CAPCO

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION

DATA MANAGEMENT

Data management: A foundation for effective data science

ALVIN TAN

DATA ANALYTICS

50TH EDITION | NOVEMBER 2019

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

CONTENTS

DATA MANAGEMENT

DATA ANALYTICS

DATA INTELLIGENCE

DEAR READER,

Welcome to the milestone 50th edition of the Capco Institute Journal of Financial Transformation.

Launched in 2001, the Journal has covered topics which have charted the evolution of the financial services sector and recorded the fundamental transformation of the industry. Its pages have been filled with invaluable insights covering everything from risk, wealth, and pricing, to digitization, design thinking, automation, and much more.

The Journal has also been privileged to include contributions from some of the world's foremost thinkers from academia and the industry, including 20 Nobel Laureates, and over 200 senior financial executives and regulators, and has been co-published with some of the most prestigious business schools from around the world.

I am proud to celebrate reaching 50 editions of the Journal, and today, the underlying principle of the Journal remains unchanged: to deliver thinking to advance the field of applied finance, looking forward to how we can meet the important challenges of the future.

Data is playing a crucial role in informing decision-making to drive financial institutions forward, and organizations are unlocking hidden value through harvesting, analyzing and managing their data. The papers in this edition demonstrate a growing emphasis on this field, examining such topics as machine learning and AI, regulatory compliance, program implementation, and strategy.

As ever, you can expect the highest caliber of research and practical guidance from our distinguished contributors, and I trust that this will prove useful to your own thinking and decision making. I look forward to sharing future editions of the Journal with you.

Lance Levy, **Capco CEO**

FOREWORD

Since the launch of the Journal of Financial Transformation nearly 20 years ago, we have witnessed a global financial crisis, the re-emergence of regulation as a dominant engine of change, a monumental increase in computer processing power, the emergence of the cloud and other disruptive technologies, and a significant shift in consumer habits and expectations.

Throughout, there has been one constant: the immense volume of data that financial services institutions accumulate through their interactions with their clients and risk management activities. Today, the scale, processing power and opportunities to gather, analyze and deploy that data has grown beyond all recognition.

That is why we are dedicating the 50th issue of the Journal of Financial Transformation to the topic of data, which has the power to change the financial industry just as profoundly over the coming 20 years and 50 issues. The articles gathered in this issue cover a broad spectrum of data-related topics, ranging from the opportunities presented by data analytics to enhance business performance to the challenges inherent in wrestling with legacy information architectures. In many cases, achieving the former is held back by shortcomings around the quality of, and access to, data arising from the latter.

It is these twin pillars of opportunity and challenge that inform the current inflection point at which the financial industry now stands. Whilst there is opportunity to improve user experiences through better customer segmentation or artificial intelligence, for example, there are also fundamental challenges around how organizations achieve this – and if they can, whether they should.

The expanding field of data ethics will consume a great deal of senior executive time as organizations find their feet as they slowly progress forward into this new territory. In my view, it is critical that organizations use this time wisely, and do not just focus on short-term opportunities but rather ground themselves in the practical challenges they face. Financial institutions must invest in the core building blocks of data architecture and management, so that as they innovate, they are not held back, but set up for long-term success.

I hope that you enjoy reading this edition of the Journal and that it helps you in your endeavours to tackle the challenges of today's data environment.

Guest Editor
Chris Probert, **Partner, Capco**

DATA MANAGEMENT: A FOUNDATION FOR EFFECTIVE DATA SCIENCE

ALVIN TAN | Principal Consultant, Capco

ABSTRACT

Data sourcing and cleansing is often cited by data scientists to be amongst the most critical, yet most time-consuming aspects of data science. This article examines how data management capabilities, such as data governance and data quality management, can not only reduce the burden of data sourcing and preparation, but also improve quality and trust in the insights delivered by data science. Establishing strong data management capabilities ensures that less time is spent wrangling data to enter into an analytics model and more time is left for actual modeling and identification of actionable business insights. We find that organizations that build analytics data pipelines upon strong data management foundations can extract fuller business value from data science. This provides not only competitive advantage through the insights identified, but also comparative advantage through a virtuous circle of data culture improvements.

