SciBite
an ELSEVIER company

Use case

Transform common business and scientific processes with a novel combination of semantic analytics and machine learning

SciBite
an ELSEVIER company

Transform common business and scientific processes with a novel combination of semantic analytics and machine learning

Pharmaceutical companies have a wealth of data at their disposal. However, the volume of data combined with limited availability of experts and data scientists constrains their ability to use it effectively. Artificial Intelligence (AI) has been touted as a way to revolutionise the entire Pharmaceutical value chain, to the extent that it will deliver a 'cure for cancer'. Despite such promises, tangible evidence of how AI is actually helping research remains elusive. Coupled with some high profile failures[1], there is growing scepticism of what AI can realistically achieve.

One of the more promising applications of AI is Machine Learning: the development of a computational model followed by 'training' using sample datasets resulting in decisions or predictions and the inclusion of a feedback loop to refine the approach based on whether a given decision is right or wrong. However, industry optimism is still tempered with caution – Gartner's Hype Cycle positions Machine Learning at the 'peak of inflated expectations', about to enter the 'Trough of Disillusionment'[2].

Looking past the hyperbole and future promises of AI, here we provide a range of real-world examples that illustrate how SciBite is pioneering the use of Machine Learning and Semantic Analytics to transform common scientific and business processes and delivering real business value.

The SciBite platform: enabling effective machine learning

SciBite's platform delivers a unique combination of semantic analytics and Machine Learning which enables the knowledge within unstructured data sources to be unlocked and used to address a broad range of complex problems.

Despite significant technology advances, the axiom 'garbage in, garbage out' cannot be ignored – quality data is fundamental to ensuring Machine Learning gives accurate and true outputs. SciBite's manually curated ontologies comprise tens of millions of synonyms and can be augmented with internal vocabularies, such as compound IDs and study codes. These ontologies provide the foundation for automated pre-processing and semantic enrichment, resulting in the high quality, contextualised data necessary for Machine Learning to be effective.

[1] For example, see https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/#6f8b23673774

