Big Data and Semantic Analysis for Innovation Indicators and Policy

Georg Licht
Centre for European Economic Research (ZEW)

CSTP-TIP Workshop: Semantic Analysis for Innovation Policy
Paris, 12-13 March 2018

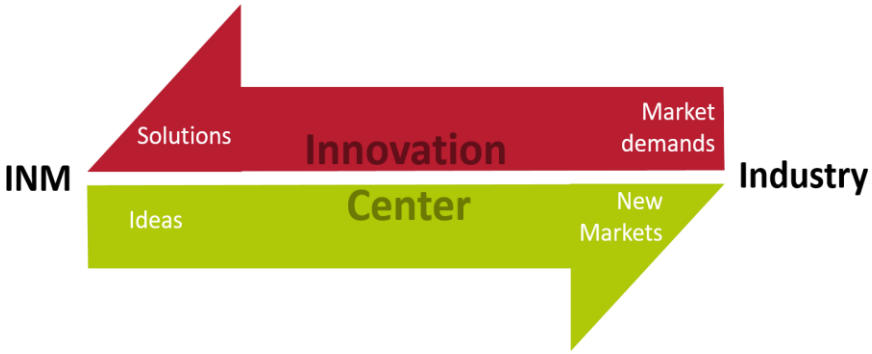
Leibniz Gemeinschaft ZEW

Agenda

- Short introduction of background
- Project 1: Partner for joint research project
- Project 2: Semantic analysis of company webpages to increase the timeliness and granularity of innovation indicators
- Some tentative learnings & next steps

