Office for PhD Education

Popular science summary of the PhD thesis

PhD student	Laurent Jan Vermue
Title of the PhD thesis	Data Integration for Industrial Big Data Applications
PhD school/Department	DTU Compute

Science summary

* Please give a short popular summary in Danish or English (approximately half a page) suited for the publication of the title, main content, results and innovations of the PhD thesis also including prospective utilizations hereof. The summary should be written for the general public interested in science and technology:

In modern applications, there are found several operational data storage systems and large amounts of heterogeneous data that is being collected, both in business and science contexts. At the same time, the data generation is in general error prone, meaning that the data entry process always will produce dirty data to some extent, either caused by human or system failure. Data integration comprises the task of cleaning dirty data and reconciling the different data sources into one homogeneous data set, which is a crucial step on the way to developing big data applications, such as machine learning models that rely on a vast amount of data. However, the pace of data creation has by far exceeded the capability of current data integration approaches, as these often rely on domain experts. As a consequence, a large fraction of valuable data is not utilized for analysis and thus leaves unused potential in every business field, which needs to be addressed.

A possible solution is to build algorithms that learn to analyze and evaluate data similar to domain experts. These algorithms could then guide data scientists in making the right data integration decisions or even make those decisions autonomously. Over time these algorithms could aggregate a large amount of knowledge about the data within a company which by far exceeds the capability of an individual human domain expert.

In this project, relational machine learning methods are developed and investigated in an effort to match the capabilities of domain experts to understand complex relational knowledge. We primarily focus on knowledge graph embedding models that can be used to predict new knowledge based on relations between objects. A data integration framework is developed based on knowledge graph embedding models that is purely based on machine learning to solve the above-mentioned data integration challenges. Overall, such endeavors rely upon complex software, which this thesis provides as published open-source frameworks to foster future research in the covered research areas and beyond as well as applications that build on it.

Please email the summary to the PhD secretary at the department