1. INTRODUCTION

In the past decade, competitive threats from new market entrants, heralded by the digital revolution, are placing ever-increasing pressures on margins within the banking industry. New arrivals from the digitally-savvy fintech sector are free from legacy thinking and infrastructure, and traditionally non-banking organizations are increasingly looking to cross-sell financial services to their large existing customer bases. Both types of entrants possess substantial comparative advantages over traditional banks, which is causing a significant disruption of the banking landscape.

Whether it is seeking to gain an advantage or simply to protect market share and keep up with the competition, this has resulted in a rapid advancement of digital agendas at more traditional banks. Increasing digitization, of course, means increasing dependence on, and generation of more, data. Combined with data from existing 'analogue' operations, as well as access to a sea of current and historic market data, banks are increasingly looking for ways to make all their data work for more than it was originally intended.

It is against this backdrop, in the hunt for net margins and differentiation of products/services, that data science is fast becoming a key capability for old and new players alike in the industry. Customers are increasingly expecting a level of servicing (in relation to, for example, accessibility, availability, privacy, security, and personalization), that can only be effectively delivered through fundamental uplifts in the way data is handled and leveraged within the organization.

However, as this article sets out, maximizing the returns on investment (RoI) in data science requires (1) a scalable means of harnessing the hidden connections, correlations, and relationships in the vast quantities of data available, and (2) a business culture that readily accepts and allows data science to influence its business strategy. It is our belief that a strong and mature data management capability is crucial in achieving both objectives.

2. WHAT IS DATA SCIENCE?

Simply put, data science is the collection of analytic methods and tools by which business insights can be extracted from statistical and semantic relationships in data. Data allows an organization to both develop a deeper understanding of what has happened, and also make stronger predictions as to what *might* happen. Drawing upon a variety of disciplines covering applied mathematics, information technology, computational theory, and data visualization techniques, these methods and tools encompass the most basic of spreadsheet-based data analyses to complex machine learning (ML) and inferential artificial intelligence applications.

Financial services organizations (FSOs) leverage data science in a variety of ways to discover new opportunities and make data-driven decisions around risk management and operational efficiency. Use-cases range from developing better customer relationships, and understanding of preferences, to predicting employee behavior and detecting financial crime – data science can be applied in any function that generates or has access to data. The overall idea is that these insights can then be turned into actionable business strategies that would otherwise not be visible to an organization.

For all the zeitgeist, however, data science, as the name would suggest, is still a data-driven discipline at heart. Regardless of method or complexity, a common process exists for all data analytics processes (Figure 1): the data must first be sourced and prepared for inputting into the analytics, and the analytics output must then be evaluated by the data scientist who then communicates any insights to decision makers.

The implication is that the intended insights and business value of the analytics can only ever be as good and reliable as the data that underpins it. Or in other words, "garbage in, garbage out", and when it comes to data science, there is more than just a nugget of truth in this well-worn cliché.

