Hard to get? Knowledge Heritage from Invention to Product Market Innovation

Using Natural Language Processing

Sheryl Winston Smith, Ph.D.
BI Norwegian Business School
sheryl.winston.smith@bi.no

January 15, 2019

Submission to 2019 Wharton Technology Conference

ABSTRACT

Patents and patent citations are a window into the innovation process. Yet even established firms have difficulty commercializing inventions. A conundrum for scholars is: does the inventive activity captured in patents and patent citations translate into innovation in the product market? In other words, how reliably can we understand the link between patents and introduction of novel products to market? Thus, unpacking the link between knowledge and innovation is crucial for innovation, entrepreneurship, and strategy scholars. This paper brings to the innovation literature an important methodological approach from computer science: application of natural language processing and the vector space model in information retrieval to the study of innovation. Specifically, this paper details a novel methodology to algorithmically compare the similarity of patent texts to the description of innovations in two distinct contexts: 1) novel products brought to market by corporate venture capital investors in the medical device industry, and 2) IPO prospectus of startups in the communications equipment industry.

Keywords: Natural Language Processing; Text Analysis; Innovation; Corporate Venture Capital; IPO; Intellectual Property; Commercialization of Technological Innovation; Technology Entrepreneurship
INTRODUCTION

Scholars studying innovation, entrepreneurship, and strategy continually seek to understand how firms cultivate, develop, and use novel knowledge in pursuit of innovation. Myriad questions exist: How efficiently does a company use scientific or technical knowledge to obtain patents and is this novel knowledge incorporated in new products? Similarly, for startups, the ability to reach significant milestones, such as an initial public offering (IPO), depends on the innovativeness of its ideas. More broadly, for established companies and startups alike, what value does the company receive in exchange for investment in innovation?

For scholars and practitioners, answering these crucial questions requires precise ways to discern how inventiveness, i.e. patenting, is reflected in observable performance outcomes such as introduction of novel products or successful launch of new ventures. This paper harnesses the power of an important set of methods from the growing computer science field of natural language processing (NLP) in a novel manner to more directly link inventive activity and innovation-related performance outcomes. From a theoretical standpoint, these methods present two types of value to researchers. In a specific sense, the methodology in this paper enables scholars to directly examine the knowledge heritage of innovative products and ideas. In a broader sense, these methods enable researchers to more fully analyze the capabilities and incentives that bring inventions to fruition, e.g., through commercialization of products or creation of new ventures.

Patents are key milestones in the innovation process. Patents offer a substantial window on invention (Griliches, 1990), and patent counts and citations are used widely to infer knowledge transfer (Hall et al., 2005; Jaffe and Trajtenberg, 2002; Trajtenberg 1990; Ziedonis, 2004). However, this inference is clouded by strategic citations and from addition of examiner citations of which the inventor was often unaware (Alcácer and Gittelman, 2006; Alcácer et al., 2009;
Lampe, 2012; Lemley and Sampat, 2012). Overall, while patents are valuable measures, they also suffer from noted limitations. Identifying knowledge heritage more directly through the computationally sophisticated textual analysis presented here can ameliorate this.

To illustrate the points above, two distinct settings are highlighted. First, the paper develops an application rooted in a key strategic question from the knowledge search and CVC literature: *How can we measure the value companies derive from knowledge search, and, more specifically, what innovation-related outcomes can companies attribute to CVC investments?* The empirical context in this paper—drawn from a novel dataset of CVC investment in the medical device industry—emphasizes the usefulness of this methodology in tracing knowledge heritage between patents and innovative products. Several factors make this a clean setting in which to focus on the novel methodology. Companies in this industry actively seek breakthrough innovation, patent heavily, and leave a paper trail of patent and pre-market approval documents that can be analyzed (Ernst and Young, 2011). Moreover, the context of CVC investment allows isolation and comparison of a distinct source of invention originating outside the company that leaves a knowledge signature on the corporate investor’s product innovations.

While the focal context is on CVC investors’ innovation outcomes, the methodology has broader potential applications. For example, this methodology is well suited to probe inside the black box of the genesis of new ventures and the commercialization of innovative products (Aggarwal and Hsu, 2009). To illustrate this, the paper further spotlights a key research question relating to the development of innovation capabilities in new ventures: *To what extent do capabilities surrounding invention contribute to the launch of startups and successful commercialization?* The potential of this methodological approach to link inventive activity, commercialization of products, and new venture launch through IPO is examined using an example of a startup initial public offering (IPO) in the communications equipment industry.
This example also points to generalizability of the methodology in a range of settings.

The paper makes three distinct contributions. Foremost, the paper pushes forward strategy scholarship by signifying the computational power associated with using language as data: specifically, this paper applies versatile IR methodology from computer science to allow strategy scholars to study a core array of questions about innovation and technological capabilities.

