INTEGRATING DIFFERENT DATA SOURCES – NEW ANALYTICAL POTENTIALS

Rainer Frietsch
Fraunhofer ISI, Karlsruhe, Germany

Image Source: istockphoto.com

Fraunhofer ISI

Structure of the talk

1. Starting Points
2. Challenges and Potentials
3. Example 1: R&D data and patents
4. Example 2: University-invented patents
5. Example 3: Patent-paper twins

Starting Points

- My **hypothesis** (better: my conviction): the available structured data is still under-explored
- **Data enrichment:** Classifications, gender information, experience (cumulated information), regionalisation/geo information/distance
- **Text mining**: finding new structures in structured data, e.g. emerging fields, classifications, "hidden information" like strategies
- **Matching data**
  - Macro data: similar classifications (mainly for academic exercises), e.g. exports and patents, R&D expenditure and patents, publications and project funding
  - Micro data: firms' and persons' names matching, e.g. CVs and patents, CVs and publications, R&D expenditure and patents, firm data and patents/publications, patents and publications
- Recent **examples**: Marie Curie fellows and publications; Hoppenstedt/Orbis and patents; EU Scoreboard and patents; DTI Scoreboard, patents, and COMPUSTAT
  German R&D survey and patents; university invented patents; patent-paper twins

Matching Patent and Firm Data – Challenges and Potentials

Fraunhofer
ISI

Major challenges

- Mergers and Acquisitions / Renaming
- International branches (not only headquarter)
- Subsidiaries might be the filing authority
- Ownership of companies

Major challenges – applicants versus companies

- Patent data are at the level of patent applicants but **patent applicants are not necessarily companies**, which leads to **several challenges**.

  - **Within** the patent database (PATSTAT) the names of applicants are in **raw data format**
    - Different **spelling variations** of the same company name.
    - might include abbreviations, special characters, typing errors, legal form etc.

  - **Which** firm level is to be covered?
    - Possible Biases:
      a) The patent applicant might be the parent company, a business unit or a subsidiary.
      b) Firm policy might state to file all patents via one single applicant (e.g. Siemens in Munich).

  - **Firms** are „changing“ over time. Mergers and Acquisitions, buy-outs and sales of subsidiaries make time-series analyses difficult.

Fraunhofer ISI

Name harmonization

- **EEE-PPAT Table by the K.U. Leuven**
  - Automated harmonization of all patent applicant names in PATSTAT
  - Based exclusively on the names available in PATSTAT (including addresses) and does not use any additional information from outside the database
  - **Stepwise validation:**
    - Character cleaning (HTML format codes, accented characters), punctuation cleaning, legal form indication cleaning (Inc., LTD, GmbH etc. = Company), common company word removal („COMPANY“, „CORP“, „CORPORATION“)
    - Spelling variation harmonization („SYSTEM“, „SYSTEMS“, „SYSTEMES“), condensing of irrelevant characters („3 COM“, „3COM“), Umlaut harmonization

  - **The OECD HAN Database**
    - Dictionary of applicant names is used
    - Identification of firms, non-business organizations and individuals
    - Name cleaning of applicant names (steps 1 and 2 of the K.U. Leuven algorithm)

