1 Elevator Pitch

Is your idea already patented? In 10 minutes, you could know! Our software-based service allows you to quickly assess your idea within the context of all U.S. patents. By simply typing in a few key words and/or uploading a description of your idea, our software will quickly and automatically perform a thorough patentability search for you. Your idea will be compared to all of those patented by the U.S. Patent office, including over 10 million patent documents, and the most relevant results will be filtered and presented to you. Then, using sophisticated network mathematics, we will compare your idea directly to the subset of the most relevant patent documents and our software will give you a top-20 of "nearest neighbors"—patents that are most closely related to your idea. From there you can determine whether your idea has already been patented either in full or in part, who your most likely competitors and/or business partners will be if you choose to patent your idea and build a business around itor perhaps you might choose sell your idea to one of the owners of the related patents identified by our software. Our software will arm you with the critical information that you need to bring your ideas to life and, more important, to capitalize on them.

Paid search services take weeks (or longer) to perform and the result is merely one professional's opinion. Current patent-search approaches are heuristically driven, described by many as an art: one requires an in-depth knowledge of patent jargon, appropriate technical keywords, and so on. The key benefit of our search service for customers would be that they could "know now" using rigorous scientific methods. In 10 minutes, using our do-it-yourself-initially search methodology, customers could be informed and educated. Minimum costs[1] for patent professionals: an hour of time ($300), a basic patentability search opinion ($2,000), a basic patent filing ($10,000), total single-patent costs ($40,000). *Before* proceeding with such costs, for less than $100, an inventor can immediately be informed about competitive patents and patent applications that already exist in relation to his/her idea. Therefore, the direct customer would be the inventor that wants to ascertain if his/her idea is worth pursuing as a patent. Indirect customers could include venture capitalists (ascertain the IP landscape of an entrepreneur's proposal), patent professionals (patent attorneys developing legal opinions) and patent managers (e.g., a university's technology transfer team needs to ascertain if a professor's invention disclosure is worth pursuing as a patent before the inventor's research is disclosed at a conference). The key features of the innovation: (1) the search is performed on an entire idea, not just a few keywords – literally, a 200-page dissertation could be copied and pasted into the search box, (2) advanced natural-language pre-processing is performed on the idea to determine key terms and concepts, (3) several independent smart searches are performed to define a subset of potential "nearest-neighbor" patent documents, (4) this subset is further analyzed using singular value decomposition to extract relevant concepts and ascertain document correlations, (5) synthesized reports based on these searches and analyses are created.

Truly, what we are proposing will make patent search *better, faster, cheaper.*

[1] Extrapolated from two sources (Quinn 2015, Auvil and Divine 2011).
Proprietary information is highlighted throughout this project proposal.

2 Commercial Opportunity

There is a broad societal need for synthesized concept search to answer a fundamental question: is my idea patentable? Or, what patent documents correlate with my idea, my patent, or my product offering? We define our potential customer as a searcher with limited resources to spend on patent analytics, but 'information on a budget' is critical to them. Is their idea patentable? Should they worry about the 'cease and desist' patent infringement letter they received?

The 'promotion of the progress of science' (U.S. Constitution Article 8, Section 1) and the 'encouragement of learning' (1790 Copyright Act) are vital pillars of this country's strategy for economic development and are embedded within the mission of the National Science Foundation. Since inception, over 8,625,000 utility / 675,000 design / 24,000 plant patents have been issued by the United States Patent and Trademark Office (USPTO). In 2013, more than 270,000 patents were issued, of which, over 75,000 were assigned to top-50 multinational companies. By the end of 2016, this innovation network will surpass 10 million patents plus nearly 4 million non-patented application documents. Objective tools to diagnose the health of this innovation pipeline by measuring dynamics of the network of patents, inventors, firms, lawyers, examiners, technologies, and countries would be instrumental in nurturing and fostering innovation in sophisticated R&D research labs and in an entrepreneur's garage. Organizing the wealth of information available about patent innovations, creating artificial-intelligent algorithms to extract meaning from this evidence, and sythesizing these analytics into actionable business intelligence has the potential to reduce asymmetry of information and bring 'transparency to intellectual property.'

Our market is a global market – anyone that has an interest in securing an idea as a patent within the U.S. patenting system. Over 250,000 new patents are issued every year, with an average patent having about three inventors. Five out of six patents are assigned to a company. Over one third of those are assigned to the top 50 multinational companies (IBM, Samsung, Canon, Sony, Microsoft). This means that the remaining two-thirds (nearly 150,000 patents every year) are assigned directly to inventors or smaller businesses. We identify this group as the un(der)served market. For each patent issued, there were two applications filed, meaning that applicants gave up on patenting their idea for several reasons; one of which, we posit, is that the patent examiner identified that this idea already is described in existing patents. Our search service could have informed the inventor much sooner *before* sunk costs were incurred.

How many entrepreneurs/inventors have a good idea, and want to know if the idea is patentable? Not every inventor (or perceived inventor) has a patent or patenting experience. Everyone perceives they have an idea new to the world, which makes 'is my idea patentable' potentially a general niche market. Every year, there are 17 million students attending U.S. colleges and universities. Some are in business schools, getting degrees in marketing or entrepreneurship. Others are in technical schools, getting degrees in engineering, biology, chemistry, and so on. Both groups have an interest in patents, patentability, and the value of innovations. Within granting agencies, like the NSF, thousands of proposals are submitted every year, and the ability to valuate the technological innovation objectively would have obvious benefits to the review process.

