# From data and information to knowledge: the Web of tomorrow

Serge Abiteboul
INRIA & ENS Cachan
Académie des sciences



INSTITUT DE FRANCE
Académie des sciences

# Organization

- The Web today
- The Web tomorrow
  - The scale
  - The new frontier: knowledge
  - How is knowledge acquired?
  - Reasoning with distributed knowledge
  - Other issues
- Conclusion

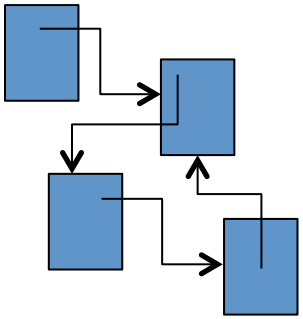
# The Web today

Graph theory

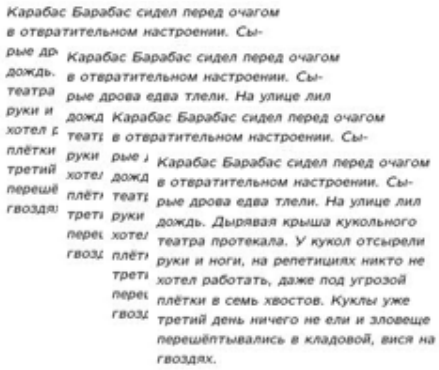
Network of machines (Internet)

Network of content (Web)

Network of people (social networks)



hypertext




universal library of text

and multimedia

personal/private data

social data

# Data & information

Originally, computers were seen as systems to compute

- Solving differential equations
- Cryptography
- Etc.

They are primarily used to store/manage **data/information**

- Accounting, banking (transactions)
- Inventory, catalogue
- Agenda, contacts
- Library, etc.
- Sciences: measurements, simulation

# What has changed recently?

- The technology
  - Larger/faster disks, memories
  - Faster networks/processors
  - Algorithmic progress, e.g., parallel computing
- The scale
  - Size of data
  - Number of machines
  - Number of users
- The Web
  - Information was living in islands with different formats, application programming languages, operating systems
  - The web brought universal standards for sharing information

