Data Sciences: From First-order Logic to the Web

Serge Abiteboul

Chaire informatique et sciences numériques

To all female students
in informatics,
in mathematics
or in the 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 ←

Two achievements for the 20th century

- Relational systems
- Web search engines

Two challenges of the 21st century

- Networks and collective knowledge
- The web of knowledge

Conclusion

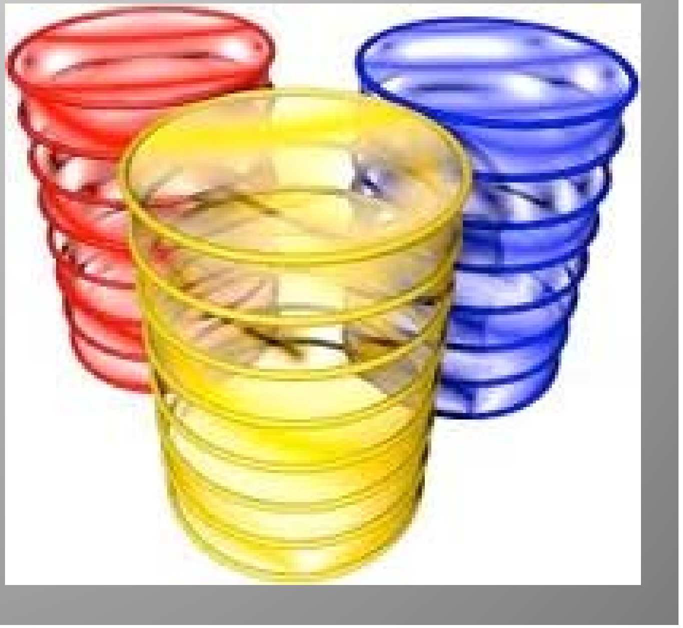
Data sciences: From first-order logic to the web

Computer systems are used to compute

- Weather simulation
- Cryptography
- Etc.

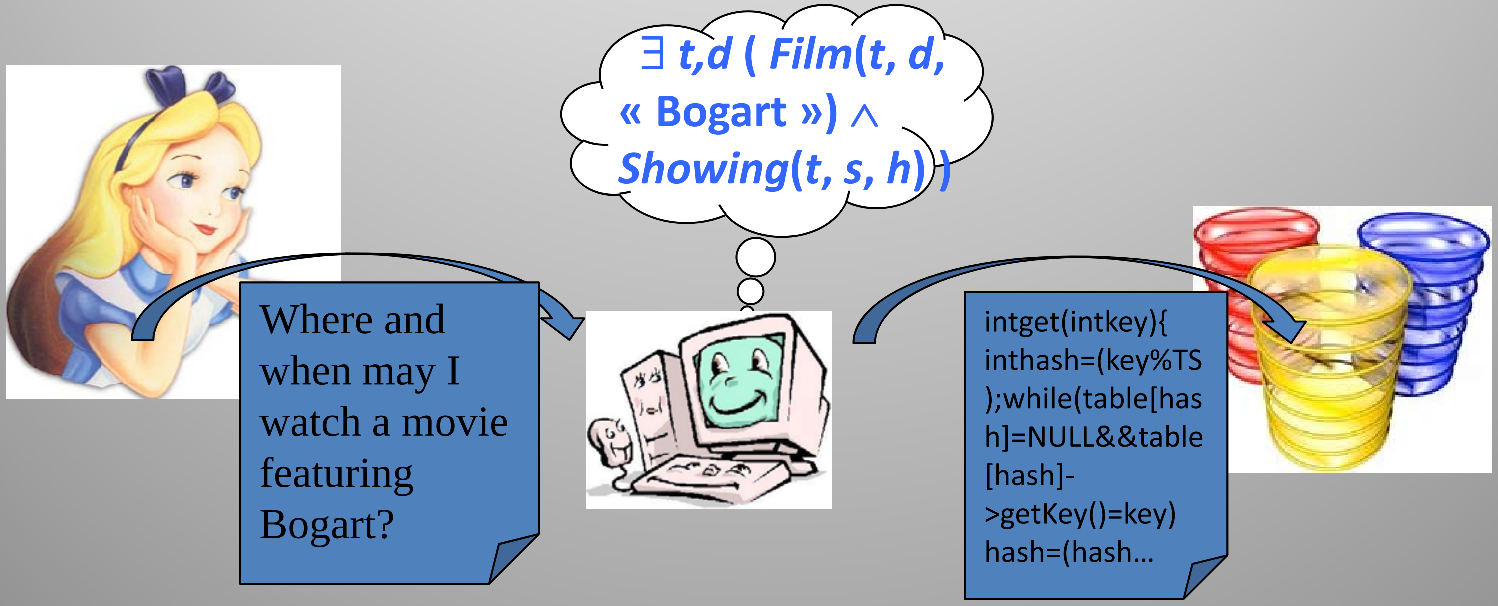
They are often used to store/manage **data**

- Accounting
- Product catalog
- Inventory
- Agenda
- Contacts
- Library
- Etc.



Data sciences:
From **first-order logic** to the web

Computer systems play the role of **mediators** between intelligent users and objects storing information



Where and when may I watch a movie featuring Bogart?

