IDENTIFYING EMERGING RESEARCH FIELDS WITH PRACTICAL APPLICATIONS VIA ANALYSIS OF SCIENTIFIC AND TECHNICAL DOCUMENTS

Patrick Thomas\textsuperscript{1}, Olga Babko-Malaya\textsuperscript{2}, Daniel Hunter\textsuperscript{3}, Adam Meyers\textsuperscript{4} and Marc Verhagen\textsuperscript{5}

\textsuperscript{1}pthomas@1790analytics.com
1790 Analytics LLC, 130 North Haddon Avenue, Haddonfield, NJ 08033, USA

\textsuperscript{2}olga.babko-malaya@baesystems.com
BAE Systems AIT; 6 New England Executive Park, Burlington, MA 01803, USA

\textsuperscript{3}daniel.hunter@baesystems.com
BAE Systems AIT; 6 New England Executive Park, Burlington, MA 01803, USA

\textsuperscript{4}meyers@cs.nyu.edu
New York University, Computer Science Dept, 719 Broadway, New York, NY 10003, USA

\textsuperscript{5}marc@cs.brandeis.edu
Brandeis University, Computer Science Dept, MS 018, Waltham, MA 02454, USA

Abstract
This paper outlines a system designed to determine whether practical applications exist for research fields, particularly emerging research fields. The system uses indicator patterns, based on features extracted from the metadata and full text of scientific papers and patents, to assess different characteristics that point to the existence of practical applications for research fields. The system may thus help determine whether a particular research field has moved beyond the early, conceptual phase towards a more applied, practical phase. It may also help to classify emerging research fields as being more 'technological' or more 'scientific' in nature. The system is tested on data from a number of research fields across a range of time periods, and the outputs are compared to responses from subject matter experts. The results suggest that the system shows promise, albeit based on a relatively small data sample, in terms of determining whether practical applications exist for given research fields. The system also shows promise in detecting the transition from absence to existence of practical applications over time, which may be of particular value in evaluating emerging technologies.

Approved for Public Release; Distribution Unlimited
Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5) and Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction
Emerging technologies are of great interest to a wide range of stakeholders. These include government agencies looking to fund promising new ideas, corporations hoping to gain a foothold in a rapidly emerging field, and investment institutions seeking returns from early investments in key innovators. Emerging technologies have also been a focus of academic research ever since Schumpeter (1912) coined the term ‘creative destruction’ to describe the emergence of new technologies, which spawned new industries while destroying old ones. More recently, Christensen & Bower (1996) used the term ‘disruptive technology’ to describe a new development that disrupts the status quo in an existing technology.

Despite this widespread interest in emerging technologies, identifying such technologies remains problematic. The problems are both theoretical and practical. The theoretical issue is how to recognize an emerging technology, without a clear definition of what constitutes such a technology. As noted by Goldstein (1999), there is a lack of precision in the meaning of emergence, and even greater ambiguity about how it occurs. The practical issues result from the sheer scale of information available, especially in the electronic age. Researchers and analysts searching for interesting new technologies face the unenviable task of locating meaningful signals among this mass of information. Their task is not aided by the fact that the number of truly emergent technologies is dwarfed by the number of mature, mundane, or failed technologies.

In an effort to address both the theoretical and practical issues associated with locating and characterizing emerging technologies, the authors are pursuing an automated system directed to these issues. The system processes very large collections of scientific publications and patents, and extracts features from the full text and metadata of these publications and patents (rather than solely from metadata, as is the case with products such as Thomson’s ESI and InCites, or Elsevier’s SciVal). It then constructs quantitative indicators based upon these extracted features. These indicators are designed specifically to locate and characterize emerging scientific and technological fields.

