Leveraging web data to inform Industrial Classifications and capturing economic dynamics

Veronika Vilgis, PhD
Dataprovider.com

Crawling the web

Monthly updates on:

> 600 million domains

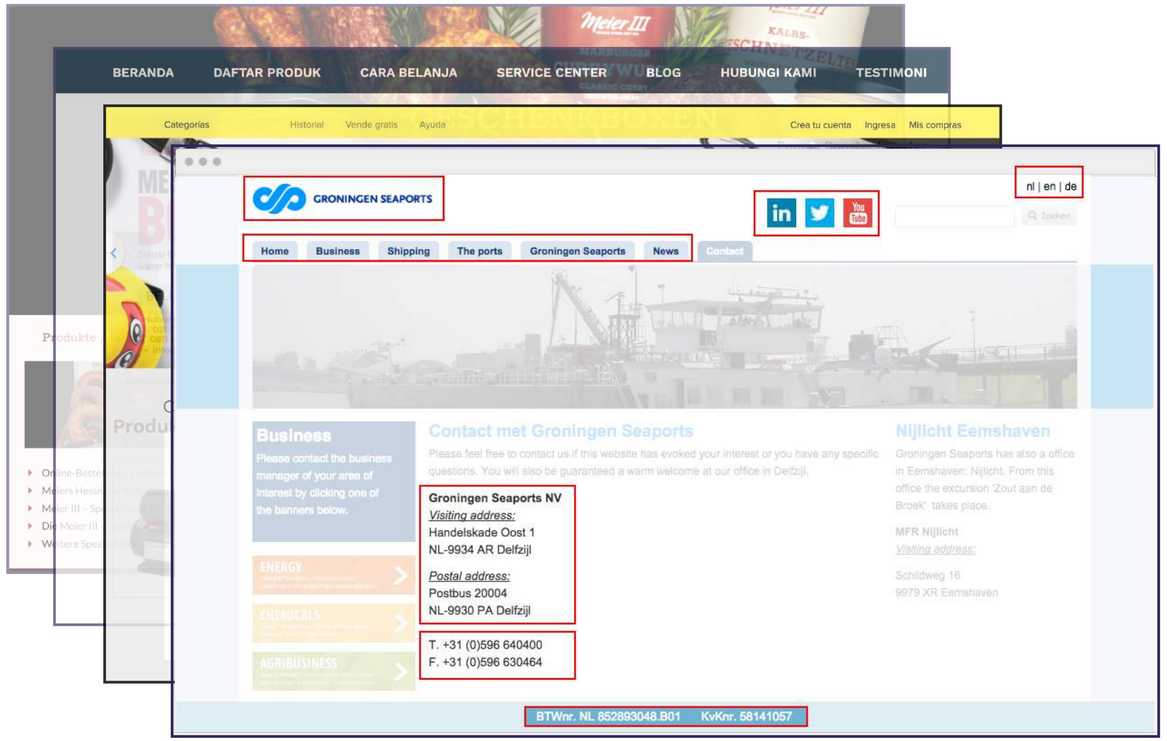