Second, ideas and knowledge are the feedstock of innovation. By applying a powerful computational approach to track the knowledge heritage of new products, this methodology permits insightful answers to key research questions: Does success in invention follow through to innovation? To what extent do companies leverage technological capabilities for market success? Third, the paper extends the literature on CVC investment by identifying a new method for tracking the essential outcome of interest—innovation—that has previously been elusive (Benson and Ziedonis, 2009; Dushnitsky and Lenox, 2005; Dushnitsky and Shaver, 2009; Katila et al., 2008; Winston Smith and Shah, 2013). This methodology provides new insight into CVC as a mechanism for external knowledge search. Finally, this paper contributes to our understanding of the extent to which IPOs draw upon the inventive activity of new ventures.

**TRACING KNOWLEDGE FLOWS BETWEEN PATENTS AND INNOVATION OUTCOMES**

The link between knowledge, i.e., fundamental scientific and engineering concepts, and innovation-related outcomes, such as introduction of novel products or launch of a new venture, is subject to substantial social science research. Appropriating benefits from inventive activity is fraught with problems, including technical uncertainty, market uncertainty, and difficulty preventing encroachment on ideas (Arrow, 1962; Levin et al., 1985). Gains from invention can happen through licensing if markets for technological ideas are well-developed (Arora et al., 2001). Startups may engage in cooperative R&D with larger firms to bring inventions to market
(Aggarwal and Hsu, 2009). However, even established firms face a difficult path to commercialize invention (Gittelman and Kogut, 2003). Thus, a fundamental question that vexes strategy scholars is: does the inventive activity captured in patents (and patent citations) translate into product innovation? In other words, how reliable is the link between innovation, often studied through patents, and innovation, i.e. introduction of novel products to market?

Patents are an enduring measure of inventive activity (Griliches, 1990). The logic of tracing knowledge flows through patent citations is straightforward. Backward citations to prior patents represent a legally enforceable boundary between existing prior art and the new invention, which must be both novel and non-obvious (Trajtenberg 1990). Thus, patent citations are widely used to measure knowledge transfer. However, the extent of knowledge flow is clouded by addition of a substantial proportion of citations by patent examiners after the patent is filed, rather than the inventor (Alcácer et al., 2009). As well, patents and citations do not indicate whether the invention results in an innovative product being introduced to market. This distinction is far from trivial, as the stark example of Kodak illustrates. Despite possessing a valuable patent stock resulting from decades of inventive activity, Kodak was forced to declare bankruptcy, largely due to the failure to introduce successful product market innovations. For example, Kodak began receiving patents related to digital photographic technology in the 1970s; however, the company did not introduce its first digital camera to market much later and then only as a reluctant market entrant (The Economist, 2012). Finally, reliance on patent protection varies significantly by industry (Levin et al., 1987).

Recognizing that patent citations and citation patterns alone do not allow a complete picture of the transfer of knowledge in the innovation process, scholars increasingly seek new approaches to probe the relationship between knowledge and innovation. Indeed, a surprising divergence exists between scientific knowledge and inventive success (Gittelman and Kogut,
Roach and Cohen (2013) link survey data with patent citations to pinpoint the contribution of knowledge flows from public institutions to industrial innovation. They find that patent citations underrepresent certain types of knowledge flows, such as contracted academic research. They also find that citation patterns are influenced by strategic concerns. Alternatively, borrowing from computer science, Kaplan and Kavili (2014) deploy close textual analysis using topic modeling based on latent Dirichlet analysis to understand the evolution of new cognitive breakthroughs based on patent abstracts in the context of nanotechnology.

**Using Natural Language Processing to Link Invention, Knowledge, and Innovation Outcomes**

Text matching algorithms can be used to quantify overlap between distinct documents reflecting inventive activity, e.g. text in patent documents, and product innovations, e.g., text in pre-market approval (PMA) documents.

The basic intuition and application of natural language processing techniques (NLP) and information retrieval (IR) are covered in detail in the computer science literature (Kwon and Lee, 2003; Manning et al., 2008; Salton and McGill, 1986; Salton et al., 1975). Briefly, IR as a field can be thought of as ‘… finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)(Manning et al., 2008, p.1). IR covers wide ground: open-ended web search, domain-specific search (such as a patent database), and personal information. Technically, IR and NLP are distinct fields, where the former focuses on probabilistic approaches and the latter on the use of “natural” (as opposed to “controlled”) queries (Lewis and Jones, 1996). However, although these fields were largely separate for decades, they are becoming increasingly interchangeable as NLP adopts probabilistic approaches and IR broadens in scope (Lease, 2007; Lewis and Jones, 1996; Manning and Schuetze, 1999; Singhal, 2001).
This paper focuses on domain-specific search (Manning et al., 2008, p. 2). Search space can be conceptualized as a query seeking answers relevant to a question at hand. In this case, the query—what is relevant to innovation?—results in looking at outcomes that reside in relatively large collections of documents—e.g., patent or PMA databases—and searching for the kernel of knowledge that is valuable to the relevant outcomes. Mathematically, IR weights the relative similarity of knowledge from two sources of text to facilitate efficient identification, based on the criteria of precision, i.e., the proportion of correct documents retrieved, and recall, i.e., the proportion of relevant documents returned (Lease, 2007; Lewis and Jones, 1996; Singhal, 2001).

