

Empower your business use cases by getting insights from text

Matteo Gabrielli
Solutions Architect @ AWS

Conf42 – Machine Learning, July 29th 2021

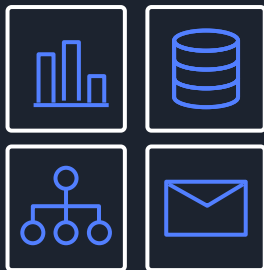
Agenda

- The need for automating document processing
- Challenges in automating documents
- How AWS can help
- Demo – Document Understanding Solution

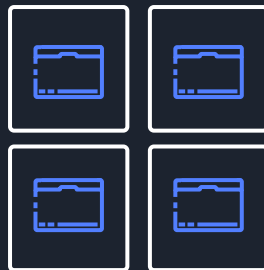
Current state of documents



Primary business tool across industries such as Finance and Marketing



Various sources and formats



They live in silos



They are growing in numbers every year

The deluge of unstructured information

“A Gartner research claims that organizations worldwide record a 25% growth in usage of paper each year. Paper continues to be a hindrance for many organizations owing to difficulties in processing and extracting information from such documents. More often than not, this process is carried out manually in organizations making it an arduous task that is prone to errors.”

Source: “Real-time Applications of Intelligent Document Processing,” February 2020.

<https://medium.com/high-peak-ai/real-time-applications-of-intelligent-document-processing-993e314360f9>

Challenges in processing documents



Extracting text manually is time-consuming, error prone, and expensive



Current rules-based systems are not intelligent and break with format changes



Extracting insights from documents requires large volumes of labeled data and ML skills



Some use cases may require human oversight; building human review workflows is complex and may increase time to market

How AWS can help

Understanding documents

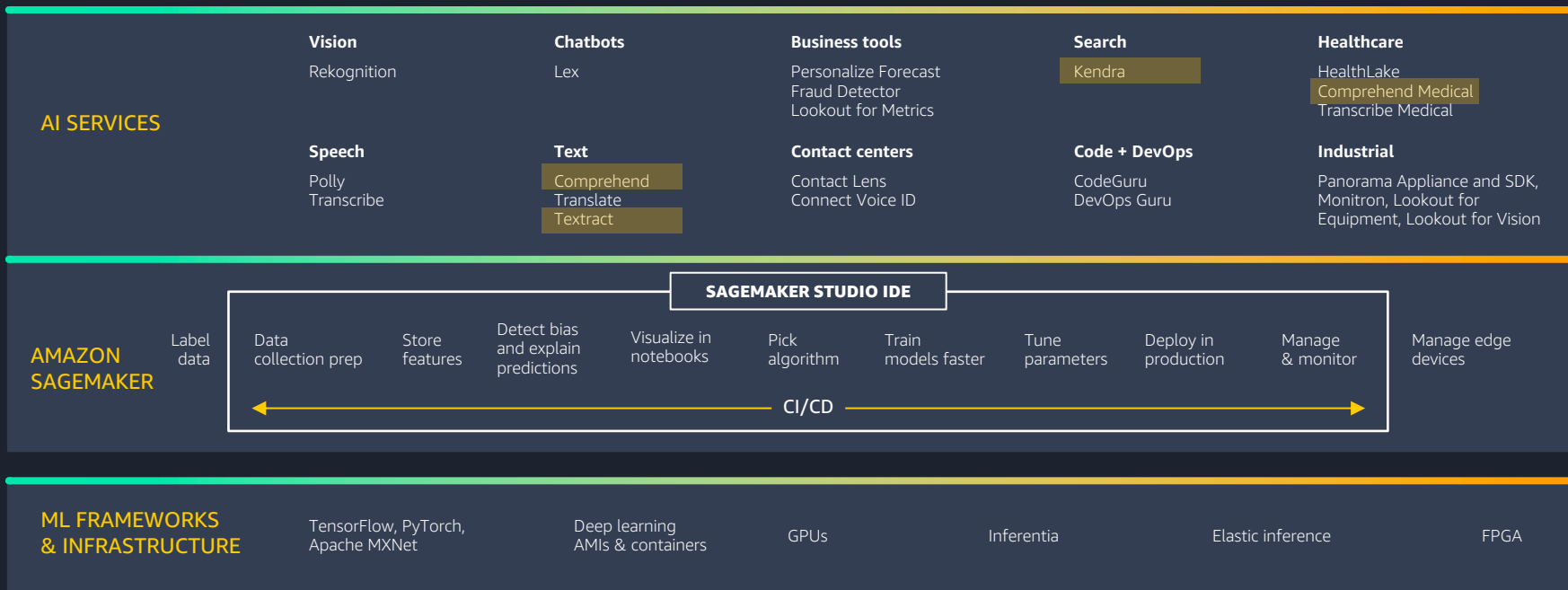
It's a process — Extract, Understand, and validate data by human oversight



AWS AI services can help with these steps

The AWS ML stack

Broadest and most complete set of machine learning capabilities



Amazon Textract

A service that extracts text, forms, and tables from documents such as pdf, picture and many more.



