Visual - NLP Classic Use Cases

Alberto Andreotti
Head of Visual NLP (Data Scientist)

Alexander Branov
Data Scientist

Aymane Chilah
Data Scientist

Nitin Kumar
Scala Developer (Data Scientist)

Motivation

- We identified the strong need for a scalable solution.
- Diversity of input formats.
- Situation is more challenging than NLP.

STARBUCKS STORE #10208
11302 EUCLID AVENUE
CLEVELAND, OH (216) 229-0749
CHK 664290
12/07/2014 06:43 PM
1912003 DRAWER: 2. REG: 2
VT PEP MOCHA 4.95
SBUX CARD 4.95
XXXXXXXXXXXX3228
SUBTOTAL $4.95
TOTAL $4.95
CHANGE DUE $0.00
---- CHECK CLOSED
12/07/2014 06:43 PM
SBUX CARD X3228 NEW BALANCE: 37.45
CARD IS REGISTERED

- 111835b(JPEG) vs. 315b, a **355** factor!!.
- Density of information is much lower in OCR than NLP.
- Handling images is challenging.

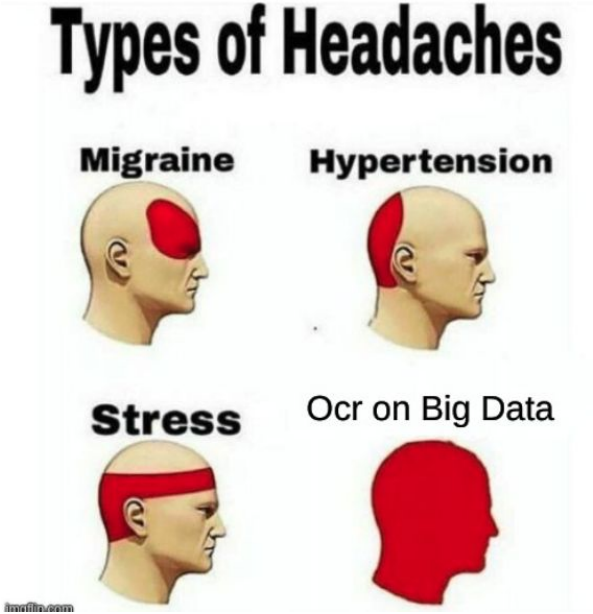
Motivation

John Snow LABS

- We provide two flavors of scalability;
  - Strong Scalability: you care about completion time of individual pieces.
  - Weak Scalability: you care about throughput.

- Checkpointing: you want to resume the computation.

- We want to solve all these problems so you don’t have to.



Introduction To Visual NLP

John Snow LABS

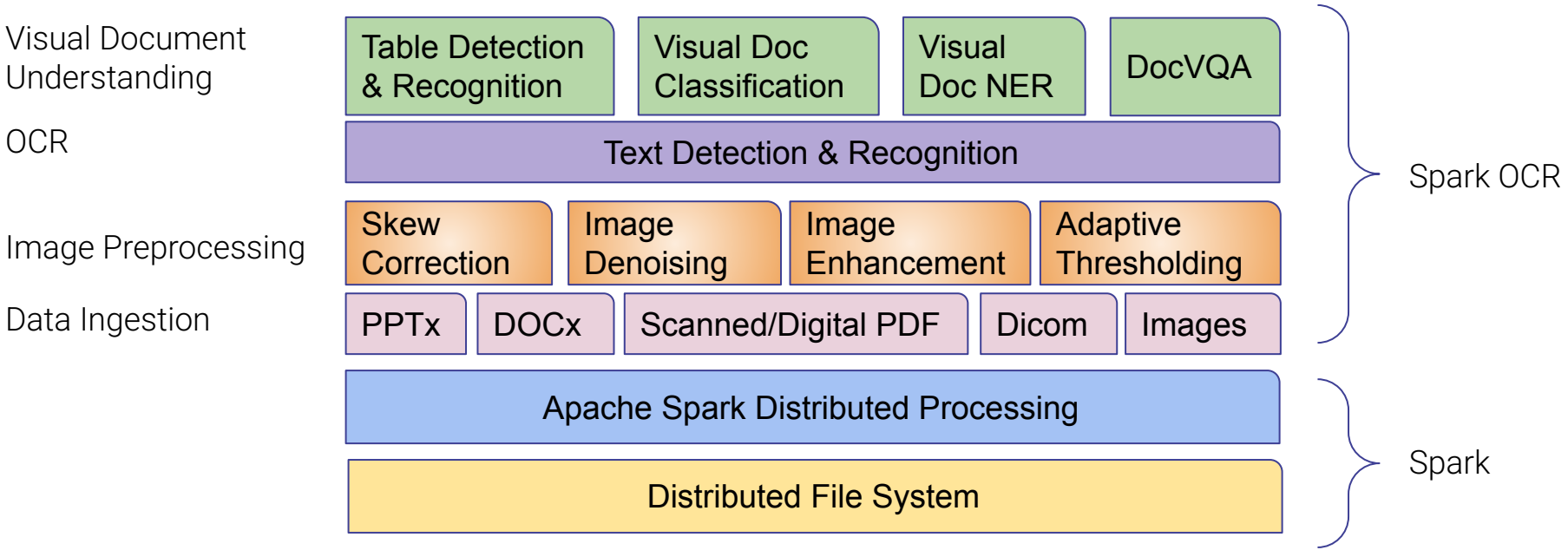
- Visual NLP is an OCR, and Visual Document Understanding library built on top of Apache Spark.
- Curated list of features -> only things that work.
- Optimized for performance and accuracy.
- Created by industry practitioners.
- Actively developed.
- Security minded.

