

AIM 202

# Automating document analysis and text extraction with Amazon Textract

Randall Hunt  
Senior technical evangelist and software engineer  
AWS

# Documents are important

Primary tool of record keeping, communicating, collaborating, and transacting



Finance



Insurance



Real estate



Accounting



Tax management



Medical



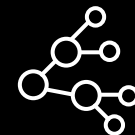
Legal



Business management



Education



And many more

# 16.3 million US mortgage applications (\$2.1 trillion) in 2016

## Uniform Residential Loan Application

This application is designed to be completed by the applicant(s) with the Lender's assistance. Applicants should complete this form as "Borrower" or "Co-Borrower," as applicable. Co-Borrower information must also be provided (and the appropriate box checked) when ☐ the income or assets of a person other than the Borrower (including the Borrower's spouse) will be used as a basis for loan qualification or ☐ the income or assets of the Borrower's spouse or other person who has community property rights pursuant to state law will not be used as a basis for loan qualification, but his or her liabilities must be considered because the spouse or other person has community property rights pursuant to applicable law and Borrower resides in a community property state, the security property is located in a community property state, or the Borrower is relying on other property located in a community property state as a basis for repayment of the loan.

If this is an application for joint credit, Borrower and Co-Borrower each agree that we intend to apply for joint credit (sign below):

Borrower		Co-Borrower	
<b>I. TYPE OF MORTGAGE AND TERMS OF LOAN</b>			
Mortgage Applied for:	<input type="checkbox"/> VA <input type="checkbox"/> FHA	<input type="checkbox"/> Conventional <input type="checkbox"/> USDA/Rural Housing Service	<input type="checkbox"/> Other (explain):
Agency Case Number		Lender Case Number	
Amount \$	Interest Rate %	No. of Months	Amortization Type: <input type="checkbox"/> Fixed Rate <input type="checkbox"/> GPM <input type="checkbox"/> Other (explain): <input type="checkbox"/> ARM (type):
<b>II. PROPERTY INFORMATION AND PURPOSE OF LOAN</b>			
Subject Property Address (street, city, state & ZIP)			No. of Units
Legal Description of Subject Property (attach description if necessary)			Year Built
Purpose of Loan	<input type="checkbox"/> Purchase <input type="checkbox"/> Refinance	<input type="checkbox"/> Construction <input type="checkbox"/> Construction-Permanent	<input type="checkbox"/> Other (explain):
Property will be:		<input type="checkbox"/> Primary Residence <input type="checkbox"/> Secondary Residence <input type="checkbox"/> Investment	

\*Mortgage Bankers Association 2016 HMDA

# About 240 million W-2 tax forms processed for FY 2018 in the US

22222		a Employee's social security number		OMB No. 1545-0008		
b Employer identification number (EIN)		1 Wages, tips, other compensation		2 Federal income tax withheld		
c Employer's name, address, and ZIP code		3 Social security wages		4 Social security tax withheld		
		5 Medicare wages and tips		6 Medicare tax withheld		
		7 Social security tips		8 Allocated tips		
d Control number		9 Verification code		10 Dependent care benefits		
e Employee's first name and initial      Last name      Suff.		11 Nonqualified plans		12a		
		13 Statutory employee      Retirement plan      Third-party sick pay		12b		
		14 Other		12c		
				12d		
f Employee's address and ZIP code						
15 State	Employer's state ID number	16 State wages, tips, etc.	17 State income tax	18 Local wages, tips, etc.	19 Local income tax	20 Locality name

Form **W-2** Wage and Tax Statement

2018

Copy 1—For State, City, or Local Tax Department

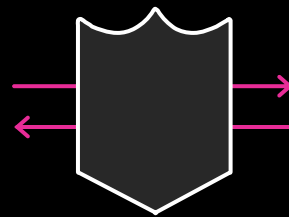
Department of the Treasury—Internal Revenue Service

\*IRS—<https://www.irs.gov/individuals/w-2-verification-code>

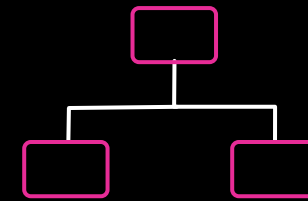
# Need for processing documents



Search  
and discovery

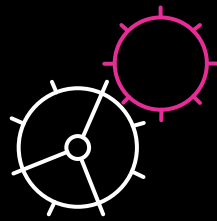


Compliance  
and control



Business  
process automation

# How documents are processed today



Manual  
processing

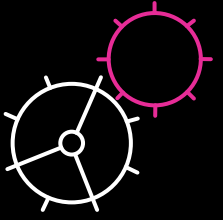


Optical character recognition  
(OCR)



Rules and  
template-based extraction

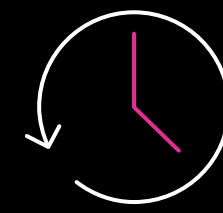
# Challenges for processing documents: Manual processing



Expensive

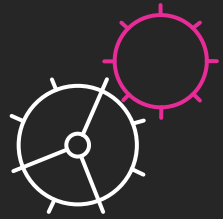


Error-prone



Time-consuming

# Challenges for processing documents: Manual processing



Employer's name – Nom de l'employeur  
AMAZON.CA

Canada Revenue Agency  
Agence du revenu du Canada

Year  
Année 2017

T4  
Statement of Remuneration Paid  
État de la rémunération payée

Employment income – line 101  
Revenu d'emploi – ligne 101 14 98031 39

Income tax deducted – line 437  
Impôt sur le revenu retenu – ligne 437 22 43,908 09

CPP/QPP EI PPIP  
28 ✓ ☐ ☐

RPC/RRQ AE RPAP

Other information (see over)  
Autres renseignements (voir au verso)

Box – Case Amount – Montant  
31 34900 00

Box – Case Amount – Montant  
32 382 00



Variable output



Inconsistent results



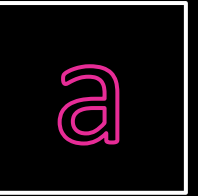
Needs multiple human reviews for consensus

## Output

1. Exempt is true
2. 28 is true
3. CPP/QPP is true
4. RPC/RRQ is true



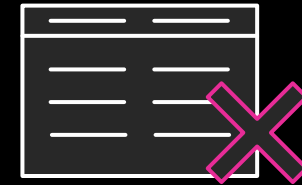
# Challenges for processing documents: OCR



Simple documents only



Error-prone



Flat bag of words

# Challenges for processing documents: OCR



## Extract data quickly & accurately

Textract makes it easy to quickly and accurately extract data from documents and forms. Textract automatically detects a document's layout and the key elements on the page, understands the data relationships in any embedded forms or tables, and extracts everything with its context intact. This means you can instantly use the extracted data in an application or store it in a database without a lot of complicated code in between

## No code or templates to maintain

Textract's pre-trained machine learning models eliminate the need to write code for data extraction, because they have already been trained on tens of millions of documents from virtually every industry, including invoices, receipts, contracts, tax documents, sales orders, enrollment forms, benefit applications, insurance claims, policy documents and many more. You no longer need to maintain code for every document or form you might receive or worry about how page layouts change over time.



