MITRE | Center for Technology & National Security

SERIES
Number 8

INTELLIGENCE AFTER NEXT

BREAKING PAST AI’S CONFIRMATION BIAS

by Mike Shea

AI is great at finding more of what you know–but not at finding things you never thought to look for. Let’s change that.

Investigating alternate hypotheses is an important part of both science and intelligence analysis. Confirmation bias pushes us to overweight evidence matching our initial hypothesis and discounting evidence to the contrary. Intelligence analysts know not to cling too tightly to a hypothesis until they acquire enough evidence to support it and find no or little evidence that contradicts it. Current artificial intelligence (AI) systems help us with finding supporting evidence but do little to help us find *contradictory* evidence.

AI excels at helping analysts automate the task of finding results in a vast sea of data based on previously known examples. It can locate heavy equipment in overhead imagery, identify documents of potential interest out of millions of others, and sort photos into like bins. It is more difficult for AI to help us find things we do not know to look for as AI applications often have the same built-in confirmation bias we try to avoid. We need systems that not only help us find supporting evidence but also potential evidence for one or more alternate hypotheses.

The challenge of AI-enabled discovery of unknowns is well recognized in the field, and there is some work to address it such as the DARPA Automated Scientific Knowlede Extraction (ASKE) project. This paper looks at the challenge when AI is applied to intelligence analysis.

There are two potential paths through which using AI applications can help analysts investigate alternate hypotheses more thoroughly. They are:

- “Widening the net” by using an ensembled approach of machine learning algorithms for clustering, classification, and word embedding to identify topics outside an initial investigation.
- Augmenting analytical structured argumentation with information retrieval systems designed to help analysts locate information supporting or contradicting their initial hypotheses.

In addition, future investments are necessary to make the process of identifying alternate hypotheses more valuable to analysts. These include providing better tools to help analysts organize their evidence and, at the same time, identify potential gaps in their arguments; building in features to highlight the analysis of competing hypotheses into search engines; and identifying ways in which AI can assist analysts in filling out a structured argument pathway or searching for evidence to support each node in their argument. AI can play an important role in strengthening the use and effectiveness of alternate hypotheses in the analytic process, leading to stronger and more accurate assessments for the Intelligence Community’s (IC) current and future customers.

Introduction

Many traditional uses of AI and machine learning (ML) focus on statistical methods to find similar data. We often build models from "training" data in order to find similar patterns in larger sets of unknown data. We train algorithms on labeled audio recordings to identify particular words or phrases found in larger sets of unlabeled audio recordings. We use document classification models to find and label documents similar to ones in a labeled training set. Typical use of ML helps us find things similar to things we already know. It is harder to use ML (and AI in general) to find things we do not know.

Consider, for example, an intelligence analyst responsible for looking at new potential weapon systems in a particular country. Part of the analyst's duties is to search for evidence in large amounts of open source reporting, imagery, and intelligence reports. The analyst receives a lead that their country of responsibility is using a chemical agent in a new weapon system and begins to conduct further investigative analysis. Over time, the analyst builds a solid amount of evidence pointing to the use of this chemical agent. An important question at this stage of analysis is whether there is sufficient evidence to prove the hypothesis true. What if the acquisition of the chemical agent was actually for another purpose? Could this evidence be pointing to a different result, or is there available evidence the analyst has not seen, indicating this hypothesis to be false? The analyst looks through the information, does not find any specific evidence of another use, writes up their intelligence report and provides it to customers, either decision-makers or warfighters, who potentially take action on the report. Many times, analysts develop a hypothesis and back it up with evidence but have no clear way to determine what evidence exists for an alternative hypothesis or if evidence exists to disprove the hypothesis.

While the IC understands and conducts analyses of competing hypotheses, the time needed, and process required for such exhaustive analysis is often cumbersome and often skipped for short-term tasks or problems with a smaller focus. Yet, understanding what hypotheses have been considered and explicitly rejected is also key to the analyst's final conclusions.

The Dangers of AI's Biases

Many implementations of artificial intelligence are at risk of building models based on biased training data. AI and ML models are only as good as the data upon which they're trained and most often these data contain flaws that bias the model. Many AI models can be easy to fool, require extraordinary amounts of training data, provide no information on how they work internally, and come nowhere near to matching the cognitive capabilities of a human being. In this article we recommend joining machine learning techniques with human cognition to offset the bias of both groups and build a partnership well beyond the sum of the parts.