> 25 million company websites

> 200 data fields

> 4 years of historical web data

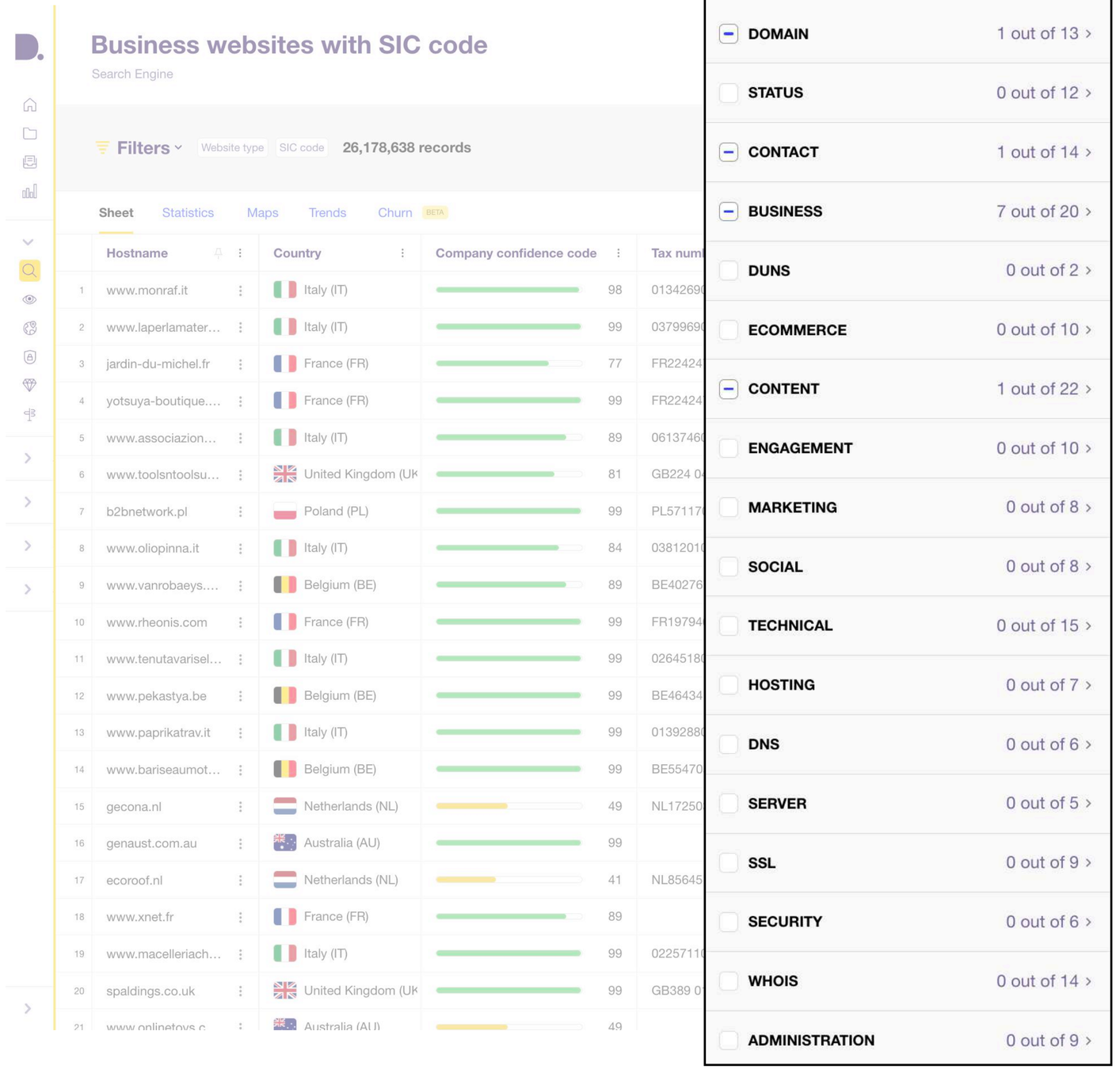
> 5 unique proprietary scores

> 50 countries

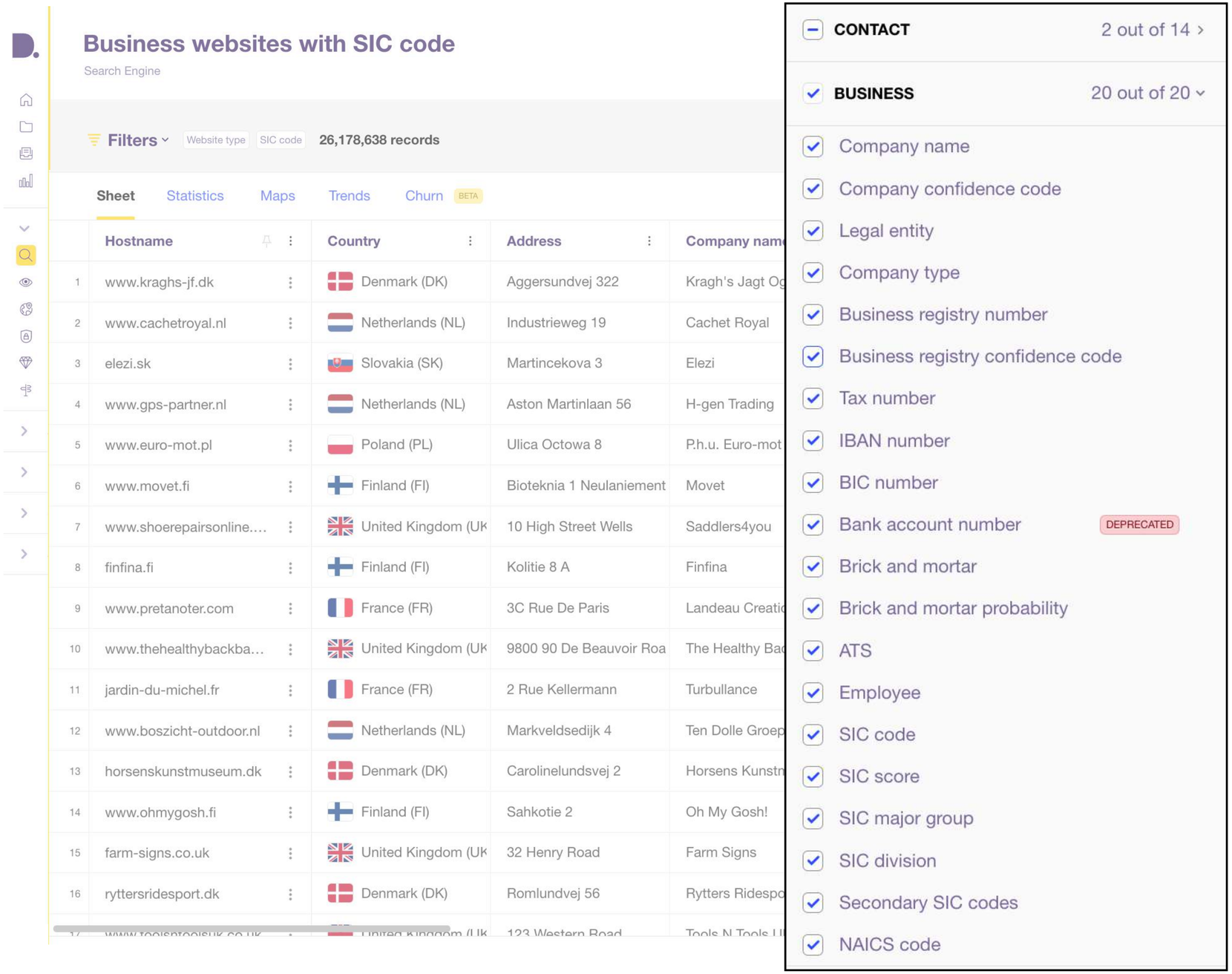


“A business registry based on the internet.”

