Mapping the Problem Space of Innovation

Extended Abstract

Fabian Gaessler

Universitat Pompeu Fabra, Barcelona · Max Planck Institute for Innovation & Competition, Munich

fabian.gaessler@upf.edu

Introduction

The idea that inventions are solutions to specific technical problems is foundational to innovation research—yet there is no systematic way of measuring which problem an invention addresses. The problem that a battery overheats, for instance, may be solved through inventions in electrochemistry, materials science, or mechanical engineering. In existing patent data these solutions appear unrelated, because the data are organized around the technology employed.

That organizing principle is shared by every established patent measure (for a recent overview, see Bergeaud, Jaffe, and Papanikolaou, 2025). Technology classes record what an invention is. Citations record which documents an examiner or an applicant considered close. Text similarity, whether from keywords or from a neural encoder, is computed over titles, abstracts, and claims, the parts of the document that describe the artifact. Every one of these measures the solution. The problem dimension goes unrecorded. Two patents that address the same technical problem through different technologies sit in different classes, do not cite each other, and use different vocabulary. Conversely, patents that employ similar technologies for different purposes appear related.

This missing problem dimension forecloses a set of questions innovation research has long wanted to ask. Which technical problems attract competing solutions from different fields, and when one field displaces another. How broadly a firm searches across alternative solution paths for a problem it has chosen to work on. Whether a firm's or an inventor's advantage lies in recognizing problems or in solving them. Each requires holding the problem fixed and observing variation in the solution, which existing data do not permit.

This paper constructs a dataset that makes the link between inventions and the problems they address visible at scale, and validates it. The unit of measurement is the applicant's stated technical problem: the shortcoming in the prior art that the invention is put forward to fix. Recovering it unfolds the single idea space of the recombination literature (Fleming, 2001; Uzzi et al., 2013; Kelly et al., 2021) into a problem space and a solution space, and gives an empirical counterpart to the theoretical treatments in which the problem is itself a choice variable (Cyert and March, 1963; Carnehl and Schneider, 2025). Earlier work has recovered problem-like structure from patent text within a single domain and on samples of a few hundred documents (Giordano et al., 2023; Trapp and Warschat, 2024). The contribution here is population scale, comparability across technological fields, and validation against an independent standard.

Data and Measurement

Every granted patent has a stated technical problem, it is written in the applicant's own words, and it is recoverable from the document alone. This holds for the European system as a matter of institutional design. Rule 42(1)(c) EPC requires the description to disclose the invention in such terms that "the technical problem, even if not expressly stated as such, and its solution can be understood", and inventive step is assessed against an objective technical problem reconstructed from the closest prior art. The applicant therefore has to articulate it.

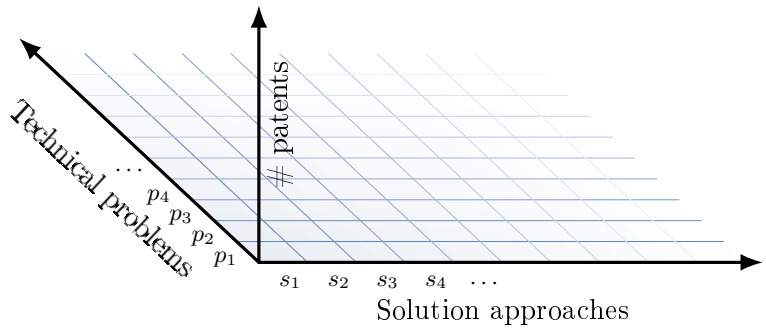


Figure 1. Stylized representation of the innovation space. The horizontal dimensions span technical problems and solution approaches; the vertical dimension indicates the number of patents for a given problem–solution combination. Established patent measures observe the solution axis alone.

Articulation still leaves the statement scattered through a long document. Fewer than one European patent in ten states the problem under a heading of its own; it typically sits in a background passage a few paragraphs into a description that runs tens of pages. Recovering it therefore calls for reading the text. A large language model reads the description and returns the problem statement together with the sentences of the source text on which it rests, so that every statement can be traced back to the document. The statements are then embedded and grouped, so that patents addressing the same or closely related problems fall together. Two properties of the output matter here: it is grounded in the applicant's own text, and it is produced identically for every patent in the corpus.

The corpus is the population of granted European patents, filing years 1978–2024, about 2.46 million documents. A random draw of 100,000 patents serves as the analysis object. Their problem statements fall into 1,231 problem clusters, with a median of 84 patents each.

