Ganglmair, Bernhard; Robinson, W. Keith; Seeligson, Michael

Working Paper

The rise of process claims: Evidence from a century of U.S. patents

ZEW Discussion Papers, No. 22-011

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Ganglmair, Bernhard; Robinson, W. Keith; Seeligson, Michael (2022) : The rise of process claims: Evidence from a century of U.S. patents, ZEW Discussion Papers, No. 22-011, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/251955

// NO.22-011 | 04/2022

DISCUSSION PAPER

// BERNHARD GANGLMAIR, W. KEITH ROBINSON,
AND MICHAEL SEELIGSON

The Rise of Process Claims: Evidence From a Century of U.S. Patents

Leibniz Association

ZEW

The Rise of Process Claims: Evidence from a Century of U.S. Patents*

Bernhard Ganglmair† W. Keith Robinson‡ Michael Seeligson§

First version: February 22, 2022
This version: April 2, 2022

Abstract

We document the occurrence of process claims in granted U.S. patents over the last century. Using novel data on the type of independent patent claims, we show an increase in the annual share of process claims of about 25 percentage points (from below 10% in 1920). This rise in process intensity is not limited to a few patent classes but can be observed across a broad spectrum of technologies. Process intensity varies by applicant type: companies file more process-intense patents than individuals, and U.S. applicants file more process-intense patents than foreign applicants. We further show that patents with higher process intensity are more valuable but are not necessarily cited more often. Last, process claims are on average shorter than product claims; but this gap has narrowed since the 1970s. These patterns suggest that the patent breadth and scope of process-intense patents are overestimated when claim types are not accounted for. We conclude by describing in detail the code used to construct the claim-type data, showing results from a data-validation exercise (using close to 10,000 manually classified patent claims), and providing guidance for researchers on how to alter the classification outcome to adapt to researchers' needs.

Keywords: innovation; patent claims; patents; patent breadth; patent scope; process claims; process intensity; R&D; text analysis.

JEL Codes: C81; O31; O34; Y10

*This project has been a few years in the making. Many of our colleagues and collaborators have contributed to this project by providing thoughtful comments and suggestions, or by simply lending us a patient ear when things did not go the way we expected. Thank you! We thank Po-Hsuan Hsu, Kenneth Huang, Aija Leiponen, Neel Sukhatme, Mike Teodorescu, and participants at PatCon7 (Northwestern University, 2017), IPSDM (Mexico City, 2017), the NBER Innovation Information Initiative Working Group Meeting (2021), and WIPIP (St. Louis, 2022) for helpful comments and suggestions. We also thank Jacob Colling, Maximilian Schneider, Lion Szlagowski, Jake Walsh, and Tianxiang Zhang for their research assistance. Institutional Review Board exemptions (on file with the authors) for data collection via Amazon Mechanical Turk have been obtained in 2016 at the University of Texas at Dallas (MR 16-217) and Southern Methodist University School of Law (2016-052-ROBW).

†University of Mannheim and ZEW Mannheim, Germany. E-mail: b.ganglmair@gmail.com.

‡Wake Forest University, School of Law, Winston-Salem, NC, USA. E-mail: robinswk@wfu.edu.

§Southern Methodist University, Cox School of Business, Dallas, TX, USA. E-mail: seeligson@gmail.com.

1 Introduction

Patents are widely used in the economics and management literature as outcome measures for innovation and R&D. Recently, scholars have discovered the text of the patent documents (beyond the title, abstract, and the references) as a source for more details on those very R&D outcomes. The content of a patent is indicative of the type of invention that is protected by the patent, such as, for instance, a process or a product. In this paper, we document the process intensity of U.S. patents between 1920 and 2020 using a novel dataset with results from a computer-assisted patent classification.[1] We show that the process intensity of patented R&D results has more than tripled. Process intensities vary across different technologies and application types. Moreover, we document that process intensity is associated with indicators of higher patent value.

For a classification of the invention, one would ideally look at the text of the patent specification. The patent statute requires that a patentee's application completely disclose her invention and enable one of ordinary skill in the art to make or use the invention without undue experimentation. These written descriptions, however, are in the form of unstructured language – in prose, with little formatting – and pose a challenge for computer-assisted classification. Instead, as our source of information we use patent claims, the object of recent work on measuring innovation outcomes (e.g., Kuhn and Thompson, 2019; Marco, Sarnoff and deGrazia, 2019). Patent claims define the patented invention and are written in more structured language. For instance, each claim must be written as a single sentence and contains three parts: a preamble, a transitional phrase, and a body. And because a patent's detailed specification must support the claim language, a patent's claims are accurate representations of the type of the patented invention.

To classify patent claims, we combine information obtained from both the preamble and the body of a claim. The preamble is a general description of the invention (e.g., a method, an apparatus, or a device), whereas the body identifies steps and elements (specifying in detail the invention laid out in the preamble) which the patentee is claiming as the invention. The combination of the preamble type and the body type provides us with a more detailed and more accurate classification of claims than other approaches in the literature. This approach also accounts for unconventional drafting approaches. We validate our classification using close to 10,000 manually classified claims.

Our paper is structured as follows: In Section 2 we provide a primer to patent claims and patent claim drafting. We give an overview of patent claims (parts of a claim, different

[1] In this paper, we describe the data for patent claims issued between 1920 and 2020. The published data files (at https://doi.org/10.5281/zenodo.6395308) also contain classification results for all independent patent claims issued between 1836 and 1919.

claim classes, and claim types), legal issues surrounding patent claims, and a summary of current practices for claim drafting given the most recent legal developments.

In Section 3 we document the rise of process claims over the last century. We show that the process-intensity of U.S. patents has more than tripled, from just below 10% in 1920 to more than 30% in 2020 (Lesson 1). Process intensity is highest in chemical and drugs & medical patents and lowest in mechanical and other patents (Lesson 2). Over the course of the century, both changes of process shares across the broad spectrum of technologies and changes in the composition of technologies (with a shift of patenting toward more process-intense patent classes) were equally important drivers of the increase of process intensity, with their respective roles changing over time (Lesson 3). We further show that different applicant types exhibit different process intensities. Patents granted to companies are more process-intense than those granted to individuals (Lesson 4); and patent granted to U.S. applicants are more process-intense than those granted to foreign applicants (Lesson 5). Patent value also depends on process intensity. The value of patents (Kogan, Papanikolaou, Seru and Stoffman, 2017) increases with process intensity (Lesson 6), and patent holders of process-intense patents have (until recently) been more likely to renew their patents (Lesson 7). Processes are also cited more often then products, although this relationship has weakened (Lesson 8). Last, we show that, until the early 1990s, independent process claims have more dependent claims (Lesson 9) and that, for the entire sample period, process claims are shorter than product claims (Lesson 10). We conclude the section with a word of caution, using claim length as a measure for patent breadth or patent scope (as in Kuhn and Thompson (2019) or Marco, Sarnoff and deGrazia (2019)) without accounting for claim type maybe yield misleading results. Such an approach would systematically over-estimate the breadth of process claims. Instead, researchers using claim length as a measure for breadth may need to normalize claim length by claim type so as to avoid comparing apples with oranges.

In Section 4, we provide a summary of our data-construction approach that combines information from both the preamble and the body of a type. We discuss how we identify different types of preambles and bodies and how in the end the two components are put together for the classification of the claim. We provide a detailed account of the underlying assumptions, the parameterization of the code, and the workflow in Appendix Section A. A description of the data files can be found in Appendix Section C.

In Section 5, we present results from a validation exercise. We use close to 10,000 manually classified claims (issued between 1976 and 2015) to assess the accuracy of our classification. Our classifier is correct (relative to the mainly classified benchmark) more than 98% of the time and outperforms a simpler approach that looks at only the keywords "method" or "process" in the claim preamble (similar to work by Angenendt (2018) or Bena and Simintzi

(2019)) by 3 percentage points. We also assess the accuracy of our classifier for different preamble-body combinations. Given this information, we show results for the *implied* accuracy and coverage (i.e., share of claims that are classified successfully) for the entire sample period. We conclude the section by providing guidance for researchers on how to adjust the classification outcome given the information provided in the published data files.

A number of recent projects show the versatility of how the information of the data files can be used in the economics and management literature. Branstetter, Chen, Glennon and Zolas (2021) empirically exploit a policy that affected the costs of production offshoring by Taiwanese firms to China for different technologies. They find that offshoring leads to a shift in innovation from products to processes, with overall innovation declining. Keum (2020) studies the link between firms' R&D and patenting outcomes and their decisions to substitute labor for capital in their production processes. He finds that process patents lead to a larger incrase in capital investments vs. non-process patents and, unlike product patents, process patents do not have a significant positive effect on employment growth. Ma (2021) studies the effect of technological obsolescence on firm growth and asset returns, finding that the effects of product innovation are more pronounced. Chen, Hsu and Wang (2022) examine the role of corporate governance for firms' innovation strategy, exploiting the introduction of staggered boards (with overlapping directors' terms) in Massachusetts in 1992. They show that the long-term orientation and managerial stability of staggered boards have a positive effect on product innovation. Babina, Fedyk, He and Hodson (2021) examine the rise of investment in AI technology by American firms and the resulting benefits from such investments. They show that increased investment in AI technology, as measured by job postings and hiring of AI-skilled employees, leads to increased product innovation (proxied by trademarks, product patents, and product updates). De Rassenfosse, Grazzi, Moschella and Pellegrino (2020) find that patent protection in a country supports increased firm exporting activity to that country, especially for product patents as the patent-industry matching is stronger for such patents (relative to process patents that potentially apply to multiple products or industries). Ganglmair and Reimers (2022) examine the patent propensity of processes and products and show that a strengthening of trade secrets protection lowers the probability of process patents being filed.

We are not the first ones analyzing the type of invention covered in a patent. There is earlier work on which we build as well as work that has been conducted in parallel. The first in line is Mike Scherer with his monumental task of manually classifying more than 15,000 U.S. utility patents by 443 large U.S. corporations issued between June 1976 and March 1977 (Scherer, 1982a,b, 1984). In a large-scale project examining historic patents, (Risch, 2012) has classified close to 2,500 U.S. handwritten patents issued between 1796 and 1839. In

another recent project, Toh and Ahuja (forthcoming) examine patents files by 101 chemical firms between 1982 and 1988. Their sample contains 44,440 patents with 534,520 claims.

Our methodology relies on an algorithmic reading of patent claims. We are aware of three other data-construction projects that performed this task on a similar scale. Crouch (2015) classifies claims as process claims if they include the word "method" or "process" within the claim. He presents the results for patents issued between 1986 and 2015, showing an increase of method patents (i.e., patents with at least one method claim) from 30% to 60% of issued patents per year. Bena and Simintzi (2019) use a similar approach. They classify process claims as starting with "A method for" or "A process for" (and minor variations) followed by a verb. They study the effect of a decrease of the cost of labor on firms' process innovation, using their claim classification for U.S. utility patents issued between 1976 and 2013. Last, in Angenendt (2018), claims that start with "A method" or "A process" are process claims, and those starting with "The use" or "The application" are use claims. The remaining claims are product claims. Studying the effect of changes in trade secrets protection on patenting strategy, he uses U.S. utility patents issued between 1976 and 2006.

Our approach goes a few steps further than the above. First, our data files cover all independent claims in U.S. utility patents issued after 1836.[2] Second, using information from both the preamble and the body of a given claim allows us to obtain better accuracy for our classifier. In Section 5, we show an overall accuracy of our classifier of 98.3%, whereas our implementations of the approaches by Angenendt (2018) and Bena and Simintzi (2019) achieve an accuracy of 95.6% and of the approach by Crouch (2015) an accuracy of 90.7%. Third, the classification in our data files is flexible in that we provide sufficient information to change (within limits) the claim classifications without running the computer code. This option will be of interest to researchers who either disagree with some of our assumptions or whose research questions require adjustments in the classification.

2 A Primer to Patent Claim Drafting

Patent claims define an invention. A patent application is required to have one or more claims that distinctly claim "the subject matter which the patent applicant ('applicant') regards as her invention or discovery."[3] Once an applicant files a patent application, a patent examiner conducts an examination of the application, including a thorough review of the patent claims. Sophisticated applicants expect that the examiner will reject some or

[2]For the analysis in this paper, we use patents issued between 1920 and 2020; the complete data files also contain claims from patents issued before 1920.

[3]Manual of Patent Examining Procedure (MPEP) §608.01(i)(a). The most up-to-date revision is available at https://www.uspto.gov/web/offices/pac/mpep/index.html.

all of the proposed claims for statutory reasons. The patent prosecution process allows for the applicant to present arguments to the examiner and/or amend the application's claims so that they will be allowed and the patent granted. The claims that appear in a granted patent are likely the result of argument, amendment, and negotiation that occurred during patent prosecution. Thus, with this process in mind, applicants should draft patent claims of varying scope.[4] This section provides a primer to patent claim drafting with an overview of patent claims (what are parts of a claim, different claim classes, and claim types), legal issues surrounding patent claims, and a summary of current practices for claim drafting given the most recent legal developments.

2.1 Parts of Claim

The patent statute does not require that patent claims be written in a specific format. However, the rules implementing the statute do provide a framework for claim formatting.[5] The rules require that, where practicable, an independent patent claim have three main parts. The preamble appears first and is a general description of the invention.[6] Second, a transitional phrase such as "comprising" or "consisting" generally follows the preamble.[7] The third part of the claim identifies elements, steps or relationships which the applicant is claiming as the invention. Best practices dictate that the applicant should separate each element or step by a line indentation.[8]

A patent may include claims that vary in scope. A patent claim is either an independent claim or a dependent claim. An independent claim does not refer back to or depend from another claim. Independent claims are less restrictive and broader in scope than dependent claims. A patent applicant should include the least restrictive independent claim as the first claim of a patent.[9] A dependent claim refers to one or more other claims (independent or dependent) and limits the subject matter in the preceding claim in various ways. A claim that depends on more than one claim is referred to as a multiple dependent claim.

[4]MPEP §608.01(m)("Many of the difficulties encountered in the prosecution of patent applications after final rejection may be alleviated if each applicant includes, at the time of filing or no later than the first reply, claims varying from the broadest to which he or she believes he or she is entitled to the most detailed that he or she is willing to accept.")

[5]37 CFR §1.75

[6]See Catalina Mktg. Int'l v. Coolsavings.com, Inc., 289 F.3d 801 (Fed. Cir. 2002) (exploring when language in the preamble will limit the claim).

[7]See MPEP §2111.03 (The word "comprising" is "synonymous with 'including,' 'containing,' or 'characterized by,' is inclusive or open-ended and does not exclude additional, unrecited elements or method steps." In contrast, consisting of "excludes any element, step, or ingredient not specified in the claim.")

[8]We provide two examples of patent claims in Table A.2 in the Appendix.

[9]37 CFR §1.759(g)

2.2 Claim Classes

Generally, claims can be divided into two main classes – product and process claims.[10] It is improper for a single patent claim to be directed to both a product and a process.[11] This section explains the types of inventions that typically fall into the classes of a product claim or a process claim.

Product Claim. Product claims include claims directed to machines, articles of manufacture, and compositions of matter. A machine is a mechanical apparatus. A manufacture is an article created from raw materials. Finally, a composition of matter is formed from the combination of two or more substances. Since machines, articles of manufacture, and compositions of matter generally embody a physical product, they are categorized as product claims.

Process Claim. Unlike product claims, process claims define a method or procedure for performing a task. For example, a process claim can explain how to make a particular item or how to perform a service.

2.3 Claim Types

Most patent claims are directed to either a product or process. Depending on the nature of the invention, patentees may claim their invention using many different types of claims. This section briefly summarizes the various types of patent claims.

Product-by-Process Claims. A product-by-process claim defines a product invention by setting forth the steps needed to create or obtain the product. The scope of a product-by-process claim is defined by the process steps.[12] Patentees may choose to use a product-by-process claim when it is difficult to describe the product in any way other than by the process in which it was created. An example of subject matter that may be claimed as a product-by-process is a chemical compound that is obtained by performing a series of steps.[13]

[10] 35 U.S.C. §101 states that a patent can be obtained for a process, machine, manufacture, or composition of matter. Machines, manufactures, and compositions of matter can be categorized as products.

[11] Ex parte Lyell, 1990 Pat. App. LEXIS 14, *12 (Bd. Pat. App. & Interferences August 16, 1990).

[12] Atl. Thermoplastics Co. v. Faytex Corp., 970 F.2d 834, 846-47 (Fed. Cir. 1992) ("In light of Supreme Court case law and the history of product-by-process claims, this court acknowledges that infringement analysis proceeds with reference to the patent claims. Thus, process terms in product-by-process claims serve as limitations in determining infringement.")

[13] See Abbott Labs. v. Sandoz, Inc., 566 F.3d 1282, 1285 (Fed. Cir. 2009) (the patent claims at issue were directed to a process for obtaining a crystalline compound).

Means-plus-Function Claims. Means-plus-function claims express a claim element as "a means or step for performing a specified function without the recital of structure, material, or acts in support thereof."[14] A patentee may use a means-plus-function claim to broadly describe a thing that performs a certain function. For example, "a means for fastening two beams together" could refer to any number of items including adhesive, nails, screws, etc.

Historically, commentators criticized means-plus-function claims for being too broad in scope. These critics argued that patent claims could be interpreted to cover products beyond what the patentee invented. However, in order for means-plus-function claims to satisfy the written description and enablement requirement the "means for" mentioned in a claim must be described in detail in the patent's specification.[15] Thus, the scope of a means-plus-function claim is limited by the patent specification.

Markush Claims. A Markush claim is a claim type used to limit the claimed subject matter to a specific list of alternatives.[16] Markush claims occur often in chemistry patents but can also be used to claim other subject matter.[17] Note that instead of claims that use the "comprising" transition phrase which denotes an open group, Markush claims are directed toward a closed system. Thus, Markush group claims commonly use the transition word "consisting" instead of "comprising."

Jepson/Improvement Claims A patentee may be awarded a patent for an improvement on an existing product.[18] A patentee can use the Jepson claim type to describe the improvement being claimed in the patent. The preamble of a Jepson claim describes what is known or in the prior art. The Jepson claim includes a transitional phrase such as "wherein the improvement comprises" to separate what is conventional or known from the subject matter that the applicant considers is new.[19] After the transitional phrase, the claim lists everything that is an "improvement" or new in the body of the claim. Jepson claims may be used to describe both product and process inventions. From the patent examiner's perspective, Jepson claims are beneficial because they identify the elements of the claim that the patentee believes are novel.