The vector-space model underlies this process and enables direct quantification of document similarity between a given pair of retrieved texts. This paper algorithmically executes a series of steps to simplify text documents, assign relative weights to words, and map them into n-dimensional vector space. A meaningful overlap between these vectors is then computed to provide a metric for document comparison (Salton and McGill, 1986; Salton et al., 1975; Singh and Dwivedi, 2012). As such, this method is considered to be a ranked retrieval model (Manning et al., 2008). These algorithms comprise the backbone of search engines (Google, 2013a, b; Kwon and Lee, 2003; Singhal, 2001).¹

Several papers in strategy research look to variants of NLP methodologies to quantify the relative importance of words within a given text. Menon et al. (2018) use NLP to analyze strategy change and differentiation from rivals. Lee and James (2007) use centering resonance analysis (CRA), which utilizes network positioning of words in text to determine influence, in this case to relate gender to media slant in the appointment of new chief officers. Ronda-Pupo and Guerras-Martin (2012) use co-word analysis, in which they measure the likelihood of two-

¹ To be precise, latent semantic analysis (also referred to as latent semantic indexing) is also based on word stemming and the vector space model, in common with the IR methodology applied in this paper. However, latent semantic analysis involves development of a matrix representation of each document, which then needs to be decomposed (using singular value decomposition) in order to compare documents (Hofmann, 1999).
words appearing together in a given document against a probabilistic backdrop, to study the evolution of research concepts in the field of strategic management. In a similar fashion, Kabanoff and Brown (2008), apply machine learning techniques based on text classifying algorithms to examine the evolution of strategic discourse by CEOs. In the context of innovation, and closely related to the methodology in this paper, Kaplan and Vakili (2014) use topic modeling to trace the evolution of breakthrough ideas in the emergent field of nanotechnology based on a form of natural language processing known as latent Dirichlet allocation (LDA) analysis of patents. This approach is based on probabilistic modeling of co-occurrence of words to infer meaning about ‘latent’ topics (Blei et al., 2003; Chang et al., 2009).

RESEARCH CONTEXT I: CORPORATE VENTURE CAPITAL AND COMMERCIALIZATION OF NOVEL PRODUCTS

Innovation propels sustainable competitive advantage over rivals. Breakthrough innovations, in particular, confer advantages in entering new product areas or creating altogether new product markets are thus a holy grail. The strategic management literature increasingly identifies a broad set of search mechanisms that firms use to achieve breakthrough innovations (Maula et al., 2012; Narayanan et al., 2009). Key among these is the practice of corporate venture capital (CVC) investment, where large corporate investors make equity investments in startups to gain access to novel knowledge (Dushnitsky and Lenox, 2005). Research spanning strategy and finance suggests that CVC results in innovation related returns to investors (Benson and Ziedonis, 2009; Chemmanur et al., 2014; Winston Smith and Shah, 2013)

While startups afford new ideas for CVC investors, this knowledge is hard to transfer back to the corporate innovation context. In part, this is a natural outcome of a variety of organizational features, such as the distinctive innovation climates in startups relative to large corporations or potential misalignment of incentives (Dushnitsky and Shaver, 2009; Lerner,
As well, the research may be more exploratory (Wadhwa and Kotha, 2006). CVC investors state strategic goals such as bringing new products to market as the reason for their investments; however, CVC is a long-term strategy, and resulting products rarely display an identifiable knowledge path back to the original investment (Lerner, 2013). Moreover, new products often draw upon a variety of complementary internal and external investments, making it difficult to attribute the value contributed by any one source. Given the marked potential attenuation of knowledge, managers and scholars require particularly precise measures to assess the path from investment to corporate innovation.

This computational approach can be implemented to link underlying scientific and technological knowledge contained in patent documents—i.e., the knowledge heritage—to introduction of novel products. Notably, the methodology can be applied in any setting where texts can be compared to one another, making it broadly applicable in strategic management scholarship. Here, the methodology is used to address a vexing question in CVC research: How can we measure the value companies derive from knowledge search, and, more specifically, what innovation-related outcomes can companies attribute to CVC investments?

