

Context – Unified Approach to Extracting, Normalizing and Analyzing Unstructured Data

Presented by:



Executive Summary

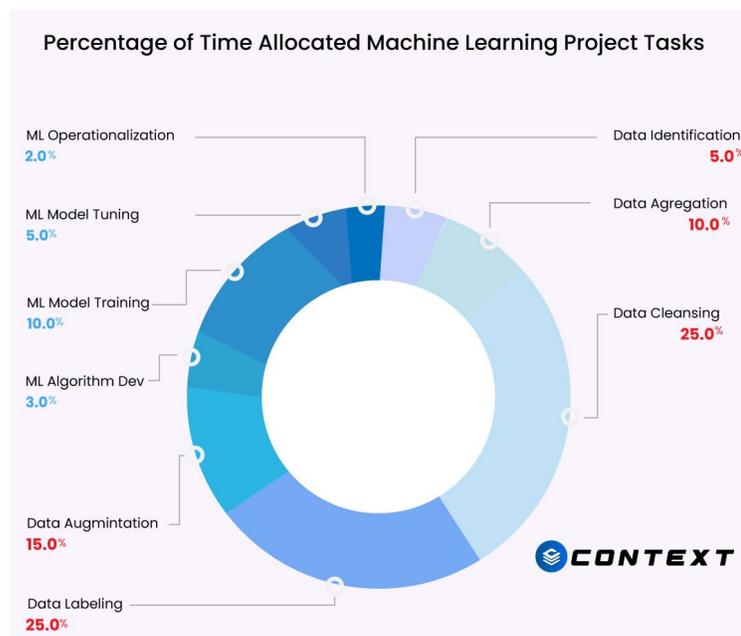
By most estimates, up to 90% of enterprise data is unstructured and never analyzed (Marr, 2019); 70% in the form of text (e.g., emails, documents). By the end of 2021, more than 80% of organizations will fail to develop a consolidated data security policy across silos, leading to potential noncompliance, security breaches, and financial liabilities. (Davis, 2019) 71% of enterprises struggle with managing and protecting unstructured data (Rizkallah, 2017), and leading Robotic Process Automation vendors estimate that ~50% of their workflows still begin with a document. In addition, many organizations fail to leverage their data in a manner that produces insights into their processes, leading to potential loss of revenue or efficiencies.

Business and Technical Challenges

There are few turn-key, out-of-the-box solutions that organizations can use to solve their problems. Most tools require custom development and API integrations that take months to implement and may not lead to a well-integrated solution. Once the data is extracted, organizing and normalizing different data structures can be daunting for most organizations and lead to cost overruns and delays. Data analysis can be challenging because the data is not organized in a single place, leading to siloed Analytics.

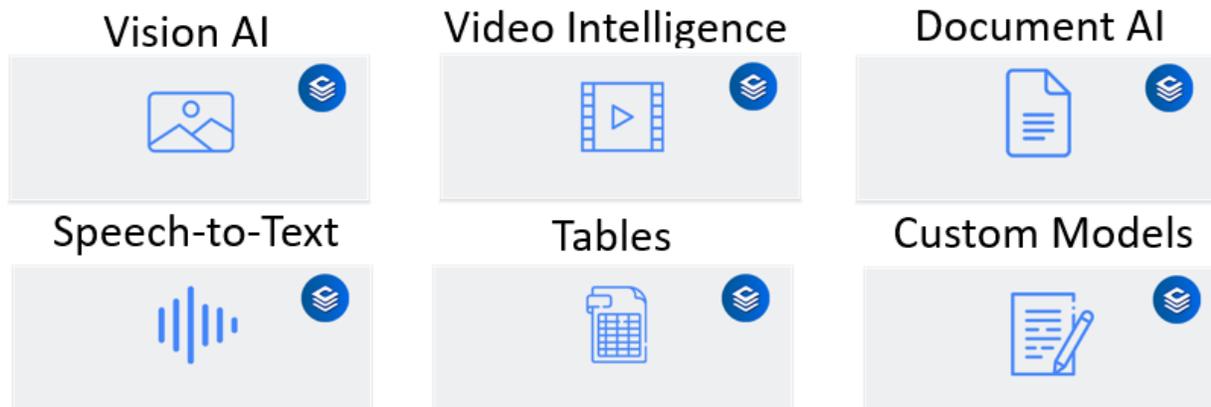
Lastly, training custom models to extract relevant information requires significant investments and resources towards efforts that lead to little to no gain for organizations. As of 2020, 80% of AI/ML projects have failed for reasons that can be summarized into the 4 points (MV, 2020):

- 1) Challenging to get AI Projects off the ground
- 2) Non-AI data tasks take the longest
- 3) Integration between technologies
- 4) Custom Development is time-consuming



Context – How Context Solves Business Problems

Our team at Aretec has developed a fully integrated AI platform called Context that takes the complexity out of AI. Context gives organizations control over their data and analytically empowers their workforces so that anyone can work with AI without understanding code. Our user-centered design, along with industry-leading AI models, allow organizations to upload, organize and analyze their unstructured datasets without any need to develop code or AI/ML models.



