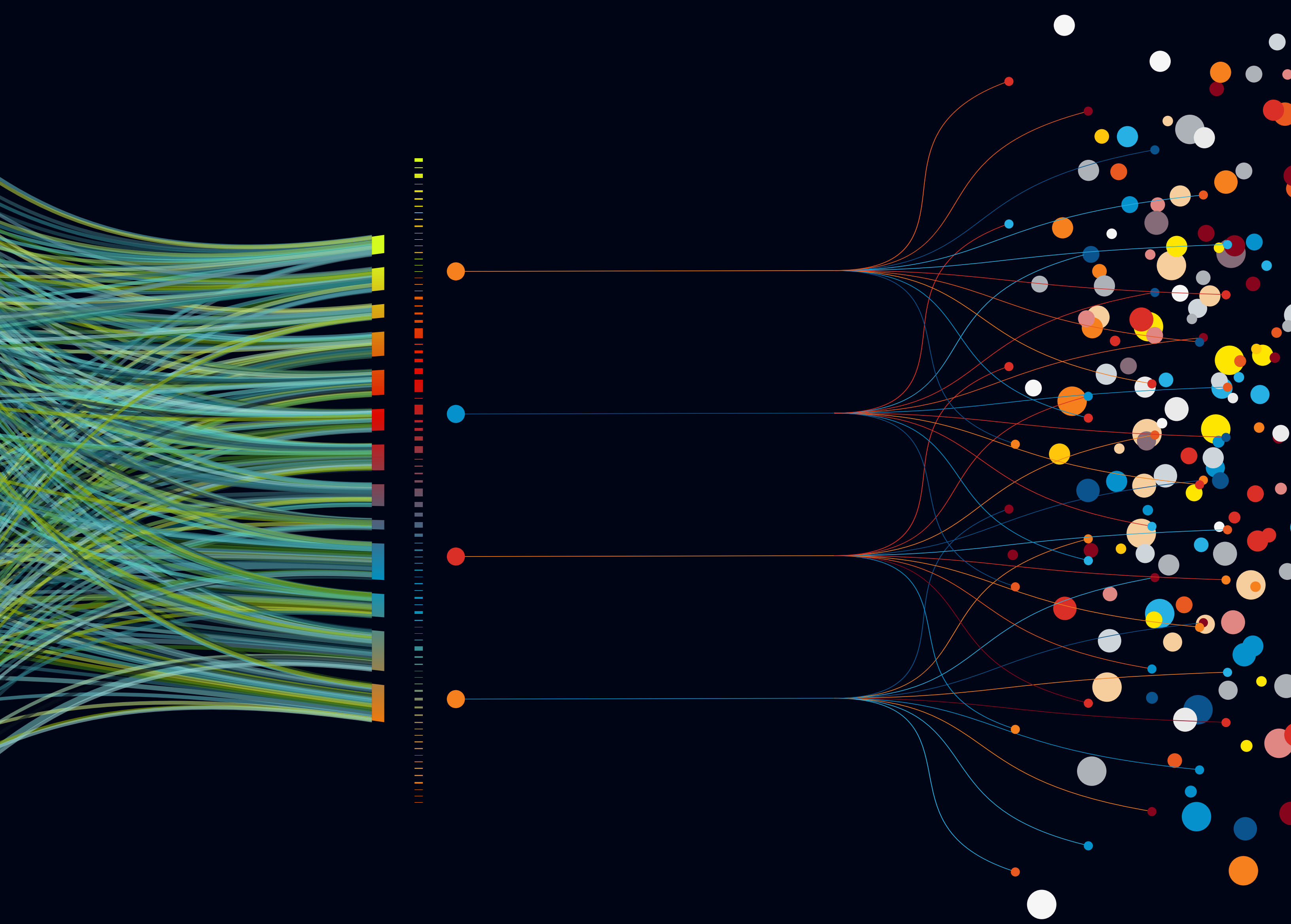
LTIMindtree



Data Extraction

- The Holy Grail of End-to-End Automation

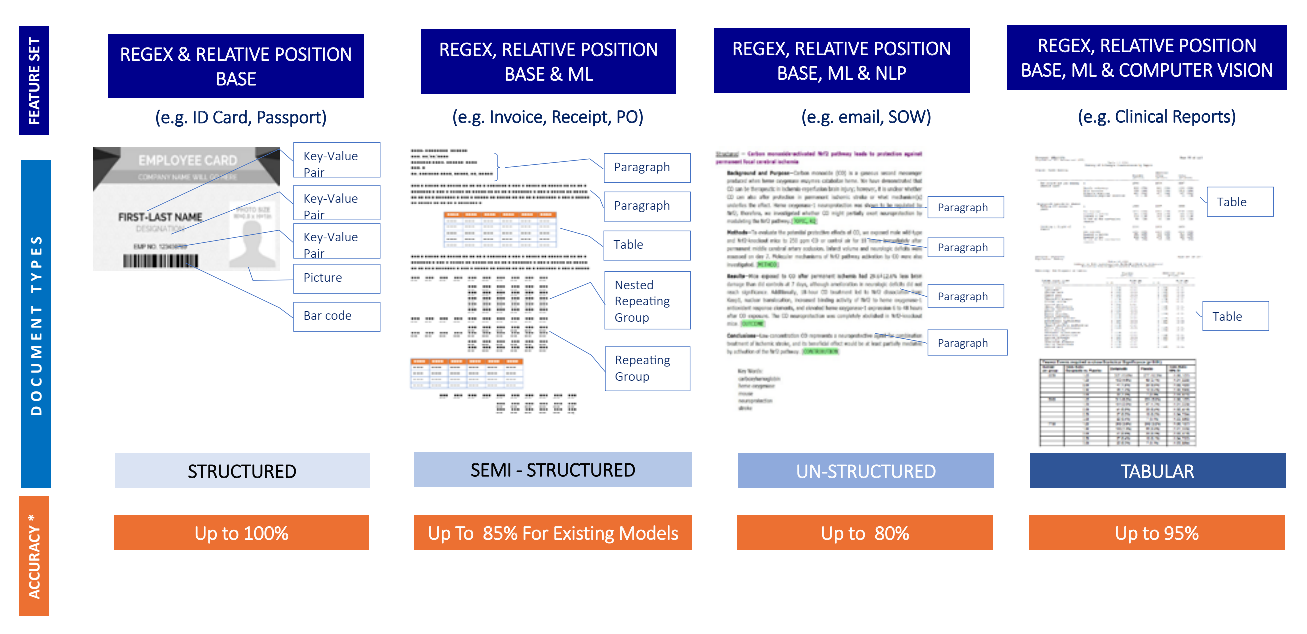
Intelligent automation has become a business imperative in the current times as more and more organizations transition from a pure-play Robotic Process Automation (RPA) strategy to hyperautomation. The typical RPA use cases, i.e., the low-hanging RPA fruits, have been consumed, creating an innate need to find newer and complex use cases to help organizations move towards the goal of end-to-end automation.

Organizations are increasingly investing in cutting-edge digital technologies such as advanced Optical Character Recognition (OCR), Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and virtual assistants (chatbots) to attain this goal. By adding an element of intelligence to existing bots, these technologies help provide actionable insights that transform a customer's experience and achieve true end-to-end automation across use cases. These technologies reduce Full-Time Equivalent (FTEs) and ensure the compliance of critical norms.

Document Classification and Data Extraction

As observed in the automation space over the last few years, the absence of a proper data extraction approach is an obstacle to achieving end-to-end automation. Furthermore, thousands of varieties of documents in paper and digital format as well as in various shapes, sizes, and forms complicate the design of an extraction strategy.

