AI Under the Hood:
What Natural Language Processing (NLP) Means for Pharma
MODERNA, COHERUS MEDICAL AFFAIRS USE CASE

*Matt Lewis[a], Richard Graves[b], Tim Mikhelashvili[c], Scott Navarro[a], Matt Booth[d], Ricky Li[a], Nga Tong[e], Paul Walden[f], Robert B Geller[f], **Sandra Carrico[b], Walter Bender[b]***

[a]MEDiSTRAVA, Boston, MA, USA; [b]Sorcero Inc., Washington, DC, USA; [c]Amedea Pharma, Inc., Westbury, NY, USA; [d]MEDiSTRAVA, London, UK; [e]Moderna Inc., Boston, MA, USA; [f]Coherus Biosciences, Redwood City, CA, USA

OBJECTIVE

- We aimed to demonstrate the benefits of transforming literature monitoring from a manual report-based process into an intuitive, semi-automated, AI-driven process.

BACKGROUND

- Intelligent Literature Monitoring (ILM) is a novel approach of augmenting and running continuous systematic literature searches around defined areas of therapeutic domains.
- ILM leverages Artificial Intelligence (AI) and its subset discipline Natural Language Processing (NLP).
- This enables the extraction of valuable insights from an expansive number of medical publications, as they are being published, into a centralized location.

In 2018, the Google AI Language Division published BERT, "Bidirectional Encoder Representations from Transformers," a pre-trained algorithm capable of accurately gauging the context of each word in a sentence. BERT uses techniques including the ability to look forward and backward across passages to create an optimal representation of language. Once that representation is developed algorithmically, it can be leveraged for applications of Language Intelligence. BERT is Open Source and allows for permanent, continuous, crowd-sourcing, and machine learning improvement. The Sorcero LI platform leverages language models including BioBERT, while adding several proprietary algorithms to generate life sciences-specific solutions.

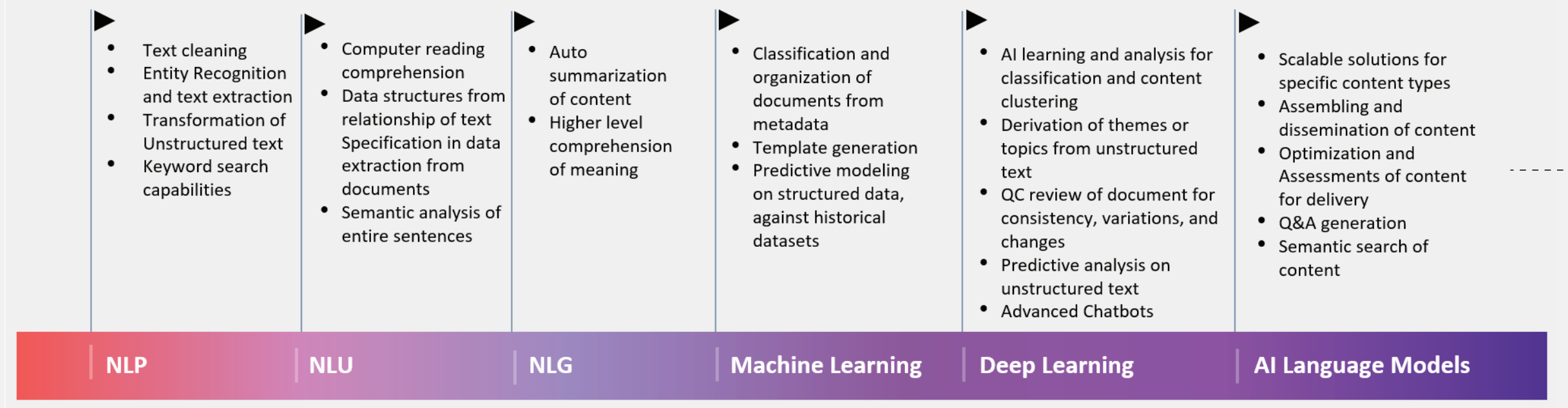
METHODS

- We assessed two different approaches to continuous literature monitoring: (1) manual search and output, (2) semi-automated search, with AI integration and digital output, applied across search scenarios producing high (>250) and low (<100) volumes of results
- ILM was performed using the Sorcero LI Platform and the BioBERT language model against identical corpora
- We compared each approach for man hours, sensitivity, specificity, versatility/utility of outputs and depth of insights
- We also qualitatively assessed the digital output from a user experience perspective

RESULTS

The ILM approach with integrated AI was the most favorable option for search strings with high volumes of search results (e.g. "oncology" or "COVID") (Figures 3 and 4).

- Compared with manually searching and reviewing the literature, this method resulted in a time reduction of 88-92%, along with 99.8% sensitivity and 95% specificity (Table 1).
- Sorcero's LI approach and continuous learning delivered double-digit absolute performance improvements across all study types, exceeding the 95% NPV threshold commonly accepted as that required for a regulatory-grade literature review solution (Table 2).
- For low-volume search results, (e.g. rare diseases) a digital tool improved accessibility, distribution, communication, archiving, and sorting/filtering.



Text cleaning
Entity Recognition and text extraction
Transformation of Unstructured text
Keyword search capabilities

Computer reading comprehension
Data structures from relationship of text
Specification in data extraction from documents
Semantic analysis of entire sentences

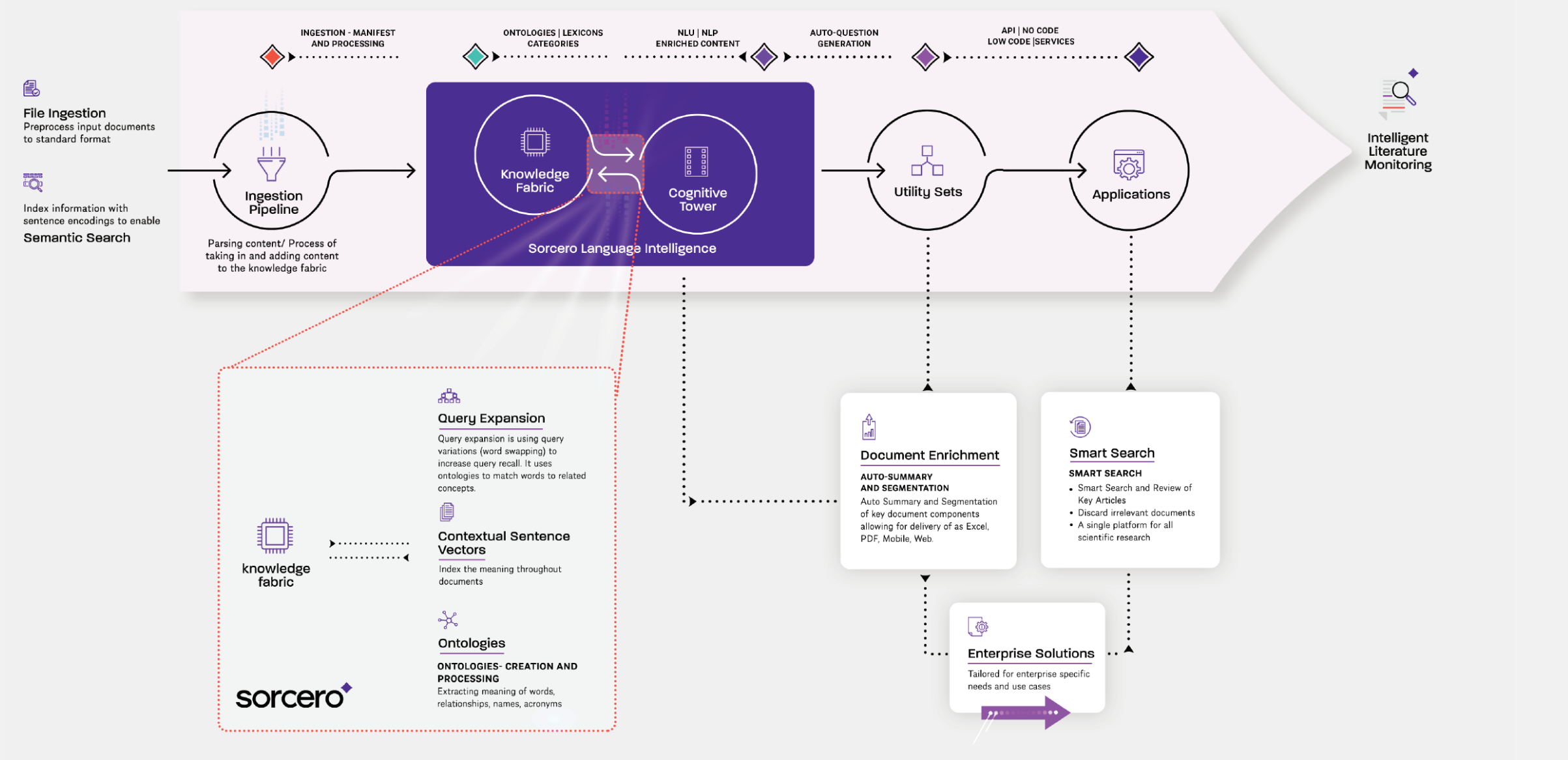
Auto summarization of content
Higher level comprehension of meaning

Classification and organization of documents from metadata
Template generation
Predictive modeling on structured data, against historical datasets

AI learning and analysis for classification and content clustering
Derivation of themes or topics from unstructured text
QC review of document for consistency, variations, and changes
Predictive analysis on unstructured text
Advanced Chatbots

Scalable solutions for specific content types
Assembling and dissemination of content
Optimization and Assessments of content for delivery
Q&A generation
Semantic search of content

Figure 1. Language Intelligent Capabilities spectrum



Preprocess input documents to standard format

Index information with sentence encodings to enable Semantic Search

Parsing content/ Process of taking in and adding content to the knowledge fabric

Query expansion is using query variations (word swapping) to increase query recall. It uses ontologies to match words to related concepts.

Index the meaning throughout documents

ONTOLOGIES- CREATION AND PROCESSING
Extracting meaning of words, relationships, names, acronyms

AUTO-SUMMARY AND SEGMENTATION
Auto Summary and Segmentation of key document components allowing for delivery of as Excel, PDF, Mobile, Web.

SMART SEARCH
- Smart Search and Review of Key Articles
- Discard irrelevant documents
- A single platform for all scientific research

Tailored for enterprise specific needs and use cases

Figure 2. The ILM Approach with Integrated AI

Results

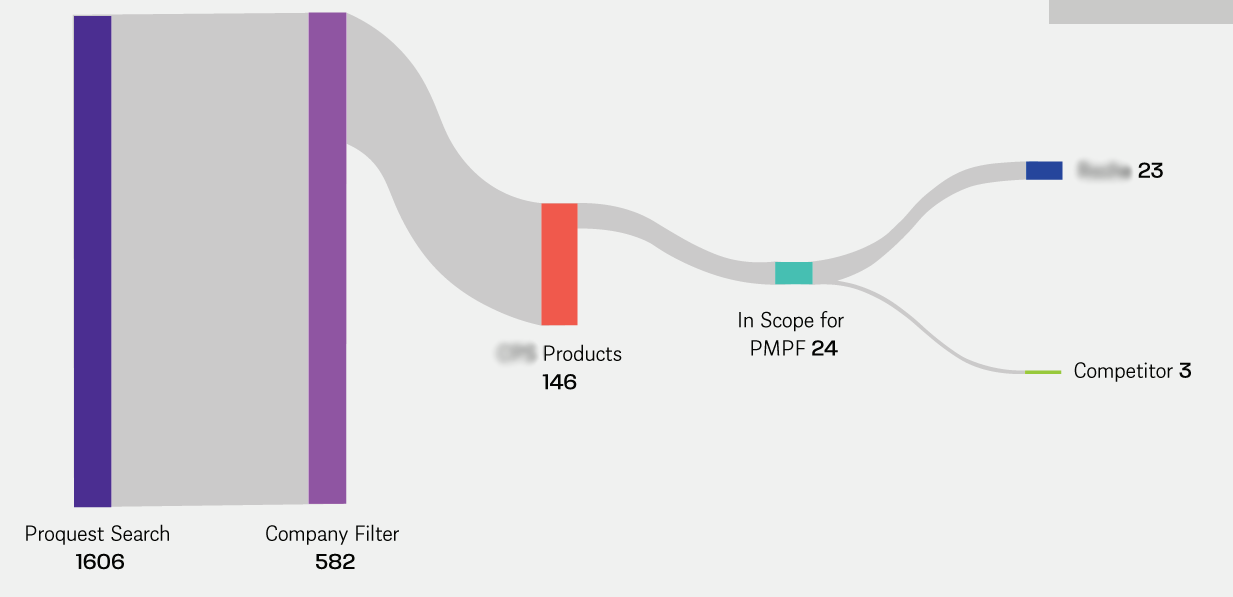


Figure 3. Smart filtering through classification

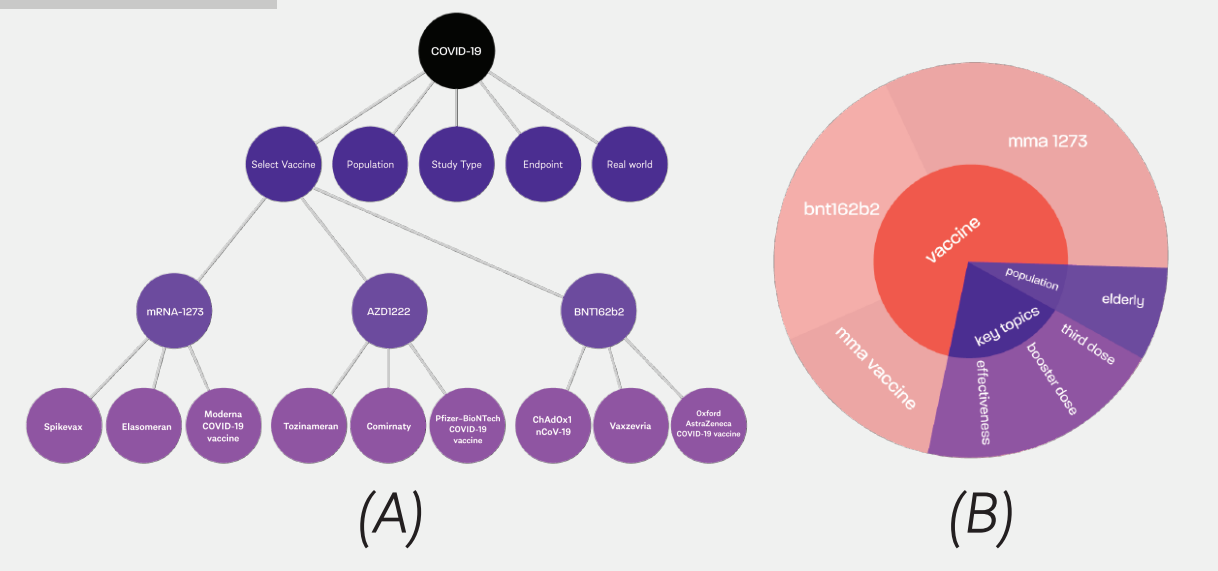
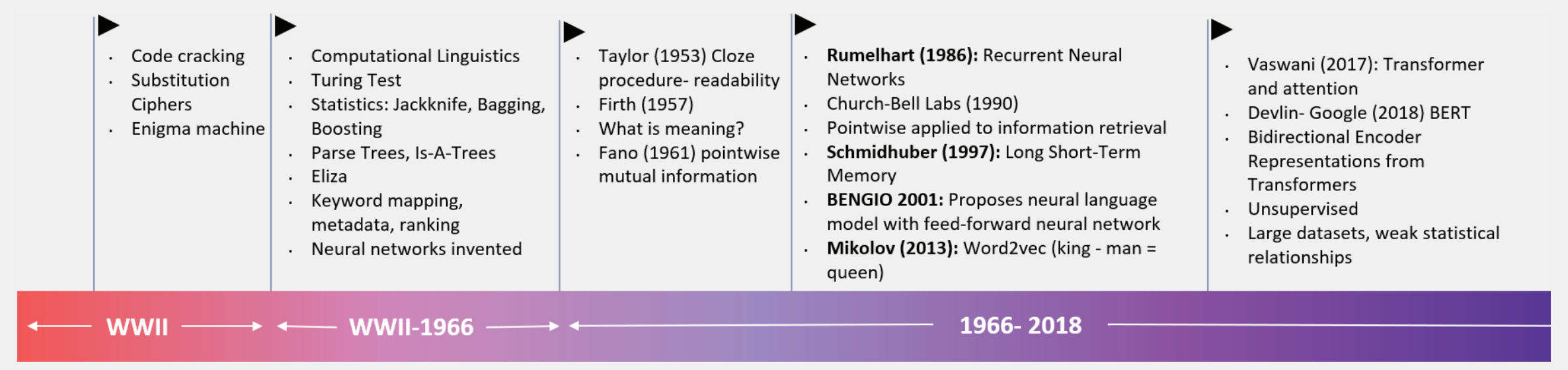