Extract data quickly
and accurately



Eliminate
manual effort



Lower document
processing costs



No ML Experience
Required

Amazon Comprehend

Discover insights and relationships in text



Documents

Email, chat,
social, phone
calls and more



Amazon
Comprehend

Automatically
extract insights
from text



Entities

+ Custom Entities



Key Phrases



PII

(Personally Identifiable
Information)



Sentiment



Document
Classification



Topics



Language



Syntax



Events

Typical Amazon Comprehend use cases



Automation of email workflows



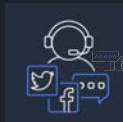
Customer support tickets routing



Documents and media tagging



Intelligent document processing



Customer sentiment analysis



Contact center call analysis



Personally Identifiable Information (PII) detection and redaction

Amazon Comprehend Medical

Medical Named Entity and
Relationship Extraction (NERe API)

Protected Health Information Identification
(PHId API)

Entities

- Medication
- Medical condition
- Test, Treatments and Procedures
- Anatomy
- Protected Health Information (PHI)

Relationship Extraction

- Medication and dosage
- Test and result
- Many more

Entity Traits

- Negation
 - Diagnosis, Sign or Symptom
-

Distill a complex process into a simple API call

Amazon Kendra

Amazon Kendra is an intelligent search service powered by machine learning that makes it easy to quickly find accurate answers in text, surface relevant FAQs and documents using Natural Language Understanding and Reading Comprehension out of the box.

Lexical
search

The screenshot displays an 'Intranet Search' interface. At the top, a search bar contains the text 'IT support desk'. Below the search bar, it indicates 'Your recent searches' and 'Not finding relevant results'. The search results are displayed in a list format, showing categories like 'Everything (21)', 'Wiki (1)', and 'Email List Archive (5)'. The results list includes items such as 'IT_Support_Training_Program.Wiki', 'Com_Support_Wiki.Web', 'OperationalBestPractices.Event...', and 'Corp_Wiki_Pending.Web'. A circular arrow icon points from the search bar area towards the results. On the right side of the interface, there is a 'RESULTS PAGE' section with a search bar containing 'Where is the IT support desk?'. Below this, it shows 'Kendra's suggested answer' as '1st floor', followed by a detailed description: '... our IT help desk, located at the 1st floor and open for support at most hours. The one in Seattle is on the 1st floor and is open from 12:30 to 5 p.m. daily.' There are also 'Frequently asked questions' listed below, including 'Where do I get IT help?', 'What are the IT support hours?', and 'Where can I get IT help corporate campus?'.

Intelligent
Search

Amazon Augmented AI (Amazon A2I)

