holi food

Funded by the European Union

BfR
German Federal Institute for Risk Assessment

Enhancing Regulatory Foresight through Weak Signal Mining: An AI-based Micro-Topic Emergence Scoring Approach

Axel Menning[1], Zsuzsa Farkas[2], Krisztián Vribék[2], Bas H.M. van der Velden[3]

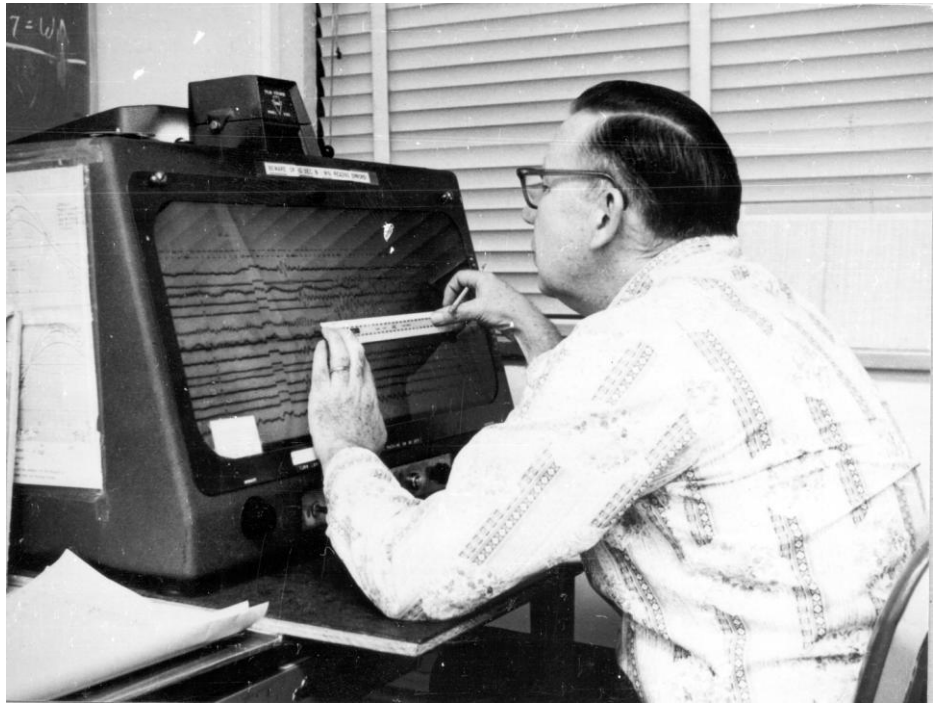
[1] German Federal Institute for Risk Assessment

[2] University of Veterinary Medicine Budapest

[3] Wageningen Food Safety Research, Wageningen, The Netherlands

Weak Signal Miner (WSM)

A seismograph records amplitude over time. The WSM records semantic momentum over time. A food and feed micro-topic becomes a potential emerging risk only if it displays an anomalous surge in magnitude and recency.

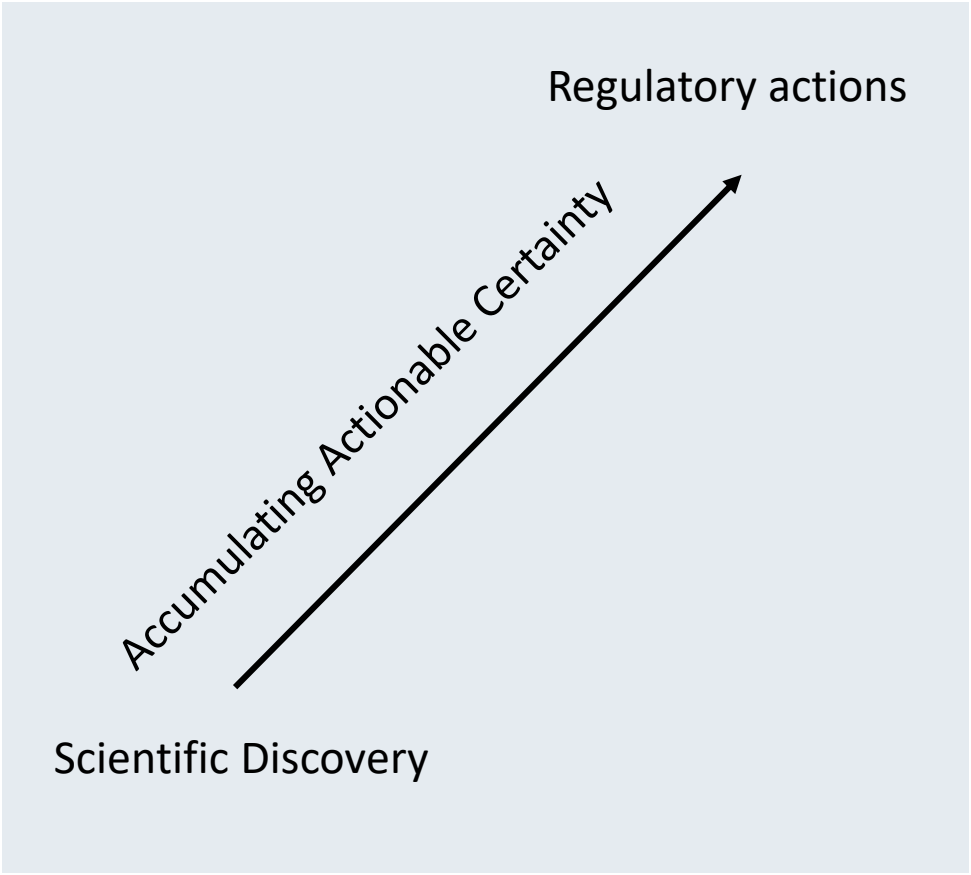


https://en.wikipedia.org/wiki/File:Viewing_of_Develocorder_Film.jpg Photo by R. McKenzie, pre-1977. Page 10 (upper photo), Earthquake Information Bulletin, v.9, no.3.

