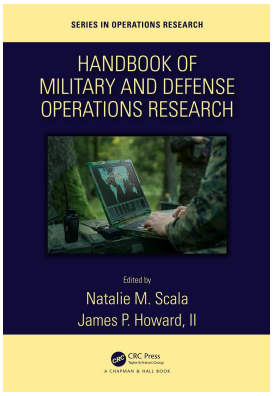

Handbook of Military and Defense Operations Research

Natalie M. Scala, James P. Howard

How Data Science Happens

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9780429467219-18
James P. Howard II
Published online on: 02 Mar 2020

Chapter 18

How Data Science Happens

James P. Howard, II

18.1 Introduction

Data science has changed how we use information (van der Aalst, 2016). Growing from roots spread across data mining, statistics, computer science, and other fields (Provost & Fawcett, 2013a), data science embraces analytic approaches applied to data generated across nearly every field. The result is a nearly unprecedented growth in the demand for data science skills among job seekers (Debortoli, Müller & vom Brocke, 2014) as every business attempts to find the truths hidden in their data. This has also expanded into government, both military and civilian sectors, as social leaders push to use data available to them (Bertot & Choi, 2013). There are many potential applications of big data to the government, regulatory, military, and law enforcement (Kim, Trimi & Chung, 2014) and the military views big data as disruptive technology Symon & Tarapore, 2015).

Despite the rush to adopt data science, there is little to tell non-practitioners and students what data science is. Data science is complicated by the lack of a single definition (Provost & Fawcett, 2013b, pp. 14–17). Everything from big data (Sagiroglu & Sinanc, 2013) and data engineering (Sharma et al., 2015) to the Internet of Things (IoT) (De Francisci Morales et al., 2016) and blockchain (Dinh & Thai, 2018) get lumped into data science.

At its core, data science allows practitioners to find deep patterns within data and turn the data into actionable information (Kaisler et al., 2013). The actionable information may be advisory, an indication of a general trend, or may be specific instructions to take. Using data visualization, advanced geospatial mapping, data mining, and predictive analytics, organizations can use data science to support their objectives and field operations.

Data science, however, also embraces the application space to generate good analysis (Jagadish, 2015). In addition to being multidisciplinary, good data science teams are also interdisciplinary, bridging the data analysis to the application space through domain expertise. To support a law enforcement mission, domain experts from criminology,

sociology, and other social science fields are necessary parts of the data science team (Lettieri et al., 2018). To manage logistics for a defense organization, a data science team needs supply chain experts and industrial engineers (Waller & Fawcett, 2013). These experts support a big data analysis by providing background, testable theory, and mission-specific guidance for data scientists coming from other fields. Without domain-specific knowledge, spurious patterns can be found in an analysis that leads analysts to unjustifiable conclusions.

In science fiction, governments use predictive analytics to identify "pre-crime" and stop crime before it happens (see, *inter alia*, Dick, 1956) or intelligent computers to fight wars (Badham, 1983). In reality, the techniques used by data scientists supporting agencies mirror the techniques those agencies already use, just ramped up in scale. At the threshold between reality and science fiction is using predictive analytics. Predictive analytics can benefit agencies by providing greater insight into their data.

However, despite the promise of big data and analytics to create intelligent machines that can solve problems, in any field, expectations must be managed. Large amounts of data are required to create statistically valid analyses. That volume of data may not be available when analyzing rare events. If analysis can be completed, it may not necessarily be valid. The discovered patterns may just reflect the biases and intuition of the data collection methods (Hajian, Bonchi & Castillo, 2016). Finally, getting to do analysis is itself a hurdle as data are often collected in forms unsuitable for analysis. In these cases, the data scientist supporting the analysis will spend more time in data preparation, cleaning the data and making it ready for use (Heer & Kandel, 2012).

