Evolution of Data Mining: From Statistical Foundations to Big Data and Deep Learning

Rajiv Chooramun
Research Scholar
Open University of Mauritius
63 Avenue Poivre Quatre Bornes Mauritius

ABSTRACT
This article traces the historical development of data mining, outlining its evolution through four phases. It begins with the inception of statistical techniques in the 18th and 19th centuries, progresses through advancements in computer technology and artificial neural networks in the mid-20th century, and moves on to the establishment of foundational concepts and algorithms in the final decades of the 20th century. Finally, it addresses the incorporation of big data and deep learning technologies in the 21st century. A comprehensive literature review was conducted to explore the historical progression of data mining. The study examines contributions from early statistical analysis, the impact of electronic computers and database systems, the formalization of data mining concepts and algorithms during the 1990s, and recent advancements driven by big data and deep learning. Each phase has significantly advanced data mining methodologies. Early statistical analysis by figures such as Bayes and Gauss provided foundational groundwork. The advent of electronic computers and database systems enhanced data processing capabilities. The formalization of data mining in the 1990s, marked by 'knowledge discovery in databases and algorithms like support vector machines, expanded its applications. In the 21st century, big data and deep learning have further elevated data mining, solidifying its importance in data science and diverse fields. While this review is limited by the scope of existing literature and historical context, it provides a comprehensive overview of data mining's dynamic evolution and its critical role in extracting valuable insights from datasets. Future research could explore emerging developments and applications in this rapidly evolving field.

General Terms
Data Mining, Machine Learning, Artificial Intelligence, Algorithms, Theory.

Keywords
Data mining, Artificial Intelligence, Machine Learning, Big Data, Knowledge Discovery.

1. INTRODUCTION
The roots of data mining (DM) can be found in three primary fields: classical statistics, artificial intelligence (AI), and machine learning (ML), each with a history extending over four centuries. Several scholarly studies have examined the historical progression of DM, typically categorizing its evaluation into four stages, commencing with the Byes theorem in 1763 and regression analysis [1-3]. The researcher Gregory Piatesky-Shapiro is credited with coining the term "DM" in 1989 [1]. DM refers to the systematic exploration and analysis of extensive data repositories in order to uncover significant connections, patterns, and trends. The process employs pattern recognition technologies along with statistical and mathematical techniques. Also referred to as data or knowledge discovery, DM is a method of extracting meaningful insights and patterns from extensive datasets. Data in DM refers to factual information, numerical values, or textual content that can be subjected to computer processing. This study provides a brief review of the evolutionary trajectory of DM which is divided into four sections. Many other authors have aided our comprehension of DM's historical development. Notable examples include He [1], Sharma [4], Nisbet, Elder [2], and Chen, Chen [3]among others.

DM is considered to be an integral component of the broader field of data science. Data science encompasses a primary emphasis on data and, consequently, incorporates the field of Statistics, which involves a methodical examination of the arrangement, characteristics, and interpretation of data, as well as its significance in making inferences, including the level of confidence associated with such inferences [5]. The discipline of DM significantly grew in the 1990s due to the advancement of relational database technology and the rising automation of commercial operations [5]. The literature on DM during the 1990s elucidated the application of diverse ML techniques to address a wide range of commercial challenges [6-8]. There was a simultaneous increase in the availability of software tools such as SPSS Clementine, SGI Mineset, IBM Intelligent Miner, and SAS Enterprise Miner, which were designed to utilize transactional and behavioural data for the goals of explanation and prediction [1] .

Both ML and AI algorithms emerged throughout the 1950s [9, 10]. According to Mitchell [11] and Teng and Gong [12], ML is of significant importance in the field of DM. Mitchell [11] posited that DM for ML algorithms encompasses various crucial stages beyond the algorithms themselves. These stages include database construction and management, data structuring and cleansing, the presentation of data and summarization, the incorporation of human expert knowledge to develop inputs and evaluate discovered empirical regularities, and determining the appropriate deployment of the obtained results. DM connects various technical areas, such as databases, human-computer communications, statistical analysis, and ML methods. In a similar vein, Teng and Gong [12] proposed that ML assumes an analogous function within the area of AI. ML is a crucial element in the field of AI research due to its ability to automate the acquisition of knowledge. A system without learning capabilities cannot be deemed genuinely intelligent.

The significance of DM has witnessed a notable rise in tandem with the emergence of big data. The process of applying DM techniques to huge, complex, and rapidly expanding data sets is commonly referred to as big DM [13].

Historically, DM techniques have been employed to uncover previously unidentified patterns and relationships of significance from datasets that were structured, homogeneous, and relatively tiny in size, as viewed from a contemporary standpoint. Variety, a fundamental attribute of big data, arises from the presence of an extensive array of diverse sources that produce or help the accumulation of large data [14].

The rest of the study is organized as follows: Section 2 outlines the inception of statistical techniques in the 18th and 19th centuries. Section 3 details the advancements in computer technology and artificial neural networks in the mid-20th century. Section 4 discusses the establishment of key concepts and algorithms in the late 20th century. Section 5 explores the integration of big data and deep learning in the 21st century. Finally, Section 6 concludes the study.

