De-identification of DICOM files

for Data Scientists

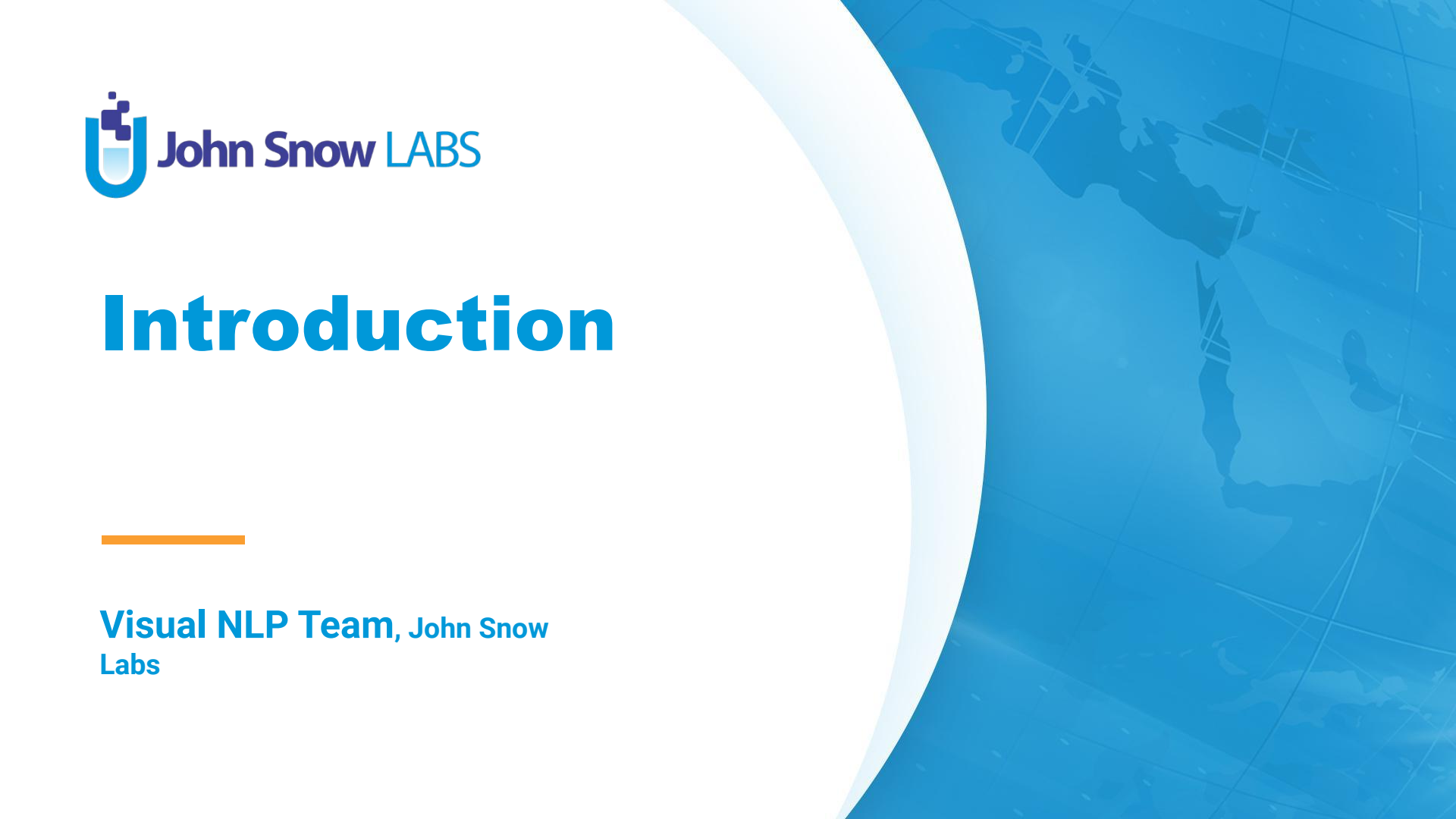
Visual NLP Team, John Snow Labs

Agenda

Main topic	Introduced Concepts
Introduction	Overview of the DICOM file format. Transfer Syntaxes. De-identification Challenges.
De-identification Targets	Burned-in text. Metadata. Overlay Annotation data. Encapsulated Documents.
Visual NLP Primer	Intro to Visual Language Models. Pipelines and Dataframes. Batch Mode vs. Streaming Mode.
Metadata De-identification	Metadata Deidentification Pipelines. User Defined Metadata Removal.
Burned-in Text Removal	Basic Pipelines for Text Removal. Text Detectors. NLP assisted Text Removal. Customized Pipelines. Dealing with big files.
Getting Started	Resources to start de-identifying yourself!