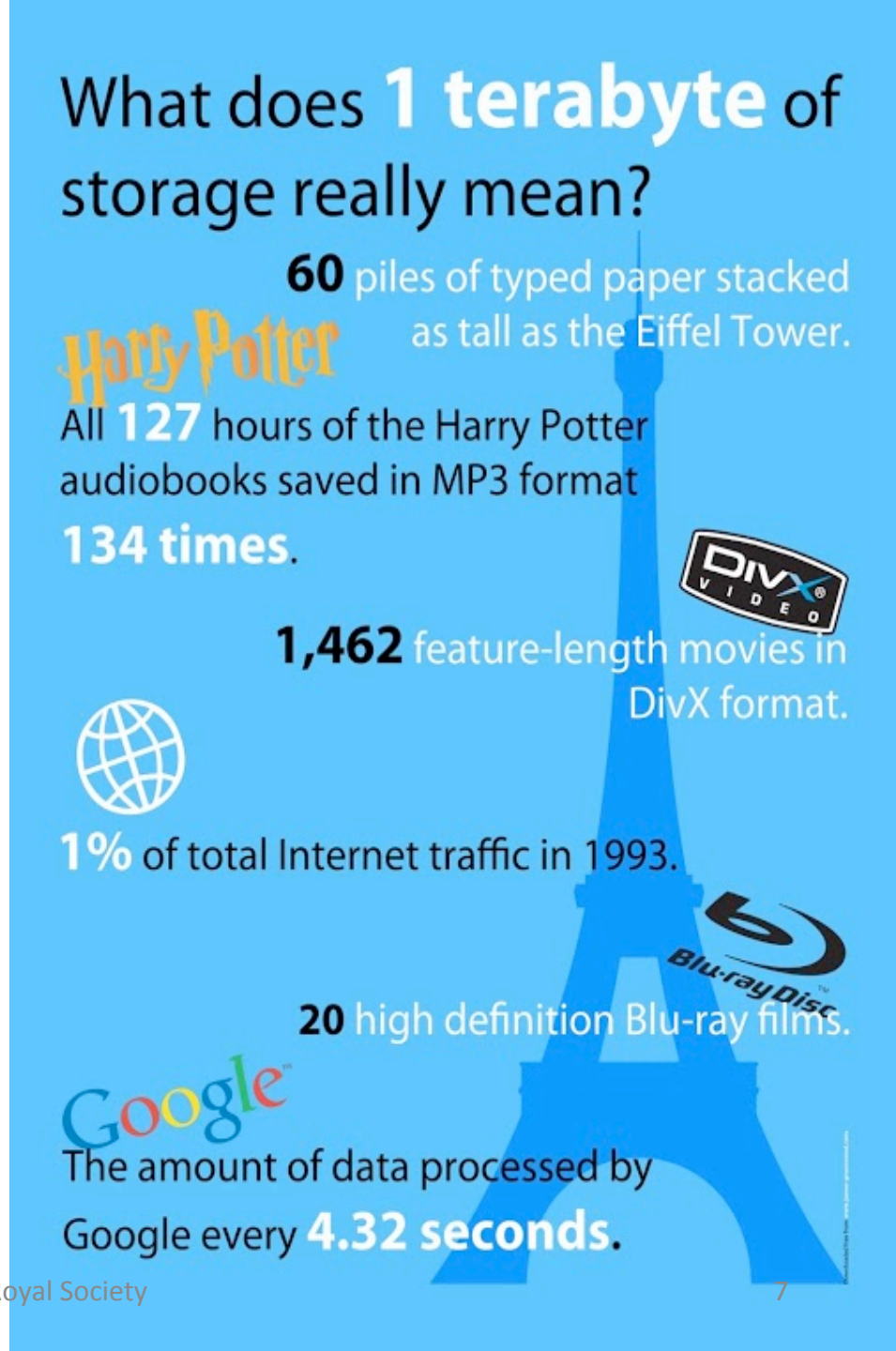
# The digital world

Hundreds of millions of web sites

Billions of communicating objects

Thousands of billions of pages

**The size of the digital world is doubling each 18 months**



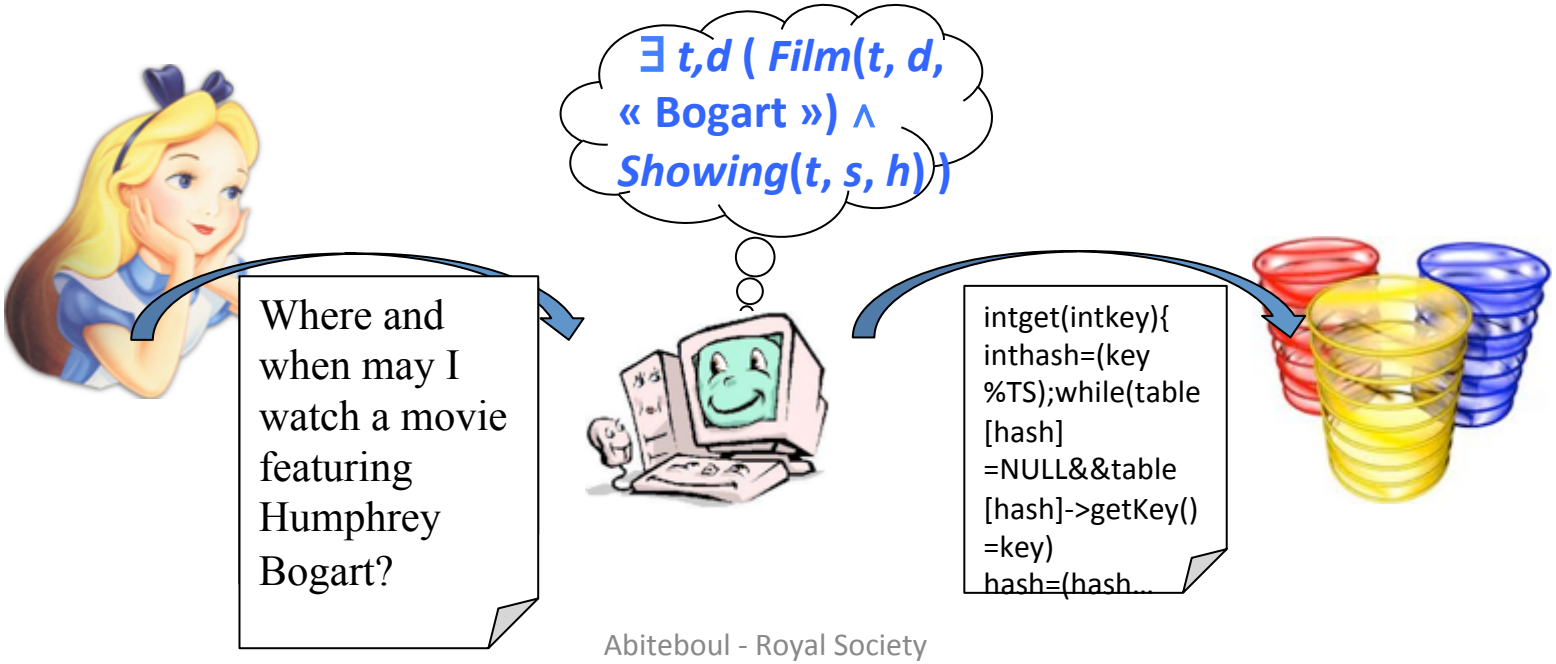
# Two main achievements of computer science of the 20th century