No multi-column detection



No rotated text detection (not shown)

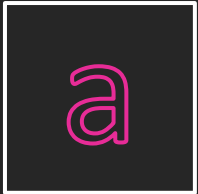


No stylized font detection (not shown)

## Output

Extract data quickly & No code or templates to accurately maintain

# Challenges for processing documents: OCR



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

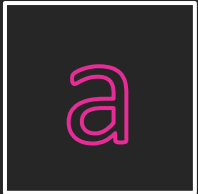


OCR reads left to right, ignoring table structure

## Output

Start Date End Date Employer Name Position Held Reason for leaving  
1/15/2009 6/30/2013 Any Company Head Baker Family relocated

# Challenges for processing documents



Full Name			Date of Birth			Gender	
John	X	Doe	01	01	1971	Male	<input checked="" type="radio"/>
First	Middle	Last	MM	DD	YYYY	Female	<input type="radio"/>

## Output

Full Name Date of Birth Gender  
John X Doe 01 01 1971  
Male  
First Middle Last MM DD YYYY  
Female



Logical groupings missed



Relationships missed

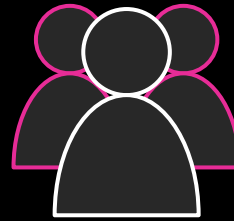


Glyphs missed

# Challenges for processing documents: Rules and template-based extraction



Limited by  
accuracy of OCR



Significant development and  
management overhead



Templates  
are brittle

# Challenges for processing documents: Rules and template-based extraction

The well-known W-2 US tax form has hundreds of variants each year

Form W2 Wage and Tax Statement 2017

d. control number 438209		Void	c. Employer's name, address, and ZIP		1 Wages, tips, other comp. \$39,010.32	2 Federal income tax withheld \$5,451.12
b Employer's FED ID number 32-3939201		a Employee's SSA number 111-22-3333		ANYCOMPANY 100 MAIN STREET ANYTOWN, USA	3 Social security wages \$39,010.32	4 Social security tax withheld \$2439.08
13 State employee	Retirement plan X	3rd party sick pay			5 Medicare wages and tips \$39,010.32	6 Medicare tax withheld \$456.90
12 See instrs. For Box 12 C \$209		14 Other			7 Social security tips	8 Allocated tips
15 State		Employer's State ID no.	16 State, wages, tips, etc.	17 State Income Tax	18 Local wages, tips, etc.	19 Local Income Tax
						20 Locality name

22222		a Employee's social security number		OMB No. 1545-0008	
b Employer identification number (EIN)		1 Wages, tips, other compensation		2 Federal income tax withheld	
c Employer's name, address, and ZIP code		3 Social security wages		4 Social security tax withheld	
		5 Medicare wages and tips		6 Medicare tax withheld	
		7 Social security tips		8 Allocated tips	
d Control number		9 Verification code		10 Dependent care benefits	
e Employee's first name and initial		Last name		11 Nonqualified plans	
				12a	
				12b	
				12c	
				12d	
f Employee's address and ZIP code		13 Statutory employee		Retirement plan	
		14 Other		Third-party sick pay	
15 State		Employer's state ID number	16 State, wages, tips, etc.	17 State income tax	18 Local wages, tips, etc.
				19 Local income tax	20 Locality name

Form **W-2** Wage and Tax Statement 2017  
Copy 1 - For State, City, or Local Tax Department

W2 **e file** 2018

d. control number 438209	Dept 8840	Corp.	Employer use only A 439
c. Employer's name, address, and ZIP code ANYCOMPANY 100 MAIN STREET ANYTOWN, USA			
e/f Employee's name, address, and ZIP code JANE DOE 123 ANY STREET ANY TOWN, USA			
b Employer's FED ID number 32-3939201	a Employee's SSA number 111-22-3333		
1 Wages, tips, other comp. \$39,010.32	2 Federal income tax withheld \$5,451.12		
3 Social security wages \$39,010.32	4 Social security tax withheld \$2439.08		
5 Medicare wages and tips \$39,010.32	6 Medicare tax withheld \$456.90		
7 Social security tips	8 Allocated tips		
9	10 Dependent care benefits		
11 Nonqualified plans	12a See instructions for box 12 C \$209		
14 Other	12b D \$395.16		
	12c		
	12d		
15 State NH	Employer's State ID no.		
17 State Income Tax	16 State, wages, tips, etc.		
19 Local Income Tax	18 Local wages, tips, etc.		
	20 Locality name		

a Employee's SSA number 111-22-3333	1 Wages, tips, other comp. \$39,010.32	2 Federal income tax withheld \$5,451.12
	3 Social security wages \$39,010.32	4 Social security tax withheld \$2439.08
b Employer's FED ID number 32-3939201	5 Medicare wages and tips \$39,010.32	6 Medicare tax withheld \$456.90
c. Employer's name, address, and ZIP code ANYCOMPANY 100 MAIN STREET ANYTOWN, USA		
e. Employee's first name and initial JANE		Last name DOE
		Suff.
123 ANY STREET ANY TOWN, USA		
f. Employee's address and ZIP code		
d. control number 438209	7 Social security tips	8 Allocated tips
9	10 Dependent care benefits	11 Nonqualified plans
12a See instructions for box 12 C \$209	14 Other	
12b D \$395.16		
12c		
12d		
13 Statutory employee <input type="checkbox"/> Retirement plan <input checked="" type="checkbox"/> Third-party sick pay <input type="checkbox"/>		
15 State NH	Employer's State ID no.	16 State, wages, tips, etc.
17 State Income Tax	18 Local wages, tips, etc.	
19 Local Income Tax	20 Locality name	