Presenter today

Alberto



Introduction

Visual NLP Team, John Snow Labs

The Dicom Format

John Snow LABS

- Digital Imaging and Communications in Medicine (DICOM)
- It turned 30 years old last year!
- Metadata is as important as image data.
- First developed for radiology and then cardiology.
- DICOM does not limit its action to images and associated information originating directly from medical devices.
- The standard is reviewed 5 times a year.

Transfer Syntaxes

John Snow LABS

1. **Compression**: Uncompressed, JPEG, JPEG Lossless, JPEG 2000, RLE Lossless.
2. **Byte Ordering (Endianness)**: Little Endian or Big Endian.
3. **Photometric Interpretation and Color Spaces**.
4. **Binary Encoding Rules**: Explicit VR (Value Representation) or Implicit VR.

Transfer Syntax UID	Transfer Syntax name
1.2.840.10008.1.2	Implicit VR Endian: Default Transfer Syntax for DICOM
1.2.840.10008.1.2.1	Explicit VR Little Endian
1.2.840.10008.1.2.1.99	Deflated Explicit VR Little Endian
1.2.840.10008.1.2.	Explicit VR Big Endian

Metadata

John Snow LABS

DICOM Metadata

PATIENT	STUDY	SERIES	IMAGE
Patient ID	Study ID	Modality	Rows
Patient Name	Study Date/Time	Manufacturer	Columns
Patient BirthDate	Study Description	Model Name/SW Version	Pixel Size
Patient Sex	Institution Name	Patient Position	Photometric Interpretation
Patient Weight	Referring Physician	Body Part Examined	Planar Configuration
...	...	+	Samples per Pixel
		Modality-Specific Attributes	...

- How important is this?: “the image can never be separated from this information”.
- (0008, 0060) -> Two hexa numbers, **(Group, Element).**
- Can also include manufacturer-specific attributes known as private data elements.

Challenges

John Snow LABS

- Dealing with multiple Transfer Syntaxes or even different versions of the standard.
- Custom Tags in Metadata.
- Volume of dataset and size of individual files.

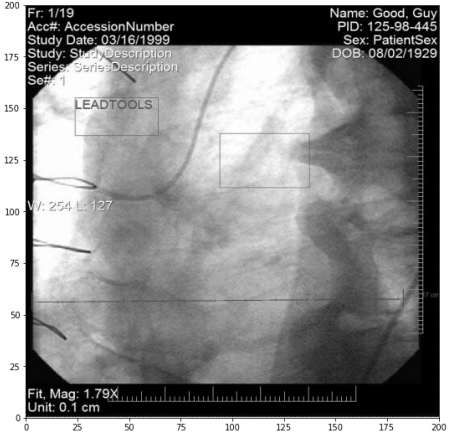
De-identification Targets

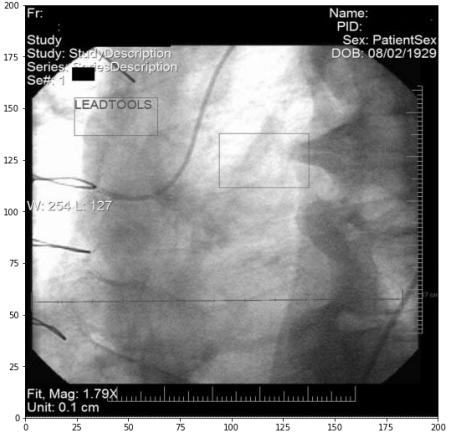
```
Field: content
                Id: 0
                Path: file:/Users/nmelnik/IdeaProjects/spar
k-ocr/workshop/jupyter/data/dicom/deidentify-medical-2.dcm
                Width: 985 px
                Height: 914 px
                Metadata:
                Dataset.file_meta -------------------------

-----
(0002, 0000) File Meta Information Group Length  UL: 196
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID         UI: 2.25.316618858989
175958452362820380094875916
(0002, 0003) Media Storage SOP Instance UID      UI: 2.25.128345240486
099325554244939980269110548
(0002, 0010) Transfer Syntax UID                 UI: Explicit VR Littl
e Endian
(0002, 0012) Implementation Class UID            UI: 1.3.6.1.4.1.30071
.8
(0002, 0013) Implementation Version Name         SH: 'fo-dicom 4.0.0'
-------------------------------------------------
(0008, 0020) Study Date                       DA: '19990316'
(0008, 0050) Accession Number                 SH: '95555556'
(0008, 0080) Institution Name                 LO: ''
(0008, 1030) Study Description                LO: 'Study Descriptio
n'
(0010, 0010) Patient's Name                   PN: 'Larisa KorsKi'
(0010, 0020) Patient ID                       LO: 'MF-5988888'
(0010, 0030) Patient's Birth Date             DA: '19280802'
(0010, 0040) Patient's Sex                    CS: 'M'
(0010, 1000) Other Patient IDs                LO: 'MF-5988889'
(0010, 1010) Patient's Age                    AS: '075Y'
(0010, 1030) Patient's Weight                 DS: '75.0'
(0020, 000d) Study Instance UID               UI: ''
(0028, 0002) Samples per Pixel                US: 1
(0028, 0004) Photometric Interpretation       CS: 'MONOCHROME2'
(0028, 0008) Number of Frames                 IS: '1'
(0028, 0010) Rows                             US: 985
(0028, 0011) Columns                          US: 914
(0028, 0100) Bits Allocated                   US: 8
(0028, 0101) Bits Stored                      US: 8
(0028, 0102) High Bit                         US: 7
(0028, 0103) Pixel Representation             US: 0
(7fe0, 0010) Pixel Data                       OB: Array of 900290 e
lements
```

```
Field: dicom_cleaned
                Id: 0
                Path: file:/Users/nmelnik/IdeaProjects/spar
k-ocr/workshop/jupyter/data/dicom/deidentify-medical-2.dcm
                Width: 985 px
                Height: 914 px
                Metadata:
                Dataset.file_meta -------------------------

-----
(0002, 0000) File Meta Information Group Length  UL: 196
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID         UI: 2.25.316618858989
175958452362820380094875916
(0002, 0003) Media Storage SOP Instance UID      UI: 2.25.128345240486
099325554244939980269110548
(0002, 0010) Transfer Syntax UID                 UI: Explicit VR Littl
e Endian
(0002, 0012) Implementation Class UID            UI: 1.3.6.1.4.1.30071
.8
(0002, 0013) Implementation Version Name         SH: 'fo-dicom 4.0.0'
-------------------------------------------------
(0008, 0020) Study Date                       DA: ''
(0008, 0050) Accession Number                 SH: ''
(0008, 0080) Institution Name                 LO: ''
(0008, 1030) Study Description                LO: 'Study Descriptio
n'
(0010, 0010) Patient's Name                   PN: ''
(0010, 0020) Patient ID                       LO: ''
(0010, 0030) Patient's Birth Date             DA: ''
(0010, 0040) Patient's Sex                    CS: 'M'
(0010, 1000) Other Patient IDs                LO: ''
(0010, 1010) Patient's Age                    AS: ''
(0010, 1030) Patient's Weight                 DS: '75.0'
(0020, 000d) Study Instance UID               UI: ''
(0028, 0002) Samples per Pixel                US: 1
(0028, 0004) Photometric Interpretation       CS: 'MONOCHROME2'
(0028, 0008) Number of Frames                 IS: '1'
(0028, 0010) Rows                             US: 985
(0028, 0011) Columns                          US: 914
(0028, 0100) Bits Allocated                   US: 8
(0028, 0101) Bits Stored                      US: 8
(0028, 0102) High Bit                         US: 7
(0028, 0103) Pixel Representation             US: 0
(7fe0, 0010) Pixel Data                       OB: Array of 900290 e
lements
```





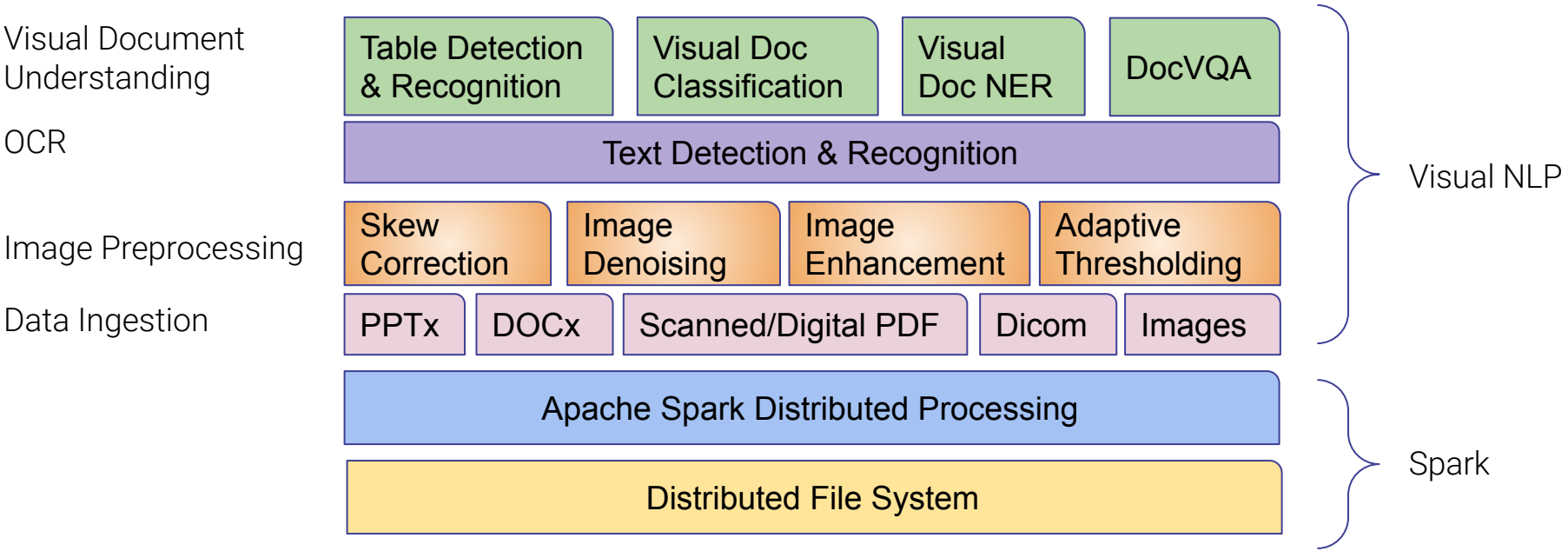
metadata

pixels

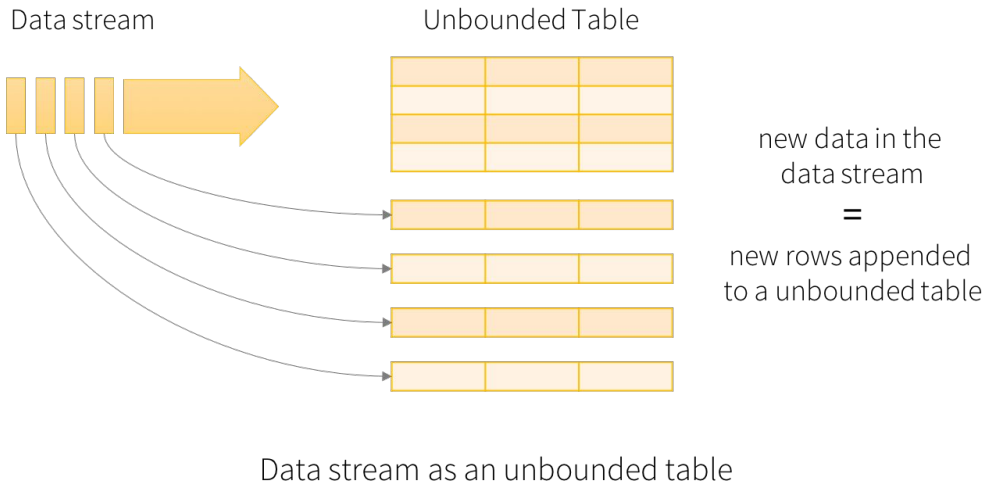
John Snow LABS

Visual NLP Primer

- Visual NLP is a Visual Document Understanding library built on top of Apache Spark.
- It’s not an API.
- Curated list of features -> only useful things that work.
- Optimized for performance and accuracy.
- Created by industry practitioners.
- Actively developed.
- Security minded.


Visual Document Understanding
OCR
Image Preprocessing
Data Ingestion
Table Detection & Recognition
Visual Doc Classification
Visual Doc NER
DocVQA
Text Detection & Recognition
Skew Correction
Image Denoising
Image Enhancement
Adaptive Thresholding
PPTx
DOCx
Scanned/Digital PDF
Dicom
Images
Apache Spark Distributed Processing
Distributed File System
Visual NLP
Spark

Streaming



- Sources and Sinks.
- Streaming Dataframe.
- Unbounded Table.
- Details: Streaming Query.

Engagement Models

John Snow LABS

Software Licensing

- Internal data science team needs to become self-sufficient with NLP and OCR
- Need partner to provide state of the art software in AI, NLP, OCR and LLM
- Want simplicity in architecture, integration, security, and scaling

Professional Services

- Data Science team is small, at capacity, or not familiar with NLP & OCR
- Time-to-market is key, and the team doesn't have enough bandwidth or experience to build a solution
- A partner can help get to market fast and skill up the internal team to become self-sufficient over time
- NLP, OCR is not a core business and it's preferred to outsource it

Managed Services

- NLP & OCR is used across the enterprise, and a robust and trusted solution is needed for scale, reproducibility and speed
- The Data Science team needs to deliver value to the broader organization
- Enterprise projects are a top focus; State-of-the-art models are a priority to not be surpassed by competitors
- Data Science team is growing rapidly and can use domain-specific experts
- Deploying, monitoring, and updating models in production is required

Getting Started

- Check Library Documentation.
- Get your trial license here.
- Check Visual NLP Workshop repo.
- Contact me or Enes: alberto@johnsnowlabs.com, enes@johnsnowlabs.com

Questions and Answers

Contact Us!

Contact us on Slack!

Emails: alberto@johnsnowlabs.com,
enes@johnsnowlabs.com