The system thus has similar goals to the European PromTech project, which also endeavours to locate emerging technologies via analysis of scientific literature (Roche et al, 2010; Schiebel et al, 2010). It is also similar to recent work by Érdi et al (2013), who used a graph-based approach to locate new connections between disparate technologies via citation links. These new connections, which can be identified at the point when patents issue, are regarded as being indicative of a possible emerging field or technology. This finding is similar to that reported in earlier research by Schoenmakers and Duysters (2010), who found that radical inventions often resulted from the combination of knowledge from domains that are not otherwise extensively connected.
In this paper, we use the system to examine one specific aspect of emerging technologies – the extent to which they have exhibited the potential for practical application. The existence of a practical application may be a reflection of the increasing maturity of a particular technology, as it moves beyond the purely conceptual stage. Conversely, the lack of a practical application may signal a technology that is still in its early, pre-emergent stage. The existence of a practical application, or the lack thereof, may also be a reflection of the research field itself. Applied research fields (such as mechanical technologies or information technologies) may inherently have greater potential for near-term practical application than theoretical and basic science research fields (such as theoretical physics or mathematics). The latter may attract increasing interest from researchers over time, and thus be defined as having emerged, without demonstrating a near-term practical application. In simplistic terms, the presence of a practical application may thus point to an emerging research field that is more ‘technological’ rather than ‘scientific’ or, in traditional terms, more ‘applied’ rather than more ‘basic’ (Stokes, 1997).

This paper contains four further sections. In the first of these sections, we outline the theoretical foundation for our system. We then describe how, guided by this foundation, we construct indicators designed to determine whether emerging technologies have demonstrated a practical application. These indicators are based on features extracted from the metadata and full text of scientific publications and patents. Having outlined the indicators, we then demonstrate how they are combined via Bayesian networks to optimize the model of practical application. Finally, we show the results of applying this model in practice to sets of scientific publications and patents associated with eight sample technologies, both emerging and non-emerging. These document sets are referred to as Related Document Groups (RDGs).

The results demonstrate the extent to which the outputs of the model concur with the opinions of subject matter experts (SMEs) regarding the presence or absence of practical applications for the eight sample technologies. The analysis is carried out across six time periods. This temporal aspect is necessary, since a practical application may not exist for a given a technology at one point in time, only to develop in a later time period.

**Theoretical Background**

The theoretical foundation for our system is provided by actant network theory (Latour, 2005). This theory provides a vision of science and technology as constituted by networks of heterogeneous elements, interconnected by disparate relationships. These networks do not just contain individuals, but also institutions, instruments, practices, terminology, materials, funders, meetings, government organizations, laws, journals, patents, publications, and so on. The membership of elements within such a network, and the nature and extent of the relationships between these elements, is dynamic and constantly changing.
In the idiom of actant network theory, the task facing our system is to identify, characterize, and evaluate over time the actant networks that comprise emerging research fields. More specifically, this task is to use indicators from the metadata and full-text of publications and patents associated with these fields in order to identify, characterize, and evaluate over time the actant networks of science and technology.

In this paper, we use indicators based on actant network theory to address the question of whether a practical application exists for a given research field at a particular point in time. As noted above, the presence of a practical application may suggest that a research field has experienced a certain degree of emergence, and has moved beyond the early, conceptual stage. This emergence may be reflected in a certain level of activity, maturity and robustness in the actant network associated with the research field.

The presence of a practical application may also help determine whether a research field is more ‘technological’ or more ‘scientific’ in nature. The processes of scientific and technological emergence are not distinct or linearly related, but are often intertwined. They also have many similar properties, and thus often have many actants in common, for example trained scientists funded and tasked to work on a given problem. However, there may be differences in the actant network that point to research fields being more (but not exclusively) technological or more (but not exclusively) scientific. For instance, commercial enterprises, particularly small companies without access to extensive funding, may be more active in the actant networks associated with technological fields that offer a greater prospect of financial return in the near term. Also, the commercial marketplace may play a greater role in the actant networks associated with technological fields. The presence of the marketplace actant in the network may in turn affect the types of outputs required from scientists. In particular, there may be a greater interest in patents (which offer the prospect of a monopoly over a given technology) rather than papers (which offer no such monopoly), especially on the part of potential investors in a given technology (Häussler, Harhoff & Müller, 2012).

