Emerald International Journal of Operations and Production Management

Capturing Value from Big Data – A Taxonomy of Data-Driven Business Models Used by Start-Up Firms

Journal:	*International Journal of Operations and Production Management*
Manuscript ID	IJOPM-02-2014-0098.R2
Manuscript Type:	Research Paper
Keywords:	start-up business model, Big data, Data driven business model, DDBM, Business model

Capturing Value from Big Data – A Taxonomy of Data-Driven Business Models Used by Start-Up Firms

Purpose – This paper aims to derive a taxonomy of business models used by start-up firms that rely on data as a key resource for business, namely data-driven business models (DDBMs). By providing a framework to systematically analyse DDBMs, the study provides an introduction to DDBM as a field of study.

Design/methodology/approach – To develop the taxonomy of DDBMs, business model descriptions of 100 randomly chosen start-up firms were coded using a DDBM framework derived from literature, comprising six dimensions with thirty-five features. Subsequent application of clustering algorithms produced six different types of DDBM, validated by case studies from the study's sample.

Findings – The taxonomy derived from our research consists of six different types of DDBM among start-ups. These types are characterised by a subset of six of nine clustering variables from the DDBM framework.

Practical implications – A major contribution of the paper is the designed framework, which stimulates thinking about the nature and future of DDBMs. The proposed taxonomy will help organisations to position their activities in the current DDBM landscape. Moreover, framework and taxonomy may lead to a DDBM design toolbox.

Originality/value – This paper develops a basis for understanding how start-ups build business models to capture value from data as a key resource, adding a business perspective to the discussion of big data. By offering the scientific community a specific framework of business model features and a subsequent taxonomy, the paper provides reference points and serves as a foundation for future studies of DDBMs.

Keywords: Data-driven business model, Big data, Business model, Start-up business model

Paper type: Research paper

1. Introduction

The exponential growth of data compounded by the Internet, social media, cloud computing and mobile devices – or big data – has an embedded value potential that must be commercialised. The widespread quote *'Data is the new oil'* (WEF, 2011; Rotella, 2012) establishes the analogy to natural resources needing to be exploited and refined to guarantee growth and profit.

Some studies estimate an increase in annually created, replicated and consumed data from around 1,200 exabytes in 2010 to 40,000 in 2020 (Gantz and Reinsel, 2012). In some industries big data has led to the creation of entirely new business models. In the retail sector, big data expedites the analysis of in-store purchasing behaviour in near real-time to adjust merchandise, stock levels and prices (Hagen et al., 2013).

A study by Kart et al. (2013) involving 720 IT and business leaders ranks the issue of monetising data over questions of technical feasibility. Hence, building on Manyika et al. (2011) and Chesbrough and Rosenbloom (2002), business models supporting data-related ventures to capture value, subsequently called data-driven business models (DDBMs), are needed. Notably, scholars have published surprisingly little on this topic. Hence, understanding the nature of business models that rely on data remains a research question.

Although companies relying on data – such as insurance companies – is not a new concept, it was only recently that companies began to make use of other data sources such as social media, smartphones or sensors, and new technologies designed to exploit this data. Hence, companies leveraging these novel forms of data and analytical methods are subject to investigation here. New technologies and innovations are often commercialised through start-up companies (Criscuolo et al., 2012). Digital start-up firms are not bound by the legacy systems of established firms built over a period of time. It might therefore be easier for newer firms to get the right infrastructure in place to exploit their data. Therefore, leveraging the advantage of starting from a blank page instead of being constrained by the existing business, these young companies create a rich variety of, presumably, purer business models. Hence, their first-mover character, unspoiled business models and population size make start-up firms a promising basis for investigating DDBMs.

Unfortunately, start-ups with DDBMs in place are still relatively young. Their business models may change rapidly in the future and, according to recent figures from venture capitalists (Rao, 2013), over 90 per cent of start-ups may fail. Distinguishing successful from unsuccessful companies is therefore currently not possible.

This paper took a snapshot of DDBMs in start-ups to obtain an understanding of current DDBMs, to build clusters of similar DDBMs and to create a reference point for future studies regarding their success and the evolution of different types of business model.

As a starting-point for the study, a framework of dimensions and features was needed to analyse systematically and describe DDBMs. Therefore, the study builds on the extant body of knowledge from business model research (Chesbrough and Rosenbloom, 2002; Osterwalder, 2004; Johnson et al., 2008; Bouwman et al., 2008; Baden-Fuller and Haefliger, 2013), specifying the more general business model frameworks to render them insightful in terms of assessing DDBMs. The resulting framework was then used to code publicly available documents describing the business models of 100 start-ups. Established clustering algorithms applied to the results of the coding process produced six DDBM types. In a series of interviews with start-up representatives from the sampled firms the patterns were confirmed by comparing the algorithmically identified clusters with the competitive landscape sketched by the interviewees.

2. Literature and research questions

2.1 Existing business model literature

The existing literature around business models has evolved significantly in recent years and the concept is now used in the context of exploring the theoretical foundations of value creation in e-business, strategy and innovation management (Zott et al., 2011). However, academic consensus on the definition, and the question of how to represent a business model, is still missing in innovation management, entrepreneurship or strategic management theory literature (Weill et al., 2011; Zott et al., 2011; Burkhart et al., 2011). The business model frameworks presented in the extant literature can be divided into static and dynamic approaches (Burkhart et al., 2011). While the static view describes the current state of a company, the dynamic view further examines the evolution of a business model (De Reuver et al., 2013; Bouwman and MacInnes, 2006; El Sawy and Pereira, 2012). While the latter will become important in future studies on the modifications that start-ups apply to their DDBMs, this paper focuses on static business model frameworks, its intention being to provide a snapshot of the DDBMs that currently create and capture value from big data in start-ups.

One of the first business model frameworks, by Chesbrough and Rosenbloom (2002), describes the concept of a business model from a functional perspective. According to them a business model articulates the value proposition, identifies a market segment and defines a company's value chain (Chesbrough and Rosenbloom, 2002). Frequently, a component-based perspective is used to describe business models (Burkhart et al., 2011). Hedman and Kalling (2003), drawing on strategy theory and business model research, propose a seven-component framework, consisting of customers, competitors, offering, activities and organisation, resources and supply of factor and production inputs, as well as a longitudinal process component to 'cover the

dynamics of the business model over time and the cognitive and cultural constraints that managers have to cope with'. Johnson et al.'s framework (2008) consists of four interlocking components: the customer value proposition, the profit formula, key resources and key processes to deliver the value proposition profitably. Bouwman et al. (2008), originating from the mobile services domain, propose the STOF-business model framework comprising four domains: the **S**ervice domain conceptualised by delivered and perceived customer value; the **T**echnology domain describing the necessary architecture to deliver the service; the **O**rganisation domain, including resources and capabilities, and company strategy; and the **F**inance domain, describing costs and revenues. Heikkilä et al. (2008) add the customer relationship perspective, placing it at the centre of their framework. Baden-Fuller and Haefliger (2013) introduce a topology with four dimensions: customer identification, customer engagement, value delivery and monetisation. Several authors propose a unified business model framework, synthesizing the existing literature (Morris et al. (2005) and Al-Debei and Avison (2010))

In the start-up and corporate world, the practitioner-oriented business model canvas of Osterwalder et al. (2010) – based on his business model ontology (Osterwalder, 2004) – is widely applied (Stuckenberg et al., 2011). The canvas consists of nine building-blocks: value proposition, key processes, key resources, key partners, customer relationships, channels, customer segment, revenue streams and cost structure.

Table 1 Review of different business model frameworks

2.2 Big data and value creation

The term 'big data' has become popular in recent years; however, its poor definition can result in ambiguous meaning. Often big data is defined in terms of data volume (Manyika et al., 2011) but there is growing awareness that this view is limited (Schroeck et al., 2012).

One of the most commonly cited definitions is proposed by Gartner (2012): 'Big data is high-**v**olume, high-**v**elocity and high-**v**ariety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.' Often a fourth dimension is added to address the uncertainty of the data, namely veracity, referring to the reliability of a certain data type (Schroeck et al., 2012). Although a useful characterisation of 'big data', this definition lacks clarity, particularly regarding explicit distinctions between 'high' and 'low' values for the respective dimensions.

