451 Research® | PATHFINDER REPORT

Unlocking the Business Value of Unstructured Data with AI/ML

COMMISSIONED BY

IRON MOUNTAIN®

JUNE 2019

About this paper

A Pathfinder paper navigates decision-makers through the issues surrounding a specific technology or business case, explores the business value of adoption, and recommends the range of considerations and concrete next steps in the decision-making process.

ABOUT THE AUTHOR

MATT ASLETT

RESEARCH VICE PRESIDENT

Matt Aslett is a Research Vice President with responsibility for 451 Research's Data, AI and Analytics Channel – including operational and analytic databases, Hadoop, grid/cache, stream processing, data integration, data governance, and data management, as well as data science and analytics, machine learning and AI. Matt's own primary area of focus currently includes distributed data management, data catalogs, business intelligence and analytics, data science management, and enterprise knowledge graphs.

PAIGE BARTLEY

SENIOR ANALYST, DATA MANAGEMENT

Paige is a Senior Analyst for the Data, AI and Analytics channel at 451 Research, covering data management, including data integration, data governance, data quality and master data management. She has experience covering a broad range of information management technologies spanning database functionality and self-service analytics to regulatory policy and compliance.

Executive Summary

Artificial intelligence (AI), including machine learning and deep learning, is set to become one of the most transformational technologies in the history of the world, affecting most aspects of our lives, whether we're conscious of it or not. The application of these technologies will likely reshape how people work, study, travel, govern, consume and pursue leisure activities.

While AI is potentially hugely transformative, it also requires specialist skills that are in short supply, and the ability of any individual organization to exploit the potential opportunities of AI will depend on its ability to hire and retain data science experts, as well as develop and deploy AI-enabled applications that enable senior decision-makers and domain experts to access and act upon the results.

Business use cases with demonstrable ROI will be the impetus that helps morph AI and machine learning from a set of isolated science experiments to a widespread driver of societal change. Fortunately, unstructured data – which nearly every organization has ample volumes of – provides a valuable and relatively unexplored corpus of information that is well-suited to be leveraged by machine learning models.

Both reactive and proactive use cases are possible. Reactive use cases will largely leverage automation and machine learning to identify, classify, govern and protect data for pressing needs such as shifting global compliance requirements. Proactive use cases, involving the leveraging of data, are even more diverse. Not only will organizations be able to automate away the tedium of repetitive tasks that generally are a poor allocation of human skills, but they will also be able to identify patterns associated with subject matter expertise: an opportunity to build better teams and curate tribal knowledge. Most importantly, these technologies will enable outcomes and insight that would not be otherwise achievable, via sheer processing power and the ability to detect and define abstract relationships.

Stronger control over internally held data will always be a net benefit to the business. With stronger mechanisms for data control and quality, downstream outcomes in initiatives such as self-service analytics and data science inherently improve. And increasingly, functionality driven by machine learning and automation will be fundamental to identifying, sorting, classifying and governing information at scale. Interpretation of that data, too, will depend on automated technologies. Both our control of data and ability to meaningfully derive value from it will be defined more by these technologies.

AI/ML for Unstructured Data

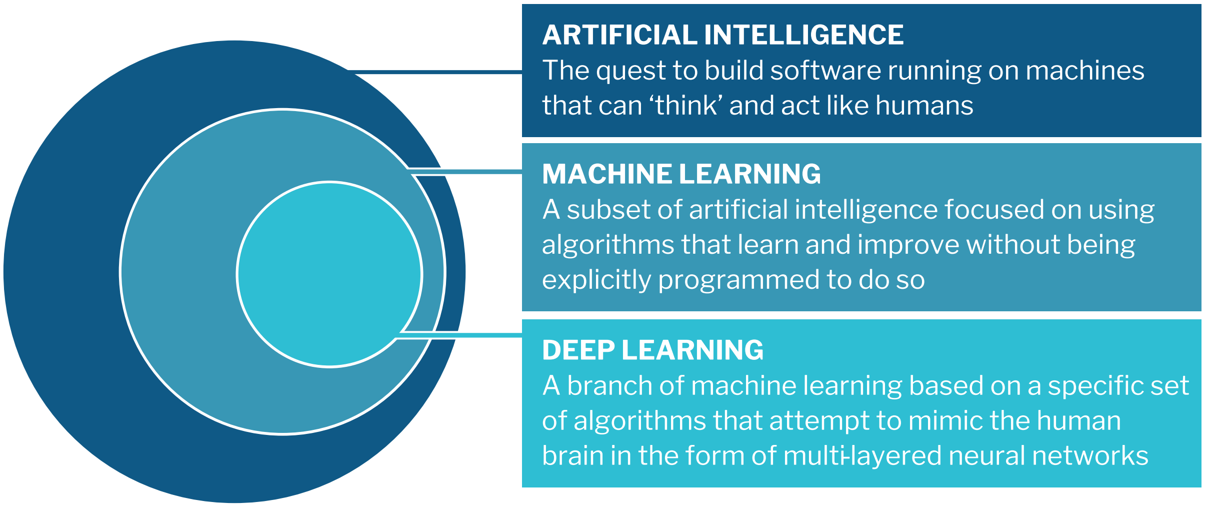
What is AI and How Does it Work?

The terms artificial intelligence, machine learning and deep learning are often used interchangeably, and not always correctly. Each term has a distinct definition, but there is some overlap. It is useful to think of AI, machine learning and deep learning as a nested hierarchy in which AI is the broadest category, representing the overall quest to build software running on machines that can 'think' and act like humans.

