A Generic Approach for Data Management combining Financial Modeling with Road Mapping

Eoin Joseph Brazil, PhD, MSc

Student ID: *10101500*

Atlantic University Alliance
National University of Ireland, Galway
Ireland

Supervisor: **Professor Conor Ryan**, University of Limerick

A research project submitted in partial fulfillment for
the award of Graduate Diploma in **Technology Commericalisation**
to the **National University of Ireland Galway**

Submitted to the National University of Ireland Galway, August, 2011.

Declaration

Author: Eoin Brazil
Title: A Generic Approach for Data Management combining Financial Modeling with Road Mapping

I hereby declare that this material is entirely my own work and does not contain material previously by any other author, except where due reference or acknowledgement has been made. Furthermore I declare that it has not been previously submitted for any other academic award or to any other university.

Signature:________________________________ **Date**:______________________

Abstract

This dissertation addresses the topic of data management from a financial model business perspective and from a road mapping perspective. The research questions in this thesis explore the subjective cost of data management as a means of calculating its costs as well as its benefits. It looks at combined these approaches with road mapping and other techniques to integrate data management within business and academic workflows. The dissertation provides a generic approach coupling financial modeling with road mapping focused at addressed the issues of data management.

Research Acknowledgements

Firstly, I would like to thank my supervisor Professor Conor Ryan for his role as advisor in my research and his guidance on this dissertation.

Secondly, I owe great and deep thanks to Dr. Jean-Christopher Desplat, my boss and collaborator on many projects.

Thirdly, I would like to thank all the members of the Irish Centre for High End Computing as I gained much from various conversations.

Acknowledgments

I could not have accomplished my research without the support, guidance, and inspiration of many people.

To my parents: Thanks for your help and support especially when I told you I was going back to university again!

To Michael, Paul,

To all on life's journey, the last verse by Alfred Lord Tennyson and his poem Ulysses passed along to me from my days in the University of Limerick.

> ... my purpose holds
> To sail beyond the sunset, and the baths
> Of all the western stars, until I die.
> It may be that the gulfs will wash us down;
> It may be we shall touch the Happy Isles,
> And see the great Achilles, whom we knew.
> Tho' much is taken, much abides; and tho'
> We are not now that strength which in old days
> Moved earth and heaven, that which we are, we are,–
> One equal temper of heroic hearts,
> Made weak by time and fate, but strong in will
> To strive, to seek, to find, and not to yield.

Contents

List of Figures

List of Tables

Part I

The Introduction

This part of the dissertation covers an introduction to the research and the relevant background material for the questions being posed.

Chapter 1

Introduction

Modern technologies are capturing an increasing volume of information due to the automation and technological-mediation of common business or organisational tasks. As a result, there has been growing interest in approaches for dealing with this growth of information. Data management is a broad term covering all the disciplines that are related to managing data and where the data itself is viewed as a valuable commodity. The problem of how to adopt the appropriate solution for data management for a particular context has lead to an increasing demand for suitable techniques and methods to answer these and related questions.

This dissertation provides a set of financial methods for use in data management and explores their use with road mapping to provide an approach for the adoption of data management within an organisation. The lack of an approach for effectively integrating data management within existing workflow is addressed in this dissertation from the viewpoint of a business financial model. The financial models presented in this dissertation can provide accurate information on the selections of data management options. The road mapping approach presented in this dissertation can be combined with the financial models to chart the appropriate technology strategy for a company or organisation seeking to utilise data management.

1.1 Dissertation Statement

Creating, evaluating and validating financial methods for use with data management is difficult, in particular the issues of estimations for specialised technology costs such as computation or storage and the calculation of time saved or benefits derived from the use of data management. There are a number of methods that could be applied to solving or minimising these issues. Road mapping is explored in conjunction with financial methods in this dissertation as a means of minimising these issues. The research questions in this dissertation will explore and defend these methods by providing answers of use to those wishing to integrate the most appropriate data management mechanisms for their particular workflow.

The work in this dissertation sought to answer the following research two questions:

RQ1 How can data management be integrated into organisation ?

RQ2 How can the benefits and costs associated with data management be calculated ?

These two research questions span a range of issues faced by those wishing to utilise data management. The methods and models presented through this work seeks to answer these questions using financial cost models and provide a road map through which these methods can be used to provide a path or map for the integration of data management.

1.2 Methodological Approach

This dissertation provides a set of financial models and methods for data management coupled to a roadmap for its adoption or for a technology strategy based on data management for a new product or service. An understanding of the financial methods is a prerequisite as is an understanding of the roadmap technique. The selection of methods used in this dissertation was based on a number of criteria including ease of use, prior similar use in the field or related field, ability to concisely present the results, and time required to use the method. The financial models and methods are outlined in chapter 2. The techniques and methods for road mapping are explored in chapter 3.

1.2.1 Overview of Scenarios

In this dissertation, the two research questions are addressed through the use of two scenarios. These studies provide results that open new avenues in the area of data management.

The first scenario looks at a small-to-medium size enterprise (SME) who wishes to move into a formal data management process within its workflow. The second scenario explores a large scale research project and its data management needs.

1.3 Contributions to the existing research in the field of data management

This dissertation has synthesised existing research on financial cost modeling and road mapping, it has applied this research to address the specific needs within the context of data management adoption of an organisation. The scenarios in this dissertation provide insights into the two research questions by outlining the approaches that can address these needs. These approaches detail a set of techniques that can assist in migrating to data management and effectively understanding its costs as well as its benefits.

The scenarios explored in this dissertation are not meant to be taken as strict implementations or guidelines but rather as a set of exemplars to provide assistance in understanding the factors and costs within data management.

1.4 Dissertation Contents

This dissertation is divided into six chapters. The first chapter is the introduction, which outlines the definitions, research questions and problems addressed by this work. The remainder of this dissertation is structured in the following way:

Chapter 2 "*Financial Methods*" - This chapter provides a background on financial modelling, it highlights a selection of methods available. The rationale and the previous work that informed to the development of methods focused on aspects relevant to data management is described. The chapter provides the necessary foundational understanding of these methods to ensure they can be used to address the cost aspects within data management adoption.

Chapter 3 "*Road Mapping*" - This chapter introduces road mapping methods and a selection of supporting techniques that are available for creating roadmaps. An example using computational chemistry is explored using these methods. The chapter shows how a roadmap can be used to provide the direction and set the priorities for a technology strategy, in this case within data management adoption.

Chapter 4 "*Empirical Analysis & Scenarios*" - This chapter introduces the two scenarios (SME and Research) and provides calculated examples of how the various financial methods can be applied to provide an empirical analysis for data management.

Chapter 5 "*A Generic Approach for Data Management combining Financial Modeling with Road Mapping*" - This chapter incorporates the methods and techniques presented into a generic approach that can be applied by organisations seeking to implement data management strategies. The chapter allows interested organisations to apply the methods presented in this dissertation to their own business problems.

Chapter 6 "*Conclusions*" - This chapter presents a summary of the work from the previous chapters and relates it back to the research questions, discussing to what extent these research questions have been answered. The limitations of this dissertation are discussed and future directions for work based on this dissertation are also outlined.

Chapter 2

Financial Methods

Financial analysis is a common means of valuing a project or new product. The methods in financial analysis are often supplemented by other techniques as financial data can be unreliable in the early stages of projects (Goffin and Mitchell, 2010). These evaluation method have their own strengths and weaknesses, and the choice of which to use depends on the level of uncertainty and the number of stages or interlocking components. It can be appropriate to use several financial methods to provide extra certainty (Cooper et al., 2001) as well scoring models such as the balanced scorecard (Kaplan and Norton, 1996), which are not based primarily on financial measures.

This chapter builds on these techniques introducing three specialised cost models relevant to data management. It further provides an example in computational chemistry to shown how financial analysis can help improvement and direct adoption of data management.

2.1 Net Present Value - NPV

A simple approach to determining value estimates the financial benefits and subtract the costs to get a net value. This approach can be improved by taking into account the time-value of money, which is the premise that income today is worth more than income in a year's time. These concepts combined provide the *Net Present Value* (NPV) approach for financial analysis.

2.1.1 Net Present Value as an Accounting Measure

Projects use NPV to reflect the present value of the expected cash flows. A positive NPV reflects a project that can increase the value of the organisation. A common investment strategy is to select projects that maximise NPV as this maximises the shareholder value of the organisation. Capital investment decisions or new project / product development selection often uses NPV as it ensures that the chosen option's return on capital will be more than the cost of the capital employed.

2.1.2 Measuring Performance using Net Present Value

NPV measures single projects or products in an organisation, however it should be used with other measures (i.e. *decision trees*) in multistage projects where dependencies can occur or where there are multiple decision points required. Decision trees and NPV are key elements in the Expected Commercial Value (Cooper, 2000) (ECV) discussed in section 2.3. The NPV measure gives a better depth of detail when combined with *sensitivity analysis*. This is achieved by re-running the NPV calculation a number of times using different assumptions within the major components such with different discount rates, product / service sales or with any other major cash factor. This provides a better understanding of the performance of NPV by highlighting any sensitivities to a specific factors, which can then be more closely monitored to ensure the product or project's success.

2.1.3 Valuing Methodologies for Net Present Value

NPV uses the concept where the value of money changes with relation to time, simply put income today is worth more than income next year because the money held today could be invested to generate a year's worth of interest. This concept is calculated using *Discounted Cash Flow (DCF)* (Brigham and Erhardt, 2002) method, as shown in equation 2.1.1, where both income and expenditure are discounted by a factor that takes account of when either occurs or by the weighted average cost of capital. An income or cost in one year's time is multiplied by a discount factor, where s is the yearly cost of money and n is the number of years: so if money costs 5 per cent a year the discount would be 1/(1.05) or 0.95. The income made in two years' time would be multiplied by a discount factor of $1/(1.05)^2$ or 0.91, and so on.

$$DCF = 1/(1+s)^n \tag{2.1.1}$$

A further enhancement of this approach is when the discount factors are included for both income and costs. This calculation is called the *Net Present Value* (Brigham and Erhardt, 2002), or NPV as shown in equation 2.1.2 where C are the cash flows (incomes or costs) for the ith time period while s and n remain the yearly cost of money and number of years, respectively. The discount factor, s, is best calculated as the average cost of the capital to the organisation.

$$NPV = C_i/(1+s)^n \tag{2.1.2}$$

The seminal work on modeling a lease or purchase decision from the view point of capital budgeting including taxation was by Johnson and Lewellen (1972) where NPV was presented as a strategy for investment between two potential cashflows, leasing or purchasing.

$$NPV = \sum_{T=1}^{Y} \frac{(P_T - L_T) - t(P_T - L_T - D_T)}{(1+k)^T} + \frac{S - t_g(S-B)}{(1+k)^T} - A \qquad (2.1.3)$$

The three terms in formula 2.1.3 are the net after-tax operating profit, the after-tax results following retirement of the asset, and the asset's initial purchase cost. P_T is the cash revenue estimation from usage of the asset in year T; L_T is the pretax operating cost for the asset in year T; D_T is the depreciation of the asset in year T; k is the cost of captal; A is the purchase price of the asset; S is the expected scrapage or re-sale value of the asset at the end of its life; B is the estimated book value of the asset at the end of useful life; t is the corporation taxation cost; and t_g is the taxation gain or loss for the capital gains tax on the disposal of the asset.

2.1.4 Limitations and Assumptions in Net Present Value

A limitation with this method is due to it being solely based on cash flows as it does not include historical accounting items which could provide a greater insight. A further assumption in this method is that of a constant discount rate over the life of the investment; this may not reflect the broader economic circumstances.

2.2 Internal Rate of Return - IRR

The Internal Rate of Return (IRR) is the rate where the Net Present Value equals zero and provides the expected rate of return under a specified set of conditions. It provides a measure for assessing the maximizing rate of return on an investment decision. The IRR of an investment is the particular discount rate where the NPV for the costs of the investment (outflows / negative cash flows) are equal to the NPV for the benefits of the investment (inflows / positive cash flows). The formula, shown previously in equation 2.1.3, applied for IRR as this is simply where the rate equals zero for the NPV calculation.

2.2.1 Internal Rate of Return as an Accounting Measure

IRR is a simpler measure compared to NPV and somewhat more popular as a result. It is often used in private equity or venture capital investment analysis as whilst these require multiple cash investments over a project's lifetime, there is only as single cash outflow at the end of the project (e.g. via the organisation's acquisition / merger or via an initial public offering).

2.2.2 Measuring Performance using Internal Rate of Return

This method does not consider the cost of capital (time based value of money) and should not be used to compare projects of different duration. IRR should also not be used to rate projects that are mutually exclusive but rather only should be used to determine the investment potential for a single project. In such situations, it could occur that there is a lower IRR (expected return), but a higher NPV (shareholder value increase) of the second project which would not be chosen, if it was limited to comparison of the projects using IRR.

2.2.3 Valuing Methodologies for Internal Rate of Return

It can be difficult to calculate the IRR as it requires a trial and error approach to find the discount rate that corresponds to a zero NPV and as such it is typically done using a computer or programmable calculator for simplicity. An example calculation is shown in Table 2.1, where 0.0547 or 5.47% is the required discount rate to return a NPV of 0.

Year	0	1	2	3	4	5	6	
Income	-1000	200	200	200	200	200	200	*NPV*
Discounted	-1000	190	180	170	162	153	145	*0*

Table 2.1: An example of the IRR calculation required to give a NPV of 0 by using a 0.0547 or 5.47% discount rate.

2.2.4 Limitations and Assumptions in Internal Rate of Return

IRR does not consider the cost of capital in its calculation but it may give a better measure of investment efficiency in capital limited investment situations. NPV is preferable over IRR as it gives a better reflection of value to the business. Another limitation with IRR is that it is impossible to provide a clear answer using this approach to projects with irregular cash flows.

2.3 Expected Commercial Value - ECV

NPV and IRR are based on the assumption that a project is an all-or-nothing proposal where there is a single decision point, options pricing theory recognises that there are many decision points in projects at incremental stages. The Expected Commercial Value (Cooper, 2000) (ECV) seeks to maximise the *expected* value, subject to certain budget constraints by recognising the incremental nature of projects with multiple decision points. ECV further addresses one specific weakness of the NPV method, which fails to consider risk as the probabilities of technical and commercial success are included in this equation.

2.3.1 Expected Commercial Value as a New Product Measure

ECV builds upon NPV by incorporating a number of additional parameters making ECV a more preferable method over NPV. ECV multiples the NPV by the probability of its commercial success, minus the cost of the commercialisation of the new product / project, and then multiplying this result by the probability for technical success, and subtracting the development cost of the new product / project.

2.3.2 Measuring Performance using Expected Commercial Value

In the case of a single stage project where there are no additional choices required over the course of the project, then ECV can be used directly. However, ECV can also be used in conjunction with *Decision Tree Analysis* (DTA) to analysis a multistage project where the ECV is calculated at each decision point to help provide even more realistic estimates. An example is shown in Figure 2.1 using a decision tree for a four stage project.

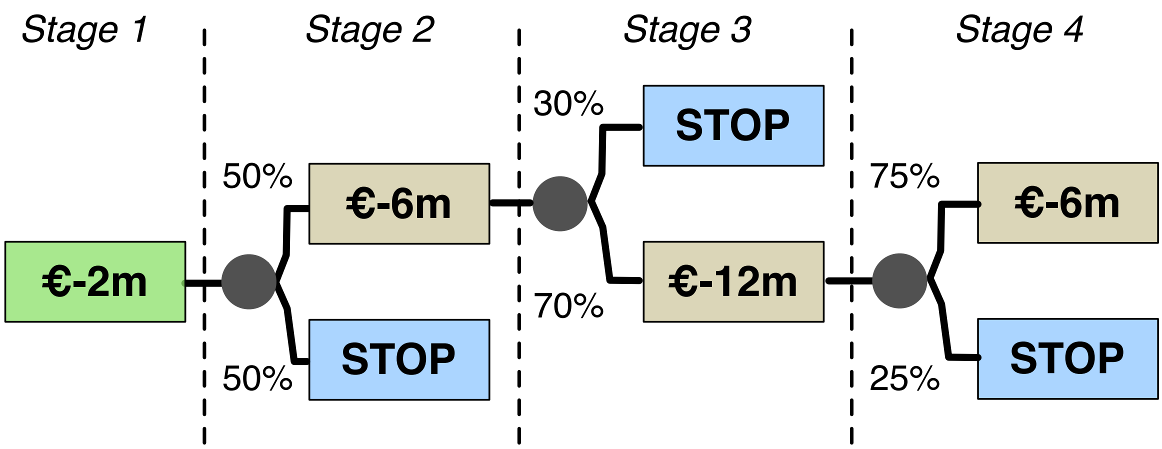

Fig. 2.1: An example of decision tree for a four stage project. Stage one costs €2 million and could only be stopped after this stage is completed, continuing to stage two costs a further €6 million. Stage three could be stopped at the beginning costing €8 million or continued to the next stage. Stage four could be stopped at the beginning costing €20 million or continued to end of the project costing €26 million.

