Introducing Amazon Textract

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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…

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

16.3M US mortgage applications ($2.1T) 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 _____

I. TYPE OF MORTGAGE AND TERMS OF LOAN

Mortgage Applied for:	☐ VA ☐ FHA	☐ Conventional ☐ USDA/Rural Housing Service	☐ Other (explain):	Agency Case Number	Lender Case Number

Amount	Interest Rate	No. of Months	**Amortization Type:**	
$	%		☐ Fixed Rate ☐ GPM	☐ Other (explain): ☐ ARM (type):

II. PROPERTY INFORMATION AND PURPOSE OF LOAN

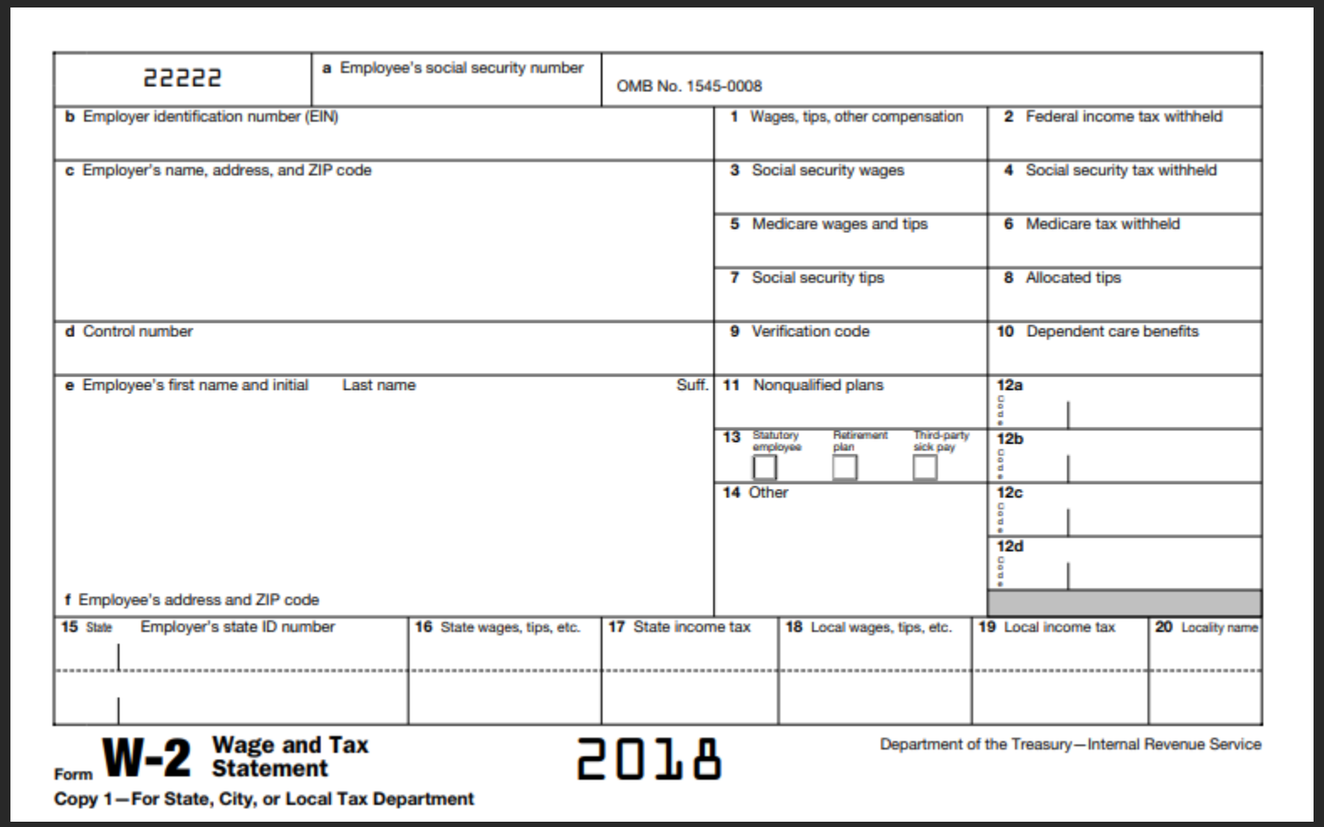
Subject Property Address (street, city, state & ZIP)	No. of Units
Legal Description of Subject Property (attach description if necessary)	Year Built

Purpose of Loan ☐ Purchase ☐ Construction ☐ Other (explain): ☐ Refinance ☐ Construction-Permanent	Property will be: ☐ Primary Residence ☐ Secondary Residence ☐ Investment

*Mortgage Bankers Association 2016 HMDA

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

About 240M W2 Tax Forms will be Processed for FY2018 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 Code

13 Statutory employee Retirement plan Third-party sick pay

12b Code

14 Other

12c Code

12d Code

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

Department of the Treasury—Internal Revenue Service

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

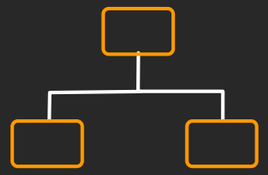
*IRS—https://www.irs.gov/individuals/w-2-verification-code

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Need for processing documents

Search
and discovery

Compliance
and control



Business
process automation

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

How documents are processed today

Manual processing

Optical Character Recognition (OCR)

Rules and template-based extraction

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents
Manual processing

Expensive

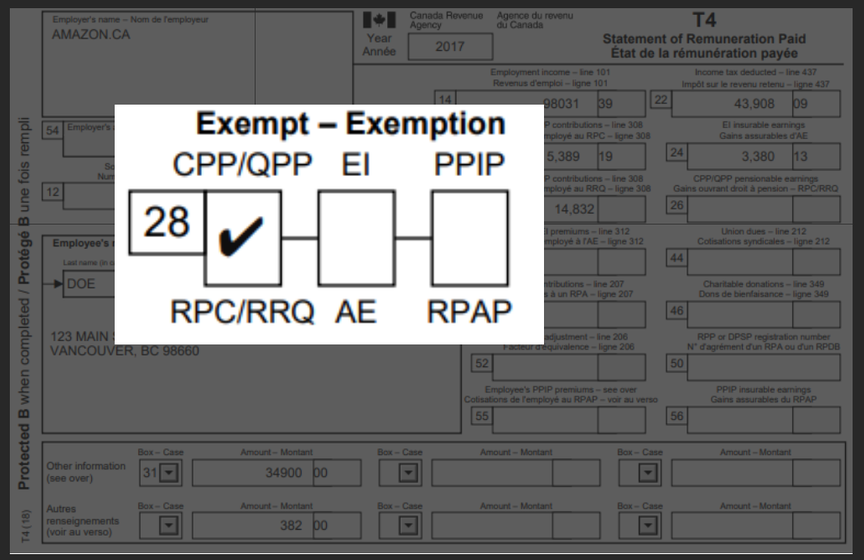
Error prone

Time consuming

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents

Manual processing



Output

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

- Variable Output
- Inconsistent Results
- Needs multiple human reviews for consensus

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents
Optical Character Recognition (OCR)

Simple Documents Only

Error prone

Flat bag of words

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents
Optical Character Recognition (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 templates to accurately maintain

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents
Optical Character Recognition (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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents

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

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Challenges for processing documents
Rules and template-based extraction

