De-Identification of Medical Images

Alberto Andreotti
Head of Visual NLP (Data Scientist)

Alexander Branov
Data Scientist

Aymane Chilah
Data Scientist

Nitin Kumar
Scala Developer (Data Scientist)

Agenda

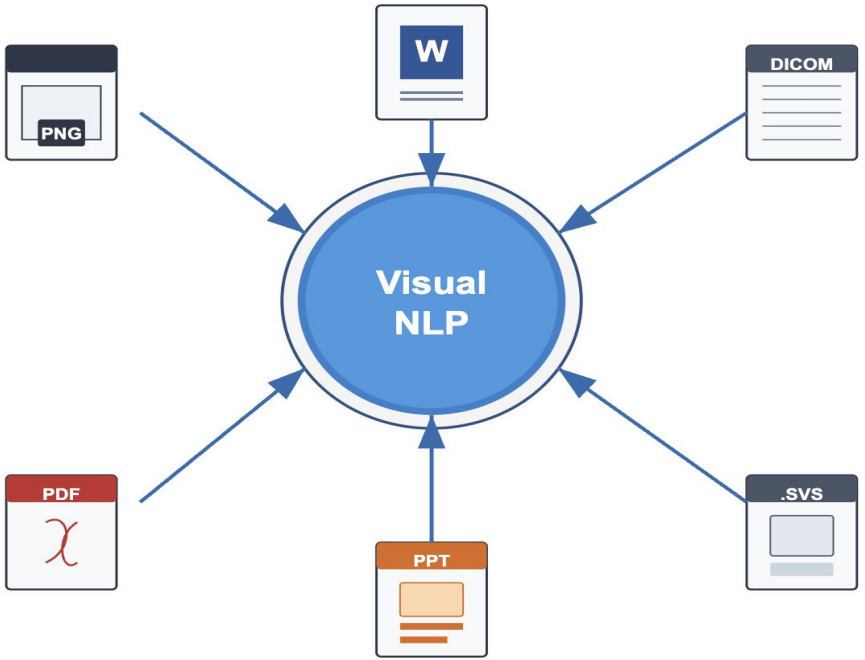
Main topic	Introduced Concepts	Pages
Introduction	Introduction To Visual NLP	3
Introduction to Dicom	Dicom, Tags, Pixels, Transfer Syntaxes	4 - 10
Dicom DeIdentification	Metadata Extraction, Metadata Deid, Metadata Methods, Pixel Deid, Pixel Deid Using Metadata	11 - 13
MIDI-B	MIDI-B Data, Validation Script, Visual NLP MIDI-B Solution	14 -16
MIDI-B Results	Study, PHI, Pixels	17 - 21
Visual NLP Dicom Pipeline	Block Diagram For Dicom Metadata & Pixel DeIdentification	22
Dicom Metrics	Dicom Dataset, Databricks/Collab Metrics, Comparison With Presidio	23 - 24
SVS	Introduction To SVS, SVS Metadata, Deid Notebook Links	25 - 27

Introduction to Visual NLP

- Visual NLP as a library helps in unblocking textual information from scanned/digital/medical documents.

- Provides high-level capabilities
  - Text Detection
  - OCR
  - Table Detection/Recognition
  - Entity Extraction
  - VQA

- Entry points into Visual NLP
  - Images
  - PDFs (Digital & Scanned)
  - Document, PPT
  - Dicom
  - SVS



What is Dicom?

John Snow LABS

- Dicom (**Digital Imaging And Communications in Medicine**)
  - International standard for medical information.
  - Allows healthcare professionals to easily view and understand medical information regardless of the equipment/software used to generate it.

- Dicom files have two fundamental parts
  - Pixels (Medical Images)
    - MRIs, CT Scans, X-Rays etc.
    - PHI burned into the pixels.
    - Overlay Pixels

  - Dicom Tags (Metadata)
    - Patient/Physician PHI
    - Equipment/Software-Related Information
    - Examination/Test-Related Information
    - Other Relevant Information (Pixels, Transfer Syntaxes, Compression level)

Dicom Tags

John Snow LABS

- Dicom Tags are essentially metadata elements associated with a dicom object.
- They carry information about the dicom object, pixels, patient demographics, equipment etc.
- Tag Structure
  - Each Dicom Tag is a unique hexadecimal number in format (XXXX, YYYY)
  - The first four digits (XXXX) represent the group number.
  - The last four digits (YYYY) represent the element number.
- Each Tag also has Value Representation (VR), which denotes the data type and format of the tag’s value (String, Integer, Date, Sequence, etc)
- To accommodate vendor-specific needs beyond the standard tags, DICOM supports 'private tags', which are always identified by odd group numbers.

Dicom Tag Types

John Snow LABS

- Dicom Tags can be 5 Specific Types
  - **Type 1**
    - Must be present and contain valid data.
    - Cannot be null.
  - **Type 1C**
    - Must be present and contain valid data, if condition applies.
    - If condition does not apply, must not be present
  - **Type 2**
    - Must be present, can be empty.
    - If redacted, set to “”
  - **Type 2C**
    - Must be present, if condition applies, can be empty.
    - If redacted, set to “”
  - **Type 3**
    - Can be present, safe to remove during anonymization.
    - Can be empty.

Dicom Pixels

John Snow LABS

- Active Photometric Interpretation (0028,0004)
  - Monochrome 1 & Monochrome 2
  - Palette Color
  - RGB
  - YBR_FULL, YBR_FULL_422, YBR_PARTIAL_420, YBR_RCT, YBR_ICT

- Pixel Representation (0028,0103) denotes whether the pixel values are signed/unsigned.

- Bits Allocated (0028,0100) is the number of bits allocated for each pixel.

- Bits Stored (0028,0101) is the actual number of bits used to represent pixel value.