Easily implement human review of machine learning predictions



Easily implement human review workflows



Reduce time to market with pre-built workflows and UIs



Multiple workforce options

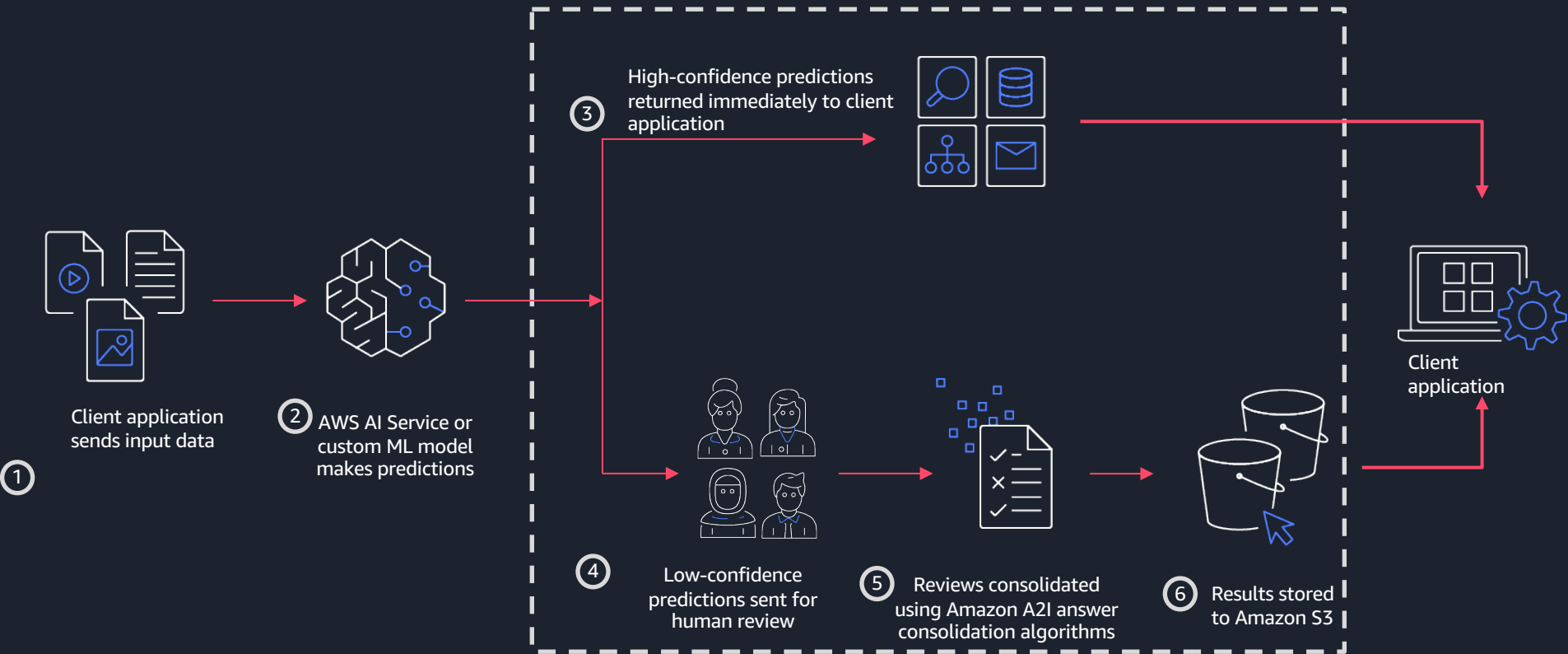


Integrate with your custom ML models



Pre-built algorithms to increase accuracy

How Amazon A2I works



Textract works with Amazon A2I for human review

Regulatory
requirements

Sensitive decisions
(e.g., issuing a loan)

Hard-to-read
documents

...more



Human review
provides oversight and
reviews for low-
confidence results.

