

# Analyse des traces numériques

**Introduction à la cartographie de l'information  
scientifique et de la presse**

# Objectifs et plan de la séance

- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- Introduction à la visualisation de l'information
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets

# Objectifs du cours

Quatre principaux objectifs :

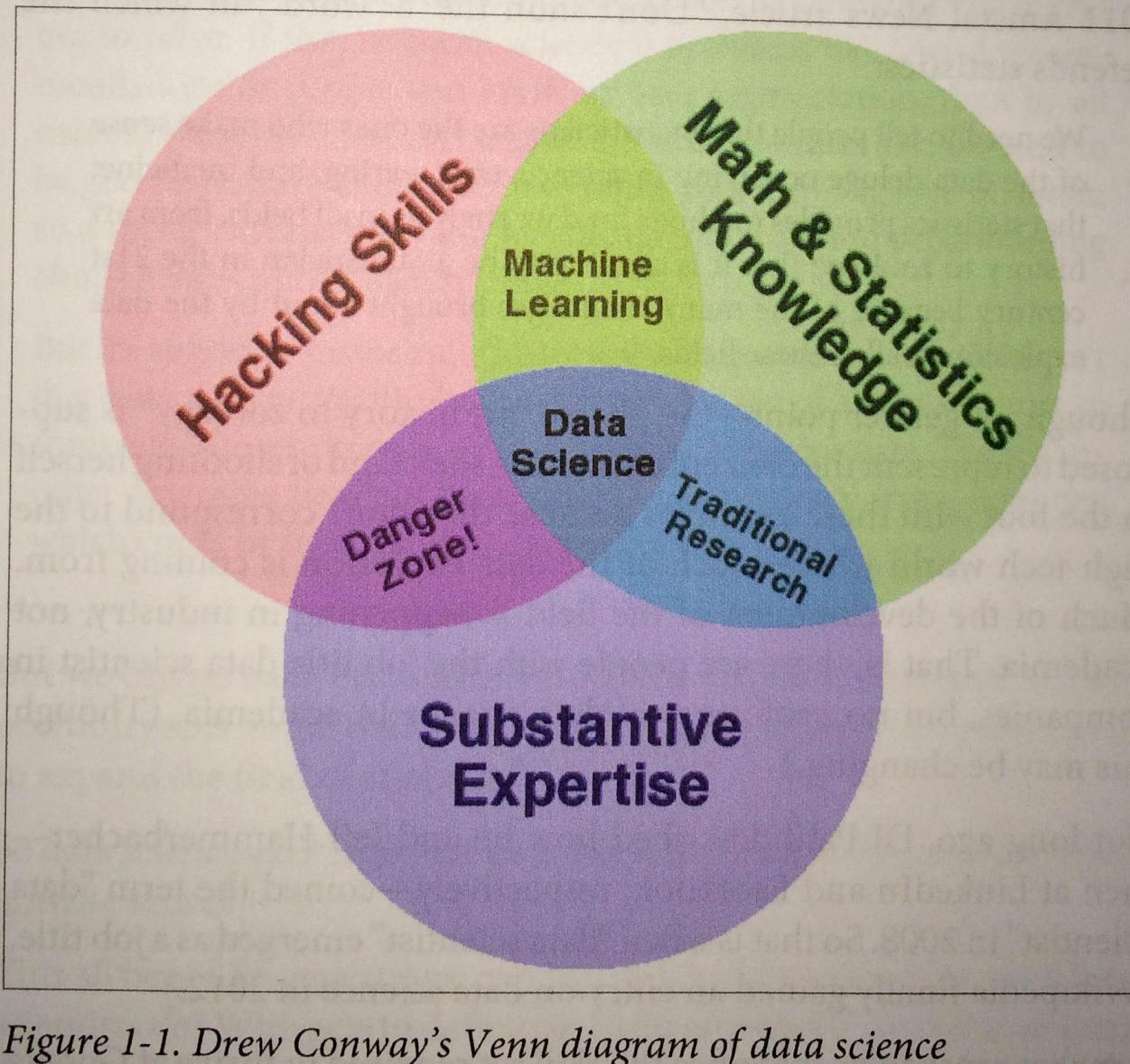
- Sensibilisation à **l'analyse de données** relatives au développement de la science et/ou au traitement médiatique de sujets débattus ;
- Cerner quelles sont **les origines** du développement de cet espace de recherche où les traitements informatiques sont au cœur des analyses conduites pour comprendre les dynamiques sociales et scientifiques ;
- Initiation aux **outils de traitements** de données textuelles et de réseaux ;
- Découvrir les notions de bases et outils de la **visualisation de l'information**.

La **cartographie de l'information ou visualisation de l'information** suppose trois ensembles de compétences :

- **Sciences de l'ingénieur** : statistiques et data mining, data management
- **Designer** : design de l'information, visualisation et infographie
- **Analyste** : expertise sur le champ étudié, interprétation

On parle souvent de **Data scientist** pour celui qui maîtrise une partie de ces trois ensembles de compétences.

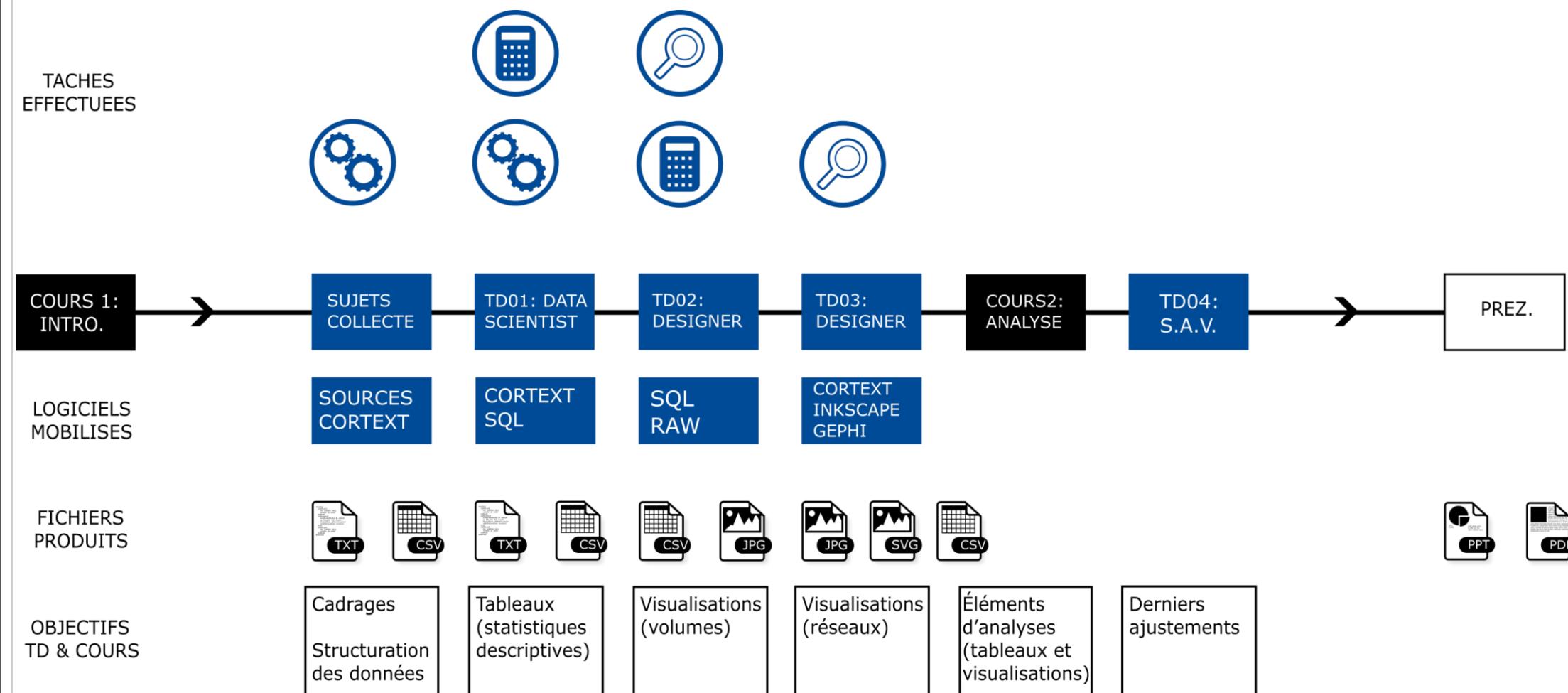
Driscoll then refers to Drew Conway's Venn diagram of data science from 2010, shown in Figure 1-1.



Votre force se trouve dans la capacité à **problématiser les questions** (amont) et à mettre **en relief les résultats** (aval).

Figure 1-1. Drew Conway's Venn diagram of data science

# Principaux jalons du cours et évaluation



# Plan des séances et évaluation

Séquencement (par groupes de 3) :

- Cours 01 (3h) : introduction, de la scientométrie aux humanités numériques
- Cours TD 02 (3h) : data scientist (tableaux descriptifs) et designer (tableau)
- Cours TD 03 (3h) : analyse et data scientist (réseaux CM et mesures)
- Cours TD 04 (3h) : designer (cartes et réseaux / CM, Gephi et Inkscape)
- SAV (4h) : SAV (préparation du rendu)
- **Présentation des analyses** (2h) : 15 min de prez par groupe et 5 minutes de questions

# Objectifs et plan de la séance

- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- Introduction à la visualisation de l'information
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets

# Les articles scientifiques comme source d'information ?

A ce titre, un **article scientifique** est considéré comme un indicateur important de la production de la recherche scientifique (mais pas le seul).

Les « **connaissances certifiées** » sont des connaissances qui ont été soumises à la critique des collègues et qui ont résistées à leurs objections (Callon, 1993).

Dès 1962, Derek de Solla Price identifie des lois générales caractérisant l'activité des scientifiques en appliquant aux articles scientifiques des **analyses quantitatives** (documents pour comprendre des dynamiques scientifiques et sociales).

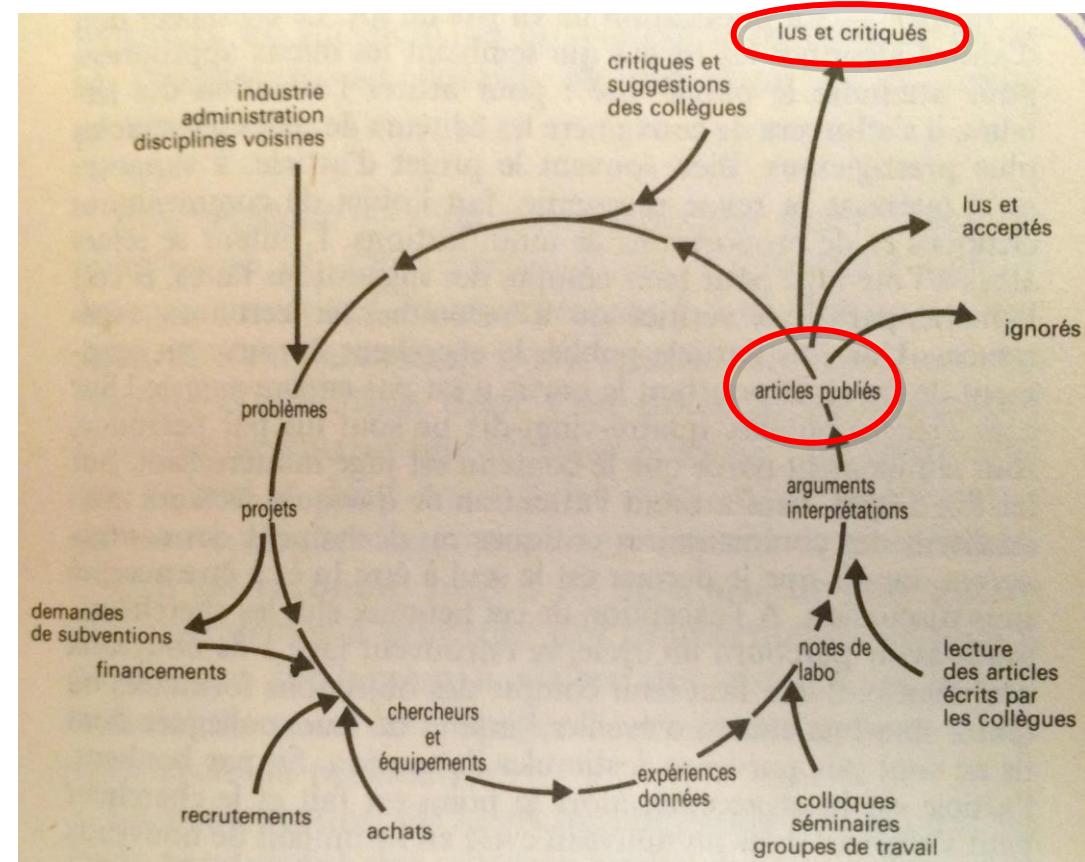


Fig. 2. — Le cycle de production des connaissances certifiées.



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Research Policy 36 (2007) 893–903



Journal

## Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking<sup>☆</sup>

Titre : haut niveau de synthèse sur le contenu de l'article

Andrei Mogoutov <sup>a,\*</sup>, Bernard Kahane <sup>b,c,1</sup>

Auteurs : collaboration scientifique

<sup>a</sup> AGUIDEL, 68 Bld de Port Royal, 75005 Paris, France

<sup>b</sup> LATTs (Laboratoire Territoires, Techniques et Sociétés), CNRS/UMLV/ENPC, École Nationale des Ponts et Chaussées,  
6-8 avenue Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne La Vallée Cedex 2, France

<sup>c</sup> ISTM (Institut Supérieur de Technologie et Management), Cité Descartes, 93162 Noisy le Grand Cedex, France

Available online 23 April 2007

Date de publication : dimension temporelle

Adresses : institutions et géographie des auteurs

### Abstract

Nanotechnology, like other emerging technologies that increasingly characterize the dynamic of our era, makes specific demands on datamining to track and interpret efficiently what is happening, through publications and other scientific output. We here propose and describe a strategy based on an automated lexical modular methodology to overcome rapidly evolving content and classification problems, which may otherwise accommodate poor quality of data and expert bias, with potential dire consequences for interpretation, decision and strategy. The proposed methodology is based on an initial nanostring enriched and screened by eight subfields, automatically identified and defined through the journal inter-citation network density displayed in the initial core nanodataset. Relevant keywords linked to each subfield are then tested for their specificity and relevance before being sequentially incorporated to build a modular query. We then, as a first test, compare the database constructed using this methodology for years 2003 and 2005 with those obtained by other approaches previously used to cover and explore the nanotechnology dynamic. Finally, using the inherent transparency, portability and replicability of our methodology, we offer, in order to help our initial query evolve and develop, a set of evaluation processes for tests by researchers in the nano field, other scientometric teams and intelligence experts involved in decision-making processes.

© 2007 Elsevier B.V. All rights reserved.

Résumé : contenu de l'article

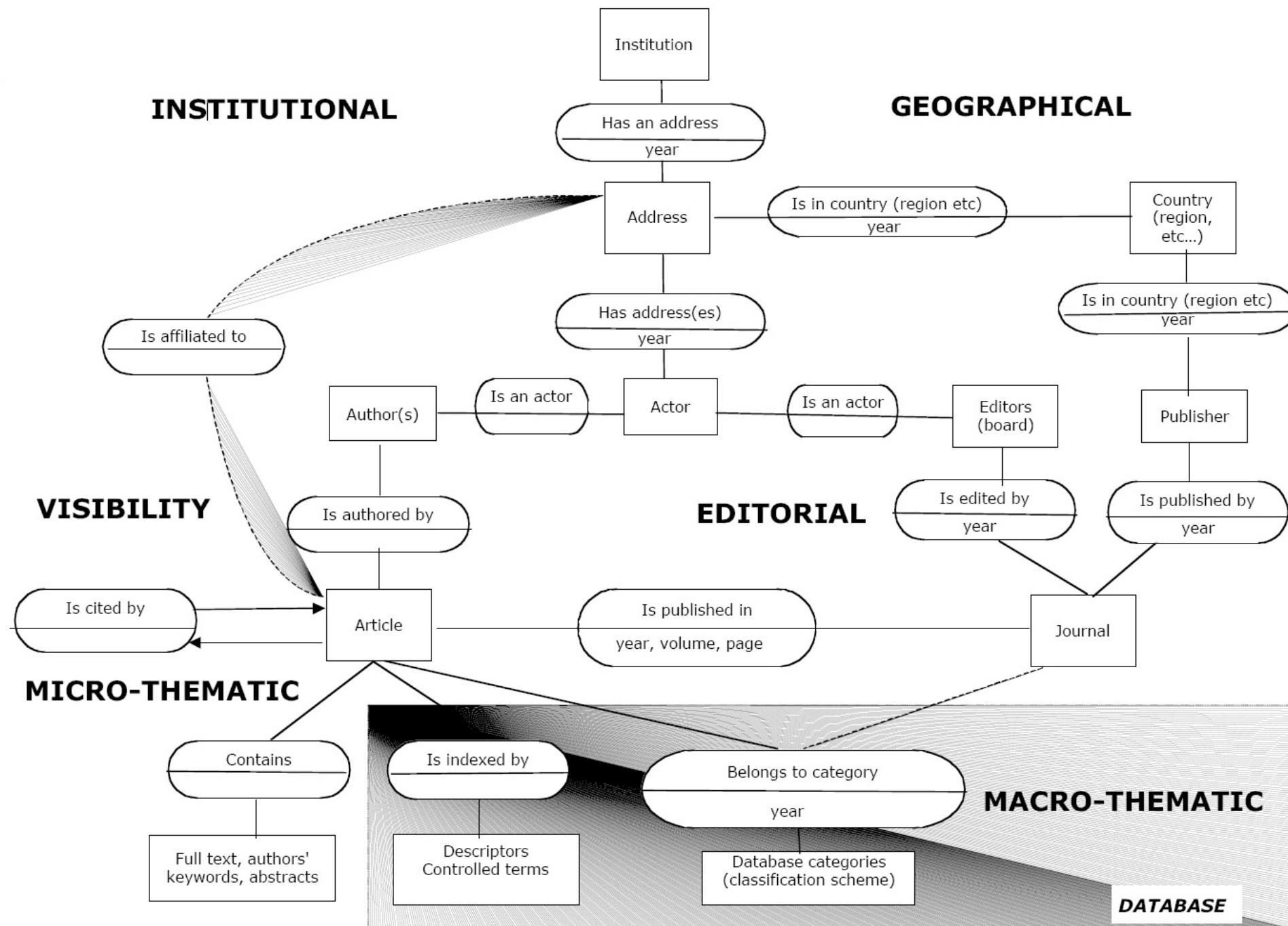
Keywords: Datamining; Nanotechnology; Emergent technologies

Mots clefs des auteurs (vision synthétique de l'article par l'auteur) :  
notions, concepts, méthodes

## References

- Cambrosio A, Keating P, Lewison G, Mercier S, Mogoutov A., in press, Mapping the emergence and development of translational cancer research; European Journal of Cancer.
- Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.K., Roco, M.C., 2003. Longitudinal patent analysis for nanoscale science and engineering: country, institution and technology field. *Journal of Nanoparticle Research* 5, 333–363.
- Noyons E.C.M., Buter B.K., Van Raan A.F.J., Schmoch U., Heinze T., Hinze S., Rangnow R., 2003, Mapping Excellence in Science and Technology across Europe, Nanoscience and Nanotechnology, Draft report of project EC-PPN CT-2002-0001 to the European Commission.
- Sampat, B.N., 2005, Examining patent examination: An analysis of examiner and applicant generated prior art., Working Paper, Columbia University.
- Zitt, M. and Bassecoulard, E., in press, “Delineating Complex Scientific Fields by A Hybrid Lexical-Citation Method: An Application to Nanosciences “Information Processing and Management”.

