

# Data Mining Curriculum: A Proposal (Version 1.0)

Intensive Working Group of ACM SIGKDD Curriculum Committee:

Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke,  
Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, Wei Wang

April 30, 2006

## 1 Introduction

Recent tremendous technical advances in processing power, storage capacity, and inter-connectivity of computer technology is creating unprecedented quantities of digital data. *Data mining*, the science of extracting useful knowledge from such huge data repositories, has emerged as a young and interdisciplinary field in computer science. Data mining techniques have been widely applied to problems in industry, science, engineering and government, and it is widely believed that data mining will have profound impact on our society. The growing consensus that data mining can bring real value has led to an explosion in demand for novel data mining technologies and for students who are trained in data mining—students who have an understanding of data mining techniques, can apply them to real-life problems, and are trained for research and development of new data mining methods. Courses in data mining have started to sprawl all over the world.

Based on this development of the field, the ACM SIGKDD Executive Committee has set up the ACM SIGKDD Curriculum Committee to design a sample curriculum for data mining that gives recommendations for educating the next generation of students in data mining. Based on feedback from researchers, educators, and students, we are convinced that it is an important task to have a carefully designed, conceptually strong, technically rich, and balanced curriculum for this discipline. A comprehensive and balanced curriculum will ensure that the education in data mining sets a solid foundation for the healthy growth of the field, and it will promote systematic training of students in computer science, information sciences, and other related fields, and it will provide guidance for the training of the next generation of data mining researchers, developers and technology users.

The Curriculum Committee is composed of university professors and researchers who have actively contributed to data mining research and education, researchers and practitioners from industry who have rich experiences in applying data mining technology, and administrators from government agencies. This report is the first draft from the Intensive Working Group of the Committee. We expect that this draft will be extensively revised and reviewed, and we are looking forward to suggestions and recommendations from the Committee and from the general data mining research, development, and application community.

The remainder of this report is structured as follows. First, we outline the principles that guided us in the selection of material in Section 2. We then give a brief description of the prerequisites that we assume students of our proposed curriculum to have in Section 3. Section 4 contains the core of this document, our curriculum proposal.

## 2 Curriculum Design Philosophy

Data mining is an interdisciplinary field at the intersection of artificial intelligence, machine learning, statistics, and database systems, and we believe that different educators will emphasize different topics in their

courses. Thus we divided this curriculum proposal into two parts. The first part titled *Foundations* contains basic material that we believe should be covered in any introductory course on data mining. The second part called *Advanced Topics* is a comprehensive collection of material that can be sampled to complete an introductory course or selections of which can form the basis for an advanced course in data mining.

We believe that the teaching of data mining should concentrate on *long-lasting scientific principles and concepts* of the field. Thus instead of covering the last details of the most recent research, we designed the basic material to lay a solid foundation that opens the door to explore more advanced material.

The core endeavor in data mining is to extract knowledge from data; this knowledge is captured in a human-understandable *structure*. The discovery of structure in data is a multifaceted problem that includes the following components:

**Database and Data Management Issues:** Where does the data reside? How is it to be accessed? What forms of sampling are needed? are possible? are appropriate? What are the implications of the database or data warehouse structure and constraints on data movement and data preparation?

**Data Preprocessing:** What are the required data transformations before a chosen algorithm or class of algorithms can be applied to the data? What are effective methods for reducing the dimensionality of the data so the algorithms can work efficiently? How are missing data items to be modelled? What transformations properly encode a priori knowledge of the problem?

**Choice of Model and Statistical Inference Considerations:** What are the appropriate choices to ensure proper statistical inference? What are valid approximations? What are the implications of the inference methods on the expected results? How is the resulting structure to be evaluated? Validated?

**Interestingness Metrics:** What makes the derived structure *interesting* or *useful*? How do the goals of the particular data mining activity influence the choice of algorithms or techniques to be used?

**Algorithmic Complexity Considerations:** What choice of algorithms based on the size and dimensionality of data? What about computational resource constraints? Requirements on accuracy of resulting models? What are the scalability considerations and how should they be addressed?

**Post-processing of Discovered Structure:** How are the results to be used? What are the requirements for use at prediction time? What are the transformation requirements at model application time? How are changes in the data or underlying distributions to be managed?

**Visualization and Understandability:** What are the constraints on the discovered structure from the perspective of understandability by humans? What are effective visualization techniques for the resulting structure? How can data be effectively visualized in the context of or with the aid of the discovered structures?

**Maintenance, Updates, and Model Life Cycle Considerations:** When are models to be changed or updated? How must the models change as the utility metrics in the application domain change? How are the resulting predictions or discovered structure integrated with application domain metrics and constraints?

The partial list above demonstrates that data mining involves many problems and many notions that have historically been studied in isolation. This necessitates a healthy coverage of a wide range of areas within the proposed curriculum.

### 3 Prerequisites

Data mining is a broad field that combines techniques from different areas in computer science and statistics. Our model curriculum assumes that students have basic background knowledge in the following areas:

**Database Systems:** Data models, query languages, SQL, conceptual database design, query processing, and transaction processing.

**Statistics:** Expectation, basic probability, distributions, hypothesis tests, ANOVA, and estimating a distribution parameter.

**Linear Algebra:** Vectors and matrices, vector spaces, basis, matrix inversion, and solving linear equations.

**Algorithms and Data Structures:** We assume familiarity with basic data structures and general maturity of students to understand algorithms written in pseudo-code.

We believe that most computer science seniors either have covered this material in previous courses, can pick up missing material in self-study, or that the missing material is introduced by the course instructor as necessary.

## 4 Course Topics and Models

Recall that we partitioned our curriculum into two parts: A course on *Foundations* and a course on *Advanced Topics*. A standard 14-week one semester introductory course on data mining (offered to either senior undergraduate or first-year graduate students) could cover all the units in Foundations and a selected set of units from the Advanced Topics. A selected set of units from the Advanced Topics can be covered in a second course.

### 4.1 Foundations (Course I)

1. **Introduction.** Basic concepts of data mining, including motivation, definition, the relationships of data mining with database systems, statistics, machine learning, different kinds of data repositories on which data mining can be performed, different kind of patterns and knowledge to be mined, the concept of interestingness, and the current trends and developments of data mining. The material can probably be introduced by showing a few case studies.
  - (a) **Concepts of data mining:** motivation, definition, the relationships of data mining with database systems, statistics, machine learning, and information retrieval.
  - (b) **Knowledge discovery process:** An overview of the Knowledge Discovery Process. Emphasis on the iterative and interactive nature of the KDD Process.
  - (c) **Mining on different kinds of data:** relational, transactional, object-relational, heterogeneous, spatiotemporal, text, multimedia, Web, stream, mobile, and so on.
  - (d) **Mining for different kind of knowledge:** classification, regression, clustering, frequent patterns, discriminant, outliers, and so on.
  - (e) **Evaluation of knowledge:** interestingness or quality of knowledge, including accuracy, utility (such as support), and relevance (such as correlation).
  - (f) **Applications of data mining:** market analysis, scientific and engineering process analysis, bioinformatics, homeland security, and so on.
2. **Data Preprocessing.** This unit will cover the following topics: (1) why preprocess the data? (2) basic data cleaning techniques, (3) data integration and transformation, and (4) data reduction methods. In particular, the following topics will be covered.
  - (a) **Descriptive data summarization:** This unit covers basic techniques for summarizing and describing data. It will cover: (1) computing the measures of central tendency such as mean, and mode, (2) computing the measures of data dispersion such as quantiles, boxplots, variances, standard deviation, and outliers, and (3) graphic display of basic statistical descriptions, such as histogram, scatter plot, boxplot, quantile-quantile plot, and local regression curves.

- (b) **Data cleaning methods:** Basic techniques for handling missing values, noisy data, and inconsistent data, including typical binning, clustering, and regression methods for data cleaning.
  - (c) **Data integration and transformation methods:** This includes data smoothing, data aggregation, data generalization, normalization, attribute (or feature) construction.
  - (d) **Basic data reduction methods:** It introduces binning (histograms), sampling, and data cube aggregation.
  - (e) **Discretization and concept hierarchy generation:** It covers discretization and concept hierarchy generation for numeric data (including binning, clustering, histogram analysis), and for categorical data (automatic generation of concept hierarchies).
3. **Data Warehousing and OLAP for Data Mining.** This unit introduces the concept of a data warehouse and its associated dimensional data model. It then introduces basic OLAP-style analysis on the data cube.
- (a) **Concept and architecture of data warehouse**
  - (b) **The dimensional data model:** including dimensions and measures; star schema, snowflake schema, and fact constellations; data cube concept; concept hierarchies in the cube.
  - (c) **OLAP Operations.** OLAP operations in the multidimensional data model (drill-down, roll-up, slice and dice, pivot)
4. **Association, correlation, and frequent pattern analysis.** This unit covers the concepts and techniques for association, correlation, and frequent pattern analysis, including the following topics.
- (a) **Basic concepts:** frequent patterns, associations, support and confidence of association rules, correlation measure, other objective functions or measures, a typical application scenario: market basket analysis.
  - (b) **Frequent pattern mining methods:** (1) the Apriori algorithm, (2) improvements to Apriori, (3) mining for max-patterns, closed patterns, and top- $k$  patterns.
  - (c) **Mining various kinds of frequent patterns:** (1) multilevel and multidimensional association rules, (2) quantitative association rules, and (3) correlation analysis.
  - (d) **Applications of association rules:** (1) Web log analysis, (2) usage of association rules as classifiers
5. **Classification.** This unit covers the concepts and techniques for classification analysis, including the following topics.
- (a) **Basic concepts:** classification
  - (b) **Evaluation of classification:** (1) evaluation metric, (2) validation for model selection, (3) overfitting, (4) comparing classifiers based on cost-benefit and ROC curves
  - (c) **Bayesian classification:** (1) foundation: Bayes theorem, (2) Naive Bayesian classification methods
  - (d) **Decision tree and decision rule induction:** (1) attribute selection and reduction, (2) basic top-down classification-tree induction schema, (3) pre/post-pruning uninformative subtrees, (4) extraction of rules from classification trees, (5) decision rule induction.
  - (e) **Linear models for classification:** (1) linear discriminant analysis, (2) classification by SVM (Support Vector Machine) analysis
  - (f) **Basic concepts of nonlinear classification:** (1) neural network, (2) SVM with nonlinear kernels

- (g) **Classification by lazy evaluation:** (1) k-nearest neighbor classifier: basic idea and error bounds, (2) locally weighted learning
  - (h) **Ensemble classifier:** Basic ideas why ensemble construction helps, basics of: weighted voting, bagging, boosting.
6. **Cluster and Outlier Analysis.** This unit covers the concepts and techniques for cluster and outlier analysis, including the following topics.
- (a) **Concept of cluster analysis**
  - (b) **Types of data and for dissimilarity computation:** Interval-scaled variables, binary variables, nominal, ordinal, and ratio-scaled variables, and variables of mixed types.
  - (c) **A categorization of major clustering methods**
  - (d) **Partition-based clustering:**  $k$ -means and  $k$ -medoids algorithms, and scalable partitioning methods.
  - (e) **Hierarchical clustering:** agglomerative and divisive hierarchical clustering methods, micro-clusters: integrated and scalable hierarchical clustering methods.
  - (f) **Density-based clustering:** concept of density-based clustering, scalable mining of clustering structures, clustering based on density distribution functions.
  - (g) **Model-based clustering:** (1) The EM Algorithm, (2) neural network approach (SOM)
  - (h) **Outlier analysis:** Concepts and basic outlier detection methods.
7. **Mining Time-Series and Sequence Data.** This unit covers the techniques for mining time-series and sequence data, with the following topics.
- (a) **Regression analysis:** (1) simple and multiple linear regression, (2) nonlinear regression, (3) logistic regression, (4) regression trees, (5) regression using Support Vector Machine, (6) other regression models.
  - (b) **Trend analysis:** A statistical approach
  - (c) **Sequential pattern mining:** Mining different kinds of sequential patterns, sequential pattern mining methods, constraint-based sequential pattern mining, closed sequential patterns, from sequential patterns to partially ordered patterns.
8. **Text Mining and Web Mining.** This unit covers the techniques for mining text and Web data, including the following topics.
- (a) **Mining text databases:** (1) Text data analysis and information retrieval, (2) keyword-based association analysis, (3) document classification, (4) text clustering analysis
  - (b) **Mining the World-Wide Web:** (1) Mining the Web's link structures to identify authoritative Web page, (2) automatic classification of Web documents, (3) construction of a multilayered Web information base, (4) mining social networks, (5) Web resource discovery, (6) Web usage mining.
9. **Visual Data Mining.** This unit covers the visual data mining techniques, including the following topics.
- (a) **Data visualization**
  - (b) **Visualization of data mining results**
  - (c) **Visual data mining:** visual classifier, projection pursuits, class-preserving projections, visualizing class-structure of high-dimensional data, class tours

## 10. Data Mining: Industry efforts and social impacts

- (a) **Social impact of data mining**
- (b) **Data mining and privacy**
- (c) **Standardization efforts**
- (d) **Data mining system products**

## 4.2 Advanced Topics (Course II)

### 1. **Advanced Data Preprocessing.** This unit will cover advanced data reduction methods.

- (a) **Advanced data reduction methods:** (1) dimensionality reduction (feature or attribute subset reduction), (2) numerosity reduction (regression, histogram, clustering, sampling, singular value decomposition (SVD), and discretization), and (3) data compression (lossless versus lossy compression, Fourier and wavelet transformation, and principal component analysis).

### 2. **Data Warehousing, OLAP, and Data Generalization:** This unit covers advanced material in data warehousing, OLAP, and data generalization

- (a) **The multidimensional data model**
- (b) **Implementations of data warehouses:** data integration, indexing OLAP data (bitmap index), efficient processing of OLAP queries, metadata repository, data warehouse back-end tools and utilities.
- (c) **Efficient computation of data cubes:** categorization of measures: distributive, algebraic, and holistic measures, cube computation methods, iceberg cubes, top-down and bottom-up computation, computing closed and approximate data cubes.
- (d) **Other data generalization approaches:** Attribute-oriented induction, mining class comparisons: discriminating between different classes.
- (e) **Exploration of data warehouse and data mining:** Discovery-driven exploration of data cubes, complex aggregation at multiple granularity, cube gradient analysis, from on-line analytical processing to on-line analytical mining.

### 3. **Advanced association, correlation, and frequent pattern analysis.**

- (a) **Advanced frequent pattern mining methods:** (1) vertical format mining, (2) pattern-growth algorithm, (3) mining closed patterns and max-patterns
- (b) **Constraint-based association mining:** (1) rule- and query-guided association mining, (2) anti-monotonicity, monotonicity, succinctness in constrained mining, (3) convertible constraints.
- (c) **Extensions and applications of frequent pattern mining:** (1) iceberg cube computation, (2) fascicles and semantic data compression, (3) frequent pattern-based classification and cluster analysis

### 4. **Advanced Classification.**

- (a) **Bayesian belief networks:** methods for (advanced) choosing BBN structure and training Bayesian belief networks
- (b) **Advanced decision tree construction:** (1) enhancements to basic classification tree induction, (2) scalable algorithms for classification tree induction, (3) integrating data warehousing techniques and classification tree induction, (4) classification with partially labeled data
- (c) **Neural network approach for classification:** (1) a multi-layer feed-forward neural network, (2) defining a network topology, (3) back-propagation, (4) interpretability of classification results.

- (d) **Kernel methods:** (1) kernel logistic regression, (2) kernel discriminant analysis, (3) advanced SVM kernel methods.
  - (e) **Introduction to learning theory:** PAC-learnability, empirical, true and structural risk, VC-theory.
  - (f) **Ensemble construction:** Weighted voting, bagging, weak learner, boosting, AdaBoost
  - (g) **Other classification methods:** (1) case-based reasoning, (2) genetic algorithms, (3) rough set approach, (4) fuzzy set approach
5. **Advanced cluster analysis.**
- (a) **Grid-based clustering:** A statistical information grid approach, clustering by wavelet analysis, clustering high-dimensional space.
  - (b) **Clustering high-dimensional data:** Subspace clustering, frequent pattern-based clustering, clustering by wavelet analysis.
  - (c) **Advanced outlier analysis:** Statistical-based outlier detection, distance-based outlier detection, deviation-based outlier detection, analysis of local outliers.
  - (d) **Collaborative filtering:**
6. **Advanced Time-Series and Sequential Data Mining.** This unit covers the advanced techniques for mining sequential data, including the following topics.
- (a) **Similarity search in time-series analysis:**
  - (b) **Hidden Markov models**
  - (c) **Periodicity analysis:** Transformation-based approach, mining partial periodicity.
  - (d) **Sequence segmentation:** Hidden Markov model and Variable Markov model for sequence segmentation.
  - (e) **Sequence classification and clustering:** (1)  $q$ -gram based methods, keyword-based methods; (2) (high order) Markov chain, hidden Markov model; (3) suffix tree, probabilistic suffix tree, and probabilistic automata.
7. **Mining Data Streams:** This unit covers the techniques for mining stream data, including the following topics.
- (a) **What is stream data?**
  - (b) **Basic tools: Chernoff bounds, reservoir sampling**
  - (c) **Stream sample counting and frequent pattern analysis**
  - (d) **Classification of data streams**
  - (e) **Clustering data streams**
  - (f) **Online sensor data analysis**
8. **Mining Spatial, Spatiotemporal, and Multimedia data.** This unit covers the techniques for mining spatial, spatiotemporal, and multimedia data, including the following topics.
- (a) **Mining spatial and spatiotemporal databases:** (1) Spatial data cube construction and spatial OLAP, (2) spatial association and co-location analysis, (3) spatial clustering methods, (4) spatial classification and spatial trend analysis, (5) spatiotemporal data mining, (6) mining moving objects and trajectories.
  - (b) **Mining multimedia databases:** (1) multidimensional analysis of multimedia data, (2) similarity search in multimedia data, (3) classification and regression analysis of multimedia data, (4) mining association and correlation in multimedia data, (5) clustering multimedia data

- (c) **Mining object databases:** (1) multidimensional analysis of complex objects, (2) generalization on complex structured and semi-structured data, (3) methodology for mining complex object databases: aggregation, approximation, and progressive refinement.
9. **Mining Biological Data:** This unit covers the techniques for mining biological data, including the following topics.
- (a) **Mining DNA, RNA, and proteins:** (1) Mining motif patterns, (2) searching homology in large databases, (3) phylogenetic and functional prediction.
  - (b) **Mining gene expression data:** (1) clustering gene expression, e.g., for gene regulatory networks, (2) classifying gene expression, e.g., for disease-sensitive gene discovery.
  - (c) **Mining mass spectrometry data**
  - (d) **Mining and integrating knowledge from biomedical literature**
  - (e) **Mining inter-domain associations**
10. **Text mining.** This module will cover work that applies known mining techniques to the text media, emphasizing the new issues which arise.
- (a) **Text representation:** Set-of-words, bag-of-words, vector-space model; the issue of large raw dimensionality
  - (b) **Dimensionality reduction:** PCA, SVD, latent semantic indexing
  - (c) **Text clustering:** agglomerative,  $k$ -means, EM; effect of a large number of noise dimensions, partial supervision
  - (d) **Feature selection in high dimensions**
  - (e) **Naive Bayes classification:** Poor density estimates, small-degree Bayesian belief network induction
  - (f) **Discriminative learning:** maximum entropy, logistic regression, and support vector learning
  - (g) **Shallow linguistics:** Phrase detection, part-of-speech tagging, named entity extraction, word sense disambiguation
11. **Hypertext and Web mining.** This module will cover work that is specific to analyzing hypermedia, i.e., involving hierarchical tagging languages and hyperlinks in conjunction with text.
- (a) **Web modeling:** The Web as an evolving, collaborative, populist social network: aggregate graph-structure of the Web, preferential attachment linking models and experimental validation
  - (b) **Link mining and social network analysis:** Links as endorsement: PageRank and HITS algorithms to identify authoritative Web pages; connections with bibliometry
  - (c) **The PageRank algorithm:** Integrating page content and page layout with link structure; topic-sensitive PageRanks; Google
  - (d) **Mining by exploiting text and links:** Exploiting text and links for better clustering and classification; unified probabilistic models for text and links
  - (e) **Structured data extraction:** Information extraction, exploiting markup structure to extract structured data from pages meant for human consumption
  - (f) **Multidimensional Web databases:** Automatic construction of multilayered Web information base; discovering entities and relations on the Web (WebKB)
  - (g) **Exploration and resource discovery on the Web:** reinforcement learning, other approaches
  - (h) **Web usage mining and adaptive Web sites:** Reorganizing Web sites by mining log data

12. **Data Mining Languages, Standards, and System Architectures.** This unit covers the issues related to data mining languages, standards, and system architectures, including the following topics.
  - (a) **Data mining primitives:** what defines a data mining task? task-relevant data, the kind of knowledge to be mined, background knowledge: concept hierarchies, user-specified constraints, interestingness measures, presentation of discovered patterns
  - (b) **Data mining languages, user interfaces, and standardization efforts**
  - (c) **Architectures of data mining systems**
13. **Data Mining Applications.** This unit covers the issues related to domain-specific data mining applications, including the following topics. (Note: Some of these themes, if concrete and good materials are available, should go into the Foundations part as case studies.)
  - (a) **Data mining for financial data analysis**
  - (b) **Data mining for the retail industry**
  - (c) **Data mining for the telecommunication industry**
  - (d) **Data mining for intrusion detection**
  - (e) **Data mining in scientific and statistical applications**
  - (f) **Data mining in software engineering and computer system analysis**
14. **Data Mining and Society.** This unit covers the issues related to social impacts of data mining, including the following topics.
  - (a) **Social impacts of data mining**
  - (b) **Data mining vs. data security and privacy**
  - (c) **Privacy-preserving data mining**
15. **Trends in Data Mining.** This unit covers the major trends in data mining, including the following topics.
  - (a) **Setting solid theoretical foundations for data mining**
  - (b) **Mining deep in specific applications**
  - (c) **Ubiquitous and invisible data mining**
  - (d) **Integrated data and information systems**

## 5 Different course modules and educational goals

Since the course can be taught in different fields, such as computer science, business, and statistics, and with different emphases, such as database, information systems, and machine learning, we should not expect the material will be covered in full spectrum with similar emphasis. We plan to insert some modules based on the feedbacks of instructors who have taught materials in specific fields.

## 6 Laboratories and exercises

Laboratories and exercises give students an opportunity to carry out experiments that illustrate topics in a realistic setting and at the same time learn the specifics of the software used. Students may also be assigned to work on projects too large to be completed during a single class period. Laboratories can provide time for independent project work and programming assignments with reporting similar to that done in other topics in computer science.

The lab projects can be categorized into several categories, and more innovative ideas and suggestions are encouraged.

1. Learn to use data mining systems by using some data mining and data warehousing softwares. Typical such softwares may include Microsoft SQLServer 2005 (Analysis manager), Oracle 10g (data mining part), IBM Intelligent-Miner, and statistics analysis software tools.
2. Implement some data mining functions, including association mining, classification, clustering, sequential pattern mining, text-mining, Web mining, bio-mining, spatial data mining packages. Some open or partially open source data mining systems, such as Weka, IlliMine, and so on can be used for data mining algorithm extension and data mining application exploration.
3. Implementation, refinement, and performance comparison of several different data mining methods.
4. Proposal, implementation and testing of new data mining algorithms and functions.
5. Using some sample data sets to implement and test data mining functions, such as KDD CUP data sets, UC-Irvine Machine Learning/KDD Repository, DBLP database, and other selected Web data sets.