

Make Web Mining Great Again

Beyond pattern extraction

Mathieu Jacomy
Sciences Po Paris médialab
Equipex DIME-SHS ANR-10-EQPX-19-01

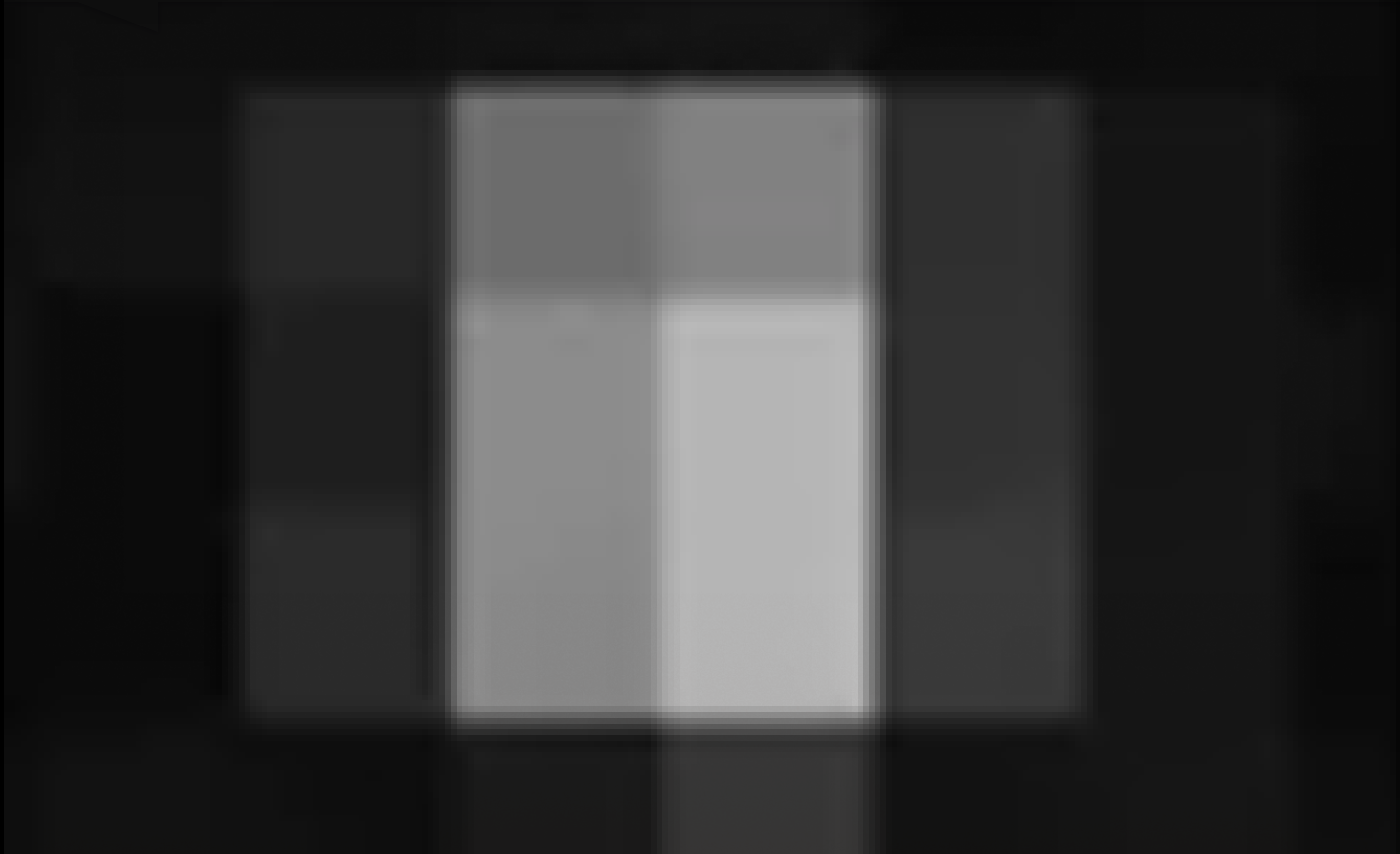


This seminary is possible thanks to



AALBORG UNIVERSITY
DENMARK

Explorable?

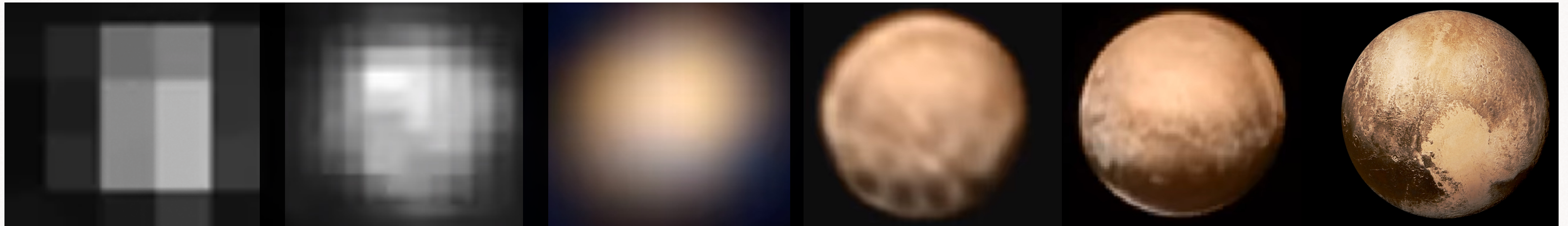


Explorable?



Poor knowledge

Rich knowledge

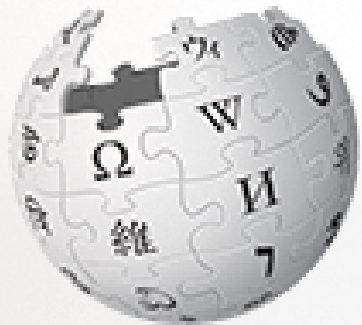


<- Different instruments ->
Different questions

Extracting patterns

Mining data

Wikipedia on "Data Mining"



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

[Interaction](#)
[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

[Article](#) [Talk](#)

[Read](#)

[Edit](#)

[View history](#)



Data mining

From Wikipedia, the free encyclopedia

Not to be confused with [analytics](#), [information extraction](#), or [data analysis](#).

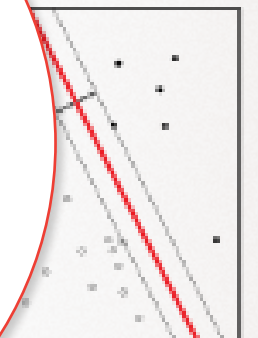
Data mining is an interdisciplinary subfield of [computer science](#).^{[1][2][3]} It is the computational process of discovering patterns in large data sets involving methods at the intersection of [artificial intelligence](#), [machine learning](#), [statistics](#), and [database systems](#).^[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.^[1] Aside from the raw analysis step, involves database and [data management](#) aspects, [data pre-processing](#), [model](#) and [inference](#) considerations, interestingness metrics, [complexity](#) considerations, post-processing of discovered structures, [visualization](#), and [online updating](#).^[1] Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.^[4]

The term is a [misnomer](#), because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data

1. Disembled
information

2. Make it
understandable

mining and
g



[\[show\]](#)

Supervised learning
([classification](#) • [regression](#))

[\[show\]](#)

Clustering

[\[show\]](#)

Dimensionality reduction

[\[show\]](#)

Structured prediction

[\[show\]](#)

Anomaly detection

[\[show\]](#)

The mining metaphor

<http://openrefine.org/>

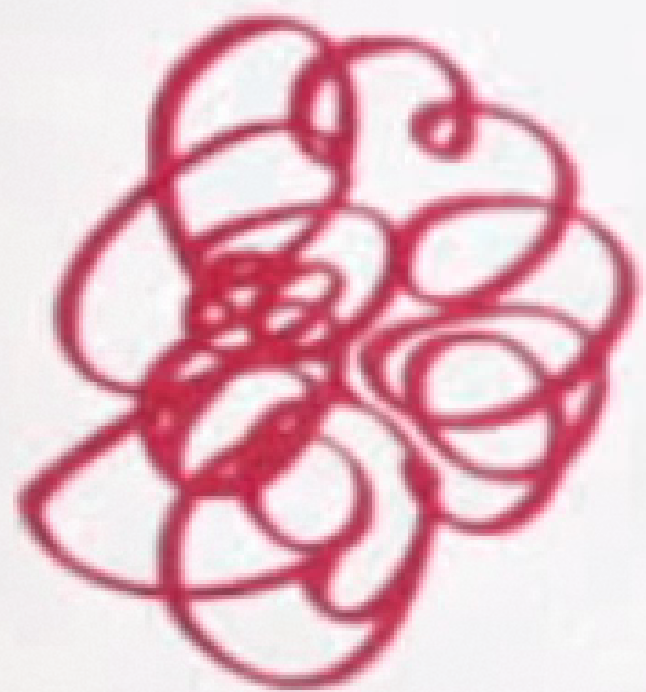


Image from
the OpenRefine
tutorial

Example: a typical data-mining paper

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13023/12752>

A recent example

Mining Pro-ISIS Radicalisation Signals from Social Media Users

Matthew Rowe & Hassan Saif

Published in ICWSM16 (International Conference on Web and Social Media)

From the abstract:

“In this paper our aim to understand what happens to Europe-based Twitter users before, during, and after they exhibit pro-ISIS behaviour (...), characterising such behaviour as radicalisation signals. We adopt a data-mining oriented approach to computationally determine time points of activation (i.e. when users begin to adopt pro-ISIS behaviour), characterise divergent behaviour (both lexically and socially), and quantify influence dynamics as pro-ISIS terms are adopted.”

Example: a typical data-mining paper

1. Obtain a corpus of Tweets and users

A seed of ~650 influent Twitter users from a previous paper. Clean, expand and filter users. Collect their 104M tweets.

2. Determine the pro-ISIS users

Create an empirically-validated lexicon of pro-ISIS keywords. Tag users if they use pro-ISIS keywords or follow pro-ISIS accounts. The moment it happens is called “activation point” for these 727 users.

3. Look at what the “activation” changes

Make 3 time slices per “activated user”: before, around and after activation.

Transliterate arabic to english. Measure variations of vocabulary, tweets sharing, and mentioning. Qualify the changing vocabulary with a sentiment analysis algorithm

4. Search what is influencing users

Set an “adoption probability” of pro-ISIS terms, using a diffusion model. Compute adoption probabilities for vocabulary, tweets sharing, and mentioning as influences. Measure their statistical accuracy.

Example: a typical data-mining paper

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13023/12752>

Findings

- 1) People with a similar profile influence you more
- 2) Becoming “pro-ISIS” is a gradual change

“Firstly, we found that social dynamics play a strong role in term uptake where users are more likely to adopt pro-ISIS language from users with whom they share many interacted users (...).

Secondly, prior to being activated, users go through a period of significant increase in [communicating with new users and adopting new terms], this clear increase suggests that users are rejecting their prior behaviour and escalating this further until becoming activated.”

Example: a typical data-mining paper

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13023/12752>

This is “mining” in the sense that it aims at retrieving “patterns” that actually exist in the data, where “actually” means “statistical evidence provided”.

“Our aim here was to **disentangle** different influence factors that govern the adoption process.”

