Cracking the code of innovation: Decoding unstructured website text for predicting company innovation in time and space

Sergey Sosnovskikh, PhD
Manchester Metropolitan University (UK)
s.sosnovskikh@mmu.ac.uk

Korneliusz Pylak, PhD
Lublin University of Technology (Poland)
korneliusz.pylak@pollub.pl

Mateusz Traczyński
Lublin University of Technology (Poland)

GEOINNO (January 2024)

Research background

Traditional methods relied on well-established secondary data to measure innovativeness and other companies' key performance indicators:

- **Patents** (Abbasiharofteh et al., 2022; Nasirov, 2020)
- **R&D projects** (Simensen & Abbasiharofteh, 2022;
- **Administrative records** (Gandin & Cozza, 2019; Maravelakis et al., 2006)
- **Scientific publications** (Marchiori et al, 2021; Cillo et al., 2019)

The landscape of innovation geography research has witnessed a transformative shift, fuelled by:

- Advancements in computational power and language modelling
- Large textual data available from diverse sources: **job postings**, **patent documents**, **web texts**, and **trademark data**.

This digital revolution has unlocked new possibilities for exploring **regional economic development**, **labour market dynamics**, and the **geographies of knowledge production and relationships** (Aweisi et al., 2021; Cetera et al., 2022; Gök et al., 2015; Skhvediani et al., 2022).

The integration of textual data analysis has been a trailblazing approach in (Abbasiharofteh et al., 2023; Ashouri et al., 2022; Daas & van der Doef, 2020; Gök et al., 2015; Kinne & Axenbeck, 2020):

- **Innovation geography** research
- Incorporating the analysis of **digital footprints of inter-firm linkages**
- **Social media data**
- **Digitized historical newspaper archives**

The use of unstructured textual data is gaining momentum, offering researchers innovative avenues to comprehend and interpret the intricate connections within innovation geography.

Our approach distinguishes itself by bridging the **contextual**, **temporal**, and **spatial aspects of innovation**. **The aim of this study is to unravel the innovation dynamics within the companies and discern early symptoms or indications that precede the official introduction of innovations.**

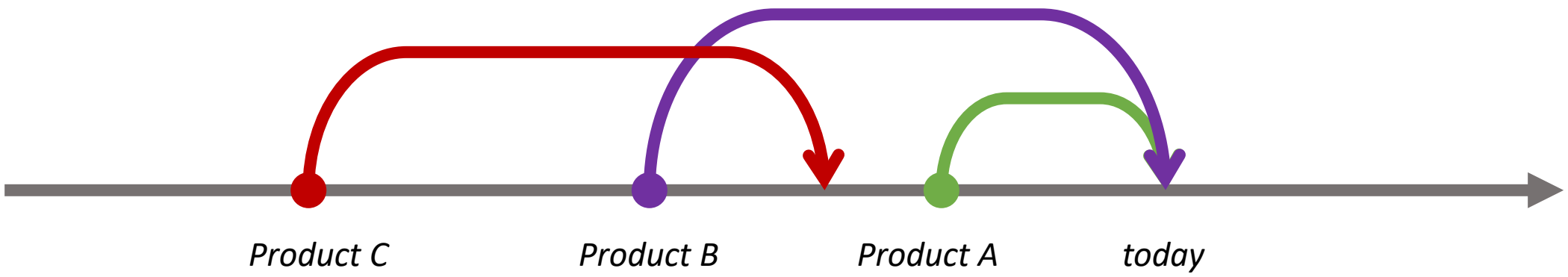
Methodology

The Aim of the Website Research

- Product extraction
- Historical analysis
- Patent correlations
- Website pre-launch examination
- Implications for strategic decision-making and future directions

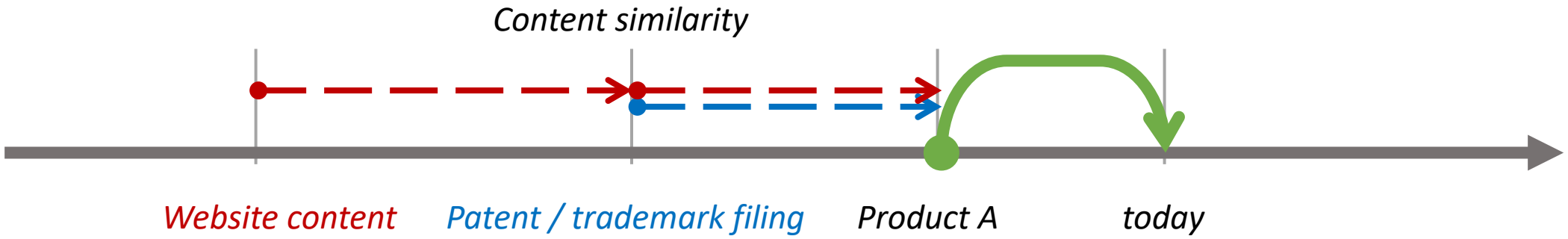
A comprehensive analysis of product introductions

- **Product extraction:** Scrapping subsites for insights
  - Identify and extract all products from the current website
  - Pinpoint subsites responsible for product storage
- **Innovations over Time:** Utilizing the Wayback Machine
  - Delve into the historical timeline to extract moments of new product introductions (innovations)
  - Explore the correlation between innovations and temporal evolution



Picking up symptoms of innovation

- **Symptom 1:** patents / trademarks vs. innovations: a synergistic analysis
  - Investigate the temporal connection between filed and received patents and product innovations
  - Evaluate patents as potential indicators of upcoming innovations
- **Symptom 2:** analyzing the website before product introduction
  - Conduct a detailed analysis of the website, focusing on the 'news' section, prior to product introductions
  - Uncover patterns and insights that may foreshadow upcoming innovations



The Website Selection Process

- **Focus on innovation**
  - Focus on companies with a patent filing history as potential indicators of innovation
  - These companies possibly do have their own products (not shops, representatives, etc.)
  - The extensive pool of corporate companies considered – 7,728 in total (years 2000-2023)
- **Diverse websites inclusion**
  - Out of the 7,728 companies, 1,450 inserted their websites into the registry *(→ good for training)*
  - The diverse industries represented within the 1,450 companies *(→ good for training)*

Analysing the websites

- We analyse 1,450 websites
- We extract all the internal hyperlinks from the websites (46,454)
- The repeating internal subsites are potential product storing websites (such as /products/ or /offer/)

Example website

1	https://vigophotonics.com
2	https://vigophotonics.com/about-us/rd-projects/poir/
3	https://vigophotonics.com/about-us/
4	https://vigophotonics.com/about-us/news/
5	https://vigophotonics.com/about-us/career/
6	https://vigophotonics.com/about-us/rd-projects/
7	https://vigophotonics.com/about-us/public-orders/
8	https://vigophotonics.com/investor-relations/
9	https://vigophotonics.com/products/
10	https://vigophotonics.com/products/epi-wafers/
11	https://vigophotonics.com/products/epi-wafers/ingaas-wafers/
12	https://vigophotonics.com/products/epi-wafers/vcsel-epi-structure/
13	https://vigophotonics.com/products/epi-wafers/qcls-epi-structure/
14	https://vigophotonics.com/products/infrared-detectors/
15	https://vigophotonics.com/products/infrared-detectors/hgcdte-mct-photoconductive-detectors/
16	https://vigophotonics.com/products/infrared-detectors/hgcdte-mct-photovoltaic-detectors/
17	https://vigophotonics.com/products/infrared-detectors/inas-and-inassb-detectors/
18	https://vigophotonics.com/products/infrared-detectors/hgcdte-mct-multi-channel-detectors/
19	https://vigophotonics.com/products/infrared-detection-modules/
20	https://vigophotonics.com/products/infrared-detection-modules/selected-line/
21	https://vigophotonics.com/products/infrared-detection-modules/configurable-line/
22	https://vigophotonics.com/products/infrared-detection-modules/inassb-affordable-detection-modules/

Useless subsite (about us)

Product subsite

Product category

Product
Product
Product

Product category

Product
Product
Product
Product

Product category

Product
Product
Product

Problems

- Products might be located on the subsites not linked to the main website (another scrapping of subsites might be needed)
- Subsites storing products are named very differently
- Not always there are separate subsites, sometimes product sites are linked to the main website
- Products might be listed on the last subsite, do not have separate subsites (plain unstructured text analysis is needed)

Solution

- Train the model to extract answers to the question: Which subsites encompass the following:
  - Product container
  - Product subsites (categories)
  - Products themselves
- Finetune, for example, the DistilBERT model by:
  - Identifying tokens in the subsites (context)
  - Addressing various questions
  - Recognizing answers that begin with the names of the subsites
- This outlines our current activities…

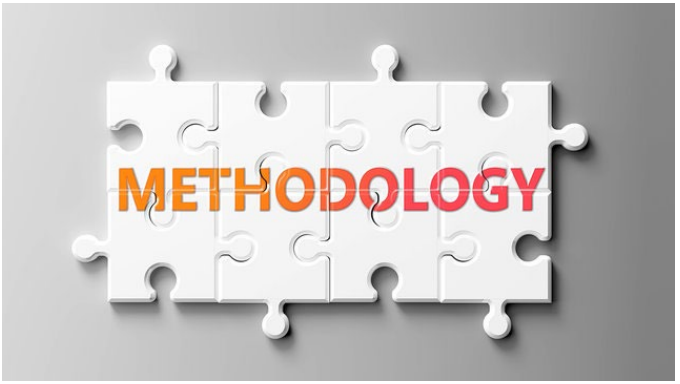
Next steps

- Each company will be analysed through the prism of its geolocation and well-defined socio-economic context, considering the **company's structure**, **economic and technological variety**, **local knowledge complexity**, etc.

How to investigate the causal connections between innovation measures over time?

Three distinct approaches:

1) Text-based R&D expenditures
2) Patents and trademarks (derived from patent office databases)
3) Text-based innovations.



- For text mining, we will utilize various topic-modelling tools, including **Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Correlated Topics Models (CTM), and word embeddings (GloVe).**
- Additionally, we will explore the transformative capabilities of **Natural Language Processing (NLP) with cutting-edge Transformer models.**

Potential Contributions (I)

- ✓ We extend the analysis of unstructured website text beyond static reflections of company websites that existing methods often rely on, leveraging advanced web scraping, social network analysis, and natural language processing techniques (Abbasiharofteh et al., 2023; Ashouri et al., 2022; Kinne & Axenbeck, 2020; Skhvediani et al., 2022).

- ✓ Instead, **we introduce a temporal dimension, anticipating a company's innovations well in advance based on changes in the textual content it publishes on its website over time.**

- ✓ While text-based innovation predictions in analyst reports have been explored over time (Bellstam et al., 2021), **our approach pioneers using unstructured website text as an early indicator of forthcoming innovations.**

- ✓ This novel temporal perspective allows us to grasp the evolution of innovative activities and provides a more comprehensive understanding of the innovation process within a single company.

Potential Contributions (II)

- Innovation capability is closely linked to a company's ability to combine existing knowledge and resources over time (Audretsch & Belitski, 2022; Bruno et al., 2022; Tomizawa et al., 2020).
- Physical proximity and inter-firm relationships are pivotal in facilitating learning and triggering innovation (Alam et al., 2022; Bailey et al., 2018; Obschonka et al., 2023; Singh et al., 2022).

- ✓ To account for this spatial dimension, **our approach goes beyond isolated predictions and considers the colocation of innovative entities in proximity.**
- ✓ By capturing the interplay between innovative companies in their spatial context, **we provide a more nuanced understanding of the geographies of knowledge production and relationships**, shedding light on how spatial dynamics shape innovation.

References

Abbasiharofteh, M., Castaldi, C., & Petralia, S. (2022). *From patents to trademarks: Towards a concordance map*. European Patent Office.

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, 1–23.

Alam, M. A., Rooney, D., & Taylor, M. (2022). From ego-systems to open innovation ecosystems: A process model of inter-firm openness. *Journal of Product Innovation Management*, *39*(2), 177–201.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., & Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, *42*, 108246.

Audretsch, D. B., & Belitski, M. (2022). The knowledge spillover of innovation. *Industrial and Corporate Change*, *31*(6), 1329–1357.

Aweisi, A., Arora, D., Emby, R., Rehman, M., Tanev, G., & Tanev, S. (2021). Using web text analytics to categorize the business focus of innovative digital health companies. *Technology Innovation Management Review*, *11*(7/8).

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, *32*(3), 259–280.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A Text-Based Analysis of Corporate Innovation. *Management Science*, *67*(7), 4004–4031.

Bruno, R. L., Crescenzi, R., Estrin, S., & Petralia, S. (2022). Multinationals, innovation, and institutional context: IPR protection and distance effects. *Journal of International Business Studies*, *53*(9), 1945–1970.

Cetera, W., Gogołek, W., Żołnierski, A., & Jaruga, D. (2022). Potential for the use of large unstructured data resources by public innovation support institutions. *Journal of Big Data*, *9*(1), 46.

Cillo, V., Petruzzelli, A. M., Ardito, L., & Del Giudice, M. (2019). Understanding sustainable innovation: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, *26*(5), 1012–1025.

Daas, P. J., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, *36*(4), 1239–1251.

Gandin, Ilaria, and Claudio Cozza. 2019. 'Can We Predict Firms' Innovativeness? The Identification of Innovation Performers in an Italian Region through a Supervised Learning Approach'. PloS One 14 (6): e0218175.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

Maravelakis, E., N. Bilalis, A. Antoniadis, K. A. Jones, and V. Moustakis. 2006. 'Measuring and Benchmarking the Innovativeness of SMEs: A Three-Dimensional Fuzzy Logic Approach'. Production Planning & Control 17 (3): 283–92.

Marchiori, Danilo Magno, Silvio Popadiuk, Emerson Wagner Mainardes, and Ricardo Gouveia Rodrigues. 2021. 'Innovativeness: A Bibliometric Vision of the Conceptual and Intellectual Structures and the Past and Future Research Directions'. Scientometrics 126 (1): 55–92. https://doi.org/10.1007/s11192-020-03753-6.

Nasirov, S. (2020). Trademark value indicators: Evidence from the trademark protection lifecycle in the U.S. pharmaceutical industry. *Research Policy*, *49*(4), 103929.

Obschonka, M., Tavassoli, S., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2023). Innovation and inter-city knowledge spillovers: Social, geographical, and technological connectedness and psychological openness. *Research Policy*, *52*(8), 104849.

Simensen, E. O., & Abbasiharofteh, M. (2022). Sectoral patterns of collaborative tie formation: Investigating geographic, cognitive, and technological dimensions. *Industrial and Corporate Change*, *31*(5), 1223–1258.

Singh, A., Chhetri, P., & Padhye, R. (2022). Modelling inter-firm competitive rivalry in a port logistics cluster: A case study of Melbourne, Australia. *The International Journal of Logistics Management*, *33*(2), 455–476.

Skhvediani, A., Sosnovskikh, S., Rudskaia, I., & Kudryavtseva, T. (2022). Identification and comparative analysis of the skills structure of the data analyst profession in Russia. *Journal of Education for Business*, *97*(5), 295–304.

Tomizawa, A., Zhao, L., Bassellier, G., & Ahlstrom, D. (2020). Economic growth, innovation, institutions, and the Great Enrichment. *Asia Pacific Journal of Management*, *37*(1), 7–31.

Manchester Metropolitan University

POLITECHNIKA LUBELSKA
LUBLIN UNIVERSITY OF TECHNOLOGY

Thank you for your attention!

Sergey Sosnovskikh, PhD
s.sosnovskikh@mmu.ac.uk