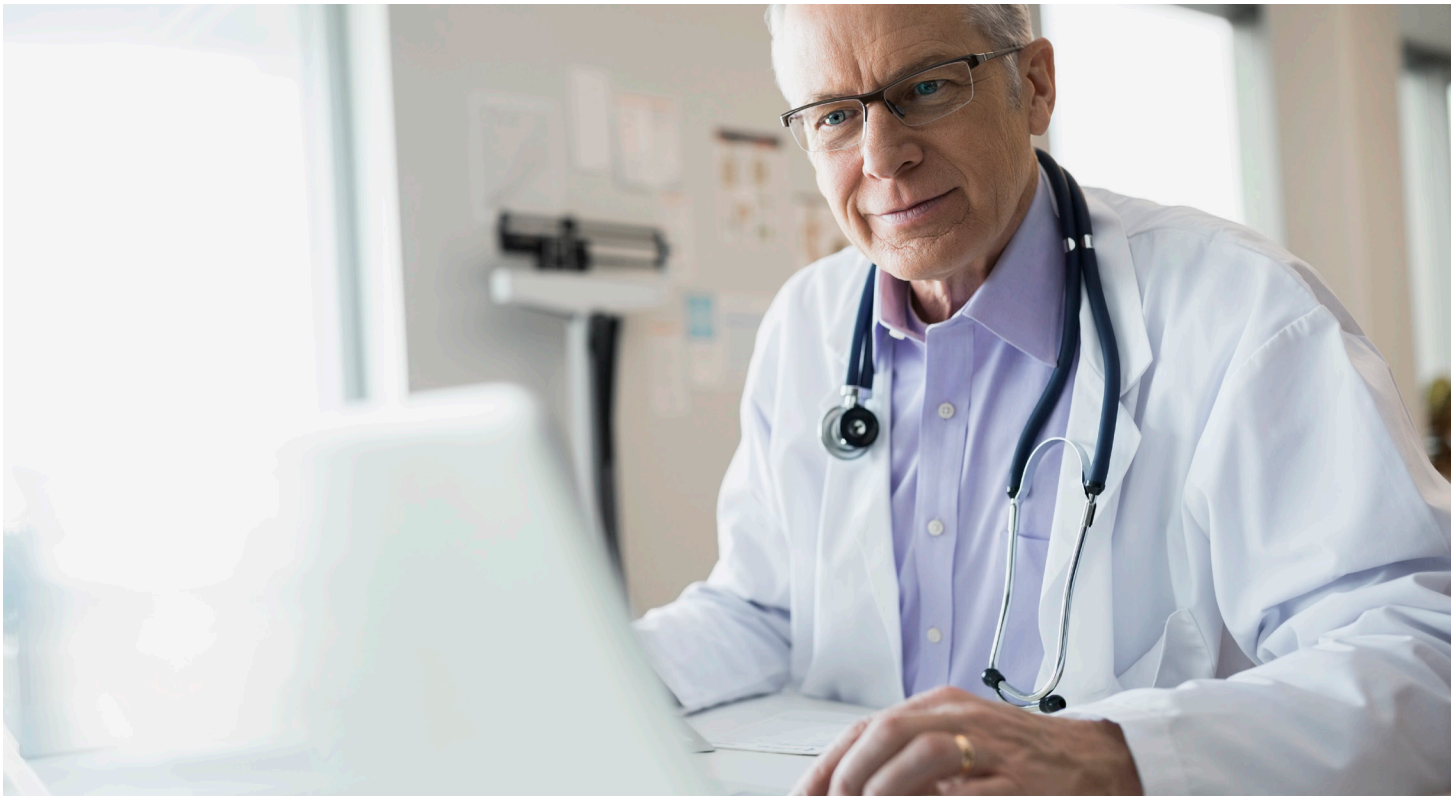
# Closing the Knowledge Gap

## Uncovering Hidden Knowledge in Scientific Literature with Machine Learning



A little knowledge may be a dangerous thing—but what happens when the problem is too *much* knowledge?

The pace of publication for new biomedical research makes it difficult, if not impossible, for researchers and practitioners to keep up with new discoveries—even within a narrowly defined field. These human limitations add to the lag time between published research and changes in clinical practice. They also make it hard to find hidden connections within the knowledge base that could lead to valuable new insights and guide the direction of future research.

This is where artificial intelligence can help. Advanced algorithms utilizing machine learning and natural language processing can tame the tsunami of scientific research and make large knowledge databases useful and usable for both researchers and clinicians.

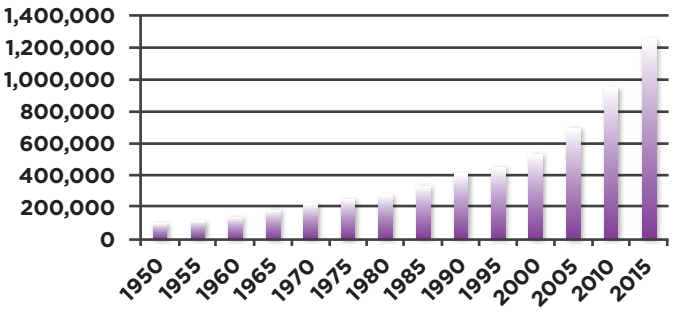
### Understanding the Knowledge Lag

There are currently more than *27 million* citations on PubMed®, which provides access to MEDLINE®, the largest digital repository for biomedical literature in the world. Every day, thousands of new citations are added. According to PubMed search results, 1,255,916 new citations were added in 2016—a 2.5-fold increase from 1990 and a six-fold increase from 1970. The pace of research continues to accelerate. And it's not just biomedical research—according to research from the University of Ottawa, more than 50 million scientific papers were published across all disciplines between 1965 and 2009, with more than 2.5 million new papers published annually at the end of study period.[1] Extrapolating from there, it is likely that there are now more than 60 million scientific papers in existence today!

[1] Jinha, A. Article 50 million: an estimate of the number of scholarly articles in existence. Learned Publishing, 2011; 23:258-63

BATTELLE

**Figure 1: New Citations in PubMed, 1950-2015**




## By the numbers:

- 60+ million scientific papers since 1965
- 2.5+ million published annually
- 27 million citations in PubMed
- 1.2 million new citations annually

All of these publications add to the overall knowledgebase in science and medicine. But the pace of publication has a dark side: data overload. Just skimming through the 13,481 new publications cited in PubMed in 2016 on the topic of diabetes alone would far exceed the number of working hours available for the year.

No wonder much of this research has limited impact on clinical practice. On average, it takes 17 years or more for research findings to translate into practice by healthcare practitioners.[2] This lag time has put tremendous barriers in place for moving to Evidence-Based Practice (EBP). Busy doctors rarely have the time to sift through mountains of published studies in order to identify the most relevant, applicable and trustworthy evidence and update their clinical practice.

This "knowledge lag" also impacts researchers, policy makers and educators. MEDLINE and other corpora of scientific papers contain valuable knowledge that could help policymakers make better decisions, professors update course curricula for medical professionals, and researchers find connections that could inform future research efforts. However, without better tools to mine and extract this knowledge, much of it remains hidden.

## The Rise of the Learning Machines

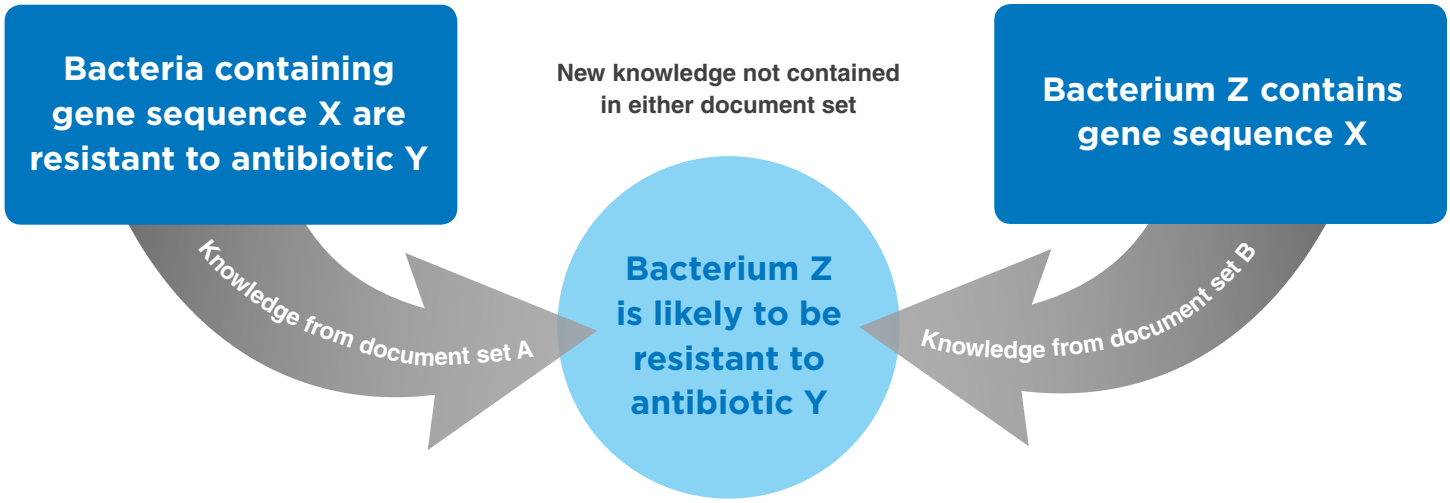
Machine learning is a type of artificial intelligence that uses classification algorithms to allow computers to detect patterns in large data sets. These algorithms give computers the ability to "learn" without being specifically programmed. Natural language processing enables the program to "understand" and extract knowledge from unstructured sources and documents written in natural (human) language, such as scientific papers. These algorithms can automate the process of collecting, filtering, comparing and synthesizing knowledge from large corpora of scientific publications.

