PREDICTING INNOVATIVE FIRMS USING WEB MINING AND DEEP LEARNING

Jan Kinne & David Lenz

May 2019

IGL 2019

MOTIVATION:
SHORTCOMINGS OF TRADITIONAL INNOVATION INDICATORS

- Timeliness
- Coverage
- Data collection costs
- Granularity

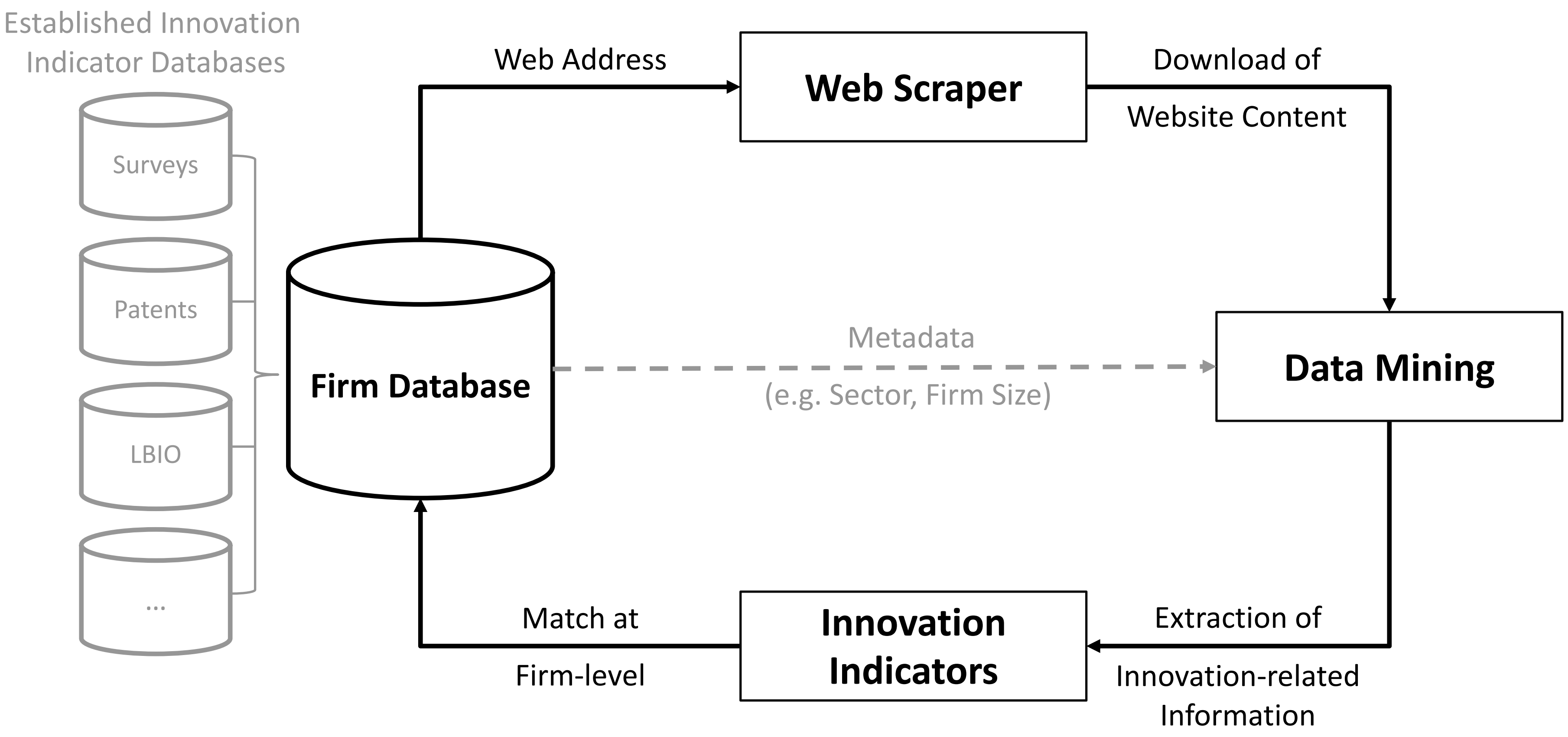
STI policy making requires an accurate and timely picture of the current state of the STI system in order to plan and evaluate policy measures in an evidence-based manner.

MOTIVATION:
WEB MINING OF FIRM WEBSITES

- Firms want us (as customers) to know how innovative they are.
- Almost all (significant) firms have websites nowadays.
- Important medium to tender and promote services and products.
- Firms have an incentive to keep their websites up-to-date.
- Firm website content is often related to product, service, and organisational innovations.

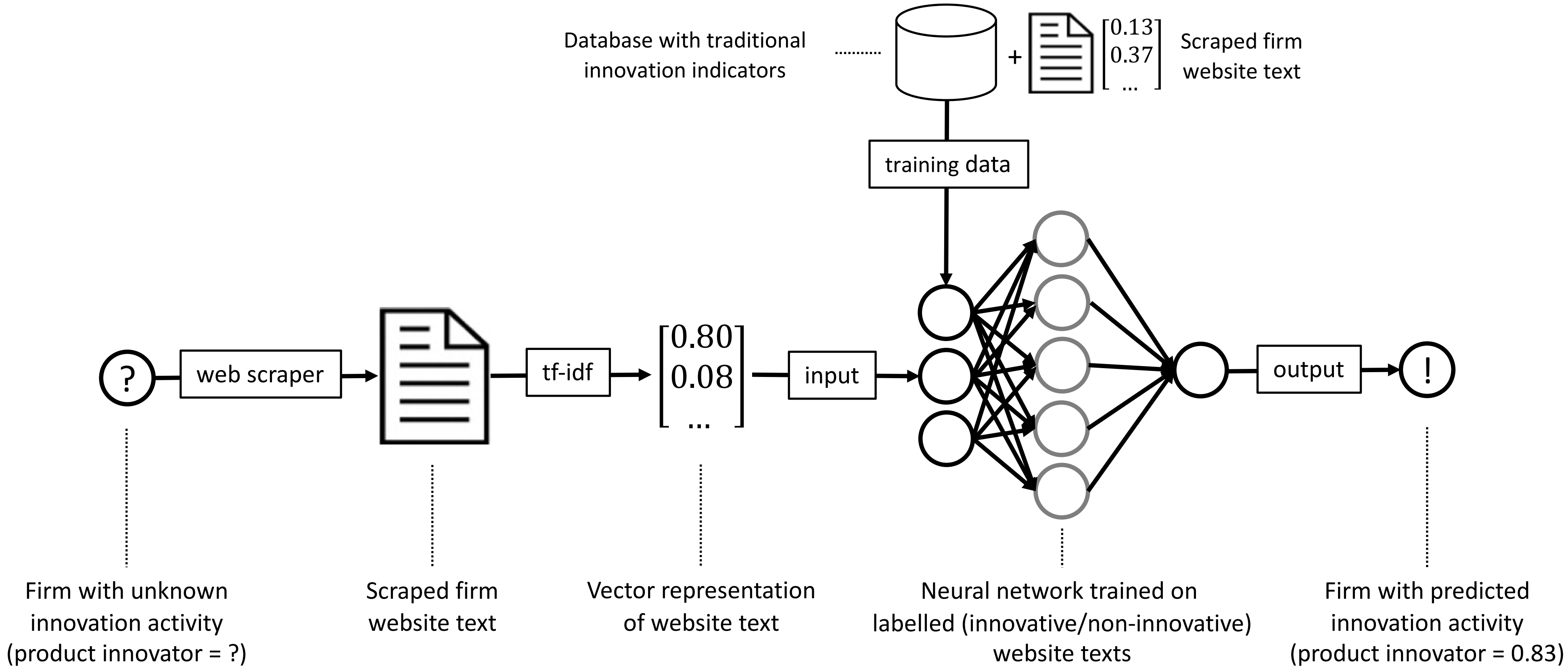
Using this openly accessible data to create timely and granular firm-level innovation indicators.

WEB MINING OF FIRM WEBSITES: ANALYSIS FRAMEWORK

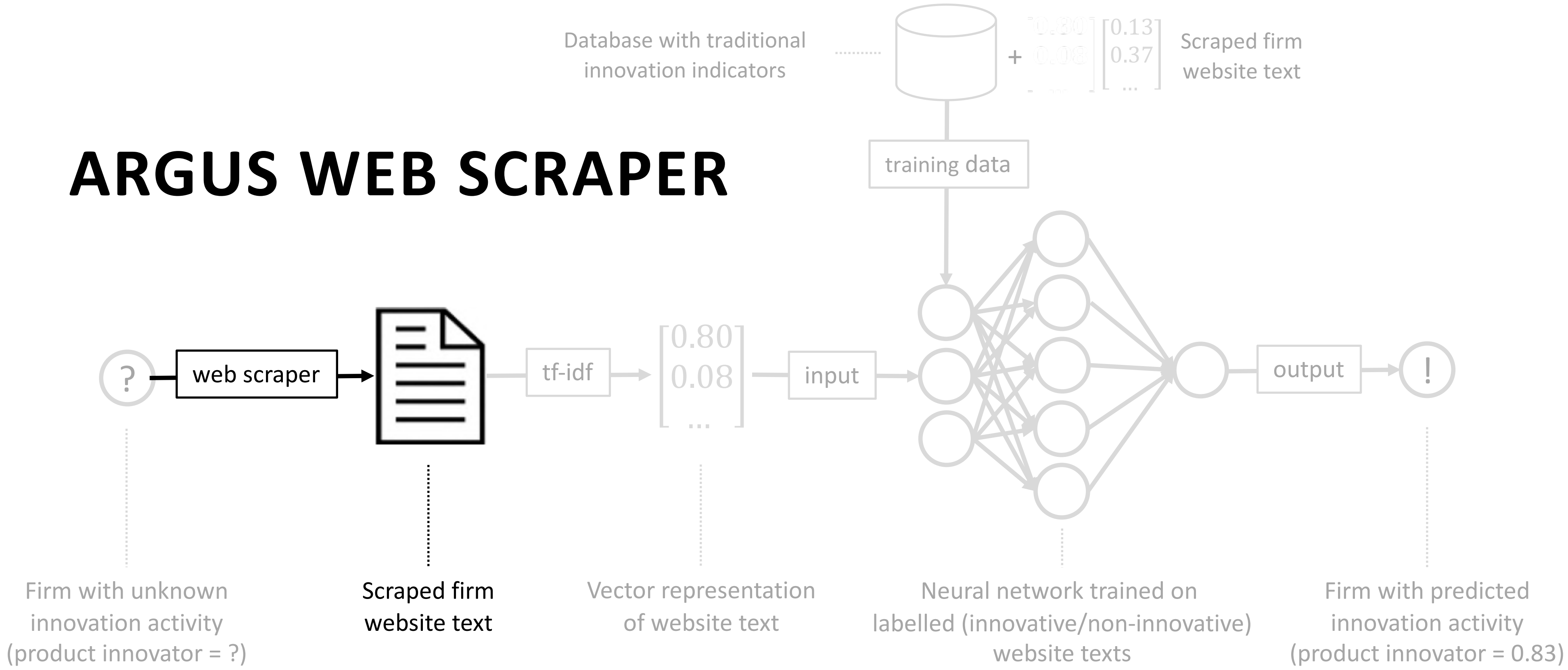


ZEW

DEEP LEARNING OF WEBSITE TEXT: FRAMEWORK



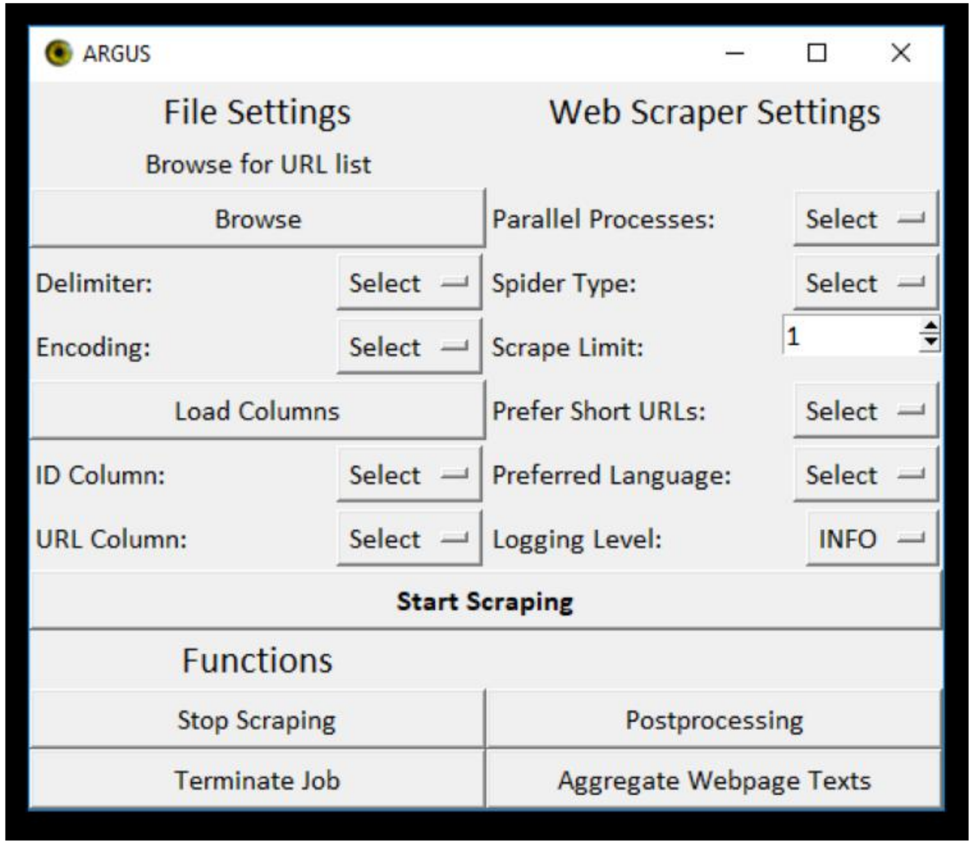
ARGUS WEB SCRAPER



ZEW

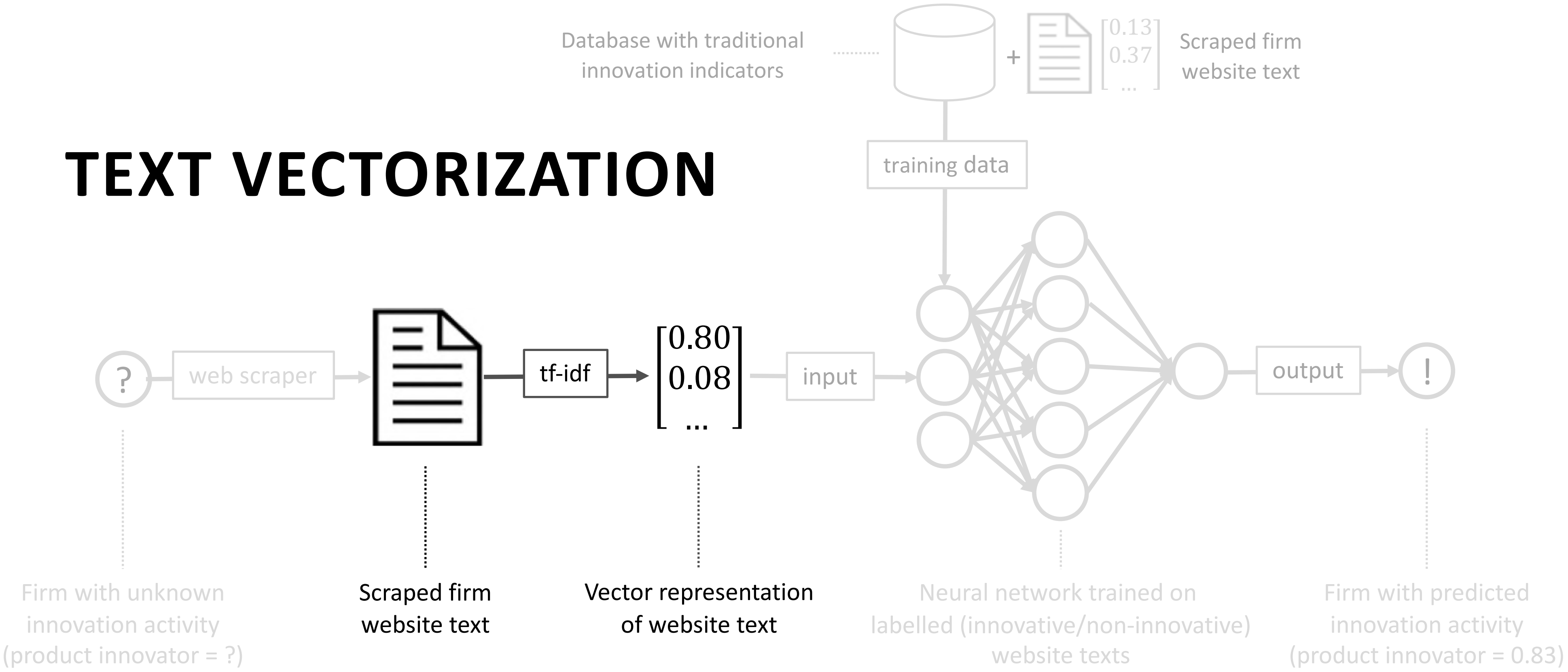
ARGUS WEB SCRAPER: OVERVIEW

- Free and easy-to-use web scraping tool with GUI [github.com/datawizard1337/ARGUS]
- ~1,000 webpages per minute per CPU core
- ~5 days for 50m webpages on an office-grade PC
- Scraping of hyperlinks and texts



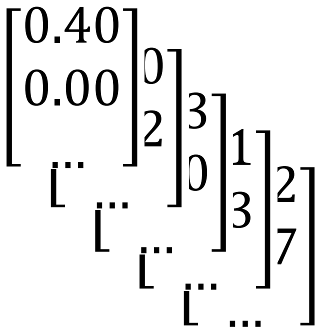
ZEW

TEXT VECTORIZATION



TEXT VECTORIZATION: WORKFLOW

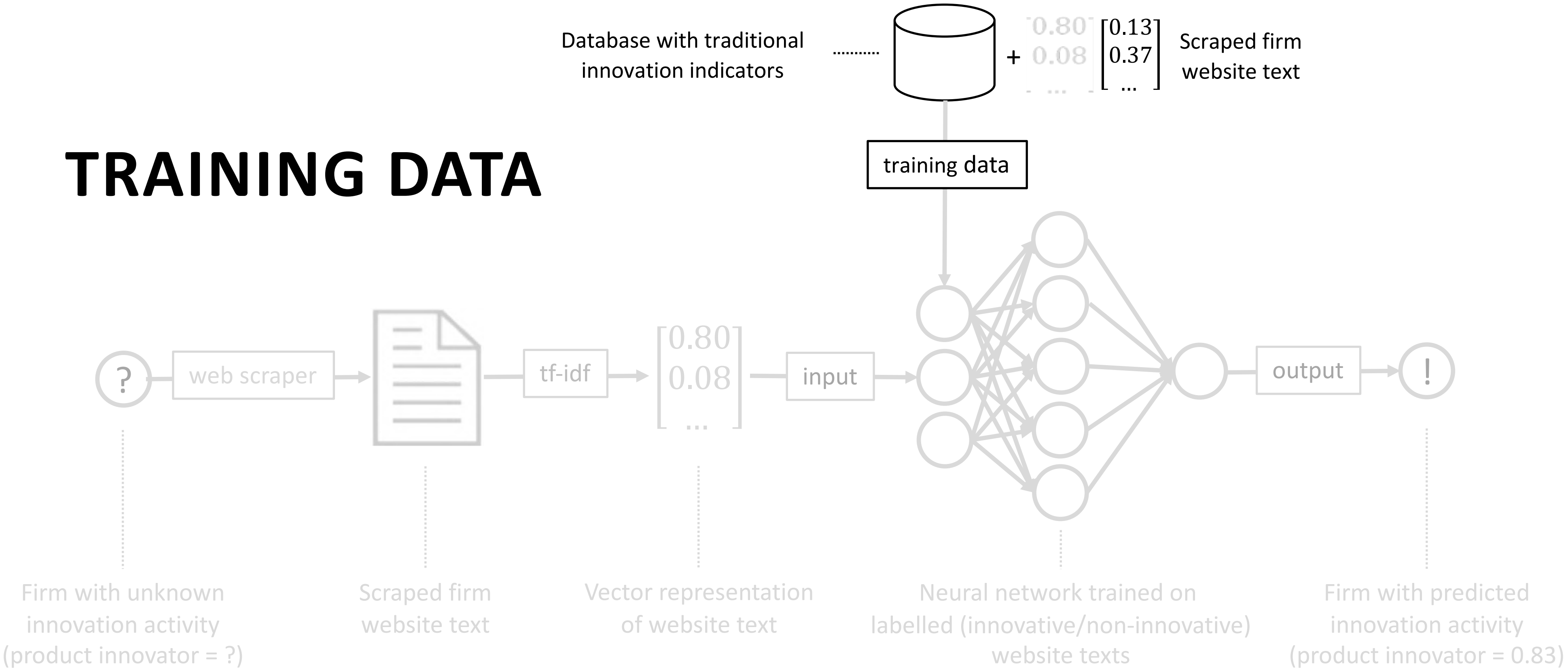
- Web scraping of 650,000 firm websites with max. 25 webpages each
- 50 GB of raw text data
- Deep learning based filtering of “bloat” webpages (e.g. imprints)
- Aggregating webpages to the website level
- Keeping max. 5,000 words per website
- TF-IDF vectorization with popularity-based filtering (dictionary of ~6,150 words)



650,000 fixed size (6,150) vectors

= 650,000 website texts of firms with unknown innovation status

TRAINING DATA



TRAINING DATA: MIP SURVEY DATA

- 2017 Mannheim Innovation Panel (MIP) traditional questionnaire-based survey of ~4,000 firms in Germany
- Binary target variable: Firm was product innovator (YES|NO)

2 Product Innovation

Product innovation describes a product (incl. services) whose components or basic characteristics (technical features, components, integrated software, applications, user friendliness, availability) are either **new or significantly improved**.

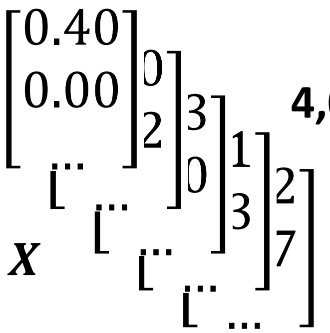
The innovation must be new to your enterprise, but it does not need to be new to your sector or market. **The sole significant factor is your enterprise's evaluation of it**. It does not matter if the innovation was developed by your enterprise alone or in collaboration with other enterprises. Purely aesthetic modifications of products (e.g. colouring, styling) are not regarded as product innovations. Selling alone of innovations that have been developed and produced entirely by other enterprises, also does not count as product innovation in this sense.

→ For examples of product innovations, see the foldout section.

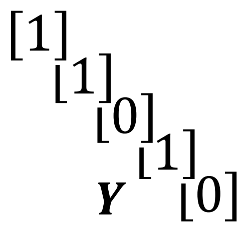
2.1 During the years 2014 to 2016, did your enterprise introduce new or significantly improved products / services?

Yes .... ☐ 1 ↓ No .... ☐ 2 *→ Please continue with **Section 3**.*

- Web scraping of website texts → TF-IDF vectorization



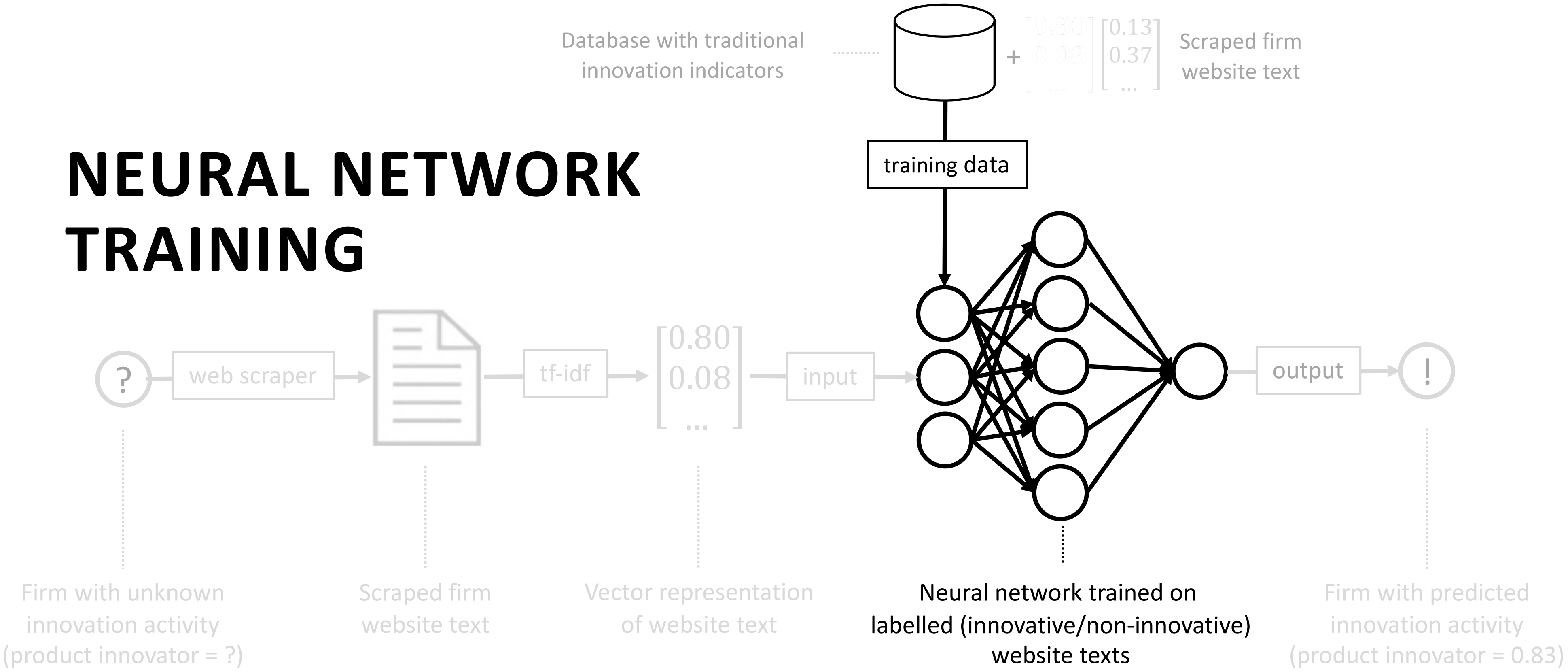
4,000 fixed size (6,150) vectors



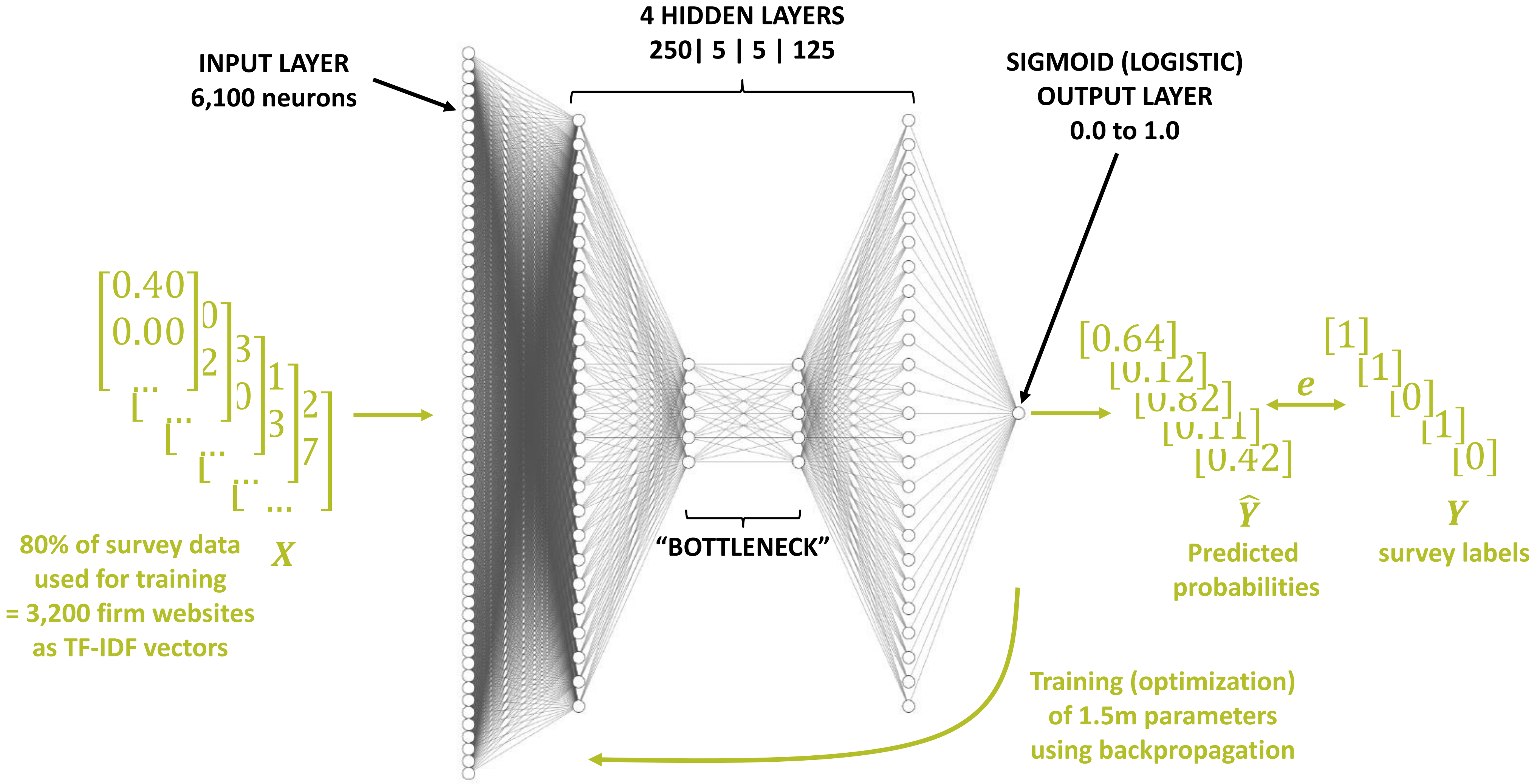
4,000 labels: product innovator yes | no

= 4,000 website texts of firms with known innovation status

NEURAL NETWORK TRAINING



NEURAL NETWORK TRAINING: UNDERCOMPLETE AUTOENCODER-LIKE ARCHITECTURE



NEURAL NETWORK TRAINING: PREDICTION PERFORMANCE

81% of firms predicted “non-innovative” are actually “non-innovative”

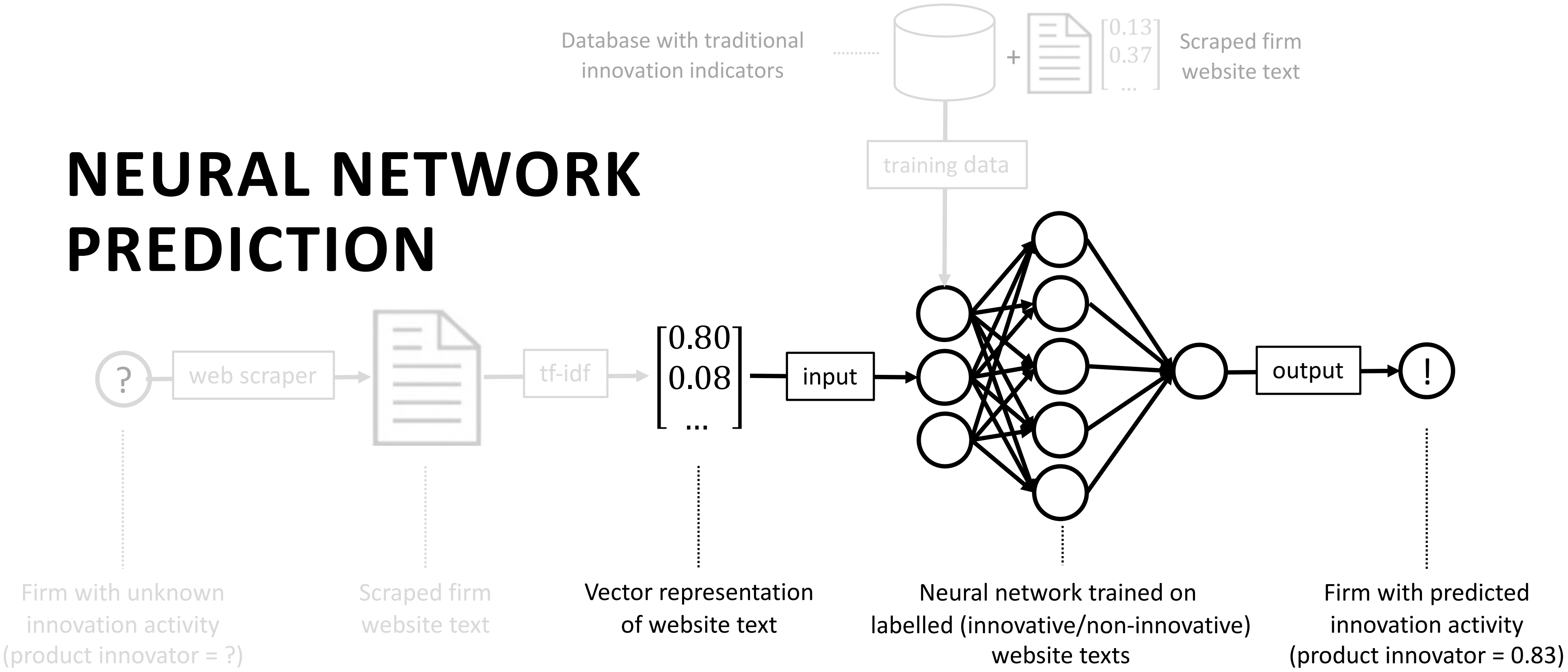
91% of actually “non-innovative” firms are recovered

Composite measure of precision and recall

label	precision	recall	f1-score	support
non-innovative	0.81	0.91	0.86	429
innovative	0.81	0.64	0.71	255
avg / total	0.81	0.81	0.80	684

20% of survey data as test set

NEURAL NETWORK PREDICTION

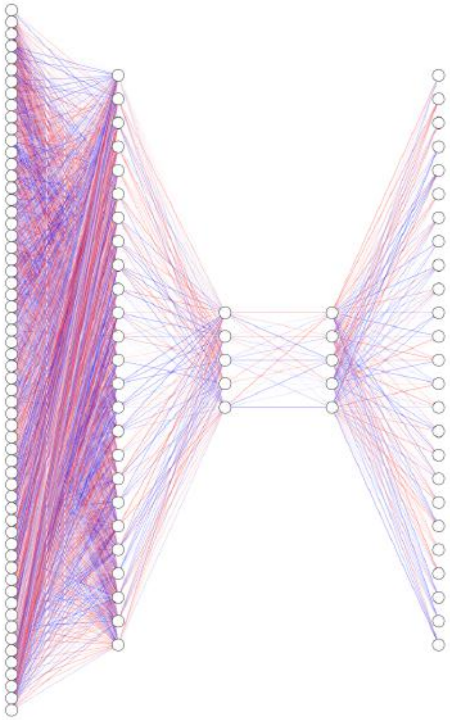


NEURAL NETWORK PREDICTIONS:
OUT-OF-SAMPLE INNOVATION PREDICTION

$$\begin{bmatrix} 0.40 \\ 0.00 \\ \ldots \end{bmatrix}$$

650.000 firm website texts as TF-IDF vectors

X



$$\begin{bmatrix} 0.64 \end{bmatrix} \; [0.12] \; [0.82] \; [0.11] \; [0.42]$$

$\hat{Y}$

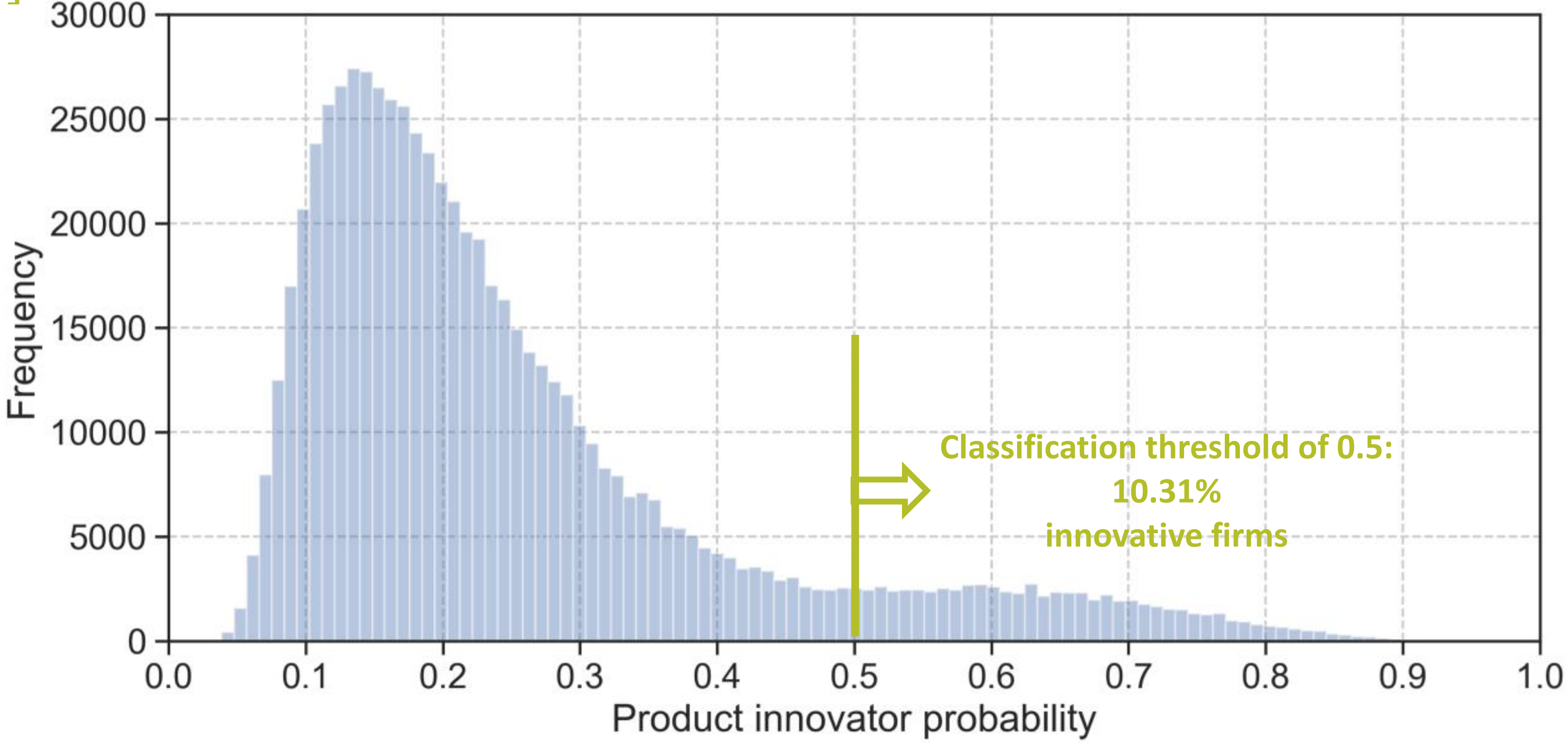
650.000 “product innovator” probabilities

ZEW

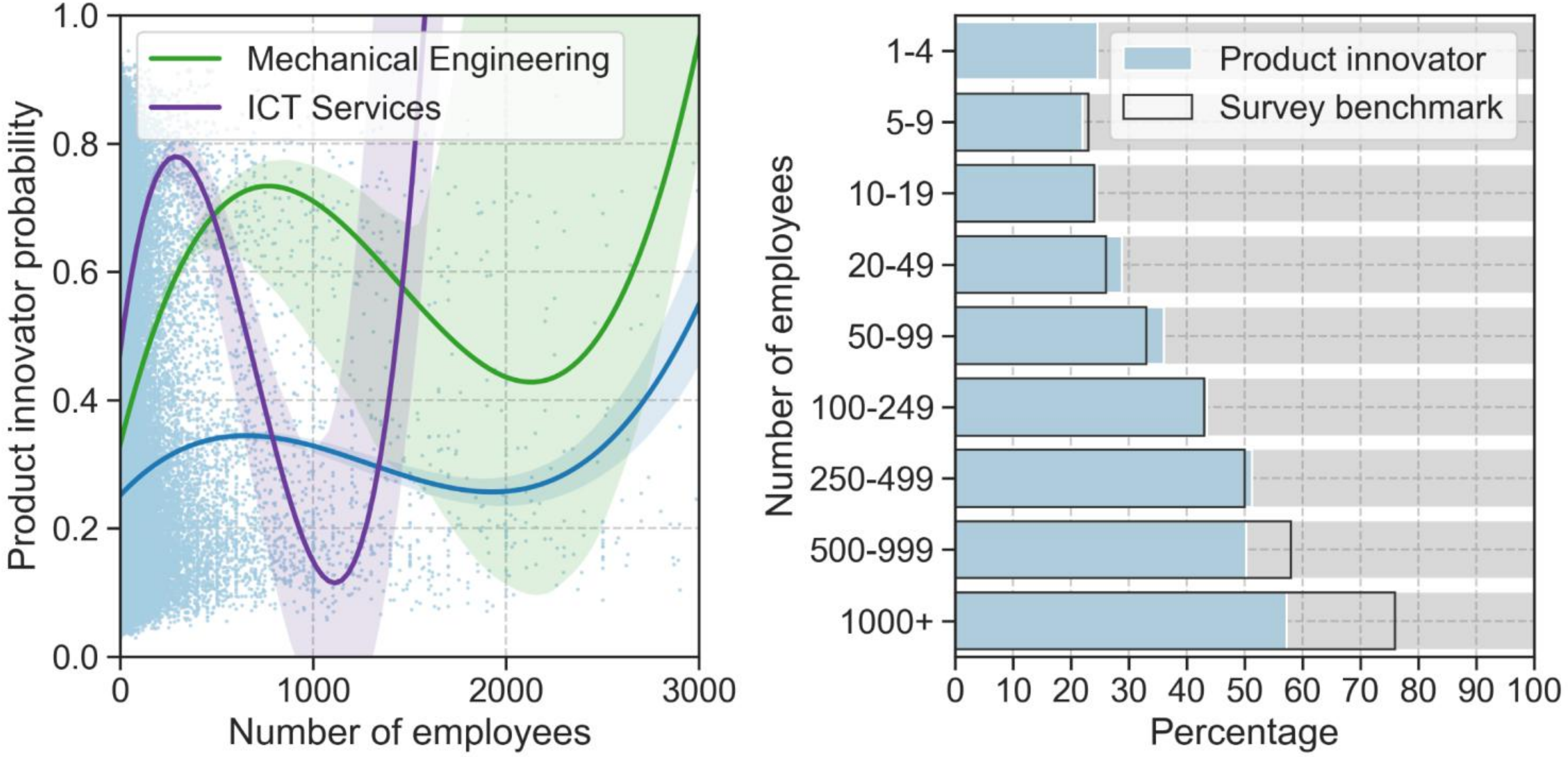
NEURAL NETWORK PREDICTIONS: A CONTINUOUS INNOVATION INDICATOR

[0.64]
[0.12]
[0.82]
[0.11]
[0.42]
$\hat{Y}$

650.000
"product innovator"
probabilities



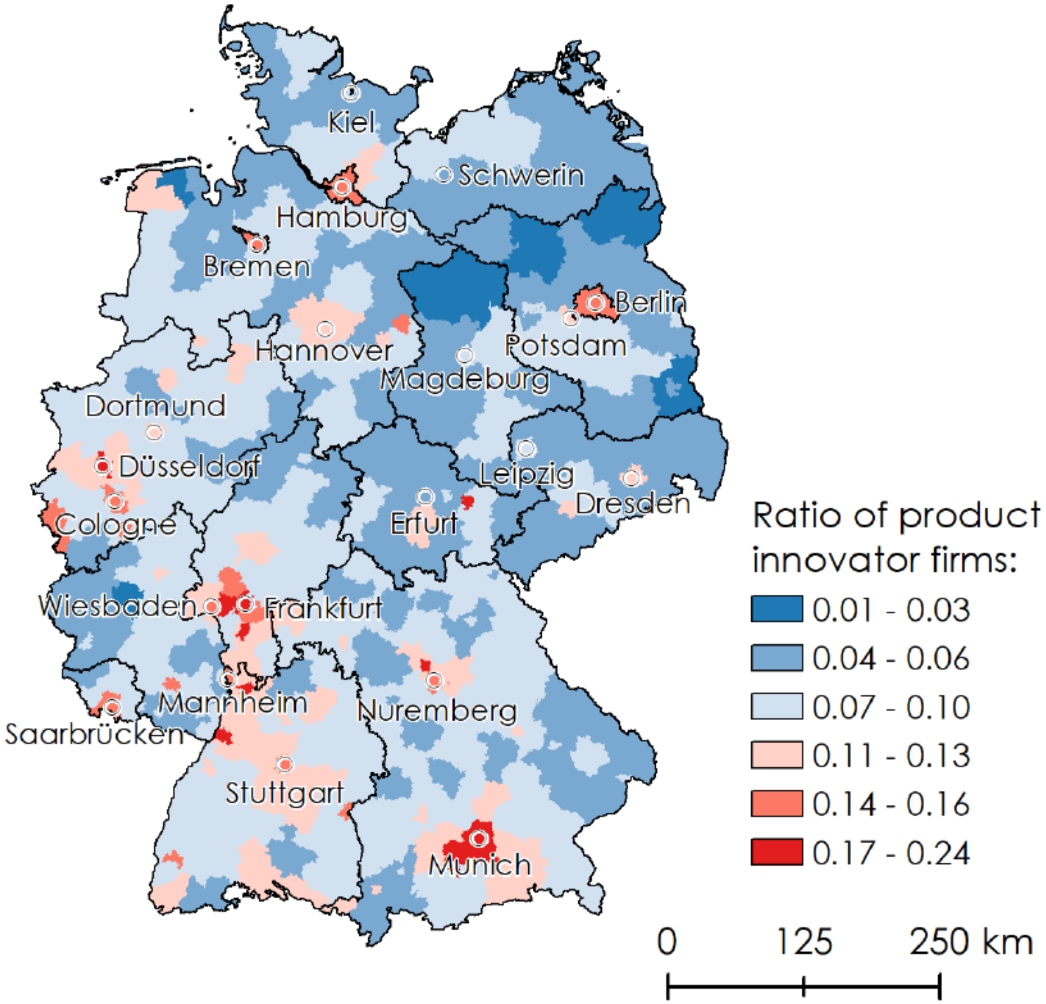
NEURAL NETWORK PREDICTIONS: INNOVATIVENESS BY FIRM SIZE



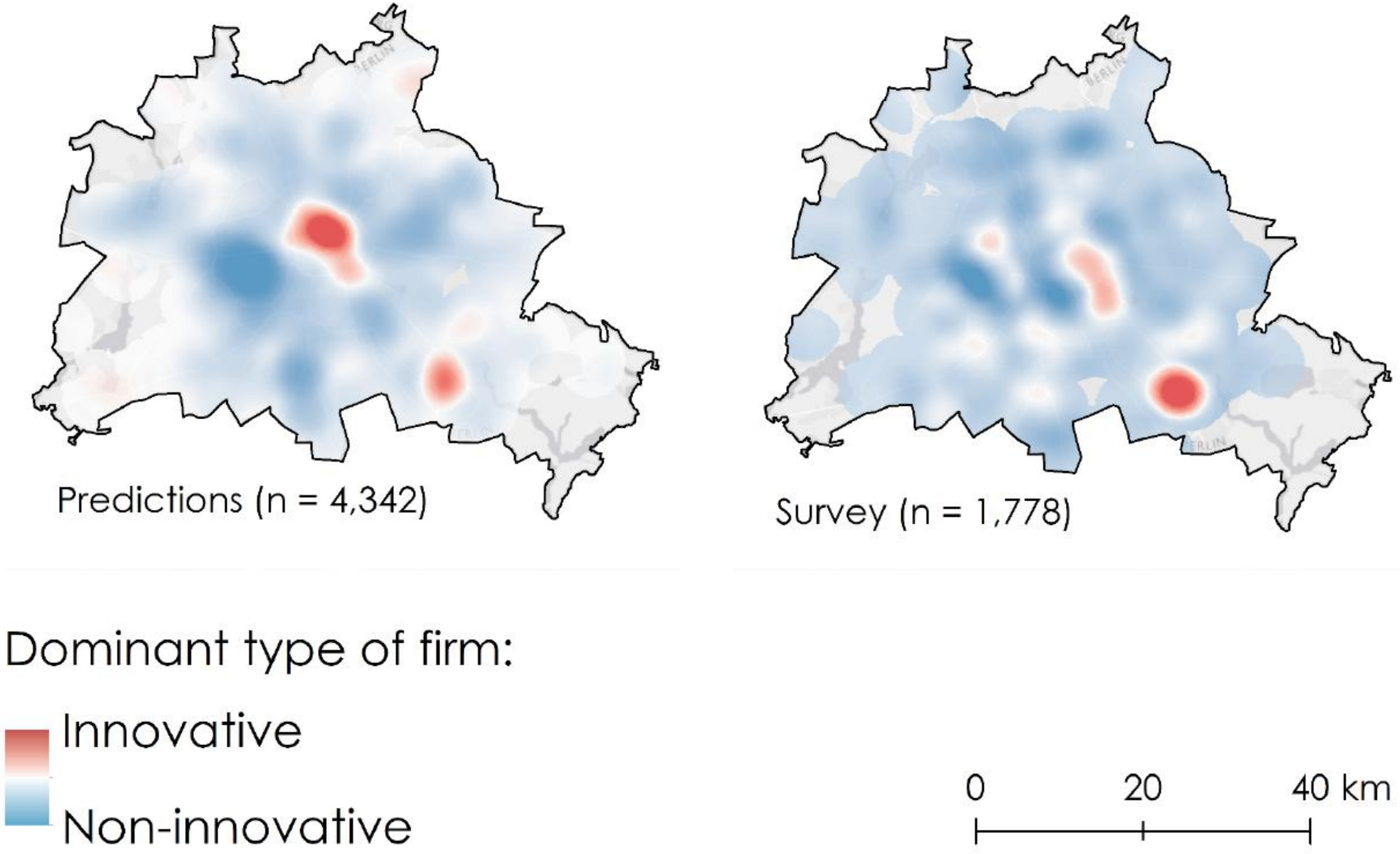
ZEW

NEURAL NETWORK PREDICTIONS: INNOVATIVE DISTRICTS

- **Higher shares of product innovators in urban areas**
  - Correlation with population density: 0.61
- **Lower shares of product innovators in East and North**



NEURAL NETWORK PREDICTIONS: MICROGEOGRAPHIC PATTERN



FUTURE DIRECTIONS: SUPPORTING BIG DATA BASED POLICY MAKING

- Industry-specific prediction models
- Developing further web-based innovation indicators
- Formalization and dissemination of a coherent methodology
- Building up a panel database of web data
- Application in policy evaluation projects
- Investigating microgeographic diffusion of innovation and technology

THANKS!

zew.de/en/team/jki

github.com/datawizard1337

twitter.com/jan_kinne