

# Text Mining Patent Data

**Sam Arts**

Assistant Professor

Department of Management, Strategy, and Innovation

Faculty of Business and Economics

KU Leuven

[sam.arts@kuleuven.be](mailto:sam.arts@kuleuven.be)

OECD workshop: Semantic analysis for innovation policy

# Text mining patent data

- Patents
  - Key source of information to study technology and innovation at the individual, firm, or regional level
- Mostly “standard” structured information
  - Patent counts
  - Patent classification, citations
  - (Disambiguated) assignees, inventors, locations
- Patent = legal document detailed technical content
  - Very rich information
  - A bit more difficult to harness (size, unstructured, different spelling, errors, punctuation, ... )

# Applications

1. Patent similarity
2. Emergence and diffusion of new technologies
3. Identifying scientific prior art
4. Patent portfolio similarity

# Patent similarity

- Critical step in many studies
  - Are knowledge spillovers geographically localized (Jaffe, Trajtenberg, and Henderson, 1993)?
  - Does inter-firm mobility of engineers influence the transfer of knowledge between firms or regions (Singh and Agrawal, 2011; Almeida and Kogut, 1999)?
  - ...
- Practitioners (inventors, attorneys, patent examiners, and R&D managers)
  - Identify closely related prior art
  - Assess the novelty of a patent
  - Identify R&D opportunities in less crowded areas
  - Detect in- or out-licensing opportunities
  - ...