This chapter will first outline the distinctions among the different ways agencies can use data science today. Second, this chapter will explain the process of how predictive analytics can be used to make guesses about the future. Third, this chapter will explain how the process is both similar to and different from existing techniques. Fourth, this chapter will analyze the legal requirements and how and when such predictive analytics can support agency activities. Finally, the chapter explains how final interpretation lies with the individual in the field. The objective of this chapter is to provide students and non-practitioners with guidance on how to use data science methods to solve problems.

18.2 Data Collection and Management

There are numerous data types that we use in data science. There are many ways to define data. One form, familiar from statistics courses, focuses on the content of the data (Stevens, 1946): count data including integers, categorical data assigned to bins, ratio measures, or numerical values. However, data science takes us to a different way of thinking about statistics. Instead of inferring on unknown populations from samples, we start looking at samples of data that are too large to look at holistically.

Much of data science revolves around "big data." Big data means different things in different contexts, and there are a number of different definitions (Morabito, 2015, pp. 23–45). Some argue that big data is defined by the number of records in the dataset or by bytes. Others define big data as data that requires certain tools or skills to analyze due to its volume and type. One strong definition defines big data based on properties and aspects of its use: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods

for its transformation into Value" (Mauro, Greco & Grimaldi, 2016). This definition captures the ideas that link big data to data science, including how it is applied to the organization. While most organizations do not produce that much data, we can use many of the same skills and techniques to analyze the data we have available to us.

"Microdata" are data that represent a single observation of something (Samarati, 2001). The speed of a vehicle at one point in time as it goes down the highway is microdata. And so is the age of the driver. This is different from the way you are used to seeing the data. News reports, briefings, and other presentations tend to focus on summary statistics, rather than the individual data elements. Summary statistics include percentages, means, and standard deviations. Microdata are the individual observations that compose those summary statistics and are a greater focus of data science than means and percentages.

When we have a lot of data, we can use statistics to make inferences about the underlying patterns. That is, we cannot establish a pattern from a single instance. A lot of data might be the speed of every vehicle traveling a certain section of the highway during an entire year. Or the ages of everyone in the organization are also a lot of data. These types of data can also be used to make inferences and predictions and collecting this information allows us to make predictions on manpower and staffing needs (Symon & Tarapore, 2015). While it is possible to use a single variable, such as speed or age, to learn something, statistics necessarily requires more information to make conclusions.

Both fingerprint (Peralta et al., 2017) and facial recognition (Samarati, 2001; Gandomi & Haider, 2015; Soyata et al., 2012) systems benefit from the big data explosion. Faster matching against existing databases is an immediate benefit as computers have grown faster, but access to more data has also improved the usability of these tools. Centralized and shared databases of fingerprints and photographs have allowed agencies to engage national and even international searches for individuals. Further, as these resources are shared, agencies can reduce costs by eliminating duplication and gain access to data they may not otherwise have acquired.

Examples like fingerprints and facial recognition provide a simple application with a single observation: a found fingerprint or a picture of someone's face. When observations of multiple variables are linked together, we can learn more information. For a vehicle's speed, we might also want information on the time of day, day of the week, weather conditions, and some sort of traffic volume metric. For age information, we might also link gender, biometric data, and other forms of data. At the very least, we might collect the date of observation, allowing us to find the age on another day. From this sort of information, we can detect patterns and perhaps empirically confirm a hypothesis. If we are luckier, we may find something we did not expect at all.

To use data in a data science analysis, the analyst must get the data to a structured format. This is often called "data cleansing" (Hernández & Stolfo, 1998). Data cleansing is the broad swath of data manipulations necessary to get into a big data analysis. This can range from date normalization, ensuring that all dates have the same format, through to extracting information from the free-form text in unstructured data.

Agencies collect information as a routine part of operations. Forms are filled out, loggers record network activity, tickets and citations are issued, and reports are written. Some of this information is collected automatically in a structured format. For instance, packages are scanned with a location, time, and data, as part of a tracking system. Those results are continuously added to a database in a structured format. However, other data, such as the contents of reports, are unstructured (Baars & Kemper, 2008). This continues to be true even if the reports are digitized.

