SciBite
an ELSEVIER company

Use case

How ontologies and machine learning work together: A critical precursor to effective artificial intelligence

SciBite
an ELSEVIER company

How ontologies and machine learning work together: A critical precursor to effective artificial intelligence

The struggle to effectively utilise the increasing volumes of data available is a common challenge in the Life Sciences research industry. Artificial Intelligence (AI) is frequently touted as a potential solution to extract valuable insights from large volumes of heterogeneous data. However, tangible successes to date have been relatively few. Areas bearing the greatest demonstrable success often utilise machine learning (ML), yet even these are at the mercy of the quality of the source data. Scientifically naive systems struggle to deal with the complexity and variability of unstructured scientific language. In a recent survey of over 16,700 data scientists, the most commonly cited challenge to undertaking machine learning was "dirty data".

According to the Pistoia Alliance, "Maybe the biggest hindrance to using AI/ML effectively is both the volume and quality of data that exists. The models that are being built will only be as good as the data that has been used... As an industry are we being overly optimistic and simplistic about the quality of data that we are feeding our smart new AI/ML pipelines?".
To illustrate this problem, the figure below lists a selection of the many varied words and phrases scientists use to refer to the term 'male'.

37 year old male	initial phase male	male fetus	six males mixed
600 yr. old male	m	male plant	stallion
adult male	make	male, 8 weeks old	steer
bull	makle	male, castrated	sterile male
castrated male	mal e	male, pooled	strictly male
cm	male	males	tetraploide male
dioecious male	male (7-2872)	man	type i males
diploid male	male (7-3074)	men	type ii males
drone	male (m-a)	normale male	virgin male
engorged male	male (m-o)	ram	winged and wingless males
fertile male	male caucasian	rooster	young male
four males mixed	male child	s1 male sterile	
individual male	male fertile	sex: male	

male (note: this sample was originally provided as a \female\ sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a male individual)

Figure 1: *A selection of the many different terms used to describe 'male'. Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI*

System designers often constrain fields to mitigate this risk but, as the next figure below shows, scientists can still be unintentionally inconsistent in the way they record their results (see Figure 2, below).

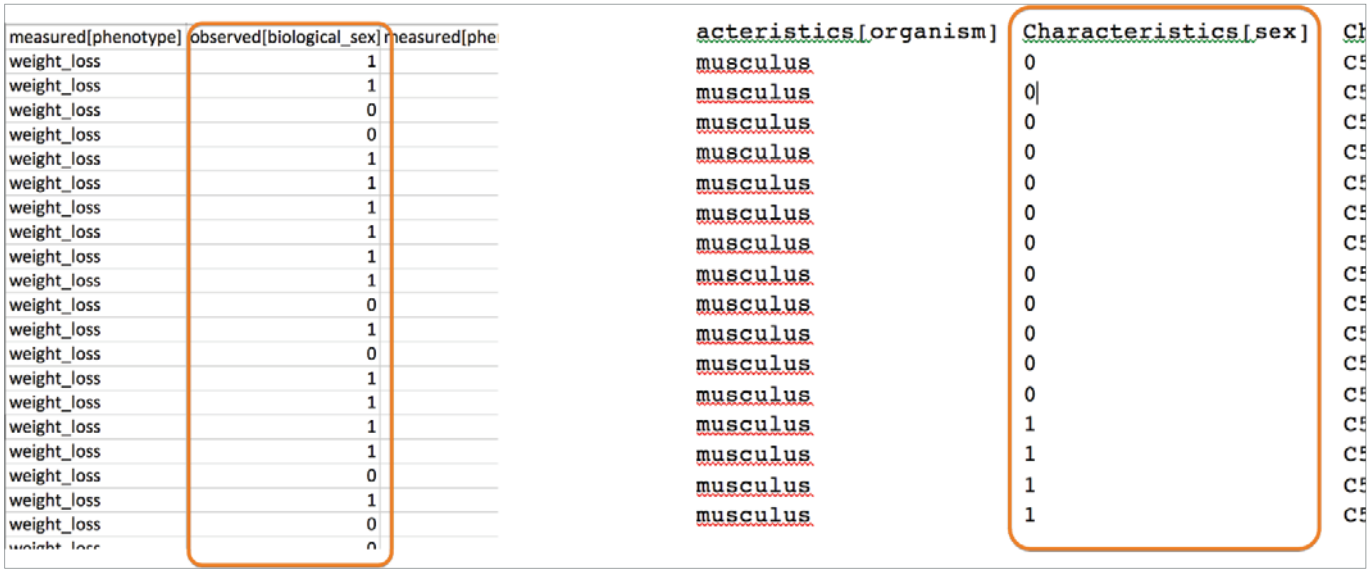
While both systems illustrated above may work well independently, but they have not been designed to be interoperable. The data is not adequately described, or 'tagged', which presents a challenge for downstream analysis. When considered in isolation, both datasets could be considered to be 'clean' but, from the perspective of a data aggregation project, the data is 'dirty'. The re-use of data outside of it's original source system is something that is often overlooked and is a problem that will get out of control very quickly if not correctly managed. Ensuring data cleanliness within a single system and ensuring it's extensibility to other systems is critical to enable multiple sources to be aggregated.

