INFORMATION RETRIEVAL & INFORMATION MINING

JOSIANE MOTHE

ASD – 2014 – HAMMAMET - TUNISIE

AND BIG DATA

Information Retrieval
Document indexing
Image indexing, …
Information access

Data mining
Multi-dimensional analysis
OLAP
Data Warehouse
Classification
Learning

Big data
Volume
Variety
Velocity

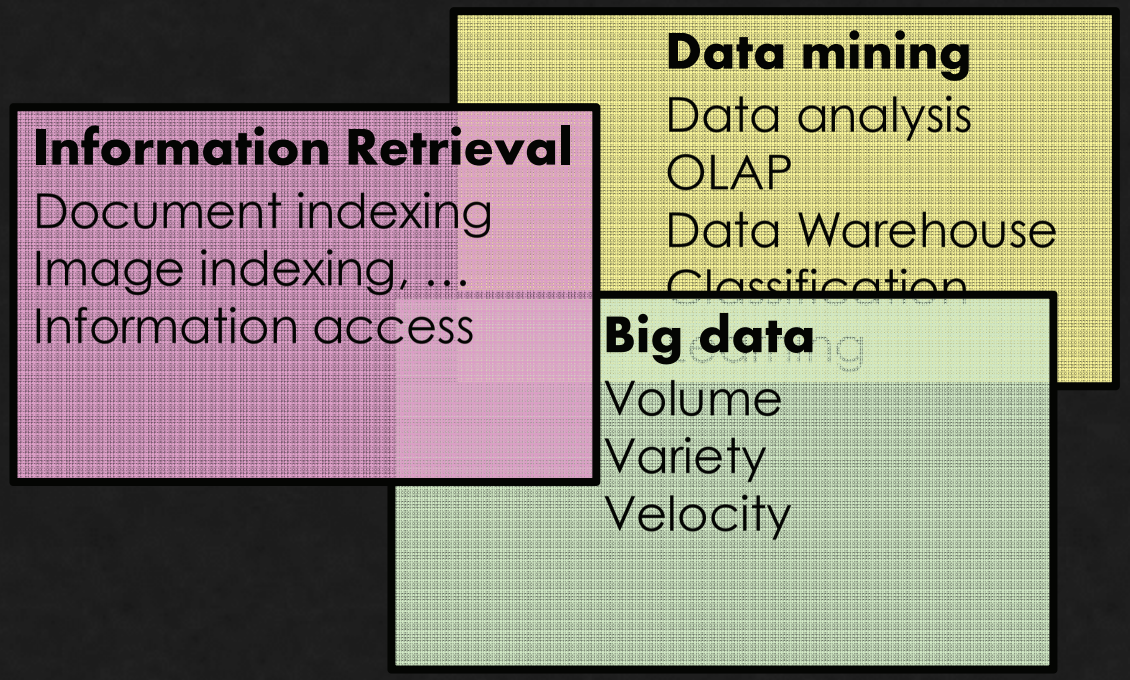
AND BIG DATA

Information Retrieval
Document indexing
Image indexing, …
Information access

Data mining
Multi-dimensional analysis
OLAP
Data Warehouse
Classification
Learning

Big data
Volume
Variety
Velocity

AND BIG DATA


Information Retrieval
Document indexing
Image indexing, …
Information access
Data mining
Data analysis
OLAP
Data Warehouse
Classification
Learning
Big data
Volume
Variety
Velocity

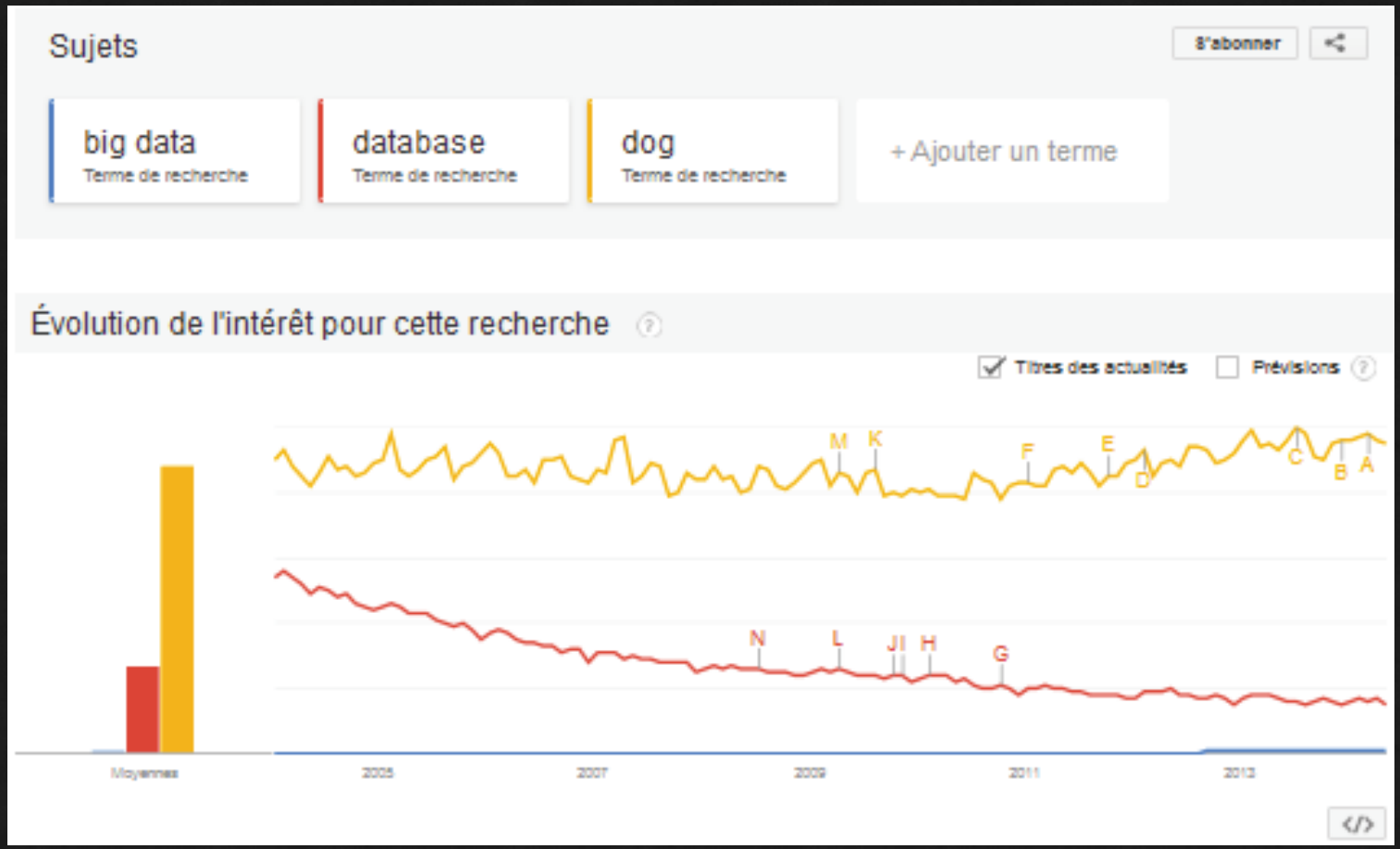
BIG DATA

Sujets
S'abonner
big data
Terme de recherche
+ Ajouter un terme
Évolution de l'intérêt pour cette recherche
Titres des actualités
Prévisions
G
F
E
D
C
B
A
2005
2007
2009
2011
2013

BIG DATA

Sujets
S'abonner
big data
Terme de recherche
database
Terme de recherche
+ Ajouter un terme
Évolution de l'intérêt pour cette recherche
Titres des actualités
Prévisions
J
I
HG
F
E
C
B
A
D
Moyennes
2005
2007
2009
2011
2013

BIG DATA


Sujets
S'abonner
big data
Terme de recherche
database
Terme de recherche
dog
Terme de recherche
+ Ajouter un terme
Évolution de l'intérêt pour cette recherche
Titres des actualités
Prévisions
M
K
F
E
D
C
B
A
N
L
JI
H
G
Moyennes
2005
2007
2009
2011
2013

BIG DATA

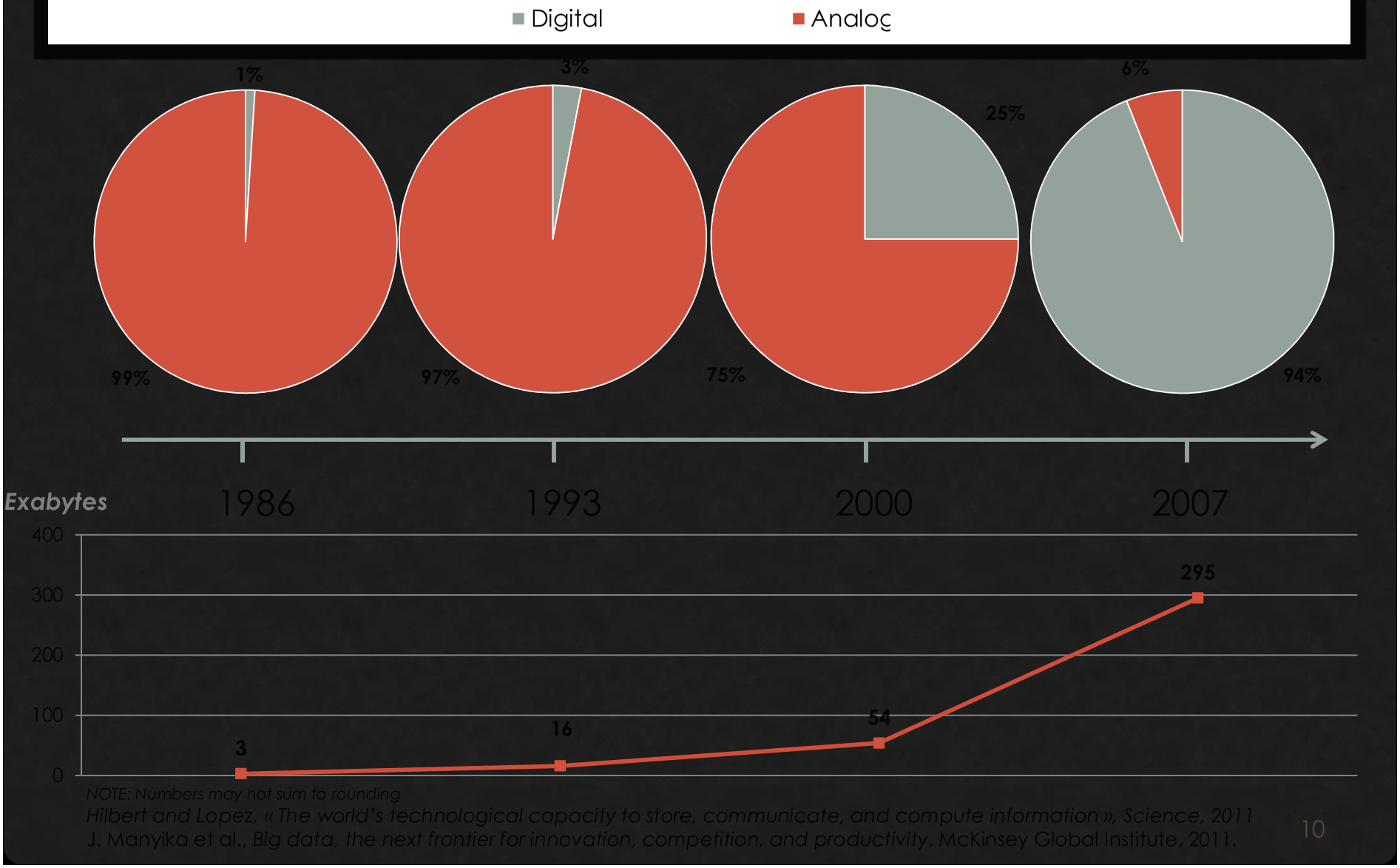
- « Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze » McKinsey 2011.
- « big data objects » : items too large to be handled using the current software and hardware
  « big data collections »: consist in small and big data objects, mostly heterogeneous (Cox et Ellsworth, 1997).
- Data too large *or* too fast *or* too heterogeneous to make ti possible to store directly in data bases or to treat it by current algorithms

Manyika J. et al., (2011). Big data, the next frontier for innovation, competition, and productivity, McKinsey Global Institute.
Cox M., Ellsworth D. (1997). Managing Big Data for Scientific Visualization, In ACM SIGGRAPH, Course 4, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management and Time-Critical Design.
http://www.dcs.ed.ac.uk/teaching/cs4/www/visualisation/lectures/98-99/lect16ref.ps.gz

BIG DATA

- Big Data
  - **Volume:** refers to the quantity of information; too large to be fully collected/ stored / analyzed / spread using current hardware and software
  - **Velocity:** refers to the dynamics of data / data is produced in flow and/or has to be treated on real time
  - **Variety:** refers to the heterogeneity of formats, types
  - **Value:** data potentiality, specifically in economic terms
  - **Veracity:** data quality, ethics

DIGITAL DATA



NOTE: Numbers may not sum to rounding

Hilbert and Lopez, « The world's technological capacity to store, communicate, and compute information », Science, 2011

J. Manyika et al., *Big data, the next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.

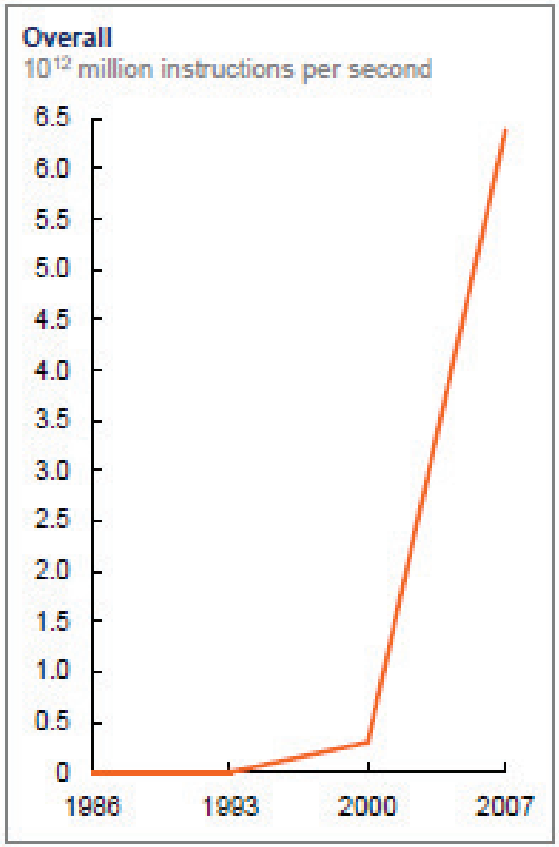
METRIC PREFIXES

Prefix	Symbol	1000^m	10^n	Decimal	English word	
					short scale	long scale
yotta	Y	1000^8	10^{24}	1 000 000 000 000 000 000 000 000	septillion	quadrillion
zetta	Z	1000^7	10^{21}	1 000 000 000 000 000 000 000	sextillion	thousand trillion
exa	E	1000^6	10^{18}	1 000 000 000 000 000 000	quintillion	trillion
peta	P	1000^5	10^{15}	1 000 000 000 000 000	quadrillion	thousand billion
tera	T	1000^4	10^{12}	1 000 000 000 000	trillion	billion
giga	G	1000^3	10^9	1 000 000 000	billion	thousand million
mega	M	1000^2	10^6	1 000 000	million	
kilo	k	1000^1	10^3	1 000	thousand	
hecto	h	$1000^{2/3}$	10^2	100	hundred	
deca	da	$1000^{1/3}$	10^1	10	ten	
		1000^0	10^0	1	one	

COMPUTING

Computation capacity has also risen sharply

Global installed computation to handle information



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

BIG DATA

- Any sector (producer / user) :
  - Sensors (cars, meters, …)
  - Linked to networks (water, roads, telecommunication, internet, …)
  - Internet: web and web 2.0 (social networks, emails, …)
  - Science: medecine, bio-informatics, astronomy, climatology, …
- Part of the global economy: societal and economic issues
- Concepts, techniques and technologies: traditional and new

TYPES

	Video	Image	Audio	Texts/ Numbers
Insurance				
Banking				
Communication and media				
Construction				
Education				
Gouvernement				
Health care				

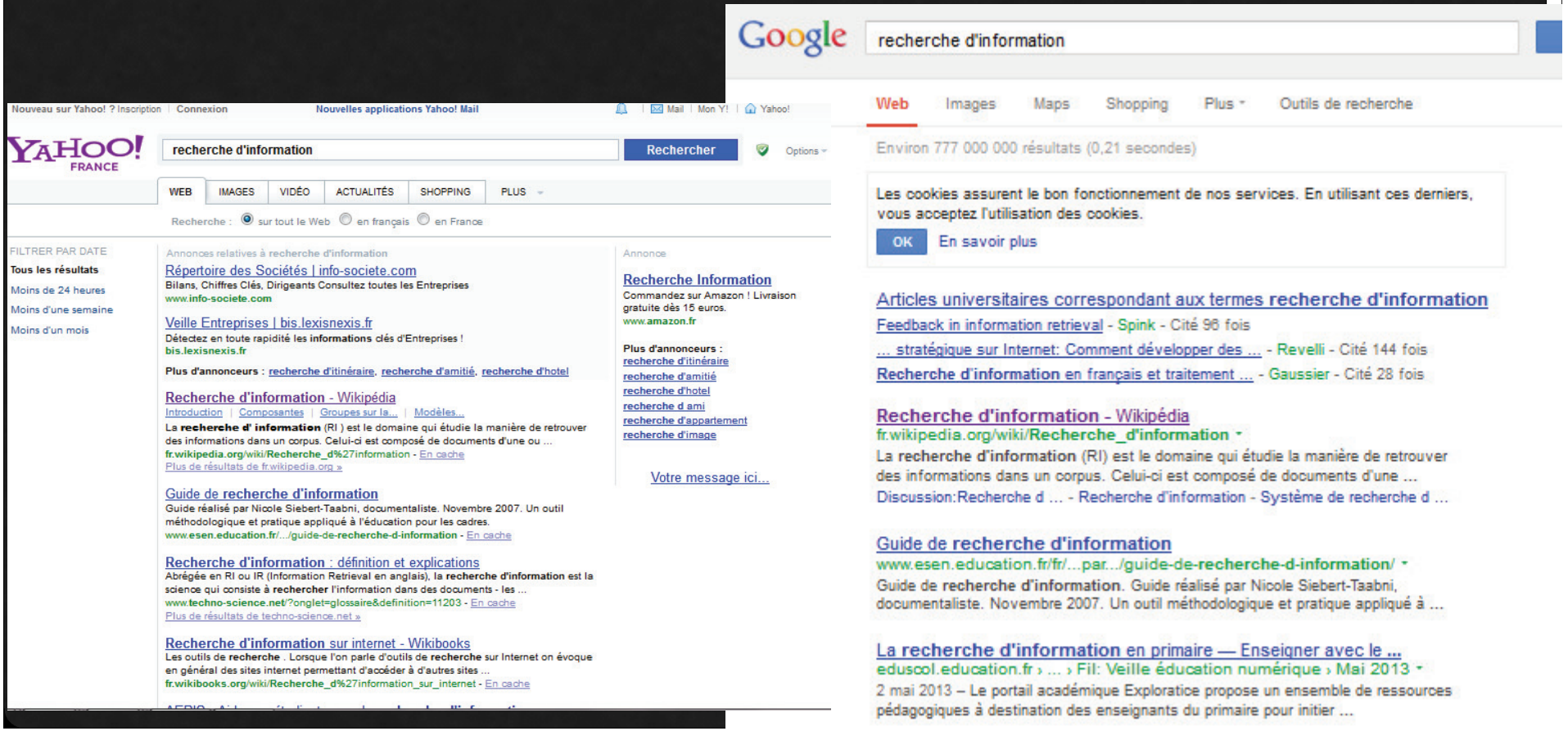
SOURCE: McKinsey Global Institute analysis
J. Manyika et al., *Big data, the next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.

BIG DATA FROM INTERNET /WEB 2.0

http://practicalanalytics.wordpress.com/2012/10/22/sizing-mobile-social-big-data-stats/

INFORMATION RETRIEVAL

- Information access : retrieve relevant documents according to a user's information need (query)



INFORMATION RETRIEVAL

- 1789
- Classifications
  - 1876 DEWEY (CDD) / 1905 CDU

classe 0 généralités
classe 1 philosophie religion
classe 2 religion
classe 3 sciences sociales
classe 4 provisoirement inoc
classe 5 sciences exactes et
classe 6 sciences appliquées
classe 7 arts-divertissement-
classe 8 littérature-linguistiq
classe 9 archéologie-géologi

Classe 6 - Sciences appliquées. Médecine. Technologie

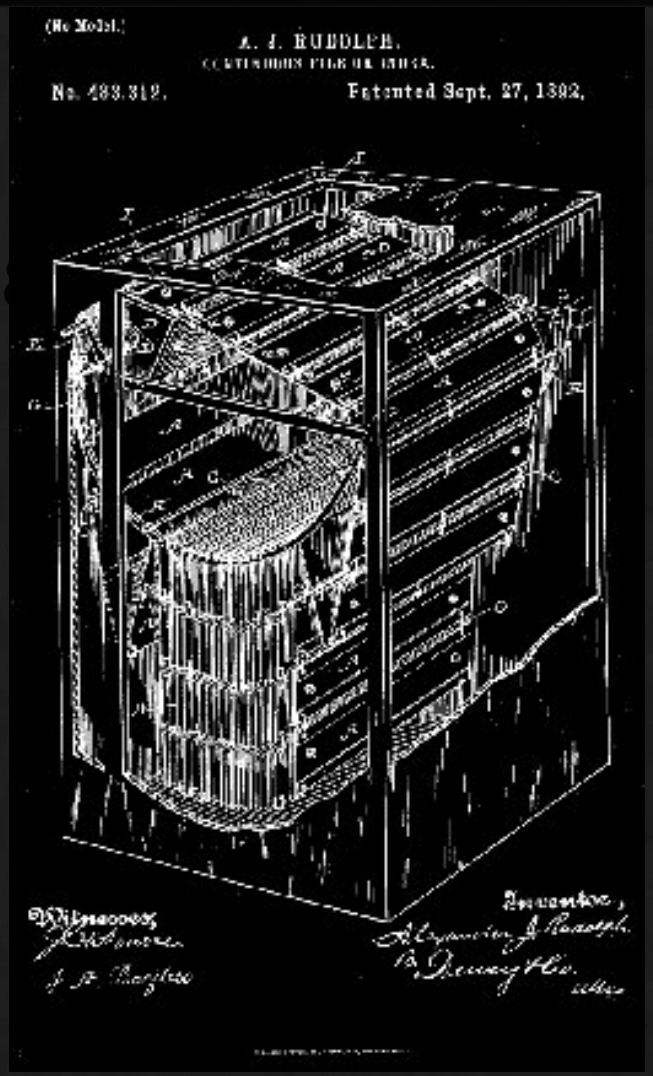
60 - Biotechnologie
61 - Sciences médicales
62 - Ingénierie. Technologie en général
63 - Agriculture, sciences et techniques relatives à l'agriculture. Sylviculture. Exploitation agricole. Exploitation de la faune et de la flore sauvage
64 - Économie domestique. Sciences ménagères
65 - Industries du transport et de la communication. Comptabilité. Gestion d'entreprise. Relations publiques.
66 - Technologie de la chimie. Industries chimiques et apparentées
67 - Industries, commerces et artisanats divers
68 - Industries, artisanats et commerces de produits finis ou assemblés
69 - Industrie de la construction. Matériaux. Méthodes et procédés

612.4 Dup

INFORMATION RETRIEVAL

- Classifications
  - 1876 DEWEY (CDD) / 1905 CDU
- Rudolph Continuous Indexer 1

http://www.libraryhistorybuff.org/rudolph.htm



INFORMATION RETRIEVAL

- Classifications
  - 1876 DEWEY (CDD) / 1905 CDU
- 40' : « Index » « Indexing »
  - Librarians : controlled language
    - 1898: Library of Congress Subject Headings
    - 1946-1962 ; 1974: Répertoires vedettes matières, RVM bibliothèque de Laval, Canada
    - 1987: RAMEAU (BnF)

INDEXATION, subst. fém.