Fraunhofer ISI

An exemplary overview – Bayer AG

PERSON NAME	DOC STANDARD NAME	EEE-PPAT NAME	HAN NAME
Bayer A.G.	BAYER AG	BAYER	BAYER AG
Bayer AC	BAYER AC	BAYER AC	BAYER AC
Bayer Adtiengesellschaft	BAYER AG	BAYER	BAYER ADTIENGESELLSCHAFT
Bayer AG	BAYER AG	BAYER	BAYER AG
Bayer Akgiengesellschaft	BAYER AKGIENGESELLSCHAFT	BAYER	BAYER AKGIENGESELLSCHAFT
Bayer Akiengesellschaft	BAYER AG	BAYER	BAYER AKIENGESELLSCHAFT
Bayer Aktlengesellschaft	BAYER AKTLENGESELLSCHAFT	BAYER	BAYER AKTLENGESELLSCHAFT
Bayer Animal Health GmbH	BAYER HEALTHCARE AG	BAYER ANIMAL HEALTH	BAYER ANIMAL HEALTH GMBH
Bayer BioScience GmbH	BAYER BIOSCIENCE GMBH	BAYER BIOSCIENCE	BAYER BIOSCIENCE GMBH
Bayer Business Services GMBH	BAYER BUSINESS SERVICES GMBH	BAYER BUSINESS SERVICES	BAYER BUSINESS SERVICES GMBH
Bayer Chemical Aktiengesellschaft	BAYER CHEMICAL AG	BAYER CHEMICALS	BAYER AG
Bayer Chemicals AG	BAYER CHEMICALS AG	BAYER CHEMICALS	BAYER CHEMICALS AG
Bayer Chemicals Aktiengesellschaft	BAYER CHEMICALS AG	BAYER CHEMICALS	BAYER CHEMICALS AG
Bayer CropScience AG	BAYER CROPSCIENCE AG	BAYER CROPSCIENCE	BAYER CROPSCIENCE AG
Bayer CropScience	BAYER CROPSCIENCE AG	BAYER CROPSCIENCE	BAYER CROPSCIENCE AG
Bayer CropScience GmbH	BAYER CROPSCIENCE GMBH	BAYER CROPSCIENCE	BAYER CROPSCIENCE GMBH
Bayer HealthCare AG	BAYER HEALTHCARE AG	BAYER HEALTHCARE	BAYER HEALTHCARE AG
Bayer Schering Pharma AG	BAYER SCHERING PHARMA AG	BAYER SCHERING PHARMA	BAYER SCHERING PHARMA AG
Bayer Schering Pharma Aktien	BAYER SCHERING PHARMA AG	BAYER SCHERING PHARMA	BAYER SCHERING PHARMA AG
Bayer Technology Services GmbH	BAYER TECHNOLOGY SERVICES GMBH	BAYER TECHNOLOGY SERVICES	BAYER TECH SERVICES GMBH

- **Two basic problems:**
  - Spelling variations
  - Parent company (ultimate owner), company, business unit, M&As

Fraunhofer ISI

Companies vs. business units

- Companies or enterprises are subject to **major changes over time.**
  - Companies are not always the patent applicant (and then also not named on the patent application)
  - Business units usually do not show up within patents
  - **Possible solutions:**
    - Identification of applicants and assignment to business units according to the address of the inventor
      - Problems: Inventors of several business units might be involved, inventors use their private addresses, external collaborations
    - Identification of applicants and assignment of technologies to business units

