Chapter 2
Data Science and Analytics

Pouria Amirian, Francois van Loggerenberg and Trudie Lang

2.1 What Is Data Science?

Thanks to advancement of sensing, computation and communication technologies, data are generated and collected at unprecedented scale and speed. Virtually every aspect of many businesses is now open to data collection; operations, manufacturing, supply chain management, customer behavior, marketing, workflow procedures and so on. This broad availability of data has led to increasing interest in methods for extracting useful information and knowledge from data and data-driven decision making. Data Science is the science and art of using computational methods to identify and discover influential patterns in data. The goal of Data Science is to gain insight from data and often to affect decisions to make them more reliable [1]. Data is necessarily a measure of historic information so, by definition, Data Science examines historic data. However, the data in Data Science can be collected a few years or a few milliseconds ago, continuously or in a one off process. Therefore, Data Science procedure can be based on real-time or near real-time data collection.

The term Data Science arose in large part due to the advancements in computational methods; especially new or improved methods in machine learning, artificial intelligence and pattern recognition. In addition, due to increasing the computational capacities through cloud computing and distributed computational models, use of data for extracting useful information even in large volume is more

P. Amirian (✉) · F. van Loggerenberg · T. Lang
University of Oxford, Oxford, UK
e-mail: Pouria.Amirian@ndm.ox.ac.uk; Pouria.Amirian@os.uk

F. van Loggerenberg
e-mail: francois.vanloggerenberg@psych.ox.ac.uk

T. Lang
e-mail: trudie.lang@ndm.ox.ac.uk

P. Amirian et al. (eds.), *Big Data in Healthcare*, SpringerBriefs in Pharmaceutical
Science & Drug Development, DOI 10.1007/978-3-319-62990-2_2

affordable. Nevertheless, the ideas behind Data Science are not new at all but have been represented by different terms throughout the decades, including data mining, data analysis, pattern recognition, statistical learning, knowledge discovery and cybernetics.

As a recent phenomenon, the rise of Data Science is pragmatic. Virtually every aspect of many organizations is now open to data collection and often even instrumented for data collection. At the same time, information is now widely available on external events such as trends, news, and movements. This broad availability of data has led to increasing interest in methods for extracting useful information and knowledge from data (Data Science) and data driven decision making [2]. With availability of relevant data and technologies, decision making procedures which previously were based on experience, guesswork or on constrained models of reality, can now be made based on the data and data products. In other words, as organizations collect more data and begin to summarize and analyze it, there is a natural progression toward using the data to scientifically improve approximations, estimates, forecasts, decisions, and ultimately, efficiency and productivity.

2.2 Methods in Data Science

Data Science is the process of discovering interesting and meaningful patterns in data using computational analytics methods. Analytical methods in the Data Science are drawn from several related disciplines, some of which have been used to discover patterns and trends in data for more than 100 years, including statistics. Figure 2.1, shows some of disciplines related to Data Science.

The fact that most methods are data driven is the most important characteristic of methods in Data Science. They try to find hidden and hopefully useful patterns which are not based on the assumption made by the data collection procedures or made by the analysts. In other words, methods in Data Science are data-driven, and mostly explore hidden patterns in data rather than confirm hypotheses which are set by data analysts. The data-driven algorithms induce models from the data. In modern methods in Data Science, the induction process can include identification of variables to be included in the model, parameters that define the model, weights or coefficients in the model, or model complexity.

Despite the large number of specific Data Science methods developed over the years, there are only a handful of fundamentally different types of analytical tasks these methods address. In general, there are a few types of analytical tasks in Data Science which can be classified as supervised or unsupervised learning.

Supervised learning involves building a model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised learning, there are inputs but no supervising output; nevertheless, we can learn relationships and structure from such data [3]. Following sections first introduce the

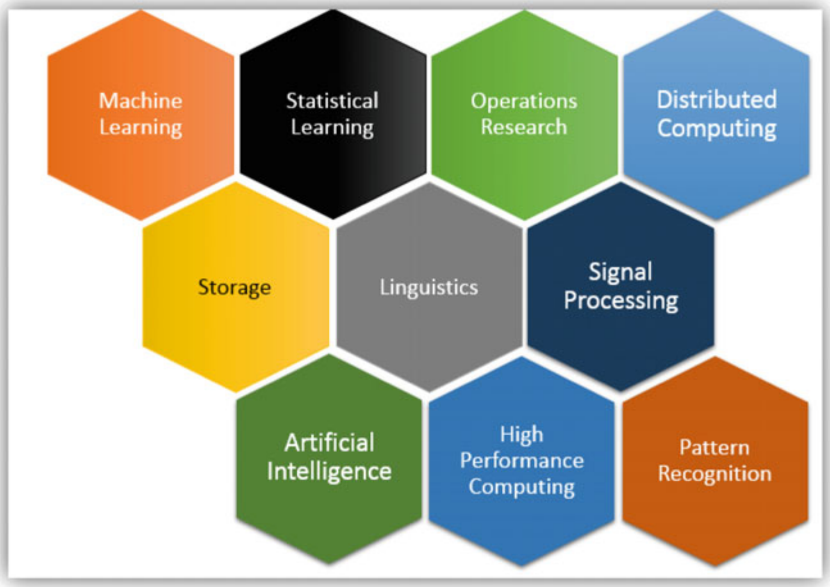


Fig. 2.1 Methods in Data Science are drawn from many disciplines

concept of supervised and unsupervised learning in more depth, and then give brief description of major analytical tasks in Data Science.

2.2.1 *Supervised and Unsupervised Learning*

Algorithms or methods in the Data Science try learn from data. Most of time, data need to be in a certain shape or structure in order to be used in a Data Science method. Mathematically speaking usually data need to be in form of a matrix. Rows (records) in the matrix represents data points or observations and columns represent values for various attributes in an observation. In many Data Science problems, the number of rows is higher than the number of attributes. However, it is quite common to see higher number of attributes in problems like gene sequencing and sentiment analysis. In some problems an attribute is called target variable since the Data Science methods tries to find a function for estimation of the target variable based on other variables in data. The target variable also can be called response, dependent variable, label, output and outcome. In this case other attributes in the data are called independent variables, predictors, features or inputs [4].

Algorithms for Data Science are often divided into two groups: supervised learning methods and unsupervised learning methods. Suppose a dataset that is collected in a controlled trail. Data in this dataset consists of attributes like id, age, sex, BMI, life style, years of education, income, number of children, and respond to drug. Consider two similar questions one might ask about a health condition of sample of patients. The first is: “Do the patients naturally fall into different groups?” Here no specific purpose or target has been specified for the grouping. When there is no such target, the data science problem is referred to as unsupervised learning. Contrast this with a slightly different question: “Can we find groups of patients who have particularly high likelihoods of positive response for a certain drug?” Here there is a specific target defined: will a newly admitted patient (who did not take part in the trial) respond to certain drug? In this case, segmentation is being done for a specific reason: to take action based on likelihood of response to drug. In other words, response to the drug is the target variable in this problem, and a specific Data Science tasks tries to find the attributes which have impact on the target variable and more importantly their importance in predicting the target value. This is called a supervised learning problem.

In supervised learning problems, the supervisor is the target variable, and the goal is to predict the target variable from other attributes in the data. The target variable is chosen to represent the answer to a question an analyst or an organization would like to answer. In order to build a supervised learning model, the dataset needs to contain both target variables as well as other attributes. After the model is created based on existing data, the model can be used for predicting a target value for a dataset without target variables. That is why sometimes supervised learning is also called predictive modeling. The primary predictive modeling algorithms are classification for categorical target variables (like yes/no) or regression for continuous target variables (numeric values). Examples of target variables include whether a patient responded to a certain drug (yes/no), the amount of a treatment (120, 250 mg, etc.), if a tumor size increased in 6 months (yes/no) and probability of increase in tumor size (0–100%).

In unsupervised learning, the model has no target variable. The inputs are analyzed and grouped or clustered based on the proximity or similarity of input values to one another. Each group or cluster is given a label to indicate which group a record belongs to.

2.2.2 Data Science Analytical Tasks

In addition to the typical statistical analysis tasks (like causal modelling) in the context of healthcare, there are several analytical tasks in healthcare from a Data Science point of view. The analytical tasks can be categorized as regression, classification, clustering, similarity matching (recommender systems), profiling, simulation and content analysis.

Regression tries to estimate or predict a target value for numerical variables. An example regression question would be: “How much will a given customer use the health insurance service?” The target variable to be predicted here is health insurance service usage, and a model could be generated by looking at other, similar individuals in the population (from health condition and records point of view). A regression procedure produces a model that, given a set of inputs, estimates the value of the particular variable specific to that individual.

While regression algorithms are used to predict target variables with numerical outcomes, classification algorithms are utilized for predicting the target variable with finite categories (classes). Classification and class probability estimation attempt to predict, for each individual in a population, which of a set of classes the individual belongs to. Usually the classes are mutually exclusive. An example classification question would be: "Among all the participants in a particular trial, which are likely to respond to a given drug?" In this example the two classes could be called "will respond" (or positive) and "will not respond" (or negative). For a classification task, the Data Science procedure produces a model that, given a new individual, determines which class that individual belongs to. A closely related task is scoring or class probability estimation. A scoring model applies to an individual and produces a score representing the probability that the individual belongs to each class. In the trial, a scoring model would be able to evaluate each individual participant and produce a score of how likely each is to respond to the drug. Both regression and classification algorithms are used for solving supervised learning problems, meaning that the data need to have target variables before the model building process begins. Regression is to some extent similar to classification, but the two are different. Informally, classification predicts whether something will happen, whereas regression predicts how much something will happen. The classification and regression compose core of predictive analytics. Nowadays, much work is focusing now on predictive analytics, especially in clinical settings attempting to optimize health and financial outcomes [5].

Clustering uses unsupervised learning to group data into distinct clusters or segments. In other words, clustering tries to find natural grouping in the data. An example clustering question would be: "Do the patients form natural groups or segments?" Clustering is useful in preliminary domain exploration to see which natural groups exist because these groups in turn may suggest other Data Science tasks or approaches. A major difference between clustering and classification problems is that the outcome of clustering is unknown beforehand and need human interpretation and further processing. In contrast, outcome of classification for an observation is a membership or probability of membership in a certain class.

The fourth type of analytical task in Data Science is similarity matching. Similarity matching attempts to identify similar individuals based on available data. Similarity matching can be used directly to find similar entities based on criteria. For example, a health insurance company is interested in finding similar individuals, in order to offer them most efficient insurance policies. They use similarity matching based on data describing health characteristics of the individuals. Similarity matching is the basis for one of the most popular methods for creating recommendations engines or recommender systems. Recommendation engines have been used extensively by online retailers like Amazon.com to recommend products based on users' preferences and historical behavior (browsing behavior and past purchases). The same concepts and techniques can be used for recommending or improving healthcare services to patients. In this case, there are two broad approaches for implementation of recommender systems. Collaboration filtering makes recommendations based on similarities between patients or services (like treatments) they used. The second class of recommendation engines can be used to make recommendations by analyzing the content of data related to each patient. In this case, text analytics or natural language processing techniques can be used on the electronic health reports/records of the patients after each visit to the hospital. Similar content types are grouped together automatically, and this can form the basis of recommendations of new treatments to new similar patients.

Profiling (also known as behavior description) tries to characterize the typical behavior of an individual, group, or population. An example profiling question would be: "What is the typical health insurance usage of this patient segment (group)?" Behavior may not have a simple description. Behavior can be assigned generally over an entire population, or down to the level of small groups or even individuals. Profiling is often used to establish behavioral norms for anomaly detection applications such as fraud detection. For example, if we know what kind of medicine a patient typically has on his/her prescriptions, we can determine whether a new medicine on new prescription fits that profile or not. We can use the degree of mismatch as a suspicion score and issue an alarm if it is too high. Also profiling can help address the challenge of health care hotspotting which is finding people who use an excessive amount of health care resources.

Simulation techniques are widely used across many domains to model and optimize processes in the real world. Engineers have long used mathematical techniques simulate evacuation planning of large buildings. Simulation saves engineering firms millions of dollars in research and development costs since they no longer have to do all their testing with real physical models. In addition, simulation offers the opportunity to test many more scenarios by simply adjusting variables in their computer models. In healthcare, simulation can be used in wide variety of applications; from modelling disease spread to optimizing wait times in healthcare settings.

Content analysis is used to extract useful information from unstructured data such as text files, images, and videos. In this context, text analytics or text mining uses statistical and linguistic analysis to understand the meaning of text, or to summarize a long text, or to extract sentiment of feedbacks (like online review for a healthcare service or center). In all these practical applications, simple keyword searching is too primitive and inefficient. For example, to detect an outbreak of a disease (like flu) from real-time feeds from a social media like twitter, with a simple keyword search it is necessary to collect and store all relevant keywords about the disease (like symptoms, treatments, etc.) and their importance. This is a manual and laborious process. Even with all relevant keywords, simple keyword search cannot offer any useful information since those keywords, can be used in other contexts. In contrast to the simple keyword search, techniques in text analytics and natural language processing can be used to filter out irrelevant contents and infer the meaning of group of words based on context. Machine learning, signal processing and computer vision also offer several tools for analyzing images and videos through pattern recognition. Through pattern recognition, known targets or patterns can be identified to aid analysis of medical images.