***Citations et références de l'article : sources scientifiques de l'article***



Zitt M.  
2004

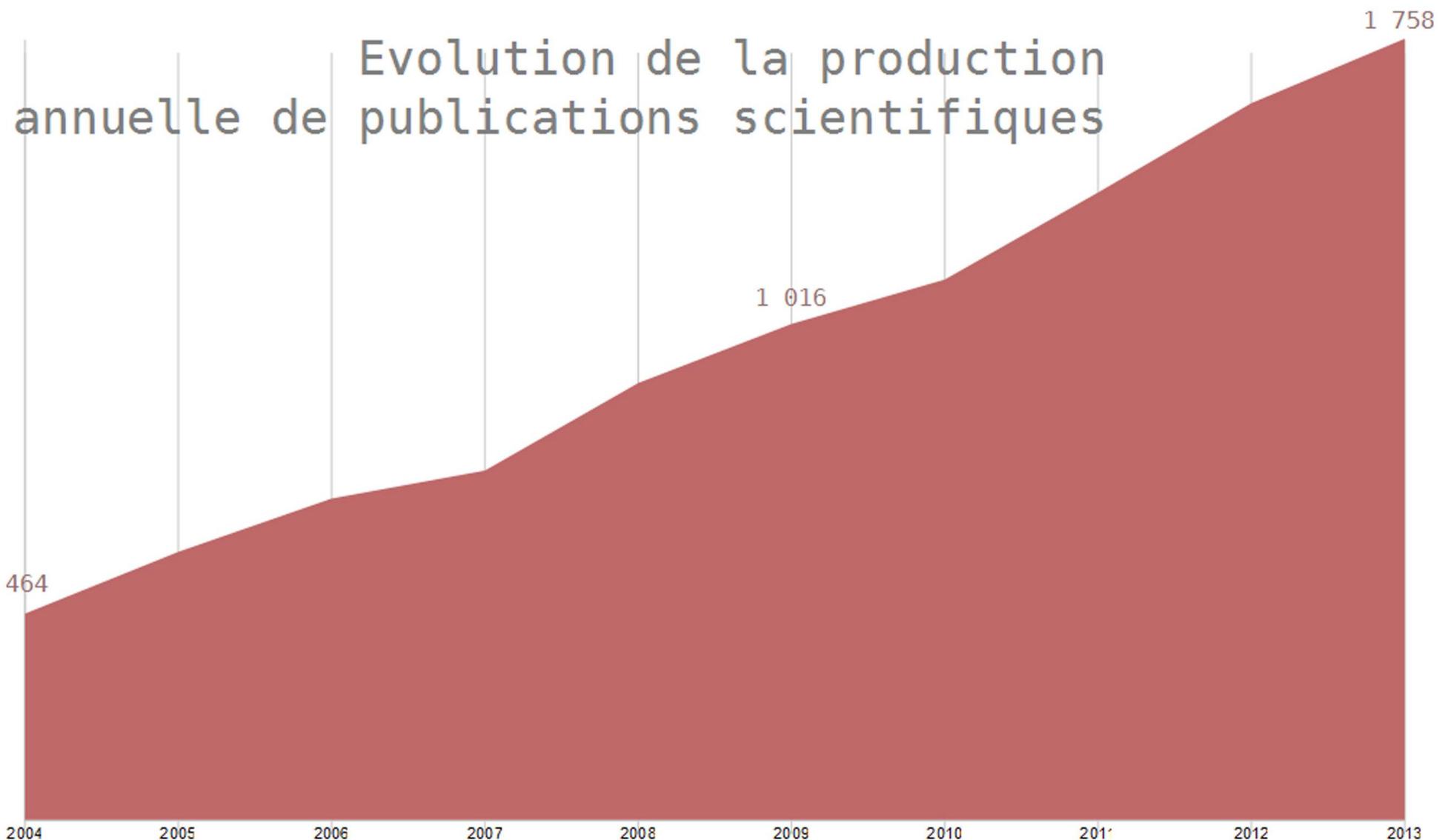
## Les indicateurs d'activité (Callon et al., 1993)

Deux grands type d'indicateurs de l'activité scientifique :

- indicateurs d'activité
- indicateurs relationnels

Les indicateurs d'activités :

- sont les plus simples
- on considère la science comme une activité ordinaire
- il s'agit bien souvent d'un comptage des publications

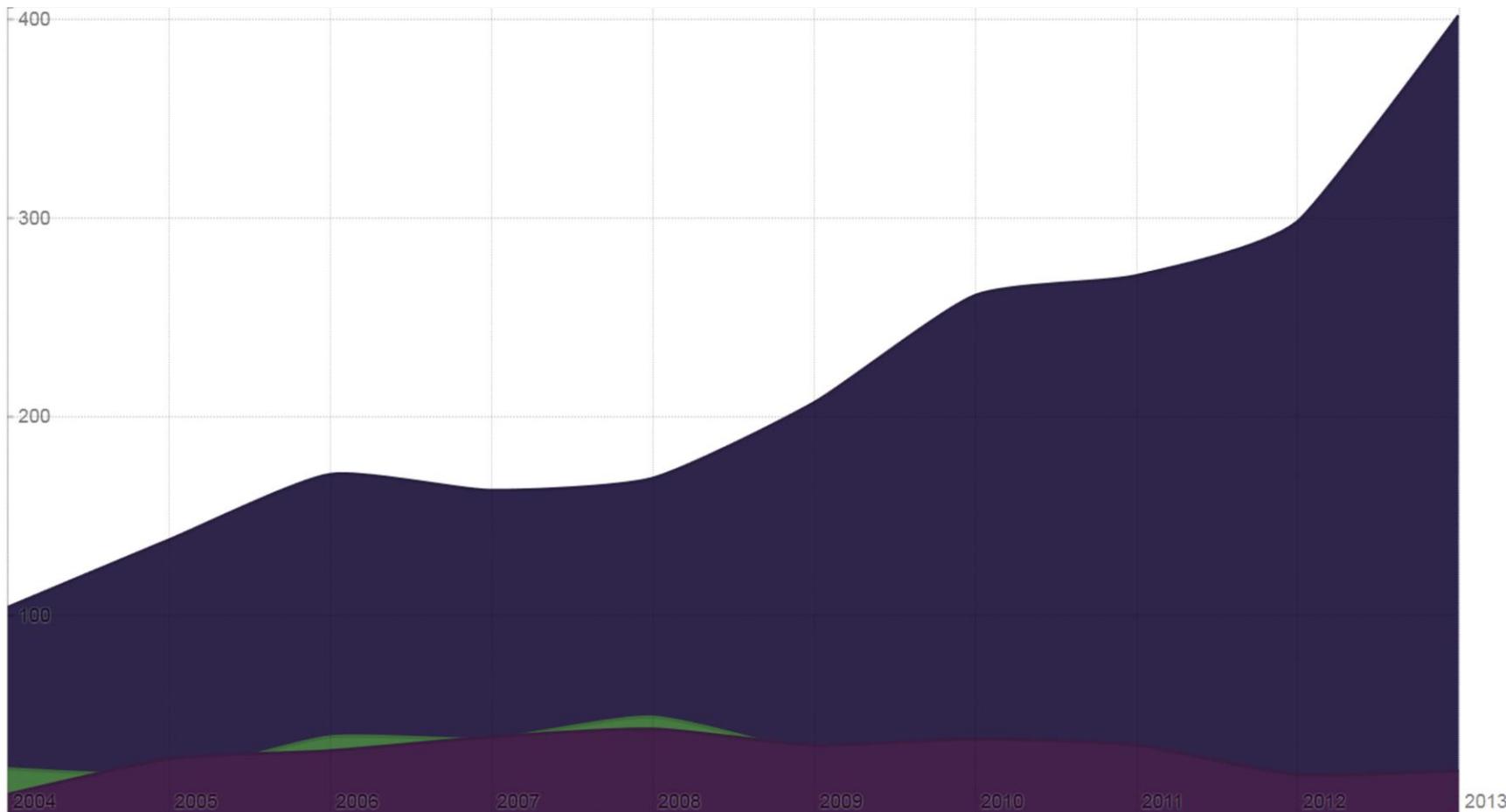


Quelle est l'évolution du champ scientifique des nanobiotechnologies en France ?

Champ en forte croissance depuis les années 2004.

### Field Evolution

- ✓ univ bordeaux 1
- ✓ univ strasbourg 1
- ✓ univ paris 07
- ✓ inra
- ✓ univ rennes 1
- ✓ aix marseille univ
- ✓ cea saclay
- ✓ european synchrotron radiat facil
- ✓ univ lyon
- ✓ univ toulouse
- ✓ univ bordeaux
- ✓ univ strasbourg
- ✓ inserm
- ✓ cea
- ✓ univ montpellier 2
- ✓ univ grenoble 1
- ✓ univ paris 11
- ✓ univ lyon 1
- ✓ univ paris 06
- ✓ cnrs



Quelle est l'évolution de la production scientifique des principaux acteurs ?

Forte croissance, avec plusieurs profils qui semblent se dessiner, principalement entre les acteurs les plus importants (croissance forte) et ceux plus marginaux (croissance nulle) dans le développement des nanobiotechnologies en France.

## Les indicateurs relationnels directs

Deux grands types d'indicateurs relationnels de l'activité scientifique :

- **directs** : relations n'entrent pas directement dans les contenus des articles (ex : adresses pour les collaborations)
- **indirects** : relations établies à partir d'une analyse du contenu des articles (ex : mots des titres, des résumés...)

## Les réseaux de collaborations

Lorsqu'on étudie un ensemble d'articles (ex : nanotechnologies), il est possible à partir des adresses des auteurs de reconstruire des réseaux de collaborations avec, par exemple :

- **Les pays** : ensemble des pays ayant collaboré avec au moins un auteur français. Cela permet dans le sujet traité de connaître quels sont les pays avec lesquels les scientifiques français ont des relations privilégiées, et avec quelle intensité.

Top	Pays	NbPublications
1	france	45712
2	usa	1894
3	germany	1721
4	italy	1689
5	spain	1499
6	uk	826
7	japan	783
8	china	655
9	belgium	614
10	poland	588
11	russia	579

### **Vue en liste**

Publications scientifiques dans le champ des nanotechnologies entre 1972 et 2015 dont au moins un des auteurs a une adresse en France.

### **Vue en matrice**

Répartition des publications de 5 pays européens (Callon et al., 1993)

Tableau 5. — Répartition des co-publications de 5 pays européens avec l'étranger (1986)

Co-publications de	Co-publications avec				Total
	Etats-Unis	CEE*	Japon	Reste du monde	
France	23,3	33,2	2,3	41,2	100
Allemagne	27,6	29,8	3,5	39,1	100
Grande-Bretagne	29,4	27,5	2,4	40,7	100
Pays-Bas	25,0	43,2	2,0	29,8	100
Italie	26,8	43,6	1,2	28,4	100
Moyenne	26,4	35,4	2,3	35,9	100

\* Estimation pour la CEE, en extrapolant à partir des 5 plus grands pays.

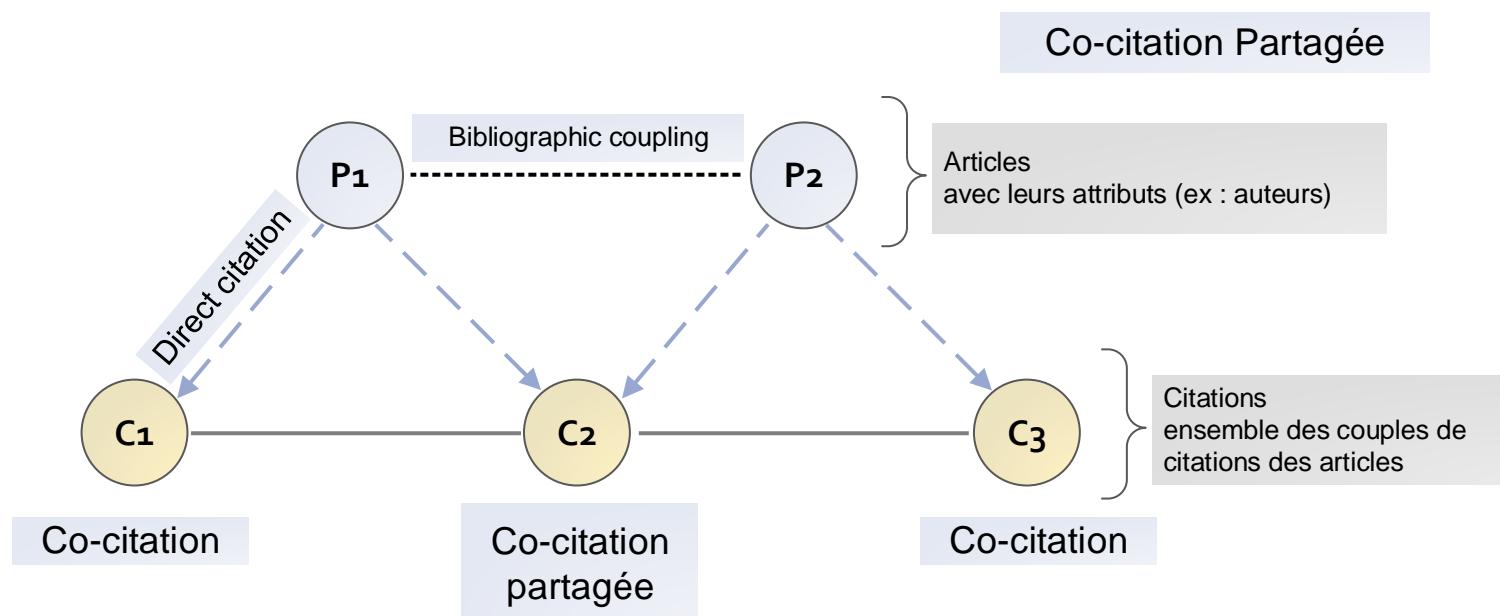
(Source : ost.)

# Identification des sources scientifiques

Le **réseau des citations** (des références d'un groupe d'articles) permet d'identifier quels sont les principaux travaux mobilisés et comment ils sont associés. C'est-à-dire les sources desquels s'inspirent les articles. Cela permet d'appréhender la visibilité par le taux de citations.

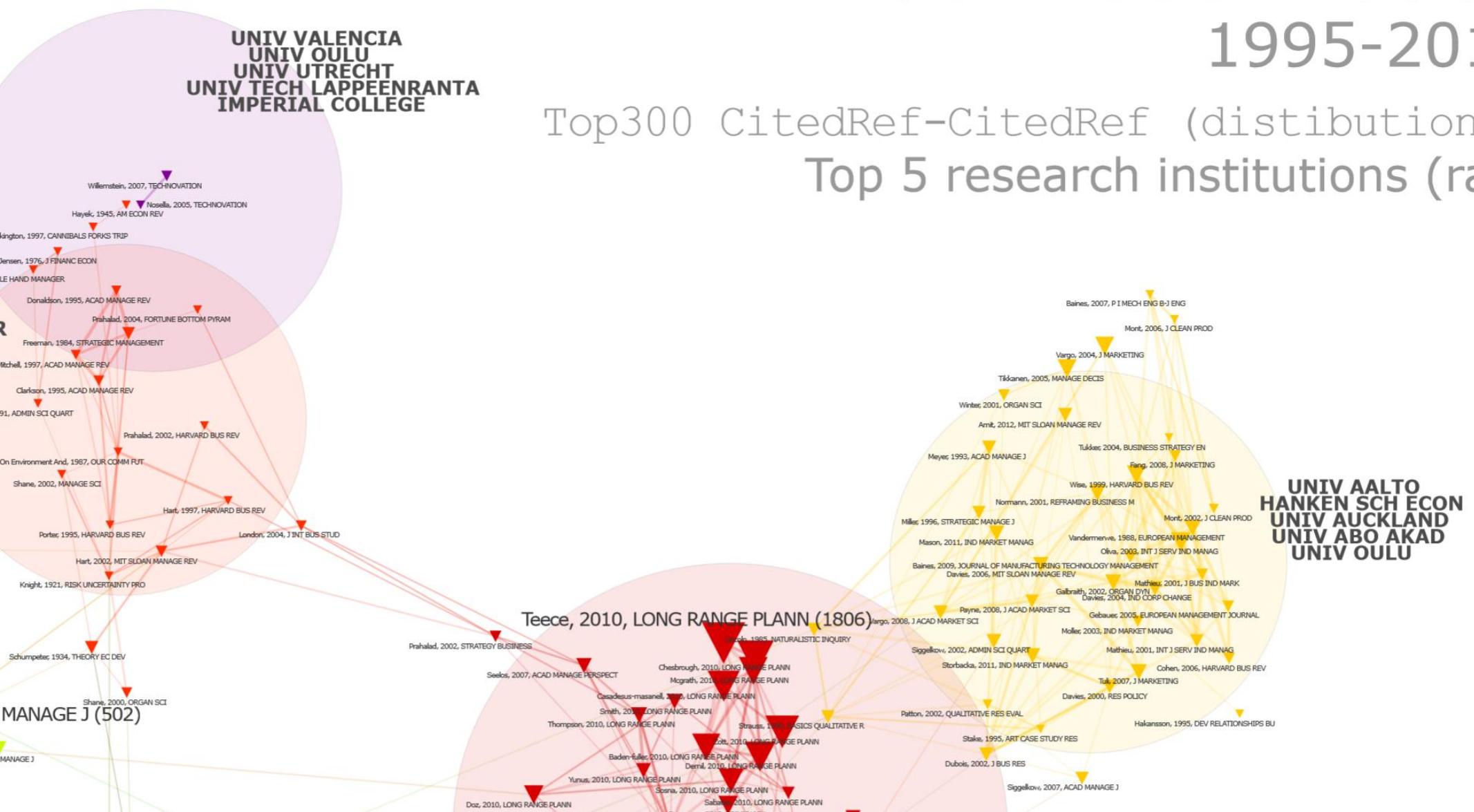
La **méthode des co-citations**, également nommé « bibliographic coupling » vise à identifier l'apparition simultanée de deux citations (couple de citations) dans l'ensemble des articles étudiés. Cette répétition dans le corpus de l'association des deux citations laisse supposer que ce couple est doté d'une signification plus précise, **plus pertinente**, que les deux citations prises indépendamment.

