MSc GFIS

Information Insights II

Objectives:

- Data mining model
- Data mining taxonomy
- Machine learning
- Clustering
- Association
- Classification
- Software tools,
- Myths, mistakes..

Data mining model

- The first step - in order to define which DM technique to use for the application, the goals of the study need identifying:
  - describing what has happened in the past **or**
  - predicting what will happen in the future:

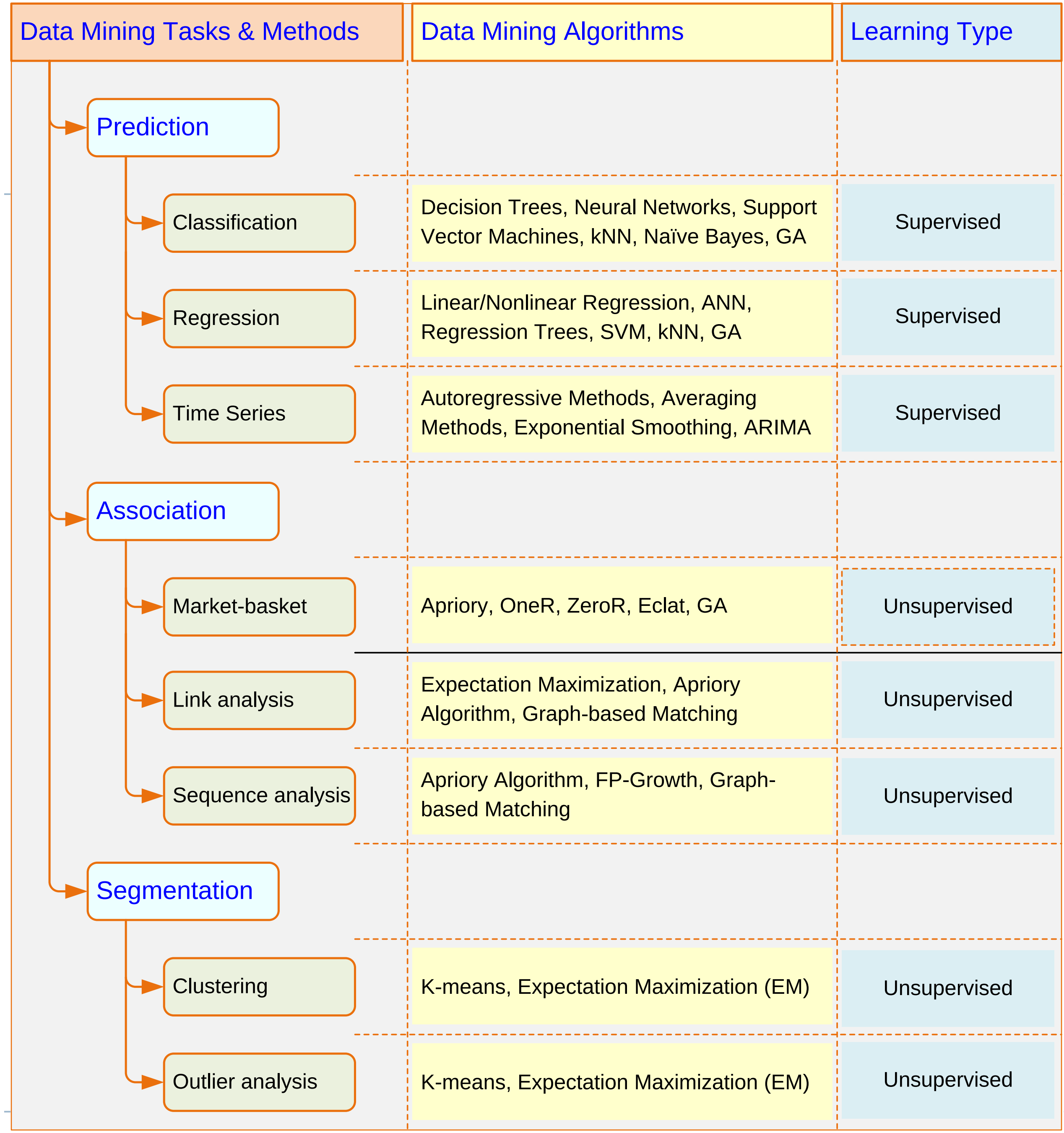
How Data Mining Works

- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

A Taxonomy for Data Mining

- A Taxonomy for Data Mining Tasks, Methods, and Algorithms



Data Mining Tasks & Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naïve Bayes, GA	Supervised
Regression	Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time Series	Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA	Supervised
Association		
Market-basket	Apriory, OneR, ZeroR, Eclat, GA	Unsupervised
Link analysis	Expectation Maximization, Apriory Algorithm, Graph-based Matching	Unsupervised
Sequence analysis	Apriory Algorithm, FP-Growth, Graph-based Matching	Unsupervised
Segmentation		
Clustering	K-means, Expectation Maximization (EM)	Unsupervised
Outlier analysis	K-means, Expectation Maximization (EM)	Unsupervised

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

Machine Learning

- Data mining is a buzzword that comes from a fascinating area of computing, called 'machine learning'.
  - There are various techniques with various implementation.
- Machine learning is about finding patterns in data, learning 'rules' that allow us to make decisions and predictions, or finding links or 'associations' between factors in situations or applications.
- Machine learning is based on self-learning or self-improving algorithms.

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

Types of Machine Learning in Data Mining

- Supervised:
  - Supervised learning is where you have input variables and an output variable and you use an algorithm to learn the mapping function from the input to the output
    - input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
      - □ Y = f(X)
  - The aim is to approximate the mapping function so that when we have new input data we can predict the output variables for that data.

- Unsupervised:
  - Unsupervised learning is where you only have input data and no corresponding output variables
    - you only have input data (X) and no corresponding output variables.
  - It is called unsupervised learning because unlike supervised learning there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

- https://www.lynda.com/Data-Science-tutorials/Machine-learning/642486/698322-4.html
  https://www.educba.com/supervised-learning-vs-unsupervised-learning/

The Data Mining Model: Describing What Happened

- **Describe** what happened
- *Descriptive techniques* are used to look for patterns
- Descriptive techniques can be of two types:
  - *Association – establishes relationships about items that occur together in a given record*
  - *Clustering – partition data into segments in which the members of data segment share similar qualities*

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things (e.g. customers)
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is no output/target variable
- In marketing, it is also known as segmentation
- Most common clustering algorithm – K-means

Clustering

Cluster Analysis for Data Mining

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Often used as an example to describe DM to ordinary people, such as the famous "relationship between diapers and beers!"
- Most common algorithm - Apriori

Association Rule Mining

- Input: the simple point-of-sale transaction data
- Output: Most frequent relationships among items
- Example: according to the transaction data…

  “Customer who bought a laptop computer and a virus protection software, also bought extended service plan 30 percent of the time" “70% of the transactions for the sale of extended service plan also purchased a laptop and virus protection”
- How do you use such a pattern/knowledge?
  - Put the items next to each other for ease of finding
  - Promote the items as a package (do not put one on sale if the other(s) are on sale)
  - Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially see and buy other items

Association Rule Mining

- Are all association rules interesting and useful?

A Generic Rule: **X ⇒ Y [S%, C%]**

X, Y: products and/or services

X: Left-hand-side (LHS)

Y: Right-hand-side (RHS)

S: Support: how often **X** and **Y** go together

C: Confidence: how often **Y** go together with the **X**

Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

Association Rule Mining

- Representative applications of association rule mining include
  - In business: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - In medicine: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)

Data Model: Predicting what will happen

- To predict what will happen means to develop a model that uses historical data to predict an outcome based on a set of input characteristics.
- Predictive techniques require the use of past history with the intent to predict future behaviour.
- DM techniques in this area serve to classify the outcome variable into predefined categories.

Data Model: Predicting what will happen

- Objective of statistical DM techniques is to find how two or more variables are related to each other.
  - **Prediction**/Forecasting estimates future values based on patterns within large sets of data, this prediction can be labeled for determining weather forecast as 'sunny' or 'rainy' . use input to produce some classification of output, e.g.
    - □ A pattern has been found already in a set of data.
    - □ Given a new set of data, you can predict which of these classes it belong too.
- **Regression** is a well-known statistical technique that is used to map data to a prediction value e.g. a real number 65°F

□ How accurate am I with this? Use classification model to give a actual measure with how close you are to the target – 95% correct

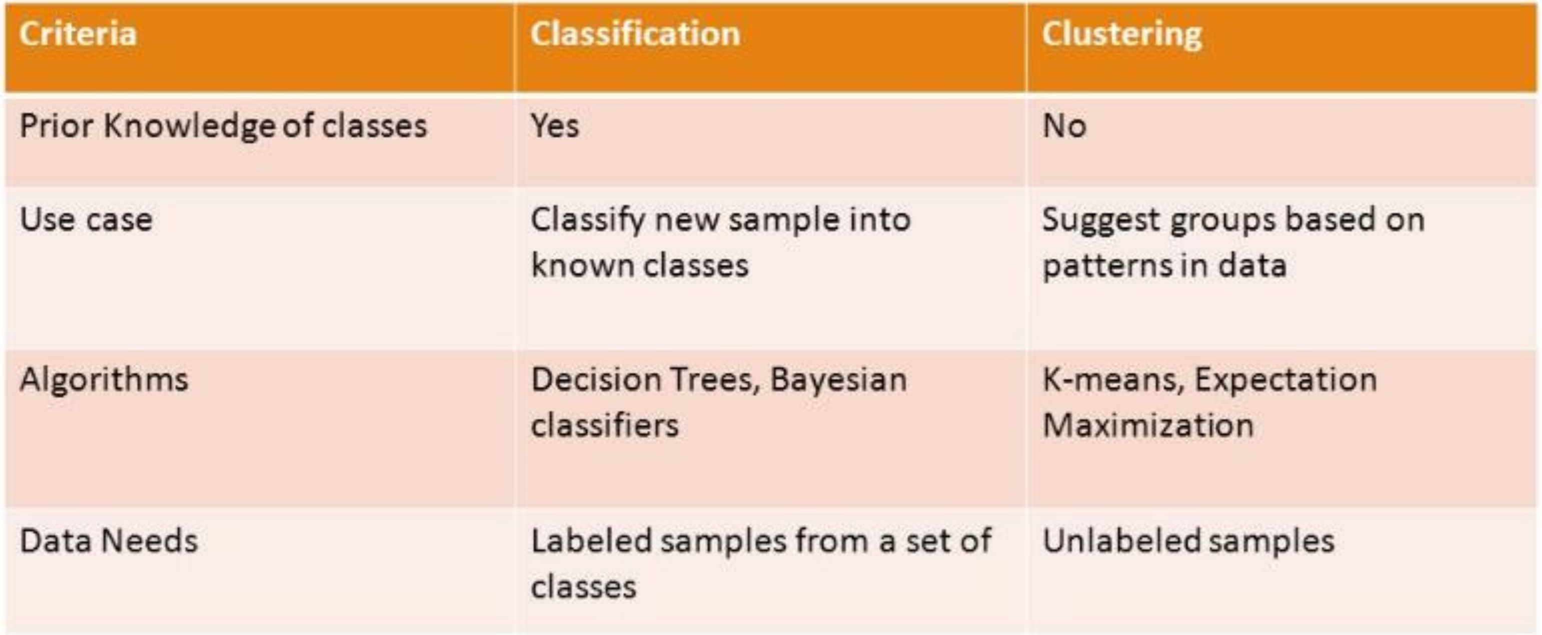
Data Mining Model - Prediction

- Most common method: Classification

  Supervised induction used to analyze the historical data stored in a database and to generate a model that can predict future behavior

- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- Most common algorithm/technique: Decision trees
- Classification versus clustering?
- Classification versus regression?

Classification vs Clustering



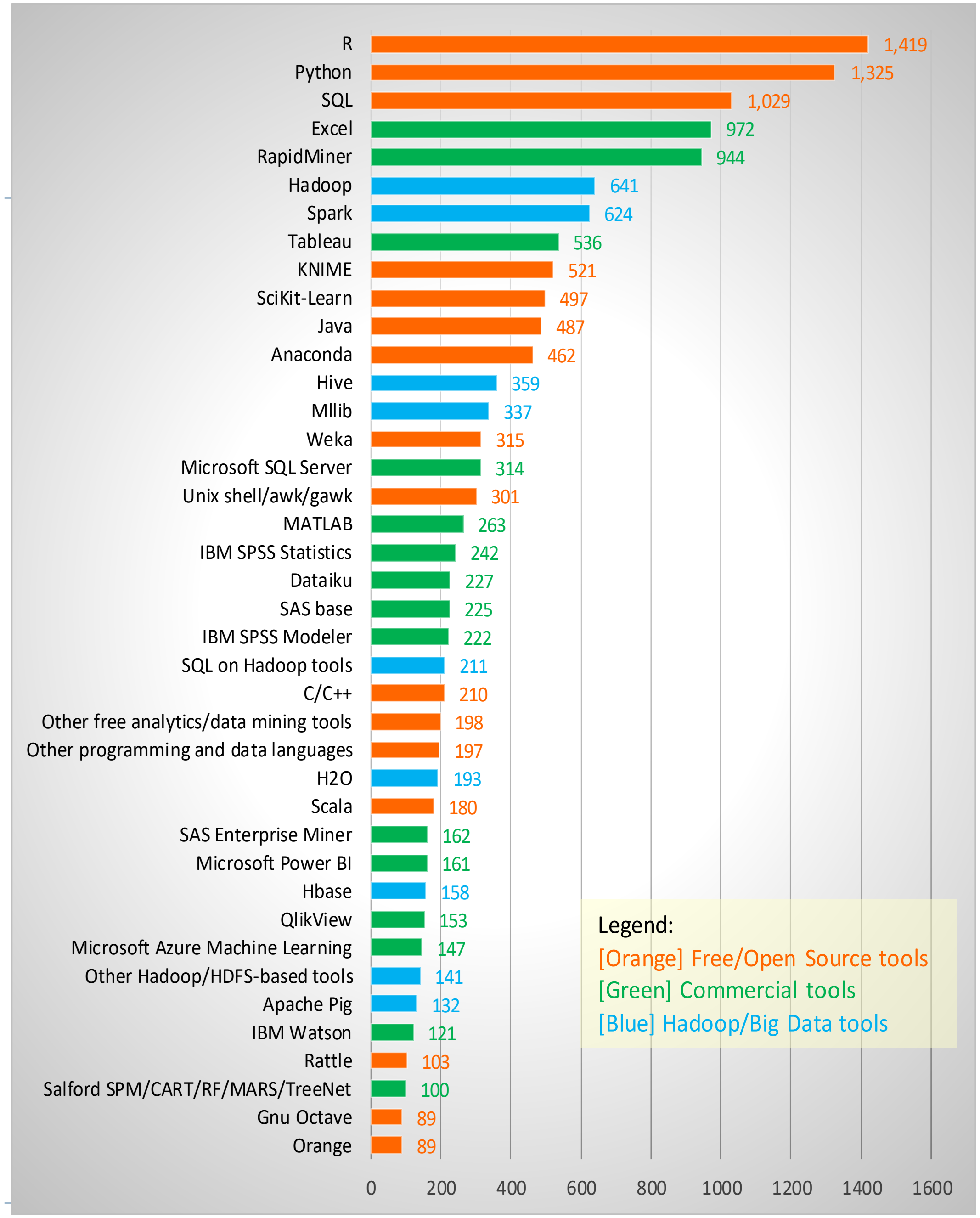
Criteria	Classification	Clustering
Prior Knowledge of classes	Yes	No
Use case	Classify new sample into known classes	Suggest groups based on patterns in data
Algorithms	Decision Trees, Bayesian classifiers	K-means, Expectation Maximization
Data Needs	Labeled samples from a set of classes	Unlabeled samples

Now Consider this Problem

- An advertising company wants to group customers based on similarities. There are no predefined labels for this group, and based on the groups on demographics and past buying behavior, they will have targeted marketing and advertising initiatives.

Data Mining Software Tools

- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS Enterprise Miner
  - Statistica - Dell/Statsoft
  - … many more
- Free and/or Open Source
  - KNIME
  - RapidMiner
  - Weka
  - R, …



Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

Data Mining Myths

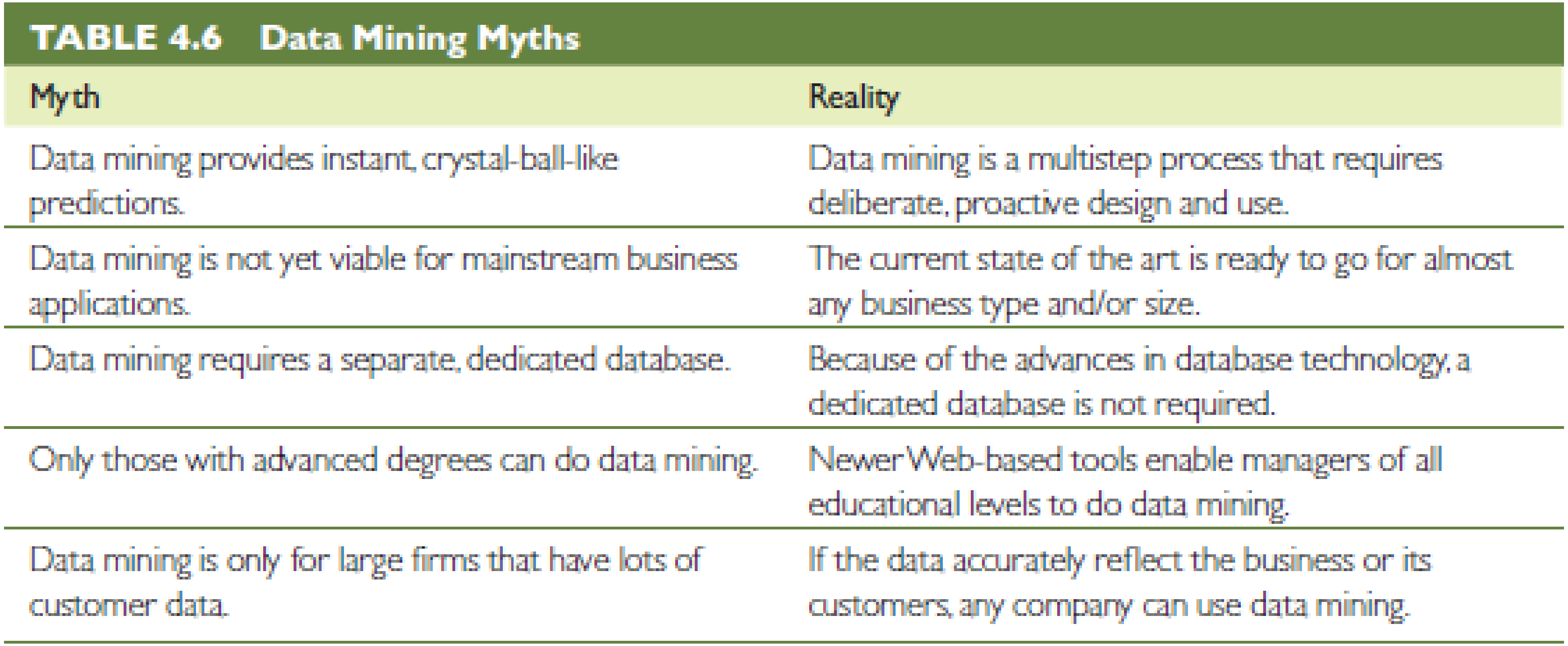


TABLE 4.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.

Data Mining Mistakes

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Beginning without the end in mind
4. Not leaving sufficient time for data acquisition, selection, and preparation
5. Looking only at aggregated results and not at individual records/predictions
6. … 10 more mistakes… in book

Sharda (2018) Business Analytics and Data Science , 4th Ed, Pearson, Chapter 4.