Issue Brief

Staying Current with Emerging Technology Trends

Using Big Data to Inform Planning

Authors
Emelia S. Probasco
Christian Schoeberl

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

December 2024

Table of Contents

Introduction

Decision-makers today are pressed to stay ahead of the tsunami of new science and technology research. Many hope that big data and artificial intelligence (AI) will help identify research evolutions and revolutions in real time, or even before they happen. As we will discuss below, data alone cannot predict scientific revolutions. Examining data to stay current with, or slightly ahead of, new technologies, however, is still valuable.

This paper proposes a human-machine teaming approach to systematically identify research developments for an organization. First, our approach starts by identifying papers that the organization has authored. Second, we use those papers to find research clusters in the Center for Security and Emerging Technology (CSET) Map of Science, which displays global academic literature clustered according to citation patterns. Third, we select a subset of clusters based on metadata that we believe indicates important research activity. Fourth, we share the selected clusters with subject matter experts (SMEs) and facilitate a discussion about the research and its potential impact for the organization.

We describe each of these steps in detail in the sections that follow and use a proof-of-concept experiment to evaluate our approach.

This paper is intended for individuals developing research or investment portfolios and priorities within their organizations. It should also be useful to SMEs interested in exploring or revealing research to which they may not otherwise be exposed as a consequence of increasing specialization.

Background

Prior Research

Technological forecasting and horizon scanning aim to identify trends and predict technological advancements. Many organizations conduct horizon scanning or forecasting projects to shape their development and investment portfolios. While some organizations employ teams of individuals to network with top researchers and attend conferences to sense emerging trends, others have employed analytical teams to provide quantitative input to make decisions. These efforts range from governments to militaries to medical institutes to smart-home vendors and presumably many others.[1]

Methods for forecasting and horizon scanning can include quantitative and qualitative approaches, individually or in combination.[2] For example, one notable qualitative approach is the Delphi method, which aims to create a consensus among experts about probable future scenarios based on repeated surveys and group discussions.[3] By contrast, quantitative approaches may rely on trend analysis and modeling of research publications, patent applications, corporate filings, and venture capital funds, either individually or in combination.[4] Analysis of past technology forecasting efforts shows that approaches that combine quantitative analysis with human judgment are, in general, more successful.[5]

Increasingly, quantitative approaches are leveraging new data analysis techniques, including natural language processing (NLP) and large language models.[6] Furthermore, the ability to easily manipulate large amounts of data has spurred policymakers and researchers alike to desire tools they might easily manipulate to pursue a specific emerging technology question or need.[7] There are many databases available that can be leveraged for this purpose: bibliometric databases such as Scopus, SciVal, or Web of Science; patent tools such as Clarivate or Quid; and funding tools such as PitchBook and Crunchbase. CSET's very own Emerging Technology Observatory (ETO) platform similarly offers publicly available tools for exploring research, patent, and funding information.*

CSET's Merged Corpus, Research Clusters, and ETO Map of Science

In this paper we used CSET's data science resources as the foundation for the quantitative part of our proposed methodology, though some of the other

* CSET's ETO is available to the public at https://eto.tech/.

aforementioned tools could have been used in a similar process. [8] Among CSET's data resources is the merged corpus of scholarly literature, a dataset with over 259 million scientific publications from around the world. The corpus combines and de-duplicates publications from Clarivate's Web of Science, OpenAlex, the Lens, Semantic Scholar, arXiv, and Papers With Code, along with metadata about individual publications, including author affiliations, funding institutions, and citation count. The merged corpus serves as the foundation for CSET's ETO Map of Science, a visualization of the global academic literature organized into clusters based on citation patterns.[9] Research papers in the merged corpus that frequently cite each other can be grouped in what CSET calls research clusters. These clusters typically emerge when papers share a common research topic, though they may be connected via citation for other reasons as well. The more citations that link papers together, the more likely they are to be grouped in the same cluster. The ETO Map of Science includes nearly 87,000 clusters available for analysis.

Cluster Features and Metadata

Each research cluster contains not only metadata about individual papers but also combined statistics and characteristics about the cluster as a whole. This aggregate information helps summarize the collective content and trends across all papers within the cluster. Cluster-level metrics include, for example, the cluster's total number of publications, the most frequent authors, author institutional affiliations, the average date of publication for papers in the cluster, frequent keywords and phrases, and funding sources. These cluster-level metrics can be used individually or in combination to identify research areas with desirable characteristics.

Methodology

The following sections describe how we used the CSET data resources described above to inform an SME discussion about emerging technologies. Our approach builds on prior work in horizon and technology scanning, in addition to prior work in structured group moderation.[10] We use our proof-of-concept experiment with U.S. Department of Defense (DOD)–affiliated papers to further illuminate our approach throughout and conclude with observations and suggestions for alternative approaches. Our proposed methodology includes five steps:

1. Identify a set of papers authored by or deemed relevant to an organization. For example, in our proof of concept, we used a previously created dataset of research papers whose authors were affiliated with DOD institutions.[11]

2. Locate the clusters in the ETO Map of Science that contain the papers identified in step 1.[12]

3. Analyze the metadata of the selected research clusters to share with SMEs. Metadata can be analyzed in several ways, and the sections that follow go into greater detail on this step.

4. Select and prepare information about a subset of clusters for SME review and discussion.

5. Using the subset of clusters, facilitate a discussion with individuals who have expertise relevant to their evaluation. In our proof of concept, we recruited AI researchers who have worked extensively with the DOD.[13]

The sections that follow go into greater depth on these steps and initial observations from our proof-of-concept exercise with SMEs.

