Model and Experimental Development for Business Data Science

Russell Newman [1], Victor Chang [2], Robert John Walters [1], Gary Brian Wills [1]

1. Electronics and Computer Science, University of Southampton, Southampton, UK
{rn2, rjw1, gbw}@ecs.soton.ac.uk
2. Xi'an Jiaotong Liverpool University, Suzhou, China
ic.victor.chang@gmail.com

Abstract

While Data Science has become increasingly significant for business strategies, operations, performance, efficiency and prediction, there is little work on this to provide a detailed guideline. We have proposed a Business Data Science (BDS) model that focuses on the model and experimental development that allows different types of functions, processes and roles to work together collaboratively for efficiency and performance improvements. Details with examples have been illustrated to show that BDS model can be a robust model. Future directions have been discussed to ensure that business intelligence, security, analytics and research contributions to BDS can be achieved.

Keywords: Business Data Science (BDS); Modelling and Experimental techniques for BDS; Future directions and contributions for Data Science

1. Introduction

Data Science is an interdisciplinary area to enable experts in different domains to study and work together (Borrego and Newswander, 2010; Provost and Fawcett, 2013 a; Ericsson, 2014). Outputs from all kinds of work can generate data in different types of formats. It has become apparently obvious that the processing, analysis and presentation of data outputs will be important to a growing number of sectors involved (Agresti and Kateri, 2011). The main reason why Data Science makes attractive to businesses is: Data Science is a study of the data that has involved processing, analysis, interpretation and making sense of the data (Han et al., 2011; Gelman et al., 2014). Businesses can understand their problems, their business performance (daily, weekly, monthly and yearly) and forecast of their business performance within a matter of minutes at any time (Provost and Fawcett, 2013 a).

The role of Data Science has become increasingly important for businesses as follows. Firstly, Data Science allows businesses to collect and analyse data about their business operations, strategies and overall performance (McAfee et al., 2012). Secondly, business can improve on their services, operations, strategies and business performance based on the outputs of analysis (Nath et al., 2010). Thirdly, businesses can improve the quality of their predictive modelling, so that decision-makers can plan for suitable strategies for their companies (Dhar, 2013). There are three major benefits of doing so, however, the ways to

execute Business Data Science (BDS) are not established as yet since existing literature does not have a conclusive guidelines or a summary of best practice approach. Although there are many organisations that have become interested in Data Science, they do not know how to operate and manage Data Science (McAfee et al., 2012). This has motivated us to present our case of BDS, particularly in the way that organisations can adopt. Furthermore, a structured guideline is useful for development of any projects and services.

In order to demonstrate effectiveness for businesses, our research is focused on development of relevant modelling and simulation techniques, to provide organisations a bridge and a smooth transition to the adoption of Business Data Science (BDS). The fundaments of these techniques are then used to construct a BDS model as explained throughout the paper. To ensure a BDS model can work effectively with business activities, modelling and simulation techniques are required to be investigated to ensure business models, functions, processes with different roles of people involved can be resilient and robust. The structure of this paper is as follows. Section 2 explains the definitions, scopes, components, functions and overall approach towards modelling and simulation techniques for BDS. Section 3 presents the experimental design for BDS model that blends business intelligence, investigation to economic bubbles and other related areas. Section 4 presents four topics of discussions and Section 5 sums up this paper with the future work described.

2. Modelling and Simulation Techniques for Business Data Science

Modelling and simulations techniques are useful for business to stay competitive, efficient and collaborative. Understanding the terminologies, including what each term means and how each terminology offers is also relevant for business growth and sustainability (Chang, 2015 a). Their definitions are as follows.

Simulating a system enables analysis of various situations by modelling them, over time, within a computer program (Banks, 1998).

A model is a “representation of an event and/or things that is real (a case study) or contrived (a use case)”. A simulation is “a method for implementing a model over time” (Banks, 2009).

A simulation may be run multiple times, to investigate how differing conditions alter the outcome. The competency to manage and master business data science has become significant for organisations that adopt business intelligence and analytics approach (Chen et al., 2012).

The word “system” denotes what, from the real world, is being simulated. A system may be broken down into its composite elements (such as people, machines and resources).

The “system model” is the simulated representation of the real-world system. System models are designed and built in such a way that a computer can perform calculations upon them and effectively run a simulation (Banks, 2009). When interpreting a system to build a model, an important consideration is how the various elements are to interact and affect one another (Cellier and Greifeneder, 2013).

The actual *method* for designing the model depends upon which simulation technique is employed. Regardless of the technique employed, a model designer must determine the level and areas of detail for the model, known as *scope*.

2.1 Scope

Scoping is the process of deciding which components should be included in a simulation models, and at what level of detail, and which components should be left out, simplified or abstracted (Sokolowski, 2009).

Recording a real-world system into a quantified model means that some concessions and assumptions must often be made. For instance, a certain component of a real-world system could be implemented fully in a simulation, resulting in a theoretically accurate simulation, at the cost of including many elements to represent the system. Alternatively, the same component may be modelled using a simpler implementation that is perhaps abstracted or makes some assumptions, without compromising the accuracy of the rest of the model (Cellier and Greifeneder, 2013).

This example demonstrates how detail or depth of scope must be decided. A model designer must choose exactly how much detail should be expressed in a model; greater detail may lead to a more accurate model, but at the same time create potentially unnecessary work in situations where a less detailed model would be sufficient.

The breadth of the model must also be scoped. A system model may contain modelled representations of many external entities that influence the core elements of the simulation. Including more of these may increase accuracy, again at the expense of time and design complexity. Alternatively, excluding more of these may result in an adequate simulation model and a saving of time.

2.1.1 System Dynamics

System Dynamics is a method of quantifiably modelling and simulating complex systems. It was developed by Jay Forrester at the MIT Sloan Management School, which was founded to exploit a fusion of engineering tools and techniques with traditional management. It was initially developed as a means of identifying the factors that make up the success or failure of a corporation or group of people (Forrester, 1997). System Dynamics models were originally processed by hand, but the technique was later adapted to take advantage of computer processing.

To build a System Dynamics simulation, a design progresses through two distinct stages: building causal loops and translating these to stocks and flows. It is possible to skip the first stage, but this would also exclude valuable analysis of the system which can lead to a higher quality simulation. Sterman's (2001) work on System Dynamics provides a succinct description of the methodology, which is used to inform the following sections.

2.1.2 Causal Links and Loops

At this first stage of design, a causal loop diagram is designed showing the various system variables and influences between each of them (Sterman, 2001). The causal loop diagram does not quantify any of the variables, but does denote whether variables positively or negatively affect each other. This is done through *feedback links*.

A feedback link is represented by an arrow. It signifies that the variable at its origin affects the variable at its destination in a positive *or* negative way relative to changes at the origin.

- A positive link means that as the item at the origin increases, the item at the destination may increase as well. If the origin were to decrease, the item at the destination would also decrease.
- A negative link means that as the item at the origin increases, the item at the destination may decrease. If the origin were to decrease, the item at the destination would increase.

Figure 1 shows a simple causal loop diagram modelling the causes and effects between variables of birth rate, death rate and living population.

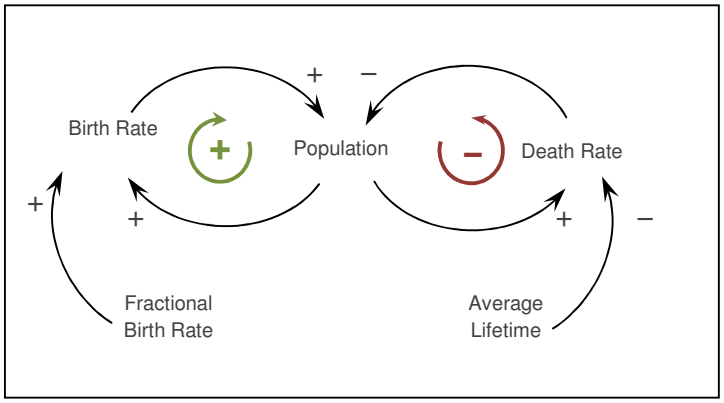

Figure 1 · Population System Dynamics Model
Derived from Fontaine et al. (2009)

Using these constructs, loops are not an uncommon feature in designs. A loop emerges when feedback links connecting variables form a closed path. Loops may be categorised according to whether they result in a positive or negative affect after all components of the loop have been evaluated.

- A positive (reinforcing) loop is denoted with a + (shown in green in Figure 1). It emerges when, after evaluating all the links in the loop, there is an overall positive effect upon all variables within. This can lead to exponential growth of variables, if not mediated with other links.
- A negative (balancing) loop is denoted with a – (shown in red in Figure 1). It emerges when, after evaluating all the links in the loop, the variables inversely affect each other.

For instance, in Figure 1, the green positive loop would cause an exponential Population increase, if the Death Rate variable was greater than the birth rate. The red negative loop will result in an eventual balance between Population and Death Rate, assuming the Birth Rate does not change.

Once the diagram is complete, it may be evaluated to ensure logical and theoretical integrity.

At this stage, the magnitude of the affect between variables is not important, and nor is the value of each variable. The aim of the exercise is to produce a reliable cause and effect structure for the model, according to reality (Coyle, 1999). Next, the model is adapted with

stocks and flows, which enable variables to be quantified. Understanding all these features may contribute to the development of Business Data Science (BDS).

2.1.3 Stocks and Flows

The causal link diagram indicates direct and inverse links between variables, but no magnitude of effect, and no value of the variables. To overcome this problem, the model may be adapted with stocks and flows (Sterman, 2001).

A “stock” is a variable that is annotated with the quantity present at any given time. The quantity may grow or shrink over time, depending upon how it is connected to other stocks in the model.

A “flow” connects stocks and is annotated with a flow rate, representing the rate of change of the stock. Items will always travel through the Flow, as fast as the rate permits (which may be zero) and provided items are present at the source stock.

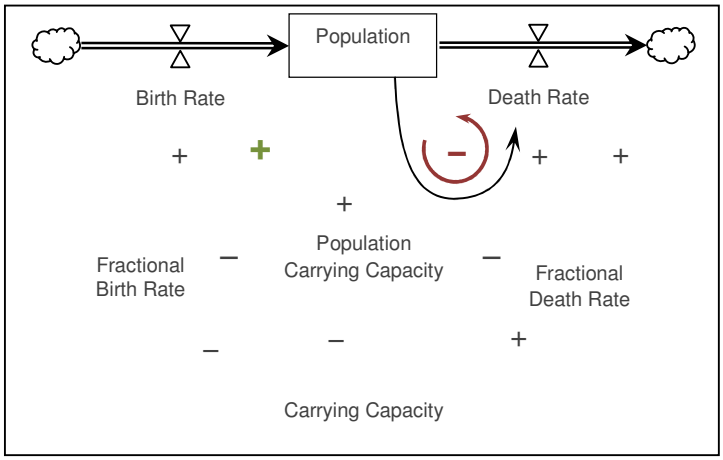

Figure 2 · Stock and Flow adaptation of Figure 1
Derived from Fontaine et al. (2009)

Figure 2 shows a stock and flow adaptation of Figure 1. The principal model variable, Population, is converted to a stock. The positive and negative loops identified in Figure 1 are shown in their new positions within the modified diagram. Some adaptations have been made to the variables. For instance, *carrying capacity* of the overall species has been added, as has the carrying capacity of the Population shown in the stock.

Forrester defines two categories of system (Forrester, 1994);

- In open systems, the outputs of the system have no effect upon its inputs, meaning the performance of the system does not result in any changes for the system.
- In closed systems, the outputs of the system do affect the inputs, so the performance of the system determines how it will behave in the future. The population models in Figure 1 and Figure 2 are closed systems.

Finding the “right” way to construct a system model involves consulting literature to discover evidence that supports design appropriate to solve the question asked of it. This is discussed further in the next section.

2.2 Proving the Design

The results of a simulation cannot be trusted unless they are accurate and reliable, and the model is correctly designed according to theory of the system subject field. To prove the model is suitable fore Business Data Science, validation is required. The successful completion of validation deems a model academically credible for analysing "what-if" scenarios.

Models may require adaptations to be properly validated. Forrester suggests that an effective approach to this is to implement improvements to a model only if they enhance the modelling of the real world, and not just for the sake of fixing a problem (Forrester, 1997).

Coyle and Exelby concur, mentioning that the model is always "a simplification of reality which is intended to serve some useful purpose". This means that a model cannot be deemed true or false; it can merely be deemed fit for purpose or not. As an example, they explain that Newton's model of gravity is adequate for many applications, but is invalidated by some modern branches of physics (Coyle & Exelby, 2000).

Models are typically proven using a two-step process involving *verification* and *validation*.

- Verification "[determines whether] an implemented model is consistent with its specification" (US Department of Defence, 1996), i.e. whether the simulated model accurately reflects the model design. Verification also determines whether a model is fit for the purpose for which it was made, and that the design transformations it has undergone (causal loops and stock/flows, for instance) are accurate (Petty, 2009).

- Validation checks to ensure that the simulation model consistently produces suitably accurate results when tested against data from trusted literature. Validation also checks that the model design is a suitably accurate reflection of the real system (Petty, 2009).

Hence, the population example model mentioned earlier (when verified and validated) may be used to analyse hypothetical fluctuations in various birth rates, and changes within other variables of the model.

Forrester (the creator of System Dynamics) described a process which is specifically designed specifically for formally validating System Dynamics models (Barlas, 1994). This process should be used in preference over a generalised two-step verification and validation process.

Forrester's process splits testing into branches of *Structure Validity* and *Behaviour Validity*. Structure Validity is broken down into two possible approaches; *Direct Structure Tests* and *Structure-oriented Behaviour Tests* (Barlas, 1996). These two approaches differ in their testing methods; one represents white-box testing, where internal components are scrutinised, and the other represents black-box testing, where outputs are scrutinised and internal components are ignored (Barlas & Kanar, 1999). They are discussed further below.

2.2.1 Direct Structure Tests

Direct Structure Tests are white-box testing methods. That is, they analyse the *internal components* of the simulation model (a "white", open box) to ensure they accurately reflect reality according to literature (Barlas, 1994, 1996).

Structure Confirmation

Structure Confirmation aims to establish the validity of the model *structure* by comparing it to what is known about the situation in reality. Each relationship (i.e. flows, feedbacks and logic) in the model must be checked against available literature. This testing is typically qualitative in nature, and requires adaptation to the model that is being validated.

Parameter Confirmation

Parameter Confirmation complements Structure Confirmation by checking that the numeric values of parameters in the model are adequately accurate, according to literature. In this case, "parameters" refers to values in the simulation model that are used to evaluate the relationships between components.

Direct Extreme Condition Testing (DECT)

DECT checks logical operations and equations within the model to ensure they handle extreme input values gracefully, without producing flawed output. Barlas provides a succinct example to help explain this; in a model representing an economy, if population is set to zero, then there should be no births, no consumption and no workers. Likewise, death rates must rise if an extreme level of pollution is simulated.

Dimensional Consistency

Checking for Dimensional Consistency involves analysing all formulae in the model to ensure that the left and right side of the equations are balanced (i.e. that they are actually equal).

2.2.2 Structure-Oriented Behaviour Tests

Structure-Oriented Behaviour Tests are a form of black-box testing. That is, they analyse results produced by the simulation model to ensure they concur with results in literature. No focus is placed upon the internal components of the model (the "black", closed box) (Barlas, 1994, 1996).

Indirect Extreme Condition Testing (IECT)

IECT is similar to Direct Extreme Condition Testing, but rather than analysing the internal components for correct behaviour under extreme conditions, only the outputs of the simulation are scrutinised. The outputs will be compared against proven results in literature (Meyers, 2010).

Behaviour Sensitivity Test (BST)

BST involves finding which parameters the system model is particularly sensitive to (i.e. those which cause large changes in the simulation output relative to the size of the parameter change). When sensitivities have been identified, they are checked against the real-world system to find if the same sensitivities are exhibited.

Modified Behaviour Prediction (MBP)

MBP involves identifying a similar but validated simulation that closely matches the system model of the one under test, and modifying the structure of both models in the same way. To pass the test, both systems should produce similar output even after the adaptation.

2.2.3 Software demonstrations for Business Data Science

We have an in-house graphical software package. Due to the agreement and ethical approval, the anonymous status of the collaborator cannot be enclosed. The purpose is to demonstrate the Business Data Science (BDS) model, which is designed for building and simulating business-related models to contribute to understanding of business functions, roles, processes, performance, efficiency and forecasting. If successful, business can gain a better understanding and a quicker response to their own business demands. Many of the findings in this section are from primary use and experimentation with the product.

The default toolset for simulation within this package is a set of components that adhere to no particular simulation theory, but are specifically designed for simulating workflows. Components are laid out in the simulation and connected with trunks.

Table 1 · Overview of Key Native Simulation Objects in Business Data Science model

Work Entry Point	Where work items enter the system, at the defined rate and distribution.
Queue	Where work items remain until they are able to be accepted at the next component in the workflow. Queues typically appear before Work Centres, and are instrumental in identifying bottlenecks in a workflow.
Work Centre	Where work items are processed by one or more Resources, according to the time taken to process each item and an optional standard deviation.
Resource	A finite collection of similar resource items that can be used to complete work. Resource is designed to simulate a human workforce, but may be used to simulate machinery as well. Items of Resource are deployed to their connected Work Centres as needed, with no further deployment possible once the Resource is depleted. Items of Resource may be released from a Work Centre once their work is complete.
Resource Pool	Resource Pools are connected to Work Centres and Resources. They enable a single Work Centre to share multiple Resource items, and to prioritise access to each type of resource. This can be used to model differing types of Resource that are all involved at a single Work Centre.
Work Exit Point	Where completed work items leave the system.
Route	Connects two of the above components showing the route taken by work through the system.

To understand how these features work together, a basic model was constructed depicting a high-traffic volume website and the servers that run it, and is shown in Figure 3. Its purpose is to identify how a hypothetical data centre handles large volumes of requests from many users.

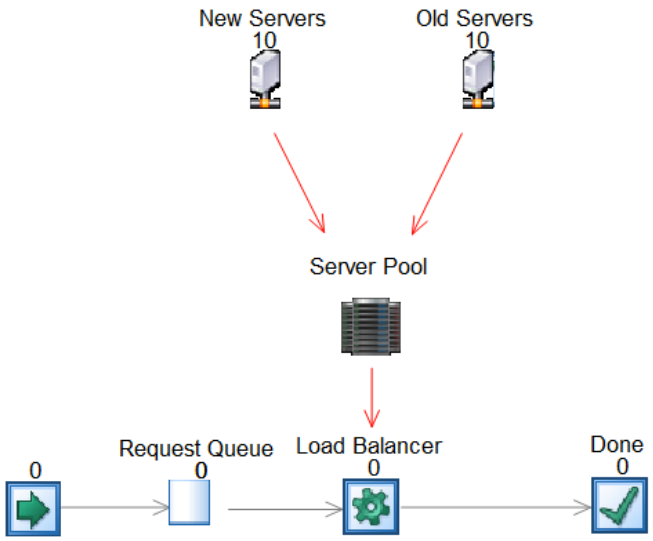

Figure 3 · Native Model of a website by recommendations of the BDS model

Requests for webpages (sent by users of the site) are modelled as Work Items, which progress through the system. They enter the model at the Work Entry Point. Work Items are then held in a Request Queue until they can be processed at the Work Centre. The Load Balancer represents the Work Centre, which distributes work items to servers.

This was initially modelled using just one type of Resource, but is now a Resource Pool of two types of Resource; *Old Servers* and *New Servers*. The *New Servers* Resource contains the same number of servers as the *Old Servers* Resource, but *New Servers* has been configured to complete work at twice the speed of the *Old Servers* Resource, simulating improved processing capacity with newer hardware. Both of these Resources are available for use in the Work Centre.

Arrows in black represent Routes along which work items flow, and are functional model components; without them, no work can flow through the model. The red arrows have been drawn on to show how Resources are linked to the Work Centre, but these have no functional involvement in the model.

This model has been scoped to the level of the data centre where the web servers reside. To increase detail within the same scope, one may model latencies and outages of specific servers. Alternatively, extra data may be sought to improve the realism of constructs within the system. For instance, the arrival times and standard deviations of request arrivals may be made more realistic by sourcing these from a relevant study in literature.

To expand the scope, one may examine the data connections and other dependencies between the requester and the data centre, which may further affect the speed of response.