Figure 1: Data analytics processes

3. LIES, DAMNED LIES, AND STATISTICS

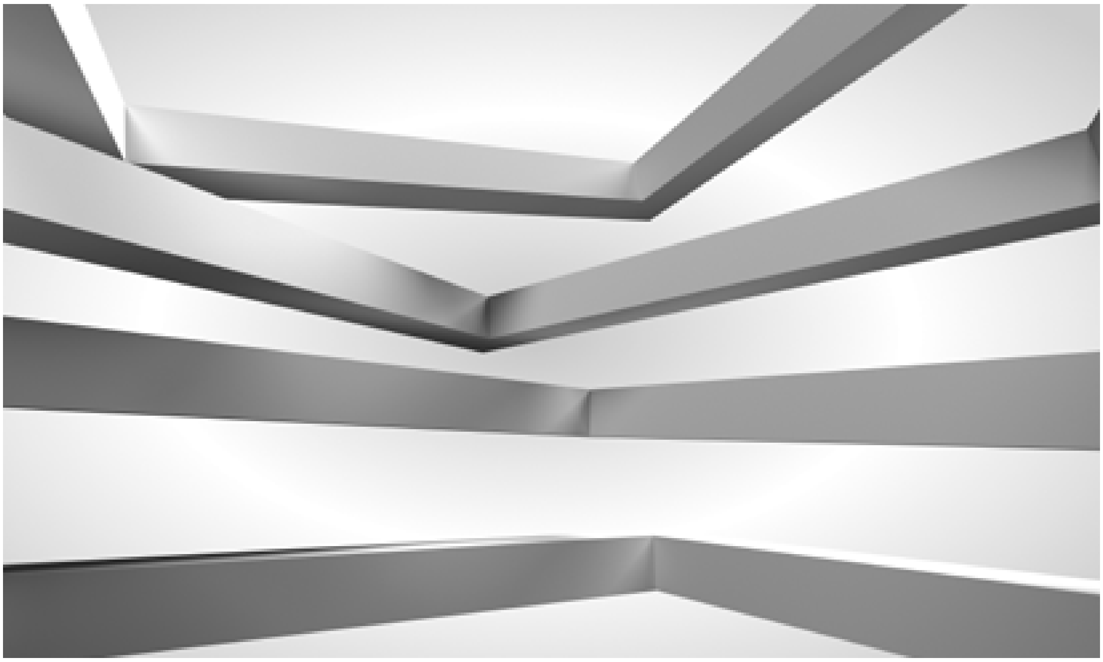
As a capability, data science is only effective if it ultimately provides positive value – the analytics results must serve a business purpose. Increasing the effectiveness of a data science capability means producing insights that can be trusted so that decision makers can turn these into strategies, which when executed produce business outcomes that are in line with the expectations set. This in turn drives a virtuous circle where data science is increasingly placed at the heart of an organization's strategic decision making.

'Garbage out' – incorrect, misleading, meaningless, or otherwise unusable data science output – causes non-optimal strategies and misguided business decisions at best, and financial and reputational damage at worst. This not only reduces the business value of the immediate results, but also erodes the trust that decision makers will have in future results, breaking the circle.

Here the issue of trust is key. Regardless of how powerful, accurate, or statistically reliable the results are, the data science capability itself needs to be trusted for decision makers to turn the analytics results into business strategies. Establishing, retaining, and growing this trust requires business outcomes that are consistently in line with the expectations set by the communicated results.

In the evaluation of data science insights, qualifying the results with a degree of *confidence* sets expectations as to how reliable the conclusions are. Confidence is provided quantitatively by an array of statistical measures, such as confidence intervals, p-values, and r-squared values. It is also provided qualitatively by descriptive interpretation of the statistical results, caveating the assessment with any risks to the reliability of results due to the assumptions made, model specification, sampling, or data quality issues. Together, these form the basis of trust between the data scientist and the decision maker.

Full and appropriate qualification of results with *known* reliability issues is simply good scientific methodology. Failure to evaluate results properly is something that should be vigorously guarded against by any number of educational, procedural, or ethical controls within the data science capability itself.



More damaging to trust, however, are the unknown unknowns – when there is an incomplete picture of the reliability and where this fact is not itself known. The causes may stem from issues associated with the scientific methodology, as well as from data scientists having misplaced assumptions in the *semantics*, *provenance*, and *quality* of the underlying data. The results cannot be qualified with something the data scientist is unaware of, and this unknowingly sets false confidence in the results.

This is even more pertinent with data science applications that involve probabilistic outcomes, such as machine learning. In such circumstances, the results are determined from a series of learned outcomes using training datasets. If confidence information is not built into the training process and the learned outcomes adjusted accordingly, the wrong outcomes are learned, and there will likely be significant systemic biases/errors in the final results.

