This Is a Publication of
The American Association for Artificial Intelligence

KDD-93: Progress and Challenges in Knowledge Discovery in Databases

Gregory Piatetsky-Shapiro, Christopher Matheus, Padhraic Smyth, and Ramasamy Uthurusamy

■ Over 60 researchers from 10 countries took part in the Third Knowledge Discovery in Databases (KDD) Workshop, held during the Eleventh National Conference on Artificial Intelligence in Washington, D.C. A major trend evident at the workshop was the transition to applications in the core KDD area of discovery of relatively simple patterns in relational databases; the most successful applications are appearing in the areas of greatest need, where the databases are so large that manual analysis is impossible. Progress has been facilitated by the availability of commercial KDD tools for both generic discovery and domain-specific applications such as marketing. At the same time, progress has been slowed by problems such as lack of statistical rigor, overabundance of patterns, and poor integration. Besides applications, the main themes of this workshop were (1) the discovery of dependencies and models and (2) integrated and interactive KDD systems.

Interest in knowledge discovery in databases (KDD) continues to increase, driven by the rapid growth in the number and size of large databases and the application-driven demand to make sense of them. The research side of KDD is of growing interest to researchers in machine learning, statistics, intelligent databases, and knowledge acquisition, as evidenced by the number of recent workshops (Ziarko 1993; Zytkow 1992; Piatetsky-Shapiro 1991a, 1991b) and special journal issues (Cercone and Tsuchiya 1993; Zytkow 1993; Piatetsky-Shapiro 1992) devoted or closely related to discovery in databases. The application side is of interest to any business or organization with large databases. KDD applications have been reported in many areas of business, government, and science (Parsaye and Chignell 1993; Inmon and Osterfelt 1991; Piatetsky-Shapiro and Frawley 1991).

The notion of discovery in databases has been given various names, including knowledge extraction, data mining, database exploration, data pattern processing, data archaeology, information harvesting, siftware, and (when done poorly) data dredging. Whatever the name, the essence of KDD is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley, Piatetsky-Shapiro, and Matheus 1992). KDD encompasses a number of different technical approaches, such as clustering, data summarizing, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies (Matheus, Chan, and Piatetsky-Shapiro 1993).

The essence of KDD is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Over 60 researchers from 10 countries took part in the Third Knowledge Discovery in Databases Workshop (Piatetsky-Shapiro 1993), held during the Eleventh National Conference on Artificial Intelligence in Washington, D.C, in the simmering July heat. A major trend evident at the workshop was the transition to applications in the core KDD area of discovery of relatively simple patterns in relational databases; the most successful applications are appearing in the areas of greatest need, where the databases are so large that manual analysis is impossible. Progress has been facilitated by the availability of commercial KDD tools for both generic discovery and domain-specific applications such as marketing. At the same time, progress has been slowed by problems such as lack of statistical rigor, overabundance of patterns, and poor integration. Besides applications, the main themes of this workshop were (1) the discovery of dependencies and models and (2) integrated and interactive KDD systems.

Real-World Applications

The applications presented at the workshop fell into three broad application areas: scientific, financial, and manufacturing. Most of the systems performed some form of classification, but two systems dealt with detecting and describing changes. In addition to talks and poster presentations, several demonstrations of research and commercial discovery systems were given at the workshop.

Scientific Applications

Two applications were presented in the area of astronomy. Usama Fayyad (Jet Propulsion Laboratory [JPL]), started the workshop with a talk on the *sky image cataloging and analysis tool* (SKICAT), an automated system for analyzing large-scale sky surveys. The multiterabyte size of the database ruled out a manual approach to image classification. Using a number

of innovative machine-learning methods, Fayyad and his colleagues were able to recognize objects at least one magnitude fainter in resolution than was previously possible and achieve an accuracy of about 94 percent. This work is noteworthy as a real application of machine learning to a difficult problem with results that are being used by scientists on a daily basis. Padhraic Smyth, also from JPL, gave a related talk on the problem of image database exploration, describing collaborative work with Usama Fayyad. Smyth described challenging issues in image analysis such as the way to measure the right attributes, the role of prior knowledge, incremental learning, and the use of multisensor data. He also examined how these issues are handled in current JPL tasks, such as the analysis of Venus images obtained by the Magellan spacecraft.