“Throughout this research we have adopted an exploratory [sic!] data mining approach by **collecting data and then analysing it** based on our **hypothesised** signals of radicalisation”

Research questions

http://archives.cerium.ca/IMG/pdf/WIKTOROW-ICZ_2006_Anatomy_of_the_Salafi_Movement.pdf

Research questions in social sciences are sensibly different.

Example from sociology

Anatomy of the Salafi Movement

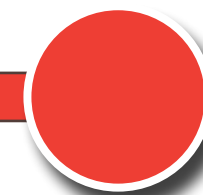
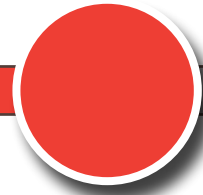
Q. Wiktorowicz

Studies in Conflict & Terrorism, 2006

From the abstract:

“This article explains the **sources of unity that connect violent extremists with nonviolent puritans**. Although Salafis share a common religious creed, **they differ over their assessment of contemporary problems** and thus how this creed should be applied. Differences over contextual interpretation have produced three major Salafi factions: purists, politicos, and jihadis.”

What is influencing Twitter users
about adopting pro-ISIS language:
vocabulary, retweets, mentions?



How a different assessment of
contemporary problems plays a role in
observed pro-ISIS behaviors on Twitter?

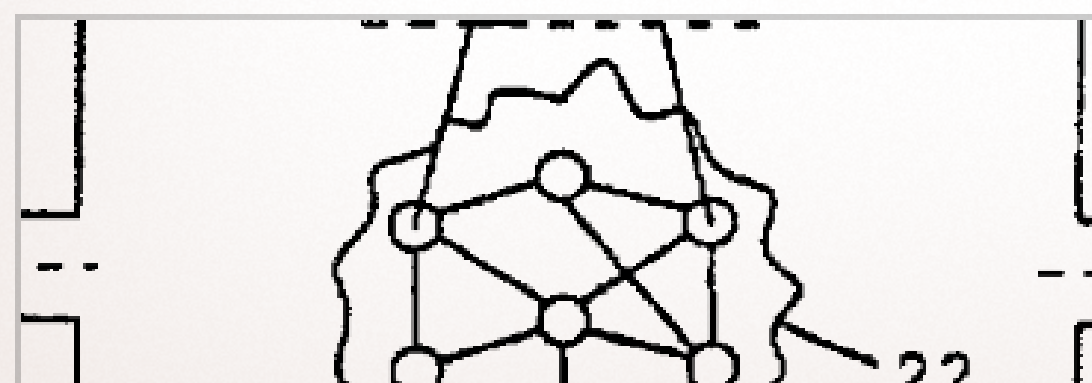
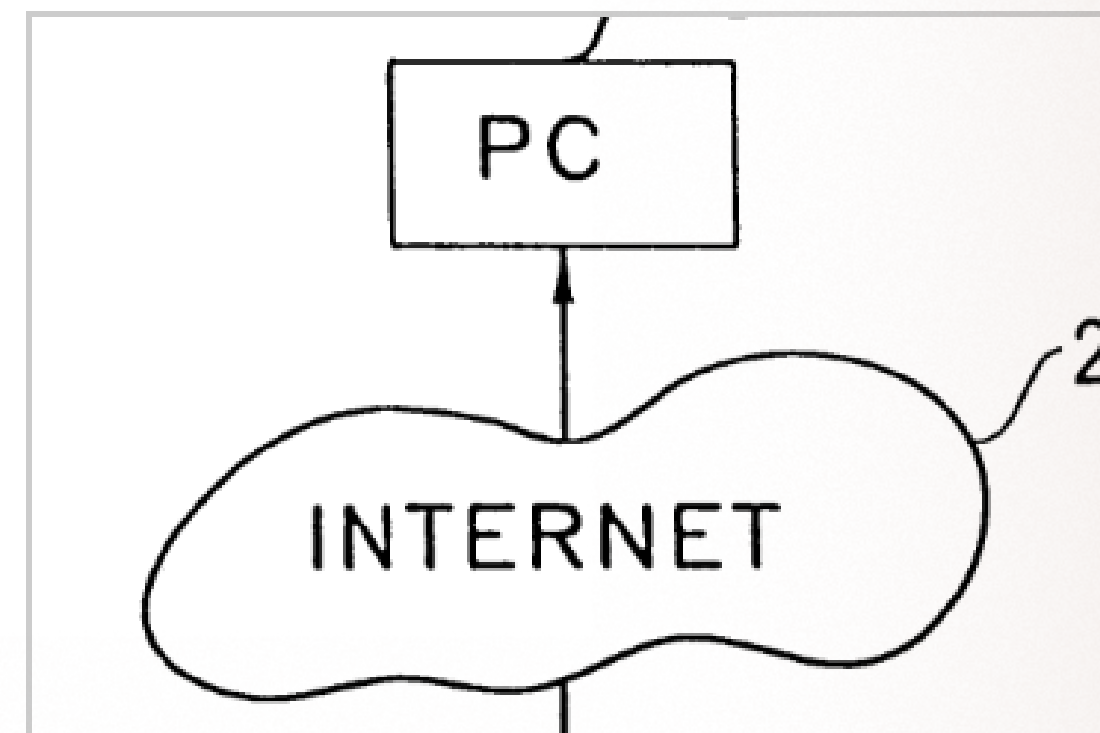
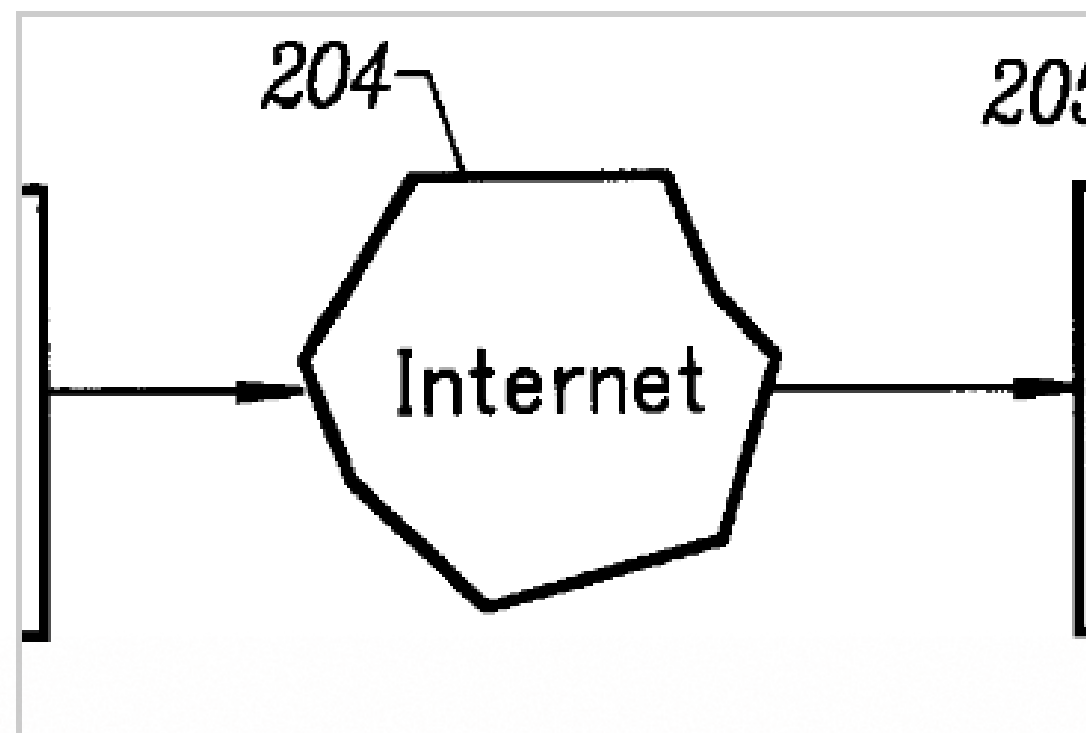
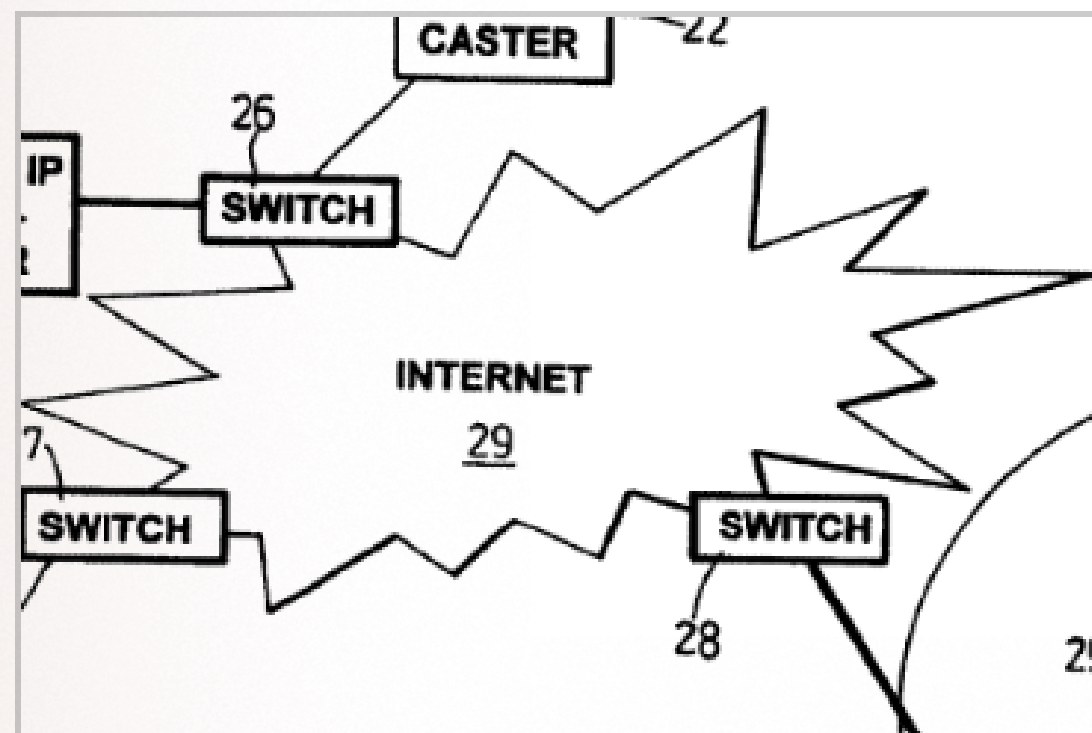
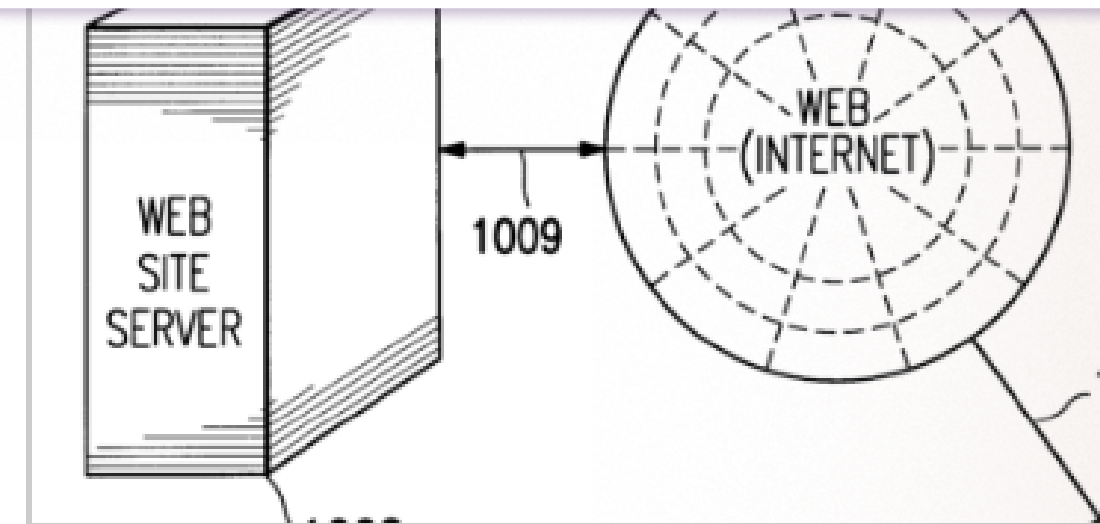
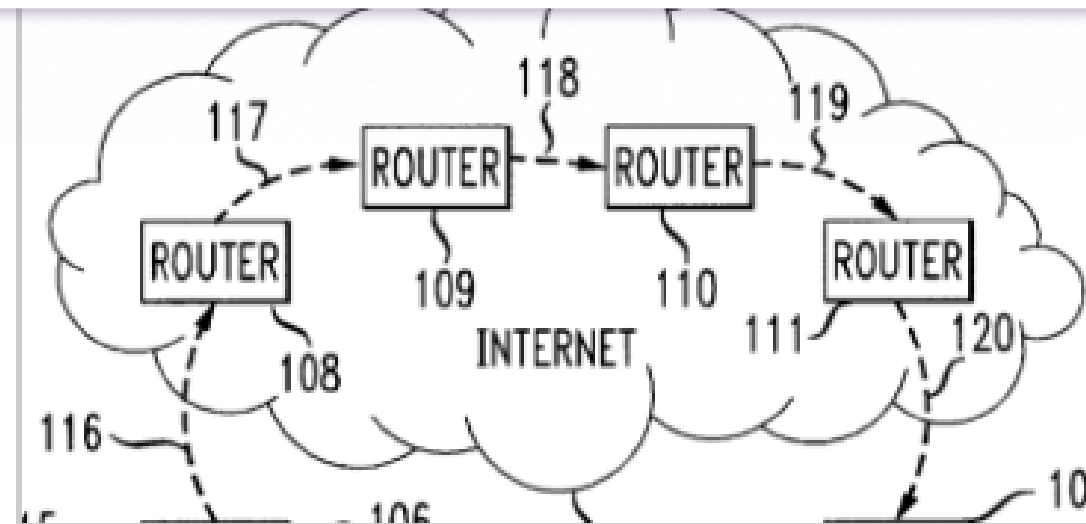
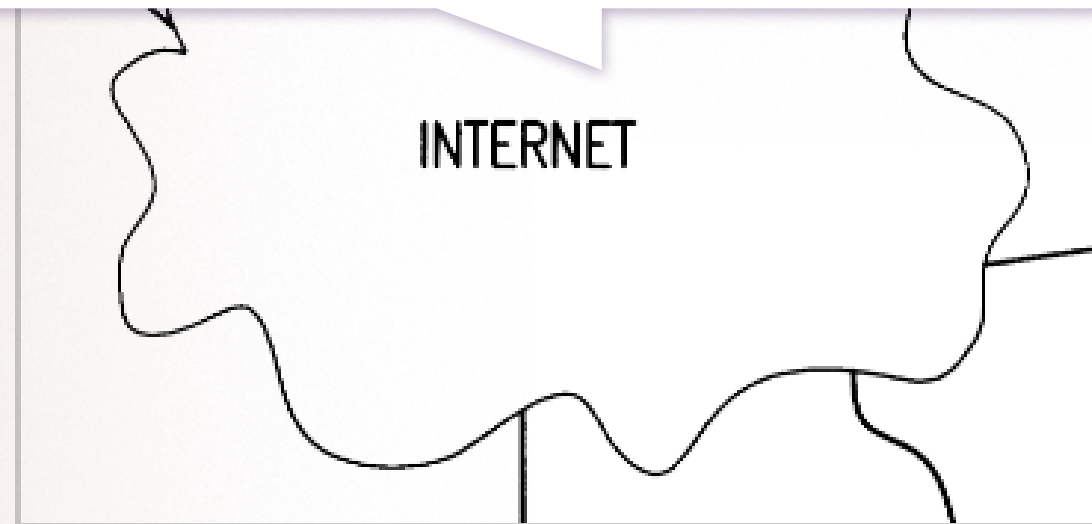
...if
we had
instruments
precise
enough!