Collecting, storing and analysing (big) data is not an end in itself for companies – they are interested in creating actual business value. Relatedly, Davenport (2006) lists several, mostly anecdotal, examples of companies drawing competitive advantage from the use of data and analytics. Moreover, an empirical study by McAfee and Brynjolfsson (2012) suggests that companies relying more on data-driven decision-making perform better in terms of productivity and profitability. Several practitioner papers and White Papers addressing the questions of 'where' and 'how' big data creates value can be obtained from consulting companies (Hagen et al., 2013; Manyika et al., 2011; Schroeck et al., 2012) and vendors (Petter and Peppard, 2012; CEBR, 2012). Two main forms, how big data creates value for companies, can be synthesised: first, (big) data is used for the incremental improvement and optimisation of current business practices, processes and services. Second, new products and business models can be innovated based on data use.

So far, few academic papers describe or analyse business models relying on data. Otto and Aier (2013) describe different business models in the business partner data domain using a case-study approach. They use the business model framework by Hedman and Kalling (2003) to analyse their cases systematically. Another stream of literature proposes *data-as-a-service* and *analytics-as-a-service* as new service types. However, most of these papers focus on technical or organisational aspects (Delen and Demirkan, 2013; Stipic and Bronzin, 2012). An exception is provided by Chen et al. (2011), who focus on the analytics ecosystem and define the two new types of business model as relying on data from a structural perspective.

Hence, most of the literature on this topic is written or commissioned by consultancies and IT vendors, who have an interest in showcasing the value-creation potential of data use. Correspondingly, the aforementioned review highlights a general gap in the business model, innovation management, entrepreneurship and strategy literature concerning 'if' and 'how' big data creates value for companies. We aim to reduce this gap by addressing the overarching research question:

What types of business model are present among companies relying on data as a resource of major importance for their business (key resource)?

Specifically:

- What does a framework look like that allows systematic analysis and comparison of DDBMs?
- What clusters of companies with similar business models exist in the identified sample?

Although the term DDBM has not yet been defined in the scholarly literature, it is commonly used by practitioners (in several blog entries, cf. Svrluga, 2012, or Diebold, 2012) and in the research community (in the British Research Council's 'New Economic Models in the Digital Economy' (NEMODE) initiative), and can therefore be considered on the edge of establishment. This paper provides a definition of a DDBM as *a business model relying on data as a key resource*.

This definition has three implications. First, a DDBM is not limited to companies conducting analytics, but includes companies that are 'merely' aggregating or collecting data. Second, a company may sell not just data or information but also any other product or service that relies on data as a key resource. Third, it is obvious that any company uses data in some way to conduct business – even a small restaurant relies on its suppliers' contact details. However, the focus is on companies using data as a key resource for their business model. Specifically, the paper uses DDBMs for business models within a big data context.

3. Research design

3.1 Overview of methodology

The objective of this paper is to build a taxonomy of business models relying on data as a key resource in the start-up world. A taxonomy is an empirically derived classification scheme used in various scientific disciplines (Hambrick, 1984). This research used cluster analysis as a numerical method for deriving taxonomies (Everitt et al., 2011). In order to systematically compare business models and enable taxonomy development, this paper proposes a DDBM framework providing a set of DDBM-specific attributes for every business model dimension. The framework was developed in two steps. First, based on a systematic literature review of existing business model frameworks, the relevant dimensions of a business model were identified. Second, for each dimension, a comprehensive set of features was identified using literature from related disciplines, including data warehousing, business intelligence and cloud-based business models. The proposed framework resembles a morphological box, where each business model can be described using the features in the different dimensions. However, a business model can have more than one feature in each dimension.

Publicly available qualitative data on the business models of 100 randomly chosen start-up companies was collected and analysed using the developed framework. To build taxonomy from this data, a *k*-medoids clustering algorithm was used.

3.2 The DDBM framework

To identify the relevant dimensions for the DDBM framework, existing static business model frameworks were systematically reviewed. Although there is no general agreement on the number and types of business model dimensions, the following six key dimensions are commonly found among various authors (cf. Table 1): key resources, key activities, value proposition, customer segment, revenue model and cost structure.

1. Key resources: Companies need resources to create value (Wernerfelt, 1984). Firm resources include 'all assets, capabilities, organizational processes, firm attributes, information, knowledge controlled by a firm' (Barney 1991). By definition a DDBM has data as a key resource, although data may not be the only key resource of the respective business model. To create the DDBM framework the types of data source used by companies need to be understood.

Several papers from the Information System field, such as data mining and data warehousing, list – mostly non-comprehensively and non-structured – potential data sources or types (Singh and Singh, 2010; Han et al., 2011; Schroeck et al., 2012; Kosala and Blockeel, 2000). More systematically, Negash (2004) proposes differentiating between structured (e.g. ERP systems) and semi-structured data (e.g. spreadsheets or videos) and distinguishing between internal and external data sources. Furthermore, Gartner (Buytendijk et al., 2013) identifies five different types of data source: 1) operational data from transaction systems; 2) dark data that you own but is currently unused (e.g. emails); 3) commercial data, including structured or unstructured data acquired from third parties (e.g. stock market data); 4) social data (e.g. Facebook); and 5) public data (e.g. socio-demographic).

Consolidating the different views, eight data sources emerge, which can be divided into internal and external data sources. Internal sources include existing data that can be drawn from IT systems but is currently unused (e.g. ERP data); self-generated data for a specific purpose, through tracking (e.g. Web-navigation or sensor data); or crowdsourced data, created by a broad set of contributors over the Web or social collaboration techniques (Gartner, 2013). External data comprises acquired data, which can be purchased from data providers; data provided by customers or business partners that is not generally available; and freely available data, which is publicly available at no direct cost. Freely available data can be further split into open data, which is downloadable, machine-readable and structured without prior processing (Lakomaa and Kallberg, 2013); social media data from websites such as Facebook; and Web-crawled data, which is publicly available but needs to be gathered electronically (e.g. blog entries).

2. Key activities: Each company performs different activities to produce and deliver its offering. For DDBMs these activities must be related to the key resource data. There are several different process models in the domain of data mining, describing the activities in the knowledge-discovery process (Cios, 2007; Han et al.,

2011). Fayyad et al. (1996) present an early, frequently cited model outlining five data-related key activities: selection of a data set for subsequent analysis; pre-processing and cleaning the data; data reduction or transformation to reduce the number of variables; data mining, namely the identification of data patterns; and the interpretation and visualisation of the mined patterns. Otto and Aier (2013) identify key activities, including retrieving data, data mining and distribution thereof based on a multi-case study. Others suggest subdividing analytics activities into descriptive, predictive and prescriptive analytics (Delen and Demirkan, 2013).

The key activities dimension of the DDBM framework results from integrating the different perspectives from the literature along the steps of the 'virtual value chain' (Rayport and Sviokla, 1995): gathering, organising, selecting, synthesising and distributing. Corresponding with the data sources dimension, data can be generated internally or acquired from external sources. The generation can be done by crawling internal sources, tracking sensors or using crowdsourcing. For further activities the data may be processed (transformed or cleaned), or aggregated (organising and selecting) from different sources. Insight is generated through analytics (synthesising), which can be subdivided into: descriptive analytics – explaining the past; predictive analytics – forecasting future outcomes; and prescriptive analytics – predicting future outcomes and suggesting decisions. Finally, the data or insight might be visualised and distributed to customers.

3. Offering/value proposition: The customer value of a product or service is the 'starting point for any business model' (Bouwman et al., 2008, p. 36). Therefore, the offering, often called the value proposition, is the central dimension of all business model frameworks (Chesbrough and Rosenbloom, 2002; Osterwalder, 2004; Johnson et al., 2008). Barnes et al. (2009) define a value proposition as the 'expression of the experience that a customer will receive from a supplier'. Thus, the value proposition is the value created for customers through the offering.

According to Fayyad et al. (1996), a company's offering can be divided into two categories: data and information/knowledge, with (raw) data being primarily 'a set of facts' without an attached meaning. When data has been interpreted it becomes information or knowledge. The output of any analytics activity considers knowledge as it attaches meaning to data. Finally, a third offering, non-data products or services, is added that accounts for companies providing a non-virtual offering.

4. Customer segment: Each company's offering targets certain customers. The most generic classification is used, differentiating businesses (B2B) from individual consumers (B2C) (Morris et al., 2005; Osterwalder, 2004). In many cases, companies could target both businesses and individual consumers.

5. *Revenue model:* In the long term, having at least one revenue stream is vital for every company. The extant literature suggests seven different revenue streams (Bouwman et al., 2008; Osterwalder, 2004; Osterwalder et al., 2010): asset sale – exchanging the ownership rights of goods or services for money; lending/renting/leasing – temporarily granting the exclusive usage right of an asset; licensing – granting permission to use protected intellectual property such as a patent in exchange for a fee; usage fee – charged per use of a particular service; subscription fee – charged for the use of the service; brokerage fee – charged for an intermediate service; or advertising.

6. *Cost structure:* To create and deliver value to customers, a firm incurs costs for labour, purchased products, and so on. In terms of a DDBM, a company's cost structure is less interesting than a specific cost advantage regarding data use. Typically, such a cost advantage would occur if the data used in its product or service were created independently of the specific offering. For instance, Twitter can use its own data without additional costs to provide an analytical service, while companies such as Gnip, a social media analytics start-up, has to buy the respective data from Twitter.

Compiling the aforementioned six dimensions and the respective features leads to the DDBM framework shown in Figure 1, which can be used to describe the business model of the sample companies comprehensively.

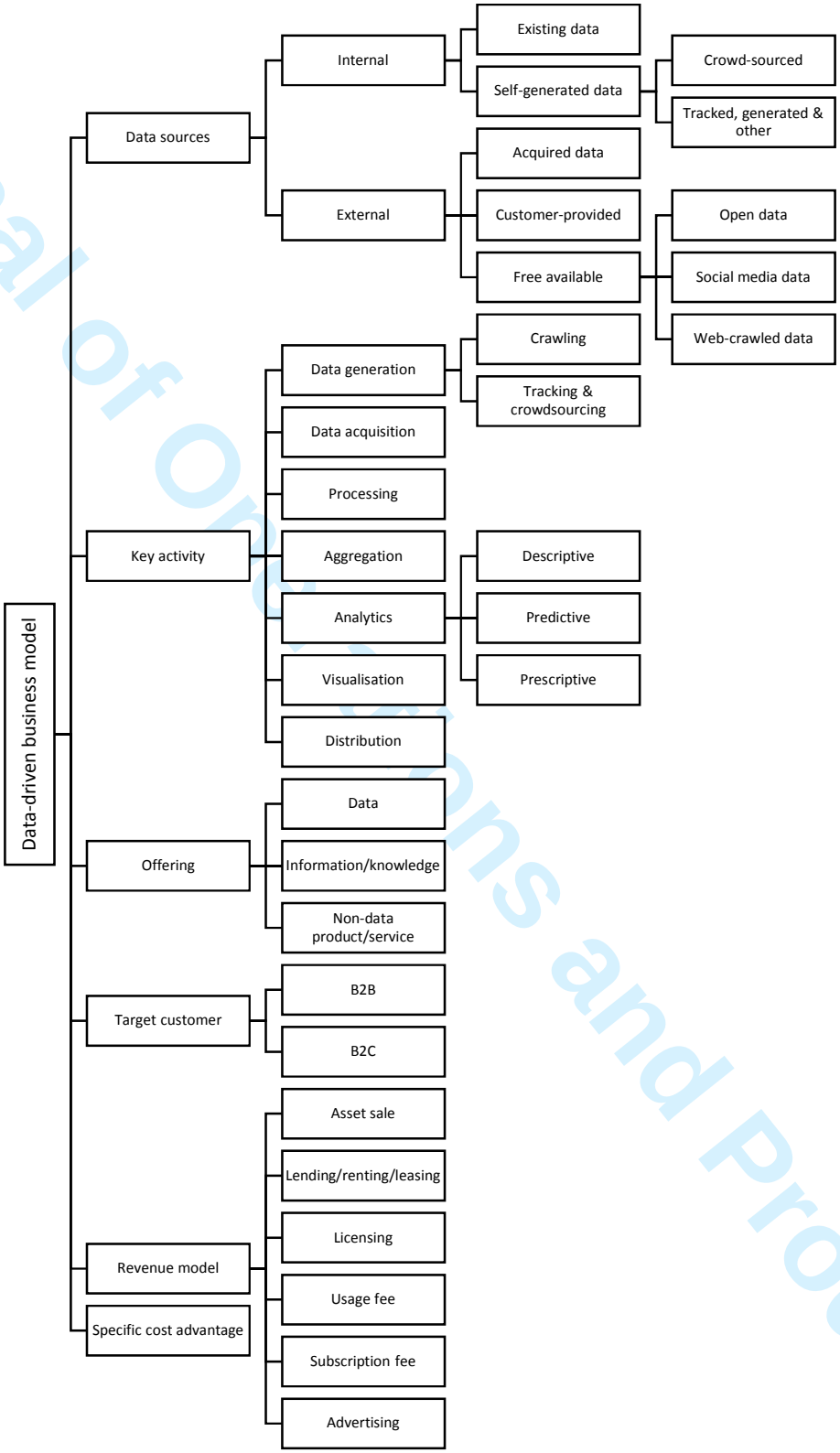

Figure 1 The DDBM framework

3.3 Sample and data collection

The sample was drawn from the website AngelList (www.angellist.com), which offers companies the possibility to create a profile to connect with investors, potential employees and interested persons (Wollan, 2011). This study focuses on companies from the categories 'big data' or 'big data analytics'. From 20 July 2013, 1,329

companies on AngelList were tagged accordingly.[1] From this list a sample of 100 start-up companies was randomly selected to prevent researcher bias through the selective choice of samples (Johnson 1997) and to achieve a representative sample allowing for generalisation (Flyvbjerg, 2006).

Subsequently, data on the business models of the selected firms was collected from reliable, publicly available sources. This type of secondary data is generally sufficient to describe business models as 'gross elements of business models are often quite transparent' (Teece, 2010), and helps to ensure descriptive validity (Tashakkori and Teddlie, 2002). Data sources comprise company websites, which provide a broad set of information about the companies, news sources, including start-up-focused online journals such as *TechChrunch* or *VentureBeat,* and traditional newspapers and magazines such as *The New York Times* or *The Wall Street Journal.* In total 303 different documents were collected to analyse the 100 different start-up business models, using on average more than three different sources for the coding of one business model.

3.4 Coding process

The data was manually analysed and coded by two independent coders, A and B, using the developed DDBM framework. While coder A was one of the authors of this paper, coder B was skilled in the relevant domain but not involved in the research. To ensure reliability of coding, the different features of the framework were clearly defined prior to coding. However, some of the more ambiguous business model dimensions required judgement and were prone to coding errors (Cooper, 1988). For instance, regarding the offering dimension, the line between data and information can be thin. Nevertheless, the case studies confirmed the reliability of the coding process. After the first coding cycle, each company was checked for sufficient coverage of all dimensions and additional data was collected, where necessary. If no information for a particular dimension was found it was coded accordingly. Concluding the coding process, both coders met with a judge (another paper author) to resolve any disagreements (Fastoso and Whitelock, 2010). In total there were 1,341 coded terms. Disagreement between the two coders focused on 35 items, resolved in all cases by investigating the context of a particular statement. The output of this process was binary feature vectors.

3.5 DDBM cluster analysis

To build taxonomy for the DDBMs of start-ups, the researchers conducted cluster analysis, following a four-step process (Ketchen and Shook, 1996; Mooi and Sarstedt, 2011): selection of clustering variables, choice of clustering algorithm and similarity measure, choice of cluster numbers, and validation and interpretation of the clustering result.

In the first step, variables determining affiliation to a group were chosen. These variables had to be relevant to the clustering process, as adding irrelevant variables can 'dramatically interfere with cluster recovery' (Miligan, 1996). Moreover, sample size constrains the number of variables since every added variable over-proportionally increases the number of required items to ensure statistical validity (Mooi and Sarstedt, 2011). It is advisable to have a sample size of at least 2^m, where m equals the number of clustering variables (Mooi and Sarstedt, 2011). Hence, for the present study (n = 100) a good number of variables was six ($2^6 = 64$) or seven ($2^7 = 128$).

The clustering variables for the paper were selected using the DDBM framework. Based on frequency analysis, the clustering variables were reduced to 'data source' and 'key activity'. For 36 per cent of the selected companies no information on revenue model was available and, for those where data was available, 83 per cent used a subscription or usage-fee-based revenue model. As this dimension lacked discriminatory power, it was not further regarded for clustering. Likewise for 'offering', 94 per cent of the companies were classified as offering information or knowledge. As no specific cost advantage could be identified, 'cost structure' was also further disregarded. Finally, the features were limited to the second level in the framework (no differentiation between the type of analytics) to reduce the number of variables. Based on this pre-selection, nine variables remained for clustering: data source – acquired data; customer provided data; free available data; tracked and generated data; crowdsourced data; key activity – aggregation; analytics; data acquisition; and data generation.

Corresponding with the intended outcome of the cluster analysis – mutually exclusive sets of similar business models to identify distinctive types – a partitioning method, the *k*-medoids clustering algorithm, was selected. The *k*-medoids algorithm groups n objects into k clusters by minimising the sum of dissimilarity between each object, p, and its corresponding representative object, o_i (medoid), for all objects in cluster C_i (Han et al., 2011):

$$\min E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i).$$

[1] The number of unique companies tagged with any of these tags. As companies can have multiple tags, the sum of the number of companies in each category is higher but includes duplicates.

The *k*-medoids algorithm was selected over the more common *k*-means algorithm, as the cluster representatives (medoids) were observed business models from the sample, making the results more meaningful. Furthermore, the *k*-medoids algorithm is less sensitive to outliers (Han et al., 2011).

One of the decisive questions when selecting a binary similarity measure is whether neither business model having a particular feature is relevant to determining their similarity, namely, if negative matches should be regarded (Everitt et al., 2011). For the present research it was assumed that the co-absence of features was relevant for the similarity of two business models. The Euclidean distance measure was used, which implicitly includes positive and negative matches by determining the distance only based on mismatches ***b*** and ***c*** (Choi et al., 2010):

$$dist = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \sqrt{b + c}$$

One fundamental question when using a partitioning clustering method is determining the number of clusters – the '*k*' in *k*-medoids. Selecting the number of clusters is a trade-off between having a reasonably large number of clusters to reflect the specific differences in the data set, and having significantly fewer clusters than data points, as this is the motivation for cluster analysis (Han et al., 2011).

Several different approaches exist to determine the number of clusters, and it is advisable to compare the results of different methods (Pham et al., 2005; Mooi and Sarstedt, 2011; Han et al., 2011).

Han et al. (2011) provide a rule of thumb to set the number of clusters to $\sqrt{\frac{n}{2}}$, with *n* being the sample size. Thus, seven is an appropriate number of clusters. Another option to determine the number of clusters is the so-called 'elbow method'[2] (Mooi and Sarstedt, 2011). Hierarchical clustering is conducted and then the number of clusters is plotted against the agglomeration coefficient – the distance at which two cases or clusters are merged to form a new cluster. Searching this plot for a distinctive break ('elbow') shows that there is no clear elbow signifying an appropriate number of clusters. However, the plot reveals there should be no more than 48 clusters in total, as clusters with zero distance are then split up. Moreover, small 'elbows' suggest 7, 16 and 23 as favourable cluster numbers. Furthermore, several statistical tests exist to determine the number of clusters (Pham et al., 2005). In this study, the silhouette coefficient was used (Rousseeuw, 1987), resulting in a value of 0.335, indicating that there should be at least six different clusters. According to Kaufman and Rousseeuw (1990), this silhouette coefficient implies weak clustering. However, this is typical for social science data and does not invalidate the clustering findings (Hambrick, 1984).

Taking together the results from the different methods, seven was identified as an appropriate number of clusters. Finally, the clustering solution must be validated to ensure the clustering result is meaningful and useful (Ketchen and Shook, 1996). First, repeating the clustering with different clustering algorithms, a hierarchical clustering method demonstrated the reliability or stability of the clustering solution, as objects were assigned to a similar cluster. Second, the clustering solution was validated in two ways: the internal cluster quality was determined using the silhouette coefficient (Han et al., 2011) and the significance of the clusters was reviewed through case studies. Finally, interpretation of the clustering solution was done through a comparison of the different clusters, as well as detailed analysis of the companies in the clusters. Consequently, six of the seven clusters proved meaningful.

4. Results and discussion

4.1 Data sample analysis

Based on the coding of the 100 companies, some general characteristics can be observed. First, in terms of data source most of the companies use external data sources (73%), 16 per cent use internal and external data sources, while 11 per cent use only internal data sources – data they create themselves. Furthermore, most of the companies in the sample conduct analytics as a key activity (76%); however, only a small number perform 'advanced' analytics, either predictive (22%) or prescriptive (6%).

The majority of all examined companies rely on a subscription (62% of all companies with information on the revenue model) or usage-fee-based revenue model (20%). Overall, a noteworthy predominance of B2B business models within the examined companies can be observed. Over 80 per cent of the companies target business customers with their offerings (70% only B2B, and 13% both B2B and B2C). The vast majority of companies in the sample offer information or knowledge, which relates to the selected sample: Web-based business models are predominant with start-ups on AngelList and therefore most of the offerings are also Web-based.

4.2 DDBM cluster results

[2] Different descriptions of this method exist (cf. Han et al., 2011; Ketchen and Shook, 1996). However, the basic idea remains the same.

As a result of clustering, seven different clusters were identified, as shown in Table 2, characterised by their respective cluster medoids. However, after analysing the seven different clusters by comparing companies in the respective cluster, only six (Types A–F) were further considered. It is important to note that, as a result of the relatively small sample size, all quantitative data on the percentage distribution of the different business model features was indicative but not statistically significant. The five companies in Cluster 3 did not show sufficient similarity and were therefore disregarded.

Table 2 DDBM cluster results and respective medoids

Comparing the six different business model types by the data source on which they rely, four distinctive patterns were identifiable. Types B and E rely on data provided by customers and/or partners. Business models under Types A and D rely on free available data. Type C relies on data generated through crowdsourcing or tracking, and Type F combines customer-provided and free available data sources. Drilling down to the specific free available data sources used by the companies shows that mostly social media data is used. Moreover, more than half of the companies of Type F obtain their data through Web crawling.

Looking at the key activities performed by the different types of company also reveals distinctive patterns. The business model types are mainly characterised by the three key activities of 'aggregation', 'analytics' and 'data generation'. Three distinctive patterns are identifiable for 'analytics' and 'aggregation'. Types A and E rely on aggregating data from different sources. Types B, C and D only conduct analytics without aggregation, and Type F conducts both aggregation and analytics. For 'data generation' the picture is less unambiguous: while all companies of Type F generate data, and this type is subsequently characterised by this activity, a fraction of companies of other types perform this activity (Type F: 57%; Type D: 50%; and Type A: 35%).

As aforementioned, most of the companies rely on a subscription or usage-fee-based revenue model. A small deviation is determinable for companies in Type C, which generate revenue from asset sales. This follows mainly from the sale of devices to generate the necessary data. Furthermore, companies in Types B and E rely entirely on a subscription or usage-fee revenue model, which emphasises the 'as-a-service' characteristic of these business models.

Slicing the data in a different way reveals that the selected revenue model depends more on the targeted customers than on business model type. While 98 per cent of B2B business models use either a subscription or usage-fee-based revenue model, those companies targeting consumers use diverse ways to generate revenue: advertising (27%), asset sales (27%) and brokerage fees (18%). While over 80 per cent of the companies target business customers with their offerings (see above), the split varies between clusters. The vast majority of companies of Types B, D and F address business customers. However, half of all companies of Type C target consumers.

In terms of the dimension offering, most companies offer information or knowledge. Some exceptions can be found in Types A and F, where companies provide raw data; and in Types C and D, where companies sell independent services related to data-generation or analytics. However, based on the difficulties of differentiating the three categories of offering, this dimension was not considered for clustering or the further analysis of clusters.

Table 3 DDBM clusters: general statistics

The six clusters consist of largely homogenous sets of companies that can be summarised through sobriquets based on their respective characteristics. The representative objects (medoids) of these six different business model types are characterised by six of the nine clustering variables, namely 'aggregation', 'analytics', 'data generation', 'free available data', 'customer-provided data' and 'tracked and generated data' in the dimensions' data source. The business model types can therefore be presented in a 3x3 matrix, as demonstrated in Figure 2.

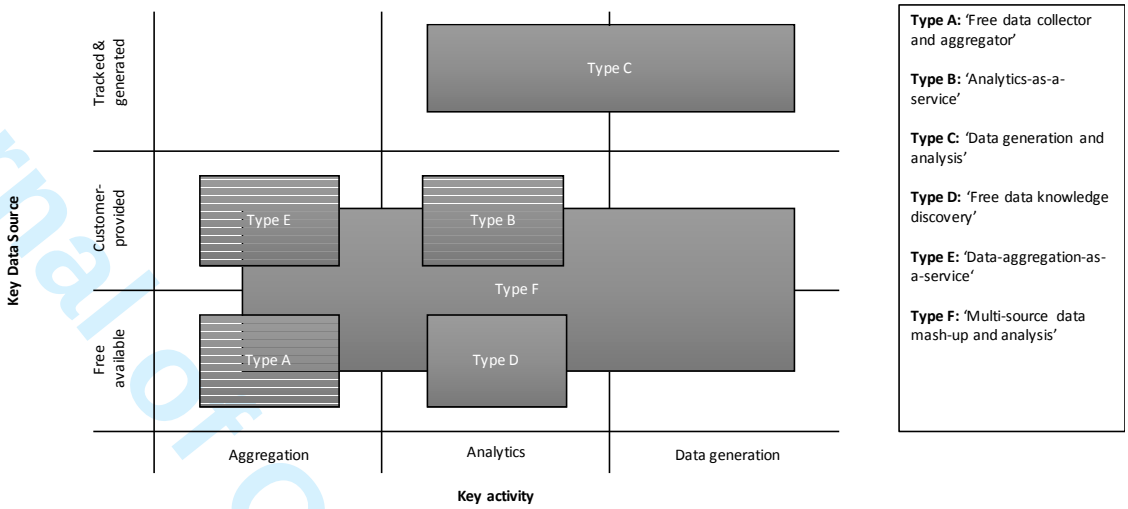

Figure 2 DDBM matrix of centroids

Type A: 'Free data collector and aggregator'

Type A companies create value by collecting and aggregating data from a vast number of different, mostly free, data sources, and then distributing it, for example, through an API. Other key activities performed by such companies are data crawling (35%) and visualisation (24%). While companies of this type are characterised by the use of free available data (100%) – mostly social media data (65%) – other data sources such as proprietary acquired data (12%) or crowdsourced data (12%) are also aggregated by some of the companies.

A relatively high share of companies targeting consumers with this kind of offering is observable. For example, companies like AVUXI and CO Everywhere aggregate data from numerous sources on local businesses, such as restaurants and bars, providing it to consumers. Accordingly, revenue models like advertising (12%) or brokerage fees (6%) – besides subscription (47%) or usage fees (12%) – are applied by these companies.

An example of such a B2B company is Gnip, which aggregates data from a wide range of different social media sources, normalises the formats, offers possibilities to filter the data and provides access to the raw data via an API. Besides providing free available social media sources, Gnip is also a premium reseller of Twitter data. Gnip's key value proposition is easy, reliable access to a large number of different data sources through a single API. Revenue is generated from a mixture of subscription and usage fees.

Type B: 'Analytics-as-a-service'

Type B companies conduct analytics (100%) on data provided by their customers (100%). Further noteworthy activities include data distribution (36%), mainly through providing access to the analytics results via an API, and visualisation of the analytics results (36%). In addition to the customer-provided data, some companies also include other data sources, mainly to improve the analytics. Sendify, a company providing real-time inbound caller-scoring, joins external demographic data with inbound call data to improve the analysis.

The scope of the different analytic services varies from fraud detection (Sift Science), improving marketing activities (7signal), improving customer service and relationships (Sendify) and increasing sales (Granify), to generic data analysis (Augify). Companies of this type primarily target business customers with their solution. Hence, the revenue model is predominantly based on subscription or usage fees.

Type C: 'Data generation and analysis'

Type C companies generate data rather than relying on existing data. Moreover, many also perform analytics on this data. Firms can be roughly subdivided into three groups: companies that generate data through crowdsourcing; Web analytics companies; and companies that generate data through smartphones or other physical sensors.

For example, Swarmly provides a smartphone application whereby users can share their current location and provide details of their sentiments about the venue. Swarmly aggregates this data to provide a real-time map of popular venues such as bars, restaurants or clubs.

The second group comprises companies such as GoSquared, Mixpanel or Spinnakr, which provide Web analytics services. They collect data through a tracking code embedded in their customers' websites and analyse it. Reports or raw data are provided through a Web-based dashboard or other interfaces.

The third group are companies that collect data through any physical device, including smartphone sensors. For example, Automatic sells a device that can be plugged into a car's data port and submits data via Bluetooth to the driver's smartphone. Automatic collects and analyses this data to provide feedback on driving styles.

Both B2C and B2B business models can be found in this cluster. As some of these companies sell physical devices for data collection (Automatic sells data loggers for cars), some generate revenue from asset sales.

Type D: 'Free data knowledge discovery'

Type D companies create value by performing analytics on free available data. Furthermore, as not all free data is available in a machine-readable format, some companies crawl data from the Web (data generation 50%). An example of this type is Gild, which provides a service for companies facilitating developer recruitment. Gild automatically evaluates the published code on open source sites such as GitHub, as well as coders' contributions on Q&A websites like Stack Overflow. A scoring mechanism allows them to identify hidden talents.

Although the companies in this cluster are homogenous regarding key data sources and activities, their offerings vary significantly: automated monitoring of review sites for hotels (Olery); recommendation of hotel deals based on analysing different booking websites (DealAngel); and identifying relevant social media influencers (Traackr, PeerIndex). Both B2B and B2C business models can be found in this cluster. The type of analytics performed by these companies ranges from descriptive analytics (the majority) to more advanced analytics techniques. TrendSpottr, for example, applies predictive analytics to identify emerging trends on real-time data streams such as Twitter or Facebook before they reach mainstream awareness.

A variety of revenue models exists within this cluster. Besides the subscription or usage-fee-based models, companies targeting consumers also rely on revenue from advertising or brokerage fees (DealAngel receives commission from booking websites).

Type E: 'Data-aggregation-as-a-service'

Companies in this cluster create value neither by analysing nor creating data but by aggregating data from multiple internal sources for their customers. This cluster can be labelled 'aggregation-as-a-service'. After aggregating the data, the companies provide it through various interfaces (distribution: 83%) and/or visualise it (33%). The areas of application focus mostly on aggregating customer data from different sources (Bluenose) or individuals (Who@) within an organisation. Other companies focus on specific segments or problems (AlwaysPrepped helps teachers to monitor students' performance by aggregating data from multiple education programmes and websites). Similar to Type B ('analytics-as-a-service'), the revenue models of such companies are primarily subscription-based and mainly business customers are targeted.

Type F: 'Multi-source data mash-up and analysis'

Type F companies aggregate data provided by their customers with other external, mostly free, available data sources, and perform analytics on this data. The offering of these companies is characterised by using other external data sources to enrich or benchmark customer data. For example, Next Big Sound provides music analytics by combining proprietary data with external data sources such as view counts on YouTube or Facebook Likes. These business models mostly target business customers and correspondingly revenue models are often subscription-based.

4.3 Validation

To validate the DDBM framework, coding and the identified business model types, case interviews were conducted with seven randomly selected companies (Radius, AGILE, GoSquared, OpenSignal, MixRank, SiQuerries and Welovroi) from the sample. We interviewed the CEOs or CTOs using a semi-structured interview covering questions about the company's value proposition, business model, core competitor, data sources, coding cross-checking, missing components in the framework, most important component/competitive advantage, mapping of key activities and challenges, data ownership and cost structure. The case studies aimed to confirm the companies' business models and further explore the different business model types in case the secondary data were not sufficient to map the company's business model accurately.

Radius provides a service for medium and large enterprises to obtain information about current and potential customers in the SME space. The CTO described the company's key activity as being to 'mine the web for tons of information', to get 'any digital footprint associated with small to medium-size businesses, whether the government has it, or some company that you have to purchase it from, or if it's available on the web'. It matches 'all those digital footprints, into a canonical record to build…the gold standard of data'. Radius is an example of a Type A DDBM with a monthly subscription-fee revenue model. It sees its key value as being able to offer this service 'better and cheaper than anybody else', as 'a marketing organization is not going to be able to do all this, and if it's doing all this it's extremely expensive'.

AGILE provides payment data analytics as a service for SME retailers and is an example of a Type B business model. The CEO stated: 'We take the complexity and hassle away from payment companies to develop a value added solution that they can charge their clients for.' In the data source dimension, AGILE currently does not use data sources other than data obtained from its customers. A subscription revenue model is generated based on data size and number of users.

GoSquared is a real-time Web analytics service classified under Type C. The CEO described the business model: 'It is correct as we save customers a lot of time and a lot of effort in understanding and analysing their data and taking all that work away from them by doing it themselves.'[3]

MixRank is a Web-based competitive intelligence tool for advertising, mapped to DDBM Type D. The CEO stated that 'MixRank tracks millions of ads through crawling public[ly] available websites. This data is automatically analysed and categorised. Customers can get access to this data for a monthly subscription fee.'

SiQuerries is a cloud-based service allowing the combination of data from multiple data sources, such as MySQL, Amazon Redshift or Google BigQuery, to build customised reports using a drag-and-drop interface. Their CEO and co-founder describes their key value as bringing 'data from multiple sources into a single system', allowing its customers to build interactive reports quickly, which can easily be shared between different teams in one company. SiQuerries generates revenue through monthly payments from its small- and medium-sized enterprises. It is an example of a Type E business model ('data-aggregation-as-a-service'). However, SiQuerries provides the means to perform analytics with the aggregated data, suggesting an overlap between Type B ('analytics-as-a-service') and Type E business models.

An example of a Type F company is Welovroi, which allows the monitoring and analysis of marketing campaigns and provides the possibility to integrate data from various sources, including external data, and facilitates benchmarking of the campaigns. The founder and CEO of Welovroi describes its value proposition as providing a 'single dashboard to monitor and improve marketing campaigns on various channels'.

Overall, the framework proved to be a suitable tool to map the business models of the case companies. All the relevant features of the companies were largely covered. However, the case interviews revealed that the value proposition of the different companies could not be represented precisely enough, an issue that should be addressed in future research. Furthermore, the case studies showed that the data collected from publicly available sources was largely accurate. Additional information could only be collected on the revenue model, when no publicly available information was available (AGILE). The clustering solution was verified through the case studies by letting the companies identify their key competitors, which had to be in the same cluster. This was true for AGILE, GoSquared and SiQuerries, while the competitors of the other cases were mostly incumbents that were not included in the sample.

5. Managerial implications

The study provides a series of implications that may be particularly helpful to companies already leveraging 'big data' for their businesses or planning to do so. The DDBM framework (Figure 1) represents a basis for the analysis and clustering of big-data-related business models. Its dimensions and features span a field of possibilities that practitioners may use to structure their own DDBMs. The framework allows the identification and assessment of potential data sources, provides comprehensive sets of potential key activities, data-related offerings and revenue models. Thus, the framework specifies existing business modelling approaches with aspects of data-driven firms.

Moreover, the identified six DDBM types provide a systematic overview of the different ways to create DDBMs. Thus, they can serve as an inspiration and a blueprint for companies considering the creation of their own DDBMs. Furthermore, practitioners face the issue of recognizing the extent to which data-driven activities are actually under their control. Therefore, the taxonomy allows practitioners to position their own business in a competitive landscape and facilitates the identification of potential gaps in the market. This can be achieved *ab initio* or with inspiration from existing DDBM examples discussed here, the latter allowing an organisation to benefit from proven guidelines in similar organisations that have been successful with DDBM implementation. Further studies, combined with an increasing sample size, may serve as market research services and suggest greater innovation and creativity.

Although this paper focusses on DDBMs in start-ups, we presume that the key findings also apply to established organisations. Key differences exist for established companies compared to start-up companies. Large existing firms have to contend with ingrained company structure, culture and traditional revenue streams. It is the competitive advantage associated with effective big data utilisation that drives the desire for existing mainstream businesses to become more data-driven. Thus, the systematic DDBM framework potentially enables established organisations and business start-ups to transform an existing business or innovative data-driven ideas into a feasible DDBM.

6. Theoretical implications

Addressing the aforementioned gaps in the literature, this paper contributes to an understanding of the nature of data-driven business models through an exploratory and descriptive study. It offers the scientific community a specific framework of business model features, comprising six dimensions with thirty-five variables, and a taxonomy. Thus, the paper may serve as a reference point for further studies and theory development in the field

[3] A further case interview was conducted with OpenSignal (also Type C)

of data-driven business models and value creation from big data. Six levels of future research can be identified: First, on an operational level, comprehending the overview on DDBMs by increasing the size and width of the sample or focusing on specific industries may lead to increasingly robust clusters. Second, evolutionary paths of companies should be analysed to understand why companies are transiting between different types of DDBM. Third, longitudinal studies, particularly those including (financial) performance indicators, may shed light on the dynamics and evolution of DDBMs. Fourth, a more theoretical grounded discussion could be considered. For instance, Markides (2013), embarking on ambidexterity literature, suggests utilising strategies like spatial, temporal and contextual separation to run conflicting business models synchronously. These strategies create quasi-start-up environments in firms with established business models. Following this idea, it might be beneficial for organisations to leverage business model patterns from the start-up scene and run them within the quasi-start-up environments. One can hypothesise that utilising the framework and taxonomy from this paper positively impacts performance variables such as 'time to market' of established organisations in the context of DDBM-related endeavours. Fifth, the study has identified clusters of currently existing DDBMs. However, future business models are not limited to these clusters. Hypothesising on successful future business models may provide valuable inspiration for research and practice. Sixth, DDBM type F is particularly interesting, as companies from this cluster seem to intelligently integrate a variety of data sources and activities - most likely from an alliance of partners and customers. Future research may investigate promising forms of alliances with complementary data sources and related activities.

7. Conclusion

This paper provides the first empirically derived taxonomy of data-driven business models (DDBMs) in start-ups. The proposed DDBM framework and the derived DDBM types can be used to create new business models for companies. While the framework is deducted from the latest research on business model innovation and documented specificities of big data, the taxonomy results from the application of the framework to over three hundred documents describing DDBMs of start-up firms. Together, framework and taxonomy add a new perspective to the – primarily technically discussed – topic of big data.

However, three types of limitations should be addressed in future research: constraints regarding (1) the diversity and size of the analysed sample; (2) the application of a framework derived from the literature; and (3) the static view of the start-up business models. In terms of sample diversity, retrieving start-up companies from further portals beyond AngelList might increase the sample's diversity and the variety of business models observed. Likewise, expanding the study's scope from start-ups to established companies incidentally creating and collecting data via their core business may also lead to a broader taxonomy of DDBMs. Moreover, the significant amount of manual work in this study limited the sample size to 100 start-up companies, prohibiting testing of the study's findings for external validity using a split sample. The framework used for coding and clustering reduces the complexity of the companies to a limited number of binary features. While this supports the study's exploratory nature, it potentially neglects dimensions that are less prominently discussed in the literature. As the purpose of this research was to take a snapshot of business models, the dynamic perspective of business models was ignored. However, business models, particularly of start-ups, frequently change and evolve over time.

References

Al-Debei, M. and Avison, D. (2010). 'Developing a unified framework of the business model concept', *European Journal of Information Systems,* Vol. 19, No. 3, pp. 359–376.

Amit, R. and Zott, C. (2001). 'Value creation in E-business', *Strategic Management Journal*, Vol. 22, No. 6, pp. 493–520.

Baden-Fuller, C. and Haefliger, S. (2013). 'Business Models and Technological Innovation', *Long Range Planning*, *46*(6), 419–426.

Barnes, C., Blake, H. and Pinder, D. (2009*). Creating & delivering your value proposition. Managing customer experience for profit,* Kogan Page, London.

Barney, J. (1991). 'Firm Resources and Sustained Competitive Advantage', *Journal of Management*, Vol. 17, No. 1, pp. 99–120.

Bouwman, H. et al. (2008). 'Conceptualizing The STOF Model', in Bouwman, H. et al., *Mobile Service Innovation and Business Models,* Springer, Berlin Heidelberg, 31–70.

Bouwman, H. and MacInnes, I. (2006). 'Dynamic business model framework for value webs'. Paper presented at the 39th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, 4 January 2006–7 January 2006.

Bouwman, H. and MacInnes, I. (2006). 'Dynamic Business Model Framework For Value Webs', *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*.

Burkhart, T., Krumeich, J., Werth, D and Loos, P. (2011). 'Analyzing the Business Model Concept – A Comprehensive Classification of Literature', *ICIS 2011 Proceedings*.

Buytendijk, F., Kart, L., Laney, D., Jacobson, S., Lefebure, S. and Hetu, R. (2013). 'Toolkit: Big Data Business Opportunities From Over 100 Use Cases', *Gartner (G00252112)*.

CEBR (2012). 'Data equity – Unlocking the value of big data', *Centre for Economics and Business Research Ltd.*

Chen, Y., Kreulen, J., Campbell, M. and Abrams, C. (2011). 'Analytics Ecosystem Transformation: A Force for Business Model Innovation', *SRII Global Conference 2011*, pp. 11–20.

Chesbrough, H. and Rosenbloom, R. (2002). 'The role of the business model in capturing value from innovation: Evidence from Xerox Corporation's technology spin-off companies', *Industrial and Corporate Change*, Vol. 11, No. 3, pp. 529–555.

Choi, S., Cha, S. H. and Tappert, C. (2010). 'A Survey of Binary Similarity and Distance Measures', *Journal on Systemics, Cybernetics and Informatics, Vol.* 8, No. 1, pp. 43–48.

Cios, K. J. (2007). *Data mining. A knowledge discovery approach*, Springer, New York.

Cooper, H. M. (1988). 'Organizing knowledge syntheses: A taxonomy of literature reviews', *Knowledge in Society 1*(1), pp. 104–126.

Criscuolo, P., Nicolaou, N. and Salter, A. (2012). 'The elixir (or burden) of youth? Exploring differences in innovation between start-ups and established firms', *Research Policy*, Vol. 41, No. 2, pp. 319–333.

Davenport, T. H. (2006). 'Competing on analytics', *Harvard Business Review*, Vol. 84, No. 1, p. 98.

Delen, D. and Demirkan, H. (2013). 'Data, information and analytics as services', *Decision Support Systems*, Vol. 55, No. 1, pp. 359–363.

De Reuver, M., Bouwman, H. and Haaker, T. (2013). 'Business model roadmapping: A practical approach to come from an existing to a desired business model', *International Journal of Innovation Management*, Vol. 17, No. 1.

Diebold, S. (2012). 'Know the difference between data-informed and versus data-driven', available at http://stevendiebold.com/know-the-difference-between-data-informed-and-versus-data-driven/ (accessed 10/06/2013).

El Sawy, O. and Pereira, F. (2012). 'Business Modelling in the Dynamic Digital Space: An Ecosystem Approach', 2012, Springer Verlag, Berlin Heidelberg.

Everitt, B., Landau, S. and Leese, M. (2011). *Cluster analysis*, Oxford University Press, London.

Fastoso, F. and Whitelock, J. (2010). 'Regionalization vs Globalization in Advertising Research: Insights from Five Decades of Academic Study', *Journal of International Management*, Vol. 16, No. 1, pp 32–42.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine,* Vol. 17, pp. 37–54.

Flyvbjerg, B. (2006). 'Five Misunderstandings About Case-Study Research', *Qualitative Inquiry 12*(2), pp. 219–245.

Gantz, J. and Reinsel, D. (2012). 'The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East', *IDC*.

Gartner (2012). 'IT Glossary – Big Data', available at http://www.gartner.com/it-glossary/big-data/ (accessed 5/06/2013).

Gartner (2013). 'IT Glossary – Crowdsourcing', available at http://www.gartner.com/it-glossary/crowdsourcing (accessed 12/08/2013).

Hagen, C., Khan, K., Ciobo, M., Miller, J., Wall, D., Evans, H. and Yadava, A. (2013). 'Big Data and the Creative Destruction of Today's Business Models', *ATKearney*.

Hambrick, D. C. (1984). 'Taxonomic Approaches to Studying Strategy: Some Conceptual and Methodological Issues', *Journal of Management,* Vol. 10, No. 1, pp. 27–41.

Han, J., Kamber, M. and Pei, J. (2011). *Data Mining. Concepts and Techniques,* Elsevier Science, Burlington.

Hedman, J. and Kalling, T. (2003). The business model concept: Theoretical underpinnings and empirical illustrations, *European Journal of Information Systems,* Vol. 12, No. 1, pp. 49–59.

Heikkilä, J., Heikkilä, M. and Tinnilä, M. (2008). 'The Role of Business Models in Developing Business Networks', in Becker, A. (ed.), *Electronic Commerce: Concepts, Methodologies, Tools, and Applications,* pp. 221–231.

Holsapple, C., Lee-Post, A. and Pakath, R. (2014). *A unified foundation for business analytics. Decision Support Systems*, 64, 130–141.

Johnson, M. W., Christensen, C. and Kagermann, H. (2008). 'Reinventing your business model', *Harvard Business Review,* Vol. 86, No. 12, pp. 57–68.

Johnson, R. (1997). 'Examining the validity structure of qualitative research', *Education,* Vol. 118, p. 282.

Kart, L., Heudecker, N. and Buytendijk, F. (2013). 'Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype', *Gartner*.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*, Wiley-Interscience, Hoboken, NJ.

Ketchen, D. J. and Shook, C. L. (1996). 'The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique', *Strategic Management Journal,* Vol. 17, No. 6, pp. 441–458.

Kosala, R. and Blockeel, H. (2000). 'Web Mining Research: A Survey', *SIGKDD Explor. Newsl.,* Vol. 2, No. 1, pp. 1–15.

Lakomaa, E. and Kallberg, J. (2013). 'Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs', *Access IEEE,* Vol. 1, pp. 558–563.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byres, A. (2011). 'Big data: The next frontier for innovation, competition, and productivity', *McKinsey Global Institute*.

Markides, C. (2013). 'Business Model Innovation: What Can The Ambidexterity Literature Teach Us', The Academy of Management Perspectives, Vol. 27, No. 4, pp. 313–323.

McAfee, A. and Brynjolfsson, E. (2012). 'Big Data: The Management Revolution', *Harvard Business Review,* Vol. 10.

Miligan, G. W. (1996). 'Clustering Validation: Results and Implications for Applied Analyses', in Arabie, P., Hubert, L. J. and de Soete, G. (eds), *Clustering and classification*, World Scientific, Singapore, pp. 341–376.

Mooi, E. and Sarstedt, M. (2011). 'Cluster Analysis', *A Concise Guide to Market Research*, Springer, Berlin Heidelberg, pp. 237–284.

Morris, M., Schindehutte, M. and Allen, J. (2005). 'The entrepreneur's business model: Toward a unified perspective', *Special Section: The Nonprofit Marketing Landscape,* Vol. 58, No. 6, pp. 726–735.

Negash, S. (2004). 'Business Intelligence', *Communications of the Association for Information Systems,* Vol. 13, pp. 177–195.

Osterwalder, A. (2004). 'The Business Model Ontology. A Proposition in Design Science Research', these. Ecole des Hautes Etudes Commerciales de l'Université de Lausanne.

Osterwalder, A., Pigneur, Y. and Clark, T. (2010*). Business model generation. A handbook for visionaries, game changers, and challengers,* Wiley, Hoboken, NJ.

Otto, B. and Aier, S. (2013). 'Business Models in the Data Economy: A Case Study from the Business Partner Data Domain', *Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI 2013) in Leipzig*, Vol. 1, pp. 475–489.

Petter, J. and Peppard, J. (2012). 'Harnessing the Growth Potential of Big Data. Why the CEO Must Take the Lead', *EMC*.

Pham, D. T., Dimov, S. S. and Nguyen, C. D. (2005). 'Selection of K in K-means clustering', *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science,* Vol. 219, No. 1, pp. 103–119.

Rao, L. (2013). 'Paul Graham: 37 Y Combinator Companies Have Valuations Of Or Sold For At Least $40M', Techcrunch.com, 26 May, available at http://techcrunch.com/2013/05/26/paul-graham-37-y-combinator-companies-have-valuations-of-or-sold-for-at-least-40m/ (accessed 07/02/2014).

Rayport, J. F. and Sviokla, J. J. (1995). 'Exploiting the Virtual Value Chain', *Harvard Business Review,* Vol. 73, No. 6, pp. 75–85.

Rotella, P. (2012). 'Is Data The New Oil?', Forbes, 2 April, available at http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/ (accessed 07/02/2014).

Rousseeuw, P. J. (1987). 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics,* Vol. 20, pp. 53–65.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. and Tufano, P. (2012). 'Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data', *IBM Institute for Business Value, Saïd Business School at the University of Oxford*.

Singh, R. and Singh, K. (2010). 'A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing', *IJCSI International Journal of Computer Science,* Vol. 7, No. 3, pp. 41–50.

Stipic, A. and Bronzin, T. (2012). 'How cloud computing is (not) changing the way we do BI', *Proceedings of the 35th International Convention MIPRO 2012*, pp. 1574–1582.

Stuckenberg, S., Fielt, E. and Loser, T. (2011). 'The impact of software-as-a-service on business models of leading software vendors: Experiences from three exploratory case studies', *Proceedings of the 15th Pacific Asia Conference on Information Systems (PACIS 2011), Queensland University of Technology*.

Svrluga, B. (2012). 'Data Data Blah Blah – It Ain't That Easy', available at http://bradsvrluga.com/2012/03/27/data-data-blah-blah/ (accessed 4/06/2013).

Tashakkori, A. and Teddlie, C. (2002). *Handbook of mixed methods*, SAGE, London.

Teece, D. J. (2010). 'Business Models, Business Strategy and Innovation', *Long Range Planning,* Vol. 43, No. 2–3, pp. 172–194.

WEF (2011). 'Personal Data: The Emergence of a New Asset Class', *The World Economic Forum*, available at http://www.weforum.org/reports/personal-data-emergence-new-asset-class (accessed 18/02/2014).

Weill, P., Malone, T. W. and Apel, T. G. (2011). 'The business models investors prefer', *MIT Sloan Management Review,* Vol. 52, No. 4, p. 17.

Wernerfelt, B. (1984). 'A resource-based view of the firm', *Strategic Management Journal,* Vol. 5, No. 2, pp. 171–180.

Wollan, M. (2011). 'Matchmaking for Web Start-Ups and Investors', *The New York Times*, 7 March, pp. B3.

Zott, C., Amit, R. and Massa, L. (2011). 'The Business Model: Recent Developments and Future Research', *Journal of Management,* Vol. 37, No. 4, pp. 1019–1042.

Table 1 Review of different business model frameworks

Author(s) Year	Value proposition/ offering	Key resource	Key activity	Market/ customer segment	Revenue stream	Cost structure	Other elements	Citations (Google Scholar, 19.01.2014)
Chesbrough and Rosenbloom, 2002	✓			✓	✓	✓	Value chain, value network, competitive strategy	1735
Hedman and Kalling, 2003	✓	✓	✓	✓			Competitors, scope of management	456
Osterwalder, 2004	✓	✓	✓	✓	✓	✓	Customer relationship, channels, key partner	1001
Morris et al., 2005	✓	✓		✓	✓	✓	Competitive strategy factors, personal factors	846
Johnson et al., 2008	✓	✓	✓		✓	✓	-	641
Al-Debei and Avison, 2010	✓	✓	✓	✓	✓	✓	Value network	116

Table 2 DDBM cluster results and respective medoids

Cluster		1	2	3	4	5	6	7
Data Source	Acquired data	0	0	1	0	0	0	0
	Customer-provided data	0	1	1	0	0	1	1
	Free available data	1	0	1	0	1	0	1
	Crowd sourced	0	0	0	0	0	0	0
	Tracked, generated & other	0	0	0	1	0	0	0
Key Activity	Aggregation	1	0	0	0	0	1	1
	Analytics	0	1	1	1	1	0	1
	Data acquisition	0	0	1	0	0	0	0
	Data generation	0	0	0	1	0	0	1
Number of companies		17	28	5	16	14	6	14
Cluster Types		**A**	**B**	**-**	**C**	**D**	**E**	**F**

Table 3 DDBM clusters general statistics

DDBM Dimensions					DDBM Cluster Types Percentages					
					A	B	C	D	E	F
					Free data collector and aggregator	Analytics-as-a-service	Data generation and analysis	Free data knowledge discovery	Data-aggregation-as-a-service	Multi-source data mash-up and analysis
	Share of companies (as percentage of total sample)				**18%**	**29%**	**17%**	**15%**	**6%**	**15%**
DDBM Dimensions	Data Source	Internal	Existing data		0%	0%	0%	0%	0%	0%
			Self-generated data	Crowd-sourced	18%	0%	31%	0%	17%	7%
				Tracked, generated and other	6%	11%	88%	0%	0%	21%
		External	Acquired data		12%	11%	0%	7%	0%	21%
			Customer provided		24%	100%	13%	0%	100%	79%
			Free available	Open data	100%	11%	19%	100%	0%	93%
				Social media data	12%	4%	0%	7%	0%	7%
				Web crawled data	71%	11%	13%	50%	0%	57%
	Key Activity	Data generation	Crawling		35%	0%	6%	50%	0%	43%
			Tracking & crowdsourcing		12%	4%	100%	0%	17%	29%
		Data acquisition			24%	21%	6%	21%	17%	29%
		Processing			100%	18%	0%	0%	100%	93%
		Aggregation			12%	82%	69%	86%	0%	93%
		Analytics	Descriptive		0%	50%	13%	21%	0%	14%
			Predictive		0%	11%	6%	7%	0%	7%
			Prescriptive		24%	39%	44%	7%	33%	36%
		Visualisation			100%	36%	19%	29%	83%	50%
		Distribution			35%	0%	6%	50%	0%	43%
	Offering	Data			12%	0%	0%	0%	0%	14%
		Information/ knowledge			88%	100%	88%	100%	83%	93%
		Non-data product/service			0%	0%	13%	7%	0%	0%
	Target Customer	B2B			71%	96%	63%	86%	83%	86%
		B2C			47%	18%	50%	21%	33%	21%
	Revenue Model	Asset Sale			0%	0%	19%	0%	0%	0%
		Lending/renting/leasing			6%	0%	0%	0%	0%	0%
		Licensing			0%	0%	0%	0%	0%	0%
		Usage fee			12%	14%	13%	21%	0%	0%
		Subscription fee			47%	46%	44%	64%	33%	36%
		Advertising			12%	0%	6%	7%	0%	7%