Machine learning is a subset of AI focused specifically on using algorithms that learn and improve without being explicitly programmed to do so, while deep learning is a further subset of machine learning and describes a specific set of algorithms that attempt to mimic the human brain in the form of multi-layered neural networks.

Figure 1: AI, machine learning and deep learning definitions

Source: 451 Research



ARTIFICIAL INTELLIGENCE
The quest to build software running on machines that can 'think' and act like humans

MACHINE LEARNING
A subset of artificial intelligence focused on using algorithms that learn and improve without being explicitly programmed to do so

DEEP LEARNING
A branch of machine learning based on a specific set of algorithms that attempt to mimic the human brain in the form of multi-layered neural networks

While potentially transformative, AI technologies are by no means novel, with conceptual and technological lineages going back decades, if not centuries. What is new, especially over the last three to five years, is the integration of machine learning functionality in business environments to improve and automate enterprise processes.

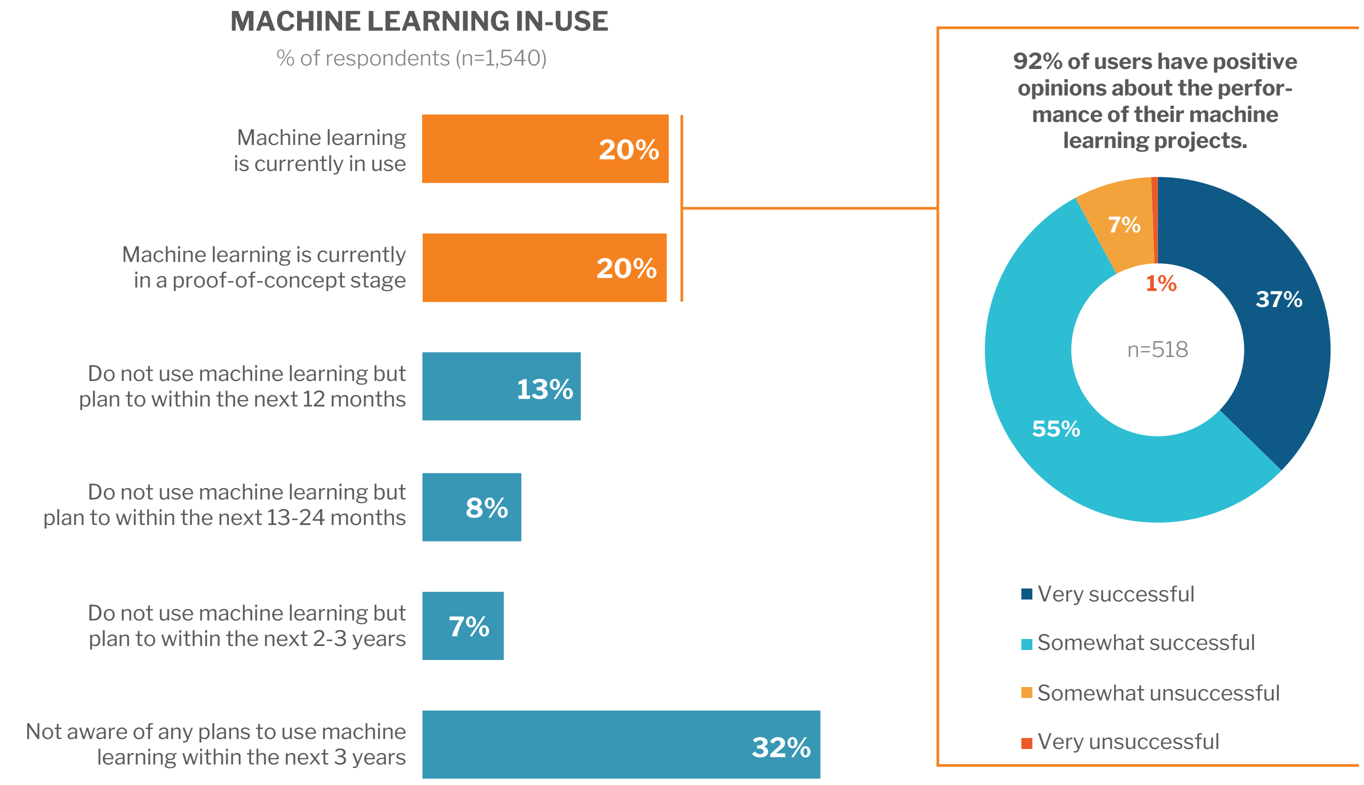
So why is machine learning popular now? There are several reasons. Machine learning lives or dies depending on the quantity and quality of the data, as well as the compute power that is available. Thanks to cloud computing, compute power is available as never before, while data is available in volumes never dreamed of. As a result, the combination of AI and machine learning is no longer a glimmer on the technological horizon or purely an academic pursuit but, rather, a real-world enabling technology with an increasing number of defined use cases and demonstrable return on investment.

Potential Enterprise Impact

The results of 451 Research's recent Voice of the Enterprise[1] AI and ML survey illustrate that the level of interest in machine learning is escalating. Machine learning is currently in use by 20% of enterprises; an additional 20% is in the proof-of-concept stage, and a further 21% are planning to adopt within the next two years.