In all cases where the reliability of outcomes is not clearly and accurately determined, significant damage to trust can happen. If analytics results are communicated and acted upon at face value, without knowledge of underlying issues in either the data or the analytics, business outcomes will likely become divergent from the expectations set.

In short, if making no prediction at all is better than providing a false one, then having no data is better than not knowing you have bad data. If data science is to be invested in as a strategic capability, then it is necessary to build trust in data science with decision makers. This not only requires the adoption of sound scientific methodologies, but also a cost-effective mechanism of ensuring data issues are managed, made known, and resolved.

These can be summarized into two key data management requirements for analytics processes: understanding and obtaining the right data, and fixing the data obtained.

3.1 Understanding and obtaining the right data

With model-led analytics (e.g., machine learning) the data scientist inputs data into an existing analytical model in order to ascertain its accuracy and viability. In this paradigm, the data scientist must first understand the *semantics* of what data is to be sourced so that the conceptual and contextual specifics of the required data can be specified. The data scientist must then determine where to source the specified data from, which requires an understanding of data *provenance* in order to ensure data is sourced appropriately.

Data semantics and data provenance are also crucial for data-led analytics such as data mining. In this paradigm, the data scientist identifies correlations within a given dataset and derives a theory or hypothesis from the observed results. As such, the semantics and provenance are not required to source the data, but to understand what and where the data has been sourced from so that the results can be appropriately understood and qualified.

In both paradigms, an understanding of data semantics and data provenance are critical for ensuring that the analytics has the *right* data:

- The data that is needed must be properly and unambiguously defined. To the uninitiated this seems like a trivial task, but the devil is in the detail and getting it wrong risks the analytics being run over the wrong data entirely. This involves identifying and establishing a shared understanding with potential data providers of what is required. If the data scientist wants 'customer name', for example, then an agreement must be made with the provider as to whether 'name of account holder' means the same thing semantically. In this example, there are many hidden nuances: does customer name include prospective or former customers? Does name of account holder cover mortgages, or current accounts, or both? Arriving at a mutual understanding is no simple task without a commonly agreed understanding of the definition, taxonomy, and *ontology* of the data.
- The data that is obtained must be representative of the population. An unrepresentative sample, for example where data obtained only represents specific subsets of the required population biases analytics outputs. As an example, if retail banking customer names are required, then it is important to ensure that the data is sourced from a provider that aggregates customers for all retail banking products, and not just, say, mortgages. Resolving this sourcing challenge requires not only accurate semantic articulation of the data required, but also an understanding of where this data can be reliably obtained.

3.2 Fixing the data obtained

Once sourced, data may still contain data quality issues that must be properly understood and resolved prior to analytics. Resolving and correcting for data quality issues is a data cleansing process that forms a critical part of the analytics preparation.

Poor quality data inputs can manifest in a variety of ways:

- Data may contain gaps, which if not corrected at source, accurately *inputted*, or omitted entirely, biases the output.
- Similarly, data may contain duplicates, which if not omitted will also result in biases.
- Data may not conform to an expected format, which if not corrected may at best break the analytics model, or at worst cause the results to become *heteroscedastic* (where the statistical results falsely suggest that the data comes from more than a single population distribution).
- Data may contain errors, which if not corrected will reduce the accuracy of the results.
- Data may be out of date, and the relationships inferred may no longer be applicable.
- Data may not be granular enough or sample size may be insufficient, both of which weaken explanatory power and the significance of outcomes.

To go back to our cliché, the 'garbage in' – incorrectly defined, inaccurate, incomplete, or otherwise poor quality data entered into an analytics process – is a primary limiting factor on the usefulness and reliability of analytics results. If providing quality inputs helps to ensure quality outputs, then having a cost-effective mechanism for understanding and resolving issues in sourced data is critical for improving the effectiveness of a strategic data science capability. This cost-effectiveness is provided by ensuring an effective centralized data management capability is in place.