2.3.3 Valuing Methodologies for Expected Commercial Value

The ECV method, shown in equation 2.3.1 on page 10, where *NPV* is the net present value, P_{CS} the probability of commercial success, P_{TS} the probability of technical success, C the commercialisation costs and D the developing costs. Using the example given in Figure 2.1, the approximate income with discounting would be €75 million with the total discount cost being €20 million and the overall confidence of success being 0.5 x 0.7 x 0.75 or 26%. On the assumption the project goes ahead the expected revenue of €75 million using 26% is approximately €19.5 million. Using NPV for this example would be result in a loss of 0.5 million and the project / product would be rejected if NPV was the sole measure used. The picture changes when using DTA and ECV where a stage gate approach facilitates better management of projects. Expanding this to concentrate on the third stage or gate in the example from Figure 2.1, the correct calculation would be €(2+ 0.5 x 6 + (0.5 x 0.7) x 12) million = €9.2 million. The example project would then have a projected of NPV of €(19.5 - 9.2) million = €10.3 million. The difference in this approach highlights its value as the example here shown how €10.8 million was added to the value of the project. This example project would end in one of four outcomes using this method. The first outcome being €55 million in profit if successfully completed, or losses of €2 million, €8 million, €20 million or €26 million depending in which intermediate stage it was stopped.

$$ECV = [(NPV \times P_{CS} - C) \times P_{TS}] - D \tag{2.3.1}$$

2.3.4 Limitations and Assumptions in Expected Commercial Value

A ECV model is useful, however it is limited to a single set of assumptions about the confidence levels in the decision tree. *Monte Carlo* simulations which run a large number of calculations with a random number generation seeding new (random) values into the various confidence levels within the decision tree can help create a more realistic ECV model by showing the full range of potential outcomes for the particular project. A problem with all of the previously mentioned financial techniques (NPV, IRR and ECV) are their financial focus which is problematic as this type of data can be incomplete at the early stages of a project.

2.4 Costing CPU Core Hours, Hard Disk Storage and Benefit-Cost Ratio

Three key elements in data management are computation, storage and the benefit-cost ratio; these can be considered using specialised financial models (Walker, 2009, 2010, Tichenor and Reuther, 2006) for each of these areas. The previous models discussed are generic financial valuation methods and many of the specialised models which will be discussed build upon their key concepts. The first model is concerned with using NPV to calculate the cost of a CPU core hour to assist in determining whether a lease or purchase option of a computer cluster is the better investment strategy for the computational element of the system. The second model is similarly concerned with using NPV to assist in determining whether to lease or purchase storage by considering the expected capital expenditure of hard drive purchase, replacement, and end-of-use salvage. This storage model further includes costs such as the projected operational expenditure for electrical consumption and the cost of staffing. The third model focuses on the utilization of a system to measure its value by capturing it return-on-investment (ROI) via a benefit-cost ratio (BCR) calculation. This model expresses the profit or cost savings divided by the sum of the investment over a given time period rather than trying to solely base measures on the usage of a cluster or larger computer system.

2.4.1 CPU Core Hour Cost Model

Walker (2009) builds upon the NPV of Johnson and Lewellen (1972) shown earlier in equation 2.1.3 on page 7 by adding two assumptions to simplify the formula and better situate it in the context of CPU hardware investment strategies. The first is the assumption that a cluster will have no salvage or resale value after its operational life due to the high depreciation of computing equipment. The second assumption is that the server cluster will have no expected cash revenue generation during its lifetime as whilst it facilitates operations or processes, it is rather a means to the business end rather the end itself.

2.4.1.1 Valuing Methodologies for CPU Core Hour Cost Model

There are three investment strategies proposed by Walker's (2009) model require complex formulas with a number of assumptions with the full detail covered in Appendix A.1.1. The strategies are the outright purchase of a cluster (see equation A.1.7 on page 58), the leasing of the necessary computational power (see equation A.1.8 on page 58), and the purchase with an annual upgrade cycle to avoid performance degradation (see equation A.1.10 on page 59).

2.4.1.2 Limitations and Assumptions in CPU Core Hour Cost Model

This model does not consider the price volatility in the online CPU market and assumes that there is no "lock-in" to a particular vendor who may require proprietary runtime environments or impose prohibition on user-developed software from running on other CPU leasing services. This model does not consider certain important peripheral issues in storage such as data uploading or downloading costs. It makes the assumption that in the distributed leased storage approach that latency is zero. The model is limited in that it uses the traditional NPV view of all or nothing with regards to investment and the assumptions about costs and CPU requirements may be incomplete at the early stages of a project or migration to a data management strategy.

2.4.2 Hard Disk Storage Cost Model

This model was designed to help users understand the premium being paid for through the use of online storage. The model allows for an understanding of the costs of these leased assets allowing for rational decision making with regards to whether this approach is an appropriate investment strategy.

2.4.2.1 Valuing Methodologies for Hard Disk Storage Cost Model

Walker (2010) builds upon his earlier CPU cost model (Walker, 2009) to explore the purchase or lease investment strategies for storage. It is focused on the SATA hard disk drive purchase, replacement, and end-of-life salvage as well as including utility / power costs and human operator costs. There are two investment strategies proposed by this model, the outright purchase of the disk storage (see Appendix A.1 and equation A.1.11 on page 59) and the leasing of the necessary storage space (see Appendix A.1 and equation A.1.12 on page 59).

2.4.2.2 Limitations and Assumptions in Hard Disk Storage Cost Model

This model does not consider certain important peripheral issues in storage such as data uploading or downloading costs. It makes the assumption that in the distributed leased storage approach that latency is zero. These points are acknowledged and further work has been proposed to modify the model but it has yet to be published. The model is limited in that it uses the traditional NPV view of all or nothing with regards to investment and the assumptions about costs and storage requirements may be incomplete at the early stages of a project or migration to a data management strategy.

2.4.3 Benefit-Cost Ratio Model

A major problem with large scale computation has been the traditional focus to measure the usage of the system, however whilst this keeps the system busy it does not capture the value of a high performance computer system in terms of benefits to the organisation. The benefit-cost ratio (BCR) was refined by Tichenor and Reuther (2006) to address this issue from an industrial / commercial perspective and as a mechanism to better capture the return-on-investment (ROI) by building on earlier work on the BCR by Kepner (2004).

2.4.3.1 Valuing Methodologies for Benefit-Cost Ratio Model

The BCR model expresses the profit or cost savings divided by the sum of the investment over a given time period. This measure better captures how these systems address the competitive edge of the organization as the previous usage model made large systems appear expensive and directed investment strategies away from hardware purchases or upgrades of these large computer systems. This model is based on a one-year time period IRR using the following formula: $BCR = 1 + IRR,\ or\ IRR = BCR - 1$. The simplest definition of the BCR ratio can be as similar to other measures such as the classic economic definition of productivity of utility divided by the cost.

$$productivity = \frac{utility}{cost}$$
$$benefit - cost\ ratio = \frac{benefit}{cost}$$

Kepner (2004) developed a productivity framework and evaluation model that influenced the BCR as it focused on productivity rather than on technical measures such as system demand or peak floating point operations per second (flops). Kepner's (2004) focused on research-oriented environments to develop equation 2.4.1, using the variables defined in Table 2.2; as a means to measure time saved in solving problems by accounting for the training of users, porting of applications to parallel systems, the launching of applications, the administration of the system and the cost of the system.

Model variable	Description
T_{SU}	Time saved by users on system
T_P	Time to parallelise for the system
T_T	Time required for training
T_L	Time to launch or begin the application using a batch resource management (job scheduler) system
T_A	Time required for system administration
C_S	Cost of the system

Table 2.2: The productivity model variables used by Kepner (Kepner, 2004) for benefit-cost ratio model in academic / research settings.

$$productivity(BCR) = \frac{T_{SU}}{T_P + T_T + T_L + T_A + C_S} \tag{2.4.1}$$

The BCR was adapted by (Tichenor and Reuther, 2006) to better reflect an industrial / commercial environment where the focus was on product design and development over basic research. Time saved by users is less of a concern in the business environment where factors such as the potential increase in market share or the importance of a job completion (i.e., how much revenue will be gained once the problem is solved or potential is there a contractual requirement to have job finished by a set time) are of greater priority. In Industrial / Commerical cases a number of changes are required as "*time to parallelize*" is often redundant as the software is typically purchased rather than developed and so in this model it is replaced by "*cost of software*". The second major change is that the launch time using resource management system (batch processing) aspect is removed as the importance of the software execution time is much greater in commercial environments.

$$productivity(BCR) = \frac{\sum P_G}{C_{AS} + C_T + C_A + C_S} \tag{2.4.2}$$

2.4.3.2 Limitations and Assumptions in Benefit-Cost Ratio Model

A BCR model broadens the focus beyond system utilisation by including factors such as training, staff costs and software costs. There are a wider range of factors that could potentially

Model variable	Description
P_G	Profit gained or maintained by the system
C_{AS}	Cost of the application software
C_T	Cost required for training
C_A	Cost required for system administration
C_S	Cost of the system

Table 2.3: The productivity model variables used by Tichenor (Tichenor and Reuther, 2006) for benefit-cost ratio model in business / commercial settings.

be added to include hosting and other miscellaneous overheads such as janitor's fees, etc. This model focused on using IRR and likewise the assumptions made by this model can be problematic as this type of data can be incomplete at the early stages of a project.

2.4.4 Computational Chemistry Example

Computational Chemistry is a wide area and covers the modeling of chemical and biochemical phenomena using computer simulations which implement the theory and principles from the field of chemistry. In this example, we will use the BCR definition for commercial productivity as shown in equation 2.4.2 on page 13. In commercial situations, we would advice that the models on computation and storage by Walker (2009, 2010) be used in conjunction with the BCR model. For the sake of simplicity, a time period of one year is assumed meaning that there is an investment at the beginning of the year and benefits by the end of the same year.

A chemical firm, ABC Chemicals Ltd., has four high priority projects that require a large cluster or HPC system to progress. The purchase of such a system will dramatically improve the probability of success for these four projects. The final IRR figure produced will be used by management to determine the cost against the benefits for the company. These projects, if successful, are expected to bring in profits of €3.2 million, €3 million, €2.2 million and €1.8 million. The costs of the system, software, training and administrator are:

- The cluster or large computer system will cost €2.5 million.
- The software licenses will cost €1.8 million.
- The training of 85 hours will be given to 100 users on the software and on the system at a rate of €100 per hour.
- The administration will require five system administrators to service both the software

and system at an individual annual salary cost (including contributions and additional employer costs) of €75,000.

$$productivity(BCR) = \frac{€3.2m + €3m + €2.2m + €1.8m}{€1.8 + (85 \times 100 \times €100) + (5 \times €75k) + €2.5m}$$

$$productivity(BCR) = \frac{€10.2m}{€5.525m} = 1.85$$

$$IRR_{1year} = BCR - 1 = 1.85 - 1 = 85\%$$

2.5 Alternative Methods to Overcome the Limitations of Financial Valuation Methods

A number of approaches to address solely financial techniques for the assessment of projects have been investigated, the Balanced Scorecard (Kaplan and Norton, 1996) is one such approach. This approach balances various dimensions such as internal and external foci, time orientation between past, present and future, current state and changes as well as both financial and non-financial measures. It takes four major perspectives on customers, financial operations, internal operations and learning & growth using SMART (specific, measurable, actionable, realistic, and target) principles. These measures balance an accounting focus by broadening it to include strategic thinking at an executive level. This type of approach is suitable for use with rolling forecast / short-term planning that are better placed to react to fast changing market conditions. The Balance Scorecard can improve performance by tying clearly defined company goals, the measurement of operational performances and the identification of areas for specific attention or additional resources into a programme that helps meet both the company's and employees goals. In the wider context of organisation acceptance and to help road mapping address fast changing markets such as those in high technology, organisations should not simply tie a data management strategy to solely financial methods but rather to broader approaches such as the Balance Scorecard.

2.6 Applying Financial Valuation Methods to Data Management

The techniques and methods in this chapter span a wide range of financial valuation methods to highlight the various concerns from computation to storage to productivity in addition to the time based value of money that management accounting uses to value or gauge the key components in data management. The models with their respective advantages and disadvantages are highlighted in Table 2.4. The key point is each model has certain advantages and disadvantages but the best approach for a data management strategy involves a combination of the specialized models with a model such as ECV to best consider the potential outcomes for the strategy.

Method	**Advantages**	**Disadvantages**
NPV	Takes account of the time-related value of money.	All or nothing view of investment, Incomplete data at early stages, Based only on cash flows not historic information, Assumes a constant discount rate for duration.
IRR	Provides measure for assessing the maximizing rate of return on an investment decision.	All or nothing view of investment, Incomplete data at early stages, Does not consider the cost of capital in its calculation, Unable to address projects with irregular cash flows.
ECV	Considers risk of both technical and commercial success in a project, Multiple decision points in projects, *Monte Carlo* simulations can be used to explore all possible decision trees.	Incomplete data at early stages, Limited to a single set of assumptions about the confidence levels in the decision tree.
CPU Core Hour	Tailors NPV to consider CPU costs and factors relevant to this type of investment.	Does not consider the cost of data uploading or downloading. Assumes no ”lock-in”. Many of the points against NPV also apply.
Hard Disk Storage	Tailors NPV to consider hard drive storage and factors relevant to this type of investment.	Does not consider the cost of data uploading or downloading. Assumes zero latency. Many of the points against NPV also apply.
BCR	Considers factors beyond system utilisation.	Incomplete data at early stages, Does not consider the cost of capital in its calculation.

Table 2.4: A summary of the advantages and disadvantages of the models presented in this chapter for application to data management.

2.7 Conclusions

This chapter presented an introduction to financial valuation methods as a method to provide a means to calculate the value for data management. Three specialised cost models focused on CPU, storage, and productivity were introduced building upon these financial valuation methods. An example of how these models were used for a simulation example in computational chemistry was explored.

The next chapter concentrates on introducing road mapping as a mean to plan for the future in a manner that promotes development of integrated technology and business strategies. It highlights the various types of road maps, the rationale behind them, a range of supplementary techniques to assist road mapping and how road mapping can help address problems in data management.

Chapter 3

Road Mapping

Roadmaps contain forecasts of possible events as well as outlining plans for action. They are best suited to dictating directions with regard to predictable rather than disruptive change and are typically subjective in balancing the events outlining those most likely and advantageous. The former chairman of Motorola, Rober Galvin defined science and technology roadmaps as:

> "*A "roadmap" is an extended look at the future of a chosen field of inquiry composed from the collective knowledge and imagination of the brightest drivers of change in that field. Roadmaps communicate visions, attract resources from business and government, stimulate investigations, and monitor progress. They become the inventory of possibilities for a particular field. In engineering, the road mapping process has so positively influenced public and industry officials that their questioning of support for fundamental technology support is muted.*"
> (Galvin, 1998)

3.1 What is Road Mapping ?

Road mapping is a process that develops technology strategy and assists in integrating business and technology. It displays the interactions between products and technologies over time by considering the short- and long-term aspects of both. A good understanding of both applications and markets is required to define the products in terms of customer requirements. The time interval of the roadmap will depend on the specific products and technologies as well as the aggregation level. Roadmaps contain elements that are focused on operational aspects, however the primary goal is to serve as a tool for business strategy planning. A technology roadmap looks forward from the present and uses a range of business experts opinions to envisage a future view with the aim of answering specific questions. A roadmap can be either technology or needs driven or can be a hybrid of both.