- **Data**: database systems
- **Information**: Web search engines

# **Data**: database systems

A **database system** plays the role of a **mediator** between an intelligent user and objects storing information

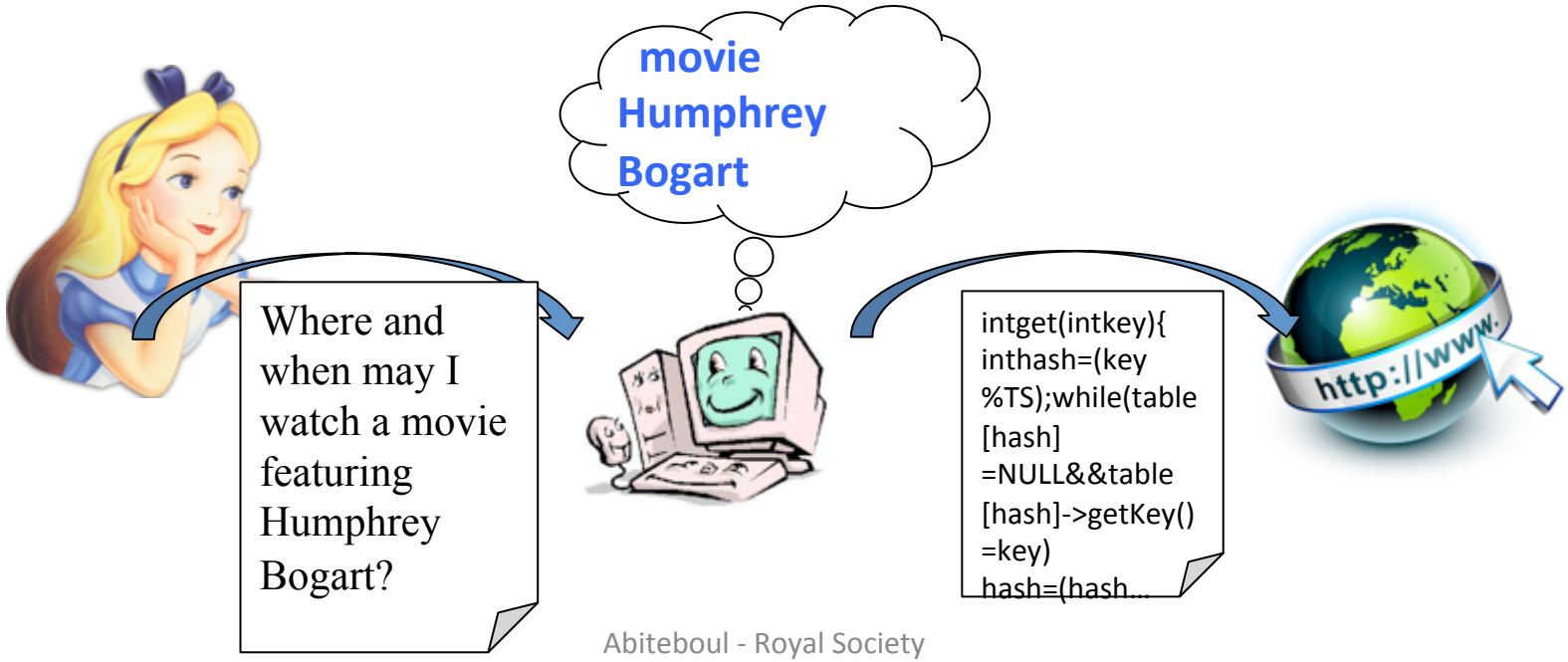
Huge impact on our lives



# **Information**: Web search engine

A **Web search engine** plays the role of a **mediator** between an intelligent user and objects storing information

Huge impact on our lives



# The secret of success

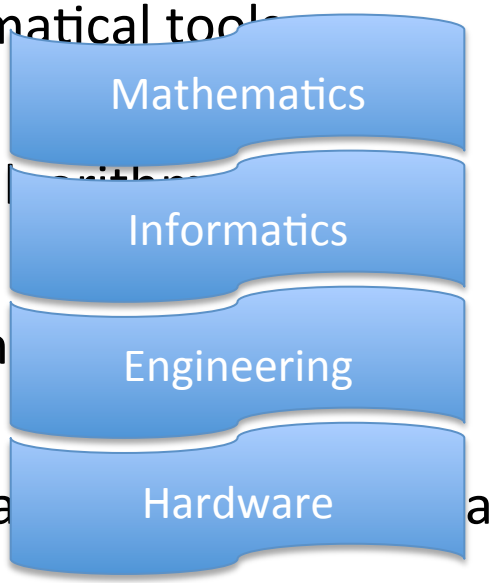
The improvement of a functionality or a new functionality

Mathematical tools

Smart algorithms

Solid engineering

Taking advantage of hardware




Better interfaces
Better ranking of pages

Logic & algebra
Probabilities & matrix multiplications

Query optimization
Parallel algorithms

Failure recovery
Clusters of thousands of machines

Disk capacity
Huge and cheap memories

# The Web tomorrow

How can the Web evolve to best serve you?


Garbage in Garbage out