Figure 4. Insight generation: auto-summary outputs from high volume ILM. A) Vaccine ontology tree; B) Keyword summary from mRNA vaccines dataset



Code cracking
Substitution Ciphers
Enigma machine

Computational Linguistics
Turing Test
Statistics: Jackknife, Bagging, Boosting
Parse Trees, Is-A-Trees
Eliza
Keyword mapping, metadata, ranking
Neural networks invented

Taylor (1953) Cloze procedure- readability
Firth (1957)
What is meaning?
Fano (1961) pointwise mutual information

Rumelhart (1986): Recurrent Neural Networks
Church-Bell Labs (1990)
Pointwise applied to information retrieval
Schmidhuber (1997): Long Short-Term Memory
BENGIO 2001: Proposes neural language model with feed-forward neural network
Mikolov (2013): Word2vec (king - man = queen)

Vaswani (2017): Transformer and attention
Devlin- Google (2018) BERT
Bidirectional Encoder Representations from Transformers
Unsupervised
Large datasets, weak statistical relationships

Figure 5. The Evolution of AI and Language Intelligence

	Manual (Control Process)	ILM Process	Results
Total Articles	1606	1606	
# of Batches	44	44	
% Reviewed	100% manual	100% AI + Manual	
Company Criteria Relevancy Filter	581	582	99.83% sensitivity
Product Relevancy Scored	146	146	
Relevancy Inclusion	23 Human Determined	23 AI Determined	>95% specificity
M.D./PhD. Time to Review	715 hours (Avg. 27 mins/articles)	55 hours Review + 25 hours of QA	88% reduction in manual time
MD/PhD Cost to Review	$286,000 (@$400/hr)	$32,000 (@$400/hr)	$254,000 reduction in cost of time to review

Table 1. Real-world results of the manual (control) process vs. ILM with integrated AI, performed for leading pharmaceutical enterprise medical affairs and publications teams

	Analyte is subject	Performance study	Human study	Intended use
Bio BERT	85.00%	75.00%	84.60%	79.70%
Sorcero LI Platform	96.50%	88.50%	96.90%	91.10%
Differential	+11.50%	+13.50%	+12.30	+11.40%

Table 2. Comparative negative predictive value (NPV) from two AI literature monitoring platforms. Machine Learning and mathematical domain representation in the LI platform automatically tune general domain language models including Bio BERT for vastly higher performance in specific biomedical domains.

CONCLUSIONS

The ILM approach to systematic literature searching:

1. Accelerates the current time-bound and episodic manual reviewprocess
2. Increases efficiency and productivity
3. Provides a deeper understanding of current complex, scientific content
4. Operates in near real-time
5. Allows automatic deployment of insights into a centralized and customized database

Digital platforms can compile and store results in a more versatile way, increasing functionality and creating efficiencies.

DISCLOSURE

"Matt Lewis[a]:Nothing to Disclose" , Richard Graves[b]:Nothing to Disclose", Walter Bender[b]:Nothing to Disclose", Sandra Carrico[b]:Nothing to Disclose"' Tim Mikhelashvili[c]:Nothing to Disclose", Scott Navarro[a]:Nothing to Disclose", Matt Booth[d]:Nothing to Disclose", Ricky Li[a]:Nothing to Disclose", Nga Tong[e]:Nothing to Disclose", Paul Walden[f]:Nothing to Disclose", Robert B Geller[f]:Nothing to Disclose"

Bibliography

Lee, J. et al. Bioinformatics 2020;36:1234–1240
Vaswani, A et al. Proceedings of the Advances in Neural Information Processing Systems. 2017; 5998–6008
Sousa, D., Couto, F.M. Adv Inf Retr 2020;12036:367
Wei, Q. et al. Proceedings of the AMIA Annual Symposium. 2019;2019:1236
Zhu X. et al. Applied Sciences 2021;11:7318