Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Quantitative Weak Signal Framework for Detecting Emerging Computer Vision Technologies Using arXiv Data

Aditya Agarwal
Vivekananda Institute of Professional Studies- Technical Campus
Siddhant Kumar Sharma
Vivekananda Institute of Professional Studies- Technical Campus
Parneeta Chaudhary
parneeta@vips.edu

Vivekananda Institute of Professional Studies- Technical Campus

Research Article

Keywords: Weak Signal Analysis, Scientometrics, Semantic Embeddings, Computer Vision, Trend Detection, arXiv

Posted Date: April 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9251652/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Computer vision research on arXiv has approximately doubled every four years, exceeding the ability of any expert to manually track nascent trends. This study employs a quantitative Weak Signal Analysis framework to detect emerging computer vision technologies using large-scale preprint data. We analyzed 8,269 cs.CV papers from arXiv (2010–2022) using sentence-level semantic embeddings (all-MiniLM-L6-v2) rather than traditional keyword matching. Visibility and diffusion were operationalized through time-weighted term and document frequency metrics. The framework integrates Keyword Emergence and Issue Maps analyze the evolution of research topics across three developmental phases. The results reveal a significant shift: over 80% of contemporary computer vision concepts were nearly non-existent in the early 2010s, yet had grown into dominant research themes by the early 2020s. We identified four lifecycle archetypes; foundational, legacy, slow-burn, and explosive, and highlighted active frontiers including vision transformers and self-supervised learning. Beyond computer vision, this framework is directly applicable to any domain with a substantial preprint record, offering a cost-effective, reproducible methodology for technology foresight.

INTRODUCTION

The rapid evolution of scientific and technological (S&T) domains poses significant challenges for long-term strategic planning. Deep learning, autonomous systems, and messenger RNA (mRNA) vaccines initially appeared marginal before undergoing rapid mainstream adoption, and the institutions that identified these advancements at a nascent stage were positioned to achieve a first-mover advantage. Identifying these technologies early allows governments, corporations, and researchers to anticipate disruptions, allocate resources effectively, and pursue timely opportunities. Traditional approaches to emerging technology detection rely heavily on expert judgment. While expert opinions are valuable, they possess inherent with natural limitations: judgments may be biased, the volume of literature that can be manually reviewed is limited, and the speed at which contemporary research appears today is overwhelming. Due to this exponential growth, identifying emerging trends through manual monitoring alone has become increasingly difficult. There is a critical need for computational methods that can process large corpora quickly, consistently, and without subjective bias.

Weak Signal Analysis (WSA), originally introduced by Ansoff, provides one such systematic approach to detecting preliminary indicators of future developments before they become evident (Ansoff, 1975). Hiltunen later refined this into a three-dimensional model involving signals, issues, and interpretations (Hiltunen, 2008). Yoon expanded this framework by introducing a quantitative layer, using time-weighted metrics such as Term Frequency (TF) to capture visibility and Document Frequency (DF) to capture diffusion (Yoon, 2012). A persistent limitation of earlier implementations has been the reliance on exact keyword matching, which misses semantically related terms, a limitation that embedding-based methodologies are better equipped to address.

Expanding upon this groundwork, Ebadi and colleagues later applied a deep-learning-augmented WSA approach to the hypersonics domain (Ebadi et al., 2022), where it successfully revealed emerging research clusters and long-term developmental patterns. Motivated by their results, we adopt a similar methodology for the field of Computer Vision (CV), which has undergone rapid growth and multiple paradigm shifts from 2010 to 2022. During this period, arXiv, one of the most influential preprint platforms, has been central to the expedited and open dissemination of CV research.

The deep learning revolution in CV was catalyzed by the success of large Convolutional Neural Networks (CNNs) trained on Graphics Processing Units (GPUs), most notably AlexNet (Krizhevsky et al., 2012), followed by deeper residual models (He et al., 2016). These advances shifted the research focus from hand-crafted features to learned representations catalyzed extensive research across diverse tasks (detection, segmentation, 3D perception, etc.). Generative Adversarial Networks (GANs) introduced a new class of generative modeling in CV (Goodfellow et al., 2014), and attention mechanisms culminating in transformer architectures further reshaped the field (Vaswani et al., 2017; Dosovitskiy et al., 2021). Adversarial examples and robustness research emerged as critical concerns as methods matured (Szegedy et al., 2014). More recently, self-supervised methods (e.g., SimCLR, BYOL) have reduced dependence on labelled data and established new directions for representation learning (Chen et al., 2020; Grill et al., 2020).

Specifically, within the CV domain, arXiv papers typically appear months before journal publication, making the corpus a preliminary vantage point into nascent research activities. Through this study, we seek to utilize these early signals to quantitatively map how the CV landscape evolves. Co-word analysis, citation bursts, and topic modeling are well-established tools for mapping scientific evolution, but they primarily function retrospectively. Few studies have combined time-weighted weak signal metrics with preprint data to identify trends as they emerge. Existing bibliometric approaches are effective in retrospectively mapping dominant research fronts; however, they often exhibit limited sensitivity to early-stage or weakly expressed signals that precede major paradigm shifts. Despite arXiv serving as the primary hub for rapidly sharing new computer vision methods, its historical data remains underutilized for early trend detection (Callon et al., 1983; Small, 2006; Rotolo et al., 2015).

Addressing the limitations of existing approaches and the increasing availability of large-scale preprint repositories, this study investigates how weak signal analysis can be applied to detect emerging technological trajectories in computer vision research. In particular, the research questions (RQs) addressed in this study are the following:

RQ1. How can weak signal analysis be implemented using large-scale preprint data and semantic embedding techniques to identify emerging research topics in computer vision?

RQ2. What patterns of topic emergence and evolution can be observed in computer vision research between 2010 and 2022 using keyword emergence maps and keyword issue maps?