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

# Assumption

1. The size will continue to grow
2. The information will continue to be managed by many systems
   - Vs. a company will conquer all the information of the world
3. These systems will be intelligent
   - In the sense that they produce/consume knowledge and not simply raw data
4. These systems will be willing to exchange knowledge and collaborate to serve you
   - Vs. capture you and keep you within islands of proprietary knowledge

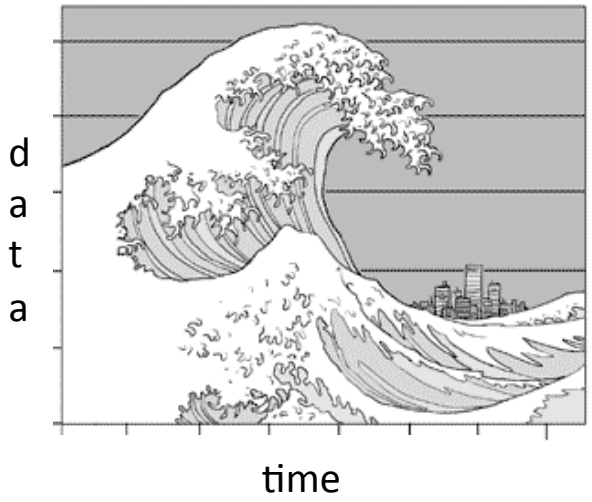
# The difficulties: The 4+1 Vs of Big data

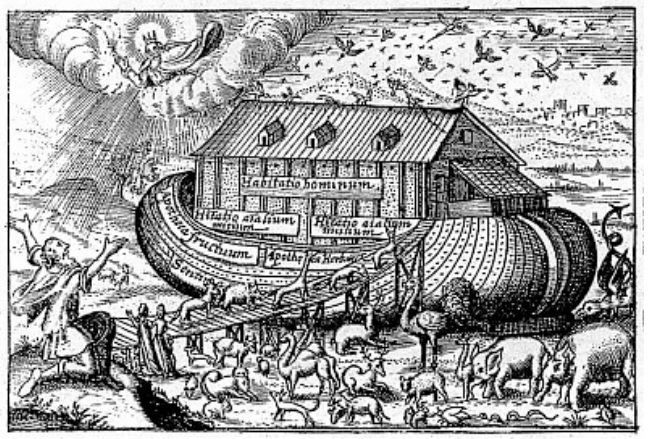
- **Volume**
  - Mass of data – more than can be stored/retrieved
- **Velocity**
  - Data streams & frequent changes
- **Variety**
  - Heterogeneity of structure, schema, ontologies…
  - Distribution of sources
- **Veracity**
  - Uncertainty, errors, imprecision, contradictions…
  - Lack of quality
  - Opinions, sentiments, and lies.
- **Value**