We offer an ultra-conservative size of our direct market to be about 4 million users and further decompose it in red-ocean (highly competitive, cost-prohibitive), blue-ocean (unserved market, see Kim and Mauborgne (2015)) and purple-ocean (partially-served markets). We want to capture a portion of this niche, un(der)served market ("do-it-yourself-initially" solutions) mainly inventors and small business owners.

Inventors	2,125,000
Attorneys	1,250,000
Academics Using Patent Data	30,000
Management Consultants	100,000
IP managers	20,000
IP Executives	50,000
Other Executives	500,000
	4,075,000

Unserved Market (Blue Ocean)
Underserved Market (Purple Ocean)
Competitive Market (Red Ocean)

Figure 1: Market Size

The inventor is on his/her computer late at night, thinking about this great idea and how it can change the world; he/she Google searches 'is my idea patentable' and sees for $99 (MyPatentIdeas.com) the question can be immediately answered using "smart-search" technologies. A small business owner receives a legal threat of patent infringement, demanding 'cease and desist' and a payment of $500,000 for infringement. This letter strikes fear in the heart of the small business owner. Before going to a patent attorney and spending thousands of dollars there as well, he/she wants to know how his/her technology is infringing. Financial stress is a key psychographic motivator (whether I am afraid I must pay a 'patent troll' or maybe I am in dire need of finding revenue streams for my technology), and market intelligence is what we offer with MyPatentIdeas.com (I want to know something RIGHT NOW that helps relieve my stress). The business owner can simply go to MyPatentIdeas.com and enter a description of his/her technology to see what's going on.

We anticipate selling our online software-as-a-service (SaaS) offerings directly to "IP knowledge workers" initially targeting those who want "do-it-yourself-initially" solutions (a la LegalZoom). Customers want initial answers immediately, independently, and anonymously.

We have performed direct market research by talking to hundreds of inventors and patent[2] attorneys. We have also talked to CEOs of several mid-range companies to ascertain their IP-management needs that can be answered using search-input/patent-document correlations. This touches on "freedom-to-operate" searches. Even offices of technology transfer at major universities have an interest in using the proposed technology if implemented. In sum, the relevant market is interested. Our goal is to launch a simple search service, one search per token (e.g., for $99). As outlined in our letters of support, we will initially give away these tokens to small group of interested stakeholders, to beta-test the system, and provide feedback for enhancements. From these conversations, we may consider other product-pricing models: bulk token purchases, per-seat subscription services, and so on. Due to the proprietary nature of the inputed searches, we will discuss with potential customers their concerns regarding privacy and confidentiality which may lead to a new manifestation of the technology in the form of an onsite, private server. The search-engine would update as new patent documents are released weekly, but all reports and search history would be stored

[2]We have a database of all registered patent attorneys [approximately 60,000], with their phone numbers and email addresses. During the summer of 2013, we had 10 sales interns, directed by Ted McGuire, call many of these patent attorneys. In this process, we flushed out MyPatentIdeas.com as a top priority for a product offering as research funds to implement the technology become available.

internally on a machine physically located within a firm's facilities. The initial "proof-of-concept" commercialization approach will go as follows: (1) we build a prototype search-reporting interface; (2) we provide tokens for free searches for interested parties: entrepreneurs, university IP managers, corporate IP managers, and patent professional associations; (3) we develop an ecommerce "pay-as-you-use" system to purchase search tokens for money; (4) we update and enhance our product description pages (e.g., http://www.mypatentideas.com/) as we implement the search-reporting technology; (5) we collect feedback on the service, and research optimal product-pricing strategies based on searcher's willingness to pay.

There are competitors in this market. There are several free search tools, such as the USPTO and Google Patents. These tools allow for advanced[3] boolean search, and Google has even suggested a "prior-art finder." These services are limited: (1) they only use one search-ranking methodology so they may miss important results, (2) they cannot account for contextual concept issues related to term ambiguity (e.g., *strike* or *car* vs. *authomobile*) hereafter described, (3) they do not synthesize the results into an informative report. Google has developed a very fast search interface for the internet, and has applied this search tool to patents. Using one search method, Google gives a searcher thousands of results in less than one second. Is this what the searcher wants? We think not. We posit that the searcher wants a comprehensive synthesis of all related documents. The searcher wants to be informed and educated about how his/her idea relates to existing patent documents. The searcher enters a document, and our system performs multiple searches to make certain all relevant documents are included; then, our service performs conceptual analyses to determine document correlations; finally, our service synthesizes the search/analysis into a meaningful report. There is another layer of competitors that offer per-seat licenses of their patent analytic products. Some of these competitors even offer some sort of concept-search feature within these enterprise-level solutions. These tools provided by companies like Thomas Innovation, IBM, PatentCafe now part of Pantros IP, and Innography are intended for large corporations based on expensive pricing structures (e.g., around $50,000 per year for a single seat). Our $99-per-search price will enable us to be a disruptive innovation and low-cost leader (Christensen 1997) thereby providing a service to the small business owners and innovators that is currently not available to them.