Étymol. et Hist. 1948 « classement sous forme d'index » *(Nouv. Lar. univ.)*; 1955 écon. (R. Barre, *Écon. pol.*, P.U.F., t. 1, p. 310 ds Quem. *DDL* t. 20). Dér. de *indexer**; suff. *-(a)tion**.

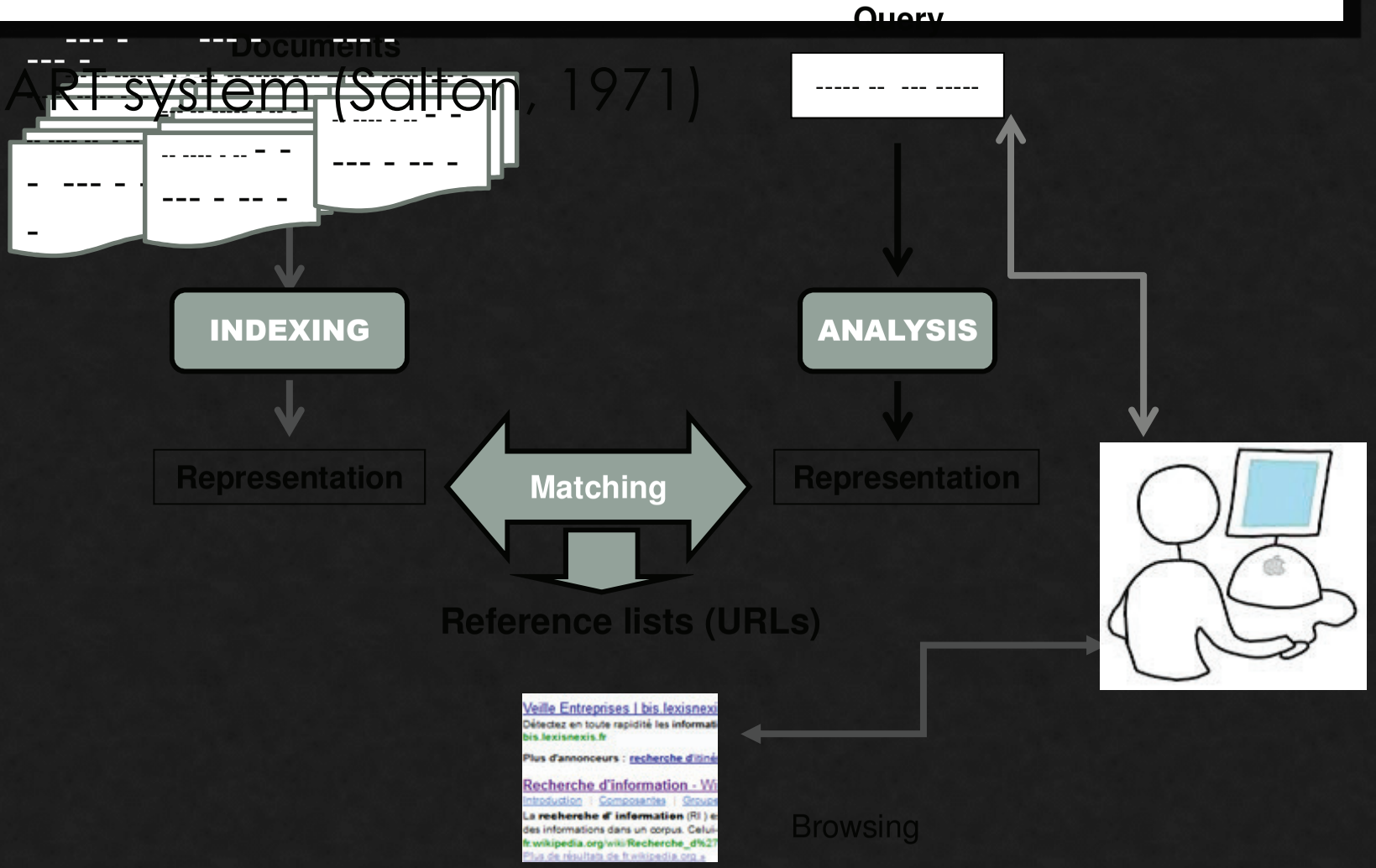
http://www.cnrtl.fr/etymologie/indexation

INFORMATION RETRIEVAL

- Catalogues (secondary documents)

INFORMATION RETRIEVAL PROCESS

- SMART system (Salton, 1971)



INFORMATION RETRIEVAL – TEST & EVALUATION

Evaluation protocol – C. Cleverdon, 60' Cranfield

- Test collection
  - Document collection
  - Query set / information needs
  - Relevance judgment
- Measures (retrieved documents vs relevance jugement)
  - Recall
  - Precision

Test collection	# documents	# queries
CACM	3 024	4 134
CISI	1 460	3 918
Medline	1 033	5 134

INFORMATION RETRIEVAL – TEST & EVALUATION

- Large scale - TREC, NIST, D. Harman, 1992

Year	Type	Name	Volume
1992	Collection	TREC1	1.26 Go 500 000 documents
1990	Driver	Normal Large	40 Mo 100 Mo
1991 1992	IBM Corsair HP C3013A kitty		1 Go 2,1 Go

  - Jugement: incomplete (pooling method)
  - New measures
    - MAP – mean average precision
    - More than 120 mesures in Trec_eval

INFORMATION RETRIEVAL – TEST & EVALUATION

- Large scale - TREC, continue

Year	Type	Name	Volume
1998	Collection	TREC7	528 155 doc
	Collection	WT10G	1 692 096 docs 50 queries
2007	Driver	Hitachi GST	1 To
2009	Collection	ClueWeb	27 To (compressed) 1 billion

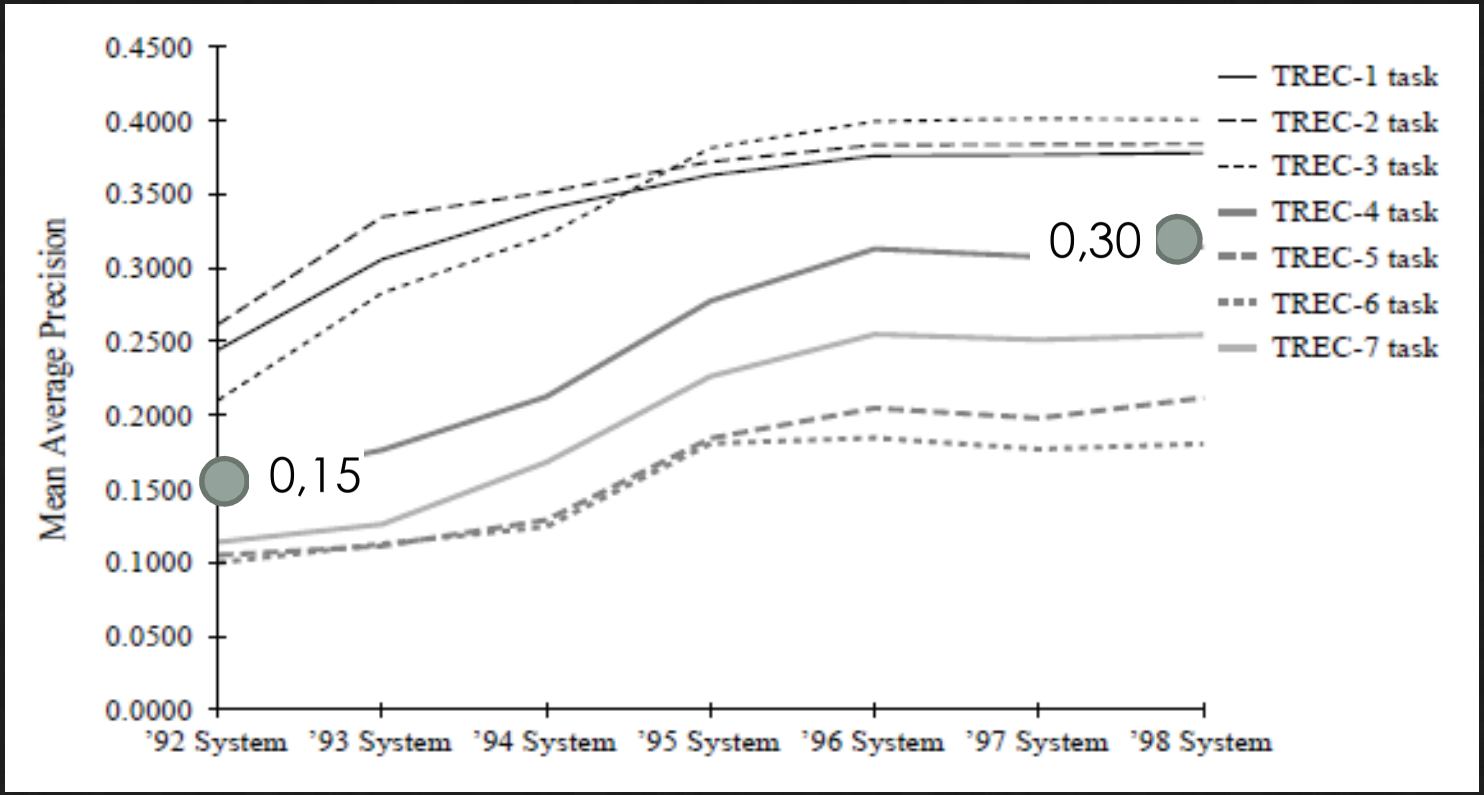
INFORMATION RETRIEVAL – TEST & EVALUATION

- TREC topics

<num> C001
<title> Architecture à Berlin
<desc> Trouver des documents au sujet de l'architecture à Berlin.
<narr> Les documents pertinents parlent, en général, des caractéristiques architecturales de Berlin ou, en particulier, de la reconstruction de certaines parties de cette ville après la chute du mur.

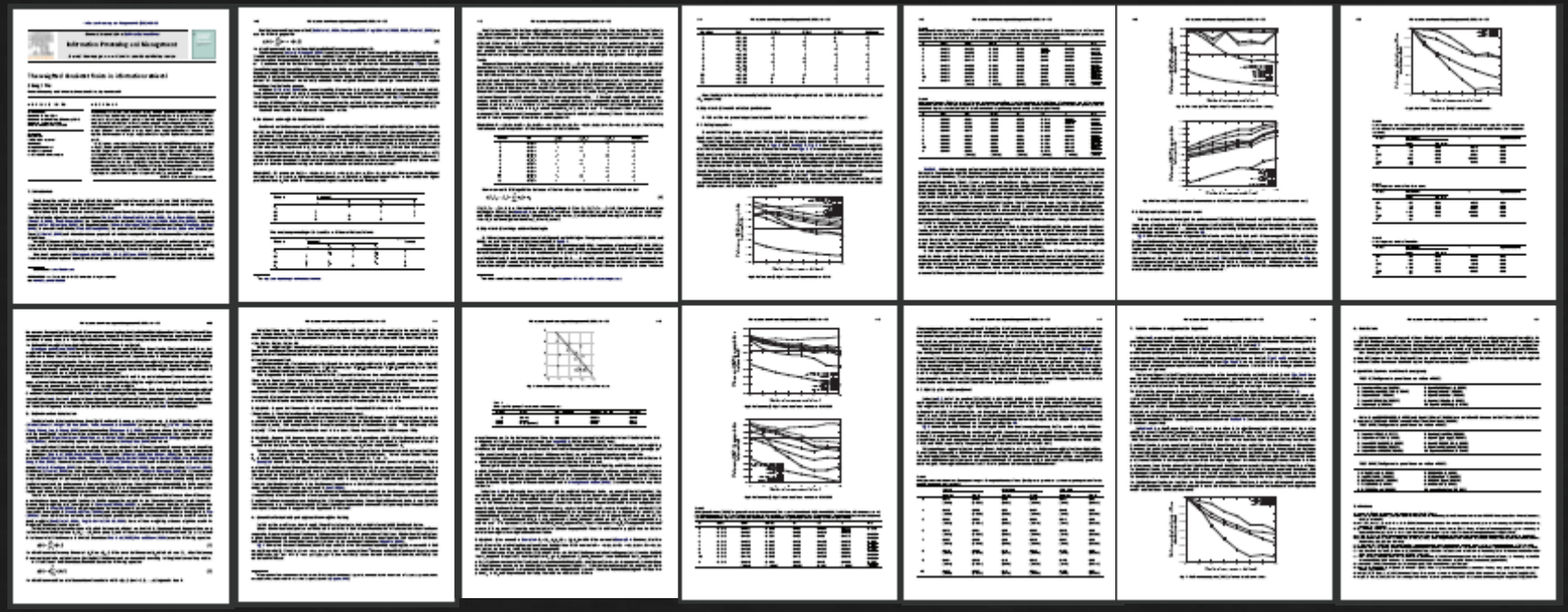
INFORMATION RETRIEVAL – TEST & EVALUATION

- Adhoc



Harman , What we have learned, and not learned, from TREC, IRSG, 2000

INFORMATION RETRIEVAL – TEST & EVALUATION

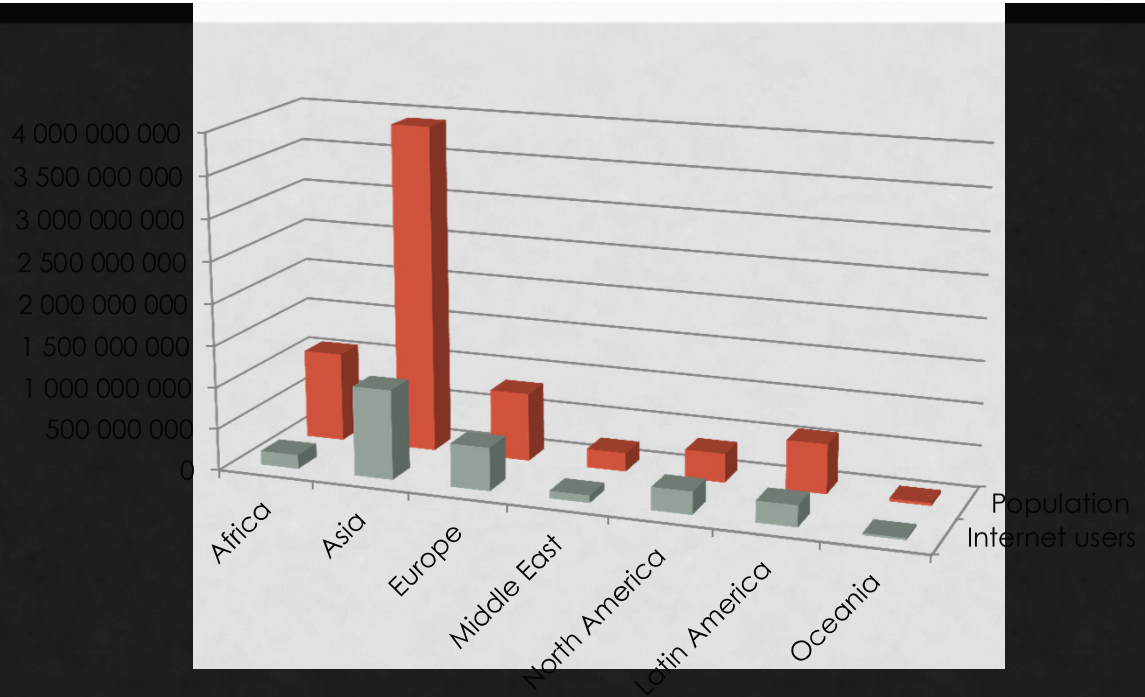


Shengli Wu, 2012, IPM, The weighted Condorcet fusion in information retrieval

MEANWHILE

- WWW
  - 1990 : Tim Berners-Lee
  - 1991: Online
  - 1993: first search engines : JumpStation, the World Wide Web Worm, Repository-Based Software Engineering (RBSE) spider
  - 1994: 10 000 servers & 10 millions of users
  - 1994: Webcrawler (indexes page content)
  - AltaVista
  - 1997: Google
  - 1998: Google index: 26 millions of pages
  - 2000: 1 billion

INTERNET USERS



- Indexed web: from 20 to 45 billion pages (Dec, 2013)
- France: 52,228,905 Internet users (juin, 2012)
- World : 2 405 518 376 (juin 2012). Population: 7 017 846 922

Internet World Stats – www.internetworldstats.com/stats.htm ; http://www.worldwidewebsize.com/

Google

- Web size:
  - > 20 à 45 billions of pages
  - indexed by Google, Yahoo! and Bing
- **Google : 78 % of the searches (Baidu 9.9%, Yahoo, Yandex, Bing)**
- Number of pages Google visits daily : 7,2 billions, 20 Po of data treated daily
- **Number of Google queries daily (2014) : 3.5 billions, about 40,000 query a second (1.20 billions in 2007)**
- About 20,000 employees
- About 450,000 servers / 25 places in the world (a large center in Irland)
- **Google data center in Mountain View built in 2005: 45,000 servers, 10 Mw**

BIG DATA IN SCIENCE

Nature, 7209(4), Sept. 2008

« Above all, data on **today's scale require scientific and computational intelligence**. Google may now have its critics, but no one can deny its impact, which ultimately stems from **the cleverness on its informatics**. The *future of science* depends in part on such cleverness again being applied to data for their own sake, complementing scientific hypotheses as a basis for exploring information cornucopia. »

SCIENCE

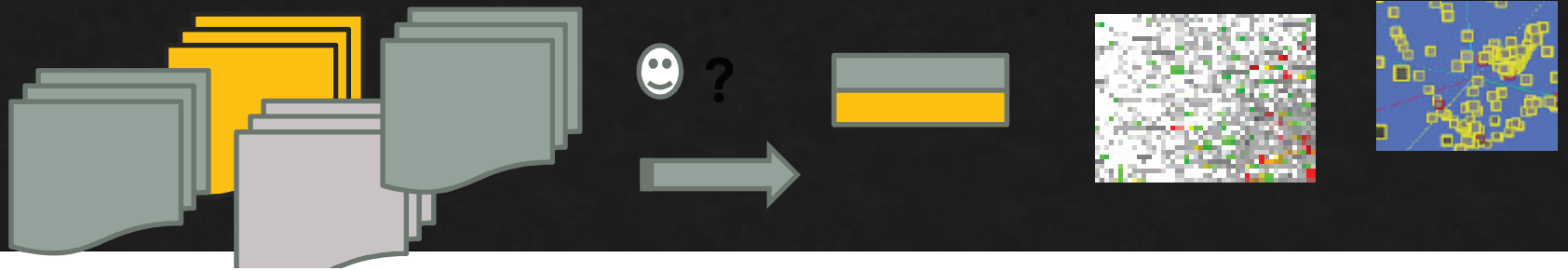
- Patents
  - 1.98 million applications in 2011 (WIPO)
- Large Hadron Collider (LHC):
  - 60 teraoctets daily / 15 peta per year
- Astronomy (OMP)
  - THéMIS solar telescope: 20 TB redu
  - European Solar Telescope: ½ Po/d
  - CDS: database of objects, related
- Medical
  - genomics
  - Images
  - video

INFORMATION ACCESS

- **either raw or created**

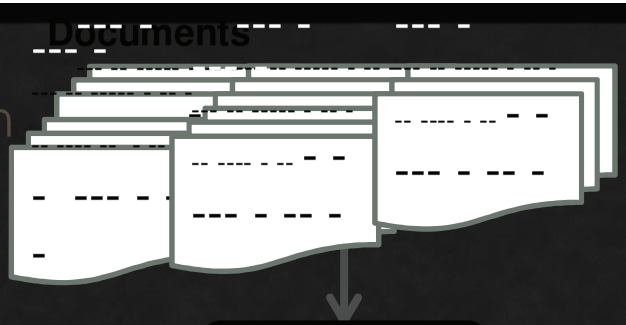
– Information retrieval: access to existing information

– Information mining and visualization: access to generated information from existing information



IR PROCESS

Indexing
- Word extraction
- « Stop word » removal
- Stemming
- Weigthing



INDEXATION

Representation

Matching

Representation

ANALYSIS

Query reformulation

Browsing

Link to retrieved documents

- Matching
- Ranking

Relevance judgment

Collecte des log de connexion

INFORMATION RETRIEVAL: ANALYSIS OF PARAMETERS

- Tuning parameters to optimize them
  - Testing various values
- Analysing parameter impact
  - Showing the results found using various values
  - Deep analysis

