Project Leadership & Society

Unique characteristics of Data Science Initiatives: Implications for Program Management

--Manuscript Draft--

Manuscript Number:	
Full Title:	Unique characteristics of Data Science Initiatives: Implications for Program Management
Short Title:	Unique characteristics of Data Science Initiatives
Article Type:	Next Practices
Keywords:	Exploratory Projects; Data Analytics; Program Management; Change Management; Delivery Framework
Corresponding Author:	Sandeep Mathur, M Tech. Transport for New South Wales Roseville, NSW AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Transport for New South Wales
Corresponding Author's Secondary Institution:	
First Author:	Sandeep Mathur, M Tech.
First Author Secondary Information:	
Order of Authors:	Sandeep Mathur, M Tech.
	Shankar Sankaran
	Samuel MacAulay
	Ivor Tsang
Order of Authors Secondary Information:	
Abstract:	Data is increasingly ubiquitous in organizational life with a large number of investments in businesses using data in ideation stage to inform a business case through to delivery of core elements of scope and finally to track benefits realization. Data Science Initiatives (DSIs) have emerged as a popular mechanism for extracting value from data. However, the track record of these initiative has drawn substantial criticism from sponsors. For example, the success rate of delivering DSIs is not perceived as high with Gartner estimating that 85% of projects fail (Asay, 2017). This paper argues that one crucial reason for this failure is that DSIs have six unique characteristics that make traditional practices for conceptualizing and managing ICT-enabled programs ineffective. We build this argument by drawing on case studies of DSIs delivered over five years at the statutory body responsible for transport in the State of New South Wales in Australia. We conclude by explaining how managing DSIs as “exploratory projects” could improve the success rate of implementations.
Suggested Reviewers:	
Opposed Reviewers:	
Additional Information:	
Question	**Response**

4th December 2021

Professor Martina Huemann
Editor-in-Chief – Project Leadership and Society,
Vienna University of Economics and Business,
Vienna, Austria

Dear Martina

We wish to submit an original research article entitled “Unique Characteristics of Data Science Initiatives: Implications for Program Management” for consideration by Project Leadership and Society.

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

Data is increasingly ubiquitous in organizational life with many investments in businesses using data in ideation stage to inform a business case through to delivery of core elements of scope and finally to track benefits realization. In this paper, we target practitioners of the emerging field of data science and analyse characteristics of Data Science Initiatives (DSIs) across all phases of a Program Lifecycle using data from six DSIs from Transport for NSW as case studies. We further validate the characteristics through semi-structured interviews with practitioners from five organisations delivering DSIs. To address high degree of uncertainty in data, the paper also identifies conditions when DSIs should be managed as Exploratory projects instead of Exploitative projects to potentially improve the success rate of implementation.

We believe that this manuscript is appropriate for publication by Project Leadership and Society because the investment in DSIs continues to grow globally and by understanding the nuances this class of programs carry, there is an opportunity to improve the success rate of the investment. The decision-makers of the organization also need to be aware how DSIs can potentially be game-changers as is evident from case studies from Transport for NSW which is pivoting from a traditional engineering to technology and data-driven business to improve services to citizens of New South Wales in Australia. These DSIs also deliver social benefits to the society as the case studies show.

We have no conflicts of interest to disclose.

Please address all correspondence concerning this manuscript to me at sandeep@mathur.com.au.

Thank you for your consideration of this manuscript.

Sincerely,

Sandeep Mathur, PgMP, FAICD, FACS

Highlights

- Large number of investments in businesses are using data from ideation stage through to delivery and benefits tracking
- Data Science Initiatives (DSI) help extract value from data
- Success rate of delivering DSIs is not perceived high with Gartner estimating that 85% of projects fail
- DSIs have unique characteristics which pose challenges when using traditional processes for delivery
- There are conditions where DSIs should be managed as Exploratory projects instead of Exploitative projects

Conflicts of Interest Statement

Manuscript title: *Unique characteristics of Data Science Initiatives (DSI)*

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

Sandeep Mathur

Shankar Sankaran

Samuel MacAulay

Ivor Tsang

The authors whose names are listed immediately below report the following details of affiliation or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript. Please specify the nature of the conflict on a separate sheet of paper if the space below is inadequate.

Author names:

Sandeep Mathur – Employee of Transport for NSW

This statement is signed by all the authors to indicate agreement that the above information is true and correct:

Author's name	Author's signature	Date
Sandeep Mathur		17h October 2021
Shankar Sankaran	Shankar Sankaran	17-Oct -2021
Samuel MacAulay		3/12/2021
Ivor Tsang		3/12/2021

Unique Characteristics of Data Science Initiatives: Implications for Program Management

Sandeep Mathur,[1] Shankar Sankaran,[2] Samuel MacAulay[3] and Ivor Tsang[4]

[1]Director – Active Transport, Transport for NSW, 22-44 Ennis Road, Milsons Point, NSW 2061, Australia. Email: Sandeep.Mathur@transport.nsw.gov.au

[2]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

[3]Senior Lecturer, Strategy and Innovation, The University of Queensland, Brisbane, QLD 4072, Australia. Email: s.macaulay@uq.edu.au

[4]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

Corresponding Author:

Sandeep Mathur

Director – Active Transport, Transport for NSW

22-44 Ennis Road,

Milsons Point, NSW 2061, Australia.

Emails: Sandeep.Mathur@transport.nsw.gov.au

Running Title: Unique Characteristics of Data Science Initiatives: Implications for Program Management

Word Count: 10498

Unique Characteristics of Data Science Initiatives: Implications for Program Management

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.

Abstract

Data is increasingly ubiquitous in organizational life with many investments in businesses using data in ideation stage to inform a business case through to delivery of core elements of scope and finally to track benefits realization. Data Science Initiatives (DSIs) have emerged as a popular mechanism for extracting value from data. However, the track record of these initiative has drawn substantial criticism from sponsors. For example, the success rate of delivering DSIs is not perceived as high with Gartner estimating that 85% of projects fail (Asay, 2017). This paper argues that one crucial reason for this failure is that DSIs have six unique characteristics that make traditional practices for conceptualizing and managing ICT-enabled programs ineffective. We build this argument by drawing on case studies of DSIs delivered over five years at the statutory body responsible for transport in the state of New South Wales in Australia. We conclude by explaining how managing DSIs as "exploratory projects" could improve the success rate of implementations.

Keywords: Exploratory Projects; Data Analytics; Program Management; Change Management; Delivery Framework

1. Introduction

Data Science Initiatives[1] (DSIs) have unique challenges that make the application of traditional program management techniques problematic. These challenges arise primarily due to uncertainty they carry in data being ingested which has a cascading impact on scope, schedule and ultimately value creation.

In all DSIs, data from various known and unknown sources is ingested into a data store and transformed and insights are generated using this transformed data. At the commencement of any DSI, the quality and structure of the data being ingested is relatively unknown. This suggests that there are occasions when DSIs need to be managed as Exploratory Projects due to limited 'information-before-action'. Lenfle (2008) describes Exploratory Projects as those for which neither technologies nor customer requirements are known at the start of the project. The uncertainty in DSIs makes it difficult to manage them as Exploitative Projects which focus on optimizing cost-quality-time triple constraints to deliver new products and services (Lenfle, 2008). The fundamental tension between exploitation of old certainties and exploration of new possibilities identified by March (1991) is relevant to DSIs.

Furthermore, the sequential and pre-defined Waterfall approaches to program management adopted by peak project and program management bodies to deliver DSIs set up structural tensions between business case development, program design, delivery and benefits realizations that undermine coherent governance across the investment life cycle. A known scope and benefits profile and old certainties allows an exploitative project to move sequentially from one gate to another seamlessly. Whereas the uncertainties in DSIs make them unsuitable to follow the sequential software development approach of analysis, design, development, testing and deployment and more aligned to iterative, innovative, and exploratory projects.

In this paper, we identify unique characteristics of DSIs that distinguish them from typical ICT-enabled programs in order to help scholars and practitioners better understand the when and why Waterfall approaches are likely to fail and what alternative might enable them to deliver more successful business outcomes. We draw on evidence from in-depth case studies of six DSIs delivered over five years at Transport for NSW (Transport). The external validity of these findings

[1] We use the term Data Science Initiative (DSI) to describe investments in Data Analytics, Business Intelligence and Data Science including Machine Learning and Artificial Intelligence technologies.

was then probed using semi-structured interviews with practitioners from five diverse organizations who are involved in delivery of DSIs. Transport is a state government enterprise responsible for delivering safe, integrated and efficient transport systems to the people of New South Wales in Australia. The organization is in early stages of delivering $72.2 billion worth of investment in transport infrastructure as outlined in its Future Transport 2056 Strategy (Transport for NSW, 2018). More than 300 initiatives have been identified for delivery in the first 10 years underpinned by data-driven technology roadmap. Most initiatives have a significant technology component including that in data and data-analytics. Transport generates significant amount of data (approximately 1TB per day) from Bus, Ferry, Light Rail, Metro, Heavy Rail and Active Transport modes every day. As big-data technology matured in past few years, it has given Transport an ability to store, transform and use the data generated from Internet of Things (IoT) devices attached to Transport assets, something it was unable to do effectively earlier. This new capability is enabling the organization to deploy data-driven decisioning both in real-time, such as managing congestion, and more traditional monthly operator contract performance.

The case studies of the six DSIs reveal a pattern of uncertainty caused by data being ingested. Key traits identified were ambiguity around goals, interdependencies, lack of skills required to deliver them and ongoing nature of product development (Mathur, 2019). Analysis of six DSIs case studies – benefits in business case vs benefits delivered (Table 1); stakeholder engagement (Table 2); slippage in schedule (*Figure 8*, *Figure 9* and Table 3); risks, dependencies and constraints (Table 5); uncertainty in requirements (Table 7); benefits mapping and dependencies (*Figure 12*); and roles & responsibilities (Table 8) shows how we arrived at those characteristics and is further elaborated in this paper. Our findings paint a picture of DSI delivery that is much more in tune with the world of exploratory projects, but also hint at the possibility that, with products and capabilities in the industry maturing, we might soon be at a transition point where the nature of DSI delivery might shift from exploration to exploitation.

We conclude by arguing that practitioners need to understand the exploratory characteristics when planning and delivering the DSIs and move away from traditional approaches which fail to account for the uncertainty and ambiguity that currently shape the delivery of DSIs.

2. Literature Review

Uncertainty in data and ambiguity of goals and means lead us to investigate how innovation, exploration and uncertainty in projects is handled in literature. We start with a review of Exploratory and Exploitative projects to see where DSIs fit into the spectrum. As DSIs are generally constructed and delivered as programs, we review Program Management literature which has matured over last two decades. We then move to the people side of initiatives and discuss Change Management as effective management of change is integral to the success of all ICT programs and DSIs. We argue Agile software development methods are more suitable for addressing the exploratory nature of DSIs and commence with a review of Scaled Agile frameworks to support large scale agile delivery. As DSIs focus on ingesting, transforming and visualizing multiple datasets, we introduce data to this paper by exploring Data Management, then close this section by discussing Data Mining and Data Science Processes.