# The scale

# Perhaps the main problem : surviving the deluge

- More and more information available
- **An issue for computer systems is to select the information the user wants to see**
  - Based on importance, quality, personal interest…

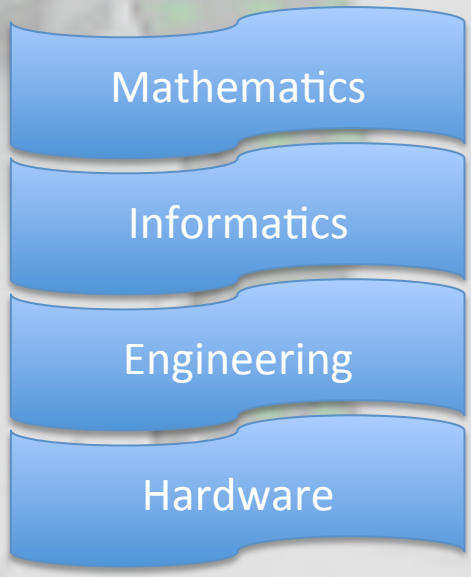


# A technical challenge
# very large scale data/information analysis

- An old problem since the early days of computer science
- Data mining, business intelligence, **big data**…

- Statistics
- Complex algorithms



*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

# The new frontier: knowledge

# 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* |

# Ontologies

Knowledge representation: logical sentences such as

***The royal society is_a An academy***

***The royal society is_a Scientific organization***

***The royal society is_a National institution***

***The royal society is_located London, United Kingdom***

***The royal society founded 1660***

***L'académie des sciences founded 1666***

***…***

A collection of such statements is called an **ontology**

# Machines prefer formatted knowledge

Machines have difficulties with **human languages**

Machines are better at handling more **formatted knowledge**

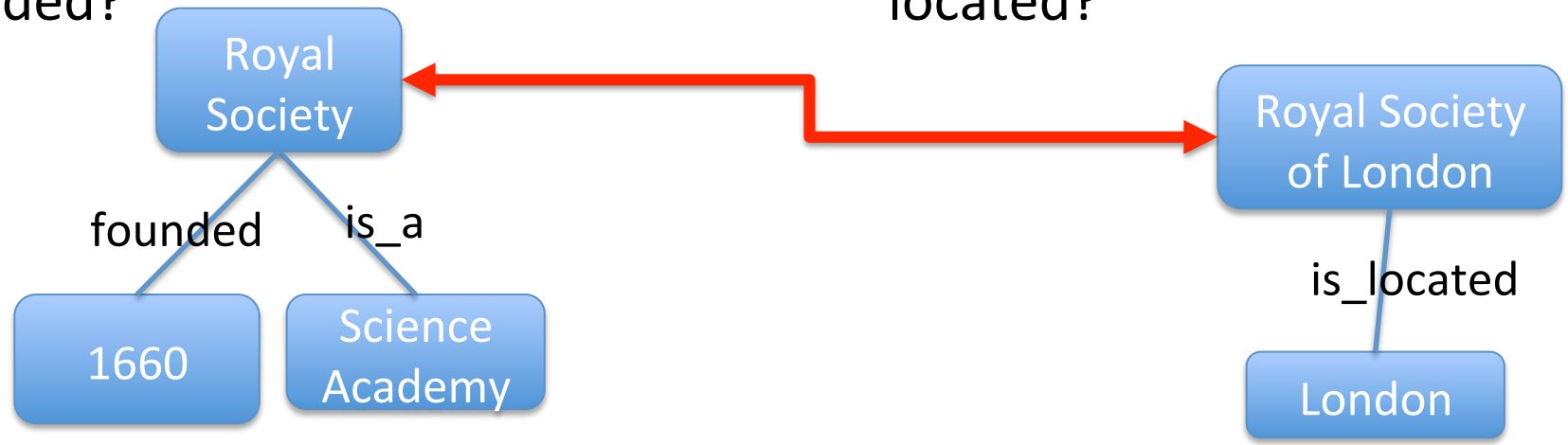
| Text (from Wikipedia) | Knowledge |
|---|---|
| The Royal Society of London for Improving Natural Knowledge, commonly known as the Royal Society, is a learned society for science, and is possibly the oldest such society still in existence… | *is_a(The royal society, Scientific organization)*<br>*founded(The royal society, 1660)*<br>*∀x,y ( founded(x,y) ∧ is_a(x, Scientific organization) ⇒ y≥1660 )* |