How we imagine the web's shape

Various representations

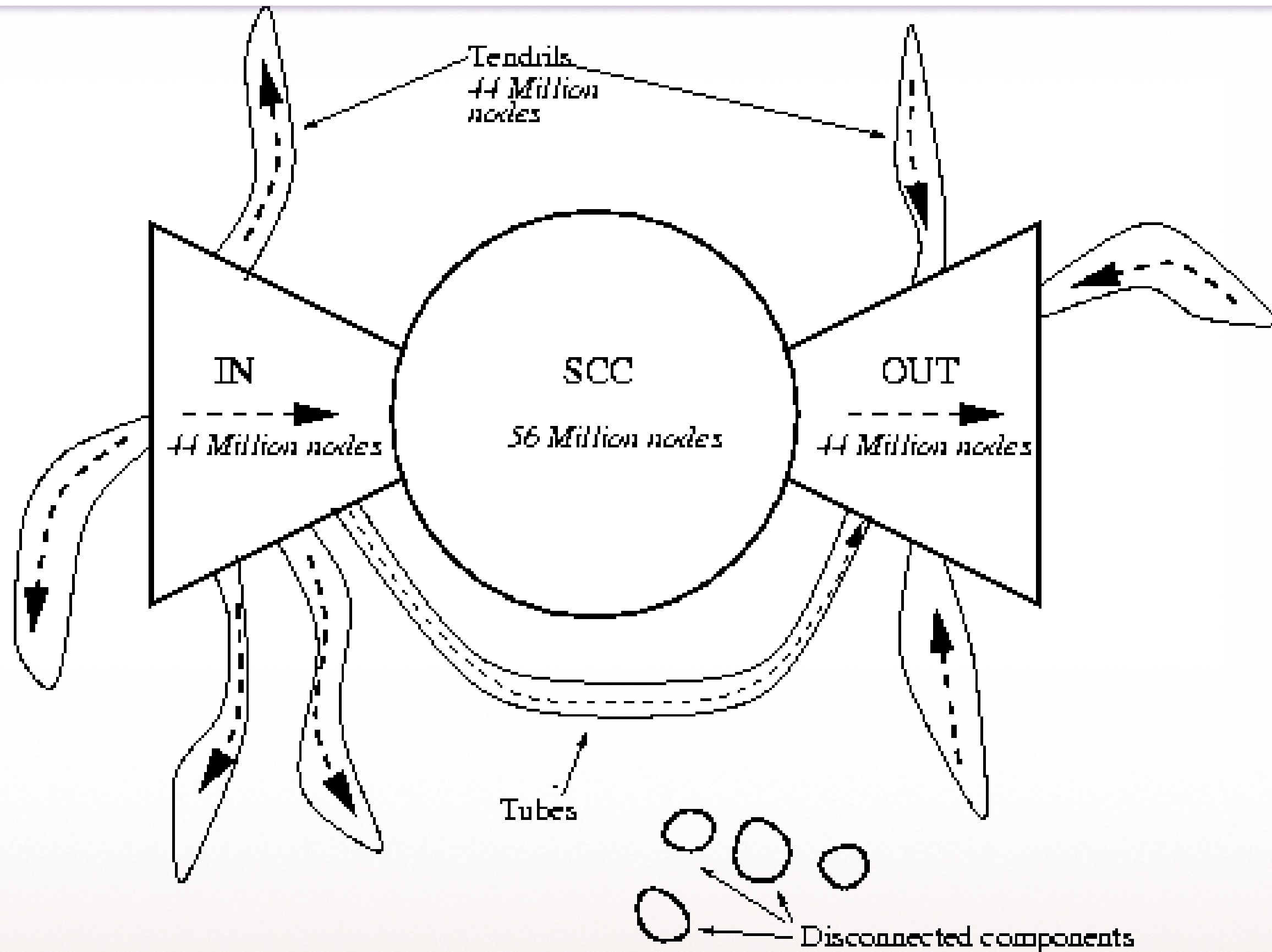
Depictions of internet in academic papers

<http://noahveltman.com/internet-shape/>



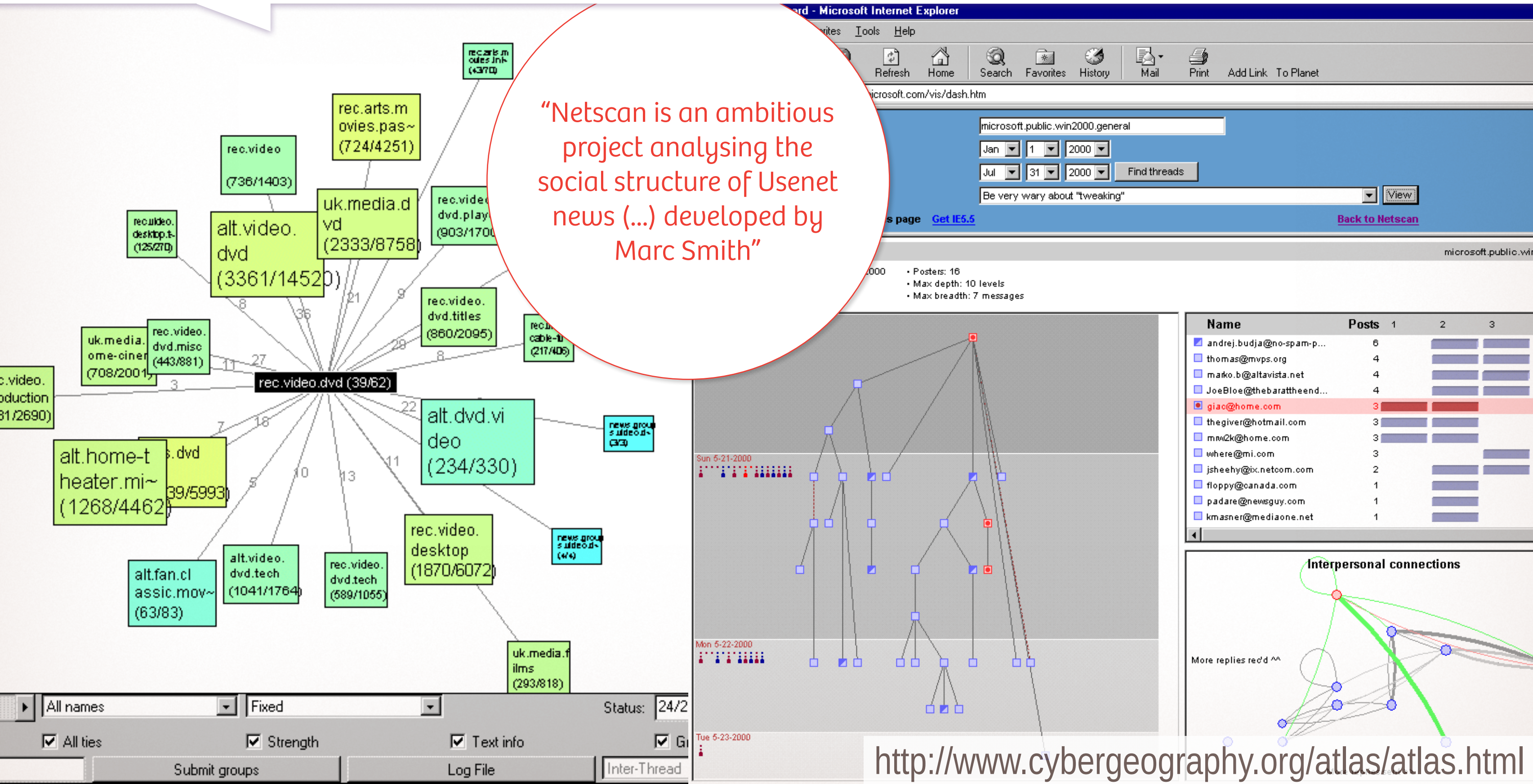
The first scientific image of the web

The bow tie,
IBM's Almaden Research, 2000



An Atlas of Cyberspaces: Early depictions of **social networks***

"Netscan is an ambitious project analysing the social structure of Usenet news (...) developed by Marc Smith"



