Provenance in Databases and Scientific Workflows

Bertram Ludäscher[1]

[1]Department of Computer Science
University of Illinois
Champaign – IL – USA

ludaesch@illinois.edu

In computer science, data provenance describes the lineage and processing history of data as it is transformed through queries or workflows. Different computer science subdisciplines have studied approaches to capture and exploit provenance, e.g., the systems and programming languages communities. In this tutorial, I will give an overview of basic research questions and results provided by the database and scientific workflow communities. Research in this area ranges from technical studies in database theory (e.g., the use of semi-ring structures to abstract and unify different types of provenance) to more applied techniques (e.g., to efficiently record, store, and query provenance), and various engineering-level questions in-between. Provenance capture and querying capabilities are also playing an increasing role in the reproducibility of scientific workflows, data science applications, the computational sciences. The first half of the tutorial will cover the different uses and types of provenance in scientific workflows, e.g., prospective vs retrospective provenance, and introduce tools that can work with or even combine both forms of provenance to support advanced uses of provenance. In the second half of the tutorial, different notions of provenance in databases will be discussed such as Why-, How-, and Why-Not (missing-answer) provenance. Provenance is a very active research area, and I will end by highlighting some questions and opportunities for future work in databases and workflows.

About the Autor

Bertram Ludäscher is a professor at the School of Information Sciences (iSchool) at the University of Illinois, Urbana-Champaign. At the iSchool, he directs the Center for Informatics Research in Science and Scholarship (CIRSS). He is also a faculty affiliate with the National Center for Supercomputing Applications (NCSA) and the Department of Computer Science. From 2005 to 2014 he was a computer science faculty at the University of California, Davis. His research interests range from scientific data and workflow management, to knowledge representation and reasoning. His current focus includes foundations of provenance and applications, e.g., for automated data quality control and data curation. Until 2004 Ludäscher was a research scientist at the San Diego Supercomputer Center (SDSC) and an adjunct faculty at the CSE Department at UC San Diego. He received his M.S. (Dipl.-Inform.) in computer science from the Technical University of Karlsruhe (K.I.T.) and his PhD from the University of Freiburg, Germany.