Machine learning can vastly reduce the time and effort required to review scientific literature, identify the most relevant citations for a specific research question, and identify hidden connections between different studies or data sets. This allows researchers and practitioners to perform queries to quickly find the information they need out of large corpora of scientific studies.

The algorithms use ontology-based categories and axioms in order to construct a knowledge base that enables meaning extraction. This allows the program to define the properties of each category within its domain and identify relationships between different categories. One way to do this is by using axioms known as "triples" within a formal structure called a Resource Description Framework (RDF). RDF is a commonly used format for representing linked data (e.g., "[the intervention] handwashing [the outcome] reduces person-to-person spread of [the target] staphylococcus").

Adding an additional "inference layer" to the program expands its ability to make connections between knowledge from different sources. This works in a similar fashion to the logic statement "If A=B and B=C then A=C." This allows the program to find hidden connections that are not directly stated in any one document. For example, researchers at Battelle used a program called Battelle Sematrix® to identify strains of bacteria that were likely to be resistant to certain antibiotics. The program first used the literature to identify known gene sequences that encoded specific proteins involved in antibiotic resistance. This knowledge was used to annotate a database containing genomic (DNA) sequences for a large number of microbial organisms. Researchers were then able to infer that if a particular organism contains a gene that encodes a protein involved in making a certain class of antibiotics ineffective, the organism is most likely resistant to the antibiotic—even though that assertion never occurred directly in either set of source documents.

[2] Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011; 104 C(12):510-520




## The Power of Natural Language Query

Sophisticated analytical programs like Battelle Sematrix can help make knowledge contained in large corpora visible and usable. Natural language query allows researchers to ask simple questions using normal language. This allows researchers and practitioners to query the knowledgebase directly, without any programming experience. For example, a query might ask "what is the evidence that handwashing reduces the spread of MRSA in hospitals?" or "does taking this statin impact the effectiveness of insulin for diabetics?" The program will be able to search through the corpora to find the most relevant source documents and represent the knowledge from those documents in a way that is easy for researchers or practitioners to analyze and understand.

This approach significantly reduces the time and human energy required to manually search through large corpora, making the knowledge contained in them more accessible. This can help users in several ways:

- Find relevant research: Researchers, policy makers and clinicians could use natural language queries to quickly identify the most relevant research to answer critical questions for guideline development, coverage and payment policy or clinical decision making. For example, a rural doctor practicing far from the nearest teaching hospital may someday be able to use a sophisticated analytical engine to find the most effective treatment options for a patient with a rare medical condition.
- Enable evidence-based practice: Automated identification and summarization of the most relevant new information could significantly reduce the time it takes for new knowledge to translate into clinical practice. Machine learning would make it easier to evaluate new evidence and make effective decisions regarding changes to treatment protocols.
- Identify promising connections: Inferential logic allows queries to be used to identify potential connections between disparate facts in order to find new insights and draw meaningful conclusions. This could be used to point to potential answers that would not have been considered otherwise.
- Guide research priorities: Sophisticated analytics will enable researchers to identify knowledge gaps as well as promising new avenues for research.

## Case Study: Reducing Research Time for CMS Measure Development

Over the past several years, the Centers for Medicare & Medicaid Services (CMS) has been moving towards a "value-based purchasing" model in which CMS pays providers, in part, based on the quality (value) of care provided rather than just the volume of care. In order to make this vision a reality, CMS needs to have quality measures that can be used to determine the value of care provided by institutions and clinicians. To develop each measure, CMS must find and evaluate the relevant research that has been published on each topic. This process can take hundreds of human hours per measure—and CMS needed to develop hundreds of measures.

Battelle and CMS partnered to find a better way. CMS is using Battelle Sematrix™ to reduce the time it takes to locate and evaluate the most relevant and usable articles for measure development from weeks or months to moments. Sematrix tagged and analyzed millions of documents from PubMed and other relevant libraries. Using a training set of 66 CMS clinical quality measures, researchers developed a system that can quickly find the most useful and relevant articles related to each measure and extract the information required for measure development into a usable form for CMS.

While human subject matter experts are still needed to read and evaluate the most relevant articles and the knowledge extracted by Sematrix, the system cuts out the massive number of hours spent reviewing abstracts and articles to find pertinent research. Measure developers can now focus on the top ten or twenty articles identified by Sematrix as most useful for the problem at hand. Ultimately, researchers use Sematrix to reduce the time it takes to locate and evaluate the most relevant and usable articles for measure development from weeks or months to moments.



**800.201.2011** | **solutions@battelle.org** | **www.battelle.org**



ID 604 08/17

BATTELLE

It can be done