[2] https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

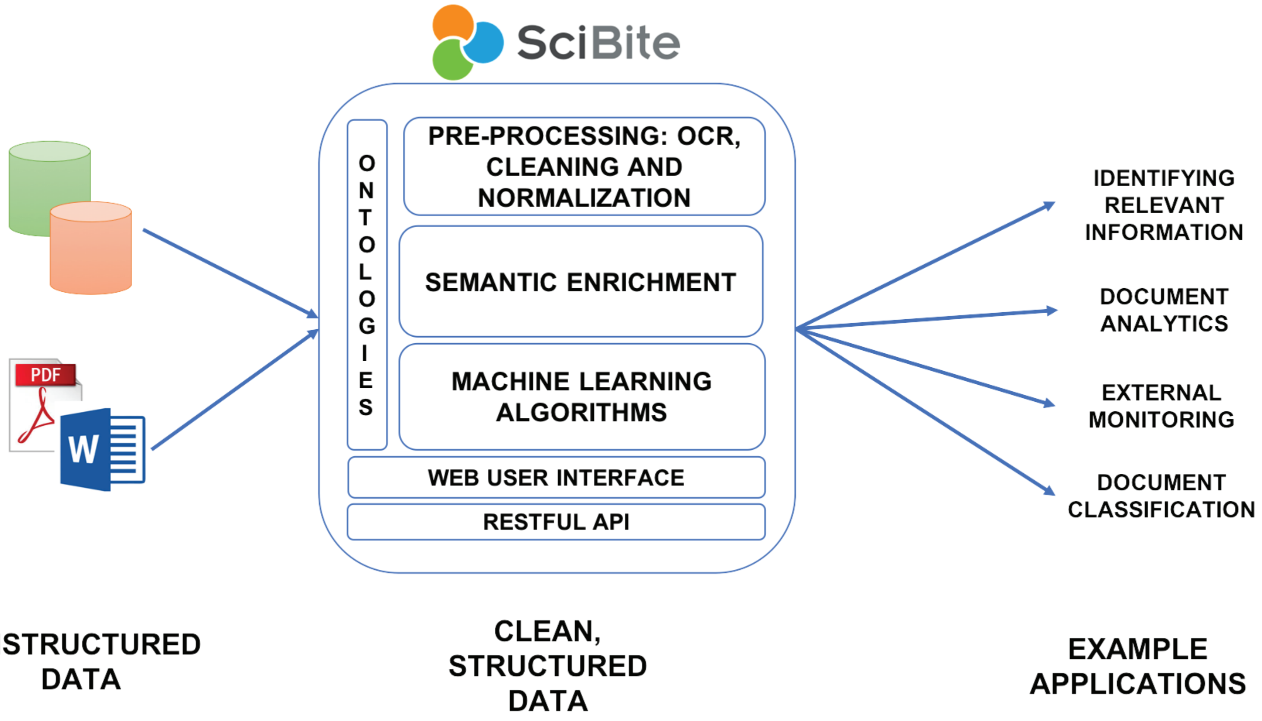

Figure 1: The SciBite platform – an integrated modular platform to enable effective machine learning

SciBite's platform provides a framework to incorporate different Machine Learning approaches, ensuring that it can be applied to a wide range of problems. The platform also includes a simple user interface to facilitate training of Machine Learning algorithms, enabling users to positively impact accuracy, and a RESTful API to enable systematic, automated analysis. Through semantic enrichment, SciBite ensures that all relevant data is found, regardless of which synonym is used as the search term. Users can create specific searches containing multiple relevant terms and entities forming so called search patterns.

Applications of the SciBite platform

Pinpointing critical information in a big data 'haystack': patent relevance

The Business Challenge: Monitoring the publication and patent activity of others is essential for several common business processes:

- Maintaining awareness of of technical developments in a particular technical field
- Identifying potential partners or in-licensing opportunities
- Early identification of potential patent infringements

However, the sheer volume of existing patents and the rate at which new ones are being published makes this a daunting task that requires experts to conduct an ongoing laborious search and review process.

Given the manual nature of this work, there is a strong possibility that something will be missed. Critical information within the patent may be hidden, or even obfuscated on purpose, making it challenging to find all relevant mentions of a specific target or indication of interest. For example, US Patent 20040009974 A1 ("Compositions useful as inhibitors of protein kinases") mentions over fifty different proteins. Without a detailed review, it is difficult to filter through the noise of irrelevant reagents and identify the target(s) that are actually relevant.

This issue is further compounded by the issue that there may not be an explicit link between terms of interest. For example, in US Patent 20040029902 A1 ("2,4-pyrimidinediamine compounds and their uses") the target (Syk Kinase) in only mentioned in the claims and body of the patent (and not in title or abstract) whereas the effector (Immunoglobulin E) is only mentioned in the abstract (and not the title or claims).

The SciBite solution:
SciBite has developed an algorithm which can analyse a patent and predict the most important target or indication mentioned within it. This is achieved by applying Semantic Analytics to extract a wide range of entities (including targets and diseases) from a patent before applying an algorithmic model, combining multiple quantitative, qualitative and linguistic features, to rank each entity based on significance. The ranked entities are then categorised into one of three levels based on their relevance within the patent: High (critical entities that are the primary focus of the patent), Medium (not critical, but likely to be significant) and Low (mentioned, not not important).

When compared with a manually curated dataset, Patent Analyzer identified 89% of targets, with 74% having a High or Medium relevance. Patent Analyser has also been independently validated through deployment in the SureChEMBL patent tool[3], where similar high levels of performance were demonstrated.

This solution is not limited to patents – SciBite has also applied similar principles to identify relevant mentions of targets and indications of interest within published literature.

Business Benefits:

- Rapid, accurate identification of the most relevant targets and indications in a collection of unstructured patents
- Significantly reduce reliance on the valuable time of experts
- Condense the patent review cycle, facilitating more frequent monitoring

Gaining scientific insight: rare disease phenotype triangulation

The Business Challenge:
In the past two decades, rare diseases (defined as those that affect less 1:1500 people in the US) have attracted increasing attention and national government support, including economic incentives and streamlined regulatory pathways to encourage pharma to invest in the development of rare disease therapies[4]. Even though genomic profiling of patients has resulted in greater numbers of people being diagnosed with rare diseases, characterization of rare diseases at the molecular level remains a challenge since the amount of research is still at a relatively low volume.
The fact that some rare diseases share similar phenotypes with common, well understood conditions, forms the basis of an inference-led approach to understanding them. However, evidence of disease similarity is often hidden within unstructured biomedical literature, necessitating a time consuming and costly review process to stand any chance of identifying relevant linkages.