In this model, peak rates for work items are experienced during the daytime. Additional work entry points may be added to simulate loads arriving from different time zones. The inflow rate of work items (i.e. requests) can be adjusted, as can their distribution.

This Business Data Science (BDS) model enables the analyst to estimate how many servers are required for a given number and distribution of requests, and to test the infrastructure under load. Similarly, the analyst may identify how much request volume the current

infrastructure can handle. This can be identified by analysing the total time taken for requests to be processed, for a specific load level, and the time a request remains in the queue. At higher load levels, the request queue grows, so items remain in the queue for longer. The analyst may run this simulation aiming to meet or exceed a request processing time target.

Given the nature of System Dynamics, these components seem inappropriate. One could theoretically implement a stock and flow using BDS model's queues and routes, but these components are not built for the purpose of System Dynamics. These components are all abstracted from the underlying simulation engine to make it easier for inexperienced users to create business simulations quickly. In comparison, System Dynamics provides a basic set of simulation components which are not individually designed to model any specific scenario, but which may be connected in different ways to model various scenarios.

Therefore, a System Dynamics simulation approach for the BDS model may contain a greater number of components than an equivalent comparative BDS model, due to the generalised nature of System Dynamics. Attempting to build a System Dynamics-style model using native BDS components would be equivalent to trying to create a generalised simulation toolset using a toolset that is already abstracted from an underlying basic set. Creating generalised components from already-abstracted ones is not desirable for reasons of reliability and suitability to purpose.

However, BDS model also contains a toolset that supports System Dynamics simulations. This toolset comprises of the Stocks (called "tanks") and Flows (called "pipes"), which are connected and utilised in a similar manner to regular BDS process objects.

2.2.4 Why businesses should know?

Business processes have been used in similar areas. Business processes can be defined in the workflows that can best represent the business activities (Scheer and Nüttgens, 2000). However, assumptions are based on the facts that processes and functions can work well (Davenport, 2013). To ensure these processes and functions can work, either detailed analysis or detailed functions of the work should be undertaken. The approach we have adopted can ensure different types of activities can stay connected and each has its own purpose. In other words, more combinations of business functions can get together and serves its own purpose. It can also reduce the possibilities that different functions or units cannot work together well.

2.2.5 Summary

This section has described System Dynamics as a method of simulating complex systems that change over time. The process of developing a System Dynamics simulation approach for the BDS model involves stages of initial design on paper to ascertain the fundamental layout and connections of the system. Systems are later converted to a fully quantified digital model, containing logical operations that define how components interact.

The process of validation explained by Forrester (1994, 1997) encompasses the traditional stages of validation *and* verification. This process aims to identify various types of flaws that may be present in a system, and should lead to a system being deemed such a reasonable simulation of the real-world system that it is fit for the purpose of its creation.

The BDS model has been investigated and tested as a software package capable of generic business process simulations, as well as System Dynamics simulations in particular.

2.3 Initial Model Development

Our previous paper (Chang et al., 2016 a) has presented an overview of findings from previous bubbles, and some indications of trends in the Web 2.0 sector (Newman et al., 2016). This section aims to produce a conceptual model based upon those findings. The purpose of this model is to consolidate findings, as a means of developing a quantified technique for answering the research question.

2.3.1 Model Specification

The model is intended to satisfy the research question for this work; to find the extent to which the Web 2.0 sector has represented a bubble during its lifetime.

To achieve this, the model will need to represent bubble scenarios. As the most recent bubble analysed, literature on the dot-com bubble will be used as the theoretical basis of the model. However, the model should also encompass other scenarios that may lead to a bubble, such as those discussed in Chang et al. (2016 a).

The following two bubble characteristics, identified in our previous work (Chang et al., 2016 a), will be taken forward as indicators the model should support.

- *When a "rapid increase in investment volume [is] not based on a corresponding increase in market knowledge or corresponding decrease in investment risk".* Source: (Valliere & Peterson, 2004)

  "Investment volume" may be measured directly against the records for a public company, or by the quantity of venture capital supplied to a private company.

  "Market knowledge" cannot be directly quantified. When interpreted in terms of investors' experience and expertise within the market, it could potentially be measured in terms of the age of the market and the number of successes or failures.

  "Investment risk" or uncontrolled risk demonstrated by Chang (2014) and Chang et al. (2016 b), cannot be directly quantified. Market analysts typically use historic data as an indicator of risk in the future. However, no such historical data exists for start-up companies, so other factors must be used to generate an indicator of survival, profitability and, therefore, risk.

- *Continued investment despite the lack of a business model or market dominance (and therefore revenue) as explained by Chang et al. (2016 b), who have explained the detailed risk analysis for investment, user satisfaction and technical efficiency by the use of Organisational Sustainability Modelling (OSM).* The presence and effectiveness of a business model is difficult to quantify, however, OSM can quantify risk into uncontrolled and controlled, as well as analysing the extent of impacts and explaining the implications to the businesses. Risks for investment can therefore be measured in the same way as above.

"Market dominance" is another concept that may be measured in terms of people using the company's product, relative to the total number of people available in the market.

2.3.2 Primary Model Variables

The Business Data Science (BDS) model is intended to represent a sector of industry. At an abstract level, it should therefore represent multiple companies within the sector and sources of financing for them. Table 2 shows the primary variables that will be used to construct a model based on the lead author's previous experience and contacts. These variables are "primary" entities because they contain data that drives the model.

Table 2 · Primary Model Variables for the BDS model

Strata	*Entity*	*Description/Purpose*
Company	*Venture Capital Investment*	It has been demonstrated that the availability of venture capital finance, or desire of such companies to make investments, is an indicator of speculation in a sector (W. A. Sahlman & Stevenson, 1985). Related companies: Eurostat, Financial Databases.
Company	*Public Investment*	As above, but for investment in a publicly traded company, through a stock exchange. Related companies: Public investment records.
Company	*Valuation*	Calculated by number of shares multiplied by share price at the given time (i.e. market cap, but calculated wherever possible for non-publically-traded companies). Related companies: Public investment records.
Company	*Product Development*	As more people use a company's product, the company may develop the product if they have available resources. This would cause a greater tie-in for users, attract more users to the product and potentially hype the company.
Company	*Product Adoption*	Number of people using the product(s) of a company.
Sector	*Technological development*	Sector-wide progress in technological resource that individual companies may choose to employ. In the dot-com bubble, for instance, companies were racing to integrate the latest available technologies in their products (Wheale & Amin, 2003).
Sector	*Demand for Product*	Total number of consumers in the sector.
Sector	*Capital Available for Investment*	The amount that investors are willing to invest in a company in the model.

2.3.3 Derived Model Variables

Some items that would be desirable as variables of the Business Data Science (BDS) model are inappropriate to build into the model structure as primary variables. This can be due to the intangible nature of the item, or because it is a function of some of one or more other variables. Thus, a derived variable is one that is wholly dependent upon others; an indicator or output of the model.

Table 3 shows the proposed derived variables.

