

Turning Data Complexity Into a Competitive Advantage

An introduction to **unsupervised AI for enterprise analytics**



Executive Summary

The introduction of AI in the Analytics and BI industries came with proclamations that organizations would now be able to unlock an unprecedented volume of insights from the vast troves of data they generated. These insights would flow through every aspect of the organization (i.e. sales and marketing, IT, supply chain, HR, finance, etc.), resulting in optimizations and efficiencies that positively impact both topline and bottomline revenue. Company cultures would be data-driven rather than steered by hunch and happenstance.

Despite heavy investment in Big Data initiatives generally, and AI technology more specifically, most organizations still report that these investments remain limited in scope and that becoming a data-driven culture continues to be an ongoing issue.¹

The majority of the investments in AI technology, especially for analytics, are in traditional BI and supervised learning, which is designed to help scale analysis of prepared, fully-structured data to provide predictions on new data.² While this method is valuable for some applications, its limitations result in frustration for those relying solely on it for analytics and business intelligence efforts. These limitations primarily being 1) the inability to ingest and analyze complex data, 2) the onerous data preparation work required from a centralized data team to perform the analysis, and 3) the requirement of pre-forming hypotheses to test prior to conducting an analysis of the data.



But the recent introduction of scaled unsupervised learning is changing how businesses approach analysis, no longer requiring pre-formed hypotheses in order to unlock insights from data. Additionally, the emergence of platforms that combine unsupervised learning with other AI techniques are able to ingest far greater volumes and more complex forms of data while substantially decreasing the amount of data preparation work required to perform analysis.

Due to these advancements enterprises are able to access a continual stream of actionable insights, so business counterparts are able to find and respond to new opportunities and risks more quickly. They're also able to track and measure a demonstrable return on investment.

Investments in AI Ramp Up, But Hit a Wall of Complexity

AI is not only a technology organizations are using to optimize their business. In many cases, AI forms the core of that business. It is how massive enterprises, such as Amazon, Google and Netflix, target their products and services to their end customers. A recent report found that nearly 65% of the 70 blue chip firms surveyed reported spending upwards of \$50 million on AI.³ Overall, total AI expenditure is predicted to reach nearly \$100 billion by 2023.⁴

Companies make these investments based on the premise that a strategic competitive advantage can be found in the data the business amasses — if only it can be extracted, translated and acted upon. AI, of course, can scale computationally in ways human cognition cannot, opening up opportunities to uncover business insights that lead to profitability, fewer risks, and greater revenue.

*"Fewer than **15%** of large scale enterprises report having their AI investments in wide-spread deployment, and less than **38%** said they have created a data-driven culture."*



Despite this heavy investment, many organizations have struggled to extract value from their data, or realize a quantifiable return on their AI investments. Fewer than 15% of large scale enterprises report having their AI investments in widespread deployment, and less than 38% said they have created a data-driven culture.⁵

This problem is particularly acute in the fields of Analytics and BI with AI. Though 83% of companies stress the importance of turning data into actionable insight, only 22% feel their company is successful at doing so.⁶ While more organizations are purchasing AI-based analytics solutions, almost all of those solutions are built on supervised learning techniques. Supervised learning is immensely helpful for building a model that will use learnings from what is known about well-structured data to make predictions on new data. This works extraordinarily well for something like identifying whether an image contains a dog vs a cat based on historical examples of such photos. The challenge, of course, is that enterprise data is rarely well structured and often the business problem is not well understood.⁷ We know what a dog is, for instance. Understanding what is driving a sales decline is usually far less clear.

Supervised AI and Unstructured Data

To begin an analysis using supervised AI, the business must first combine all the data into a well-structured format. Put simply, the machine learning needs the data to be properly labeled in order to perform the analysis.