The SciBite Solution:
SciBite has developed a method which quantifies disease similarities identified within the biomedical literature based on their phenotypes. As a first step, Semantic Analytics was used to extract co-occurring pairs of conditions and clinical signs from over 25 million MEDLINE abstracts. Machine Learning algorithms were used to rank these relationships and predict how scientifically interesting they are, for example based on how often the diseases co-occur compared to how often they appear independently. The resulting information was used to create a knowledge graph representing the strength of connectivity between diseases based on shared phenotypes or ‘phenotype signatures’. Where there is strong overlap in phenotype signatures, we can hypothesize that a disease pair could share an underlying mechanistic relationship and use this to classify poorly characterised diseases.

SciBite’s algorithms can extract themes from the scientific literature without any human intervention and have been proven to successfully predict disease relationships without any prior knowledge of the diseases and to infer relationships without any direct evidence, based on the strength of indirect evidence.

SciBite has extended this approach by including additional data sources, such as gene association data from DisGeNET and protein-protein interaction data from iRefIndex[5] to go beyond phenotypes and classify diseases based on richer signatures comprised of genomic, proteomic and phenotypic information.

[3] http://surechembl.org

[4] Tambuyzer E. (2010). Rare diseases, orphan drugs and their regulation: questions and misconceptions.Nat. Rev. Drug Discov. 9:921-929; Melnikova I. (2012). Rare diseases and orphan drugs. Nat. Rev. Drug Discov. 11(4):267-8

[5] See www.disgenet.org/ and http://irefindex.org/

Business Benefits:

- Readily find direct links between diseases in the biomedical literature
- Discover new mechanistic links between diseases which may have never been previously (or explicitly) stated
- Gain a better understanding of the mechanisms behind disease
- Identify opportunities for drug repositioning

Achieving comprehensive and timely monitoring: automating pharmacovigilance screening

The Business Challenge:
Pharmaceutical companies are obliged to monitor a range of credible data sources for possible causal relationships between their drug(s) and an adverse events, rapidly assess their validity and notify regulatory bodies within a reasonable timeframe. Adverse event 'signal detection' needs to cover related compounds, such as those from the same class and/or with the same target. Due to the sheer volume of new scientific content published each week, a typical weekly monitoring process requires several days of laborious and subjective manual review to complete.

As a consequence, many Pharmaceutical companies outsource the pre-screening of documents, but this has a significant drawback: reviewers simply don't have the time to read all of the available content relevant to the drug(s) of interest and the the serious implications of missing a critical signal result in an incredibly risk averse process, generating many false positives.

Pharmaceutical companies need a strategy that minimises or eliminates the risk of missing a serious adverse event or reporting deadline.

The SciBite Solution:
SciBite has developed a novel integrated approach, combining Semantic Analytics and Machine Learning, to automate the identification and prioritisation of potential adverse events. Firstly, a rules-based approach was used to automatically process the full text of a large volume and diversity of biomedical content and rapidly identify all mentions of relevant drugs and adverse events, without being limited by the indexing terms used by the data source.

The resulting documents were used as input for an algorithm that had been trained with a range of phrase constructs that describe adverse events. The algorithm not only identifies possible connections between drug and adverse events but also calculates an associated confidence value, predicts if the identified relationship describes a positive or negative association. Using an iterative approach to adapt the search patterns, the algorithm achieved over 90% accuracy.

By combining this with classification algorithm, and business rules defining how documents containing potential safety signals should be prioritized and routed, the right people are alerted to important safety signals in near real-time without the need to wait for the earlier stages of triage process to be finished.

