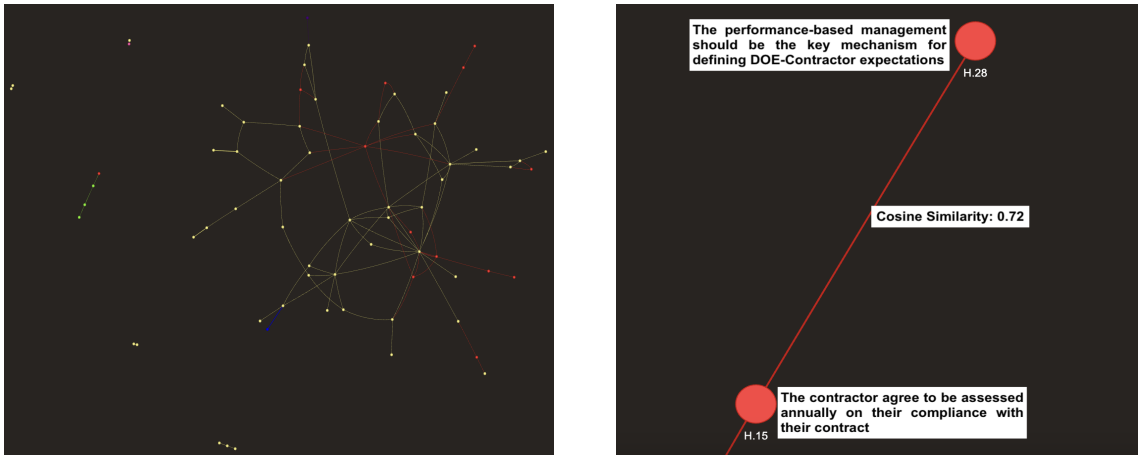
Data Science Clinic – Argonne National Laboratory

Argonne has thousands of policies and procedures and lacks an efficient way to identify or communicate requirements when they need to be updated. For instance, reviewing Argonne's policies related to performance standards is a tedious task as the relevant policies are spread across several sections identifiable by only a few experts. An interactive model of Argonne's policies was built to help improve operational efficiency, using machine learning to model the relationships between documents and visually depict them in a knowledge graph.

Firstly, the text data was pre-processed in order to prepare the corpus in a way that would optimize the model. Any overly specific text that does not add contextual value (eg, dates, addresses, or dollar amounts) was tokenized and replaced with a noun representing it. Clause numbers that reference other parts of the corpus were also tokenized and substituted out for the actual referenced text itself, which gives the model more semantic meaning. Finally, the preceding "parent" header text was inserted into the start of the current clause to provide further context for each document.

Embedding models were used to understand the relationship between documents. The text data was converted into vectors and cosine similarity was applied to these vectors to produce a score for each pair of documents ranging from -1 to 1. Several models were trained and fine-tuned using domain-specific legal data. The model which was found to have the most potential was fine-tuned using self-labeled pairs of Argonne documents as an additional layer of data on top of the pre-trained model. If more self-labeled data is generated, then a model fine-tuned on a subset of documents could be applied to understand the relationship between all Argonne policies.



A pipeline was created such that any of the models can be used to generate a knowledge graph (above left). A zoomed-in snapshot of the knowledge graph (above right) depicts documents as nodes colored according to the clause or section. Each edge indicates the connection between two documents, including the cosine similarity score. This knowledge graph provides the foundation for Argonne to understand the dependencies and interactions of the policies and procedures. For example, if the DOE changes their "performance-based" structure, which is discussed in H.28, Argonne could now see which other documents might also be affected by this change.