

# ISTARI.AI

DEEP MARKET INTELLIGENCE

---

## Web Mining of Firm Websites

Jan Kinne

ZEW



# **istari.ai**

---

- Data Intelligence startup
- Founded 2019
- Academic spin-off
- Web-based company and  
market information in real-time
- Development of webAI

# The founders

---

**David Lenz**



- PhD in statistics and econometrics
- Specialized in machine learning and NLP

**Jan Kinne**



- PhD in geoinformatics and economic geography
- Specialized in firm data and location analysis

# webAI

Manual analysis...

...of traditional data sources.

- Slow
- Outdated info
- Not flexible
- Expensive



AI-based analysis...

...of web data.

- High-frequency
- Up-to-date
- Flexible
- Large-scale



> **Deep Market Intelligence**

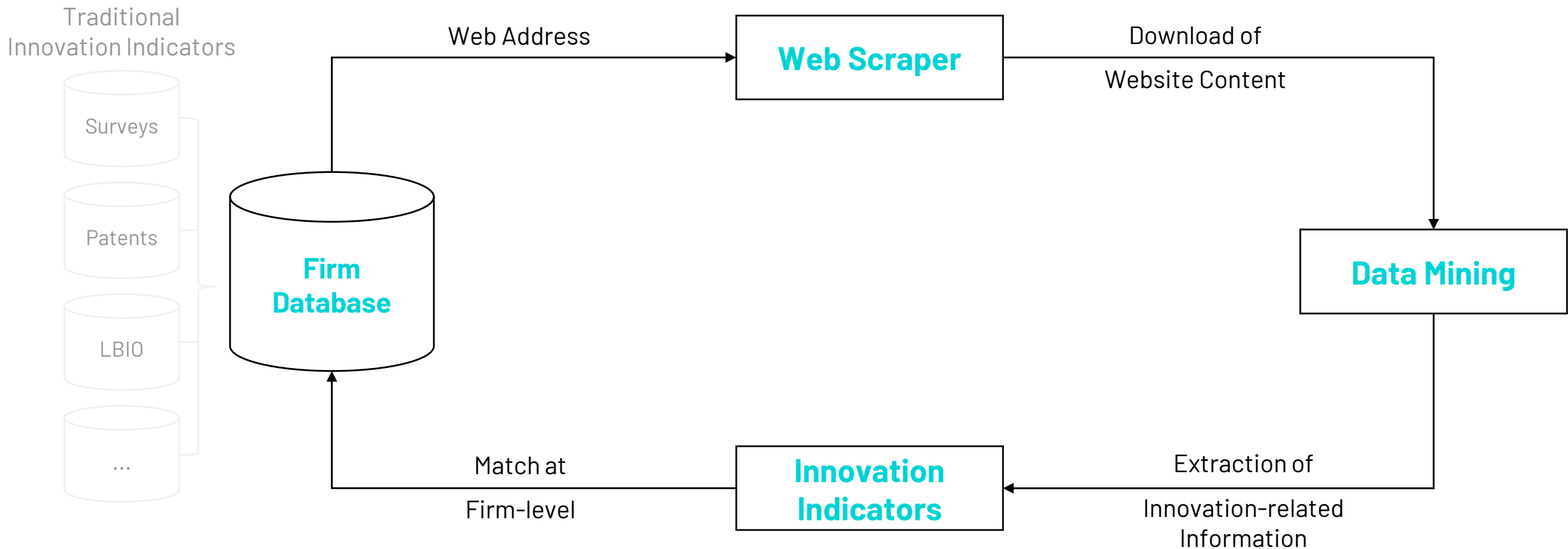
# Our research

Extensive joint research on the topic of **web-based innovation indicators** in the context of our PhDs.