2.1 Exploration vs Exploitation

While exploration focusses on innovation, discovery, experimentation and risk taking; exploitation focusses on production, efficiency, implementation and execution. Balancing exploration and exploitation is key to organizational success and survival (March, 1991). Projects often act as the organizational vehicle for delivering change on these dimensions (Brady & Davies, 2004). Exploitative projects focus on optimizing cost-quality-time triple constraints to deliver new products and services whereas in exploratory projects neither the goals nor the means of attaining them are clearly defined from the outset (Lenfle, 2008). Therefore a traditional view of project management as the accomplishment of a clearly defined goal with cost-quality-time triple constraints is unlikely to fit neatly with the logic of innovation that is first and foremost characterized by discovery (Van de Ven, Polley, Garud, & Venkataraman, 1999) and unforeseeable uncertainty (Loch, DeMeyer, & Pich, 2006).

2.2 Program Management

We see DSIs being typically implemented as a Program on a continuous spectrum rather than a single one-off project and focus on Program Management rather than Project Management processes. *A program is defined as related projects, subsidiary programs, and program activities managed in a coordinated manner to obtain benefits not available from managing them individually* (Project Management Institute, 2017). *Program Management is defined as the application of knowledge, skills, and principles to a program to achieve the program objectives*

and to obtain benefits and control not available by managing program components individually (Project Management Institute, 2017). A review of Program Life Cycle reveals some gaps in using it for delivery of DSIs (Table 9). Project Management Institute's standard defines program life cycle as containing three phases - Program Definition, Program Delivery and Program Closure of a program as per *Figure 1*.

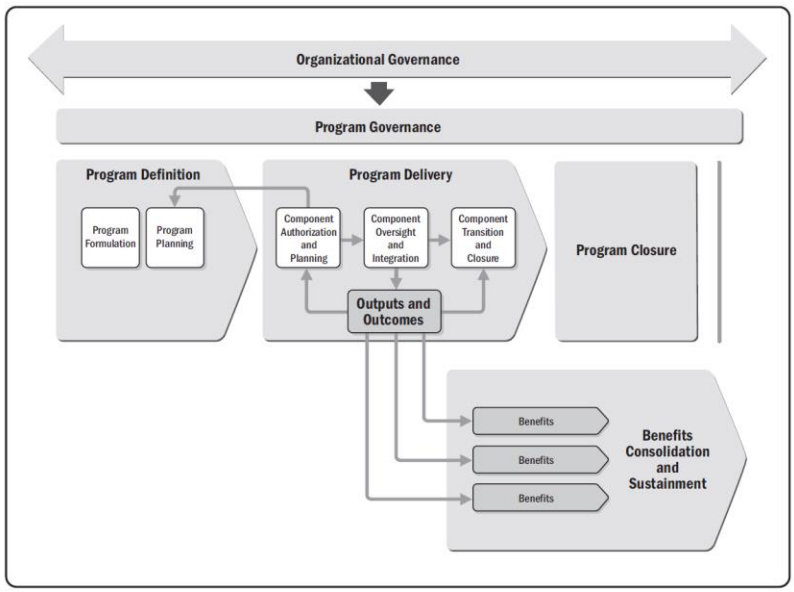


Figure 1. Program Life Cycle (Project Management Institute, 2017)

Managing Successful Programmes (MSP) is another program management framework whereby large, complex change can be broken down into manageable, inter-related projects (Axelos, 2020). The MSP framework is based on three core concepts: MSP Principles; MSP Governance Themes; and MSP Transformational Flow as per *Figure 2*.

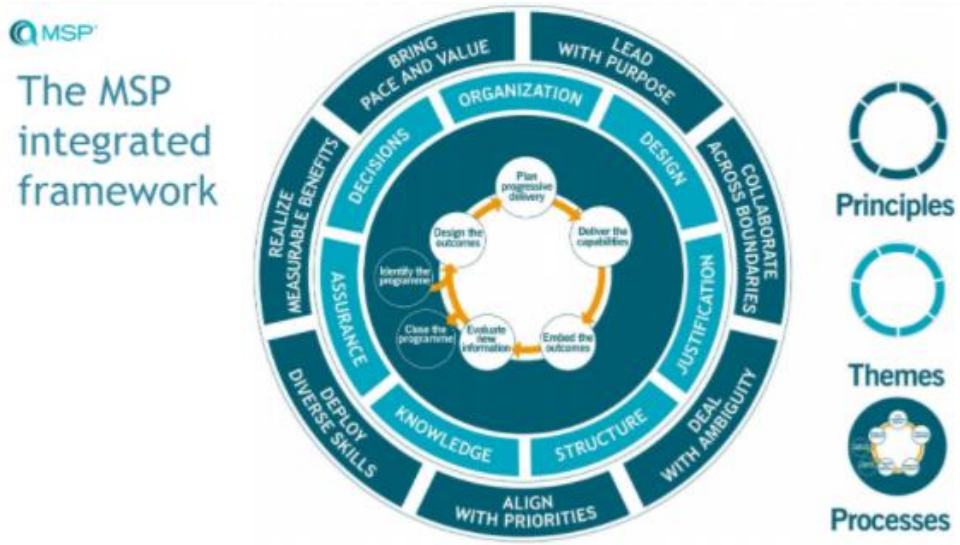


Figure 2. The MSP integrated framework (Axelos, 2020)

While our case studies used PMI's standard, we see both PMI and Axelos standards candidates to be used for DSIs with PMI standards as being principle-based while the Axelos standards provide more detailed guidance on program management.

2.3 Change Management

Value realization for any program occurs when the product and service created is adopted by users successfully. This is usually achieved when changes required to adopt the outputs from the projects create desired outcomes. 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 widely advocated change management models. Of the four Change Management models, the models proposed by ADKAR and Kotter continue to be referenced and widely used by practitioners when delivering ICT change programs. During organizational change such as restructuring, mergers and acquisitions, the Kübler-Ross model is often preferred. All the four Change Management models are popular because of the simplicity in understanding and implementing them. We see effective use of a change management model in DSI as essential to value realization as each dataset brings in additional insight, complexity and need for change which needs to be embedded in the organization.

2.4 Scaled Agile

A need for large projects which are often globally distributed with teams requiring collaboration and coordination has led to the emergence and use of scaled-agile frameworks such as Scaled Agile Framework (SAFe), Large-Scale Scrum (LeSS) and Lean Scalable Agility for Engineering (LeanSAFE). (Ebert & Paasivaara, 2017; Leffingwell, 2007). In the context of DSIs, we also see the relevance of scaling is high as often multiple geographically spread teams within an organization are involved in delivering its outcomes.

2.5 Data Management

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). *Figure 3* defines the eleven data management knowledge areas with data governance at the center of wheel and other knowledge areas necessary to be implemented at different times during the delivery depending upon the requirements of the organization. The Environmental Factors hexagon in *Figure 4* shows the relationship between people, process, and technology with goals and principles at the center for people to execute activities and use the tools required for successful data management. We see eleven knowledge areas of data management as foundational elements for the DSI delivery team unlike other ICT-enabled programs.

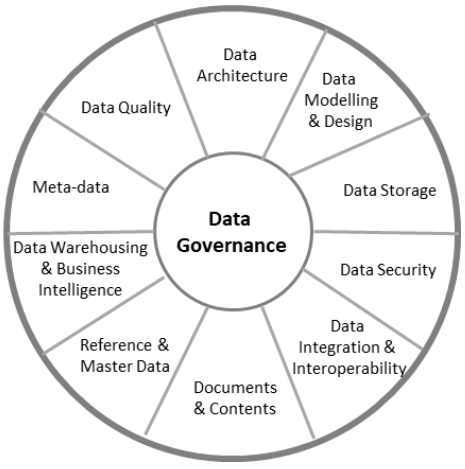


Figure 3. Data Management Framework (Earley, 2017)

Figure 4. Environmental Factors Hexagon (Earley, 2017)

2.6 Data Science Processes

Delivery of DSIs requires good understanding of Data Mining and Data Science delivery processes. Knowledge Discovery in Databases (KDD) model (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), Cross-Industry Standard Process for Data Mining (CRISP-DM) as per *Figure 5* (Chapman et al., 2000), Sample, Explore, Modify, Model and Assess (SEMMA) (SAS Institute, 2009), Obtain, Scrub, Explore, Model and Interpret (OSEMN) (Mason & Wiggins, 2010), Team Data Science Process (TDSP) (Severtson, Franks, & Ericson, 2017), Foundational Methodology for Data Science (FMDS) (Rollins, 2015) are some of the widely used frameworks in delivery of DSIs.

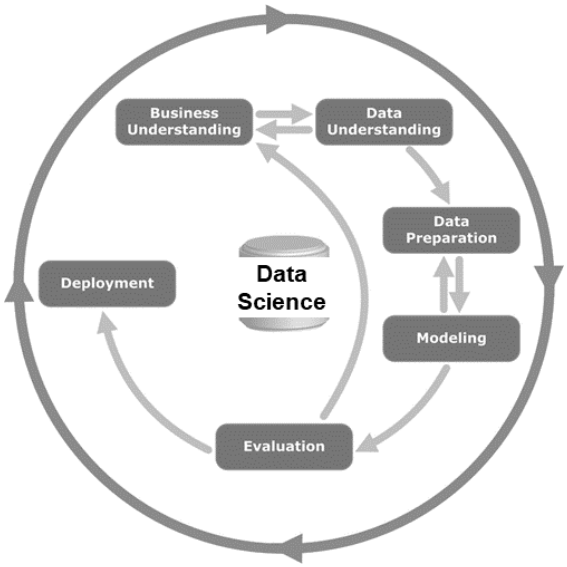


Figure 5. CRISP-DM Data Science Processes (Chapman et al., 2000)

Foroughi and Luksch (2018) compared KDD, CRISP-DM, FMDS and TDSP processes against four common iteratives stages of Problem Definition / Formulation; Data Gathering; Data Modelling and Data Production as per *Figure 6*. They found that KDD process does not cover the business understanding and deployment phases of CRISP-DM methodology. CRISP-DM does not have the analytic approach; identification of suitable data collection strategy and data resources; and feedback phases of FMDS. While FMDS and TDSP are very similar, the detailed stages of FMDS could be more useful for a wide range of projects but TDSP uses a specific set of Microsoft tools and infrastructure to deliver intelligent applications by deploying machine learning or AI models.

KDD Process	CRISP-DM	FMDS	TDSP
-	Business Understanding	Business Understanding	Business Understanding
	-	Analytic Approach	
	-	Data Requirements	
Selection	Data Understanding	Data Collection	Data Acquisition and Understanding
Pre Processing		Data Understanding	
Transformation	Data Preparation	Data Preparation	
Data Mining	Modelling	Modelling	Modelling
Interpretation/ Evaluation	Evaluation	Evaluation	
-	Deployment	Deployment	Deployment
-	-	Feedback	Customer Acceptance

Figure 6. Summary of Data Science methodologies and their phases (Foroughi & Luksch, 2018)

2.7 Summary of Literature Review

In the previous section, we investigated exploratory and exploitative projects to handle innovation, exploration and uncertainty in projects. We then reviewed program, data and data science domains which are essential to understanding characteristics of DSIs. We reviewed Program Management frameworks from PMI and Axelos as DSIs are not one-off projects and require management as programs. We reviewed Change Management to ensure that investment in DSI delivers value to the organization. We reviewed several Scaled Agile frameworks in order to assess their suitability for both smaller and large scale DSI implementations. An understanding of Data Management is essential to all DSIs and we reviewed DAMA's DMBoK. We then reviewed several Data Mining and Data Science methods used in last twenty-five years to find that FMDS and TDSP from IBM and Microsoft respectively are built upon the solid foundations from KDD and CRISP-DM.

