Leiden University

ICT in Business and the Public Sector

A Framework for Evaluating Innovation and Data Science Initiatives

Name:	Ayush Sharma
Student ID:	s3840603
Date:	09/08/2025
1st supervisor:	dr. A.P. Pereira Barata
2nd supervisor:	Prof.dr.ir. J.M.W. Visser

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: Innovation and Data Science (IDS) teams are increasingly recognized as key components in driving technological advancements and facilitating data-driven decision-making within organizations. These teams operate on different principles from *traditional* technology teams. As such, traditional evaluation practices may not fully capture the impact of IDS teams. Therefore, there is interest in establishing a structured approach to assess the value of these teams. By establishing a comprehensive evaluation framework, valuable insights can be gained on project outcomes, organizational efforts, and strategic goals.

Aim: This research aims to deliver a validated evaluation framework designed with the goal of better understanding the impact of a project by rigorously evaluating it at both the inception of a project and its conclusion, directly addressing three objectives: (1) to design a structured assessment tool that captures both the value and feasibility of projects; (2) integrate an ML–based validation technique to refine and support the assessment tool; and (3) provide visualization that supports decision making, helping streamline future project workflows.

Materials and Methods: The research subject department is the *Innovatie- en Datalab* within the *Inspectie Leefomgeving en Transport*, part of the *Ministerie van Infrastructuur en Waterstaat* in the Netherlands. A dual-phase assessment scorecard was developed using qualitative insights and root cause analysis guided by the design science research methodology, and then validated through a constrained autoencoder model to optimize and confirm its structure.

Results: The final framework consists of a pre- and post-assessment scorecard structured around two validated dimensions: Innovation and implementation. The ML approach supported by a constrained autoencoder was able to successfully reduce responses to an interpretable 2D Innovation–Implementation quadrant (Diamond Model), capturing pre-to-post shifts that clarified differentiation between innovation potential and delivery feasibility. Per-question contribution analysis further validated the refinement of the scorecard questions, and Brier-RMSE analyses indicated consistent, accurate mappings. These insights help guide strategic decisions on choosing the right delivery tool and setting realistic expectations.

Conclusion: This research aimed to tackle the current challenges in assessing the impact of IDS teams. To do so, a framework was tailored and thoroughly validated, using a balanced scorecard with a constrained autoencoder for practicality. The resulting approach bridges structured evaluation with actionable management insights, ensuring that assessments directly inform strategic alignment. The framework demonstrated that effective innovation endeavours require management techniques specifically adapted to team culture and technological focus, while maintaining the flexibility necessary for creative exploration. Ultimately, this work contributes to the growing field of innovation management.

Acknowledgements

I would like to express my deepest gratitude to my first supervisor, Dr. A.P. Pereira Barata, for your unwavering support during challenging moments and for the valuable suggestions that have elevated my thesis to a higher level. Your guidance has undoubtedly helped me grow throughout this period, and you have consistently pushed me to think in terms of opportunities and possibilities, rather than limitations.

I am also sincerely grateful to my second supervisor, Dr. J.M.W. Visser, whose supervision helped me bring together the loose ends in my thoughts into a clear and comprehensive whole.

A special thanks goes to Inspectie Leefomgeving en Transport (ILT) for hosting my thesis internship. I am especially thankful to all my colleagues at the IDLab for their time, support, and generosity in sharing their expertise with me.

Lastly, I am deeply grateful to my family and my girlfriend for their unwavering encouragement and support throughout this research. I could not have achieved this without you guys.

पापा, ये आपके लिए है। मुझे उम्मीद है कि ये आपके चेहरे पर मुस्कान ला देगा।

Contents

Chapter 1

Introduction

In today's rapidly evolving digital era, Innovation and Data Science (IDS) teams have emerged as critical drivers of change across industries [Kane et al., 2015]. Emerging literature underscores the growing importance of embedded data science teams which forms the base of IDS teams to support organizations responding to technological advancements, improving decision making, and developing new services [Luo, 2022]. Organizations across the public and private sectors are leveraging data insights, automation, and experimental technologies to stay competitive and adapt to technological advancements [Waller and Fawcett, 2013]. However, this comes with significant challenges concerning the strategic management of these teams, such as those outlined by the "black box" theory of Kline and Rosenberg [1986] about the uncertain and non-linear nature of innovation processes. To lower these risks, there is an underlying need for developing a structured approach that can aid in systematically evaluating, selecting, and measuring the impact of IDS projects across these teams.

1.1 Background

The rise of artificial intelligence (AI) and the increased funding towards it have accelerated the adoption of IDS teams across the private and public sectors, to remain competitive with industry standards and major technology companies. For example, in 2021, AI investment grew by 108%, from $32.1 billion to $66.8 billion [CB Insights, 2021], with a projected contribution to the global economy of $15.7 trillion by 2030 [Banerjee et al., 2023].

In the public sector, IDS teams are increasingly seen as enablers of strategic transformation, playing a vital role beyond just experimenting with new technologies. Their role extends to addressing broader organisational goals, such as improving oversight, enhancing decision-making, and driving policy innovation. Considering their diverse capabilities, the motivation behind the establishment of these teams also varies significantly, both in terms of the type of their innovative products and their approach. As Hauser et al. [2006] noted, there is no one-size-fits-all approach; organisations differ widely in what they innovate and how they do it. In public administration, this often translates into a dual mission: to deliver operational efficiency while also generating long-term societal value. These teams have shown the potential to not just reshape operations but also drive fundamental long-term societal benefits, e.g., improved decision-making through transparent and data-backed evidence. In addition, they have the potential to transform digital interactions between citizens and the government while lowering operational costs [Mergel, 2016].

Despite the benefits and steady increase in adoption of these teams [PWC Data Science Group, 2023], many face challenges in maintaining value beyond initial pilots. History has shown that numerous once-innovative teams eventually lose their edge, with some organizations ultimately failing altogether [Christensen, 1997].

Without defining success metrics, evaluation criteria, or clear intake systems, teams are left to function in an uncertain environment. According to the Global Innovation Survey by Boston Consulting Group [2015], barriers to innovation are most often related to poor selection and implementation of concepts, with organizational culture also cited among the top six impediments. These issues contribute to what

Mazzucato [2018] describes as an innovation imbalance, where ideas are pursued without alignment on feasibility or value, leading to long timelines and bad execution.

Realizing innovation's true potential requires more than technical capability; it demands strategic alignment, clear value assessment, and effective execution. In practice, many public sector innovation teams struggle to meet these expectations due to foundational gaps in how innovation projects are selected, managed, and evaluated [Organisation for Economic Co-operation and Development (OECD), 2017, Mergel et al., 2019]. Table 1.1 below outlines the common risk traits identified during the interviews and stakeholder sessions of this research. These findings forms the basis of the framework and are further analysed in 3.2. These areas highlight the shortcomings in strategic alignment, evaluation, and visibility.

1.	Unclear project value at intake
2.	Lack of criteria to assess feasibility and viability
3.	No clear way to match projects with Agile or traditional methods
4.	Limited visibility of project results
5.	No shared definition of success across projects
6.	Weak communication of value to stakeholders
7.	Unstructured decision-making in project selection

Table 1.1: List of risk areas

Several European institutions and organizations, since at least the 1990s [Gray, 2014], have focused on data provision and sharing for various purposes. Specifically in the Netherlands, the *Inspectie Leefomgeving en Transport* (ILT) is the legal supervisory authority under the Ministry of Infrastructure and Water Management. It is tasked with safeguarding safety, sustainability, and confidence in domains such as transportation, infrastructure, the environment, and housing. Within this broad regulatory scope, ILT inspectors carry out monitoring and inspection duties in the field, assisted by the data made available through companies and other regulated entities. However, often the data collected is too complex for manual analysis by inspectors.

In recognition of this challenge and to uncover the potential of IDS, the ILT formed the *Innovatie-en Datalab* (IDLab) in 2017. The purpose of IDLab is to strengthen the ability of the inspectorate to act as a modern data-informed regulator by exploring innovative tools and methods. Importantly, the goal is not to replace inspectors but to enhance their work through smart assistance tools. Since its inception, the ID Lab has explored and tested innovative ideas such as risk-based inspection models for inland shipping, data dashboards to aid in real-time decision-making during inspections, and tools that automatically identify unusual patterns in environmental reports. These systems allow inspectors to focus their efforts where they are most needed, increasing the overall effectiveness of ILT's regulatory work. In upcoming Chapters 2 & Section 3 we review the technologies used and examples of some projects in more depth.

Although the adoption of IDS teams has grown rapidly, the challenges in managing these efforts with organizational goals, selecting the right projects, etc., have also persisted for a long time.

1.2 Problem statement

The core issues across innovation teams lie in how projects are selected, managed, and evaluated. This research uses IDLab as a representative case to explore the challenges faced by the team. Although the team is equipped with technical skills, it faces challenges in systematically demonstrating the value of its initiatives and selecting projects that maximize strategic and social impact. Although innovation projects are not always intended for deployment, a surprising number of IDLab initiatives end in a prototype or exploratory stage with no clear follow-up strategy. The absence of a structured approach to assess the project's feasibility and strategic value early on thus makes it difficult to translate the lessons learned and often gets lost.

Lack of clear guidance on how to assess the value of new ideas or demonstrate tangible results can cause the team to disconnect from organizational goals, operational support, and miss opportunities for broader impact. Based on these observations, the following problem statement is proposed.

PS: *In what way can innovation teams be supported by a structured, data-driven framework that evaluates and guides decision making at the inception of a project?*

To address the Problem Statement, we will decompose it into three tractable RQs.

1.3 Research questions

Towards answering our problem statement, we propose the following decomposition into three RQs:

RQ.1 How can structured pre- and post-assessment questions be designed for IDS teams to assess the value, viability, and feasibility of smart technology projects?

RQ.2 How can machine learning be used to validate and improve the questions of a project assessment scorecard?

RQ.3 How can dimensionality reduction help simplify complex project data and guide project management decisions?

To answer these research questions, we study the IDLab as a specific case. The research assesses how innovation projects are currently managed, and then designs a data-driven evaluation framework to improve project intake and classification, validate the usefulness through team members interviews.

1.4 Framework Objectives

Most current innovation frameworks are either too general or designed with the private sector in mind, with a strong focus on speed to market and return on investment [Pisano, 2015] [Goffin and Mitchell, 2016]. They offer little assistance in aligning innovation projects with more general societal goals, public value, or guidance for early-stage decision making. Furthermore, not many frameworks provide useful resources for categorising different types of projects or choosing suitable management techniques according to project attributes.

Current frameworks are too qualitative and rigid. No integration of ML with human input. Our framework helps bridge this gap by introducing a learning component that maps structured human input (questionnaire scores) to a project typology model in a dynamic and generalizable manner. The purpose of this research is to address these shortcomings and develop a framework to support innovation teams in selecting, managing, and evaluating projects more effectively. To help them deliver greater value, align with user needs, and communicate their impact more clearly.

To realise this goal, the Design Science Research Methodology (DSRM) [Peffers et al., 2007] is used to guide the development of this framework. Given its structured approach to problem solving in information systems, it follows a structured six-step process: (1) problem identification; (2) definition of solution objectives; (3) design and development; (4) demonstration; (5) evaluation; and (6) communication.

This framework is designed to meet the following objectives:

- **Design a Structured Project Assessment Instrument:** Drive go no-go decisions by providing clear, actionable measures to evaluate the viability and feasibility by developing a scorecard consisting of pre and post-questions.
- **Ensure Consistency and Reliability in Project Evaluation:** Validate and strengthen the scorecard by applying an autoencoder-based model, improving the accuracy of the questions, and checking whether the questions are grouped and interpreted consistently across all projects.
- **Enhance Visibility of Project Outcomes:** Simplify complex assessment data into interpretable dimensions (e.g., Innovation and Implementation) and map projects within a 2D strategic quadrant, helping teams better understand how their perception of a project compares to its actual strategic position, spot misalignments, and tailor project management approaches accordingly.

By addressing these objectives, the framework aims to assist the team in avoiding scope creep, reducing intake of unclear and personally-driven projects, while establishing clear assessment criteria that protect capacity for working on the most promising work.

1.5 Overview of the thesis

In Chapter 1, we introduce the accelerating adoption of IDS teams and then narrow our focus to the specific challenges of managing these teams within the public sector. Key issues are identified that form the basis for our problem statement and are later decomposed into three research questions. We conclude the chapter by outlining the three main contributions of this thesis.

The remainder of the thesis is structured as follows. Chapter 2 covers the related work & theoretical background, including types of innovation teams, workflow of these teams, and current evaluation frameworks, followed by dimensionality reduction techniques. Chapter 3 presents the research design, application of DSRM, and approaches to form the scorecard, are how to interpret it is described. Chapter 4 reports the evaluation results, including validation with the autoencoder and pre/post scorecard in action. The discussion in Chapter 5 interprets the findings, links them to the research questions, and discusses implications for practice. Finally, Chapter 6 summarizes the problem statement, the approach, the conceptual solution, and future research opportunities.

Chapter 2

Literature Review

The capability of IDS teams to explore potential applications of state-of-the-art technologies identifies them as the leaders in driving digital transformation efforts across organizations [Konopik et al., 2022]. However, these teams are linked to several challenges [Christensen et al., 2018], of which we mention two: (1) lack of a shared or consistent definition of innovation; (2) poor understanding of how innovation processes unfold in practice.

First, the term "innovation" is very multifaceted, and there is no general definition [Kogabayev and Maziliauskas, 2017]. When not clearly understood, this leads to ambiguity around project goals, viability, and success KPIs. Uncertainty about what is truly novel complicates the evaluation of project proposals. In this thesis, we do not assess ideas solely based on their novelty but place greater emphasis on their strategic fit and readiness within the organization.

Second, innovation work often tends to be messy and iterative. With uncertainties involved, it is hard to follow a strict approach for different experiments. Critical obstacles occur during the project, such as: technological constraints, insufficient support to innovation teams, and inappropriate evaluation frameworks [Pisano, 2019]. Without a clear understanding of the project value at hand, teams risk being caught in a trade-off of either rushing to deliver without validation or over-experimenting without producing tangible results.

This chapter examines existing research and literature across four key areas: (a) characterization and formulation of IDS, (b) life cycles of data science projects, (c) assessment practices for IDS teams, and (d) evaluation and dimensionality reduction techniques.

These concepts are structured to address two key challenges introduced earlier and form the basis for the analysis. These topics are critical in defining and understanding how innovation teams have evolved, their most critical bottlenecks, and how their value can be assessed.

Section 2.1 describes the practical landscape for innovation and data science teams. It begins by categorizing various types of innovation teams (2.1.1) and then investigates the distinguishing features of data-driven innovation teams (2.1.2). Then examines the application of IDS teams, their structure (2.1.3), as well as the technological architecture and workflows (2.1.4). Together, these sections define what innovation looks like in practice and highlight the differences in how teams are structured and operate.

Section 2.2 addresses the challenges in evaluating innovation by examining existing frameworks and methods. The subsections address both the theoretical foundation for defining innovation success (2.2.1) and the practical evaluation and assessment frameworks (2.2.2) proposed in the literature.

In Section 2.3, we look at techniques that can transform high-dimensional datasets into lower-dimensional representations while preserving the most relevant information, and how this can be crucial for assessing a real-world team.

Finally, Section 2.4 summarizes the identified gaps in current approaches and emphasizes the potential for a structured framework. These findings directly inform the design principles and objectives outlined in Chapter 3, which uses the Design Science Research Methodology (DSRM) to guide the creation of a practical framework tailored to the IDS teams.

2.1 Innovation and Data Science in Practice

We begin by examining innovation and data science as distinct practices, before delving into how their convergence has resulted in a new interdisciplinary field within applied science.

Innovation teams focus on exploring and testing novel ideas that can drive systemic or organisational change [Bason, 2018] [Kattel et al., 2020]. Often working in uncertain and complex environments using hypothesis-driven methods to navigate strategic challenges and design novel solutions.

Data science teams, in contrast, specialize in extracting insights from data through techniques such as machine learning, statistical modelling, and predictive analytics, these teams enable more informed decision making, improve risk assessment, and help evaluate the impact of innovation efforts.

The core difference lies in their orientation: While innovation teams are solution-driven and forward-thinking, data science teams are insight-driven and grounded in data analysis [Rainer Kattel, 2019]. The convergence of these capabilities forms Innovation and Data Science (IDS) teams, hybrid units that combine exploration with analytical foundation to experiment and solve complex organizational problems.

Therefore, this section first examines the definition of innovation and the types of innovation teams commonly found across organizations. Then it explores the role of data-driven innovation teams as an overarching theme, highlighting how data science capabilities are integrated into innovation practices. Finally, it looks at their core focus areas, technological approaches, and typical workflows, providing a foundational understanding of how these teams operate and create value in complex environments.

2.1.1 Innovation & Team Types

Innovation is a broad concept defined and applied differently across disciplines. It has evolved through continuous contributions from early economic and social theorists, making it challenging to arrive at a single, unified definition. Gabriel Tarde was among the first to define innovation as a socially driven process that drives societal changes by using new tools and patterns of behavior. He also highlighted the role of imitation in accelerating the spread of innovation, recognizing that innovation can produce uneven benefits across social groups [Tarde, 2001]. Joseph Schumpeter advanced Tarde's theory by linking it directly to economic development. Schumpeter, who is also termed the "Father" of innovation theory, emphasized the role innovation plays in generating economic growth and as a tool that improves competitive advantage [Vanderburg, 2005]. Crucially, he laid a foundational distinction between invention, i.e, the creation of a new idea, and innovation that he termed as the application of that idea to generate economic value [Schumpeter, 1939].

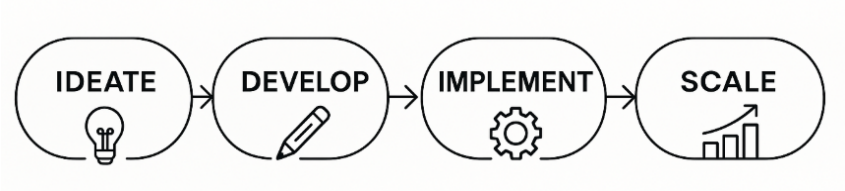


Figure 2.1: Conceptual Flow of an Innovation Team

This definition has further evolved as its application across different domains has expanded. As Fagerberg [2005] emphasizes, the literature on innovation is scattered, reflecting its evolution across fields such as economics, management, and public policy. O'Connor and Rice [2008] further claims that "[y]our academic knowledge of the phenomenon is not complete." In his study, another key concept is innovativeness, which is the ongoing ability to create, test, and improve ideas, even in the face of failure. Emphasising that it is never just about launching new products or services.

Innovation teams take on different forms and transition to different phases as shown in Figure 2.1, adapted from the study of Du Preez and Louw [2008], ranging from an internal R&D team to cross-functional labs or external accelerators. An effective innovation team systematically aligns with strategic goals, fosters experimentation, manages resources and interfaces, and scales validated ideas to deliver tangible results. Continuous learning and adaptation are essential throughout this process.

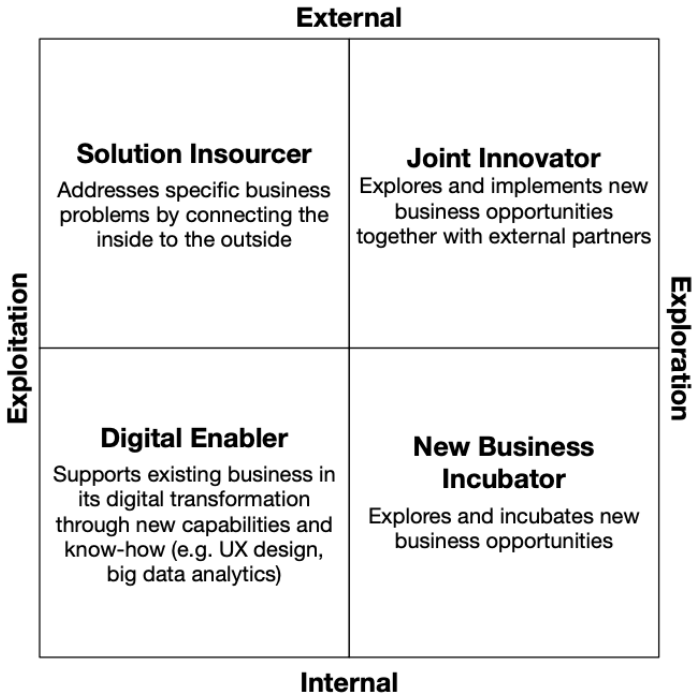


Figure 2.2: Innovation Team's Types

The tasks these teams take on differ, but they all share a common goal of driving sustainable and meaningful change through collaborative efforts. Recent research has identified four common types of innovation teams, each aligned with different goals and ways of working: Digital Enablers, Solution Insourcers, New Business Incubators, and Joint Innovators [Björk et al., 2023]. As seen in Figure 2.2, these types are organized along two key dimensions: whether they focus on improving what already exists (exploitation) or creating something entirely new (exploration), and whether they work mainly within the organization or with external partners. For example, Digital Enablers help organizations modernize by using technologies like AI and big data, while New Business Incubators and Joint Innovators focus on exploring new markets or services, often in collaboration with outside actors.

Based on a similar concept, we also see the emergence of data-driven innovation (DDI) labs, a hybrid team that integrates both exploitative and exploratory methods. These teams leverage advanced machine learning algorithms to power probabilistic and predictive models that help organizations streamline current operations. Their experimental capabilities and their ability to test and drive solutions that can be incorporated into daily operations to support decision-making position them as highly valuable [Mikalef et al., 2019] [Wamba et al., 2017]. We explore these teams in more detail in the next section.

2.1.2 Data-driven Teams

The goal of innovation, which is to generate novel ideas and propel change through creative experiments, gives rise to the idea of data-driven teams. However, this desire for uniqueness and inventiveness continuously adds ambiguity to the innovation process. Companies usually struggle with the fundamental questions of what they should innovate and how to approach the process most efficiently [Søndergaard et al., 2021].

In today's era, nearly every aspect of our life and work continuously yields data related to our preferences, behavior towards technologies, and services we use. As more data is made available, it also gives rise to creative ideas to comprehend and process data efficiently, empowering organizations to better understand customers, predict their behaviors, and tailor services that are geared to their needs. As more and more organizations realize this, their ability to collect, analyze, and act on this data has become a cornerstone of innovation.

This is the point at which data-driven innovation, an approach that draws inspiration and insight from

vast amounts of data from users, technologies, stakeholders, and environments, becomes essential. The concept of data-driven teams gained traction when Davenport and Patil [2012] called the data scientist 'the sexiest job of the 21st century,' highlighting the rise of roles focused on extracting value from messy, unstructured datasets. By exploring these rich data streams, organizations can reduce uncertainty and spark creativity throughout the innovation process [Luo, 2022].

Unlike traditional innovation methods that rely primarily on human intuition and expert judgment, this approach leverages cutting-edge technologies like machine learning and data science to automatically evaluate, validate, and generate innovative ideas. As illustrated in Table 2.1 from Luo's study, various approaches to data-driven innovation can be categorized based on their methodologies and application contexts. It demonstrates how data-driven and human-social methods differ when it comes to innovation activities like finding opportunities, evaluating them, coming up with ideas, and judging designs. On the one hand, data-driven methods use algorithms to speed up and expand the process of creating new ideas. On the other hand, human-centered methods focus on understanding user needs, putting themselves in their shoes, and using what they already know [Brown, 2009] [Björgvinsson et al., 2012].