Since AI systems generally learn by example, 'dirty' data creates a barrier to their effective implementation.

For example, without comprehensive coverage of all possible variants of a term of interest in a training data set, the output of a machine learning algorithm will include a large number of outliers.

Similarly, ambiguity can lead to a term being mistaken as the wrong type of entity. For example, ALS is used as an abbreviation for the disease Amyotrophic lateral sclerosis. However, it is also used as a synonym of the gene 'Insulin Like Growth Factor Binding Protein Acid Labile Subunit' (IGFALS). Thus, if we use just "leave it to the computer" a machine learning algorithm could mistakenly incorporate literature related to IGFALS into a model for the disease ALS, resulting in an incorrect link between ALS and insulin signaling with potentially dangerous consequences.

High quality, contextualised data is not only key to ensuring experiments are reproducible but is also a necessary prerequisite for AI to be effective and deliver accurate and true outputs.



measured[phenotype]	observed[biological_sex]	measured[phe
weight_loss	1	
weight_loss	1	
weight_loss	0	
weight_loss	0	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	0	
weight_loss	1	
weight_loss	0	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	1	
weight_loss	0	
weight_loss	1	
weight_loss	0	

acteristics[organism]	Characteristics[sex]	Ch
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	0	C5
musculus	1	C5
musculus	1	C5
musculus	1	C5
musculus	1	C5

Figure 2: *Inconsistent coding of data. In the example on the left, Male =1 and 0=Female. In the example on the right Female =1 and 0=Male*

SciBite
an ELSEVIER company

SciBite: Enabling machines to learn more effectively

Giving AI a 'Head Start'

SciBite contextualises unstructured scientific text and transforms it into clean, 'self-describing', machine-readable data. It uses established, controlled ontologies to apply an explicit, unique identifier, meaning and description to scientific terms. This ensures that, for example, the terms PSEN1, Presenilin-1, AD3 and PSNL1 are all interpreted as the same Alzheimer's-related gene.

SciBite harmonises data by exploiting ontologies to automate semantic enrichment and annotation, whilst also coping with ambiguities such as synonyms, typographic errors or cryptic data, such as project codes, cell line IDs, and internal drug abbreviations.

Without using ontologies, an AI model needs to learn everything from scratch. It's akin to going into a library to find a book without some form of classification or indexing to guide you to its likely location. The Dewey Decimal system was developed to help people find the right information faster and ontologies serve a similar purpose in science.

Since ontologies encapsulate a common model of knowledge associated with a given domain, they can give AI a 'head start' by identifying what is known in context of an existing scientific framework. For example, a machine learning algorithm doesn't need to be 'taught' that Crohn's is an inflammatory disease if this relationship is already encapsulated in the ontology that is used to enrich the input. Similarly, by annotating content with the MedDRA ontology, the annotated concepts will be known to be adverse events, rather than simply predictions of something that might be.

More sophisticated pattern recognition

Many applications of AI involve pattern recognition, but their accuracy is highly dependent on the data being unambiguous. This problem can be illustrated using the example of training a machine learning algorithm to identify phrases in the biomedical literature that are indicative of a protein-protein interaction. For example, an algorithm can be trained to correctly identify that "...the binding of repaglinide to HSA in human plasma..." refers to an interaction between two entities, but one is a drug and the other is a protein.

Through semantic enrichment, scientific terms can be explicitly tagged with a Uniform Resource Identifier (URI) as representing a particular type of entity, enabling machine learning algorithms to be trained to recognise that phrases such as "Cdc53 contains independent binding sites for Cdc34 and Skp1 suggesting it functions as a scaffold protein within an E2/E3 core complex" specifically refer to a protein-protein interaction. Normalisation of data through semantic enrichment reduces ambiguity and amplifies the signal, increasing the accuracy of pattern recognition.

A solid foundation for knowledge graphs

Knowledge graphs provide essential context for many AI applications. Clean, accurately described data becomes interoperable, provides a solid foundation from which to build graph models without having to worry about the original naming convention used by each source. For example, a gene mentioned in a document can be linked to related quantitative gene expression data via its GENE URI. The workflows and rules used to query the knowledge graph will also be simpler and quicker to process. The resulting reduction in complexity coupled with increased scientific rigour makes the automated AI process less of a 'black box' and provides increased confidence in the decisions that are made.

Increased resolution for neural networks

AI approaches typically require numerical input. However, a method known as word2vec can overcome this limitation by encoding words as numerical vectors. This enables queries to expressed as a mathematical operation, for example: melanoma – skin + brain = glioma. The accuracy of word2vec depends on how well the source text is standardised. Semantic tagging increases the resolution of a numeric vector for a particular word. For example, the word 'hound' is used less frequently than 'dog' so it's numeric vector would tell you very little about its semantic relationships. However, if the numeric vectors of 'hound' and 'dog' are combined then the resolution, or semantic knowledge, of the vector is increased significantly.