4. MANAGING THE INPUTS

If what you get out of an analytics process is only as good as what you put in, then producing good outputs at scale requires cost-effective ways of controlling the inputs. For effective data science, it is just as critical to understand whether or not bad inputs exist, as it is to remediate them.

Good data scientists already know this.

Due to the criticality of ensuring an analytics process is provided with good inputs, data science projects often allocate a seemingly disproportionate amount of time, effort, and resources to simply preparing data for the analytics. The required data needs defining and describing semantically, trusted sources need to be identified, data quality needs to be measured, and issues identified and controlled. As we have already discussed, these are necessary activities to ensure that the end results are reliable and that decision makers continue to trust in the results.

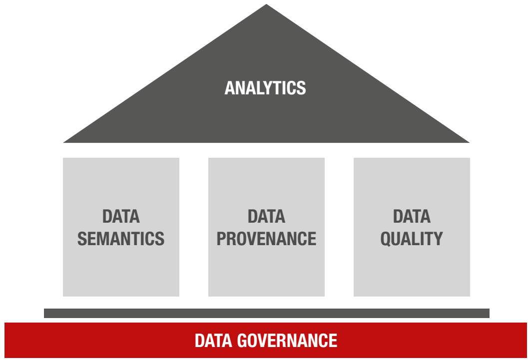
Defining what is needed, identifying where to get it, and data cleansing are, therefore, the data management requirements of analytics processes.

However, these are also hugely time and resource intensive activities. By some estimates, 80% of project time is typically spent preparing data for an analytics project.[1] Even for an organization actively seeking to become more data-driven, this is difficult to scale across more than just a handful of projects, and significantly raises the bar for a data science project to be viable through its benefits. In the bigger picture, organizations must find ways to minimize bad data provided to their data science projects, while also minimizing the marginal cost of doing so.

The answer is to ensure an effective data management capability is in place, providing the scale economies necessary for making more data science projects cost-effective.

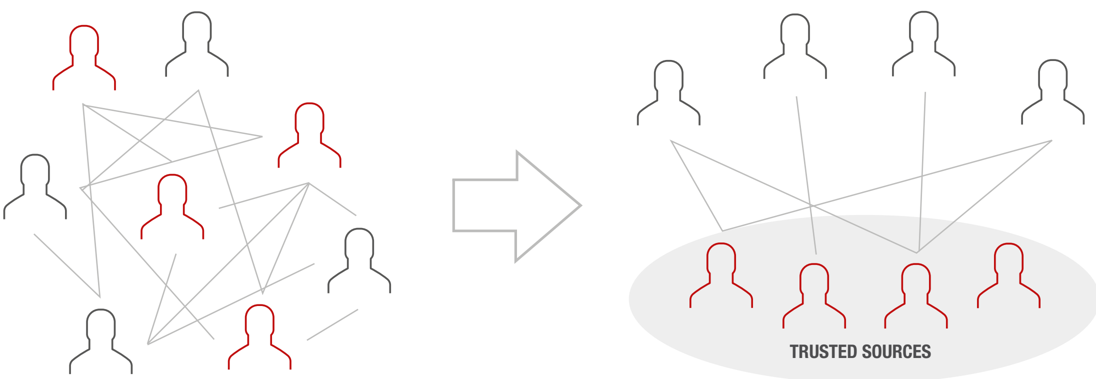
An organization's data management capability provides a set of centralized, scalable services for describing what the data means, for understanding and recording where the data comes from, for maintaining good quality data, and for ensuring the roles and responsibilities for data management are effectively discharged (Figure 2). Briefly, this includes:

Figure 2: Organizational data management capability



- **Semantics**: data is given commonly agreed and understood definitions, is placed in a commonly known taxonomy and ontology so it can be categorized accordingly, and semantic relationships between data is clear. Defining the semantics of data can also include conceptual modeling of data in order to understand the hierarchy, ordinality, and cardinality of data relationships with business concepts and data domains.
- **Provenance**: the sources of data, and path taken to where it is consumed, are identified and documented. Depending on the granularity at which this *lineage* is captured, this can involve identifying the aggregations and transformations en-route. Under provenance, sources of data can be certified as 'trusted' if applicable governance (see below) criteria are met.