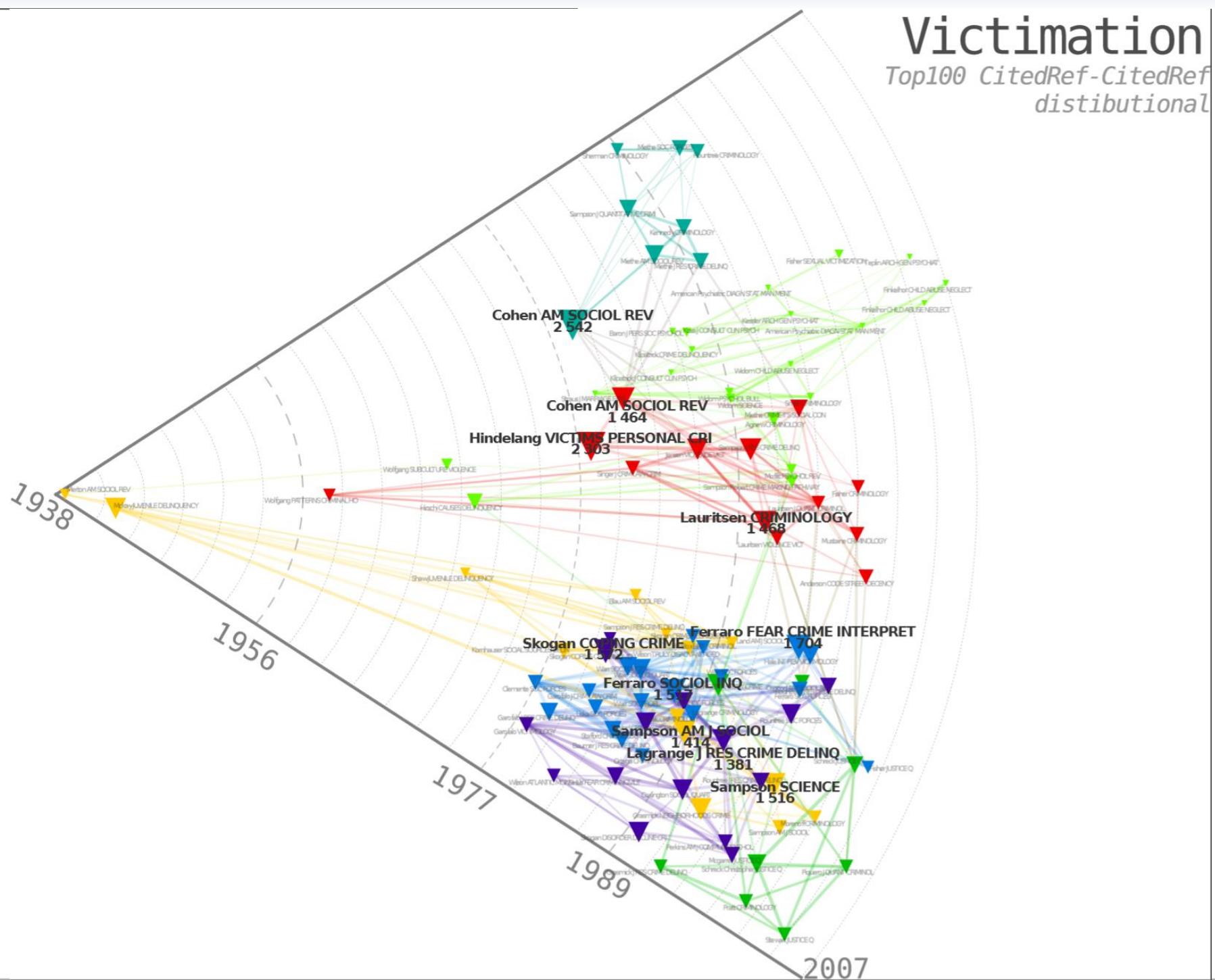
Aussi, les articles fréquemment co-cités par un groupe d'auteurs peuvent laisser supposer que ces auteurs partagent les mêmes sources scientifiques, les mêmes articles fondateurs, et peuvent témoigner d'une **communauté de chercheurs** partageant la même vision de leurs travaux.



# Business Model 1995-2014

## Top300 CitedRef-CitedRef (distributional) Top 5 research institutions (raw)





## Les indicateurs relationnels indirects

**indirects** : relations établies à partir d'une analyse du contenu des articles (mots des titres, des résumés...)

Dans la filiation de l'analyse des co-citations, il s'agit ici d'identifier les paires de mots fréquemment répétées dans les textes des articles : permet d'analyser **la signification des textes**.

Méthode du CorText Manager : à partir d'une analyse grammaticale des phrases d'un article, on identifie les **groupes nominaux**.

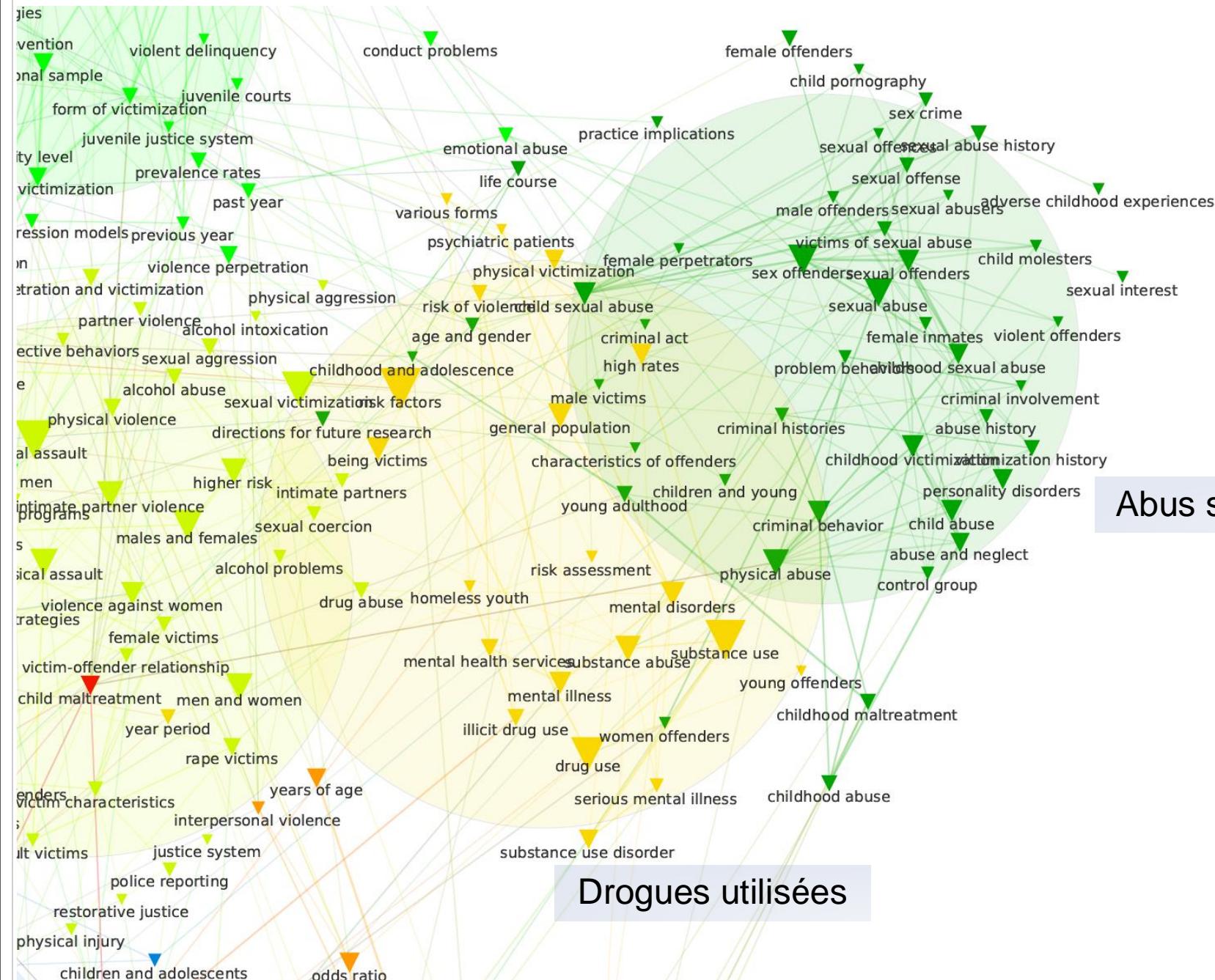
The phylogenetic position of the elephant shark (*Callorhinichus milii*) is particularly

DT JJ NN IN DT NN NN ( NNS NN )VBZ RB

relevant to study the evolution of genes and gene regulation in vertebrates.

JJ TO VB DT NN IN NNS CC NN NN IN NNS

Les groupes nominaux identifiés **co-occurrent** tous les uns avec les autres au sein d'un article. La répétition de cette opération pour l'ensemble des articles étudiés permet d'obtenir les fréquences d'apparition des couples de groupes nominaux et offre une vue synthétique des principaux thèmes traités dans l'ensemble des articles.



## *Articles sur les enquêtes portant sur la victimisation*

## Abus sexuels

## Drogues utilisées

# Objectifs et plan de la séance

- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- Introduction à la visualisation de l'information
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets

# Au delà de la scientométrie

<http://digitalmethods-seminar.org/06-03-14-mapping-st-through-structured-data-1400-1730-ensci/>

## Première vidéo (17min)

4min 20s -> 12min 30s (8min30s) : Historical background

- Derek de Solla Price
- Barabási–Albert
- PageRank

12min 30s -> 17min 0s (4min30s) : Cartographies

- Collaborations
- Co-word analysis
- Cluster' identification
- Filtering

17min 0s -> 23 min 0s (6min): Théorie de la percolation

- Graph et seuils : identification de moment de basculement pour dégager éléments structurels du réseau

A partir de 28min 0s (8min): Construction des clusters

- Modularité
- Clique

## Seconde vidéo (4min56s)

0s -> 3min 40s (3min 40s) : Construction des clusters

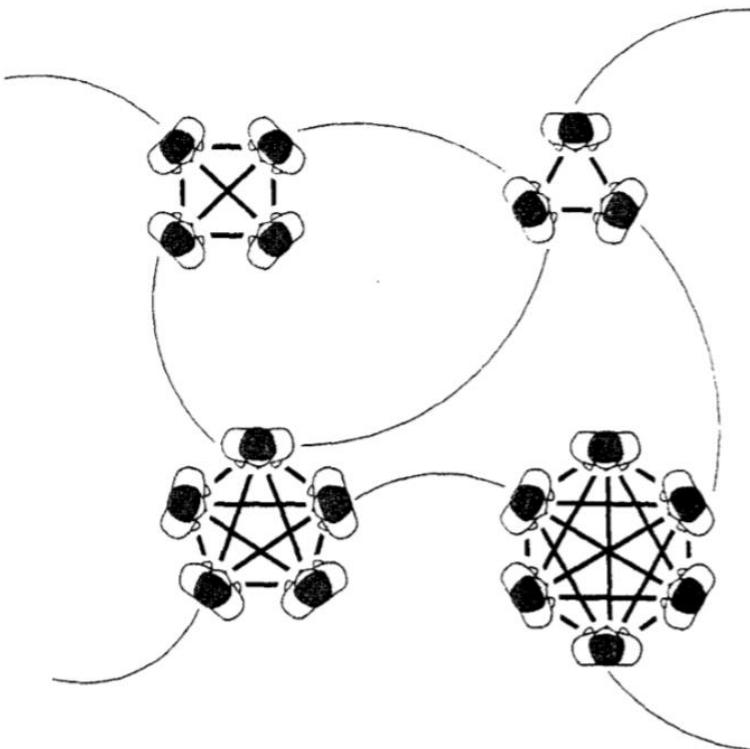
- Clique
- Maps of Random walks

3min 40s -> 4min 56s (1min16s) : Projection du réseau

- Force-based layout



# Complexité et réseaux sociaux



*Albert Lazlo Barabasi (2002) reprenant Granovetter*

*Les liens faibles jouent un rôle important dans la circulation d'informations, de ressources, rumeurs, entre différentes communautés.*

*Exemples : un réseau Facebook ou linkedin*

**Figure 4.1 Strong and Weak Ties.** In Mark Granovetter's social world, our close friends are often friends with each other as well. The network behind such a clustered society consists of small, fully connected circles of friends connected by strong ties, shown as bold lines. Weak ties, shown as thin lines, connect the members of these friendship circles to their acquaintances, who have strong ties to their own friends. Weak ties play an important role in any number of social activities, from spreading rumors to getting a job.

## *Identification des **hubs** et des **connecteurs***

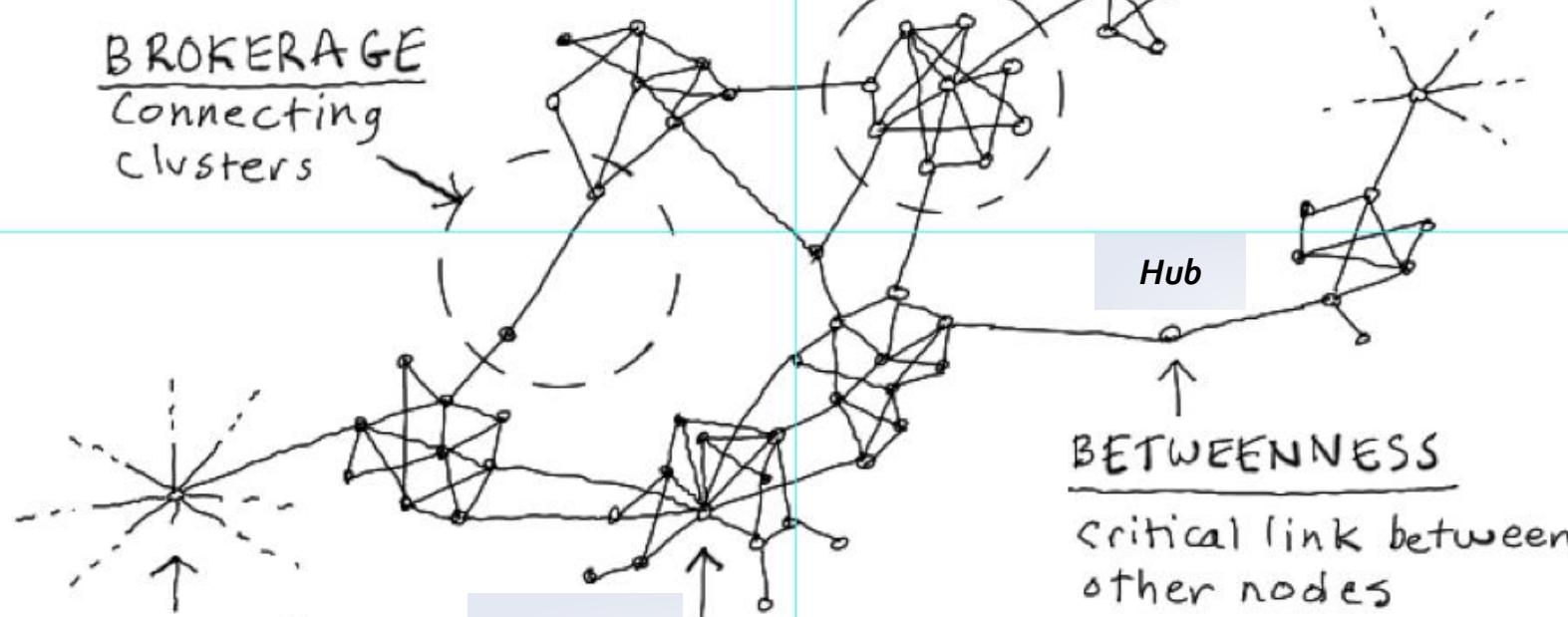
*Suivant la structure du réseau (collaborations d'auteurs, réseaux Facebook...), les individus appartenant au réseau ne se trouvent pas dans des situations identiques.*

- *Un **connector** est un nœud (individu) connecté de manière anormalement élevée à un ensemble d'autres nœuds*
- *Un **hub** est un nœud (un individu) dont les liens permettent de passer d'une première communauté à une seconde (et qui ne sont pas forcément densément liées)*

# ANATOMY OF A SOCIAL NETWORK

## BROKERAGE

Connecting clusters



## DEGREE

Number of connections

## CLOSURE

How easily a node can make connections

## CLOSURE

Building trust within a cluster

Hub

## BETWEENNESS

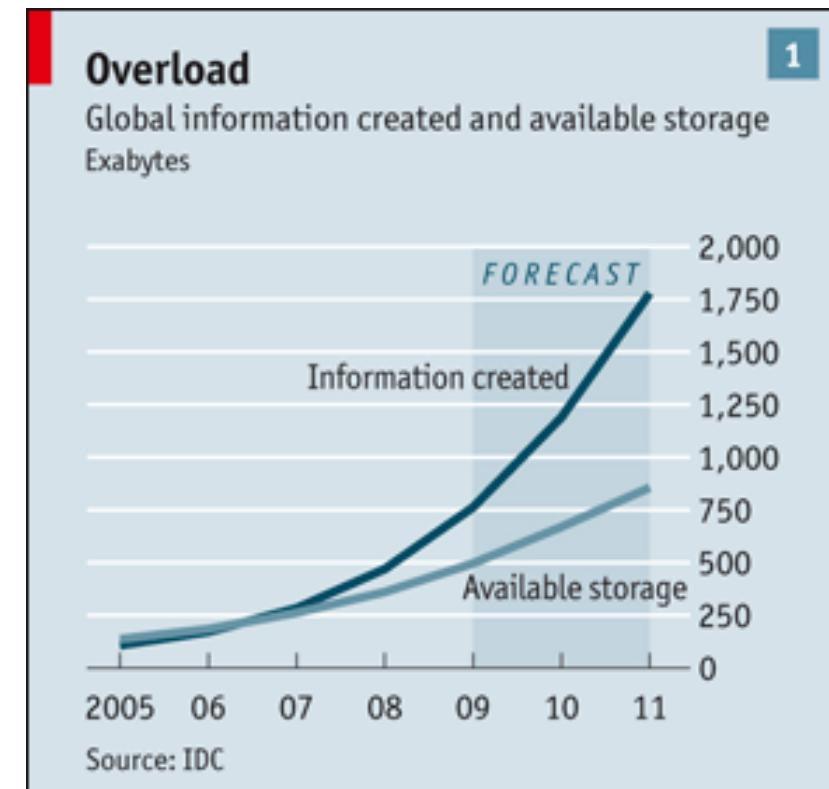
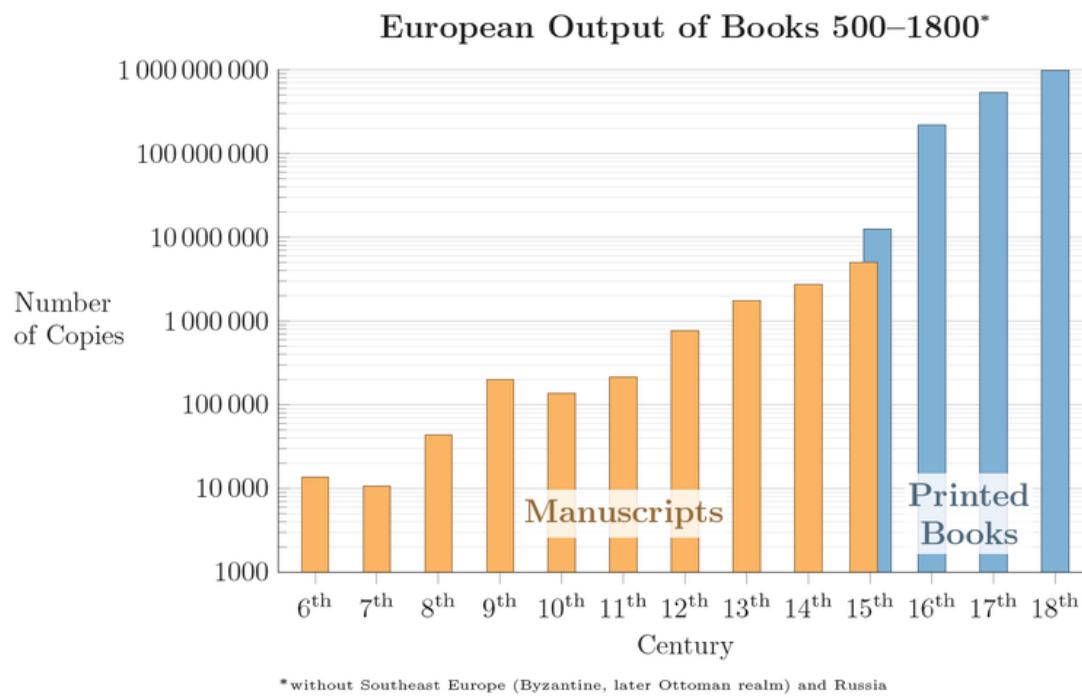
critical link between other nodes

# Objectifs et plan de la séance

- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- Introduction à la visualisation de l'information
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets

# Nouvelles sources de données et visualisation

Les activités humaines, la **vie en société, génèrent des données et des informations**. Il en a toujours été ainsi. Plus les sociétés se développent, plus le volume créé est important.



*History of books*, Wikipedia , visité le 06/02/2010,  
[http://en.wikipedia.org/wiki/History\\_of\\_books](http://en.wikipedia.org/wiki/History_of_books)

*Data, data everywhere*, The economist, 25/02/2010,  
[www.economist.com/node/15557443](http://www.economist.com/node/15557443)

Mais, l'information n'est pas connaissance. Et **plus le volume de données auquel nous avons accès augmente, plus il est difficile de les trier, de les organiser, et finalement de les comprendre.**

## Data deluge



Ces données sont parfois si complexes que **le meilleur moyen de les expliquer est d'utiliser l'image.**

« *Une représentation graphique n'est pas un simple dessin, elle implique souvent une grande responsabilité au moment de décider comment procéder. Il ne suffit pas de 'dessiner', une représentation graphique dans une forme solide. Il faut la construire et l'organiser jusqu'à ce que toutes les relations entre les données en soient révélées.* »

Jacques Bertin, *La graphique et le traitement graphique de l'information*, Flammarion, 1977

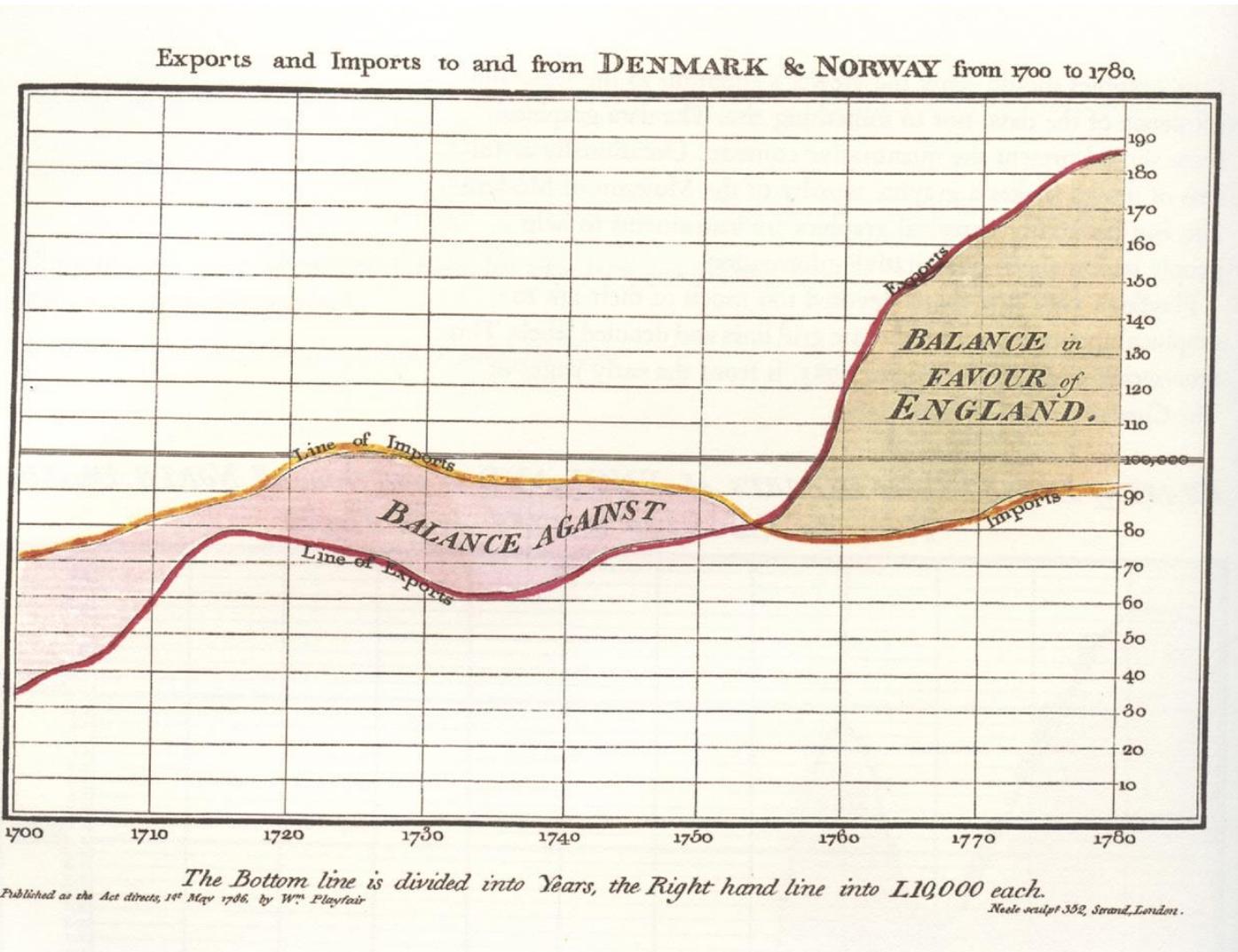
# Comment et pourquoi visualiser des informations ?



## Le monde visible

La mappemonde d'Ebstorf (Nord de l'Allemagne, 1300), montre l'ensemble du monde connu (éléments géographiques, historiques et religieux).

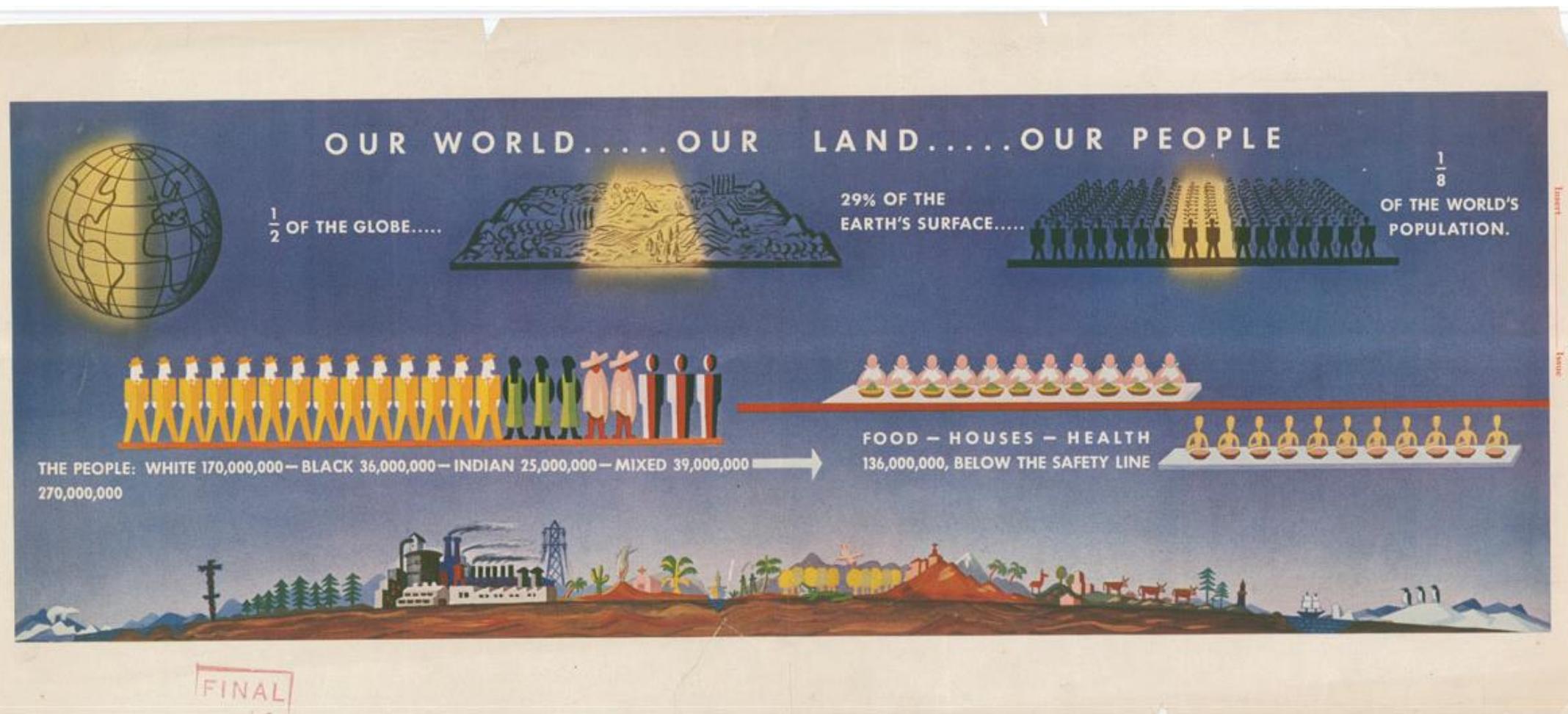
Décrire le monde connu, pour le savant, le scientifique.



## Les schémas statistiques

William Playfair, 1786, diagramme représentant les importations et exportations de l'Angleterre en fonction du temps pour montrer la balance commerciale.

Chiffres confinés à un public d'experts, le schéma et les diagrammes éclairent les décisions politiques.



## Informer le public

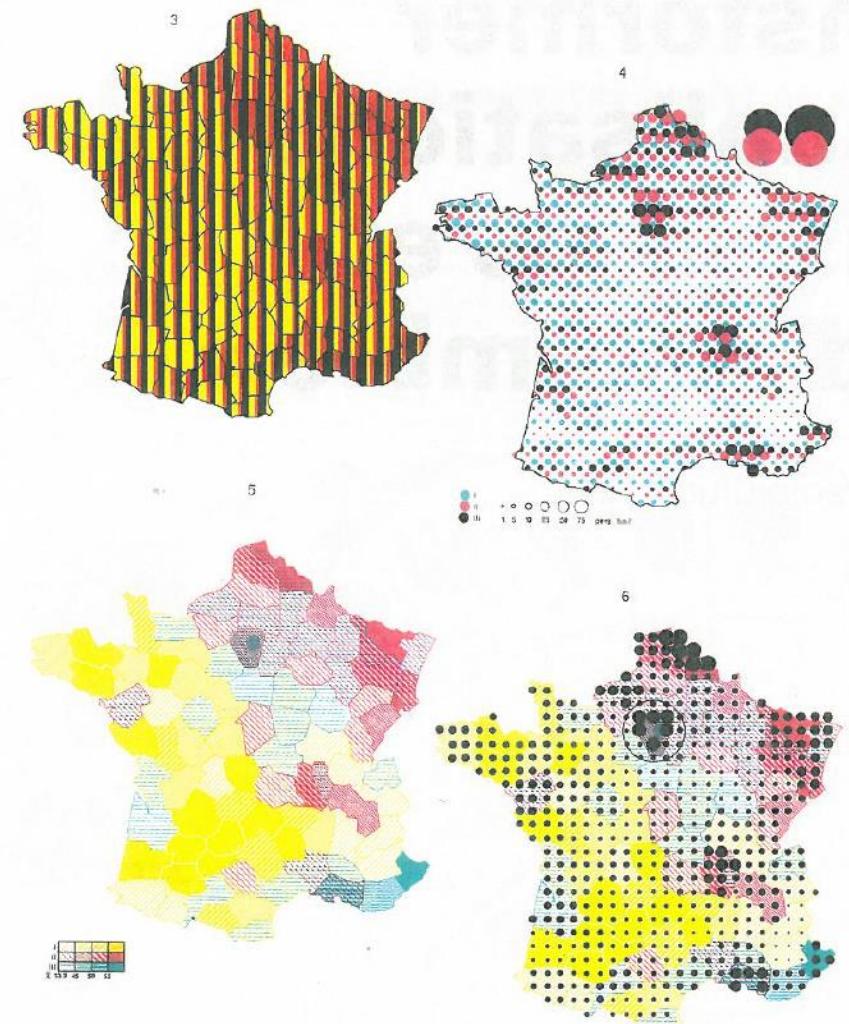
Analyse de la population des Amériques (Irving Geis, 1940, McCall) : composition ethnique et pauvres aux USA.

Avec le développement des journaux, de la presse et des magazines, les représentations graphiques d'information et de données se popularisent.

# Massification et théorisation

**Massification** de l'accès à l'information et donc aux données et visualisations.

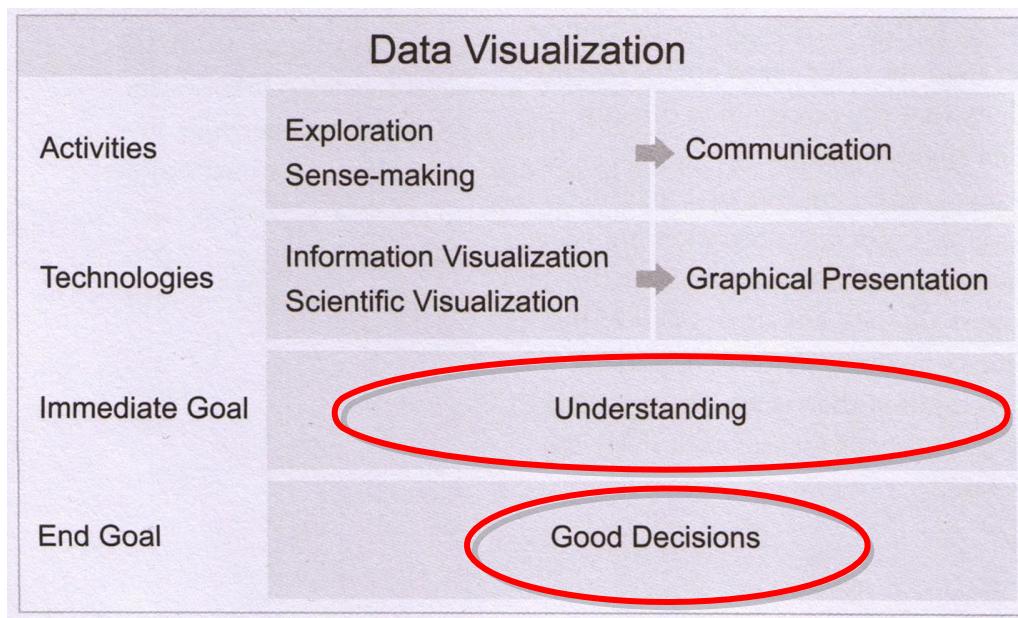
Mais qui s'accompagne d'une **utilisation parfois détériorée** (publicités, presse...).



1967

→ Formalisation mathématique de la projection de données sur des cartes (Jacques Bertin, 1967)

# Mais pourquoi cela est-il intéressant ?



Stephen Few, *Simple Visualization Techniques for Quantitative Analysis, Now you see it*, Analytics Press, 2009

Il est alors possible de comprendre visuellement, plus immédiatement, des situations - des problèmes souvent complexes.

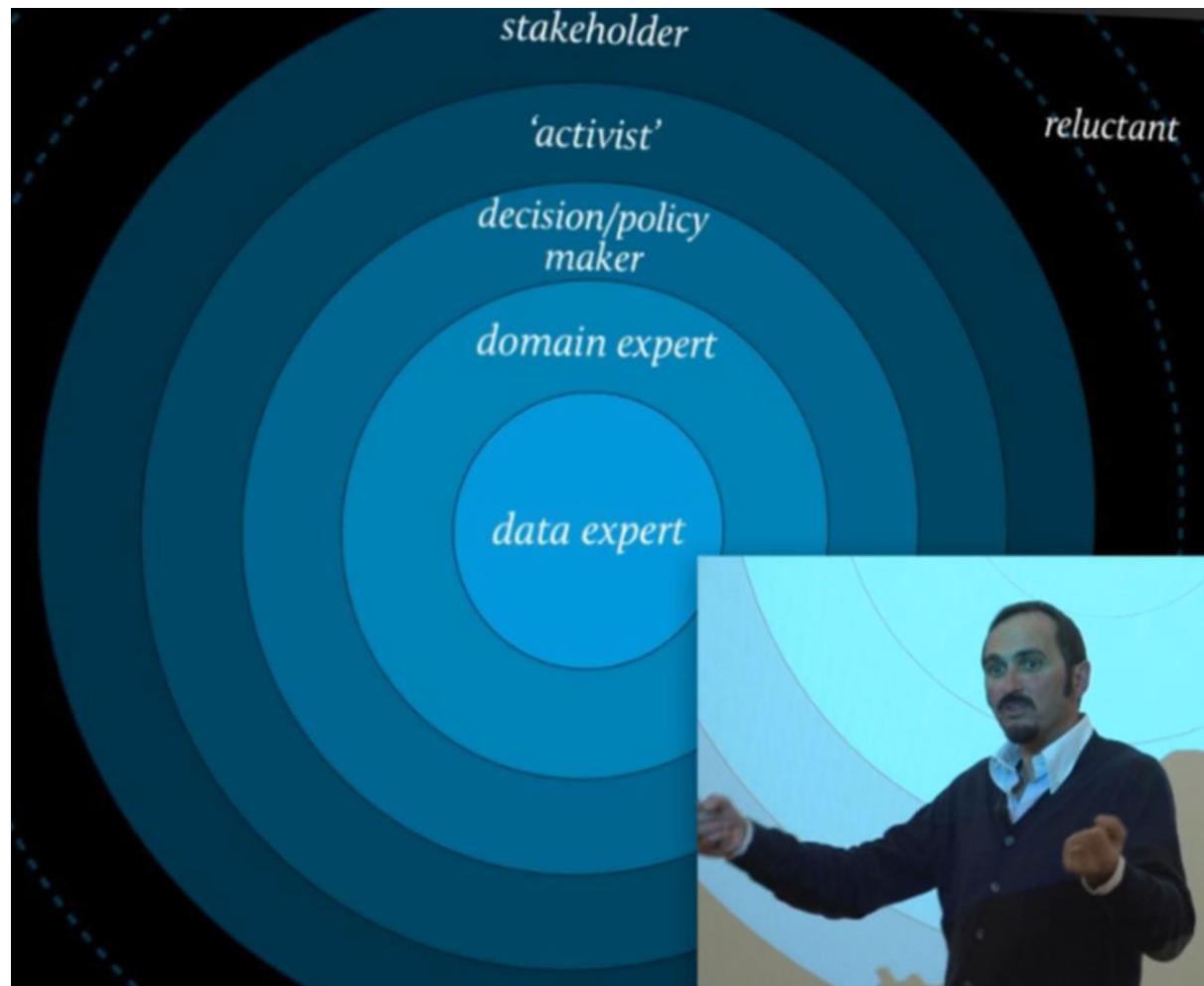
- **Présentation synthétique** des différentes dimensions d'un problème, d'un sujet
- La **position des thématiques, des acteurs, ...**
- Permet au lecteur (scientifique, citoyen, politique) d'obtenir rapidement des informations sur les **résultats synthétisent des situations complexes, et de comprendre les relations** qui associent ces éléments

Paolo Ciuccarelli : *Beyond Visualization, Designing meaning through data experience*, ENSCI, 2014



[http://www.dailymotion.com/video/x1ji4y4\\_paolo-ciuccarelli-20-mars-a-l-ensci\\_tech](http://www.dailymotion.com/video/x1ji4y4_paolo-ciuccarelli-20-mars-a-l-ensci_tech)

(17:40-> 28:15)

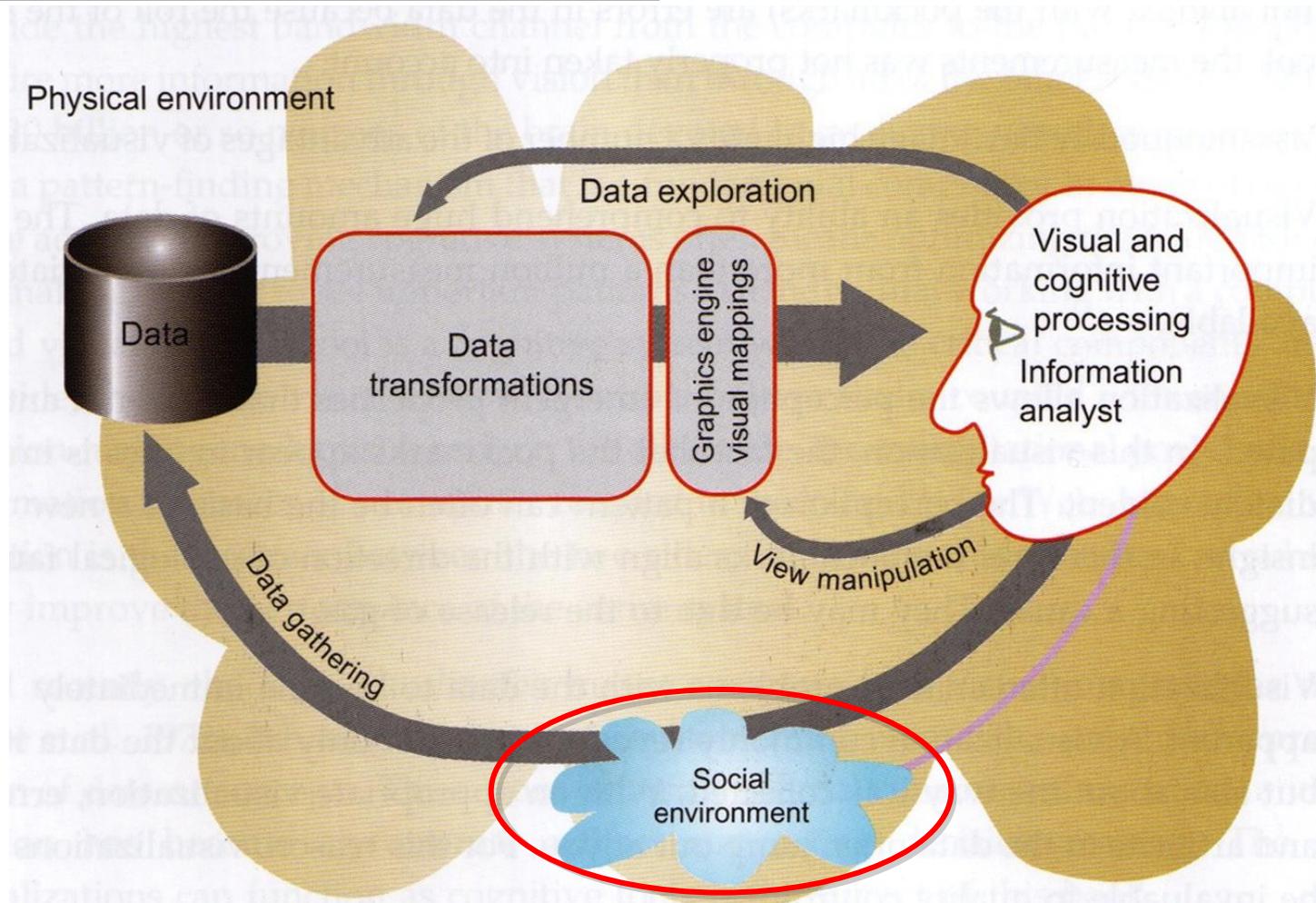


Paolo Ciuccarelli plaide pour une meilleure articulation : moins celui qui reçoit le message connaît le sujet, plus il est nécessaire de produire des visualisations intégrant des éléments de contexte (éléments narratifs).



# Objectifs et plan de la séance

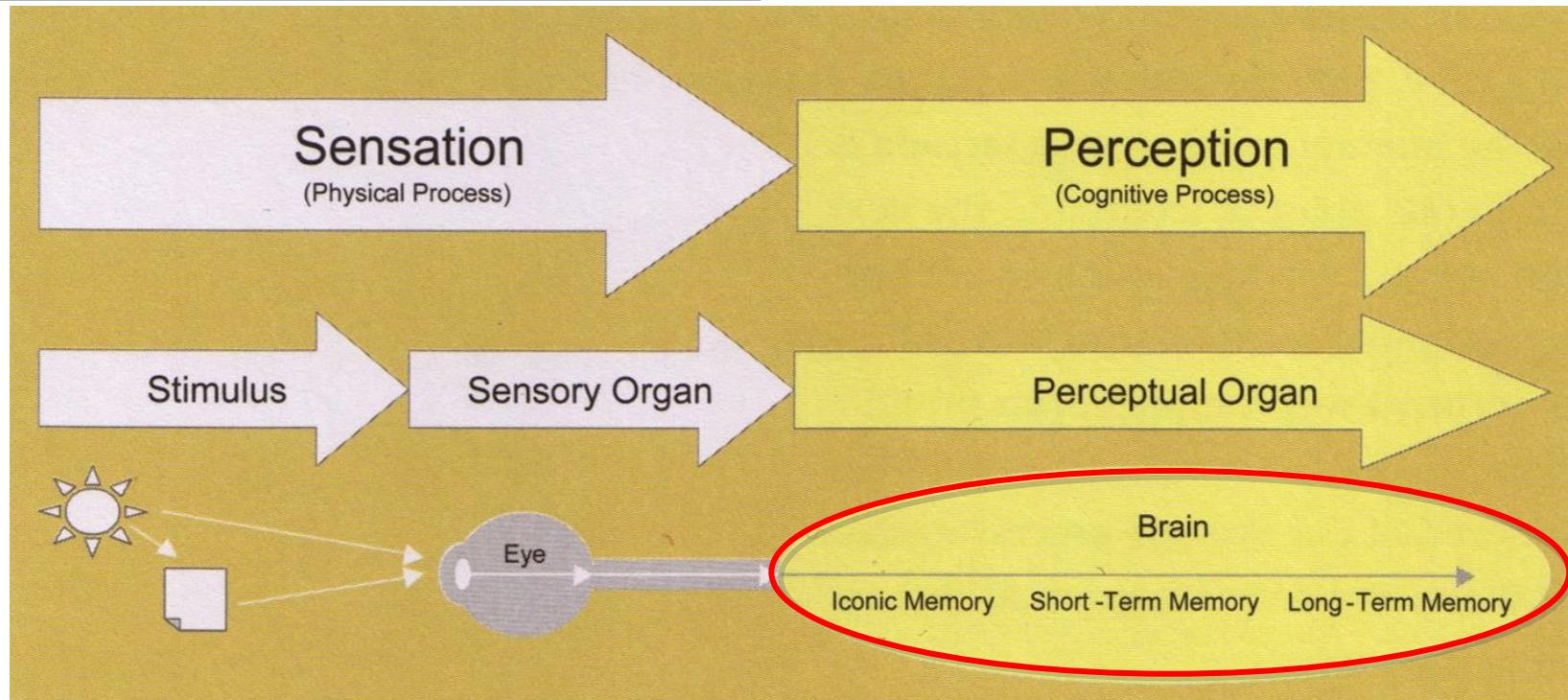
- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- **Introduction à la visualisation de l'information**
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets



Colin Ware, *Information Visualization, perception for design*, MK, 2012

L'environnement **social** (histoire, culturel...) : compréhension du sujet, choix du périmètre d'étude...

et l'environnement **physique** (sources accessibles, moyens techniques...) joue un rôle important à la fois dans la collecte de la données, et dans la bonne compréhension des principaux messages des visualisations.



Stephen Few, *Show me the Numbers, Designing Tables and Graphs to Enlighten*, Analytics Press, 2004

La perception visuelle dépend de nombreux facteurs :

- **Physiques** : lumière (chaleur de la lumière, intensité, reflet...), structure organique de l'œil
- **Cognitifs** (mémoire, expériences...)

Le **processus cognitif** mobilise trois types de mémoire :

- **Mémoire visuelle**: automatique et inconsciente (très rapide), utilisée pour reconnaître les objets (symboles...), saisir les mouvements...
- Mémoire **court-terme** : temporaire et capacité limitée en quantité d'information, c'est ce qu'il est possible de retenir comme information lorsqu'on parcours une visualisation
- Mémoire **long-terme** : permet de saisir un contexte entre deux situations ou faits, reconnaître des images, les motifs principaux (moins importante pour nous)...

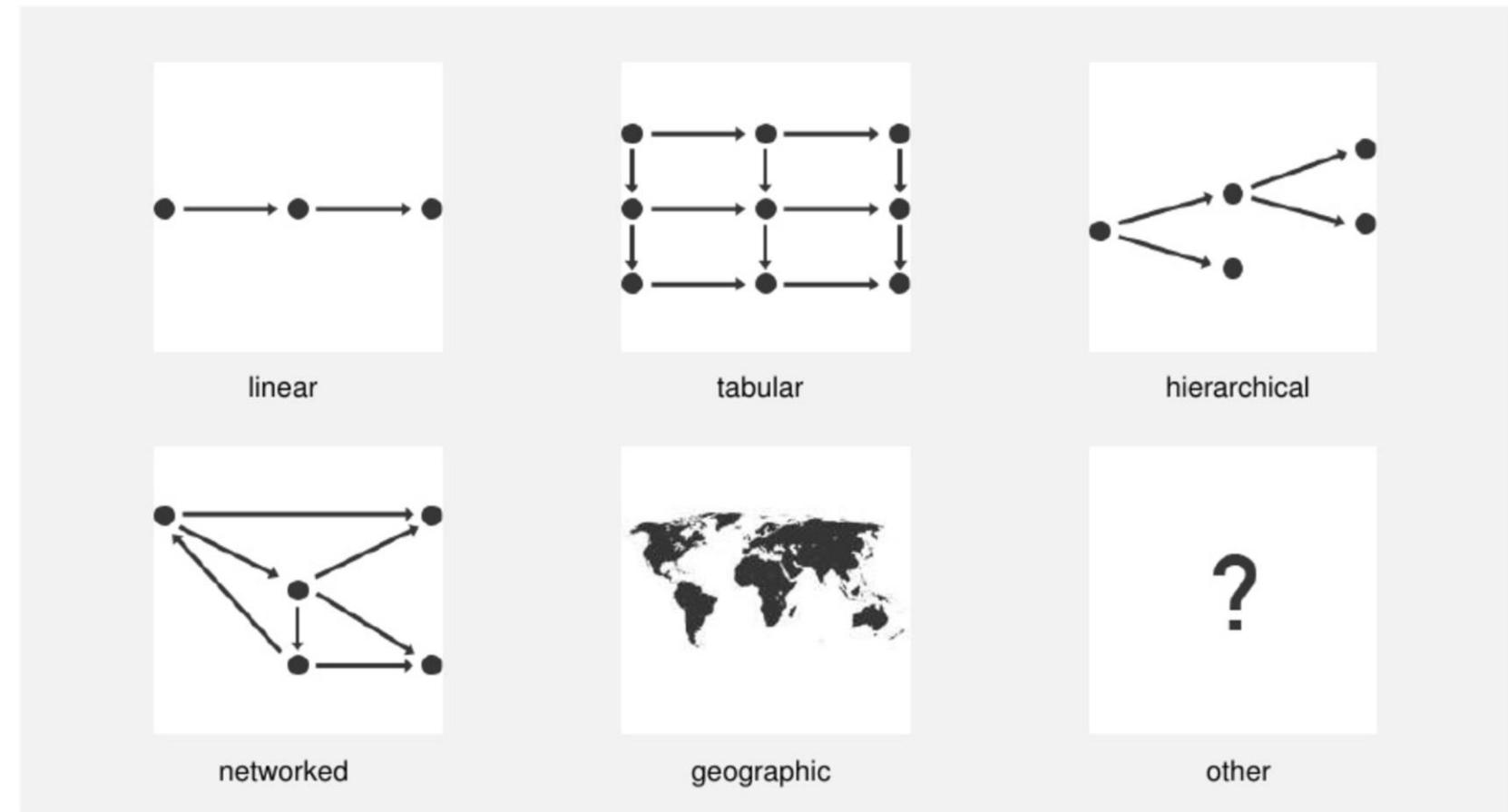
## Etape 1 / Comment cartographier des données ? Des données aux variables visuelles !

property	marks	ordinal/nominal mapping	quantitative mapping
shape	glyph	O □ + △ S U	
size	rectangle, circle, glyph, text	● ● ● ●	● ● ● ● ● ● ● ● ● ●
orientation	rectangle, line, text	— — /   \ —	— — — / / / / / / / /
color	rectangle, circle, line, glyph, y-bar, x-bar, text, gantt bar	■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ...	min max

Chris Stolte, Diane Tang and Pat Hanrahan, *Polaris : A system for query, analysis, and visualization of multidimensional relational databases*, IEE, 2002

- Alors que les symboles sont adaptés aux **variables qualitatives** (textes, types, classification...),
- faire varier un cercle, par exemple, est adapté aux **variables quantitatives** (entier, décimale...).

## Etape 2 / Examine the Data



Trisnadi Kurniawan, *Infographics and Data Visualisation*, <http://fr.slideshare.net/trisnadi/infographics-data-visualisation>, 2009

Combien d'éléments peuvent être reliés pour construire un réseau ?  
Quels sont les liens entre ces éléments dans les données ?

## / Etape 3

# Six Graphic Variables

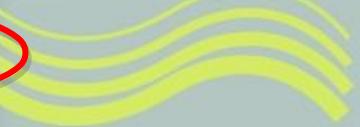
By: Dave Melsness

Differentiated

Ordered

### Graphic Variables

The six graphic variables displayed below are the cartographer's key to success. How they are used determines the effectiveness of a graphic, or in this case a map, and how easily it is interpreted by the audience. Refer to the chart below and consider the data you wish to display with each before determining which graphic variable and symbolization (point, line, polygon/area) is most suitable.

	Point	Line	Polygon/Area
Shape			Not Effective
Pattern/Texture	Not Effective		
Orientation			Not Effective
Size			Not Effective
Color			
Value			

Similar objects

An action / mouvement  
(Links in our visualization)

Quantify a phenomenon  
(proportionality)

A change of a position  
(cons and pros) or to group and categorize

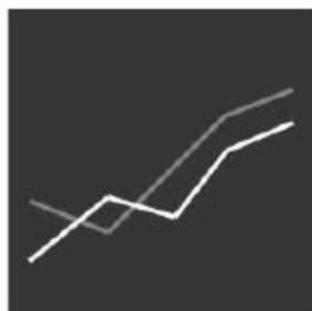
# Etape 4 / Data Visualisation Patterns

INDEPENDENT



Bar charts

CONTINUOUS



Line graphs



Stacked area

PROPORTIONS

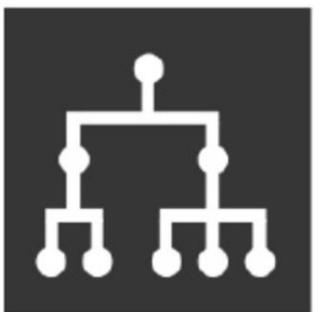


Pie charts



Ring charts

HIERARCHIES



Tree diagrams

NETWORKS



Diagram map

CORRELATIONS



Scatterplots

CARTOGRAPHICS



Bubble charts

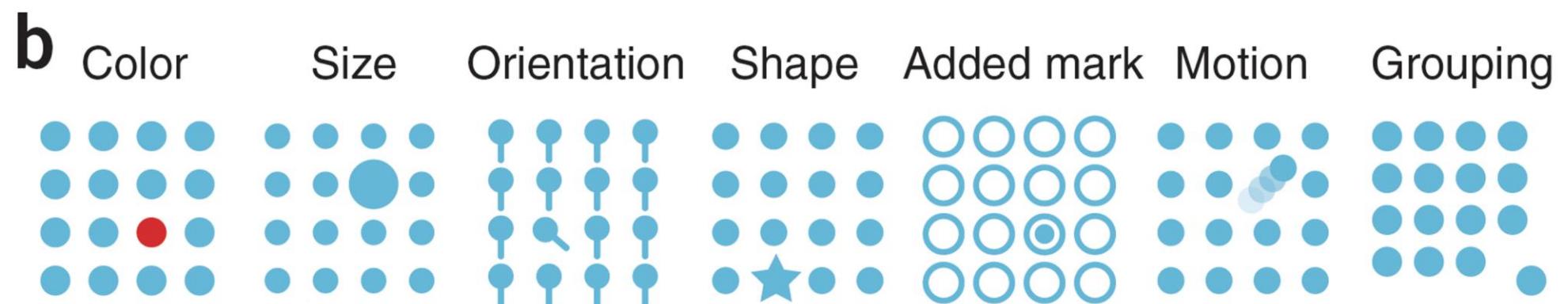
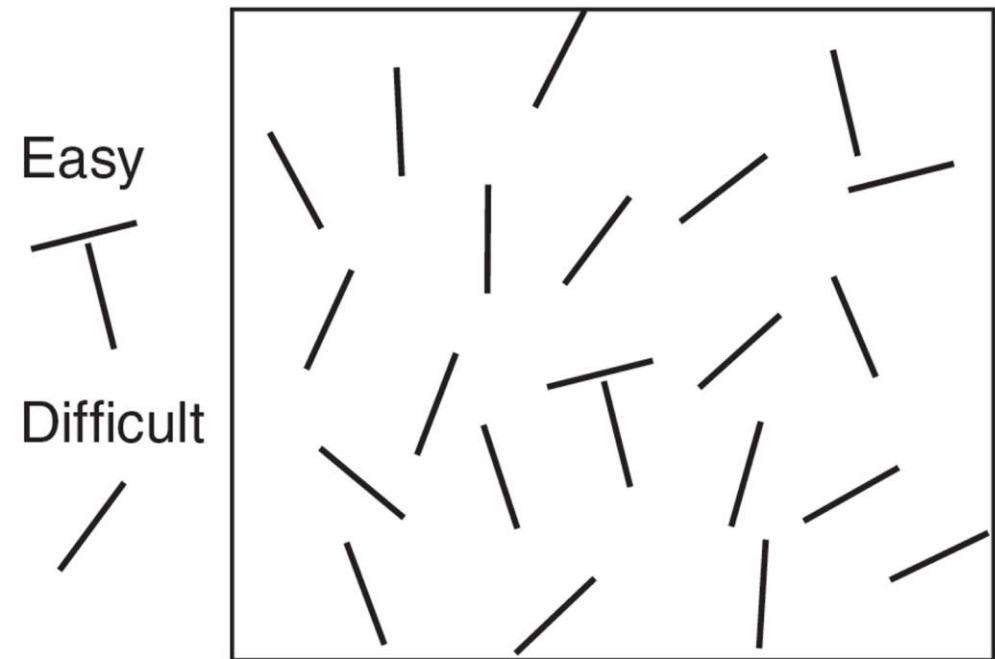


Maps

Trisnadi Kurniawan, *Infographics and Data Visualisation*, <http://fr.slideshare.net/trisnadi/infographics-data-visualisation>, 2009

## Mémoire visuelle : effet d'**identification immédiate** (ou pas !)

<b>a</b>	MSVTLHTVFCERTPKTC
Easy	EMESRCVPQEGVQWRDL
A	<b>GSA</b> LQPGFGGFKQVFCL
Difficult	SLPRTGRGGNSIWWGKK
P	FEDEYSEYSEYLKH <b>A</b> VR
	GVVSMSNNGPNTNGSQF
	FITYGKQPHLDMDKYTVF
	GKVIDGLEK <b>A</b> PVNEKTY
	RPLNDVHIKDITIHNPF



Bang Wong, *Points of View: Salience*, 2010, Nature Methods, p 7-773

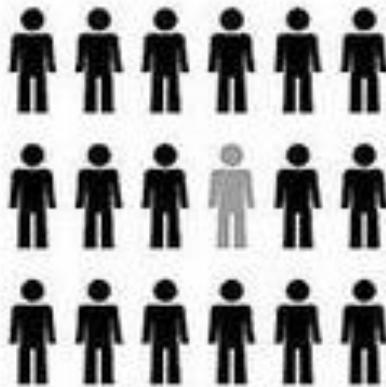
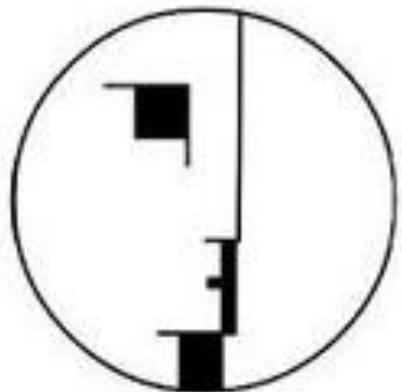
- (a) Certains éléments peuvent être vus d'**un seul regard**, d'autres sont difficilement identifiables
- (b) Exemples où **certaines objets sont isolés** en utilisant les variables graphiques et cet effet

## Théorie de la Gestalt (1935)

Le postulat de base est le suivant : devant la complexité de notre environnement, le cerveau va chercher à mettre en forme, à donner une structure signifiante à ce qu'il perçoit, afin de le **simplifier et de l'organiser**. Pour cela, il structure les informations de telle façon que ce qui possède une signification pour nous, se détache du fond pour adhérer à une structure globale.

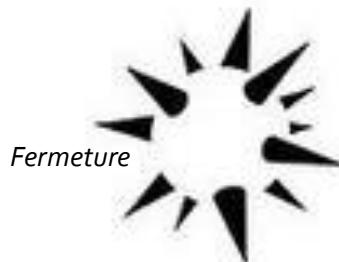
Les lois :

- **Proximité** : nous regroupons les points d'abord les plus proches les uns des autres.
- **Fermeture** : lorsque des images sont imparfaites ou qu'une suite d'événements est incomplète, notre cerveau a tendance à combler les vides afin de percevoir ces informations dans leur totalité.
- **Similarité** : si la distance ne permet pas de regrouper les points, nous nous attacherons ensuite à repérer les plus similaires entre eux pour percevoir une forme.
- **Symétrie** : des éléments symétriques sont perçus comme une forme globale.
- **Continuité** : des points rapprochés tendent à représenter des formes lorsqu'ils sont perçus, nous les percevons d'abord dans une continuité, comme des prolongements les uns par rapport aux autres.
- **Familiarité** : les formes les plus familières sont les formes les plus rapidement identifiables.

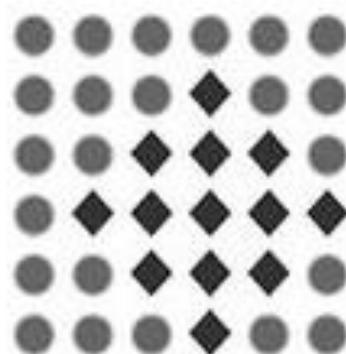


Fermeture

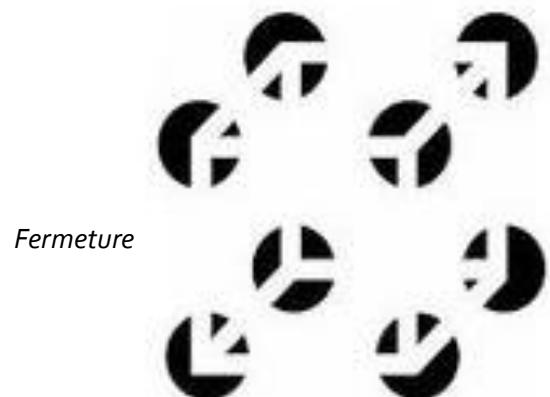
Symétrie



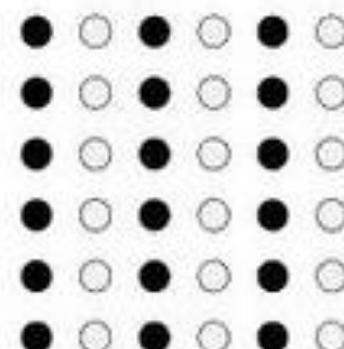
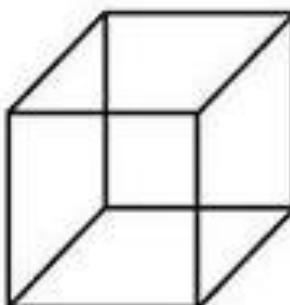
Fermeture



Proximité



Fermeture



Similarité



# Objectifs et plan de la séance

- Introduction du cours
  - Objectifs
  - Plan des séances et évaluation
- Aux origines de l'analyse des traces numériques : la scientométrie
  - Articles scientifiques
  - Les indicateurs d'activité
  - Réseaux de collaborations (1.1)
  - Réseaux de citations et de co-citations (1.2)
  - La co-occurrence des mots
- Au delà de la scientométrie
  - Scientometrie et au-delà
  - Complexité et réseaux sociaux
- Nouvelles sources de données et visualisation
  - Le déluge de données
  - Image et complexité
  - Des statistiques visuelles aux infographies
- Introduction à la visualisation de l'information
  - Les étapes du processus de visualisation
  - Données et variables visuelles
  - Les opérations
  - Mémoire visuelle et effets visuels
- Echelles et dimensions d'analyse
  - Echelles et type d'analyses
  - Typologie des informations représentées
  - Illustrations par des projets

# Echelles et types d'analyses

La **précision du sujet** consiste aussi à bien définir l'envergure du sujet !

## Les nanotechnologies

Nanobiologie

Nanomatériaux

Nanoélectronique

Nanorobots

Ingénierie  
tissulaire

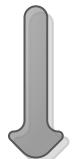
...

...

...

...

+ général



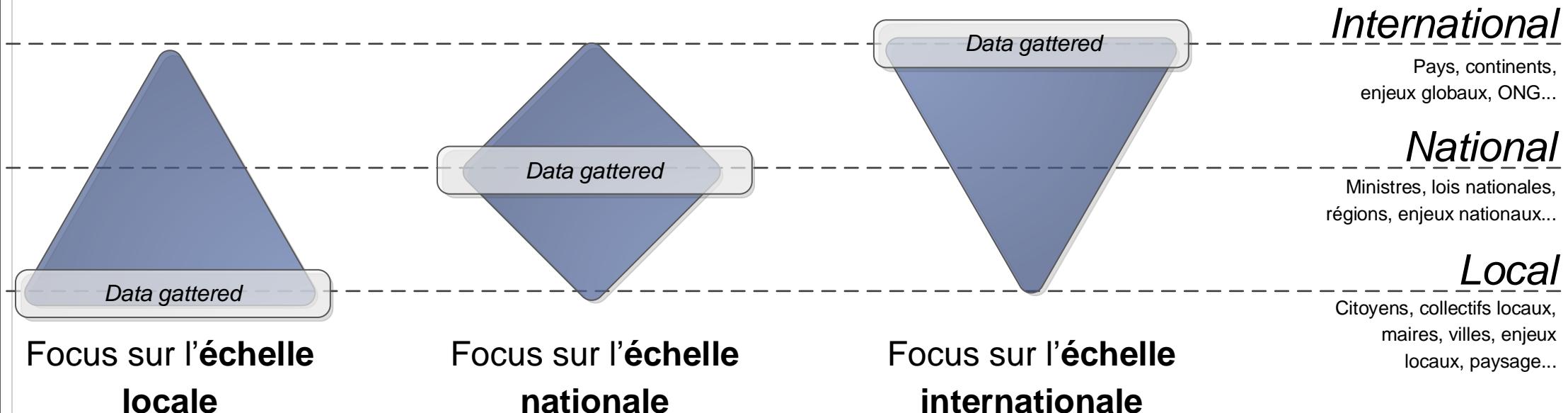
+ précis

Le volume d'information disponible est en général fonction de la précision du sujet !

- *Général : plus d'information, mais nécessite de mieux trier, de mieux synthétiser*
- *Précis : moins d'information, nécessite de fouiller d'avantage*



## Exemples d'échelles géographiques



Focus sur l'échelle  
locale

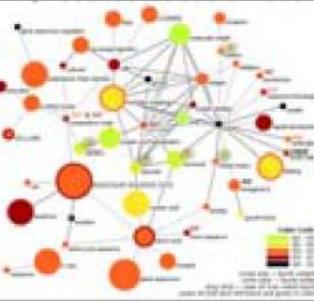
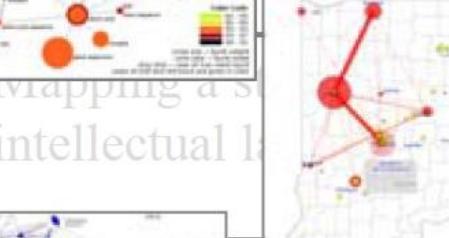
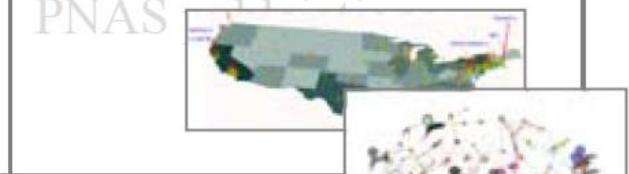
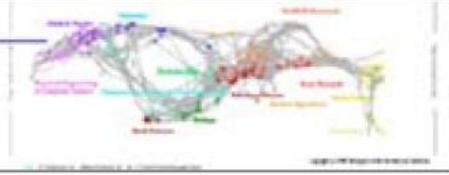
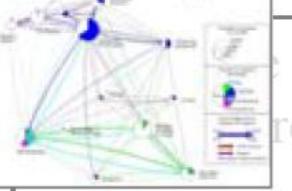
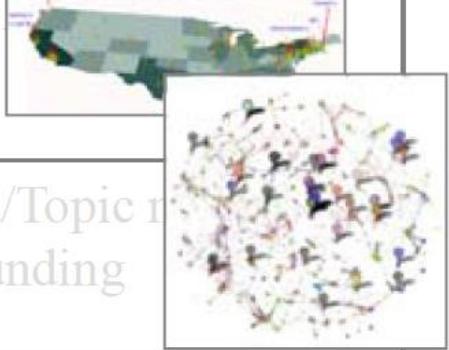
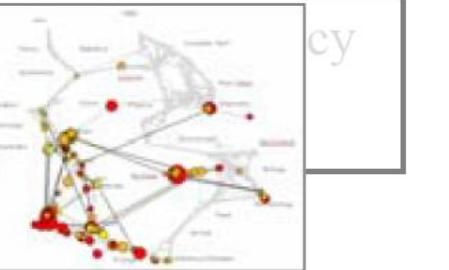
Focus sur l'échelle  
nationale

Focus sur l'échelle  
internationale

### Exemple gaz de schiste

- International : film Gasland, environnement et réchauffement climatique...
- National : indépendance énergétique...
- Collectifs locaux, citoyens, maires...



	<i>Micro/Individual (1-100 records)</i>	<i>Meso/Local (101–10,000 records)</i>	<i>Macro/Global (10,000 &lt; records)</i>
<b>Statistical Analysis/Profiling</b>	Individual person and their expertise profiles	Larger labs, centers, universities, research domains or states	All of NSI, all of science, SA, all of science
<b>Temporal Analysis (When)</b>	Funding portfolio of one individual		
<b>Geospatial Analysis (Where)</b>	Career trajectory of one individual		
<b>Topical Analysis (What)</b>			
<b>Network Analysis (With Whom?)</b>	NSI's work of one		

# Types d'information des visualisations

*Information Graphics*, Sandra Rendgen, 2012, Taschen, 480 p.

**Lieu** : les éléments sont organisés de façon spatiale (ex : carte géographique)

- Les évènements se déroulent à un certain endroit
- Les objets sont organisés de façon spatiale, il est donc possible de placer des évènements simultanés dans un espace.

**Temps** : les éléments sont organisés de façon chronologique (ex : graphique d'une variable par année).

- Il y a une séquence temporelle fixe, chronologique
- Permet de décrire un phénomène, une tendance (mais n'explique pas les causes)

**Catégorie** : les éléments sont divisés en classes (ex : une classification présentant des groupes de données, de phénomènes)

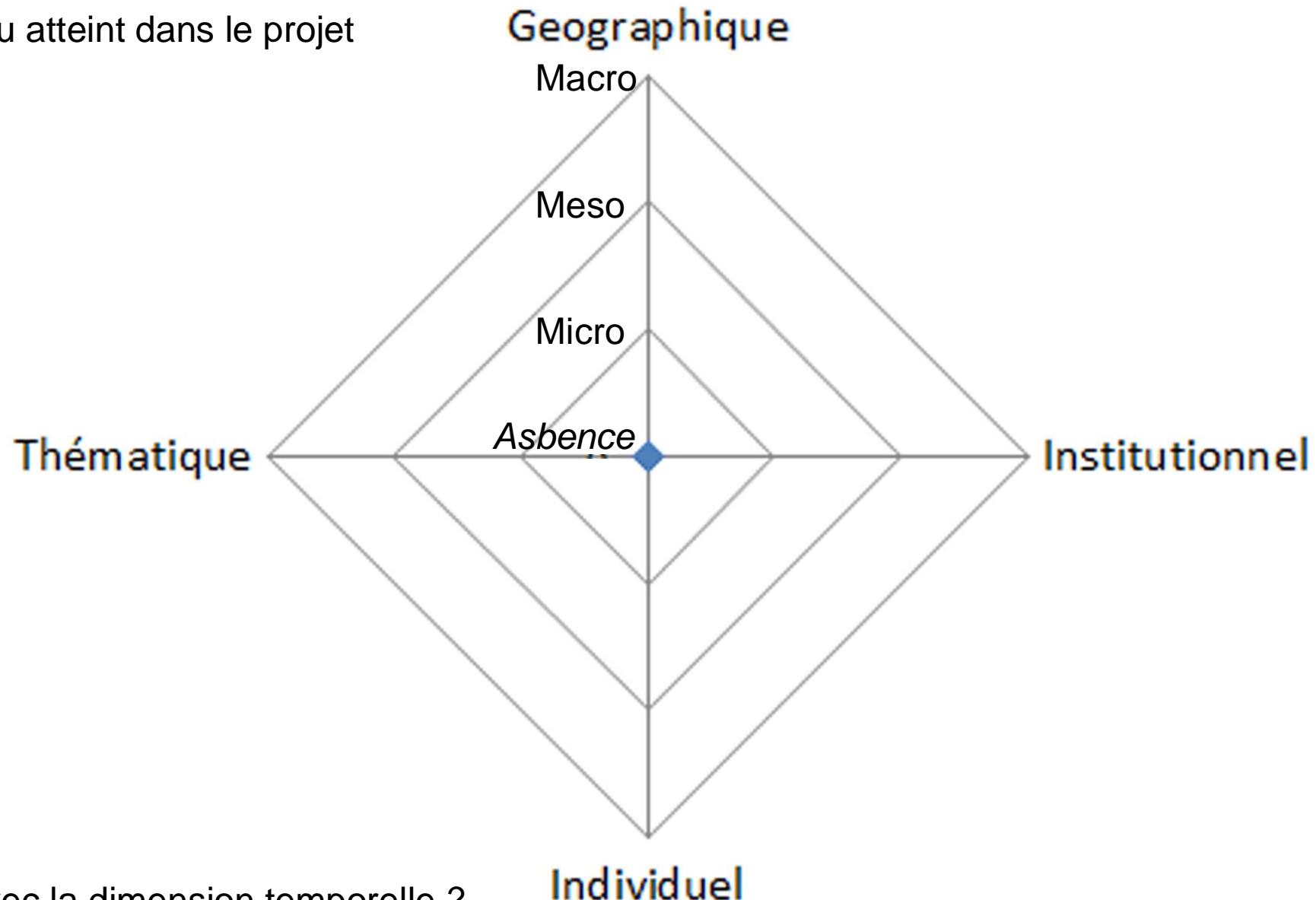
- Par type, par caractéristiques communes
- Il y a un ordre

**Hiérarchique** : les éléments sont organisés verticalement (ex : treemap, camembert...)

- Éléments triés par ordre de priorité, par ordre d'importance
- Présente un classement des éléments, par quantité

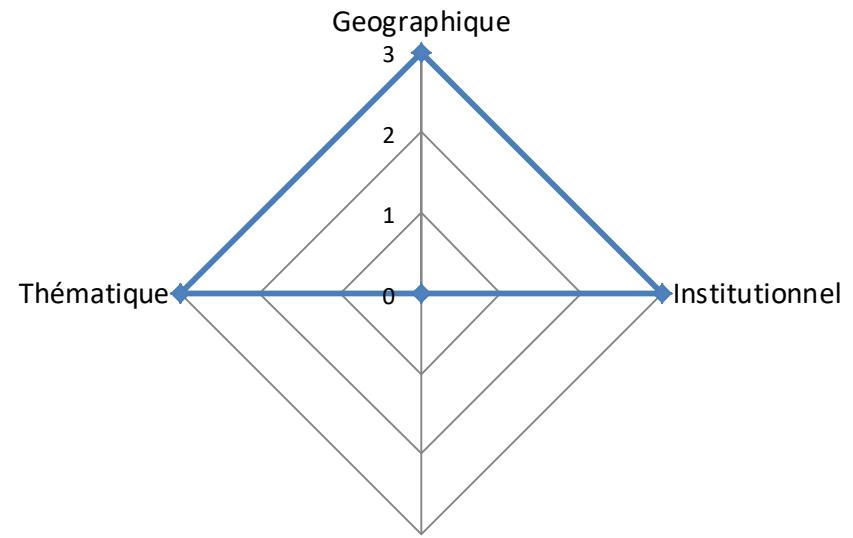
## Grilles de lecture des projets :

- Dimensions d'analyse
- Echelles : niveau atteint dans le projet



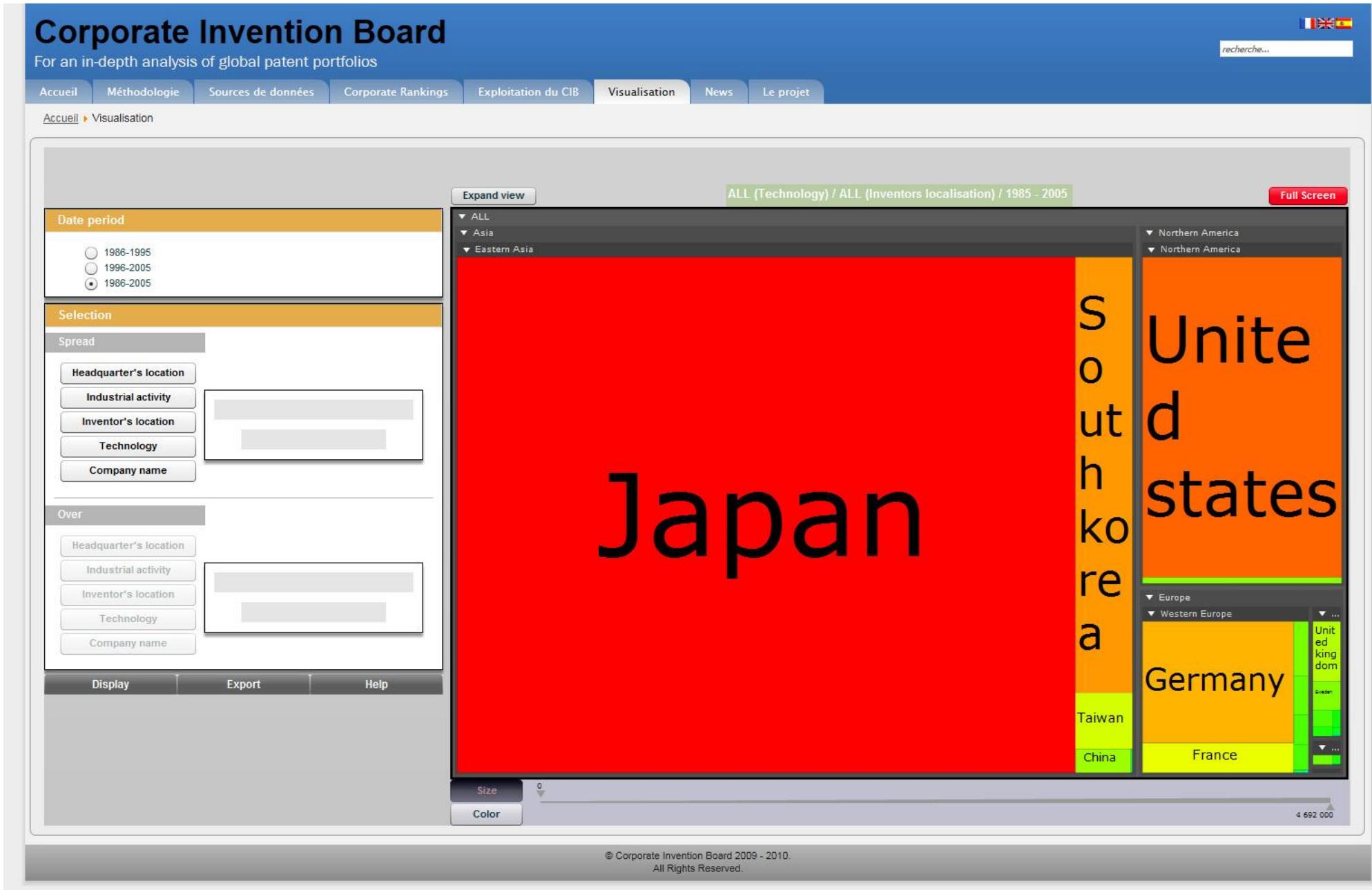
# Globalisation de la R&D des 2 400 plus grands groupes mondiaux en investissements de R&D

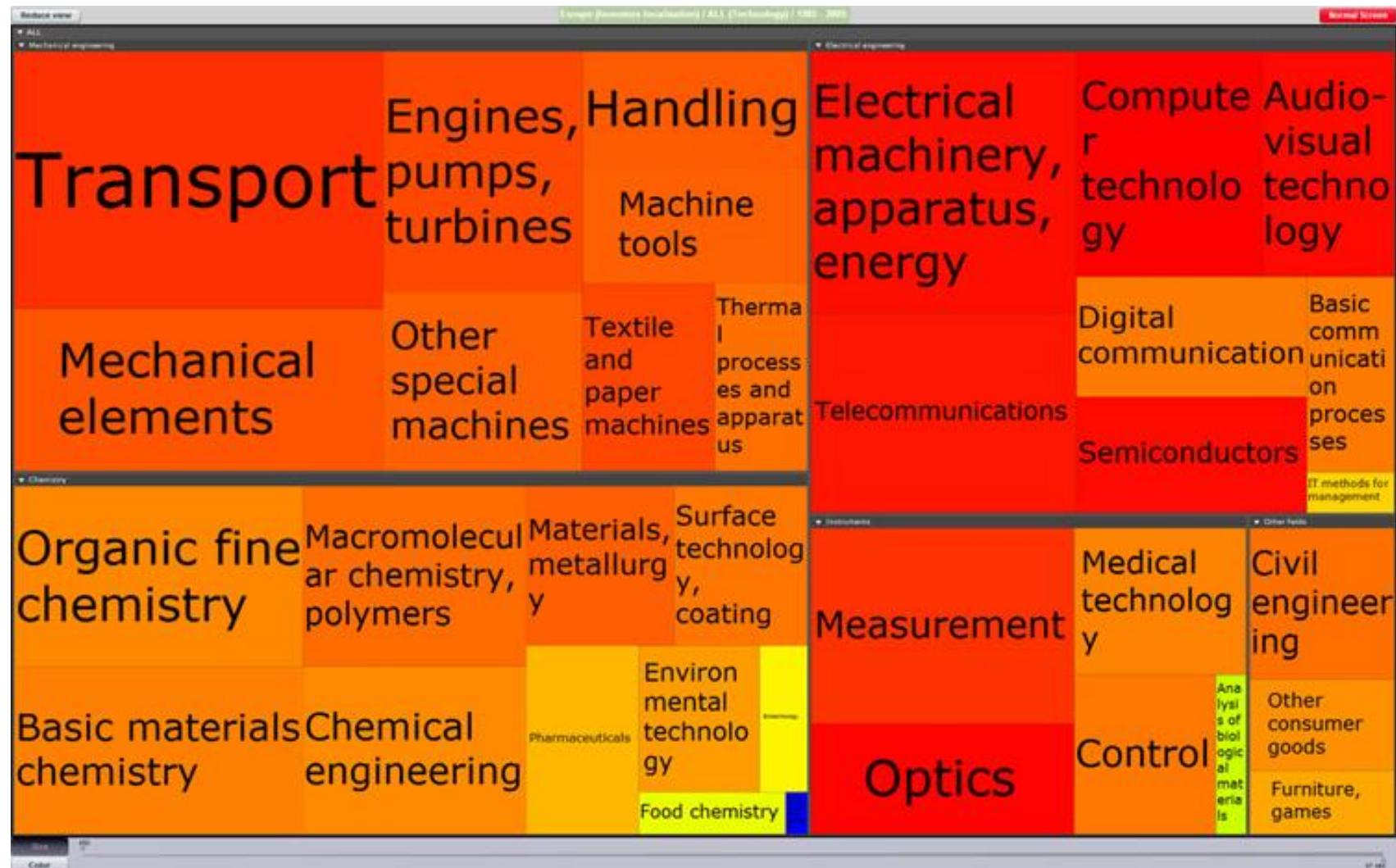
**Corporate Invention Board**



## Données collectées

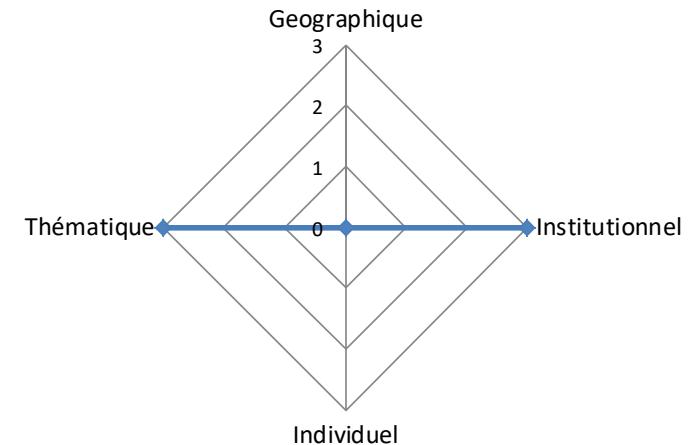
- 2 000 groupes du scoreboard (IPTS)
- 400 filiales chinoises et indiennes
- 170 000 filiales (ORBIS)
- 6 millions de brevets prioritaires (Patstat 2009)





# Global Map of Technology

Map of Technology

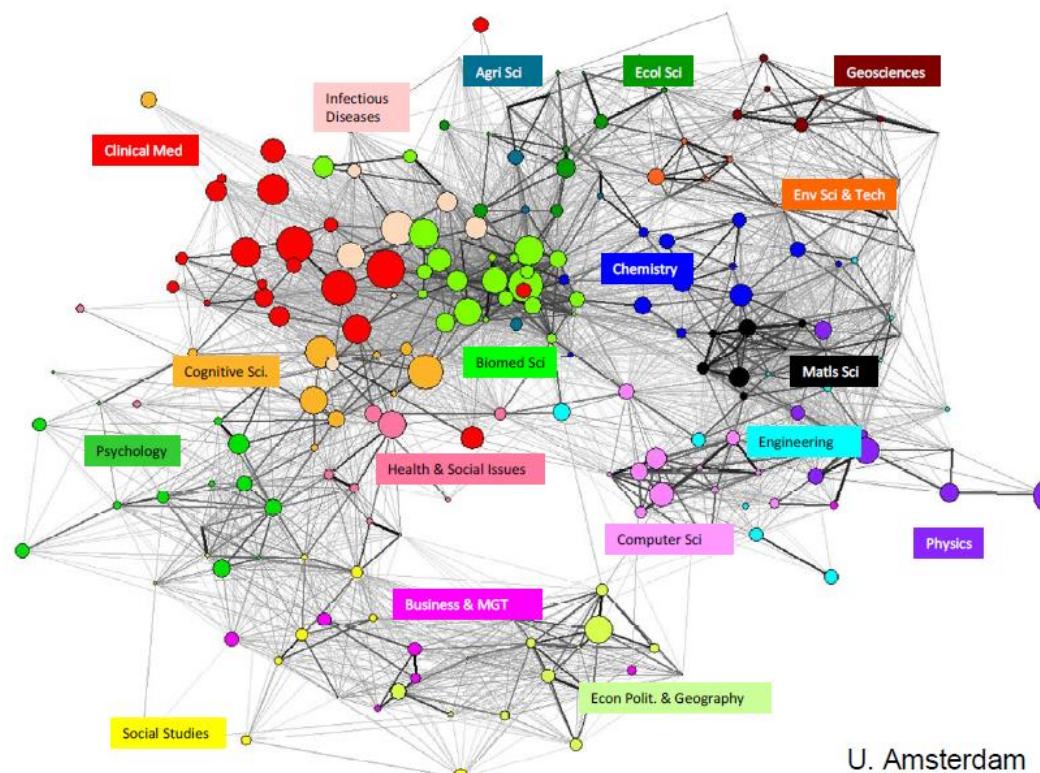


## Périmètre initial

- Ensembles des **dépôts prioritaires** (6 millions) disponibles

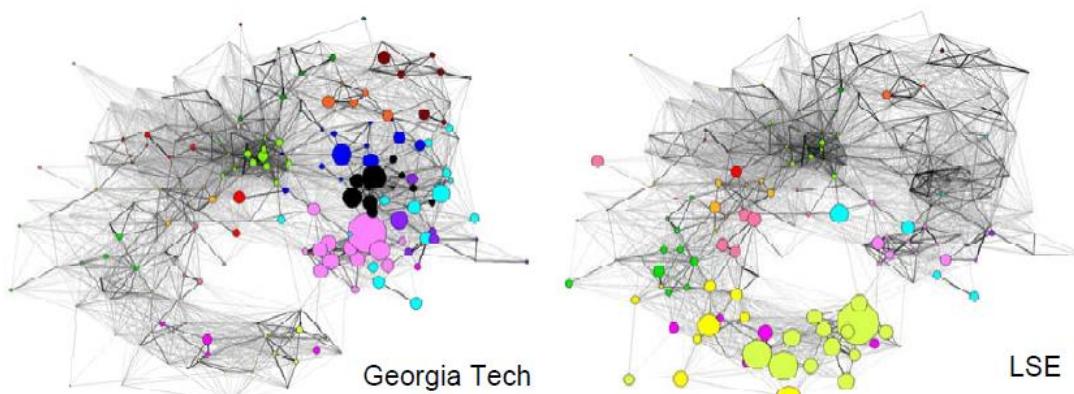
## Principaux éléments de méthode

- Classification de technologie produite la WIPO (Ulrich Schmoch, 2008), enrichie : 5 domaines, 35 classes et **389 sous-champs technologiques** basées sur le IPC
- **Co-occurrences d'IPC** et clustering



Motivations ?

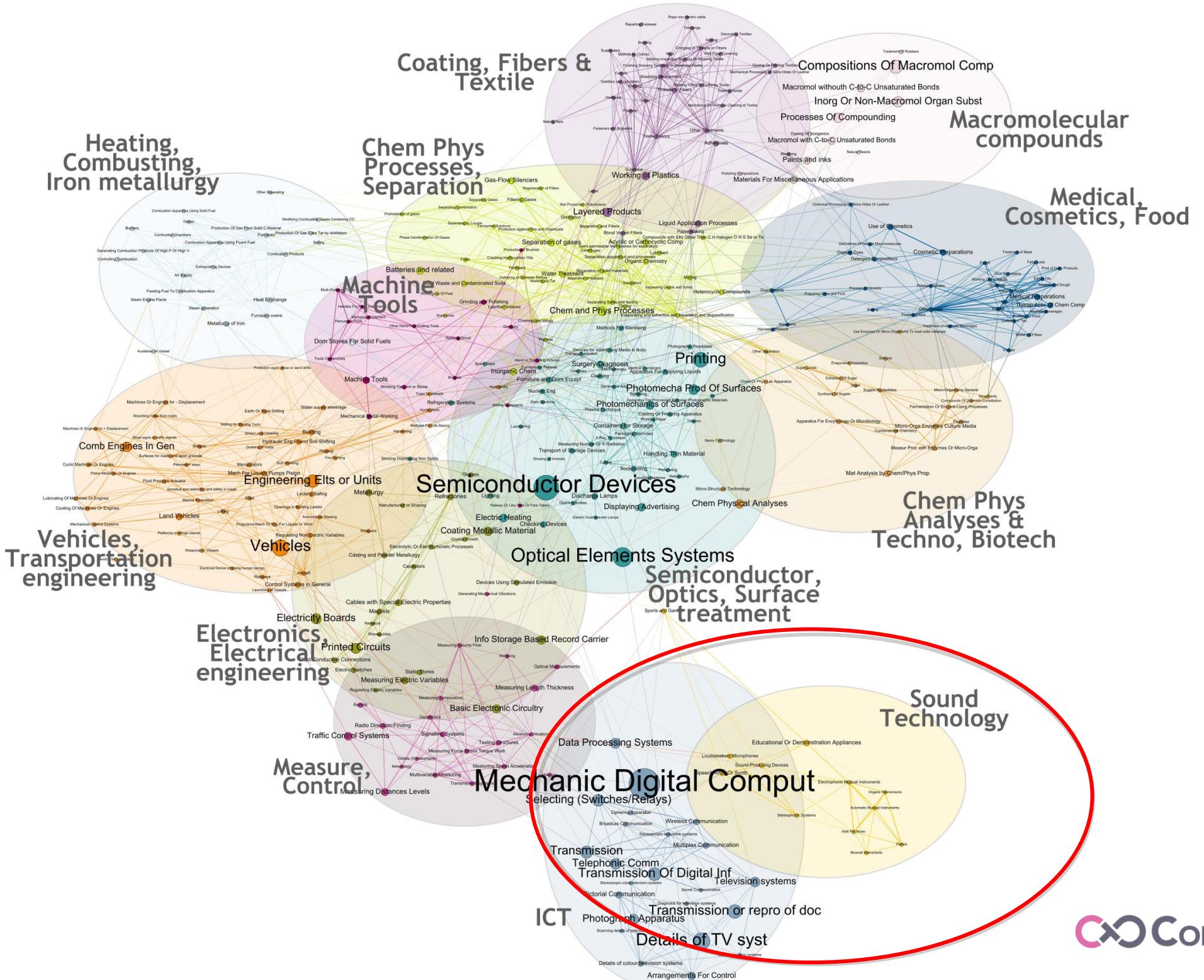
**Carte globale de la science:**  
 Publication scientifiques  
 Citations journals-journals  
 Catégories ISI



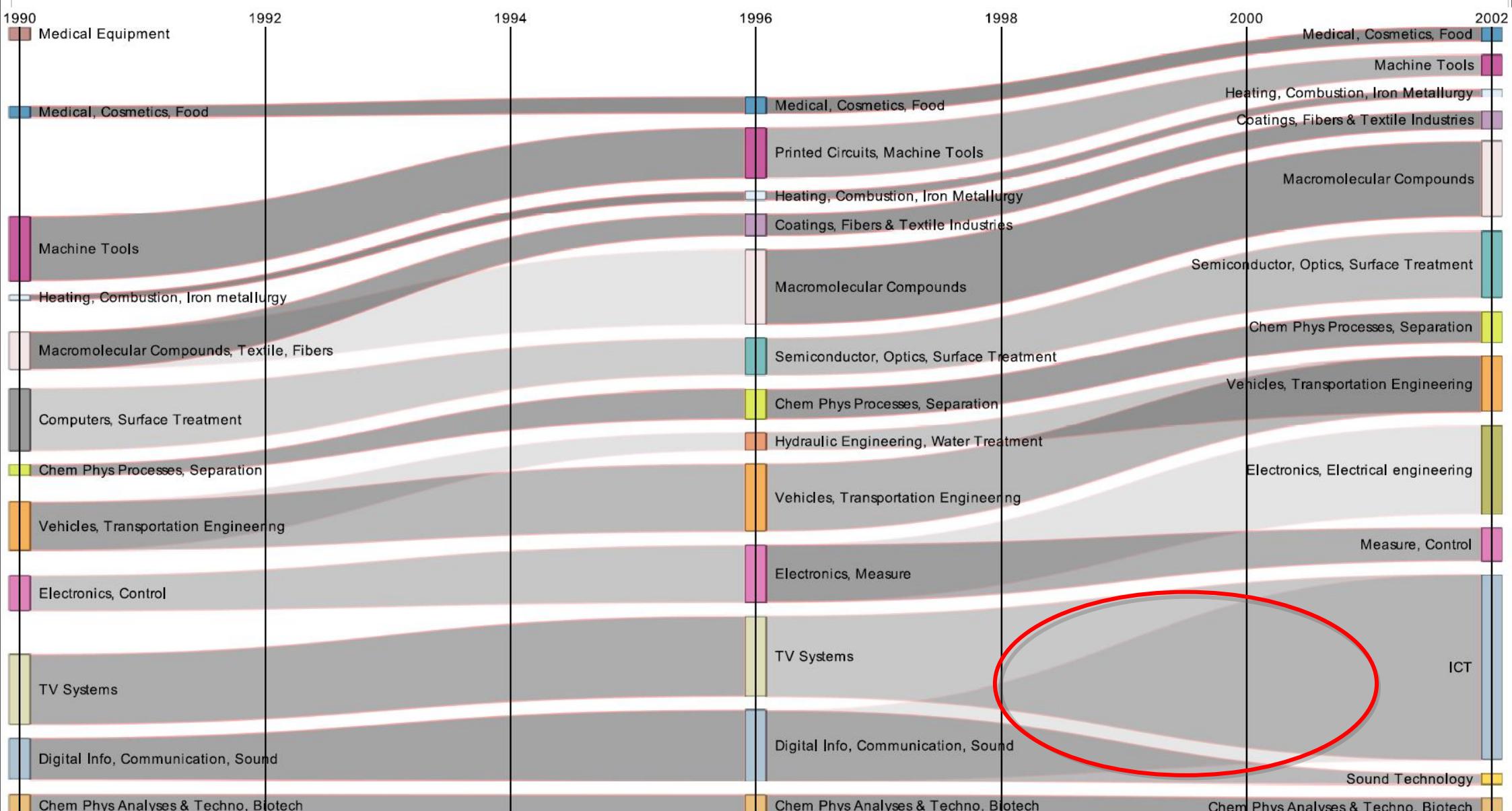
Projections sur cette carte

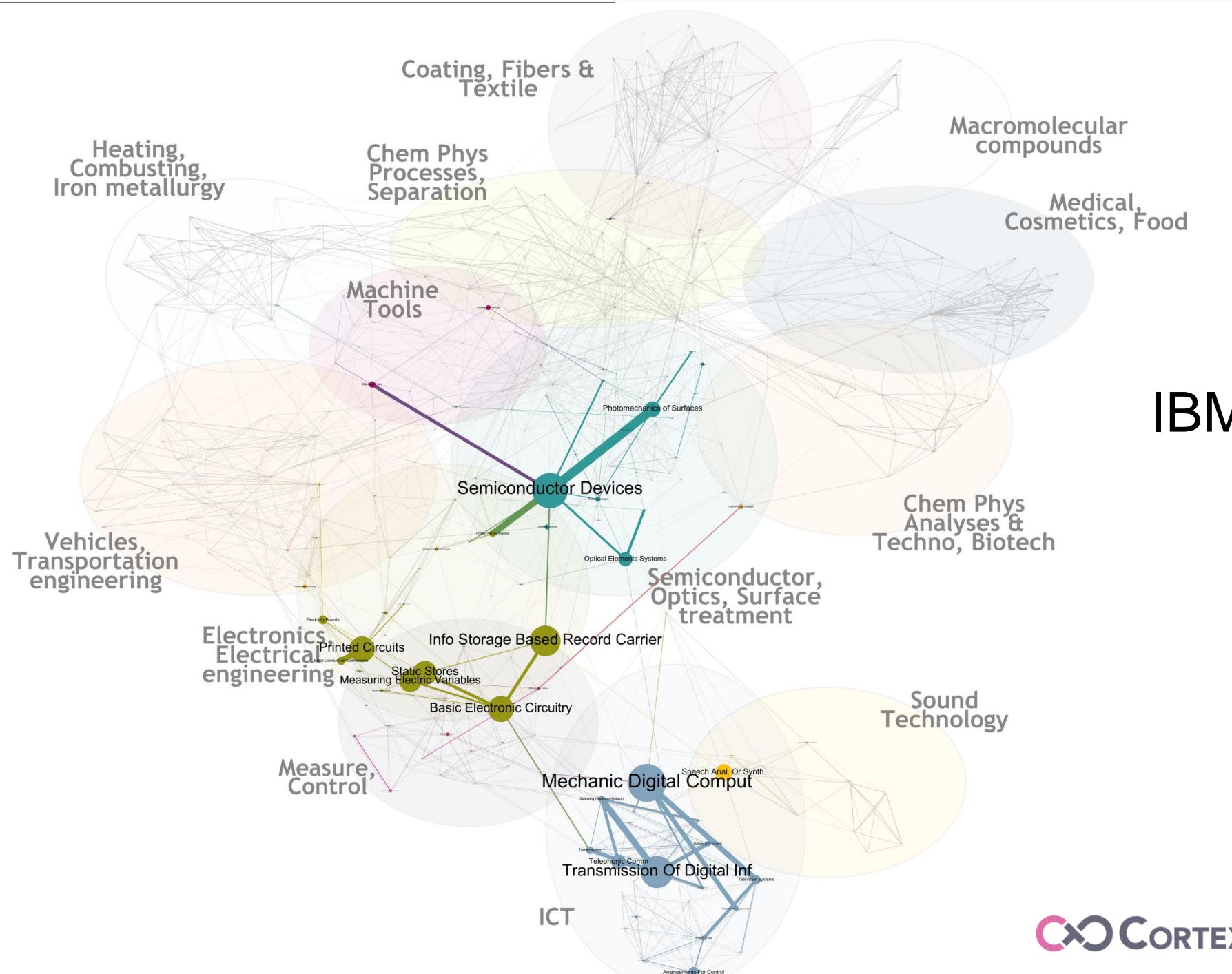
Figure 3. Publications profiles of the University of Amsterdam, Georgia Tech and London School of Economics (LSE) overlaid on the map of science.

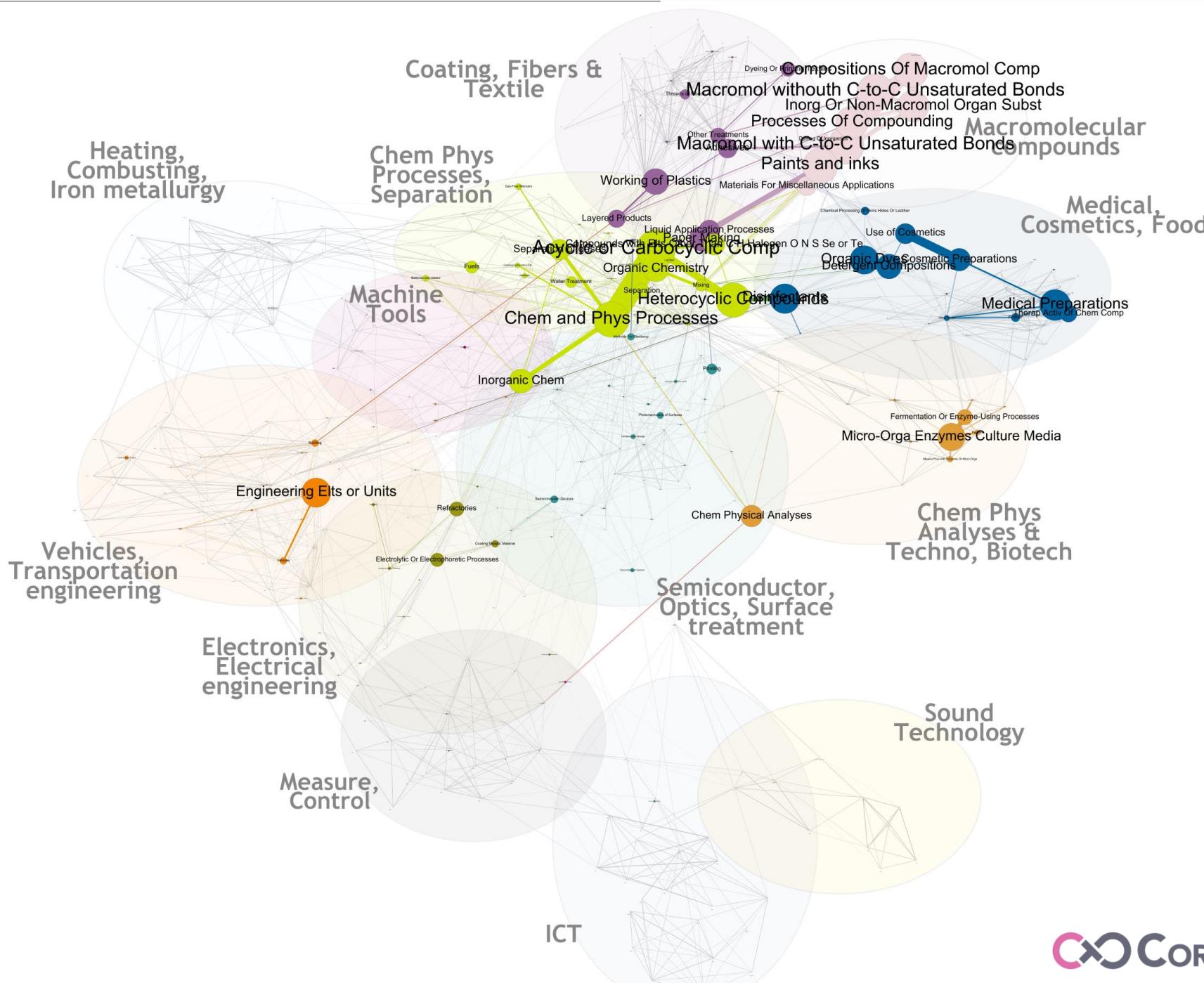
Science overlay maps: a new tool for research policy and library management, Ismael Rafols, Alan L. Porter, Loet Leydesdorff, 2010, ASIS&T



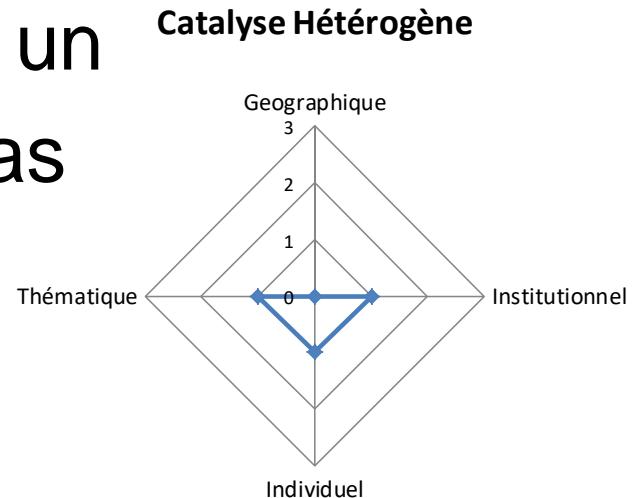
# Introduction – Scientométrie – Et au-delà – Données numériques – Cartographie de l'information – Echelles et analyse







# Co-activité (recherches et inventions) dans un champ scientifique et technologique : le cas de la catalyse hétérogène

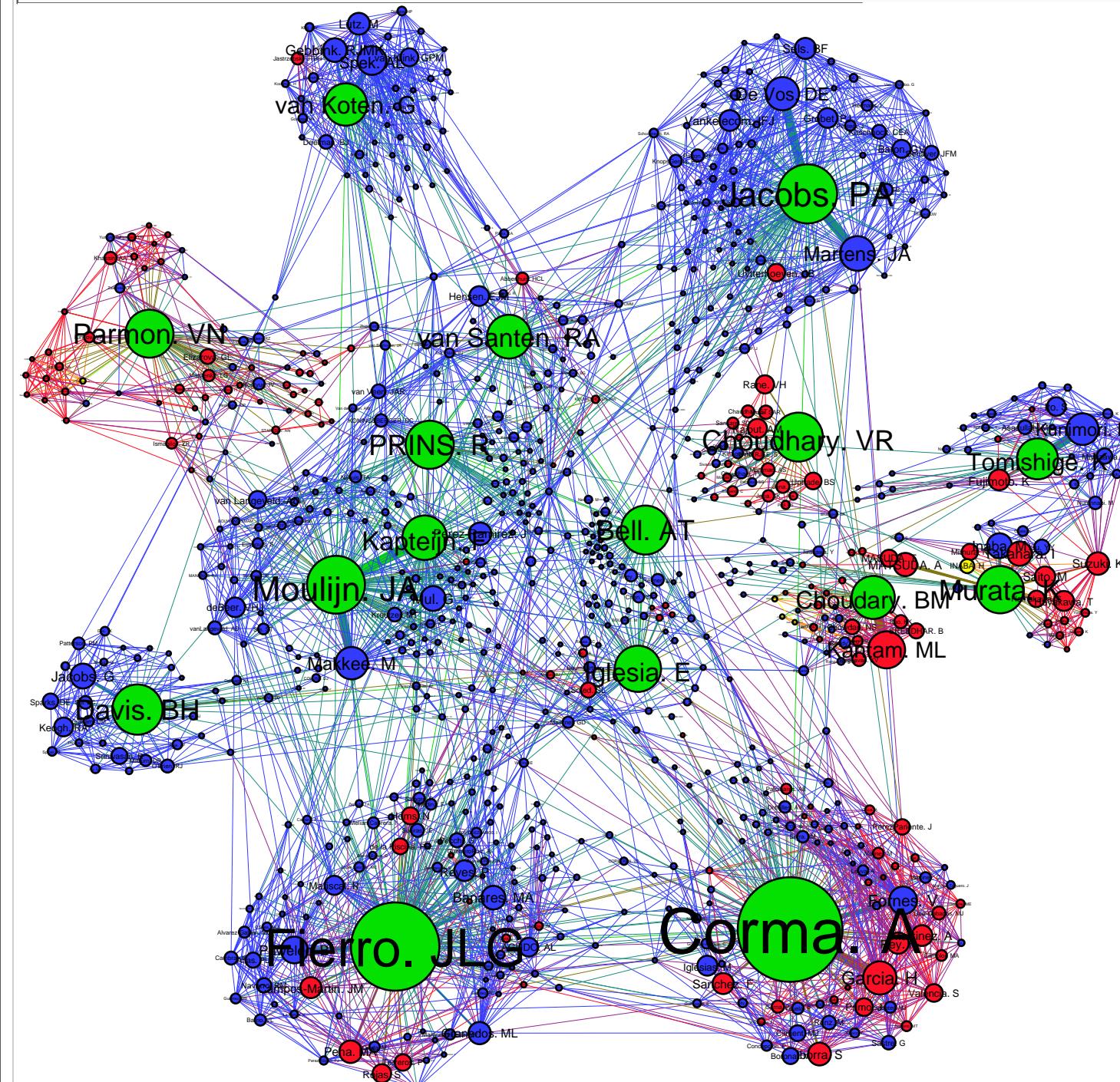


## Délinéation

- Identification d'un top16 d'auteurs centraux dans le champs scientifique de la catalyse hétérogène par enquêtes auprès d'experts.
- Publications et brevets associés

## Eléments principaux de méthode

- Réseau de collaborations de données hétérogènes



Analyse temporelle

Propension dans le temps à la co-activité en fonction de la structure du réseau de collaborateurs immédiats (brevets>publications)

# Bibliographie et ressources

## Références cartographie de l'information

Sémiologie graphique – Les diagrammes – les Réseaux – les cartes, Jacques Bertin, 2005 4ème ed., ed EHESS, 452 p.  
*Visual complexity, Mapping patterns of information*, Manuel Lima, 2011, Princeton Architectural Press, 272 p.  
*Atlas of science - Visualizing What We Know*, Katy Borner, 2010, The MIT Press, 272 p.  
*Information Graphics*, Sandra Rendgen, 2012, Taschen, 480 p.  
*Information Visualization, perception for design*, Colin Ware, MK, 2012  
*Simple Visualization Techniques for Quantitative Analysis, Now you see it*, Stephen Few, Analytics Press, 2009

## Références scientométrie et analyse de réseau

*Linked : How everything is connected to everything else and what it means for business, science, and everyday life*, Albert-Laszlo Barabasi, 2003, ed A Pulme Book, 294 p.  
*Réseau lexicométrique et réseau de citation pour la structuration de corpus*, M. Zitt, E. Bassecoulard Lereco, Observatoire des Sciences et de Techniques (OST), (2004) Paris  
*La scientométrie*, Michel Callon, Jean-Pierre Courtial, Hervé Penan, Que sais-je ?, PUF, 127p  
*Doing Data Science*, Cathy O'Neil & Rachel Schutt (2013)

## Références web

*History of books*, Wikipedia , visité le 06/02/2010, [http://en.wikipedia.org/wiki/History\\_of\\_books](http://en.wikipedia.org/wiki/History_of_books)  
*Data, data everywhere*, The economist, 25/02/2010, [www.economist.com/node/15557443](http://www.economist.com/node/15557443)  
*Crunching the numbers*, The economist, 19/05/2012, [www.economist.com/node/21554743](http://www.economist.com/node/21554743)  
*Beyond Visualization, Designing meaning through data experience*, Paolo Ciuccarelli, ENSCI, 2014, [www.dailymotion.com/video/x1ji4y4\\_paolo-ciuccarelli-20-mars-a-l-ensci\\_tech](http://www.dailymotion.com/video/x1ji4y4_paolo-ciuccarelli-20-mars-a-l-ensci_tech)  
*Infographics and Data Visualisation*, Trisnadi Kurniawan, 05/04/2014, <http://fr.slideshare.net/trisnadi/infographics-data-visualisation>, 2009

## Autres ressources web

*Sciences, technologies et visualisations*, [www.sciences-technologies.eu](http://www.sciences-technologies.eu)  
*Corporate Invention Board*, [www.corporateinventionboard.eu](http://www.corporateinventionboard.eu)  
*CorText Manager*, <http://manager.cortext.net>

Twitter : @ScTechViz (visualisation, Data mining, analyse de la science et de la technologie)