Harnessing the Power of Unstructured Data with NLP and NLU

expert.ai

Introduction

Data has always been integral to enterprise business, but the rise of digital transformation has created a new sense of urgency around the ways in which it is managed, analyzed and governed. This newfound emphasis around data has placed the role of chief data officer (CDO) squarely in the business spotlight. In 2012, only 12% of large, data-intensive firms employed a CDO, whereas 65% do today, according to a NewVantage Partners survey.

With data being created and consumed at unprecedented rates, how you prepare for and address it today will define the future success of your organization. While the data boom does pose a daunting challenge, it also offers a unique opportunity to establish a competitive edge in the marketplace, especially when it comes to unstructured data (e.g., emails, PDFs, contracts, etc.).

Unstructured data is the white whale of the business world. It represents the large majority of enterprise data but is extremely difficult to extract value from. However, those that can do so effectively put themselves in a prime position to make more intelligent business decisions and operate with greater efficiency.

To make the most of your unstructured data, you must make artificial intelligence and natural language processing (NLP) top priorities. However, historical approaches to AI and NLP no longer suffice. To succeed, you need the right approach, the right expertise and a focus on the right data.

We know this because at expert.ai, we have been developing natural language solutions for more than 30 years. We have always recognized the need for the enterprise to access and understand unstructured language data at a human-like level of comprehension. The ongoing proliferation of unstructured data has created a newfound urgency among organizations to make this process a priority.

Our innovative technology and unique natural language understanding (NLU) platform address broad language data challenges, but for them to be most effective, we must first understand the natural language challenges and opportunities from the viewpoint of the CDO. This survey enables us to gain clarity on their challenges.

Marco Varone

*Founder & Chief Technology Officer, **expert.ai***

Executive Summary

This report, fueled by research of CDOs provided by Sapio Research, reveals how data teams are faring as they guide their companies towards AI success.

What are the opportunities and threats organizations face on the road towards becoming data-led businesses? What solutions are they putting in place to mitigate risk and improve efficiency? And, importantly, how are they measuring the success of their AI projects?

The vast majority of AI and machine learning technology adoption has been focused on leveraging structured data (highly specific data stored in a predefined format). Data teams have been reluctant to address the challenges presented by unstructured data and, more specifically, unstructured language data.

With that said, management of unstructured language data, by necessity, must be a top priority for chief data officers. Natural language technologies have advanced and CDOs are moving beyond being considered an analytics resource to finally partnering with the business to build scalable strategic capabilities.

The time to act is now. Unstructured data is growing, and this trend is not going away. You can expect more of everything from email to social media posts to digital business files. According to IDC, unstructured data is growing in volume by more than 50% every year, and by 2025 it will form as much as 80% of all data.

Analytics tools powered by AI have been created specifically to access the insights available from unstructured language data. Their ability to use the breadth of unstructured data to help organizations know themselves better and become more efficient is a revolutionary step towards becoming a data-led business.

Many data teams are clearly just embarking on their journeys into NLP and NLU capabilities. As this discovery continues, they need to understand AI vendor hype and identify proven approaches that can deliver business impact now. The data from this report will help them do exactly that.

Key Findings

The Business Imperative to Deliver Usable AI Solutions Quickly

CDOs have a clear mandate from their business counterparts – help us deliver business impact now by leveraging AI. Nearly all (96%) CDOs plan to cut through the AI hype and promises in the next 12 months to help their companies drive business value in the near-term without having to wait for AI technologies to mature. Not surprisingly, improving data security (92%) and getting value from unstructured language assets (91%) are also top priorities.

Unstructured Language Data Can Be Hard to Access and Understand

Most business-relevant information originates from unstructured data, much of which is text based (e.g., emails, support call transcripts, product feedback, customer survey comments, contracts, etc.). It presents a unique challenge for multiple reasons. First, it cannot be stored the same way as structured data (e.g., collecting it in column-row databases, spreadsheets, etc.). Second, language data requires an understanding of domain specific jargon and spans many languages globally. As a result, it is far more difficult to analyze and search through, making its value to organizations much harder to capture. That is, until very recently.