RQ3. What lifecycle trajectories characterize CV technology development, and do foundational, legacy, slow-burn, and explosive patterns manifest distinctly within the corpus?

CONCEPTUAL AND METHODOLOGICAL BACKGROUND

Bibliometric and Scientometric Approaches to Trend Detection

Understanding the evolution of scientific fields is a fundamental objective of scientometrics, with researchers developing a range of tools to track the trajectory of knowledge production. Co-word analysis captures semantic relationships between concepts through keyword co-occurrence patterns, facilitating the identification of dominant and emerging themes (Callon et al., 1983; Small, 2006; Rotolo et al., 2015).

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), have facilitated the automated discovery of latent thematic structures within large text corpora. In practice, however, LDA-derived topics tend to reflect dominant, well-established themes, the model requires a sufficient document density per topic to achieve stability, which often results in early-stage signals being obscured as noise rather than meaningful signal. Citation-based indicators, including citation bursts and citation velocity, provide valuable insights into scientific impact but inherently lag nascent research activity due to the time required for citation accumulation.

Weak Signal Analysis in Technology Foresight

Weak Signal Analysis is rooted in strategic management and futures studies, where it was developed to identify subtle and early indicators of potential disruption. Ansoff conceptualized weak signals as fragmented and ambiguous information preceding transformative shifts, while Hiltunen formalized the concept along the dimensions of visibility, diffusion, and interpretation (Ansoff, 1975; Hiltunen, 2008).

In order to facilitate large-scale application, subsequent studies quantitatively operationalized WSA quantitatively by integrating text mining techniques with time-weighted indicators. Yoon proposed utilizing term frequency and document frequency as proxies for visibility and diffusion, respectively, allowing weak signals to be systematically detected within large textual datasets (Yoon, 2012). Recent research has expanded this framework through the integration of machine learning and deep learning techniques to enhance signal extraction and clustering, demonstrating the applicability of WSA to extensive scientific and technological corpora (Ebadi et al., 2022; Yoon & Kim, 2012).

Preprint Repositories as Early Scientometric Indicators

Preprint repositories such as arXiv have emerged as integral components of scientific communication, particularly within the domain of computer science and artificial intelligence. Several studies have demonstrated that trends observed in preprint repositories often precede those detected in citation-based databases (Larivière et al., 2014; Wang et al., 2020). In fast-moving fields such as computer vision, where conference and preprint dissemination are primary modes of communication, arXiv provides a preliminary vantage point into emerging research trajectories. However, preprint data remains underutilized in scientometric analyses of technological emergence, particularly in combination with weak signal frameworks.

METHODOLOGY

The corpus was conducted from the arXiv OAI metadata snapshot, filtering for cs.CV papers published between 2010 and 2022. This study employs an end-to-end Python pipeline for data ingestion, preprocessing, keyword extraction, metric computation, and visualization. The implementation details (including packages, parameters, and source code links) is summarized in **Online Resource 1.** Additional processed datasets (DoV, DoD, and signal classifications) are provided as Supplementary Information as well in **Online Resources 2–4**.

Corpus Curation

We utilized the arXiv Open Archives Initiative (OAI) metadata snapshot (the arXiv Dataset) and filtered for papers categorized under Computer Vision (cs.CV) from 2010–2022. After de-duplication and quality filtering (removing non-English or extremely short abstracts), the final curated dataset contained 8,269 unique publications. To support both fine-grained and coarse-grained longitudinal views, we partitioned the data into 13 annual slices (t_1=2010,. . ., t_{13}=2022) and aggregated them into three macro-periods:

- P1: 2010–2013 (early CV period)
- P2: 2014–2017 (deep learning emergence)
- P3: 2018–2022 (deep learning maturity)

NLP Processing

For each paper we consolidated the title and abstract and applied:

- Lowercasing and punctuation removal
- Tokenization and lemmatization utilizing the spaCy library
- Stopword removal and numeric filtering
- Extraction of representative uni-grams and bi-grams (*n*- grams)
- Manual filtering and normalization to standardize variants (e.g., unifying “3D reconstruction” vs “3-d reconstruction”)

To facilitate more sophisticated semantic extraction, we emplyoed the all-MiniLM-L6-v2 model from the sentence-transformers library. We selected this distilled model over the standard BERT architecture because it retains the majority of BERT's semantic accuracy while offering a 5–10x improvement in inference speed. Additionally, MiniLM generates 384-dimensional embeddings (in contrast to the 768 dimensions of BERT-base), which mitigated memory constraint issues on standard hardware and ensured efficient processing of the large-scale corpus.

Preprocessing produced a vocabulary of normalized keywords for Degree of Visibility (DoV) and Degree of Diffusion (DoD) computation; manual normalization consolidated lexical variants. While manual intervention introduces a degree of subjectivity, it is commonly employed in co-word–based scientometric studies to ensure conceptual coherence and reduce noise arising from linguistic variation (Callon et al., 1983).

Time-weighted Signal Metrics

Following Yoon and Ebadi, we operationalize the visibility and diffusion dimensions via time-weighted metrics:

$$DOV_{ij} = \frac{TF_{ij}}{N_j} \times (1 - (w \times (n - j)))$$

$$DOD_{ij} = \frac{DF_{ij}}{N_j} \times (1 - (w \times (n - j)))$$

where TF_{ij} is the term frequency of keyword *i* in year *j*, DF_{ij} is the document frequency (number of documents containing keyword *i* in year *j*), N_j is the total document count for year *j*, n = 13 (years), and w = 0.025 is a discount factor that minimally reduces the contribution of older years.