Financial Applications

Two systems were presented for detecting and describing changes in large business databases. Tej Anand described A.C. Nielsen's recent work on a commercial product called OPPORTUNITY EXPLORER. This system is a redesign and extension of its SPOTLIGHT (Anand and Kahn 1992) product for identifying and reporting on trends and exceptional events in the extremely large supermarket-sales databases. An innovative feature of SPOTLIGHT is the automatic explanation of relationships between key events. OPPORTUNITY EXPLORER is a more general tool for developing interactive, hypertextual reports using knowledge discovery templates that convert a large data space into concise, interlinked information frames. It is marketed to help sales analysts and product managers of consumer packaged-goods companies develop better sales strategies.

Christopher Matheus and Gregory Piatetsky-Shapiro, both of GTE Laboratories, presented their key-findings reporter (KEFIR), a system for discovering and explaining key findings in large relational databases. Although the system's design is domain independent, the current focus is on trend and normative analysis of health-care information. KEFIR performs an automatic drill down on the data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to norms or previous values. It then identifies explanatory relationships between findings and generates a report using natural language templates and graphics. A prototype of KEFIR has been implemented in C and TCL with an embedded SQL interface.

Three other financial applications—in the areas of insurance, marketing, and stock market analysis—used classification methods. John Major (Travelers Insurance) analyzed the important problem of selecting the most interesting rules among those discovered in data. He presented a rule-refinement strategy that defined rule interestingness by rule accuracy, coverage, simplicity, novelty, and significance. His method gave preference to rules not dominated in these measures by other rules and removed those rules that were potentially redundant. In an application of the method to a tropical storm database, the system reduced 161 rules generated by IXL (a product of IntelligenceWare, Inc.) to the 10 most interesting ones that were meaningful to a meteorologist.

Wojtek Ziarko and Rob Golan (both of University of Regina) presented an application of the Reduct Systems DATALOGIC/R discovery tool to identify strong predictive rules in stock market data. Monthly data collected over a 10-year period were analyzed to identify dominant relationships among fluctuations of market indicators and stock prices. Evaluation of the results (including both precise and imprecise, strong and weak rules) by a domain expert revealed that the strong rules confirm expert's experiences, but weak rules were difficult to interpret. DATALOGIC/R, a commercially available tool that derives rules using the variable-precision rough set approach, was also demonstrated at the workshop.

Pierro Bonissone and Lisa Rau, both of General Electric Research and Development, presented preliminary results of applying decision trees and logistic regression to a database of accounting, customer, and sales information. Initial results that suggest emerging markets and provide feedback on sales performance are encouraging enough to warrant further pursuit of this work.

Manufacturing Applications

Two applications dealt with semiconductor manufacturing and software engineering. Sharad Saxena of Texas Instruments presented his approach to fault isolation during semiconductor manufacturing using automated discovery from wafer-tracking databases. These databases contain the history of the semiconductor wafers as they undergo various processing steps. A generate-and-test approach is taken for using such databases for automated diagnosis. Based on prior manual analysis of such databases, classes of queries to the database as well as patterns in the responses to these queries that are useful for fault isolation are identified. Diagnosis is accomplished by automating the query generation and the detection of potentially useful patterns. A prototype system was implemented and tested on real data, finding both known and previously unknown faults.

Inderpal Bhandari of IBM presented *attribute focusing*, a method for exploratory analysis of attribute-valued data intended for use by domain experts who do not have a background in data analysis. The approach uses a model of interestingness based on the magnitude of data values, the association of data values, and basic knowledge of the limits of human information-processing capabilities as well as a model of interpretation to guide the domain specialist to discover knowledge from attribute-valued data. This approach has been used successfully by software managers, developers, and testers at IBM to make real-time improvements in their products as well as in their process of production. The attribute-focusing approach is being used in several IBM laboratories, with reported net savings of hundreds of person-days. A personal computer–based implementation of attribute focusing

was demonstrated by Bhandari and Michael Herman, also of IBM.

Discovery of Dependencies and Models

The second major theme of the workshop was discovery of dependencies and models. The workshop provided clear evidence of the diversity of technical approaches that are being applied to the general KDD problem. The focus was on the use of particular mathematical and statistical methods for the induction of qualitative relationships directly from data.