If you currently perform a Google search "is my idea patentable" you will find several resources enabling the searcher to begin the patenting process (patenting kits). You will also find information on "do-it-yourself" patentability search using basic text-matching search services such as the USPTO or Google. However, if you carefully scrutinize the market need and the resulting products and services, you can readily see that the fundamental question is not answered–I have an idea and I want to know if the idea is already patented *before* I buy a patenting kit. I am not an expert on patent-language jargon, nor do I want to be. I merely want to enter my idea into a search box, and be informed and educated about my idea in relationship to existing patents. Now, the market does not readily offer a document-correlation patent-search service. Five years ago, this was also the case. Five years from now, we anticipate this will still be the case. The large incumbents are

[3] We again emphasize that there is a limitation of terms allowed in a query: see `https://support.google.com/gsa/answer/4411411#requests`.

not incentivized to target this un(der)served market. Part of our commercialization effort will be to get organic traffic to find our website. For example, using various search-engine-optimization[4] tools, we conclude that conservatively, there are 20,000 daily searches for this phrase. This search space is highly competitive, but none of the competition answers the question with an online web-application solution. We believe this implies a strong market potential. First, can we achieve 1000 visitors per day through organic search? This would save us daily about $6000 over having to pay for placed advertisements using Google AdWords. Second, from those who visit our page, can we get a 10% conversion rate? We have developed some live-chat functionality into our website, so the searcher can ask questions to a sales team enabling us to close the deal. If the price per unit is $99, with 5% + 10% sales commissions paid to an independent online sales staff, after the technology is developed, we foresee $84 of revenue per sold search. In the above goal-based scenario. 1000 visitors leads to 100 daily searchers, generating $8400 in daily revenues which is $3M in annual revenues. Obviously, if we create repeat business, and offer other product-pricing options, we believe that the *initial* impact of this technology could be $10M per year.

We anticipate that we have all of the necessary technical resources to develop this prototype. We own five servers with a total storage capacity of 150TB which can easily store all 10 million patents and the millions of patent-application documents. Across these machines we have over 100 cores (CPUs) and either 64MB or 128MB of RAM in each machine. Additionally, we have one GPU card (which has thousands of computing cores) which can be used for general-purpose computing of matrix computations. On the software side of things, we have developed alpha prototypes of all of the elements needed to make this product offering possible. We will utilize this project resources to scale these alpha components into an integrated system.

3 Innovation

Published patent documents represent "prior art" in the determination of future patent publications. As such, understanding patent-document correlations is important to stakeholders in innovation and its commercialization. Many times, these stakeholders have a document of interest (e.g., an invention disclosure or a patent publication) and they want to understand which patent documents are most related to this focal document (idea). Such a technology would enable them to address fundamental strategic questions: 'is my idea patentable?' or 'what is the competitive landscape of this focal patent?' These questions have implications for: (1) patentability, (2) infringement, (3) freedom-to-operate, and other important issues.

[4]If you do a current Google search of this phrase "is my idea patentable", you should see our blog website appearing on page 2, with a few pages ranked therein. We anticipate enhancing these SEO tactics to improve our overall placement thereby driving additional traffic.

To illustrate the utility of our methodology, consider the following sentences:

1. The workers went on *strike* due to unsafe work conditions.
2. The boxers *strike* one another incessantly for 12 rounds.
3. The newcomers *strike* up a conversation with their neighbors.

The word *strike* has a different meaning depending on context. In the first sentence, we infer that *strike* (noun) is the refusal to work as a form of protest. In the second sentence, we infer that *strike* (verb) means to hit. In the third sentence, we infer that *strike* (verb) means to initiate. If we were to use Google's *text-matching* search for the "strike," the search engine has no way to distinguish the multiple meanings from the single search term. Additionally, if two words have a similar meaning (e.g. *car* and *automobile*), results may be missed using Google's search approach. A more robust approach accounts for basic word-sense amibiguity[5] issues.

Latent semantic analysis (LSA) has been developed (Deerwester et al. 1990, Landauer et al. 1998) enabling the extraction of latent (or hidden) semantic structure thereby addressing these word-sense amibiguity issues that text-matching search cannot. Albeit robust, this approach for large document collections is not tractable due to the high-complexity computational restrictions. We overcome this problem by creating a manageable subset of potentially relevant documents (through multiple-search strategies), and then perform a custom LSA on this subset corpus.

For a document collection of n documents and m terms, the singular value decomposition (SVD) necessary to perform LSA has a complexity defined to be: $O(n^2m+m^3)$. The complexity of the post-filter correlation computations using cosine similarity is $O(n^2-n+m)$. As a result of this complexity, performing a full or true SVD on a large document collection (corpus) is intractable (Hofmann 1999, Blei et al. 2003, Laura et al. 2005) meaning in the traditional form, it is unsuitable to search a large collection of documents. Consider, for example $n = 10$ million documents with a total of $m = 250,000$ terms; the complexity of the needed SVD is about 25 quintillion. Our innovation is to use a search methodology based on the idea inputed to drastically reduce n and m. For example, if we can create a subset of $\tilde{n} = 2,000$ documents with a total of $\tilde{m} = 200,000$ terms reduced[6] to 10,000 terms, the new complexity is 24 million *times* more efficient. Something that was intractable can now be delivered to the searcher in a few minutes. If our proposed search-subset LSA (SSLSA) can be performed in one minute, it would take over 63 years to perform a realtime, full-document-corpus LSA.