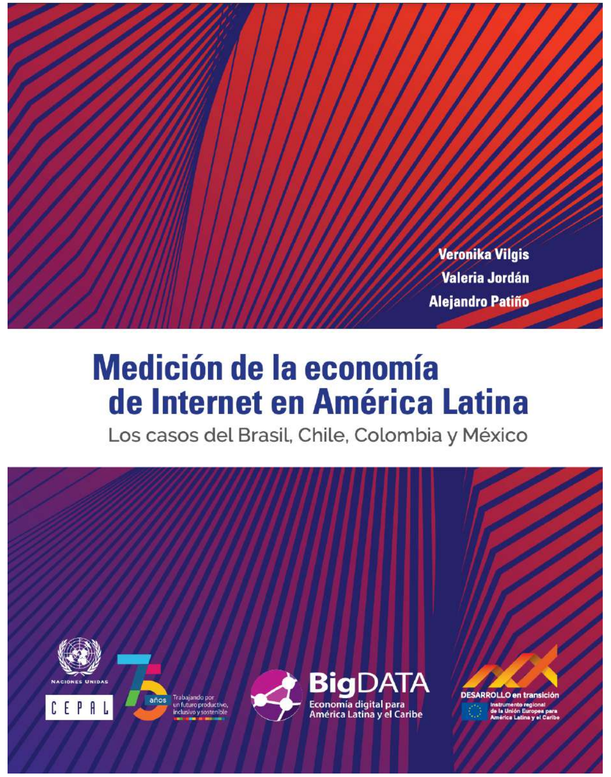
> 200 variables per websites



> business information

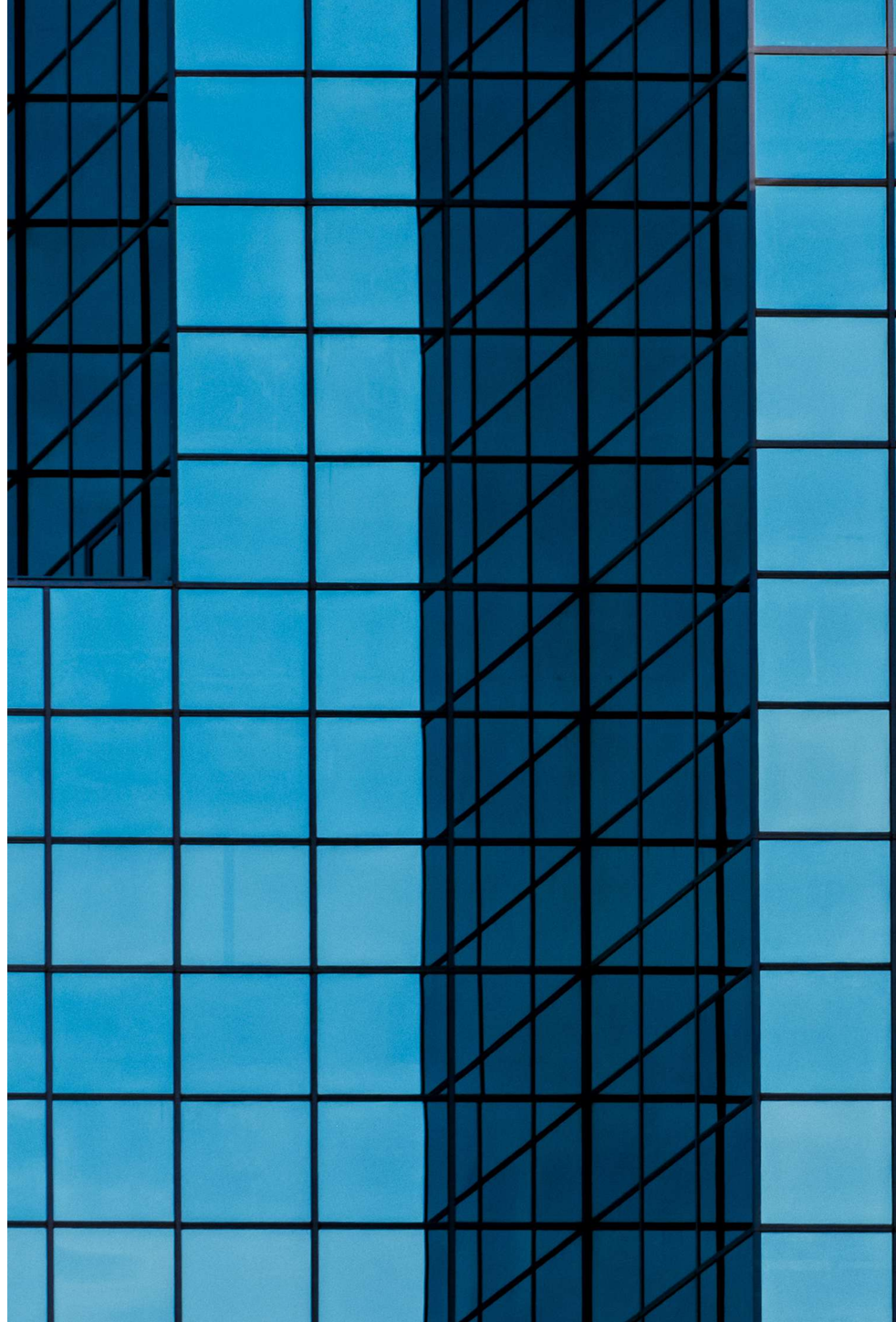


Merging web data with business registry data



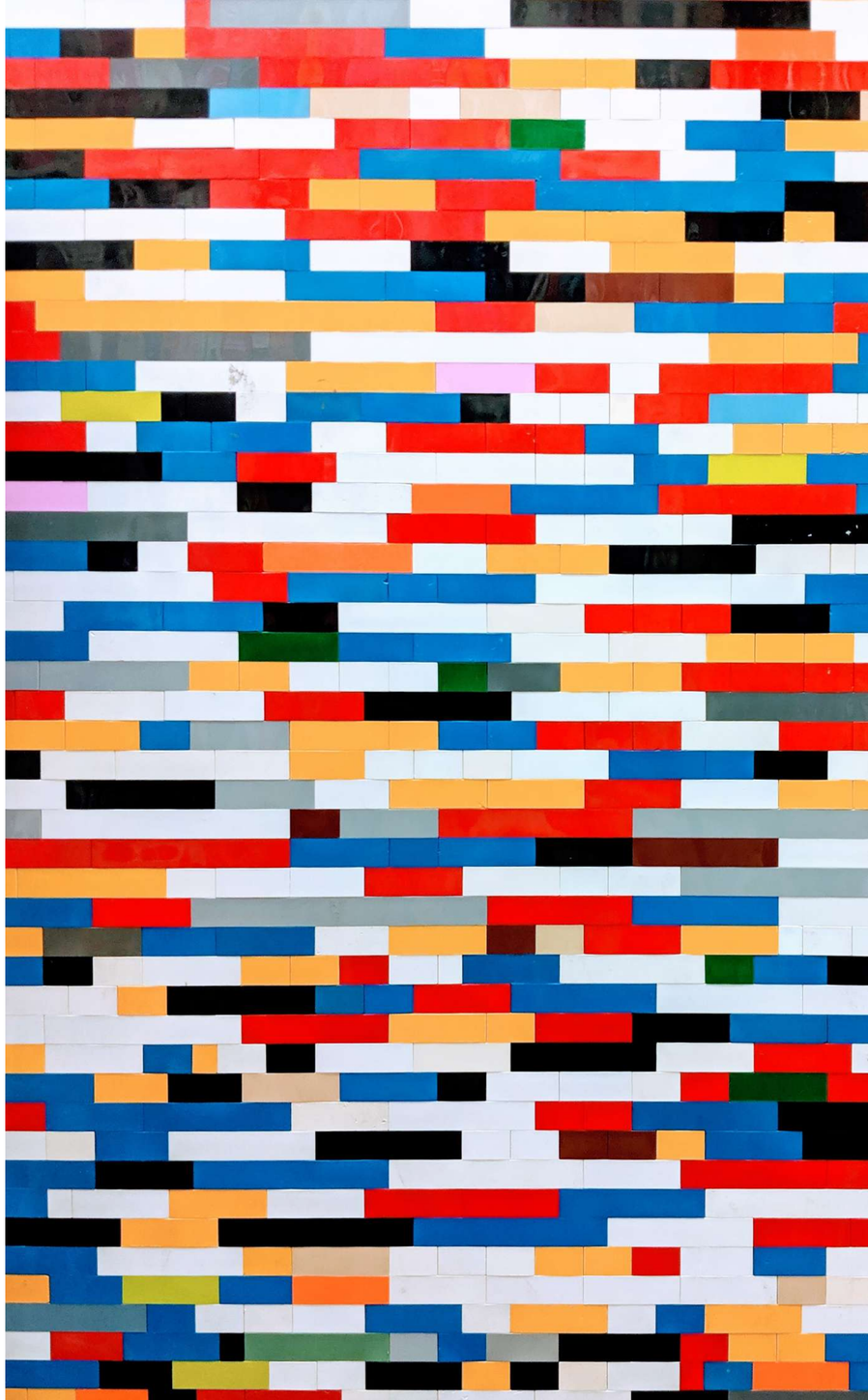
Classifications of websites

- Website type
- Online store
  - Shopping cart, payment methods, delivery methods
- Brick and mortar
- SIC



SIC Classification: Methodology

- Input fields are:
  - HTML
  - Language
  - Country
- Vectorizing the content:
  - HTML parsing and text extraction
  - Vectorisation: determining the most relevant and distinctive words
- Training the classifiers
  - SIC from Dun & Bradstreet
  - Multilayer perceptron (MLP), artificial neural network to train the relationship between the vector of the website content and the SIC code.
  - Separate classifiers (and vectorizers) for different languages