INFORMATION RETRIEVAL: ANALYSIS OF PARAMETERS

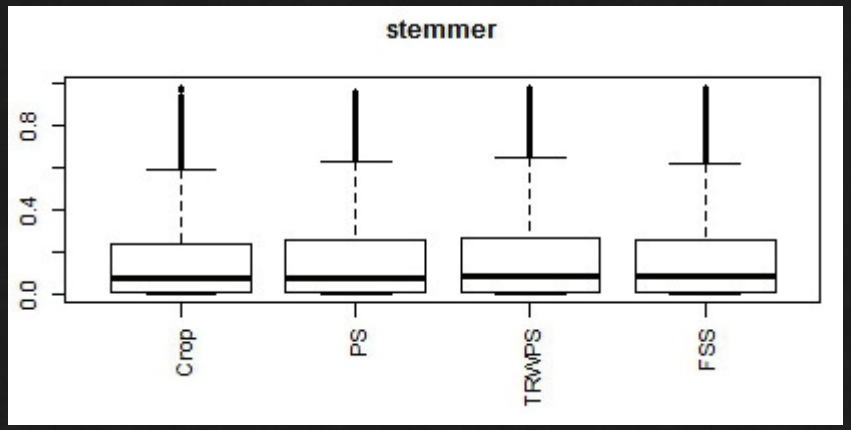
- Runs: 80,000 different combinations / 100 queries
- 8 parameters :

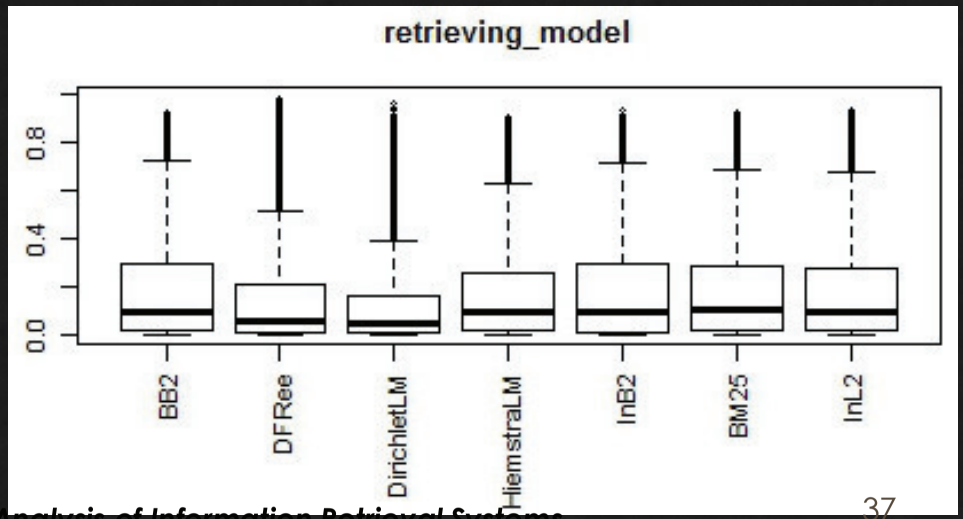
Name	Explanation	Possible Values
stemmer	name of the stemming algorithm	Crop, FSS, PS, TRWPS
retrieving_model	name of the retrieving algorithm	BB2, BM25, DFRee, DirichletLM, HiemstraLM, InB2, InL2
TrecQueryTags_process	what topic fields to use	N T T,D T,D,N T,N
parameter_free_expansion	is the parameter for expansion model free?	TRUE, FALSE
trec_qe_model	model of query reformulation	Bo1, Bo2, Info, KL, KLCom, KLCor
expansion_documents	number of top documents to use for query reformulation	2, 5, 10, 20, 50, 100
expansion_terms	number of highest weighted terms to use	2, 5, 10, 15, 20
expansion_mindocuments	number of documents a term should appear in to be eligible for expansion	2, 5, 10, 20, 50

J Ayter, C Desclaux, A Chifu, S Déjean, J Mothe. *Performance Analysis of Information Retrieval Systems. Spanish Conference on Information Retrieval,2014*

INFORMATION RETRIEVAL: ANALYSIS OF PARAMETERS

- Boxplots

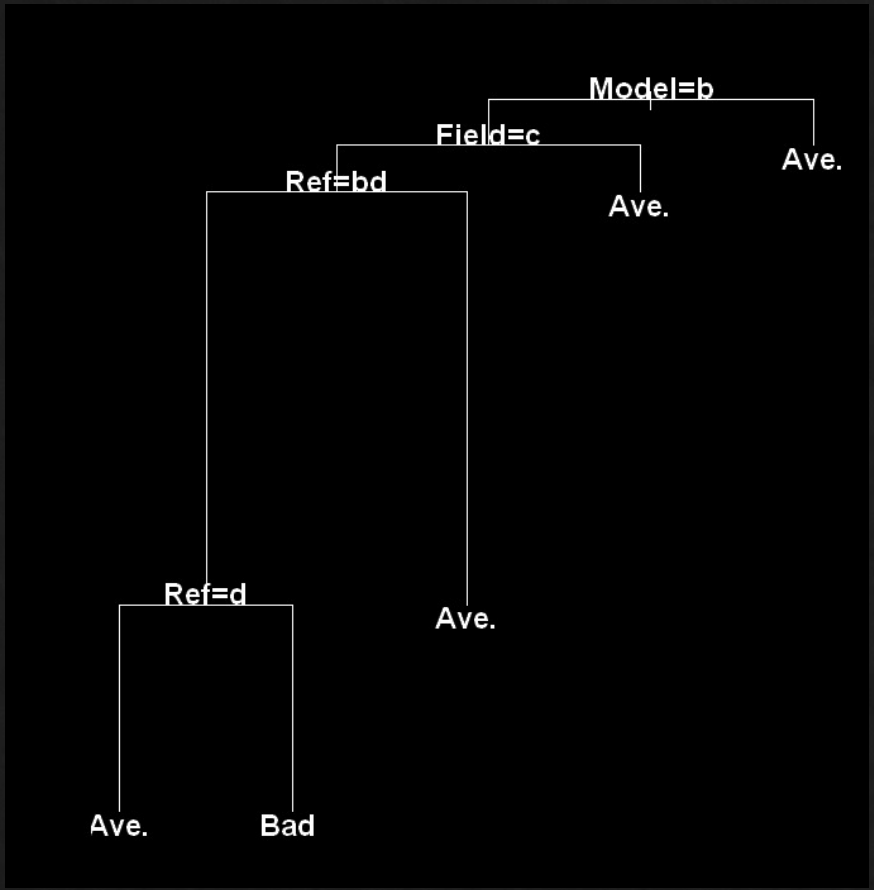




J Ayter, C Desclaux, A Chifu, S Déjean, J Mothe. *Performance Analysis of Information Retrieval Systems. Spanish Conference on Information Retrieval,2014*

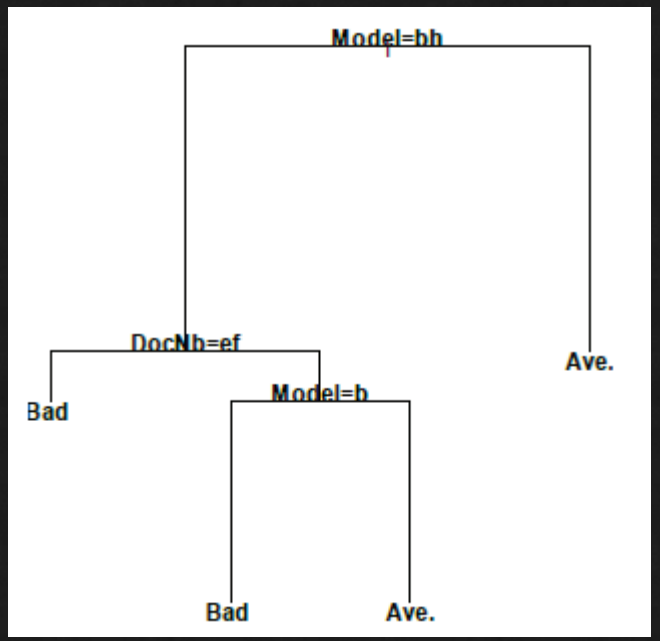
INFORMATION RETRIEVAL: ANALYSIS OF PARAMETERS

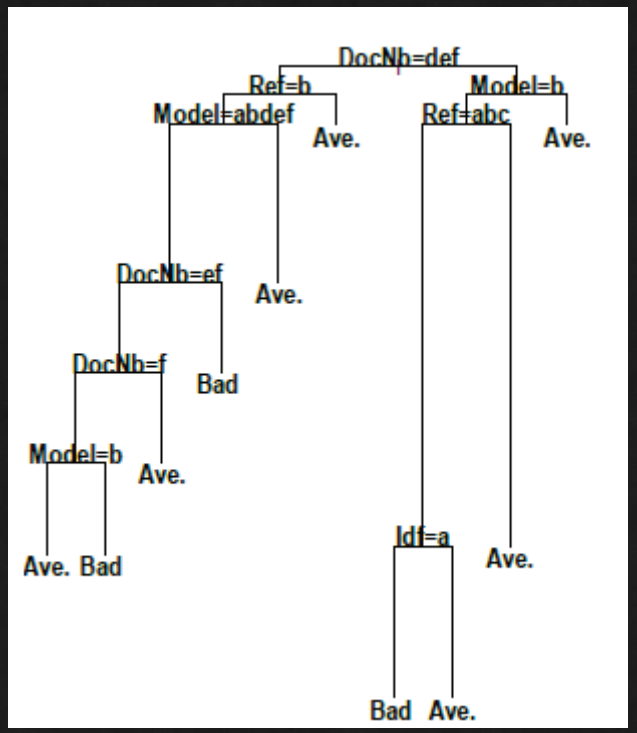
- Classification tree
  - All queries



INFORMATION RETRIEVAL: ANALYSIS OF PARAMETERS

- Classification tree (Easy / Hard queries)





INFORMATION RETRIEVAL: ANALYSIS OF EVALUATION MEASURES

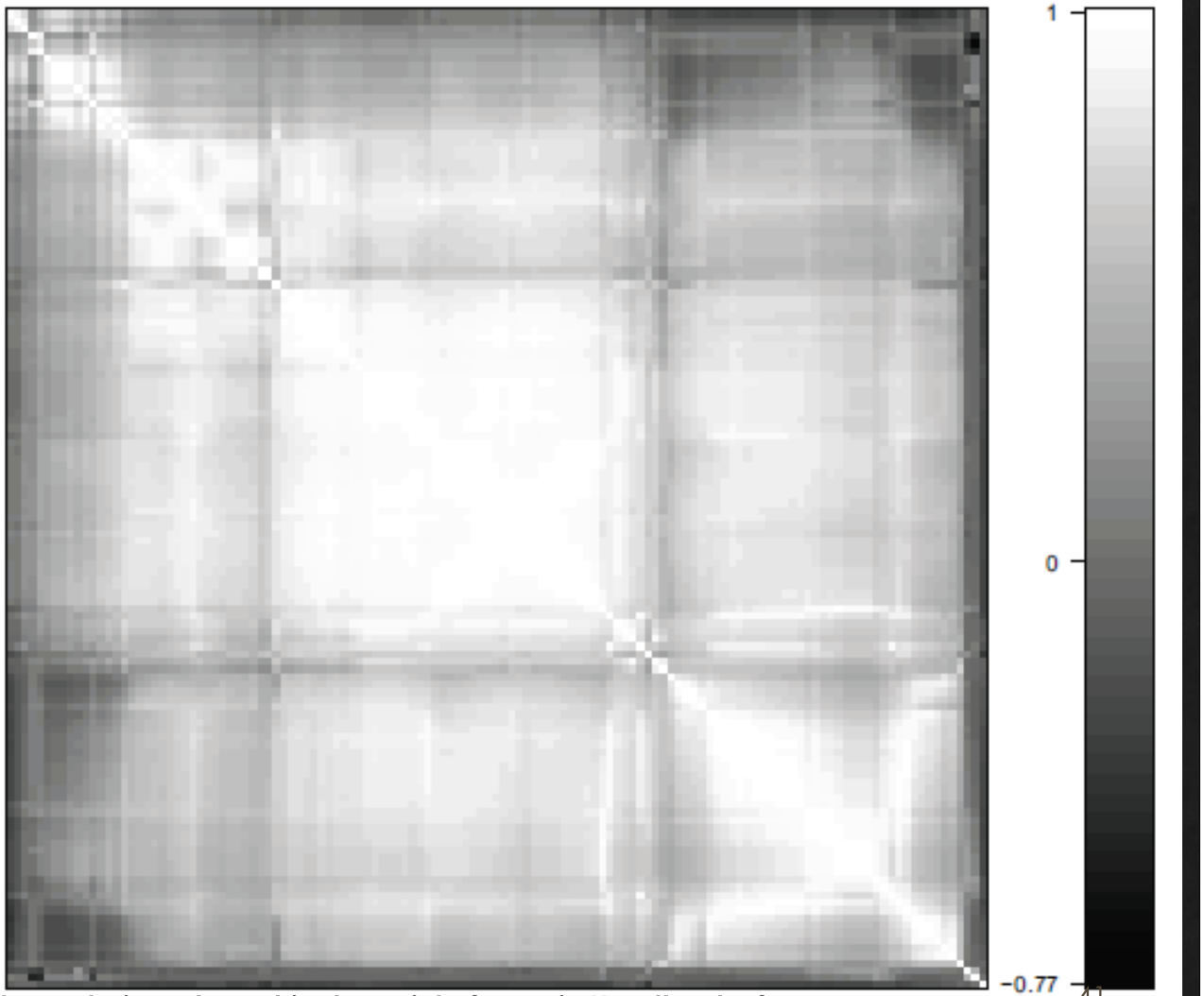
- Official runs (7 TREC adhoc, 50 topics):
  - 23 518 rows, 130 colomns

Line	Year	System	Topic	0.20R.prec	0.40R.prec	0.60R.prec	...
1	TREC 1993	Brkly3	101	0.2500	0.1250	0.1111	...
2	TREC 1993	Brkly3	101	0.3077	0.2692	0.3077	...
3	TREC 1993	Brkly3	101	0.4737	0.4474	0.4211	...
...	...	...	...	...	...	...	...
23516	TREC 1999	weaver2	448	0.0000	0.0000	0.0357	...
23517	TREC 1999	weaver2	449	0.0000	0.0000	0.0000	...
23518	TREC 1999	weaver2	450	0.7627	0.6864	0.5966	...

S. Déjean, J. Mothe. Visual clustering for data analysis and graphical user interfaces, in Handbook of Cluster Analysis (C. Hennig, M. Meila, F. Murtagh, R. Rocci Ed.), to appear 2014.

INFORMATION RETRIEVAL: ANALYSIS OF EVALUATION MEASURES

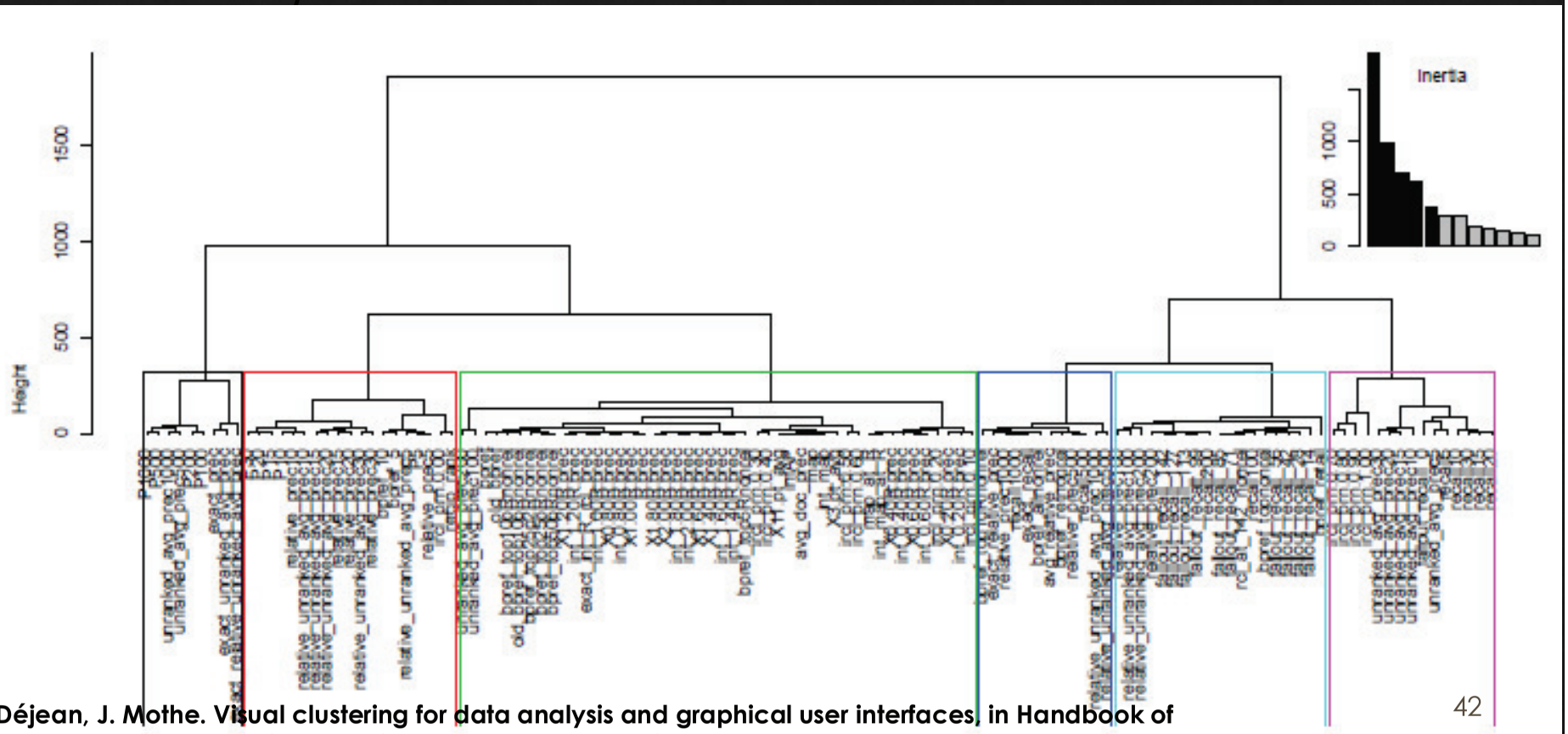
- Correlation
- 130 x 130
- the higher the correlation, the lighter the pixel
- optimal reorganization
- strong redundancy
- a "good" system is good whatever the performance measure used



S. Déjean, J. Mothe. Visual clustering for data analysis and graphical user interfaces, in Handbook of Cluster Analysis (C. Hennig, M. Meila, F. Murtagh, R. Rocci Ed.), to appear 2014.

INFORMATION RETRIEVAL: ANALYSIS OF EVALUATION MEASURES

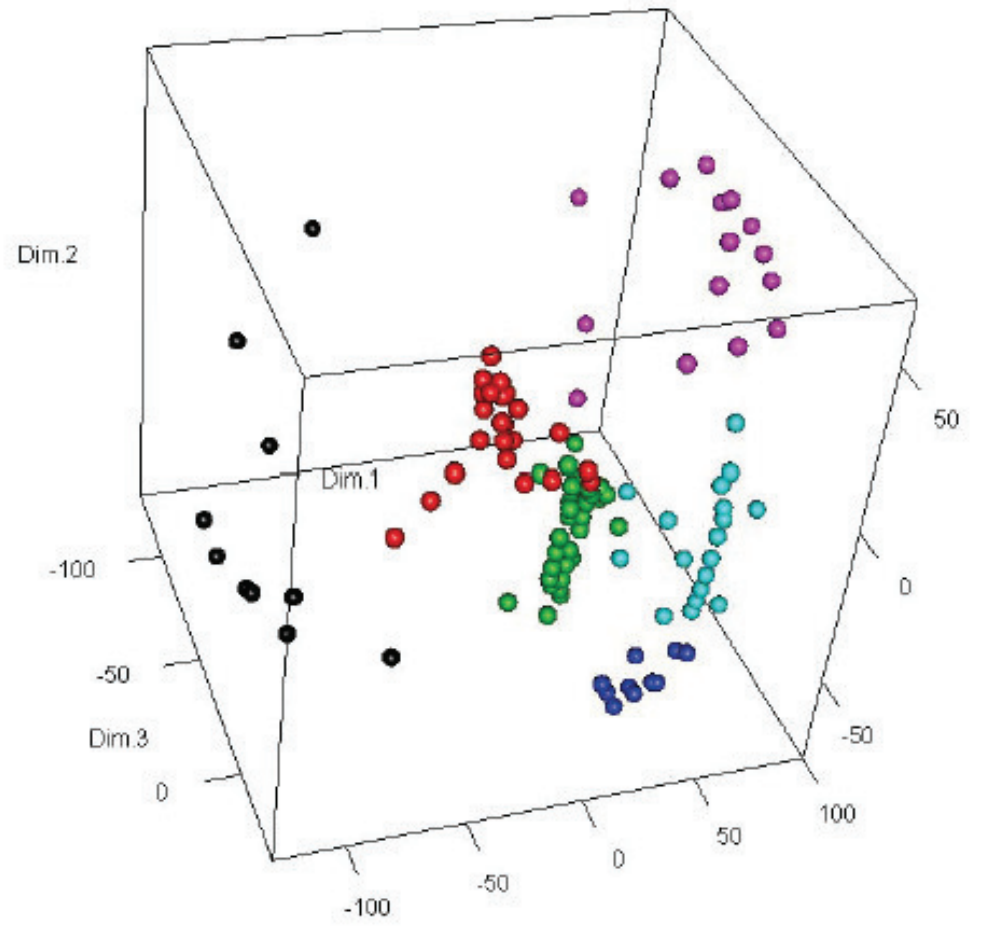
- Hierarchical clustering (Euclidean distance & Ward's criterion)



S. Déjean, J. Mothe. Visual clustering for data analysis and graphical user interfaces, in Handbook of Cluster Analysis (C. Hennig, M. Meila, F. Murtagh, R. Rocci Ed.), to appear 2014.