2. FIRST GENERATION (1763-1930)

Since DM is focused on understanding past events by grouping data or taking averages, it can be said that DM existed about 1000 BC in China. DM often utilizes statistical techniques as a fundamental component of its methodology. Statistical techniques are essential in various stages of the DM process, helping to analyze and interpret patterns in data [15]. The roots of DM can be linked to early statistical analysis. In the 17th and 18th centuries, individuals like John Graunt and Thomas Bayes made contributions to the analysis of demographic data and probability theory, setting the stage for later developments in statistics. Thomas Bayes introduced in Bayes theorem in 1763. The 18th and 19th centuries saw the development of probability theory and statistical methods, with figures like Pierre-Simon Laplace and Sir Francis Galton contributing to the understanding of probability distributions and statistical inference. Carl Friedrich Gauss introduced the method of least squares in [16]. Sir Francis Galton introduced the concept of regression in [17]. Karl Pearson introduced the concept of correlation in [18].

The idea of using machines to automate tasks related to data analysis has early roots. For example, Herman Hollerith's invention of the tabulating machine for the 1890 U.S. Census marked a significant development. This electromechanical device was designed to process punch cards and automate the tabulation of census data, representing an early form of data processing. Charles Babbage's Analytical Engine, conceived in [19], had concepts that foreshadowed aspects of computation and data processing. Astronomers and navigators have long used observational data to make predictions and understand patterns in celestial movements. Ptolemy's work in ancient times and later astronomers' use of tables and charts to predict celestial events can be seen as an early form of extracting knowledge from data. Further, demographers and public health officials started collecting and analyzing data related to populations and diseases. John Snow's work during the 1854 cholera outbreak in London, where he used a map to identify patterns in the geographic distribution of cases, is an early example of spatial analysis and pattern recognition in data [20]. Early forms of quality control involved the inspection of products and the analysis of defects. Walter A. Shewhart, in the 1920s and 1930s, developed statistical methods for quality control, laying the groundwork for later advancements in statistical process control [21].

The concept of a general-purpose electronic computer as we understand it today did not exist until the 1940s. However, there were mechanical and electromechanical devices used for specific computational tasks before the 1930s. These early computing devices were not programmable in the way modern computers are, and they were often specialized for specific tasks. The development of electronic computers, capable of general-purpose computation and programmability, began in the 1940s with machines like the ENIAC (1946) and the Harvard Mark I (1944). In summary, while there were mechanical and electromechanical devices used for computational purposes before the 1930s, the era of electronic computers did not begin until the 1940s.

3. SECOND GENERATION (1940-1988)

During the second stage of DM, notable progress was made worldwide. Key developments include: advancements in computer technology and database management systems, leading to increased processing power and storage capabilities [22]. The emergence of the artificial neural network (ANN) algorithm, enabling more sophisticated pattern recognition and prediction tasks [23]. The evolution of statistical learning theory (SLT), providing a framework for analyzing and interpreting data-driven models [24]. The classification of analytical techniques into supervised and unsupervised learning categories, allowing for more targeted and efficient data analysis [25]. The development of decision tree models, which offer a simple yet effective approach to data classification and decision-making [26].

The rise of computers from the 1940s marked a transformative era in the history of technology, ushering in a new age of computation, data processing, and automation. Completed in 1946, the electronic numerical integrator and computer (ENIAC) is recognized as one of the first general-purpose electronic digital computers. Designed to perform complex numerical calculations, ENIAC was crucial for military computations, especially during World War II [27]. IBM's 700 series, introduced in the early 1950s, included a range of mainframe computers. These machines are widely used in business, scientific research, and government applications [28]. The development of transistors in the 1950s, replacing vacuum tubes, resulted in smaller, more reliable, and energy-efficient computers. This development significantly contributed to the miniaturization of electronic components [29].

The introduction of the IBM System/360 marked a significant milestone. It was a family of compatible mainframe computers, offering different models to meet various computing needs. This approach of compatibility across a product line was influential. The invention of the microprocessor in the early 1970s, such as the Intel 4004, marked the beginning of the era of personal computing. Microprocessors allowed for the integration of CPU functions on a single chip. Personal computers, such as the Apple II (1977) and IBM PC (1981), brought computing capabilities to individuals and small businesses. he advent of graphical user interfaces, popularized by the Apple Macintosh (1984) and Microsoft Windows (1985), made computers more user-friendly and accessible to a wider audience [30].

In addition to the aforementioned advancements in computers, numerous statistical software packages were developed during this age of DM. These software programs enabled the researcher to fully investigate the capabilities of statistical approaches used in DM. Only the most notable software packages are mentioned here. The BASIC Statistical Package was developed in the 1960s at the University of Michigan. This statistical package was one of the early statistical software systems developed for mainframe computers [31] .

The SAS (Statistical Analysis System) was introduced in the late 1960s, with the first version, SAS 1, developed between

1966 and 1968 at North Carolina State University. SAS is a comprehensive statistical software suite that includes modules for data management, statistical analysis, and reporting. It has been continuously updated and expanded over the years. SPSS (Statistical Package for the Social Sciences) was first introduced in 1968 at Stanford University. SPSS was initially designed for social science researchers and gained popularity for its user-friendly interface and statistical analysis capabilities. Linear structural relation (LISREL) was first developed by [32]. LISREL is statistical software designed for structural regression modelling, using systems of linear equations known as structural equation models. Minitab was initially introduced in 1972 by Minitab; Inc. Minitab is a statistical software package designed for quality improvement and statistical analysis. It became popular in academia and industry for its ease of use and statistical capabilities [33].

Further, the 1970s set the stage for the foundational technologies in database management and data retrieval. The subsequent decades would witness the integration of these technologies into more comprehensive DM practices. The 1970s witnessed the development and adoption of early database management systems. These systems, such as IBM's Information Management System (IMS) and the CODASYL database model, provided a structured way to organize and manage large volumes of data. SQL, developed in the early 1970s, became a standardized language for interacting with relational databases. This standardization made it easier for users to retrieve and manipulate data, laying the groundwork for more sophisticated data analysis. The concept of relational databases gained prominence during this time, with Edgar F. Codd publishing his influential paper on the relational model in 1970. Relational databases provided a more flexible and efficient way to organize and relate data, setting the stage for future DM activities. The 1970s saw the development of decision support systems, which were digital tools created to aid in the decision-making process. While not explicitly focused on DM, these systems often involved the analysis of structured data to support managerial decisions [34].

With the utilization of previously developed statistical technique such as Correlation, regression models, PCA, Factor Analysis etc., and the advancements in computers and databases, the applications of DM gained momentum. Artificial Neural Networks (ANNs) and Decision Trees have played significant roles in advancing DM by providing powerful tools for extracting meaningful patterns, making predictions, and gaining insights from large datasets [15].

3.1 Artificial Neural Networks (ANNs)

Warren McCulloch and Walter Pitts introduced the concept of ANNs in 1943. Later, in 1957, Frank Rosenblatt introduced the perceptron, a specific type of neural network. Warren McCulloch and Walter Pitts's[23] article, titled "A Logical Calculus of the Ideas Immanent in Nervous Activity," is a landmark contribution to the fields of neuroscience and AI. McCulloch and Pitts introduced a simplified model of a neuron, which they referred to as a "formal neuron" or "McCulloch-Pitts neuron." A neural network is a computational model modeled after the human brain, made up of layers of interconnected nodes (neurons) that are structured to analyze and learn from data using weights and activation functions. It excels in tasks like pattern recognition and decision-making by adapting and improving based on feedback (training).

The McCulloch-Pitts neuron model can be represented mathematically with a simple threshold function. For example, if there is a single McCulloch-Pitts neuron with n binary input signals $x_1, x_2, \cdots, x_n$ and corresponding weights $w_1, w_2, \cdots, w_n$. The output of the neuron, denoted as y, is determined by comparing the weighted sum of inputs to a threshold T:

$$y = \begin{cases} 1 & \text{if} \sum_{i=1}^{n} w_i \cdot x_i \geq T \\ 0 & \text{otherwise} \end{cases} \quad 8.1$$

where, w_i represents the weight associated with the input x_i, and the threshold T is a parameter that determines the firing threshold of the neuron. If the weighted sum of inputs is equal to or exceeds the threshold, the neuron produces an output of 1 (firing); otherwise, it produces an output of 0 (not firing). This binary model is a simplified binary model, and the weights and thresholds are typically set manually in this original formulation. Modern neural networks, on the other hand, often use continuous activation functions and involve a training process to adjust weights based on input-output pairs during learning.

This idea of threshold logic laid the foundation for later developments in artificial neural networks. The authors demonstrated that networks of these artificial neurons could be configured to perform logical operations, such as AND, OR, and NOT. This showed that simple computational units, connected in specific ways, could carry out complex information processing tasks.

An AND gate produces an output of 1 only if all of its inputs are 1.

$$y = \begin{cases} 1 & \text{if} x_1 = 1 \; and \; x_2 = 1 \\ 0 & \text{otherwise} \end{cases} \quad (8.1)$$

In the context of McCulloch-Pitts neurons, the weights ($w_1 \; and \; w_2$) are set such that the threshold is reached only when both inputs are 1.

An OR gate produces an output of 1 if at least one of its inputs is 1.

$$y = \begin{cases} 1 & \text{if} x_1 = 1 \; or \; x_2 = 1 \\ 0 & \text{otherwise} \end{cases} \quad (8.2)$$

In the McCulloch-Pitts neuron model, the weights ($w_1 \; and \; w_2$) are set such that the threshold is reached when at least one input is 1.

A NOT gate produces an output that is the opposite of its input.

$$y = \begin{cases} 1 & if \; x_1 = 0 \\ 0 & if \; x_1 = 1 \end{cases} \quad (8.3)$$

In the McCulloch-Pitts neuron model, this could be achieved by setting a negative weight and a threshold such that the neuron fires when the input is 0 and remains inactive when

the input is 1

While the formal neurons in their model were abstractions, McCulloch and Pitts aimed to draw connections between their logical calculus and the behaviour of real neurons in the brain. This interdisciplinary approach laid the groundwork for future research at the intersection of neuroscience and AI. Today's deep learning landscape, encompassing convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other advanced architectures, is based on the foundational work of McCulloch and Pitts. The simplicity of their model demonstrated the feasibility of using artificial systems to perform computations inspired by the functioning of biological neurons, setting the stage for the development of increasingly complex and powerful artificial neural networks.

3.2 Statistical learning theory (SLT)

SLT, as a field of study, is attributed to Vladimir Vapnik and Alexey Chervonenkis in [35]. SLT emerged as a distinct field of inquiry, initially focused on the rigorous analysis of function estimation from available data. SLT underwent a significant transformation in the 1990s with the introduction of novel ML algorithms, specifically support vector machines (SVMs), which were rooted in the theoretical foundations of SLT. This development marked a shift from a purely theoretical framework to a practical tool for estimating multidimensional functions, with SLT now serving as both a theoretical analysis framework and a source of practical algorithms for function estimation [24].The empirical risk minimization (ERM) principle is fundamental in SLT. It involves minimizing the empirical risk, which is the average loss over the training dataset. For a hypothesis *h* and training set (x_i, y_i) the empirical risk $R_{emp\,(h)}$ is often defined using a loss function *L* as:

$$R_{emp\,(h)} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} L(h(x_i), y_i) \qquad (8.5)$$

The generalization error, also known as the expected risk, represents how well a model trained on a specific dataset performs on unseen data. It represents the gap between the actual risk and the observed risk:

$$R_{gen}(h) = R(h) - R_{emp\,(h)} \qquad (8.6)$$

The VC dimension quantifies the capacity of a hypothesis class. The VC dimension d_{VC} of a hypothesis class H is defined as the highest number of points that H can shatter. The growth function $m_H(N)$ represents the number of distinct labeling for *N* points:

$$m_H(N) \leq 2^N \qquad (8.7)$$

The VC dimension is related to the growth function as

$$d_{VC} = sup\{N : m_{H(N)} = 2^N\} \qquad (8.8)$$

Rademacher complexity evaluates the complexity of a hypothesis class relative to a set of random variables called Rademacher variables. Given a hypothesis class *H* and a sample $(x_1, y_1), \ldots, (x_N, y_N)$, the empirical Rademacher complexity $R_{emp}(H)$ is defined as

$$R_{emp}(H) = \left(\frac{1}{N}\right) E_\sigma \left[sup_{h \in H} \sum_{i=1}^{N} \sigma_i\, h(\boldsymbol{x}_i)\right] \qquad (8.9)$$

where σ_i are Rademacher variables taking values +1 or -1 with equal probability.

In the context of SVM, which is closely related to SLT, margin refers to the space between the decision boundary and the closest data point. For a hyperplane defined by $\boldsymbol{w}.\boldsymbol{x} + b$, the margin (M) is given by:

$$M = \frac{1}{\|\boldsymbol{w}\|} \qquad (8.10)$$

3.3 Supervised and Unsupervised Learning Techniques

Supervised and unsupervised learning concepts emerged early in the development of machine learning (ML), evolving through the mid-20th century. The term "ML" was first introduced by Arthur Samuel in 1956, who is regarded as one of the pioneers in the field. Samuel described ML as “the ability of a machine to acquire knowledge from experience without direct programming” [36].

During the 1950s and 1960s, researchers started investigating the use of ML algorithms for classifying and predicting continuous outcomes, such as values of continuous variables. This exploration gave rise to supervised learning methods, which entails training an ML model using labelled data to make inferences about new data [25]. By the 1960s and 1970s, researchers expanded their focus to use ML algorithms for uncovering patterns and structures in data without prior knowledge of expected outcomes. This resulted in the creation of unsupervised learning methods, where algorithms learn from unlabeled data to identify patterns and connections in the data [37].

Although certain statistical methods were formulated prior to their current classification under the umbrella of supervised and supervised learning techniques, it is now commonplace to categorize all statistical methods within these two broader categories. Some of these techniques are mentioned below.

Supervised learning techniques started with linear regression in 1795 (least squares method), but formalized statistical methods emerged later (see, Chapter 2). The logistic function was introduced in the 19th century, but logistic regression gained popularity in statistics in the mid-20th century. Discriminant analysis was introduced in the 1930s. It is used for classification and dimensionality reduction by finding linear combinations of features that best discriminate between predefined classes or groups (see, Chapter 4). Decision trees were introduced in the 1960s (early forms), but gained popularity in ML in the 1980s. The concept of k-Nearest Neighbors (k-NN) is old, but the k-NN algorithm gained popularity in the 1970s. Support vector machines (SVM) was introduced in 1992 (original formulation), but it was popularized in the 1990s by Vapnik and Cortes. Neural networks, random forests, gradient boosting machines (xgboost), and naive bayes are also supervised leaning techniques [37].

Principal component analysis (PCA) is a technique for dimensionality reduction that transforms the original variables into a new set of independent variables known as principal components. It is often used to simplify the dataset by capturing the most important information and reducing noise [38]. Similar to PCA, factor analysis seeks to uncover underlying factors or latent variables that elucidate patterns in the observed variables. It is particularly useful when there are latent variables that cannot be directly measured [39]. Cluster analysis groups similar observations or variables into clusters based on their similarities. Techniques like K-means

clustering or hierarchical clustering can be applied to multivariate datasets [37].

Canonical correlation analysis (CCA) is also anunsupervised learning technique. CCA explores the relationships between two sets of variables and identifies linear combinations of variables that maximize the correlation between the sets. It is useful for understanding the associations between two sets of variables [37]. Association rule mining identifies relationships and dependencies between variables in transactional databases. It is important to note that the distinction between supervised and unsupervised SEM is based on the researcher's objectives and the nature of the analysis, rather than being an intrinsic property of SEM itself [40].

3.4 Decision Trees

One of the earliest works on decision trees is that of Morgan and Sonquist, who published a paper titled "Problems in the Analysis of Survey Data, and a Proposal" in [41]. However, the formalization and popularization of decision trees in the field of ML can be credited to the work of Leo Breiman, J. Friedman, R. Olshen, and C. Stone, who published the influential book "Classification and Regression Trees" in [26]. CART is a non-parametric statistical method that can be used for both classification and regression tasks. The key idea behind CART is to recursively partition the data into subsets based on the values of input features, creating a tree-like structure, as shown in Figure 1.

At each node in the tree, a binary splitting rule is applied to divide the dataset into two subsets based on a selected feature and a threshold value. This can be represented as:

Splitting Rule: $x_i < \text{threshold}$

where $x_i$is the value of the selected feature.

For classification tasks, CART often uses Gini impurity as an impurity measure. The Gini impurity for a node *t* is calculated as follows:

$$G\ (t) = 1 - \sum_{i=1}^{c} p\ (i/t)^2 \qquad (8.11)$$

where *c* is the number of classes and $p\ (i/t)$is the proportion of class *i* samples in node *t*.

For regression tasks, the mean squared error (MSE) is commonly used as a splitting criterion:

$$MSE(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \bar{y}_t)^2 \qquad (8.12)$$

where $N_t$is the number of samples in node *t*, $y_i$is the target value for the *i*-th sample, and $\bar{y}_t$ is the mean target value in node *t*.

The recursive tree-building process involves iteratively applying the binary splitting rule to create child nodes until a stopping criterion is met. This process aims to minimize the impurity measure or the MSE. After tree construction, pruning involves removing branches to avoid overfitting. This can be achieved by introducing a cost-complexity term that penalizes the complexity of the tree.

4. THIRD GENERATION (1989-2000)

The period between 1989 and 2000 marked a crucial phase in the evolution of DM. The field saw the establishment of foundational concepts, algorithms, and applications, setting the stage for further growth and development in subsequent years. The term "DM" was coined in 1989 by Gregory Piatesky-Shapiro [1].

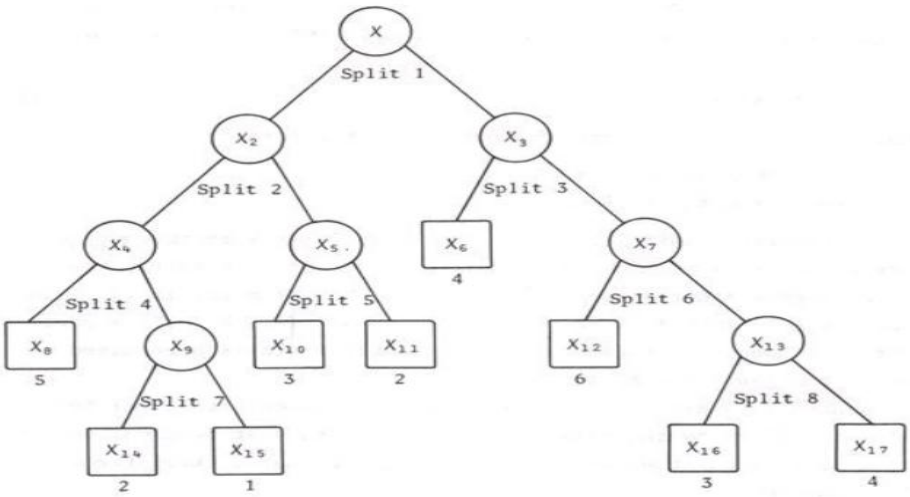

Fig 1: Structured Classifiers

Source: Breiman [42]

4.1 Decision Trees

The term KDD was popularized in the late 1980s and early 1990s, particularly through the work of Gregory Piatetsky-Shapiro and others [43]. While there may not be a specific individual who "introduced" KDD in a single moment, Piatetsky-Shapiro is often credited with coining the term. In 1989, Gregory Piatetsky-Shapiro, along with William J. Frawley and Christopher J. Matheus, organized the first workshop on "knowledge discovery in databases" at the AAAI-89 conference. This workshop is considered a key event in the early development of the KDD field. The term KDD gained popularity as a result of this workshop, and it became associated with the systematic process of extracting useful knowledge from large datasets. Piatetsky-Shapiro continued to contribute significantly to the field of KDD, co-founding the journal "DM and Knowledge Discovery" in 1997. The KDD process has since become a fundamental aspect of DM and ML, encompassing various techniques and methodologies for extracting meaningful patterns, knowledge,

and insights from data [44, 45].

4.2 Support Vector Machine (SVM)

The early 1990s saw the introduction of the SVM model. The foundational work on SVMs is often attributed to Vladimir Vapnik and his colleagues. In [24], Vapnik and Corinna Cortes published a paper titled "Support-Vector Networks" that introduced the concept of SVMs. SVMs are supervised ML algorithms suitable for classification and regression tasks. They excel in high-dimensional spaces and are ideal for tasks requiring clear decision boundaries. SVMs are extensively applied in image classification, text categorization, and bioinformatics.

In the SVM technique, the basic idea is to find a hyperplane that best separates the data points of different classes. For example, the decision function for a linear SVM is given by:

$$f(x) = \boldsymbol{w} . \boldsymbol{x} + b \quad (8.13)$$

where: $f(x)$refers to the decision function; **w** is the weight vector that is perpendicular to the hyperplane; **x** is the vector of input features; and *b* stands for the bias term.

The equation of the hyperplane is expressed as:

$$\boldsymbol{w} . \boldsymbol{x} + b = 0 \quad (8.14)$$

This equation represents the decision boundary that separates the data points of different classes.

The margin (M) is the distance between the hyperplane and the nearest data point from each class. It is given by the following formula:

$$\mathrm{M} = \frac{2}{\|\boldsymbol{w}\|} \quad (8.15)$$

Here, $\|w\|$is the Euclidean norm of the weight vector.

The objective function for linear SVM in its primal form is to maximize the margin while ensuring that all the data points are correctly classified. It is often expressed as a minimization problem:

$$min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad (8.16)$$

subject to the following constraints:

$$y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq 1$$

for all training samples $(\boldsymbol{x}_i y_i)$.

The Lagrangian for the optimization problem is defined as:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\| - \sum_{i=1}^{N} \alpha_i [y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) - 1] \quad (8.17)$$

Here, $\alpha_i$are the Lagrange multipliers.

The dual form of the optimization problem involves maximizing the Lagrangian with respect to α:

$$max_{\alpha}\, W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j \ (\boldsymbol{x}_i . \boldsymbol{x}_j) \quad (8.18)$$

subject to the following constraints:

$$\alpha_i \geq 0 \ and \ \sum_{i=1}^{N} \alpha_i y_i = 0$$

Once the optimal α values are determined, **w** can be calculated as:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i \quad (8.19)$$

The bias term *b* can be computed using any support vector:

$$b = y_i - \boldsymbol{w} . \boldsymbol{x}_i \quad (8.20)$$

These equations capture the essence of a linear SVM. In practice, when dealing with non-linearly separable data, the kernel trick is often employed to map the input features into a higher-dimensional space, introducing non-linearity to the decision boundary. The equations are then adapted accordingly to incorporate the chosen kernel function.

5. FOURTH GENERATION (2001-PRESENT)

In the fourth generation of DM, three significant advancements occurred. The first is the increased use of "deep learning" in the context of artificial neural networks with multiple layers, which became more prevalent in the 2000s. The second is the concept of "big data," which gained popularity in the early 21st century to refer to datasets that are too vast and complicated for traditional data analysis methods. Doug Laney, a researcher and analyst, is credited with defining the three Vs of data—volume, velocity, variety—in a paper published in 2001. Finally, the emergence of "data science" as a field, which gained prominence around 2008 [46-48].

5.1 Deep Learning

In the early 2000s, researchers began to create advanced AI algorithms known as deep learning neural networks. These algorithms are capable of learning multiple layers of representation from data, making them highly versatile and effective in numerous applications, such as image and speech recognition, natural language processing, and self-driving vehicles [49].

As shown in Figure 2, the development of deep learning techniques began in the 1940s to 1960s with the introduction of artificial neural networks and early models. During this time, foundational work laid the groundwork for advancements in neural network research. In the 1980s, a crucial algorithm for training neural networks, backpropagation, was created. However, limitations, such as the vanishing gradient problem, hinder the effectiveness of deep neural networks during this time [50]. In the late 1990s and early 2000s, Geoffrey Hinton and his colleagues introduced deep belief networks (DBNs), a type of deep neural network architecture. A significant breakthrough in deep learning occurred in 2012 when a deep neural network achieved remarkable performance in image classification through the ImageNet Large Scale Visual Recognition Challenge. Since then, there has been a rapid progression of deep learning research and applications spanning multiple domains, such as computer vision, natural language processing, and speech recognition, with significant advancements taking place since 2012 onward [51].

While the term "deep learning" gained popularity in the early 2000s, it is important to recognize that the foundations of this field were laid much earlier. The recent surge in interest in deep learning is the outcome of major progress in algorithms, computational capacity, and access to extensive, labelled datasets. These milestones, which occurred around the turn of the century, have contributed to the ongoing development and

innovation in deep learning. In other words, the field is still evolving, with new breakthroughs and improvements being made regularly [52].

5.2 Big Data

In the early 21st century, the term "big data" emerged to refer to datasets that were too vast and intricate for conventional data processing techniques. This concept has been around for some time, but the term gained popularity in the mid-2000s [53]. In 2001, researcher and analyst Doug Laney defined the three key characteristics of big data, known as the three Vs: volume, velocity, and variety. These refer to the large volumes of data, fast data processing speeds, and diverse data types that are characteristic of big data. In 2008, the open-source framework Apache Hadoop became a significant milestone in the processing of massive amounts of data across distributed clusters. Based on the MapReduce programming model developed by Google, Hadoop enables the processing of big datasets with ease [54].

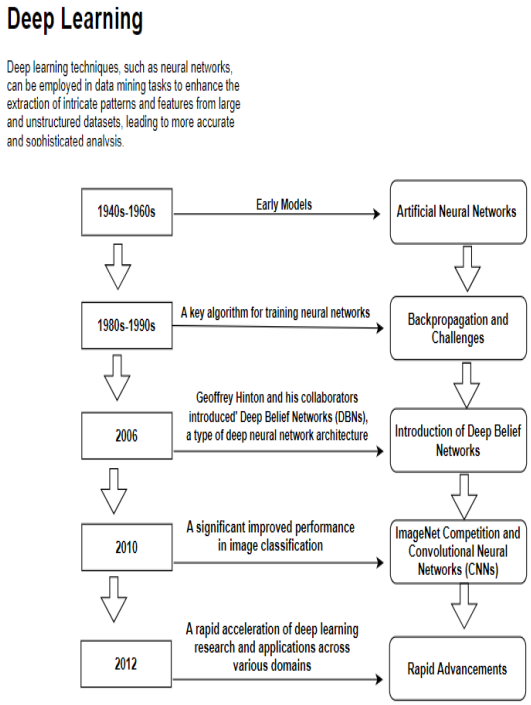

Figure 1 Development of Deep Learning
Source: Author

In the early 2010s, big data gained widespread recognition among industries as a valuable tool for data analysis, business intelligence, and decision-making. By 2012, big data had become a mainstream topic, with businesses and sectors across various industries exploring how to harness large datasets for insights and innovation. Since then, the field of big data has continued to evolve, with advancements in technologies, tools, and methodologies for collecting, storing, and analysing massive volumes of data. While the term "big data" has only recently emerged, the challenges and opportunities associated with large datasets have been present for decades, and the term simply provides a convenient label for addressing these challenges [55, 56].

5.3 Data Science

The term "data science" has evolved from various fields, with its origins difficult to pinpoint to a specific year. However, the term gained prominence and recognition in the early 21st century as the field expanded to encompass a wide range of skills and techniques for extracting insights from data. The foundational principles of data science draw upon statistics, computer science, and domain-specific expertise. Statistical methods for data analysis have been used for many years, and computer science has long been involved in data processing and analysis. The terms "DM" and "knowledge discovery in databases (KDD)" were commonly used in the late 1990s and early 2000s to describe the process of uncovering patterns and insights from large datasets [57].

The concept of "data science" gained significant attention around 2008, as statisticians, computer scientists, and practitioners recognized the need for a more integrated approach to unlocking value from data. This growing recognition culminated in the term "data scientist" gaining further prominence in 2012, following the publication of an article in the Harvard Business Review, which highlighted the importance of interdisciplinary skills in statistics, programming, and domain expertise [58]. As the field evolved, data science became more formally established in the 2010s, with the development of dedicated academic programs and increased demand from businesses for professionals with data science skills [59]. This shift toward a more holistic and interdisciplinary approach to data analysis and management has led to the emergence of data science as a distinct field with its own methodologies, tools, and best practices.

6. DISCUSSION AND CONCLUSION

6.1 Discussion

The historical trajectory of DM is a testament to the relentless pursuit of knowledge extraction from data. Starting from the rudimentary statistical techniques of the 18th century, such as Bayes' theorem and Gauss' least squares method, the field laid its foundational principles on probability and inference. These early efforts underscored the importance of structured data analysis, setting a precedent for future developments. The mid-20th century saw a paradigm shift with the advent of electronic computers, which exponentially increased data processing capabilities. This era marked the birth of ANN and database management systems, revolutionizing data handling and pattern recognition. The integration of statistical learning theory provided a robust framework for analyzing data-driven models, enhancing the precision and applicability of DM techniques.

The formalization of DM in the 1990s, epitomized by the term "KDD" and the creation of support vector machines, signified a pivotal moment. These advancements facilitated the adoption of DM across diverse domains, from finance to healthcare, demonstrating its versatility and potential. The 21st century heralded the age of big data and deep learning, further propelling DM into new dimensions. The capability to handle large and intricate datasets has made DM indispensable in contemporary data science, driving innovations in AI, predictive analytics, and beyond. As we look ahead, the continuous evolution of DM promises to unlock unprecedented insights, shaping the future of technology and decision-making in profound ways.

6.2 Conclusion

The historical development of DM highlights its profound impact on the extraction and analysis of data. Beginning with early statistical methods by pioneers like Bayes and Gauss, the field has continuously evolved, incorporating advancements in computing and database technologies. The mid-20th century introduced ANN and SLTs, significantly

enhancing data processing and pattern recognition capabilities. The formalization in the 1990s, marked by the introduction of "KDD" and algorithms such as support vector machines, facilitated widespread adoption across various sectors.

Entering the 21st century, the emergence of big data and deep learning has further revolutionized DM, making it a cornerstone of modern data science. These advancements have enabled the handling of vast, complex datasets, driving innovations in fields like AI and predictive analytics. As data continues to grow exponentially, the ongoing evolution of DM will remain critical, unlocking new insights and shaping future technological advancements.

7. ACKNOWLEDGMENTS

The author extends sincere gratitude to Dr. Prakash Gorroochurn, Associate Professor of Biostatistics, Columbia University, for his invaluable guidance, and to the DRC members for their constructive feedback and support throughout the development of this work.

8. REFERENCES

[1] He, J. *Advances in data mining: History and future*. in *2009 Third International Symposium on Intelligent Information Technology Application*. 2009. IEEE.

[2] Nisbet, R., J. Elder, and G.D. Miner, *Handbook of statistical analysis and data mining applications*. 2018: Academic press.

[3] Chen, G., et al., *A Review of the Development and Future Trends of Data Mining Tools.* Innovative Computing: IC 2020, 2020: p. 113-119.

[4] Sharma, M., *Data mining: A literature survey.* International Journal of Emerging Research in Management & Technology, 2014. **3**(2).

[5] Dhar, V., *Data science and prediction.* Communications of the ACM, 2013. **56**(12): p. 64-73.

[6] Piatetsky-Shapiro, G., *Discovery, analysis, and presentation of strong rules.* Knowledge Discovery in Data-bases, 1991: p. 229-248.

[7] Linoff, G.S. and M.J. Berry, *Data mining techniques: for marketing, sales, and customer relationship management*. 1997: John Wiley & Sons.

[8] Tukey, J.W., *Exploratory data analysis*. Vol. 2. 1977: Reading, MA.

[9] Fradkov, A.L., *Early history of machine learning.* IFAC-PapersOnLine, 2020. **53**(2): p. 1385-1390.

[10] Delipetrev, B., C. Tsinaraki, and U. Kostic, *Historical evolution of artificial intelligence.* 2020.

[11] Mitchell, T.M., *Machine learning and data mining.* Communications of the ACM, 1999. **42**(11): p. 30-36.

[12] Teng, X. and Y. Gong. *Research on application of machine learning in data mining*. in *IOP conference series: materials science and engineering*. 2018. IOP Publishing.

[13] Wu, X., et al., *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2013. **26**(1): p. 97-107.

[14] Che, D., M. Safran, and Z. Peng. *From big data to big data mining: challenges, issues, and opportunities*. in *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, April 22-25, 2013. Proceedings 18*. 2013. Springer.

[15] Han, J., J. Pei, and H. Tong, *Data mining: concepts and techniques*. 2022: Morgan kaufmann.

[16] Gauss, C.F., *The theory of the combination of observations least subject to errors*. 1795: H. W. Miller.

[17] Galton, F., *Typical laws of heredity.* Nature, 1877. **15**: p. 492-495.

[18] Pearson, K., *Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London, 1896. **60**(1): p. 489-498.

[19] Babbage, C., *On the economy of machinery and manufactures*. 1832: Charles Knight.

[20] Snow, J., *On the mode of communication of cholera*. 1855: John Churchill.

[21] Shewhart, W.A., *Economic control of quality of manufactured product*. 1931: D. Van Nostrand Company, Inc.

[22] Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and.* Techniques, Waltham: Morgan Kaufmann Publishers, 2012.

[23] Pitts, W. and W.S. McCulloch, *A logical calculus of the ideas immanent in nervous activity.* Bulletin of Mathematical Biophysics, 1943. **5**(4): p. 115-133.

[24] Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**: p. 273-297.

[25] Mitchell, T.M. and T.M. Mitchell, *Machine learning*. Vol. 1. 1997: McGraw-hill New York.

[26] 26. Breiman, L., *Classification and regression trees*. 1984: Routledge.

[27] Goldstine, H.H., *The computer from Pascal to von Neumann*. 1993: Princeton University Press.

[28] Bashe, C.J., et al., *IBM's early computers*. 1986: MIT press.

[29] Riordan, M. and L. Hoddeson, *Crystal fire: The birth of the information age*. 1997: WW Norton & Company.

[30] Ceruzzi, P.E., *A history of modern computing.* 2003: MIT press.

[31] Nie, N.H., D.H. Bent, and C.H. Hull, *SPSS: Statistical Package for the Sciences*. 1970: McGraw-Hill.

[32] Jöreskog, K., *A general method for estimating a linear structural equation system.* ETS Research Bulletin Series, 1970. **1970**(2): p. i-41.

[33] Raykov, T. and G.A. Marcoulides, *A first course in structural equation modeling*. 2012: routledge.

[34] Codd, E.F., *A relational model of data for large shared data banks*. Communications of the ACM, 1970. **13**(6): p. 377-387.

[35] Vapnik, V. and A. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities.* Theory of Probability and Its Applications,

1971. **16**(2): p. 264-280.

[36] Samuel, A.L., *Some studies in machine learning using the game of checkers.* IBM Journal of research and development, 1959. **3**(3): p. 210-229.

[37] Hastie, T., Tibshirani, R., and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.

[38] Jolliffe, I.T., *Principal component analysis for special types of data*. 2002: Springer.

[39] Everitt, B. and T. Hothorn, *An introduction to applied multivariate analysis with R*. 2011: Springer Science & Business Media.

[40] Hastie, T., et al., *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. 2009: Springer.

[41] Morgan, J.N. and J.A. Sonquist, *Problems in the analysis of survey data, and a proposal.* Journal of the American statistical association, 1963. **58**(302): p. 415-434.

[42] Breiman, L., *Classification and regression trees*. 2017: Routledge.

[43] Azevedo, A., *Data mining and knowledge discovery in databases*, in *Advanced methodologies and technologies in network architecture, mobile computing, and data analytics*. 2019, IGI Global. p. 502-514.

[44] Piateski, G. and W. Frawley, *Knowledge Discovery in Databases.-MIT Press, Cambridge.* MA, USA, 1991.

[45] Salzberg, S.L., *On comparing classifiers: Pitfalls to avoid and a recommended approach.* Data mining and knowledge discovery, 1997. **1**: p. 317-328.

[46] LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.

[47] Laney, D., *3D data management: Controlling data volume, velocity, and variety.* META Group, 2001.

[48] Davenport, T.H. and D. Patil, *Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce.* Harvard Business Review, 2012: p. 70.

[49] Sarker, I.H., *Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions.* SN Computer Science, 2021. **2**(6): p. 420.

[50] Pires, P.B., J.D. Santos, and I.V. Pereira, *Artificial Neural Networks: History and State of the Art.* Encyclopedia of Information Science and Technology, Sixth Edition, 2024: p. 1-25.

[51] 51. Chhabra, P. and S. Goyal. *A Thorough Review on Deep Learning Neural Network*. in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*. 2023. IEEE.

[52] Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural networks, 2015. **61**: p. 85-117.

[53] Halevi, G. and H.F. Moed Dr, *The evolution of big data as a research and scientific topic: Overview of the literature.* Research trends, 2012. **1**(30): p. 2.

[54] Vignesh, P., et al. *Research in Big Data Analytics Utilizing Simulations*. in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2023. IEEE.

[55] Jebbu, A., R. Kumari, and T. Pati, *Big data analytics: Concepts and techniques*. 2008: McGraw-Hill.

[56] Saeed, I. and R. KUMAR, *Challenges and Emerging Patterns in Big Data Analytics.* Authorea Preprints, 2023.

[57] Iskamto, D., *Data science: Trends and its role in various fields.* Adpebi International Journal of Multidisciplinary Sciences, 2023. **2**(2): p. 165-172.

[58] 58. Alvarado, R.C., *Data Science from 1963 to 2012.* arXiv preprint arXiv:2311.03292, 2023.

[59] O'Regan, G., *Introduction to Data Science*, in *Mathematical Foundations of Software Engineering: A Practical Guide to Essentials*. 2023, Springer. p. 385-398.