Table 3 · Derived Model Variables for Business Data Science

Strata	*Derived Entity*	*Description/Purpose*	*Function of*
Company	*Investment Volume*	Investment volume is measured with different variables for private and public companies. This variable will be a compound of both, to simplify the model architecture.	• Investment (Venture Capital) • Investment (Public)
Company	*Market Share*	Proportion of consumers using a company's product.	• [Company] Product Adoption • [Sector] Demand for Product
Company	*Perceived Risk*	The risk of investing in a company, as perceived by investors.	• [Company] Market Share • [Company] Valuation
Sector	*Market Share Fragmentation*	Indicator of whether the market is comprised of several similarly-performing competitors, or whether a dominant company has emerged.	• [All Companies] Market Share
Sector	*Perceived Sector Experience*	The degree to which investors think they understand the behaviour and business models of the sector they are working in.	• [Sector] Technological Development • [Company] Product Development • [All Companies] Valuation
Sector	*Speculation*	Momentum of investors involved in a sector they do not fully understand, or otherwise neglect/ignore due diligence.	• [Sector] Demand for Product • [Sector] Market Share Fragmentation • [All Companies] Perceived Risk • [Sector] Perceived Sector Experience

2.3.4 Model Design

In order to demonstrate the relationship between different stakeholders and tasks to achieve, the Business Data Science (BDS) model is designed based upon the specification and variables noted so far in this paper.

The BDS model shown in Figure 4 is split into three sections, representing the overall Market or Industry, Investors and Companies. Variables have been linked in the diagram using the following notation method:

- A line with a + indicates a direct (positive) relationship between variables.
  The destination variable will increase or decrease with the origin variable.

- A line with a – indicates a negative (inverse) relationship between variables.
  The destination variable is inversely affected by the origin variable.

- A line with "Fn." Indicates a more complex relationship between variables.
  The effect of the relationship may depend upon rates of change at the origin variable, for instance.

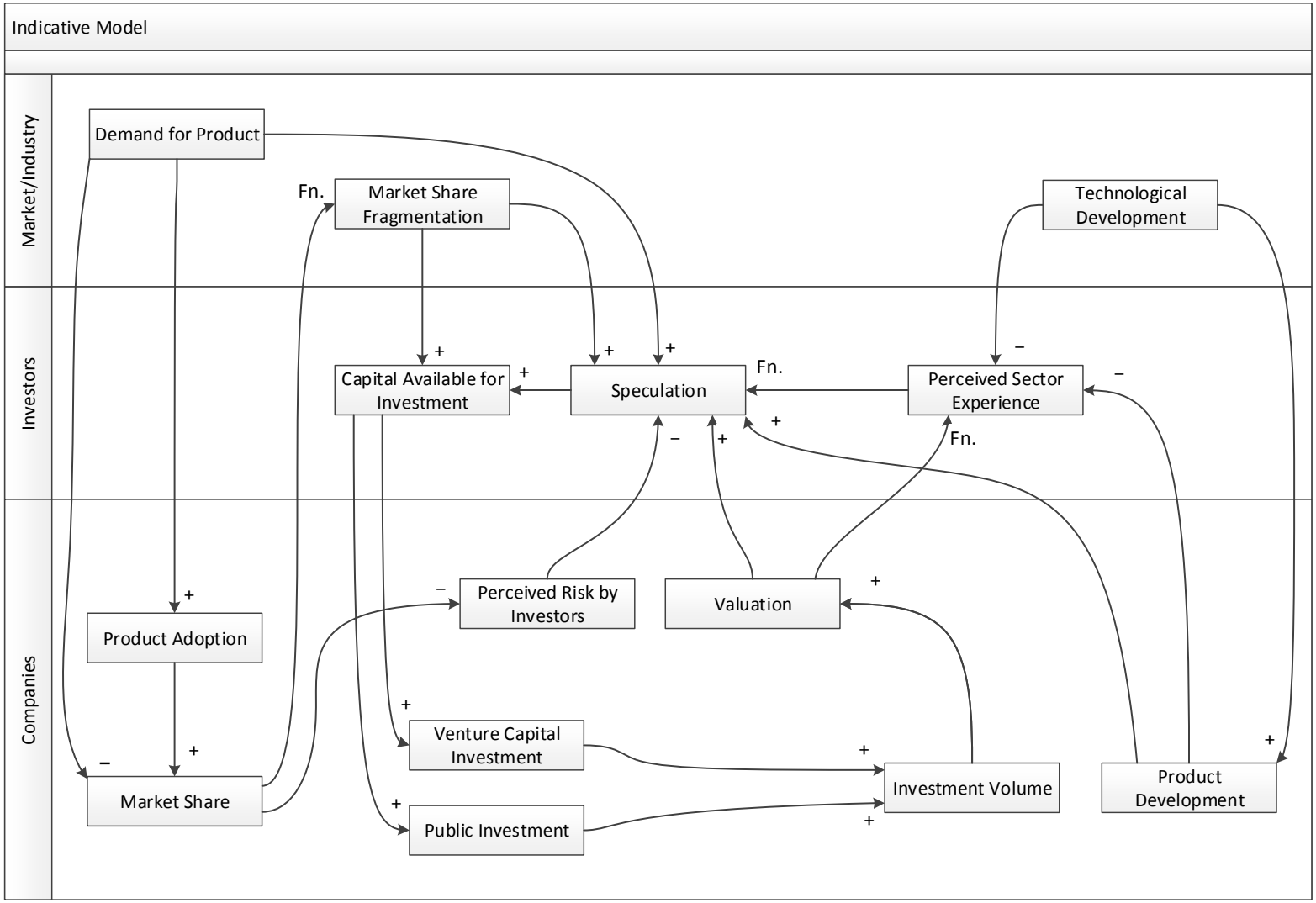

Figure 4 · Indicative Model Design for Business Data Science

3. Experimental Design for Business Data Science

The purpose of setting experimental design is to ensure all businesses can achieve resiliency in their strategies, operations and response to emergencies. Companies may use high performance stress test to validate their models (Chang, 2014). However, the approach we have recommended is to use a systematic way of verification without the need to undertake high performance stress tests. At this point, it is helpful to remind the overarching research question related to our previous work:

> *This research aims to quantitatively identify the degree to which companies involved with the web sector have exhibited a repeat of the bubble-like state that was observed during 1999-2001 (Chang et al., 2016 a).*

The BDS model shown in Figure 4 is a tool to aid development of this research. It presents an interpretation and combination of findings from the various literature that has been consulted for this work. Some elements are based on actual data, while others are conceptual and not quantifiable. This model is, intentionally, not computable and cannot answer the research question *directly*.

The literature review has revealed key metrics that may aid in the identification of a bubble state, and these are illustrated in the model.

3.1 Conceptual Design

This section aims to establish a potential approach to answer the research question, so that it can contribute to the development of the BDS model. It works at a high level, and an actual technique to implement our proposal will be presented in the next subsection.