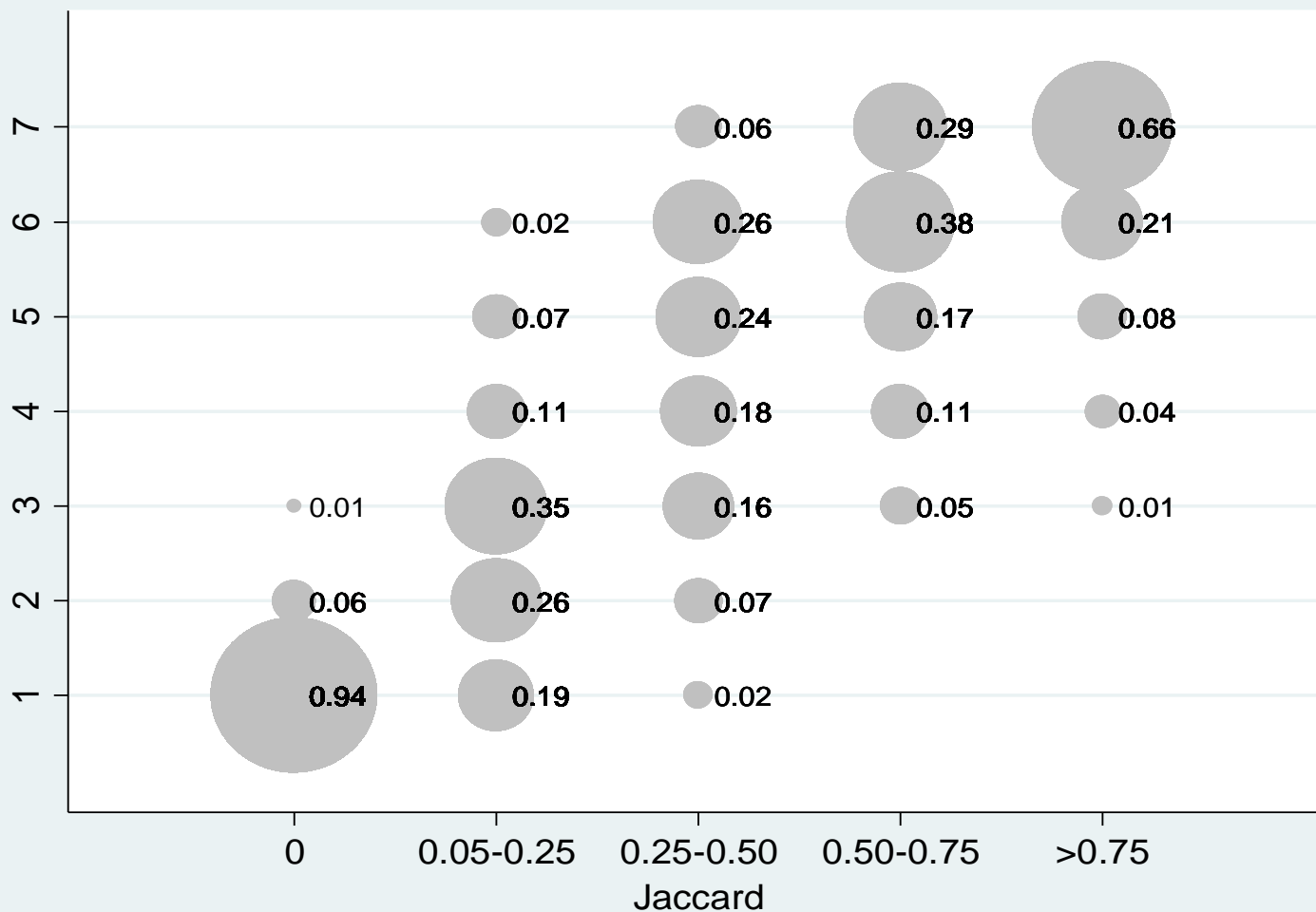
# Patent similarity

- Develop new measure of patent similarity based on text
- Validate new measure
  - External validation: expert assessments (13 experts from 5 fields)
  - Face validity: same patent family, assignee, inventors, cite each other
- Validate improvement over patent classification
  - External validation: expert assessments
  - Face validity: same patent family, assignee, inventors, cite each other
- See Arts S., Cassiman B., Gomez J. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39 (1), 62-84.

# Patent similarity

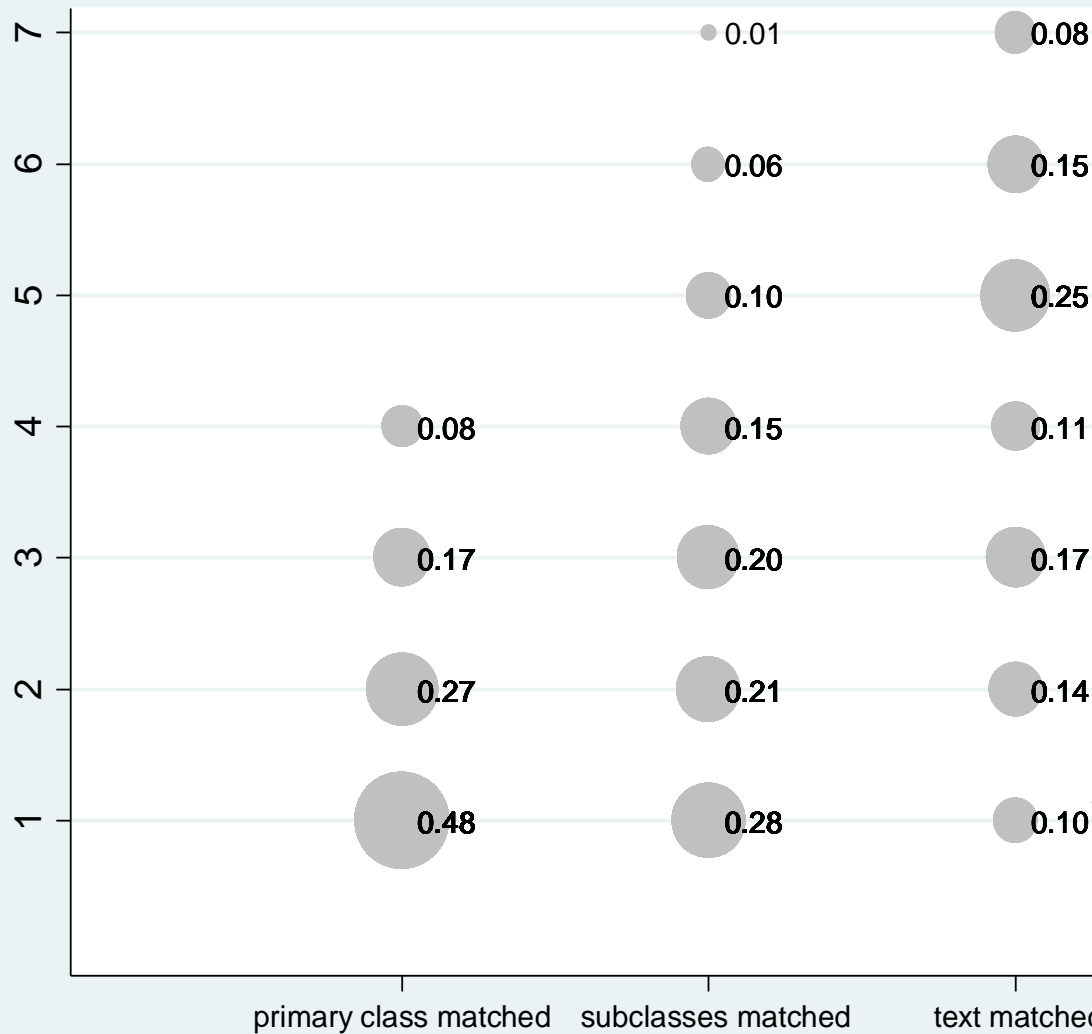
- Title and abstracts from all US utility patents granted between 1976-2013 (4.4 million)
- Concatenate title and abstract, lowercase, eliminate stop words (SMART system >600 words), words with 1 character, numbers, words which appear only once
- Each patent bag of unique keywords
- 526,561 keywords; avg 37 per patent

# Patent similarity: validation



Patent pairs with a larger Jaccard are also more like to belong to same patent family (docdb), inventor(s), assignee(s), and more likely to cite each other

# Validation improvement over patent classification

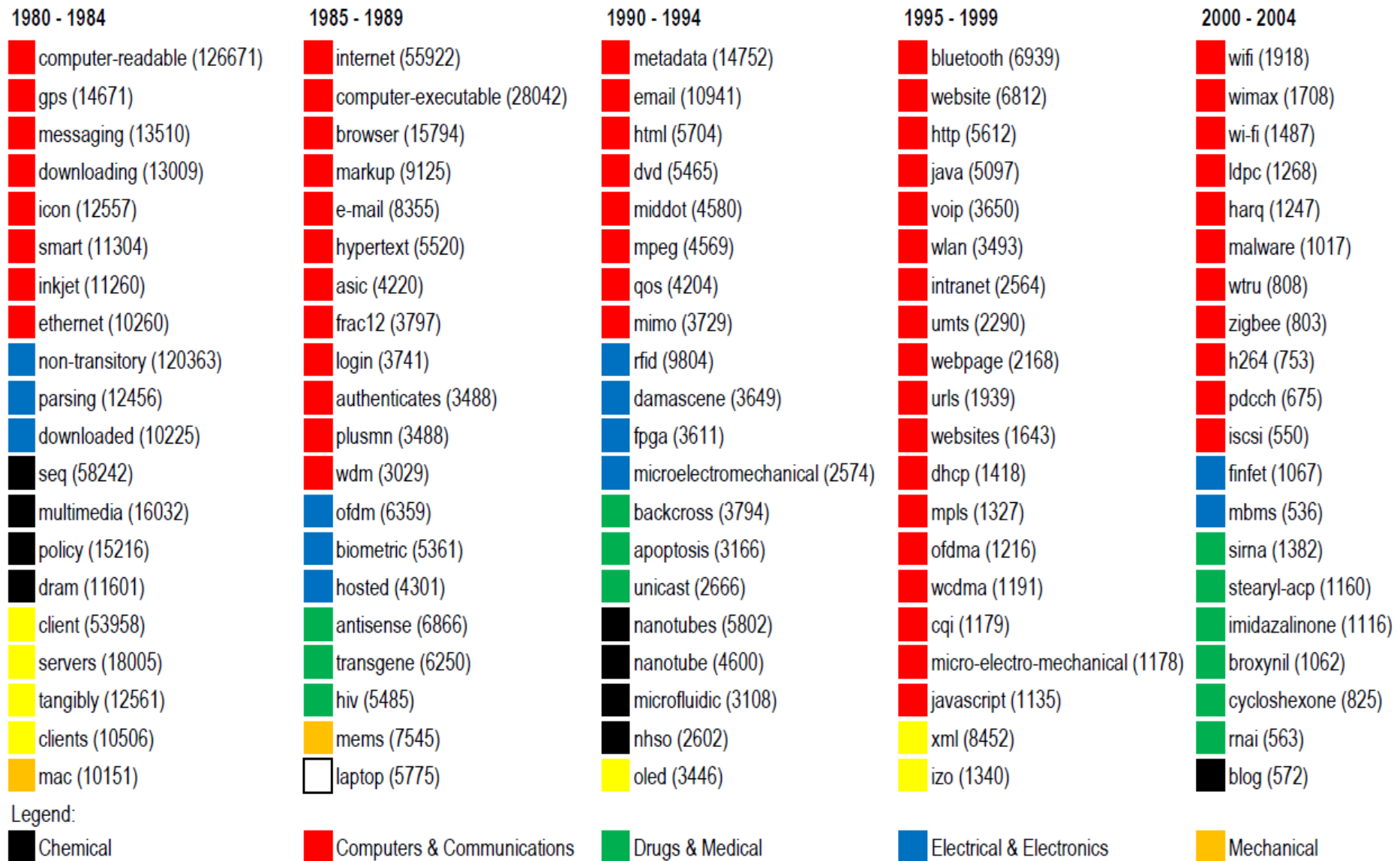


matched on general keywords (method, system, device, apparatus, process, material, image, sensor, ...)

**Text-matched patents are also more likely to belong to same patent family (docdb), inventor(s), assignee(s), and more likely to cite each other**



# Emergence and diffusion of new technologies



- Most reused novel words by five-year periods and colored by NBER technology category, 1980-2014.
- See Balsmeier et al. (2018). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. Forthcoming *Journal of Economics & Management Strategy*

# Emergence and diffusion of new technologies

- Industrial scientists and engineers who engage in academic boundary spanning (bench-level collaboration with academics) create more novel and valuable technology in industry
- See Arts S., Veugelers R. (2018). Taste for Science, Academic Boundary Spanning and Inventive Performance of Scientists and Engineers in Industry. *CEPR Discussion paper DP12704*

**TABLE 5: Mediation of Motives by Co-publications, Time Spent on Research, and Hours Worked**

	(1) Citation- weighted patents	(2) Citation- weighted patents	(3) Citation- weighted patents	(4) Citation- weighted patents	(5) Citation- weighted patents	(6) New words	(7) New words	(8) New words	(9) New words	(10) New words
Taste for science	0.612** (0.249)	0.461** (0.215)	0.508** (0.247)	0.595** (0.243)	0.380* (0.204)	0.718*** (0.223)	0.634*** (0.209)	0.672*** (0.229)	0.721*** (0.221)	0.610*** (0.217)
Taste for salary & career	-0.802* (0.423)	-0.388 (0.361)	-0.829* (0.425)	-0.741* (0.404)	-0.373 (0.373)	-0.302 (0.348)	-0.103 (0.335)	-0.305 (0.364)	-0.243 (0.336)	-0.065 (0.337)
Co-publications		0.034*** (0.007)			0.031*** (0.008)		0.028*** (0.008)			0.025*** (0.009)
Research			0.011* (0.006)		0.007 (0.008)			0.006 (0.007)		0.003 (0.007)
Hours worked				0.035** (0.015)	0.022 (0.015)				0.036* (0.019)	0.025 (0.020)
Log likelihood	-5383.628	-4656.485	-5205.093	-5268.610	-4548.289	-1212.691	-1156.376	-1205.659	-1193.241	-1145.243

Notes: The sample includes 464 industrial scientists and engineers. Models are estimated with Poisson quasi-maximum likelihood, estimated with exposure to account for differences in job tenure. All models include controls for age, age<sup>2</sup>, female, Belgian, married, children, pre sample patents, pre sample publications, pre sample publication citations, Ph.D. scholarship, time to Ph.D., Ph.D. abroad, firm patents, firm co-publications, firm name missing, job tenure > 10, and Ph.D. field fixed effects. Robust standard errors in brackets, clustered at firm level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

# Identifying scientific prior art in patents

- Machine learning techniques to classify non-patent references as scientific (publications, conference proceedings)
  - See Callaert, J., Grouwels, J., & Van Looy, B. 2011. Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, 91(2): 383-398.
- Match NPRs to WOS publications

# Identifying scientific prior art in patents

- Scientific publications are particularly important for inventors who explore new fields and lack prior field-specific expertise
- See Arts S., Fleming L. (2018). Paradise of Novelty – or Loss of Human Capital? Exploring New Fields and Inventive Output. Provisionally accepted *Organization Science*

**TABLE 2: Inventor-firm fixed effects models exploring new fields**


	(1)	(2)	(3)	(4)	(5)
	Forward cit. (ln)	Forward cit. (ln)	Forward cit. (ln)	Forward cit. (ln)	Forward cit. (ln)
Exploring new fields	-0.0104*** (0.0016)	-0.0118*** (0.0018)	-0.0306*** (0.0023)	-0.0250*** (0.0019)	-0.0407*** (0.0024)
Exploring new fields*Expert team			0.0793*** (0.0030)		0.0757*** (0.0030)
Exploring new fields*Science				0.0586*** (0.0033)	0.0525*** (0.0033)
Expert team			0.0423*** (0.0023)		0.0435*** (0.0023)
Science				0.0673*** (0.0022)	0.0687*** (0.0022)
R-squared	0.036	0.066	0.067	0.067	0.068

