Amazon Textract

Specify and extract information using the Queries feature in Amazon Textract

Navneeth Nair

Senior Product Manager, Amazon Textract
AWS

Martin Schade

Senior ML Product SA, Amazon Textract
AWS

Agenda

Overview of AI/ML at AWS

Issues with legacy solutions for document processing

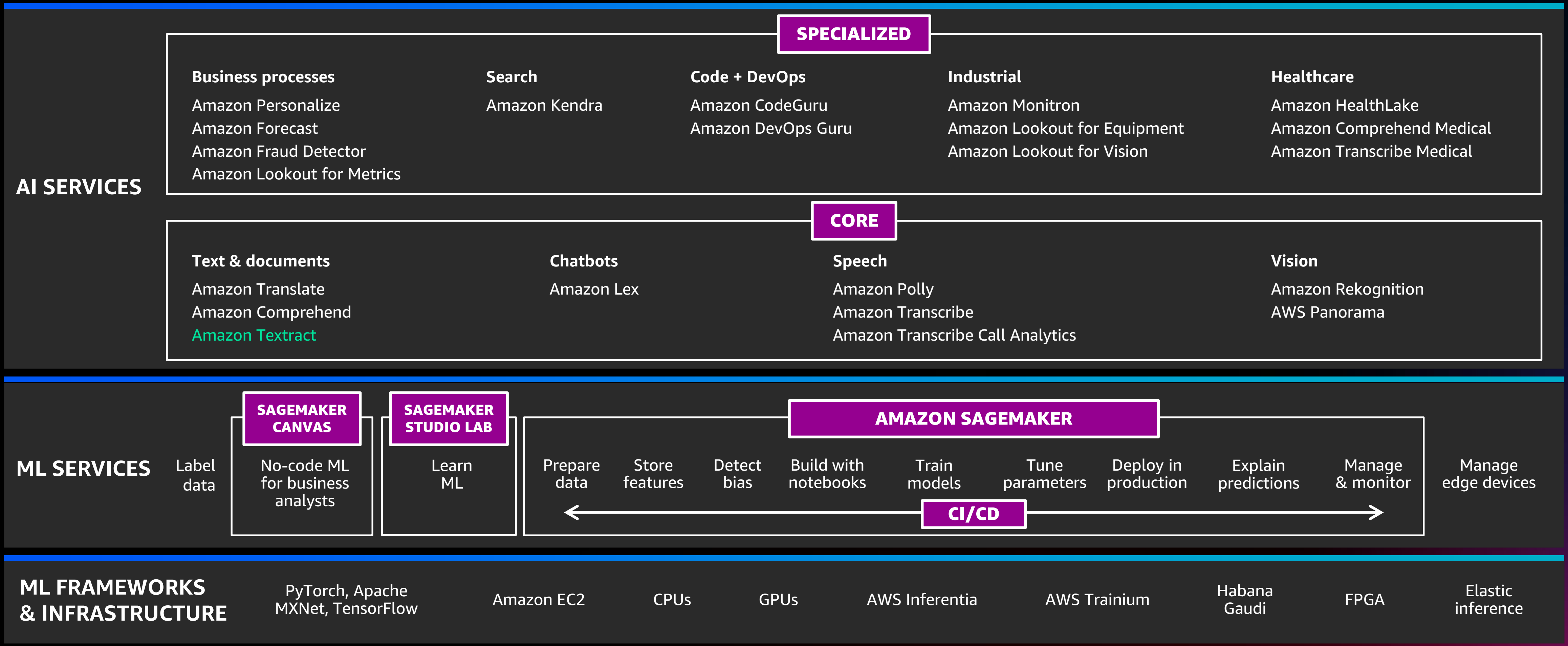
Amazon Textract’s current capabilities

Textract Analyze Document – “Queries”

Demo: Examples of automated data extraction from mortgage and insurance industry documents

The AWS AI/ML stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES



AI SERVICES

SPECIALIZED

Business processes
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

Search
- Amazon Kendra

Code + DevOps
- Amazon CodeGuru
- Amazon DevOps Guru

Industrial
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

Healthcare
- Amazon HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical

CORE

Text & documents
- Amazon Translate
- Amazon Comprehend
- Amazon Textract

Chatbots
- Amazon Lex

Speech
- Amazon Polly
- Amazon Transcribe
- Amazon Transcribe Call Analytics

Vision
- Amazon Rekognition
- AWS Panorama

ML SERVICES

Label data

SAGEMAKER CANVAS
No-code ML for business analysts

SAGEMAKER STUDIO LAB
Learn ML

AMAZON SAGEMAKER
- Prepare data
- Store features
- Detect bias
- Build with notebooks
- Train models
- Tune parameters
- Deploy in production
- Explain predictions
- Manage & monitor