AI's Current Role

Given this scenario, machine learning applications help analysts discover information related to their initial hypothesis. Search engines bring back documents based on the analyst's queries. Recommendation engines, a common application of AI, present documents similar to ones the analyst already acquired. Image recognition algorithms identify subjects based on pre-defined and pre-labeled training data.

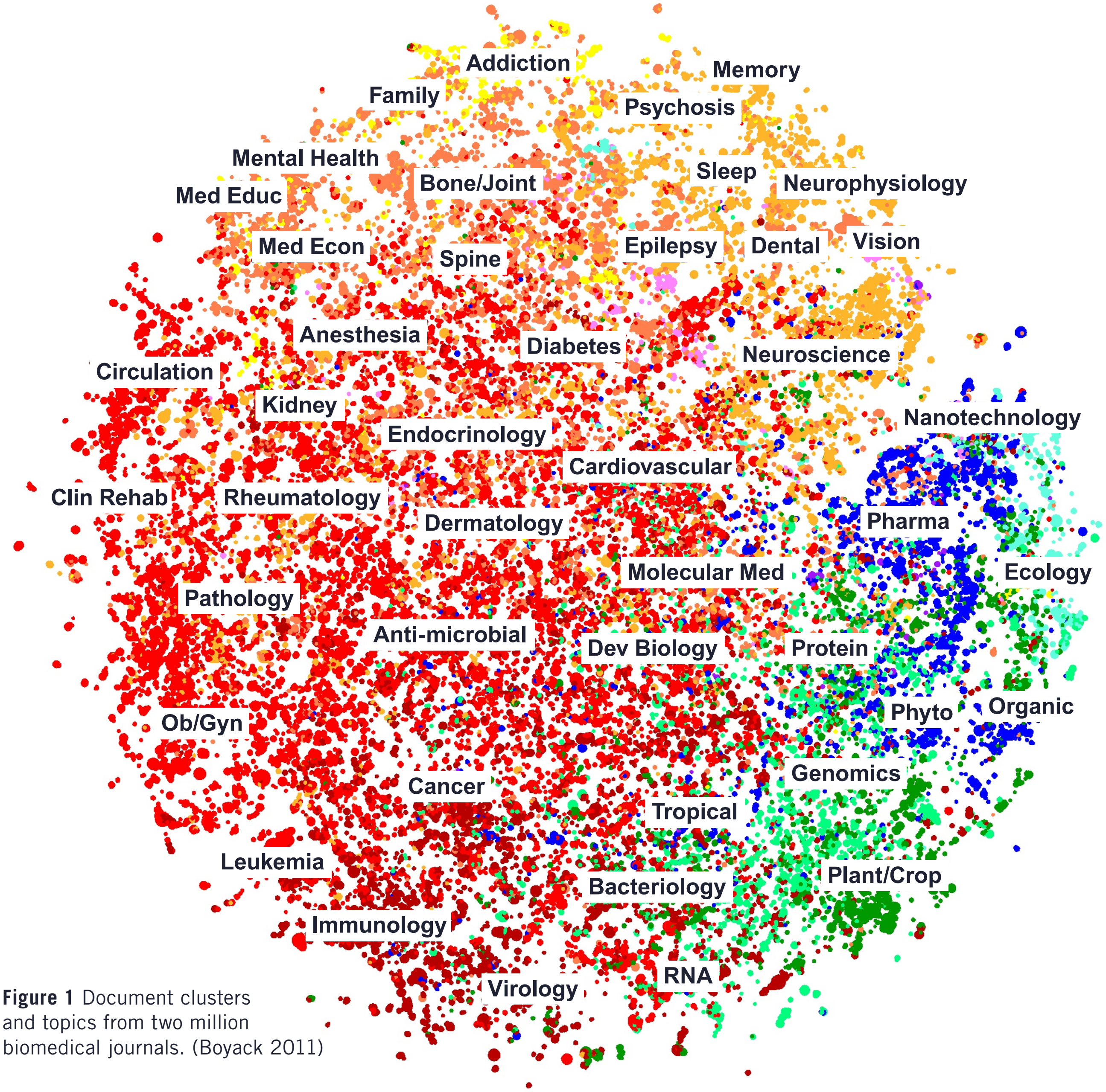

Figure 1 Document clusters and topics from two million biomedical journals. (Boyack 2011)

Most statistical-based machine learning algorithms act upon examples of what we already know. They do not help identify “unknown unknowns.” One way to attempt to identify “unknown unknowns” is by using unsupervised learning algorithms. For example, clustering algorithms can look for differences among a large set of data and bin the data based on these differences without needing to be trained. An analyst could run thousands of documents through a clustering algorithm to identify groups of documents with similar word usage in them. The tools would then help the analyst identify the top words per cluster to see which words pulled it into one bin or another and identify potential new topics from these words. Figure 1 shows an example of document clustering and topic identification for 2 million biomedical journals.

Clustering like this has a few drawbacks. First, it often clusters on words with no actual contextual meaning since no human told it what topics to track. Second, it often requires identifying a fixed number of clusters which can be a shot in the dark when you are not sure how many topics exist in the corpus. Often continual iterative approaches are required to find optimal clusters.

There is another disadvantage for finding "unknown unknowns" using clustering: you must define a boundary around the whole corpus you intend to cluster. An analyst cannot run clustering algorithms against all possible documents because the total size is too large, and the number of topics becomes so large it becomes meaningless. Instead, the analyst must define an outer boundary of documents to investigate. How does the analyst know which documents to include? If they run queries to draw a larger set of data, the results are still bound by the original queries. They are not truly "unknown unknowns."

Widening the Net: Word Embedding for Query Expansion

One way to find potential "unknown unknowns" is to widen the net. Instead of pulling back a large number of documents based on known queries, various tools can help analysts discover new queries based on the documents returned from the initial set of queries. One method to do this is by using a technique called "word embedding."

Word embedding is a form of unsupervised learning in which mathematical word vectors are created by identifying the use and position of words within sentences. Fed by very large sets of sentences, models based on word embeddings act as a domain-specific thesaurus to identify alternate terms or language often used in a similar context in a large corpus.

Using word embedding models, analysts can run queries and, instead of getting back documents, they see a list of other commonly used terms related to those queries. A mixture of positive and negative queries can work together to identify entirely new areas of potential investigation. These new queries can thus widen the net for a larger set of documents, which can then be binned using clustering methods described above.

While many large-scale models are built on general purpose corpora, such as training on billions of Google News articles, word embedding models show great value when trained on a large set of domain-specific documents such as intelligence message traffic. This prevents receiving general-purpose results for a given query and instead the word embedding algorithm "learns" from the domain-specific text and returns more relevant results. Mustard, for example, will be more likely associated with "mustard gas" and "chemical weapons," and less often with "ketchup and mustard" or "picnic," when trained on a large corpus of material based on weapons of mass destruction.

By mixing word-embedding models, supervised document classification models, and unsupervised document clustering along with traditional data mining techniques, analysts and data scientists can work in partnership cyclically to expand the net and focus on possible evidence of an alternate hypothesis. Iterative steps include:

- Using search queries to recover a large set of documents related to their topic
- Clustering these documents to see what new topics might reveal themselves
- Running queries through a large subject-focused word embedding algorithm to discover more potential terms of interest based on the topics
- Building a larger set of documents to run through clustering and so-on

Analysts partnered with data scientists throughout this process can use data mining techniques to filter out noise and help identify potential new lines of analysis.

Building Off of Structured Argumentation

Structured argumentation is an analytic process breaking down and investigating a hypothesis holistically. Structured argumentation drives the analyst's thought process through a series of logical steps, each backed by its own evidence, tracking both positive and negative paths. It intends to help analysts think through all of the facets of the hypothesis, including those they may either skip over or not see to begin with. Such a process usually produces a visual map of the argument, with each piece of the argument backed by its own array of questions and evidence (Tanner 2019).

This process flow is largely manually constructed and can be documented using off-the-shelf office applications. Although the method of structured argumentation is sound, the added burden of the process tends to make it less popular among analysts. The act of working through all facets of the argument may push an analyst to discover previously unknown and alternate hypotheses, however. These alternative hypotheses would likewise be documented in the structured argumentation process.

There is a potential to augment structured argumentation with AI and ML applications. Once an analyst has developed a structured argument workflow, individual queries could search for evidence to support each node in the workflow, including seeking out evidence for nodes in the argument that *disprove* the hypothesis.

Potential future investments in this area could include an investigation of ways to make the process of structured argumentation more facile and valuable to analysts. Better tools, for example, may help analysts better organize their evidence while, at the same time, identifying potential gaps in their arguments. Methods to promote an analysis of competing hypotheses could be built into the very search tools analysts typically use. Another line of investment might focus on helping an analyst fill out the structured argument pathway or search for evidence to support each node in the argument.

