Past, Present, and Future of Data Science

Fundamentals of Data Science

This module provides an overview of the leading-edge methods, methodologies, and technologies used in data science and examines trends in their usage. Several key trends to watch include:

* Improvement in prediction accuracy
* Improvement in computational efficiency (e.g. through cloud analytics)
* Development of better model explanations
* Automatic enrichment of the data collected for analytics

Data science is here to stay. Since the collection of data is not likely to stop and the need for smarter and faster decision making is only poised to increase, the need for data science and its practitioners (data scientists) will only increase concomitantly in the coming years. The underlying goal of turning data (“big” data) into actionable insight will remain critical for every business.

The main goal of this last module is to provide a broad awareness of the origins of data science (the past), the current state of the art (the present), and coming trends (the future). The focus is specifically on business applications or *business analytics* as it is called.

# A Longitudinal View to Business Analytics and Data Science

Note: The discussion in this section is based on (Delen, 2020).

While the application of data science to business (business analytics) has been attracting a certain amount of “buzz” in recent years, it actually has a rather long history. It is possible to find references to corporate analytics as far back as the 1940s, during the World War II era, when more effective methods were needed to maximize production with limited resources. Many optimization and simulation techniques were developed back then.

The notion of analyzing business processes to improve productivity is even older. One example of this is the time and motion studies initiated by Frederick Winslow Taylor in the late 19th century. Another example is Henry Ford’s development of the assembly line, which gave rise to mass production thanks in part to Ford’s analytical approach for optimizing the assembly process. Analytics began to command more attention in the late 1960s when computers were starting to be used in decision support systems. Since then, analytics has evolved with the development of enterprise resource planning (ERP) systems, data warehouses, and a wide variety of other hardware and software tools and applications.

The table below shows the terminology applied to data analytics since the 1970s. During the early days of analytics, prior to the 1970s, data was often obtained from domain experts using manual processes (i.e., interviews and surveys) and used to build mathematical or knowledge-based models for solving constraint optimization problems. The idea was to do as best as possible with limited resources. Such decision support models typically fall under the term *operations research* (OR). The problems that were too complex to solve using linear or non-linear mathematical programming techniques were tackled using heuristic methods such as simulation models.

| **Decade** | **Terminology Used** | **Type** | **Subtypes** |
| --- | --- | --- | --- |
| 1970s | Decision Support Systems | Decision Support System | AI & Expert Systems, Routine Reporting |
| Enterprise/Executive IS |
| 1980s | Integrated Information Systems | On-Demand Static Reporting, Enterprise REsource Planning |
| 1990s | Executive Information Systems | Dashboards & Scorecards, Data Warehousing |
| Business Intelligence |
| 2000s | Business Intelligence | Data/Text/Web Mining, Cloud Computing, Saa5 |
| Analytics |
| 2010s | Business Analytics & Big Data | Social Network-Media Analytics, In-Memory & In-Database |
| Big Data |
| 2020s | Automation | Automated Analytics | Robotics, Smart Robo-Assistants, AI, Deep Learning, IoT, Sensors |

It’s difficult to predict what the next decade will bring and what the new analytics-related terms will be. The intervals between paradigm shifts in information systems and particularly in analytics have been shrinking, and this trend is likely to continue for the foreseeable future. Even though the practice of analytics is not new, the dramatic increase in its popularity is very new.

The explosion in recent years of “big data”—large volumes of structured or unstructured data—has given rise to new ways of collecting and storing this data along with intuitive software tools that yield data-driven business insight. This has been a boon to business professionals who are taking advantage of the many benefits it offers. In the midst of increasing global competition, a huge opportunity exists for using the power of data analytics to:

* Make better managerial decisions
* Increase revenue
* Decrease costs
* Improve product quality
* Improve the customer experience through customization
* Stop fraud before it happens

More and more companies are now preparing their employees with a knowledge of business analytics so they can drive effectiveness and efficiency in their day-to-day decision-making.

Following are several of the leading-edge enablers and enhancers of data science.

# Deep Learning

Note: The discussion in this section is based on (Delen, 2020).

At the moment, *deep learning* is one of the newest and perhaps most popular subfields within the areas of artificial intelligence (AI) and machine learning. Like other machine learning methods that preceded it, deep learning seeks to mimic the thinking processes of human beings through the use of mathematical algorithms that learn from data in ways similar to how the human brain learns.

What makes deep learning so different? Traditional machine learning algorithms such as those used in decision trees, support vector machines, logistic regression, and neural networks rely heavily on the manner in which the data is presented to them. In other words, analysts must provide the data in a format that allows those algorithms to “learn” (identify patterns in the data) and carry out their functions (i.e., prediction, classification, estimation, clustering, or association) with an acceptable degree of accuracy. Humans must manually identify or derive features that are theoretically or logically relevant for the objectives of the problem at hand and then feed these features into the most appropriate algorithm.

For example, if a marketing manager wants to use a decision tree to predict whether a particular customer will return (or churn), they must provide the algorithm with relevant information (e.g., income, occupation, educational level, records of past interactions with the company, etc.). However, the algorithm by itself cannot identify the nature of the information provided to it. It doesn’t know, for instance, whether a data item appearing in a customer survey form or on social media is a demographic characteristic or a socioeconomic attribute—or entirely something else. This is where data preparation is critically important.

While a structured, human-mediated approach generally works for abstract and formal tasks, it is much more challenging to apply it to informal, seemingly easy (at least to humans) tasks such as face identification and speech recognition. Carrying out these kinds of tasks requires a significant amount of contextual knowledge. It is not straightforward to train a machine-learning algorithm to recognize the nuanced meaning of a sentence by simply providing it with rules of grammar and semantics. The “deep” contextual knowledge about the world that is needed is not easily formalized and is certainly not explicitly available. Deep learning’s main contribution to classic machine learning methods is the ability to automatically acquire the contextual knowledge needed to accomplish these kinds of informal tasks and consequently extract advanced features that contribute to superior system performance.

To develop an intimate understanding of deep learning, it is useful to see where it fits within the artificial intelligence family of learning methods. A simple hierarchical or taxonomy-like representation may in fact provide such a holistic view. In this taxonomy, deep learning is a subdomain within a broader area called *representation learning*. In turn, representation learning techniques lie within the broader domain of machine learning, in which the emphasis is on automatically discovering system features and mapping those features to specific targets or outputs. Finally, machine learning is simply a subset of the broader discipline of artificial intelligence.

# Explainable AI

Note: The discussion in this section is based on (Delen, 2020).

In data science (business analytics, predictive analytics, machine learning, etc.), a trade-off exists between model complexity and model performance. The more complex a model is, the more accurate the prediction outcomes become. As it becomes simpler, the model becomes more transparent and interpretable—but at a cost in predictive performance. As interest in machine learning methods has grown thanks to the advent of model ensembles and deep learning, model interoperability has become a fast-growing field of research. Today, this trend is manifested as *explainable AI* and *human-interpretable machine learning*.

Data science methods and their underlying machine learning algorithms are especially effective at capturing complex relationships between input and output variables but are not nearly as effective at *explaining* what they do. In other words, the prediction models are accurate but lack transparency and explainability.

In order to mitigate this deficiency (which is also known as the “black-box syndrome”), the machine learning community has proposed several methods, most of which are characterized as sensitivity analyses. Some of these methods are global (they provide explanations based on average scores of all data samples) and others are local (they provide explanations at the level of individual samples). In the context of predictive modeling, a sensitivity analysis usually refers to an exclusive experimentation process designed and executed to discover the cause-and-effect relationships between input variables and output variables.

Some of the methods for evaluating the importance of a model’s variables are applicable to specific models or algorithms (e.g., decision trees, neural networks, or random forests) while others are model/algorithm-agnostic (i.e., they can be applied to any predictive model). Here are the most commonly used variable importance methods employed in machine learning and predictive modeling:

1. **Developing and observing a well-trained decision tree model to examine the relative discernibility of the model’s input variables:** Variables that are split closer to the root of the tree are relatively more important and make a relatively greater contribution to the prediction model.
2. **Developing and observing a rich and large random forest model, and assessing the variable split statistics:** If the ratio of a given variable’s selection into candidate counts is larger, then its importance and relative contribution are also greater. This process can be extended to the top three layers of the trees to generate a weighted average of the split statistics and can lead to a measure of variable importance for random forest models. (Note: The ratio of a given variable’s selection into candidate counts is the number of times a variable is selected as the level-0 splitter divided by the number of times it was picked randomly as one of the split candidates.)
3. **Performing a sensitivity analysis based on the perturbation of input values, where the input variables are gradually and systematically changed/perturbed one at a time and the relative change in the output is observed:** Larger changes in the output to the perturbation of a variable reflect a greater degree of importance for that variable. This method is often used in trained feed-forward neural network modeling where all of the input variables are numeric and standardized/normalized.
4. **Sensitivity analysis based on leave-one-out methodology:** This method can be applied to any type of predictive analytics method (i.e., it is predictive-model-agnostic). Because of its general applicability and ease of implementation, this sensitivity analysis method is used in several commercial and free/open source analytics tools by default and therefore is further explained below.
5. **Sensitivity analysis based on developing a surrogate model to assess the variable importance of a single record/sample using LIME or SHAP methodologies**: While the previous methods are considered global interpreters, LIME and SHAP are local because they explain the importance of variables at the sample level (as opposed to the average of all samples). Because of recent increased interest in these two methods, they are explained below.

# AutoML

AutoML stands for *automated machine learning*. The ambitious goal of AutoML is to significantly reduce or eliminate the need for experienced and skillful data scientists to build machine learning and deep learning models. Instead, an automated software system—the AutoML system—allows analysts to simply provide the labeled training data as input and receive an “optimized” model as output.

According to the experts in the field, there are a few different ways of achieving this goal. One approach is for the software system to train (and test) every possible viable type of model on the data and pick the one that works best. A refinement of this is to build one or more ensemble models that combine the other models, which sometimes (but not always) gives better results. Another technique is to optimize the hyperparameters (the specific parameters that ML techniques use) of the best model or models to train an even better model. Feature engineering (creating new features or transforming existing features to a better representation) is a valuable addition to any model training.

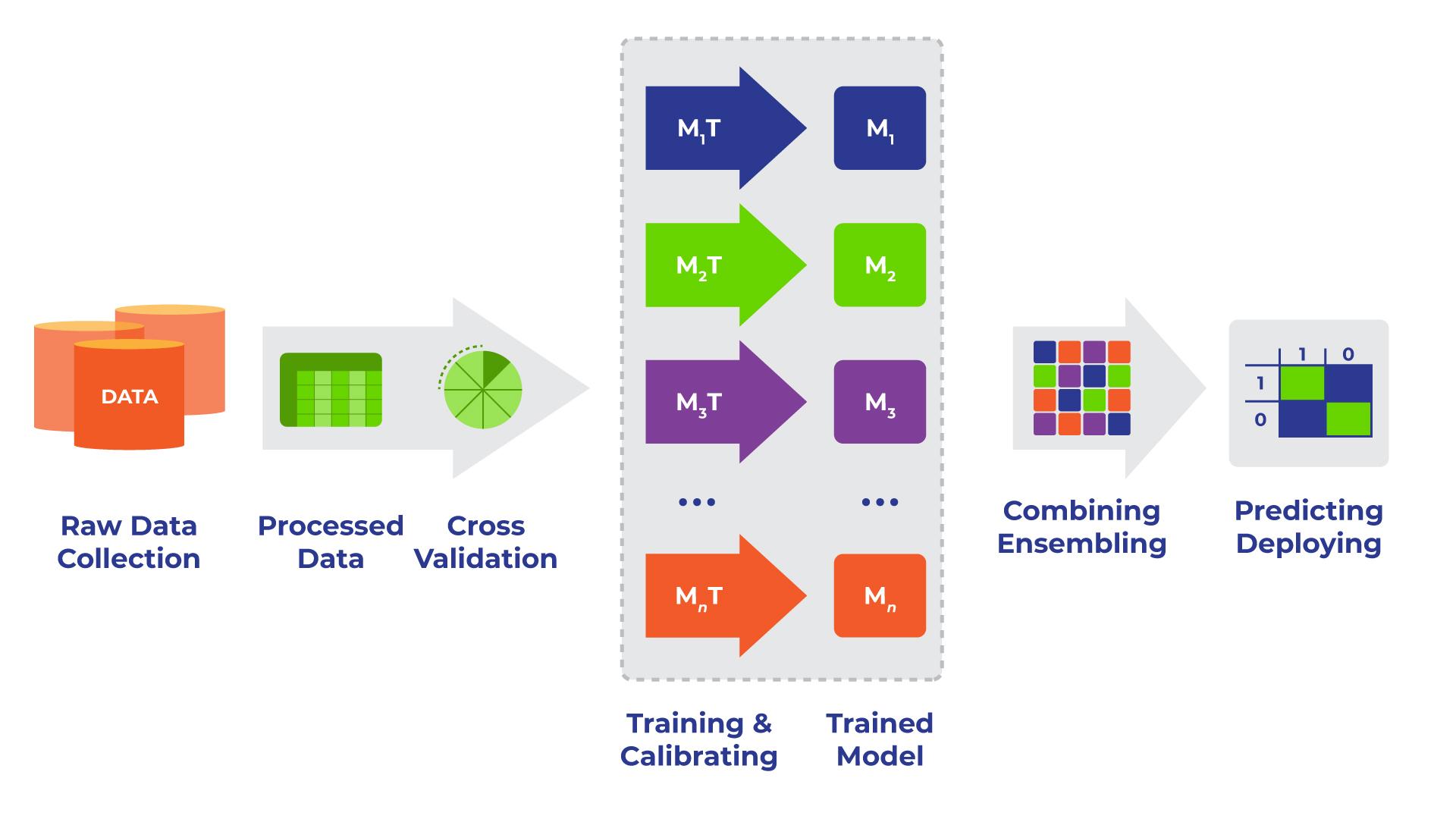
Although the idea is enticing and perhaps viable for ordinary problems with standardized data sets, AutoML is still a dream for novel and complex business problems.

# Model Ensembles

Note: The discussion in this section is based on (Delen, 2020).

*Ensemble modeling*, which is often referred to as *model ensembles* or simply *ensembles*, is a relatively new practice in data science and predictive modelling in which outcomes produced by two or more analytics models are combined into a compound output. Ensembles are primarily used for prediction modeling where the scores of two or more models are combined to produce a better prediction. This prediction can be of either a classification type (predicting a class label) or a regression/estimation type (estimating a numerical value).

Although the use of ensembles has been dominated by prediction type modeling, it can also be used for other analytics tasks such as clustering and association rule mining. In other words, model ensembles can be used for supervised as well as unsupervised machine learning tasks. Traditionally, these machine learning procedures focused on identifying and building the best possible model (often the most accurate predictor on the holdout data) from a large number of alternative model types. To do so, analysts and scientist use an elaborate experimental process that mainly relies on trial-and-error to improve each single model’s performance (defined by some predetermined metrics, e.g., prediction accuracy) to its best possible level so that the best of the models can be used/deployed for the task at hand. As shown in the following figure, the ensemble approach turns this thinking around—rather than building models and selecting the single best model to use, it proposes to build many models and use them collectively for the task they are intended to perform.



Usually researchers and practitioners build ensembles for two main reasons: better accuracy and outcomes that are more stable, robust, consistent, and reliable. Numerous research studies and publications over the past two decades have shown that ensembles almost always improve predictive accuracy for a given problem and rarely yield worse predictions than single models. Ensembles began to appear in the data mining and data analytics literature in the 1990s, motivated by the limited success obtained by earlier work on combining forecasts that dated back a few more decades. By the early-to-mid 2000s, ensembles had become quite popular and almost essential for winning data mining and predictive modeling competitions. One of the most popular examples of a competition being won by ensemble modeling is the famous Netflix prize, which was an open competition that asked researchers and practitioners to predict user ratings of films based on historical ratings. The prize was one million US dollars for a team that could reduce the root-mean-squared error (RMSE) of the then-existing Netflix internal prediction algorithm by the largest margin but no less than 10 percentage points. The winner, runner-up, and nearly all the teams at the top of the leaderboard used model ensembles in their submissions. The winning submission was the result of an ensemble containing hundreds of predictive models.

# Sensor Technologies and IoT

One of the most promising data acquisition approaches involves fully automating the complete process. This is especially appropriate for the collection of environmental data such as temperature, wind speed, and barometric pressure; physiological data such as blood pressure, heart rate, and blood sugar level; and status information such as the amount of fuel left in a tank or whether a trash receptacle is empty or full. The process of collecting data from these devices can easily be automated.

But how can we transfer the data from these devices to an application that stores it or processes it? The answer is through a concept called the *internet of things* (IoT). It’s simple: every device is connected to the internet with its own unique internet address. *Sensors* make the measurements and the data is then automatically transmitted to online applications that can analyze it or store it. The data can even be shared among devices. Rather than requiring people to manually collect, prepare, transmit, and load data into analysis tools, the IoT allows devices to do this automatically.

There are countless applications for analytics and the IoT. Doctors can monitor health data for a large number of patients and use analytics to identify trends that could impact public health. An electric power utility can monitor customer power consumption and automatically increase generating capacity when the temperature rises. Weather forecasters can use real-time weather data to make accurate predictions when conditions change from minute to minute. Traffic sensors can monitor traffic flow throughout a city and adjust traffic signal timings to optimize vehicular throughput.

Many examples also exist at the personal level. Your refrigerator can monitor how quickly you’re depleting its contents so it knows when to schedule another grocery delivery. Trash dumpsters can detect when they are full and notify the trash collection company that they need to be emptied. A fitness tracker can monitor your physical activity—walking, running, sleeping—and suggest ways to improve your health. A self-driving car—with its plethora of sensors—can detect nearby vehicles along with the car’s speed, direction, road surface conditions, etc. to steer the car and adjust its speed.

Several companies are already marketing IoT products. For instance, Smartbin™ has introduced waste receptacles for many different kinds of waste byproducts (regular trash, recyclables, waste oil, etc.) that sense when they are full. Waste disposal companies monitor these receptacles and schedule service calls when the receptacles need to be emptied. Also, Clorox recently introduced a Britta® water filter system that connects to your home Wi-Fi network, monitors its own usage, and orders its own filter replacements when necessary.

In all these examples, people do not need to interact with other people or even with machines. The machines themselves do all the “talking.”

According to recent industry research, more than 38 billion “things” will be connected to the internet in the next few years. This is on top of tablets, smartphones, and PCs. There are many reasons why IoT is growing exponentially:

1. The hardware is getting smaller, more affordable, and more powerful. The cost of actuators and sensors has decreased significantly in the last 10 years, resulting in much cheaper sensors, overall. In addition, costs associated with data processing, bandwidth, and mobile device hardware have decreased by 97 percent since the last decade.
2. Business intelligence (BI) tools are increasingly available. Today, the number of companies that offer BI tools either on their premises or in the cloud continues to increase while the costs are declining. The prevalence of BI along with BI tools
3. offering their BI tools both on premise and in the cloud at cheaper rates. Big Data and BI tools are widely available and are highly sophisticated.
4. New and interesting use cases are emerging nearly every day.

# Geospatial Analytics

Note: The discussion in this section is based on (Sharda, Delen, and Turban, 2019).

A consolidated view of the overall performance of an organization is usually represented through the visualization of data science models that provide actionable information. The information may include current and forecast values of various business factors and key performance indicators (KPIs). The presentation of these KPIs in both numeric and graphic formats can be overwhelming and can even lead to a risk of missing potential growth opportunities or overlooking problematic areas. As an alternative, many organizations prefer to use visualizations that are geographically oriented (GeoMaps) and based on traditional location data such as postal codes. Geographic visualizations can lead to location-based insight that would not be readily forthcoming using other visualization approaches. Location-based data is readily available from geographic information systems (GIS), which capture, store, analyze, and manage the data. Through the use of Integrated sensor technologies, nearly any data measurement can be supplemented with geographic location information. For instance a person with a smartphone might be walking near a clothing store when the phone’s global positioning system (GPS) triggers a visual alert that the store is having a sale. Additionally, radio frequency identification (RFID) chips can be attached to merchandise for easy tracking and identification. This capability is especially important in healthcare settings when medications and medical equipment need to be carefully monitored and tracked.

## Real-Time Location Intelligence

Many devices that we commonly use are constantly sending out their location information. Cars, buses, taxis, mobile phones, cameras, and personal navigation devices all transmit their locations thanks to network-connected positioning technologies such as GPS, Wi-Fi, and cell tower triangulation. Millions of consumers and businesses use location-enabled devices for finding nearby services, locating friends and family, navigating unfamiliar streets, tracking pets, dispatching mobile service technicians, and even participating in sports, games, and other hobbies. This has led to a surge in the use of location-enabled devices, which has concomitantly created massive databases of historical and real-time streaming location information.

The movement of people and mobile objects can be optimized by automatically processing geo-positioning data without needing any cameras to observe the movements visually. Such analysis can help determine the best layout for products or optimal routing for public transportation. The automated data collection enabled through capture of cell phone and Wi-Fi hotspot access points presents an interesting new dimension in nonintrusive market research data collection and, of course, microanalysis of such massive data sets.

# Cloud Computing and Data Science

Note: The discussion in this section is based on (Sharda, Delen, and Turban, 2019).

Another emerging technology trend of which business analytics and data science users should be aware is *cloud computing*. Wikipedia (en.wikipedia.org/wiki/cloud\_computing) defines cloud computing as “a style of computing in which dynamically scalable and often virtualized resources are provided over the Internet. Users need not have knowledge of, experience in, or control over the technology infrastructures in the cloud that supports them.” This definition is broad and comprehensive. In some ways, cloud computing is a new name for many previous, related trends: utility computing, application services, grid computing, on-demand computing, software as a service (SaaS), and even older, centralized computing with dumb terminals. But the term cloud computing originates from a reference to the internet as a “cloud” and represents an evolution of all of the previously shared/centralized computing trends. The Wikipedia entry also recognizes that cloud computing is a combination of several information technology components as services. For example, infrastructure as a service (IaaS) refers to providing computing platforms as a service (PaaS) along with all of the basic platform provisioning such as management administration, security, and so on. It also includes SaaS, which includes applications delivered through a Web browser while the data and the application programs are on a remote server “in the cloud.”

This module introduced several of the leading-edge enablers of data science. Beyond these, the future of data science is unpredictable. Regardless of the direction data science takes, the future will be assuredly fast-paced, exciting, and for practitioners, very rewarding.

# References

Delen, D. (2020). *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners, 2nd Edition,* Pearson FT Press, Upper Saddle River: New Jersey.

Sharda, R., Delen, D. and Turban, E. (2019). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support, 11th Edition*, Pearson Higher Education, ‎London, England.