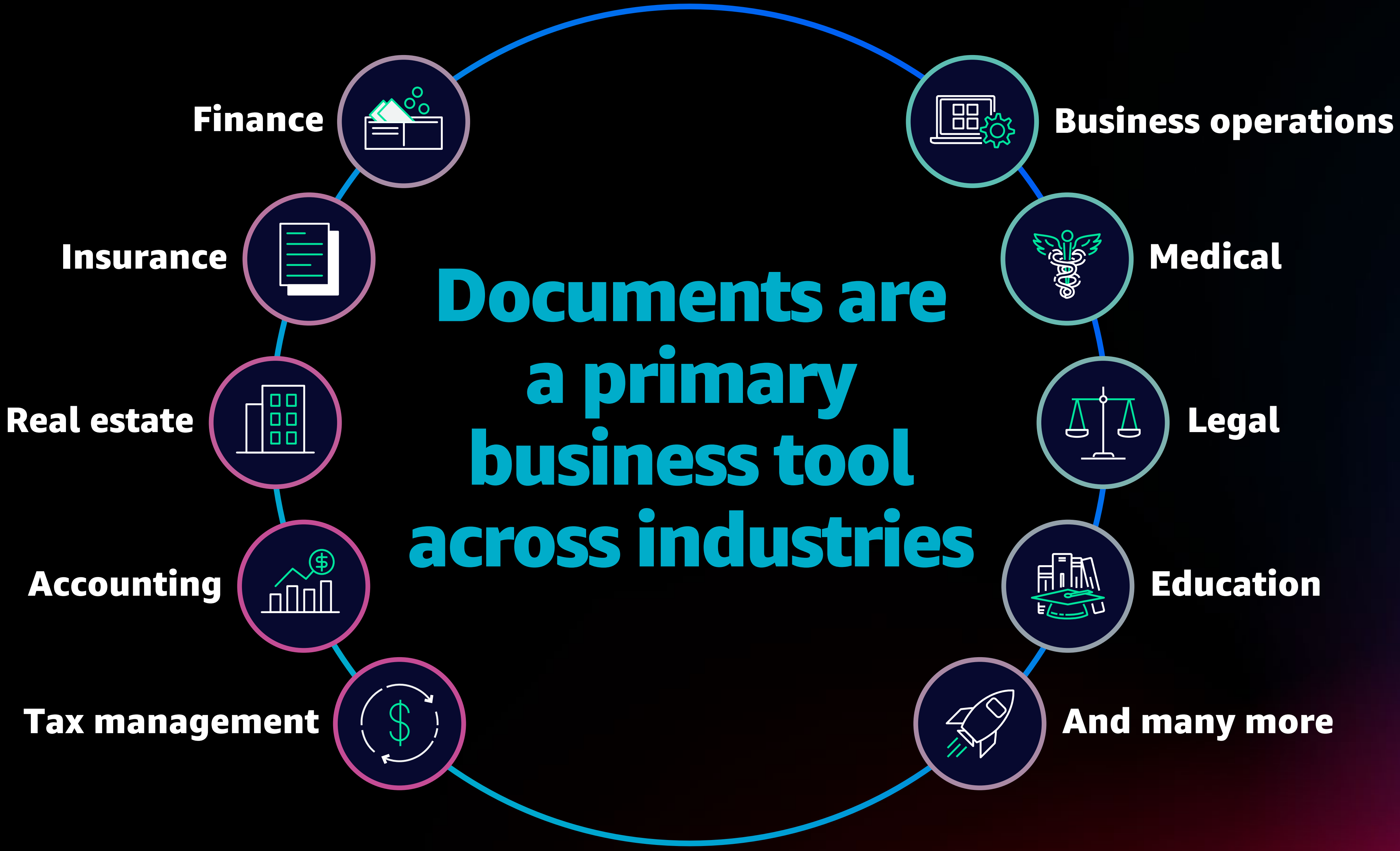
CI/CD

Manage edge devices

ML FRAMEWORKS & INFRASTRUCTURE

- PyTorch, Apache MXNet, TensorFlow
- Amazon EC2
- CPUs
- GPUs
- AWS Inferentia
- AWS Trainium
- Habana Gaudi
- FPGA
- Elastic inference

Documents are a primary business tool across industries



Finance

Insurance

Real estate

Accounting

Tax management

Business operations

Medical

Legal

Education

And many more

Legacy document processes do not meet today’s needs

Legacy OCR and manual processes are time-consuming, error-prone, and expensive

Manual processes do not scale easily with document volume

Difficult to find useful information needed for business decisions

Faster document processing shortens decision cycles so you can serve more customers and have people do higher-value tasks

Amazon Textract

Go beyond OCR with accurate, versatile information extraction

Amazon Textract

Automatically extract printed text, handwriting, and data from any document

No ML experience required

Reduce manual effort

Lower costs

Key verticals and business outcomes

Financial services

Accurately extract data and insights from pay stubs, loan applications, tax forms, driver's licenses, invoices, titles, etc.

Manufacturing and retail

Analyze invoices, bills of materials, contracts, licenses, warranties, and other agreements

Healthcare and life sciences

Process insurance claims and quickly get to clinical insights faster from healthcare and patient intake forms

Public sector

Improve speed and quality of services by gathering data from healthcare records, government forms, public records, taxes, securities filings, and unstructured documents

→

Productivity

Improve accuracy and process velocity and reduce costs

End-user experience

Automate user data capture and provide a better end-user experience through convenience, speed of service, and improved availability

Decision-making

Improve decision-making through richer inferences to scale business processes more effectively

aws

Improving loan underwriting process productivity

Problem

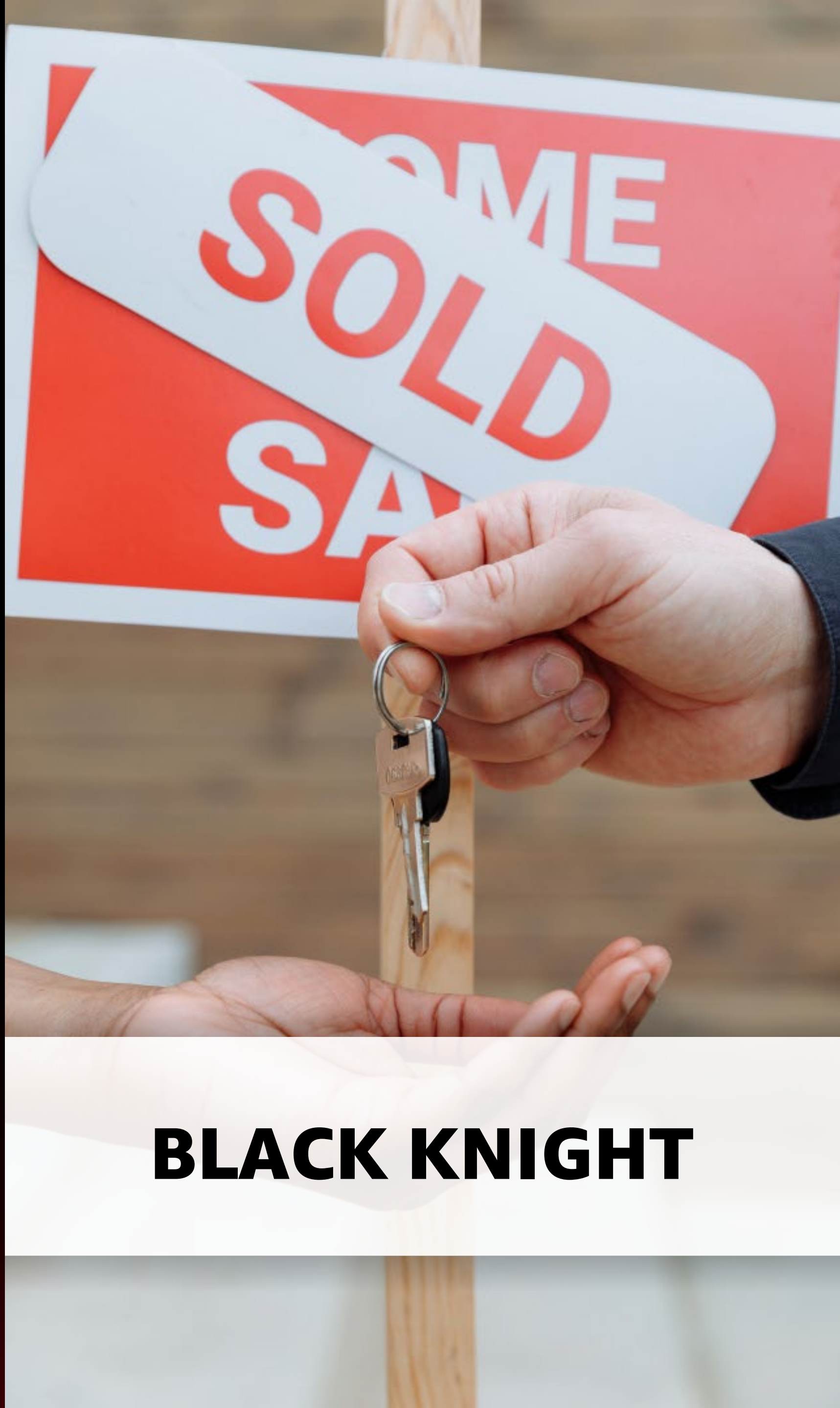
Incomplete loan packages, missing data, and the discovery of additional documentation requirements during the underwriting process often creates more work and slows the process down, resulting in fewer loans reviewed by underwriters and added lender costs

