SUMMER School '24

DATA UNMASKS BIAS

Emory CXR
Data Preprocessing

Theo Dapamede, MD, PhD

08/19/2024

HITI LAB

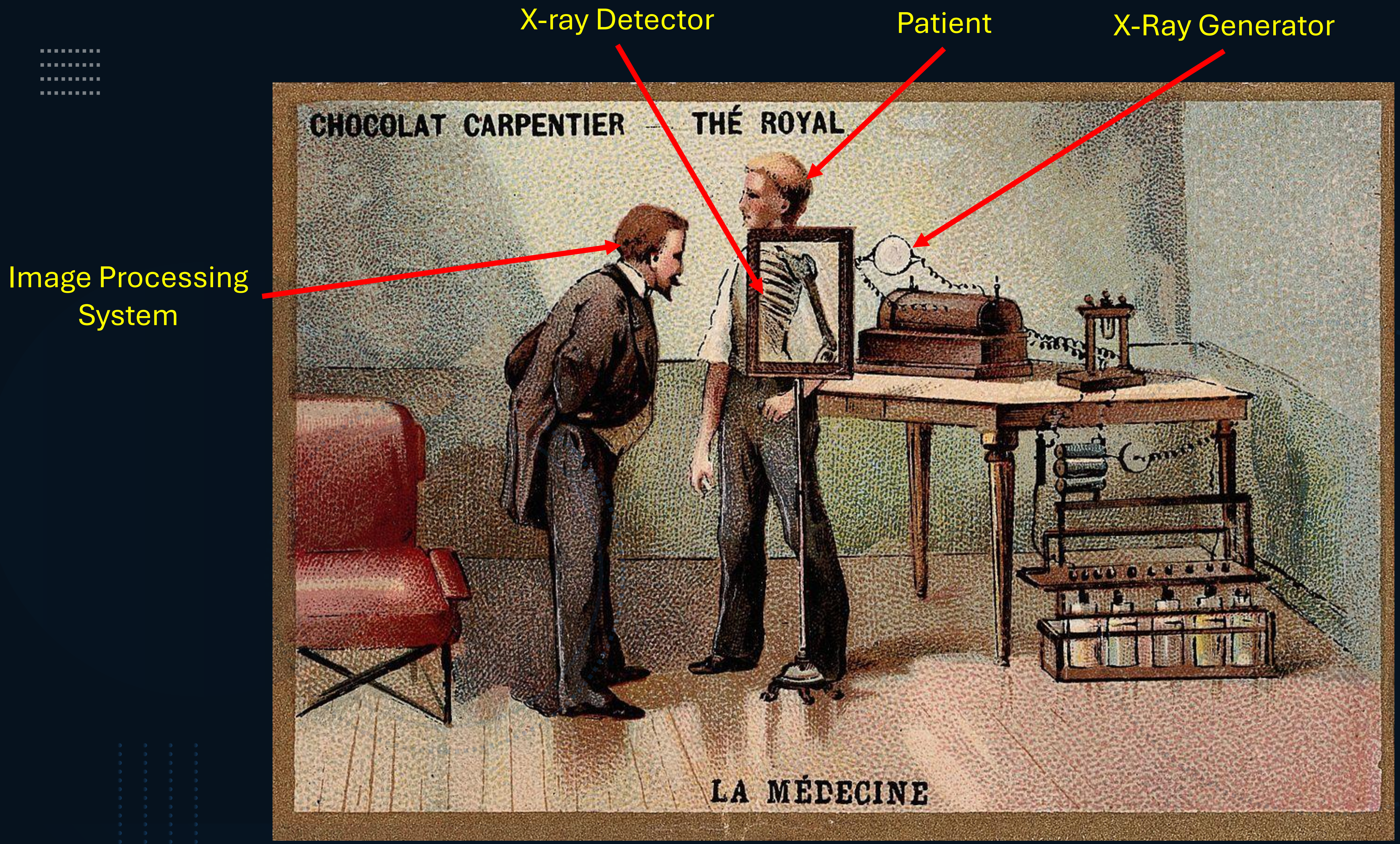
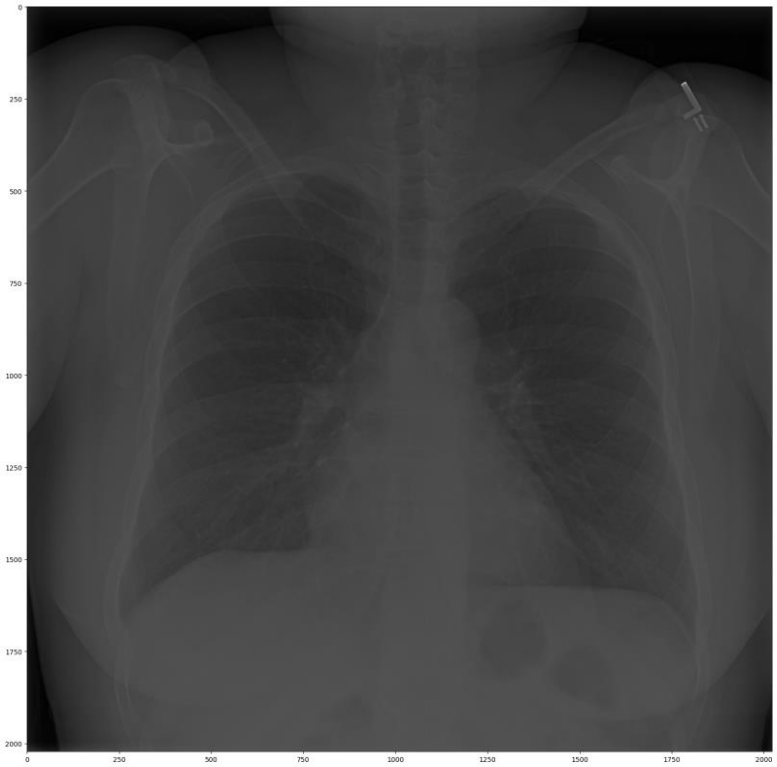
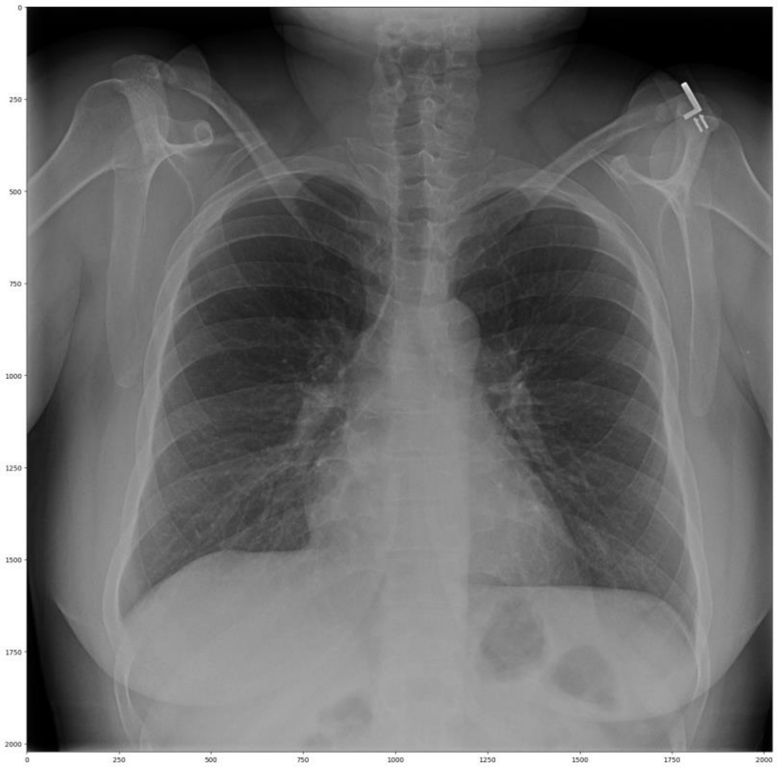