- Image Pixel Data (7FE0,0010)
  - Uncompressed Transfer Syntaxes
  - Compressed Transfer Syntaxes (Lossy / Lossless)

- Visual NLP does not alter the Photometric Interpretation / Color Space of DICOM pixels in any way. The final DICOM file retains the same characteristics as the input.

```
Dataset.file_meta -------------------------------
(0002,0000) File Meta Information Group Length  UL: 186
(0002,0001) File Meta Information Version       OB: b'\x00\x01'
(0002,0002) Media Storage SOP Class UID         UI: Digital X-Ray Image Storage - For Presentation
(0002,0003) Media Storage SOP Instance UID      UI: 2.25.463628102274132074849128424375172598
(0002,0010) Transfer Syntax UID                 UI: Explicit VR Little Endian
(0002,0012) Implementation Class UID            UI: 1.3.6.1.4.1.22213.1.143
(0002,0013) Implementation Version Name         SH: '0.5'
(0002,0016) Source Application Entity Title     AE: 'POSDA'
-------------------------------------------------
(0008,0005) Specific Character Set              CS: 'ISO_IR 100'
(0008,0008) Image Type                          CS: ['ORIGINAL', 'PRIMARY', '']
(0008,0016) SOP Class UID                       UI: Digital X-Ray Image Storage - For Presentation
(0008,0018) SOP Instance UID                    UI: 2.25.463628102274132074849128424375172598
(0008,0020) Study Date                          DA: '20010705'
(0008,0021) Series Date                         DA: '20010705'
(0008,0022) Acquisition Date                    DA: '20010705'
(0008,0023) Content Date                        DA: '20010705'
(0008,0024) Overlay Date                        DA: '20010705'
(0008,0025) Curve Date                          DA: '20010705'
(0008,002A) Acquisition DateTime                DT: '20010705'
(0008,0030) Study Time                          TM: ''
(0008,0032) Acquisition Time                    TM: ''
(0008,0033) Content Time                        TM: ''
(0008,0050) Accession Number                    SH: '20010706E403961'
(0008,0060) Modality                            CS: 'DX'
(0008,0070) Manufacturer                        LO: 'GE MEDICAL SYSTEMS'
(0008,0080) Institution Name                    LO: 'Mccoy Medical Clinic'
(0008,0081) Institution Address                 ST: '45166 Morgan Walks Suite 852 East Aaron, KS 93919'
(0008,0090) Referring Physician's Name          PN: 'EATON^WILLIAM'
(0008,0092) Referring Physician's Address       ST: '12148 Donna Overpass Apt. 115 North Sherry, TN 60972'
(0008,0094) Referring Physician's Telephone Num SH: '752.671.3789x973'
(0008,1050) Performing Physician's Name         PN: 'ELLIS^PAUL'
(0008,1090) Manufacturer's Model Name           LO: 'Revolution XQi ADS_28.4'
(0008,1155) Referenced SOP Instance UID         UI: 2.25.181816462680784373513761473735268351478
(0009,0010) Private Creator                     LO: 'GEMS_IDEN_01'
(0009,1027) [Image actual date]                 SL: 20000101
(0010,0010) Patient's Name                      PN: 'MARTIN^CHAD'
(0010,0020) Patient ID                          LO: '339833062'
(0010,0030) Patient's Birth Date                DA: ''
(0010,0040) Patient's Sex                       CS: ''
(0010,1010) Patient's Age                       AS: ''
(0010,1040) Patient's Address                   LO: '68265 Mark Bridge Suite 049 Robinsonville, OK 55335'
(0010,21D0) Last Menstrual Date                 DA: '20010705'
(0013,0010) Private Creator                     LO: 'CTP'
(0013,1010) Private tag data                    LO: 'Pseudo-PHI-DICOM-Data'
(0013,1013) Private tag data                    LO: '87009668'
(0018,0015) Body Part Examined                  CS: 'CHEST'
(0018,0060) KVP                                 DS: '125'
(0018,1020) Software Versions                   LO: 'Ads Application Package VERSION ADS_28.4'
(0018,1110) Distance Source to Detector         DS: '1800'
(0018,1111) Distance Source to Patient          DS: '1750'
(0018,1150) Exposure Time                       IS: '43'
```

Transfer Syntaxes

John Snow LABS

- Transfer Syntaxes are the language translators of Dicom objects, ensuring different systems understand each other.
- They dictate how data is stored and transmitted within Dicom objects.
- Without correct transfer syntax, data, especially pixel data can be misinterpreted, leading to processing errors.
- Compression Options:
  - Lossless preserves all image details, with no reduction in dicom object size.
    - JPEG-LS, RLE, JPEG 2000 Lossless
  - Lossy reduces file size but with some quality loss.
    - JPEG, JPEG 2000
- There are limitations to using a specific transfer syntax.
  - Not all Transfer Syntaxes support variations in Photometric Interpretation, Bits Stored, Pixel Representation, or their combinations.

Transfer Syntax	Color Space	Visual NLP Support	
		Encoding	Decoding
RLE Lossless	Monochrome 1, Monochrome 2, PALETTE COLOR, RGB, YBR_FULL	Yes	Yes
JPEG Baseline 8-Bit	Monochrome 1, Monochrome 2, RGB, YBR_FULL	Yes	Yes
JPEG 2000	Monochrome 1, Monochrome 2, RGB, YBR_FULL, YBR_RCT, YBR_ICT	Yes	Yes
JPEG 2000 Lossless	Monochrome 1, Monochrome 2, PALETTE COLOR, RGB, YBR_FULL, YBR_RCT	Yes	Yes
JPEG LS Lossless	Monochrome 1, Monochrome 2, PALETTE COLOR, RGB, YBR_FULL	Yes	Yes
JPEG LS Near Lossless	Monochrome 1, Monochrome 2, RGB, YBR_FULL	Yes	Yes
HTJ2K	Monochrome 1, Monochrome 2, YBR_ICT, YBR_RCT, RGB, YBR_FULL	No	Yes
HTJ2K Lossless	Monochrome 1, Monochrome 2, PALETTE COLOR, YBR_ICT, YBR_RCT, RGB, YBR_FULL	No	Yes
HTJ2K Lossless RPCL	Monochrome 1, Monochrome 2, PALETTE COLOR, YBR_ICT, YBR_RCT, RGB, YBR_FULL	No	Yes

Visual NLP Dicom Pipeline

- **Dicom Metadata Only De-Identification Pipeline**
  - This pipeline demonstrates how to extract metadata from a DICOM file and apply de-identification techniques specifically on metadata tags.