Limited by accuracy of OCR

Significant development and management overhead

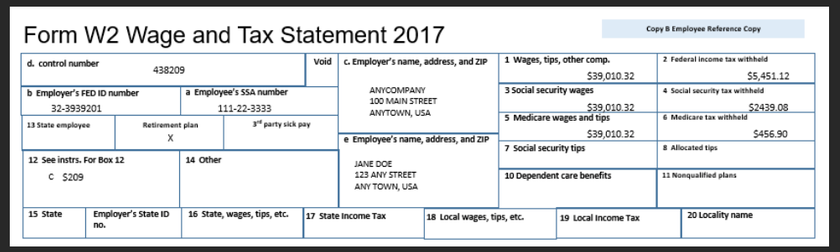
Templates are brittle

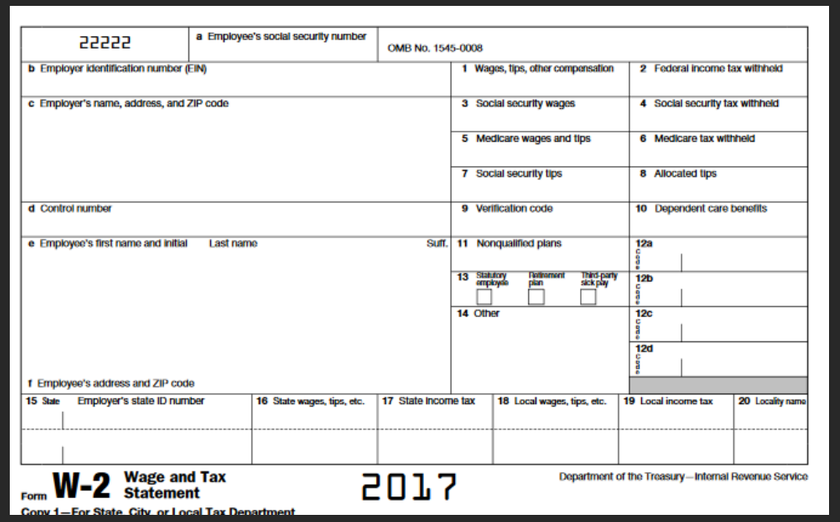
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

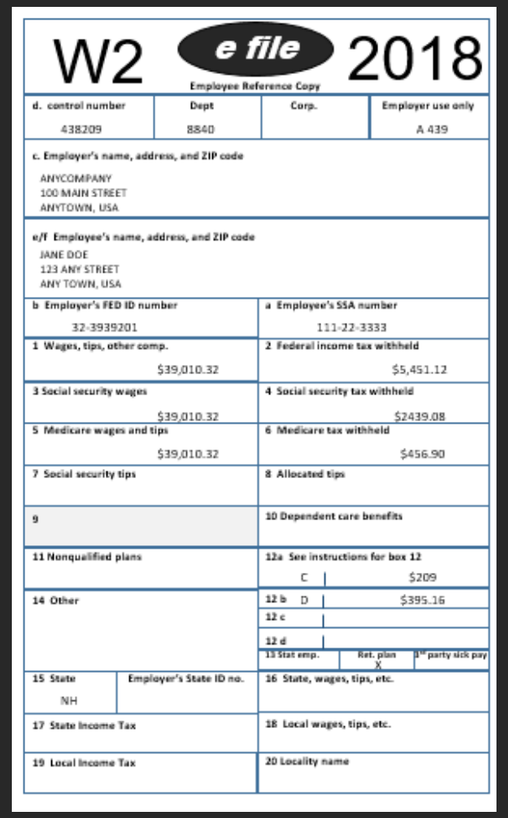
Challenges for processing documents

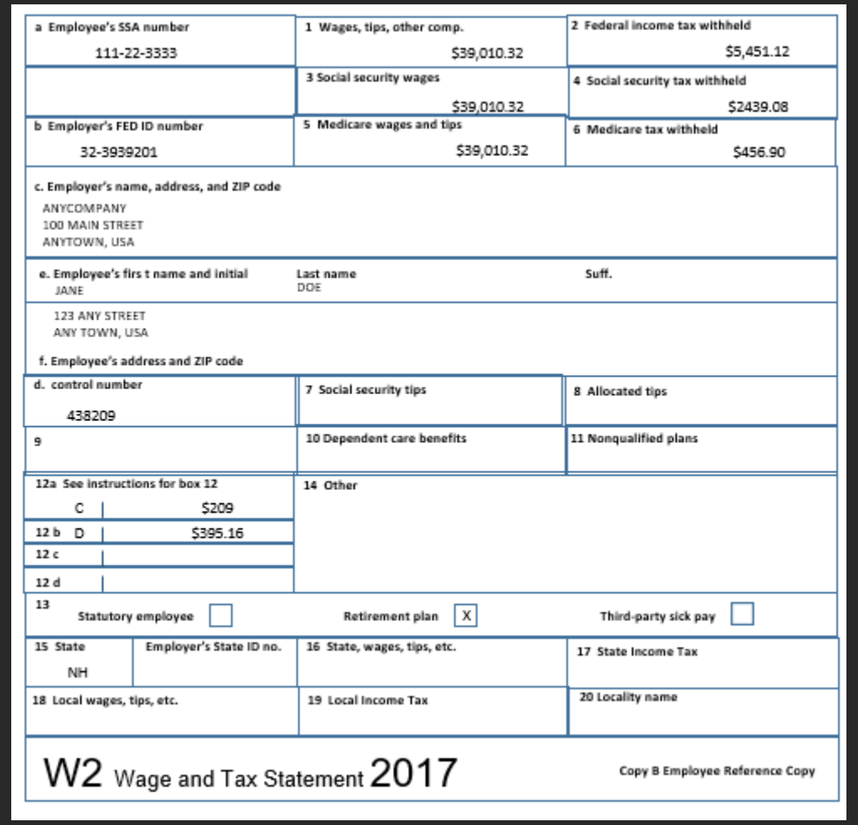
Rules and template-based extraction

The well-known W2 US tax form has 100s of variants each year



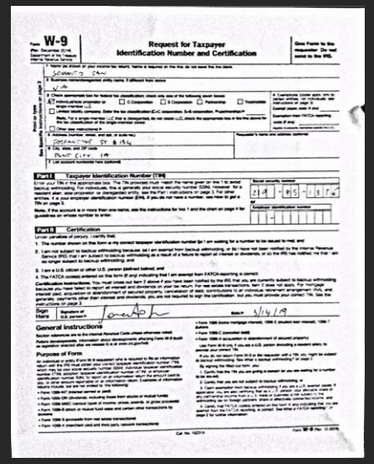
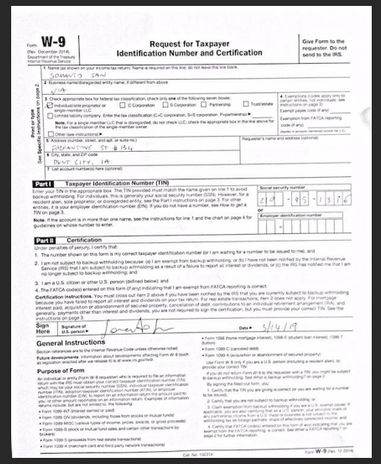
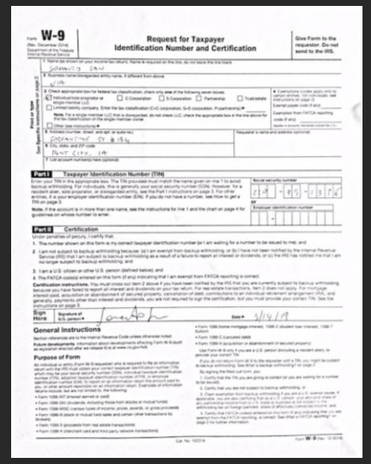
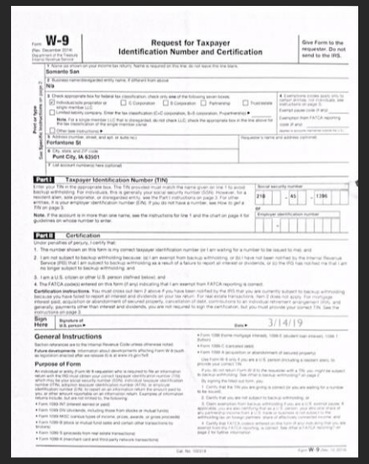