Solution

Black Knight developed Underwriter Assist (UA), an AI-powered mortgage solution that leverages Amazon Textract and other machine learning technologies; UA performs document recognition, data extraction, and analysis of income, assets, and property appraisal

Impact

Using advanced machine learning services like Amazon Textract, Black Knight has increased lender productivity, driven operational efficiency, and reduced costs, resulting in higher customer satisfaction



BLACK KNIGHT

Scaling claims processing through automation

Problem

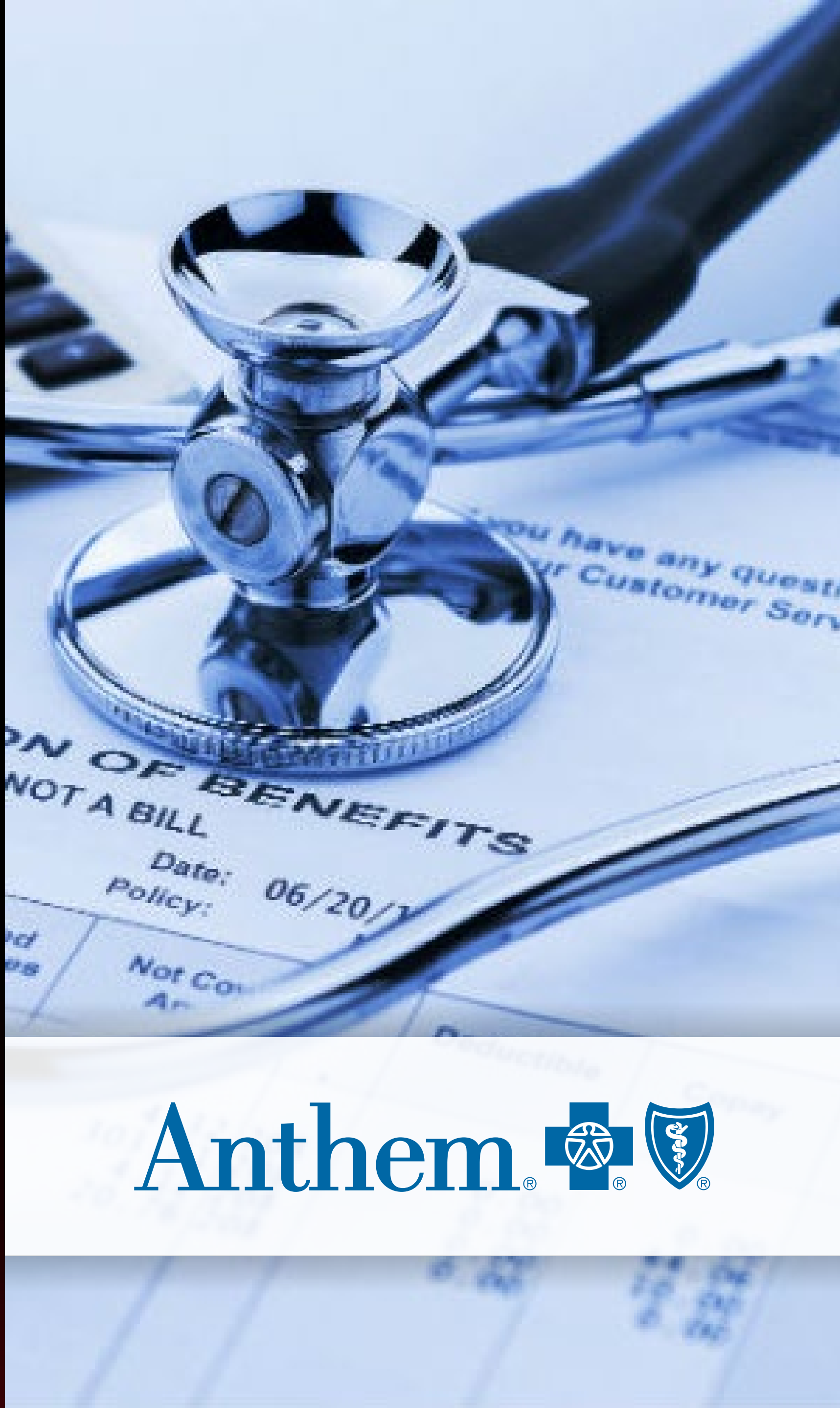
Health insurance companies spend millions of dollars to extract sensitive information from claims forms and accompanying attachments to perform their business operations; to reduce manual labor, the company wanted to automate the process

Solution

Anthem chose Amazon Textract to digitize and automate its claims process, for its image-processing capability, ability to detect tables and forms, and adherence to security and compliance standards

Impact

Using Amazon Textract, Anthem can extract and digitize data to quickly process thousands of claims each day; having already automated 80% of their claim processing workflow, Anthem expects to reach 90% automation levels or higher on AWS



Amazon Textract customers

3M	Anthem	ARQ	Assent	AXRAIL	bakertilly
BLACK KNIGHT	BDO	beyond Lucid technologies	BlueVine	Broadridge	Canara HSBC OBC Life Insurance
CHANGE HEALTHCARE	CHISEL	COX AUTOMOTIVE	THE GLOBE AND MAIL*	healthfirst	HELLOSIGN
HIVEO	IHS Markit	intuit	LUMIQ	Met Office	MOODY'S ANALYTICS
NHS	PennyMac	pfs TECH	QL Resources Berhad	Roche	rekeep
salesforce	TC Energy	Wrapped Insurance	FRED HUTCH	LexisNexis	FINRA

