The Risk Management Process for Data Science: Gaps in Current Practices

Sucheta Lahiri
Syracuse University
sulahiri@syr.edu

Jeffrey S. Saltz
Syracuse University
jsaltz@syr.edu

Abstract

Data science projects have unique risks, such as potential bias in predictive models, that can negatively impact the organization deploying the models as well as the people using the deployed models. With the increasing use of data science across a range of domains, the need to understand and manage data science project risk is increasing. Hence, this research leverages qualitative research to help understand the current practices concerning the risk management processes organizations currently use to identify and mitigate data science project risk. Specifically, this research reports on 16 semi-structured interviews, which were conducted across a diverse set of public and private organizations. The interviews identified a gap in current risk management processes, in that most organizations do not fully understand, nor manage, data science project risk. Furthermore, this research notes the need for a risk management framework that specifically addresses data science project risks.

1. Introduction

The field of data science has many related terms that are extensively used in academia and industry, all conveying the creation of new value via the analysis of data. Some popular terms include big data, artificial intelligence, big data analytics, text mining, business intelligence, business analysis, machine learning, and finally, data science. Independent of the term used, the field of data science draws interest from various other fields such as statistics, informatics, computing, communication, management, and sociology [24].

With the contribution of various disciplines, particularly statistics, informatics, and computing, the entire cycle of the data science process can become obscure and complex. These unforeseen challenges of obscurity and unpredictability can creep in at any step of the data science project, that is, from the data capture to interpreting the results of the project [25].

With respect to these challenges, an attempt is made by the authors with a Systematic Literature Review (SLR) that explores the risks arising during the risk management processes of data science projects [17]. The SLR review calls for a further in-depth qualitative study to better understand the entire risk management process of data science projects, and the success criteria related to mitigating the risks that emerged during the execution of the projects.

To enhance the field's understanding of risk management for data science projects via a qualitative study involving 16 organizations, this paper explores how organizations identify and mitigate potential risks introduced due to data science projects. In short, this paper aims to analyze how organizations across various sectors identify, manage, and mitigate data science project risks. The paper explores the following overarching research questions:

- **RQ1**: What are the most effective risk management processes deployed in public and private organizations to manage the risk of data science projects?
- **RQ2**: How are the risk elements identified, monitored, and mitigated in public and private organizations?

The rest of this paper first provides a brief background on risk management processes in general and for data science projects. Then, the methodology of this research is explained. This is followed by research findings and finally a discussion of the conclusions and possible next steps.

2. Background

To provide some background context, this section starts with the definition of risk. Then, provides a review of organizational risk management and Risk Management Frameworks (RMF). This is followed by exploring data science project risks and uncertainty.

2.1. Risk Management Framework (RMF)

According to classical decision-making theory, risk is defined as "reflecting variation in the distribution of possible outcomes, their likelihoods, and their subjective values" [1]. Another definition of risk, by the Project Management Institute (PMI), is "an uncertain event or condition that, if it occurs, has a positive or negative effect on a project objective" [2]. Independent of the specific definition used for risk, the

URI: https://hdl.handle.net/10125/79479
978-0-9981331-5-7

intent or objective of risk management is to "reduce or neutralize potential [risks], and simultaneously offer opportunities for positive improvement in performance" [3][4].

In practice, organizations develop and maintain processes to try and handle visible risks that are measurable. For example, in healthcare where slow, small, and steady changes are desired, a logical incrementalism risk management model is used [5]. On the other hand, for construction projects where risk is a part of everyday life, a clear and exhaustive classification of risk that provides a comprehensive analysis is deemed appropriate [6].

Financial and regulatory constraints are some of the factors that compel organizations to manage their risk strategically. Enterprise Risk Management (ERM) is a process that acts as an umbrella term to monitor and manage all the major risks (such as Operational Risk, Market Risk, and Liquidity Tisk) in a top-down approach. To help address these risks, published frameworks have been created to evaluate, streamline, and monitor the processes within organizations. For example, in 2004, the first directive on an integrated framework was published by the Committee of Sponsoring Organizations (COSO) to acknowledge the governance control and enterprise-level monitoring of risk elements [7]. More recently, in 2015, the National Institute of Standards and Technology (NIST) Privacy Framework was created to identify and address the harmful consequences of privacy breaches [8]. Another RMF provides a lexicon of privacy risks, models, and objectives [9]. Several other RMFs have been defined for the cloud such as NIST-FISMA, and the SaaS model on ISO/IEC 27005 standard [10].

