Transforming Science Prioritization Processes Using Artificial Intelligence

Harley A. Thronson[a], Brian A. Thomas[b], Louis Barbier[c], & Anthony Buonomo[d]

[a]NASA GSFC (retired), [b]NASA GSFC, [c]NASA HQ OCS, [d]Senior Consultant

Artificial Intelligence (AI) and Machine Learning (ML) have potential to augment significantly the current labor-intensive processes of science prioritization, specifically by the National Academies' Decadal Survey on behalf of NASA and NSF. Here we summarize what we believe to be the first exploratory demonstration-of-concept results from an application of AI/ML to Survey science prioritization. Specifically, we applied Latent Dirichlet Allocation (LDA) and Natural Language Processing (NLP) to reveal trends in published astrophysics research that may indicate science priorities and which could be applied to strategic planning. For the purpose of the work that we summarize here, AI/ML is able to analyze – that is, to "understand," in a manner of speaking – a vast amount of text to reveal complex relationships among research topics, including the growth or decline of science community activities in those topics over time. We trained ourselves and AI/ML algorithms by using ~400,000 abstracts in the period 1998 to 2010 to "forecast" the Academies' Astro2010 recommendations and compare with the solicited white papers. Comparing our results with actual Astro2010 recommendations allowed us to identify candidate metrics that better predicted the actual results of the Survey. We found, for example, that Compound Annual Growth Rate (CAGR) of papers published in a topic area is a good proxy measure for importance of this topic area of research. With this training complete, we identified candidate astrophysics astrophysics science priorities for the 2021+ period using the research during 2007 - 2019 . We conclude that appropriate application of AI can potentially significantly reduce the current workload of the Decadal Survey processes and reveal otherwise unrecognized characteristics in the body of astronomical research.

We emphasize throughout the exploratory nature of our work, encouraging colleagues to pursue promising results further. Our most critical governing assumption was that increased (or decreased) research activity can be used to identify scientific or technology topic areas worthy of increased (or decreased) future emphasis. We discuss advantages, limitations, and recognize the "black box" nature of our technique. We note ethics issues associated, for example, with using AI/ML to reveal "hidden" meanings and biases in published work. Furthermore, inevitable improvements in AI may soon enable widespread and welcome identification of and advocacy for science and technology priorities by disparate and diverse groups and organizations. Consequently, we continue to urge a near-term, in-depth evaluation of appropriate applications of AI, including implications and consequences, as well as support for multiple follow-on assessments, of which ours is only a beginning.

One of the critical planning activities in the sciences is the establishment of credible priorities that will govern future investment priorities. The most highly regarded of the national processes of scientific prioritization for America's funding agencies is the National Academies' Decadal Surveys. Starting with the first decadal survey in astronomy and astrophysics in 1964, the range of science encompassed by the

Decadal Surveys has grown to include planetary sciences, heliophysics, Earth science, and biological sciences.

Today, the Decadal Surveys are the "gold standard" for strategic planning in the sciences (and elsewhere) largely because of a handful of highly respected characteristics: (1) significant, multi-year commitment to their creation by the Academies, the funding agencies, and the individual participants; (2) participation of highly regarded subject matter experts chosen to represent key demographics and professional specialties; (3) transparency and openness to community input, including mid-decade reviews; and (4) a record of success, as shown by the national commitment to fulfill the recommendations of previous Surveys.

One of the characteristics of the Decadal Surveys is how remarkably little they have changed fundamentally over the past half-century (Dressler 2016): a central steering committee of a couple dozen members supported by several specialty panels that evaluate candidate priority science and technology goals. One of the concerns is that the process may no longer be optimal for identifying the key science to be undertaken by the community.

Among the principal challenges faced by this process stem from Survey panelists' necessity to assess a large -- and rapidly growing -- amount of relevant information, specifically many tens of thousands of research papers. If this seems to be an exaggeration consider that scientific research publication is no longer constrained to just a handful of premier research journals. In the past few decades the rise of open access journals has greatly increased the number of refereed journals. For example, Bornmann and Mutz (2015) have estimated the global rate of increase in scientific publication at about 8 to 9% yearly, which means the number of research papers published in a given year doubles in less than a decade.

In addition, the Surveys have for some time been formally soliciting white papers from the broad community and receiving in response several hundred, largely advocating on behalf of mission concepts, candidate new science and technology goals, and improvements to the state of the professions. A further complication is that these contributions seem not intended to represent the full range of activities of the research community. Various areas may be over represented while others are absent. In recent years, some Survey panels have solicited inputs from the community when they determined areas were absent.

In the context of the Surveys, AI brings a major advantage: it can vastly increase the number of sources of information (inputs) while providing only a few highly condensed materials (outputs) for review by the Surveys' panels. Many of the candidate inputs may be entirely novel for the Survey, which will increase the breadth and diversity of factors under consideration by AI for making strategic recommendations. AI can also augment the existing inputs, for example, by providing narrative summaries and statistics on the contents of the numerous white papers. Similar efforts are underway in other disciplines, for example, the SemNet project for Quantum Physics literature (Krenn et al. 2019).

The potential input materials have thus increased vastly over the years in both variety and quantity, while the basic processes of the Surveys – and other strategic planning activities – are little changed over the same period. This leads to our central concern: Does the current process actually identify the best science to be performed? We believe that the Survey process is at a watershed moment: the abundance of information must be better assessed and existing methods of panel work should be augmented. We believe that artificial

intelligence (AI) and Machine Learning (ML) offer a possible solution to the current challenges and is sufficiently mature to be considered seriously as a new tool for the Survey.

