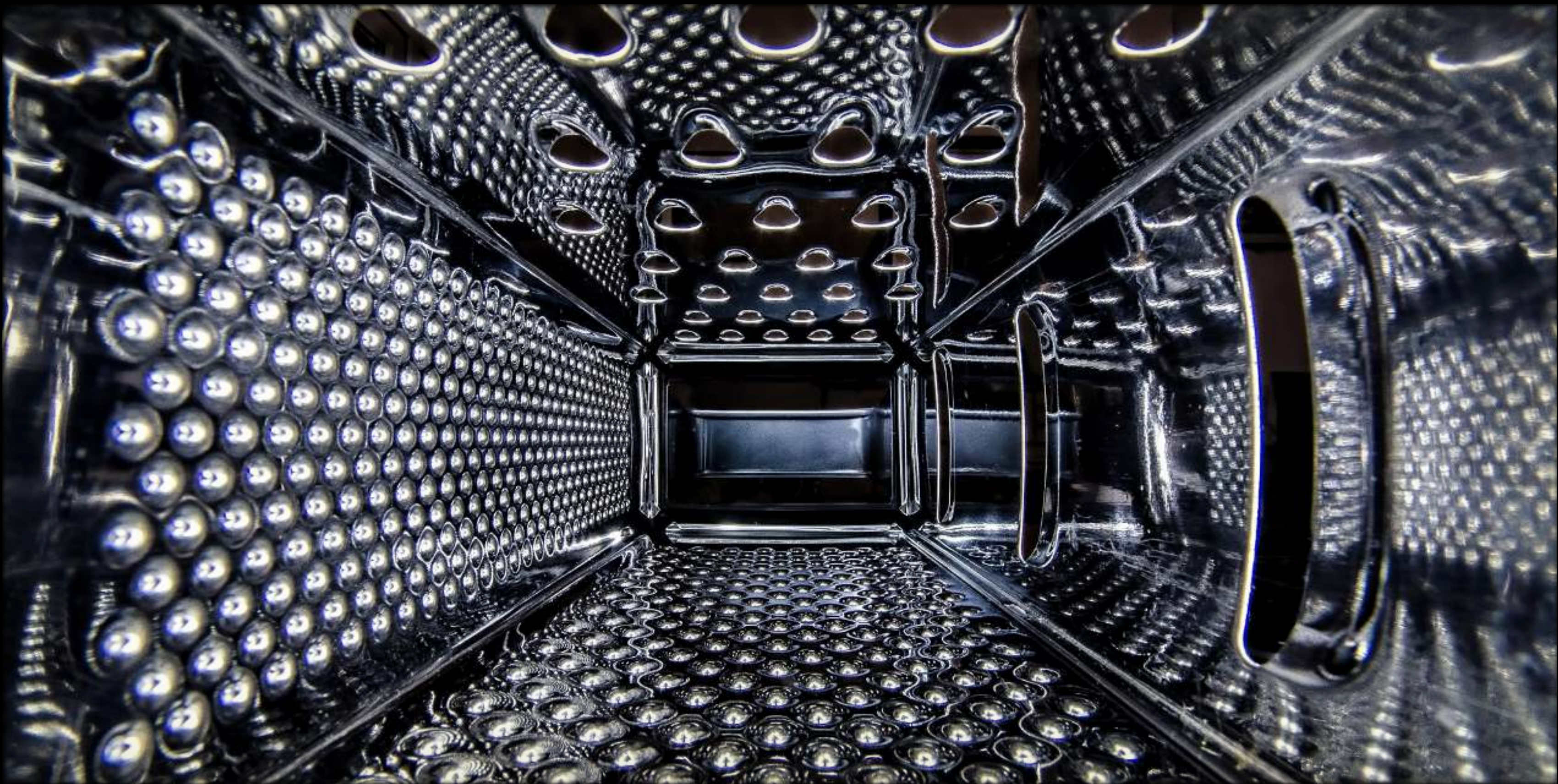


Explorer les internets avec le crawler Hyphe

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



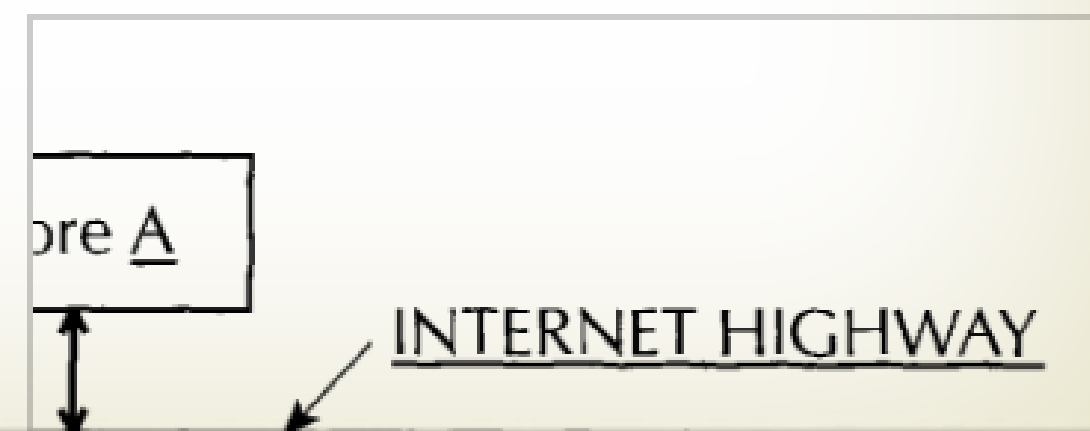
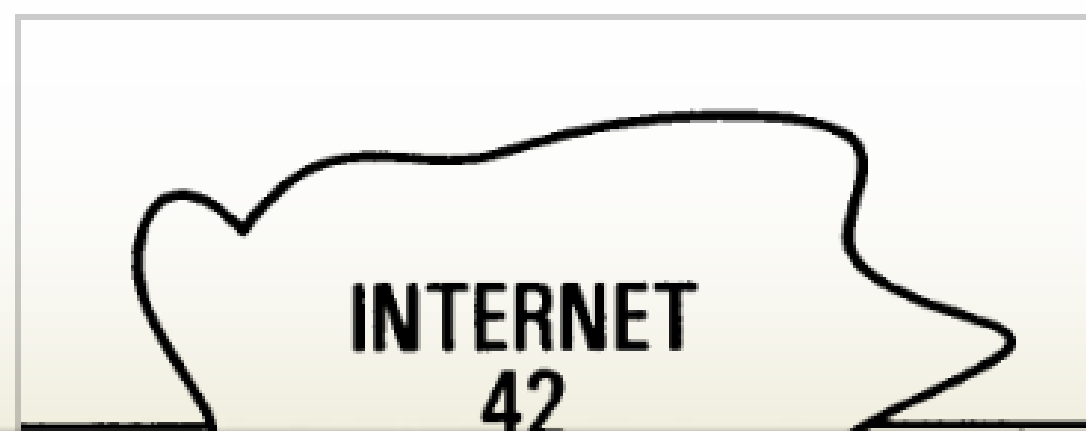
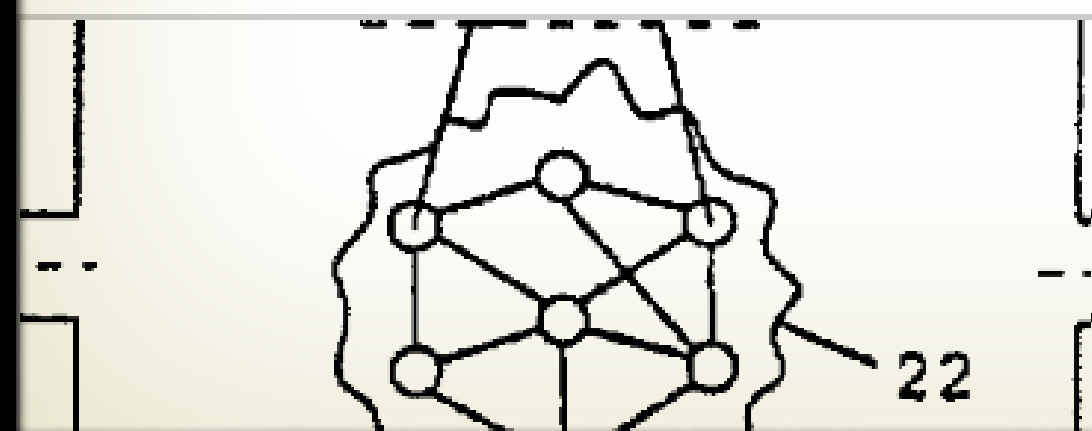
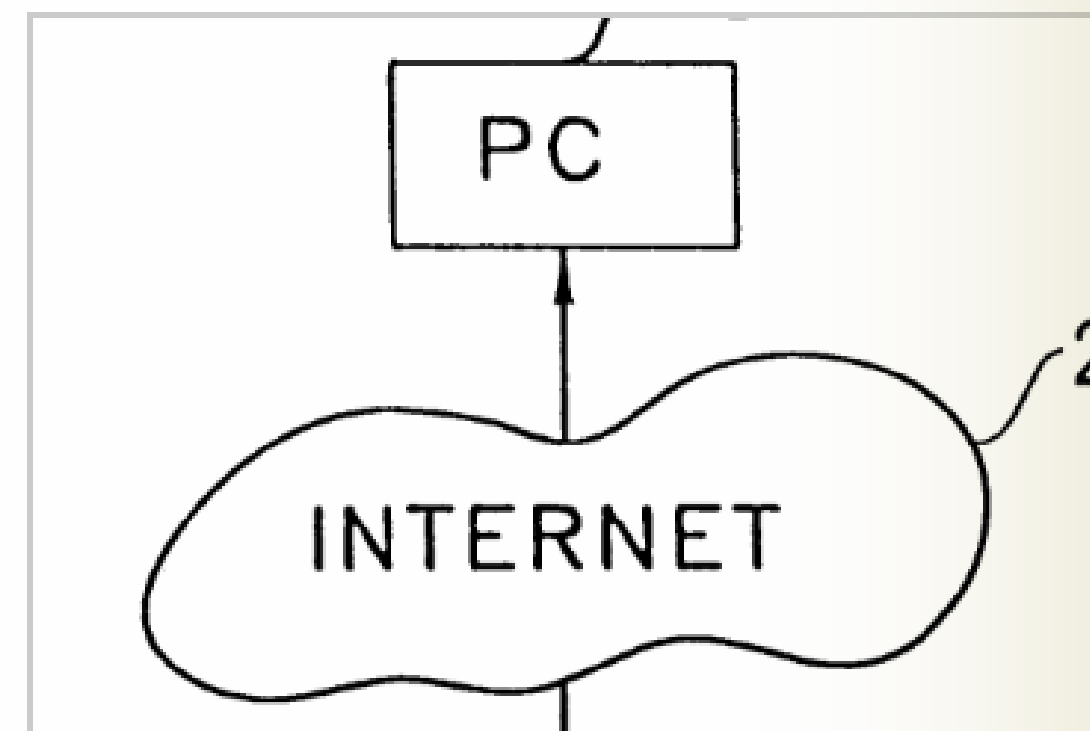
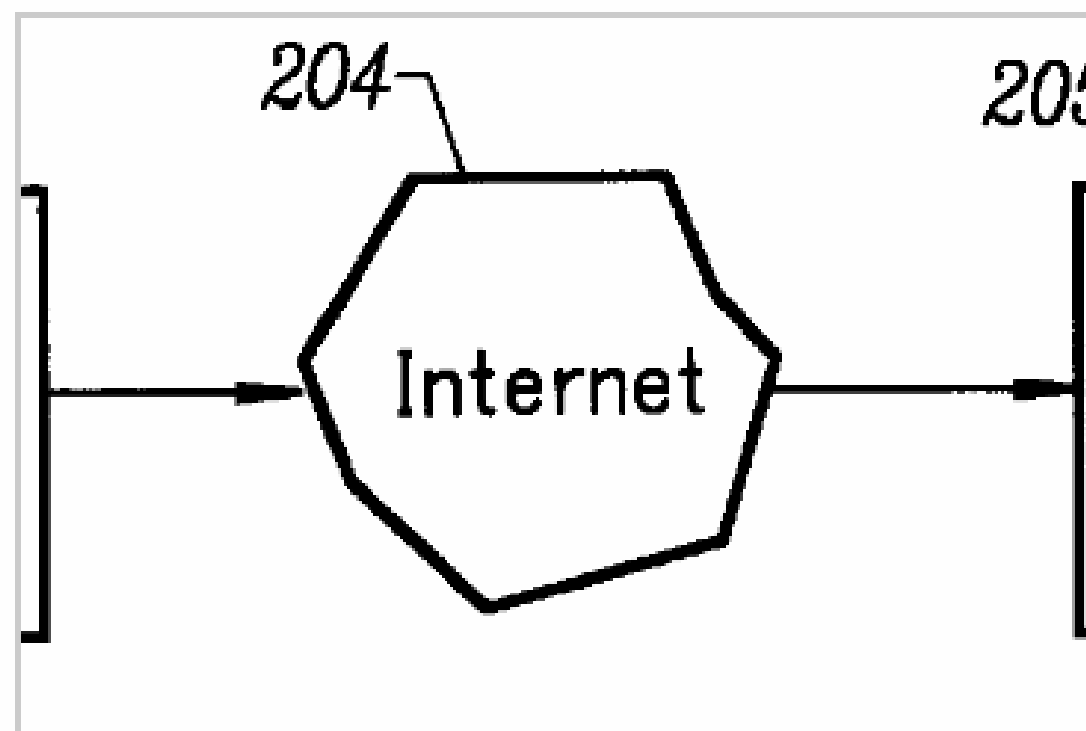
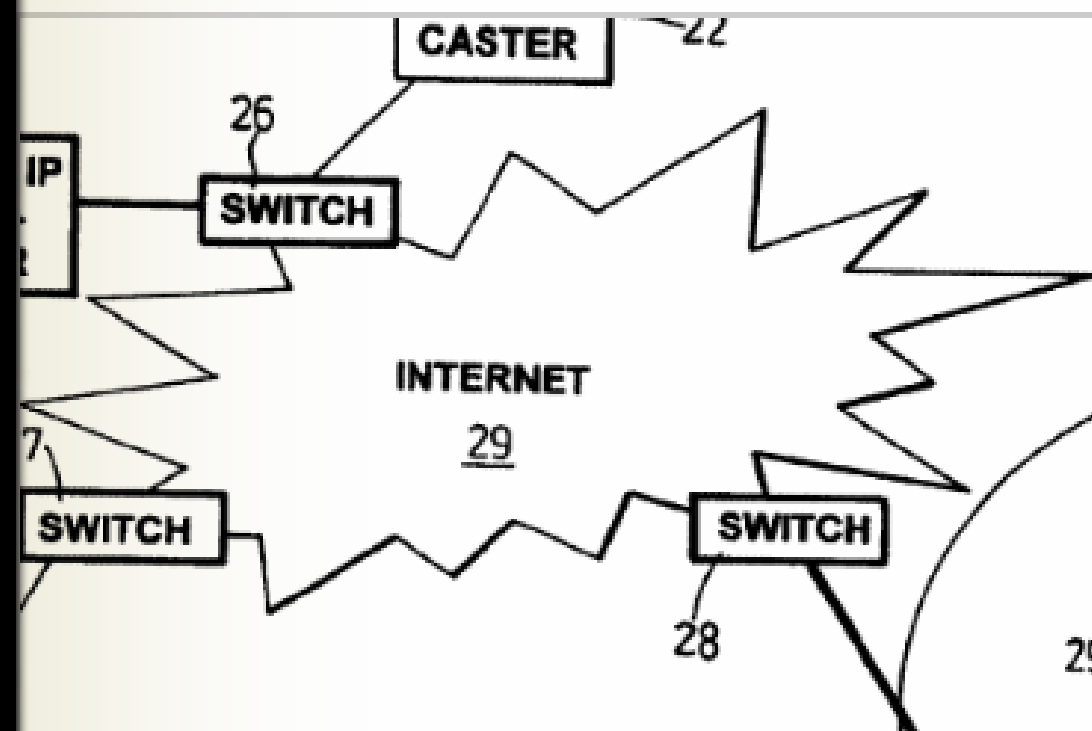
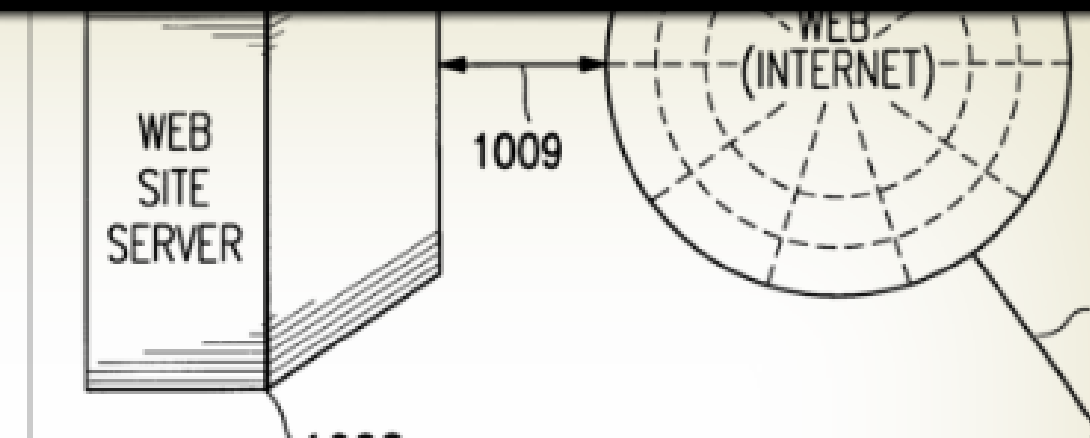
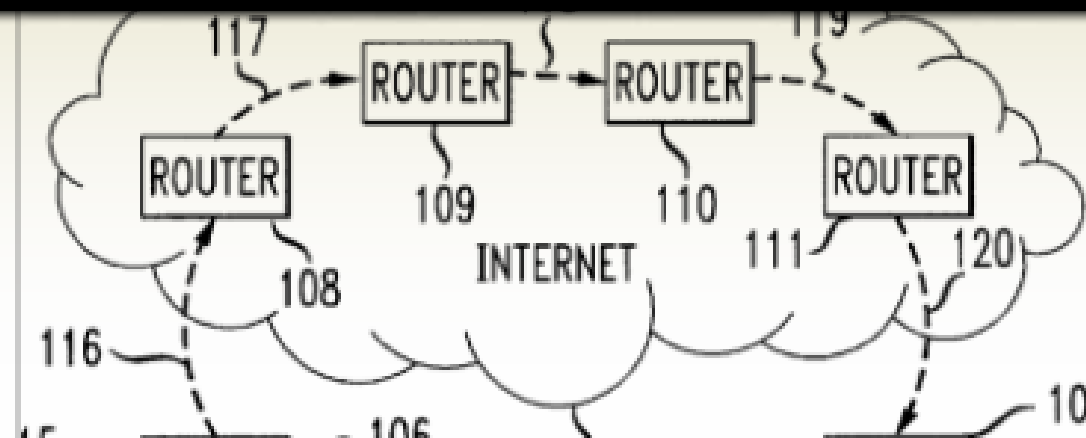
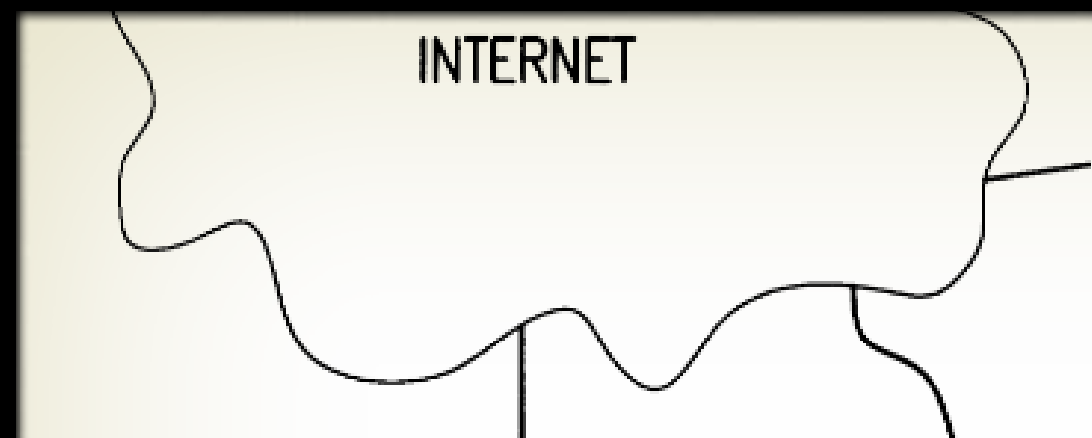
Le médialab