Each business process uses a different set of unique documents. These documents can be broadly classified as structured, semi-structured, and unstructured, as shown below.



FEATURE SET	REGEX & RELATIVE POSITION BASE (e.g. ID Card, Passport)	REGEX, RELATIVE POSITION BASE & ML (e.g. Invoice, Receipt, PO)	REGEX, RELATIVE POSITION BASE, ML & NLP (e.g. email, SOW)	REGEX, RELATIVE POSITION BASE, ML & COMPUTER VISION (e.g. Clinical Reports)
DOCUMENT TYPES	STRUCTURED	SEMI - STRUCTURED	UN-STRUCTURED	TABULAR
ACCURACY *	Up to 100%	Up To 85% For Existing Models	Up to 80%	Up to 95%

The illustration depicts different types of documents that fall under each category. Clearly, a one-size-fits-all approach for data extraction will not work.

The process of standardizing document templates is an integral part of a data extraction strategy, as evidenced in the path-breaking OCR-based data extraction product used in the US healthcare industry a decade ago. The Health Care Finance Administration (HCFA) forms developed by CMS are an excellent attempt at standardization. An OCR product scans the data on the HCFA forms and ignores the printed red overlay, increasing the speed and accuracy of data extraction. This standardization can help digitize the claims process from end-to-end, reduce the overall turnaround time to settle a claim, and minimize the volume of healthcare claims handled by Third-Party Administrators (TPA).

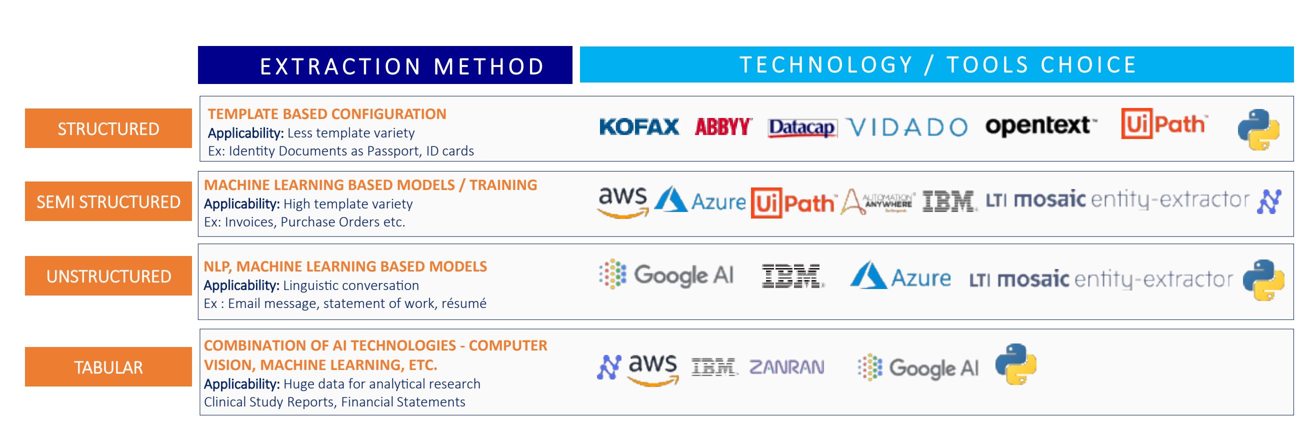
One Size Does Not Fit All

Over the last few years, digital documents such as Portable Document Formats (PDFs) and spreadsheets have ushered in different dynamics. As organizations embrace digitization, IT systems have resorted to creating digital documents to replace traditional paper-based documents. Since paper-based documents are unavoidable, a data extraction approach needs to strike a balance between both formats.

A finely tuned OCR-based application can help extract data from images of paper-based documents, while technologies such as ML, relying on a key-value-based search, can be leveraged to extract data from digital documents. One such tool is LTIMindtree's Mosaic Agnitio. Such ML-based tools also increasingly learn, making data extraction faster and accurate.

Another alternative to data extraction is to develop light-weight data extraction components using Python, AI, and ML for specific needs. LTIMindtree has developed one such solution to split PDFs that contain thousands of supplier invoices into individual invoices to send them to respective business entities for processing. By employing innovative approaches, organizations can minimize costs and manage the technical debt that will arise from automation initiatives in the future.

The illustration below shows some of the data extraction tools that LTIMindtree has explored to build expertise in devising data extraction strategies.



	EXTRACTION METHOD	TECHNOLOGY / TOOLS CHOICE
STRUCTURED	TEMPLATE BASED CONFIGURATION **Applicability:** Less template variety Ex: Identity Documents as Passport, ID cards	KOFAX ABBYY Datacap VIDADO opentext UiPath
SEMI STRUCTURED	MACHINE LEARNING BASED MODELS / TRAINING **Applicability:** High template variety Ex: Invoices, Purchase Orders etc.	aws Azure UiPath AUTOMATION ANYWHERE IBM LTI mosaic entity-extractor
UNSTRUCTURED	NLP, MACHINE LEARNING BASED MODELS **Applicability:** Linguistic conversation Ex : Email message, statement of work, résumé	Google AI IBM Azure LTI mosaic entity-extractor
TABULAR	COMBINATION OF AI TECHNOLOGIES - COMPUTER VISION, MACHINE LEARNING, ETC. **Applicability:** Huge data for analytical research Clinical Study Reports, Financial Statements	aws IBM ZANRAN Google AI

Choosing a Data Extraction Method

A data extraction approach is necessary to march towards the goal of complete automation. In fact, pioneering organizations have made much progress in this space. COTS tools are a good alternative for those that lack the resources to explore and design options since these tools can integrate with RPA and Business Process Management (BPM) tools and transform and automate business processes.

Author profile

Praveen Naregal
Associate Director – Digital Engineering Group, LTIMindtree

Praveen has 18+ years of extensive experience in Product Engineering, Automation and Digital Transformation. He is experienced in building multiple enterprise scale products & solutions for various industry verticals. He is passionate about building high-performing teams and solving complex customer problems by intelligent application of technology. He is currently responsible for delivery of Intelligent RPA and Digital solutions to the Oil & Gas vertical.

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 750 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by nearly 90,000 talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — combines the industry-acclaimed strengths of erstwhile Larsen and Toubro Infotech and Mindtree in solving the most complex business challenges and delivering transformation at scale. For more information, please visit **www.ltimindtree.com.**