

De-Identification of Medical Images in Dicom Format

Alberto Andreotti
Head of Visual NLP (Data Scientist)

Alexander Branov
Data Scientist

Aymane Chilah
Data Scientist

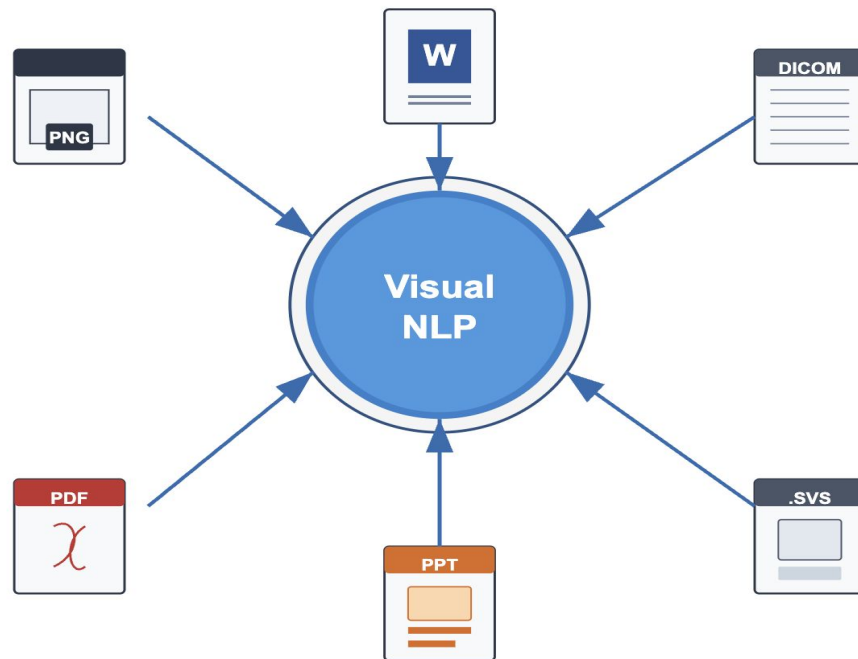
Nitin Kumar
Scala Developer (Data Science)

Agenda

Topic	Slides	Reference Notebook / Links
Introduction to Visual NLP	3	Visual NLP
What is Dicom?	4	Dicom Standard
Dicom Tags	5	Private Tags
Dicom Pixels	6 - 7	Dicom Standard - Photometric Interpretation
Transfer Syntaxes in Dicom	8 - 9	Dicom Standard - Transfer Syntax
Visual NLP Dicom Pipeline	10 - 13	N/A
Metrics	14 - 15	Evaluation Repo

Introduction to Visual NLP

- Visual NLP as a library helps in unlocking 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?

- 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 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, etc)
- To accommodate vendor-specific needs beyond the standard tags, DICOM supports 'private tags', which are always identified by odd group numbers.

- 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'
(0010,2100) 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 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

- **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: [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 (Named Entity Recognition), and finally redacting identified PHI regions and metadata tags.
- Notebook: [Dicom_Deidentification.ipynb](#)

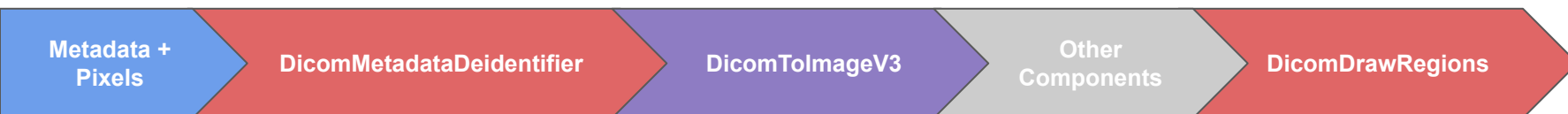
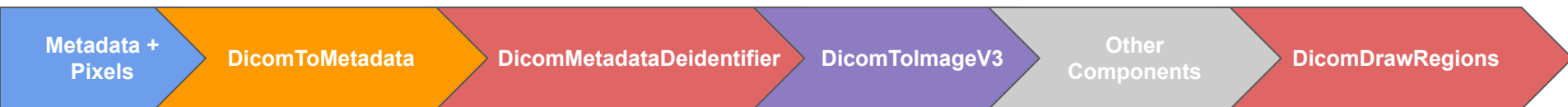
- **Dicom Pixel De-Identification With Metadata Pipeline**

- 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 (Named Entity Recognition) 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: [Dicom_Deidentification_Using_Metadata.ipynb](#)