The critical distinction between structured and unstructured data is what the nature of the data is (Park & Song, 2011). If the data is free-form, the data is unstructured. If the data are constrained, the data is structured. This can be compared to an exam. Multiple-choice and true/false questions are structured whereas short answer and essay responses are unstructured, regardless of how well they are written. If it can be put into a spreadsheet in an orderly way, it is probably structured data.

Data cleansing is not an exact science and is the first part of data preparation. The second part is about creating data aggregates. Data aggregates are new data elements created from within an existing dataset. For example, the number of packages moving through a processing center is an aggregate. We can use this sort of information to find examples of unusual spikes or delays in the supply chain. IoT devices may only report aggregates, like the average number of observations per second, rather than reporting each individual observation. In this case, the microdata is the number of observations per second, rather than "true" microdata. In many ways, the terminology is circumstance-specific.

Finally, there are outside datasets available that can be integrated with locally produced datasets to improve analysis. The World Bank provides economic aggregates that can be used to support analysis. The Global Database of Events, Language, and Tone (GDELT) provides publicly available data on news events and governments worldwide that can be used for regional analysis (Parrish et al., 2018). Analysts have combined unrelated datasets for decades but the massive merging of datasets for inference and prediction is the hallmark of data science.

While the mathematics of data cleansing and data aggregation are relatively simple, identifying what to cleanse and how to aggregate is not. Data aggregates should be justified based on an operating theory that explains the potential relevance of the data to the question. For instance, it is not immediately clear there is a reason to connect the number of soldiers in a unit with who won the World Series. Less outlandishly, there is a reason to potentially connect unemployment data for a nation to an area seeing an increase in unrest (Pervaiz, Saleem & Sajjad, 2012). Once data is cleansed and aggregated, it is possible to start analyzing to search for underlying patterns. In the next two sections, we will explore a variety of methods and show they can be used to reveal hidden information.

18.3 Data Exploration

Once data has been cleansed and prepared, we want to move into a data exploration phase. The data exploration phase allows us to examine the data and see what patterns are obvious (Tukey, 1977). In addition, a data scientist will want to have some idea of the data structure before moving too deeply into the analysis, because the structure of the data will dictate the analytic constraints. Later, in Section 18.4, we will use the scientific method to find meaning from the data. Using hypothesis testing, predictive modeling gives the analyst the ability to propose a relationship in the data and determine, within some degree of error, whether or not the relationship actually exists.

Using exploratory data analysis (EDA), we can see what information and insights are in the data, distinct from the yes and no responses from hypothesis testing. There are several types of exploration ranging from visualization, which allows us to examine the data via charts and graphs, to cluster analysis, which groups observations from

the dataset based on how similar they are to each other. This section will discuss these methods and others as part of data exploration.

Visualization is often the first step in a data exploration program (Larson & Chang, 2016). At best, a table of numbers is inconvenient to review. At worst, there is more data, both by observation count and variables in each observation, than we can meaningfully understand by looking at the numbers alone. Data visualization is a suite of techniques, beyond the ubiquitous bar charts and pie charts, we can use to better understand our data.

Modern techniques include a standardized language for data plotting, called the grammar of graphics. The grammar of graphics creates a standard approach for defining the elements of a data graph (Wilkinson, 2012). On the back end, standardized presentations of data allow the viewer to understand the meaning and intent faster. Newer tools can create animations to show the change in data over time, for instance, though animation may not be an effective data analysis tool (Robertson et al., 2008). Interactive charts routinely grace the front page of major newspaper websites.

There are a number of plotting tools within Excel. Other data visualization tools are available in statistical software like R (Wickham, 2016) and some general-purpose scripting languages, like Python (Milovanovi, 2013). Many of these tools are freely available and create publication-quality graphics. Both R and Python have vibrant user communities creating a new visualization and analytic tools that address new problems and old problems in novel ways.