Image source: Wikipedia Commons

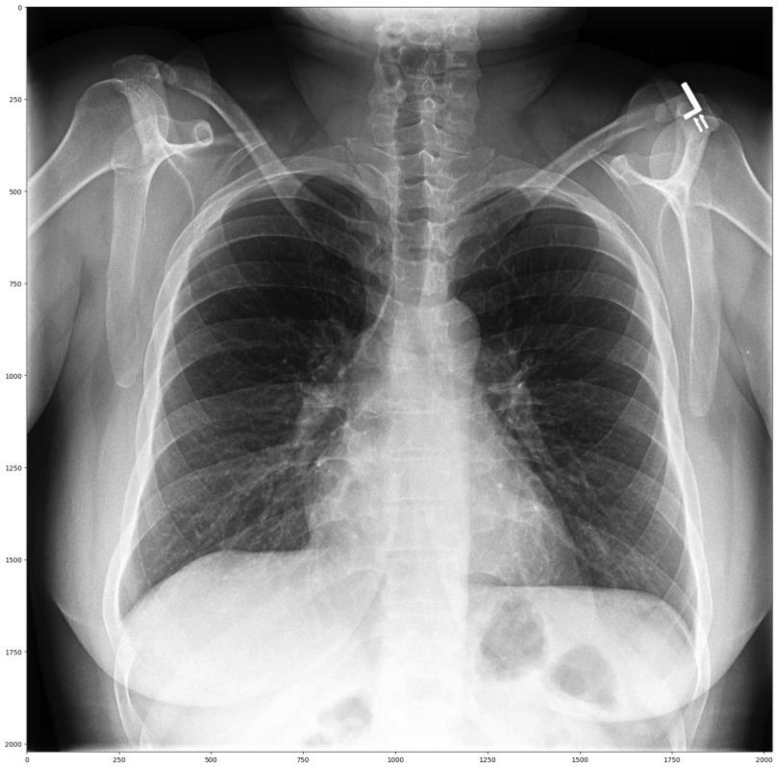
A

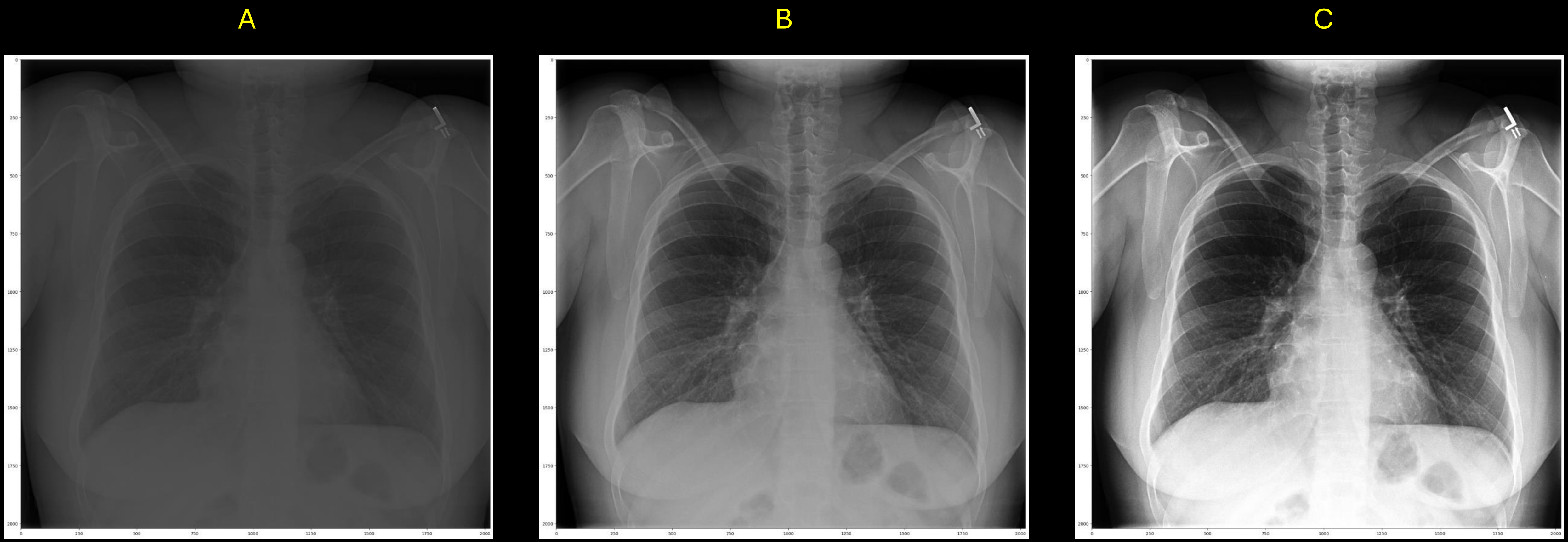

0
250
500
750
1000
1250
1500
1750
2000
0
250
500
750
1000
1250
1500
1750
2000

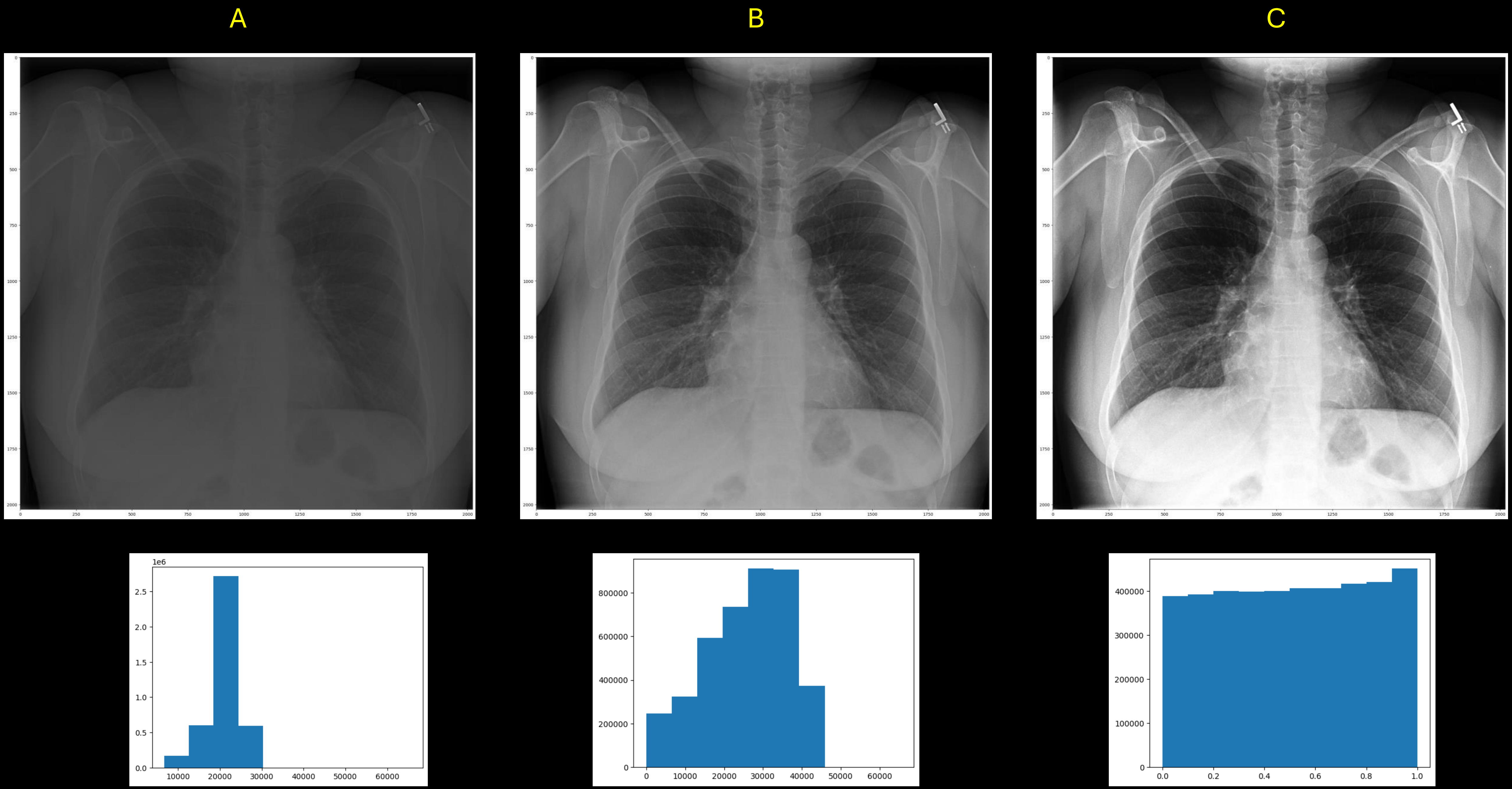
B

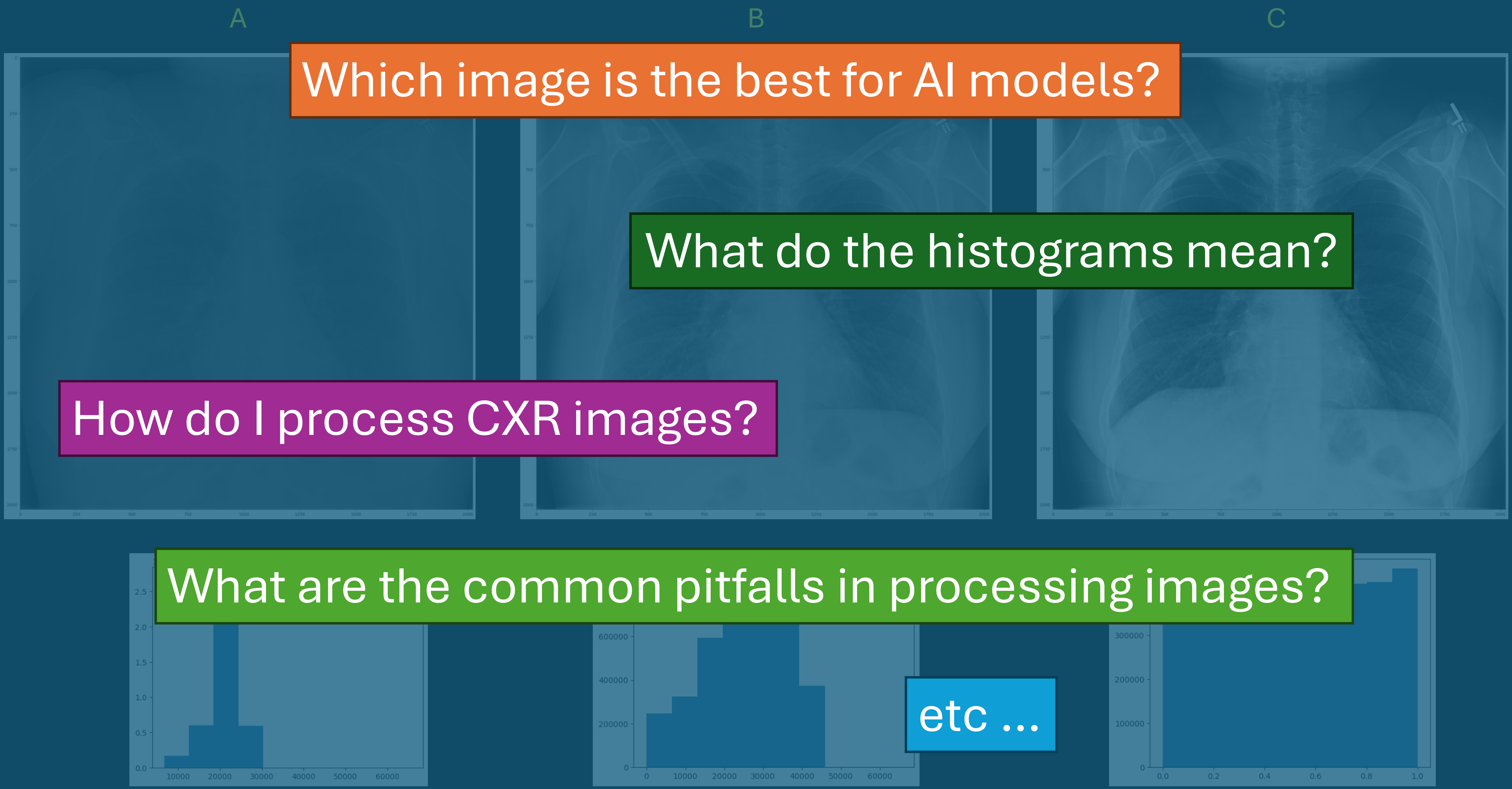

0
250
500
750
1000
1250
1500
1750
2000
0
250
500
750
1000
1250
1500
1750
2000

C


0
250
500
750
1000
1250
1500
1750
2000
0
250
500
750
1000
1250
1500
1750
2000


A
B
C
0
250
500
750
1000
1250
1500
1750
2000


A
B
C
0
250
500
750
1000
1250
1500
1750
2000
1e6
2.5
2.0
1.5
1.0
0.5
0.0
10000
20000
30000
40000
50000
60000
800000
600000
400000
200000
0
0
10000
20000
30000
40000
50000
60000
400000
300000
200000
100000
0
0.0
0.2
0.4
0.6
0.8
1.0



Which image is the best for AI models?

What do the histograms mean?

