La cuisine des données du Web

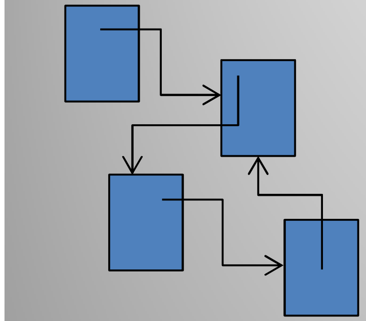
Serge Abiteboul

INRIA Saclay & ENS Cachan

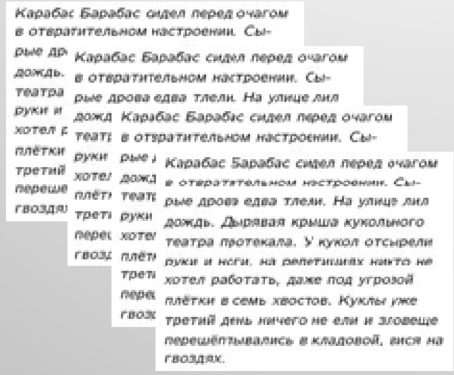
Network of machines (Internet)

Network of content (Web)

Then came the network of people



hypertext



Карабас Барабас сидел перед очагом в отвратительном настроении. Сырые дрова едва тлели. На улице лил дождь. Дырявая крыша кукольного театра протекала. У кукол отсырели руки и ноги, на репетициях никто не хотел работать, даже под угрозой плётки в семь хвостов. Куклы уже третий день ничего не ели и зловеще перешёптывались в кладовой, вися на гвоздях.

universal library of text

and multimedia

personal/private data

social data

What has changed

- The scale
- The encounter between humans and machines
  - Opinions vs. facts
  - Beliefs
  - Trust
- The imprecision
  - Missing information (open world)
  - Imprecision & probabilities
  - Errors & contradictions

Acquiring knowledge

Wide variety of approaches of collectively acquiring knowledge on the Web

(*) knowledge = formal/numerical knowledge

- Web graph analysis
- Collaboration
- Recommendation
- Web scale knowledge extraction

- Main issue: Evaluation of the quality

Web graph analysis

Skill and magic of Web search engines

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

Wrong: The more information, the more complicated it is to find the right information; what matters is how to choose between the results

The skill: indexing billions of pages

- Using techniques such as hashing

The magic: finding what you want (in general)

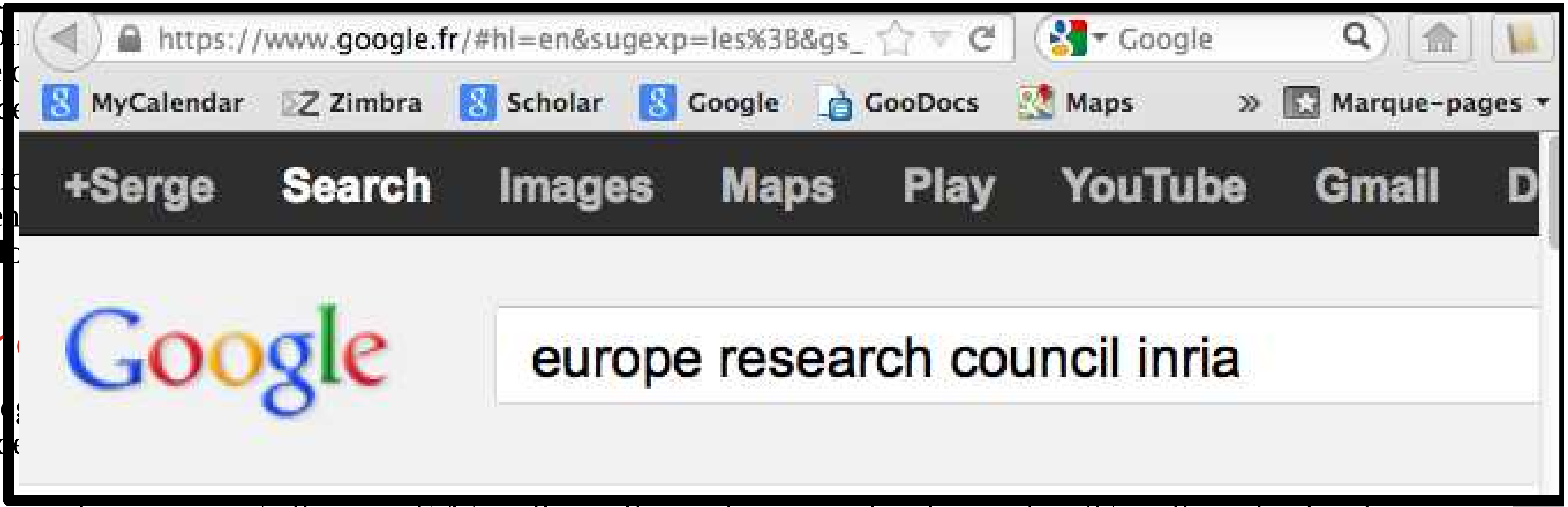
- Using "measures" to rank pages such as TFIDF and
- PageRank: mathematically-based popularity measure