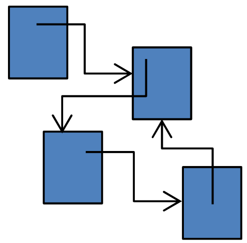
$\exists\, t,d\ (\mathit{Film}(t, d, \text{« Bogart »}) \land \mathit{Showing}(t, s, h))$

```
intget(intkey){
inthash=(key%TS
);while(table[has
h]=NULL&&table
[hash]-
>getKey()=key)
hash=(hash...
```

Data sciences:
From first-order logic to the **web**

Today information is found on the « World Wide Web », a public hypertext system (*) on the Internet (**) that allows consulting, using a browser, pages usually found with web search engines

(*) Hypertext



(**) Internet

A network that allows exchanging flows of information between machines

Success stories on the web

Google: web pages

Facebook: personal data

Wikipedia: encyclopedia

Amazon, eBay: web catalog

YouTube, Dailymotion: v

Twitter: communicat

Flickr, Picasa: photos

iTunes, Kazaa, Emule, Batanga, BearShare: music

Myspace: personal pages

Wikileaks: state secrets

They are all about data management

What do they have in common ?

The numerical world

Billions of communicating objects

Hundreds of millions of web sites

1000 billions of pages (September 2008)

More than 10 billion searches/month on the web (April 2008)

We are living in a gigantic numerical world

Measuring it with coffee spoons

8 bits = **1** byte

1 terabyte = **10^{12}** bytes

- 200 terabytes = all the books ever written

1 petabyte = **10^{15}** bytes

- 100 petabyte = the volume of data produced by the CERN particle collider in a minute

1 exabyte = **10^{18}** bytes

- 5 exabytes = le volume of the data corresponding to all the words ever pronounced by humans

1 zettabyte = **10^{21}** bytes

- **½ zetta = the Internet traffic in 2012** – 0.5.10^{21}
- 66 zetta: the visual information sent to the brain in a year

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

Data, information, knowledge

Data	Elementary description of some reality	*Temperature measurements in a weather station*
Information	Data with a meaning (to construct a representation of some reality)	*A curb giving the evolution of the average temperature in a place throughout the year*
Knowledge	Information equipped with some notion of truth and more generally some general laws that have been inferred	*The fact that temperature on the earth is augmenting because of human activity*

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

Introduction

Two achievements for the 20th century

- **Relational systems** ←
- Web search engines

Two challenges of the 21st century

- Networks and collective knowledge
- The web of knowledge

Conclusion

Classic data management

A great success of the 20th century

- Academic and industrial research
- Theoretical foundations
- Commercial systems such as Oracle, DB2, SQL Server
- Open-source software such as mySQL

Relational model, Edgar F. Codd-1970

- Strongly inspired by First-order logic
- Developed in the late 19th century by mathematicians to formalize the language of mathematics

Relations

Film

Title	Director	Actor
Casablanca	M. Curtiz	H. Bogart
Casablanca	M. Curtiz	P. Lore
Les 400 coups	F. Truffaut	J.-P. Leaud
Star Wars	G. Lucas	H. Ford

Showing

Title	Theater	Time
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

Queries are expressed in relational calculus

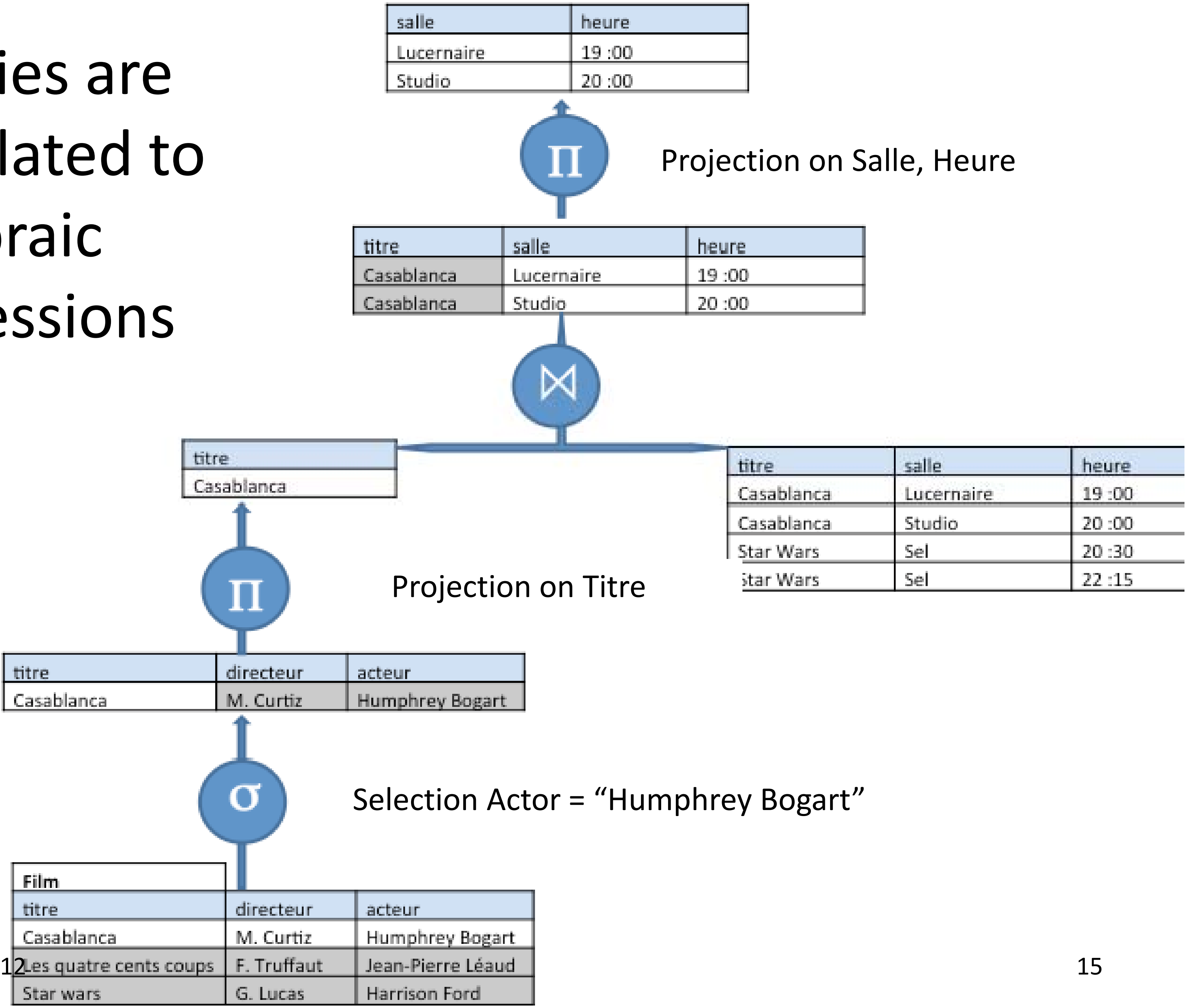
q_{HB} = { *Theater, Time* | $\exists$ *Director, Title*
(Film(*Title, Director*, « Humphrey Bogart ») $\wedge$
Showing(*Title, Theater, Time*) }