The Latency Gap in Emerging Risk Detection

Weak signals circulate for years before reaching institutional articulation.

Manual synthesis is no longer scalable for modern complex global supply chains.



Locating the weak signal miner in the risk analysis process

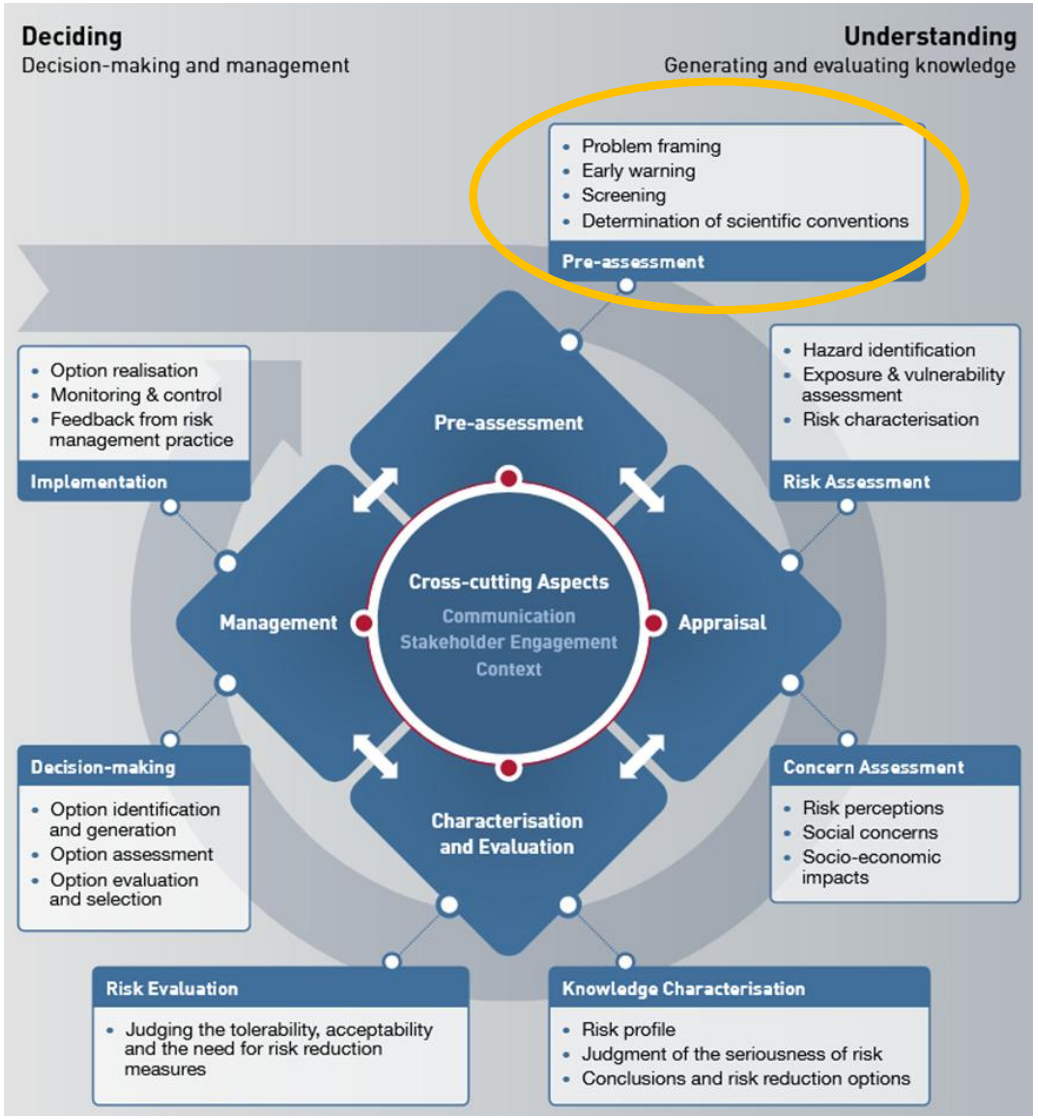


Image Source: IRGC.org; risk governance framework

Weak Signals: Signs or indicators of possible changes

Origins: „Managing strategic surprise by response to weak signals“ (Ansoff, 1975)