The primary empirical context here is the medical device industry, an industry driven by innovation and hallmarked by extensive CVC activity (Standard and Poor's Corporation, 2007). The medical device industry is an excellent setting to demonstrate the power of these IR techniques. The medical device industry provides text to analyze that reflects both invention and innovation outcomes. Firms in this industry nearly universally rely on patents, which codify the underlying scientific and technological knowledge, to define intellectual property rights (Levin et al., 1985). Firms must file for PMA by the U.S. Food and Drug Administration (FDA) to introduce a new medical device in the marketplace, thus documenting the innovation. Moreover, clinical trials are costly—over $100 million by some estimates—and PMA approval equates with
introduction to market (Ernst and Young, 2009).

The medical device industry also is characterized by fast innovation cycles. Achieving and sustaining an edge in innovation is crucial to the introduction of new devices to market. (Ernst and Young, 2011). To this end, device companies investment routinely in startups for access to external sources of innovative ideas (Ernst and Young, 2011; PricewaterhouseCoopers, 2006). From a methodological perspective, CVC investment in startup companies provides a transparent case to track knowledge heritage: the patents are filed by and belong to separate companies but the knowledge can be linked to the investors’ PMAs. Finally, from a strategy perspective, the medical device industry continues to be fertile setting for strategy scholars to examine important aspects of innovation (Chatterji et al., 2008; Karim and Mitchell, 2000; Winston Smith and Shah, 2013; Wu, 2013). For policy makers, understanding the link between knowledge and innovation helps save lives and decrease costs (Winston Smith and Sfekas, 2013).

Selection of text documents to analyze

The text matching approach above is applied to a unique dataset consisting of all CVC investments in the medical device industry made by the 4 largest medical device-CVC investors in this industry (Medtronic, Boston Scientific, Johnson and Johnson, and Guidant) over the period 1978-2007. One hundred and twenty-four unique startup companies received at least one round of CVC investment by at least one of these four incumbents over the sample period. To capture inventive activity, the full text of all focal patents (i.e., patents filed by the four incumbents that cited the startup companies) and all of the startup patents filed between 1978-2010 were downloaded using the US Patent and Trademark Office database. Patents must cite related prior patents, enabling us to link each incumbent’s patents to the startup companies’ patents. To trace innovation, all PMAs filed by these CVC investors were extracted from the full U.S. FDA Pre-market Approvals database. Together, these four companies filed 170 PMAs
between 1978 and 2007. The PMAs were then matched to the 1,954 patents filed by the startups. (An illustrative case is extracted from this sample to demonstrate the steps below.)

To illustrate the approach, patent texts from Image-Guided Neurologics, a startup in which Medtronic invested, and Medtronic’s own patents are compared to texts describing innovations that were commercialized and introduced to market by Medtronic: 1) Activa for Deep Brain Stimulation (DBS); 2) Interstim Sacral Nerve Stimulator; 3) Synchromed Infusion Pump; and, 4) Intrel Spinal Cord Stimulator.

The methodological goal is to create an analytical comparison between two sets of text documents, those representing inventions (in this case, patents) and product innovations (in this case, PMA applications), to allow quantification of how closely related these texts are to one another. In this manner, applying IR theory and techniques creates an original measure that links the knowledge heritage of product market innovations by the CVC investor back to the startups in which it invested. Practically, these steps are accomplished using the open-source programming language Python and associated modules. Python is fast, well-documented, and continually augmented through user-generated modules for a variety of tasks undergirding IR and ‘big data’ analysis (Prechelt, 2000).

**Algorithmic calculation of knowledge similarity between pairs of text documents**

The vector-space model identifies common elements between document pairs. Once the texts are processed, the intersection of the two dictionaries (in this case, the ‘patent dictionary’ and the ‘PMA dictionary’) is generated. The number of words, \( N \), in each dictionary defines an \( N \)-dimensional vector. The intersection vector is composed of elements whose magnitude is given by the tf-idf score. Practically, this is accomplished using Python code for precise vector computations to calculate the term frequency (tf) and inverse document frequency (idf) measures described above. At this point, each document (i.e., the patent document and the PMA
document) is now represented as a vector in which individual words are the vector components and the tf-idf weights give the magnitude of each component. The final computational task involves quantifying the extent to which the two documents are related by computing the cosine distance between the two vectors in common space. This information is used to calculate overlap between each patent document relative to each PMA document.