Choosing Your Starting Point: Platform vs. Cloud Solution vs. Open Source

There are different ways to handle unstructured language data. Just over one-quarter of data teams rely on a multi-vendor/multi-platform approach that combines open-source, platform and cloud provider solutions. Each approach has pros and cons related to costs, siloed efforts, and levels of quality and explainability. Our survey results show that companies take many approaches when working with unstructured data:

Platform Solutions: 44%

Open-Source Solutions: 34%

Cloud Vendor Solutions: 34%

The AI Knowledge Skills Gap Exists

Keeping up with quickly advancing AI technology and understanding whether their capabilities are fact or fiction takes a commitment from data team members. To close the AI knowledge gap and identify best practices, 48% of teams are hoping to upskill through training courses, followed by 31% who plan to recruit staff with expertise. When you look at data teams currently deploying AI-driven business solutions, 55% plan to recruit team members with expertise. Once planning is over, real-world deployment experience seems to be preferred by the majority of respondents.

How to Measure AI Project Success

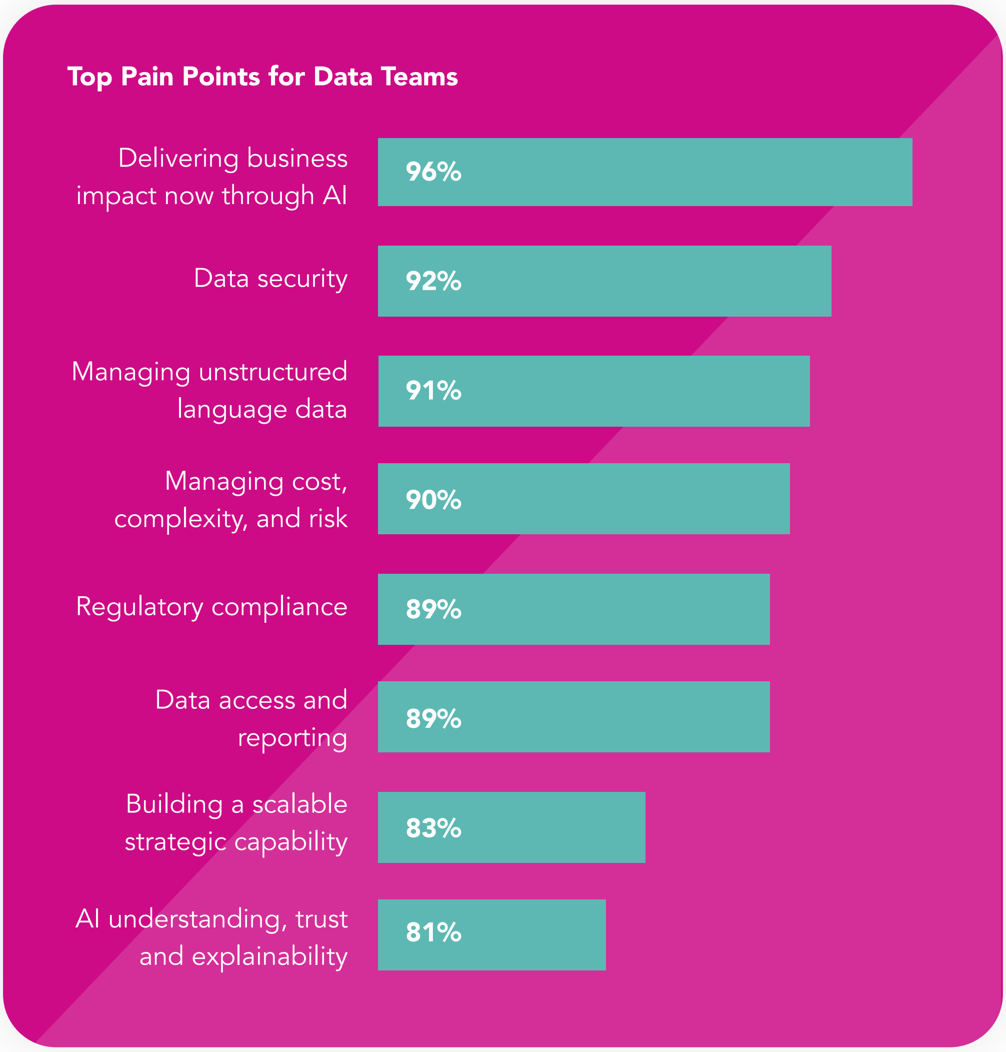
Data teams are delivering technical solutions to solve business problems and need their business counterparts to understand how success is measured. Almost 75% of data teams currently deploying AI projects are measuring project impact based on cost savings. Less than 5% of teams are without a plan to measure ROI and communicate business value.