Use of AI and Implications for Science Prioritization

Although the work that we report on here is unambiguously exploratory in nature, some likely implications seem obvious, which we discuss below in turn as examples of issues that the funding agencies and National Academies should consider, specifically

- *Significant reduction in workload for science and technology prioritization*
- *Revelation of otherwise unknown science, technology, facility, and cultural relationships*
- *Proliferation of science and technology priority surveys and possible mitigation of biases*

Workload Reduction. The premier science and technology prioritization process for much of the federal government is the National Academies' Decadal Surveys, which – for the purpose of demonstration – consume roughly 200 subject matter experts (SMEs) intermittently for, say, a year's work for one-half day per week or about 15 work-years. [Some participants work less; some very much more. The reader is encouraged to adopt alternative estimates.]

The initial work that we present here consumed four non-SMEs for six months at about one-quarter time on average, or about 0.5 work-year.

The comparison between these two estimates is scarcely reasonable, as the Decadal Surveys encompass much more than simple science and technology prioritization. They make every effort to involve a large community of colleagues, undergo in-depth review and costing exercise, and prepare and present their extensive findings. That said, with plausible advances in AI in the coming years, it seems reasonable to expect a reduction in workload for Astro2030 approaching an order of magnitude in comparison with Astro2020 *for a similar product.* Consequently, any savings in workload may be – should be? – accompanied by additional new tasks not able to be undertaken by previous Surveys that were limited to current processes.

We note that an AI-enabled substantial reduction in the workload to produce one of the core goals of the Decadal Surveys, science prioritization, means that some interesting analyses might be possible that heretofore were vastly too time-consuming. We identify some of these studies in our "Next Steps" section.

Revelations. Artificial Intelligence is notorious for producing unanticipated or mysterious results and the same is true for the initial work that we report on here. As noted in the discussion of our training activity in the next section, our process produced some high-priority topics that were made up of individual terms whose relationship was difficult for us to understand: "star formation" with "climate change" or "black hole" with "impact crater." These unusual combinations may be revealing of an underlying relationship that required more time and/or sophistication than we possess. Or, equally likely, it is the algorithmic product of the requirement in our adopted algorithm to produce a specific number of topic areas, one way or another, as we discuss in the next section.

Note that we removed, or "blacklisted," a few hundred terms that we forbade the AI algorithms to include in the analysis, which included, for example, specific missions, astronomical facilities and instruments, and organizations of any kind. In subsequent work, it would be interesting to include these terms to examine what relationships AI finds among them and science topics, and how it prioritizes them. It is this type of result which might be most useful to funding agencies attempting to improve their portfolio of investments.

Proliferation of science priority surveys. The relative ease of producing at least an initial identification of candidate future science priorities as we identify them here suggests a possible "bias-free" process that uses only the growth (or decline) in cumulative interest in a topic to represent the science community's growing (or declining) priorities. [The authors are not naïve to believe that the process we describe here is entirely and truly "bias-free."]

Some aspects of science prioritization become more transparent by appropriate application of AI, although there remains significant opaqueness in current AI processes to the point that the algorithms are often referred to as "black boxes." As we point out in our section on science prioritization for 2021+, some of the results produced by our AI applications were ambiguous to us and difficult to interpret. There has been notable and increasing success in the development of programs to "explain" AI results. However, we did not have the resources yet to apply this.

Perhaps the most consequential implication for science prioritization results from how straightforward and inexpensive – by any measure – use of AI is, as we demonstrate by our study in a preliminary way. That is, a dozen or so SMEs, equivalent to a modest-sized astronomy department, could approximately duplicate some of the Decadal Survey results (and much else) over a few months of part-time effort. Indeed, there is no reason that multiple Decadal Survey-type reports or equivalents could not be produced as often as resources and motivation permit. Competition among such surveys would presumably produce improved community-based science and technology goals. Coordination and competition among such products will be interesting and should be productive.

AI Training: "Predicting" the Astro2010 Science Recommendations

Artificial Intelligence (AI) systems typically require training of some kind for successful application. In our exploratory work, we trained using custom software in conjunction with open-source AI algorithms and supporting materials to "predict" the outcome of the Astro2010 Decadal Survey. We emphasize again that our essential assumption was that, properly analyzed, the published body of research accurately reflects the interests and priorities of the community of active astronomers.

We downloaded for analyses the metadata from the Astrophysics Data System (ADS) Abstract Service for all papers from 1998 to 2010. These metadata have several fields, including title, abstract, year, authors, affiliations, and database. We applied three filters to the data to remove (1) papers that are not unambiguously on astronomy, (2) papers which do not have a valid year, and (3) papers with abstracts less than one hundred characters in length (including whitespaces). The resulting data set consists of about 425,000 abstracts. About filter #1, planetary science papers in the ADS, as well as heliophysics and geophysics studies, generally were included in our analysis, as we will show.

