Foreword

Recent development in computer technology has significantly advanced the generation and consumption of data in our daily life. As a consequence, challenges, such as the growing data warehouses, the needs of intelligent data analysis and the scalability for large or continuous data volumes, are now moving to the desktop of business managers, data experts or even end users. Knowledge discovery and data mining (KDD), grounded on established disciplines such as machine learning, artificial intelligence and statistics, is dedicated to solving the challenges by exploring useful information from a massive amount of data.

Although the objective of data mining is simple—discovering buried knowledge, it is the reality of the underlying real-world data which frequently imposes severe challenges to the mining tasks, where complications such as data modality, data quality, data accessibility, and data privacy often make existing tools invalid or difficult to apply. For example, when mining large data sets, we require mining algorithms to scale well; for applications where getting instances is expensive, the mining algorithms must manipulate precious small data sets; when data suffering from corruptions such as erroneous or missing values, it is desirable to enhance the underlying data before being mined; in situations, such as privacy preserving data mining and trustworthy data sharing, it is desirable to explicitly and intentionally add perturbations to the original data such that sensitive data values and data privacy can be preserved. For multimedia data such as images, audio and videos, data mining algorithms are severely challenged by the reality: finding knowledge from a huge and continuous volume of data items where the internal relationships among data items are yet to be found. When data are characterized by all/some of the above real-world complexities, traditional data mining techniques often work ineffectively, because the input to these algorithms is often assumed to confirm to strict assumptions, such as having a reasonable data volume, specific data distributions, no missing and few inconsistent or incorrect values. This creates the challenges between the real-world data and the available data mining solutions.

Motivated by these challenges, this book addresses data mining techniques and their implementations on the real-world data, such as human genetic and medical data, software engineering data, financial data and remote sensing data. One unique features of the book is that many contributors are the experts of their own areas, such as genetics, biostatistics, clinical research development, credit risk management, computer vision and applied computer science, and of course, traditional computer science and engineering. The diverse background of the authors renders this book a useful tool of overseeing real-world data mining challenges from different domains, not necessarily from the computer scientists and engineers' perspectives. The introduction of the data mining methods in all these areas will allow interested readers to start building their own models from the scratch, as well as resolve their own challenges in an effective way. In addition, the book will help data mining researchers to better understand the requirements of the real-world applications and motivate them to develop practical solutions.

I expect that this book will be a useful reference to academic scholars, data mining novices and experts, data analysts and business professionals, who may find the book interesting and profitable. I am confident that the book will be a resource for students, scientists and engineers interested in exploring the broader uses of data mining.

Philip S. Yu
IBM Thomas J. Watson Research Center

Philip S. Yu *received a BS in electrical engineering from National Taiwan University, MS and PhD in electrical engineering from Stanford University, and an MBA from New York University. He is currently the manager of the software tools and techniques group at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. His research interests include data mining, data stream processing, database systems, Internet applications and technologies, multimedia systems, parallel and distributed processing, and performance modeling. Dr. Yu has published more than 450 papers in refereed journals and conferences. He holds or has applied for more than 250 U.S. patents. Dr. Yu is a fellow of the ACM and the IEEE. He is associate editor of* ACM Transactions on the Internet Technology *and* ACM Transactions on Knowledge Discovery in Data*. He is a member of the IEEE Data Engineering steering committee and is also on the steering committee of IEEE Conference on Data Mining. He was the editor-in-chief of* IEEE Transactions on Knowledge and Data Engineering *(2001-2004), an editor, advisory board member and also a guest co-editor of the special issue on mining of databases. He had also served as an associate editor of* Knowledge and Information Systems*. In addition to serving as program committee member on various conferences, he will be serving as the general chair of 2006 ACM Conference on Information and Knowledge Management and the program chair of the 2006 joint conferences of the 8th IEEE Conference on E-Commerce Technology (CE '06) and the 3rd IEEE Conference on Enterprise Computing, E-Commerce and E-Services (EEE '06). He was the program chair or co-chairs of the 11th IEEE International Conference on Data Engineering, the 6th Pacific Area Conference on Knowledge Discovery and Data Mining, the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, the 2nd IEEE Intl. Workshop on Research Issues on Data Engineering: Transaction and Query Processing, the PAKDD Workshop on Knowledge Discovery from Advanced Databases and the 2nd IEEE International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems. He served as the general chair of the 14th IEEE International Conference on Data Engineering and the general co-chair of the 2nd IEEE International. Conference on Data Mining. He has received several IBM honors including two IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, two Research Division Awards and the 85th plateau of Invention Achievement Awards. He received an Research Contributions Award from IEEE International Conference on Data Mining (2003) and also an IEEE Region 1 Award for "promoting and perpetuating numerous new electrical engineering concepts" (1999). Dr. Yu is an IBM master inventor.*