W2 Wage and Tax Statement 2017  
Copy B Employee Reference Copy

# It looks easy, but ...

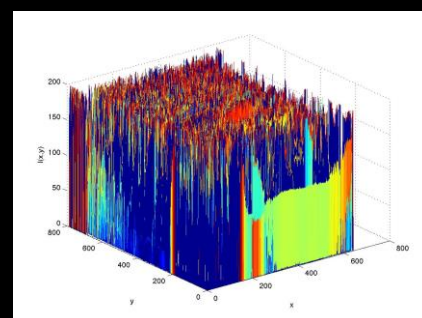
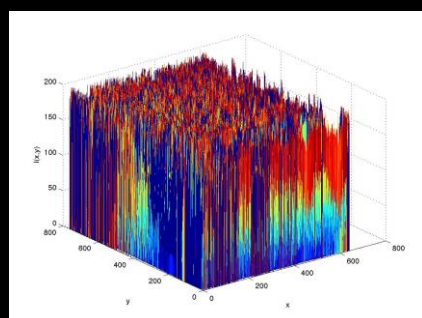
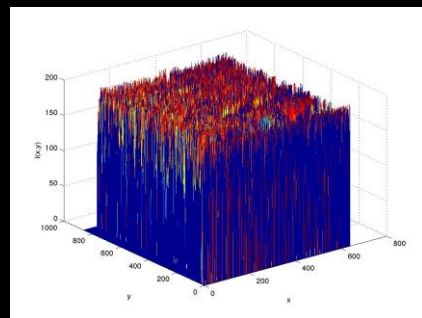
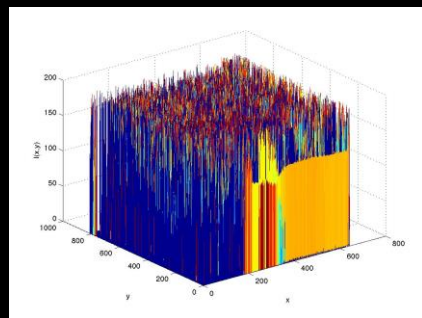
Form W-9, Request for Taxpayer Identification Number and Certification. The form is filled out with handwritten information. The taxpayer is identified as 'J. A. Smith' with a TIN of '12-34-567890'. The form includes sections for certification, general instructions, and a signature block.

Form W-9, Request for Taxpayer Identification Number and Certification. The form is filled out with handwritten information. The taxpayer is identified as 'J. A. Smith' with a TIN of '12-34-567890'. The form includes sections for certification, general instructions, and a signature block.

Form W-9, Request for Taxpayer Identification Number and Certification. The form is filled out with handwritten information. The taxpayer is identified as 'J. A. Smith' with a TIN of '12-34-567890'. The form includes sections for certification, general instructions, and a signature block.

Form W-9, Request for Taxpayer Identification Number and Certification. The form is filled out with handwritten information. The taxpayer is identified as 'J. A. Smith' with a TIN of '12-34-567890'. The form includes sections for certification, general instructions, and a signature block.

