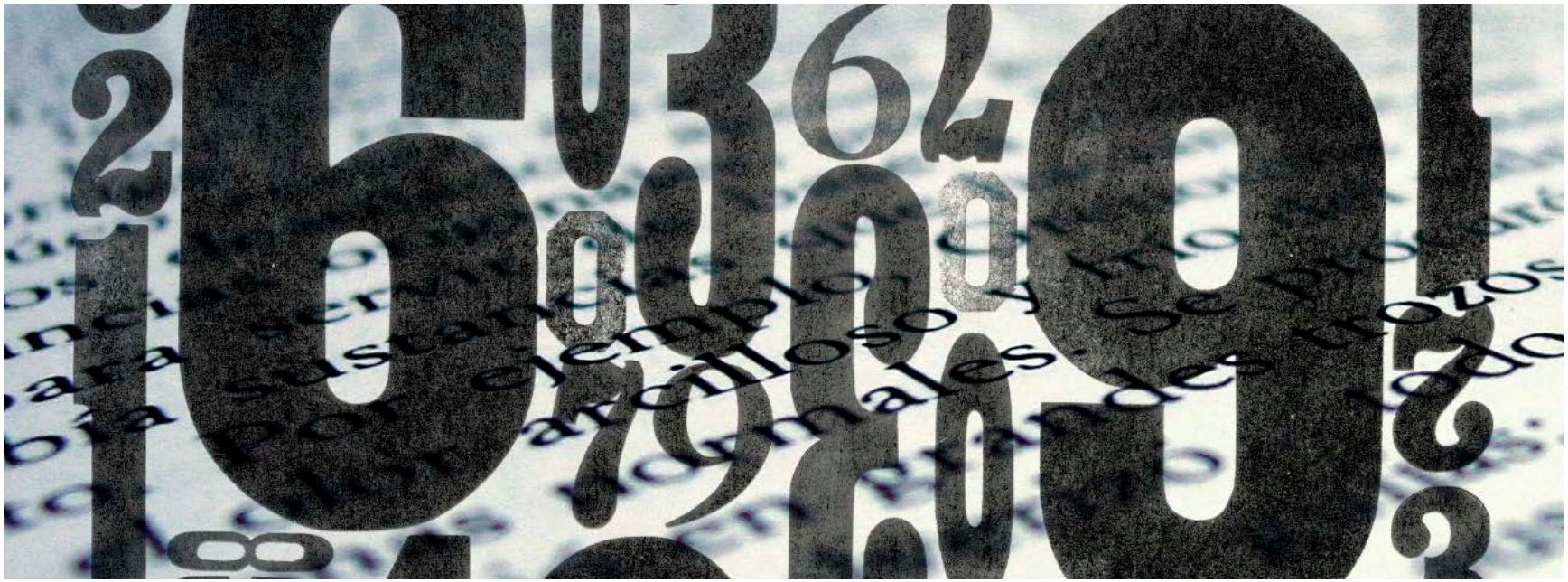
IDENTIFYING THE TECHNOLOGY PROFILES OF R&D PERFORMING FIRMS – A MATCHING OF R&D AND PATENT DATA