There are a range of types of roadmaps that fit within the technology roadmap category, eight have been identified by Phaal et al. (Phaal et al., 2004). Roadmap types include product planning, service/capability planning, strategic planning, long-range planning, knowledge asset planning, program planning, process planning and integration planning. A roadmap can take a range of graphical formats which facilitate the technique being adaptable to fit a broad set of problems. The technology road mapping technique can be seen as a needs-driven planning process to identify, select and help develop technology alternatives that can address a set of product or process needs. Phaal et al. (Phaal et al., 2004) described a generic roadmap structure as shown in Figure 3.1.

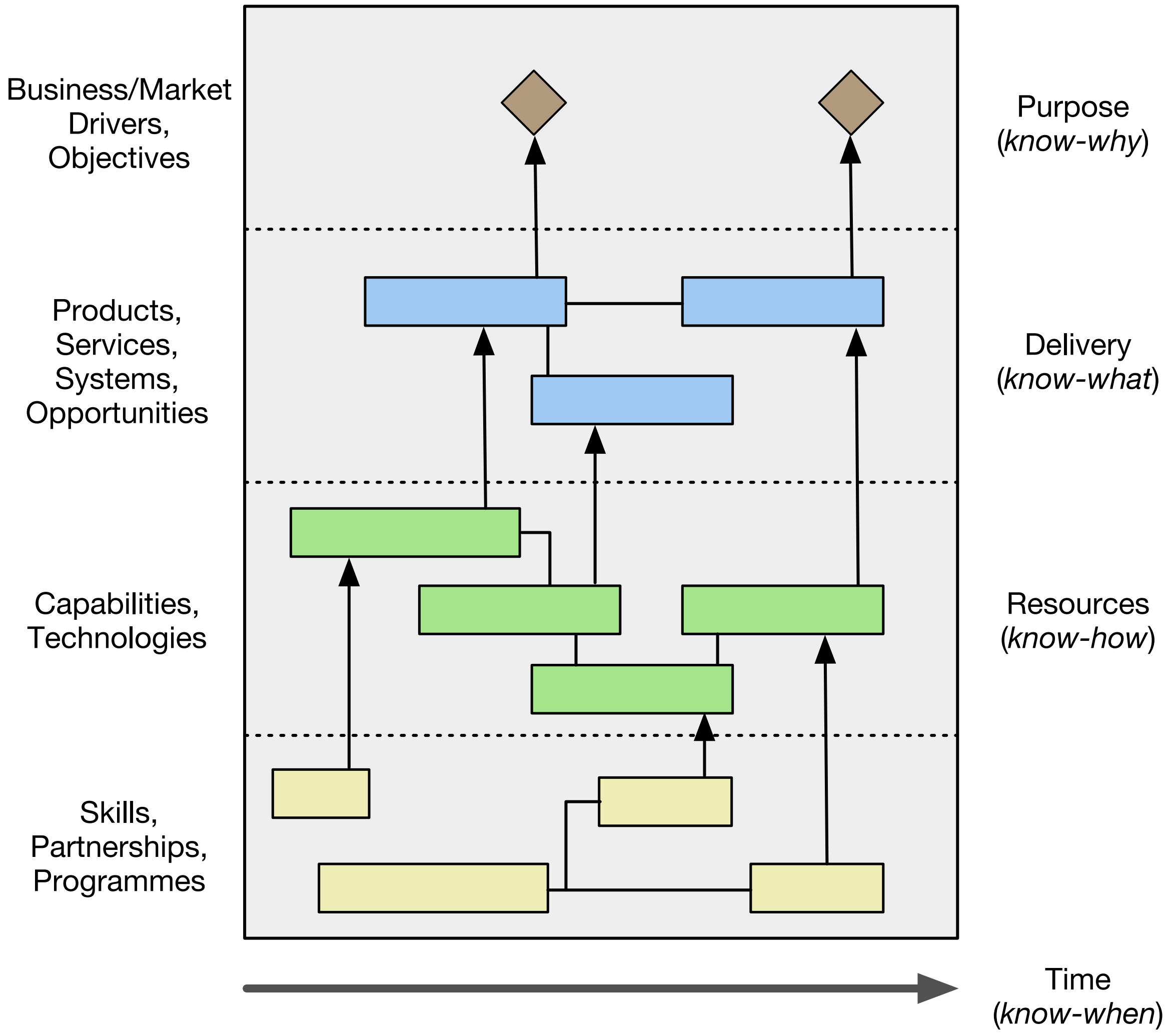

Fig. 3.1: The generic roadmap structure (Phaal et al., 2004).

A roadmap is in many ways like a Gantt chart at a higher level of abstraction, where the emphasis is on the logical structure and interdependencies over its total completeness. Phaal et al. (2003) defined a process of a generating a roadmap quickly using four stages as shown in Figure 3.2. The first three stages identify the key market drivers, product features and technologies with the fourth stage using the information gathered in the first three stages to create the first roadmap. The first stage selects a prioritised set of business and market drivers using a weighting from 0 to 10. The second stage considers the features or benefits of the product / service that drive competitive advantage. A *cross-impact matrix* is used to check how well these support the key market and business drivers, an example is shown in Figure 3.3 and in the larger context of road layers is shown in Figure 3.4. This is similar to the Quality Function Deployment (see section 3.1.1.1) approach where each feature is given a score according to the strength of its influence on each driver multiplied by a weighting factor to give an overall score. The overall score displays how well the features satisfy the set of drivers ensuring that the most appropriate features are chosen to address the drivers.

The third stage performs a similar analysis that links the product / service features (these have weights derived in the first stage) to the competences or technologies that the company has available. This provides a linkage between the necessary competencies and how they contribute to the performance of the products / services. In the fourth stage, the information from the first three stage is used to create the roadmap and it also needs to incorporate the key market events, their timings, the number of product / service lines and the frequency of new product / services, etc. This stage requires involvement from management in the organisation to link current performance and market trends.

The architecture or structure of a roadmap should be aligned to match the scope and focus of the issue being addressed as it will then allow a common language to support the desired innovation, strategy or policy. The two key dimensions of a roadmap, shown in Figures 3.1, 3.2 and 3.4, are time and layers (with optional sub-layers). The granularity allows a group of roadmaps to be developed that can span the different levels required.

3.1.1 Supporting Tools

Road mapping has a range of techniques that have been developed to help in creating and eliciting the information necessary to build the map. Two of the most widely used of these techniques are Quality Function Deployment and the Innovation Matrix.

3.1.1.1 Quality Function Deployment

Quality Function Deployment (QFD) (Goffin and Mitchell, 2010) is a technique that helps ensure the customer's needs and desires are what drives the choices made for a product's design through the concept and idea phase of the product creation process. It brings a focus to

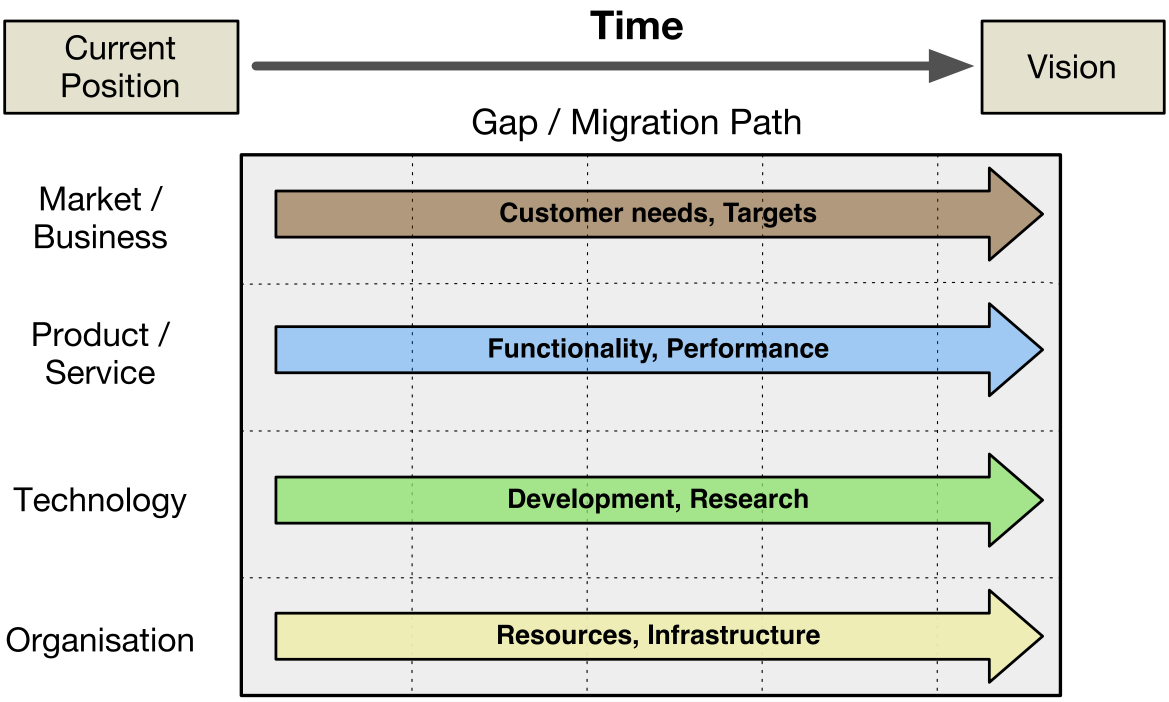

Fig. 3.2: The generic roadmap service / product planning structure (Phaal et al., 2003), identifying the products / services required for meeting strategic objectives, and the competences needed to achieve them.

Driver Weight	3	4	10	2	4	8	6
	Market Driver 1	Market Driver 2	Market Driver 3	Market Driver 4	Market Driver 5	Market Driver 6	Market Driver 7
Feature 1		**	*				
Feature 2			**	*	*		
Feature 3				**			
Feature 4	*					*	*
Feature 5				*	*		
Feature 6			**			*	*
Feature 7	*			**			
	6	**8**	**50**	**8**	**8**	**16**	**12**

Fig. 3.3: An example of a cross-impact matrix (Phaal et al., 2003) for the first stage relating market drivers to product / service features. Each star represents a weight of one and are used with the driver weight to ensure higher priority features are focused on. In Market Driver '3' its weight of 10 is multiplied by 6 (1 star + 2 stars + 1 star + 2 stars) to give a final result of 50. A feature without any star is where there is no priority or relation from the perspective of the business between this feature and the specific driver.

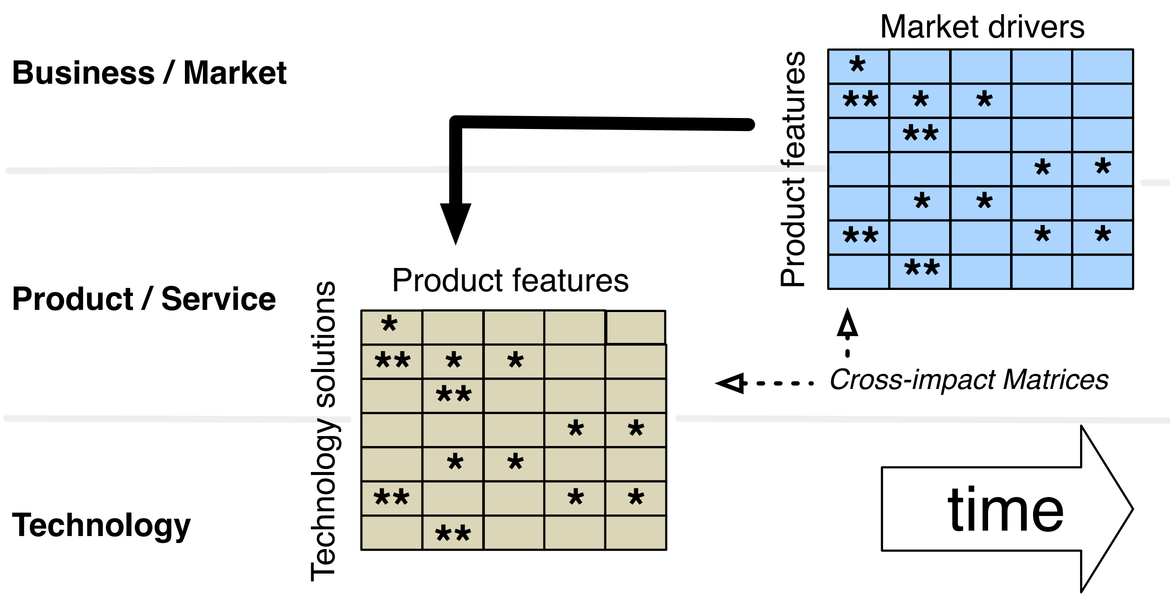

Fig. 3.4: An example of how the linked grid analysis can be used to understand the relationships between roadmap layers and sub-layers.

the customer requirements, the functionality or the what; these are then codified into technical product characteristics, the mechanisms or the how. QFD is a simpler process when compared to road mapping as the specific process technologies and time dependencies are less important as it can be seen as an element within the roadmap planning phase with the goal of determining the technologies in the area of interest. It feeds into roadmap scenarios showing how product functionality may evolve over time and which technologies will be required or will potentially influence product functionality. The results assembled using this technique are divided into five to eight sections or *rooms* to create a *House of Quality* (Clausing, 1994) as shown in Figure 3.5.

The first section in Figure 3.5 is concerned with the customer needs or requirements, these are weighted as they will all not be of the same importance. A Kano analysis (Matzler and Hinterhuber, 1998) of these features breaking them into basic, performance and excitement categories can assist in determining the weighted to attached to these needs. A basic category contains the features that a product or service would simply be unacceptable, e.g. a car must start readily or a shampoo should not cause immediate hair loss. A performance category lists features that provide a real benefit to the customer and each improvement brings greater satisfaction, e.g. improved car fuel economy, improved ease of use, or a lower cost. These features are the long-term focus for competition within the specific market. A excitement category contains features that respond to hidden needs or desires and would not often be

