There are three important roles for machine learning.

1. Data Mining: this is using historical data to improve decisions. An example is looking at medical records and applying it to medical knowledge when making a diagnoses.

2. Software applications that we cannot program by hand: Examples of this are autonomous driving and speech recognition

3. Self-customizing programs: An example of this is a newsreader that learns a readers particular interests and highlights these when the reader visits the site.

Machine Learning

A scientific field is best defined by the central question it studies. The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” This question covers a broad range of learning tasks, such as how to design autonomous mobile robots that learn to navigate from their own experience, how to data mine historical medical records to learn which future patients will respond best to which treatments, and how to build search engines that automatically customize to their user’s interests. To be more precise, we say that a machine *learns* with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics. We

might say the defining question of Computer Science is “How can we build machines that solve problems, and which problems are inherently tractable/intractable?” The question that largely defines Statistics is “What can be inferred from data plus a set of modeling assumptions, with what reliability?” The defining question for Machine Learning builds on both, but it is a distinct question. Whereas Computer Science has focused primarily on how to manually program computers, Machine Learning focuses on the question of how to get computers to program themselves (from experience plus some initial structure). Whereas Statistics has focused primarily on what conclusions can be inferred from data, Machine Learning incorporates additional questions about what computational architectures and algorithms can be used to most effectively capture, store, index, retrieve and merge these data, how multiple learning subtasks can be orchestrated in a

larger system, and questions of computational tractability.

Data Mining:

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as *models* or *patterns*. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis (for example, they may have been collected in order to maintain an up-to-date record of all the transactions in a bank). This means that the objectives of the data mining exercise play no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as "secondary" data analysis.

When we are faced with large bodies of data, new problems arise. Some of these relate to housekeeping issues of how to store or access the data, but others relate to more fundamental issues, such as how to determine the representativeness of the data, how to analyze the data in a reasonable period of time, and how to decide whether an apparent relationship is merely a chance occurrence not reflecting any underlying reality. Often the available data comprise only a sample from the complete population (or, perhaps, from a hypothetical super population); the aim may be to *generalize* from the sample to the population. For example, we might wish to predict how future customers are likely to behave or to determine the properties of protein structures that we have not yet seen. Such generalizations may not be achievable through standard statistical approaches because often the data are not (classical statistical) "random samples," but rather "convenience" or "opportunity" samples. Sometimes we may want to summarize or *compress* a very large data set in such a way that the result is more comprehensible, without any notion of generalization. This issue would arise, for example, if we had complete census data for a particular country or a database recording millions of individual retail transactions.

Data mining is often set in the broader context of *knowledge discovery in databases,* or KDD. This term originated in the artificial intelligence (AI) research field. The KDD process involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures. Once again the precise boundaries of the data mining part of the process are not easy to state; for example, to many people data transformation is an intrinsic part of data mining. In this text we will focus primarily on data mining algorithms rather than the overall process.

KDD:

KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets. The data-mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data-mining step of the KDD process. A natural question is, How is KDD different from pattern recognition or machine learning (and related fields)? The answer is that these fields provide some of the data-mining methods that are used in the data-mining step of the KDD process. KDD focuses on the overall process of knowledge discovery from data, including how the data are stored and accessed, how algorithms can be scaled to massive data sets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported. The KDD process can be viewed as a multidisciplinary activity that encompasses techniques beyond the scope of any one particular discipline such as machine learning.

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimatly understandable patterns in data (Fayyad,Piatetsky-Shapiro, and Smyth 1996).

Here, *data* are a set of facts (for example,cases in a database), and *pattern* is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, in our usage here, extracting a pattern also designates fitting a model to data; finding structure from data; or, in general, making any high-level description of a set of data.



The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process. Here, we broadly outline some of its basic steps: First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint. Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed. Third is data cleaning and preprocessing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes. Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found. Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as in Fayyad, Piatetsky- Shapiro, and Smyth (1996). Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities). Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps. Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models. Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge. The KDD process can involve significant iteration and can contain loops between any two steps. The basic flow of steps (although not the potential multitude of iterations and loops) is illustrated in figure 1. Most previous work on KDD has focused on step 7, the data mining. However, the other steps are as important (and probably more so) for the successful application of KDD in practice.