To appear in: *Technological Forecasting and Social Change*

Emergence Scoring to Identify Frontier R&D Topics and Key Players

Alan L. Porter, Jon Garner, Stephen F. Carley, Nils C. Newman

Abstract

Indicators of technological emergence promise valuable intelligence to those determining R&D priorities. We present an implemented algorithm to calculate emergence scores for topical terms from abstract record sets. We offer a family of emergence indicators deriving from those scores. Primary emergence indicators identify "hot topic" terms. We then use those to generate secondary indicators that reflect cutting edge organizations, countries, or authors especially active at frontiers in a target R&D domain. We also flag abstract records (papers or patents) rich in emergent technology content, and we score research fields on relative degree of emergence. This paper presents illustrative results for example topics -- Nano-Enabled Drug Delivery, Non-Linear Programming, Dye Sensitized Solar Cells, and Big Data.

Keywords

R&D assessment; R&D Indicators; Technology Emergence Indicators; Tech mining; Emerging technology

1. Introduction

Attention to emerging technologies is increasing. Such indicators can address varied subjects, ranging from breakthrough science to novel technology and on to commercial innovation. Foci can range from 'micro' (e.g., treating specific sub-topic activity patterns) through systematic 'macro' indications (e.g., disruptive technological system emergence). We focus at the micro level, seeking practical measures to distinguish "hot" R&D sub-topics. Such indicators of emergence can contribute to R&D policy and portfolio management, technology opportunities analyses [1], and management of innovation.

The U.S. Intelligence Advanced Research Projects Activity (IARPA) Foresight and Understanding from Scientific Exposition (FUSE) Program drew attention to the value of technology emergence indicators [http://www.iarpa.gov/index.php/research-programs/fuse]. FUSE supported four teams that explored ways to derive indicators via text analyses of Science, Technology & Innovation (ST&I) data resources. We have been involved in conceptualizing bases for emergence and framing candidate indicators [2]. We continue to work to generate viable indicators [3]. This paper carries that effort forward to offer "emergence scores" for terms appearing in R&D abstract records. It goes on to use those emergent terms to distinguish cutting edge "players" -- research organizations, countries, or individuals based on their engagement of emerging technology content.

We seek to distinguish topics drawing accelerating attention in research publication or patenting within a target domain. *Our aim is to provide practical means to separate cutting edge research from other research ongoing in a target domain*. We go a second step to tally R&D activity on those cutting edge topics by "players" – countries, organizations, or individuals. This can provide vital intelligence on who are leading the way to advance these frontiers.

We are 'micro' in another way – we seek to identify emergent R&D within particular topical domains – i.e., abstract record datasets retrieved by searching for a given topic, such as "Big Data," in suitable databases. Our approach is to treat and analyze topical terms extracted from such datasets – i.e., text mining. We operationalize a conceptual model of technical emergence using a combination of thresholds and activity trend calculations to generate Emergence Scores. We extend those to measure the players most active in pursuing such topics, thereby providing a suite of emergence indicators. The paper presents results from several case analyses and considers a range of potential indicator developments.

Section 2 offers a brief introduction to various perspectives on "technical emergence." Section 3 describes the data and our analytical approach. Section 4 generates our R&D emergent terms, with emergence scoring, for the case analyses. It points toward validation in the form of testing their predictive performance. Section 5 uses that term emergence scoring to generate indicators of the extent to which players are active at the cutting edge in the target domain. Section 6 explores use of these emergence indicators to spotlight cutting edge papers, countries or authors, and probe their features. It also investigates cross-domain emergence comparison. Section 7 offers conclusions, and discusses limitations and future opportunities.

2. Literature Concerning Emergence Indicators

2.1. Emergence as a Property

We start with a brief, broad consideration of "emergence" from different perspectives [12], then point to our interest within that. Many scholarly fields consider the notion of emergence in various contexts. Rotolo, Hicks, and Martin [5] explore conceptual foundations for "emergence" along multiple dimensions. They note the sharp growth in research and popular press publications addressing aspects of emergence. Rotolo et al. [5] and Li et al. [8] consider emergence in complex systems – arising from a coming together of components to offer new "emergent" properties not easily predicted as the sum of the parts. Li et al. [8] review various approaches in seeking to discern distinctions between emergent and disruptive technologies.

"Emergence" can be associated with radical change in science, tracing back to Kuhn [7], to differentiate from ordinary scientific progression. "Emerging technologies" range from incremental to radical innovations, covering a wide spectrum. These can well reflect differences in emergence processes (e.g.,

in biomedicine [6] vs. semiconductors or such). FUSE interests concern early identification of novel advances portending scientific and technological opportunities across a gamut of R&D areas.

Rotolo et al. [5] track the evolution of "emergent technologies" in the social science literature. Three distinct facets in thinking about emergence are sources, characteristics, and effects. We see value in being able to distinguish various "emergent" entities, including:

- converging research streams
- technology currently at an early stage of development, showing high growth potential
- hot sub-technologies within a target domain
- radical or discontinuous innovation
- potential enhanced economic influence.

Emergence can be treated at numerous levels, emphasizing some or all of these aspects (c.f., [9], considering "Lab on a Chip" with distinct developmental pathways as a platform technology – or not). As we develop indicators of emergence, it is cautionary to keep in mind that the concept is inherently complex. Multiple approaches have been explored to measure emergence [10,11]. We don't attempt to treat those comprehensively here [see 5]. But to set this work in context, we note that the scope can range from measuring macro-level players – e.g., national propensities [12, tracking longer term country R&D activities] – to micro-level scientific topics [13]. Temporal perspective can range from recent monthly analyses to decades-long time frames [14]. Data and methods employed can range from expert opinion [12] to bibliometric analyses of R&D publications or patents [15]. The last noted article is highly salient as an exemplar of bibliometric analyses, such as this paper represents too. But, we especially contrast Glanzel and Thijs' [15] approach to identify emergence based on citation patterns, whereas we pursue lexical change measures (term activity patterns).

Technology (or science) growth can take place in many pertinent dimensions:

- Within the technology space overall or of various components
- Into other technology spaces
- Within the R&D community

Our approach focuses within a particular scientific or technological domain, not emergence cast at the level of the whole scientific enterprise. This limits our window to that given domain (to be implemented as a search set on a topic), so we won't see the spread of initiatives from that domain to others. Conversely, we should detect intrusion of novel concepts/findings/methods from other domains. That said, our definition accepts that sharp intrusion as emergence.

Technical emergence is an important facet of foresight [16] and technology intelligence and forecasting [17 -- that text treats growth curves and predictive capabilities]. Interest can range from technical emergence to innovation – i.e., breakthrough commercial or military applications arising from new technical capabilities. Focus may be general or domain oriented [18]; we fit the latter interest. We focus on technical emergence indicators for R&D activity. Several approaches contribute diverse formulations and approaches. Staudt and colleagues [19] target high-impact and transformative science metrics. An et al. [20] compare national level contributions to emerging themes.

2.2. Criteria for Technical Emergence Indicators

On a conceptual level, we followed FUSE by focusing on criteria for technical emergence indicators – namely, Novelty, Persistence, Growth and Community. We also track with FUSE in that we are probably tracking prominence rather than emergence of ideas. Where we depart from FUSE is in how we operationalize these core concepts. The FUSE program had access to several full databases (patent and publications). We note two dimensions to this – enhancing abstract records to full text (articles or patents) and ability to process entire global databases in one's calculations. This unprecedented level of data access permitted the creation and testing of models at a scale not seen before. But the vast data access also pushed the FUSE program to solutions which required immense data access and powerful computing. Our approach operates under the assumption of a different reality, a reality where an end user would have limited access to data – namely, search results usually drawn from one database of field-structured abstract records and attendant metadata. Our approach is not better (full text offers richer potential); it is just designed for different, practical operational environments.[1]

Importantly for us, Rotolo et al. [5] consolidate FUSE and other prior research to identify key attributes of technological emergence. Our model keys on four attributes -- *novelty, persistence, community,* and *growth*. These attributes do not directly translate into unambiguous indicators (measures). Some points of interest regarding the four criteria:

- **Novelty** – newness; can pertain to technologies, to technical sub-systems, functions, and/or uses
- **Persistence** – indicating some identity and momentum -- e.g., shared use of acronyms, ongoing community interest [FUSE explorations treated "cold fusion" as a vivid counter-example]
- **Community** – as in "community of practice," implying multiple players, not all within some single unit, and connecting in some manner -- e.g., citation connections in R&D literature or patent analyses;
- **Growth** – pertaining to increasing R&D outputs and/or to gains in other facets (e.g., funding, players).

We note potential clashes – e.g., growth reflecting in upward, relatively continuous trends vs. novelty embodied in discontinuous R&D activity onset and spiking. Likewise, persistence and novelty pull against each other – persistence implies ongoing multi-year activity, whereas novelty watches for relatively short-period, abruptly increasing activity. "Community" poses multiple dimensions (c.f., [21]).

"Growth," in particular, points toward trend analyses of time series data with an eye toward projecting likely future activity trajectories (i.e., technology [17]). We confront tradeoffs such as "novelty" favoring

[1] The Authors are familiar with FUSE as participants in the program but are not engaged with the final implementation of SRI's Copernicus system by Meta. Nor do we have knowledge of any changes to UI/UX and data access now that Meta is part of the Chan Zuckerberg Initiative. Thus we cannot speak to how Meta's system would perform compared to a "smaller data" approach. On a conceptual level, we are encouraged by the advances in a large data approaches; however, as practitioners in this space, we are curious as to how the META system will be able to provide end users access to information that is covered by copyright from corporate publishers (Clarivate, Elsevier, Digital Science, etc…). In the absence of operational access to such full data systems, we offer this generation of emergence indicators from "smaller" (but not so small – see Table 1) sets of abstract records.

detection in short time series to stress recency and disruptive change patterns vs. "persistence" seeking sustained growth patterns. Growth can be modeled multiple ways as well – e.g., fitting logistic or exponential curves to time series to project future trending [17].

2.3. Methodological Roots of our R&D Emergence Indicators

Focusing in, we look at R&D advances using "tech mining" [22] to treat topical content compiled from ST&I abstract records. Tech mining just combines bibliometric and text analytic methods to track activity patterns. Important general tools include routines to clean text fields – namely, fuzzy matching routines and thesaurus application to consolidate name variations. Given that text in ST&I discourse is messy – lots of ways to depict a closely related thing – consolidation of variants is vital. This potentially calls upon a range of clustering and topic modeling tools. Kontostathis et al. [23] overview roles for text mining algorithms to assist in detecting and tracking topical emergence. The last section of this paper returns to compare our approach to some others.

Small, Boyack, and Klavans [13] present an intriguing approach to identify emerging S&T topics using literature data. Drawing on the *Scopus* database, they identify topical clusters using direct citation and co-citation, and track temporal patterns. Their results treat the four emergence criteria just noted nicely in distinguishing emergent topics for all of science (a macro approach). In contrast, we seek to distinguish emergent topics within target research fields (a micro approach). For instance, one of Small et al.'s [13] 71 emergent topics is "cloud computing"; we might search and download S&T abstract records relating to that topic, then seek topics meeting the four criteria *within* the resulting "cloud computing" dataset.

2.4. Summing Up

To reiterate, our aim is to operationalize the four traits in analyzing S&T literature and patent abstract datasets to devise practical emergence indicators. Our strategy has two stages. First, we seek to identify emergent *terms* – i.e., topical content that evidences the four attributes. To do so we extract topical content from downloaded abstract record sets to discern terms or phrases that show high growth, along with evidence of novelty, persistence, and community. Second, we then strive to get at "who" is most active in pursuing research that uses those terms in the available text data (abstract records). For instance, which research organizations most actively include the high emergence terms in their publications? [Saying that, we recognize one of many challenges – should one look for the greatest publication rate or the highest concentration of publishing relating to those topics?] Term emergence scoring also enables us to generate useful information regarding the degree of technical emergence of particular records and research fields.

