HEIBRIDS
Helmholtz Einstein International
Berlin Research School in Data Science

Admission Round 2019

Project Title	**Optimization of Data Science Processes**
Project leads / supervisors	Prof. Dr. Volker Markl (TU Berlin, ECDF) Prof. Dr. Uwe Ohler (MDC, Helmholtz)
Project description	Applying data science methods typically involves a tedious, iterative process of specifying and executing complex data analysis pipelines. These pipelines comprise of preprocessing steps, model building, and performance evaluation. Heterogeneous data sources and systems for pipeline execution often introduce complex dependencies on input data and processing infrastructure. For instance, when training a neural network, data transformations and preprocessing steps may be carried out with custom scripts or on a scalable dataflow engine such as Spark or Flink, while the actual training may be executed on specialized systems such as TensorFlow or MXNet, potentially on custom designed hardware. In order to simplify and accelerate this tedious process of data analysis, it would be highly beneficial to enable the declarative specification of such pipelines. This lets end users specify *what* they want a computer system to compute, but leave the decisions on *how* to efficiently compute the result to the executing system. This principle brings opportunities for workflow optimization and is the basis of many successful technologies, e.g., execution of database queries specified in SQL or the training of neural networks specified as computational graphs. There are existing data analysis pipeline abstractions such as scikit-learn pipelines, SparkML pipelines, or the KeystoneML project. However, many of these systems lack a holistic declarative specification of a data science workflow. They are not designed to incorporate a schema of the processed data and do not support different environments for execution. Another important lack of functionality is support for tedious, orthogonal tasks like tracking the metadata of the pipeline (e.g., the hyperparameters of ML models), checking for common data errors (missing values, wrong data types), or recording data lineage (e.g., the datasets used for training and evaluation). In order to overcome these deficiencies and challenges, we propose the following research directions: • Design of a holistic declarative specification for data science pipelines, which addresses the aforementioned requirements of declarativity, support for different execution environments, automatic data validation and recording of metadata. • Implementation of a system for the optimized execution of pipelines expressed in the declarative specification, with support for different runtimes, e.g. translation to a mixed Spark/tensorflow workload with

	experiment tracking enabled or translation to transactions inside a database with an ML extension • Utilization of an experiment database to automatically suggest tests for potential data errors in the pipeline e.g., wrong data types, missing normalization of the data) Complex data processing pipelines arise virtually all application domains of data science. In this work, we focus on the end-to-end management of machine learning tasks involving high-throughput molecular data. Tackling the problems outlined above with database systems research inspired optimization and automation techniques represents a highly interdisciplinary endeavor: It requires understanding the requirements and challenges of a data science problem in one or more domains in order to generalize, abstract, and implement a pipeline abstraction and potential optimizations. In turn, the application domain benefits from simplified programming and accelerated experimentation.
References	[AKK15] A. Alexandrov, A. Kunft, A. Katsifodimos, Felix S., L. Thamsen, O. Kao, T. Herb, V Markl, Implicit Parallelism through Deep Language Embedding, SIGMOD Conference 2015 [AMS16] Abiteboul, Serge; Miklau, Gerome ; Stoyanovich, Julia ; Weikum, Gerhard: Data, Responsibly (Dagstuhl Seminar 16291). Dagstuhl Reports 6(7): 42-71 (2016) [BCF17] D. Baylor, E. Breck, H. Cheng, N. Fiedel, et al., The Anatomy of a Production-Scale Continuously-Training Machine Learning Platform, KDD, 2017 [SVK17] E. R. Sparks, S. Venkataraman, T. Kaftan, M. J. Franklin, B. Recht, Keystoneml: Optimizing pipelines for large-scale advanced analytics, ICDE 2017 [VBP12] J. Vanschoren, H. Blockeel, B. Pfahringer, G. Holmes, Experiment databases: A new way to share, organize and learn from experiments, Machine Learning 87.2 (2012): 127-158, 2012 [GWCV+17] Gönen M, Weir BA, Cowley GS, et al. A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of Cancer Cell Lines. Cell Syst, in press. [DWBB+17] Daneshjou R, Wang Y, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat. 2017 38(9) [GFYW+17] Grüning BA, Fallmann J, Yusuf D, Will S, et al. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Res, in press [CMWU16] Calviello L, Mukherjee N, Wyler E, Ohler U. Detecting actively translated open reading frames in ribosome profiling data. Nature Methods 2016. [KOD16] Kassuhn W, Ohler U, Drewe P. Cseq-simulator: A data simulator for clip-seq experiments. Pac Symp Biocomput. 21:433-44, 2016.