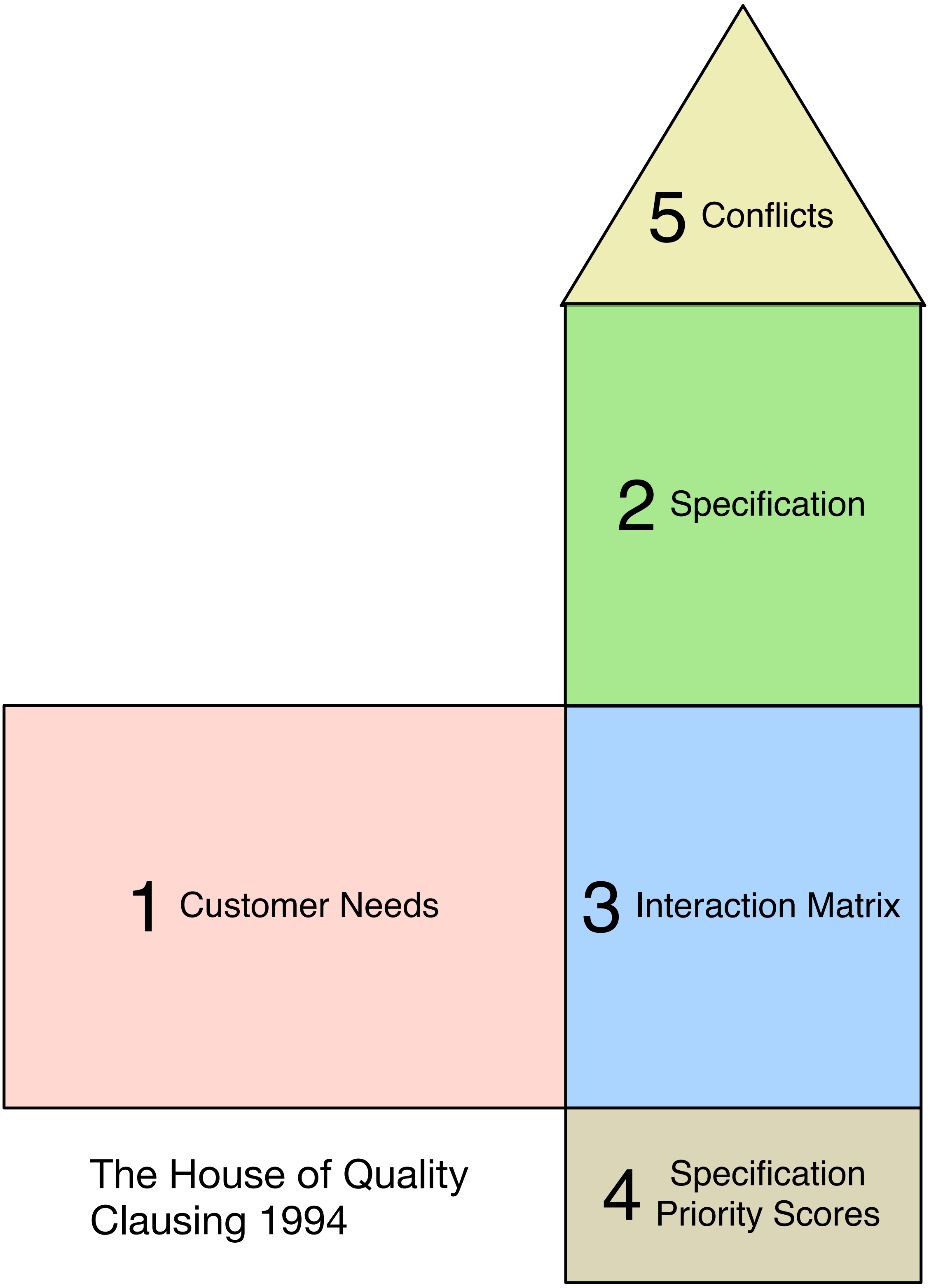

Fig. 3.5: The House Of Quality model in QFD (Clausing, 1994).

deliberately included in the product or service. Examples of these excitement or delighter features can be seen with the introduction of the first TV remote or mobile phones with touch screen interfaces. The second section in Figure 3.5 are the main elements or aspects of the product/process being provided by the organisation. It is a technical specification from the view of the provider of the item/service and takes into considers factors that such an organisation might need to provide. These factors combine with the needs from customer requirements analysis to create an interaction matrix which determines the impact between the customer and organisational factors using three levels of impact (designated with one, two or three stars in the matrix being weighted as 1, 3 and 9 respectively). The interactional matrix shows which of the features contribute the most to satisfying the requirements of the customers. The weighting given by the customer are multiplied by these organisational factors using the impact levels in the interaction matrix to give the specification priority scores for the various features, shown in the fourth section of Figure 3.5. The fifth and final section in Figure 3.5 deals with conflicts arising between the various features. These conflicts are those that cannot be resolved and indicate where one feature or benefit must be traded off against one another.

An example is shown in Figure 3.6 using this technique for a data management scenario. The particular areas of impact shown in this figure highlight automation, integration with the existing workflow, backup and the process complexity within the software infrastructure. A key advantages of this technique are the compact form of presentation and the ability to highlight that the detailed features of the product or process are actually aligned to the needs / requirements of the customers. The QFD technique focuses on customer needs or functional requirements (the what) and translates these into the appropriate technical characteristics of the product or services (the how). The disadvantage of this technique is that it does not have any concept of time or time based dependencies. Time is a key element in road mapping, however the QFD provides a good foundation for beginning a roadmap and can be used as part of the roadmap planning phase. This planning phase helps determine the necessary technologies and then projects forward over at least a 4-6 year period considering how these technologies will develop, what technologies will be required and how technology might influence the product or service functionality.

3.1.1.2 Innovation Matrix

The Blue Box model (Groenveld, 2007) is a three stage model developed by Philips company addresses the time aspect and is shown in Figure 3.7. It highlights the relationship between uncertainty of product feasibility and the demands of resource allocation. The phases are indicated by the two circles and the square. The first phase (upper circle) asks where it is physically possible, the second phase (the square) asks is it something attractive to the

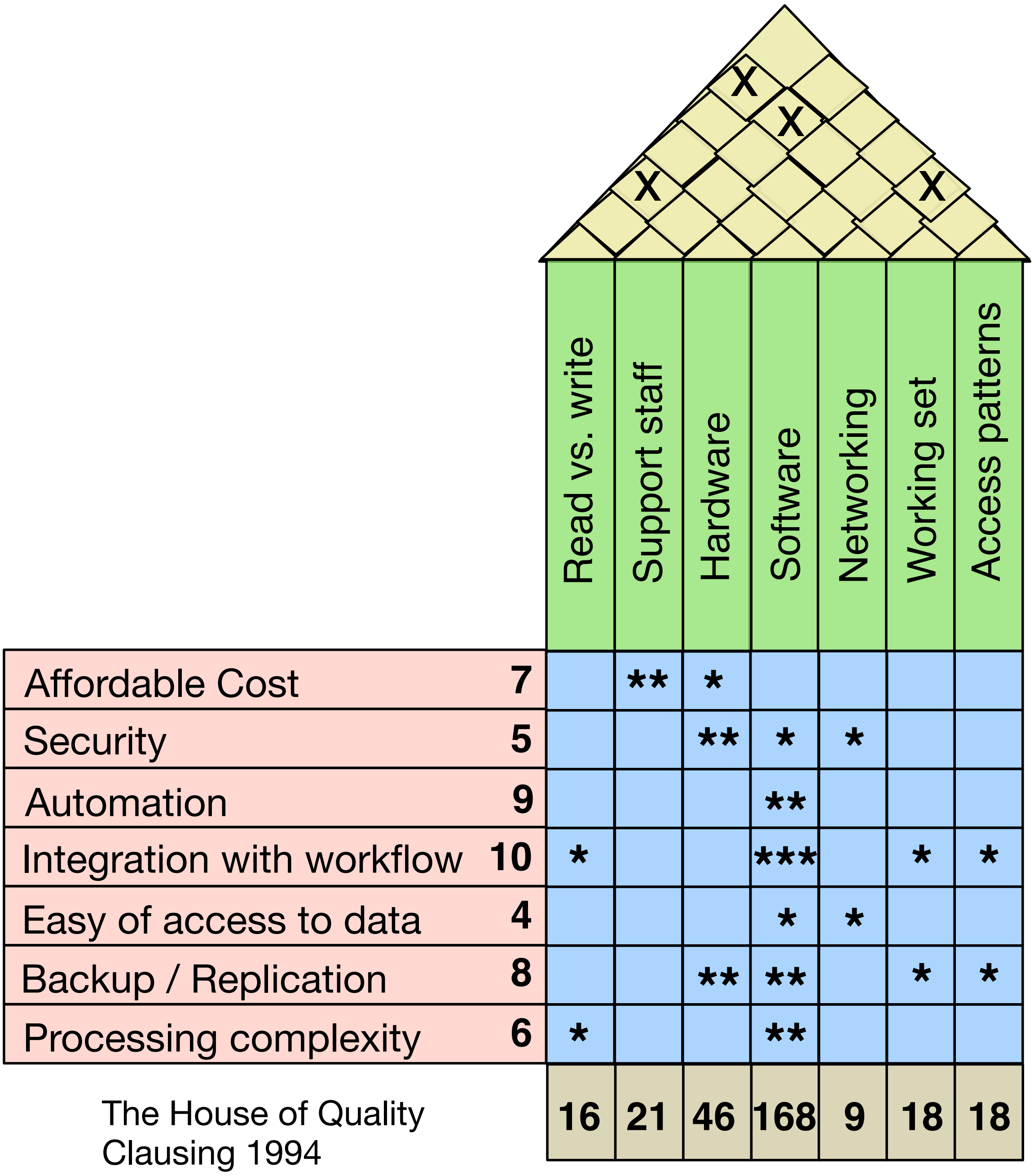

Fig. 3.6: An example of applying 'The House Of Quality' to the needs and desires for a hypothetical Data Management scenario (Clausing, 1994).

company where it addresses product portfolio issues and the third phase (lower circle) asks how it can be achieved by addressing product development activities.

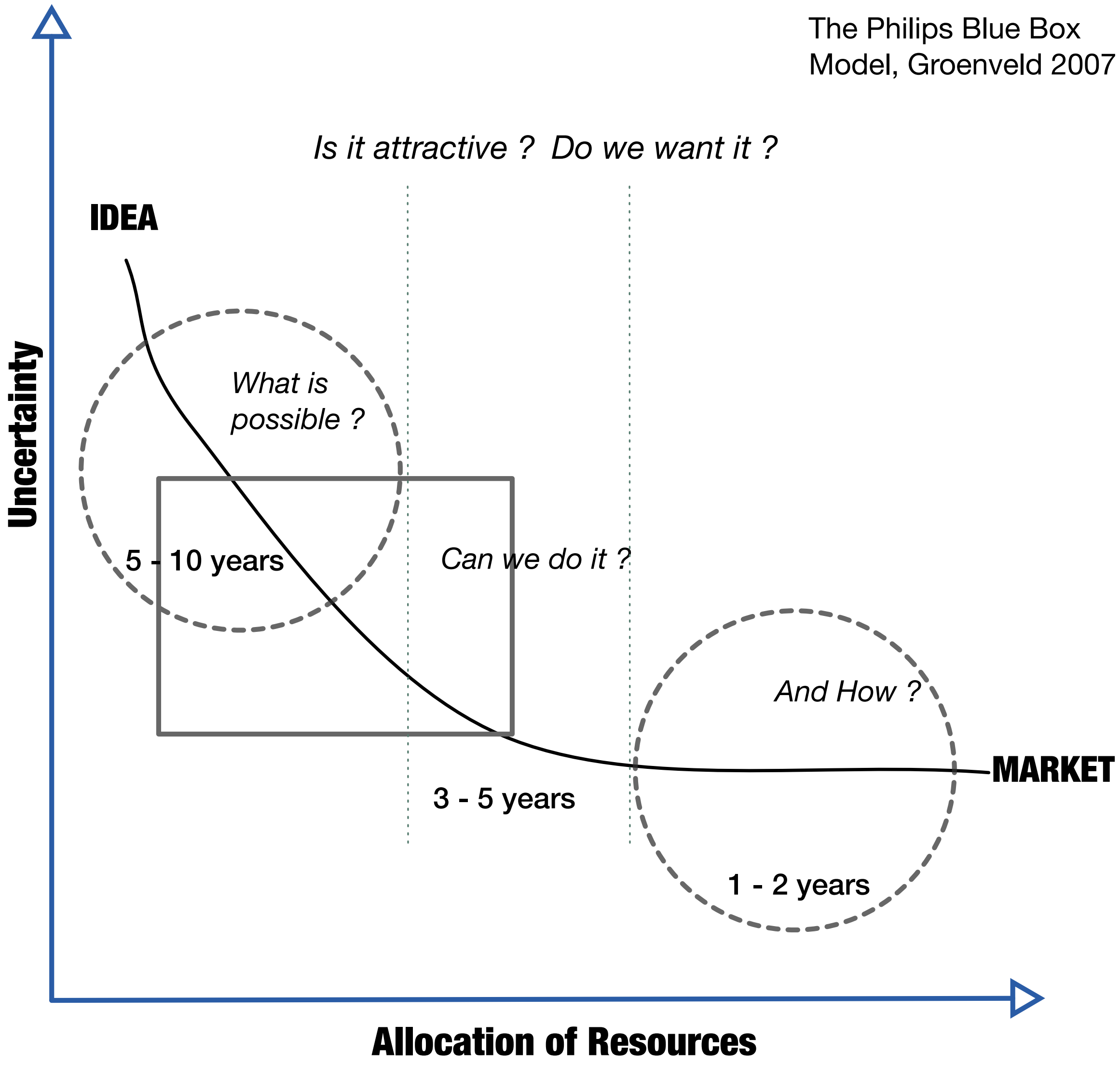

Fig. 3.7: The Blue Box model (Groenveld, 2007) highlights the uncertainty of product feasibility and resource allocation demands using three different time stages to map an idea to a market.

The Innovation Matrix (Groenveld, 2007) shown in Figure 3.8 and uses the inputs from the Blue Box model to support road mapping. It highlights the uncertainty on the vertical axis and the availability of the technologies or products on the horizontal axis. The model used similar time frames to the Blue Box model considering 5 - 10 years, 3 - 5 years and 1 - 2 years periods. Products and technologies are placed into nine sectors of the technology/product matrices based on the parameters of ‘availability’ and ‘uncertainty’. Sectors include categories such as missed opportunities, the problem (where there is a short term requirement but where the feasibility of the technology is unproved), your future is the sector where the best balance between required availability and degree of uncertainty for the product/technology can be managed.

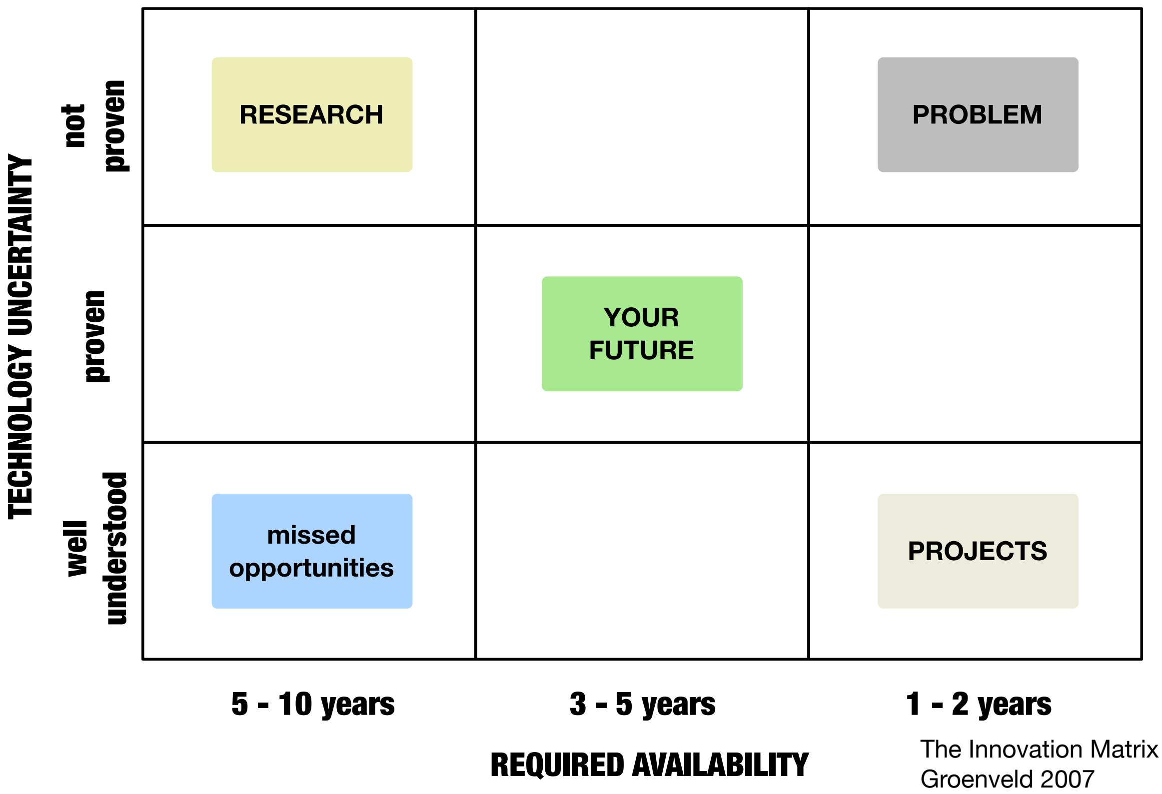

Fig. 3.8: The Innovation Matrix model (Groenveld, 2007).

The results of these three techniques provide the necessary details to take further actions to investigate specific problem areas or to being the road mapping process. A seven stage process developed at Philips (Groenveld, 2007) can be used to create a three layer roadmap of markets/applications, of products/services and of technologies and consists of:

1. Problem recognition by management.
2. Development of the three layer roadmap by specialists.
3. Roadmap discussion and information-gathering by small team.
4. Broader workshops to incorporate multidiscipline participation in the roadmap generation.