The research question asks whether a repeat of the 1999-2001 bubble-like state has ever occurred. Thus, a benchmark of “the 1999-2001 bubble-like state” is required. Subsequent time periods may then be compared against this benchmark, to quantifiably report how similar a time period is to the benchmark.

Broadly speaking, the process is as follows:

1. Collect data.
   Actual metrics to be derived from model.
2. Run analysis to establish a quantified benchmark for the dot-com bubble period.
3. Run the same analysis on a subsequent period.
    a. Compare results for this period to the benchmark created in step 2. Any similarities/differences in the results may be indicative of the presence/absence of a bubble state.
4. Return to step 3 for additional time periods *(optional)*

This approach offers the potential to explore results for more than one time period after the bubble. By analysing two or more periods after the bubble, a convergence/divergence from the benchmark may be observed.

The metrics employed to do this, time periods analysed and results interpretation are subject to limitations and merits of the selected analytical method.

3.2 Real Design

Section 3.1 proposed a conceptual design for an experiment that would answer the research question. This section aims to identify a suitable technique to convert the pseudo design into a feasible experiment.

The conceptual design and model make some implications of the data that will be analysed:

1. One axis of data (probably rows) will represent periods of time.
2. One axis of data (probably columns) will represent metrics of the period.
   These will reflect the metrics shown in the model.
3. The relationship between metrics in each analysed period will be the primary output.

Factor Analysis can be particularly suitable to meet all these requirements (O'Rourke et al., 2013). Given rows of metrics, a Factor Analysis will try to form "components" around these metrics. Components represent groupings of metrics with common correlative properties within the analysed period. Metrics are given a value of "loading" to each component. As a method of dimension reduction, this analysis will reveal underlying commonalities between the metrics.

The Factor Analysis can be run in a binned fashion. That is, an initial analysis may be run for the dot-com bubble period, and subsequent analyses may be run for other time periods. The results from different time periods will be comparable. Changes in the structuring of components within each time period will enable detailed analysis of the results.

This approach for BDS implies the following schedule of tasks:

Gather data for analysis

This section presents an example about how to collect and analyse data for the BDS model. Data gathering and collections are based on the combination of the BDS model, the lead author's experience with industry and the primary/derived variables presented in Table 2 and Table 3, data for the following variables will be sought.

Table 4 · Input variables for Factor Analysis

Metric	*Source*
Venture Capital Investment (by phase if available)	Eurostat *or* Datastream *or* NVCA
Online Population (by geographic area if available)	Datastream *or* World Bank
Company Share Price	Datastream
Company Volume Traded	Datastream
Company Sales	Datastream
Company R&D Spend	Datastream

Company Market Capitalisation (i.e. valuation)	Datastream
Company Market Share	Datastream

The minimum reporting period sought for each variable in Table 4 shall be one year, but data of greater detail will be retained. The actual analysis periods will be the lowest common denominator of available reporting periods across all required data.

Wherever possible, venture capital and company data will be limited to activity within the technology/web sector.

Bin Data by Time Period

As explained earlier, a benchmark is required of the dot-com bubble period. This may then be tested against subsequent analyses for other time periods.

Four bins are defined as thematic ranges of years, so that each bin represents a different time period within 1995 – 2012.

Figure 5 shows a graph of the NASDAQ index for the relevant time period, and illustrates the proposed binning periods. Gaps of one year were left between each bin to ensure greater contrast between them. This is another demonstration for our BDS model, which can be adapted by businesses.

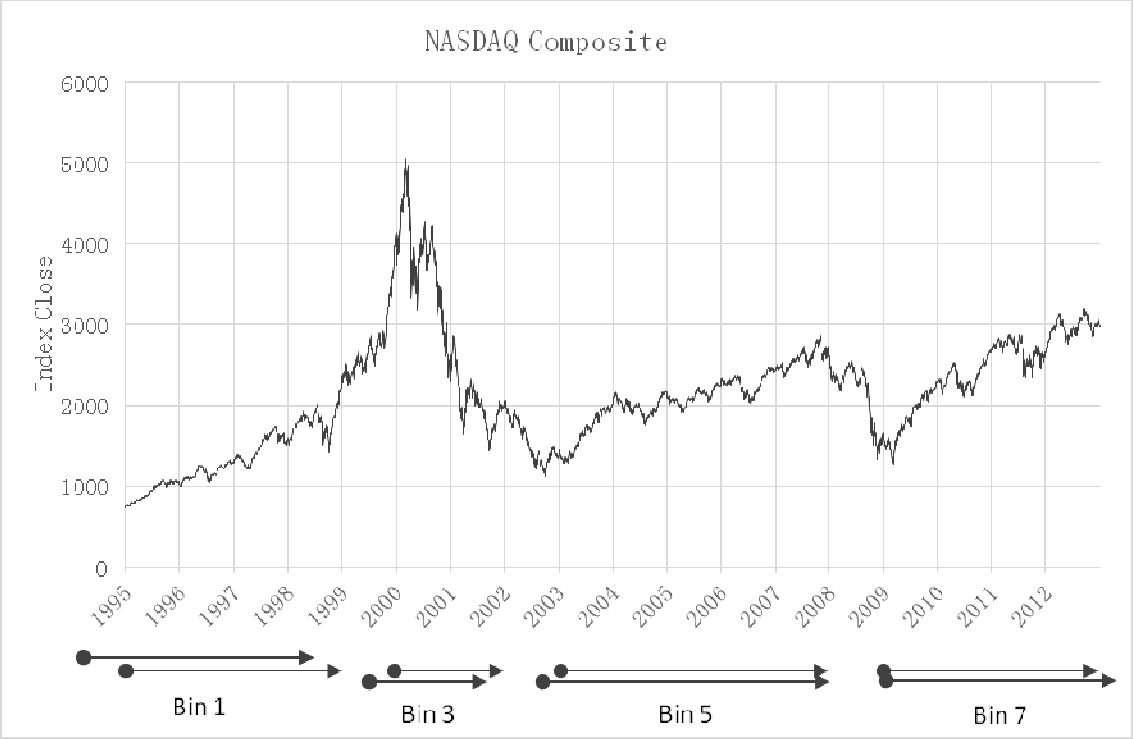

Figure 5 · Data bins shown against NASDAQ Composite index for relevant years

Detailed dates for the same bins are presented in Table 5.

Table 5 · Binning Periods

Bin	*Beginning*	*End*	*Description*	*Included*
1	01.01.1995	31.12.1998	Nominal activity prior to bubble	Yes
2	01.01.1999	31.12.1999		
3	01.01.2000	31.12.2001	Bubble Collapse	Yes
4	01.01.2002	31.12.2002		
5	01.01.2003	31.12.2007	Recovery	Yes
6	01.01.2008	31.12.2008	Subprime market crash	
7	01.01.2009	31.12.2012	Contemporary	Yes

If data is available as early as 1995, Bin 1 will present results prior to the bubble collapse. Two post-bubble bins are designed, as 5 and 7.

Run Factor Analyses

A factor analysis will be run once for each included bin. A Varimax rotation will be applied to the result. This adjusts the axes of the initial results to optimise loadings and simplify the final factor structure. It assists in emphasising relationships in the data, and outputs “factors”.

Factors are similar in appearance to components, and retain the same characteristics of metrics and loadings.

Compare Factor Analysis Results

The factors extracted from each bin will be compared side-by-side. Results for the bin during the bubble will be compared to those after it. Particular attention will be paid to the structuring of the factors in each bin, as this indicates how the underlying relationships between the metrics changes over time.

If a bubble state has occurred since the dot-com bubble burst, then similar factors with similar weightings will be observed during and after the bubble.

4. Discussion

Section 3 presents examples about how to collect and analyse data, how to perform simulations and experiments relevant to the businesses by the use of the proposed Business Data Science (BDS) model. Since more examples will be useful for its further development, this section presents four topics for discussion relevant for BDS future directions as follows.

4.1 Business Intelligence

Business intelligence plays an important role to allow business understand the performance of their business activities through a series of data collection, queries and analysis. There are different types of emphasis. Chen et al. (2012) explain that businesses can adopt business intelligence to enhance their businesses and deliver the positive impacts with the use of big data approaches. Chang (2013) presents his business intelligence model through the quantitative study of companies using SAP and investigation of their business performance during the economic recession. Chang (2014) then analysed the factors that

caused financial crisis in 2008-2009 period and used the Heston Model to analyse market volatility. The contributions included that firstly, the use of Heston Model can be simulated in the Cloud to allow thousands of computations can be achieved within seconds; secondly, the risk can be visualised to make the tracking and monitoring much easier; and finally, the predictive modelling can be used to forecast the movements of selected shares that can achieve about 95% accuracy. Ramachandran and Chang (2014) present their financial cloud solutions that provide accurate business intelligence services that can calculate the status of risk and return. Business intelligence can be applied to the BDS model to ensure that the status of risk and return can be analysed in real-time to provide the decision-makers vital information prior making decisions for buy or sell.

4.2 Security and privacy

Security and privacy always remain a priority and challenge since businesses constantly face off cyber crimes, hacking and unauthorised access. The UK Government (2015) has estimated £27 billion of loss due to cyber crimes and hacking in the UK. Businesses should invest more to improve the quality of security services and ensure all data can be protected against threats, hacking and unauthorised access. To ensure all the services can be resilient against attacks, a large scale penetration testing and ethical hacking should be undertaken to verify that services can be robust. Chang and Ramachandran (2016) present their large scale penetration testing in 10 petabytes of data centres and conclude that the time to recover from the impact of hacking may require a minimum of 125 hours. They use their multi-layered security to demonstrate. Chang et al. (2016 c) then improve their multi-layered security that includes integration of three different security solutions on top of NoSQL databases. Hacking on NoSQL database can be minimised without the attack from SQL injection. Chang et al. (2016 d) also conduct large scale surveys with 400 professionals returned their full feedback. They have concluded that privacy is the number one factor for businesses between Year 2016 and 2019. Their results also show that more than 50% of businesses are willing to spend £1 million and above for the following three years and invest for better services and infrastructures against cyber hacking. Hence, all the businesses should improve their level of security, privacy and trust to ensure that their BDS model can protect the clients' safety of their personal data and information about their investment.

4.3 Analytics

Analytics can help businesses to stay competitive since a lot of numerical inputs can be processed and presented the outputs as visual analytics, including charts, graphs, gadgets and reports. Technologies include the integration of artificial intelligence, machine learning, data warehouse, business intelligence, visualisation and web technologies (Provost and Fawcett, 2013 b, Ramanchandran and Chang, 2014; Chang, 2016). Chang (2016) also demonstrates the use of emerging analytics that can work in several disciplines such as healthcare, finance, education, natural science and security. Natural science simulations such as weather visualisation, sandstorm, tsunami and air pressures can be processed, visualised and presented within seconds and minutes. Anyone without the scientific training can understand the outputs with ease and allow them to plan their activities ahead. Analytics can be useful for businesses to process a large amount of data and blend with Business Data Science. The services include analysis of computational simulations and forecasting. Provost and Fawcett (2013 b) also assert that the use of big data analytics can help businesses to make better decisions based on the facts and key figures they have received.

4.4 Contributions to Data Science

Data Science includes the common 5Vs: volume, velocity, variety, veracity and value. Volume refers to the size and quantity of data involved and normally includes terabytes, petabytes and zetabytes of data in BDS. Velocity is the rate in which the data has been created, developed and stored for businesses. Variety includes different formats and types of data involved in business activities. Veracity is the extent of accuracy in analyzing the data and interpreting results, which can be instrumental and influential to business development. Value refers to the added value created by the adoption of BDS model that can result in creation of positive outputs and best practices for a large number of organisations. The use of BDS can contribute to the existing knowledge in Data Science. For example, disaster recovery based on Chang (2015 c) can be implemented to ensure that a large quantity of data can be processed and also be retrieved within one to two hours when the state of emergencies has happened to allow business continuity. Terabytes and petabytes of data can be processed, analysed and stored on daily basis smoothly with automation in place (Chang and Wills, 2016). Waller and Fawcett (2013) also explain that the full adoption of Data Science for businesses can create added value for supply chain management supported by the literature and their examples. Additionally, frameworks can also be fully integrated with BDS model to provide organizations a list of guidelines, best practices and recommendations to follow (Chang et al., 2013). Any errors and success made in the past can be presented as case studies, so that organizations that plan to adopt BDS can read about what to do and what not to do and have a better knowledge about how to reduce risk and how to enhance profitability, opportunities and benefits of adopting BDS.

5. Conclusion

In this paper, a conceptual model for Business Data Science (BDS) has been specified and illustrated. The model is designed and built using insights from the literature review. As a conceptual model, it's not intended to be computable. The aim is to help businesses to integrate their resources, processes and activities so that they can maximise their efficiency and performance, which are essential for business data science. As a tool for furthering this research, the presented model hypothesises ways in which data may be related. This also reinforces discoveries for better techniques for modelling and simulations. With improved techniques, businesses can obtain their analytics outputs faster with a better quality.

Following the model design, an experimental design identified several metrics that should be analysed, and selected Factor Analysis as the means of analysing them to answer the research question. At a high level, Factor Analysis will identify correlative relationships between key metrics before, during, and after the dot-com bubble. The comparison of outputs from these periods should provide an answer to the research question. All these issues are important to the development of BDS since a list of structured guidelines can be presented and lessons learned can be instrumental for businesses to stay up-front with BDS. The future directions for our BDS have been discussed in details as follows. Firstly, business intelligence can integrate with BDS to create greater positive impacts of analysing a large quantity of data and making outputs into visualisation and interactive analysis. Secondly, security and privacy for BDS should always be improved and maintained to ensure that all the data can be kept safe and all services can be resilient to different forms of attacks. Thirdly, business analytics can always present outputs in a way that the stakeholders can understand without technical backgrounds. Fourthly, BDS can further contribute to volume,

velocity, variety, veracity and value for businesses to improve efficiency, collaboration, business performance and opportunities.