2.2. Data Science Project Frameworks

Launched in 1999, CRISP-DM (CRoss-Industry Standard Process for Data Mining) is the "*de facto* standard" for developing data mining and knowledge discovery projects [29]. While CRISP-DM defines five phases for a project, as well as documents to create during each phase, the framework does not explicitly discuss potential project risks and mitigating those risks. Refined methodologies, such as IBM's ASUM-DM, and SAS' SEMMA have tried to address the intricacy and heterogeneity of data science but have also not addressed project risk and the mitigation of that risk.

2.3. Data Science Project Risks

Data science projects introduce additional risks within an organization. A classic example of a business strategy that generated risks to an organization was the well-publicized reaction to Target's prediction of a teenager's pregnancy and promotional advertisements sent to the teenager's family [11, 12]. Another example was the application of machine learning to criminal justice, where a Florida county used the COMPAS recidivism prediction score to determine sentencing. This algorithm had false-positives and false negatives that created a disparate impact for Black Americans [13], which raised questions with respect to the notions of algorithmic fairness [14].

In yet a different potential incident, a data science team might be asked to develop a model to predict the healthcare cost of a prospective employee, by tracking and analyzing eating habits and exercise routines [15]. To do this project, the team needs to address questions that might increase the risk for the organization, such as the potential for an unfair algorithm (e.g., gender bias), and the risk of not adhering to data privacy concerns (e.g., mining "public" social media data to train models to infer personal attributes and identity).

These examples demonstrate some of the data science specific risks that a data science project might create for an organization. Thus, it is not surprising that many argue that Machine Learning / Data Science projects introduce new classes of risk to organizations [16]. These risk classes are different from the risk that are already visible and known to businesses. The unforeseeable and unmeasurable risks can be termed as uncertainty. Economists have studied uncertainty in terms of vulnerable macroeconomic variables that are subject to unforeseen shocks in the market. The unpredictability in the markets and business has helped economists create a clear distinction between risk and uncertainty. According to [27]:

"Uncertainty must be taken in a sense radically distinct from the familiar notion of risk [...] It will appear that a measurable uncertainty, or 'risk' proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all."

In other words, uncertainty is "unmeasurable, random and unpredictable" [26].

In our paper, we explore and explain the risk management processes that are operating in public and private organizations to manage known and measurable risk. The authors delineate the current risk management processes currently operational in private and public firms and direct attention to the risk that is uncertain, unknown, and unpredictable. Without exploring these risks, the use of data science could impact the reputational and economic well-being of that organization. Additionally, it has been found that the risks that might crop up with the use of data science

have often been neglected [17]. A different study highlighted the absence of robust methodology to develop analytic models [18].

3. Methodology: Inductive Approach

In this study, a general inductive approach was used to investigate how organizations identified and managed big data science risk. According to [28], "inductive analysis refers to approaches that primarily use detailed readings of raw data to derive concepts, themes, or a model through interpretations made from the raw data by an evaluator or researcher." To achieve this inductive approach, we conducted semi-structured interviews with 16 data scientists from private and public organizations. An interview protocol was drafted that kept in mind the diversity of the profiles of participants. Questions were generic and related to understanding the organizational hierarchy, processes to capture, manage and mitigate risks, and overall opinion on the deployed risk management process.

The recording feature of Zoom was used to record and auto-transcribe the interview. Text mining tool Otter was used to validate the transcribed data. As suggested by [26], the authors let the data speak for itself that eventually gave rise to dominant themes. No other "structured methodologies" were used that may have skewed the data with external biases. Specifically, themes of each participant were derived, which were connected with similar themes from other participants. Also, the themes that were important with specific sectors (conglomerates, manufacturing, sales) and business were captured.

3.1. Data Collection

Sixteen organizations were selected from private and public (federal government) sectors to be part of the study. The project was duly approved by the Institutional Review Board (IRB) with the permission to conduct only virtual meetings with the participants owing to COVID-19 social distancing constraints. A careful selection of interviewees via selective sampling was made to make sure that there was diversity across several theoretically salient factors [21]. Hence, a description of a potential interviewee's role, business domain, and professional experience was examined prior to sending a request for participation for the interview. The minimum years of work experience for the participants were two years, as the authors did not want to interview newly employed professionals owing to the nature and specificity of questions related to critical risks and the mitigation process.

Contextual details about each interviewee are summarized in *Table 1*. Specifically, *Table 1* shows the role of the person interviewed, aligned industry segment, the number of years of experience in that role (or similar roles) and risk managed.