# ...not a single corresponding pixel value in common

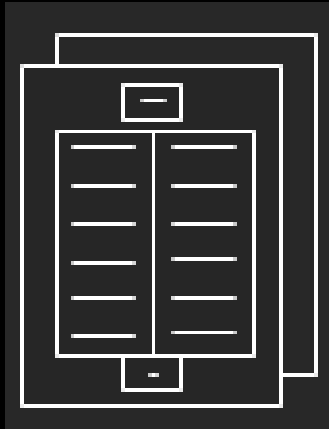




# Introducing Amazon Textract: Extract text and data from virtually any document



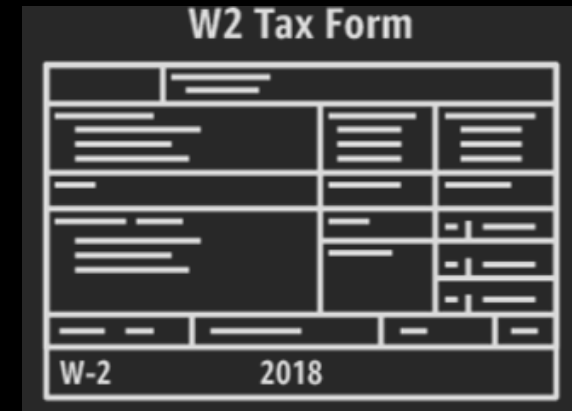
# Amazon Textract features



Text extraction



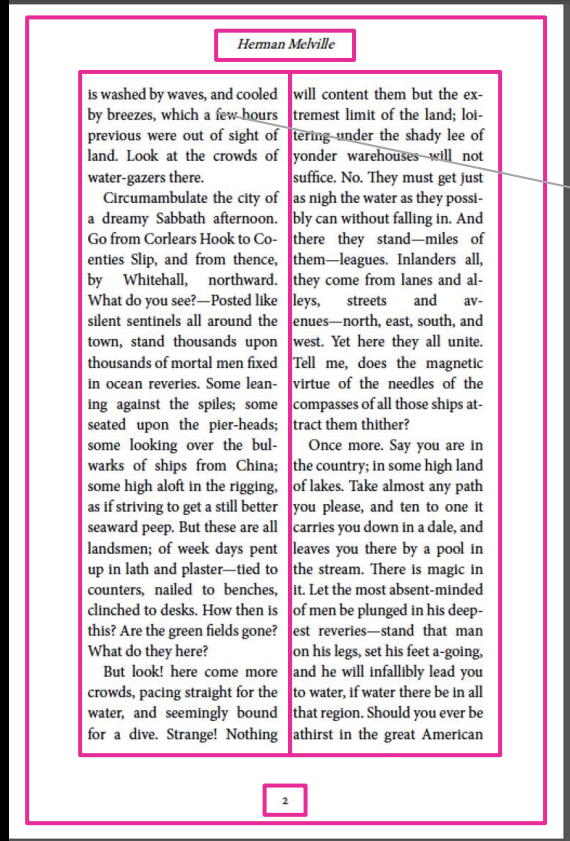
Table extraction



Form extraction

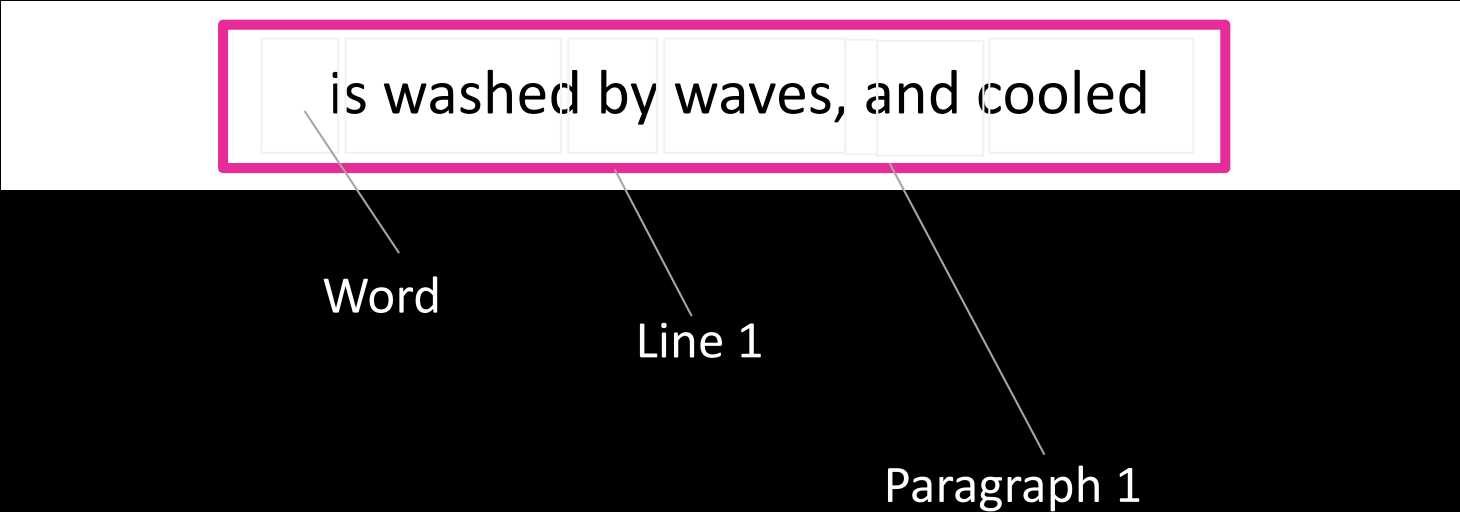
# Amazon Textract: Text extraction

## Document



## Output

Blocks: PAGE, PARAGRAPH, LINE, WORD



# Amazon Textract text extraction API: DetectDocumentText

## Request

Name	Description
Document	Blob or Amazon S3 object

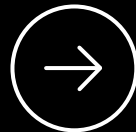
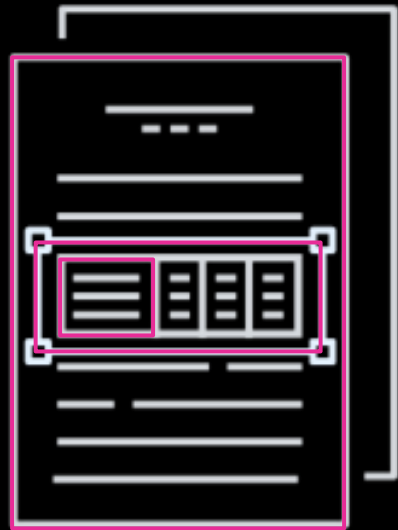
## Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	CHILD
Block type	PAGE, PARAGRAPH, LINE, WORD
Pages	Contains number of pages in the document

# Amazon Textract: Table extraction

## Document

---



## Output

---

Blocks: PAGE, TABLE, CELL

For each block, you get

- Text
- Confidence score
- Block relationships (e.g., cells within a table)

# Amazon Textract table extraction API:

## Analyze document with tables as FeatureTypes parameter

### Request

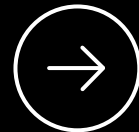
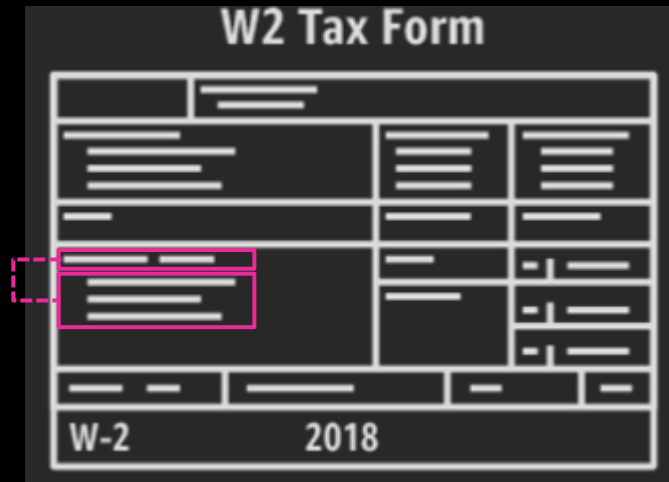
Name	Description
Document	Blob or Amazon S3 object
FeatureTypes	TABLES

### Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	CHILD
Block type	PAGE, TABLE, CELL
Pages	Contains number of pages in the document

# Amazon Textract: Form extraction

## Document



## Output

Blocks: PAGE, KEY\_VALUE\_SET

For each block of your document

- Form field name (key) and field value (value) association
- Confidence score
- Page number
- Block relationships

# Amazon Textract forms extraction API:

## Analyze document with forms as FeatureTypes parameter

### Request

Name	Description
Document	Blob or Amazon S3 object
FeatureTypes	FORMS

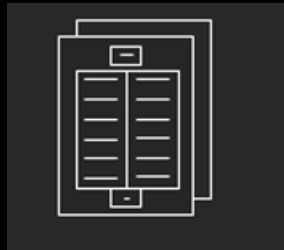
### Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	KEY, VALUE, CHILD
Block type	PAGE, KEY_VALUE_SET
Pages	Contains number of pages in the document

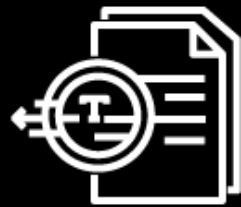
# Amazon Textract: Sync and async

## Synchronous

---



Document



Amazon Textract

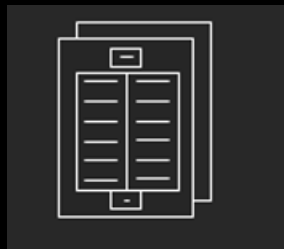


Get results

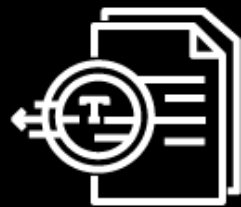
Supports single-page documents such as images (e.g., mobile capture)

## Asynchronous

---



Document



Amazon Textract



Notification



Get results

For multi-page documents, up to 3,000 pages



# Amazon Textract: Text extraction simplified

## Extract data quickly & accurately

Textract makes it easy to quickly and accurately extract data from documents and forms. Textract automatically detects a document's layout and the key elements on the page, understands the data relationships in any embedded forms or tables, and extracts everything with its context intact. This means you can instantly use the extracted data in an application or store it in a database without a lot of complicated code in between

## No code or templates to maintain

Textract's pre-trained machine learning models eliminate the need to write code for data extraction, because they have already been trained on tens of millions of documents from virtually every industry, including invoices, receipts, contracts, tax documents, sales orders, enrollment forms, benefit applications, insurance claims, policy documents and many more. You no longer need to maintain code for every document or form you might receive or worry about how page layouts change over time.



Multi-column detection

## Output

Extract data quickly & accurately

No code or templates to maintain

# Amazon Textract: Table extraction simplified

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



Table recognized



Words grouped by cell

## Output

{

Start Date: 1/15/2009  
End Date: 6/30/2013  
Employer Name: Any Company  
Position Held: Head Baker  
Reason for leaving: Family relocated

}

# Amazon Textract: Form extraction simplified

Full Name			Date of Birth			Gender	
John	X	Doe	01	01	1971	Male	<input checked="" type="radio"/>
First	Middle	Last	MM	DD	YYYY	Female	<input type="radio"/>

## Output

Full Name:  
First: John  
Middle: X  
Last: Doe

Date of Birth:  
MM: 01  
DD: 01  
YYYY: 1971

Gender:  
Male: True  
Female: False

- ✓ Logical groupings captured
- ✓ Relationships captured
- ✓ Glyphs captured

# Amazon Textract: Under the hood

# Text extraction: OCR reimagined

## Orientation


Jane Doe  
123 Any Street  
Any Town, USA

101

DATE \_\_\_\_\_

PAY TO THE ORDER OF \_\_\_\_\_ \$ \_\_\_\_\_

\_\_\_\_\_/100 DOLLARS

 AnyCompany

"000": 000 000 0000 0000

01/01/190005:12

SALES DRAFT

AnyCompany

Amazon Texttract Cafe

123 Any Street  
Any Town, USA

CASHIER: John Doe  
TERMINAL: 512  
RRN: 12345678901

CREDIT CARD

NAME: Mary Major  
NUMBER: XXXXXXXXXXXXXXX0000  
EXPIRY: 01/99  
AUTH: 123456  
AMOUNT: 123.45

TABLE: 12  
CHECK: 1234

AMOUNT: 123.45

GRATUITY: \_\_\_\_\_

TOTAL: \_\_\_\_\_

I agree to pay the above total amount in accordance with my credit card issuer agreement

X \_\_\_\_\_

# Text extraction: OCR reimaged

## Structure variability

Form W2 Wage and Tax Statement 2017										Copy B Employee Reference Copy	
d. control number 438209		Void		c. Employer's name, address, and ZIP		1 Wages, tips, other comp. \$39,010.32		2 Federal income tax withheld \$5,451.12			
b Employer's FED ID number 32-3939201		a Employee's SSA number 111-22-3333		ANYCOMPANY 100 MAIN STREET ANYTOWN, USA		3 Social security wages \$39,010.32		4 Social security tax withheld \$2439.08			
13 State employee Retirement plan X		3rd party sick pay		e Employee's name, address, and ZIP		5 Medicare wages and tips \$39,010.32		6 Medicare tax withheld \$456.90			
12 See instrs. For Box 12 C \$209		14 Other		JANE DOE 123 ANY STREET ANYTOWN, USA		7 Social security tips		8 Allocated tips			
15 State		Employer's State ID no.		16 State, wages, tips, etc.		17 State Income Tax		18 Local wages, tips, etc.			
						19 Local Income Tax		20 Locality name			

Form W-2 Wage and Tax Statement 2017										Department of the Treasury — Internal Revenue Service	
22222		a Employee's social security number		OMB No. 1545-0008		1 Wages, tips, other compensation		2 Federal income tax withheld			
b Employer identification number (EIN)		3 Social security wages		4 Social security tax withheld		5 Medicare wages and tips		6 Medicare tax withheld			
c Employer's name, address, and ZIP code		7 Social security tips		8 Allocated tips		9 Verification code		10 Dependent care benefits			
d Control number		11 Nonqualified plans		12a		12b		12c			
e Employee's first name and initial		Last name		Suff.		13 Statutory employee		14 Other			
f Employee's address and ZIP code		15 State		Employer's state ID number		16 State, wages, tips, etc.		17 State Income tax			
		18 Local wages, tips, etc.		19 Local income tax		20 Locality name					

W2 e file 2018			
Employee Reference Copy			
d. control number 438209	Dept 8840	Corp. A 439	Employer use only
c. Employer's name, address, and ZIP code ANYCOMPANY 100 MAIN STREET ANYTOWN, USA			
e/f Employee's name, address, and ZIP code JANE DOE 123 ANY STREET ANYTOWN, USA			
b Employer's FED ID number 32-3939201		a Employee's SSA number 111-22-3333	
1 Wages, tips, other comp. \$39,010.32		2 Federal income tax withheld \$5,451.12	
3 Social security wages \$39,010.32		4 Social security tax withheld \$2439.08	
5 Medicare wages and tips \$39,010.32		6 Medicare tax withheld \$456.90	
7 Social security tips		8 Allocated tips	
9		10 Dependent care benefits	
11 Nonqualified plans		12a See instructions for box 12 C \$209	
14 Other		12b D \$395.16	
15 State NH		Employer's State ID no.	
17 State Income Tax		18 Local wages, tips, etc.	
19 Local Income Tax		20 Locality name	


a Employee's SSA number 111-22-3333	1 Wages, tips, other comp. \$39,010.32	2 Federal income tax withheld \$5,451.12
b Employer's FED ID number 32-3939201	3 Social security wages \$39,010.32	4 Social security tax withheld \$2439.08
c. Employer's name, address, and ZIP code ANYCOMPANY 100 MAIN STREET ANYTOWN, USA		6 Medicare tax withheld \$456.90
e. Employee's first name and initial JANE		
Last name DOE		
Suff.		
123 ANY STREET ANYTOWN, USA		
f. Employee's address and ZIP code		
d. control number 438209	7 Social security tips	8 Allocated tips
9	10 Dependent care benefits	11 Nonqualified plans
12a See instructions for box 12 C \$209		14 Other
12b D \$395.16		
12c		
12d		
13 Statutory employee <input type="checkbox"/> Retirement plan <input checked="" type="checkbox"/> Third-party sick pay <input type="checkbox"/>		
15 State NH	Employer's State ID no.	16 State, wages, tips, etc.
17 State Income Tax		18 Local wages, tips, etc.
19 Local Income Tax		20 Locality name

# Text extraction: OCR reimagined

## Document variability

22222		a Employee's social security number		OMB No. 1545-0008	
b Employer identification number (EIN)		1 Wages, tips, other compensation		2 Federal income tax withheld	
c Employer's name, address, and ZIP code		3 Social security wages		4 Social security tax withheld	
		5 Medicare wages and tips		6 Medicare tax withheld	
		7 Social security tips		8 Allocated tips	
d Control number		9 Verification code		10 Dependent care benefits	
e Employee's first name and initial		Last name		Suft.	
		11 Nonqualified plans		12a	
		13 Statutory employee Retirement plan Third party sick pay		12b	
		14 Other		12c	
				12d	
f Employee's address and ZIP code					
15 State Employer's state ID number		16 State wages, tips, etc.		17 State income tax	
		18 Local wages, tips, etc.		19 Local income tax	
				20 Locality name	

Form **W-2** Wage and Tax Statement 2017 Department of the Treasury—Internal Revenue Service  
Copy 1—For State, City, or Local Tax Department

Jane Doe		101	
123 Any Street			
Any Town, USA		DATE _____	
PAY TO THE ORDER OF _____		\$ _____	
_____		/100 DOLLARS	
 AnyCompany			
"000": 000 000 0000 0000			

01/01/1900		05:12	
SALES DRAFT			
AnyCompany			
Amazon Textract Cafe			
123 Any Street			
Any Town, USA			
CASHIER: John Doe			
TERMINAL: 512			
RRN: 12345678901			
CREDIT CARD			
NAME: Mary Major			
NUMBER: XXXXXXXXXXXX0000			
EXPIRY: 01/99			
AUTH: 123456			
AMOUNT: 123.45			
TABLE: 12			
CHECK: 1234			
AMOUNT:		123.45	
GRATUITY:		_____	
TOTAL:		_____	
I agree to pay the above total amount in accordance with my credit card issuer agreement			
X _____			
Signature			
Customer Copy			

Herman Melville

is washed by waves, and cooled by breezes, which a few hours previous were out of sight of land. Look at the crowds of water-gazers there.

Circumambulate the city of a dreamy Sabbath afternoon. Go from Corlears Hook to Coenties Slip, and from thence, by Whitehall, northward. What do you see?—Posted like silent sentinels all around the town, stand thousands upon thousands of mortal men fixed in ocean reveries. Some leaning against the spiles; some seated upon the pier-heads; some looking over the bulwarks of ships from China; some high aloft in the rigging, as if striving to get a still better seaward peep. But these are all landsmen; of week days pent up in lath and plaster—tied to counters, nailed to benches, clinched to desks. How then is this? Are the green fields gone? What do they here?

But look! here come more crowds, pacing straight for the water, and seemingly bound for a dive. Strange! Nothing will content them but the extremest limit of the land; loitering under the shady lee of yonder warehouses will not suffice. No. They must get just as nigh the water as they possibly can without falling in. And there they stand—miles of them—leagues. Inlanders all, they come from lanes and alleys, streets and avenues—north, east, south, and west. Yet here they all unite. Tell me, does the magnetic virtue of the needles of the compasses of all those ships attract them thither?

Once more. Say you are in the country; in some high land of lakes. Take almost any path you please, and ten to one it carries you down in a dale, and leaves you there by a pool in the stream. There is magic in it. Let the most absent-minded of men be plunged in his deepest reveries—stand that man on his legs, set his feet a-going, and he will infallibly lead you to water, if water there be in all that region. Should you ever be athirst in the great American

2



# Beyond OCR: Segmentation and rectification

## Photometric

**CONSOLIDATED DIAGNOSTIC PATHOLOGY FORM\***

Microscopic Appearance:

1. *Histological pattern:*

CELL DISTRIBUTION		+	-	STRUCTURAL PATTERN		+	-
Diffuse			<input checked="" type="checkbox"/>	Streaming			
Mosaic		<input checked="" type="checkbox"/>		Storiform			
Necrosis			<input checked="" type="checkbox"/>	Fibrosis			
Lymphocytic Infiltration		<input checked="" type="checkbox"/>		Palisading			
Vascular Invasion			<input checked="" type="checkbox"/>	Cystic Degeneration			
Clusterized		<input checked="" type="checkbox"/>		Bleeding			
Alveolar Formation			<input checked="" type="checkbox"/>	Myxoid Change			
Indian File			<input checked="" type="checkbox"/>	Pseudomoma/Calcification			

2. *Cellular features:*

Squamous	+	-	Adenomatous	+	-	Sarcomatous	+	-	Lymphomatous	+	-
Squamous Cell			Glandular cell	<input checked="" type="checkbox"/>		Round Cell			Large Cell		
Spindle Cell			Cell Stratification	<input checked="" type="checkbox"/>		Fibroblast			Small Cell		
Keratin			Secretion	<input checked="" type="checkbox"/>		Osteoblast			RS Cell/RS Like		
Desmosome			Intracyt. Vacuole	<input checked="" type="checkbox"/>		Lipoblast			Inflam. Cell		
Pearl			Gland formation	<input checked="" type="checkbox"/>		Myoblast			Plasma Cell		

Otherwise Specified: D1 75%, D2 75%, D3 75%, D4 75%

2. *Cellular Differentiation:*

Well	Moderately	Poor
		<input checked="" type="checkbox"/>

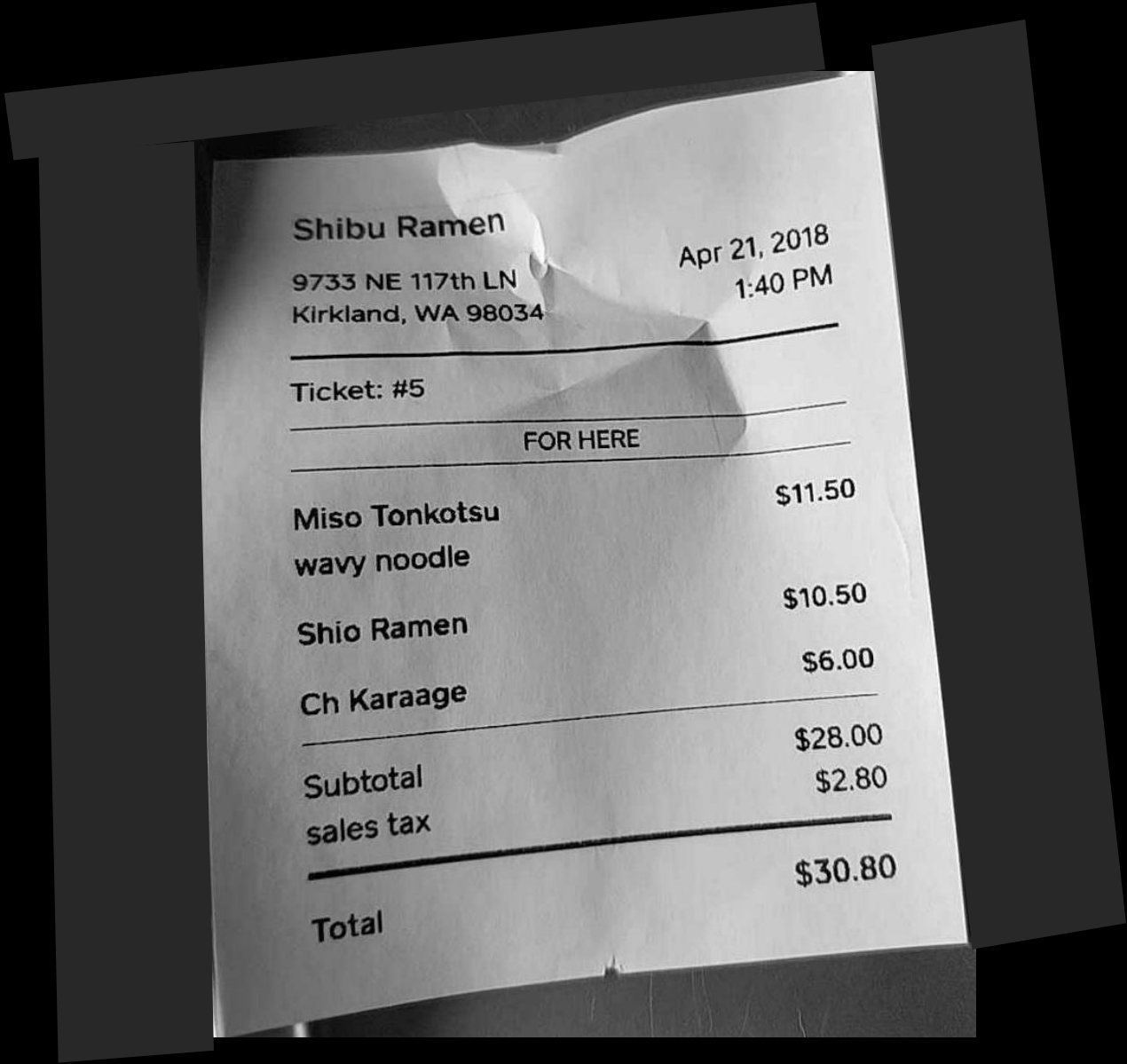
3. *Nuclear Atypia:*

Nuclear Appearance	0	I	II	III
Aniso Nucleosis				<input checked="" type="checkbox"/>
Hyperchromatism				<input checked="" type="checkbox"/>
Nucleolar Prominent				<input checked="" type="checkbox"/>
Multinucleated Giant Cell				<input checked="" type="checkbox"/>
Mitotic Activity				<input checked="" type="checkbox"/>
Nuclear Grade				



# Beyond OCR: Segmentation and rectification

Geometric



# Beyond OCR: Table and cell detection

Understand document structure and context to find tables

Employment Application

Applicant Information

Full Name:

Jane Doe

Phone Number:

555-0100

Home Address:

123 Any Street, Any Town, USA

Mailing Address:

same as home address

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

# Beyond OCR: Table and cell detection

Understand cells even without explicit boundaries

Annual expenses:

	Previous Year	Current Year
Transportation	\$9,044	\$10,028
Accommodations	\$47,329	\$86,345
Childcare	\$3,222	\$4,149
Total	\$78,886	\$120,962

# Beyond OCR: Table and cell detection

## Variable-sized rows and columns

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

# Beyond OCR: Field name (key) and value extraction

Detect phrases or groups of words

Full Name			Date of Birth			Gender	
John	X	Doe	01	01	1971	Male	<input checked="" type="radio"/>
First	Middle	Last	MM	DD	YYYY	Female	<input type="radio"/>

## Output

Full Name:  
First: John  
Middle: X  
Last: Doe

Date of Birth:  
MM: 01  
DD: 01  
YYYY: 1971

Gender:  
Male: True  
Female: False

# Beyond OCR: Inferring key/value association

## Detect structures of the same form without templates

Form W2 Wage and Tax Statement 2017										Copy B Employee Reference Copy																								
d. control number 438209					c. Employer's name, address, and ZIP code PARIS BAKERY 123 MAIN STREET HANOVER, NH 03755					1 Wages, tips, other comp. \$39,010.32					2 Federal income tax withheld \$5,451.12																			
b. Employer's FED ID number 32-3939201					a. Employer's SSA number 111-22-3333					3 Social security wages \$39,010.32					4 Social security tax withheld \$2439.08																			
11 Stat employee					Retirement plan X					3* party sick pay					5 Medicare wages and tips \$39,010.32					6 Medicare tax withheld \$456.90														
12 See instrs. For Box 12 C \$209 D \$395.16					14 Other					e. Employer's name, address, and ZIP code LILLIAN CRANE 1893 ORCHARD RD WHITE RIVER JUNCTION, NH 03789					7 Social security tips					8 Allocated tips														
13 State NH					Employer's State ID no.					16 State wages, tips, etc.					17 State Income Tax					18 Local wages, tips, etc.					19 Local Income Tax					20 Locality name				

22222										a. Employee's social security number OMB No. 1545-0008																													
b. Employer identification number (EIN)										1 Wages, tips, other compensation					2 Federal income tax withheld																								
c. Employer's name, address, and ZIP code										3 Social security wages					4 Social security tax withheld																								
										5 Medicare wages and tips					6 Medicare tax withheld																								
										7 Social security tips					8 Allocated tips																								
d. Control number										9 Verification code					10 Dependent care benefits																								
e. Employee's first name and initial Last name Suff.										11 Nonqualified plans					12a																								
f. Employee's address and ZIP code										13 Statutory employee Retirement plan Third-party sick pay					12b																								
										14 Other					12c																								
															12d																								
15 State NH										Employer's state ID number					16 State wages, tips, etc.					17 State Income Tax					18 Local wages, tips, etc.					19 Local income tax					20 Locality name				

Form **W-2** Wage and Tax Statement 2017 Department of the Treasury—Internal Revenue Service  
Copy 1—For State, City, or Local Tax Department

W2 <i>e file</i> 2017																																		
Employee Reference Copy																																		
d. control number 438209					Dept 6840					Corp. A 439					Employer use only																			
c. Employer's name, address, and ZIP code PARIS BAKERY 123 MAIN STREET HANOVER, NH 03755										e/f. Employee's name, address, and ZIP code LILLIAN CRANE 1893 ORCHARD RD WHITE RIVER JUNCTION, NH 03789																								
b. Employer's FED ID number 32-3939201					a. Employer's SSA number 111-22-3333					1 Wages, tips, other comp. \$39,010.32					2 Federal income tax withheld \$5,451.12																			
3 Social security wages \$39,010.32					4 Social security tax withheld \$2439.08					5 Medicare wages and tips \$39,010.32					6 Medicare tax withheld \$456.90																			
7 Social security tips					8 Allocated tips					9					10 Dependent care benefits																			
11 Nonqualified plans					12a See instructions for box 12 C \$209 D \$395.16 12c 12d					13 Stat. emp. Ret. plan 3* party sick pay					14 Other																			
15 State NH					Employer's State ID no.					16 State wages, tips, etc.					17 State Income Tax					18 Local wages, tips, etc.					19 Local Income Tax					20 Locality name				

a. Employer's SSA number 111-22-3333					1 Wages, tips, other comp. \$39,010.32					2 Federal income tax withheld \$5,451.12									
b. Employer's FED ID number 32-3939201					3 Social security wages \$39,010.32					4 Social security tax withheld \$2439.08									
					5 Medicare wages and tips \$39,010.32					6 Medicare tax withheld \$456.90									
c. Employer's name, address, and ZIP code PARIS BAKERY 123 MAIN STREET HANOVER, NH 03755										e. Employer's first name and initial Last name Suff. LILLIAN CRANE 1893 ORCHARD RD WHITE RIVER JUNCTION, NH 03789									
d. control number 438209					7 Social security tips					8 Allocated tips									
9					10 Dependent care benefits					11 Nonqualified plans									
12a See instructions for box 12 C \$209 12b D \$395.16 12c 12d					14 Other														
13 Statutory employee <input checked="" type="checkbox"/> Retirement plan <input type="checkbox"/> Third-party sick pay <input type="checkbox"/>																			
15 State NH					Employer's State ID no.					16 State wages, tips, etc.					17 State Income Tax				
18 Local wages, tips, etc.					19 Local Income Tax					20 Locality name									

W2 Wage and Tax Statement 2017 Copy B Employee Reference Copy

# Beyond OCR: Inferring key/value association

## Key/value association

Name	Social Security #	Date of Birth	Relationship
<div>↓</div> John Doe	000-00-0000	01/01/1972	

a. Are you making support payments for a dependent noted above or on your attachment(s)? ☒ Yes ☐ No

<div>↑</div> John Doe	100 Main Street	Any town
Name	Address	City

# Beyond OCR: Inferring key/value association

## Infer empty values

Full Name			Date of Birth			Gender	
John		Doe	01	01	1971	Male	<input checked="" type="radio"/>
First	Middle	Last	MM	DD	YYYY	Female	<input type="radio"/>

## Output

Full Name:

First: John

Middle: null

Last: Doe

Date of Birth:

MM: 01

DD: 01

YYYY: 1971

Gender:

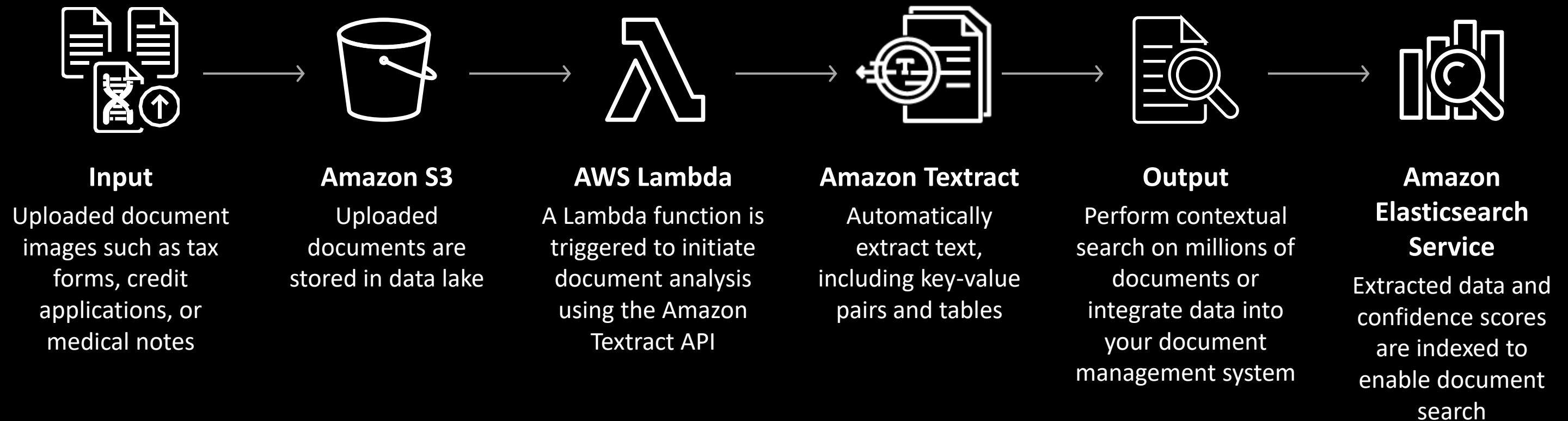
Male: True

Female: False

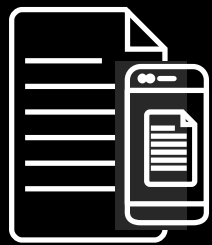


# Amazon Textract: Reference implementation

# Reference architecture: Index and search documents

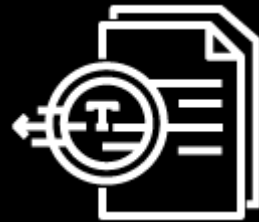


# Reference architecture: Form capture



## Input

Customer uses mobile app to capture a photo of a W-2 form



## Amazon Textract

The Amazon Textract API is integrated into the end-user application to automatically extract text from the W-2 form and auto-populate the form fields



## Customer application

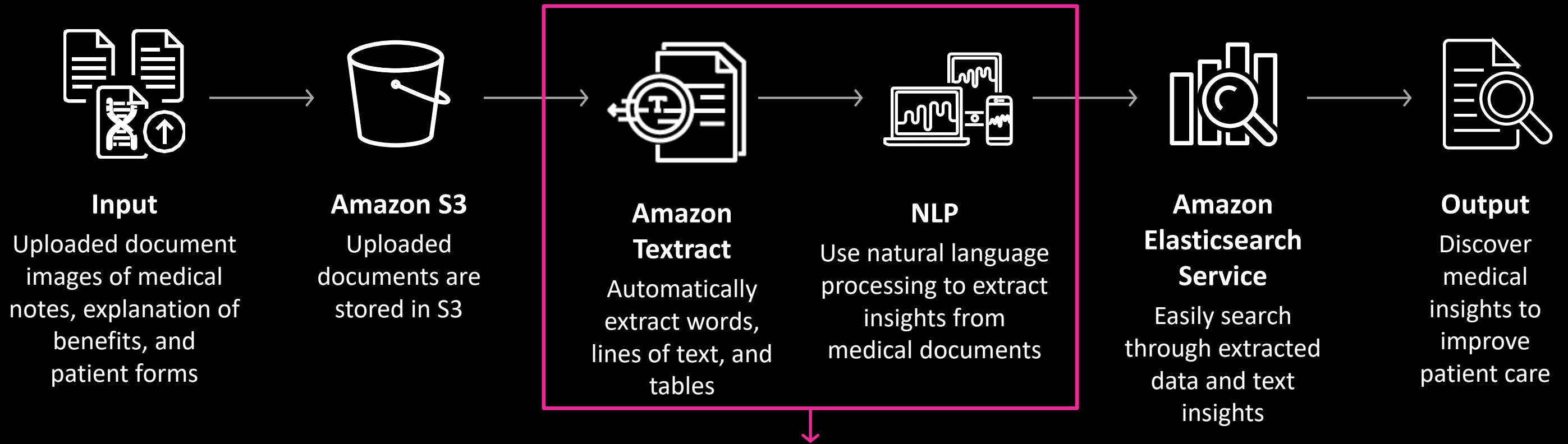
Customers experience real-time capture of tax information by taking a photo instead of performing manual data entry



## Database

User-submitted data is loaded into a database for use in tax preparation

# Reference architecture: Extract for NLP



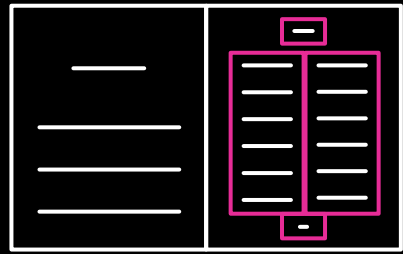
Quickly turn extracted text and data into actionable insights

# Amazon Textract: Customers

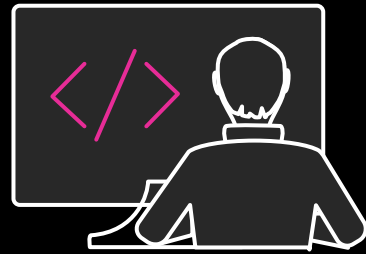
# Amazon Textract: Launch customers



# Amazon Textract: Benefits



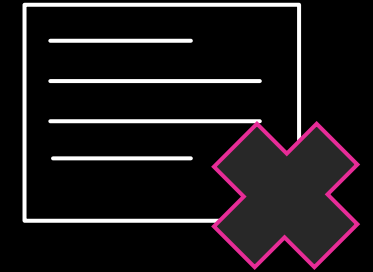
Extract data quickly and accurately



Eliminate manual effort



Lower document processing costs



No ML experience required

# Thank you!

Randall Hunt