INFORMATION RETRIEVAL: ANALYSIS OF EVALUATION MEASURES

- Principal Component Analysis
- Representation of variables on the first two principal components PC1 and PC2
- Respectively explaining 34% and 15% of the total variance
- Colors reveal the cluster the variables belong to after partitioning



S. Déjean, J. Mothe. Visual clustering for data analysis and graphical user interfaces, in Handbook of Cluster Analysis (C. Hennig, M. Meila, F. Murtagh, R. Rocci Ed.), to appear 2014.

INFORMATION RETRIEVAL: ANALYSIS OF QUERY LOGS

- Revue.org logs
- Sessions
- Queries / browsing
- Reformulation

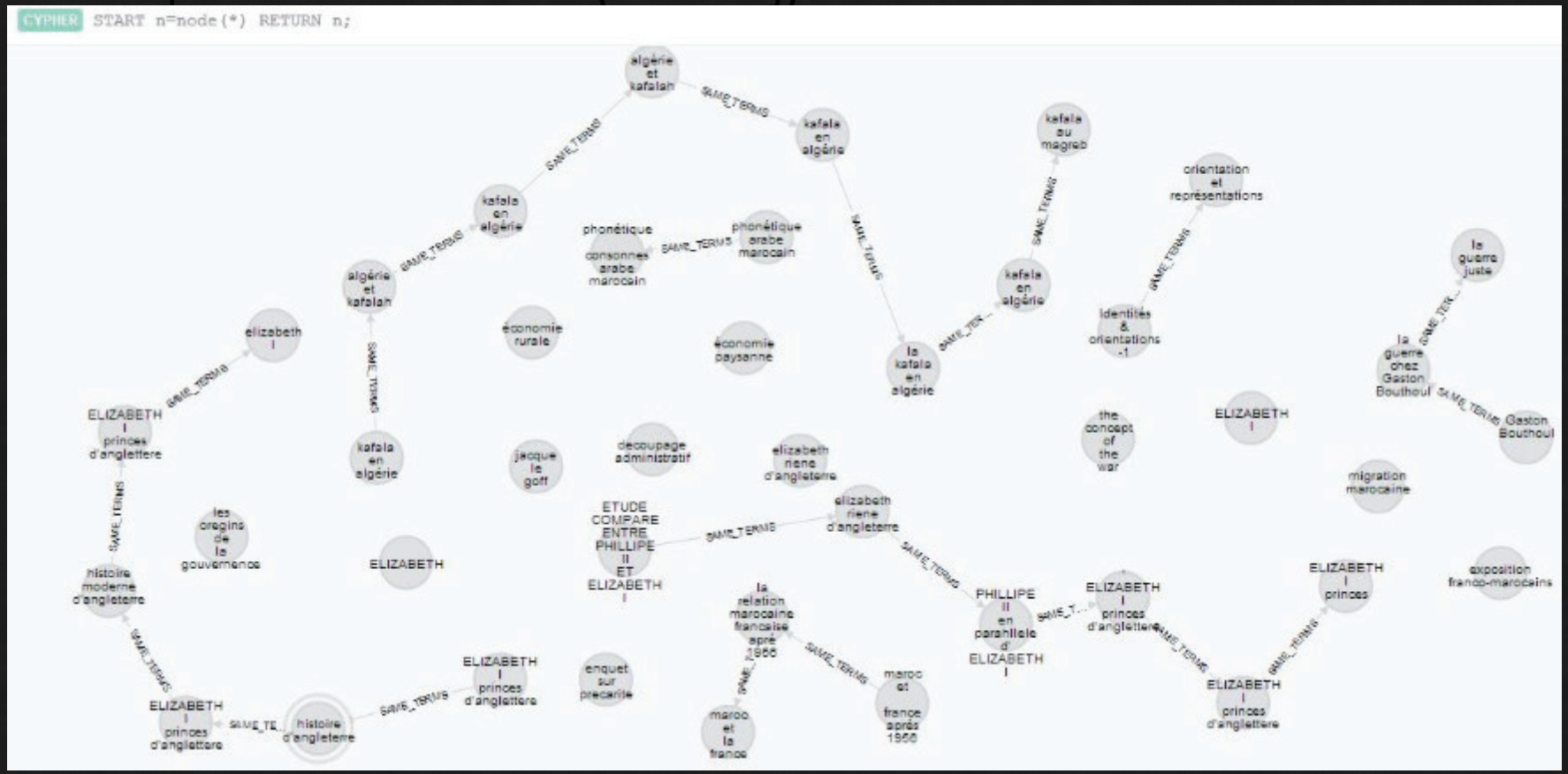
```
IP1   1   date1-08:15:56   NA    REQ[tourism]   Pub[Articulo - Journal of Urban Research]  -
IP1   1   date1-08:18:42   166   NA    NA    http://www.openedition.org/10031
IP1   1   date1-08:20:25   103   REQ[Subaltern Urbanisation in India]   NA
IP1   1   date1-08:21:15   50    ---->DOC[http://leo.hypotheses.org/8075]   NA
IP1   1   date1-08:22:54   99    ---->DOC[http://suburbin.hypotheses.org/297]   NA
IP1   1   date1-08:34:08   674   ---->DOC[http://suburbin.hypotheses.org/315]   NA
IP1   1   date1-08:45:47   699   ---->DOC[http://suburbin.hypotheses.org/302]   NA
IP1   1   date1-08:46:24   37    ---->DOC[http://suburbin.hypotheses.org/302?lang=en_GB]  NA
IP1   1   date1-08:48:48   144   NA    NA    http://www.openedition.org/8991
IP1   1   date1-08:54:27   339   REQ[tourism]   NA
IP1   1   date1-08:57:00   139   REQ[rural tourism]   NA

-
IP2   1   date2-23:36:29   NA    ---->DOC[http://droitcultures.revues.org/2118]   NA
IP2   1   date2-23:37:26   57    REQ[la kafala en algérie]  Pub[Droit et Cultures]
IP2   1   date2-23:38:08   42    REQ[kafala en algérie]   NA
IP2   1   date2-23:38:41   33    ---->DOC[http://anneemaghreb.revues.org/97]   NA
IP2   1   date2-23:38:43   2     ---->DOC[http://anneemaghreb.revues.org/476]   NA
IP2   1   date2-23:40:04   81    ---->DOC[http://anneemaghreb.revues.org/95]   NA
IP2   1   date2-23:40:26   22    ---->DOC[http://clio.revues.org/10293]   NA
IP2   1   date2-23:43:03   157   ---->DOC[http://anneemaghreb.revues.org/72]   NA
IP2   1   date2-23:44:07   64    ---->DOC[http://anneemaghreb.revues.org/394]   NA
IP2   1   date2-23:45:07   60    REQ[algérie et kafalah]   NA
IP2   1   date2-23:45:37   30    REQ[kafala en algérie]   NA
IP2   1   date2-23:45:42   5     REQ[algérie et kafalah]   NA
IP2   1   date2-23:45:43   1     REQ[kafala en algérie]   NA
```

On going work

INFORMATION RETRIEVAL: ANALYSIS OF QUERY LOGS

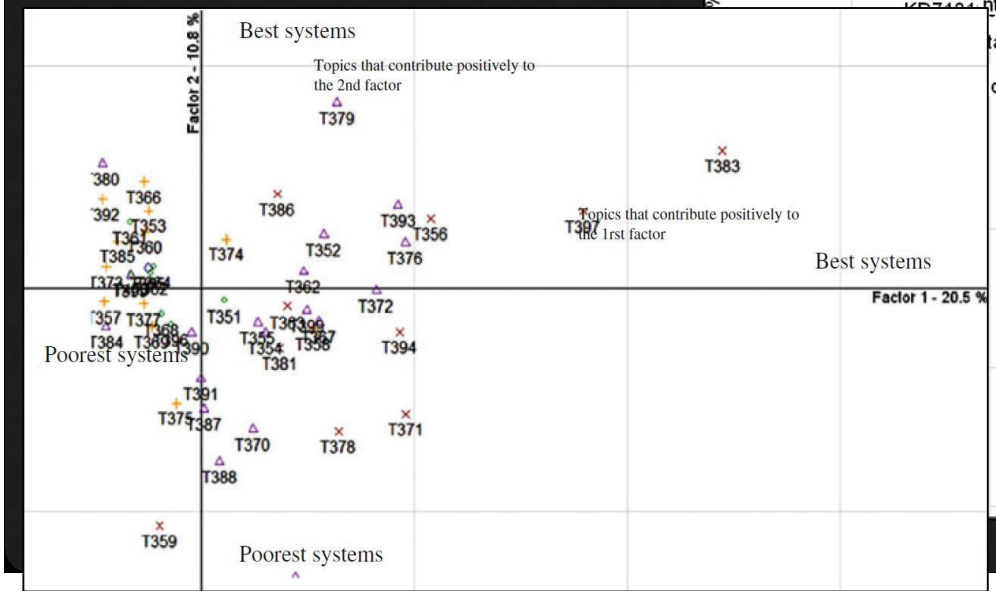
- Graph data base (Neo4j)

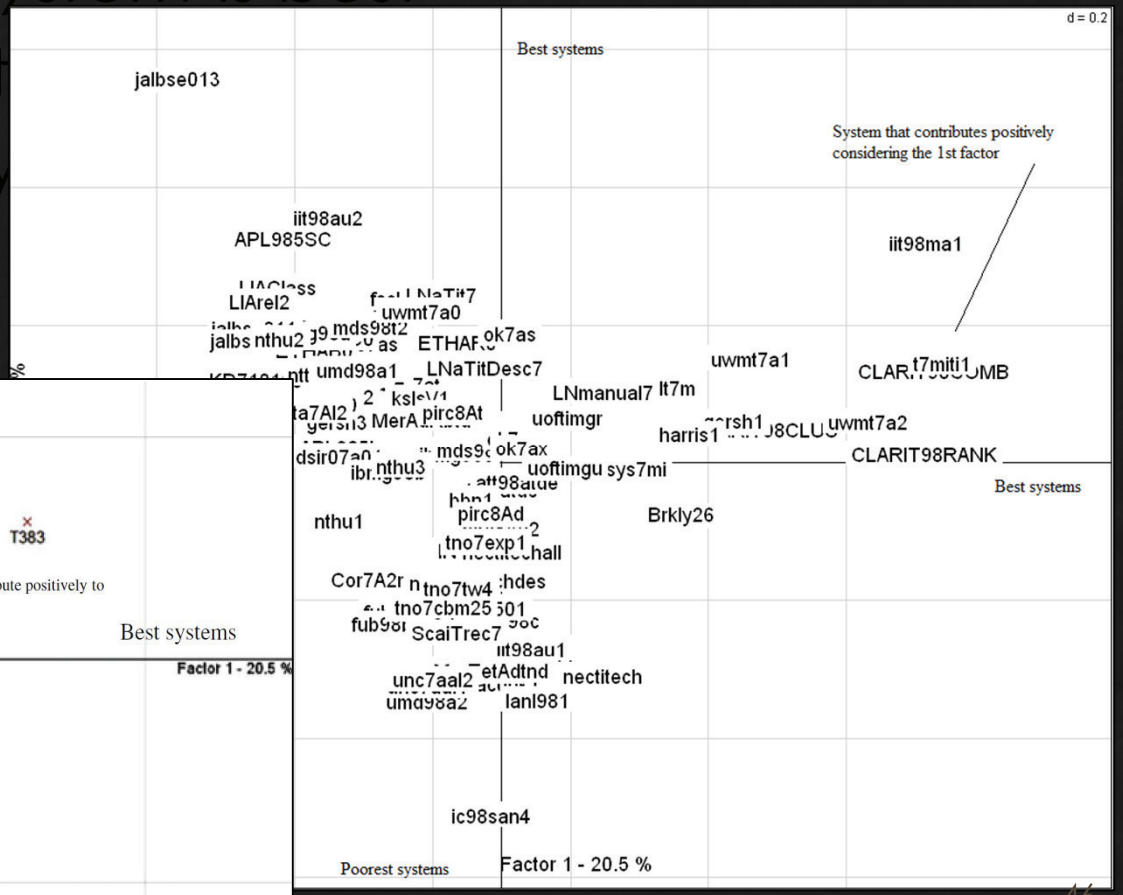


On going work

INFORMATION RETRIEVAL: SELECTIVE FUNCTIONS

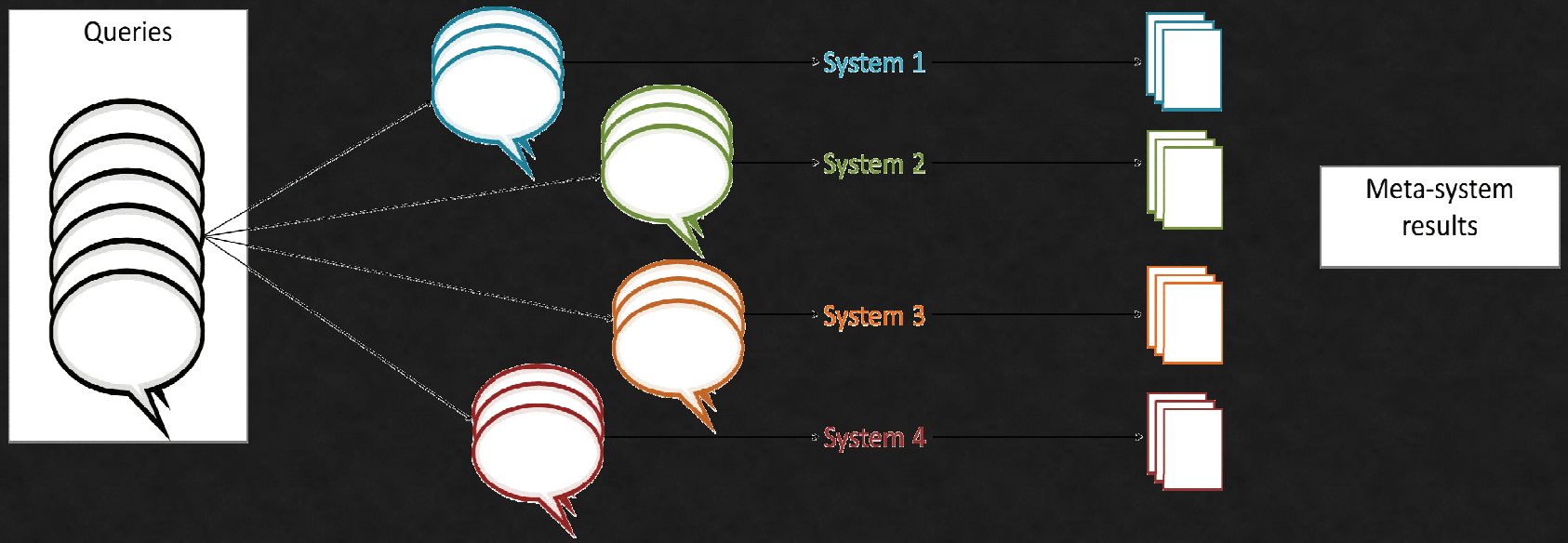
- Learning which *system* is best
  - System variabilit
  - Query variability





INFORMATION RETRIEVAL: SELECTIVE FUNCTIONS

- Learning which *system* is best



INFORMATION RETRIEVAL: SELECTIVE FUNCTIONS

- Based on training on a sub-set of documents
  - For any (repeated) query : learn the best system on a set of documents (given relevance judgement)

Method	MAP imporvment*:
a system <-> a query	+21%
a system set <-> a query set	+24%

  - Data: Trec 7
  - Using the participating systems
  - Comparison with the best participating system

A. Bigot, C. Chrisment, T. Dkaki, G. Hubert, J. Mothe. Fusing Different Information Retrieval Systems according to Query-Topics. Information Retrieval Journal, Vol. 14 N. 6, p. 617-648, 2011.

INFORMATION RETRIEVAL: SELECTIVE FUNCTIONS

- Based on training on query features
  - Pre-retrieval features
    - WordNet Senses
    - Inverse Document Frequency
  - Post retrieval
    - Standard Deviation
    - Query Feedback
- Learning using SVM
  - Training / testing query sets

Adrian Chifu, Josiane Mothe. Expansion sélective de requêtes par apprentissage, Conférence francophone en Recherche d'Information et Applications (CORIA 2014), march 2014.

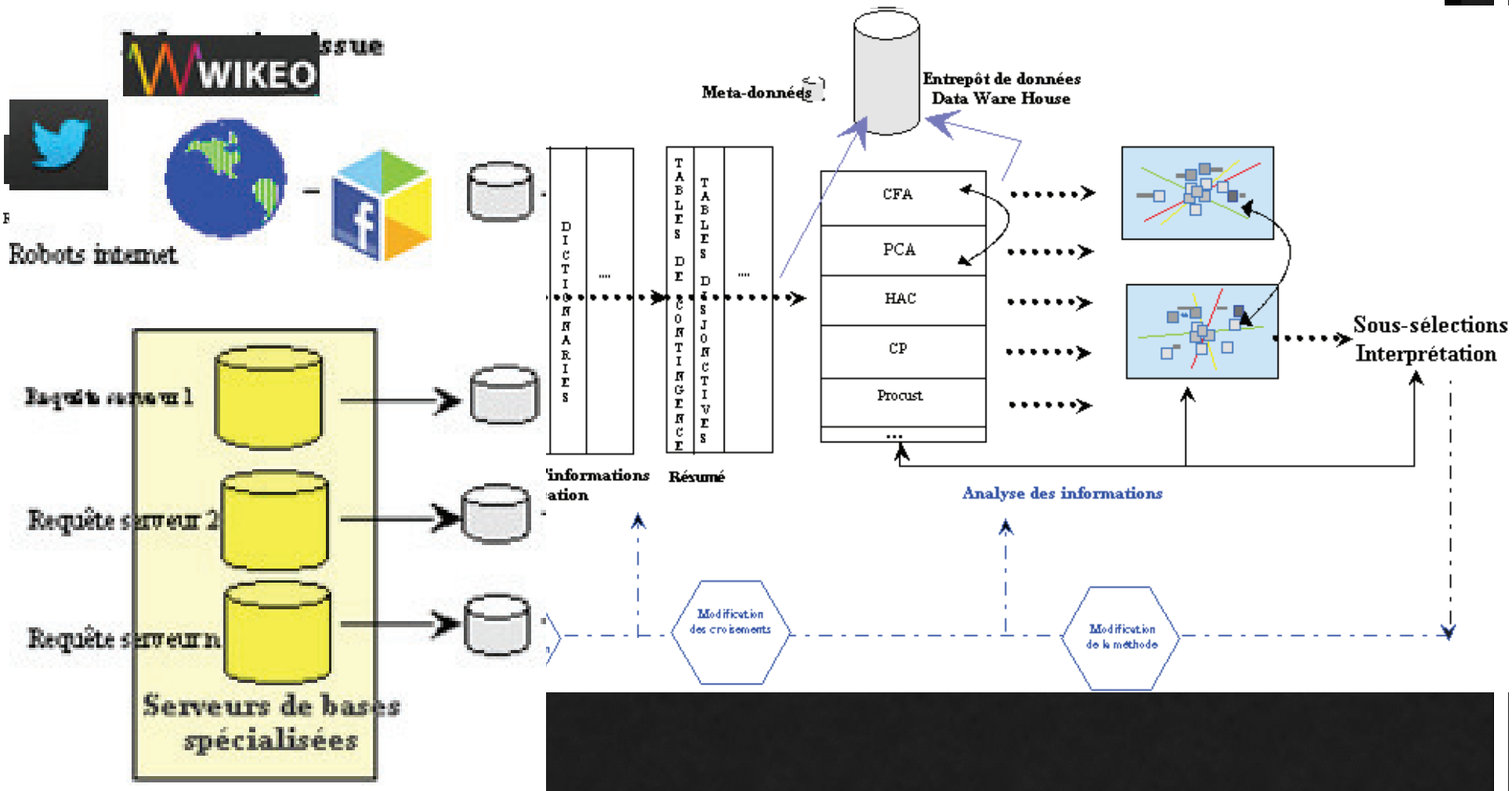
INFORMATION RETRIEVAL: SELECTIVE FUNCTIONS

- Based on training on query features

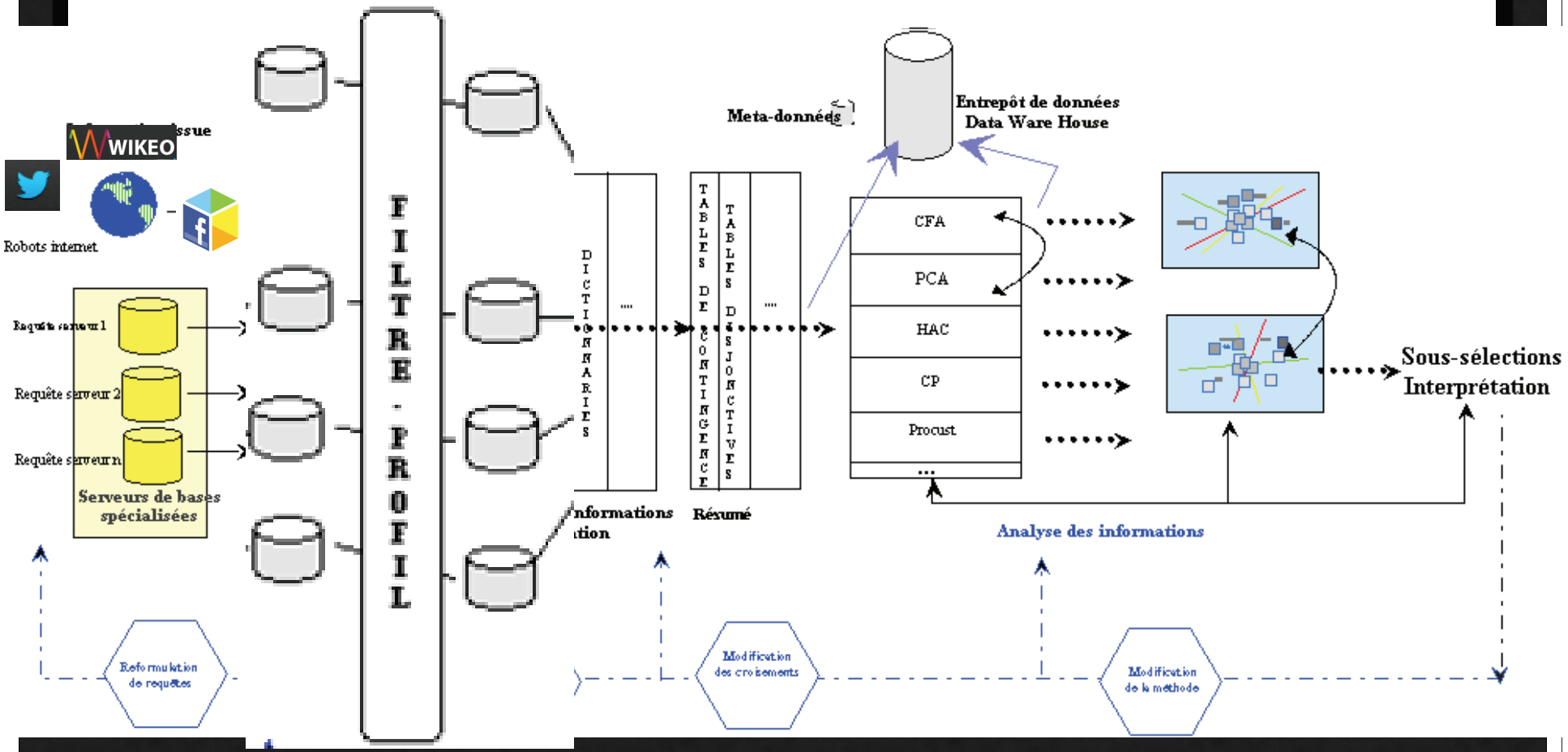
	Robust	WT10G	GOV2
Baseline TRF	0.265	0.234	0.331
Baseline TDRF	0.305	0.264	0.334
Result with decision	**0.317**	**0.296**	**0.372**
Improvement TRF	**19.41%**	**24.65%**	**12.59%**
Improvement TDRF	**3.90%**	**12.27%**	**11.35%**
p-value TRF			
p-value TDRF			
SVM accuracy	97.20%	100%	100%

Adrian Chifu, Josiane Mothe. Expansion sélective de requêtes par apprentissage, Conférence francophone en Recherche d'Information et Applications (CORIA 2014), march 2014.