[14] 35 U.S.C. §112 (f).

[15] See 35 U.S.C. §112 (f) ("such claim shall be construed to cover the corresponding structure, material, or acts described in the specification and equivalents thereof").

[16] Gillette Co. v. Energizer Holdings, Inc., 405 F.3d 1367, 1372 (Fed. Cir. 2005) ("Claim drafters often use the term "group of" to signal a Markush group. A Markush group lists specified alternatives in a patent claim, typically in the form: a member selected from the group consisting of A, B, and C").

[17] 3 Chisum on Patents §8.06 (2018).

[18] 35 U.S.C. §101.

[19] 37 C.F.R. §1.75(e).

Software and Beauregard Claims. Patentees have used creative claiming techniques in response to the proliferation of computer-related technology. For example, patentees have claimed patented software as both a product and as a process. As a process, software is a series of steps performed by a computer processor. As a product, patentees have claimed the physical device (e.g., computer registers or switches) that stores and executes the software.

Beauregard claims are directed toward the "product" of a data storage device ("a computer readable storage medium") having software code to cause a computer processor to perform certain method steps.[20] Despite Beauregard claims being directed to a computer readable storage medium, the Federal Circuit treats Beauregard claims as process claims when considering whether a particular claim is eligible for patenting.[21]

Chemical & Pharmaceutical Claims. Claims to chemical compounds may also have a unique form. Chemical claims may be directed to a chemical compound or the method of making a chemical compound. The chemical compound claim typically lists a number of elements, materials, and other chemicals that are used to make the claimed compound. Markush groups are often used in chemical compound claims.

Pharmaceutical claims directed to a composition use a similar approach. Pharmaceutical claims may also be directed to a treatment method for using a pharmaceutical. Treatment method claims follow a similar format to that of conventional process claims but may provide for dosages, times, and conditions of treatment (Rosenberg, 2017:§2.10).

Business Method Claims. Another claim type that covers a specific type of process is known as a business method claim. Business method claims are directed toward steps for carrying out a function not directly tied to a particular machine. Business method claims became more prevalent with the creation of the Internet and the explosion of e-commerce. Business method claims are controversial because they are vulnerable to invalidity challenges related to patent eligibility. For example, in *Bilski v. Kappos*, the Supreme Court held that a method for hedging risk was an abstract idea and therefore not eligible for patenting.[22] Despite their vulnerability, business method claims that satisfy the current two-step *Mayo* test for abstract ideas are eligible for patenting.[23]

[20] CyberSource Corp. v. Retail Decisions, Inc., 654 F.3d 1366, 1373 (Fed. Cir. 2011).

[21] CyberSource Corp. v. Retail Decisions, Inc., 654 F.3d 1366, 1375 (Fed. Cir. 2011).

[22] Bilski v. Kappos, 561 U.S. 593, 611 (2010)("The concept of hedging, described in claim 1 and reduced to a mathematical formula in claim 4, is an unpatentable abstract idea...").

[23] See Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66 (2012).

2.4 Claim Issues

Patentees' use of various claim types has given rise to numerous legal disputes. Within the last forty years, U.S. Courts have decided cases that have had a significant impact on how patentees draft claims and how claims are interpreted by the courts. Legal issues that have had a significant impact on claim drafting and interpretation include patent eligibility, written description, enablement, and the doctrine of equivalents. This section briefly summarizes some of the more significant legal issues that have had an impact on how product and process claims are drafted and interpreted.

Patent Eligibility. The question of what subject matter should be eligible for patenting has had a major impact on how patentees draft product and process claims. As set forth in the patent statute, any new and useful, machine, manufacture, composition of matter or process is eligible for patenting.[24] In contrast, things occurring in nature, natural phenomena, and abstract ideas are not patent eligible. The legal interpretation of these principals has evolved over decades and has affected how both product and process claims are drafted.

Things found in nature are not eligible for patenting. However, naturally occurring organisms that have been modified by man are considered manufactures or compositions of matter and are therefore patentable. For example, bacteria that have been genetically modified to possess characteristics not normally found in their natural state are patent eligible.[25] In addition, man-made organisms are eligible for patenting. For instance, the Supreme Court has held in *Myriad* that synthetically created DNA is patent eligible.[26]

Claims directed to software, business methods and computer-implemented inventions have been significantly affected by legal developments in recent years. Patent claims directed to this subject matter are commonly challenged on the grounds that they are directed to an abstract idea and therefore not patentable. The Supreme Court has held in *Mayo* that patent claims that incorporate abstract ideas are eligible for patenting if the claim amounts to significantly more than the abstract idea itself.[27] However, claims directed to an abstract idea implemented on a general purpose computer are not patent eligible.[28] The legal framework for determining whether a claim containing an abstract idea "amounts to significantly more" continues to evolve.

[24] 35 U.S.C. §101.
[25] See Diamond v. Chakrabarty, 447 U.S. 303 (1980)
[26] See Ass'n for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576 (2013).
[27] Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66 (2012).
[28] Bilski v. Kappos, 561 U.S. 593 (2010).

Written Description and Enablement. The patent statute requires that a patentee's application completely disclose her invention and enable one of ordinary skill in the art to make or use the invention without undue experimentation. Both the written description and enablement requirements are primarily directed at the patent specification, not the claims. Thus, a detailed discussion of these requirements is beyond the scope of the paper.

However, both the written description and enablement requirements suggest that patent claims be drafted in a clear and precise manner. The complexity of claim language is caused by several factors including the nature of the subject matter, claim drafting strategies and the requirement that each claim is a single sentence.[29] Nevertheless, the terms used in the claims should be supported by the patent specification and, with reasonable certainty, inform those with skill in the art about the scope of the invention.[30]

In addition, to satisfy the enablement requirement, a patent application must disclose the invention such that a person of ordinary skill in the art could make or use the invention without undue experimentation. In order to satisfy the enablement requirement, the patent claims must recite how the claimed elements or steps work together. Further, the claims must be directed to an operable invention (Rosenberg, 2017:§8.02).

2.5 Claim Drafting

Given the legal issues discussed above, patent stakeholders have developed best practices for drafting claims. Unlike claim drafting fundamentals, best practices evolve with technology and the law. For example, new technologies such as computer software ushered in new practices for claiming software inventions. This section briefly describes some current claim drafting considerations and resources.

USPTO Guidance. Periodically, the USPTO issues guidance to the public. A discussion of all the USPTO resources available is beyond the scope of this paper. The important point is to understand that claim drafting and interpretation is a rapidly changing activity.

For example, the USPTO maintains a website that provides guidance related to issues of subject matter eligibility.[31] As of this writing, the subject matter eligibility website was updated on June 7, 2018, to include a memorandum on a recent Federal Circuit opinion. The site also includes a "quick reference sheet" that summarizes Federal Circuit decisions that have found patent claims eligible for patenting. For practitioners that draft and prosecute

[29] Fressola v. Manbeck, 36 USPQ2d 1211 (D.D.C. 1995).

[30] Nautilus, Inc. v. Biosig Instruments, Inc., 134 S. Ct. 2120, 2124 (2014).

[31] https://www.uspto.gov/patent/laws-and-regulations/examination-policy/subject-matter-eligibility

patents in the software, business method, and medical treatment areas, the USPTO website is an invaluable resource. Patent prosecutors use this resource to counsel their clients on issues related to claim drafting.

Practitioner Considerations. Over time, experienced patent prosecutors develop best practices for drafting claims. These practices are shared internally within companies and law firms. The practices can range from general drafting advice to specific advice for how to claim a particular technology.

Patent practitioners must consider statutory and strategic issues when drafting claims. Most practitioners attempt to draft broad claims that still define the invention. However, a practitioner's strategic perspective on patenting may be different from the lay view of patenting. For example, a practitioner's purpose in drafting claims may be to prevent competitors from entering the marketplace occupied by the claimed invention.

In deciding between product and process claims, practitioners must also consider how difficult it will be to determine infringement. In other words, how much investigation will it take to identify an infringing product or method? A practitioner must also consider who will potentially infringe the claim and where that infringement may take place. For example, claims that require more than one actor to perform the claimed steps are difficult to enforce (Robinson, 2012). Further, for U.S. patents, a system claim with a component located outside the U.S. may still be enforceable,[32] while the infringement of a process claim must take place entirely within the U.S.[33]

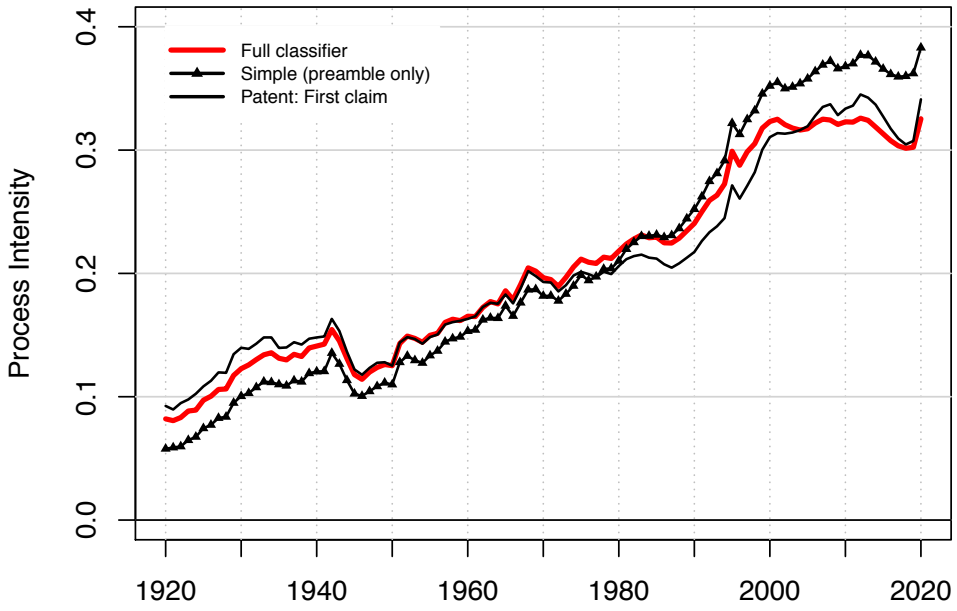
3 The Rise of Process Claims

In this section, we document the development of the process-intensity of U.S.-issued patents using the outcome from our ***pat****ent* ***c****laim* ***c****lassification by* ***a****lgorithmic* ***t****ext-analysis* (patccat). We describe the data construction in detail in Section 4. We show results for the universe of U.S. utility patents issued between 1920 and 2020, for the aggregate sample as well as broken down by technology categories and different applicant types. We then ask whether claim and patent breadth and patent value vary with the process-intensity of a patent and provide descriptive evidence.

[32]NTP, Inc. v. Research In Motion, Ltd., 418 F.3d 1282, 1317 (Fed. Cir. 2005) (holding that the location of a relay in Canada did not preclude infringement of a system claim).

[33]Id. at 1318 (holding that "a process cannot be used 'within' the United States as required by section 271(a) unless each of the steps is performed within this country.")

Figure 1: The Rise of Process Claims



Notes: We plot the share of process claims and process patents (process intensity) by application year for 1920 to 2020. We show the share of process claims scaled by all independent claims from full patccat classification (red line); the share of patents with process claim (from patccat) as their first claim (black line); and the share of process claim from simple classification approach (process claim if the words "method" or "process" are used anywhere in the preamble of the claim). *Data source: patccat.*

3.1 Process Claims over Time and Across Technologies

In Figure 1, we depict the process-intensity of U.S. patents. The red line captures the results from the full patccat classifier. It represents the annual share of process claims scaled by the total number of independent claims that year; equivalent to the average number of process claims per patent filed in a given year. The solid black line depicts the share of patents with a process claim as their first independent claim, which is presumably the broadest and most important claim in a patent (Kuhn and Thompson, 2019). Last, we plot the results from a simple classification exercise in which we classify a process claim if it uses the term "method" or "process" anywhere in the preamble (similar to the approaches in Angenendt (2018) and Bena and Simintzi (2019)). For all three time series, we show annual averages by the patents' application year and refer to these averages as *process intensity.*

Lesson 1. *The process-intensity of U.S. patents has increased by 25 percentage points, from an average of just below 10% in 1920 to more than 30% in 2020.*

Process intensity of U.S. patents has steadily increased throughout our sample period. We observe a few short periods of decline, most notably during World War II and in the mid

2010s.[34] After a fairly constant growth rate of process intensity between 1945 and 1985, we observe faster growth until early 2000. The increase of process intensity has slowed down since then. During the last few years in our sample period, we can observe a slight reversal, with the share of process claims and process patents lightly decreasing.

The observed patterns are similar for all three measures of process intensity. The simple approach underestimates process intensity (relative to the full classifier) until early 1980s and overestimates the share of process claims in patents since then (with a gap of approximately 6 percentage points in 2020). As we will see in Figure 2, a part of this gap is explained by our product-by-process claim type.[35] We also observe different levels of process intensity comparing the overall share of process claims with the share of patents holding a process claim as their first claim. Between the early 1970s and mid 2000s, the share of process claims is higher than the share of first-claim process patents.

Our results for the process intensity of patents compare well with empirical patterns of the overall R&D mix of firms. For instance, Scherer (1967, 1984) finds that about 25% of R&D efforts are process-related. Hall et al. (2013b) find (for Italian manufacturing, 1996–2005) that 24.0% of firms engage only in process innovaton, and 26.9% of firms in both process and product innovation. Using data from the German Innovation Survey, Rammer et al. (2016:59ff) that in 2014, firms devoted 27% of their total innovation budget to process innovation. For the same year, Kindlon and Jankowski (2017) show that 31.7% of U.S. businesses are process innovators and 42.3% are active in both product and process innovation.

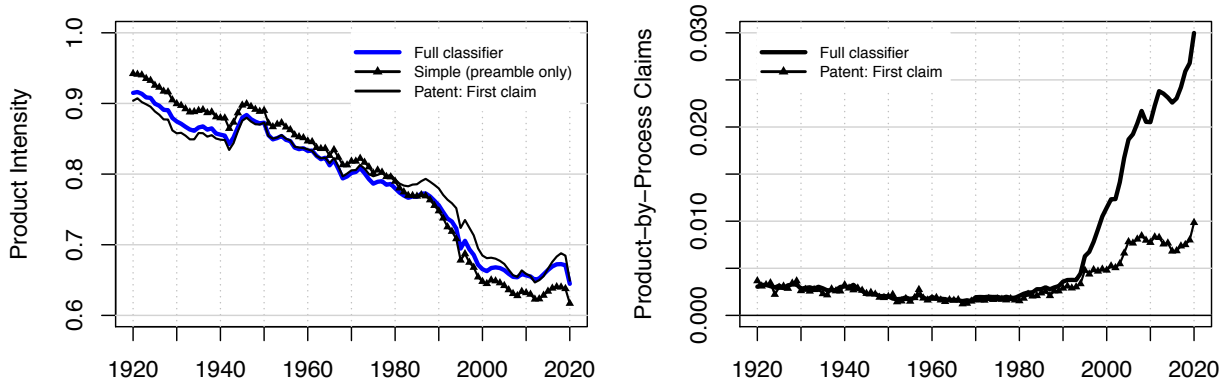
While the close relationship of our results with survey evidence is reassuring, we should not necessarily expect an increase in innovation and R&D to translate to one-to-one to an increase patents. Not all patentable inventions are patented (Mansfield, 1986) and the legal institutions governing the protection of intellectual property (such as patent law or laws protecting trade secrets) affect firms' incentives to seek protection in patents (Png, 2017a,b) and, particularly, process patents (Graham and Hegde, 2015; Ganglmair and Reimers, 2022). Also, not all results from firms' R&D efforts are patentable.

In Figure 2, we show the product intensity of U.S. patents (right panel) and the use of product-by-process claims (left panel) for our sample period. For a large part of of our sample period, product-intensity is the mirror image of process intensity as product-by-

[34]The uptick at the end of our sample period is likely due to truncation. Our sample contains all patents issued until the end of 2020. The number of issued patents that were applied for in 2020 (as depicted in the figure) is relatively low given examination delays.

[35]Claims that use the word "process" in the preamble in what we label a by-process phrase – indicating that something is performed "by a process" – are process claims following the simple approach but product-by-process claims according to our full classifier. See Section 4.

Figure 2: Product and Product-by-Process Claims



Notes: We plot the share of product claims (product-intensity) on the left and the share of product-by-process claims on the right by application year for 1920 to 2020. For product claims, we show the results from the full patccat classifier (blue line); the share of patents with product claims (from patccat) as their first claim (black line); and the share of product claims (1 - share of process claims) from the simple classification approach. For product-by-process claims, we show results for the full patccat classifier and the share of patents with a product-by-process claim as their first claim. *Data source: patccat.*

process claims are a marginal claim category – about 0.25% of all claims are of that type – until the mid 1990s. We observe a slow increase starting in 1980 and an acceleration in the relative numbers of product-by-process claims in 1995. The overall share reaches 3% in 2020.[36]

The rise in process claims and increase in process intensity is likely not uniform across technologies. In some technologies, process innovation is more prevalent or applicants are more likely to patent their processes than in other technology areas. We examine anticipated differences using the NBER technology categories, a coarse classification of patents into 6 categories and 37 sub-categories (Hall, Jaffe and Trajtenberg, 2001).[37,38]

[36]We will see the source of this strong increase in product-by-process claims in Figure A.3. A considerable number of these claims are *Beauregard claims*.

[37]Patents are assigned to one of six categories (with their subcategories in brackets): Chemical (1) [Agriculture, Food, Textiles - 11; Coating - 12; Gas - 13; Organic Compounds - 14; Resins - 15; Miscellaneous - 19], Computers & Communications (2) [Communications - 21; Computer Hardware & Software - 22; Computer Peripherals - 23; Information Storage - 24; Electronic business methods and software - 25], Drugs & Medical (3) [Drugs - 31; Surgery and Medical Instruments - 32; Genetics - 33; Miscellaneous - 39], Electrical & Electronic (4) [Electrical Devices - 41; Electrical Lightning - 42; Measuring & Testing - 43; Nuclear & X-rays - 44; Power Systems - 45; Semiconductor Devices - 46; Miscellaneous - 49], Mechanical (5) [Material Processing & Handling - 51; Metal Working - 52; Motors & Engines + Parts - 53; Optics - 54; Transportation - 55; Miscellaneous - 59], and Others (6) [Agriculture, Husbandry, Food - 61; Amusement Devices - 62; Apparel & Textile - 63; Earth Working & Wells - 64; Furniture, House Fixtures - 65; Heating - 66; Pipes & Joints - 67; Receptacles - 68; Miscellaneous - 69]. Appendix 1 in Hall, Jaffe and Trajtenberg (2001) lists the respective USPC main classes (version 1999) for each sub-category.