The same applies to genes: BRCA1 is used far more frequently than it's synonyms PPP1R53 or BROVCA1. Semantic enrichment normalises and categorises entities prior to training word vectors, resulting in better resolution and improving its discoverability. For example, it is possible to ask questions such as 'find the indications most semantically similar to BRCA1'.

Accurate document classification

The combination of semantic enrichment and AI can also be applied to document classification problems. Semantic enrichment not only provides additional context to enable more accurate classification, but can also reveal the most common scientific metadata tags and entity types associated with documents in a given class. For example, one class might include mentions of many biological processes but only one compound, where as another class might have a totally different entity profile. This helps users to understand why the classification process has assigned a given document to a specific category, making the automated process less of a 'black box' and providing increased confidence in the decisions that are made.

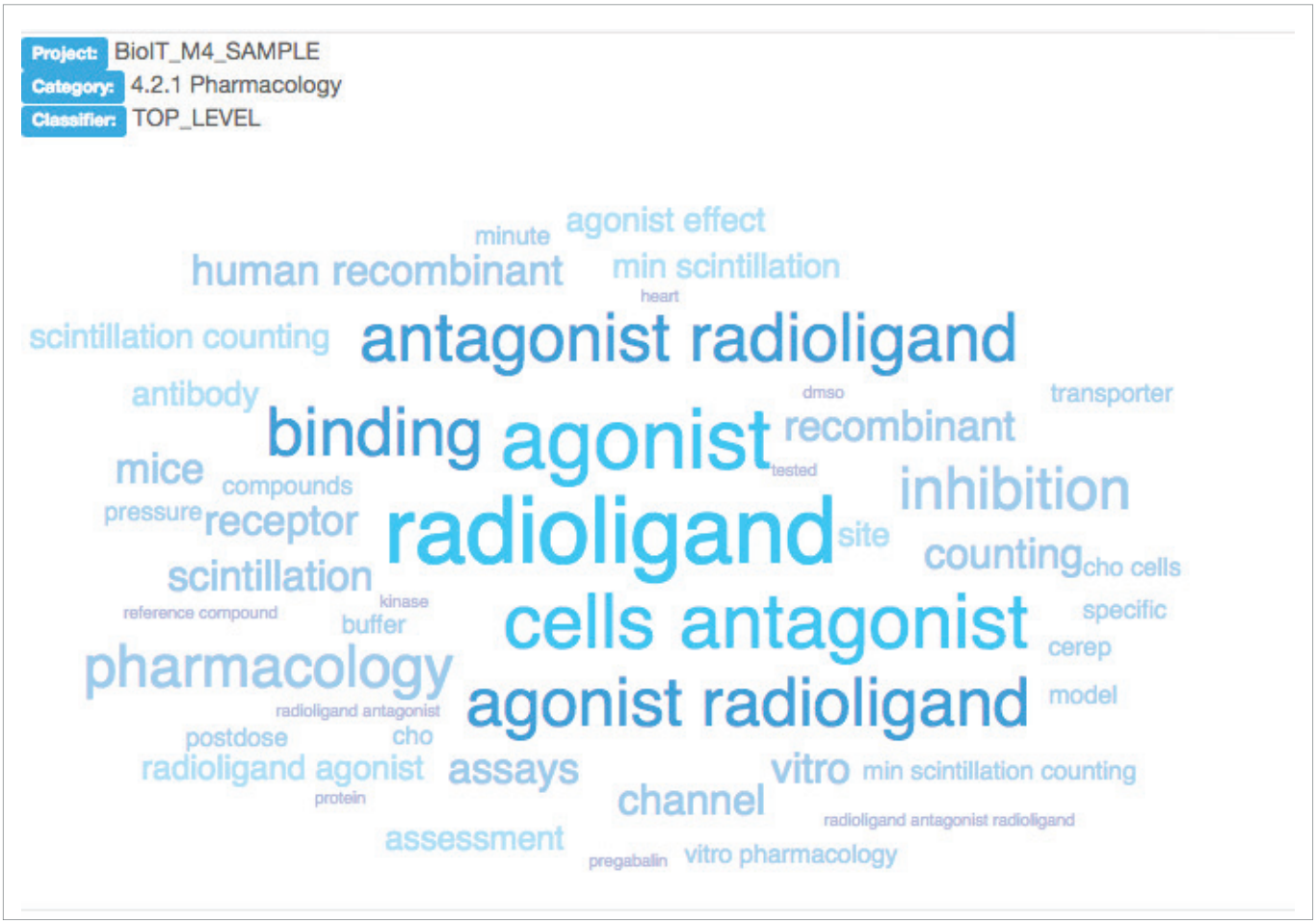


Figure 3: A word cloud illustrating the most discriminative features found in documents assigned to the pharmacology category of the FDA's electronic Common Technical Document (eCTD) Module 4 hierarchy using SciBite's ClassifR application

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