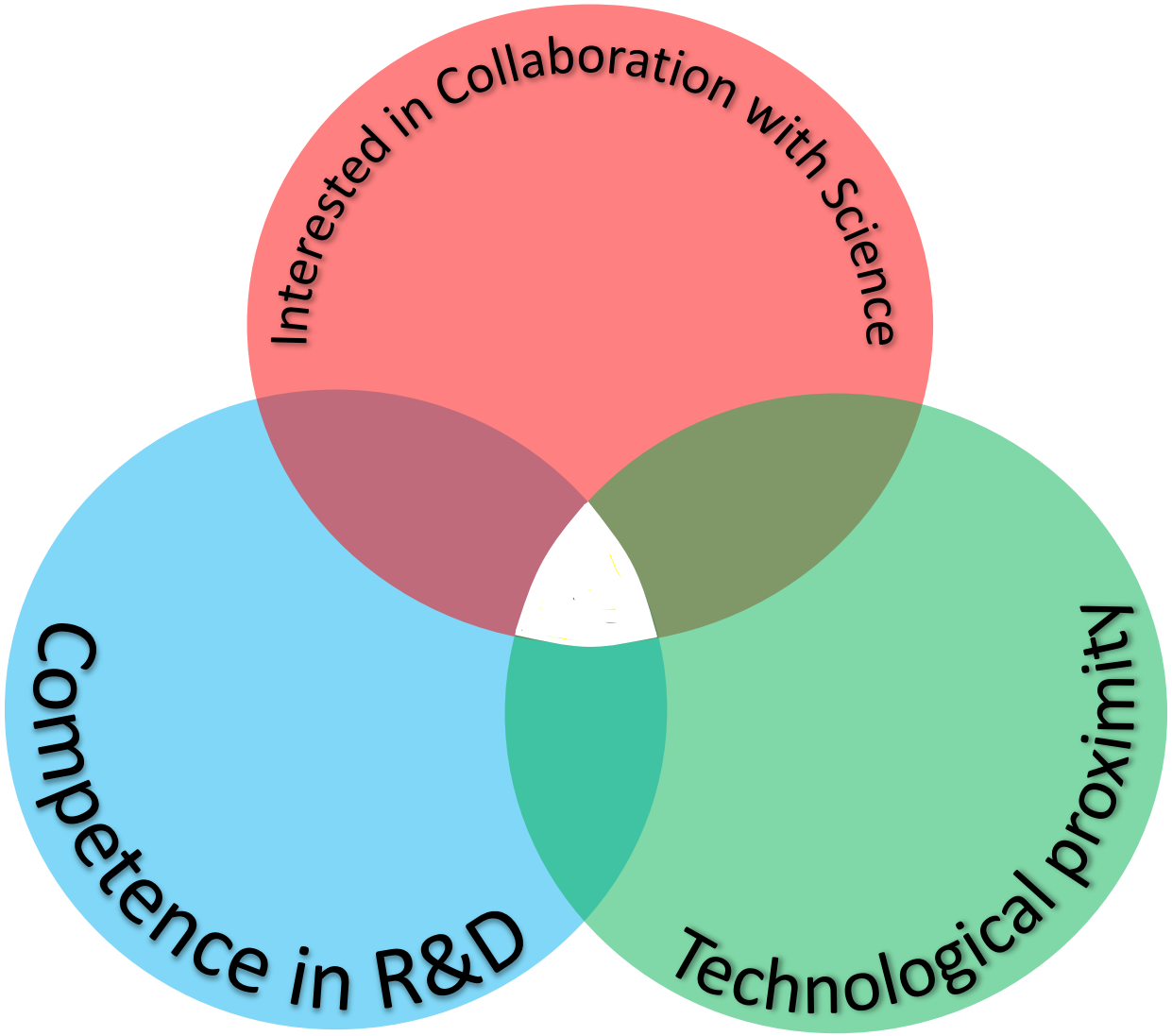
Background and objective

- **Institute for New Materials (INM)** looks for industry partners (SMEs) to further develop and turn INM inventions into innovations
- INM: Basic research (typical up to Technology Readiness Level 5)
- Wide portfolio of technologies:
  Surface technology,
  physical/chemical technologies,
  biotechnology



- INM tech transfer unit
  to perform applied research and innovation „Innovation Center“
- Task of the project: Identification of potential partners based on ZEW‘s data for INMs Innovation Center

Theory: The Optimal Partner for INM



Approach

(1) **Neuronal network** to link competencies of firms, collaborative abilities and behaviour and technological proximity

(2) **Daten:** Mannheim Enterprise Panel (MUP), Mannheim Innovation Survey (German CIS), Linked Patent application database link to MUP, Data on publicly funded, collaborative R&D projects

(3) **Result:** List of firms with firm specific scores for being a partner for INM in applied research and development projects

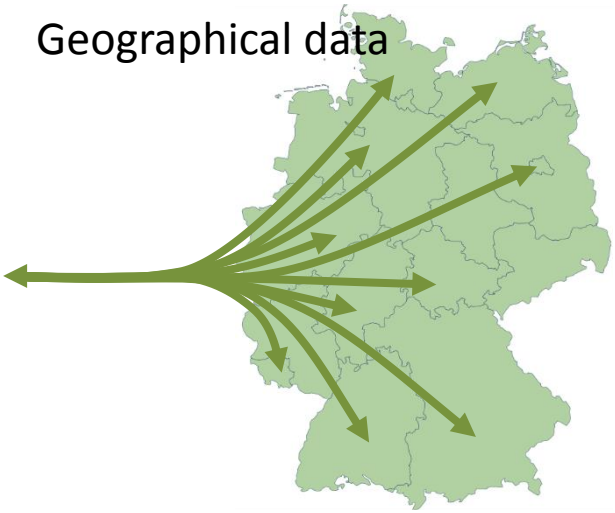
(4) **Verification of list** (done by INM)

Firm level data & information

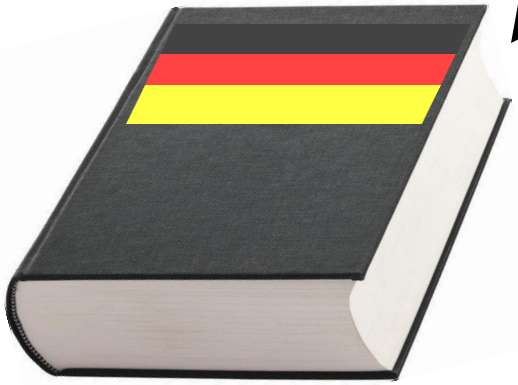
Patents

Firm level data

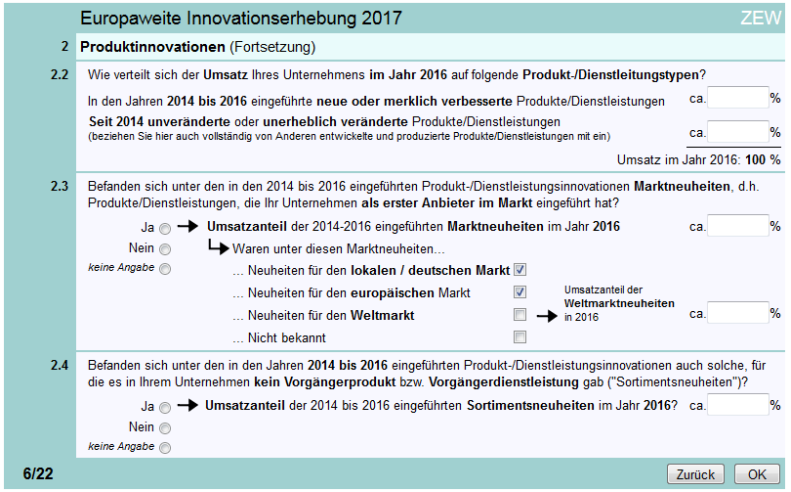
Geographical data



Publicly Funded R&D projects



Innovation Survey Data



Europaweite Innovationserhebung 2017 ZEW

2 **Produktinnovationen** (Fortsetzung)

2.2 Wie verteilt sich der **Umsatz** Ihres Unternehmens **im Jahr 2016** auf folgende **Produkt-/Dienstleitungstypen**?

In den Jahren **2014 bis 2016** eingeführte **neue oder merklich verbesserte** Produkte/Dienstleistungen ca. %

Seit 2014 unveränderte oder **unerheblich veränderte** Produkte/Dienstleistungen
(beziehen Sie hier auch vollständig von Anderen entwickelte und produzierte Produkte/Dienstleistungen mit ein) ca. %

Umsatz im Jahr 2016: **100 %**

2.3 Befanden sich unter den in den 2014 bis 2016 eingeführten Produkt-/Dienstleistungsinnovationen **Marktneuheiten**, d.h. Produkte/Dienstleistungen, die Ihr Unternehmen **als erster Anbieter im Markt** eingeführt hat?

Ja → **Umsatzanteil** der 2014-2016 eingeführten **Marktneuheiten** im Jahr **2016** ca. %

Nein → Waren unter diesen Marktneuheiten...

keine Angabe

... Neuheiten für den **lokalen / deutschen Markt**

... Neuheiten für den **europäischen** Markt

... Neuheiten für den **Weltmarkt** → **Umsatzanteil der Weltmarktneuheiten in 2016** ca. %

... Nicht bekannt

2.4 Befanden sich unter den in den Jahren **2014 bis 2016** eingeführten Produkt-/Dienstleistungsinnovationen auch solche, für die es in Ihrem Unternehmen **kein Vorgängerprodukt** bzw. **Vorgängerdienstleistung** gab ("Sortimentsneuheiten")?

Ja → **Umsatzanteil** der 2014 bis 2016 eingeführten **Sortimentsneuheiten** im Jahr **2016**? ca. %

Nein

keine Angabe

6/22 Zurück OK

Industry-Science collaboration

Neuronal Network for probability of collaboration

- Based on German Innovation Survey Data (Mannheim Innovation Panel)
- Mannheim Enterprise Panel: E.g. sales, employment, industry, creditrating, ownership structure
- Local environment of firms (other firms within 5/10 kilometers distance
- Patent application data: Stock of patent application & patents applied for in last three years

Technological competencies of firms

Neuronal Network for probability of hold publicly funded R&D project

- Link Mannheim Enterprise Panel with patent application data
- Link Mannheim Enterprise Panel with data of public R&D project funding data base

Technological proximity

Distance between patent portfolio of INM and firms

- Jaccard-Index based on IPC (International Patent Classification)

$$Jaccard(IPC_U, IPC_{INM}) = \frac{|IPC_U \cap IPC_{INM}|}{|IPC_U \cup IPC_{INM}|}$$

  - MAXIPC: Patent with maximum similarity
    AVGIPC: Average similarity using INM and firms patent stock

- Similarity between patent title (Inverse Document Frequency)
  - MAXTITLE: Patent with maximum similarity
    AVGTITLE: Average similarity using INM and firms patent stock

ZEW

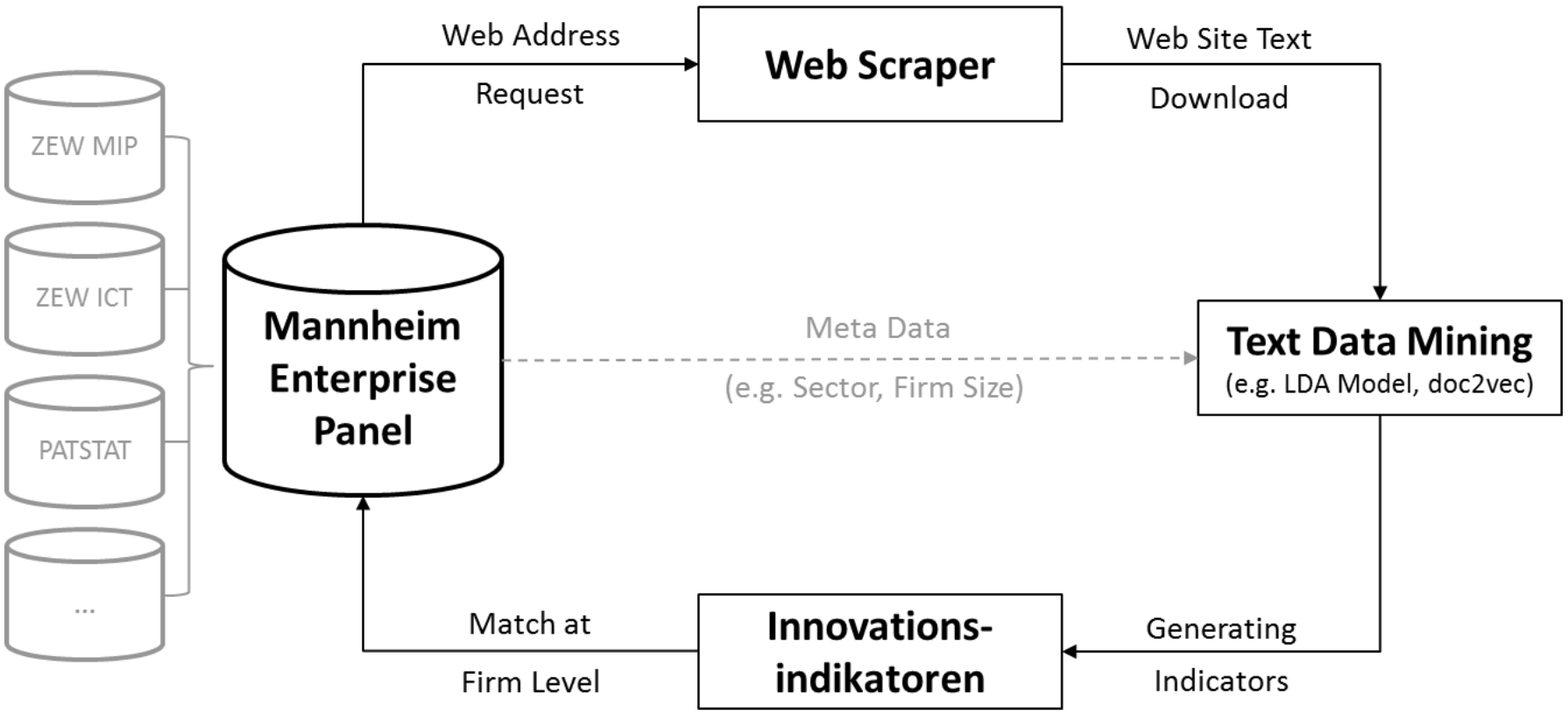
List of firms

firma	branche	patstock	koop	oefp	maxipc	avgipc	maxtitle	avgtitle	inmdist
proRheo GmbH	electro	0.14	0.0000	0.0465	1.00	1.00			142.10
INMATEC Technologies GmbH	rubber	0.85	0.0005	0.9986	1.00	1.00			155.24
GALAB Technologies GmbH	glas	0.89	0.0000	0.0002	1.00	1.00			520.46
Kongsberg Maritime Contros GmbH	electro	1.38	0.0163	1.0000	1.00	0.75			605.77
SECURETEC Detektions-Systeme AG	electro	3.62	0.0468	0.9214	1.00	0.75			363.22
A.M. Ramp & Co. GmbH	chemistry	0.32	0.3724	1.0000	0.75	0.75			140.49
ICT Integrated Circuit Testing Gesellschaft fü	electro	47.10	0.1429	0.1836	1.00	0.72	88.68	1.61	368.39
micro resist technology Ges. für chem. Mate	chemistry	0.76	0.8650	1.0000	1.00	0.70	99.96	49.98	583.26
Viramed Biotech AG	chemistry	0.52	0.0000	0.0000	1.00	0.67			348.06
Attomol GmbH Molekulare Diagnostika	chemistry	0.41	0.8843	1.0000	1.00	0.67			558.35
CeramTec-Etec GmbH	glas	1.41	0.0001	0.9973	1.00	0.67			182.06
Bruker Nano GmbH	electro	2.89	0.2598	1.0000	1.00	0.65			580.13
ERWEKA GmbH	machine	1.66	0.8732	0.9947	1.00	0.61			156.54
Esko-Graphics Imaging GmbH	machine	9.06	0.5367	1.0000	1.00	0.59			549.11
Transfertex GmbH & Co. Thermodruck KG	print	0.73	1.0000	0.9995	0.50	0.50			169.91
BYK-Gardner GmbH	electro	0.90	0.9001	0.0000	0.50	0.50			364.11
Hemmelrath, Lackfabrik GmbH	chemistry	0.14	0.9602	1.0000	0.50	0.50			167.28
Krögel Maschinenbau GmbH & Co. KG	machine	0.32	0.9903	1.0000	0.50	0.50			295.29
Bte Bedampfungstechnik GmbH	glas	0.32	1.0000	1.0000	0.50	0.50			166.75
Human Gesellschaft für Biochemica und Dia	electro	0.29	0.5076	0.6560	0.50	0.50			131.08
LAR Process Analysers AG	electro	1.32	0.9622	1.0000	0.50	0.50			577.90
Nanion Technologies GmbH	furniture	0.44	0.0000	0.0000	0.50	0.50			353.89
Tritron GmbH	chemistry	0.32	0.9025	0.9993	0.50	0.50			227.78
emtec Electronic GmbH	electro	0.32	0.8968	0.9996	0.50	0.50			450.50
Union Instruments GmbH	electro	0.38	0.0344	0.1612	0.50	0.50			570.44

Project 2: Motivation and objectives

- Improving the current methodology for gaining firm level innovation indicators
- Improved **timeliness**, higher geographic and sectoral **granularity** with comparatively **low data collection costs**
- Mapping the **diffusion** of technology (as web based information collection allow for high frequency)
- Identification of current technological **trends** using endogenic categorisation

Firm level Innovation Indicators from text mined firm websites



WEBSCRAPER

- Scrapes webpage text based on input list of firms' web addresses and stores them to a database

is_start_url	mup_url	redirect	start_url	text	timestamp	url
NaN	sill-lighting.com	False	http://www.sill-lighting.com/	Zurück zur Übersicht 1424 Laurel Drive Sewickl...	Wed Dec 27 13:59:25 2017	http://www.sill-lighting.com/contact-amerika
1.0	sill-lighting.com	False	http://www.sill-lighting.com/	das licht von morgen Die Firma SILL Leuchten G...	Wed Dec 27 13:59:25 2017	http://www.sill-lighting.com/

- Number of sub-pages per website (firm) is restricted by *limit* parameter

- Fast (~10.000 URLs/minute) and based on open source software (Python Scrapy package)

Some tentative learnings & next steps

- **NEXT STEPS**
  - Project 1:
    - Including patent abstracts to improve similarity analysis
    - Testing the model by contacting firms
  - Project 2:
    - Topic modelling on corpus based on websites
    - At which stage and how to include background information on firms

- **LEARNINGS**
  - Neuronal network delivers better results than individual probit equations
  - Downloading webpages of firms
    - Number of subpages is highly skewed; Limit the number of sub-page to 30 provide sufficient information
    - High frequency of downloading is feasible