**Indicators and Indicator Patterns**

Based on the theoretical foundation discussed in the previous section, we define three characteristics of actant networks that may point to a practical application existing for a given research field in a particular time period: (1) the extent of commercial involvement, (2) the presence of significant patenting, and (3) the degree of maturity of the field.

Each of these characteristics is addressed by a different set of indicators, referred to hereafter as an ‘indicator pattern’. These indicator patterns are based on features extracted from full text and metadata features of both scientific papers and patents. Scientific paper collections we currently process include Elsevier (full text articles from 438 journals over 1980-2011, ~4M records), Thomson Reuters’ Web-Of-Science® (abstracts of journals and conference proceedings for
the same time period, ~40M records) and Pub-Med Central (full text articles from biomedical journals, ~250k records, dominantly 2008-present). We also processed Lexis-Nexis Patent data which includes granted patents and published patent applications from multiple national patent offices. Each of the indicator patterns, and the indicators contained therein, are described below.

**Indicator Pattern 1: Commercial Involvement**

This pattern measures the extent to which commercial organizations are active in the actant network associated with a given research area, relative to academic, government and non-profit organizations. The rationale for this pattern is that research areas with near-term practical applications are particularly likely to attract the attention of commercial organizations. Meanwhile, research areas for which practical applications are far in the future, or have yet to be defined, may be characterized by a lower level of activity from commercial organizations. This is not to say that commercial organizations will be entirely absent from these research areas, since many large corporations have specific departments devoted to blue-sky research. However, their activity may be at lower than in research areas that already have practical applications. The indicators in the Commercial Involvement pattern are:

*Indicator 1.1 - Percentage of Researchers Affiliated with Commercial Organizations* - this indicator is calculated by dividing the number of distinct authors affiliated with commercial organizations by the total number of authors publishing in a given Related Document Group (RDG) and time period. In order to count individual researchers accurately, all author references are disambiguated. The rationale for this indicator is that industrial researchers may be particularly likely to focus on technologies with near-term practical applications. The presence of a large number of such researchers in a given research area, relative to the overall number of researchers active in the area, may indicate that the field has demonstrated promise in terms of practical application.

*Indicator 1.2 - Percentage of Publishing Organizations defined as Commercial* – this indicator is similar to the industrial researchers indicator described above, but is calculated at the level of organizations, rather than individual researchers. Specifically, this indicator counts the number of organizations that published at least one scientific article in a given RDG and time period, and then calculates the percentage of these organizations that are commercial, rather than academic, non-profit or government. In order to count organizations accurately, all organization references are disambiguated.

*Indicator 1.3 - Percentage of Funding Organizations defined as Commercial* - this indicator starts by collating all funding organizations acknowledged in publications in a given RDG and time period. It then determines the percentage of these organizations that are defined as commercial, rather than academic, non-profit or government. In order to count funding organizations accurately, all organization references are disambiguated. The rationale for this indicator is that,
while commercial organizations do fund basic research, they may be especially interested in research efforts related to more applied technologies. The presence of extensive funding from commercial organizations may thus indicate that a technology is more applied, and has potential near-term practical applications.

*Indicator 1.4 - Percentage of Funding from Non-Academic Organizations* - this indicator is similar to Indicator 1.3, but divides the organization types differently. Specifically, rather than isolating commercial funders, it instead isolates academic funders, and counts the percentage of funding organizations that are defined as non-academic. The rationale for this indicator is that, if a large proportion of funders are academic, the field may be more likely to be at a basic science stage of development, rather than an applied technology stage. Conversely, if a large proportion of funders are non-academic, the field may be more applied, and closer to near-term practical application. These non-academic funders may be corporations, government agencies, or non-profit organizations.

*Indicator 1.5 - Percentage of Patents with Company Names in their Background Section* – this indicator computes the percentage of patents that include company names within the Background section of their Specification. The Specification section of a patent often (but by no means always) contains a description of the background of the claimed invention, including a discussion of prior research. If this Background section contains a reference to a company name, this may be because that company is responsible for an element of this prior research. The existence of numerous patents with such references to companies suggests that a research area may be close to practical application, or has already exhibited such an application.

*Indicator 1.6 Percentage of Patents with Company Names in the Example Section* - this indicator computes the percentage of patents that include company names within the Example section of the Specification. In order to demonstrate the usefulness of their invention, patent applicants may list examples describing practical experiments or applications of the invention. These examples are contained in the Specification section of the patent, typically under the heading ‘Example’ or ‘Examples’. These examples may contain references to company names, for example suppliers of testing equipment, raw materials, or diagnostic tools. The presence of numerous patents with such references to company names may be indicative of a technology in which researchers are undertaking practical experiments using equipment and supplies sourced externally. In turn, this suggests that they foresee near-term applications for the invention that would justify this expense.

*Indicator Pattern 2: Significant Patenting*

The presence of the marketplace in the actant networks associated with more ‘technological’ research fields may affect the types of outputs required of the scientists working in these fields. In particular, there may be a greater interest in patents (which offer the potential of a monopoly over a given technology) rather
than papers (which offer no such monopoly), although it should be recognised that the propensity to patent may to vary across fields. For example, patents may play a greater role in fields such as semiconductor and pharmaceuticals than they do in fields such as retail and finance. This pattern includes the following indicators:

**Indicator 2.1 Percentage of Patents** - the percentage of patents is determined by dividing the number of patents by the combined number of patents and papers published in a given RDG and time period. If a technology demonstrates a near-term practical application, particularly an application with possible commercial value, researchers may be more likely to protect their innovations via patents, which offer monopoly rights over these innovations. Conversely, if a field is at a more basic, theoretical stage, the possibilities for patenting may be reduced, and researchers may be more likely to publish their findings in scientific journals. A high percentage of patents among the documents in an RDG may thus be indicative of a technology that is demonstrating potential near-term practical application.

**Indicator 2.2 Extent of Highly Cited Patents** - if a technology is shown to have near-term practical application, it may attract the attention of researchers interested in developing improved versions of the technology. The patents from these researchers will often cite key early patents describing the technology (Breitzman & Mogee, 2002). Hence the presence of numerous highly cited patents suggests that a technology has been the subject of extensive research, which may be due to its potential practical applications. Highly cited patents are defined as those with a Citation Index greater than one. The Citation Index is derived by dividing the number of citations received by a patent by the mean number of citations received by all patents from the same Patent Office Classification (POC) and issue year. The expected Citation Index for an individual patent is one, and a Citation Index greater than one shows that a patent has been cited more often than peer patents from the same year and POC.

**Indicator 2.3 Extent of Citation in Highly Cited Patents** - this indicator measures the soft max of the Citation Index values of the most highly cited patents in a given RDG and time period. As noted above, the presence of highly cited patents suggests that a technology has been the subject of extensive research, which may be due to its potential practical applications. If the most highly cited patents within a technology have extremely high citation rates associated with them, this may be a particularly strong indicator that a practical application has been identified for the technology.

**Indicator 2.4 Percentage of Commercial Patent Assignees** - this indicator counts the percentage of patent assignees (i.e. owners) in a given RDG and time period that are classified as commercial, rather than individual, academic, government or non-profit. The rationale for this indicator is that, while commercial organizations do fund basic research, they may be especially interested in research efforts
related to more applied technologies. The presence of a high percentage of commercial patent assignees may thus indicate that a research field has potential near-term practical applications.

_Indicator 2.5 Percentage of Commercial and Individual Patent Assignees_ – this indicator is similar to Indicator 2.4, but counts both commercial and individual patent assignees, rather than just the former. The rationale for counting commercial patent assignees as a possible signal for practical application is the same as in Indicator 2.4 – i.e. companies may be especially interested in research efforts related to more applied technologies. Meanwhile, patents assigned to individuals often describe specific practical applications, such that these individuals see sufficient near-term commercial potential in their invention to justify the expense of filing and prosecuting the patent. The presence of a high percentage of commercial and individual patent assignees - relative to academic, government and non-profit assignees - may thus indicate that a research field has potential near-term practical applications.

_Indicator 2.6 Average Generality Index_ – the Generality Index measures the extent to which the patents citing a given starting patent are dispersed across technologies (Hall, Jaffe & Trajtenberg, 2001), as defined by Patent Office Classifications (POCs). The rationale for this indicator is that, once an innovation has been shown to have practical use, it may draw the attention of researchers in other disciplines, who consider its potential application elsewhere. This indicator is normalized in the same way as the Citation Index (i.e. by POC and year) and has an expected value of one.

_Indicator 2.7 Extent of patents with high Generality Index scores_ - This indicator computes the number of patents in a given RDG and time period with Generality Index scores greater than 1.5 (i.e. they are at least 50% more generally applicable than peer patents from the same issue year and technology). The indicator is similar to Indicator 2.6, but focuses on individual, high scoring patents, rather than the mean Generality Index across all patents in a given RDG and time period.

_Indicator Pattern 3: Maturity of Field_

This pattern analyzes the maturity of a given research field, with the purpose of identifying fields that have moved beyond the initial, pre-emergent phase. The indicators in this pattern analyze the types of documents present in the RDG (e.g. the existence of product reviews), as well as the availability of technologies cited by documents in the RDG (i.e. whether the technologies are only available on an experimental basis or are readily available for use). Other indicators measure the extent of characteristic terminology in the claims section of patents, which tells us whether the terminology of the RDG has been legally accepted; as well as the extent of specific lexical phrases often used to talk about practical applications, which may be indicative of a relatively mature field.
Indicator 3.1 Extent of product review articles - we created an ontology of genres (interpreted as the term for any category of scientific literature characterized by a particular style and form), and applied it to the documents within a given RDG and time period. The ontology includes 21 types that are arranged in a shallow hierarchy. Some of the most frequent genres are Research Article, Review Article, Report, Book Review, Product Review, Commentary, Abstract, Letter, Correction, Case Report, Editorial, Short Communication, and Discussion. For this indicator, we computed the number of Product Reviews in a given RDG and time period. By their nature, product review articles are related to a product that is available, or will soon be available, in the commercial marketplace. As such, the product has – or is purported to have – a practical application. The presence of numerous product review articles suggests a research field is connected to existing, or forthcoming, products.

Indicator 3.2 Maturity of technologies referenced in patents – this indicator measures the maturity of the technologies referred to in patents, using a classification scheme in which these referenced technologies are defined as unavailable, immature or mature. The classification system uses an ontology that starts with a set of seed patterns and seed technologies, and employs machine learning tools and other heuristics to create the classifications from these seeds. The rationale for this indicator is that research fields with a practical application will tend to reference more mature technologies than research fields that are still at a conceptual stage.

Indicator 3.3 Density of Manufacture relations – the Manufacture relation links arg1 (person, organization or document) with arg2 (document or term), such that arg1 is the manufacturer of arg2. For example, in their Materials and Supplies sections of scientific articles, authors may provide details of items they used and where they obtained them. Similarly, patents may refer to practical experiments using particular tools or equipment. This indicator computes the density of Manufacture relations, i.e. the average per document in a given RDG and time period. The presence of Manufacture relations suggests that researchers in a research field are discussing relevant manufacturing tools and techniques. In turn, this suggests that the research field has moved beyond the purely conceptual or theoretical stage, and has entered a more applied stage where specific methods of manufacture are being considered. A high density of Manufacture relations may thus be indicative of an applied technology with potential near-term practical application.

Indicator 3.4 Density of Practical relations – the Practical relation refers to a particular item either being used, or being useful in some way. It often involves patterns in which trigger words (verbs like use or utilize or instrumental prepositions such as by, via or with) are in close proximity to either citations or terminology describing tools, methods or other descriptors of technology. This indicator computes the density of Practical relations identified in a given RDG and time period. A high density of such relations may be indicative of a research
field in which practical applications are a major focus, and are being actively discussed. In turn, this suggests that such practical applications may exist, or are forthcoming.

Indicator 3.5 Percentage of patents with Examples in the Description of patents - this indicator computes the percentage of patents that include Example headings within their specification section. To demonstrate the usefulness of their invention, patent applicants may list examples describing practical experiments or applications of the invention. These examples are contained in the Specification section of the patent, typically under the heading ‘Example’ or ‘Examples’. The presence of numerous patents containing such an ‘Example’ section may be indicative a research field where it has been possible for many researchers to report the results of practical experiments and applications.

Indicator 3.6 Percentage of patents that include references to trademarks - this indicator computes the percentage of patents that include references to trademarks in their Specification sections. While the purpose of patents is to protect innovations, the purpose of trademarks is to protect products or services. Patents may make reference to trademarks in their Specifications, for example when referring to a potential application for an invention, a component used in the invention, or a piece of equipment used to test the invention. The presence of numerous patents with references to trademarks may be indicative of a research field that is closely related to existing products, or is being tested in practical experiments, and may thus have near-term practical application.

Model Description
As outlined above, there are three indicator patterns directed to the existence of practical applications for given research fields in particular time periods. These patterns are combined in our model, which then produces an output in the form of a Yes/No response to the question: ‘Was there a practical application for <concept> during <time period>?’. In order to test our model, the responses are compared to those from subject matter experts (SMEs) as to whether practical applications existed for given research fields in particular time periods. There are a total of eight such technologies (DNA Microarrays; Genetic Algorithms; Cold Fusion; Steganography; RF Metamaterials; Horizontal Gene Transfer; Tissue Engineering; and RNA Interference) and six time periods (1981-1985; 1986-1990; 1991-1995; 1996-2000; 2001-2005; 2006-2010). For example, the SMEs might be asked whether a practical application existed for Genetic Algorithms in 1996-2000, or Tissue Engineering in 2001-2005.

Before reporting the results from comparing the outputs of our model with the responses from the SMEs, it is first useful to outline details of the model. Our model is based on Bayesian networks, which are probabilistic graphical models. Probabilistic relationships among variables are captured by a directed acyclic graph. In this graph, the nodes are variables together with a specification of the probability distribution for each variable conditional on values for its parent.
variables (i.e. variables having edges to the given variable). Such a representation allows for a decomposition of the joint probability distribution over the variables that supports efficient computation of probabilities.

The model is hierarchical, with indicator variables linked to pattern variables, which are in turn directly linked to the question variable. The pattern variables represent abstractions or summaries of related indicator variables. Such a hierarchical structure has a number of advantages. First, it allows the use of many correlated variables without danger of over-counting. For example, percentages of commercial researchers and percentages of commercial organizations will be highly correlated, so that treating them as independent pieces of evidence would overstate their influence. If, however, we take these two variables to be manifestations of a more general commercial involvement variable, and link this variable to the question variable, then the correlation between the commercial researcher and commercial organization variables will be properly accounted for.

**Figure 1 – Model for Existence of Practical Application**

A second advantage of the hierarchical approach is that it enables more accurate modeling when training data is sparse. With sparse data, estimating the correct, or even an approximately correct, probability distribution for each indicator variable conditional on the question variable is very error prone. With intermediate pattern variables, theoretical considerations can be brought to bear to help shape the
distribution in conjunction with whatever ground truth data is available. A third advantage to a hierarchical structure is that explanations for answers are more comprehensible when framed in terms of meaningful patterns above the level of individual indicators. Finally, the intermediate patterns will often be of interest in and of themselves, independently of how the question is answered.

Figure 1 shows the model for addressing the question ‘Was there a practical application for <concept> during <time period>?’. This model includes the three indicator patterns outlined in the previous section – i.e. commercial activity, extensive patenting, and field maturity. It should be noted that other indicator patterns may also have informational value regarding this question. However, the additional patterns we considered were not used, either because analysis showed their information value to be low, or because modeling their relationship to the existence of a practical application proved difficult and error-prone.

Each subgraph in Figure 1, consisting of a pattern node and its children, is a naïve Bayes model. Updating in a naïve Bayes model is particularly simple. Evidence takes the form of an assignment of a value to a child variable and each piece of evidence contributes independently to the posterior distribution over the parent variable.

For our model of practical application, we used the following Boolean conditions over the pattern variables as conditions for inferring the existence of a practical application:

1. Commercial Involvement is TRUE.
2. Patenting is TRUE.
3. Maturity is TRUE.

Each of these conditions is considered ‘sufficient to some degree’ for a practical application to exist. The Boolean conditions are combined using a weak disjunctive model known as the “noisy-or” distribution. The probability of the existence of a practical application is greater than 0.5 when at any one of the three conditions is satisfied, but the conditions differ in their strength of support for a practical application. Finding a significant patent presence provides very strong support (greater than 0.9) for the existence of a practical application independently of the other pattern variables while maturity provides only weak support (0.55) when the other two pattern variables are false. When all pattern variables are false, the probability of a practical application is low (0.1).

Results

We ran the model on the RDGs for each of the eight sample technologies in each of the six time periods listed above. Where the model output a probability value greater than 0.5 for the existence of a practical application, this was considered a positive answer to the question (i.e. the model states that a practical application existed in that research field and time period). Conversely, a value lower than 0.5 was considered a negative output (i.e. the model states that a practical application did not exist for the field during the time period). These outputs were then
compared against the responses from the SMEs, and the results are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>RF Metamaterials</th>
<th>Tissue Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td>27</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>False positives</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>True negatives</td>
<td>11</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>False negatives</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Recall</td>
<td>0.90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>0.84</td>
<td>0.66</td>
<td>1</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.82</td>
<td>0.83</td>
<td>1</td>
</tr>
</tbody>
</table>

There are a total of 46 answers in the ‘All’ column of Table 1. This represents eight technologies times six time periods, minus two time period/technology pairs for which the data were too sparse for the model to run (both pairs were from the earliest time period, 1981-1985). The results in Table 1 show that the model worked consistently well with respect to the SME responses. Recall is 0.90 (i.e. 90% of time period/technology pairs with positive SME responses to the question of whether a practical application existed were also given positive answers by the model). Meanwhile, Precision is 0.84 (i.e. 84% of time period/technology pairs marked with a positive answer by the model were also marked positive by the SMEs); and Accuracy is 0.82 (i.e. in 82% of cases, the responses from the model and the SMEs matched, whether these responses were positive or negative).

Tables 2 and 3 provide more detailed views of the different indicator patterns incorporated in the model, and how they contribute to the responses generated by the model. Table 2 shows results for Tissue Engineering, while Table 3 shows results for RF Metamaterials.

<table>
<thead>
<tr>
<th></th>
<th>SMEs</th>
<th>Model</th>
<th>Commercial Involvement</th>
<th>Patents</th>
<th>Maturity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981 - 1985</td>
<td>NO</td>
<td>NO</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1986 - 1990</td>
<td>YES</td>
<td>YES</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1991 - 1995</td>
<td>YES</td>
<td>YES</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1996 - 2000</td>
<td>YES</td>
<td>YES</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>2001 - 2005</td>
<td>YES</td>
<td>YES</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>2006 - 2010</td>
<td>YES</td>
<td>YES</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

In Table 2, it is possible to see the transition from entirely negative answers to the practical application question from both the model and the SMEs in the earliest time period (1981-1985), to entirely positive responses in the more recent time periods. This reflects the widespread application of tissue engineering techniques.
in recent years. The responses from the SMEs and the model match for each time period, so precision, recall and accuracy are all one for this RDG.

Table 3 shows the SME responses and model outputs for the RF Metamaterials RDG. This table reveals that a practical application for RF Metamaterials has been identified much more recently, according to both the SMEs and the model. For the first four time periods, covering 1981 to 2000, the SMEs gave a negative response to the question of whether a practical application existed for RF Metamaterials. They gave a positive response to this question for both 2001-2005 and 2006-2010. The model, meanwhile, suggested that there was a practical application for RF Metamaterials earlier than the SMEs, and switched from a negative to a positive response in the 1996-2000 time period. This is largely due to the presence of a high percentage of industrial researchers during this period. The positive response from the model in 1996-2000 results in a false positive for this period, reflected in the figures for precision (0.66) and accuracy (0.83).

<table>
<thead>
<tr>
<th></th>
<th>SMEs</th>
<th>Model</th>
<th>Commercial Involvement</th>
<th>Patents</th>
<th>Maturity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981 - 1985</td>
<td>NO</td>
<td>NO</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1986 - 1990</td>
<td>NO</td>
<td>NO</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1991 - 1995</td>
<td>NO</td>
<td>NO</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>1996 - 2000</td>
<td>NO</td>
<td>YES</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>2001 - 2005</td>
<td>YES</td>
<td>YES</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>2006 - 2010</td>
<td>YES</td>
<td>YES</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

The results in the tables above suggest that our model shows promise in terms of determining the existence of practical applications for given research fields. This determination is based solely on the content of scientific and technical documents, and without access to product or market data. Given that product and market data are often difficult to collate, or are expensive to source, a model such as this that does not require access to such data may be a useful tool.

Practical applications do not only exist for emerging research fields, but also for mature fields. Such fields may also exhibit the characteristics covered by the indicator patterns included in our analysis – i.e. commercial involvement, significant patenting, and maturity. Hence, it is not necessarily the existence of a practical application that is interesting from an emerging technology standpoint, but the existence of such an application for the first time. As a result, the promise shown by the model in recognizing the transition from absence to existence of practical applications for given research fields may be of particular interest when evaluating emerging technologies.

Although the results are promising, it should be noted that they are based on a very small data set, largely due to the time-consuming nature of surveying SMEs. Additional research using more extensive data sets may thus be instructive, in
order to determine whether the promising results reported here are repeated for a larger sample of technologies. It also needs to be recognized that, as with any human-based response data, there is the possibility of bias or misunderstanding in the responses from the SMEs. It is thus possible that the model outputs are being compared against erroneous responses from the SMEs.

Conclusions
This paper outlines a system directed to the determination of whether practical applications exist for research fields, particularly those that are emerging. The system uses indicator patterns - based on features extracted from the metadata and full text of scientific papers and patents - to assess different characteristics that point to the existence of practical applications. This system may help determine whether a particular field has moved beyond the early, conceptual phase towards a more applied, practical phase. It may also help to classify emerging research fields as more 'technological' or more 'scientific' in nature.

The results reported in this paper suggest that the system shows promise in determining whether practical applications exist for given research fields in particular time periods, based on comparisons with responses to the same question from subject matter experts. Perhaps more interestingly, the system shows promise in detecting the transition from absence to existence of practical applications over time, which may be of particular value in evaluating emerging technologies. It should be noted, however, that the results are based on a small sample of research fields, due to the time consuming nature of obtaining responses from subject matter experts. More extensive research using a larger sample of research fields may be worthwhile, in order to determine whether the results are repeated for this larger sample. Also, in this paper, the system is applied to pre-determined research fields. It may be interesting to apply the system to a broader corpus of documents covering multiple research fields, to determine the extent to which it is able to locate those fields that are emerging, and also appear to have a practical application.

Acknowledgments
Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References


