

May 8th , 2017



DAEN
690

MACHINE LEARNING ANALYSIS OF RADIOLOGY SCANS FOR AUTOMATED DECISION SUPPORT IN LUNG CANCER DETECTION

-LUNG CANCER SLEUTHS

The Lung Cancer Sleuths



ANIRUDH



KC



VAMSHI



PRAGNAKAR

OUTLINE

DAEN
690

- ▶ Introduction
- ▶ Problem
- ▶ Objectives
- ▶ Data
- ▶ Models

INTRODUCTION

- ▶ In the United States, lung cancer strikes 225,000 people every year, and accounts for \$12 billion in health care costs.
- ▶ Early detection is crucial to give patients the best chance at recovery and survival.
- ▶ Cancer Moonshot, a new initiative to make a decade's worth of progress in cancer prevention, diagnosis, and treatment in just 5 years.
- ▶ Kaggle in collaboration with Cancer Moonshot is convening the data science and medical communities to develop lung cancer detection algorithms.

PROBLEM?

- ▶ Present image assessments are identifying non cancerous lung lesions as potentially cancerous leading to
 - ▶ Unnecessary patient anxiety
 - ▶ Additional follow-up Imaging and interventional treatments
- ▶ I.E the current semi automated-detection technology has a very high False positive rate

OBJECTIVES

- ▶ Using a data set of thousands of high-resolution lung scans provided by the National Cancer Institute, develop algorithms that accurately determine when lesions in the lungs are cancerous.
- ▶ Build models that can act as decision support to the doctors in lung cancer detection
- ▶ By this we intend to get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

DATA

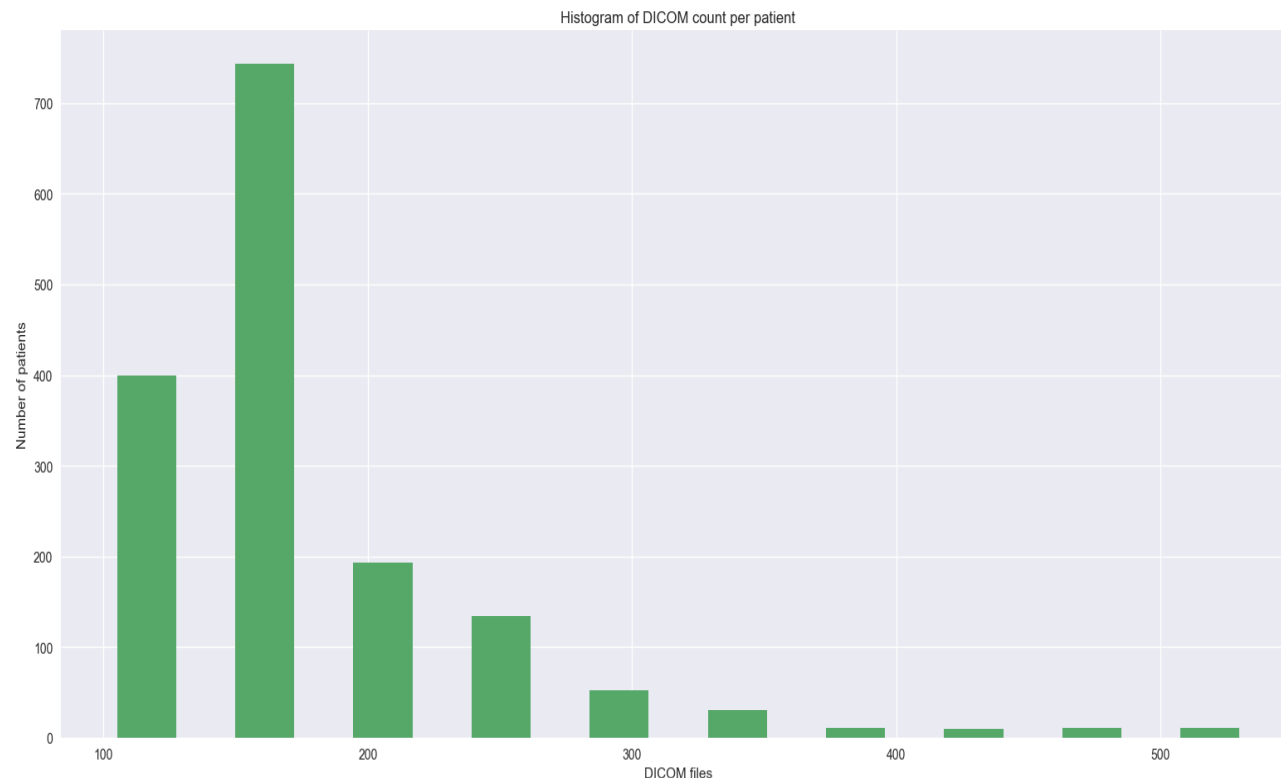
- ▶ Data source
 - ▶ Provided by Kaggle as part of competition.
- ▶ CT scans in DICOM(Digital Imaging and Communications in Medicine) format
 - ▶ Sample data
 - ▶ 20 patient records. Approx. (1.5 GB)
 - ▶ Master copy
 - ▶ 140 GB (approx.)
 - ▶ 1596 patients records provided
 - ▶ 1398 training data set
 - ▶ 198 trial data set

