

Leveraging web data to inform Industrial Classifications and capturing economic dynamics

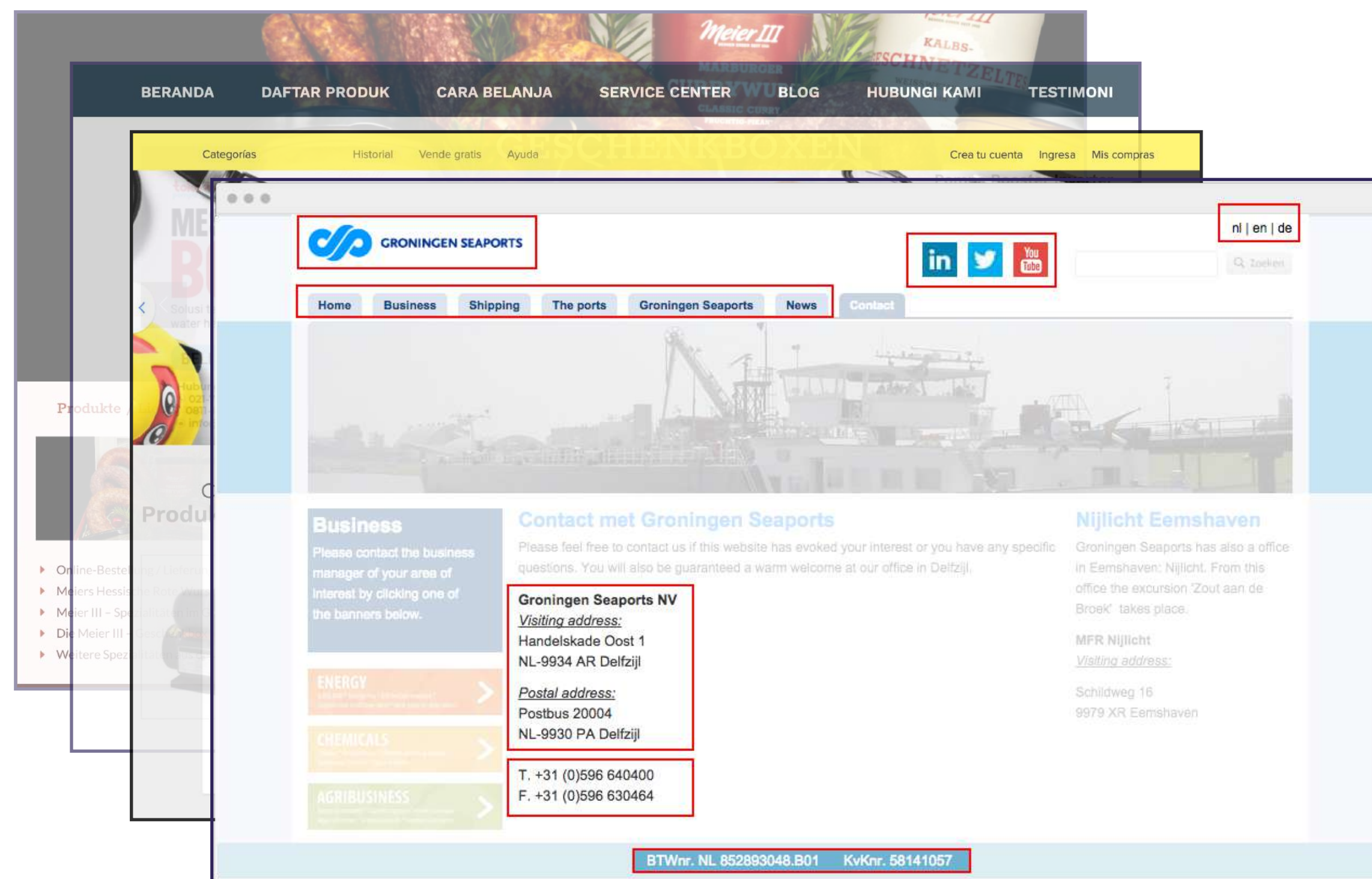
Veronika Vilgis, PhD
Dataprovider.com



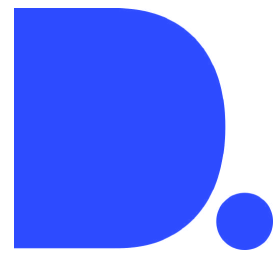
Crawling the web

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 websites with SIC code

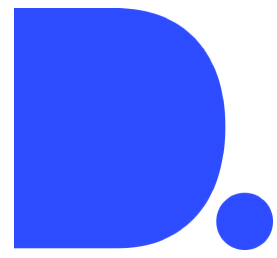
Search Engine

Filters Website type SIC code 26,178,638 records

Sheet Statistics Maps Trends Churn BETA

	Hostname	Country	Company confidence code	Tax num
1	www.monraf.it	Italy (IT)	98	01342690
2	www.laperlamater...	Italy (IT)	99	03799690
3	jardin-du-michel.fr	France (FR)	77	FR22424
4	yotsuya-boutique...	France (FR)	99	FR22424
5	www.associazion...	Italy (IT)	89	06137460
6	www.toolsntoolsu...	United Kingdom (UK)	81	GB224 0
7	b2bnetwork.pl	Poland (PL)	99	PL57117
8	www.oliopinna.it	Italy (IT)	84	03812010
9	www.vanrobaeys...	Belgium (BE)	89	BE40276
10	www.rheonis.com	France (FR)	99	FR19794
11	www.tenutavarisel...	Italy (IT)	99	02645180
12	www.pekastya.be	Belgium (BE)	99	BE46434
13	www.paprikatrav.it	Italy (IT)	99	01392880
14	www.bariseaumot...	Belgium (BE)	99	BE55470
15	gecona.nl	Netherlands (NL)	49	NL17250
16	genaust.com.au	Australia (AU)	99	
17	ecorooft.nl	Netherlands (NL)	41	NL85645
18	www.xnet.fr	France (FR)	89	
19	www.macelleriach...	Italy (IT)	99	02257110
20	spaldings.co.uk	United Kingdom (UK)	99	GB389 0
21	www.onlinetovs.c...	Australia (AU)	49	

- DOMAIN 1 out of 13 >
- STATUS 0 out of 12 >
- CONTACT 1 out of 14 >
- BUSINESS 7 out of 20 >
- DUNS 0 out of 2 >
- ECOMMERCE 0 out of 10 >
- CONTENT 1 out of 22 >
- ENGAGEMENT 0 out of 10 >
- MARKETING 0 out of 8 >
- SOCIAL 0 out of 8 >
- TECHNICAL 0 out of 15 >
- HOSTING 0 out of 7 >
- DNS 0 out of 6 >
- SERVER 0 out of 5 >
- SSL 0 out of 9 >
- SECURITY 0 out of 6 >
- WHOIS 0 out of 14 >
- ADMINISTRATION 0 out of 9 >



> business information

Business websites with SIC code

Search Engine

Filters Website type SIC code 26,178,638 records

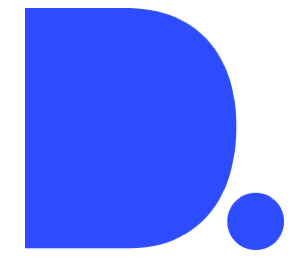
Sheet Statistics Maps Trends Churn BETA

	Hostname	Country	Address	Company name