Matching of R&D survey data and patents

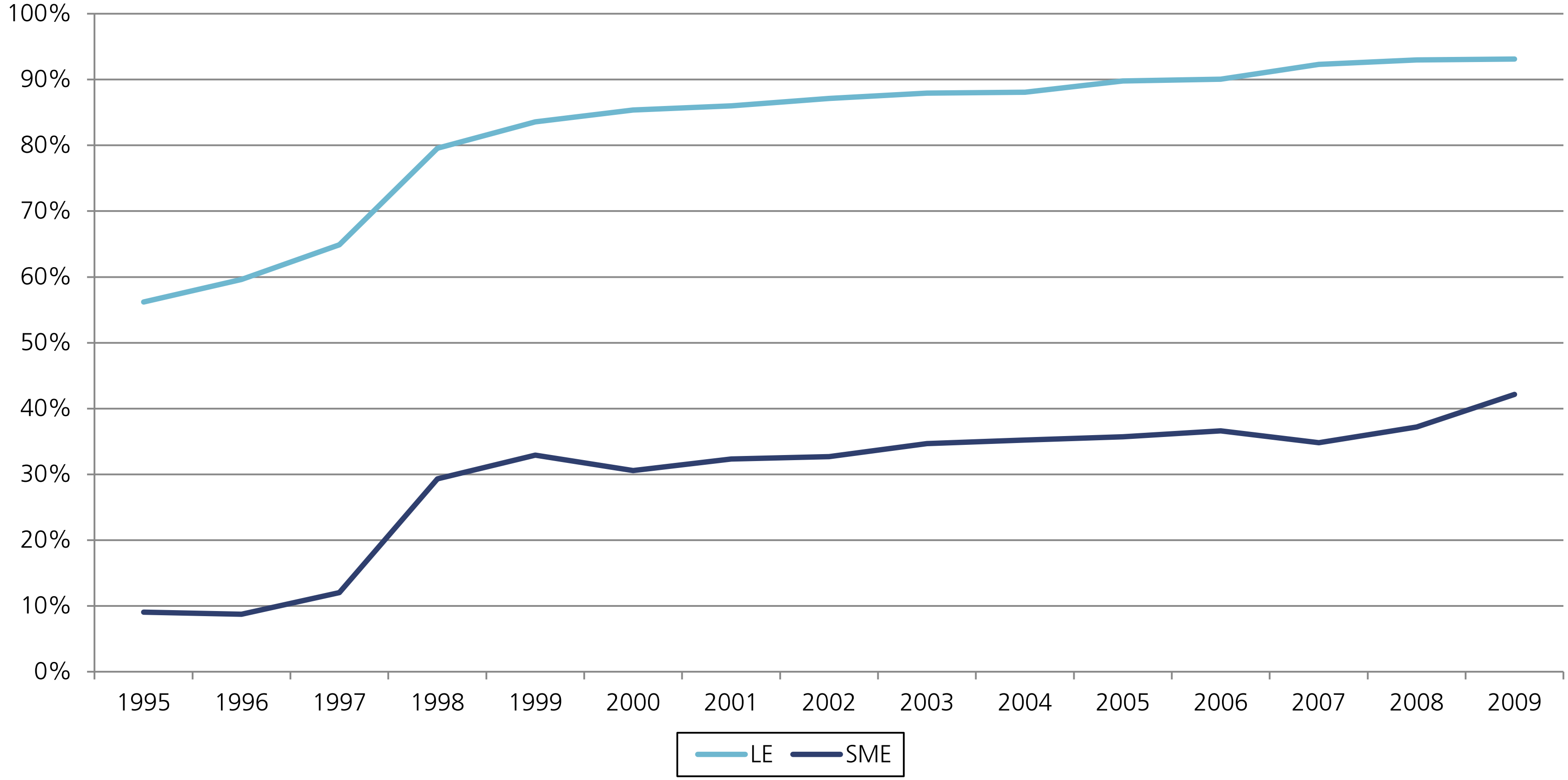
The matching procedure

- **Aim**
  - Finding information of patent applicants in PATSTAT, which fit (or are similar) to a firm/branch in the German R&D survey by Stifterverband

- Name cleaning
  - Cleaning of different spellings: use of small letters, "umlaute" and special characters, blanks, deletion of legal forms
- Similarity between names
  - **Levenshtein-Distance** of names: minimal number of editing steps to make the two texts identical
  - If the first three digits of the zip code do not match (given they are available), then similarity = 0
- Selection of matches
  - Is the similarity higher than the defined threshold, then we define this as a match. The threshold is empirically defined by recall and precision

Fraunhofer ISI

Coverage by type of applicants (share of matched applications in total applications)



Quelle: EPO – PATSTAT, calculations by Fraunhofer ISI,

Fraunhofer ISI

Reasons for incomplete coverage

- Not all patenting companies are covered by the company database
  - For example: BSH BOSCH UND SIEMENS HAUSGERAETE, HARMAN BECKER AUTOMOTIVE SYSTEMS, OSRAM
  - → 10.4% of all companies with more than 100 transnational patents between 2005 and 2009.
  - → Partial assignment of the missing firms to enterprises (e.g. OSRAM, BSH).

- Matching algorithm only for the priority years 2005-2009 (reduction of data), but patent data is used for the period 1995-2009 → increased error rate in earlier years

- **F-Score matching** cannot reach 100%

Fraunhofer ISI

Identifying university-invented patents (instead of only university owned patents)

Dornbusch, F.; Schmoch, U.; Schulze, N.; Bethke, N. (2013): Identification of university-based patents: A new large-scale approach. In: Research Evaluation, 22 (1), S. 52-63.