Table 1: Interviewee Summary

ID	Role	Years Exp	Industry	Risk managed
1	Data scientist	25	Gov. - Defense	Tactical, Strategic
2	Data scientist	32	Conglomerate	Ethical, Budgetary, Timeline
3	Data scientist	22	Private supply chain (previously marines)	Model
4	Data scientist	30	Healthcare	Budgetary, Timeline, Headcount, Opportunity
5	Data scientist, Risk Manager	29	IT and Services	Market
6	Data scientist	13	Management Consulting	Headcount, Data Quality, Data Privacy
7	Data scientist	7	Finance	Data Privacy, Policies, Technology, Model
8	Enterprise Risk Manager	25	Gov. Veteran's affairs	Healthcare, Privacy, Finance, Policy
9	Data Scientist, Risk Manager	12	Finance	Model
10	Credit Risk Manager	16	Finance	Credit
11	Enterprise Audit / Risk Manager	30	Finance – Asset Management	Advisory, Business Continuity
12	Data Scientist / Project Manager	20	Private	Compliance, Data Validation, Data Quality
13	Risk Manager for Data Science Projects	5	Finance	Corporate, Credit, Model
14	Data Science Manager	8	Private	Business, Technology, Personnel
15	Data Scientist	30	Manufacturing (automaker)	Quality improvements, Cost, predictive maintenance
16	Data Scientist	19	Entertainment	Recommendation, Personalization

One-on-one semi-structured interviews were conducted via phone/video zoom calls. Open-ended, semi-structured interviews enabled the authors to ask probing and follow-up questions, allowing for a better understanding of the phenomenon under investigation. Each interview lasted for 30 to 120 minutes. Several times, participants agreed or requested an extended session to address all the questions of the interview.

The objective of these interviews was to collect information about how the organization identified and managed data science project risk. During each

interview, the initial questions covered the participants' background, roles, and responsibilities. The focus then shifted to understanding the interviewees' thoughts and practices concerning how their teams identified risk, and in general, the types of risk that were typically identified and then managed. In addition, the process of how teams identified risk as well as the process of how the team executed their data science project was explored. The interview ended with questions related to the sustenance and motivation of remote work during the COVID-19 pandemic.

3.2. Data Analysis

The analysis of the interviews leveraged the guidelines suggested by Braun and Clarke [22] for thematic analysis of qualitative data, which involved six steps: familiarizing oneself with the data, generating initial code, searching for themes, reviewing themes, defining and naming themes, and producing a report. The thematic analysis was conducted by both the authors who generated initial themes independently. The themes were divided into similar and dissimilar themes in a matrix across each sector by each author. Both matrices were later matched, and smaller clusters of themes were subsumed under broader themes. Brainstorming zoom calls were scheduled on a weekly basis between two authors to discuss obscure and outlier themes. Finally, a table of mutually agreed upon themes was created (*Table 2*) and dominant themes were retained.

4. Findings –Identify and Minimize Risk

Table 2 provides a high-level summary of how data science risk is managed.

Table 2: Currently Data Science Risk Process

ID	How Data Science Risk is Managed
1	Enterprise-level risk assessment via project management
2	7-step success criteria
3	Understand data accuracy, timeliness, repeatability, ANSII, OSHA
4	Define business goals, process improvement, saving cost, understanding the right project padding time
5	Understand model risk via backtesting, regression and classification algorithms, correlation testing.
6	Governance control, management, and quality of data
7	Calculating daily or weekly volatility, amount of leverage being used on investment. Also, look at previous project risks.
8	672-step risk management process integrated with CRM and internal control process across finance and IT
9	Monitor model performance in repository with incremental data
10	Get an understanding of the risk at the top level and then explore the more granular level of the identified risk. Basis the risk of the portfolio, client's requirements, approvals, credit scores are analyzed. Dodd-Frank capital planning regulation provides structure to report and manage risk.
11	A general assessment of risk within the organization – nothing specific for data science. Follows COSO lines of defense, ISO17009, COBID, SOX ISAE, SAS16
12	Monitor the risk levels (protect assets and client's interests). Not focused on model risks such as bias.
13	Highly structured process to evaluate and mitigate model risk for the firm (including model bias).
14	Ad-hoc risk management with informal meetings with data scientists, data analysts, business stakeholders, self-accountability
15	Four areas are addressed to mitigate risk: quality improvements, throughput improvements, cost reduction strategies, predictive maintenance. There is no formal way of addressing risk.
16	Define success and success criteria, minimum viable product to track success and goal.

Furthermore, as shown below in *Table 3* below, the analysis identified six common actions across two categories (identifying risks, minimizing risks). Each of these are actions are discussed in this section.