Peter Neuhäusler, Rainer Frietsch, Carolin Michels, Verena Eckl

4th Global Tech Mining Conference September 9th, 2014



Fraunhofer ISI

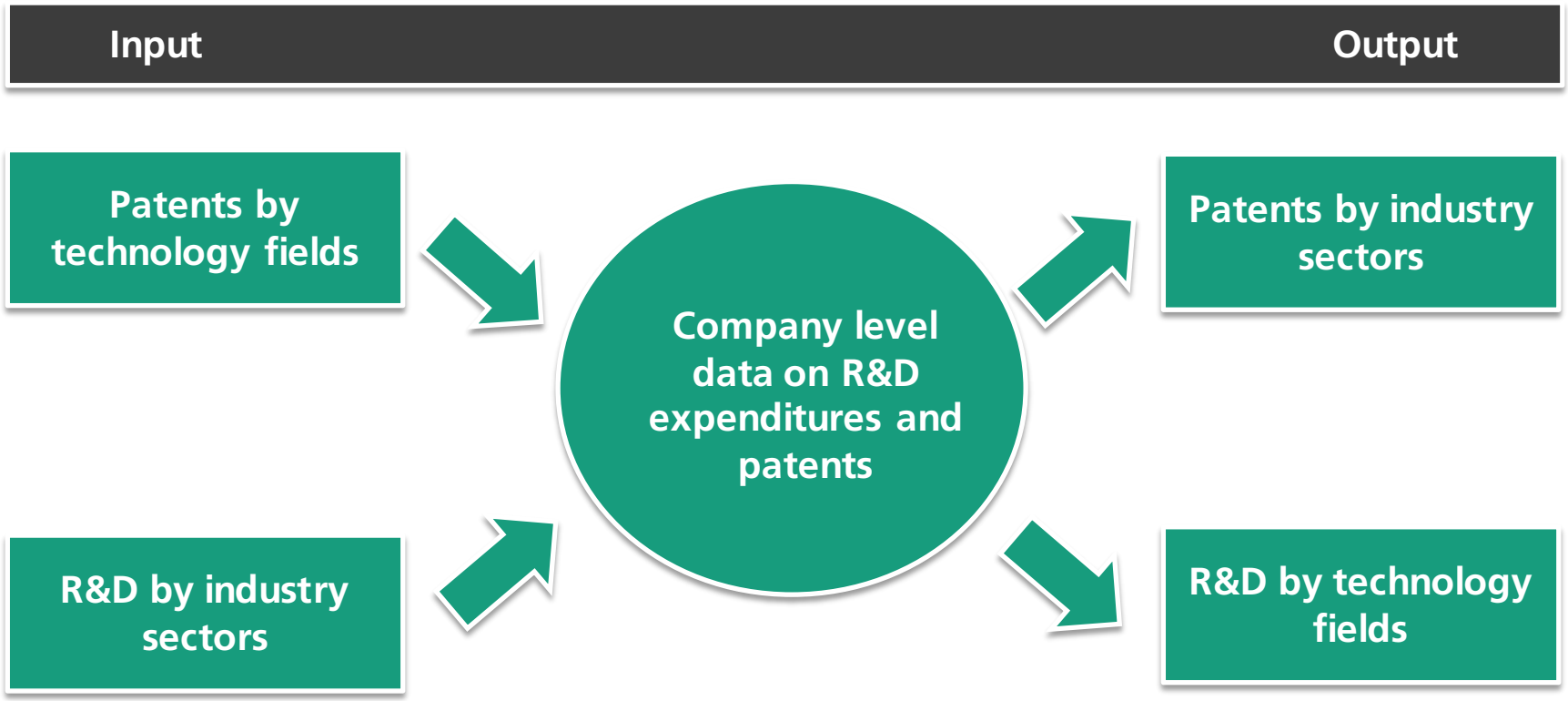
From industries to technologies – The motives of the research project

- **R&D expenditures** of companies are reported **by industries/sectors (NACE)** and cannot be provided at the level of technology fields.
  - Numerous **other indicators are only available at the level of technology fields** (e.g. support or funding programs or patents).
- **Problems**
  - Companies themselves can rarely provide information on expenditures by technology fields.
  - Large enterprises often are technologically heterogeneous → we do not know how much R&D is spent for certain kinds of technologies.
- **But…**
  - …micro data (company level) on R&D expenditures including a sector allocation
  - and micro data (applicant level) on patent filings including a technology field information are available.

➢ **Both micro databases can be merged** in order to identify R&D expenditures by technology fields and patents by sectors.

