eBook

Automation buyer's guide:

What you *really* need to know about structured and unstructured data technologies

The Unstructured Data Company

INDICO DATA

Structured, unstructured and robots, AI!

U.S. President Herbert Hoover famously campaigned on the commitment to voters for "a chicken for every pot." Now, nearly a century later, enterprises are beginning to realize the promise of a process for every bot.

Gartner lists "Hyperautomation" as one of its Top Strategic Technology Trends for 2021, ushering in an era in which "anything that can be automated in an organization should be automated."[1] At the same time, IDC predicts that within the next two years, half of knowledge workers will regularly interact with their own AI-enhanced robot assistant, which will help identify and prioritize tasks, collect information, and automate repetitive work.[2]

"The global pandemic has accelerated AI adoption, and it is becoming ubiquitous across all business processes," says Ritu Jyoti, program vice president, Artificial Intelligence Research at IDC. "AI solutions powered by machine learning, conversational AI, and computer vision are at the forefront of business resiliency, accelerated innovation, and transformative customer and employee experiences."

Given that the landscape of documents and data that users are dealing with range from highly structured W-2 forms to ungainly, inherently unstructured financial reports and video content, it's no surprise that enterprises often struggle as they assess automation alternatives. And, as with most technology categories, there is a lot of noise out there: competing (and sometimes misleading) claims; complex and confusing jargon; a veritable alphabet soup of three-letter acronyms to decipher.

This buyer's guide aims to eliminate confusion and give buyers clarity to select the right tools for the right data and the right tasks. Read on for first-hand experience and recommendations from automation center of excellent (COE) leaders who share what you really need to know.

> "AI solutions powered by machine learning, conversational AI, and computer vision are at the forefront of business resiliency, accelerated innovation, and transformative customer and employee experience."
>
> – IDC

Define:

Structured vs. semi-structured vs. unstructured data

As you delve into automation, before long you'll learn about the three basic forms of data and why they matter when it comes to automation. These are: structured, unstructured and semi-structured data. In a nutshell, you can automate processes involving structured data with simple tools, but when it comes to unstructured and semi-structured, you'll need a more intelligently capable automation platform.

Structured data: RPA, BI, analytics

What it is: As its name implies, structured data is highly organized, typically in a database or spreadsheet with rows and columns. Essentially both the data and the data's schema are present together, and as a result, each piece of data can be mapped to a specific, fixed field or location. Structured data is often managed using the Structured Query Language (SQL), a common programming language for relational databases. With relational databases, it's possible to view data by various criteria, such as customers by region, and to answer queries such as, "customers who spent more than $500 with us last year."

How to use it: It's relatively easy to automate processes that involve structured data. Robotic process automation (RPA) tools or solutions that use connectors and screen scraping tools work well with structured data. You can build automation routines that tell the tools exactly where the data they need resides in any given document. So long as there's no deviation from that norm, the tools should work well to automate simple, repetitive tasks, such as extracting data from a spreadsheet and entering into a customer relationship management (CRM), enterprise resource planning (ERP) or other downstream system.

Unstructured data: requires "cognitive" capabilities

What it is: Unstructured data, on the other hand, adheres to no specified format. Types of unstructured data include the text in an email message, PDFs, Word files, photos, videos, presentations, call center or legal transcripts and more. Given that it follows no predetermined format, it's much more difficult to automate processes involving unstructured data. The data is presented without any schema present that can yield vital context for understanding the data.

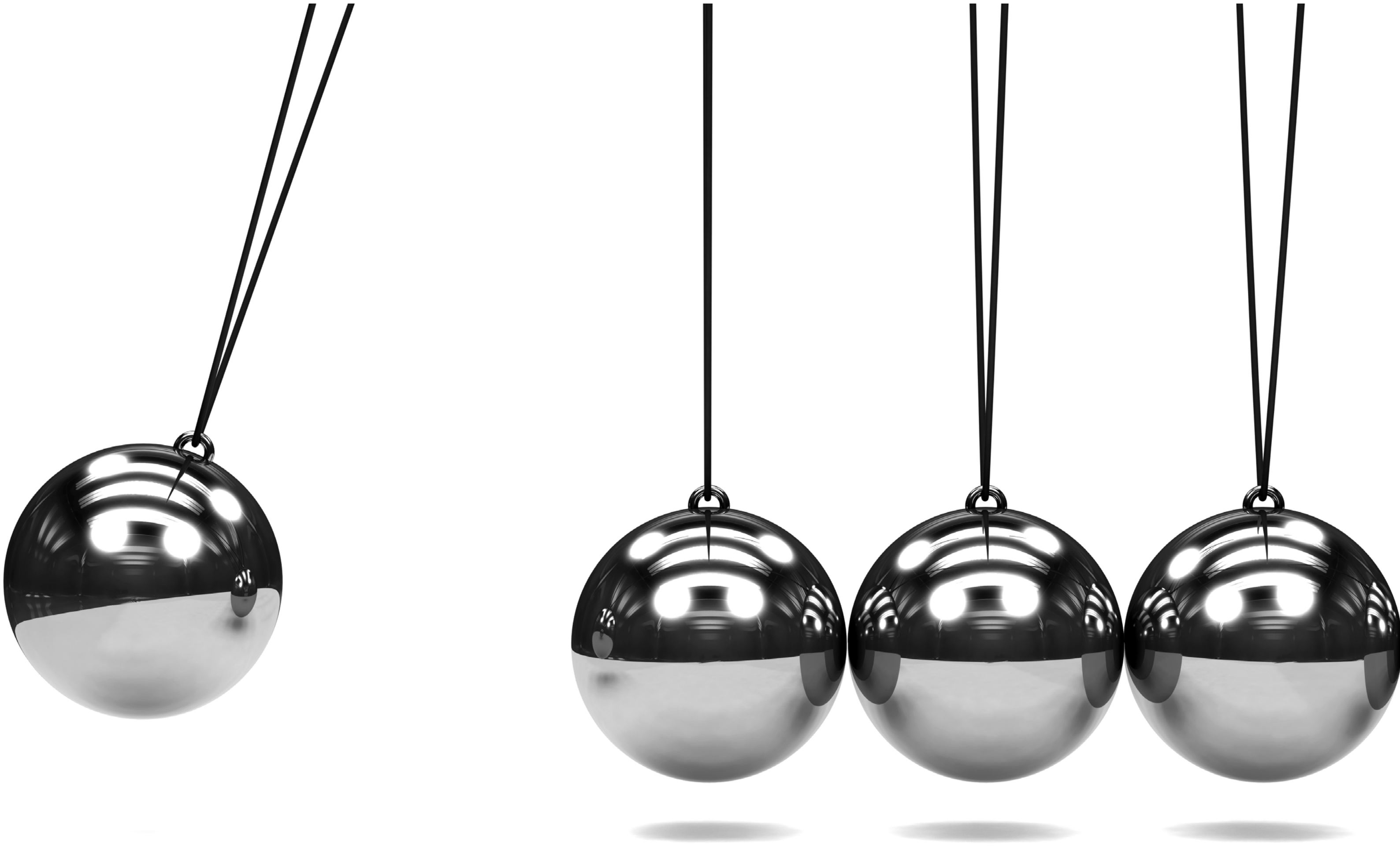
How to use it: But AI changes the game. With enough data, we can now train models to "read" unstructured data much like a human does, complete with an understanding of the context behind any given document or image. The models can extract key data elements required to automate a given process, such as financial figures, social security numbers, names, addresses and so on. Or, a model may be fed an image of a damaged car and be smart enough to know, "This car has been in an accident and has damage to the right front fender."

So what is "semi-structured" data?

What it is: Semi-structured data falls somewhere in between. Consider an email: while the text of the email is unstructured, the header contains structured elements – the "to" and "from" fields, date and time, for example. So, as a whole, an email may be considered an example of semi-structured content. Digital photos are another example. Typically, they will also contain a date, time and perhaps location where the photo was taken – all structured elements, although the image is wholly unstructured.

How to use it: In some instances, it's possible to use an RPA or templated tool to automate some of a process for handling these data types – such as categorizing an asset by date. But you'll still need a more intelligent automation tool to find and extract relative data. Invoices are often touted as an example of semi-structured content. That may be the case if your company gets invoices from only four or five suppliers, and they consistently use the same format. In that case, it's conceivable you could train an RPA or templated tool to extract key data elements to automate invoice processing.

But large companies likely receive invoices from dozens if not hundreds of companies that use many different formats. Invoices are often mis-characterized as semi-structured. In reality you'd be hard-pressed to create templates to handle each of them, and would forever be troubleshooting them as they change over time. Here again, it makes more sense to treat the invoices as unstructured content and use an intelligent data processing tool to automate invoice processing.



Automate:

Attack of the three letter acronyms

Business and technology love their three-letter acronyms (aka TLA). As you review automation solutions, you'll undoubtedly be exposed to them; you've already seen several in this guide. Here's a brief glossary of essential abbreviations and their implications for automation.

OCR

Among the first tools you're likely to encounter while researching solutions for automating processes is optical character recognition, or OCR. OCR is most often used to handle scanned documents. This means the documents are effectively images, even if they are saved in PDF format. Images, of course, are not readily machine-readable, so computers can't immediately process what humans can clearly see as text in the document. OCR addresses that issue by identifying text in such documents and converting it to a digitized format that computers can manage and then automate processes with.

RPA

One of the most talked about and pervasive automation technologies is RPA, robotic process automation. RPA is a technology that enables enterprises to use software robots that can interface with digital systems and software to perform tasks in a manner similar to humans.

RPA tools often work well when they know what's coming. If you have a series of documents that are all formatted exactly the same – like W-2s and other IRS forms, statements from the same bank, or a website "Contact Us" form – then a templated RPA approach should serve you well. Such an approach often involves using OCR technology to identify the text within an image. Then a template is used to indicate exactly where in the document the data you care about is located and the data can be extracted.

An RPA tool may then be employed to take the resulting data and put it into some other downstream system for processing, relieving a human from performing these same tedious steps over and over. So long as there is no variation anywhere in the process, whether in the documents or in the steps required to get the job done, it should work well.

RPA is an effective tool for high-volume, high-throughput, repetitive tasks that rely on structured or rules-based data, such as templated documents, spreadsheets or databases. For example:

- Customer service, such as automatically sorting and categorizing inquiries and shipping them to the most appropriate customer service agent for a quick response. An RPA system may also be able to offer an automated answer to simple FAQs, such as product return policies.
- Many HR tasks can be automated with RPA. To help with recruiting, software robots can collect resumes from various online platforms by matching predefined keywords to suggest promising candidates. RPA can also aid with employee onboarding, where multiple pieces of information need to be collected and input into an HR management system.
- In insurance and healthcare, some aspects of the claims management process can be automated with RPA, including inputting data from various sources and ensuring all required information is entered.

Serious challenges with RPA begin when unstructured data enters the equation. Processes which require nuanced decision making or involve non-repetitive/non-standardized tasks bring RPA to rapid halt. In more complex applications, RPA can also be brittle, requiring significant time and investment supporting production.

As soon as any form of unstructured data is introduced, the RPA tool needs to be supplemented with other technology, such as natural language processing (NLP).

Buyer beware:

Solution providers are increasingly playing fast and loose with their claims around unstructured data support. The vast majority of these are referring to basic template and rule-based approaches. When projects move from POC to production and encounter real-world variability, they fail...and fail hard.

NLP

Natural language processing tools are artificial intelligence technologies that read, analyze, and respond to text or speech like human beings would. NLP combines linguistic rules with machine learning and deep learning models to "understand" the meaning of language with intent and sentiment.

NLP technologies are increasingly being deployed in a wide array of use cases, including:

- Consumer digital assistants, voice-operated GPS apps, dictation software, and chatbots
- Virtual customer service agents and chatbots
- Sentiment analysis, which assess favorable, unfavorable or neutral perceptions from customers in everything from customer service/support conversations to social media chatter
- Fraud and spam detection
- Language translation (such as from Japanese to English)
- Text summarization

NLP technologies can also be a powerful tool when used in concert with RPA and OCR solutions. Take, for example, the situation outlined above, where an RPA tool might struggle with unstructured content. Now, imagine that an OCR tool can "read" scanned documents, and NLP then analyzes and understands the scanned text in the documents by taking context into consideration. An unstructured data platform can then structure that information, opening the door for RPA to now do what it does well.

Buyer beware:

NLP technology may be able to solve your unstructured data use case. But, in many cases, you’ll need an army of data scientists and a multi-million dollar compute budget in order to be able to successfully implement.

OOTB

One four-letter addition to the automation mix is OOTB. Many technology vendors offer point solutions for automation – or so-called “out-of-the-box” models. These offerings address a specific function for a business. Contract analytics. Accounts payable. Customer relationship management. Apps abound.

OOTB solutions can be very attractive; however, they come with a big buyer beware. It may seem convenient to have a ready-made solution for a particular business process – and it could likely please a single line of business leader who wants automation yesterday. If you truly have a one-off challenge to address, an OOTB solution might be fit for purpose.

That said, OOTB point solutions present a number of visible and potentially unseen pitfalls:

- As the point solutions rack up, so do the administrative responsibilities, from management to updates to training and more
- Support is equally complex when you have to work with an array of technology vendors and manage multiple relationships
- Point solutions seemingly never quite meet all of an organization’s needs and are inflexible to accommodate new requirements
- By the same token, many OOTB solutions aren’t really “out-of-the-box” when they may require extensive customization/configuration to meet an organization’s specific needs – which means it takes a very long time to unbox
- Point solutions don’t scale; you can get much more value from a standard, repeatable, and flexible solution that give you the agility to adapt as new uses, new users, and new process requirements arise

Change:

What makes Indico Data different

Enterprises have long struggled with their unstructured data. Though effective with structured data challenges, RPA vendors and point solutions have fallen down or fallen short with traditional approaches to automation. But now, the tide is turning thanks to breakthrough NLP and **deep learning** technologies.

At its simplest, deep learning is a type of machine learning that simulates the behavior of the human brain, allowing it to be trained and learn from very large data sets. It can adapt and recognize patterns in unstructured data in ways that RPA can't – enabling it to take unstructured documents and then restructure them for utilization.

At the forefront of this revolution is Indico Data and its pioneering Unstructured Data Platform. Through its innovative NLP, AI and ML software, the Indico Platform allows enterprises to ingest unstructured data at massive scale and add structure, enabling them to do what's been impossible with traditional automation and analytics tools: realize the unlimited potential of their unstructured data. And the key to this is a breakthrough approach that Indico has developed called "Composite AI."

The power of composite AI

Composite AI is about combining modern AI approaches including neural networks with a range of other AI approaches like rule-based reasoning, graph analysis, transfer learning, and machine teaching. The goal is to enable AI solutions that require less data and energy to learn and which embody more "common sense" approaches to model creation. Composite AI recognizes that no single AI technique is a silver bullet. Indico's approach leverages a powerful set of techniques including:

Multimodal fusion

Even the most sophisticated traditional AI systems are constrained to a single data mode, such as text or image. Indico can reason across data modes, incorporating visual information alongside semantic information to make decisions. This results in significantly higher accuracy for multimodal use cases like document understanding.

Transfer learning

Indico is the industry leader in the adoption and application of Transfer Learning. Transfer Learning allows customers, for the very first time, to build high-quality custom machine learning models with as few as 200 labeled examples. Transfer Learning allows customers to effectively "amplify" their small labeled datasets to perform as if they have labeled tens of thousands of samples.

Machine teaching & human in the loop

One of the biggest challenges in AI is training data. The ability to accurately capture real-world data is critical to the creation of high performance machine learning models. Machine teaching is, at its core, a human-centric approach to AI. Rather than asking "how do we learn?", we ask "how do we teach?" Further, enabling customers to reach the highest levels of efficacy often involves a human review at the end of a process, and Indico has developed the market's leading human in the loop interface to enable customers to capture efficiency gains while also saving time and money.

Empowering citizen data scientists

As a result, the process through which companies use Indico to build data models is simple and highly effective. Business subject matter experts label the data points they deem most important to whatever process they're looking to automate. As they apply labels, the model is updated on the fly and will start to show predictions on subsequent datasets. Once you're comfortable with the predicted results, you're done building your model.

The beauty of this approach is that the people who understand the business problem and the desired results – those on the business side of the house – are the ones who train the model. With Indico, there's no need to try to explain to a data scientist what you're after and then hope you get the appropriate results. Citizen data scientists can create models themselves. Everything is in plain English and users can have fully working models in a matter of hours or days, not weeks or months. With the Indico Platform, enterprises can:

Automate

Automate the intake of unstructured documents, emails, CSVs, images, videos, audio and more.

UiPath™

AUTOMATION ANYWHERE

Analyze

Analyze unstructured data, extracting actionable business insights and intelligence.

Power BI

tableau

Apply

Apply this data, creating new application experiences to transform manual processes into automated advantages.

SAP

salesforce

The Indico platform can handle the gamut of document processing needs, whether it involves highly structured documents, completely unstructured or something in between. It's effective because it's built on a database of several billion words and images, terabytes of data, and hundreds of pre-trained machine learning "tasks", providing a deep base of knowledge that gives it the context required to "read" and understand virtually any type of content.

Checklist:

Knowing what you need

Now you know the data. You know the lingo. You even know about a new data platform that can augment or outseat existing automation technologies in order to unlock the value of unstructured data. What else should you know? Here are best practices and recommendations to help you evaluate your automation needs – and to assess the automation vendors chasing your business.

1. Begin with the business, not the tech

Savvy automation leaders uniformly advise to start with understanding the needs of the enterprise and the individual line of business (LOB) first. Too often, organizations look straight to the technology without first understanding the business needs or human activities involved in a given process they may seek to automate. Or, a technology vendor may present a shiny new object instead of helping to properly assess what the customer's real pain point might be.

Conduct due diligence with LOB leaders and the financial lead of a particular business. Complete an assessment of the feasibility of automating the process and clearly articulate the business value. Then select the right tool(s) to match the task – and ensure you have strong feedback loops with your line of business partners. Too often, automation teams create business requirements in a black box and return solutions that yield a response from the business along the lines of: "That's not what I asked for." Create ample visibility, exposure and collaboration.

2. Identify areas for automation

Keep all your options open when you begin to assess technologies. Ask yourself:

- Are we the only ones or one of a handful of companies who deal with these documents/data this way?
- Does a solution already exist?
- Is this a custom workflow?
- What kind of automation solution do we need to buy or build?
- If we're building, what technologies do we need?

Recalling what we've covered about RPA, it could be a good tool for work that is, as one Indico customer puts it: "Very copy-and-paste, click here, click there, read from the same tables every time. Unless you can count on one hand the number of sub processes for the RPA, that's not the right tool for the job."

Something else to keep in mind: automation often reveals fundamental flaws in a business' process. If the process is broken or inefficient by nature, no amount of technology will fix it; it will only amplify the issue. Consider the workflow and address issues in the human activities involved before seeking to automate with a technology.

3. Make vendors prove their mettle

In the technology industry, it's not unusual for vendors to want to latch on to the latest trend and claim to have a product or service that fits the category. Artificial intelligence and unstructured data are no exceptions, which means you need to be vigilant about querying vendors to ensure their technology can really be classified. Point of fact: the London-based venture capital firm MMC found in a 2019 survey that of 2,830 startups in Europe that were classified as AI companies, only 1,580 – about 56% – actually offered AI technology.[3]

Here are three critical questions to ask prospective AI and automation solution providers to determine whether the technology they're offering is legitimate

What's your algorithm strategy?

Some AI vendors have their own, homegrown algorithms that they've trained over time. Others use open source frameworks such as TensorFlow, that are open to the general public and are constantly being improved. In Indico's case, the company constantly benchmarks the best in class transformer-based architectures in the market such as BERT, RoBERTa, and GPT2. When a new algorithm arrives and outperforms its predecessors, it is added to the platform and made available to customers, effectively future proofing them from new innovations. Either way, the vendor should be able to explain what its algorithm does and where it came from. If not, disqualify.

What's your data strategy?

Here the answers may vary. For Indico, for example, the answer lies in our generalized model that is the baseline for all of its tools. The Indico base model consists of billions of worlds and images and terabytes of data, which is enough to enable it to understand human language and context broadly. Users can then customize the model to take on whatever task they're trying to tackle, but using 100x to 1000x less data than would normally be required. Indico also never pools customer data together to ensure 100% information security. Here again, if a prospective vendor can't articulate its own data story, perhaps they're part of the 40% that aren't really selling AI.

What's your application strategy?

For AI to be useful, it has to come with some form of application that makes it accessible to those who want to employ it, whether data scientists, IT, or business people. Indico, for example, has a point-and-click user interface that makes it simple for anyone to build effective process automation solutions and models, without the need for data science expertise. Without some sort of application like that to make the AI technology useful, you're not going to realize the full value of it.

Ready to unlock the value of your unstructured data?

With the Indico Unstructured Data platform, big data has never been more unlimited or more powerful. To learn more about how you can put all your data to use, request a demo and no-cost consultation at IndicoData.ai

About Indico Data

Indico Data transforms unstructured data into actionable insights. With the Indico Unstructured Data Platform™, enterprises of all sizes can automate, analyze, and apply unstructured data –– documents, emails, images, videos and more –– to a wide range of enterprise workflows. This enables them to gain rich insight and maximize the value of their existing software investments, including RPA, CRM, ERP, BI, by enabling these systems to work with unstructured data.

For more information, visit IndicoData.ai

Sources

1. Gartner, https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021/

2. IDC, "FutureScape: Worldwide Artificial Intelligence 2021 Predictions"

3. The Verge, https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report

The information in this eBook has been prepared by Indico Data and is for informational and marketing purposes only. It does not constitute legal, financial, business or investment advice of any kind and is not a substitute for qualified professional advice. You should not act or refrain from acting on the basis of any content included in this eBook without seeking the appropriate professional advice. The contents of this eBook may not reflect current developments or address your specific situation.

Indico disclaims all liability for actions you take or fail to take based on any content in this eBook. Although the information in this eBook has been gathered from sources believed to be reliable, no representation is made as to its accuracy. This eBook is not an endorsement or recommendation of any third-party products or services of any kind.