EARLY DETECTION OF LUNG CANCER

Deep learning AA 2021/2022

DATA SCIENCE & SCIENTIFIC COMPUTING

Why deep learning for cancer detection

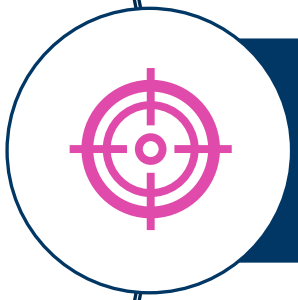
Detecting lung cancer has a huge impact on survival rate

Currently the work of reviewing the data must be performed by highly trained specialists

Potential for missed warning signs, particularly in the early stages.

LUNA Grand Challenge

LUNA Grand Challenge is the combination of an open dataset with higly-quality labels of patient CT scans



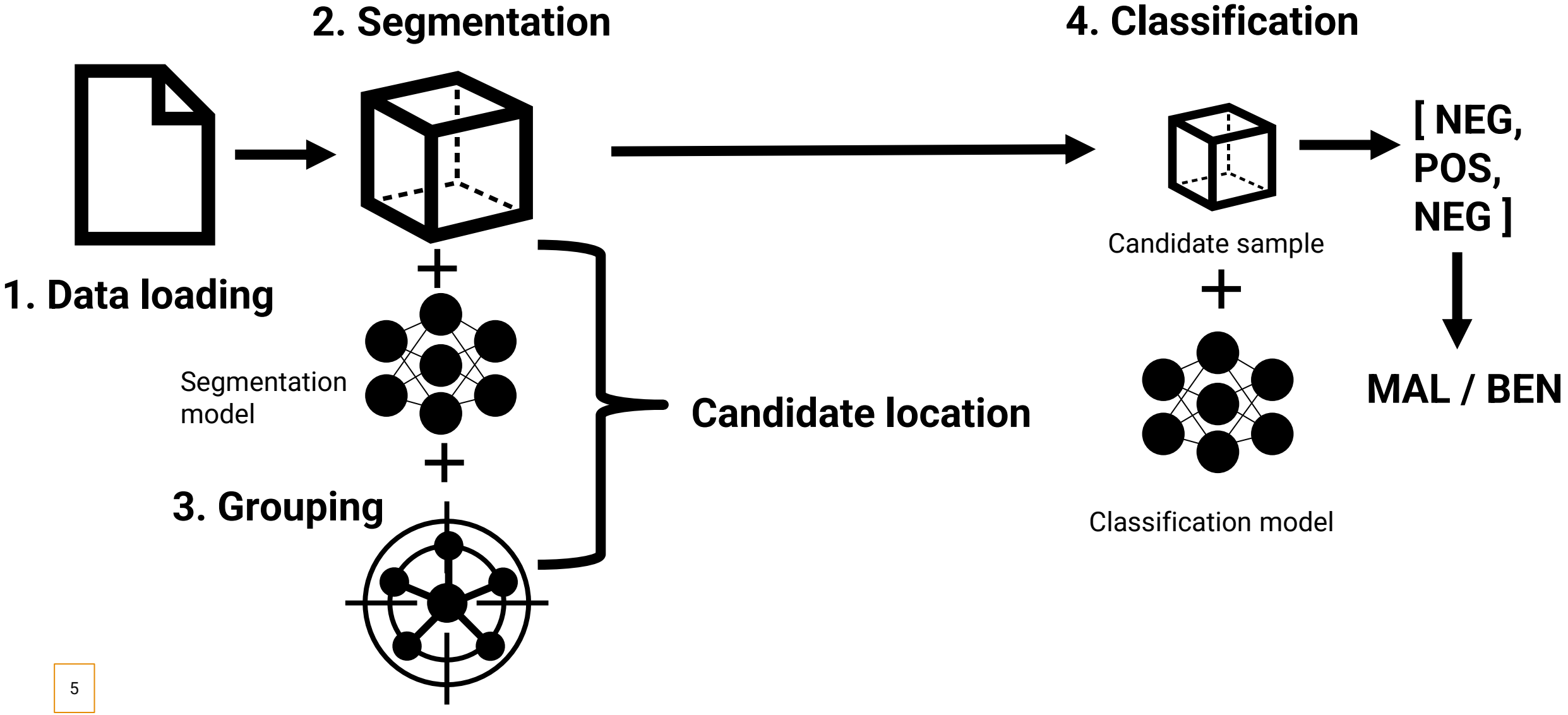
The goal is to encourage improvements in nodule detection

A total of 120GB of data

What is a nodule?

- Is any of the lumps and bumps inside someone's lung
- The definition limits the size of the nodule to 3cm or less
- A nodule can turn out to be benign or a malignant tumor

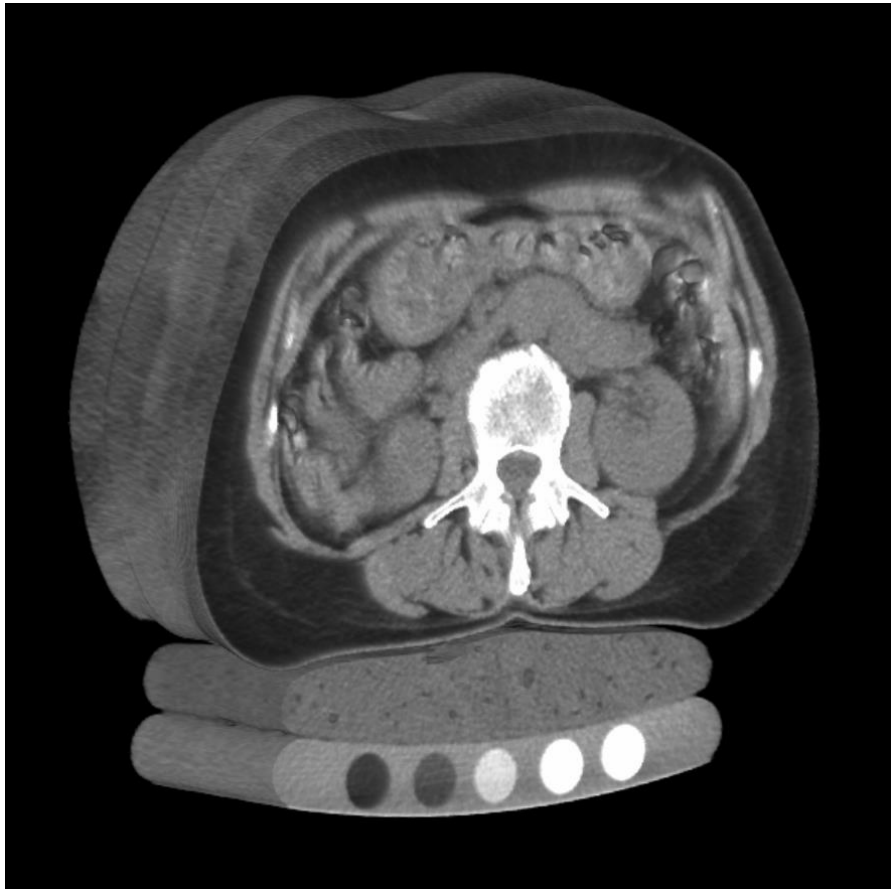
Brief model overview



DATASET

What is a CT scan?

- CT scans are essentially 3D X-rays, represented as a 3D array of single-channel data
- The main difference from X-rays is that CT scan retains the third dimension of the dat
- CT scan voxels are expressed in Hounsfields units, air is -1,0HU, water is 0HU, bone +1,0HU



A CT scan of human torso showing, from the top, skins, organs, spine

How CT scan are acquire

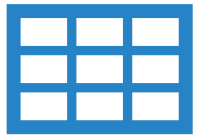
The bed the patient is resting on moves back and forth, allowing the scanner to image multiple slices of the patient

The way in which CT scanner measures distance along the head-to-foot axis is different than the other two axes

LUNA Dataset

Subset*

Zip files which contain all CT images. In each subset CT images are stored in MetaImage. Each .mhd is stored with a separate .raw binary file for the pixeldata



Candidates.csv

Csv file that contains nodule candidate per line. Each line holds the scan name, the x, y and z position of each candidate in world coordinates.

Annotations.csv

A csv file that contains one finding per line. Each line holds the SeriesUID of the scan, the x,y and z position of each finding in world coordinates, and the correspoding diameter in mm.