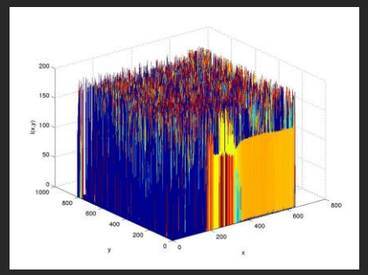
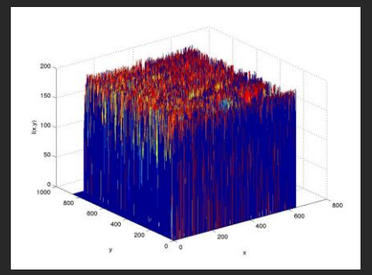
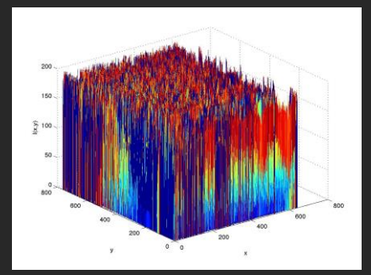
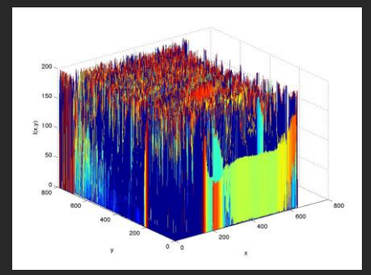


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

It looks easy, but …

…not a single corresponding pixel value in common

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Introducing Amazon Textract

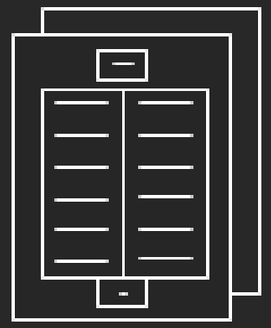
Extract text and data from virtually any document

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

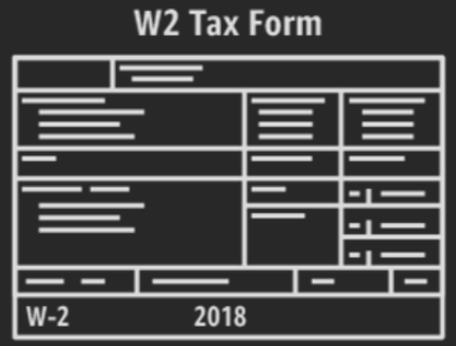
aws

Amazon Textract Features



Text extraction

Table extraction

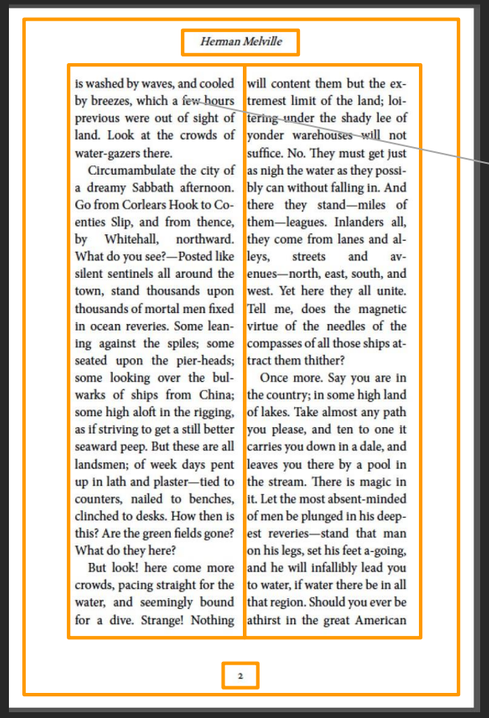


Form extraction

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

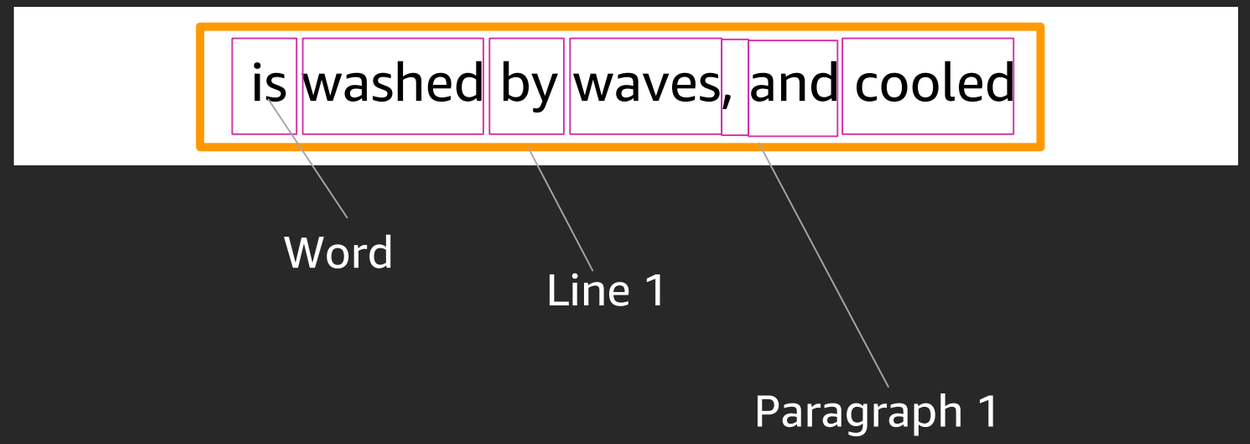
Amazon Textract - Text Extraction

Document



Output

Blocks: PAGE, PARAGRAPH, LINE, WORD



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract - Text Extraction API
DetectDocumentText

Request

Name	Description
Document	Blob or Amazon S3 object

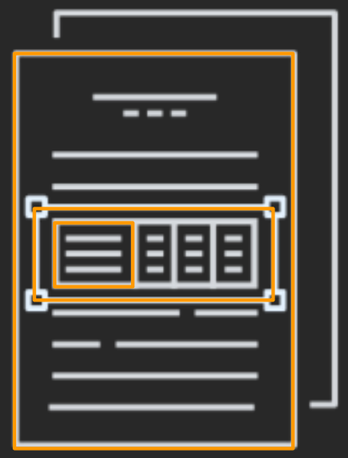
Response

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract - Table extraction API

AnalyzeDocument with “table” as FeatureTypes parameter

Request

Name	Description
Document	Blob or Amazon S3 object
FeatureTypes	TABLES

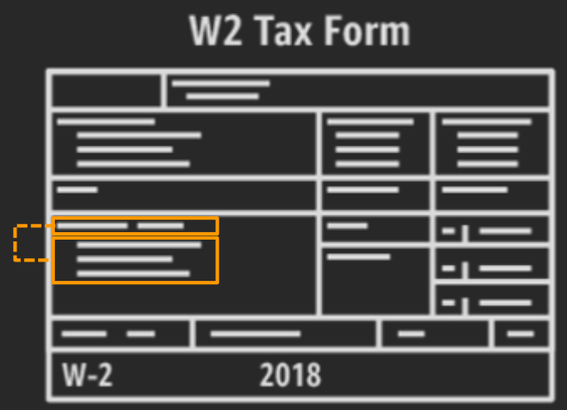
Response

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract - Forms Extraction API

AnalyzeDocument 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

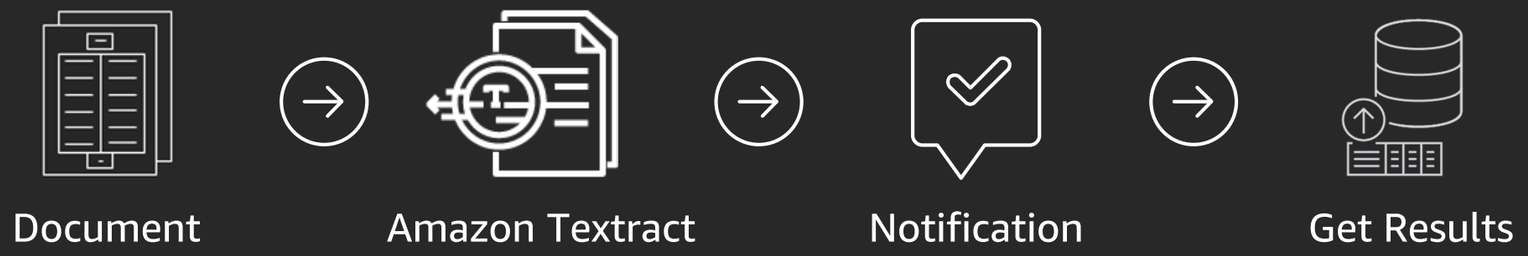
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract
Sync and Async

Synchronous

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

Asynchronous



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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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.

