# Data-driven NLP plus machine learning equals better drug-discovery insights

*Machine learning is generating considerable excitement in the biopharmaceutical community due to its potential to revolutionize pattern identification, predict successes and failures, and improve research decision-making.*

The application of machine learning to drug discovery has recently been an area of increasing focus, with new, practical examples of use cases further fueling industry attention and interest.

To fast track new drug development via early predictions of drug success or failure, make effective use of real-world data (RWD) in pharmacovigilance studies, or find new uses for existing drugs—and ultimately reduce costs and improve health outcomes—life science organizations must make effective use of the wealth of unstructured data available, to understand patterns and trends via machine learning. **Effective use of natural language processing (NLP) text mining on this unstructured data is critically important as a first step.**

**The application of machine learning in life science is dependent on having access to good quality data upon which to train algorithms.** Structured data can provide valuable knowledge, but up to 80% of the data that can drive drug discovery insights is unstructured—for example in ClinicalTrials.gov records, MEDLINE abstracts, or from RWD sources such as medical records or "voice of the customer" (VoC) feeds. This unstructured data must be extracted via NLP text mining and, if desired, combined with structured records to obtain a detailed, comprehensive view.

## Selecting the right tool for life science NLP

The biopharmaceutical community is increasingly interested in building machine learning models to develop solutions to challenges across the drug discovery pipeline. The data needed to build such models can be extracted from different text-based sources via NLP text mining. There is a variety of options for NLP tools, and choosing which software to use presents important considerations.

| | |
|---|---|
| **STATISTICAL NLP SYSTEMS** | These depend on example data to identify patterns in new data. This can be challenging in commercial settings where good quality, large-scale, and representative annotated data is rarely available. Annotating a gold standard is expensive, requiring development of detailed annotation guidelines and use of multiple annotators to judge the reproducibility of results (measured by inter-annotator agreement). |
| **RULE-BASED NLP SYSTEMS** | These rely on a specialist to enumerate the types of language rule or pattern that represent drug discovery concepts, instead of inferring the presence of the concepts from labeled data. Since the rules are derived from human understanding of language, they can be specific and accurate in a way that machine-learned models cannot, and may also do better when applied to new data. Large numbers of handwritten rules can be hard to maintain, however, and will only capture the patterns that the specialist has thought of. |

# A new type of data-driven, rule-based NLP

The Linguamatics NLP platform represents an agile new approach to rule- or pattern-based NLP, making the process of identifying drug research concepts simple for life science researchers. Patterns can be easily created, edited and reused in a rapid, iterative process, even when the user is not a specialist in computational linguistics or NLP. Since it is not primarily a statistical NLP system, the NLP platform does not require labeled data, and it is able to make effective use of dictionaries and ontologies to increase recall.

Linguamatics NLP has a transparent, easily editable query language for expressing extraction rules; and a search engine architecture that allows fast, data-driven methodology for refining queries.

This approach was used to provide one of the top results in a 2015 i2b2 challenge (http://bit.ly/2r1rXKH).

# The NLP platform and downstream machine learning use-case examples

Ultimately, the NLP platform queries can produce data, or features, to be used in downstream machine learning models, and this is an approach often used by Linguamatics customers. For example, in a 2017 journal paper (https://doi.org/10.7717/peerj.3154), **Eli Lilly** researchers described how they have extracted potential new uses for existing drugs by mining adverse event data in ClinicalTrials.gov, to calculate ranking statistics for the treatment-indication association.

Another **top 10 pharma company** uses the NLP platform to annotate and categorize VoC call feeds for pharmacovigilance. VoC call transcripts are a rich seam of potential patient-reported outcomes, side effects, drug interactions and more. Researchers built an agile text-mining workflow to process and make sense of the unstructured call feeds. The extracted features are used as the structured substrate for machine learning algorithms, to assist in categorizing the call feeds and to build predictive models around the different products.

In a 2016 publication, researchers from **Roche and Humboldt University of Berlin** described how they used NLP to systematically identify all MEDLINE abstracts containing both the protein target and the specific disease indication of a known set of successfully approved or failed cancer therapeutics (for example, abstracts containing both “Her2” and “breast cancer,” or “c-Kit” and “gastrointestinal stromal tumor”). The researchers applied machine learning classifiers and found that the NLP-extracted data features could be used to predict success or failure of target-indication pairs, and hence, approved or failed drugs.

---

*“These patterns allow predicting success of drugs in Phase II or III with remarkably high accuracy.”*

*— Heinemann, F., Huber, T., Meisel, C., Bundschus, M. and Leser, U. (2016) “Reflection of successful anticancer drug development processes in the literature,” Drug Discovery Today, Vol. 21(11), pp. 1740–44. Available at: http://bit.ly/2eQGIuX*

## How does the Linguamatics NLP platform accelerate the development of supervised machine learning techniques?

- **No need for analyst annotations of raw data:** Supervised machine learning techniques require good quality training data with standardized annotations; these are not easy to obtain for most tasks, can take months to generate and incur significant costs due to the expert scientific knowledge required. The NLP platform can be used to produce training data with much less human attention in much less time. Evaluating samples of NLP platform query results and comparing differences between query runs requires specialist time measured in hours rather than months.

  Machine learning also suffers when the distribution of annotations across key concepts is not even. Any skewed distribution of annotations means that machine learning may have very few training instances, for example for specific risk factors. Using the NLP platform it is possible for researchers and data scientists to incorporate prior knowledge of the types of linguistic construction that might occur.

- **Dealing with uncertain business problems:** To be useful in a business context, machine-learned models must function well with metrics that are ill defined and subject to change. Since the human annotation of data and training of the models can take months, this can deter organizations from taking advantage of machine learning.

  With the Linguamatics NLP platform, flexibility is at the heart of it—modifications can be made, and results seen, in seconds, and new features can be created quickly and incorporated into models. Often it takes a few rounds of using the NLP platform for users to understand their data well enough to be able to specify what they want to extract, and what distinctions they want to make. For example, lists of parameters created by researchers will often be incomplete, or there will be subtleties that only arise once you start looking at the real data. This can be accommodated easily when using the NLP platform.

- **Shortening development time:** Creating features for machine learning models is time-consuming and technically challenging; even highly qualified workers with postgraduate degrees routinely fail to execute them effectively. Not surprisingly, in industry, machine learning-based systems are often deemed risky to adopt, and difficult to understand and maintain. This is largely due to the opaque nature of the models and the infeasibility of gathering annotated data in many real-world scenarios. The NLP platform simplifies feature engineering by making it easy to harvest new keywords and phrases, as well as using existing terminologies to speed up the process.

## The NLP platform and machine learning: The best of both worlds

This analysis has shown that the Linguamatics NLP platform is highly complementary to the development of machine learning algorithms. Machine learning models benefit from a clear, systematically-extracted, comprehensive set of data features; to obtain these from unstructured text needs powerful NLP. The NLP platform accelerates access to such data features, giving machine learning projects a much higher chance of success when using unstructured data.

**CONTACT US**
+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)
nlp@iqvia.com
**linguamatics.com**