Investigate Relevant IARPA Projects

A number of Intelligence Advanced Research Projects Activity (IARPA) projects have touched on using automated systems for the discovery of "unknown unknowns" and may be worth further investigation. These tend to have a heavy focus on crowdsourcing the discovery of future scientific trends. These include the IARPA Foresight and Understanding from Scientific Exposition (FUSE), Forecasting Science & Technology (ForeST), and Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) programs. FUSE, for example, processes a vast array of scientific publications looking for emerging technology areas of potential high interest. ForeST seeks to identify and forecast emerging trends by aggregating crowdsourced expert predictions. The CREATE program continues this idea through crowdsource-structured analytical techniques; seeking "unknown unknowns" through the aggregate predictions of tens of thousands of forecasters; an approach deemed superior to the opinions of a smaller number of experts (Tetlock 2017). These promising research activities should be further investigated to determine their value and relevance to this issue.

Conclusion

Computers are great at helping us automate manually intensive tasks. They look tirelessly over millions of images to detect potential weapons in photographs or identify text in road signs. It's much harder, however, for computers to show us something entirely new and meaningful; something useful that we have not previously considered. Just as it is harder to prove causality than it is to show statistical correlations, it is similarly challenging for computers to think outside of the boundaries we set for them–the boundaries of our own experience. Investigating ensemble mixtures of supervised and unsupervised systems and better human/computer teaming in structured argumentation are two potential paths to expand our reach for alternate hypotheses. There is no silver bullet. No single tool or algorithm will likely be able to detect a meaningful alternate hypothesis on its own. Ultimately, analysts will need to embrace the core concepts of alternative hypothesis analysis and use AI and ML-based tools to help them search for evidence of those alternate hypotheses. Tools are the lever, but the required force still comes from us.

References

Boyack, Kevin W.; David Newman; Russell J. Duhon; Richard Klavans; Michael Patek; Joseph R. Biberstine; Bob Schijvenaars; André Skupin; Nianli Ma; Katy Börner. March 17, 2011. "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches." 2011. https://doi.org/10.1371/journal.pone.0018029

Elliott, Joshua. Downloaded March 17, 2021. "Automated Scientific Knowledge Extraction." Defense Advanced Research Projects Agency. https://www.darpa.mil/program/automating-scientific-knowledge-extraction

Extance, Andy. September 10, 2018. "How AI technology can tame the scientific literature." Nature. https://www.nature.com/articles/d41586-018-06617-5

Flemming, Nic. May 30, 2018. "How artificial intelligence is changing drug discovery." Nature. https://www.nature.com/articles/d41586-018-05267-x

Daugherty, Paul R. and Wilson, H. James. August 2018. "Collaborative Intelligence: Humans and AI Are Joining Forces." Harvard Business Review. https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces

Fung, Isaac; Guerra, Santiago; Kamar, Ece Semiha; Lasecki, Walter S.; Liu, Anthony; Matute, Gabriel. April 2020. "Towards Hybrid Human-AI Workflows for Unknown Unknown Detection." WWW '20: Proceedings of The Web Conference 2020. https://doi.org/10.1145/3366423.3380306

Heuer, Richards J. Jr. 1999. "Psychology of Intelligence Analysis." Central Intelligence Agency. https://www.cia.gov/resources/csi/books-monographs/psychology-of-intelligence-analysis-2/

Stech, Frank. 2004. "ACH: Analysis of Competing Hypotheses–Almost Cause for Hope DRAFT." The MITRE Corporation.

Tanner, Michael. January 7, 2019. "Structured Argumentation Tools Survey." The MITRE Corporation.

Tetlock, Philip. 2017. *Expert Political Judgement*. Princeton University Press.

Authors

Mike Shea is a writer, software engineer, and data scientist. Mike has worked at MITRE since 1997 supporting multiple intelligence community customers on topics including the counterproliferation of weapons of mass destruction, media exploitation, and foundational intelligence.

Intelligence After Next

MITRE strives to stimulate thought, dialogue, and action for national security leaders developing the plans, policy, and programs to guide the nation. This series of original papers is focused on the issues, policies, capabilities, and concerns of the Intelligence Community's analytical workforce as it prepares for the future. Our intent is to share our unique insights and perspectives surrounding a significant national security concern, a persistent or emerging threat, or to detail the integrated solutions and enabling technologies needed to ensure the success of the IC's analytical community in the post-COVID-19 world.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

MITRE | Center for Technology & National Security