# What are ontologies useful for?

**To answer** queries more precisely

When was the Royal Society founded?

Where is the Royal Society located?




**To integrate** data from several sources

When was the science academy located in London founded?

# How is knowledge obtained?

# How is knowledge obtained?

- Humans specify explicitly the knowledge
  - Standard in science, much less in “real-world”
- But
  - People like to talk/write in their natural languages
  - People like to publish on the web: web pages, blogs, tweets, votes, recommendations, opinions…
  - They do not appreciate the constraints of a knowledge editor
  - They want to keep their visibility
- Machines will have to obtain knowledge

# Automatic and collective acquisition of knowledge

(*) knowledge = formal/digital knowledge EARLIER

- Extraction from text
- Web graph analysis
- Collaboration
- User's opinion
- Recommendation
- Reasoning and inference

Main issue: quality

# Knowledge extraction from text

Construction of large knowledge bases

- Complex linguistic problem with inaccuracies and errors
- E.g.: Yago (from Wikipedia) at Max Plank Institute

Search for syntactic patterns such as

- Woody Allen is married to Soon-Yi Previn
- Woody Allen got hooked with Soon-Yi Previn

**You may think that the problem is to find such sentences**

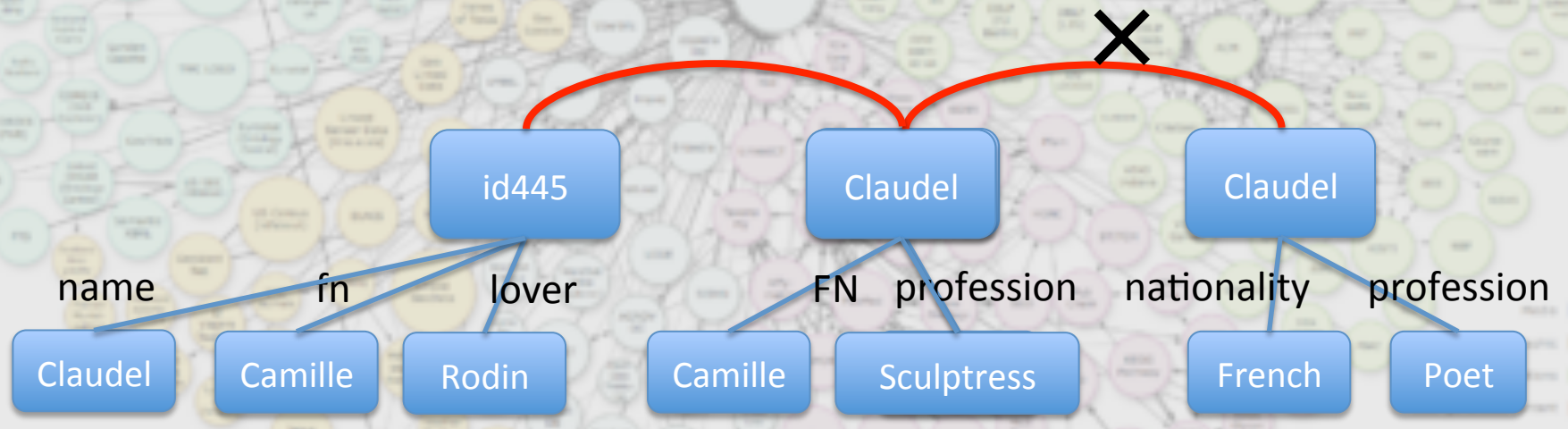
**Wrong:** The problem is to keep a balance between

precision: fraction of results that are correct

recall: fraction of correct results that are retrieved

# Aligning ontologies

- Manually – see Linked data further
- Automatically
  - E.g.: Paris system at Inria & Telecom ParisTech