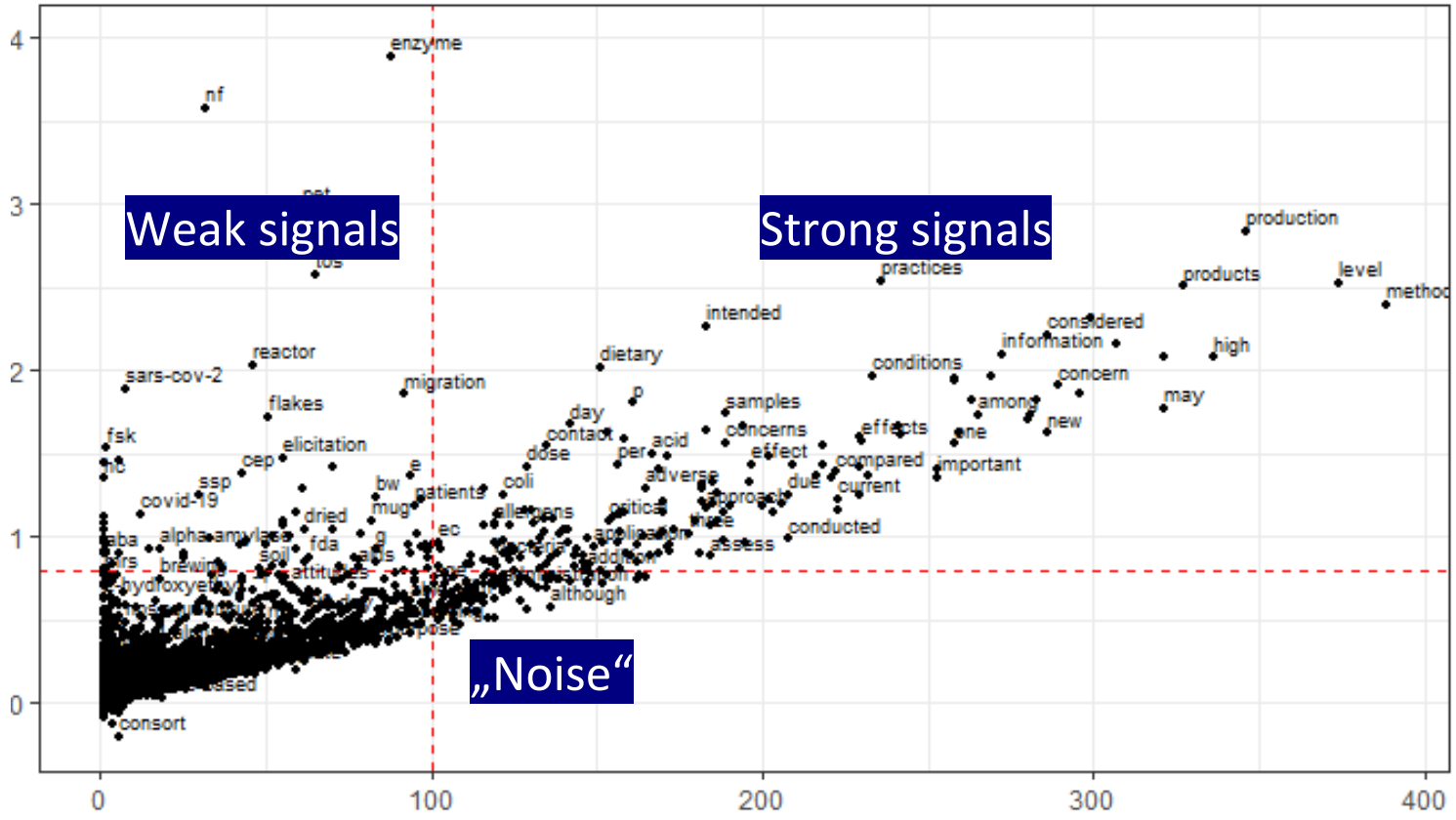
- Strategic surprises rarely occur without warning. Instead, they are preceded by early, vague, and fragmented pieces of information: "weak signals.“
- If an organization waits until a signal becomes "strong" before taking action, it will likely be too late to mount a proactive response. Organization will be forced into crisis management.

Operations: “Detecting weak signals for long-term business opportunities using text mining of Web news” (Yoon, 2012)

- Weak signals: terms that have a low frequency of occurrence but an above-average rate of increase over time.
- Weak signal mining is a text mining technique to distinguish emerging concepts (weak signals) from well-known concepts (strong signals) and non-evolving concepts (noise).

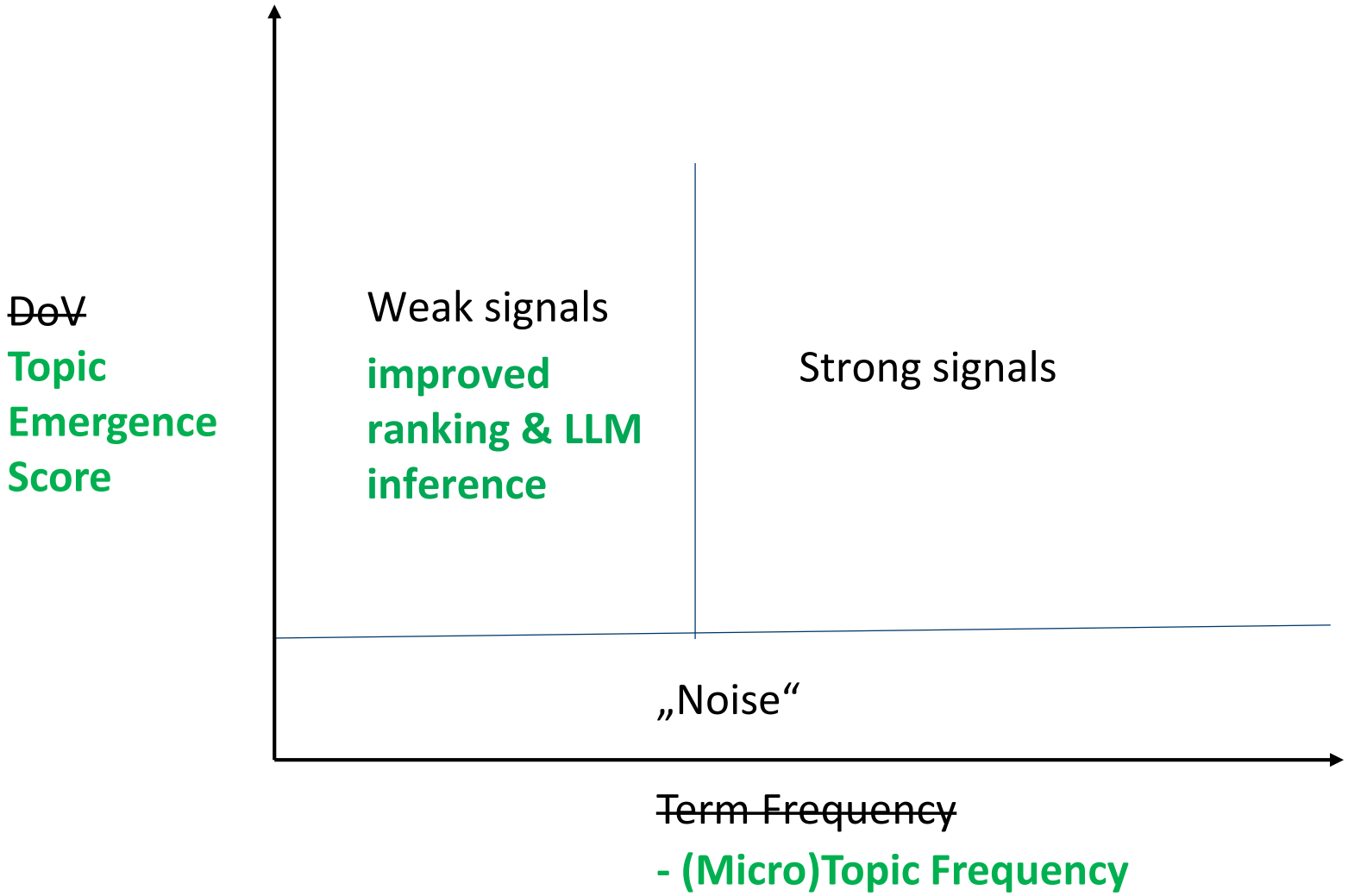
Keyword Emergence Matrix

Degree of Visibility

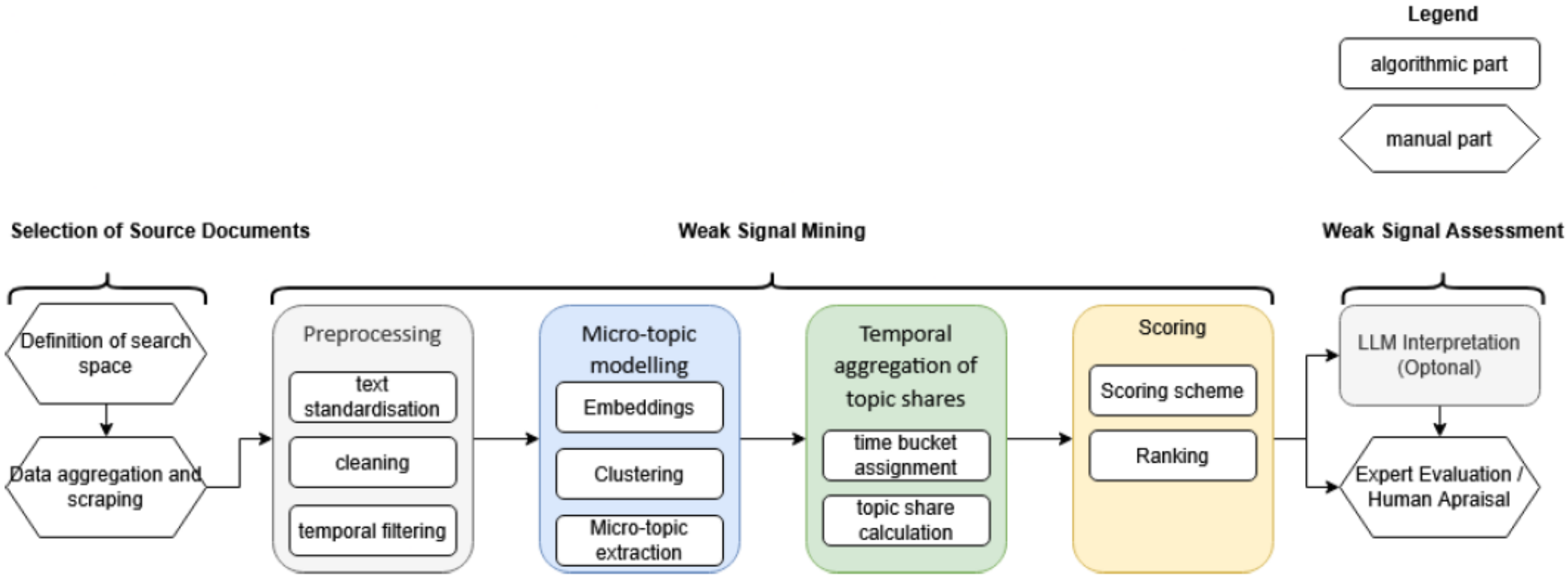


Term Frequency

Areas of improvement for Weak Signal Detection



Weak Signal Mining Workflow



Temporal aggregation and normalization

Time bucket parameter defines the „refresh rate“ of our regulatory radar.

Parameter tuning depends on:

- input data (social media versus academic literature)
- kind of risk (acute risk vs systemic shifts)

Options:

- Auto = selects the finest granularity with sufficient density
- Manual options: Y, HY, Q, M, W, D
- Count topic occurrences per bucket
- Compute topic shares per bucket
- Topic-by-time dataframe

OUTPUT: Topic Share Matrix S				
	Topic1	Topic2	Topic3	...
Bucket 1	0.67	0.33	0.00	...
Bucket 2	0.00	0.50	0.50	...
Bucket 3	1.00	0.00	0.00	...
Bucket 4	0.50	0.50	0.00	...

Topic emergence score

Recency score $\mathbf{R_k}$:
How active /"warm" is the discussion on topic k?

- T: Most recent time bucket
- $\mathrm{f}_{k,t}$: Frequency of topic k in time bucket t
- $e^{-\Delta(T-t)}$: exponential decay function

$$\mathrm{R}_k = \sum_{t=1}^{T} (\mathrm{f}_{k,t} \cdot e^{-\Delta(T-t)}$$

Magnitude Score $\mathbf{M_k}$:
Flags sudden, unnatural spikes, regardless how small the absolute numbers are.

- $f_{k,T}$: The frequency of the topic in the current time bucket
- $f_{k,T-1}$: The frequency of the topic in the previous time bucket

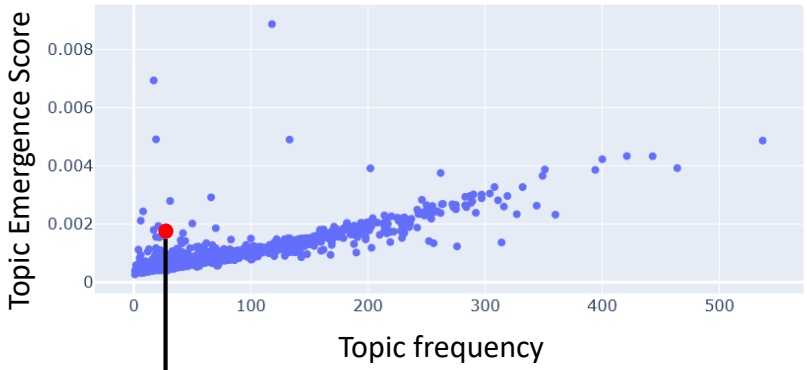
$$\mathrm{M}_k = f_{k,T} - f_{k,T-1}$$

Topic Emergence Score $\boldsymbol{C_k}$:
Blends normalized $\mathbf{R_k}$ and $\mathbf{M_k}$

- α: The weighting parameter
  - 0.5: R and M are treated equally

$$C_k = \alpha \cdot R_k^{norm} + (1 - \alpha) \cdot M_k^{norm}$$

Reading Weak Signals



LLM Option:

- Topic representations are sent to LLM
- LLM infers
  - Intuitive topic label
  - Novelty score + reasoning
  - Severity score + reasoning

Story instead of keywords:
['cruzi', 'chagas', 'trypanosoma', 'triatomines', 'triatomine', 'acai', 'oral', 'acd', 'attalea', 'trypomastigotes']

Always human in the loop

Requires

- domain knowledge
- imagination

Potential emerging issue:
Further deforestation of Attalea palms can lead to an increase in chagas disease.

Risk Assessment

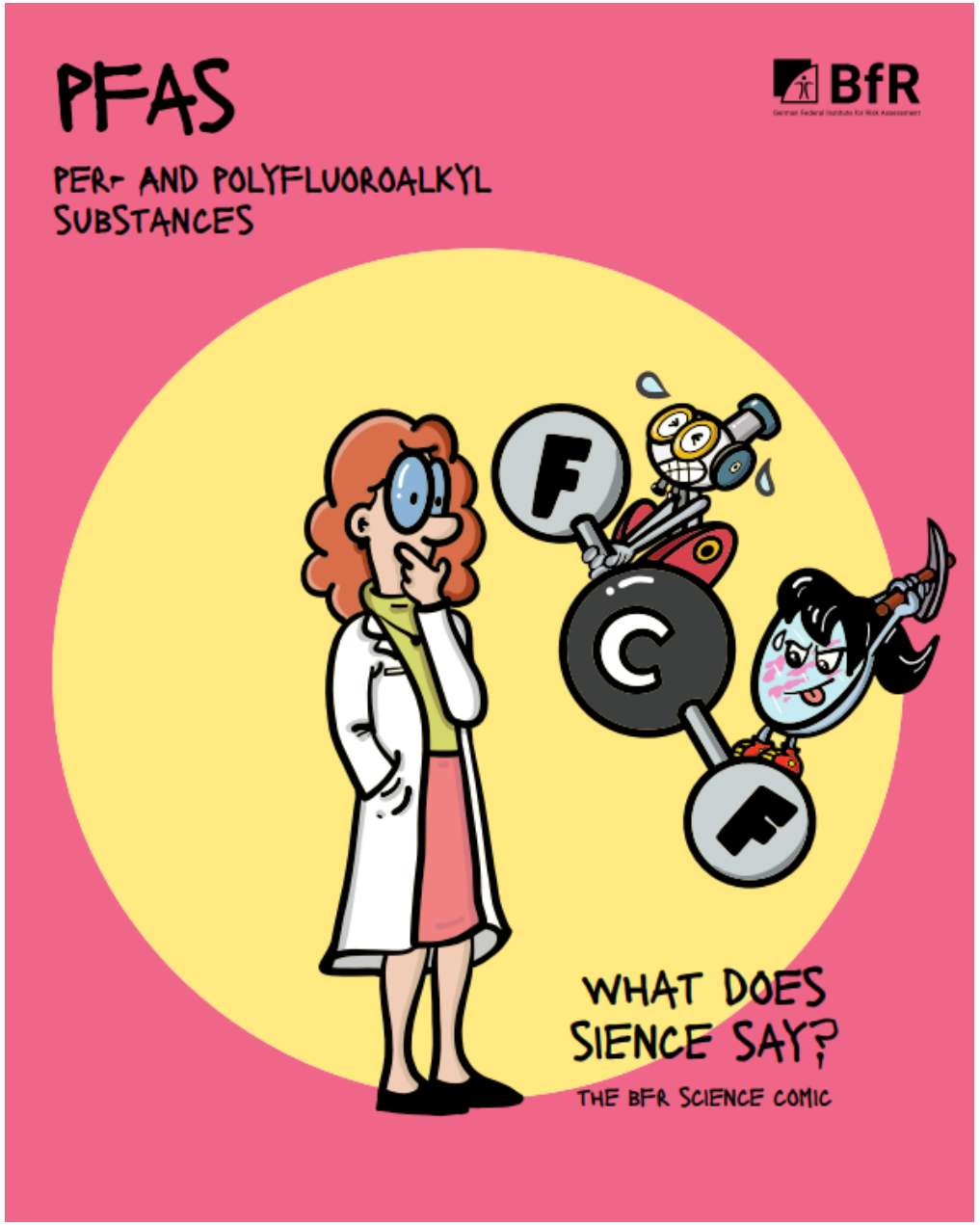
Retrospective Analysis: PFAS

EFSA's first dedicated opinion on PFOS/PFOA dates to July 2008.

Time frame for data collection: 2004–2007

PubMed query: (food[Title/Abstract] OR feed[Title/Abstract] OR diet*[Title/Abstract]) AND (contamination[Title/Abstract] OR residue*[Title/Abstract] OR toxic*[Title/Abstract] OR accumulation[Title/Abstract] OR exposure[Title/Abstract] OR persistent[Title/Abstract])

Number of documents: 18,179



https://www.bfr.bund.de/

Citrinin in Red Yeast Rice (RYR) Supplements

The first official EFSA publication 13 explicitly addressing citrinin in RYR dates to 2018 (EFSA ANS Panel, 2018).

Time frame for data collection: 2010–2017

PubMed query: ("dietary supplement*"[Title/Abstract] OR "food supplement*"[Title/Abstract] OR "nutritional supplement*"[Title/Abstract] OR "nutrition supplement*"[Title/Abstract] OR "functional food*"[Title/Abstract] OR "health supplement*"[Title/Abstract] OR nutraceutical*[Title/Abstract]) AND (risk[Title/Abstract] OR hazard[Title/Abstract] OR safety[Title/Abstract] OR toxic*[Title/Abstract] OR exposure[Title/Abstract])

Number of documents: 7,068

https://upload.wikimedia.org/wikipedia/commons/1/16/Red_yeast_rice.jpg

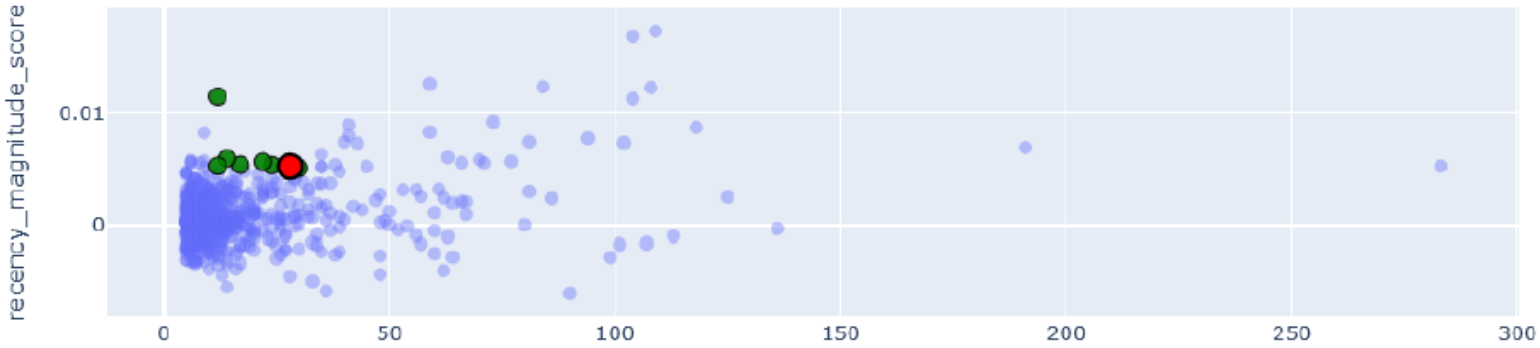
PFAS Results

N of micro-topics: 638

Count: 28 ; **Representation:** ['pfos', 'pfoa', 'pfcs', 'apfo', 'perfluorinated', 'perfluorooctane', 'linearbranched', 'pfosa', 'branched', 'sulfonate‘] → the main PFOS/PFOA chemical family discourse

Count: 9 **Representation:** ['pfos', 'quail', 'pfoskg', 'mallards', 'birds', 'mallard', 'geese', 'bobwhite', 'ld50', 'hmx‘] → PFOS in wildlife ecotoxicology

With 28 occurrences in 18,179 documents (≈0.15% of the corpus), → rare but semantically coherent micro-topic- The topic density indicates a suitable granularity regime: clustering is fine-grained enough to capture PFOS as a distinct discussion node, but not so fragmented that topics become trivial or uninterpretable.



Rank	Count	Recency-and-magnitude score	Weak signal score	Representation
0	12	0.011479	0.950000	[vanadium, den, hepatocarcinogenesis, 8ohdgs, bicyclol, deninduced, dna, dpcs, p00001, preneoplastic]
1	14	0.005924	0.526004	[acacia, gum, colostrum, gymnema, gurmarin, arabic, tuberomammillary, rk9, rk2, senegal]
2	10	0.004407	0.523799	[abcg1, abca1, macrophages, ldlr, macrophage, cholesterol, microm2, apoe, fdg, lesion]
3	17	0.005355	0.412669	[aps, endocrine, disruption, car, dieldrin, ru486, estriol, teleost, eds, marine]
4	16	0.004726	0.395317	[ifnalpha, serotonin, 13cisra, pica, mirtazapine, sickness, sleep, mcpp, antidepressant, kaolin]
5	18	0.004775	0.348595	[ergot, fescue, tall, ewes, hie, alkaloids, endophyteinfected, loe, alkaloid, toxicosis]
6	22	0.005657	0.307974	[cvd, counseling, gujaratis, visits, lifestyle, mental, management, cholesterol, ghq, caseness]
7	21	0.004532	0.257265	[egcg, tea, catechins, spicederived, green, gcbe, caffeine, theanine, epigallocatechin, gte]
8	24	0.005320	0.235325	[cd, cadmium, g1, dry, crabs, gill, fish, microg, midgut, macrocopa]
9	28	0.005228	0.129111	[pfos, pfoa, pfcs, apfo, perfluorinated, perfluorooctane, linearbranched, pfosa, branched, sulfonate]

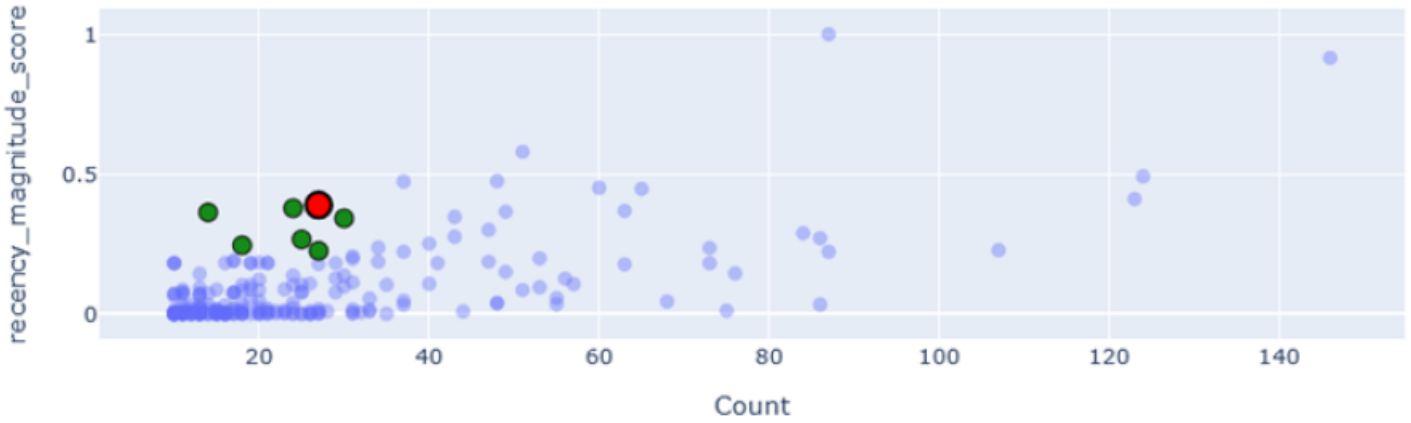
RYR Results

N of micro topics: 234

Count: 27

Representation: ['rice', 'red', 'yeast', 'monacolin', 'lovastatin', 'citrinin', 'ryr', 'monacolins', 'fermented', 'pcsk9']

With 27 occurrences in a corpus of 7,068 documents (≈0.38%), the RYR–citrinin topic is low-frequency but semantically well-defined. It is not diluted into broader categories such as "dietary supplements"



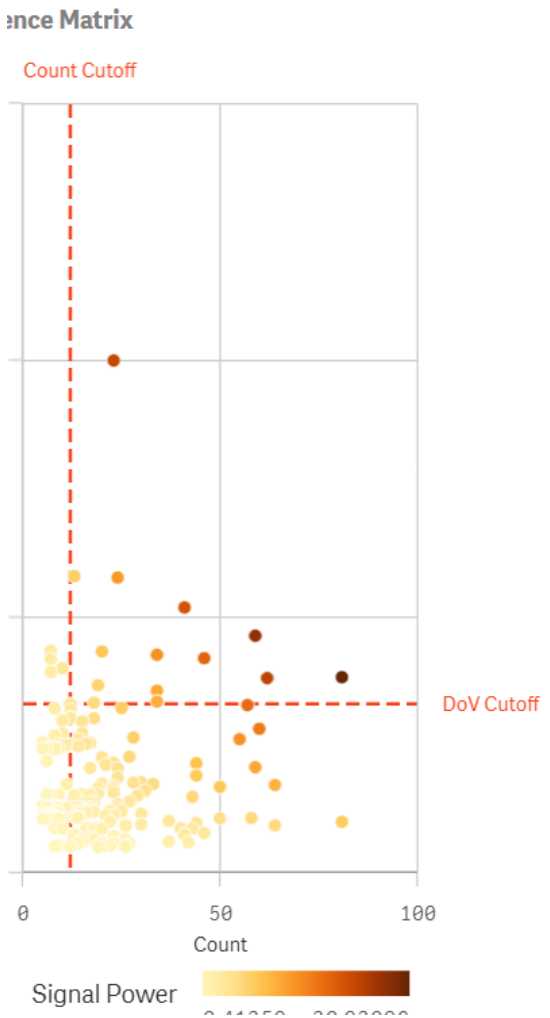
Rank	Count	Recency-and-magnitude score	Weak-Signal Score	Representation
0	14	0.362719	0.918237	[fracture, hip, steroid, anabolic, versus, rehabilitation, trial, trials, complications, albumin]
1	24	0.378124	0.652592	[kidney, ckd, caz, potassium, stones, da, piglets, llcpk1, hyperkalemia, zea]
2	27	0.389601	0.593750	[rice, red, yeast, monacolin, lovastatin, citrinin, ryr, monacolins, fermented, pcsk9]
3	18	0.244968	0.435088	[creatine, monohydrate, cr, cnh, cm, cnl, rhabdomyolysis, 2g, training, exercise]
4	30	0.341775	0.354535	[nanoparticles, nanotechnology, tq, nanoparticle, ra, solubility, size, food, stability, av]
5	25	0.266500	0.281831	[pfs, alkenylbenzenes, botanical, pfss, weps, alkenylbenzene, kcl, poisons, plant, concern]

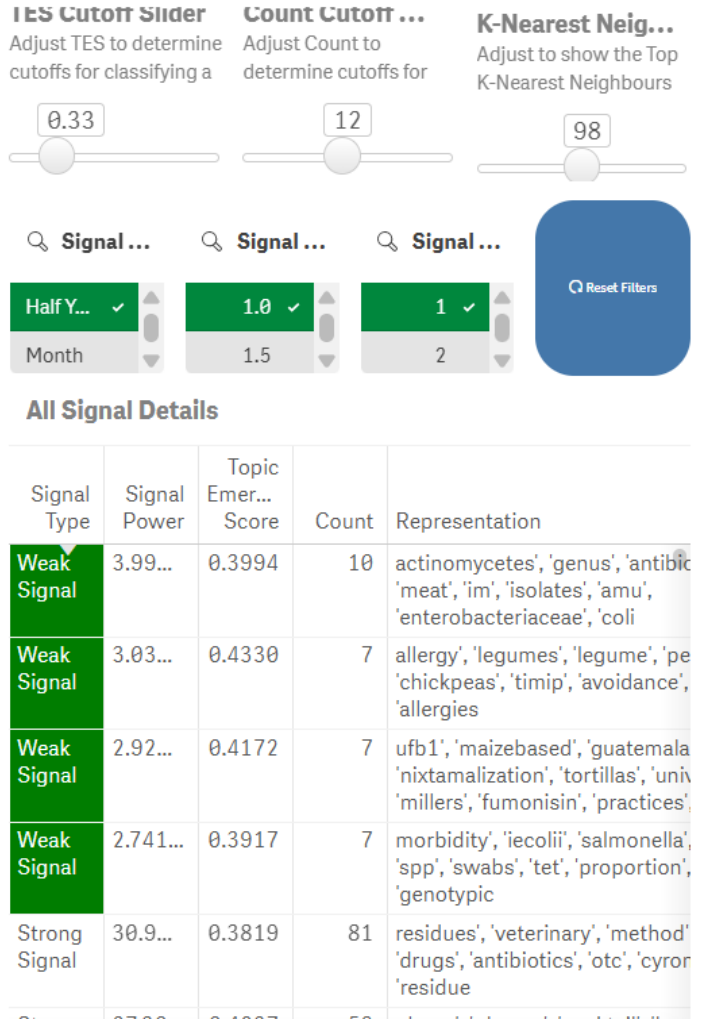
Resume

Validation shows that the Weak Signal Miner:

- ✓ Isolates (some) rare and semantically coherent micro-topics for very different types of emerging risks
- ✓ Identifies temporal escalation patterns in these topics well before they enter the regulatory mainstream.
- ✓ Ranks them among the top weak signals in their respective corpora, despite low absolute frequencies.
- ✓ proves to be a valid approach to shrink the search window for emerging risks.

- ❖ Tool will not detect "synthetic risks"or "unknown unknowns"but rather systemic risks
- ❖ Emerging risk detection still requires creativity, expertise and human-in-the-loop





Visit: github.com/HOLiFOOD/WSM

HoliFood Dashboard implementation

This research is part of a broader set of work packages within the HOLiFOOD project which has received funding from the European Union's Horizon Research and Innovation Actions, Grant Agreement no. 101059813 (https://holifoodproject.eu/about/).

German Federal Institute for Risk Assessment

Dr. Axel Menning

German Federal Institute for Risk Assessment

Innovation Centre Food Chain Modelling and Artificial Intelligence

Department Information Technology

German Federal Institute for Risk Assessment

bfr.bund.de/en

BfR | Identifying Risks – Protecting Health

Consumer health protection to go

BfR2GO – the BfR Science Magazine

bfr.bund.de/en/science_magazine_bfr2go.html

Follow us

- @bfrde | @bfren | @Bf3R_centre
- @bfrde
- youtube.com/@bfr_bund
- social.bund.de/@bfr
- linkedin.com/company/bundesinstitut-f-r-risikobewertung
- soundcloud.com/risikobewertung