In practice, relational systems use a simpler syntax

SQL:

select *Theater, Time*
from Film, Showing
where Film.*Title* = Showing.*Title* **and** *Actor*= «Humphrey Bogart»

Queries are translated to algebraic expressions



salle	heure
Lucernaire	19 :00
Studio	20 :00

Projection on Salle, Heure

titre	salle	heure
Casablanca	Lucernaire	19 :00
Casablanca	Studio	20 :00

titre
Casablanca

titre	salle	heure
Casablanca	Lucernaire	19 :00
Casablanca	Studio	20 :00
Star Wars	Sel	20 :30
Star Wars	Sel	22 :15

Projection on Titre

titre	directeur	acteur
Casablanca	M. Curtiz	Humphrey Bogart

Selection Actor = “Humphrey Bogart”

Film

titre	directeur	acteur
Casablanca	M. Curtiz	Humphrey Bogart
Les quatre cents coups	F. Truffaut	Jean-Pierre Léaud
Star wars	G. Lucas	Harrison Ford

Query optimization

Goal: choose the execution plan with the lowest possible cost (typically time) to evaluate the query

Problem

- The "search space", that is to say the space in which we search for an execution plan, is potentially enormous. Use "heuristics" to reduce it
- One must be able to estimate very quickly the cost of each candidate plan (to find the least expensive)

Optimizers of relational systems such as Oracle or DB2 perform very well on simple queries

- In practice, most queries are simple

On the complexity of queries

Some logical statements can be neither proven nor disproved & some problems cannot be solved [Church-Turing]

Some problems can be solved but solutions are simply too expensive

- For example, factoring a large integer into prime numbers

Complexity of a task based on the size of the data

- Time: how long it takes
- Space: how much disk space (or memory) it requires

Example

- Linear time: if I double the size of the data, I double the time
- P: time in n^k where n is the size of the data
- EXPTIME: time in k^n

Why these systems are so successful

Queries are expressed in relational calculus

- A logical language, simple and understandable especially in variants such as SQL

Calculus queries can be translated into algebraic expressions

- That are easy to evaluate; Codd's theorem

The evaluation of algebraic expressions can be optimized

- Because it is a limited model of computation

Parallelism allows scaling to very large databases

- Relational calculus queries are in the complexity class AC^0
- Highly parallelizable

An open problem

Any relational calculus query can be evaluated in P

Conversely, can all queries computable in P be expressed using relational calculus? No!

- Given a graph G, and two points a, f of the graph, is there a path from a to f?
- One can ask whether there is a path of length 3 or even k, for k fixed

Problem: Find a logical language that would express all queries computable in polynomial time but that would express only queries computable in polynomial time

This would provide a bridge between what is logically expressible and what is easily computable

*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 and Larry Page,
founders of Google.
Interview in *Playboy Magazine*, 2004

