A Data-Driven Framework for Business Analytics in the Context of Big Data

Jing Lu(✉)

University of Winchester, Winchester SO22 5HT, UK
Jing.Lu@winchester.ac.uk

Abstract. A vast amount of complex data has been generated in every aspect of business and this enables support for decision making through information processing and knowledge extraction. The growing amount of data challenges traditional methods of data analysis and this has led to the increasing use of emerging technologies. A data-driven framework is therefore proposed in this paper as a process to look at data and derive insights in a procedural manner. Key components within the framework are data pre-processing and integration together with data modelling and business intelligence – the corresponding methods and technology are discussed and evaluated in the context of big data. Real-world examples in health informatics and marketing have been used to illustrate the application of contemporary tools – in particular using data mining and statistical techniques, machine learning algorithms and visual analytics.

Keywords: Business analytics · Conceptual modelling · Data pre-processing
Information visualisation · Data mining · Business intelligence
Analytical tools · Big data applications · Decision support

1 Introduction

The digital economy has facilitated an explosion in the data available to the world. This has affected businesses, jobs, education and healthcare. The term 'Big Data' refers to datasets so large and complex that it would be impossible to analyse them using traditional methods. Big data has been defined by Gartner in 2001 as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [3]. *Volume* is the amount of data that's created; *Velocity* refers to the speed at which new data is generated and the speed it moves around – this can help to appreciate the difference between large datasets and big data; *Variety* is the number of types of data and places that are creating it. The 3 Vs can be used to set up common ground and also point out where big data challenges and opportunities arise.

Contributing significantly to the explosive growth in volume is the Internet of Things (IoT), which is now driving big data to the next level with regards to the enabling technologies and also the future possibilities. With many forms of big data, quality and accuracy are less controllable. IBM data scientists break big data down into four dimensions by adding a 4th V as *Veracity*, referring to the trustworthiness of the data. However all this volume of fast-moving data of different variety and veracity has

A. Benczúr et al. (Eds.): ADBIS 2018, CCIS 909, pp. 339–351, 2018.
https://doi.org/10.1007/978-3-030-00063-9_32

to be turned into *Value*, which leads to the 5th V for big data [15]. Some analysis must be applied to the data as the value is not in raw bits and bytes, but rather the insights gathered from them. Big data analytics technology is generally available today, stretching from simple statistical tools to more sophisticated machine learning approaches, with deep learning among the latest trends.

Big data can deliver value in almost any area of business or society: to better understand and serve customers; to optimise their processes; to improve healthcare and security etc. The term big data remains difficult to understand because it can mean so many different things to different people. The understanding will be different depending on the perspectives from technology, business or industry [14]. Analysts at Gartner estimate that upwards of 80% of enterprise data today is unstructured. Most of it is irrelevant noise so, unless non-technical business people are clear about the kinds of data being gathered and how to make practical use of it, they will be overwhelmed. Despite the huge volume of data generated, only 0.5% of all data is currently analysed. Organisations are aware that there are growing opportunities to use big data to make better decisions, but there is a significant gap between collecting the data and making the decisions.

This paper will focus on approaches to extracting value and generating insights from complex data by using advanced data analytics techniques, e.g. predictive analytics or customer behaviour analytics. In particular, it will include details of a corresponding methodology highlighting relevant tools and technologies while illustrating sample real-world applications in health informatics and marketing.

2 Methodology

2.1 (Big) Data-Driven Framework

A data-driven framework is shown in Fig. 1 which covers aspects of the business analytics life cycle from data management, data pre-processing and integration through to data modelling and business intelligence, culminating in the essential tasks of insight management – indicated on the right-hand side of the framework. The 5 Vs of big data are linked within the framework and labelled on the left-hand side of Fig. 1. A 6th V also features across the main stages of the framework, i.e. *Visualisation*.

Data-driven means that algorithms derive key characteristics of the models from the data itself rather than from the hypotheses/assumptions of the analyst. The process starts with business knowledge and market understanding – growing business without understanding the market and competitors is risky. Domain expertise is needed to frame business goals in a way that provides value to the organisation.