Pain Points: The Issues that Hurt and How Companies Prioritize Them

The data-driven economy will be the defining force of the 2020s and beyond. Fueled by the growth of data-powered startups, its impact will be felt across the enterprise landscape.

Startups are disrupting established industries by employing the latest data-driven technologies to bring their innovative solutions to market. Meanwhile, established companies are competing with these newcomers by using data to help them to test and iterate potential opportunities more rapidly and at far broader scale.

McKinsey has stated that businesses built on the back of analytics outperform those without a data strategy by nearly two times. This is why so many companies are now leveraging data to earn or keep their place in this digital economy. The opportunities are endless, but with this very real revolution comes a raft of growing pains.



Top Pain Points for Data Teams

Delivering Business Impact Through AI

Nearly all CDOs (96%) see delivering business impact as their top concern in the next 12 months. This is not terribly surprising when you consider all the hype for AI in recent years accompanied by a lack of tangible results.

Business stakeholders are looking for CDOs to cut through the AI hype and noise and deliver immediate business impact. They are looking for real-world solutions that offer near-term benefits using proven AI technologies that are available today.

Data Security

The second most prevalent pain point is data security, with 92% stating that it is high on their 12-month agenda. Whether transactional, contractual, customer interaction, social or sensor data, companies have ownership of an increasing number of hugely valuable data sets – ones which are increasingly being targeted by criminals.

No surprise then that data security has become paramount to the drastically increasing number of data teams and projects. Data security is also the leading issue (56%) that data teams are seeking to resolve/improve upon with their current AI projects. Not too surprisingly data security is a critical aspect of any technology project, given the increased regulatory focus on data governance and the nearly daily stream of cybersecurity and ransomware attack news.

Managing Unstructured Language Data

More than 90% of CDOs agreed that management of unstructured language data (e.g., text from business documents, emails, etc.) must be addressed in the next 12 months. Your ability to extract value from this data is what will separate you from your competition via better NPS scores and reduced manual document handling and extraction costs. NLU technology is a key to doing so.

Managing Cost, Complexity and Risk

As companies generate more data and the ways they want to use it become more complex, data operations teams face an increasing number of challenges. More often than not, these challenges can be boiled down to three things: cost, complexity and risk. When asked which issues they are seeking to resolve with the AI projects they are currently undertaking, cost, complexity and risk were cited by 90%.

Building a Scalable Strategic Capability

Data teams are integral to executing core business initiatives across numerous business functions (e.g., increasing ROI in day-to-day operations, uncovering new opportunities, driving strategic initiatives, etc.). However, for them to meet the growing needs of an organization, they need a scalable alternative to hiring more knowledge workers to process documents and unlock unstructured data. More than four in five CDOs (83%) agree this should be addressed within the next 12 months.

Approaches to Handling Unstructured Data

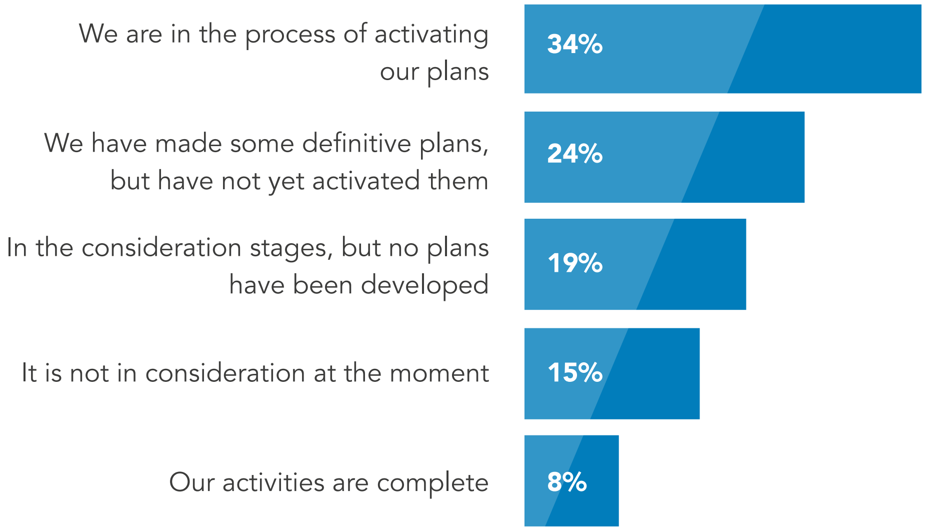
By definition, unstructured data is any information without a predefined data model or that is not organized in a predefined manner. It is typically text heavy but may also include information such as dates, monetary figures and more (e.g., email, client contracts, customer service interactions, etc.).