Innovation Activity	**Data-Driven Methods**	**Human-Centered Methods**
Identifying Opportunities	- Unsupervised learning - Analyze data and its sources - Uncover hidden patterns - Widen the experiment space	- Interviews - Observations - Identify user needs, pain points, and challenges
Evaluating Opportunities	- Supervised learning - Process data to rapidly assess - Score potential solutions	- Intuition - Qualitative assessment to interpret needs and shape design direction
Generating Ideas	- Generative algorithms - Identify Patterns - Structure datasets to build concepts	- Brainstorming - Crowd-sourcing - Collaborative ideation techniques.
Evaluating Ideas	- Machine learning models - Train on previous experiments data to simulate or predict performance	- User testing - Feedback sessions - Iterative prototyping with stakeholders

Table 2.1: Comparison of Data-Driven and Human-Centered approaches in Innovation activities

It is critical to differentiate the concept of *data-driven innovation* from other terms such as data-based innovation. While *data-based innovation* typically focuses on leveraging data to develop new products and deliver value directly to users, the focus here is on making the innovation process itself more data-informed, iterative, and strategic. In this context, data is not merely used for reporting or compliance, but as a proactive tool to identify patterns, solve problems, and guide decision making at every stage of the innovation lifecycle [Lycett, 2013]. This approach is termed as *Data Science*, i.e., a multidisciplinary field that uses data engineering, machine learning, and statistical analysis to gain valuable insights from complicated and frequently unstructured datasets [Dhar, 2013]. The inventor of the term *Data science*, D.J. Patil, also reinforces that data science emphasizes more on predictive and prescriptive modeling than traditional analytics, which frequently concentrates on descriptive or diagnostic tasks. Many researchers like Hayashi [1998] have also explained it as a way of thinking that uses data to analyze real-world phenomena by combining statistics, data analysis, machine learning, and related approaches [Alzubi et al., 2018]. When innovation teams adopt these capabilities, they can automate complex processes and make forward-looking decisions [Cockburn et al., 2018]. This adoption is universal. For instance, innovation teams are increasingly using operational, behavioral, and contextual data patterns to define problems, rank opportunities, and refine possible solutions rather than depending only on gut feeling or static planning. For example, in the public sector, past data on environmental violations or inspection results can predict the non-compliance areas and create more informed regulatory processes. The US Chamber of Commerce Foundation has identified four key categories of data that are key in value creation across industries, including the government: demographic, economic, geographic, and transportation. This study on digitalizing public services highlights the capabilities of data and how it leads public sector innovation efforts by engaging various stakeholders, leading to improved organizational processes

and decision-making.

The concept of Innovation and Data Science (IDS) teams arises from the merger of the two well-established organisational forces: innovation teams and data science teams explored in this study above. In such teams, the exploratory competencies of innovation teams are coupled with the analytical, problem-solving capabilities of data science, enabling the group to both envision and implement solutions that are strategically novel and operationally viable. Through this review, we shed light on how an IDS team functions at the intersection of an analytical mindset and iterative problem-solving, efficiently using data as a catalyst for both operational and strategic innovation. The difference between human-centered and data-driven approaches highlights how innovation works in complex, data-rich environments that are constantly changing. The following section 2.1.3 expands on this fundamental knowledge by discussing the key areas on which a data-driven team focuses and the technological strategies that allow these teams to produce practical benefits.

2.1.3 Team Dynamics & Workflow

This section describes the structural composition and primary focus areas of Innovation and Data Science (IDS) teams, expanding on the idea that these teams are multidisciplinary enablers for innovation [Li et al., 2023]. IDS teams operate at the intersection of possibility and complexity, where success depends not just on technical competencies, but on how well people understand the purpose, collaborate across boundaries, and learn through iteration. The 5 P's of Data Science—Purpose, Plan, Process, People, and Performance [Richard, 2024] formed a basis for this idea. These components are essential to developing a useful framework that can evaluate, assist, and scale innovation projects, which requires a thorough understanding of the team's structure and workflow.

As highlighted by Uysal [2022], the structure of IDS teams typically comprises data scientists, ML engineers, domain experts, and a manager, each contributing to transforming abstract ideas into scalable and data-driven solutions. This cross-functional composition is not incidental but essential; as emphasized by Passi and Jackson [2018], embedding data science into organizational workflows requires effective collaboration between team members with divergent competencies and motivation. IDS teams do not follow a rigid structure or a top-down, bottom-up approach as several studies like Drach-Zahavy and Somech [2001] suggest. Depending on the project's scope and organizational procedures, the precise roles involved may change, but they typically involve data scientists and data or machine learning engineers. This structural flexibility, combined with their ability to manage uncertainty and use technologies, makes them central to driving operational changes.

Based on previous studies and several industry reports, these are some of the areas that reflect how an IDS generates value across different organizations:

- **Process Improvement**: Enhancing operational capabilities and driving automation (a method by which computers carry out tasks that were previously completed by humans [Parasuraman et al., 2000] is critical for any growing organisation.

  Davenport and Kirby [2016] estimates that this transition in automation, powered by AI systems, will be critical in supporting industries such as manufacturing, finance, and healthcare.

- **Operational forecasting and Predictive maintenance**: Forecasting tools for maintenance are widely used across sectors such as manufacturing, logistics, and energy. Machine learning libraries and time-series models become critical in predicting equipment failures or demand fluctuations, allowing organisations to optimize inventory, workforce, and maintenance schedules [Waller and Fawcett, 2013].

- **Product Development**: Data-driven teams increasingly fuel product development by turning historical data into tailored, customer-centric solutions. Big product-based companies make use of user-generated content, feedback loops, and product reviews. Researchers utilize these data as they develop Neural networks to mine them and look out for hidden customer needs, which later can be used to guide technical design choices [Luo, 2022].

- **Real-Time Monitoring**: Being able to gather, examine, and respond to real-time data is revolutionizing a variety of industries, from manufacturing and logistics to public transport and infrastructure governance. According to Kambatla et al. [2014], technology like computer vision,

real-time data analytics facilitates the quick processing of data as it is fed, flagging any discrepancy and enabling organizations to react to alerts immediately. This strategy exhibits how data can spur creativity and shorten product development cycles, and it is backed by the marketing and design literature [Yin et al., 2021].

To support these focus areas, the diverse roles work explored earlier in the section work together to create and execute creative solutions using data, analytics, and technology.

Workflow

The effectiveness of IDS teams is deeply tied to their ability to manage end-to-end workflows cleanly across the data science lifecycle, right from data acquisition and exploration to model deployment and monitoring. These workflows are not linear but iterative, requiring flexible structures that support continuous learning and adaptation [Saltz and Shamshurin, 2016]. A typical data science lifecycle followed by IDS teams involves six key stages: problem formulation, data collection, data wrangling, modeling, evaluation, and deployment [Sculley et al., 2015b]. This section provides an exhaustive description of how IDS teams work and the tools they use.

Core Technology Stack:
IDS teams rely on a modern data architecture to support the entire lifecycle of data science and machine learning (ML) projects, which consists of a sophisticated and multi-layered technological ecosystem. From data collection and preprocessing to modeling, deployment, and impact monitoring, their workflows cover every stage of the data lifecycle.

Key technologies span the full data pipeline:

- **Big Data Frameworks**: Technologies such as Apache Hadoop and Spark enable scalable and distributed data processing workflows [Chen et al., 2012].
- **Cloud Infrastructure**: Platforms such as AWS, Azure, and Google Cloud provide tailored AI/ML environments (e.g., SageMaker, Azure ML, and Vertex AI) that enable iterative development, scalability, and reproducibility.
- **Machine learning libraries**: These cloud services work with popular ML libraries (e.g., TensorFlow, PyTorch) to enable end-to-end data workflows, from processing to deployment.
- **Data Preparation Tools**: Libraries such as pandas, Dask, and Spark DataFrames are critical for efficient data cleaning and transformation, which is frequently cited as the most time-consuming step in the pipeline.

Davenport and Patil [2012] famously described the data scientist as "the sexiest job of the 21st century," emphasizing the role's growing importance across industries. Building big data capabilities is now widely recognized as a driver of organizational and process innovation [Ramadan et al., 2020]. Thus, understanding these technologies is critical to see how innovation teams build, iterate and deploy data-driven solutions in complex, rapidly changing environments.

Artificial Intelligence and Machine Learning:
AI and ML concepts are central to the innovation team's work. For over 60 years, computer scientists, engineers, researchers, students, and industry professionals have studied, implemented, and evolved these technologies [Alzubi et al., 2018]. John McCarthy first proposed AI in the 1950s, defining it as the science of creating machines with human-like intelligence. AI has evolved into a multidisciplinary field that includes computer science, mathematics, psychology, neuroscience, and others [Ongsulee, 2017].

Machine learning, a subset of artificial intelligence, allows systems to learn from data without explicit programming. It developed from pattern recognition and computational learning theory [Alzubi et al., 2018]. ML approaches range from traditional regression and decision trees to advanced neural networks and deep learning models.

Data Science Life Cycles:
Studies highlight that access to influential technologies is necessary, but not sufficient. Several articles in the literature explore the importance of incorporating AI/ML into decision-making processes and coordinating technical efforts with business or policy goals [Waller and Fawcett, 2013]. This has also been seen by the growing investments in both infrastructure and human capabilities [Ramadan et al., 2020].

Effective IDS teams connect technical functions (such as data ingestion, modeling, and deployment) with overarching innovation goals. They operate within a fundamentally different logic than traditional IT or software development units. Their work is characterized by uncertainty, experimentation, and iterative discovery [Ranawana and Karunananda, 2021]. Managing these dynamics requires workflows and lifecycles that can flexibly accommodate change and learning. Traditional software development lifecycles, such as the SDLC (Figure 2.3) involve completing each stage before moving on to the next [Alazzawi et al., 2023]. This lifecycle has further been adapted and led to the development of more mature and agile frameworks. For innovation-focussed projects, the linearity does not work as they benefit more from an adaptive planning approach. Müller and Turner [2007] emphasizes that innovation and R&D projects require a different ownership style and hierarchy than traditional projects, such as vision and adaptability. More recent research Lappi [2022] also demonstrates that an agile management style is better suited to projects with fixed goals and changing priorities. In this view, innovation is framed as a learning process, one that establishes direction while being flexible to adapt to new changes.

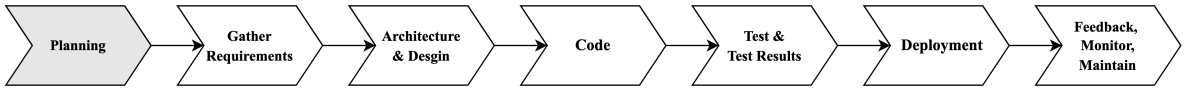


Figure 2.3: SDLC

While public sector teams are not frequently tasked with producing tangible business results, their importance is key for crucial tasks like inspection and automation, which necessitate more structure in their operations while maintaining considerable agility [McGrath, 2010]. Several data science lifecycle frameworks exist, each providing structured yet adaptable processes. Methodologies such as CRISP-DM, as seen in Figure 2.4, a six-phase data science lifecycle that focuses on a lot more business understanding, are crucial when measuring progress. Similarly, OSEMN, a leaner iterative model, offers a solid structure. Saltz and Shamshurin [2016] observe that teams frequently combine these approaches, beginning with CRISP-DM and then adapting agile elements or inserting steps such as "*Problem Framing*" (i.e., Defining the project scope) as necessary. Cockburn [2006] contends that no single method works in all situations, so teams combine elements based on context.

Figure 2.4: CRISP-DM Model

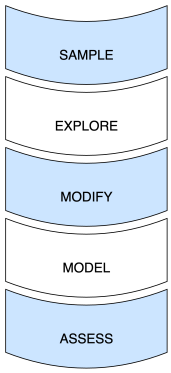


Figure 2.5: SEMMA Life Cycle

Data science workflows focus on continuous experimentation, cycling through data, feature engineering, model training, and evaluation in response to new insights. These projects usually have an iterative, cross-functional workflow that differs for each team. Singla et al. [2019] investigate these differences further, demonstrating how these practices differ between ML and non-ML teams. According to their findings, ML teams frequently build internal components that support larger systems, whereas non-ML teams typically develop software directly for end users. Several data science frameworks exist, but innovation-focused data science teams rely on well-established process models when creating or developing an AI/ML solution. Many teams use structured approaches that are comparable to popular frameworks explored above such as CRISP-DM (Cross-Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, and Assess), despite variations in implementation [Amershi

et al., 2019]. These approaches emphasize iterative feedback loops throughout the process and are essentially data-driven, despite minor variations in their phases and terminology.

Figure 2.6 depicts a typical data science workflow adapted from Blitzstein & Pfister's work [Mayo, 2016] where data preparation, feature engineering, model training, and evaluation are examples of tasks that frequently require going back and reviewing previous phases in light of new information or changing goals. This flexibility is crucial for data-driven innovation teams because the solutions they create are usually exploratory and flexible. Traditional software development techniques are inadequate for AI systems due to their inherent complexity and unpredictability, which emphasizes the need for adaptable, non-linear workflows that facilitate quick experimentation and ongoing improvement [Amershi et al., 2019].

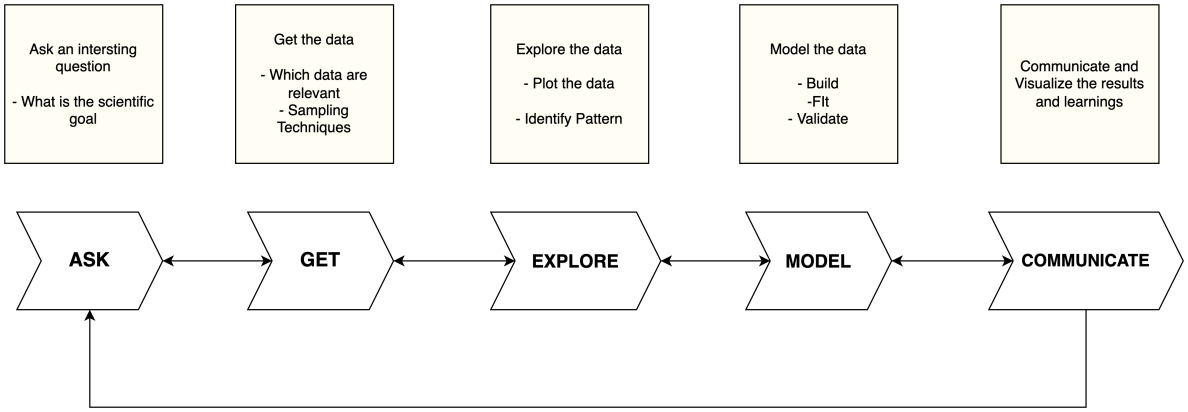


Figure 2.6: Blitzstein and Pfister's Data Science Process

The literature also emphasizes on tailoring the management approach to the type of project. Mergel [2016] and Schwaber and Beedle [2002] promote Scrum methodology for innovation projects, highlighting the circular nature of development and stakeholders' feedback. Whereas, traditional waterfall methods are best suited to projects with a more linear and fixed scope that are well understood and have very few changes. Ries [2011] Lean Startup and Blank [2013] build on these ideas, recommending hypothesis-driven build-measure-learn cycles for uncertain ventures. Many organizations have been using hybrid models that incorporate agile principles (frequent demo/review) along with some formal project planning. The main takeaway is that strict plan-driven management often clashes with the exploratory nature of innovation work.

2.2 Evaluating Innovation

Here, we introduce the concepts of value, impact, and outcomes in the context of innovation. In Section 2.2.1, we define and distinguish these terms. Section 2.2.2 discusses existing evaluation approaches and how structured frameworks can help decision making, tracking progress, and facilitating meaningful innovation.

2.2.1 Understanding Outcome and Value

Evaluation of results is a significant challenge and poses uncertainties across public sector innovation teams. There is a gap in the way these teams are perceived. An innovation-driven initiative typically requires more than seven months to progress from conception to end-user implementation. It should be noted that learning is a significant outcome of innovation efforts; a failed experiment often yields valuable insights that can potentially contribute to future success. Thus, it is first important to address what value means for an innovation team and how different scholars have measured it in the past.

According to Diazbeltran [2023]'s Data Science Dynamics: Data science teams typically move from Effort (planning, coding, data wrangling) to Outputs (models, documentation, APIs), which in turn enable Outcomes (business decisions or product improvements), ultimately generating Impact (such as increased revenue or reduced churn). Figure 2.7 showcases this as a sequential model, which goes from

effort to output to outcome to impact, and provides a useful mental framework for understanding the value chain of data science work.

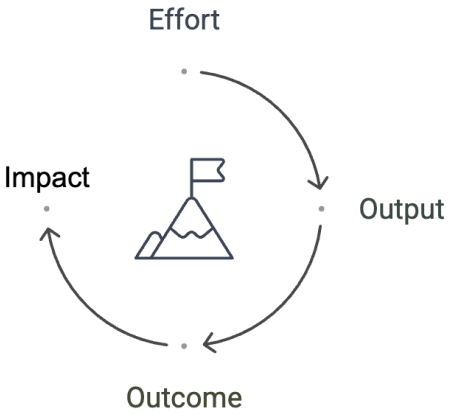


Figure 2.7: Mental Model of transition from work to change

Similarly, concepts like value proposition, viability, and feasibility of an idea are essential in innovation planning. These three lenses are specifically mentioned in design thinking literature as well [Brown, 2009], where feasibility refers to technical and operational capability, viability is seen across business or systemic sustainability, and desirability by the user and internal stakeholders to inspect the mission fit. Similar to this, Osterwalder and Pigneur [2010] Business Model Canvas subtly addresses viability by framing projects in terms of value propositions and revenue/cost structures.

However, defining value itself becomes part of the framework because public sector innovation frequently has several non-monetary value dimensions (equity, trust, and transparency). The definition of Impact within public-sector innovation is multifaceted (economic, social, and democratic) and changes over time, according to the common theme. To put it briefly, research on design and business models emphasizes how crucial it is to make these aspects clear; as a result, a structured framework ought to direct teams in defining success criteria (metrics & KPIs) and identifying feasibility constraints at an early stage.

2.2.2 Project Evaluation Frameworks & Related Approaches

Across the private and public sectors, several R&D evaluation frameworks exist with a focus on estimating and measuring the impact of their innovation efforts. Most of these frameworks are specific to the structure of the organisation and its capabilities, while others are more generic. In private enterprises, a project's benefits are typically measured in terms of financial value [Kohli and Grover, 2012] [Davern and Kauffman, 2000] or organizational performance (Sabherwal and Jeyaraj, 2015; Tambe and Hitt, 2012). Unseen gains, such as lower chance of reducing staff and greater innovation potential [Kleis et al., 2012] [Otim and Grover, 2012] are often overlooked. However, innovation efforts in the public sector generate not only the benefits mentioned earlier but also social and political benefits that add value to the community. The different types of benefits in the public sector make it difficult to quantify benefit targets during planning.

Corporate frameworks in the private sector, like the Ambition matrix, focus on the financial aspect, where it identifies what are the core, distinctive transformational innovation initiatives. Based on the same idea, Nagji and Tuff [2012] recommends a 70–20–10 split in innovation investment (70% core, 20% adjacent, 10% transformational). This approach highlights the monetary aspect where teams must focus on tangible benefits. This is further supported by the Theory of Constraints (TOC) by Goldratt and Cox [1992], which talks about the organisation driving innovation by identifying the bottlenecks and working on improving only those processes. As per Jurczyk-Bunkowska [2010], this approach ensures that investments made to improve the identified areas will benefit the business in measurable ways. Innovation-related solutions are also linked to improvement.

In another relevant concept, Chwastyk [2015] emphasizes the importance of evaluating innovation projects by recognizing the unique nature of their underlying tasks and stages. Unlike standard business operations, innovation processes are distinguished by their high degree of uniqueness, uncertainty, and inter-

disciplinarity. Each innovation initiative typically involves a non-repeatable set of tasks spread across multiple teams and timelines, with little precedent to guide success. This complexity, characterized by the need to balance creative ideation in the early stages with structured execution and decision-making in the later stages. This concept requires that innovation efforts be evaluated not only for their outcomes, but also for the structure and quality of the process.

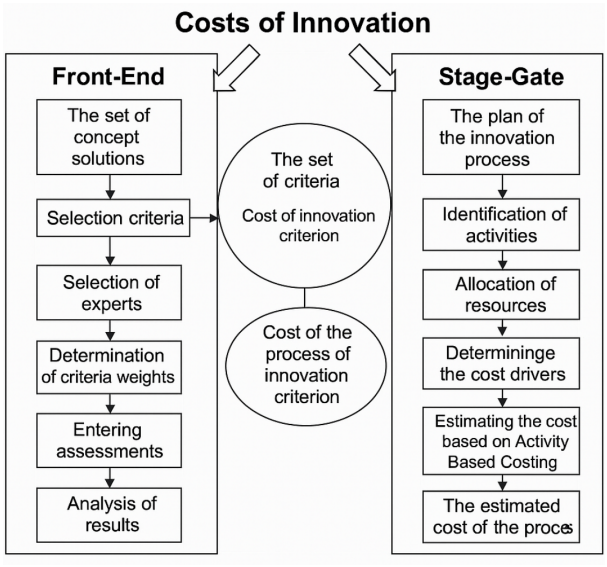


Figure 2.8: Phase-based evaluation of Innovation

As seen in Figure 2.8, the author Chwastyk [2015] breaks down this dual evaluation approach (also termed Activity-based costing) by separating the innovation process into two distinct but interconnected stages: the front-end and the stage-gate. The Front-End focuses on ideation and concept development, which is where the seeds of innovation are planted. Regardless of the limited information available at this point, early evaluations are critical. The assessment focuses on the potential benefits of proposed solutions, both tangible (e.g., cost savings, increased efficiency) and intangible (e.g., user satisfaction, strategic alignment). Because measurable outcomes are still a ways off.

As the process progresses to the Stage gate phase, the focus shifts from abstract potential to executional effectiveness. This stage is more structured and operational, focusing on speed, resource allocation, and discipline of decision making. The evaluation criteria in this phase expand to include feasibility, performance tracking, and risk management. Here, innovation is more than a concept; it is a project that must be managed, monitored, and piloted.

These two stages, each with its own set of characteristics, priorities, and challenges, form a comprehensive lens through which the value and viability of innovation initiatives can be meaningfully assessed. In summary, modern innovation frameworks combine financial planning, human factors, and evidence. While initial frameworks across private organisations focused on the monetary aspects, they provide guidance on how much budget to allocate, but later research shows how a team must be assessed for non-monetary benefits too. With a combination of these multidimensional evaluation practices, public innovation teams can balance dollars and motivation while building up proof that their efforts have an impact.

Assessment Scorecards for Evaluation

Assessment scorecard, a structured evaluation technique for project proposals, is proven to reduce subjectivity and promote transparent decision making [Kaplan et al., 1996]. Unlike informal reviews, a scorecard serves as an early warning system for problems and "offers a quantitative, standardized method for evaluating projects" [Kaplan et al., 1996].

This term has been further improved and referred to as "Balanced Scorecards" which is another prominent method widely used as a structured tool to support transparent project evaluation [Kopecka, 2015]. Cobbold and Lawrie [2002] defines it as a management tool that private or public sector companies can use to measure the performance of their enterprises, particularly concerning the growth strategies they have put in place. Kaplan et al. [1996] presented how scorecards can reduce subjectivity in decision-making

by offering standardized, quantitative assessments that can be used at the inception or post-completion of projects.

These scorecards incorporate characteristics such as strategic alignment or operational feasibility to assess feasibility and fit with organizational goals. Recent studies [Taufik et al., 2021] also emphasize capturing underrepresented dimensions such as user commitment, learning intent, and project maturity—factors depending on the organisational structure, often overlooked in traditional assessments that focused solely on technical or strategic fit. Some of the attributes that inspire the design of these scorecards are:

- **Strategic fit**: Alignment with organizational goals and innovation strategy.
- **Feasibility**: Access to the right technologies and data sources.
- **Project Understanding**: Expected benefit and Objective definition.
- **Complexity/risk**: Estimated technical or execution risk.

In the context of innovation project management, scorecards have not been explored in depth and are often seen as ineffective due to unclear evaluation criteria and a lack of contextual relevance [Spanò et al., 2016]. Yet, in high-uncertainty environments of innovation teams, structured evaluation is still possible if grounded in a deep understanding of team dynamics and challenges. Designing the right questions is key to surfacing these insights and enabling more informed decision-making.

Shehnar's Diamond Model:
Effective evaluation of innovation projects often requires multi-dimensional frameworks. Another influential approach is the Diamond Model by Shenhar and Dvir [2007] that classifies projects based on uncertainty and expected impact. It was originally proposed in the context of project management and innovation, and it defines four key dimensions to profile a project: Novelty, Technology, Complexity, and Pace. Each dimension captures a critical aspect of uncertainty and challenge in the project. It can be used as a tool to analyze the expected benefits and risks of a project and develop a set of rules and behaviors for each type of project. Figure 2.9 shows a representation of the model in terms of innovation.