Table 3: Risk Identification and Minimization

Theme/ Category	Action	Identified in Organization	ID
Identifying Risks	Reviewing documentation (e.g., mapping pitfalls of previous projects)	Veterans' affairs, Finance, Media), and the conglomerate	ID8, ID10, ID12, ID2, ID13
	Asking questions (e.g., to identify risk, to define the success or decision criteria, risk and strategic outcome of the project)	Consulting, Finance-Bank, Media, Marine Corps, Entertainment, Healthcare, Defense	All except ID14
	Identifying Regulatory and Ethical Constraints	Finance (e.g., model risk, credit risk), Consulting, Conglomerate, Manufacturing	ID6, ID7, ID2, ID15, ID6, ID11
Minimizing Risks	Reducing Project Timeline / Cost Risk / Opportunity Risk	Consulting, Finance-Bank and Media, Marine Corps, Entertainment, Healthcare, Defense	ID2, ID4, ID16, ID3,
	Reducing Operational Risk (e.g., data errors, data labeling)	Finance - Model Risk, Private-supply chain, Finance -asset management	ID5, ID3, ID11
	Understanding & Distinguishing Data Science risks and general IT Risks	Finance and Media Entertainment, Finance - Banking	ID12, ID4, ID16

4.1. Identifying Risks

Many of the organizations focused on identifying project risk, but often not data science specific risks.

This section highlights the common approaches organizations used in terms of how they understand project risk.

4.1.1 Reviewing documentation

One approach an organization used to identify risk was to create and review documentation. For example, reviewing previous projects could be useful in terms of making sure mistakes were not repeated and previously identified risks were considered for future projects. In this situation, organizations viewed the risk management process as creating a knowledge base, such that all the teams could leverage the information. For example, ID2 stated:

"In my organization, we try to err on the side of caution in the sense that we would rather duplicate information than lose it. And we are quite heavy on documentation. So, there is a process that we have for doing a data science project. It has got seven major stages in it. And each stage requires documentation."

ID13, who worked at an investment bank as a Model Risk data scientist, focused on documentation in terms of how the model should work and then validating that behavior. Hence, ID13 noted:

"The first thing that one needs to identify is to do a risk assessment and to see where are the risks in this model…there is some sort of documentation associated with it, which kind of details the model scope, the criteria for success, the model performance, any benchmark model has been varied, and how the team is planning to properly monitor the models"

In a different example, for ID11's organization, the information on each project was captured and then shared with the senior stakeholders. The taxonomy was stored in a central repository that recorded all the risks of the data science projects run so far. This was demonstrated by ID11's statement that:

"…The central repository and system we use to manage and record all risks. And then we actually use data analytical tools to mine data from that system to help us…"

4.1.2 Asking questions

Another way to identify risk was via asking questions (either internally within the team, or an external group asking questions of the team) before the start of the project. The questions could be related to budgetary constraints, available resources, data availability, ethical concerns, and/or timelines.

For example, ID2's process of identifying risk started with seven steps before the commencement of the data science project, and the first step started with reviewing some of the important questions such as:

"What exactly is the problem… is this ethical? Do we have the resources to do this? Do we have the data to do this? Do we have the budget? Does the customer have the budget to pay for us?"

In a different example, the data scientist ID12 noted that the questions to be asked were dependent upon the objective of the data science project but the risks were identified with frequent discussions as noted by ID12:

"…it's all about communication with the business stakeholders … in that process [of discussing with stakeholders], there is a way to identify risk…"

Regarding the success criteria via a set of questions, ID16 noted:

"You set up the goal of the Big Data project, in very explicit terms, a very well-established definition of what that success what that project intends to do. What/How would you define success? … So, I start from there. Then another big component that I look at [ask questions] is with respect to the state of the enterprise data. What is the level of maturity? What is the quality? Is it siloed all over the enterprise?"

Finally, according to ID1, the following are some of the questions that were asked to understand risk elements for project risk management included:

- *Why are we doing this project? What do we expect to get from it?*
- *What is our strategy for rolling back the projects that don't work?*
- *What kind of measures do you put in place to do the rollback?*

4.1.3 Identifying Regulatory Constraints

Risk management is also driven by ensuring that the team adheres to laws and regulations. For example, the Federal Reserve's authority in the USA defines Matters Requiring Attention (MRA), which are informal practices and processes that "constitute matters that are important and that the Federal Reserve is expecting a banking organization to address over a reasonable period of time" [23]. ID10 specifically mentioned MRA and that the team followed the latter as a supervisory action to manage qualitative risk assumptions for capital planning purposes.

A different example, which was mentioned by ID6, was focused on the risk relating to data privacy and GDPR. With the importance of GDPR, and the nuances around it within many regions, the risk of how to keep the data, how people used the data, who used the data, and how long would the data be used was important to understand and manage in the risk management process. In order to address these challenges, the data scientist ID6 noted the process of impact assessment that was done for every data science project:

"…the impact assessment is predominantly looking at risk management, like understanding the

parameters of usage, understanding if data is properly classified, you know, either, if it's meant to be public if it's a private type of data. And if indeed, if it's restricted or internal. All of those things are becoming huge, including even technology around what you call cloud infrastructure. Where is the servicer sitting? What country? Does it align with the legislation and legislation as well as the policies government policy around data? And so really, it's about protection…".

4.2. Minimizing Risks

Many of the organizations focused on minimizing project risk, but often not data science specific risks. This section highlights the key areas of focus identified across the organizations, in terms of where they focus their efforts to minimize risk.

4.2.1 Reducing Project Timeline / Cost Risk

As noted by authors, within ID1's organization, minimizing project risk was focused on minimizing financial risk:

"*The project level risk management is typically related to identifying, and addressing the risk related to the financial investment and overall management of the project…*"

ID4's organization was focused on project timeline risk. Specifically, a strategy of time padding was used by ID4, who worked with a healthcare company. According to ID4, the time delay of a project was a key risk and the participant did not ask its dedicated team managers to assign a padding time. Rather, ID4 asked for the true estimated time and then added 15 percent on top of that time estimate to manage the time delay risk structurally. Another risk ID4 noted, in terms of timeline and cost, was the availability of knowledgeable resources:

"If you don't have skilled resources, it has a time effect. If you have steady resources, and they're sitting idle while you're doing the planning, there's a cost impact, you see".

In a different example focused on cost risk, ID1's organization risk management process was part of their internal control program. In short, their entire process of risk management was mandated by their Office of Management Budget. Under the mandate, as noted by ID1 below, all the projects were auditable that went through risk assessment criteria:

"So, one of the things that we are required to do, as part of what we call our internal controls program, monitors our risks from a fiscal perspective, mainly, and it's kind of done through the lens of auditability. So, there's a government-wide mandate to make sure all of our activities are auditable, which includes going through a risk assessment that we do on the front end for each project that we run"

4.2.2 Reducing Operational Risk

One of the risks that are often not visible but can become an impediment for the execution of a data science project is Operational Risk (OR). The importance of understanding the challenges of OR was highlighted by ID3, who used operational risk matrix approach to help manage their risk. In a different example, some of the ID1's questions specifically focused on minimizing operational risk. For example, ID1 focused on a fallback strategy:

"What is our strategy for rolling back the projects if it doesn't work? And what kind of measures do you put in place to do it? And then there are longer-term risks things like, you know, if we pivot to this approach, and we forget how we used to do things, and something goes wrong, and we have to go back to the old way, can we even reconstitute what we used to do to put that back in place?"

Managing risk by having a rolling back strategy highlights the concern with respect to implementation risk and the quality of the implementation.

Mitigation of OR was mentioned by ID10, with respect to deciding the appropriate monitoring infrastructure:

"You could have infrastructures where you, you could say, look, I use the underlying infrastructure … now suddenly [you] don't need a lot of expensive infrastructures, so the Operational Risk is sort of mitigated. So, I think there will be a lot of these interesting dynamics that will sort of play out in the future."

4.2.3 Understanding Data Science Specific Risks

Many organizations did not specifically mention data science risk. However, finance organizations typically discussed Model Risk. For example, ID10 reported this view with respect to model risk:

"…once you say something as a model … you have to document it. And so that is basically what the developers do…"

Another example of a data science specific risk, which was not noted by most organizations, was the risk noted by ID1 on the potential dependence on an automated process that replaced a human process. As this comment showed, understanding this dependence is a part of their strategic risk:

"…if we ever have to go to court, or there's a lot of people that [can] sue us in court, over personnel decisions made by the machine, what do we have to do to make sure that what we are doing is both ethical and legal? And what kind of strategic risk does that engender for the agency? So those are the kinds of federal levels of risk I have to consider when we apply projects…"

5. Findings – Organization Attributes

As summarized in *Table 4*, two organizational attributes were identified that impacted the data science risk management process: the organization's risk management process and the organizational structure. Each is discussed below.

Table 4: Organization Attributes

Theme / Category	Alternatives	Identified in Organization
Risk Management Process Maturity	Ad-hoc	ID12, ID14
	Well-defined (based on a standard framework)	ID6, ID8
	Well-defined (custom framework)	ID1, ID8, ID12
Organizational Structure	Centralized / Hierarchical	ID1, ID8, ID9, ID13
	Hybrid	ID6, ID11
	Decentralized	ID2

5.1. Risk Management Process

An organization's process for managing risk was one of three identified alternatives, each of which is described below.

5.1.1 Ad-hoc

An ad-hoc process is when the organization defines a risk management process framework for each project. For example, ID12 described their process, which focused on communication with stakeholders to identify risks:

"So, there's no specific common way to identify risk solutions in each individual project …. It's all about communication with the business stakeholders and defining requirements and gathering that internally developed delivering outcomes. So, in that process, there is a way to identify risk."

For ID14, there was also no particular risk management process for the team, however, the team did realize the importance of identifying and monitoring the team. So, there were ad-hoc meetings scheduled with the relevant team members who discussed the process as the need arose.

"It's probably talking amongst themselves the data scientists and data analysts, the data governance, people, the technologies, the business stakeholders, all of them. So, there is no systematic way of establish[ing] any kind of risk management process per se. Its heavily ad hoc meaning varies from project to project, initiative to initiative. So by and large, what I would say is there is a broad recognition of the idea of risk meaning deploying, say, for instance, the wrong model, but there has been no concerted effort to put together a systematic or a methodical process to follow"

ID15 who worked for a manufacturing company was not particularly focused on project risks. ID15's team did not have any RMF to follow. The data science team and the stakeholders worked in tandem with each other. The stakeholders' handed over a project to the data science team to execute or the data science team proposed a project to the stakeholders, but the risk was never a key aspect of their discussions.

5.1.2 Well-defined (based on a standard framework)

While no organization used a standard framework, there were several organizations that used a custom framework that was a refinement of one (or more) existing standard frameworks. Teams often thought of their frameworks as enhancements to those existing frameworks.

For example, in the governmental agency of ID8, the framework was proprietary but was based on existing frameworks. In one instance, ID8 mentioned that their framework borrowed from several standard risk management models (e.g., parts of OMB, COSO, hybrid model, NIST 30310, and NIST 3000). As this organization combined several risk management frameworks to create their specific framework. As noted by ID8, it was typically a proprietary and centralized risk management framework:

"We have one framework that is used across the entire organization … I built the one that is out of headquarters and supported by a chief of staff, and the Secretary's mandate, that we'll do enterprise risk management for the director".

In a different example, ID6 noted that having RMF that was based on standards but refined for the organization, was more effective:

"If you start to introduce standards that are more kind of interpretable to businesses and process, as well as the guidelines, then you start to get to the nitty-gritty of day to day in terms of how to actually have an impact".

ID12's organization maintained a guide that was highly customized for their projects. ID12 mentioned NIST, however, and also emphasized that the standard was not replicated as was, but highly tailored with the needs of the projects.

"We basically have incorporated a lot of those standards already into our process, and then we enhance them or adjust them to our needs …And so within the silos of each of the individual departments like cyber risk, we would adapt and be closer in line with us international standard versus operational risk, which is mostly my focus, that would be loosely coupled to a specific standard because we find that a lot of the things that we want to accomplish, especially

around data-driven decision making, and automated data driven decision making, doesn't necessarily exist in the industry".

5.1.3 Well-defined (custom framework)

Several organizations used a custom framework. ID8, from Veterans Affairs, described their risk management process, which was a well-defined process with 672 steps:

"…believe it or not, it's a 672-step process that goes through all the phases of it. But it's really broken down and who, what, where, when, and why leads to a racy chart of input, and has hyperlinks to the actual artifacts and deliverables. So, no matter what phase you're in, you can at least do what I call a good 80%..."

So, for example, their risk assessment process included all the directives, guidelines, templates, deliverables, all built into the process. Initially, their risk policy was such that all identified risks were supposed to be mitigated, but the executives realized that if the cost to mitigate was more than the risk, they would keep that risk (these identified risks were never data science risks, such as bias in their models).

"[the] first thing we do … is we do an overall risk assessment program, every administration and staff office, so we can level set our risk tolerance, and appetites".

The risk management framework described by ID1 did not follow any particular international standard but was a well-defined custom framework. ID1 explained that the overall risk management framework started at the senior level within the organization. Their risk management framework had an internal control program (mandated by the Office of Management budget that monitors risk through the audit process).

5.2. Organizational Structure

The structure within the organization, for how they executed their data science risk management process, was one of three identified alternatives, each of which is described below.

5.2.1 Centralized / Hierarchical.

ID1 and ID8 both followed a centralized process. One of the reasons for having a centralized risk management process, according to ID8, was to facilitate a single meeting across all the teams. This approach is explained as follows:

"We have one framework that is used across the entire organization ...It's kind of a parent-child relationship."

A top-down hierarchical approach was also used by ID1, ID9 and ID13. For example, data quality rules were created by business that created the data.

5.2.2 Hybrid

ID6 followed a hybrid risk management approach with respect to centralization. As noted by ID6, the reason for having a hybrid approach was to keep in mind the nature of the company:

"…it's usually hybrid, because… many companies under one umbrella. So, a hybrid works, because it's the best way to kind of cater for different types of the business"

ID11's organization was domiciled in the Netherlands; however, the data science team was located in North America. The team, therefore, created a synergy between the European framework and idiosyncrasies the commitment of local regulations.

"I work for a company based in the Netherlands. So, a lot of our frameworks are inherently European…that's the basis of European models…COSO, but again, based on local needs and commitments…it's very applicable to the to the US firm."

5.2.3 Decentralized

Within ID14's organization, the idea of risk management was decentralized. According to ID14:

"I think most of the risk management ideas are completely decentralized."

ID2's organizational structure was also decentralized:

"*It's a distributed structure that we have in place because our department is quite big. So, we have about 100 people… distributed over 14 countries. And so, time zones play a role as well… so, it gets…it gets complicated real fast. But ultimately, the risks, of course, percolate up to the top. But yes, the individual risks are distributed"*

Upon inquiry if the decentralized structure worked, ID12 highlighted the niche expertise that each individual required to vet the project. The decentralized structure provided that room and scope of ownership and accountability:

"I think it [distributed structure] does work. And because people see it from their point of view, and I'm a big believer in expertise. And so, I think, for example, the budgetary risk should be borne by those people who are actually doing the job. So, I'm so far removed from writing lines of code, that I can't tell you whether it takes 100 hours or 200 hours to write something anymore… but the programmers know that very well. So, I have to delegate that risk to them, and simply asked, so how long is it? And I'm going to have to trust their answer. And so, I lack the expertise of making that call. And, on the other hand, you know, the developers are always trained to solve whatever problem is given to them"

6. Discussion

6.1. Summary

The main contribution of this research is the consolidation of implicit and explicit knowledge and insights, across various range of business domains, that are leveraged to manage data science risk. Specifically, in this study, we report on interviews with people across 16 organizations, where the focus of the interviews was to understand how the risks were identified and managed in data science projects.

One key finding is that many organizations, especially outside of the financial service domain, use an ad-hoc approach to identify and mitigate data science project risks. Furthermore, no framework was identified that enabled teams to easily identify and mitigate potential data science-specific risks. Another key finding is that many organizations that do focus on risk, think of risk in terms of project costs or technical implementation issues.

In short, data science specific risks are typically not a focus when teams execute data science projects. In fact, with respect to data science specific risks, if the risks were explored, that analysis was due to a data scientist taking self-accountability to make sure that the risks were understood and monitored. In other words, the risk was centralized on an employee level but decentralized on a team level.

Perhaps due to this lack of applicable framework, when teams did try to identify risk, an approach that was often used was to ask questions. These questions were asked, either as a part of the success criteria or more specifically, focused on identifying risks and viability of the projects. These discussions were useful to identify and manage risk. Hence, one of the key results of this study was to draft a set of questions that would be used for an organization's data science projects.

6.2. The Impact of COVID

As the interviews were conducted at the pinnacle of COVID-19 in North America, the authors enquired the data scientists how they motivated themselves at the time of global crisis and how that impacted their risk management. The consumption of professional network and knowledge sharing was the source of motivation for ID8:

"We share models, and that keeps my brain working… I do a lot of time on the virtual stuff… I'll build a model. And I'll put it up and like, let other guys tear it apart who are smarter than me"

ID4 talked about how nimble the companies were in terms of adapting the work from home set-up and how that introduced minimal risk. However, the concerns over mental fatigue were also brought up during the interview, and how that might create additional risks:

"I actually thought technology would be a barrier. Because all of a sudden, you know, organizations that are not capable of handling 1000 people to work from home sometimes your VPNs and whatever else, right? I was surprised that companies quickly adjusted and nimble enough. And zoom of the world made it so easy to meet online …So the productivity actually went up very much. I was actually happy. But I do think now I'm beginning to see the signs of this. I see the fatigue taking place".

6.3. Limitations & Potential Next Steps

One of the limitations of this study was the technical challenge of the audio quality during the interviews with some of the participants. For example, on a couple of occasions, the authors had to reconnect to receive a better audio quality of zoom calls, and the transcription of the discussions was sometimes a challenge due to the recording quality.

Another limitation was the availability of only male participants for the interviews. While female participants were identified to be interviewed, for various reasons, no females actually completed the interview. In addition, there were only 16 people in this study, all geographically located in North America. Hence, one next step will be a comparative study to explore if the opinions, challenges, and concerns of the risk management process during the execution of a data science project changed with gender or across other parts of the world.

Yet another limitation was that there were only sixteen organizations represented within our study. Other organizations might have had different practices. Hence, another next step is to conduct additional interviews to validate our initial findings. A different next step could be to conduct a quantitative survey across a wide range of participants, exploring these research questions across a broacher audience.

Finally, future work could be focused on developing a framework and methodology for identifying explicit data science risks as well as developing management processes at the individual, project, and executive levels for mitigating or eliminating data science risks. This future work could leverage an existing risk management framework and/or data science process framework. As a first step towards this framework, a survey could be conducted to understand the viability of a risk management framework for data science projects.

7. References

[1] March, J. G., & Shapira, Z. (1987). Managerial perspectives on risk and risk -taking. *Management science*, *33*(11), 1404-1418.

[2] PMI (2000). "A Guide to the Project Management Body of Knowledge." PMBOK Guide, 2000 Ed. Project Management Institute, Pennsylvania.

[3] Ward, S., & Chapman, C. (2003). Transforming project risk management into project uncertainty management. *International journal of project management*, *21*(2), 97-105.

[4] Rosemann, M., & Zur Muehlen, M. (2005). Integrating risks in business process models.

[5] Wiboonrat, M. (2011). Risk management in healthcare services. In *Proceedings of the International Conference on Security and Management (SAM)* (p. 1).

[6] Szymański, P. (2017). Risk management in construction projects. *Procedia Engineering*, *208*.

[7] Prewett, K., & Terry, A. (2018). COSO's Updated Enterprise Risk Management Framework—A Quest For Depth And Clarity. *Journal of Corporate Accounting & Finance*, *29*(3), 16-23.

[8] Hiller, J. S., & Russell, R. S. (2017). Privacy in crises: The NIST privacy framework. *Journal of Contingencies and Crisis Management*, *25*(1).

[9] Garcia, M., Lefkovitz, N., Lightman, S., Brooks, S., & Nadeau, E. (2015). *Privacy risk management for federal information systems* (No. NIST Internal or Interagency Report (NISTIR) 8062 (Draft)).

[10] Alosaimi, R., & Alnuem, M. (2016). Risk management frameworks for cloud computing: a critical review. *International Journal of Computer Science & Information Technology*, *8*(4), 1.

[11] Asadi Someh, I., Breidbach, C. F., Davern, M. J., & Shanks, G. (2016). Ethical implications of big data analytics.

[12] Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of business research*, *69*(2), 897-904.

[13] Angwin, J., &Larson, J. (2016). Bias in criminal risk scores is mathematically inevitable, researchers say. Retrieved Feb 1, 2017, from:, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

[14] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).

[15] Andra Gumbus and Frances Grodzinsky. 2016. Era of big data: Danger of discrimination. SIGCAS Comput. Soc. 45, 3, 118–125.

[16] Tiell, S., & Metcalf, J. (2016). The Universal Principles of Data Science Ethics. Accenture Labs.

[17] Saltz, J., & Lahiri, S. (2020). The Need for an Enterprise Risk Management Framework for Big Data Science Projects. In *9th International Conference on Data Science, Technology and Applications, DATA 2020* (pp. 268-274). SciTePress.

[18] Grossman, R. L. (2018). A framework for evaluating the analytic maturity of an organization. *International Journal of Information Management*, *38*(1), 45-51.

[19] Irani, Z., Ezingeard, J., Grieve, R. and Race, P. (1999). Case study approach to carrying out information systems research: a critique, *International Journal of Computing Applications and Technology*, 12(2), pp. 190-198.

[20] Cornford, T. and Smithson, S. (2006). *Project Research in Information Systems: A Student's Guide*, 2nd ed., Palgrave Macmillan, NY.

[21] Eisenhardt, K. (1989). Building theories from case study research, *Academy of management review*, vol. 14, no. 4, pp. 532-550.

[22] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

[23] Eisenbach, T. M., et al. (2017). Supervising large, complex financial institutions: What do supervisors do? *Economic Policy Review*, (23-1).

[24] Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM*, 60(8).

[25] Turkay, C., Pezzotti, N., Binnig, C., Strobelt, H., Hammer, B., Keim, D.A., Fekete, J.D., Palpanas, T., Wang, Y. and Rusu, F., (2018). Progressive data science: Potential and challenges. *arXiv*:1812.08032.

[26] Müllner, J. (2016). From uncertainty to risk—a risk management framework for market entry. *Journal of World Business*, *51*(5), 800-814.

[27] Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

[28] Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, *27*(2).

[29] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.