To summarize, we defined term Data Science Initiative (DSI) to include investments in Data Analytics, Business Intelligence and Data Science including Machine Learning and Artificial Intelligence technologies. We identified that this class of programs has challenges in delivery due to uncertainty they carry in data which has a cascading impact on scope, schedule and ultimately value creation. To deal with uncertainty and innovation, we investigated what literature on exploratory project offers. We further explored Program Management, Change Management, Data Management and Data Science Process domains to see how they influenced delivery of DSIs. This view of the literature motivated us to ask the following research question:

"What unique characteristics cause DSIs to face challenges delivering envisaged value when using traditional processes for managing ICT-enabled programs?"

3. Research Setting and Methods

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 continues to deliver DSIs. This gave him good access to data to conduct the case studies.

A multi-methods approach (Hunter & Brewer, 2015; Straits & Singleton, 2018) combining case studies and semi-structured interviews with practitioners was used. To obtain granularity of program life-cycle as well as variation for analytical comparisons, an embedded case design was selected (Yin, 2018) to analyze six DSIs in Transport, each of which provided a unique scope and opportunity to understand characteristics of DSIs. Our interest was to understand characteristics of DSIs as experienced by the organization's participants themselves and identify uniqueness with this class of initiatives to bring in improvements within the organization. While we collected data for all six DSIs, the pair of first and sixth DSI serve as extreme cases reflecting both chronology and the maturity of Transport in delivering DSIs. We saw specific patterns emerging as we progressed through DSIs and by the time we reached the sixth, we saw stability in patterns and consistency in characteristics. The author of this paper continues to deliver more DSIs further validating the findings but for the purpose of this paper, we stop at the sixth DSI. The number of cases selected follows the recommendation made by Small (2009) and Yin (2018) with the extreme cases expected to contribute to theoretical replication (predicting contrasting results) and the semi-structured interviews with practitioners providing literal replication (finding similarities) to setting up a multiple case study design.

Using an interpretive research tradition associated with case-studies, ontological and epistemological assumptions on DSI characteristics emerged which were externally validated with practitioners from five organizations delivering DSIs using semi-structured interviews. Informed consent was sought from interviewees by carefully explaining the study and its aims, as well as their ethical rights during interviews. Ethics approval was granted by the home university of the researchers. The interviews used open-ended questions to gain lived experienced of interviewees. Interpretive approach (Sandberg, 2005) to justify knowledge produced was adopted by analyzing interview transcripts leading to coherent interpretations of DSI characteristics. Each of the

characteristics was either confirmed or altered as per responses to the questions. A new characteristic emerged after the second interview and was further validated by remaining interviewees. Follow-up questions were asked especially when interviewees mentioned DSIs they delivered and associated ambiguity and complexity on some of the characteristics.

Iterating among in-depth analysis of each case, comparisons across cases, and connections to the literature (Eisenhardt, 1989), we reviewed diverse stakeholder groups as well as formal and informal interactions as per Table 2 and how they influenced the delivery of DSIs that led us to further analysis and theorizing (Agar, 1986). For majority of our stakeholders in Transport, this was their first experience working on or interacting with a DSI. It was a steep learning curve for them and our interactions with them not only educated them but also informed us of the nuances of DSIs. We also used supplementary evidence such as documents and participant observations from each DSI. The evidence included Program Management artefacts such as Program Management Plan, Risk Register, Schedule and Communications, Minutes of governance and working group meetings and Stories and Burndown Charts in Jira and Confluence software. The observations took place throughout the delivery of six DSIs specially in ceremonies such Daily Standup; Sprint Planning; Backlog Grooming and specially Sprint Retro which gave a good reflection on what the team felt at the end of each two-weekly Sprint. Our research question emerged over time by reviewing the challenges faced by the delivery teams backed by evidence from the case studies.

The primary author was the Program Manager of the six DSIs chosen as case studies which were delivered between January 2017 to December 2020 and thus brings in-depth insights of the program life cycle.

The practitioners for semi-structured interviews represented five Australian-based organizations covered DSIs delivered in Australia, New Zealand and USA. While our sample size was small, the five interviews helped us to validate the trustworthiness and reliability of the DSI characteristics by probing whether geographical or organizational boundaries might induce variation in the findings.

3.1 Research Setting

Our research was situated within Operational Systems division of Transport for NSW (Transport), a state government enterprise that leads the development of safe, integrated and efficient transport systems for the people of NSW in Australia. Transport's functions include

transport planning, strategy, policy, procurement and other non-service delivery functions across all modes of transport - roads, rail, ferries, light rail, metro and point to point. The organization is in early stages of executing $72.1 billion worth of Future Transport 2056 Services and Infrastructure Plans (SIPs) which has set out more than 300 initiatives to be delivered in the first 10 years of the 40-year vision underpinned by data-driven technology roadmap (Transport for NSW, 2021). Most initiatives have a significant technology component including data and data-analytics.

Transport generates significant amount of data from Bus, Ferry, Light Rail, Metro, Heavy Rail and Active Transport (Walking and Cycling) modes every day. The real-time data includes timetable information, position of every transport vehicle and predicted time of arrival at the next transit stop which is shared on all passenger Apps; event information from sensors such as doors opening & closing information, temperature, speed, number of passengers etc.; ticketing information such as Opal cards (smartcard tickets) being tapped on and off at the gates, checking of tickets; monitoring the use of cycleways and pedestrian walkways through to crime and incident information on the transport network. While the realization of using the data for improving customer service in their multi-modal journeys and managing performance of Transport Operators was always there - lack of technology, skills and investment prevented Transport from mining and capturing value from the data it was generating. For example, Transport continued to rely on Operators to tell us through their monthly reports if services delivered met contractual performance requirements. This realization has further led to Future Transport Strategy 2056 to include a Data ecosystem within Transport to provide continuous improvements on its asset performance and improved customer and operational information. The journey of primary author of this paper mirrors that of Transport as an organization. When he joined Transport in late 2016 as a Program Manager, he had significant experience in delivering ICT Transformation Programs using both Waterfall and Agile methods but none in delivering data related Programs. This paper tracks the delivery of DSIs, the challenges he faced in understanding some of the unique characteristics of DSIs, his ability to influence near real-time data-driven decision-making in Transport and growth in maturity of Transport to harness the value of data underpinned by his own ability to deliver DSIs.

The research method uses participant-observation technique and multiple case studies over full program life cycle covering a period of five years collecting DSI data. Out of the six potential

sources of data - documentation, archival records, interviews, direct observations, participant-observation, and physical artifacts; five have been used except for interviews.

4. Data Collection and Analysis

The six DSIs chosen as case studies represent contemporary phenomenon in depth and within its real-world context (Yin, 2018) at Transport that was particularly useful for our research question because the organization needs to better understand the unique characteristics of DSIs. Table 1 provides a summary of the six DSIs.

Table 1

Summary of six Transport DSIs

Program	Description	Period	Business Case Benefits	Benefits Delivered	Budget (AUD)
Vanguard	Consolidate and disseminate data & information to contribute to a public transport network where customers and staff feel safe & always travel with a valid ticket.	Jan 2016 - Mar 2019	Increase revenue through improved fare compliance & improve Customer Satisfaction & Security outcomes.	Delivered dashboards to paint a picture of fare evasion & security by ingesting six of possible twenty-one data sources. Also, laid the foundation of Data Management & DSI delivery	$5.14m
Ferry	Implement evidence-based Ferry Contract Management & improved customer experience.	Apr 2018 - Dec 2020	Deliver five dashboards to monitor operator performance. Also, deliver Microsoft Azure-based Operational Data Lake (ODL) platform to current and future needs.	Program has been delivered with a new Operator On-Boarded on TfNSW's systems. Performance Reporting was also delivered.	$4.8m
CTABS	Enable data analytics and verification of Provider self-reporting.	Oct 2017 - Mar 2018	Obtain visibility of community transport services in NSW; Understand the customers (Who/How/Why/Where); Understand the trips & travel patterns; Assess service quality; Investigate opportunities to improve service delivery; (vi) Determine if CTABS has resulted in operational efficiencies; and (vii) Assist in managing contracts	Project terminated as both solution and benefits could not be delivered.	$289k

PTIPS Analytics	Conduct a proof of concept of Azure big-data platform by using PTIPS (Public Transport & Information Priority System) which supports operational requirements of all public transport buses in Metropolitan NSW.	Apr 2019 - Jun 2019	Validate analytics solution using Azure Operational Data Lake; Provide self-service capability to Bus Contract Managers & Operators with minimum six months of PTIPS data; and Determine the Operational Expenditure (OPEX) requirements.	All benefits delivered including ten complex PowerBI dashboards with high stakeholder satisfaction	$357k
Light Rail Priority	Provide priority to Light Rail at traffic intersections shared with other road users.	Feb 2019 - Mar 2020	Support optimising Sydney Light Rail journey time; Provide light rail, enhanced level 3 priority at intersections; Increase visibility of Light Rail vehicles to TMC, RMS and SCATS; Support decrease in Sydney congestion; and Implement a hardware free solution for all SCATS intersections.	Technology solution has been delivered but other business benefits are dependent on other systems and cannot be directly attributed to this project. This project is an enabler project.	$1.49m
MPR	Ensure data management and architectural consistency of Operational Data Lake (ODL) across multiple performance reporting business cases.	Jan 2020 - Jun 2021	Delivery of consistent ODL architecture and Bus (Metro), Bus (Regional), Ferry, Light Rail, Sydney Metro, Community Transport, OnDemand and Zero Emission Buses Performance Reporting.	Program was delivered consisting of ten projects. Performance dashboards delivered are already being used by Contract Management teams to identify and resolve operational issues.	$4.4m

The chronology of six DSIs has been bracketed into three stages: Exploration, Transition and Exploitation that Transport went through as the six DSIs were delivered. When the author commenced delivering his first DSI (Vanguard) as traditional ICT Program, he faced challenges in managing the schedule The planned milestones were not met. In hindsight, the organization was not aware of the exploratory nature of DSIs. However, as we progressed on Vanguard, we started acknowledging the unique characteristics and making changes to the delivery processes. At macro level, we map this initial stage to Transport's "exploration" stage. Transport's "Transition" stage maps to the organization accepting the uniqueness of DSIs, adapting to delivery processes, and building skills to deliver DSIs successfully. "Exploitation" stage refers to a mature state where organization accepts that datasets come with uncertainty; agile methods are practiced and management accepts DSI business cases without measurable benefits.

In context of Transport, the stages can be roughly mapped to DSIs delivery timeline of Exploration stage mapping to Vanguard & CTABS; Transition stage mapping to Ferry, Light Rail Priority and PTIPS Analytics; and Exploitation stage mapping to MPR. *Figure 7* 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 for the uncertainty and engage the stakeholders effectively. While each of the six DSIs supported different parts of Transport with different requirements, 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 Exploitation stage.

Figure 7 provides an overview of the six DSIs used as case studies which includes the delivery timeline, complexity, and project-stages.