Figure 3: Moving to centralized data management for data science



[1] CrowdFlower, "2016 data science report," https://bit.ly/2TtLN2c

- **Quality**: various quality dimensions such as completeness, conformity, consistency, validity, accuracy, and timeliness of data are measured and published/reported on a periodic basis. Issues are formally tracked, often against service level agreements defined against the material criticality of the data/process being impacted.
- **Governance**: the policies, processes, accountabilities, and responsibilities by which effective data management is defined, monitored, and enforced. Governance acts as a demand-management mechanism for ensuring data management activities are prioritized. Moreover, data governance provides an assurance to data consumers (such as data scientists) that governed data taken from trusted sources is well defined, meets minimum thresholds for data quality, and that data quality issues are formally managed and remediated.

Without a vision for streamlining the servicing of these requirements, an organization's data science can easily devolve into a web of hit-and-miss, fact-finding engagements between analytics projects and potential providers, as each project independently seeks to find the right data from the right sources.

A centralized data management capability provides the *hub* of data services and expertise that effectively allows all processes, analytics or not, to outsource their data management requirements. In such a setup, the centralized capability actively maintains a library of semantically defined data along with their trusted sources, allowing service users to quickly understand what they need and where to get it, avoiding unnecessary fact finding (Figure 3).

There are several benefits to this. Firstly, the data semantics (definition, taxonomy, ontology, and modeling) and data provenance (lineage and trusted sources) services offered not only free valuable time and effort for data scientists to focus on the actual analytics, but also ensure more reliable and explainable analytics results.

Secondly, it acts as a governing body for all data management in the organization and ensures that the outcomes are available for all processes. This allows for incremental gains as the knowledge (semantics, provenance, and quality) built from one project adds to the existing body of knowledge from others. From the data science perspective, the cost of data management is greatly reduced as data science projects benefit from the efforts of not only other data science projects, but also of the entire gamut of regulatory and transformational programs that occur in a modern FSO. For example, bad quality data is no longer remediated at the point of consumption by each data science project, but at the point of origination, therefore benefitting all consumers (data science and non-data science alike).

Thirdly, a centralized data management capability allows analytics processes and models to be defined in terms of a globally accepted semantic model. This allows for analytics results to be defined and communicated in a common business language, which in turn enables better interpretation and understanding of results amongst decision makers.

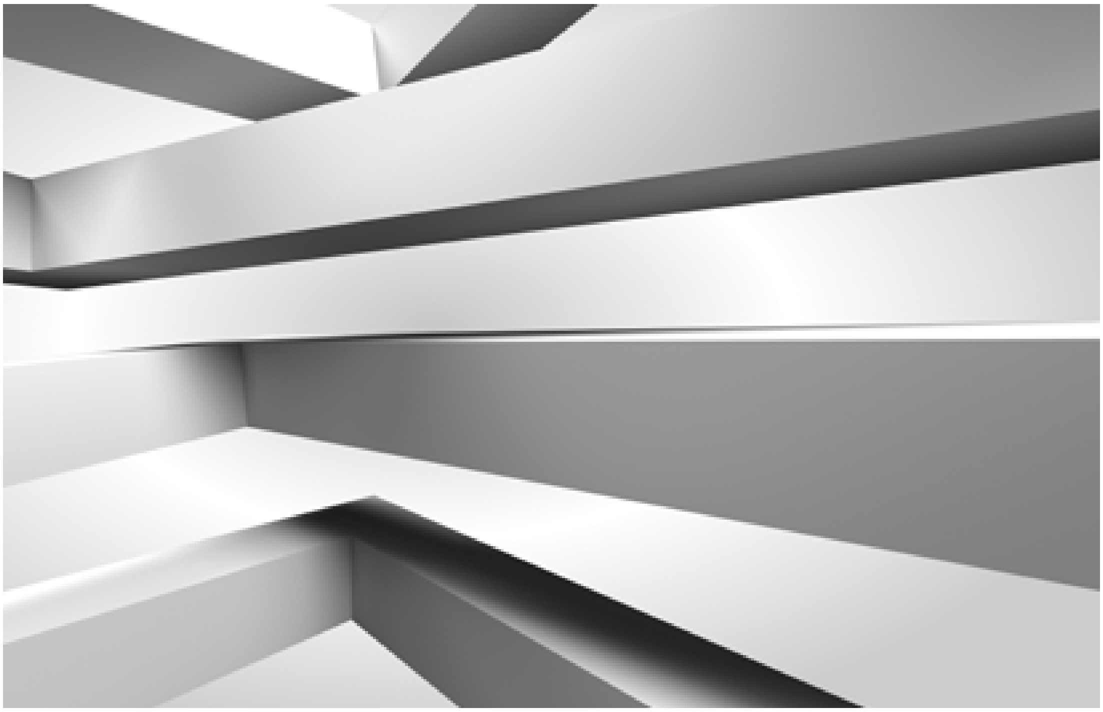
Figure 4: Building a strong data culture

