Several classes of applications have recently emerged that require database systems to manage and process *probabilistic data*. For example, large-scale data integration systems have to cope with several issues of data quality that can often be expressed naturally using probabilistic information; data acquisition systems such as sensornet or RFID deployments can produce huge volumes of raw data that is often error-prone and is best described through a probabilistic model; data extracted automatically from large text corpora is often imprecise and best modeled probabilistically. There have been attempts in the past to extend database management systems with the capability of storing and processing probabilistic data, but the technical challenges involved are enormous and remain, for the most part, unaddressed.

The collection of eight articles in this special issue represents an interesting sample of different research perspectives, efforts, and novel applications in the vibrant area of large-scale *probabilistic data management*. The first four papers in the issue discuss novel algorithmic and systems aspects of probabilistic databases. The paper by Benjelloun et al. presents a prototype probabilistic database system developed at Stanford University that adds both data uncertainty and data lineage as first-class concepts. The authors give a thorough discussion of various facets of the system (e.g., data model, query language, lineage management) in a unified framework, and describe their current implementation effort. The paper by Green and Tannen offers a concise, yet very thorough, comparison of the various proposed models for both *incomplete* and *probabilistic* databases, thus providing an excellent entry point to the rich literature on both topics. In the third article, Ré et al. discuss the evaluation of SQL queries over probabilistic databases, and give interesting theoretical results demonstrating the boundary between "easy" and "hard" queries. In a nutshell, they show that while some queries can be evaluated by simple manipulation of probabilities (a process that can be easily pushed inside relational database engines), others have high theoretical complexity making evaluation intractable. Unfortunately, much of the database research on probabilistic databases has evolved independently from the much richer area of *probabilistic models* in machine learning and knowledge representation. The paper by Getoor tries to fill this gap, by providing a very accessible discussion of probabilistic graphical models and their extensions to relational databases.

The final four articles in the issue discuss four novel application domains and the associated requirements on a probabilistic data management system. Jayram et al. describe **AVATAR**, a high-precision, rule-based system for extracting information from text documents developed at IBM Almaden, that relies on probabilistic data management techniques. Choudhury et al. give a thorough survey of recent work done at the Intel Research Seattle lab on automatic human activity recognition from sensor data, and discuss desirable features for a database system that can support such applications. Garofalakis et al. discuss the novel probabilistic data management challenges faced by the *Data Furnace* project that aims to provide database support for pervasive computing (e.g., "smart home") environments. Finally, the article by Doan et al. surveys the authors' recent efforts on building a new data management platform that allows effective information sharing across web communities, arguing convincingly for probabilistic database tools.

We believe that you will find this diverse collection of research papers on probabilistic data management stimulating and thought-provoking, and hope that this special issue will spark even more interest in this exciting new area. Our sincerest thanks to all the authors for their contributions.