**Illustrative example: Medtronic and Imaged-Guided Neurologics**

The tension between the quest for relevant external knowledge and the challenge of incorporating that novel knowledge in a way that benefits the investor lies at the heart of CVC investing (Dushnitsky and Shaver, 2009; Wadhwa and Kotha, 2006). In an overview of CVC practices, Lerner (2013) observes: ‘Knowledge doesn't automatically flow from start-ups to the large organizations that have invested in them—at least not in a timely manner.’ Nonetheless, he concludes: ‘For companies that have found traditional in-house research unequal to the task of generating valuable insights into next-generation technologies or the movements of the market, the creation of a venture fund might well prove to be what executives are always looking for—the breakthrough idea that changes everything.’

The methodology in this paper opens a wider window on tracing the often-oblique path from breakthrough ideas to novel innovations. The surrounding context is the explosive potential of the neuromodulation market for Medtronic and rival device makers (Arndt, 2005b; Medtronic Inc., 2014a). To elucidate the power of the textual analysis methodology, four devices for which Medtronic filed PMA applications (Activa for Deep Brain Stimulation (DBS), Intrel Spinal Cord Stimulator, Interstim Sacral Nerve Stimulator, and Synchromed Infusion Pump) are analyzed in comparison to two sets of patent documents: 1) cited patents of Image-Guided Neurologics (IGN), a startup company, and 2) the focal Medtronic patents that cite the IGN patents. These devices share common scientific and technologic principles with each other (and with cardiac
pacemakers), i.e., implantation of tiny electrical stimulators in the body. At the same time, each organ system in the body requires specific expertise and capabilities.

In 2003, Medtronic invested in IGN, a startup with cutting-edge development of a ‘frameless’ stereotactic head frame. The investment represented a continuing foray into the expanding area of neuromodulation, which relied on implantation of precisely targeted stimulators in the brain to treat multiple issues, including Parkinson’s disease and epileptic seizures (Arndt, 2005b). Neuromodulation represented a growing area of exploration for Medtronic. On one hand, the underlying principles were complementary to Medtronic’s core expertise in cardiac rhythm management—for example, in online Q&A Medtronic notes that DBS is ‘sometimes called a brain pacemaker’ (Medtronic Inc., 2014b). On the other hand, in response to the question: ‘The pacemaker has been around for more than 40 years. Why has it taken so long to find other applications for this device?’ then CEO Dr. Steven Oesterle commented on the difficulty of succeeding in this market: ‘... the barriers to entry to the IPG [implanted pulse generators] market are huge. These things are extremely complicated and difficult to make, so you don't see people say, "Gee, I could go do that," raise $10 million, and get working. The info-tech hurdles are as big as they get.’ (Arndt, 2005a). In the same interview, however, he presciently observed: ‘This is a huge market. It is underdeveloped and underpenetrated, and it's where growth will be in the medical-device industry over the next decade.’ (Arndt, 2005a). In other words, this represented an area of technology that was both difficult to master but crucial to future growth, i.e., the ultimate logic underlying CVC investing.

Table 1 presents cosine similarity matches between the IGN patents cited by focal Medtronic patents and each of the Medtronic PMAs. These scores show different degrees of connection to each of the PMAs. Typically, a medical device builds on multiple patents. The values suggest the highest relative similarity is between IGN’s patent #720480 (‘Deep organ
access device and method”) and Medtronic’s Activa for Deep Brain Stimulation PMA. Interestingly, the degree of similarity (cosine similarity score 0.109) is close to that of Medtronic’s focal patent #736651 (“Robotic trajectory guide”) that cited this IGN patent to the same PMA (cosine similarity score 0.115). Moreover, the average cosine similarity of all cited IGN patents (0.08118) exceeds the average score for the focal Medtronic patents (0.06617), suggesting relatively stronger knowledge linkage between IGN’s patents. Table 2 provides a fuller flavor of the analysis by showing underlying representative word stubs and associated tf-idf weights for PMA 960009 associated with Medtronic’s Activa Neurostimulator and IGN’s cited patents.

-----------------------------------------

Insert Table 1 & Table 2 about here

-----------------------------------------

Common knowledge heritage amongst the devices is also reflected in the scoring. The next highest set of scores for IGN patents is with Medtronic’s PMA 970004, the Interstim device for treatment of incontinence (0.07119); again the IGN patents score higher on average than those of Medtronic (0.05932). The average score for PMA 860004, the Synchromed Infusion Pump (0.05775) is slightly lower than that for Medtronic (0.07949) but is similar to IGN’s match with Activa and Interstim. Interestingly, these three devices share more common physiological properties than does the Intrel spinal stimulator (Arndt, 2005a; Medtronic Inc., 2014a). Patents from both IGN and Medtronic are relatively less connected with the spinal stimulator than the other three devices. Moreover, IGN’s score (0.03891) is relatively lower than Medtronic (0.04002).