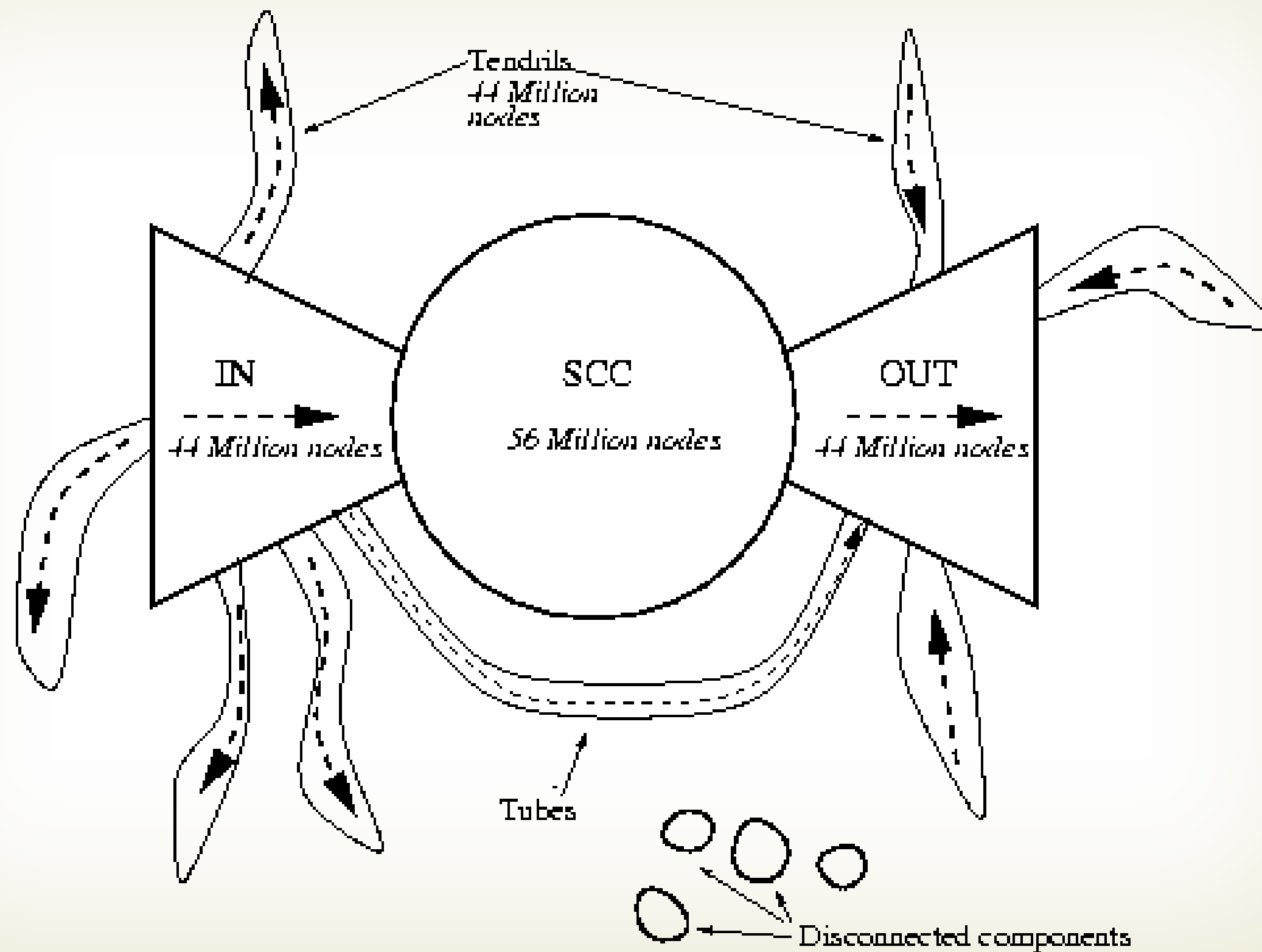
I.

Visualiser le web comme un réseau

Les internets dans les publications académiques



Les internets dans les publications académiques

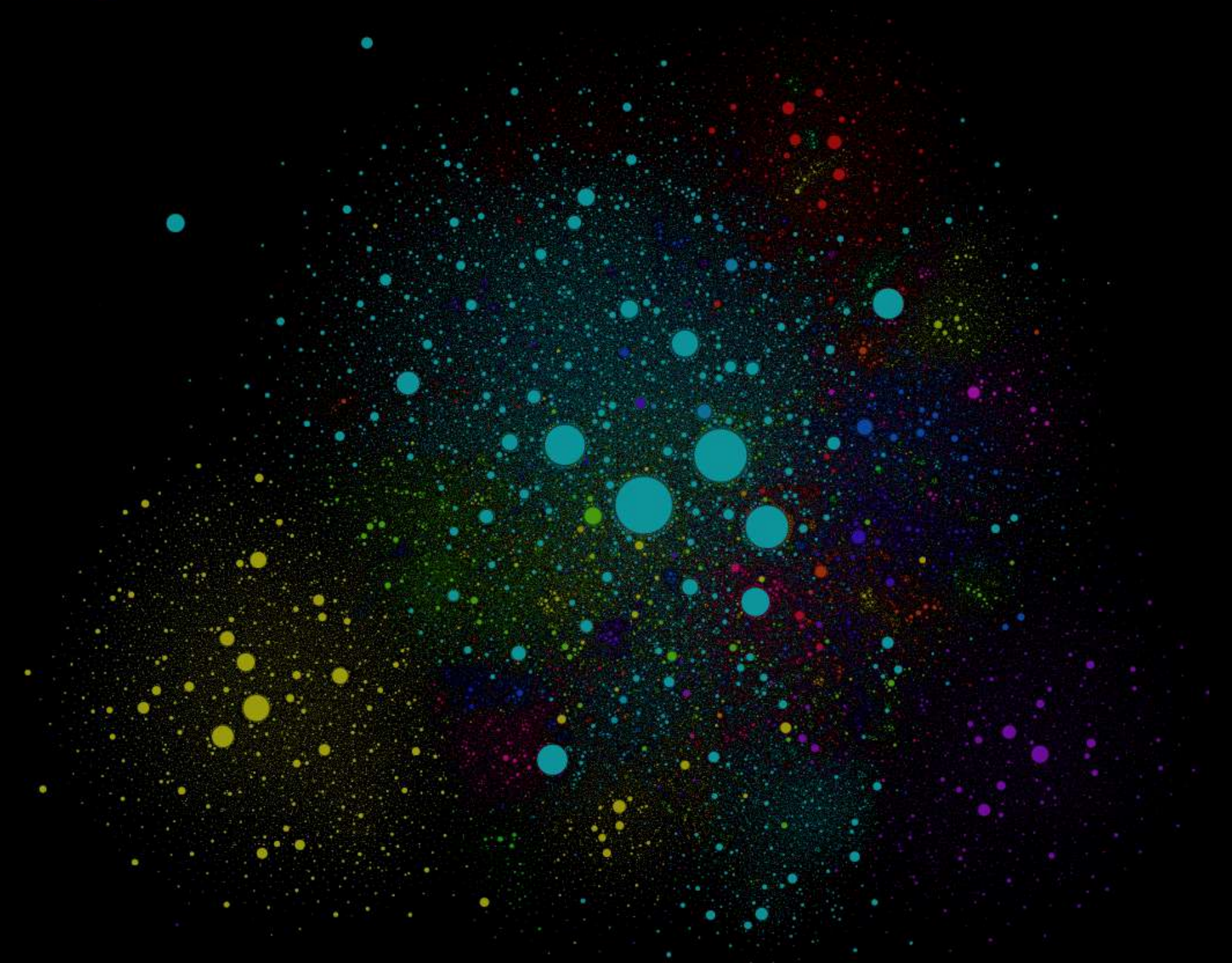


The bow tie, IBM's Almaden Research, 2000

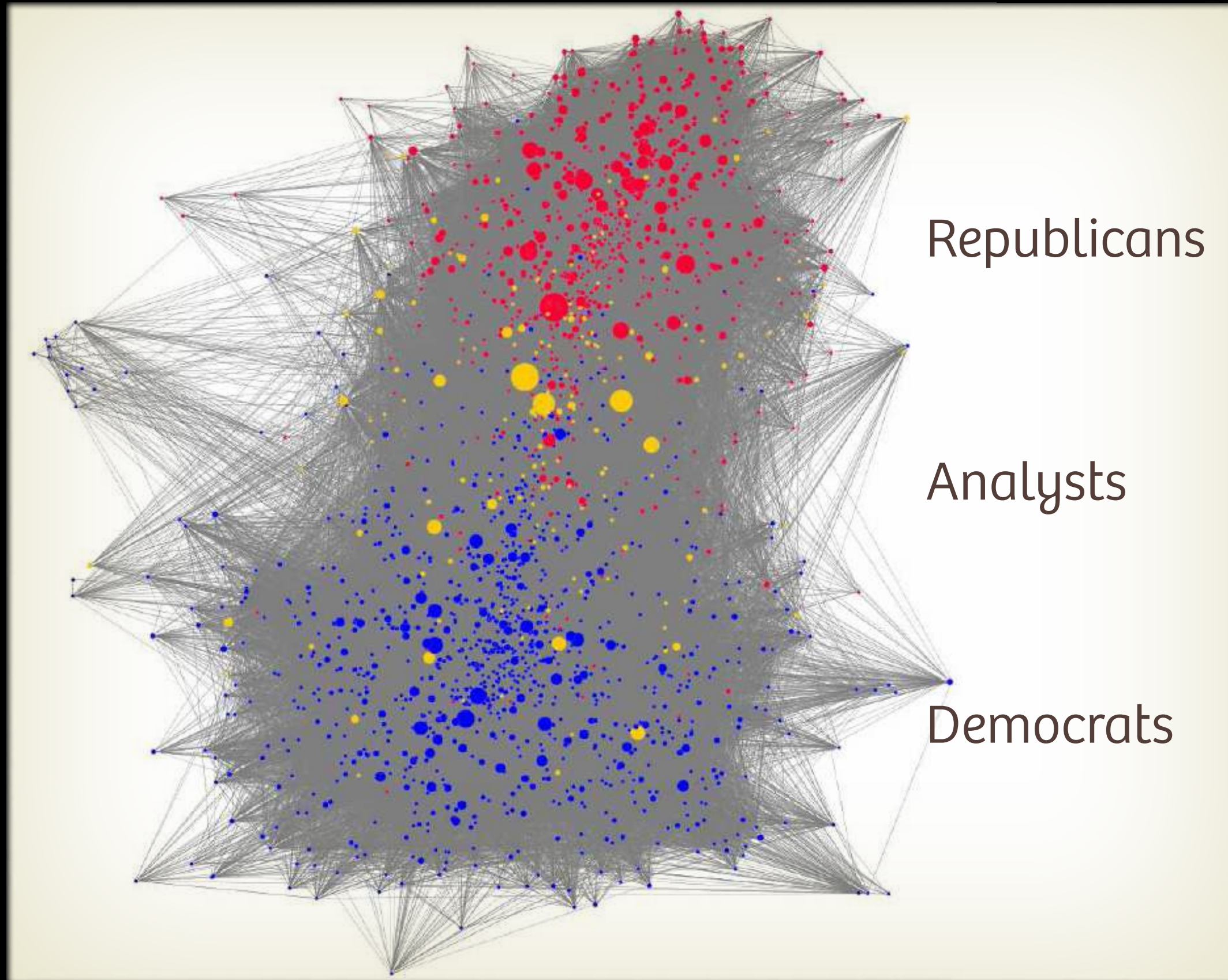
Les 10000 sites les plus visités

The Internet map

[About](#) [Blog](#) 

Un exemple où les sous-corpus sont identifiés

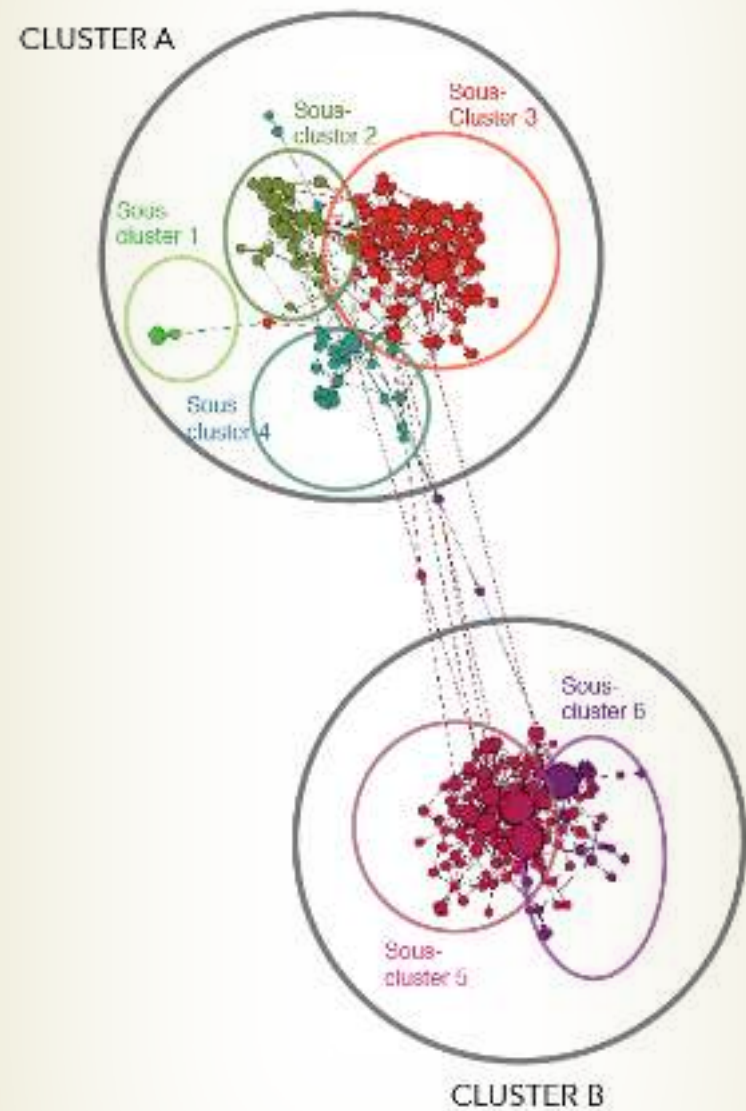


US blogosphere Linkfluence (2006)

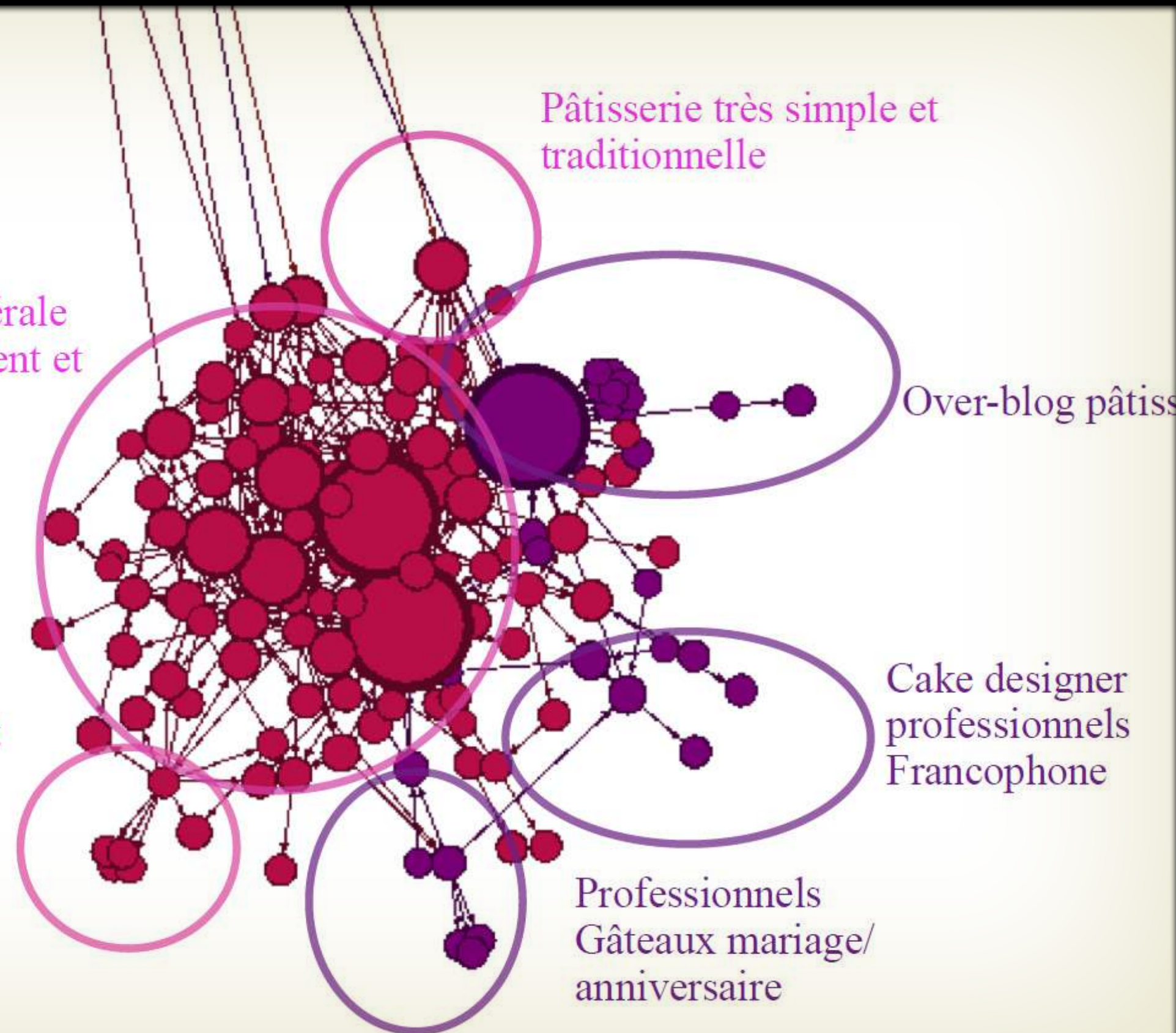
Exemple tiré d'un cours sur Hyphe

La Pâtisserie

Anglophone



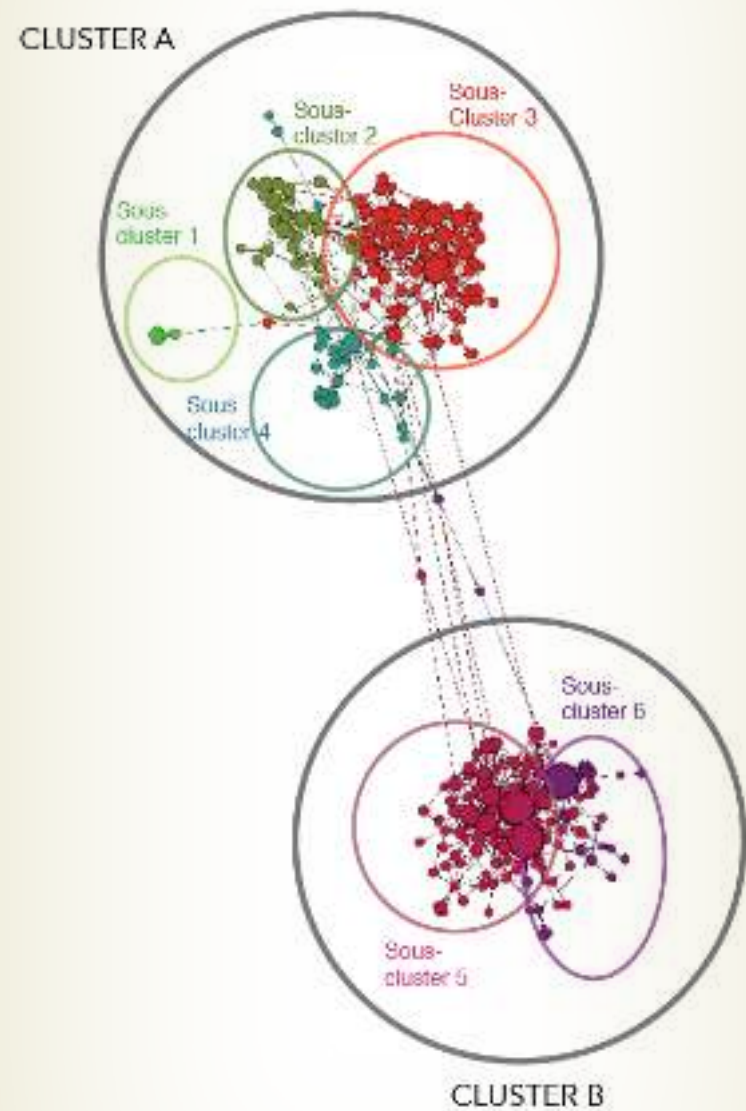
Francophone



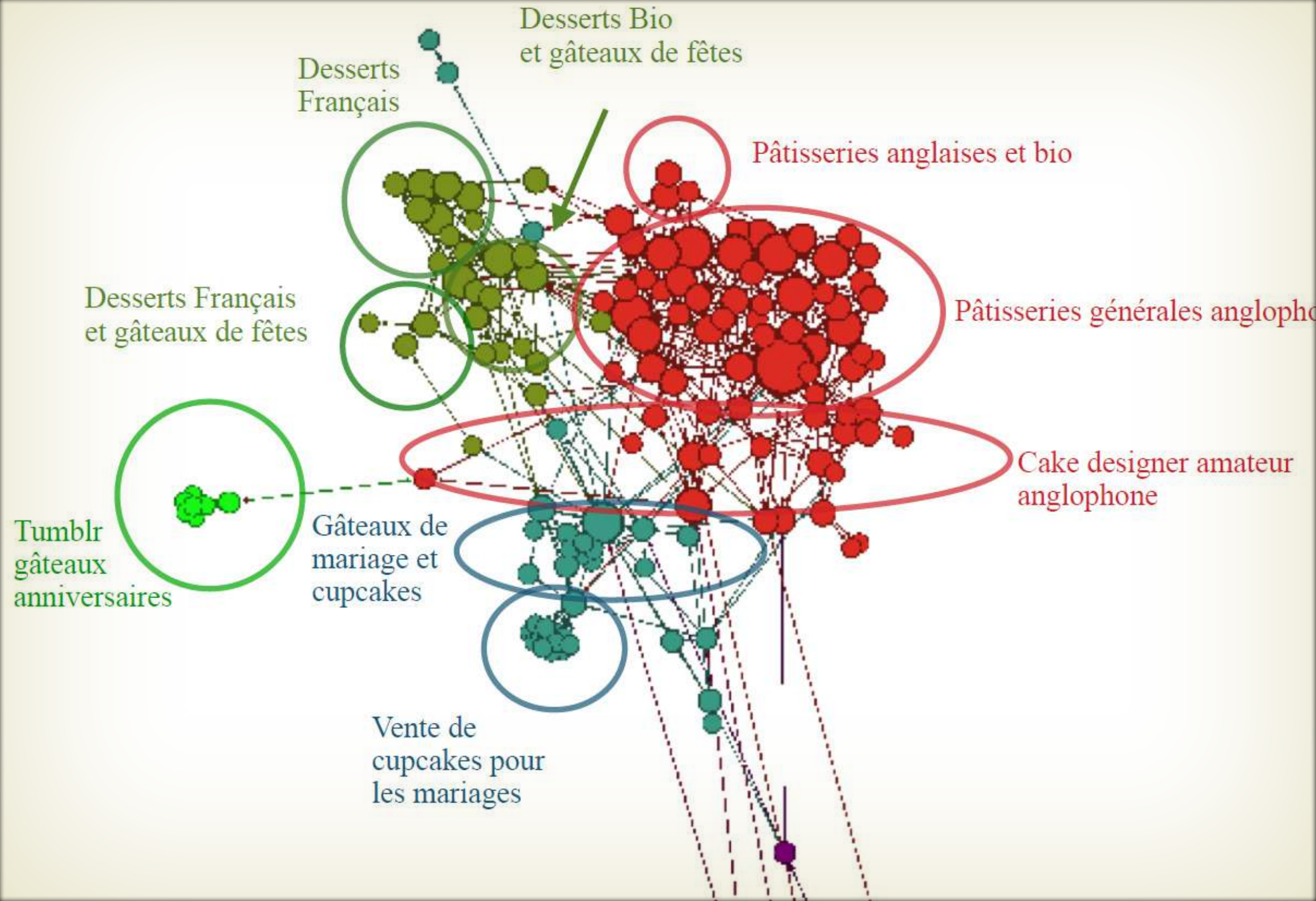
Exemple tiré d'un cours sur Hyphe

La Pâtisserie

Anglophone



Francophone



II.

Les sites web
ça n'existe pas

Une chaîne méthodologique axée sur l'exploration

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

Une chaîne méthodologique axée sur l'exploration

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

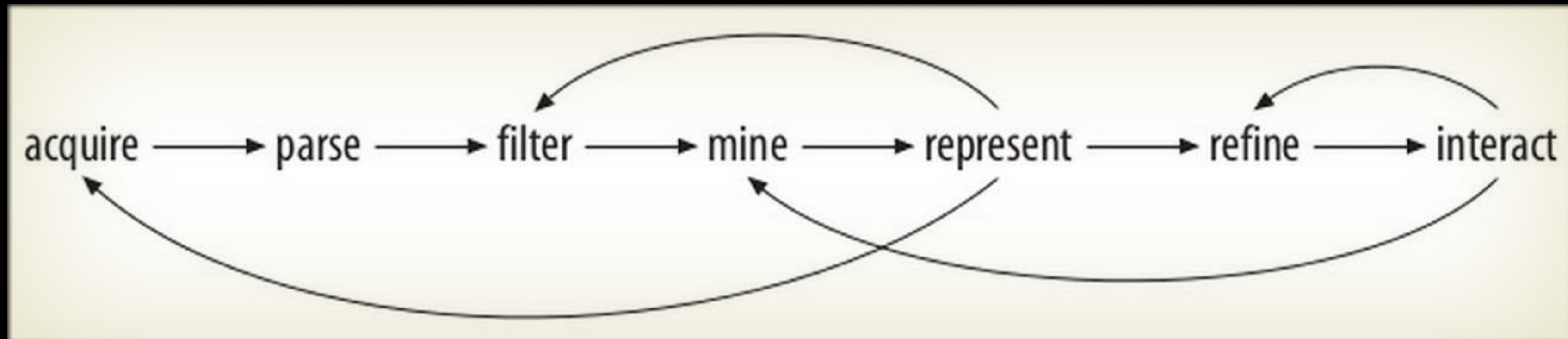
5. Finalization

Validate general quality
and export corpus



iterative
curation

Une chaîne méthodologique axée sur l'exploration



Définir des entités web

<https://en.wikipedia.org/wiki/Snail>

<https://fr.wikipedia.org/wiki/Escargot>

<https://en.wikipedia.org/wiki/Slug>

<https://en.wikipedia.org/wiki/Slug#Behavior>

<https://en.wikipedia.org/wiki/Lettuce>

<https://en.m.wikipedia.org/w/index.php?title=Lettuce>

Définir des entités web

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php ?title=Lettuce

Définir des entités web

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	/w	/index.php	?title=Lettuce

TLD

Définir des entités web

https://	en.	wikipedia	.org	/wiki	/Snail		
https://	fr.	wikipedia	.org	/wiki	/Escargot		
https://	en.	wikipedia	.org	/wiki	/Slug		
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior	
https://	en.	wikipedia	.org	/wiki	/Lettuce		
https://	en.	m.	wikipedia	<u>.org</u>	/w	/index.php	?title=Lettuce

Domain

TLD

Définir des entités web

https://		en.	wikipedia	.org	/wiki	/Snail	
https://		fr.	wikipedia	.org	/wiki	/Escargot	
https://		en.	wikipedia	.org	/wiki	/Slug	
https://		en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://		en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Définir des entités web

https://	en.	wikipedia	.org	/wiki	/Snail		
https://	fr.	wikipedia	.org	/wiki	/Escargot		
https://	en.	wikipedia	.org	/wiki	/Slug		
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior	
https://	en.	wikipedia	.org	/wiki	/Lettuce		
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Path

Définir des entités web

https://	en.	wikipedia	.org	/wiki	/Snail		
https://	fr.	wikipedia	.org	/wiki	/Escargot		
https://	en.	wikipedia	.org	/wiki	/Slug		
https://	en.	wikipedia	.org	/wiki	/Slug		
https://	en.	wikipedia	.org	/wiki	/Lettuce		
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Path

Query

#Behavior

Définir des entités web

https:// en. wikipedia .org /wiki /Snail

https:// fr. wikipedia .org /wiki /Escargot

https:// en. wikipedia .org /wiki /Slug

https:// en. wikipedia .org /wiki /Slug

https:// en. wikipedia .org /wiki /Lettuce

https:// en. m. wikipedia .org /w /index.php ?title=Lettuce

Subdomains

Domain

TLD

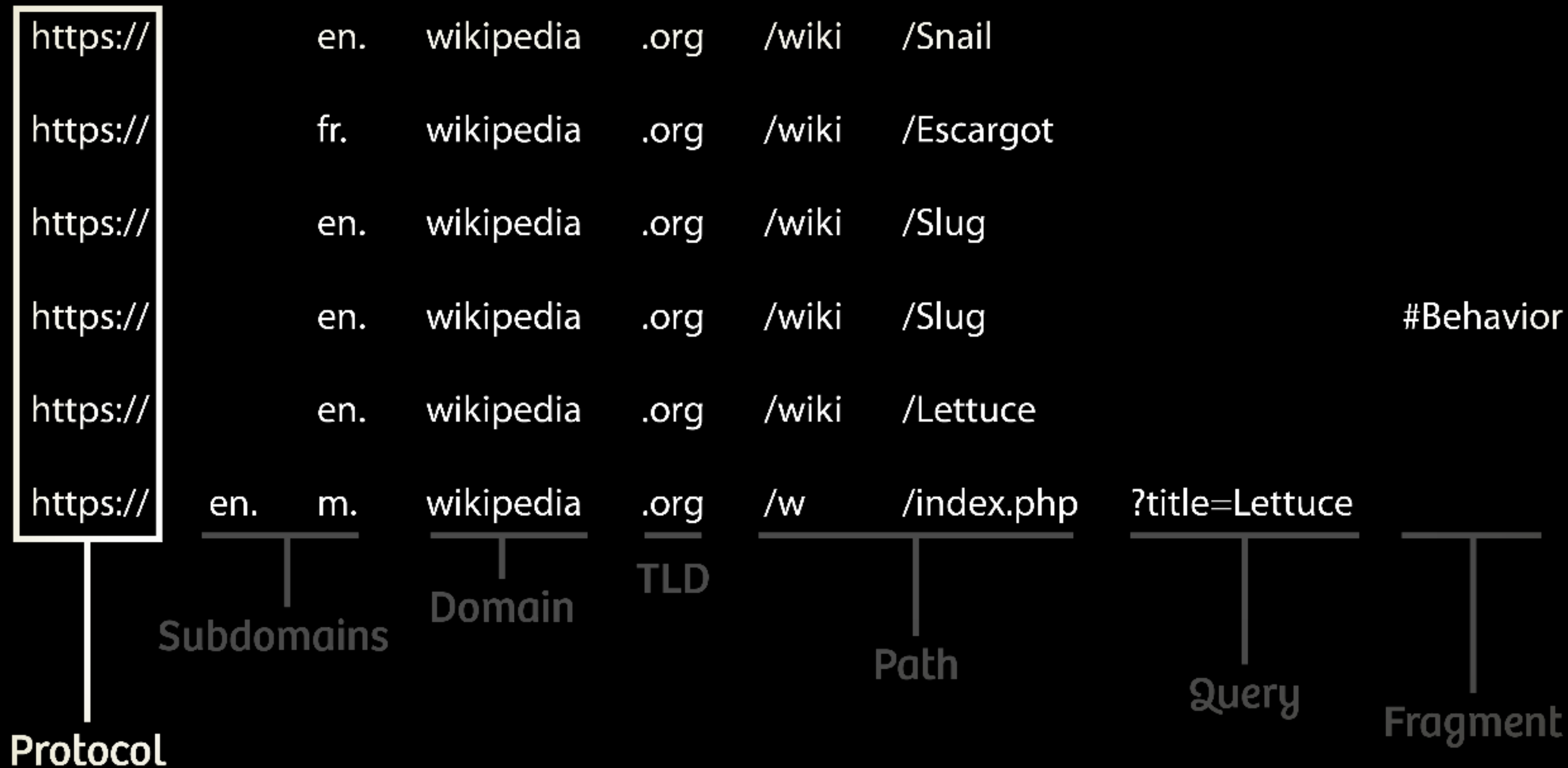
Path

Query

Fragment



Définir des entités web



Définir des entités web

.org	wikipedia	en.	/wiki	/Snail			https://	
.org	wikipedia	fr.	/wiki	/Escargot			https://	
.org	wikipedia	en.	/wiki	/Slug			https://	
.org	wikipedia	en.	/wiki	/Slug		#Behavior	https://	
.org	wikipedia	en.	/wiki	/Lettuce			https://	
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce	https://	
<u>TLD</u>	<u>Domain</u>	<u>Subdomains</u>		<u>Path</u>		<u>Query</u>	<u>Fragment</u>	<u>Protocol</u>

Définir des entités web

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **WIKIPEDIA**

Rule: must be prefixed by **".org | wikipedia"**

Définir des entités web

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **ENGLISH WIKIPEDIA**

Rule: must be prefixed by “.org | **wikipedia** | en.”

Définir des entités web

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **ENGLISH WIKIPEDIA** (IMPROVED)

Rule: must be prefixed by **".org | wikipedia | en."** OR **".org | wikipedia | m. | en."**

Définir des entités web

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR THE SLUG PAGE

Rule: must be prefixed by “.org | wikipedia | en. | /wiki | /Slug”

Définir des entités web

.org	wikipedia	en.	/wiki	/Snail		https://	
.org	wikipedia	fr.	/wiki	/Escargot		https://	
.org	wikipedia	en.	/wiki	/Slug		https://	
.org	wikipedia	en.	/wiki	/Slug	#Behavior	https://	
.org	wikipedia	en.	/wiki	/Lettuce		https://	
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce	https://

ALL THE RULES CAN **COEXIST**

Example: **WIKIPEDIA** + **SLUG PAGE**

The rules apply in order: every page is in **ONE and ONLY ONE** web entity

Définir des entités web

.org	wikipedia	en.	/wiki	/Snail		https://	
.org	wikipedia	fr.	/wiki	/Escargot		https://	
.org	wikipedia	en.	/wiki	/Slug		https://	
.org	wikipedia	en.	/wiki	/Slug	#Behavior	https://	
.org	wikipedia	en.	/wiki	/Lettuce		https://	
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce	https://

YOU CAN DEFINE **EACH WIKI PAGE** AS A DIFFERENT ENTITY
Hyphe can learn **automatic rules** that apply to future crawls

Trouver de l'information sur Hyphe

Site officiel

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

Démo

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

Code source et installation

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

Reporter des bugs

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

Publication

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

Poster ICWSM

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

III.

Faire pousser
un corpus web

Statut des entités web dans Hyphe



IN

Connu
et accepté
dans le corpus



OUT

Connu
et rejeté



UNDECIDED

Connu mais on
ne sait pas s'il
faut le garder
ou non



DISCOVERED

Déecté par
Hyphe mais
non connu de
l'utilisateur

Faire pousser un corpus web



UNCHARTED



This web entity
is the starting point

Faire pousser un corpus web



UNCHARTED



Crawling the starting point detects its neighbours, but we do not know if they are relevant.

Faire pousser un corpus web



UNCHARTED

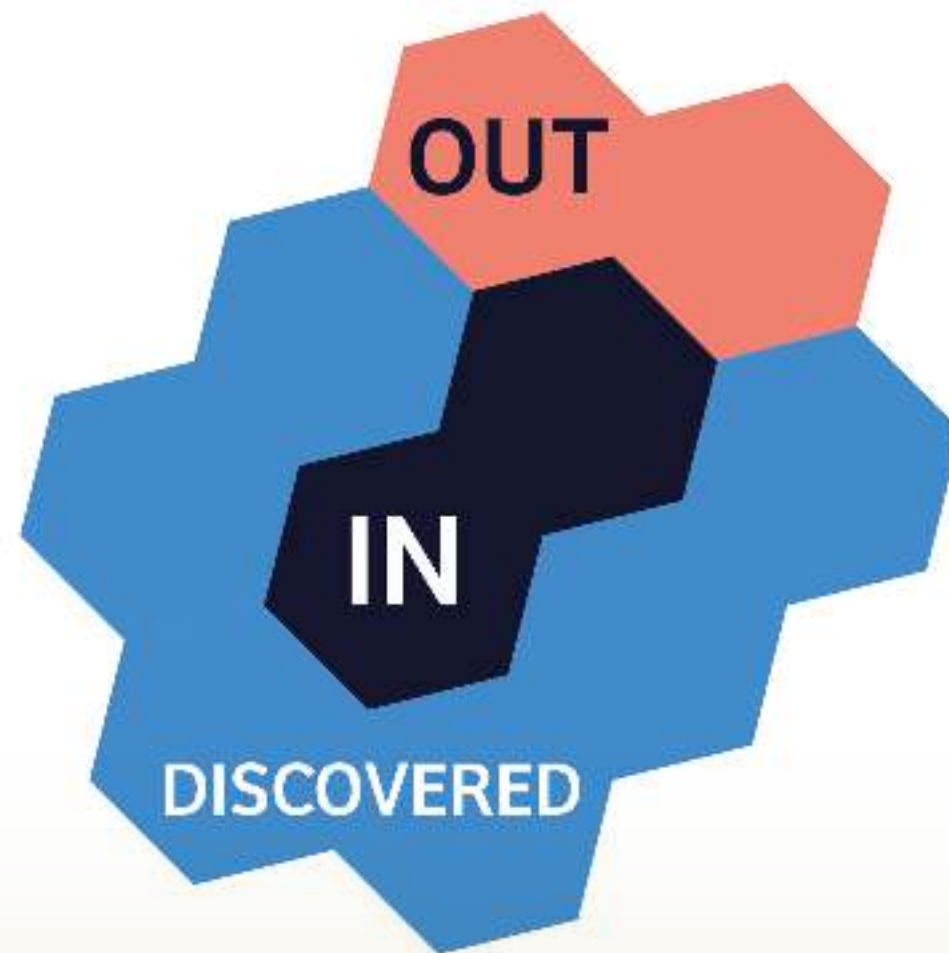


Adding and crawling more web entities expands the corpus.

Faire pousser un corpus web



UNCHARTED

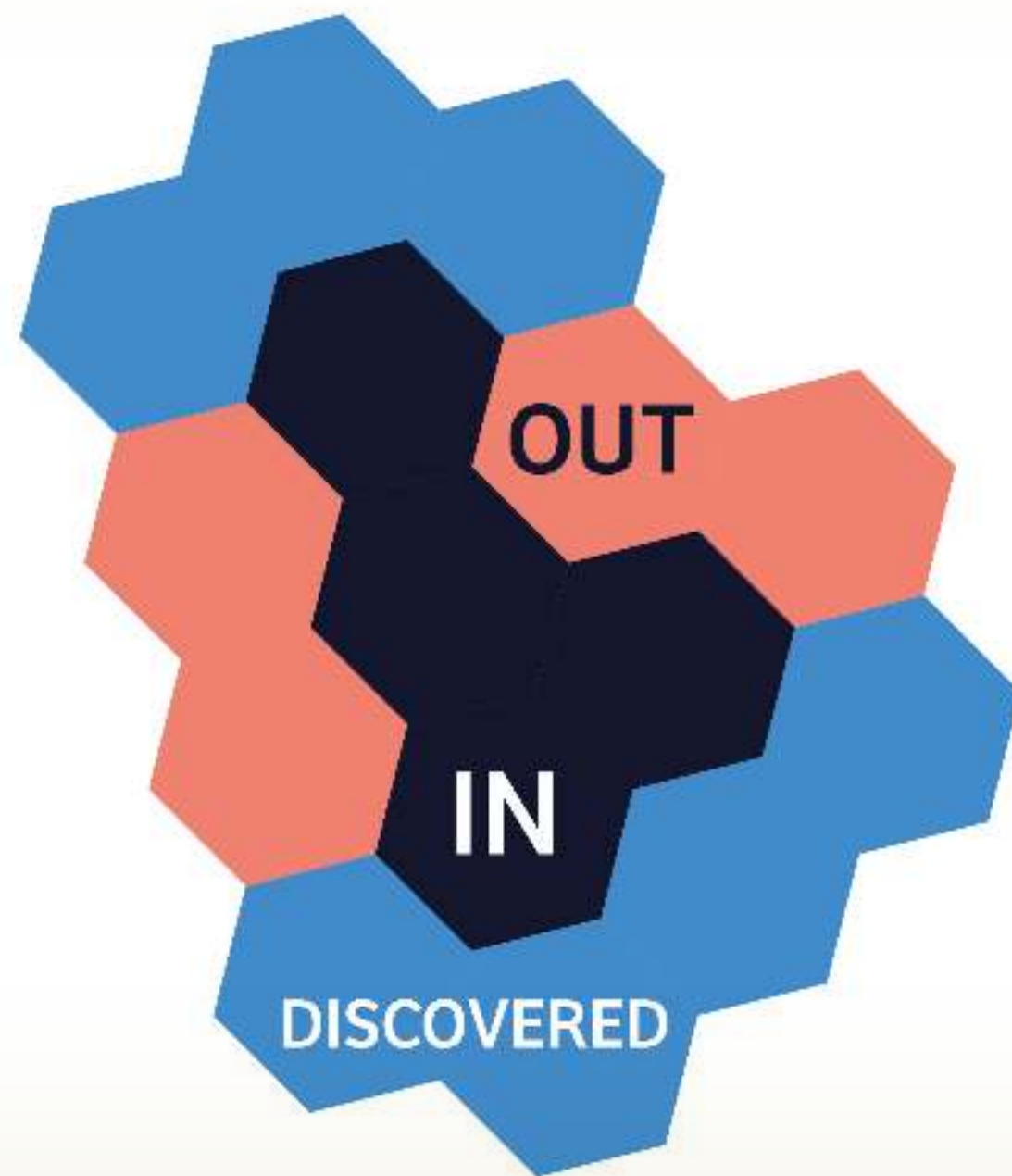


Some web entities are not relevant for the corpus and are rejected.

Faire pousser un corpus web



UNCHARTED



Faire pousser un corpus web



UNCHARTED



Faire pousser un corpus web



UNCHARTED



OUT

IN

DISCOVERED

Faire pousser un corpus web



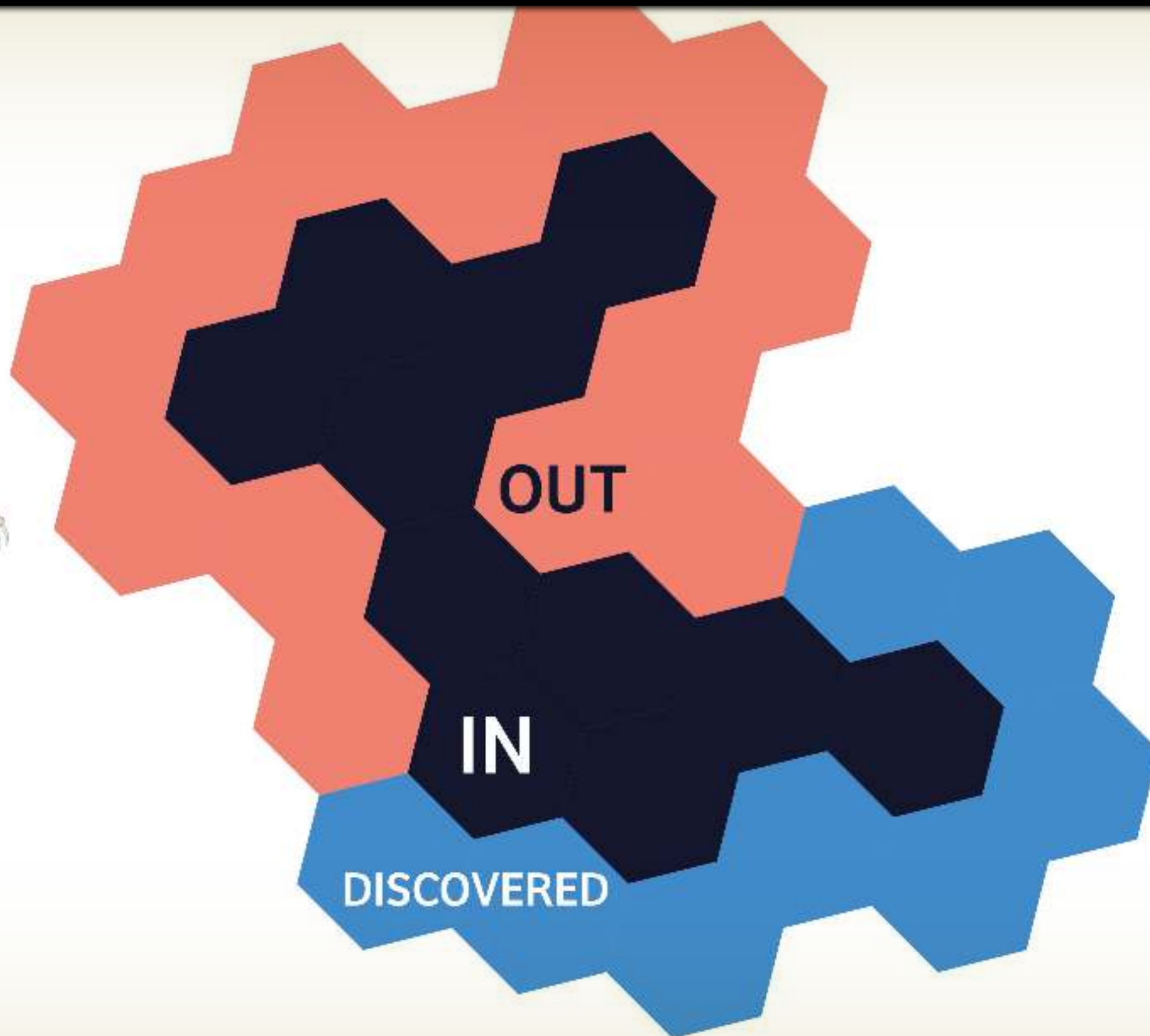
UNCHARTED



Faire pousser un corpus web



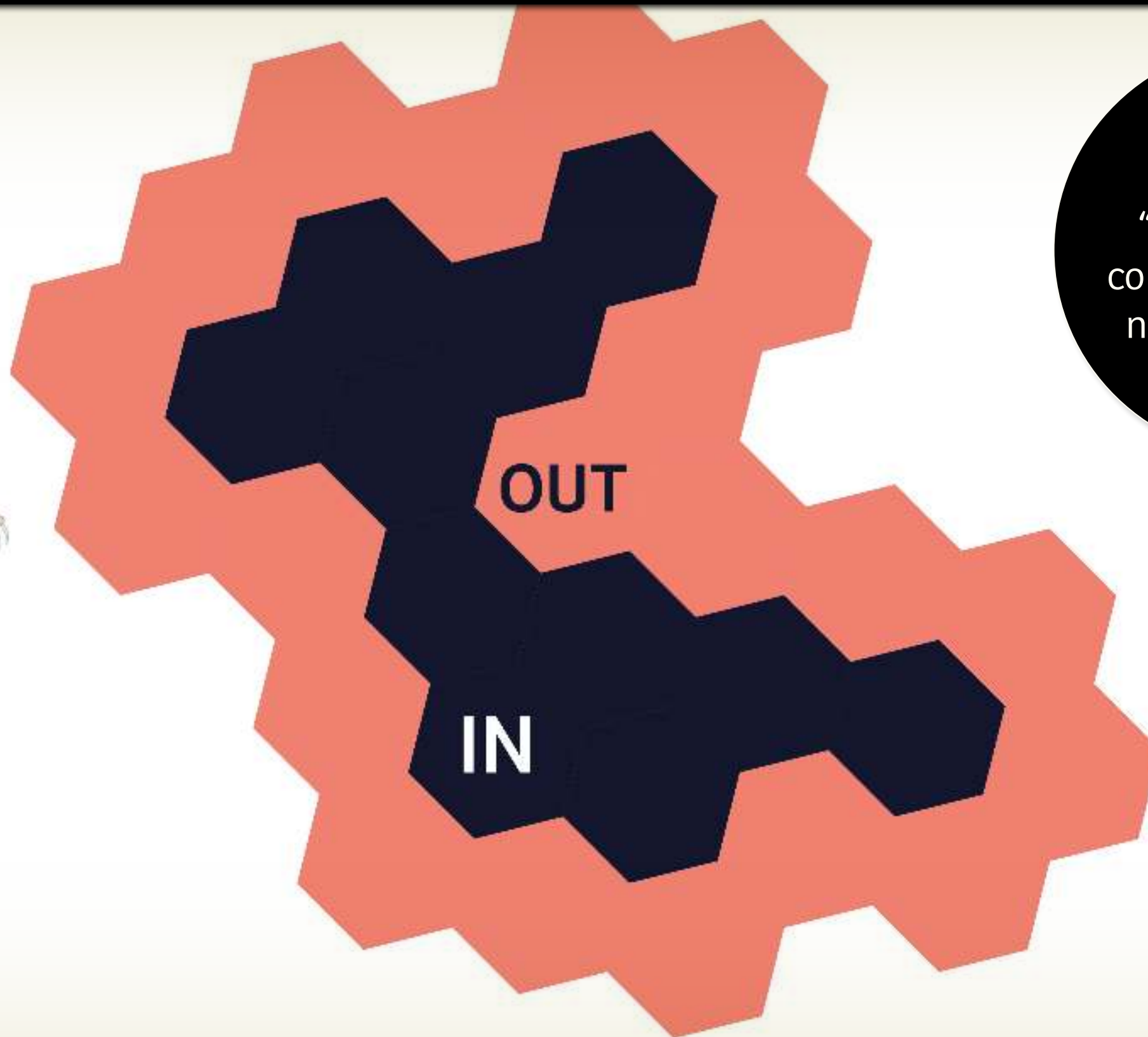
UNCHARTED



Faire pousser un corpus web



UNCHARTED



Au final le
"OUT" apparaît
comme la frontière
naturelle du "IN"

Faire pousser un corpus web

However, the web is not a flat space.
Growing a corpus actually looks a little different.

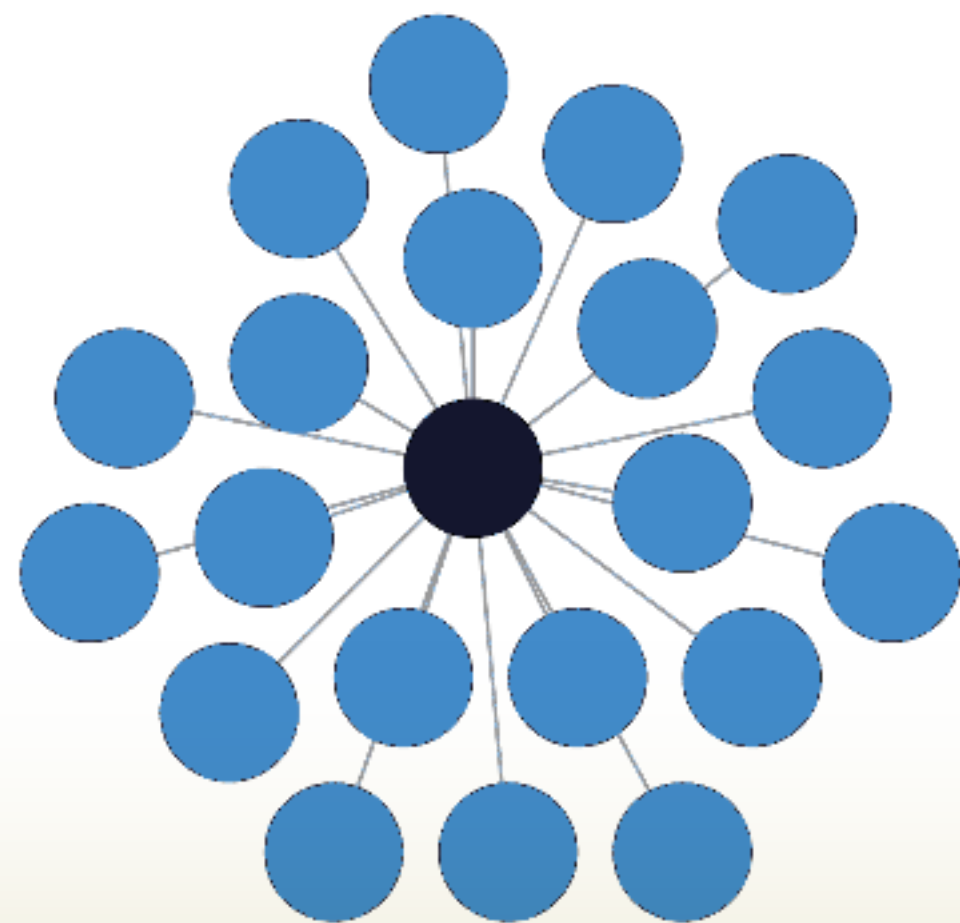


This web entity
is the starting point

Faire pousser un corpus web

Difference #1:

There are much more neighbors
per web entity in the corpus

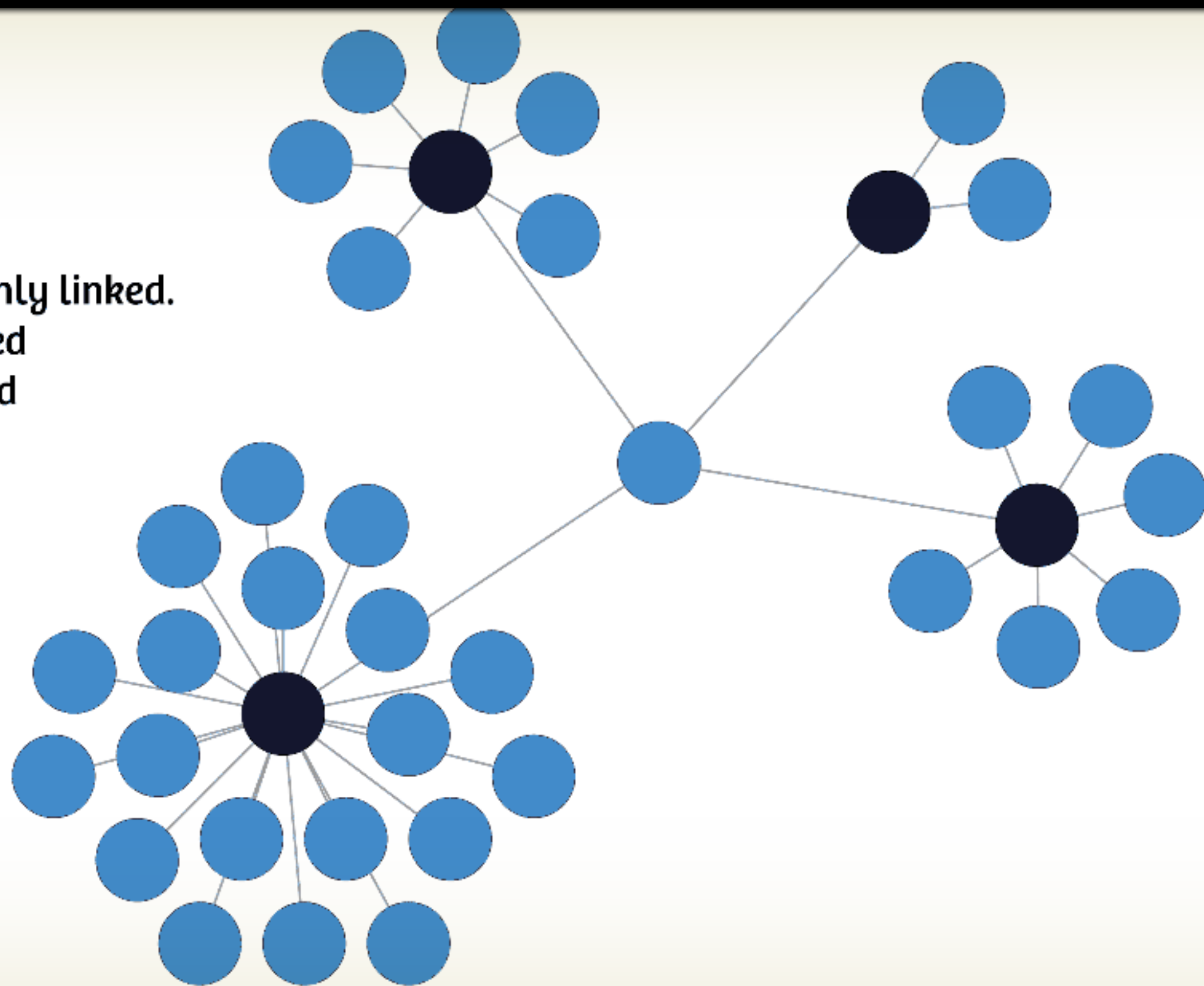


Faire pousser un corpus web

Difference #2:

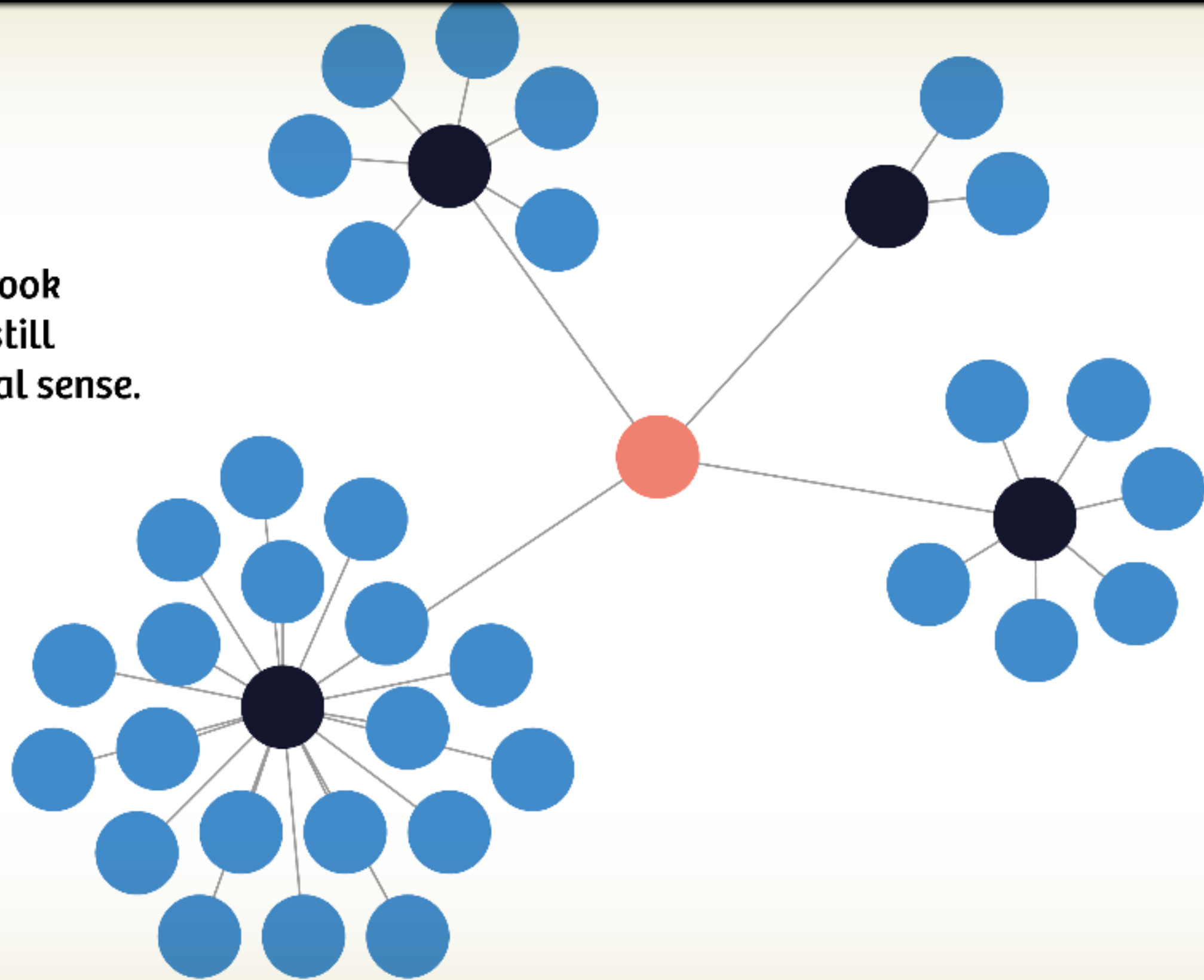
Neighbors are VERY unevenly linked.

- A few are highly connected
- Most are poorly connected

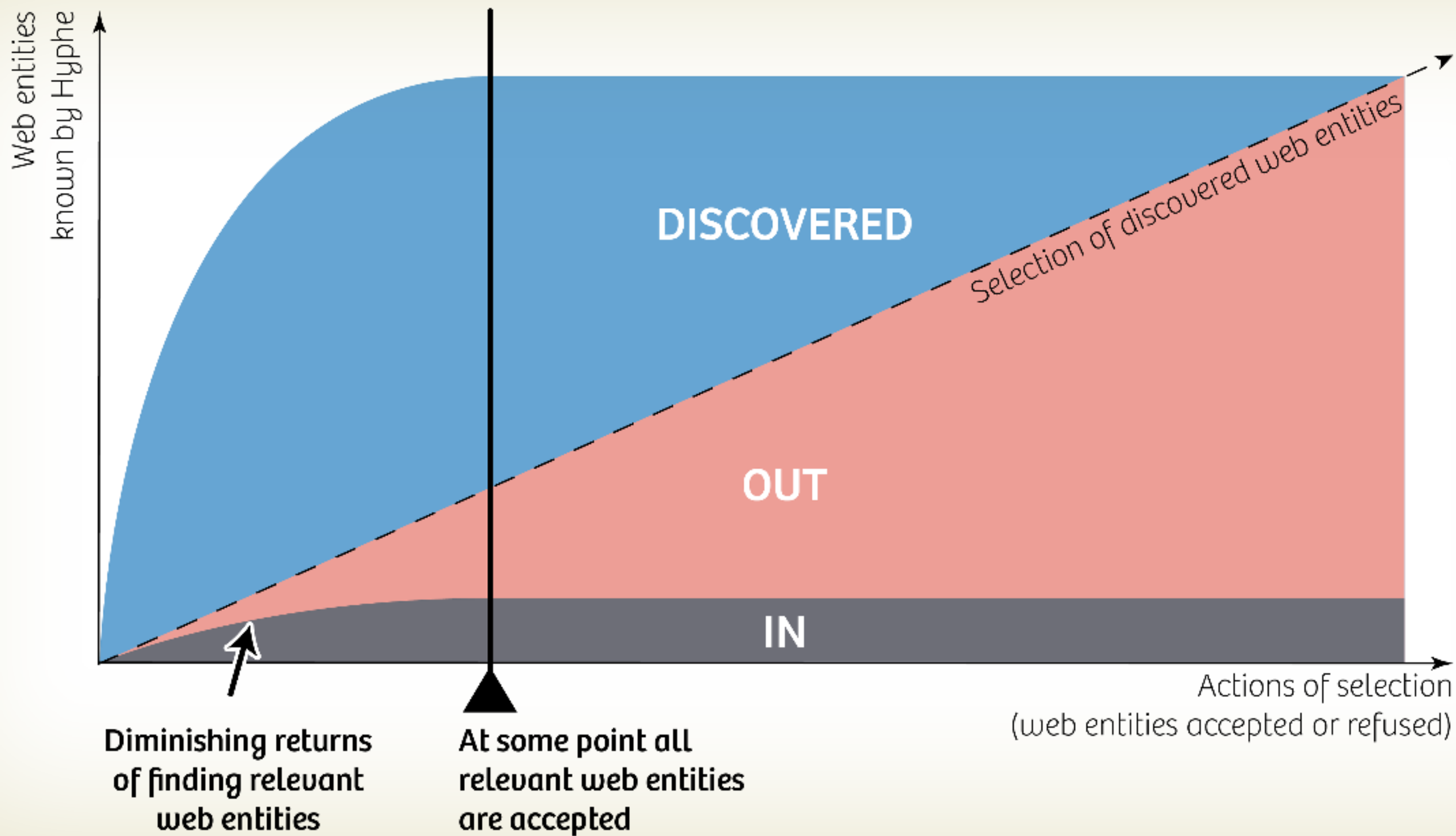


Faire pousser un corpus web

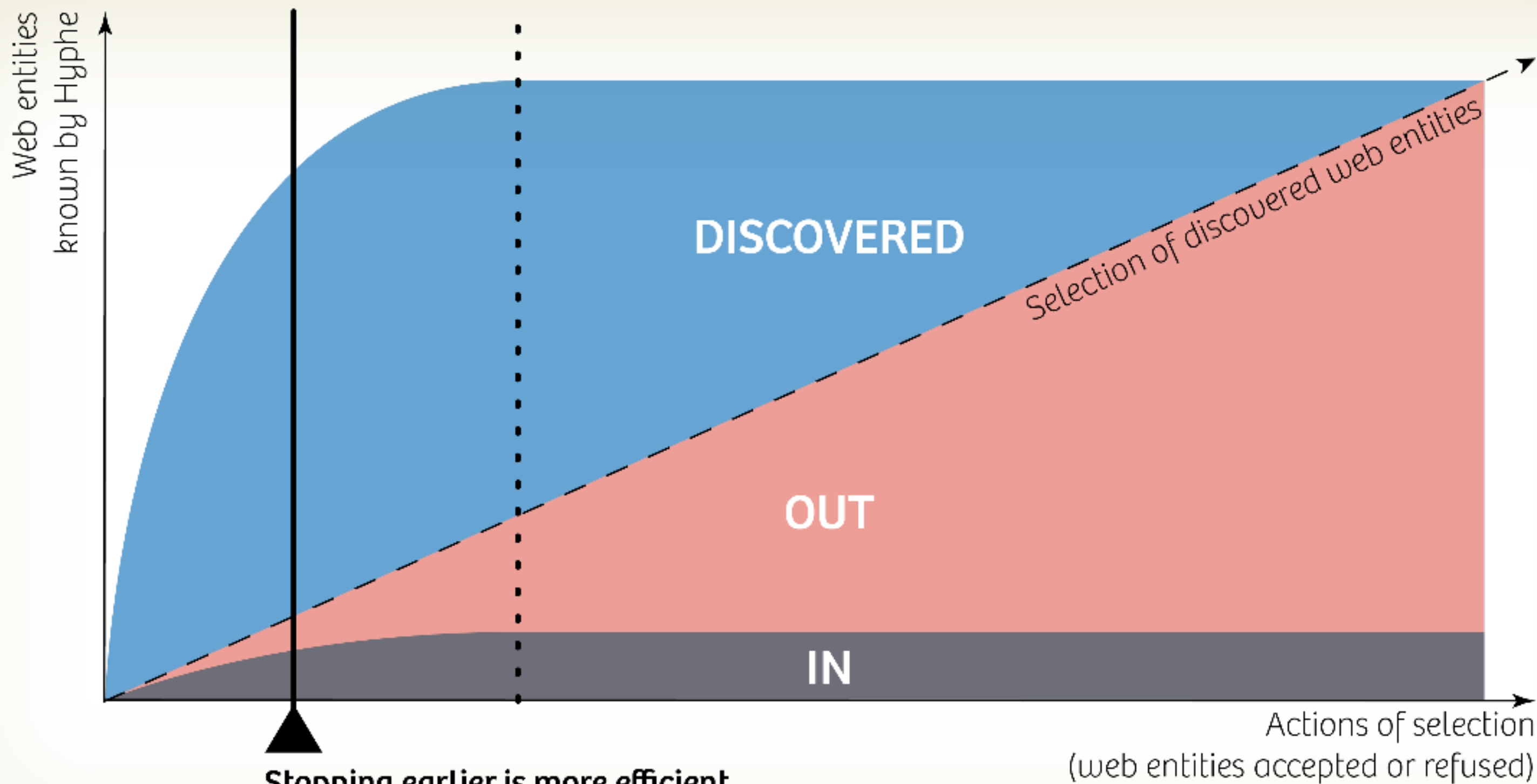
Difference #3:
Often the border does not look
like it is "around", but it is still
the border in the topological sense.



Faire pousser un corpus web



Faire pousser un corpus web

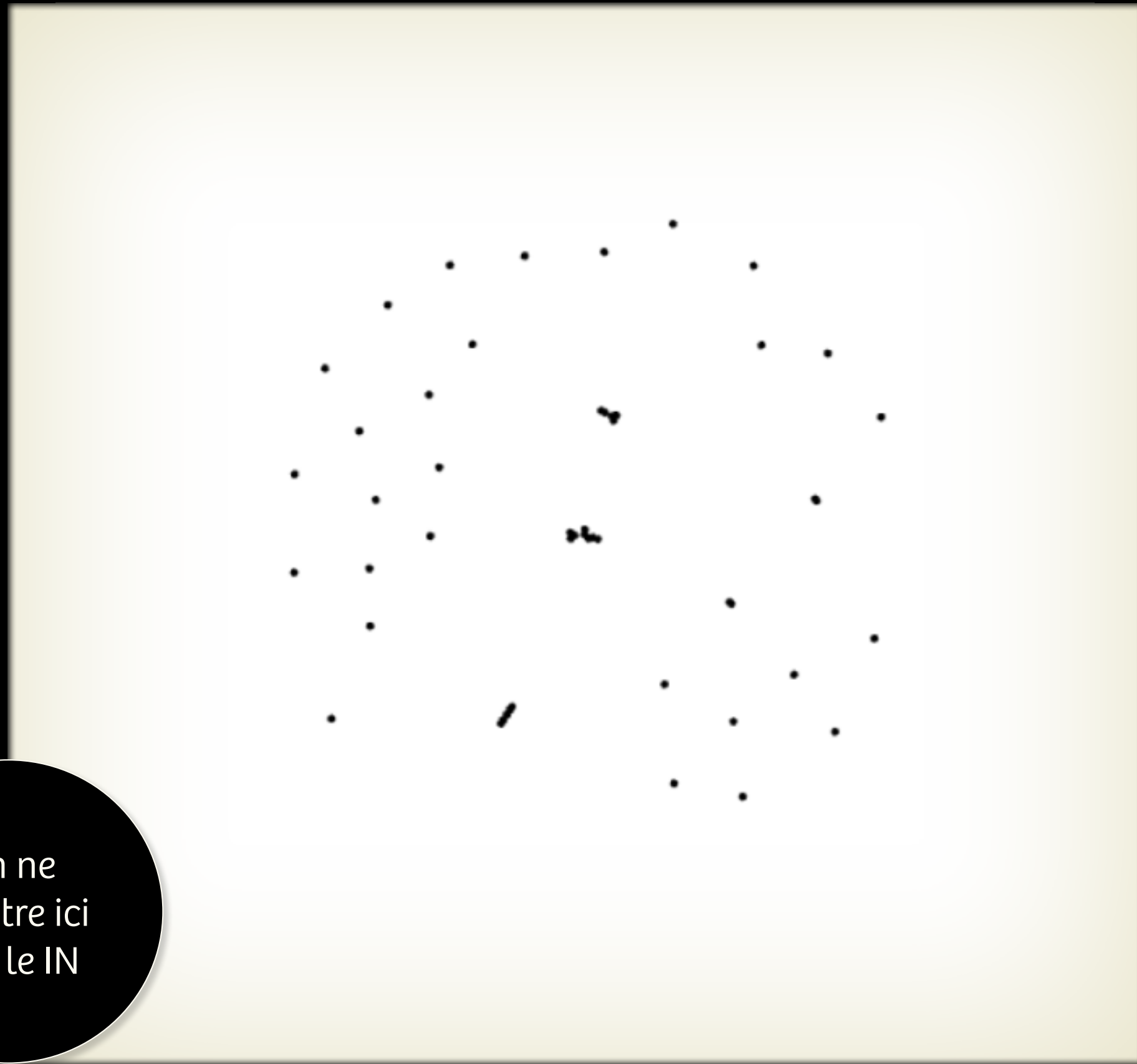


**Stopping earlier is more efficient,
when MOST relevant web entities
are accepted in the corpus**

Actions of selection
(web entities accepted or refused)

Un exemple empirique : l'avortement (abortion)

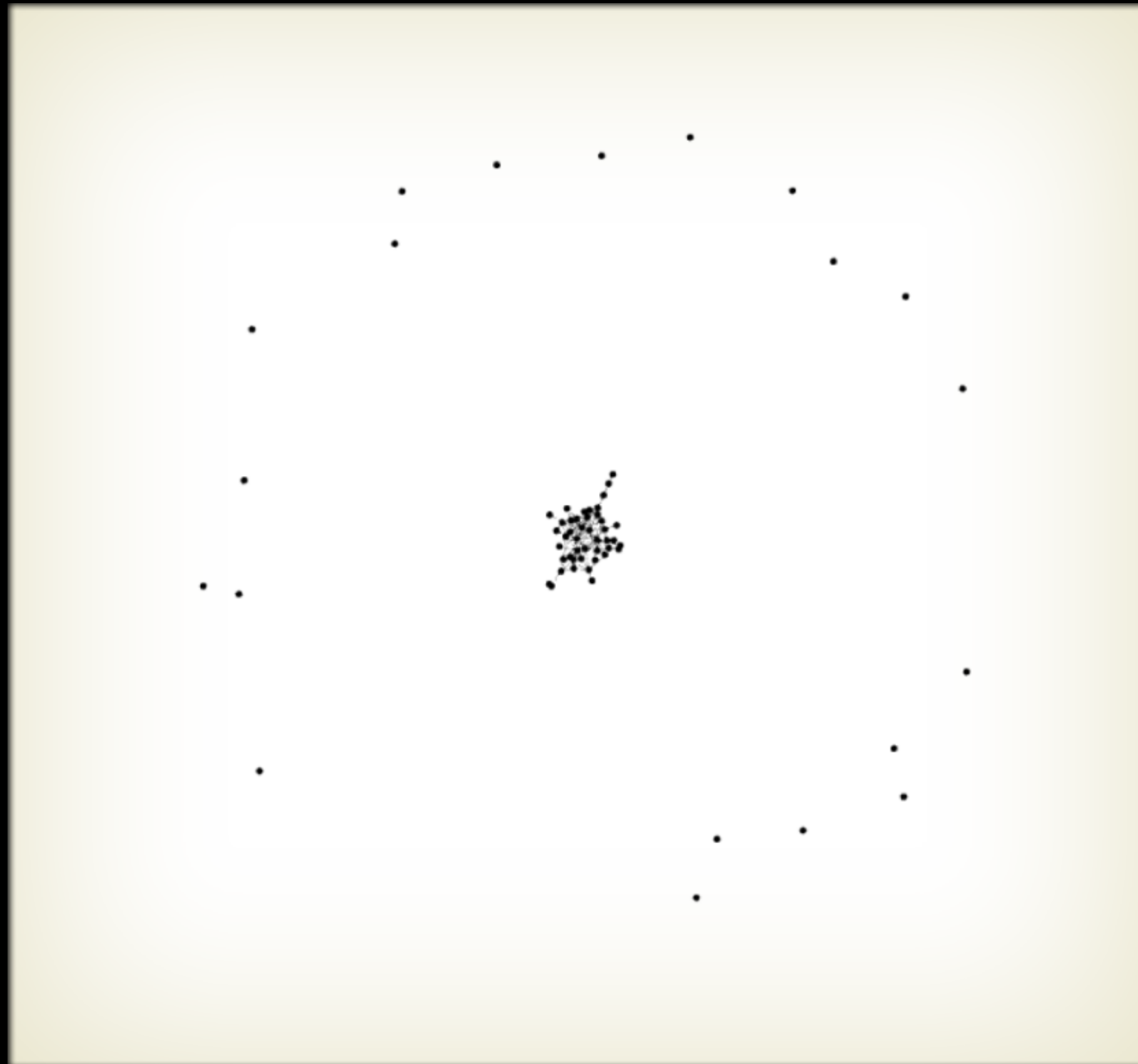
Ceci est mon ensemble de départ :
toutes les entités ne sont pas
connectées.



On ne
montre ici
que le IN

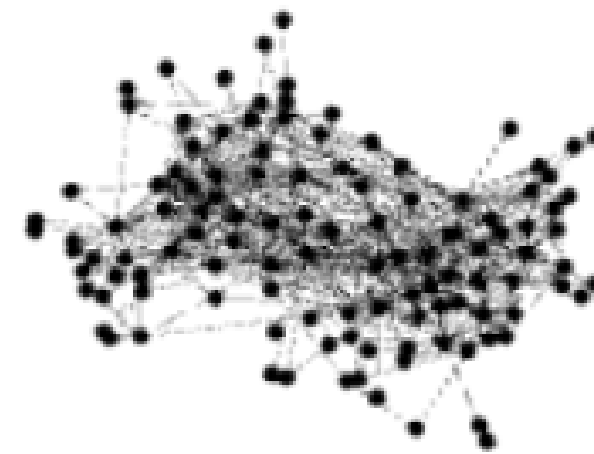
Un exemple empirique : l'avortement (abortion)

Ajouter plus d'entités agrège des entités précédemment déconnectées.



Un exemple empirique : l'avortement (abortion)

A partir d'un moment les entités s'agrègent en un «strongly connected component».

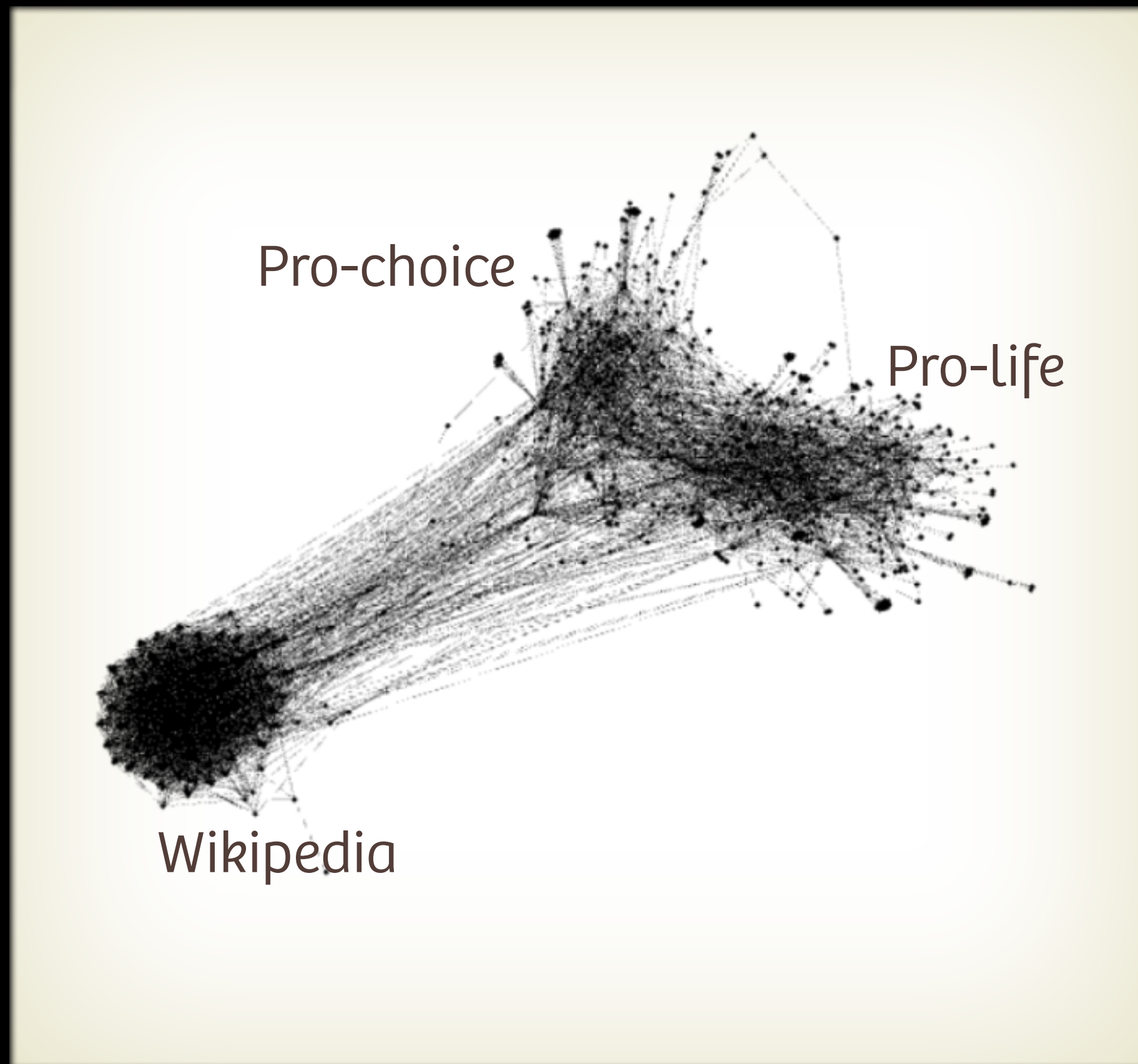


Des pôles
apparaissent
déjà

Un exemple empirique : l'avortement (abortion)

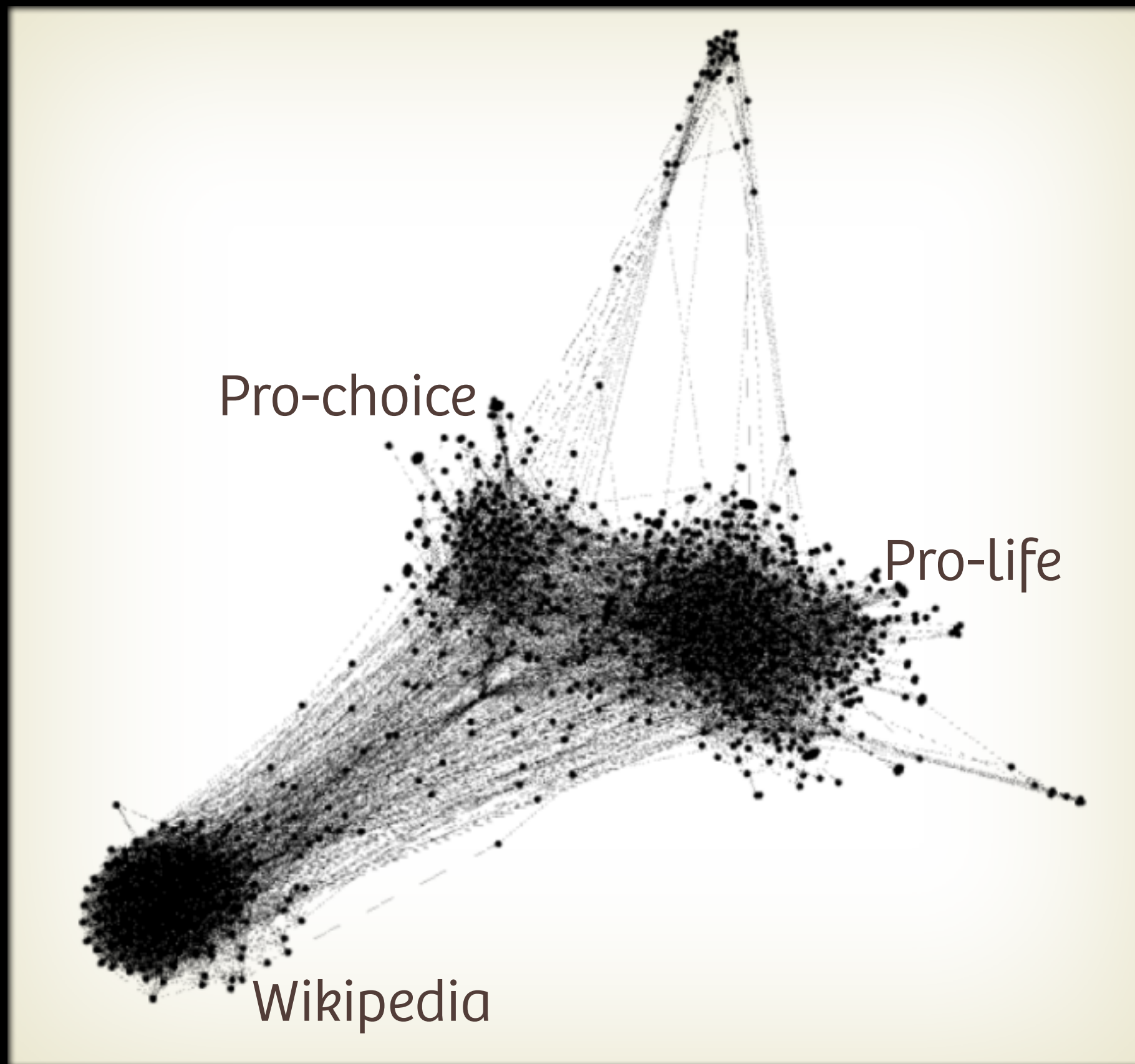
A ce point le corpus est plus complet.
Des clusters apparaissent plus
clairement.

NB: j'ai choisi ici de considérer les
pages Wikipedia comme des entités
distinctes. Ce contexte est important
pour interpréter le gros cluster.



Un exemple empirique : l'avortement (abortion)

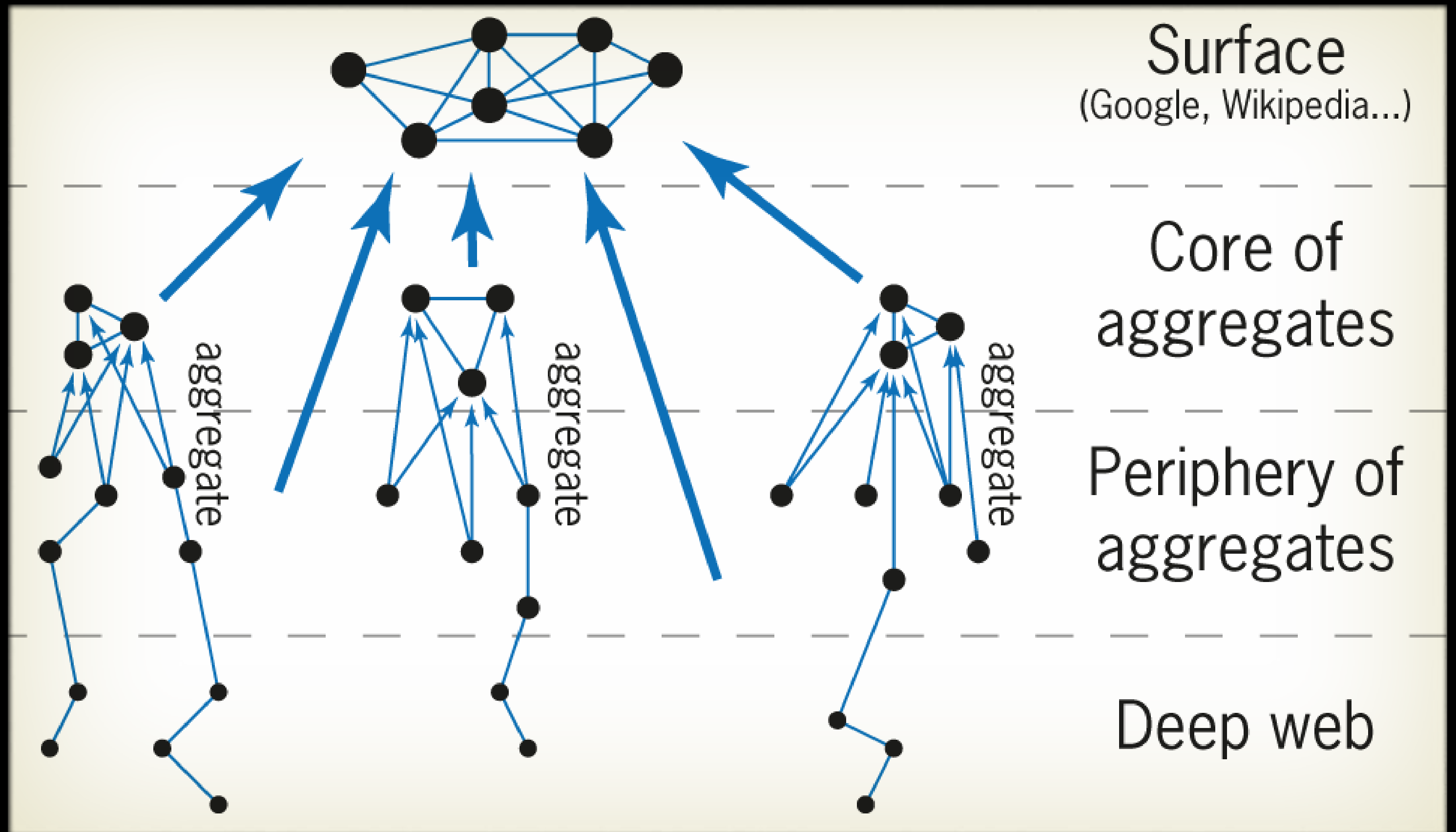
Une fois atteinte une relative exhaustivité, le corpus peut être considéré comme terminé.



IV.

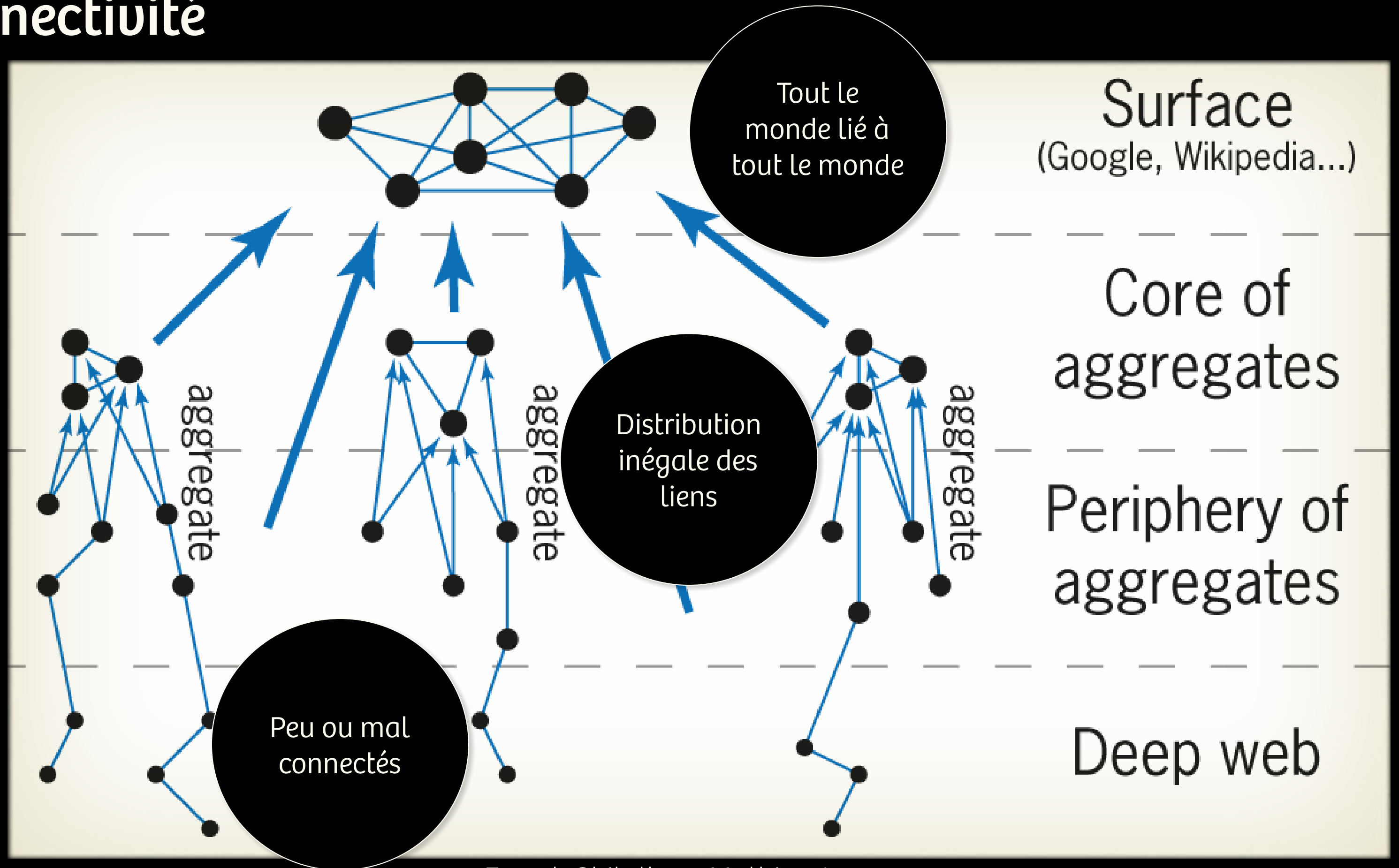
Le modèle du
web en couches

Le web en couches



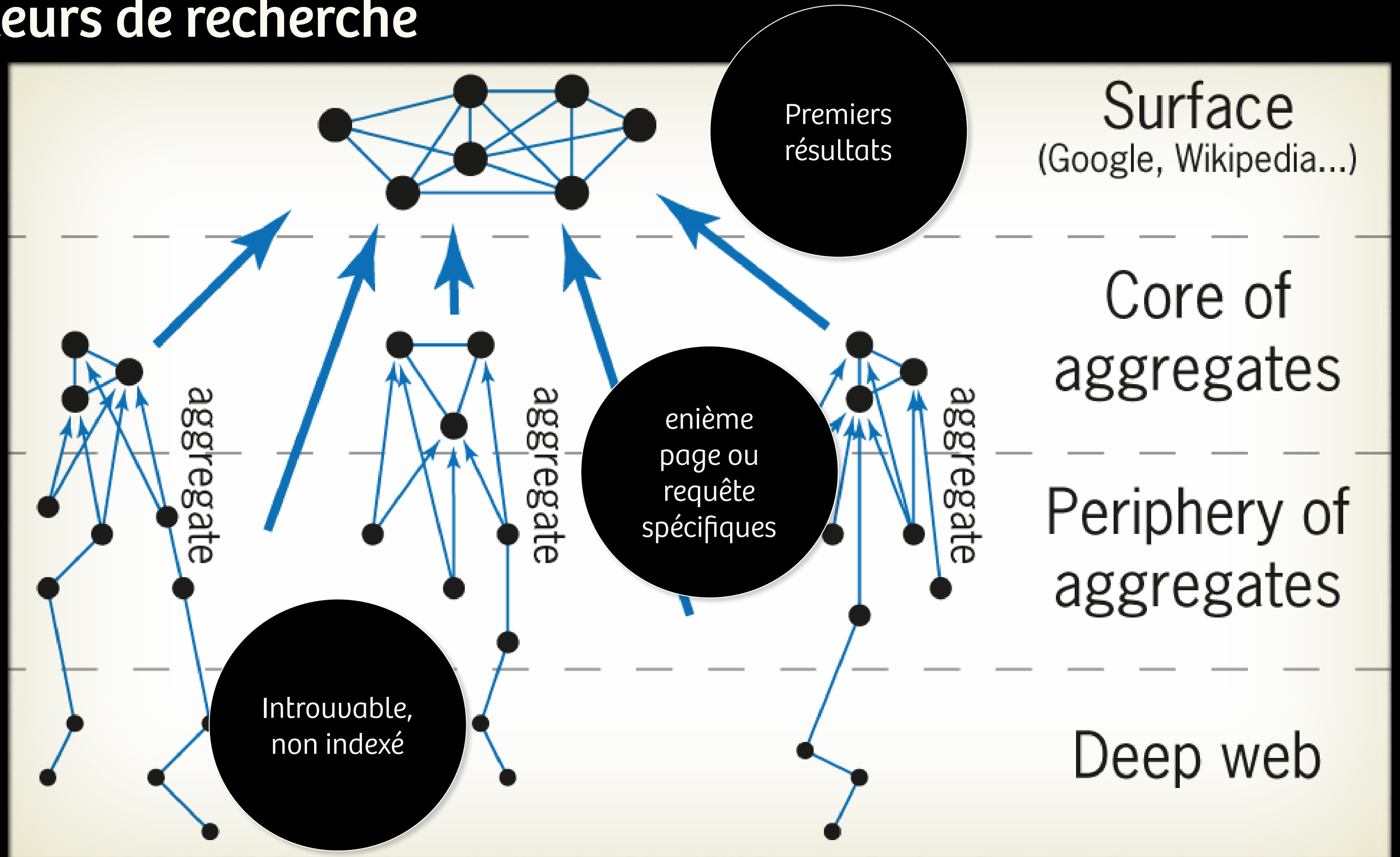
Le web en couches

Connectivité



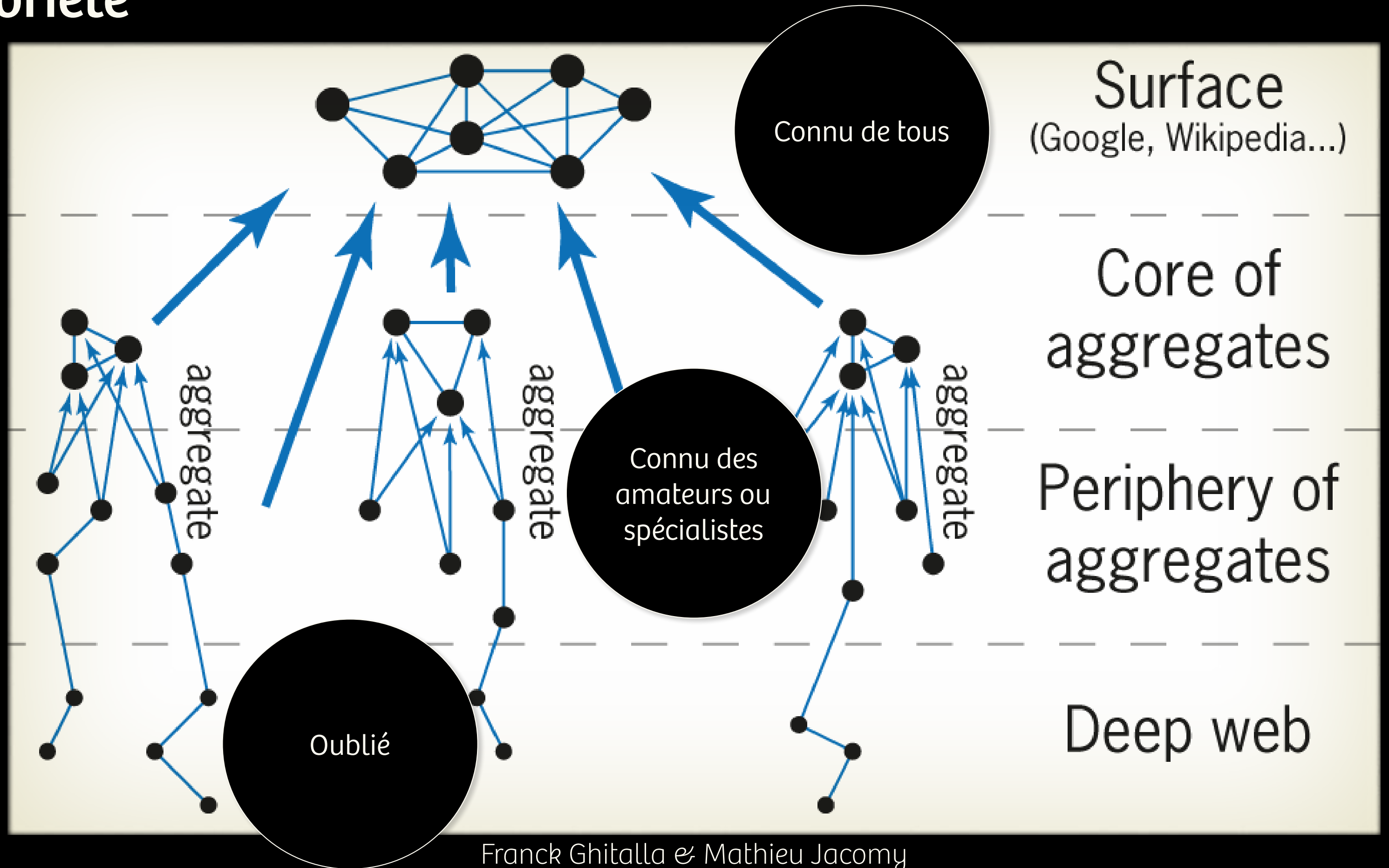
Le web en couches

Moteurs de recherche



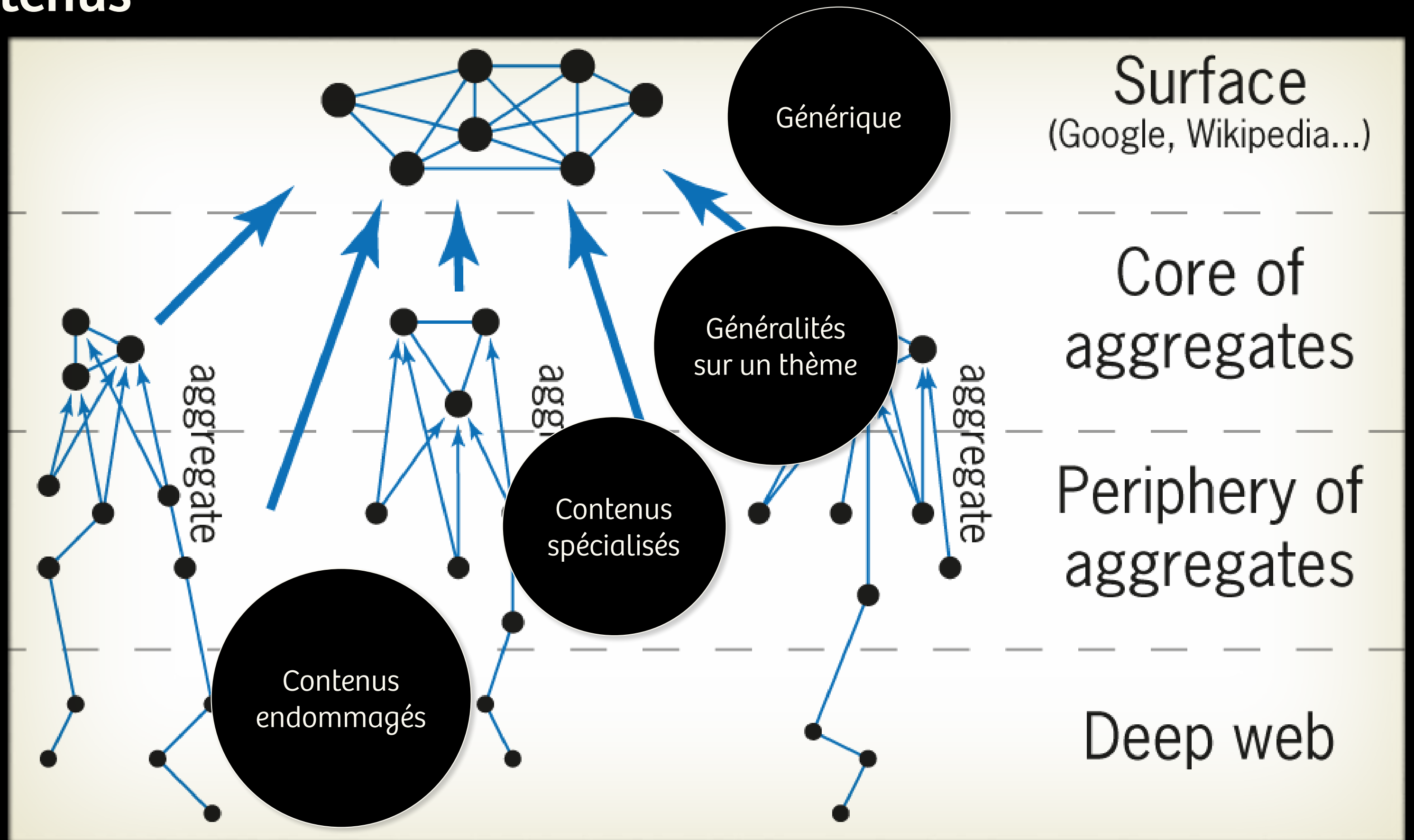
Le web en couches

Notoriété



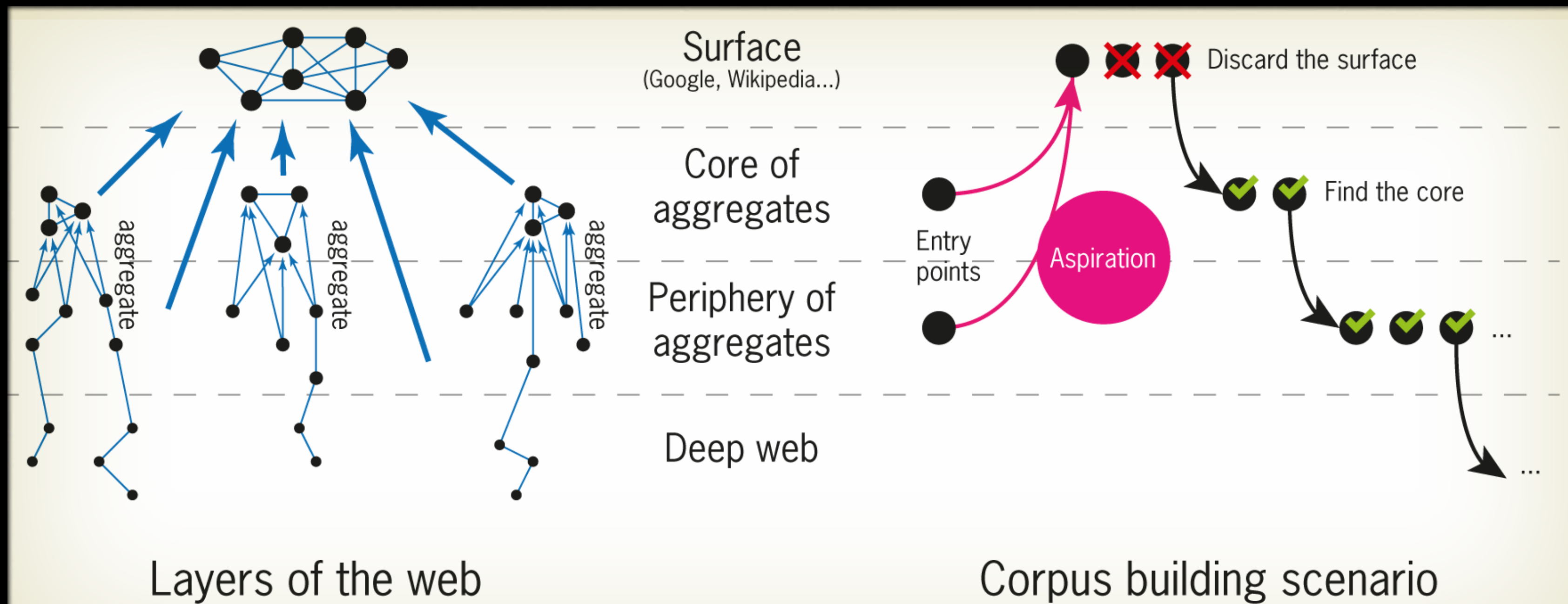
Le web en couches

Contenus



Le web en couches

Contenus



Merci de votre attention

@jacomya

Mathieu.Jacomy@sciencespo.fr

SciencesPo
MÉDIALAB