Amazon Textract:

Generally available features overview

Amazon Textract capabilities

Text

Loan application

Forms

Invoices and receipts

Identity documents

Specialized documents

Handwriting

Tables

Earnings

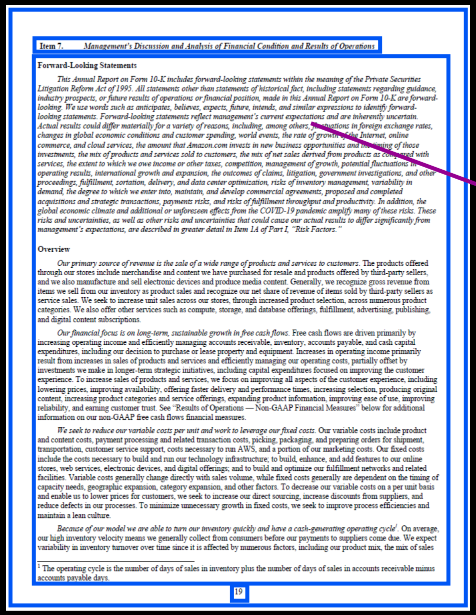
Paystubs

Queries

Text extraction

OPTIMAL FOR DENSE TEXT EXTRACTION WITH INDUSTRY-LEADING OCR ACCURACY

Document



Item 7. *Management's Discussion and Analysis of Financial Condition and Results of Operations*

Forward-Looking Statements

This Annual Report on Form 10-K includes forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995. All statements other than statements of historical fact, including statements regarding guidance, industry prospects, or future results of operations or financial position, made in this Annual Report on Form 10-K are forward-looking. We use words such as anticipates, believes, expects, future, intends, and similar expressions to identify forward-looking statements. Forward-looking statements reflect management's current expectations and are inherently uncertain. Actual results could differ materially for a variety of reasons, including, among others, fluctuations in foreign exchange rates, changes in global economic conditions and customer spending, world events, the rate of growth of the Internet, online commerce, and cloud services, the amount that Amazon.com invests in new business opportunities and the timing of those investments, the mix of products and services sold to customers, the mix of net sales derived from products as compared with services, the extent to which we owe income or other taxes, competition, management of growth, potential fluctuations in operating results, international growth and expansion, the outcomes of claims, litigation, government investigations, and other proceedings, fulfillment, sortation, delivery, and data center optimization, risks of inventory management, variability in demand, the degree to which we enter into, maintain, and develop commercial agreements, proposed and completed acquisitions and strategic transactions, payments risks, and risks of fulfillment throughput and productivity. In addition, the global economic climate and additional or unforeseen effects from the COVID-19 pandemic amplify many of these risks. These risks and uncertainties, as well as other risks and uncertainties that could cause our actual results to differ significantly from management's expectations, are described in greater detail in Item 1A of Part I, "Risk Factors."

Overview

Our primary source of revenue is the sale of a wide range of products and services to customers. The products offered through our stores include merchandise and content we have purchased for resale and products offered by third-party sellers, and we also manufacture and sell electronic devices and produce media content. Generally, we recognize gross revenue from items we sell from our inventory as product sales and recognize our net share of revenue of items sold by third-party sellers as service sales. We seek to increase unit sales across our stores, through increased product selection, across numerous product categories. We also offer other services such as compute, storage, and database offerings, fulfillment, advertising, publishing, and digital content subscriptions.

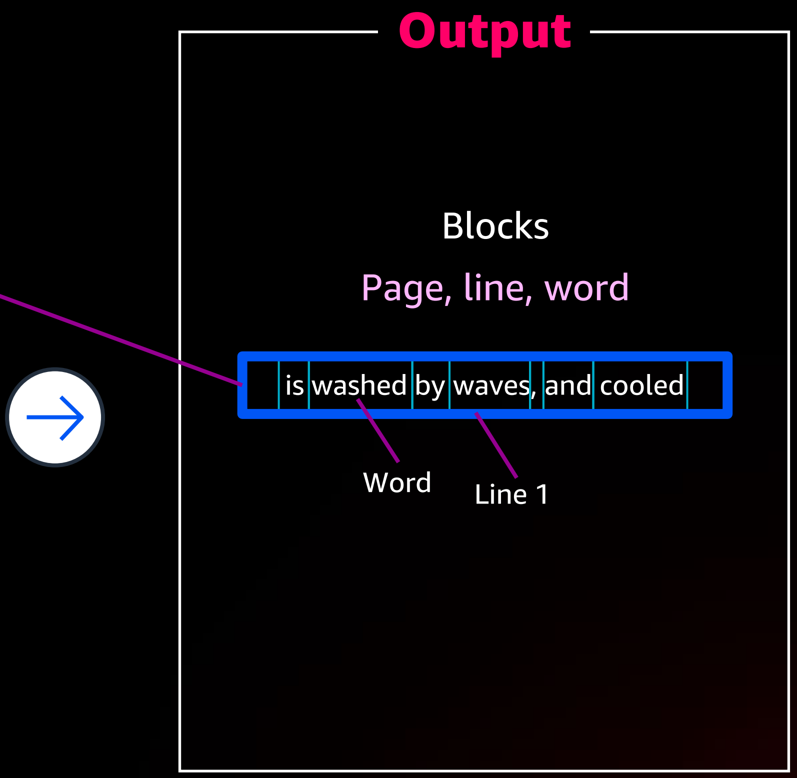
Our financial focus is on long-term, sustainable growth in free cash flows. Free cash flows are driven primarily by increasing operating income and efficiently managing accounts receivable, inventory, accounts payable, and cash capital expenditures, including our decision to purchase or lease property and equipment. Increases in operating income primarily result from increases in sales of products and services and efficiently managing our operating costs, partially offset by investments we make in longer-term strategic initiatives, including capital expenditures focused on improving the customer experience. To increase sales of products and services, we focus on improving all aspects of the customer experience, including lowering prices, improving availability, offering faster delivery and performance times, increasing selection, producing original content, increasing product categories and service offerings, expanding product information, improving ease of use, improving reliability, and earning customer trust. See "Results of Operations — Non-GAAP Financial Measures" below for additional information on our non-GAAP free cash flows financial measures.