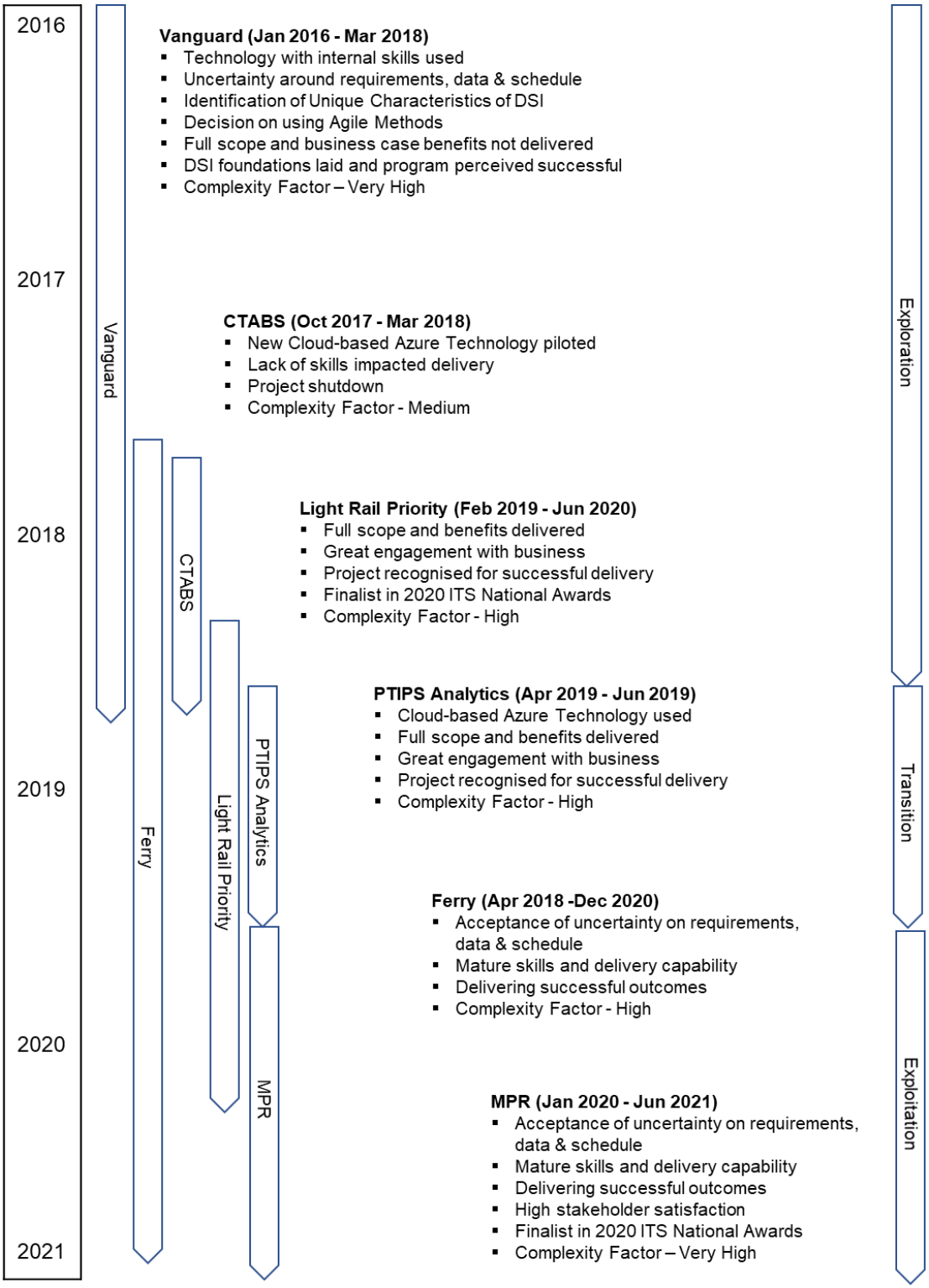


Figure 7. Overview of six Transport for NSW DSIs

The delivery of DSIs required continuous engagement with various stakeholder groups through formal and informal interactions. Table 2 provides a summary of stakeholders across along with number of governance meetings which were used to inform data for this paper.

Table 2

Stakeholder Summary of six DSIs

	No of Stakeholders[a]							
Program	Total	Management	Governance	Support	Technical	Business	Governance Meetings Participated[b]	Working Group Meetings Participated[c]
Vanguard	67	7	15	5	32	8	36	702
Ferry	62	5	13	5	21	18	15	310
CTABS	17	7	0	5	5	0	5	35
PTIPS Analytics	34	7	5	5	12	5	4	152
Light Rail Priority	83	16	6	5	49	7	13	418
MPR	68	4	29	6	21	4	24	427

(a) Stakeholders include Program Steering Committee, Operational Systems Management, Business Users, Delivery & Support Team
(b) Working Groups include Scrum Team, Architecture, Change, Data Management and Program/Project Working Groups
(c) Governance Meetings include Program Board and Project Control Group Meetings

Case studies from Transport as per Table 1 and Figure 7 demonstrated our inability to baseline the scope and schedule. As per Figure 5, each DSI went through steps of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment using one or more datasets. Vanguard which was the first DSI, ingested only six out of possible twenty-one data-sets. Ingestion of two key datasets ORPA (tickets scanned) and ePIN (fines issued) was dependent on a third-party vendor which took more than eighteen months to sign contracts, schedule and deliver the datasets. Even when the initial data arrived, the team found integrity issues with data such existence of fines issued record without mandatory tickets scanned record which required going back to the vendor to remediate continuously impacting delivery schedule. As we progressed the delivery of our first DSI Vanguard in order to provision for uncertainty, not only did we adopt agile ways of working but created a high-level flexible schedule which acknowledged the complexity of data. The shift in approach to scheduling is evident from a Waterfall-Agile Hybrid timeline in the Vanguard business case (Figure 8) to a data-centric timeline in January 2018 (Figure 9). Figure 8 shows two sequential activities of “Write Agile Stories” where requirements were captured and “Build + System Testing Iterations” where system is built with overall schedule taking 16 months to deliver. In reality, Vanguard took 27 months to deliver (Figure 7) and *Figure 9* shows a portion of the overall schedule (January 2018 – October 2018) reflecting how the schedule is broken down by various data-sets (ORPA, STA, SRS, ePIN, FCS) showing data-centric approach.

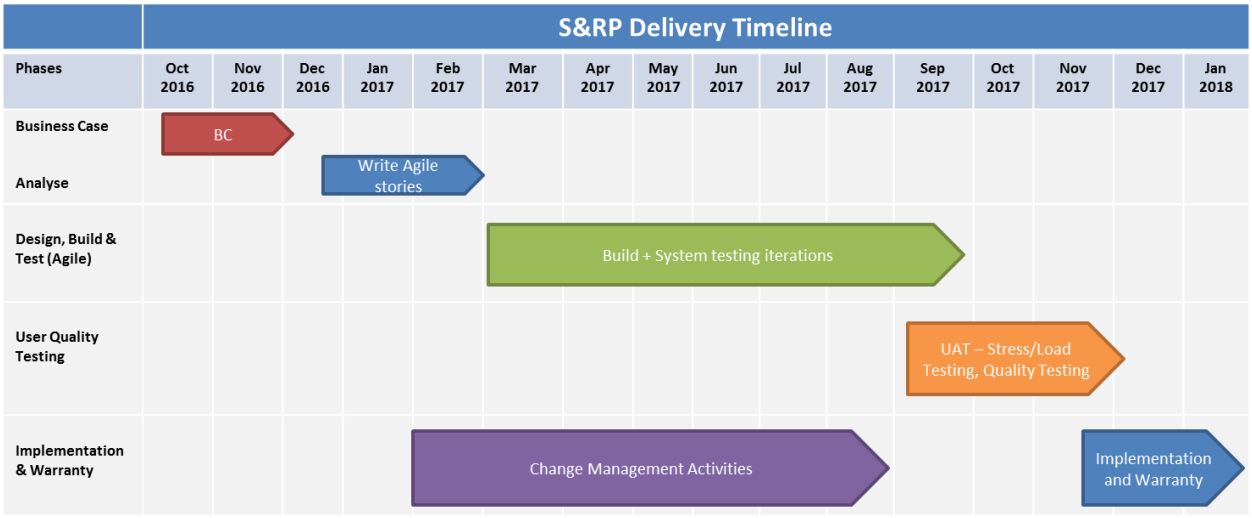
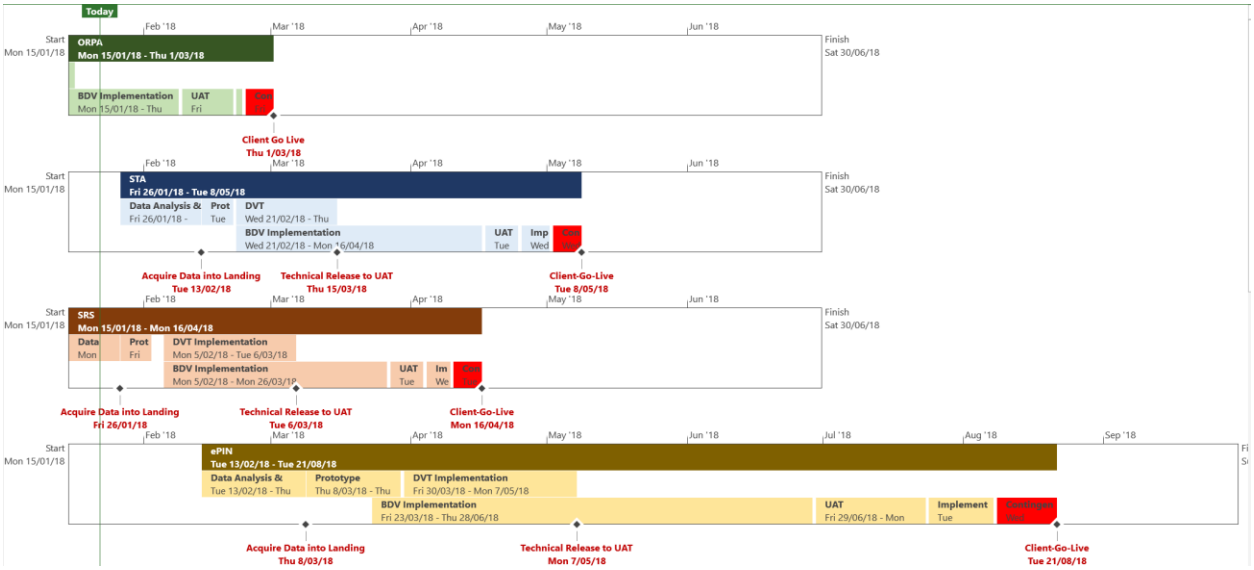


Figure 8. Vanguard Delivery Timeline in Business Case (Oct 2016)



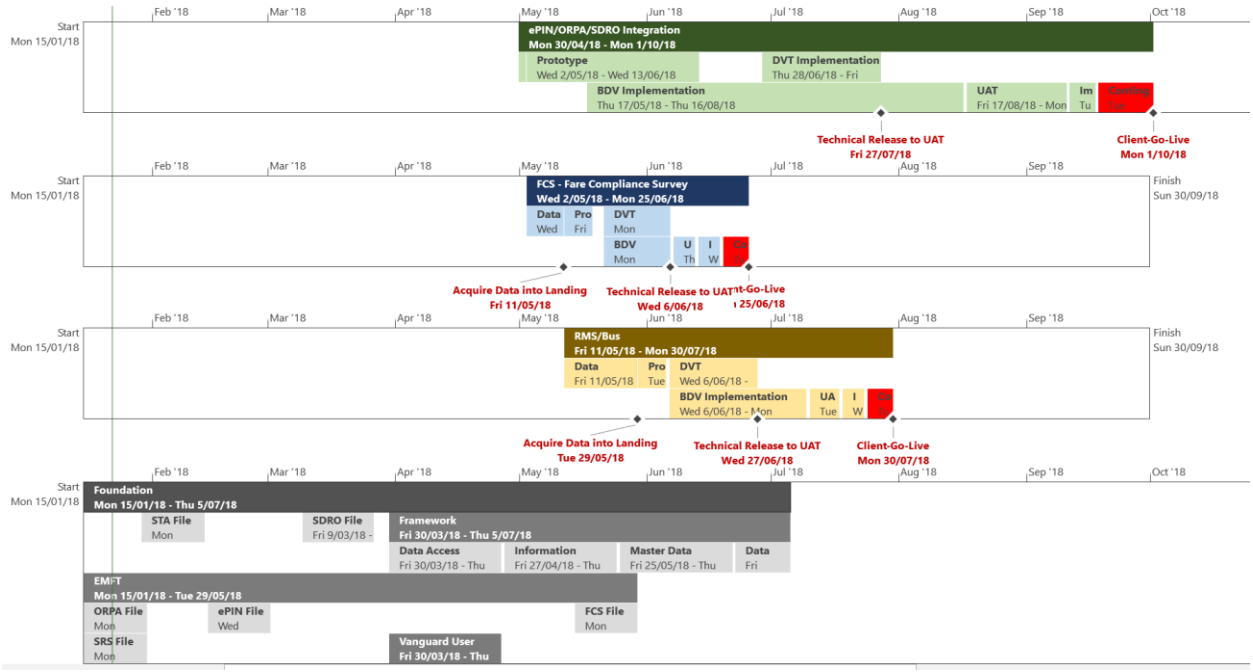


Figure 9. Vanguard Delivery Timeline using Data-Centric approach (Jan 2018)

The uncertainty is further highlighted by comparing the planned vs delivered dates of various datasets some of which were delivered almost a year late on a two and a half year Vanguard program as per Table 3. Specifically, ORPA dataset was delivered 6-months and ePIN more than a year late. This had a cascading impact on fines collection data from RNSW which was further delayed by 7-months.

Table 3

Vanguard - Planned vs Actual Delivery of data-sets

Data-Set	Description	Planned	Delivered	Variance (Days)
BOCSAR	Crime data on Transport network	1/05/2018	17/11/2017	-165
STA Incidents	Incident data from State Transit Authority	16/10/2018	18/07/2018	-90
SRS Incidents	Incident data from State Rail Services	16/10/2018	27/03/2019	162
ORPA	Tickets checked data from Opal Revenue Protection App	31/10/2017	18/05/2018	199
ePIN	Electronic penalty infringement notice data	31/10/2017	8/11/2018	373
RNSW	Fines collection data from Revenue NSW	6/02/2018	26/09/2018	232

Our sixth DSI, Multi-Modal Performance Reporting (MPR) Program commenced in January 2020 to "*ensure data management and architectural consistency of Operational Data Lake across multiple performance reporting business cases"*. Initial scope included Bus (Metro), Ferry, Light Rail and Bus (Regional) Performance Reporting and scope was extended in June 2020 to include Sydney Metro Performance Reporting as well as Data Ingestion and Self-Service projects. Our organization and teams had transitioned to Exploitation stage by the time MPR

delivery took place. Learnings from previous DSIs were applied in identifying core roles required for successful DSI delivery. These include Cloud Architect, Information Architect, Data Architect, Scrum Master and Change Manager roles at the Program level to bring in governance, architectural and change management consistency and each project having Product Owner, Business Analyst, PowerBI and ETL Developers and Tester roles to bring in delivery consistency.

MPR Program had six projects being executed in parallel and was the most mature DSI delivered at Transport. While uncertainty around data and schedule was still there, the team had ability to manage risks better. Daily 15-minutes Standups ensured that any blockers are addressed promptly. Fortnightly Agile ceremonies such as Retrospectives, Planning and Reviews took place without fail. Regular backlog grooming allowed us to plan the next sprint. Monthly Program Risk Review Meetings were scheduled with the full team to update Risks, Assumptions, Issues and Dependencies (RAID) Register. The communication with the stakeholders worked well with frequent showcases with product owners giving them ability to provide feedback. All the lessons from earlier implementations were applied to MPR and the program was able to retain key resources from other five DSIs. The schedule was being managed using quarterly iterations and Harvey balls to show progress of each deliverable as per *Figure 10*. We no longer did detailed scheduling at individual dashboard and dataset level; and instead identified user stories for each sprint till the dashboard and associated datasets went through User Acceptance Testing and deployed in production environment.

	Iteration 1 (Jan – Mar)	Iteration 2 (Apr – Jun)	Iteration 3 (Jul – Sep)	Iteration 4 (Oct – Dec)
MPR Program	• Technical Assurance Plan Approval by OTASC	• MPR Solution Architecture (Initial) Approval by TDA & OTASC • System Classification • Cloud Risk Assessment • Business Impact Assessment	• Masterdata HUB Solution • Service Design (Application Layer) • Information Security Management Plan	• Service Design (Platform) & Transition • MPR Solution Architecture (Final) Approval • Penetration Testing & Cyber Security Report • Business Glossary
Ferry	• Patronage / Opal data Analysis • Data Modelling for GTFS feed	• On-Time Running Report	• Service Availability Dashboard • Service Interruption Dashboard • Service Investigator Dashboard • Unscheduled Dashboard	• Executive Dashboards • Customer Feedback (Salesforce) • Customer Satisfaction Index • Customer Complaints • Patronage Dashboard • Training Videos • Productionise Dashboards
Light Rail	• Journey Time & Frequency Report • Service Availability Report	• Service Availability Report • Headway Performance • Service Adjustments • Service Interruption	• Productionise four Dashboards • Executive Dashboards (incl. 9:00am Minister Report) • Traffic Signal Delay (TSD) Dashboard	• Training Videos • Productionise Dashboards
TCB	• Training for PTIPS Analytics Reports	• On-Time Running Dashboard	• Productionise OTR & Passenger Load Dashboard • Passenger Load Dashboard	• Training Videos • Passenger Load Dashboard • Cancelled & Incomplete Trips
Sydney Metro			• Journey Time & Frequency Report • Headway Performance • Service Availability Report	• Service Interruption • Patronage Report • Police Reported Crime (BOCSAR) • Asset Failures • Asset Management • Safety Incidents • Training Videos • Productionise Dashboards
Data Ingestion		• Ingest Timetable & Trip Data - All modes • OpenData Self-Service - Ferry data	• Productionise Timetable & Trip Data Ingestion • OpenData Self-Service – API Access to ODL	• OpenData Self-Service – Direct Access to ODL • Training Videos
ODL Foundation	• Non Prod Environment Design & Setup (Initial) • Data Acquisition & Processing Contracts • ODL Foundation Contract	• Non Prod Environment Design & Setup (Final)	• MPR MVP Solution Architecture • MPR Prod Environment Build & Migration • Initial ODL Platform Solution Review with Microsoft & Databricks	• ODL Platform Solution Architecture (Final) Approval by TDA & OTASC • Prod Environment Build • External Review (Microsoft++) Platform

Figure 10. MPR Delivery Roadmap

Transport as an organisation matured in delivery of DSIs by the six DSI. When Harris, the Test Lead on Ferry Performance Reporting Project raised risk MPR-R-008 "*Risk that business requirements are not delivered* "; Sophie, our Portfolio Manager requested in our regular Program Risk Review meeting that we accept the risk with a note "... requirements will evolve with data discovery and development of dashboards.".

Our data-related processes matured as Harvey, our developer commented in 20th April 2020 Retrospectives "*Some nice design in data loading to standardise the data loading.*". He further added "*Consider data management is part of instead of addition to the implementation. One goal is to be able to hand over to somebody else to support as well as share the best practice and design. The value is beyond the project scope.*". Mo, our Information Architect commented in the same Retrospectives "*CSELR Automation have gone well. The data being refreshed on a daily basis*". In 4th May, 2020 Retrospectives, Mo commented "*TCB Data ingestion went quite good*".

Deep engagement with business is essential for Agile delivery. Team acknowledged this requirement several time in Retrospectives – "*Involve business stakeholders in dashboard testing*" on 10th September 2020; "*More frequent checking of report with certain stakeholders or points of contact, especially when requirements are not clear*" on 27th July 2020; "*Being proactive in gathering AND finalising Requirements – at least one sprint prior*", "*Engage with stakeholders at an early stage*" and "*Stakeholder engagement during daily stand-up*" on 27th July 2020; "*Stay nimble and agile in how we deliver.*" on 13th July 2020; and "*Work closely with business stakeholders*" on 5th April 2020.

Showcases with stakeholders and team velocity were acknowledged as key success-factors. Abhra, our Scrum Master lauded "*Team pace*" on 5th April 2020; "*Improved performance. Appreciate the entire team*" on 20th April 2020; "*Inclined growth on Productivity*" on 4th May 2020; "*Maintain Sprint Velocity*" on 1st June 2020; and "*Good pace of overall team velocity.*" on 13th July 2020 Retrospectives. "*Continue Show cases with business*" was noted on 5th April 2020; "*Showcase MPR Program to (non-Program) OS & CST Stakeholders*" on 4th May 2020 and "*Showcase (sell) MPR outcomes to wider TfNSW Community.*" on 15th June 2020 Retrospectives.

A significant change took place as Sponsors accepted that uncertainty exists in delivery of right quality of data with appropriate business rules and visualisations; and that the MPR Program team will deliver the best possible outcome in shortest amount of time. This was a marked shift from Waterfall oriented fixed time-cost-scope mindset.

The MPR Program closed in June 2021 with component projects closing between October and May 2021. Scalable big-data platform was delivered to store, process and service analytics needs of Contract Managers, Operators and Data Analysts. The Agile delivery allowed business value to be delivered consistently and incrementally. Our ability to deliver value allowed us to add more projects to the program.

Terry, the Light Rail Performance Reporting Business Sponsor nominated the project team for Operational Systems Rewards & Recognition. The MPR Program was a finalist in Intelligent Transport Systems Australia Awards in "Excellence in Transport Data Award" category. The MPR Success Story was published by our delivery partners showcasing the strength of partnership and using technology to deliver data-driven business outcomes.

Looking back at our first and the sixth DSI, risks around delivery processes reduced between Vanguard and MPR as the organization and delivery teams better understood the "how to" of DSIs and moved from Exploration to Exploitation stage. However, what did not change was the uncertainty around datasets due to the number of datasets being ingested, source, inherent quality, and transformation complexity resulting in very high data complexity as per Table 4. The uncertainty was lower when the datasets were internal or known but in majority of DSIs, the datasets were external and the team did not know about them until we started ingesting them.

Table 4

Source Data Summary of six DSIs

		Count by Type			Count by Source		Count by Transformation Complexity				
Program	No of Data Sources	Reference	Master	Transaction	Internal	External	Very High	High	Medium	Low	Data Complexity[(b)]
Vanguard	14	8	0	6	8	6	3	3	0	8	Very High
Ferry	10	5	1	4	6	4	0	3	2	5	Low
CTABS	16	2	6	8	0	16	0	0	16	0	Medium
PTIPS Analytics	7	5	1	1	5	2	1	0	1	5	Very High
Light Rail Priority	7	5	1	1	5	2	1	0	1	5	High
MPR	20	5	3	12	8	12	3	6	6	5	Very High

(a) Transformation Complexity is determined by application of business rules
(b) Data Complexity is based on number of Data Sources, Source and Transformation Complexity

A review of Risks, Dependencies and Constraints of six DSIs in Table 5 shows majority of the ratings fell into high and very high category thus contributing to higher complexity and uncertainty in delivery. The ratings followed Transport Enterprise Risk Management (TERM) Framework and were based on monthly review of risks by the DSI delivery team.

Table 5

Risks, Dependencies and Constraints of six DSIs

Program	Risks	Dependencies	Constraints
Vanguard	Very High	Very High	High
Ferry	Medium	Medium	Medium
CTABS	Medium	Medium	Medium
PTIPS Analytics	High	Medium	High
Light Rail Priority	High	High	High
MPR	Very High	Very High	High

This analysis leads us to our first characteristic that DSIs carry high degree of uncertainty right from initiation through to closing phases.

A comparison of Business Case Benefits and Benefits Delivered from six DSIs in Table 1 shows the benefits to Transport were enablers for decision-making.

Table 6

Vanguard Business Case Quantified Benefits

Benefit Type	Current	Target	Benefit Measure	Benefiit Owner
Improved Fare Compliance	6.40%	5%	Non-Compliance Rate	Security and Revenue Protection
Improved fine payment	32%	45% (in Y7)	% paid on time	Sydney Trains, TfNSW.

The first DSI had "*Improved fare compliance*" as Business Case Benefit as per Table 6 however, the delivery team was unable to establish the direct benefit contribution. Though the team was able to provide analytics on fare-evasion to internal Security & Revenue Protection team, Transport Officers, and to the NSW Police they were not accountable for what was done to reduce the evasion including changing the behavior of travelling public. Team subsequently identified deeper socio-economic factors which lead to fare-evasion in certain parts of Sydney and NSW and even finding repeat offenders. In the Vanguard Success Story video, Tony, who became the sponsor in 2018 talks about "*Vanguard Program provided an innovative opportunity to consolidate transport crime statistics, security incidents and fare compliance data into one system. The challenge was that data sets were not centralised and held by several Government agencies and key transport operators. File structures differed and data collation validation systems were not automated. There were some inefficiencies within the Security and Revenue Protection teams where time was lost collecting, cleaning and collating data rather than conducting analysis and developing strategies and operational outcomes in the Transport Cluster. So, we looked to find a better solution.*". Tony closes with "*Vanguard has consolidated key datasets. It has greatly enhanced our capability to analyse trends, identify hot spots and share these insights with our partners.*". Vanguard laid the foundation of Data Management and DSI delivery and became a show-case for people, process and technology outcomes. Vanguard was nominated for Operational Systems Rewards and Recognition, 2018 Transport (Cluster) Awards under Safety Category and was a finalist in 2019 Project Management Institute's "Innovation in Project Management" Awards category. In the nomination for the awards, we called out the direct and indirect outcomes Vanguard delivered as per *Figure 11*:

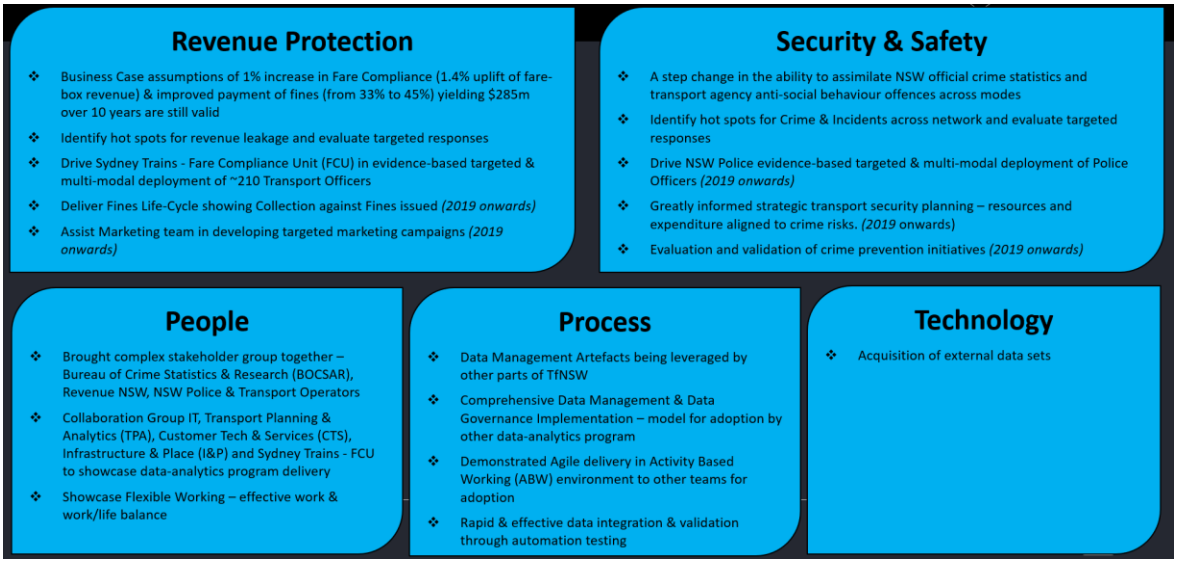


Figure 11. Outcomes Delivered by Vanguard

All other subsequent Business Cases showed benefits as enablers. The sixth DSI gave us insights into Bus, Ferry, Light Rail and Sydney Metro performance however they at best allowed Contract Management teams to identify and resolve operational issues. We were unable to quantify any hard benefits such as performance penalties from the insights we provided in sixth DSI. An example of Benefits outlined in Light Rail Performance Reporting Business case is:

- Evidence-based and data-driven contract management for KPIs and performance management
- Avoid costs of stand-alone data repositories and analytical services
- Reduce system complexity as a result of standardized data exchange

This is different from ICT Business Cases where there is typically a link to measure increase in revenue, customer satisfaction, efficiency, or reduction in costs. This leads us to our second characteristic that DSIs are enablers for decision making and may not have a direct benefit contribution.

A review of Business Case Benefits in Table 1 and our assessment of Benefits Complexity and Requirements Uncertainty in Table 7 shows a general lack of clarity in what we want to achieve out of these investments and how we want to achieve them. The ratings are based on judgement of the team and reflect relative complexity and uncertainty of the six DSIs.

Table 7

Benefits Complexity & Requirements Uncertainty of six DSIs

Program	Benefits Complexity	Requirements Uncertainty
Vanguard	Very High	Very High
Ferry	High	Low
CTABS	Medium	Low
PTIPS Analytics	High	Medium
Light Rail Priority	Medium	Low
MPR	High	High

Whether it was Vanguard or MPR, we did not have the certainty that with the datasets identified during Business Case development, we will achieve the identified outcomes. This leads us to our third characteristic that neither the goals nor the means of attaining them are clearly defined from the outset for a DSI.

Our case study of six DSIs showed the dependency. The first and second DSIs built our foundational knowledge and capability to deliver DSIs. From third DSI onward, each added additional services to the technology platform, brought in additional datasets and improved our DSI delivery processes. The Benefits Mapping and Dependencies of the six DSIs are shown in *Figure 12*.

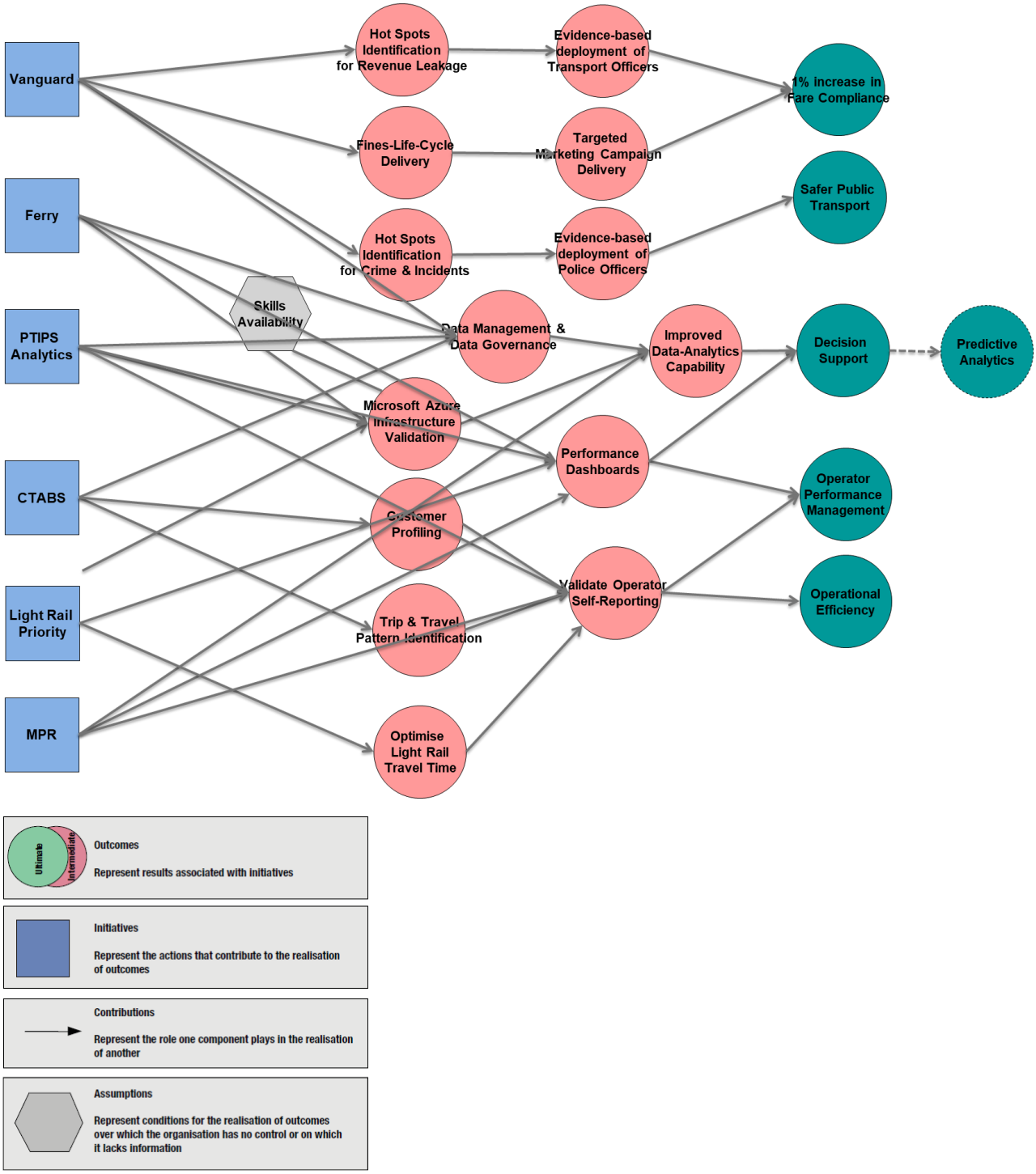


Figure 12. Benefits Mapping and Dependencies of six DSIs

We saw that sixth DSI used the foundational layers delivered by previous DSIs as well it reused the datasets ingested by earlier DSIs. While often there was no hard dependency on datasets previously ingested, the insights got richer in each subsequent DSI. This leads us to our fourth characteristic that DSIs are not independent of each other and act as an enabler to next one.