# Knowledge acquisition: collaboration

Internauts perform collectively tasks they cannot solve individually

- No payment & little management

**Wikipedia: encyclopedia**

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

**You probably heard that this is the work of amateurs and thus that it cannot be correct**

**Wrong:** the main issue is the stronger presence of professionals with personal agendas

# Collaboration (end)

**Linked data project**

- Publish ontologies : Publish facts/relations and links between them
  - (2011) 31 billion facts/relations;
- Align them with other ontologies: Publish links
  - (2011) 504 million links

**Open source software** E.g., Linux: operating system

- **You may think that the Internet and the Web are major successes of the industry notably US**
- **Wrong:** They are mostly running on open-source code

**Crowdsourcing** Publish questions ☛ Internauts answer

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

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

# Opinions & recommendations

**Learn the opinions** of the Internauts

- Quantitative (***)
- Qualitative ("dating spot")

Increasingly standard

- Movies in eBay
- Lodging in Airbnb
- Food in TripAdvisor…

Users are taking over the role of specialists - spamming

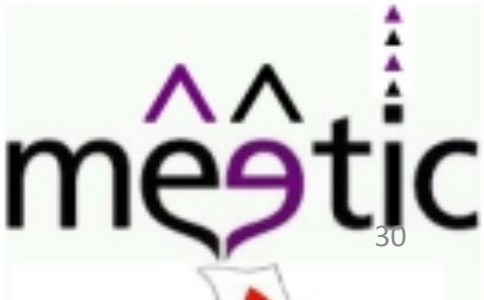
From users opinions/purchases, **derive recommendations**

- 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



# Opinions: voting with Web graph analysis

**You were perhaps told that a search engine is great because of the amount of information it indexes**

**Wrong:** indexing billions of Web pages is simple compared to selecting the links that appear on the first page of answers, which has become a business and even political issue

- Based on the opinions of Web users

# Reasoning with distributed knowledge

# Inferring knowledge

Using facts such as

Psycho is a Hitchcock movie

And rules such as

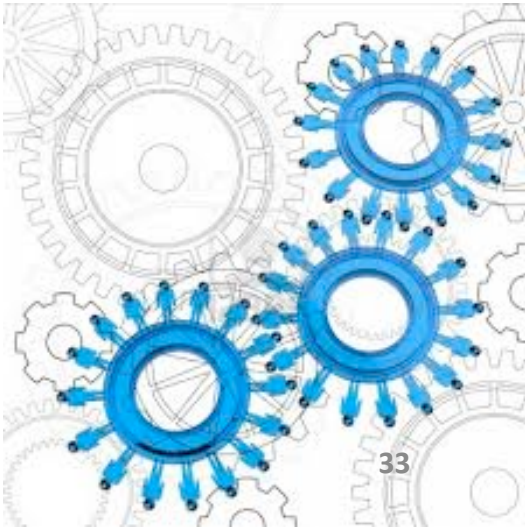
WantsToSee( Alice, *t* ) ← Film( *t*, Hitchcock, *a* ), *not* Seen( Alice, *t* )

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

Alice would like to see the movie Psycho

Inference is complicated

- Facts on the Web are uncertain: probabilities
- Facts on the Web may contain contradictions
- Scale
  - Avoid inferring all that may be inferred – too costly



# Where is the truth?

- **Is everything true?**
  - People rarely publish that something is false, e.g., "Elvis was not French"
    - There are too many false statements to state
  - One can use inference: "Elvis was American"; so it is likely that "Elvis was not French"
- We have to deal with contradictions
  - Elvis Presley died on August 16, 1977 & The King is alive
  - Using voting, a machine may derive that "it is very likely that Elvis died"
- Difficulties
  - Reason with inconsistencies ⇒ problems with logic
  - Truth becomes uncertain ⇒ measure uncertainty with probabilities
  - There may be different truths (viewpoints)