The algorithmic textual analysis used in this example underscores more general points. First,
the relative magnitude of the cosine similarity scores provides basis for comparison between different sets of documents, e.g. IGN patents matched to a Medtronic PMA, relative to Medtronic patents matched to the same PMA. In this manner, the scores permit inference about the relative strength of the linkage. Alternatively, the scores could also be used to set a threshold level. Second, as given, these examples provide a substantive view inside the scoring and the deep knowledge it can provide. However, it is worth noting that knowledge-match scores are computed for all potential CVC investor-startup dyads, resulting in a full matrix that quantifies how closely every CVC investor's PMA is related in context-dependent textual analysis space to the patents of each startup. This matrix can then be used as an input for econometric analysis.


**RESEARCH CONTEXT II: COMMERCIALIZATION AND NEW VENTURE CREATION**

To illustrate the versatility of this methodology beyond the medical device setting (where the existence of PMA documentation provides a textual record of innovation), a second example is provided below. The knowledge heritage of Vocera, a startup in the communications equipment industry is linked from its patents to textual measures of innovation leading to commercialization (product reviews) and new venture creation (IPO prospectus) to demonstrate the broader usefulness of this approach.

**Text Matching Vocera patents, IPO prospectus, and product reviews**

A growing question in the strategy and CVC literature focuses on the impact of CVC investment prior to IPO on the startup (Chemmanur et al., 2014; Katila et al., 2008; Park and Steensma, 2012). In order to elucidate generality of this methodology beyond the medical device
setting, the second empirical context applies this methodology to analyzing the knowledge heritage in product innovations and the IPO prospectus of Vocera (VCRA), a communications and software company. Prior to its IPO in 2012, Vocera received CVC investment from Motorola, Cisco, and Intel (Albenesius, 2005; Lipset, 2003). This context illustrates how this methodology can be used to more deeply probe the relationship between CVC investment and startup innovation, including the impact on the commercialization by the startup and launching of a new venture.

Table 3 presents representative results from this analysis applied to Vocera’s patents and to textual measures of innovation leading to commercialization (product reviews) and new venture creation (IPO prospectus). As well, this table points to the variety of documents that can be mined for insights into innovation outcomes. As in the prior example, these results suggest that Vocera leaned on its technical knowledge heritage in commercialization of new products. Columns 1-3 analyze the degree of similarity between Vocera’s 16 patents and innovation outcomes, represented through different product reviews of newly commercialized products. The results suggest that Vocera introduced product innovations that heavily encompass the underlying technical knowledge, as evidenced by substantially large cosine similarity scores between the patent portfolio and each of the three technical reviews. Column 4 also tracks the knowledge relationship to the products, but uses a professional medical journal article for the textual record of the product.

Column 5 applies this semantic matching methodology to trace the knowledge heritage
embedded in Vocera’s S-1 form filed with the SEC. Specifically, the IPO prospectus is compared to Vocera patent portfolio. In this case, the cosine similarity results suggest that Vocera leveraged its technical knowledge in reaching IPO. The literature suggests that patents help new ventures launch successfully. These patents may serve as a signaling mechanism for new firms (Hsu and Ziedonis, 2013). An additional explanation may be that the patents represent the knowledge that forms the backbone of the new venture. This example thus provides insight into questions of scholarly significance that are distinct from those presented in the main analysis. Future research can include broader and comparative analytics to further address these questions.

DISCUSSION AND CONCLUSION

This paper presents a significant advance in integrating computer science IR methods with questions at the frontiers of strategy research. Strategic management scholars have focused substantial attention on the relationship between knowledge and innovation for firms’ competitive survival. An important strand within the literature examines external knowledge sourcing as a key component of the innovation process. Yet scholars have also been limited in trying to link the knowledge that results from inventive activity to introduction of novel products that result in advantage over rivals. This methodology provides scholars with new tools to reliably and reproducibly disentangle knowledge that is inherently embedded in text, and to deploy this insight to more deeply probe a full range of the innovation process and outcomes.

More broadly, this methodology enables strategy scholars to effectively incorporate language-based metrics as data. This question-focused paper details an innovative methodology to objectively link the underlying knowledge components in patent documents with product innovation. This paper brings an intersection of the insights and techniques from IR—that language itself can be valuable data through mathematical manipulation— with the insight from strategic management that innovation is a complex process involving far from transparent flows
of knowledge from a variety of sources.

