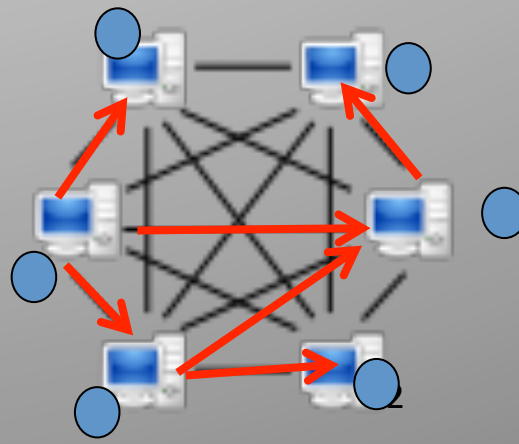
Quel parallélisme?

- 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





Index 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**

Prouesse et magie

On vous a dit

- La Toile est extraordinaire par la quantité d'informations qu'elle contient

Non

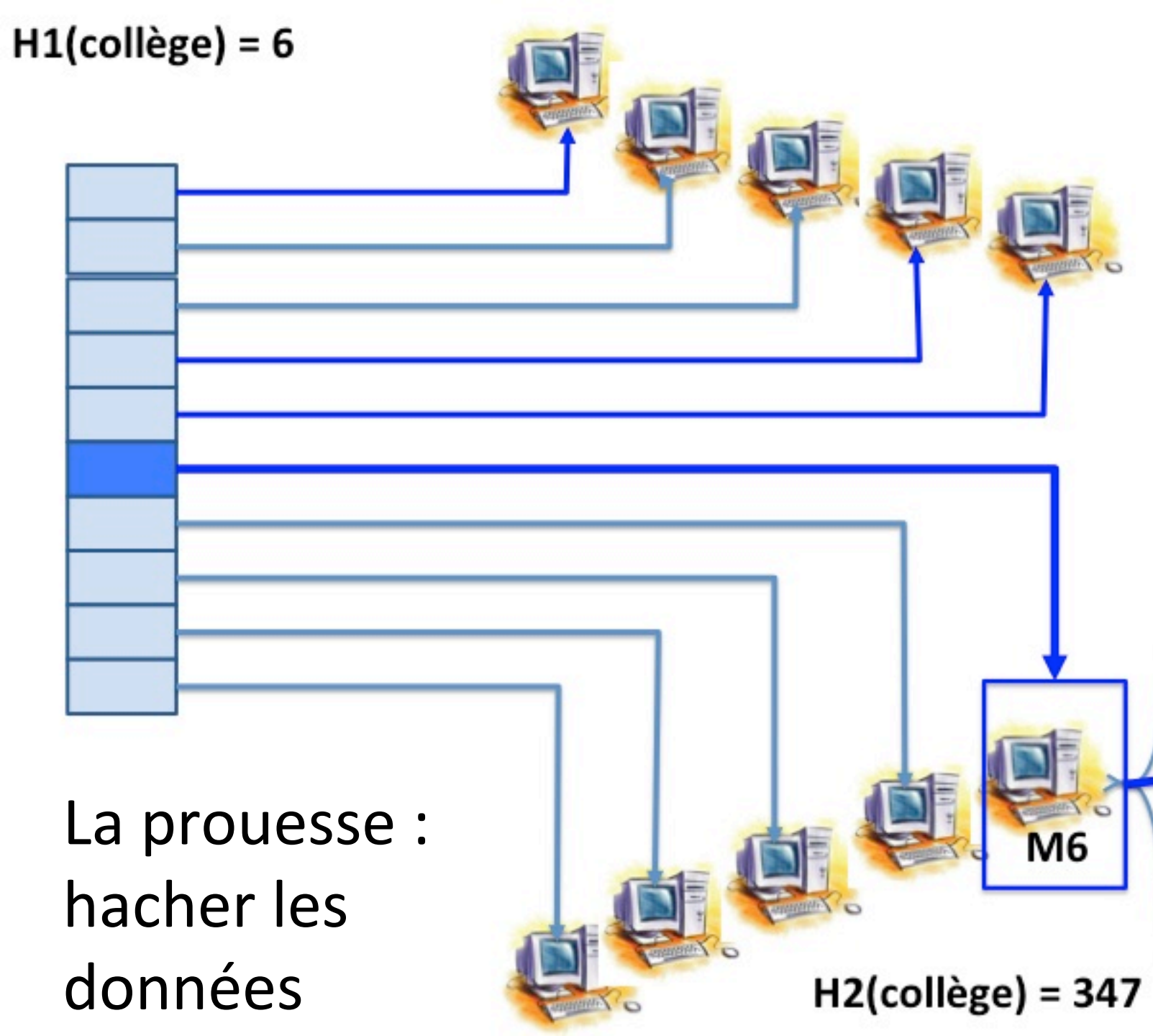
- Plus il y a d'informations, plus c'est compliqué de trouver la bonne information

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 prouesse : hacher les données

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 Madame 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 sûrement possible de faire mieux

Simplisme de PageRank

- Privilégie la popularité; encourage l'uniformité
- Ne tient pas compte des opinions négatives

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

Les systèmes relationnels
comment on en est arrivé là

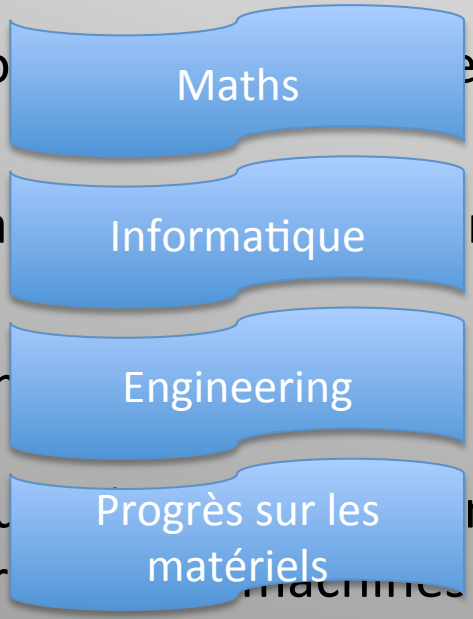
L'amélioration d'une fonction existante ou une nouvelle fonctionnalité

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

Notamment, les techniques de point fixe

Des a…nts

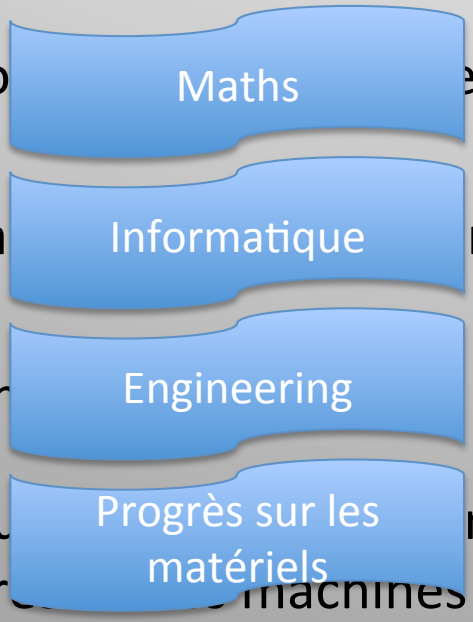
Notamment, l'utilisation du parallélisme massif

Un er…

Notamment, faire fonctionner des fermes de machines

Et tou…mpte des progr…machines

Baisse du prix des mémoires

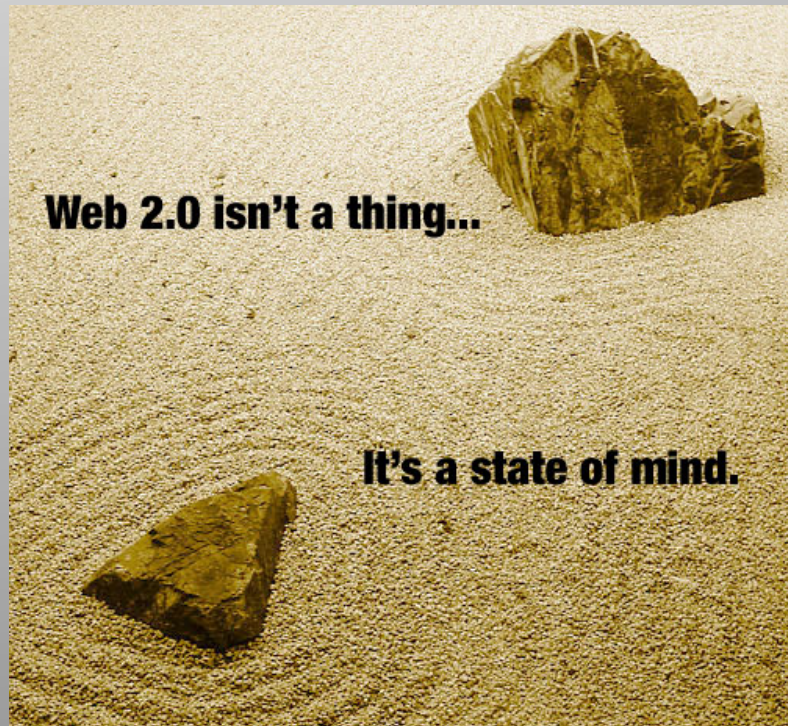


Après les réseaux de machines, puis de contenus, les **réseaux d'utilisateurs**

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



Connaissances collectives

Plusieurs approches

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

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

Des problèmes ouverts

Analyses statistiques

- Grand nombre de personnes et gros volumes de données
- Nécessité de vérifier l'information, évaluer sa qualité, résoudre les contradictions