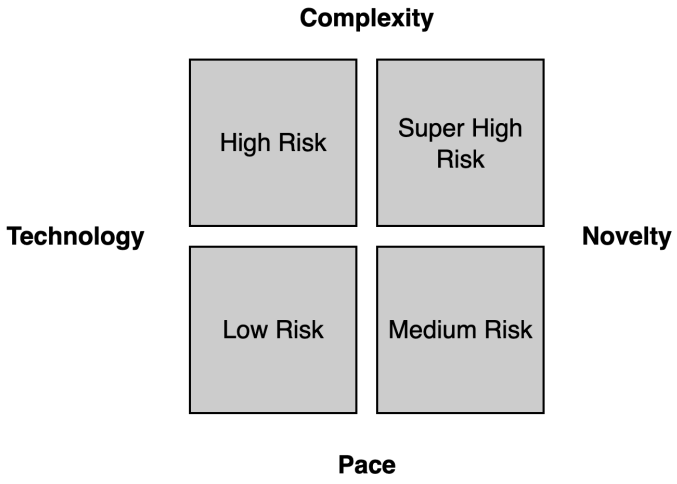


Figure 2.9: Shenhar's Diamond Model

Novelty here measures how new the product or innovation is to the market or users (ranging from incremental improvements to breakthrough innovations); Technology assesses the level of technological readiness or newness involved (from low-tech to high-tech ventures); Complexity gauges the complexity of the project's scope, which often correlates with how difficult an innovation is to implement in practice. The Pace lastly represents the urgency or time pressure for project's completion.

While the Diamond Model uses four axes (often visualized as a radar or diamond shape) to classify projects, many researchers and practitioners have also distilled innovation positioning into simpler two-dimensional matrices (quadrant plots) for ease of visualization. In the next chapter, we detail how we

adapt a similar method for the quadrant model for innovation assessment.

2.3 Dimensionality Reduction Techniques

As seen in literature on innovation evaluation frameworks, this process often involves analyzing complex data (i.e., numerous questions or indicators/patterns) and reducing it to key factors. This section explores key analytical techniques that support such process and assist with dimensionality reduction and pattern recognition.

Principal Component Analysis:
Principal Component Analysis (PCA) is a widely used unsupervised dimensionality reduction method that transforms high-dimensional data into a lower-dimensional space by identifying orthogonal directions known as principal components, with the aim of retaining most of the original variance [Jolliffe and Cadima, 2016b]. Originally developed by Pearson [1901] and later formalized by Hotelling [1933].

The main goal of PCA is to reduce the number of variables by projecting the original variables onto a new coordinate system where the axes (components) are ordered by the amount of variance they explain [Jolliffe and Cadima, 2016a]. In practice, PCA helps simplify complex datasets and is effective for visualization, which is crucial for this research. It is frequently used in both supervised and unsupervised learning tasks [Abdi and Williams, 2010]. However, a limitation is that the resulting components can be linear combinations of the original variables and may include negative weights, which complicates interpretability in certain contexts (e.g. a negative weight on a survey question is hard to explain). In our case, applying PCA to reduce the innovation scorecard data may not be ideal, as it could produce components that mix unrelated question categories or assign negative weights, making them harder to interpret. Our objective instead is to group questions into clear, positive-valued dimensions that align with meaningful and distinct innovation concepts.

Nonnegative Matrix Factorization:
To address the limitations of PCA, researchers have turned to Nonnegative Matrix Factorization (NMF), another widely used technique for dimensionality reduction and pattern discovery. Unlike PCA, which can produce negative values in its components and loadings, NMF factorizes a data matrix into lower-dimensional matrices under a non-negativity constraint. , making it particularly useful in domains where data naturally takes only non-negative values (e.g., document-term matrices, pixel intensities, or survey responses) [Sculley et al., 2015a]. This yields a "parts-based" representation where each component combines inputs additively and no negative weights appear. This makes NMF intensively being used for producing interpretable, lower-dimensional representations and is often preferred over PCA when all features are non-negative (as in survey scores) [Mihelcic and Miettinen, 2025].

In the context of innovation surveys, an NMF factor model would ensure that all question-to-factor weights are positive, making each derived factor easier to label (e.g., a factor may represent a theme like "technical complexity" with only contributing items and no contradictory negative signs). However, one of the main challenges lies in its non-convex optimization objective (i.e., objective function may have multiple local minima (or maxima), making finding the global optimum significantly more challenging than in convex optimization), which can lead to solutions that are not globally optimal or reproducible [Gillis, 2014]. Considering this drawback, standard NMF still does not fully meet our needs: it does not guarantee that each survey question loads primarily onto only one latent factor (i.e., it lacks dimension specificity). In other words, a single question could contribute to multiple factors in the NMF, blurring the conceptual separation between innovation dimensions.

Autoencoder:
Autoencoders are a type of unsupervised learning model that aim to discover meaningful low-dimensional representations of high-dimensional data patterns that traditional linear methods like PCA or NMF may overlook. An autoencoder is a type of neural network that learns to compress data into a low-dimensional latent representation and then reconstruct the original data from that encoding [Hinton and Salakhutdinov, 2006b]. Unlike PCA, autoencoders are capable of modeling non-linear relationships, enabling them to preserve more intricate structures in the data. In recent years, their application has expanded beyond image processing and natural language understanding into domains such as public policy, innovation analytics, and recommendation systems [[Bengio et al., 2013]. Figure 2.10 shows a typical workflow of an autoencoder. Here, the encoder is the layer that encodes a compressed representation of the input data

by dimensionality reduction. As data goes through the encoder layers, it is compressed by the process of "squeezing" itself into fewer dimensions.

Next, is the bottleneck (or "code") which contains the compressed representation of the input: this acts as both the output layer of the encoder network and the input layer of the decoder network. Here, the main goal of the design and training of an autoencoder to discovering the minimum number of features or dimensions that are crucial for the reconstruction of the input data. The representation, code from this layer goes to the decoder, which comprises of hidden layers with a larger number of nodes that or decode the encoded representation of data, this steps is crucial as it reconstructs the data back to its original form. This output is then compared to the "ground truth" which is basically the original input set to measure the effectiveness of the autoencoder. One key term here is the difference between the output and ground truth, which is called the reconstruction error.

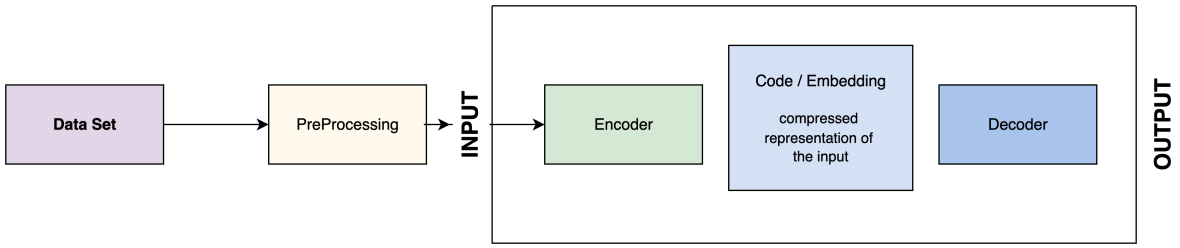


Figure 2.10: Workflow for Autoencoder

Variants like sparse or denoising autoencoders introduce architectural constraints that can enhance interpretability and robustness, making them especially suitable when domain-specific structure or feature disentanglement is important [Vincent et al., 2008]. Moreover, because autoencoders are trainable with custom loss functions and regularization, they provide a flexible framework for integrating task-specific objectives, such as preserving semantic groupings or enforcing sparsity in factor loading matrices—thus bridging the gap between statistical learning and domain-driven evaluation models.

Evaluation Metrics in Dimensionality reduction

1. **Reconstruction loss:** In unsupervised learning approaches, particularly autoencoders and NMF models, reconstruction loss serves as a key metric to understand how accurately the compressed term holds the original input information. It quantifies the difference between the input and its reconstruction, commonly using Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). This is crucial for validating whether the compressed latent features retain a meaningful structure for understanding or prediction [Hinton and Salakhutdinov, 2006a].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

2. **Root Mean Squared Error (RMSE):** RMSE penalizes large deviations more heavily, making it effective for models where extreme errors are undesirable. In the evaluation of innovation frameworks or project-based forecasting, RMSE provides a quantifiable indicator of predictive quality and accuracy of the reconstruction [Chai and Draxler, 2014].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

3. **Brier Score:** Used to assess the accuracy of probabilistic predictions. It is particularly relevant when models produce probability estimates for binary outcomes. As it takes into account both calibration and refinement, the Brier score has been found helpful in policy or innovation contexts where prediction confidence is just as crucial as accuracy. [Siegert, 2016].

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2$$

Where $\hat{p}_i \in [0, 1]$ is the predicted probability, and $y_i \in \{0, 1\}$ is the true label.

2.4 Gaps and Opportunities for a Framework

This review of the existing literature looked at the technology architecture, team structure dynamics, management practices, and structural challenges that come with IDS teams in public sector. Each of these methods provides valuable information on gaps and provides opportunities to develop a more structured framework.

Existing research on these practices reveals a mismatch between traditional project governance models and the exploratory and often unpredictable nature of data-driven innovation. A major case is the absence of structured intake procedures. Without clear criteria, ideas can overwhelm teams, making it difficult to prioritize and align with strategy [Van der Meer et al., 2021] [Kattel et al., 2018]. This highlights the need for an initial filtering or scoring mechanism.

Another gap is a mismatch between project types and delivery models. Scholars argue that Data and AI projects frequently necessitate iterations, ethical constraints, or adaptable governance, which rigid IT or stage-gate processes can restrict [Zhang et al., 2020]. We also explored how overly applying agile methodologies does not work well for all types of projects and is not a one-size-fits-all approach. This need is seen across the final phase of projects; when they mature, their ability to evaluate becomes critical. Quinn Patton and Patrizi [2010] proposes "developmental evaluation" to handle risks and unrealistic expectations in innovative projects. Many innovation labs lack measures for defining success metrics beyond broad objectives, leading to poor ROI calculations, a poor indicator of team performance. However, strategic tools like the Balanced Scorecard, are known to enable both a structured intake process and, together with the diamond model approach, can be used for visualising project positioning.

In addition, while dimensionality reduction techniques such as PCA and NMF have been used to simplify complex project data, they often struggle to assess the nuances of innovation teams and projects. This presents an opportunity to use constrained autoencoders, which can preserve these complexities while mapping projects into much simpler dimensions.

In conclusion, the literature on data-driven innovation teams demonstrates established practices and gaps. Each section explored above identifies opportunities for a structured framework that strikes a balance between structure and flexibility, providing templates and decision guides to enhance the project intake and assessment planning while being highly exploratory and diverse in nature of innovation work.

A clear, upfront framework will enable the team to foresee a project's trajectory, identifying early whether it will remain an exploratory proof-of-concept or be deployed, and also analyzing it once it is completed. Ensuring alignment between all stakeholders on what is strategically important for the team, prioritizing the right projects at the right time, and establishing explicit go/no-go checkpoints that protect capacity for the most promising work. In short, knowing "where the project will land" shifts innovation from ad hoc experimentation to a focused, value-driven pipeline that delivers measurable impact more quickly.

Chapter 3

Methodology

In this chapter, the structured approach used to develop the framework is outlined. First, the Design Science Research Methodology is described, followed by details about the techniques used for generating the questions and typology classification.

3.1 Design Science Methodology

We adopted the Design Science Research Methodology (DSRM), a problem-solving paradigm that focuses on developing and evaluating innovative artifacts [vom Brocke et al., 2020]. These artefacts include models, algorithms, frameworks, design principles, and methodologies.

Given the applied nature of this work, which is conducted in a real-world innovation lab, DSRM was an appropriate choice, as it emphasises problem-solving and practical relevance. The framework created by Peffers et al. [2007] divides the research process into six stages. We incorporate all the steps from the DSR process as shown in Figure 3.1 (adapted from Peffers et al. [2007]):

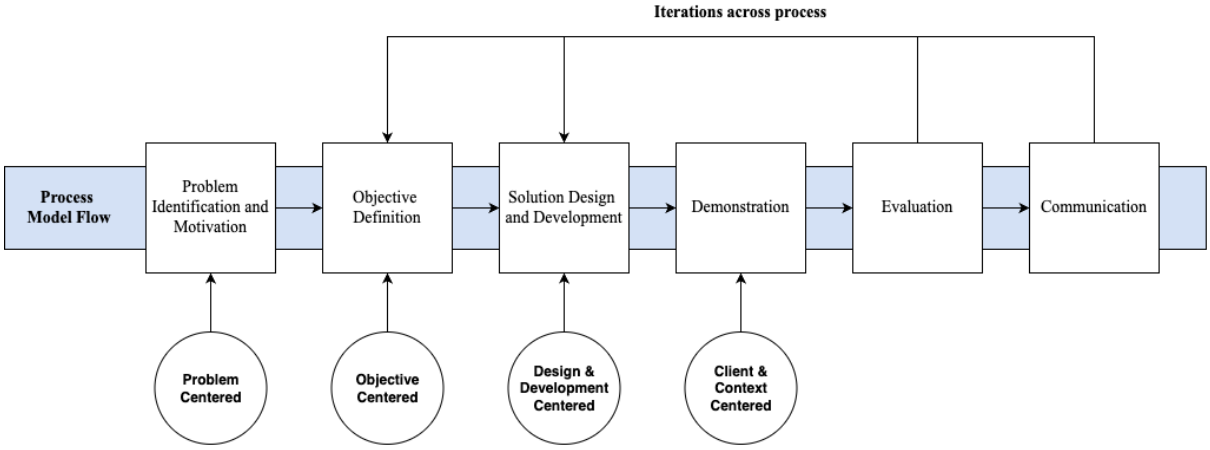


Figure 3.1: DSRM Model

1. **Problem Identification and Motivation:**

   "Define the research problem and justify the value of the proposed solution"

   The research began by identifying key challenges limiting the innovation team's effectiveness, most notably a lack of clarity about the impact of their work and the absence of a structured project intake and management approach. Through interviews with team members and analysis of past projects revealed a fragmented process was revealed in which innovation efforts were pursued longer than expected, often lacked clear exit strategies or alignment with larger strategic goals. Insights from literature also revealed key gaps in supporting the uncertain nature of innovation work. This phase contributed to the definition of the core problem statement and revealed the importance of

developing a framework to guide decision-making, increase visibility, and improve the long-term value of innovation projects.

2. **Define the objectives for a solution:**

   "Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible"

   Following the problem identification phase, and drawing on internal challenges and relevant literature, this study proposes a framework to support three critical stages of innovation work: project selection, execution, and post-project assessment. Each phase contains elements such as clear go or no-go criteria, success metrics, stakeholder alignment tools, and knowledge capture mechanisms.

3. **Design & Development:**

   During this phase, the desired functionality and architecture of the artifact are determined, followed by the creation of the artifact itself. This phase's main output is a management framework with components for intake scoring, execution guidance, and post-project reflection, which combines literature insights, internal pain points, and team feedback.

4. **Demonstration:**

   This exercise focused on using the framework to solve the problem. Following its concept, we applied the framework to previous innovation projects to assess fit, identify gaps, and demonstrate how it improves project focus, decision-making, and traceability.

5. **Evaluation:**

   The evaluation determines how well the artifact supports a problem solution by comparing the objectives to the actual results. We gathered feedback on usability and impact from team members and leaders, and we refined framework elements to improve prioritization, structure, and learning.

6. **Communication:**

   In this phase, the final tailored framework is shared with the IDLab team to improve their current workflow, encourage adoption, and ensure consistent application across future projects.

Below is a breakdown of some of the key data collection and analysis techniques used to design the scorecard, and the phase of DSRM they relate to

3.2 Scorecard Design Process

To precisely assess project viability at inception and accurately reflect on outcomes post-execution, a systematic approach is used to generate structured questions. The method combines qualitative and quantitative data from DSRM-driven steps, each detailed below.

In this section, we explore how each component provides essential insights that directly influenced the development of scorecard questions.

Data Source	**Purpose**	**Analytical Method**
Root Cause Analysis	Understand historical pain points and key causes/reasons	Root cause analysis (Fishbone, 5 Whys)
Semi-structured interviews (n = 15)	Identify key challenges and success factors	Thematic analysis and coding
Past project portfolio (n = 22)	Identify recurring patterns and bottlenecks	Pattern mapping

Table 3.1: Overview of data sources, their purpose, and the analytical methods used.

3.2.1 Interviews

During the initial phase of this research, semi-structured interviews were conducted with team members of the IDS team at the ILT, each team member specializing in AI/ML, data science, and agile method-

ologies. All 15 members across the department, including Data scientists, engineers, and Managers, were interviewed using a flexible question guide. In total, 15 in-depth interviews were conducted, guided by approximately 24 open-ended questions. This semi-structured format helped strike a balance between consistency, achieved through a predefined question guide which can be found in Appendix A, allowing for probing rich, contextual details. The questions explored the experiences and motivation of the participants when working in an exploratory setting, the specific challenges they encounter, and their strategies to overcome those obstacles. This technique directly contributes to the identification of the main issues and establish goals for the solution i.e. Step 1 and 2 of DSRM.

This promoted narrative responses. Follow-up questions ("How does your team adjust planning to accommodate uncertainties?") resulted in concrete examples and in-depth knowledge. These conversations also revealed a recurring pattern: individuals were highly motivated and skilled, but their working styles differed significantly. Some preferred structured plans, while others favored spontaneous, problem-driven exploration. This lack of alignment occasionally caused friction between vision and execution, particularly when moving between the ideation and delivery phases.

Furthermore, additional team role mapping exercises were conducted to delve deeper into participants' responses and gain a comprehensive understanding of their perspective on the team and each member's diverse competencies. Another theme discovered was a lack of clear user engagement. Many participants stated that end-users were either unclear or absent entirely, particularly in early-stage proof-of-concept projects. This limited the team's ability to validate usefulness while also reducing stakeholder visibility.

3.2.2 Thematic Coding

Following the interviews, a thematic analysis method is used to find trends in the transcripts. All interviews were recorded with participant consent, transcribed verbatim, and systematically coded. To ensure clarity, we focus on capturing the essence of the text rather than a word-for-word representation. Each phrase was assigned a short code. For example, the statement "We are given freedom, but that also means no one is protecting us; we have to prove our worth constantly" was coded as #proving_value, capturing the team's ongoing need to prove its worth and show its impact. Each code was categorized into more general subthemes that were connected to important categories relevant for problem identification and objective definition, actions related to step 1 & 2 of DSRM. Based on how frequently they occurred and how relevant they were to the framework's main goals, these subthemes were then either improved, combined, or removed.

This analysis facilitates the identification of two primary dimensions: (1) the team's perceptions of its structure, goals, and operational approaches, and (2) systemic conflicts and barriers hindering effectiveness. Together, these themes and dimensions provided a solid foundation for clearly defining the problem and guiding the subsequent solution objectives. These themes are discussed in more detail in the coming chapter and results. In Figure 3.2, the coding process adapted from [Williams and Moser, 2019] is shown.

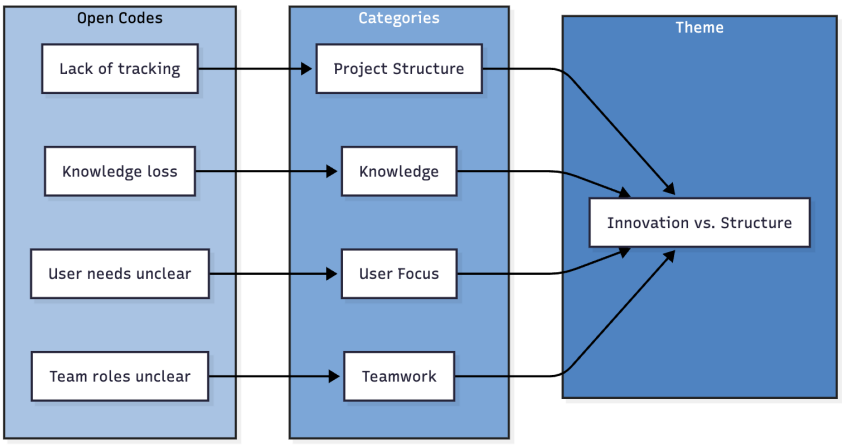


Figure 3.2: Coding Process

3.2.3 Deducing information from Interview data

The goal of this phase was to translate the insights into assessment questions, done not simply by summarizing common problems, but by identifying which pain points must surfaced early or late in a project lifecycle to increase project success. This is done in 3 key steps:

(A) Identifying core dimensions:

First, the entire set of codes and interviews is grouped to understand the characteristics that team members share as the reason for project's under-performance. This is motivated by two key dimensions: (1) strategic misalignments in how projects are planned and executed, e.g., goal ambiguity, unclear ownership, undefined success criteria, and project adoption; and (2) systemic conflicts that hinder project effectiveness, such as innovation potential, learning, and reusability, short-term experimentation vs. long-term value. These formed the conceptual backbone of the scorecard and were translated into questions that would allow these conflicts to surface early in the project lifecycle.

(B) Mapping dimensions to codes:

Each cluster (theme) of codes within the two dimensions is further broken down into specific groups of diagnostic gaps, i.e., the identifiable points of failure or delays that a question should address and help surface during project inception or final review. By linking real quotes and observed working habits (breakdown is provided in Appendix C), we turned common internal challenges into concrete questions. This made it easier for the team to reflect on how well-structured and innovation-ready the projects are. The table below demonstrates the complete mapping.

(C) Question Design:

Once the diagnostic gaps are defined, the next step is to translate them into questions. This step correlates with step 3 of the DSRM, which focuses on the design and development of the artifacts. For each dimension, the corresponding interview codes are reviewed to understand the specific challenges or gaps the team faced. The question is designed such that it surfaces these challenges in a project setting, either during planning (Pre) or at evaluation (Post). Later, we define what needs to be measured. For each dimension, we ask: What would a team need to know in order to judge whether this issue is being handled well in a project? For example:

If the challenge is "projects lack milestones," the relevant measurement is: Is there a clear time estimation for each phase?

Similarly, If the team struggles with "proving value," then we ask: How significant is the expected impact on ILT's operations, insights, or decision-making if the project succeeds? (Where, 0 = Minimal; 5 = Transformational). Some general rules for questions are:

- Questions must be rooted in a specific diagnostic gap/challenge, such as "redundant efforts," "lack of ownership," or "unclear success criteria."
- They must use a 0–5 scale to make teams reflect rather than using a yes/no binary.
- Encourage conversations, scorecard sessions should lead to some dialogue and discussion.
- Differentiate between types of projects, For instance, the challenge of "unclear success criteria" led to the question: "To what extent are success criteria loosely defined or expected to evolve during the project?" This helps differentiate research-oriented innovation projects from those with well-defined KPIs.

3.2.4 Root Cause Analysis

Following qualitative coding, the identified categories were then examined through a root cause analysis, motivated by techniques such as "5 Whys." The aim was to grasp the fundamental challenges and key causes behind them. This analysis also helps in planning what to fix and what changes to make.

This analysis helps systematically move from superficial symptoms to deeper underlying issues. Iteratively asking "why" enables us to identify underlying issues and trace back causal relationships [Andersen and Fagerhaug, 2002]. With each iteration, another layer of the issue is removed, ultimately revealing a core cause or group of causes that need to be addressed.

The detailed outcomes of the root cause analysis and how it is used are elaborated in the next chapter.

3.2.5 Stakeholder Identification

Across the interviews, a significant portion of the research focused on mapping key stakeholders involved in the IDLab's overall operation. This approach primarily belongs to the first two steps of DSRM and was crucial to better understand the current workflow, what each role entails, and how the team supports inspectors. It also helped to define who the framework is ultimately intended to serve, both directly and indirectly, as shown in Figure 3.3

The stakeholder mapping process was an important step in aligning the framework's structure with practical realities. Interviews and internal observations were used to categorize stakeholders based on their roles, responsibilities, and proximity to the operational, managerial, and implementation aspects of innovation efforts.

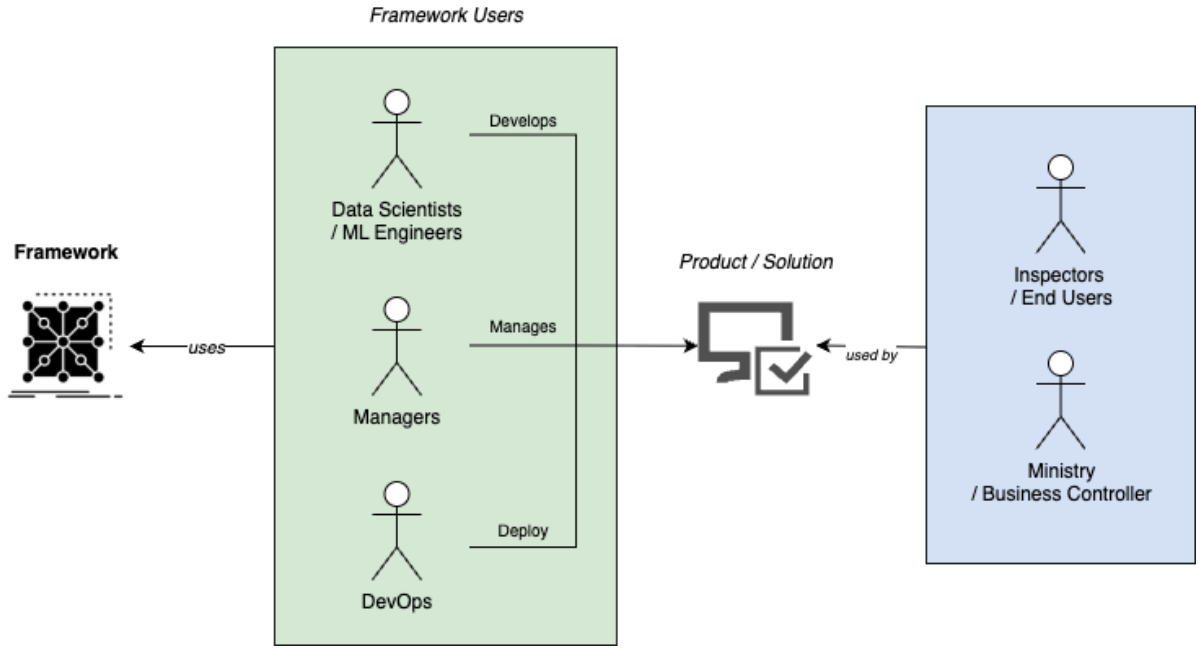


Figure 3.3: Stakeholders IDLab

The three primary stakeholder groups that were identified are:

1. Core Team Members: These individuals serve as the IDLab's operational backbone. They are the framework's most frequent and direct users, using it to organize their work, manage uncertainties, and align technical progress with overall project goals. However, rather than being a homogeneous team, the core members bring a wide range of competencies from various technological and disciplinary domains.

   Roles in the core team are differentiated based on their technical competencies and tasks, including:
   - DevOps, ICT, and Software Engineering Specialists
   - AI, machine learning, and applied data scientists
   - Privacy, Behavioral, and Social Data Scientists
   - Traditional Data Science Practitioners

2. Project managers and Department Head: These roles are key to bridging stakeholders, managing scope, ensuring cross-functional alignment, and advocating for the team's achievements. They are expected to use the framework to help make go/no-go decisions for new projects, track progress, and provide feedback on the expected outcomes.

3. Inspectors and Ministry Stakeholders: These include domain experts, inspectors, and policy officers from the Ministry who have an indirect impact on the outcomes of IDLab projects. Their physical work area is different from the core team members of the IDLab. While they do not directly interact with the framework, their expectations, feedback, and strategic needs shape how projects are scoped and evaluated, making them indirect contributors to the framework's success.

3.2.6 Project Portfolio Analysis

In addition to interviews, a structured study of 22 innovation projects was conducted. These projects ranged across domains and technologies, from AI/ML prototypes to dashboards for inspections. This exercise looked to reveal deeper patterns by scrutinizing:

- The type of innovation the team is working on, Radical or Incremental. This helps to determine which strategy is best suited for them.
- The maturity phase the project reached, which includes exploration, prototyping, validation, and deployment.
- Comparisons of expected and actual timelines to identify common bottlenecks, slowdowns, and accelerators.
- Team-level impact, considering not only the deliverables but also whether the team's work resulted in follow-up initiatives, knowledge reuse, or organizational change.

Each project was assessed against a consistent set of success metrics in dimensions such as value delivered, stakeholder adoption, and social impact, and analyzed using a pattern mapping approach. These were mapped using a binary (0–1) matrix and referenced back with interview findings to ensure that they reflected the real challenges and opportunities noted before. Recurring traits associated with either successful outcomes (e.g., high implementation uptake, measurable savings, reusability) or project challenges/failures (e.g., no stakeholder engagement, lack of defined success criteria) were identified.

This portfolio analysis was critical in ensuring that the framework responds not only to individual project needs but also to support teams to work on clear, meaningful, and impact-driven projects with a better understanding of where they're going and how to get there. We use this concept in the design chapter for the development of the scorecard.

Correlation Analysis: To further validate the consistency and relevance of the metric-focused dimensions, a correlation analysis was performed using these data from past innovation projects (See in Appendix B). We explore the interrelationships between various project dimensions, particularly those captured in the post-assessment phase.

3.2.7 Defining Scoring Mechanism

A 0–5 Likert scale is used to measure each question at both the start of the project (pre-project) and the end of the project (post-project). This scale was influenced by literature on technology readiness levels (TRL), maturity model scoring [Paulk et al., 1993], and prior evaluation frameworks in public sector innovation and digital transformation projects. The structure ensures that responses are not random but grounded in defined performance gradients.

Nemoto and Beglar [2014] advocates that scales with four to six points are best for balancing the accuracy of the measurements with the ease of use for the respondents. Scales with more than six points are less useful because they are harder for people to understand.

The six-point scale is interpreted as follows, and an example is shown in 4.4:

- 0 indicates the complete absence of the trait or an unresolved, high-risk condition.
- 1–2 represents initial, weak, or inconsistent evidence of the trait being present.
- 3 reflects a moderate, functional level, this is also the minimum threshold expected for a project to be reasonably scoped or ready.
- 4–5 reflects a strong presence of the trait in the project's pitch or outcome.

As described in chapter 2 on related work, where we outline the concept of innovation and data science teams as separate entities, we break down how innovation-focused teams are characterized by their capacity for novelty and experimentation, which we call innovation novelty. In contrast, data science or implementation-oriented teams excel at structured execution and using data to support decision making (traits that align with implementation readiness). This conceptual separation was further supported by insights during interviews with stakeholders and through analysis of project portfolios. We observed that some exploratory projects tend to fall along a spectrum: some demonstrate high, strategic novelty but lack clear plans for completion; others show operational strength but limited strategic ambition.

These observations highlighted the need for an assessment tool that could analyze innovation capacity and implementation readiness independently, rather than collapsing both into a single composite score.

3.3 Dimensionality Reduction and Latent Representation

To translate high-dimensional scorecard responses into an interpretable two-dimensional space, we apply dimensionality reduction techniques that preserve the conceptual meaning of each axis. Our aim is not merely to compress the data, but to map each project into a space where one axis reflects its innovation profile and the other its implementation profile.

Based on the 0–5 scoring scale of the scorecard, we aggregate 16 project features into two core conceptual dimensions: **Innovation Novelty** and **Implementation Difficulty**. This adaptation builds on the Diamond Model by Shenhar and Dvir [2007] (discussed in Chapter 2) and is informed by patterns consistently observed in our historical portfolio review, interview data, and expert input from stakeholders in IDLab.

Innovation Novelty (X-axis) Degree of novelty, ambiguity, and exploratory character of a project.

Implementation Difficulty (Y-axis) Clarity of the pathway to implementation, including planning, data readiness, and stakeholder support.

This two-dimensional typology allows projects to be located in one of four quadrants, each representing a qualitatively different challenge. The quadrant view supports visual profiling of project teams, identification of strengths and weaknesses, and tracking changes in perception between pre- and post-assessments.

The subsequent subsections describe two methods for positioning projects in this quadrant:

1. A simple, transparent *average method*.
2. A data-driven *grouped autoencoder* that learns the mapping from responses to the quadrant axes.

Both methods produce Innovation (X) and Implementation (Y) coordinates for each project, but differ in their assumptions and flexibility.

3.3.1 Average Method

As a straightforward reference method, we calculate each project's position in the typology by averaging responses within each conceptual group of questions. The Innovation coordinate is the unweighted mean of all items assigned to the innovation group; the Implementation coordinate is the mean of all items in the implementation group.

This method directly reflects the scorecard structure without any learned parameters. It serves as an intuitive point of comparison for the more flexible autoencoder approach.

3.3.2 Grouped Autoencoder

To derive the same two-dimensional representation while allowing for more flexibility in modelling relationships between questions, we use a **grouped autoencoder** (GAE). This neural network compresses questionnaire responses into a two-dimensional latent space and then reconstructs them, learning patterns directly from the data.

The model has two components:

1. **Encoder** – maps the original response vector $\mathbf{x}_i$ to a 2D latent vector $\mathbf{h}_i$:

$$\mathbf{h}_i = g(\mathbf{x}_i W_e)$$

   where W_e is the encoder weight matrix and $g(\cdot)$ is the *softplus* activation, ensuring non-negative values.

2. **Decoder** – reconstructs the original responses from $\mathbf{h}_i$ using the transpose of the encoder's weights:

$$\hat{\mathbf{x}}_i = g(\mathbf{h}_i W_e^\top)$$

   Tied weights reduce the number of parameters and improve interpretability.

The "grouped" structure incorporates prior knowledge: each question belongs to one of the two conceptual groups (innovation or implementation). A regularisation term, controlled by parameter θ, encourages each question to connect primarily to the latent dimension for its group. This maintains the interpretability of the two axes in terms of the scorecard design.

The model is trained to minimise a loss combining:

- Reconstruction error — encouraging accurate reconstruction of original responses.
- Group regularisation — encouraging alignment of feature–component connections with the predefined grouping.

Once trained, the encoder outputs Innovation and Implementation coordinates for each project, directly aligned with the conceptual axes of the typology.

3.3.3 Evaluation Measures

Two quantitative measures are used to assess the quality of the learned two-dimensional representation.

Reconstruction Error (RMSE) Let $X \in \mathbb{R}^{n \times d}$ be the original response matrix and $\hat{X} \in \mathbb{R}^{n \times d}$ the reconstruction obtained from the autoencoder. The reconstruction error is:

$$\text{RMSE} = \sqrt{\frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(X_{ij} - \hat{X}_{ij} \right)^2}$$

Lower values indicate that the two-dimensional representation retains more of the original information.

Average Method Reconstruction For the Average method, each project's two-dimensional coordinates are computed as the mean of its innovation-group questions and the mean of its implementation-group questions. Because the groups may have different sizes, the group means are obtained by normalising once by the number of questions in each group. The reconstruction $\hat{X}$ is then formed by assigning each project's group mean back to all questions in that group. RMSE is computed in the same way as above, now measuring the average deviation of each score from its group mean.

Brier Score Let $y_j \in \{0, 1\}$ be the known group label for question j and $p_j \in [0, 1]$ be the predicted probability that question j belongs to the innovation group. The Brier score for a set of m questions is:

$$\text{Brier} = 1 - \frac{1}{m} \sum_{j=1}^{m} \sqrt{(y_j - p_j)^2}$$

In our application, we first normalise the learned weights for each question so that they sum to 1 across the two latent components, allowing them to be interpreted as probabilities.

Each axis (innovation and implementation) has its own Brier score, computed by taking only the questions assigned to that axis and comparing their predicted probabilities to the "ideal" value of 1 for their own group. For example, for innovation questions ($y_j = 1$ for the innovation axis), $y_j - p_j$ is small when the axis assigns high probability to its own group; for implementation questions ($y_j = 1$ for the implementation axis), the same logic applies.

Higher Brier scores (closer to 1) indicate that the axis clearly distinguishes its intended group of questions from the other group.

Reconstruction Error (RMSE) quantifies how accurately the autoencoder is able to recreate the original questionnaire responses from the two-dimensional latent representation. Lower RMSE indicates greater accuracy of compression and suggests that key information is retained.

Brier score is used to evaluate how well the model separates the two questions bags, innovation and implementation. It checks whether the reduced data space reflects this distinction clearly, without needing a manual threshold. Full implementation of this approach can be found on Git here [Barata].

The results and findings derived from this validation approach are detailed and discussed in the subsequent Results section.

Chapter 4

Results

This chapter reports the validation and evaluation of the Project Assessment Scorecard using the GAE. We first present the results from each method that is used to design the scorecard and then present the refined set of questions that are produced for both the pre- and post-project scorecards, guided by the design principles and approaches explored in Section 3.2. We then compare GAE with a baseline average method, assess question-specific contributions, and examine the project trajectories in the Innovation–Implementation space for both pre- and post-assessments.

4.1 Project Assessment Scorecard Design

This section presents the results of each analysis (method) and how they led to the formation of each question on the scorecard.

1. Initial diagnosis

Root cause analysis, as shown in Figure 4.1, revealed two major bottlenecks. Each shows a blind spot because of which projects falter, and informs the design of questions to diagnose that problem.

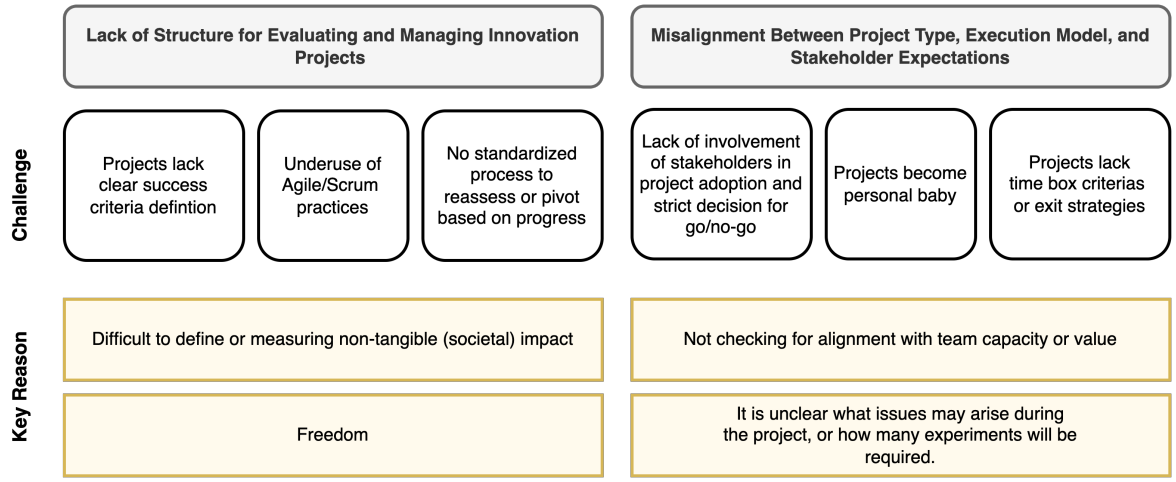


Figure 4.1: Cause analysis

(A) Lack of structure for evaluating and managing innovation projects:

This key challenge comprises three interconnected operational challenges:

- Unclear Success Definition: Making it difficult to understand project KPIs and manage it's scalability.

- Missing Agile/Scrum: Although not a necessity, it provides a structure for innovation while being flexible.
- Lack of standardized review processes: results in loss of knowledge built into previous projects.

(B) Misalignment between project type, execution, and stakeholder expectations:

This challenge sheds light on the difficulty in mapping the nature of the project, how they are executed, and what stakeholders expect in terms of results. It also manifests itself in three sub-areas:

- Low managerial involvement in the early stages, especially during adoption and critical go/no-go decisions, results in a disconnect between strategic intent and project execution.
- Project ownership has become overly personalized, with individual team members treating projects as their own "pet project".
- No defined exit criteria, resulting in projects/experimentation with unclear boundaries for success or discontinuation.

Two deeper underlying causes emerge through root cause analysis. First, innovation projects frequently aim for benefits that have a non-tangible societal impact, which, as seen in Chapter 2, is difficult to define and measure within traditional frameworks, resulting in uncertainty in planning and evaluation. Second, teams operate within unstructured independence, resisting ideas that may hinder autonomy in their approach towards experiments. This gives the idea that strategic value is not explicitly reviewed, complicating planning and the expected project outcome.

These challenges lead to the core problem to tackle: *the lack of a common, organized method for assessing and directing innovation projects.* Identifying causes helps map the assessment focus. For instance, many issues in the "Lack of Structure" category, such as poorly defined success criteria or lack of a pivoting mechanism, can be addressed by better early planning and criteria definition (PRE). For each root cause, a corresponding diagnostic question was developed as seen in Table 4.1.

Root Cause	**Frequency (n=22)**	**Example Projects Impacted**	**Diagnostic Need Identified**	**Linked Scorecard Question(s)**
Lack of goal clarity	15	Geo Chatbot, Putin's Shadow Fleet Risk Model, Manage O: Disk	Ensure project has a clearly articulated objective	Are the project's goals explicitly defined and understood by the team?
Undefined success criteria	14	GeoAI Big Bags Image Recognition, Web Scraping BRL 100/200 – Pilot	Define what success looks like before development	Are there clear and measurable success criteria defined prior to execution?
Redundancy or unclear value	13	Water Quality Webscrape, Content Scrape Innovation Transport	Demonstrate added value or novelty compared to existing solutions	To what degree does this project avoid duplication and offer something new or needed?
Weak implementation strategy	11	Geo-webapp Small Aviation, Drone Tracks #2 (Std Dev)	Assess whether an implementation path exists	Is there a clear plan or pathway toward implementation beyond prototype stage?
Ambiguous stakeholder ownership	10	Temporal Design from Dynamic Networks, Geo Chatbot, O: Disk	Ensure responsible stakeholder engagement and ownership	Is there a clearly identified stakeholder who owns or has requested the project?

Table 4.1: Root Causes Identified Across IDlab Projects

2. Framing Questions from Interview data

As explored in 3.2.2, qualitative insights from interviews revealed recurring challenges within IDlab. The interviews were coded thematically and analyzed using an inductive approach. Rather than starting with

predefined dimensions, the analytical process was bottom-up, allowing patterns and tensions to emerge from the data.

Table 4.3 shows sample questions used in the final scorecard, demonstrating how they correspond to the diagnostic gaps identified.

3. Portfolio pattern mapping

An analysis of 22 historical projects, grouped by success and failure outcomes, was conducted to identify critical project traits that correlate with project performance. The complete mapping can be seen in Appendix B.

This metric-based approach allowed for capturing nuances. We observed:

- Projects that aligned with ILT or ministry-wide themes (Column R: "Did the project align with Ministry priority areas or team goals?") were more likely to reach production or be reused (Column L/K).
- This pattern validated a diagnostic gap identified in previous steps (projects with low strategic fit often failed to progress).
- Leading us to design a question to ensure early alignment and avoid repeating projects that failed due to low strategic relevance.

Correlation Analysis: The resulting correlation matrix (Appendix B) provided several meaningful insights.

Firstly, a moderately strong positive correlation ($r = 0.63$) was observed between projects that had a clear and structured planning approach and those that were completed within the estimated time from idea to outcome. This finding suggests that projects with well-defined plans and structured execution strategies are more likely to adhere to their timelines, supporting the inclusion of structured planning as a key pre-assessment criterion.

Interestingly, strategic alignment showed weak negative correlations with documentation (-0.34), cost and time savings (-0.24), and time prediction (-0.29). This indicates that while alignment with ministry priorities remains essential, it does not inherently predict operational success and may, in some instances, be associated with more complex or exploratory projects that pose delivery challenges.

4. Scoring

Scorecards are operationalized using the 0-5 scale both at the beginning and end of a project. The scoring is interpreted in table 4.4. It is key to note that even though the same scoring scale is used at both phases, the interpretation changes. For example, a score that moves from 1 to 4 suggests significant organizational learning or successful implementation. Similarly, if a project starts at 4 and ends at 2, it indicates a disconnect between expectations and reality. suggesting that despite initial confidence, the project did not deliver as expected.

Trait Being Assessed	What a 0 Means	What a 5 Means
Problem/Need Clarity	No problem identified or vague idea	Clearly defined and well-understood problem or need
Strategic Alignment	Project conflicts with ILT's mission or goals	Strong alignment with ILT's strategic direction
Novelty / Added Value	Fully redundant or repeating existing work	Entirely new approach or significant improvement
Impact Potential	Minimal or unclear expected benefit	High potential to transform operations or decisions
Access to Tech/Data	No access or knowledge of needed tools/data	Full access and readiness to use tools/data
Innovation Level	Routine improvement	Radical or novel approach
Exit Strategy	No plan for handover, scale, or closure	Clear exit or transition strategy in place
Time Definition	Timeline vague or undefined	Timeline broken into clear, realistic phases
Level of Iteration	One-off delivery expected	Designed to evolve through multiple test-and-learn cycles
Experimentation Openness	Fixed outputs, minimal learning flexibility	Highly exploratory, open-ended learning path
Definition of Success	Success is emergent or undefined	Clear KPIs and success criteria from start
Compliance / Governance Readiness	Legal or ethical risks unresolved	Fully compliant and reviewed

Table 4.4: Scorecard traits with scale definitions.

Building on the concept of separation of questions explored in 3.2.7, each question was classified into one of these two categories, which we call "bags". For example, questions assessing the novelty of the idea, its potential to shift paradigms, or its alignment with long-term strategic vision were grouped under Innovation. In contrast, items focusing on technical feasibility, rigid planning, or stakeholder alignment were grouped under Implementation. Importantly, no item was intended to simultaneously contribute to both dimensions.

By establishing these two “bags” of questions, the scorecard becomes more than a collection of questions; it becomes a structured assessment tool capable of revealing not just whether a project is promising, but in what way.

4.2 Results

4.2.1 Final Scorecards

As a key output of this research, the scorecard was created to support structured evaluation of innovation projects. This section presents the final form of the scorecard, grounded in the iterative process defined by Design Science Research Methodology (DSRM) as outlined in the methods 3.2. The scorecard was then applied and validated using an autoencoder-based evaluation setup. The scorecard captures two key dimensions:

Innovation potential: the extent to which a project introduces novel ideas or approaches.

Implementation feasibility: the perceived ease or readiness with which the project can be executed.

The scorecard supports two different critical areas for project understanding: the exploration area, which is supported by the pre-questionnaire intended to guide evaluation at project inception, and the reflection aspect, catered by the post-questionnaire, designed to support reflection after project implementation.

Each question was refined based on triangulated evidence from three core qualitative sources: root cause

analysis (RCA), thematic coding from stakeholder interviews, and cross-case pattern mapping. The table below captures the final form of these questions.

Table 4.5: Final Pre Scorecard Questions by Dimension

#	**Question**	**Innovation**	**Implementation**
1	To what degree does this project avoid duplication and offer something new or needed beyond existing tools/processes?	Yes	
2	How much novelty or innovation does this project introduce to ILT's existing operations or knowledge base? *(0 = Routine improvement; 5 = Radical or novel direction)*	Yes	
3	Is the project expected to go through multiple development or research cycles? *(0 = One-shot delivery; 5 = Iterative with frequent adjustment)*	Yes	
4	To what extent is the project designed for experimentation and iteration? *(Scale: 1 = Minimal, clear outcomes / 5 = High, open-ended exploration)*	Yes	
5	Are the problem, data, and solution space well-defined and scoped?	Yes	
6	What is the level of uncertainty or risk involved in delivering intended value? *(Scale: 1 = Low risk, known path / 5 = High risk, unproven concept)*	Yes	
7	How much does the project aim to generate learning that can be shared or reused by other teams or future projects?	Yes	
8	How much of the project's value lies in learning, discovery, or generating new insights?	Yes	
9	To what extent does this project align with ILT's strategic direction, goals, or organizational priorities?		Yes
10	Is the problem or opportunity clearly defined and understood?		Yes
11	Do we have access to the right technologies and data to implement this solution effectively?		Yes
12	To what extent are success criteria loosely defined or expected to evolve during the project? *(0 = Success is emergent and learning-focused; 5 = Clear KPIs and outputs are defined)*		Yes
13	What is the level of uncertainty in integrating the solution into existing systems or workflows? *(0 = High uncertainty or major unknowns; 5 = Integration pathways are well-understood and feasible)*		Yes
14	Is the project currently in a validation phase with clear hypotheses to test?		Yes
15	What is the likelihood of this project producing a tangible output or deliverable?		Yes
16	How strong is the level of stakeholder commitment or sponsorship? *(0 = None; 5 = Strong champions with ongoing support)*		Yes

Table 4.6: Post-Project Scorecard Questions by Dimension

#	**Question**	**Innovation**	**Implementation**
1	Did the project generate internal learning or insights that influenced future projects or ways of working within ILT?	Yes	
2	To what extent did the project offer a novel approach or significantly improve upon existing tools, processes, or knowledge?	Yes	
3	How innovative or original was the final approach or solution relative to ILT's existing practices and knowledge base?	Yes	
4	Did the project involve iterative development or experimentation over time, rather than a single fixed delivery cycle?	Yes	
5	Was the project primarily focused on open-ended exploration or discovery, rather than predefined outcomes?	Yes	
6	Was the project designed for long-term impact or institutional change, rather than short-term results only?	Yes	
7	Did the project result in measurable improvements or influence in ILT's operations, decision-making, or data capabilities?		Yes
8	Were the original project objectives clearly defined and appropriately scoped from the outset?		Yes
9	Did the project result in a tangible output that was deployed, adopted, or reused in other teams or contexts?		Yes
10	Was the project primarily focused on developing a concrete solution based on a well-understood business or user need?		Yes
11	Were the major risks and uncertainties identified and effectively managed throughout the project lifecycle?		Yes

The final version, filled with respondents' input, was subsequently applied in the experimental phase to evaluate its ability to measure innovation potential and implementation readiness of projects.

4.2.2 Dimensional Mapping and Scorecard Validation

Theta Sweep:

Before comparing the mapping methods used for validation, the GAE described in Section 3.3.2 was evaluated across a range of θ values to identify the value that best balances the reconstruction accuracy (RMSE) and separation of question groups (Brier score).

Figure 4.2 shows the trade-off between RMSE and Brier score for θ values in the range $[0, \ldots, 1]$. Higher Brier scores indicate a better separation of Innovation and Implementation bag questions, whereas a lower RMSE indicates a more accurate reconstruction of responses.

The results show that $\theta = 1$ provides both perfect bag separation (Brier = 1.000) and a low reconstruction error and was therefore selected for all future experiments.

Comparison with Baseline Method:

Using $\theta = 1$, we begin to compare the autoencoder reconstruction quality with the simple baseline average method described in Section 3.3.1. Figure 4.2 presents the Brier Score vs. Reconstruction Error (RMSE) trade-off curves for both methods, separately for the pre- and post-assessment questionnaires.

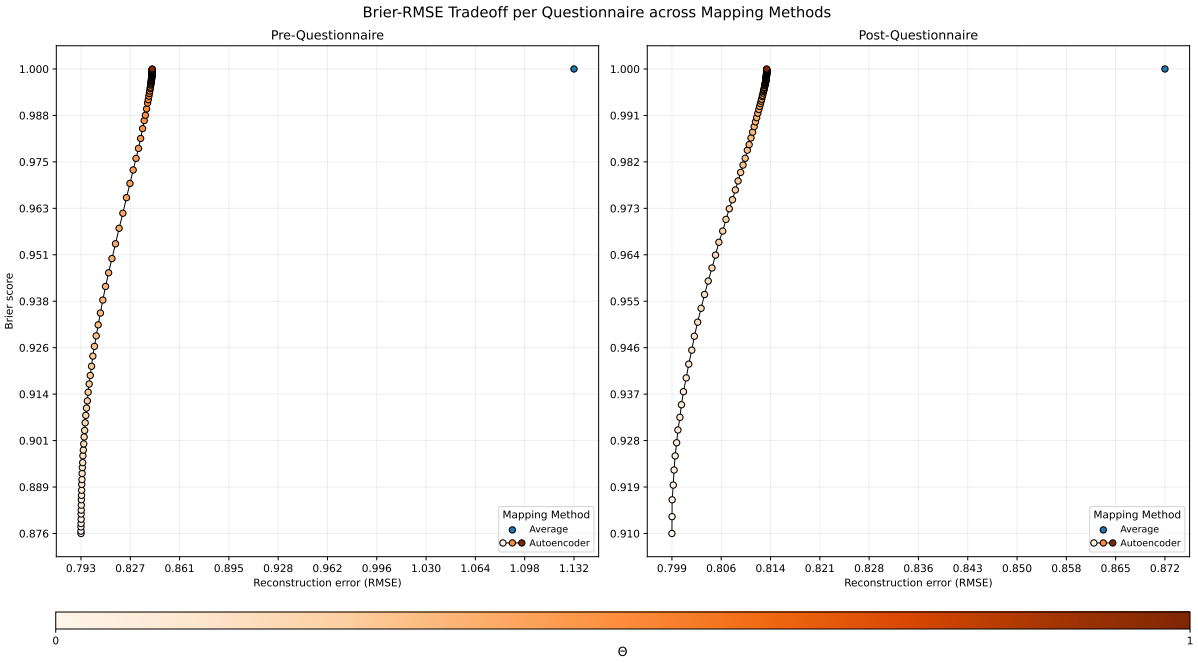


Figure 4.2: Brier-RMSE Tradeoff per Questionnaire across Mapping Methods

Table 4.7 summarizes the performance of our autoencoder method against the average baseline method. We focus on RMSE value, and as seen in the figures across both, the Autoencoder achieves lower RMSE values and outperforms the baseline in all cases. In particular, the pre-assessment RMSE improves from 1.13 (average method) to 0.84 (autoencoder), while the post-assessment RMSE improves from 0.87 to 0.813. This validates our unsupervised learning approach. By learning optimized weights from the input questions, the model produces a more accurate and meaningful representation of projects, effectively capturing the underlying themes.

Mapping Method	RMSE (Pre)	RMSE (Post)
Average	~ 1.132	~ 0.872
Autoencoder	~ 0.84	~ 0.813

Table 4.7: Comparison of Autencoder RMSE with Baseline

Autoencoder Mapping and Question Contributions

Next, we analyze the per-question contribution to each bag produced by the autoencoder. In figure 4.3, each bar represents a question, and the height indicates its relative contribution to the dimensional mapping. The color coding (red and blue) reflects alignment with either the innovation or implementation axis, and a higher bar means the question was more significant in determining a project's position along that axis. It is crucial to note that the Brier score for both axes is 1.000, indicating that the forced separation of bags was considered, confirming perfect alignment between the learned mapping and the predefined question groupings.

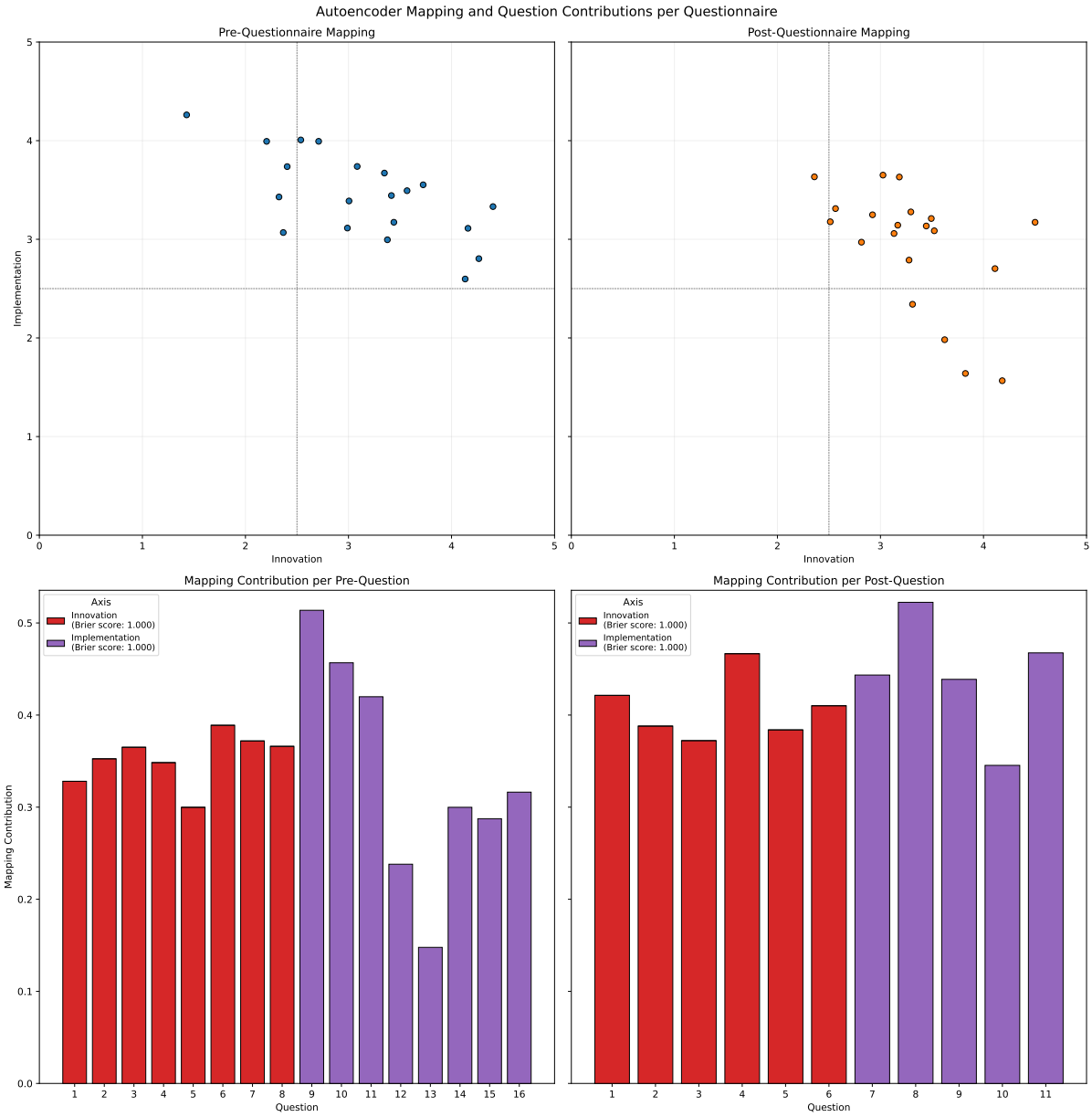


Figure 4.3: Autoencoder Mapping and Question Contributions per Questionnaire

It conveys that questions designed for innovation predominantly contribute to the Innovation axis, and similarly to Implementation.

Project Mapping in Quadrant

Finally, using the scores generated by the autoencoder for both pre and post scorecard, each project was plotted within a learned 2D quadrant space (Innovation $\times$ Implementation), with arrows indicating the

shift from pre to post evaluations (Figure 4.4).

The average shift across all projects is $+0.17$ in Innovation and -0.51 in Implementation, with an average Euclidean distance of 0.65 between the pre and post positions. We further interpret these results in the discussion and break down what this shift reveals about the team's evolving perception and its implications for future projects.

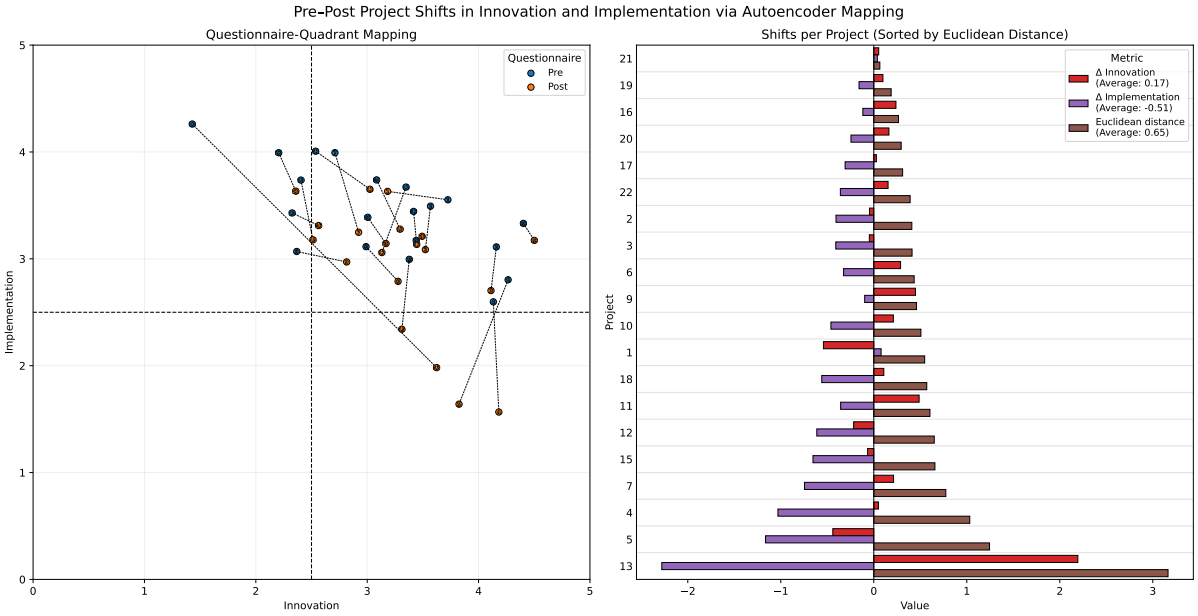


Figure 4.4: Project Shifts over time via Autoencoder Mapping

Dimension	Coded Themes & Quotes	Diagnostic Gap	Example Scorecard Question	Stage
Project Structuring	#unstructured_project_lifecycles "We've worked on projects for months without knowing when to stop."	No defined milestones, time-boxing, or stopping criteria	To what extent are milestones, deliverables, and exit points defined at project start?	PRE
Strategic vs. Exploratory Tension	#diverse_approaches, #exploratory_vs_usability "We call it 'organized chaos'... too much user focus limits innovation."	Confusion between exploratory vs user-driven goals	Is the project type (e.g., exploratory vs applied) clear, and is the approach appropriate to this type?	PRE
Impact & Evaluation	#lack_of_impact_measures "We don't have a clear way to quantify success... failure is a success too."	Absence of outcome-based metrics and post-hoc evaluation practices	Are intended outcomes and success metrics defined clearly and revisited post-completion?	PRE + POST
Ownership & Guardrails	#autonomy_without_guardrails, #self_sufficiency "We got an infinite amount of freedom... nobody else was taking responsibility for us."	Team autonomy without shared accountability or explicit ownership structures	Are roles, responsibilities, and ownership clearly defined and aligned within the team?	PRE
Intake Clarity & Role Understanding	#innovation_understanding, #project_selection_criteria "People think we are just data collectors... that's someone else's job."	External misunderstandings about IDLab's role; poor intake filtering	Has the project been selected based on a clear fit with the team's innovation mandate and expertise?	PRE
Internal Alignment	#cognitive_diversity, #team_perspectives "This diversity is our strength, but it's hard to align."	Strategic/technical/user perspectives misaligned at the start	Were project goals discussed and co-framed across strategic, technical, and user lenses early in the process?	PRE
Growth Linked to Value Proof	#incremental_team_growth, #proving_value "Every time we proved value, we asked for more people."	Organizational growth and resourcing tied to showing tangible value	Was the project's value communicated internally or externally to support long-term investment or adoption?	POST
Learning & Reuse	(Implied from structural model and lack of structured evaluation)	Lack of post-project reflection and reuse	Are project outputs, learnings, or prototypes documented and available for reuse across other initiatives?	POST

Table 4.2: Mapping IDLab Interview codes to diagnostic gaps & questions by project stage.

Diagnostic Gap	**Scorecard Question**	**Scale Type**
Lack of time-boxing or milestones	There's a clear time estimation done for each phase of the project	0 = Ambiguous → 5 = Well-defined
Difficulty proving value	How significant is the expected impact on ILT's operations, insights, or decision-making if the project succeeds?	0 = Minimal → 5 = Transformational
Objective/Goal Ambiguity	How clearly defined is the problem or opportunity being addressed?	0 = Unclear → 5 = Well articulated
Innovation misunderstood as data support	To what extent does this project align with ILT's identity (strategic direction, organizational goals and culture)?	0 = Not aligned at all → 5 = Fully aligned and reinforces core priorities
Tension between exploration and user delivery	Expected level of experimentation and iteration?	1 = Minimal → 5 = High, open-ended

Table 4.3: Mapping of diagnostic gaps to scorecard questions and scoring scales.

Chapter 5

Discussion

In this chapter, we interpret the results from the Autoencoder experiments and discuss their implications in terms of innovation project assessment and decision-making. The discussion is structured in three phases: (1) Framework design intent and how the results support it, (2) Autoencoder-based validation and insights, and (3) Strategic implications for innovation management. We explicitly connect the findings with the three research questions outlined in Chapter 1 and examine how the results of the final Project Scorecard, in combination with dimensionality reduction techniques, can guide innovation teams.

5.1 Supporting Innovation Teams with Structured Tools (RQ1)

The overarching goal of this research was to explore how innovation teams can be supported by a structured framework that informs decision-making across the project life cycle from inception to closure, enhancing the team's ability to assess the value and viability of the project and its outcome. We are addressing a clear problem: innovation teams often make early-stage project decisions based on subjective perception rather than structured evidence. These decisions, while strategically critical, are often overlooked and frequently unstructured.

To address this gap, we developed the Project Assessment Scorecard, a structured tool consisting of a pre-assessment and post-assessment scorecard. This design directly responds to the first research question, RQ1. Its development was grounded in prior research, analysis of past projects, team member insights, and a root cause analysis, each of which is detailed in Chapter 3.

Scorecard
The initial version of the two scorecards comprised 23 and 11 questions, respectively, which were iteratively refined and reduced to 16 (for pre-) based on evaluation outcomes from the autoencoder-based validation approach. The model highlighted questions that added little value, suggesting that they could be removed or rephrased to improve alignment across both scorecards. The final version of the two scorecards is shown in Table 4.5 and Table 4.6. The questions were designed around four criteria: Strategic fit, Feasibility, Project understanding, and Complexity or risk associated (explored in Chapter 2).

Each question was carefully grouped to serve a clear purpose i.e., either capturing the innovation potential or assessing implementation readiness. We ended with 16 questions on two conceptual "bags" representing these two strategic dimensions:

Innovation: perceived novelty, originality, and strategic value of the project.

Implementation: perceived feasibility and implementation readiness.

Autoencoder Validation
This grouping of questions was further validated via the autoencoder approach; Figure 4.3 provides evidence that these groupings were meaningful. We can interpret that during pre-assessment, contribution weights were more dispersed, indicating that respondents' mental models did not yet fully align with the intended separation (between innovation and implementation features at that stage), whereas, in post-assessment: contributions became more concentrated within their assigned axes, suggesting that

teams learned to differentiate between innovation potential and feasibility of implementation through project experience.

This confirms that the structured bag approach is not just based on theory (conceptual clarity in measurement tools is essential to ensure consistent evaluation [Mankins, 2009]) but also backed by the model's results, directly addressing RQ1.

Impact
Using a structured scorecard early in the project helps teams think more clearly about the core strategic fit of their work [Spanò et al., 2016]. It encourages them to consider not just what they aim to achieve, such as exploration, novelty, or added value, but also how realistic it is to build, deliver, given the available resources, stakeholder support, technical setup, and any legal or ethical constraints.

This framework will support teams by enabling:

1. Clearer expectation setting – Across projects, while managers have a more general goal ("this project should help us improve customer experience"), developer teams frequently move with a hypothesis or a set of assumptions ("if we can train this model to 90% accuracy, it should work"). Without having a structured mechanism to align both their perspectives often drift apart. The framework encourages both parties to define success and key results in a quantifiable, mutually understood way. By starting a conversation that brings technical viability and managerial intent into alignment it helps teams avoid the common pitfalls of pursuing different objectives all along. Clearly stating the expectation from the project.
2. Early discovery of blind spots – Teams often fall in love with the solutions they envision at the beginning of a project, sometimes missing critical nuances (missing right data, ethical clearance was missing) associated with that approach, and only discovering them later. These failures are not unique; they occur frequently in industry pilots and innovation labs. Such "obvious-in-hindsight" risks are brought to light earlier by the framework, which functions as a structured checklist. Teams are prompted to pause and identify gaps that they might otherwise overlook in their enthusiasm by asking focused questions across dimensions.
3. More grounded planning – Innovation teams operate with a lot of freedom and are often very optimistic. Teams trust their approach to be the most novel, believing that they can create value and scale quickly. However, this results in overcommitment, such as scheduling a project for four months when integration alone actually takes six. By clearly linking goals across each sprint, iteration, and forcing to definition of an exit strategy, the framework introduces accountability without reducing creativity and freedom of innovation teams. It assists groups in defining the boundaries between their desired and achievable goals. There have been promising prototypes that failed due to ill-defined timelines and inadequate follow-up phases. This framework provides visibility into both aspects, understanding where the project will go to plan better and also analyzing the approach at the end of the project.

By applying the scorecard before and after the project (completion), this design of the scorecard also enables the move from static evaluation to a reality check, helping teams reflect on expectations, adjust plans, and track how project perceptions change over time. This shift is reflected upon in the autoencoder results, which are discussed in the following section.

5.2 Validation and Strategic Mapping of the Scorecard (RQ2, RQ3)

This section interprets the results from the Autoencoder validation of the project scorecard and the mapping of projects. Section 5.2.1 compares model performance against a baseline and interprets the results achieved (RQ2), while in Sections 5.2.2 and 5.2.3 we interpret the reduced Innovation and Implementation space to understand project trajectories, question contributions, and portfolio patterns (RQ3).

5.2.1 Method Comparison

The Autoencoder was compared against a simple average baseline to understand whether our designed scorecard structure could be empirically validated and whether ML could offer any considerable improve-

ments in reliability and interpretability. This directly addresses RQ2 — How can machine learning be used to validate and improve the questions of a project assessment scorecard?

As illustrated in Methods and Results, the baseline represented the basic manual approach to interpret scorecards, where each question within a "bag" is weighted equally and averaged to produce the X and Y scores. While transparent, this method assumes equal item contribution and offers no means to verify the groupings' consistency. The autoencoder, by contrast, learns optimal weights for each question by forcing a constraint. In this, the model doesn't freely reassign questions; it works within the predefined bag constraints. In turn, making sure the evaluation assesses both the model's performance and the framework's grouping logic.

Performance
The Brier-RMSE trade-off curves in Figure 4.2 clearly demonstrate that the autoencoder achieves much better reconstruction performance compared to the baseline in both pre- and post-assessment datasets. In the pre-set, RMSE drops off from 1.13 (average) to 0.84 (GAE). In the post-set, the RMSE fell from 0.87 to 0.813. These differences indicate that the GAE is better able to preserve the information contained in the original question responses when compressing them into the two latent dimensions (X and Y).

These results confirm two things. First, the scorecard grouping is consistent between projects. Because if the predefined groupings were poorly aligned with actual patterns in the data, forcing the model to stay within them would have caused noticeably worse reconstruction (higher RMSE) and lower association with a particular bag. Since the performance stayed strong, it suggests the groupings are not an artificial separation; they reflect meaningful structure in the dataset. Meaning the model could still capture and preserve variance crucial for defining projects while respecting the Innovation/Implementation split. This reinforces its credibility as a structured assessment tool and supports its potential for consistent application across diverse projects.

Second, the autoencoder's performance compared to the simple average approach highlights the value of a data-driven approach for handling multidimensional inputs. It reconstructs better while preserving the intended Innovation/Implementation split. This indicates that the method is not only effective at reducing dimensionality but also in retaining the integrity of the underlying constructs, ensuring that the compressed dimensions still reflect the bags. This preservation is essential for such an evaluation framework, where interpretability and alignment to core dimensions are essential for project adoption.

To address the framework's second objective, we tested consistency and reliability by comparing the autoencoder to the simple average baseline. The strong performance under $\theta = 1$ indicates that the scorecard captures meaningful, consistent patterns across projects. In the next section, we further explore how the question contributions and autoencoder mappings reinforce the clarity of the Innovation–Implementation groupings across both pre and post phases, supporting the scorecard's robustness and interpretability.

5.2.2 Autoencoder Mapping and Question Contributions

The contribution plot in Figure 4.3 shows for each scorecard, how much each question contributes to the final X or Y score in both phases. First, since θ-regularization was used as a constraint, we examine its rationale and implications.

The decision to use $\theta = 1$ is related to the role of the simple baseline average method. However, earlier experiments were also conducted with multiple θ values, including $\theta = 0.5$ and $\theta = 0.0$.

We observed,

- In Lower θ (e.g., 0.0–0.5): The model achieved lower RMSE in some cases, but at the expense of conceptual interpretability. Questions began to load on both axes, while reconstruction improvements were seen, but it violated the separation of innovation and implementation axes critical for the scorecard. It would have deviated a lot from the theoretical foundation of the scorecard and made it difficult for managers to interpret.
- $\theta = 1$: The model was forced to obey the original bag grouping, making sure it only learns the relative importance of questions within their assigned bag. This is crucial for the baseline method. Now, each question within a bag has equal weight, and the final score is a straightforward mean. This makes the comparison between the autoencoder and the baseline structurally fair: (a) Both

methods operate under the same conceptual grouping constraint. (b) Only difference is that the autoencoder learns optimal weights from the data, whereas the baseline fixes them equally.

This is important because if θ was relaxed, we would no longer be testing "can learned weights outperform equal weights within the same conceptual structure?" Instead, we would just be testing a different model structure altogether.

Interpretation
This regularization and these findings led to our actual contribution mapping. Several patterns are evident from Figure 4.3. We interpret these by exploring two perspectives: first, what these findings mean as a practical tool for innovation teams, and second, as a validation tool for researchers and managers developing structured project evaluation frameworks.

First, From the practical perspective of the teams that use the framework, the contribution analysis transforms the scorecard from a static measurement tool into an interactive diagnostic instrument. By quantifying the influence of each question on the final Innovation (X) and Implementation (Y) scores, the model reveals the specific levers that determine a project's strategic position within the two-dimensional mapping seen in the top half of Figure 4.3. This transparency allows teams to move beyond a score: it shows which items drove their position, so they can target feasibility gaps or strengthen the value proposition where it matters.. For example, a project with a low Innovation score can identify whether this is driven by weak novelty in the technological approach, redundant approach, or lack of experimental fit, each linked to specific questions in the "Innovation" bag. Similarly, a drop in Implementation score post-assessment can be traced to elements related to operational readiness or stakeholder support, or adoption. This interpretability directly supports the overarching goal of the framework, that is, enabling informed and targeted decision-making at both the inception and post-assessment stages of innovation projects.

Across the contributions observed, we also see some discrepancies. In some places, contributions were more diffuse (suggesting less stable mental models at inception), and few questions influencing that particular axis less than the others. Some pre-set questions contribute disproportionately (e.g. Q5, Q6 in Innovation; Q9, Q13 in Implementation), while others show more or less equal influence. This highlights specific items and is useful for refining the scorecard further in future iterations. In contrast, in the post-set, contributions become very similar, suggesting that the learning during the project helps teams distinguish more clearly between the two axes. We bind this with the actual shift observed and explore further in 5.2.3. For researchers and managers, this change in contribution patterns can serve as markers of maturity during project evaluation, giving teams useful feedback on their development.

Secondly, from the framework developer's perspective, the contribution analysis serves as a quality assurance method for the scorecard itself. As described in section 5.1, we can employ this technique to iterate the question framing, taking away the low-contribution ones (which may be redundant, poorly understood, or misaligned with their intended dimension). Similarly, stable high-influence questions validate their critical applicability in defining the dimensions well, reinforcing their retention in the scorecard. The relevance of this approach and impact is further shown at the end of section 5.2.3, where we explore how the combination of project mapping (shift) and the question-level contributions offers a powerful decision-support tool for innovation teams.

5.2.3 Project Trajectories Pre to Post

The last key element of the framework's third objective is to make project traits and outcomes more visible by mapping them into a 2D space defined by the latent dimensions of Innovation and Implementation. The pre–post trajectory plots in Figure 4.4 display the position of each project before and after completion. Each point represents a project in the latent space, and connecting lines indicate the direction and magnitude of change over time. Blue dot is the pre-placement, and orange signifies the outcome.

While we explored the observed shifts in the results, the average directional movement suggests that they are consistent across projects and large enough (mean distance = 0.65) to matter in practice, suggesting that these movements are not just noise but reflect significant changes in how projects are perceived over their lifecycle, or that teams are not estimating properly what/where a project entails from inception to conclusion. We interpret these shifts through the lens of perception change documented across innovation projects, as literature suggests this pattern is consistent with known dynamics in innovation projects

that teams often begin with optimistic assumptions about feasibility, only to revise them downward as practical challenges emerge [Dan Lovallo, 2003].

Because this study was based within a live operational environment of the IDLab team, the observed pre–post project shift carries immediate organisational relevance. They are not hypothetical projections, but instead a mirror to how the team perceives and navigates its own innovation portfolio over time.

The downward shift in implementation scores across most projects is particularly telling. It is probably the case of optimism bias [Lovallo and Kahneman, 2003] that appears to influence responses, technical feasibility challenges are underestimated, and are rated more generously. As projects progress, the realities of data complexities, availability, regulatory compliance, and organisational alignment become clearer. The drop in Implementation, thus, reflects experience-based adjustments, not necessarily project failure.

Conversely, the slight upward drift in Innovation scores suggests that during execution, teams become better able to articulate and defend the novelty of their work. This may be due to the refinement of the objective, insights from prior experiments, and a proof-of-concept through early prototypes or stakeholder feedback. For the IDS team, this is an encouraging sign that team members underestimate how innovative something will actually end up being and overestimate how easy it will be to implement.

Example from IDLab portfolio:
Project 13: Content Scraper: Initially positioned as a High Implementation / Low Innovation initiative, the project's pre-assessment indicated strong delivery feasibility and a modest novelty profile. However, post-assessment revealed a steep drop in Implementation score, moving the project toward the low–low quadrant. This means that although it seemed simple at first, unexpected issues (like problems getting the right data or meeting compliance rules) made it much harder to complete. The main takeaway is that future projects like this should test technical feasibility early on to avoid nasty surprises later.

Project 7: Manage O disk (Document Categorization (NLP)): That focused on exploring NLP techniques and categorizing documents started out strong in both innovation and implementation, and it mostly stayed that way. Its innovation score stayed the same, and its implementation score only dipped slightly. It stands as a positive control case within the projects tested, demonstrating how well-planned projects from the beginning, with the right people and resources in place, can keep a project on track without losing its innovative edge.

Similarly, other IDS teams can use this framework and the trajectory plots as a practical tool to monitor project progress and identify potential issues early. For example, if a project's Implementation score shows a sharp decline, it signals the need to reassess feasibility assumptions and resource allocations before risks escalate. Similarly, a decrease in Innovation may indicate challenges such as scope creep, loss of originality, or a shift toward more conventional approaches. For managers, it is a means to gain insight into the overall health of the organisation's innovation efforts. A consistent downward shift in implementation, for example, across projects might relate to broader capability gaps or an overly ambitious project pipeline.

Tailoring Project Management Approaches
These shifts are also critical because they inform not just where a project currently sits, but how it should be managed going forward. Literature in adaptive project management and innovation governance [Cicmil et al., 2006] emphasizes that the project lifecycle must be aligned with the project core vision and its task structure. This project-process matching also follows Shenhar and Dvir [2007]'s advice: the life-cycle model should be chosen based on the project's quadrant. In other words, type drives process.

Our framework operationalises this principle by suggesting predefined project management approaches for projects that fall under the four quadrants that are visible in the plots. The justification for the recommended approaches for the identified project types is shown in Table 5.1. The selection is guided by project traits and supported by literature.

(1) **Exploratory Research**:

Projects falling under this quadrant are highly novel and expected to follow a circular and iterative project lifecycle. With an aim to explore new technologies and methods without having a rigid and fixed scope, the problem is not clearly defined, and the objective often changes as new information emerges.

For these types of projects, [Gothelf and Seiden, 2021] proposes a dual focus agile approach that allows for early exploration (problem-solution fit) as well as lean validation and structured hypothesis testing.

This approach includes concepts like Timeboxing, low-cost prototypes, and discovery sprints that helps to manage risk while fostering creativity.

(2) **Technology Adaptation**:

These project's score highly on the implementation axis. Their primary tasks revolve around the integration and adoption of new technologies to transform the existing workflows.

Since the innovation risk is low and execution clarity is high, traditional waterfall or phase-gate models work best, but due to the uncertain nature of data sources and frequently shifting needs it is recommended to adopt scrum, as its iterative cycles provide structure and adaptability while still supporting efficiency, compliance, and timely delivery [Cooper, 2007].

(3) **Process Innovation**:

Projects placed in this quadrant possess a good mix of both innovation and implementation traits. These projects follows a systematic approach to novelty within established operational contexts, implementing new ideas or method and often rethinking internal workflows. These projects evolve through iteration but require alignment with strict operational guidelines that are more linear.

A hybrid approach is appropriate, combining Agile's flexibility with the discipline of stage-based reviews [Boehm and Turner, 2013].

(4) **Incremental Innovation**:

A project that evolves continuously while having a fixed objective falls into this criterion. These projects aim to support existing processes or methods in place, improve performance, and maintain scalability of the solutions. Some projects may also involve automating repetitive tasks, but they are usually short in duration, low in risk, and have minimum overhead. For these projects, a lightweight project management approach with quick cycles and direct ownership is ideal. Thus, we suggest Scrum, Kanban boards with strict timelines to track progress.

Table 5.1: Justification of tailored project management approaches

Project Type	Suggested PM Approach	Key Characteristics	Justification from Literature
Exploratory Research	Agile Exploration (Lean Startup, Design Thinking + Scrum)	• High uncertainty, exploratory scope • Iterative sprints, Circular tasks • Problem statement loosely defined	• Managing innovation uncertainty through iterative loops [Ries, 2011] • Managing ill-defined problems [Brown, 2009] • Agile Scrum planning for learning based projects [Highsmith, 2004]
Technology Adaptation	Scrum / Stage-Gate	• Well-defined scope • Low novelty, high implementation clarity • Documentation-heavy, structured delivery	• Stage-gate for controlled implementation [Cooper, 1990] [Cooper, 2007] • Scrum for rigid scope [Highsmith, 2004]
Process Innovation	Hybrid Agile (Agile Delivery, SAFe)	• Moderate uncertainty and iteration • Needs structure for delivery • Cross-functional coordination	• Hybrid models to balance agility and uncertainty [Ambler and Lines, 2016] • SAFe enables scalable coordination in innovation [Suomalainen, 2016]
Incremental Innovation	Kanban / Continuous Improvement	• Small, operational enhancements • Repeatable linear tasks, low complexity • Focus on flow efficiency • Minimal planning required	• Kanban emphasizes visual workflow and flow control [Anderson, 2010]

5.3 Limitations

5.3.1 Limitations in Scorecard and Question Generation

The development of the scorecard was intrinsically shaped by the operational environment of the IDLab team in which it was deployed. The conceptual "bags" of questions, as well as the measurement and wording of individual questions, were informed by stakeholder interviews, and the team's current workflow and processes. While this ensured good contextual fit, it also introduced limited generalization of the scorecard as the questions may not directly transfer to other organisational settings without modification.

Additionally, the composition of the question set is constrained by practical considerations, including survey size, stakeholder interest, and willingness for assessment. These constraints may have led to under-representation of certain other strategic verticals (e.g., organisational readiness (ORL)), potentially limiting the scorecard's broader influence.

Depending on self-reported data also imposes some limitations. Scores for Innovation and Implementation

reflect team perceptions, which can be influenced by optimism bias, overstated novelty, and Unclear understanding. Although the pre–post design was designed to capture directional changes and it does mitigate some bias, the final values, however, may still carry some perception-driven deviation.

Finally, the framework validation didn't include any formal psychometric tests to check if it measures what it's supposed to or if results stay consistent over time. Although we indirectly tested consistency using the autoencoder, adding traditional validation methods can help strengthen the scorecard's reliability even more.

5.3.2 Limitations in Evaluation Using the Autoencoder

The autoencoder used to analyze and interpret multidimensional data from the scorecard has its own set of methodological constraints. This section examines the practical limitations of the framework.

First, the model was trained exclusively on data from the IDLab team's portfolio, limiting its external validity. Without retraining on more other project sets, the latent space and learned weight structure might not generalize because they are optimized for the specific distribution of responses in this dataset.

Second, the sample size (number of projects) and project diversity (placement on quadrant) used sufficed to demonstrate proof-of-concept, but it might be a bit modest for ML standards. This limits the statistical power to identify other, more detailed patterns.

Third, while the choice of $\theta = 1$ successfully enforced the separation between Innovation and Implementation questions, it also constrains the model to serve and work according to the predefined conceptual idea of the scorecard. The interpretability does appear high, but may limit the model's ability to discover other new emerging relationships that do not align with the original "bag" definitions. highlighting the importance of considering other strategic dimensions revealed during the pattern mapping exercises (Appendix B).

The autoencoder is also optimized for reconstruction instead of prediction. While it does not estimate the outcome, it does still capture structural consistency within the data. Predictive modeling could be added to the framework to increase its usefulness in further decision support. Also, as seen previously, because the autoencoder relies on self-reported data, it's subject to bias.

5.4 Future Work

This section will first explore some ways to further refine the framework, and then offer ideas for expanding the research into new operational contexts and applications.

5.4.1 Scorecard Design Refinements

Two major areas that can be refined in this framework revolve around the limitation of generalization of the scorecard and developing it into a working product for the ID Lab.

Firstly, as the key future direction lies in generalising the scorecard better, so it can be applied beyond IDLab. The structuring of question sets and weighting could be modularised for different sectors, project types, and team focus. Adjustment to the scorecard to accommodate the team's technical language and perceived organisational support would allow the framework to retain the same performance. Enhanced by more rigorous psychometric validation tools like reliability testing, will further ensure that the scorecard's structure remains robust across different team/project settings.

Secondly, beyond methodological improvements, future work will also focus on translating the framework into a functional product for the ID Lab's operational use. The envisioned solution, prototyped during this research (shown in Appendix D), is designed to integrate directly with the Confluence workflow. This setup will allow project teams to complete the scorecard digitally prior to any pitch, with instant autoencoder-based mapping, quadrant classification, and recommended project management strategies available within the same interface. Such integration will ensure that the framework becomes not only a research contribution but also a living decision-support tool embedded in the Lab's innovation workflow process.

5.4.2 Enhancing Machine Learning Evaluation and Prediction

On the machine learning side, there is potential to extend the current autoencoder approach into a hybrid evaluation–prediction tool. One interesting direction is to add a supervised prediction layer, such as a linear regression model or a more sophisticated algorithm (e.g., gradient boosting, random forests), trained on both autoencoder-derived dimensions and project metadata. This would enable the framework not only to map Innovation and Implementation trajectories but also to predict key project outcomes and the likelihood of achieving stated goals at the pitch.

Finally, this model can be made more powerful by expanding the dataset to include projects from multiple teams and organisations, allowing the autoencoder to learn more general patterns while fine-tuning to specific contexts. This can transform the scorecard to become more universal in assessing innovation, making it much easier to apply at a larger scale.

Chapter 6

Conclusion

This research aimed to design, validate, and operationalise a structured framework for evaluating innovation projects at both their inception and their conclusion. Many existing evaluation techniques often fail either because they're not based on a clear, well-defined idea of what they're trying to measure, or because no one has properly tested whether they work in practice. Using IDLab as a case study, and by understanding their operations and managerial needs, the framework developed here aims to address that exact gap by focusing on two specific dimensions critical for the IDS teams. It combines a scorecard with ML–based validation and reduction techniques to guide the team and tailor management strategies accordingly.

To develop this framework, a design science research methodology was applied, which led to three key components: (1) an insightful structured scorecard, (2) a machine learning–based validation and dimensionality reduction approach, and (3) quadrant-specific project management recommendations.

The three research questions were addressed as follows:

- RQ1: How can structured pre- and post-assessment questions be designed for IDS teams to assess the value, viability, and feasibility of smart technology projects?
  This was achieved by the development of the scorecard, building around the Innovation and Implementation dimensions and structuring into clear "bags" of questions. Together, we capture both value and feasibility, providing a consistent way to assess projects.
- RQ2: How can machine learning be used to validate and improve the questions of a project assessment scorecard?
  The Autoencoder, with $\theta = 1$, showed lower reconstruction error and similar predictive performance compared to a simple average baseline. This showed that the scorecard's structure is consistent and also highlighted specific questions that could be improved or deleted.
- RQ3: How can dimensionality reduction help simplify complex project data and guide project management decisions?
  The autoencoder's 2D mapping made it possible to visualise the placements of projects, shifts in their positioning from pre- to post-assessment. This visualization provides deeper insights at the individual project level, which, when studied properly, reveals characteristics that can be linked to project management approaches that best fit that project.

From a scientific perspective, this work has combined a constrained autoencoder model with a qualitative insight-based assessment tool, creating a clear and interpretable connection between ML tools and established concepts in innovation management. From a more general practical standpoint, it offers a ready-to-use decision-support system with a clear route to being deployed in the IDLab's daily operations. The framework is designed so that assessment leads directly to action, building a strong workflow procedure that measures a project, diagnoses its strengths or weaknesses, and recommends tailored management approaches. Although it was developed for a specific team, the underlying methods can be applied more widely.

Future research could focus on adapting the scorecard for other functional departments, expanding the dataset to support cross-team learning, and building the planned digital prototype for seamless

integration into existing team workflows. Taken together, these steps would position the framework not only as a solid academic contribution but also as a scalable, practical tool for managing innovation portfolios.

Bibliography

Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2 (4):433–459, 2010. doi: https://doi.org/10.1002/wics.101. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.101.

Abdulbasit Alazzawi, Qahtan Yas, and Bahbibi Rahmatullah. A comprehensive review of software development life cycle methodologies: Pros, cons, and future directions. *Iraqi Journal for Computer Science and Mathematics*, 4:173–190, 07 2023. doi: 10.52866/ijcsm.2023.04.04.014.

Jamil Alzubi, Nour Ababneh, and Riyad Mohammad. Machine learning and big data analytics: A survey on security and privacy. *Journal of Big Data*, 5(1):1–22, 2018. doi: 10.1186/s40537-018-0124-5. URL https://doi.org/10.1186/s40537-018-0124-5.

Scott Ambler and Mark Lines. The disciplined agile process decision framework. In *The Disciplined Agile Process Decision Framework*, pages 3–14, 01 2016. ISBN 978-3-319-27032-6. doi: 10.1007/978-3-319-27033-3_1.

Saleema Amershi, Daniel S. Weld, Mark B. Hoffman, and Ece Kamar. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. doi: 10.1145/3290605.3300233. URL https://doi.org/10.1145/3290605.3300233.

Bjorn Andersen and Tom Fagerhaug. Root cause analysis: simplified tools and techniques. *The Journal for Healthcare Quality (JHQ)*, 24(3):46–47, 2002.

D.J. Anderson. *Kanban: Successful Evolutionary Change for Your Technology Business*. Blue Hole Press, 2010. ISBN 9780984521401. URL https://books.google.nl/books?id=RJ0VUkfUWZkC.

Anita Banerjee, Smeeta Kabadi, and Dostonbek Karimov. The transformative power of ai: Projected impacts on the global economy by 2030. *Review of Artificial Intelligence in Education*, 4:e020, 09 2023. doi: 10.37497/rev.artif.intell.educ.v4i00.20.

António Pereira Barata. GitHub - pereirabarataap/grouped$_a$*utoencoder* : *APyTorch−basedautoencoderthatsupportsgroupe*

Christian Bason. *Leading Public Sector Innovation: Co-creating for a Better Society*. Policy Press, Bristol, UK, 2nd edition, 2018. ISBN 9781447336242. 10.51952/9781447336259. URL https://doi.org/10.51952/9781447336259.

Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013. 10.1109/TPAMI.2013.50.

Erling Björgvinsson, Pelle Ehn, and Per-Anders Hillgren. Design things and design thinking: Contemporary participatory design challenges. *Design Issues*, 28(3):101–116, 2012. 10.1162/DESI$_{a0}$0165.*URL*.

Jennie Björk, Johan Frishammar, and Lina Sundström. Measuring innovation effectively—nine critical lessons. *Research-Technology Management*, 66(2):17–27, 2023. 10.1080/08956308.2022.2151232. URL https://doi.org/10.1080/08956308.2022.2151232.

Steve Blank. Why the Lean Start-Up Changes Everything, 5 2013. URL https://hbr.org/2013/05/why-the-lean-start-up-changes-everything.

Barry Boehm and Richard Turner. *Balancing Agility and Discipline: A Guide for the Perplexed*. Addison-Wesley, 07 2013. URL https://www.researchgate.net/publication/245579630_Balancing_Agility_and_Discipline-A_Guide_for_the_Perplexed.

Tim Brown. Change by design: How design thinking creates new alternatives for business and society. *Harvard Business Review Press*, 2009. URL https://www.researchgate.net/publication/338394954_Tim_Brown_Change_by_Design_How_Design_Thinking_Transforms_Organizations_and_Inspires_Innovation_2009. Book.

CB Insights. State of ai 2021 report. https://www.cbinsights.com/research/report/ai-trends-2021/, 2021. Accessed Month Day, Year.

Tianfeng Chai and R.R. Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7:1247–1250, 06 2014. 10.5194/gmd-7-1247-2014.

Min Chen, Shancang Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19 (2):171–209, 2012. 10.1007/s11036-013-0489-0. URL https://doi.org/10.1007/s11036-013-0489-0.

Clayton M. Christensen. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press, Boston, MA, 1997.

Clayton M. Christensen, Rory McDonald, Elizabeth J. Altman, and Sarah Palmer. Disruptive innovation: An intellectual history and directions for future research. *Journal of Management Studies*, 55(7): 1043–1078, 2018. 10.1111/joms.12349. URL https://doi.org/10.1111/joms.12349.

Piotr Chwastyk. Costs of innovation as a factor in the choice of innovative solutions. 11 2015.

Svetlana Cicmil, Terry Williams, Janice Thomas, and Damian Hodgson. Rethinking project management: Researching the actuality of projects. *International Journal of Project Management*, 24:675–686, 11 2006. 10.1016/j.ijproman.2006.08.006.

Ian M. Cobbold and Gavin J.G. Lawrie. The development of the balanced scorecard as a strategic management tool. In *Proceedings of the Third International Conference on Performance Measurement and Management (PMA 2002)*, Boston, MA, July 2002.

Alistair Cockburn. *Agile software development: the cooperative game*. Pearson Education, 2006.

Iain M Cockburn, Rebecca Henderson, and Scott Stern. The impact of artificial intelligence on innovation. Working Paper 24449, National Bureau of Economic Research, March 2018. URL http://www.nber.org/papers/w24449.

Robert Cooper. Managing technology development projects. *Engineering Management Review, IEEE*, 35:67–67, 02 2007. 10.1109/EMR.2007.329141.

Robert G. Cooper. Stage-gate systems: A new tool for managing new products. *Business Horizons*, 33(3):44–54, 1990. ISSN 0007-6813. https://doi.org/10.1016/0007-6813(90)90040-I. URL https://www.sciencedirect.com/science/article/pii/000768139090040I.

Daniel Kahneman Dan Lovallo. Delusions of Success: How Optimism Undermines Executives' Decisions, 7 2003. URL https://hbr.org/2003/07/delusions-of-success-how-optimism-undermines-executives-decisions.

Thomas H. Davenport and Julia Kirby. *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*. HarperBusiness, New York, NY, 2016. ISBN 9780062644185.

Thomas H. Davenport and D. J. Patil. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(10):70–76, October 2012. 10.1007/s10912-021-09728-8. URL https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Michael J. Davern and Robert J. Kauffman. Discovering potential and realizing value from information technology investments. *Journal of Management Information Systems*, 16(4):121–143, 2000.

Vasant Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013. 10.1145/2500499. URL https://doi.org/10.1145/2500499.

Jaime Diazbeltran. Machine learning model for last mile logistics: A case study on eta prediction in startups in latam. Medium blog post, April 2023. URL https://medium.com/@jaime.diazbeltran/machine-learning-operations-for-last-mile-logistics-a-case-study-on-eta-prediction-in-startups-in-ed0af7c33269.

Anat Drach-Zahavy and Anit Somech. Understanding team innovation: The role of team processes and structures. *Group Dynamics: Theory, Research, and Practice*, 5(2):111, 2001.

Niek D Du Preez and Louis Louw. A framework for managing the innovation process. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*, pages 546–558. IEEE, 2008.

Jan Fagerberg. Innovation: A guide to the literature. *Oxford Review of Economic Policy*, 21(2):1–23, 2005. 10.1093/oxrep/gri013. URL https://doi.org/10.1093/oxrep/gri013.

Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12, 01 2014.

Keith Goffin and Rick Mitchell. *Innovation Management: Strategy and Implementation.* Palgrave Macmillan, Basingstoke, UK, 3rd edition, 2016. ISBN 9781137373434.

Eliyahu M. Goldratt and Jeff Cox. *The Goal: A Process of Ongoing Improvement.* North River Press, Great Barrington, MA, 2nd edition, 1992. ISBN 9780884270614.

