Getting Rid of Data

Tova Milo

Tel Aviv University
School of Computer Science
Tel Aviv, Israel
Email: `milo@cs.tau.ac.il`

Abstract. We are experiencing an amazing data-centered revolution. Incredible amounts of data are collected, integrated and analyzed, leading to key breakthroughs in science and society. This well of knowledge, however, is at a great risk if we do not dispense with some of the data flood. First, the amount of generated data grows exponentially and already at 2025 is expected to be more than five times the available storage. Second, even disregarding storage constraints, uncontrolled data retention risks privacy and security, as recognized, e.g., by the recent EU Data Protection reform. Data disposal policies must be developed to benefit and protect organizations and individuals. Retaining the knowledge hidden in the data while respecting storage, processing and regulatory constraints is a great challenge. The difficulty stems from the distinct, intricate requirements entailed by each type of constraint, the scale and velocity of data and the constantly evolving needs. While multiple data sketching, summarization and deletion techniques were developed to address specific aspects of the problem, we are still very far from a comprehensive solution. Every organization has to battle the same tough challenges, with ad hoc solutions that are application specific and rarely sharable. In this talk I will discuss the logical, algorithmic, and methodological foundations required for the systematic disposal of large-scale data, for constraints enforcement and for the development of applications over the retained information. I will overview relevant related work, highlighting new research challenges and potential reuse of existing techniques, as well as the research performed in this direction in the Tel Aviv Databases group.