We seek to reduce our variable costs per unit and work to leverage our fixed costs. Our variable costs include product and content costs, payment processing and related transaction costs, picking, packaging, and preparing orders for shipment, transportation, customer service support, costs necessary to run AWS, and a portion of our marketing costs. Our fixed costs include the costs necessary to build and run our technology infrastructure; to build, enhance, and add features to our online stores, web services, electronic devices, and digital offerings; and to build and optimize our fulfillment networks and related facilities. Variable costs generally change directly with sales volume, while fixed costs generally are dependent on the timing of capacity needs, geographic expansion, category expansion, and other factors. To decrease our variable costs on a per unit basis and enable us to lower prices for customers, we seek to increase our direct sourcing, increase discounts from suppliers, and reduce defects in our processes. To minimize unnecessary growth in fixed costs, we seek to improve process efficiencies and maintain a lean culture.

Because of our model we are able to turn our inventory quickly and have a cash-generating operating cycle[1]. On average, our high inventory velocity means we generally collect from consumers before our payments to suppliers come due. We expect variability in inventory turnover over time since it is affected by numerous factors, including our product mix, the mix of sales

[1] The operating cycle is the number of days of sales in inventory plus the number of days of sales in accounts receivable minus accounts payable days.

19



✓ Outputs detected text in 3 hierarchy blocks: page, line, and word

✓ Bounding box for each line and word provide visual cues for post-processing

✓ Included confidence scores enable informed decision-making for your workflows

Table extraction

EXTRACT TABLES FROM DOCUMENTS WHILE PRESERVING DATA STRUCTURE AND RELATIONSHIPS

Document

Previous employment history				
Start date	End date	Employer name	Position held	Reason for leaving
1/15/2009	6/30/2013	Any company	Head Baker	Family relocated
8/15/2013	Present	Example corp.	Baker	N/A, current employer

→

Output

Blocks

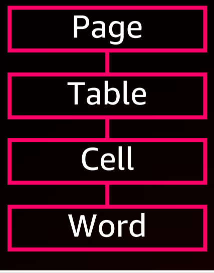
Page, table, cell

For each block

Text

Confidence score

Block relationships
(for example, cells within a table)



✓ Outputs recognized tables with relationships data intact

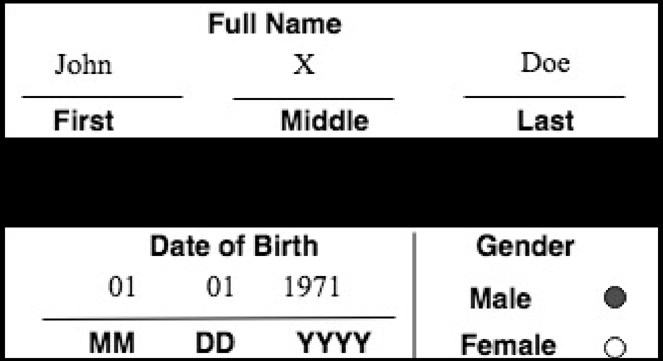
✓ Intelligently groups cells within tables and words within each cell

✓ Output also includes confidence scores, geometry info, and row/column indexes

Form extraction

EXTRACT FORM DATA FROM DOCUMENTS AS KEY-VALUE PAIRS TO PRESERVE STRUCTURE

Document



Full Name

John	X	Doe
First	Middle	Last

Date of Birth			Gender	
01	01	1971	Male	●
MM	DD	YYYY	Female	○

Output

Blocks

Page, key-value set

Example output

First: John

Middle: X

Last: Doe

MM: 01
DD: 01
YYYY: 1971

Male: True
Female: False

✓ Outputs form field name (key) and field value name (value) with relationship data intact

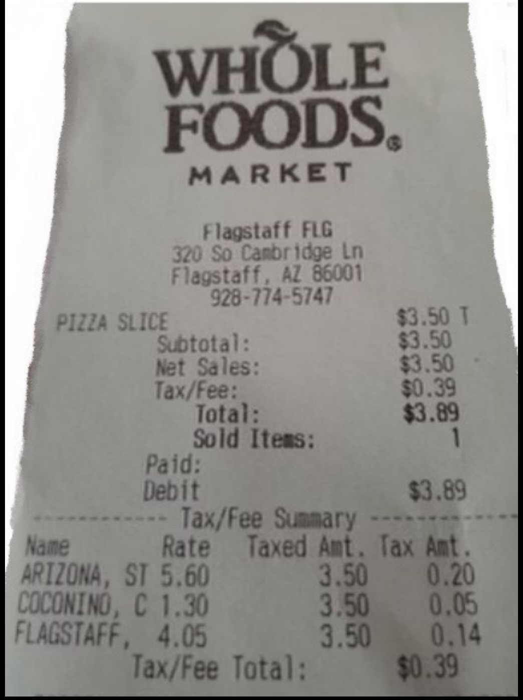
✓ Captures logical groupings, relationships, and glyphs

✓ Output also includes confidence scores and geometry info

Invoices and receipts

SPECIALIZED SUPPORT TO PROCESS INVOICES AND RECEIPTS AT SCALE

Document



WHOLE FOODS MARKET

Flagstaff FLG
320 So Cambridge Ln
Flagstaff, AZ 86001
928-774-5747

PIZZA SLICE	$3.50 T
Subtotal:	$3.50
Net Sales:	$3.50
Tax/Fee:	$0.39
Total:	$3.89
Sold Items:	1
Paid:	
Debit	$3.89

------------ Tax/Fee Summary ------------

Name	Rate	Taxed Amt.	Tax Amt.
ARIZONA, ST	5.60	3.50	0.20
COCONINO, C	1.30	3.50	0.05
FLAGSTAFF,	4.05	3.50	0.14
Tax/Fee Total:			$0.39

Output

Summary fields

Vendor Name: WHOLE FOODS MARKET
Subtotal (SUBTOTAL): $3.50
Net Sales (OTHER): $3.50
Tax/Fee (TAX): $0.39
Sold Items (OTHER): 1
Paid (OTHER):
Debit (OTHER): $3.89
Tax/Fee Total (TAX): $0.39
Total (TOTAL): $3.89

Line items

ITEM: Pizza Slice
PRICE: $3.50

✓ Outputs headline amounts, line item details, and inferred fields (like Vendor Name)

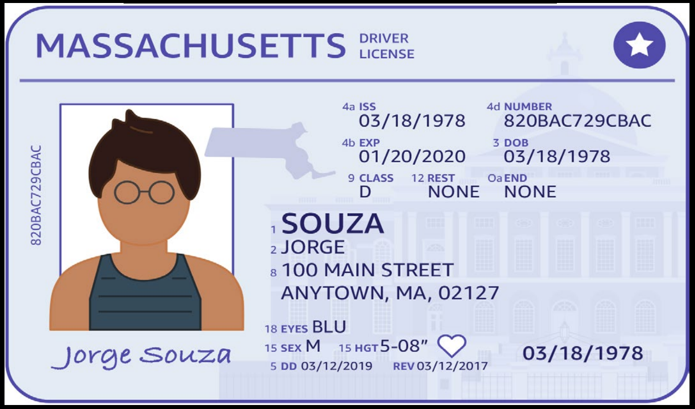
✓ Supports any style of invoice or receipt