Our delivery of six DSIs showed that specific skills in each of the Program Management, Change Management, Scaled Agile, Data Management and Data Science domains are required as per Table 8. Depending upon the type of DSI, the skill level and amount of time required for each of the role varies but they are all essential to successful delivery.

Table 8

Roles & Responsibilities across DSI Domains

Domain	Role	Responsibility	DSI Essential	ICT Essential
Program Management	Business Owner	Have the primary business and technical responsibility for governance, compliance, and return on investment (ROI) for a	Yes	Yes
Program Management	Product Manager	Owns Vision and Roadmap and defines features and releases	Yes	Yes
Program Management	Program Manager	Responsible for overseeing the achievement of larger organizational goals by coordinating efforts between different	Yes	Yes
Change Management	Customer	Buyers of a solution. Internal customers are part of the enterprise whereas external customers are outside the enterprise and can be Business-to-Business (B2B), Business-to-Professional (B2P) or Business-to-Consumer (B2C)	Yes	Yes
Change Management	Change Manager	Lead change management and adoption of product and processes	Yes	Yes
Change Management	Change Analyst	Support Change Manager by deep-diving into stakeholder identification, impact assessment and taking them through the	Yes	Yes
Scaled Agile	Domain Architect	A specialist with deep knowledge within a particular domain of their expertise. A domain could be 'Data Services,' 'Process	Yes	Yes
Scaled Agile	Product Owner	Responsible for defining and prioritising stories to streamline the execution of program priorities while maintaining the	Yes	Yes
Scaled Agile	Scrum Master	Servant leaders and coaches for an agile team who help remove impediments and foster an environment for high-	Yes	Yes
Scaled Agile	Business Analyst	Guide businesses in improving processes, products, services and software through data analysis. They act as an interface between IT and the business to help bridge the gap and improve efficiency.	Yes	Yes
Scaled Agile	Tester	Responsible for the quality of software development and deployment and performing automated and manual tests to	Yes	Yes
Scaled Agile	Cloud Engineer	Responsible for duties associated with cloud computing, including design, planning, management, maintenance and support. Can encompass a few different roles such as cloud architect, cloud software engineer, etc.	Yes	No
Scaled Agile	Security Engineer	Identify threats and vulnerabilities in systems and software, develop and implement solutions to defend against hacking, malware and ransomware, insider threats and all types of cybercrime.	Yes	No
Data Management	Data Owner	Has the authority and accountability for the information assets. Decides who has the right to access and edit data and how it's used and be responsible for overseeing and protecting a data domain.	Yes	No
Data Management	Data Custodian	Responsible for the safe custody, transport, storage of the data and implementation of business rules and is generally a technical role. Can also be called Database Administrator (DBA), Data Modeller, ETL Developer. Acts as the proxy for the	Yes	No
Data Management	Data Steward	Responsible for data content, context, and associated business rules and is generally a business role	Yes	No
Data Science	Information Architect	Implements information structure, features, functionality, UI and focuses on structural design and implementation of an	Yes	No
Data Science	Data Architect	Responsible for data architecture and data integration and may work at the enterprise level or functional level. Work on the structural design of an infrastructure specific to collecting data, pulling it through a lifecycle and pushing it into other	Yes	No
Data Science	Data Scientist	Analytical data experts who have the technical skills to solve complex problems and curiosity to explore what problems need to be solved. They're part mathematician, part computer scientist and part trend-spotter straddling both the business	Yes	No
Data Science	Machine Learning (ML) Engineer	Focuses on researching, building and designing self-running artificial intelligence (AI) systems to automate predictive models and deploy into production environment	Yes	No
Data Science	Data Engineer	Works with multiple databases to capture and process live, streaming, and distributed data. Designs and develops data collection, management, and search-and-retrieval systems in order to support the collection, processing, exploitation,	Yes	No
Data Science	Data Analyst	Responsible for providing descriptive statistics, probability models, and other quantitative assessments of raw, processed, and generated data. Employs a combination of traditional statistical and machine learning/artificial intelligence techniques	Yes	No
Data Science	Data Communications	Responsible for communicating and presenting summaries of structured and unstructured data in visual, text-based and interactive formats. Requires strong technical knowledge for implementing data visualizations using technologies such as	Yes	No

We also note that an ICT Program using agile methods will have some of the roles defined in Table 8 but there are skills in Data Management and Data Science domains which are specific to DSIs. The last two columns compare which roles are essential in DSI vs ICT Programs. This leads us to our fifth characteristic that skills required to deliver a DSI are different to a typical ICT program.

All six DSIs are now closed, and this has allowed us to carry out both real-time and retrospective data collection. While the scale of the DSIs is different, together they have allowed us to identify characteristics of DSIs which brought in uncertainty in their management and governance. Table 9 shows the gaps and issues observed across three program phases by the author pre-delivery, during and post the delivery of the DSIs.

Table 9

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.

The six DSIs gave us a good starting point of their unique characteristics. These were further validated with semi-structured interviews with practitioners from five organizations. Table 10 provides a background of semi-structured interview participants who are all practitioners delivering DSIs.

Table 10

About Semi-Structured Interview Participants

Name	Title	Organization	Years of Experience	DSI Delivery Experience	Interview Date
Abhijit Pattnaik	Practice Director | Data, Analytics and AI	Wipro	13	13	22/05/2021
Mo Rashid	Information Architect	Suncorp	27	27	28/05/2021
Iman Eftekhari	Founder and Director	Agile Analytics	20	18	31/05/2021
Kale Temple	Co-CEO	Intellify	7	7	23/05/2021
Rodney Joyce	Managing Director	Data-Driven AI	23	5	1/07/2021

A questionnaire consisting of seventeen questions was prepared for qualitative data collection using open questions and all participants were asked same set of questions in 1-hour interviews. The objective of the questionnaire was to gain an appreciation from practitioners on challenges they faced in delivery of DSIs and explore the connection of the characteristics identified through our case studies from Transport with their cross-industry experience. The first section was knowing more about the participants, their organization and typical projects they have delivered. The second section was structured to reveal participants views of DSIs, how they compared with other ICT Programs and the frameworks/methodologies they have used. The third section was used to validate each of the five characteristics on DSIs which had emerged from case study of six DSIs from Transport. The last section allowed any other comments or thoughts to be gathered. The participants were asked in this section if they would like to add other characteristics. Based on the response to this question in Closing section, a sixth characteristic emerged. The questions organized across four sections are listed in Table 11.

Table 11

Interview Questions

Section 1 – Interviewee Background	
1	Could you please tell me about your background? How long have you been working in ICT sector? What type of industries have you worked in?
2	What is a typical week in your role look like? What are the challenges you face?
Section 2 – Comments on DSIs	
1	What is DSI to you? Please share your experiences of a DSI Program you have worked on. What was your role and what was the size of Team/Budget/Duration?
2	How long have you been managing DSIs? What has changed in terms of people, process and technology?
3	How do you compare an ICT Program with DSI? What similarities do you see? What differences do you see? Has this changed over time? If so, why?
4	What Delivery Frameworks/Methodologies have you used for DSIs. Why did you choose them? Is there a criterion you use when choosing the Delivery Framework / Methodology for a DSI?
5	Were the DSIs you delivered considered successful by Sponsor/Client. What made them successful? When were they considered not to be successful?
6	What key risks and issues have you seen in DSIs?
7	Anything you could have done differently to deliver DSIs?
Section 3 – Validation of DSI Characteristics	
1	DSIs carry high degree of uncertainty right from initiation through to closing phases.
2	DSIs are enablers for decision making & may not have a direct benefit contribution
3	Neither the goals nor the means of attaining them are clearly defined from the outset for a DSI
4	DSIs are not independent of each other and act as an enabler to next one
5	Skills required to deliver a DSI are different to typical ICT program
6	DSIs do not end and after initial delivery convert into continuous business improvement initiative
Section 4 – Closing	
1	Is there any other part of DSI which you would like to discuss?
2	Can you think of other people I should talk to about DSIs?

In semi-structured interviews, we wanted to probe the themes which consistently emerged in six case-studies within Transport. When asked about the uncertainty, four out of five responses

from semi-structured interviews agreed with this characteristic and shared what they are doing to address it. However, one response highlighted the increasing maturity associated with DSIs and availability of "*out of box models*" which reduce uncertainty in specific use cases. Iman commented on existence of unknowns in DSIs and "*There is [a] cone of uncertainty and for DSIs the variation from beginning [to end] could be plus minus 99%."*. Iman justified using Agile methods "*to reduce level of uncertainty by developing, showing and getting feedback in agile and iterative way*". Kale acknowledged the inherent risks in most DSIs and mentioned that his consulting organization has a high bar in selection of client DSIs for delivery. His organization does not accept DSIs for delivery which carry high risk in order to maintain high customer satisfaction and as a result "*We have got most initiatives to probably a 98% success rate and very low failure rates and adoption rates from projects at the moment."*. He emphasized the risk-averse approach of *"if we don't think we can do it, we won't take it*". An obvious downside of this approach is that Kale is walking away from some business opportunities in a market where the demand is high, and supply of skilled resources is comparatively low. Rodney agreed that within an organization, the degree of uncertainty starts to drop once the organization has done a few DSIs validating the exploratory nature with "*you learn lot of hard lessons on the first one*". However, Abhijit pointed out that DSI industry is maturing with the availability of "*plug-and-play use cases"*. The increasing availability of such solutions allow system integrators reuse of models *"it's become a 60-40 mix, 60% of it is pre-built stuff that people are just coming and plugging in for you now"*. Abhijit acknowledged "*when people are doing exploration, your assumption is 100% right"*.

This leads us to conclude that when the data is from a well-defined and structured source such as Salesforce, SAP or Peoplesoft applications, the degree of uncertainty is reduced, and organizations have a choice to use pre-built solutions. Hence this characteristic needs to be qualified with type and complexity of data-source and model which is being built. At the same time, the responses validate our hypothesis of emerging shift from exploration to exploitation as the organization and industry matures in delivery of DSIs. This allows us to finalize our first characteristic that "*DSIs carry high degree of uncertainty right from initiation through to closing phases except for when the data is from a well-defined and structured source*".

While DSIs as enablers characteristic was true for the DSI case studies from Transport, a different story emerged from the interviews. Mo acknowledged the characteristic but also highlighted an example from work he did with traders in Energy NZ where "…*there can be*

situations where you can actually allocate benefits". Iman agreed that DSIs often do not deliver a tangible outcome and noted that DSIs validate a few things and "*At the end we have learning and knowledge, and that knowledge is the outcome of DSI project*". Kale's experience was like Mo's where "*There are cases when they can have a direct benefit contribution"* citing examples where business derived five to ten times of the investment through DSI. Rodney categorized the DSIs – one where you are doing a proof of value (PoV) or proof of concept (PoC) and second where you are operationalizing a solution and commented "*...it wouldn't be fair to compare the operationalization of the product to the PoV.*". Rodney's view was that in PoV and PoC you're testing out an idea or theory to move forward or not and but at the same time mentioned *"...seen examples where by implementing a DSI, business were able to actually save dollars"*. Abhijit also echoed similar sentiments "*I think it's bit of half and half.*".

Hence, we can say that the validity of this characteristic is based on context in which a DSI is delivered. This allows us to finalize our second characteristic that "*DSIs are often enablers for decision making & may not have a direct benefit contribution*".

The lack of clarity on goals and means of attaining them characteristic was generally ratified by the interview responses. Mo agreed with the characteristic that you iterate and adjust what you are delivering "*...need to go through the discovery process and understand what your final goal is".* Iman mentioned that unlike an engineering design there are lot moving parts and "*In DSIs you don't know that it is a 4-bedroom, 2-bathroom and a backyard house".* On managing expectations, he added *"You don't expect any set delivery from the beginning, even as you get to know better about technology possibility.*". Kale also highlighted the fluidity in starting DSIs "*What you end up with is not the same thing, and so what that means is that very few organizations know the proper requirements of what they're looking to achieve when they approach an ML initiative"*. Kale emphasized that unlike building software where you are lot more definitive around ways and outcomes you are not never sure of where you will end up in DSIs. Rodney agreed with the characteristic with a caveat that as market matures the goals will be clear "*The cloud vendors are packaging these services up and making you pay by the time you use it"*. Abhijit reiterated that with the emergence of pre-built solutions for standard data-sources, the "*how to*" is becoming clearer however the definition of success of a DSI engagement is one of the biggest challenges today with "*The customer sometimes would not know what the end game is they're looking for"*.

The exploratory nature of DSIs was highlighted in this characteristic and ratified by interviews with a caveat that as the market matures, the emergence of pre-built solutions will reduce the uncertainty. This allows us to finalize our third characteristic that "*Neither the goals nor the means of attaining them are clearly defined from the outset for a DSI with the caveat that as the market matures, the emergence of pre-built solutions will reduce the uncertainty*".

The interdependency of DSIs was emphasized through the interview responses. Mo agreed "*The sequencing is very important as it gives the foundation to the subsequent ones.".* Iman further elaborated that this is a knowledge building process where you are adding layers "*You build the first knowledge set and based on that you determine the next level on top of that"*. Kale commented on both technology and people. On technology aspect he mentioned breaking *DSIs* into comparable streams of similar use cases *"You can use a template from the first one to solve the second one, assuming you're trying to solve similar use cases like Customer Churn, Cross Sell".* On people side, he commented *"They get more knowledge every single time they do it, so there are incremental gains.".* Rodney agreed that there is a layered approach in DSIs where you keep iterating and "*It's kind of the agile continuous improvement idea to a delivery - fail fast and move on as opposed to fail slow".* Abhijit agreed that the DSIs can help you drive towards a business outcome and often are quite related "*You move from one use case to another use case, which is essentially taking you from where you are to your mission and vision of where you want to be in two or three years".*

In summary the characteristic was validated and allows us to finalize our fourth characteristic that "*DSIs are not independent of each other and act as an enabler to next one*".

The skills requirement characteristic was apparent in all interview responses reflecting the evolving nature of ICT sector and especially Data Science. Iman called out skillset in delivery and implementation domains and the need for cross-functional team which understands end-to-end delivery process "*The same person may play different role at different stages of DSI depending on the complexity and size of project".* Kale called out different skill sets with reference to software engineering versus data science in terms of programming, programming languages, techniques and even mindset. He commented *"When you talk about DSIs, it is hard problem-solving using data techniques and I'm not saying that ICT doesn't have to do that, but you are architecting and reengineering business processes at the same time which could be a very different thing to - I'm going to manage security or networking or infrastructure.".* Rodney elaborated that five years ago

some of these roles did not even exist. Abhijit went further that besides technical skills "*I think skill varies significantly in understanding and appreciating the business context".*

In summary, this characteristic was validated quite emphatically in all responses and allows us to finalize our fifth characteristic that "*Skills required to deliver a DSI are different to those required for a typical ICT program*".

Our sixth characteristic came up when question was asked if interviewees have seen any other characteristic. Iman who first raised this saw a pattern that once you show value and what is possible, DSIs go from one phase to another as the business wish to go further or expand it to other functional areas "*It has to be ongoing and a mindset of not making it finite in terms of budget and delivery".* Kale agreed that the business context changes overtime and what that means is that once a DSI is delivered, the end product needs to be monitored and improved overtime. "*I've got a saying that we don't do projects, we build products, and it's exactly that point which is once you've built something, a model will change overtime".* Rodney concurred with the evolution of software "*That's the nature of software now. […] the software is never finished, it's just continually changing.* Abhijit mentioned his experience in pharmaceutical device manufacturing industry where "*They have grown and scaled that model out.*"

While this characteristic was thought of later, it was still validated and shows how DSIs deliver products with data and models both of which evolved over time. For the six Transport DSI Case Studies, five products are still being used and continuously improved validating this characteristic. This allows us to finalize our sixth characteristic that "*DSIs do not end and after initial delivery convert into managing the product, model and data*".

5. Conclusion and Recommendation

In this section, we review our research question: *"What unique characteristics cause DSIs to face challenges delivering envisaged value when using traditional processes for managing ICT-enabled programs*?" and summarize our conclusion. We first start with conclusions from this paper, discuss the limitations of our research, the implications on business managers and practitioners in understanding characteristics of DSIs and end with the recommendations.

5.1 Conclusion

We conclude that current literature does not adequately cover unique characteristics of DSIs and the business managers and practitioners need to be informed about the differences between DSIs and ICT-enabled programs so that they adapt methods to improve the chance of successful business outcomes.

Program Management for ICT-enabled Programs has rich literature and proven delivery frameworks which have matured over the past three decades (Axelos, 2020; Project Management Institute, 2016, 2017). This paper makes a significant contribution to the practice of the emerging field of data science and program management.

The current Program Management literature does not adequately support delivery of innovative and exploratory DSIs and instead focuses on risk elimination and rapid delivery of business outcomes of exploitative initiatives. We identify six unique characteristics of DSIs to be used in delivery of DSIs and complement domains identified in literature review and practice - 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.

We also conclude that organization and teams go through stages of Exploitation, Transition and Exploration in delivery of DSIs. The process of delivering DSIs becomes efficient as they deliver more of them. With exception of data from well-defined and structured source, every new dataset carries uncertainty in scope and quality. This brings in framing of an underpinning exploration component to a DSI combined with a shift to exploitative as the organization and teams mature.

5.2 Limitations and implications of research

This research has used six DSIs from one public sector organization in Australia as case studies to identify unique characteristics and validated with semi-structured interviews with practitioners from five organizations. The authors of this paper are undertaking development of a DSI Delivery Framework incorporating program management, change management, agile delivery, data management, and data science domains to assist practitioners. Sandberg (2005) notes that truth is always something unfinished within the interpretive tradition, the criteria proposed do not enable researchers to generate absolute truth claims. We believe that the DSI characteristics we have identified do not present an exhaustive and universal set and more may emerge as the field of data science advances. Future research can include validating the characteristics with other public and private sector organizations delivering DSIs in other countries. 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 as some of the characteristics will change as the maturity DSIs changes from being exploratory to exploitative.

5.3 Implications for practice

Limited availability of methods and standards in delivery of DSIs has caused the business managers and practitioners 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.

5.4 Recommendations

We started with five unique characteristics and validated them with using semi-structured interviews with five practitioners. As a result, the characteristics have been updated and a sixth one has emerged. The six characteristics are summarized in Table 12:

Table 12

Unique Characteristics of DSIs

No	Description
(i)	DSIs carry high degree of uncertainty right from initiation through to closing phases except for when the data is from a well-defined and structured source
(ii)	DSIs are often enablers for decision making & may not have a direct benefit contribution.

(iii) Neither the goals nor the means of attaining them are clearly defined from the outset for a DSI with the caveat that as the market matures, the emergence of pre-built solutions will reduce the uncertainty.

(iv) DSIs are not independent of each other and act as an enabler to next one

(v) Skills required to deliver a DSI are different to those required for a typical ICT program

(vi) DSIs do not end and after initial delivery convert into managing the product, model and data

We suggest additional research to validate the characteristics with other public and private sector organizations delivering DSIs. As the field is evolving rapidly, the authors believe that the six characteristics identified in this paper will also evolve. It is possible that new characteristics may emerge and some identified here are no longer treated as unique as DSIs go from exploratory to being exploitative. With the size of investment underway in DSIs and need for data-driven decision making in organizations, additional research is essential in the field of DSIs. The characteristics identified in this paper will deliver a small but significant contribution to the body of knowledge for Program Management relevant to both literature and practitioners. Without this understanding, 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 improving data-driven decisioning capability.

References

Agar, M. H. (1986). *Speaking of ethnography* (Vol. 2): Sage.

Asay, M. (2017). 85% of big data projects fail, but your developers can help yours succeed. *Big Data Tech. Rebuplic, 11*, 1-5.

Axelos. (2020). *Managing successful programmes* (5th ed. ed.): TSO.

Brady, T., & Davies, A. (2004). Building project capabilities: from exploratory to exploitative learning. *Organization studies, 25*(9), 1601-1621.

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*.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review, 14*(4), 532-550.

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.

Foroughi, F., & Luksch, P. (2018). Data Science Methodology for Cybersecurity Projects. *arXiv preprint arXiv:1803.04219*.

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

Hunter, A., & Brewer, J. D. (2015). Designing multimethod research. In *The Oxford handbook of multimethod and mixed methods research inquiry*.

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.

Loch, C. H., DeMeyer, A., & Pich, M. (2006). Managing the unknown: A new approach to managing high uncertainty and risk in projects 2006. *Hoboken, NJ*.

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

Project Management Institute. (2016). *Governance of Portfolios, Programs, and ProjectsGlobal standards* (pp. 1 online resource (130 p).

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*.

Sandberg, J. (2005). How do we justify knowledge produced within interpretive approaches? *Organizational research methods, 8*(1), 41-68.

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/

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

Small, M. L. (2009). How many cases do I need?' On science and the logic of case selection in field-based research. *Ethnography, 10*(1), 5-38.

Straits, B. C., & Singleton, R. (2018). *Social research: Approaches and fundamentals*: Oxford University Press.

Transport for NSW. (2018). Future Transport Strategy 2056. Retrieved from https://future.transport.nsw.gov.au/sites/default/files/media/documents/2021/FutureTransportStrategy2056.pdf

Transport for NSW. (2021). Future Transport Technology Roadmap 2021-2024. Retrieved from https://future.transport.nsw.gov.au/sites/default/files/media/documents/2021/NSW_Future_Transport_Technology_Roadmap_2021-2024.pdf

Van de Ven, A. H., Polley, D. E., Garud, R., & Venkataraman, S. (1999). The innovation journey *Oxford University Press, New York*.

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