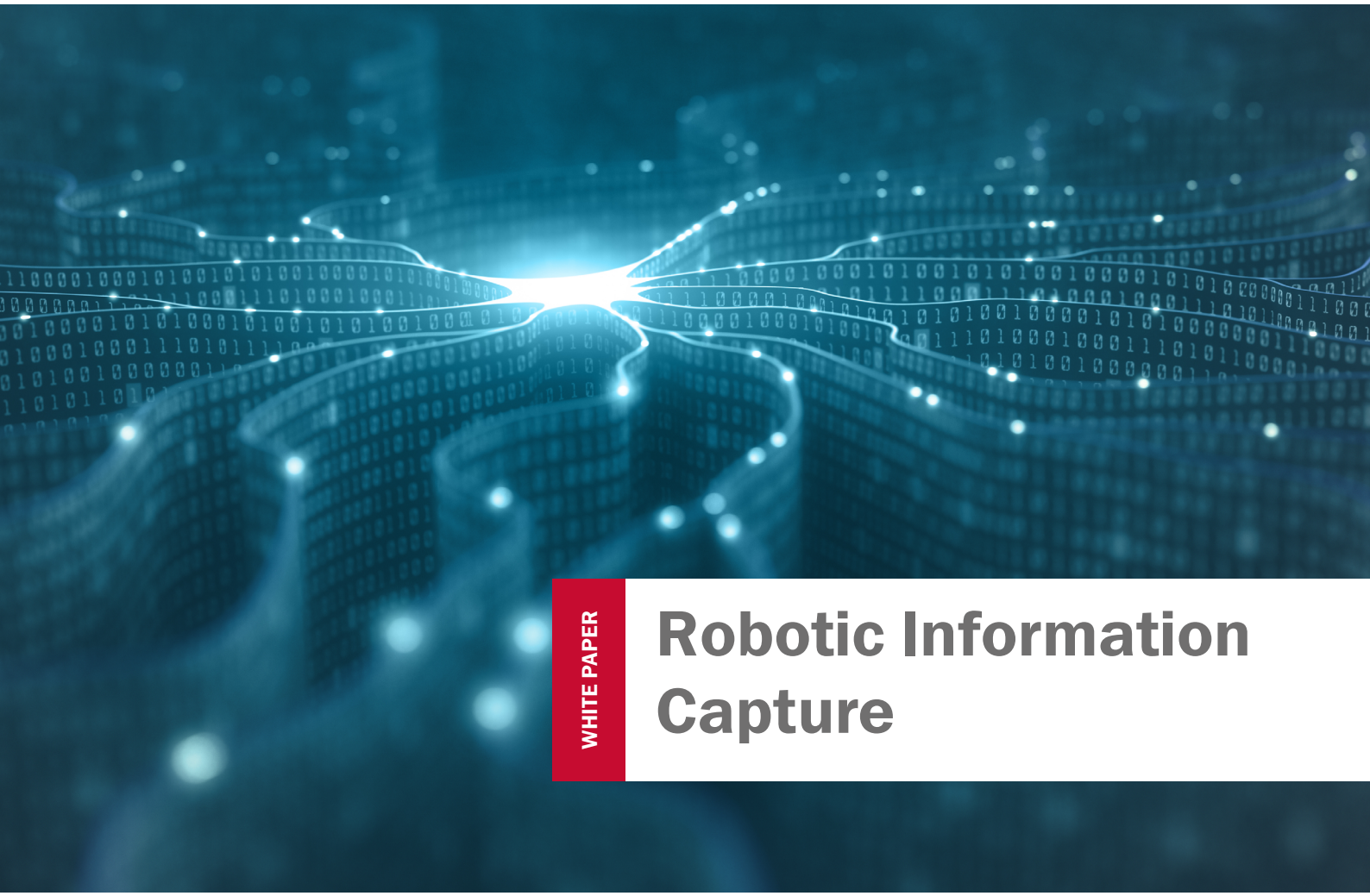
ABBYY®



WHITE PAPER

Robotic Information Capture

DATA ENABLEMENT FOR MACHINE LEARNING THROUGH ROBOTIC INFORMATION CAPTURE

Content

01 Introduction

In our global information-driven economy today, how well an organization can capture, utilize and benefit from information has a direct influence on their business and overall company performance. Many organizations still face challenges when it comes to effectively acting on this information because their data lacks context. Petabytes of information are being stored every day, which creates an amazing opportunity to ignite machine learning through the informational data we already receive from mobile, the cloud, and existing backend solutions. But the issue is how to understand it and use it in a cost effective manner. Overwhelmed, firms now have opportunities to train Artificial Intelligence (AI) with these massive data stockpiles, and other untapped resources to build relevant contextual information from large data sets and images on-demand.

Those who have generated a thoughtful and strategic approach to Big Data and AI will have a distinct advantage over their competition, but getting there is the challenge. AI training is typically messy and complicated, requiring large amounts of human guidance with tens of thousands of data samples at a time to properly train. The big question is, "how do we teach AI?"

"The labor involved in preparing an automated capture solution is currently very high. An AI-driven Robotic Information Capture (RIC) approach will reduce the cost and bring sophisticated flexible capture systems online quicker. Future capture solutions will require flexible architectures built on smart microservices enabled as cloud REST APIs."

Harvey Spencer, HSA[1]

Robotic Information Capture (RIC) is the answer to the transition and data enablement for machine self-learning. These systems help to collect and prepare AI training data from growing backlogs and current everyday processes using AI technology, avoiding the need for highly-skilled technical professionals to provide complex training. Thus, it enables common office workers and casual users to train the system on-demand when needed. Powered by machine learning, these robotic systems automatically develop document classification, data extraction and validation rules by analyzing an operator's actions as well as document structure, layout, keywords and other parameters. After several iterations, these systems can now repeat these actions on their own, letting the operators simply verify the results and easily handle exceptions moving forward.

To gain more insight from these unstructured or semi-structured inputs, the user simply adds more context to the learning experience.

Notes: 1. HSA Harvey Spencer Associates Inc. (www.hsassocs.com)

02 The AI Advantage

AI provides computers with the ability to learn and make decisions, without being explicitly programmed by a human being. A new report from Forrester predicts that investment in Artificial Intelligence will grow by 300% in 2017 as executives and IT leaders invest in AI to reduce costs, scale faster, and make better business decisions faster.[2] Analysts say that by providing access to powerful insights, never before available, AI will, "drive faster business decisions in marketing, e-commerce, product management and other areas by helping close the gap from insights to action."

If you thought that AI was something for the future, think again. Banks are using AI technology to handle activities like investing in stocks or managing different financial operations in real-time. AI technology is also changing the face of healthcare with the development of virtual personal health care assistants and a variety of "healthcare bots" that schedule appointments, providing more responsive patient support. Large manufacturers use AI, in not only in production but also in management; extracting and analyzing data for decision-making and product strategy design.

03 AI Requires Data

Making the most of AI requires more than just the purchase of the technology. The proper preparation of training data is essential. It is important to remember that there is a significant difference between training a human operator and training an AI robot, in terms of the amount of data required. One example of this is self-driving cars. A human operator learns rather quickly whereas it can take years to "train" a self-driving car. Humans typically learn by driving around the block, parking a few times, and then after a few miles on the freeway they are ready to drive almost anywhere. Machines, on the other hand, require data from hundreds of thousands of miles driving in various locations and under different conditions to be able to learn all possible scenarios, before they can drive on their own. That is because humans understand the context of traffic, signs and lights since childhood, while machines have none of that background.

Notes: 2. Forrester "Predictions 2017: Artificial Intelligence Will Drive the Insights Revolution".

This is why collecting and preparing training data is the number one challenge for any organization working to adopt AI. Even though AI algorithms are publicly available, each organization will have to develop its own set of training data in order to create a viable AI system specific for its own tasks. This situation works in favor of large-scale digital organizations like Google or Amazon, which have already collected vast amounts of data in actionable electronic formats. They can easily apply it for training generic AI systems.

The trouble for organizations lacking a vast info-infrastructure like Google or Amazon is that most companies store their data inside diverse information systems – a variety of ERP, CRM and case management databases, not to mention legacy document management systems, electronic archives, and even paper archives. Some pieces of this data are more available and actionable than others. Strategically important information is never shared deliberately. Nevertheless, in almost all cases data, which was sufficient for human-driven procedures, is not enough for an AI-driven approach.

04 Information Capture is the First Step to Training AI

We already know that the first step of implementing any AI solution is collecting data. For the majority of small and mid-sized organizations, this means that they must execute major data enabling initiatives to extract valuable information intelligence from traditional archives, scanned image-only documents, and file storages. This is a mandatory first step of each AI project.

However, data extraction and preparation requires a lot of time and effort. This process could be significantly automated, but using a traditional data capture system is also challenging for smaller, resource-stressed organizations. Typical automated data capture is an enterprise-scale process that involves a combination of scanning and computing hardware, automation software and human operators. In addition to costly hardware and software, modern automated data capture solutions typically require expensive highly experienced engineers or professional services to set up the systems and define classification and data extraction rules.

In order to illustrate this challenge, let us consider an example. As a part of a new automation project, a personal software company wants to implement an AI assistant offering custom services by the city that a potential customer works in. Unfortunately, the customer's business address information is not always available in the CRM, and they decided to fill the gaps from stored business cards, collected at industry events.

To set up such a project, we need to specify classification rules such as how to separate business cards from other documents collected. Using the size and format of the document for image-based classification could be one way; but we can still confuse a ticket with a business card, so we need to use text-based classification and look for certain keywords or specific data formats, such as phone numbers, emails, etc., which are typical for business cards.

You may already feel the challenge. A specially trained professional will have to visually analyze documents, define how they could be differentiated, then set up these rules manually, test them, handle exceptions and repeat the process.

Data extraction rules are even more complicated. They require you to specify fields you are looking for and describe rules of how to find these fields using keywords or layout information from the document.

05 Getting Started with Robotic information Capture

Imagine that the capture software is powered by machine learning and does not require complex training to get started. Instead, it can learn the rules of document classification and data extraction by analyzing operator's actions in real-time with the user.

Now you do not need a trained engineer to setup these rules. The operator selects the first document from the batch and defines its class (e.g. receipt.) Then the operator picks the next document. The system assumes that it is also a receipt since no other classes are defined. The operator either confirms it or creates a new class (e.g. ticket). The system now analyzes a large set of parameters to find and internalize the differences between two classes. For the next document, it suggests a class, which is more appropriate according to the previous training and the operator either confirms it or corrects. Through this simple procedure the operator can do both tasks at the same time – process the documents and train the system, so the future documents require less work.

Automatic data extraction training looks similar. The operator defines a list of fields to extract for a particular document type, then loads the document and selects the area on the screen for each field as its location. Since documents are already pre-OCRed, the system understands the document's structure and layout, knows the location of text versus images, alongside other objects on the page, so the operator does not need to be pixel-level accurate while selecting fields.

During this process, the system automatically learns how to find each field and applies this learning to the next document, suggesting the location of the fields. The operator either confirms or corrects it. After the few iterations, the system can locate fields and extract data automatically, so the operator may just verify and confirm extraction.

We call this new approach, "Robotic Information Capture, or RIC." It allows regular operators without any special skills to train the data capture system by continuously replacing and automating routine data capture operations, which usually require either manual data re-typing or complex pre-training by experienced and expensive professionals. RIC lowers the barrier for adopting capture automation for small and medium-sized organizations, by leveraging AI technology for automatic training. It also expands the use of the technology beyond extracting data from a page to capturing meaning through natural language processing.

06 Moving Forward with Robotic information Capture and AI Training

AI is changing IT systems and processes in the organizations. It opens new opportunities for organizations to collect, extract and leverage their data for automatic decision making, resulting in higher operational efficiency and profitability.

After the transition period, when AI-driven business processes will be commonly adopted, larger enterprises will receive further advantages because they have more identified data, making their AI-driven systems more powerful and better at making decisions.

Smaller companies simply do not have enough data to take advantage of AI technology, however, tagged data and identified information is the new world currency as service providers can potentially host and train cloud-based AI systems for others by leveraging aggregated data from multiple companies. By bridging a gap between AI technology, providers, and customer organizations, service providers will play an even more significant role in the future of business IT ecosystems.

Request an ABBYY FlexiCapture demo now to see RIC system in action.

ABBYY
North American Headquarters
880 N. McCarthy Blvd.
Suite 220 Milpitas. CA 95035, USA
Tel+1866.463.7689
Fax +1 408.457.9778
sales@abbyyusa.com

www.ABBYY.com

ABBYY®

Request an ABBYY FlexiCapture demo