[5] *Polysemy* is when a single term has multiple meanings. *Synonymy* is when multiple terms share a single meaning. Word Sense Disambiguation (WSD) is one of the fundamental problems in the field of Natural Language Processing (NLP) (Navigli 2009, Agirre et al. 2014).

[6] We further reduce to $\tilde{m}' = 10,000$ terms by constructing the term-document matrix $\tilde{A}$ such that only terms from the query input are included (not the entire corpus). The *query* phase reduces $\tilde{n}$ significantly and $\tilde{m}$ slightly. The *prune* phase reduces $\tilde{m}'$ significantly: we prune to only include terms found in the inputed query. Such pruning defines the concepts of the analysis *based* on the natural language of the author of inputed idea.

The first aspect of our innovative process is the search input, or query. Normally, this is a few key words, or maybe a few attributes. In our proposed technology, the input is an entire document. If you are an inventor with a research paper, you can literally copy and paste the entire document into the search box as the query. If you are a university IP manager ascertaining whether an academic's research paper is worth pursuing as a patented technology, you can copy and paste the entire document into the search box as the query. If you are reviewing a NSF grant for technical merit, you can copy and paste the entire proposal into the search box as the query. The input (*focal query document*) is an entire document, and the fundamental output will be the most closely related patent documents, which we define as *document correlations*. We posit that our specific SSLSA approach is unique and novel to provide valid results within a reasonable response time.

We however, do not suppose that "concept search" for patents is exclusively our idea. In fact, we are aware of several enterprises making such search available by creating approximation-indices of the entire patent corpus, since full-SVD is not possible. They offer search services to large enterprises, and even to patent examiners to determine prior art during the examination process. The traditional approximation technique: (1) an index is created about once a month on a random subset of the entire patent-document collection (and takes several days to compute), (2) the index defines for the entire collection term-concept maps that are applied to the inputed idea, (3) the most relevant documents are retrieved.

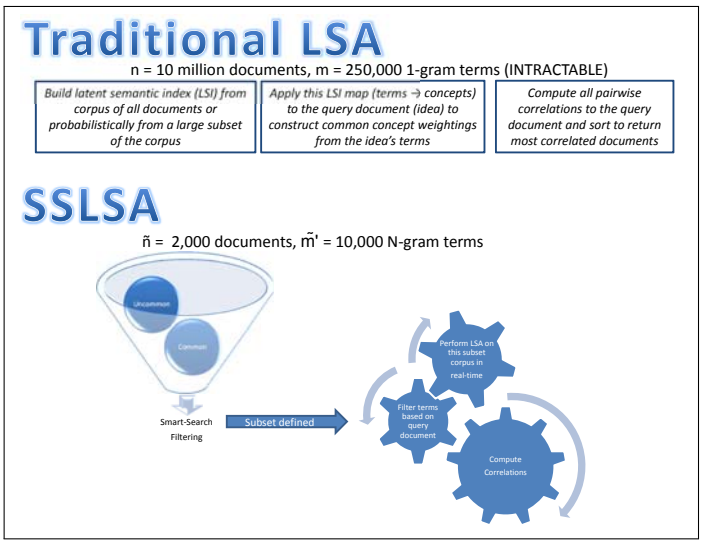

Figure 2: Traditional "tractable" LSA approaches vs. SSLSA

With that disclosure, we emphasize that no one is offering a low-cost, do-it-yourself initially service that will empower the small business innovation researcher. This is the essential aim of our endeavor. We introduce SSLSA as an automated method to determine document correlations from an input query document. The resulting analysis is anchored to the natural language of the inventor, not the overarching language structure of all patent documents. This process is a two-stage approach. In the first stage, we search the entire corpus using several independent N-gram queries and merge these results into a subset. From this subset, we perform SVD on all query-defined N-grams using various filtering techniques. Defining the concepts using SVD based on the focal query document is fundamental to this approach. Rather than having the corpus define the concepts, the focal query (inputed idea) defines the concepts. This is an important nuance as we apply semantic analysis to innovation documents to ascertain a query's novelty. The overall research plan has been designed, and the component elements of the technology associated with the innovation

have been alpha prototyped. The seed funding from this grant would be to: (1) empirically assess the validity of the methodology based on optimal input parameters, (2) develop the updating search methodologies, (3) develop beta reports to the searcher, (4) receive feedback from interested parties to enhance the reports and search interface, (5) develop a working ecommerce platform for the pay-per-search offering, (6) consider other payment options (bulk purchases, subscription based, and so on).

A preliminary provisional patent application was filed in Janaury 2014 regarding a less refined description of this project. We are in the process of filing for a traditional application of this technology. In addition, our cleansed patent-data platform, in its organization, represents a component of a potential competitive advantage. In our Mergent Patent Archive description, we delineate how we have data more comprehensive in its preparation than both the USPTO and the Google patent-search services. Ultimately, our ability to create custom reports based on sophisticated mathematics and sound statistics so that the average inventor can be impressed with both the depth and simplicity of the report, will determine the success of this venture in relation to competitors.