(i) Extracting information-rich features from the abstracts. Next, we extract so-called "terms" from the titles and abstracts of these papers using the textacy implementation (6) of *SingleRank* (Wan and Xiao, 2008 (7)) powered by scispaCy. Studies have shown that *SingleRank* performs better than other unsupervised term extraction algorithms on short, abstract-length texts (Papagiannopoulou and Tsoumakas (8). As our process uses it here, a "term" is (typically) a one- to three-word science noun. Examples include "black hole," "bl lac," "star formation rate," "agn," and so on. This process produces several hundred candidate "terms" from the data. To be clear at this point, the authors did not in any way choose the astrophysics terms used in our analyses. Those choices were purely a result of application of the AI algorithms to the ADS abstracts.

Note that for our analysis, we used only abstracts, in part because of the smaller size of the abstract relative to an entire article, although mainly because abstracts typically have far more concentrated useful information to our process than does an entire paper.

The set of terms is then curtailed based on frequencies and *SingleRank* scores.

Artificial Intelligence systems are famously devoid of common sense and are entirely ignorant of the subject matter when training is initiated. Therefore, we manually had to inspect the most common terms initially identified by our algorithms and remove those that are not plausibly valid astronomy terms. Examples of these excised terms, which can number 300+ in our processes, include "new result," "main characteristic," and "observational program." We recognize that in many cases, it is open to discussion whether a candidate term should be excised. Examples of this are "high resolution," "large structure," and "spectral resolution." In general, we tended to be strict about allowing only terms that were unambiguously scientific.

(ii) Topic modeling and the stochastic nature of the analysis. The primary output of our adopted process is a large number of "topics," each of which is a mathematical compilation of many terms derived as described above (see, e.g., Figure 1). Using the hundreds of terms as features, we train a set of topic models using gensim's implementation of Latent Dirichlet Allocation (LDA (9). *Note that this algorithm is stochastic*. That is, re-running the algorithm on the same data with the same parameters, changing only the required random initialization, leads to slightly different results.

Our implementation is stochastic because it relies upon Variation Bayesian Inference. Exact *post hoc* inference of LDA is intractable to compute (10). Each trained topic model may have a differing number of topics, but all other parameters are the same. A topic model is chosen for continued analysis by identifying the model which has the highest *UMass* (11) coherence score. For our analyses, we eventually adopted topic numbers of 350.

Because our implementation is stochastic, we cannot unambiguously identify candidate science priorities from a single run of our adopted algorithms. Instead, as we describe near the end of the next section, we have adopted a method that uses multiple runs.

Nota bene: *Topics are not and should not be confused with, for example, "keywords" or other similar elements produced for scientific text.*

(iii) Calculating time series. Using the chosen topic model, we calculate the "topic evolution with time" using the ~425,000 abstracts in our dataset. These distributions allow us to create time series for each topic by summing each abstract's fractional contribution to each scientific topic for each year. We present the output of this process more fully in the next section on how we derive candidates for priority science in the 2021+ time period.

For example, assume our model has three scientific topics, each consisting of a number of terms. Assume abstract A has a fractional contribution for those three topics of 0.8, 0.2, and 0.1 "counts," respectively, and paper A was published in 2000. In this case, paper A contributes 0.8 counts to topic 1 for the year 2000, 0.2 to topic 2, and 0.1 to topic 3. Our algorithm then sums the relative contribution of counts to each topic from of all 425,000 abstracts. [How we use "counts" is described more fully in the following section.]

Finally, we calculate several characteristics of these time series, including the Compound Annualized Growth Rate (CAGR) of counts, a measure of the growth – or decline – of a topic over time in the ADS compilation. As a result, we can now explore topics, their top (i.e., most-frequent) contributing terms, and measures of their growth or decline within the abstracts over time.

Identifying Candidate Astrophysics Priorities for 2021+

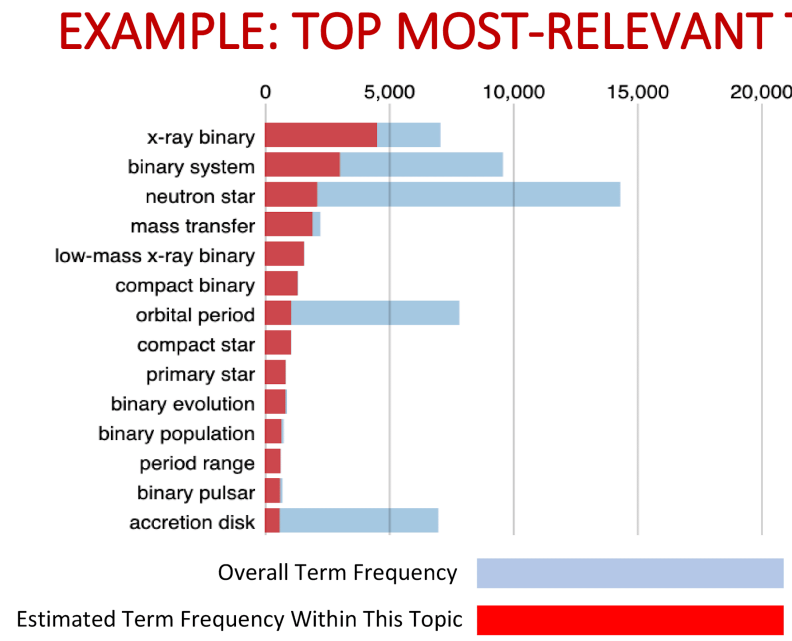
Based on the training by "predicting" the priorities for the Astro2010 Decadal Survey, we used two techniques to seek candidate astrophysics science priorities for 2021+.