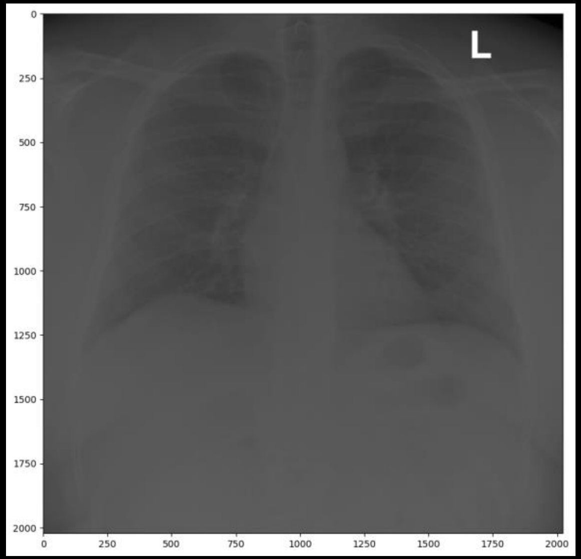
How do I process CXR images?

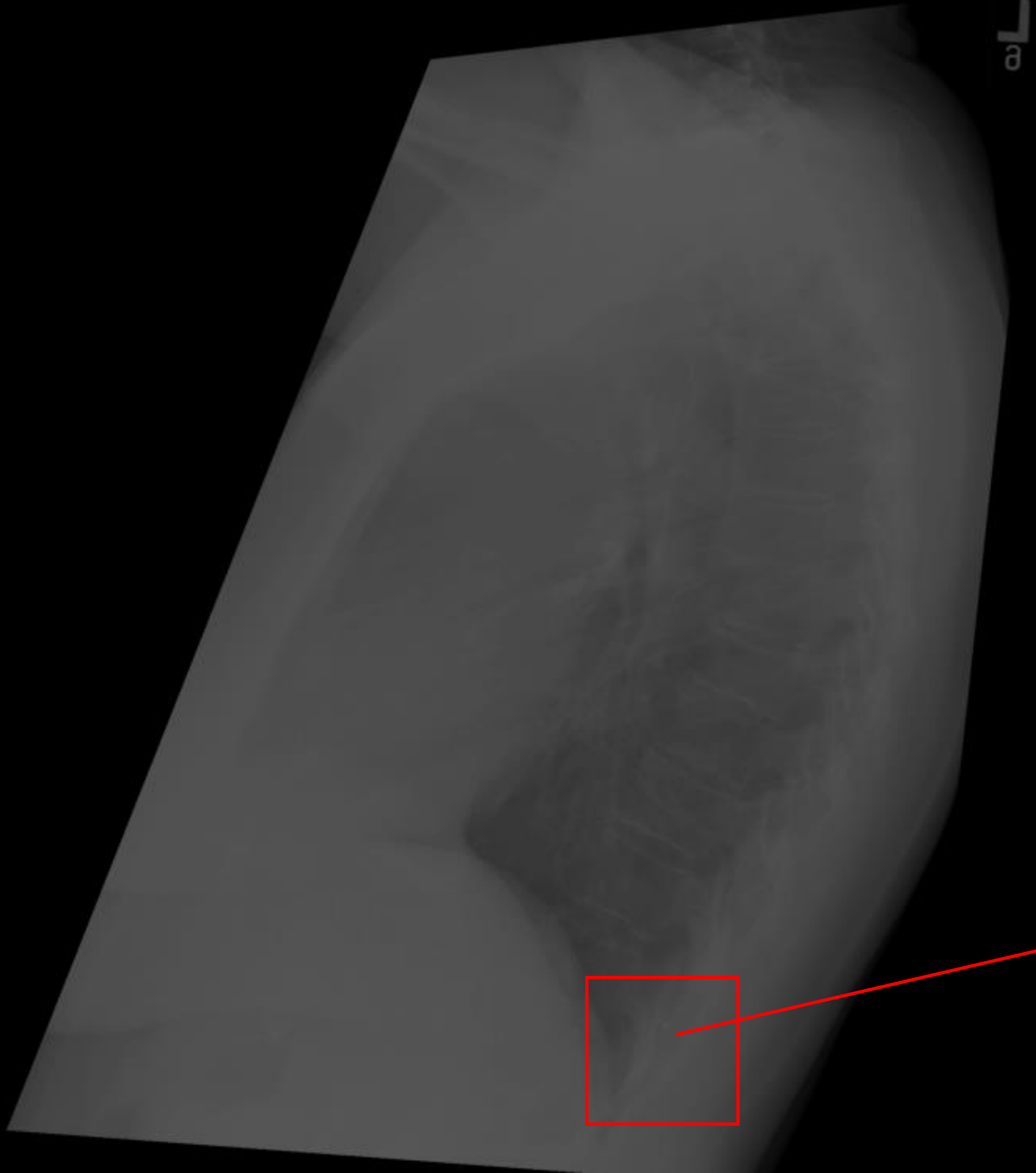
What are the common pitfalls in processing images?

etc ...

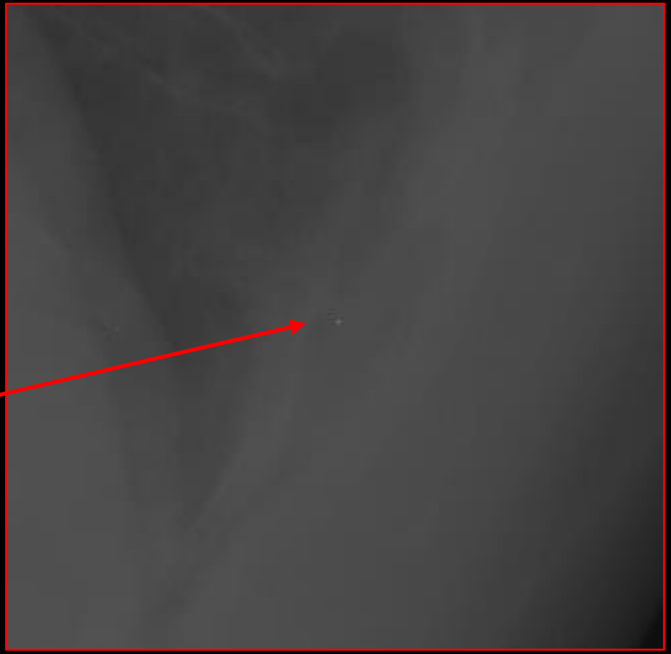
Our experience

- Pixel distribution in **PNG images** show anomalies:
  - Some pixels have values >>> 100 SD
- Equalisation, normalisation or standardisation methods performed on the PNG images don’t result in the optimal output
  - Hence, datapoints are deleted from the dataset