Our R&D emergence indicators should meet several objectives:

- Generalizability across S&T domains (i.e., not relying on domain-specific thresholds)
- Database independence (i.e., trying to avoid reliance on fields or data elements particular to one or a few databases; aiming to work with both research publication and patent data)
- Ease of use so that an analyst can generate useful indicators of cutting edge R&D activity

- An algorithmic (reproducible) approach.

2. Data and Methods

2.1.Data

The process begins by retrieving a set of research publication or patent abstract records from a suitable database. Section 3.2.1 addresses search queries for the topical datasets used here. The datasets addressed would usually be topical (e.g., resulting from a search on, say, "graphene"), but could be organizational (e.g., a search for Georgia Tech authored papers), or universal for a given data source (e.g., all European Patent Office patents over an extended period).

Select fields are used in these analyses: 1) topical information is based on title and abstract Natural Language Processing (NLP) to extract terms and phrases; 2) player information is extracted as follows – authors, from the Authors field; organizations, from Author Affiliations; and countries, from Author Affiliations. Section 3.2 goes into the extraction, cleaning, and consolidation processes further, especially for terms.

We experimented with the generation of indicators using data on four technologies in six datasets (Table 1). One potentially vital characteristic of the dataset being analyzed is growth rate. Detecting emerging topics in the context of rapidly growing record sets could differ from doing so in relatively stable sets (i.e., low annual growth rate). Our listed datasets give us a rapidly growing science/technology (NEDD), two rapidly growing technologies (DSSCs and Big Data), and a relatively slow-growing, applied mathematics research area (Non-Linear Programming).

Table 1. R&D Emergence Indicator Test Datasets

Dataset	Source	10-Year Test Period	Full Period Available	Total #
Nano-Enabled Drug Delivery (**NEDD**)	MEDLINE	2001-2010	2000-2013	10354
NEDD	Web of Science (**WoS**)	2000-2009	2000-2012	50745
Non-Linear Programming (Non-Linear)	WoS	2003-2012	2003-2015	3225
Dye-Sensitized Solar Cells (**DSSCs**)	WoS	2000-09	2000-2012*	8053

DSSCs	PatStat	2001-2010	1957-2013 (early years inappropriate)	4872
Big Data	WoS	2004-2013	2003-2016 (partial year)	13349

To treat these field-structured text records, we employ VantagePoint (www.theVantagePoint.com) for text processing and emergence indicator calculations, in conjunction with MS Excel. However, the algorithm does not require particular software.

Herein, we analyze these topical time series datasets to illustrate the R&D emergence indicators. We use a 10-Year test period consisting of a base period (3 years) plus an active period (7 years), and, for validation purposes, a follow-on period (an additional 3 years). In arriving at these periods, we tried many variations. Shorter time periods yield much noisier results (overly sensitive to specific processing). Longer periods run counter to the aim of seeking currently emerging topics. The 3 + 7 model offered a reasonable middle ground. In future research we will investigate using other time periods. In particular, we see promise in a 7-period active period with a 5-period base using quarterly data.

The total record number varies over an order of magnitude in these cases.

2.2.Methods

This paper reflects a bibliometric-based approach to measure emergence – i.e., we tabulate and track patterns in R&D literature and/or patents. "Tech Mining" is a term that we have adopted to describe our use of text mining tools to extract useful intelligence from ST&I information resources [22]. Tech Mining combines bibliometrics with text analyses to draw inferences from sizable record sets. It favors searching for work on a topic of interest in ST&I global databases, retrieving abstract record sets on a topic – e.g., Table 1. Those records provide convenient compilations of field-structured information. We note these attributes to distinguish this work from text mining of unstructured text such as news feeds or social media compilations.

Tech Mining furthers various analytical aims. A number of those relate to our purpose of measuring R&D topical emergence. We note a few exemplars here, not a comprehensive review. Much such work seeks to generate Competitive Technical Intelligence (CTI) by tracking "who's doing what?" [22, 24]. Variants of such text analyses can associate actors and technologies, exploring related factors such as R&D funding [25].

Tracing technological trends is especially relevant; this is fostered by text analyses of topical content to consolidate important terms and phrases relating to a given concept or theme [26, 27]. That topical content can then be tracked over time to get at evolution pathways [28].

We now turn to five specific methodological steps that comprise our efforts in generating R&D emergence indicators. Figure 1 shows these as a basic flowchart.

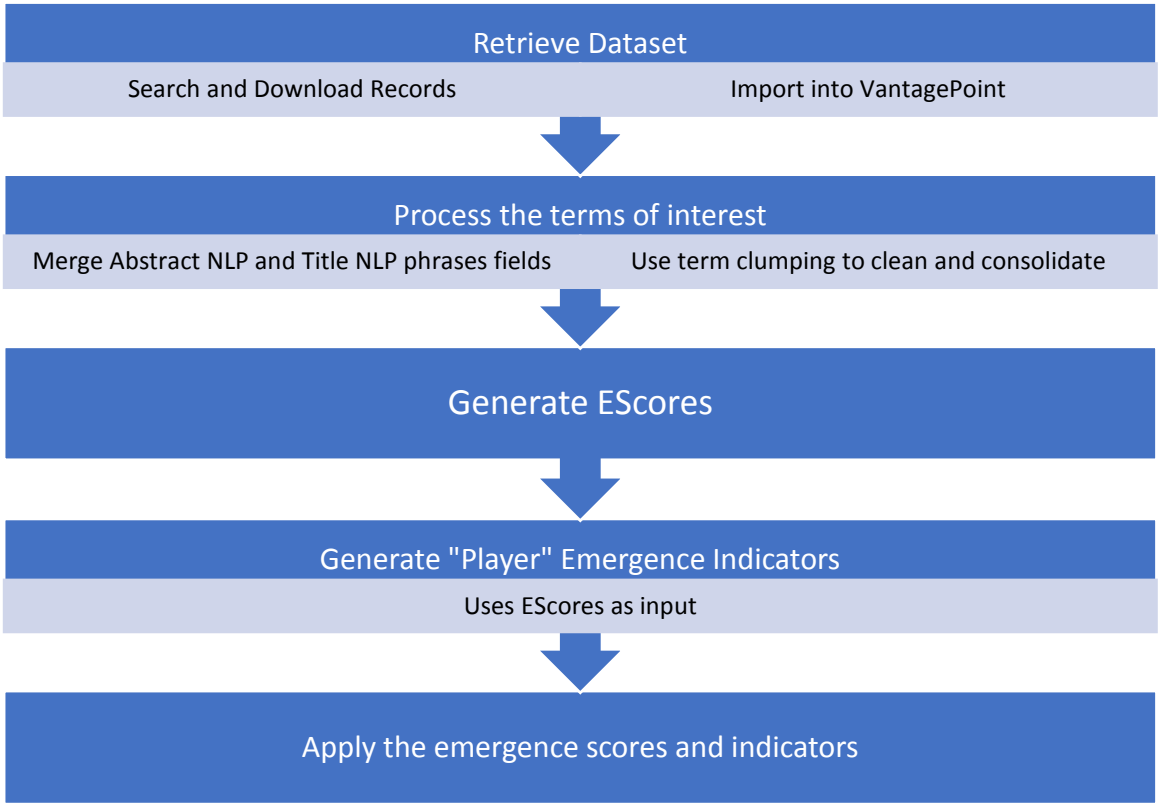

Figure 1. The Basic Process of Generating R&D Emergence Indicators

2.2.1. Retrieve Dataset

Step 1 has been introduced under "Data." As per Table 1, one could search in various databases for R&D on a target domain – and that selection can greatly affect results. One's search algorithm also makes a big difference in the resulting content and scope. Those sensitivities are not of primary concern here, as we use the topics as case studies to illustrate emergence scoring, rather than to advise on R&D priorities per se. However, we note that search formulation warrants serious attention. Indeed, we have published articles on the search strategy development for nanotechnology [29, 30] that undergirds the current DSSC and NEDD searches. We have also elaborated the development of three of these topical search developments in extensive detail [31, 32, 33], with consideration of all four searches as well [3].

For readers interested in the distinct search strategies for the four topics (NEDD, Non-Linear Programming; DSSCs; Big Data), the Supplemental Materials[2] to this article provide considerable detail in Section A. Exact replication would be extremely challenging, but one could use the search framework and queries to approximate them. For example, Figure 1-S there presents the Big Data 4-part search

[2] Supplemental Materials for this article are available at: http://hdl.handle.net/1853/59335. These are too extensive to warrant appending to the article. They provide details on search strategies; figures showing the additional topics not shown in the article comparing total and normalized emergence scoring results for countries, organizations, or authors; and other supporting details.

framework: core lexical query, expanded lexical query, specialized journal search, and cited reference analysis. Table 2-S lists the search terms comprising the core and expanded lexical queries. The intricacies in applying these are laid out elsewhere [28].

One should be wary of the likely incompleteness of the most recent period(s) of data (e.g., the last year). Possible recourses are to delete the most recent period data (but we seek to be as current as viable), collapse incomplete recent periods together, or normalize for partial last period data.

2.2.2. Process the terms of interest

Here, we seek to provide the essence of the treatment approach; more complete details are available on request. The fundamental notion is to extract informative *topical content* in a reproducible, efficient way.

Topical content in R&D abstract records varies by source database. After comparing the effectiveness of alternative fields and manipulations of them (e.g., merging fields), we have chosen to extract, and combine, noun phrases (including single word terms) from titles and abstracts. To devise topical emergence indicators, we elect NOT to use various keyword fields (e.g., MEDLINE MeSH terms, WoS author keywords or Keywords Plus, patent class codes, patent claim NLP phrases) in favor of fields that are more generally available in various ST&I database records. We have calculated emergence indicators and scores using various combinations of topical content to compare. Results generally track reasonably well.

Table 2 summarizes the steps we take in processing abstract and title phrases to get at topical content.

Table 2. Standard Term Cleaning/Consolidation Process

1. Import the Abstract Natural Language Processing (**NLP**) phrases and the Title NLP phrases fields into the analytical software (e.g., VantagePoint).
2. Merge those two fields; then remove terms appearing in just one instance to yield an "Abs+Ti>=2" field (e.g., for Big Data -- 197,960 items reduced by eliminating terms appearing in just one instance to give 31,348 terms).
3. Apply the five standard thesauri in ClusterSuite[3]; then separately run VantagePoint's List Cleanup (general fuzzy matching routine) (yielding 22,474 terms for the Big Data test set).
4. Split that term set into unigrams (single words) and multi-word noun phrases, treating each subset as follows:
    - Unigrams – run a WoS stopword thesaurus of 786 terms [for scientific data or use patents stopwords if patent data], thereby removing many general technical terms.
    - Multiwords – Run the ClusterSuite "Fold NLP Terms" algorithm ("Folding" counts occurrences of a shorter term appearing in longer phrases, and it augments record and instance counts; it does not remove terms).
5. Merge the resulting unigram and multiword lists and manually screen out a few very frequent and consolidated noise terms to input the remaining terms to the EI script (which offers further cleaning routine options as well) (for Big Data, 22,425 terms).

[3] We have consolidated these thesauri along with various fuzzy matching and other cleaning and text consolidation routines into a script called "ClusterSuite," developed by J.J. O'Brien [4] and available at www.VPInstitute.org.

As mentioned, and explored further in Supplemental Materials, Section A, topical terms can be drawn from various abstract record fields. Those include keywords, index terms, and class codes (especially for patents). For generalizability, we work here with abstract and title phrases, using VantagePoint's NLP routines to extract those from the abstracts and titles. This NLP formulation is conceptually akin to Princeton Word net, drawing on semantic and syntactic rule sets, for English language. This NLP is further trained to identify blocks of text that do not adhere to normal English rules – e.g., chemical formulas – to better extract meaning from scientific discourse.

This routinized term consolidation process aims to facilitate reproducibility and comparisons among different datasets. Our experience with identification of emergent terms is that users are put off to see noisy terms included; hence, the extensive attention to data cleaning (Table 2). While the resulting topical term sets are far from perfect, they do appear valid to knowledgeable domain experts.[4] Comparisons of alternative term sets in generating emergent terms underpin Table 2.

2.2.3. Generate EScores

We have developed a custom "Emergence Indicator" (**EI**) script for VantagePoint software (Figure 2). The script first separates terms (generated via Steps A & B) that meet these thresholds:

a) Appear in records from at least 3 years
b) Appear in at least 7 records
c) The ratio of records containing the term in the active period to those in the base period must be at least 2:1
d) The term cannot appear in 15% or more of the base period records
e) Terms are also required to have more than one author that doesn't share the same record set

The thresholds aim to achieve the desired ET emergence criteria of novelty (c & d), perseverance (a & b), and research community (e). The particular values are based on our experience with test cases. The EScoring script allows users to alter these values at will.

[4] We thank Dr. Natalie Abrams (National Cancer Institute) and Jing Ma (Beijing Institute of Technology), for in-depth exploration of the NEDD content. We thank Prof. Gary Parker and Prof. Anton Kleywegt (Georgia Tech) for review of our Non Linear Programming term sets.

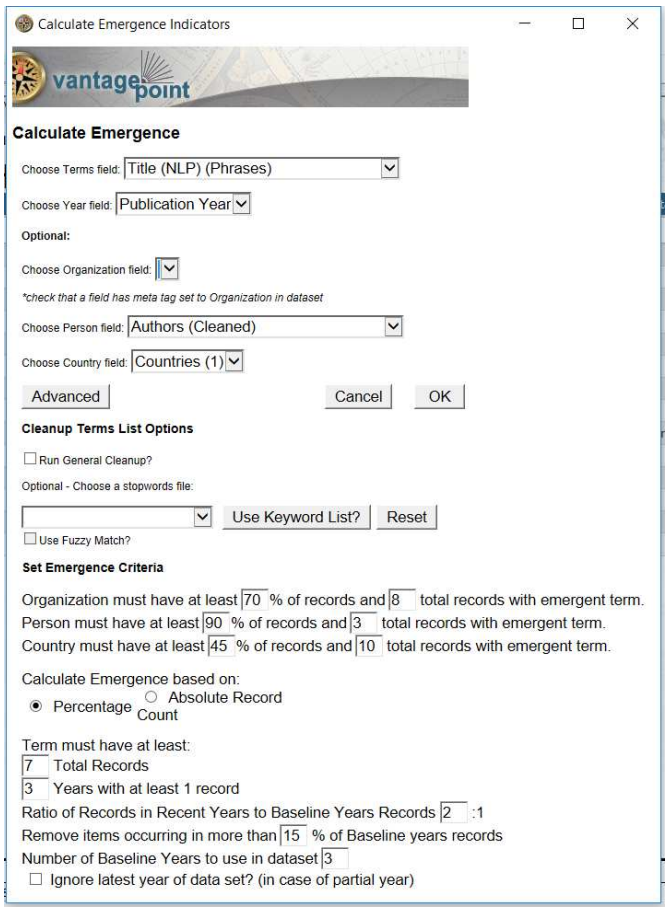

Figure 2. Emergence Indicator Script Control Panel

These thresholds target the four attributes of emergence that we pursue: Novelty, Persistence, Community, and Growth. Thresholds a) and b) aim to assure a level of Persistence (i.e., that the topic is not a "one-hit wonder." Thresholds c) and d) support Novelty and Growth; the term is appearing increasingly often later in the data period. Threshold e) assures that multiple authors not all within one research group have engaged the topic. The specific levels chosen are based on our experimentation with the test datasets described here and several others. However, the script enables a user to vary the thresholds. E.g., for a small size dataset, one might reduce the requirement of at least 7 total records containing the term.

We initially developed a set of routines to tag "emergent terms" from these candidate terms. Those were binary – either emergent or not [3]. This paper presents an advance to generate "Emergence Scores" (**EScores**) that provide continuous, numerical scale values for the candidate terms. We examined various EScore formulations – e.g., differential weighting of title vs. abstract terms; different combinations of trend components; multiplicative vs. additive component weighting; tiered term levels; and so on. We selected an additive model incorporating three of four available component trends[5]:

[5] Calculations are incorporated in VantagePoint's "Emergence Scoring" script as per Figure 2.

- Active Period Trend - comparing the change from the most recent 3 years to the first 3 years of the active period.
- Recent Trend -- comparing the change from the most recent 2 years to the 2 years prior.
- Slope from the mid-year of the active period to the most recent year.(usually would be Year 7 to Year 10) [presuming a 3-year base period followed by a 7-year active period]
- Slope from first point to mid-point (not included in "EScore5," our favored formulation)

We based this EScore on observed behavior, considering trend plots and selectivity (how many terms score as 'emergent').

*EScore = 2 * Active Period Trend + (Recent Trend + Mid-Year to Last Year Slope).*

For a given term with 7 periods of active data (the default), the calculations would be:
Active Period Trend = Terms Record Count of period 5, 6, 7/Summation(Square Root(Total Record Counts in period 5, 6, 7)) - Terms Record Count of period 1, 2, 3/ Summation(Square Root(Total Record Counts in period 1, 2, 3))
Recent Trend = 10 * (Terms Record Count of period 6, 7/Summation(Square Root(Total Record Counts in period 6, 7)) - Terms Record Count of period 4, 5/ Summation(Square Root(Total Record Counts in period 4, 5)))
Mid-Year To Last Year Slope = 10 * (Terms Record Count of period 7/ Square Root(Total Record Counts in period 7) - Terms Record Count of period 4/Square Root(Total Record Counts in period 4) / Change in Time (e.g. period 7 - period 4)).

We examined term sets, finding strong correspondence between the binary emergent term (**ET**) sets for a given test dataset and the high EScore terms. For instance, for DSSCs, of 90 ETs, 88 have EScores >1; 45 have EScores >2. After numerous comparisons for the several test datasets, we settled on a threshold for EScores of 1.77 (square root of Pi). These EScores provide the bases for secondary emergence indicators – see "D."

2.2.4. Generate "Player" Emergence Indicators

Emergent terms point to cutting edge R&D activity, but are quite sensitive to term cleaning and consolidating. For instance, the "Big Data" set includes many variations on data analysis terms – which particular ones "make the cut" depends on nuances in the term consolidation. In essence, we use them as pointers, but place more stock in the "player" emergence metrics.

The EScores offer many options to measure the degree of emergence of individual records; record compilations (e.g., to compare domains of interest) and "players" – i.e., organizations, individuals, and/or countries on their extent of incorporation of highly emergent terms in their R&D activity data being considered.

We experimented with various ways to use the EScores to gauge these secondary indicators. Some keyed on tallying emergent term use (e.g., by an organization) vs. others that count records with substantial emergent term content. We considered alternative modes to normalize for different attributes (e.g., record length). We determined not to normalize on the dataset so that one would get a set percentage of terms above threshold in any target domain; instead we favored an absolute mode that enables cross-dataset comparison (e.g., to gauge relative domain emergence of NEDD vs. DSSCs).

Table 3 presents results that led to our determining a threshold value of 1.77 for high emergence EScores. After empirical comparisons, we determined to set aside terms below that EScore. Put another way, we do not factor in "less emergent or non-emergent" terms in calculating indicators of leading cutting edge players (i.e., those most actively writing on emergent topics) in target R&D domains.

A vital option is whether to use those terms per se, or to use records, as the basis of determining cutting edge players. Our test analyses led us to prefer to use terms to distinguish cutting edge organizations (or countries or individuals). However, we do see end-use value in tallying EScores (>1.77) to identify research publications or patents with high emergence content.

We considered many alternative term counting approaches to identify highly cutting edge organizations. These varied counting, summing up, or averaging of raw or transformed EScores. We compared relative rankings of the top organizations by various measures. Outlier terms (e.g., one Big Data term had an EScore of 46.7, far above the next highest term at 8.6) posed concern. That led to trying logarithm and square root transformations. We adopted the square root (SQRT) as providing somewhat wider range without concern for ln (0).

Experimentation led us to adopt two measures to compare players (organizations, countries, individuals):

1) **Total** = Summation of SQRT (EScores) above the chosen threshold [SQRT (Pi) = 1.77], counting each time a term was used in a distinct record; but not crediting multiple occurrences within a record[6]
   ∑ (SQRT(ESc) x # records for that ESc term) – summation over all high ESc terms

2) **Normalized** = Summation (as in "1") divided by the SQRT of the number of records..
   [∑(SQRT(ESc) x # records for that ESc term) – summation over all high ESc terms] / SQRT (# of records of that player)

The Total measure credits overall organizational use of the high EScore terms, but in a way that does not unduly favor extremely high scoring terms. The Normalized measure is an attractive option in discerning certain differences. Section 4 reports results for test cases.

2.2.5. Apply the emergence scores and indicators.

This 'section' just completes the five process elements. EScores provide a resource to enrich R&D management in various ways. Section 4 presents case analyses and the final section offers ideas on potential uses.

[6] The use of "square root of pi" here is incidental – we chose it in that we were inclined toward taking the square root of EScores as a suitable transformation. More critically, we decided that a threshold between 1.5 and 2 was desirable. Importantly, note that the 1.77 threshold is set for the EScores (not for the SQRT of those EScores). In other words, we first screen out EScores <1.77; then we take the square root of those above-threshold EScores.

4. Validation

How effective are these emergence indicators? To address this question, we follow the IARPA lead and start with trying to assess how well our emergence indicators predict sustained R&D emphases in future periods. We selected a 3-year test period. Fewer than three years would seem apt to reflect "more of the same" – terms appearing in many papers (or patents) in the preceding few years would likely remain 'hot' for a year or so at least. More than three years would be problematic to expect high continuity; by definition, "emergence" reflects rapid change.

We decided to focus on the research activity (publications) of the last 3 years of the 10-year period analyzed and compare those to the following 3 years (the test period). Using similar period durations holds appeal and the most recent years are most determinant of the trends constituting the Emergence score (ESc5 version[7]), as well as most relevant for the ETs.

To set the validation stage, recall that we calculate ETs and EScores using 10 years of data, divided into a 3-year base period followed by a 7-year active period. Now we augment that with an additional 3-year test period (recall Table 1), and we draw on the last 3 years of the active period for comparison. *Our core question is whether designation as emergent foretells high R&D activity in the test period?*

We considered various ways to measure high test period R&D activity, focusing on the number of papers (or patents) in which the ETs or high EScore terms appear. We did not formulate the validation as strict hypothesis testing, but rather as an exploratory approach. So, we examined various metrics, such as: relative trending for those terms in the test period vs. the prior 3 years, and various ratios of test period to prior period. We noted that the overall domain growth rate pattern affected such comparisons – e.g., contrast the relatively stable Non-Linear Programming publication trend vs. super growth, tapering off, for Big Data.

All said, we decided that term prominence in the test period was a suitable measure of emergence [in this, we are following the FUSE project that had adopted prominence three years later as a key criterion]. That is, the primary comparison would be between candidate emergent terms' publication activity in the test period vs. that of other terms (i.e., other candidate emergent term formulations and non-emergent term benchmarks). Table 3 consolidates key results.

[7] This version is used throughout the paper; we retain the "ESc5" designation to keep our records in order.

Table 3. Predictive Utility of Candidate Emergence terms in Three Test Datasets

Dataset		ESc5<0	ESc5<1 & >0	ESc5<2 & >1	ESc5>2
a) **BD**	#	505	486	50	29
b)Test Period	Ave. 2014-16	16.0	17.1	52.2	**208.7**
c)Prior Period	Ave. 2011-13	10.7	13.1	38.1	120.9
d) **Non-Linear**	#	129	79	35	25
e)Test Period	Ave. 2013-15	4.65	5.28	8.23	**15.8**
f) Prior Period	Ave. 2010-12	3.48	5.05	7.43	13.8
g) **DSSCs**	#	683	--	--	70
h)Test Period	Ave. 2010-12	37.1			**149.7**
i)Prior Period	Ave. 2007-09	14.5			48.2

Notes: Table 3 summarizes results in terms of our favored Emergence Scoring algorithm (shorthand is "Esc5" for the fifth one investigated). ESc scores are partitioned by size in the 3rd—6th columns; higher values indicate greater emergent attributes. Results are presented for three of the four datasets investigated – BD = Big Data; Non-Linear = Non-Linear Programming; DSSCs = Dye-Sensitized Solar Cells. The "#" indicates the number of terms scoring in the Escore range indicated. For example, 29 BD terms score >2. The Test Period results indicate the mean number of records containing each of the terms in the given EScore range in that period. The text discusses implications.

Table 3 is excerpted from a working table that includes relative values for ETs and various combinations of EScores and ETs [e.g., counts of occurrences for terms with EScore >2 & also an ET (the binary measure)]. Our comparisons led to the conclusion that the high EScore terms were distinctly superior in performance to the binary "ETs." So, to simplify, we focus on EScores here.

Consider Table 3. Rows present counts for each of three datasets: Big Data (BD), Non-Linear Programming (Non-Linear), and Dye-Sensitized Solar Cells (DSSCs). For each of those, the first row shows the number of terms fitting the criteria designated by the columns. For example, for BD, 505 terms had an EScore <0, whereas 29 terms achieved EScore >2. The following row tallies the average number of records containing each of those terms in the test period. Our prime emergence characteristic of note is the relative level of activity in the test period (i.e., prominence). For BD, compare the 208.7 average # of records for ESc5 > 2 terms to the average of 16 records for EScore < 0 records. The high

EScore terms also tend to be more active in the prior 3-year period, suitably reflecting their "emergent" characteristics. E.g., for BD, row c) shows the EScore <0 terms averaging 10.7 records vs. 120.9 for the EScore >2 terms for 2011-13. [Note that the prominence in the test period is understated in that it represents a shorter period than the prior comparison period because our 2016 data are incomplete (due to lag in database indexing).]

Likewise, one can compare prominence for Non-Linear terms as somewhat higher in the test period than in the prior period (15.8 vs. 13.8), but notably higher than EScore <0 terms in the corresponding periods. The intermediate EScore value terms are notably less prominent. Results for DSSCs also support high EScore terms tending to be prominent R&D interests in the following test period (149.7 vs. 37.1 for EScore <0 terms.

Table 3 also shows that the emergence scoring algorithm is selective – only 29 BD terms, 25 Non-Linear, and 70 DSSC terms make the cut at >2. In selecting a threshold for EScores, we weighed the appeal of a higher degree of emergence for the highest scoring terms (here, consider EScores >2) vs. appeal in having a larger number of terms. These terms were drawn from, respectively, 26,093 BD, 10,768 Non-Linear, and 29,121 DSSC terms (abstract & title phrases, treated as described previously). Our compromise was to select the square root of pi (1.77 – i.e., a value between 1.5 and 2). Counts for that level rise modestly to 36, 29, and 81, respectively, for BD, Non-Linear, and DSSCs. Ergo, our process is very selective in what it identifies as high emergence terms. We considered relaxing thresholds (e.g., to a shorter period of 2 base years and 6 active years, thereby enabling less long-lived terms to qualify), but decided not to do so in favor of better persistence in considering the lengthier time series.

5. Emergence Indicator Results for Case Analyses

5.1. Emergent Topics (Terms)

Note our caveat that term formulation is quite sensitive, so that some term variants (e.g., of "Big Data") make the threshold whereas other associated terms do not. So, the particular emergent term sets are somewhat fragile and should be considered with caution. Nonetheless, we feel they provide topics showing accelerating R&D attention recently.

Table 4 indicates the number of terms with EScores >2. Using our threshold of 1.77 increases that number modestly (last paragraph of prior section). To give the flavor of these high EScoring terms, Table 4 shows the top 10 for each dataset, giving the term's EScore and the # of records in the 10-year dataset in which it appears. [The Supplemental Materials for this article list all terms with EScores >1.77.]

Table 4. Top 10 High Emergence Score Terms in Three Test Datasets

Topic	Term	EScore	#
DSSCs	power conversion	13.1	179
	power conversion efficiency	12.0	174
	organic dye	9.3	94
	electrochemical impedance	8.5	121
	photovoltaic performance	8.2	197
	electron microscopy	7.1	128
	TiO(2)	7.1	68
	extinction coefficient	6.7	51
	TiO(2) film	6.4	46
	density functional theory	6.2	71
Non-Linear	mixed integer	5.3	106
	operating cost	4.3	25
	Mixed Integer Non-Linear Program MINLP	4.1	18
	linear behavior	4.0	17
	novel approach	3.3	22
	model results	3.2	12
	non-linear function	3.1	16
	mixed integer linear program MILP	2.9	7
	non-linear behavior	2.8	15
	scheduling problem	2.7	16
BD	big data	46.7	622
	data analytics	8.6	119
	MapReduce	7.9	600
	big data analytics	6.6	80
	Hadoop	6.5	436
	social media	5.5	73
	Big Data process	4.9	61
	framework MapReduce	4.4	151
	social network	4.3	130
	Hadoop cluster	4.2	66

5.2. Cutting Edge "Players"

"Secondary" indicators, in the sense that they build from the primary indicators -- the high EScoring terms -- offer considerable appeal. The calculations are parallel in generating cutting edge R&D organizations, authors/inventors, and/or countries – i.e., those prolificly using ETs. Here, we focus

attention on such organizations. The aim is to identify organizations whose cutting-edge R&D in a target domain stands out, using an algorithmic process that is easy to apply.

The secondary indicators lend themselves to visualization variations. Here, we plot various pairs of Total EScores, Normalized EScores, and Number of publications. For some purposes, plotting transformed values enhances one's ability to discern contrasts of interest. The Supplemental Materials present a large family of these for Big Data, DSSCs, and Non-Linear Programming, in turn, showing organizations, authors, or countries positioning based on EScores of their publications. Here, Figures 3-4-5 present logarithmic transformations, as seem most suitable, for organizations publishing on DSSCs.

\ compares the log (Total) against Normalized EScore measures for DSSC organizations. We first apply the Total (measure "1" from above; displayed along the horizontal axis) as a signal of research organizations most actively publishing (or if using patent data, patenting) on "hot" topics in the domain. So, were one seeking a collaborating organization, or a target university program to which to apply, this could be especially helpful. The top organization here is the Chinese Academy of Sciences (CAS) with a total EScore approaching 1,000. [No absolute meaning is attached to this sum of square roots of EScores; higher is more.]

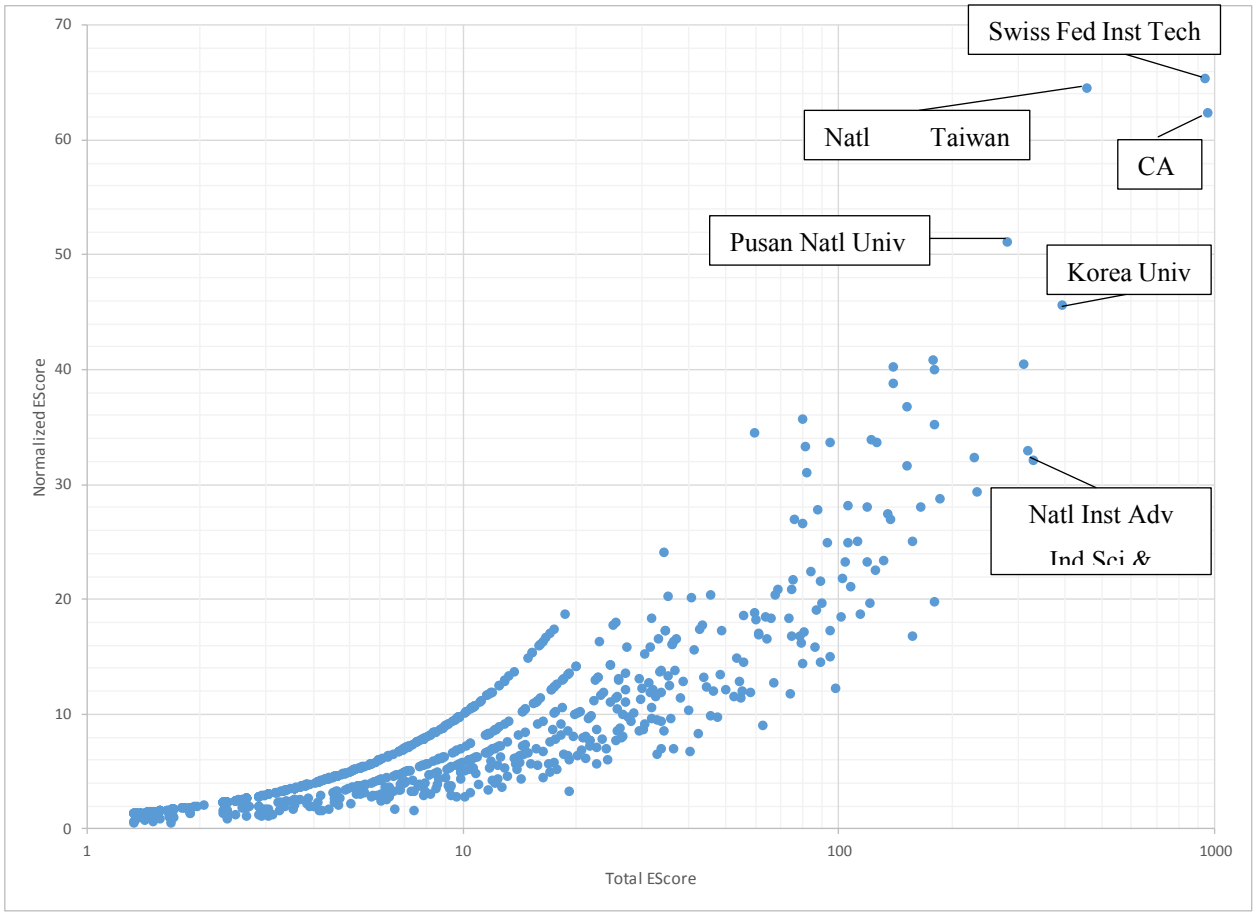

Figure 3. Normalized vs Log (Total EScores) for DSSC Organizations

As displayed in Figures 4 and 5, the leading DSSC publishers for 2003-2009, as indexed in WoS, are CAS, the Swiss Federal Institute of Technology (we consolidate with its French name, Ecole Polytechnique Federale Lausanne, to get 207 publication records), and the National Institute of Advanced Industrial Science & Technology (105). Fig. 4 presents log (Total EScores) (here on the vertical axis) vs. log (number of publications). Fig. 5 presents Normalized EScores vs. log (number of publications). The Normalized EScores counterbalance high publication rate as a major contributor to Total EScores. In Fig. 5, note that National Taiwan University, with its 50 publications, scores slightly higher than CAS and the Swiss Federal Institute of Technology, with over 200 each. So if you are looking for a group keying on DSSC emergent topics, this would point to them. These data don't speak to the networking within organizations (likewise, for countries), so DSSC R&D within an organization could be well-connected or quite dispersed. Consider CAS as an exemplar with over 100 institutes; we can't tell from the present data treatment to what degree those are co-located. However, one can readily list the players with their EScores and record counts to facilitate perusal within organizations.

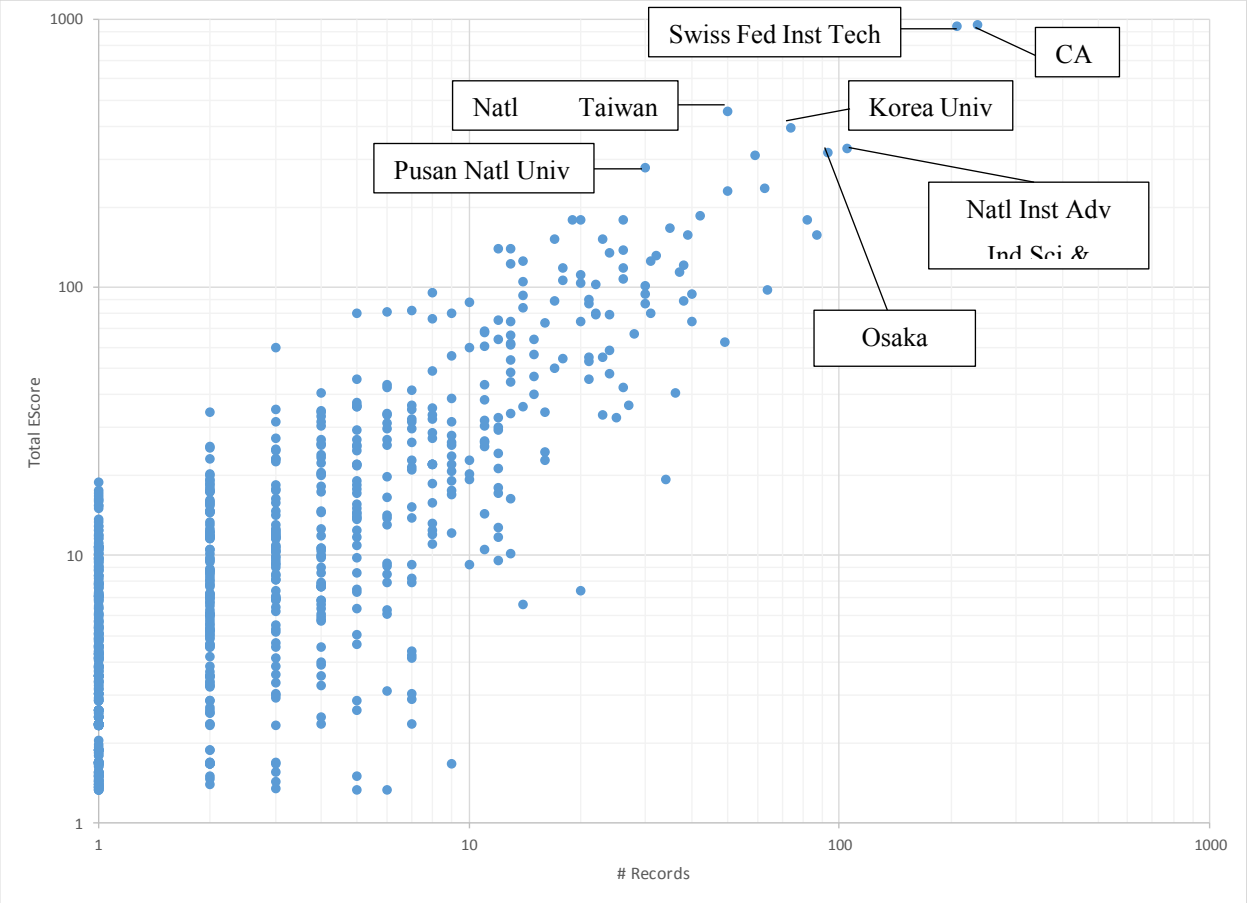

Figure 4. Log (Total EScores) vs Log (Publication Counts) for DSSC Organizations

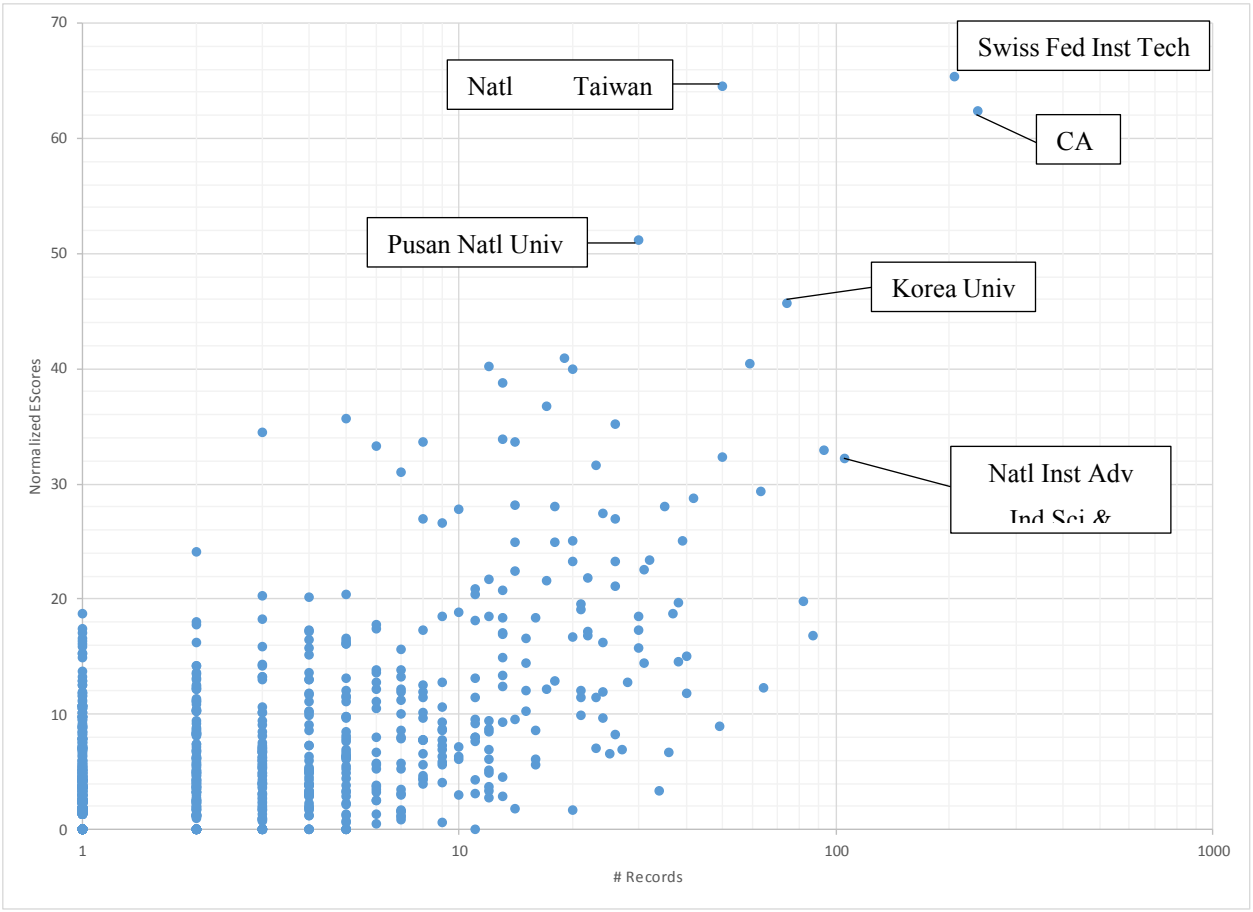

Figure 5. Normalized EScores vs Log (Publication Counts) for DSSC Organizations

We generally recommend discarding low record counts. The EI Script (Fig. 2) default settings are 10 records for countries, 8 for organizations, and 3 for authors. Consider two high scorers on the Normalized EScore measure – Bannari Amman Institute of Technology on Big Data (Fig. 6, 2 records) and Ocean University of China on Non-Linear Programming (Fig. 7, 2 records). For most purposes, these would not be of interest. Again, we nominate the Total EScore measure ("1") as dominant for most purposes. However, the EI script offers flexibility so that one could explore such very low record count organizations to pursue particular interests.

As mentioned, Figures analogous to Figures 4 and 5 are available (Supplemental Materials) for Big Data and Non-Linear Programming (raw data, not log transformed). Those complement the plots of Normalized vs. Total EScores by explicitly showing the numbers of publications in the target domain by particular organizations. Furthermore, tables can provide full details for further probing of particular cutting edge organizations (those most actively addressing ETs).

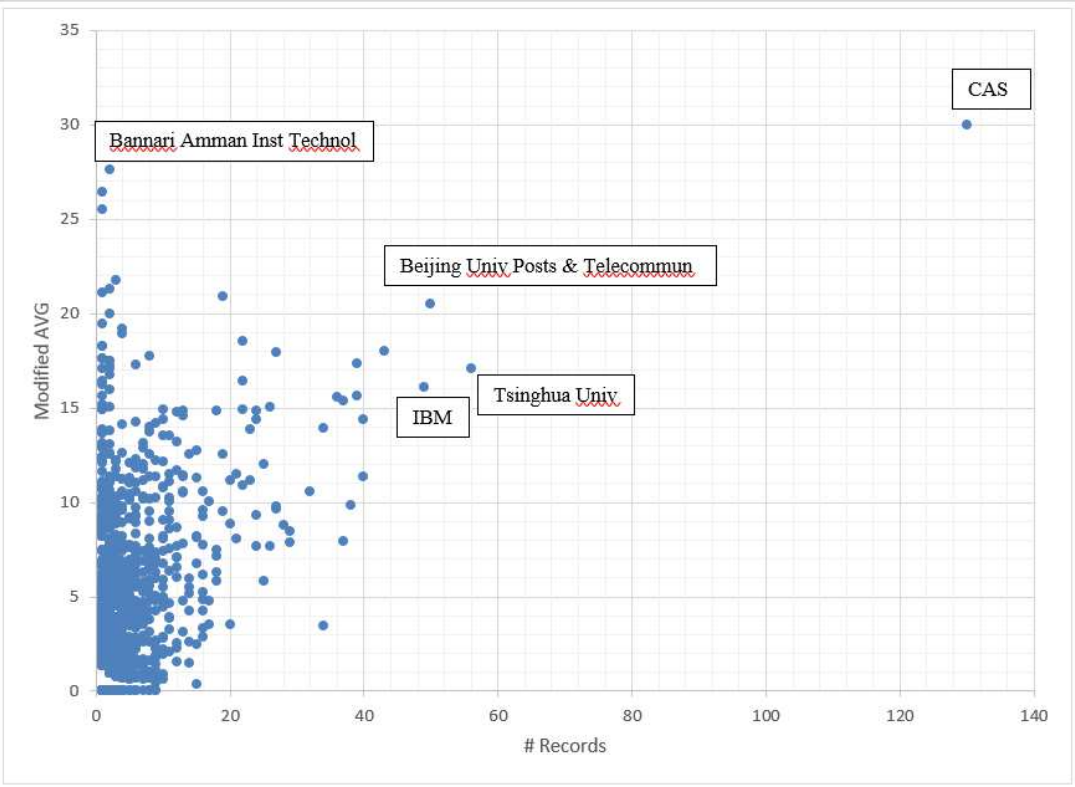

Figure 6. Normalized vs. Total EScores for Organizations Publishing on Big Data

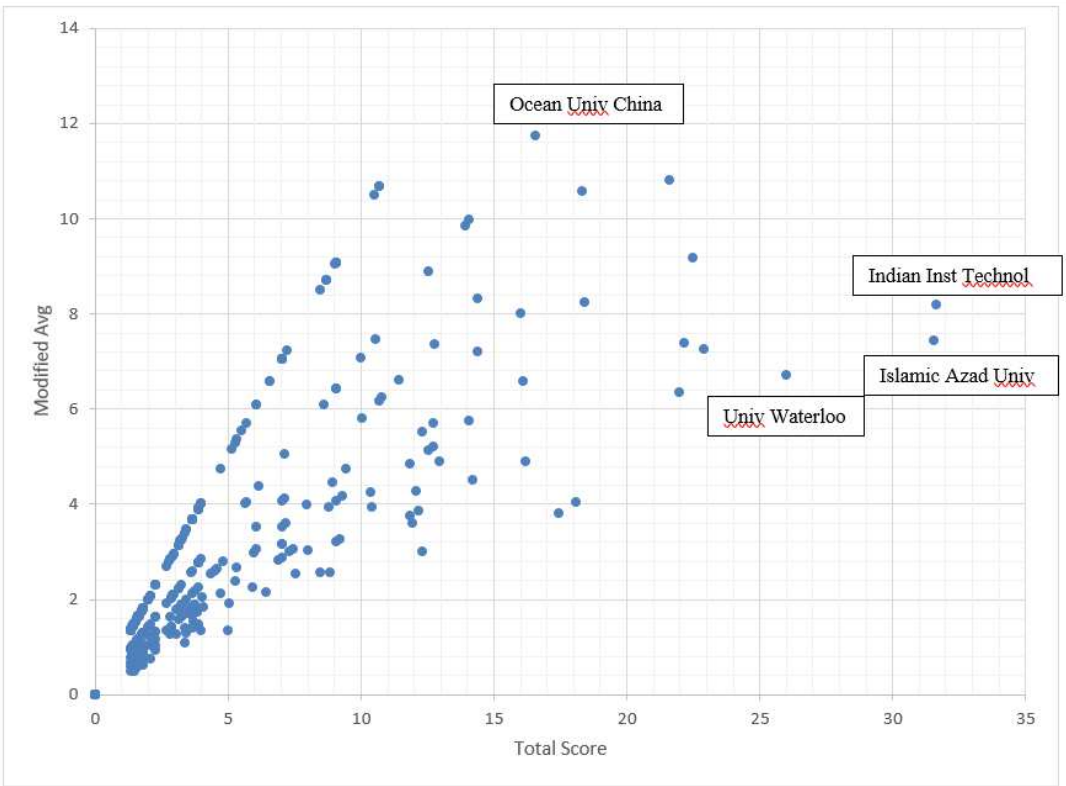

Figure 7. Normalized vs. Total EScores for Organizations Publishing on Non-Linear Programming

6. Further Analyses: Toward Applications of the R&D Emergence Indicators

R&D emergence scoring can serve multiple purposes. To provide focus, consider an analyst addressing issues concerning R&D management; mergers & acquisitions; product development, or ST&I policy. Technical Emergence Indicators, derived from searches in global databases, could support such an analyst by identifying:

1) Component **technologies** within a domain that warrant attention as "emerging" [based on the set of emergent terms generated]
2) High priority research **papers or patents** within that domain deserving special attention [based on emergent records]
3) Key **organizations** active at the frontier of R&D in the domain -- to monitor as high priority (potential collaborators or competitors) [based on calculations of organizations with high Total (and/or Normalized) EScores]
4) **Countries** to track [based on high Total (and/or Normalized) EScores, analogous to the treatment of organizations just illustrated]
5) Cutting-edge **authors** [likewise based on high Total (and/or Normalized) EScores]
6) Relative emergent R&D activity level of different **technical domains** [based on field level calculations, to be described below]

Given these six targets for EScore application, there are many options possible. Section 5 just illustrated #3 – "cutting edge organization" calculations. We presently offer observations on targeting the other five, with further exploratory notions offered in the final Section.

6.1. Emerging Technologies within a Domain

The base measure here is to generate high EScore **terms**, as described previously. Those high EScore terms are not, however, neat sets of "the" hot topics of note. At this juncture we support a process of:

- Calculating EScores (the ESc5 measure) for the "qualifying terms" (using the EI Script)
- Applying a threshold of 1.77 to distinguish high EScore terms
- Presenting those high EScore terms for a target domain dataset to knowledgeable colleagues to stimulate selection of emerging technologies for further analyses tailored to one's driving research questions.

6.2. Priority Papers (or Patents)

What papers should researchers, analysts, or managers concerned with the cutting-edge in a technical domain scrutinize? The answer is first determined by the type of papers most suited to those needs – perhaps, foresight studies, technology roadmaps, and/or technology assessments of the technology; and/or heavily cited, recent review papers; and/or heavily cited, classic research studies? For WoS records, one could use document type and "times cited" to screen for high priority publications.

To that, we add an indication of the **high EScore papers**. Those address cutting-edge topics, potentially providing novel concepts, methods, or applications. Our proposed measure to identify high EScore papers is

** **Total** = Summation of SQRT (EScores) above 1.77 – for the qualifying papers

Given that EScores reflect a combined measure based on trends and thresholds, they are absolute in nature. So, one would expect more papers above a given EScore (based on the high EScore terms appearing in a paper's title and abstract) in a hot area (e.g., BD) than in a relatively staid one (e.g., Non-Linear Programming). Accordingly, depending on the purpose in mind, one might point to the "Top N" emergent papers for the domain under study to scan for content of special importance.

To illustrate the possibilities, take the Big Data case. In VantagePoint, we first make a field of the 36 terms with an EScore >1.77. We then tally the sum of the square roots of those EScores for each record. Next, set a threshold either on how many records we want (e.g., Top 10) or what score value to use. [In the BD case, the term "Big Data" is an outlier with a huge EScore, so we consider setting that aside as both overly general and overloading.] The result is a collection of research papers whose abstracts contain a relatively high amount of emergent terms. [Here also we have not investigated all options – e.g., instead of basing the selection on sum of the square root of high EScore terms appearing in each record, one might, instead, select records containing the most high EScore terms.]

One then has a rich analytical resource in the tabulation of high EScore papers to use in distinguishing cutting edge authors (or inventors), organizations, or countries to explore those entities further. As noted in Section 5, in identifying "cutting edge organizations," we measured their emergent term content, instead of emergent records. That does not preclude examining the set of "emergent records" of, say, a BD R&D organization. For instance, imagine a Chrysler Competitive Technical Intelligence (**CTI**) study of Toyota's high EScore papers on Electric Vehicles to help gain a sense of its frontier R&D interests in that domain.

Conversely, one could use information on authors, organizations, or countries that are most actively addressing ETs to screen for their emergent papers. For example, recall the DSSC data that showed CAS and the Swiss Federal Institute of Technology as most cutting edge organizations in this domain. One might therefore key on papers authored by their researchers.

To illustrate this process, we explored the Big Data research using Total EScores to identify:

- Top 12 Organizations (chosen as the Total EScores had a large jump between #12 and #13)
- Top 11 Authors
- Top 11 Countries

Results are interesting:

- 6 of the 12 high EScoring organizations matched with 7 of the high EScoring papers
- No matches of any of those top 11 authors with the 36 high Total EScoring papers
- 5 Top 11 countries associate with some of the 36 papers: France (1), Germany (1), India (5), China (10), and the US (12) – with none of these papers showing co-authors from multiple Top 11 countries

Multiple measures are vital in analyzing R&D data. Again, to illustrate the potential of combining emergence scoring with other measures, we form a matrix of the 36 high Total EScoring BD papers by Times Cited for each paper (as provided in the WoS records). Given that more than half of the 13,349

BD papers in our dataset were published in 2014 or later, citation data are constrained. Nonetheless, we note that only 5 of the 36 high Total EScore papers have received multiple citations – 2 with 2 cites; 1 with 3; 1 with 16; and 1 with 18. So, were the analyst seeking influential, cutting-edge papers (s)he might point to those two that have high EScore and are highly cited:

- Bian et al. (2012), Towards Large-scale Twitter Mining for Drug-related Adverse Events, *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, Maui
- Lee et al. (2011), YSmart: Yet Another SQL-to-MapReduce Translator, *IEEE International Conference on Distributed Computing Systems,* Minneapolis

6.3. Cutting Edge Countries

Our first focus in comparing countries is on propensity to publish papers (or patents) within a target technical domain. We don't believe that calculating EScores for countries on all topics offers much value. So focusing on a target domain (e.g., DSSCs), we tally high EScore terms to benchmark countries (e.g., the leading countries in the domain in terms of articles or patents treating ETs).

Our emergence scoring for countries mirrors that for organizations or authors, using:

- "**Total EScores**" = Summation of SQRT (EScores) above 1.77
- "Normalized EScores" = Summation (as in "1") divided by the SQRT of the number of records [an alternative measure]

Visualizations of the Total and Normalized EScores, along with Number of publications, for countries appear in the Supplemental Materials. Those include plots for countries analogous to those for organizations illustrated by Figures 3, 4 & 5. We offer observations to suggest potential utility of analyzing data in this way.

DSSC research spans 25 years. It is a well-connected community with extensive cross-citation of research papers. Leading authors have produced enormous numbers of research publications. At the country level, one could well begin investigation by tallying the number of publications and the extent to which those are cited. Our contribution is to offer additional metrics that focus on R&D emergence within the domain.

DSSC research was initiated in Switzerland, and Swiss authors and, especially, one organization (Swiss Federal Institute of Technology, Lausanne) continue to publish much highly emergent research. However, on the national level, Fig. 21 (Supplementary Materials) shows China, Japan, and South Korea out in front. Taiwan, Switzerland, and the USA stand forth as a second tier. Fig. 22 points out high Total EScores with high publication counts. EScoring here presents an interesting comparison between the USA and South Korea – the USA publishes more on DSSCs, but South Korea shows higher emergence scoring.

Big Data research publication shows dominance by the USA and China (Figures 24-A and 25-A). Somewhat surprising, the emergence scoring does not find the next most prolific countries (Germany

and the UK) quite as strong. On total EScoring, India appears next [followed by Japan, Germany, Taiwan, Australia, South Korea, Canada, and then the UK (not shown)].

Non-Linear Programming – our "less emergent" domain presents a surprise. Fig. 27-A shows the USA leading on Total EScore, followed by Iran. On Normalized EScore, Iran leads, with the USA second. As mentioned, emergence scoring here seems to be flagging a potentially interesting intelligence finding. In other analyses of nanotechnology data focusing on a national level, Iran also shows strongly (#5 for 2006-2015)[8].

The next tier of Non-Linear Programming countries is led by China, followed by India, Canada, and Brazil – the 'BRIC's" without Russia, plus Canada. This interesting research emergence pattern may reflect an applied math field that is highly respected for sophisticated contributions, but not demanding heavy technical infrastructure.

An additional, distinct option would be to identify emergent technologies, papers, organizations, and/or authors **within a country**, within a domain. This entails different indicators using component data provided by the EI Script – a future research target to build on emergence scoring. To help gauge the resourcefulness of the target country we could investigate whether many or few distinct research groups are actively investigating emergent topics.

For instance, in the Non-Linear test case, Iran's activity is notable. We separate the Iranian Non-Linear records; then examine the EScores of the terms used. Whatever the extent of those reaching our general thresholds, we could characterize the leading foci within Iranian R&D on Non-Linear Programming. Here, we want to identify concentrations of relatively emergent R&D and the authors and organizations performing that R&D. Of interest, would be the degree of R&D concentration in certain organizations within Iran.

6.4. Cutting-Edge Authors

Our approach here parallels that for cutting edge organizations previously detailed. Our primary measure is the same "Total" Summation of SQRT (EScores) >1.77, augmented by "Normalized" Summation (as in "Total") divided by the SQRT of the number of records. Figures like Figures 3, 4 & 5 are included as Figures 11-A through 19-A in the Supplemental Materials for this article.