  - Notebook: Visual_12_Dicom_Metadata_Only.ipynb

- **Dicom Pixel De-Identification Pipeline**
  - This pipeline focuses on de-identifying Protected Health Information (PHI) within the image pixels as well as metadata.

  - It involves extracting metadata, detecting text regions in pixel data, extracting text, performing NER, and finally redacting identified PHI regions and metadata tags.

  - Notebook: Visual_13_Dicom_Deidentification.ipynb

Visual NLP Metadata DeIdentification Methods

Action	VR	Description
replaceWithRandomName	PN, LO	Fake pair while preserving consistency per input if a seed is provided.
patientHashID	LO	Deterministic, numeric pseudonym based on the original Patient ID.
hashID	UI, LO, SH	Generate a deterministic UID.
ensureTagExists	LO, DS, IS, CS, SS	Create tag if missing.
replaceWithLiteral	All VR	Replace tag with a literal value.
shiftUnixTimeStampRandom	SL, FD	Generates a realistic, shifted past timestamp based on the current UNIX time.
shiftDateRandomNbOfDays	DA, DT, AS	Deterministic date shift by random number of days.
shiftDateFixedNbOfDays	DA, DT	Deterministic date shift by fixed number of days.
remove	All VR	Remove tag value, with placeholder.
delete	All VR	Delete Tag

Visual NLP Dicom Pipeline (Metadata - Deidentification)

- **Dicom Pixel De-Identification With Metadata Pipeline**
  - This pipeline focuses on de-identifying Protected Health Information (PHI) within DICOM image pixels using metadata as a supporting source.
  - Metadata extracted from the DICOM file is used to generate additional NER entries, which can be combined with entities detected by Healthcare-NLP models.
  - This hybrid approach improves PHI detection by capturing entities that may be missed by the models but are present in the metadata.
  - The remaining steps follow the standard pixel de-identification process.
  - Notebook: Visual_14_Dicom_Deidentification_Using_Metadata.ipynb

Introduction to MIDI-B

- Benchmark dataset for medical imaging de-identification.
- Contains DICOM files with PHI in both metadata and pixel data.
- Dataset available via TCIA:
  - MIDI-B Collection
  - Validation Script
  - TCIA Download Script
- Dataset Structure:
  - Two sets: Validation and Test.
  - Each set includes:
    i. Synthetic set → contains DICOM files with PHI in Pixels and Tags.
    ii. Curated set → same files manually de-identified by TCIA.
- Validation Actions Used in MIDI-B:
  <date_shifted>, <uid_changed>, <pixels_retained>, <text_removed>, <patid_consistent>, <text_retained>, <text_notnull>, <pixels_hidden>, <uid_consistent>, <tag_retained>
- PHI Formats in Pixel Data:
  - Typical structure: ***FirstName LastName [Gender] Date\nDate***
  - Other PHI: *JT, SWU, JKR, MWF, ICG, NKF, YH, TJN*

Validation Run

Action	Total	Failed
<tag_retained>	1,101,091	0
<text_notnull>	618,539	0
<date_shifted>	139,774	0
<uid_changed>	234,418	0
<pixels_retained>	23,886	0
<uid_consistent>	234,418	0
<patid_consistent>	23,921	0
<pixels_hidden>	35	5

Test Run

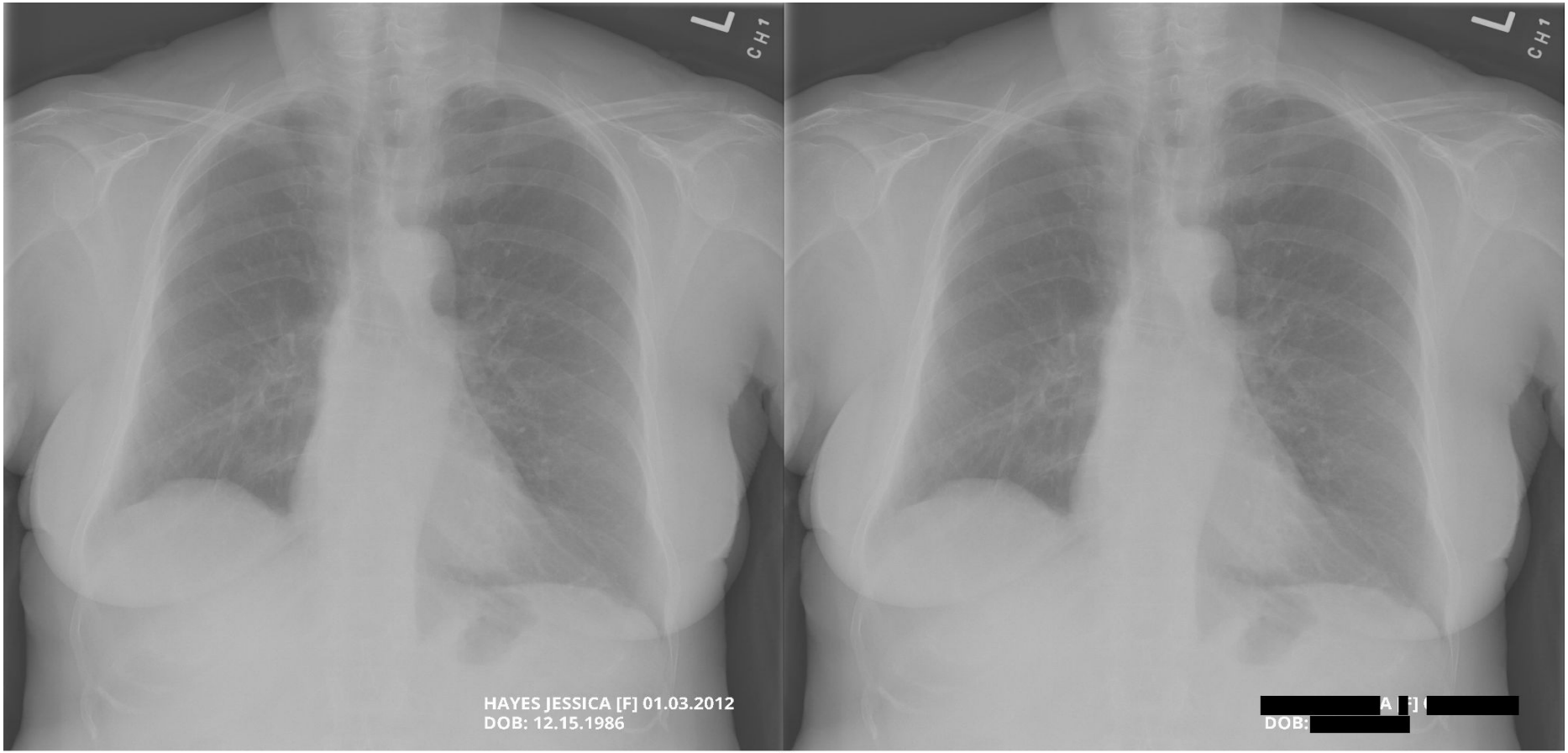
Action	Passed	Failed
<tag_retained>	1,325,259	0
<text_notnull>	741,498	0
<date_shifted>	171,930	0
<uid_changed>	280,275	0
<pixels_retained>	29,633	0
<uid_consistent>	280,275	0
<patid_consistent>	29,660	0
<pixels_hidden>	27	0