Introduction

Two achievements for the 20th century

- Relational systems
- **Web search engines**🡐

Two challenges of the 21st century

- Networks and collective knowledge
- The web of knowledge

Conclusion

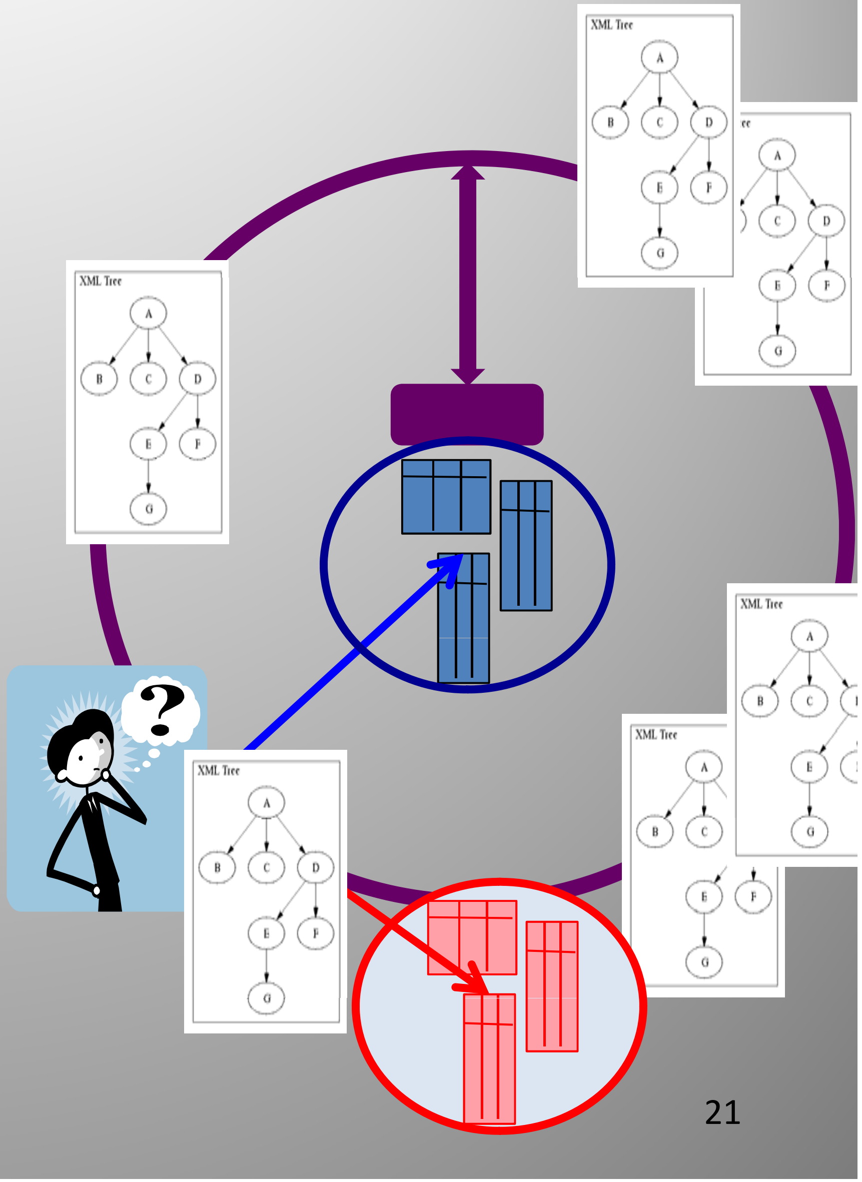
What changed with the web

Information was living in islands with different formats, application programming languages, operating systems

The web brought universal standards for sharing information

We now have:

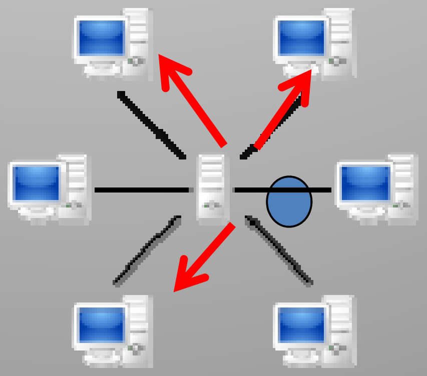
- Uniform and universal access to information
- Access to huge volumes of information

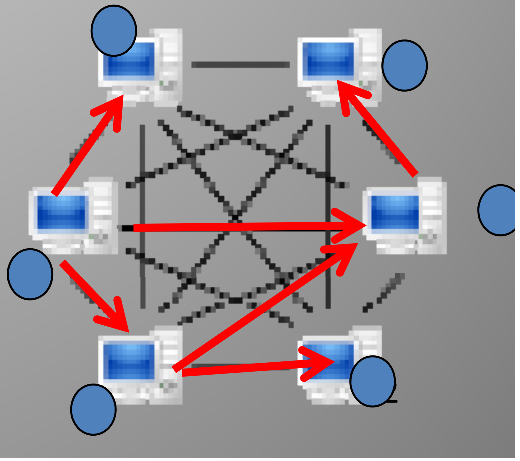


Parallelism

- Essential to manage large volumes of data
  - Improve availability, performance, etc..
- What kind of parallelism?
  - Machines are increasingly multi-processors
  - Collaboration between servers at different sites of a company
  - Hundreds or thousands of servers in a "cluster”
  - Millions of web servers

- Example: two organizations for movie distribution
  - Each movie on a single server
    - If the movie is too popular, the server saturates
  - Peer-to-peer architecture each machine is a client and server





A web index

The index gives, for each word, the list of pages that contain this word

Word	Page number
…	
collège	34,56,223,9900,111111…
…	
france	56,778,6560,9900,9999…
…	
informatique	9890,11122290…
…	

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

Scaling

The more pages are indexed, the larger the index grows

- Billions of pages
- The index size is roughly the size of the indexed data
- Each query is becoming increasingly expensive to evaluate

The more users the engine has, the more queries it receives

- Tens of billions of search queries per month

Solution: parallelism

Skill and magic

You were told that the web is extraordinary because of the amount of information it contains