2.3 Data Science, Analytics, Statistics, Business Intelligence and Data Mining

2.3.1 Data Science and Analytics

In general, Data Science, analytics and even data mining are the same. Data Mining is considered the predecessor to Analytics and Data Science. Data Science has much in common with data mining since the algorithms and approaches for preparation of data and extracting useful insights from data in both, are generally the same. Analytics, on the other hand, is more focused on the methods for finding and discovering useful patterns in data and has less coverage about data preparation [6, 7]. In this case Analytics is an important part of any Data Science procedure. However, one can argue that in order to do Analytics, data need to be collected and prepared before the modelling stage. In this context, Analytics is the same thing as the Data Science. In this book, Data Science and Analytics are used interchangeably.

2.3.2 Statistics, Statistical Learning and Data Science

Data Science and statistics have considerable overlap with statisticians even arguing that Data Science is an extension of statistical learning. In fact, statistical learning and machine learning methods are highly similar and in most cases the line between these two has been blurred recently. In a nutshell, differences between Data Science and statistical learning are highly related to the mindset of analyst and their background.

However, as the core of statistical learning, statistics is often used to perform confirmatory analysis where a hypothesis about a relationship between inputs and an output is made, and the purpose of the analysis is to prove or reject the relationship and quantify the degree of that confirmation or denial using some statistical tests [8]. In this context, many analyses are highly structured, such as determining if a drug is effective in reducing the incidence of a particular disease.

In statistics, controls are essential to ensure that bias is not introduced into the model, thus misleading the interpretation of the model. Most of the time, interpretability of statistical models and their accuracy are important in understanding what the data are saying, and therefore great care is taken to transform the model inputs and outputs so they comply with assumptions of the modeling algorithms. In addition, much effort is put into interpretting the errors as well [9].

Data Science, on the other hand, often shows little concern for final parameters in the models except in very general terms. The key is often the accuracy of the model and, therefore, the ability of the model to have a positive impact on the decision making process [10]. In contrast to the structured problem being solved through confirmatory analysis using statistics, Data Science often attempts to solve less structured business problems using data that were not even collected for the purpose of building models; the data just happened to be around [1]. Controls are often not in place in the data and therefore causality, very difficult to uncover even in structured problems, becomes exceedingly difficult to identify.

Data Scientists frequently approach problems in more unstructured, even casual manner. The data, in whatever form it is found, drives the models. This is not a problem as long as the data continues to be collected in a manner consistent with the data as it was used in the models; consistency in the data will increase the likelihood that there will be consistency in the model's predictions, and therefore how well the model affects decisions.

In summary, statistical learning is more focused on models but in Data Science, data are driving the modelling procedure [11].

2.3.3 Data Science and Business Intelligence

Another field which has a considerable overlap with Data Science is Business Intelligence (BI). The output of almost all BI analyses are visualizations, reports or dashboards that summarize interesting characteristics and metrics of the data, often described as Key Performance Indicators (KPIs). The KPI reports are user-driven and case-based and determined by a domain experts to be used by the decision makers. These reports can contain simple descriptive summaries or very complex, multidimensional measures about real-time events.

Both Data Science and BI use statistics as a computational framework. However, the focus on BI is to explain what was happened in the business or what is happening in the business. Based on these observations, decision makers can take appropriate actions.

Data Science also uses historic data or data that have been collected. In contrast to BI, Data Science is focused more on finding patterns in terms of models for describing the target variable based on inputs. In other words, predictive analytics is not part of BI but is at the heart of Data Science. This leads to the fact that Data Science can provide more valuable insights for decision makers than BI can.

2.4 Data Science Process

The procedure of a Data Science project need to be structured and well defined in order to minimize the risks. As it mentioned before, the goal of Data Science is to find useful and meaningful insight from data. This goal also is goal of Knowledge Discovery in Databases (KDD) process. KDD is an iterative and interactive process

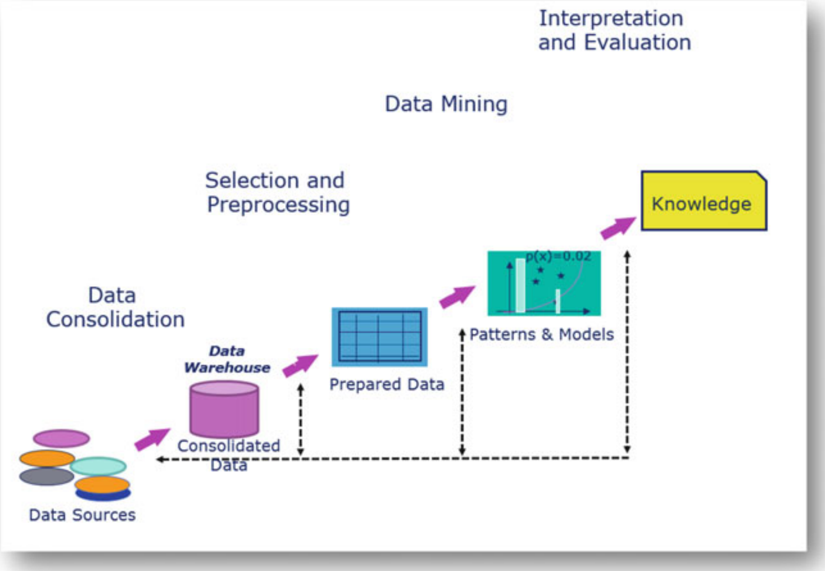


Fig. 2.2 Knowledge Discovery in Databases (KDD) Process

of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in massive databases [12]. Fortunately, both Data Science and KDD have well-defined steps and tasks for conducting projects.

Like Data Science, KDD includes multidisciplinary activities. Activities in KDD entail integrating data from multiple sources, storing data in a single scalable system, preprocess data, apply data mining methods, visualization and interpreting results. Following figure illustrates multiple steps involved in an entire KDD process.

As it illustrated in Fig. 2.2, data warehousing, data mining, and data visualization are major components of a KDD process.

2.4.1 CRISP-DM

Similar to KDD process, CRISP-DM (CRoss-Industry Standard Process for Data Mining) process defines and describes major steps in a Data Science process. The CRISP-DM is the most widely used data mining process model since its inception in the 1990s [13].

For Data Scientists, the step-by-step process provides well-defined structure for analysis and not only reminds them of the steps that need to be accomplished, but also the need for documentation and reporting throughout the process.

The documentation in Data Science process is highly valuable because of multidisciplinary nature of it; as serious Data Science projects are done in a Data Science team composed of team members with different backgrounds. In addition, the CRISP-DM provides common terminology for Data Science teams.

The six steps in the CRISP-DM process are shown in Fig. 2.3: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These steps, and the sequence they appear in the Fig. 2.3, represent the most common sequence in a Data Science project.

Data is at the core of CRISP-DM process. In a nutshell, the process starts with some questions which need domain understanding to define the scope, goal and importance of the project. Then relevant data are collected and examined to identify the potential problems in the data as well as to understand the characteristics of data. Before doing any analytics, the data need to be prepared to identify and fix problems and issues in the data. At this stage data are ready to be used in Data Science process. Data Scientists often use various models for a same analytical tasks. So based on the questions and its required performance, models are generated, evaluated and then expected effects and limitations of each model are documented. Finally, the best model based on success criteria, is going to be deployed in production environment to be used in real-world applications.

Note the feedback loops in the figure. These indicate the most common ways the typical Data Science process is modified based on findings and results of each step

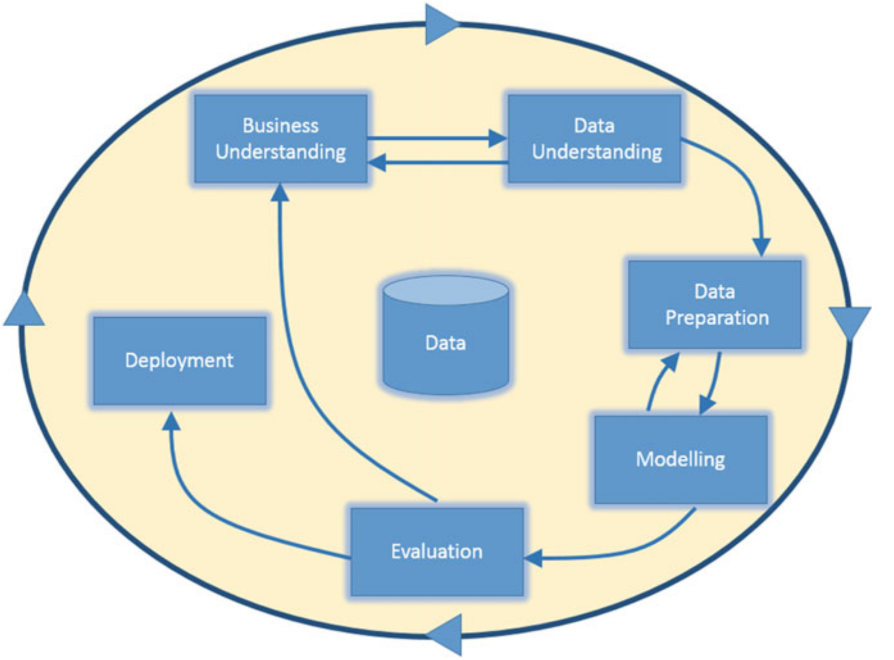


Fig. 2.3 CRISP-DM Process

during the project. For example, if process objectives have been defined during business understanding, then data are examined during data understanding. At this stage, if it turns out that there is insufficient data quantity or data quality to build predictive models and it is not feasible to collect more data with higher quality, business objectives must be redefined with the available data before proceeding to data preparation and modeling. As another example, if a built models have insufficient performance, data preparation task need to be done again to create new derived variables based on transformation on or interactions between existing variables to improve the models' performance.

2.4.2 Domain Knowledge and Business Understanding

Every Data Science project needs objectives before any data collection, preparation, and modelling tasks. Domain experts who understand needs, requirements, decisions, strategies and can understand the value of data must define these objectives. Data Scientists themselves sometimes have this expertise, although most often, managers and directors have a far better perspective on how models affect the organization [14]. In research settings, researchers always understand the problems therefore with enough domain knowledge they can define objectives of a Data Science project. Domain knowledge in this step is very important. Without domain expertise, the definitions of what models should be built and how they should be assessed can lead to failed projects that don't address the key business concerns [1, 15].

2.4.3 Data Understanding and Preparation

Unfortunately, most of data in healthcare industry are not suitable for many kinds of analytical tasks. Often 90% of the work in a Data Science project (especially in healthcare) is getting the data in a form in which it can be used in analytical tasks. More specifically, there are two major issues associated with existing data in healthcare. First, a large number of medical records are still either hand-written or in digital formats that are slightly better than hand-written records (such as photographs or scanned images of hand-written records or even scanned images of printed reports). Getting medical records into a format that is computable is a prerequisite for almost any kind of progress in current state of healthcare settings from analytical point of view [16]. The second issue related to isolated state of the existing data sources. In other words, existing digital data sources cannot be combined and linked together. These two issues can be resolved with standard electronic health records concept that is patient data in a standard form that can be shared efficiently between various electronic systems and that can be moved from one location to another at the speed of the Internet [16]. While there are currently hundreds of different formats for electronic health records, the fact that they are electronic means that they can be converted from one form into another. Standardizing on a single format would make things much easier, but just getting the data into some electronic form is the first step. Once all data are stored in electronic health records, it is feasible to link general practitioners' offices, labs, hospitals, and insurers into a data network, so that all patient data are immediately stored in a logical data store (but physically multiple data stores). At this point data is ready to be prepared for the analytical tasks.

Most analytical tasks need data in two-dimensional format, composed of rows and columns. Each row represents what can be called a unit of analysis. This is slightly different than unit of observations and measurements. Generally, data are collected from different sources with unit of observation in mind but then in the data preparation step, transformed to units of analysis. In healthcare, a unit of analysis is typically a patient, or test results for patients [17]. The unit of analysis is problem-specific and therefore is defined as part of the business understanding step of Data Science process.

Understanding data entails generating lots of plotting and examining the relationship between various attributes. Columns in the data are often called attributes, variables, fields, features, or just columns. Columns contain values for each unit of analysis (rows). For almost all Data Science methods, the number of columns and the order of the columns must be identical from row to row in the data. In data understanding step, missing values and outliers need to be identified. Typically, if a feature has over 40% of missing values, it can be removed from dataset, unless the feature conveys critical information [18]. For example, there might be a strong bias in the demographics of who fills in the optional field of "age" in a survey and this is an important piece of information. There are several ways for handling missing values. Typically, the missing values can be replaced with the average, median or even some other computations based on values of the same features in other records. This is called feature imputation. Some important models in Data Science (like tree-based ensemble models) generally can handle missing values. Similar to missing values there are some standard statistical ways for identification and handling outliers in data. It is important that identification and handling missing values and outliers are documented in this step of Data Science process. Also data type of attributes determines necessary steps in their preparation. For predictive modelling (supervised learning) it is necessary to identify one or more attributes as target variable. Identification of target variable is usually done in first step of a Data Science process (business understanding). The target variable can be numeric or categorical depending on the type of model that will be built in next step. At the end of this step, data is ready to be used for building models and testing their performance.

2.4.4 Building Models and Evaluation Metrics

Based on type of questions, analytical tasks of the Data Science project (classification, clustering, simulation, regression and so on) can be determined. For example, if there is a target variable in the question ("which participants are likely to respond to a given drug in a trial?”), the business question need to be answered with a supervised learning task. If the target variable is of type categorical, learning problem is a classification (“positive/negative response to the drug”). If the target variable is numeric, the learning problem is regression. There are many algorithms that can be used in classification or regression or in both. Each algorithm has its own assumption. Since the most widely used types of Data Science tasks in healthcare are classification and regression [19] following part of this section focuses on predictive analytics.

Regardless of algorithm for predictive analytics task, the data are split into two sets; a training set and a test set. A training set is used for building the model (for example finding the coefficients of features which best describe the variability in training set). A test set is used for evaluation of performance of the built model. Percentage of splitting of the data depends on the size of data. If the dataset is large enough, training and test sets can have a similar number of rows. Typically, 60–80% of data is used for training the model.

As it mentioned before, a predictive model is built with values of the training set. For evaluating the performance of model, a test set is used. In other words, the result of applying model building step to training set is a trained model which can be used for prediction. The test set is not used in the model building step. For evaluating the model performance, the test set is used as input for the model. After applying the model to the test set, the test set has two values (two columns) for the target variable; one is the actual value and the other is the result of applying the predictive model (predicted value). At this stage (which is called scoring), differences between actual and predicted values for test set can be used for evaluating performance of the model.

Often algorithms in Data Science have hyper parameters. Values of hyper parameters impact the model performance. In the Data Science process, finding a good value for a hyper parameter (model tuning) is done by examining different values for each hyper parameter and then calculating the model performance. Usually a range of values needs to be tested for various hyper parameters (for example using exhaustive grid search or random search). This process of building model is iterative (Fig. 2.4). This step typically results in evaluating many models based on their performance. However, the performance of the model, is one element of success criterion of the Data Science process. Hyper parameters will be discussed later in the context of a regression task.

Most of time, in Data Science projects, the success criterion is more important than the model assumptions. In other words, the determination of what is considered a good model depends on the particular interests of the project and is specified as the success criterion. The success criterion needs to be converted to a quantifiable metric so the Data Scientist can use it for selecting models. Often success criterion is a quantity for percentage of improvements in a previous modelling process like 10% improvements in prediction of malignant tumor with 30% less cost. Sometimes the success criterion is doing a task automatically using a Data Science method and the success metric for that is whether the Data Science process is computationally and economically feasible.

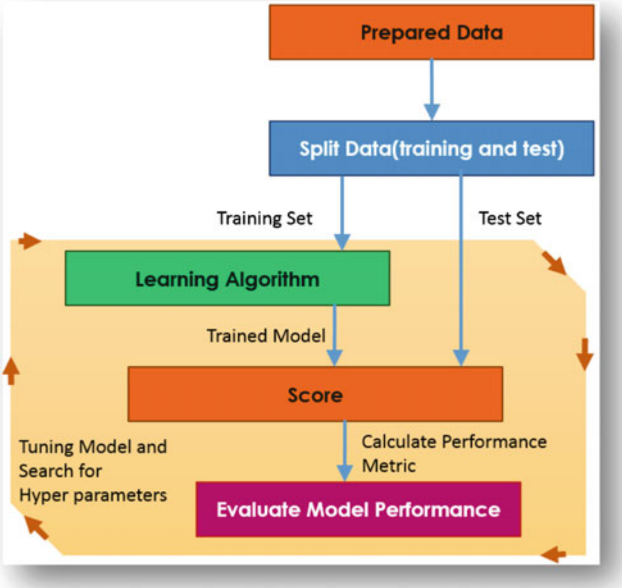
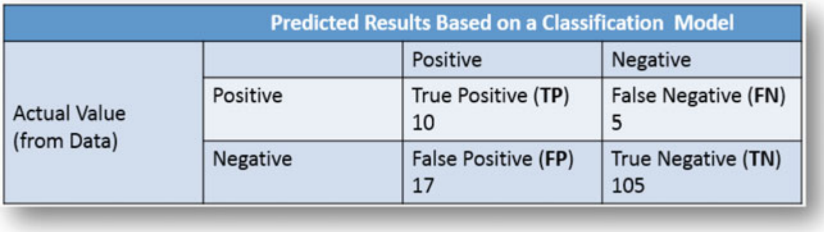

Prepared Data
Split Data(training and test)
Training Set
Test Set
Learning Algorithm
Trained Model
Score
Tuning Model and Search for Hyper parameters
Calculate Performance Metric
Evaluate Model Performance

Fig. 2.4 Building model procedure

If the purpose of the predictive model is to provide highly accurate predictions or decisions to be used by the decision makers, measures of accuracy (performance) will be used. If interpretation of the model is of most interest, accuracy measures will be used for certain models which are interpretable. In other words, not all models in Data Science have meaningful interpretations. In this case, higher accuracy models with difficult (or no) interpretation will not be included in final model evaluation if transparency and interpretation are more important than accuracy of prediction. In addition, subjective measures of what provides maximum insight may be most desirable. These subjective measures are often defined based on ease of implementation (from development time, expenses and migration of existing platforms points of view) and ease of description of the model. Some projects may use a combination of both so that the most accurate model is not selected if a less accurate but more transparent model with nearly the same acceptable accuracy is available.

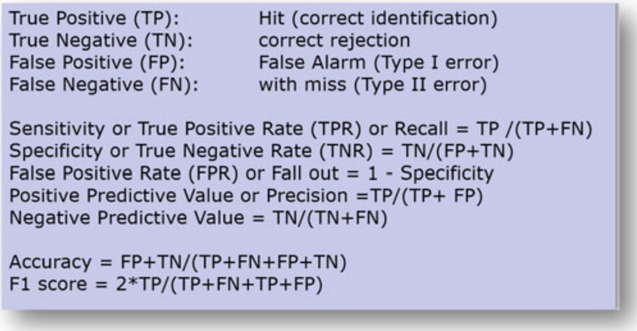
For classification problems, the most frequent metric to assess model performance is accuracy of the model which is percentage of correct classification without regard to what kind of errors are made. In addition to the classification model, another result of applying a classification model is the confusion matrix. Figure 2.5, illustrates a result of confusion matrix for detection of malignant tumor.

In this case the overall accuracy (or accuracy) of model is (10 + 105)/(10 + 5 + 17 + 105) = 84%. In addition to overall accuracy, the confusion matrix can



Predicted Results Based on a Classification Model			
Actual Value (from Data)		Positive	Negative
	Positive	True Positive (**TP**) 10	False Negative (**FN**) 5
	Negative	False Positive (**FP**) 17	True Negative (**TN**) 105

Fig. 2.5 Confusion matrix in a classification problem



True Positive (TP):	Hit (correct identification)
True Negative (TN):	correct rejection
False Positive (FP):	False Alarm (Type I error)
False Negative (FN):	with miss (Type II error)

Sensitivity or True Positive Rate (TPR) or Recall = TP /(TP+FN)
Specificity or True Negative Rate (TNR) = TN/(FP+TN)
False Positive Rate (FPR) or Fall out = 1 - Specificity
Positive Predictive Value or Precision =TP/(TP+ FP)
Negative Predictive Value = TN/(TN+FN)

Accuracy = FP+TN/(TP+FN+FP+TN)
F1 score = 2*TP/(TP+FN+TP+FP)

Fig. 2.6 Various Performance Metrics based on Confusion Matrix

provide a different measure of performance, like sensitivity, precision, fall out and F1 score. Figure 2.6, illustrates calculation of various performance measures based on the confusion matrix. The performance metrics from confusion matrix are good when an entire population must be scored and acted on. For example, for making decision about providing customized service for all hospital visitors.

If the classification model intended for a subset of the population, for example by prioritizing patients, by sorting the patients based on a model score and acting on only a portion of those entities in the selected patients, other performance metrics can be accomplished such as ROC (Receiver Operator Characteristics), and Area under the Curve (AUC). ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the ideal point for classification (a false positive rate of zero, and a true positive rate of one). The area under the ROC curve, is AUC. A larger AUC usually means higher performance. The steepness of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate. Figure 2.7, shows the ROC diagram for the classification problem.

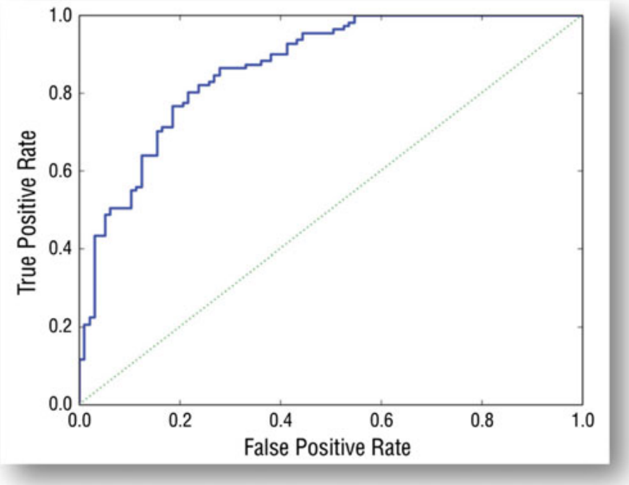


Fig. 2.7 ROC curve for tumor identification problem (AUC = 0.83)

For regression problems, the model training and scoring method are similar to the classification problems. In the following paragraphs, model building, hyper parameter identification and performance metrics calculation are described using simple linear regression and a powerful penalized linear regression model.

As mentioned before, regression problems are classified as supervised learning or predictive analytics problems. In supervised learning, the initial dataset has labels or known values for a target variable. The initial dataset is usually divided to training and test datasets for fitting the model to data and assessing the accuracy of prediction respectively. Linear regression or Ordinary Least Squares (OLS) is a very simple approach for predicting a quantitative response. Linear regression has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern approaches in Data Science, linear regression is still a useful and widely used statistical learning method. It assumes that there is approximately a linear relationship between X and Y. Mathematically, the relationship between X and Y can be wrote as Eq. 2.1. In Eq. 2.1, given a vector of features $X_T = (X_1, X_2, \ldots, X_p)$, the model can predict the output Y (also known as response, dependent variable, outcome or target) via the model:

$$Y = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.1}$$

Equation 2.1 Linear Regression Model (Ordinary Least Squares).

The term β_0 is the intercept in statistical learning, or bias in machine learning. The β_j's are unknown parameters or coefficients. The X_j are used for make prediction and are known as features, predictors, independent variables or inputs. The variables X_j can be quantitative inputs (such as measurements or observations like brain tumor size, type, and symptoms), transformations of quantitative inputs (such as log, square-root or square of observations inputs), or basis expansions, such as $X_2 = X_1^2$, $X_3 = X_1^3$, leading to a polynomial representation or dummy variables for representing categorical data (like gender Male/Female) or interactions between variables, for example, $X_3 = X_1 \cdot X_2$. Also it might seems that the model can be non-linear (by including X_1^2 or X_1^3), no matter the source of the X_j, the model is linear in the parameters [3, 9]. The OLS is widely used method for estimating the unknown parameters in a linear regression model by minimizing the differences between target values in test dataset and the target values predicted by the linear approximation function. In other words, the least squares approach chooses $\hat{\beta}_j$ to minimize the RSS (Residual Sum of Squares of errors).

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \cdot x_i \right) \right)^2 \tag{2.2}$$

Equation 2.2 Residual Sum of Squares of errors (n is the number of observations or rows in training dataset).

In Eq. 2.2, the $\widehat{y_i}$ is the predicted (estimated) value for x_i vector $(x_1, x_2, \ldots x_p)$. Residual Standard Error (RSE) is an estimate of the standard deviation of errors. More specifically, it is the average amount that the response will deviate from the true regression line. It is computed using the following formula:

$$RSE = \sqrt{\frac{1}{n-2} RSS} \tag{2.3}$$

Equation 2.3 Residual Standard Error (RSE).

The RSE is considered a measure of the lack of fit of the model to the data. If the predictions obtained using the model are very close to the true outcome values then RSE will be small, and it can be concluded that the model fits the data very well. On the other hand, if $\widehat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the training data well. The RSE provides an absolute measure of lack of fit of the model to the data. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE especially when comparing performance of the same model on different datasets. The R^2 (R squared or coefficient of determination) statistic provides an alternative measure of fit. It takes the form of a proportion and is independent of the scale of Y.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.4}$$

Equation 2.4 R^2 statistics or coefficient of determination.

In Eq. 2.4, TSS is the total sum of squares which can be calculated with Eq. 2.5.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.5}$$

Equation 2.5 Total Sum of Squares of errors (TSS).

TSS measures the total variance in the response Y, and is the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS – RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X [3]. A R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A R^2 near 0 indicates that the regression did not explain much of the variability in the response.

While easy to solve the minimization problem of linear regression, it is very prone to overfitting (high variance). In order to overcome the overfitting potential of linear regression, in penalized linear regression an additional penalty term is added to Eq. 2.1, which force the problem to balance the conflicting goal of minimizing the squared of errors and the penalty term. As an example of penalized linear regression LASSO (Least Absolute Shrinkage and Selection Operator) adds a penalty term that is called $\ell 1$ norm (Eq. 2.6). The penalty term is sum of absolute values of coefficients. The $\ell 1$ norm provides variable selection and results in sparse coefficients [20] (some of unimportant features might have coefficient value of zero).

$$Y = f(X) = \beta_0 + \sum_{j=1}^{p} \left(X_j \beta_j\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{2.6}$$

Equation 2.6 LASSO penalized linear regression model.

The LASSO algorithm is computationally efficient; calculating the full set of LASSO models requires the same order of computation as ordinary least squares however it provides higher accuracy than the OLS regression [21]. In Eq. 2.6, the λ is a hyper parameter. As it mentioned before, many algorithms in Data Science have hyper parameters. In order to find a good value for the hyper parameters, usually a range of values need to be tested for various hyper parameters (for example using exhaustive grid search or random search). Scatter plots of metrics (like errors) and values of a hyper parameter, can be useful for identifying a potential good range of hyper parameters. Figure 2.8, shows error plot for a hyper parameter for a LASSO

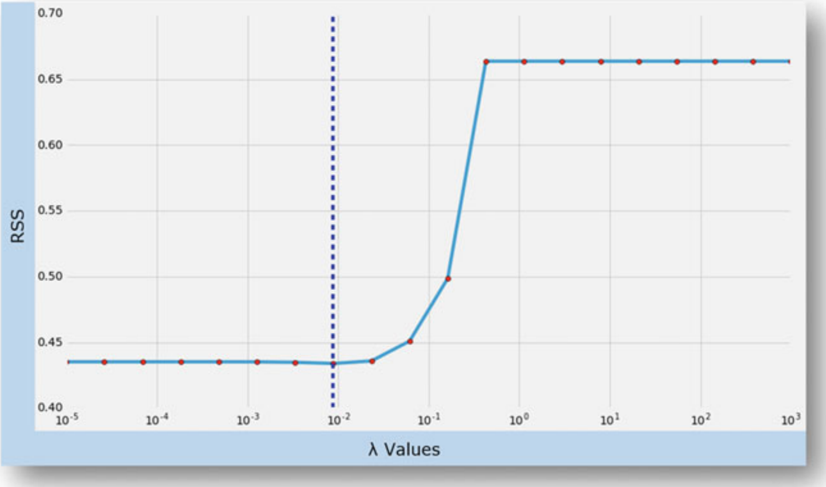


Fig. 2.8 RSS for a regression problem. In this figure, a penalized regression model (LASSO) is used for estimating (predicting) survival rate based on tumor measurements in a certain type of brain cancer. *Red dots* show the values tested for hyper parameter. *Vertical blue line* shows the minimum value for RSS and its corresponding λ

model. As you can see in the Fig. 2.8, values around 0.01 for λ results in considerably lower RSS.

2.4.5 *Model Deployment*

Once the best model based on success criteria is found (built), the final model has to be deployed in production where it can be used by other applications to drive real decisions. It is worth noting that after tuning the model (in the previous step), in building the final model, all data are used for training the model. In other words, for building the model and evaluating the model performance the whole dataset needs to be divided into training and test sets. After identification of the best model (by building various models and assessing the accuracy metrics like R^2 for regression and accuracy for classification), the whole dataset will be used for building the final model.

Models can be deployed in many different ways depending on the hosting environment. In most cases, deploying a model involves implementing the data transformations and predictive algorithm developed by the data scientist in order to integrate with an existing information management system or a decision support platform.

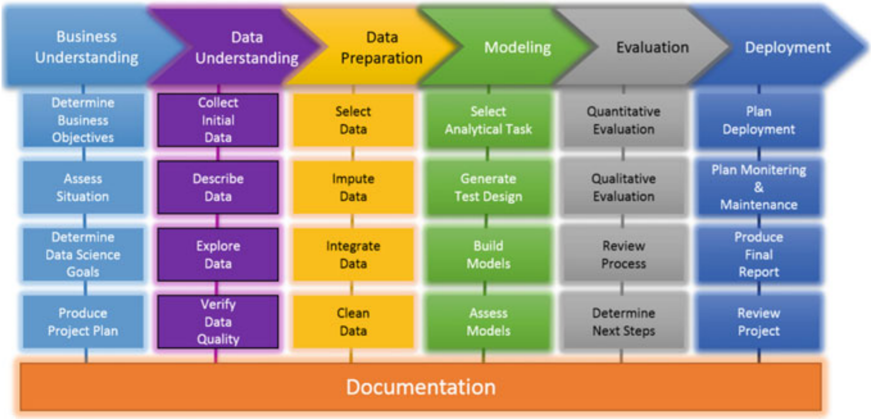


Fig. 2.9 CRISP-DM Steps and Tasks

Model deployment usually is a cumbersome process for large projects. Developers are typically responsible for deploying the model and translating the Data Science pipeline to production ready code. Since developers and Data Scientists usually work with different programming languages, development environments, coding lifecycle and mindset, the model deployment can be error prone and cumbersome. It needs careful testing procedures to prevent wrong translation of a Data Science pipeline and at the same time ensuring about non-functional requirements of the system like scalability, security and reliability.

Recently, some cloud computing providers have extended their service offering to Data Science. For example, Microsoft's Azure Machine Learning (AzureML) [22–24] dramatically simplifies model deployment by enabling data scientists to deploy their finial models as web services that can be invoked from any application on any platform, including desktop, smartphone, mobile and wearable devices. Figure 2.9 summarizes major steps and activities in CRISP-DM process.

2.5 Data Science Tools

There are a large number of programming languages, software and platforms for performing various tasks in a Data Science project. Based on Oriely's Data Science Survey 2015, Python, R, Microsoft Excel and Structured Query Language (SQL) are most widely used tools among data scientists [25]. In addition to R and Python, other popular programming languages in Data Science projects are C#, Java, MATLAB, Perl, Scala and VB/VBA. Relational databases are the most common systems for storage, management and retrieval of data (using SQL or SQL-based languages like T-SQL). Most popular relational databases in Data Science are MySQL, MS SQL Server, PostgreSQL, Oracle and SQLite. In addition

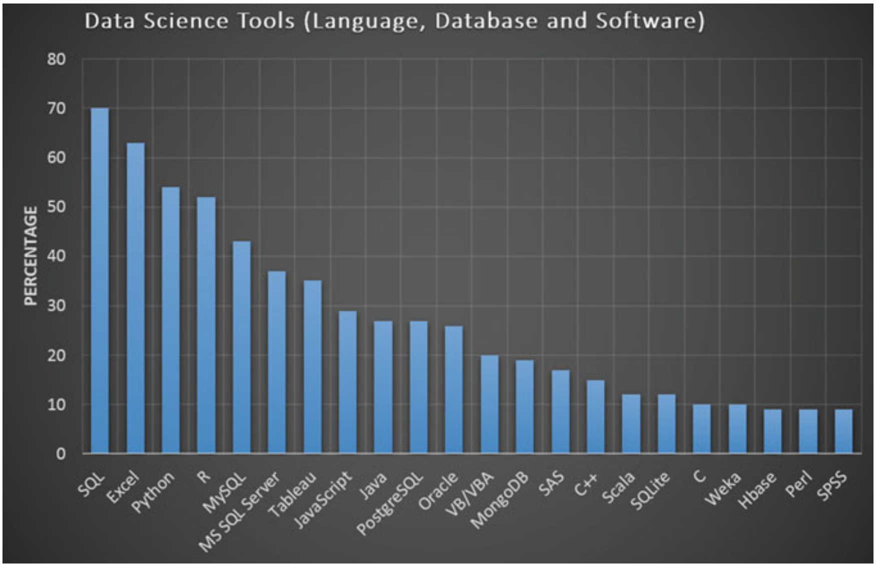


Fig. 2.10 Most widely used tools (programming languages, software and data storage solutions)

to relational databases, NoSQL systems like MongoDB, Cassandra, HBase, Redis, Vertica, Neo4j and CouchBase have been widely used especially for storage and processing semi-structured or highly connected data. Figure 2.10 shows some of the most widely used tools in Data Science.

2.6 Summary

This chapter briefly explained Data Science and its foundation in the context of healthcare. Applications of Data Science in healthcare were illustrated as analytical tasks in regression, classification, clustering, similarity matching, content analysis, simulation and profiling categories. Then Data Science process and steps were discussed in the context of CRISP-DM process. Afterwards, important concepts of success criteria and model performance were illustrated thoroughly in the context of predictive analytics and finally Data Science tools, environments and software mentioned concisely. Many experts believe that data science has the potential to revolutionize healthcare. Availability of large amounts of data from different sources is a major driving force for this revolution. The medical industry has had large amount of data from various sources such as clinical studies, hospital records, electronic health records and insurance data for generations. Today, with the growing quantity of data from traditional sources as well as rather new medical data sources like gene expression and next generation DNA sequence data and other data sources like social media, healthcare is now awash in data in a way that it has never been before. With the availability of scalable data analytics methods in Data Science, it is feasible to make sense of all the accessible data to ask important questions such as what treatments work, and for whom. There is a wide spectrum of opportunities for using Data Science methods for improving the healthcare systems; from entrepreneurs, data scientists and researchers looking to use their skills to build cutting edge services for monitoring patients, identifying high risk populations, predicting outbreaks to existing companies and organizations (including health insurance companies, biotech, pharmaceutical, and medical device companies, hospitals and other care providers) that are looking to restructure/rebuild their products and services. Next chapter is about closely related topic of Big Data.

References

1. Abbott, D.: Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Wiley (2014)
2. Provost, F., Fawcett, T.: Data Science for Business. O'Reilly Media (2013)
3. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics (2009)
4. Kelleher, J.D., Namee, B. Mac, D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics. The MIT Press (2015)
5. Hersh, W.R.: Healthcare data analytics. In: Hoyt, R., Yoshihashi, A. (eds.) Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, 6th edn, pp. 2629–2630 (2014)
6. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan. Manag. Rev. **52**, 21 (2011)
7. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. **35**, 137–144 (2015)
8. Vapnik, V.: The nature of statistical learning theory. Springer Science & Business Media (2013)
9. Gareth, J., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer (2014)
10. Waller, M.A., Fawcett, S.E.: Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J. Bus. Logist. **34**, 77–84 (2013)
11. Amirian, P., Van Loggerenberg, F., Lang, T., Varga, M.: Geospatial Big Data for Finding Useful Insights from Machine Data. In: GISResearch UK 2015 (2015)
12. Piateski, G., Frawley, W.: Knowledge Discovery in Databases. MIT press (1991)
13. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide. (2000)
14. Schutt, R., O'Neil, C.: Doing Data Science. O'Reilly Media (2013)
15. Amirian, P., Basiri, A., Van Loggerenberg F., Lang, T., Varga, M.: Geocomputation as a Service : Geospatial Big Data in Healthcare
16. O'Reilly, T., Steele, J., Loukides, M., Hill, C.: How Data Science Is Transforming Health Care Solving the Wanamaker Dilemma, pp. 1–29 (2012)
17. Amirian, P., Basiri, A., Van Loggerenberg, F., Moore, T., Lang, T., Varga, M.: Intersection of Geospatial Big Data, Geocomputation and Cloud Computing. In: 1st ICA European Symposium on Cartography, pp. 72–74 (2015)
18. Fontama, V., Barga, R., Tok, W.H.: Predictive Analytics with Microsoft Azure Machine Learning, 2nd edn. Apress (2015)

19. Madsen, L.: Data-Driven Healthcare: How Analytics and BI are Transforming the Industry. Wiley (2014)
20. Teppola, P., Taavitsainen, V.-M.: Parsimonious and robust multivariate calibration with rational function Least Absolute Shrinkage and Selection Operator and rational function Elastic Net. Anal. Chim. Acta **768**, 57–68 (2013)
21. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)
22. Amirian, P., Loggerenberg, F., Lang, T., Thomas, A., Peeling, R., Basiri, A., Goodman, S.: Using big data analytics to extract disease surveillance information from point of care diagnostic machines. Pervasive. Mob. Comput. ISSN: 1574–1192. http://dx.doi.org/10.1016/j.pmcj.2017.06.013 (2017)
23. Barnes, J.: Azure Machine Learning Microsoft Azure Essentials. Microsoft Press (2015)
24. Mund, S.: Microsoft Azure Machine Learning. Packt Publishing (2015)
25. King, J., Magoulas, R.: 2015 Data Science Salary Survey. O'Reilly (2015)

Springer

http://www.springer.com/978-3-319-62988-9

Big Data in Healthcare
Extracting Knowledge from Point-of-Care Machines
Amirian, P.; Lang, T.; van Loggerenberg, F. (Eds.)
2017, VII, 100 p. 46 illus., 45 illus. in color., Softcover
ISBN: 978-3-319-62988-9