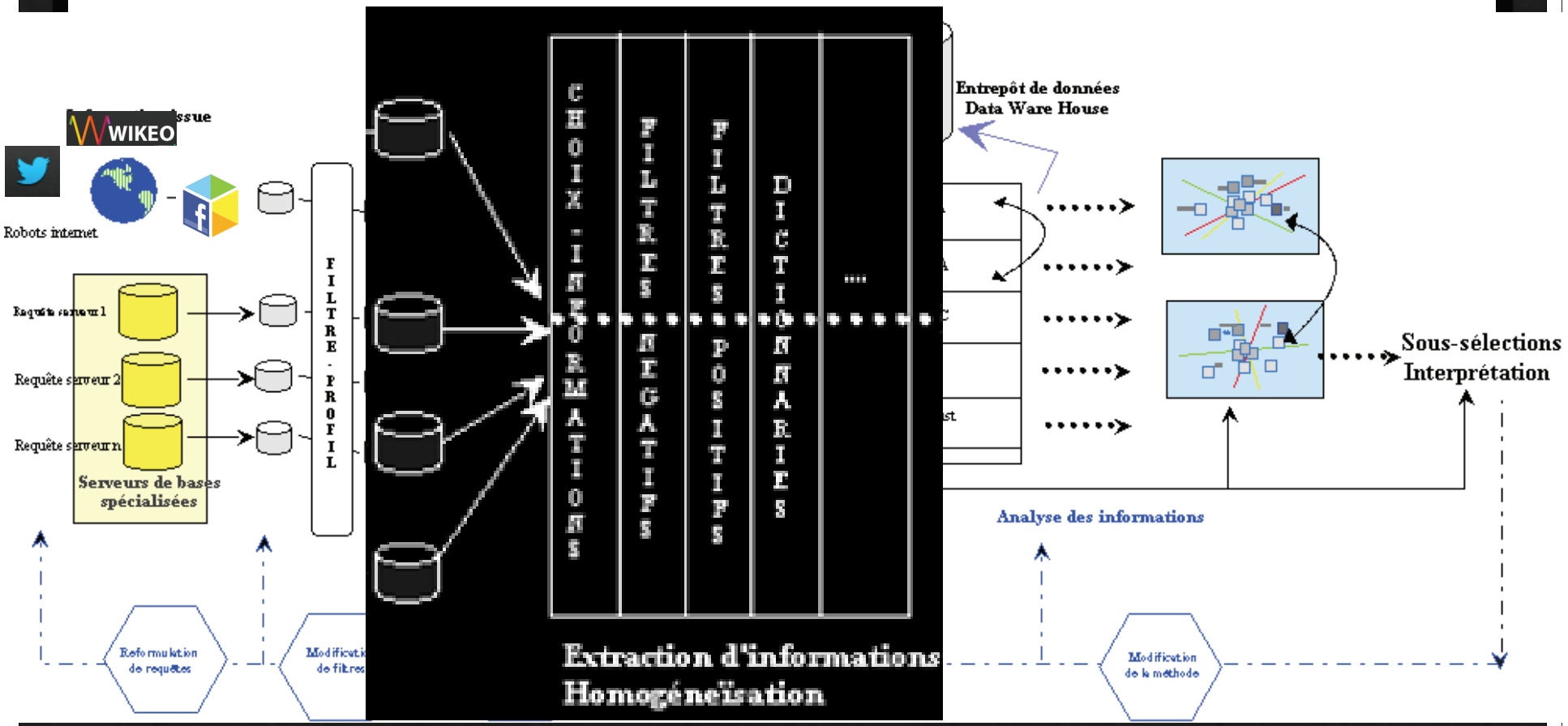
INFORMATION MINING PROCESS



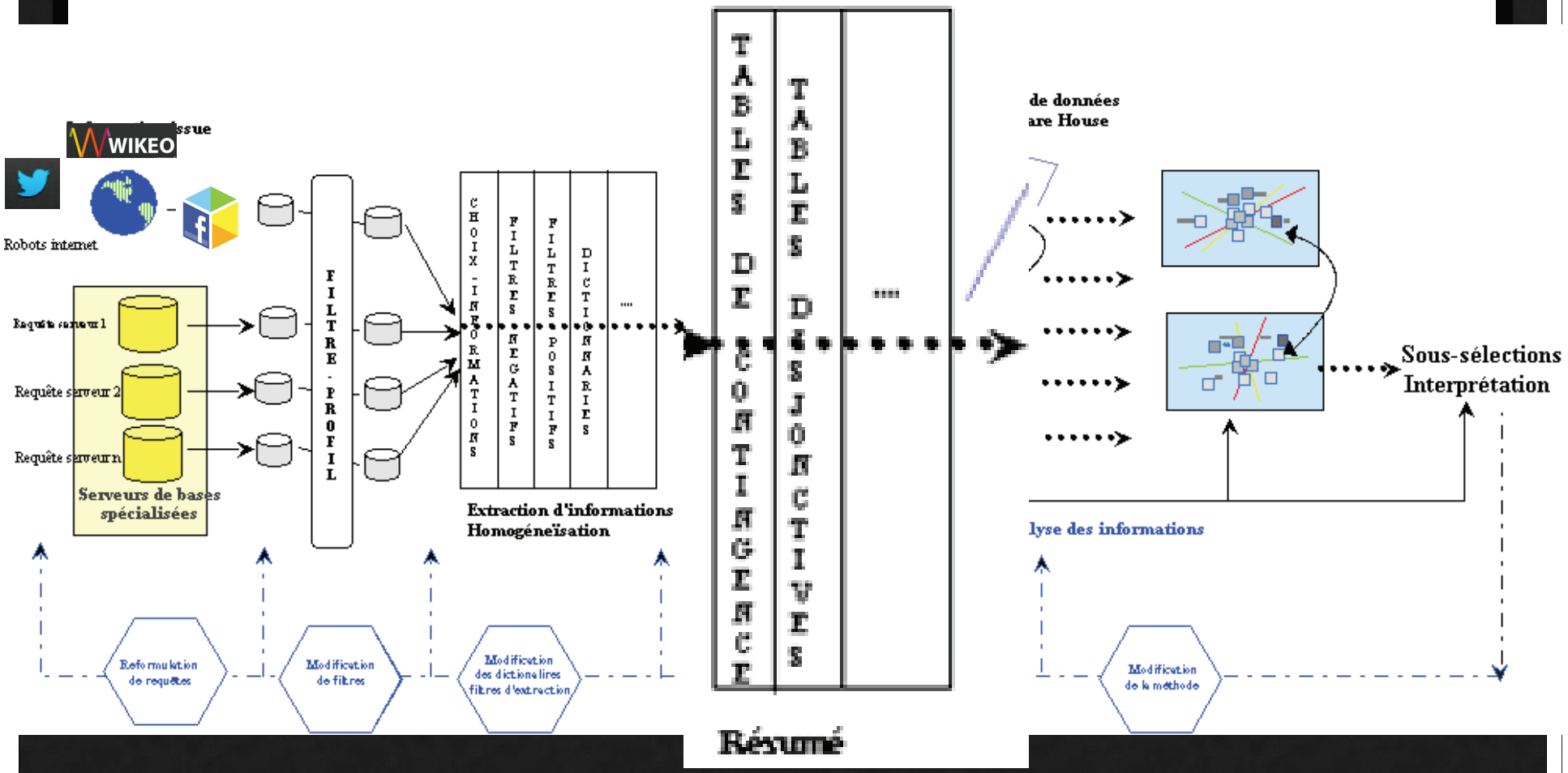
INFORMATION MINING PROCESS



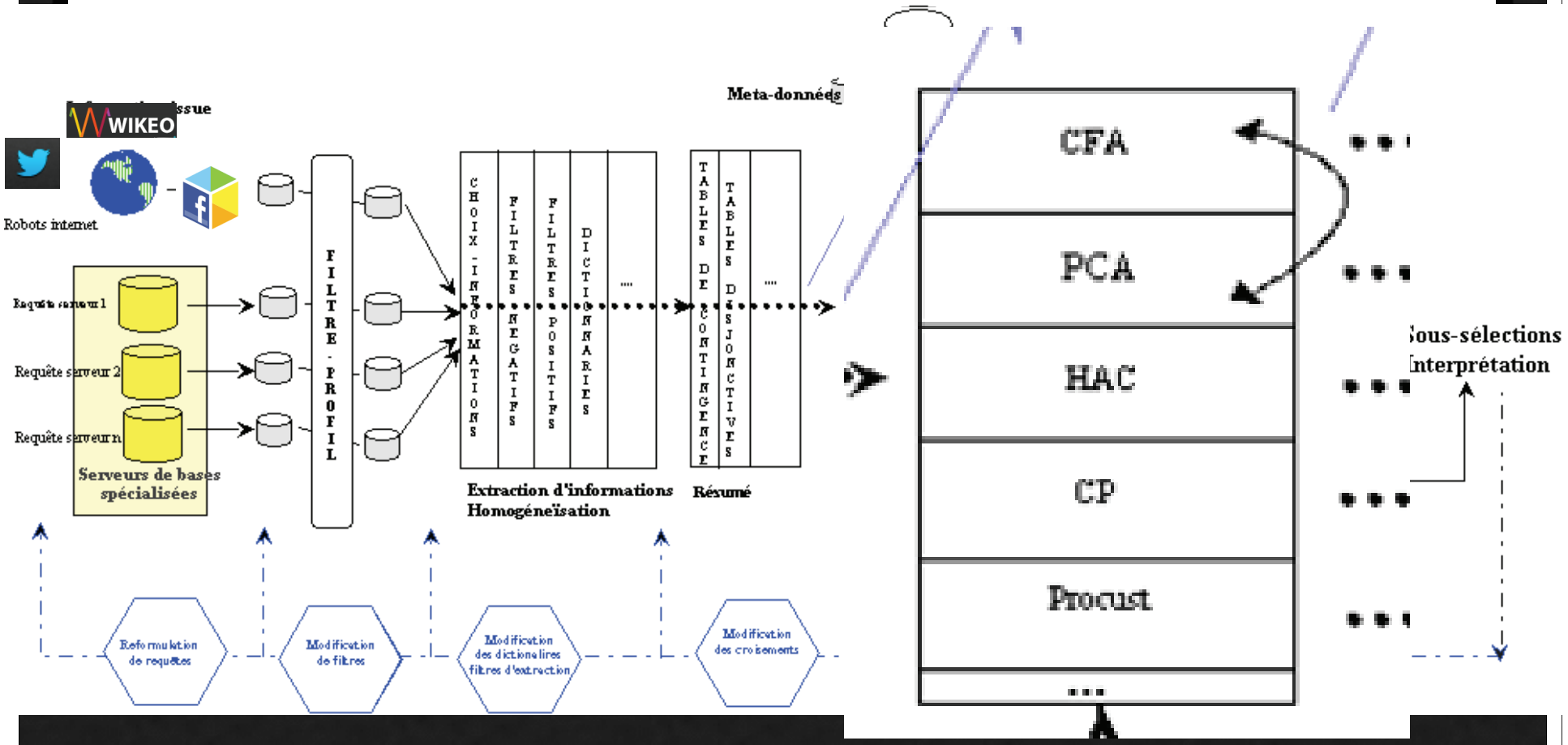
INFORMATION MINING PROCESS



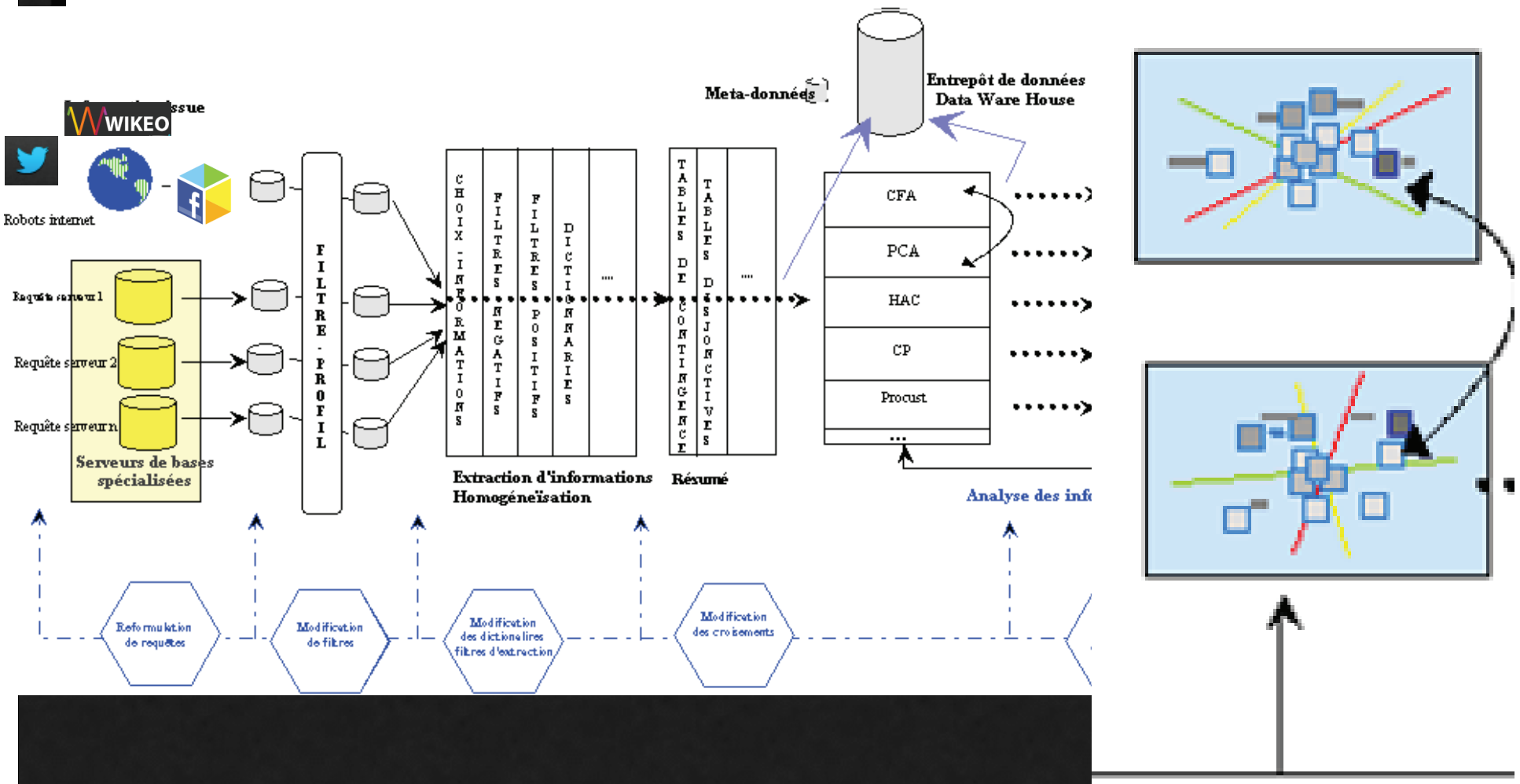
INFORMATION MINING PROCESS



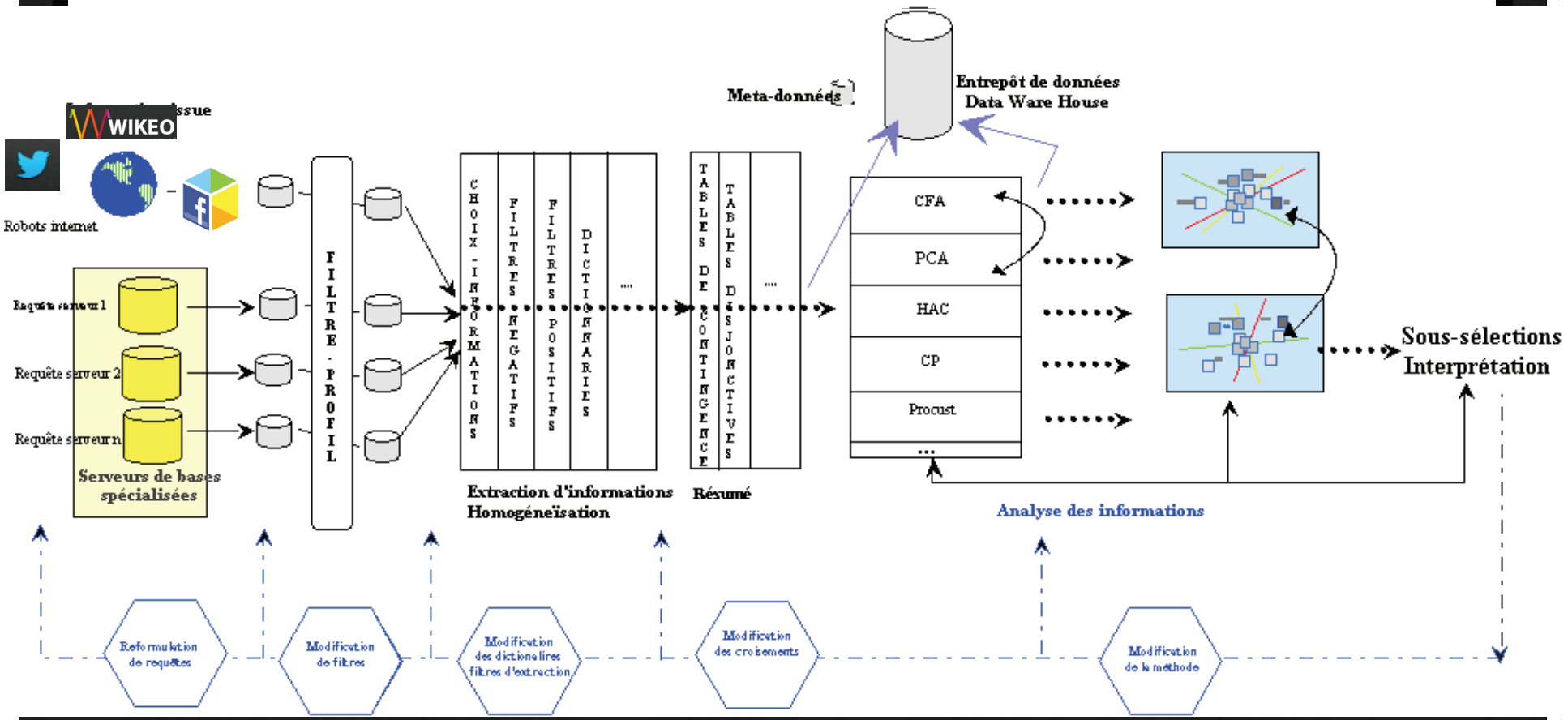
INFORMATION MINING PROCESS



INFORMATION MINING PROCESS



INFORMATION MINING PROCESS

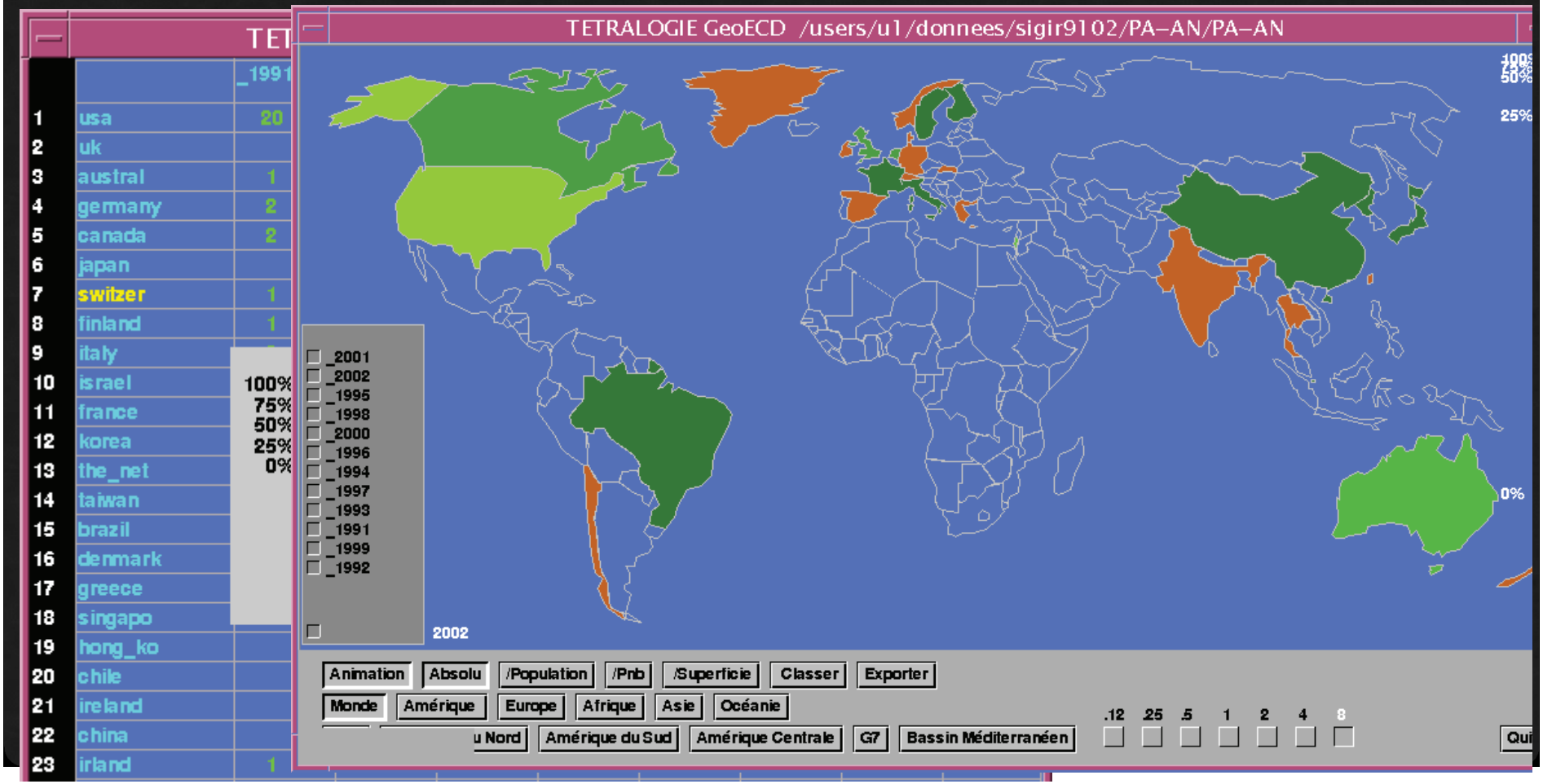


ANALYSIS OF TEXTS

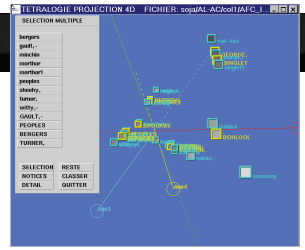
- Data: scientific publications, patents, blogs, tweets, …
- Frequency: major actors, terms, …
- Relationships
  - Collaborative networks (authors, …)
  - Geographic networks
  - Large network
  - Temporal evolution
- Tétralogie (IRIT)