We treated the time-decay factor "w" as a critical hyperparameter to control the rate at which historical signals diminish. After evaluating a grid of values ($w \in \{0.025, 0.05, 0.075, 0.1\}$), we selected w = 0.025. This value maximized the signal-to-noise ratio by effectively dampening obsolete terms while keeping foundational concepts intact. This balance between tracking recent developments and maintaining historical continuity aligns with prior weak signal research (Ebadi et al., 2022; Yoon, 2012).

Aggregation, Mapping, and Validation

Yearly DoD (**Online Resource 2**) and DoV (**Online Resource 3**) scores were aggregated into period-level summaries for the three macro-periods (P1, P2, P3). Specifically, yearly values were averaged within each period to obtain stable estimates of keyword visibility and diffusion across time windows.

Two visualization-based analytical frameworks were then constructed to analyze the evolution of research signals:

Keyword Emergence Map (KEM).

The KEM plots average term frequency (TF) on the x-axis as a measure of keyword visibility and the growth rate of DoV on the y-axis as an indicator of emergence. Following prior weak signal studies, the map is divided into four quadrants using the median values of the two axes. These quadrants correspond to different signal types:

- Strong signals: high visibility and high emergence rate
- Weak signals: low visibility but high emergence rate
- Latent signals: low visibility and low emergence rate
- Well-known but not strong signals: high visibility but low emergence rate

Keyword Issue Map (KIM).

The KIM is constructed similarly, but replaces term frequency with average document frequency (DF) to measure diffusion across the corpus. The x-axis represents average DF and the y-axis represents the growth rate of DoD. As with KEM, the map is divided into four quadrants using median thresholds to identify strong, weak, latent, and well-known signals based on their diffusion patterns. To increase robustness, signals were extracted by identifying keywords that occupy the same quadrant in both KEM and KIM. This intersection ensures that selected signals exhibit consistent patterns in both visibility and diffusion dimensions.

From these visualizations, we selected 95 high-impact keywords for expert validation. Each keyword was assigned one of five states within each period: Strong Signal, Weak Signal, Latent Signal, Well-known (but non-emergent), or Not Present. Several limitations should be noted. First, reliance on arXiv data may introduce disciplinary and geographic biases associated with preprint adoption. Second, the analysis is restricted to English-language abstracts, which may exclude contributions published in alternate languages.

RESULTS

This section represents frequency tables, vocabulary evolution, signal distributions, and detailed trajectories for important keywords.The list of final high impact signals can be accessed in **Online Resource 4.** The results are organized around three primary objectives: identifying the thematic composition of the field, analyzing how this vocabulary shifted over time, and tracking the transition of individual topics through various signal states across the three periods.

Top keywords by frequency and diffusion

Table 1 presents the top keywords by total term frequency (TF) from 2010–2022, while Table 2 summarizes the leading keywords by document frequency (DF).

Table 1
Top Keywords by Term Frequency (2010–2022)

Keyword	Total TF
Action recognition	446
3D Shape	175
3D Reconstruction	143
3D Point Cloud	110
Action Human Recognition	106
3D Face	104
Active Learning	94
3D Human	90
Activity Recognition	84
3D Detection Object	40

Table 2
Top Keywords by Document Frequency (2010– 2022)

Keyword	Total DF
Action recognition	164
3D recognition	109
3D Shape	80
3D Cloud points	60
Action Human Recognition	57
Activity Recognition	43
Activity Learning	41
3D Human	41
3D Face	30
3D Convolutional Neural	17

High term frequency reflects sustained topical attention, whereas high document frequency indicates broad diffusion across the research community. Together, these measures distinguish foundational research themes from narrowly specialized or nascent topics.

Vocabulary growth and stability

Figure 1 illustrates the number of unique normalized keywords per year. The corpus expanded from 25 documents in 2010 to 1,831 in 2022; concomitantly, the vocabulary of distinct semantic concepts rose sharply from 2010 through 2015 before stabilizing between 210–236 unique normalized keywords per year. This trend suggests that subsequent growth was driven primarily by the recombination and novel application of existing concepts rather than the introduction of entirely new fundamental terms.

Signal distribution across periods

Table 3 summarizes the classification of the 95 validated keywords across the three study periods. The data indicate a significant transition: in P1, 77 of these keywords were not present (81.1%); by P3, 69 had emerged as strong signals (72.6%). Figure 2 visualizes the evolution of signals across the three macro-periods (P1 → P2 → P3) using a Sankey diagram, illustrating how keywords transition between signal states over time.

Table 3
Signal Distribution of 95 Validated Keywords

Category	P1 (2010–13)	P2 (2014–17)	P3 (2018–22)
Strong Signal	12	43	69
Latent Signal	0	9	13
Well-known	5	18	5
Not Present	77	18	5
Not Validated	1	7	3

DISCUSSION

The maps and aggregated measures reveal several important narratives about evolution in CV.

Deep learning as the structuring paradigm

During macro-period P1 (2010–2013), many strong signals corresponded to classical machine learning techniques and task-specific engineering approaches, such as AdaBoost and Expectation-Maximization (EM). The emergence of convolutional neural networks (CNNs) and GPU-enabled deep learning models during P2 triggered a rapid paradigmatic reorientation of the field. Following the breakthroughs demonstrated by AlexNet and subsequent deep architectures (Krizhevsky et al., 2012; He et al., 2016),

model architectures and representation learning became the central drivers of progress in computer vision.

By P3, deep learning and its variants, including CNN-based architectures, attention-based models, and transformer adaptations, constituted the majority of strong signals, indicating a fundamental shift in how CV problems are conceptualized and addressed. The KEM and KIM for P3 (Fig. 3, Fig. 4) exhibit dense clustering in the high-visibility, high-diffusion quadrant, consistent with a field that has converged upon a shared methodological toolkit. These clusters correspond to mature and widely adopted research topics that now define the foundational core of contemporary computer vision.

