Explainable Intelligence

New approaches to artificial intelligence have the potential to eradicate some of the challenges associated with dealing with large volumes of data

By Amy Guarino at Kyndi

"Remember the Three Princes of Serendip who went out looking for treasure? They didn't find what they were looking for, but they kept finding things just as valuable. That's serendipity, and our business (drugs) is full of it." – George W Merck

Drug hunting, or drug innovation, has improved since George was made President of the family business in 1925. That said, it is widely recognised that there is a tremendous amount of value that could be unleashed by exploiting existing knowledge locked in internal and external research documents, corrective and preventative action documents, and audit findings.

This article will review the challenges of the current processes for scientists and researchers, the key requirements for new processes and capabilities, the pros and cons of different approaches (especially those involving artificial intelligence [AI]), and the benefits for pharmaceutical companies, patients, and the industry overall.

Keywords
Data handling
Drug development
Natural language processing

Current State

Developing a new drug requires significant investment in time and resources. According to the May 2016 *Journal of Health Economics,* the pre-tax capitalised cost per drug approval was US $2.6 billion (in 2013 dollars) and was increasing annually by 8.5%. The average time for that development is 10-12 years. Much of the insights and information developed during this time is collected in various text documents, which can be difficult to utilise across a broad set of people due to the use of alternative terms and concepts, as well as limitations around technology to quickly analyse the information and understand the most critical aspects or insights without requiring a subject matter expert to personally review the information. Some progress is being made by utilising machine learning capabilities to assess the structured data available, but those results are the product of a 'black box' system that is opaque and hidden from understanding or justifiability.

Discovery often begins with target identification: choosing a biochemical mechanism involved in a disease condition. Thousands of molecules for each prospective drug are screened and experimented with to validate the potential impact. These experiments are documented in reports that are reviewed and audited for process validity.

Once a candidate molecule shows potential promise, then it must be characterised with specific attributes around size, shape, strengths, and weaknesses. Considerations and concerns about potential drug formulation and delivery are also reviewed. These efforts denote the preclinical trial efforts, after which, a series of clinical trials take place that further confirm the efficacy and safety of a potential drug. All of these efforts spin out dramatic reams of text data that could be utilised in the future to improve the efficiency of the drug innovation process.

Challenges in dealing with this volume of data include:

- The volume of information available is overwhelming – more than anyone has time to read
- The data of interest is buried in this information
- The ways in which data is presented makes it hard to 'connect the dots' regarding trends and relationships

As a result, text information often languishes in data repositories and is not effectively utilised across the larger organisation. If it is utilised, it requires significant time and energy to review and gain insights.

New Approaches

By providing ways for scientists to share or search this rich store of data by enabling them to quickly analyse this text and extract key concepts and their connections, the scientists and researchers can discover new connections and concepts and, hopefully, reduce the investment in time and money to develop new and more effective drugs.

Novel approaches in AI are unifying probabilistic and logical methods with explainability as a core requirement. A critical function of this technology is to answer questions, recognise similarities, and find analogies rapidly. These features enable the technology to build models that are made up of a series of questions, for which the software attempts to generate answers from the data.

The reasoning can then be justified by pointing to specific instances in the data and highlighting the relevant words and phrases. By providing auditability, scientists and researchers can confidently assess the results when applying them to further analysis or to make immediate decisions.

These types of solutions augment human intelligence by using natural language processing (NLP) and applying symbolic representation of information in a compact knowledge graph and allowing users to search, explore, and interact with information in ways that reduce the time to results for almost any knowledge- or information-heavy task.

NLP is an AI technique that allows computers to be able to understand language. The development of NLP applications is challenging because computers traditionally require humans to 'speak' to them in a programming language that is precise, unambiguous, and highly structured, or through a limited number of clearly enunciated voice commands. However, human speech is not always precise – it is often ambiguous, and the linguistic structure can depend on many complex variables, including slang, regional dialects, and social context.

As defined in the Semantic Web, a knowledge graph is "a graph, composed of a set of assertions (edges labelled with relations) that are expressed between entities (vertices), where the meaning of the graph is encoded in its structure, the relations and entities are unambiguously identified, a limited set of relations are used to label the edges, and the graph encodes the provenance, especially justification and attribution, of the assertions…" Knowledge graphs are now used by Google, Apple, Amazon, Microsoft, and Facebook as ways to organise data and create or expose relationships. This approach represents a new way to think about data and questions that can be asked of that data.

Using these different AI techniques, users can:

- Identify, classify, and extract entities, insight, and connection from unstructured and structured text documents
- Create and use the knowledge graph technology without extensive preparation, including labelled datasets or model training data
- Search the combined knowledge compiled from both structured and unstructured data
- Create and use knowledge graphs composed from documents in multiple data formats, jargons, and languages
- Identify key people, places, and facts, along with their connections, using the knowledge graph

Improving Innovation

By leveraging these new techniques and approaches, drug innovators will quickly be able to analyse these vast stores of information and knowledge to quickly highlight key insights and connections which will lead to a more efficient and potentially more creative drug innovation process. Organisations are estimated to see between 10-100 times more efficiency. By addressing the challenges of building these solutions without requiring large labelled datasets, and by explaining the outcomes in human understandable language, the data and the approaches will be trusted.

'The Three Kings of Serendip' will rapidly find what they were looking for, but, also, they will soon discover "things that are just as valuable." Everyone in the world will benefit from this increase in drug innovation and efficiency, as costs for new treatments are reduced, time to benefit reduced, and ability to support treatment for orphan diseases increases.

Over the past 25 years, Amy Guarino has been an accomplished business leader who uses sales, marketing, and management methods learned at International Business Machines to quickly drive results at start-up and mid-size companies. Prior to joining Kyndi as Chief Operating Officer, she spent eight years at Marketo as the Vice President of Business Development. Amy has a BA from the University of Notre Dame, US, was selected for the Forbes 40 over 40 list, and is recognised as one of the Top 100 Women of Influence in the Silicon Valley. She has also served on numerous boards in addition to advising Silicon Valley start-ups.
Email: amy@kyndi.com