[38]This classification is based on the United States Patent Classification (USPC) System. Because it was

Table 1: Process Intensity by Technology

NBER Category	1920–39	1940–59	1960–79	1980–1999	2000–14
Chemical	0.393	0.429	0.380	0.420	0.418
Computers + Communications	0.068	0.033	0.084	0.308	0.368
Drugs + Medical	0.165	0.196	0.296	0.347	0.379
Electrical + Electronic	0.053	0.052	0.101	0.236	0.296
Mechanical	0.077	0.080	0.136	0.199	0.231
Others	0.065	0.077	0.115	0.158	0.196

Notes: We report the average share of process claims for the 20-year time windows 1920–39, 1940–59, 1960–79, 1980–1999, and 2000-14 (by application year) for the NBER technology categories. *Data source: patccat,* https://patentsview.org *and Google Patents Public Datasets (for the NBER categories).*

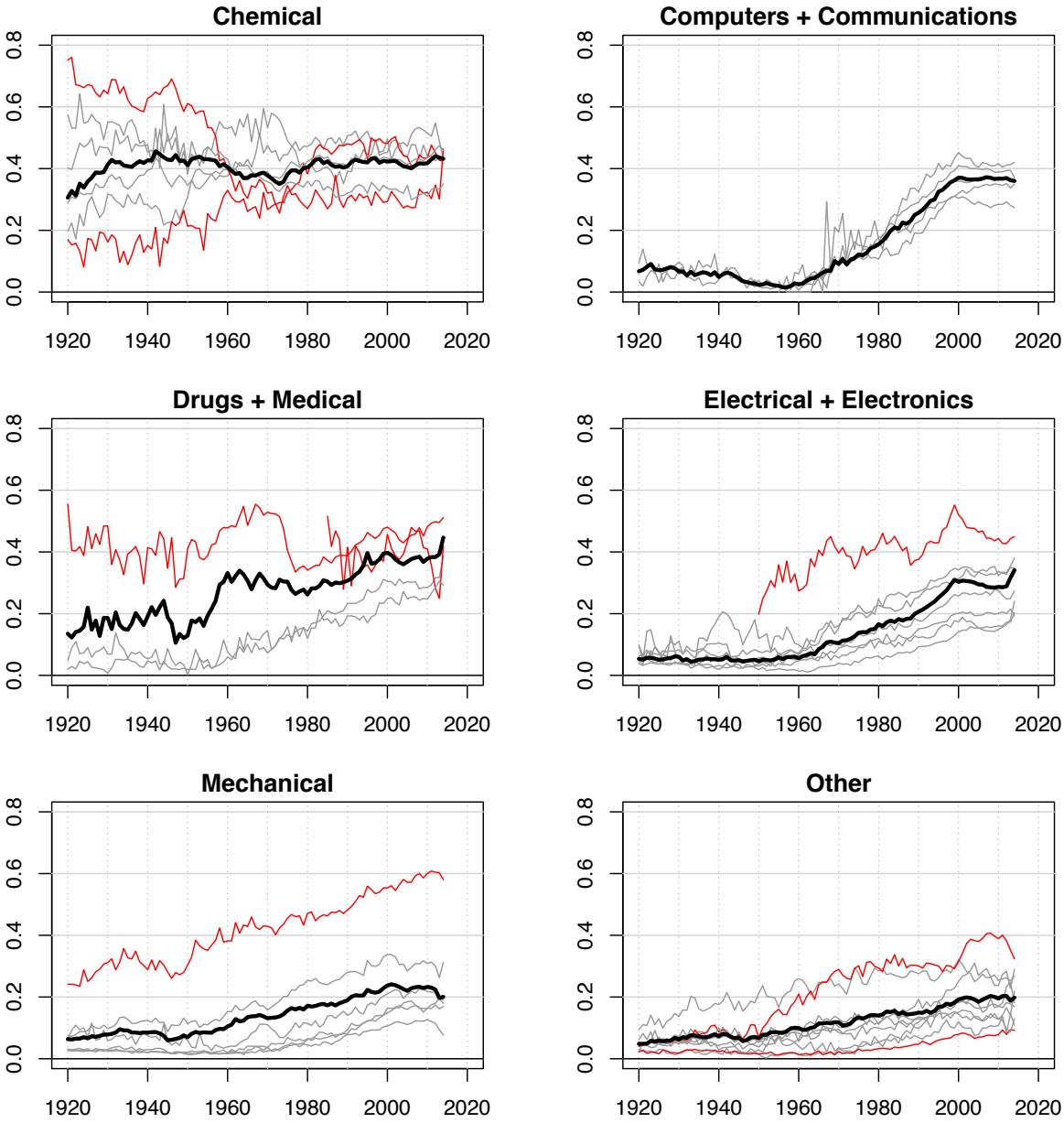
Lesson 2. *Process intensity is highest in chemical and drugs & medical patents and lowest in mechanical and other patents. In chemical patents, it has been fairly constant since the 1940s, whereas in computers & communications, electrical & electronics, and mechanical patents it was constant until the mid 1960s and has since then seen a steady increase.*

In Figure 3, we show the share of process claims (scaled by all patent claims filed in a given year) by application year (1920 to 2014) for the six NBER technology categories. We see a considerable amount of heterogeneity of the level and the increase of process intensity across technologies. Patents in the chemical category (with the highest level of process intensity), and to some extent, in drugs and medical exhibit constant process intensities, whereas patents in all other categories experienced increases in process intensity to varying degrees. We observe the strongest increase in computers & communications patents, followed by electrical & electronics. Process intensity in both mechanical and other patents has increased, but at a much lower pace. We summarize the average process intensities for 20-year time windows in Table 1.

Technologies are not homogeneous with respect to their process intensity. We observe a considerable amount of variation in Figure 3 where we depict the process intensity for sub-categories by thin lines. For **Chemical** patents, we see high variation in process intensities until around 1980, with convergence toward the overall average. Patents in Organic Compounds (sub-category 14, in red) were initially highly process intense but exhibited a decline until the 1960s. Patents in Gas (sub-category 13, in red) have the lowest process intensity among Chemical patents, with increases in the 1940s and 1950s. Patents in **Computers** have experienced a homogeneous development across all sub-categories. The sub-category

discontinued in May 2015, we use patents issued through May 2015 and filed trough the end of 2014 for our analysis.

Figure 3: Process Intensity by Technology



Notes: We plot the share of process claims (process intensity) by application year for 1920 to 2014 for six different NBER technology categories (black lines) and 37 sub-categories (thin grey and red lines) (Hall, Jaffe and Trajtenberg, 2001). We depict sub-categories discussed in the main text using red lines. We trim the time-series for sub-categories 23, 25, 33, and 46 because of low numbers of observations. *Data source: patccat,* `https://patentsview.org` *and Google Patents Public Datasets (for the NBER technology categories).*

with the lowest process intensity is Computer Peripherals (sub-category 23) at 30% in 2015, roughly 10 percentage points lower than the overall average.

Category **Drugs & Medical** comprises four sub-categories. Surgery and Medical Instruments (32) and Miscellaneous (39) have low process intensities, whereas Drugs (31) and Genetics (33) exhibit process intensities consistently above the average levels. Patents in **Electrical & Electronics** are rather homogeneous, showing similar levels and trends across sub-categories. This is for all but Semi-Conductor Devices (sub-category 46, in red) that exhibits noticeably higher levels of process intensity. We observe similar patterns for **Mechanical** patents, where process intensity is among the lowest. Patents in Metal Working (sub-category 52, in red) are the exception. Their levels of process intensity are up to three times as high as the average levels, and among the highest of all NBER technology sub-categories.

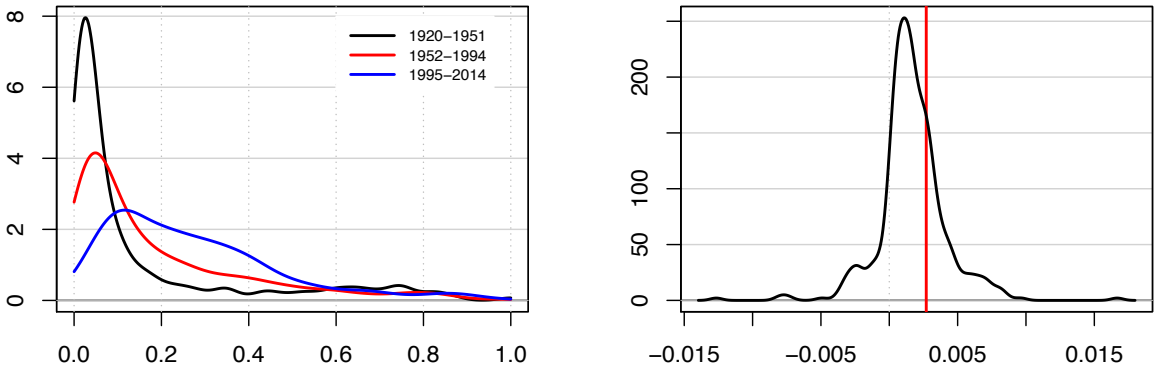
In Figure 4 we provide a more granular picture of the trends across different technologies. We plot the distributions of levels and slopes of process intensities for 430 UPSC main classes. In the left-hand side panel, we show the distribution of average levels of process intensities across USPC main classes for three time periods (by application year): 1920–1951 (black), 1952–1994 (red), and 1995–2014 (blue).[39] In the right-hand side panel, we show the distribution of the slope of linear trend lines from simple OLS regressions of the annual share of process claims for 1920 to 2014 in each of the USPC main classes.

The depicted distributions (kernel density plots) in the left-hand side panel show both a general upward trend of levels of process intensity and widening of the spread across technologies. In other words, for a uniform upward trend across all patent classes, we would expect a horizontal shift of the kernel density plots. But while the peak of the density shifts to the right, we also observe a change in the shape of the distributions. These patterns are a first piece of the puzzle of the main contributors of the overall increase in process intensity.

The unweighted mean slope of the estimated trend lines in the right-hand side panel is 0.00116 (sd=0.0024). The slope coefficient for the full sample is 0.0027, translating to a predicted increase of process claims by 25.7 percentage points. The picture depicts an increase of the process intensity for a large majority of the USPC classes: 373 out of the 430 main classes (86.7%) exhibit a positive slope coefficient. The rise in process claims is therefore not limited to just a few USPC classes and technologies; we observe it across a wider technological spectrum.

[39] We choose 1952 and 1995 as our cutoffs because of the major legislative changes in those years. The year 1952 marks the birth of modern patent law with the Patent Act of 1952. In June 1995, two major provisions of the *Uruguay Round Agreements Act* of 1995 went into effect, extending the maximum validity of a patent to 20 years from filing and introducing provisional patent applications.

Figure 4: Process Claims: Distribution of Levels and Slopes



Notes: In the left-hand side panel, we show the distribution (kernel density plots) of average levels of process intensity across 430 USPC main classes for three time periods: 1920–1951 (black curve), 1952–1994 (red curve), and 1995–2014 (blue curve) by application year. In the right-hand side panel, we show the distribution of the linear slope parameters from OLS regressions for each of the 430 USPC main classes for 1920–2014 by application year. The vertical red line marks the slope for the full sample. *Data source: patccat, https://patentsview.org and Google Patents Public Datasets (for the USPC main classes).*

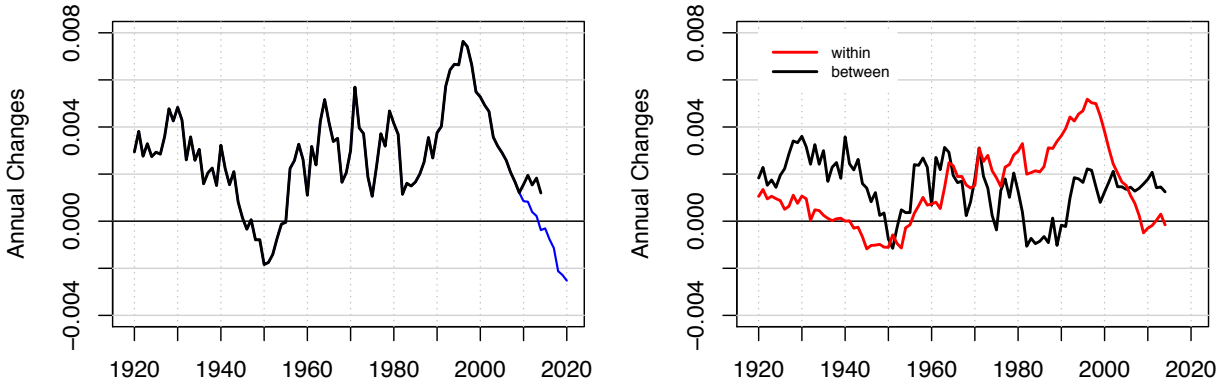
In Figure 4 we see a rise in process claims across larger number of technologies (patent classes), at varying degrees. We also see that the process intensity of some technologies has decreased. Is the rise in process claims driven by a small number of patent classes (with high levels or increases) whose relative numbers have increased, or do we observe the increase in process intensity across many technologies and patent classes, as a systemic change? To shed additional light on this question, we look at how process intensity has changed year-by-year.

First, note that the average share of process claims in Figure 1 is a weighted average of the share of process claims in each of the 430 USPC main classes in our sample, where the weights are the per-year number of claims in a given USPC main class in a given year scaled by the total number of claims in that year. Similarly, the annual changes of the share of process claims (we show 10-year rolling averages) depicted in the left-hand side panel of Figure 5 are the weighted average changes in each of the 430 USPC main classes.[40]

Are changes in the share of process claims due to changes in the share of process claims *within* each USPC class or due to changes in the composition of USPC classes? Asked differently, do we see an increase because the share of process claims increases for USPC classes on average, or because USPC classes with higher process intensity grow (faster) and those with low process intensity shrink? To answer these questions, we decompose the change

[40]For this analysis, we use patents granted up to and including 2014. The information for USPC main classes is incomplete for 2015, and no longer available for more recent years. The black line in the figure depicts the rolling means up to 2014. The blue line is the rolling mean for the complete sample with all patents issued until 2020.

Figure 5: Decomposition of Annual Changes



Notes: In the left panel, we show the annual changes of the share of process claims by grant year for 1920 to 2014 (black line), and extended to 2020 (blue line). In the right panel, we show the annual change in the share of process claims due to variation in process intensity (Δ within, red line) and variation in USPC class composition (Δ between, black line) for 1920 to 2014. The cross term (Δ cross term) is negligible in size and omitted from the figure. *Data source: patccat, https://patentsview.org and Google Patents Public Datasets (for the USPC main classes).*

in the weighted average process intensity, denoted by $\Delta\mu_t$ (as a change from a time $t-1$ to t), into the component that is due to the change in process claims in a given class c, and the component that is due to the reallocation of patents towards higher-intensity patent classes, captured by a change in the USPC densities $\Delta\gamma_{ct}$.[41] We can write the change in process intensity as follows:

$$\Delta\mu_t = \underbrace{\sum_c \gamma_{c,t-1}\Delta\mu_{ct}}_{\Delta \text{ within}} + \underbrace{\sum_c \mu_{c,t-1}\Delta\gamma_{ct}}_{\Delta \text{ between}} + \underbrace{\sum_c \Delta\mu_{ct}\Delta\gamma_{ct}}_{\Delta \text{ cross term}}. \tag{1}$$

It is equal to the change in process claims for each class c weighted by the relative size of the respective class in the previous period, $\gamma_{c,t-1}$ (changes *within* classes) plus the change of USPC composition, $\Delta\gamma_{ct}$ holding the share of process claims constant at the previous period's levels (changes *between* classes).

In the right-hand side panel of Figure 5, we plot 10-year rolling averages for the within and between effects. The respective graphs depict the change in overall process intensity attributed to changes within USPC classes and between USPC classes. The cross term is negligible and omitted from the picture.

The figure paints a mixed picture. In the earlier years of our sample (1930s and 1940s), the change in USPC class composition was the main driver of changes in overall process

[41] We follow the approach in De Loecker, Eeckhout and Unger (2020).

Table 2: Decomposition by Time Period

Time period	Change ($\Delta\mu_t$)	Δ within	Δ between	Δ cross term
1920–29	2.743	1.053	1.582	0.108
1930–39	3.061	0.502	2.830	-0.271
1940–49	1.234	-0.902	2.293	-0.157
1950–59	0.673	-0.122	1.072	-0.277
1960–69	5.364	2.405	3.124	-0.165
1970–79	0.980	1.661	-0.120	-0.561
1980–89	1.899	2.209	-0.731	0.415
1990–99	6.756	4.565	1.869	0.297
2000–09	2.864	1.601	1.225	0.048
2010–14	0.734	0.142	0.545	-0.031
1920–2014	26.308	13.114	13.689	-0.594

Notes: We report the annual changes (in percentage points) of the share of process claims by decade for 1920 to 2014. We further report the contribution of the variation in process intensity (Δ within), the variation in USPC class composition (Δ between), and the composite effect (Δ cross term). *Data source: patccat, https://patentsview.org and Google Patents Public Datasets (for the USPC main classes).*

intensity. During the 1960s and 1970s, both effects were equally relevant. Increases in process intensity across the broad spectrum of technologies was the main driver in overall changes in later years of our sample period, particularly during the 1980s and 1990s.

In Table 2, we aggregate the contribution of each of the effects by decade, and the numbers confirm the preliminary insights from Figure 5. First, in aggregate, both *within* and *between* have equally contributed to changes in process intensity. The overall increase of 26.5 percentage points stems from a 13.1 percentage point increase within and a 13.7 percentage point increase between USPC classes. The relative role of the effects, however, varies over time. During the 1930s, 1940s, and 1950s, the between effect dominates. In these years, overall process intensity changes because of a redistribution of patents to process-intensive patent classes. During the 1970s, 1980s, and 1990s, the within effect dominates. The rise of process claims in these years is a result of a general shift of patent claims toward process claims across all patent classes.

Lesson 3. *In the 1930s, 1940s, and 1950s, changes of process shares across the broad spectrum of technologies were more important a driver of the rise of process intensity than changes in the composition of technologies with a shift of patenting toward more process-intense patent classes. In the 1970s, 1980s, and 1990s, these technological changes were the main driver. Over the last century, the two effects played on average equally important roles.*

3.2 Patent Assignees and Their Locations

In a next step we ask who are the applicants that file process-intense patents. In Figure 6, we plot the share of process claims (left-hand side) and the share of process patents by their first independent claim (right-hand side) for different applicant types and the applicants location (U.S. or foreign) for the years 1975 through 2020.

Lesson 4. *Patents granted to companies and government entities are more process-intense than those granted to individuals.*

Panel (a) in Figure 6 shows that patents filed by individuals are the least process-intensive, followed by those filed by companies and corporations and then, following closely, government entities. We can observe these patterns for both the overall share of process claims and for first-claim process patents. In fact, for both measures of process intensity we find that companies file patents twice as process intense as patents granted to individuals.

A possible explanations for these patterns is offered by Abernathy and Utterback (1978). They argue that as firms mature, their innovative activity shifts from products to processes. Klepper (1996) offers a similar explanation, arguing a firm's incentives to invest in cost-reducing processes are increasing in the volumne of production (i.e., its size). Many of the individual applicants are likely firms in their infancy, and the low process-intensity of their patents (relative to – larger – companies) comports with the life-cycle hypothesis. Scherer (1991), Cohen and Klepper (1996a,b) or Huergo and Jaumandreu (2004) offer (partial) empirical support for the hypothesis, whereas McGahan and Silverman (2001) find no evidence of a shift of innovation from products to processes as industries mature. Cucculelli and Peruzzi (2020) conclude that the empirical evidence of the life-cycle hypothesis is far from settled and identifies measurement difficulties (for both maturity and innovative activity) as a possible obstacle.

Another potential explanation is a size advantage for companies. They are better equipped to benefit from scale economies and thus are more likely to focus on the development and commercialization side of R&D that comes with more process innovation. Large players in the pharmaceutical and chemical industries are often said to acquire their product innovation through acquisitions of smaller firms while developing cost-saving processes in-house. Related results are offered by Link (1982) or Lunn (1986) who find that increased industry concentration (with fewer but larger firms) is correlated with a higher proportion of R&D expenses dedicated to process innovation and a higher concentration of process claims in patents, respectively.

Differences in the patenting behavior of individuals and companies might also be responsible for the patterns in Panel (a) of Figure 6. Ample survey evidence suggests that

the propensity to patent is higher for products than processes (Levin, Klevorick, Nelson and Winter, 1987; Cohen, Nelson and Walsh, 2000; Arundel, 2001; Hall, Helmers, Rogers and Sena, 2013a). A potential source of this discrepancy are the costs associated with the monitoring of infringement that are said to be a major contributor to the costs of patent enforcement (Hall, Helmers, Rogers and Sena, 2014). Individuals (and small firms) may be less inclined to patent their process inventions because they lack the means and resources to enforce their patents. As Goldstein (2013:64) has pointed out: "A patent claim whose infringement is very hard to discover is a claim with low or no value." In line with this reasoning, Hall, Helmers, Rogers and Sena (2014) conclude that trade secrets are more important for small firms than large firms, Crass, Garcia-Valero, Pitton and Rammer (2019) find stronger degree of substitutability between secrecy and patents for small firms than large firms. Also, Ganglmair and Reimers (2022) find that individuals (and small firms) are more likely to reduce their process patenting when the secrecy option is more attractive, whereas larger companies are not affected by these changes, continuing to patent processes at the same rates.

Lesson 5. *Patents granted to U.S. applicants are more process-intense than those granted to foreign applicants.*

Panel (b) of Figure 6 shows that patents by foreign applicants are less process intense than those by U.S. (domestic) applicants. Differences in the propensity to patent deliver one possible explanation. Similar to the arguments above, enforcement of process patents may be more difficult for foreign firms (in the U.S. market) than it is for domestic firms as their are additional geographic impediments to monitoring. As a consequence, foreign firms are less inclined to file for U.S. patents (and disclose their processes in U.S. patents).

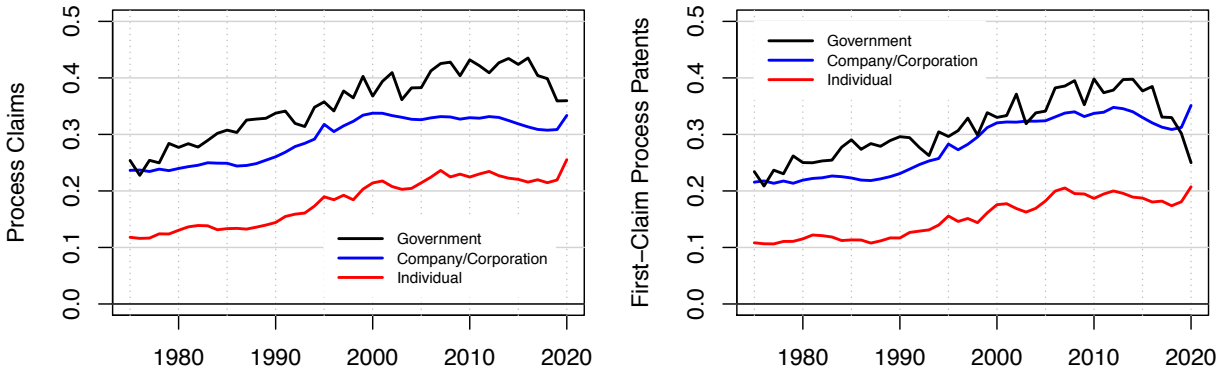
Allison and Lemley (2000:2124) offer a value-based explanation: "Foreign inventors might not file applications in the U.S. for inventions they considered less valuable, or inventions for which the U.S. was unlikely to be a large market." We return to the relative values of process-intense patents further below.

3.3 The Value of Process-Intense Patents

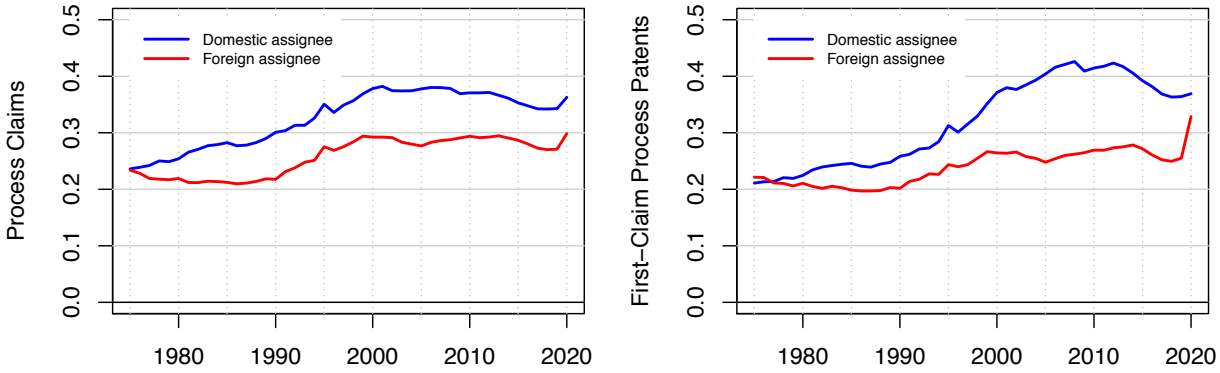
In this section, we turn to the relationship between process intensity and patent value, using three different measurements for patent value (other than the number of independent claims): patent value from stock-market responses (Kogan, Papanikolaou, Seru and Stoffman, 2017), payment of patent maintenance fees (Pakes, 1986), and number of forward citations (Jaffe and Trajtenberg, 1999). We present our results in Figure 7. The left-hand side panels show the value measures for medium process-intensity patents (blue line) and

Figure 6: Domestic and Corporate Applicants File More Process Claims

(a) Assignee Type (Organization)



(b) Geographic Origin



Notes: We show the process intensity for different assignee types (government, company/corporation, or individual) the origin (domestic or foreign) by application year for 1975 to 2020. In the panels on the right we plot the average share of process claims; in the panels on the left we plot the average number of patents with a process claim as their first claim. *Data source: patccat and* https://patentsview.org.

high process-intensity patents (red line) relative to low process-intensity patents. The graph for medium-process intensity patents also provides insights for "mixed patents" with a similar number of product and process claims.[42] The right-hand side panels show the value measures for patents with a first process claim relative to first product claim. All figures plot the value measure by grant year.

Lesson 6. *Process-intense patents are of higher value than their product-intense counterparts.*

The graphs in Panel (a) of the figures show for patents issued between 1926 and 2020 that patent value (Kogan, Papanikolaou, Seru and Stoffman, 2017; Stoffman, Woeppel and Yavuz, 2021) increases in process intensity.[43] We define patents as low process intensity if their share of process claims is less than 1/3. A high process intensity patent has a share of process claims greater than 2/3. Medium process intensity patents are defined as all other patents that are neither low nor high process intensity. The relative value of medium (blue) and high process-intensity patents (red) in the left-hand side panel are above unity, indicating higher values than for low process intensity. Moreover, the patent values for high process-intensity lie above those for medium process-intensity for most of the sample period. The more independent process claims a patent has, the higher its value. The analogous is true when we consider only the first claim of a patent. The value of patents with process claims as their first claims lies consistently above that of patents with a first product claim.

In both panels, we see a convergence of patent values in early to mid 1960s. Whereas in the 1920s to 1950s and then again in the 1970s to the present day, the value of process patents was anywhere between 20% to 60% higher than that of product claims, this wedge temporarily disappeared in the 1960s.

Lesson 7. *Process-intense patents are renewed and their fourth-year maintenance fees paid at higher rates, but have fallen behind in the last decade.*

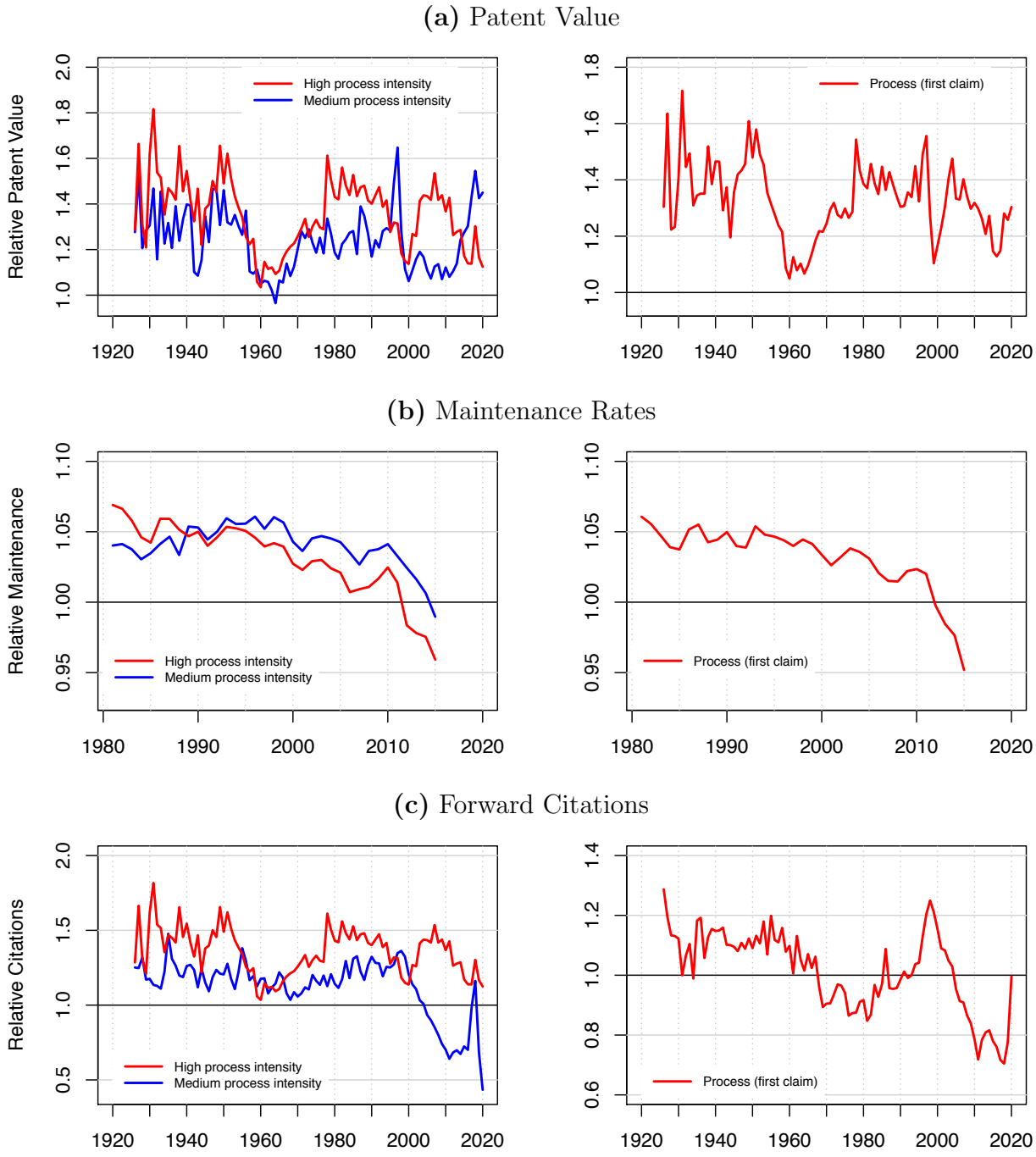
For our second measure of patent value, we use patent holders' maintenance decisions, following Pakes (1986) and others who have pioneered the use of maintenance fee payments to evaluate patent value. Schankerman and Pakes (e.g., 1986); Pakes and Simpson (e.g., 1989); Lanjouw, Pakes and Putnam (e.g., 1998).[44] The payment of patent maintenance fees suggests that patent holders consider their patent valuable or important enough to make that

[42] We define medium process-intensity patents as those with a share of process claims between 1/3 and 2/3.

[43] The updated data for patent value is available for download at https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data.

[44] The USPTO patent maintenance fee events and description files. Download at https://developer.uspto.gov/product/patent-maintenance-fee-events-and-description-files.

Figure 7: Process Claims Have Higher Value



Notes: We show the (relative) value of patents for different levels of process intensity over time (by grant year). For the panels on the left, we define patents as low process intensity (share of process claims less than 1/3), high process intensity (share of process claims more than 2/3), and medium process intensity (all other patents). We plot the value of high process intensity patents (red) and medium process intensity patents (blue) relative to low process intensity patents. For the panels on the right, we define process patents by the type of their first claim and plot the value of process patents relative to product patents. We consider three different measures of patent "value:" patent value as estimated in Kogan, Papanikolaou, Seru and Stoffman (2017) (panel (a)); a patent's fourth year maintenance status (panel (b)); and the number of forward citations (panel (c)) from Kogan, Papanikolaou, Seru and Stoffman (2017). *Data source: patccat, Kogan, Papanikolaou, Seru and Stoffman (2017), and the USPTO maintenance fee events.*

payment (or, rather, pay attention to renewal schedule). To avoid long truncation windows, we use only the fourth-year maintenance payments for our graphs in Panel (b) of Figure 7 and plot the relative maintenance rates for the years 1981 through 2015. In the left-hand side panel, the red line depicts the annual share of renewed high process-intensity patents relative to the annual share of renewed low process intensity-patents. Values above unity imply that high process-intensity patents are renewed more often (analogously for medium process-intensity patents depicted by the blue line). In the right-hand side panel, we depict share of renewed process patents (by first claim) relative to renewed product patents. Again, values above unity imply higher maintenance rates for process patents.

We make two main observations from Panel (b). First, between 1981 and the early 2010s, process-intense patents were more likely renewed than patents without process claims. Beginning in 2010, the maintenance advantage for process-intense patents (LHS) and process patents (RHS) disappeared; and process patents granted in 2015 were up to 5% less likely renewed than product patents. Second, since the early 1990s, medium-process intense patents were consistently more often renewed than high process-intense patents. This means that in the 1990s and 2000s, "mixed patents" with similar numbers of process and product claims had the highest renewal rates and value, higher than patents that were either predominantly process or predominantly product. This observation is related to the results in Toh and Ahuja (forthcoming) who find that higher product-process integration (in the form of mixed patents) increases firm performance.

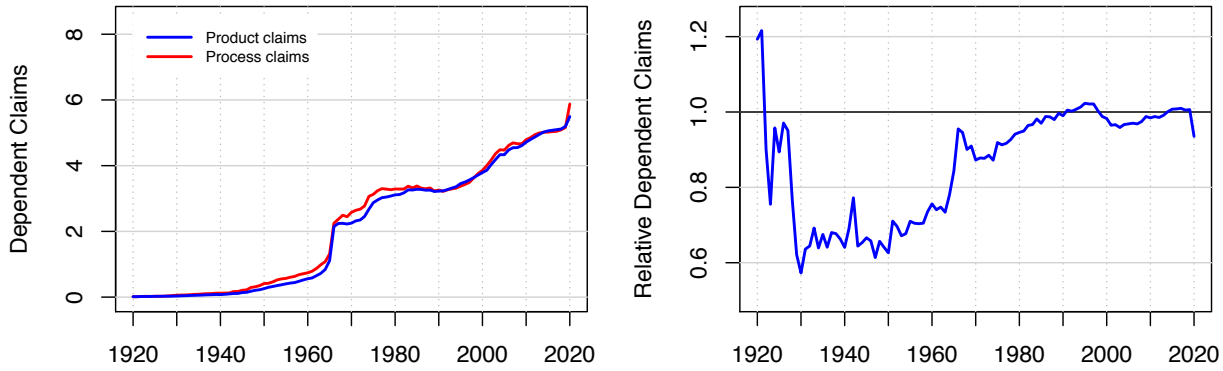
Lesson 8. *Process-intense patents are cited more often by other patents. Patents with a mix of process and product claims have been the least cited over the last two decades. Similarly, process patents (with process claim as their first claim) are less cited than product patents in the 1970s and 1980s and again since the mid 2000s.*

As our third value measure, we look at the number of forward citations to capture the impact of a patent invention (Jaffe and Trajtenberg, 1999; Hall, Jaffe and Trajtenberg, 2005).[45] Panel (c) presents our results for different levels of process intensity (LHS) and process patents by first claim (RHS).

First, until the early 2000s, process-intense patents were cited more often than product-intense patents, hinting at a wider technological impact of process inventions. Until the 1950s and again in the 1980s, high process-intense patents were cited 50% more often than low process-intense patents. Second, since the early 2000s, mixed patents (with medium process intensity) have been the least cited patents. This suggests that in the last 20 years,

[45] We use the number of forward citations in the data files provided by Kogan, Papanikolaou, Seru and Stoffman (2017).

Figure 8: Independent Process Claims Used To Have More Dependent Claims



Notes: We show the number of dependent claims following an independent claim by application year. In the panel on the left, we plot the average number of dependent claims for all independent claims of a patent. For the panel on the right, we plot the number of dependent claims per independent process claim relative to the number of dependent claims per independent product claim. *Data source: patccat.*

specialization or, rather, specialized patents (patents predominantly process or product) have had a greater impact than patents that claim a mix of processes and products. Third, similar to our observations in Panel (a), process intensity had a minor effect on citations (as relative citations are close to unity for both medium and high process-intense patents). Fourth, process patents (with a process claim as their first claim) were more cited than product patents until the 1960s and again in the 1990s – by up to 20% more. In the 1970s and 1980s and since the mid 2000s, process patents are cited less often than product patents.

3.4 Number of Dependent Claims

Our data files contain information only for independent claims. They are stand-alone and do not reference other claims. In that sense, independent claims are less restrictive and broader in scope than dependent claims. A dependent claim refers to one or more other claims (independent or dependent) and limits the subject matter in the referenced claim(s) in various ways. The number of dependent claims per independent claim can therefore provide us with a measure of patent scope (for more on this, see the next section).

Lesson 9. *The number of dependent claims following an independent claim was consistently higher for process claims until 1990.*

In Figure 8, we plot the number of dependent claims referencing a given independent claim. In panel (a) of the figure, we show the absolute number of dependent claims by the type of the independent claim. Dependent claims were not widely used until 1965 when

we see doubling of the number of dependent claims within a year. Since then, the average number of dependent claims per independent claim has increased from just above two to almost six.

In panel (b) of the figure, we show the number of dependent claims per independent product claim relative to the number of dependent claims per process claim. These annual ratios are fairly noisy until the mid 1960s after which we observe on average two dependent claims per independent claim. From the early 1970s until 1990s, the ratio was below 1 and the number of dependent claims for independent product claims was lower than for process claims.

3.5 The Scope of Process Claims

We conclude our analysis by taking yet another look at how patent scope differs by claim type, using the length of claims. Osenga (2012:626–629), Kuhn and Thompson (2019), or Marco, Sarnoff and deGrazia (2019) have argued for an interpretation of shorter claims as broader claims or claims with a wider scope. In Figure 9, we show the length (in number of words) of process and product claims for the full sample period of 1920 to 2020.

Lesson 10. *Process claims are shorter than product claims. Both types become longer over time.*

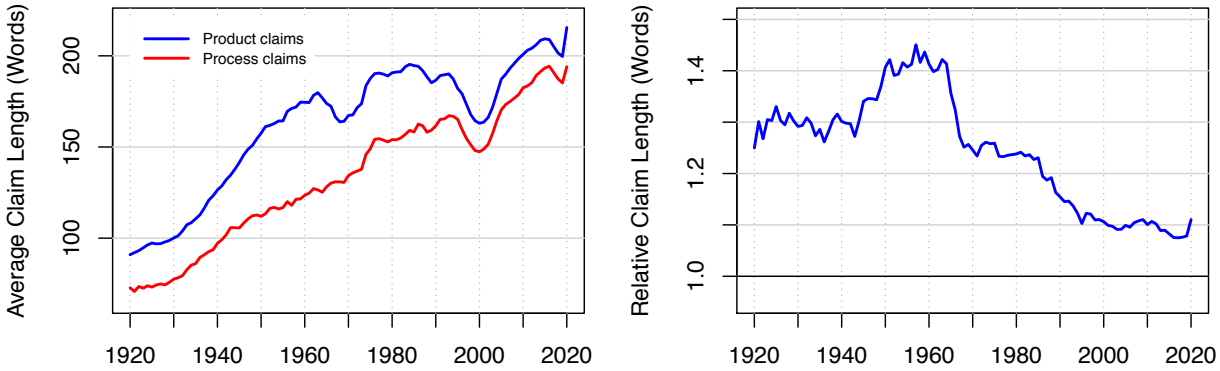
In the left-hand side panels of the figure, we plot the length of claims, in the right-hand side panel we plot the length of product claims relative to process claims. Panel (a) shows the result for all independent claims, whereas Panel (b) shows the results for the first claim of a patent.

First, as predicted by Osenga (2012), we find a positive trend in the length of both types of claims. Process claims have more than doubled in length between 1920 and 2020. Moreover, we find the process claims have consistently been shorter than product claims. This insight is best seen in the right-hand side panels of the figure. Until the 1980s, product claims were at last 20% longer than process claims. We see the largest differences in the late 1940s through the mid 1960s, with a stark decrease around 1965. The difference in length has further decreased since. The average independent product claim filed after 2000 is 5–10% longer than an average process claim. When considering only the first claim of a patent, the length of claims seems to have converged.

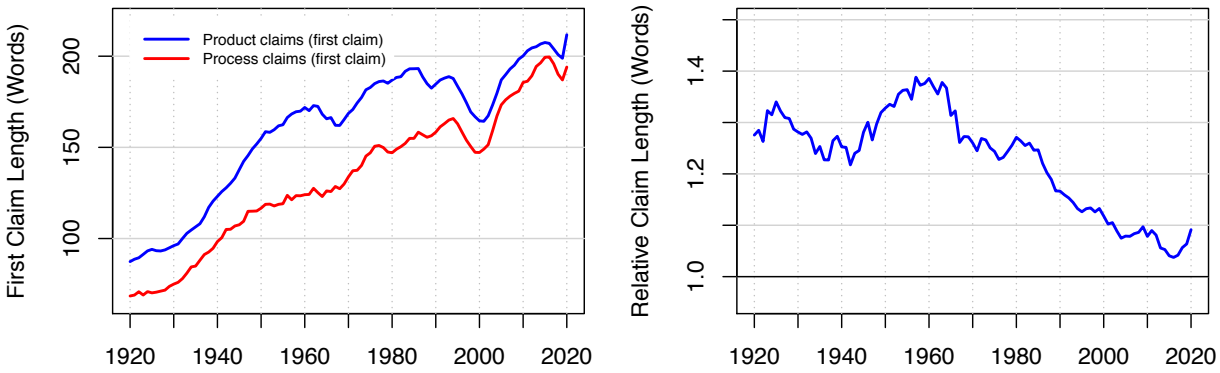
Our results suggest that different claim types are of different shape. In this context, Osenga (2012:644) writes: "It is possible that method claims are consistently of a different shape than machine claims." If such systematic differences exist, then using claim length as a measure of claim breadth (or patent breadth) may be misguided. Such an approach would

Figure 9: Process Claims are Shorter (and Broader)

(a) All Independent Claims



(b) First Claim



Notes: We show the length of independent claims by application year. In the panels on the left, we plot the average length of claims (in words) for all independent claims of a patent (in (a)) and the first claim of a patent (in (b)). For the panels on the right, we plot the length of process claims relative to product claims. *Data source: patccat.*

systematically over-estimate the breadth of process claims. Instead, researchers using claim length as a measure for breadth may need to normalize claim length by claim type so as to avoid comparing apples with oranges.

4 Patent Claim Classification

In this section, we describe our approach for claim classification. The goal is to provide a brief introduction into the general approach and classification rules. In Appendix Section A, we give a more detailed account of each of the steps, including information on data sources and data pre-processing. In Section 5, we describe how we assess the quality of our classifier and present results from this validation exercise.

4.1 Process and Product Claims

For our main classification of patent claims, we combine information obtained from both the preamble and the body of a claim. The preamble is a general description of the invention (e.g., a method, an apparatus, or a device), whereas the body identifies steps and elements (specifying in detail the invention laid out in the preamble) which the applicant is claiming as the invention. The combination of the preamble type and the body type provides us with a more detailed and more accurate classification of claims that also accounts for unconventional drafting approaches.

4.1.1 The Preamble

The classification of the preamble is based on a simple keyword search. We use two lists of keywords, one to identify *method preambles* (with keywords such as *method* or *process*, among others) and one to identify *product preambles* (with keywords such as *apparatus*, *machine*, or *device*, among others). In addition to these two main types, we also identify *by-process preambles* and *empty preambles*. Throughout the paper and in the published data files (in the `label` of a claim), we use capital letters to identify the preamble type.

Method preambles ('M') list a process keyword in the first few words of the claim's preamble.[46] If that first part of a preamble also contains a product keyword, then the process keyword is mentioned before the product keyword.

[46] We use the first eight words of the preamble for this keyword search. See Section A for more information on the parameterization of our classifier.

By-process preambles ('B') use a by-process phrase, such as "by the following process" or "by the following method," but do not use a process keyword prior to the by-process phrase.

Product preambles ('P') list a product keyword in the first few words of the claim's preamble. If that first part of a preamble also contains a process keyword, then the product keyword is mentioned before the process keyword. Also, the preamble is not a product preamble if it uses a by-process phrase.

Empty preambles ('E') do not fall into one of the three above types.

Our classification approach generally prioritizes the preamble type and uses the body type as tie-breaker or when the information obtained from the preamble is inconclusive. We take a conservative view of this "preamble-first" approach by keeping the list of process words and product words relatively short. Both lists comprise statutory terms in addition to a few other terms (see Section A for full lists).

4.1.2 The Body

We complement the preamble type with information on the type of the body of the claim. The body is the part of the claim that describes the individual components or steps of the invention. For the classification of the body, we take a parts-of-speech approach, analyzing the linguistic structure of each (indented) line or bullet point in the body of the claim. The steps of a method or process (in a process claim) are listed using the gerund form of a verb, whereas the elements of an apparatus or device (in a product claim) are listed using nouns. The classification of each line primarily depends on whether a noun occurs before a gerund or whether a gerund occurs before a noun. The keywords themselves could even be the same: "applying paint using a brush" should be interpreted as a step in a process claim, but "a brush for applying paint" should be interpreted as an element in a product claim.

Method body ('m'): The predominant share of lines in the body constitutes steps.[47]

Product body ('p'): The predominant share of lines in the body constitutes elements.

Mixed body ('x'): At least one step or element line can be identified, but neither steps nor elements rise to the level of predominant.

Empty body ('e'): None of the individual lines in the body can be identified as either a step or element.

[47] We use 90% as the critical share of lines (for both method bodies and product bodies). We also use a shorter section of the line for the identification of steps. See Section A for more information on the parameterization of our classifier.

Natural language processors may occasionally misclassify words; context and sentence structure are important, and while patent attorneys deploy a more predictable set of linguistic patterns than would be found in works of literature, errors may occasionally arise. It is possible, even less common, that those errors will lead to a misclassification of a line as a step instead of an element, or vice versa. Our approach requires that the *predominant* share of lines in the body constitute either steps or elements. We believe this minimizes the impact of such rare cases.

4.1.3 Preamble-Body Combinations Determine the Claim Type

In a final step, we combine the classifications of the preamble and the body to obtain the type of the claim. This classification follows five rules:

1. A claim with a method preamble is a process claim, regardless of the body type.

2. A claim with a product preamble is a product claim except when the body is a method body (claim label *Pm*). In this latter case, the claim is a product-by-process claim.

3. A claim with a by-process preamble (i.e., a preamble that explicitly refers to a process or method by which something is made or implemented) is a product-by-process claim except when the body is a product body (claim label *Bp*). In this latter case, the body does not describe the process or method announced in the preamble, and we consider the preamble misleading.

4. A claim with an empty preamble takes the body's informative type as its claim type. That means, a claim with a method body is a process claim (claim label *Em*), and a claim with a product body is a product claim (claim label *Ep*). For a mixed body or an empty body (claim labels *Ex* and *Ee*) we assume that not enough information is available to classify the claim. Such a claim has no claim type.

5. A claim without a body (i.e., a *single-line claim* for which preamble and body cannot be separately identified[48]), takes as its type the type of the preamble (where the entire claim text is treated as the preamble text). For empty preambles (and no body), not enough information is available to classify the claim (no claim type).

[48] As part of the data pre-processing, we convert single-line claims into multi-line claims with a proper preamble-body format. Claims for which this conversion fails are single-line claims to which this classification rule applies. For more details on the single-line claim conversion (and pre-processing steps more generally), see Section A.

4.2 Simple Approach for Process Claims

In addition to the full preamble-body classification, we also construct claim types using a simple keyword approach: a claim is a process claim if it uses the terms *method* or *process* either in the preamble or in the body. It is a product claims otherwise. We consider two different versions of this classification (and we will see in the next section that the former outperforms the latter in terms of accuracy):

1. In the first (stringent) approach, a claim is a process claim if in the preamble either *method* or *process* is used (and a product claim otherwise). This approach requires separate processing of the preamble and the body of the claim. This approach is closely related to those in Angenendt (2018) and Bena and Simintzi (2019).

2. In the second (relaxed) approach, a claim is a process claim if *method* or *process* is used anywhere in the claim – preamble or body (and a product claim otherwise). This is the approach taken by Crouch (2015).

Note that unlike our full preamble-body classification, these two approaches yield full coverage, meaning that all claims can be classified.

5 Validating the Claim Classifier

How well does our automated classifier perform? Our approach is meant to capture conventions and rules of patent-claim drafting. In order to assess the quality of our results, we use a manually curated sample of patent claims as a benchmark. We describe the construction of this sample and present results on the accuracy and coverage of our classifier. We conclude this section with some guiding notes for researchers on how to use the information in the data files to change and adapt the classification to one's needs.

5.1 Benchmark Sample

We use a sample of 9,830 manually classified patent claims drawn from patents issued between 1976 and 2015. We selected an equal number of claims (approximately 250) from each patent grant year. Given the ever-increasing number of patent grants per year, patents with earlier grant years are over-represented in our sample, whereas more recent patents are under-represented. Within each year, our claims, however, are (approximately) representative of the technology distribution.[49] For the manual classification, we hired individuals on

[49]Within each year, we selected patents according to their within-year distributions of NBER technology categories (Hall, Jaffe and Trajtenberg, 2001).

Amazon Mechanical Turk, having each claim manually classified twice – as process claim, product claim, or product-by-process claim. The claims for which we saw disagreement we classified manually for a final classification. This manually classified sample contains 2777 process claims (28.25%), 7024 product claims (71.45%), and 29 product-by-process claims (0.30%).

5.2 Quality Assessment of the Classifier

In Table 3, we summarize the results from our quality assessment exercise. We ask how our automated classification compares to the manually classified benchmark sample. To illustrate the performance of the classifier for different formats of the raw data, we use both the data from the USPTO bulkdata site (multi-line sample) and the preprocessed data from PatentsView (single-line sample).[50]

The overall accuracy of our approach lies at 98.3% and 98.4% for the multi-line and the single-line samples, respectively. This is more than 2.7 percentage points more accurate than the preamble-only simple approach with an accuracy of 95.6% and 95.7%, respectively. Our combined preamble-body approach improves accuracy by filling about two thirds of the gap to 100% accuracy. However, this improvement comes at a cost in terms of coverage. Claims for which insufficient information (preamble or body) is available, cannot be classified. Overall, we classify between 98.3% and 98.5% of the claims in our benchmark sample.

Notice, however, that having claims in multi-line format does not necessarily improve the quality of our classification. Claims that are converted from single-line format to multi-line format exhibit higher accuracy than original multi-line claims.[51] We achieve the lowest accuracy for single-line claims that cannot be converted.[52] Jepson claims are also claims that undergo a conversion step. The peculiar structure of their preamble makes them particularly difficult to classify (using the preamble-body approach). We see this in the noticeably lower accuracy rate for Jepson claims compared to other claims.

[50] In single-line formatted claims, the preamble and all lines of the body are concatenated; they appear as one line or paragraph in the data. In multi-line formatted claims, on the other hand, the preamble-body structure is preserved; the paragraphs and individual lines (bullet points) in the body are separated. In a data pre-processing step, we convert (when possible) single-line claims into multi-line claims. See Appendix A for more details.

[51] For converted claims, we assume that all lines in the body are at the same level of indentation. In original multi-line format claims (as obtained from the USPTO's raw data files), lines in the body are of various levels of indentation. For the classifier, we use only the second (and in some cases the third) level of indentation.

[52] The number of these not converted claims is low in both samples. In the multi-line sample, 16 claims (1.1% of all single-line claims in that sample) cannot be converted. In the single-line sample, 24 claims (0.25% of that sample) cannot converted.

Table 3: Quality Results for Process/Product/Product-by-Process

	Multi-line sample		Single-line sample	
	Accuracy	**Coverage**	**Accuracy**	**Coverage**
Overall results	0.983	0.983	0.984	0.985
Original multi-line	0.984	0.985		
Converted	0.991	0.973	0.987	0.987
Partially converted	0.977	0.979	0.977	0.985
Not converted	0.857	0.438	0.9	0.417
Jepson claims	0.911	0.891	0.901	0.868
Non-Jepson claims	0.985	0.986	0.986	0.989
Simple approach (preamble only)	0.956	1	0.957	1
Simple approach (full claim)	0.907	1	0.906	1
Conditional on manual classification				
Process	0.979	0.989	0.974	0.983
Product	0.986	0.981	0.99	0.987
Product-by-Process	0.821	0.966	0.519	0.931
Conditional on automated classification				
Process	0.995	1	0.994	1
Product	0.994	1	0.99	1
Product-by-Process	0.176	1	0.173	1
Claim counts				
All claims		9830		9830
Single-line claims		1492		9830
Regular claims		8338		0
Jepson claims (regular or single-line)		265		303

We further summarize the accuracy and coverage of our approach, conditional on the classification source. The figures conditional on manual classification reflect the accuracy of our approach for all manually classified process, product, and product-by-process claims, respectively. For instance, we correctly classify 97.8% of all manually classified process claims. The figures conditional on the automated approach reflect the accuracy of our approach for all algorithmically classified process, product, and product-by-process claims, respectively. For instance, the manual classification agrees with 99.5% of all algorithmically classified process claims. In both sets of numbers, we can see that we do well with process and product claims. Product-by-process claims, however, exhibit low accuracy. In fact, our approach seems too aggressive: Only 17.4% of all claims that we classify as product-by-process claims are indeed of that type in the benchmark sample.

In Table 4, we report accuracy and coverage (multi-line sample) for different technology categories, following the categorization in Hall, Jaffe and Trajtenberg (2001). We obtain the

Table 4: Quality by (NBER) Technology Category

NBER Technology Category	Accuracy	Coverage
Chemical	0.987	0.973
Computers + Communications	0.961	0.988
Drugs + Medical	0.995	0.989
Eletrical + Electronic	0.990	0.986
Mechanical	0.996	0.979
Others	0.996	0.979

Notes: We provide results for the accuracy and coverage of our claims classification for the benchmark sample, broken down by the NBER technology categories from Hall, Jaffe and Trajtenberg (2001). *Accuracy* is the share of correctly classified patent claims (conditional on being classified). *Coverage* is the share of process claims for which the classifier determines a claim type.

lowest coverage (97.3%) for claims in patents related to chemicals. The lowest accuracy is for claims in patents in the computers and communications category, which including software patents and business method patents.

5.3 Adjustable Classifications

Researchers using the data may have different needs or prefer a different accuracy-coverage balance. There are a number of ways of adjusting the claim classification with the information in the published data files.

First, the preamble-body labels allow for easy changes of the claim types. Our classification follows the rules laid out in Section 4. One possible source for guidance for reclassification is Table A.5. Changing the rules of classification can affect both accuracy and coverage of the classifier. Note, however, that higher accuracy will come at the cost of coverage, and vice versa. For instance, classifying claim labels *Ex*, *Ee*, and *E-* will increase coverage, but lower accuracy because the accuracy of these labels is at beast 0.690, 0.952, and 0.444, respectively.

Second, in the data files, we provide the keyword that determines the classification of a given preamble as product or method preamble alongside the keyword of the other preamble type (if it exist). Researchers who prefer to *drop* a keyword from one of the keyword list can easily change the preamble type to empty preamble or, if an alternative keyword exist change the type from product preamble to method preamble (or vice versa). In addition to the keywords, we also include the first 15 words of the preamble in the data files. Researchers who prefer to add a keyword can easily do so without having to process the raw data.

Third, we include the number of steps, elements, and total number of lines in the body for each claim. With this information, researchers can readily change the classification of the body by changing the step and element thresholds for method and product preambles.

For our classification, we choose thresholds of 90% for both types of bodies. Lowering the thresholds (symmetrically or asymmetrically) will reduce the number of mixed bodies.

6 Concluding Remarks

We document the use of process claims in the U.S. over the last century. Using novel data on the type of independent patent claims, we show an increase of the annual share of process claims of about 25 percentage points (from below 10% in 1920). This rise in process intensity is not limited to a few patent classes but can be observed across a broad spectrum of technologies. Process intensity varies by applicant type: companies file more process-intense patents than individuals, and U.S. applicants file more process-intense patents than foreign applicants. We further show that patents with higher process intensity have indicators of higher value as compared to other patents but are not necessarily cited more often. Last, process claims are on average shorter than product claims; but this gap in length has narrowed since the 1970s. These patterns suggest that the patent breadth and scope of process-intense patents is overestimated when claim types are not accounted for. We conclude by describing in detail the code used to construct the claim-type data, showing results from a data-validation exercise (using close to 10,000 manually classified patent claims), and providing guidance for researchers on how to alter the classification outcome to adapt to individuals' needs.

References

Abernathy, William J. and James M. Utterback, "Patterns of Industrial Innovation," *Technology Review*, 1978, *80* (7), 40–47.

Allison, John R. and Mark A. Lemley, "Who's Patenting What? An Empirical Exploration of Patent Prosecution," *Vanderbilt Law Review*, 2000, *53*, 2099–2174.

Angenendt, David T., "Easy to Keep, But Hard to Find: How Patentable Inventions are Being Kept Secret," 2018. Unpublished manuscript, University of Bologna.

Arundel, Anthony, "The Relative Effectiveness of Patents and Secrecy for Appropriation," *Research Policy*, 2001, *30* (4), 611–624.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," 2021. Unpublished manuscript, available at https://ssrn.com/abstract=3651052.

Bena, Jan and Elena Simintzi, "Machines Could not Compete with Chinese Labor: Evidence from US Firms' Innovation," 2019. Unpublished manuscript, available at https://ssrn.com/abstract=2613248.

Branstetter, Lee G., Jong-Rong Chen, Britta Glennon, and Nikolas Zolas, "Does Offshoring Production Reduce Innovation: Firm-Level Evidence from Taiwan," NBER Working Paper 29117, National Bureau of Economic Research, Cambridge, Mass. 2021.

Chen, I-Ju, Po-Hsuan Hsu, and Yanzhi Wang, "Staggered Boards and Product Innovations: Evidence from Massachusetts State Bill HB 5640," *Research Policy*, 2022, *51* (4), 104475.

Cohen, Wesley M. and Steven Klepper, "A Reprise of Size and R&D," *Research Policy*, 1996, *106* (437), 925–951.

Cohen, Wesley M. and Steven Klepper, "Firm Size and the Nature of Innovation within Industries: The Case of Process and Product R&D," *Review of Economics and Statistics*, 1996, *78* (2), 232–243.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh, "Protecting their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)," NBER Working Paper 7552, National Bureau of Economic Research, Cambridge, Mass. 2000.

Crass, Dirk, Francisco Garcia-Valero, Francesco Pitton, and Christian Rammer, "Protecting Innovation Through Patents and Trade Secrets: Evidence for Firms with a Single Innovation," *International Journal of the Economics of Business*, 2019, *26* (1), 117–156.

Crouch, Dennis D., "Method Patent Claims," 2015. PATENTLY-O Blog (Sept. 29, 2015) at https://patentlyo.com/patent/2015/09/method-patent-claims.html (last visited: February 14, 2022).

Cucculelli, Marco and Valentina Peruzzi, "Innovation Over the Industry Life-Cycle. Does Ownership Matter?," *Research Policy*, 2020, *49* (1), 103878.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, May 2020, *135* (2), 561–644.

De Rassenfosse, Gaétan, Marco Grazzi, Daniele Moschella, and Gabriele Pellegrino, "International Patent Protection and Trade: Transaction-Level Evidence," 2020. Unpublished manuscript, available at https://ssrn.com/abstract=3562618.

Ganglmair, Bernhard and Imke Reimers, "Visibility of Technology and Cumulative Innovation: Evidence from Trade Secrets Laws," 2022. available at https://ssrn.com/abstract=3393510.

Goldstein, Larry M., *True Patent Value: Defining Quality in Patents and Patent Portfolios*, Chicago, Ill.: True Value Press, 2013.

Graham, Stewart and Deepak Hegde, "Disclosing Patents' Secrets," *Science*, 2015, *347* (6219), 236–237.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498, National Bureau of Economic Research, Cambridge, Mass. 2001.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg, "Market Value and Patent Citations," *RAND Journal of Economics*, 2005, *36* (1), 16–38.

Hall, Bronwyn H., Christian Helmers, Mark Rogers, and Vania Sena, "The Importance (or Not) of Patents to UK Firms," *Oxford Economic Papers*, 2013, *65* (3), 603–629.

Hall, Bronwyn H., Christian Helmers, Mark Rogers, and Vania Sena, "The Choice between Formal and Informal Intellectual Property: A Review," *Journal of Economic Literature*, 2014, *52* (2), 375–423.

Hall, Bronwyn H., Francesca Lotti, and Jacques Mairesse, "Evidence on the Impact of R&D and ICT Investments on Innovation and Productivity in Italian Firms," *Economics of Innovation and New Technology*, 2013, *22* (3), 300–328.

Huergo, Elena and Jordi Jaumandreu, "How Does Probability of Innovation Change With Firm Age?," *Small Business Economics*, 2004, *22* (3), 193–207.

Jaffe, Adam B. and Manuel Trajtenberg, "International Knowledge Flows: Evidence From Patent Citations," *Economics of Innovation and New Technology*, 1999, *8* (1-2), 105–136.

Keum, Daniel, "Firing Costs and the Decoupling of Technological Invention and Post-Invention Investments," 2020. Unpublished manuscript, available at https://ssrn.com/abstract=3774703.

Kindlon, Audrey E. and John E. Jankowski, "Rates of Innovation among U.S. Businesses Stay Steady: Data from the 2014 Business R&D and Innovation Survey," InfoBrief 17/321, U.S. National Science Foundation, Washington, D.C. 2017.

Klepper, Steven, "Entry, Exit, Growth, and Innovation Over the Product Life Cycle," *American Economic Review*, 1996, *86* (3), 562–583.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

Kuhn, Jeffrey M. and Neil C. Thompson, "How to Measure and Draw Causal Inferences with Patent Scope," *International Journal of the Economics of Business*, 2019, *26* (1), 5–38.

Lanjouw, Jean O., Ariel Pakes, and Jonathan Putnam, "How to Count Patents and Value Intellectual Property: The Uses of Patent Renewal and Application Data," *Journal of Industrial Economics*, 1998, *46* (4), 405–434.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter, "Appropriating the Returns from Industrial Research and Development," *Brookings Papers on Economic Activity*, 1987, *3*, 783–831.

Link, Albert N., "A Disaggregated Analysis of Industrial R&D: Product versus Process Innovation," in Devandra Sahal, ed., *The Transfer and Utilization of Technical Knowledge*, Lexington, Mass.: Lexington Books, 1982.

Lunn, John, "An Empirical Analysis of Process and Product Patenting: A Simultaneous Equation Framework," *Journal of Industrial Economics*, 1986, *34*, 319–330.

Ma, Song, "Technological Obsolescence," NBER Working Paper 29504, National Bureau of Economic Research, Cambridge, Mass. 2021.

Mansfield, Edwin, "Patents and Innovation: An Empirical Study," *Management Science*, 1986, *32* (2), 173–181.

Marco, Alan C., Joshua D. Sarnoff, and Charles deGrazia, "Patent Claims and Patent Scope," *Research Policy*, 2019, *48*, 103790.

McGahan, Anita M. and Brian S. Silverman, "How Does Innovative Activity Change as Industries Mature?," *International Journal of Industrial Organization*, 2001, *19* (7), 1141–1160.

Osenga, Kristen J., "The Shape of Things to Come: What We Can Learn from Patent Claim Length," *Santa Clara Computer & High Technology Law Journal*, 2012, *28* (3), 617–656.

Pakes, Ariel, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 1986, *54* (4), 755–784.

Pakes, Ariel and Margaret Simpson, "Patent Renewal Data," *Brookings Papers on Economic Activity: Microeconomics Annual*, 1989, pp. 331–401.

Png, Ivan P.L., "Law and Innovation: Evidence from State Trade Secrets Laws," *Review of Economics and Statistics*, March 2017, *99* (1), 167–179.

Png, Ivan P.L., "Secrecy and Patents: Theory and Evidence from the Uniform Trade Secrets Act," *Strategy Science*, 2017, *2* (3), 176–193.

Rammer, Christian, Torben Schubert, Paul Hünermund, Mila Köhler, Younes Iferd, and Bettina Peters, "Dokumentation zur Innovationserhebung 2015," ZEW Dokumentation 16-01, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany 2016.

Risch, Michael, "America's First Patents," *Florida Law Review*, 2012, *64* (5), 1279–1336.

Robinson, W. Keith, "No 'Direction' Home: An Alternative Approach to Joint Infringement," *American University Law Review*, 2012, *62*, 59–122.

Rosenberg, Morgan D., *Patent Application Drafting: A Practical Guide*, New York, N.Y.: LexisNexis, 2017.

Schankerman, Mark and Ariel Pakes, "Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period," *Economic Journal*, 1986, *96* (384), 1052–1076.

Scherer, Frederic M., "Research and Development Resource Allocation Under Rivalry," *Quarterly Journal of Economics*, 1967, *81*, 359–394.

Scherer, Frederic M., "Inter-Industry Technology Flows and Productivity Growth," *Review of Economics and Statistics*, 1982, *64* (4), 627–634.

Scherer, Frederic M., "Inter-Industry Technology Flows in the United States," *Research Policy*, 1982, *11* (4), 227–245.

Scherer, Frederic M., "Using Linked Patent and R&D Data to Measure Interindustry Technology Flows," in "R&D, Patents, and Productivity," Chicago, Ill.: University of Chicago Press, 1984, pp. 417–464.

Scherer, Frederic M., "Changing Perspectives on the Firm Size Problem," in Zoltan J. Acs and David B. Audretsch, eds., *Innovation and Technological Change: An International Comparison*, New York, N.Y.: Harvester Wheatsheaf, 1991.

Stoffman, Noah, Michael Woeppel, and M. Deniz Yavuz, "Small Innovators: No Risk, No Return," 2021. Unpublished manuscript, available at https://ssrn.com/abstract=3291471.

Toh, Puay Khoon and Gautam Ahuja, "Integration and Appropriability: A Study of Process and Product Components within a Firm's Innovation Portfolio," *Strategic Management Journal*, forthcoming.

A Data Construction: Assumption and Code

In this appendix section, we provide a detailed account of our claim classification. We list our data sources, describe the data preparation procedure, and introduce the functions used for the classification.

A.1 Data Sources and Preparation

Our main data sources (Table A.1) are the USPTO's Patent Grant Full Text Data files available at https://bulkdata.uspto.gov (Bulk Data Storage System BDSS), PatentsView at https://patentsview.org/download/data-download-tables, and the Google Patents Public Data.

Table A.1: Data Sources

Source	Format and notes	Sample
USPTO	XML format	2002–2020
USPTO	fixed-width text (APS, Green Book)	1976–2001
PatentsView	data tables (claims in single-line formatting only)	1976-2020
Google	BigQuery	1920–1975

Notes: In the main text, we present results for patents granted between 1920 and 2020. The complete data also contain a data file with patent claim information for patents issued between 1836 and 1919. The raw data for these historic patents we obtained from the Google Patents Public Data. For more detailed information on the raw data format for patents issued between 1976 and 2001 (Green Book), see https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/1976.

For initial processing of the raw data (i.e., full-text documents), we extracted the claims text, determined dependency relationship between individual claims, and assigned values for the level of indentation for each line in a claim.[53]. For patent claims that were initially filed and published with indented lines, we are able to preserve this multi-line structure and utilize it in our body classification. The claims in Table A.2 from U.S. patents 6,009,555 and 6,635,133 are two such examples of claims in multi-line format. The column `varPatentClaim` contains the patent-claim identifier we use in our data (8-digit patent number and 4-digit claim number); the column `varLevel` indicates the indentation level of a line. The claims of patents, for which this multi-line structure does not apply or is no longer preserved (the claims obtained from the Google Patents Public Data and PatentsView), our classifier converts (when possible) from single-line format to multi-line format (see details below).

[53]The code for the data download and the pre-processing steps is available at https://gitlab.com/lion-sz/dependency-scraper.

Table A.2: Two Examples for Patent Claims

U.S. Patent 6,009,555: Multiple component headgear system		
`varPatentClaim`	`varText`	`varLevel`
06009555-0001	1. A headgear apparatus comprising:	1
06009555-0001	a headband member having a frontal portion;	2
06009555-0001	a visor member removably secured to said frontal portion of said headband; and	2
06009555-0001	an eye shield member removably secured to said frontal portion of said headband.	2

U.S. Patent 6,635,133: Method for making a multilayered golf ball		
`varPatentClaim`	`varText`	`varLevel`
06635133-0001	1. A method of making a ball, comprising:	1
06635133-0001	forming an inner sphere by forming an outer shell with a fluid mass center;	2
06635133-0001	forming a plurality of core parts;	2
06635133-0001	arranging and joining the core parts around the inner sphere to form an assembled core;	2
06635133-0001	molding a cover around the assembled core.	2

In Table A.3, we list the input variables used for the classifier. Required inputs is the minimum information necessary for the classifier to function (i.e., a patent-claim identifier and the claim text). Additional variables (i.e., information on the indentation level and the sequence in which the lines of a claim are printed in the patents) are used for some sub-routines of the classifier. If missing, the classifier constructs the necessary variables.[54]

In a last pre-processing step, we drop all dependent claims from the claims text data. We have developed the classifier with independent claims in mind (and we have tested using only independent claims). Our classifier can be applied to any English-language claims text file with a unique patent-claim identifier and the (single-line or multi-line) text of the patent claim.

The function `-fn.patccat-` performs the classification. It calls on a number of other sub-routines that perform individual steps of our classifier. Function `-fn.patccat-` takes as input an R object `data.frame -data-` (see Table A.3) with the claims text, the column

[54]If a patent number and a claim number are provided, the classifier function `-fn.patccat-` will first construct a patent-claim identifier using the individual identifiers. If the text is of multi-line format, an additional variable that indicates the level of indentation (where level 1 is the preamble and levels 2 or higher are for the body) can be included. If this variable is not included, then the first line for a given patent claim is taken to be the preamble, and all other lines are body lines at the same level of indentation (level 2).

Table A.3: Input Data (`data.frame -data-`)

Variable	Description	Format
	Required Inputs	
`varPatentClaim`	Unique identifier for a patent claims (as patent-claim combination)	integer, string
`varPatent`	*Patent number/identifier*	*integer, string*
`varClaim`	*Claim number/identifier*	*integer, string*
`varText`	The text of the given line of the claim, including the leading outline designation (1., a., A., etc.)	string
	Additional Inputs	
`varLevel`	Level of indentation of a line. For a multi-line formatted claim, the preamble as highest-order line of a claim has a value of `varLevel` = 1; its body lines that are indented once have a value of `varLevel` = 2. All body lines with further indentations have higher values for `varLevel`. For a single-line formatted claim, the single line has a value of `varLevel` = 1.	integer
`varSequence`	Ordered sequence number of lines within a `varPatent`. `varSequence` equal to 1 is the first line (preamble if multi-line claim) of the first claim of the patent. The first line of a claim (the preamble if multi-line claim or the entire claim if single-line claim) is the smallest value of `varSequence` for a given `varClaim`.	integer
`varID`	Unique line identifier. It is not directly used by the classifier, but serves as a useful identifier for each row.	integer ≥ 0

name for the patent-claim identifier (`-varPatentClaim-`), the column name for the claim text (`-varText-`), the column name for the level of indentation (`-varLevel-`, not required), the column name for the sequence in which the individual lines of a claim appear in the patent (`-varSequence-`, not required), and the column name of a simple running index (`-varID-`, not required).

A.2 Reformat (`-fn.reformatdata-`)

Our keyword search approach for the preamble classifier requires that the signal terms in the list of process words and product words appear at the beginning of the preamble. Likewise for the body classifier that searches for nouns or gerund forms of the verb at the beginning of each line of the body. Most of the claims in our data come well formatted. For the rest, we perform three steps to convert the claim text into a format we can process. Function `-fn.reformatdata-` performs this conversion. The function takes a `data.frame` with the claims text as input. The function calls on three sub-routines: `-fn.jepsonreformat-`, `-fn.singlesplitter-`, and `-fn.beginWithIn-`. We describe each below.

A.2.1 Jepson Claim Reformatting (`-fn.jepsonreformat-`)

The preamble of a Jepson (or improvement) claim first describes what is known (or in the prior art). We refer to this first component as "prior-art part." It is followed by a transitional phrase (such as "wherein the improvement comprises"). After this transitional phrase, the claim lists everything that is considered an improvement. This latter component we refer to as "improvement part." For our preamble classifier, we use only the improvement part. We split the preamble at the transitional phrase and treat the text of the improvement part as the text of the preamble. We do not use the prior-art part for our analysis.

A.2.2 Single-Line Claim Splitting (`-fn.singlesplitter-`)

Claims that are in single-line format (where the preamble and all lines of the body are concatenated and appear as one line or paragraph), are not ready for our preamble-body approach. We first convert such claims into multi-line claims before applying our classifier. For this, we take two steps: First, we split the claim at the transitional phrase (e.g., "comprising") or certain punctuation characters to obtain the preamble and the body. Second, we identify enumeration counters in the text of body and use these to split the body into individual lines. We refer to three types of converted single-line claims:

1. *Fully-converted* claims have a preamble and a multi-line body. We treat them as proper multi-line claims and apply the baseline version of our classifier (as with all other multi-line claims in the data).

2. *Partially-converted* claims have a preamble and a single-line (non-converted) body. We treat this single-line body as the only line in the body and apply the baseline version of the classifier.

3. *Non-converted* claims are single-line claims that cannot be converted. By definition, they have an empty body. We apply a single-line version of our classifier.

A.2.3 In-Environment Claims (`-fn.beginWithIn-`)

A third (and relatively small category) of claims are those beginning with the word "in" or "for" followed by a statement of the environment. Such claims have the structure of Jepson claims, but do not explicitly specify an improvement (and lack the respective transitional phrase). We trim the beginning of the preamble up to the first comma and use the text following the first comma for our preamble classifier.

A.3 Preamble Type (`-fn.preambletype-`)

For the classification of the preamble, we use two keyword lists: a list with terms identifying a process preamble (`-processwords-`) and a list with terms identifying a product preamble (`-productwords-`). We choose short lists with statutory terms and a few strong terms to prioritize the preamble information in our claim classifier. To identify the preamble as either a process preamble or product preamble, a word from the process list or product list must appear in the first 8 words of the preamble. For single-line claims (after the reformatting steps described above), this means that the respective words appear in the first 8 words of the full claim.

Function `-fn.preambletype-` performs the preamble classification. It takes several items as inputs: `data.frame` with the claims text, list of process words (`-processwords-`), list of product words (`-productwords-`), number of parameter values used for the classification, and `TRUE` or `FALSE` flag of whether the claims are single-line claims. The parameter values (such as number of words within which a keyword must be used) are specified in `-my.params-`.

In the sequel, we provide classification details for *method (or process) preambles* and *product preambles*. We also describe a third preamble type – *by-process preambles* – that we use for the classification of product-by-process claims.

A.3.1 Method Preamble ('M')

A method preamble ('M') is the preamble of a process claim. It names a method or process as the invention described by the claim. The two main keywords of the list of process words are *process* and *method* as the most widely used terms to describe a process claim (e.g., in patent 6,635,133). A preamble is a method preamble if, within the first 8 words of the preamble, one of the process words is used, but none of the following applies: (1) one of the product words is used before the process word, (2) the process word is immediately followed by a noun, or (3) the terms *for* or *by* are used within 3 words (before) of the process word. Terms such as "computer-implemented method" or "machine-controlled process" are exempt from (1).

The list of process words comprises the following terms: "method", "process", "approach", "manner", "practice", "recipe", "scheme", "technique", and "treatment".

A.3.2 Product Preamble ('P')

A product preamble ('P') is the preamble of a product claim. It names a machine, an apparatus, or a device (a "thing") as the invention described in the claim. For our list of

product words, we use statutory terms and a short list of very common terms used to describe "things." A preamble is a product preamble if, within the first 8 words of the preamble, one of the product words is used, but none of the following applies: (1) one of the process words is used before the product word, or (2) the preamble uses a by-process phrase (see below). The list of product words comprises the following terms: "system", "apparatus", "device", "machine", "computer", "assembly", "circuit", "data", "semiconductor", "composition", "medium", and "means".[55]

A.3.3 By-Process Preamble ('B')

A preamble is a by-process preamble ('B') if it uses a by-process phrase. Such a phrase is "by [up to 3 words] process" or "by [up to 3 words] method." The preamble is not a by-process preamble if the preamble is a method preamble (i.e., when a process word is used before a by-process phrase). In single-line claims, we search for by-process phrases in the first 50 words of the claim.[56] In multi-line claims, we impose no such word limit but consider the text of the entire preamble.

A.3.4 Empty Preamble ('E')

A preamble that is neither method preamble, product preamble, or by-process preamble is classified as an empty preamble ('E').

A.4 Body Type (`-fn.bodytype-`)

The body of a claim contains a number of (indented) lines of text, where lines describe steps (of a method or process) or elements (of an apparatus, device, or machine). Our approach for the classification of lines is a parts-of-speech approach. We classify the body in two steps. First, we identify each line in the body as either a (1) step, (2) element, or (3) a line that refers to other parts of the claim via the terms "said," "wherein," "whereby," or similar. Second, if the lines of a body are predominantly steps, then the body is a method body ('m'); if the lines are predominantly elements, then the body is a product body ('p').

Function `-fn.bodytype-` performs the body classification. It takes several items as inputs: a `data.frame` `-data-` and a number of parameter values used for the classification. The parameter values are specified in object `-my.params-`. The function `-fn.bodytype-`

[55] We also include a longer list of product words that comprises the 100 most frequently used product words in a sample of 1% of all claims for patents issued between 1976 and 2015.

[56] We restrict the number of words for single-line claims to minimize the noise from the language of the body in such claims.

also calls on the subroutine `-fn.POStagger-` that performs the parts-of-speech tagging. We use the openNLP POS tagger in the Apache openNLP library.[57]

A.4.1 Steps

In most cases, a line is a step if the gerund form of a verb occurs within the first 2 words of the line (e.g., "forming," "arranging," and "molding" in patent 6,635,133) and none of the following applies: (1) the gerund form of the verb is from a list of commonly used words that do not describe steps of a method or process;[58] (2) a noun is used before the gerund form; (3) the word "means" is used in the line; (4) the line begins with words "said," "when," "wherein," "whereas," or similar; or (5) the line begins with a cardinal number or a determiner.

A.4.2 Elements

A line is an element if a noun occurs within the first 10 words of the line (e.g., "headband," "member," or "visor" in patent 6,009,555) and none of the following applies: (1) the line is a step; or (2) the line begins with words "said," "when," "wherein," "whereas," or similar. A line is also an element if it uses the word "means" (indicating a means-plus-function claim) or if one of the following constructions are used: (1) a noun is sandwiched between a gerund form of a verb and another form of a verb; (2) a noun is immediately preceded by a pairing of a cardinal number or determiner and a gerund; or (3) a noun is immediately preceded by a triple of a cardinal number or determiner, an adjective, and either an adjective or a gerund.

A.4.3 Predominant Line Types: Classifying the Body

For the classification of the body, we aggregate the information obtained for each line. We consider four different body types. (1) The body is a method body ('m') if 90% or more of all lines of a body (classified as either steps or elements) are steps. (2) The body is a product body ('p') if 90% or more of all lines of the body are elements. (3) The body is a mixed body ('x') if at least one line is either a step or an element, but neither line type dominates the body. (4) The body is empty ('e') if it has neither steps nor elements.

[57] For more information on the R interface to the Apache OpenNLP tools, see the library's CRAN page at https://cran.r-project.org/web/packages/openNLP/index.html.

[58] This list contains the terms "being", "comprising", "consisting", "including", "having", "depending", "indicating", "representing", "containing", and "housing".

Table A.4: Classification Table

Preamble type	Body type	Claim	Label
Method	Method	Process	Mm
Method	Product	Process	Mp
Method	Mixed	Process	Mx
Method	Empty	Process	Me
Method	—	Process	M–
Product	Method	Product-by-Process	Pm
Product	Product	Product	Pp
Product	Mixed	Product	Px
Product	Empty	Product	Pe
Product	—	Product	P–
By-Process	Method	Product-by-Process	Bm
By-Process	Product	Product	Bp
By-Process	Mixed	Product-by-Process	Bx
By-Process	Empty	Product-by-Process	Be
By-Process	—	Product-by-Process	B–
Empty	Method	Process	Em
Empty	Product	Product	Ep
Empty	Mixed	No Claim Type (–)	Ex
Empty	Empty	No Claim Type (–)	Ee
Empty	—	No Claim Type (–)	E–

A.5 Claim Type (`-fn.claimtype-` and `-fn.claimtype.single-`)

In a final step, we combine the information for preambles and bodies. Each preamble-body combination corresponds to a claim type as specified in Table A.4. The labels indicate the respective preamble-body combination (capital letters for the preamble type and lower-case letters for the body type). For single-line claims (so that the body type is '–') the claim type follows the preamble type. For more information on the general approach behind the classification table, see the main text. Functions `-fn.claimtype-` (for multi-line claims) and `-fn.claimtype.single-` (for single-line claims) perform the claim classification, using output from functions `-fn.preamblemtype-` and `-fn.preambletype-` (only for multi-line claims).

B Further Data Validation

B.1 Further Analysis of Data Accuracy

For a better understanding of where the accuracy of our classifier comes from, we provide additional information in Table A.5. We break down all claims in our benchmark sample

Table A.5: Drafting Quality: Classification by Preamble-Body Combination

						Benchmark		
Label	**Preamble**	**Body**	**Claim**	**Count**	**Accuracy**	**Process**	**Product**	**Product-by-Proc.**
Mm	Method	Method	Process	1737	1	1	0	0
Mp	Method	Product	Process	251	0.9960	0.996	0.004	0
Mx	Method	Mixed	Process	621	1	1	0	0
Me	Method	Empty	Process	17	1	1	0	0
M-	Method	-	Process	3	0.6667	0.667	0	0.333
Pm	Product	Method	Prod-by-Proc.	112	0.0357	0.170	0.795	0.036
Pp	Product	Product	Product	3454	0.9988	0.001	0.999	0
Px	Product	Mixed	Product	155	0.9613	0.039	0.961	0
Pe	Product	Empty	Product	28	1	0	1	0
P-	Product	-	Product	4	1	0	1	0
Bm	By-Proc.	Method	Prod-by-Proc.	12	1	0	0	1
Bp	By-Proc.	Product	Product	3	1	0	1	0
Bx	By-Proc.	Mixed	Prod-by-Proc.	5	1	0	0	1
Be	By-Proc.	Empty	Prod-by-Proc.	2	1	0	0	1
B-	By-Proc.	-	Prod-by-Proc.	0		0	0	0
Em	Empty	Process	Method	74	0.8514	0.851	0.122	0.027
Ep	Empty	Product	Product	3188	0.9906	0.009	0.991	0
Ex	Empty	Mixed	–	71		0.31	0.69	0
Ee	Empty	Empty	–	84		0.048	0.952	0
E-	Empty	-	–	9		0.444	0.444	0.111

by their preamble-body combinations, identifying them by their respective two-letter labels in the first column. The first letter captures the preamble type (M for process or method; P for product; B for product-by-process; and E for empty), the second letter captures the body type (m for process or method; p for product; x for mixed; and e for empty; and '–' when no body is available).

We make a number of observations. First, for claims where both preamble and body "concur" on the type (preamble-body combinations *Mm* for process and *Pp* for product), our classifier achieves an accuracy of 99.9%. Moreover, an empty preamble may be viewed as a strong indication for a product preamble. In this particular case, we have agreement when the body is a product body (*Ep*). Accuracy in this case is at 99.0%. These three cases of *narrow concurrence* (*Mm*, *Pp*, and *Ep*) make up 85.2% of all claims in the benchmark sample with an accuracy of 99.59%. We consider process preambles strong indicators for the overall type of the claims. All claims with process preambles (combinations *Mm*, *Mp*, *Mx*, *Me*, and *M-*) are process claims. Our concurring claims and all process-preamble claims together make up 94.3% of all claims with an accuracy of 99.61%. Last, by-process preambles are strong indicators for product-by-process claims, except when in combination with a product body (*Bp*). In this latter scenario, the body does not describe a process as *announced* by the

by-process preamble. We therefore consider such claims product claims. Adding by-process preamble claims to our selected claims (*wide concurrence*), we obtain a sample that makes up 94.5% of all claims in our benchmark sample at an accuracy of 99.61%.[59]

Product preambles in combination with process bodies (*Pm*) exhibit a noticeably lower accuracy (3.57%). Their numbers, however, are small, making up only 1.1% of the entire sample. A second area of relatively low accuracy (85.14%) are empty preambles in combination with process bodies (*Em*). Again, their numbers are low, making up only 0.7% of the entire sample. Last, claims with empty preambles and mixed bodies (*Ex*), empty bodies (*Ee*), or no bodies (*E-*) do not reveal sufficient information for classification. From our benchmark sample, we know that 81.1% of these claims are product claims. For the results presented in this paper, however, we will consider them unclassified.[60]

In Table 4 we reported the accuracy and coverage of our benchmark sample for different NBER technology categories. In Figure A.1 we examine the technology-specific accuracy and coverage over the benchmark-sample period. We observe a noticeably lower level of accuracy for claims in chemical patents and computers & communications. The accuracy in chemical patents improves whereas it drops in computers & communications from 100% to below 95% (and to a lesser extent in electrical & electronics. The simple approach (preamble only) performs well in electrical & electronics, mechanical, and other, but does particularly poorly in chemical, computers & communications, and drugs & medical (until the early 2000s).

The differences in the performance of the classifier across different technology categories is the result of differences in claim-drafting patterns across the technologies. We explore these patterns further in the next section.

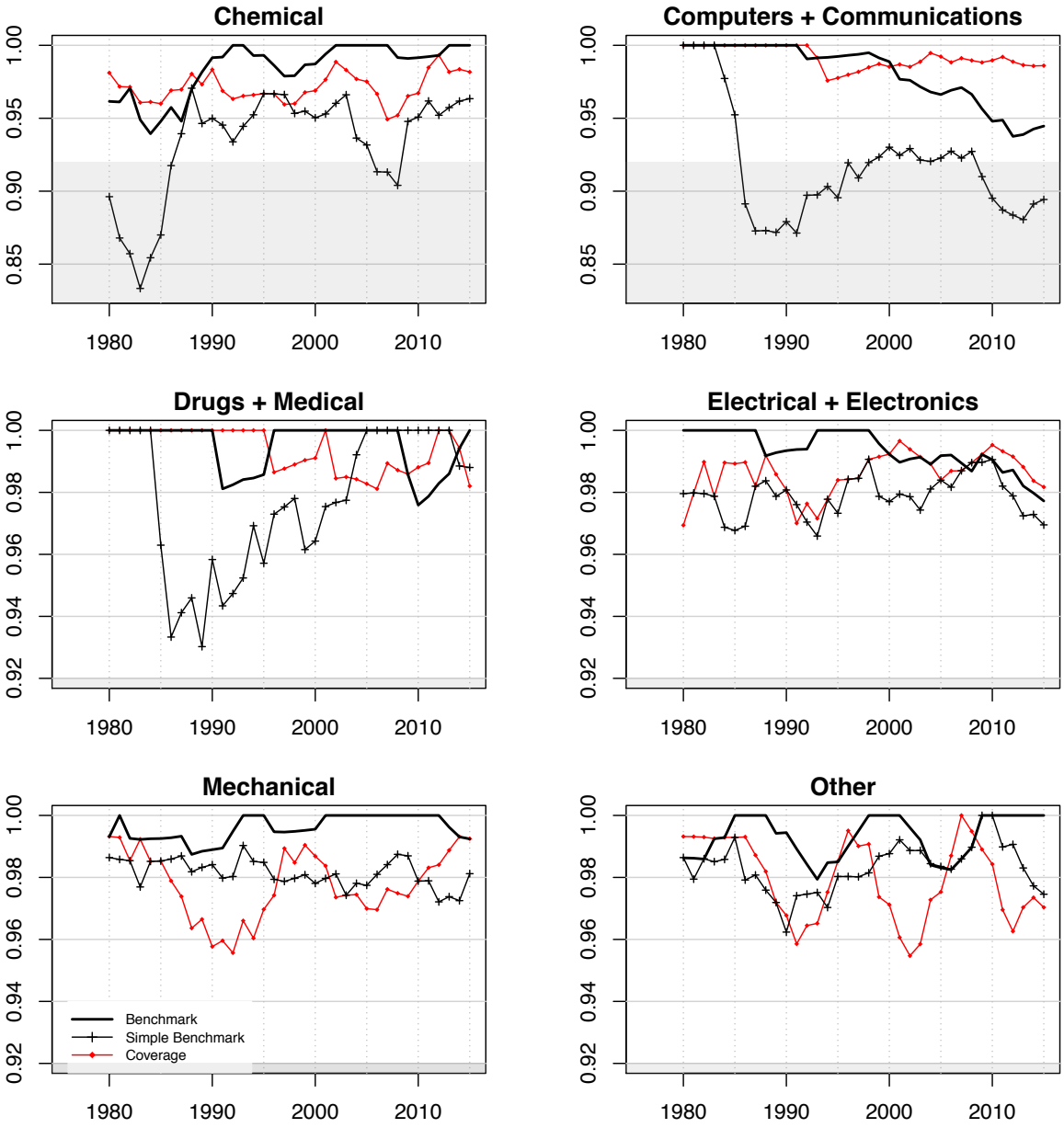
B.2 Implied Accuracy and Coverage in the Full Sample

The accuracy results in the previous section are for the benchmark sample of manually classified claims – claims from patents issued between 1976 and 2015. To get a better picture of the data quality for the full sample (1920 – 2020), we calculate the *implied accuracy*. We take the accuracy for each preamble-combination from Table A.5 and apply it to the claims with the respective preamble-body combinations in the full sample. Under the assumption that for a given combination the accuracy does not change over time (or, rather, is the same for the pre-1976 era as the benchmark sample period), this approach gives us a reliable

[59] In our entire data sample (with all patents issued between 1920 and 2020) these three subsamples make up 82.6%, 90.1%, and 90.2%, respectively. We will present results conditional on these concurrence statuses below.

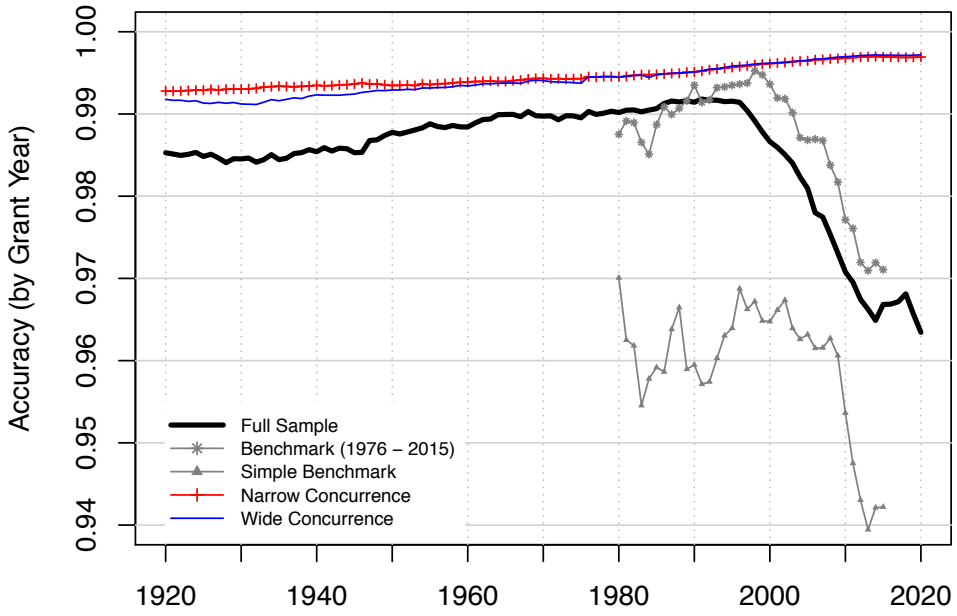
[60] Researchers feeling more optimistic about the classification as product, or for whom full coverage is important, can easily adjust the classification. Our output file includes the labels (and therefore preamble-body combination) used in Table A.5.

Figure A.1: Accuracy for the Benchmark Sample



Notes: We plot 5-year moving averages of the accuracy of our classification for the benchmark sample (black line) for different technology categories (NBER categories), following Hall, Jaffe and Trajtenberg (2001). We also plot the accuracy of the simple approach (preamble only) (thin black line), and the coverage of the full approach (red line). *Data source: patcat and* `https://patentsview.org`.

Figure A.2: Implied Accuracy for the Full Sample



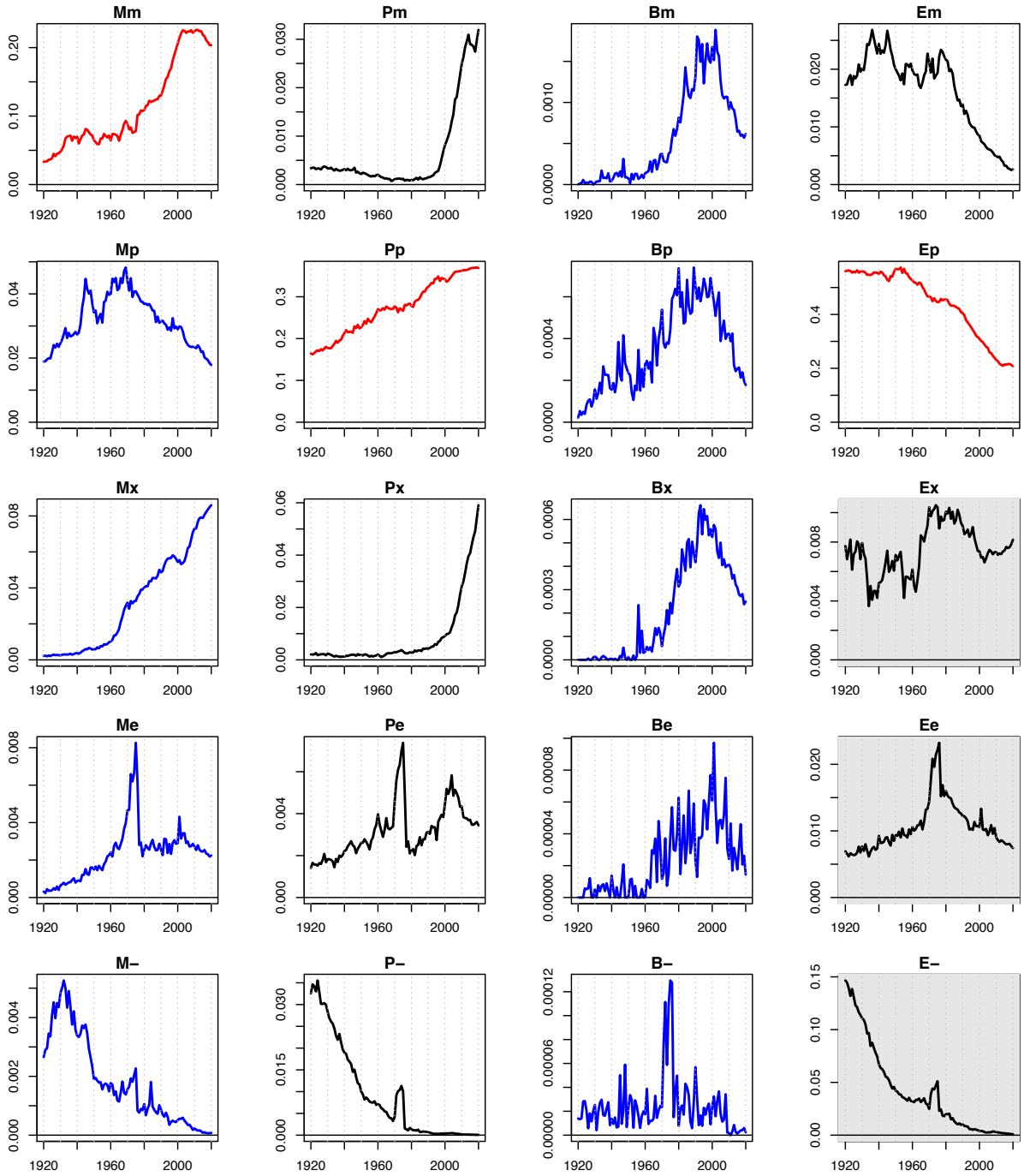
Notes: We plot the implied accuracy for the full sample (black solid line), the narrow-concurrence claims (red line), and the wide-concurrence claims (blue line) for the years 1920 through 2020 (by grant year). The grey lines depict the actual accuracies for the benchmark sample (full classification and simple approach). *Data source: patccat.*

picture of the overall accuracy of our classifier and data, taking changes in patent-claim drafting and changes in preamble-body combinations into account.

We present the results of our implied accuracy calculations in Figure A.2. The red and blue lines depict the results for the narrow-concurrence and wide-concurrence samples, respectively. For both sub-samples, accuracy is steadily increasing over time, exceeding 99.5% for the last three decades in the sample period. The solid black line depicts the accuracy for all claims in the full sample. Here we see an increase from 98.5% to about 99.2% in 1996. Over the following 20 years, the implied accuracy drops to about 96.5%. We see a corresponding decline of accuracy in the benchmark sample for both the full approach and the simple approach (with a delay). Claims in patents related to computers & communications and electrical & electronics are responsible for this drop in accuracy. We saw a decline in accuracy in these NBER technology categories in Figure A.1.

In Figure A.3, we examine the source of these patterns. We plot the distribution of each preamble-body combination – their respective annual shares over all claims in a given year over time. The graphs in red are the combinations for the narrow-concurrence sub-sample, the graphs in blue represent the additional combinations for the wide-concurrence

Figure A.3: Distribution of Preamble-Body Combinations



Notes: We plot the annual shares of preamble-body combinations over all claims in a given year, for the years 1920 through 2020 (by grant year). The graphs in red are the combinations for the narrow-concurrence sub-sample, the graphs in blue the additional combinations for the wide-concurrence sub-sample. Grey-shaded panels indicate preamble-body combinations for which a claim is not classified. *Data source: patccat.*

sub-sample. Panels with a grey-shaded background are with combinations for which the classifier does not assign a claim type.

Preamble-body combinations with low accuracy (per Table A.5) and whose relative frequency increases are responsible for a drop in accuracy. Claims with a product preamble and a method body (combination *Pm*) have a low accuracy (3.57%) and exhibit a stark relative increase in numbers starting around 1995, mirroring the decrease in overall accuracy as depicted in Figure A.2.[61] We find this increase most pronounced in patents related to computer hardware & software (NBER sub-category 22), information storage (24), and electronic business method & software (25) – and to a lesser extent in communications (21), computer peripherals (23), and semiconductor devices (46).

In Figure A.4, we plot the coverage of our classifier for the full sample and the two concurrence sub-samples. While we see a steady increase of coverage for the full sample of claims (except for the years 1971–1975, see below), both the narrow-concurrence (red) and wide-concurrence (blue) sample exhibit declining coverage starting in the mid 1990s. The reason for this decline in coverage is related to the decline in accuracy. As we observe in Figure A.3, combinations *Pm* and *Px* that are not part of the sub-samples gain in weight whereas the relative numbers of combinations *Mm* and *Ep* (both in the narrow-concurrence sample) decrease.

The dip in coverage in the years 1971 through 1975 also appears in Figure A.3. We see spikes in the time series of preamble-body combinations with empty bodies or no bodies, particularly for *P-*, *Ee*, and *E-*. The latter two are not classified, and increases in their shares will result in a decrease in coverage across all samples.

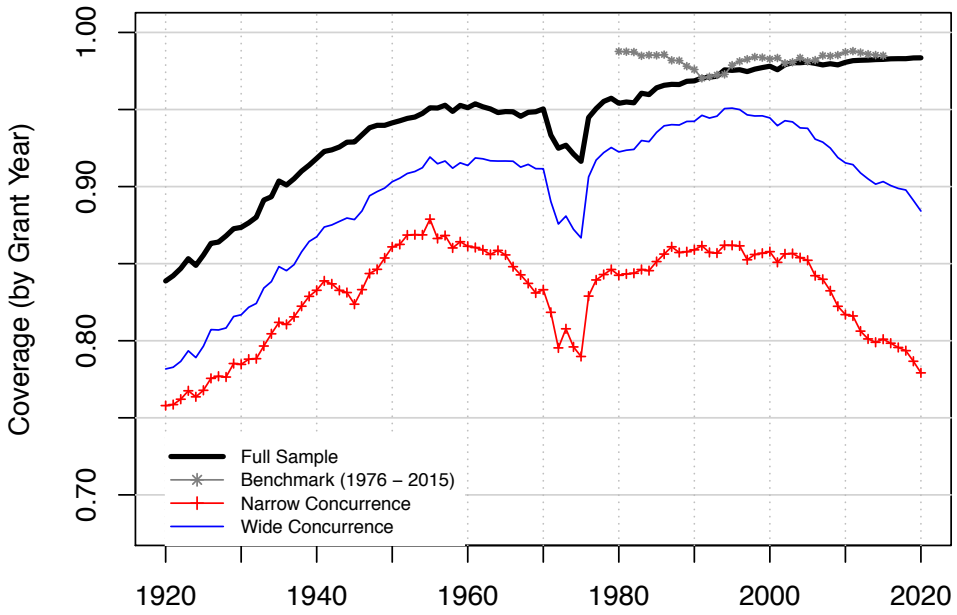
The source of the coverage patterns in the 1970s are single-line claims for which our conversion to a multi-line format (with a proper preamble-body structure) fails (or is limited to a single-line body). All our claims prior to 1976 come in single-line format, and the conversion outcomes have improved between 1920 and 1970. In Figure A.5 we plot preamble types (left) and body (types) over time. In 1920, 20% of all claims had no body. This number has decreased since, with the exception between 1971 and 1975.

C Data in the `patccat` Database (Zenodo)

The data files for utility patents granted between 1836 and 2020 are available for download at Zenodo.org: `https://doi.org/10.5281/zenodo.6395308`. In Table A.6, we provide the list of variables (and short descriptions) for the data files with patent-level information.

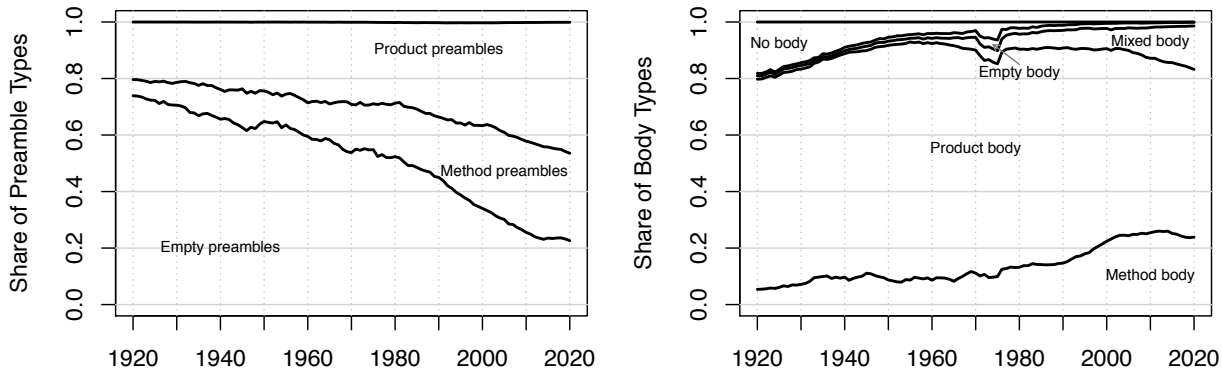
[61] We see a similar increase for claims *Px*. Such claims, however, have an accuracy of 96.13% and given their low numbers cannot be responsible for the drop in overall accuracy.

Figure A.4: Coverage



Notes: We plot the coverage for the full sample (black line), the narrow-concurrence sample (red), and the wide-concurrence sample (blue) for the full sample period (by grant year). For comparison, we also plot the coverage of our benchmark sample for the years 1976 through 2015 (by grant year). *Data source: patccat.*

Figure A.5: Preamble and Body Types



Notes: We plot the share of preamble types (left panel) and body types (right panel) for the full sample period (by grant year). We do not include by-process preambles as their numbers are negligible for the purpose of this graph. *Data source: patccat.*

The data file with information on patents granted between 1836 and 1919 contains 1,038,041 observations; the file with patents granted between 1920 and 2020 contains 9,102,807 observations. in Table A.7, we provide the list of variables (and short descriptions) for the data files with claim-level information (for all independent claims). The file with patents granted between 1836 and 1919 contains 4,324,148 independent claims; the file for patents from 1920 through 2020 contains 27,585,398 independent claims.

Table A.6: Patent-Level Information

Variable Name	Description	Values
`patent_id`	USPTO patent number	string
`claims`	Number of independent claims; the sum of `processClaims`, `productClaims`, `prodByProcessClaims`, and `noCategory`	integer
`noCategory`	Number of independent claims without a claim type (`claimType`= 0)	integer
`processClaims`	Number of process claims (`claimType`= 1)	integer
`productClaims`	Number of product claims (`claimType`= 2)	integer
`prodByProcessClaims`	Number of product-by-process claims (`claimType`= 3)	integer
`firstClaim`	`claimType` of the first independent claim of the patent	integer
`simpleProcessClaims`	Number of process claims by simple approach (`processSimple` = 1)	integer
`simpleProcessPreamble`	Number of process claims by simple approach, preamble only (`processPreamble` = 1)	integer
`meansClaims`	Number of means-plus-function claims	integer
`meansFirst`	Is first claim a means-plus-function claim?	0 = no; 1 = yes
`JepsonClaims`	Number of Jepson claims	integer
`JepsonFirst`	Is first claim a Jepson claim?	0 = no; 1 = yes

Table A.7: Claim-Level Information

Variable Name	Description	Values
PatentClaim	Patent-claim identifier of the form [patent number]-[claim number]	string
singleLine	Is claim in the input data in single-line format?	0 = no; 1 = yes
singleReformat	Format of the claim after conversation of function -fn.singlesplitter-	0 = multi-line claim (original); 1 = multi-line claim (converted); 2 = two-line claim (preamble and single-line body); 3 = single-line claim (not converted)
Jepson	Is claim a Jepson claim?	0 = no; 1 = yes
JepsonReformat	Format of the claim after conversion of the function -fn.jepsonreformat-	0 = not Jepson claim; 1 = Jepson claim (converted)
inBegin	Does claim begin with an "in" phrase?	0 = no; 1 = yes
wordsPreamble	Length of the text of the preamble (in words); length of claim if single-line format	integer
wordsBody	Length of the text of the body (in words); no value if claim is single-line format and body does not exist (or not converted)	integer
dependentClaims	Number of dependent claims following (and referrig to) the independent claim	integer
isMeansPreamble	Does preamble use a means-plus-function phrase?	0 = no; 1 = yes
isMeansBody	Does body use a means-plus-function-phrase?	0 = no; 1 = yes
isMeans	Is claim a means-plus-function claim?	0 = no; 1 = yes
processPreamble	Does preamble use terms "method" or "claim" (simple classifier)?	0 = no; 1 = yes
processBody	Does body use terms "method" or "claim" (simple classifier)?	0 = no; 1 = yes
processSimple	Does claim use terms "method" or "claim" either in the preamble or the body (simple classifier)?	0 = no; 1 = yes
claimType	Invention type of the claim	0 = no category; 1 = process; 2 = product; 3 = product-by-process
preambleType	Preamble type	0 = empty; 1 = process; 2 = product; 3 = by-process
preambleTerm	Keyword used to classify the preamble as product or process preamble	string
preambleTermAlt	Keyword used for the preamble classification if type changes from process to product or vice versa	string
preambleTextStub	First (approximately 15) words of the preamble	string
bodyType	Body type	0 = mixed; 1 = process; 2 = product; 3 = empty
bodyLinesStep	Number of step lines in the body	integer
bodyLinesElement	Number of element lines in the body	integer
bodyLinesTotal	Number of lines in the body	integer
label	Preamble-body type combination	string

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de