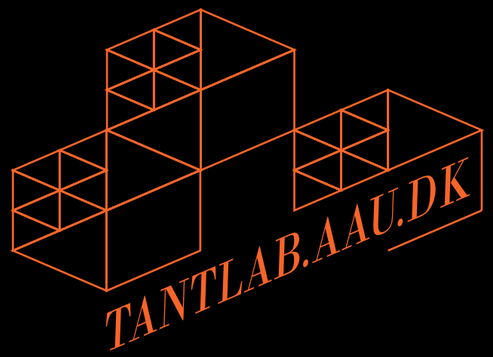


Growing a web corpus

Hyphe’s iterative curation on a web topology

Mathieu Jacomy
Aalborg University TANTLab

Web entity status in Hyphe

IN

Known and accepted in the corpus

OUT

Known and refused

UNDECIDED

Known but we cannot decide to accept or refuse it

DISCOVERED

Detected by Hyphe but not known by the user

Growing a web corpus in Hyphe

UNCHARTED

IN

This web entity
is the starting point

Growing a web corpus in Hyphe

UNCHARTED

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

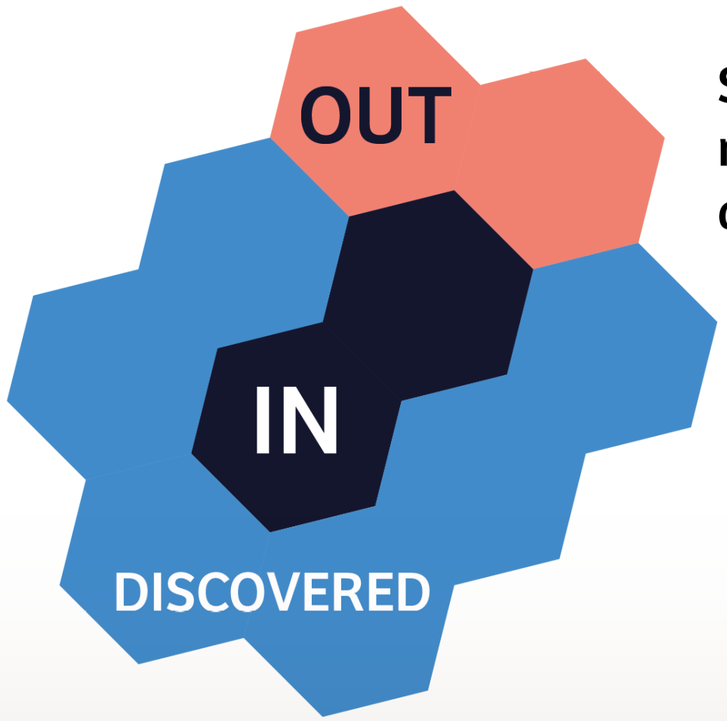
Growing a web corpus in Hyphe

UNCHARTED

Adding and crawling more web entities expands the corpus.