We label a few leading authors in Figure 11-A. The top three in terms of Total EScore are all affiliated with the Swiss Federal Institute of Technology, Lausanne. Ho and Lee are associated with the National Taiwan University. In terms of number of publications, Graetzel and Nazeeruddin lead, with one other (Hagfeldt at Uppsala University, 227 papers, for a Total EScore summation of 265 and a Normalized EScore of 26.2) preceding Zakeeruddin and Ho. Lee trails with 44 (in 73d place), so emergence scoring differs from basic publication quantity. This analysis points to the Swiss group as leaders in the field.

[8] Paper under review.

Publications per author for Big Data are far fewer than for DSSCs. BD is far newer and far more dispersed across authors (a much less cohesive research community). That pattern can be discerned, for instance, in Supplement Fig. 15– where the striation by number of papers published is pronounced.

Jinjun Chen and Xuyun Zhang, of the University of Technology, Sydney (**UTS**) are interesting. All 7 of their Big Data papers included in this high EScoring set are jointly authored by them. Indeed, 18 of 20 Zhang papers are co-authored by Chen; likewise, 18 of 20 Chen papers, by Zhang. [This identifies the potential utility in applying VantagePoint's "Combine Author Networks" script to consolidate multiple authors who are heavily collaborating, to treat the group as a single entity.]

The top emergence scoring author, Chang Liu, is also affiliated with UTS. Two other authors have more papers in the Big Data document set, but do not achieve high emergence (Lizhe Wang with CAS and Wei Wang with Tianjin Normal University).

Big Data is an explosive dataset – growing rapidly with wide participation over a short lifespan (given our search criteria). Our 13,349 BD papers have 34,779 authors (2.6 authors/paper). Collaboration among the most prolific authors (>= 12 papers or in the group of 11 high Total EScores) are shown in Supplement Fig. 20. The Figure shows a pattern of limited networking. It also shows some strong teaming, especially at UTS. Fig. 8 shows a portion of that figure to give the flavor of a sparse overall network, with some tight local collaborative networks.

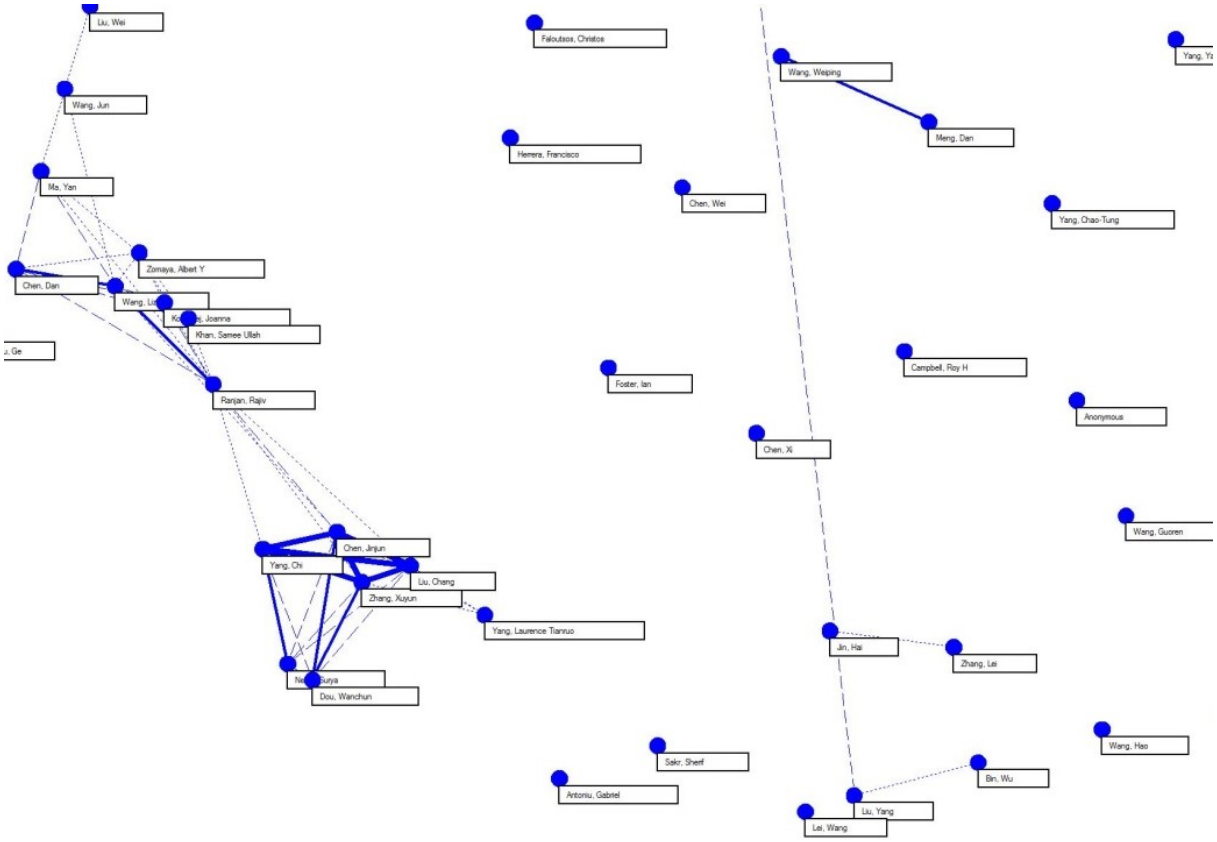

Figure 8. Collaboration among Prolific Big Data Authors [partial]

Publications per author for Non-Linear Programming are even fewer. More interesting, Normalized EScores for authors differ notably – peaking at 63 for DSSCs, 47 for Big Data, and only 14 for Non-Linear. The fields differ on degree of emergence. Based on Normalized EScoring, the top Non-Linear Programming author would reflect a single paper – reinforcing our preference for Total EScoring as a primary indicator of author emergence.

6.5. Comparing Research Domains based on Degree of Emergence

How do we propose to assess domain (or field) emergence? We view this as an open research question that we would approach by devising metrics based on EScores.

Here are ideas on composing a model to compare technical domains. For each technical domain being addressed, gather suitable 10-year datasets. For each of those – e.g., the current examples: Big Data, DSSCs, Non-Linear Programming – calculate measures such as:

- # of high EScoring terms
- # of records above a Total EScoring threshold; here we demonstrate a threshold of "10"; two of the three case examples in Table 5 show marked activity
- # of cutting edge organizations (rationale is to have more than a couple of organizations actively pressing a frontier); here two of the three domains show activity above the "100" threshold posed
- # of cutting edge authors – here the threshold is sensitive (key rationale is to have enough individuals to constitute a community); as with organizations, Non-Linear shows none.
- # of countries above threshold – here the differences among the three test domains are muted.

This formulation suggests that experimentation is warranted to identify suitable thresholds to use – given the purpose of comparing the degree of emergence of technical domains. The variations between the young, hot field (BD) and the 25-year-old one (DSSCs) are notable. DSSCs show more emergent terms and organizations and significantly more authors. That would seem to reflect the building of a substantial research community more than frantic activity at the frontier. By all five component measures, Non-Linear Programming is not an emergent technology.

Table 5. Illustrative Domain Emergence Measures

Measure \ Tech Domain	Big Data	DSSCs	Non-Linear Programming
ESc terms > 1.77	36	80	29
# of ESc records > 10	257	293	7
Organizations Total EScore >100	7	39	0
Authors Total EScore > 40	22	246	0
Country Total EScore > 40	30	26	10

7. Conclusions and Discussion

To sum up, this paper offers a viable R&D emergence indicator set based on four criteria: novelty, persistence, growth, and community. The paper introduces "Emergence Scoring" – providing a quantified metric to distinguish terms evidencing sharp, recent R&D activity, within a dataset under study. We provide a script to calculate EScores, as well as options regarding term manipulations, weights placed on components, thresholds (e.g., term appearance in how many records spanning how many years; high EScores being >1.77), etc. We demonstrate the process for three different topics – Big Data, Non-Linear Programming, and Dye-Sensitized Solar Cells. The process described yields viable indicators of 1) emergent terms and of 2) cutting edge players, plus ways to measure degree of record and technical domain emergence.

Section 6 explores ways that researchers and research program managers could gain value from application of these within-domain R&D emergence indicators. EScoring can point attention to papers that address cutting edge topics. It can be used to aggregate such research activity to focus on countries, organizations, or authors that are especially actively engaging these hot topics. At the end of this section, we note further uses that could be pursued – for one, to compare research domains to see which may warrant special attention or funding. "Zooming in" to use the R&D emergence indicators to identify cutting edge players, and then pursue further analyses of their research can inform competitive technical intelligence ends.

As noted in the Literature section, a number of approaches to measure "emergence" are being explored these days. Our approach focuses on R&D, with emergence based on four criteria (novelty, persistence, growth, and community) taken into consideration. This contrasts to approaches that measure change on a single dimension – e.g., burst analysis. Burst analysis is viable with scientific literature -- our target (along with patents) [34]. Chen notes that a burst detection algorithm [35] can detect sharp increases in a specialty of interest. While Kleinberg's algorithm [35] was devised to detect activity bursting of single terms, it can be applied to multiword phrases as well. It uses a probability density function to distinguish terms that present more rapidly than expected. Chen's CiteSpace II (2006) identifies research fronts based on bursts of terms extracted from titles, abstractors, and keywords in bibliographic records. The key distinction here is our positing of novelty, persistence, and community criteria. It would be interesting to compare results on given abstract record topical sets.

A new paper by Qi Wang [36] offers a multi-criteria approach to identify emerging research topics. That differs from our own in scope. Wang analyzes an encompassing swath of research literature – 10 years of WoS. "Research topics" are relatively "macro" in comparison to our terms, using Waltman and Van Eck's [37] direct citation based identification of some 4,000 topics. Those are gauged on Novelty, Growth, Coherence, and Scientific Impact. Sample topics measured include "graphene" and "solar cell." Contrast this to our "micro" exploration of ETs within domains such as DSSCs – a subset of solar cells. Again, comparison of approaches in detail would be intriguing.

We note *limitations* of our EScoring. Our approach relies on search and retrieval of abstract records, with their metadata, from R&D databases. Excepting very few (notably, MEDLINE), most ST&I databases require licensing. Search queries of a given domain by independent, experienced searchers tend to vary considerably, so resulting emergence indicators may differ accordingly. Indexing by those databases adds delay to the lags from discovery to publication or patenting. And, the content is inherently limited, as compared to full text resources. Further, patent abstracts' topical content is not inherently presented to communicate fully (Derwent second level data do strive to improve this).

How best to extract topical content? This can be done at a very discrete level, as we do with ETs, or at more aggregated levels. "ETs" provide great specificity, but at the cost of noisiness – i.e., many related term variations are distinct. As a consequence, in our present formulation, we don't emphasize exact terms highly. Many aggregation options can be explored, including topic modeling [38], clustering approaches, and factor analyses [e.g., we see promise in using VantagePoint's Principal Components Analysis (PCA) to consolidate ETs after they are generated, or prior to running the EScoring script]. Inherent in any of these is term processing. Here, we present an explicit approach to extract, clean, and consolidate terms. The aim is replicability, but there are many dimensions to consider in how best to treat terms. For instance, Table 2 incorporates "term folding up" to consolidate phrase variations, but one might thereby mask emergent variations.

We offer promise of predictive validity in two cases presented (BD, DSSCs) in the form of checking that ETs tend to increase in research activity in the following 3-year period. Non-Linear Programming ETs also remained active (Table 3). However, this needs to be extended to other topics and situations. In current studies we are pursuing on FLASH memory and Light-Emitting Diodes (LEDs), preliminary results of these commercialized technologies don't show strong future research activity for ETs.