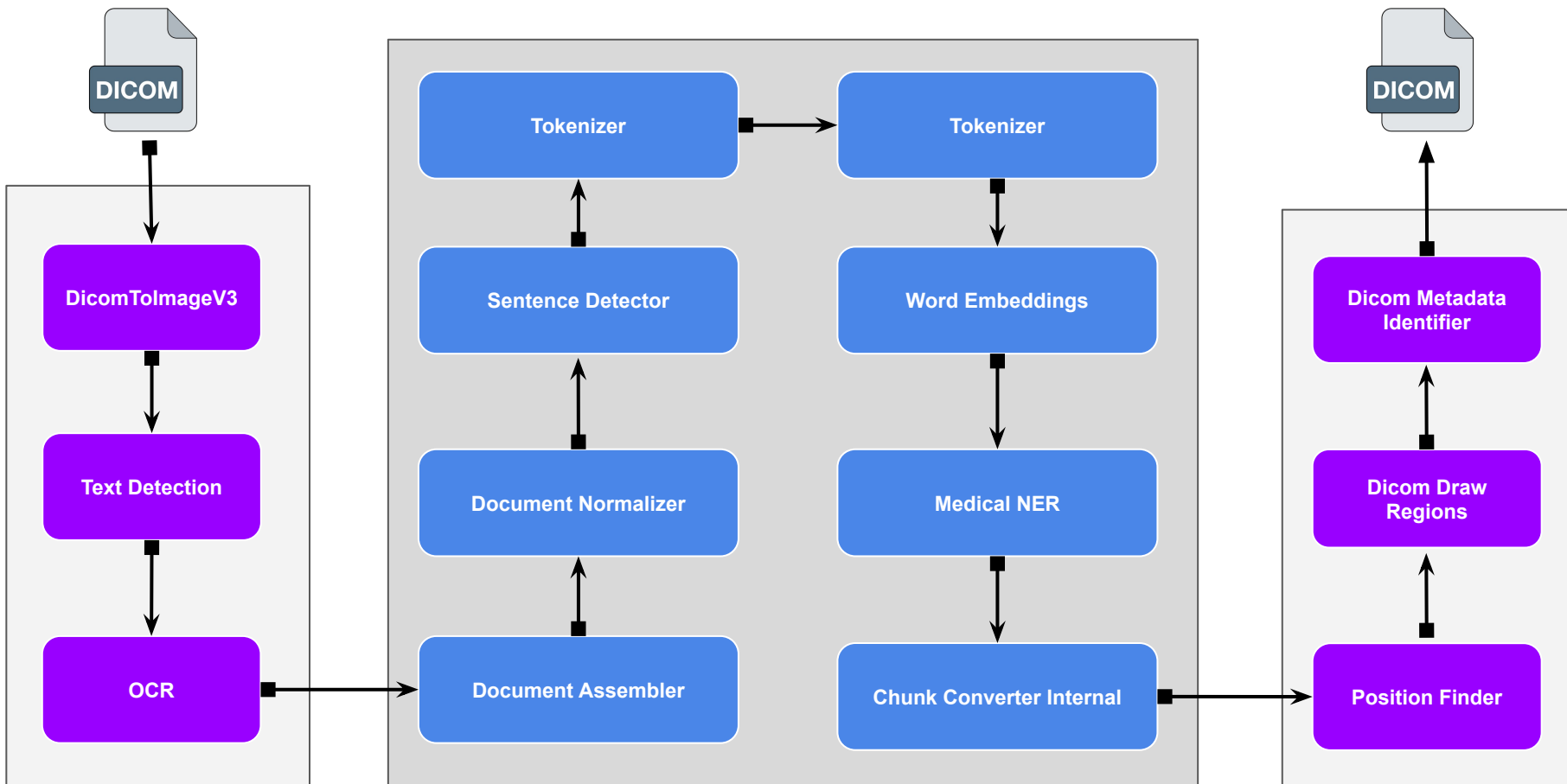
- **Dicom Pixel Obfuscation Pipeline**

- This pipeline focuses on replacing original PHI text with newly rendered obfuscated text directly within the image.
- Instead of redacting PHI regions using black boxes, it extends the concept of de-identification by generating visually consistent patches that contain obfuscated text, seamlessly overlaid in place of the original content.
- What is Obfuscation?
 - {original Text : "John Snow", obfuscated Text : "Jane Knox" }
 - {original Text : "02/02/1999", obfuscated Text : "20/12/2015" }
 - {original Text : "22", obfuscated Text : "33" }
- Notebook: [Dicom_Obfuscation_Methods.ipynb](#)

Visual NLP Dicom (Pipeline Flows)



Visual NLP Dicom Pipeline



- **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	Precisi on	Recall	F1-Score	Google Collab (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	0.07	0.128	0.091	0.54	N/A	N/A

- **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>