No

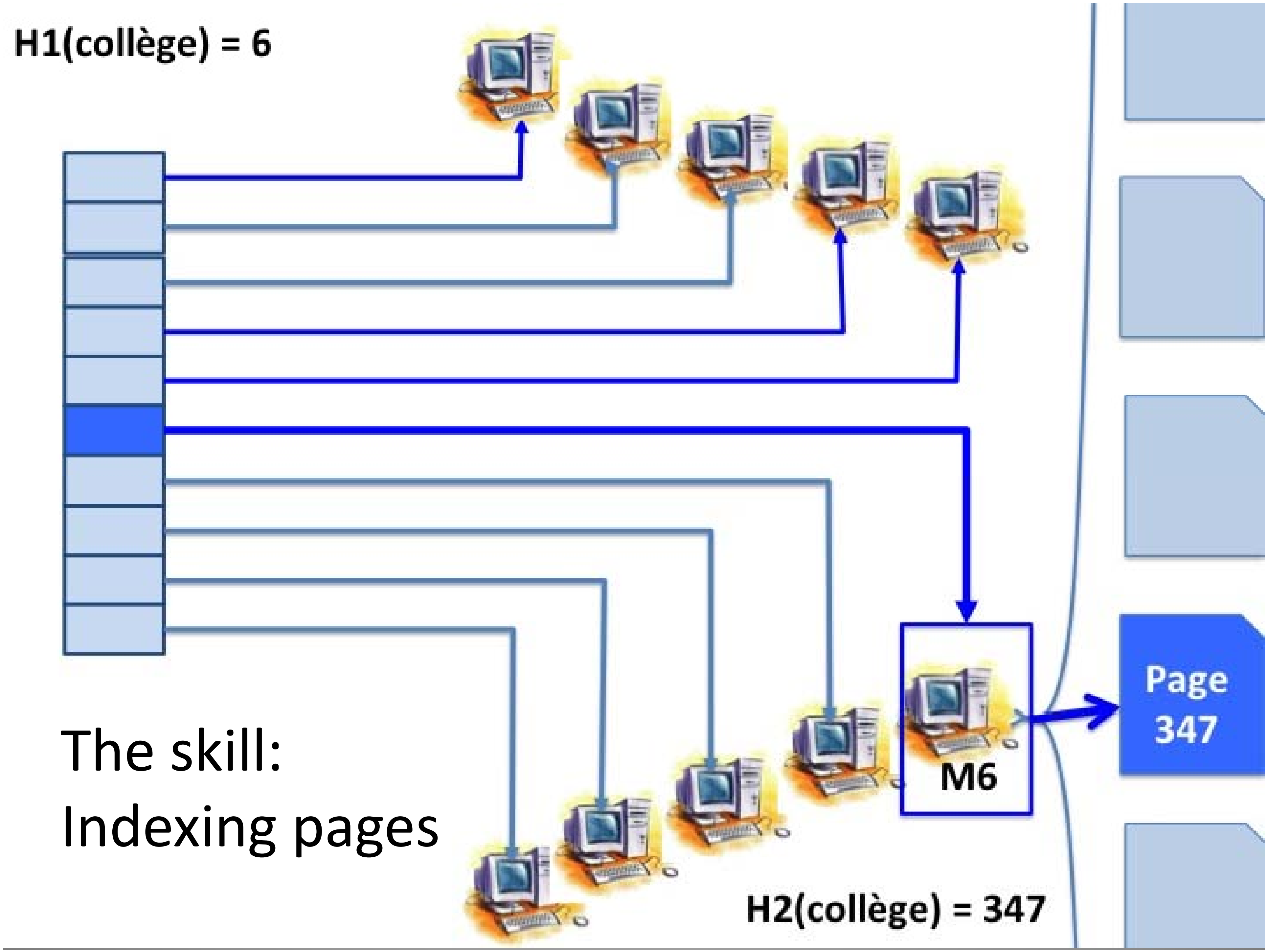
- The more information, the more complicated it is to find the right information
- What matters is the quality of information

The skill: indexing billions of pages

- Using techniques such as hashing

The magic: finding what you want (in general)

- Using "measures" to rank pages such as TFIDF and PageRank



The skill:
Indexing pages

The magic: ranking pages

Random surfer on the web

- Popularity = probability to be in a page
- The probability to be in lemonde.fr is larger than that of being on the personal page of Madame Michu

Turn this into an equation: $pop = \Theta \times pop$

And then solve the equation

- pop_0 defined by $pop_0[i] = 1/N$
    - All pages are assumed to be equally popular
- $pop_1 = \Theta \times pop_0$
- $pop_2 = \Theta \times pop_1$
- $pop_3 = \Theta \times pop_2 \ldots$

The fixpoint gives the popularity

Open problems

Queries are too simplistic

- Primitive language with almost no grammar: list of keywords
- Imprecise result: list of pages
- We can do better

PageRank is too simplistic

- Based solely on popularity; how about originality?
- Does not take into account negative opinions

Too much power for a few search engines

And why this secret on the ranking criteria?

Relational systems

How did we get there?

The improvement of a
function
function

Mathematics

Mathem

Informatics

Smart al

Engineering

Solid engineering

Hardware progress

Taking i re
progress

Abstract data model

Relational calculus and algebra

Query optimization and concurrency control

Failure recovery

Disk capacity

Web search engines

How did we get there?

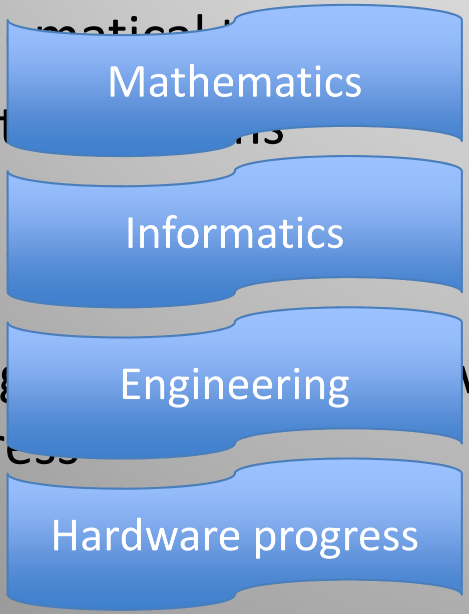
The improvement of a functionality or a new functionality