Output

Extract data quickly & accurately

No code or templates to maintain

- Multi-column detection

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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

}

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Amazon Textract –

Form Extraction simplified

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

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

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Amazon Textract
Under the Hood

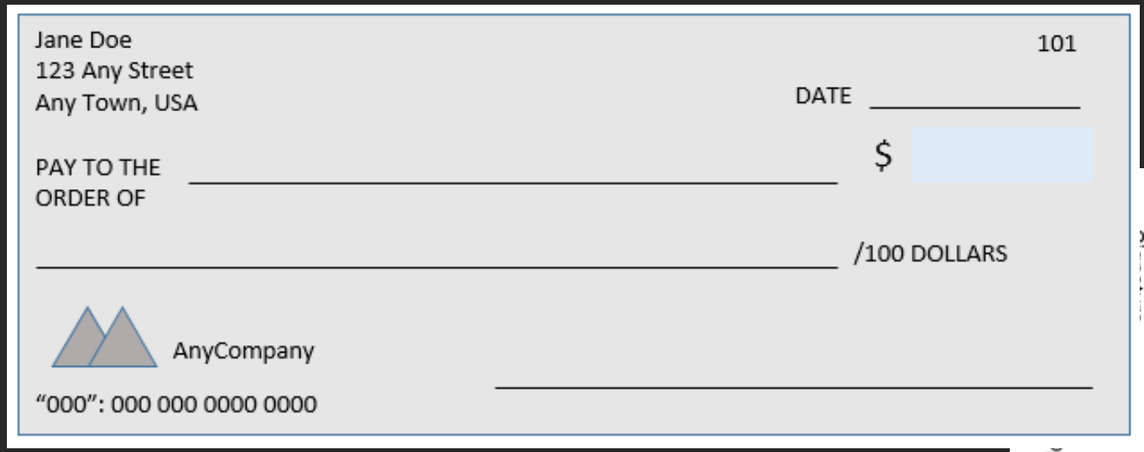
AWS re:Invent

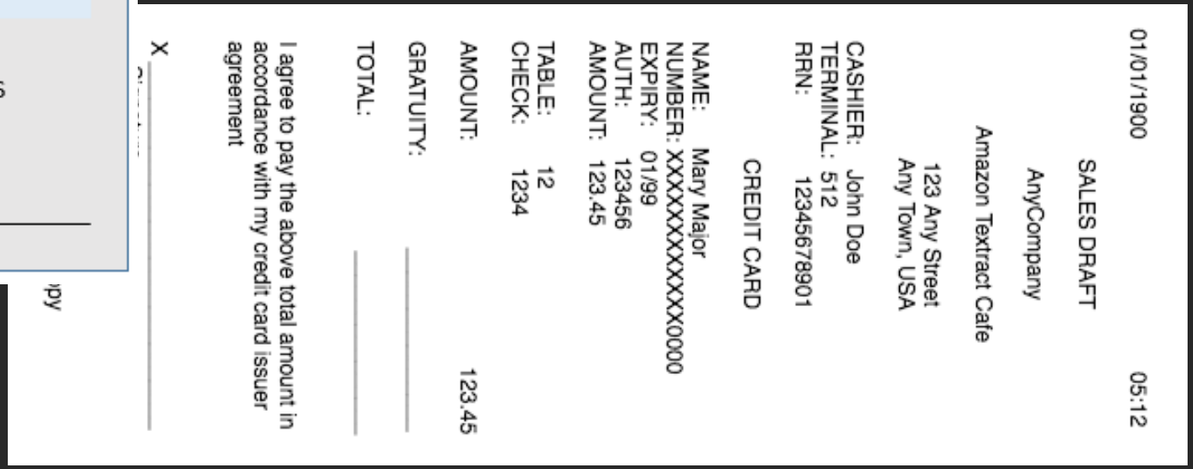
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Text Extraction: OCR reimagined

Orientation

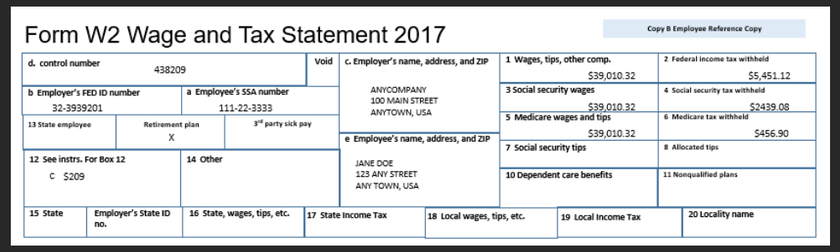


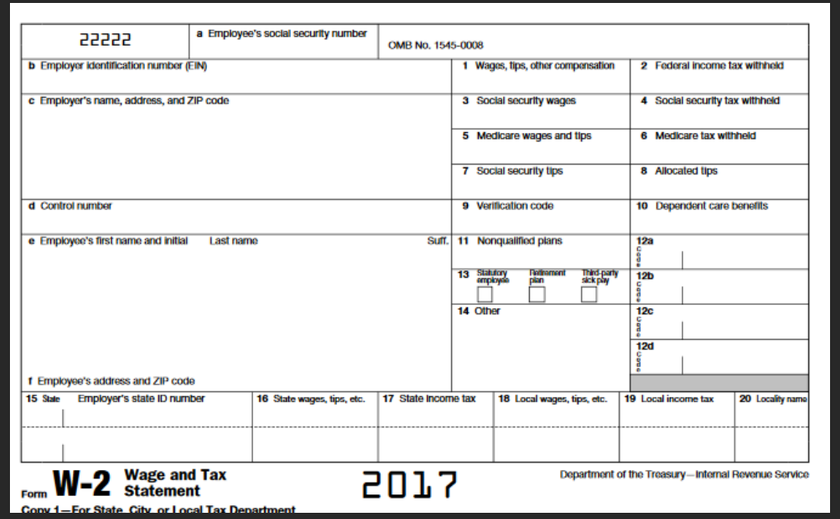


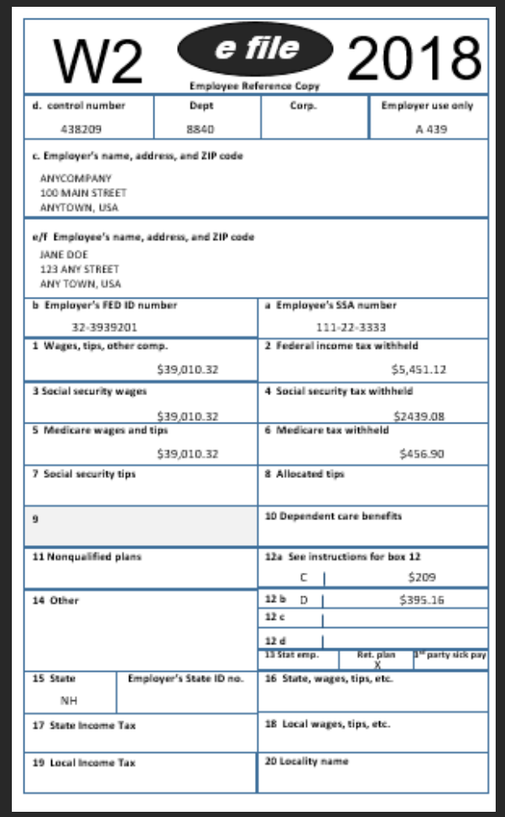
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