Jan Zytkow outlined the latest developments in his joint research with Robert Zembowicz (both of Wichita State University) on deriving equations from data. He proposed a computationally simple test for the absence of functional dependency that can eliminate the much more expensive search to determine the form of dependency. The test relies on search for the breakdown of data into optimal intervals. Initial experiments with their 49ER system showed that the test significantly reduces the computation time but loses only a few actual equations, typically those with a particularly poor fit with the data.

Dependency networks are an important form of discovered knowledge, and recent progress in this field (Spirtes, Glymour, and Scheines 1993; Pearl 1992) is encouraging for KDD. Probabilistic networks are a powerful knowledge representation medium, providing a bridge between the power of explicit knowledge representation in graphic form and more subtle (but robust) quantitative statistical methods. Greg Cooper (University of Pittsburgh) presented the latest results of his research on the use of Bayesian statistical methods for learning causal probabilistic network models that contain hidden variables. In earlier work, Cooper and Herskovits (1992) demonstrated that networks with hidden variables can directly be inferred from data. In this talk, he showed how to structure the calculations to dramatically speed up the computation.

Cooper also summarized recent research progress relevant to the discovery of directed probabilistic networks from data: There is a greater understanding of what relationships can be captured from data by directed acyclic graphs (DAGs) and which DAGs are indistinguishable based only on data. New methods have been developed for the discovery of probabilistic networks with measured and possibly unmeasured (latent) variables; these methods have been applied to real data with promising results. Major improvements needed for applications to real databases include better computational (search) efficiency, integration of different methods (especially those dealing with discrete and continuous variables), and estimation of the confidence and the stability of the output.

Saŝo Dŝeroski (Jozef Stefan Institute) gave an invited overview of inductive logic-programming (ILP) methods for KDD. ILP is an important paradigm that goes beyond the typical attribute-value relations (which are the limit of what can be learned by most current machine-learning methods) to the more general language of first-order relations. The field has developed rapidly in recent years (Muggleton 1992) and now boasts relatively sophisticated algorithms and methods for handling a variety of problems, with great potential for KDD applications (Lavrac and Dzeroski 1993). Dzeroski outlined the motivation for ILP and proceeded in his talk from early work through more recent extensions to successful applications. He described a particularly successful experiment in the prediction of protein secondary structure, where not only was the ILP method better in terms of predictive accuracy than alternative published methods, but perhaps more significantly, it yielded new domain knowledge. Still, much work remains to be done in handling noisy probabilistic concepts, especially in dealing with large databases.

The workshop revealed that much work is in progress in the knowledge discovery area that promises to take us beyond the discovery of relatively simple representations such as conjunctive probabilistic rules or linear models. However, as one broadens the search space to allow for more expressive languages of knowledge representation, there comes an inevitable computational penalty in terms of the scaling complexity of the algorithms. The talks showed that although the underlying models might be different, for each class of models, steady progress is being made in whittling down impractical algorithms to practical ones by taking advantage of particular structural characteristics of the methods and the representation being used. We hope to see some of the presented techniques at future workshops as standard workhorses of successful applications.

Integrated and Interactive Systems

The third theme of the workshop was integrated and interactive systems. The two areas are closely related because multimethod, integrated discovery systems frequently rely on human expertise to select the next discovery method, and interactive systems frequently offer a choice of multiple discovery algorithms.

Ron Brachman (AT&T Bell Laboratories) started the session with a talk entitled "Integrated Support for Data Archaeology." *Data archaeology* is the skilled human task of interactive and iterative data segmentation and analysis. He presented a system, called IMACS, that supports a data archaeologist with a natural, object-oriented description of an application domain; a powerful query language; and a friendly user interface that supports interactive exploration. IMACS is built on CLASSIC, a formal knowledge representation system.

Willi Kloesgen (GMD, Germany) described rule-refinement and rule-optimization strategies in EXPLORA, an interactive system for the discovery of interesting patterns in databases. The number of patterns presented to the user is reduced by organizing the search hierarchically, beginning with the strongest, most general hypotheses. An additional refinement strategy selects the most interesting statements and eliminates the overlap-

ping findings. The efficiency of discovery is improved by inverting the record-oriented data structure and storing all values of the same variable together, which allows the efficient computation of aggregate measures. Different data subsets are represented as bit vectors, making the computation of logical combinations of conditions efficient. EXPLORA, a publicly available system that runs on a MACINTOSH, was demonstrated at the workshop.