Jeff Gothelf and Josh Seiden. *Lean UX: designing great products with agile teams.* O'Reilly Media, Inc., 2021.

Paul Gray. *Data, Information and Knowledge in the Age of Digital Reason.* Digital Press, 2014.

The Boston Consulting Group. The most Innovative Companies 2015. Technical report, BCG, 12 2015. URL https://media-publications.bcg.com/MIC/BCG-Most-Innovative-Companies-2015-Nov-2015.pdf.

John R. Hauser, Gerard J. Tellis, and Abbie Griffin. Research on innovation: A review and agenda for marketing science. *Marketing Science*, 25(6):687–717, 2006. 10.1287/mksc.1050.0144. URL https://doi.org/10.1287/mksc.1050.0144.

Chikio Hayashi. What is data science ? fundamental concepts and a heuristic example. In Chikio Hayashi, Keiji Yajima, Hans-Hermann Bock, Noboru Ohsumi, Yutaka Tanaka, and Yasumasa Baba, editors, *Data Science, Classification, and Related Methods*, pages 40–51, Tokyo, 1998. Springer Japan. ISBN 978-4-431-65950-1.

Jim Highsmith. Agile project management: Creating innovative products. *The Agile Software Development Series*, 01 2004.

G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313:504–7, 08 2006a. 10.1126/science.1127647.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006b. 10.1126/science.1127647. URL https://doi.org/10.1126/science.1127647.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933. URL https://api.semanticscholar.org/CorpusID:144828484.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2065), 2016a. 10.1098/rsta.2015.0202. URL https://doi.org/10.1098/rsta.2015.0202.

Ian Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202, 04 2016b. 10.1098/rsta.2015.0202.

Magdalena Jurczyk-Bunkowska. Metoda planowania procesu innowacji dla małych przedsiębiorstw. *Zarządzanie Przedsiębiorstwem*, (1):14–25, 2010. Polskie Towarzystwo Zarządzania Produkcją.

Karthik Kambatla, Arto Kollias, Venkata Kumar, and Ananth Grama. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7):2561–2573, 2014. 10.1016/j.jpdc.2014.01.003. URL https://doi.org/10.1016/j.jpdc.2014.01.003.

Gerald C. Kane, Doug Palmer, Anh Nguyen Phillips, David Kiron, and Natasha Buckley. Strategy, not technology, drives digital transformation. Technical report, MIT Sloan Management Review and Deloitte University Press, 2015.

Robert S Kaplan, David P Norton, et al. Using the balanced scorecard as a strategic management system. *Harvard Business Review*, 1996. URL https://download.microsoft.com/documents/uk/peopleready/Using%20the%20Balanced%20Scorecard%20as%20a%20Strategic%20Management%20System.pdf.

Rainer Kattel, Mariana Mazzucato, and Josh Ryan-Collins. The economics of change: Policy and practice for a new economy. *UCL Institute for Innovation and Public Purpose Working Paper*, IIPP WP 2018-03, 2018. URL https://www.ucl.ac.uk/bartlett/public-purpose/wp2018-03.

Rainer Kattel, Mariana Mazzucato, and Josh Ryan-Collins. Challenge-driven innovation policy: Towards a new policy toolkit. *Journal of Industry, Competition and Trade*, 20(2):421–437, 2020. 10.1007/s10842-019-00329-w. URL https://doi.org/10.1007/s10842-019-00329-w.

Roberto Kleis et al. Information technology and intangible output: The impact of it investment on innovation productivity. *Research Policy*, 41(3):547–560, 2012. 10.1016/j.respol.2011.11.004. URL https://doi.org/10.1016/j.respol.2011.11.004.

Stephen J. Kline and Nathan Rosenberg. An overview of innovation. In Richard S. Rosenbloom and William J. Spencer, editors, *The Positive Sum Strategy: Harnessing Technology for Economic Growth*, pages 275–305. National Academy Press, Washington, D.C., 1986.

Timur Kogabayev and Antanas Maziliauskas. The definition and classification of innovation. *Holistica*, 8, 04 2017. 10.1515/hjbpa-2017-0005.

Rajiv Kohli and Varun Grover. Cocreating it value: New capabilities and metrics for multifirm environments. *MIS Quarterly*, 36(1):225–232, 2012. 10.25300/MISQ/2012/36.1.12.

Jens Konopik, Christoph Jahn, Tassilo Schuster, Nadja Hoßbach, and Alexander Pflaum. Mastering the digital transformation through organizational capabilities: A conceptual framework. *Digital Business*, 2(2):100019, 2022. ISSN 2666-9544. https://doi.org/10.1016/j.digbus.2021.100019. URL https://www.sciencedirect.com/science/article/pii/S2666954421000181.

Nattarinee Kopecka. The balanced scorecard implementation, integrated approach and the quality of its measurement. *Procedia Economics and Finance*, 25:59–69, 2015. ISSN 2212-5671. https://doi.org/10.1016/S2212-5671(15)00713-3. 16th Annual Conference on Finance and Accounting, ACFA Prague 2015, 29th May 2015.

Otto Lappi. Gaze strategies in driving – an ecological approach. *Frontiers in Psychology*, 13:821440, 2022. 10.3389/fpsyg.2022.821440. URL https://doi.org/10.3389/fpsyg.2022.821440.

Keyao Li, Mark A. Griffin, Tamryn Barker, Zane Prickett, Melinda R. Hodkiewicz, Jess Kozman, and Peta Chirgwin. Embedding data science innovations in organizations: a new workflow approach. *Data-Centric Engineering*, 4:e26, 2023. 10.1017/dce.2023.22.

Dan Lovallo and Daniel Kahneman. Delusions of success. *Harvard business review*, 81(7):56–63, 2003.

Jianxi Luo. *Data-Driven Innovation: What Is It.* arXiv preprint arXiv:2201.08184, 2022. URL https://www.researchgate.net/publication/357987427_Data-Driven_Innovation_What_Is_It.

Mark Lycett. Data science: An action plan for expanding the technical areas of the field of data science. *International Journal of Data Science and Analytics*, 1(1):5–13, 2013. 10.1007/s41060-013-0004-0. URL https://doi.org/10.1007/s41060-013-0004-0.

John Mankins. Technology readiness assessments: A retrospective. *Acta Astronautica - ACTA ASTRONAUT*, pages 1216–1223, 11 2009. 10.1016/j.actaastro.2009.03.058.

Matthew Mayo. The data science process, 2016. URL https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html.

Mariana Mazzucato. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths.* Penguin Books, Harlow, England, 10th anniversary edition edition, 2018. ISBN 9780241305591.

Rita Gunther McGrath. *Discovery-Driven Growth: A Breakthrough Process to Reduce Risk and Seize Opportunity*. Harvard Business Review Press, Boston, MA, 2010. ISBN 9781422174859.

Ines Mergel. Agile innovation management in government: A research agenda. *Government Information Quarterly*, 33:516–523, 09 2016. 10.1016/j.giq.2016.07.004. URL https://www.researchgate.net/publication/308003888_Agile_innovation_management_in_government_A_research_agenda.

Ines Mergel, R. Karl Rethemeyer, and Kimberley R. Isett. The product owner in agile development: An empirical study of roles and role ambiguity. *Government Information Quarterly*, 36(2):301–309, 2019. 10.1016/j.giq.2018.12.004. URL https://www.researchgate.net/publication/307431462_Big_Data_in_Public_Affairs.

Matej Mihelcic and Pauli Miettinen. Finding Rule-Interpretable Non-Negative Data Representation . *IEEE Transactions on Knowledge & Data Engineering*, 37(05):2538–2549, May 2025. ISSN 1558-2191. 10.1109/TKDE.2025.3538327. URL https://doi.ieeecomputersociety.org/10.1109/TKDE.2025.3538327.

Patrick Mikalef, Rogier Van de Wetering, and John Krogstie. From big data analytics to dynamic capabilities: The effect of organizational inertia. In *PACIS 2019 Proceedings*, page 198, Beijing, China, 2019. URL https://aisel.aisnet.org/pacis2019/198. Case study of 27 firms.

Ralf Müller and Rodney Turner. The influence of project managers on project success criteria and project success by type of project. *European Management Journal*, 25(4):298–309, 2007. 10.1016/j.emj.2007.06.003.

Bansi Nagji and Geoff Tuff. Managing your innovation portfolio. *Harvard Business Review*, 90(5):66–74, 2012.

Tomoko Nemoto and David Beglar. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, volume 108, pages 1–6, 2014.

Gina Colarelli O'Connor and Mark P. Rice. Major innovation as a dynamic capability: A systems approach. *Journal of Product Innovation Management*, 25(4):313–330, 2008. 10.1111/j.1540-5885.2008.00304.x. URL https://doi.org/10.1111/j.1540-5885.2008.00304.x.

Sivadon Ongsulee. *Big Data Management: Concepts, Techniques, and Applications*. CRC Press, Boca Raton, FL, 2017. ISBN 9781498715135.

Organisation for Economic Co-operation and Development (OECD). The innovation imperative in the public sector: Setting an agenda for action. Technical report, OECD Publishing, Paris, 2017. URL https://www.oecd.org/governance/ăĽμæŰř-the-innovation-imperative-in-the-public-sector-9789264270879-en.htm.

Alexander Osterwalder and Yves Pigneur. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. Wiley, Hoboken, NJ, 2010. ISBN 9780470876411.

Stephen Otim and Varun Grover. Resolving uncertainty and creating value from the exercise of e-commerce investment options. *Information Systems Journal*, 22(4):261–287, 2012. 10.1111/j.1365-2575.2011.00385.x. URL https://doi.org/10.1111/j.1365-2575.2011.00385.x.

Raja Parasuraman, Thomas Sheridan, and Christopher Wickens. A model for types and levels of human interaction with automation. ieee trans. syst. man cybern. part a syst. hum. 30(3), 286-297. *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans: a publication of the IEEE Systems, Man, and Cybernetics Society*, 30:286–97, 06 2000. 10.1109/3468.844354.

Samir Passi and Steven J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–28, November 2018. ISSN 2573-0142. 10.1145/3274405. URL http://dx.doi.org/10.1145/3274405.

Mark Paulk, William Curtis, Mary Beth Chrissis, and Charles Weber. Capability maturity model for software (version 1.1). Technical report, Software Engineering Institute, Carnegie Mellon University, Feb 1993. URL https://www.sei.cmu.edu/library/capability-maturity-model-for-software-version-11/.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3): 45–77, 2007.

Gary P. Pisano. You need an innovation strategy. *Harvard Business Review*, 93(6):44–54, June 2015. 10.5465/amp.2013.0197. URL https://hbr.org/2015/06/you-need-an-innovation-strategy.

Gary P. Pisano. *Creative Construction: The DNA of Sustained Innovation*. PublicAffairs, New York, NY, 2019. ISBN 9781610398770.

PWC Data Science Group. AI adoption in the business world: current trends and future predictions. Technical report, PWC, 4 2023. URL https://www.pwc.com/il/en/mc/ai_adopion_study.pdf.

Michael Quinn Patton and Patricia A. Patrizi. Strategy as the focus for evaluation. *New Directions for Evaluation*, 2010(128):5–28, 2010. https://doi.org/10.1002/ev.343. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ev.343.

Rainer Kattel. Innovation bureaucracies: How agile stability creates the entrepreneurial state. https://www.ucl.ac.uk/bartlett/publications/2019/dec/innovation-bureaucracies-how-agile-stability-creates-entrepreneurial-state, December 2019. Bartlett Faculty of the Built Environment.

Muawia Ramadan, Hana Shuqqo, Layalee Qtaishat, Hebaa Asmar, and Bashir Salah. Sustainable competitive advantage driven by big data analytics and innovation. *Applied Sciences*, 10:6784, 09 2020. 10.3390/app10196784.

Romesh Ranawana and Asoka Karunananda. An agile software development life cycle model for machine learning application development. In *2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1–6, 12 2021. 10.1109/SLAAI-ICAI54477.2021.9664736.

Kelsey Richard. The 5P Framework - Trust Insights Marketing Analytics Consulting, 4 2024. URL https://www.trustinsights.ai/blog/2024/04/the-5p-framework/.

Eric Ries. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown Business, 2011. ISBN 9780307887894. URL https://www.amazon.com/Lean-Startup-Entrepreneurs-Continuous-Innovation/dp/0307887898.

Jeffrey S. Saltz and Ivan Shamshurin. Big data team process methodologies: A literature review and the identification of key factors for a project's success. In *Proceedings of the 2016 IEEE International Conference on Big Data*, pages 2872–2879, 2016. 10.1109/BigData.2016.7840936. URL https://doi.org/10.1109/BigData.2016.7840936.

Joseph A. Schumpeter. *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*. McGraw-Hill, New York, NY, 1939.

Ken Schwaber and Mike Beedle. *Agile Software Development with Scrum*. Series in Agile Software Development. Prentice Hall, 2002. ISBN 9780130676344. URL https://www.pearson.com/en-us/subject-catalog/p/agile-software-development-with-scrum/P200000003387.

D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, and Dan Dennison. Hidden technical debt in machine learning systems. *NIPS*, pages 2494–2502, 01 2015a.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*, pages 2503–2511, 2015b. URL https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

Aaron J Shenhar and Dov Dvir. *Reinventing project management: the diamond approach to successful growth and innovation*. Harvard Business Review Press, 2007.

Stefan Siegert. Simplifying and generalising murphy's brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143, 12 2016. 10.1002/qj.2985.

Rajiv Singla, Ankush Singla, Yashdeep Gupta, and Sanjay Kalra. Artificial intelligence/machine learning in diabetes care. *Indian Journal of Endocrinology and Metabolism*, 23(4):495–497, 2019. $10.4103/\text{ijem.IJEM}_2 28_1 9.URL$.

Rosanna Spanò, Fabrizia Sarto, Caldarelli Adele, and Riccardo Viganò. Innovation & performance measurement: An adapted balanced scorecard. *International Journal of Business and Management*, 11: 194–194, 05 2016. 10.5539/ijbm.v11n6p194.

Tanja Suomalainen. *Changing the planning for agile and lean software development: From roadmapping to continuous planning*. PhD thesis, University of Oulu, 09 2016.

Helle Søndergaard, Mette Knudsen, and Nicolai Laugesen. The catch-22 in strategizing for radical innovation. *Technology Innovation Management Review*, 11:4–16, 04 2021. 10.22215/timreview/1425.

Gabriel Tarde. *Penal Philosophy*. Transaction Publishers, New Brunswick, NJ, 2001.

Deni Ahmad Taufik, Humiras Hardi Purba, and Hasbullah Hasbullah. Balanced scorecard: Literature review and implementation in organization. *Operations Excellence Journal of Applied Industrial Engineering*, 13:111–123, 03 2021. 10.22441/oe.2021.v13.i1.012.

Mustafa Uysal. Big data analytics for smart cities: A review. *Journal of Big Data*, 9(1):1–24, 2022. 10.1186/s40537-022-00567-7. URL `https://doi.org/10.1186/s40537-022-00567-7`.

Frans Van der Meer, Jakob Edler, Luke Georghiou, and John Yeow. Mission-oriented innovation policy: Practices and challenges. *Research Policy*, 50(4):104190, 2021. 10.1016/j.respol.2021.104190. URL `https://doi.org/10.1016/j.respol.2021.104190`.

Willem H. Vanderburg. *Living in the Labyrinth of Technology*. University of Toronto Press, Toronto, ON, 2005. 10.3138/9781442657298. URL `https://doi.org/10.3138/9781442657298`.

Pascal Vincent, Hugo Larochelle, Y. Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 01 2008. 10.1145/1390156.1390294.

Jan vom Brocke, Alan Hevner, and Alexander Maedche. Introduction to design science research. In Jan vom Brocke, Alan Hevner, and Alexander Maedche, editors, *Design Science Research. Cases*, pages 1–13. Springer, 2020. $10.1007/978\text{-}3\text{-}030\text{-}46781\text{-}4_1.URL$.

Matthew A. Waller and Stanley E. Fawcett. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84, 2013. 10.1111/jbl.12010. URL `https://doi.org/10.1111/jbl.12010`.

Samuel Fosso Wamba, Angappa Gunasekaran, Shahriar Akter, Steven Ji-Fan Ren, Rameshwar Dubey, and Stephen J. Childe. Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70:356–365, 2017. 10.1016/j.jbusres.2016.08.009. URL `https://doi.org/10.1016/j.jbusres.2016.08.009`.

Michael Williams and Tami Moser. The art of coding and thematic exploration in qualitative research. *International Management Review*, 15:45, 2019. URL `https://api.semanticscholar.org/CorpusID:198662452`.

Zhigang Yin, Hüseyin Kaynak, and Shuxia Wang. Data analytics and big data: An overview. *Journal of Industrial Information Integration*, 22:100187, 2021. 10.1016/j.jii.2020.100187. URL `https://doi.org/10.1016/j.jii.2020.100187`.

Zihan Zhang, Simon S. Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning, 2020. URL `https://arxiv.org/abs/2010.05901`.

Appendix A

Interview Protocol

The first set of interviews are conducted in semi-structured individual settings, allowing open discussions about the nature of work and existing practices in IDS teams. In total, 15 interviews were held. This setting encouraged deeper reflection on topics and bottlenecks, while also breaking down how projects are initiated, performed, and evaluated in terms of their value and alignment with the larger organizational goals.

Introduction

- What is your current role and what type of projects do you typically work on?
- How did you start working on your current project?
- How would you describe your area of expertise within the team?

Project initiation and workflows

- How do new projects typically get started in the team?
- What are the main ways projects are initiated? (Personal project, user requests, data-driven ideas)
- How is a project's value or feasibility assessed before development begins?
- Are there any formal intake or review steps you follow?

Motivation and ownership

- What motivates you personally to work on innovation or exploratory projects?
- How do you decide which projects are worth developing?
- Have there been times when a project felt too personal to let go? How do you handle that?

Success and Evaluation

- What does a "successful project" look like to you?
- Are there any defined success metrics or evaluation criteria you use?
- Do you typically know when to exit a project? How is that decided?

Collaboration

- How often does collaboration happen in projects? Do you work more individually or in pairs/teams?
- How do you share updates and get feedback from others (e.g., Managers, Paul, or Stephanie)?
- Is there any informal or formal accountability structure?

Project Management Practices

- Do you follow any kind of project planning methodology (e.g., Kanban, Scrum)?
- How do you feel about introducing more structured tools like roadmaps or time-boxed sprints?

- Do you have regular check-ins or defined phases for every project? Does it feel restrictive and why?

Innovation and Project types

- Do you see different projects, some more experimental/innovation projects, and others more implementation-focused work?
- Is it helpful to categorize projects and define goal/results prior?
- What kind of structure (if any) would help innovation projects succeed without reducing flexibility?

Challenges and Gaps

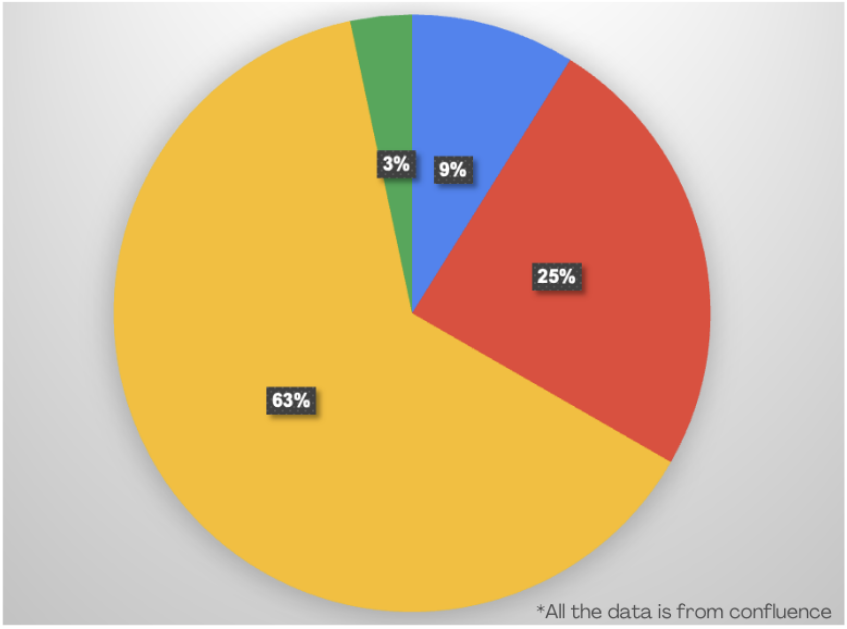
- What issues currently hold projects back in terms of projects being adopted (selected) or measured (success)?
- If you could improve something about how the team manages its innovation projects, what would it be?

Appendix B

Questions mapping with sources

Project Pattern Mapping

This step aims to assess the historical innovation portfolio against a consistent set of success metrics in dimensions such as value delivered, stakeholder adoption, social impact, and progress in terms of scaling. The analysis is implemented as a pattern mapping exercise: Each project is tested and assigned to its appropriate life cycle phase (Experimentation, Prototype, Pilot, or Usage) based on the observed results, as illustrated in Figure 4.3



Yellow: Experiment phase
Red: Prototype phase
Blue: Pilot phase
Green: Production

Figure B.1: Product Maturity Phases — IDLab

Every project from the IDlab portfolio was compared against a set of success criteria to identify patterns and characteristics (scored in binary yes/no) of what makes a project successful, scalable, or impactful. The metrics and questions associated with them to identify are:

Implementation Success

- Deployment status: Was the project delivered into production or operational use?
- User adoption rate: Are end-users actively using the solution or insights produced?

- Reuse: Has the tool, model, or method been reused in other teams or projects?
- Documentation quality: Does it have usable documentation for others to pick up?

Business Value and Strategic Impact

- Decision support: Did the output inform or shape any real business or policy decisions?
- Cost/time savings: Was there measurable reduction in time, effort, or cost?
- Process improvement: Did it automate or simplify a previously manual or complex process?
- Alignment with ministry/team priorities: Did it contribute to a recognized strategic objective?

Learning and Success

- Knowledge creation: Did it produce internal learning (e.g., methods, pipelines, experiments)?
- Novelty: Was the approach, model, or data use novel compared to ILT's standard practices?
- Scalability: Can the outcome or approach scale to other domains or units?
- Documentation of failures/lessons: Are learnings from unsuccessful or exploratory paths documented?

Project Management and Process

- Time to outcome: Was the work completed in a timely manner compared to expectations?
- Defined scope and Definition of Done (DoD): Was there a clear definition of done or evaluation criteria up front?
- Iteration quality: Did the project improve through feedback loops or prototyping?
- Stakeholder engagement: Were stakeholders involved throughout (not just at the end)?
- Exit strategy used: Was there a structured process for stopping or pivoting if needed?

Long-Term Impact and Sustainability

- Sustained usage or value: Is the output still in use or producing value after project close?
- Internal influence: Did the project shift thinking or routines within the team/org?
- External recognition: Has the project been referenced or reused by other ministries or partners?

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Project Name	Did the project transition succesfully from research to implementation?	Did the project inform any key decision-making? Any actionable business insight?	Were the results actionable for stakeholders or decision-makers?	Were stakeholders actively engaged throughout the project lifecycle?	Are end-users actively using the solution or output of the project?	Did the project lead to measurable cost or time savings?	Did the project replace or automate any manual processes?	Did the project lead to the creation of new methods, models, or internal learning?	Does the project have clear documentation for knowledge sharing?	Has any part of the tool, method, or process been reused in other projects?	Was the project deployed into production or active use?	Was the time from idea to outcome within estimated time agreed on?	Were any success metrics defined for the project? A DoD ?	Has the project demonstrated long-term impact or sustained change?	Did the project follow a standardized structure or project management process?	Any challenges with Data availability?	Did the project align with Ministry priority areas or team goals?	Was the problem statement or research approach novel?
Noise Pollution Web Scrape	No	No	Yes	Partial	No	No	No	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes
Sight Line Model Applied in Practice	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Detecting Std Dev in Drone Tracks	No	No	Partial	No	No	No	No	Yes	No	No	No	No	Yes	Partial	No	No	Yes	Yes
Inland Ship Degassing Detection	No	Yes	No	No	No	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	Yes
ABM for Nanomaterials	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	No	Partial	No	Yes	Yes
Drinking Water & Legionella Pilot #1	No	No	No	Yes	No	No	No	Yes	No	No	No	No	Yes	No	Partial	Yes	Yes	Yes
Manage O:\ Disk	Yes	Yes	Yes	Yes	Yes		Yes	Yes	No	No	Partial		No	Yes	Partial	No	Yes	Yes
Putin's Shadow Fleet Risk Model	No	No	No	Yes	Partial	No	No	Partial	No	No	No	No	No		No	No	Yes	Yes
Web scrape water quality	No	No	No	No	No	No	No	Yes	No	No	No	Yes	No	No	No	No	Yes	Yes
Detecting Std Dev in Drone Tracks #2	No	No	No	No	No	No	No	Yes	No	No	No	No	Yes	Partial	No	No	Yes	Yes
Geo-webapp Small Aviation	No	No	No	Yes	No	No	No	Yes	No	Yes	No	No	No	Partial	No	No	Yes	Yes
NLP pipeline	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Geo Chatbot	No	No	Partial	Yes	No	No	No	Yes	Yes	No	No	Yes	No	No	Partial	Yes	No	Partial
Content scrape innovation transport	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
GeoAI Kickstart: Big Bags Image Recognition	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Partial	Partial	Yes	No	Yes	No	No	Yes	Yes
Web scraping Supervision of BRL 100/200 companies (Air conditioning services) - **Pilot**	Yes	Yes	Yes	Yes	Partial	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Yes
Temporal design choices for feature engineering from dynamic networks	No	No	No	No	No	No	No	Yes	Yes	Yes	No	Yes	No	No	No	No	No	Yes
Geo Webapp Schiphol TMA Vertical Deviations	No	Partial	No	Yes	No	No	Partial	Yes	No	No	No		Partial	No	No	Yes	Yes	Yes
Shipbreaking Model	Partial	Yes	Partial	Yes	No	No	Partial	Yes	Yes	No	No		No	Partial	No	Yes	Yes	Yes

Figure B.2: Portfolio Analysis - Pattern Mapping

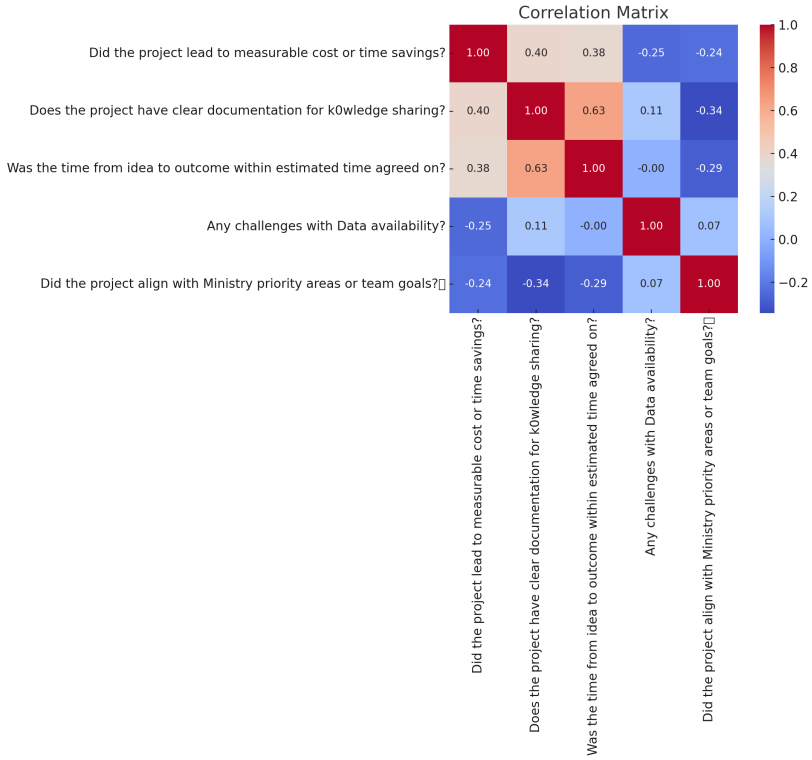


Figure B.3: Correlation Matrix - Questions Project Mapping

Appendix C

Codebook

Table C.1: Codebook

Code Group	**Code**	**Gr**	**Co**
Team Identity & Origin	Origin of ID Lab as a data-driven shift	1	1
	Incremental team growth based on proven value	1	1
	Proving value under autonomy	1	1
	Scaling introduces scrutiny	1	1
Processes	Lack of project tracking	8	15
	Potential for lightweight agile adoption	6	8
	No defined roles or process lead	5	9
	Stand-ups and retrospectives not used	4	13
Organizational Culture	Freedom and responsibility tension	2	2
	Resistance to structured project management	3	10
	Risk-averse culture focused on avoiding publicity	1	1
	Theory over execution / implementation resistance	1	1
	Scaling impact causes external pressure	1	1
Innovation Culture	Freedom vs. discipline tension	6	5
	Projects drifting from original pitch	4	3
	No pressure to deliver	3	3
Interdepartmental Processes	Agile methods used in other teams	3	2
	Stronger success measurement in other teams	2	2
	Structured project phases comparison	2	2
	Scrum adoption and Deciated roles	2	4
Knowledge & Learning	Failure is success	5	14
	No central repository for lessons learned	5	4
	Knowledge loss across transitions	4	3
	New ideas emerge from building	3	2
Project Management & Structure	Missing exit strategy and timeboxing	2	2
	No clear milestones or success criteria	2	2
	Overextended projects without review	1	1
	Failure to reassess direction	1	1
	Lack of project tracking	1	1
Project Intake & Selection	No formal intake process	9	6
	Unstructured project selection	7	5
	Balancing innovation vs. demand-driven work	6	5
	Need for smart backlog	4	3

Continued on next page

Table C.1 – continued from previous page

Code Group	Code	Gr	Co
Team beliefs w.r.t Projects	Project initiation based on stakeholder balance	1	1
	Innovation mandate – avoiding repetitive tasks	1	1
	Technology exploration for learning	1	1
	Avoiding misuse of team expertise	1	1
	Exploratory work accepted as valid mode	1	1
	User vs innovation tension	1	1
	Balancing short and long term projects	1	1
Project Lifecycle & Management	No defined success metrics	6	5
	Lack of exit strategy	5	4
	Informal transition between phases	4	3
	Need for timeboxing	3	2
Team Dynamics	Diverse approaches and team perspectives	1	1
	Need for defined roles	5	3
	Lack of accountability	4	3
	Managerial involvement in project selection	3	2
User-Centricity	Gap in understanding user needs	6	5
	Need for user validation during pilots	5	4
	Not all projects need user focus	2	2
Impact & Metrics	Lack of explicit success metrics	1	1
	Incremental societal progress emphasis	1	1
	Challenges of measuring impact	1	1
	Metrics misuse danger	1	1
	Misalignment with end-user effort	1	1

Gr = The total number of times the code has been applied
Co = The number of interviews the code surfaced in

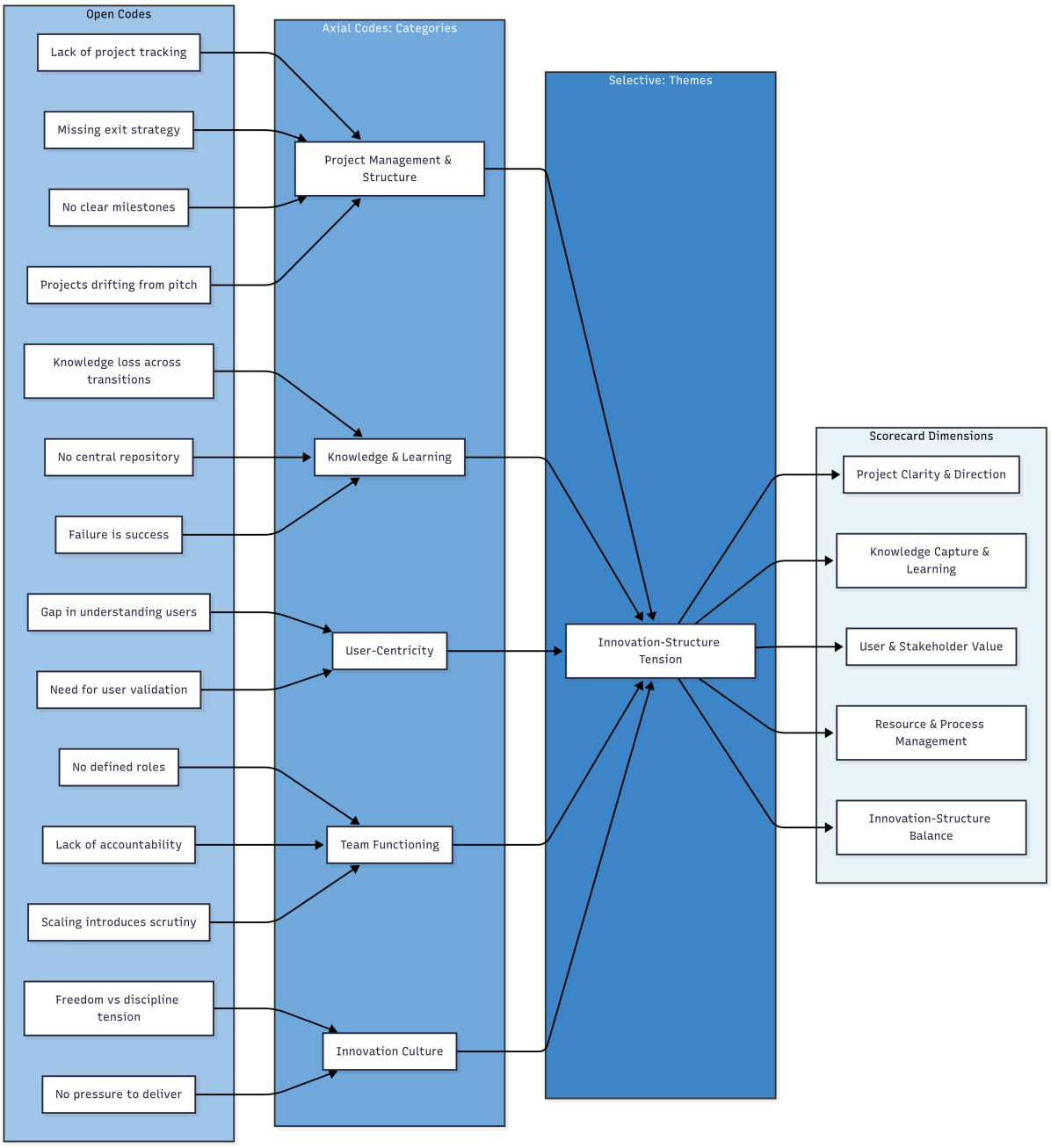


Figure C.1: Selective Coding Process

1. How IDlab sees itself:

IDlab positions itself at the center of innovation within the Ministry. Established in 2017, the team does not follow strict workflows; instead, it focuses on organic growth, experimentation, and learning, an approach strongly supported by the freedom in which they operate. Some team members describe their work as "organized chaos", where freedom, autonomy, and curiosity fuels innovation.

Projects are typically initiated from one of three sources: addressing inspector needs, leveraging available data, or testing emerging technologies. These projects are aimed at (1) making inspectors happy, (2) helping management solve strategic problems, and (3) keeping data scientists engaged with innovative challenges. Unlike other data teams, IDlab explicitly avoids tasks such as basic dashboarding and instead follows a four-phase approach for its projects: Experimentation → Prototype → Pilot → Usage to deliver value.

The table below summarizes the key beliefs, behaviors, and structural choices that emerged during the

interviews and offers insight into how IDlab functions today.

Theme	**Code**	**How the Team Operates**
Data as direction	#innovation_mandate #origin_of_IDLab #data_driven_transformation	Projects are chosen for novelty and impact, not routine analysis. "The ID Lab was started because leadership wanted a data-driven organization, and the existing analysis teams couldn't achieve that."
Proving worth constantly	#self_sufficiency #proving_value	The team has wide freedom, but must consistently prove its value. "We are given freedom, but that also means no one is protecting us, we have to prove our worth constantly."
Incremental development	#incremental_team_growth	The lab expands only when value is demonstrated through action. "We grew stepwise, every time we proved value, we asked for more people. Proof drives expansion here"
Balance of perspectives	#diverse_approaches #team_perspectives	Teams often start without knowing exactly what the output will be. "We have different mindsets in the team where some focus on innovation, some on users."
Multi-Faceted Projects selection	#project_selection_criteria	Projects must meet one or more: inspector value, strategic fit, or innovation challenge. "We have three reasons for taking projects: (1) making inspectors happy, (2) helping management solve strategic problems, (3) keeping data scientists engaged with innovative challenges."

Table C.2: IDLabs Operational beliefs

This operation model reflects IDlab's non-linear approach to innovation, characterized by iterations and the circular nature of data-driven experiments. However, when this approach encounters uncertainties and more rigid aspects of the organization, it can cause misalignment and problems. We explore some of these challenges uncovered from the interviews in the section below.

2. Challenges:

Along with mapping IDlab's current strengths, six key challenge areas were identified through the interviews, ranging from a lack of structured project evaluation to misaligned stakeholder expectations, which the proposed framework aims to address. Table C.3 presents a summary of these challenge areas and the respective quotes of the members of the team.

Theme	Challenge	Quote	Indication
Unstructured Project Lifecycles	Projects lack milestones, time-boxing, or exit criteria, leading to stagnation.	“We’ve worked on projects for months without knowing when to stop.”	Introduce evaluation and time-boxing practices in the workflow.
Exploratory vs. Usability focus	The conflict between exploratory innovation and user-driven design.	“We call it ’organized chaos’—sometimes we don’t know what we’re looking for, but we still find useful things, and too much focus on users often limits innovation.”	Support project typology tool that differentiates between initiatives and identifies their type.
Lack of Impact measures	Teams find it difficult to measure impact, which results in external pressure.	"We don’t have a clear way to quantify success. Failure is a success, too. It’s a combination of strengths and weaknesses."	Incorporate lightweight outcome-based KPIs.
Autonomy without Guardrails	High independence, but lack of structured direction or shared accountability.	"We got an infinite amount of freedom. But with this freedom, we also got responsibility. Nobody else was taking responsibility for us."	Define ownership and priorities for the team.
Innovation understanding & Role clarity	Strong push for novelty, but misunderstood by others as a data support team.	“"We often have to say no because people think we are just data collectors. We are not here to fetch data or do dashboarding. That’s someone else’s job.”	Create clearer intake filters and communication standards for engagement.
Cognitive Diversity within a team	Diverse perspectives (strategic versus technical, user versus novelty) create conflict but also strengthen a team.	“This diversity is our strength, but it’s hard to align.”	Facilitate multi-perspective project framing and alignment rituals.

Table C.3: Challenges Summary from the Interviews

The interviews showed that although IDlab is known for its high level of autonomy, some of these characteristics can pose operational difficulties if not controlled. For example, unstructured project lifecycles are frequently the result of what the team refers to as "organized chaos". Projects run the risk of stale if there are no clear exit criteria. In a similar scheme, many people consider the team’s cognitive diversity to be one of its main advantages. However, individual project management preferences (where some team members prefer informal, self-directed methods, others prefer structured approaches like Agile or Kanban) can actually cause conflict and hinder collaboration.

It is key to note that these tables do not display contradictions, but highlight the fundamental conflicts between structure and intention. These findings underscore the need for a framework that fosters freedom while enhancing ownership, structure, and clarity in project evaluation.

Pre

Question	Root Cause Analysis (RCA)	Interviews + Coding	Pattern Mapping
To what extent does this project align with ILT's strategic direction, goals, or organizational priorities?	✓	✓	
To what degree does this project avoid duplication and offer something new or needed beyond existing tools/processes?	✓	✓	✓
Is the problem or opportunity clearly defined and understood?	✓	✓	✓
Do we have access to the right technologies and data to implement this solution effectively?	✓	✓	✓
How much novelty or innovation does this project introduce to ILT's existing operations or knowledge base?	✓	✓	✓
Is the project expected to go through multiple development or research cycles?	✓	✓	✓
To what extent is the project designed for experimentation and iteration?	✓	✓	✓
Are the problem, data, and solution space well-defined and scoped?	✓	✓	✓
What is the level of uncertainty or risk involved in delivering intended value?	✓	✓	✓
How much does the project aim to generate learning that can be shared or reused by other teams or future projects?	✓	✓	✓
To what extent are success criteria loosely defined or expected to evolve during the project?	✓	✓	✓
How much of the project's value lies in learning, discovery, or generating new insights?	✓	✓	✓
What is the level of uncertainty in integrating the solution into existing systems or workflows?	✓	✓	✓
Is the project currently in a validation phase with clear hypotheses to test?	✓	✓	✓
What is the likelihood of this project producing a tangible output or deliverable?	✓	✓	✓
How strong is the level of stakeholder commitment or sponsorship?	✓	✓	✓

Figure C.2: Validated Questions with source

Appendix D

Future work

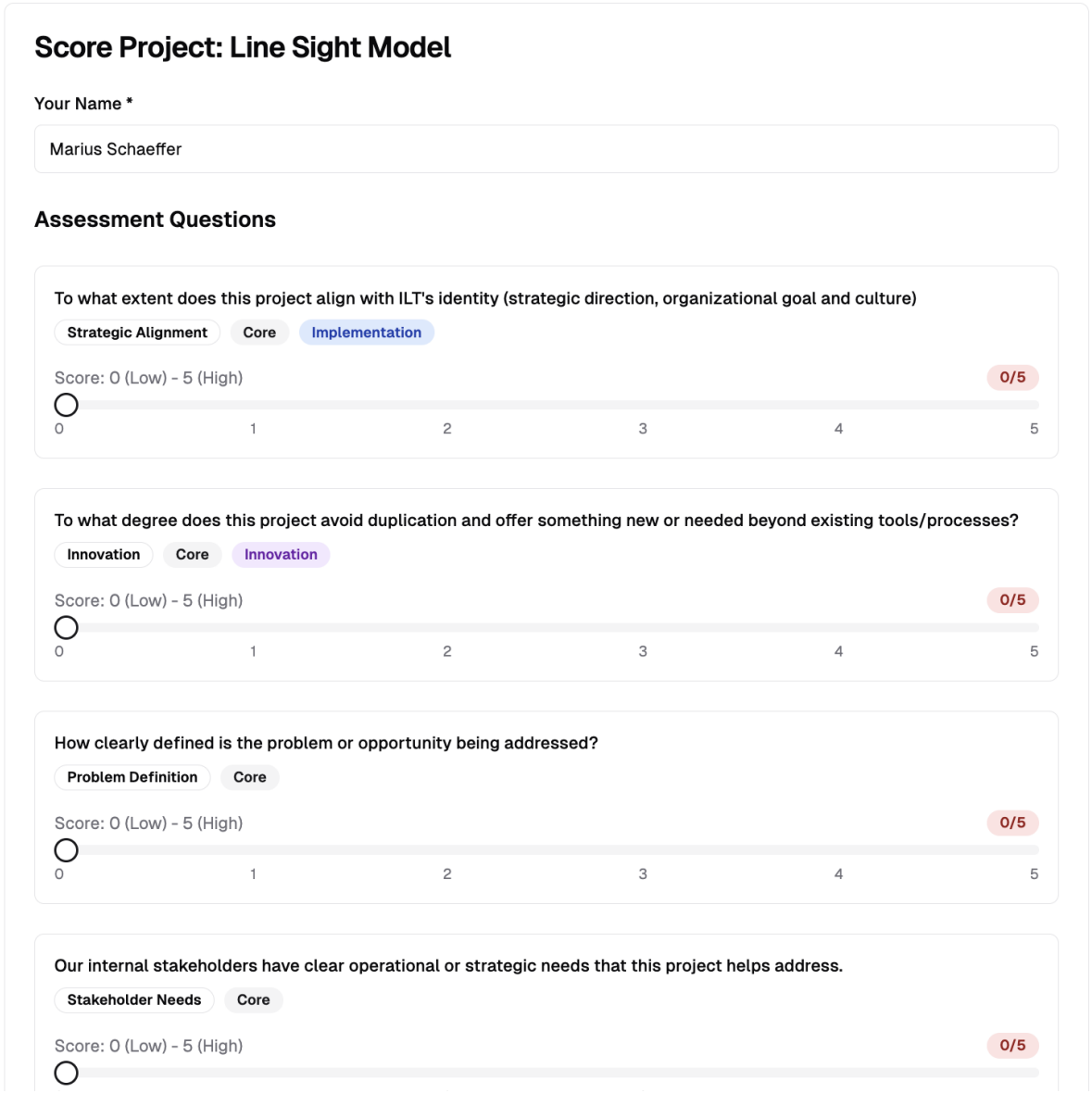
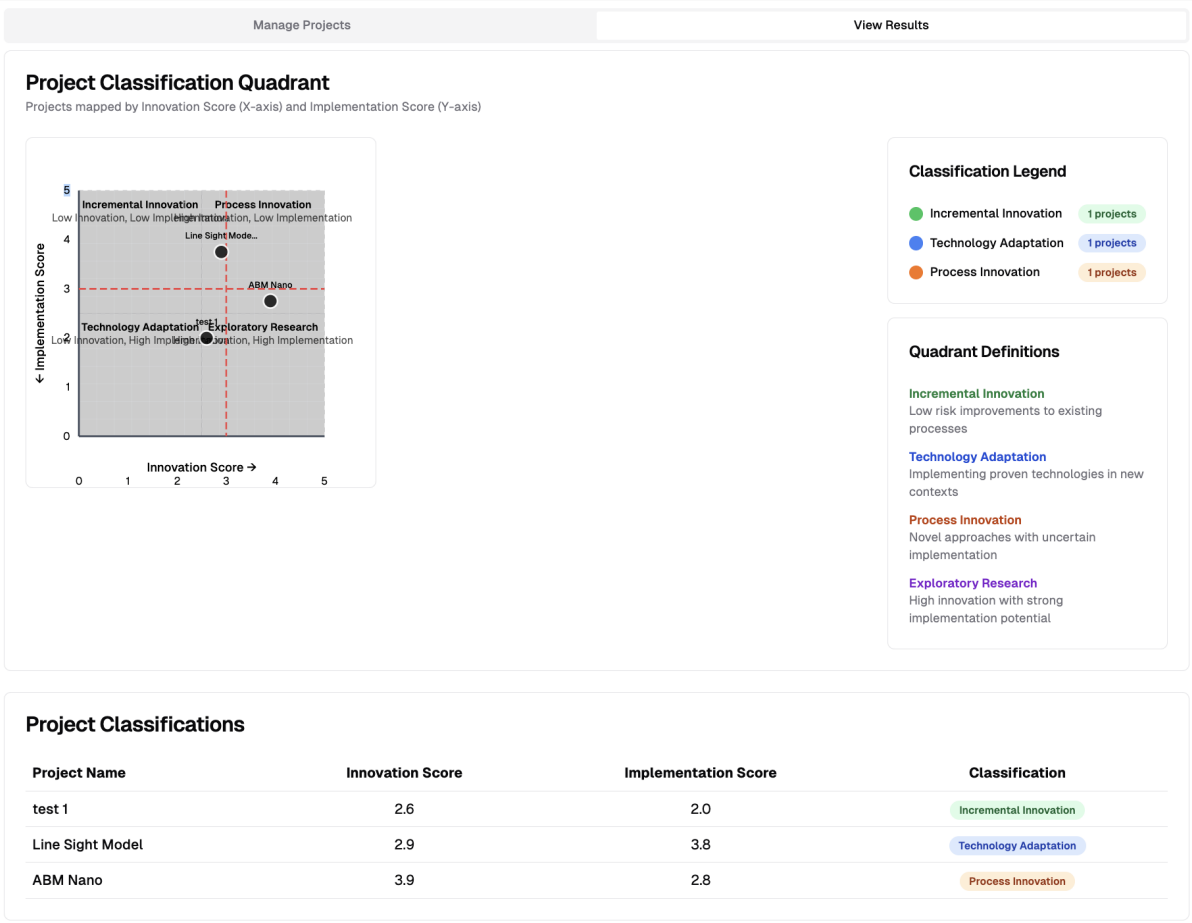


Figure D.1: Framework Protoype 1



Project Classifications

Project Name	Innovation Score	Implementation Score	Classification
test 1	2.6	2.0	Incremental Innovation
Line Sight Model	2.9	3.8	Technology Adaptation
ABM Nano	3.9	2.8	Process Innovation

Figure D.2: Framework Protoype 2