Note that in this exploratory work, our results should not be interpreted as predicting the priority science recommendations for the current Astrophysics Decadal Survey, Astro2020, which incorporates substantially more into its considerations than the data that we used to the exclusion of other material. At the same time, we remain confident that the work that we describe here is the forerunner of more ambitious application of Artificial Intelligence to future Surveys and other planning activities.

We analyzed the metadata from the Astrophysics Data System (ADS) Abstract Service for all abstracts from 2006 to 2019, inclusive, assuming that this was a good representation of active research – and its behavior – during this period. We used the same three criteria as described for our analysis in the left column to prune the abstracts. The abstracts totaled about 450,000 for this period.

An example of one product of our analysis is shown in Figure 1, derived for our adopted number of 350 topics This bar graph shows what our AI algorithms identify as, in this example, the 14 most-relevant terms, as judged by "counts," that make up the topic, which we identify as evolution of a close binary system that results in an x-ray binary. Note again that we *neither* chose *a priori* the topics nor the terms that make up the topic. They were produced solely by the AI algorithms that we used for this analysis.

Our choice of 350 topics meant, in this case, that the output our AI process was 350 bar charts such as this one, a daunting task for manual analysis and far too much to display here. [See the definition of "count" in the previous section.] Moreover, note that subsequent iterations of our algorithms produce slightly different outcomes from the same data, as emphasized in the previous section.



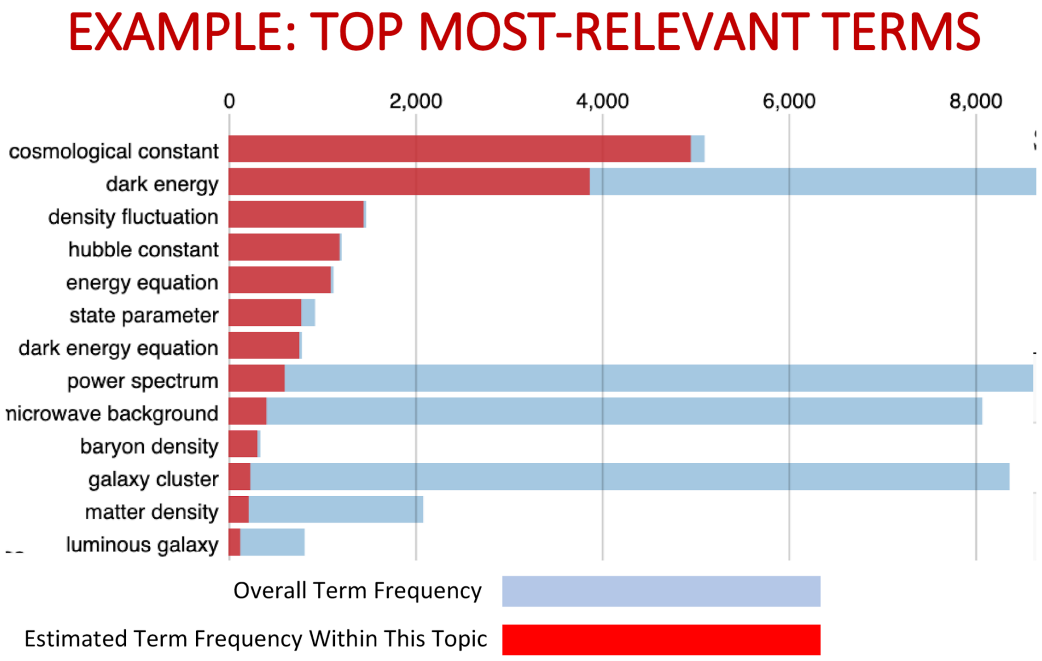


Fig. 1. Examples of the relative contribution of the 14 (top) and 13 (bottom) most-frequent terms to two of the 350 topics produced by our process. The graphs shows both the estimated frequency in terms of counts (x-axis) of each term in the topic, as well as the overall term frequency within all 350 topics.

(ii) First method to identify scientific priorities based on evolving activity by the research community. In schematic Figure 2, we plot the Calculated Annual Growth Rate (CAGR) versus the count of papers, as defined in the previous section, in which the 350 topics appear. There are four distinct regions in our figures of this type that contribute to our conclusions: (1) candidate future priorities in the upper left, topics with high CAGR (aka, growing number of publications over time that include this topic); (2) popular (i.e., high-count) topics widely included in the abstracts, although little CAGR growth or decline (i.e., 0.04 > CAGR > -0.04 over 14 years); (3) topics with negative CAGR in the lower left, suggesting declining contributions to the astrophysics literature; and (4) the bulk of the topics, those points with both modest counts (<2500) and relatively flat CAGR (i.e., 0.04 > CAGR > -0.04 over 14 years). Note that a generally similar analysis has been undertaken within the last few years by R. Zeinio (private communication) for the Office of Naval Research.