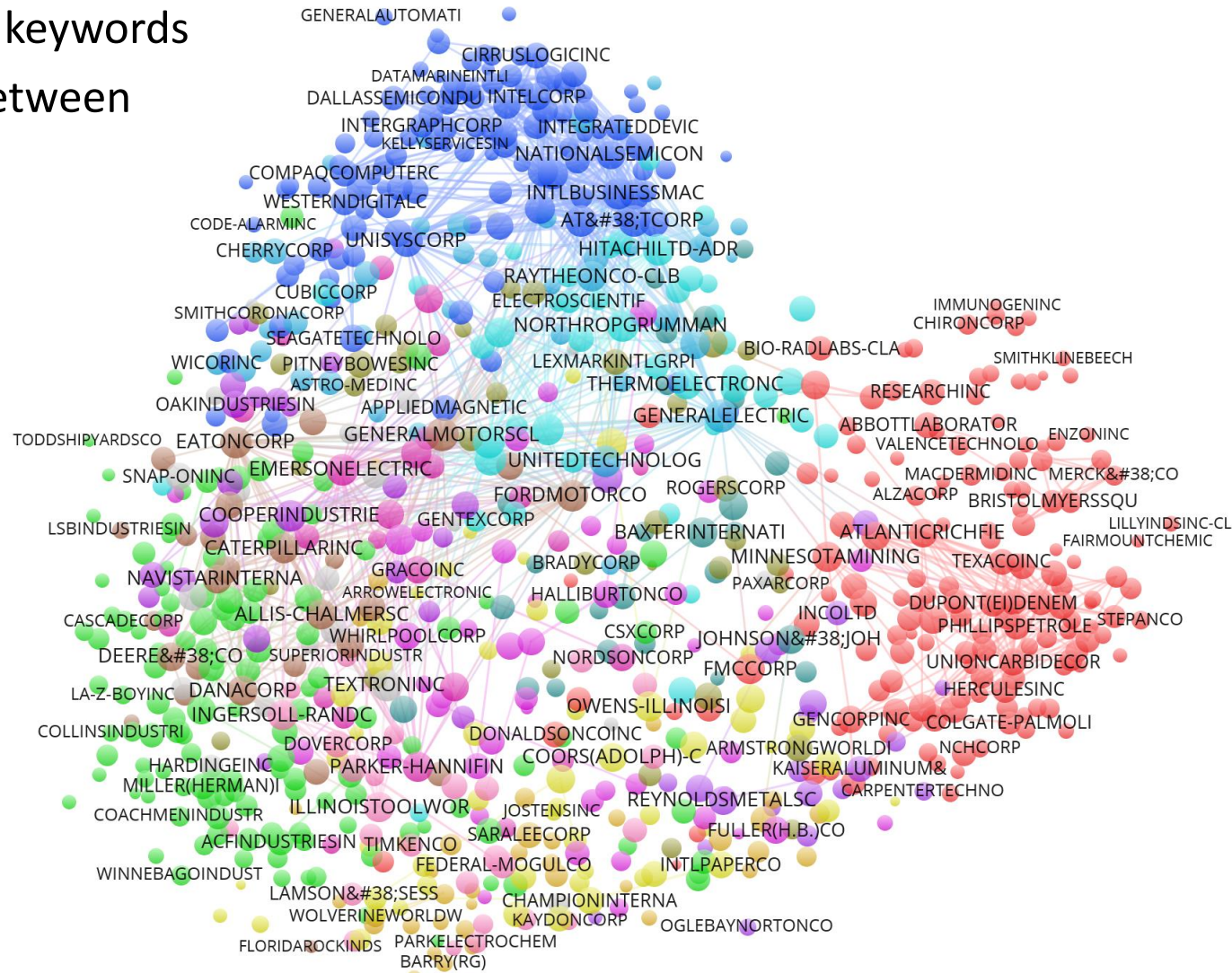
*Notes:* The sample includes all patents of inventors that are assigned to a firm, filed and granted between 1975 and 2002. The sample is restricted to inventors with at least one prior patent (and hence are at the risk of exploring new fields), and have at least two patents assigned to the same firm. The sample includes 2,705,431 inventor-patent observations. All models include controls for prior patents (ln), specialization, prior collaborations (ln), move, prior move, days since last patent (ln), days since first patent (ln), team, team prior patents (ln), team specialization, firm prior patents (ln), number of classes, number of subclasses, year and technology class indicators, and inventor-firm level fixed effects. Robust standard errors in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

# Patent portfolio similarity

- Critical step in many studies
  - How do knowledge spillovers affect R&D and productivity of firms (Bloom, Schankerman, and Van Reenen, 2013)?
  - Are M&As between firms with similar technological knowledge more successful (Makri, Hitt, and Lane, 2010)?
  - ...
- Practitioners (inventors, R&D managers)
  - Map companies in technology space
  - Find acquisition targets
  - ...

## Patent portfolio similarity

- firm  $i = (v_{i1}, v_{i2}, \dots, v_{in})$
  - $v_{in}$  = number of patents of firm  $i$  with keyword  $n$
  - $n = 526,561$  unique keywords
  - Cosine similarity between two firms
- 



Items: 826 firms

Clusters: 13

Min. cluster size: 25

Links: 340,206

Weights: Total link strength

# Text mining patent data

- Open access to all code and cleaned data
  - <https://dataverse.harvard.edu/dataverse/patenttext>
  - [https://github.com/jcgcarranza/smj\\_code](https://github.com/jcgcarranza/smj_code)



Thank you!

[sam.arts@kuleuven.be](mailto:sam.arts@kuleuven.be)