External validity of the text matching methodology in related disciplines

While application of this methodology is novel in the strategic management literature, the techniques themselves have been validated in computer science (Manning et al., 2008). Researchers in domains relevant to strategic management recently have started to draw inspiration from this approach as well. In the finance literature, Hoberg and Phillips (2010) develop a similar vector space model to identify product market synergies in mergers and acquisitions, and Hanley and Hoberg (2010) analyze the relationship between IPO prospectuses and pricing. A similar approach was used to quantify firm fundamentals and investor sentiment (Tetlock, 2007; Tetlock et al., 2008). Recently, scholars in the management of information systems literature have turned to vector space techniques to link the similarity between consumer preferences and online consumer rankings (Archak et al., 2011) and to study the boundaries of online communities (Van Alstyne and Brynjolfsson, 2005).

Strengths and limitations

This methodology has both strengths and limitations. The methodology works well to study innovation in a setting such as the medical device industry, where documents are available to trace the knowledge heritage from patent to product innovation. More broadly, this methodology is well suited to industries that utilize patenting and where there are readily available descriptions of product innovations. Thus, this might easily be deployed in the context of the oft-studied pharmaceutical and biotechnology industry (Cardinal, 2001; Dunlap-Hinkler et al., 2010; Higgins and Rodriguez, 2006; Rothaermel, 2001).

On the other hand, in industries that do not rely on patents, or in complex technology industries, measures of invention take greater care to identify. However, while this same criticism might be applied to existing approaches that rely on patent counts and citations, it is
helpful to also note that the flexibility of the methodology here is that it can be applied to any text document (not just patents). The second example provides insights in this direction. As the expansion of digitized text documents continues to explode, it is easier to trace the digital trail of knowledge flows beyond patents, e.g., in technology blogs, online forums, LinkedIn data, interview transcripts, IPO documents, and so on. Likewise, the methodology can be extended to scientific, technical, and professional publications to capture measures of invention outside of patent documents.
REFERENCES


Google I. 2013b. Natural language processing at Google.


Medtronic Inc. 2014b. Questions and answers about DBS for OCD.


Singh JN, Dwivedi SK. Analysis of Vector Space Model in Information Retrieval. In Proceedings of the National Conference on Communication Technologies & its impact
on Next Generation Computing CTNGC 2012. Available at.


## Table 1. Cosine Similarity Scores for Medtronic PMAs: Medtronic and Image-guided Neurologics Patents

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6902569</td>
<td>0.06751</td>
<td>0.03692</td>
<td>0.07123</td>
<td>0.05452</td>
<td>7497863</td>
<td>0.03312</td>
<td>0.02061</td>
<td>0.03025</td>
<td>0.03394</td>
</tr>
<tr>
<td>6902569</td>
<td>0.06751</td>
<td>0.03692</td>
<td>0.07123</td>
<td>0.05452</td>
<td>7517337</td>
<td>0.09224</td>
<td>0.06819</td>
<td>0.07859</td>
<td>0.19924</td>
</tr>
<tr>
<td>6793664</td>
<td>0.08074</td>
<td>0.04134</td>
<td>0.08441</td>
<td>0.07092</td>
<td>7497857</td>
<td>0.02437</td>
<td>0.02113</td>
<td>0.02625</td>
<td>0.02083</td>
</tr>
<tr>
<td>7204840</td>
<td>0.10897</td>
<td>0.04047</td>
<td>0.05788</td>
<td>0.05105</td>
<td>7366561</td>
<td>0.11495</td>
<td>0.05016</td>
<td>0.10219</td>
<td>0.06396</td>
</tr>
<tr>
<td>Average</td>
<td>0.08118</td>
<td>0.03891</td>
<td>0.07119</td>
<td>0.05775</td>
<td></td>
<td>0.06617</td>
<td>0.04002</td>
<td>0.05932</td>
<td>0.07949</td>
</tr>
</tbody>
</table>

* Image-guided Neurologics patent titles cited by Medtronic patents
  - 6902569 Trajectory guide with instrument immobilizer
  - 6793664 System and method of minimally-invasive exovascular aneurysm treatment
  - 7204840 Deep organ access device and method

* Medtronic focal patent titles
  - 7497863 Instrument guiding stage apparatus and method for using same
  - 7517337 Catheter anchor system and method
  - 7497857 Endocardial dispersive electrode for use with a monopolar RF ablation pen
  - 7366561 Robotic trajectory guide
Table 2. Word Stem tf-idf Scores for Medtronic PMA 960009 and Activa-Image-Guided Neurologics Pair