The future work will include a detailed account of how the metrics/data can be gathered and prepared for detailed BDS analysis, so that organisations that have adopted our BDS model can find it useful for their day-to-day operations and business strategies.

References

Agresti, A., & Kateri, M. (2011). *Categorical data analysis* (pp. 206-208). Springer Berlin Heidelberg.

Åström, K. J., & Wittenmark, B. (2013). *Computer-controlled systems: theory and design*. Courier Corporation.

Banks, C. (2009). What is Modeling and Simulation? In J. Sokolowski & C. Banks (Eds.), Principles of Modeling and Simulation (pp. 3–24). Wiley Publishing.

Banks, J. (Ed.). (1998). Handbook of Simulation: Principles, Methodology, Advances, Applications and Practice. New York: Wiley. Retrieved from https://www.scribd.com/doc/53073919/Handbook-of-Simulation

Barlas, Y. (1994, July). Model validation in system dynamics. In *Proceedings of the 1994 International System Dynamics Conference* (pp. 1-10). Sterling, Scotland.

Barlas, Y. (1996). Formal aspects of model validity and validation in System Dynamics. System Dynamics Review, 12(3), 183–210.

Barlas, Y., & Kanar, K. (1999). A Dynamic Pattern-oriented Test for Model Validation. In Proceedings of 4th Systems Science European Congress

Borrego, M., & Newswander, L. K. (2010). Definitions of interdisciplinary research: Toward graduate-level interdisciplinary learning outcomes. *The Review of Higher Education*, *34*(1), 61-84.

Cellier, F. E., & Greifeneder, J. (2013). Continuous system modeling. Springer Science & Business Media.

Chang, V. (2013). Business integration as a service: computational risk analysis for small and medium enterprises adopting SAP. *International Journal of Next-Generation Computing*, *4*(3).

Chang, V., Walters, R. J., & Wills, G. (2013). The development that leads to the Cloud Computing Business Framework. *International Journal of Information Management*, *33*(3), 524-538.

Chang, V. (2014). The business intelligence as a service in the cloud. *Future Generation Computer Systems*, *37*, 512-534.

Chang, V. (2015 a). *A proposed cloud computing business framework*. Nova Science Publisher.

Chang, V. (2015 b). Towards a Big Data system disaster recovery in a Private Cloud. *Ad Hoc Networks*, *35*, 65-82.

Chang, V. (2016). An overview, examples and impacts offered by Emerging Services and Analytics in Cloud Computing. *International Journal of Information Management*, in press.

Chang, V., & Ramachandran, M. (2016). Towards achieving Data Security with the Cloud Computing Adoption Framework.

Chang, V., & Wills, G. (2016). A model to compare cloud and non-cloud storage of Big Data. *Future Generation Computer Systems*, *57*, 56-76.

Chang, V., Newman, R., Walters, R. J., & Wills, G. B. (2016 a). Review of economic bubbles. *International Journal of Information Management*, *36*(4), 497-506.

Chang, V., Walters, R. J., & Wills, G. B. (2016 b). Organisational sustainability modelling—An emerging service and analytics model for evaluating Cloud Computing adoption with two case studies. *International Journal of Information Management*, *36*(1), 167-179.

Chang, V., Kuo, Y. H., & Ramachandran, M. (2016 c). Cloud computing adoption framework: A security framework for business clouds. *Future Generation Computer Systems*, *57*, 24-41.

Chang, V., Ramachandran, M., Yao, Y., Kuo, Y. H., & Li, C. S. (2016 d). A resiliency framework for an enterprise cloud. *International Journal of Information Management*, *36*(1), 155-166.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

Coyle, G., & Exelby, D. (2000). The validation of commercial system dynamics models. System Dynamics Review, 16(1), 27–41. doi:10.1002/(SICI)1099-1727(200021)16:1

Davenport, T. H. (2013). *Process innovation: reengineering work through information technology*. Harvard Business Press.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

Ericsson, K. A. (2014). *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games*. Psychology Press.

Forrester, J. W. (1994). System dynamics, systems thinking, and soft OR.System Dynamics Review, 10(2‐3), 45-256.

Forrester, J. W. (1997). Industrial dynamics. *Journal of the Operational Research Society*, *48*(10), 1037-1041.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. *The management revolution. Harvard Bus Rev*, *90*(10), 61-67.

Meyers, R. A. (Ed.). (2010). *Complex systems in finance and econometrics*. Springer Science & Business Media.

Nath, P., Nachiappan, S., & Ramanathan, R. (2010). The impact of marketing capability, operations capability and diversification strategy on performance: A resource-based view. *Industrial Marketing Management*,*39*(2), 317-329.

Newman, R., Chang, V., Walters, R. J., & Wills, G. B. (2016 a). Web 2.0 – the past and the future. *International Journal of Information Management*, in press.

O'Rourke, N., Psych, R., & Hatcher, L. (2013). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Sas Institute.

Petty, M. D. (2009). Verification and validation. *Principles of modeling and simulation: A multidisciplinary approach*, 121-149.

Provost, F., & Fawcett, T. (2013 a). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Provost, F., & Fawcett, T. (2013 b). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51-59.

Ramachandran, M., & Chang, V. (2014). Financial software as a service–a paradigm for risk modelling and analytics. *International Journal of Organizational and Collective Intelligence*, *4*(3), 65-89.

Sahlman, W. A., & Stevenson, H. H. (1985). Capital market myopia. Journal of Business Venturing, 1(1), 7–30. doi:10.1016/0883-9026(85)90004-7.

Scheer, A. W., & Nüttgens, M. (2000). *ARIS architecture and reference models for business process management* (pp. 376-389). Springer Berlin Heidelberg.

Sokolowski, J. (2009). Simulation and Data Dependency. In J. Sokolowski & C. Banks (Eds.), Principles of Modeling and Simulation (pp. 91–119). New York: Wiley Publishing.

Sterman, J. (2001). System Dynamics Modeling: Tools for Learning in a Complex World. California Management Review, 43(4), 8–25.

The UK Government (2015), The Cost of Cyber Crime, Government Report, retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60943/the-cost-of-cyber-crime-full-report.pdf , accessed on 1 April, 2016.

US Department of Defence. (1996). Modeling and Simulation. U.S. Department of Defence. Retrieved from http://www.dtic.mil/whs/directives/corres/pdf/500061p.pdf, accessed on 1 April 2016.

Valliere, D., & Peterson, R. (2004). Inflating the Bubble: examining dot-com investor behaviour. Venture Capital, 6(1), 1–22.

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

Wheale, P., & Amin, L. (2003). Bursting the dot.com "Bubble': A Case Study in Investor Behaviour. Technology Analysis & Strategic Management, 15(1), 117–136.