DeepFiltering

Leverage the World's Knowledge in your Analytics.

Cerenode Inc.

cerenode.io
@cerenodeio
info@cerenode.io

Introduction & Background

The ability to filter data in order to minimize noise and limit extraneous data points is fundamental to any effort in data analyses. However, in analytics today, data is mostly filtered through variables that are present in your data.

Consider a simpified example from drug discovery, where a user has a dataset with compounds and corresponding IC50 values from a drug/compound efficacy trial. As a first cut, to identify interesting compounds, you could filter this dataset based on the IC50 values present in the dataset (**Table 1**). However, in most cases, the dataset does not contain information on everything that you might need to make the decisions pertinent to your analysis. That information is probably present as part of another data source. Ideally, you would like to filter or slice your data based on all the information that is relevant to your analyses – irrespective of where the data is. This information could reside in experiments you did in your lab, public ontologies or even licenced databases. To gather relevant insights from an analyses, it is crucial that the data can be sliced by all relevant variables and not just those that are natively present in the dataset. **Box. 1** outlines some typical question users ask around their data.

S No.	Compound	IC-50 (uM)
1	C1	0.8 (±0.1)
2	C2	0.6 (±0.1)
3	C3	1.7 (±0.2)
4	C4	1.4 (±0.7)
5	C5	1.3 (±0.5)
6	C6	1.4 (±0.4)
7	C7	2.2 (±0.9)
8	C8	1.3 (±0.4)
9	C9	1.3 (±0.1)
10	C10	1.1 (±0.2

Table 1: Example compound toxicity data (IC-50) from a screen.

Box 1

Q1. *Which compounds in the dataset are relevant to a specific set of pathways or genetic variants?*

Q2. *How do the IC50 values fare for compounds that have been tested against certain disease types?*

Q3. *How do the IC50 values fare for compounds that are approved drugs and sold by the Top 10 pharma?*

It is clear that most pieces of information that are required to answer real-world questions are not present natively in the dataset. So how does one answer questions that require users to leverage facts and knowledge gathered elsewhere? This is where Cerenode's DeepFiltering technology comes in.

DeepFiltering

Cerenode has built a vast network of facts – a knowledge graph. In this network, every piece of information is represented as a node and relationships they share are represented as connections (or edges) between the nodes. This knowledge graph can be described as a map of how everything relates to each other in biology and medicine (**Figure 1**). The knowledge graph already indexes hundreds of publicly available data sources including ontologies and taxonomies such as SNOMED, ICD10, MeSH, GO, Orphanet, GARD and others. The knowledge graph is very versatile and allows