Patent output of universities

- Since the end of the 1990s, most European countries have been moving away from the individual ownership of academic patents towards systems of **institutional ownership** by the universities (e.g. Geuna/Rossi 2011; Lissoni et al 2008).
- Germany had abolished the so called Professors Privilege in 2002
- However, there are still some ways of "**bypassing**" the university ownership
- In addition, contract and collaborative research may not appear as university patents
- Collaboration structures could be detected by analyzing the full scale of university patents

➢ **University owned vs. university invented**

- **Problem**: inventor affiliations are not listed on the patent
- **Solution 1**: adding affiliations by a name matching of authors and inventors
- **Solution 2**: tracking all inventors on university-owned patents by their IDs in the database

Fraunhofer ISI

1. Step: The matching algorithm - Identification of academic patents

- An approach for the identification and analysis of academic patents
- Basic idea: Match identical names of authors with university affiliation and inventors
  - Data sources: PATSTAT and SCOPUS

2) Organization | 3) Names | 4) Time | 5) Location | 6) Subject

$$x_{\text{uni-inv}} = 1 \text{ if } (a_{\text{names match}} + b_{\text{time match}} + c_{\text{location match}} + d_{\text{subject match}})$$

	Organization matching	Name matching	Time window matching	Location matching	Classification matching
