Now that we have gathered so much data, what do we do with it? This question has become common in many organizations. The culprit is the digital revolution; digitized information is easy to capture and fairly inexpensive to store. But why do people store so much data?

Besides the fact that it is easy and convenient to do so, people store data because they think some valuable assets are implicitly coded within it. In scientific endeavors, data represents observations carefully collected about some phenomenon under study. In business, data captures information about critical markets, competitors, and customers. In manufacturing, data captures performance and optimization opportunities, as well as the keys to improving processes and troubleshooting problems.
“I never waste memory on things that can easily be stored and retrieved from elsewhere.”

—Albert Einstein
1879–1955
Raw data is rarely of direct benefit. Its true value is predicated on the ability to extract information useful for decision support or exploration and understanding the phenomena governing the data source. Traditionally, analysis was strictly a manual process. One or more analysts would become intimately familiar with the data and—with the help of statistical techniques—provide summaries and generate reports. In effect, the analysts acted as sophisticated query processors. However, such an approach rapidly breaks down as the quantity of data grows and the number of dimensions increases. Who could be expected to “understand” millions of cases, each having hundreds of fields? Further complicating this situation, the amount of data is growing so fast that manual analysis (even if possible) simply cannot keep pace.

A community of researchers and practitioners interested in the problem of automating data analysis has grown steadily under the label knowledge discovery in databases (KDD) and data mining. The first KDD workshop was held in 1989; it has evolved into an annual international conference, most recently “KDD-96: The Second International Conference on Knowledge Discovery and Data Mining,” which attracted more than 500 attendees. Our intent in assembling this special section is to give an overall KDD field through articles introducing and defining its constituent areas, including the perspectives of the core fields of statistics and databases, as well as a representative set of applications and challenges. The article by Fayyad, Piatetsky-Shapiro, and Smyth defines and explicates our process-centric view of the field and outlines challenges as yet unmet.

Statistics is at the heart of the problem of inference from data. Through both hypothesis validation and exploratory data analysis, statistical techniques are of fundamental importance. The article by Glymour, Madigan, Pregibon, and Smyth gives a statistical perspective, identifying a wealth of statistical results KDD can benefit from as well as some caveats and directions to consider.

In statistics, pattern recognition, and artificial intelligence (machine learning), algorithms are based on the assumption that data can be loaded into a computer’s main memory. A wealth of interesting issues arise when the data is too large to fit in main memory. A perspective from databases, a field fundamental to KDD, is provided by Imielski and Mannila, who identify challenges posed by KDD for database technology and postulate a new direction and view for both.

Quality of data is critical in data analysis. Inmon outlines the importance of and need for a data warehousing step in the KDD process. A representative set of industrial applications is described by Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, and Simoudis, who outline how KDD influences the way companies do business and the challenges to their practical use in KDD applications.

Issues specific to KDD applications in scientific data analysis are elaborated with illustrative examples by Fayyad, Haussler, and Stolorz, who make the case that by using KDD to analyze massive datasets, scientists are free to focus on tasks for which machines are poorly suited, namely, creative data analysis, theory and hypothesis formation, and deriving insights into underlying phenomena.

Etzioni explores the challenges and opportunities presented in discovering useful knowledge in the vast resources of the Internet, concluding that effective Web mining is feasible in practice.

Work addressing the core problems in KDD is ongoing; most problems of representation, search complexity, and use of prior knowledge to help search and statistical inference remain open and require serious attention. Nevertheless, successful applications continue to appear, driven mainly by the glut of database content that has clearly surpassed raw human processing abilities. Driving the growth in the field are the strong economic and social forces resulting from the data overload phenomenon most readers are familiar with.

At this stage of its development, the KDD field shows promising signs of yielding significant payoffs. More importantly, however, we see the beginnings of a new science and the foundations of what we hope will become a theory for efficient inference from massive datasets. We hope these articles foster a proper understanding of the objectives, promises, and challenges of this young and exciting field. Most of all, we hope they enable you to see past the inevitable hype toward more realistic expectations for the promise of KDD and data mining.

USAMA FAYYAD is a senior researcher at Microsoft Research and a distinguished visiting scientist at the Jet Propulsion Laboratory of the California Institute of Technology. He can be reached at fayyad@microsoft.com.

RAMASAMY UTHURUSAMY is manager of knowledge and decision support at General Motors Corp. He can be reached at samy@gmr.com.