Towards multiple ontologies in science mapping

A tribute to Loet Leydesdorff

Ismael Ràfols

INGENIO (CSIC-UPV), Universitat Politècnica de València

UNESCO Chair on Diversity and Inclusion in Global Science

Centre for Science and Technology Studies (CWTS), Leiden Univ.

The argument

- bottom-up algorithmic approaches, lack conceptual consistency regarding the type of categories used
  - in a same classification a category captures methods, another one has materials, a third one contains empirical objects and a fourth is focused on theories instead of having a single logic
- more useful representations by using ontologies based on a single logic that aligns with the particular conceptual needs of the analysis.
- Novel classification methods based on machine learning and language models hold promise to produce these tailored, question-driven ontologies.

CWTS

On taxonomies

"This book first arose out of a passage in Borges [which] quotes a 'certain Chinese encyclopaedia' in which it is written that 'animals are divided into:

> (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) *et cetera,* (m) having just broken the water pitcher, (n) that from a long way off look like flies'.

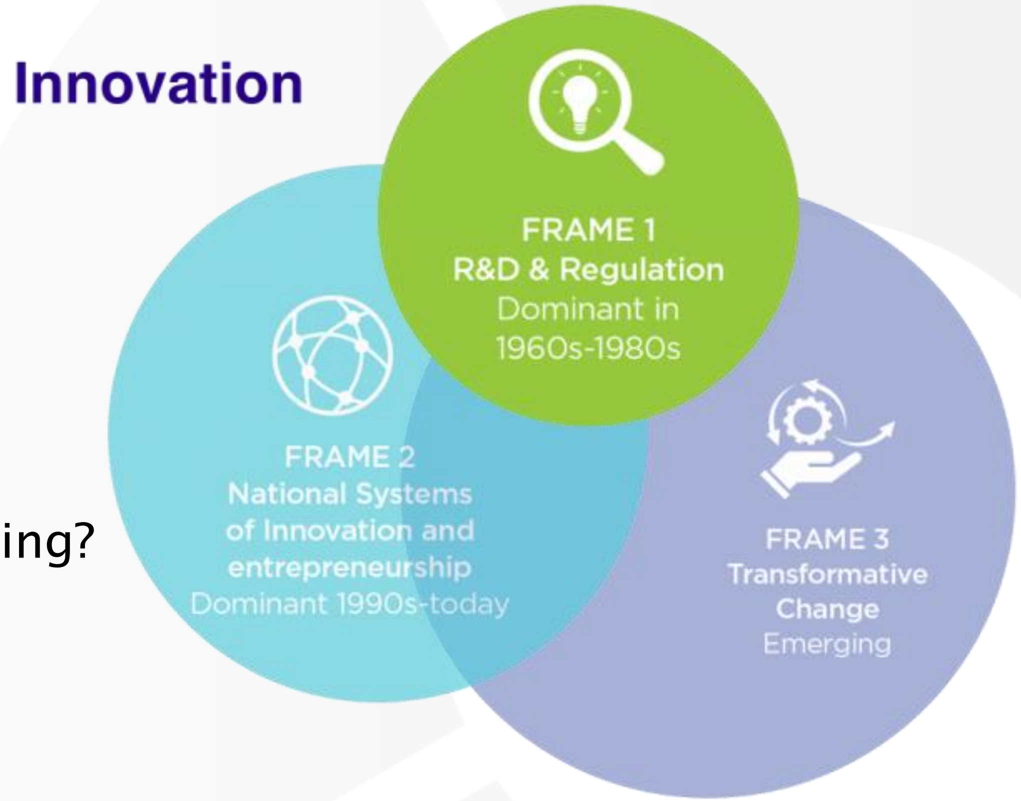
In the wonderment of this taxonomy, the thing we apprehend in one great leap (…) is the limitation of our own, the stark impossibility of thinking *that*."

Michel Foucault, *The order of things*

The motivation

Rethinking the role of science in society

- The social function of research?
  - Linear Model (1960s-80s)
  - Innovation systems (1990s-20s)
  - Transformative Innovation (emergent…)
- Social impact of research
  - What is the contribution of science to wellbeing?
- Mission- Oriented Research and Transformative Innovation
  - Impact is NOT enought – What TYPE of impact?



Transformative Innovation Policies Consortium
https://www.tipconsortium.net

CWTS

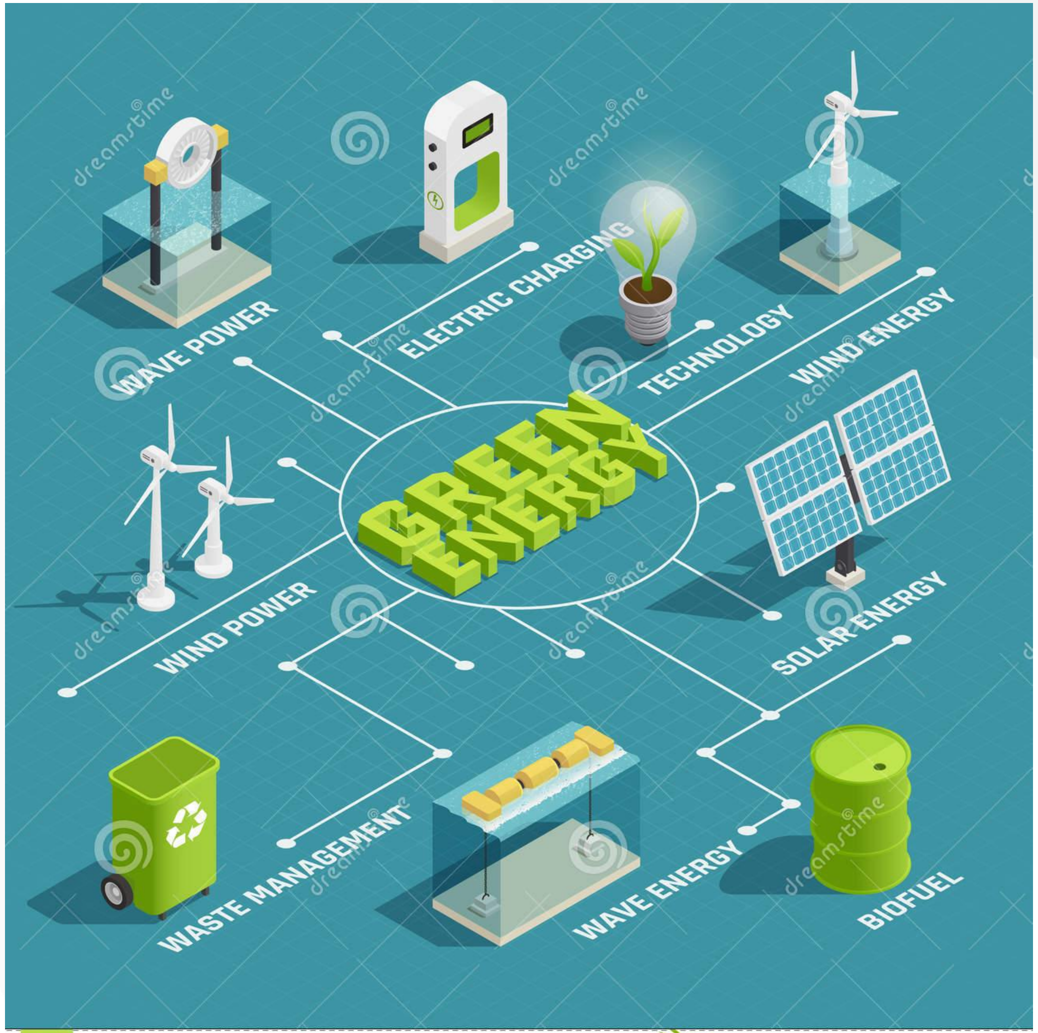
What type of research should be conducted?