Math

Smart

Solid

Taking ware progr



Mathematics

Informatics

Engineering

Hardware progress

Better ranking of pages

PageRank fixpoint definition

The use of massive parallelism

Running clusters of thousands of machines

Huge and cheap memories

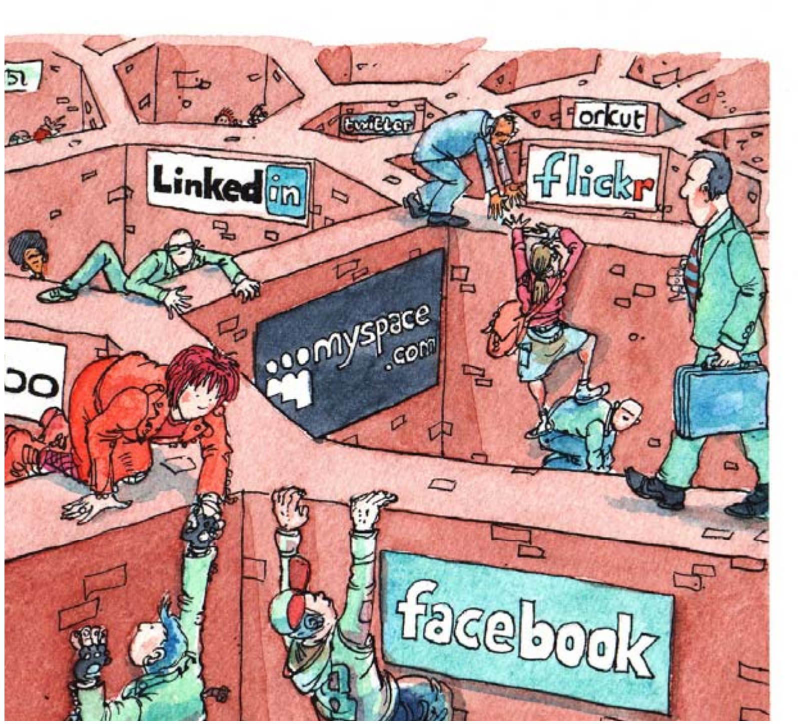
- Introduction
- Two achievements for the 20th century
    - Relational systems
    - Web search engines
- Two challenges of the 21st century
    - **Networks and collective knowledge** ←
    - The web of knowledge
- Conclusion

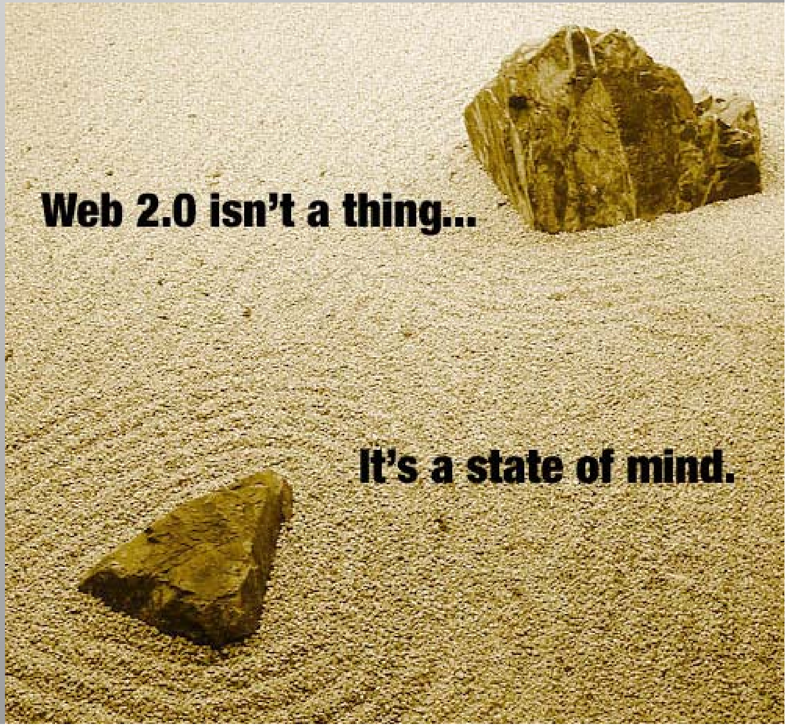
After the networks of machines, the networks of contents, the networks of users

The web is not just for obtaining data

Everyone can participate: tweets, Wikipedia, mashups

Keywords: interaction, communities, communication, networks





Collective knowledge

Different approaches

- Grading
- Expertise evaluation
- Recommendation
- Collaboration
- Crowdsourcing

Grading

Learn the opinion of the internaut

- Quantitative (***)
- Qualitative (“perfect for dating”)

eBay: customers grade vendors

Increasingly standard

- Movie in allocine
- Restaurant in ViaMichelin
- Webpage annotations in Delicious

Expertise evaluation

Evaluate

the quality of information

the quality of information sources

Illustration: recent work on corroboration

How is expertise built on the web ?

- Specialized blogs for law, medicine, art…
- Citizen blogs in Tunisia or Syria

Will reputation be some day determined by programs ?