Visual NLP MIDIB Pixel DeIdentification

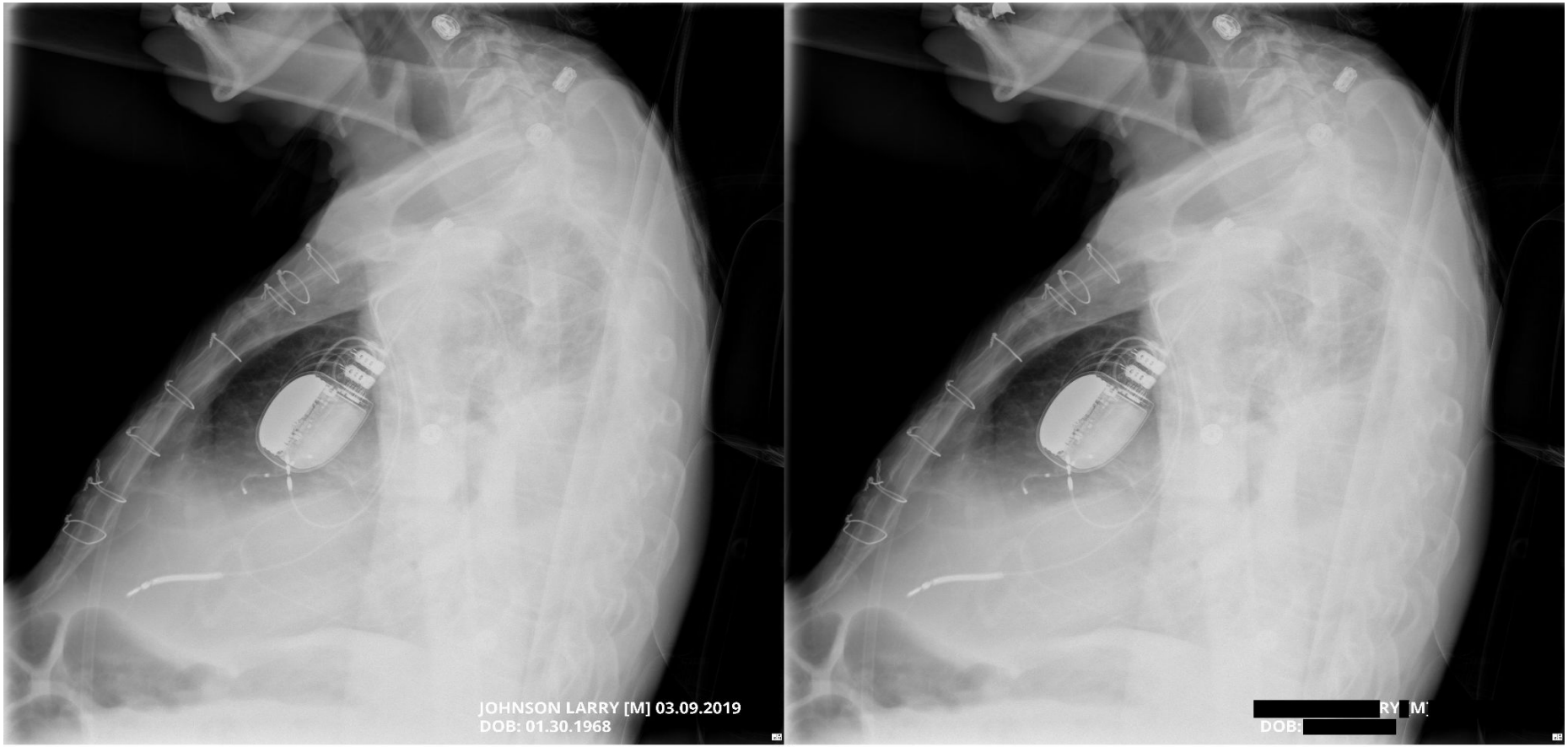
- The DICOM files originate from the MIDI-B dataset, which contains several studies; among these, 25 files were identified as having meaningful PHI within the pixel data such as patient names, gender, or dates of birth.
- A new pixel de-identification pipeline was developed and optimized specifically for these cases.
- Additionally, Excel report files have been shared for both test and validation sets using the MIDI-B PHI validation script.
  - **dicom/validation_report**
- Examples of PHI text that were not successfully redacted in the Validation Set:
  - ACO, JKR, TJN, JGR, YH
- Notebook: Visual_15_Dicom_Midib.ipynb