What type of technological interventions for energy?

- Wind turbines?
- Biofuels?
- Nuclear plants?
- Solar panels?
- Network managements (centralised? decentralised)?
- Thermic isolation materials?
- Reduction of transportion?

Portfolio – managing diversity:
what is an appropriate balance between research options?

But how can we classify research according to consistent categories?



CWTS

The problem of classification

A tribute to Loet Leydesdorff

‘What are the natural structural units of science?

How are these structural units related to one another?

What are the forces which determine these structural units and their interrelations?

How does the structure of science change over time, both at a macro- and at a micro-level?’

(Leydesdorff, 1987, p. 294).

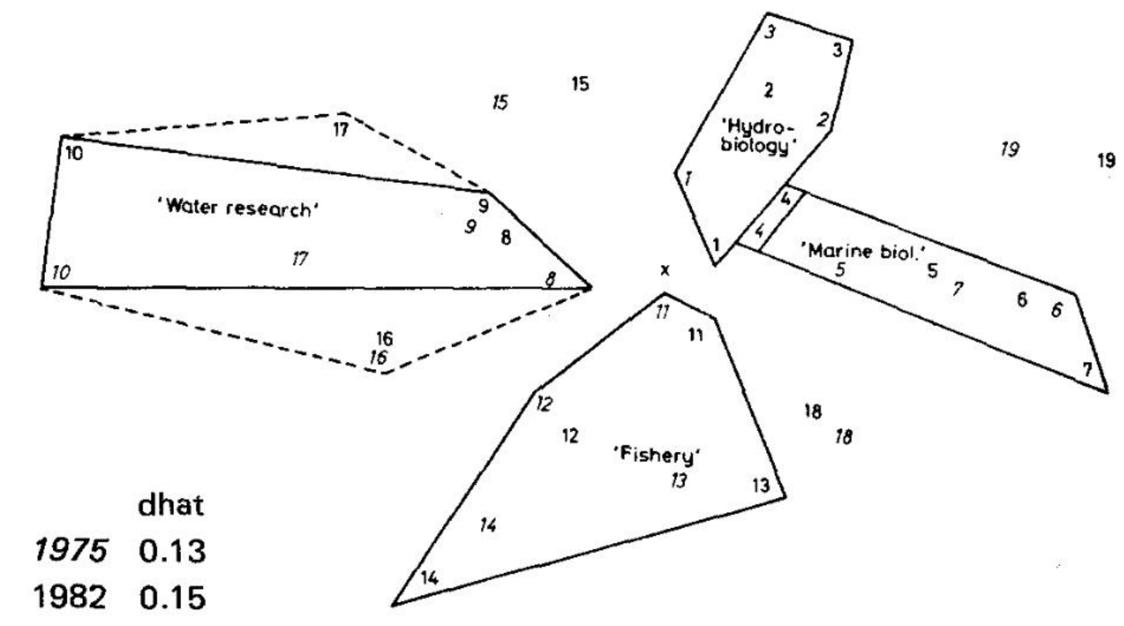


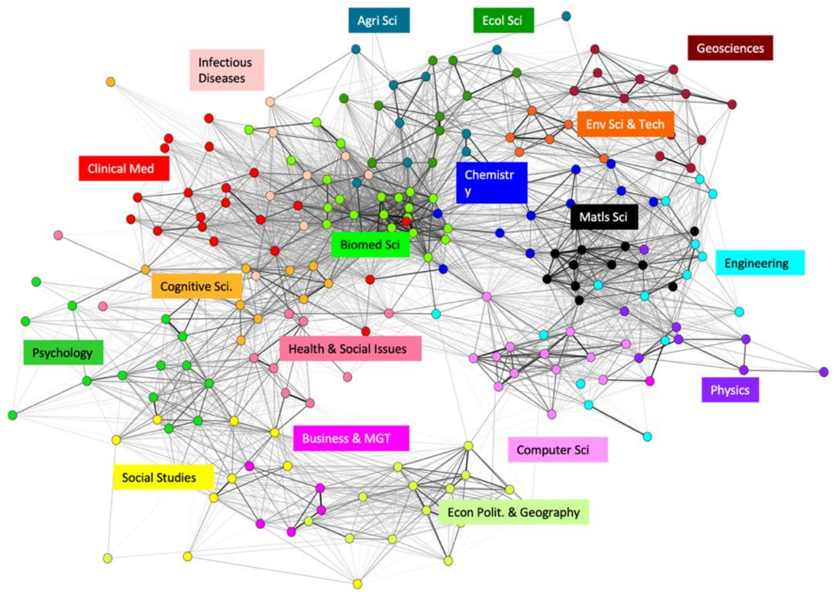
Figure 1. Epistemic distances among 19 ‘aquatic ecology’ journals in 1975 (*italics*) and 1982. Source: Leydesdorff (1986).

CWTS

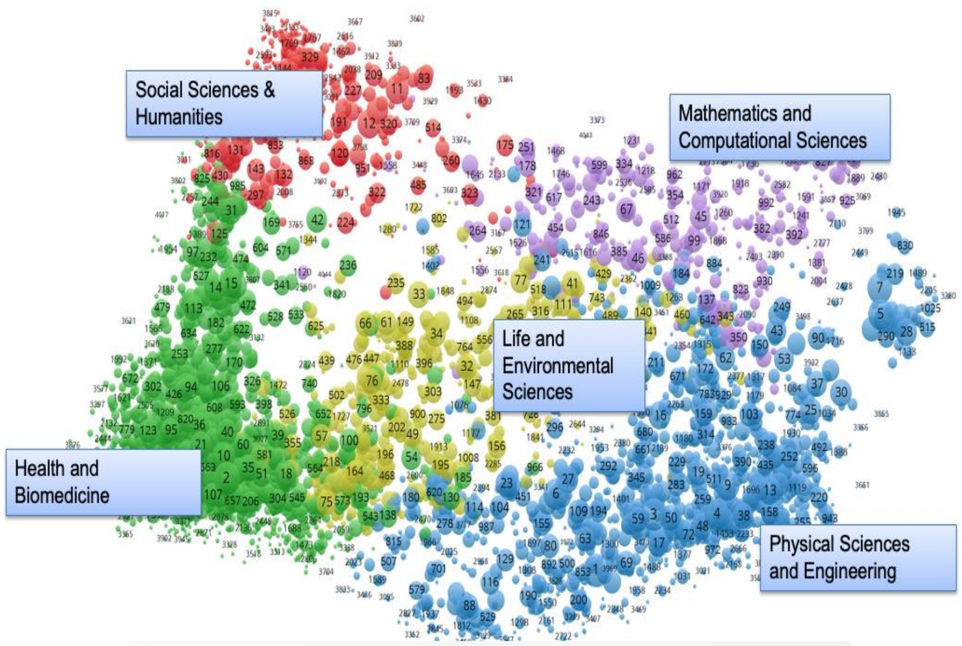
From 250 WoS Subject Categories to 4,000 article-level clusters (ALCs)