Common histograms are excellent for showing distinctions provided the data has already been binned by class (Freedman, Pisani, & Purves, 2007, pp. 31–56). However, our most powerful general-purpose visualization is the box plot (McGill, Tukey, & Larsen, 1978), also known as the box-and-whiskers plot, which shows several key facts of a dataset. While some packages produce slightly different versions of box plots, the basic box plot shows the mean of a dataset, the minimum, the maximum, and how the middle 50% straddle the mean. If two box plots representing different classes are presented together, any distinctions are immediately revealed.

Heat maps provide a different type of visualization (Bojko, 2009). At its core, a heat map is a two-dimensional form of a histogram. A heat map works by creating a two-dimensional matrix across two classes of the dataset, for instance, gender and age group, and provides a statistical measure for each matrix entry. In other words, a heat map is a visual representation of a cross-tabulation. Typically, color intensity is used to demonstrate values. These can be stylized such as using color intensity on a map of the world to show the relative values of some measure across counties, commonly seen in reports and the news.

Explicit classes can exist in data. For instance, we can categorize people by gender, race, age group, or where they went to college. These classes are explicit if there is variable within the dataset that provides information about the class. Box plots, histograms, heat maps, and similar are excellent for identifying distinctions when there are a small number of classes in the data and one or two variables of interest in the data. Our real-world data often exceeds these numbers. There can be multiple classes of data and dozens or more distinct variables within a dataset. Further, the distinctions between classes may not be limited to one parameter but may instead be spread among multiple parameters simultaneously.

Cluster analysis, part of a family of techniques called unsupervised learning, captures relations within the data by discovering latent classes within the data (Jain, 2010). Latent classes will exist within the data if there is a distinct grouping within

the data, but the explicit class information is not available within it. An easily understood example comes from education and income. We know from numerous studies that more education generally leads to higher household income. However, if we do not have data on education attained, there is likely a latent grouping within the data by income. Cluster analysis can reveal these groupings.

Mathematically speaking, cluster analysis finds the closest observations to each other by measuring the distance among all data points, across n dimensions, where n is the number of variables in the dataset. Then, the clusters are found by finding the middle points necessary to minimize the total distance. Each middle point represents the prototypical element of the cluster, and its associated data points will lie around the midpoint. The process is complicated to visualize because it takes place across n dimensions.

However, the results of cluster analysis are not perfect. First, the results can be difficult to interpret. While the clusters may exist, it is probably not evident what the clusters represent. While the example given for income and education may be clustering based on income level, there is no reason that some other group affinity has been discovered by the process. Second, most algorithms require specifying the number of clusters to search for. Limiting the number of clusters to 3–4 can decrease the number of computations, but it may not necessarily reflect the real groupings within the data, leading to groups merged which may be distinct. Over-specifying the number of the clusters can have the opposite effect, drawing distinctions where none exist.

Beyond these elementary data processing techniques are several approaches to extract meaning loosely connected data. One example is geographic information system (GIS) (Star & Estes, 1990). GIS can take other data, however it was acquired or of whatever type, and link it to whatever geographic information is available. During our data explorations, natural clusters may emerge, especially when there are fine controls on what data is presented. For instance, the locations of terrorist events can be better understood by overlaying them on a map. In one sense, GIS is the digital equivalent of a map on the wall covered with pins.

The more detailed location is specified is generally better. But, for rarer events monitored at the national or global level, it may be sufficient only to acknowledge what jurisdiction an activity happened in. More complicated might be when proximity is not necessarily defined by the straight-line distance. One example might be a subway line. Two stations may only be a short drive apart but at many stops distant from each other. Understanding distance in the context of the relevant geography helps frame the understanding. For instance, consider a string of security violations at transportation facilities. If there is a sudden spike in violations in one area, it may be indicative of penetration testing as a prelude to a larger attack.

Whatever the geographic level of detail, policymakers and commanders will be interested in knowing if an area has become a hotbed of activity. It may not be possible to predict exactly where something will happen next, but it may be possible to document organized activity or just show where opportunistic activity is occurring. In that case, traditional sources and methods can be informally approached for support and patrolling increased in the area.

A different analytic framework for extracting meaning from data is called social network analysis (SNA) (Scott, 1988). This method is not related to social media, like Facebook and Twitter. However, social media can provide a framework for understanding social network analysis. SNA shows connections between individual observations in a population. It can be used to show who knows whom, who may be influencing whom,

and who is within a certain number of connections, called hops, of someone in particular. Further, like a geographic mapping of events, the individual units of analysis may be connected to other data elements of interest. SNA reveals these patterns visually providing strong clues about where we should apply formal analytic techniques.

All of these exploratory methods can advance an analysis. GIS mapping and SNA may illuminate a relationship that was not previously available but suddenly becomes obvious. But, EDA does not, by itself, provide evidence of conclusions. There is little, if any, statistical validity in EDA. However, that does not diminish the value EDA provides to informing the modeling process and communicating results. By highlighting potential relationships, the analyst can use EDA to prepare hypotheses for testing. At the most elementary, EDA can lead to new variables, generated from the relationship data, that can be used by more advanced modeling methods.

18.4 Making Predictions

The techniques described so far allow one to examine historical data. While historical evaluation is useful to provide understanding of what has already happened, the real power of data science is the power to use historical data to make actionable predictions. This process is called predictive analytics and uses techniques from supervised learning and machine learning to produce statistical and semi-statistical models (Shmueli & Koppius, 2011). Predictive analytics provides many different approaches for making predictions, but all of them result in the basic output of answering a question. From elementary logistical regression models (DeMaris, 1992) to random forests (Breiman, 2001) to the latest deep learning networks (Goodfellow, Bengio, & Courville, 2016), all of these models take the input and put it into a bin.

These methods have become common in many fields, such as retail. Amazon.com, the large online retailer, uses its vast store of information about prior purchases by all its customers to provide the suggestions in its "Related to items you've viewed …" list. This task is sometimes known as classification because it places inputs into classes based on how similar each is to an example provided previously (Kwak & Choi, 2002). Classes may be as simple as true or false for some property we are interested in. For instance, a simple true or false could ask a system to identify whether or not an identification card is likely valid or not. Or it could be a more complicated multiclass classification scheme that bins inputs into new categories, based, again, on similarity to previous examples. From an image, what kind of vehicle is detected is a multiclass problem (Wang & Gao, 2005). However, similarity may be defined in a number of ways and this is where data science occasionally becomes an art.

Throughout this chapter we have discussed discovering and capturing relationships within the data. Predictive analytics is where we apply the results of this information-gathering by creating statistical models that reflect the data. In practice, the process is similar to the process for data analysis presented in a research methods course, but there are many different algorithms we can chose from, as opposed to the standard linear model. These models come with names like logistic regression, naïve Bayes (Lewis, 1998), boosted decision trees (Roe, Yang & Zhu, 2006), random forest, and artificial neural networks, to name a few, and each has very different internal operations. However, despite any differences, they all function in the same basic way, accepting an input of observations and producing an output of a prediction. Fortunately, most

of these model algorithms are widely available in numerous statistical packages and programming languages.

Data, called training data, is presented to the modeling algorithm that has already been class-labeled. This may be because we already know the classifications or it was hand-labeled into its distinctive classes. This sample data also includes all of the variables that will define our model. These are those elements we will have access to later, to make predictions on, but will include both source data and derived data, based on the outcomes of the exploratory analysis. Regardless of which model is used, the modeling tool will "study" the data to create a prediction engine. Depending on the amount of data and which modeling algorithm is used, the process may take from a few minutes to several hours.

The result should be a tool that can make predictions, though its usefulness is still in question. We can evaluate its validity through a process called cross-validation (Shao, 1993). Like the rest of data science, there are many options, though only a couple are commonly used. The first applies the model to test data taken from the training data. To be most effective, the test should be excluded from training. Then model accuracy can be measured by seeing how many of the test examples are correctly classified. The second option applies the model back over all of the original training data. In this option, our goal is to again ask how many are correctly identified during the test phase. A third method, called multifold cross-validation divides a training dataset into k subsets and uses all but one for training and tests on the remaining. This process is repeated for each subset, with the results of each training averaged. However, this is computationally expensive as it requires each model to be calculated k times.

During the evaluation process, regardless of which cross-validation option is used, models that are two-class classifiers use two key metrics to score the results (Baeza-Yates & Ribeiro-Neto, 1999, p. 75). The first is precision, which is the identified class that are identified correctly. Mathematically, this is,

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}.$$

The second is recall, which measures the false negative rate. Mathematically, this is,

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

The metrics together tell the story of how good a model is. More complex metrics are available for multiclass models, but at their core, they evaluate the same error rates. Typically, many different models are used with the same data and the predictive outcomes for each scored to determine the model's effectiveness. Multiclass generalizations of these measures are also available.

By comparing the results of different models, we can determine which provides the best fit. Assuming a model with a good fit has been produced, it can be used to make predictions about data upon request. However, the most interesting possibility is to make the model and its results available through the organization's information technology infrastructure. In that case, the model can provide real-time predictions of events as they happen and make the results available to field operations (Howard & Beaumont, 2015).

18.5 In the Field

Automated analysis and predictions can provide a great advantage to decision-makers. However, in practice, the insights and predictions of data science only provide guesses. These can be used to supplement the analyst's work, but they cannot replace the sound judgment of a human. The methods themselves only provide information that can be used to point an individual in the right direction. The decisions of a predictive analytics system alone should not be used for an automated targeting or automated action system.

There are several reasons for this relating to statistical validity. The underlying predictive models are based on statistics and therefore statistical validity applies. The first of these is internal validity. Internal validity refers to the causal connection between the explanatory variables and the outcome variable (Brewer & Crano, 2000). For these sorts of predictive models, internal validity is how well the prediction is explained by the data we have given the model. In one sense, we can measure this statistically. But in a deeper sense, a subject matter expert can reasonably say the explanation makes sense. For instance, there is little reason to believe that the weather in Los Angeles would have any bearing on a baseball game in Boston. On the other hand, the weather in Boston would reasonably affect a game in Boston.

The second major reason is external validity, which describes how well a prediction based on the training data will hold up against members of the larger population (Mitchell & Jolley, 2012, p. 56). Like internal validity, we can estimate this statistically, but a subject matter expert can use their expertise to advance the analysis. An automated prediction algorithm will be inherently biased by the data it is trained on (Hajian, Bonchi & Castillo, 2016). That is, any underlying bias in the training data will be replicated in the model. For instance, if drivers of red cars are more likely to receive a ticket for speeding, and this was reflected in the training data, then a predictive model is more likely to flag red cars for speeding. That bias can be reduced or eliminated using statistical techniques and any good model will account for that. However, the model cannot account for cases that are insufficiently similar to those that it has already seen. This means that generally, a behavioral model that only used men as the training set is unlikely to be applicable to an instance where the subject was a woman.

Once these validity concerns are alleviated, an analyst must make sure that the application of the model makes sense. This means ensuring the results align with objectives and align with expectations. It does little good and erodes public goodwill to hear the words, "the computer says …" when following up on a model-generated action (Wihlborg, Larsson & Hedström, 2016). At the same time, the model may be producing unexpected results because the model has identified a pattern previously unknown. Knowing how to use the results of a predictive model fall squarely within the range of human judgment. Ultimately, the robots will not replace humans but will work hand-in-hand with them.

18.6 Summary

This chapter has outlined the fundamentals of data science to give students or non-practitioners the vocabulary and knowledge to understand basic analysis. This chapter provides for where data may come from and how data can be extracted from other

sources for analysis. The chapter has also provided for using EDA to gain insight. Finally, the chapter has explained how to create actionable predictions and put those prediction-making engines into the field. While the data can provide a great deal of insight and suggestions, it is ultimately up to the person in the field to determine how to interpret and use the information effectively (Lipsky & Hill, 1993).