Fraunhofer ISI

Schematic representation of the research output



Fraunhofer ISI

The data matching – a brief overview

- **Aim**
  - Find information on patent applicants in PATSTAT that match (or are similar to) a company name in the R&D database and link the information from both databases.
  - R&D data by companies is provided by the R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009)

- **Approach**
  - Calculate the similarity of company entries from the R&D database to all applicants' names in PATSTAT (2005-2009). A certain degree of similarity determines the selection of the respective pair of R&D and PATSTAT entries as a "match".

Fraunhofer ISI

The matching procedure – Three steps

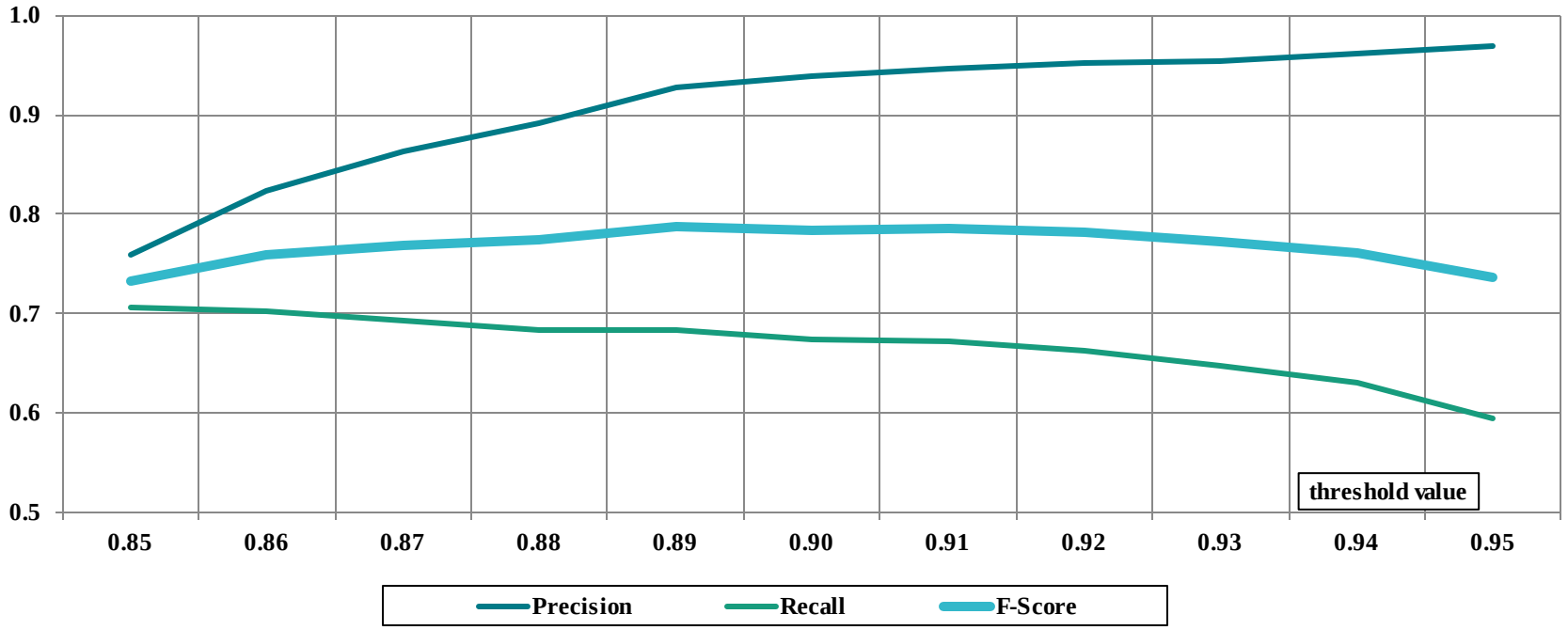
- **Cleaning the data**
  - Cleanup of different **spelling variations:** lowercase letters, special characters, spaces, removal of legal form of companies etc.

- **Similarity calculation**
  - **Levenshtein-Distance**: minimum number of edits to align two text strings.
  - In case the first three digits of the **ZIP code** (if available) do not match: similarity is set to 0.

- **Selection of the matched entries**
  - If the similarity of two text strings is higher than a predetermined threshold value, this is interpreted as a "match". The threshold is determined by Recall and Precision using a manually matched random dataset (N=1,000) as gold-standard.

Fraunhofer ISI

Selection of the threshold value

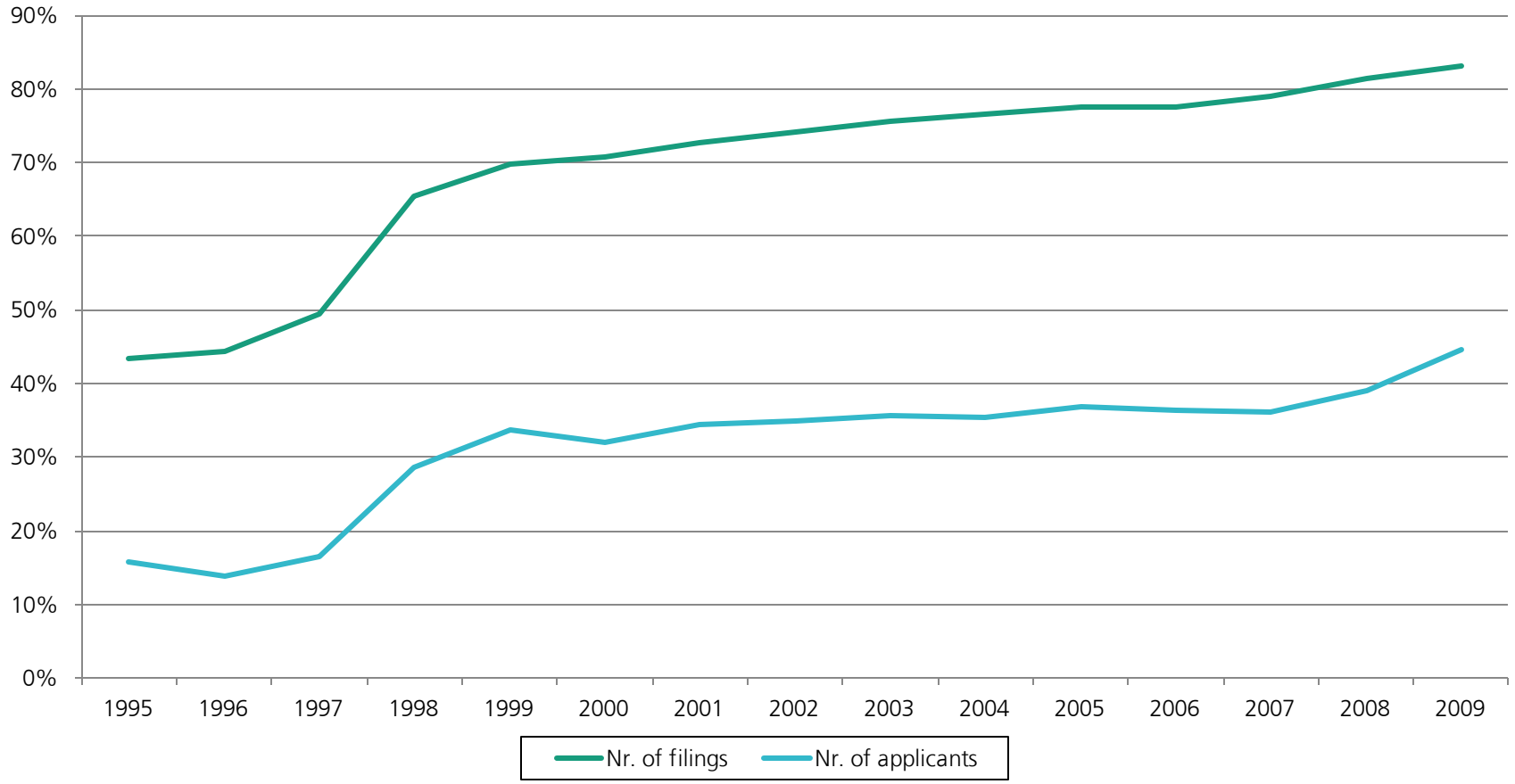


Source: R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT, calculations by Fraunhofer ISI.

F-score is highest at a threshold of t = 0.89. This threshold is the optimal compromise between Recall and Precision and is therefore used for the matching.

Fraunhofer ISI

Dataset coverage - proportion of matched filings in all filings, GPTO, 1995-2009



Source: R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT, calculations by Fraunhofer ISI.

Fraunhofer ISI

Weighting of the patent data

- It is not plausible that any patent requires the same amount R&D expenditures. **Differences between technology fields** are particularly evident.
  - The relation between patented and non-patented research results is different between the technology fields (difference in the propensity to patent).
  - R&D is expensive in some technology fields and less expensive in others.
- Especially large companies, with large amounts of R&D expenditures and patent filings, often are technologically **heterogeneous.**
- **Solution**: weighting of patents per technology field using an earlier estimation of the patent intensity based on Schmoch et al. (2003) and Schmoch and Gauch (2004).

Fraunhofer ISI

Weights of patents in technology fields – Indexed on the weight in the field "Transport" (= 100)

Technology field	Weight Transport=100	Equal weight
Electrical machinery, apparatus, energy	19.6	1
Audio-visual technology	21.5	1
Telecommunications	181.5	1
Digital communication	181.5	1
Basic communication processes	57.6	1
Computer technology	47.2	1
IT methods for management	47.2	1
Semiconductors	47.2	1
Optics	25.2	1
Measurement	38.5	1
Analysis of biological materials	38.5	1
Control	26.7	1
Medical technology	29.5	1
Organic fine chemistry	57.8	1
Biotechnology	116.1	1
Pharmaceuticals	261.3	1

Technology field	Weight Transport=100	Equal weight
Macromolecular chemistry, polymers	57.8	1
Food chemistry	57.8	1
Basic materials chemistry	57.8	1
Materials, metallurgy	82.9	1
Surface technology, coating	57.8	1
Micro-structural and nano-technology	289.0	1
Chemical engineering	57.8	1
Environmental technology	8.1	1
Handling	8.1	1
Machine tools	8.1	1
Engines, pumps, turbines	20.8	1
Textile and paper machines	51.7	1
Other special machines	8.8	1
Thermal processes and apparatus	8.8	1
Mechanical elements	8.8	1
Transport	100.0	1
Furniture, games	15.2	1
Other consumer goods	3.8	1
Civil engineering	8.8	1

Source: Schmoch and Gauch (2004), R&D survey by Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT, calculation by Fraunhofer ISI.

Fraunhofer ISI

Results of the weighting – the aggregate level

- **Two matrices** (weighted and unweighted) with the number of patent applications per technology field within the given economic sector.

- **For each economic sector:**
  - Calculation of the share of patents in a given technology field on all patents of a given sector (weighted and unweighted)
  - With the help of these shares, the aggregated R&D expenditures per sector can be distributed alongside the technology fields
  - Sum up the R&D expenditures per technology field

- **Finally:** Sum of R&D expenditures for each technology field (weighted and unweighted) as well as the matrices for the conversion → transferability to other R&D data sets

Fraunhofer ISI

Total business R&D expenditures of the German economy in 2009 by technology fields

Technology field	Weighted	Unweighted
Electrical machinery, apparatus, energy	1.851.049	4.109.213
Audio-visual technology	482.407	1.042.977
Telecommunications	2.730.352	716.179
Digital communication	2.963.248	832.286
Basic communication processes	380.063	280.259
Computer technology	1.697.118	1.804.062
IT methods for management	435.133	500.724
Semiconductors	1.220.529	1.276.022
Optics	413.107	866.413
Measurement	2.880.518	3.445.193
Analysis of biological materials	166.395	254.055
Control	770.377	1.303.496
Medical technology	1.001.225	1.632.027
Organic fine chemistry	1.713.663	2.276.137
Biotechnology	1.144.435	644.455
Pharmaceuticals	4.933.701	1.575.775

Technology field	Weighted	Unweighted
Macromolecular chemistry, polymers	660.088	707.858
Food chemistry	248.097	236.662
Basic materials chemistry	1.554.572	1.816.982
Materials, metallurgy	1.273.309	731.799
Surface technology, coating	1.028.208	801.933
Micro-structural and nano-technology	1.316.374	222.986
Chemical engineering	1.727.901	1.331.950
Environmental technology	175.333	1.016.653
Handling	321.576	1.512.352
Machine tools	372.674	1.925.051
Engines, pumps, turbines	1.988.154	4.489.604
Textile and paper machines	1.839.802	1.186.204
Other special machines	352.546	1.769.254
Thermal processes and apparatus	266.086	1.192.356
Mechanical elements	747.133	3.606.001
Transport	16.660.697	7.961.966
Furniture, games	269.898	657.272
Other consumer goods	164.747	1.082.777
Civil engineering	316.611	1.258.195

Source: R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT,calculations by Fraunhofer ISI.

Fraunhofer ISI

Qualification of the results

- To **qualify the results of the conversion**, available external information on R&D expenditures for certain technologies were used.

  - **A report by the BMBF (2011) about nanotechnologies in Germany** contains information about R&D expenditures of surveyed companies. Extrapolated values show expenditures of 1.3 billion Euros (compared to 1.316 in the field of "micro and nano structures" from our calculations).

  - **Information from business unions**: For Biotechnology between 780 million and 1.05 billion Euros compared to 1.1 billion Euros and for the pharmaceutical industry 5.1 to 5.4 billion Euros compared to 4.9 billion Euros.

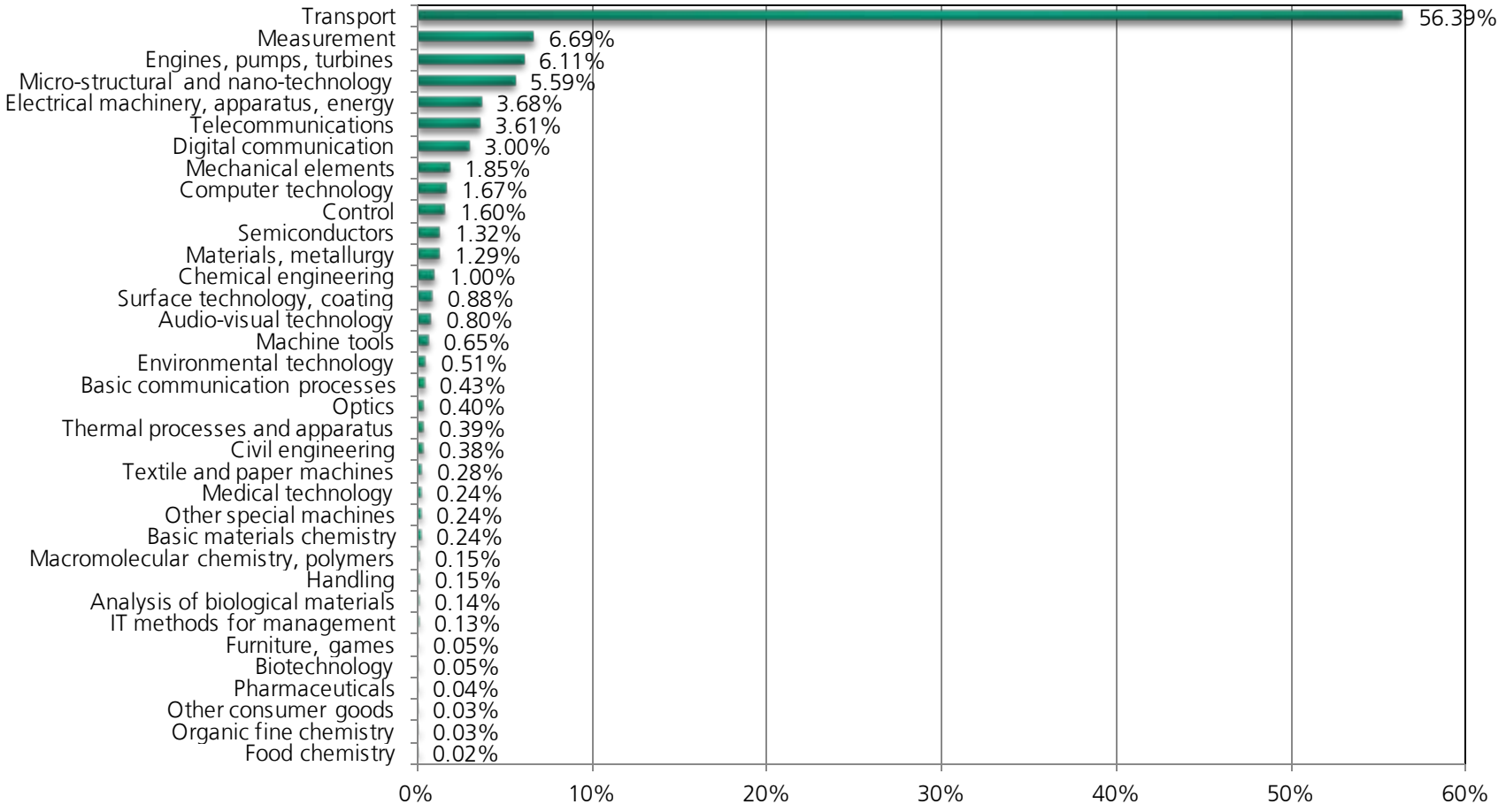
  - Comparison with the **applicant survey of the European Patent Office** (EPO 2010) reveals similar patent intensities in chemistry, biotechnology, polymers and electronics. Deviations in the fields of handling, human necessities and civil engineering. Also, higher R&D intensities in the transport sector (effect of industrial structure in Germany).

Fraunhofer ISI

Further opportunities arising from the conversion

Fraunhofer ISI

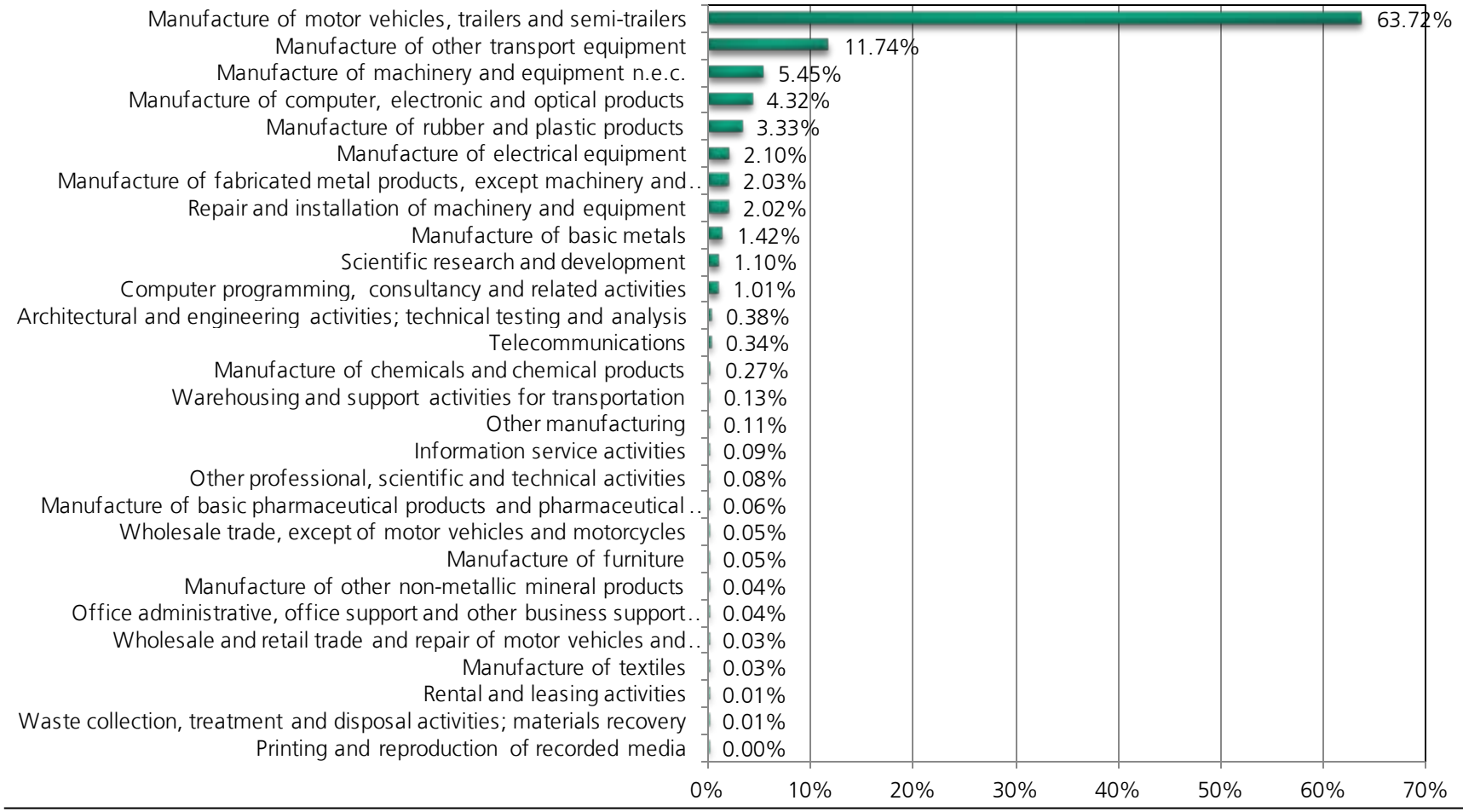
Total R&D expenditures in sector 29 "Manufacture of motor vehicles, automotive parts" by technology field



Source: R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT, calculations by Fraunhofer ISI.

Fraunhofer ISI

Total R&D expenditure in the technology field "transport" by economic sector



Source: R&D survey of the Stifterverband für die Deutsche Wissenschaft, 2007 and 2009, EPO – PATSTAT, calculations by Fraunhofer ISI.

Fraunhofer ISI

Caveats

- Sometimes **multiple entries of patent applicants** need to be assigned to one company entry → algorithm allows for that, but the error is increased. For companies that have several divisions but file their patents via their headquarters this is even amplified.
- Different matching algorithms (n-gram, soundex) could be used → tests revealed most to be very time- and resource intensive, Levenshtein distance lead to plausible results

- The **assignment of companies to sectors** is somewhat artificial. Especially larger "manufacturing" firms are often assigned to "wholesale trade" → manual recoding
- **Implicit assumption** that the relationship between R&D expenditures and patents within technology fields is more or less constant over time → regular updates are necessary

- **Different weighting schemes** could be used → weighting with the share of patent filings by field in all patent filings, however, lead to implausible results

- Applying the matching algorithm to **other sources** is possible, e.g. other company databases, patents/publications. Yet, the R&D data at the company level is confidential.

Fraunhofer ISI

Thank you!

Dr. Peter Neuhäusler

Fraunhofer Institute for Systems and Innovation Research ISI

Breslauer Straße 48
76139 Karlsruhe
www.isi.fraunhofer.de

Peter.Neuhaeusler@isi.fraunhofer.de

Fraunhofer ISI