Science structure:	Number of documents per category per year	**Evolution:**
	10^6	
10 fields (e.g. medicine)		No morphological change
	10^5	
500 disciplines	10^4	+99% stability
8,000 specialties	10^3	+95% stability
	10^2	
100,000 problems		56% stability (cluster survival rate)
	10^1	

Map of WoS Subject categories



Map of 4,000 article-level clusters

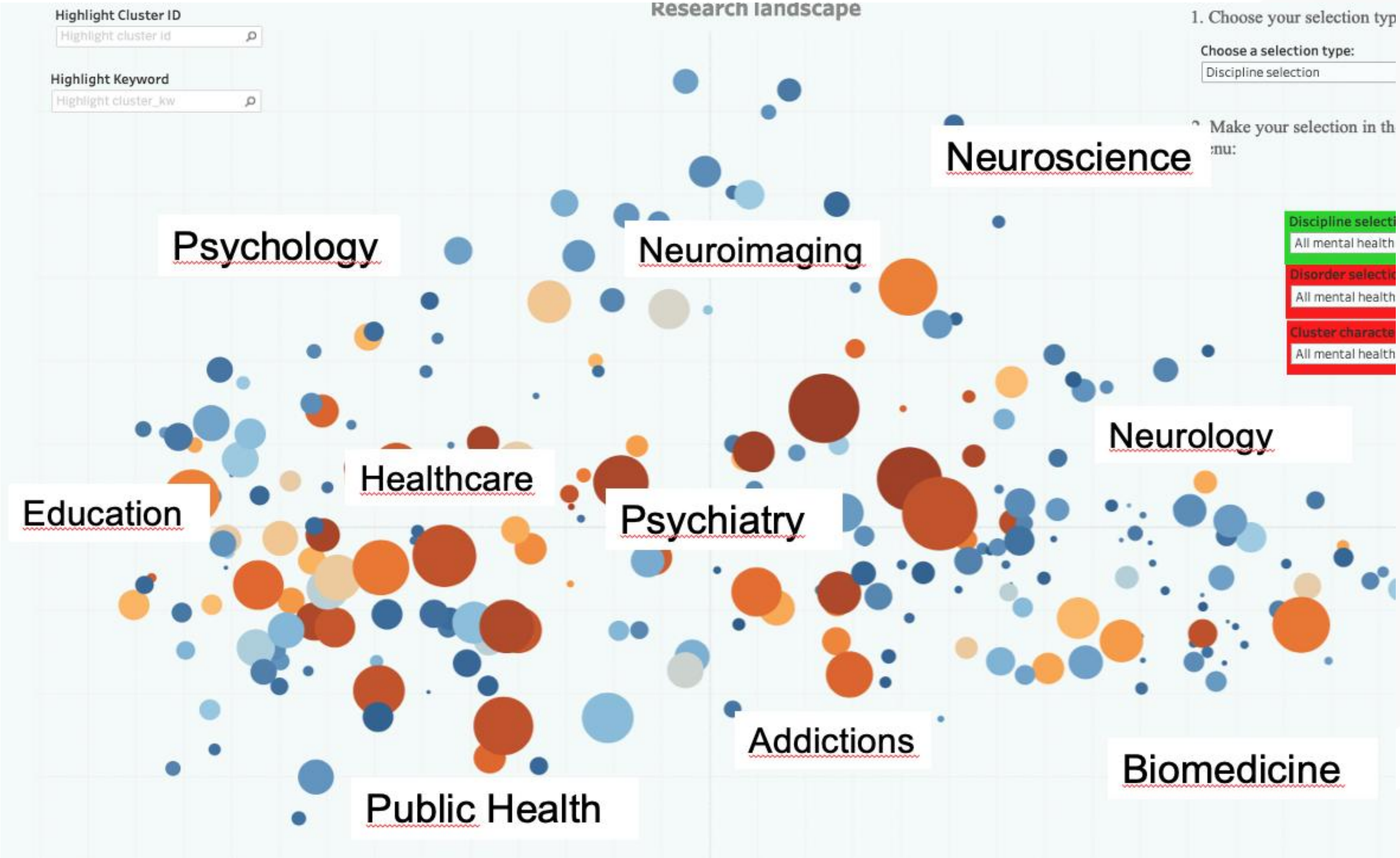


Mapping mental health for the Swedish government

- National Strategy on Mental Health in Sweden
  - Report from 26 authories in mid-2021. Estrategy to be decided by 2023
  - Mental Health research agenda of Swedish funding agency for health and welfare (FORTE)
- Vinnova's programme on 'Sustainable Precision Health'
  - Beyond Mental health. Parallel study on 'Cardiometabolic research'
- Health related to environmental and social factors
- Beyond medicine and biology, consider research relevant for a 'healthier society'
- Interest in themes/topics that can be narrowly targeted
- Explicit interest in prevention, rehabilitation, social determinants of mental health, Artificial Intelligence

The research landscape of mental health
-- fails to signal prevention, social determinants, etc.

Van De Klippe et al. (SPP, 2023)



The use of micro-clusters aims to capture both:

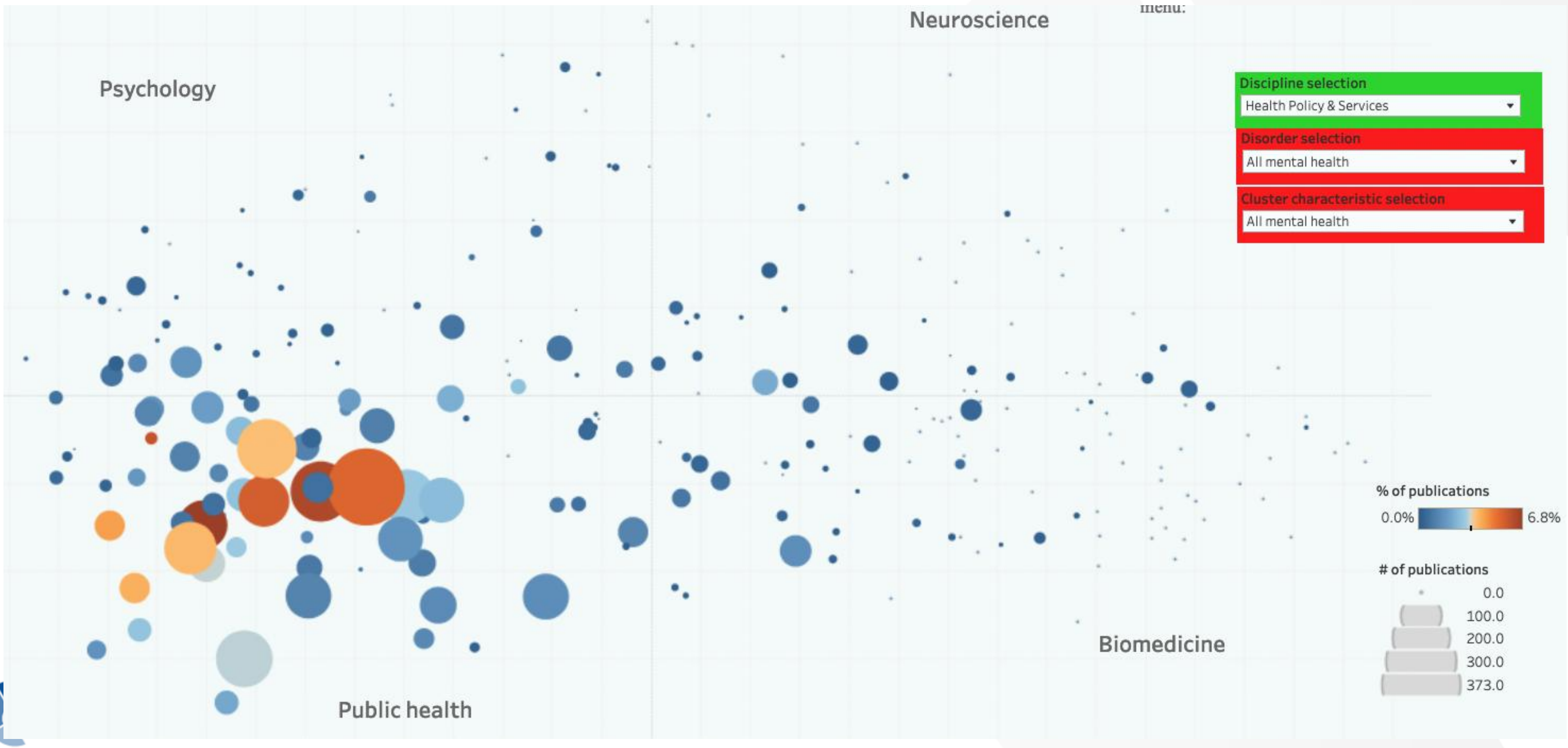
- Clusters specialised in MH (e.g. autism, schizophrenia, with >90% related to MH)
- Clusters where MH is a secondary but relevant topic (e.g. HIV stigma, child overweight and counselling, with about 10-20% related to MH)
- Number of pubs – size
- Colour: % of pubs related to MH

even if we include all clusters (380) with more than 10% of publications related to mental health, about 20% (!!) of the corpus remains spread over the other 3,633 clusters

The cluster fail to distinguish prevention, rehabilitation, etc. → but help guess patterns

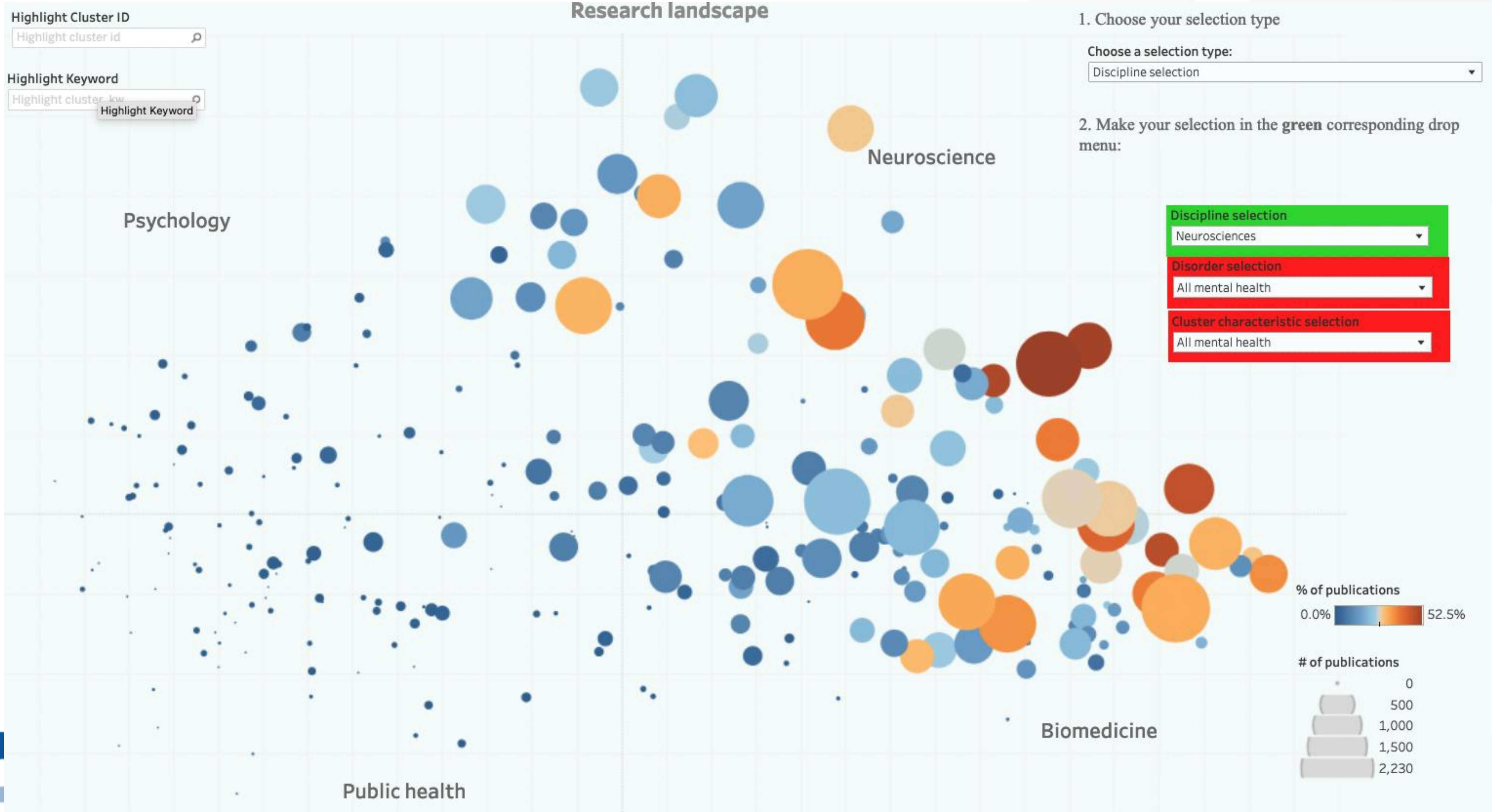
Disciplines fall in certain areas, but most clusters are multidisciplinary

Disciplinary Category Wos: ‘Health Policy and Services’