Figure 2: Current uses of machine learning

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Adoption and Use Cases 2H 2018



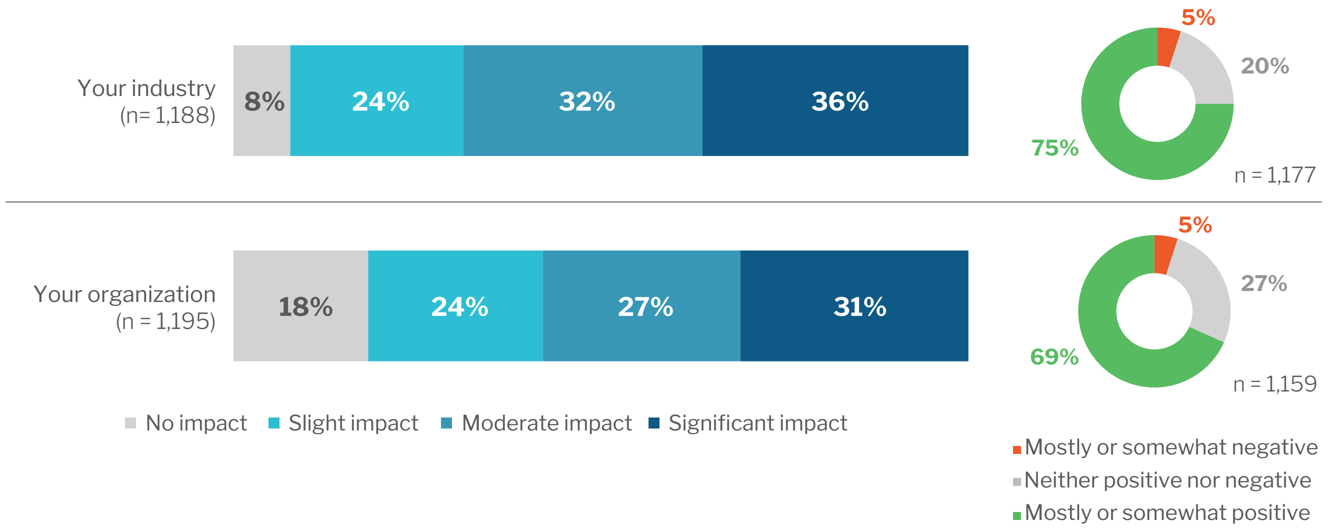
Significantly, 92% of those that are currently using machine learning believe that their projects have been successful. The survey results also show that respondents are largely bullish about AI and about its impact across multiple domains – whether that is the potential impact on individual respondents, the companies they work for, the industry in which their company operates or society as a whole.

For example, 58% of respondents said they expect that machine learning will have a moderate or significant impact on their organization, while 69% expect that any impact will be positive. Meanwhile, 68% expect that machine learning will have a moderate or significant impact on their industry, and 75% expect that any impact will be positive.

[1] This paper references data from 451 Research's Voice of the Enterprise (VotE) quarterly and semi-annual surveys. VotE surveys are fielded against 451 Research's proprietary panel of 60,000+ global enterprise IT decision makers, spanning numerous industries and IT roles.

Figure 3: Potential impact of machine learning

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Adoption and Use Cases 2H 2018

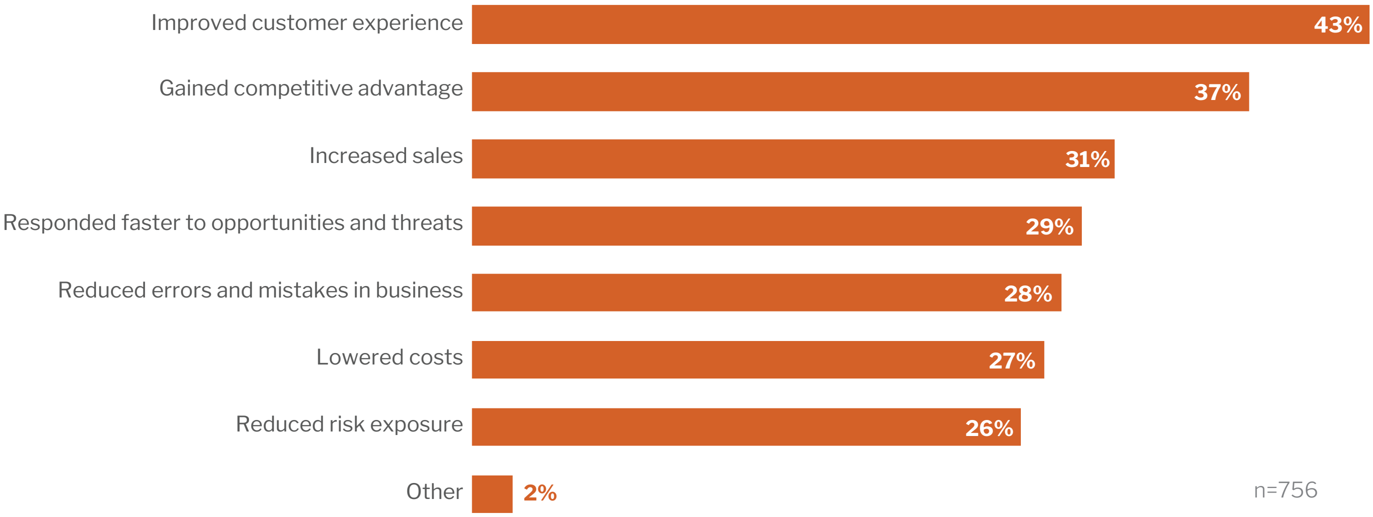


There are multiple benefits that organizations can potentially accrue from machine learning that lead to this positive perception. The most significant, according to 451 Research survey results, is improving customer experience, followed by gaining competitive advantage and increased sales. Other potential benefits include the ability to respond faster to opportunities and threats, as well as fewer errors, lower costs and reduced risk exposure.