Step 1. Finding Papers Relevant to an Organization

To find research clusters relevant to an organization, we start with a collection of research papers of interest. Collecting the initial set of papers could have been achieved in several ways, such as searching for clusters that have a concentration of selected keywords in paper titles or abstracts.[14] But selecting keywords requires assumptions about which words are most relevant and could bias the results toward what is already known, as opposed to helping SMEs discover new, rapidly changing research.

Instead, we chose to use a set of papers presumed to be relevant because they were published or funded by the organization itself. This approach can support analysis for large institutions conducting and funding research, such as the National Institutes of Health (NIH), tech companies, or large nonprofit research labs. While collecting papers is a relatively straightforward task in principle, collecting and completing a quality analysis of the publications is time-consuming in practice. For the purpose of our proof of concept, we relied on prior CSET work that established a dataset of papers by the DOD and its affiliated organizations.[15]

With the initial list of papers in hand, we can identify research clusters that contain one or more of the papers and begin cluster-level analysis. Figure 1 displays the ETO Map of Science with all research clusters on the left (85,643 clusters) and with clusters containing DOD-affiliated papers colored in black on the right (34,387 clusters).

Figure 1. DOD-Affiliated Clusters Within the ETO Map of Science

Source: Authors' analysis.

Steps 2 and 3. Analyzing the Clusters

Once we have the clusters with papers from the initial set (in our case, DOD-affiliated papers), we can examine how the papers appear in the cluster to help determine which clusters to examine more closely. For example:

- Some clusters may contain many papers from the initial set, while others only have a few.
- Some papers from the initial set may be frequently cited within the cluster or simply highly cited papers generally (what we call "core" or "highly cited"), while others have fewer citation links within the cluster.
- Some papers from the initial set may be cited by papers in other clusters (what we call "exporting"), while other papers are exclusively cited by papers within their own cluster.[16]

For our proof of concept we focused on clusters that had features from the second and third bullet: they contained DOD-affiliated papers that were core or highly cited and DOD-affiliated papers that were exporting (some papers fell into both categories). For more on the advantages and disadvantages of these three features and how they were used in our proof of concept, see Appendix A.

Finding Candidate Clusters of Interest

Beyond identifying clusters that matter to the organization, we wanted to select clusters that are changing in a way that indicates that the research could be on the cusp of, or actively contributing to, new applications or products. Choosing what might qualify for this review is a judgment call, and many metadata indicators are available to choose from (see Table 3 for a list). For our proof of concept, we chose to focus on the clusters with the most papers published in the last five years (which we call "growth") and the clusters that appear to be transferring knowledge to other clusters as indicated by export citations (which we call "export activity"). For both of these metrics, we normalized the data to account for the different sizes of clusters as well as for the average ages of papers in a cluster. In our conclusion, we address some of the other metadata we could have used.

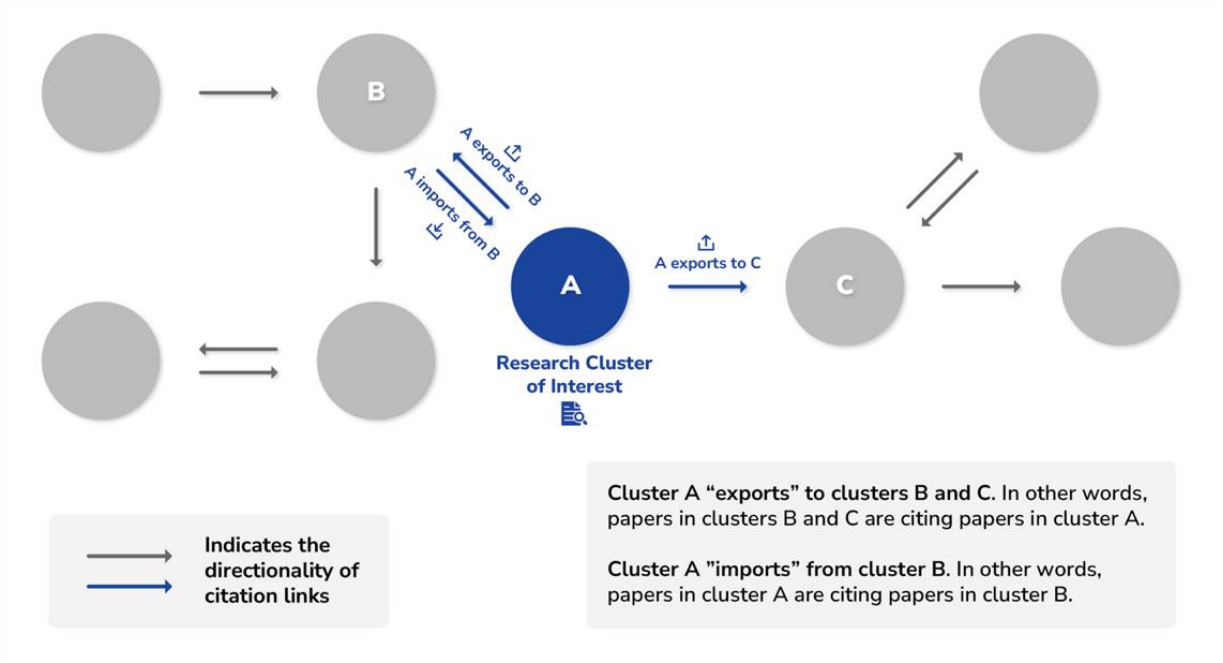
Growth

Given our goal to support discussions around emerging technology, we wanted to select clusters that include a lot of recently published research. Accordingly, we calculated the number of papers published each year in the last five years for each cluster (2019–2023) and then calculated how substantial that quantity was relative to other clusters of the same size and age. This gave us percentile ratings for each cluster in terms of recent growth. For example, a cluster with 90th percentile growth has added more papers in the past five years than 90 percent of all other comparable clusters.