Business Benefits:

- Identify and prioritise potential adverse events with high confidence in near-real time
- Avoid the trade-off between the volume of content and the time available to review it
- Eliminate the significant costs associated with outsourcing pre-screening of document for signal detection
- Automate pre-screening, enabling the time of expert resources to be used more productively

Automated classification of documents: streamlining knowledge transfer

The Business Challenge:
The acquisition and integration of a company or in-licensing of a compound requires the transfer of associated content, typically in the form of a large volume of unstructured documents that lack a consistent format or layout. The cataloguing of such documents and subsequent filing into an existing repository is a time consuming and error-prone manual process and is a barrier to successful knowledge transfer.

For example, the FDA requires pharmaceutical companies to assign non-clinical study reports to a single-sub category of the electronic Common Technical Document (eCTD) Module 4 hierarchy[6]. For example, Pharmacology reports must be assigned to one of the three Pharmacology sub-categories (Primary, Secondary or Safety Pharmacology). This requires each one of potentially hundreds of documents to be opened, reviewed and curated by an expert who assigns the document to a specific classification and records key metadata, including the compound and species.

[6] See https://www.fda.gov/downloads/drugs/ucm163175.pdf

The manual classification of nonclinical study reports for an acquired compound can take hundreds of hours and the subjective nature of this process can result in inconsistencies when different people are responsible for the review process. Even when correctly classified, the sheer volume of content makes it impossible to find critical information, such as safety data.

The SciBite Solution:
SciBite has used two complementary approaches to achieve automated document classification. The first approach applies a rule-based approach to classification based on the title of the document (e.g. if 'safety pharmacology' appears in the title then it should be classified as a Safety Pharmacology document). Intuitively, an initial approach might be to base classification on the title of the document. However, the computational extraction of titles from PDF documents is a notoriously difficult task, since the title of a document is not always explicit and its location and format is rarely consistent. To overcome this challenge, SciBite developed and trained an algorithm to understand the structure of a document and detect and extract the title and other important elements and metadata, such as the synopsis and author.

Title-based classification was augmented with a secondary Machine Learning approach in which entities are extracted from the entire document and assign it to a category with highest possible confidence. The predicted classification and other extracted metadata automatically populates fields in a document summary form for entry into the document management system. In validation testing, SciBite achieved greater than 85% classification accuracy.

Business Benefits:
- Repeatable and automated classification of documents, with improved quality and consistency
- Automate the alignment of key incoming documents against classification standards
- Reduce tens or hundreds of hours of manual effort to minutes
- Eliminate bottlenecks and increase capacity of the acquisition team by over 70%
- Ability to provide timely and accurate responses to urgent regulatory and safety questions

Summary

The use cases outlined above highlight the great potential of combining Semantic Analytics with Machine Learning. SciBite's platform can address a broad set of business and scientific problems, including the ability to automate processes related to:
- Pinpoint relevant information
- Monitoring external information
- Document classification

Ultimately, SciBite delivers robust and repeatable results and conserves the valuable time of experts.

SciBite's award-winning platform[7] is the culmination of tens of years of experience applying Semantic Analytics to Pharmaceutical data. Global pharmaceutical companies and emerging biotechs have partnered with SciBite and are leveraging our unparalleled know-how to unlock the potential of the wealth of unstructured biomedical literature that is now at their disposal.

To learn how SciBite can transform your scientific and business processes with a flexible, easy to use, accessible environment, speak to one of our experts today or email us at contact@scibite.com.

[7] SciBite has been recognised with a series of awards, including Bio-IT World's Best of Show 2017 and the British Chamber of Commerce in Japan's 2017 British Business Award for Innovation. The latter is in recognition of our transformation of data management in the life sciences, and the opportunity this has brought for Japan to gain a global advantage in the sphere.

SciBite
an ELSEVIER company

SciBite's data-first, semantic analytics software is for those who want to innovate and get more from their data. At SciBite we believe data fuels discovery and we are leading the way with our pioneering infrastructure that combines the latest in machine learning with an ontology-led approach to unlock the value of scientific content. Supporting the world's leading scientific organisations with use-cases from discovery through to development, SciBite's suite of fast, flexible, deployable API technologies empower our customers, making it a critical component in scientific, data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com

Head Office:
SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

www.scibite.com
✉ contact@scibite.com
LinkedIn: SciBite
Twitter: @SciBite
+44 (0)1223 786 129