5. IMPROVING THE DATA CULTURE

We have already described how trust in analytics outputs is key for driving an effective data science capability, and that significant components of this trust are reliant on the cost-effectiveness of ensuring analytics processes have 'good' inputs.

However, regardless of how trustworthy analytics results are, decision makers do not habitually act on these insights. This is especially the case with data mining insights that are often produced in financial services with little business sponsorship and poorly defined/planned business implementation.

What is often missing, therefore, is not just trust, but also the *willingness* of decision makers to take the insights on board and operationalize them. This willingness stems from an



inherent mindset or *culture* for data-driven decision making, where decision makers actively drive the data science process and are invested and interested in the outcomes. In a strong data culture, decision makers place data science output on equal footing to more traditional mechanisms, which are more reliant on experience and intuition.

An effective data management capability helps to foster a strong data culture. As previously described, data governance is a key data management service that ensures the effective discharge of data management roles and responsibilities. Crucially, this involves ensuring data owners and stewards are not only identified but are actively engaged in the governance and management of data. These data owners typically include the same decision makers that analytics projects provide insights to.

In this way, an effective and mature data management capability helps strengthen the data culture of an organization by actively involving decision makers in the governance of the very data that is used to provide insights back to the decision maker (Figure 4). This completes the circle – not only is trust greatly enhanced, becoming an implicit outcome rather than an explicit result of the data science, but it also helps to engender the data culture where decision makers are willingly at the heart of data-driven decision making.

6. CONCLUSION

Data science is effective when decision makers regularly make business decisions from the analytics insights, and the business outcomes are consistently in line with the expectations. These goals require trust and willingness on the part of the decision maker to operationalize the business insights provided by the analytics.

A data management capability helps build the willingness by fostering a data culture that puts decision makers at the forefront of data-driven decision making, and not data scientists. This is done through actively involving data owners in the governance of the data, which is used to provide insights to them.

Trust is built by ensuring business outcomes are consistently in line with expectations. This requires expectations to be properly set, which in turn requires the semantics, provenance, and quality of data inputs to the analytics be defined and known – 'good' inputs. While very time-consuming and resource intensive to perform for each data project in a silo, economies of scale are achievable by outsourcing these data management requirements to a centralized data management function.

Figure 5: Hierarchy of needs for data-driven decision making

In summary, more cost-effective, more reliable, and better understood analytics results build trust in the data science capability. Coupled with improving willingness of decision makers to operationalize analytics insights through mature data governance, implementing a mature data management capability is, therefore, essential in ensuring data science is cost-effective and has scalable impact.

In the hierarchy of needs, therefore, data management is the foundational layer for good data science and data-driven decision making (Figure 5).

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Hong Kong
Kuala Lumpur
Pune
Singapore

EUROPE
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO