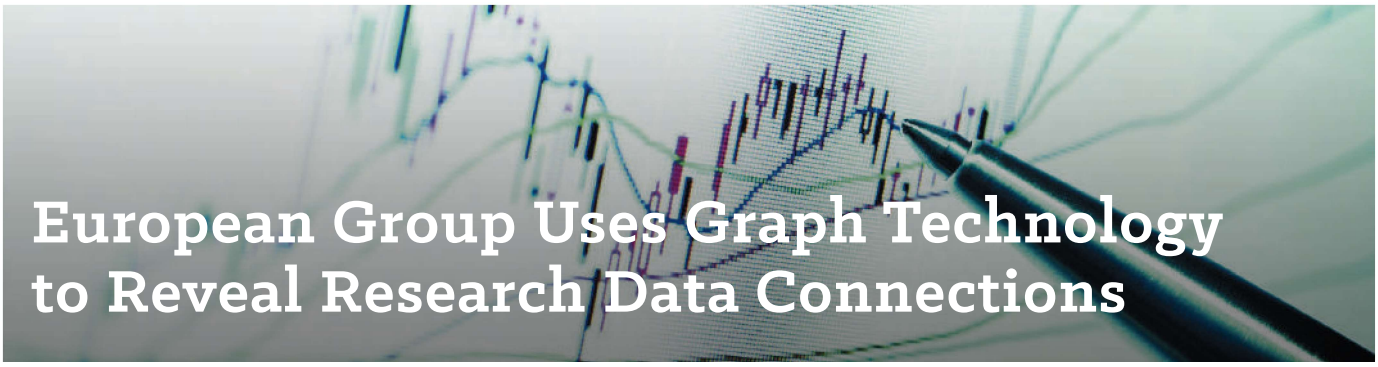

European Group Uses Graph Technology to Reveal Research Data Connections

A leading European research group is using graph-based data management techniques to make research inroads – and support efforts to find drugs and vaccines for COVID-19. Neo4j's Alicia Frame reports.

The amount of data in scientific research is growing exponentially. Big Data and other advances in data science hold huge potential for medical researchers to gain previously unattainable insights that could yield medical and pharmaceutical breakthroughs. But the life sciences sector is failing to maximise the potential.

The interdisciplinary nature of life sciences research and the highly heterogeneous nature of medical data present a real challenge when it comes to modelling scenarios and making connections. Classic analysis tools of the past 30 years – spreadsheets and relational databases – have reached their limits. They cannot cope with the huge heterogeneity associated with the vast amounts of data, let alone the complexity of the multiple sources data scientists need to explore. The problem is intensified by the fact that a significant part of today's data is highly unstructured and highly connected.

Researchers need a dynamic way to leverage and connect big data to provide new, valuable insights. Database software called graph technology has appeared as a viable and powerful alternative. One prominent convert to the graph technology approach is the German Centre for Diabetes Research, DZD.

DZD brings together experts from across Germany to develop effective prevention and treatment measures for diabetes across multiple disciplines. DZD's research network accumulates a huge amount of data from clinical trials and patient information covering a host of disciplines, distributed across various locations.

To answer challenging biomedical questions about diabetes, DZD must connect data from many different studies, reports and surveys. The organisation needs to incorporate research projects from multiple locations in Germany, including 500 researchers and 10 university hospitals.

The data covers various disciplines, from studies on a molecular level to pathway analyses and animal models. This encompasses data from clinical trials, multi-omics experiments and patient information. Clearly, it is increasingly necessary to integrate and link more and more data. Doing that will be the next step in biomedicine. In addition, the healthcare sector, which is increasingly turning away from general blockbuster drugs and moving to individualised, precision medicine or treatment, needs better visibility of data.

DZD decided it needed a better and more complete way of seeing all this data. Now, DZD uses graph technology in combination with techniques such as Artificial Intelligence (AI) to make connections that no-one else can. This innovative approach to working with large, complex datasets could play a vital role in helping in the prevention, discovery of new subtypes, early diagnosis and treatment of major illnesses.

DZD Head of Data and Knowledge Management, Dr Alexander Jarasch, and his team are using Neo4j's graph software to find connections across all its data, speeding up data analysis dramatically. DZD has built a 'master database' to consolidate data and provide its 500-strong team of scientist peers with a holistic view of available information.

Cross-disciplinary Integration

DZD's research focusses on diabetes, a metabolic disease. It is not enough for researchers to only look at metabolic data. They must take into account data from disciplines such as genomics or proteomics, through to environmental influences. In the human body, everything is connected to metabolic pathways. A gene encodes a protein that is active in a metabolic pathway and metabolises a metabolite, which in turn is able to regulate another gene.

Human metabolism is a network of thousands of components that are connected with each other, similar to a graph data model. That's why it's so important to be able to uncover these connections and to create a new layer of analysis on top of this data using graph technology.

DZD is using graph database software to mine deeper into its diabetes 'map' to seek out hidden connections and relationships, allowing its researchers to examine new avenues of research. It is also looking to build new data models and better compare animal and human data. In a graph representation, abnormalities, patterns or connections can be easily identified and interrogated.

DZD plans to roll out graph technology further, to target discoveries in other clinical studies. In future, data from diabetes research could be integrated with that of cancer and Alzheimer's research, for example, to discover possible connections. DZD is also looking to exploit the combination of machine learning with graph technology to identify new subtypes of diabetes.

Knowledge Graph to Fight COVID-19

Neo4j graph software is also at the heart of a new Knowledge Graph to help fight COVID-19. This open-source initiative of DZD data scientists, developers and data people from academia and industry connects data from a range of well-established public sources and links them in a searchable database. This is helping researchers and scientists throughout the world to navigate through the 128,000-plus publications on the disease and related disease areas such as SARS, plus over 32,000 relevant patents and 1700 clinical trials.

The Knowledge Graph allows researchers and scientists to create new hypotheses by querying not only clinical information but also data on a gene or protein, clinical trial, drug and drug targets. This is a critical capability in the absence of long-term clinical trials and with the minimal peer-reviewed research available in the current pandemic.

Researchers tend to have only a passing awareness of research outside their field. Yet life sciences companies face the challenge of distilling the findings of many papers across multiple disciplines and assimilating all that information, to create effective COVID regimes and develop a vaccine as quickly as possible.

Normally, it would be necessary to carry out searches on the patent database, the publication database and the gene database, and then make the connections manually. Researchers create Excel sheets, a list of identifiers, and then they go to the database and type in these identifiers to get further information. This yields limited results because of the lack of connections. It is very manual work, error-prone, extremely inefficient and slow. Indirect connections may be missed.

In contrast, the COVIDgraph.org is a graph database that allows researchers to structure this data and to connect it to fundamental things from biology, including genes, the proteins and their functions. A clinical trials database allows researchers to understand what kind of COVID-19 clinical trials are out there. The dataset specifies typical inclusion criteria, such as people under a certain age, or a specific risk group, like diabetic patients. This is valuable information that is usually scattered across different databases.

ACE2 Breakthrough

An early breakthrough has been around ACE2, the host cell receptor that mediates infection by SARS-CoV-2, the coronavirus responsible for COVID-19. The assumption might be made that the receptor ACE2 is just active in lung tissue, because one of the most vulnerable groups to the virus is patients with lung disease. In fact, of 55 human tissues in DZD's database, the receptor is active in 53 of them. This means the ACE2 receptor can attack almost every tissue of the body. As a result, scientists quickly realised that any potential vaccine will need to be able to fight the virus in all of these different tissue areas.

Graph technology is playing a key role in surfacing details like these to support the race to find a COVID-19 vaccine – as well as take the life sciences sector to the next level in precision medicine, prevention and treatment of diabetes. Graph technology's ability to discover relationships between data is taking life sciences research to the next level and supporting researchers as they make discoveries that will have a major clinical impact.

Alicia Frame

Alicia Frame is currently the Lead Product Manager and Data Scientist at Neo4j, where she works on the company's product management team to set the roadmap and strategy for developing graph-based machine learning tools. She earned her PhD in computational biology from the University of North Carolina at Chapel Hill and a BS in biology and mathematics from the College of William and Mary in Virginia, and has over eight years of experience in enterprise data science at BenevolentAI, Dow AgroSciences, and the EPA.

Email: alicia.frame@neo4j.com