Set	Total	Passed	Failed	Score
Test Set	27	27	0	1.00
Validation Set	35	30	5	0.87

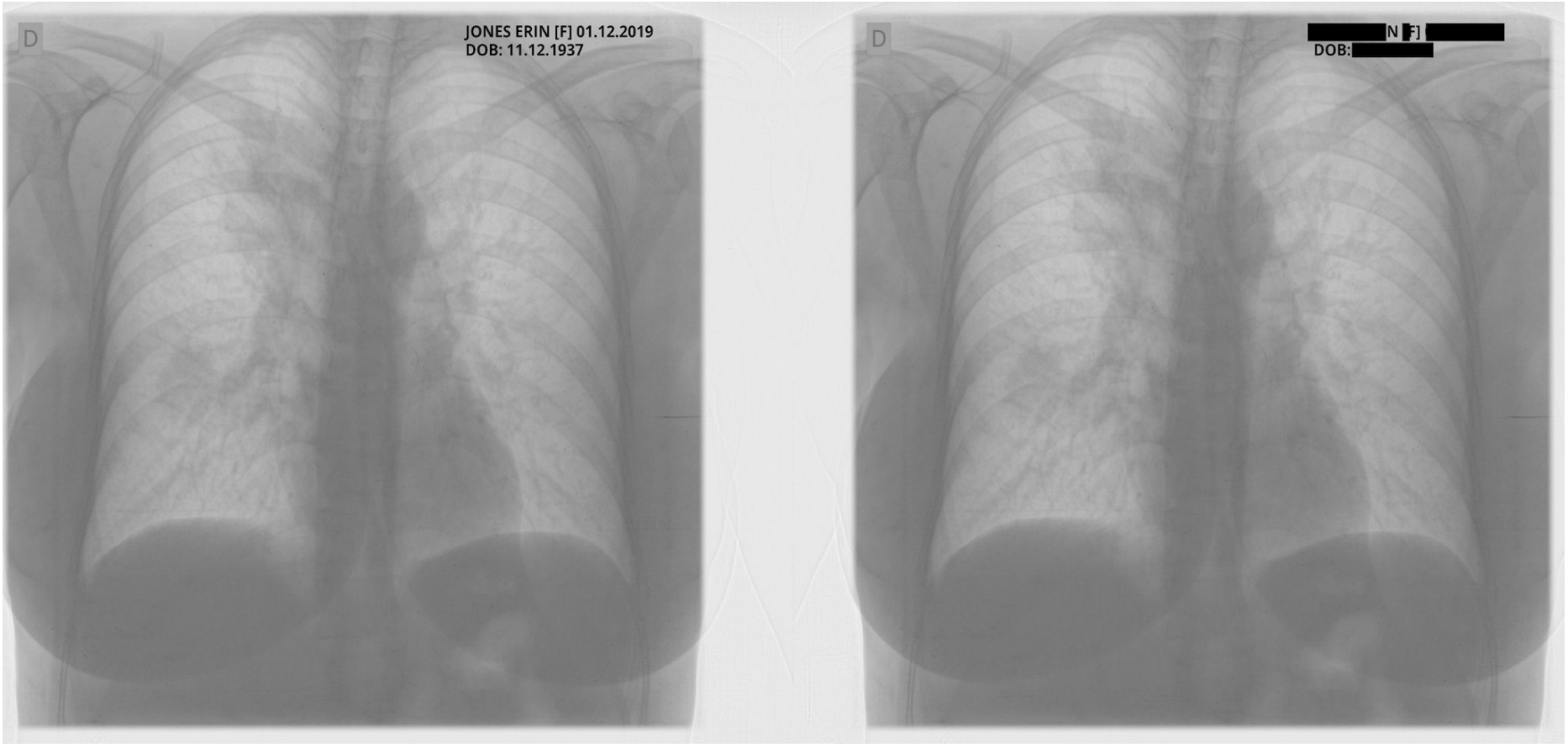
UID: 2.3.957.1.1.1204841.4.221.1244497244821555782
MIDIB ORIGINAL PHI: HAYES JESSICA [F] 01.03.2012 DOB: 12.15.1986



