



coreon

Knowledge meets language.

# ML-powered Taxonomization: AI Lends Taxonomists a Hand

Alena Vasilevich  
Computational Linguist@Coreon



[alena@coreon.com](mailto:alena@coreon.com)



<https://www.linkedin.com/company/coreon-gmbh>

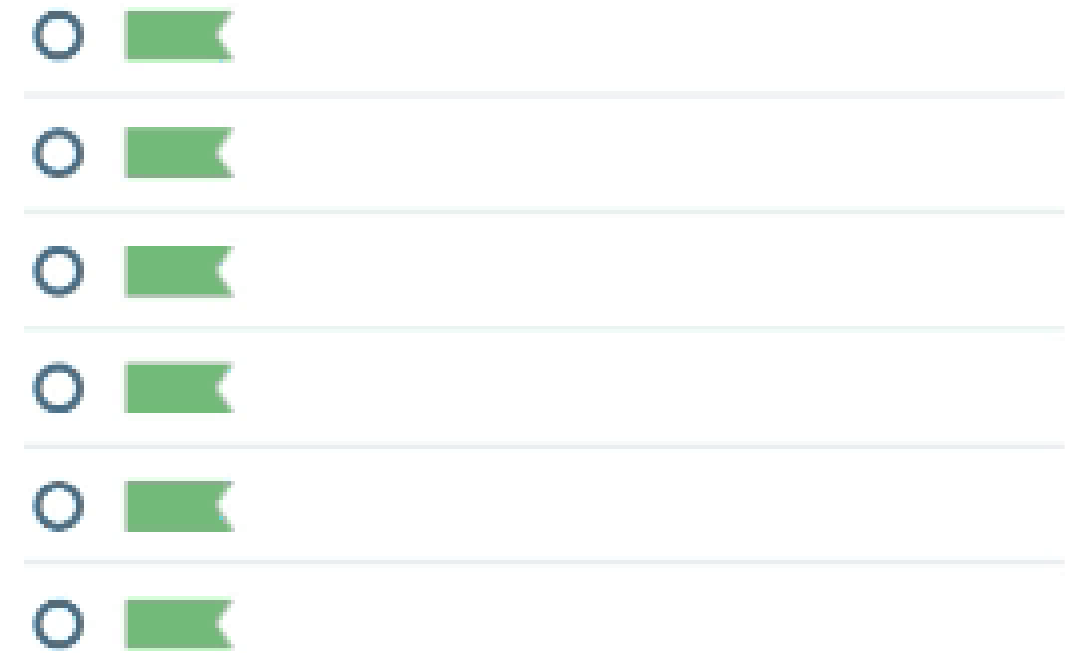


connecteddata**london**

# Coreon: Backbone for the Data



- ⌘ convert data into knowledge
- ⌘ incorporate **terminologies**  
**taxonomies**  
**ontologies**  
**thesauri**  
**vocabularies**  
into one **Knowledge Graph**
- ⌘ concept-oriented and language-agnostic data model
- ⌘ intuitive, lightweight **UI**



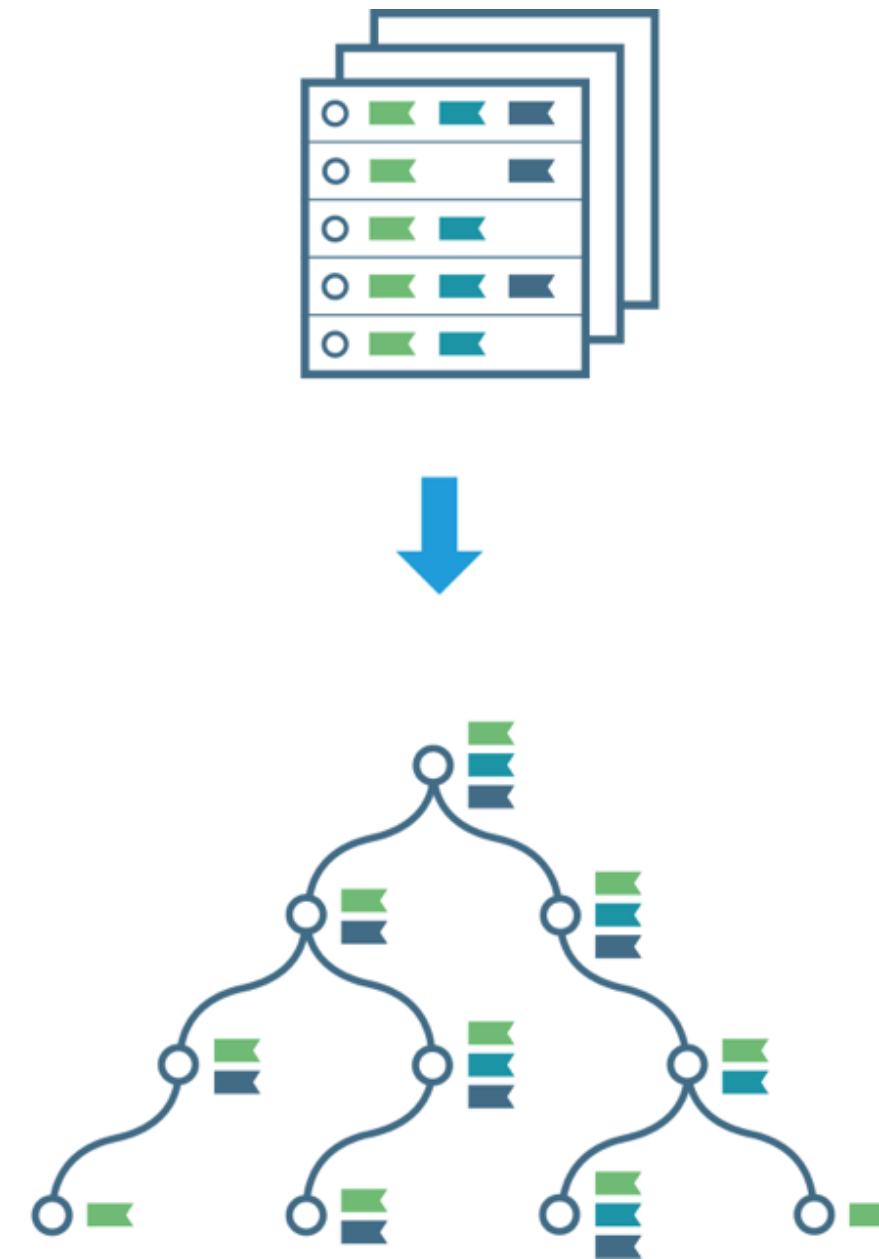
# Agenda



- ⌘ Structured data and IATE as a resource
- ⌘ Manual Taxonomization
- ⌘ Automatic Taxonomization
- ⌘ Collaborative-AI Approach
- ⌘ Industry use case: topic classification

# Perks of Structured Data

- ⌘ A Powerful Resource for AI/ML projects
- ⌘ Cross-lingual Data Analysis
- ⌘ Enterprise Search
- ⌘ Actionable intelligence
- ⌘ Cross-border Interoperability



# Interactive Terminology for Europe, IATE

- ⌘ Introduced in 2004, used by most EU Institutions, covers all EU domains
- ⌘ Recent focus on healthcare, financial crisis, environment, fisheries, and migration
- ⌘ EuroVoc for domain classification system



- ⌘ Number of concepts: **961 116**
- ⌘ Number of terms: 7 992 325
- ⌘ New terms last week: 1 646

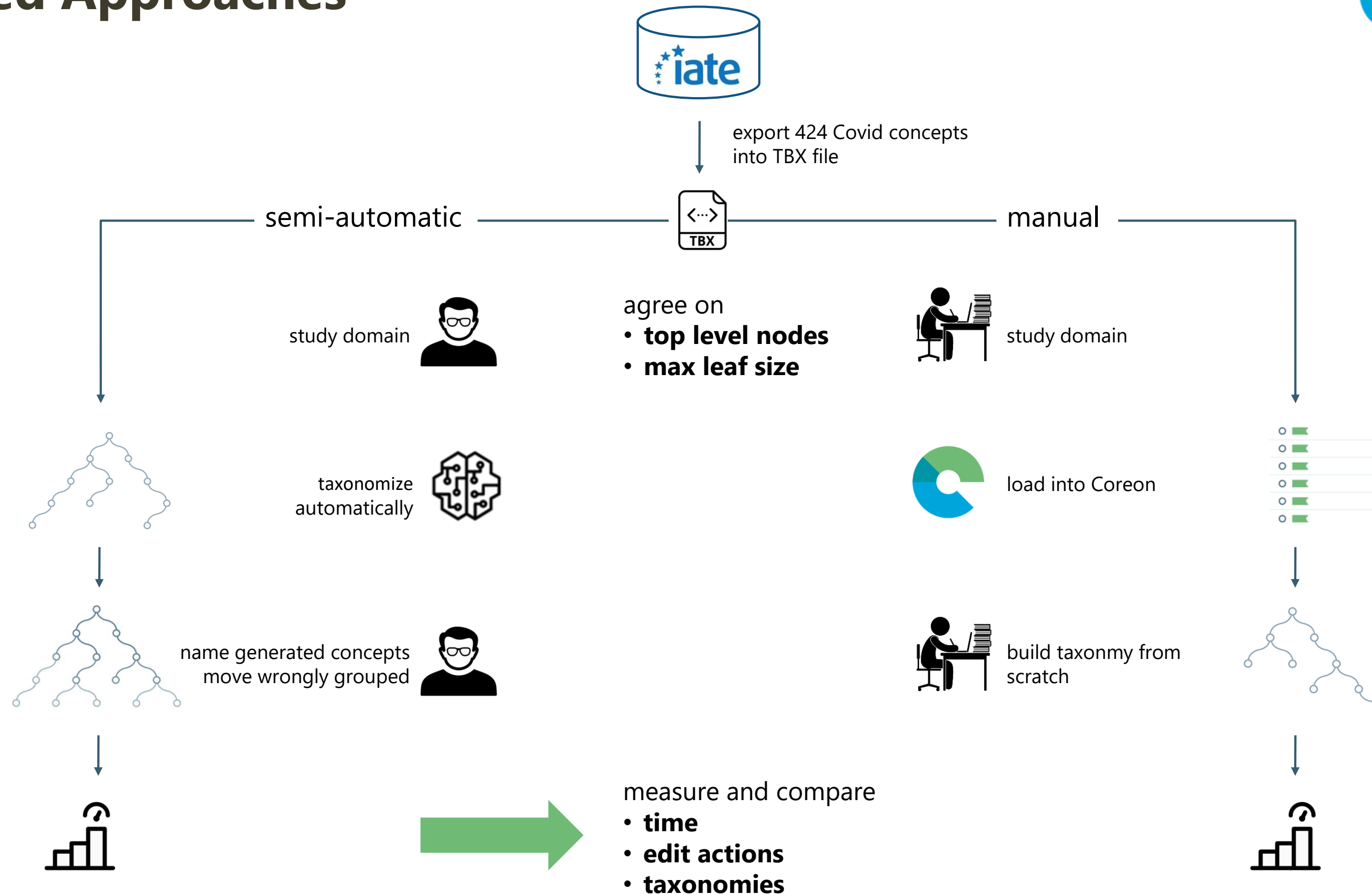


**TODO:** draft a deeply-structured taxonomy, **\_fast\_** 

**INPUT:** a flat set of COVID concepts, no relations between them

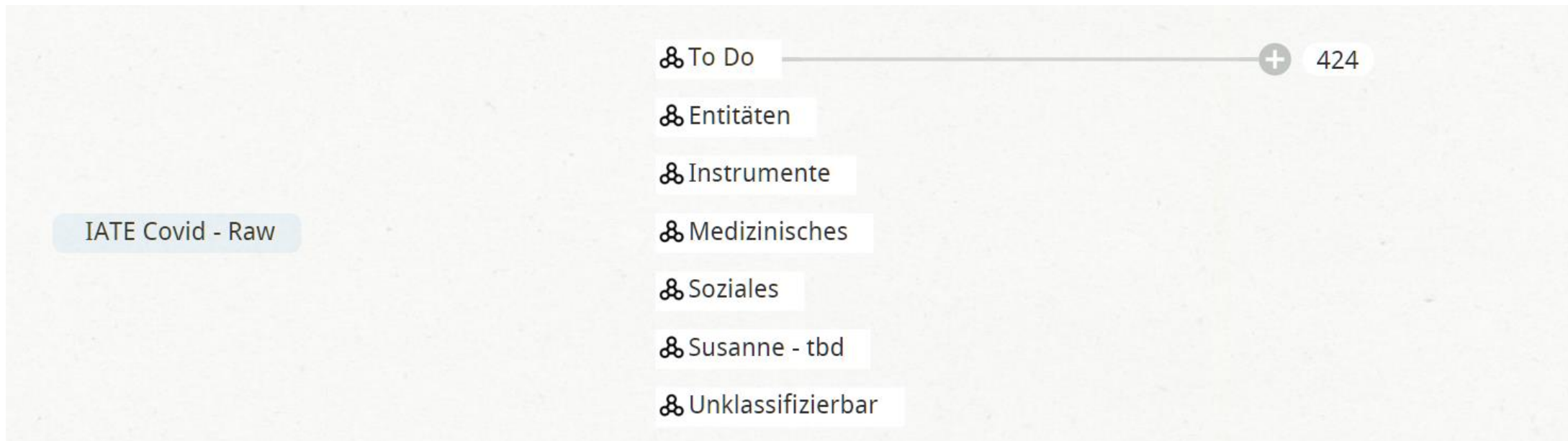
**OUTPUT:** a hierarchical knowledge graph,  
consumable via REST/SPARQL

# Two Tested Approaches

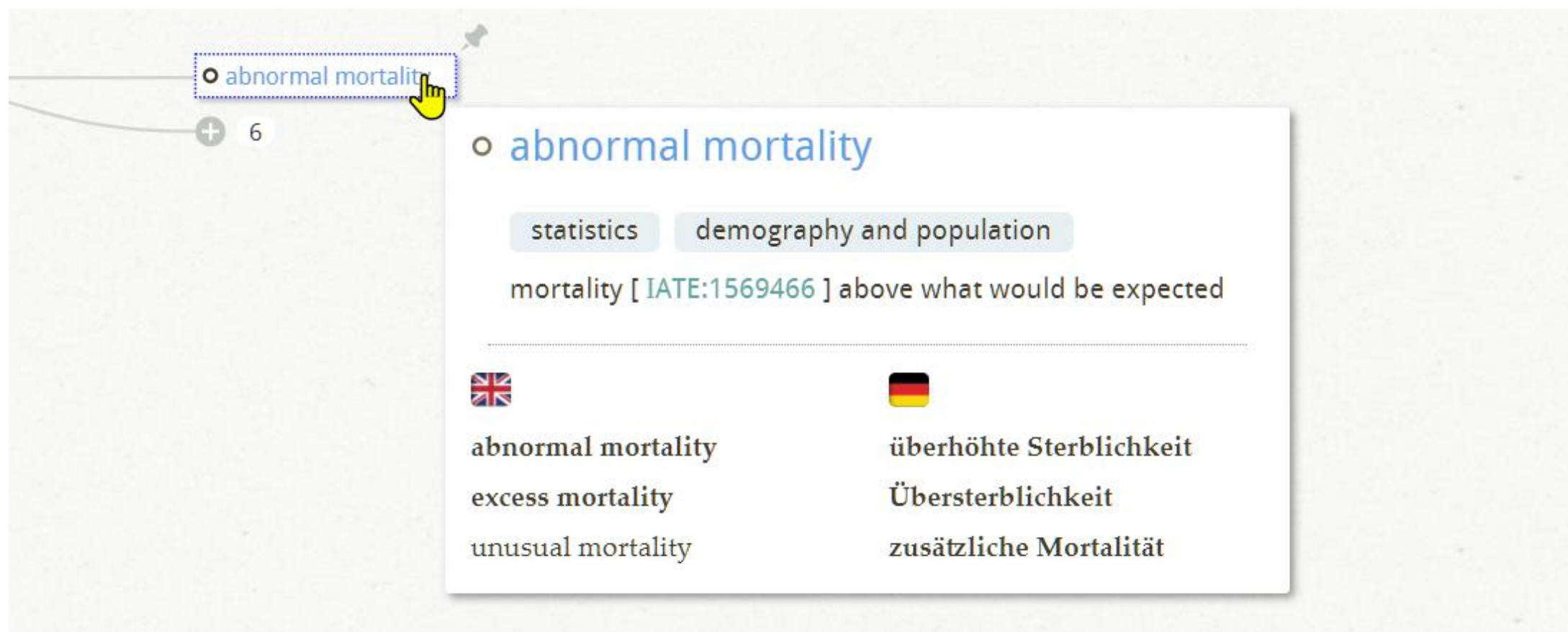




# Manual Taxonomization



- ▶ Top level nodes, temporary helper buckets, and lots to do...

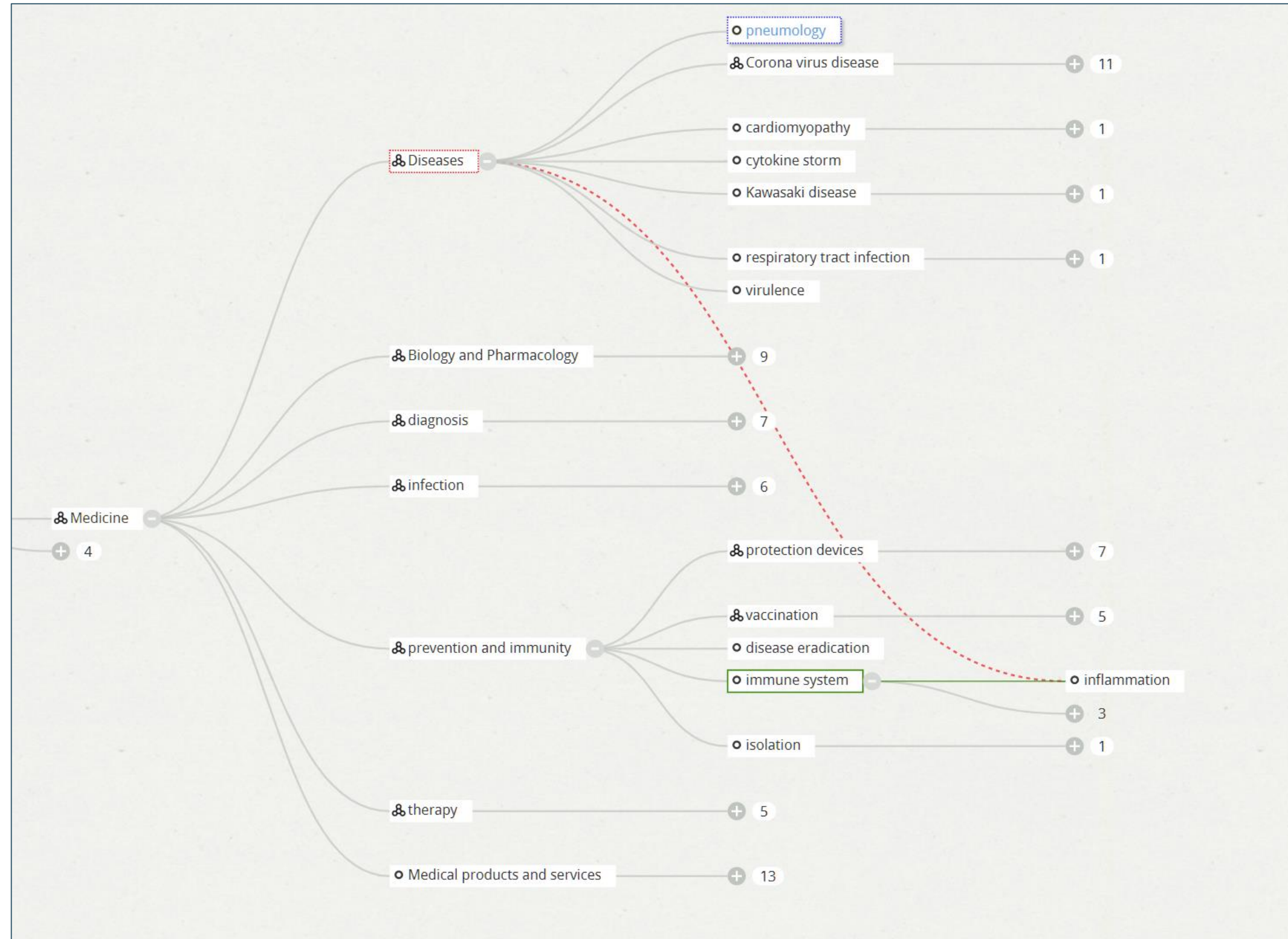


- ▶ Concept card displaying important metadata




# Manual Taxonomization: Editing Actions

- ⌘ drag'n'drop
- ⌘ pin
- ⌘ filter

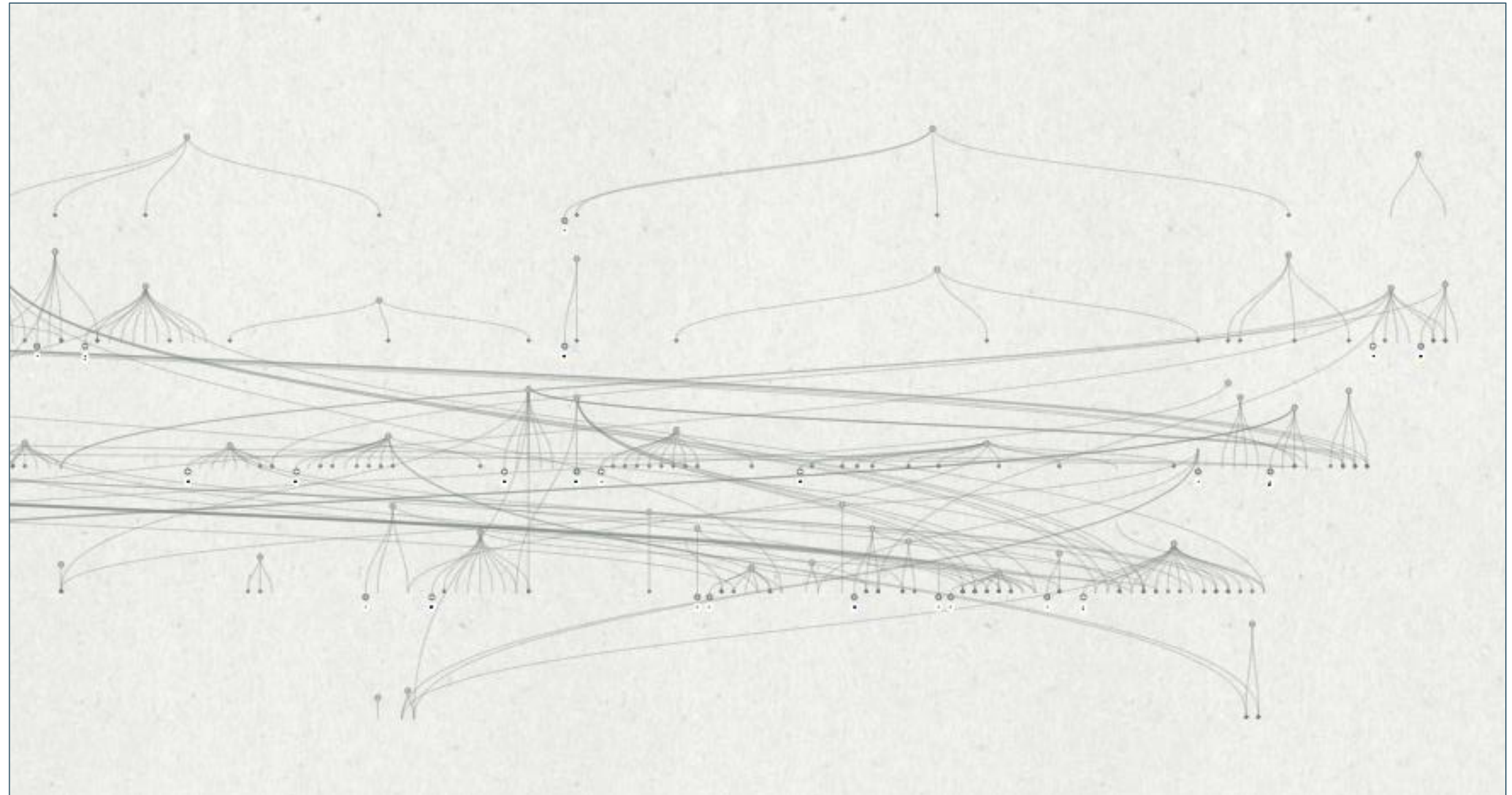


dragging 'inflammation'  
from 'diseases'  
to 'immune system'

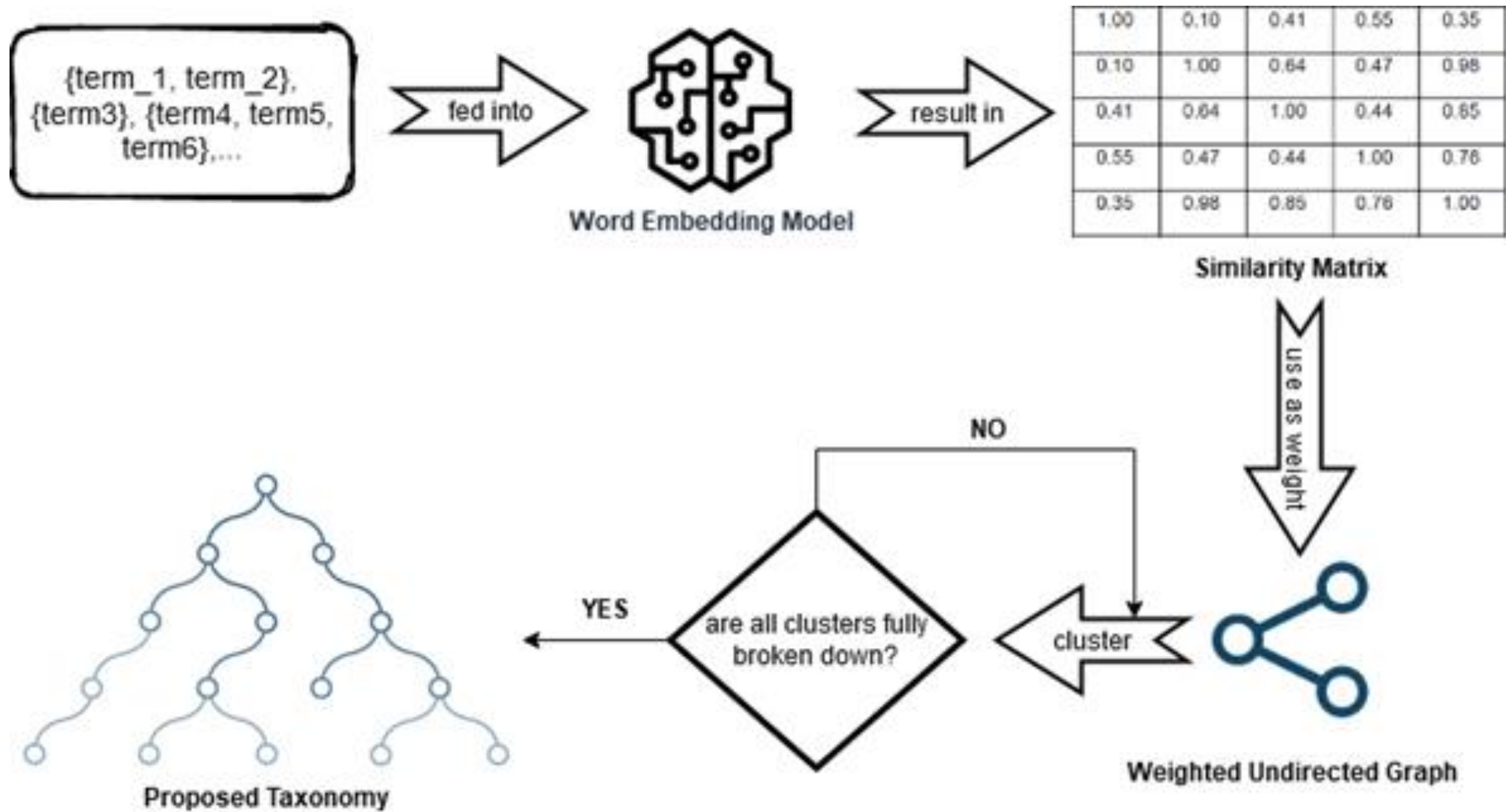
# Manual Taxonomization: Result

 load into Coreon

 build taxonomy  
**from scratch**

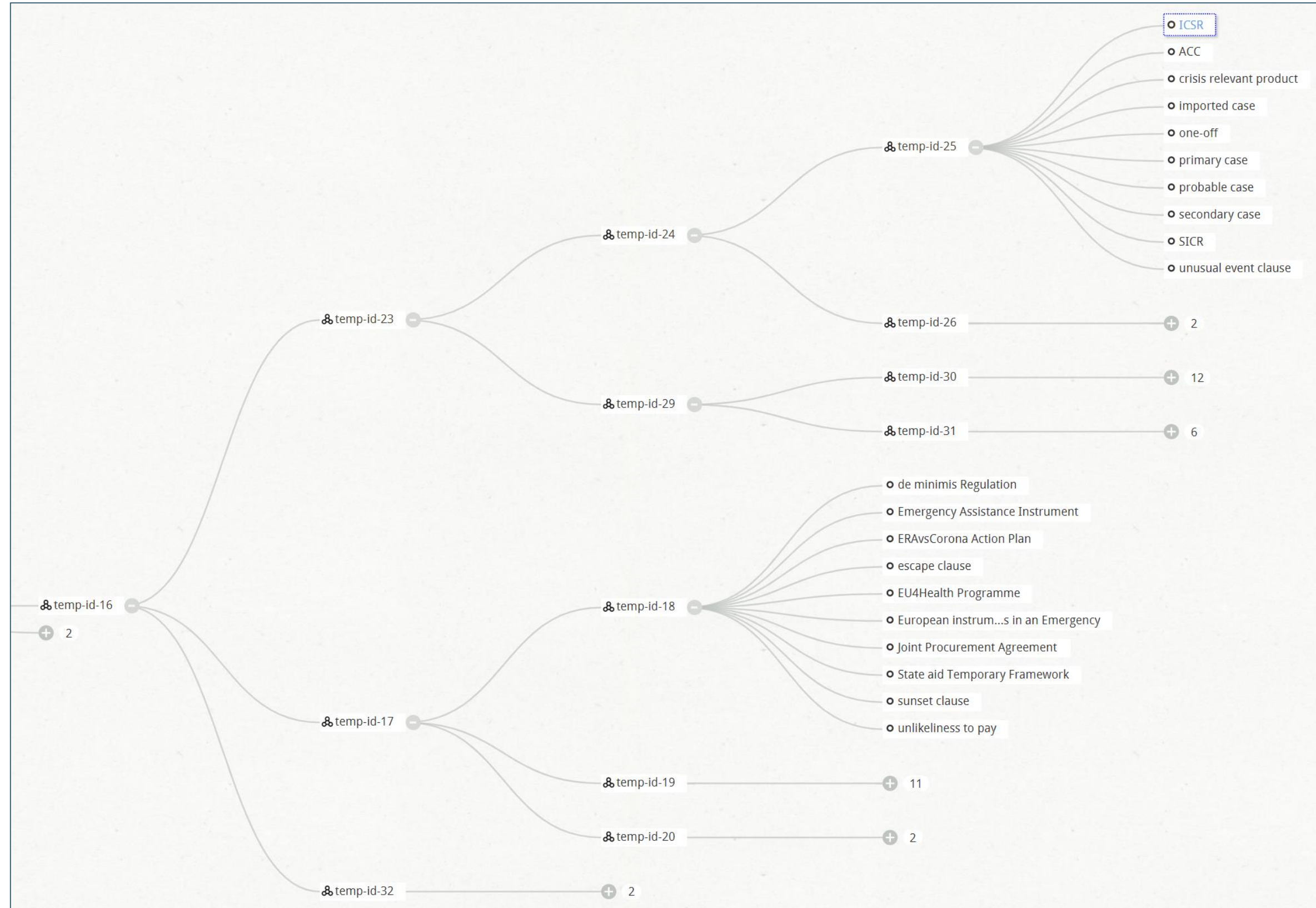


# Auto-Taxonomization: Data + WE + Community Detection Algorithm





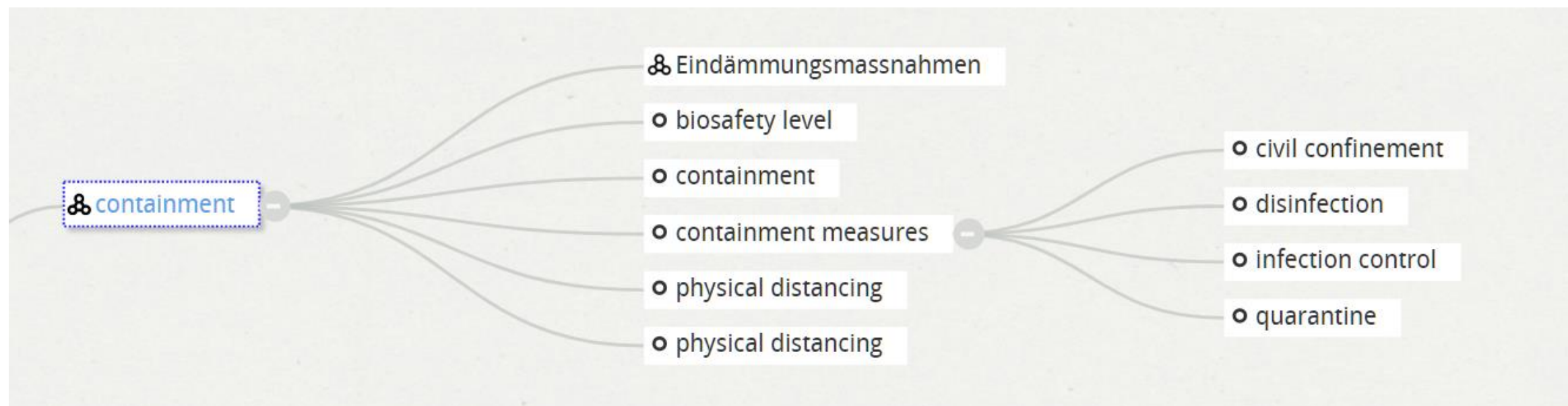
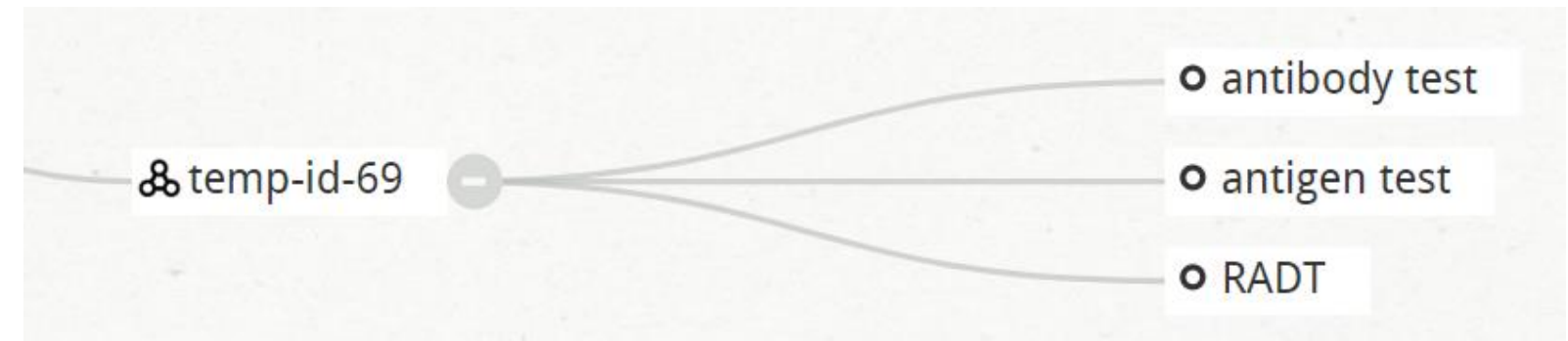
# Human Revision of AI-Drafted Taxonomy



initial situation  
after automatic  
taxonomization

# Good Clustering, Bad Clustering?

- ▶ 55 clusters, majority pretty accurate
- ▶ some clusters are off, and we blame **WE**:
  - ▶ ‘interstitial space’ and ‘hospital pharmacy’
  - ▶ spaces appearing in similar “semantic neighborhoods”
- ▶ some existing IATE concepts became parents of concept clusters



# Collaborative Taxonomization: Result



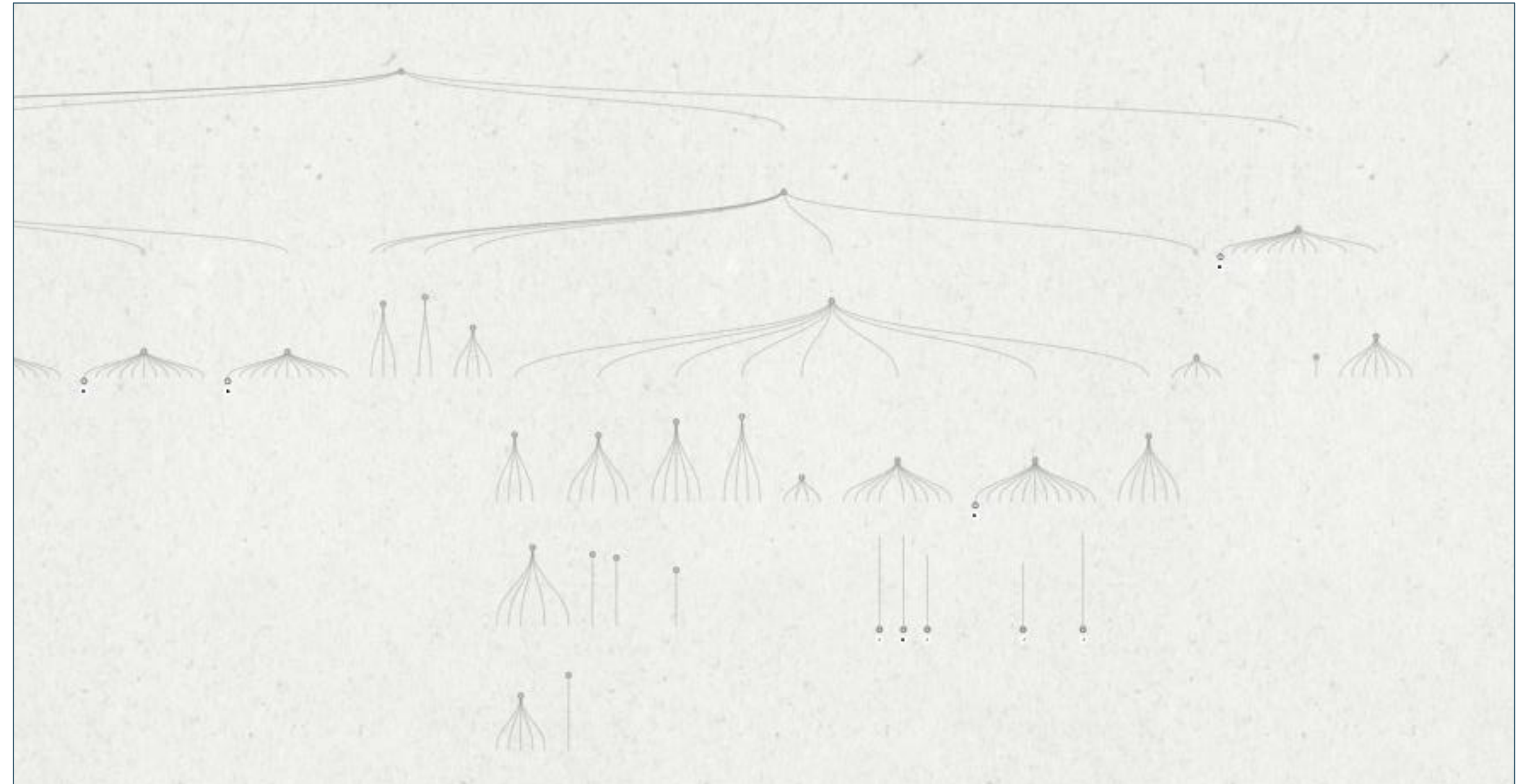
**automatic**  
taxonomization using  
ML algorithm



load into Coreon



name auto concepts  
move wrong concepts





Metrics	Taxonomization	
	Manual	Semi-Automatic
Curator's recorded time (hours)	<b>40</b>	8
Relations created / changed	<b>1 147</b>	432
Concepts created	115	28
Intermediate structural nodes renamed	—	45
Overall relations	679	470

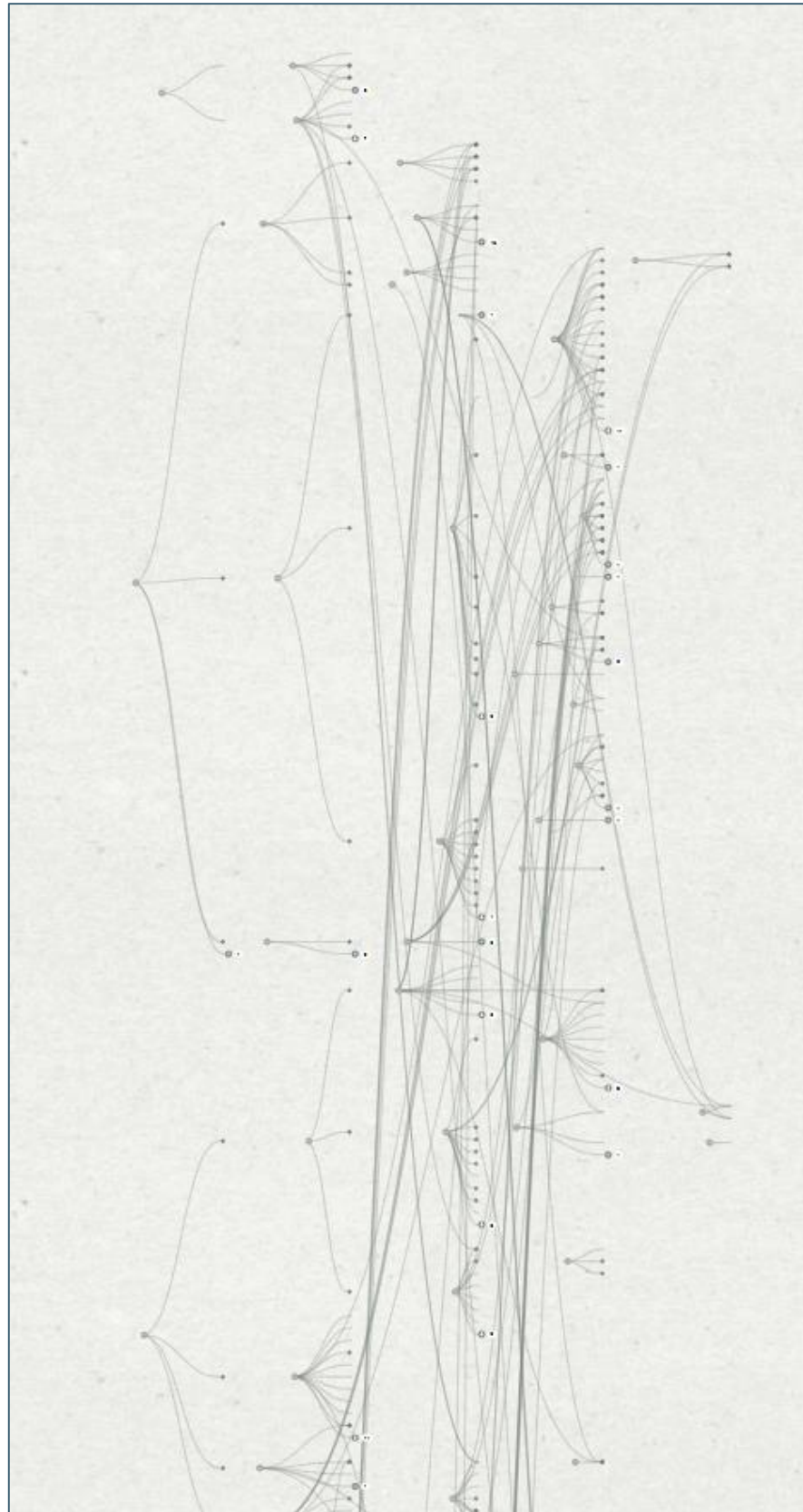
# Resulting Taxonomies



load into Coreon



build taxonomy  
from scratch



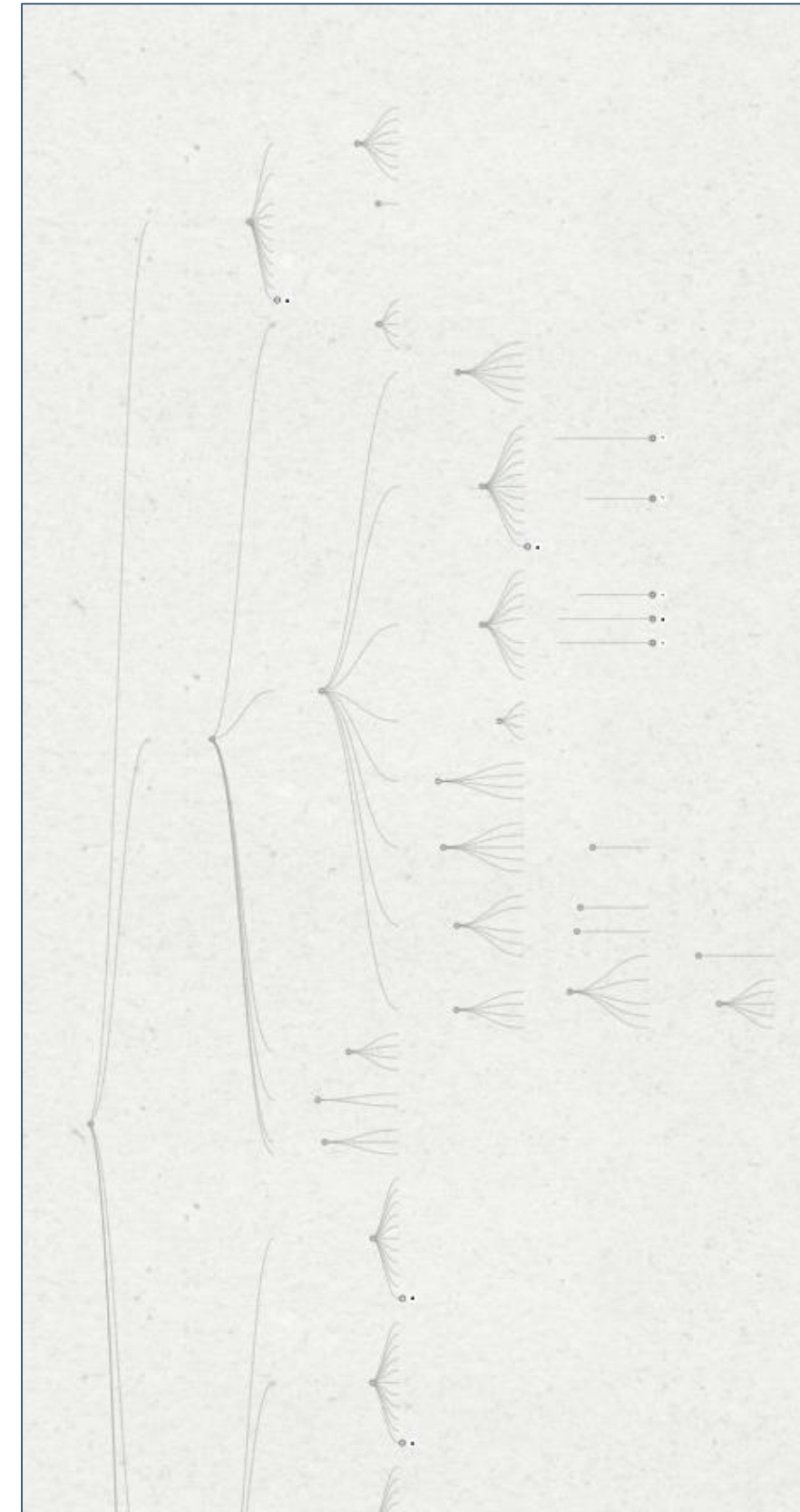
**automatic**  
taxonomization using  
ML algorithm



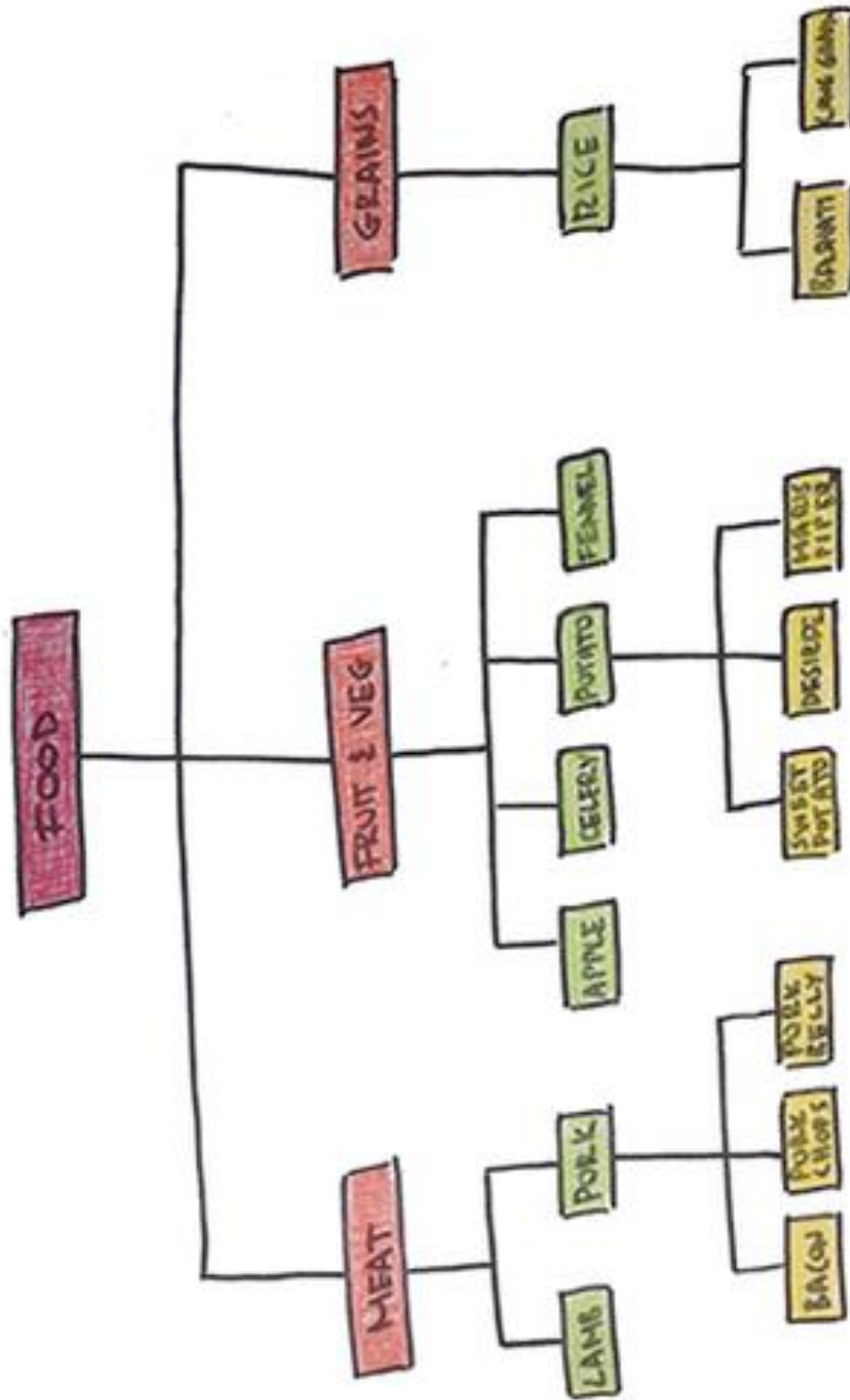
load into Coreon



name auto concepts  
move wrong concepts



# Taxonomization Benefits

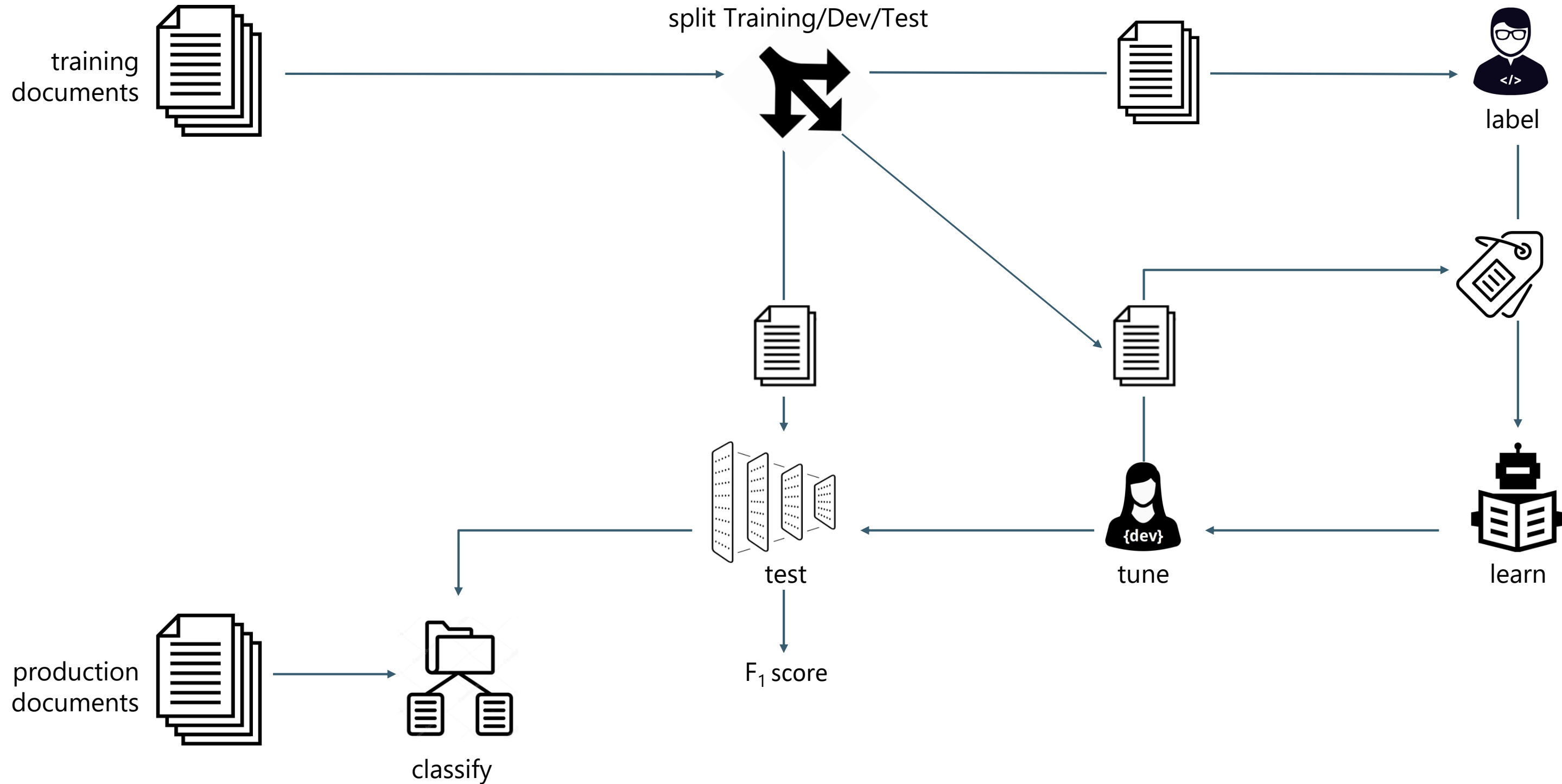


- ⌘ Effective way to add structure to data
- ⌘ Improve data quality
  - ⌘ avoid duplicates and overlapping concepts
  - ⌘ associative relations
- ⌘ Easier and safer data maintenance
- ⌘ Formalize multilingual knowledge, make it machine-digestible
- ⌘ Boost performance of AI algorithms, priming them with structured data

# Text Classification

1500 / 300 / 800



split Training/Dev/Test





# CNN Predictions and True Labels / Test Set



	True labels 	Predicted labels 
<p>The spread of COVID-19 displayed a characteristic sub - exponential linear growth ( mean DGR : 0.06 ; range : 0.01 , for Kerala , to 0.11 , for Gujarat , 95 % CI : 0.04 - 0.08 , 11 April 2020 to 3 May 2020 ) that deviates from exponential growth estimates ( mean DGR : 0.1643 ; range : 0.1163 , for Kerala , to 0.2175 , for Tamil Nadu , 95 % CI : 0.1392 - 0.1894 , 2 March 2020 to 3 May 2020 ) , as a consequence of lockdown strategies ( ) . Reduction of R , from a range of 1.35 - 2.86 ( pre - lockdown 95 % CI : 1.49 - 2.21 ) to 1.13 - 1.67 , in first phase , and to 1.08 - 1.63 , in second phase of lockdowns ( post - lockdown 95 % CI : 1.20 - 1.42 )...</p>	health management, epidemiology	medicine
<p>"In late 2019 , a cluster of severe respiratory disease cases occurred in Wuhan , China ( Hubei Daily ( Reporter Yu Jinyi ) , 2019 ; World Health Organization , 2020c ) . By January 2020 , a novel coronavirus was sequenced , human - to - human transmission was clearly documented ( Centers for Disease Control and Prevention , 2020 ; , and additional cases identified in other countries caused by the same virus now known as SARS - CoV-2 , resulting in the identification of a new disease syndrome , . As the virus continued to spread , resources were deployed rapidly to identify and develop prophylactic and therapeutic agents for patients . However , these initial activities were uncoordinated leading to haphazard testing of non - optimal or potentially harmful interventions . One reason for the lack of coordination was the lack of an infrastructure and well vetted guidelines to facilitate optimized drug testing...</p>	medicine, medical products and services	medicine, medical products and services
<p>The number of the novel coronavirus disease 2019 ( COVID-19 ) cases and mortality grow rapidly around the world . Aged individuals with preexisting medical conditions are at increased risk for severe acute respiratory syndrome coronavirus 2 ( SARS - CoV-2 ) infection , with disease and mortality occurring predominantly in this group . It is thus conceivable that patients with MS or neuromyelitis optica spectrum disorders ( NMOSDs ) who are receiving disease - modifying treatments have increased susceptibility to COVID-19 infection and disease . However , evidence supporting this notion is currently lacking . We have conducted a survey through the Chinese Medical Network for Neuroinflammation ( CMNN ) ...</p>	medicine, diseases	medicine, diseases

Classifiers	Metrics (micro-averages), %		
	Precision	Recall	F-1
Non-initialized CNN	81.6	75.4	78.4
<b>Initialized CNN</b>	<b>82.5</b>	<b>78.8</b>	<b>80.6</b>
🙄 Zero-shot 0.95 threshold	15.3	37.8	21.7
🙄 Zero-shot 0.97 threshold	15.0	26.3	19.1
🙄 Zero-shot 0.99 threshold	12.0	10.2	11.0



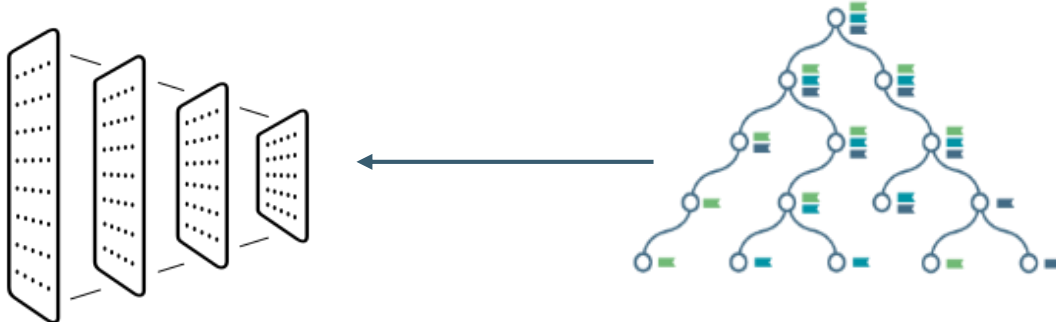
# Taxonomized Data to Enhance AI Performance



Label documents automatically



Boost CNN with taxonomy



Enjoy finer granularity in document classification



Get multilingual for free





# coreon

Knowledge meets language.

## Thank You!



@coreonapp  
@lennyvasilevich



[alena@coreon.com](mailto:alena@coreon.com)



<https://www.linkedin.com/in/alenasvasilevich>



connecteddata**london**