NBER WORKING PAPER SERIES

WORDS IN PATENTS:
RESEARCH INPUTS AND THE VALUE OF INNOVATIVENESS IN INVENTION

Mikko Packalen
Jay Bhattacharya

Working Paper 18494
http://www.nber.org/papers/w18494

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2012

The authors thank Darius Lakdawalla, Dana Goldman, Alan Garber, Richard Freeman, John Ham, Josh Graff Zivin, David Blau, Joel Blit, Subhra Saha, Tom Philipson, Neeraj Sood, Pierre Azoulay, Grant Miller, Jeremy Goldhaber-Fiebert, and Gerald Marschke for their comments on early drafts of this paper and for their encouragement. We also thank seminar participants at the Harvard Business School, Stanford School of Medicine, University of Guelph, and especially Bruce Weinberg's working group on innovation and science at the NBER for excellent feedback. Dr. Bhattacharya's work on this paper was partially funded by the National Institute on Aging. Despite all this help, the authors are responsible for all errors in the paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Words in Patents: Research Inputs and the Value of Innovativeness in Invention
Mikko Packalen and Jay Bhattacharya
NBER Working Paper No. 18494
October 2012
JEL No. I1,O31,O32,O33

ABSTRACT

Intelligently allocating research effort and funds requires deciding whether to build on recent advances or on more established knowledge. When recent advances create superior opportunities for invention, their adoption as research inputs in the invention process promotes technological progress. The gains from pursuing such innovative research paths may, however, be very limited, due to the undeveloped nature of new knowledge, quick obsolescence of fast-improving knowledge, and the vast scope of the existing knowledge base. In this paper, we first develop a new approach to identifying research inputs in invention. Next, we estimate the value of pursuing innovative research paths that are created by the arrival of new research inputs. We identify research inputs based on a natural language analysis of 10 billion word and word sequence patent pairs in 6 million patents granted during 1920-2010. This novel textual analysis empirically reveals which single and general purpose technologies and scientific discoveries have been popular as research inputs in invention. We estimate the value of innovative research by comparing patents that mention these research inputs early against the value of other patents. For this comparison, we develop also a new measure of patent value. The measure distinguishes between citations that reflect the cumulative nature of invention and citations that may merely reflect similarity.

Mikko Packalen
University of Waterloo
Department of Economics
200 University Avenue West
Waterloo, ON N2L 3G1
Canada
packalen@uwaterloo.ca

Jay Bhattacharya
117 Encina Commons
Center for Primary Care
and Outcomes Research
Stanford University
Stanford, CA 94305-6019
and NBER
jay@stanford.edu

1 Introduction

Anyone involved in research must choose whether to build their work on recent advances or rely on more established knowledge. This is a choice faced by scientists and inventors as well as private and public financiers of research, such as pharmaceutical firms and the National Institutes of Health (NIH). When recent advances create superior opportunities for invention, innovative research that pursues those new opportunities promotes technological progress. Many of the potential benefits of such innovative research may, however, never be realized as risk aversion, the principal-agent problem, limited rationality, and entrenched interests may bias researchers, firms, and research agencies against innovative research paths (e.g. Kuhn 1962; March 1991; Ahuja and Lambert 2001).

Yet, favoring less innovative research directions is not necessarily foolish, as the private and social benefits of innovative research may be quite limited. When the knowledge base in recombinant innovation is already expansive enough, new advances that add to it have little impact on what can be achieved with invention (Weitzman, 1998). Knowledge created by recent advances may also be initially too shallow, or organizations may lack the appropriate complementary capabilities, for the advances to spur useful invention (e.g. Nerkar, 2003). In addition, knowledge about the properties of recent advances may be initially progress so fast that inventions building on it soon become obsolete. That the benefits of innovative research may be small is not a mere theoretical possibility. There exists both anecdotal and quantitative evidence (Utterback, 1996; Fleming, 2001) suggesting that knowledge needs to mature and deepen before it becomes most useful in spurring subsequent inventions.

The benefits of innovative research may thus be large, small, or even non-existent. Quantitative evidence on the relative benefits of different research directions can guide research decisions and policy. In this paper, we develop a new approach to identifying pieces of knowledge that are recombined in the invention process–*research inputs*–and estimate the benefits of innovative research that builds on recent advances.

We measure research inputs and innovativeness based on text in patents. We first index the words and 2- and 3-word sequences that appear in 80+ years of US patents. We refer to these words and word sequences (e.g. *microprocessor*, *polymerase chain reaction*) as concepts. For each concept, we then determine its year of first appearance and track its subsequent mentions. This textual analysis is important in itself. It reveals which single and general purpose technologies and scientific discoveries have been the most popular as research inputs in invention. Having indexed text in patents, we construct a measure of innovativeness for each patent based on whether the patent includes an early mention of a concept that later becomes popular.

We measure each invention's value from received patent citations. Because some citations reflect inventions that are merely similar to the cited invention, we develop a new measure that reflects only inventions that build upon the cited invention. The new measure is the count of citations received from patents that advance technologies that are distinct from the technologies advanced by the cited patent. In technical terms, the new measure is the count of citations received from patents which novel components have been assigned by patent examiners to technology categories that do not overlap with any of the technology categories assigned to the novel components of the cited invention.

To examine the benefits of innovativeness in technological innovation we compare citations for innovative patents against citations for other patents. Patents that build on any popular new concept are compared against other patents granted in the same technology class in the same year.

Our approach reveals whether and to what extent the pursuit of the best innovative research directions results in inventions that are better than the average invention. If innovative research actually turns out to lead to valuable patents, this fact would provide inventors and organizations who pursue and fund innovative research a quantitative rationale for their research strategy. A finding of no relationship would indicate that the value

of early research opportunities created by the arrival of new research inputs is quite limited, perhaps due to one or more of the aforementioned potential mechanisms. Our estimates on the benefits of innovativeness are thus an important input to discussions about whether extent inventors and research organizations should pursue and fund innovative ideas rather than projects that recombine only well-known existing ideas.

Our analysis also contributes to a better understanding of the drivers of technological progress by identifying concepts that have served as popular research inputs in technological innovation. While any advance may spur invention, these popular research inputs are the most likely drivers of technological progress.

We contribute to several strands of literature. Our estimates of the benefits of innovativeness add to the heretofore sparse evidence–anecdotal (Utterback, 1996) and quantitative (Fleming, 2001; Ahuja and Lampert, 2001; Nerkar, 2003; Schoenmakers and Duysters, 2010)–on the benefits of innovative research.[1] By advancing the measurement of the knowledge base in invention and the measurement of how it evolves, our analysis complements the recombination theory of invention (e.g. Usher, 1922, Schumpeter, 1939, Weitzman, 1998). While new knowledge has a central role in this dominant theory of invention, there is little systematic evidence on what new knowledge and matter is recombined in technological innovation and on how important new knowledge is as a research input.

Our analysis also advances the methods for measuring innovativeness from text (Evans, 2011; Grodal and Thoma, 2009; Azoulay et al., 2011; Bhattacharya and Packalen, 2011).[2]

[1]Fleming (2001) uses subclass information in patents to measure the mean age of an invention's components and relates that mean age to the total number of forward citations. Ahuja and Lambert (2001), Nerkar (2003), and Schoenmakers and Duysters (2010) relate the mean age of cited patents to the total number of forward citations. In comparison, we rely on text to measure research inputs, which yields a more detailed and a more easily interpretable list of research inputs, and our measure of patent value distinguishes between citations that may merely reflect similarity and citations that reflect the cumulative nature of invention.

[2]Evans (2010) indexes 400,000 appearances of 28,000 terms in 18,000 scientific articles related to the Arabidobsis plant to measure the exploratiVeness of publications based on the use of new words or word combinations, finding that industry ties lead to more explorative research. Azoulay et al. (2011) index mentions of 25,000 expert-assigned MeSH keywords in 26,000 publications by 465 scientists to calculate measures of novelty based on the mean age of keywords and the overlap between pre- and post-award keywords, finding that rewarding long-term success encourages more explorative research. Grodal and Thoma

Compared with these existing textual approaches, we analyze considerably more text and concepts, yielding a more comprehensive account of new advances and their timing. Existing analyses have also relied on predefined word lists, whereas we index all available words. Each textual approach, including ours, has its limitations, and we consider the approaches complementary.[3]

Our measurement of single and general purpose technologies advances the literature aimed at identifying general purpose technologies and their effects on technological progress (e.g. Bresnahan, 2011). We also contribute to the empirical literature on how science benefits technological innovation (e.g. Fleming and Sorenson, 2004; Grodal and Thoma, 2009) by examining the relationship between innovativeness and the use of science.[4] We measure the use of science based on citations in patents to the scientific literature.

An additional but important contribution of our analysis is our novel measure of patent value. While it is well-known that patent citations may reflect similarity rather than the cumulative nature of invention, to our knowledge no previous study has developed a measure to address this issue. We also contribute by organizing and examining patent-level data for 1920-2010. To our knowledge existing large-scale patent-level analyses have focused on the post-1975 time period.

The balance of the paper proceeds in a familiar order–methods, data, analysis, conclusion.

(2009) index keywords in scientific articles to construct a list of 130,000 bio- and nanotechnology words and track their 240,000 appearances in 1,500 patents and 2,800 press releases, finding that scientific concepts that arise at the two fields' intersection are more likely to appear in patented and commercialized inventions. Bhattacharya and Packalen (2011) index the mentions of 1,800 FDA-approved active ingredients in 16 million biomedical publications to determine research inputs and the quality of the associated research opportunities.

[3]A limitation of keyword lists (used in Grodal and Thoma, 2009, and Azoulay et al., 2011) is the small number of keywords included in each publication. Author-specified keywords (used in Grodal and Thoma, 2009) may include or exclude keywords for strategic reasons but authors may also have little incentive to consider which keywords are appropriate. Vocabularies and predefined lists (such as the list of FDA approved ingredients used in Bhattacharya and Packalen, 2011) concern only certain types of advances and their scope is thus limited. Keywords based on a vocabulary (such as the MeSH vocabulary used in Azoulay et al., 2011) may not include all important advances or may include advances only after a considerable lag. Expert-curated word lists (used in Evans, 2010) are necessarily biased toward the types of knowledge that conform to the experts' training and beliefs about what types of advances the research builds upon.

[4]Fleming and Sorenson (2004) shows that science facilitates recombination of unfamiliar combinations. Grodal and Thoma (2009) examine the transfer of concepts from science to nanotechnology.

2 Methods

We first propose a new way to identify research inputs in technological innovation. The cumulative nature of invention and the view of invention as a recombination process both suggest that research inputs are an important driver of technological progress. We then present our approach to measuring the value of inventions based on patent citations. This approach distinguishes between citations that may merely reflect similarity and citations that reflect the cumulative nature of invention. Finally, we explain how we construct the measure of innovativeness based on the identified research inputs and how we estimate the value of pursuing innovative research directions in technological innovation.

2.1 Identifying Research Inputs from Text in Patents

By design, patents distribute information about advancements in knowledge. Each patent describes an invention and, in the process, reveals the components of the invention and some of the knowledge and matter that served as research inputs in the invention process that led to the invention.

Existing analyses of research inputs have taken advantage of the information that is revealed by patent subclasses and citations. Patent subclasses are a subjective delineation of the components of an invention. A patent examiner assigns each patent to one or more technology classes and subclasses based on what the examiner perceives to be the components of the invention. Fleming (2001) uses this subclass information to examine what knowledge and matter is recombined in each invention.[5] Citations in patents reveal additional information about what knowledge was used as inputs in the invention process (e.g. Caballero and Jaffe, 1993; Popp, 2002), but this citation information comes with certain caveats that we discuss below in Section 2.2 and footnote 23. Both of these existing approaches to measuring

[5]In related work, Alexopoulos (2011) uses classification information for technical books to examine the extent of innovation across technologies.

research inputs have their advantages. We pursue a distinct, complementary, approach.

We measure inputs to the invention process directly from the patent text by indexing all words and all 2- and 3-word sequences in each patent. Many words and word sequences represent important prior inventions and scientific discoveries. This is especially true for popular new words and word sequences that first appear in patents well into the time period covered by our patent sample. Even a non-expert recognizes many of these words and word sequences as representing knowledge that has driven technological change. Because so many of the popular new words and word sequences in patents describe important prior inventions and scientific discoveries, we believe that our textual approach reveals important components of inventions and, more broadly, important *research inputs*–pieces of knowledge that were recombined in the invention process that led to the patented invention.[6]

There is, of course, some noise in patent texts. Not all words in patents, and not even all new words that appear in patents, represent knowledge that are either components of the invention or relevant research inputs. However, the purpose of the patent text is to describe the invention, rather than describe other inventions. Thus, there does not appear to exist much reason for inventors to include word sequences that do not reflect the components of the invention.[7]

This textual approach complements the subclass and citation based approaches to measuring the knowledge that served as the basis of an invention. A non-expert will often find it easier to understand the meaning of popular words and word sequences that appear in patents than subclass names or titles of cited patents or scientific references, as the subclass names and citation titles are often narrow and technical. Also the noise in patent-to-patent

[6]Some less important innovations are also named, so not all new concepts represent important advances, and some are either re-named or first named only after proven valuable (e.g. drugs). It is also possible that a new concept is an output of a patent, as opposed to an input. However, this property does not drive our results. For a given concept, the number of such patents is at most one, whereas the number of patents we consider innovative because of the specific concept is generally much higher. Moreover, our results are robust to reassigning for each concept the innovative patent with the most citations as not innovative.

[7]However, similar to the incentive to not include competitors' patents among cited patents (Lampe, 2012), there may be an incentive not to include certain words.

citations (e.g. Alcacer et al., 2009; Lampe, 2012), the fact that patent-to-patent citations can reflect similarity rather than the cumulative nature of invention (see Section 2.2 and footnote 23), the sparsity of patent-to-science citations especially in older data, and the fact that only patented inventions can be reflected in citations to previous patents, increase the value of using textual analysis to complement a citation based approach to identifying research inputs.

The patent data we index spans 80+ years of patents. For each patent, the indexed text includes the title, abstract, body, and claims. The indexed text does not include text in citation fields (newer patents), nor the text in the end-of-patent reference section (older patents). By *word* we mean a character sequence that is separated from other character sequences with whitespace. Before indexing the data, we replace various special characters with the space character and erase other special characters. We do not index words and word sequences which length falls outside certain limits, words that include certain special characters and words that do not include any alphabet characters, words that reflect changes in the presentation of patents rather than changes in the nature of inventive activity, and certain very common words and word sequences. Please see the Data Appendix for details.

We interchangeably refer to the research inputs revealed by indexing the words and word sequences in patents as *concepts*. To determine when each concept was a new research input, we determine the year in which the concept was first mentioned in some patent. We refer to this year of arrival as the concept's *cohort*.[8]

In determining the cohort of each concept and the timing of inventive activity in general, we rely on grant years of patents. While the application year of a patent represents the timing of inventive activity better than its grant year, application year is not always unambiguous and readily available. For newer patents (1976-2010) the data list multiple application years

[8]There are potential benefits from using other approaches to determining the cohort of each concept, as patents have typos and the older patent data (1920-1975) include many errors due to the nature of the optical character recognition (OCR) method used by the USPTO in extracting the data from the original patents. A more finely-tuned approach, based on the 11th mention for instance, is left for future research.

when a patent is a continuation patent of one or more previous applications. For older patents (1920-1975) the application year must be extracted from OCR text, which less than state of the art quality prevented us from extracting an application year for all older patents. Thus, in practice the advantages of using application rather than grant years are more limited.

Having organized concepts by cohort, we construct a popularity ranking of concepts in each cohort based on the number of patents that mention each concept. For each concept we also construct a simple measure of whether the concept is an important general purpose technology ("GPT"). This measure is based on the concept's ranking in each of 6 technology categories.[9] For each concept, we first determine in how many technology categories the concept is a top 1-10 concept in its cohort and in how many technology categories the concept is a top 11-100 concept in its cohort. A raw GPT score of a concept is then calculated by adding together the number of categories where the concept is a top 1-10 concept in its cohort and 0.5 times the number categories where the concept is a top 11-100 concept in its cohort. Concepts for which this raw GPT score is 4 or higher are listed with the text "GPT+" to signify that the concept is important in multiple technology categories.