# Where is the truth? (end)


Illustration: Recent work on corroboration

To determine the true/false facts:

- Use voting, get an estimate of the "truth value" of facts
- Based on that, determine the "quality" of the data sources
- Using this quality, get a better estimate of the "truth value" of facts
- Then, get a better estimate of the "quality" of data sources…

A human learns the difference between a newspaper and a tabloid

On the Web, too many sources of information – machines will have to separate the wheat from the chaff


In a classical database: everything that is not in the database is assumed to be false – ***closed world*** *assumption*

On the Web, if you don't know something, it may be out there… Or not – ***open world*** *assumption*

Difficulties

- We cannot bring all the Web locally
- We cannot visit all the Web
- How do I know where to find something?
- How do I know it is not out there?

# Explaining

- **Users want to understand the information they see, the answers they are given**

  In their professional life and

  in their social life as well

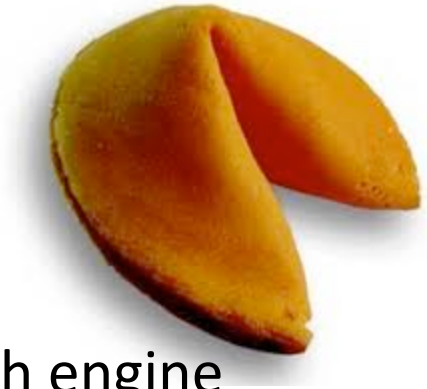
  - E.g., Alice has a new boyfriend – How was this determined?
- Analogy: in sciences, knowing the result of an experiment without knowing its conditions is often useless
- Difficulties
  - Reasoning with large number of facts
  - Information is often probabilistic and not public
  - Requires knowing how the information was obtained (its *provenance*)

# Other issues

# Privacy

- Users want to control their data
- They have **personal data**
  - They want to share with their friends
  - They are willing to give to systems to get more personalized services
- The systems want to get their personal data
  - For personalized ads
  - To sell to other business
- **How to do we that?**
  - Better systems
  - Better laws
  - Better users (digital literacy)

# Serendipity



- You may hear by chance a song that is going to totally obsess you
- A librarian may suggest your reading an article that will transform your research

This is serendipity

- A perfect search engine
- A perfect recommendation system
- A perfect computer assistant

Such systems are boring

They lack serendipity

Design programs that would **introduce serendipity** in our lives

- Random algorithms & sentiment analysis

# Hypermnesia

*Exceptionally exact or vivid memory, especially as associated with certain mental illnesses*

For a user: We cannot live knowing that any word, any move will leave a trace?

For the ecosystem: We cannot store all the data we produce – lack of storage resources

**Forgetting is Key to a Healthy Mind**
*Scientific American*
Image: Aaron Goodman

- **A main issue is to select the information we choose to keep**

# The Babel of human-machine-interaction

- Each time a user interacts with a data source, does he have to use the ontology of that source ?
  - How the source organizes information
  - The particular terminology it uses
  - Its sequencing of tasks, its codes…
- No!
- Instead of a user adapting to the ontologies of the N systems he uses each day
- **We want the N systems to adapt to the user's ontology**

# Religion…science…machines

- Knowledge used to be determined by religion
- Knowledge determined scientifically
- **Knowledge determined by machines**
  - Match making
  - Medical diagnosis…
- Decisions are increasingly made by machines
  - Stock market (automatic trading)
  - Fully automated factory
  - Fully automated metros
  - Death penalty (killer drones)…

# Conclusion


# to the digital world!

- The massive use of digital information has modified in depth all facets of our life : work, science, education, health, politics, etc.
- We will soon be living in a world surrounded by machines that
  - acquire knowledge for us
  - remember knowledge for us
  - reason for us
  - communicate with other machines at a level unthinkable before
- What will we do with that technology?
- Will we become smarter ?
- Will we become master or slave of the new technology?

# How can we get prepared to these changes?

Informatics and digital humanities are at the cross road of these questions

**Learn informatics**

- To understand the digital world you are living in
- To decide your life in the digital world
- To participate in/contribute to the digital world


informatiques mathématiques

CACHAN

INSTITUT DE FRANCE
Académie des sciences