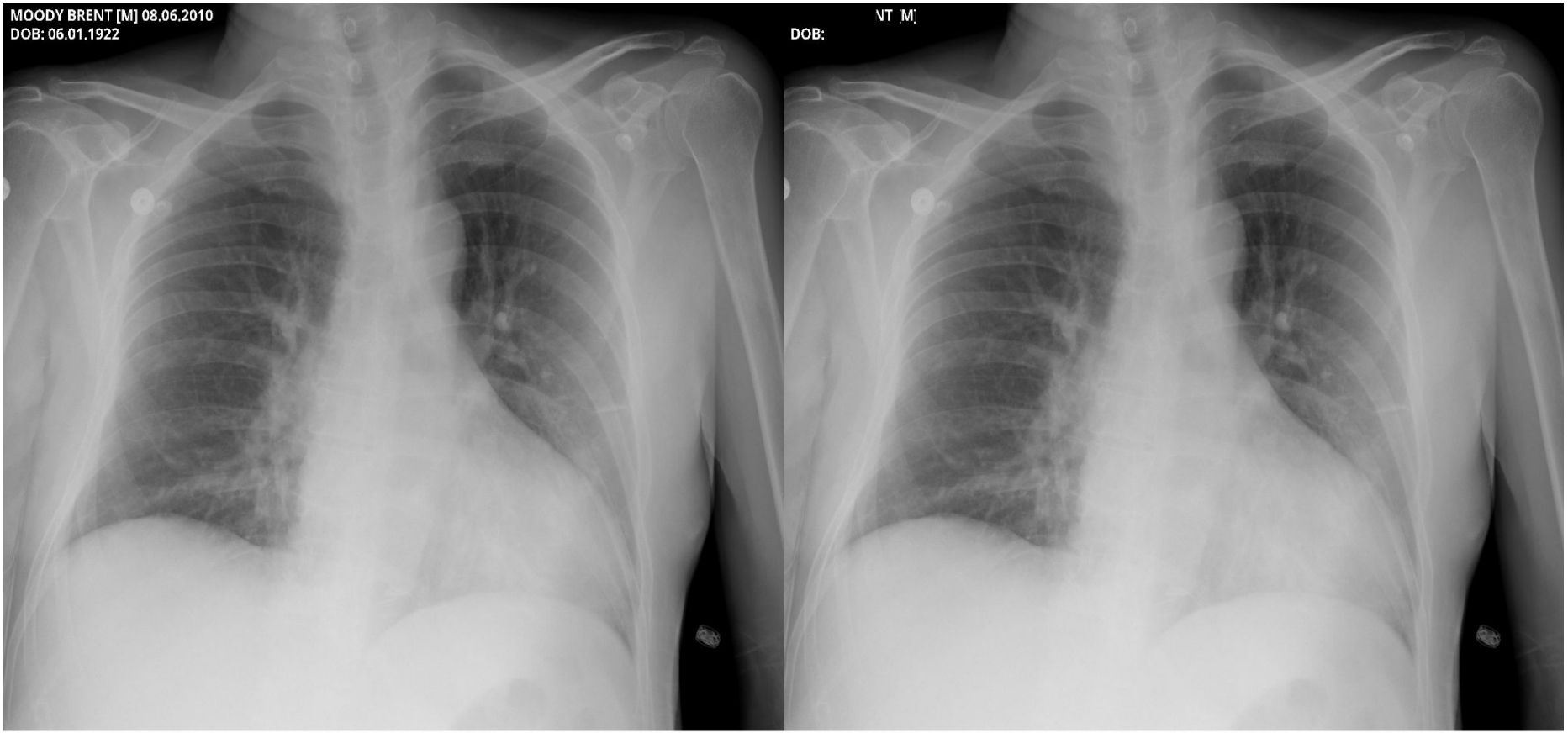
UID: 2.1.239.0.1.5573651.5.438.2234045507897739313
MIDIB ORIGINAL PHI: JOHNSON LARRY [M] 03.09.2019 DOB: 01.30.1968



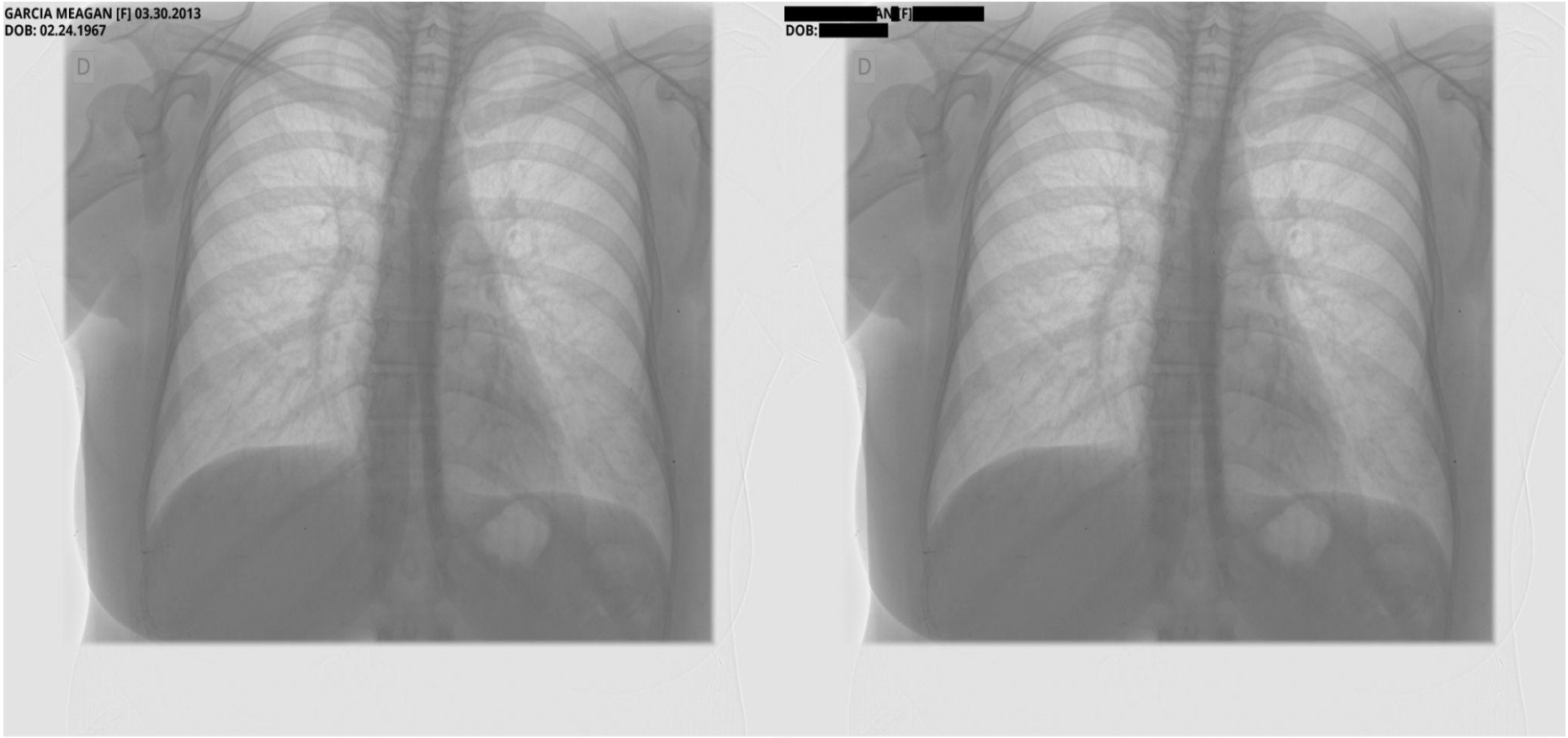
UID: 2.4.569.0.1.8829737.8.747.1088525486356273997
MIDIB ORIGINAL PHI: JONES ERIN [F] 01.12.2019 DOB: 11.12.1937



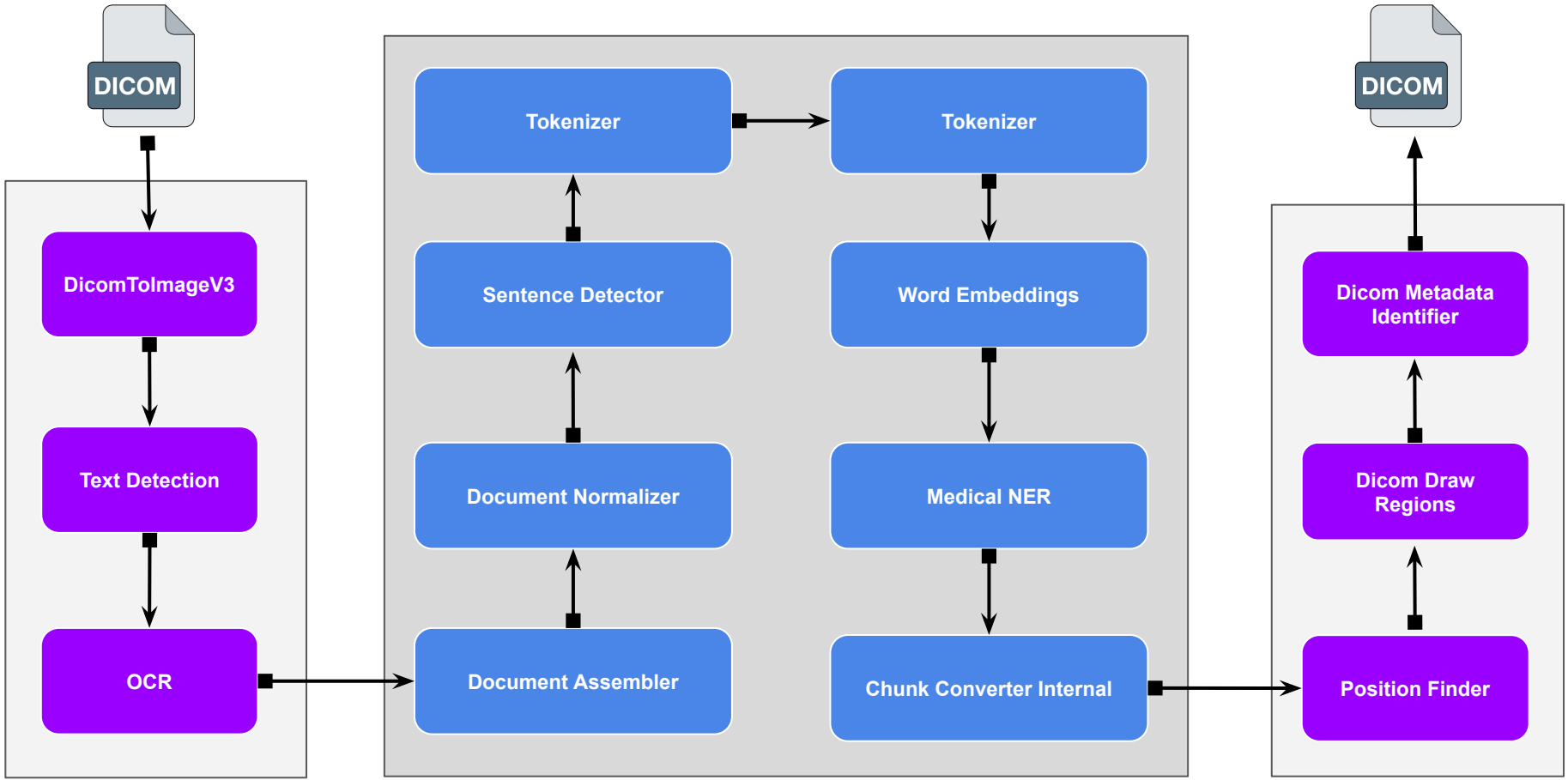
UID: 3.4.732.1.3.0861594.8.259.2531881142306101964
MIDIB ORIGINAL PHI: MOODY BRENT [M] 08.06.2010 DOB: 06.01.1922



UID: 3.3.608.1.1.2993058.8.105.13430594626591622 80
MIDIB ORIGINAL PHI: GARCIA MEAGAN [F] 03.30.2013 DOB: 02.24.1967



Visual NLP Dicom Pipeline



Dicom Metrics

- **Repo Link** https://github.com/JohnSnowLabs/dicom-deid-dataset
- **Google Collab**
  - CPU - HIGH RAM [8 Cores] - 0.18 Credits/hr
  - GPU - HIGH RAM [8 Cores] 1 X A100 GPU (40 GB) - 7.62 Credits/hr
- **Databricks**
  - **Cluster**
    - CPU - Driver 64 GB [16 Cores] m4.4xlarge, with minimum & maximum 8 Executors 32GB [8 Cores] m4.2xlarge - 15 dbu/h
    - GPU - Driver 64 GB Single GPU g4dn.4xLarge[T4], with minimum & maximum 2 Executors 16GB Single GPU g4dn.xLarge[T4] - 4.27 dbu/h
  - **Standalone**
    - CPU - Driver 64 GB [16 Cores] m4.4xlarge - 3 dbu/h
    - GPU - Driver 64 GB Single GPU g4dn.4xLarge[T4] - 2.85 dbu/h

Model	Precision	Recall	F1-Score	Google Collab GPU (s)	Databricks Cluster CPU (s)	Databricks Cluster GPU (s)
ImageTextDetector + ImageToTextV2 (Base)	0.871	0.800	0.834	3.63	2.94	2.76
ImageTextDetector + ImageToTextV2 (Large)	0.892	0.822	0.856	4.06	3.59	3.2
ImageTextDetector + ImageToTextV3	0.741	0.433	0.547	0.68	1.83	1.0
ImageToText	0.436	0.289	0.348	0.31	0.85	0.89
Presidio (CPU Only)	0.07	0.128	0.091	0.54	N/A	N/A

SVS (Scanned Virtual Slide)

John Snow LABS

- SVS (Scanned Virtual Slide) is a proprietary Whole Slide Image (WSI) format developed by Aperio.
- Each image level is stored as a separate TIFF image directory (IFD) with metadata.
- It’s used in digital pathology to store extremely high-resolution scanned microscope slides.
  - **File Extension:** .svs
  - **MIME Type:** image/tiff (it’s a specialized multi-page TIFF)
  - **Compression:** JPEG or JPEG2000 compression inside TIFF tiles
  - **Storage:** Pyramidal TIFF structure (multi-resolution levels for zooming)
- Structure of an SVS File
  - Level 0 (Full-resolution scan) : Base level, 40× magnification
  - Level 1 (Used for faster visualization) : Downsampled, 20× magnification
  - Level 2 (For overviews) : Thumbnail-level, 10x magnification
  - Level 3 (Low resolution preview) : Macro image, 1x magnification
  - Level N

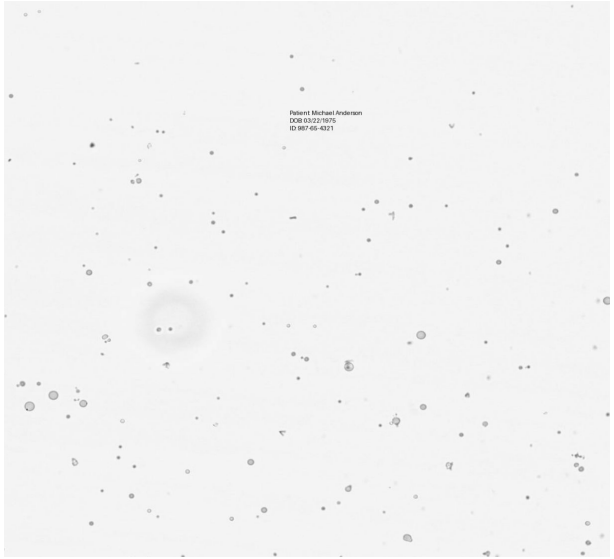
SVS Metadata

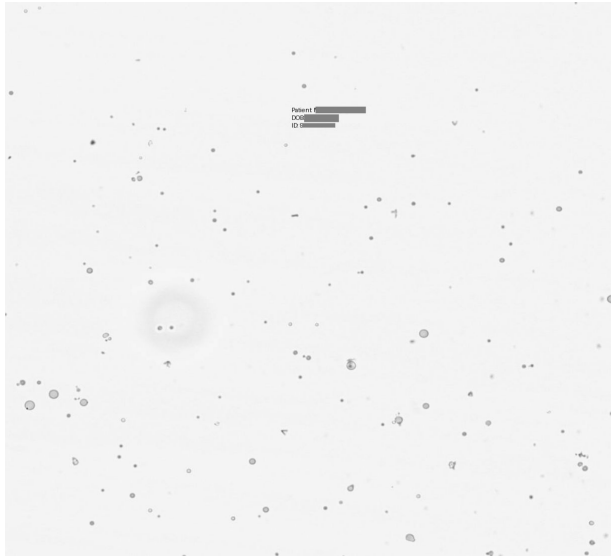
John Snow LABS

Category	Example Key	Description
Magnification	`Aperio.AppMag`	Objective magnification (20× / 40×)
Pixel Scale	`Aperio.MPP`	Microns per pixel
Scanner Info	`Aperio.ScanScopeID`	Device ID of scanner
Image Size	`ImageWidth` , `ImageLength`	Full slide dimensions
Tile Size	`TileWidth` , `TileLength`	Patch size used during scanning
Compression	`Compression=JPEG`	Compression type per tile
Color Model	`PhotometricInterpretation=RGB`	Color encoding used
Specimen ID	`Aperio.Barcode` , `Aperio.SlideId`	Links slide to specimen
Date & Time	`Aperio.Date` , `Aperio.Time`	Scan timestamp

SVS Pixel & Metadata DeIdentification

- SVS Pretrained Pipeline : Link
- Sagemake Listing: Link
- Notebooks:
  - SparkOcrWSIDeidentification_folder.ipynb : Link
  - SparkOcrWSIDeidentification.ipynb : Link





Next Steps

- Addressing text_retained and text_removed tags.
- The actions <text_retained> and <text_removed> require enhanced handling through the addition of NER stages, Regex rules, and LLM-based metadata de-identification, all integrated within Spark processing stages.
- While Regex rules are effective for structured or pattern-based PHI, they are limited when dealing with unstructured text or contextual entities. To achieve comprehensive coverage, the pipeline should leverage Named Entity Recognition (NER) and Large Language Models (LLMs) for detecting and redacting sensitive information within textual DICOM metadata fields.
- To streamline this process, pretrained de-identification pipelines can be developed combining both pixel-level and tag-level de-identification into a unified Spark workflow. These pipelines would ensure consistent, scalable, and fully in-Spark processing for both structured metadata and embedded image PHI.

Questions and Answers

Resources

- **Visual NLP:**
  - https://www.johnsnowlabs.com/visual-nlp/
- **Visual NLP Workshop:**
  - https://github.com/JohnSnowLabs/visual-nlp-workshop/tree/master
- **Healthcare NLP Workshop:**
  - https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/healthcare-nlp
- **Spark NLP Workshop:**
  - https://github.com/JohnSnowLabs/spark-nlp-workshop
- **Visual NLP Pipeline Components:**
  - https://nlp.johnsnowlabs.com/docs/en/ocr_pipeline_components
- **Visual NLP Speed Benchmarks:**
  - https://nlp.johnsnowlabs.com/docs/en/ocr_benchmark
- **Visual NLP Helpers:**
  - https://nlp.johnsnowlabs.com/docs/en/ocr_structures
- **Dicom Deidentification Metrics:**
  - https://github.com/JohnSnowLabs/dicom-deid-dataset