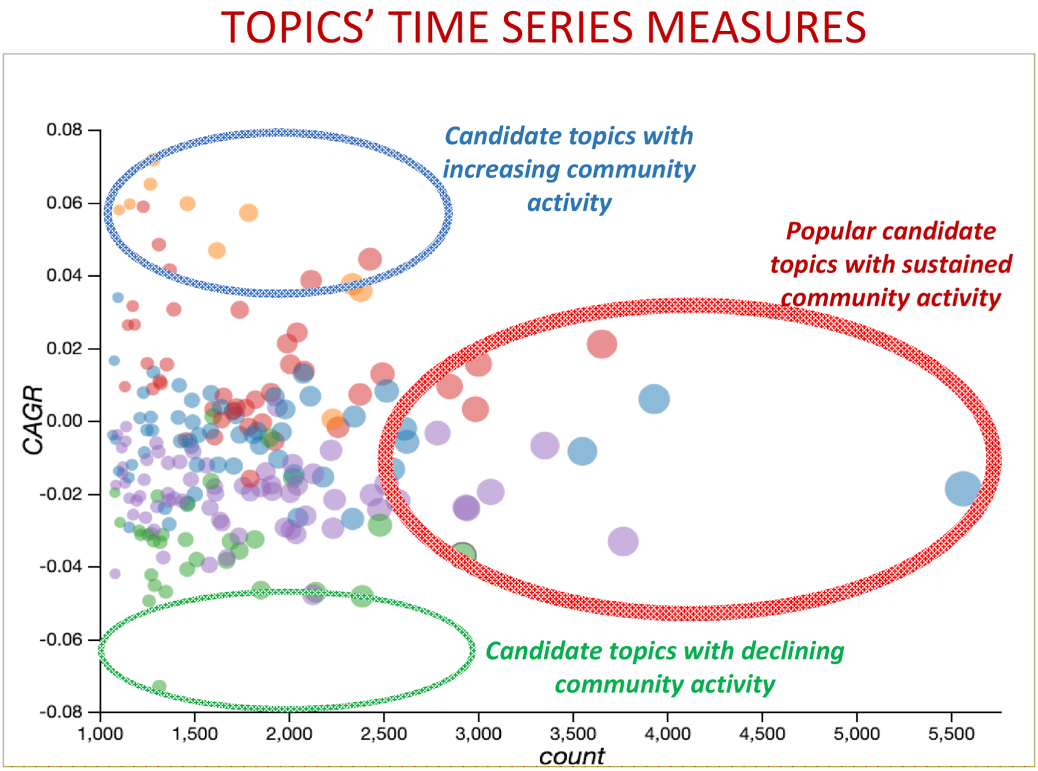


Figure 2. Schematic CAGR versus counts for those topics (of the 350 total in our analysis) with counts > 1000. Approximate regions of increasing, declining, and strong, but static, scientific activity are shown. We interpret that topics to the right in this chart are more frequently found in published research. Although the chart shown here should be interpreted by the reader as schematic,

the data plotted here was produced via the analysis described in our poster so is typical of the product of our adopted algorithms..

For the purpose of this demonstration, we adopt the view that science topics worthy of consideration for new investments, including missions and facilities, will be found in the first region (upper left in Figure 2). Those topics worthy of consideration for sustained investments, including support for missions and facilities, are found in the second region, whereas science topics found in the third region seem to be of declining community activity, as judged via use of the CAGR.

Among multiple interesting results of our analysis is that for our Astro2010 "training" analysis (previous section), we found an average CAGR of about 0.06; that is, ~ 6% annual growth rate in counts for 350 topics in our sample. Interestingly, this impressive growth rate appears not to have been sustained through the years preceding the analysis that we undertook for the period 2006 – 2019 (Figure 2). However, note that the 350 topics produced by our adopted AI algorithms for the 1998 – 2010 period are not the same as the 350 topics produced by the algorithms for 2006 – 2019. [Again, a consequence of the stochastic nature of these algorithms.]

The following three figures, of 350 produced each time our algorithms were run, exemplify the CAGR behavior of counts versus time for topics in the first three regions identified in Figure 2.

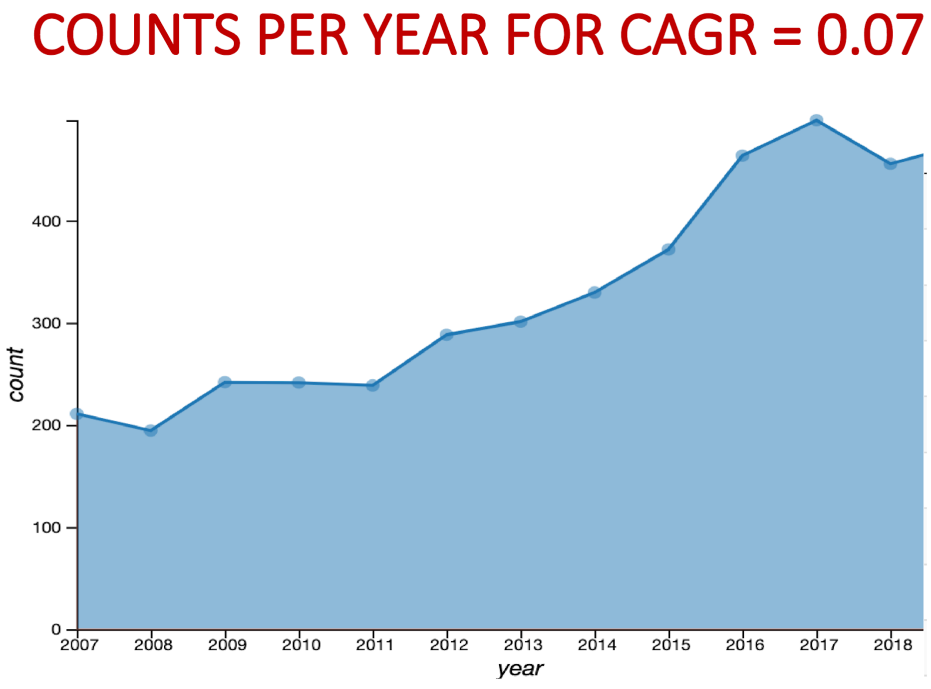


Fig. 3. Compound Annual Growth Rate (CAGR) of counts for a topic with CAGR = 0.07, which suggests a growing community research activity.