Clusters Exporting Ideas

We also hypothesized that clusters that are highly cited might be indicative of knowledge transfer or application and therefore important for SME review.[17] We consider these citation relationships analogous to sourcing: authors import information from other papers to build on prior work (import citations), and they export their information to papers published at a later date that cite them (export citations). Figure 2 provides an example of a research cluster of interest (colored in blue) and tracks its citation linkages to other clusters.

Figure 2. Import and Export Citation Relationships Between Research Clusters



Source: CSET.

A high export citation percentile signifies that the cluster's research is impacting other clusters of research. (It could be just a few other clusters, or it may be many; for this analysis we did not differentiate.)

We calculate citation percentiles for each cluster as a whole. The export citation metric is equal to the average export citation percentile of the individual articles within the cluster that are published in the past five years.

For more on our percentile calculations, see Appendix B.

Step 4. Select and Present a Subset of Clusters for Human Review

Steps 1 to 3 help us identify and quantitatively analyze clusters. Step 4 prepares this quantitative analysis for SME review and contextualization. SMEs contextualize research when they provide insights into how that research might impact society. The challenge lies in finding and presenting the right data to SMEs in a setting that will prompt useful insights. This requires the selection of a realistic number of clusters for SME review and a clear presentation of the quantitative analysis.

The complexity of the data for each cluster limits the number that any SME can reasonably consider within a set period or in a single meeting. Experimenting with the number of clusters shown to SMEs was beyond the scope of this analysis. Instead, for our proof of concept we chose a set of key attributes based on our prior experience. We also established a percentile cutoff based on those criteria and our intuition that SMEs could review only 20 or fewer clusters. The research cluster inclusion criteria and rationale for our proof of concept are included in Table 1. The final set of clusters is presented in Table 2. Our choices simply reflect our prior experience working with research clusters as well as our prior work on DOD and AI research. The metadata we used can support future research efforts with different interests or goals, which we believe is an overall strength. Given different goals, other criteria could easily be selected based on the available metadata (for a list of available metadata, see Table 5).

Table 1. Goals and Cluster Selection Criteria for Our DOD Proof of Concept

Intended Goal:	Selection Criteria:
Identify clusters most likely related to the organization's interests	Select clusters with core or highly cited paper(s) authored or funded by the organization
Identify clusters where the organization's research may be translating to applications	Filter clusters that have one or more of the initial set of papers exported to a different cluster
Identify clusters more likely to have recently been translating research to applications, and reduce the overall number of clusters the workshop participants will have to review	Down select to clusters that have the highest percentile citation exports to other clusters. *Note that we chose the 80th percentile for our proof of concept, and this selection criteria correlated with clusters that had high growth (all greater than 70th percentile).*
Identify suitable clusters for review based on the expertise of the workshop participants	Down select to clusters that have a percentage of papers related to a given subject based on an NLP classifier. *In our proof of concept, we chose to use CSET's AI classifier to select clusters with 50 percent or more papers classified as about AI (more in Appendix C).*

Source: Authors' analysis.

Presenting Clusters and Communicating Data Limitations

Once selected, the research clusters must be presented to SMEs in a way that adequately communicates what the clusters are and how they can and cannot be used in answering certain questions. In addition to the spreadsheet shown in Table 2, each listed cluster had a link to the research cluster web page containing additional information (see an example in Figure 3). Of the information available on the cluster web page, SMEs mostly relied on the list of highly cited or core articles in the cluster to develop a deeper understanding of the subject areas covered in the cluster. Table 2 displays the clusters as they were sent to the SMEs in advance of the facilitated discussion.

Table 2. Clusters Presented to SMEs for Proof-of-Concept Event

Cluster	Average age of papers (years)	Papers classified as AI	Papers classified as NLP	Papers classified as robotics	Papers classified as computer vision	Extracted key phrases	Number of papers published in last 5 years (percentile)	Papers exporting to other clusters (percentile)	Unique clusters importing (percentile)
5167	3.24	90%	0%	5%	85%	neural radiance fields, point cloud, neural scene representation, neural rendering, point cloud generation	1.00	1.00	0.99
9301	4.30	93%	2%	0%	4%	deep neural networks, neural networks trained, deep learning, gradient descent, training deep neural	0.99	0.99	0.98

2658	5.77	81%	0%	77%	0%	soft robotics, soft pneumatic actuators, soft materials, soft robotics applications, soft crawling robots	1.00	0.99	0.99
64740	5.84	76%	0%	1%	1%	gradient descent, minimax optimization problems, convergence, optimistic gradient method, algorithms	1.00	0.99	0.93

22884	3.50	53%	0%	1%	1%	physics-informed neural networks, neural networks, partial differential equations, neural network approximation, Fourier neural operator	1.00	0.98	1.00
43601	9.70	60%	0%	41%	0%	model predictive control, stochastic optimal control, path integral control, optimal control problems, differential dynamic programming	0.96	0.98	0.92

416	11.23	76%	0%	20%	56%	point cloud registration, point cloud, cloud registration methods, iterative closest point, unsupervised point cloud	0.98	0.98	0.82
3205	8.21	57%	0%	0%	18%	low-rank matrix recovery, matrix completion, matrix recovery problems, robust principal component, principal component analysis	0.94	0.97	0.93

51390	6.04	73%	0%	0%	33%	convolutional sparse coding, deep neural networks, sparse coding, deep learning, dictionary learning algorithm	0.98	0.96	0.93
59024	7.62	61%	1%	0%	10%	hypergraph neural networks, hypergraph representation learning, hypergraph node classification, real-world hypergraphs, dynamic hypergraph learning	0.95	0.96	0.92

7185	6.40	84%	0%	5%	79%	multiple object tracking, tracking multiple objects, object detection, object tracking methods, tracking performance	0.98	0.94	0.82
2812	11.95	92%	0%	91%	0%	quadruped robot, legged robots, robot leg design, dynamic multilegged robots, dynamic locomotion control	0.92	0.94	0.74

6265	7.72	95%	0%	93%	1%	continuum robots, soft continuum manipulators, soft robot, cable-driven continuum robot, soft robots modeling	0.98	0.94	0.90
45546	8.05	81%	0%	65%	1%	informative path planning, Gaussian process, multi-robot information gathering, Bayesian optimization, reinforcement learning	0.97	0.93	0.93

46074	6.65	61%	0%	1%	28%	principal component analysis, linear discriminant analysis, robust principal component, proposed method, component analysis based	0.99	0.91	0.84
1994	7.94	71%	0%	0%	67%	image forgery detection, image copy-move forgery, copy-move image forgery, forgery detection methods, image splicing	0.98	0.90	0.74

1125	11.78	76%	0%	0%	59%	hyperspectral unmixing, unmixing methods, real hyperspectral data, proposed method, spectral unmixing problem	0.92	0.90	0.74

Source: Authors' analysis of CSET's research clusters.

Note: All clusters have a core paper from the DOD-affiliated corpus and the DOD-affiliated papers are cited by other clusters.

Figure 3. ETO Map of Science Research Cluster Web Page Excerpt

Source: ETO Map of Science.

In addition to the spreadsheet and web page, SMEs must be alerted to important features of the clusters that can sometimes be confusing:

- Because clusters are papers connected by citations, rather than papers categorized as being about a specific topic, understanding the subject of research in a cluster and the trends within it takes some study. The research clusters do not represent traditional research areas.
- Research clusters are not dynamically generated. Due to resource constraints, the ETO Map of Science is clustered every two years. This means that extremely nascent research may not yet have its own cluster but instead be grouped with an existing research cluster.
- Clusters selected do not fully account for research funded by a given institution. Data related to the funders of research publications is inconsistent and infrequently included in bibliometric databases. The sparseness of funder data leads to undercounting a given organization's investment within the research landscape. Also, certain organizations are less likely to make their research

publicly available, which impacts the ability to find highly active and emerging areas of research.[18]

- The merged corpus of papers on which the research clusters are based does not represent research in all languages equally. For example, Chinese-language research is underrepresented as a consequence of restrictions on the use of Chinese-language datasets such as the China National Knowledge Infrastructure.

SMEs should study this background information and the spreadsheets in detail before participating in the workshop described in step 5.

Step 5. Engaging Subject Matter Experts

The final step in our proposed methodology is to gather SMEs for a facilitated discussion after they have had a chance to review the selected clusters. Facilitated conversations, like those based on the Delphi method, are frequently used but, to our knowledge, seldom applied to bibliometric-centered approaches to examining emerging technology.

For this step, it is important to invite SMEs that are well positioned to evaluate the clusters based on their technical or application domain knowledge. For our proof of concept, we invited individuals who had national security expertise, given our original corpus of DOD-affiliated papers. Furthermore, because the DOD has broad technology interests, we selected clusters related to artificial intelligence using CSET's AI classifier and also sought out SMEs with knowledge of AI research relevant to the DOD.[19] Accordingly, our workshop included six individuals who had experience and knowledge of the DOD and its applications of AI.

We shared the clusters and associated metadata (Table 2) with workshop participants one week in advance of the meeting and asked them to consider questions such as if they felt the research clusters were relevant to military interests, if there were any clusters that were surprising, or if there appeared to be research clusters missing. The spreadsheet also contained brief descriptions of the metadata columns and a link to further background information on the ETO Map of Science. Our workshop discussion followed a semi-structured group interview approach that echoed the questions sent in advance (the full set of questions can be found in Appendix D).

Proof-of-Concept Discussion Results

To understand what our method might produce, we include here a synopsis of our proof-of-concept discussion and conclude with general observations.

The participants in our proof of concept discussed all of the clusters we shared but became most interested in one cluster they found surprising and a collection of several clusters they saw as related. The surprising cluster, 1994, has many articles related to image forgery and forgery detection, and the SMEs were largely unfamiliar with the research within the cluster. They found the cluster's high growth and exports interesting and surmised that the research could be relevant to national security concerns around adversarial attacks, clandestine operations, and disinformation. Closer examination of this cluster could support evolving research in the creation and detection of forged images, as well as ways in which forgeries might be used.

The SMEs also noted that a number of clusters that surfaced in our analysis were related to robots and robot navigation (including cluster numbers 5167, 2658, 43601, 416, 7185, 2812, 6265, and 45546 shown in Table 2). The combination of these research advances could substantially affect the DOD's interests in robotics and robotic navigation, especially for drones on land, at sea, and underwater, which are related to the DOD's recent focus on drones via the Replicator program.[20]

SME Workshop Participant Observations and Takeaways

Over the course of the discussion, the participants and authors observed several advantages to our approach of leveraging data to inform SME discussions.

- **Our method prompted participants to consider research and emerging technologies with which they were not familiar**, adding to—rather than duplicating—what the SMEs already knew about developments in AI. Participants shared that our data-informed approach was unique in their experience and could complement established approaches (such as networking or conference attendance). One participant also felt this could be especially useful for organizations with limited time or money for extensive tech scanning efforts. However, the participants with the most extensive personal networks, which they had established specifically for the purposes of horizon scanning, had less use for the data-informed approach.

- **The data prompted a robust conversation about what technologies seemed to be missing from the list of clusters.** In evaluating the list, the participants discussed the research areas they would have expected to see with high growth or exporting knowledge based on DOD priorities. Participants discussed whether the absence of certain research was concerning, especially large language models.

- **Finally, the data prompted the participants to consider how the clusters could be interacting as a set.** This caused an exploration of how emerging techniques or discoveries might influence one another or lead to application. In our particular workshop, for example, the participants connected several clusters to important gains in ground robotics and terrestrial navigation.

While several of the participants remarked that these benefits were helpful and the discussion was worthwhile, the workshop also made clear certain limitations of our approach:

- **The discussion prompted by the clusters was still speculative.** While the data and discussion surfaced relevant technologies that a decision-maker may want to pursue, their inclusion or exclusion from the cluster list is not definitive and is shaped by the strategy for selecting clusters.

- **Some clusters were so foundational and broadly applicable that the SMEs did not derive any insights from their inclusion** (for example, cluster 9301, related to deep neural networks). Clusters that seemed more focused on specific applications, such as cluster 2658 (related to soft robotics) or cluster 1994 (related to image forgery detection) were viewed as more valuable, especially for SMEs not already familiar with these application areas. Clusters containing foundational research may have been more present in our selection of research clusters because the DOD tends to publish basic research. Also, our results may have skewed toward more foundational research because we chose clusters with at least 50 percent of the papers classified as "AI." Clusters with a smaller proportion of papers classified as AI may be more relevant to applications.[21]

- **Research papers (and the accumulation of new research papers into clusters) lag research discoveries.** As a consequence, some of the latest innovations with AI (e.g., large language models or multimodal AI) were not represented in the selected data. However, while our selection criteria omitted this research, the SMEs noted the absence in the discussion.

The participants also noted that the phrases extracted to describe each cluster (see Table 2) were useful only as shorthand. The participants said they had to read the titles and abstracts of a subset of the cluster's papers, especially the highly cited papers, to better understand the subject areas covered by the cluster.

Conclusion and Future Applications of This Approach

Beyond the specific results of our proof-of-concept workshop, the data we analyzed and presented to the SMEs facilitated a novel discussion because it illuminated new areas of research for the SMEs to consider and brought recent research activity to their attention. Overall, participants characterized our method as a "bottom-up" approach that facilitated a robust conversation based on data.

"This is a well-rounded approach." -Workshop participant

The metrics we developed and used to guide our identification of clusters of interest are useful beyond our proof of concept. Future researchers may use our approach to investigate other application areas, for example by using collections of papers from medical institutions to investigate emerging applications of AI in health care. Alternatively, a project could use a collection of papers from a private company or foundation to identify related or especially active areas of interest to a company. Identifying clusters using a different core set of papers may yield results more closely tied to applications than the results of our experiment with the DOD.[22] Indeed, it occurred to us through the course of this project that we may have chosen the hardest use case for our experiment, as the military tends to publish mostly basic research and not applied research.

Beyond starting with a different original set of papers, the metrics and tools developed for our experiment can be manipulated in a myriad of ways. A subset of the metrics we created for our investigation are included in Table 3. For example, future investigations may emphasize clusters with a higher concentration of papers from the original set of papers, or select only clusters where between 25 percent and 75 percent of the papers are classified as AI by CSET's classifier, or use a different classifier altogether. Researchers may wish to explore the interrelationships between different metrics as well.

Table 3. Cluster Metadata Available for Future Explorations (* indicates those metrics used in our proof of concept)

● Average age of papers in cluster*
● Average age of papers, categorization (old, adult, young)
● Size of cluster categorization (small, medium, large)
● Percent papers classified as AI related*
● Percent papers classified as NLP related*
● Percent papers classified as robotics related*
● Percent papers classified as computer vision related*
● Extreme growth predicted (designation based on extreme growth calculation)[23]
● Growth percentile. Out of all clusters, the number of papers in this cluster published in the last five years as a percentile based on this cluster's age and size (a higher percentile means that the cluster has more papers from the last five years compared to other clusters).*
● Percent related to original corpus (percent of papers in the cluster that are from the original corpus)
● Original corpus paper core/highly cited (if a paper from the corpus is considered a core or highly cited paper in the cluster)*
● Original corpus paper is cited by (exports to) a different cluster*
● Identified key phrases (a.k.a. CSET extracted phrases)*
● Export percentile (number of times papers in this cluster are cited by papers in different clusters, as a percentile based on this cluster's age and size)*
● Export diversity percentile (number of other clusters that cite papers in this cluster as a percentile based on this cluster's age and size of cluster)*

Source: Authors' analysis.

Finally, whereas our approach identified individual clusters based on a set of papers and metadata thresholds, it may be equally useful to take the papers funded from a research investor (e.g., the NIH or a nonprofit research funder) and evaluate the clusters as if they were part of an investment portfolio, with some transitioning to application more rapidly than others. Alternatively, instead of using metadata to evaluate clusters, large language models could be used to summarize selected research clusters, and those summaries might better support an SME's evaluation of the potential application of the cluster.