The major technical challenge will be the optimal formulation of the multiple-search process to guarantee a valid nomological capture of potentially relevant patents before SVD is performed. That is, can we perform searches to get a subset of $\tilde{n} = 2000$ documents such that the top, say 200, documents after search-subset SVD (SSLSA) would also be found in the full-SVD document collection? Our simulated experiments detailed in the Technical Discussion will further outline how this validity challenge will be addressed.

4 Company/Team

This research will be carried out by Entrepreneurial Innovation, LLC. (DBA Patent Rank). This company was formed as a Nevada-based LLC, recently converted to a Washington-based LLC, with foreign-entity operating permissions in the state of Kentucky. The company has a NSF code of 6250026089, a DUNS number of 078286980, a tax-identification number of 452562977, and a CCR PIN of PatentR99. We believe the overall Phase I objective can be readily achieved within the twelve-month performance period.

The company was founded by Monte Shaffer, the principal investigator (PI) in 2011. Since inception, Ted McGuire has been the acting CEO of this company. Both the PI and CEO will devote a significant level of effort to the proposed Phase I activities. Monte's efforts will focus on managing and directing the technology development and the overall R&D plan. Ted's efforts will focus on the development of market opportunites: direct B2C sales, relationships with venturing groups and programs, feedback from patent professionals and technical-merit managers at universities and grant-funding institutions. Monte's expertise in internet marketing will support Ted's efforts (e.g., see the market-facing informational website already developed: http://www.mypatentideas.com).

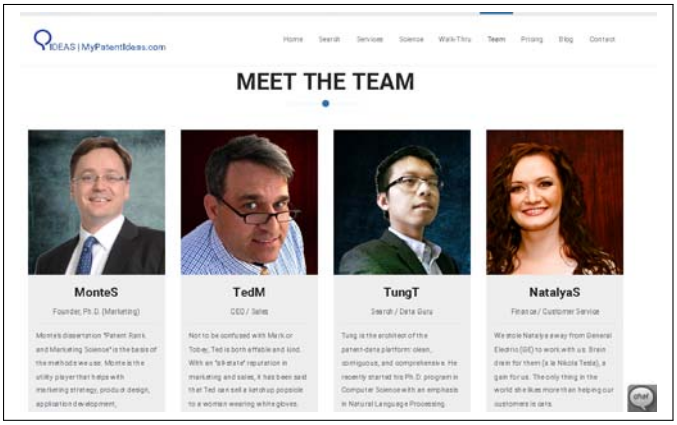

Figure 3: Personnel to Develop Service

The vision of the company Patent Rank is to bring transparency to patent analytics through the development of services that objectively inform stakeholders using sophisticated mathematics and statistics (see branded website: http://www.patent-rank.com). The company has existing operations. The company's primary technology is Patent Rank, an objective value of a patented technology over time. MyPatentIdeas.com (the value of an idea) fits well within the overall vision and direction of the company. Patent Rank is involved as an industry partner with an academic initiative known as CRIE (Commercialization Research on Innovation and Entrepreneurship) - making cleansed patent data more readily available for academic consumption to advance the study of innovation and entrepreneurship (see http://crie.patent-rank.com). Additionally, using the same patent repository, Patent Rank has developed a patent-archival module[7] within Mergent's Archives. Within this aforementioned product, Patent Rank has developed a robust boolean-search methodology that can be leveraged in the development of this concept-search technology.

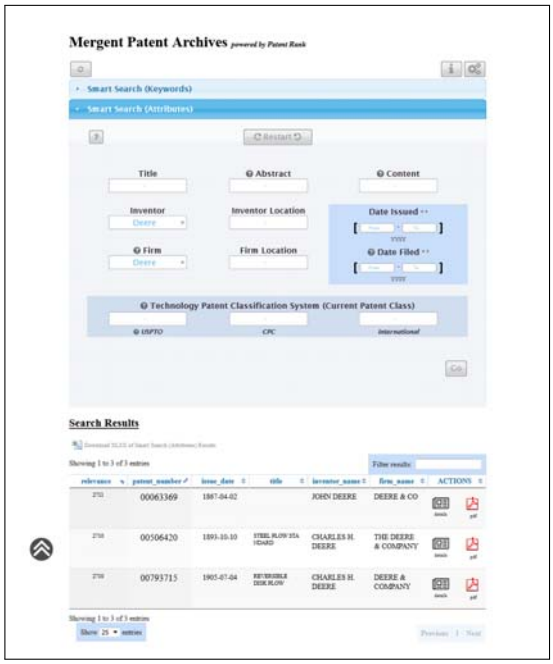

Figure 4: Module of Mergent Archives

Since 2008, Monte has been cleansing and augmenting the patent-data platform. The technical and commercialization team has developed since that time. Mirek Truszczynski has facilitated the advancement of computer-science algorithms into the data cleansing procedures and also facilitates efficient programming objectives performed by the computer-science research students also included on the proposal. Tung Tran is in his second year of a Ph.D. program in Computer Science and has been working on the patent data since the Fall of 2013. His major responsibilities have been the core patent data objects, the search-engine implementations, and the development of natural-language processing methods. Orion Fisher is planning on beginning his first year of a Ph.D. program in Computer Science and has been working on the patent data since the Summer of 2014. Orion has been working on the fundamental SVD matrix computations with efforts of making these computations as efficient as possible utilizing various approaches.

[7] http://www.mergent.com/solutions/print-digital-archives/mergent-archives

The company has had some limited revenues over the past three years. The core valuation technology, Patent Rank, received a NSF SBIR Phase I/Ib Award IIP #1315850 and external funding from Kentucky's Innovation Network to support the development of the patent-data platform. When Monte Shaffer completed his dissertation (Shaffer 2011) and founded this company, he had less than 2 terabytes (TB) of patent data. Now, the company has over 50 TBs of patent data, and the repository is growing weekly as new patent data is made available through the United States Patent and Trademark Office (USPTO). In addition to the grant-funds as revenues, through the development of the relationship with Mergent, patent archives as a module of Mergent Archives has been purchased and installed at major universities in the United States, generating marketing-facing revenues.

5 Technical Discussion / R&D Plan

We define the following:

Stopwords Words filtered out before the processing of natural language text. For example, the "snowball" list: *a, about, above, after, again, against* ...

N-gram A contiguous sequence of N items from a given sequence of text unaffected by the stopword filtering.

Term Any unique N-gram.

Document Sequence of N_t terms denoted by $\boldsymbol{d} = (t_1, t_2, \ldots, t_{N_t})$ where t_i represents the ordinal i^{th} term in the sequence of all terms.

Corpus Collection of n documents denoted by $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n\}$ where d_j represents any specific element in D.

Query Unstructured text q that is stopward filtered and defined using the N-gram terms as the focal input document $\boldsymbol{d}_q$. This document may or may not be part of the corpus.

Common A term t_i from the query that is frequent within the corpus. The most common term t_{C_1}, the second most common term t_{C_2}, and so on.

Uncommon A term t_i from the query that is rare within the corpus. The most uncommon term t_{U_1}, the second most uncommon term t_{U_2}, and so on.

Index A ranked term-frequency scheme derived from the corpus to determine the (un)common nature of an individual term. The index returns the rank (or quantile) if the term exists. If the term is not found, null is returned.

Subset Collection of $\tilde{n}$ documents denoted by $\tilde{D}$ retrieved in (un)common searches where $\tilde{n} \ll n$

[We consider this entire section to be proprietary.]

Traditional LSA (latent semantic analysis) is intractable with millions of documents defined within an information retrieval system. Other approaches that anchor to LSA-methodologies, most notably the probabilistic approximation techniques (Hofmann 1999), build an approximate concept-map. We, however, do not want to lose the ability to determine concepts based on the inputed query document, so we want to perform[8] SVD (singular value decomposition) based on the key phrases from the query *NOT* from the overall concepts of the entire patent collection. For example, if 400 concepts were defined from a corpus of 10 million documents across 250,000 terms, the dimensionality of each concept explains less and less variance. The first concept explains the most variance, the second concept explains the next most variance, and so on. The top concepts in this construction would be more related to the overall requirements of natural language within patent documents. On the other hand, if we were to reduce the document set using our subset corpus methodology *and* we also include the terms based on those that exist within the query document, all concepts will be uniquely defined based on the inputed idea. Although subtle, this is an important feature of our document-correlation strategy. The searcher is informed based on the inputed text, and all concepts are anchored to that custom query request. Our method, SSLSA (subset-search LSA), reduces the dimensionality of the computation significantly. Ideally, we want to demonstrate that this method works without loss of generality. That is, if we performed full LSA and identified the top-20 most correlated documents (based on terms and concepts) relative to the query document, would we get the same top-20 results using our SSLSA methodology. As previously mentioned above, doing so would make the computational complexity *24 million times* more efficient. Mathematically, we define SVD (Deerwester et al. 1990) as:

$$A = USV^T \tag{1}$$

where $A_{n\times m}$ represents the term-document matrix[9] that can be decomposed as the product of three matrices U, S, and V^T where matrix S is a diagonal matrix representing eigenvalues for the term-document system. Necessarily, each eigenvalue is necessarily smaller than the previous and the inclusion of all eigenvalues would perfectly represent a linear combination of a weighting schema of all terms mapped to all documents. Based on these mathematics, one can consider or a fixed number of eigenvalues within S, hereafter defined as k concepts. This is equivalent to zeroing-out all additional singular values of S which also correspondingly affect U and V^T. That is, we can observe U_k, S_k, V_k^T. Furthermore, if we truncate A as $\tilde{A}$ based on our proposed methodologies, we can similarly observe $\tilde{U}_k$, $\tilde{S}_k$, $\tilde{V}_k^T$.

[8] SVD is mathematically equivalent to PCA (principal components analysis).

[9] The matrix has terms as columns, documents as rows. The values of $a_{ij} \in A$ can be either frequency counts (how many times the term appears within the given document) or some normative weighting based on within-document and within-corpus frequencies. The rank of S is $rk(S) = min(n, m)$. Based on these deconstructions and conforming properties of matrix multiplication: $U_{n\times rk(S)}$ represents the document-concept matrix and $V_{m\times rk(S)}$ represents the term-concept matrix.

Innovation Details

Can we develop "smart-search" technologies to ascertain a subset corpus of the entire patent collection such that the subset $\tilde{n} \ll n$, yet the document correlations are just as valid as if we did perform semantic analyses on the entire collection? Although most semantic analyses utilize the "bag-of-words" assumption (Blei et al. 2003), we utilize key N-grams or phrases to more carefully define the subset through an initial meta-search protocol (the *granularity* hypothesis).

We identify phrases as common or uncommon by indexing all phrases in the entire patent collection. For example, the 1-gram phrase "claim" is a required word in all patent documents, so it is very common. A 2-gram phrase "semantic analysis" is likely a term that is not common within the patent collection. A 3-gram phrase "natural language processing" is even more likely a term that is not common within the patent collection. Furthermore, this 3-gram, can be decomposed into smaller N-grams: 1-grams {natural, language, processing} and 2-grams {natural language, language processing}. If we define our search to include all 3-grams and the relevant decompositions, our multiple-search methods should be very precise as a prefilter to create a subset corpus. We posit that *common* searches will capture the general technological space of the idea and that *uncommon* searches will capture the technological specificity of the idea.

For each relevant N-gram, we can query the patent-term index and return quantile thereby ascertaining its (un)commonality. We can similarly perform N-gram analysis of the inputed idea as a query document.

procedure SSLSA(q, D, r, k)
  $\tilde{D} = \emptyset$
  $d_q \leftarrow$NGRAM(q, r)
  while $r \neq 0$ **do** ▷ For r-gram terms
    $s_c \leftarrow$SEARCH$(d_q, r, N_{C_{(r)}}$,common$)$
    $s_u \leftarrow$SEARCH$(d_q, r, N_{U_{(r)}}$,uncommon$)$
    $\tilde{D} = \tilde{D} \cup s_c \cup s_u$
    $r = r - 1$
  end while ▷ $\tilde{D}$ is the subset
  $\tilde{D} = \tilde{D} \cup d_q$;
  $\tilde{A} \leftarrow$FILTER$(d_q, \tilde{D})$
  Compute SVD($\tilde{A}$) for k concepts
  Note $\tilde{U}_k$, $\tilde{S}_k$, $\tilde{V}_k^T$
  Compute pairwise similarities (d_i, d_j) in $\tilde{D}$
end procedure

procedure NGRAM(text,r)
  remove stopwords
  build frequency for each unaffected r-gram
  return d ▷ Document as term-frequency vector
end procedure

procedure SEARCH(d_q, r, n,method)
  (common) from index: OR top-n r-grams
  (uncommon) from index: OR bottom-n r-grams
  return s ▷ Most relevant results (results ≤ 250)
end procedure

procedure FILTER($d_q, \tilde{D}$)
  Construct Term-Document Matrix of $\tilde{D}$
  Filter by terms found in d_q
  return $\tilde{A}$
end procedure

Figure 5: SSLSA Algorithm

Now, we can describe commonality and uncommonality at the patent-collection level (the corpus) and at the inputed-idea level (the query-document). Utilizing these features, can we develop independent search methodologies to nomologically capture all relevant patent documents such that the most correlated documents (e.g., top-20) are always obtained? This is the optimality constraint that will make this technological innovation very meaningful to the market. Once we have appropriately captured the subset, we can create a custom LSA based on the inputed idea, and report the most-related documents. The terms of the custom LSA are defined by the inputed idea, and various weighting mechanisms can be developed based on local/subset/global corpus frequencies.

Above, in Figure 5, we outline the basic algorithm design for our prototype fast-document correlations innovation with inputs: query q, document corpus D, number of N-grams r, number of concepts k. First, we filter out very common words, known as stop words. Next, we create N-grams based on the query input. For each N-gram, we perform an independent (un)common search based on a predefined policy. We perform independent searches, and merge[10] the resulting set as subset corpus $\tilde{D}$. At this stage, we can also summarily report on these multiple-search[11] results. We can also build nomological word clouds based on the (un)commonalities of the idea searched and patent documents. Finally, we proceed to perform SVD to identify concepts of this subset corpus. In the synthesized report, we can create word clouds describing the most important concepts educating the searcher on the technological space using terms from patented technology descriptions. In addition, we can perform pairwise correlations which will enable us to visually graph the result set (based on say the first two dimensions or concepts). From this graph, we can create a document map. On this map, we can place a star (your idea is here) and overlay technology classifications of patent documents. We can then do an inset zoom feature, and describe the most relevant documented patents.

Research Objectives and Technical Tasks

The overall research objective is to validate the nomological capture of this SSLSA approach. If successful, the web application can be prototyped.

[Task 1:] **Data feeds.** Automation of downloading new patent data (weekly, every Tuesday); parsing and cleansing the data; Automation of downloading new patent applications (weekly, every Thursday); parsing and cleansing the data; update search and commonality indices.

[Task 2:] **Optimal Nomological Search (Experiment).** Run simulated experiments to ascertain the optimal nomological capture. More details follow.

[Task 3:] **Optimal SVD / correlation computations.** Optimize matrix computations across multiple machines (develop a master/slave work flow system

[10] The subset corpus will be defined as a union of all independent searches.

[11] This is, as-if we went to Google and performed multiple searches and combined all top results.

so results can be returned in two stages, with a goal that the total response time for this custom information-retrieval process be less than 10 minutes).

[Task 4:] **Synthesized reports (e.g., Mirek's personality[12] report).** Develop a summary report of the search results and an advanced report educating the searcher on the relevant concepts and nearest-neighbor patent documents.

[Task 5:] **Ecommerce web application** Build a website that is secure (SSL) with payment options, and the search-reporting interface.

Optimal Nomological Search (Experiment).

The entire U.S. patent corpus contains approximately 10 million patent documents. The fundamental validity question is: "if we use a search methodology to reduce the SVD dimensions, do we still capture all potentially correlated documents?" We define this validity question as *comprehensive nomological capture.* To demonstrate the validity of SSLSA, as part of our simulated-experiment design, we have created a patent corpus: we first searched for all patents invented by Kia Silverbrook, the most prolific inventor in the world. We obtained over 4500 documents. We next randomly selected a utility patent granted in the same year as a corresponding Kia patent. Taken together, we define this sample as our corpus D. We posit that the Kia patents should necessarily be more closely correlated to each other, since they represent patents from the same inventor which we posit uses the same natural-language preferences. They represent an adjacent neighborhood of similar patents. Conversely, the random sample should consist of patents not closely correlated.

Within this sample corpus, we have over 9000 documents, and full SVD can be computed (that is, it is tractable). We will perform full SVD to define "true similarity" among documents. This introduces the first parameter selection r, the number of N-grams to include in defining the term-document matrix A. We will run simulations to test $r = 1$ (similar to bag of words) up through $r = 5$ where 5-grams and corresponding 4-grams, 3-grams, 2-grams, and 1-grams are defined as terms (the *granularity* hypothesis of linked phrases). The second parameter selection k represents the number of concepts to be extracted. Prior research suggests that an optimal range for $k = [50, 1000]$ (Landauer et al. 1998); in our simulated design, we will loop over values for $k = 50, 100, 250, 500, 1000$. At the full-SVD level, we therefore have 25 simulation-combinations to perform (5 levels of N-grams and 5 levels of concepts).

Several features within a simulation can also be manipulated, in search of optimality. Features include: the minimum frequency for a term to be included in the matrix, whether the terms are pruned based on the query or if all terms are included (the *inventor's-language* hypothesis), how the term-document matrix is constructed (various weighting methods), how truth is defined (top-10, top-20, top-50, top-100), the number of results returned from a single search (top-50, top-100, top-250, top-500), the number of (un)common search terms that are included in a single search

[12] We have experience building advanced reports: http://www.profiletraits.com/a/report/self/RkouG9janggAAOoEhvs/

(current options selected are the same for both (un)common searches: $N_{C_{(r)}} = N_{U_{(r)}} = 10$ or $N_{C_{(r)}} = N_{U_{(r)}} = 5r + 5$ or $N_{C_{(r)}} = N_{U_{(r)}} = 2^r + 5$), and so on. In total, we have over 5000 simulation-scenarios. For a single simulation-scenario, we will perform 100 simulations (a total of 500,000 trials). We estimate that running on 90+ cores (CPUs), these simulation-scenarios will take approximately 20 days to run.

We randomly select a document from the corpus D, and label it the search query. We note if this document was part of the Kia subset or not. Given the specific parameters of the scenario, we perform the necessary SSLSA and we compute all pairwise correlations for the query document q and the subset corpus $\tilde{D}$. Sorting on these correlations, we identify the nearest patent neighbors. At the full-SVD layer, we do simular correlations to identify what we define as *true* similarity. Did we lose any top results in our SSLSA approach? To ascertain this, we define potential "nomological capture errors." We define Type I error as the percentage of excluded documents found in nearest neighbors of D but not found in $\tilde{D}$. We define Type II error as the percentage of included documents found in $\tilde{D}$ but not found in nearest neighbors of D. As previously mentioned, we define these errors for various levels of correlation (top-10, top-20, top-50, top-100). We may attempt to develop a hypothesis test of such misclassifications as an extension of the work of Dasgupta et al. (2015).

Timeline

We anticipate this research and development (R&D) project to take 12 months from July 1, 2016–June 30, 2017. The first two months will be used to prepare the initial data feeds, and design the experiments to ascertain the optimal parameter selection for nomological capture. Thereafter, the ecommerce web-application will be built in tandem with the custom reporting features. We anticipate that about six months into this project (December 2016), we will have a working prototype that we can allow interested parties to beta-test. In the first six months, we believe over 80% of the project will be complete, but common to pareto problems, the refinement to finish the prototype (remaining 20%) will take an additional six months. We have already developed a commercial application called Patent Rank database and are aware of the difficulties and obstacles in developing a new commercial application. Also, we succeeded in developing our commercial application with a budget that was slightly less than the current proposed one. Hence, we are confident that our team can overcome the various obstacles that invariably occur in this process.

6 Conclusion

We have a plan. We have the patent data. We have the computing resources. We have the know-how. We have the ambition. We now need the funding to make this happen. With the appropriate funds, we will be able to accelerate the research and commercialization process to make MyPatentIdeas.com a successful and valid patent-search service.