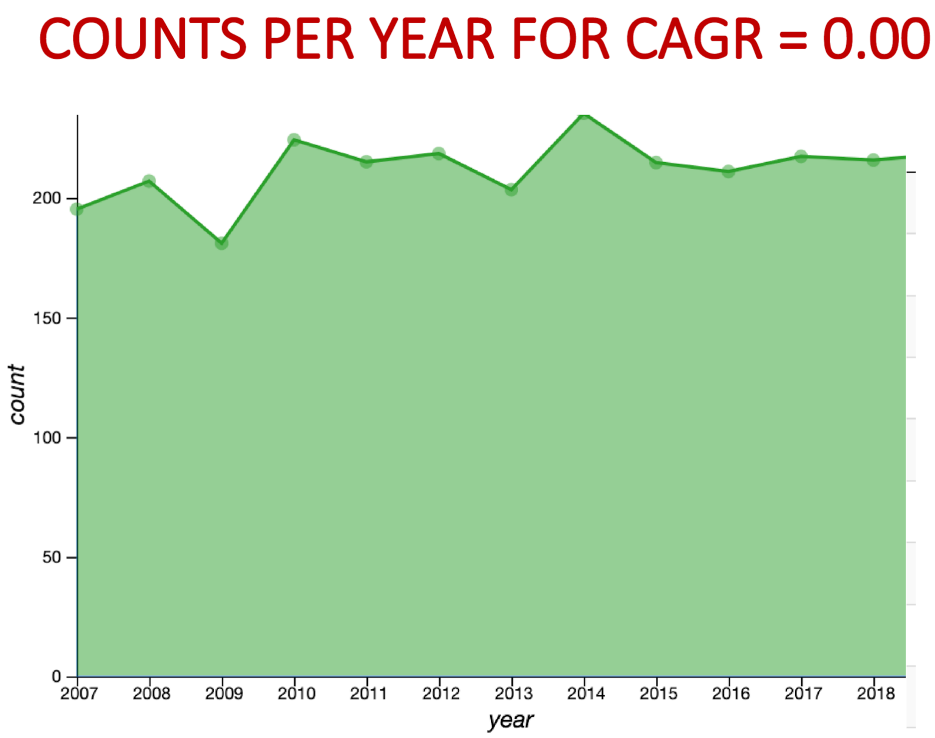


Fig. 4. Compound Annual Growth Rate (CAGR) of counts for a topic with CAGR ~ 0.00, which suggests a sustained, relatively long-term community research activity, which is also fairly broad-based, as the counts for this topic was high.

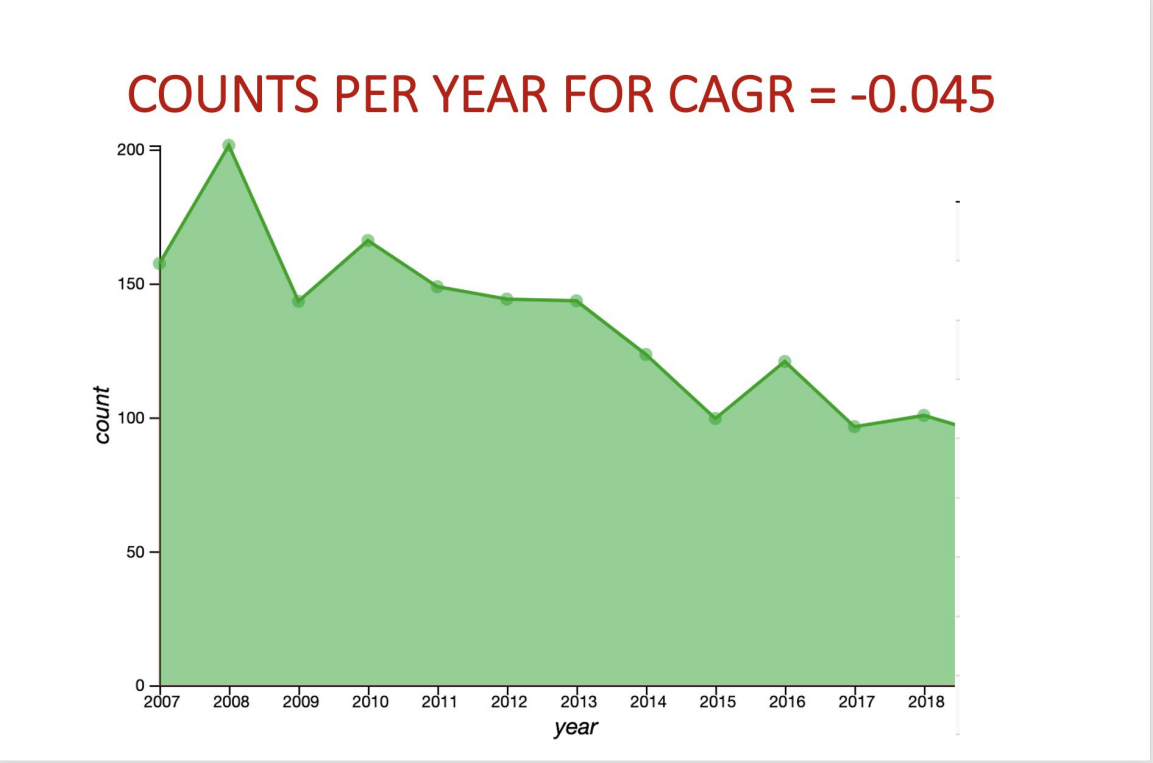


Fig. 5. Compound Annual Growth Rate (CAGR) of counts for a topic with CAGR = -0.045, which suggests a declining community research activity.

(iii) Second method to identify scientific priorities based on evolving activity by the research community. For our exploration of AI and science prioritization, we evaluated a second and complementary method to produce the results in our initial results in the next section.

The method that we summarize in the previous subsection has the advantage of easy visualization of the evolution over time of research activity; that is, by examining the CAGR for, perhaps, several dozen science topics. The challenge to such a brute-force method is having to examine at least several dozen products of our algorithms similar to Figure 1. And, moreover, to do it multiple times, as repeated iterations of our adopted algorithms have a stochastic component, producing slightly different output each time. Consequently, we examined how to combine algorithmically (i.e., without significant manual effort) multiple, repeated runs of our AI algorithms.

We utilized Cosine Similarity for this activity, where term counts in each topic formed the vectors to determine which topics in each run were repeatedly present in other runs. In order to form these vectors, we needed to prune the list of topic terms to only the most relevant. We utilized the "logprob" metric, which measures "term relevancy" (Sievert and Shirley (12)) to filter terms after we inspected a selection of topics, terms, and papers to identify the best threshold value for logprob (-2.5). Using these term vectors, we then inter-compared topics from separate runs to find the best match of topics among different runs. The value of the Cosine Similarity, which varies from 0 to 1, was noted for each best match topic pair and then added

to a running sum for topics with the identified terms. This sum is what we are referring to as "topic inter-run relatedness" (or just "relatedness"). In theory, the maximum relatedness score for any set of terms that identify topics would be $N^2 - N$, where N is the number of runs. In practice, because of term list degeneracies caused by term relevancy filtering, we may combine some topics within a run which were formerly distinct. For the work presented in this poster N = 4, with much larger values of N under construction.

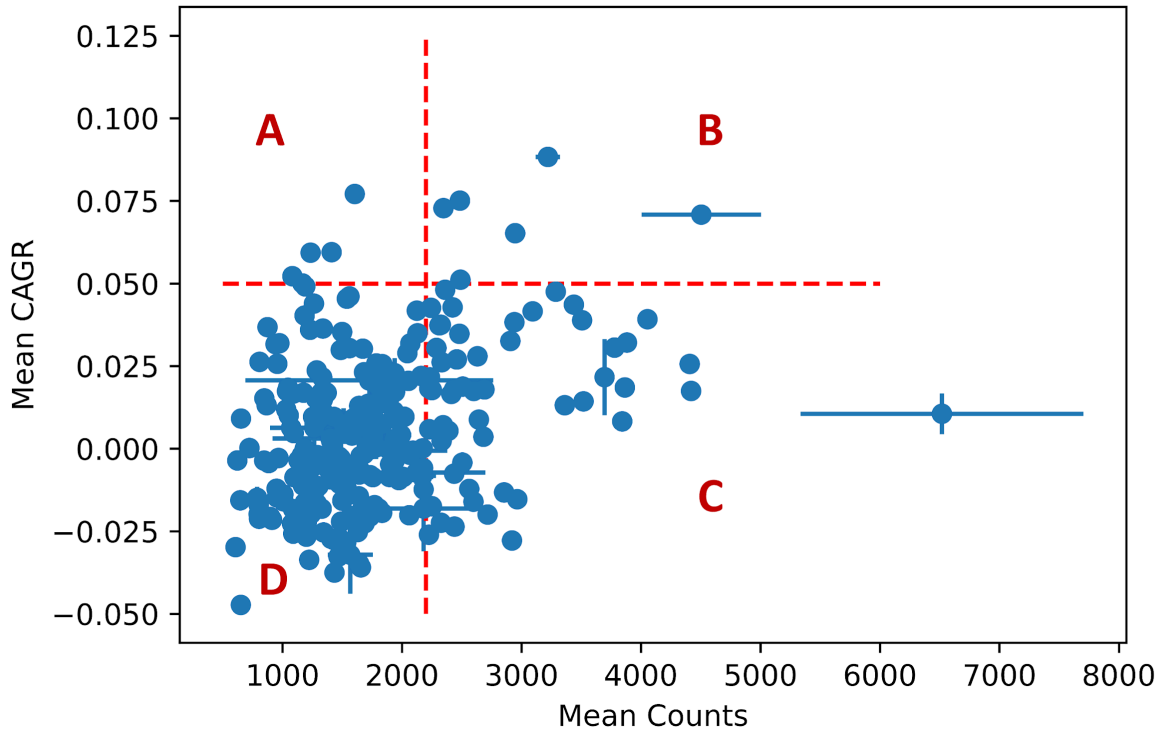


Fig 6. Example plot of the topics with highest interrelatedness, produced as described in the text. Two-sigma uncertainties are shown in cases where they are significantly larger than the dot.