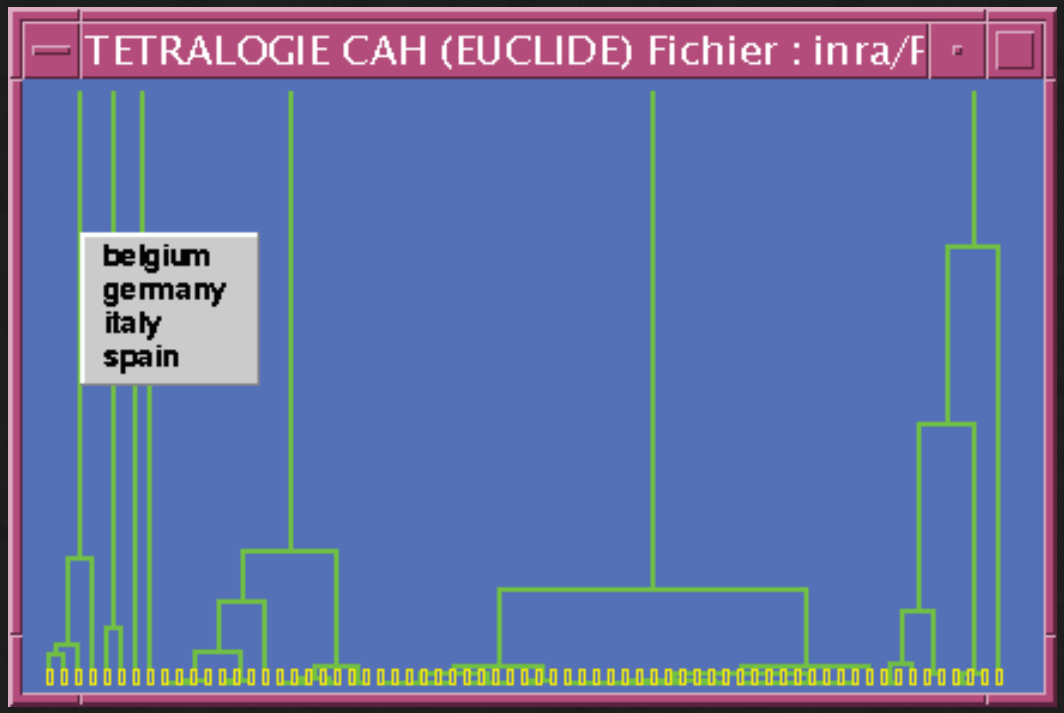
GRAPHES ET RÉSEAUX DE COLLABORATION

- Analysis of frequency [Dousset, 2012]



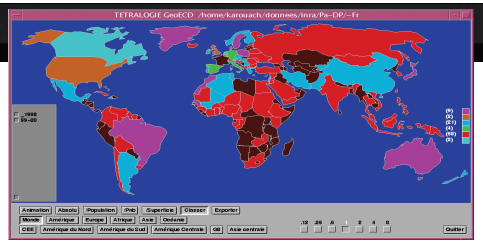
TÉTRALOGIE



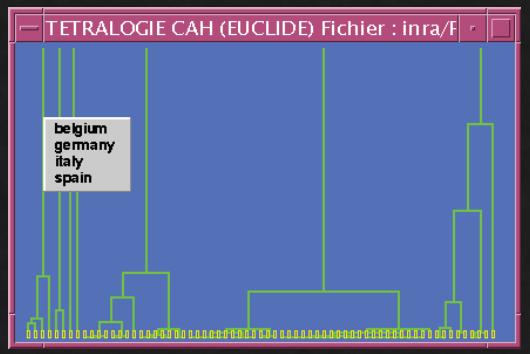


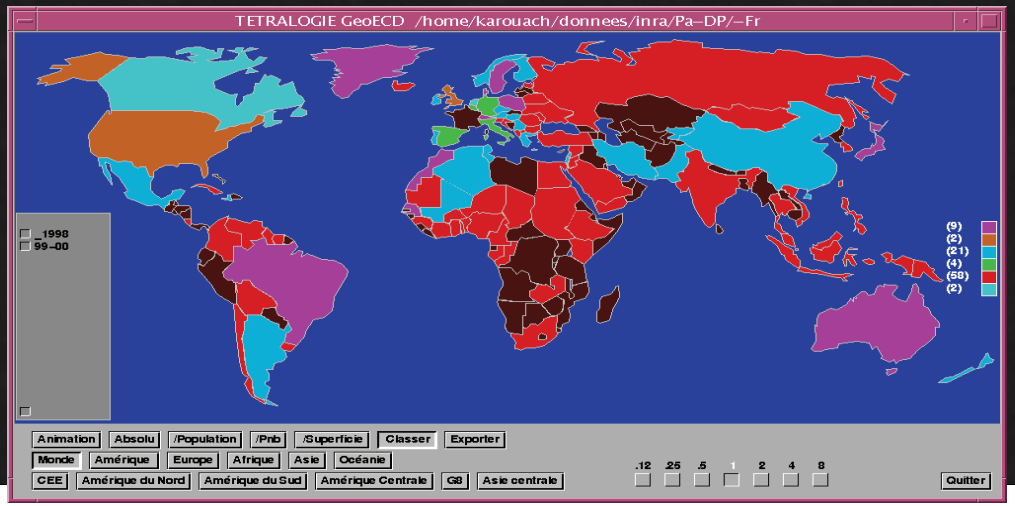
© Dousset

GEOECD



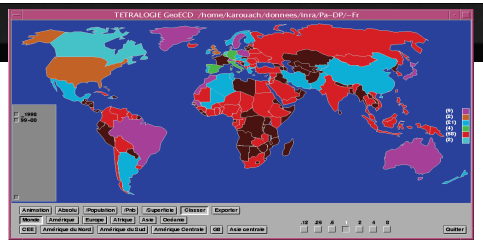
- Geo-referenced data



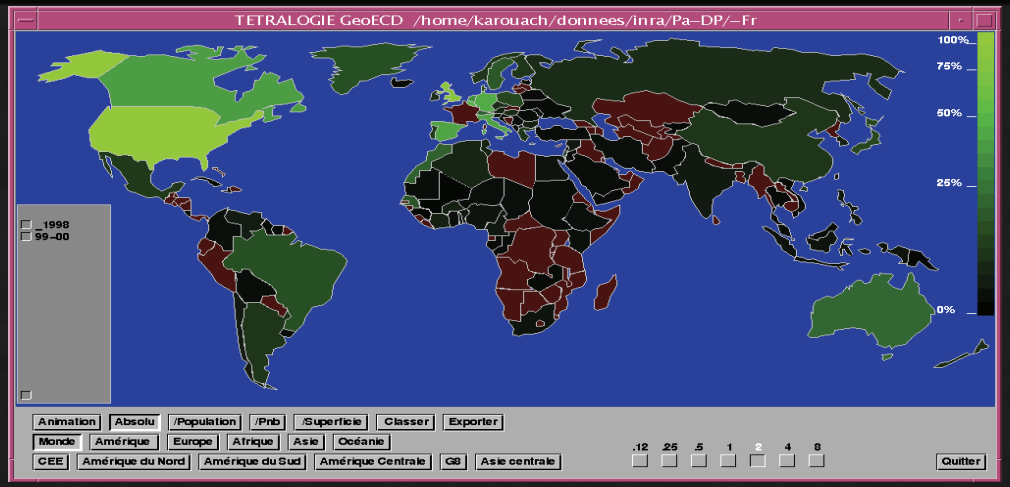


© Karouach

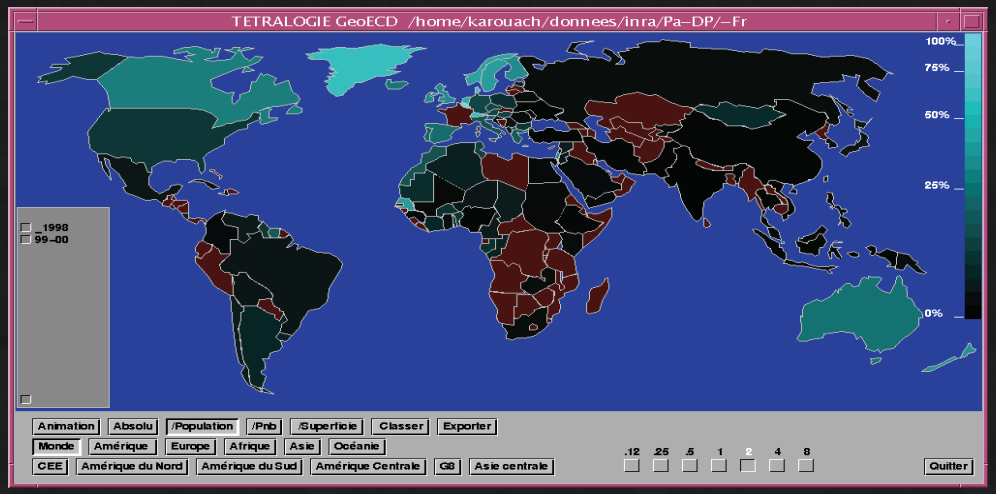
GEOECD



- Absolute mode

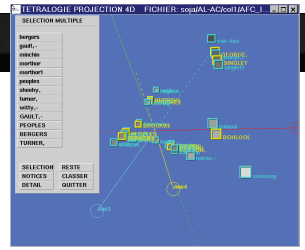


- Relative mode



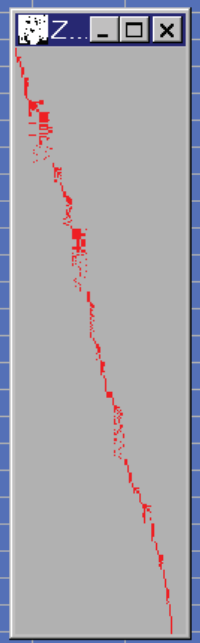
© Karouach

TÉTRALOGIE



TETRALOGIE TABLEUR FICHIER: AL-AC-

		ANGLE,-	BALATTI	MONTALD	PUEPPKE	BETHLEN	BROWN,-	BOLLICH	DUNIGAN	MILLHOL	BROCKWE
1	angle,-	4									
2	hashem,	2									
3	chaney,	1									
4	heckman	1									
5	heger,-	1									
6	madaria	1									
7	balatti		4	3	1						
8	montald		3	3							
9	guiamet		2	2							
10	pueppke		1		5						
11	krishna				2						
12	heron,-				1						
13	zdor,-r				1						
14	bethlen					5	4				
15	brown,-1					4	4				
16	franson					1	1				
17	mihara,					1	1				
18	staffor					1	1				
19	paul,-e					1					
20	pacovsk					1					
21	bollich							5	5		
22	dunigan							5	6		
23	hicks,-							2	2		
24	millhol							2	2		
25	habetz,							1	1		
26	griffin							1	1		
27	kitchen							1	1		
28	harvill							1	1		
29	mitchel							1	1		
30	mitchel1							1	1		
31	newsom,							1	1		
32	taylor,1							1	1		
33	rabb,-j								1		
34	frazier								1		
35	william										
36	bauer,-										
37	brockwe										

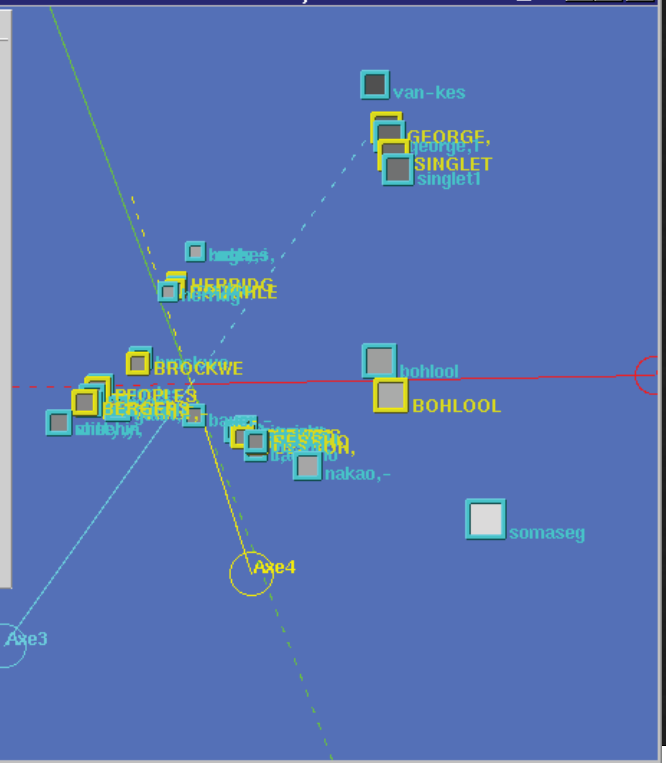


TETRALOGIE PROJECTION 4D FICHIER: soja/AL-AC/col1/AFC_I...

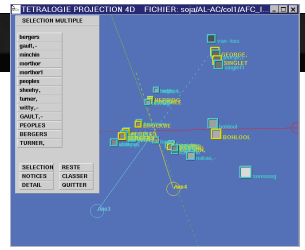
SELECTION MULTIPLE

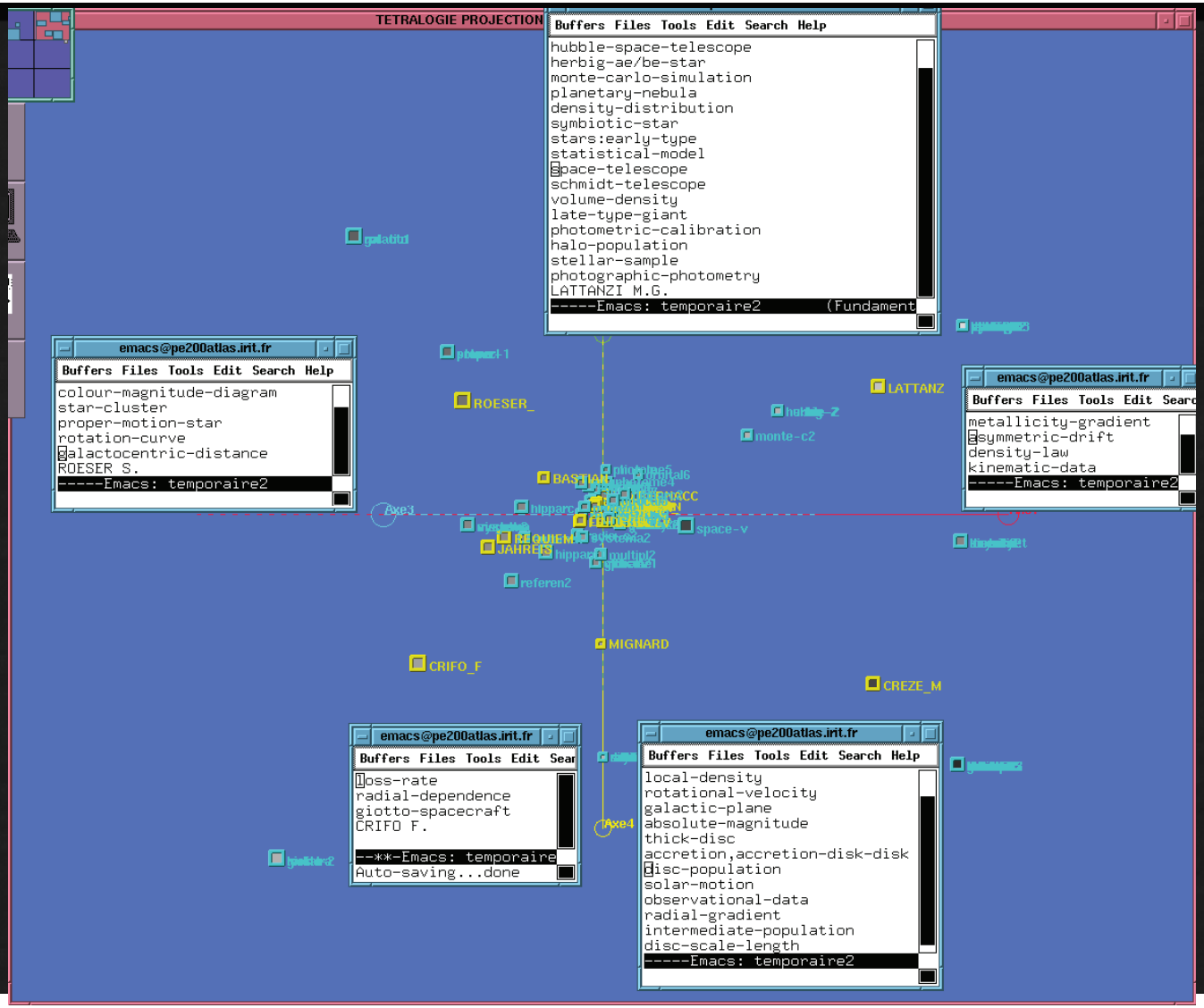
bergers
gault,-
minchin
morthor
morthor1
peoples
sheehy,
turner,
witty,-
GAULT,-
PEOPLES
BERGERS
TURNER,