Manque d'explication

- Les systèmes ne savent pas expliquer des choix qui font intervenir de gros calculs

Atteintes à la confidentialité : conflit entre

- L'utilisateur qui veut protéger ses données confidentielles
- Les systèmes qui veulent ces données pour offrir un meilleur service, plus personnalisé

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 réalisations 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 un document

auteur = Serge Abiteboul ; titre = Sciences des données

nature = leçon inaugurale ; date = Mars 2012 ; 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

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

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

- Introduction
- Deux réalisations 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

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

Vision dépassée 1 : hypertexte

Vision dépassée 2 : bibliothèque universelle de documents

Et tous les Web

- Web des téléphones « intelligents »
- Web des réseaux sociaux (d'avant 7 à plus de 77 ans)
- Web sémantique
- Web des objets communicants et intelligence ambiante
- Web des mondes virtuels (jeux 3D)
  - …

Risques
Dangers
Pièges
Excès
Chausse-trappes,
Dangers
…

Les écueils de la Toile

Eviter la noyade dans un océan de données

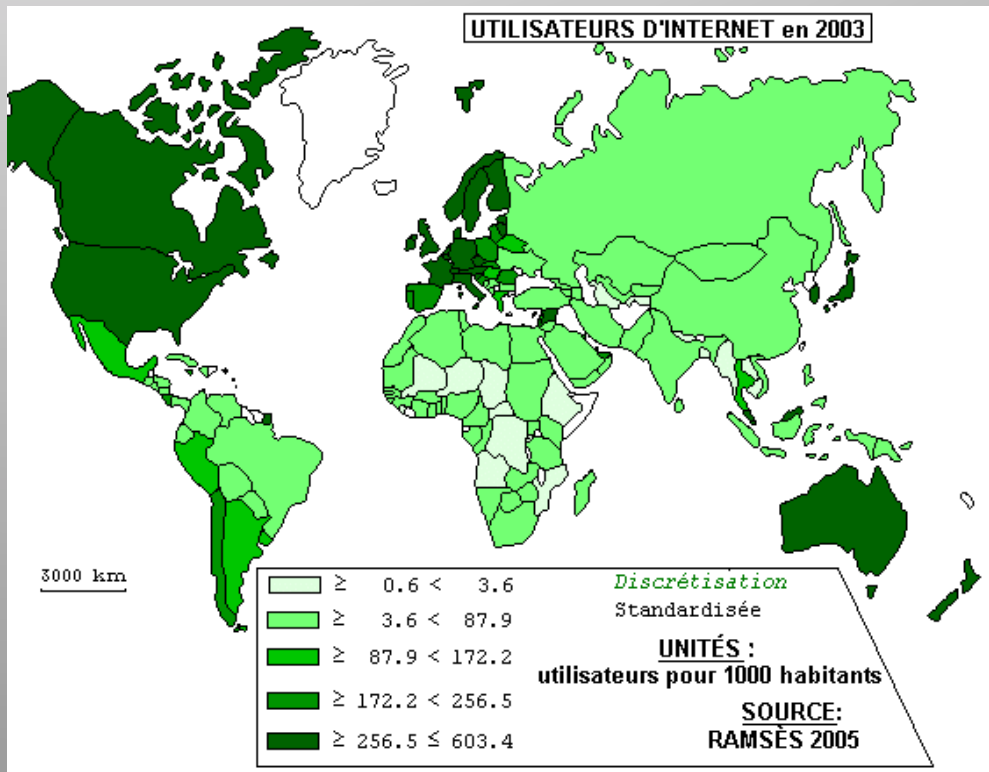
- Un des fils conducteurs de cette présentation

Accès à l'information pour tous

- Fracture sociale

  CREDOC 2009 : en France,
  40% de la population n'utilise jamais l'informatique

- Nord/Sud
- **Enseignement**



Les écueils de la Toile (fin)

Démocratie ou pas ?

Et la vie privée ?

Pour des individus meilleurs ou pires ?

Je veux continuer à croire que la Toile participera à féconder un meilleur futur

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

Et demain…

Choix politiques

Nouveaux outils informatiques qui restent à inventer

Et pour ce qui concerne les aspects scientifiques

La prochaine étape des sciences des données, que l'on retiendra, a déjà commencé : c'est

la Toile des connaissances

Des données,

à l'information,

aux connaissances…

Remerciements : Martín Abadi, Jérémie Abiteboul, Manon Abiteboul, Gilles Dowek, Emmanuelle Fleury, Laurent Fribourg, Sophie Gamerman, Bernadette Goldstein, Florence Hachez-Leroy, Tova Milo, Marie-Christine Rousset, Luc Segoufin, Pierre Senellart et Victor Vianu

informatiques mathématiques
inria

COLLÈGE DE FRANCE
1530

ENS
CACHAN