Validation

A measure produced by a large language model carries no weight until it is shown to track a judgment made independently of it. The reference instrument is a frozen benchmark of 27,410 patent pairs, drawn along citation and technology-class axes. Each pair is labeled by a blind same-problem judge shown only each patent's title, abstract, and problem evidence, with the verdict Different, Borderline, or Same.

The representation orders those verdicts at AUC 0.956, against 0.926 for the same problem statements under a conventional embedder. The ordering reproduces under three independent judge lineages from different organizations, and it holds *within* a technology field, which rules out the judge and technological distance as explanations. A model-free cross-check, where an applicant places a citation in the document, corroborates the ordering.

Figure 2 sets the measure against the established alternatives on the same pairs. Conventional patent measures predict a shared problem weakly: technology classes reach an average precision of 0.25, keyword overlap 0.40, and a title-and-abstract neural encoder 0.46, against a base rate of 0.14. Working from the problem text raises this to 0.85. That margin is the quantity the dataset supplies.

Descriptive Evidence

A problem grouping invites the immediate objection that it reproduces the technology classification under another name. Placing each problem cluster on an axis anchored at both ends answers it: zero if all its patents share a single technology subclass, one if it covers as much technology as a size-matched random draw of patents. The observed median is 0.41 $[0.39, 0.43]$, well away from

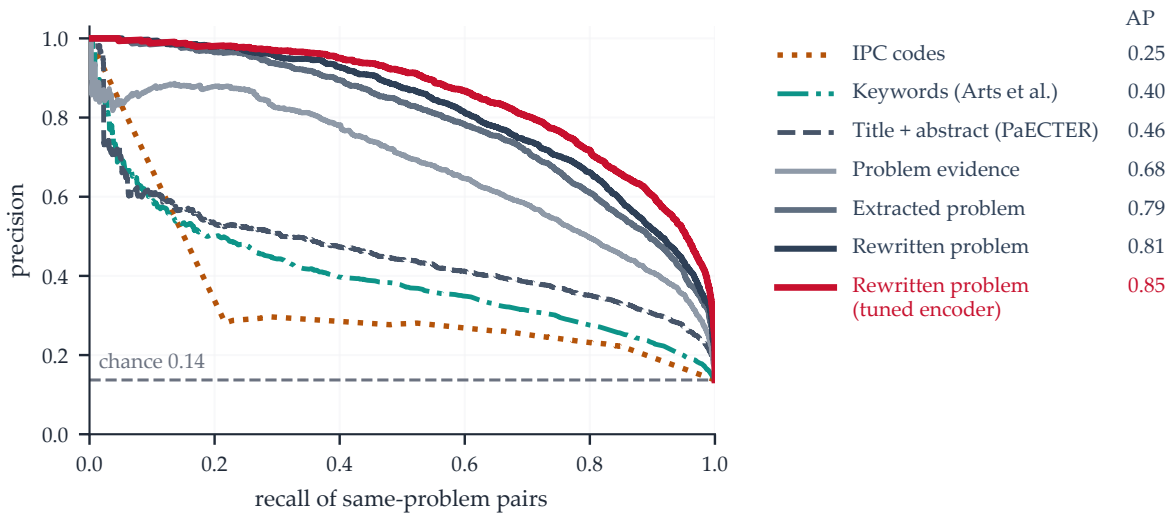


Figure 2. Predicting same-problem pairs. Precision–recall over the frozen benchmark, all series computed on the same 23,839 pairs; AP is average precision. The upper curves are built from the extracted problem text, the lower ones from information already in patent databases.

both anchors and closer to the technology-blind end. A median cluster spans 26 of 633 technology subclasses where a size-matched random draw spans 72.

Holding a problem fixed makes technological substitution directly observable. In one cluster read patent by patent, persistent electrostatic charge on moving or contacting components in imaging devices, filings are almost entirely photographic-film chemistry until the late 1990s and almost entirely camera-module electronics thereafter, with the two sets of firms barely overlapping. The problem persists while the technology answering it is replaced. The pattern is invisible to any measure organized on the solution axis.

At the firm level the two axes move together without collapsing into one (Figure 3). Firms that share technology usually share problems, at a rank correlation of 0.85. The pairs off that diagonal are the ones no technology-based measure can reach. Airbus and Yamaha Motor, aircraft and motorcycles, share almost no technology and a great many problems, among them vibration, fatigue, and lightweight structures. Fujifilm and Ethicon sit the other way round, close in medical instrument technology and far apart in the problems they address.

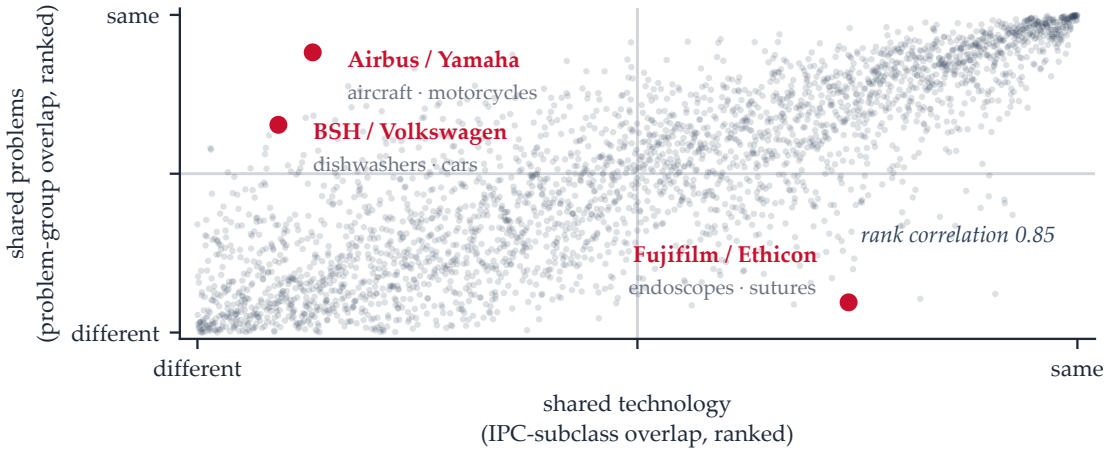


Figure 3. Shared technology against shared problems, for pairs of firms. Each point is one of the 2,919 pairs among the 96 firms holding at least 100 patents in the random 100,000-patent analysis object, placed by where its overlap ranks among all pairs on each axis.

Conclusion

The paper makes three contributions. It introduces a representation of invention that adds the problem dimension to the solution dimension patent data already record (Figure 1). It delivers a validated, large-scale dataset in which the European patent population is organized by the problems its inventions address. And it yields empirical measures without equivalent in existing data, at the three natural levels of analysis: technological substitutes at the problem level, inventions that compete because they answer the same problem, whatever technology they employ; the breadth of a firm's search across alternative approaches to a problem it has chosen; and, at the inventor level, specialization in recurring problems as against recurring methods, which speaks to team composition and the division of inventive labor.

Beyond academic research, knowing which technical problems attract many competing solutions and which remain underserved bears on the allocation of public R&D funding, on the design of mission-oriented innovation programs, and on how competition authorities define markets in technology-intensive industries.

The dataset will be made publicly available. The release comprises the problem statement extracted for each patent, the evidence sentences that support it in the source document, and the problem clusters together with their labels, keyed to patent publication numbers so that they link to standard patent data.

References

Arts, S., B. Cassiman, and J.C. Gomez (2018). Text Matching to Measure Patent Similarity. *Strategic Management Journal*, 39(1): 62–84.

Bergeaud, A., A.B. Jaffe, and D. Papanikolaou (2025). Natural Language Processing and Innovation Research. *NBER Working Paper* No. 33821.

Carnehl, C. and J. Schneider (2025). A Quest for Knowledge. *Econometrica*, 93(2): 623–659.

Cyert, R.M. and J.G. March (1963). *A Behavioral Theory of the Firm*. Prentice-Hall.

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science*, 47(1): 117–132.

Giordano, V., G. Puccetti, F. Chiarello, T. Pavanello, and G. Fantoni (2023). Unveiling the Inventive Process from Patents by Extracting Problems, Solutions and Advantages with Natural Language Processing. *Expert Systems with Applications*, 229: 120499.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, 3(3): 303–320.

Trapp, S. and J. Warschat (2024). LLM-Based Extraction of Contradictions from Patents. In *World Conference of AI-Powered Innovation and Inventive Design*, IFIP Advances in Information and Communication Technology 735. Springer, 3–19.

Uzzi, B., S. Mukherjee, M. Stringer, and B. Jones (2013). Atypical Combinations and Scientific Impact. *Science*, 342(6157): 468–472.