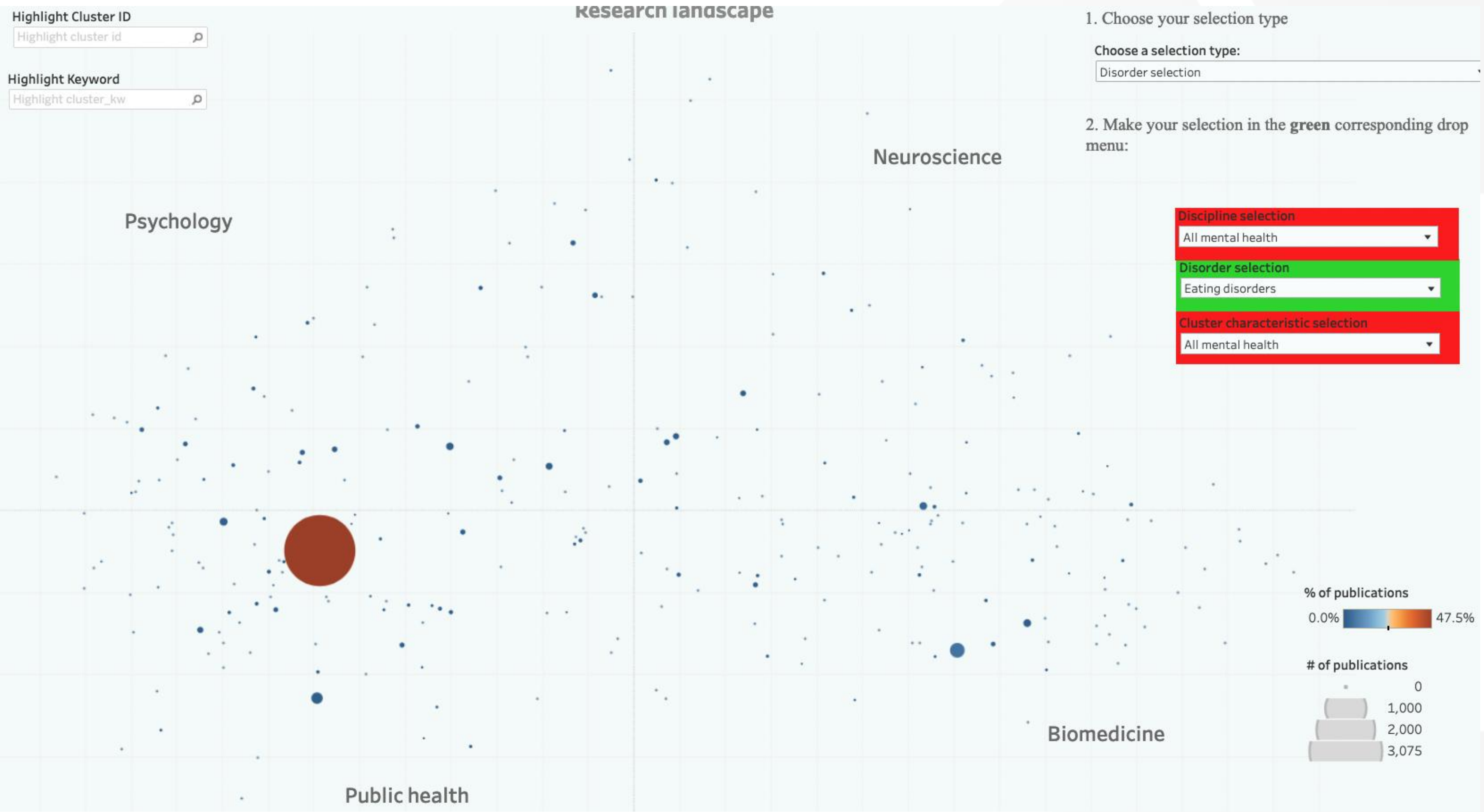


Disciplines fall in certain areas, but most clusters are multidisciplinary
Not a nested structure.

Disciplinary Category Wos: ‘Neurosciences’



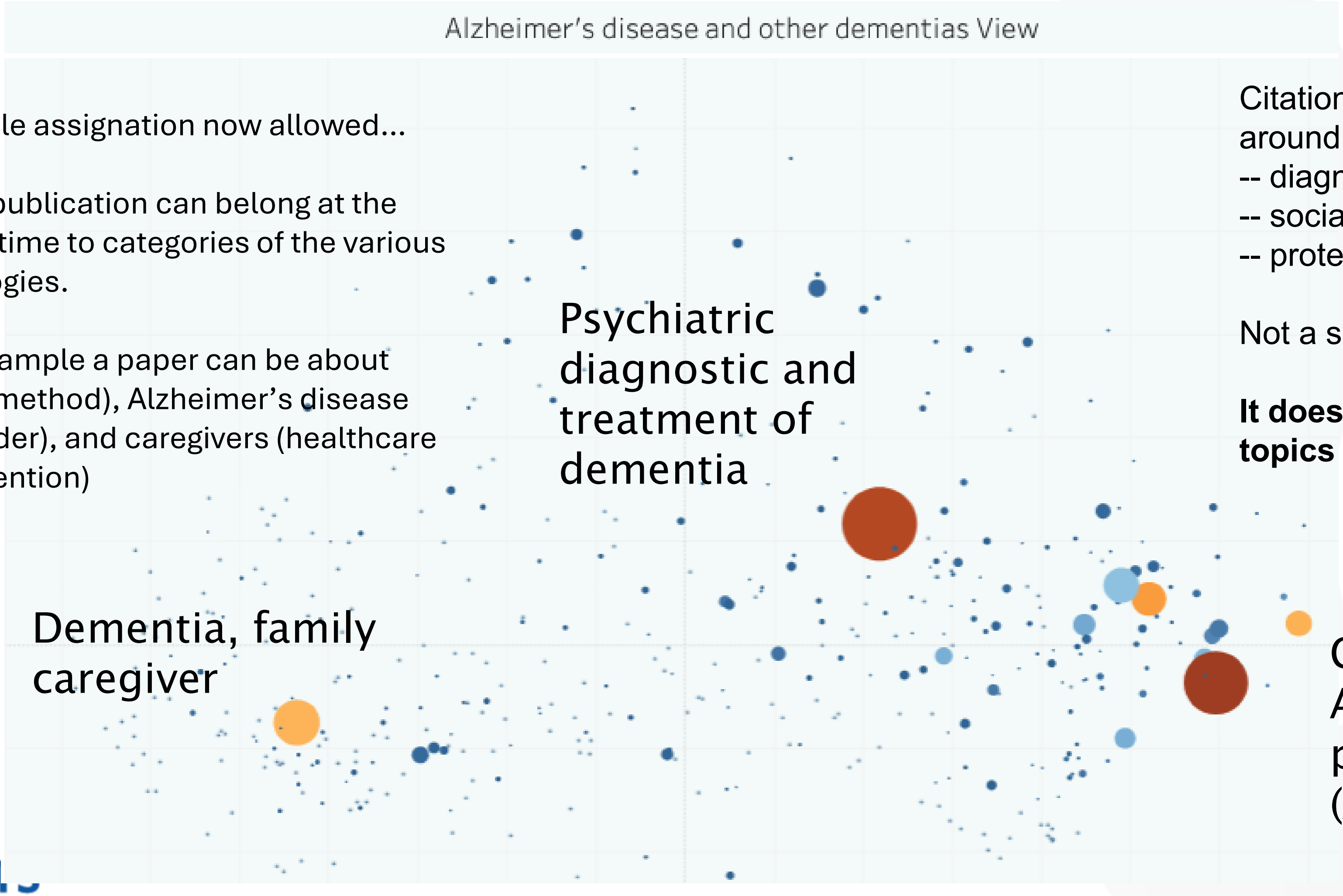
Topics in ALCs are mainly empirical objects of study
Eating disorders



Problems with ALCs:
1) Multiple logics for creating a cluster, but seldom the logics of interest for the policy question: prevention, healthcare, diagnosis, treatment…

Multiple assignation now allowed…
BUT
each publication can belong at the same time to categories of the various ontologies.

For example a paper can be about fMRI (method), Alzheimer's disease (disorder), and caregivers (healthcare intervention)



Citation-based clusters group around different types of topics:
-- diagnostic and treatment
-- social healthcare (caregiver)
-- protein

Not a systematic classification

It does not account for belonging to various topics

On taxonomies

"This book first arose out of a passage in Borges [which] quotes a 'certain Chinese encyclopaedia' in which it is written that 'animals are divided into:

(a) belonging to the Emperor, (b) embalmed, (c) tame, (d) suckling pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) *et cetera,* (m) having just broken the water pitcher, (n) that from a long way off look like flies'.

In the wonderment of this taxonomy, the thing we apprehend in one great leap (…) is the limitation of our own, the stark impossibility of thinking *that*." (Foucault, 2012)

On taxonomies

MICHEL FOUCAULT

“This book first arose out of a passage in Borges [which] quotes a ‘certain Chinese encyclopaedia’ in which it is written that ‘animals are divided into:

> (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) *et cetera,* (m) having just broken the water pitcher, (n) that from a long way off look like flies’.

In the wonderment of this taxonomy, the thing we apprehend in one great leap (…) is the limitation of our own, the stark impossibility of thinking *that*.”

Foucault, *The order of things*

“…when we say that a cat and a dog resemble each other less than two greyhounds do, even if both are tame or embalmed, even if both are frenzied, (…) what is the ground on which we are able to establish the validity of this classification with complete certainty? On what ‘table’, according to what grid of identities, similitudes, analogies, have we become accustomed to sort out so many different and similar things?…”

CWTS

tel gallimard

Problems with ALCs:
2) Long tails due to single assignation of paper to cluster

Obviously, the ordering of publications by empirical object (…) is only one of potentially many useful thematic perspectives (…). An alternative thematic perspective that focuses on methods might distinguish field-observational studies, experimental studies, macroecological studies, and policy-oriented social studies (…). In principle, by use of different approaches and types of data one should be able to tease out different aspects (…).

(Held & Velden, 2022, p. 666)

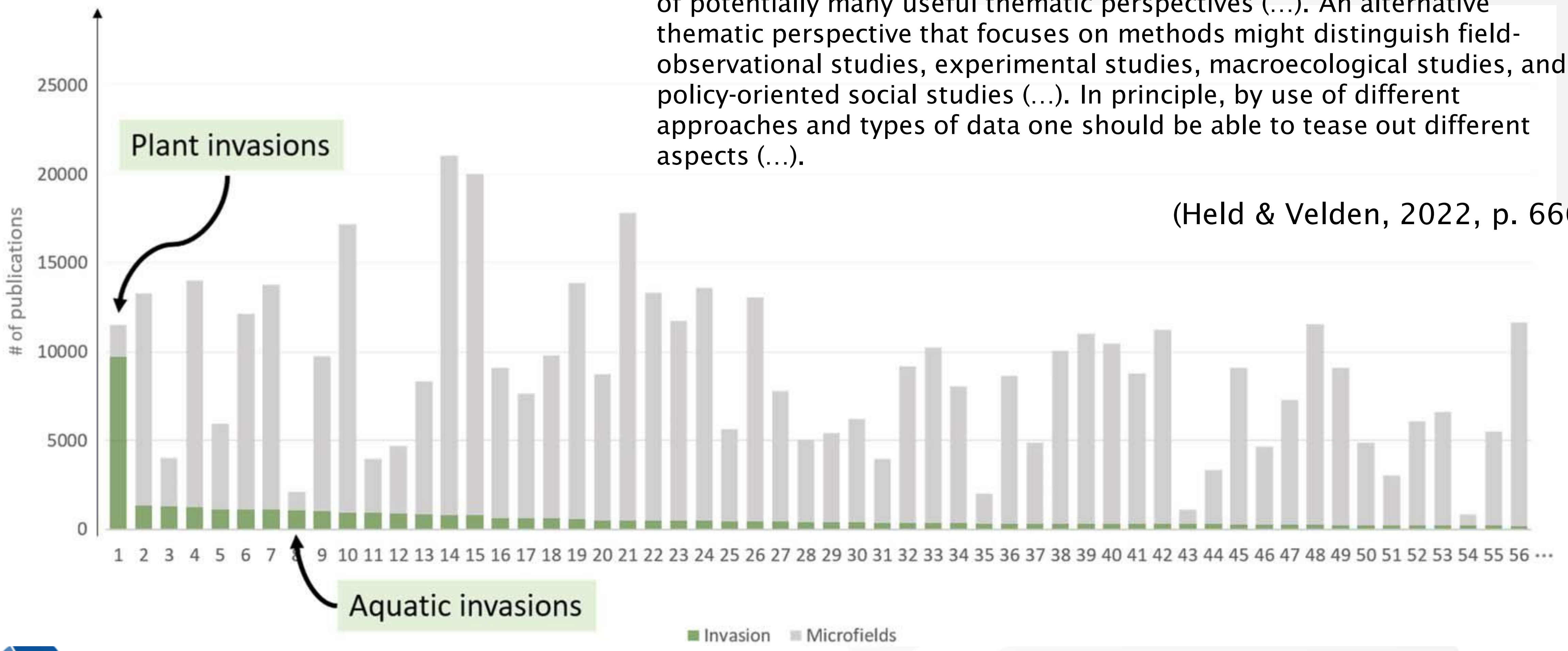


Figure 5. Distribution of 'invasion biology' publications. Source: Held & Velden (2022).

Classifications are never ‘natural’

Choosing relevant ontologies

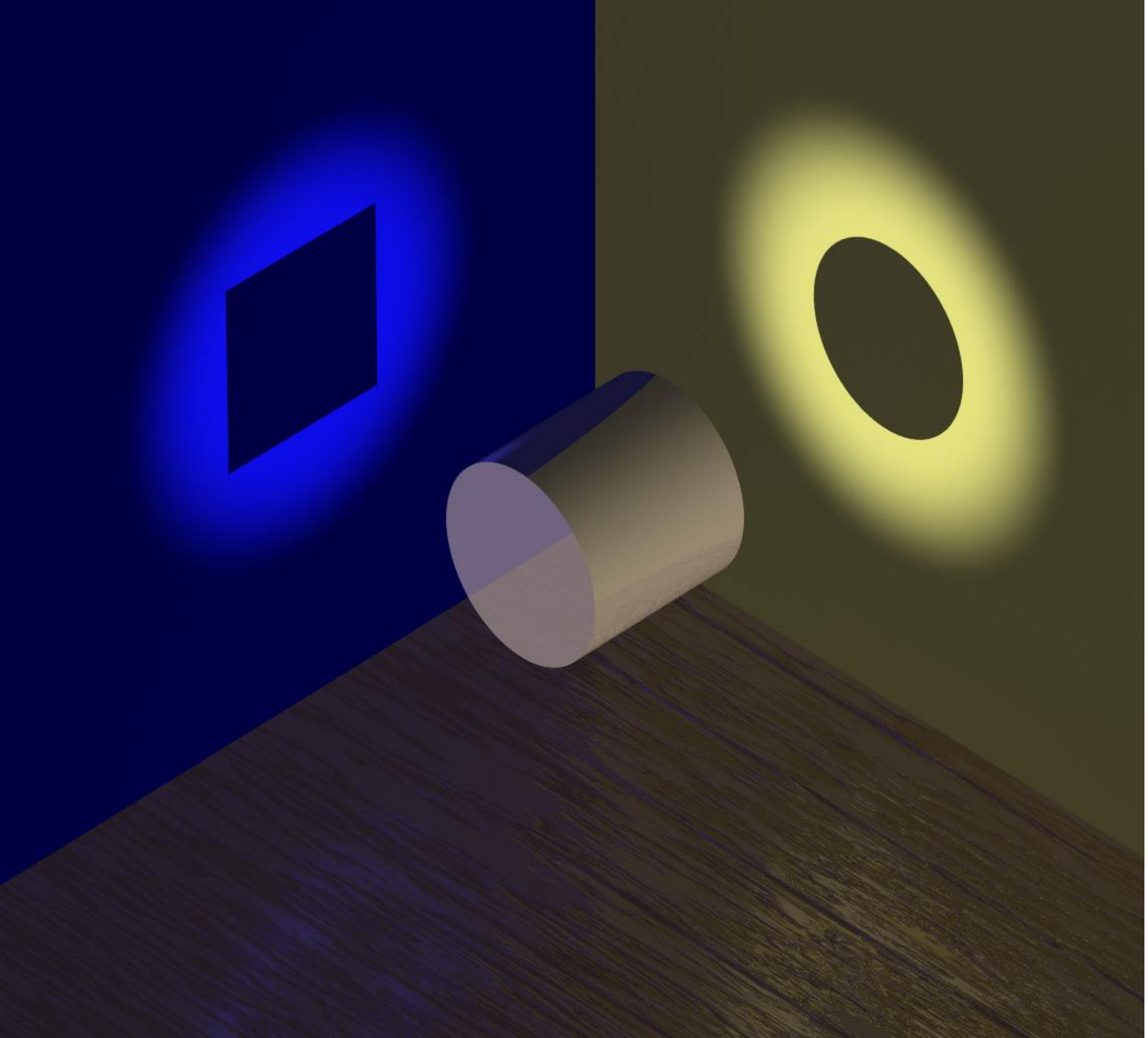
Classifying wines: terroir (terrain) or cépage (grape)?

Mary Douglas: how ‘institutions do the classifying’.

1) Changes in the job classification in the the late 18^{th}, when guilds were the dominant institution, to mid 19^{th} century, under the influence of free trade institutions after the industrial revolution.

2) Wine classification shifted from territory-based classification in France to grape types in California (i.e. *terroir* vs. *cépage)*, after the internationalisation of wine markets in the late 20^{th} century meant that buyers lost the detailed territorial context.

https://www.ivinio.com/fr/cepage-ou-terroir-/



From ‘maps of science’, to ‘mappings of science’

according to particular logics

- The analyst can decide which dimensions of science she wants to make visible by choosing to map through particular characteristics of some specific units of analysis.

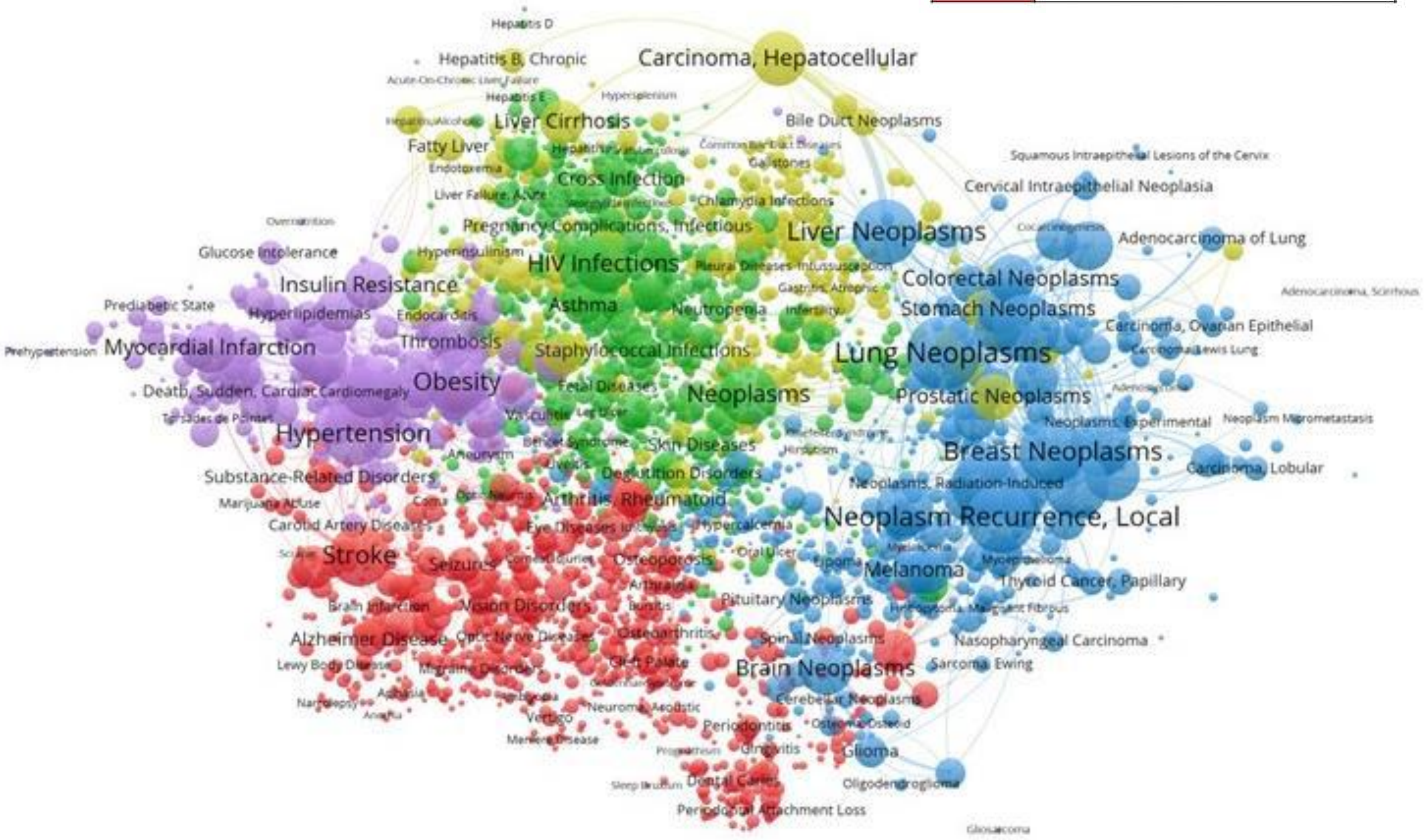
- The key questions then are about which logics, characteristics and relations are relevant for particular analyses?

- The answer to these questions will determine which ontologies and classification to use, which epistemic spaces to generate.

Map of diseases

Coburn et al. (2025)

Colour	Disease group
	Cardiometabolic diseases
	Infectious diseases
	Digestive system diseases
	Cancers
	Neurological and dental diseases



Questions on how science evolves in relation to societal demands and needs may be better posed with ontologies expressing these demands and needs, rather than internalist ontologies (such as disciplinary).

Multi-layered Topic Modelling

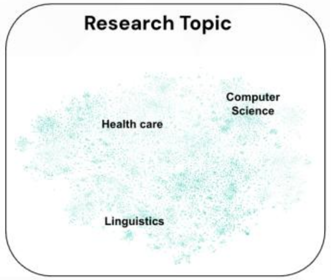
Berta Grimau and Nicolau Duran-Silva at SIRIS Academic (unpublished)

Corpus of scientific articles

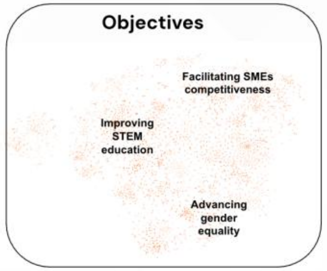
Step 1: Based on each title+abstract, generate three layers according to a specific logic with the help of an LLM.

Research topic	'The research topic is the role of neutrophil extracellular traps (NETs) in the progression of COVID-19.',
Methods	'Analyze autopsy results to identify lung infiltration of neutrophils.'
	'Review literature to link aberrant NET formation to pulmonary diseases and other COVID-19 related symptoms.'
Objectives	'Investigate the presence of neutrophil extracellular traps (NETs) in COVID-19 patients.'
	'Examine the potential role of NETs in organ damage and mortality in COVID-19.'

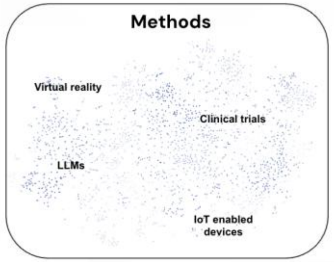
Step 2: For each layer, generate a topic model* to identify the basic pool of clusters.



The general research field to which the research belongs.

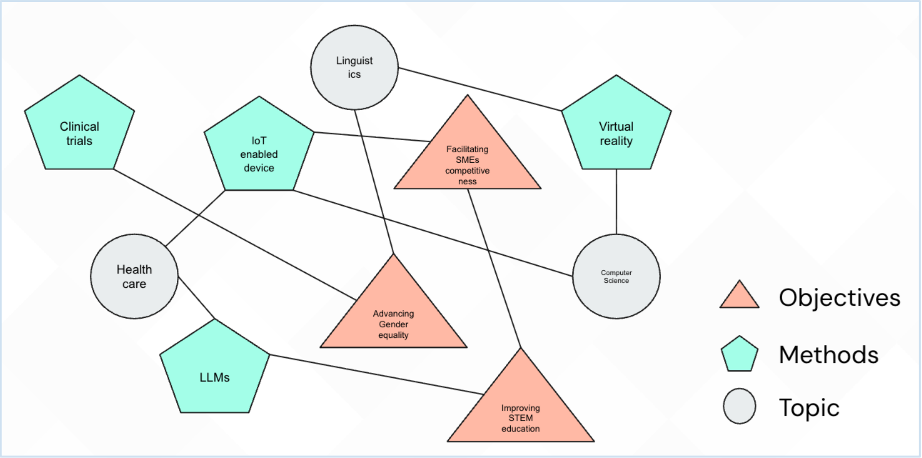


The objectives or expected impact of the research.



The methods and techniques employed to pursue the research.

Step 3: Find connections between the three layers to identify heterogeneous clusters.



*The topic models are performed by vectorizing title+abstracts with the BERT-based model Specter2 and the HDBSCAN clustering algorithm.

Summary

Towards multiple ontologies in science mapping

- There are a variety of logics in the classifications of scientific docs
  - Disciplinary (conventional)
  - Objects of study: diseases, animals,
  - Methods
  - Theories
  - Goals of research
- Many logics create crossing cats – heterogeneous networks if in same layer
- Shifts in science policy associated with changes in classifications
  - Instead of perspective of researchers, perspective of users/beneficiaries
  - Standpoint theory: 'Strong objectivity' recognises that different viewpoints are legitimate. Making explicit the viewpoint in the construction of classification.

CWTS

Summary

- bottom-up algorithmic approaches, lack conceptual consistency regarding the type of categories used
  - in a same classification a category captures methods, another one has materials, a third one contains empirical objects and a fourth is focused on theories instead of having a single logic

- more useful representations by using ontologies based on a single logic that aligns with the particular conceptual needs of the analysis.
- novel classification methods based on machine learning and language models hold promise to produce these tailored, question-driven ontologies.

CWTS

Mapping methodology based on bibliometrics

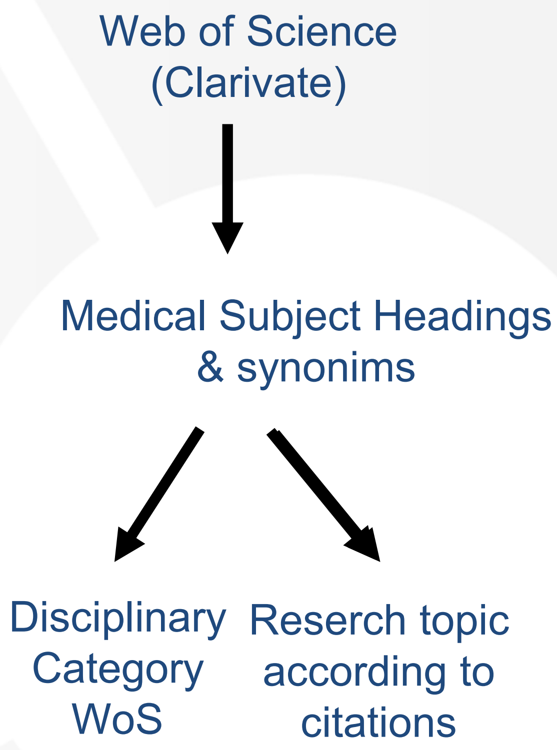
- Which description (problematic choices!)?
  - **Coverage: Which database?**
    Web of Science (Bias towards biomedicine, English).
  - **Definition: How are publications selected?**
    Using classification and synonims of PubMed:
    Medical Subject Headings (Bias towards biomedicine!).
  - **Classification: How are publicacions classified?**

    Two approaches

    1) Large disciplines: groupings of Web of Science Categories.

    2) Research Micro-areas: publications topics according to citation clusters

    (*Both problematic → trying to check robustness of insights)*



Questionable choices → Problematic mapping → Only exploratory!!