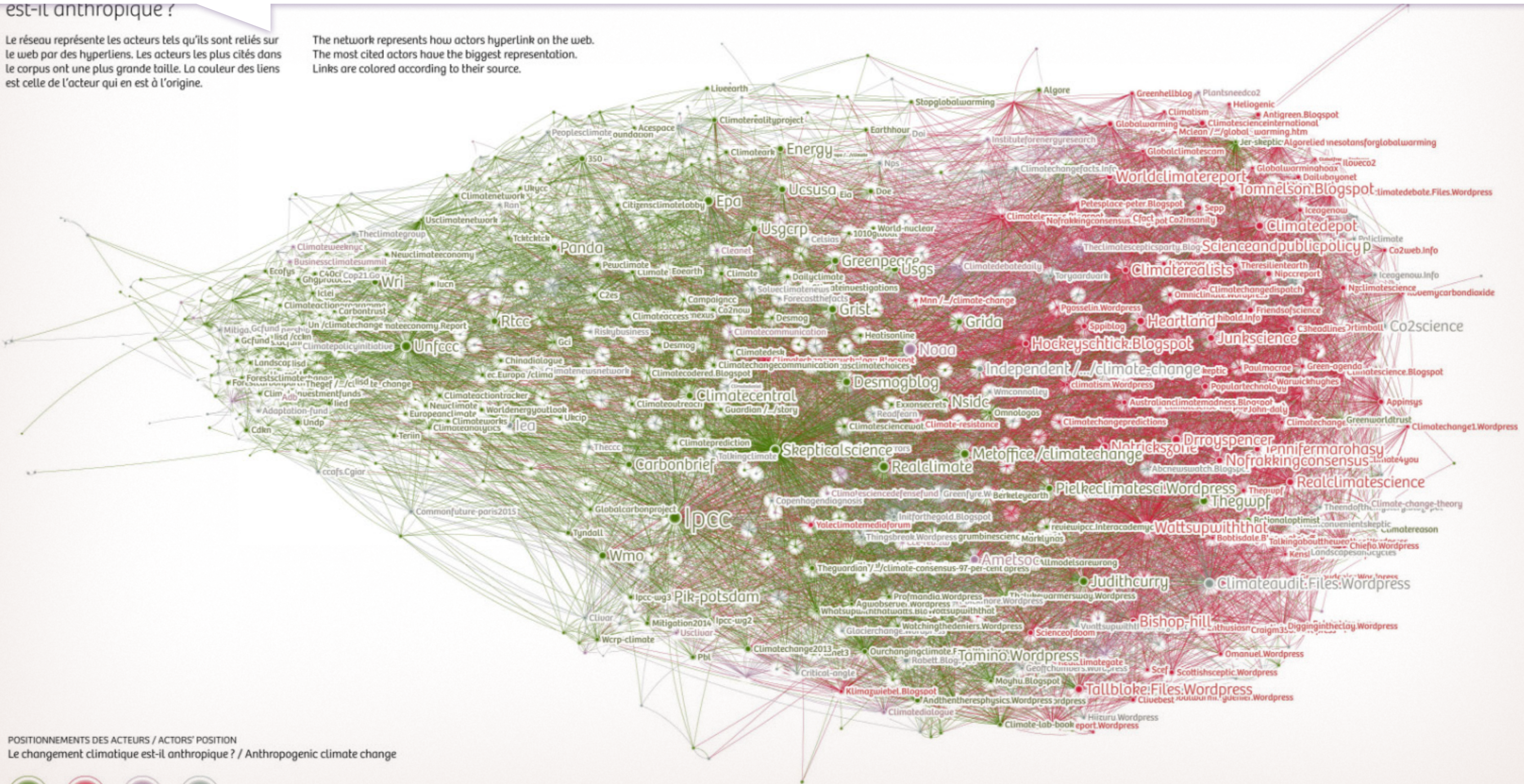
A fully analyzed corpus of a domain (climate change)

Climate change debate on the web
Sciences Po médialab (2015)

est-il anthropique ?

Le réseau représente les acteurs tels qu'ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l'acteur qui en est à l'origine.

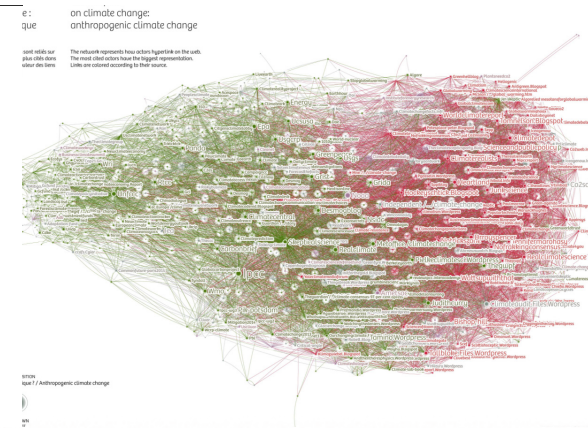
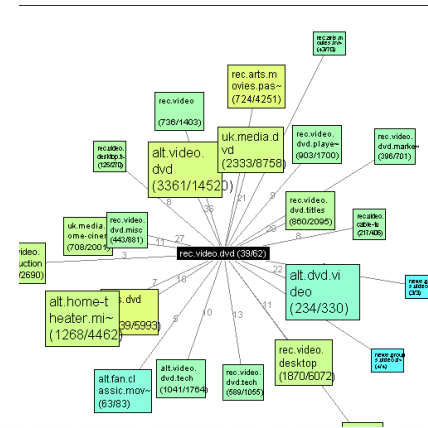
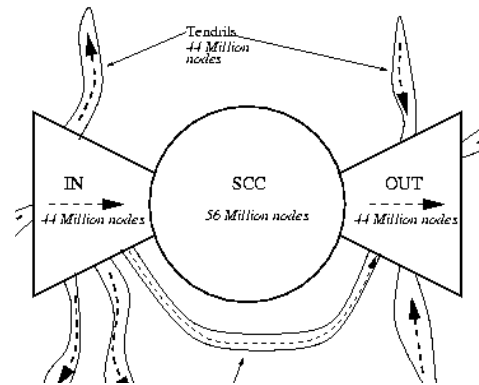
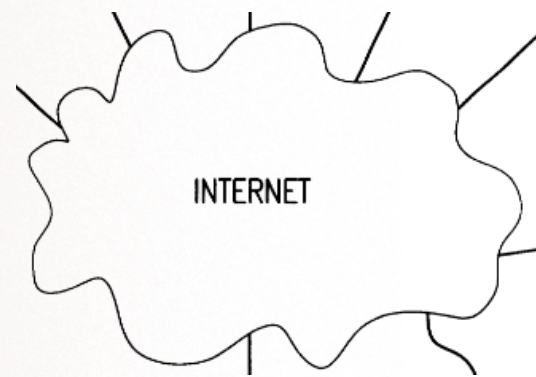
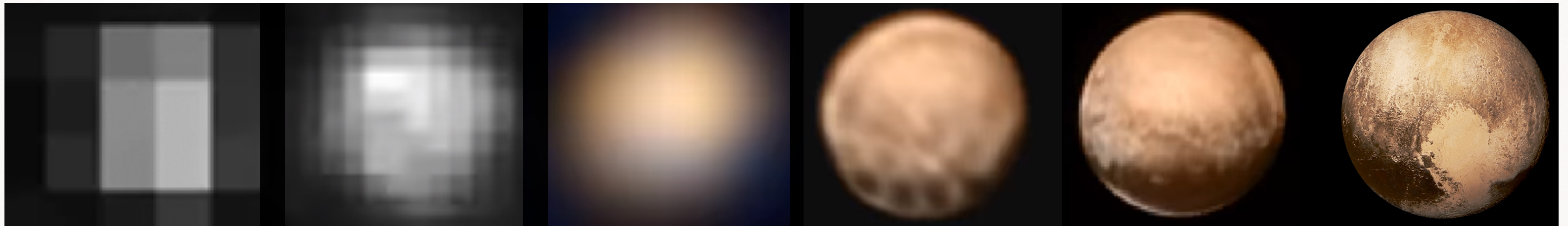
The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.



Which degree of knowledge?