► Meta data included in one DICOM File.

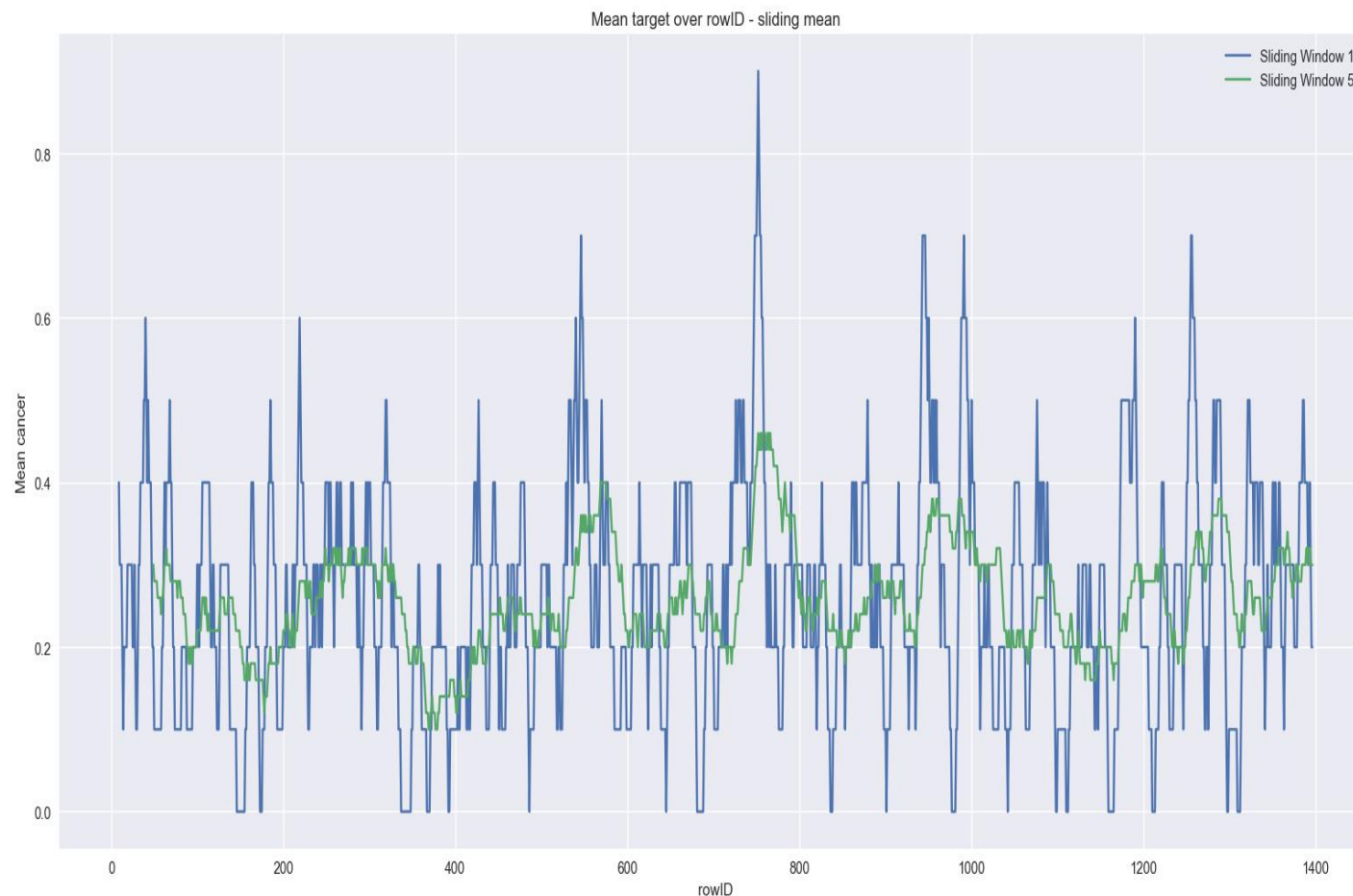
1. Patient ID
2. ~~Patient's Birth date~~
3. Pixel Data
4. Image Position, Pixel Spacing
5. Rescale Intercept, Rescale slope