In Section 4 we list the most popular concepts of each cohort. We also list by decade the 40 concepts with the highest GPT scores among concepts that first appered in that decade. These lists of popular new concepts reveal which new research inputs have been important in technological innovation over time. To compare this approach with citation based approaches, we also list the most cited patents and scientific references in patents. We also examine how the total number of new concepts in each cohort has evolved and how their subsequent mentions in patents are distributed. A particular focus is then placed on the top 10,000 concepts in each cohort as our measures of innovativeness of each patent are constructed based on mentions of these most popular concepts.

[9]Technology category specific rankings are available upon request. In assigning patents to technology categories, we rely on the Hall et al. (2001) mapping of 3-digit technology classes to the following 6 technology categories (number of classes in parentheses): 1. Chemical (80), 2. Computers & Communications (48), 3. Drugs & Medical (15), 4. Electrical & Electronics (58), 5. Mechanical (118), and 6. Others (125).

2.2 Measuring the Value of an Invention

We measure an invention's value from the citations the patent has received from other patents. Received patent citations are a commonly used measure of patent value (e.g. Harhoff et al., 1999, Hall et al., 2005) as well as knowledge flows (e.g. Jaffe et al., 1993).

Citations serve as a useful measure of an invention's value to the extent that they reflect the cumulative nature of invention. However, while citations disclose relevant prior art which the citing inventions build upon, the main purpose of patent citations is to delimit the scope of the patent by indicating which parts of the citing invention are not novel and therefore not covered by the patent (e.g. Jaffe et al., 1993; Strumsky et al., 2010). A citation may thus merely indicate that the citing and cited inventions are similar, or that some of their components are similar, in the sense that the inventions or some of their components are near one another in the technology space. Consequently, a patent may receive many citations not because other inventions either rely on or improve upon the cited invention but because the citing patents cover inventions or components that are similar to the cited invention or some of its components (e.g. Jaffe et al., 2002).

The concern that citations may merely reflect the similarity of inventions potentially weakens the case for using citations to measure an invention's value. We address this concern by measuring patent value by the number of citations for which the novel parts of the citing invention are not anywhere near the novel parts of the cited invention in the technology space. Such citations will likely only reflect the cumulative nature of invention, whereas others may reflect mere similarity.[10]

In this approach, we first determine how close the citing and cited patents' novel parts are in the technology space. A delineation of the technologies advanced by each invention is revealed by technology codes. Claims in a patent specify the novel parts of the invention, and the primary and multiple secondary technology classification codes assigned to the patent

[10] At the very least, this type of citations should be much more likely than other citations to reflect the reliance of the citing invention on a technology covered by the cited patent.

delineate what types of technologies are covered by the claims (Strumsky et al., 2010; U.S. Patent and Trademark Office, 2005).[11] We infer from the technology codes of each citing and cited patent pair whether the novel parts of the citing invention are anywhere near the novel parts of the cited invention in the technology space. The technology space is specified by patent examiners who maintain the classification and assign the codes to patents.

At the 3-digit level, the classification system used in patents has over 400 technology classes. Different 3-digit codes may cover closely related technologies. A citing invention may thus be near a cited invention even when the two do not share a 3-digit technology code. Consequently, borders within this technology space are better determined based on the mapping of the 3-digit technology classes to 6 technology categories (as well as 37 sub-categories) developed by Hall et al. (2001). Claims in a pair of citing and citing patents are unlikely to cover similar technologies when the technology categories spanned by the 3-digit codes of the citing and cited patent do not overlap.

In our novel approach to measuring patent value, we thus first determine for each patent the primary and all secondary 3-digit technology codes assigned to the invention. Next, we determine for each patent which technology categories are spanned by these technology classes.[12] We then calculate for each patent the number of citations received from patents which technology categories do not overlap with any of the technology categories of the cited patent. We refer to the count of such received citations as "No-Overlap Citations"; it is one of our two preferred measures of patent value.

Based on the No-Overlap Citations, we construct our second preferred measure of patent value: an indicator variable that measures whether a patent is among the top 5% most cited patents granted in the same technology class and in the same year.[13] We refer to this measure

[11] Primary and secondary classes are also called as *original* and *cross-reference* classes, respectively.

[12] 26% of patents granted during 1920-2010 have technology codes in multiple categories. Our approach is thus distinct from counting citations for which the technology category of the primary technology class is different for the citing and cited patents.

[13] Singh and Fleming (2010) use top 5% most cited status as a measure of breakthrough invention. Their substantive focus differs from ours and they only consider citation measures constructed from total citations.

as "Top 5% by No-Overlap Citations". We also report results based on the total number of citations and the top 5% most cited status in terms of total citations. We refer to these secondary measures as "Total Citations" and "Top 5% by Total Citations", respectively.

2.3 Measuring Innovativeness and Its Impact on Patent Value

We construct binary measures of innovativeness for each patent. These variables measure whether a patent mentions new concepts that later become popular, capturing which patents are innovative the sense that they take advantage of the early opportunities created by the arrival the best new research inputs.[14] We consider a research input to be new for the first 10 years following its first appearance in a patent (cohort).

In some analyses we employ a single innovativeness measure, a measure that captures whether the patent mentions a concept that is new and among the top 100 most popular concepts in its cohort. In other analyses we employ multiple binary innovativeness measures, with one measure capturing, for example, whether the patent mentions any new top 10 concepts, and another measure capturing, for example, whether the patent mentions any new top 11-20 concepts. We vary the set of popular new concepts considered from the top 10 to the top $10,000$ concepts in each cohort.

To evaluate the benefits of innovative research we compare received citations to patents that mention a new research input against citations to patents that do not mention any such new inputs. Patents in the latter category–the control group–are obviously either patents with no new concepts or patents with less popular new concepts. In these analyses we regress citations on one or multiple binary measures of innovativeness, which are constructed as mentioned above. The specifications with multiple innovativeness measures

[14] Given our focus on popular research inputs, we uncover how useful are innovative inventions that are based on the best new research inputs. Inventors and scientists who are considering pursuing research that relies on a new research input may often have private information about the input's long-term potential. Thus, evidence on the value of inventions that take advantage of the opportunities created by the arrival of the best research inputs can be more important than evidence on the quality of inventions that take advantage of opportunities created by new research inputs in general.

are designed to examine whether concepts in higher ranked concept groups are more potent than concepts in lower ranked concept groups in terms of creating valuable opportunities soon after their arrival.[15] In both sets of analyses, we control for year and technology class effects by comparing innovative patents only to patents that were granted in the same year and in the same technology class (within estimation).

In most analyses we include patent length, measured by the number of characters, as a control variable.[16] We obtain also technology category specific and time period specific estimates. As our dependent variables are count and binary variables, we employ Poisson and logit models in addition to linear regression models.

3 Data and Descriptive Statistics

The data consist of US patent documents granted during 1920-2010. Figure 1.1 shows by grant year the number of patents included in the patent document data. The figure shows also the number of patents listed in the November 2011 version of the USPTO Patent Master File. The Master File lists the patent number and grant year of each granted patent but not the patent documents themselves. The figure shows that with the exception of 1971-1975, the patent document data cover over 99.99% of granted patents.

The Master File also lists the current primary and secondary technology classes assigned to each patent. We use the main technology class (and grant year) of each patent in determining the comparison group for each patent. In constructing our two preferred patent value measures, we use all listed technology classes to determine whether patents in each citing and cited patent pair span overlapping technology categories. Figure 1.2 shows the number

[15] In these specifications, we include for each concept group also a continuous explanatory variable that measures how many additional (above 1) concepts in the group are mentioned in a given patent. This way, the coefficients on the binary innovativeness measures should not be larger for higher ranked concept groups only because a given patent is more likely to mention multiple concepts from a higher ranked concept group than multiple concepts from a lower ranked concept group.

[16] Longer patents are more likely to include *any* concept. Estimates of the impact of innovativeness on patent value are larger when patent length is not included as an explanatory variable.

of patents granted in each of the 6 technology categories during our sample period. Though patenting has increased in every category, the composition of invention across technology categories has changed markedly over the years, particularly in the form of an increased share for Computers & Communications and Drugs & Medical categories. Accordingly, it is important to employ also category specific analyses to distinguish effects driven by changes in the composition of invention from other effects.

For 1976-2010, the patent document data are a machine-readable transfer from the original patents. In these data different fields such as title, abstract, claims, patent-to-patent citations, and patent-to-non-patent-reference citations are clearly indicated. For 1920-1975, the data are an OCR transfer from the original patents. In these data, only patent number and grant year are separately indicated. Elements such as title, application year, claims, and references must be determined by searching the ASCII scan of each patent for markers that reveal the desired information.[17] Please see the Data Appendix for details on our data organization, extraction and disambiguation efforts.

We analyze the textual content in patents to capture research inputs. Figure 2.1 depicts by technology category the median number of non-whitespace characters in patents each year. The figure indicates that patent length has increased over time, but this finding comes with the caveat that the information in the older and newer data are different in terms of data quality and data coverage because the older (pre-1976) data are an OCR scan. Figure 2.2 depicts the median number of unique words in patents each year. Here, "word" refers to *any* character sequence that is separated from others by whitespace; the numbers describe the raw data before the replacement of special characters and other pre-processing that we do for the concept analysis (see the Data Appendix). The drop in the number of unique words in 1976 is indicative of the less than ideal quality of the OCR scan applied to the pre-1976 data. Further descriptive analysis of patent texts is postponed until Section 4.

[17]For example, to determine the title, we search the ASCII scan for capitalized words near the beginning of the scan. The searches are complicated by the less than state of the art nature of the OCR scan.

We index patent-to-patent citations to list the most cited patents and to measure patent value. Patents granted since 1947 include a references section (we do not index in-text citations). Figure 3.1 depicts the mean number of citations by the grant year of the citing patent. Figure 3.2 depicts the mean of Total Citations by the grant year of the cited patent, based on citations in all patents and based on citations in the newer patents. The figures give an indication of how much information is added by extracting also citations in the older data. Figure 3.3 depicts the means of Total Citations and No-Overlap Citations by technology category. The figure shows that patents receive No-Overlap Citations in all technology categories. Figure 3.4 depicts the share of No-Overlap Citations captured by patents with the Top 5% by No-Overlap Citations status as well as the share of Total Citations captured by patents with the Top 5% by Total Citations status. Within each technology category, No-Overlap Citations are even more concentrated than Total Citations.

We index citations in patents to the scientific literature to list the most cited scientific references and to examine the relationship between the use of science and the use of new concepts in the invention process. For non-patent references in the newer data, we discern whether a citation is to a scientific reference and disambiguate the scientific references. For the older data, we only determine whether a non-patent reference is present and use it as a proxy for a scientific reference. Please see the Data Appendix for the details.

Figure 4.1 depicts the number of patents with a non-patent reference and the number of patents that cite a science. Over time it has become more common to cite science in patents but it is unknown to what extent the trend reflects changing citation patterns rather than an increased reliance on science in invention. By comparing patents to other patents granted in the same year (and in the same technology category or class), secular trends in citing behavior should not bias our findings. Figure 4.2 depicts the share of patents that cite science within each technology category. The stark differences across categories highlight the importance of employing category specific analyses in this context.

4 New Research Inputs in Technological Innovation

One important output of our methods is the organic identification of new research inputs based on their actual use, rather than expert judgement or citations. In this section, we describe some basic information about the new concepts that are identified by the approach.

4.1 Top 20 Concepts in Each Cohort

Table 1 lists the top 20 most popular new concepts in each decade from 1920s to 2000s. For this descriptive summary table, new concepts are grouped by the decade of their cohort. Popularity of each concept is determined based on the number of patents that mention the concept, among patents granted during 1920-2010. The colored squares affixed to each concept name indicate the technology category with the most patents that mention the concept. With a handful of exceptions, these assigned technology categories remain the same when the mapping is instead based on how the first 100 mentions of the concept are distributed across technology categories.

Two additional tables in the Appendix present more detailed information on the identity of top new concepts and their use. Table A1 lists information for the top 20 most popular new concepts in each cohort from 1921 to 2010. Table A2 lists for each decade the 40 new concepts with the highest GPT scores.

Results in Table 1 and in Tables A1 and A2 show that the textual approach identifies many concepts which even a casual observer recognizes as representing single and general purpose inventions and scientific discoveries that have served as important research inputs in technological innovation. The presence of many important multi-word concepts in these tables shows that indexing multi-word concepts in addition to single-word concepts is valuable. Our approach is informative across different time periods, with the possible exception of the last 5 years or so. Even the top concept lists for the early (1920s and 1930s) cohorts appear to be informative despite the fact that only the data for patents granted in 1920 was

used to discern which words form the base vocabulary that does not reflect new knowledge. Due to OCR errors in the older patent data, the assigned cohort of some concepts listed in Table 1 and in Tables A1 and A2 is not the true year in which they were first mentioned in patents (e.g. *laser*, *internet*) and some important concepts are altogether missing from these tables because they were assigned the cohort 1920 (e.g. *email*). We do not explicitly address these concerns, because they are to a significant degree artifacts of the less than state of the art nature of the OCR scan.

The assigned technology categories in Table 1 demonstrate several patterns, which are supported by the more detailed results in Tables A1 and A2 (column 7 in these tables lists the assigned technology category for each concept). While concepts mapping to the Computers & Communications category have long been important, they have come to dominate the list of most popular concepts in recent decades. This is indicative of the emergence of computers as an important general purpose technology. Concepts mapping to the Drugs & Medical category in turn enjoyed a boom in the 1980s, and have all but disappeared from the top 20 list since. Two technology categories, the Electrical & Electronics category and the Chemical category, appear frequently on the list in the early to mid 20th century and have all but vanished from the list in recent decades. We leave for future research to examine whether the disappearances reflect a stagnation in certain types of innovation, or variation in the speed at which different types of concepts are adopted as research inputs in technological innovation, or something else.

The changes in the extent to which the different technology categories appear in Table 1 and in Tables A1 and A2 are, of course, linked to changes in the share of patents that are granted in each category. To which extent such linkages are driven by changes in the fertility of research inputs, by demand-induced changes in research effort, and by other factors, is also left for future research. A specific motivation for the present analysis is that methods that identify research inputs are themselves an important research input to analyses of such

linkages (see Popp, 2002; Bhattacharya and Packalen, 2011).

Tables A1 and A2 also give an indication of how quickly new concepts are adopted, as the entries in columns 5 and 6 of these tables list the number of patents that mention each concept during years 0-4 and years 5-9 after the concept's arrival. Concepts in post-1960s cohorts have received much more early mentions than concepts in older cohorts, suggesting in particular that the pace at which new advances are adopted as research inputs increased throughout the 1970s, 1980s, and 1990s. We return to this issue in Section 4.4.

4.2 Comparison with Top 20 Patents and Scientific References

To provide an illustration of how the textual approach complements citation based approaches, we now compare research inputs identified from text with research inputs identified from citations. Table A3 in the Appendix lists by grant year the 20 most cited patents for 1920-2010. The list is constructed based on citations in US patents during 1947-2010. Table A4 in the Appendix lists by publication year the 20 most cited scientific references for 1900-2010. The list is constructed based on citations in US patents during 1976-2010.

A comparison of the number of citations that the top patents and scientific references receive (Tables A3 and A4) with the number of mentions that top concepts receive (Table 1 and Tables A1 and A2) shows that top concepts receive orders of magnitudes more mentions than top patents and scientific references receive citations. Therefore, when the intention is to identify important single or general purpose technologies, or to examine their adoption as research inputs, or to examine the impact of specific research inputs on technological innovation, or to identify which scientific discoveries have had the biggest impacts on technological innovation, tracking concepts is likely to be a more fruitful approach than tracking citations.

Comparison of patent and scientific reference titles in Tables A3 and A4 with concept names in Table 1 and in Tables A1 and A2 shows that top concepts capture technologies that are broad enough for their names to be informative even to a casual observer. By contrast,

the titles of the most cited patents and scientific articles are so narrow and technical that they are much harder for a non-expert to understand. Consequently, social scientists examining science or invention will likely often find it much easier to discern the context of findings that are derived using concepts than the context of findings that are derived using citations.

4.3 Frequency of New Concepts

We now broaden the analysis to consider also concepts outside the top 20 concepts in each cohort. Figure 5.1 shows the number of concepts and the number of total mentions for single-word concepts (1-grams) in each cohort. In this figure, concepts are grouped based on whether they are mentioned once, 2 to 10 times, or more than 10 times. The results in the left panel show that across cohorts the vast majority concepts are mentioned only once. The number of concepts that are mentioned 2 to 10 times far exceeds the number of concepts that are mentioned more than 10 times. The results in the right panel show that concepts mentioned more than 10 times still capture a significant share of total concept mentions. Figure 5.2 extends the analysis to multi-word concepts (2- and 3-grams). The results are similar to the results for single-word concepts. Comparison of Figure 5.2 with Figure 5.1 also reveals that the number of concepts and the number of concept mentions are both an order of magnitude greater for multi-word concepts than single-word concepts.

As Figures 5.1 and 5.2 indicate, the number of concepts in each cohort is very large. Focusing on a smaller subset of concepts can thus greatly reduce computational costs. These figures also imply that focusing the analysis to a subset of all concepts, such as concepts mentioned more than 10 times, will still capture a meaningful share of the information contained in text. Another factor against including all concepts in an analysis is that, due to OCR errors and typos, it is hard to disentangle which concepts among the many rarely mentioned concepts contain meaningful information.

Figures 5.3 and 5.4 show the number of concepts and concept mentions for concepts

mentioned more than 10 times. The results show that the number of concepts remains very large also when one focuses the analysis only on concepts mentioned more than 10 times. In part to limit computational cost and analytical complexity, the analysis in the next section (on the impact of innovativeness on patent value) relies on measures of innovativeness that are constructed based on mentions of top 10,000 concepts in each cohort. We next describe how often the top 10,000 concepts in each cohort are used.

4.4 Mentions of Top 10,000 New Concepts

The extent of variation in whether patents mention a top new concept is the key descriptive statistic for the analysis of the impact of innovativeness on patent value. Figure 6.1 shows by grant year the share of patents that mention at least one new concept. We consider a concept new during years 0-9 following its first mention in patents (cohort year). For this figure, concepts are divided into four groups: top 1-10, top 11-100, top 101-1,000, and top 1,001-10,000 concepts in each cohort. Each subfigure demonstrates that there is variation in terms of whether a patent mentions a popular new concept from the concept group, at least when comparisons are conducted at the grant year level. Such variation is present also within each technology category, as is shown by Figure 6.2. The regression analyses in the next section show that there is sufficient identifying variation also when patents are compared only to patents granted in the same technology class in the same year.

Another notable patent-level descriptive statistic is the relationship between the use of new concepts and the use of science. Figure 6.3 shows for each technology category how the use of new concepts in a patent depends on the use of science in the patent. Use of science is determined based on presence of one or more scientific references in the patent. Within each technology category, the share of patents that mention a new concept is greater for patents that cite a scientific article compared to patents that do not cite science. This is quantitative evidence that science plays an important role in the introduction and adoption

of new research inputs in technological innovation.

At the concept level, the two most noteworthy observations about the mentions of the top $10,000$ concepts concern the rate of early adoption and the ratio of early vs. total mentions. Figure 6.4 shows by cohort for the four concept groups the number of patents in which the concepts are mentioned on average when they are new. Figure 6.5 shows the corresponding information by technology category for the top $10,000$ concepts, with concept mentions and cohorts determined separately for each category. These figures suggest that since the 1970s there has been a considerable increase in the pace at which new research inputs are adopted in invention.[18] Figure 6.6 in turn shows for each concept group the ratio of per-year mentions for the concepts when they are new (years 0-9 after the cohort year) and per-year (since cohort) total mentions for the concepts. The low ratios indicate that mentions during years 0-9 are generally but a small share of the total mentions for a concept. Patents that mention a concept during years 0-9 after its cohort thus generally engage in a relatively early use of the concept, which supports considering such patents to be innovative.

5 The Value of Innovative Patents

5.1 Results

We first consider the value of pursuing innovative research that takes advantage of the early opportunities created by the arrival of the top 100 new concepts in each cohort. The measure of innovativeness is a dummy variable that captures whether a patent mentions any new top 100 concept, with concepts considered new during the years 0-9 after their arrival. The results are show in Table 2. Columns 1 and 2 show the results with each of our two preferred measures of patent value as the dependent variable, and Columns 3 and 4 show the results

[18] This finding is present also when the number of mentions for each concept when it is new is normalized by the total number of patents granted during the same period, and when concept rank is determined based on how many patents mention the concept when it is new.

for the other two measures of patent value. The reported estimates are incidence rate ratio estimates (Columns 1 and 3) and odds ratio estimates (Columns 2 and 4). The estimates are above 1 across the four columns, indicating that the best new research inputs create valuable research opportunities soon after their arrival.

A positive relationship between the innovativeness measure and patent value is present in different time periods. Time period specific estimates are shown in the first column of Table 3. Figure 7.1 further illustrates this with grant year specific comparisons of the value of innovative patents against the value of other patents.[19]

The positive relationship between the innovativeness measure and patent value is also robust to employing alternative measures of innovativeness. The second column in Table 3 shows that the relationship remains positive when the most cited patent associated with each concept is reassigned from the innovative group of patents to the control group of patents.[20] The third column in Table 3 shows that the relationship remains positive when the measure of innovativeness is constructed with the assumption that concepts are new during years 0-4 years after their arrival.

The best new concepts create valuable research opportunities soon after their arrival across all 6 technology categories. This is shown by the technology category specific estimates in Table 4. Figure 7.2 further illustrates this finding by depicting grant year specific comparisons of the value of innovative patents against the value of other patents for each technology category, using the approach that corresponds to a linear fixed effects model.

[19]The comparisons in Figure 7.1 correspond to a linear fixed effects model with grant year and technology class pair specific fixed effects. Within each grant year, the observations on the outcome variable are first normalized so that the average value of the variable is the same in all technology classes.

[20]The estimates are obtained after reassigning for each top 100 concept in each cohort one of the patents that mentions the new concept, namely the patent that has received the most number of Non-Overlap Citations relative to the number of Non-Overlap Citations received by the patent's control group, as not innovative (i.e. the indicator variable measuring innovativeness is set to 0 for that patent). This reassignment analysis addresses the concern that the estimates in Table 2 might be driven by citations received by a patent that covers the concept i.e. a patent which the concept in question and for which the concept is a research output as opposed to a research input. The approach is aimed at establishing a lower bound for the patent value impact of using the new concept as a research input; the approach is not meant to suggest that the reassigned patent–or any patent for that matter–necessarily covers the concept in question.

We now consider estimates from empirical models with multiple measures of innovativeness, and also extend the analysis to the top $10,000$ most popular concepts in each cohort. Table 5 depicts estimates for the top 100 new concepts from a specification for which separate innovativeness measures were constructed based on mentions of top 1-10, top 11-20, ..., top 81-90, and top 91-100 new concepts in each cohort. Table 6 depicts estimates from the corresponding specification for the top $10,000$ concepts, with separate innovativeness measures constructed based on mentions of top 1-$1,000$, top $1,001$-$2,000$, ..., top $8,001$-$9,000$ and top $9,001$-$10,000$ new concepts in each cohort

Estimates in Tables 5 and 6 are generally higher for the top two concept groups (i.e. top 1-10 and top 11-20; top $1-1,000$ and top $1,001$-$2,000$) than for lower ranked concept groups. Concepts that are most popular in the long run thus appear to be also the most potent in terms of creating valuable early research opportunities soon after their arrival. At the same time, estimates in Table 6 indicate that even new concepts below the top $2,000$ concepts create valuable research opportunities soon after their arrival. Even the use of new concepts in the top $9,001$-$10,000$ concept group is associated with higher patent value compared to patents that mention no concepts that are new and among the top $10,000$ concepts in their cohort.

To quantitatively assess the relative value of inventions that take advantage of the best new concepts, the last three rows in Tables 5 and 6 show average predicted values of the dependent variable for different patent groups.[21] The last three rows in Table 5 [Table 6] show the average predicted value (1) for patents that mention a new top 10 [top $1,000$] concept, (2) for patents that do not mention a new top 10 concept but do mention a new top 11-100 concept [do not mention a new top $1,000$ but do mention a new top $1,001$-$10,000$

[21] For the predicted values to be as comparable with one another as possible, we construct them as follows. We first estimate the linear fixed effects model corresponding to each column. Next, we calculate for each grant year and technology class pair the average predicted value, with patent length replaced by its median value. We then calculate the average of these average predicted values over such grant year and technology class pairs that have patents in all three patent groups.

concept], and (3) for patents that do not mention any new top 100 concept [top 10,000 concept]. The results for our first preferred patent value measure (Column 1) in Table 6 reveal that patents that mention a new top 1,000 concept are on average over two times more valuable than patents that do not mention any new top 10,000 concept.

5.2 Discussion

The regression results show that inventions that build on the best new research inputs are much more valuable than the average invention. This finding gives inventors and organizations who pursue and fund innovative research a quantitative reason to expect that the early opportunities created by the arrival of the best new research inputs are fruitful. Their pursuit yields inventions that are more valuable than the average invention.

The result indicates that new knowledge often is not too superficial for it to spur useful inventions, somewhat in contrast with the findings in Fleming (2001) which suggested that knowledge may need to mature first. A conclusion in Fleming (2001) is that "Organizations that seek technological breakthroughs should experiment with new combinations, possibly with old components." We interpret our findings as suggesting that organizations should include relatively new components too in the recombinant inventive search process.[22,23] Our

[22]The difference in results may be due to the methods used. We consider research innovative when it includes even one new concept whereas component familiarity in Fleming (2001) measures how familiar is the average component, and we infer components from words whereas Fleming (2001) infers them from subclasses.

[23]Ahuja and Lampert (2001), Nerkar (2003) and Schoenmakers and Duysters (2010) find a positive relation between the mean age of cited patents and the total number of forward citations, suggesting that the use of new knowledge leads to more valuable inventions. Nerkar (2003) and Schoenmakers and Duysters (2010) also relate the age spread of the cited patents to the total number of forward citations; the positive empirical relation suggests that mixing new and mature knowledge is beneficial. One caveat to these analyses arises from the use of cited patents as a measure of research inputs. Because citations can be indicative of either similarity with the cited patents or dependence on the cited patents as building blocks, a patent may cite recent patents even when it only builds on older ideas. This caveat could potentially be addressed by excluding citations for which the cited and citing patents or their components lie in same technology categories, similar to the approach that we developed here for measuring of patent value. However, relying exclusively on this approach to measure research inputs would leave one with research inputs that are very limited in their number and in what the inputs cover. Another caveat to these analyses–a caveat that applies also to Fleming (2001)–stems from the reliance on all forward citations as a measure of patent value.

finding also suggests that knowledge base does not yet appear to be so expansive that additions to it would not significantly enhance the opportunities for recombinant innovation, contrary to the scenario raised in Weitzman (1998), and that the knowledge does not progress so fast that quick obsolescence would render innovative inventions to be of relatively little value.

As existing patent-level analyses have recognized (e.g. Fleming 2001; Fleming and Sorenson 2004; Singh and Fleming 2010), one important caveat to patent-level analyses is that patents do not capture the outcomes of all inventive efforts. Some inventive efforts do not produce any result or produce only a result that does not warrant patenting. Furthermore, not all successful inventive efforts are patented, and some patented inventions are the result of more effort than others.

For the present analysis, this caveat implies that the empirical results measure the impact of innovativeness on the value of invention conditional on inventive efforts leading to a patented invention. Accordingly, while we find that the innovativeness has a large positive impact on the value of patented inventions, innovativeness can still have a negative impact on the value of inventive effort if innovativeness has a sufficiently large negative impact on the probability that the inventive effort leads to a patented innovation. However, as Fleming (2001) notes, the force of this caveat is weakened by the fact that a large share of patented inventions receive no or only a few citations. Given that the bar to patent is so low and many patents cover the outcomes of innovative efforts that were essentially unsuccessful, the use of patents to measure inventive effort does not necessarily suffer from much truncation.[24]

The caveat notwithstanding, our analysis of the value of innovativeness shows that the type of innovativeness considered here matters as it changes the distribution of outcomes and

[24]This caveat could potentially be tackled by tracking patents for an inventor, organization, or field, to measure how innovativeness impacts the extent of invention for an entity or field. We leave such analyses, which come with their own caveats, for future research. A related topic for future research is contrasting multiple forms of innovativeness, such as the type of innovativeness considered here and the type of innovativeness considered in Fleming (2001).

that words in patents are an important predictor of patent value.[25] Moreover, the caveat does not impact the interpretation of our other results (the identity of popular research inputs, the increase in the pace at which new research inputs have been adopted, and the interdependence of innovativeness and the use of science).

6 Conclusion

We have shown that in technological innovation the pursuit of innovative research directions can have benefits: inventions that build on the best new research inputs are much more valuable than the average invention. New knowledge thus does not appear to be too superficial or improve too fast to prevent the knowledge from spurring valued inventions. And, the existing knowledge base does not appear to be so large that additions to it would not significantly enhance opportunities for recombinant invention. Our methodology and findings are an important input to discussions on to what extent researchers and research organizations should pursue and fund innovative vs. more established research directions, a key issue in research resource allocation.

To identify research inputs, we developed and employed a new large-scale text based approach. The textual analysis enabled us to list the single and general purpose inventions and scientific discoveries that have served as important research inputs and drivers of technological innovation during 1920-2010, the period currently covered by digitized patent data. Our approach complements citation based approaches to capturing research inputs. Given the central role of research inputs and the quality of the associated research opportunities as drivers of technological and scientific progress–according to both theory (e.g. Weitzman, 1998) and evidence (e.g. Popp 2002; Bhattacharya and Packalen, 2011)–the new approach

[25]The latter finding–that textual content is a predictor of patent value–complements earlier analyses that have linked made citations in patents to patent value. Text and citations in patents and other research publications thus reveal not only which existing ideas created the opportunities that were pursued in the inventive activity that led to a given research publication but also which existing ideas have been the most useful as research inputs by having served as the basis of the more valuable inventions and discoveries.

is a valuable input to analyses of the drivers and consequences of inventive activity.

We also advanced the methods for measuring patent value. We developed and employed a new citation based measure that distinguishes between citations that may merely reflect similarity and citations that should only reflect the cumulative nature of invention. The new approach addresses a central caveat that applies to analyses of technological innovation that employ existing citation based measures of patent value.

Finally, we showed quantitatively that science plays a role in the introduction and adoption of important new research inputs in technological innovation. The finding complements Fleming and Sorenson (2004) which found that science serves as a map that allows inventors to combine components that have been rarely used together. Both analyses point to mechanisms that researchers should strive to account for when estimating the practical benefits of science.

While in this paper the focus has been on technological invention, the approach can also be applied to study research inputs and innovativeness in science. Extending the analysis to science is a particularly intriguing direction for future research because incentives are different in science and invention (see e.g. Aghion et al., 2008). The direction of invention is largely disciplined by the for-profit motive of firms, whereas scientific researchers generally do not risk failure if they shun innovative ideas to protect the value of their own human capital and past ideas. Entrenched interests can thus exclude innovative ideas with relative ease in science. Hence, the pursuit of innovative research directions may have more limited private benefits in science than in technological innovation. Application of the textual approach to science is unfortunately hindered by the lack of public access to large-scale digitized data on scientific articles. Only for biomedical sciences are data widely available but the data are limited to abstracts (see Bhattacharya and Packalen, 2011). Making complete publication texts widely available for research purposes would open important new avenues for the study of science and its links to technological innovation.

References

Aghion, P., Dewatripont, M. and J. C. Stein, 2008, "Academic Freedom, Private-Sector Focus, and the Process of Innovation," *RAND Journal of Economics*, vol. 39, pp. 617-35.

Ahuja, G. and C. M. Lampert, 2001, "Entrepreneurship in a Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Invention," *Strategic Management Journal*, vol. 22, pp. 521-43.

Alcacer J., Gittelman M. and B. N. Sampat, 2009, "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis," *Research Policy*, vol. 38, pp. 415-27.

Alexopolos, M., 2011, "Read All about It!! What Happens Following a Technology Shock?" *American Economic Review*, vol. 101, pp. 1144-79.

Azoulay, P., Manso, G. and J. Graff Zivin, 2011, "Incentives and Creativity: Evidence from the Academic Life Sciences," *RAND Journal of Economics*, vol. 42, pp. 527-54.

Bhattacharya, J. and M. Packalen, 2011, "Opportunities and Benefits as Determinants of the Direction of Scientific Research," *Journal of Health Economics*, vol. 30, pp. 603-15.

Bresnahan, T., 2011, "General Purpose Technologies," in Hall., B. and N. Rosenberg (eds.) *Handbook of the Economics of Innovation*, North-Holland Elsevier.

Caballero, R. J. and A. Jaffe, 1993, "How High are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth," *NBER Macroeconomics Annual*, vol. 8, pp. 15-83.

Evans, J. A., 2010, "Industry Induces Academic Science to Know Less about More," *American Journal of Sociology*, vol. 116, pp. 389-452.

Fleming, L., 2001, "Recombinant Uncertainty in Technological Search," *Management Science*, vol. 47, pp. 117-32.

Fleming, L. and O. Sorenson, 2004, "Science as a Map in Technological Search," *Management Science*, vol. 25, pp. 909-28.

Grodal, S. and G. Thoma, 2009, "Cross-Pollination in Science and Technology: Concept Mobility in the Nanobiotechnology Field," *Annals of Economics and Statistics*, vol. 93.

Hall, B., Jaffe, A. and M. Trajtenberg, 2001, "The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools," NBER Working Paper No. 8485.

Hall, B. H., Jaffe, A. and M. Trajtenberg, 2005, "Market Value and Patent Citations," *RAND Journal of Economics*, vol. 36, pp. 16-38.

Harhoff, D., Narin, F., Scherer, F. M. and K. Vopel, 1999, "Citation Frequency and the Value of Patented Innovations," *Review of Economics and Statistics*, vol. 81, pp. 511-5.

Jaffe A., Trajtenberg, M. and M. S. Fogarty, 2002, "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Study of Patentees," in Jaffe, A. and M. Trajtenberg (eds.) *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, MIT Press.

Jaffe, A., Trajtenberg, M. and R. Henderson, 1993, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, vol. 108, pp. 577-98.

Kuhn, D., 1962, *The Structure of Scientific Revolutions*. University of Chigaco Press.

Lampe, R., 2012, "Strategic Citation," *Review of Economics and Statistics*, vol. 94, pp. 320-33.

March, J. G., 1991, "Exploration and Exploitation in Organizational Learning," *Organizational Science*, vol. 2, pp. 71-87.

Nerkar, A., 2003, "Old Is Gold? The Value of Temporal Exploration in the Creation of New Knowledge," *Management Science*, vol. 49, pp. 211-29.

Popp., D., 2002, "Induced Innovation and Energy Prices," *American Economic Review*, vol. 92, pp. 160-80.

Schoenmakers, W. and G. Duysters, 2010, "The Technological Origins of Radical Inventions," *Research Policy*, vol. 39, pp. 1051-9.

Schumpeter, J., 1939, *Business Cycles*. McGraw-Hill: New York.

Singh, J. and L. Fleming, 2010, "Lone Inventors as Sources of Breakthroughs: Myth or Reality?" *Management Science*, vol. 56, pp. 41-56.

Strumsky, D., Lobo, J. and S. van der Leeuw, 2010, "Using Patent Technology Codes to Study Technological Change," Santa Fe Institute Working Paper 10-11-028.

Utterback, J. M., 1996, *Mastering the Dynamics of Innovation*, Harvard Business School Press.

Usher, A. P., 1922, *History of Mechanical Inventions*, McGraw-Hill Book Company, New York.

U.S. Patent and Trademark Office, 2005, *Handbook of Classification*.

Weitzman, M., 1998, "Recombinant Growth," *Quarterly Journal of Economics*, vol. 113, pp. 331-60.

Figures

Figure 1.1: Patents in the Master File vs. Patents in the Document Data.

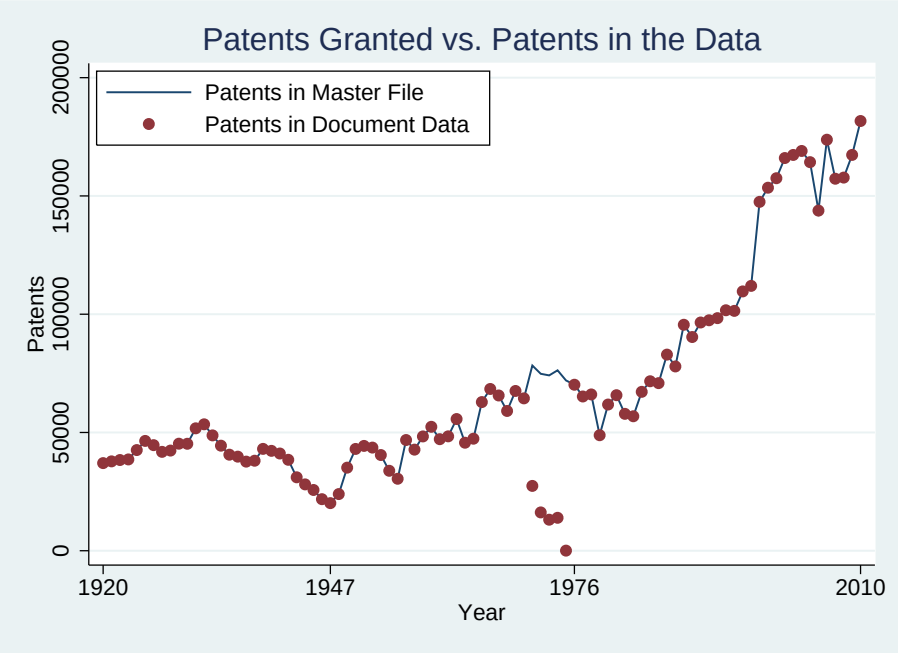


Figure 1.2: Patents by Technology Category.

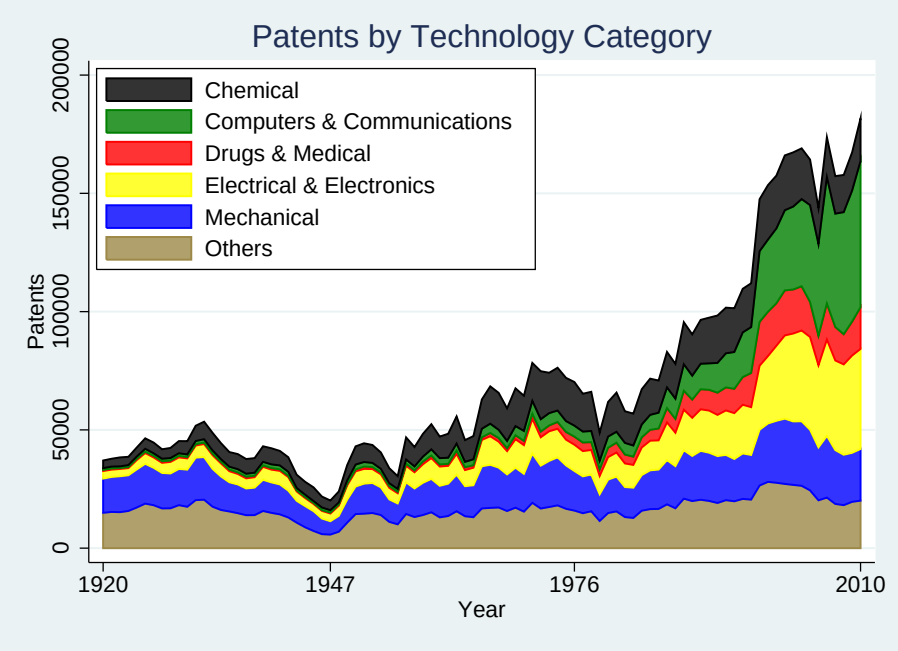


Figure 2.1: Number of Characters in Patents.

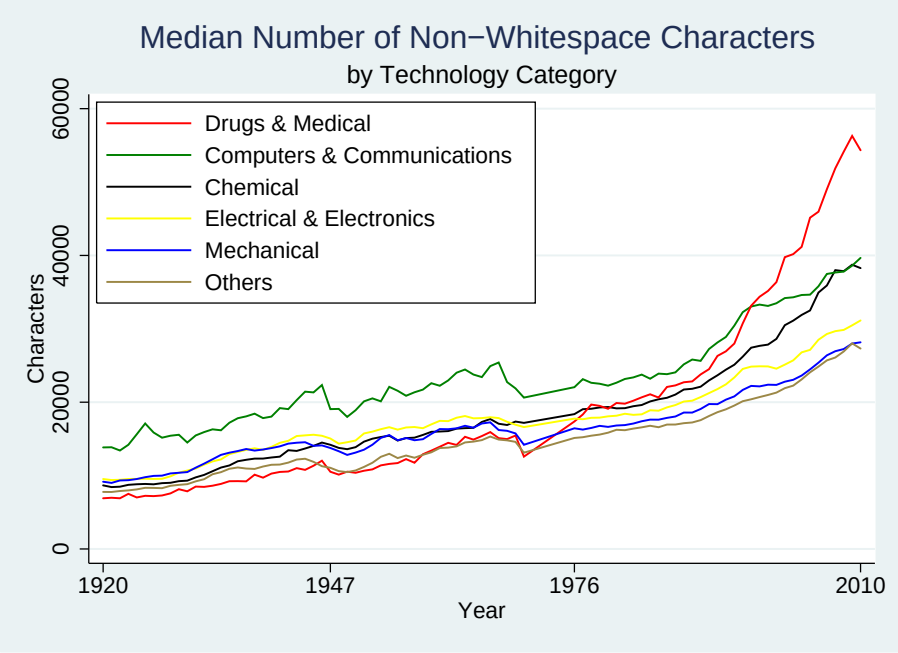


Figure 2.2: Number of Words in Patents.

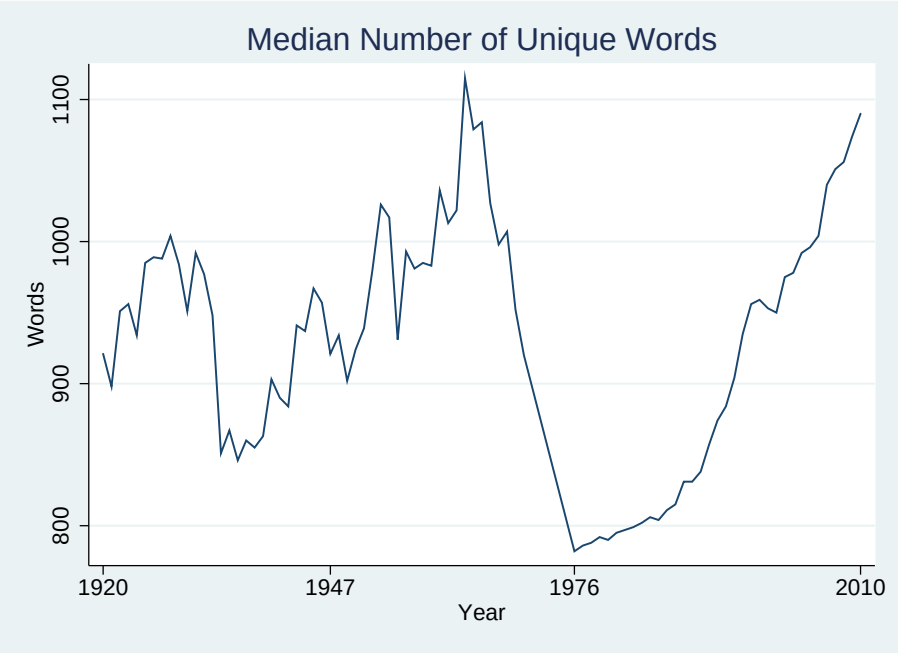


Figure 3.1: Citations to Patents: In Newer and in Older Data.

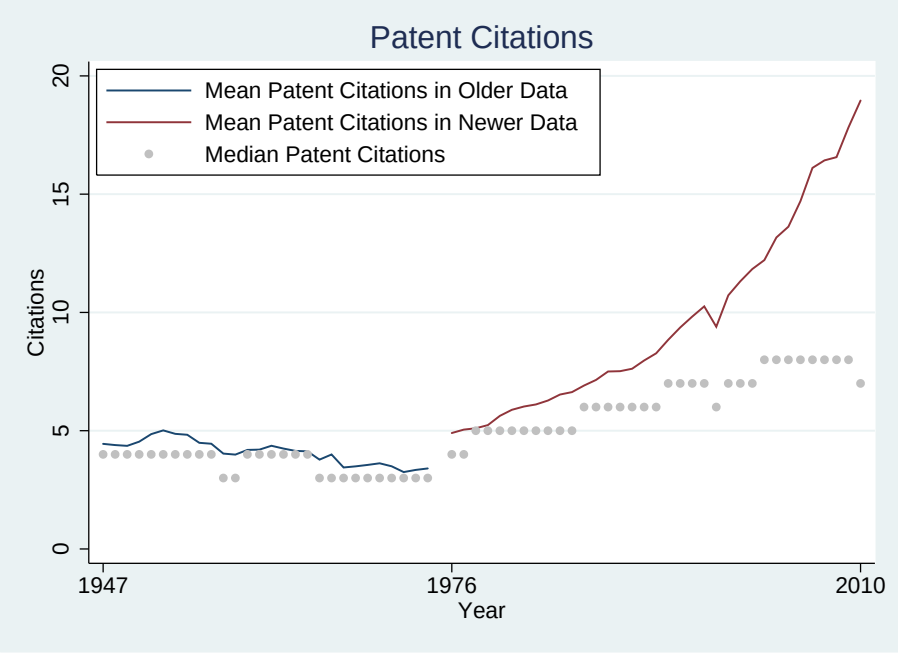


Figure 3.2: Received Citations Based on All Data vs. Newer Data Only.

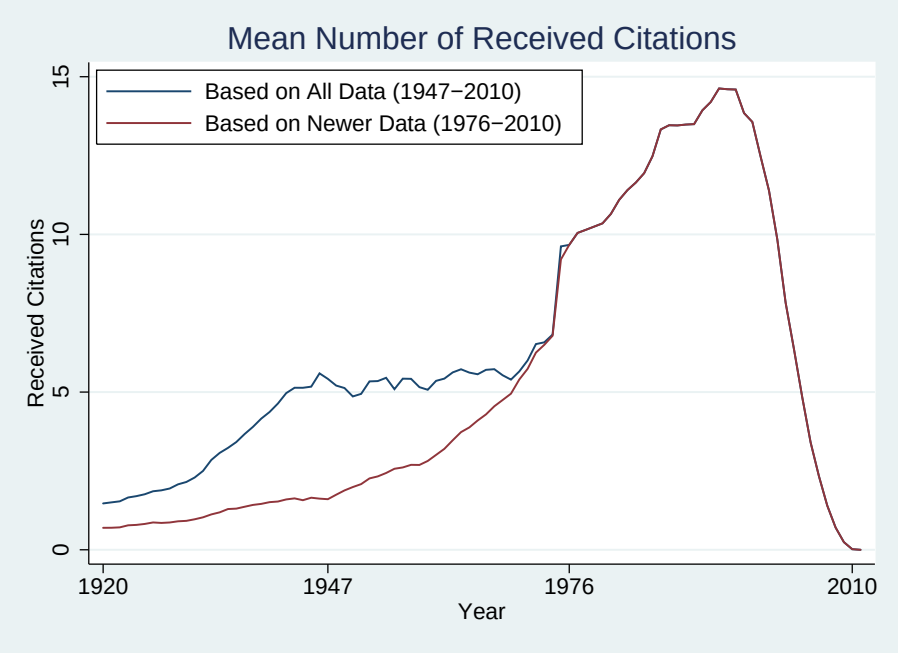
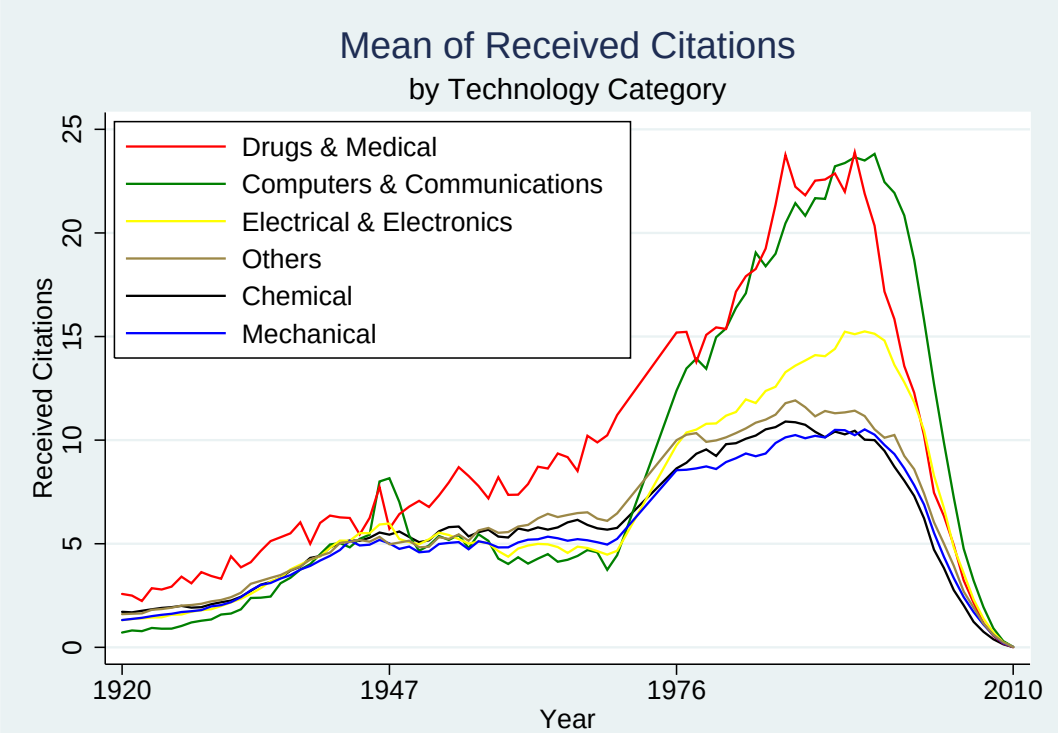


Figure 3.3: Received Total and No-Overlap Citations, by Technology Category.



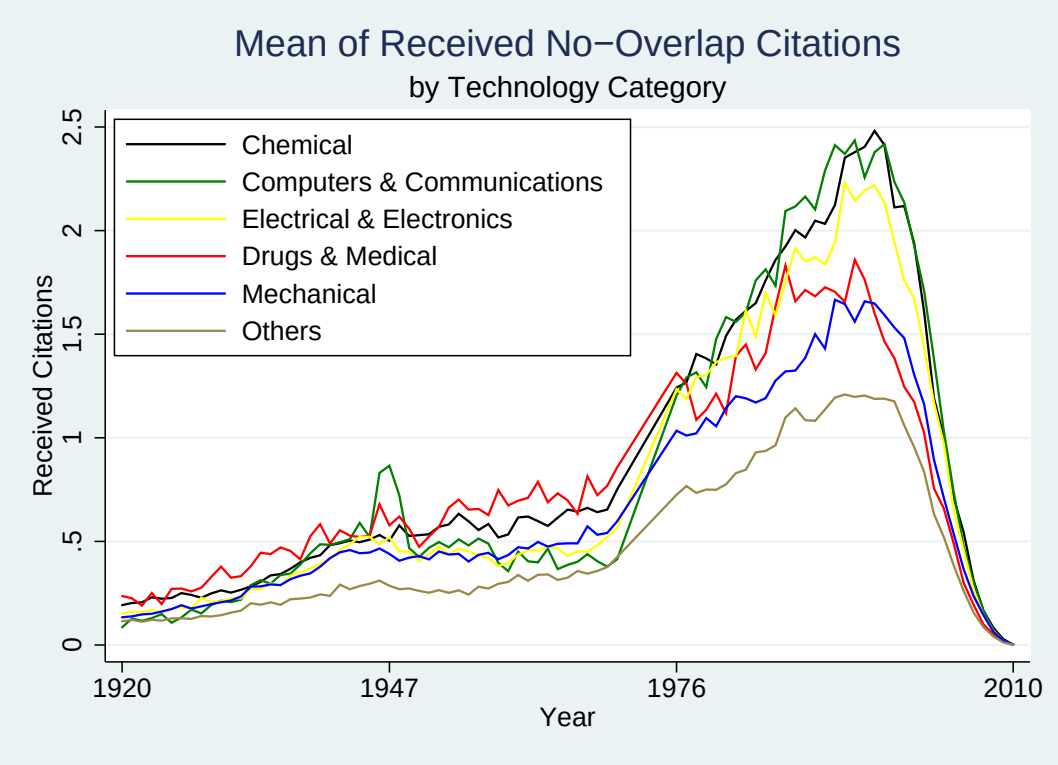


Figure 3.4: Share of Citations Received by Patents with Top 5% Status, by Technology Category.

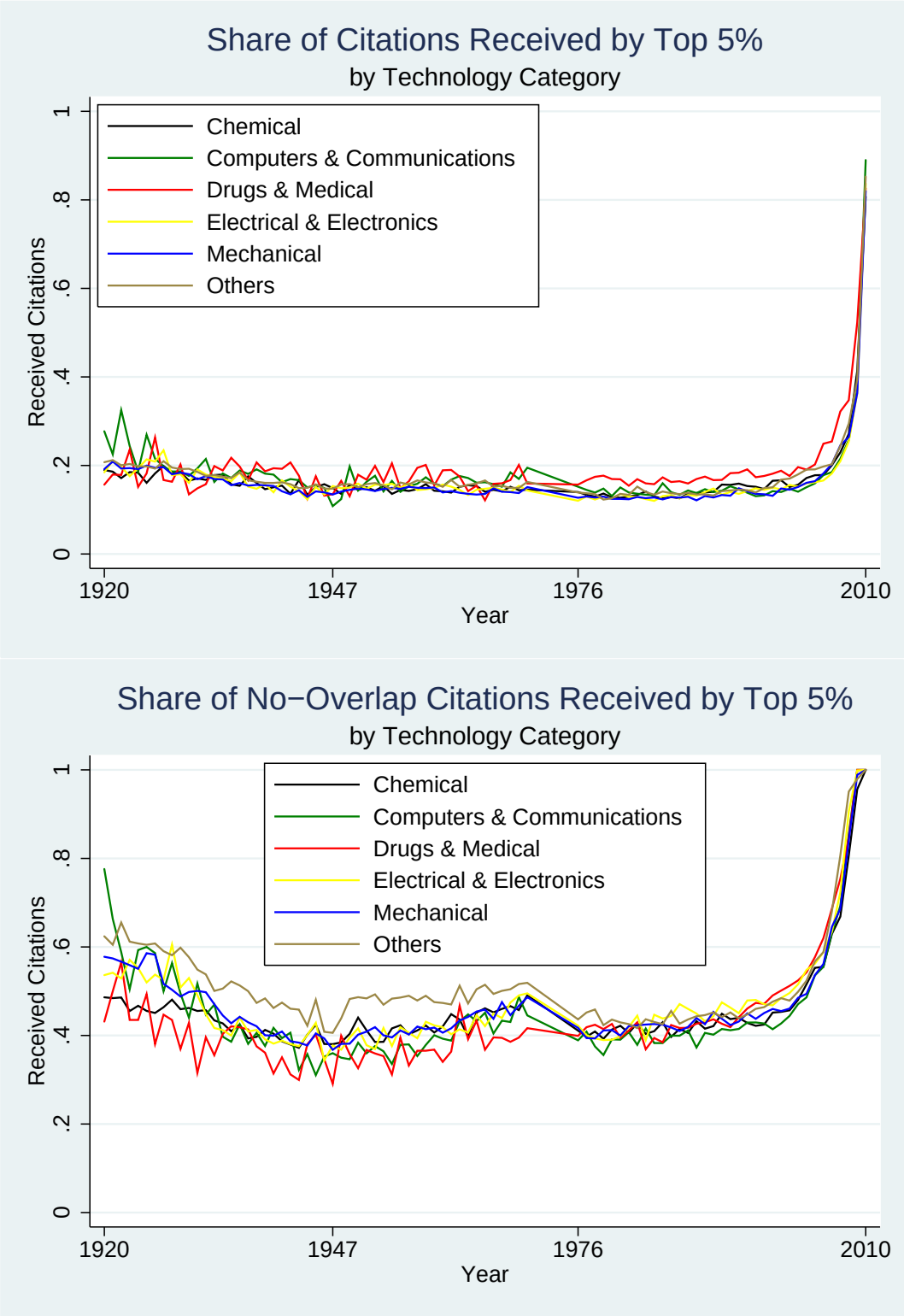


Figure 4.1: Patents with a Scientific Reference and Patents with any Non-Patent Reference.

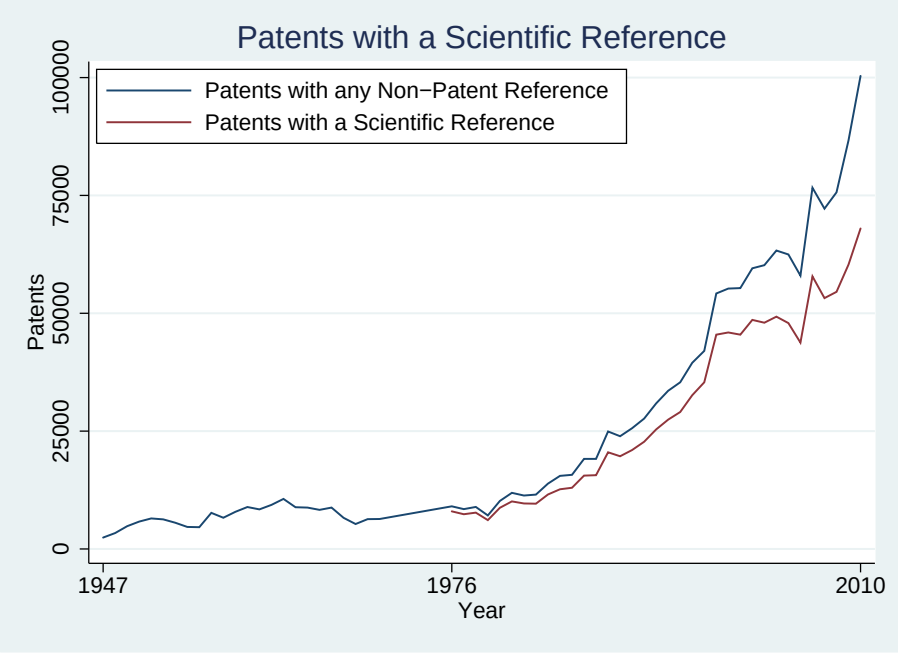


Figure 4.2: Share of Patents that Cite Science, by Technology Category.

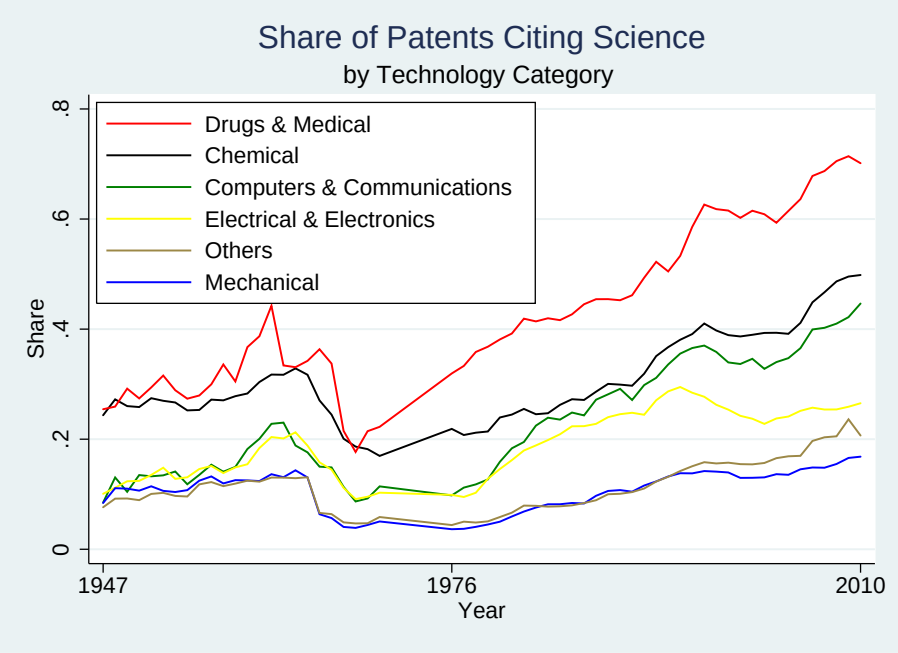


Figure 5.1: Number and Mentions of New Concepts: 1-grams.

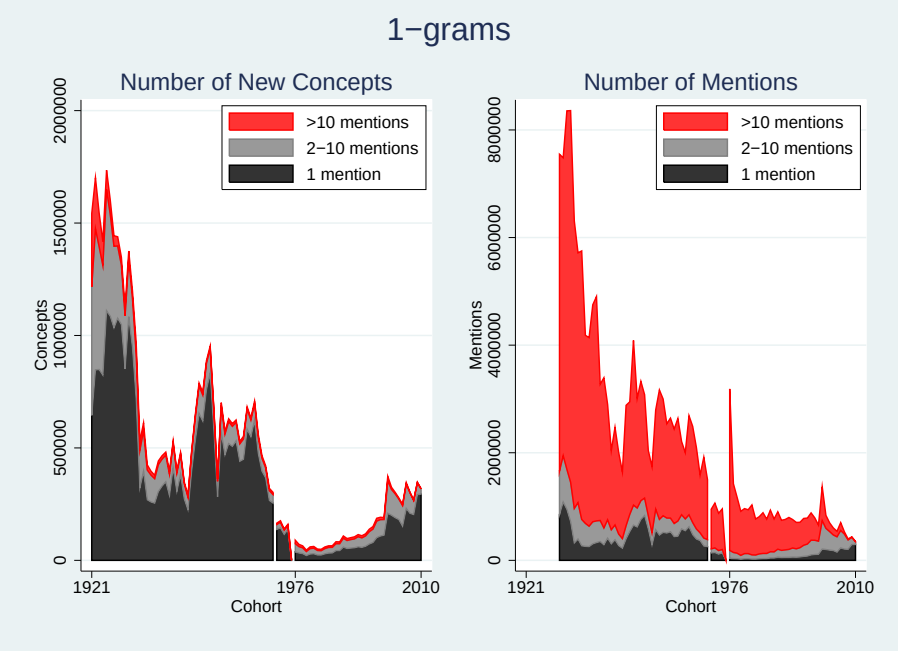


Figure 5.2: Number and Mentions of New Concepts: 2- and 3-grams.

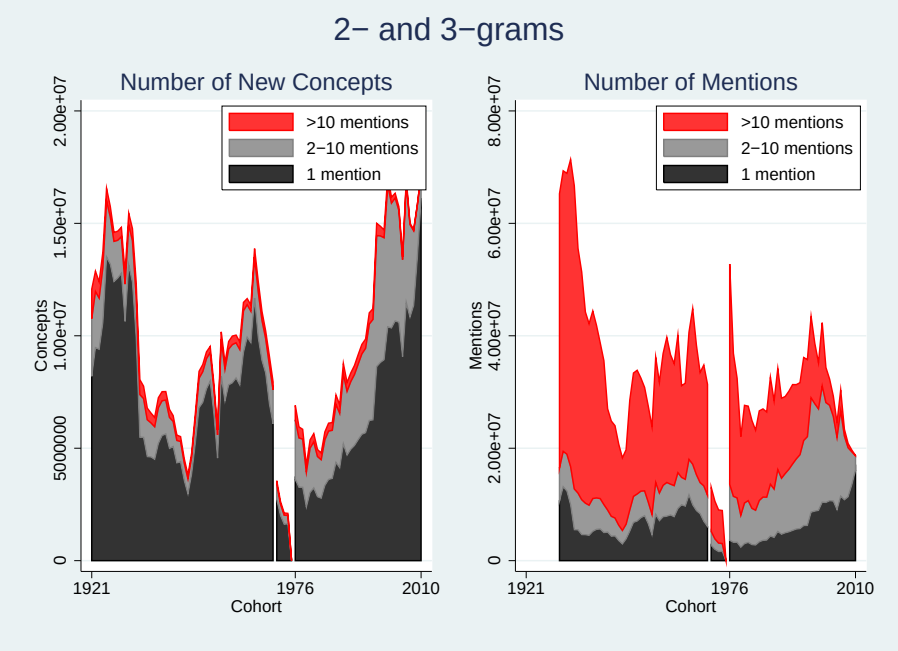


Figure 5.3: Number and Mentions of New Concepts: 1-grams with more than 10 Mentions.

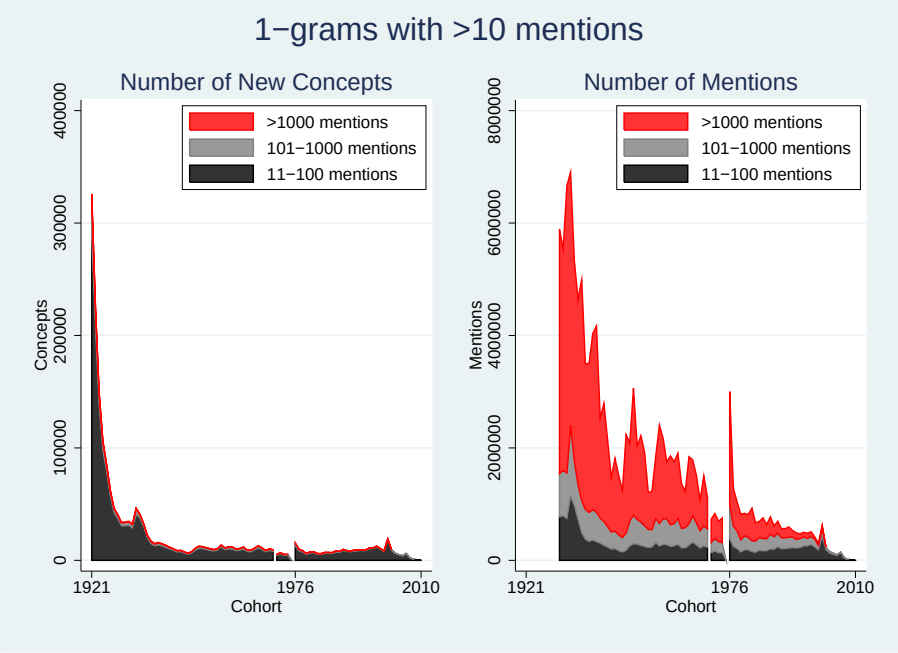


Figure 5.4: Number and Mentions of New Concepts: 2- and 3-grams with more than 10 Mentions

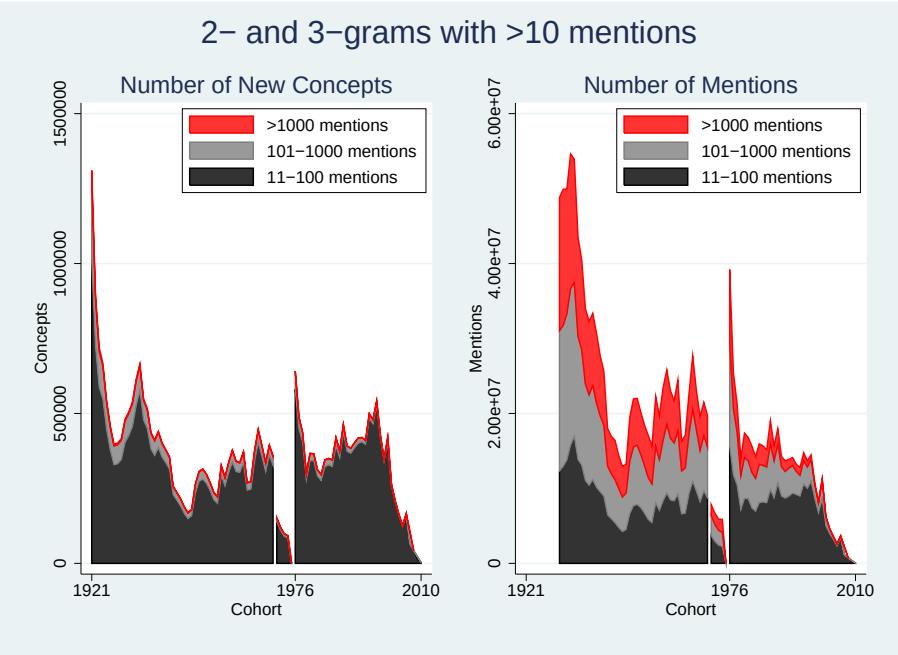


Figure 6.1: Share of Patents that Mention a New Concept, by Concept Group.

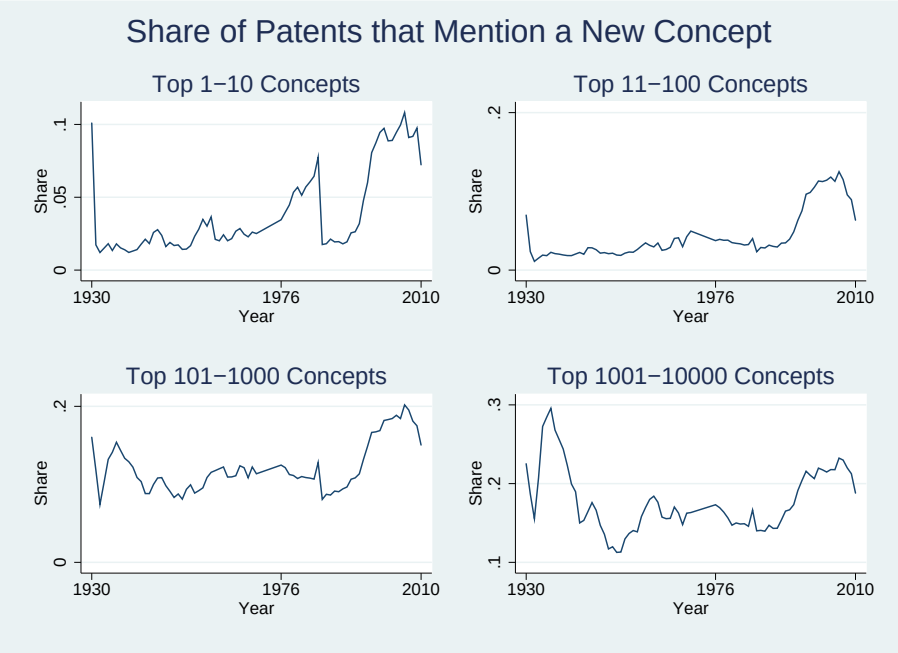


Figure 6.2: Share of Patents that Mention a New Top 10000 Concept, by Technology Category.

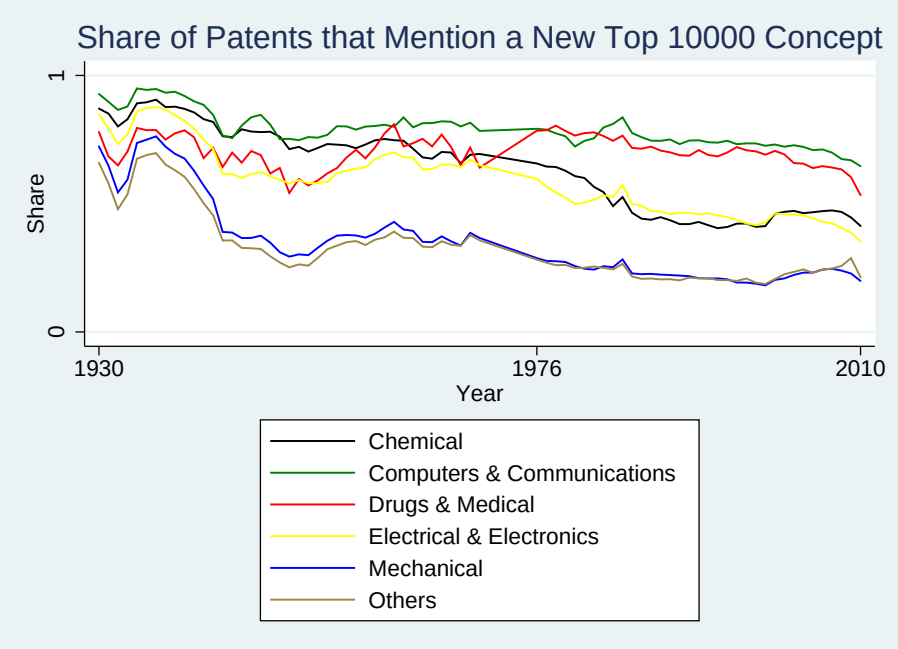


Figure 6.3: Share of Patents that Mention a New Top 10000 Concept, by Science-Citing Status.

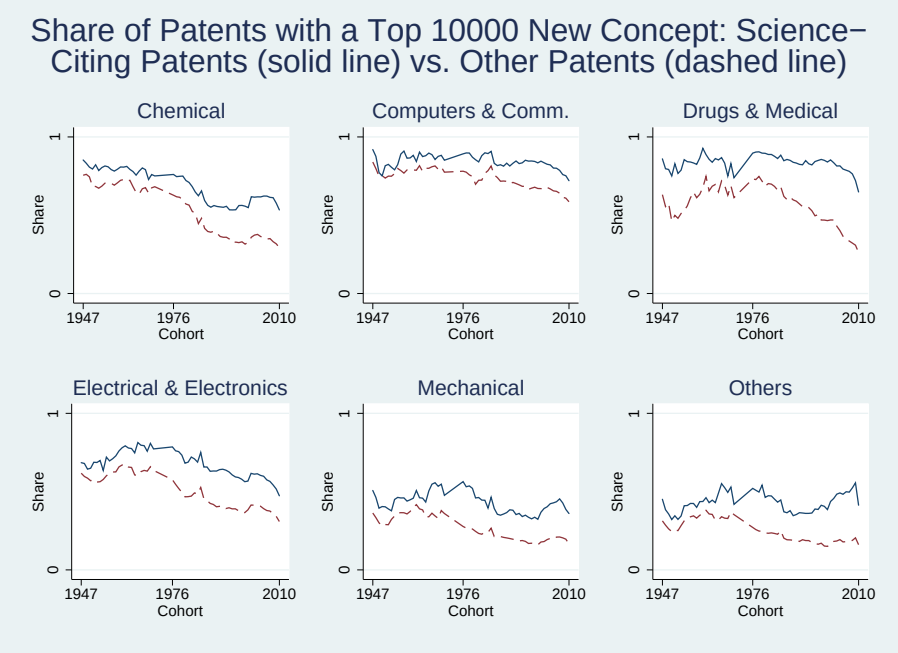


Figure 6.4: Number of Mentions for Top Concepts when New (years 0-9), by Concept Group

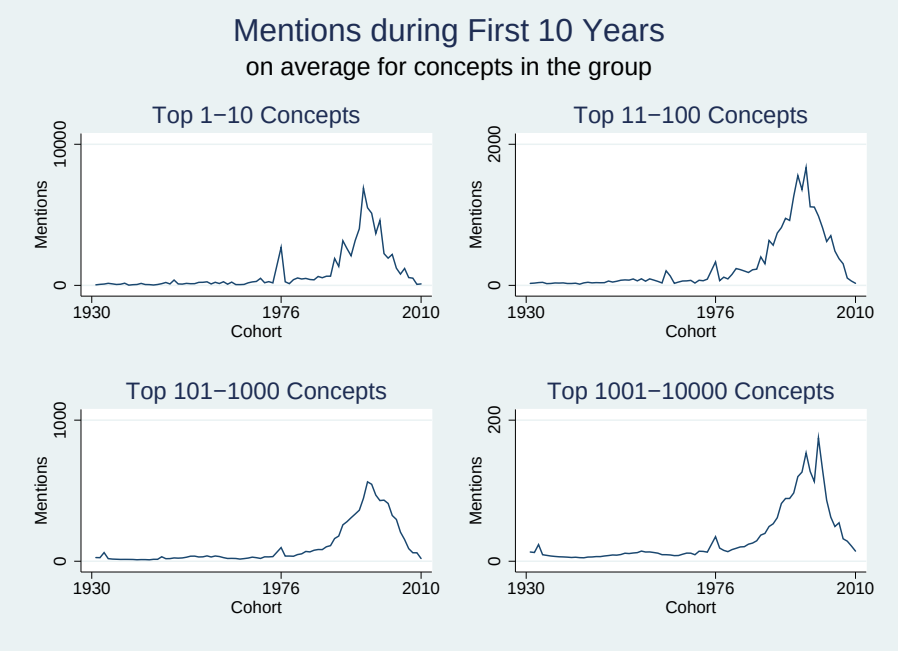


Figure 6.5: Number of Mentions for Top Concepts when New (years 0-9), by Technology Category

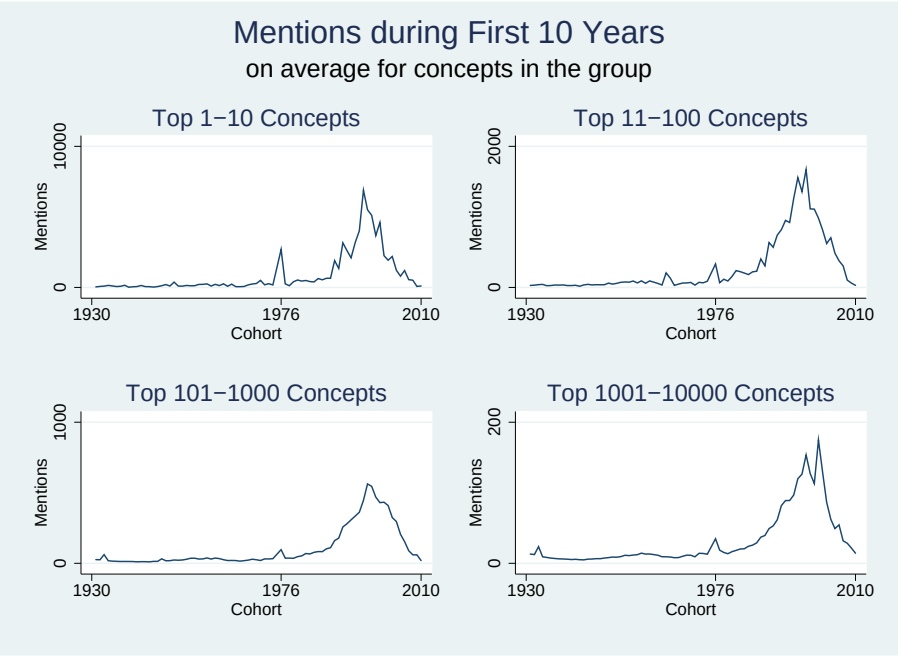


Figure 6.6: Ratio of Number of Mentions when New (years 0-9) vs. when Older (years 10-).

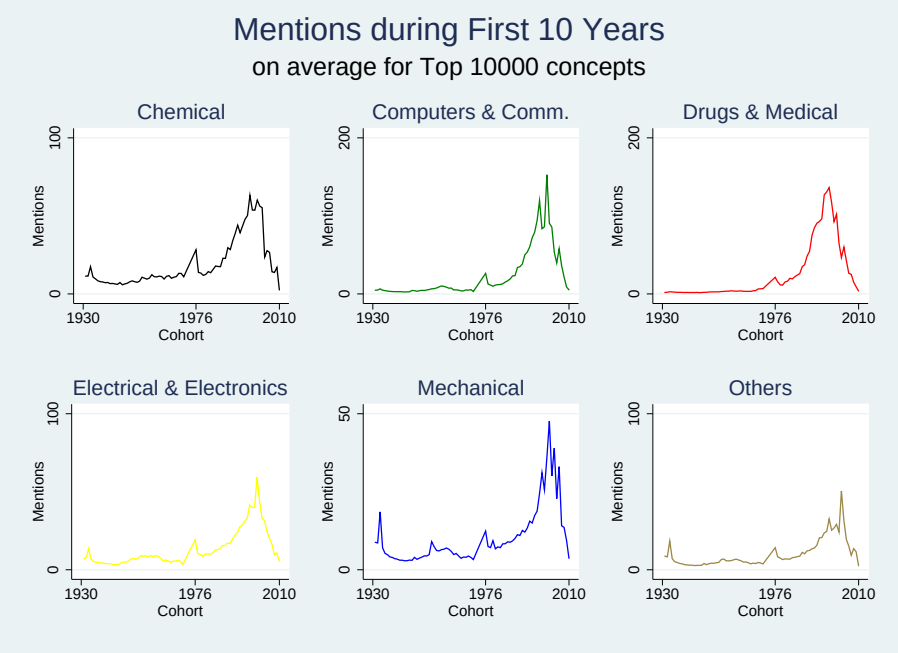


Figure 7.1: Citations to Innovative Patents vs. Other Patents, by Citation Measure.

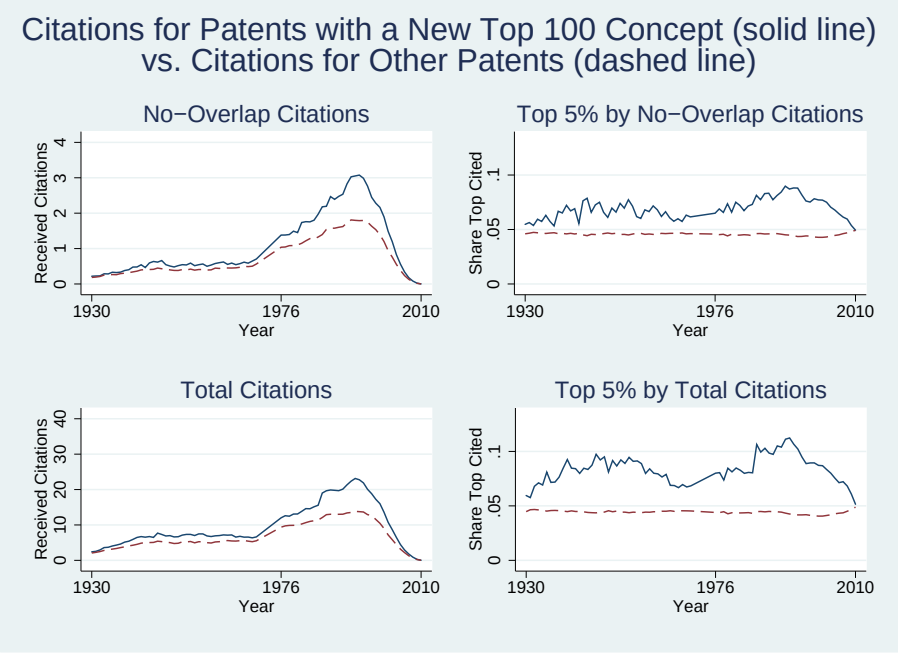
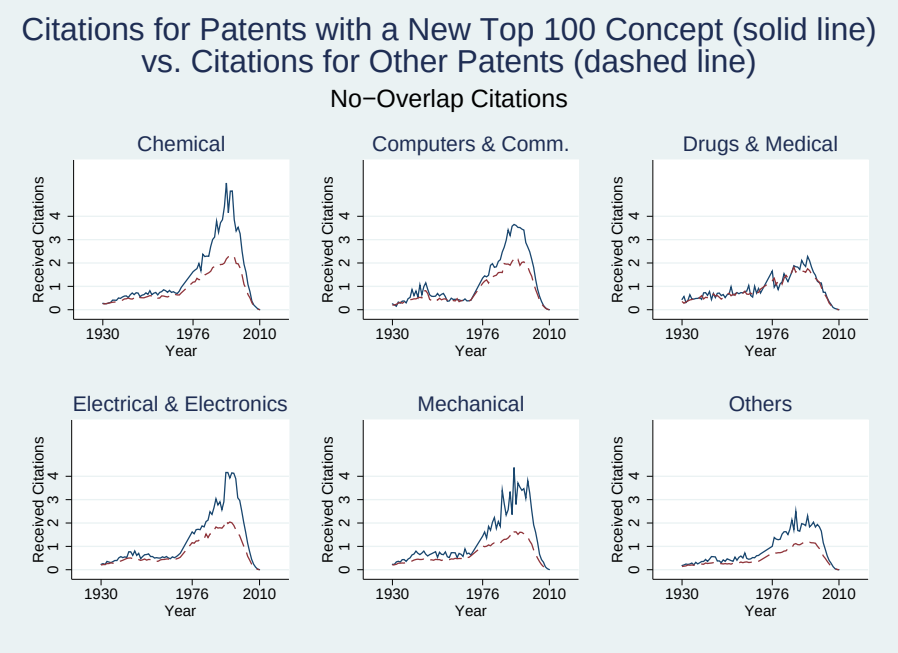


Figure 7.2: Citations to Innovative Patents vs. Other Patents, by Technology Category.



Tables

Table 1: Top 20 Most Popular Concepts by Decade of Cohort.

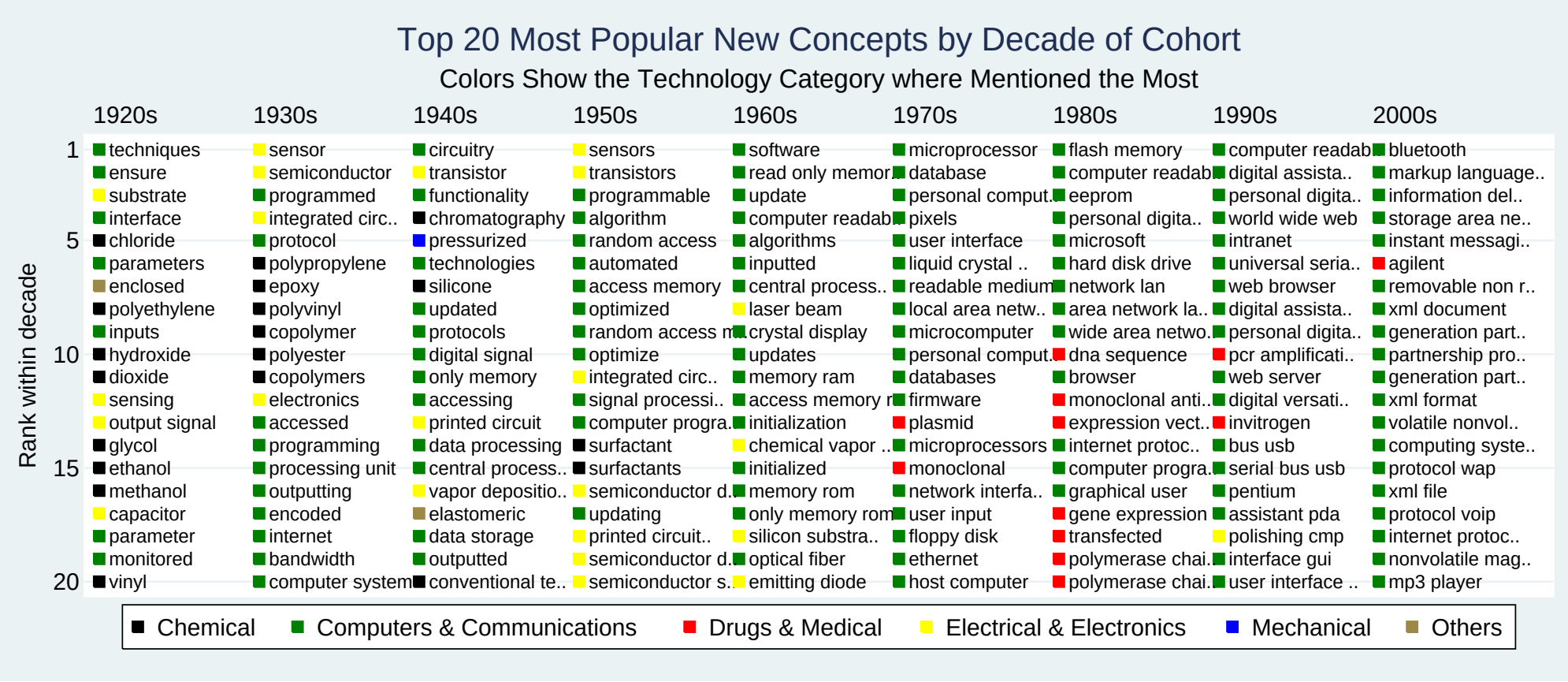


Table 2: Estimates of the value of using top 100 new concepts as research inputs.

Dependent variable:	(1) No-Overlap Citations	(2) Top 5% by No-Overlap Citations	(3) Total Citations	(4) Top 5% by Total Citations
Model:	Poisson	Logit	Poisson	Logit
Time Period: 1930-2005				
Technology Categories: All				
Top 1-100	1.59(0.02)[0.000]	1.70(0.02)[0.000]	1.51(0.01)[0.000]	2.10(0.02)[0.000]
Patent length*	1.28(0.00)[0.000]	1.31(0.00)[0.000]	1.25(0.00)[0.000]	1.53(0.00)[0.000]
Fixed effects (Year×Tech Class)***	included (28497)	included (29563)	included (29549)	included (29563)
Observations	4906312	4823706	4913729	4823706

(Footnotes to Tables 2-6 are presented after Table 6.)

Table 3: Estimates of the "Top 1-100" coefficient by time period (column 1), and with two alternative definitions of innovative patents (columns 2-3). Model the same as in Column 1 of Table 2.

Time Period:	(1) As above	(2) With most cited assigned as not innovative	(3) With concepts considered new only during years 0-4 after arrival
1930-1975	1.30(0.02)[0.000]	1.28(0.02)[0.000]	1.31(0.02)[0.000]
1976-2005	1.63(0.02)[0.000]	1.61(0.02)[0.000]	1.63(0.03)[0.000]

Table 4: Estimates of the "Top 1-100" coefficient by technology category. Model the same as in Column 1 of Table 2.

Time Period:	Chemical	Computers & Communications	Drugs & Medical	Electrical & Electronics	Mechanical	Others
1930-1975	1.25(0.02)[0.000]	1.21(0.03)[0.000]	1.11(0.05)[0.026]	1.29(0.02)[0.000]	1.40(0.06)[0.000]	1.42(0.04)[0.000]
1976-2005	1.69(0.03)[0.000]	1.54(0.02)[0.000]	1.12(0.07)[0.049]	1.76(0.03)[0.000]	2.03(0.06)[0.000]	1.78(0.05)[0.000]

Table 5: Estimates of the value of using top 100 new concepts as research inputs, by concept rank.

Dependent variable:	(1) No-Overlap Citations	(2) Top 5% by No-Overlap Citations	(3) Total Citations	(4) Top 5% by Total Citations
Model:	Poisson	Logit	Poisson	Logit
Time Period: 1930-2005				
Technology Categories: All				
Top 1-10	1.30(0.02)[0.000]	1.46(0.02)[0.000]	1.28(0.01)[0.000]	1.67(0.02)[0.000]
Top 11-20	1.23(0.02)[0.000]	1.26(0.02)[0.000]	1.21(0.01)[0.000]	1.43(0.03)[0.000]
Top 21-30	1.12(0.02)[0.000]	1.19(0.02)[0.000]	1.12(0.01)[0.000]	1.29(0.02)[0.000]
Top 31-40	1.09(0.02)[0.000]	1.10(0.02)[0.000]	1.11(0.01)[0.000]	1.17(0.02)[0.000]
Top 41-50	1.12(0.02)[0.000]	1.11(0.02)[0.000]	1.09(0.01)[0.000]	1.21(0.02)[0.000]
Top 51-60	1.03(0.02)[0.109]	1.07(0.02)[0.001]	1.08(0.01)[0.000]	1.15(0.02)[0.000]
Top 61-70	1.13(0.02)[0.000]	1.20(0.03)[0.000]	1.13(0.01)[0.000]	1.29(0.03)[0.000]
Top 71-80	1.17(0.02)[0.000]	1.14(0.03)[0.000]	1.11(0.01)[0.000]	1.27(0.03)[0.000]
Top 81-90	1.09(0.02)[0.000]	1.10(0.03)[0.000]	1.10(0.01)[0.000]	1.25(0.03)[0.000]
Top 91-100	1.14(0.02)[0.000]	1.14(0.03)[0.000]	1.13(0.01)[0.000]	1.22(0.03)[0.000]
Patent length*	1.28(0.00)[0.000]	1.31(0.00)[0.000]	1.25(0.00)[0.000]	1.25(0.00)[0.000]
Number of mentions for each concept group**	included	included	included	included
Fixed effects (Year×Tech Class)***	included (28497)	included (29563)	included (29549)	included (29563)
Observations	4906312	4823706	4913729	4823706
Mean of Predicted Value from Linear Model:				
w/ top 10	1.75	.082	16.0	.103
w/ top 11-100 but w/out top 10	1.24	.058	10.9	.067
w/out top 100	.91	.044	7.76	.040

Table 6: Estimates of the value of using top 10,000 new concepts as research inputs, by concept rank.

Dependent variable:	(1) No-Overlap Citations	(2) Top 5% by No-Overlap Citations	(3) Total Citations	(4) Top 5% by Total Citations
Model:	Poisson	Logit	Poisson	Logit
Time Period: 1930-2005				
Technology Categories: All				
Top 1-1000	1.29(0.01)[0.000]	1.37(0.01)[0.000]	1.19(0.00)[0.000]	1.43(0.01)[0.000]
Top 1001-2000	1.15(0.01)[0.000]	1.19(0.01)[0.000	1.14(0.00)[0.000]	1.27(0.01)[0.000]
Top 2001-3000	1.11(0.01)[0.000]	1.16(0.01)[0.000	1.11(0.00)[0.000]	1.22(0.01)[0.000]
Top 3001-4000	1.10(0.01)[0.000]	1.13(0.01)[0.000	1.11(0.00)[0.000]	1.22(0.01)[0.000]
Top 4001-5000	1.11(0.01)[0.000]	1.12(0.01)[0.000	1.10(0.00)[0.000]	1.21(0.01)[0.000]
Top 5001-6000	1.07(0.01)[0.000]	1.11(0.01)[0.000	1.10(0.00)[0.000]	1.19(0.01)[0.000]
Top 6001-7000	1.13(0.01)[0.000]	1.15(0.01)[0.000	1.12(0.00)[0.000]	1.22(0.01)[0.000]
Top 7001-8000	1.08(0.01)[0.000]	1.11(0.01)[0.000]	1.10(0.00)[0.000]	1.19(0.01)[0.000]
Top 8001-9000	1.10(0.01)[0.000]	1.12(0.01)[0.000]	1.12(0.00)[0.000]	1.25(0.01)[0.000]
Top 9001-10000	1.09(0.01)[0.000]	1.12(0.01)[0.000]	1.11(0.00)[0.000]	1.22(0.01)[0.000]
Patent length*	1.19(0.00)[0.000]	1.20(0.00)[0.000]	1.16(0.00)[0.000]	1.34(0.00)[0.000]
Number of mentions for each concept group**	included	included	included	included
Fixed effects (Year×Tech Class)***	included (28497)	included (29563)	included (29549)	included (29563)
Observations	4906312	4823706	4913729	4823706
Mean of Predicted Value from Linear Model:				
w/ top 1000	1.51	.052	12.4	.082
w/ top 1001-10000 but w/out top 1000	.99	.044	8.75	.048
w/out top 10000	.67	.041	5.92	.022

FOOTNOTES TO TABLES 2-6: Estimates are incidence rate ratios (Poisson models) and odds ratios (logit models). Standard errors are in parentheses (clustered by grant year and technology class pair); the associated p-values are in brackets. For logit models, observations for grant year and technology class pairs with more than 3500 observations are excluded (21 groups); the high number of positive outcomes in these groups prevents the conditional logit algorithm from converging.

(*) The variable measuring patent length is based on the number of non-whitespace characters in a patent. To address outliers, patent lenghts which are in the top 5% (bottom 5%) among patents granted the same year are replaced with the patent length in the 95th percentile (5th percentile). Patent lengths are then standardized so that the explanatory variable measures distance from the mean patent length in terms of standard deviations, relative to other patents granted the same year.

(**) For each concept group (e.g. Top 1-10 concepts), an explanatory variable is constructed for each patent, measuring the number of times the patent mentions a new concept in the corresponding concept group, minus one (see Section 2.3).

(***) Each grant year and technology class pair has a separate fixed effect (within groups estimation).

Appendix: Additional Tables

Tables A1-A4 are embedded as PDF files. Clicking on a colored link opens a table in a new window.

Table A1: Top 20 Most Popular Concepts by Cohort Year (Click here to open an embedded PDF).

By column number, the table lists the following items:

1. Cohort (i.e. the year the concept first appeared in a US patent).
2. Popularity rank among concepts in the cohort (based on column 4).
3. Concept name.
4. Number of patents in which concept mentioned during 1920-2010.
5. Number of patents in which concept mentioned during years 0-4 after the cohort year.
6. Number of patents in which concept mentioned during years 5-9 after the cohort year.
7. Technology category with the most patents that mention the concept.
8. Summary of how popular the concept has been in each of the 6 technology categories relative to other concepts in the same cohort. Here, "A", "B", "C" and "-", respectively, indicate "top 10", "top 100", "top 1,000", and "not top 1,000". Characters 1-6 in the entry are, respectively, for [1] Chemical, [2] Computers & Communications, [3] Drugs & Medical [4] Electric & Electronics, [5] Mechanical, and [6] Others.
9. GPT score (see Section 2.2).
10. Reports "GPT+" for each concept that is a General Purpose Technology in that the concept is very popular across sufficiently many technology categories (the GPT score is 4 or higher).

Table A2: Top 40 Concepts with the Highest GPT-scores by Cohort Decade (Click here to open an embedded PDF).

By column number, the table lists the following items:

1. Decade of Cohort (with Cohort in parentheses)
2. GPT rank among concepts with the same decade of cohort (based on column 10).
3. Concept name.
4. Number of patents in which concept mentioned during 1920-2010.
5. Number of patents in which concept mentioned during years 0-4 after the cohort year.
6. Number of patents in which concept mentioned during years 5-9 after the cohort year.
7. Technology category with the most patents that mention the concept.
8. Summary of how popular the concept has been in each of the 6 technology categories relative to other concepts in the same cohort. Here, "A", "B", "C" and "-", respectively, indicate "top 10", "top 100", "top 1,000", and "not top 1,000". Characters 1-6 in the entry are, respectively, for [1] Chemical, [2] Computers & Communications, [3] Drugs & Medical [4] Electric & Electronics, [5] Mechanical, and [6] Others.
9. GPT score (see Section 2.2).

10. Reports "GPT+" for each concept that is a General Purpose Technology in that the concept is very popular across sufficiently many technology categories (the GPT score is 4 or higher).

Table A3: Top 20 Most Cited Patents by Grant Year, 1920-2010. (Click here to open an embedded PDF).

By column number, the table lists the following items:

1. Grant year of cited patent
2. Popularity rank among patents with the same grant year (based on Column 3)
3. Number of citations from patents granted during 1947-2010
4. Patent title (the information for years 1920-1975 includes other information too because title is extracted from the OCR scan)
5. Patent number

Table A4: Top 20 Most Cited Scientific References by Publication Year, 1900-2010. (Click here to open an embedded PDF).

By column number, the table lists the following items:

1. Publication year of cited scientific reference
2. Popularity rank among scientific refereces with the same publication year (based on Column 3)
3. Number of citations from patents granted during 1976-2010
4. The scientific reference, as written in one of the citing patents (for references with double quotes, the text inside the quotes is shown first)

Data Appendix

1. **Obtain and organize Table of Issue Years and Selected Document Types Issued Since 1836 *(Grant Year File)*.** The table is at http://www.uspto.gov/web/offices/ac/ido/oeip/taf/issuyear.htm. These data specify the number of the first utility patent granted each year. We use these data to determine the grant year of each patent in the Master File.
2. **Obtain and organize U.S. Patent Grant Master Classification File (*Master File*) and link to Grant Year File.** The data are at http://www.google.com/googlebooks/uspto-patents-class.html. The Master File specifies the patent number, a primary technology classification, and multiple secondary technology classifications for granted patents. The assigned classifications are updated as the classification system changes (i.e. as new technology classes are introduced). Only the current version is available. We use version 1110 (mcfcls1110.txt, November 2011). We combine these data with the Grant Year File to obtain a list of patents granted each year, which enables us to examine the completeness of the Document Data. We use the assigned primary technology class (and grant year) of each patent in the Master File in determining the comparison group for each patent in the analyses of the value of innovativeness. There are 1,028 patents with no assigned primary technology class; we assign each such patent to the technology class that appears most often among the secondary technology classes assigned to the patent. The primary technology class is also combined with the Category File (see below) to assign each patent to one of 6 technology categories. We use the primary and all secondary assigned technology classes of citing and cited patents in determining which citations are Non-Overlap Citations.
3. **Obtain Mapping of Technology Classes to Technology Categories and Technology Category Labels (*Category File*).** The category labels and mapping are at https://sites.google.com/site/patentdataproject/Home/downloads/patn-data-description/classification_06.xls. These data specify 6 broad technology categories and map each 3-digit technology class to one category. Technology class 850 is not mapped in these data; we map it to technology category 4 (and to subcategory 43 though we do not use subcategory information). We use the mapping to assign each patent to a technology category based on the patent's primary technology class; this enables us to conduct technology category-specific analyses of the value of innovativeness and to determine how the mentions for a concept are distributed across technology categories. We also use the mapping to determine the technology categories spanned by the primary and all secondary technology classes of each citing patent and each cited patent; this information is used in determining which citations are Non-Overlap Citations.
4. **Obtain and organize Patent Document Data.**
   - **4.1. Download and unzip data**. The older data (1920-1975) are at http://www.google.com/googlebooks/uspto-patents-grants-ocr.html. The newer data (1976-2010) are at http://www.google.com/googlebooks/uspto-patents-grants-text.html. We use these data to determine (1) patent text, (2) citations to patents, and (3) citations to the scientific literature and other non-patent references.
   - **4.2. Store information on each patent in patent-specific files.** The extracted data have multiple patents in each file. Individual patents within each file are clearly indicated. Information on some patents appears in multiple places. The data format changes in the beginning of 1976, 2001 and 2005.
5. **Determine which patents in the Document Data appear also in the Master File.** We only analyze the information (text and citations) in those patents in the Document Data that appear also in the Master File.
6. **Determine patent text.** In the newer data, the fields we consider as patent text (title, abstract, brief summary, description, claims) are clearly indicated. In the older data, the fields are only indicated

among the scanned text. For these data, we determine where patent text ends and citations begin by searching for indications of the presence of phrases such as “CITED REFERENCES” and “following references are of”. Any text considered as being part of a bibliography is not included in the textual analysis.

7. **For each patent, index all words and 2- and 3-word sequences (*Concepts*) that appear in it.**
   - **7.1. Construct a list of all text in a patent.** In constructing this list, we add a space character between text from different text fields and paragraphs.
   - **7.2. Replace special characters with the space character.** Exceptions: parentheses, brackets, and braces are deleted; period, comma, colon and semi-colon are replaced (1) with “ X ” (space character, X, space character) when followed by whitespace (so that the indexed word sequences do not contain words from different sentences or words from different independent clauses of a sentence) and (2) with the space character when not followed by whitespace (as in those cases the characters may reflect something other than punctuation that separates two sentences or two independent clauses within a sentence).
   - **7.3. Change all alphabetic characters to lowercase.** In principle, analysing to what extent mentions of a given concept begin with an upper case letter could be used to exclude concepts such as “Microsoft” that do not represent innovation inputs in the traditional sense. However, such an approach would also exclude important inventions such as “Teflon” that have been important innovation inputs.
   - **7.4. Eliminate possessive case.** Character sequence “`s “ is replaced with the space character. Character sequence “ s’ “ is replaced with “s “
   - 7.5. **Replace with whitespace certain words that are likely typographical or other control sequences.** We replace with whitespace words in the set ['lpar','centerdot','largecircle','circleincircle','nbsp','amp','prt','num','plusmn','verbar','ensp','boxh','middot','thinsp','ltoreq','gtoreq','times','tau','phi','omega','beta','gamma','multidot','rho','dbd','plus','minus','prime','theta','eta','vertline','tab','alpha','equals','lambda','delta','epsilon','sigma','term','reg','rtm','sup','sub','xae' ,'lsqb','lcub','rcub','mdash','rsqb'], words that include a character sequence in the set ['dquo','squo','emsp','apos'], words that begin with a character sequence in the set ['po','pp','p','x','ijm','mems','art','str','equ','spc','pro'] *and* are followed by a number (the appearance of such sequences as new concepts typically reflect changes in how patents are recorded rather than new invention inputs), words that are 3 or 4 characters long and begin with “bis”, words that are 3 characters long and begin with “x”, followed by a letter and a number, words that are 3 or 4 characters long and end with “gr”, as well as words that begin with “fra” and are followed by a letter and a number.
   - **7.6. Replace with whitespace character sequences that have two or more consecutive numeric characters.** Concepts with multiple consecutive numeric characters are often page numbers, publication years and control sequences.
   - **7.7. Eliminate excess whitespace.** All whitespace longer than one character is replaced with a single space character.
   - **7.8. From the list of text, extract all words and all 2-, and 3-word sequences that satisfy character length limits on concept length and on length of individual words within multi-word concepts.** We only extract 1-grams with 3-29 characters, 2-grams with 7-59 characters and 3-grams with 11-89 characters. We only extract 2- and 3-grams for which each word is at least 3 characters long.
   - **7.9. Exclude concepts with DNA or RNA sequence information.** We exclude all concepts that include one or more words that consist only of letters in the set ['a', 'c', 'g', 't'] or in set ['a', 'c', 'g', 'u'].

7.10. Exclude concepts that include certain words or character sequences that reflect common words or patent terminology (which appearance as new concepts reflects changes in how patents are recorded). We exclude concepts that include a word in the set ['the','www','jan','feb','mar','apr','may','jun','jul','aug','sep','oct','nov','dec','therein','wherein','hereby','thereby','thereof','being','does','into','both','second','third','whose','therefore','last','this','from','that','such','which','with','will','then','also','means','when','between','other','said','from','more','have','these','thus','while','each','their','either','further','first','above','below','ser','art','filed','valid','invalid','novel','novelty','german','germany','japan','japanese','korea','korean','nos','priority','prior','fig','figs','figure','figures','table','tables','federally','sponsored','federal','examined','intellectual','shown','herein','apparent','readily','obvious','anticipated','heretofore','step','steps','applicability','technical','exemplary','known','ordinary','inventive','various','overview','abandoned','limiting','discussed','expressly','listed','laid','brief','reexamined','unexamined','identified','provisional','background','related','appln','applicable', 'harbor','wiley'.]
We exclude concepts that include a character sequence in the set ['apparatus','application','aspect','claim','compris','configur','depict','describ','descript','diagram','disclos','drawing','element''equivalent','embodiment','envisi','example','feature','flowchart','formula','illustrat','implemen','includ','invalidat','invention','limitation','methodol','modific','other','patent','particular','prefer','present','publication','reference','referri','significant','steps','summar','subclaim','typical','unillustrated','validat'].

7.11. Exclude multi-word concepts that include certain common words or word sequences, exclude multi-word concepts that include word sequences which appearance as new concepts reflects changes in how patents are recorded. We exclude multi-word concepts for which either the first or last word is a word in the set ['and','than','proc','its',',','from','that','with','are','can','has','via','for','was','nor','but','all','rev','were','using','press','having','have']. We exclude multi-word concepts which contain a character sequence in the set ['root over','manual cold','proc natl','acad sci','natl acad','sci usa','flow chart','utility model','subject matter'].

8. **Construct a list of the concepts that appear in the patents.**
9. **For each concept, index Cohort, Total Mentions, and Popularity Rank among concepts with the same cohort.**
   - **9.1. Determine the Cohort of each concept.** The cohort of a concept is the year in which the concept first appeared in any patent.
   - **9.2. Index the number of patents (total and in each technology category separately) in which each concept is mentioned, and rank concepts in each cohort according to the number of patents in which they appear.** We index the number of patents in which a concept appears, not the total number of times a concept appears in patents.
10. **For top 10,000 concepts, exclude concepts which likely reflect changes in how patents are recorded.** The newer patents are recorded using three different recording schemes, one for each time period 1976-2000, 2001-2004, and 2005-2010. In constructing the list of top 10,000 concepts, we exclude concepts that are mentioned during each year for one of these time periods but are not mentioned (or are mentioned only a few times) outside the time period.
11. **Construct patent-specific variables measuring mentions of top concepts.** For example, for the analyses reported in Table 2 we determine for each patent whether the patent mentions any of the 1,000 concepts that are top 100 ranked among concepts in their cohort *and* have a cohort year that is 0-9 years apart from the grant year of the patent.
12. **Index citations to patents.** In the newer Document Data, patent citations are indicated in a separate field. In the older Document Data, patent citations are among the scanned text. We extract these patent citations by first searching for indications of the presence of phrases such as “CITED

REFERENCES" and "following references are of" and then analyzing the text that follows. We extract the number, grant year and inventor name in each reference. We use grant year and inventor name information in these citations to compensate for OCR errors in the following way. We only include citations for which either the cited grant year is within 10 years of the actual grant year of the cited patent or the first letter of the cited inventor name matches the first letter of the actual inventor name of the cited patent (the actual grant year is determined from the Master File and Grant Year File; the actual inventor name is determined from citations to pre-1976 patents in the newer patent data).

13. **Index citations to scientific and other non-patent references.** In the newer Document Data, non-patent references are indicated in a separate field (there are additional non-patent references in the patent text but we do not consider them to limit the scope of the analysis). To distinguish scientific references from other non-patent references, we first search the non-patent references for terms that would indicate that the reference is to a patent reference, technical publication, marketing material, or web page (the searched terms include terms such as "ser. no.", "patent", "pat. appl", "derwent ", "database wpi", "search report", "office action", "advertisement", "ibm technical bulletin", "disclosure", "language abstract", "withdrawn", "JP", "EP", "english translation", "www. ", "website ", etc.). We designate as potential scientific citations all references for which such terms are not found. Among the potential scientific citations we then search for an indication of a publication year (we first search the reference for parentheses, and then search inside the parentheses for a 4-digit number between 1500 and 2015; when a such sequence is not found we search for 2-digit numbers that follow either the character " ' " or the character " / ".) The citations among the potential scientific citations for which a publication year is found are considered scientific references. We also disambiguate these scientific references to display the list of most cited scientific references in Table A4. To disambiguate the scientific references, we only seek to match scientific references that have the same publication year. After indexing the scientific references in patents by publication year, we seek citations that have two double quotations (as they often surround a title); such citations are matched based on words inside the double quotations (title), other citations are matched based on all words in the reference. Matching is attempted only for references with the same publication. Before seeking matches, we exclude certain character sequences such as "pages " that can be expected to be present in some citations to a scientific reference but not in other citations to the same reference. We also exclude character sequences that include non-alphabetic characters and character sequences that are shorter than 3 characters. We do not disambiguate references to Chemical Abstracts and references to GenBank accession numbers when the references only specify numerical information (as is typically the case). We consider two citations to be to the same scientific reference (i.e. a match) when the SequenceMatcher comparison in Python returns a value above 0.9. Before using this comparison algorithm, we organize all words in each reference alphabetically. In the older Document Data, non-patent references are among the scanned text. For these older data, we extract non-patent references by searching for indications of the presence of the phrase " OTHER REFERENCES " within the " CITED REFERENCES " section and then search for publication years (a 4-digit number). Older patents for which such publication year is found among other references are assigned as having a non-patent reference. The search for an non-patent reference section is stopped when an indication is found for the presence of phrases as "CERTIFICATE", "FOREIGN PATENTS" or "CORRECTION".
14. **Index patent titles.** In the newer Document Data, patent titles are indicated in a separate field. For the older Document data, we extract patent title based on the appearance of capital letters near the beginning of the text. The data on patent titles is used in constructing the list of most cited patents in Table A3.