Recommendation

Use web data for deriving recommendations

- Meetic organizes dates
- Netflix suggests movies
- Amazon suggests books

Statistical analysis to discover “proximities”

- Between customers in Meetic
- Between customers and products in Netflix or Amazon

amazon

Collaboration

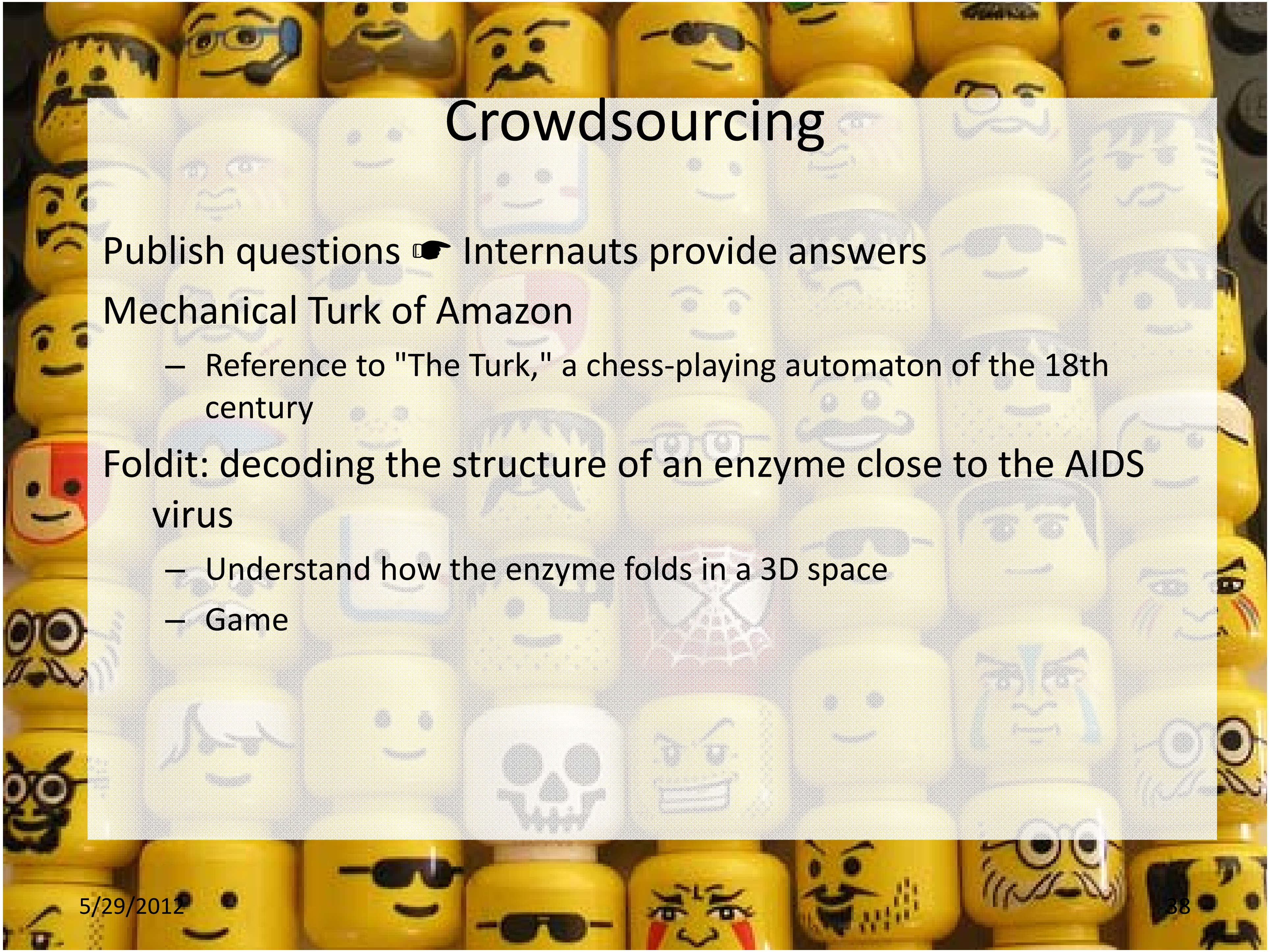
Internauts perform collectively tasks they cannot solve individually

Wikipedia: encyclopedia

- 281 editions; 3 millions articles for the English version
- Extremely used
- Covers a much larger spectrum than a traditional encyclopedia
- Controversial quality

Linux: operating system in open source

Linked data: corpus of open data



Crowdsourcing

Publish questions ☛ Internauts provide answers

Mechanical Turk of Amazon

- Reference to "The Turk," a chess-playing automaton of the 18th century

Foldit: decoding the structure of an enzyme close to the AIDS virus

- Understand how the enzyme folds in a 3D space
- Game

Open problems

Statistical analysis

- On large volume of data and users
- Need to verify information, evaluate its quality, resolve contradictions

Lack of explanation

- Systems are bad at explaining choices resulting from complex computations

Confidentiality issues

- Conflict between the user who wants to protect its confidential data and systems that want this data to provide better and more personalized services (in the best case)

But of the tree of the knowledge of good and evil, thou shalt not eat of it: for in the day that thou eatest thereof thou shalt surely die.

Genesis 2:17

- Introduction
- Two achievements for the 20th century
    - Relational systems
    - Web search engines
- Two challenges of the 21st century
    - Networks and collective knowledge
    - **The web of knowledge ←**
- Conclusion

From text to knowledge

The web of text is based on the fact that people like to read, write, speak, listen to text

Machines understand better more formatted **knowledge**

Text	Knowledge
Je suis presque certain que Bob est amoureux d'Alice	Likes(Bob, Alice, 95%)

Semantic web

Add semantic annotations to specify the meaning of documents on the web

Example

author= Serge Abiteboul ; title = Sciences des données

nature = leçon inaugurale ; date = Mars 2012 ; language = français

Inside a document

Woody Allen <*dbpedia:Woody_Allen*> was at Cannes <*geo:city_France*> for the kickoff of …

Knowledge bases such as dbpedia are called **ontologies**

Ontologies

Logical sentences such as:

- ***classes*** *sa:Person, sa:Director, sa:Film*
- *sa:Director* ***subclass of*** *sa:Person*
- *Sa:Movie* ***synonymous of*** *sa:Film*
- *sa:Woody_Allen* ***is a*** *sa:Director*
- ***relation*** *sa:directed*
- *sa:Woody_Allen sa:directed sa:movie_Manhattan*

What is this useful for?

- **To answer** queries more precisely
- **To integrate** data from several data sources and eventually integrate all the data of the web

Problem: knowledge acquisition

Internauts

- like to publish on the web in their natural languages
- do not appreciate the constraints of a knowledge editor
- want to keep their visibility

Knowledge will often be generated automatically

- Search for syntactic forms such as
- Napoléon died in Sainte-Hélène

Construction of large knowledge bases

- Complex
- Language understanding
- The web is full of inaccuracies and errors

Problem: distributed reasoning

Using facts such as

Psycho is a Hitchcock movie and Alice saw Psycho

And rules such as

WantsToSee(Alice, *t*) ← Film(*t*, Hitchcock, *a*), *not* Seen(Alice, *t*)

We can **infer "intentional"** facts such as

Alice would like to see the movie Psycho

Answering a query becomes more complicated

- Infer new facts but avoid inferring all that may be inferred – too costly
- Collaboration between systems that possess and infer facts

Totally new context

- We are surrounded by systems that possess and infer facts
- Modifies the way we think

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

Introduction

Two achievements for the 20th century

Relational systems

Web search engines

Two challenges of the 21st century

Networks and collective knowledge

The web of knowledge

Conclusion 🡐

The web is multiform

Industry, health, culture, government, science, ecology …

Unavoidable

- To find work, to work, to find an apartment, to manage bank accounts, to be part of an association, to have friends…

Hosting all human knowledge

- The most horrific fantasies, all the violence
- Inaccuracies, errors

The web is multiform

Outdated view 1: hypertext

Outdated view 2: universal library

And the other webs

- Smartphone web
- Social network web
- Semantic web
- Communicating objects web and ambient intelligence
- Virtual world web (3D games)
  - …

risks
dangers
traps
excess
pitfalls

The pitfalls of the web

Avoid drowning in an ocean of data

This has been one the threads of this presentation

Access to information for all

- Social divide

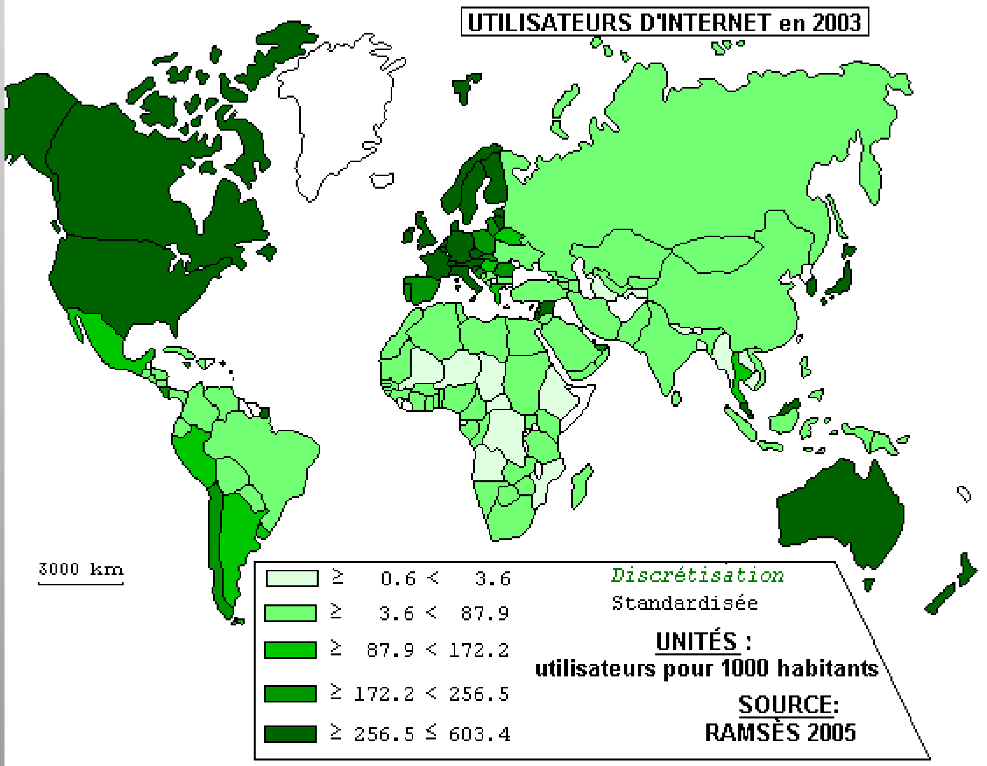
  CREDOC 2009: in France,

  40% of the population never uses computers

- North/south

- Teaching of

  computer science



The pitfalls of the web (end)

Democracy or not?

And private life?

For better or worse people?

I want to continue believing that the web will contribute to building a better future

And tomorrow…

Political choices

New software tools that are yet to be invented

And for scientific aspects

The next stage of data sciences has already started:

The web of knowledge

From data
to information,
to knowledge…

Acknowledgements: 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 and Victor Vianu