(0008, 0005) Specific Character Set	CS: 'ISO_IR 100'
(0008, 0016) SOP Class UID	UI: CT Image Storage
(0008, 0018) SOP Instance UID	UI: 1.2.840.113654.2.55.1609788432169499538152465
51971491067836	
(0008, 0060) Modality	CS: 'CT'
(0008, 103e) Series Description	LO: 'Axial'
(0010, 0010) Patient's Name	PN: '0a38e7597ca26f9374f8ea2770ba870d'
(0010, 0020) Patient ID	LO: '0a38e7597ca26f9374f8ea2770ba870d'
(0010, 0030) Patient's Birth Date	DA: '19000101'
(0018, 0060) KVP	DS: ''
(0020, 000d) Study Instance UID	UI: 2.25.1314839497924593748416559454002566892718
9308811493143066650	
(0020, 000e) Series Instance UID	UI: 2.25.5329856372890633558583375240523888470449
8238267638676785109	
(0020, 0011) Series Number	IS: '2'
(0020, 0012) Acquisition Number	IS: '1'
(0020, 0013) Instance Number	IS: '26'
(0020, 0020) Patient Orientation	CS: ''
(0020, 0032) Image Position (Patient)	DS: ['-172.500000', '-176.100006', '-34.540001']
(0020, 0037) Image Orientation (Patient)	DS: ['1.000000', '0.000000', '0.000000', '0.000000', '1.000000', '0.000000']
(0020, 0052) Frame of Reference UID	UI: 2.25.1164299566818961382999986368386572209817
7327703743652111991	
(0020, 1040) Position Reference Indicator	LO: 'SN'
(0020, 1041) Slice Location	DS: '-34.540001'
(0028, 0002) Samples per Pixel	US: 1
(0028, 0004) Photometric Interpretation	CS: 'MONOCHROME2'
(0028, 0010) Rows	US: 512
(0028, 0011) Columns	US: 512
(0028, 0030) Pixel Spacing	DS: ['0.625000', '0.625000']
(0028, 0100) Bits Allocated	US: 16
(0028, 0101) Bits Stored	US: 16
(0028, 0102) High Bit	US: 15
(0028, 0103) Pixel Representation	US: 1
(0028, 0120) Pixel Padding Value	US: 63536
(0028, 1050) Window Center	DS: '40'
(0028, 1051) Window Width	DS: '350'
(0028, 1052) Rescale Intercept	DS: '-1024'
(0028, 1053) Rescale Slope	DS: '1'
(7fe0, 0010) Pixel Data	OW: Array of 524288 bytes

- ▶ Each patient scanned on different machines
- ▶ Different no.of slices of CT scan for different patients
- ▶ Image resolution remained same for all slices and all patients at (512X512)
- ▶ A scan may have a pixel spacing of [2.5, 0.5, 0.5], which means that the distance between slices is 2.5 millimeters. For a different scan this may be [1.5, 0.725, 0.725]



- ▶ Checking if there is relationship between patient ID and cancer
- ▶ Number of training patients: 1398
Cancer rate: 25.91%