Philip Chan (Columbia University) proposed metalearning as a general technique to integrate a number of distinct learning processes. He examined several techniques of learning arbiters that select among independently learned classifiers. Such strategies are especially suitable for massive amounts of data that main–memory-based learning algorithms cannot handle efficiently. Preliminary results are encouraging, showing that parallel learning by metalearning can achieve comparable prediction accuracy in less time and space than purely serial learning.

An important design issue discussed at the workshop was the use of an internal versus an external database. Both IMACS and EXPLORA use an internal database approach of preloading relevant parts of the data and transforming them into their internal and efficient format. This approach generally speeds up discovery for small- or medium-sized databases. However, it limits the system ability to work with large external databases. An external database approach, taken in discovery systems such as SKICAT, SPOTLIGHT, and KEFIR, is to build an interface, usually based on SQL, to a database management system. This approach has its difficulties, such as dealing with communication problems and having to fit the discovery requests into the Procrustean bed of SQL. Retrieval from an external database can take longer because in addition to a communication delay, the physical database organization can be suboptimal for discovery system requests. However, this approach allows handling large external databases that would not fit in memory and avoids duplicating the code for database management system operations such as joins or aggregations. We expect the coming advances in database technology, such as faster hardware, SQL servers, and forthcoming powerful SQL2 and SQL3 standards, to make the external database approach more attractive.

Other related issues were discussed at the summary session. Larry Kerschberg (George Mason University) observed that analysts frequently need to track hypotheses in multiple databases. He proposed a mediator agent between an analyst and different discovery algorithms on the one hand and multiple data and knowledge sources on the other hand. Metalearning might offer a way to develop such mediator agents.

Zytkow proposed an agenda for integration. The first part is integration of different forms of knowledge: contingency tables, rules, decision trees, and equations. Each form has different strengths, and a limited conversion is possible from one form to another. The second part is the integration of search in different spaces of new terms, equations, and rules. Such integration is required in a machine discovery system to match human flexibility in detecting patterns of different type. The multiple searches should globally be controlled and guided by a combination of data conditions, background knowledge, and user preferences.

Advances and Difficulties

The workshop and the ensuing discussion on the KDD NUGGETS e-mail list highlighted several difficulties in application development (to subscribe, send e-mail to kdd@gte.com).

Insufficient statistical awareness: Some KDD experiments are performed without sufficient awareness of statistical theory. The classical example of this problem is testing N independent patterns for deviation from the norm, each test having a significance of α. Then, $N\alpha$ patterns are likely to pass the test purely because of chance. Eliminating such random discoveries requires statistical controls, such as Bonferroni adjustments, which in this example means reducing the significance level for each test to α/N to assign the final discovery the significance of α. Other ways to eliminate chance discoveries include randomized testing procedures (Jensen 1991). At the summary session, Major estimated that only about half of the work presented at the workshop dealt adequately with this problem. Hopefully, raising this issue will increase proper statistical awareness.

Overabundance of patterns: As many pioneers of KDD have found, even with proper statistics, it is all too easy to find many statistically significant patterns that are either obvious, redundant, or useless. A common approach to reducing the number of obvious discoveries (such as only women have pregnancies) is to focus on changes because obvious patterns will not change. Redundant discoveries can be eliminated by using rule-refinement methods such as those presented by Major or Kloesgen or by using some findings to explain others. The more difficult task of separating the important patterns from the useless ones requires domain knowledge. A general heuristic here is that rules and patterns are important to the degree that they can lead to a useful action. This heuristic suggests a decision-theoretic framing of the problem of evaluating the usefulness of discovered patterns. The utility of a particular pattern should not be measured in isolation, but instead, it should be evaluated in the context of a set of possible actions.

Integration: Even if a perfect discovery system is built, it needs to be integrated with other existing hardware-software systems to be useful. As expert system developers discovered years earlier, usually only a small part of the deployed system is new technology; the rest is interfacing and system integration, mundane but critical steps in moving from prototype stage to deployment.

Privacy versus discovery: Discovery in social or business data can raise a number of legal, ethical, and privacy issues. In 1990, Lotus was planning to introduce a CD-ROM with data on more than 100 million American households. The stormy protest led to

the withdrawal of this product (Rosenberg 1992). Recent conferences on computers, freedom, and privacy have also increased awareness about issues of privacy and data ownership.

These difficulties are compensated for by a number of important advances in areas relevant to KDD. Here we list only a few:

Multistrategy systems: Several recent comparisons of different learning and discovery algorithms have showed that different methods are superior for different types of problems (Brodley 1993); no single method is best across a range of problems. As a result, there is a movement toward multistrategy learning methods, especially for classification, that apply a number of different methods to the same task and select rules from the best method. Multistrategy systems is an area of active research interest, with recent progress reported in Michalski and Tecuci (1993).

Large-scale databases: Because most learning algorithms cannot handle large data sets, it is usually necessary to reduce the size of the data on which learning is performed. One way is to eliminate irrelevant data using the data dependencies. This method has been shown (Almuallim and Dietterich 1991) to increase the performance of the classifier methods. Other methods rely on various forms of data sampling. Catlett (1991) used an intelligent sampling approach to make a sublinear algorithm for decision tree induction. His method has been used to efficiently learn decision trees from databases with hundreds of thousands of records.

Overall, the workshop reflected measurable progress in developing and deploying KDD applications.

Acknowledgments

The authors thank all the workshop participants, especially Inderpal Bhandari, Greg Cooper, Saso Dzeroski, Usama Fayyad, Willi Kloesgen, Ryszard Michalski, Evangelos Simoudis, and Jan Zytkow for their comments and contributions to this article. We thank Peter Patel-Schneider for his editorial guidance and rapid processing of this article.

References

Almuallim, H., and Dietterich, T. 1991. Learning with Many Irrelevant Features. In Proceedings of the Ninth National Conference on Artificial Intelligence, 547–552. Menlo Park, Calif.: American Association for Artificial Intelligence.

Anand, T., and Kahn, G. 1992. SPOTLIGHT: A Data-Explanation System. In Proceedings of the IEEE Conference on AI for Applications, 2–8. Washington, D.C.: IEEE Computer Society.

Brodley, C. 1993. Addressing the Selective Superiority Problem: Automatic Algorithm-Model Class Selection. In *Proceedings of the Tenth Machine Learning Conference*, 17–24. San Mateo, Calif.: Morgan Kaufmann.

Catlett, J. 1991. Megainduction: A Test Flight. In *Proceedings of the Eighth Machine Learning Conference*, 596–599. San Mateo, Calif.: Morgan Kaufmann.

Cercone, N., and Tsuchiya, M., eds. 1993. *IEEE Transactions on Knowledge and Data Engineering* (Special Issue on Learning and Discovery in Databases).

Cooper, G., and Herskovits, E. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347.

Frawley, W.; Piatetsky-Shapiro, G.; and Matheus, C. 1992. Knowledge Discovery in Databases: An Overview. *AI Magazine* 14(3): 57–70.

Inmon, W., and Osterfelt, S. 1991. *Understanding Data Pattern Processing: The Key to Competitive Advantage*. Wellesley, Mass.: QED.

Jensen, D. 1991. Knowledge Discovery through Induction with Randomization Testing. In Proceedings of the 1991 Knowledge Discovery in Databases Workshop, ed. G. Piatetsky-Shapiro, 148–159. Menlo Park, Calif.: American Association for Artificial Intelligence.

Lavrac, N., and Dzeroski, S. 1993. *Inductive Logic Programming: Techniques and Applications*. Chichester, U.K.: Ellis Horwood.

Matheus, C.; Chan, P.; and Piatetsky-Shapiro, G. 1993. Systems for Knowledge Discovery in Databases. *IEEE Transactions on Knowledge and Data Engineering* 5(6): 903–913.

Michalski, R. S., and Tecuci, G., eds. 1993. *Machine Learning: A Multistrategy Approach, Volume 4*. San Mateo, Calif.: Morgan Kaufmann.

Muggleton, S. 1992. *Inductive Logic Programming*. San Diego, Calif.: Academic.

Parsaye, K., and Chignell, M. 1993. *Intelligent Database Tools and Applications*. New York: Wiley.

Pearl, J. 1992. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* 2d ed. San Mateo, Calif.: Morgan Kaufmann.

Piatetsky-Shapiro, G., ed. 1993. Proceedings of Knowledge Discovery in Databases 1993: Papers from the AAAI Workshop, Technical Report WS-93-02. Menlo Park, Calif.: American Association for Artificial Intelligence.

Piatetsky-Shapiro, G., ed. 1992. *International Journal of Intelligent Systems* (Special Issue on Knowledge Discovery in Databases and Knowledge Bases) 7(7).

Piatetsky-Shapiro, G. 1991a. Knowledge Discovery in Real Databases: A Workshop Report on KDD-89. *AI Magazine* 11(5): 68–70.

Piatetsky-Shapiro, G. 1991b. Report on AAAI-91 Workshop on Knowledge Discovery in Databases. *IEEE Expert* 6(5): 74–76.

Piatetsky-Shapiro, G., and Frawley, W., eds. 1991. *Knowledge Discovery in Databases*. Menlo Park, Calif.: AAAI Press.

Rosenberg, M. 1992. Protecting Privacy, Inside Risks Column. *Communications of ACM* 35(4): 164.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Berlin: Springer-Verlag.

Ziarko, W. 1993. *Rough Sets, Fuzzy Sets, and Knowledge Discovery*. Berlin: Springer-Verlag.

Zytkow, J., ed. 1993. *Machine Learning (*Special Issue on Machine Discovery) 12(1–3).

Zytkow, J., ed. 1992. *Proceedings of the Machine Discovery Workshop*. Wichita, Kans.: Wichita State University.

Gregory Piatetsky-Shapiro is a principal member of the technical staff and a principal investigator for the Knowledge Discovery in Databases Project at GTE Laboratories, Waltham, Massachusetts. He works on developing tools for intelligent data analysis, with current focus on a system for finding and reporting key changes and deviations in health care data. Previously, he worked on intelligent interfaces to heterogeneous databases and the development of database systems for financial applications. Piatetsky-Shapiro received his Ph.D. in 1984 from New York University. He organized and chaired the 1989, 1991, and 1993 Knowledge Discovery in Databases (KDD) Workshops. He edited (with William J. Frawley) *Knowledge Discovery in Databases* (AAAI Press) as well as special issues on KDD of the *International*

Journal of Intelligent Systems and the *Journal of Intelligent Information Systems*. He moderates the KDD NUGGETS electronic news list.

Christopher J. Matheus is a senior member of the technical staff at GTE Laboratories. He completed his Ph.D. in computer science at the University of Illinois at Urbana-Champaign in 1989 with a dissertation on machine learning. He joined GTE Laboratories shortly thereafter and is now working on developing AI techniques for extracting meaningful information from large databases. His current research interests include knowledge discovery in databases, machine learning, and INTERNET discovery agents. Matheus is a former Thomas J. Watson Fellow and was a research associate for several years at Oak Ridge National Laboratories in the areas of telerobotics, expert systems, and artificial neural networks.

Padhraic Smyth received a Ph.D. in electrical engineering in 1988 from the California Institute of Technology. He joined the Jet Propulsion Laboratory (JPL) in 1988 and is currently a technical group leader and a principal investigator on projects involving the application of statistical pattern-recognition techniques. Smyth has received 10 certificates of appreciation for technical innovation from the National Aeronautics and Space Administration since 1988 and was awarded the Lew Allen Prize for Excellence at JPL in 1993. His research interests include statistical pattern recognition, decision theory, information theory, source coding, telecommunications, and the application of probability and statistics to problems in AI.

Ramasamy Uthurusamy is machine-learning project leader at General Motors (GM) Research Laboratories. His research interests include theoretical and computational issues in the application of AI to real-world problems. Currently, his research focuses on the area of machine learning and its applications. Prior to joining GM, he was with Exxon Production Research Company. He has taught at Purdue University and at the University of Idaho. He obtained his Ph.D. from Purdue, an M.S. in computer science from Washington State University, and a B.E. in electrical engineering from the University of Madras, India.