PATSTAT	?	Full strings of last- and first name	Priority year	First two digits of the postcode*	IPC classification = WIPO 34
SCOPUS	Author affiliation = university	Full strings of last- and first name	Publication year: One year time-lag and time-window	First two digits of the postcode*	Scopus classification: fine- / coarse-grained

*= meanwhile NUTS3 Codes and distance matrix applied

See also: Dornbusch et al. 2013. Identification of university-based patents: A new large scale approach. Research Evaluation 22, 52-63.

Recall & Precision in identification of academic patents

Verification of matching results → Precision and Recall analysis:

- Recall → Percentage of university-owned patents covered by the algorithm:
- Precision → Online-Survey covering all authors for whom academic patents have been identified:
  - 1,681 person with 2,782 filings addressed
  - 435 exploitable answers (26%) received

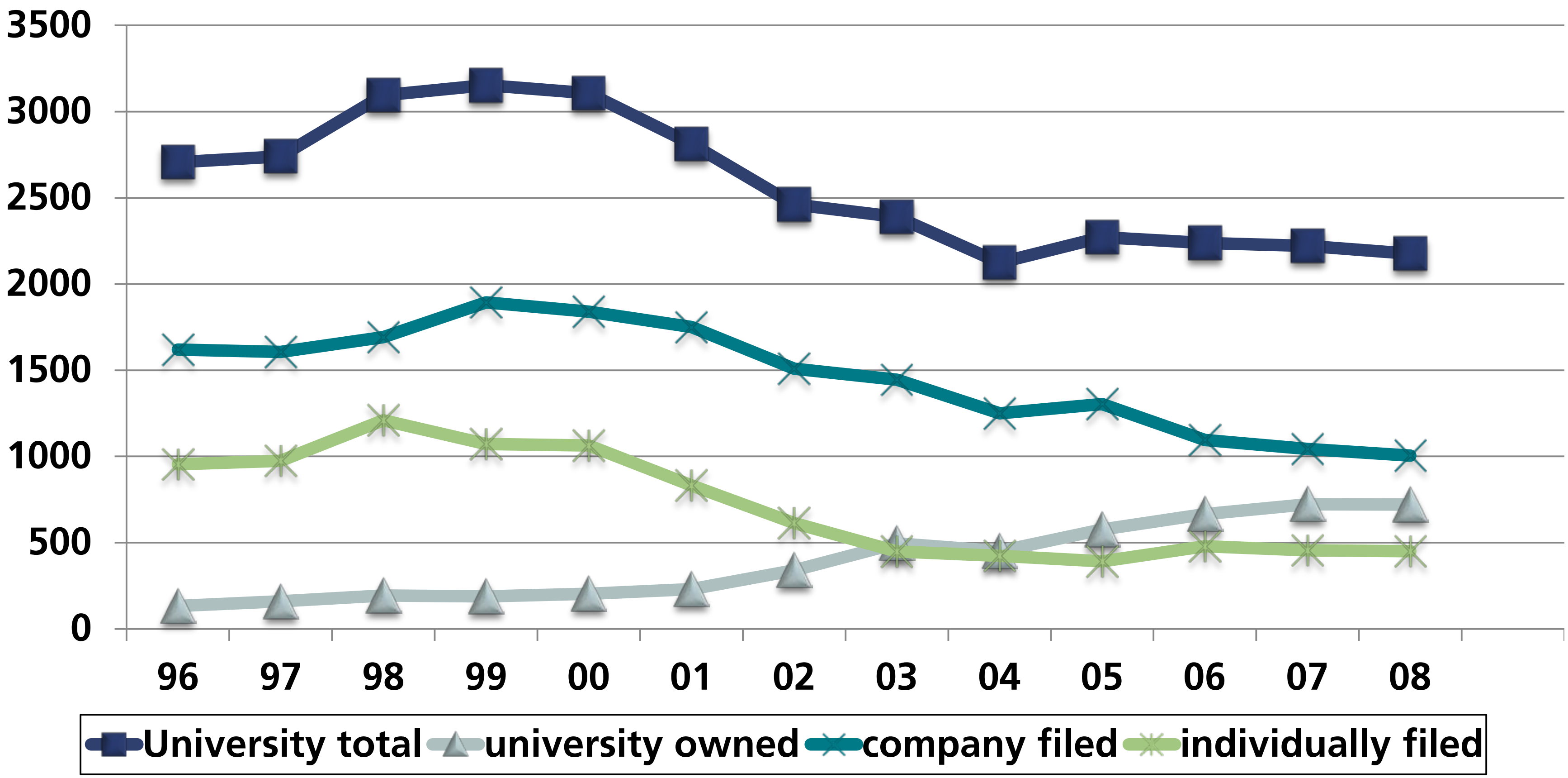
	Selection criteria	Recall	Precision	F-Scores		
				R=P (F_1)	P>R ($F_{0,5}$)	R>P (F_2)
	1-digit pc*	0,76	0,63	0,69	0,65	0,73
Standard criterion	**2-digit pc ***	0,71	0,77	$\boxed{0,74}$	0,76	0,72
	F-conc	0,71	0,52	0,60	0,55	0,66
	1-digit pc*, F-conc	0,64	0,82	0,72	0,78	0,67
High precision	**2-digit pc*, F-conc**	0,59	0,93	0,72	$\boxed{0,83}$	0,64
High recall	**2-digit* OR (1-digit* pc + F-conc)**	0,74	0,72	0,73	0,72	$\boxed{0,74}$

*= meanwhile NUTS3 Codes and distance matrix applied

Dornbusch et al. 2013. Identification of university-based patents: A new large scale approach. Research Evaluation 22, 52-63.

Fraunhofer ISI

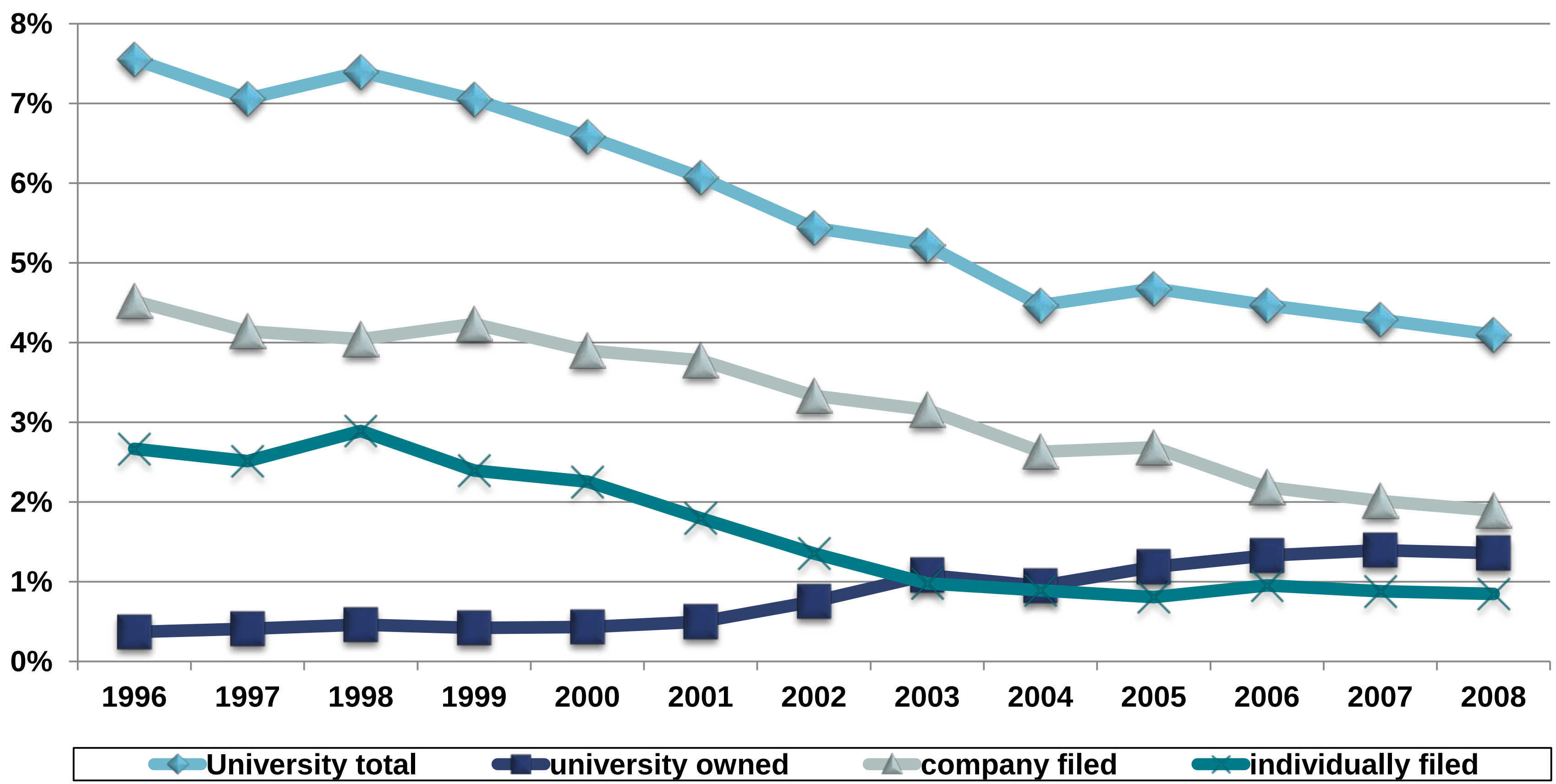
Absolute number of university patents in Germany



Source: EPO – PATSTAT; Elsevier – SCOPUS; Fraunhofer ISI calculations.

Fraunhofer ISI

Shares of university patents in Germany



Source: EPO – PATSTAT; Elsevier – SCOPUS; Fraunhofer ISI calculations.

Fraunhofer
ISI

Example 4:
Patent-Paper Twins

Background and motivation

- There are several studies that try to find similarities in patents OR publications or patents AND publications to identify similar scientific or technological fields
- Some studies are on the level of researchers/inventors (e.g. Meyer 2006)
- Some try to find twins based on general similarities (e.g Magermann et al. 2010; Magermann et al. 2012)

- **Technically speaking**: If you compare all patent abstracts with all publication abstracts, you will find a lot of similarities, but you might not be able to pin it to the same origin
- Therefore, we used a two stage approach to figure out what comes (probably) really out of the **same piece of research**

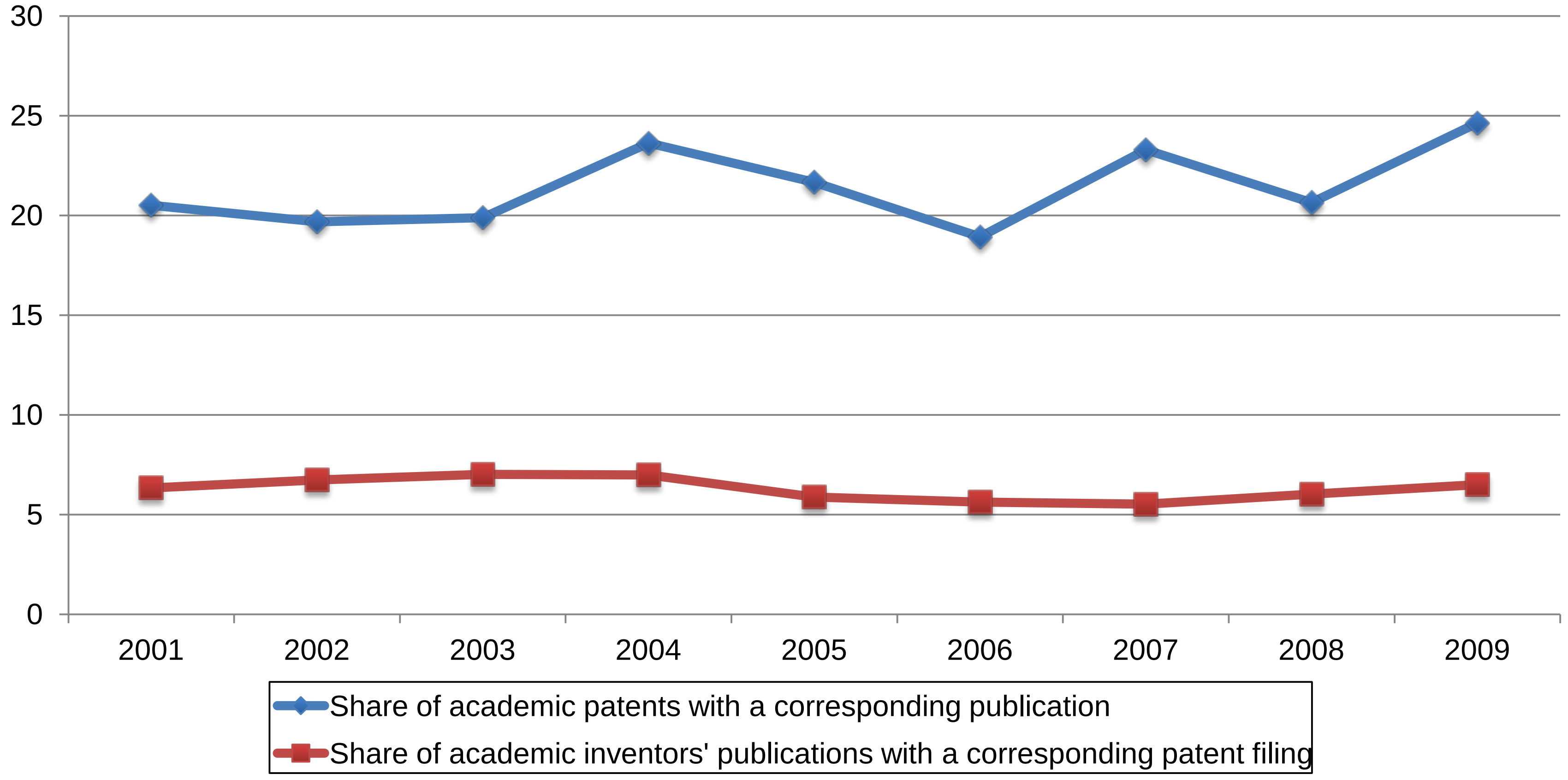
- Using the link on the inventor/author level, we identify **similar patents and publications** by the same inventors/authors
- We end up with **two datasets**
  - **one for patents** to address the first research question and
  - **one for publications** to address the second research question

Fraunhofer ISI

Content (cosine) similarity

- **Stop-word removal**: Common words having no distinctive meaning are removed
- **Stemming**: Stripping word-suffixes to combine word variants with shared meanings → „Porter Stemmer“ (van Rijsbergen et al. 1980; Porter 1980) applied
- **Cosine-similarity** between term vectors calculated: Inner product of two vectors divided by the product of their Euclidean norms → 1= similar vectors; 0 = unrelated vectors
- Patent-paper pairs of three author-inventors independently evaluated by three researchers → **Threshold** for cosine similarity used here is 0.6

Shares: academic patents with corresponding publications - and vice versa



Source: EPO – PATSTAT; Elsevier – SCOPUS; Fraunhofer ISI calculations.

Are academic publications with correspond. patents scientifically more valuable?

Publications

dV	Scientific regard		Int. allignment		No. of citations	
	β	sig	β	sig	∂y / ∂x	sig
patent_dummy	**0.056**	***	**-0.095**	***	-0.255	
Field_controls	YES		YES		YES	
Year_dummies	YES		YES		YES	
N	44262		49975		57278	
pseudo R^2					0.010	
R^2	0.007		0.112			

OLS & Neg.-bin regression
Source: EPO – PATSTAT, own calculations.
Significance Level: ***p<0.01, **p<0.05, *p<0.1, robust standard errors.

Source: EPO – PATSTAT; Elsevier – SCOPUS; Fraunhofer ISI calculations.