DATA PREPROCESSING

DAEN
690

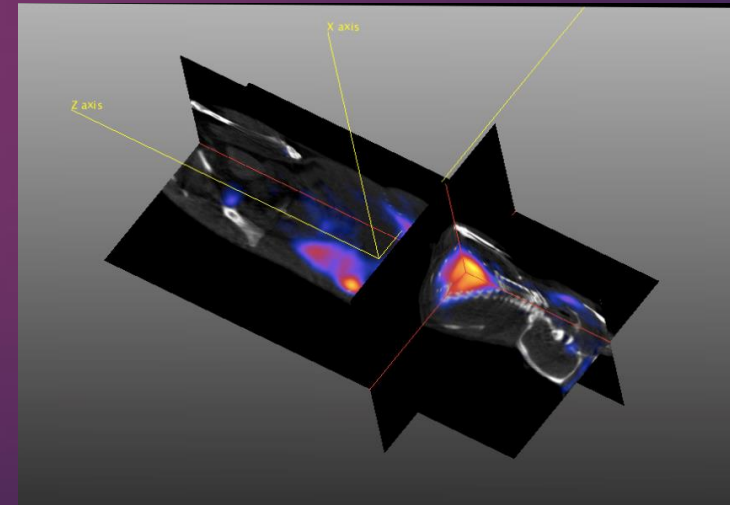
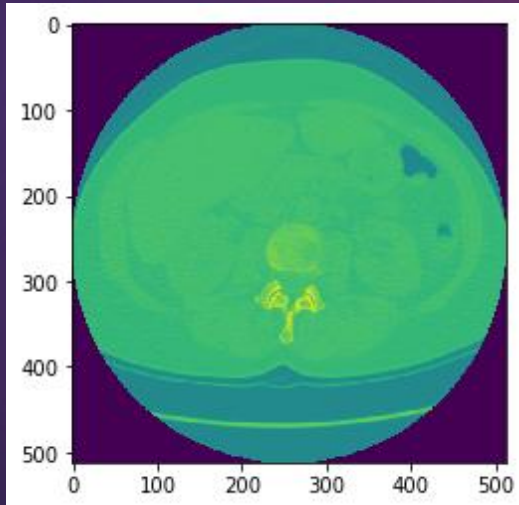
(Initial image Pixel size) No . of slices

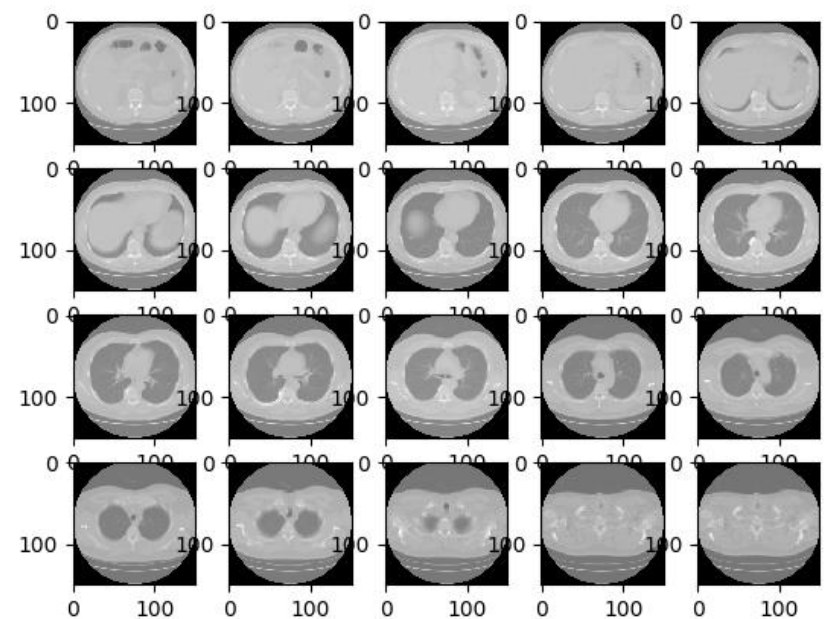
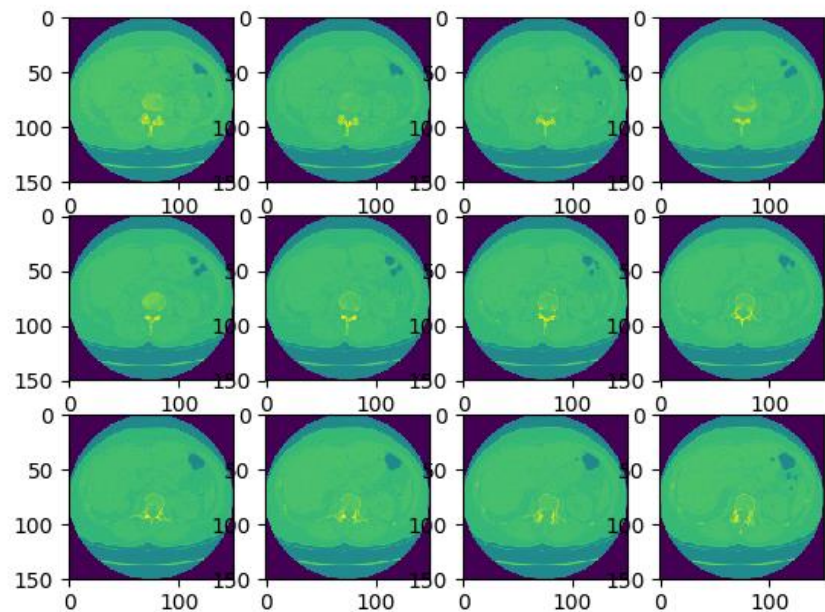
- ▶ (512, 512) 134
- ▶ (512, 512) 128
- ▶ (512, 512) 133
- ▶ (512, 512) 110