Distinct lifecycle archetypes

Tracing the 95 validated keywords across P1–P3 reveals four archetypal trajectories:

Foundational: Concepts such as 3D reconstruction, and attention mechanisms maintained strong signal status across all three periods, suggesting they function as foundational pillars rather than passing trends. Action recognition, for instance, persists consistently under both handcrafted feature methods in P1 and attention-based architectures in P3; the terminological label remains constant while the underlying methodologies undergo total transformation. Attention mechanisms illustrate this most vividly: what began as a sequence modeling component became the architectural basis for vision transformers, carrying the same keyword across a complete paradigm shift.

Legacy (Sunset)

Methods such as AdaBoost and Expectation-Maximization were prominent strong signals in P1 but had largely disappeared from the corpus by P2. The decline occurred within a single period; the unprecedented performance of AlexNet (2012) was sufficiently disruptive that the research community largely abandoned boosting-based and EM-based CV methods almost immediately. The transition of AdaBoost from a top-ranked P1 signal to total absence in P2 does not represent a gradual decline, but rather a definitive discontinuation. This suggests that in fast-moving technical fields, paradigm shifts can obviate research directions entirely rather than simply pushing them to the margins.

Slow-burn (Latent → Strong)

GANs (Goodfellow et al., 2014) and 3D detection exemplify topics that were absent or latent in P1, appeared as weak signals in P2, and only consolidated into strong research fronts by P3. In the case of GANs, the original theoretical framework generated immediate interest, but practical CV applications; such as image synthesis and style transfer, required several years of architectural refinement and compute availability to become mainstream research pursuits. Slow-burn signals are the most actionable output of WSA, becoming visible early enough to inform funding or positioning decisions while being backed by enough cumulative evidence to be credible rather than merely speculative.

Explosive (Emergent): Adversarial attacks (Szegedy et al., 2014) and autonomous driving represent a different dynamic: topics that transitioned from effectively absent to strong signals within a single period. Explosive emergence is often triggered by discrete events: a single influential paper, a high-profile failure, or a sudden surge in industrial investment. Adversarial vulnerability, for instance, became an urgent concern almost immediately after Szegedy et al. demonstrated that imperceptible perturbations could fool state-of-the-art classifiers, a result with direct implications for deployed systems. Such signals resist prior prediction but tend to leave a clear retrospective trigger point, which is itself analytically useful. Table 4 provides representative examples.

Table 4
Representative Evolutionary Pathways (examples)

Keyword	P1	P2	P3
Action recognition	Strong	Strong	Strong
AdaBoost	Strong	Not Present	Not Present
3D detection	Not Present	Latent	Strong
GAN	Not Present	Latent	Strong
Adversarial attack	Not Present	Strong	Strong

Semantic Drift and Concept Fragmentation in Keyword-Based Signal Detection

An important limitation was observed during inspection of the keywords related to semantic consistency over time which has been discussed further.

The keyword *“attention”* illustrates semantic drift. It appears as a strong signal across P1–P3 (Fig. 5), suggesting stability; however, its meaning shifts substantially. In early computer vision, attention referred to saliency and region-weighting mechanisms, whereas after the introduction of transformer architectures in “Attention Is All You Need” (Vaswani et al., 2017), it became a core architectural operation enabling global feature interactions. Thus, a stable keyword masks a significant conceptual transition.

Self-supervised learning presents the opposite problem: it does not appear as a strong signal in any period (Fig. 5), despite being one of the most consequential developments in recent CV. The absence is a measurement artifact. Unlike “attention”, self-supervised learning shows concept fragmentation, wherein a single conceptual paradigm is expressed through multiple loosely related terms, including “contrastive learning,” “representation learning,” “pretraining,” and “unsupervised learning.”

These two phenomena represent complementary challenges in longitudinal text analysis. Together, they highlight a fundamental limitation of keyword-centric weak signal analysis, even when enhanced with semantic embeddings. Lexically stable strong signals require qualitative inspection, as they can possibly

mask conceptual discontinuity, whereas diffused paradigms are better tracked by clustering semantically related terms than by monitoring any single keyword.

Emerging fronts and weak signals

Using the Weak Signal Strength Matrix (Fig. 5), we identify several promising research directions that appeared as weak or latent signals in P2 but strengthened significantly in P3, or remain weak signals while exhibiting clear growth trends. Notable examples include Vision Transformers (ViT) (Vaswani et al., 2017; Dosovitskiy et al., 2021), self-supervised learning approaches such as SimCLR and BYOL (Chen et al., 2020; Grill et al., 2020), spiking neural networks, and panoptic segmentation. These areas represent active frontiers where methodological innovation and increasing research attention suggest strong potential for future impact.

Methodological validation

The framework reproduces several well-documented historical developments in computer vision, including the CNN-driven deep learning revolution (Krizhevsky et al., 2012), the emergence of GAN-based generative modeling (Goodfellow et al., 2014), and the growing influence of attention-based architectures (Vaswani et al., 2017). Simultaneously, our analysis highlights early-stage research directions that are just beginning to take shape. This dual capacity to validate historical trends and detect upcoming shifts, underscores the value of weak signal analysis for scientific foresight. Furthermore, sensitivity analysis confirmed that the identified signal trajectories remained stable across small variations in the time-decay parameter ($w \in 0.025–0.1$). This stability suggests that the framework is robust to moderate changes in parameter selection and that the detected trends reflect genuine patterns in the underlying research corpus rather than artifacts of the weighting scheme.

From a scientometric perspective, the identified lifecycle archetypes provide insight into the temporal dynamics of scientific innovation. Weak signal trajectories capture early-stage research activity that is not yet visible through citation-based indicators, highlighting the complementary value of preprint-based analyses for technology foresight and research policy (Rotolo et al., 2015).

Conclusion and Future Work

We applied a time-weighted TF/DF-based Weak Signal Analysis to 8,269 arXiv CV papers (2010–2022) and documented a quantifiable paradigm shift: the majority of modern CV concepts were absent in the early 2010s and became dominant by the early 2020s. We characterized lifecycle archetypes (foundational, legacy, slow-burn, explosive) and identified nascent fronts (e.g., ViT, self-supervised learning) deserving of continued monitoring.

Future work will focus on automating the continuous monitoring of preprint streams to provide real-time weak-signal dashboards. Furthermore, we plan to apply hierarchical clustering to the semantic embeddings to aggregate isolated signals into broader, more cohesive thematic groups. Although demonstrated within the context of computer vision, the proposed framework is fully applicable to other

rapidly evolving scientific domains, including natural language processing, robotics, and interdisciplinary AI research for cross-domain foresight.

Declarations

Data and Code Availability

We retrieved the primary dataset for this analysis via the arXiv Open Archives Initiative (OAI) metadata snapshot [https://www.kaggle.com/datasets/Cornell-University/arxiv].To facilitate reproducibility and future research, we have made the custom Python code for our Weak Signal Analysis framework openly available at [https://github.com/sidsharmaa/weak-signals-new].

The datasets generated during this study are available as Supplementary Information including Degree of Diffusion per Year, Degree of Visibility per Year, and High Impact Signals Classification.

Competing Interests

Authors declare no competing interests

Funding

No funding was received or utilized for this study

CRediT Authorship Contribution Statement

Conceptualization: Aditya Agarwal, Siddhant Kumar Sharma, Parneeta Chaudhary; Methodology: Aditya Agarwal, Siddhant Kumar Sharma; Formal analysis and investigation: Aditya Agarwal, Siddhant Kumar Sharma, Parneeta Chaudhary; Data curation: Aditya Agarwal, Siddhant Kumar Sharma; Visualization: Aditya Agarwal, Siddhant Kumar Sharma; Writing – original draft preparation: Aditya Agarwal, Siddhant Kumar Sharma; Writing, review and editing: Parneeta Chaudhary; Supervision: Parneeta Chaudhary; Project Administration: Parneeta Chaudhary.

Conflict of Interest

Authors declare no conflict of interest.

References

1. Ansoff, H. I. (1975). Managing strategic surprise by response to weak signals. *California Management Review*, 18(2), 21–33. https://doi.org/10.2307/41164635
2. Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235. https://doi.org/10.1177/053901883022002003

3. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning*, 119, 1597–1607.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations (ICLR)*.
5. Ebadi, A., Auger, A., & Gauthier, Y. (2022). Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *Journal of Informetrics*, 16(4), 101344. https://doi.org/10.1016/j.joi.2022.101344
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
7. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., … Valko, M. (2020). Bootstrap your own latent: A new approach to self–supervised learning. Proceedings of the International Conference on Machine Learning, 119, 3711–3721.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90
9. Hiltunen, E. (2008). The future sign and its three dimensions. *Futures*, 40(3), 247–260. https://doi.org/10.1016/j.futures.2007.08.021
10. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.
11. Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6), 1157–1169. https://doi.org/10.1002/asi.23044
12. Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, 44(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006
13. Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, 68(3), 595–610. https://doi.org/10.1007/s11192-006-0132-y
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. Proceedings of the International Conference on Learning Representations (ICLR).
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
16. Wang, Z., Chen, Y., & Glänzel, W. (2020). Preprints as accelerator of scholarly communication: An empirical analysis in mathematics. *Journal of Informetrics*, 14(4), 101097. https://doi.org/10.1016/j.joi.2020.101097
17. Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39(16), 12543–12550.

https://doi.org/10.1016/j.eswa.2012.04.059

18. Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 90(2), 445–461. https://doi.org/10.1007/s11192-011-0543-2

Figures

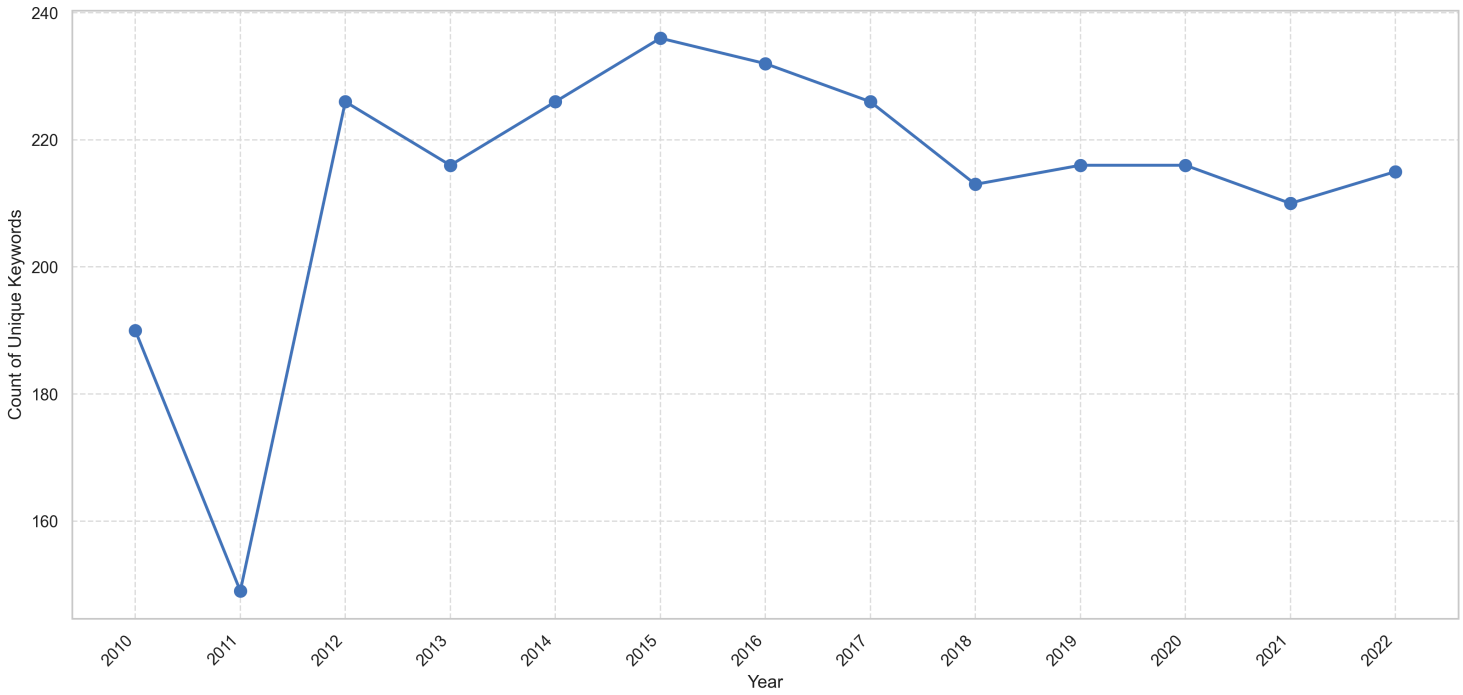


Figure 1

Number of Unique Normalized Keywords per Year (2010–2022)

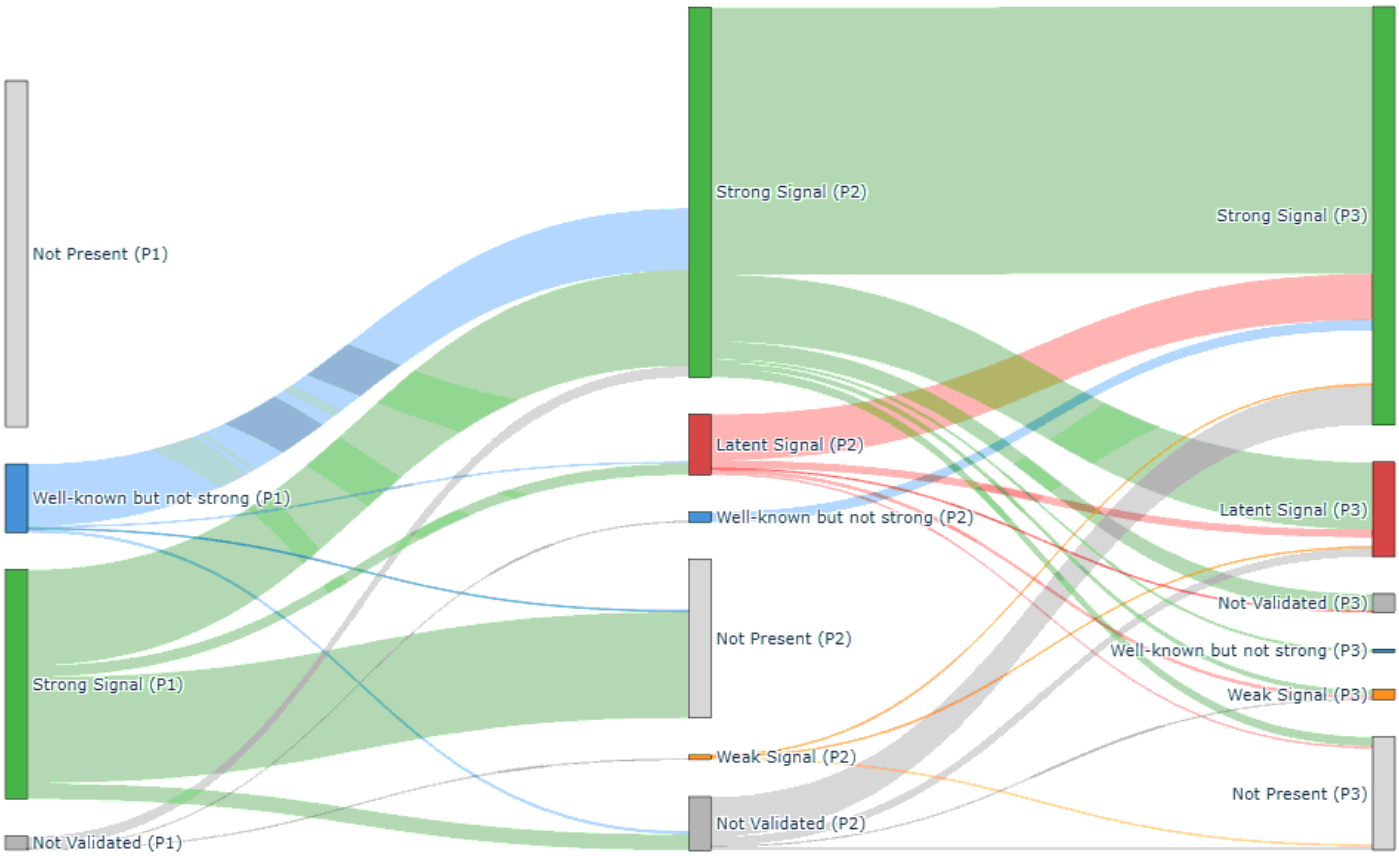


Figure 2

Signal evolution across periods (P1 → P2 → P3)

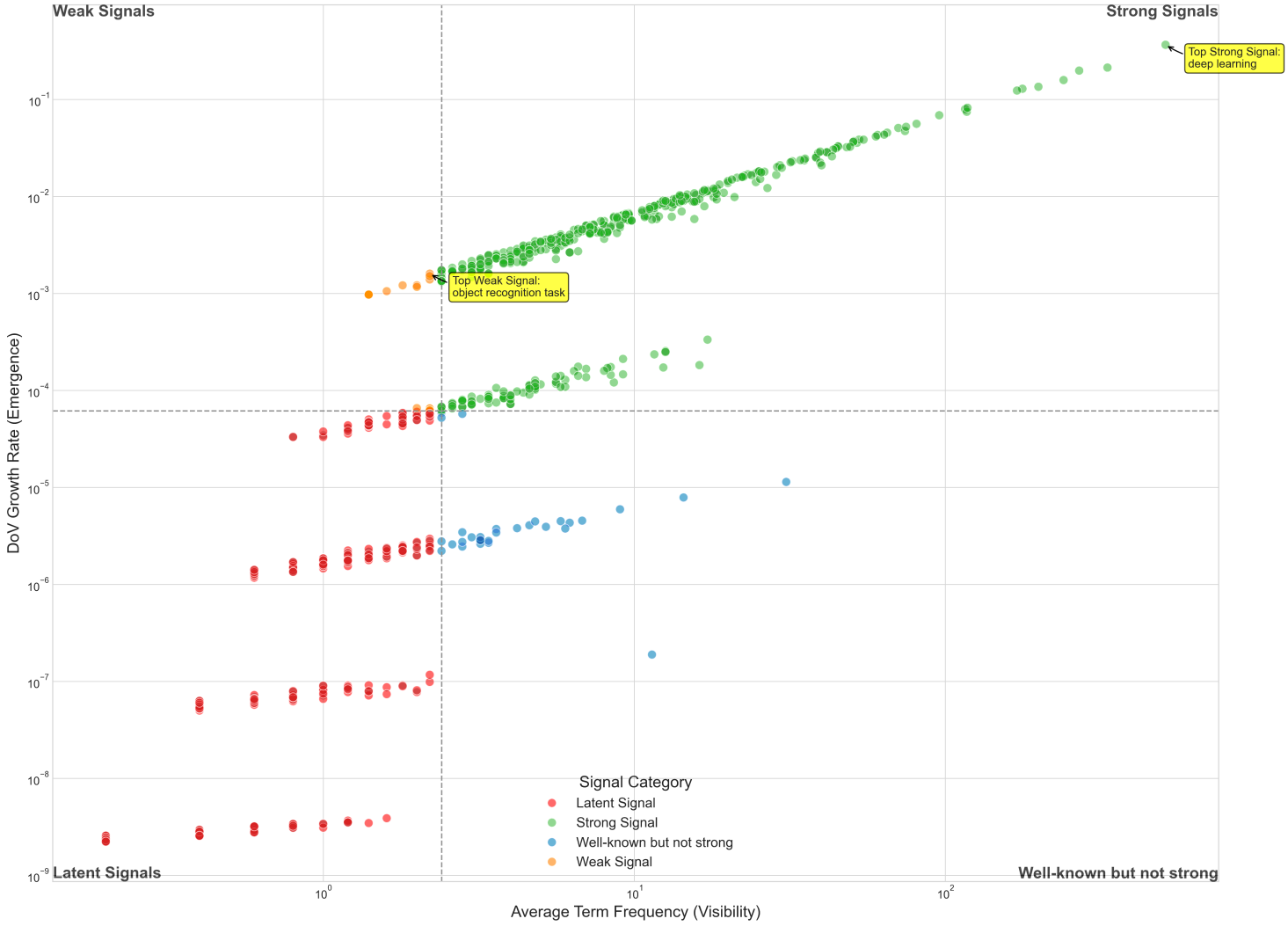


Figure 3

Keyword Emergence Map (KEM) for P3 (2018–2022)

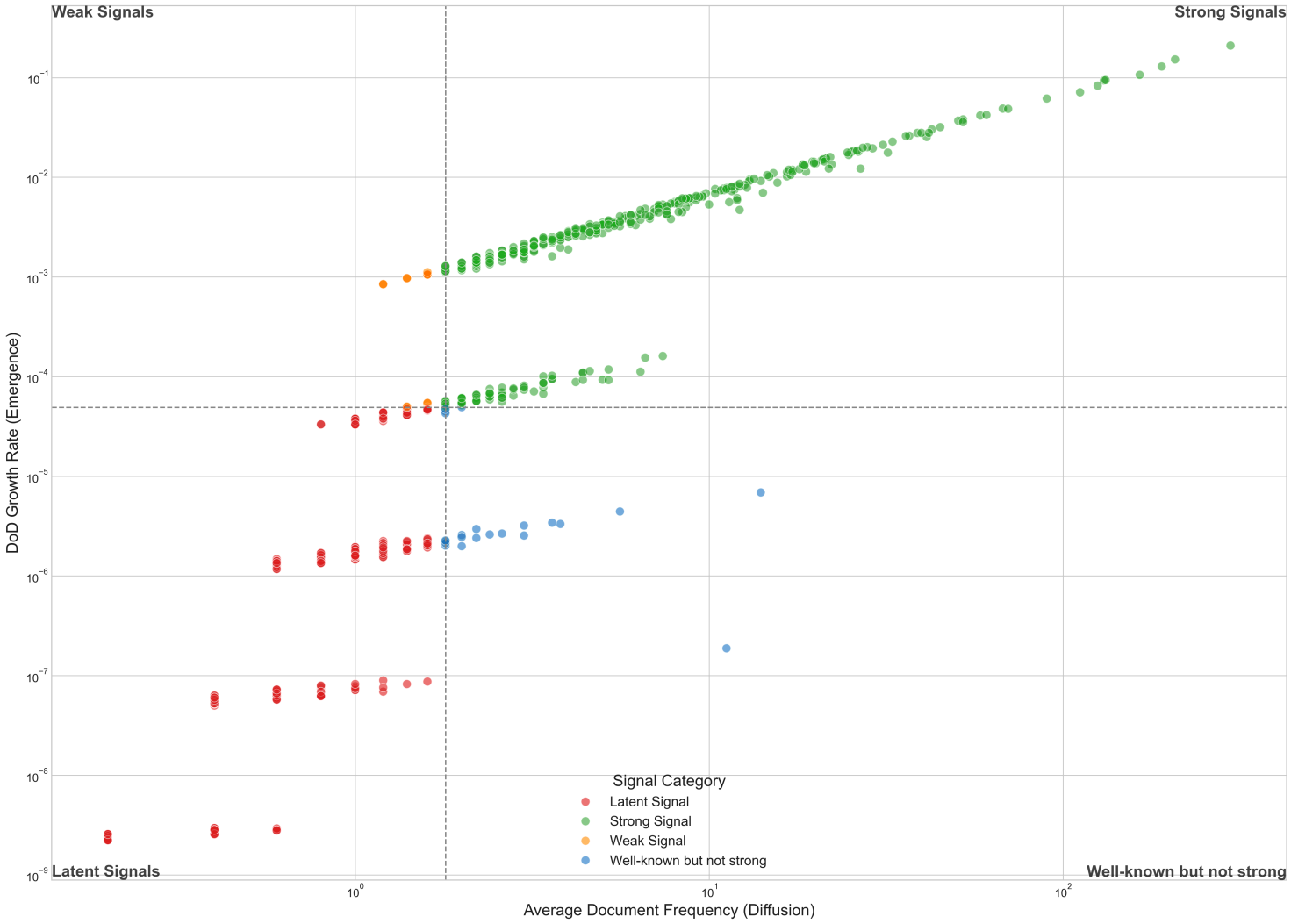


Figure 4

Keyword Issue Map (KIM) for P3 (2018–2022).

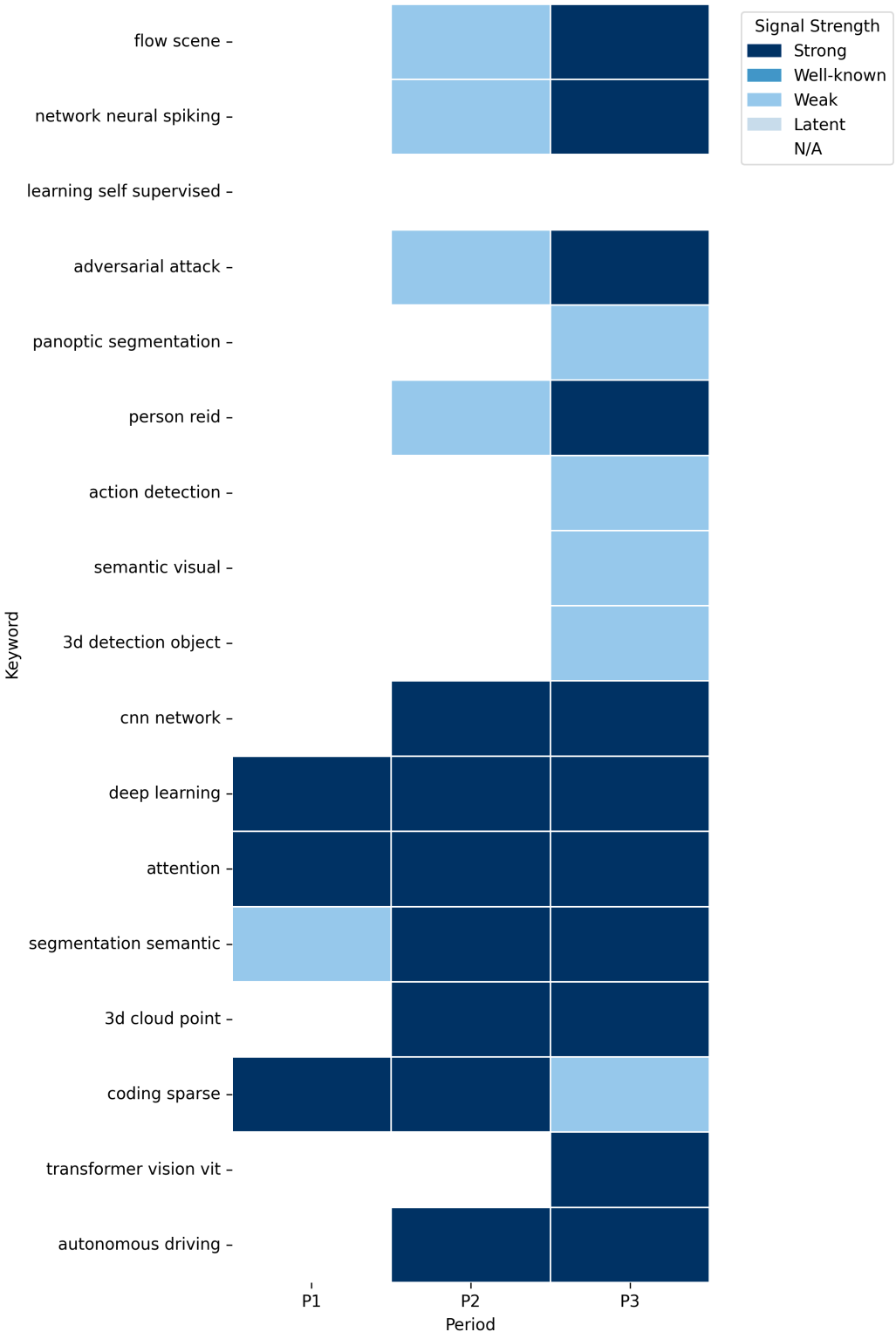


Figure 5

Weak Signal Strength Matrix (P1–P3) for selected keywords

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.csv
- ESM3.csv
- ESM4.xlsx