PhoenixD
Photonics · Optics · Engineering
Innovation Across Disciplines

Leibniz Universität Hannover

Invitation to the PhoenixD Colloquium

Monday, November 1st, 2021, 10.00 – 12.00 am
Room M11, Building 1104

"Semantic Research Data Management in the National Research Data Initiative (NFDI)"

Prof. Dr. Sören Auer

Technische Informationsbibliothek (TIB)

In this talk we will give an overview on the concepts and implementation of semantic Research Data Management for the National Research Data Initiative (NFDI). We will introduce vocabularies and ontologies for establishing a common understanding of research data and showcase their use in the context of the NFDI initiatives NFDI4Ing, NFDI4Chem and NFDI4DataScience. We give an overview on three open technology components, ready to be used in PhoenixD:

- Terminology service for the collaborative creation of terminologies, vocabularies and ontologies: https://service.tib.eu/ts4tib/index
- Open Research Knowledge Graph (ORKG) for organising scientific contributions in a knowledge graph: https://www.orkg.org
- Leibniz Data Manager as a meta-data repository for research data: https://labs.tib.eu/info/projekt/leibniz-data-manager

---

„Democratizing Data Science through Example-Driven Data Preparation"

Prof. Dr. Ziawasch Abedjan

FG Datenbanken und Informationssysteme, LUH

Data scientists spend about 80 % of their time preparing data. Data preparation encompasses various tasks including discovery, extraction, transformation, and cleaning.
Most of these tasks need heavy user supervision in the form of predefined configurations, such as rules, parameters, or patterns. Defining these type of configurations makes most data preparation tools hard to use and less accessible. Machine learning techniques provide the opportunity to learn the configurations and ultimately the preparation task itself. However to define data preparation as a machine learning task, a model is required that can generalize from a small training dataset to very large data. In this talk, I will surface the state-of-the-art in learning-based tools to support the data scientists with a special focus on data cleaning. I will discuss how user supervision can be reduced to a handful of example corrections using effective feature representation, label propagation, and transfer learning methods. Our cleaning systems internally leverage an automatically generatable set of base detectors and correctors and learn to combine them for highest utility. In practice, with a small number of 20 user-annotated tuples, our systems outperformed state-of-the-art techniques.