Images after resizing & Normalizing

- ▶ (150,150) 20
- ▶ (150,150) 20
- ▶ (150,150) 20
- ▶ (150,150) 20

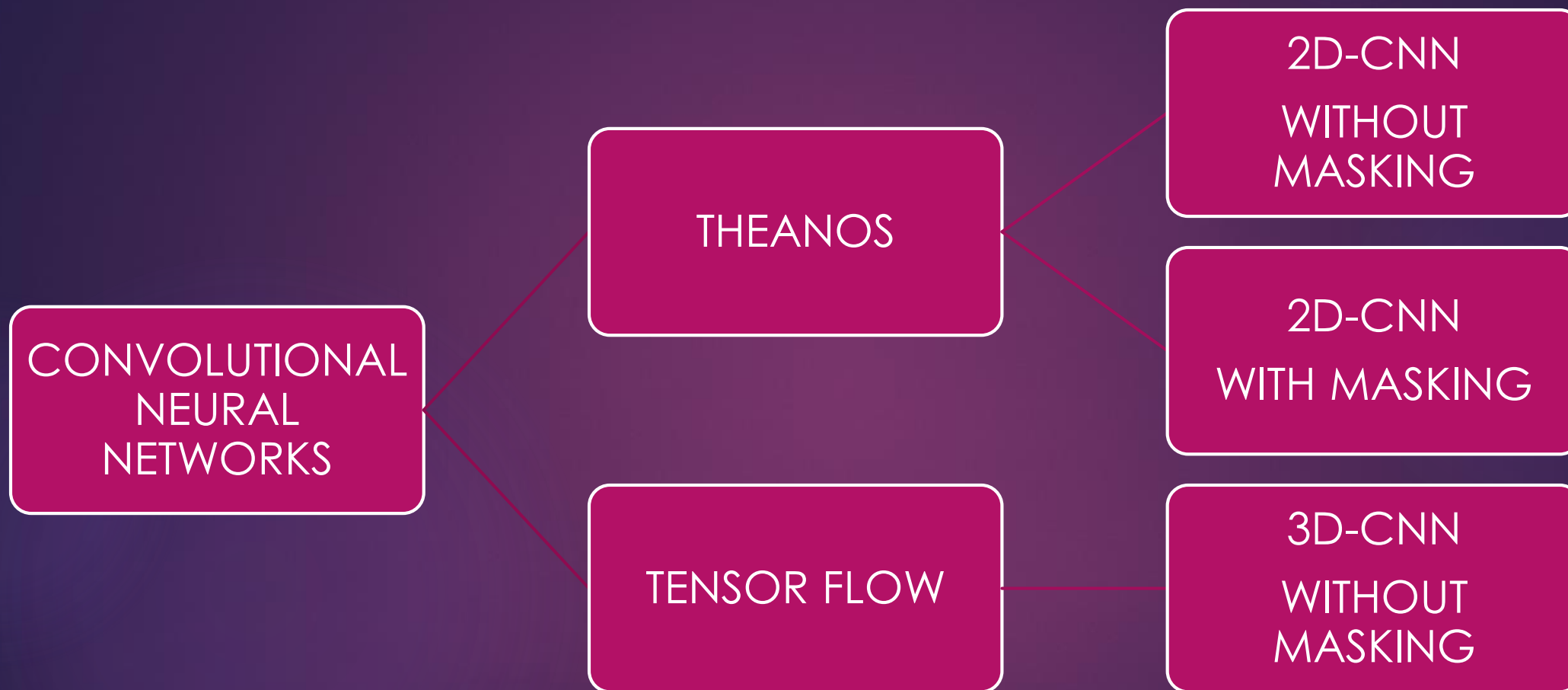


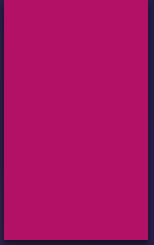


- Changed to greyscale in order to skip the complexities with colors during model building