Further research is in order to consider indicator behavior and alternative formulations. It would be interesting to compare, and possibly combine, lexical, term-based analyses with citation pattern analyses [15] to identify R&D topical emergence. We apply thresholds to meet novelty, persistence, and community criteria, but our EScores just reflect growth patterns -- see Wang [36] for exploration of measures of multiple criteria, also drawing on Rotolo et al. [5]. We have tracked emergent topics and cutting edge players over rolling increments. That is, for a target technology, we calculate the emergence indicators for, say, 1991-2000; then calculate them for 1992-2001; and so on. We find reasonable "staying power" as time advances [39].

Systematic examination of EScoring behavior for topics in different arenas is needed – i.e., to compare physical sciences, engineering, natural and biomedical sciences, and social sciences.

Among the options that warrant further assessment, we mention:

- Systematic comparison of a given topic based on searches in multiple databases, possibly combining such results as input to the EI script

- Exploration of shorter time periods – can we devise reliable emergence indicators from fewer than 10 years of data?
- Combining with other data for enriched technology and/or organization profiling. As touched upon here, publication activity and citation activity metrics complement the emergence indicators. One potential application might be inclusion of emergence indicators in the "research landscaping" service provided by the Chinese Academy of Sciences to help program managers judge the merits of research proposals.

We also are excited about new applications of these emergence indicators, such as:

- Comparing sets of related technologies using EScoring – e.g., various types of solar cells – to help gauge relative growth trajectories and innovation potential.
- Studying technology growth for a target technological domain by scoring various sub-systems and component technologies.
- Developing technological emergence workbooks – following the Clarivate Analytics offering of semi-automated Derwent patent profiling provided from a search set by the software as an MS Excel workbook.
- CTI use, as in profiling a target organization's patents with multiple measures, including EScoring on its various technologies with substantial R&D activity, to spotlight strengths and future potentials. Organizational emergence profiles could also be scripted akin to technological profiles noted in the previous bullet item. [Note that this implies analyzing search sets based on the organization instead of topical search.]
- Contributing to Technology Readiness Assessment (TRA), a metrics-based process used particularly in U.S. Defense Acquisitions to gauge the maturity of, and the risk associated with, a target technology [www.acq.osd.mil/chieftechnologist/publications/docs/TRA2011.pdf]. We see potential in emergence indicators contributing to Technology Readiness Levels to help benchmark the target technology's status and prospects.
- We envision diverse users of such technical emergence indicators. In addition to several suggested throughout the paper, one could see private equity decision processes benefiting from such empirical indicators to help sift through more vs. less attractive investment opportunities.
- We see opportunity in developing empirical indicators. However, we recognize irreplaceable value in engaging domain technical and market experts to help focus and scope such emergence profiling, check results (i.e., review high emergence term sets), and interpret findings.

References

[1] A.L. Porter, M.J. Detampel, Technology opportunities analysis, *Technological Forecasting and Social Change*, 49 (1995) 237-255.

[2] J. Alexander, J. Chase, N. Newman, A. Porter, J.D. Roessner, Emergence as a conceptual framework for understanding scientific and technological progress, *PICMET (Portland International Conference on Management of Engineering and Technology)*, Vancouver (2012); https://www.researchgate.net/profile/Jeffrey_Alexander/publication/248391856_Emergence_as_a_Conceptual_Framework_for_Understanding_Scientific_and_Technological_Progress/links/550847050cf26ff55f80c7b9.pdf

[3] S.F. Carley, N.C. Newman, A.L. Porter, J. Garner, An indicator of technical emergence, *Scientometrics*. (to appear).

[4] J.J. O'Brien, S. Carley, A.L. Porter, ClusterSuite [software script available at www.VPInstitute.org].

[5] D.D. Rotolo, D. Hicks, B.R. Martin, What is an emerging technology? *Research Policy* 44 (2015) 1827-1843.

[6] W. Boon, E. Moors, Exploring emerging technologies using metaphors: A study of orphan drugs and pharmacogenomics, *Social Science & Medicine,* 66 (9), 1915–1927, 2008.

[7] T.S. Kuhn, *The structure of scientific revolutions*. Chicago: University of Chicago Press (1962).

[8] M. Li, A.L. Porter, A. Suominen; Relationships between emerging technology and disruptive technology/innovation: A bibliometric perspective, *Technological Forecasting and Social Change*, (under review).

[9] R.O. van Merkerk, D.K.R. Robinson, Characterizing the emergence of a technological field: Expectations, agendas and networks in Lab-on-a-chip technologies. *Technology Analysis & Strategic Management*, 18 (3-4) (2006) 411-428.

[10] J. Goldstein, Emergence as a construct: History and issues. *Emergence* 1 (1) (1999) 49-72.

[11] J. de Haan, How emergence arises. *Ecological Complexity* 3 (4) (2006) 293–301.

[12] A.L. Porter, J.D. Roessner, X-Y. Jin, N.C. Newman, Measuring national emerging technology capabilities, *Science and Public Policy,* 29 (3) (2002) 189–200.

[13] H. Small, K. Boyack, R. Klavens, Identifying emerging topics in science and technology, *Research Policy*, 43 (8) (2014) 1450–1467.

[14] I.D. Lacasa, H. Grupp, U. Schmoch, Tracing technological change over long periods in Germany in chemicals using patent statistics, *Scientometrics*, 57 (2) (2003) 175–195; doi:10.1023/a:1024133517484.

[15] W. Glänzel, B. Thijs, Using hybrid methods and 'core documents for detecting and labelling new emerging topics, *Scientometrics,* 91 (2) (2012) 399–416.

[16] B.R. Martin, Foresight in science and technology, *Technology Analysis and Strategic Management,* 7 (2) (1995) 139–168.

[17] A.T. Roper, S.W. Cunningham, A.L. Porter, T.W. Mason, F.A. Rossini, J. Banks, *Forecasting and Management of Technology*, 2d edition, New York: John Wiley (2011).

[18] B.C. Stahl, What does the future hold? A critical view on emerging information and communication technologies and their social consequences. In M. Chiasson, O. Henfridsson, H. Karsten & J. I. DeGross (Eds.), *Proceedings Researching the Future in Information Systems: IFIP WG 8.2 Working Conference, Future IS*. Turku, Finland (2011) 59–76; Heidelberg: Springer.

[19] J. Staudt, H. Yu, R.P. Light, G. Marschke, K. Borner, B.A. Weinberg, High-impact and transformative science (HITS) metrics: Definition, exemplification, and comparison, *Blue Sky Forum on Science and Innovation Indicators*; (2016) cns.iu.edu/docs/publications/2016-weinberg-bluesky.pdf.

[20] L. An, X. Lin, C. Yu, X. Zhang, Measuring and visualizing the contributions of Chinese and American LIS research institutions to emerging themes and salient themes, *Scientometrics* 105 (3) (2015) 1605-1634.

[21] M. Girvan, M.E.J. Newman, Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99 (12) (2002) 7821–7826.

[22] A.L. Porter, S.W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage* New York: John Wiley & Sons (2005).

[23] A. Kontostathis, L.M. Galitsky, W.M. Pottenger, W. M., S. Roy, D.J. Phelps, A survey of emerging trend detection in textual data mining. Survey of Text Mining, 185-224 (2004).

[24] C. Choi, Y. Park, Monitoring the organic structure of technology based on the patent development paths, *Technological Forecasting and Social Change,* 76 (6) (2009) 754–768; doi:10.1016/j.techfore.2008.10.007.

[25] M.M. Hopkins, J. Siepel, Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better)? *Technology Analysis and Strategic Management,* 25 (6) (2013) 655–685; http://dx.doi.org/10.1080/09537325.2013.801950

[26] P-L. Chang, C-C. Wu, H-J. Leu, Using patent analyses to monitor the technological trends in an emerging field of technology: A case of carbon nanotube field emission display, *Scientometrics*, 82 (1) (2010) 5–19; doi:10.1007/s11192-009-0033-y.

[27] J. Yoon, K. Kim, Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks, *Scientometrics*, 88 (1) (2011) 213–228; doi:10.1007/s11192-011-0383-0.

[28] Y. Huang, D. Zhu, Y. Qian, Y. Zhang, A.L. Porter, Y. Liu, Y. Guo, A hybrid method to trace technology evolution pathways: A case study of 3 D printing, *Scientometrics*, 2017; doi: 10.1007/s11192-017-2271-8.

[29] A.L. Porter, J. Youtie, P. Shapira, D.J. Schoeneck, Refining search terms for nanotechnology, *Journal of Nanoparticle Research*, 10 (5), 715-728 (2008); Online First: 10.1007/s11051-007-9266-y.

[30] S.K. Arora, A.L. Porter, J. Youtie, P. Shapira, Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs, *Scientometrics*, 95 (1), 351-270 (2013); DOI: 10.1007/s11192-012-0903-6.

[31] Y. Guo, C. Xu, L. Huang, A.L. Porter, Empirically informing a technology delivery system model for an emerging technology: Illustrated for dye-sensitized solar cells, *R&D Management*, 42 (2) (2012) 133-149.

[32] Y. Huang, J. Schuehle, A.L. Porter, J. Youtie, A systematic method to create search strategies for emerging technologies based on the web of science: illustrated for Big Data, *Scientometrics*, 105 (3) (2015) 1-18; doi: 10.1007/s11192-015-1638-y.

[33] X. Zhou, A.L. Porter, D.K.R. Robinson, M.S. Shim, Y. Guo, Nano-enabled drug delivery: A research profile, *Nanomedicine: Nanotechnology, Biology and Medicine.* 10 (5) (2014) 889-896; http://dx.doi.org/10.1016/j.nano.2014.03.001.

[34] C.M. Chen, C.M., CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*. 57 (3): 359-377 (2006); see also http://cluster.cis.drexel.edu/~cchen/citespace/.

[35] J. Kleinberg, Bursty and hierarchical structure in streams. *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (91–101), Edmonton, Alberta, Canada: ACM Press (2002).

[36] Q. Wang, A bibliometric model for identifying emerging research topics, *Journal of the American Society for Information Science and Technology*, 69 (2), 290-304 (2017); https://doi.org/10.1002/asi.23930

[37] L. Waltman, L., N.J. Van Eck, N.J., A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63, 2378–2392 (2012).

[38] S. Ranaei, A. Suominen, A. Porter, S. Carley, Identifying technological emergence using text mining and machine learning, *Technological Forecasting and Social Change* (in process).

[39] S. Carley, A.L. Porter, NC. Newman, J.G. Garner, A measure of staying power: Is the persistence of emergent concepts more significantly influenced by technical domain or scale? *Scientometrics,* 111 (3) (2017) 2077-2087.

Acknowledgements

This research has been supported by the National Science Foundation under EAGER Award #1645237 for a project, "Using the ORCID ID and Emergence Scoring to Study Frontier Researchers." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

This paper is closely based on Garner, J., Carley, S., Porter, A.L., Newman, N.C. (2017), Technological emergence indicators using emergence scoring, *Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR.

We thank colleagues at Georgia Tech's Program in Science, Technology & Innovation Policy (STIP) for their contributions to the project noted, and, particularly, to reflection on tech emergence (especially, visiting scholar, Serhat Burmaoglu, with a paper devoted to that in preparation), and development of emergence indicators (especially, Jan Youtie). We are hopeful about pursuing development of "Indicators of Technological Emergence" via a project presently under consideration by NSF.

Supplemental Materials

Extensive supplemental materials are available at: http://hdl.handle.net/1853/59335. These are too extensive to warrant appending to the article. They provide details on search strategies; figures showing the additional topics not shown in the article comparing total and normalized emergence scoring results for countries, organizations, or authors; and other supporting details.