Le progr
explorato
nationale
au cœur d

La sélecti
L'excellen
méthodol

Comm

Deux cat
création d



europe research council inria

Avec une enveloppe pouvant aller jusqu'à 1,5 millions d'euros (« jeunes chercheurs ») ou 2,5 millions (« chercheurs confirmés »), les heureux élus ont les moyens de recruter l'équipe de leur choix et de mettre en oeuvre les moyens nécessaires pour mener à bien leur projet.

Entretien avec Jean-Pierre Banâtre

Jean-Pierre Banâtre est professeur émérite à l'Université de Rennes 1 et conseiller auprès de la direction de l'Institut pour le programme ERC.

Quelle est la spécificité de ce programme européen ?

J.P. Banâtre : Le programme ERC est consacré à la recherche fondamentale. Au fil du temps, ce programme inédit qui encourage des projets à risque a pris une place de plus en plus importante. Il a permis de donner leur chance à des chercheurs venus du monde entier désireux de poursuivre leurs travaux en Europe. Je gage que, bientôt, les instituts feront figurer le nombre de lauréats ERC dans leurs indicateurs.

Quelles sont les caractéristiques d'un bon projet pour l'ERC ?

J.P. Banâtre : Un bon dossier est d'abord porté par un leader scientifique déjà reconnu, ou très prometteur (pour les