5. Modify the roadmap to reflect any changes from these discussions and workshops.
6. Updating / improvement of the supporting tools (QFD, Blue Box and Innovation Matrix).
7. Repeat stages 2-6 to stimulate learning.

3.1.2 Limitations and Assumptions in Road Mapping

Road mapping is a useful technique with regards to planning for markets/applications, products/services and technologies; however, it is best suited to the incremental developments of existing aspects. Disruptive technologies (Christenson, 1997) or new kinds of breakdowns (Spinosa et al., 1997) are ill-suited to road mapping as they often cannot be accurately forecasted. Time can be an issue for fast moving organsiations such as customer facing businesses as the process of gathering and analysing the data required to create a roadmap. Roadmaps can suffer from a limited field of view where only serious consideration is mainly focused on the next product or service generation.

3.2 Supplementary Methods to Accompany Road Mapping

Road mapping is one approach to developing technology strategy but there are a wide variety of other approaches. A sample of the diversity can be seen by considering a small subset including scenario planning (Cornelius et al., 2005, Wack, 1985), Delphi (Rowe and Wright, 1999), information markets (Hahn and Tetlock, 2006), TRIZ (Altshuller, 1996) and service blueprinting (Bitner et al., 2008).

Scenarios (Cornelius et al., 2005) have been used in business to provide coherent but credible alternative narratives about the future. Scenarios have been used for strategic planning to challenge their assumptions, to develop alternative strategies, and to test both their current plans and alternative plans within the context of alternative or hypothetical futures. Shell has used scenarios for over thirty years to help plan and anticipate shifts in the world energy market. Pierre Wack, an early advocate of scenarios for planning at Shell found scenarios (Wack, 1985) to be a better tool than forecasting which often failed "*in anticipating major shifts in the business environment that make whole strategies obsolete*".

The Delphi (Rowe and Wright, 1999) method uses a coordinator to iteratively send a set of questions to a panel of anonymous experts, after each round the coordinator circulates the results and gives the experts a chance to reconsider their replies in light of the opinions of the other experts. Answers far from the general responses are allowed to justify their opinions as these might be due to lack of knowledge or more importantly due to specialist knowledge. The next round and following rounds follow this process moving towards a consensus.

Information markets (Hahn and Tetlock, 2006) or prediction markets use the idea of virtual trading of idea stocks to represent new product ideas in virtual marketplaces with participants suggesting and trading shares with the resulting stock prices providing indicators of the idea's potential success. These idea markets source, filter and evaluate new product ideas or technologies.

TRIZ (Teoriya Resheniya Izobretatelskikh Zadatch) (Altshuller, 1996) is a creative problem solving approach based on the notion of a database of ideas, it was originally developed using patent databases. Problems and solutions are grouped into generic patterns, providing a means to search all the generic ways a specific problem could be solved. Design trends can be identified using TRIZ and opportunities for improvement can be determined. This approach offers insights into how design tradeoffs are managed by showing how previous products managed to overcome particular design problems. The key concept in this approach is to avoid reinventing the wheel by summarizing a large body of knowledge applied to similar problems and utilizing this knowledge to solve the particular problem.

Service blueprinting (Bitner et al., 2008) is related to QFD with an additional emphasis on time. The service blueprint is a flow diagram of the critical interactions between the provider of the service and the end customer. It clarifies how at each stage of the process how the organization contributes to the customer's experience during the service interaction. The approach of documenting the service can help encourage ways of thinking about improving the customer experience.

These approaches can be used to accompany a road mapping process within an organisation and questions various aspects for future planning and strategy. The most appropriate combinations for an organisation depend upon its business and environment but a multi-faceted view can improve strategic decision making within an organisation. An example is where scenario planning can be combined with road mapping to help look beyond a restricted number of likely future scenarios by exploring alternative possibilities in a disciplined way.

3.3 Applying Road Mapping to Data Management

Data management definition in this dissertation covers all disciplines related to the management of data and where it is viewed as a valuable commodity. In the context of road mapping, the particular approach used in this dissertation is shown in Figure 3.9 in conjunction with the techniques and methods from section 3.1 and section 3.1.1. In particular, the current growth of huge data volumes in many business sectors requires new solutions to tackling the problems arising with these volumes. A customer-centric or end user-view of technology in the context of big data and utility computing (Carr, 2008) as shown in Figure 3.10 can help understand the different types of technology required in addressing big data. Utility computing (Carr, 2008) uses the metaphor of a traditional public utility (such as electricity, water or telephone network) and applies it to a consolidation of computing resources such as computation, storage, networking and services, within a metered or pay-per-use service. Big data has been defined by Jacobs (2009) as "*data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time*". These are data sets whose large size move them out of the ability of commonly used software tools or applications to process, capture and manage the data set within a reasonable period of time.

It has been reported that in 2008 approximately 9.57 zettabytes (957 billion terabytes or 957,000 billion gigabytes) of information were processed by the world's servers (Short et al., 2011). In the context of this report, businesses were defined those conducting work whose units were performed by servers for both heuristic and practical reasons. These enterprise information flows are producing a torrent of transactional data from customers, suppliers, operations; as too are sensors and devices which are recording everything from mobile phones to smart energy meters. In particular, the digital data or "exhaust" of an individual's interactions with services, businesses and organisations are generating a large volume of data which is a by-product of these activities. Big data and data management are creating value by discovering new needs, better segmenting populations, innovating new products or services, and allowing for analytics to inform or potentially automate decision making. The aim of the road mapping approach and scenarios described in this section is to help inform organisations, allowing them to better plan their responses to these aspects of a changing organisational environment.

3.3.1 Applying Road Mapping to Data Management - An example of a multinational enterprise in the chemical sector and its data management needs

An example of using the linked grid analysis (Phaal et al., 2003) for a large scale research project is shown in Table 3.1 and in Figure 3.11. In the context of this scenario, the project is exploring computational chemistry with the goal of solving complex equations describing the natural phenomena using millions of calculations to effectively describe the phenomena. An example usage case within this scenario is concerned with the modeling of electron repulsion in a complex molecule where the application is used to predict the behavior and characteristics of the chemical system.

In Table 3.1, a range of market drivers are shown in the context of a multinational enterprise in the chemical sector. In this example, the overall scores range from the lower range of multimedia context to the upper range scores such as transactional growth or in-vito simulation. The process of road mapping moves to Figure 3.11 by concentrating on one of the market drivers, that of in-vito simulation. The main market driver for this type of business is

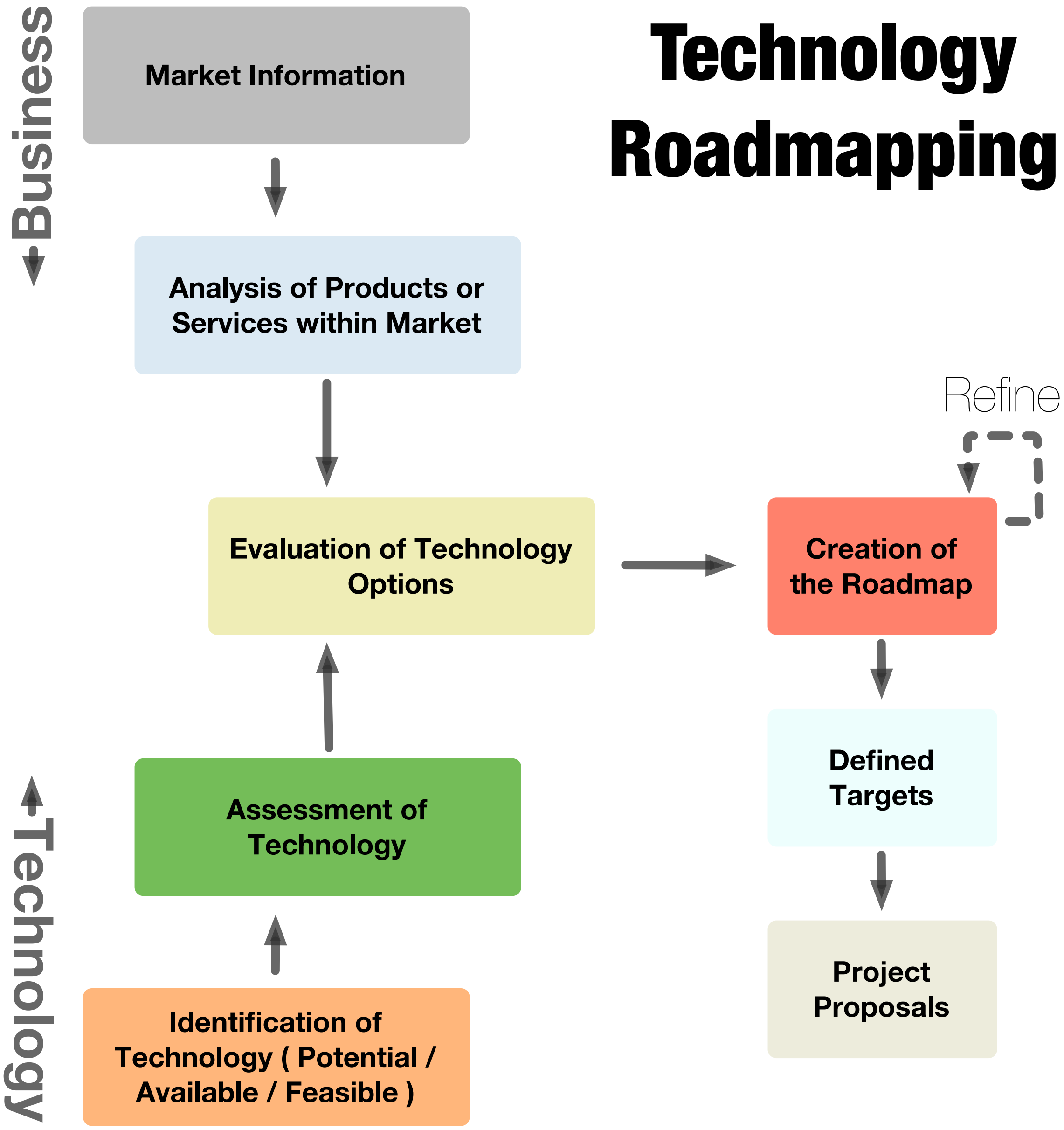

Fig. 3.9: The technology road mapping approach taken in this dissertation

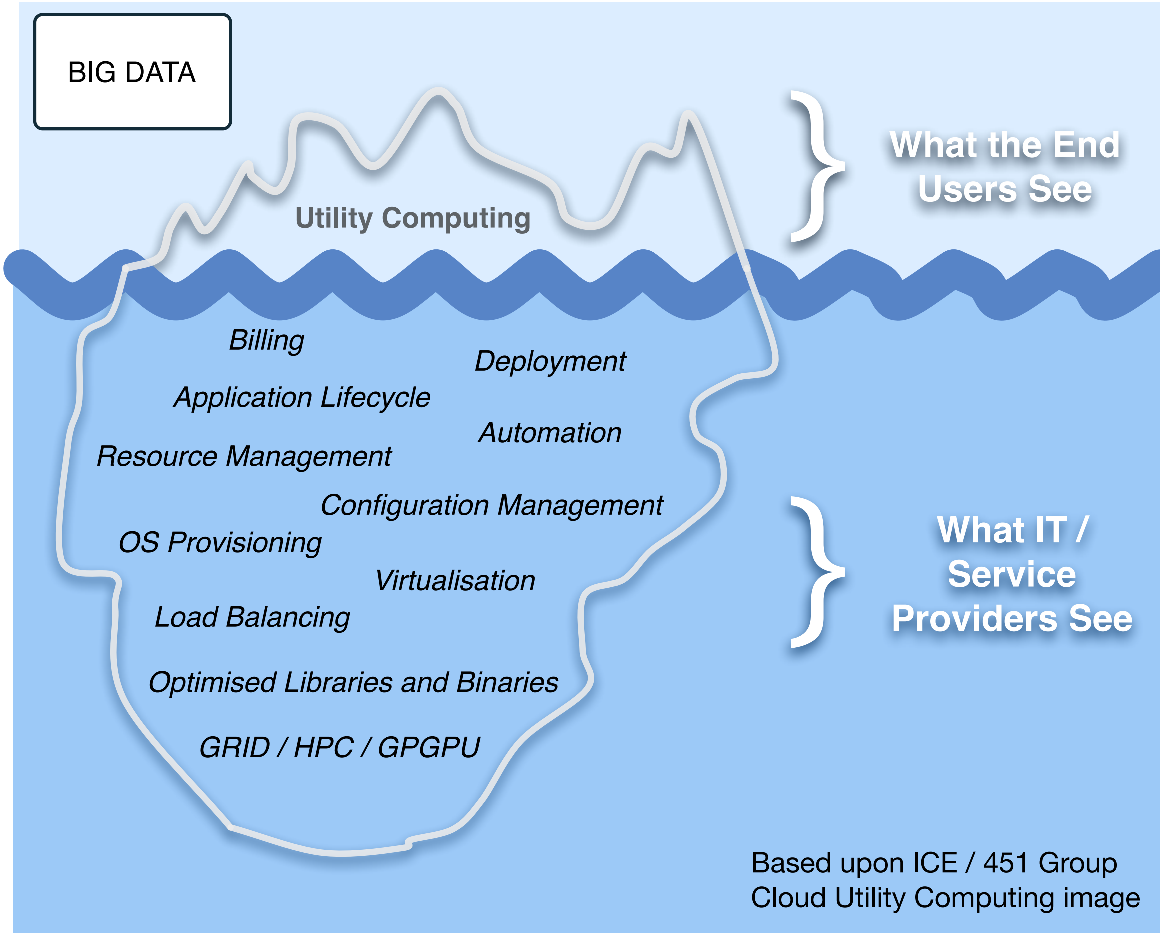

Fig. 3.10: Customer-centric / end-user of technology within the context of big data

Driver Weight	**7**	**3**	**2**	**3**	**6**	**3**	**9**
	Transactional Growth	Multimedia Content	Social Media	Sensor Networks	Customer Profiling	Mobile Devices	In-vito Simulation
Processing Complexity	***	**	*		***	**	***
Automation	**		**	*	*	**	**
Workflow Integration	**			**	*		***
Data Compatibility	*	**				*	*
Security	*			*	*		
Cost	**	**	**	**	**	*	*
Backup / Replication	*			**	**		**
	84	18	10	24	60	18	108

Table 3.1: Business / market cross-impact matrix with market drivers and product features.

the generation of new products, hence the ranking of 9 to the weight for the in-vito simulation category. The key product features for this type of simulation are the processing complexity, the workflow integration, the automation of these simulations as well as backup / replication. These factors and weightings help understand the relationship between the roadmap layer, as illustrated previously in Figure 3.4. In Figure 3.11, the process drills down firstly highlighting the main product features such as processing complexity and workflow integration. Drilling down with the linked grids identifies that problem partitioning and latency scheduling as the key areas required for technology solutions.

These technology solutions need to be viewed in term of the context which is the in-vito simulation of a new compound within the chemical sector. Problem partition is where the simulation algorithm would require that each processor core used would have full information on the aspect it was modeling, e.g. a particular atom. The effective design of a partition will ensure neighboring atoms can be kept close in computer memory allowing for improved performance due to the locality of the data. Latency scheduling is where a deliberate approach is made to address the issue that communication between the many cores on a computer can cause a delay in processing the simulation. Latency scheduling can be used to ensure that data transfers are hidden by ensuring the data is preemptively loaded while the computer is processing an earlier stage of the simulation. This approach hides the delay between copying from potentially another machine to a local machine if running a distributed simulation or if a standalone machine from the disk to memory as there are exponential time differences in these copying operations. The minimisation of such delays helps speed up the processing of a simulation.

The matrices in Figure 3.11 highlight the necessary solutions that provide immediate path towards data management, particularly with the context of computational chemistry where dividing the problem to provide for its efficient solving and controlling / hiding latency through effective scheduling are important approaches to managing the complexity and workflow for this type of scenario.

Driver Weight	9	3	6	3	4	3	4
	Processing Complexity	Automation	Workflow Integration	Data compatability	Security	Cost	Backup / Replication
Problem Partitioning	***	*	*			**	**
Load Balancing	**		**	*		**	*
Effective checkpointing	**		**		*	**	**
Data distribution	*	**	*	*	*	*	*
Locality & data movement	**				*		
Scheduling for latency	***		**	*	**	*	***
Concurrency	*	*	*	**	**		**
	126	**12**	**54**	**15**	**28**	**24**	**44**

Fig. 3.11: Product / service to technology cross-impact matrix with product features and technology solutions.

The final stage of roadmap involves creating a roadmap from the cross-impact matrix information. This maps the market opportunities / triggers to a feature to a technology response with related issues is shown in Figure 3.12.

3.4 Conclusions

This chapter presented an introduction to the road mapping and the supporting techniques used to assist this process. An example using a chemical multinational enterprise in the area of computational chemistry was explored using these techniques from a perspective of data management.

In the following chapters, the methods and scenarios used to explore the research ques-

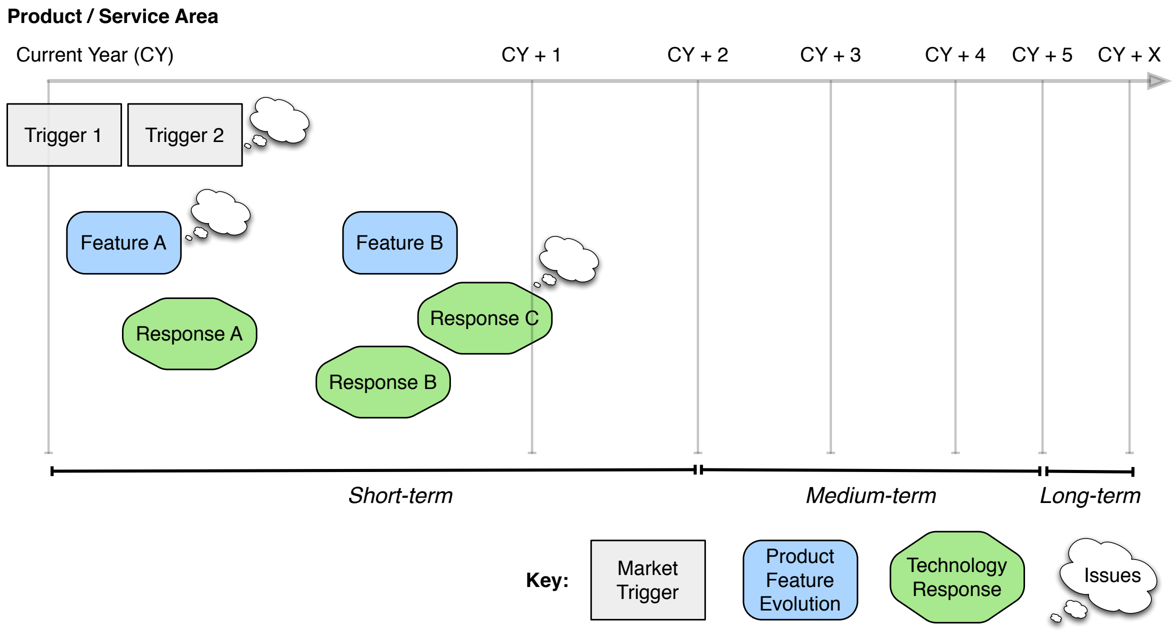

Fig. 3.12: A linked grid analysis produces information required for a roadmap as illustrated here.

tions introduced in chapter 1 will be discussed in greater detail. The next chapter focuses on an introduction of the methodologies used in this dissertation. It uses two scenarios to work through examples of the methodologies in action.

Chapter 4

Empirical Analysis & Scenarios

Scenarios (Cornelius et al., 2005) develop coherent narratives about alternative future possibilities. They are typically used in strategic planning to challenge assumptions, test plans, and develop alternative strategies to address these possibilities. Shell used this approach to plan and anticipate shifts in the world energy markets. The com pany found scenarios (Wack, 1985) to be a better tool than forecasting which often failed "*in anticipating major shifts in the business environment that make whole strategies obsolete*". The use of scenarios in this dissertation aim to test data management strategies and develop adoption strategies for the three cases discussed in this chapter.

This chapter uses two scenarios to explore data management strategies. The first scenario looks at a small-to-medium size enterprise (SME) who wishes to move into a formal data management process within its workflow. The second scenario explores a large scale research project and its data management needs.

4.1 Scenario One - A SME wishing to move to data management

In this scenario, a SME that provides reporting and data services to financial services companies is set to incorporate a data management strategy to protect and enhance the value of its core asset. The company first part of this process is to determine the two cross-impact matrices as shown in Table 4.1 and in Table 4.2. This follows the same process illustrated previously in section 3.3.1.

The SME determines the key market drivers for its product and what are the product features that data management can support. In this scenario, it identifies *processing complexity*, *workflow integration*, *security*, and *cost* as the key features that need to be considered. These features are then considered in terms of technology solutions and the SME identifies *concurrency*, *latency scheduling*, *load balancing*, *fault tolerance*, and *hardware scalability* as the potential solutions for its data management strategy. Features such as *concurrency* (the simultaneously execution of computation) and *latency scheduling* (the management of processing time delays using various mechanisms) are particularly important given the nature of clients that the SME serves and whom expect accurate prompt data.

Driver Weight	**3**	**5**	**6**	**2**
	Charts & Graphs	Real-Time Processing	Transaction Growth	Mobile Devices
Processing Complexity	*	**	***	
Workflow Integration		***	**	*
Security			*	*
Cost		*	*	
	3	30	42	4

Table 4.1: The cross-impact matrix mapping the market drivers and product features for scenario one.

Driver Weight	**4**	**5**	**3**	**3**
	Processing & Complexity	Workflow Integration	Security	Cost
Concurrency	***	**		*
Latency Scheduling	***	**		*
Load Balancing	**	***		**
Fault Tolerance	*	*	*	
Hardware Scalability	***	***	*	**
	48	55	6	18

Table 4.2: The cross-impact matrix mapping the product features to technology solutions for scenario one.

The information that the linked grid analysis produces using the cross-impact matrix approach in Table 4.1 and in Table 4.2 create the roadmap for this scenario, as shown in Figure 4.1. This figure highlights how the issues can be mapped to a feature evolution (for the product or service) to a technology response and where potentially issues may arise. In this scenario, a short four year road map is produced but in practice this would cover a longer period.

Following this example of the SME and using the BCR formula discussed in chapter 2 allows for a financial analysis to be carried out on whether it is of value for this company to adopt this data management approach using the technology solutions proposed in Table 4.2.

The company will require a new cluster with storage to provide for the fault tolerance, scalability, load balancing as well as to improve the concurrency of the data delivery to clients.

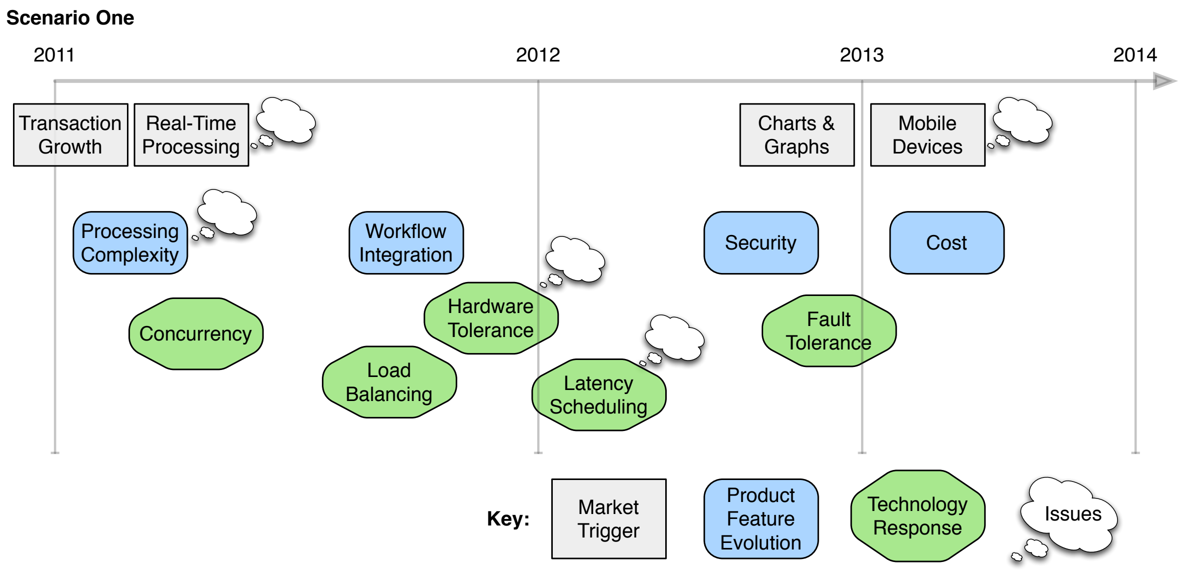

Fig. 4.1: Using the linked grid analysis and the two cross-impact matrices produces the roadmap for scenario one.

The bespoke nature of this project will require both software development and a number of licenses for various packages required within the workflow of the company. The changes to the system will require training and the new hardware will require the recruitment of two additional staff to support the new data management driven workflow. These new features will increase the profits of the company to €2 milion from €750k by facilitating expansion and new clients. The variables for this model are shown in Table 4.3 and the costs for the entire strategy are:

- The cluster or large computer system will cost €500k.
- The software licenses and development costs are estimated at €300k.
- The training of 15 hours will be given to 10 users on the software and on the system at a rate of €100 per hour.
- The administration will require two system administrators to service both the software and system at an individual annual salary cost (including contributions and additional employer costs) of €75,000.

Model variable	Value	Description
P_G	€2m	Profit gained or maintained by the system
C_{AS}	€300k	Cost of the application software
C_T	€15k	Cost required for training
C_A	€150k	Cost required for system administration
C_S	€500k	Cost of the system

Table 4.3: The productivity model variables used by Tichenor (Tichenor and Reuther, 2006) for benefit-cost ratio model as applied in this scenario.

$$productivity(BCR) = \frac{\sum P_G}{C_{AS} + C_T + C_A + C_S}$$

$$productivity(BCR) = \frac{€2m}{€300k + (15 \times 10 \times €100) + (2 \times €75k) + €500k}$$

$$productivity(BCR) = \frac{€2m}{€965k} = 2.07$$

$$IRR_{1year} = BCR - 1 = 2.07 - 1 = 107\%$$

The results of the data management would provide a one year IRR for 107% and therefore justify the cost of the investment. In this example scenario, these figures probably understate the potential for data management as the BCR is calculated for a single year return on investment and it may require additional time to see the full value of this strategy. These results are dependent on the particular figures being accurate which may or may not be the case, a simple technique to gauge the range of these figures is sensitivity analysis. In this scenario, we could increase the cost of software development by 20% and reduce the profits by 20% to see how this would change the final BCR results. The results of this sensitivity analysis show that there is still a IRR of 56%, a range of such analysis should be carried out to test the range for the project to ensure what are the key factors for that this SME needs to be aware of. A wider analysis called scenario analysis can be used to calculate the BCR (or other financial analysis method) where three cases are investigated. The best case uses high revenues with low costs, the worst case uses low revenues with high costs, and the base case is calculated with the given data or initial estimates.

$$productivity(BCR) = \frac{€1.6m}{€360k + (15 \times 10 \times €100) + (2 \times €75k) + €500k}$$
$$productivity(BCR) = \frac{€1.6m}{€1.025m} = 1.56$$
$$IRR_{1year} = BCR - 1 = 1.56 - 1 = 56\%$$

4.2 Scenario Two - A large scale research project and its data management needs

This scenario is for a large scale data management of research data where the commercial aspects are less important when considered with productivity in terms of research. The organisation would still undertake the cross-impact matrix analysis, however the emphasis differs as it is research rather than market drivers that should influence the matrices for this scenario. The research to product features matrix is shown in Table 4.4 and the product features to technology solutions matrix is shown in Table 4.5. A typical example Irish research project requiring data management due to the large volume of data produced include work by the Neonatal Brain Research Group[1] in UCC on the recording of neonatal electroencephalogram (EEG) for the detection of neurological problems in infants. Similar projects with data management requirements can be found in astrophysics at the Dublin Institute for Advanced Studies (DIAS)[2] on stellar evolution or at a number of the Science Foundation Ireland (SFI) research clusters such as at the Biomedical Diagnostics Institute (BDI)[3] on cancer diagnosis or at Systems Biology Ireland (SBI)[4] on cell simulations.

In the case of this scenario, the research workflow being supported is within the bioinformatics domain and targeted at the requirements for high throughput sequencing. New hardware advances are allowing full genomic sequences to be obtained, however this is still the raw data of the organisms DNA which does not include any useful analysis of what the data means. This means that whilst hardware can produce large amounts of sequence data, it must be managed to ensure that it can be effectively analysed for particular drug or clinical needs.

In this scenario, it identifies *processing complexity*, *workflow integration*, *statistical testing*, *quality assessment*, *sequence manipulation* and *sequence targeting* as the key features

[1] `http://www.ucc.ie/en/neonatalbrain/`
[2] `http://www.dias.ie/`
[3] `http://www.bdi.ie/`
[4] `http://www.ucd.ie/sbi/`

that need to be considered. These features are then considered in terms of technology solutions and the research project identifies *automation*, *concurrency*, *latency scheduling*, *data compatibility*, *fault tolerance*, and *software scalability* as the potential solutions for its data management strategy.

Driver Weight	**2**	**5**	**7**	**3**
	Visualisation	Data Quality	Usability	Interoperability
Processing Complexity		*		***
Workflow Integration	**	*	*	**
Statistical Testing	*			*
Quality Assessment	*	***		*
Sequence Manipulation		*	**	
Sequence Targeting	*		**	
	10	30	35	21

Table 4.4: The cross-impact matrix mapping the research drivers and product features for scenario two.

Driver Weight	**7**	**9**	**4**	**2**	**5**	**3**
	Processing Complexity	Workflow Integration	Statistical Testing	Quality Assessment	Sequence Manipulation	Sequence Targeting
Automation	**	***	**	*	**	*
Concurrency	**	**			*	
Latency Scheduling	**	*			*	
Data Compatibility	***	***	*	*	**	*
Fault Tolerance	**	***			***	*
Software Scalability	***	**			***	**
	98	126	12	4	60	15

Table 4.5: The cross-impact matrix mapping the product features to technology solutions for scenario two.

The information that the linked grid analysis produces using the cross-impact matrix approach in Table 4.4 and in Table 4.5 create the roadmap for this scenario, as shown in Figure 4.2. This figure highlights how the issues can be mapped to a feature evolution (for the product or service) to a technology response and where potentially issues may arise. In this scenario, a short four year road map is produced but in practice this would cover a longer period.

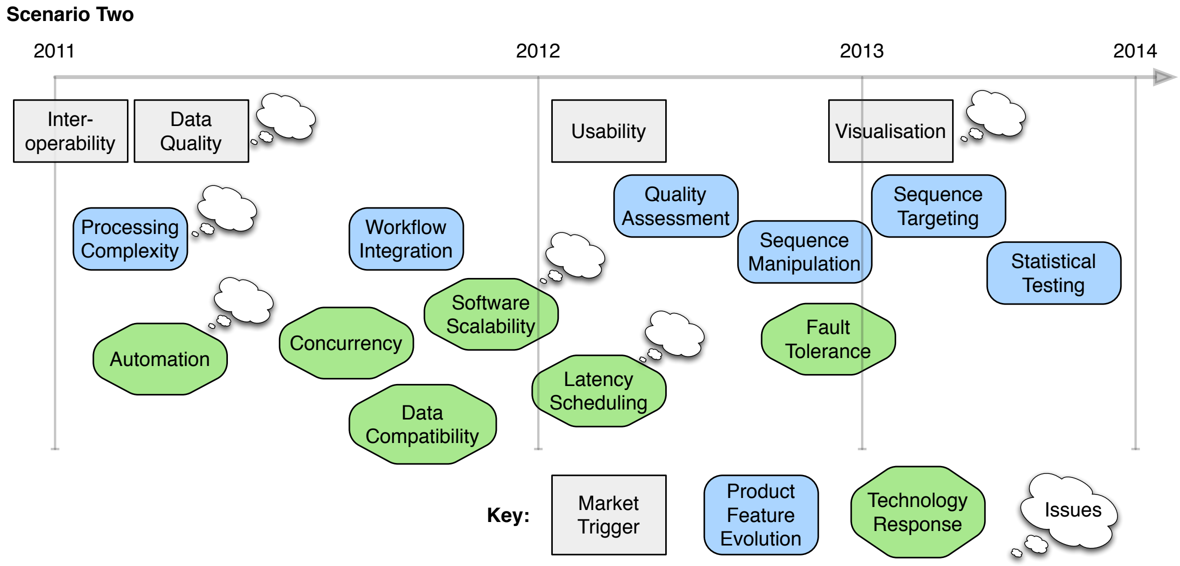

Fig. 4.2: Using the linked grid analysis and the two cross-impact matrices produces the roadmap for scenario two.

Following this example of this bioinformatics research data management issue and using the research oriented BCR formula (Kepner, 2004) discussed in equation 2.4.1 on page 13. This formula allows for a financial analysis (focused more towards time than cost) to be carried out on whether it is of value for this research group to adopt this data management approach using the technology solutions proposed in Table 4.5.

The research group will require a cluster with storage to provide the hardware necessary to run the technology solutions highlighted in Table 4.5. This research will require software development but as bioinformatics codes are generally open-source or developed in-house there will not be any license costs. Parallelisation will be required withing the application development to scale it to deal with the large volumes produced by high throughput sequencing. The proposed system would consist of 100 CPUs (with six cores per CPU) giving 600 CPU cores or $TCPU$, it would require one day a month maintence, and we further assume a 99% operational reliability and a 90% CPU utilization. The $TC = TCPU \times H \times \mu = 600 \times ((365-12) \times 24) \times (0.99 \times 0.90) = 4.52$ million CPU hours annually or 7,548 hours. Kepner (2004) proposes a conservative measure for time saved as $(timesystem\ is\ in\ use) \times (average\ number\ of\ users) \times (1 - \frac{1}{Average\ number\ of\ CPUs\ per\ job})$. If we further assume that the average number of users is 6 and the average number of CPUs required per job is 90 then $(7548) * (6) * (1 - \frac{1}{90}) = 44,784.8$ hours saved. The variables for this model are shown in Table 4.6 and the costs, taking a individual annual salary cost (including contributions and additional employer costs) as €75,000, for the entire strategy are:

- In terms of the time saved by users, a system evaluation yielded that approximately 44,784.8 hours of user time would be saved by the system.
- The cluster or large computer system will cost €200k or the equivalent of 2,000 staff hours.
- The parallelization time is estimated to take 620 hours for 20 users' algorithms.
- Taking a 8 second application launch time with an estimated 3,000 parallel job launches per year, the launch delay would be 6 hours and 40 minutes of application launches that would not be experienced through the use of serial desktop machines.
- The training will take approximate four hours per user so training will take 80 hours in total.
- The administration will require one system administrator or 1,000 staff hours.

Model variable	**Value**	**Description**
T_{SU}	44,784.8 hours	Time saved by users on system
T_P	620 hours	Time to parallelise for the system
T_T	80 hours	Time required for training
T_L	6 hours 40 minutes	Time delay to launch or begin the application using a batch resource management (job scheduler) system
T_A	1,000 hours	Time required for system administration
C_S	2,000 hours	Cost of the system

Table 4.6: The productivity model variables used by Kepner (Kepner, 2004) for benefit-cost ratio model as applied in this scenario.

$$productivity(BCR) = \frac{T_{SU}}{T_P + T_T + T_L + T_A + C_S}$$
$$productivity(BCR) = \frac{€75k \times 44,784.8}{€75k \times (620 + 80 + 6.667 + 1000 + 2000)}$$
$$productivity(BCR) = \frac{€3358869}{€278000} = 12.08$$
$$IRR_{1year} = BCR - 1 = 12.08 - 1 = 1108\%$$

In this example the one-year IRR is 1108%, this highlights an example of how their users are more productive when they use a large scale system for their bioinformatics processing. In this example scenario, these figures probably understate the potential for data management as the BCR is calculated for a single year return on investment and it may require additional time to see the full value of this strategy. It would also be appropriate to conduct an sensitivity analysis to determine the factors influencing the BCR in this scenario.

4.3 Conclusions

This chapter presented two scenarios using the road mapping and the financial analysis methods presented in this dissertation as a means for calculating the value for data management in these scenarios.

In the following chapter, the methods and scenarios introduced in this dissertation are summarised to provide a generic approach for those wishing to calculate the value of data management within their organisation.

Chapter 5

A Generic Approach for Data Management combining Financial Modeling with Road Mapping

This chapter summarises the methods, scenarios and approaches covered previously within a generic approach for road mapping that incorporates financial modeling.

5.1 Tailoring a Generic Approach for Data Management

This dissertation breaks down the creation of a roadmap into three phases of cross-impact matrix mapping, productivity / cost modeling, and roadmap generation, as shown in Figure 5.1. The first phase (stages 1 to 3) help generate the necessary information to underly the roadmap and ensure that the appropriate choices are considered from the market, the product features, the technological solutions as well as from the financial / productivity aspects. The second phase (stage 4) uses this information to create a roadmap. The third and final phase (stage 5) uses this information combined with financial modeling to estimate the costs of this roadmap.

5.1.1 Cross-impact Matrix Mapping

In chapter 3, linked grid analysis using cross-impact matrix mapping was introduced to provide a mechanism to highlight the relationships between the layers of a road map. It involves a four stage process, shown in Figure 3.2, the first three stages select and prioritise the business, product / services, market drivers and technology or competences of the company. These stages define the necessary competencies and link them to the performance of the product / services. An example of the cross-impact matrix mapping is shown in Table 5.1. The fourth stage of creating the road map is described below in section 5.1.2, however in our road mapping approach we add a fifth stage of cost modeling to ensure that the productivity and cost aspects are calculated providing measurable costs or benefits to accelerate the adoption of the roadmap.

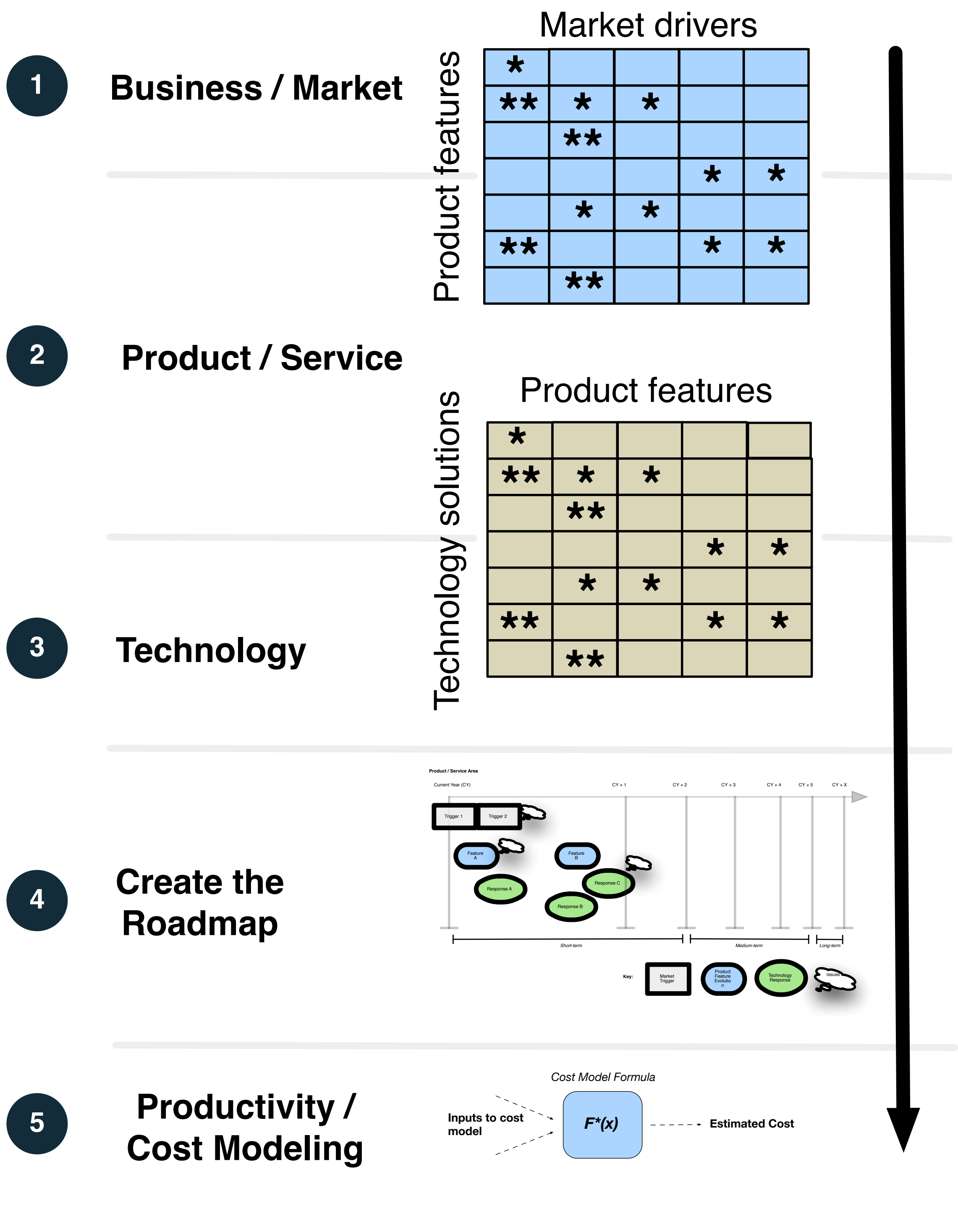

Fig. 5.1: The proposed generic approach to road mapping coupled an additional stage of financial modeling.

Driver Weight	(1-10)
	Market Drivers / Product Features
Product Features / Technology Solutions	(*/**/**) (1/3/7)
	Driver Weight × *Product Features or Technology Solutions*

Table 5.1: Business / market cross-impact matrix with market drivers, product features or technology solutions.

5.1.2 Create the Road Map

A linked grid analysis using the cross-impact matrix approach, as shown Table 5.1, provides the necessary information to create a roadmap. The mapping of market opportunities / triggers to a feature to a technology response with related issues is shown in Figure 5.2. The roadmap provides the basis of a cohesive technology strategy identifying a plan with the market, product / service, and technology highlighted within a defined timeframe.

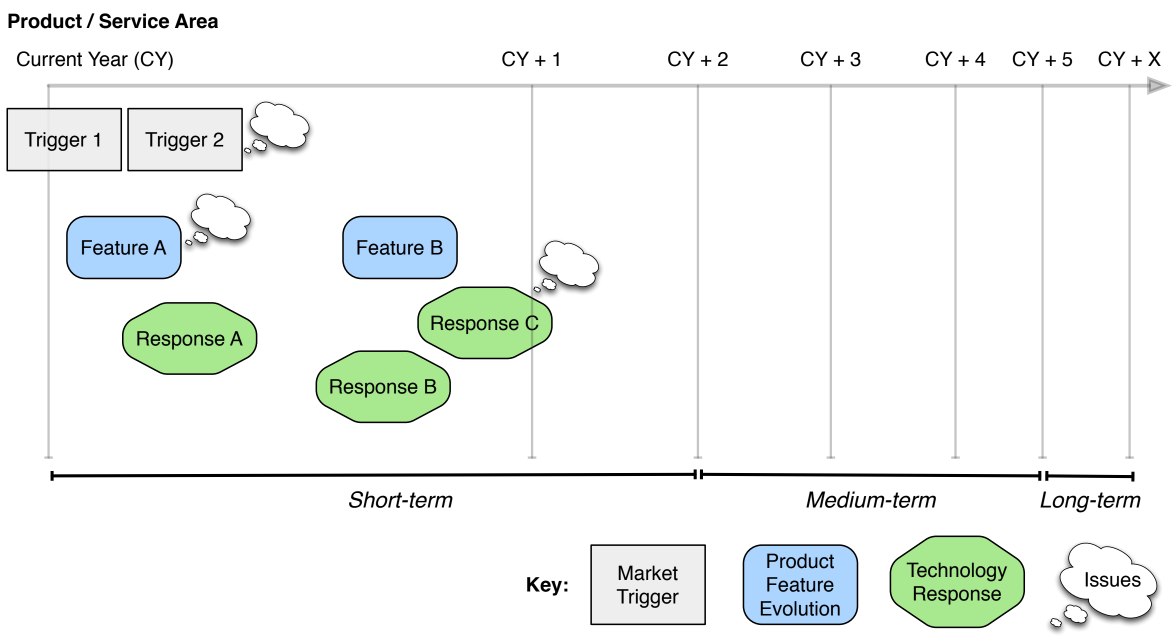

Fig. 5.2: A linked grid analysis produces information required for a roadmap as illustrated here.

5.1.3 Productivity / Cost Modeling

The creation of a roadmap does not include any financial information. In this dissertation we purpose the use of a range of financial modeling approaches to provide tangible estimates that can be linked to the roadmap. A number of approaches were suggested in chapter 2 such as the benefit cost ratios by Kepner (2004) and by Tichenor and Reuther (2006) or the Expected Commercial Value (Cooper, 2000) (ECV). The benefit cost ratio uses this simple model, $BCR = 1 + IRR,\ or\ IRR = BCR - 1$, it has two derivations either for academic/research or for business/commercial environments. The difference between these derivations is mainly concerned with the priorities of the environments, so time costs is prioritised over monetary for research and vice-verso for business.

$$productivity\ academic\ /\ research\ (BCR) = \frac{T_{SU}}{T_P + T_T + T_L + T_A + C_S}$$
$$productivity\ business\ /\ commercial\ (BCR) = \frac{\sum P_G}{C_{AS} + C_T + C_A + C_S}$$

The ECV formula is similar to the net present value formula with modifications that maximises the *expected* value whilst recognising that projects are incremental in nature and contain multiple decision points within the course of the project.

$$ECV = [(NPV \times P_{CS} - C) \times P_{TS}] - D$$

The incorporation of financial measures into the road mapping process provide added insights and an estimation with regard to the value that a successful implementation of the roadmap will bring to the organisation or company.

5.2 Conclusions

This chapter has briefly summaries the methods, scenarios and approaches to provide a generic approach for combining road mapping and financial modeling to provide a deeper insight into the technology strategy whilst providing estimates of the cost benefits to the business.

Chapter 6

Conclusions

This final chapter gives a summary of the work undertaken in this dissertation, it further explores a number of future directions based on the research presented. From the earlier research and literature presented in this dissertation, it is clear that road mapping when combined with financial modeling requires additional research. The combination of approaches explored in this dissertation offers new insights when combined. A technology strategy plan can help provide a more detailed explanation utilising the information from both the roadmap and from the financial modeling. This detail can ensure that management can make more informed business decisions in the area of data management.

6.1 Research Problems

The initial motivation for this thesis was to explore two broad issues regarding the adoption of data management in organisations. These issues were:

1. How can an organisation best plan their technology strategy to adopt data management
2. How can the value of data management to an organisation be best conveyed to encourage its adoption

These issues were focused into two narrower research questions:

RQ1 How can data management be integrated into organisation ?

RQ2 How can the benefits and costs associated with data management be calculated ?

The next section addresses these questions and shows where the work in this dissertation help in answering the questions.

6.1.0.0.1 RQ1 — How can data management be integrated into organisation ? RQ1 explored the issue of integrating data management into an organisation. Chapter 3 investigated RQ1 and explored road mapping as a method that provides a plan for integration of data management. This method allows for the issues, drivers, and potential solutions for the integration of data management to be structured and provides a visual means to represent them.

The best approach for integrating data management would be to ensure that the approaches covered in this dissertation are explored with regards to the particular context and organisation. This work has addressed this question by providing the necessary methods to integrate data management.

6.1.0.0.2 RQ2 — How can the benefits and costs associated with data management be calculated ? RQ2 explored the issue of benefits versus costs for data management. Chapter 2 investigated RQ2 and explored a range of general financial modeling methods before presenting three specific cost models relevant for data management. These specific models allow for the costing and benefits arising from the adoption of data management to be estimated.

This work has explored this question by using focusing on specialised financial models that incorporate the necessary technical complexities within the cost calculations to accurately model the benefits and costs associated with data management.

A brief discussion on generalising the work in this dissertation is given in the next section 6.2 and situates the views derived from both approaches in a broader context.

6.2 Generalising the Results

This dissertation has synthesised two disparate streams of research from strategic management in the form of road mapping and from financial accounting in the form of financial modeling. The proposed approach for combining road mapping and financial modeling within the context of data management provides a structured approach that includes the development of a technology strategy and the estimation of the costs plus benefits of implementing such as strategy. A typical organisation will often address these two elements separately however by combining them the potential costs and/or benefits can be established in parallel with the development of a strategy. This type of parallel development combining both elements would help ensure a better costing of the strategy by ensuring those that cost the elements are more familiar with the issues and constraints, particularly in the complex financial models for data management that require a certain degree of understanding of this activity. This proposed combination of financial modeling and road mapping builds upon prior research as discussed in chapter 2 and in chapter 3, respectively. This combination itself will need further real world studies to address its limitations.

6.3 Limitations

There are a number of limitation with this work that should be noted. The work in this dissertation does not validate the existing models for cost or benefit modeling. The length of time required to successfully plan, implement and review a technology strategy based on the road mapping approach is approximately three to five years. This has prevented a comprehensive practical case study from being implemented and reviewed to gauge how this proposed work would compare to using both of these methods in isolation.

6.4 Future Work

This section provides several areas for additional research stemming from this dissertation that can be further developed to improve this work.

6.4.1 Comprehensive cost models

The existing cost models focus on single aspects relevant to data management, however to provide a better estimate new models that link compute, network, and storage aspects should be combined. These models would still lack integration to the benefit cost model approach discussed in this dissertation so both types of model should be used to provide detailed information on both aspects. This combination would add greater depth and coverage to the models. These models are all based on initial estimations and further use in practice should help in improving their accuracy and validity.

6.4.2 Integrating supplementary methods into this approach

A range of supplementary methods were described in chapter 3 such as scenario planning (Cornelius et al., 2005, Wack, 1985), Delphi (Rowe and Wright, 1999), information markets (Hahn and Tetlock, 2006), TRIZ (Altshuller, 1996) and service blueprinting (Bitner et al., 2008). The current approach could be accompanied by one or more of these methods to improve future planning and strategy. Strategic decision making would be improved by providing a multi-faced view on the issues, opportunities and technologies faced when implementing a technology strategy.

6.4.3 Exploiting related technology solutions

In the case of data management there are related technology solutions that could be used in conjunction with the proposed roadmap but that are specific focused at the infrastructural aspects of data management. This plumbing layer may not be considered when focusing specifically on the issues or problems facing a particular organisation and its business challenges. Examples of this type of infrastructural support include the use of automated deployment systems such as Puppet (Turnbull and McCune, 2011) or Chef (Nelson-Smith, 2011) or the use of de-duplication filesystems such as lessfs[1]. These technologies can support a data management strategy by providing infrastructural support that is not directly targeted at the immediate market drivers but which nevertheless can underpin the technology solutions being used to address these drivers.

6.5 Final Remarks

Combining financial cost models and road mapping can offer several advantages; however there are several issues that need to be considered when designing and presenting them. This work presented a body of work exploring road mapping for technology strategy development, a variety of cost models that are relevant for data management, as well as providing examples for how to utilise these models in a realistic scenario. The body of work presented in this dissertation has summarised the work of road mapping and cost modeling and applied it to the specific challenge of data management.

[1] `http://www.lessfs.com`

References

Altshuller, G. (1996). *And Suddenly the Inventor Appeared*. Technical Innovation Center Inc, Worchester, MA, USA.

Bitner, M. J., Ostrom, A. L., and Morgan, F. N. (2008). Service blueprinting: A practical technique for service innovation. *California Management Review*, 50(3):66–94.

Brigham, E. F. and Erhardt, M. C. (2002). *Financial Management: Theory and Practice*. Thomson Learning.

Carr, N. (2008). *The Big Switch: Rewiring the World, from Edison to Google*. W. W. Norton and Co.

Christenson, C. (1997). *The Innovator's Dilemma*. Harvard Business School.

Clausing, D. (1994). *Total Quality Development*. ASME Press, New York, USA.

Cooper, R. G. (2000). *Product Leadership: Creating and Launching Superior New Products*. Perseus Publishing.

Cooper, R. G., Edgett, S. J., and Kleinschmidt, E. J. (2001). *Portfolio Management For New Products*. Perseus Publishing, 2nd edition.

Cornelius, P., Van de Putte, A., and Romani, M. (2005). Three decades of scenario planning in shell. *California Management Review*, 48(1):92–109.

Galvin, R. (1998). Science roadmaps. *Science*, 280:803.

Goffin, K. and Mitchell, R. (2010). *Innovation Management: Strategy and Implementation Using the Pentathlon Framework*. Palgrave Macmillan, 2nd edition.

Groenveld, P. (2007). Roadmapping integrates business and technology. *Research-Technology Management*, 50(6):49–58.

Hahn, B. and Tetlock, P. (2006). *Information Markets- A New Way of Making Decisions*. AEI-Brookings Press.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 8(52):36–44.

Johnson, R. W. and Lewellen, W. G. (1972). Analysis of lease-or-buy decision. *Journal of Finance*, 27(4):815–823.

Kaplan, R. S. and Norton, D. P. (1996). *The Balanced Scorecard - Translating Strategy into Action*. Harvard Business School, MA, USA.

Kepner, J. (2004). High performance computing productivity model synthesis. *International Journal of High Performance Computing Applications*, 18(4):505–516.

Matzler, K. and Hinterhuber, H. (1998). How to make product development products more successful by integrating kano's model of customer satisifaction into quality function deployment. *Technovation*, 18(1):25–38.

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8).

Nelson-Smith, S. (2011). *Test-Driven Infrastructure with Chef*. O'Reilly Media, 1st edition.

Phaal, R., Farrukh, C. J. P., Mitchell, R., and Probert, D. R. (2003). Starting-up roadmapping fast. *Research-Technology Management*, pages 28–52.

Phaal, R., Farrukh, C. J. P., and Probert, D. R. (2004). Technology roadmapping - a planning framework for evolution and revolution. *Technological Forecasting and Social Change*, 71:5–26.

Ranganathan, P. (2011). From microprocessors to nanostores: Rethinking data-centric systems. *IEEE Computer*, 44(1):39–48.

Rowe, G. and Wright, G. (1999). The delphi technique as a forecasting tool- issues and analysis. *International Journal of Forecasting*, 15(4):353–375.

Short, J. E., Bohn, R. E., and Baru, C. (2011). How much information? 2010 report on enterprise server information. Technical report, Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego.

Spinosa, C., Flores, F., and Dreyfus, H. (1997). *Disclosing New Worlds*. MIT Press.

Tichenor, S. and Reuther, A. (2006). Making the business case for high performance computing: A benefit-cost analysis methodology. *CT Watch Quarterly*, 2(4A):2–8.

Turnbull, J. and McCune, J. (2011). *Pro Puppet*. Apress, 1st edition.

Wack, P. (1985). Scenarios: Uncharted waters ahead. *Harvard Business Review*, 63(5):72–89.

Walker, E. (2009). The real cost of a cpu hour. *IEEE Computer*, 42(4):35–41.

Walker, E. (2010). To lease or not to lease from storage clouds. *IEEE Computer*, 43(4):44–50.

Part II

The Appendices

This part of the dissertation covers a range of additional but relevant background material for the questions being posed.

A.1 Appendix A - Chapter 2 - Additional Details

This appendix provides the full detail on two complex cost models, a CPU core hour model and a Hard Disk storage model.

A.1.1 Valuing Methodologies for CPU Core Hour Cost Model

The assumptions in this model means that there is no salvage or no taxation issues given there is no profit generated. This is shown where the cash revenue is zero in equation A.1.1.

$$NPV = \sum_{T=1}^{Y} \frac{-L_T}{(1+k)^T} - A \tag{A.1.1}$$

Taking a recap on the discounting of cash with a rate of interest to incorporate the time value of money, the cost of capital k can be taking by this formula for present value (PV) for a future value (FV) in year T as shown in equation A.1.2.

$$PV = \frac{FV}{(1+k)^T} \tag{A.1.2}$$

In the context of an investment with an annual amortized cash flow (the profit minus the cost) C_T for Y years, the NPV can be calculated using this equation:

$$NPV = \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T} \tag{A.1.3}$$

In computing, Moore's law (Moore, 1965) refers to the biennial halving of CPU performance, however whilst this is no longer true in the context of transistors being doubled per chip but other advances such as the doubling of cores per chip, developments in photonics, in nonvolatile memory and in 3D stacking (Ranganathan, 2011) have ensured that this law still holds after a fashion. In this work, we will assume this trend is continuing which allows for the prediction of the future capacity FC of a T-year-old CPU to be discounted to its present capacity PC using the following equation:

$$PC = \frac{FC}{(\sqrt{2})^T} \tag{A.1.4}$$

The maximum CPU hours consumed annually by users of a cluster can be defined as the total useful capacity (TC) of the cluster. An example of a 256 CPU cluster with a 60% utilization can have its TC calculated using the PC formula as shown in equation A.1.4. The TC for this example would be 256 CPUs x 365 days x 24 hours x 0.6 CPU hours per year. The net present capacity (NPC) of a cluster over an operational life of Y years can be defined by:

$$NPC = TC \times \sum_{T=0}^{Y-1} (\frac{1}{\sqrt{2}})^T \Rightarrow NPC = TC \times \frac{1-(\frac{1}{\sqrt{2}})^Y}{1-\frac{1}{\sqrt{2}}} \quad \text{(A.1.5)}$$

This makes an assumption that the $TC = TCPU \times H \times \mu$, where $TCPU$ is the total CPU cores in the cluster; where H is the expected number of operational hours the cluster provides on an annual basis, and where μ is the expected server utilization. Walker (2009) builds on this model to propose the NPV for a CPU hour as:

$$R = NPV/NPC \quad \text{(A.1.6)}$$

Exploring Walker's (2009) model for the case of an outright purchase of a cluster using substitution in equation A.1.6, taking equation A.1.3 on page 57 for NPV and equation A.1.5 for NPC provides the definition for the purchase case as:

$$R(purchase) = \frac{(1-\frac{1}{\sqrt{2}}) \times \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{(1-(\frac{1}{\sqrt{2}})^Y) \times TC} \quad \text{(A.1.7)}$$

Building on the earlier example of a small cluster with a cost of capital of 8%; an operational lifespan of four years; and where the computational capacity of the cluster is 256 CPUs (assuming six cores per CPU) gives 1536 CPU cores. A cluster requires preventive maintenance and upkeep so it would be unavailable for at least one day a month; a further assumption is that in the available time there will be 99% operational reliability and a 90% CPU utilization. This would give a $TCPU$ of 1536 CPUs and the $TC = TCPU \times H \times \mu = 1536 \times ((365-12) \times 24) \times (0.99 \times 0.90) = 11.59$ million CPU hours annually.

The second type of investment strategy is to lease the necessary CPU power. The major difference in this approach is that there is an underlying assumption that there is no depreciation of the computational capacity due to the competitive nature of the market which will allow the lessee to always utilise the latest technology for their required capacity. This allows the lease case to be defined as:

$$R(lease) = \frac{\sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{Y \times TC} \quad \text{(A.1.8)}$$

A third investment strategy involves purchasing the cluster and upgrading it annually with the latest CPUs to avoid any performance degradation costs. This is factored into the equation by taking an approximation of the CPU cost to that of the original cluster purchase price and incorporating it into the annual operating cost. This requires a modification to the NPV formula as shown here:

$$NPV = C_0 + \sum_{T=0}^{Y-1} \frac{C_T - A}{(1+k)^T} \quad \text{(A.1.9)}$$

The purchase upgrade strategy for a server means that CPU performance degradation of the cluster is no longer relevant and the purchase upgrade case can be defined as:

$$R(purchase\ upgrade) = \frac{C_0 + \sum_{T=0}^{Y-1} \frac{C_T - A}{(1+k)^T}}{Y \times TC} \quad \text{(A.1.10)}$$

A.1.2 Valuing Methodologies for Hard Disk Storage Cost Model

Walker (2010) builds upon his earlier CPU cost model (Walker, 2009) to explore the purchase or lease investment strategies for storage. It is focused on the SATA hard disk drive purchase, replacement, and end-of-life salvage as well as including utility / power costs and human operator costs. The net present value accounting for the time based value of money is the key to this cost model and is simplified as follows:

$$NPV = \sum_{T=0}^{N} \frac{P_T - C_T^P}{(1+I_K)^T} + \frac{S}{(1+I_K)^N} - E, \quad \text{(A.1.11)}$$

where P_T is the annual profit from the purchased asset in year T; C_T^P is the projected operating cost of the asset in year T, I_K is the cost of capital; N is the total operational lifespan of the asset in years; S is the salvage or resale value of the asset after N years; and E is the purchase (capital) cost of the asset. This can be further modified to take into account the lease case as follows:

$$NPV = \sum_{T=0}^{N} \frac{P_T - C_T^L}{(1+I_K)^T} - \sum_{T=0}^{N} \frac{L_T}{(1+I_R)^N}, \quad \text{(A.1.12)}$$

where C_T^L is the projected operating cost of the leased asset in year T, L_T is the lease payment at year T; and I_R is the interest rate of the lease payments. The structure will generally regard the leasing rate as being smaller than the cost of capital I_K due to the predictability of this payment.

The NPV for purchase or NPV for lease of assets allows for the decision to be made using the following criteria: If the $NPV(\Delta NPV) \geq 0 \Rightarrow$ purchase; if $\Delta NPV < 0 \Rightarrow$ lease, where $\Delta NPV = NPV_P - NPV_L$. In Table A-1, the variables used for the model are listed. Walker's (Walker, 2010) model had a number of assumptions including where $\lceil V_T \rceil_\Omega$ was taken as an operator returning the minimum number of Ω-sized disk drives required to store V Gbytes of data. The derived terms $S, C_T, and E_T$ represent the expected end-of-life sale / salvage value of the disks, the operating cost in year T, and the capital cost in year T,

Model variable	Description
δ	Cost of electric utility ($/kilowatt hour)
Ω	Size of purchased disk drives (Gbytes)
ρ	Proportional difference between human effort in maintaining a purchased versus a leased storage infrastructure
Υ	Used disk depreciation factor on salvage ($[0.0, 1.0]$)
C	Disk controller unit cost ($)
H_T	Annual human operator salary ($)
I_F	Risk-free interest rate (percent)
κ	Current per-Gbyte storage price ($/Gbyte)
L_T	Expected annual per-Gbyte lease payment ($/Gbyte/year)
P_C	Disk controller power consumption (kW)
P_D	Disk drive power consumption (kW)
V_T	Expected storage requirement in year T (Gbytes)

Table A-1: The storage model variables used by Walker (Walker, 2010) for the storage purchase-or-lease model.

respectively.

$$\Delta NPV = \sum_{T=0}^{N} \frac{C_T - E_T + L_T}{(1+I_F)^T} + \frac{S}{(1+I_F)^N} - C \tag{A.1.13}$$

$$S = \gamma \times \Omega \times \lceil V_T \rceil_\Omega \times \kappa \times e^{-0.438T}$$

$$C_T = -\rho \times H_T - (365 \times 24) \times \delta \times (P_C + P_D \times \lceil V_T \rceil_\Omega)$$

$$E_T = (1.03 \times \lceil V_T \rceil_\Omega - \lceil V_{T-1} \rceil_\Omega) \times \Omega \times \kappa \times e^{-0.438T}$$