Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

Enabling Theory-guided Data Science for Scientific Workflows

Theory-guided Data Science (TGDS) [1] is the concept of combining scientific knowledge and data science models. While standard machine learning methods learn exclusively from the provided data, TGDS intends to improve results by including existing insights from the problem domain. One possible approach is to utilize so-called theory-based models. This family of models represents cause-effect relationships between variables that were either proven empirically or deduced theoretically from first principles. This definition includes a whole set of methods, from simple closed-form expressions up to complex simulations of dynamic systems.

These simulations belong to the class of scientific applications that demand large amounts of processing power, produce similarly high volumes of data, and tend to require several days to run on standard hardware.

They also belong to the type of applications regularly run on high-performance computing infrastructure to gain insights faster on even larger scales. For such programs, it is a good practice to employ software that monitors their processes and orchestrates the different stages of its life cycle, such as data loading, data preparation, computation, and analysis. This is the purpose of scientific workflow management systems.

The goal of this thesis is to develop an application or add-on for a scientific workflow management system that enables the use of TGDS models. We are currently working with material scientists, and have done so for quite some time, in an effort to improve accuracy and efficiency of their research while also aiming for a reduction of computational effort. The focus of this thesis will be to integrate a variety of TGDS approaches around material science simulations which are provided. This variety of possible approaches in itself poses a challenge to be addressed. Further requirements regarding the solution envisioned are easy integration of existing simulations, clean and structured collection of any associated data for every simulation run and a simple user interface to start workflows, monitor their state and visualize results.

This results in the following tasks:

- Literature review focusing on TGDS and workflow applications.
- Proposal and implementation of a workflow application for several TGDS use cases.
- Iterating with researchers over feedback concerning their respective use cases.

To successfully conduct this thesis project, the student must possess knowledge of python or java and basic knowledge of web development.

To support you with this task, we offer thorough mentoring and weekly meetings with your advisor, access to our institute's computing infrastructure, code reviews and guidance to improve your software development skills.

Throughout this work, you will acquire knowledge about various theory-guided data science techniques, thorough practical experience with workflow technology and hands-on experience with interdisciplinary work.

[1] A. Karpatne et al. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data". en. In: *IEEE Transactions on Knowledge and Data Engineering* 29.10 (Oct. 2017), pp. 2318–2331. arXiv: 1612.08544.

Ansprechpartner

Daniel Betsche, M.Sc. daniel.betsche@kit.edu Raum: 364

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34

Master Thesis