Data management is the data layer of the framework, which may include the internal data environment within an organisation and the external data sources as necessary. Data exists in different formats and has various types – data or database experts are needed to identify what data is available for modelling and how that data can be accessed and normalised. Data analysts are needed later to build the model that achieves the business objectives [17]. The Computing Research "Big Data & IoT Review 2017" shows that currently *structured* data from internal and external sources

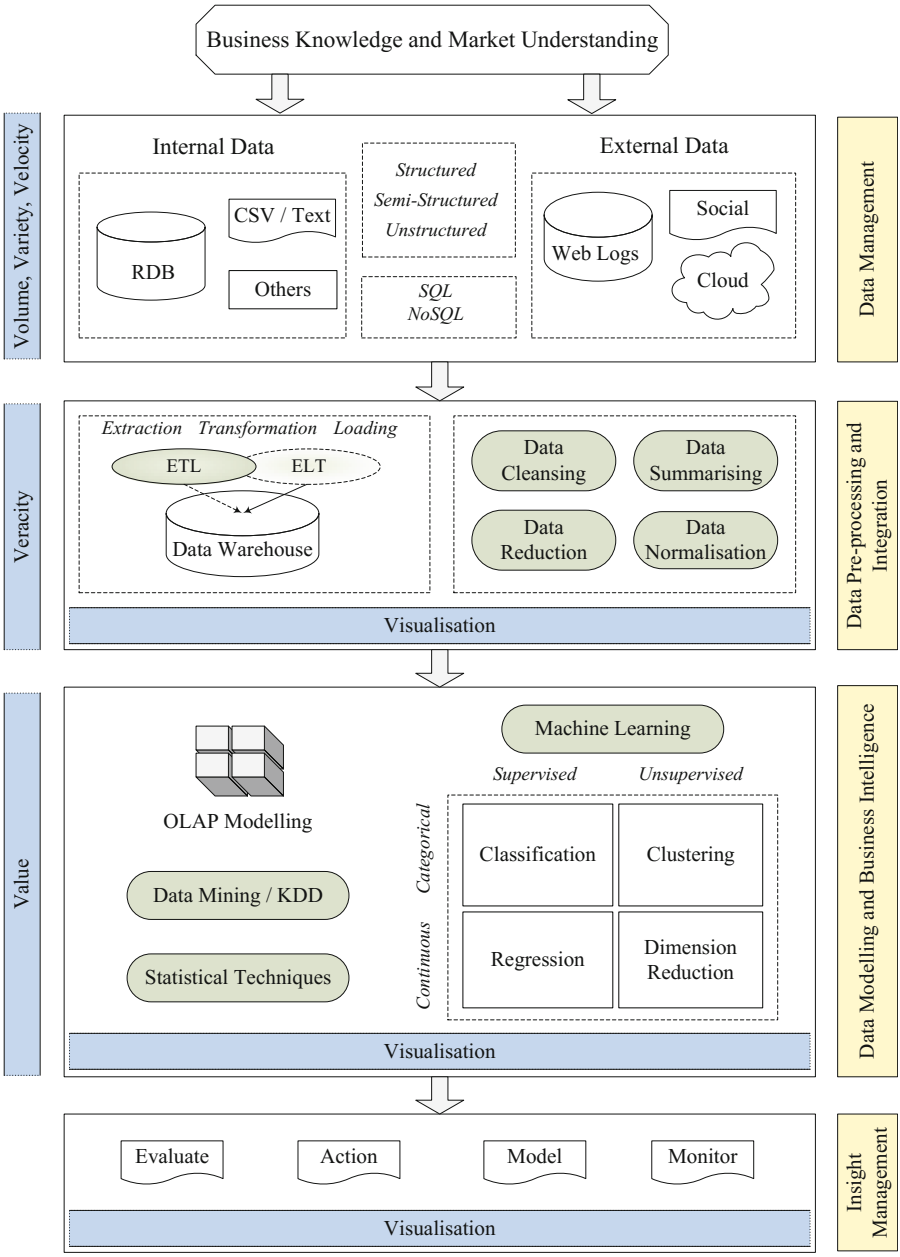

Fig. 1. Data-driven framework for business analytics.

is most likely to be analysed. The review also predicts the largest expected growth area will be the analysis of structured data from external sources [2].

Not every data management/analysis problem is best solved using a traditional RDBMS. The NoSQL database, originally referred to as *Non*-SQL and lately *Not only* SQL, is an alternative to established SQL approaches and is used in web and cloud applications [19]. NoSQL databases typically fall into one of four categories: Key-value Stores (based on Amazon's Dynamo paper), Document Databases (JSON and XML are popular formats), Column Family Stores (based on the BigTable paper from Google) and Graph Databases (inspired by mathematical graph theory).

2.2 Data Pre-processing and Integration

Big data is not just about the 'Big'. Data quality can be poor, e.g. it can contain redundant information and missing (or nonsense) values; data can be in an unsuitable format; data may need to be transformed into a form that can be used for analysis; data can be in multiple pieces, files or databases, and will need to be checked to ensure records match or are accurately collated into a single dataset. The challenge of addressing these problems can be equal to, or even greater than, that of performing the analysis. If they are not addressed, then the analysis is likely to be less valuable.

The quality of the data in terms of measurement accuracy, corruption and data entry errors can significantly affect the analytical results. If there are multiple sources collecting or hosting the data of interest (e.g. different departments in an organisation), it can be useful to compare the quality and choose the best data (or even combine the sources). However it is important to keep in mind that, for ongoing analytics, data collection is not a one-time effort. Additional data will need to be collected again in future from the same and/or other sources [18].

Data pre-processing is used to clean raw data and prepare the final dataset for use in the later stage of data modelling and business intelligence – it typically consists of data cleansing, data summarising, data reduction and data normalisation. Extraction, Transformation and Loading (ETL) will be used when data warehousing is needed: extracting data from various sources then transforming it through certain integration processes before finally loading the integrated data into a data warehouse [10].

Traditional ETL is well known throughout the database industry. In contrast, ELT has become more common recently due to the introduction of Hadoop technology and hardware/storage becoming much less expensive [8]. After extracting data instead of transforming it, first loading the raw data into (e.g.) a Hadoop Distributed File System and, when the data is needed, a schema will be built to transform the data as required. Data quality (complete and accurate) is also important during the integration stage, e.g. through capabilities such as profiling, validation, cleansing and enrichment.

2.3 Data Modelling and Business Intelligence

The success of most organisations is highly dependent on the quality of their decision making and Business Intelligence (BI) focuses on supporting and improving the decision-making process. BI can be defined as "a set of methodologies, processes, architectures and technologies that transforms raw data into meaningful and useful

information which enables effective strategic, tactical and operational insights and decision making" [11]. OLAP (On-Line Analytical Processing) is a term used to describe a technology that takes a *multi-dimensional* view of aggregate data and provides quick access to information for the purpose of advanced analysis. OLAP and SQL searches on databases are descriptive in nature and based on business rules set by the user, but don't involve statistical modelling or automated algorithmic methods.

Data mining is the essential part of *knowledge discovery in databases* (KDD) – the overall process of converting raw data into useful information and derived knowledge – one definition being "the science of extracting useful information from large datasets or databases" [6]. After data preparation, some data mining methods can be applied, e.g. considering examples in the market basket analysis area: (1) Association rules – which items are commonly bought together? (2) Sequential patterns mining – what are common purchase sequences in which customers buy products across time? (3) Classification – how likely is a customer to respond to a marketing campaign? (4) Clusters and outliers – what cohesive groups of customers do we have?

Both machine learning and statistical inference are fundamentally involved with extracting information from a dataset and for this reason there is a significant overlap between the fields. Analytics problems can be broadly segregated by whether the output is continuous or categorical (classes) and whether it is supervised (includes desired outputs) or unsupervised. Data-driven decision making is becoming the norm for analytics and business intelligence, with research showcasing to what extent supervised and unsupervised learning are underlying this.

2.4 Insight Management

Insight is information that can make a difference – insight management is about understanding information needs and then managing the way that information flows through so that it has a positive effect. Once a model is built from a data analysis perspective, it is important to evaluate the results and take action, reviewing the steps executed to construct the model.

Data is the most valuable asset of many organisations today, but only if its interpretation and impact delivers competitive advantage. Arguably the key difference between data and insight is that the latter resonates with senior stakeholders within the client business, enabling them to make decisions. The process of trying to generate insight from information is not just a matter of using algorithms to analyse data. Customers and other stakeholders increasingly participate in improving existing products and services – without customers engaging with organisations it will continue to be a struggle to develop insight. Consumers are expecting ever-increasing degrees of personalisation of goods and services they purchase – both driving the creation of data and the requirement to be able to draw actionable insights from it.

3 Tools and Technology

3.1 Overview

Technologies related to collecting, cleansing, storing, processing, analysing and visualising big data are evolving at a fast pace. Table 1 provides an outline description of some analytical tools in the broad technology categories associated with pre-processing, data mining, statistics and visualisation.

Table 1. Analytical tools and technologies.

Tools	Technologies		
	Pre-processing	Data mining and statistics	Visualisation
Excel - electronic spreadsheet program	Show missing data in pivot table	Analysis ToolPak, XLCubed, PowerPivot	Graphs/charts, PivotCharts
Alteryx - data preparation, blending and analysis	Drag and drop tools to eliminate SQL coding/formulae	Pre-packaged tools and procedures for predictive analytics	Workflow for self-service data analytics
SPSS - statistical analysis software package	Validate data, unusual cases and optimal binning	Descriptive statistics, inferential analysis, prediction	Chart builder, Graphboard Template Chooser
R - statistical computing and graphics environment	Raw => correct => consistent data	Statistics, time series, classification, clustering	Base, grid, lattice and ggplot2
iNZight - data exploration and insight generation	Quick explore => missing values	Relationships, estimation, time series	Visual Inferential Tools (VIT)
Weka - machine learning and data mining software	Discretisation, normalisation, attribute selection	Classification, clustering, association rules and sequential patterns mining	Plot, ROC, tree/graph/boundary visualiser
Tableau - data visualisation and analytics	Joins, unions, splits and pivots	Segmentation and cohort analysis, predictive analysis	Interactive and visual analytics

3.2 Data Extraction and Preparation

A study of European decision-makers' attitudes to data and analytics in modern business was conducted in 2016 (Alteryx) – 500 organisations were surveyed in all. Surprisingly this research found that while more data sources, systems and applications were being deployed, Excel spreadsheets were still used for analysis across 58% of businesses [1]. Excel can be used for data entry, manipulation and presentation, but it also offers a suite of statistical analysis functions and other tools that can be used to run descriptive statistics and to perform inferential statistical tests. Even if using alternative analytical software, Excel is often helpful when preparing data for processing by those packages.

Alteryx is a tool especially made to extract, transform and load data. Its key capabilities for data preparation include: connecting to and cleansing data; improving data quality; offering repeatable workflow design to assist with data integrity. Alteryx will be used in the health informatics application below to illustrate the extraction of data from a relational database into a dimensional data warehousing model.

3.3 Statistical Techniques

IBM SPSS was originally a widely used program for statistical analysis in the social sciences. It is now used by market researchers, health researchers, survey companies, government, education researchers, marketing organisations, data miners and others. An SPSS Python plug-in has been developed which connects SPSS with Python and thus makes everything in it available to SPSS and conversely. IBM SPSS Modeler is a data mining and text analytics software application used to build predictive models and conduct other analytical tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming [7]. The IBM SPSS Direct Marketing option enables advanced analysis of customers or contacts, potentially improving marketing campaigns and maximising return on investment.

R is a language and environment for statistical computing and graphics, with RStudio providing a user-friendly interface to analyse and manipulate data (https://www.r-project.org). R is commonly used for big data management and analysis – it is widely accepted by the data science area and has a very active support community. Developed using R, iNZight can generate insights into real-world data by producing graphs and summaries through statistical analysis. The iNZight desktop software and iNZight Lite web-based version are comparatively simple menu-driven (point and click) systems which are gaining popularity.

3.4 Machine Learning and Visual Analytics

Weka (Waikato Environment for Knowledge and Analysis) is open source software which offers a wide range of statistical inference and machine learning algorithms [5], primarily for data pre-processing, classification, regression, clustering, association rules, sequential patterns mining and visualisation. It can be applied to real-world problems and also used to analyse big data. Weka's main user interface is Knowledge Explorer, which features several panels giving access to key components of the

workbench – e.g. the Classify panel can apply classification and regression algorithms to estimate the accuracy of predictions and visualise models through decision trees.

As shown in Fig. 1, visualisation is an important aspect of data pre-processing and integration, data modelling and business intelligence as well as insight management. Tableau Software (https://www.tableau.com) supports an integrated iterative approach that combines data analysis with data visualisation and human interaction. It provides a collection of interactive visualisation products designed for business intelligence. Advanced functionalities include cohort analysis via drag-and-drop segmentation, what-if analysis of scenarios and predictive analytics using (e.g.) forecasting models – an R plug-in allows integration with other platforms and handles statistical needs.

4 Applications

4.1 Health Informatics

A case study from breast cancer research considered one of the electronic patient records systems within a University Hospital which contained over sixteen thousand patient records [12]. The objectives defined for this case study and the follow-on investigation [13] were as follows: (1) evaluate emerging database technologies and analytical tools, especially the application of data mining and visual analytics within the healthcare domain, (2) test the ability of selected analytical software to derive useful information from the extracted and anonymised cancer patient records, and (3) showcase the utilisation of tools and technologies to visualise the analytical results in order to maximise insight, evaluating outcomes in conjunction with domain experts.

Data cleansing was carried out through SQL, Alteryx and Weka. When patient records with errors or missing information are excluded, associated records from other workflows would be removed from corresponding data tables. For instance, where patient details were incomplete and excluded from the patients dataset, records for these patients could then also be excluded from the diagnosis events dataset. The Alteryx workflow in Fig. 2 removes duplicate records relating to the same diagnosis so that the event table only contains one record per event. It also shows the filtering of records to remove diagnosis events prior to 1960.

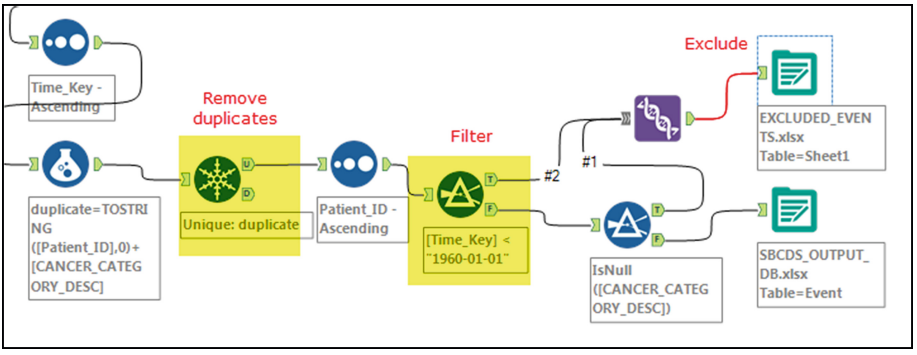

Fig. 2. Integration tool filtering records – Alteryx.

The software selected for visual analytics was Tableau, which presents information using a variety of chart types – its ability to support the analysis and visualisation of multi-dimensional models made Tableau an appropriate choice. Dashboards are typically used as a means of displaying live data where each dashboard is a collection of individual indicators, designed in such a way that their significance can be understood intuitively. Figure 3 gives an example of a diagnosis dashboard from Tableau which contains five charts each demonstrating a different situation [13].

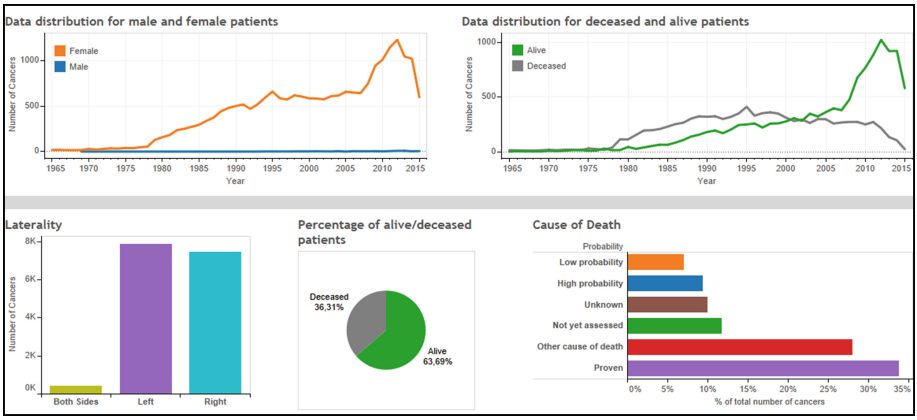

Fig. 3. Visual analytics dashboard – Tableau.

Several data mining techniques were implemented for the case study by using Weka with the cancer patient records. In particular, Generalized Sequential Patterns mining has been used to discover frequent patterns from disease event sequence profiles based on separate groups of living and deceased patients. A directed acyclic

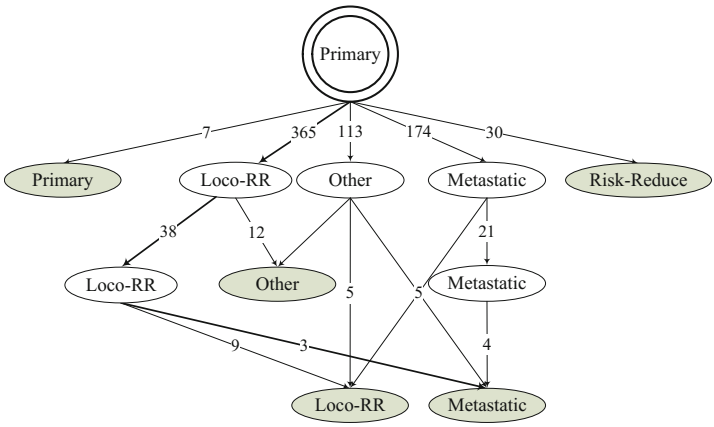

Fig. 4. SPG for maximal sequential patterns when *minsup* = 0.5% (alive patients).

Sequential Patterns Graph (SPG) represented the maximal sequential patterns found from subsets of the breast cancer data [12]. Figure 4 shows the SPG for alive patients only, when the minimum support threshold is 0.5%, where it can be seen that nodes of SPG correspond to elements (or disease events) in a sequential pattern and directed edges are used to denote the sequence relation between two elements.

4.2 Direct Marketing in Banking

This case study is based on a common business scenario: a bank was interested in targeting marketing material more effectively and aimed to use the data to build a predictive model that could identify customers likely to accept a direct marketing offer over the telephone. Typically there is some cost associated with making an offer, such as call centre employees, design and printing, or perhaps even customer email fatigue. This cost motivates the construction of the model which seeks to minimise the number of people contacted who are unlikely to accept the offer [16].

"Bank Customer" used here is a representative dataset from Github (2017) – there are 32,127 instances and 21 attributes with 20 predictor variables (features) and one response variable [4]. The features are a mix of customer demographics, customer marketing history and financial services market indicators.

A key part of data preparation is creating transforms of the dataset, such as rescaled attribute values. For the classification shown below, certain external factors were eliminated. Partly assisted by using the 'BestFirst' algorithm in Weka, the following variables were selected for the predictor set: *length* (in seconds of the call), *prev.sales* (number of sales to this customer in the past) and *age* (age of customer). The outcome status is either 'Yes', i.e. a sale was made during the call, or 'No'.

During the pre-processing stage, the *length* variable has been divided into 5 groups: <0.5 h, 0.5–1 h, >1–1.5 h, >1.5–2 h, >2 h. And customer age has been divided into 3 groups of <42 years, 42–67 years and >67 years. Several of the classification algorithms have been evaluated using Weka and compared according to a number of measures: (1) accuracy, sensitivity and specificity, (2) *n*-fold cross validation and (3) receiver operating characteristic. In this case study, a predictive model was created by applying the Weka Decision Tree J48 algorithm to the prepared banking customer dataset.

The complete decision tree is shown in Fig. 5 with seven rectangular boxes representing the leaf nodes (i.e. classes). Each internal node represents a test on the variable and each branch represents an outcome for the test. There are two numbers in parentheses in the leaf nodes: the first shows how many customers reached this outcome and the second shows the number for whom the outcome was not predicted.

Cluster analysis is a data mining technique used to group instances of data based on a similar metric. The following example demonstrates the method using the banking customer dataset and the most popular clustering algorithm, *k*-means. The dataset is loaded into Weka and the 'simplekmeans' algorithm is selected – the same three variables are considered: *length*, *prev.sales* and *age*. Three classes were selected and the output is shown in Fig. 6. Clustering requires the user to have sufficient expertise in the domain, as the number of classes has to be entered manually.

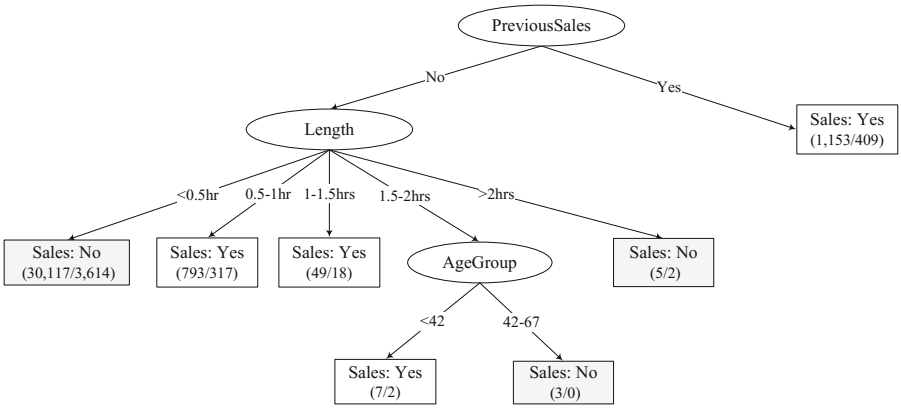

Fig. 5. Decision tree example for banking/marketing analysis.

Cluster centroids:

	Cluster#			
Attribute	Full Data	0	1	2
	(32127)	(4872)	(9962)	(17293)
prev.sales	0.0701	0.2943	0.0304	0.0298
length	516.6318	932.5649	447.0528	439.5324
age	40.0164	40.6232	50.9184	33.5651
y	no	yes	no	no

Fig. 6. Clustering example for banking/marketing analysis.

Cluster 0 – these are the 15% of customers who purchased the products. It shows the average previous sales for this group was the highest at $\sim$0.3, compared with the other two clusters which were $\sim$0.03. In terms of the length of phone call, this cluster is also the highest, $\sim$15 min – twice as long as the other two groups. Finally, the average age within this group is about 40 years.

Cluster 1 – the number of customers in this group is more than twice that of Cluster 0, although they didn't accept the sales offer – average age is around 50 years.

Cluster 2 – the final cluster is the largest (54%) – this youngest group with average age about 33 years did not purchase any products from the bank.

5 Conclusion

Businesses have generated and developed a vast amount of complex data over the years. This data is a precious resource and could play a vital role enabling support for decision making through insight generation and knowledge extraction. The growing amount of data exceeds the ability of traditional methods for data analysis and this has led to the increasing use of emerging technologies. A data-driven framework for business analytics

has been proposed in this paper which integrates the sources of data, stages of processing, methods and technologies within the context of the 5 Vs of big data.

Big data is overwhelming not only because of its volume, diversity of data types and the speed at which it must be managed, but also the trustworthiness of the data – most important is how to create *value* from all this data. Using big data technology has the potential to lead to more efficient and flexible business applications. However, there are several issues that need to be addressed to maximise the benefits of big data analytics across commerce and industry. Wyatt (2016) cites the five big challenges for healthcare as: data quality; developing reliable inference methods to inform decisions; implementation of trusted research platforms; data analyst capacity; and clear communication of analytical results [20].

Following the proposed framework, analytical methods and tools, this paper has illustrated two real-world case studies – in particular through the application of visual analytics, data mining techniques and machine learning algorithms. Khan *et al.* (2016) suggested that, while many organisations still build market value and advantage over their rivals through traditional means, algorithms have emerged as a better way to change the business and gain a more competitive edge [9]. If algorithms are applied correctly they can provide insights that make a business process more profitable and highlight new ways of doing business as well as new opportunities for growth.

Data analytics algorithms vary significantly in capability and scope. Consequently there are algorithms which aim to find the '*known* knowns' – e.g. OLAP analysis; then there are algorithms which are able to discover 'known *unknowns*' – e.g. through data mining and machine learning; and recently there are algorithms which are even able to extract the '*unknown* unknowns' from datasets – e.g. deep-learning algorithms. By applying such digital innovation, the ultimate goal is improving decision making in the business environment harnessing the full potential of big data.

References

1. Alteryx: The business grammar report: a study of european decision-makers' attitudes to data and analytics in modern business (2016). https://www.alteryx.com/resources/the-business-grammar-report-a-study-of-european-decision-makers-attitudes-to-data
2. Computing Research: Big Data & IoT Review 2017 (2017). https://www.computing.co.uk/ctg/news/3010002/computing-big-data-iot-review-2017
3. Gartner IT Glossary (2001). https://www.gartner.com/it-glossary/big-data
4. GitHub (2017). https://github.com/QUT-BDA-MOOC/FLbigdataStats
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explor. **11**(1), 10–18 (2009)
6. Hand, D.J., Smyth, P., Mannila, H.: Principles of Data Mining. MIT Press Cambridge, USA (2001)
7. IBM: IBM SPSS Statistics for Windows, Version 22.0. IBM Corporation, Armonk, NY (2013)
8. IBM developerWorks: Hive as a tool for ETL or ELT (2015). http://www.ibm.com/developerworks/library/bd-hivetool

9. Khan, I., Gadalla, C., Mitchell-Keller, L., Goldberg, M.S.: Algorithms: The new means of production. Digitalist Magazine (2016). www.digitalistmag.com/executive-research/algorithms-the-new-means-of-production
10. Kimball, R., Ross, M.: The Data Warehouse Toolkit – The Definitive Guide to Dimensional Modeling. Wiley, New York (2013)
11. Lans, R.: Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses, Morgan Kaufmann Publishers Inc. (2012)
12. Lu, J., et al.: Data mining techniques in health informatics: a case study from breast cancer research. In: Renda, M.E., Bursa, M., Holzinger, A., Khuri, S. (eds.) ITBAM 2015. LNCS, vol. 9267, pp. 56–70. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22741-2_6
13. Lu, J., Hales, A., Rew, D.: Modelling of cancer patient records: a structured approach to data mining and visual analytics. In: Bursa, M., Holzinger, A., Renda, M.E., Khuri, S. (eds.) ITBAM 2017. LNCS, vol. 10443, pp. 30–51. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64265-9_4
14. Marr, B.: Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. Wiley, Chichester (2015)
15. Marr, B.: Big Data In Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley, Oxford (2016)
16. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)
17. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehouse **5** (4), 13–22 (2000)
18. Shmueli, G.: Practical Time Series Forecasting with R: A Hands-on Guide. Axelrod Schnall (2016)
19. Wiese, L.: Advanced Data Management: For SQL, NoSQL, Cloud and Distributed Databases. De Gruyter Textbook (2015)
20. Wyatt, J.: Plenary Talk: Five big challenges for big health data. In: 8th IMA Conference on Quantitative Modelling in the Management of Health and Social Care, London (2016)