Context AI Solutions

- 1) **Document AI Parsers:** Parsers are specialized machine learning algorithms that have been productized for specific kinds of documents, like bank statements, invoices, etc. Out-of-the-box, Context has integrated parsers developed by Google, based on their years of experience in the field. Document AI provides support for general document parsers, industry-specific parsers (e.g., 1040, W2), or custom parsers we train specifically against customer needs. Aretec offers specific services to accelerate parser training and deployment for custom use cases.
- 2) **Video Intelligence / Vision AI / Speech-to-Text:** Context’s out-of-the-box Video, Vision, and Speech processing capabilities allow users to derive insights using pre-trained models. Object detection, audio transcription, labeling, and classification are some of the insights gained by users out of the box. This is also an area where Aretec can custom-develop models based on customer use cases.
- 3) **Enterprise Knowledge Graph:** This exciting feature helps in entity enrichment, entity linkages and normalization of the extracted data from the documents. Our proprietary software allows



Document image



```
{Class: /us/gov/ID/Driverslicense
State: Ohio
Name: Jane Doe
Issued: 7/2/2018
Expires: 1/21/2021 }
```

Structured data

organizations to leverage out-of-the-box knowledge graphs or build their own knowledge graphs based on their understanding of linkages between data sets.

- 4) **Human in the Loop (HITL):** HITL is an assistive tool that allows customers to introduce human reviewers in the digitization process where necessary, verifying and validating document metadata extracted by Document AI. It also allows for in-line updating of the captured metadata or suggestions for fields that might have been missed while continuously training underlying models to increase the accuracy of future predictions. Reviewers may accept or reject documents based on confidence/accuracy scores, triggering further activities in the document processing workflow.
- 5) **Data Warehouse:** As part of the extraction, Context inserts the document metadata inserted into a Data Warehouse where it can be enriched with other sources and aligned with data governance initiatives. Once in the data warehouse, organizations can to run machine learning models using an easy SQL-like syntax to uncover deeper insights and predictions from the data.
- 6) **Search:** Our advanced search feature allows users to search for keywords and values throughout their datasets, whether structured or unstructured.

Aretec's Context Accelerators

Aretec helps our clients accelerate their journey to modernizing their data and gaining insights by leveraging our experience in data science and consulting with our professional services. Some of our accelerators include:

- 1) Heuristic / Fuzzy Matching – Match key/value pairs using custom models
- 2) Custom Models – Our data scientists can build custom models for requirements not met by out-of-the-box models
- 3) Labeling-as-a-Service – Data Labeling and Annotation services for images, videos, documents, forms and other niche use cases.
- 4) Model Uptraining – Uptraining of existing models for more accuracy
- 5) Workflow Automation – Custom automation using RPA + Context
- 6) Custom Integrations – Integrate with leading providers such as ServiceNow, Salesforce, AWS, Azure, Dropbox, Postgres, Oracle and many others

About Aretec

Aretec, Inc. (Aretec) is a data science firm specializing in application development, quantitative research, and tailored advanced analytics platforms that transform data into actionable insights to aid federal government decision-makers. Aretec specializes in developing creative solutions to technical problems using appraised Capability Maturity Model Integration for Development Level 3 (CMMI-DEV/3) and CMMI for Services Level 3 (CMMI-SVC/3) engineering practices focused on customer requirements. Aretec is also appraised for ISO 9001:2015, ISO/IEC 20000-1:2011, and ISO/IEC 27001:2013.

Bibliography:

Davis, D. (2019). AI Unleashes the Power of Unstructured Data. Retrieved 28 July 2021, from <https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html>

Marr, B. (2019). What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone. Retrieved 28 July 2021, from <https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/?sh=1e91500015f6>

MV, P. (2020). Top 10 Reasons Why 87% of Machine Learning Projects Fail - DZone AI. Retrieved 28 July 2021, from <https://dzone.com/articles/top-10-reasons-why-87-of-the-machine-learning-proj>

Rizkallah, J. (2017). Council Post: The Big (Unstructured) Data Problem. Retrieved 28 July 2021, from <https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/?sh=72e0ad51493a>