1	www.kraghs-jf.dk	Denmark (DK)	Aggersundvej 322	Kragh's Jagt Og
2	www.cachetroyal.nl	Netherlands (NL)	Industrieweg 19	Cachet Royal
3	elezi.sk	Slovakia (SK)	Martincekova 3	Elezi
4	www.gps-partner.nl	Netherlands (NL)	Aston Martinlaan 56	H-gen Trading
5	www.euro-mot.pl	Poland (PL)	Ulica Octowa 8	Ph.u. Euro-mot
6	www.movet.fi	Finland (FI)	Bioteknia 1 Neulaniement	Movet
7	www.shoerepaironline....	United Kingdom (UK)	10 High Street Wells	Saddlers4you
8	finfina.fi	Finland (FI)	Kolitie 8 A	Finfina
9	www.pretanoter.com	France (FR)	3C Rue De Paris	Landeau Creatio
10	www.thehealthybackba...	United Kingdom (UK)	9800 90 De Beauvoir Roa	The Healthy Bac
11	jardin-du-michel.fr	France (FR)	2 Rue Kellermann	Turbulance
12	www.boszicht-outdoor.nl	Netherlands (NL)	Markveldsedijk 4	Ten Dolle Groep
13	horsenskunstmuseum.dk	Denmark (DK)	Carolinelundsvej 2	Horsens Kunstr
14	www.ohmygosh.fi	Finland (FI)	Sahkotie 2	Oh My Gosh!
15	farm-signs.co.uk	United Kingdom (UK)	32 Henry Road	Farm Signs
16	ryttersridesport.dk	Denmark (DK)	Romlundvej 56	Rytters Ridespo
17	www.toolsn...co.uk	United Kingdom (UK)	123 Western Road	Tools N Tools LI

CONTACT 2 out of 14 >

BUSINESS 20 out of 20 v

- Company name
- Company confidence code
- Legal entity
- Company type
- Business registry number
- Business registry confidence code
- Tax number
- IBAN number
- BIC number
- Bank account number DEPRECATED
- Brick and mortar
- Brick and mortar probability
- ATS
- Employee
- SIC code
- SIC score
- SIC major group
- SIC division
- Secondary SIC codes
- NAICS code



Merging web data with business registry data

Discussion Paper

Measuring the internet economy in The Netherlands: a big data analysis

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2016 | 14

Lotte Oostrom
Adam N. Walker
Bart Staats
Magda Slootbeek-Van Laar
Shirley Ortega Azurduy
Bastiaan Rooijackers

Paper

Measuring the internet economy in the Netherlands 2016-2018

A big data analysis

Lotte Oostrom
Jaap Jansen
Raymond Kleingeld
Andries Kuipers
Magda Slootbeek-Van Laar
Joram Vuik
Adam N. Walker

May 2020

Veronika Vilgis
Valeria Jordán
Alejandro Patiño

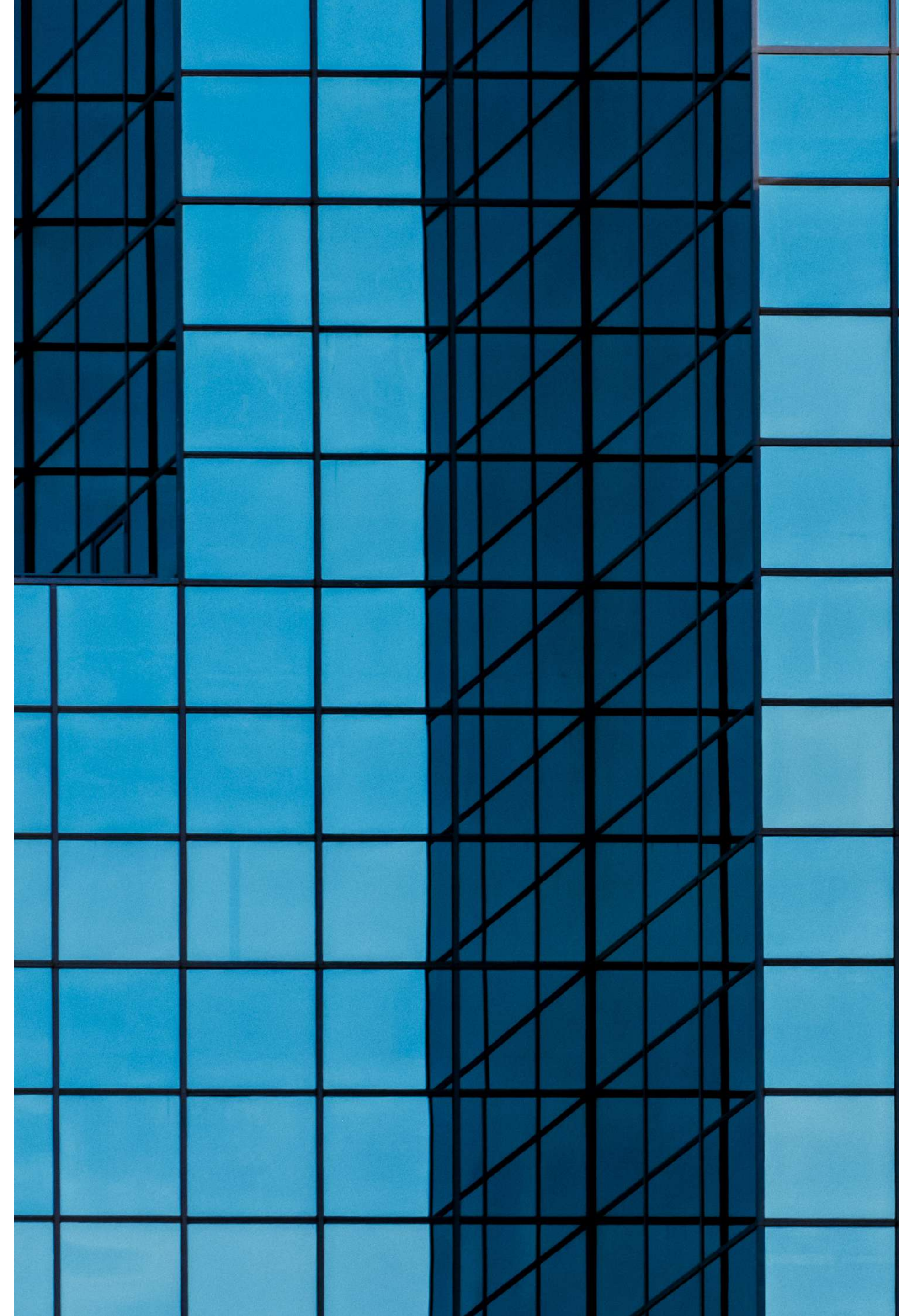
Medición de la economía de Internet en América Latina

Los casos del Brasil, Chile, Colombia y México

CEPAL
75 años
Trabajando por un futuro productivo, inclusivo y sostenible
BigDATA
Economía digital para América Latina y el Caribe
DESARROLLO en transición
Instrumento regional de la Unión Europea para América Latina y el Caribe

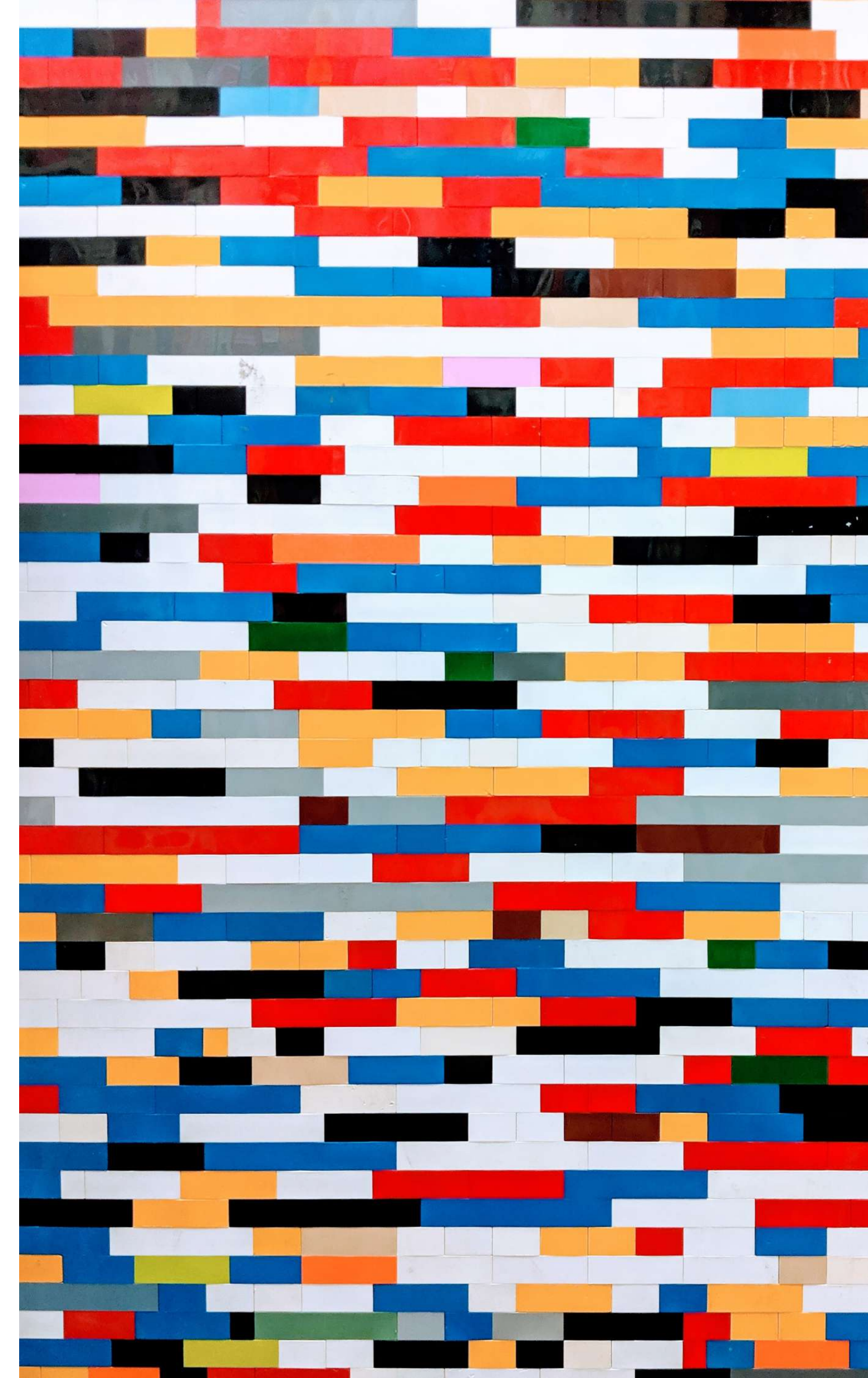
Classifications of websites

- 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

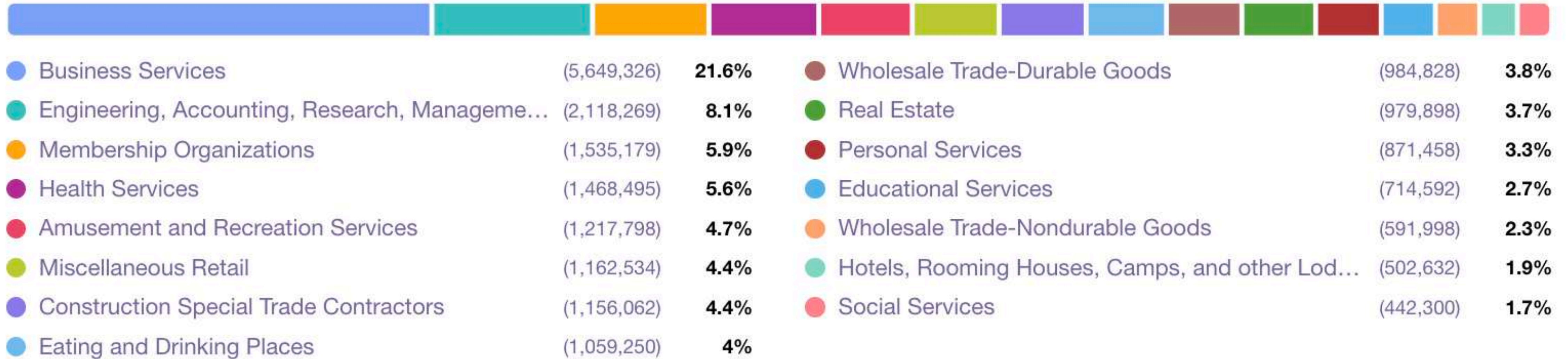


D. SIC Division



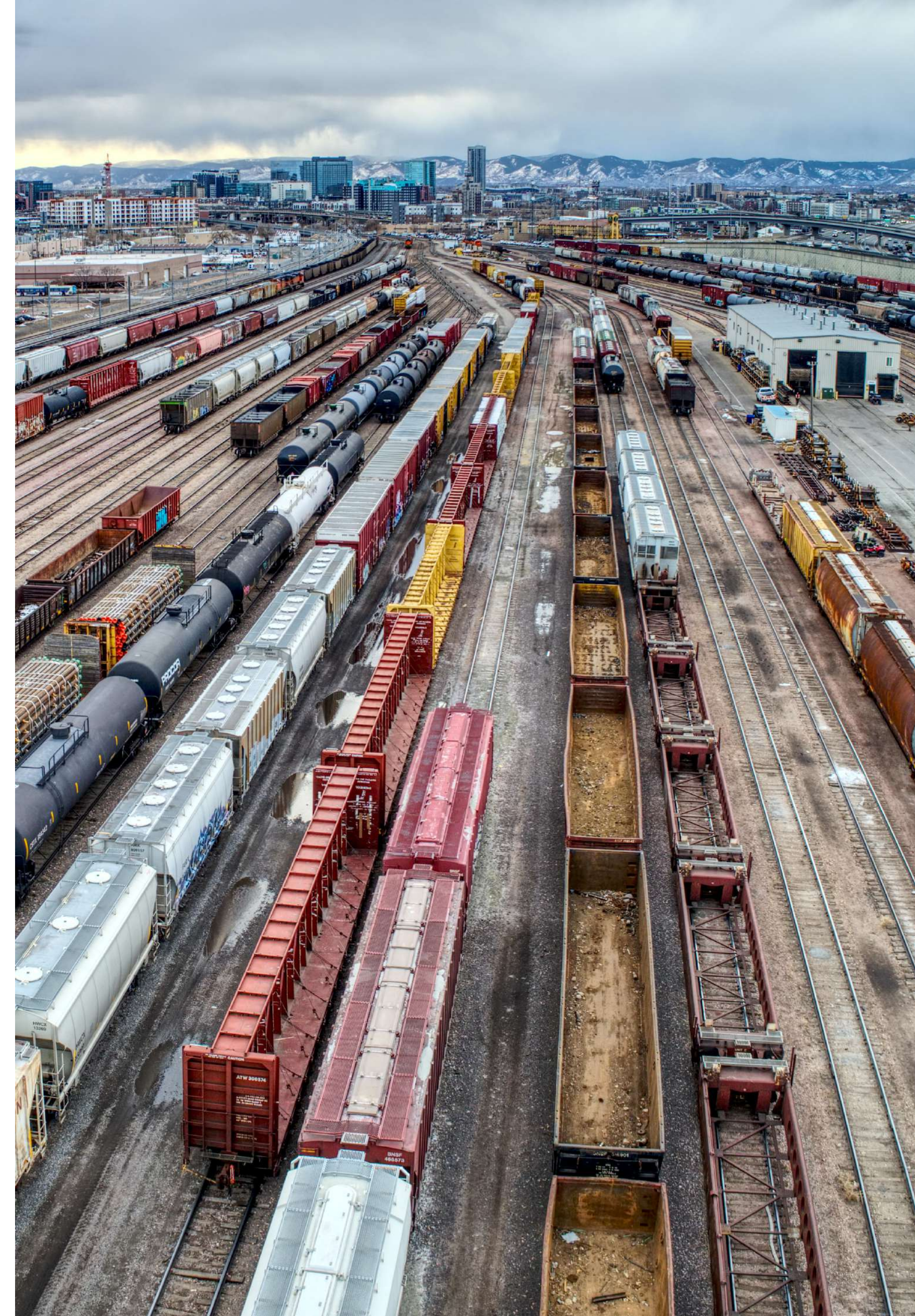
Services	(15,731,495)	60.1%	Manufacturing	(1,326,845)	5.1%
Retail Trade	(3,123,436)	11.9%	Transportation, Communications, Electric, Gas a...	(758,160)	2.9%
Wholesale Trade	(1,576,826)	6%	Agriculture, Forestry and Fishing	(519,170)	2%
Construction	(1,550,919)	5.9%	Public Administration	(137,459)	0.5%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

SIC Major Group



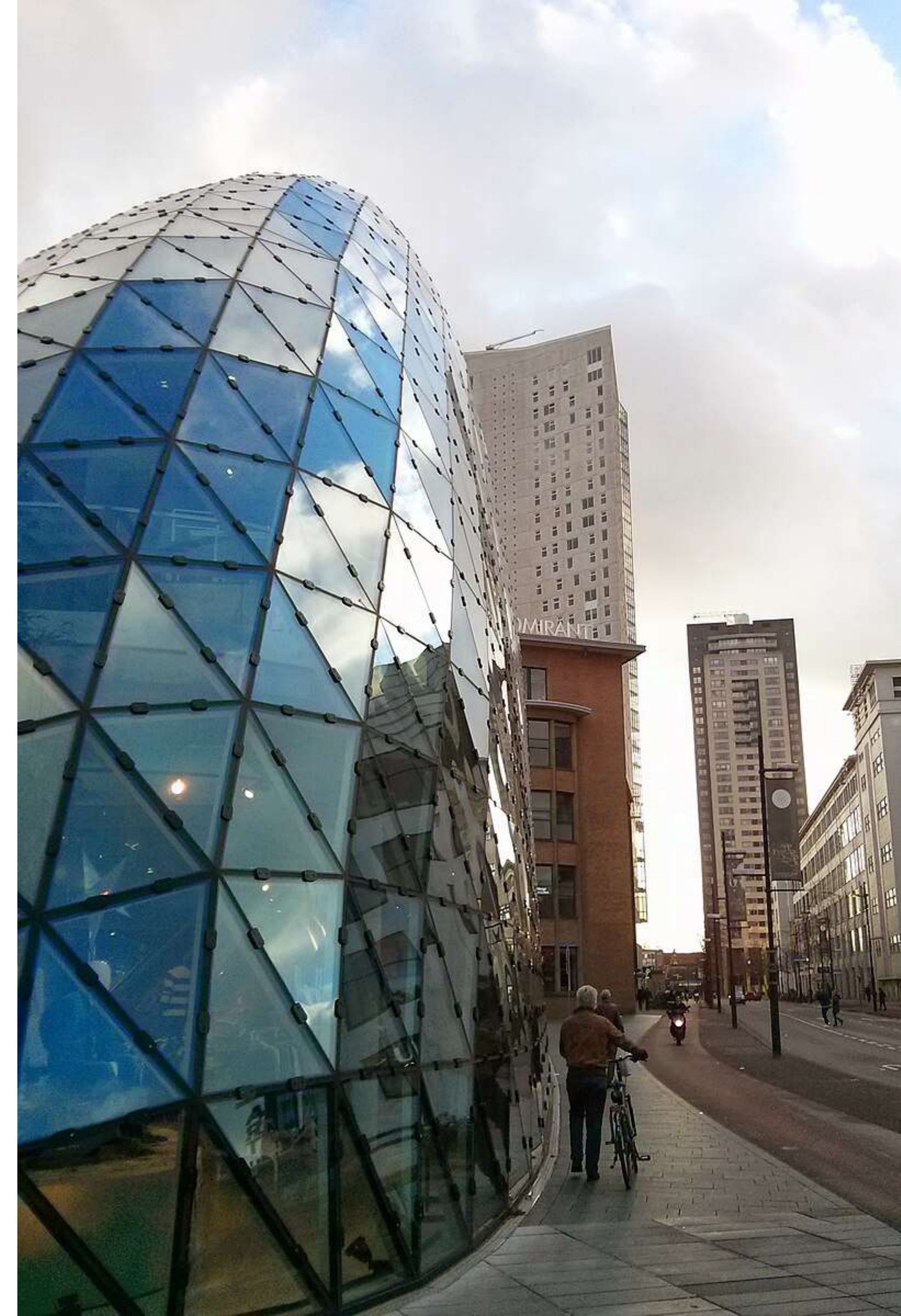
Internet categories

- 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



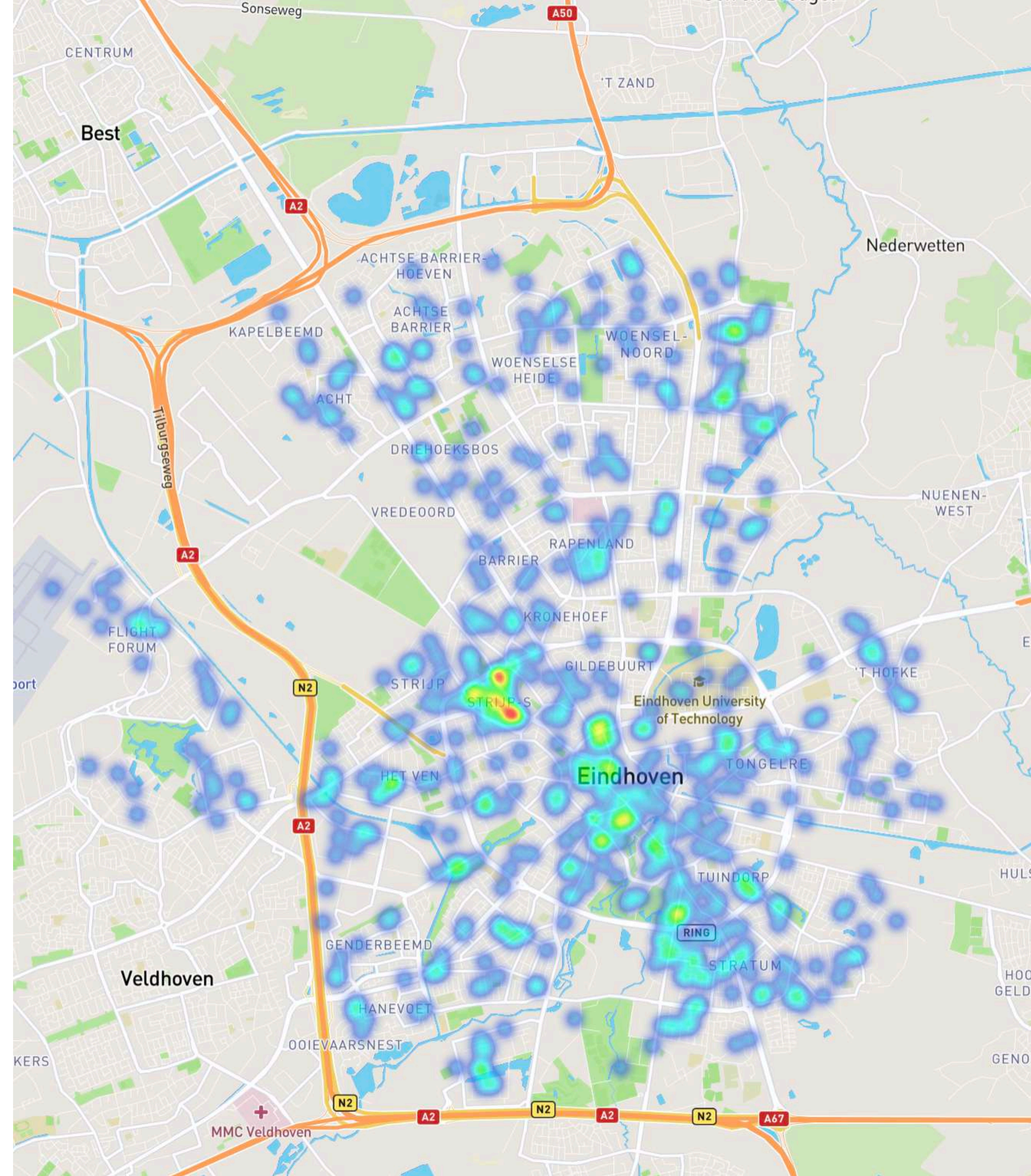
Case Study

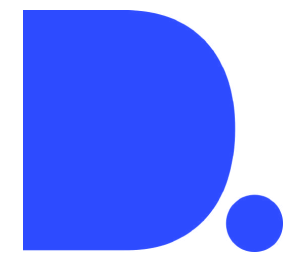
- City of Eindhoven
- Tech hub
- SIC's Major Groups
 - Business Services (73) and
 - Engineering, Accounting, Research, Management, and Related Services (87)
- 3000 business websites
- Topic modeling



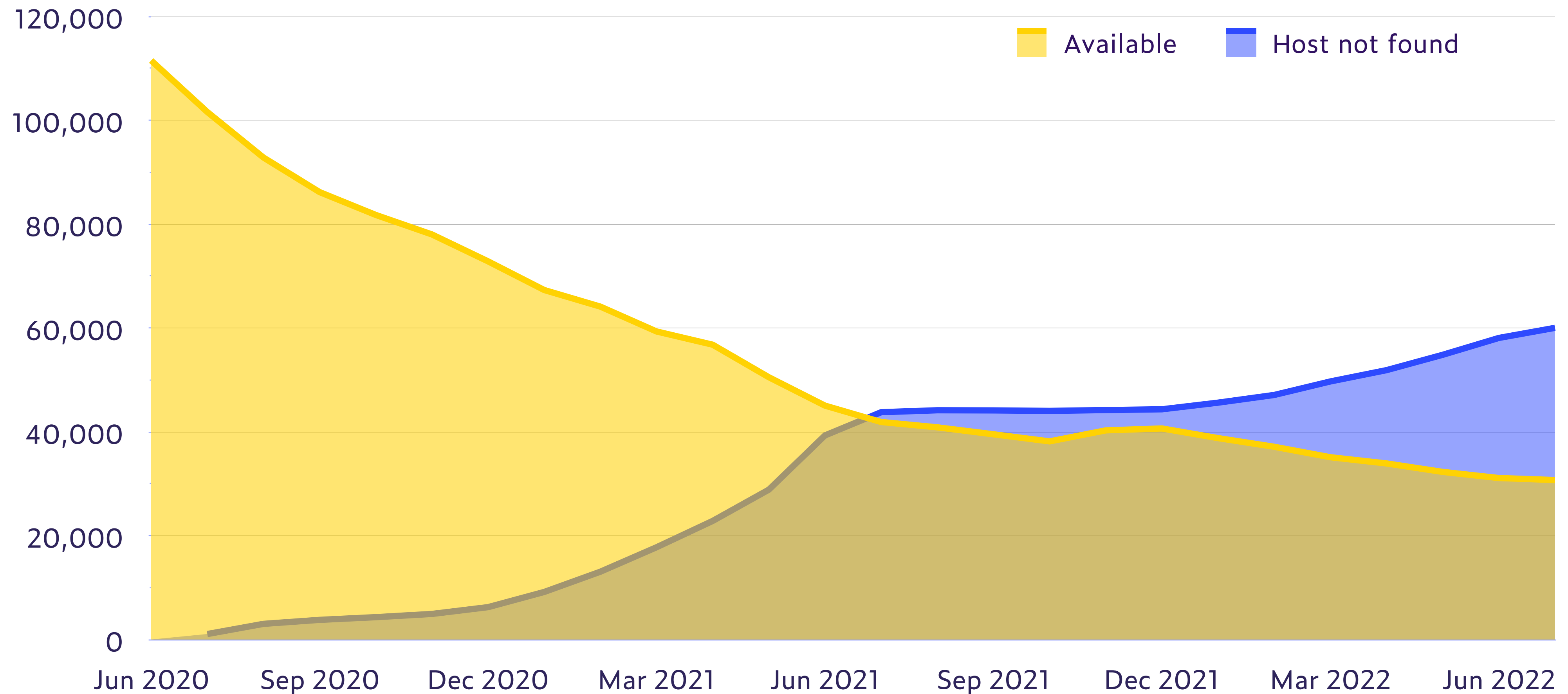
Results

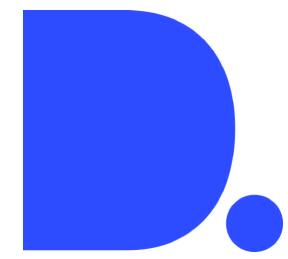
- Topics of interest
 - Digital marketing and SEO optimisation services
 - Services relating to green energy and sustainable construction
 - Services offering coach workshops for personal development



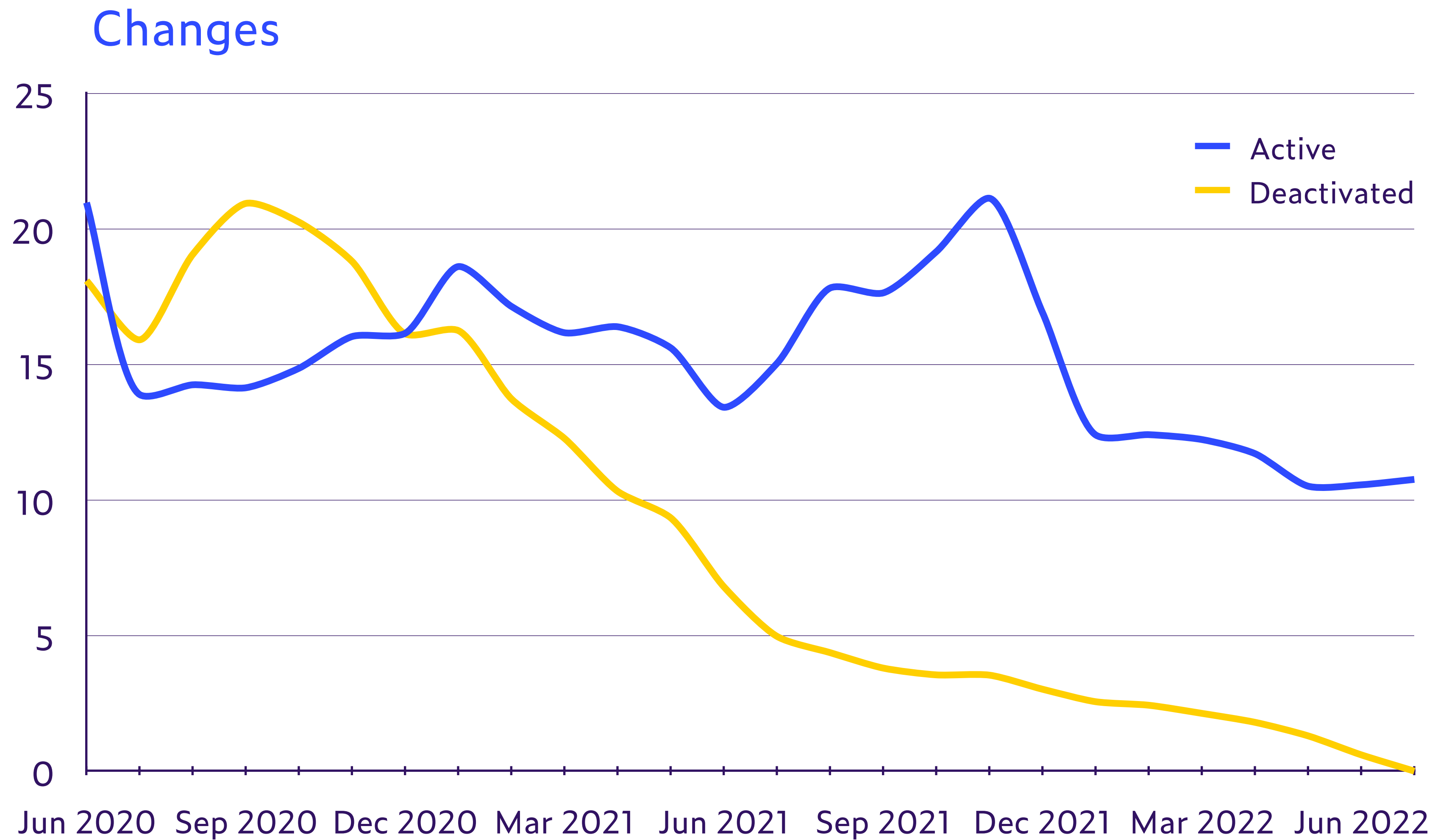


Number of US eCommerce sites by returned response: June 2020 - July 2022





Performance of 'winners' and 'losers' over time



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