The fact that relatively few respondents cited lower costs illustrates that, for most organizations, machine learning is not primarily about cutting people's jobs. Automating tasks will likely play a part in the future, but the indications at this early stage are that automation need not necessarily lead to wholesale job losses.

Figure 4: Benefits of machine learning

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Adoption and Use Cases 2H 2018



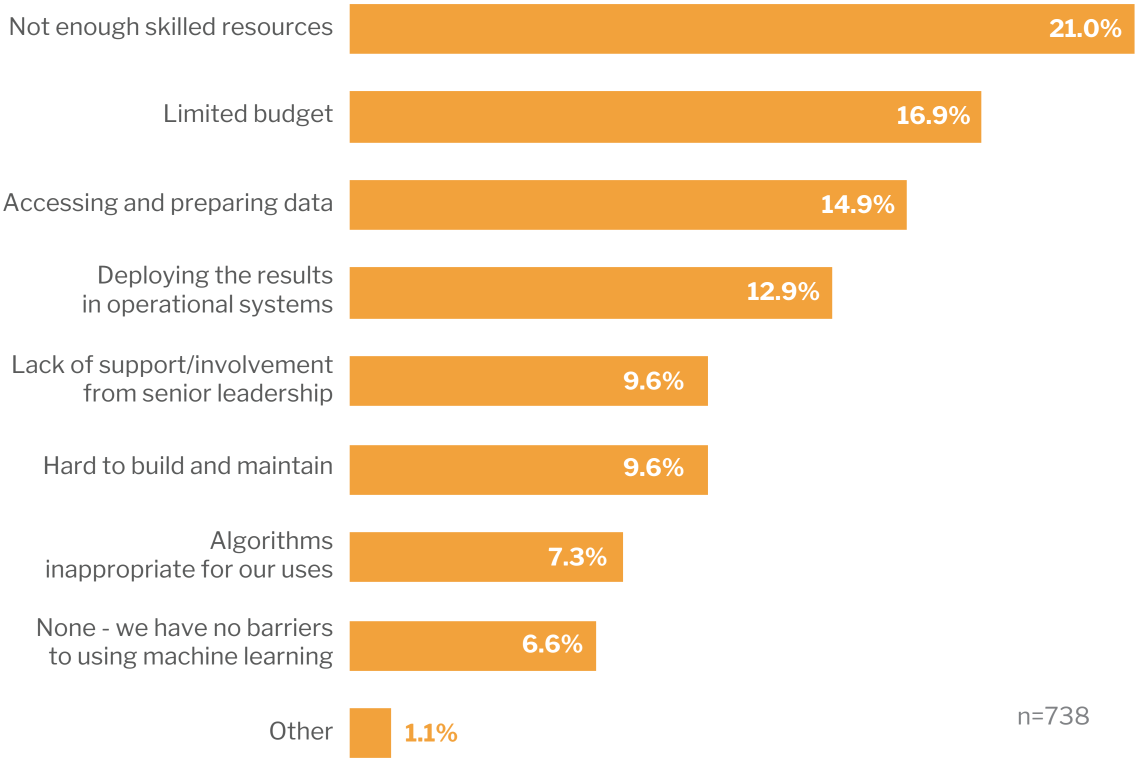
Barriers to Machine Learning

It is one thing to identify the potential opportunities to extract value from information using machine learning. It's another to actually deliver on that potential and create viable ML-driven business initiatives. Despite the success noted above, there are multiple barriers to machine learning adoption.

The most significant of these, according to 451 Research survey results, is the lack of skilled resources. Most obviously, this relates to the ability of an organization to find and hire developers and data scientists with the requisite skills to develop, test and maintain machine learning models.

Figure 5: Barriers to machine learning

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Adoption and Use Cases 2H 2018



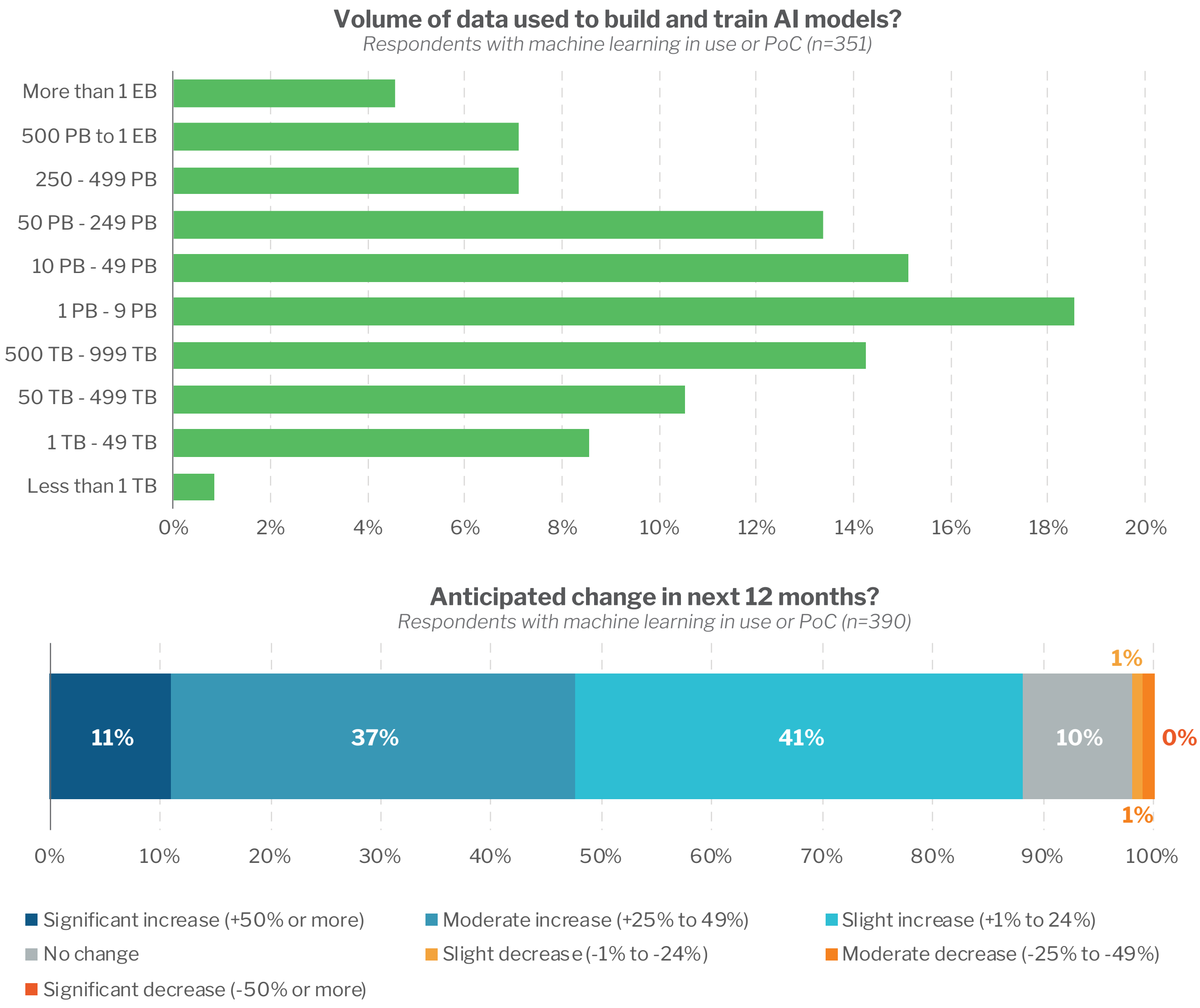
The skills challenge also relates to the domain expertise required to train machine learning models and interpret the results, however. 451 Research estimates that there are currently as many as two million data scientists in the world. While this number is increasing, there are arguably greater potential benefits to be had in getting the results of machine learning into the hands of the estimated 200-250 million knowledge workers in the world. This explains why we see the highest proportion of enterprises saying that their primary strategy for machine learning involves purchasing software applications with built-in machine learning capabilities.

The skills challenge involves the ability to build and maintain the associated infrastructure, which increases the attractiveness of cloud-based software delivery models. In addition to limited budgets, accessing and preparing data scored highly as a barrier to machine learning. The volume of data used in AI and machine learning projects is already significant: 451 Research survey results indicate that 80% of respondents are using more than 500TB of data to build and train their AI models, while two-thirds are using more than 1PB of data.

The results also show that 89% of respondents expect that the volume of data used will have increased 12 months from now, with 11% expecting a significant increase (of 50% or more), 37% of respondents expecting a moderate increase (25-49%), and 42% anticipating a slight increase (1-24%). As the volume of data being used to build and train AI models increases, this will likely exacerbate data access and preparation challenges.

Figure 6: Volume of data used in AI/ML workloads today and anticipated change 12 months from now

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019



Volume of data used to build and train AI models?

Respondents with machine learning in use or PoC (n=351)

Anticipated change in next 12 months?

Respondents with machine learning in use or PoC (n=390)

Data for AI and ML: Structured or Unstructured?

Not only is the volume of data growing, but the range of data types that are being used for machine learning is growing as well. Initial AI and machine learning projects have largely focused on generating insight from structured data sources – particularly data that has been generated by enterprise applications and already resides in enterprise databases.

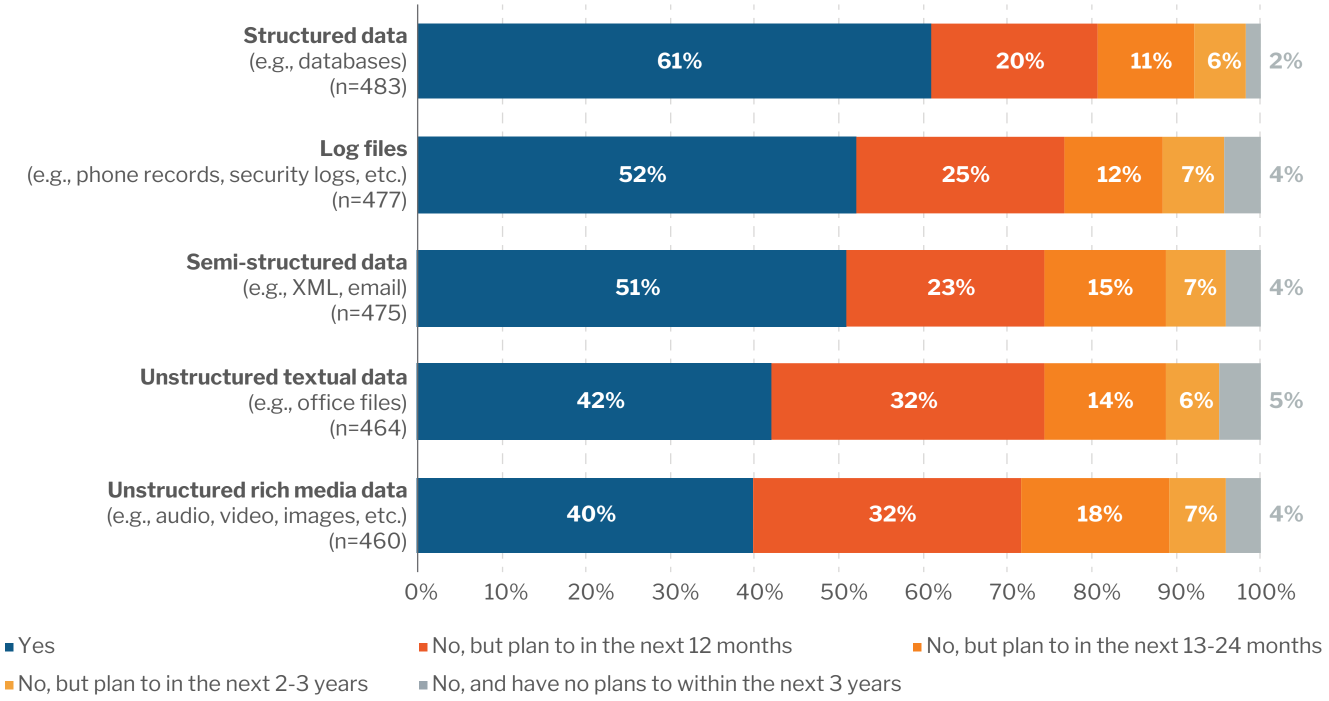
451 Research survey data indicates that 61% of respondents are using structured data to feed their AI/ML projects today, while 51% are using semi-structured data (such as email and XML documents). In comparison, 42% are currently using unstructured textual data (such as office files) and 40% are using unstructured rich media data (audio, video, images).

However, the survey results indicate that a significant proportion of respondents are looking to bring the benefits of machine learning to semi-structured and unstructured data. Nearly one-third (32%) of respondents are planning to use both unstructured textual data and unstructured rich media data for AI/ML projects in the next 12 months, while 23% are planning to use semi-structured data.

Looking further ahead, 50% of respondents plan to use unstructured rich media data in the next 24 months. During the same time frame, 46% plan to use unstructured textual data, and 38% plan to use semi-structured data. If respondents follow through on their plans, the use of semi-structured and unstructured data will be almost as widespread as the use of structured data from machine learning projects within the next two years.

Figure 7: Data types used for AI/ML projects

Source: 451 Research, Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019



The potential to unlock business insights from unstructured and semi-structured data with AI and machine learning is significant given the amount of business information that exists in unstructured form, including business contracts, official letters, designs and illustrations, and photos and videos.

This information could be in its original physical format or stored on digital formats, including video tapes and tape storage, as well as microfilm or microfiche. Many enterprises have sought to reduce the amount of paper floating around their offices by digitizing documents that need to be kept for regulatory reasons, for example.

However, archiving that information can make it more difficult to find and extract business value from it. Machine learning technology – specifically document-understanding models – can be used to scan these physical and digital documents and images to identify valuable business information and make it available for analytics and regulatory purposes.

Additionally, the combination of entity extraction and metadata enrichment can enable the creation of search filters to provide easy access to this information. This combination of document understanding and metadata-driven search filters can be used by enterprises in several ways: for example, to gain a better understanding of their business decisions, improve engagement with customers, and ensure compliance with regulatory requirements (see use cases below).

AI Use Cases for Unstructured Data

AI, broadly speaking, and specifically machine learning as a vehicle for execution, are potentially transformative technologies, with a caveat: they need viable use cases to demonstrate business value. Without demonstrating business value and ROI in the real world, machine learning and associated technologies such as neural networks are likely to remain more akin to science experiments rather than meaningful drivers of societal change.

Fortunately, unstructured data is still a largely unexplored frontier in this regard, and it is easy to identify several valuable use cases based on the application of machine learning to textual data alone, which could potentially slash the amount of time and tedium involved in duplicative efforts and repetitive work. Add the capability to mine and leverage other sources such as video, audio and other multimedia data, and the possibilities become even greater.

The following represents a list – by no means exhaustive – of potential use cases:

CONTRACT ANALYSIS

Contracts are often constructed of clauses that have been repurposed and adapted from document to document, and the manual identification of changes can take significant time and effort. The ability to quickly and accurately identify and extract key elements and terms of a contract is critical to negotiation and favorable outcomes for the business, and doing so requires the ability to leverage advanced natural language processing, such as the interpretation of grammar and key terminology.

Contract analysis, and management of contract workflows, is already an area of legal technology. But being able to do so at scale consistently, highlighting and coding relevant terms, is what matters and what automated textual analysis is best suited to do. The intent isn't to replace legal professionals but, rather, to accelerate their work processes and allow them to focus on higher-level tasks such as negotiation instead of the rudimentary combing through and interpretation of documents.

These capabilities have applications for all types of contracts, ranging from customer contracts to those with third parties and partners. As long as key elements can be identified and extracted at scale, the organization with the more sophisticated automation will have an advantage in terms of securing more favorable agreements, assessing risk exposure, and identifying patterns such as non-compliance or characteristics of successful sales approaches that could be duplicated.

HUMAN RESOURCES OPTIMIZATION

Talent is often an organization's greatest asset, and failure to allocate that talent effectively is a wasted investment, leading the organization to underperform relative to its potential. Human workers generate enormous amounts of semi-structured and unstructured content in the course of daily work, and this data can provide insight into skills, expertise and existing efforts and initiatives within the organization. Machine learning can be used to identify the associated patterns, with several potentially impactful use cases.

The ability to consistently identify real-world skills based on work output rather than skills that are self-declared or identified with static HR assessments would allow organizations to quickly build 'dream teams' for projects or initiatives by grouping individuals with highly complementary talents. This capability would also lend itself to helping prevent organizational 'brain drain' as individuals retire or move on; as soon as someone departs, machine learning could be used to quickly identify internal talent with a similar skill set, as well as help provide a condensed dossier of the projects and tasks the departing worker had been involved in. Likewise, the ability to analyze unstructured data for recurring patterns would help workers avoid reinventing the wheel and duplicating work products, such as templates and spreadsheets, which others had already created for similar tasks or initiatives.

Recruitment use cases also abound. The ability to analyze, via machine learning, the work products of existing high-performing employees would help HR teams build better profiles for ideal candidates when the time comes to hire. Having a better understanding of the real skills that lead to high performance in a given role would lead to more accurate job postings, and the ability to have more accurate, data-driven interviews and recruitment assessments.

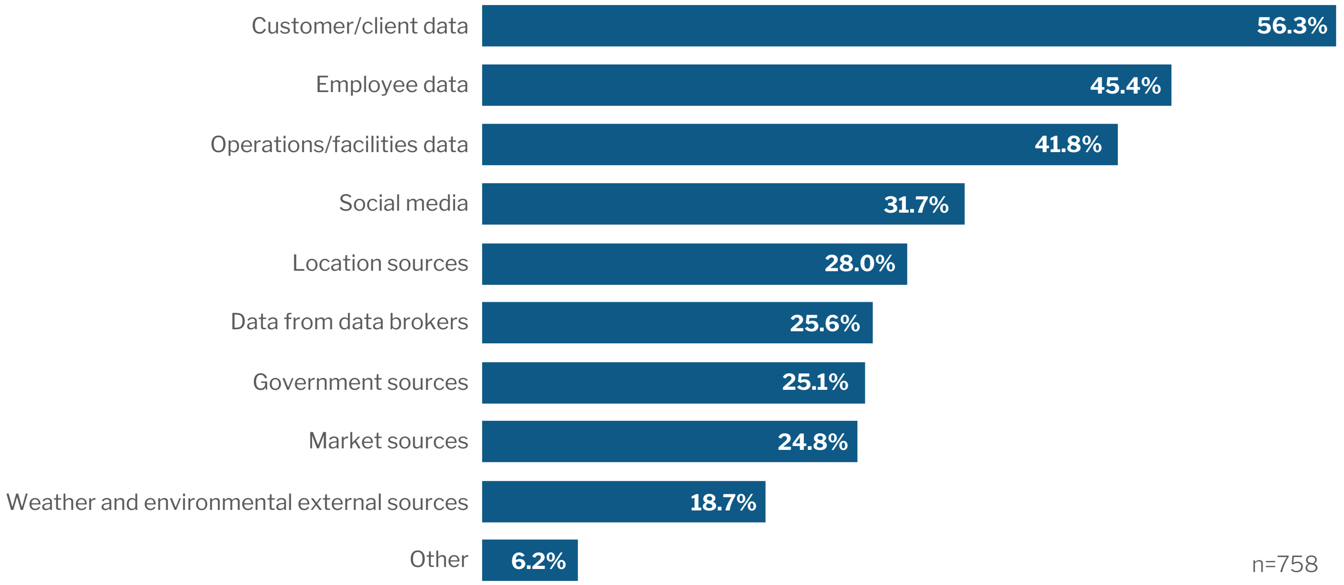
RISK REDUCTION, COMPLIANCE AND DATA PRIVACY

The majority of enterprise data is often unstructured data, yet it's often the least managed data type. This is a problem because many types of sensitive data – such as personal data protected by regulations such as GDPR – often are in an unstructured format. But it's not just a regulatory concern; unstructured data needs to be identified and governed for reactive use cases, as well as leveraged for more proactive ones. The first step to either protecting or leveraging data is simply knowing what data you have and where it is, which is a task machine learning excels at.

Additionally, the most sensitive data types are often also the most commonly used and valuable to the enterprise, warranting more rigorous governance. In 451 Research's Voice of the Enterprise: AI/ML survey, enterprise respondents reported that customer data and employee data were the two most common categories of data being used in machine learning initiatives and model training.

Figure 8: Data sources that are used (or will be used) as inputs to machine learning

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Adoption and Use Cases 2H 2018



Customer and employee data is rife with examples of data, often unstructured, that could be considered 'personal' under today's evolving regulatory landscape. And this definition of 'personal data' is becoming more expansive and ambiguous given the growing variety and scope of legal frameworks.

Personal data no longer always fits into neat hard-coded and easily identified templates such as Social Security and phone numbers. What is needed is technology that can associate subtle patterns, particularly in unstructured data sources, that are affiliated with these data types. Identifying sensitive and personal data, however, is just the first step. To protect long-term, the enterprise needs the capability to automatically classify and apply policy to this data at scale because human efforts would never be sufficient to monitor and evaluate modern enterprise volumes of data. In this sense, automation is also needed for classification and policy application.

Privacy use cases are not all reactive. Once personal/sensitive data has been identified, classified and had appropriate policy applied to it, it can also be used more proactively. Having a strong data governance and privacy program that leverages automation provides the opportunity to build trust with customers via awareness campaigns and PR. It also, perhaps paradoxically, can liberate additional data for analysis. With granular capabilities to mask, redact or anonymize sensitive data automatically, these datasets can often be leveraged defensibly in analytics and data science efforts in situations where they may have formerly been entirely off-limits or locked down for fear of misuse. More control of data always has positive downstream benefits, and automation plays a key role in controlling sensitive data at scale.

Conclusion

Artificial intelligence in general has huge potential to improve business decision-making, with machine learning techniques providing multiple opportunities for enterprises to unlock the value inherent in their data and associated documents. The combination of low-cost computing power and large volumes of data is enabling machine learning to become an enabler of improved customer experience, competitive advantage, and accelerated sales, rather than the academic pursuit it has been previously.

As volumes of enterprise data continue to grow, particularly in relation to machine learning efforts, barriers to success such as the lack of skilled resources, limited budgets, and accessing and preparing data are likely to be exacerbated, increasing the attractiveness of managed cloud services that mask the complexity.

While the majority of machine learning efforts to date have involved structured data, there is a clear growing appetite to bring the benefits of machine learning to unstructured and semi-structured data, including email and XML documents, as well as paper and digitized office files, and also audio, video and images.

The potential to unlock business insights from unstructured and semi-structured data with AI and machine learning is significant given the amount of business information that exists in unstructured form. Machine learning models, such as document understanding, text extraction and search filters can be used to identify and provide easy access to unstructured and semi-structured data, whether it is in its original physical format or stored on digital formats, as well as video, audio and other multimedia data.

This can be for both proactive analytics and reactive regulatory purposes – but the two are not mutually exclusive – machine learning can be used not only to automatically locate personally identifiable data, for example, but also to proactively mask, redact or anonymize sensitive data to enable it to be used for analytical purposes.

Recommendations

- Act quickly. The positive opinions about the performance of machine learning projects conducted to date suggest that those organizations that don't embrace machine risk being left behind.
- While experimentation is an important first step, focus on developing business use cases with demonstrable business value and the potential for return on investment.
- Machine learning and automation will be fundamental to identifying, sorting, classifying and governing information at scale, as well as the interpretation of that data. In addition to automating repetitive tasks, focus on enabling outcomes and insight that would not be otherwise achievable.
- Be cognizant of the breadth and depth of potential use cases. Narrowly focused products and services might satisfy individual use cases, but investment in broader platforms can enable multiple use cases and the opportunity for enterprises to apply their own models and domain expertise.
- Unstructured and semi-structured data is a relatively untapped resource, but that won't stay the case for long. Enterprises should act now to identify relevant data sources as well as desired business outcomes.

With the exponential growth of unstructured data across industries, most organizations recognize the potential of their data but struggle to uncover its full value because they have too much unstructured and unclassified information, lack the internal resources and skills to analyze it, or both. Much of this content is archived in storage boxes, backup tapes and file systems, which makes it difficult to extract critical business information quickly and easily and leverage that information to drive value, achieve greater efficiencies and reduce costs/increase revenue. Even if the content is accessible, it takes considerable time, resources and cost to search, retrieve and understand what information is available. Simply storing information is not enough to remain competitive in today's digital world. You need to harness the power of your data content regardless of its format or storage location.

Iron Mountain addresses these data challenges with Iron Mountain InSight™, a cloud-native, subscription-based intelligent content platform that combines Google Cloud's AI and machine learning with Iron Mountain's document management and data governance expertise to classify, enrich, and extract metadata from documents, images, and videos, making them available through a powerful visual search UI that allows users to take full advantage of the rich metadata. Additionally, Iron Mountain InSight incorporates a market leading policy and privacy engine to help companies meet existing and emerging retention and privacy regulations, such as GDPR and California Consumer Privacy Act.

With Iron Mountain InSight, data becomes more valuable using:

- Quick analysis and classification of information – Employing a robust process for sequencing machine learning across specific sets of data and information that executes high fidelity optical character recognition, visual similarity searches, facial and speech recognition, natural language processing, and other functions;
- Flexibility & versatility – Iron Mountain InSight delivers this capability across a variety of content types (documents, maps, spreadsheets, videos, etc.) and physical and digital sources (hardcopy, backup tape, SharePoint, network file share, etc.), applying machine learning and AI-based technology to uncover trends, insights and indicators;
- Security, ease of implementation & scale – Delivered as a subscription-based, cloud-native platform – Iron Mountain InSight enables quick implementation and delivers the scale to fit an organization's growing need for analysis, classification and data value discovery.

Iron Mountain InSight automates the process of data capture by connecting to all applicable data sources, including off-site physical records and backup tapes, data stored in the cloud, and digital files hosted onsite or archived at a remote location. This process creates structure for unstructured content and reveals relationships across disparate formats and sources, while applying consistent compliance policies across all of the content.

Iron Mountain InSight enables companies to automate data understanding, enable compliance and business value, and deliver peace of mind. For more Information on Iron Mountain InSight, go to: www.ironmountain.com/insight.

Content Provided by

IRON MOUNTAIN®

About 451 Research

451 Research is a leading information technology research and advisory company focusing on technology innovation and market disruption. More than 100 analysts and consultants provide essential insight to more than 1,000 client organizations globally through a combination of syndicated research and data, advisory and go-to-market services, and live events. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

The information contained herein has been obtained from sources believed to be reliable. 451 Research disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although 451 Research may discuss legal issues related to the information technology business, 451 Research does not provide legal advice or services and their research should not be construed or used as such.

451 Research shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.

NEW YORK
1411 Broadway
New York, NY 10018
+1 212 505 3030

SAN FRANCISCO
505 Montgomery,
Suite 1052
San Francisco, CA 94111
+1 212 505 3030

LONDON
Paxton House
30, Artillery Lane
London, E1 7LS, UK
+44 (0) 203 929 5700

BOSTON
75-101 Federal Street
Boston, MA 02110
+1 617 598 7200

451 Research®