This puts a tremendous burden and cost on the data team to conduct data preparation work upfront to test even a single pre-formed hypothesis. As much as 80% of enterprise data is unstructured.⁸ Creating a model that incorporates any of this unstructured data results in three different forms of inefficiencies: 1) the data team is saddled with an overwhelming amount of data preparation work; 2) the cognitive limitations when dealing with so much complex data inevitably leads to the majority of unstructured data being excluded from the analysis; and 3) the data team structures the data to test the hypothesis, and any change to the hypothesis requires re-structuring the data.

"96% of enterprises reported running into data quality and labeling issues, and 78% said their AI/ML projects have stalled at some point."

The Rise of the Data Preparation Slog

The inability for supervised learning techniques to analyze unstructured data means that substantial data preparation.⁹ This creates both resource constraints, added costs and slower time to insight.

Estimates range, but McKinsey finds that fully half of data science work is spent on this kind of data preparation. This creates not only constraints in establishing a pipeline of valuable insights into the business, but also increases cost that limit the return on investment. The average wage of data scientists has grown by 16% in just two years, which far outstrips the average wage growth across all occupations, and the gap between talent and demand is expected to reach 250,000.¹⁰ Nationally, the majority (50%) of businesses with data science investments have 10 or fewer data scientists on staff, meaning that a small group is responsible for huge quantities of data preparation for analysis across multiple functions.¹¹ This is largely driven by the need to revisit data structure as hypotheses are changed or new hypotheses are introduced.

The proliferation of traditional supervised learning within the enterprise has also meant the proliferation of data preparation work. In fact, 96% of enterprises reported running into data quality and labeling issues, and 78% said their AI/ML projects have stalled at some point. This means many projects are bottlenecked, blocked or even abandoned.

But this resource constraint doesn't just lead to frustrated efforts and reduced ROI. It also impacts the analysis that does get completed.

The Glut of Unanalyzed Data in the Enterprise

The value of machine learning is to scale the cognitive load related to analysis of large sets of data. But the data preparation required by supervised learning means data teams are forced to select which data will be included in the analysis. Inevitably, the vast amount of unstructured data in need of labeling is culled into what is manageable for a human team.

This leads to large quantities of unanalyzed data. Research firm Gartner estimates that up to 50% of unstructured data has indeterminate value, resulting in wasted storage costing millions of dollars per year per enterprise.¹² But this is just an estimation of the cost of storing unstructured data. There's tremendous opportunity costs to consider.

Since supervised AI requires upfront data preparation, naturally much of the available unstructured data doesn't get included in the actual analysis, resulting in both missed insights and lowered confidence. Feature generation with supervised learning, for instance, requires that the analyst first label any unstructured data prior to analysis. This limits the number of new features derived by the time a human can spend on data preparation in order to feed the machine.

In these scenarios, the features — and the insights they inform — are not autonomously generated since they are dependent on the manual work of an individual's or team's time and cognitive capacity.

While the structure of data represents significant constraints on analysis when using supervised ML, it feeds into a larger issue for how data analysis is run within the enterprise: the methodology.



How the Traditional Analytical Process Blocks Data-Driven Cultures

Supervised machine learning, when employed in the proper context and investigation, can provide meaningful impact to a business. This is especially true when it comes to predictive analysis on known questions using structured data.

As we've demonstrated, these two pre-conditions are difficult and time-consuming to meet.

The reliance on supervised AI means that the traditional analysis process applies an hypothesis-first approach. In other words, the organization involved must start analysis with pre-formed hypotheses on what types of questions they'll need data to answer, and also what types of insights they'd like to derive from that data.

The hypothesis-first analytical process spans multiple teams and typically includes the following workflow:

- 1 The team hypothesizes on the types of needs that will be supported through analysis of the data.
- 2 The data science team then joins disparate data sources. It's during this stage that data labeling occurs, and much of the unstructured data will be culled from analysis.
- 3 The data science team manually generates new features from combined data.
- 4 Data analysts form their own hypotheses on what insights they need from the prepared data.
- 5 The data analysts test the hypothesis using supervised machine learning.
- 6 If not enough insight is gleaned, the data analysts repeat the hypothesis process. If the insights are deemed sufficient, they're passed to a business analyst.
- 7 The business analyst must prioritize the insights based on their estimated impact and translate for the business counterpart, who in turn will choose whether to act on them.

Traditional Analytical Process

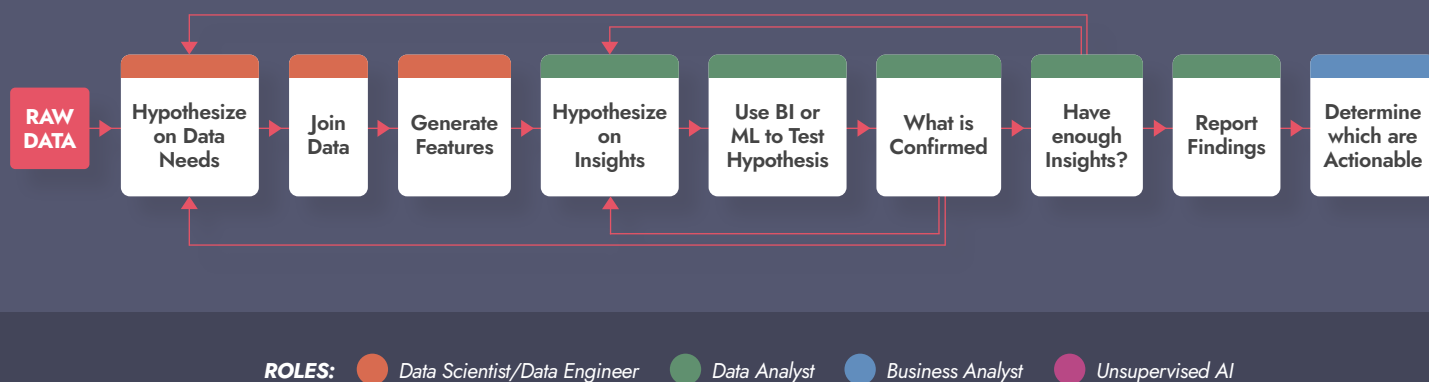


Figure 1a: The traditional analytical process within enterprise companies.

“The amount of digital data created by 2020 was expected to reach 44 zettabytes, the equivalent of 40X more bytes of data than there are estimated stars in the observable universe.”

Beyond the considerable time required to fully complete this process, it also introduces bias and is difficult to scale, particularly without ever expanding investments in data science and engineering teams.

Bias is introduced at the outset of the process due to an individual or team beginning with a hypothesis that isn't fully formed by an analysis of the data. Instead, the data scientist or engineer needs to begin with assumptions about what the analysis could say or should say. Much of the data goes unanalyzed as a result, limiting the output of the analysis.

The other issue is scalability. As discussed previously, an enterprise's data is already large and complex. The volume of data increases exponentially. (The amount of digital data created by 2020 was expected to reach 44 zettabytes, the equivalent of 40X more bytes of data than there are estimated stars in the observable universe.¹³)

And the complexity of that data continues to grow as new technologies emerge and extend across business and consumer applications (ex. wearables, voice-recognition, etc.). Most of this is amassed in an unstructured form, and therefore either unanalyzed or only partially analyzed using supervised learning and a hypothesis-first methodology.

As an example, if an organization wanted to look at very simple patterns in the data, it can look at a single column on their own. In the table titled “Single Variable Analysis,” we provide a sense of how many possibilities the organization would need to review. This is typically how a BI application would be used: a chart of multiple values for one column would allow the organization to get through the 11 column example in only 11 screens.

Single-Variable Analysis

Columns w/5 Values	Possible Hypothesis	Comparison
11	55	Working hours in 1.5 weeks
30	150	Working hours in 3.7 weeks
115	575	Working hours in 14 weeks

Figure 2a: Single variable analysis.

Multi-Variable Analysis

Columns w/5 Values	Possible Hypothesis	Comparison
11	48,000,000	Human lifespan in minutes
30	10 ²¹	Grains of sand on Earth
115	10 ⁸⁰	Atoms in the Universe

Figure 2b: Multi variable analysis.

However, by 115 columns, even single column analysis starts getting substantial. Yet the story is far worse with multi-column patterns — patterns that are about several columns having particular values (ex. Customers over 45 in Oregon). The complexity grows far beyond what people can ever handle quickly.

A better method would be to form hypotheses once the raw data is analyzed and the patterns are surfaced. This would 1) increase confidence in the findings of analysis because it encompasses all the available data, 2) automate hypothesis testing and feature generation, 3) eliminate the need for upfront data preparation work, and 4) provide a larger and more frequent number of insights into the business.

With traditional techniques such an approach wasn’t possible. However, the recent introduction of scaled unsupervised learning is changing the way enterprises analyze and take action on their insights, allowing them to embrace complex data fully and turn what had been an obstacle into a competitive advantage.

Unsupervised Transforms Complexity from Cost to Competitive Edge

What is Unsupervised Learning?

Unsupervised learning is not new. It's been used as long as supervised machine learning. One of the reasons that unsupervised learning is less prominent within the enterprise, however, is that supervised learning algorithms are simpler — but that is also why they don't scale when dealing with complex enterprise data.

With supervised learning, the AI must be trained with historical data in order to analyze new, fully structured data. After multiple iterations, its main output are models which are then used to provide predictions on future outcomes so long as the data is consistent with prior structured and historical data.

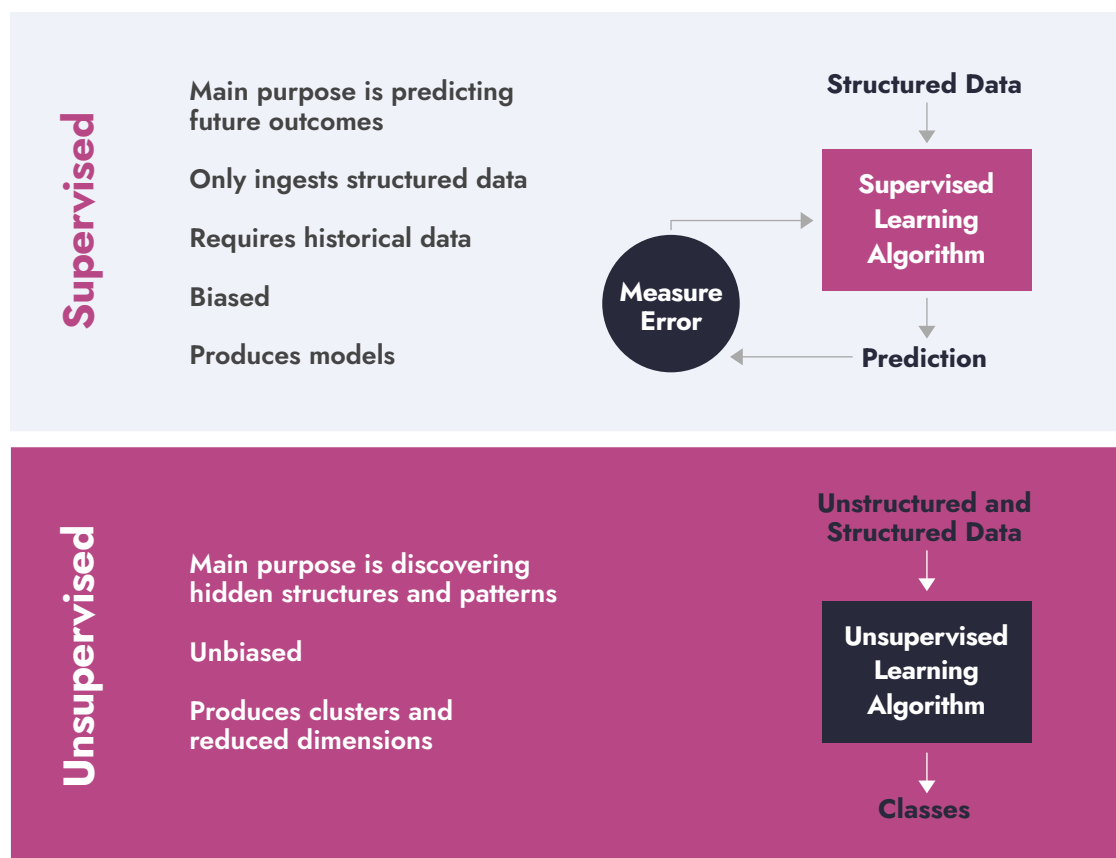


Figure 3: Supervised learning to unsupervised learning comparison.



By contrast, unsupervised learning doesn't require labeled datasets for training. Instead, the algorithms ingest raw, unclassified data and provide hidden structures and useful patterns from that data. This inferred structure allows the AI to identify groups or clusters within the data that exhibit similar behavior, all of which is accomplished without human supervision. This provides several key advantages when dealing with complex and large data:

- ▶ ***Data labeling is not required upfront***
- ▶ ***The algorithm reduces bias associated with the hypothesis-first methodology***
- ▶ ***Far vaster capabilities for scaling across larger and more complex datasets***
- ▶ ***Greater pattern recognition from the data***
- ▶ ***More confidence in the analysis since it includes more data than supervised learning***

Simply put, unsupervised learning provides a tested and highly valuable approach to analyzing complex data, identifying hidden structures and patterns, and surfacing new insights previously inaccessible.



“The next revolution of AI will not be supervised.”

Applications of Unsupervised Learning in the Enterprise

Scaled unsupervised learning is gaining prominence across several applications as forward-thinking enterprises seek to drive more value and scalability from their increasingly complex data.

In particular, leading thinkers within the AI industry see unsupervised learning as the path forward for extracting value and insight from data, and automating much of the manual work currently involved with supervised learning methodologies. As Yann LeCun, chief AI scientist at Facebook and NYU professor, stated: “The next revolution of AI will not be supervised.”¹⁴

Recent applications of unsupervised learning within a business context include using it to help shorten learning cycles and improve performance among self-driving vehicles.¹⁵ HSBC has used unsupervised learning techniques to identify patterns that suggested fraudulent activities, substantially cutting down on the number of investigations.¹⁶ UnitedHealthcare used unsupervised learning to identify new types of fraud that would be difficult to identify without historical data using supervised learning.¹⁷

The foremost field experiencing a transformation through the increased application of unsupervised AI is BI and Analytics.

Transforming Analytics with Unsupervised

The market most primed for transformation by the proliferation of unsupervised AI is the BI and Analytics space. The sector, already undergoing fast-change in recent years as more vendors sought to incorporate AI, could have the most immediate impact for the enterprise companies investing in it.

Specifically, the use of unsupervised learning can empower analytics to significantly scale across disparate data sources, greater volumes and complexity of data, surface far more insights while reducing the human effort and cycle times to access them.

Today, there is only one platform in the analytics industry that is built entirely on unsupervised AI. That platform is Unsupervised, which has established itself not only as the sole provider in the analytics space, but also as the largest provider of unsupervised software globally.

Unsupervised's AI allows it to analyze unprecedented amounts of complex data for enterprises, identify patterns that would be hidden or undiscoverable with traditional machine learning, and then surface insights directly to business and data teams in an intuitive UI.

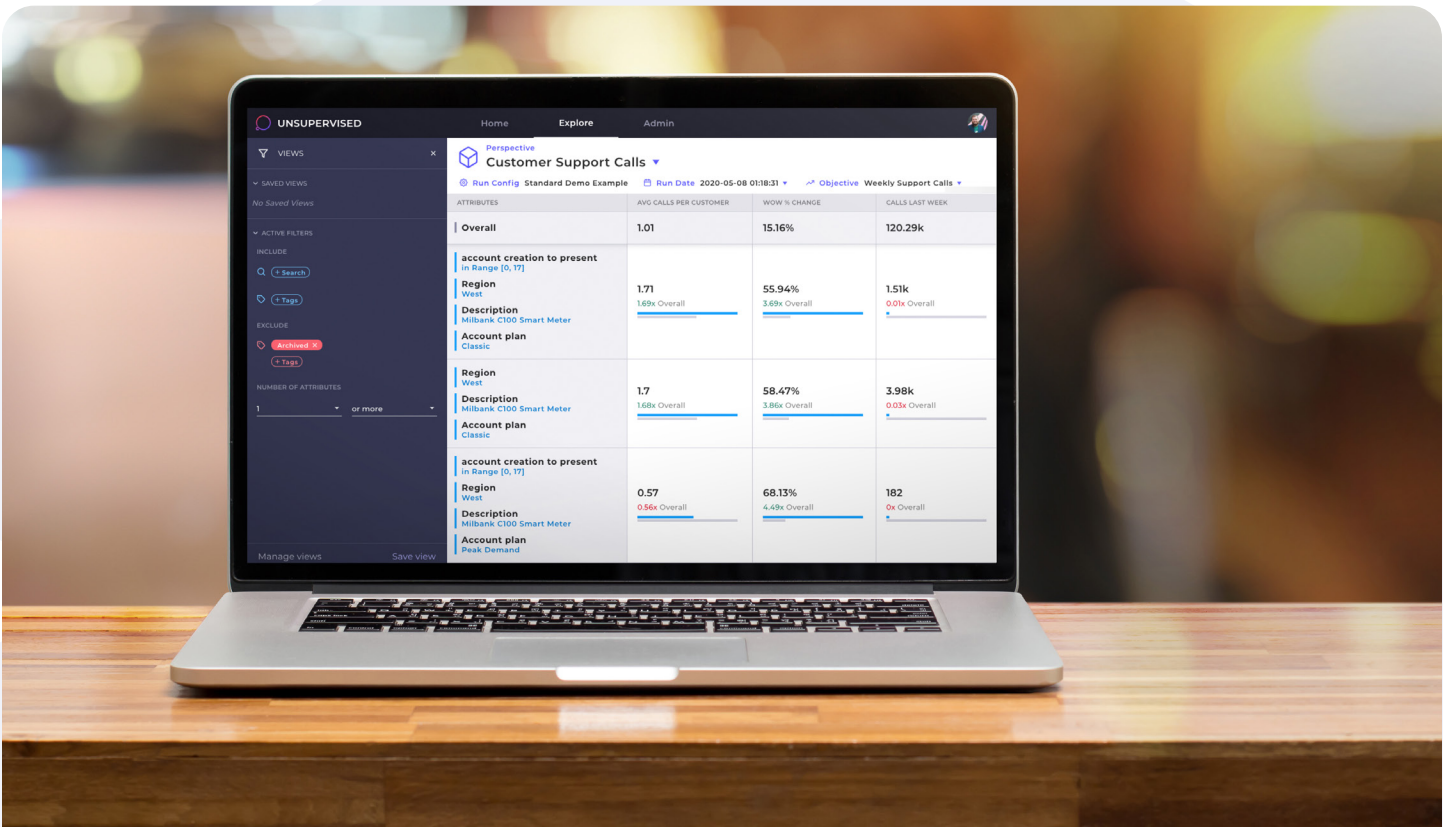
While the backbone of the platform is scaled unsupervised learning, the process for how the platform ingests the raw data, identifies relevant patterns, ranks them based on impact to core business goals, and then translates them into clear, human language requires a multi-step process involving multiple forms of AI.

First, the platform ingests data from multiple sources. This can include data warehouses, data lakes, log files, or a multitude of other sources. Importantly, the data in these can be structured, unstructured, or a combination of both. Next, the semantic data layer of the platform is configured to understand the meaning and rule for each piece of data informing how it can be handled. Given that knowledge, the data layer can create millions of features as needed on the data. These are not just on single tables, but the system can join and aggregate across many tables — following relationships to extract the full complexity hidden in your data. Furthermore, this layer can integrate external, third-party data to increase the comprehensiveness.

The AI's discovery engine then analyzes data from this data layer, looking at tremendous numbers of combinations of features and finding the patterns that exist in the data across the base data and the derived features. The most interesting of these are explored by the AI with information on those and other related patterns getting stored into a library of such patterns for humans to review.

When humans use the user interface of the system, they are presented with the most important insights from the software related to the business metrics they care about. This means that months of back and forth between end-consumers of insights and data experts are short-cut with those users immediately presented with the most important things in the data.

Importantly to the above, once the configuration of the system is done, it is able to handle new data as well. Thus every week (or however often the business chooses), the AI can keep delivering the freshest insights from an ever-changing business.



A New Analytical Process

As discussed previously, the traditional analytical process is a hypothesis-first approach that requires starting with a theory, long data preparation cycles, multiple tests and handoffs, and lots of unstructured data on the cutting room floor.

The introduction of scaled unsupervised learning, and the Unsupervised platform, empowers enterprises to take a different approach to analysis. In this framework, the Unsupervised platform ingests the raw data and autonomously joins the data and generates up to a million new features. It also discovers patterns and insights that most impact the pre-identified KPI, surfacing them in a single UI.

This process reduces the manual data preparation work, as well as the requirements of ranking and translating the most relevant insights — all of which is managed by the AI.

What it fundamentally shifts is the expectations of the human team. With Unsupervised, the AI is focused on what AI is best at: autonomously preparing and analyzing large, complex data. And in this process the human team is focused on what they do best: making decisions based on the provided data.

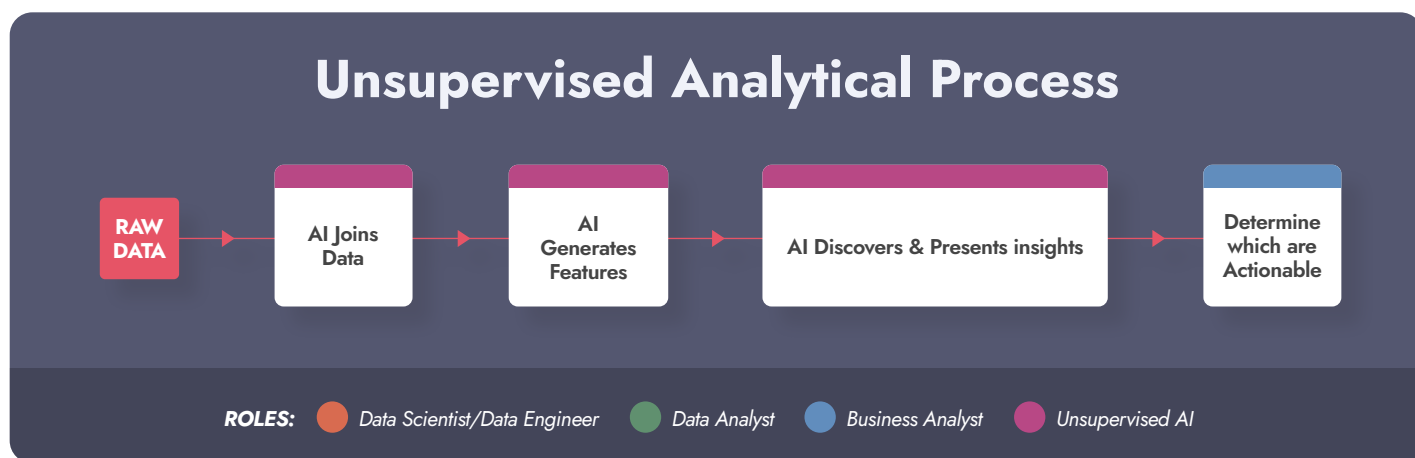


Figure 1b: Analytical process using Unsupervised platform.

The outcomes of using Unsupervised among complex data are multiple, as we've outlined in this report. **However, the three most prominent values are:**



Boundless analysis

Because the platform is able to ingest both structured and unstructured data, the enterprise is not forced into choosing which data to include and long, manual data preparation cycles.



Repeatable ROI

Insights are provided in a continual stream and, as data continues to flow through, the return on those insights are trackable and the entire process can be repeated without starting the entire analytical process over again.



















Complexity Transformed Into Competitive Advantage

By embracing the full complexity of an enterprise's data, new insights that would be difficult or impossible to discover are unearthed. This provides the enterprise opportunities and optionality that can only be found within their data, and not available to competitors.

Enterprises are unlocking new insights and value with unsupervised AI that wouldn't have been possible before. It represents the next phase of AI in the Analytics and BI space. As Yann LeCun said, unsupervised learning is "what's going to allow our deep-learning systems to go to the next level."¹⁸

With Unsupervised platform, that next level is accessible today. ■

- 1 *Big Data and AI Executive Survey 2020*; [New Vantage Partners](#) 
 - 2 *Supervised learning* entry, [Wikipedia](#) 
 - 3 *Big Data and AI Executive Survey 2020*; [New Vantage Partners](#) 
 - 4 *Worldwide Artificial Intelligence Spending Guide*, [IDC](#) 
 - 5 *Big Data and AI Executive Survey 2020*; [New Vantage Partners](#) 
 - 6 *Overcoming Barriers to Data Impact: New Tools and a New Data Mindset Can Bring About Real-Time Decision-Making*, [Business Review](#) 
 - 7 *The issue of having clear articulation and understanding of the business problem the analysis is intended to shed light on has only been of greater concern after events like the COVID-19 pandemic and social unrest. As an example, the models airlines use to adjust pricing — often by the minute — have been upended due to an unprecedented drop in demand and no historical data to adjust to. See: “Coronavirus Has Upended Everything Airlines Know About Pricing,”* [The Wall Street Journal](#) 
 - 8 *Organizations Will Need to Tackle Three Challenges to Curb Unstructured Data Glut and Neglect*, [Gartner](#) 
 - 9 *The Age of Analytics: Competing in a Data-Driven World*, [McKinsey&Company](#) 
 - 10 *Ibid*
 - 11 *2020 State of Enterprise Machine Learning*, [Algorithmia](#) 
 - 12 *Organizations Will Need to Tackle Three Challenges to Curb Unstructured Data Glut and Neglect*, [Gartner Research](#) 
 - 13 *“How much data is generated each day?”* [The World Economic Forum](#)
 - 14 *“The AI technique that could imbue machines with the ability to reason,”* [MIT Technology Review](#) 
 - 15 *“The Future of AI is Unsupervised,”* [Forbes](#) 
 - 16 *“HSBC partners with AI startup to combat money laundering,”* [Reuters](#) 
 - 17 *“Data-Crunching Technology Spurs Insurance Dollars,”* [The Wall Street Journal](#) 
 - 18 *“The Future of Deep Learning Is Unsupervised, AI Pioneers Say,”* [The Wall Street Journal](#) 
-



Unlock the power of scaled unsupervised learning.

Schedule a demo of Unsupervised today and
turn complexity into your competitive advantage.

unsupervised.com