Amazon A2I with Amazon Textract: Defining conditions

Confidence score



Trigger human review for form fields retrieved with less than 90% confidence

Important keys



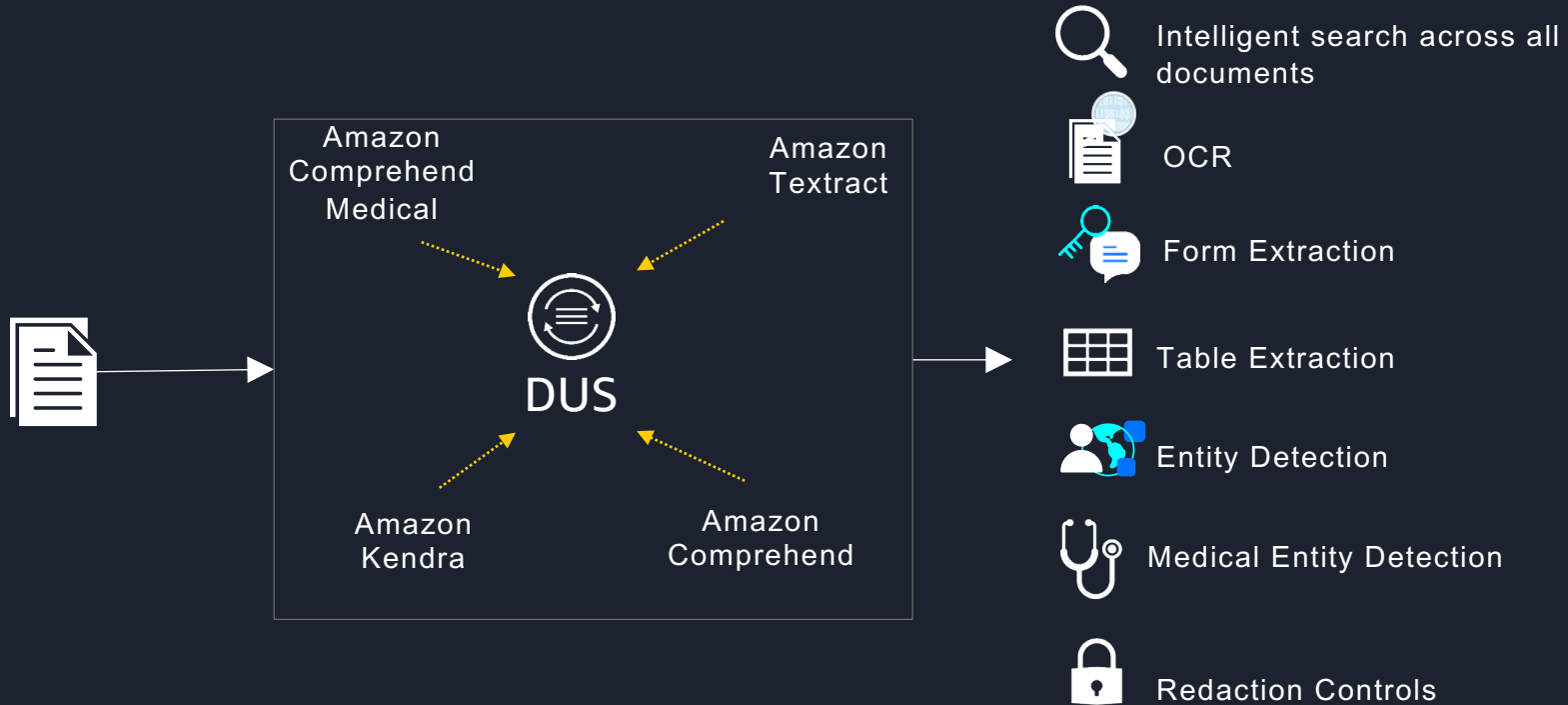
Trigger human review if you don't find a field called Adjusted Gross Income

Random sample



Trigger human review for a 10% sample of all documents

Document Understanding Solution (DUS)



Demo

DUS details

Know more about Document Understanding Solution

DUS Demonstrates Value of Intelligent Search

Q what percentage of FTE used in trade promotion goal ? X Search

Elasticsearch
Keyword Search Results

Amazon Kendra
Semantic Search Results

Elasticsearch and Amazon Kendra
Compare Search Technologies

Elasticsearch

Keyword search results

FOUND ABOUT 50 RESULTS ACROSS 10 DOCUMENTS

[management.png](#)

...to transforming our economy, Slightly Below 3 1 5 8 4 fostering U.S. competitiven...
...Improved 1 1 development of new businesses (USPTO, EDA....
...Not Met 5 6 2 2 2 6 8 10 NIST, and NTIA) See Appendi A: Performance and Reso...
...includes all of the Bureau of Industry and Security (BIS), and portions of ITA. 17 F ...

[management.png](#)

...to transforming our economy, Slightly Below 3 1 5 8 4 fostering U.S. competitiven...
...Improved 1 1 development of new businesses (USPTO, EDA....
...Not Met 5 6 2 2 2 6 8 10 NIST, and NTIA) See Appendi A: Performance and Reso...
...includes all of the Bureau of Industry and Security (BIS), and portions of ITA. 17 F ...

[management.png](#)

...to transforming our economy, Slightly Below 3 1 5 8 4 fostering U.S. competitiven...
...Improved 1 1 development of new businesses (USPTO, EDA....

Amazon Kendra

Semantic search results

1-10 OF 83 RESULTS

Amazon Kendra suggested answers [More info](#)

management

The Trade Promotion goal accounted for 13 percent of FTE and 13 percent of the theme funding This goal includes all of the Bureau of Industry and Security (BIS), and portions of ITA. 17 F Y o P E R F O N C E N A C C O U N A B I L T O T

[management](#)

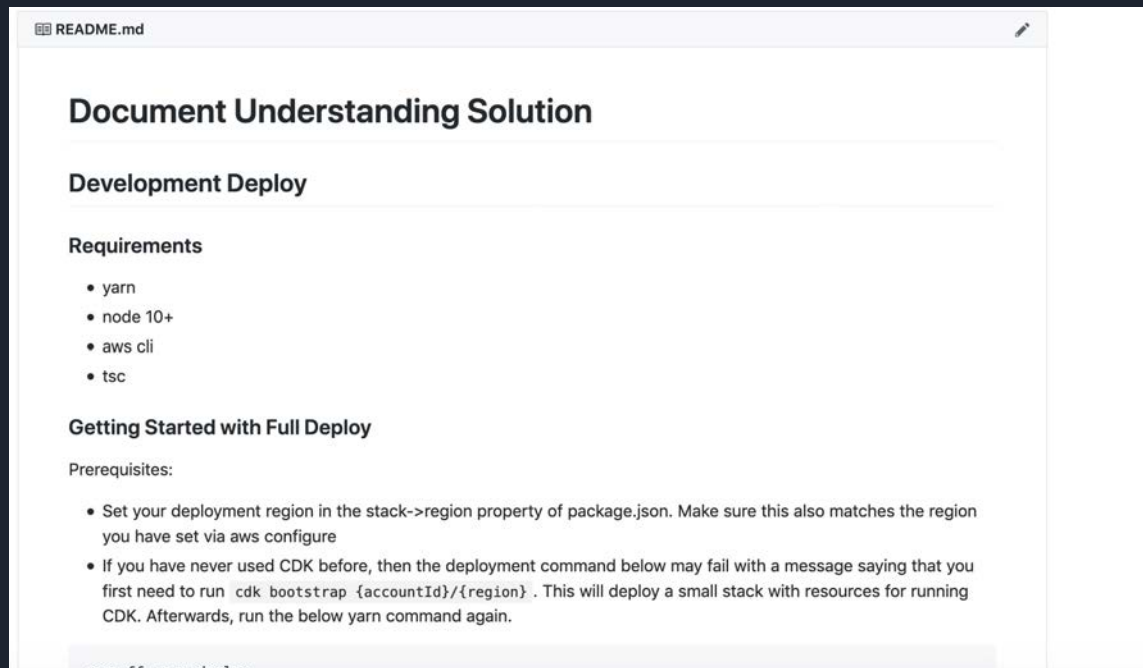
...the three goals in terms of FTE and funding accounted for 3 percent of FTE and 9 percent of the theme funding. This goa includes all of the Minority Business Development Agency (MBDA) and portions of EDA the International Trade Administration (ITA), and NIST. The Trade Promotion goal accounted...

Technical Details

- Follows the Well Architected Framework
- Can process over hundreds of thousands of documents
- Open source
- Customizable for use-case specific needs
- Built on popular technologies like CDK and React making it easier to extend the solution

Installation Guide

<https://github.com/awslabs/document-understanding-solution>



The screenshot shows a GitHub README file for the 'Document Understanding Solution'. The title is 'Document Understanding Solution'. Below the title, there are sections for 'Development Deploy', 'Requirements', and 'Getting Started with Full Deploy'. The 'Requirements' section lists: yarn, node 10+, aws cli, and tsc. The 'Getting Started with Full Deploy' section includes a 'Prerequisites:' list with two items: setting the deployment region in package.json and running a bootstrap command if using CDK for the first time.

README.md

Document Understanding Solution

Development Deploy

Requirements

- yarn
- node 10+
- aws cli
- tsc

Getting Started with Full Deploy

Prerequisites:

- Set your deployment region in the stack->region property of package.json. Make sure this also matches the region you have set via aws configure
- If you have never used CDK before, then the deployment command below may fail with a message saying that you first need to run `cdk bootstrap {accountId}/{region}`. This will deploy a small stack with resources for running CDK. Afterwards, run the below yarn command again.

Getting Started with DUS

- The Demo can be downloaded here:
<https://github.com/awslabs/document-understanding-solution>
- Free tier available
- Business Development and Solution Architects can help with integration

Certifications

- Amazon Textract

- PCI
- ISO
- HIPAA BAA

- Amazon Comprehend Medical

- SOC 1, 2, 3
- ISO
- HIPAA BAA
- HITRUST CSF

- Amazon Kendra

- HIPAA
- ISO, SOC coming Q1-21

- Amazon Comprehend

- SOC 1, 2, 3
- ISO
- FedRAMP Moderate (East/West)
- FedRAMP High (GovCloud)
- DoD CC SRG IL2 (East/West)
- DoD CC SRG IL2 (GovCloud)
- HIPAA BAA
- MTCS (Singapore)
- C5
- ENS High

Thank you!

ml.aws