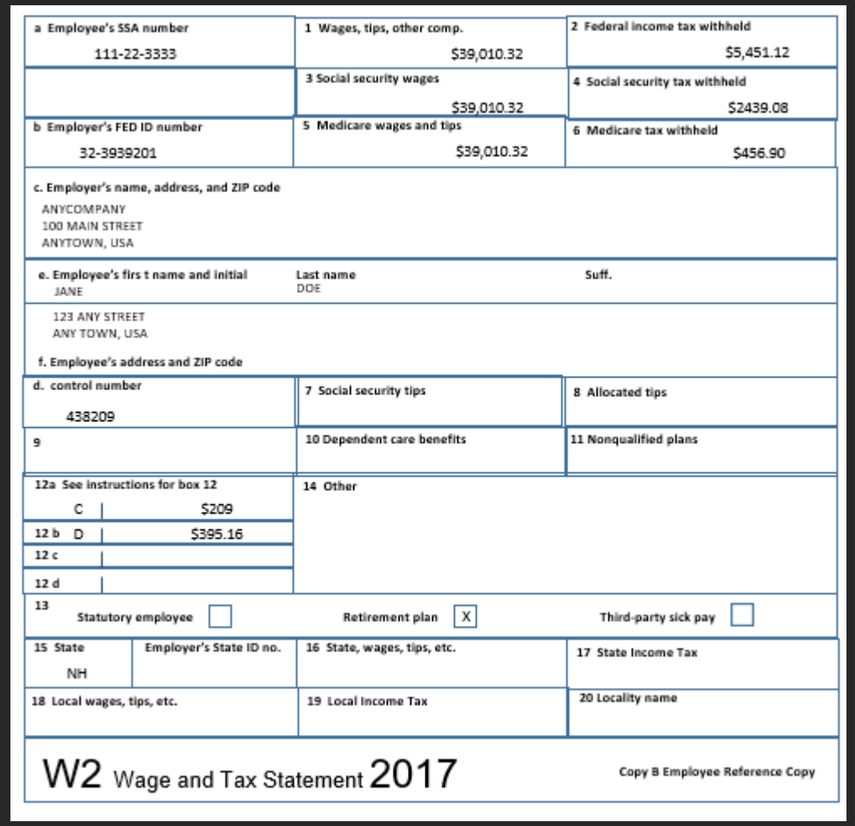
Text extraction: OCR reimagined

Structure variability





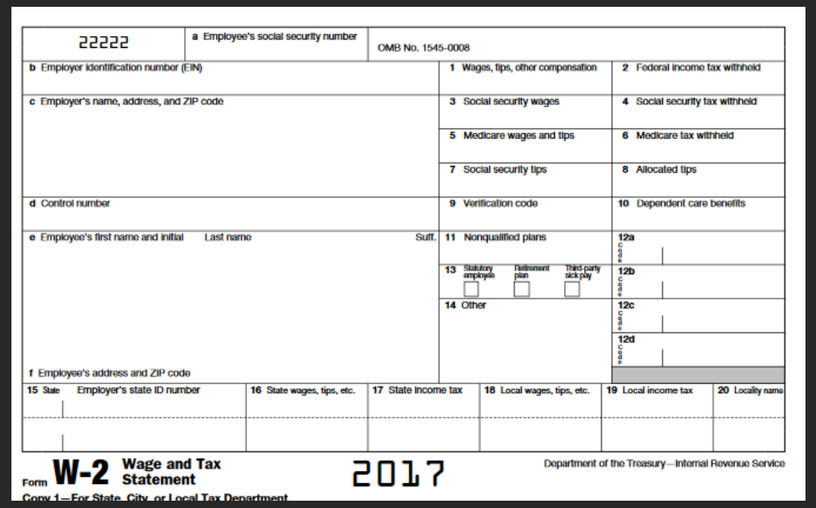




© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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 | Suff. | 11 Nonqualified plans | 12a

13 Statutory employee | Retirement plan | Third party sick pay | 12b

14 Other | 12c

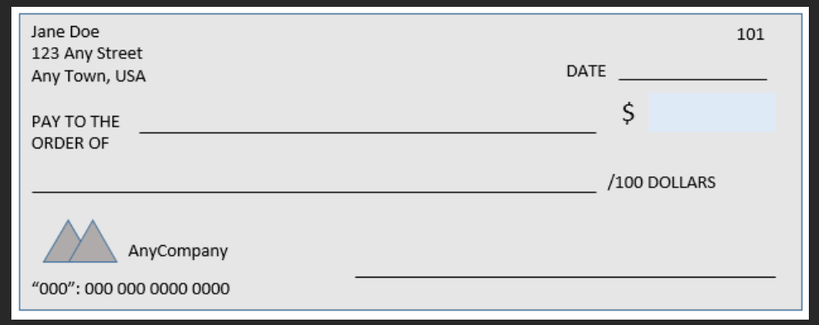
12d

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 2017 Department of the Treasury—Internal Revenue Service

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



Jane Doe
123 Any Street
Any Town, USA

101

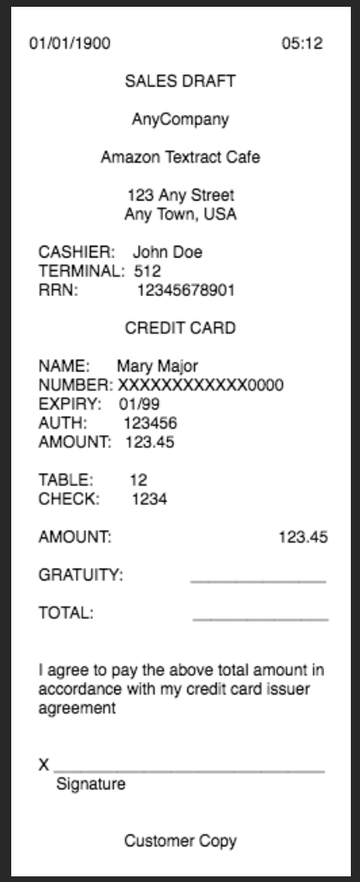
DATE _____

PAY TO THE ORDER OF _____ $

_____ /100 DOLLARS

AnyCompany

"000": 000 000 0000 0000



01/01/1900 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

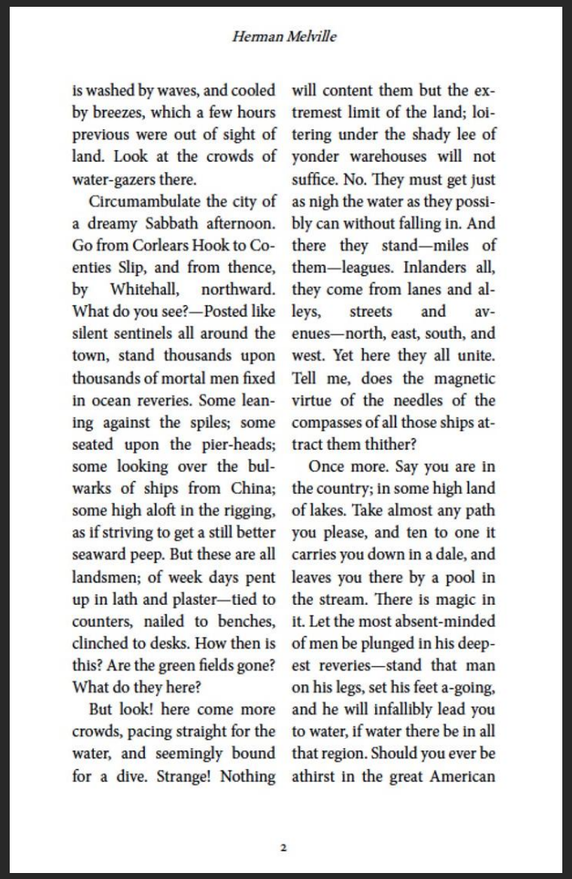
GRATUITY: _____

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?