Collaboration

Example: Wikipedia

Internauts perform collectively tasks they cannot solve individually

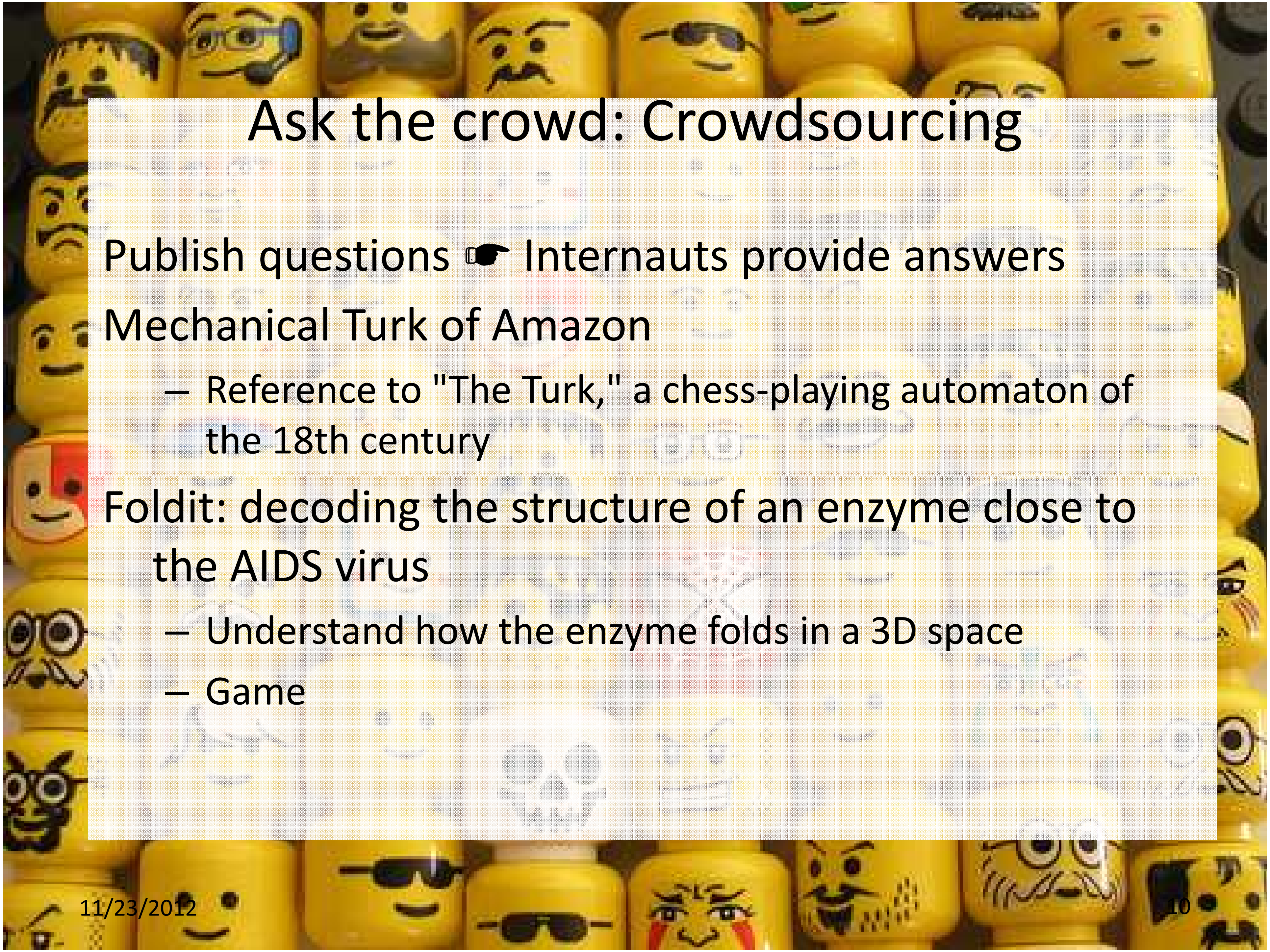
Wikipedia: encyclopedia

- 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

Other examples: open-source software (Linux), open data



Ask the crowd: 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

Crowdsourcing experiment

Which of these statements are true?

1. JPB has been a school teacher
2. JPB has been a fireman
3. JPB has had Yves Cochet as teaching assistant
4. JPB has been the companion of Carla Bruni

Recommendation

Recommendation

Use web data for deriving recommendations

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

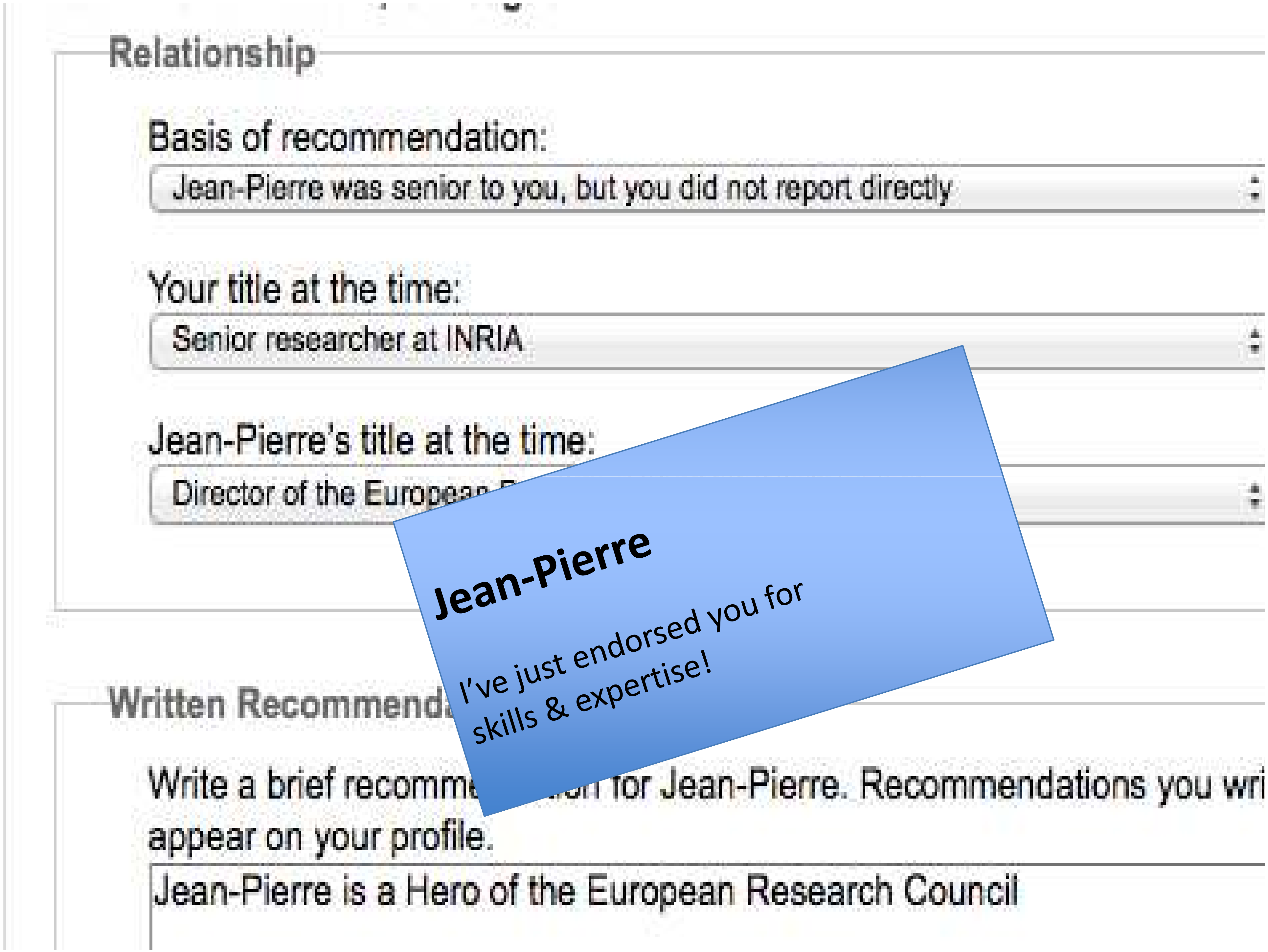
Statistical analysis to discover “proximities”