As we said at the beginning, our methodology will not reliably predict the future applications of new technologies, but neither will purely expert opinion or AI prove to be a crystal ball. The best we may be able to hope for is to support human judgment about emerging technologies and their applications by illuminating data that either supports or challenges expert knowledge and intuition. Our method described herein is simply one approach intended to help those trying to stay afloat in a tsunami of new research and development.

Authors

Emelia S. Probasco is a senior fellow at CSET.

Christian Schoeberl is a data research analyst at CSET.

Acknowledgments

The authors would like to thank Alan Brown, Drew Calcagno, Deji Coker, Igor Mikolic-Torreira, Jane Pinelis, and Stu Rogers for sharing their time and expertise. For feedback and assistance on the paper, we would like to thank Catherine Aiken, Zachary Arnold Kevin Boyack, Alan Brown, Kathleen Curlee, James Dunham, Shelton Fitch, Jessica Ji, Autumn Toney, Igor Mikolic-Torreira, and Dewey Murdick.

Document Identifier: doi: 10.51593/20240015

Appendix A: Using the Characteristics of Papers Within Clusters to Identify Relevant Clusters

Concentration Metric Limitations

The most straightforward approach to surfacing relevant clusters for a given audience is to find those clusters with a high concentration of relevant papers. The presence of relevant publications can signal that the cluster as a whole is relevant; however, this approach has several limitations. For example, concentration is calculated from a set of publications that may not include all the relevant papers. While we gathered publications from the DOD's research website (the Defense Technical Information Center) and by a search for papers authored by individuals at DOD-affiliated institutions, the collection of papers is almost certainly smaller than the total number of publications relevant to the DOD.[24] For example, our corpus includes only those papers authored by DOD organizations and not those papers that might have been funded by the DOD but authored by a researcher at a university.[25] Last, the DOD may be researching areas in which there are already numerous papers from industry or academia. As a result of the strong global interest in AI, the overall concentration of one organization's papers will be lower despite constituting a meaningful presence for our research goals. Therefore while the concentration metric may be used to identify clusters, it also has important limitations.

The other reason to avoid relying on concentration is that in our proof of concept we were trying to discover emerging research. When we examined the clusters with the highest concentration of U.S. DOD-authored papers, we found them to be largely predictable and well-known areas of research for the DOD. To demonstrate, these clusters are listed in Table 4.

Table 4. Clusters with More Than 50 Percent of Papers Classified as AI, Sorted by Concentration of DOD-Affiliated Papers

Cluster	Extracted Key Phrases
7068	hyperspectral anomaly detection, anomaly detection method, target detection, detection method based, based hyperspectral anomaly
41876	atmospheric turbulence, turbulence mitigation, image restoration, image quality, proposed method
74405	hyperspectral data, hyperspectral image classification, dimensionality reduction, SOM and DBN, MLP SOM
66451	humanoid robot, mobile robot, UGV, autonomous unmanned system, unmanned ground vehicle
48313	target recognition, infrared images, object detection, infrared target detection, convolutional neural network
25612	unmanned aerial vehicles, cooperative search algorithm, target search, UAV cooperative search, UAV search planning
81063	unmanned aerial vehicles, CARLA, individual factors, teleoperation performance, obstacle avoidance
34365	graph Laplacian, manifold learning, data points, Laplacian semi-supervised learning, learning algorithms

70963	data fusion, semantic information fusion, semantic data, search and rescue, probabilistic semantic data
78366	land cover classification, deep learning, remote sensing images, cover classification performance, land cover

Source: CSET research clusters.

Core or Highly Cited Papers

Core publications are individual papers that are most highly connected to the other papers within the cluster through shared citations. These core publications can represent key research techniques, questions, or applications that others within the cluster heavily draw on or relate to. Additionally, we can see which of the constituent publications within a research cluster are in the top 10 for most citations overall, within or outside of the cluster (in other words, are highly cited). Clusters that have papers from our original list that are core or highly cited could be especially relevant for SME review because this may indicate that the cluster is closely related to the interests of our chosen organization.

Corpus Paper Exporting

Concentration and core or highly cited approaches to finding clusters focus on the internal composition of a given cluster. We can also leverage citations from a single paper to find those papers that are connecting to papers in different research clusters (in other words, the papers are linked, but not so closely linked via citation that they are in the same cluster). As stated previously, the citation relationship between publications is bidirectional. Therefore, a publication can "import" information from other clusters as well as "export" its information to other clusters. For our collection of papers, we analyzed which papers were exporting to other research clusters.

This export relationship to other research clusters is not definitive but could be important, as it may signal potential translational impact and knowledge sharing, and we could aim to find the clusters with the DOD-affiliated papers that export to other DOD or non-DOD clusters at a relatively high rate.[26]

Appendix B: Establishing Percentiles

In order to accurately compare the recent activity of research clusters, we classify each research cluster according to both the years since the average publication date (age) of its papers and the number of papers in the cluster (size). These classifications stem from our expectation that the cluster-level metadata is comparable only to clusters with similar characteristics. For example, research clusters that are in more established research areas may have more publications, as they have had more time for research to accumulate.

Without benchmarking the clusters, smaller and more recent research clusters would be difficult to surface through these size and age metrics. To address this issue, we categorize each cluster as either young (0–5), adult (6–11), or old (12+) based on the average age of its papers. Each cluster also receives an assignment of small (0–199), medium (200–999), or large (1,000+) based on the numbers of papers within the cluster. We use these age and size requirements to establish our percentile calculations.

Appendix C: CSET's AI Classifier

Using the AI Classifier

Machine learning gives us the ability to rapidly assess an enormous corpus of research papers to determine which papers may be relevant to AI. To classify individual research publications, we deploy a set of subject-specific machine learning models trained on data from arXiv, a repository of preprint research publications. These subjects include artificial intelligence, computer vision, natural language processing, and robotics. Authors self-label the publications they submit to arXiv, and these labels are reviewed by SMEs serving as editors. By treating these labels as ground truth regarding the relevance of publications to specific subjects, we can learn to classify each publication based on its title and abstract. Of note, however: the predictions from these subject-specific models are only available for publications with an English title and abstract.[27]

Once the research clusters are formed, we can observe the cluster-level proportion of publications relevant to a given subject. This cluster-level proportion can signify key attributes of the given cluster. For example, a cluster with a high composition of AI-relevant publications may focus on emerging algorithms, distributed computing techniques, or other areas at the core of AI. A cluster with a smaller composition of AI-relevant publications may be an application area of an AI technique in a new field of research, such as object detection for medical diagnostics. From previous research, we consider a research cluster with more than half of its constituent publications flagged as AI-relevant to be an AI-relevant research cluster for current purposes.[28]

Appendix D: Discussion Guide

1. Were any of the clusters surprising to you? If so, in what way? What were your expectations when reviewing the clusters?
2. Do you feel that relevant, emergent research related to AI is missing from this list? What would you say are the most important/relevant research areas that are missing?
3. Do you believe these clusters are representative of the military's research interests? Or top military research interests? If so, why? If not, why not?
4. When reviewing these clusters, what factors led you to classify clusters as relevant to the military?
    a. Cluster subjects? Extracted phrases? Prominent papers? Countries and organizations? Background info, the type of research, where it happens…?
    b. Is there any other information you would have liked to see?
5. How would you describe the subject matter of these clusters, or are the five key concepts sufficiently explanatory for this area of research?
6. Are there any clusters that you are particularly knowledgeable about? Any that you know extremely little about?
7. What applications do you think are relevant to these clusters?
    a. How soon do you see this military application emerging? From whom, how?
8. Which of these clusters is producing research that is closest to transitioning to application? What leads you to this conclusion? Based on your review, how might you categorize this research cluster on a TRL (technology readiness level) scale?
    a. Do you foresee any obstacles that might stand in the way of a transition to application?
9. Do you have a hypothesis or evidence that might explain why these clusters are in the 70th percentile or above in terms of accumulated papers over the past five years?

Endnotes

[1] Beat Habegger, "Horizon Scanning in Government: Concept, Country Experiences, and Models for Switzerland" (Center for Security Studies, 2009), https://www.semanticscholar.org/paper/Horizon-Scanning-in-Government%3A-Concept%2C-Country-Habegger/c01221d8a2f64e6bc36b1ea4b961325f06a3f3c3; Maryse Penny, Tess Hellgren, and Matt Bassford, "Future Technology Landscapes: Insights, Analysis and Implications for Defence" (RAND Corporation, December 5, 2013), https://www.rand.org/pubs/research_reports/RR478.html; Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," *Frontiers in Research Metrics and Analytics* 6 (April 6, 2021), https://doi.org/10.3389/frma.2021.630124; Ramphul Ohlan and Anshu Ohlan, "A Comprehensive Bibliometric Analysis and Visualization of Smart Home Research," *Technological Forecasting and Social Change* 184 (November 1, 2022): 121975, https://doi.org/10.1016/j.techfore.2022.121975.

[2] Committee on Forecasting Future Disruptive Technologies, *Persistent Forecasting of Disruptive Technologies* (Washington, DC: National Academies Press, 2010), https://doi.org/10.17226/12557; Karel Haegeman et al., "Quantitative and Qualitative Approaches in Future-Oriented Technology Analysis (FTA): From Combination to Integration?," *Technological Forecasting and Social Change* 80 (March 1, 2013): 386–397, https://doi.org/10.1016/j.techfore.2012.10.002.

[3] Norman C. Dalkey, "Delphi" (RAND Corporation, 1967). See also Norman C. Dalkey, Bernice B. Brown, and S. W. Cochran, "The Delphi Method, III: Use of Self-Ratings to Improve Group Estimates" (RAND Corporation, 1969).

[4] Ulrich Schmoch, "Double-Boom Cycles and the Comeback of Science-Push and Market-Pull," *Research Policy* 36, no. 7 (September 1, 2007): 1000–1015, https://doi.org/10.1016/j.respol.2006.11.008; Steven R. Walk, "Quantitative Technology Forecasting Techniques," in *Technological Change*, ed. Aurora Teixeira (London: IntechOpen, 2012), http://www.intechopen.com/books/technological-change/quantitative-technology-forecasting-techniques; Heini M. Järvenpää, Saku J. Mäkinena, and Marko Seppänena, "Patent and Publishing Activity Sequence Over a Technology's Life Cycle," *Technological Forecasting and Social Change* 78, no. 2 (February 2011): 283–293, https://doi.org/10.1016/j.techfore.2010.06.020; Murat Bengisu and Ramzi Nekhili, "Forecasting Emerging Technologies with the Aid of Science and Technology Databases," *Technological Forecasting and Social Change* 73, no. 7 (September 1, 2006): 835–844, https://doi.org/10.1016/j.techfore.2005.09.001.

[5] Carie Mullins, "Retrospective Analysis of Technology Forecasting: In-Scope Extension" (Tauri Group, August 13, 2012), https://doi.org/10.21236/ADA568107; Carie Mullins, "Retrospective Analysis of Long-Term Forecasts" (Bryce, July 20, 2018), https://www.openphilanthropy.org/files/Blog/Mullins_Retrospective_Analysis_Longterm_Forecasts_Final_Report.pdf.

[6] Jing Ma et al., "Identifying Translational Indicators and Technology Opportunities for Nanomedical Research Using Tech Mining: The Case of Gold Nanostructures," *Technological Forecasting and Social Change* 146 (2019): 767–775, https://doi.org/10.1016/j.techfore.2018.08.002.

[7] Emily Sylak-Glassman, Sharon Williams, and Nayanee Gupta, "Current and Potential Use of Technology Forecasting Tools in the Federal Government" (Institute for Defense Analysis, March 2016), https://www.ida.org/-/media/feature/publications/c/cu/current-and-potential-use-of-technology-forecasting-tools-in-the-federal-government/d-5735.ashx.

[8] Any cluster listed in this report is publicly accessible via the ETO Map of Science web page. Our approach to identifying clusters, however, is not currently accessible through the web interface and involves bespoke analysis of proprietary data, as is detailed in the report.

[9] For more information on how we generated our merged corpus of scholarly literature, see "Documentation: Merged Academic Corpus," ETO, https://eto.tech/dataset-docs/mac/. For more information on how we generated the Map of Science, see "Documentation: Map of Science," ETO, https://eto.tech/tool-docs/mos/.

[10] See, for example, Mullins, "Retrospective Analysis of Technology Forecasting," and Mullins, "Retrospective Analysis of Long-Term Forecasts"; Committee on Forecasting Future Disruptive Technologies, *Persistent Forecasting*; Dalkey, Brown, and Cochran, "The Delphi Method, III."

[11] Emelia Probasco and Autumn Toney, "A Quantitative Assessment of Department of Defense S&T Publication Collaborations" (CSET, June 2024), https://cset.georgetown.edu/publication/a-quantitative-assessment-of-department-of-defense-st-publication-collaborations/.

[12] Of note, this feature is not available in the public interface for the ETO Map of Science.

[13] Our decision to conduct a facilitated discussion was somewhat arbitrary and based on our own resource constraints. Other, more extensive methods could be appropriate, such as the Delphi method.

[14] Autumn Toney and Emelia Probasco, "Who Cares About Trust?" (CSET, July 2023), https://doi.org/10.51593/20230014b; Joanna Lewis, Autumn Toney, and Xinglan Shi, "Assessing the Global Research Landscape at the Intersection of Climate and AI" (CSET, November 2023), https://cset.georgetown.edu/publication/assessing-the-global-research-landscape-at-the-intersection-of-climate-and-ai/.

[15] Our method for identifying DOD-affiliated papers is described in Probasco and Toney, "A Quantitative Assessment." We also include here a set of papers identified in the Defense Technology Information Center website.

[16] Caleb Smith, Richard Klavans, and Kevin W. Boyack, "A Bibliometric Solution to the Problem of Translational Science," in *26th International Conference on Science and Technology Indicators*, eds. N.

Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (STI, 2022), https://doi.org/10.5281/zenodo.6946189.

[17] Autumn Toney and Melissa Flagg, "Analyzing the Directionality of Citations in the Map of Science" (CSET, March 2023), https://cset.georgetown.edu/publication/analyzing-the-directionality-of-citations-in-the-map-of-science/.

[18] This is particularly true in our case for DOD research publications, where a majority of extremely relevant research is likely classified or otherwise not released for public consumption. Other more public-facing institutions like the NIH may have an easier time finding these highly active and relevant research areas.

[19] Autumn Toney-Wails, Christian Schoeberl, and James Dunham, "AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications," arXiv preprint arXiv:2403.09097 (March 14, 2024), https://doi.org/10.48550/arXiv.2403.09097.

[20] "Replicator," Defense Innovation Unit, accessed September 13, 2024, https://www.diu.mil/replicator.

[21] For more information see Autumn Toney, "Locating AI Research in the Map of Science" (CSET, July 2021), https://cset.georgetown.edu/publication/locating-ai-research-in-the-map-of-science/, and Probasco and Toney, "A Quantitative Assessment."

[22] For more information on why the original corpus of DOD publications may skew toward basic instead of applied research, see Probasco and Toney, "A Quantitative Assessment."

[23] Based on Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," *Frontiers in Research Metrics and Analytics* 6 (April 6, 2021), https://doi.org/10.3389/frma.2021.630124.

[24] The Defense Technical Information Center maintains a list of papers funded or authored by the DOD, which can be found at https://discover.dtic.mil/.

[25] Missing funding data is high across all available sources. See, for example: N. Smirnova and P. Mayr, "A Comprehensive Analysis of Acknowledgement Texts in Web of Science: A Case Study on Four Scientific Domains," *Scientometrics* 128, (2023): 709–734, https://doi.org/10.1007/s11192-022-04554-9. See also Belén Álvarez-Bornstein and Michela Montesi, "Funding Acknowledgements in Scientific Publications: A Literature Review," *Research Evaluation* 29, no. 4, (October 2020): 469–488, https://doi.org/10.1093/reseval/rvaa038.

[26] We chose to focus on exports in this experiment because of our interest in understanding potential applications of AI. It is reasonable to consider, however, that an organization funding basic research with translation impact might be more interested in imports instead of exports, since those imports might identify foundational research worth supporting.

[27] Toney-Wails, Schoeberl, and Dunham, "AI on AI."

[28] Autumn Toney, "Data Snapshot: Locating AI Research in the Map of Science," CSET (blog), July 14, 2021, https://cset.georgetown.edu/publication/locating-ai-research-in-the-map-of-science/.