Despite the warning, data use will grow and become more powerful as time goes on. Tools considered state of the art ten years ago are considered quaint by today's standards. But the increase in power, driven largely by the widespread availability of specialized hardware, has opened many new applications, from self-driving vehicles to advanced image recognition. As these applications expand, so will the amount of data available via resource sharing. Further, specialized methods invented to solve a problem unique to one sector will prove their value in others. A data scientist will have to keep abreast of new methods since options for application are not always obvious.

While new methods and datasets are developed, there is likely always a place for the elementary statistical methods on smaller datasets. Sometimes big data is not available or distinctions are clear enough that statistical learning methods are sufficient to make predictions. The analyst will have to make complex and only partially informed decisions when building data science models, but the good news is the practitioner will have a rich set of options available.

References

Baars, H. & Kemper, H.-G. (2008). Management support with structured and unstructured data: an integrated business intelligence framework. *Information Systems Management*, *25*(2), 132–148.

Badham, J. (1983). *Wargames*. Hollywood: United Artists.

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston: Addison-Wesley Longman Publishing Co., Inc.

Bertot, J. C. & Choi, H. (2013). Big data and e-government: issues, policies, and recommendations. In *Proceedings of the 14th annual international conference on digital government research* (pp. 1–10). dg.o '13. Québec, Canada: ACM.

Bojko, A. A. (2009). Informative or misleading? Heatmaps deconstructed. In *International conference on human-computer interaction* (pp. 30–39). Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brewer, M. B. & Crano, W. D. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16).

De Francisci Morales, G., Bifet, A., Khan, L., Gama, J., & Fan, W. (2016). IoT big data stream mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2119–2120). ACM.

Debortoli, S., Müller, O., & vom Brocke, J. (2014, October). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, *6*(5), 289–300.

DeMaris, A. (1992). Logit modeling: practical applications. Sage.

Dick, P. K. (1956, January). The minority report. *Fantastic Universe*, 4–36.

Dinh, T. N. & Thai, M. T. (2018). Ai and blockchain: a disruptive integration. *Computer*, *51*(9), 48–53.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W. W. Norton & Company.

Gandomi, A. & Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, Massachusetts: MIT press.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: from discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125–2126). ACM.

Heer, J. & Kandel, S. (2012). Interactive analysis of big data. *XRDS: Crossroads*, *19*(1), 50–54.

Hernández, M. A. & Stolfo, S. J. (1998). Real-world data is dirty: data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*(1), 9–37.

Howard, J. & Beaumont, S. (2015). Analysis as a service. In *Digital leaders* Richards, J. (Ed.) (pp. 20–21). London: BCS, the Chartered Institute of IT.

Jagadish, H. (2015). Big data and science: myths and reality. *Big Data Research*, *2*(2), 49–52. Visions on Big Data.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31* (8), 651–666.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: issues and challenges moving forward. In *2013 46th Hawaii international conference on system sciences* (pp. 995–1004). IEEE.

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78–85.

Kwak, N. & Choi, C.-H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, *13*(1), 143–159.

Larson, D. & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, *36*(5), 700–710.

Lettieri, N., Altamura, A., Giugno, R., Guarino, A., Malandrino, D., Pul- virenti, A., … Zaccagnino, R. (2018). Ex machina: analytical platforms, law and the challenges of computational legal science. *Future Internet*, *10*(5), 37.

Lewis, D. D. (1998). Naive (bayes) at forty: the independence assumption in information retrieval. In *European conference on machine learning* (pp. 4–15). Springer.

Lipsky, M. & Hill, M. (1993). Street-level bureaucracy: an introduction. In *The policy process: a reader* Hill, M. (Ed.) (pp. 381–385). New York: Routledge.

Mauro, A. D., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*(3), 122–135.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, *32*(1), 12–16.

Milovanovi, I. (2013). *Python data visualization cookbook*. Birmingham: Packt Publishing Ltd.

Mitchell, M. L. & Jolley, J. M. (2012). *Research design explained*. Cengage Learning.

Morabito, V. (2015). *Big data and analytics*. Cham, Switzerland: Springer.

Park, B.-K. & Song, I.-Y. (2011). Toward total business intelligence incorporating structured and unstructured data. In *Proceedings of the 2nd international workshop on business intelligence and the web* (pp. 1219). ACM.

Parrish, N. H., Buczak, A. L., Zook, J. T., Howard, J., Ellison, B. J., & Baugher, B. D. (2018). Crystal cube: multidisciplinary approach to disruptive events prediction. In J. I. Kantola, S. Nazir, & T. Barath (Eds.), *Advances in human factors, business management and society* (pp. 571–581). Cham, Switzerland: Springer International Publishing.

Peralta, D., Garca, S., Benitez, J. M., & Herrera, F. (2017). Minutiae-based fingerprint matching decomposition: methodology for big data frameworks. *Information Sciences*, *408*, 198–212.

Pervaiz, H., Saleem, M. Z., & Sajjad, M. (2012). Relationship of unemployment with social unrest and psychological distress: an empirical study for juveniles. *African Journal of Business Management*, *6*(7), 2557–2564.

Provost, F. & Fawcett, T. (2013a). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59. PMID: 27447038.

Provost, F. & Fawcett, T. (2013b). Data science for business: what you need to know about data mining and data-analytic thinking. Sebastopol, California: O'Reilly Media, Inc.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325–1332.

Roe, B. P., Yang, H.-J., & Zhu, J. (2006). Boosted decision trees, a powerful event classifier. In L. Lyons & M. K. Ünel (Eds.), *Statistical problems in particle physics, astrophysics and cosmology* (pp. 139–142). World Scientific.

Sagiroglu, S. & Sinanc, D. (2013). Big data: a review. In *2013 international conference on collaboration technologies and systems* (pp. 42–47). IEEE.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, *13*(6), 1010–1027.

Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109–127.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Sharma, S., Tim, U. S., Gadia, S., Wong, J., Shandilya, R., & Peddoju, S. K. (2015). Classification and comparison of NoSQl big data models. *International Journal of Big Data Intelligence*, *2*(3), 201–221.

Shmueli, G. & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 553–572.

Soyata, T., Muraleedharan, R., Langdon, J., Funai, C., Ames, S., Kwon, M., & Heinzelman, W. (2012). Combat: mobile-cloud-based compute/communications infrastructure for battlefield applications. In E. J. Kelmelis (Ed.), *Proceedings of SPIE* (Vol. 8403). Modeling and Simulation for Defense Systems and Applications VII.

Star, J. & Estes, J. E. (1990). *Geographic information systems: an introduction*. Englewood Cliffs, New Jersey: Prentice Hall.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Symon, P. B. & Tarapore, A. (2015). Defense intelligence analysis in the age of big data. *Joint Force Quarterly*, *79*(4).

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley Publishing Company, 4–11.

van der Aalst, W. (2016). Process mining: data science in action. Berlin: Springer.

Waller, M. A. & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84.

Wang, J.-H. & Gao, Y. (2005). Multi-sensor data fusion for land vehicle attitude estimation using a fuzzy expert system. *Data Science Journal*, *4*, 127–139.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Cham, Switzerland: Springer.

Wihlborg, E., Larsson, H., & Hedström, K. (2016). "The computer says no!"-a case study on automated decision-making in public authorities. In *2016 49th Hawaii international conference on system sciences* (pp. 2903–2912). IEEE.

Wilkinson, L. (2012). The grammar of graphics. In J. E. Gentle, W. K. Hrdle, & Y. Mori (Eds.), *Handbook of computational statistics* (pp. 375–414). Springer.