MODELS

DAEN
690

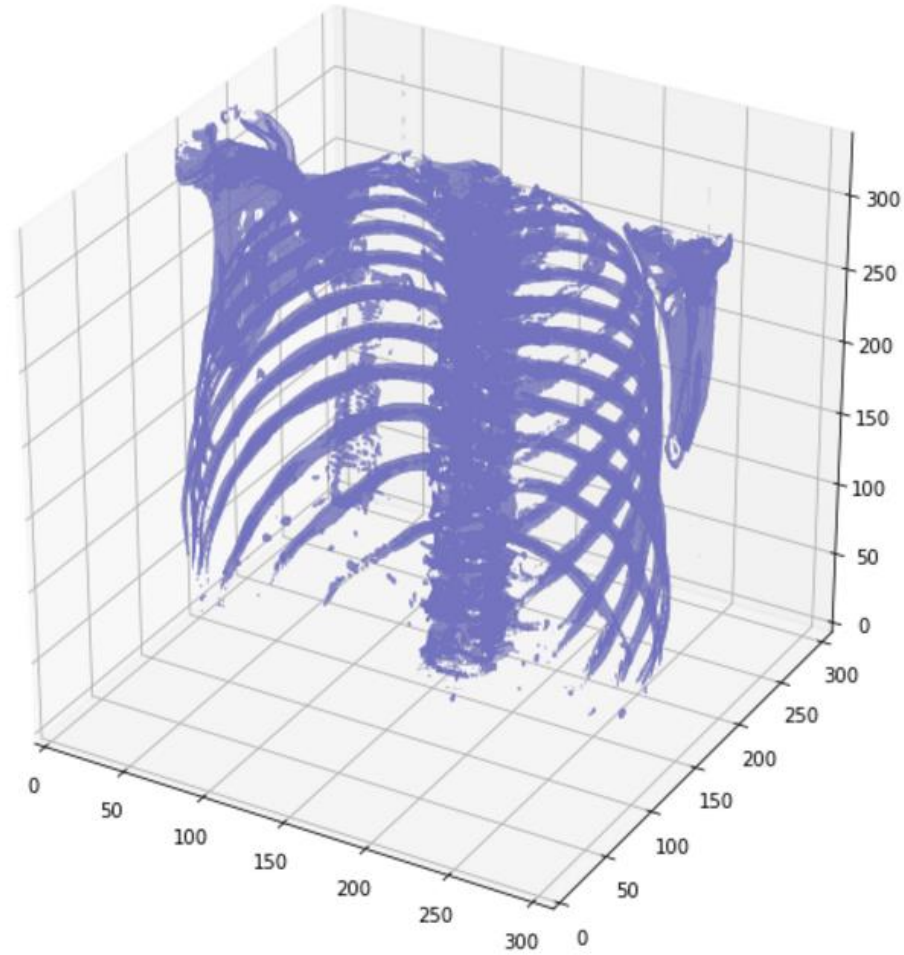


- 
- ▶ Identify Lung tissues
 - ▶ Standardize the slices
 - ▶ Sampling

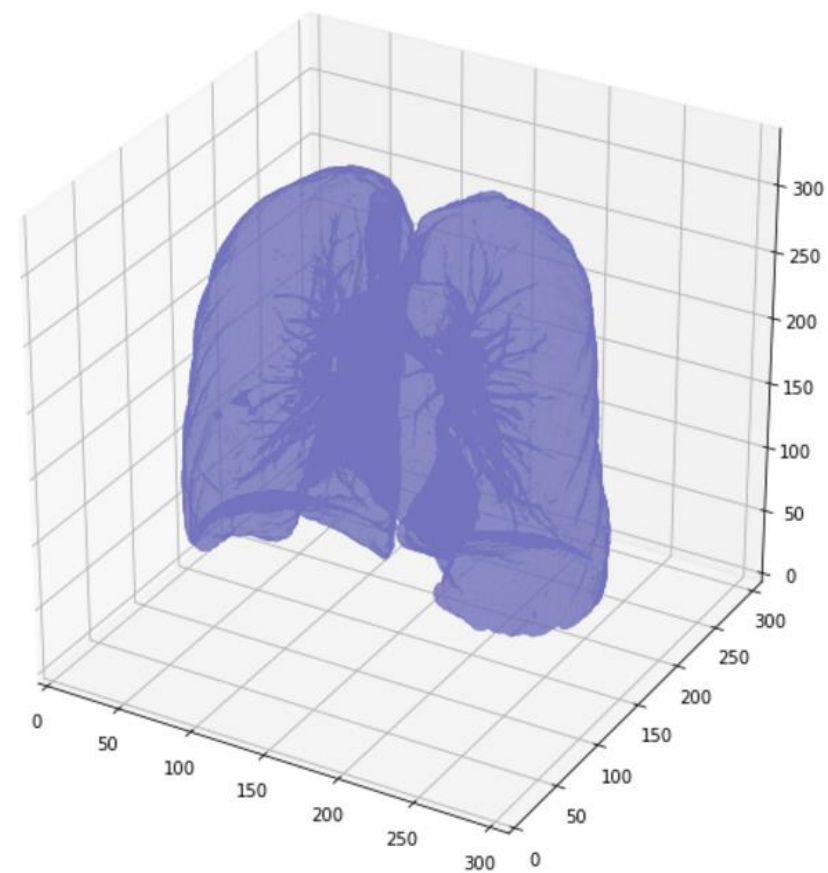
- ▶ Basics knowledge needed to understand CT scan and how 3D model generated
 - ▶ Hounsfield Unit (HU)
 - ▶ How to get 3D model from 2d samples provided
 - ▶ Segmentation

Substance	HU
Air	-1000
Lung	-500
Fat	-100 to -50
Water	0
CSF	15
Kidney	30
Blood	+30 to +45
Muscle	+10 to +40
Grey matter	+37 to +45
White matter	+20 to +30
Liver	+40 to +60
Soft Tissue, Contrast	+100 to +300
Bone	+700 (cancellous bone) to +3000 (cortical bone)

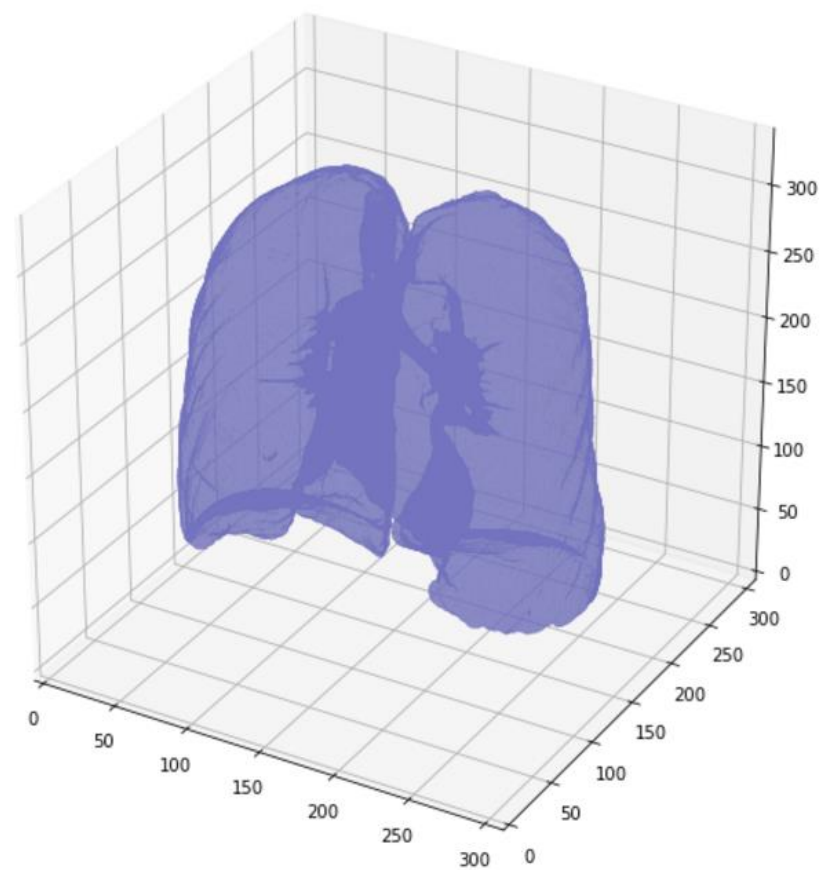
- Modeling the DICOM images to 3D by using HU unit



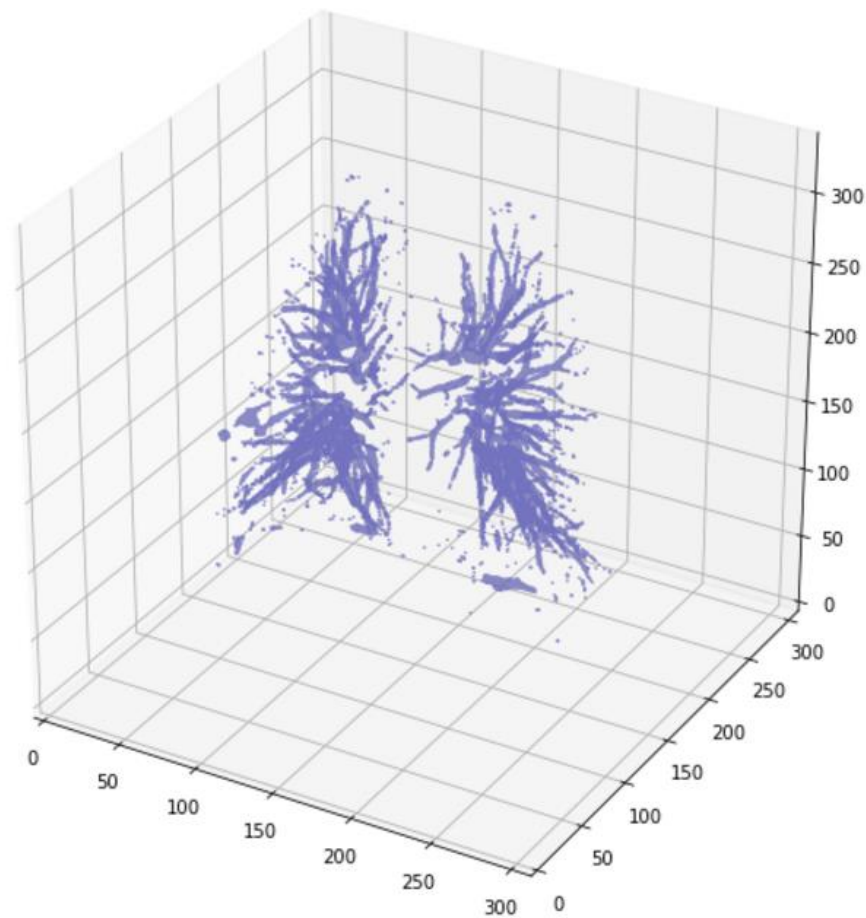
► Look at the insides



► Look at the inside



- ▶ This is the difference between the last 2 images and data points are extracted.



We would like to thank:

Professor F. Brett Berlin, Term Professor DAEN 690, GMU.

For guiding us at every step for successful completion and extending full support for providing high computational resources. Without him we wouldn't meet the deadline.

Dr. Jayshree Sarma, Interim Director, Research Computing, GMU.

For providing cluster access and for teaching us how to use the cluster efficiently for our needs. She spent hours sitting along and had been guiding us through obstacles.

Md Anindya Prodhan, Research Assistant, UVA.

For driving all the way from Charlottesville, to help us gain access to XCG, and helping us when we were struck with technical complications.

THANK YOU...

DAEN
690

*WE ARE OPEN FOR
QUESTIONS & DISCUSSION*

Happy Graduation...!!!*