Our initial result of applying this combination heuristic is shown in Figure 6, where the topics from the four separate runs were intercompared and those topics that had mean Cosine Similarity of 0.7 or greater kept. Estimated two-sigma uncertainties are shown. The two dashed red lines are drawn in the figure: horizontal at mean CAGR = 0.05 and vertical at mean counts = 2100. Both lines are about two-sigma above the mean of each parameter. We have identified four quadrants that possess characteristics relevant to prioritization of science topics: (A) modest term counts with high CAGR (i.e., growth in science activity); (B) high term counts with high CAGR; (C) high term counts with modest, low, or negative mean CAGR (< 0.05); and (D) modest term counts and modest, low or negative mean CAGR (<0.05).

We interpret the content of the four quadrants as (A) modest, but growing, science community activity; (B) strong and growing science community activity; (C) high and static (mean CAGR with two-sigma of 0.0) science community activity; and (D) modest and stable science community activity. That is, the topic areas of highest priority for guiding future investments seem likely to be found in the upper right, although

promising topics should be found in quadrant A. Topic areas most likely to be candidates for sustained investment seem likely to be generally found in quadrant C.

Initial Derived Science Priorities for 2021+ and Next Steps

(i) Initial derived science priorities for 2021+. It is the information in Figure 6 that we now use to derive candidate science priorities for 2021+, again bearing in mind the exploratory nature of our study.

These topics, and their associated relevant terms from quadrants A - C are respectively identified in Tables 1 - 3 below. Table 3 lists those twenty topics in quadrant C with the largest counts. We emphasize that the terms listed in the table were taken for demonstration purposes verbatim from the output of our process and, consequently, require further analysis. In addition, the terms appearing for each topic are a shortened list of the most-frequent terms. [See Figure 1 for an example of the large number of terms that may make up a topic.]

We note, for example, that Solar System objects (Mars, Vesta, etc.) appear prominently as for future or sustained astrophysics priorities. This simply reflects the significant appearance of Solar System objects in the ADS, as important astronomy target objects. It is straightforward to remove such terms in future analysis, if desired. Furthermore, although we removed a long list of missions and facilities from permitted terms, we overlooked a few in the iteration presented here: MSL and the Auger Observatory.

The behavior here of the Auger Observatory, incidentally, is an instructive example of one common and useful aspect of our type of analysis. Observations began about the first year of the ADS data that we assessed here (2006) and continues today. Thus, the topic that includes that observatory shows a very high CAGR, although a modest total number of counts so far, as expected. It will be interesting to observe the time-evolution of this topic – or any! – which is relatively straightforward via analyses of the type we describe here.

Table 3 High Counts, Modest CAGR (Quadrant C)

(2) Next steps

Although the work that we summarize here is purely exploratory, we are confident that Artificial Intelligence (AI) and Machine Learning (ML) will be applied extensively within very few years to analyze science activities. With that in mind, there are some activities that seem worthwhile to pursue in the near term, activities which we have suggested to NASA:

1. Building upon the work initiated here, a National Academies' assessment of applications of AI and ML for science and technology prioritization, among other applications.
2. Solicitation for proposals to fund multiple and more extensive follow-on to the work discussed in this paper.
3. A series of community workshops focused on use of AI and ML to augment current processes of science and technology prioritization and other applications.
4. AI-based analysis of effects, for example, of sky surveys, Great Observatory-type missions, theory programs, and professional conferences on growth of topics, which some of our work suggests may be both over- and under-appreciated.
5. Analysis of biases inherent in the type of analyses that we initiated here.
6. How to mitigate the pervasive "black box" characteristics of AI and ML.

ACKNOWLEDGEMENTS

This work was supported by the NASA Headquarters Science Mission Directorate and Dr. Michael Seablom, the Division's Chief Technologist. We benefitted significantly by Dr. Giulio Varsi's early recognition of the potential value of Artificial Intelligence for the purposes that we discuss here. Dr. Ryan Zeinio very helpfully discussed with us his analysis of candidate technology investment priorities for the Office of Naval Research.

---

[1] Author Contact: hthr0ns0n1@gmail.com
[2] Retired
[3] Heliophysics Science Division, NASA Goddard Space Flight Center

[4] Office of Chief Scientist, NASA Headquarters
[5] Senior Consultant
[6] https://chartbeat-labs.github.io/textacy/build/html/api_reference/information_extraction.html#module-textacy.ke.textrank
[7] Wan, X. and Xiao, J. (2008) Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008, 855–860. URL: http://www. aaai.org/Library/AAAI/2008/aaai08-136.php.
[8] Papagiannopoulou, E, Tsoumakas, G. A review of keyphrase extraction. WIREs Data Mining Knowl Discov. 2020; 10:e1339. https://doi.org/10.1002/widm.1339
[9] https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html#sphx-glr-auto-examples-tutorials-run-lda-py
[10] Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, *23*, 856-864.
[11] https://radimrehurek.com/gensim/models/coherencemodel.html
[12] Sievert, C. & Shirley, K. E., "LDAvis: A Method for Visualizing and Interpreting Topics," *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces,* p. 63 – 70, 2014.