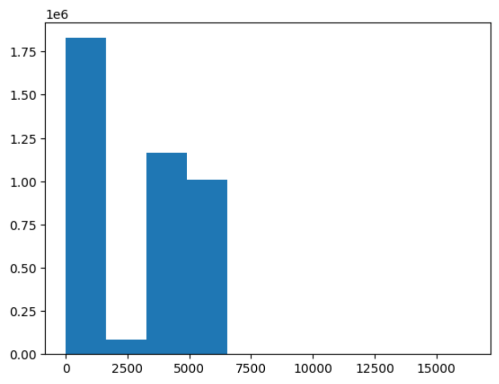


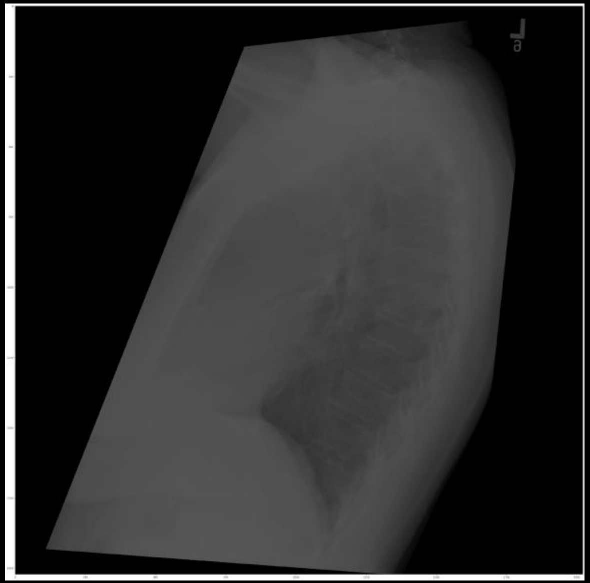


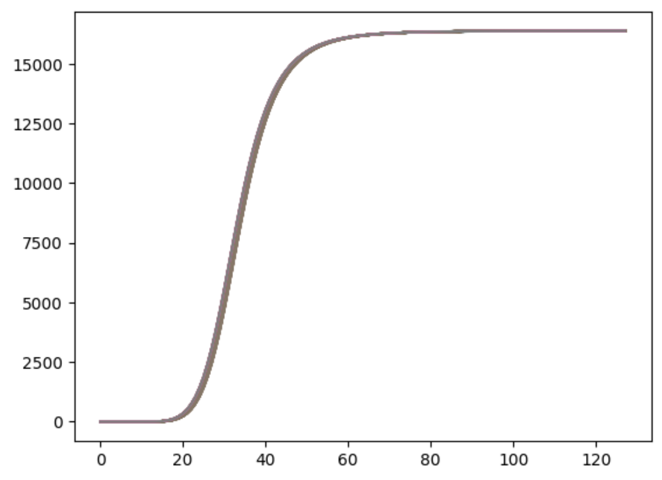
- Found 1 pixel with extremely high value;
- Not always found in the marker as previously thought;