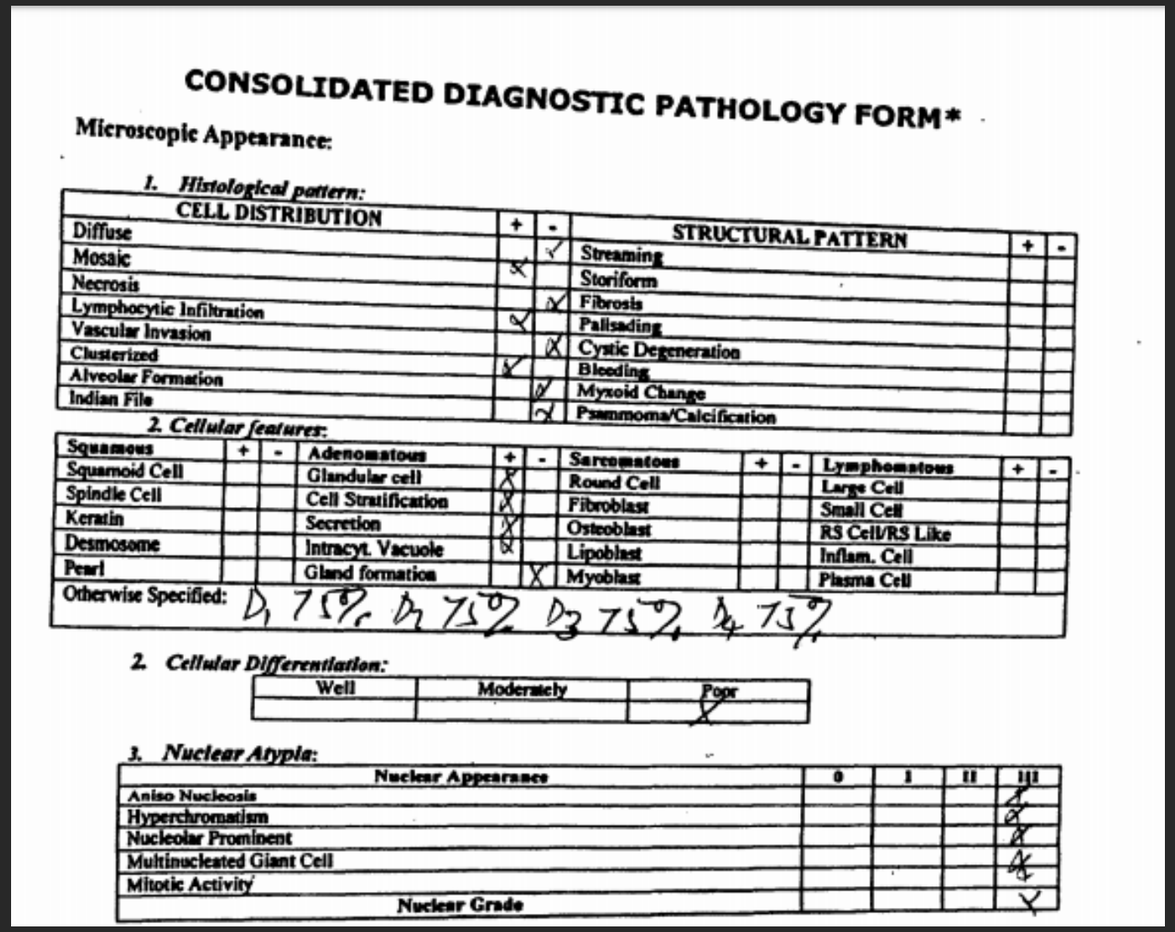
Once more. Say you are in the country; in some high land of lakes. Take almost any path you please, and ten to one it carries you down in a dale, and leaves you there by a pool in the stream. There is magic in it. Let the most absent-minded of men be plunged in his deepest reveries—stand that man on 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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Beyond OCR: Segmentation and rectification

Photometric



CONSOLIDATED DIAGNOSTIC PATHOLOGY FORM*

Microscopic Appearance:

1. Histological pattern:

CELL DISTRIBUTION	+	-	STRUCTURAL PATTERN	+	-
Diffuse		✓	Streaming		
Mosaic	✗		Storiform		
Necrosis		✗	Fibrosis		
Lymphocytic Infiltration	✗		Palisading		
Vascular Invasion		✗	Cystic Degeneration		
Clusterized	✓		Bleeding		
Alveolar Formation		✗	Myxoid Change		
Indian File		✗	Psammoma/Calcification		

2. Cellular features:

Squamous	+	-	Adenomatous	+	-	Sarcomatous	+	-	Lymphomatous	+	-
Squamoid Cell			Glandular cell	✗		Round Cell			Large Cell		
Spindle Cell			Cell Stratification	✗		Fibroblast			Small Cell		
Keratin			Secretion	✗		Osteoblast			RS Cell/RS Like		
Desmosome			Intracyt. Vacuole	✗		Lipoblast			Inflam. Cell		
Pearl			Gland formation		✗	Myoblast			Plasma Cell		

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

2. Cellular Differentiation:

Well	Moderately	Poor
		✗

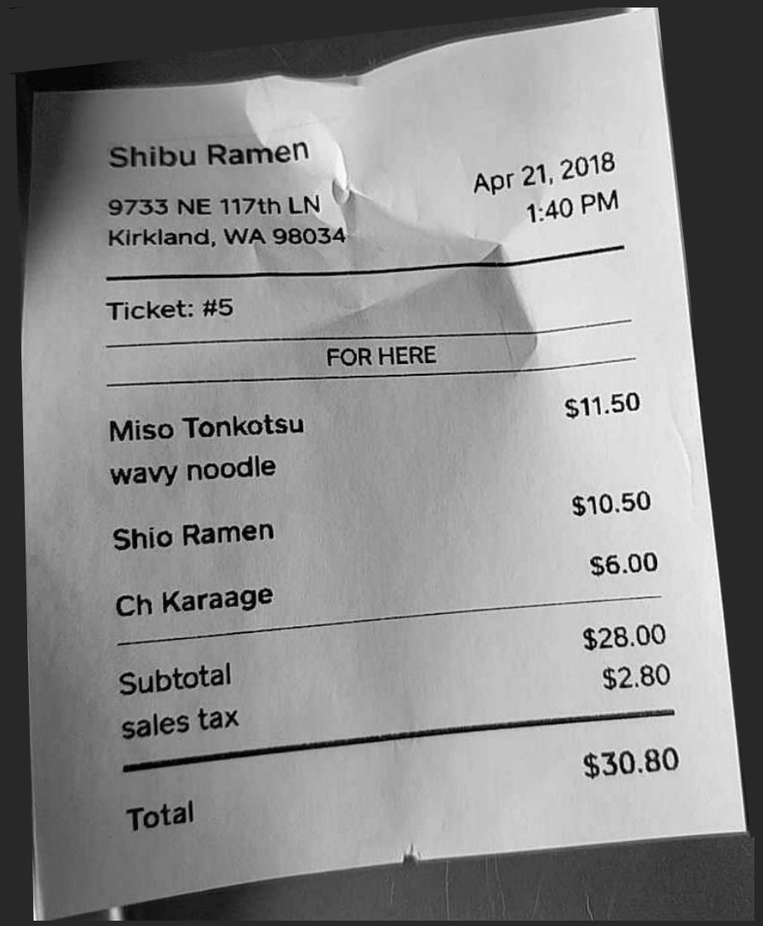
3. Nuclear Atypia:

Nuclear Appearance	0	I	II	III
Aniso Nucleosis				✗
Hyperchromatism				✗
Nucleolar Prominent				✗
Multinucleated Giant Cell				✗
Mitotic Activity				✗
Nuclear Grade				✓

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Beyond OCR: Segmentation and rectification

Geometric



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Beyond OCR: Field name (key) and value extraction

Detect phrases or groups of words

Output

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

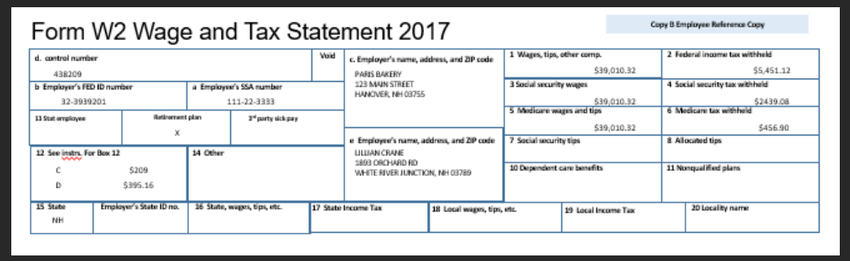
Date of Birth:
- MM: 01
- DD: 01
- YYYY: 1971

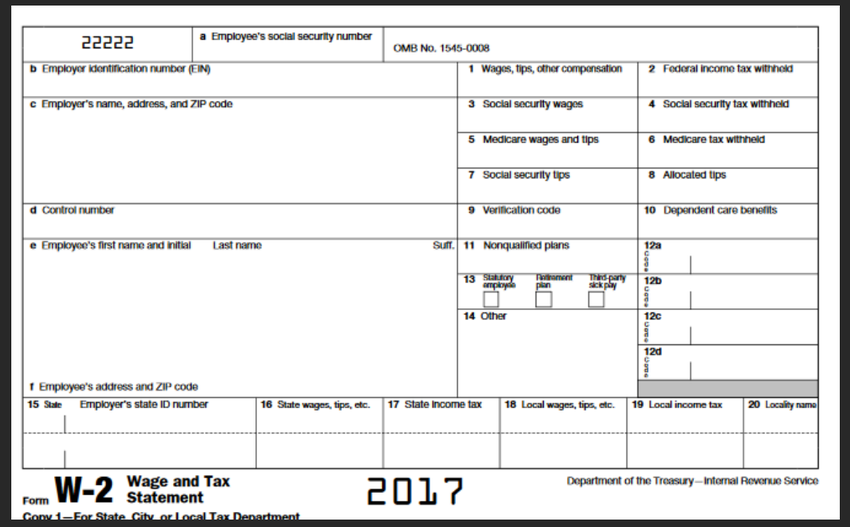
Gender:
- Male: True
- Female: False

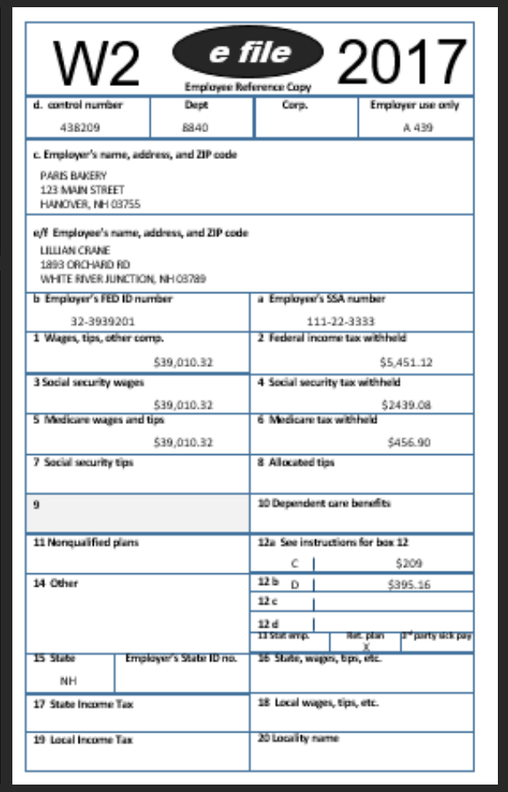
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

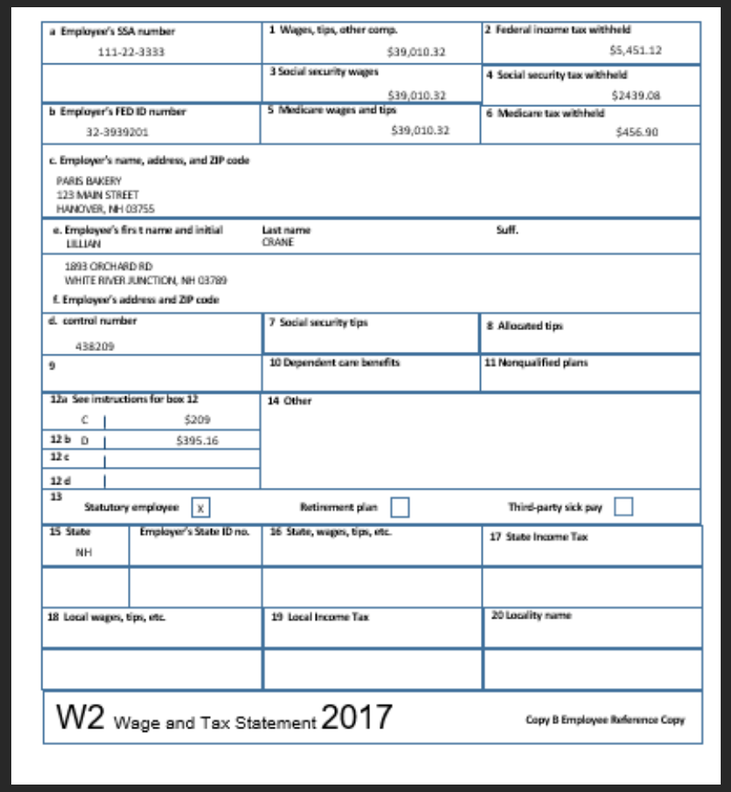
Beyond OCR: inferring key/value association

Detect structures of the same form without templates









© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Beyond OCR: Inferring key/value association

Key and value association

Name	Social Security #	Date of Birth	Relationship
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

John Doe	100 Main Street	Any town
Name	Address	City

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Beyond OCR: Inferring key/value association

Infer empty values

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

Output

Full Name:
- First: John
- Middle: null
- Last: Doe

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

Gender:
- Male: True
- Female: False

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

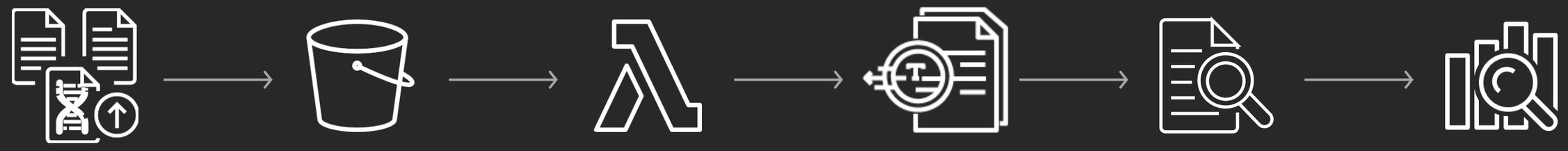
Amazon Textract
Reference Implementation

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Reference architecture—Index and search documents



Input
Uploaded document images such as tax forms, credit applications, or medical notes

Amazon S3
Uploaded documents are stored in data lake

AWS Lambda
A Lambda function is triggered to initiate document analysis using the Amazon Textract API

Amazon Textract
Automatically extract text, including key-value pairs and tables

Output
Perform contextual search on millions of documents or integrate data into your document management system

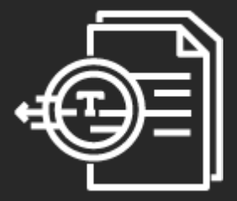
Amazon Elastisearch Service
Extracted data and confidence scores are indexed to enable document search

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Reference architecture—Form capture

Input

Customer uses mobile app to capture a photo of a W2 form



Amazon Textract

The Amazon Textract API is integrated into the end-user application to automatically extract text from the W2 form and auto-populate the form fields

Customer Application

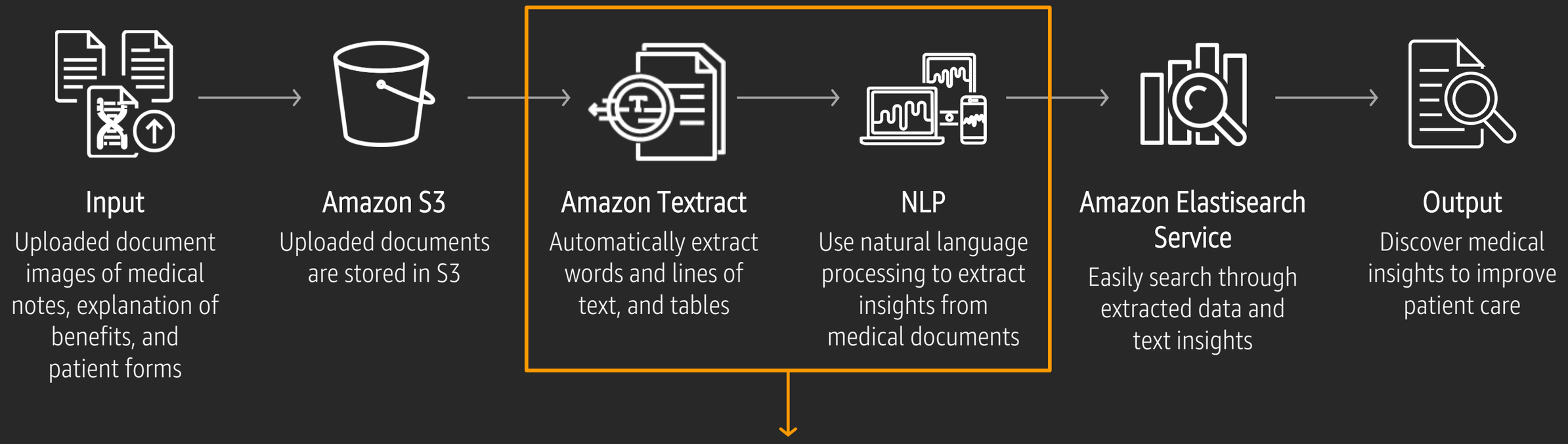
Customers experience real-time capture of their tax information by taking a photo instead of manual data entry

Database

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Reference architecture—Extract for NLP



Quickly turn extracted text/data into actionable insights

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract
Customers

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

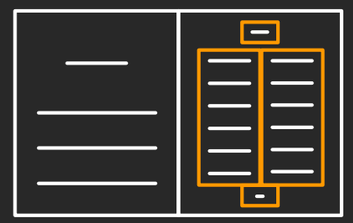
aws

Amazon Textract

Launch Customers

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract
Benefits



Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No ML Experience Required

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract

Pricing

Per 100 pages or images processed	Up to 1M pages/month	1M+ pages/month
Text Detection	$0.15	$0.06
Table Detection (Text Detection included)	$1.50	$1.00
Key-Value Detection (Text Detection included)	$5.00	$4.00
All	$6.50	$5.00

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract

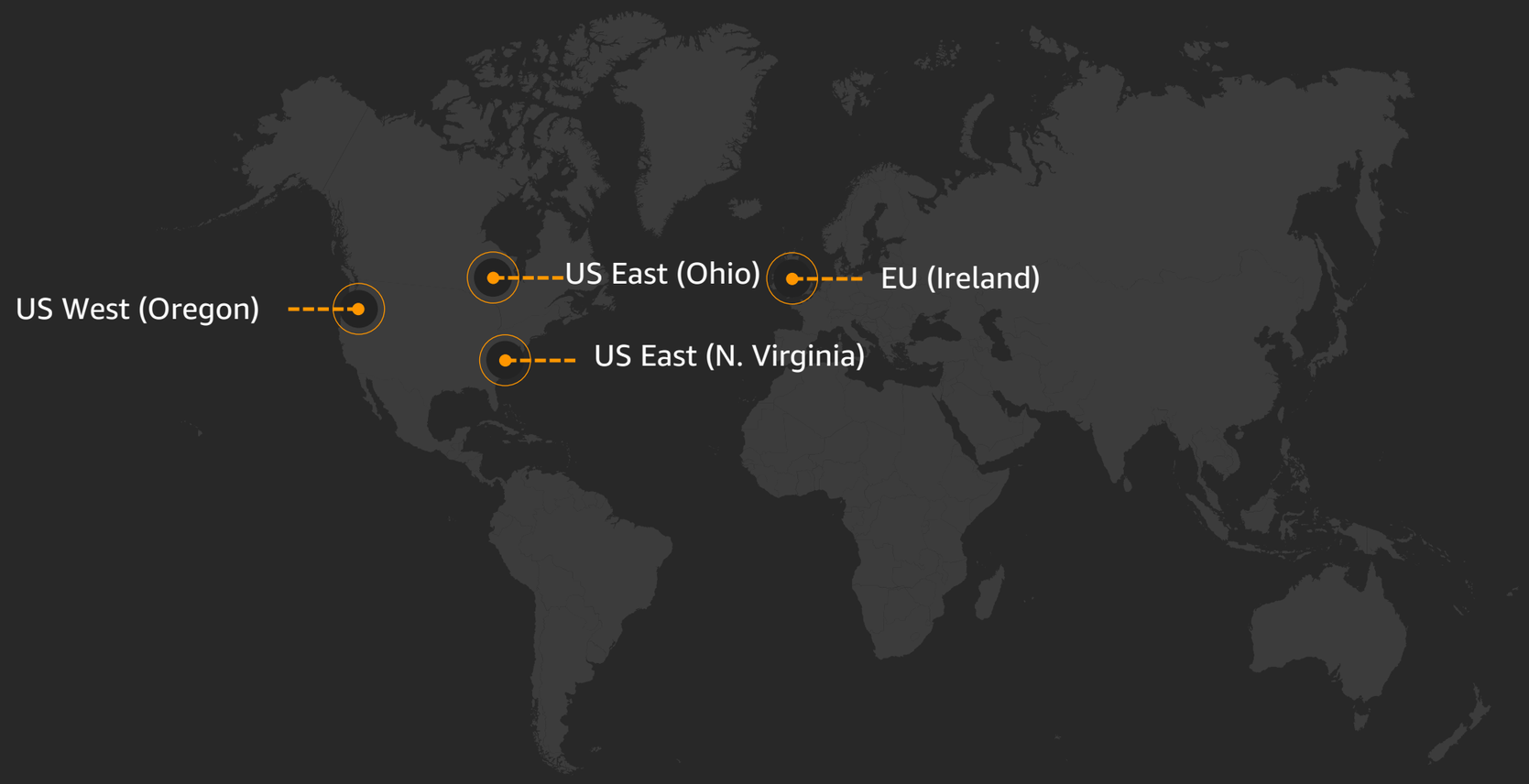
Free Tier

Features	Free for first three months
Text Detection	Up to 1,000 pages or images
Table Detection	Up to 100 pages or images
Key-Value Detection	
Text, Table, and Key-Value Detection	

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract

Regions



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract

Preview

To	Visit
LEARN MORE	
or	https://pages.awscloud.com/textract-preview.html
SIGN UP	

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark