A Framework to Manage Data Science Initiatives

Sandeep MATHUR[1], Shankar SANKARAN[2], Samuel MACAULAY[3] and Ivor TSANG[4]

Director, Transport for NSW, Level 1, 22 Ennis Road, Milsons Point, NSW 2061, Australia. email: Sandeep.Mathur@transport.nsw.gov.au [1]
Professor Organizational Project Management, University of Technology Sydney, School of the Built Environment, UTS City Campus, PO Box 123, Broadway, NSW 2007, Australia. email: Shankar.Sankaran@uts.edu.au [2]
Senior Lecturer, University of Technology Sydney, UTS Business School, UTS City Campus, PO Box 123, Broadway, NSW 2007, Australia. email: Samuel.Macaulay@uts.edu.au [3]
Professor of Artificial Intelligence, University of Technology Sydney, Australian Artificial Intelligence Institute, UTS City Campus, PO Box 123, Broadway, NSW 2007, Australia. email: Ivor.Tsang@uts.edu.au [4]

ABSTRACT

Data is increasingly ubiquitous in organizational life and Data Science Initiatives (DSIs) have emerged as a popular mechanism for extracting value from it. However, the track record of these programs has drawn substantial criticism. For example, the success rate of delivering DSIs is not perceived as high with Gartner estimating that 85% of projects fail. DSIs have unique characteristics and pose challenges delivering the envisaged value when using traditional processes for managing ICT-enabled programs. There are occasions when DSIs should be managed as Exploratory Projects.

In this theoretical paper, we review the related delivery frameworks and propose a framework synthesizing program management, change management, scaled agile, data management and data science domains. The framework covers people and processes and specifically excludes products and technologies. The framework may enable consistency in how the practitioners plan and execute the initiatives potentially leading to an improvement in the success rate of DSI implementations.

KEYWORDS

Exploratory Projects; Program Management; Change Management; Scaled Agile; Data Management; Delivery Framework.

Declaration of Conflicting Interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: This research is supported by an Australian Government Research Training Program Scholarship.

INTRODUCTION

Data Science Initiatives (DSIs) have unique characteristics and pose challenges delivering the envisaged value when using traditional processes for managing ICT-enabled programs. Due to uncertainty they carry in data, scope and schedule, DSIs often present themselves as candidates to be managed as Exploratory Projects. Furthermore, the Waterfall approaches to program management adopted by peak bodies, set up structural tensions between business case development, program design, delivery and benefits realizations that decouple value creation from capture and thus undermine coherent governance across the investment life-cycle. In this paper, we expand on existing frameworks proposed in the program management, change management, scaled agile, data management, and data science domains and propose a synthesized framework to deliver DSIs as Exploratory projects. In context of this paper, we use the term Data Science Initiative (DSI) to include investments in Data Analytics, Business Intelligence and Data Science including Machine Learning and Artificial Intelligence.

Balancing exploration and exploitation is key to organizational success and survival (March, 1991). Exploitative projects focus on optimizing cost-quality-time triple constraints to deliver new products and services whereas exploratory projects are those projects where neither the goals nor the means of attaining them are clearly defined from the outset (Lenfle, 2008).

In this paper, we draw on in-depth case-studies of six Data Science Initiatives (DSIs) delivered over last four years at Transport for NSW (Transport). Transport is a state government enterprise responsible for delivering safe, integrated and efficient transport systems to the people of NSW. *Figure 1* provides an overview of the six DSIs used as case studies which includes the delivery timeline and complexity. Case Study of six DSIs has showed that they have unique characteristics (Mathur, 2019) around degree of uncertainty; enablers for decision-making; unclear goals; interdependency and skills requirement.

We aim to contribute to project and program management research by proposing a synthesized framework explicating the different logic (Lenfle, 2016) required to deliver DSIs as exploratory projects and incorporating program management, change management, agile delivery, data management, and data science domains.

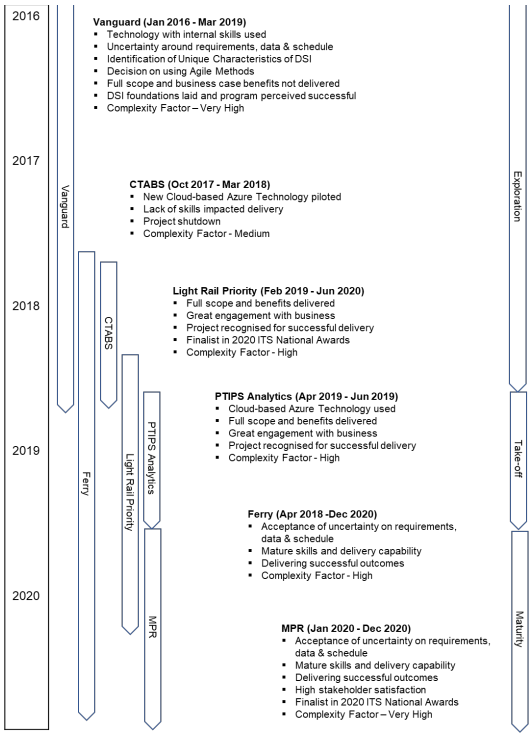

Figure 1. **Overview of six Transport for NSW DSIs**

MOTIVATION

Creation of a comprehensive program delivery framework for DSIs requires understanding of several domains and this section attempts to cover all such domains.

We see DSIs being typically implemented as a Program on a continuous spectrum rather than a single one-off project and will focus on Program Management rather than Project Management processes for delivery. A review of Program Life Cycle has identified gaps in using it for delivery of DSIs. Both PMI (Project Management Institute, 2017) and Managing Successful Programmes (MSP) (OGC, 2004) standards are widely accepted and used in the industry with PMI being principle-based and OGC providing detailed guidance on program management.

Value realization for any program occurs when the product and service created is adopted by users. Change management is a systematic approach that includes dealing with the transition or transformation of organizational goals, core values, processes or technologies. Kotter's Change Management Model (Kotter, 2007); McKinsey's 7-S Change Management Model (Lorenzi & Waterman, 1985); ADKAR Change Management Model (Hiatt, 2006) and Kübler-Ross Five Stage

Change Management Model (Kübler-Ross, 2009) are some of the popular models used because of the simplicity in understanding them.

A need for large projects which are often globally distributed with teams requiring collaboration and coordination has led to popularity of scaled-agile frameworks such as Scaled Agile Framework (SAFe), Large-Scale Scrum (LeSS) and Lean Scalable Agilty for Engineering (LeanSAFE) (Ebert & Paasivaara, 2017; Leffingwell, 2007). In context of DSIs, we see the relevance of scaling is high as often multiple geographically spread teams within an organization are involved in delivering data science outcomes. A comparison of various five scaled agile framework shows each of them have strengths depending upon the use case in an organization.

Data Management is the development, execution, and supervision of plans, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles (Earley, 2017).

Development of a DSI Delivery Framework requires good understanding of Data Mining and Data Science delivery processes. The Knowledge Discovery in Databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996); Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000); Sample, Explore, Modify, Model and Assess (SEMMA) model (SAS Institute, 2009); OSEMN model (Mason & Wiggins, 2010); Team Data Science Process (TDSP) (Severtson, Franks, & Ericson, 2017) and Foundational Methodology for Data Science (FMDS) methodology (Rollins, 2015) are models considered appropriate in this context.

This view of the literature motivated us to ask the following research question: *"What design principles should be incorporated in a Data Science Initiative (DSI) Delivery Framework so that program managers can adopt a predictive path to realize value from such investments*?".

RESEARCH SETTING AND METHODS

Our research insights emerged from primary author's desire to deliver DSIs effectively underpinned by Future Transport 2056 Strategy to embed technologies such as big-data, internet of things, machine learning and artificial intelligence to deliver and improve customer journeys. Taking a practice lens on delivery of DSIs guided us to focus on full life-cycle of DSIs. Such a focus requires deep engagement in the field, observing and interacting with decision-makers, business stakeholders, program managers and delivery team members. As a result, we chose to

study delivery of DSIs within a single organization (Transport) where the primary author of this paper is employed full-time and is setting up DSI delivery capability while delivering DSIs. This gave him access to data to conduct the case studies. To obtain granularity of program life-cycle as well as variation for analytical comparisons, we used an embedded case design (Yin, 2018) to track the unfolding of six DSIs in Transport, each of which provided a unique scope and opportunity to build DSI delivery capability. The six DSIs provide us with an opportunity to use an embedded case study research method covering all three purposes – exploratory, descriptive and explanatory (Scholz & Tietje, 2002; Yin, 2018). Our interest was to understand DSI delivery as experienced by the organizational participants themselves and identify uniqueness with this portfolio of initiatives to bring in improvements within the organization.

We used a variety of evidence including documents, artifacts, and participant observations from each DSI. Consistent with inductive research approaches, our research question emerged over time, as we engaged iteratively with evidence from the field and extant research that helped us make sense of what we had found.

The primary author is the program manager of the six DSIs chosen as case studies which were delivered between January 2017 to December 2020 or are still being delivered and thus brings in-depth insights of the program life-cycle.

This paper organizes the case by bracketing it into three project-stages: Exploration, Takeoff and Maturity stages Transport went through while six DSIs were delivered. The stages can be roughly mapped to DSIs delivery timeline of Exploration stage mapping to Vanguard & CTABS; Take-off stage mapping to Ferry, Light Rail Priority and PTIPS Analytics; and Maturity stage mapping to MPR. *Figure 1* shows the timeline and highlights of the six DSIs indicating author's journey from uncertainty and frustration of not being able to deliver program outcomes as per the schedule to acceptance of exploratory nature of DSIs and ability to plan the uncertainty and engage the stakeholders effectively. While each of the six DSIs were unique, this paper focuses on first (Vanguard) and sixth (MPR) as they represent boundary conditions of story presented here i.e., we present details of initial Exploration stage and close with that of Maturity stage.

DATA COLLECTION AND ANALYSIS

Six DSIs from Transport managed by the author have been used to collect data. Four DSIs are closed and two are still being delivered which has allowed us both real-time and retrospective data collection. While the scale of the DSIs is different, together they paint a good picture of unique characteristics and business cases. Table 1 shows the gaps and issues identified across three program phases.

Table 1. Program Life-Cycle Deliverables & DSI Gaps & Issues

Key Phase Deliverables	Gaps & Issues for DSIs
Program Definition Phase	
Key deliverables of this phase are Business Case, Program Charter and Program Management Plan.	• For DSIs, risks associated with both costs and benefits are high. Considering the time it takes to develop and get a Business Case approved in both public and private sectors, the accuracy of the documents is questionable. • Unless the Program Management Plan stays at a high level, the accuracy of scope and schedule is low. The delivery mechanism will evolve as the Components are identified and executed.
Program Delivery Phase	
In this phase, individual Components are initiated, planned, executed, transitioned, and closed while benefits are delivered, transitioned and sustained in accordance to the Program Management Plan.	• For DSIs, identification of all Components upfront is difficult at the time Program Management Plan is developed and hence only limited planning can be done due to high degree of uncertainty • The Benefits will be discovered as the Components are planned & executed again due to high degree of uncertainty
Program Closure Phase	
In this phase, the Program Benefits are transitioned to sustaining organization and program is closed.	• While sponsor and stakeholders are continuously communicated and kept informed on both the costs and benefits delivered, for an un-initiated stakeholder the value delivered by the program may be questionable. The outcomes are often enablers to organizational decision-making capability rather than absolute financial and non-financial metrics.

DSI DELIVERY FRAMEWORK

In this section we review the design principles used to build the framework, the framework, and processes of the framework.

The Design Principles

Considering the exploratory nature of DSIs, the framework will conform to the following design principles:

- End-to-end delivery of solution and value realization;
- Core and non-core domains identification;
- Use agile methods to support exploratory nature of DSIs instead of waterfall;
- Support both single team and scaled agile delivery of data science capability;
- Specify people, process and deliverables; and
- Agnostic to tools and technologies

The Framework

The proposed framework has five core domains and integrates PMI's The Standard for Program Management (Project Management Institute, 2017) for program management; Proscii Framework (Hiatt, 2006) for people change management; Scaled Agile (SAFe) (Scaled Agile, 2020) for solution delivery; DAMA's DMBoK (Earley, 2017) for data management; and CRISP-DM (Chapman et al., 2000) for data science processes as per *Figure 2* representing methods for the five domains.

As the domains are modular, it allows organizations to replace methods. For example, in program management domain PMI's methods (Project Management Institute, 2017) can be replaced with MSP (OGC, 2004). Furthermore, the framework is flexible to allow integration with other organizational domains such as Risk Management, Procurement Management, Asset Management, etc. Each process in the domain has been described in detail for the framework to be adopted by an organization implementing DSIs.

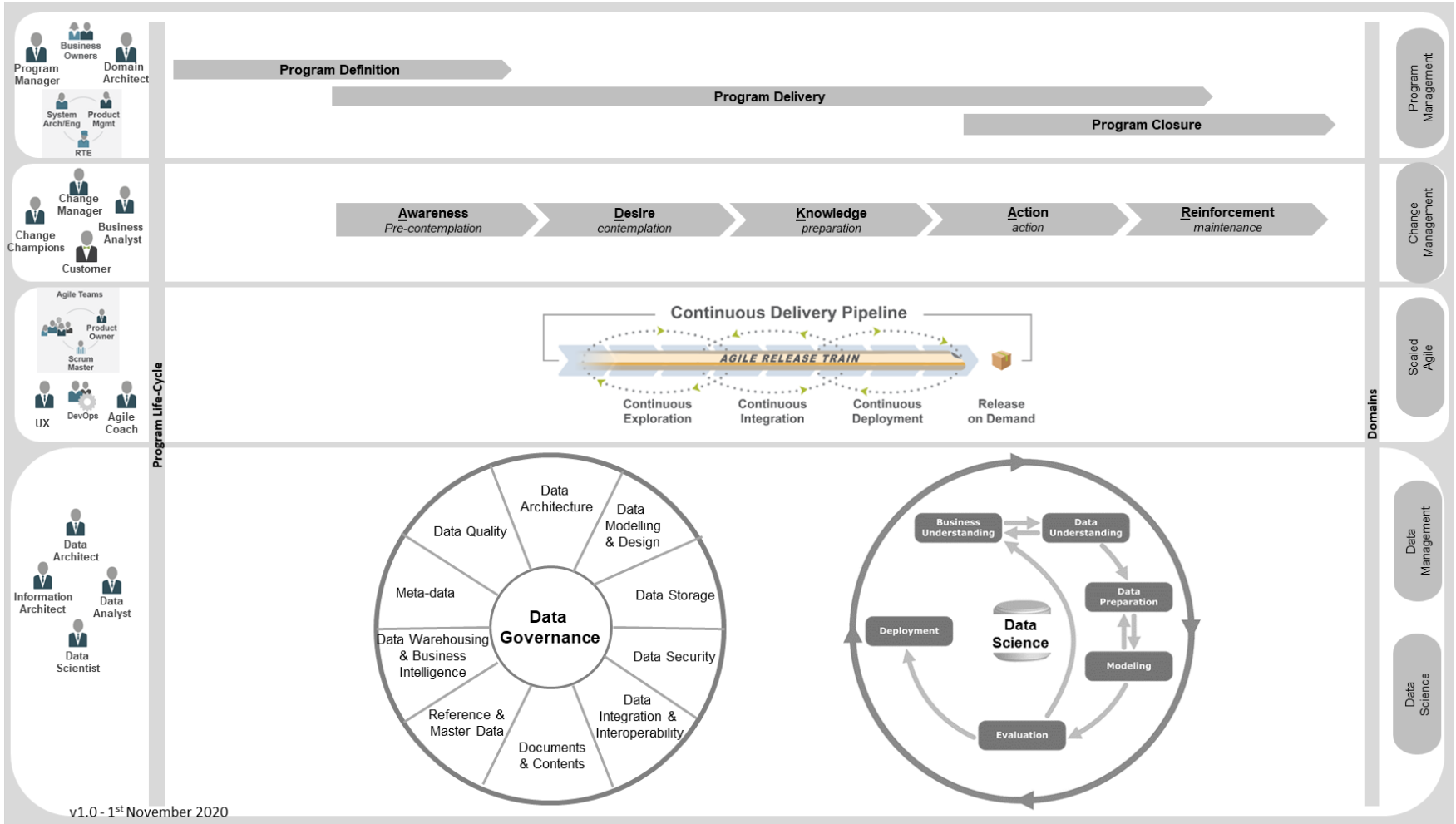

Figure 2. **DSI Delivery Framework**

~~DISCUSSION AND CONCLUSION~~

~~In this section, we review our research question: *"What design principles should be incorporated in a Data Science Initiative (DSI) Delivery Framework so that program managers can adopt a predictive path to realize value from such investments*?" and summarize our conclusion. We first start with the limitations of our research and then discuss implications on business managers and program managers from the proposed DSI Delivery Framework. We conclude that current literature does not adequately cover delivery of DSIs the proposed framework is a step in the right direction to assist practitioners.~~

~~Limitations and Implications of Research~~

~~This research has used six DSIs from one public sector organization as case study to develop the DSI Delivery Framework using five domains. Future research can include validating the framework with other public and private sector organizations delivering DSIs. Another aspect is that DSIs are a more recent phenomenon and sit in a rapidly evolving technology and delivery space. This has an impact on currency of the research work being done.~~

~~Implications for Practice~~

~~Limited availability of methods and standards in delivery of DSIs has caused the business managers and program managers to chart their own path and thus introduce inconsistency in how DSIs are treated and delivered in different organizations. With emergence of research such as this, it is expected that the standardization on DSIs will increase and provide guidance to the practitioners in efficient delivery of the DSIs.~~

Conclusion

Program Management for ICT-enabled Programs has rich literature and proven delivery frameworks which have matured over the past three decades (Project Management Institute, 2016, 2017; OGC, 2011). This paper makes a significant contribution to the theory and practice of the emerging field of data science.

The current Program Management literature does not adequately support delivery of DSIs and instead focuses on risk elimination and rapid delivery of business outcomes.

We propose a DSI Delivery Framework which has five core domains and integrates PMI's The Standard for Program Management (Project Management Institute, 2017) for program management; Proscii Framework (Hiatt, 2006) for people change management; Scaled Agile (SAFe) (Scaled Agile, 2020) for solution delivery; DAMA's DMBoK (Earley, 2017) for data management; and CRISP-DM (Chapman et al., 2000) for data Science processes as per *Figure 2* representing methods for the five domains.

~~We suggest additional research to fine tune the proposed DSI Delivery Framework which currently has been used for one public sector organization (Transport). The authors of the paper already intend to validate the trustworthiness and reliability of the framework through monitoring the use of framework at Transport as well as semi-structured interviews with practitioners and portfolio managers from other organizations. The framework proposed in this research will deliver a significant contribution to the body of knowledge for Program Management relevant to both literature and practitioners. Without this work, there will be more failed programs, dissatisfied sponsors and delay much needed investment in this emerging field as well as delay the benefits that will flow from harnessing the data and the nuggets in it.~~

REFERENCES

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 9*, 13.

Earley, S. E. (2017). DAMA-DMBOK: Data management body of knowledge. *Technics Publications*.

Ebert, C., & Paasivaara, M. (2017). Scaling Agile. *IEEE Software, 34(6), 98-103*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Hiatt, J. (2006). *ADKAR: a model for change in business, government, and our community*: Prosci.

Kotter, J. (2007). Leading change: Why transformation efforts fail. *Harvard business review, 86*, 97-103.

Kübler-Ross, E. (2009). *Death: The Final Stage*: Simon and Schuster.

Leffingwell, D. (2007). *Scaling software agility : best practices for large enterprisesAgile software development series* (pp. 1 online resource (xxix, 349 pages)).

Lenfle, S. (2008). Exploration and project management. *International Journal of Project Management, 26*(5), 469-478.

Lenfle, S. (2016). Floating in space? On the strangeness of exploratory projects. *Project Management Journal, 47*(2), 47-61.

Lorenzi, P., & Waterman, R. H. (1985). In Search of Excellence. *Journal of Organizational Behavior, 6*(3), 246-247.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization science, 2*(1), 71-87.

Mason, H., & Wiggins, C. (2010). A taxonomy of data science. *Dataists. com*(9).

Mathur, S. (2019). Unique Characteristics of Data-Analytics Programs. *PMI Australia Conference, Gold Coast*. Retrieved from https://www.pmiaustraliaconference.org.au/wp-content/uploads/2019/06/Sandeep-Mathur.pdf

OGC. (2004). *Managing successful programmes*: TSO.

Project Management Institute. (2017). *The standard for program management,*(Fourth edition. ed.).

Rollins, J. (2015). Foundational methodology for data science. *Domino Data Lab, Inc., Whitepaper*.

SAS Institute. (2009). *Getting Started with SAS Enterprise Miner 6.1*: SAS Institute.

Scaled Agile. (2020). SAFe for Lean Enterprises 5.0. Retrieved from https://www.scaledagileframework.com/

Scholz, R. W., & Tietje, O. (2002). *Embedded case study methods: Integrating quantitative and qualitative knowledge*: Sage.

Severtson, W. R. B., Franks, L., & Ericson, G. (2017). What is the team data science process? *Microsoft Azure*.

Yin, R. K. (2018). *Case study research and applications : design and methods* (Sixth edition. ed.). Thousand Oaks, California: SAGE Publications, Inc.