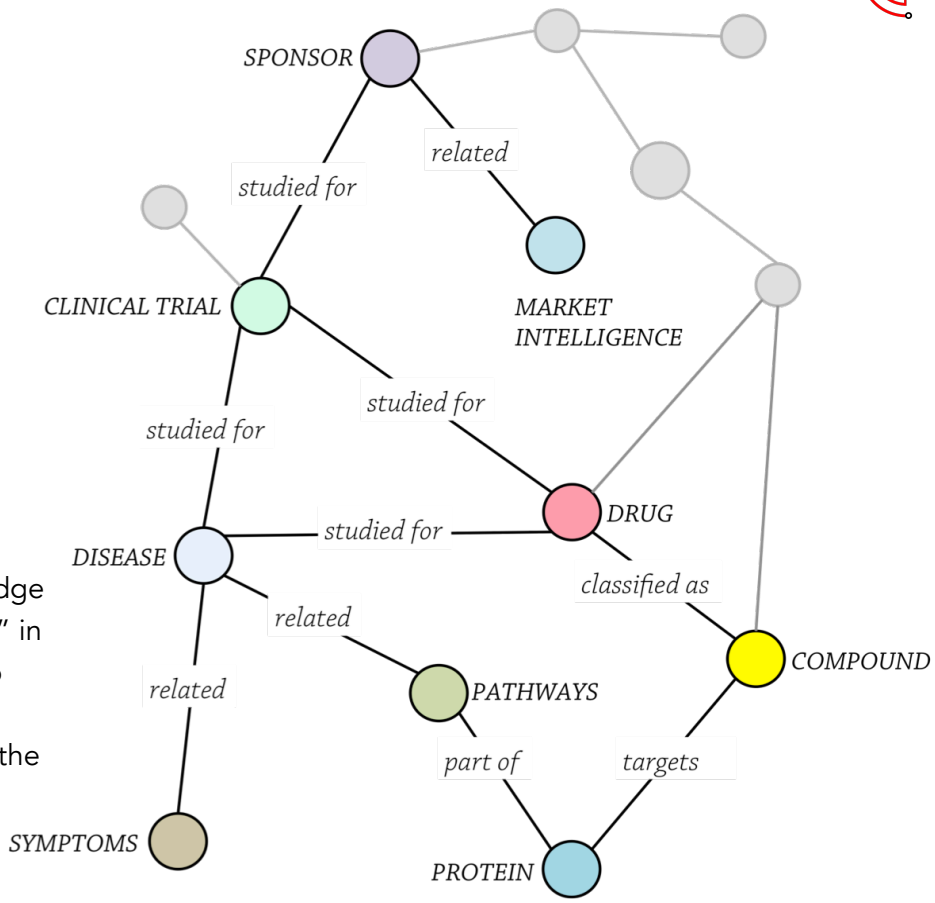


Figure 1: Cerenode's knowledge graph is a map of how "things" in biology and medicine relate to each other. Nodes represent "things" and edges represent the relationships.

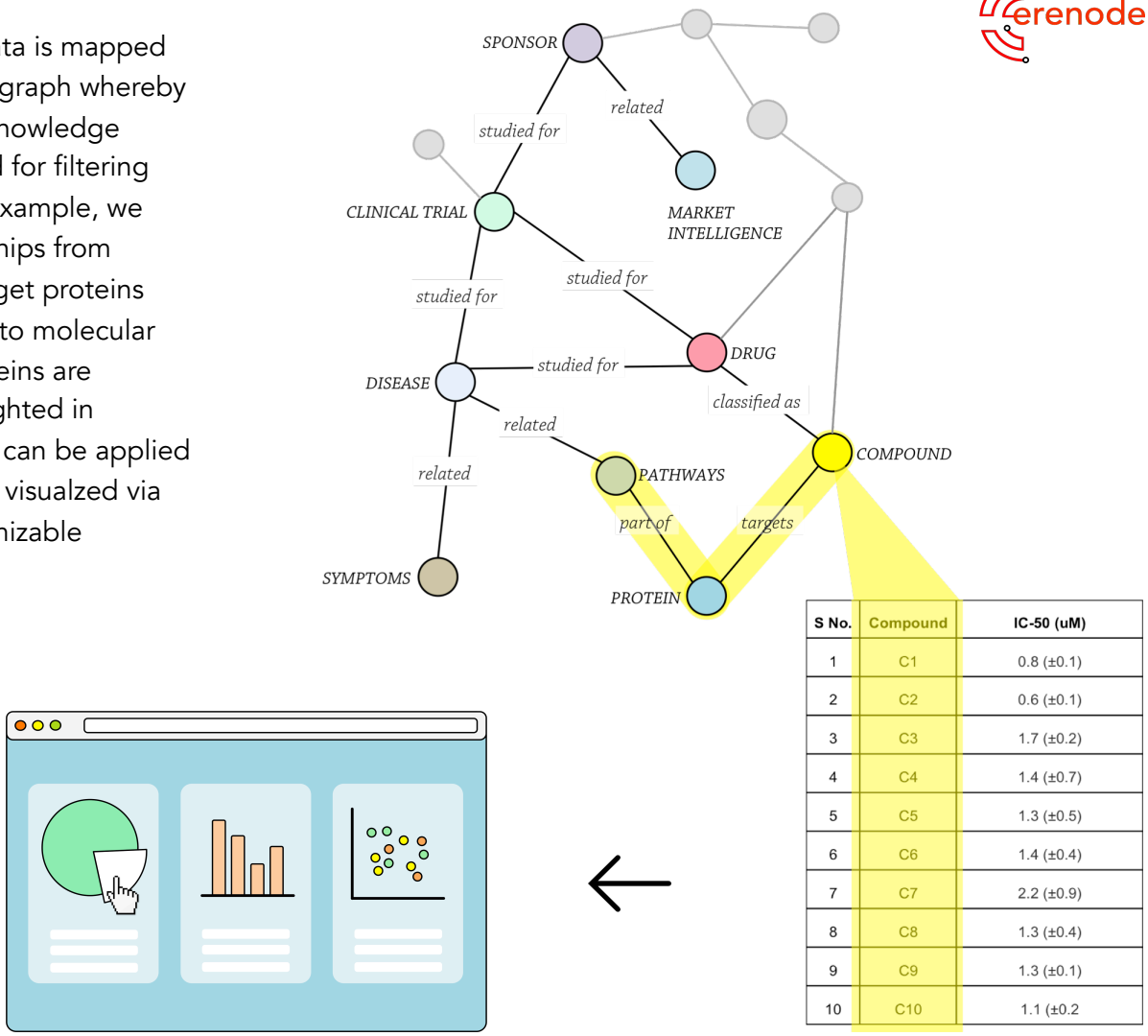
indexing of private data in a secure manner. Private data can include sources as diverse as ELM (Electronic Lab Managements systems), EHR (Electronic Health Records), R&D experiments and even spreadsheets. Once your data is indexed, Cerenode can leverage any fact in the knowledge graph seamlessly by tracing how nodes are connected. By simply connecting their data to any number of relevant nodes, users can start filtering their data. Effectively, this brings multiple dimensions to your analyses that are not natively present in your data. Users are able to slice and dice their data using any fact indexed in the knowledge graph. We call this DeepFiltering.