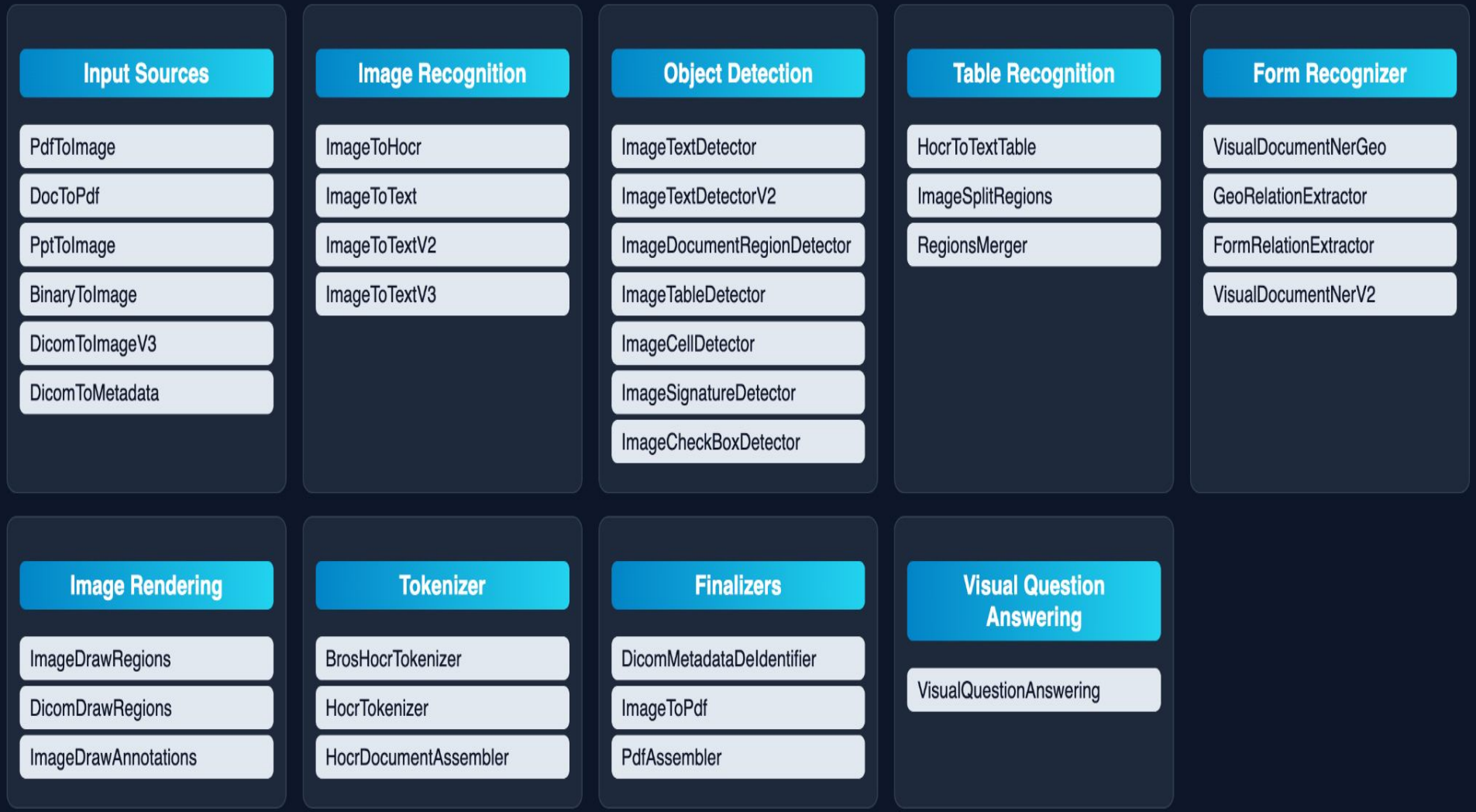




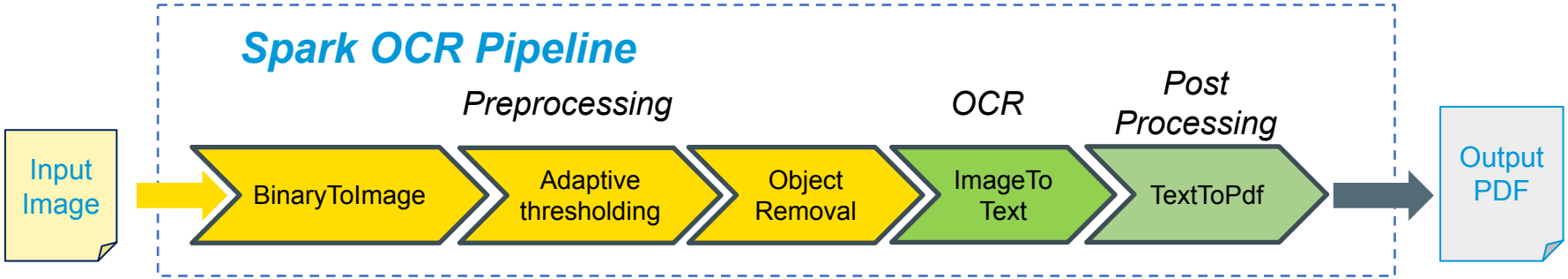


Visual NLP Architecture



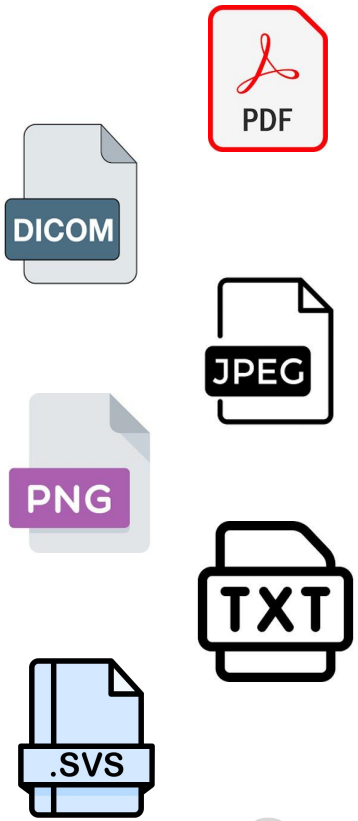

Input Sources
PdfToImage
DocToPdf
PptToImage
BinaryToImage
DicomToImageV3
DicomToMetadata
Image Recognition
ImageToHocr
ImageToText
ImageToTextV2
ImageToTextV3
Object Detection
ImageTextDetector
ImageTextDetectorV2
ImageDocumentRegionDetector
ImageTableDetector
ImageCellDetector
ImageSignatureDetector
ImageCheckBoxDetector
Table Recognition
HocrToTextTable
ImageSplitRegions
RegionsMerger
Form Recognizer
VisualDocumentNerGeo
GeoRelationExtractor
FormRelationExtractor
VisualDocumentNerV2
Image Rendering
ImageDrawRegions
DicomDrawRegions
ImageDrawAnnotations
Tokenizer
BrosHocrTokenizer
HocrTokenizer
HocrDocumentAssembler
Finalizers
DicomMetadataDeIdentifier
ImageToPdf
PdfAssembler
Visual Question Answering
VisualQuestionAnswering

Sample Visual NLP workflow



Input Types : Entry Points

- **Images - (PNG, GIFs, JPEG, etc..)**
  - BinaryToImage
    - Input : BinaryContent
    - Output: Image Object
    - Use Cases : Text Detection, OCR, De-Identification, Obfuscation
- **PDF - (Digital / Scanned)**
  - PdfToImage
    - Input : BinaryContent
    - Output: Image Object
    - Use Cases : De-Identification, Obfuscation
- **Dicom**
  - DicomToImageV3
    - Input : BinaryContent
    - Output : Image Object
    - Use Cases : Medical De-Identification
  - DicomToMetadata
    - Input : BinaryContent
    - Output : Dicom Tags
    - Use Cases : Tag De-Identification
- **SVS**
  - Pure python implementation
    - Input : SVS
    - Output : SVS
    - Use Case : PHI De-Identification



PDF Processing Stages

- **PdfToImage**: Converts each page of a PDF into an image.
- **PdfToForm**: Extracts key–value structured data (form fields) from PDFs.
- **PdfToText**: Retrieves embedded text from digitally generated (non-scanned) PDFs.
- **PdfToHocr**: Generates OCR text in HOCR format with layout, page number, and coordinates.
- **PdfToTextTable**: Extracts tables from PDF reports page by page, producing structured table data with text chunk coordinates.
- **PdfDrawRegions**: Draws or overlays regions on the existing PDF.
- **ImageToTextPdf**: Recognizes text from embedded images and renders the recognized text on top of the original PDF pages.
- **PdfAssembler** / **ImageToPdf**: Combines multiple single-page PDFs or rendered images into a multi-page PDF document.

Image Transformations

John Snow LABS

CPU

ImageTransformer

- Erosion
- Dilation
- Scaling
- Otsu Thresholding
- Adaptive Thresholding
- Median Blur
- Blur
- Remove Objects

GPU

GPUImageTransformer

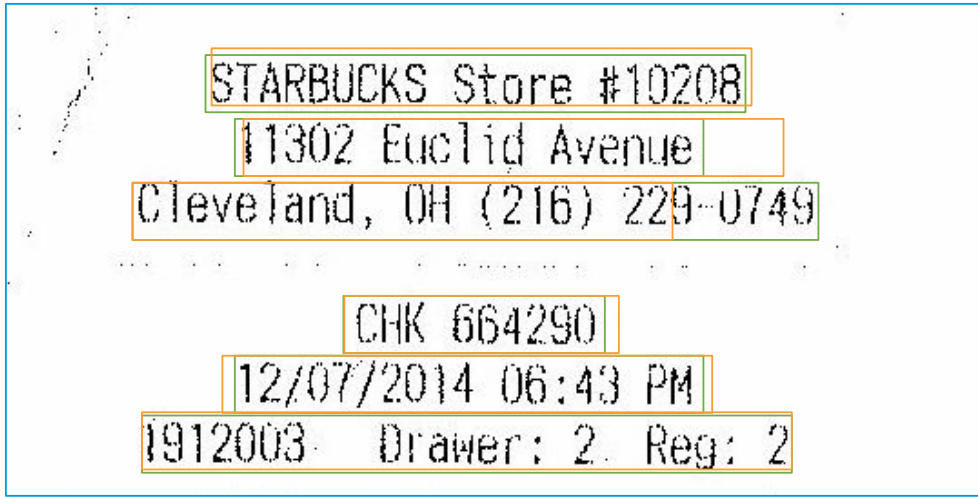
- Erosion
- Dilation
- Scaling
- Otsu Thresholding
- Huang Thresholding

Example Notebook : Visual_01_Image_Processing.ipynb

Image Text Detection

- ImageTextDetectorV2
  - Implementation : Python
  - Input : Image Object
  - Output : Regions
  - Hardware : CPU/GPU

- ImageTextDetector
  - Implementation : Scala
  - Input : Image Object
  - Output : Regions
  - Hardware : CPU/GPU
  - Average 25% Speed up on GPU Hardware



Example Notebook : Visual_02_Text_Detection.ipynb

Image Text Recognition

- ImageToText
  - Implementation : Scala
  - Input : Image Object
  - Output : Text
  - Hardware : CPU
  - Support for Multiple Languages
  - Faster ; Low accuracy for noised images
  - Case Sensitive

- ImageToTextV2
  - Implementation : Scala
  - Input : Image Object, Image Crop Regions
  - Output : Text
  - Handwritten and Printed Text Recognition
  - External Text Detector is required
  - Hardware : CPU/GPU
  - Accuracy & Speed Benchmarking Article

- ImageToTextV3
  - Implementation : Scala
  - Input : Image Object, Image Crop Regions
  - Output : Text
  - Hardware : CPU/GPU
  - External Text Detector is required
  - Case Sensitive

Example Notebook : Visual_03_Text_Recognition.ipynb

Image Text Recognition : Benchmarks

John Snow LABS

Type	Dataset	Model	Caching	CPU			GPU		
				TimeTaken (seconds)	Average Time	CER	TimeTaken (seconds)	Average Time	CER
Printed	FUNSD	ocr_base_printed_v2	No	1709.24	0.01514	0.07333	405.61	0.00359	0.07333
		ocr_base_printed_v2_opt	No	1608.29	0.01425	0.07351	1072.87	0.00951	0.07351
		ocr_large_printed_v2	No	3216.36	0.02850	0.06786	544.25	0.00482	0.06786
		ocr_large_printed_v2_opt	No	3505.75	0.03106	0.06787	1245.23	0.01103	0.06787
		ocr_base_printed_v2	Yes	1559.14	0.01381	0.07333	351.01	0.00311	0.07333
		ocr_base_printed_v2_opt	Yes	1348.17	0.01194	0.07371	732.45	0.00649	0.07371
		ocr_large_printed_v2	Yes	2857.73	0.00771	0.06786	518.96	0.00460	0.06786
		ocr_large_printed_v2_opt	Yes	3111.66	0.02757	0.06773	904.11	0.00801	0.06773
	SROIE	ocr_base_printed_v2	No	3182.52	0.00823	0.01471	867.77	0.00224	0.01471
		ocr_base_printed_v2_opt	No	2839.97	0.00734	0.01489	2410.47	0.00643	0.01489
		ocr_large_printed_v2	No	5440.10	0.01406	0.01428	1093.20	0.00283	0.01428
		ocr_large_printed_v2_opt	No	5063.07	0.01309	0.01463	3111.29	0.00804	0.01463
		ocr_base_printed_v2	Yes	2108.02	0.00545	0.01470	810.26	0.00209	0.01470
		ocr_base_printed_v2_opt	Yes	1907.55	0.00493	0.01488	1608.62	0.00493	0.01488
		ocr_large_printed_v2	Yes	4311.84	0.00330	0.01428	993.43	0.00257	0.01428
		ocr_large_printed_v2_opt	Yes	3864.54	0.00999	0.01434	2035.05	0.00526	0.01434
Handwritten	IAM	ocr_base_handwritten_v2	No	2094.36	0.00372	0.06478	617.20	0.0013	0.06478
		ocr_base_handwritten_v2_opt	No	1885.96	0.00335	0.06548	1588.70	0.0031	0.06548
		ocr_large_handwritten_v2	No	3153.85	0.00560	0.04645	1492.10	0.0027	0.04645
		ocr_large_handwritten_v2_opt	No	3109.64	0.04529	0.04502	2669.50	0.0039	0.04502
		ocr_base_handwritten_v2	Yes	1412.95	0.00251	0.06478	390.72	0.0008	0.06478
		ocr_base_handwritten_v2_opt	Yes	1304.18	0.00232	0.06566	1076.24	0.0023	0.06566
		ocr_large_handwritten_v2	Yes	2281.14	0.00405	0.04645	363.07	0.0006	0.04645
		ocr_large_handwritten_v2_opt	Yes	2137.57	0.00380	0.04487	1325.86	0.0132	0.04487

De-Identification/Obfuscation

- De-Identification is supported for Image, PDF, Dicom and SVS input formats.
- Obfuscation is supported for Image and PDF.

Healthcare Recovery Trauma Center

Kimberly Lawrence 24/05/1977
Sierra Valley Medical Institute INC

Patient Summary

Kimberly Lawrence, born on 24/05/1977, is a 46-year-old Female. Her medical journey indicates a diagnosis of Type 2 Diabetes Mellitus accompanied by Peripheral Neuropathy. Current management focuses on glycemic control through medication and lifestyle adjustments, with regular monitoring for neuropathy progression. Overall, the patient health status is stable with ongoing management of chronic conditions.

Patient Demographics

Name: Kimberly Lawrence

DOB: 24/05/1977

Age: 46

Sex: Female

SSN: 567-45-5412

Hospital ID: HOSP26508961

Patient Lifestyle

Smoking Status: Never

Alcohol Consumption: Rarely

Diet Preference: Diabetic-friendly

Exercise Habits: Mild

Patient Vitals

Heart Rate: 72

Respiratory Rate: 16

Temperature Celsius: 36.7

Oxygen Saturation Percent: 98

Blood Pressure: 130/85 mmHg

Recorded Date: 15/11/2024

Doctor Information

Doctor Name: Cheryl Blankenship

Doctor Unique ID: DR14144B

Sierra Valley Medical Institute INC (402) 738-5912 22/04/2025

Healthcare Recovery Trauma Center

[redacted]

Patient Summary

[redacted]orn on [redacted], is a [redacted]year-old Female. Her medical journey indicates a diagnosis of Type 2 Diabetes Mellitus accompanied by Peripheral Neuropathy. Current management focuses on glycemic control through medication and lifestyle adjustments, with regular monitoring for neuropathy progression. Overall, the patient health status is stable with ongoing management of chronic conditions.

Patient Demographics

Name: [redacted]

DOB: [redacted]

Age: [redacted]

Sex: Female

SSN: [redacted]

Hospital ID: [redacted]

Patient Lifestyle

Smoking Status: Never

Alcohol Consumption: Rarely

Diet Preference: Diabetic-friendly

Exercise Habits: Mild

Patient Vitals

Heart Rate: 72

Respiratory Rate: 16

Temperature Celsius: 36.7

Oxygen Saturation Percent: 98

Blood Pressure: 130/85 mmHg

Recorded Date: [redacted]

Doctor Information

Doctor Name: [redacted]

Doctor Unique ID: [redacted]

[redacted] [redacted] [redacted]

Healthcare Recovery Trauma Center

Hortense Heinrich 26/05/1977
REHABILITATION HOSPITAL OF SAVANNAH

Patient Summary

Hortense Heinrich orn on 26/05/1977 is a 51 ear-old Female. Her medical journey indicates a diagnosis of Type 2 Diabetes Mellitus accompanied by Peripheral Neuropathy. Current management focuses on glycemic control through medication and lifestyle adjustments, with regular monitoring for neuropathy progression. Overall, the patient health status is stable with ongoing management of chronic conditions.

Patient Demographics

Name: Hortense Heinrich

DOB: 26/05/1977

Age: 51

Sex: Female

SSN: 789-67-7634

Hospital ID: SDNM48720183

Patient Lifestyle

Smoking Status: Never

Alcohol Consumption: Rarely

Diet Preference: Diabetic-friendly

Exercise Habits: Mild

Patient Vitals

Heart Rate: 72

Respiratory Rate: 16

Temperature Celsius: 36.7

Oxygen Saturation Percent: 98

Blood Pressure: 130/85 mmHg

Recorded Date: 17/11/2024

Doctor Information

Doctor Name: Joseph Friedlander

Doctor Unique ID: IE36306Q

REHABILITATION HOSPITAL OF SAVANNAH (624) 950-7134 24/04/2025

PDF De-Identification Dataset

John Snow LABS

- Fully synthetic dataset of 50 medical-style PDF documents generated using Faker and GEMINI API.
  - Divided into Easy (30), Medium (10), and Hard (10) levels with increasing layout complexity and noise.
  - Contains PHI in multiple layouts (header, footer, free-text, key-value, and tables)
  - Contains an average of 40–50 PHI entities per document, mainly Dates, Names, Gender, Age, Phone Number, SSN and Hospital/Organization identifiers.

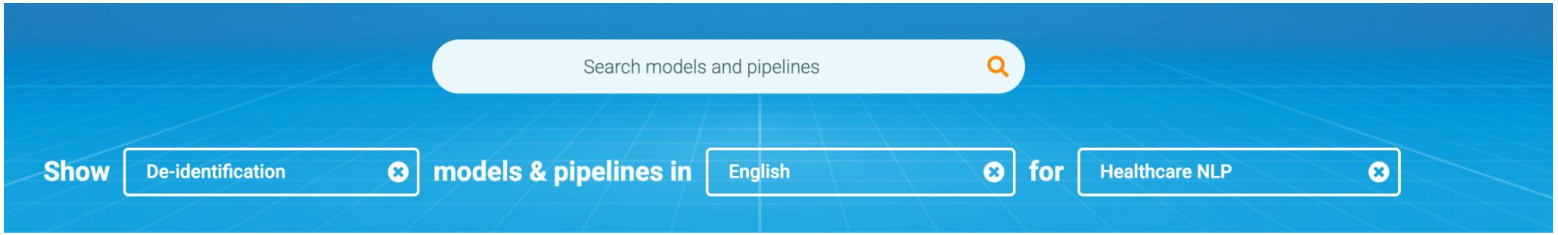
- Designed for benchmarking NER solutions on PHI de-identification tasks.

- Performed two experiments: Complete De-Identification and Zero-Shot De-Identification.

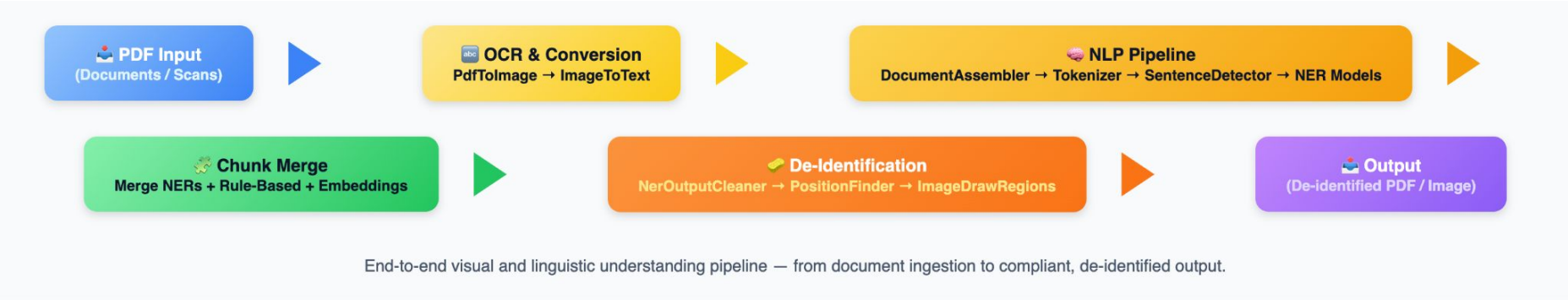
Difficulty Level	Precision	Recall	F1-Score	Total Files
Easy	0.9851	0.9799	0.9825	30
Medium	0.9800	0.9575	0.9686	40
Zero Shot Medium	0.9861	1.0000	0.9930	10
Hard	0.9561	0.9290	0.9424	50

Anonymization Pipeline Builder

- Unified framework for **De-Identification** and **Obfuscation** tasks.
- Supports Image, PDF, Dicom Input Sources.
- Compatible with multiple OCR versions ImageToText, ImageToTextV2 and ImageToTextV3.
- Provides configurable parameters for individual stage behavior, thresholds, and entity matchers.
- Generates outputs in **PDF**, **Image**, **Dicom** which destroys the intermediate results as its an aggregation stage, or returns **Dataframe** which keeps intermediate stages.
- All Healthcare De-Identification Pipelines are supported, so users can experiment with all variations.
    - Go To Models Hub
    - Select **De-Identification** option in **Show** Drop Down
    - Select Language **EN** or any other in **Language** Drop Down
    - Select **All Healthcare NLP Version** in **For** Drop Down



De-Identification/Obfuscation



- The core workflow is common across all inputs only the Input/Output stages change by file type (PDF, DICOM, Image).
- The process flows from Text Extraction → Healthcare NLP → Visual Rendering to identify and remove PHI.
- Obfuscation replaces PHI text, while Image Inpainting optionally improves visual quality of the output.

- De-Identification & Obfuscation Pretrained Pipelines are available in Models Hub.
    - De-Identification pretrained pipelines come in two variations with and without signature removal.
    - Obfuscation pretrained pipelines do not include signature removal.
- PDF De-Identification solutions are available on Amazon Sagemaker Endpoints.
    - Refer to Slide 26 for PDF De-Identification and Obfuscation Links
- Any PDF De-Identification pretrained pipeline can be converted for image-based de-identification using the example provided in the De-Identification Notebook.

Example Notebook : Visual_04_AnonymizationPipelineBuilder.ipynb

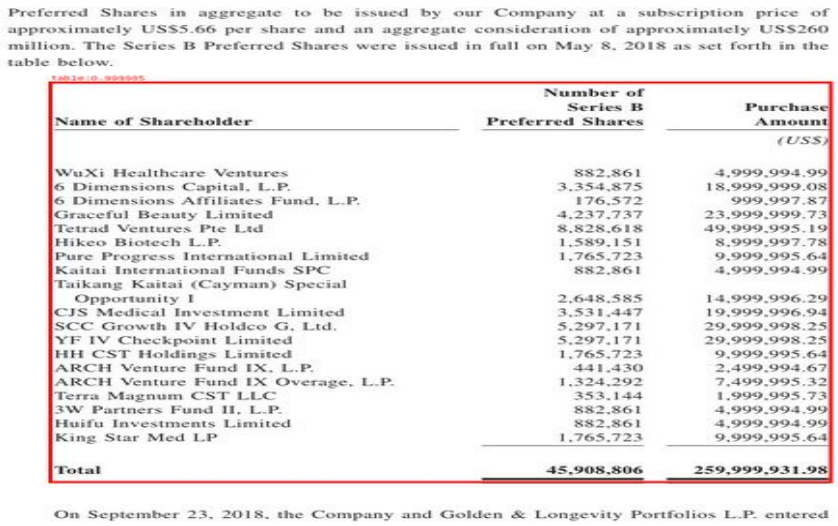
Example Notebook : Visual_05_DeIdentification.ipynb

Example Notebook : Visual_06_Obfuscation.ipynb

Table Detection & Recognition

- Table present in documents can have some metadata about their positions, contents and extracting those tables is a matter of reverse engineering. But if those tables are present as images then we will need to apply Computer Vision and OCR and complex algorithms to extract the accurate text from the image.

- General Workflow:
    - Table Detection Identify and isolate regions in the image where tables are present.
    - Cell Detection and HOCR Generation for each detected table region, perform cell segmentation and generate HOCR to capture text layout and structure.
    - Table Reconstruction combine the detected cell regions with the HOCR output to rebuild the final structured table representation.



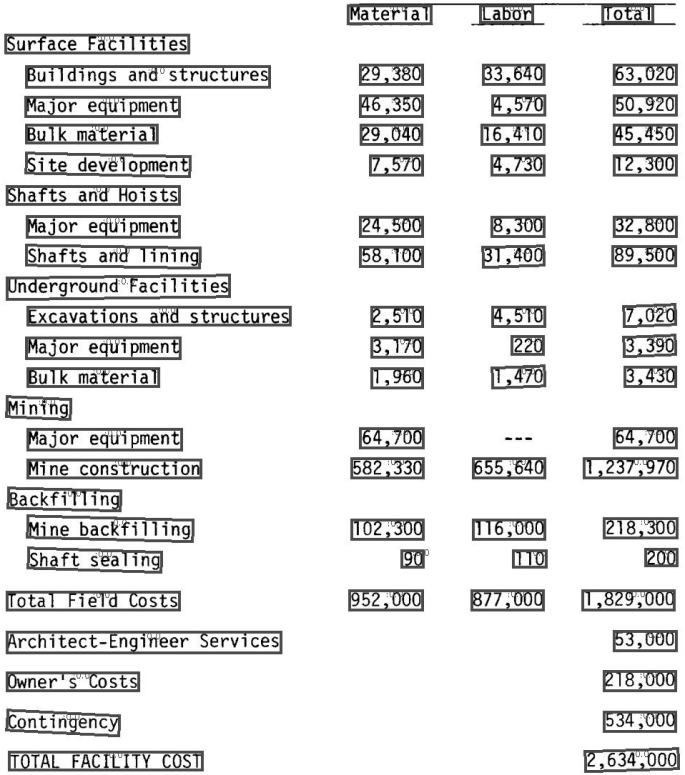
Preferred Shares in aggregate to be issued by our Company at a subscription price of approximately US$5.66 per share and an aggregate consideration of approximately US$260 million. The Series B Preferred Shares were issued in full on May 8, 2018 as set forth in the table below.

Name of Shareholder	Number of Series B Preferred Shares	Purchase Amount (US$)
WuXi Healthcare Ventures	882,861	4,999,994.99
6 Dimensions Capital, L.P.	3,354,875	18,999,999.08
6 Dimensions Affiliates Fund, L.P.	176,572	999,997.87
Graceful Beauty Limited	4,237,737	23,999,999.73
Tetrad Ventures Pte Ltd	8,828,618	49,999,995.19
Hikeo Biotech L.P.	1,589,151	8,999,997.78
Pure Progress International Limited	1,765,723	9,999,995.64
Kaitai International Funds SPC	882,861	4,999,994.99
Taikang Kaitai (Cayman) Special Opportunity I	2,648,585	14,999,996.29
CJS Medical Investment Limited	3,531,447	19,999,996.94
SCC Growth IV Holdco G, Ltd.	5,297,171	29,999,998.25
YF IV Checkpoint Limited	5,297,171	29,999,998.25
HH CST Holdings Limited	1,765,723	9,999,998.64
ARCH Venture Fund IX, L.P.	441,430	2,499,994.67
ARCH Venture Fund IX Overage, L.P.	1,324,292	7,499,995.32
Terra Magnum CST LLC	353,144	1,999,995.73
3W Partners Fund II, L.P.	882,861	4,999,994.99
Huifu Investments Limited	882,861	4,999,994.99
King Star Med LP	1,765,723	9,999,995.64
Total	45,908,806	259,999,931.98

On September 23, 2018, the Company and Golden & Longevity Portfolios L.P. entered

Example Notebook : Visual_07_Table_Recognition.ipynb

Metrics: Table Detection and Recognition



	Material	Labor	Total
Surface Facilities			
Buildings and structures	29,380	33,640	63,020
Major equipment	46,350	4,570	50,920
Bulk material	29,040	16,410	45,450
Site development	7,570	4,730	12,300
Shafts and Hoists			
Major equipment	24,500	8,300	32,800
Shafts and lining	58,100	31,400	89,500
Underground Facilities			
Excavations and structures	2,510	4,510	7,020
Major equipment	3,170	220	3,390
Bulk material	1,960	1,470	3,430
Mining			
Major equipment	64,700	---	64,700
Mine construction	582,330	655,640	1,237,970
Backfilling			
Mine backfilling	102,300	116,000	218,300
Shaft sealing	90	110	200
Total Field Costs	952,000	877,000	1,829,000
Architect-Engineer Services			53,000
Owner's Costs			218,000
Contingency			534,000
TOTAL FACILITY COST			2,634,000

- *generate a list of all adjacency relations between each content cell and its nearest horizontal and vertical neighbours.*
- *An adjacency relation is a tuple containing the textual content of both cells, the direction and the number of blank cells (if any) in between.*
- *This 1-D list of adjacency relations can be compared to the ground truth by using precision and recall measures.*
- *Example: (Material, Labor), (Labor, Total), (Surface Facilities, Building and Structures), etc.*

Metrics: Table Detection and Recognition

John Snow LABS

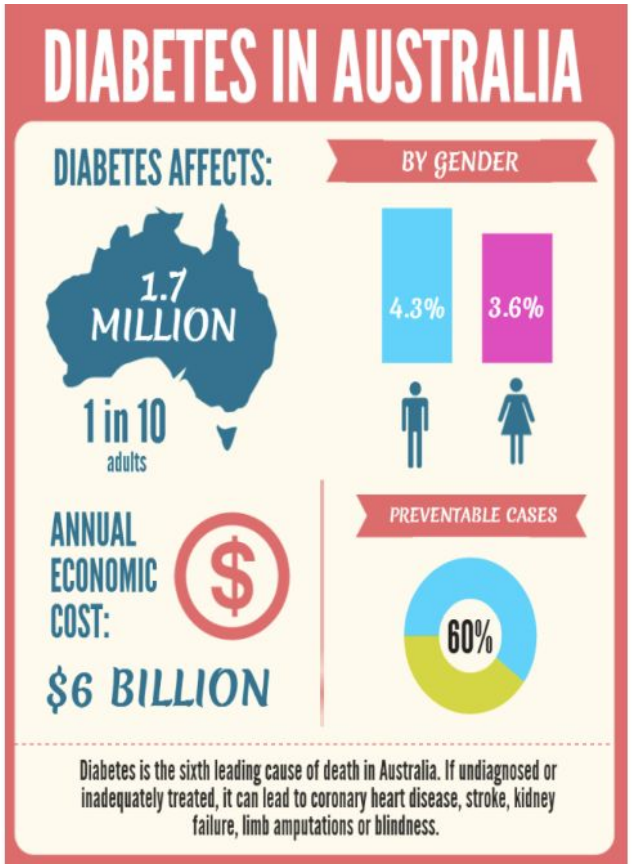
Table Detection Model Comparison

Model	general_model_table_detection_v2 (0.5)	general_model_table_detection_v2 (0.8)	table_detection_v3 (0.5)	table_detection_v3 (0.8)
IOU @ 0.6 precision	0.93394	0.93995	0.96637	0.98140
IOU @ 0.6 recall	0.91314	0.90646	0.95991	0.93987
IOU @ 0.6 f1	0.92342	0.92290	**0.96313**	**0.96018**
IOU @ 0.7 precision	0.91116	0.91917	0.96188	0.97674
IOU @ 0.7 recall	0.89087	0.88641	0.95546	0.93541
IOU @ 0.7 f1	0.90090	0.90249	**0.95866**	**0.95563**

OCR Model Comparison

Model	JSL image2text	AWS Textract	JSL image2textV2 with regions merger 5.4.0	JSL image2text 5.4.0
IOU @ 0.6 precision	0.74357	0.77286	0.50647	0.72612
IOU @ 0.6 recall	0.68153	0.65669	0.49873	0.73121
IOU @ 0.6 f1	0.71120	0.71006	0.50257	**0.72866**
IOU @ 0.7 precision	0.64142	0.65244	0.27167	0.59709
IOU @ 0.7 recall	0.58790	0.55414	0.26752	0.60127
IOU @ 0.7 f1	**0.61349**	0.59917	0.26957	0.59917

Visual Question Answering



- Extracts and interprets not only the textual content whether handwritten, typewritten, or printed but also a wide range of visual cues such as page layout and structure (including forms and tables), non-textual components (like marks, checkboxes, separators, and diagrams), and stylistic features such as fonts, colors, and highlights.

How many females are affected by diabetes? 3.6%

What percentage of cases can be prevented? 60%

What could lead to blindness or stroke? Diabetes

What is the Annual Economic Cost (in billion) ? 6

Example Notebook : Visual_08_Question_Answering.ipynb

Medical Assistant

John Snow LABS

- It works on Databricks with usage of Databricks Container Service.
- It has multimodal LLM under the hood.
- It may be used to extract any medical info from docs, especially prescriptions and medical test.
- It can process handwritten text.
- It is possible to set separate questions/prompt for each image or use one question for all.
- Our approach supports Sparks and Databricks strong features like scalability.

Extracted Medical Tests (JSON Format)

Prompt: Extract medical tests with their attributes and return the result as JSON. If no tests are found, return an empty JSON.


```
{
  "tests": [
    {"name": "BILIRUBIN, TOTAL", "reference_range": "0.2-1.2 mg/dL", "value": 0.6},
    {"name": "ALKALINE PHOSPHATASE", "reference_range": "31-125 U/L", "value": 33},
    {"name": "AST", "reference_range": "10-30 U/L", "value": 14},
    {"name": "ALT", "reference_range": "6-29 U/L", "value": 9},
    {"name": "CHOLESTEROL, TOTAL", "reference_range": "<200 mg/dL", "value": 166},
    {"name": "HDL CHOLESTEROL", "reference_range": "≥50 mg/dL", "value": 62},
    {"name": "TRIGLYCERIDES", "reference_range": "<150 mg/dL", "value": 46},
    {"name": "LDL-CHOLESTEROL", "unit": "mg/dL(calc)", "value": 90},
    {"name": "CHOL/HDLC RATIO", "reference_range": "<5.0(calc)", "value": 2.7},
    {"name": "NON HDL CHOLESTEROL", "reference_range": "<130 mg/dL(calc)", "value": 104},
    {"name": "HEMOGLOBIN A1c", "reference_range": "<5.7% of total Hgb", "value": 5.0},
    {"name": "MAGNESIUM", "reference_range": "1.5-2.5 mg/dL", "value": 2.2}
  ]
}
```


Example Notebook : Visual_09_Medical_Assistant.ipynb

Light Pipelines

Light Pipeline

Create LightPipeline

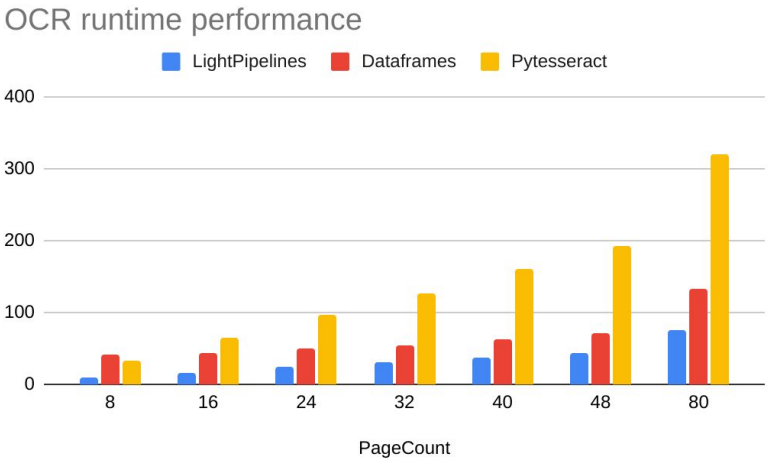
```
from sparkocr.base import LightPipeline

lp = LightPipeline(pipeline)

%%time
lp.fromLocalPath(pdfs_path)
```

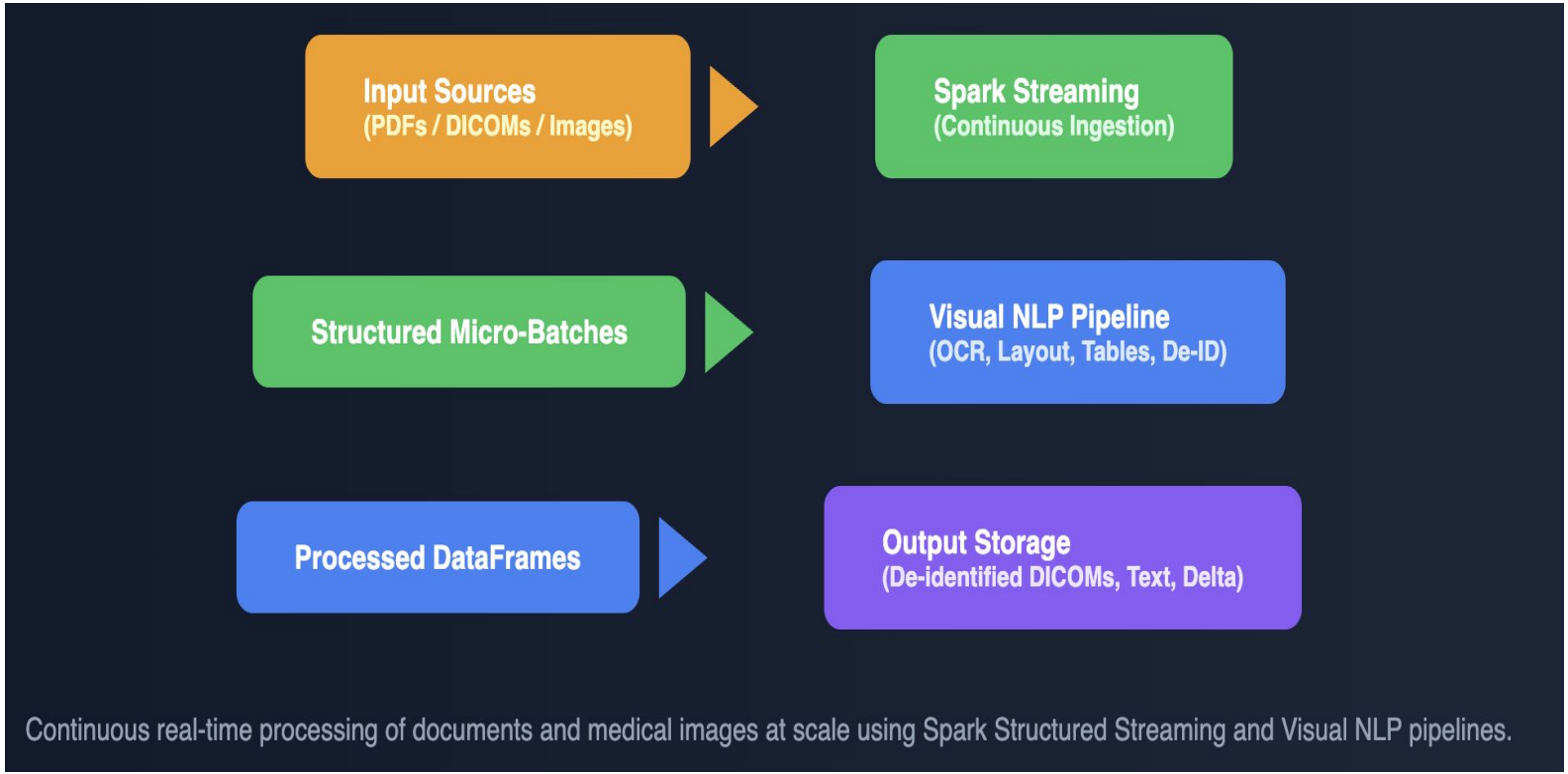
```
CPU times: user 4.58 ms, sys: 6.95 ms, total: 11.5 ms
Wall time: 6.24 s
```

Spark ML pipeline



Example Notebook : Visual_10_Light_Pipeline.ipynb

Streaming



Example Notebook : Visual_11_Dicom_Streaming.ipynb

Visual NLP Sagemaker Listing

John Snow LABS

- Clinical De-identification for PDF (EN) – Link
- Clinical Obfuscation for PDF (EN) – Link
- Extract Digital Text from Handwritten content – Link
- PHI leakage detection for DICOM – Link
- DICOM Images De-identification Full – Link
- DICOM Images De-identification Base – Link
- DICOM Images De-identification Alias – Link
- SVS Images De-identification – Link

Next Steps

John Snow LABS

- Add compatibility for small vision-language models in GGUF format, allowing seamless execution directly on Spark without taking data out from the DataFrame.
- Improve support for running advanced models like Medical Assistant VLMs across diverse compute environments and deployment platforms.
- Bridging layers to ensure smooth interoperability between older pipelines and the new VLM architecture.
- Integrate a new OCR engine delivering higher accuracy, better layout retention, and optimized performance for large-scale document processing.

Questions and Answers

Resources:

- **Workshops & Training**
  a. Visual NLP Workshop - Link
  b. Spark NLP Workshop - Link

- **Documentation**
  a. Healthcare NLP Documentation - Link
  b. Spark NLP Documentation - Link
  c. Visual NLP Documentation - Link

- **Repositories**
  a. John Snow Labs Repository - Link
  b. Dicom De-Identification Repository - Link
  c. PDF De-Identification Dataset - Link

- **Blogpost & Benchmarks**
  a. PDF De-Identification Benchmarks - Link
  b. OCR Benchmarks - Link
  c. Visual NLP Speed Benchmarks - Link

- **Notebooks & Tutorials**
  a. De-Identification / Obfuscation Notebooks - Link

- **Cloud Integrations**
  a. AWS SageMaker Listings - Link
  b. Integration Docs - Link