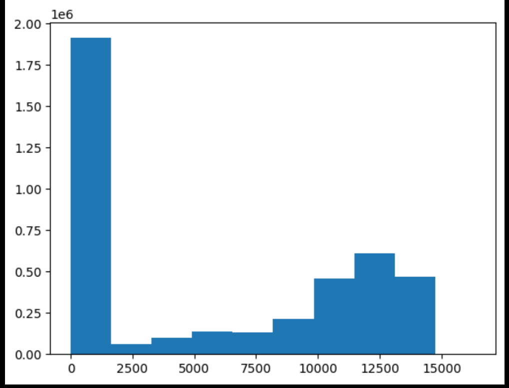
From PNGs Back to DICOMs

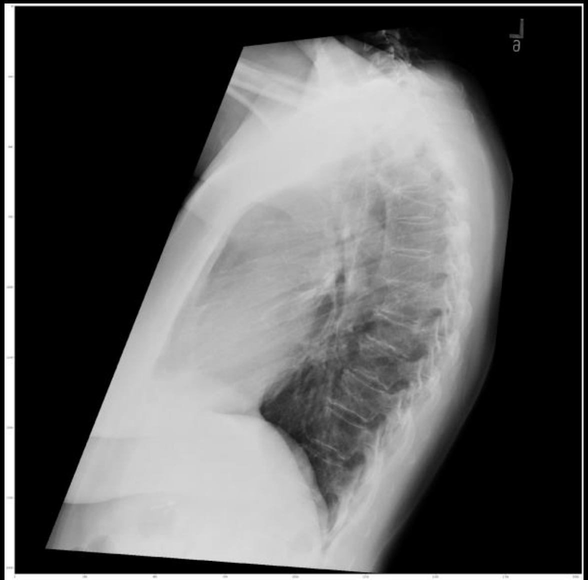




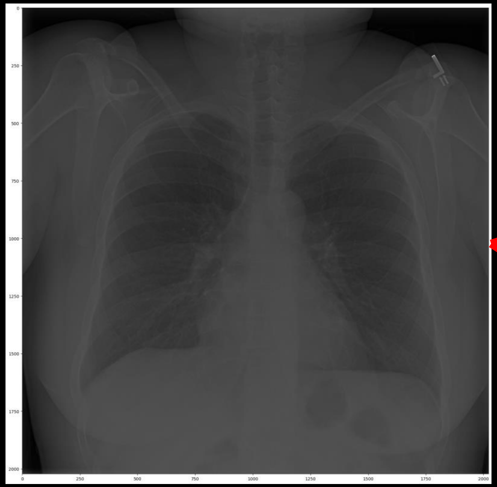


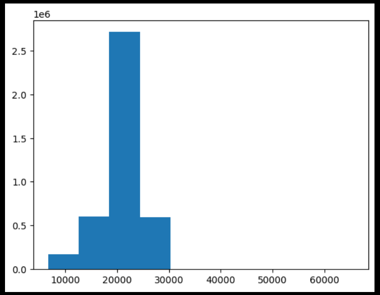
VOI LUT



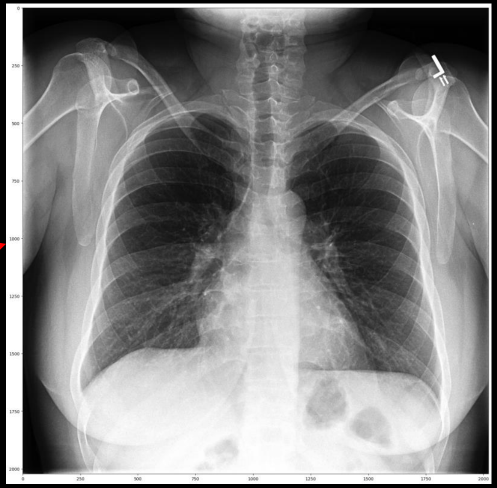


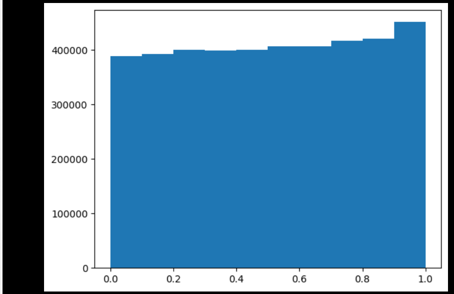
Raw Pixel Data



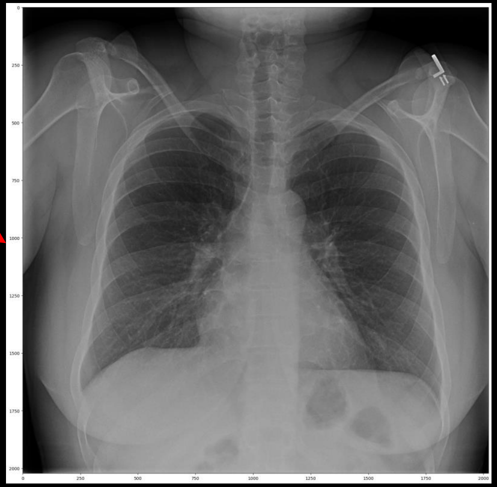


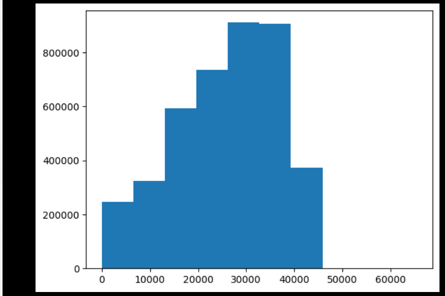
Histogram Equalization

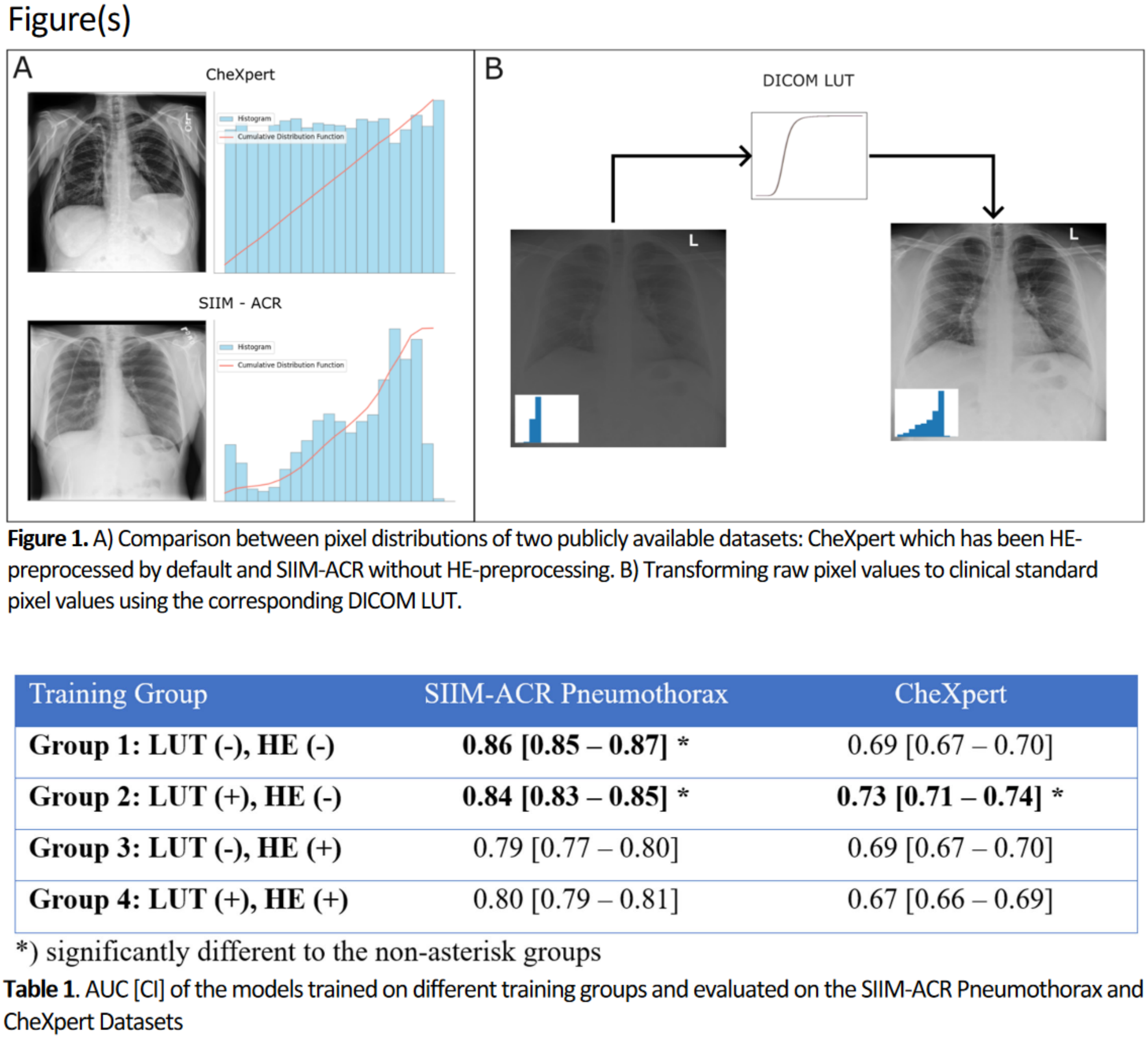




DICOM VOI LUT (Values Of Interest)







Figure(s)

A
CheXpert
Histogram
Cumulative Distribution Function
SIIM - ACR
Histogram
Cumulative Distribution Function
B
DICOM LUT
L
L

Figure 1. A) Comparison between pixel distributions of two publicly available datasets: CheXpert which has been HE-preprocessed by default and SIIM-ACR without HE-preprocessing. B) Transforming raw pixel values to clinical standard pixel values using the corresponding DICOM LUT.

Training Group	SIIM-ACR Pneumothorax	CheXpert
Group 1: LUT (-), HE (-)	**0.86 [0.85 – 0.87] ***	0.69 [0.67 – 0.70]
Group 2: LUT (+), HE (-)	**0.84 [0.83 – 0.85] ***	**0.73 [0.71 – 0.74] ***
Group 3: LUT (-), HE (+)	0.79 [0.77 – 0.80]	0.69 [0.67 – 0.70]
Group 4: LUT (+), HE (+)	0.80 [0.79 – 0.81]	0.67 [0.66 – 0.69]

*) significantly different to the non-asterisk groups

Table 1. AUC [CI] of the models trained on different training groups and evaluated on the SIIM-ACR Pneumothorax and CheXpert Datasets

DICOM VOI LUT

1. Increases model performance
2. Increases model generalizability

Dapamede, et al. 2024

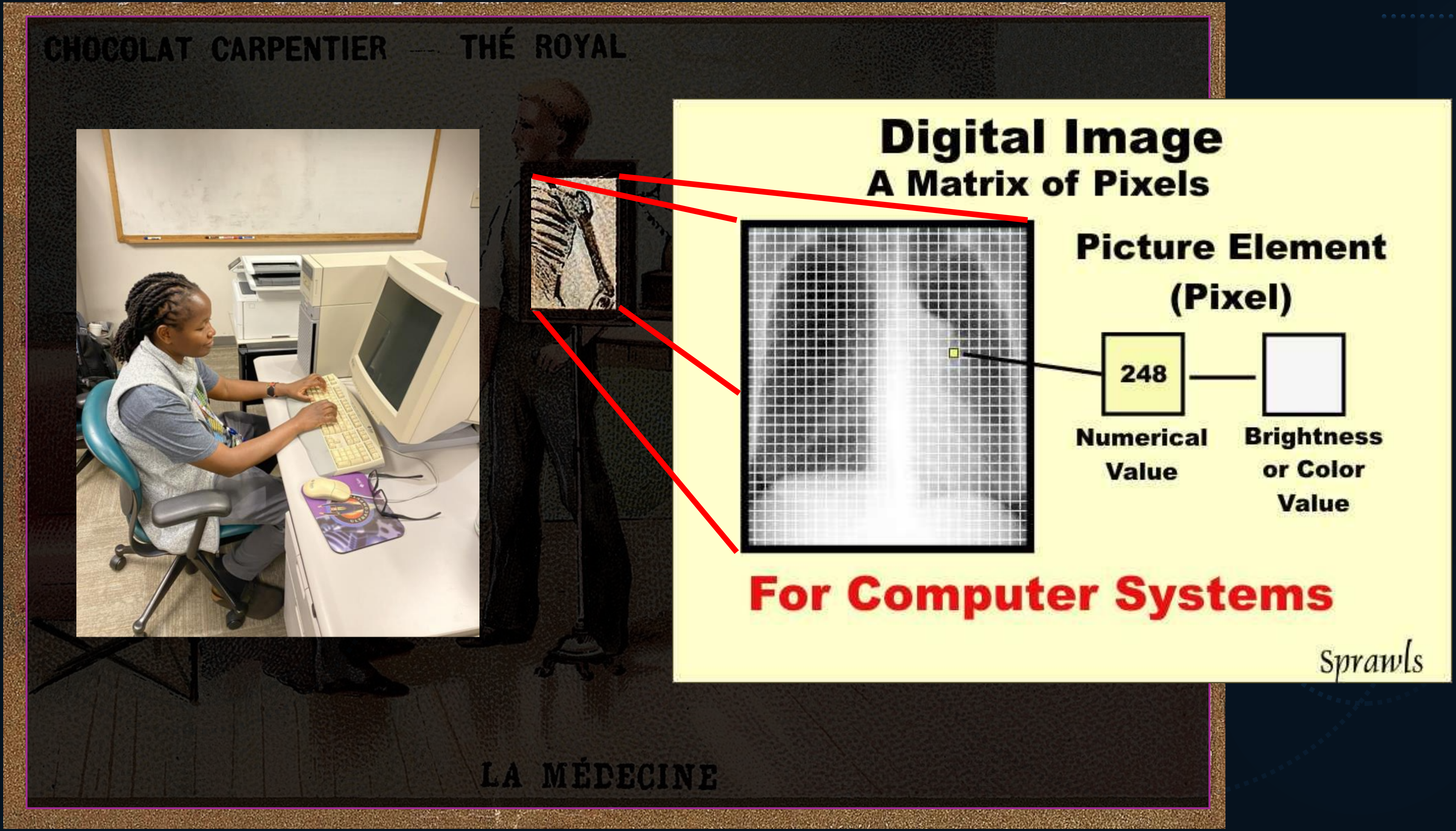


Image source: Wikipedia Commons, Sprawls

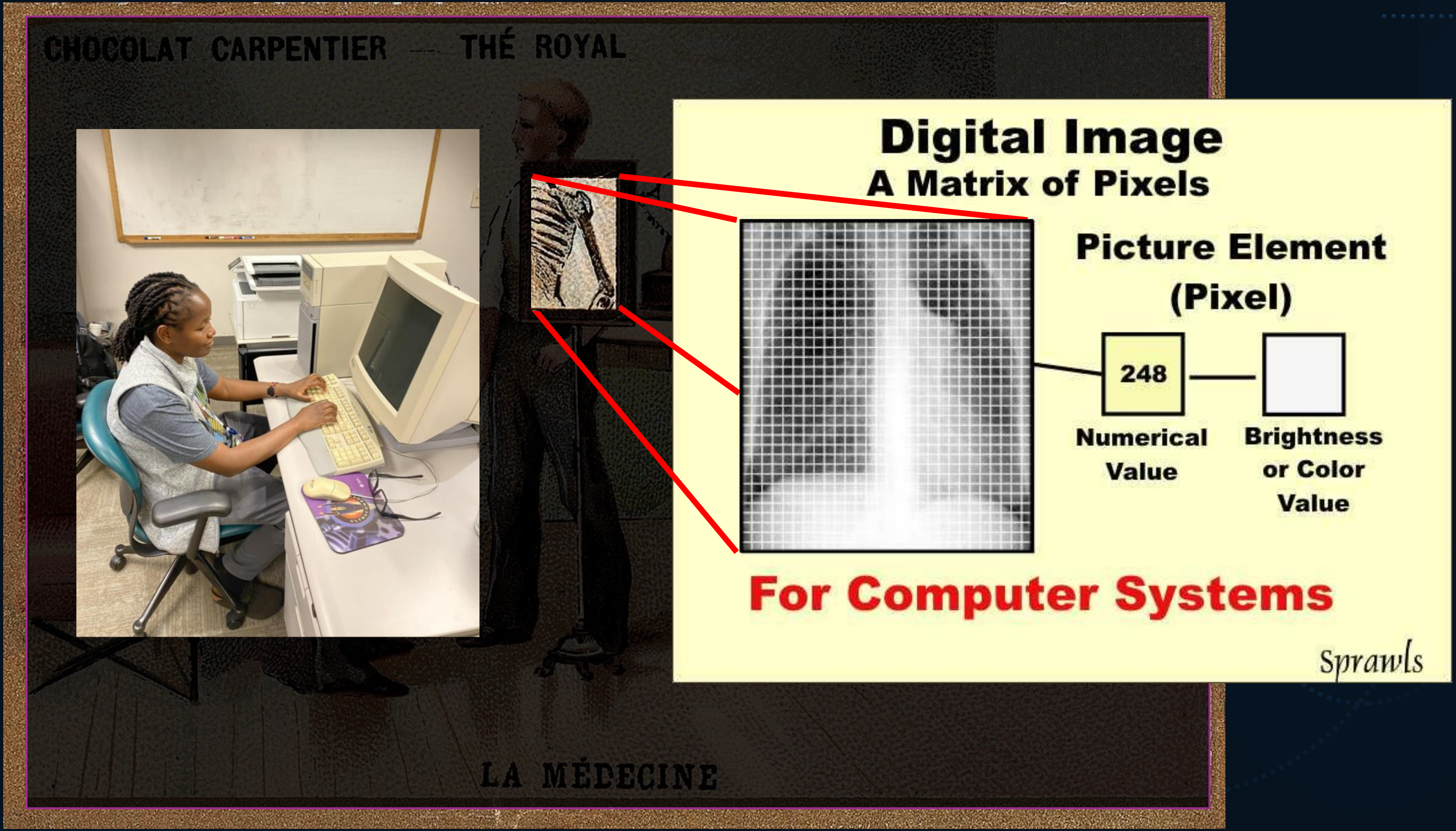
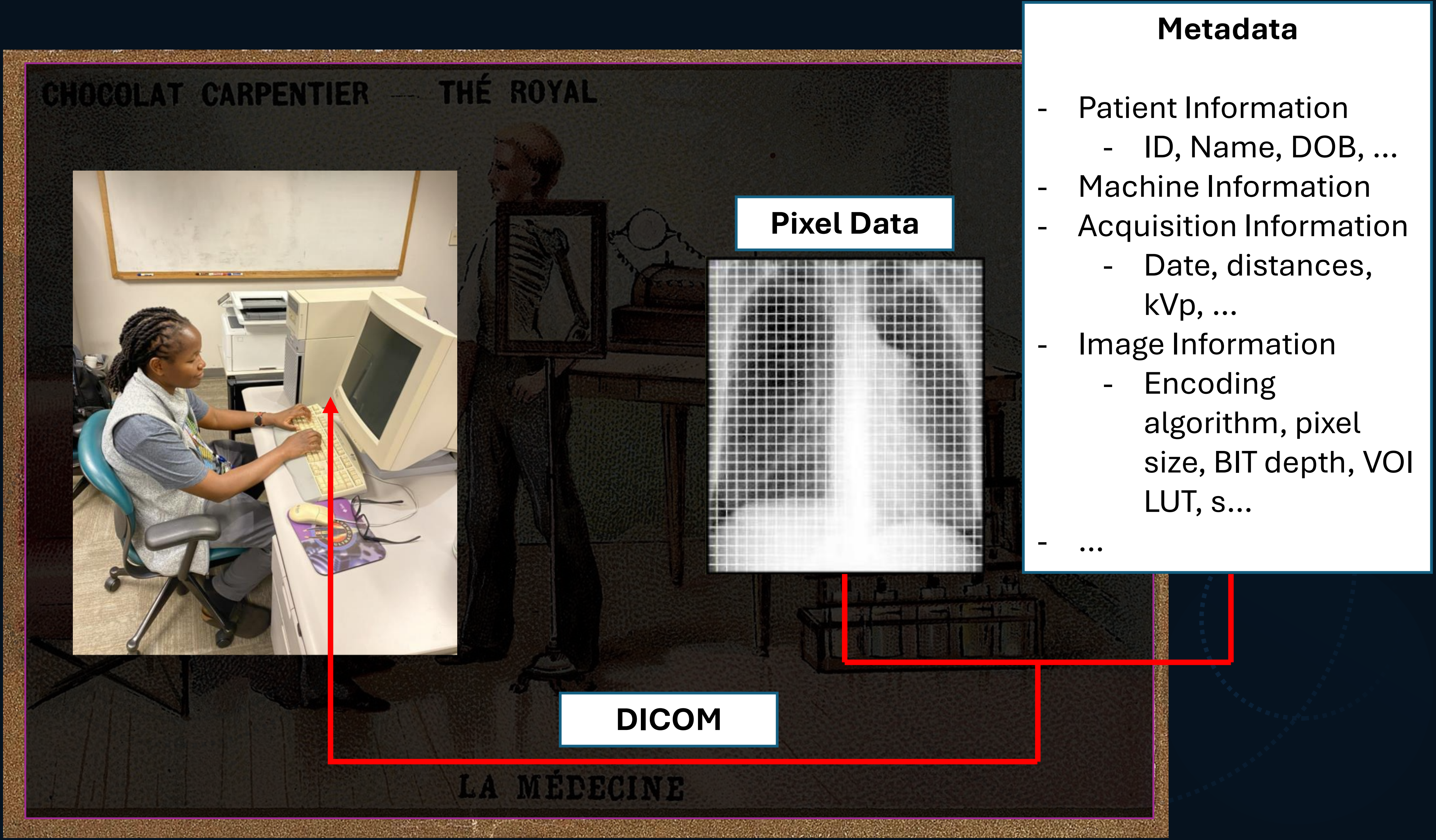


Image source: Wikipedia Commons, Sprawls



Pixel Data

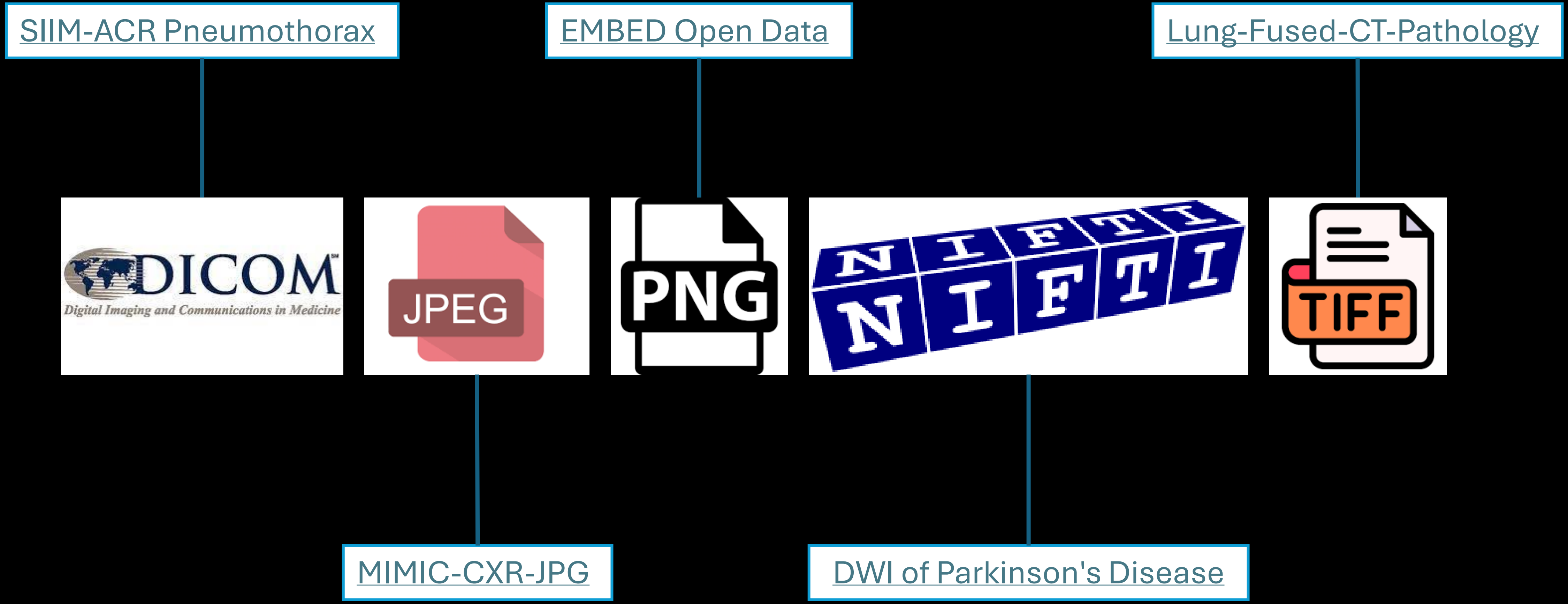
Metadata

- Patient Information
  - ID, Name, DOB, ...
- Machine Information
- Acquisition Information
  - Date, distances, kVp, ...
- Image Information
  - Encoding algorithm, pixel size, BIT depth, VOI LUT, s...
- ...

DICOM

Image source: Wikipedia Commons, Sprawls

Different File Types (with examples from public repositories)



SIIM-ACR Pneumothorax

EMBED Open Data

Lung-Fused-CT-Pathology

MIMIC-CXR-JPG

DWI of Parkinson's Disease

Hands On

1. Basics of working with DICOM files
2. DICOM Image Preprocessing
3. Standard Normalization Techniques