Data slices extracted using DeepFiltering can be directly visualized to identify interesting patterns and insights using Cerenode's customizable dashboards. Using Cerenode's powerful APIs, the advantage of DeepFiltering can be leveraged within an organization's analytics pipeline.

Box 2

*To answer questions in **Box 1**, DeepFiltering will look up proteins that are targeted by the compounds in the study and identify corresponding molecular pathways. Further, it can trace the genes and gene products involved in these pathways and also the genetic variations. Similarly, it can also trace relationships with disease classifications or custom classifications such as Top10 Pharma. Based on the traced relationships, DeepFiltering will segment your data to reveal insights or enrich it for further analyses. In this way, any external knowledge (facts or data points) can be leveraged – maximising the insights in user's data.*

Figure 2: Your data is mapped to the knowledge graph whereby the nodes in the knowledge graph can be used for filtering your data. In our example, we trace the relationships from compounds to target proteins and subsequently to molecular pathways the proteins are involved in (highlighted in yellow). The filters can be applied and results can be visualzed via Cerenode's customizable dashboards.



S No.	Compound	IC-50 (uM)
1	C1	0.8 (±0.1)
2	C2	0.6 (±0.1)
3	C3	1.7 (±0.2)
4	C4	1.4 (±0.7)
5	C5	1.3 (±0.5)
6	C6	1.4 (±0.4)
7	C7	2.2 (±0.9)
8	C8	1.3 (±0.4)
9	C9	1.3 (±0.1)
10	C10	1.1 (±0.2

Specific and enriched slices of data from DeepFiltering provide a powerful advantage to any analytics pipeline as all Statistics and Machine Learning methods depend wholly on the quality of data.

DEEPFILTERING IN HEALTHCARE

For the success of personalized medicine, the key is stratification of patients i.e. identifying the right treatment that can benefit specific segment of patients. Similarly, clinical decision support systems are key for efficient healthcare. The data needed for these purposes is largely locked in EHRs typically across many hospitals or care centres. Success largely depends on identifying relevant biomarkers and confounding factors that need to be accounted for while treating patients. Biomarkers and other indicators can be found in diagnostic reports, genomic reports, recorded disease phenotypes, demographic data and many other sources.

DeepFiltering can pull in data from all these sources and compliment them with medical ontologies, taxonomies and public data. DeepFiltering can slice and dice your EHR data using all these sources of knowledge. The data slices can be visualized or subjected to further analyses to identify confounding factors, patterns and outliers in the data. The key advantage of DeepFiltering in this context is its ability to pull in any source of information to filter the EHR data – resulting in a flexible and agile environment for EHR data analyses.