Are Graph Databases the Key Ingredient AI is Missing in Drug Discovery?

The drug discovery process is hugely data-intensive, making it an ideal application for artificial intelligence (AI) and machine learning. But it isn't as straightforward as it sounds. Gaining valuable insight can be complex. Database pioneer Emil Eifrem explains why graph software could be the missing link in better understanding data so that the power of AI can put it into context, and pull out the most salient information for drug discovery researchers.

Pharma companies are looking to tap into powerful new technologies like machine learning to streamline labour-intensive parts of drug discovery in a bid to increase efficiencies, better control costs and bring products to market more quickly.

The way knowledge is organised and represented in AI-powered systems can have a significant impact on how and what exactly can be learnt from it. Representing these relationships in a graph database can enable life scientists to spot hidden connections and shed light on cause and effect more quickly and accurately.

Today most models and techniques that provide the foundations for AI systems are not optimised for detecting or traversing relationships within datasets. Graph databases, however, have shown they have the power to link complex relationships – making them the ultimate data structure to power machine learning models, as the German Center for Diabetes Research (DZD) has recently spotlighted: "[Graph technology] enables a new dimension of data analyses to fight diabetes by helping us to connect highly heterogeneous data from various disciplines, species and locations to build an invaluable body of knowledge," according to Dr Alexander Jarasch, Head of Bioinformatics and Data Management at the DZD. "By applying modern machine learning techniques to our Neo4j graph, we are getting closer to understanding this complex disease to help diabetics and those with prediabetes."

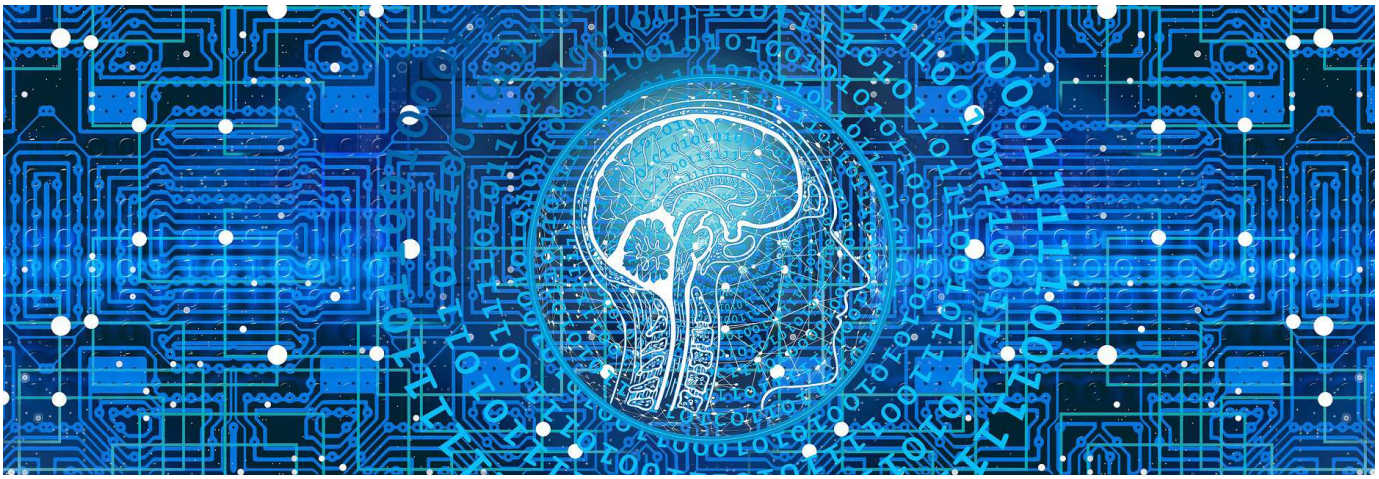
Graph database technology provided DZD with a whole new measure of data analyses to help in the fight against diabetes by contributing to the connection of highly heterogeneous data from various disciplines, species and locations to create a hugely valuable body of knowledge. By applying advanced machine learning techniques to a graph database, DZD's research team have become much closer to understanding the complexities of the disease and have moved a step forward in being able to help diabetics and those with prediabetes.

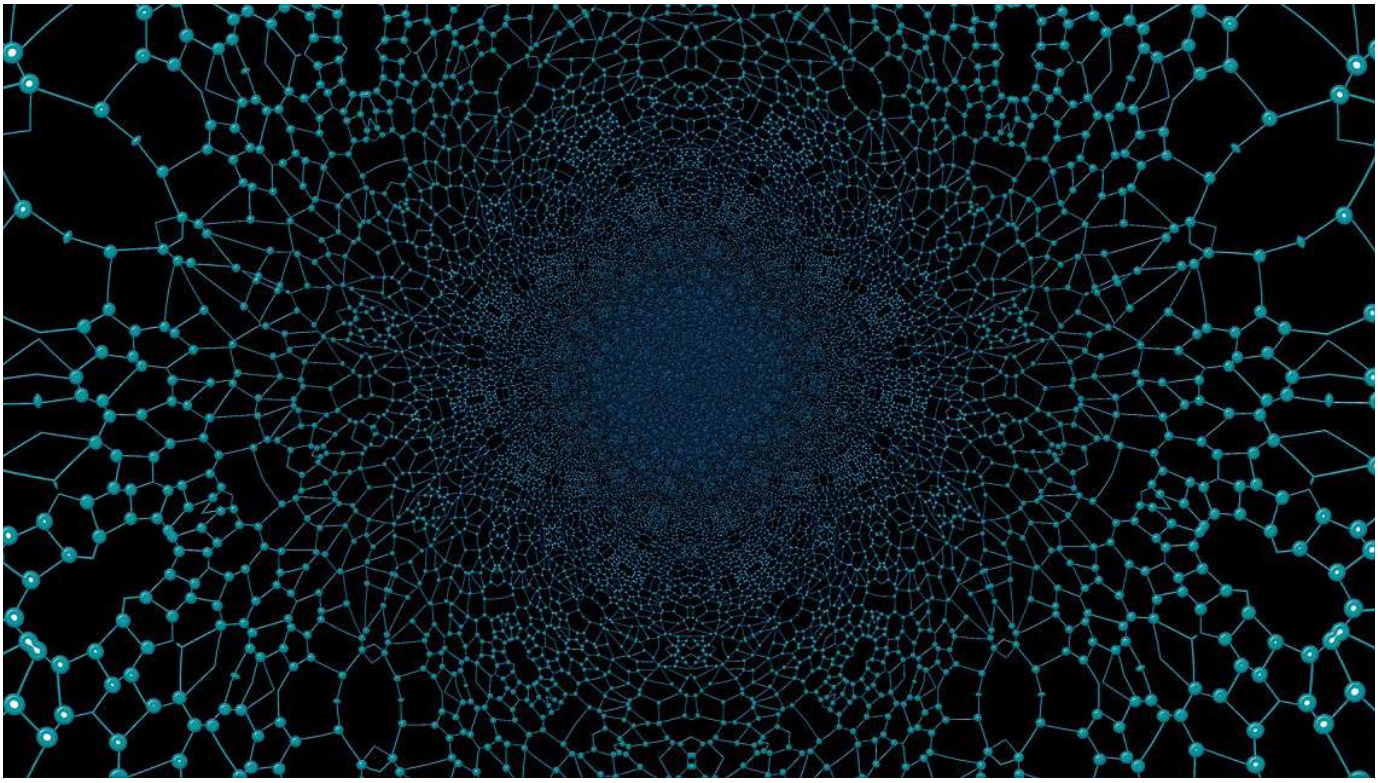
As we know, medical data is highly heterogeneous by its nature, ranging from cell-level to incredibly detailed data, often contained within the same study group. All too often, researchers want to link either end of the scale – the meeting point between different data sets – which is where the compelling results tend to sit. But this can be a gargantuan challenge.

In addition to this data complexity, researchers are also looking at huge amounts of data, often running into exabytes, which need to be distilled. Trawling through such large amounts of data, which put in perspective is the equivalent of more than 25 million whitepapers, is pretty much impossible for human teams. Working out how multiple researchers can access and collaborate on the data is also a challenge.

With data at this scale, which often comes in an unstructured format, data needs to be turned into a valuable research ingredient as quickly as possible – not just simply initially analysed and stored. Deep data mining and pattern detection with graph technology can provide a path to invaluable insight.

Graph software also has the innate power to collaboratively filter data. Collaborative filtering is the process of filtering information, using the information gathered by many users. Collaborative filtering is actually a technique used by recommendation engines. Information or patterns can be filtered via data sources, viewpoints, multiple agents and so forth. This approach allows research teams to work on data at the same time.





Three Big Issues

Traditional databases struggle to cope with three of the biggest issues in the life science sector – heterogeneity, complexity and scale. This is why we need to reconsider the way data has been historically modelled. Traditional SQL and relational databases find the volume and the unstructured nature of the data extremely difficult to handle. Scientists need diverse quality data to develop new drugs. When data is missing or there is insufficient data to feed AI algorithms to train and test accurate models, this makes it an impossible goal. In addition, this data is often stored in disparate database silos across different departments and formats, creating yet another bottleneck to research advancement.

The Novartis Institutes for BioMedical Research, the research arm of Novartis, has shown how it has harnessed the power of graph technology to get over these hurdles. Novartis has used graph software to help create a system of scalable biological knowledge. This isn't just about connecting huge amounts of heterogeneous data – it is also about enabling researchers to create a query for a particular kind of triangular relationship. These nodes are made up of chemical compounds, specific biological entities and diseases as described in specific research literature. The system needs to use uncertainty in key links as part of the query.

Graph technology allows the system to capture the strength of the relationship between the medical research text by encoding it in the properties of a graph and connecting or joining the dots between these terms. This then provides a foundation for later queries that link the literature to observed chemical or biological data. Results can be tested by removing links and monitoring how results differ.

Pharma, animal health and agrichemical giant Monsanto is another example. It has been working to get better inferences out of its plant genomics pipeline data. Before it recognised the power of graph technology it used relational databases, which resulted in millions of rows of data, which was a difficult read and made connecting relationships extremely difficult and time-consuming.

Monsanto's research teams wanted a tool that would enable it to provide an entire tree of ancestors from a single plant. They tested graph technology and can now process arbitrary-depth trees at what approximates to scale-free performance. This can not only be done quickly, but they can now analyse the ancestry of one million plants, instead of just one plant, in minutes.

The graph tool has totally changed Monsanto's research horizon. One of the graph software applications it developed, for example, touched every seed in its pipeline. The research time observed a 10x performance increase in this application, and a particular data scientist was able to replace one month of manual number-crunching with three hours of analysis using the graph-based system.

Discovering Hidden Patterns

These companies, amongst others, are exploring new frontiers in what can be achieved in graph-powered AI, which is essentially an intelligent system of connections.

This relationship-first approach to data puts it in real context and provides a foundation for accurate, smart predictions and ultimately informed decision-making. This is helping life sciences make real advances that will contribute to the discovery of therapies in the future.

Emil Eifrem

Emil Eifrem is CEO and co-founder of Neo4j. Emil famously sketched out what today is known as the property graph model on a flight to Mumbai in 2000. Since then Emil has devoted his professional life to building and evangelising graph databases, and is a frequent conference speaker and a well-known author and blogger on NoSQL and graph databases.