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

Stop emailing

Experiment with Meetic?

No with Linkedin



Relationship

Basis of recommendation:

Jean-Pierre was senior to you, but you did not report directly

Your title at the time:

Senior researcher at INRIA

Jean-Pierre's title at the time:

Director of the European

Written Recommenda

Write a brief recommendation for Jean-Pierre. Recommendations you wri appear on your profile.

Jean-Pierre is a Hero of the European Research Council

Jean-Pierre

I've just endorsed you for skills & expertise!

Issues

Statistical analysis on large volume of data & number of users

- Need to verify information, evaluate its quality, resolve contradictions

Lack of explanation

Systems are bad at explaining choices

Lack of serendipity

Quickly boring?

Lack of privacy

But user likes personalization

Web scale knowledge extraction

Ontologies

Basis of knowledge: logical sentences such as

sa:Jean-Pierre_Banatre yago:wrote "Generalized multisets for chemical programming"

sa:Jean-Pierre_Banatre yago:profession yago:chimiste

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

What are ontologies useful for?

- **To answer** queries more precisely
- **To integrate** data from several data sources

Illustration [work of Suchanek]:

1. Yago: a system developed at MPI to extract knowledge from Wikipedia
2. Paris: a system developed at INRIA to align two ontologies

A lot of knowledge is present in texts

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

Machines understand better more formatted **knowledge**

Text	Knowledge
In 2008, JPB has called me twenty times to convince me to submit a stupid ERC proposal.	responsabilité(2008, JPB, Chargé des affaires Européennes, INRIA)

Main issue: evaluation

1. Quality of the data
2. Quality of the source

Issue: is everything true?

People on the Web rarely publish that something is wrong

- There are too many wrong statements

A fact may contradict some known facts

- JPB ***is not*** born in Cancale (because some sites say he is born in Saint Malo and people are born in a single place)

Closed world sometimes exists

- JPB has not been a companion of Carla Bruni because he does not appear in any list of her companions found on the Web

EXPERIMENT: stop email and publish a new such list with JPB in it

Corroboration

When two facts are contradicting, use voting

- Count how many sites say New York is the capital of US and how many say it is Washington

Can we do better?

Yes we can by learning about the expertise of sites

Use this to evaluate the quality of sources

Get a better estimate of the truth value of facts; loop…

Today: personal evaluation of a source of information

Tomorrow: will reputation be determined by programs ?

Conclusion

Let’s imagine the future

The Web will turn into a distributed knowledge base with billions of users supported by billions of systems analyzing information, extracting knowledge, exchanging knowledge, inferring knowledge

From closed-world and precise to open-world and imprecise

We will soon be living in a world

- surrounded by machines that acquire knowledge for us, remember knowledge for us, reason for us
- communicating with others at a level unthinkable before

Main issues: choosing, filtering…

- How do we find information/knowledge?
  - To take advantage of the available resources
  - Quality evaluation is a key issue
- How do we choose among all the knowledge that can be obtained? What is of interest ?
  - Of course when the user asks a query
  - Notifications & serendipity

Other issues

- How do we accept some particular knowledge?
  - Need for explanations
- How do we keep control over our own data?
  - Protecting our private life
- Will a Web of knowledge move us away from reading text/literature
  - More precise but dry
  - I doubt it…

INRIA

Merci !
FRAPAR.