Poor knowledge

Rich knowledge



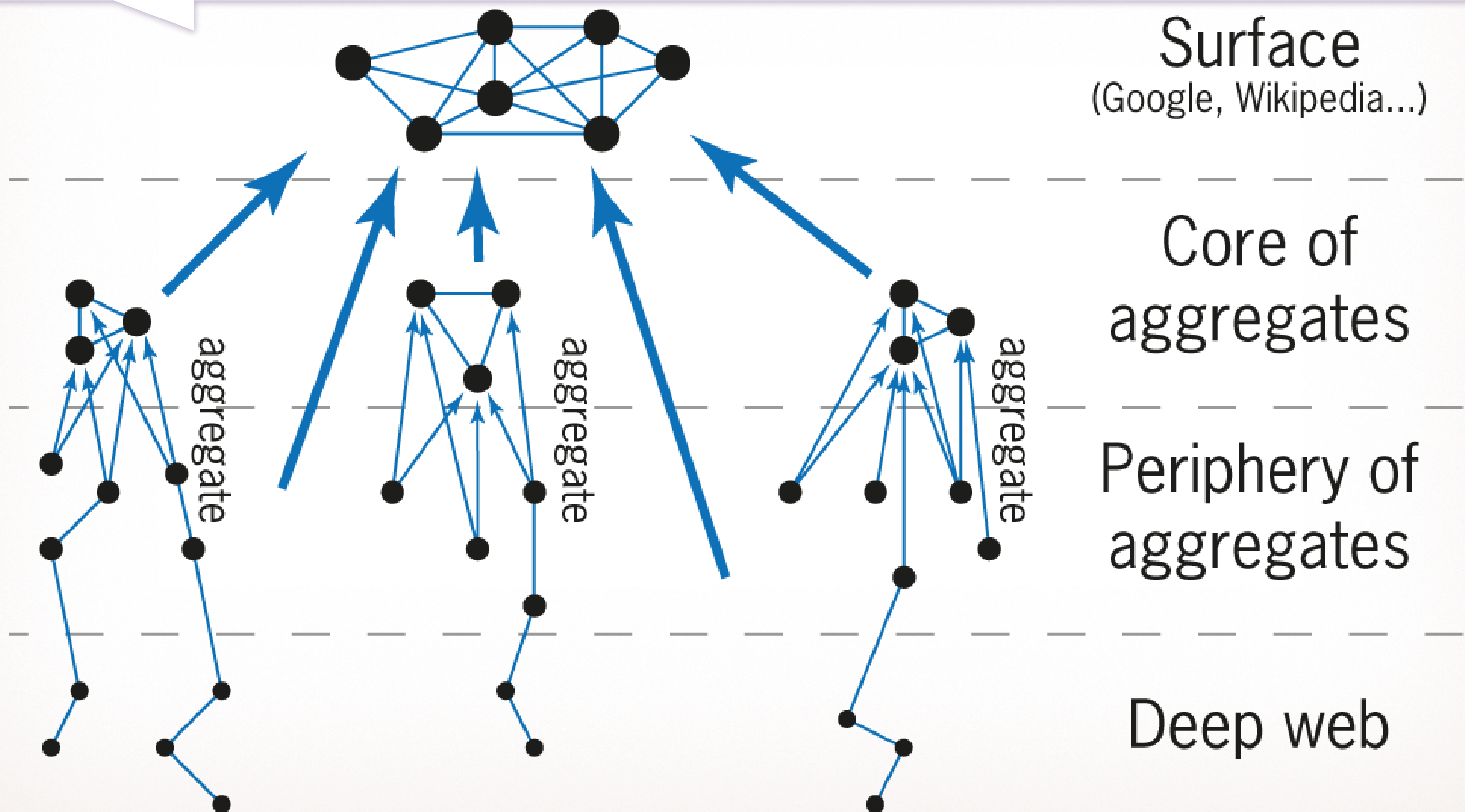
...

<- Different questions ->

The web as layers

Hyphe's model of the web

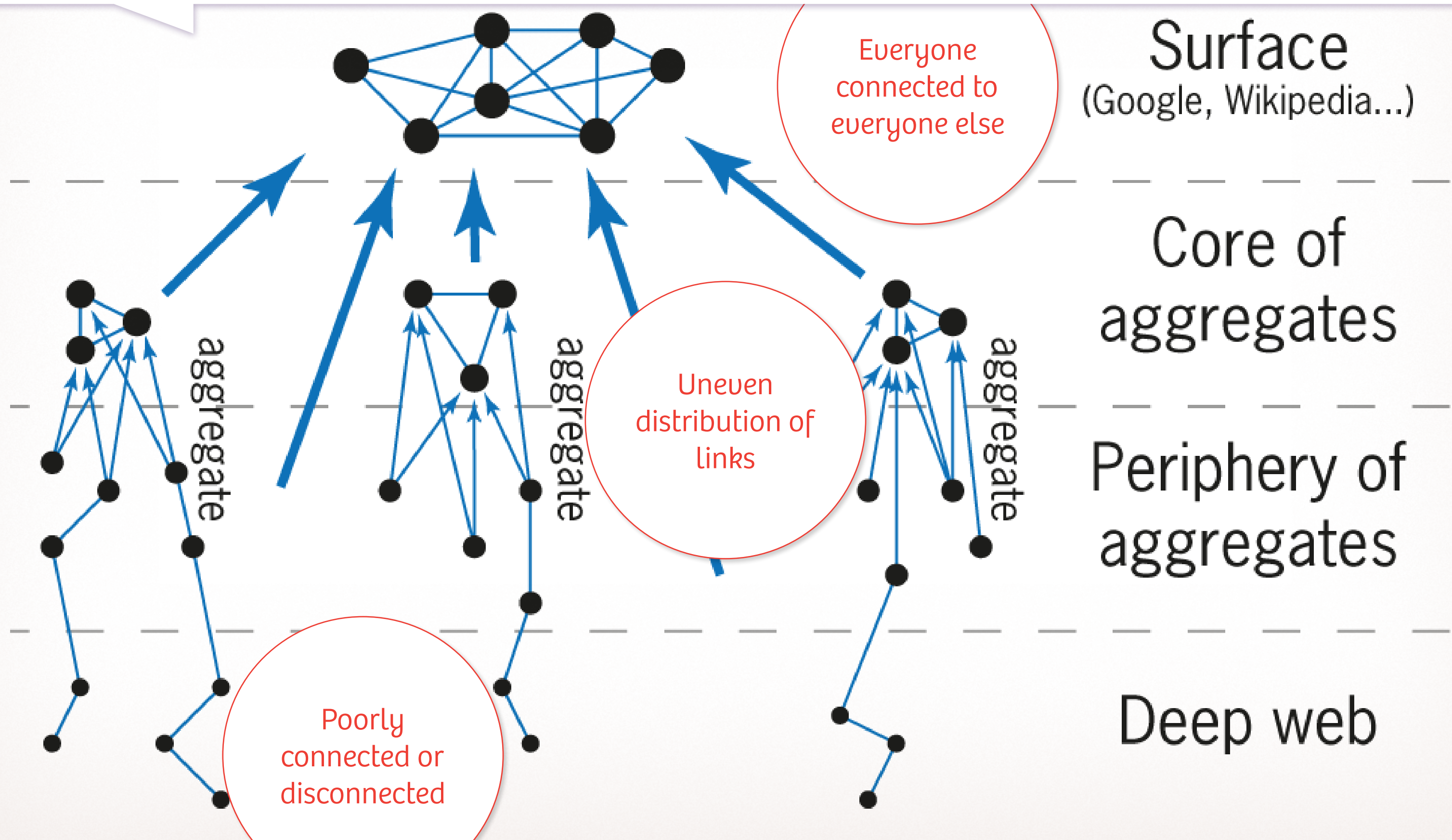
Web as layers



Web as layers

Connectivity

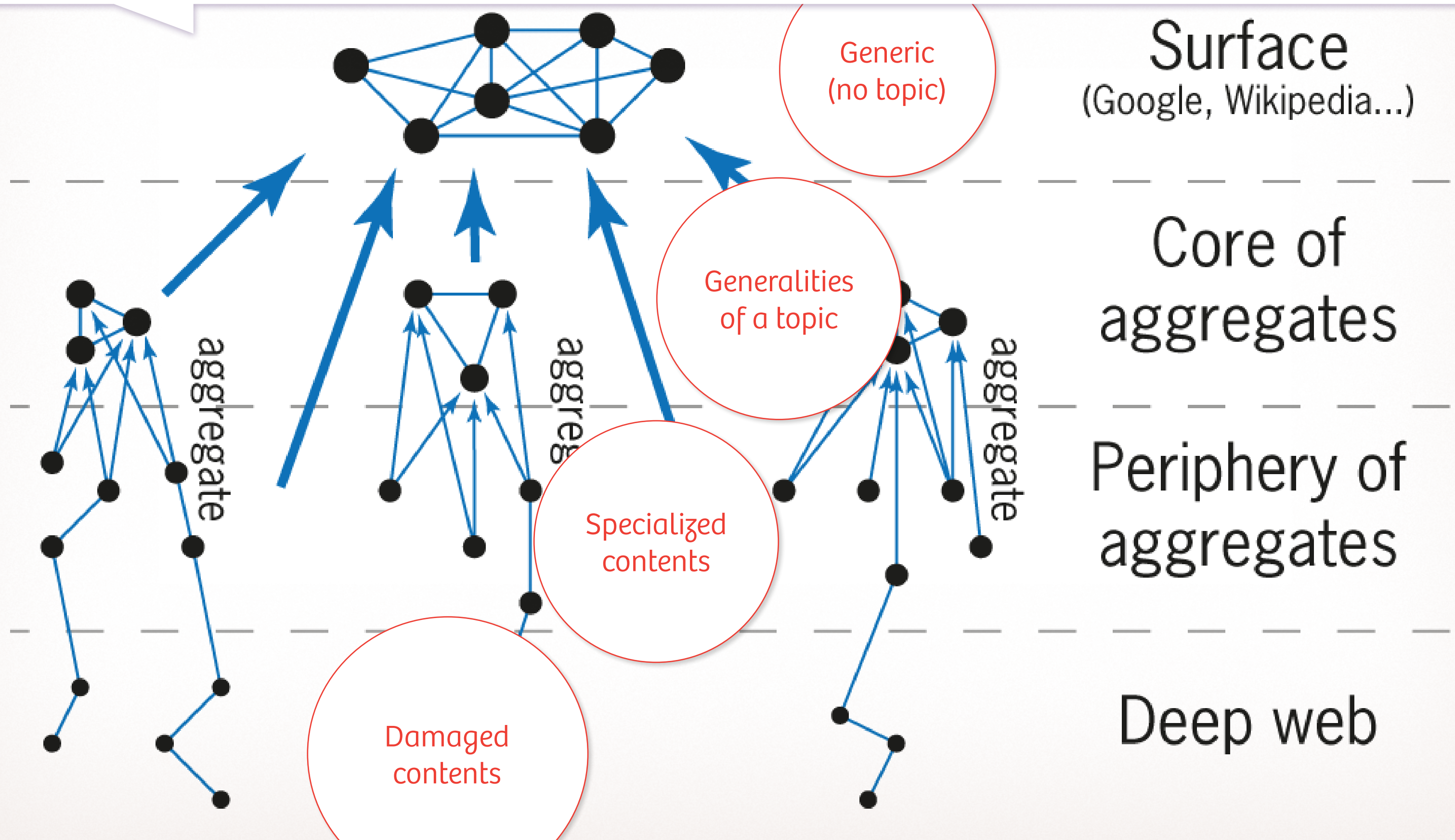
Franck Ghitalla & Mathieu Jacomy
<https://ateliercartographie.wordpress.com/>



Web as layers

Contents

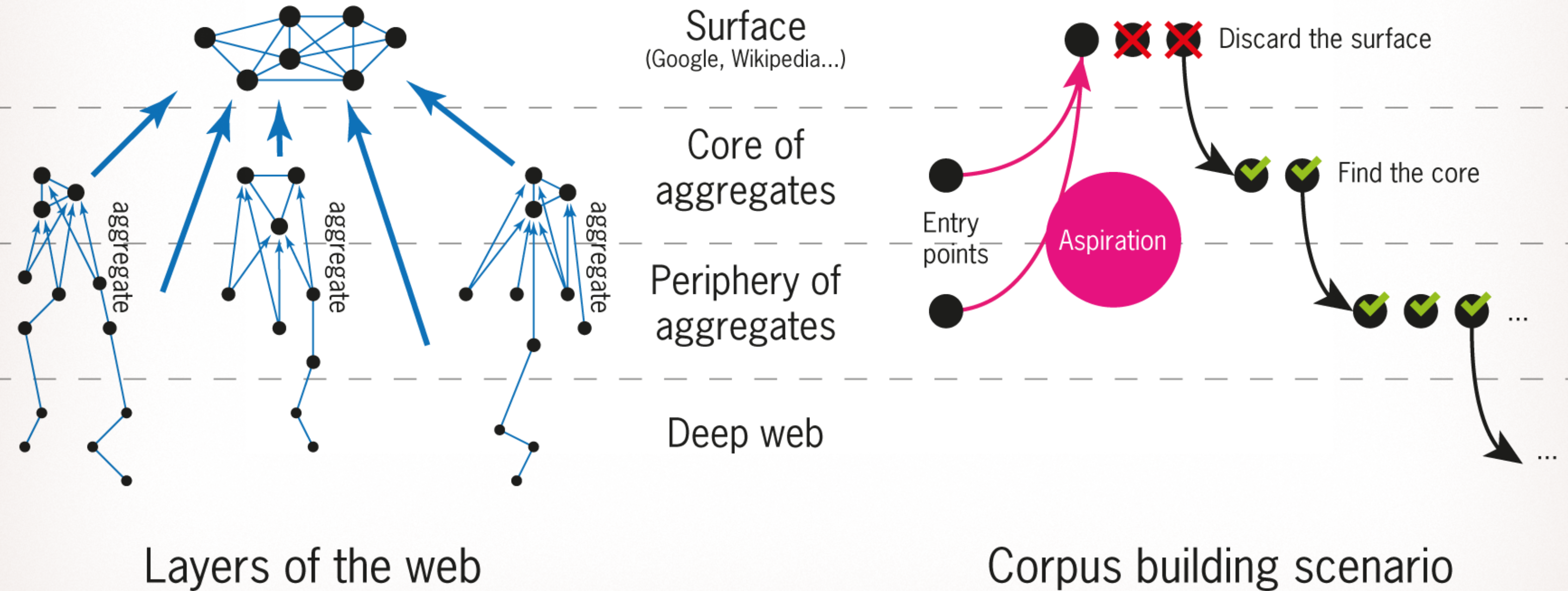
Franck Ghitalla & Mathieu Jacomy
<https://ateliercartographie.wordpress.com/>



Web as layers

Curation process

Franck Ghitalla & Mathieu Jacomy
<https://ateliercartographie.wordpress.com/>



Hyphe's method

Exploratory web-mining methodological chain

1. Sourcing

Define your field a priori
and gather starting points

2. Harvesting (crawl)

Download the data
with a crawler

3. Monitoring

Visualize corpus
and monitor its properties


4. Curation

Select documents to limit
topic drifting and adjust
corpus boundaries

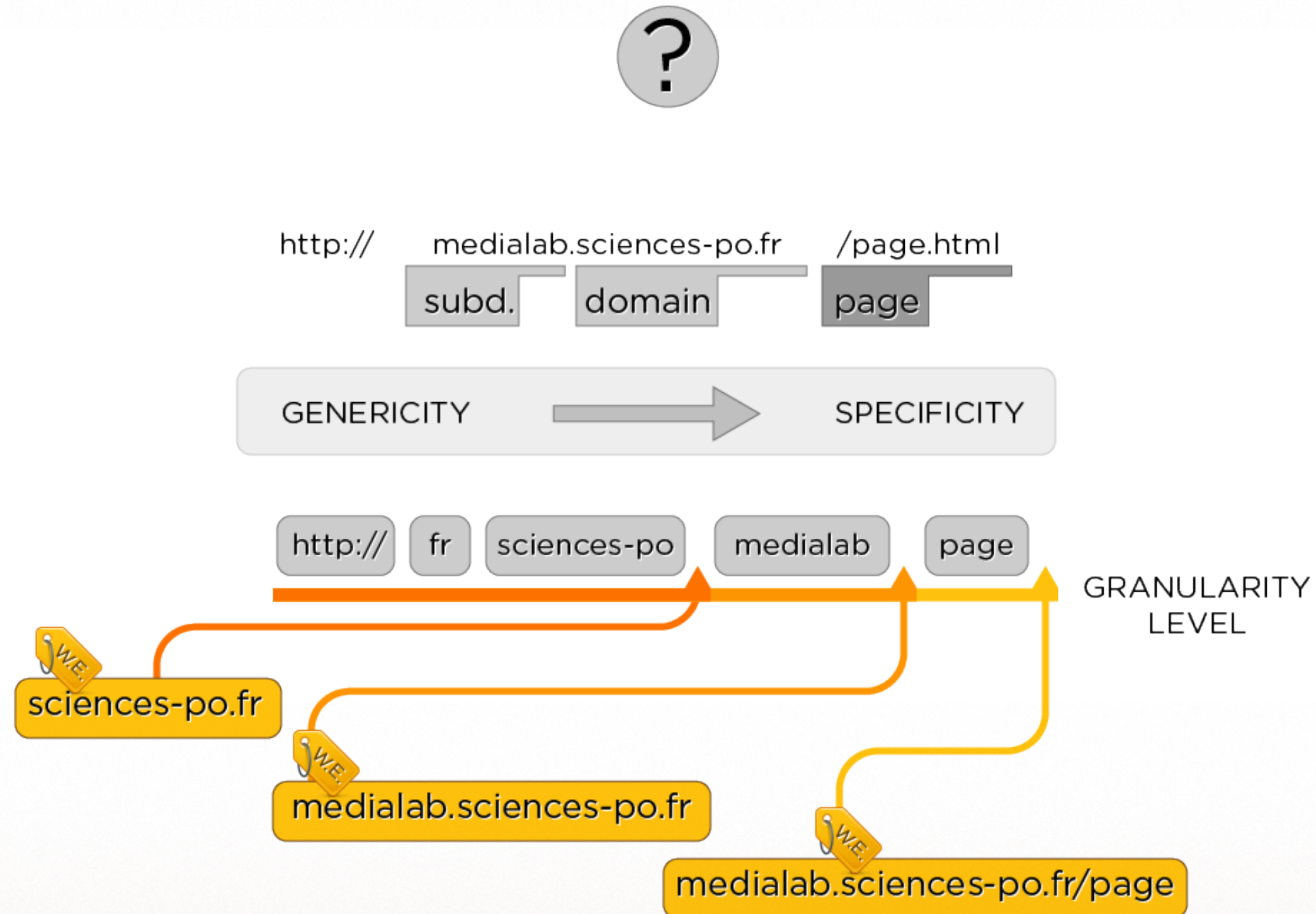
5. Finalization

Validate general quality
and export corpus

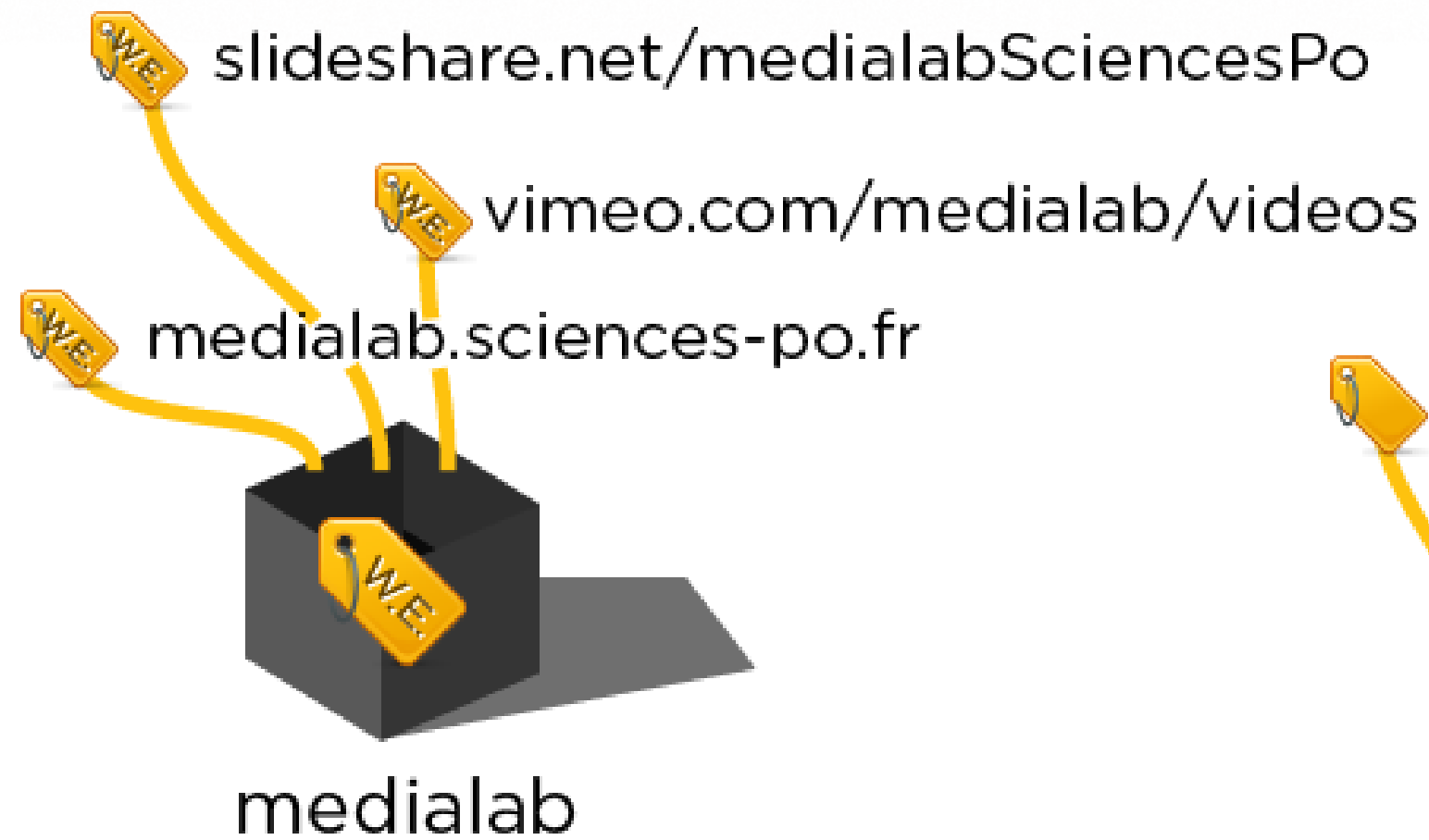
Exploratory web-mining methodological chain

- 
1. Sourcing
Define your field a priori
and gather starting points
 2. **Harvesting** (crawl)
Download the data
with a crawler
 3. **Monitoring**
Visualize corpus
and monitor its properties
 4. **Curation**
Select documents to limit
topic drifting and adjust
corpus boundaries
 5. Finalization
Validate general quality
and export corpus

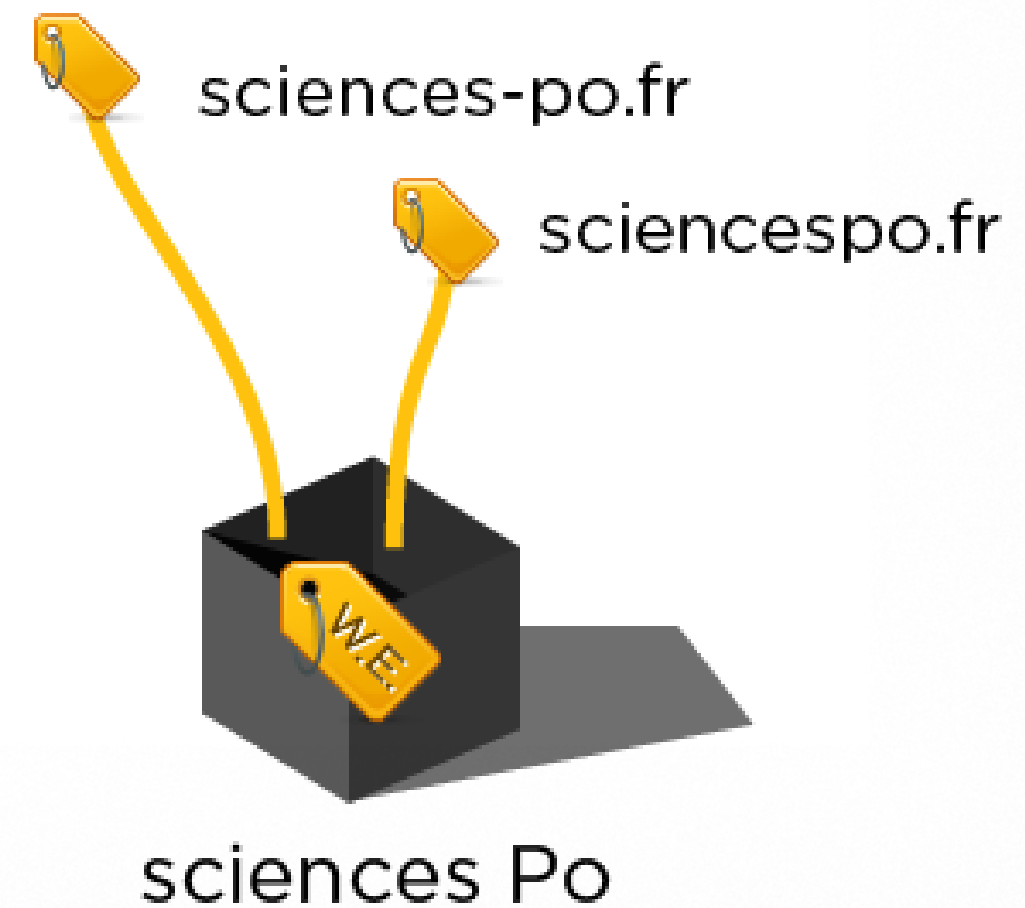
Web entities



Web entities

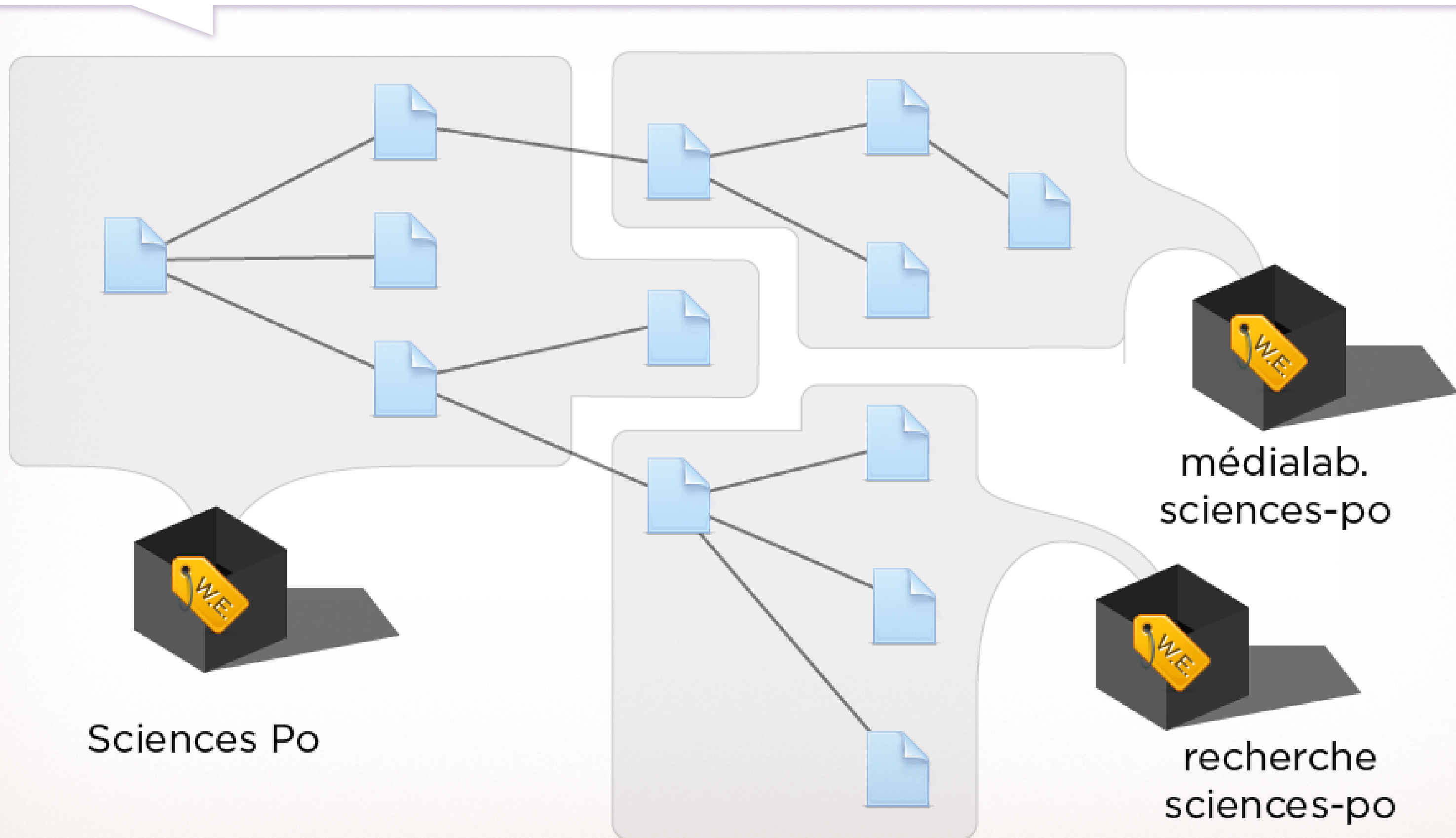


ACTOR'S PRESENCE
ON THE WEB



ALIASES

Web entities



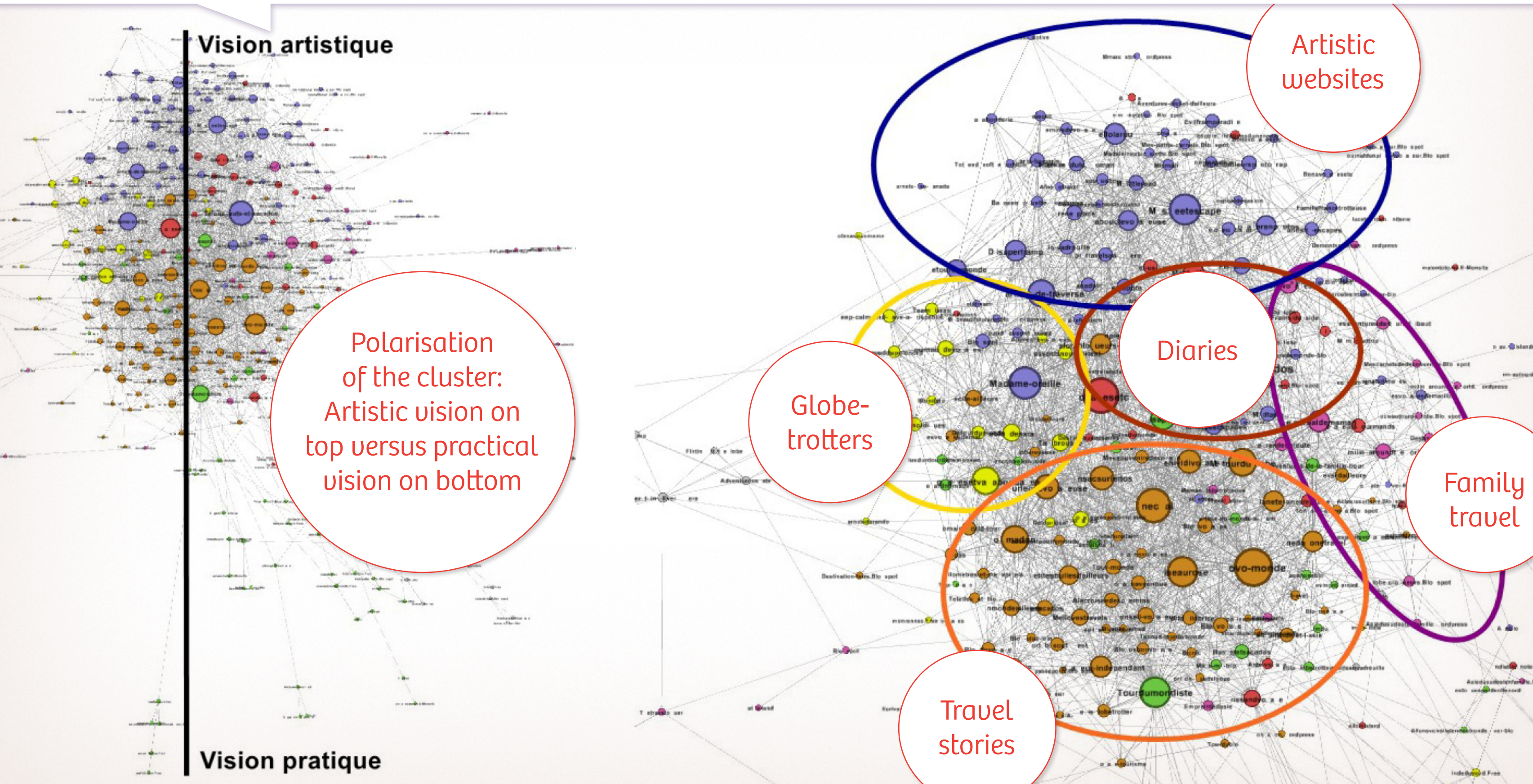
Students using Hyphe

Examples

Examples from Hyphe teachings

Travel

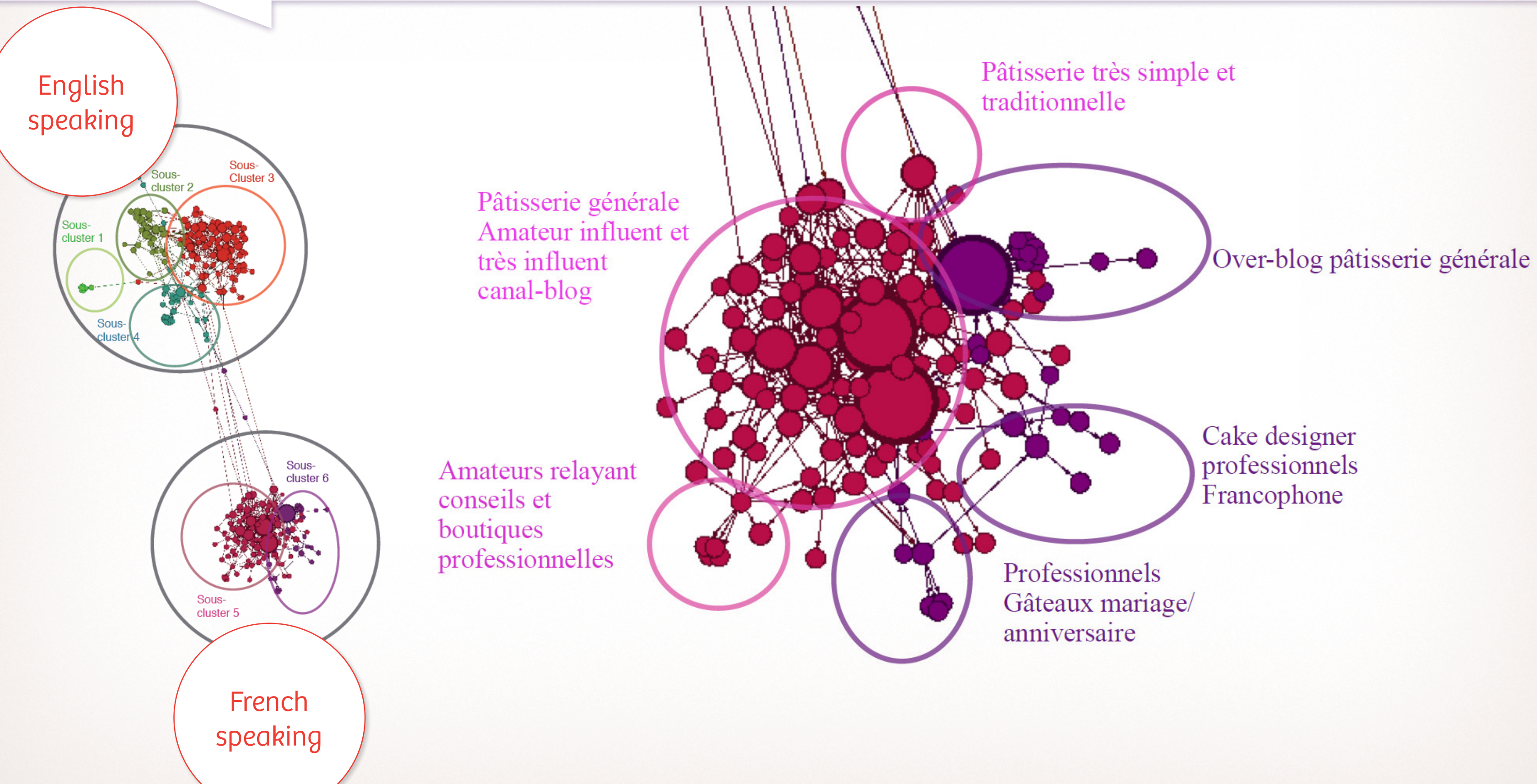
Course in Université Paris Descartes
“Méthodes d'enquête complémentaires”



Examples from Hyphe teachings

Bakery (pâtisserie)

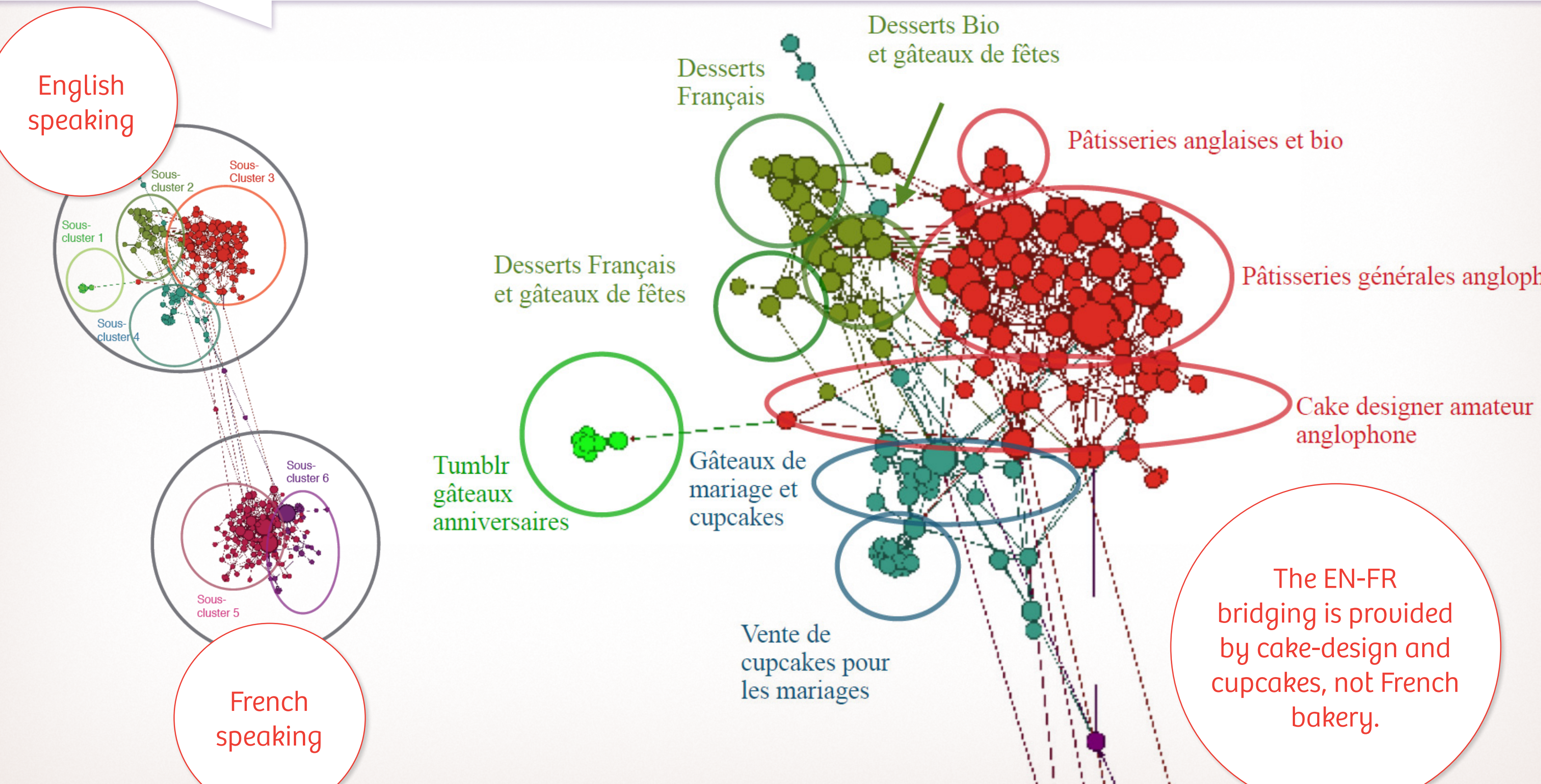
Course in Université Paris Descartes
"Méthodes d'enquête complémentaires"



Examples from Hyphe teachings

Bakery (pâtisserie)

Course in Université Paris Descartes
"Méthodes d'enquête complémentaires"



Examples from Hyphe teachings

Soccer (football)

Course in Université Paris Descartes
"Méthodes d'enquête complémentaires"

Research
question: are
football fans and
football games
fans the same
community?

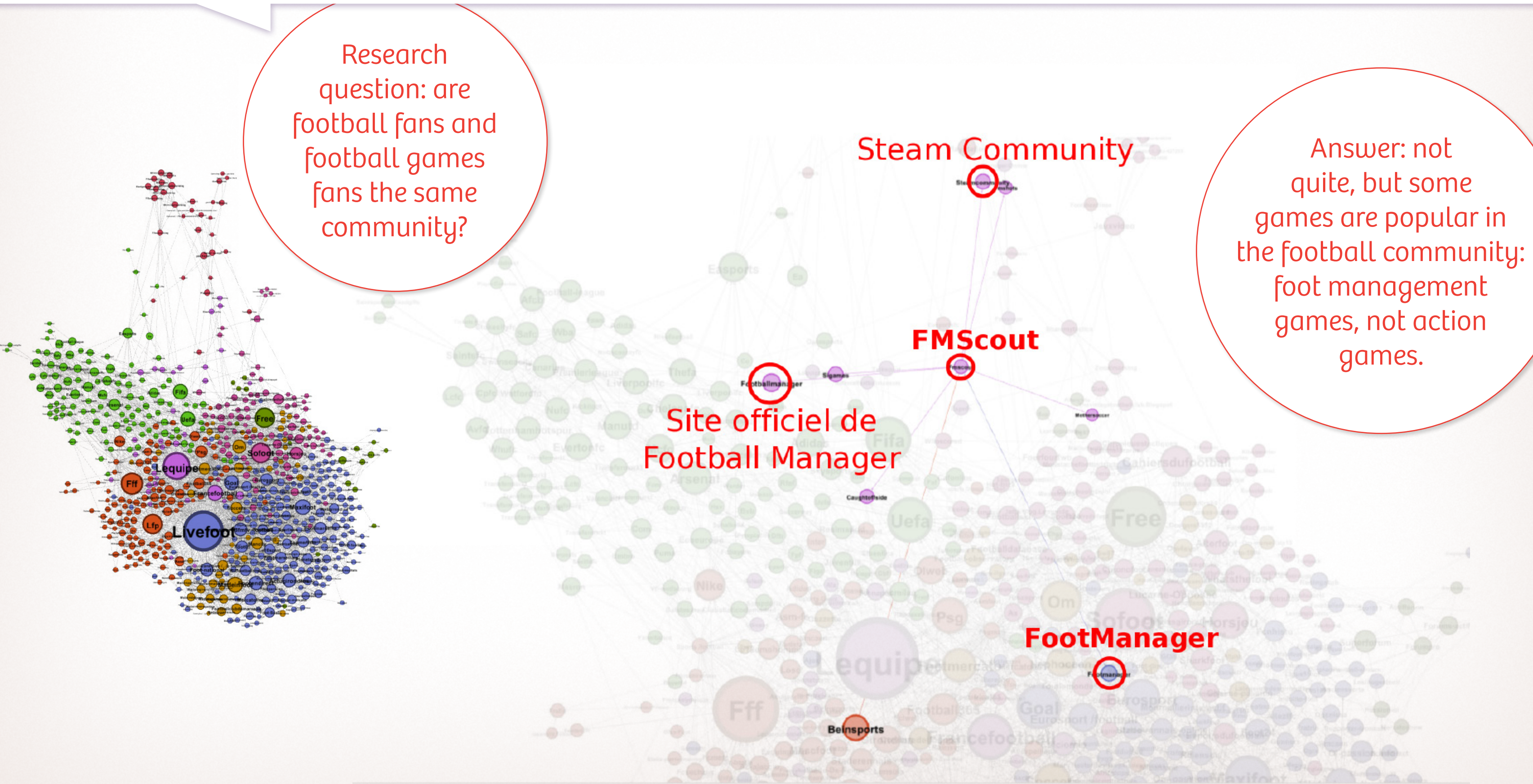
Steam Community

Answer: not
quite, but some
games are popular in
the football community:
foot management
games, not action
games.

FM Scout

Site officiel de
Football Manager

FootManager



Finding informations on Hyphe

Official website

<http://hyphe.medialab.sciences-po.fr/>

Demo

<http://hyphe.medialab.sciences-po.fr/demo/>

Source code and install

<https://github.com/medialab/hyphe>

Bug reporting

<https://github.com/medialab/hyphe/issues>

Paper

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13051/12797>

ICWSM poster

<http://www.medialab.sciences-po.fr/wp-content/uploads/2016/05/Hyphe-ICWSM-A3.pdf>

Thanks for your attention

@jacomy
Mathieu.Jacomy@sciencespo.fr

SciencesPo
MÉDIALAB

<http://medialab.sciences-po.fr>