Title : Utilizing SMT-Based Data-Integrity Constraints to Estimate Data-Quality and Compliance

Authors : Eric W.D. Rozier (Iowa State University of Science and Technology, U.S.A.)

Speaker : Eric W.D. Rozier

Abstract: Increased reliance on machine learning and automated reasoning has led to increased vulnerability to data-integrity violations. In order to ensure next-generation data-driven systems and infrastructure are reliable and provide trustworthy services better mechanisms must be developed to reason about whether the data being utilized is compliant, and has integrity. In this paper we discuss the concept of data integrity faults, their potential impact to data-driven reasoning, and introduce the use of SMT-based reasoning about integrity and compliance. We demonstrate this novel approach using real data on nutrition information, providing examples of real data-integrity faults in the USDA’s National Nutrient Data Base for Standard Reference Release 28, and in crowd sourced data collected from MyFitnessPal for an ongoing patient study about the relation of blood glucose and exercise and eating habits. We demonstrate high rates of data-integrity faults in crowd sourced data, with nearly 27% of our data failing one or more SMT-based constraints. Similarly, in federally published data we find nearly 10% of data published by the USDA is non-compliant, and features data-integrity faults. We discuss these results, and the need for more formal checks to help safe-guard machine-learning and automated reasoning from data with low-integrity.