Unstructured data can be extremely valuable to any organization, but its value is notoriously difficult to extract. This is because machines inherently struggle to make sense of ambiguous language and domain-specific terminology within your data.

Thus, it is critical that organizations manage and analyze their unstructured data effectively so they can make more informed business decisions. This will enable them to prosper within their highly competitive environments. If they ignore this information, they miss out on insight that could potentially transform their business and lead to higher levels of success.

Our data shows that **34% of teams are activating their plans for NLP projects while another 24% are still defining their plans but are not quite ready to activate them**. As data teams move from consideration to planning and ultimately activation, there are many decisions to be made. Two of the most important are:

- What AI approach are we going to use to unlock the value of our unstructured language data?
- What type of NLP software will provide the best quality, lower costs and speed for our project?



How would you describe where you are in terms of your thinking about NLP and NLU?

Unlocking Language Data: Combining Machine Learning and Rules-Based AI Techniques

Most of the natural language approaches available today rely heavily on machine learning. While this is the most well-known method, it is not always the best solution. A growing number of data teams are starting to recognize that there are other options. For example, 21% of CDOs are finding success with hybrid natural language platforms that enable them to combine AI techniques in a more customized approach.

Machine learning requires significant amounts of data, time and processing resources. It also has a knowledge problem, meaning it can learn patterns within data but struggles to detect nuances in language, limiting accuracy potential. In addition, the rationale behind a machine learning model's output is unexplainable and often subject to bias. All that said, machine learning does possess the invaluable ability to quickly process large quantities of data. Makes it easy to scale.

Another method for working with unstructured language data is to leverage rules-based techniques known as symbolic AI. This knowledge-based approach relies on domain expertise from human subject matter experts and as a result, symbolic models provide complete transparency into how the outcomes are achieved. Data teams often choose this approach because a knowledge-based approach delivers higher levels of accuracy and lower computational costs, since rules-based techniques don't require large quantities of data. However, since this approach is based on human expertise, it can be more challenging to scale and not as fast as machine learning methods.

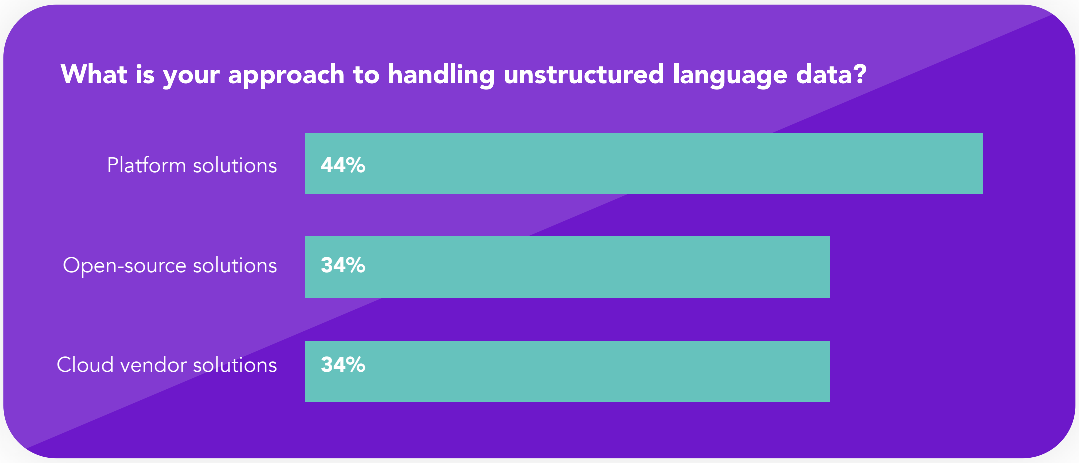
Options for Working with Unstructured Data

There are many types of NLU solutions with which to address unstructured data. According to our survey, the three most popular choices are platform, open-source and cloud vendor solutions. Interestingly, there is not a huge disparity between the usage of each solution type.

Platform Solutions: 44% are using platforms, including machine learning and hybrid NL solutions such as expert.ai

Open-Source Solutions: 34% are using open-source platforms such as Open NLP, SpaCy and Huggingface

Cloud Vendor Solutions: 34% are using cloud vender solutions such as Microsoft, AWS, Google and IBM



What is your approach to handling unstructured language data?

There are pros and cons to every solution. Some of the primary considerations include cost, the quality of data available, and levels of expertise required to work with the tools.

Platform Solutions

Pros

- Certain platforms support hybrid AI, combining machine learning and symbolic approaches
- Low-code/no-code solutions make platforms accessible to business users, not just data scientists
- High level of customer support

Cons

- Machine learning-only approaches can create high computational costs associated with managing large quantities of data
- High initial investment relative to other solutions
- Compatibility issues with existing workflows and infrastructure

Open-Source Solutions

Pros

- Open-source code is freely available
- Many libraries available
- Access to a community of peers

Cons

- No product support
- Steep learning curve and time-consuming to set up (e.g., an interface must be created by data team with coding skills and AI knowledge)
- Results may be difficult to explain and are dependent on the quantity/quality of data available

Cloud Vendor Solutions

Pros

- Simple, easy-to-use tools if you can code in Python
- Capacity to expand quickly
- Addresses specific use cases rather than IT infrastructure as a whole

Cons

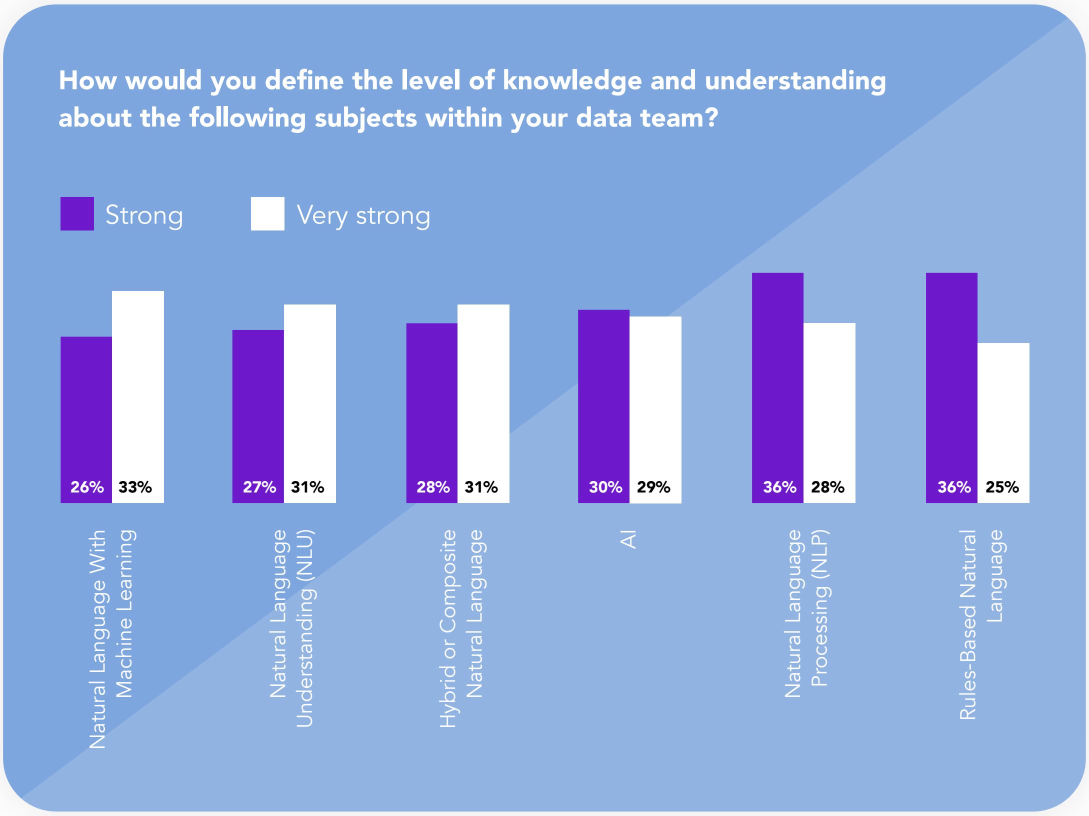
- Creates data silos which need to be managed independently of on-premise solutions and may require different software stacks
- Limited tools can make design cumbersome
- Results may be difficult to explain and are dependent on the quantity/quality of data available

Levels of Knowledge in NLP and NLU and Plans for Upskilling

There is a growing realization across the enterprise that unstructured language data is not merely a by-product of their operations but a critically important resource that must be mined for actionable insights. While two-thirds of organizations claim to be knowledgeable about AI, they often lack employees with the tangible skills to build and execute successful AI programs.

Unfortunately, organizations have found it difficult to acquire the talent they require, whether it be through internal training or external recruitment. In fact, over the past two years, almost half (46%) of UK businesses have struggled to recruit for roles that require data skills, according to the UK Data Skills Gap 2021 report.

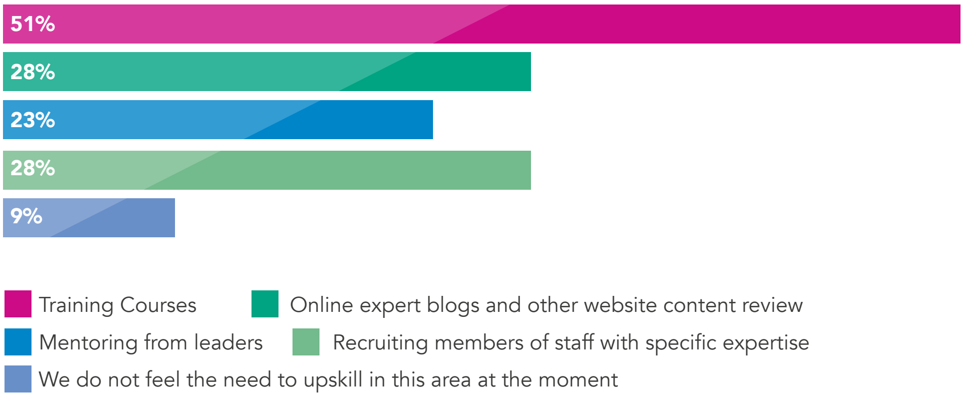
Given the low supply of data skills on the open market, it is no surprise that organizations looking to fill their skills gaps choose not to seek external expertise as their first option. With that said, the timing of an AI project is an important consideration. Companies that have made definitive AI plans but have not yet activated them are more likely to look for external expertise (58%) versus any other upskilling methods.



How would you define the level of knowledge and understanding about the following subjects within your data team?

In general, though, when data teams have gaps in their team members' knowledge and understanding, the most common solution is to upskill the team through training courses. This was the primary method for every specific knowledge area including AI (51%), NLP (41%) and NLU (35%). Few organizations felt online content or mentoring from team leaders were viable methods to bridging the skills gap.

How do you plan to upskill your team in AI where they have gaps in their knowledge and understanding?

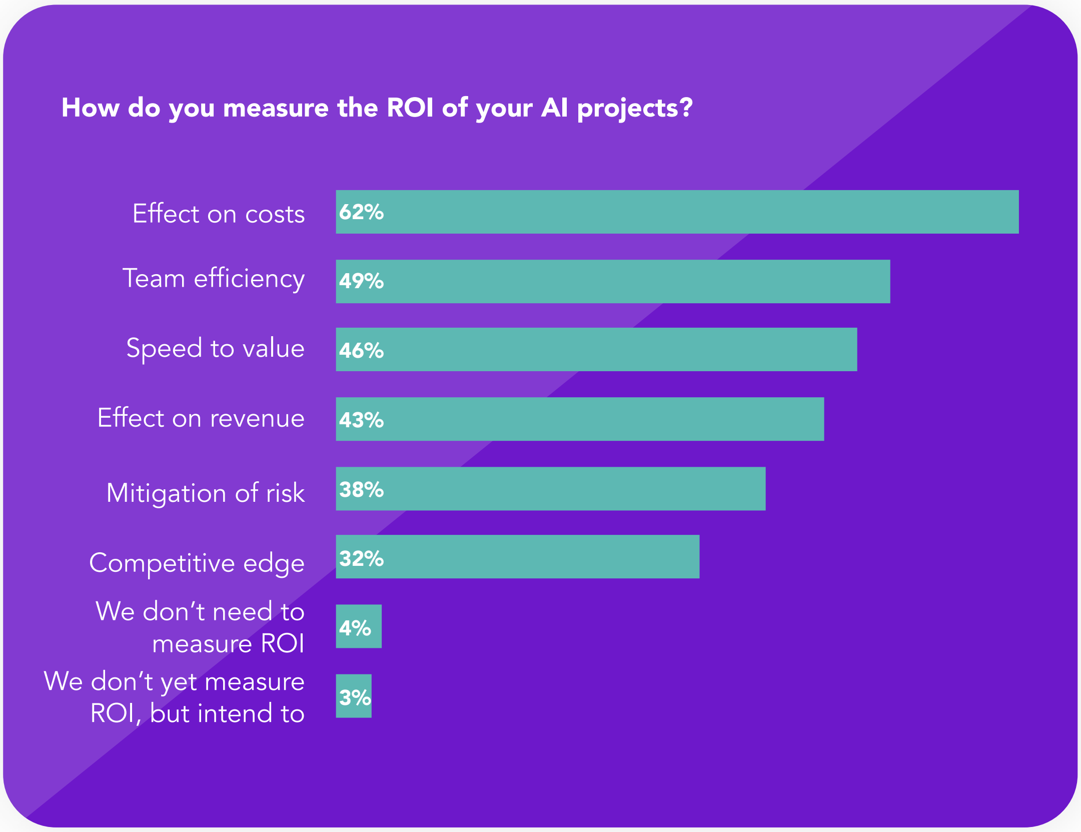


ROI: Challenges with Measurement

AI is transforming the enterprise in every industry and we have only scratched the surface of its potential. In many ways, what prevents organizations from reaching that potential is a lack of clear, standardized success measurements.

According to McKinsey's The State of AI in 2020, just 22% of organizations reported quantified value from AI. What many of those companies have in common is they aligned their AI projects to business value from the outset. In fact, the same survey explains that high performers are nearly twice as likely as others to clearly prioritize AI initiatives linked to business value across organizations.

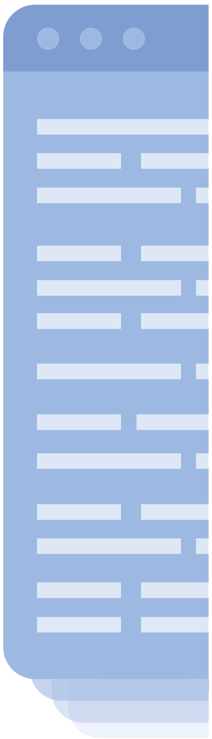
Ultimately, the key to achieving ROI is first defining how to measure it. This will, of course, differ from one organization to the next depending on their specific business goals. So, while there is no overarching set of performance metrics by which to measure success, there are some metrics that CDOs leverage more frequently.



Effect on Costs

Investments are monetary by nature, so it is no surprise that the most common ROI measurement for AI projects (used by 62% of CDOs) is how they impact costs within the organization. Can labor costs be reduced by making current employees more efficient at manual, language-intensive processes? If your data team can clearly attribute its AI initiative(s) to hard costs, you have set yourself up for success.

Team Efficiency



Much of what an organization seeks to accomplish with AI is greater process and team efficiency. This could mean removing repetitive, labor-intensive tasks (e.g., email management, claims processing, etc.) from the day-to-day of employees and freeing them up to take on higher-value tasks. It could also mean enabling employees to execute those tasks more quickly. With enhanced employee efficiency, organizations can make more timely business decisions. Nearly half (49%) of CDOs choose to leverage the metric.

Speed to Value

Time is not a luxury every business has when it comes to developing innovative products and getting them to market. Thus, speed to value is important to many CDOs (46%) who seek to establish immediate trust in their initiatives as well as continued funding for them. This metric must go hand in hand with another such as cost and efficiency, as it needs a specific value with which to measure speed.

Effect on Revenue

Savings on costs are not the only potential financial benefit of an AI project. The effect on business revenue is also a valued KPI for 43% of CDOs. For instance, media organizations reliant on customer engagement to drive advertising revenue could measure the impact content personalization has on time spent on the site which directly impacts revenue generation. Likewise, information providers that deliver access to specialized data on financial markets can create new revenue streams with higher margins by automating the collection and summarization of content as a paid service.

Mitigation of Risk

Though it may not be your first consideration when it comes to measuring ROI, 38% of CDOs point to risk mitigation as a core KPI. Particularly for organizations in high-risk industries such as financial services and insurance, AI can be a catalyst for reducing risk exposure that can lead to lost profits or even bankruptcy.

What Does This Tell Us?

This report reveals key insights that all data teams should evaluate as they drive their businesses toward becoming data-led companies. It is no surprise that artificial intelligence is viewed as the vehicle for addressing data-centric challenges, especially those involving natural language. What is less clear to data teams is how to get started and what resources they need to ensure success. With that in mind, we have three recommendations for data teams as they look to optimize their unstructured data.

Select a Proven Vendor

With so many vendors to choose from in the AI and NLP space, finding the right one can be a daunting task. Each has unique benefits to offer, but whether they actually follow through on them is another matter. To ensure you get started on the right foot, choose a vendor with a proven track record. While their technology alone does not ensure AI success, it does provide a solid foundation with which to build your model.

Explore a Hybrid AI Approach

Traditionally, natural language solutions were either built via symbolic AI or machine learning. While both offer their own capabilities and benefits, neither are without their shortcomings. A hybrid approach has emerged over the past year as the most viable solution, as it offers benefits of both symbolic and ML while limiting their weaknesses. Rather than sacrifice core capabilities up front, seek a hybrid solution that offers them all.

Address a Single Business Case Then Build a Strategic Capability

As the saying goes, you need to crawl before you walk. This holds true when it comes to building AI capabilities at your organization. Prove the value of your AI model by focusing on a single business case. Once you have established a sufficient level of success and gained the trust of key stakeholders, you can look to broaden your AI foothold across the organization to address further challenges and create additional value.

Closing Thoughts

Our recommendations are designed to help you make the best decisions along your AI journey. While these apply to every organization, no organization is exactly the same in its needs and circumstances. Before diving headfirst into your AI journey, be sure to take stock of your own organization's situation. These questions are a good place to start:

1. How seriously are you taking the issue of unstructured language data? What technologies are you utilizing to draw as much information from unstructured data, and could those technologies be doing more?
2. Where are you focusing the majority of your time, and how could resources be better deployed? What technologies can help you optimize time-intensive tasks?
3. What is the expertise level of your data team regarding NLP and NLU? What knowledge or skill gaps does your team have and how do you plan to fill them?
4. What KPIs are you using to measure ROI and do they support the goals of your organization?

Survey Methodology

The survey was conducted among 116 decision makers where data / analytics is a large part of their role. Research took place across the USA and Europe. The interviews were conducted online by Sapio Research in October 2021.

Get **Started**

Interested in learning how NLU AI will transform your company? Get started here.

See what expert.ai can do for you!

About us
Expert.ai is the premier artificial intelligence platform for language understanding that augments business operations, scales data science capabilities, simplifies AI adoption and provides the insight required to improve decision making throughout organizations. The expert.ai brand is owned by Expert System (EXSY:MIL), that has cemented itself at the forefront of AI-based natural language solutions across Insurance, Banking, Publishing, Defence & Intelligence, Life Science & Pharma, Oil Gas & Energy, and more.

www.expert.ai **info@expert.ai**