Growing a web corpus in Hyphe

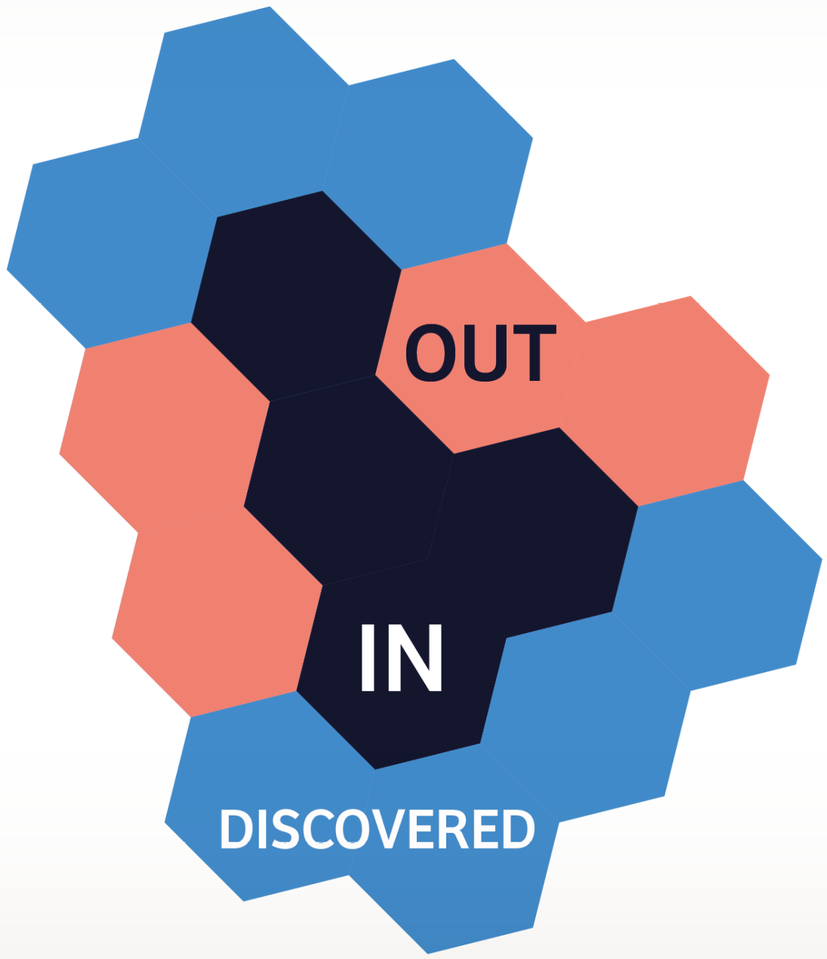
UNCHARTED



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

Growing a web corpus in Hyphe

UNCHARTED



Growing a web corpus in Hyphe

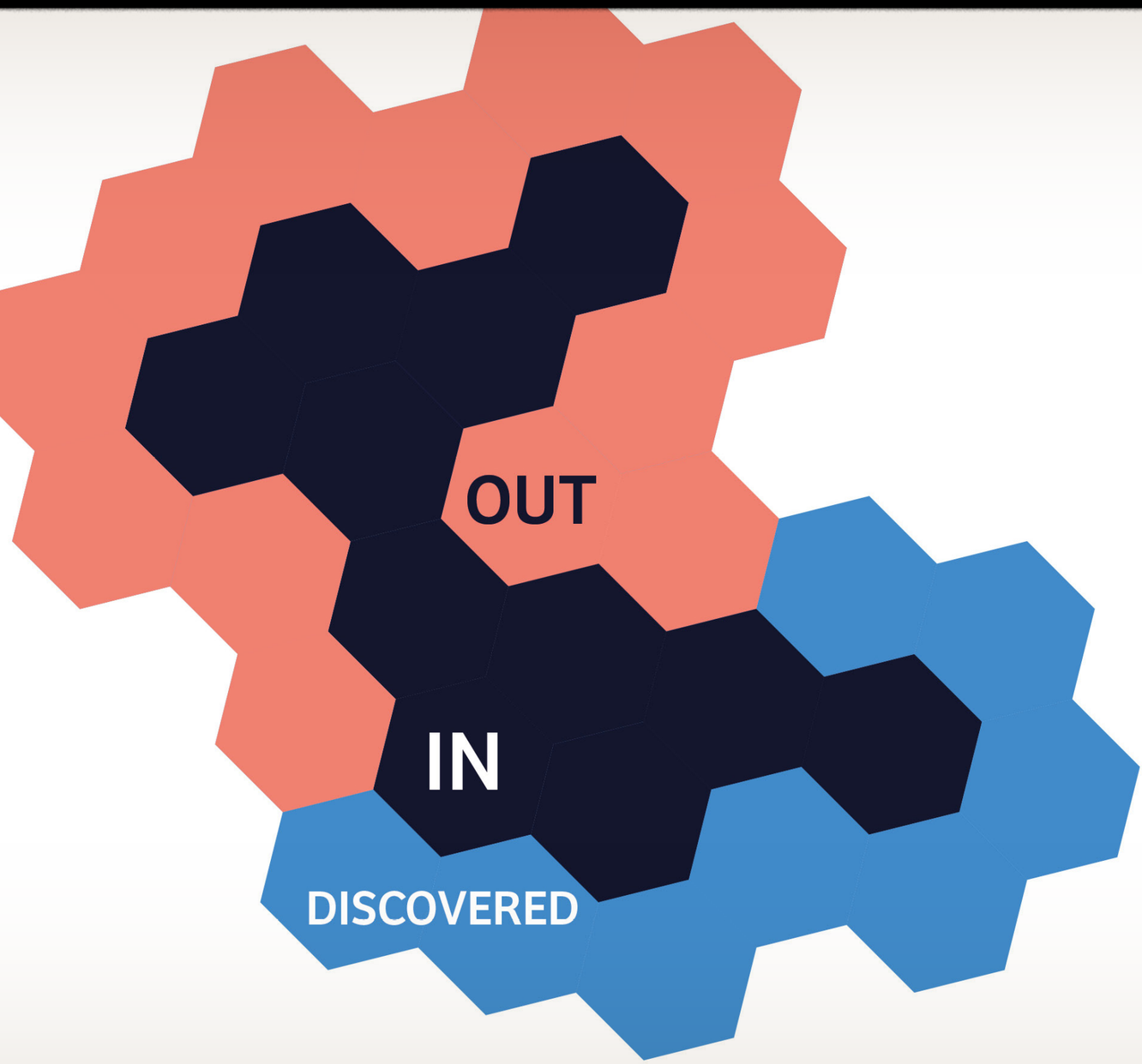
UNCHARTED

Growing a web corpus in Hyphe

UNCHARTED

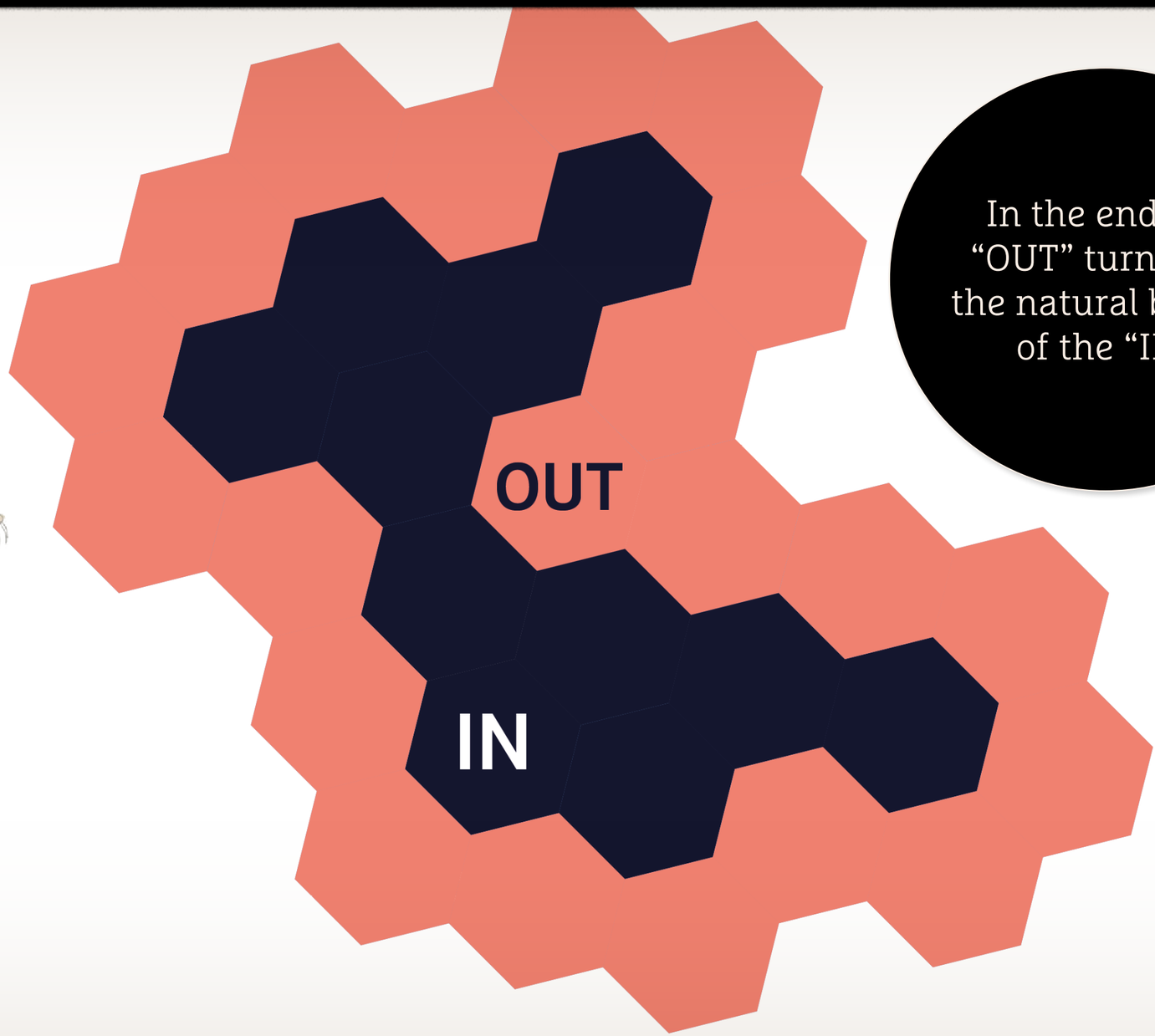
Growing a web corpus in Hyphe

UNCHARTED



Growing a web corpus in Hyphe

UNCHARTED



In the end the “OUT” turns out the natural border of the “IN”

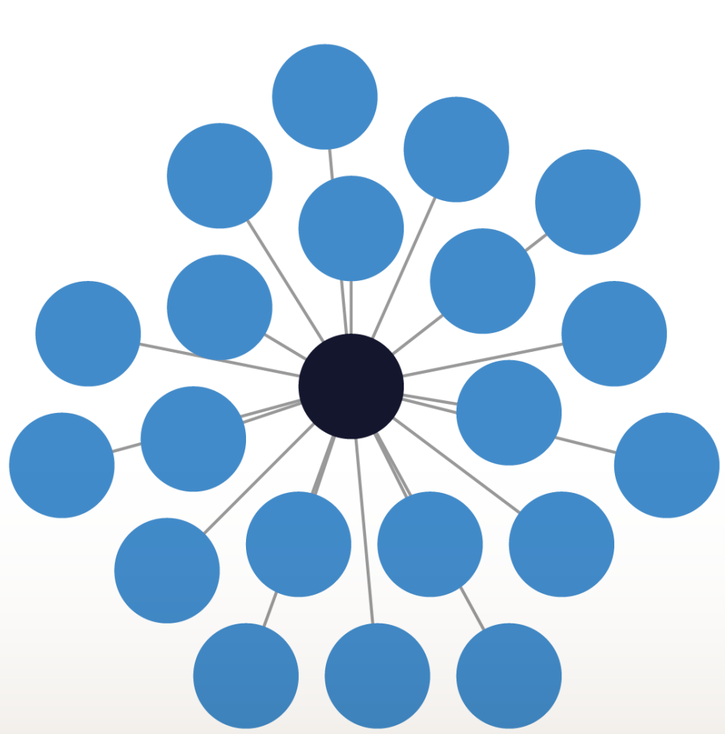
On actual web topology

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

This web entity
is the starting point

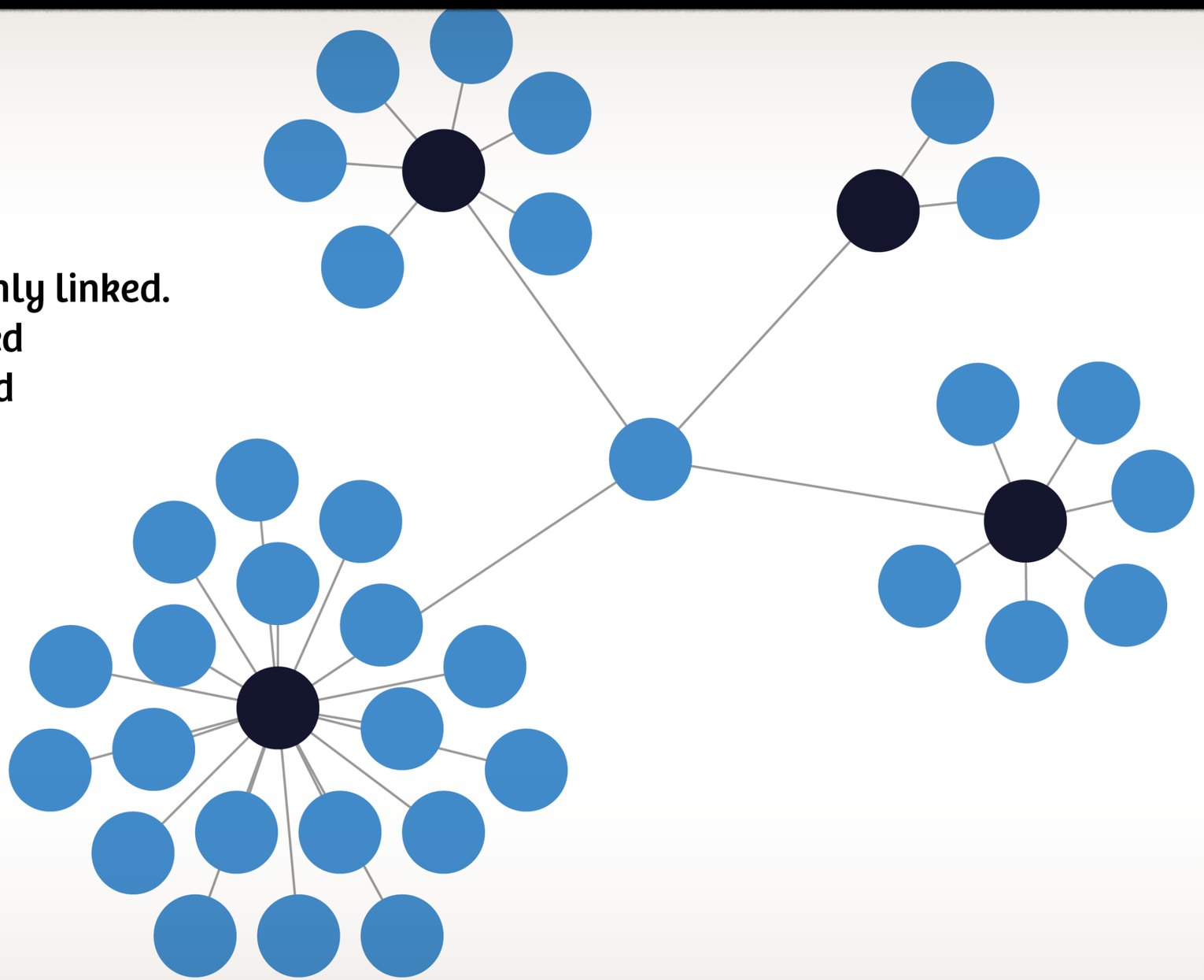
On actual web topology

Difference #1:
There are much more neighbors per web entity in the corpus



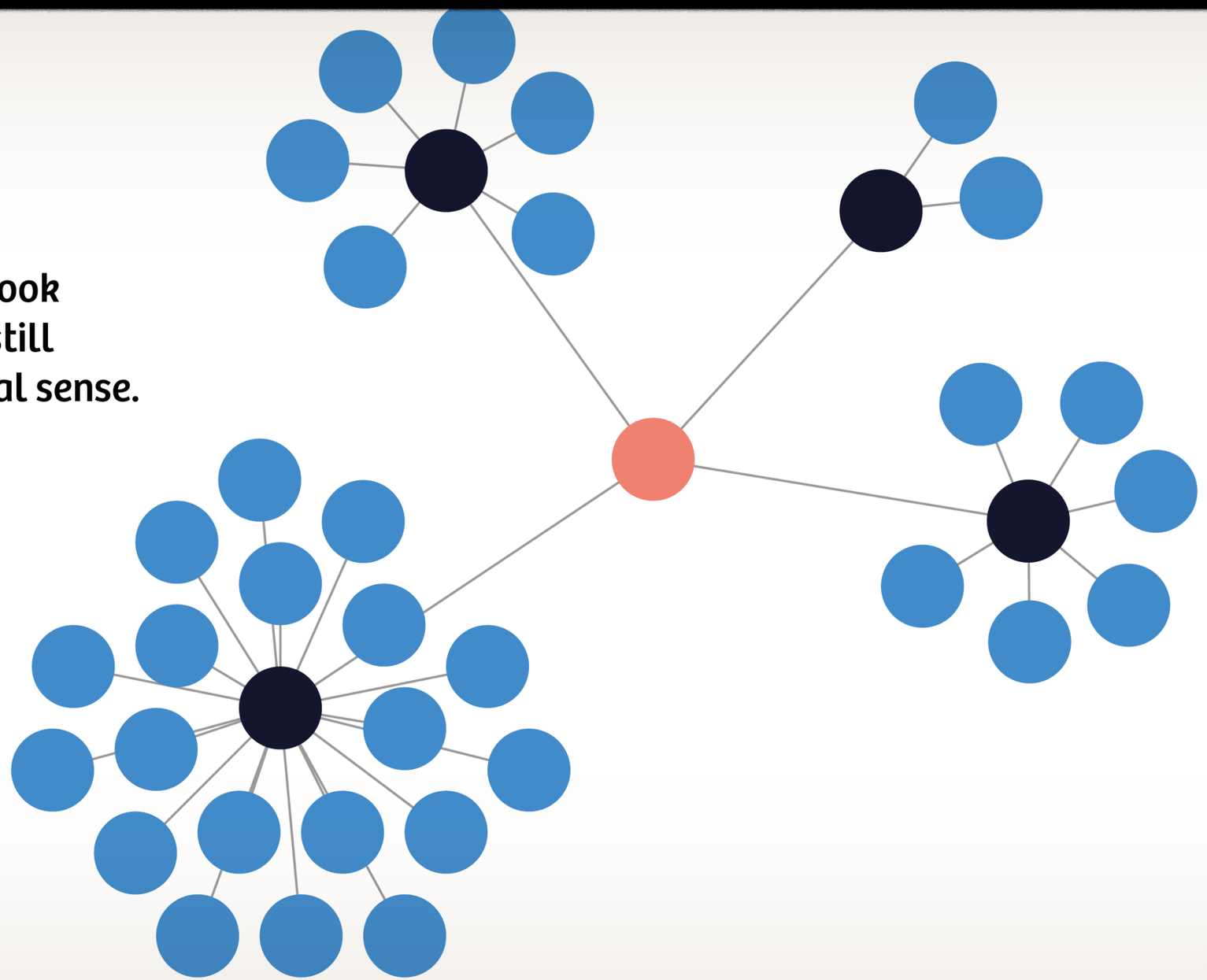
On actual web topology

Difference #2:
Neighbors are VERY unevenly linked.
- A few are highly connected
- Most are poorly connected

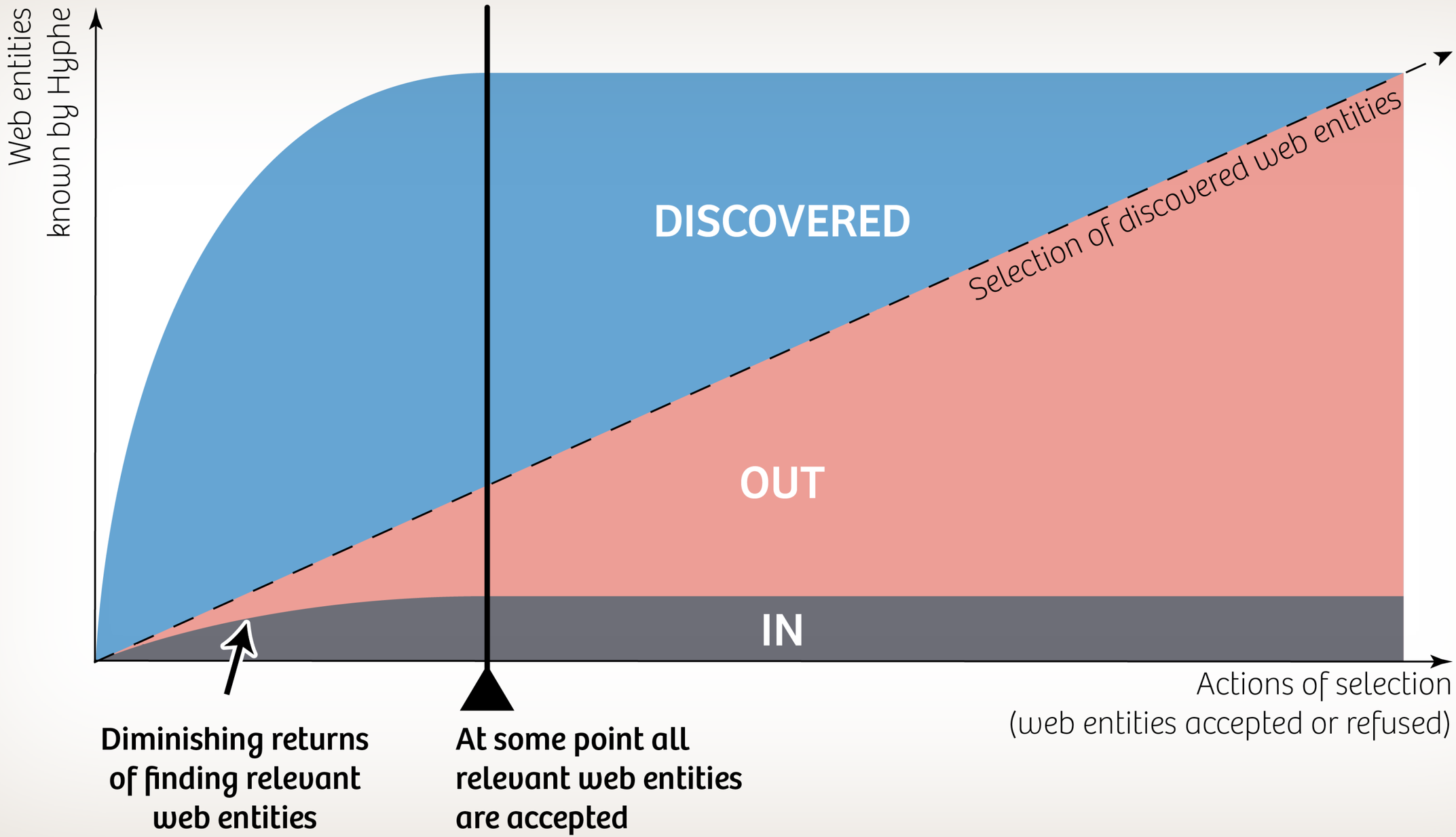


On actual web topology

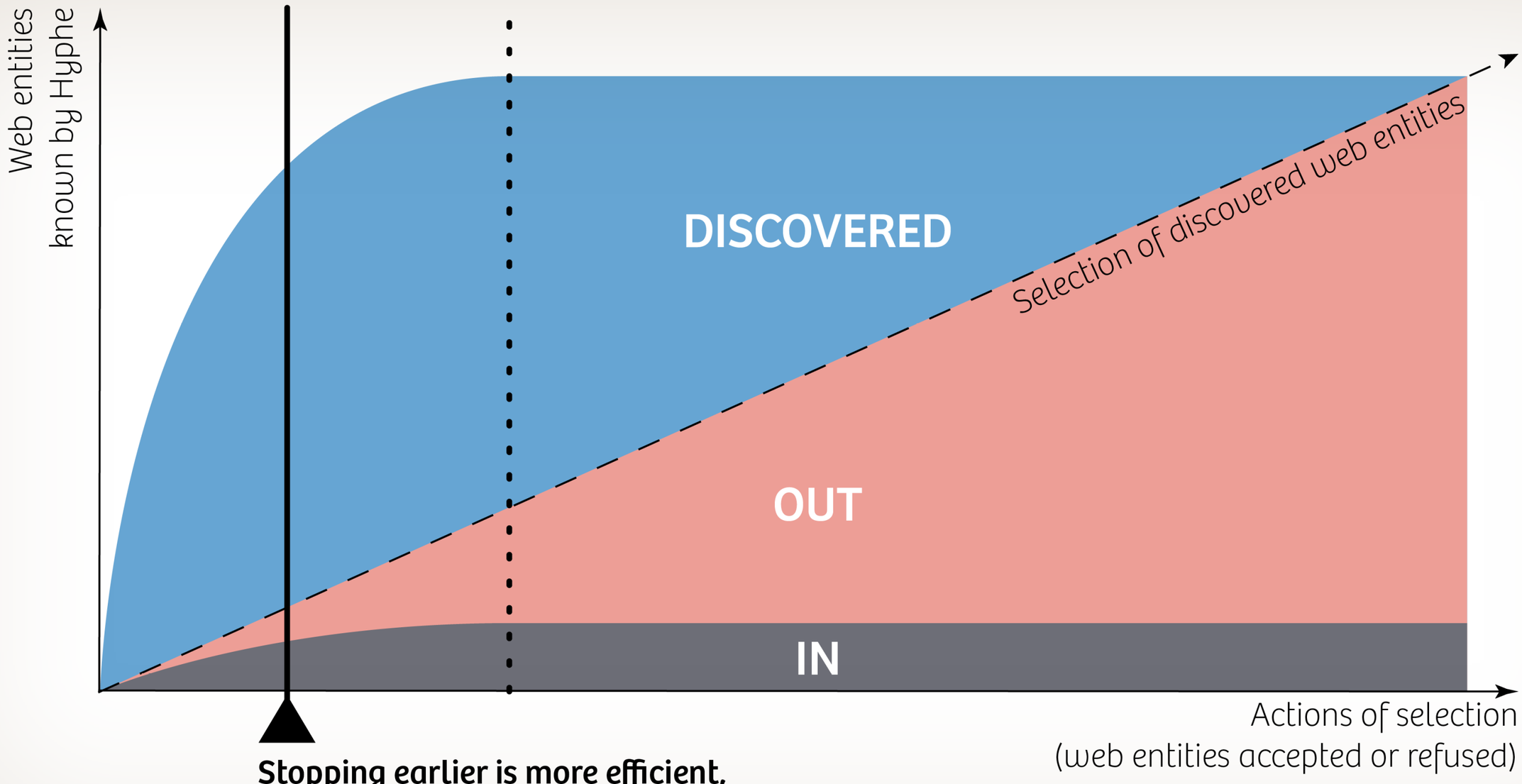
Difference #3:
Often the border does not look like it is “around”, but it is still the border in the topological sense.



Managing the prospection process



Managing the prospection process



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

Thank you for your attention

@jacomyma
reticular.hypotheses.org
Mathieu.Jacomy@gmail.com