<table>
<thead>
<tr>
<th>PMA 960009</th>
<th>Image-guided Neurologics Patent #</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Activa-Deep Brain Stimulation)</td>
<td>6793664</td>
</tr>
<tr>
<td><strong>Word stub</strong></td>
<td><strong>w_i = tf-idf</strong></td>
</tr>
<tr>
<td>stim</td>
<td>0.53366</td>
</tr>
<tr>
<td>system</td>
<td>0.47544</td>
</tr>
<tr>
<td>impl</td>
<td>0.29030</td>
</tr>
<tr>
<td>deep</td>
<td>0.26538</td>
</tr>
<tr>
<td>control</td>
<td>0.26363</td>
</tr>
<tr>
<td>brain</td>
<td>0.26140</td>
</tr>
<tr>
<td>elect</td>
<td>0.21676</td>
</tr>
<tr>
<td>trem</td>
<td>0.21628</td>
</tr>
<tr>
<td>lead</td>
<td>0.11395</td>
</tr>
<tr>
<td>neurostim</td>
<td>0.08967</td>
</tr>
<tr>
<td>extend</td>
<td>0.08887</td>
</tr>
<tr>
<td>chang</td>
<td>0.06958</td>
</tr>
<tr>
<td>kit</td>
<td>0.06789</td>
</tr>
<tr>
<td>ex</td>
<td>0.06172</td>
</tr>
<tr>
<td>access</td>
<td>0.05273</td>
</tr>
<tr>
<td>process</td>
<td>0.04687</td>
</tr>
<tr>
<td>ad</td>
<td>0.03703</td>
</tr>
<tr>
<td>mod</td>
<td>0.03555</td>
</tr>
<tr>
<td>magnet</td>
<td>0.03555</td>
</tr>
<tr>
<td>new</td>
<td>0.03555</td>
</tr>
</tbody>
</table>

Cosine similarity
PMA 960009: DBS | 0.08074 | 0.06751 | 0.10897
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6892083</td>
<td>0.46859</td>
<td>0.23934</td>
<td>0.15776</td>
<td>0.19883</td>
<td>0.17822</td>
</tr>
<tr>
<td>6901255</td>
<td>0.46376</td>
<td>0.25666</td>
<td>0.20004</td>
<td>0.22942</td>
<td>0.24150</td>
</tr>
<tr>
<td>7190802</td>
<td>0.13127</td>
<td>0.06790</td>
<td>0.06440</td>
<td>0.07134</td>
<td>0.08875</td>
</tr>
<tr>
<td>7206594</td>
<td>0.25715</td>
<td>0.19701</td>
<td>0.17371</td>
<td>0.18186</td>
<td>0.23226</td>
</tr>
<tr>
<td>7248881</td>
<td>0.46395</td>
<td>0.23926</td>
<td>0.16564</td>
<td>0.21143</td>
<td>0.18847</td>
</tr>
<tr>
<td>7257415</td>
<td>0.46780</td>
<td>0.24362</td>
<td>0.16834</td>
<td>0.21270</td>
<td>0.19417</td>
</tr>
<tr>
<td>7310541</td>
<td>0.46140</td>
<td>0.23772</td>
<td>0.16063</td>
<td>0.20489</td>
<td>0.17919</td>
</tr>
<tr>
<td>7457751</td>
<td>0.32279</td>
<td>0.18118</td>
<td>0.13118</td>
<td>0.15673</td>
<td>0.18608</td>
</tr>
<tr>
<td>7764972</td>
<td>0.33194</td>
<td>0.22582</td>
<td>0.16818</td>
<td>0.20421</td>
<td>0.21323</td>
</tr>
<tr>
<td>7953447</td>
<td>0.45202</td>
<td>0.24303</td>
<td>0.16810</td>
<td>0.21724</td>
<td>0.18957</td>
</tr>
<tr>
<td>7978619</td>
<td>0.25356</td>
<td>0.25417</td>
<td>0.17135</td>
<td>0.18720</td>
<td>0.21074</td>
</tr>
<tr>
<td>8098806</td>
<td>0.44184</td>
<td>0.23946</td>
<td>0.13303</td>
<td>0.16868</td>
<td>0.17696</td>
</tr>
<tr>
<td>8121649</td>
<td>0.44336</td>
<td>0.23905</td>
<td>0.16898</td>
<td>0.22102</td>
<td>0.18754</td>
</tr>
<tr>
<td>8175887</td>
<td>0.32470</td>
<td>0.18820</td>
<td>0.13722</td>
<td>0.15894</td>
<td>0.18541</td>
</tr>
<tr>
<td>8498865</td>
<td>0.42122</td>
<td>0.26889</td>
<td>0.16400</td>
<td>0.22929</td>
<td>0.21349</td>
</tr>
<tr>
<td>8626246</td>
<td>0.45254</td>
<td>0.24387</td>
<td>0.16924</td>
<td>0.21915</td>
<td>0.19203</td>
</tr>
</tbody>
</table>

Average Patent similarity: 0.38487 | 0.22282 | 0.15636 | 0.19206 | 0.19103

Vocera Communications IPO Prospectus: 0.18105 | 0.27202 | 0.25765 | 0.13687 | 1.00000