SIC Division

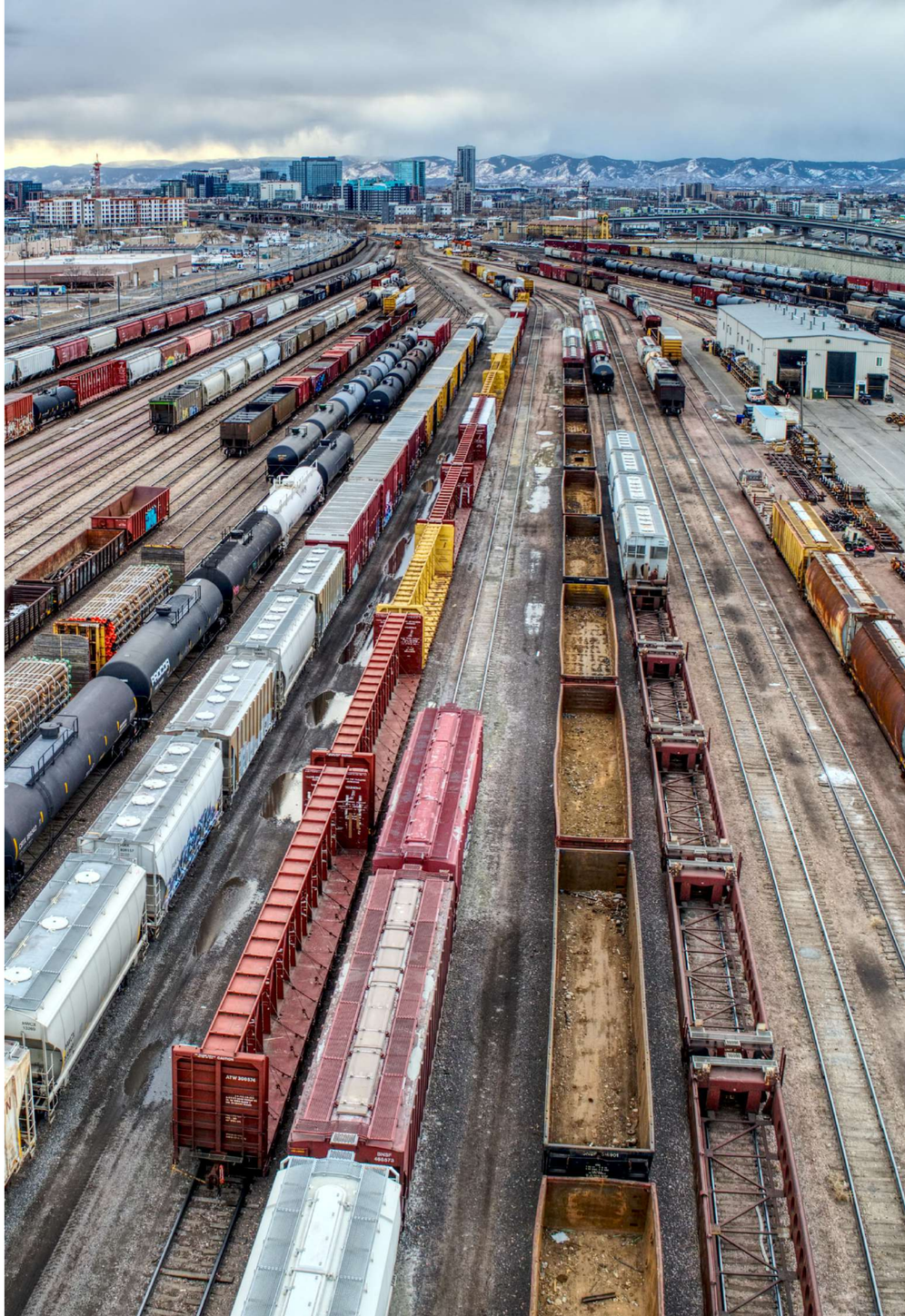
Services	(15,731,495)	**60.1%**	Manufacturing	(1,326,845)	**5.1%**
Retail Trade	(3,123,436)	**11.9%**	Transportation, Communications, Electric, Gas a…	(758,160)	**2.9%**
Wholesale Trade	(1,576,826)	**6%**	Agriculture, Forestry and Fishing	(519,170)	**2%**
Construction	(1,550,919)	**5.9%**	Public Administration	(137,459)	**0.5%**
Finance, Insurance and Real Estate	(1,444,198)	**5.5%**	Mining	(10,130)	**< 0.01%**

SIC Major Group

Business Services	(5,649,326)	21.6%	Wholesale Trade-Durable Goods	(984,828)	3.8%
Engineering, Accounting, Research, Manageme…	(2,118,269)	8.1%	Real Estate	(979,898)	3.7%
Membership Organizations	(1,535,179)	5.9%	Personal Services	(871,458)	3.3%
Health Services	(1,468,495)	5.6%	Educational Services	(714,592)	2.7%
Amusement and Recreation Services	(1,217,798)	4.7%	Wholesale Trade-Nondurable Goods	(591,998)	2.3%
Miscellaneous Retail	(1,162,534)	4.4%	Hotels, Rooming Houses, Camps, and other Lod…	(502,632)	1.9%
Construction Special Trade Contractors	(1,156,062)	4.4%	Social Services	(442,300)	1.7%
Eating and Drinking Places	(1,059,250)	4%			

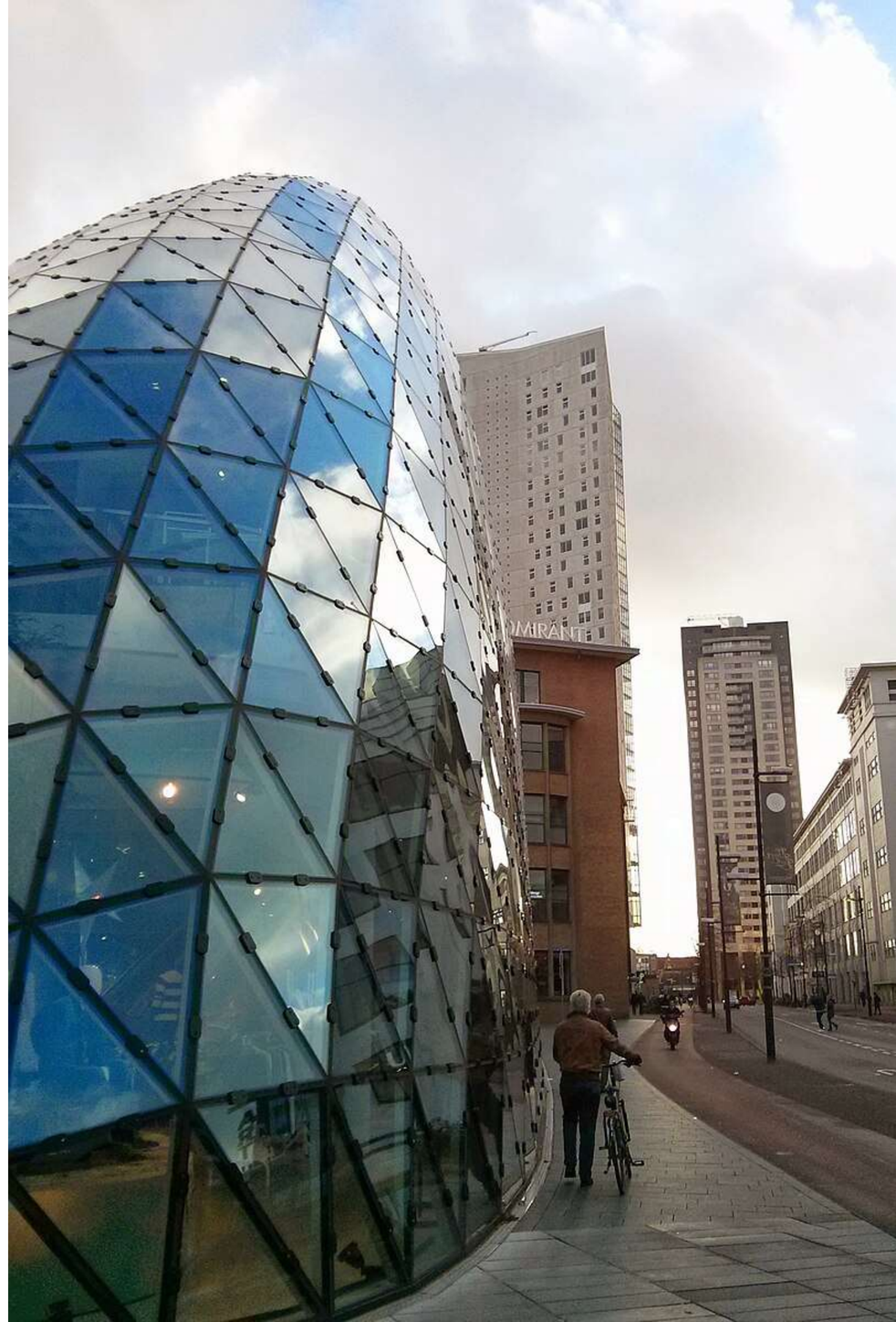
Internet categories

- Adult
- Advertising & Marketing
- E-commerce and Shopping
- Entertainment
- Information & Education
- Infrastructure & Telecom
- Jobs and Career
- Marketplace & Sharing economy
- Online Service/Product
- Social Service/Product
- Software Development
- Web/Cloud Hosting
- Website Creator



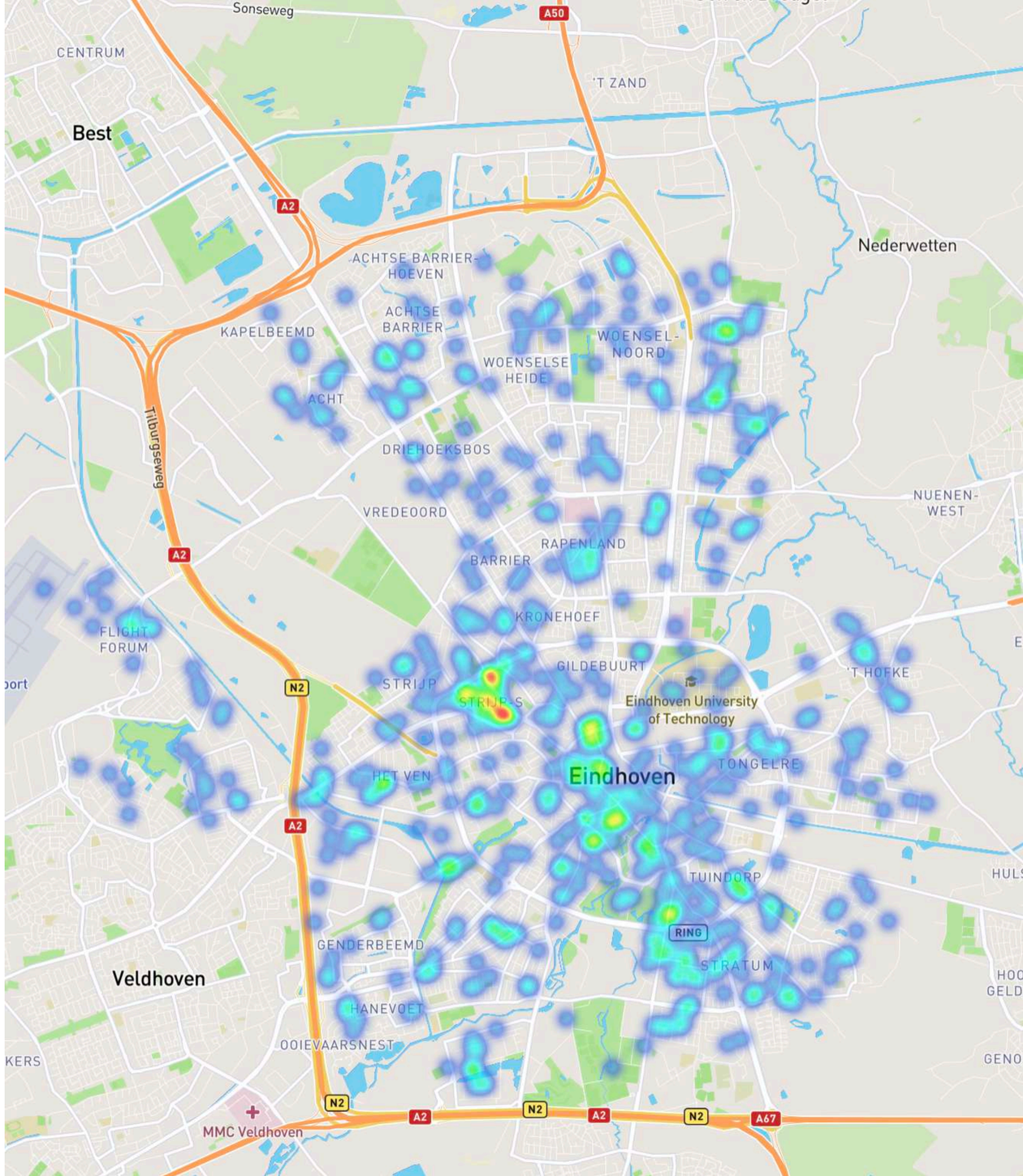
Case Study

- City of Eindhoven
- Tech hub
- SIC's Major Groups
  - Business Services (73) and
  - Engineering, Accounting, Research, Management, and Related Services (87)
- 3000 business websites
- Topic modeling



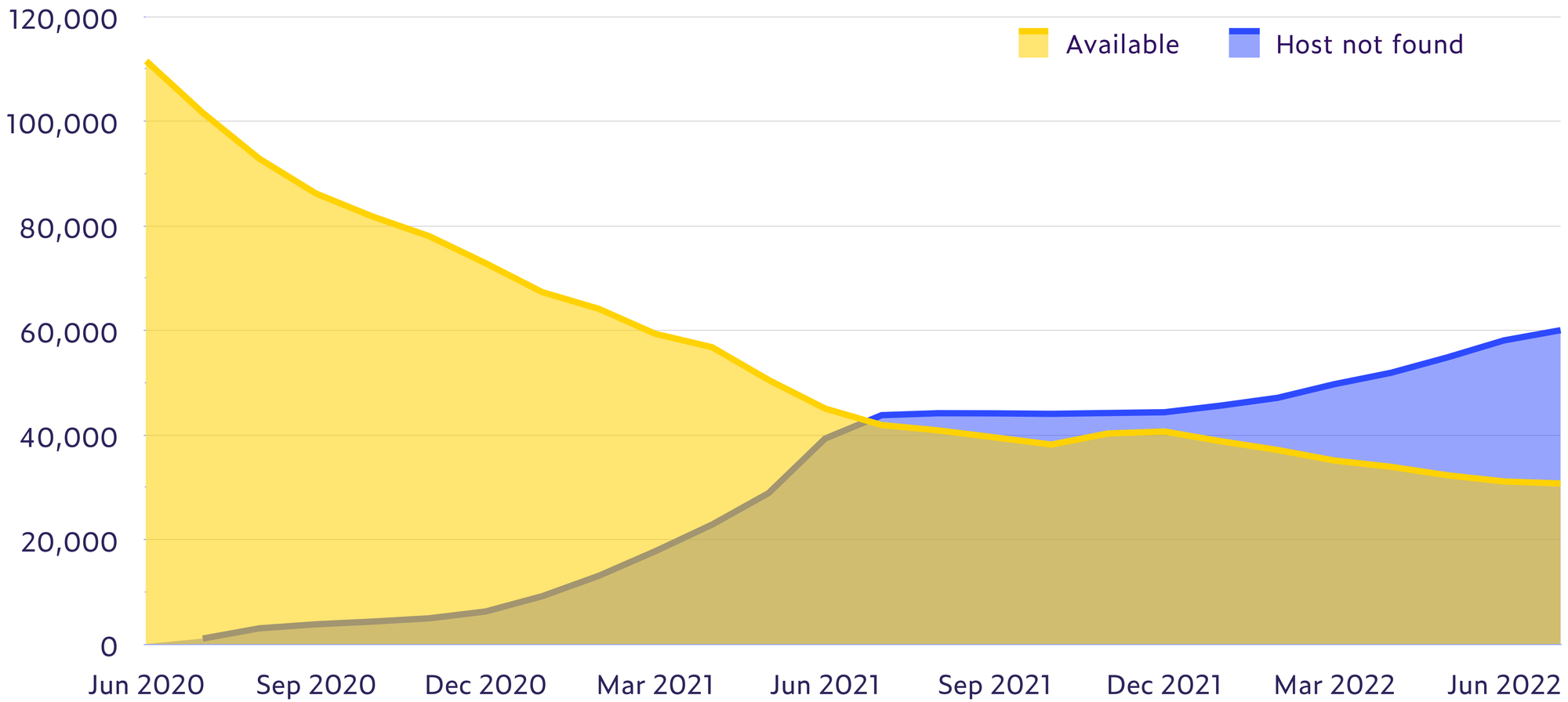
Results

- Topics of interest
  - Digital marketing and SEO optimisation services
  - Services relating to green energy and sustainable construction
  - Services offering coach workshops for personal development



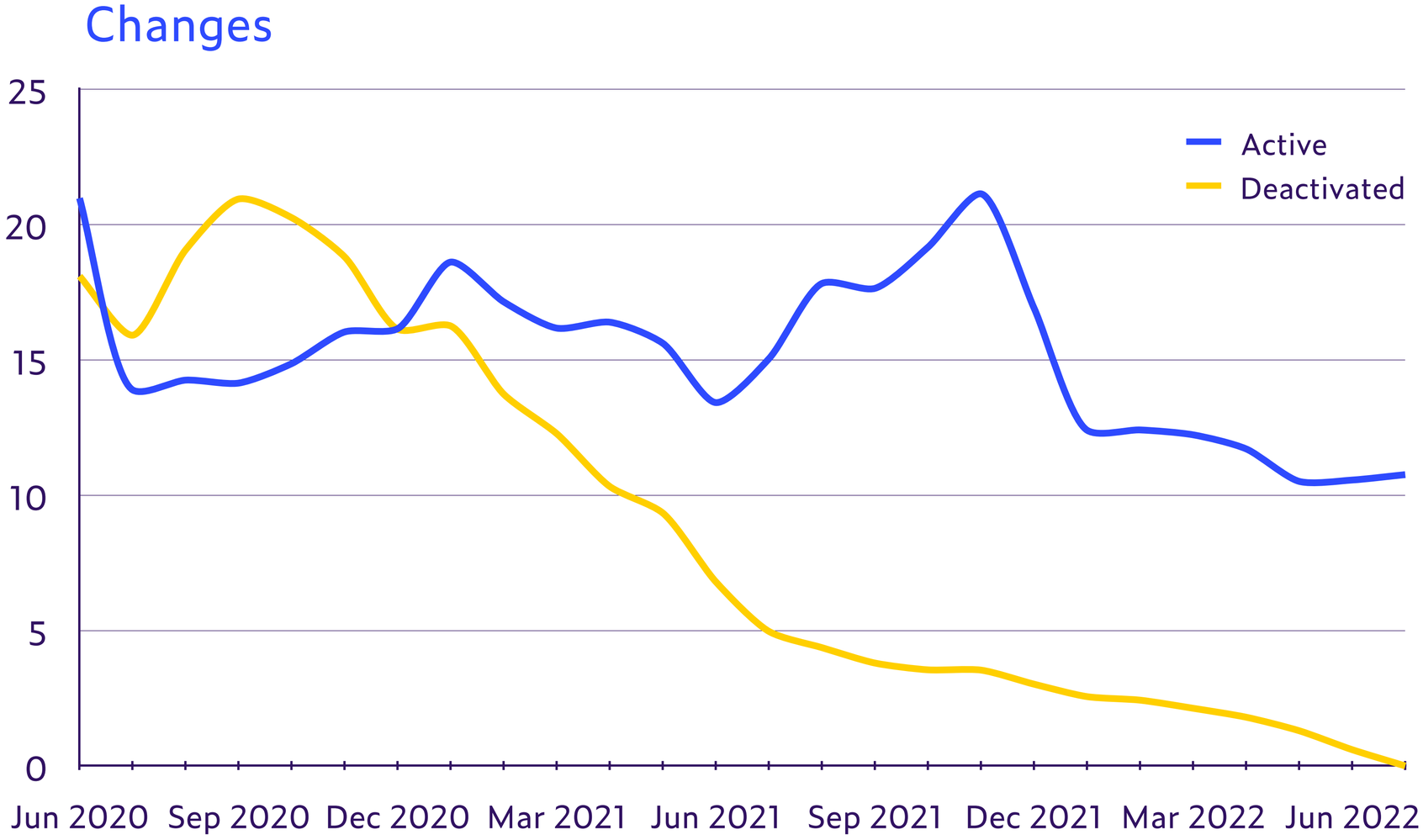
Number of US eCommerce sites by returned response:

June 2020 – July 2022



Source: Dataprovider.com, September 2022

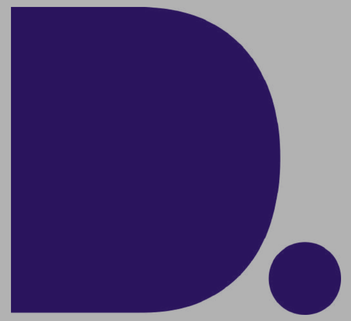
Performance of ‘winners’ and ‘losers’ over time



Source: Dataprovider.com, September 2022

Summary

- Existing high-level categories
- Data driven approach to identify topics / subcategories
- Monitor over time and across geographies
- Discover new industries as they are emerging



Control tomorrow

Dataprovider | Groningen
Helperpark 292
9723 ZA Groningen
The Netherlands

Dataprovider | Amsterdam
Weesperplein 4b
1018 XA Amsterdam
The Netherlands