SELECTION | RESTE
NOTICES | CLASSER
DETAIL | QUITTER



TÉTRALOGIE



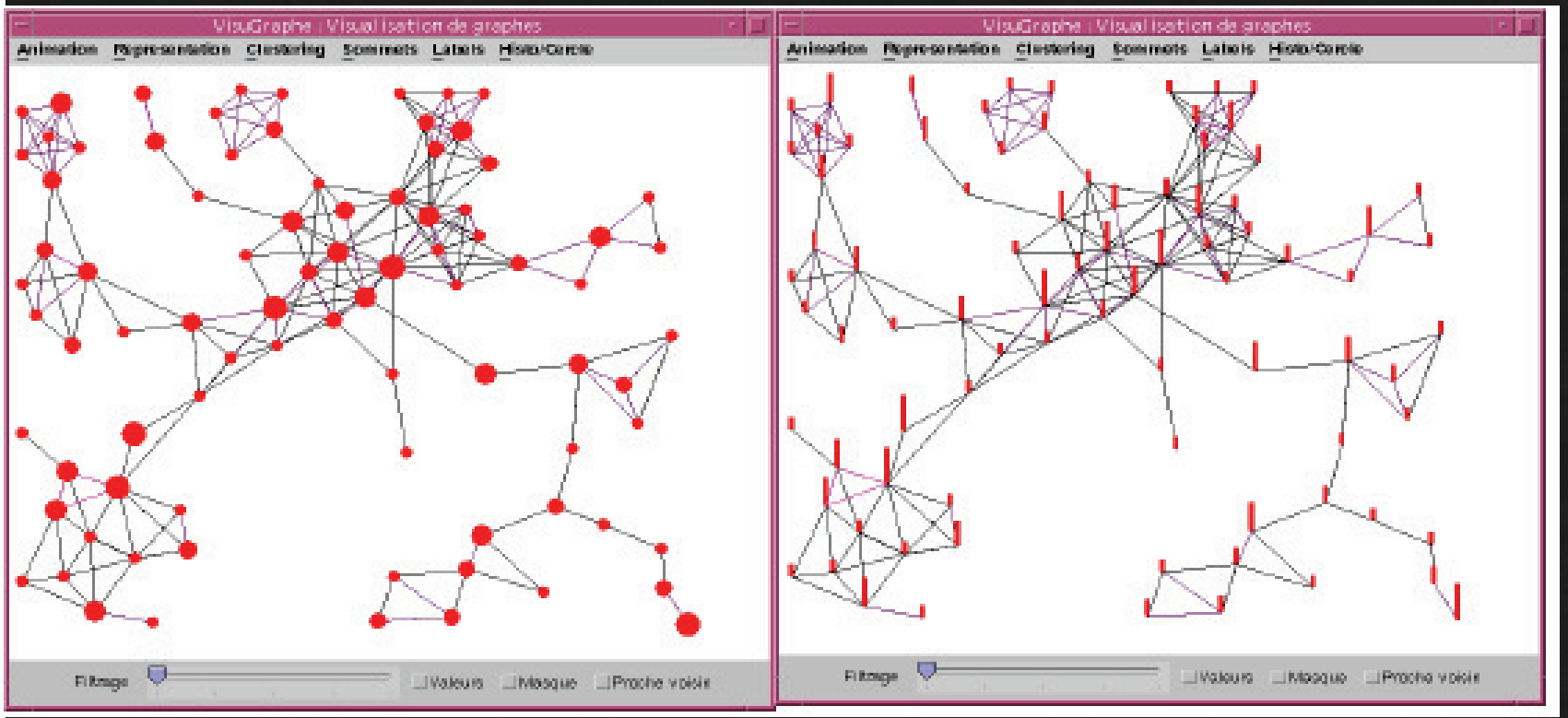


© Dousset

GRAPHS AND COLLABORATIVE NETWORKS

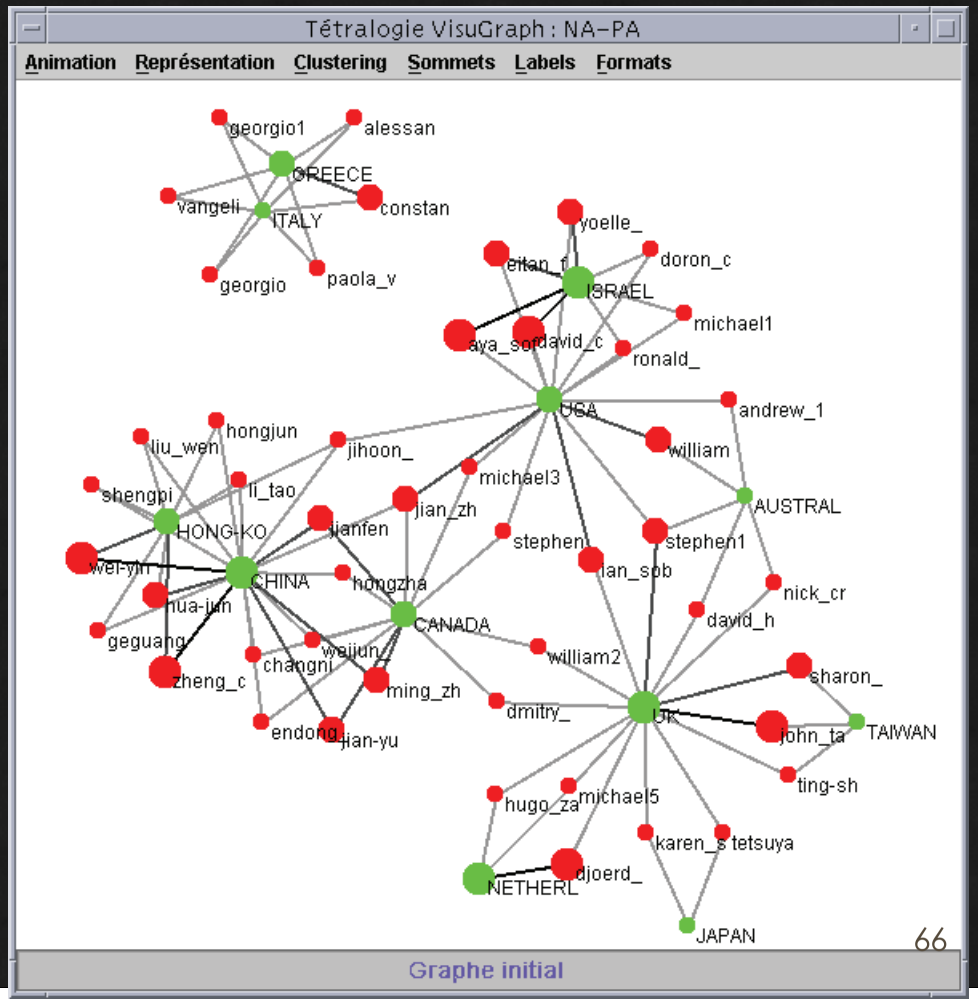
- Publication co-authors

[Dousset, 2012]



GRAPHS AND COLLABORATIVE NETWORKS

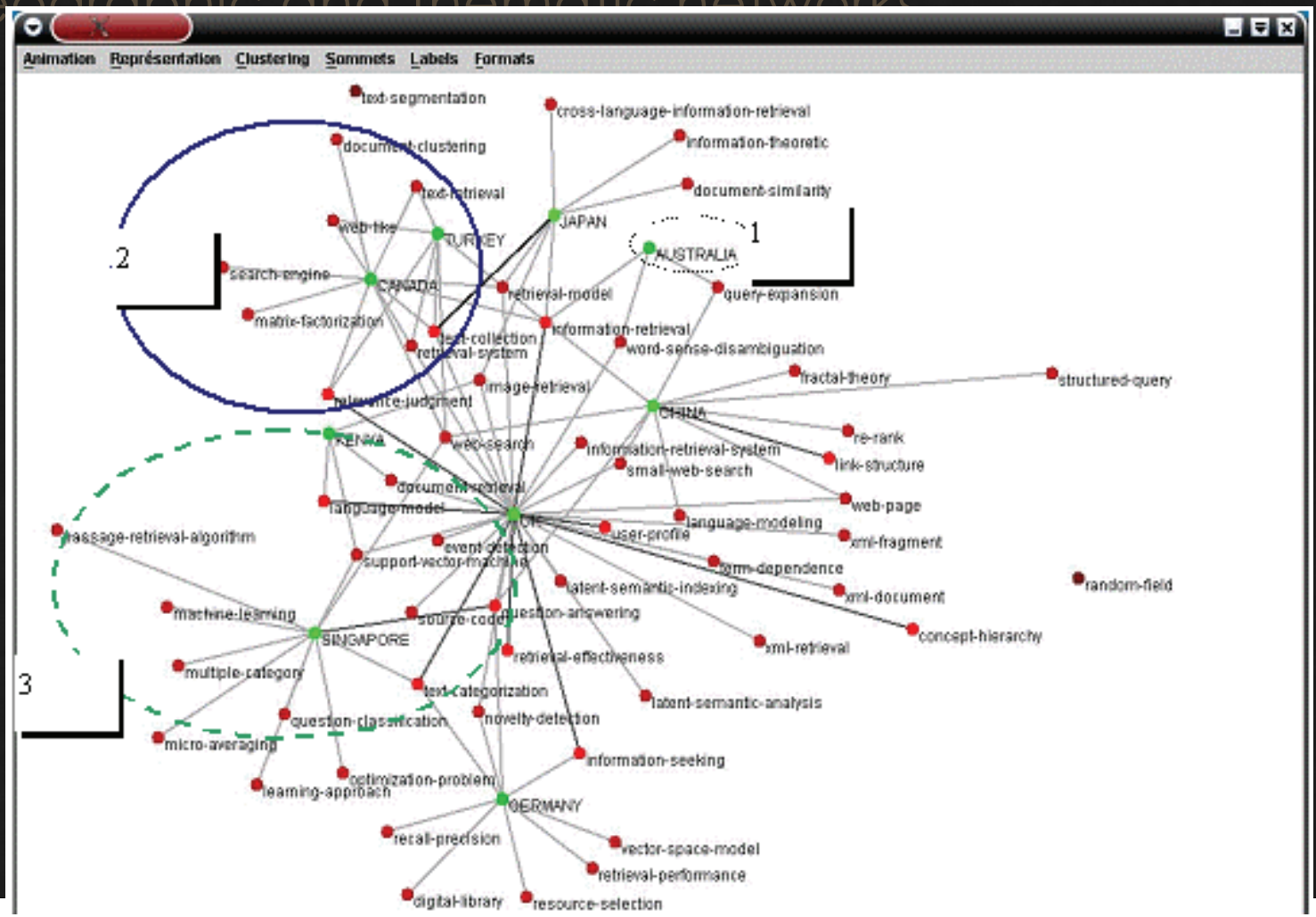
- Geographic networks



[Mothe et al.]

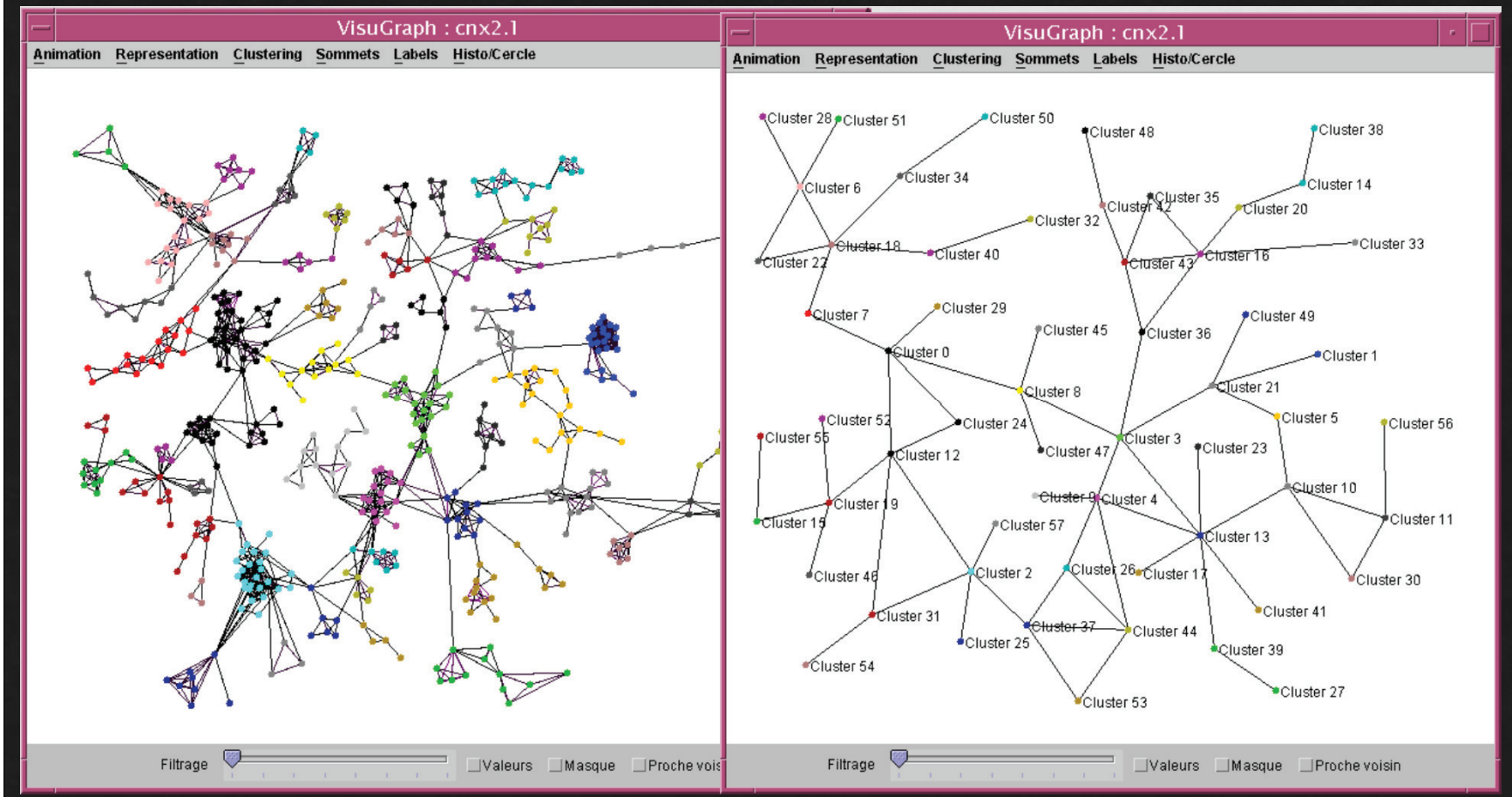
GRAPHS AND COLLABORATIVE NETWORKS

- Geographic and thematic networks



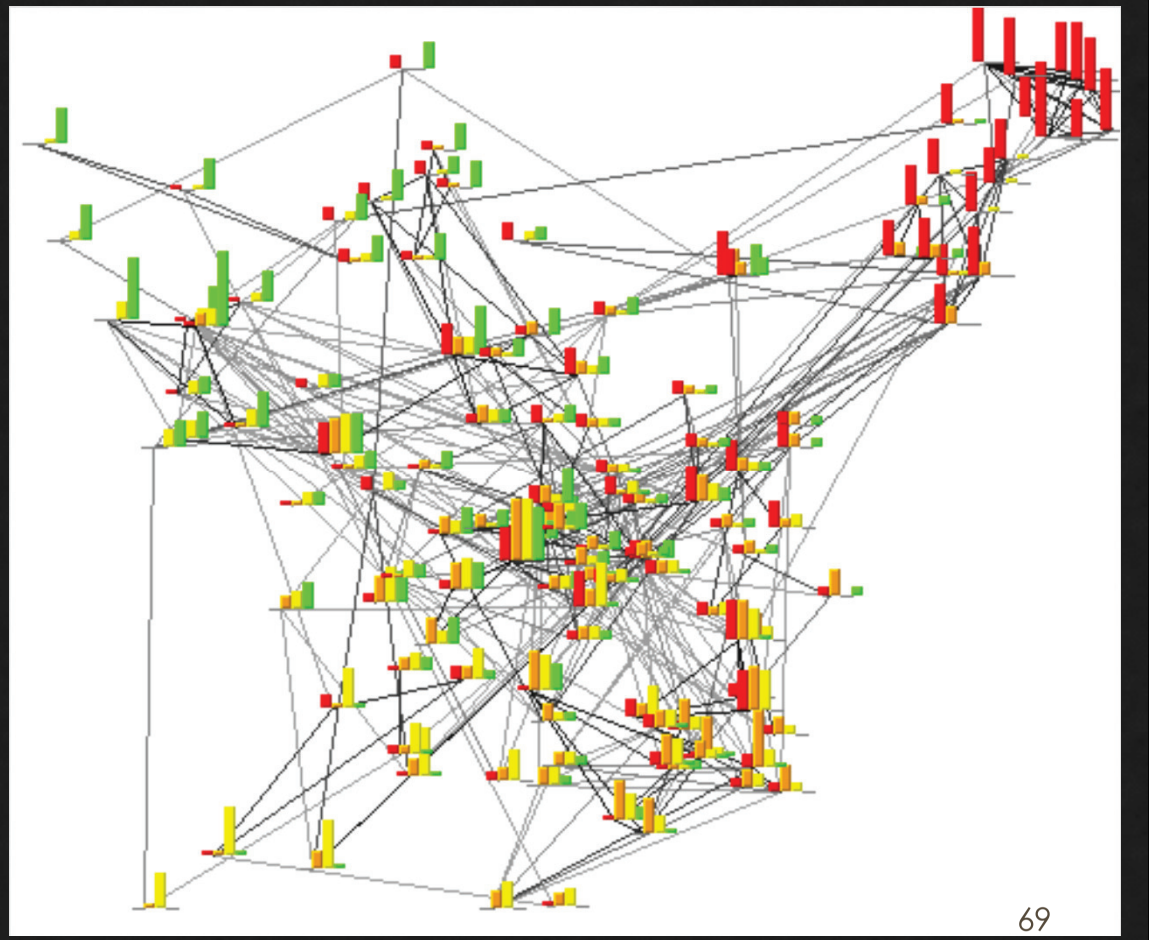
GRAPHS AND COLLABORATIVE NETWORKS

- Large networks



GRAPHS AND COLLABORATIVE NETWORKS

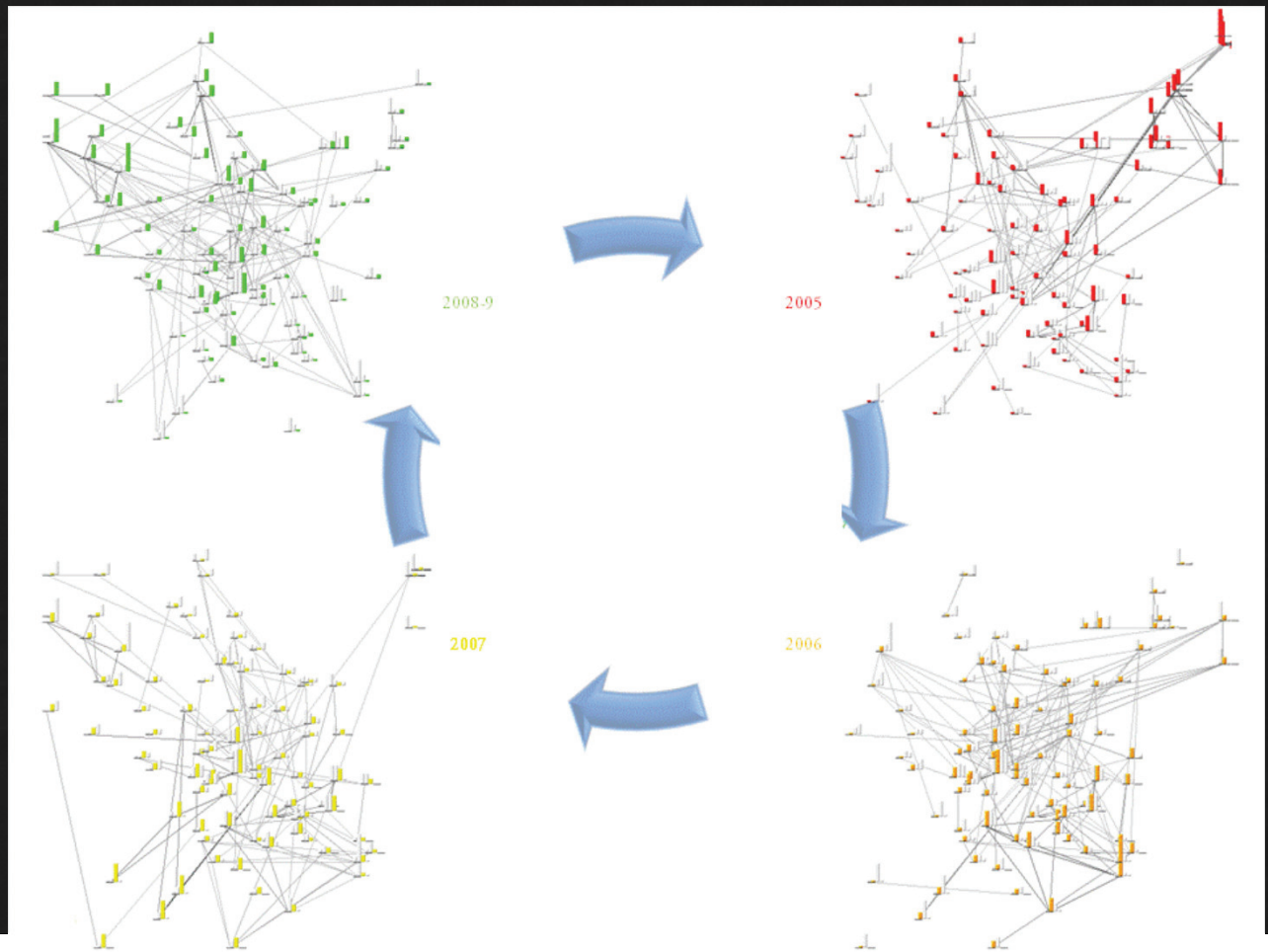
- Temporal networks



[Loubier, 2009]

GRAPHS AND COLLABORATIVE NETWORKS

- Temporal networks

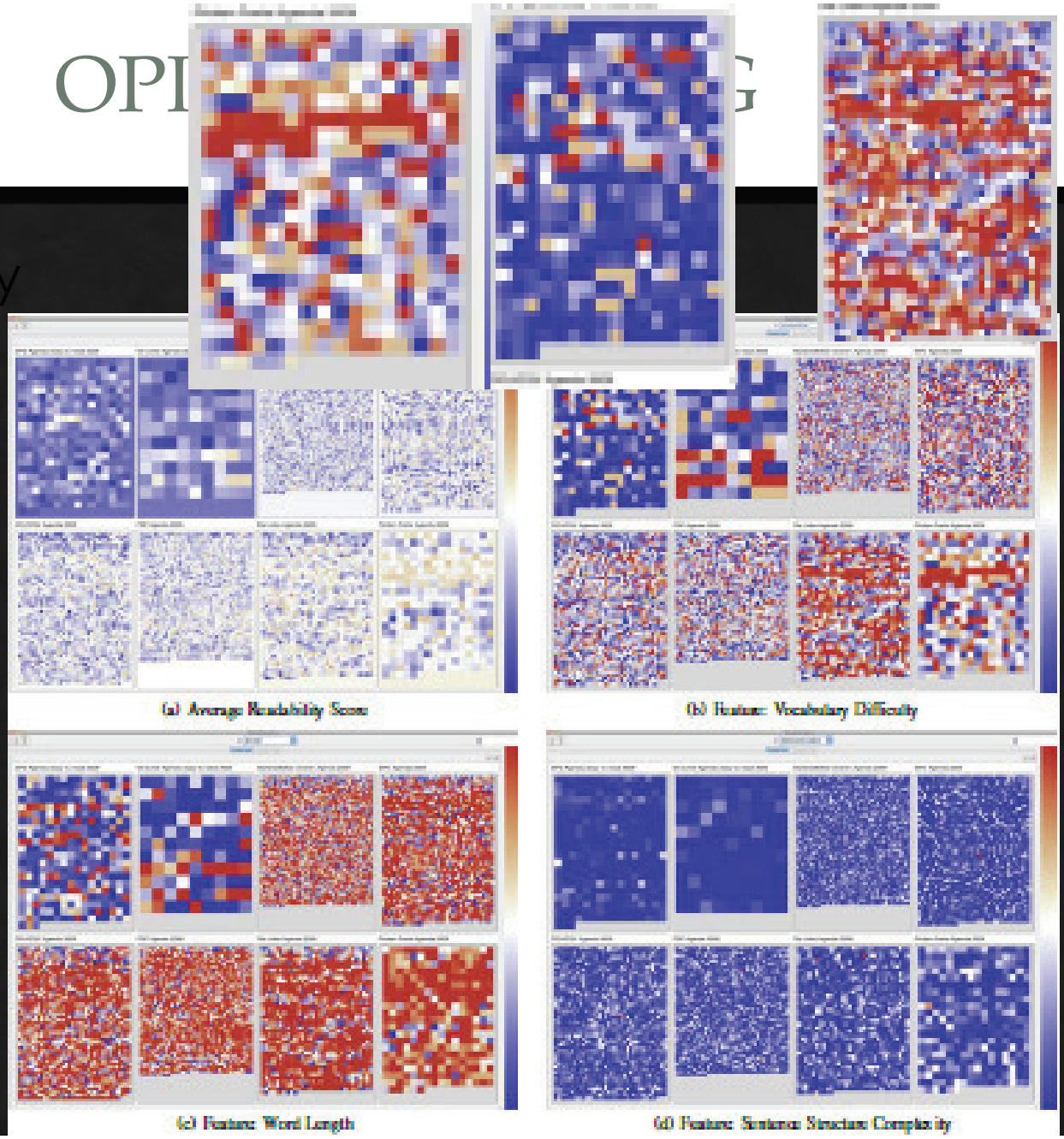


[Loubier, 2009]