✓ No templates or configuration required

Identity documents

SPECIALIZED SUPPORT FOR IDENTITY DOCUMENTS

Document



MASSACHUSETTS DRIVER LICENSE

820BAC729CBAC

4a ISS 03/18/1978

4d NUMBER 820BAC729CBAC

4b EXP 01/20/2020

3 DOB 03/18/1978

9 CLASS D

12 REST NONE

Oa END NONE

1 SOUZA

2 JORGE

8 100 MAIN STREET
ANYTOWN, MA, 02127

18 EYES BLU

15 SEX M

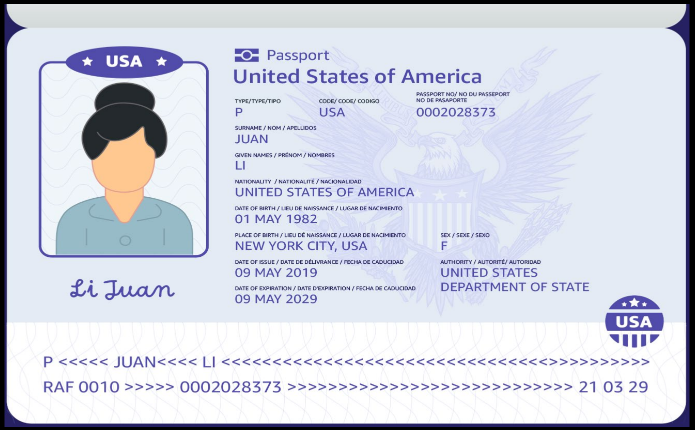
15 HGT 5-08"

5 DD 03/12/2019

REV 03/12/2017

03/18/1978

Jorge Souza



USA

Passport

United States of America

TYPE/TYPE/TIPO P

CODE/ CODE/ CODIGO USA

PASSPORT NO/ NO DU PASSEPORT NO DE PASAPORTE 0002028373

SURNAME / NOM / APELLIDOS JUAN

GIVEN NAMES / PRÉNOM / NOMBRES LI

NATIONALITY / NATIONALITÉ / NACIONALIDAD UNITED STATES OF AMERICA

DATE OF BIRTH / LIEU DE NAISSANCE / LUGAR DE NACIMIENTO 01 MAY 1982

PLACE OF BIRTH / LIEU DE NAISSANCE / LUGAR DE NACIMIENTO NEW YORK CITY, USA

SEX / SEXE / SEXO F

DATE OF ISSUE / DATE DE DÉLIVRANCE / FECHA DE CADUCIDAD 09 MAY 2019

AUTHORITY / AUTORITÉ/ AUTORIDAD UNITED STATES DEPARTMENT OF STATE

DATE OF EXPIRATION / DATE D'EXPIRATION / FECHA DE CADUCIDAD 09 MAY 2029

Li Juan

USA

P <<<<< JUAN<<<< LI <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<>>>>>>>>>

RAF 0010 >>>>> 0002028373 >>>>>>>>>>>>>>>>>>>>>>>>>>> 21 03 29

→

Output

First Name: JORGE
Last Name: SOUZA
Middle Name:
Address Line1: 100 MAIN STREET
Address Line 2:
City: ANYTOWN
State: MA
Document Number: 820BAC729CBAC
Expiration Date: 01/20/2020
Date of Birth: 03/18/1978
ID Type: Driver License
Date of Issue: 03/18/1978
Issued By: MASSACHUSETTS
Class: D
Restrictions: NONE
Endorsements: NONE

✓ Over 95% accuracy for US driver licenses and passports

✓ No templates or configuration required

✓ Outputs normalized field names and supports implied elements

Textract Queries

Data extraction challenges

Data Variations

(SSN vs. Social Security Number vs. Tax ID)

Post processing

Expensive

Time consuming

Data Structure

(Table, Form, Implied Fields)

Post processing

Custom models

Manual

Nested Data

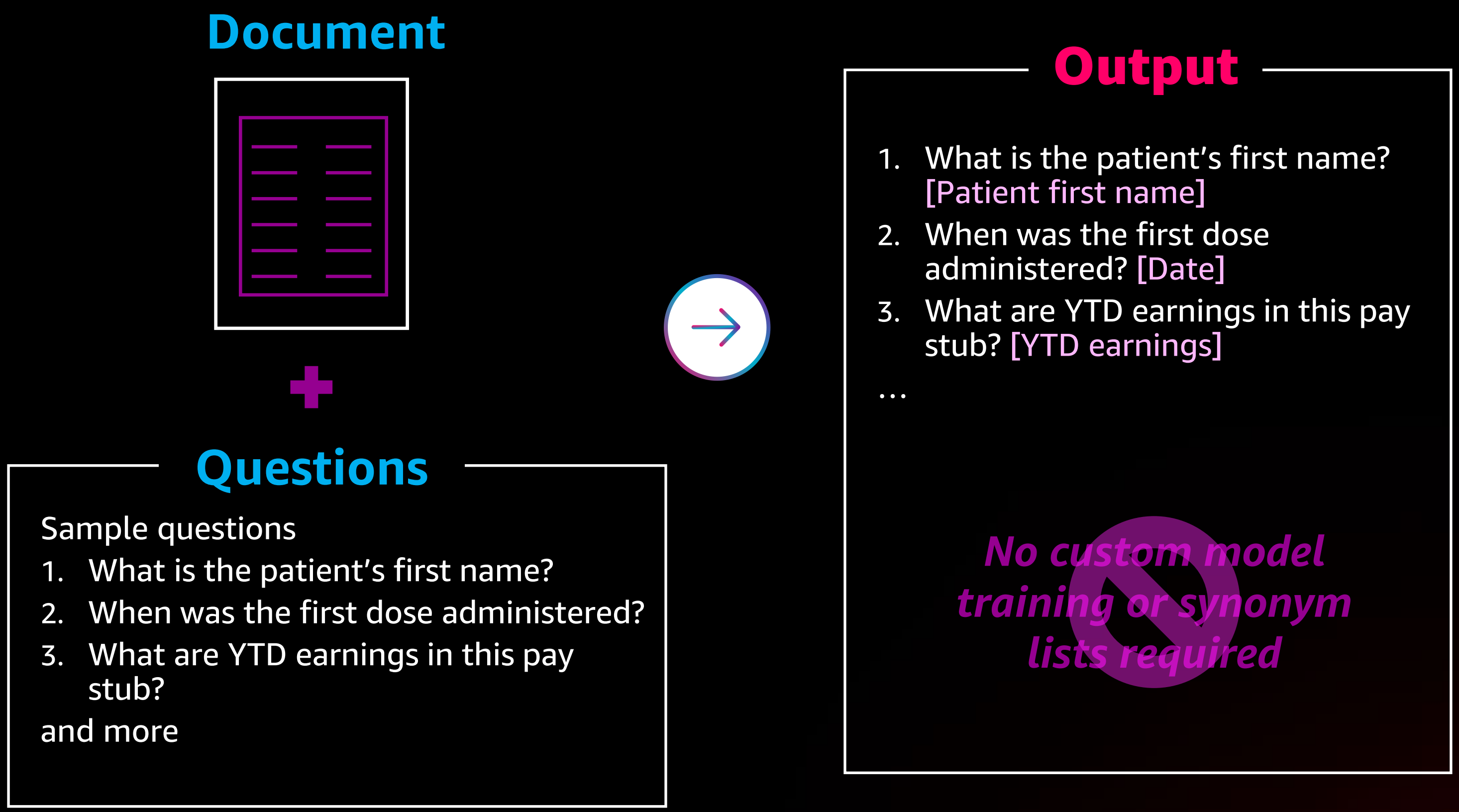
(page sections, duplicate fields)

Development and management overhead

Templates are brittle

Textract queries

EASILY SPECIFY AND EXTRACT VALUABLE PIECES OF INFORMATION FROM DOCUMENTS



Document

+

Questions

Sample questions

1. What is the patient's first name?
2. When was the first dose administered?
3. What are YTD earnings in this pay stub?

and more

Output

1. What is the patient's first name? [Patient first name]
2. When was the first dose administered? [Date]
3. What are YTD earnings in this pay stub? [YTD earnings]

…

No custom model training or synonym lists required

✓ Combines visual, spatial, and textual cues to provide higher accuracy

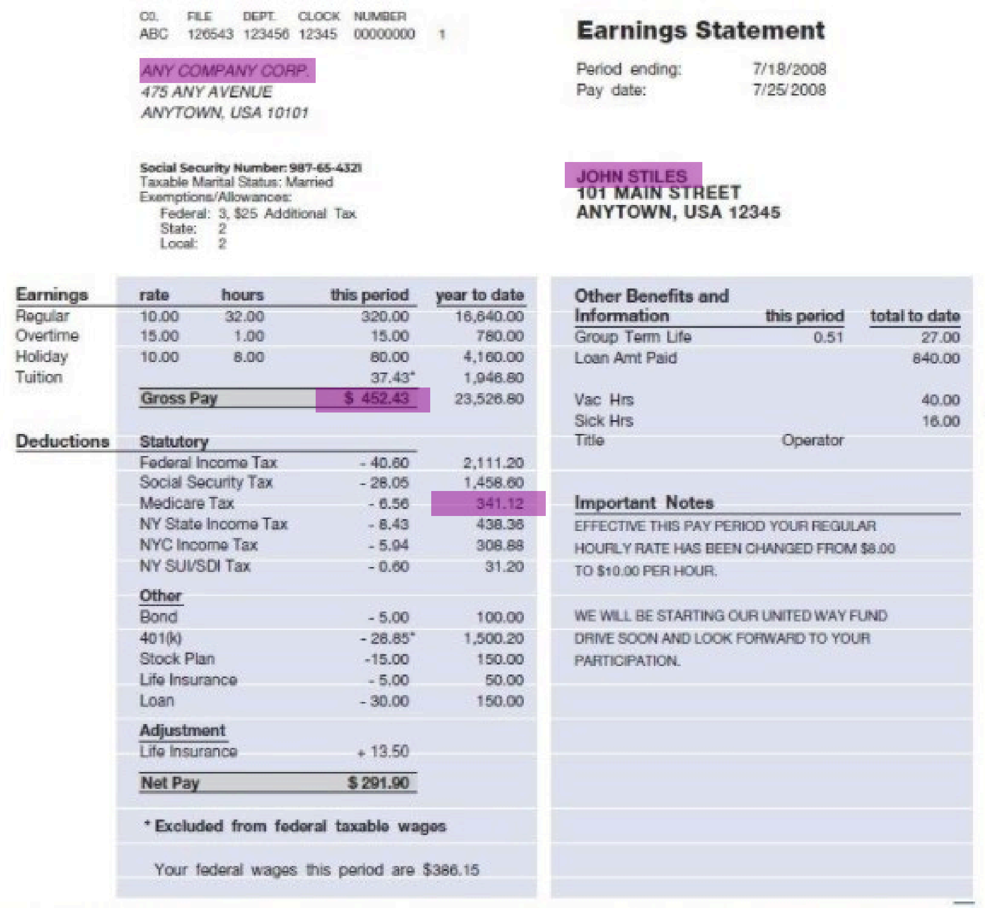
✓ Simple Q&A response reduces post-processing

✓ Extract only the specific fields that matter

✓ Easy integration with existing Amazon Textract Analyze Document API

Example: Amazon Textract Queries on pay stubs

Pay stubs



CO.	FILE	DEPT.	CLOCK	NUMBER	
ABC	126543	123456	12345	0000000	1

ANY COMPANY CORP.
475 ANY AVENUE
ANYTOWN, USA 10101

Earnings Statement

Period ending:	7/18/2008
Pay date:	7/25/2008

Social Security Number: 987-65-4321
Taxable Marital Status: Married
Exemptions/Allowances:
Federal: 3, $25 Additional Tax
State: 2
Local: 2

JOHN STILES
101 MAIN STREET
ANYTOWN, USA 12345

Earnings	rate	hours	this period	year to date
Regular	10.00	32.00	320.00	16,640.00
Overtime	15.00	1.00	15.00	780.00
Holiday	10.00	8.00	80.00	4,160.00
Tuition			37.43*	1,946.80
Gross Pay			$ 452.43	23,526.80

Deductions	**Statutory**		
	Federal Income Tax	- 40.60	2,111.20
	Social Security Tax	- 28.05	1,458.60
	Medicare Tax	- 6.56	341.12
	NY State Income Tax	- 8.43	438.36
	NYC Income Tax	- 5.94	308.88
	NY SUI/SDI Tax	- 0.60	31.20
	Other		
	Bond	- 5.00	100.00
	401(k)	- 28.85*	1,500.20
	Stock Plan	-15.00	150.00
	Life Insurance	- 5.00	50.00
	Loan	- 30.00	150.00
	Adjustment		
	Life Insurance	+ 13.50	
	Net Pay	**$ 291.90**	

*** Excluded from federal taxable wages**

Your federal wages this period are $386.15

Other Benefits and Information	this period	total to date
Group Term Life	0.51	27.00
Loan Amt Paid		840.00
Vac Hrs		40.00
Sick Hrs		16.00
Title	Operator	

Important Notes

EFFECTIVE THIS PAY PERIOD YOUR REGULAR HOURLY RATE HAS BEEN CHANGED FROM $8.00 TO $10.00 PER HOUR.

WE WILL BE STARTING OUR UNITED WAY FUND DRIVE SOON AND LOOK FORWARD TO YOUR PARTICIPATION.

Example – tabular data extraction

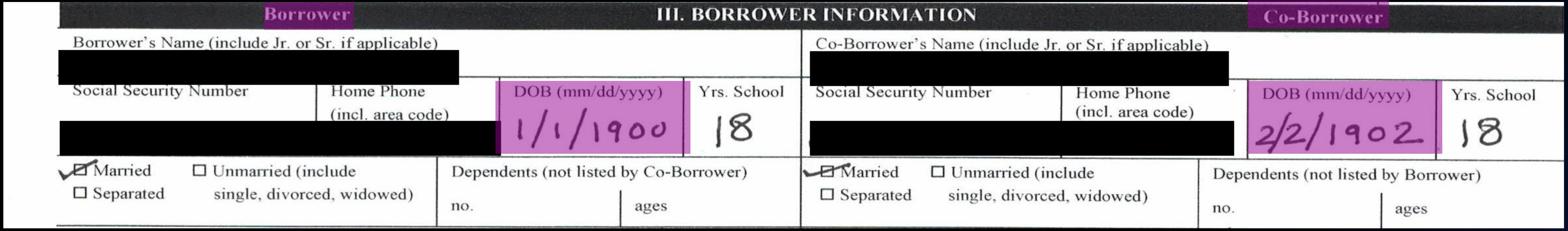
- What is the gross pay this period? $452.43
- What is the gross pay YTD? $23,526.80
- What is the Medicare tax YTD? $341.12

Example – implied fields

- What is the company name? ANY COMPANY CORP.
- What is the employee name? JOHN STILES

Example: Textract Queries on Fannie Mae Form 1003

Uniform Residential Loan Application (Form 1003)



Borrower | III. BORROWER INFORMATION | Co-Borrower

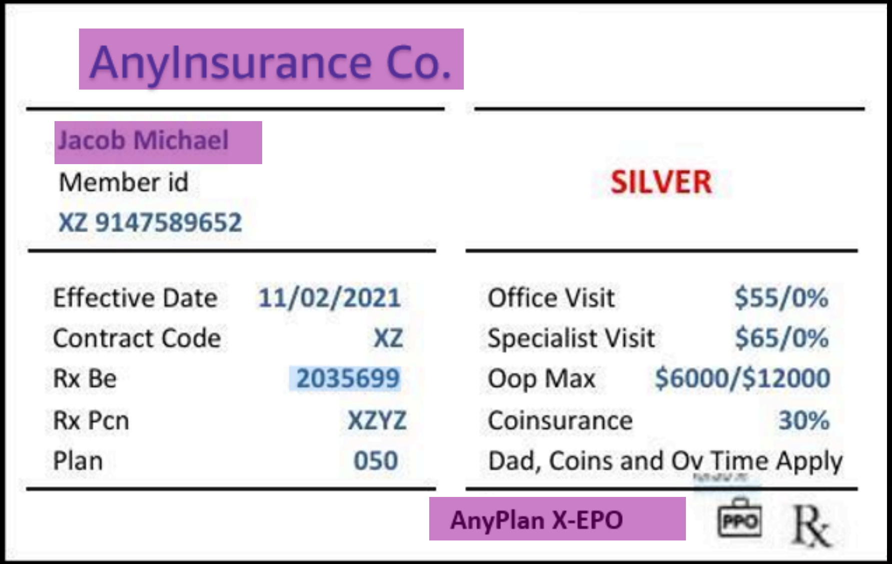
Borrower's Name (include Jr. or Sr. if applicable) [redacted]				Co-Borrower's Name (include Jr. or Sr. if applicable) [redacted]			
Social Security Number	Home Phone (incl. area code)	DOB (mm/dd/yyyy) 1/1/1900	Yrs. School 18	Social Security Number	Home Phone (incl. area code)	DOB (mm/dd/yyyy) 2/2/1902	Yrs. School 18
[redacted]				[redacted]			
☑ Married ☐ Separated	☐ Unmarried (include single, divorced, widowed)	Dependents (not listed by Co-Borrower) no.	ages	☑ Married ☐ Separated	☐ Unmarried (include single, divorced, widowed)	Dependents (not listed by Borrower) no.	ages

For example

- What is the borrower's date of birth? **01/01/1900**
- What is the co-borrower's date of birth? **02/02/1902**

Example: Textract Queries on health insurance cards

Health insurance card



AnyInsurance Co.

Jacob Michael
Member id
XZ 9147589652

SILVER

Effective Date	11/02/2021	Office Visit	$55/0%
Contract Code	XZ	Specialist Visit	$65/0%
Rx Be	2035699	Oop Max	$6000/$12000
Rx Pcn	XZYZ	Coinsurance	30%
Plan	050	Dad, Coins and Ov Time Apply	

AnyPlan X-EPO

PPO ℞

For example

- What is the member name? **Jacob Michael**
- What is the plan type? **AnyPlan X-EPO**
- What is the insurance provider? **AnyInsuranc Co.**

Demo

Resources

Amazon Textract service

- https://aws.amazon.com/textract/

Blogs

- Automatically extract text and structured data from documents with Amazon Textract – https://go.aws/3mmDXUI
- Specify and extract information from documents using the new Queries feature in Amazon Textract – https://go.aws/3xpBShn
- Announcing support for extracting data from identity documents using Amazon Textract – https://go.aws/3Q1eWvX
- Announcing specialized support for extracting data from invoices and receipts using Amazon Textract – https://go.aws/3miPzIn

GitHub repository

- Mortgage – https://bit.ly/3thI7Bs
- Insurance – https://bit.ly/3mqEV2d
- Vaccination cards – https://bit.ly/3aJJlPo
- Constructs Hub: https://constructs.dev/packages/amazon-textract-idp-cdk-constructs/
- GitHub Constructs: https://github.com/aws-samples/amazon-textract-idp-cdk-constructs/
- GitHub Samples: https://github.com/aws-samples/amazon-textract-idp-cdk-stack-samples

Thank you!