Amazon Textract Overview

Hello!

I am Ian Mckay

Lead DevOps Engineer at Kablamo,
AWS APN Ambassador

@iann0036
ian.mckay@kablamo.com.au

General Overview

- Textract is *really good* OCR
- Upload your document bytes (or specify an S3 location) and get a JSON structure of your document
- Advanced form and table detection features (at a price)
- Available today in N. Virginia, Oregon, Ohio, Ireland and London regions

Form Data Extraction

- Detection of form-based (Field -> underline) data
- Returns key-value pairings and bounding boxes

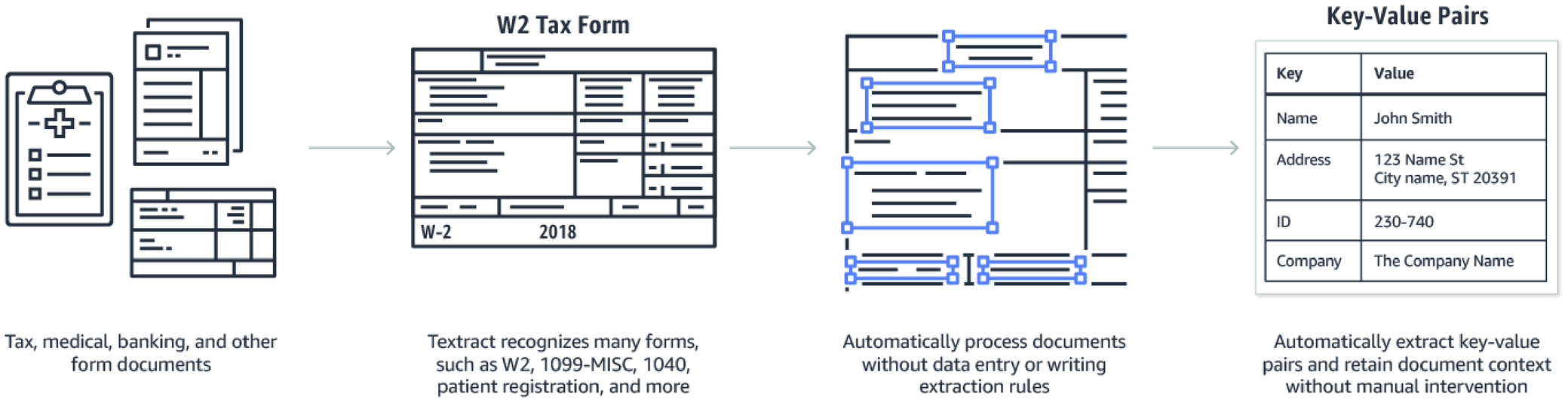


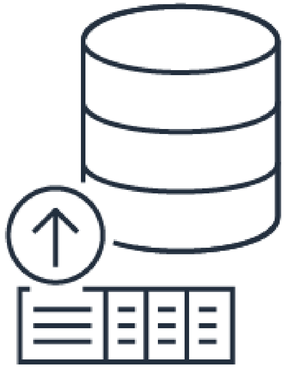
Table Data Extraction

- Detection of tabular data when held in enclosed grid
- Returns text data and row / column position and bounding boxes

When extracting text from documents and forms, Textract automatically detects and extracts structured data

Department	Budget	Actual	Difference
Accounting	$12,500.00	$11,293.12	$1,206.88
Finance	$24,000.00	$23,203.29	-$796.71
Human Resources	$13,000.00	$11,832.76	$1,167.24
Marketing	$82,500.00	$84,049.47	-$1.549.47
Sales	$76,000.00	$77,019.38	-$1,019.38

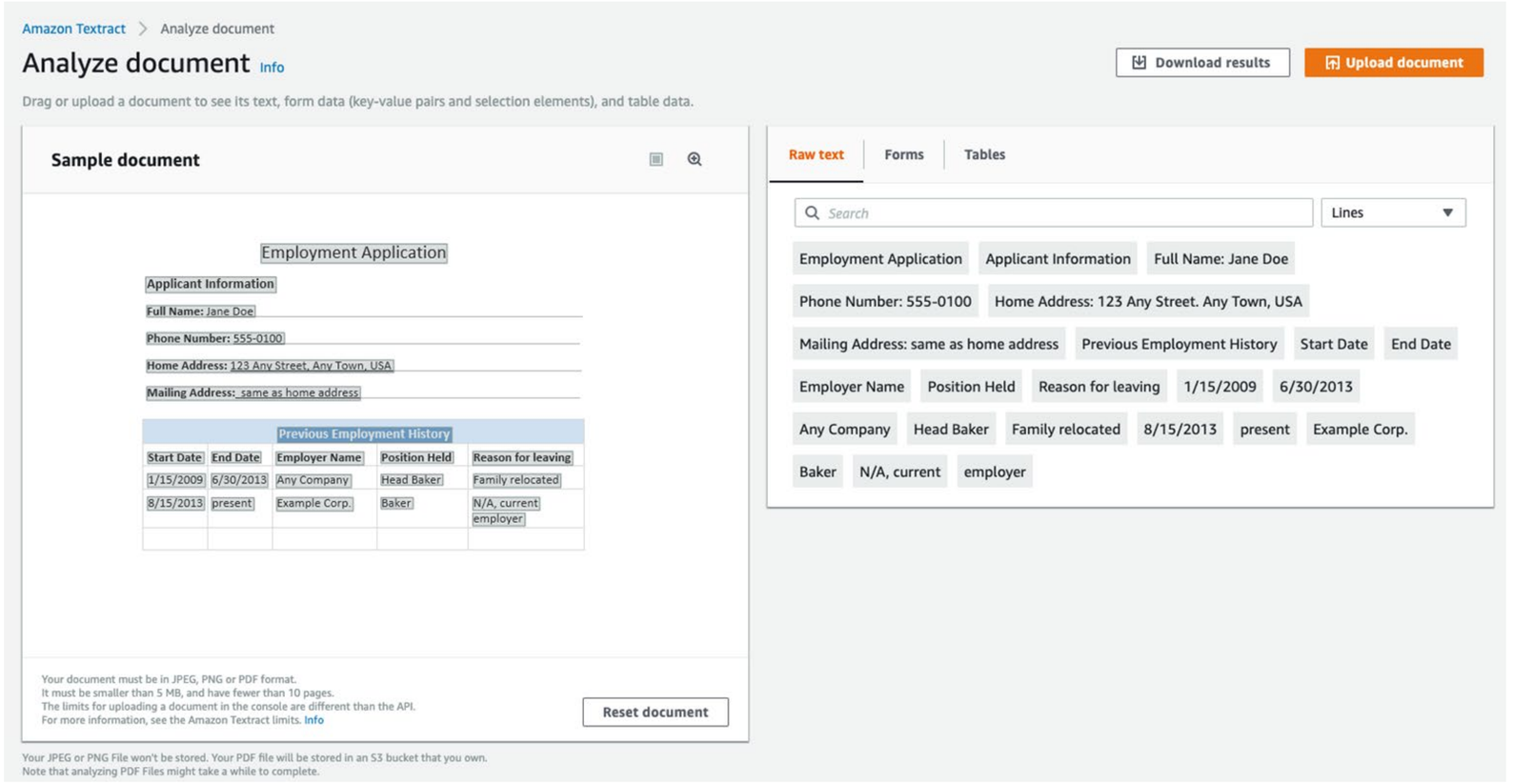
Textract preserves the tabular structure of extracted data, so that text remains grouped within each cell



With the tabular format of the data intact, easily upload extracted data into a database

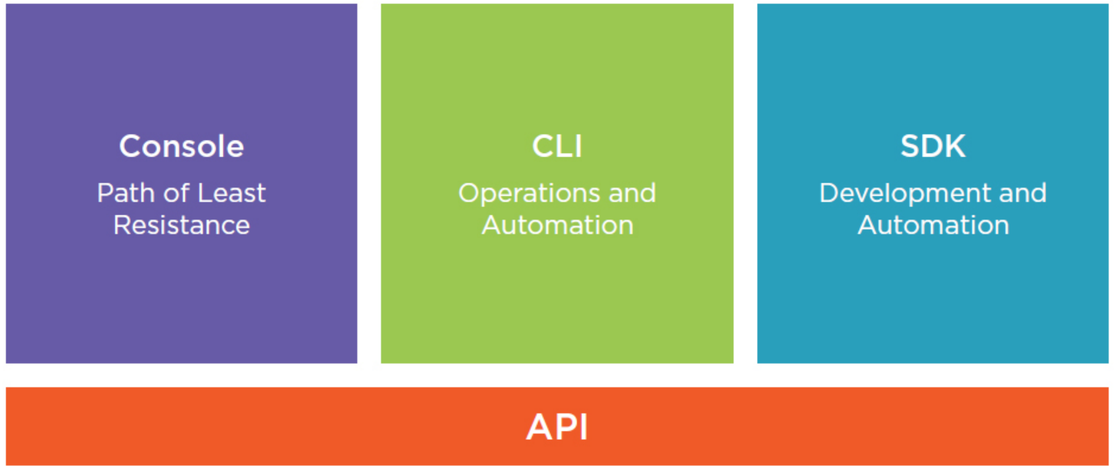
Usage (Console)

- Easy to test new document patterns
- Gets the API JSON results



Usage (CLIs / APIs / SDKs)

- Publicly available API, CLI and SDKs in most major programming languages
- Main calls are AnalyzeDocument (sync) and StartDocumentAnalysis / GetDocumentAnalysis (async)



JSON Format


```
{
 "DocumentMetadata": {
  "Pages": 1
 },
 "Blocks": [
  ...
  {
   "BlockType": "LINE",
   "Confidence": 99.9095458984375,
   "Text": "Phone Number: 555-0100",
   "Geometry": {
    "BoundingBox": {
     "Width": 0.2554837763309479,
     "Height": 0.025423433631658554,
     "Left": 0.056762322783470154,
     "Top": 0.30236420035362244
    },
    "Polygon": [
     {
      "X": 0.056762322783470154,
      "Y": 0.30236420035362244
     },
     {
      "X": 0.3122461140155792,
      "Y": 0.30236420035362244
     },
     {
      "X": 0.3122461140155792,
      "Y": 0.3277876377105713
     },
     {
      "X": 0.056762322783470154,
      "Y": 0.3277876377105713
     }
    ]
   },
   "Id": "ed701149-8ed2-4ae2-a641-6871a048d516",
   "Relationships": [
    {
     "Type": "CHILD",
     "Ids": [
      "16bba43f-2ce9-4965-b03d-d5456cb5e427",
      "784b7bcf-176a-4620-8dac-fc3110b56506",
      "960d5486-336c-455a-b4d3-f8ae2f10c32e"
     ]
    }
   ]
  },
  ...
}
```

Demos

- Console usage
- Data extraction from PDF with Boto3

Rough Edges

- Occasional wrong words (depends on source quality)
  - Filter on confidence scores
- 90° Rotated Text not detected
- Won’t detect single-row or single-column tables
- Does not support handwriting
- Badly contrasting colours have worse detection
- Checkmarks ‘X’ sometimes not detected
- English language only

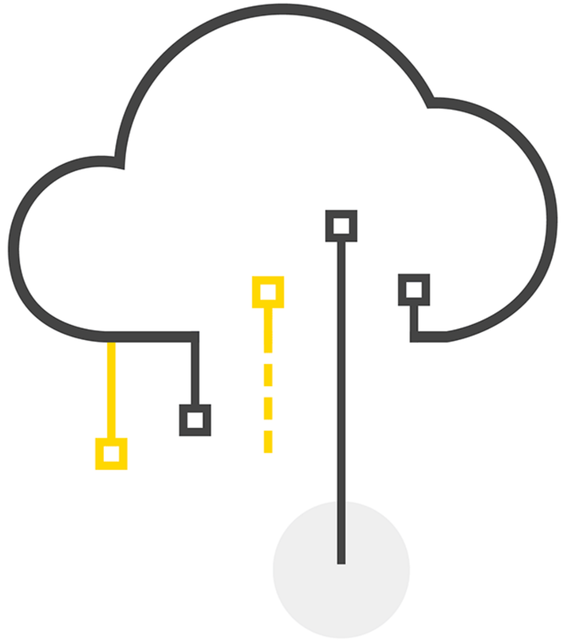
Pricing

	0-1M pages	1M+ pages
Text only	$1.50 USD / 1k pages	$0.60 USD / 1k pages
Text + Tables	$15.00 USD / 1k pages	$10.00 USD / 1k pages
Text + Forms	$50.00 USD / 1k pages	$40.00 USD / 1k pages
Text + Tables + Forms	$65.00 USD / 1k pages	$50.00 USD / 1k pages

Adding all features = 43x increase in price!

More Resources

- User Groups
- Slack
- Reddit / Twitter / LinkedIn
- Free AWS training and engagement days
- AWS Partner Network



aws partner network

Thanks!

Any questions?

github.com/iann0036/textract -demo

ian.mckay@kablamo.com.au

@iann0036