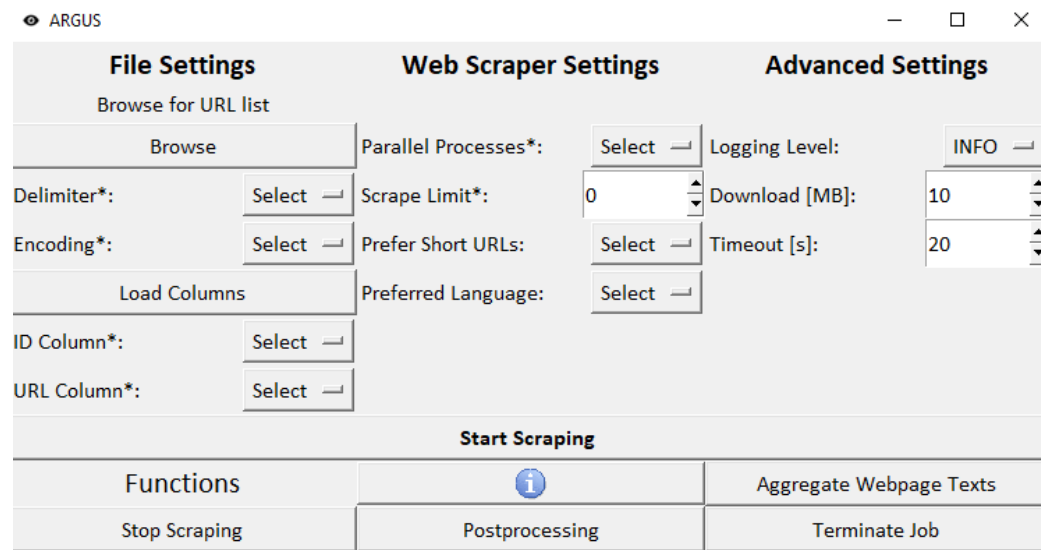
- Traditional innovation indicators are often not granular, outdated and costly to collect.
- Use of company websites as data source.
- Extraction of relevant information using text mining.



# Web Mining Framework



- A open-source web scraping tool for “broad crawls”.
- Based on the Python Scrapy framework.
- Crawling of millions of different (company) websites.
- Extraction of texts and hyperlinks.
- Quite fast even on office-grade computers.

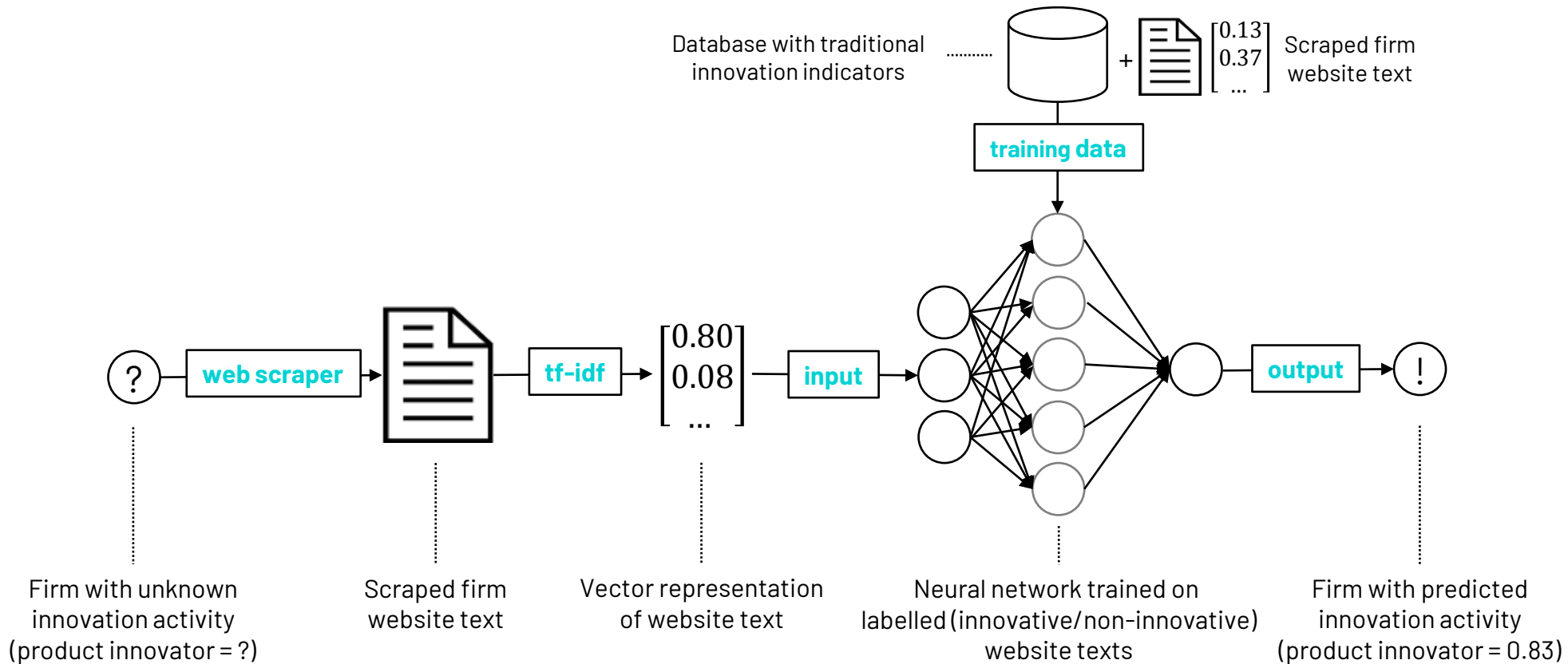


The screenshot shows the ARGUS web scraping tool interface. It has a title bar with 'ARGUS' and standard window controls. The main area is divided into three sections: 'File Settings', 'Web Scraper Settings', and 'Advanced Settings'. 'File Settings' includes a 'Browse for URL list' button and a 'Load Columns' section with dropdowns for 'ID Column\*' and 'URL Column\*'. 'Web Scraper Settings' includes 'Parallel Processes\*' (a dropdown), 'Scrape Limit\*' (a numeric input set to 0), 'Prefer Short URLs' (a dropdown), and 'Preferred Language' (a dropdown). 'Advanced Settings' includes 'Logging Level' (a dropdown set to 'INFO'), 'Download [MB]' (a numeric input set to 10), and 'Timeout [s]' (a numeric input set to 20). At the bottom, there is a 'Start Scraping' button and a 'Functions' section with buttons for 'Stop Scraping', 'Postprocessing', and 'Aggregate Webpage Texts'.

| ID | dl_rank | dl_slot | alias  | error | redirect | start_page | title               | keywords  | description | language  | text  | links   | timestamp               | url                     |
|----|---------|---------|--------|-------|----------|------------|---------------------|---|-------------|---|-------|---|-------------------------|-------------------------|
| 0  | 1       | 0       | zew.de | NaN   | None     | False      | https://www.zew.de/ | ZEW – Leibniz-Zentrum für Europäische Wirtscha... | NaN         | Aktuelle Meldungen, Pressemitteilungen und Inf... | de-DE | [<-p<-] Mit der Digitalisierung verändert sich... | Tue Aug 4 08:42:26 2020 | https://www.zew.de/     |
| 1  | 1       | 1       | zew.de | NaN   | None     | False      | https://www.zew.de/ | ZEW – Leibniz Centre for European Economic Res... | NaN         | Current issues, press releases and information... | en-US | [<-p<-] The economic sentiment in the informat... | Tue Aug 4 08:42:26 2020 | https://www.zew.de/en/  |
| 2  | 1       | 2       | zew.de | NaN   | None     | False      | https://www.zew.de/ | Mitarbeiterinnen und Mitarbeiter des ZEW Mitar... | NaN         |   | de-DE | [<-p<-] Das ZEW – Leibniz-Zentrum für Europäis... | Tue Aug 4 08:42:26 2020 | https://www.zew.de/team |

# InnoProb

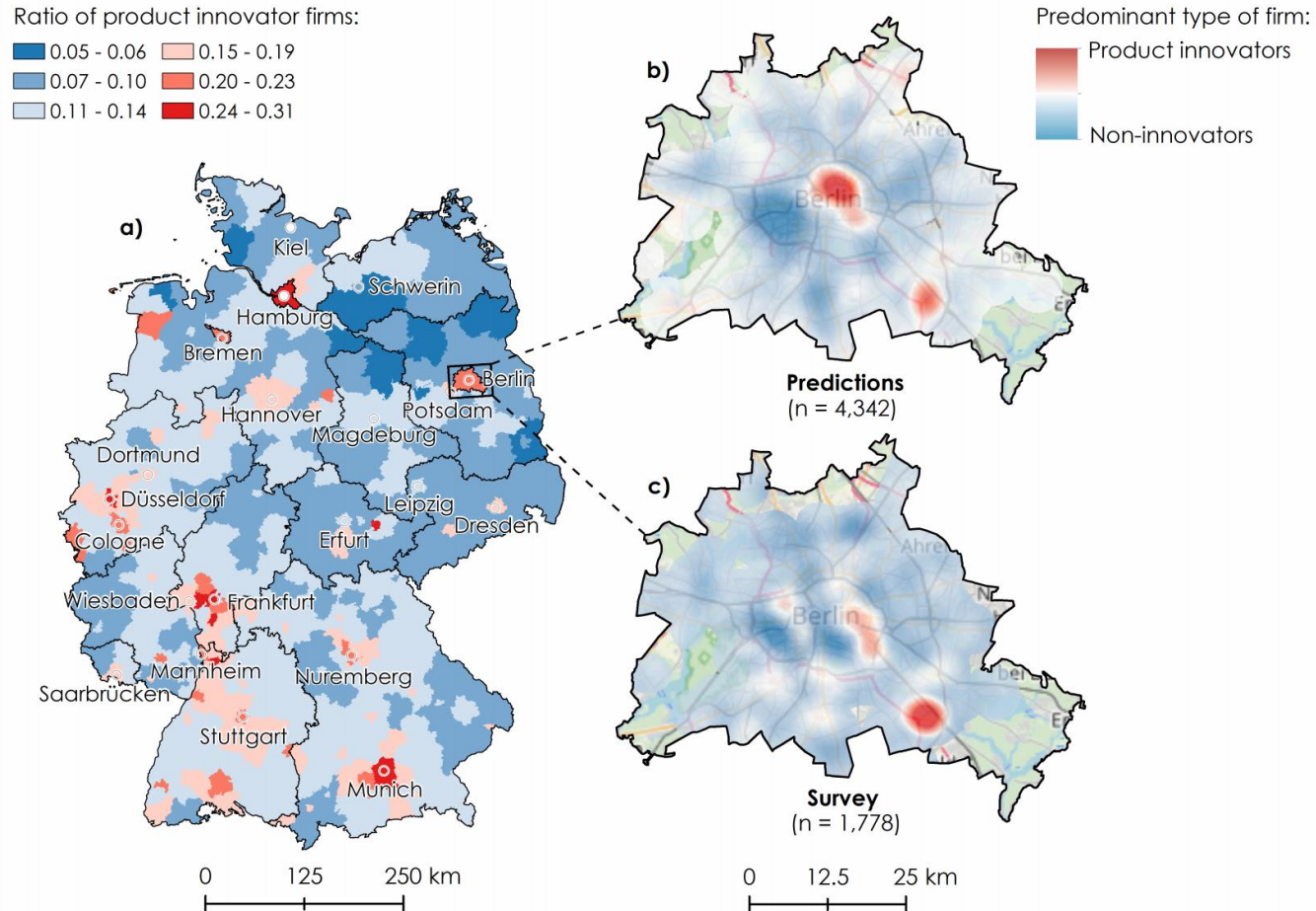
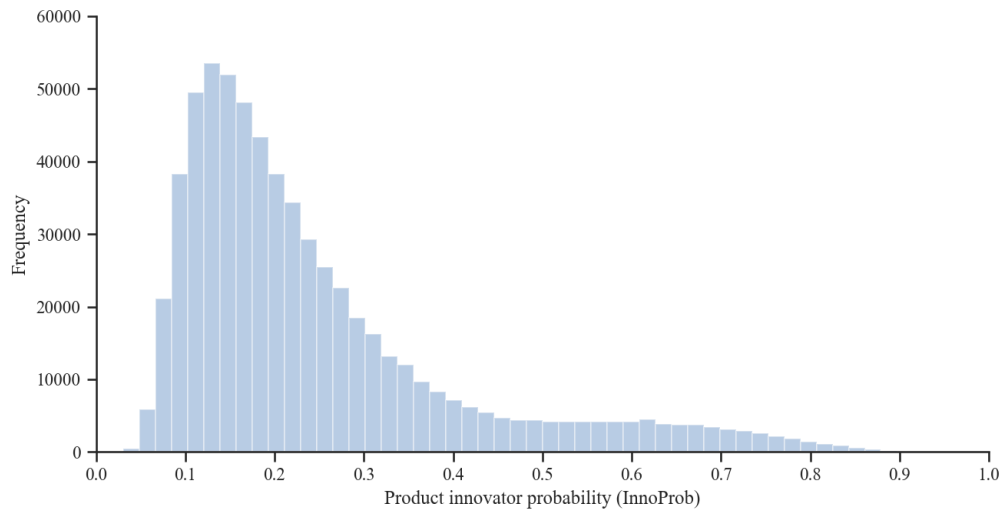
A deep learning model for **predicting product innovator firms** based on their website texts.





# InnoProb

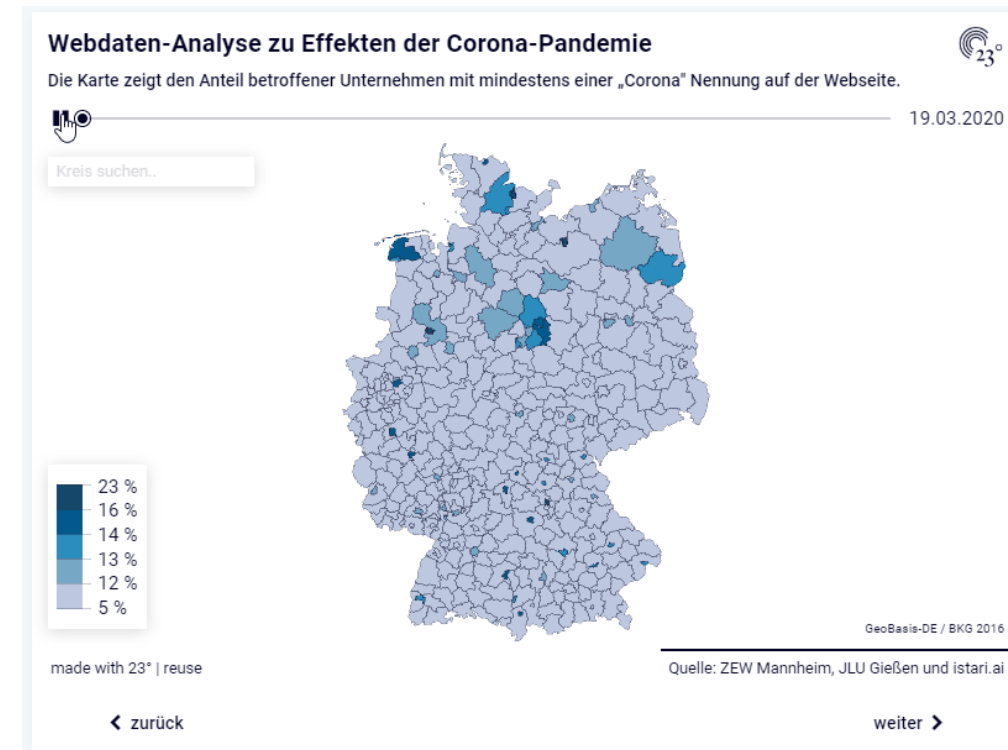
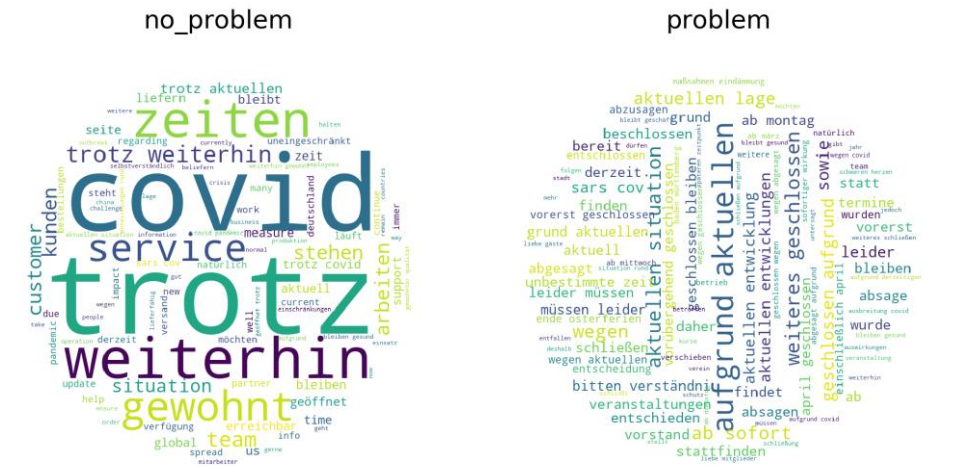
We use InnoProb to calculate **product innovator probabilities** for 600,000 German firms with websites. Robustness tests using traditional indicators.



# Pandemic analysis

Web-based analysis for the German Research Ministry on the **effect of the coronavirus pandemic on German firms**.

- Pandemic requires quick, yet evidence-based actions.
- Traditional data gathering procedures way too slow.
- webAI scans all German firm websites 2x per week.
- Identification and classification of “coronavirus” references.



## More...

---

Check out our website: [www.istari.ai](http://www.istari.ai)

Follow us on Twitter: [@istari\\_ai](https://twitter.com/istari_ai)

Follow me on Twitter: [@jan\\_kinne](https://twitter.com/jan_kinne)

Check out our papers:

- Paper on web mining basics: <https://link.springer.com/article/10.1007/s11192-020-03726-9>
- InnoProb paper: <http://ftp.zew.de/pub/zew-docs/dp/dp19001.pdf>
- Digital Layer B2B hyperlink paper: <http://ftp.zew.de/pub/zew-docs/dp/dp20003.pdf>
- Certification adoption paper: <https://ieeexplore.ieee.org/abstract/document/908286>
- Coronavirus pandemic report: [http://ftp.zew.de/pub/zew-docs/ZEWKurzexpertisen/ZEW\\_Kurzexpertise2005.pdf](http://ftp.zew.de/pub/zew-docs/ZEWKurzexpertisen/ZEW_Kurzexpertise2005.pdf)