TÉTRALOGIE AND LANGUAGES

OPI G

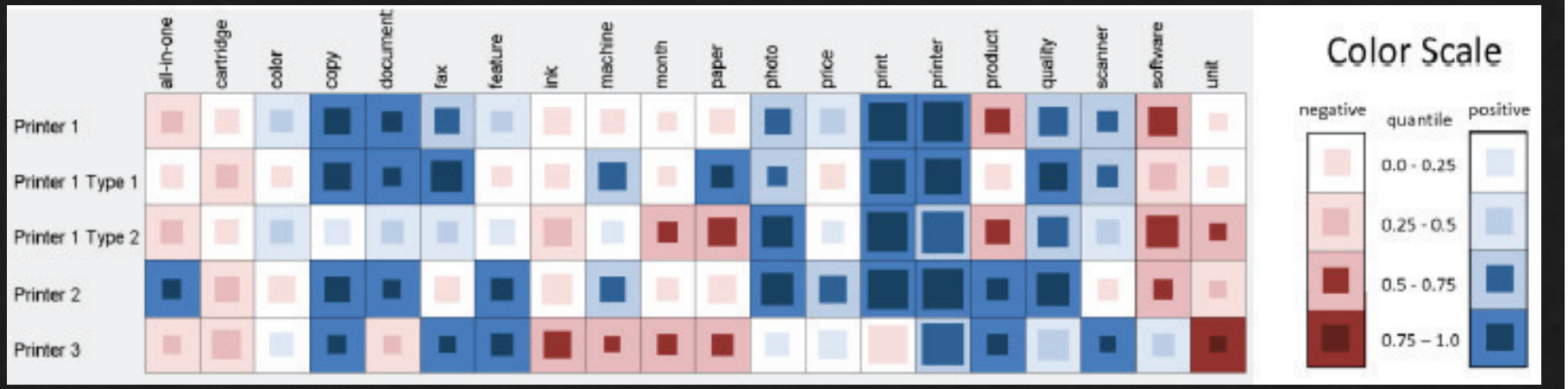
- Readibility



[Oelke et al., 2012]

OPINION MINING

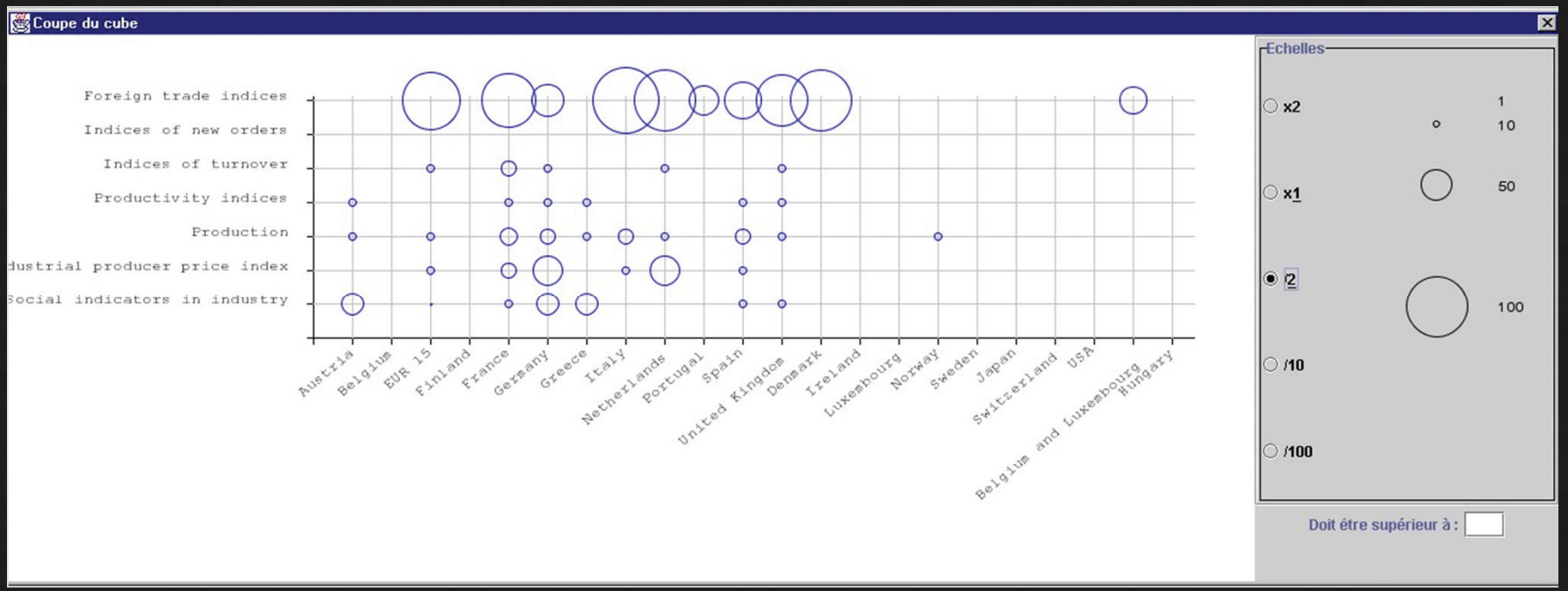
- On products [3]



[Oelke et al., 2009]

DOCCUBE

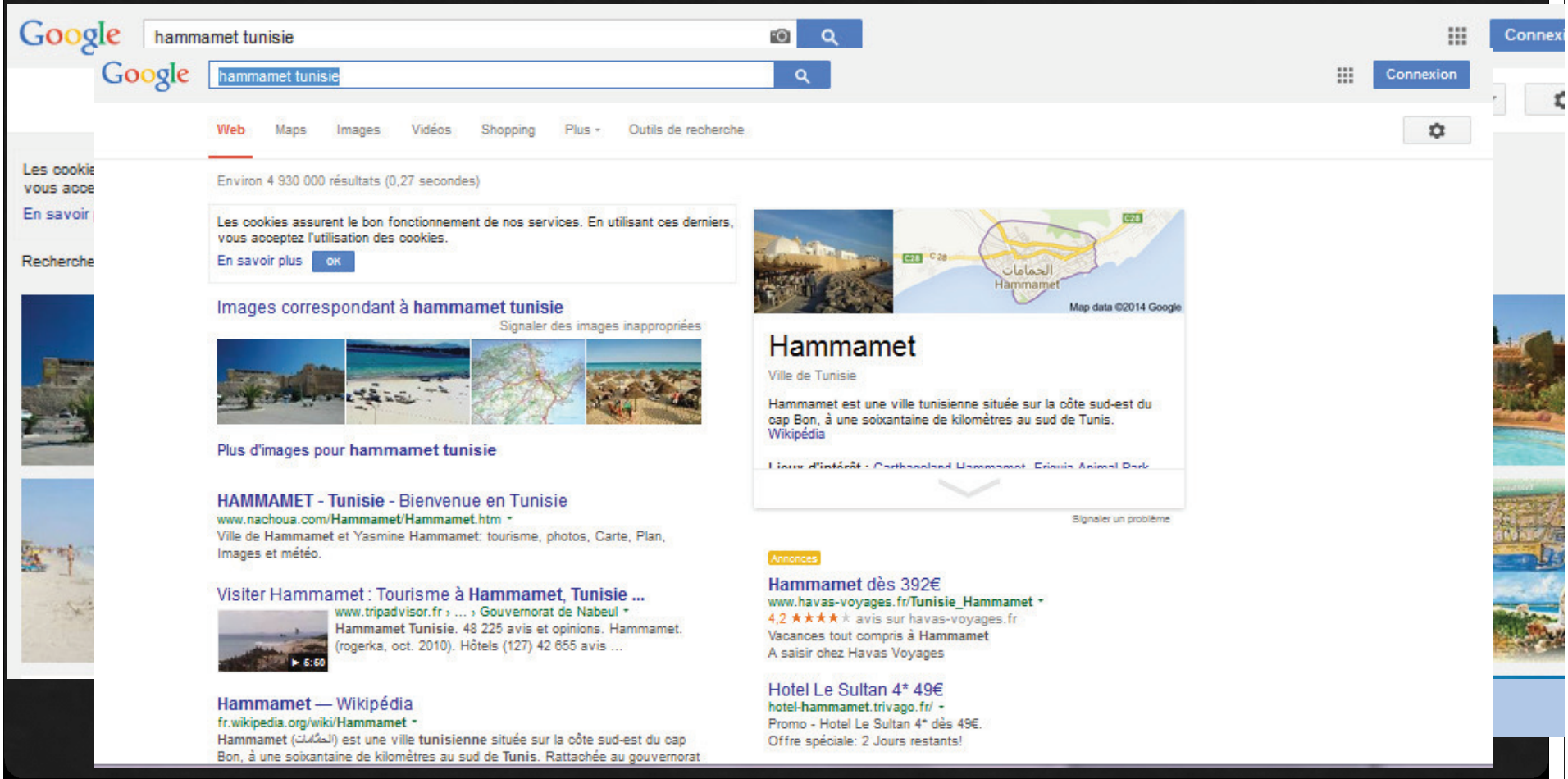
- OLAP visualization
  - Operator: slice



DocCube

INFORMATION RETRIEVAL AND MINING

- Variety



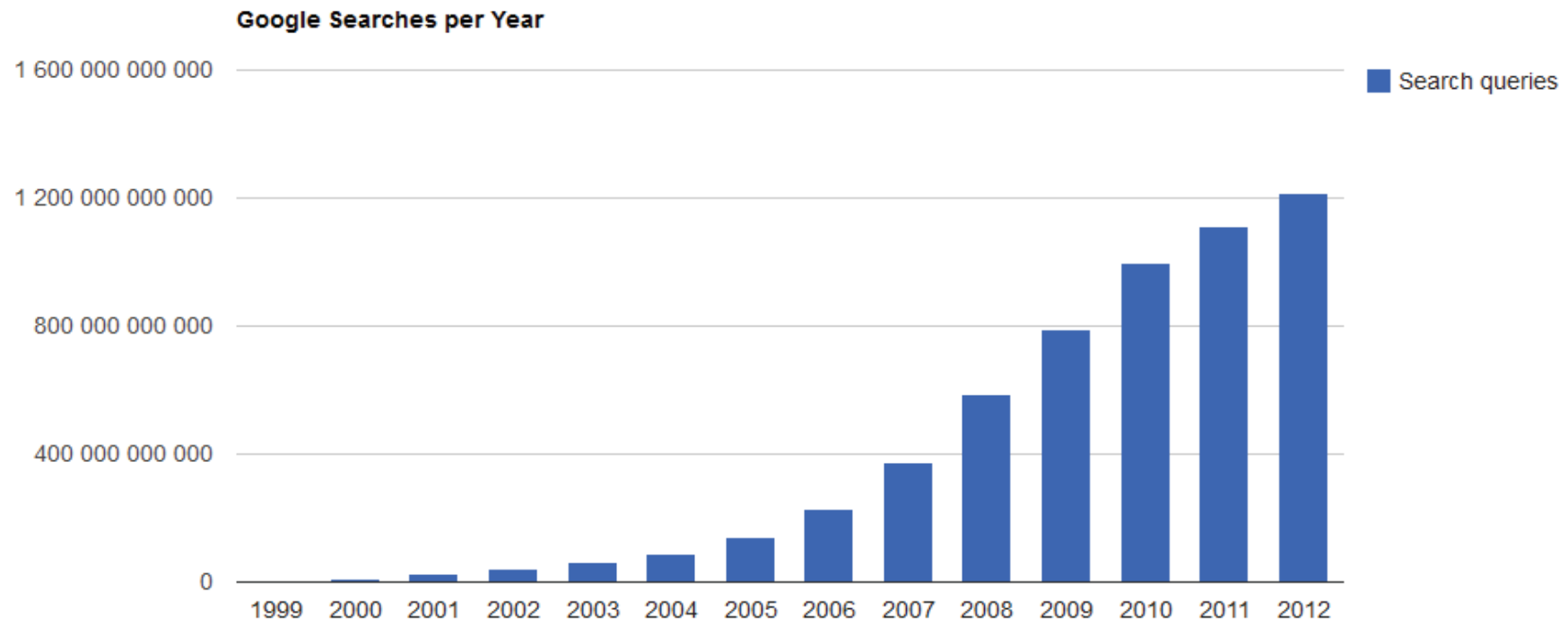
INFORMATION RETRIEVAL AND MINING

- Velocity and flows

Twitter Company Statistics	Data
Total number of active registered Twitter users	645,750,000
Number of new Twitter users signing up everyday	135,000
Number of unique Twitter site visitors every month	190 million
Average number of tweets per day	58 million
Number of Twitter search engine queries every day	2.1 billion
Percent of Twitter users who use their phone to tweet	43 %
Percent of tweets that come from third party applicants	60%
Number of people that are employed by Twitter	2,500
Number of active Twitter users every month	115 million
Percent of Twitters who don't tweet but watch other people tweet	40%
Number of days it takes for 1 billion tweets	5 days
Number of tweets that happen every second	9,100

http://www.statisticbrain.com/twitter-statistics/

INFORMATION RETRIEVAL AND MINING



http://www.internetlivestats.com/google-search-statistics/

INFORMATION RETRIEVAL AND MINING

- Velocity and flows

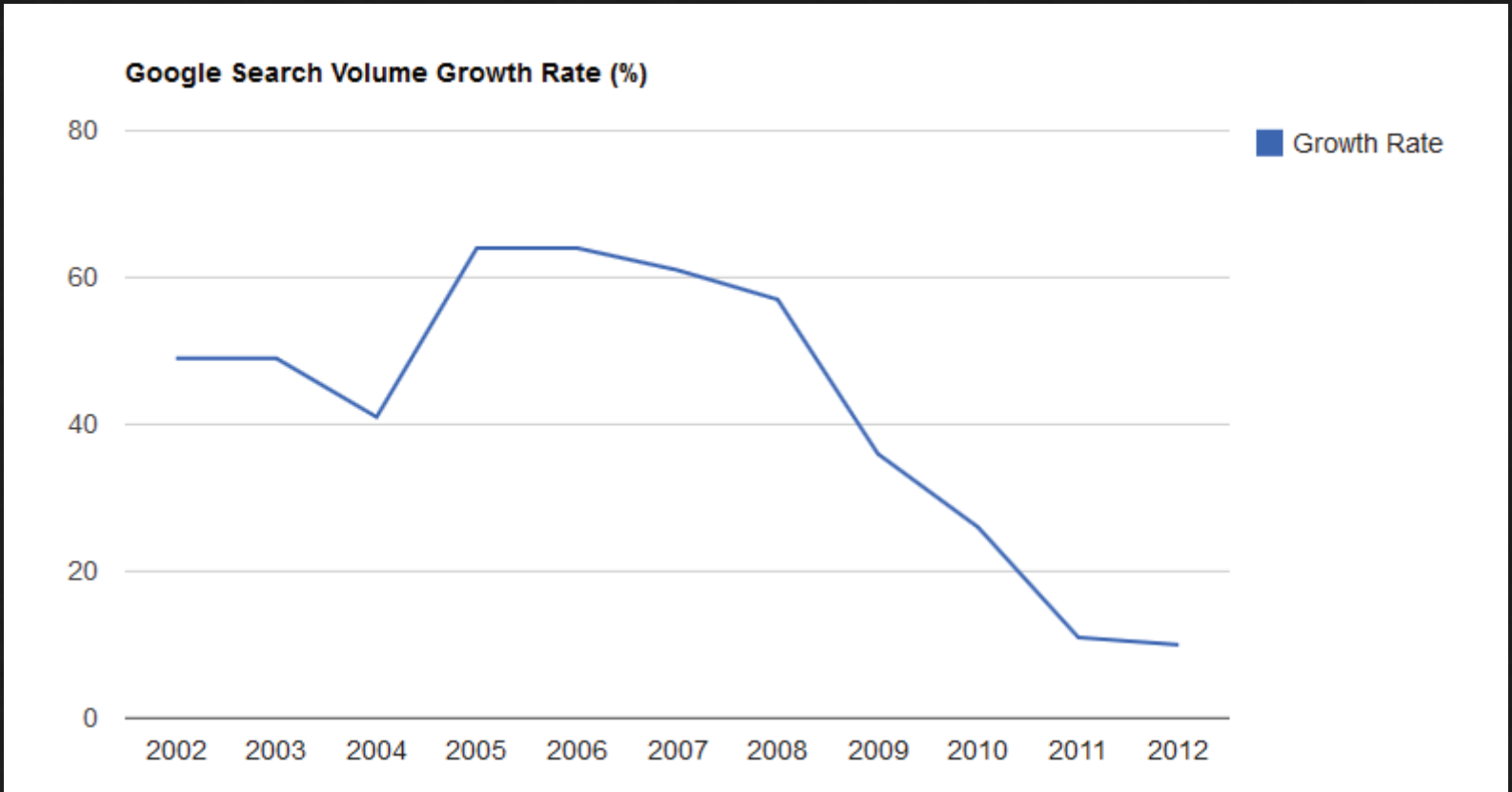
Curious facts

- In 1999, it took Google one month to crawl and build an index of about 50 million pages. In 2012, the same task was accomplished in less than one minute. [4]
- 16% to 20% of queries that get asked every day have never been asked before. [4]
- Every query has to travel on average 1,500 miles to a data center and back to return the answer to the user. [4]
- A single Google query uses 1,000 computers in 0.2 seconds to retrieve an answer. [7]

http://www.internetlivestats.com/google-search-statistics/

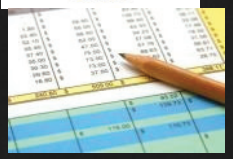
INFORMATION RETRIEVAL AND MINING

- Velocity and flows



http://www.internetlivestats.com/google-search-statistics/

SCIENTIFIC ORIENTATIONS FOR THE FUTURE



Indexing

Indexing

Indexing

Indexing

Indexing

MSC. ERMAKOVA CONGRES EGC 2013	300,00
MSC. LEVA CONGRES EGC 2013 TLSE	350,00
PRESTATION DE CONSEIL DONT APPRO	1 010,00
Inscription INFORSID2013 RIGOTANTH	440,00
RIGOT ANTHONY PARIS-BORDEAUX 27	137,80
Conférence ICEIS 2013	595,00

collaboration	URL2
information	URL2
information	URL1
national	URL1
partenaire	URL1
recherche	URL1

Analyzing

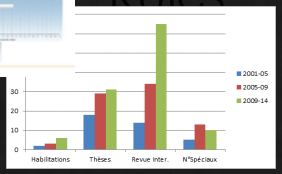
Selecting

Ceci est un texte court qui correspond à un résumé il resprésente juste ce que pourrait donner un résumé automatique de texte.

Analyzing

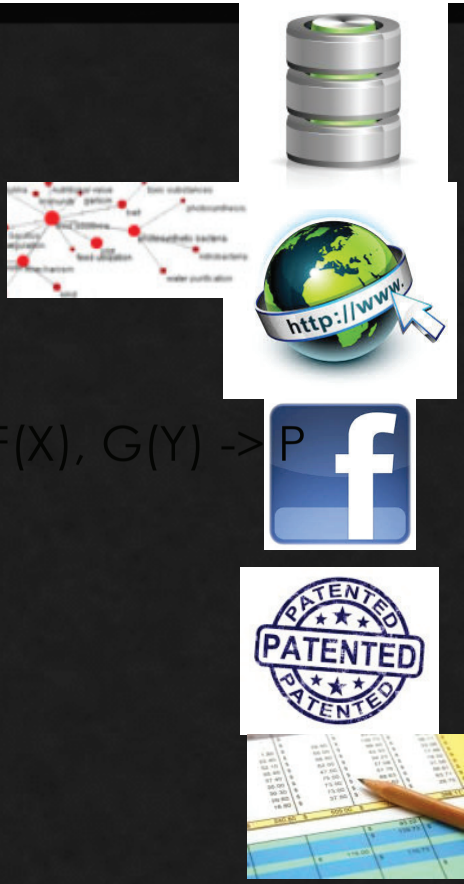
Selecting

Rules



Ceci est un texte court qui correspond à un résumé il resprésente juste ce que pourrait donner un résumé automatique de texte.

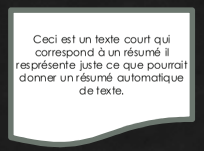
SCIENTIFIC ORIENTATIONS FOR THE FUTURE

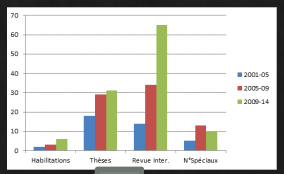


F(X), G(Y) -> P

Representing/ Analysing/ Aggregating/Visualizing

F(X), G(Y) -> P





SCIENTIFIC ORIENTATIONS FOR THE FUTURE

- Semantic search
- Selective approaches
  - Selective Query Expansion
  - Other types of selective approaches
    - Based on difficulty
    - Based on type (navigational, transactional)
- Fusing more specialized search engines
  - Agregate search

INFORMATION RETRIEVAL AND MINING & BIG DATA

JOSIANE MOTHE

ASD - 2014