Exploring csv files

The candidates.csv file contains information about all lumps that potentially look like nodules, wether thos lumps are malignant or benign tumors.
The number of lines is 551066 which 1351 indicating malignancy (class = 1)

```
(base) silviobaratto@xps-13-9310:~/OneDrive/github/LUNA$ head candidates.csv
seriesuid,coordX,coordY,coordZ,class
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-56.08,-67.85,-311.92,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,53.21,-244.41,-245.17,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,103.66,-121.8,-286.62,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-33.66,-72.75,-308.41,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-32.25,-85.36,-362.51,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-26.65,-203.07,-165.07,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-74.99,-114.79,-311.92,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-16.14,-248.61,-239.55,0
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,135.89,-141.41,-252.2,0
```

Exploring csv files

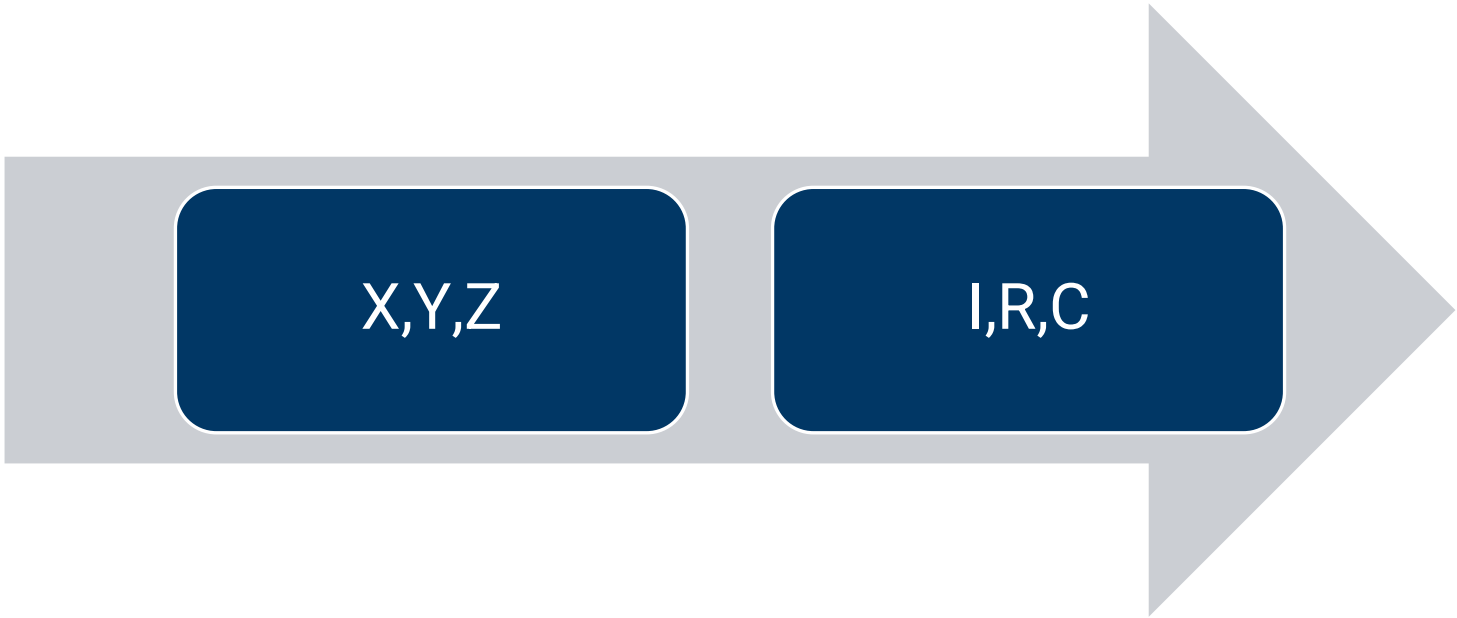
The annotations.csv file contain information about some of the candidates that have been flagged as nodules. We are interested in the diameter_mm information since we can include a representative spread of nodule sizes.

```
(base) silviobaratto@xps-13-9310:~/OneDrive/github$ head LUNA/annotations.csv
seriesuid,coordX,coordY,coordZ,diameter_mm
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,-128.6994211,-175.3192718,-298.3875064,5.651470635
1.3.6.1.4.1.14519.5.2.1.6279.6001.100225287222365663678666836860,103.7836509,-211.9251487,-227.12125,4.224708481
1.3.6.1.4.1.14519.5.2.1.6279.6001.100398138793540579077826395208,69.63901724,-140.9445859,876.3744957,5.786347814
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405,-24.0138242,192.1024053,-391.0812764,8.143261683
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405,2.441546798,172.4648812,-405.4937318,18.54514997
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405,90.93171321,149.0272657,-426.5447146,18.20857028
1.3.6.1.4.1.14519.5.2.1.6279.6001.100621383016233746780170740405,89.54076865,196.4051593,-515.0733216,16.38127631
1.3.6.1.4.1.14519.5.2.1.6279.6001.100953483028192176989979435275,81.50964574,54.9572186,-150.3464233,10.36232088
1.3.6.1.4.1.14519.5.2.1.6279.6001.102681962408431413578140925249,105.0557924,19.82526014,-91.24725078,21.08961863
```

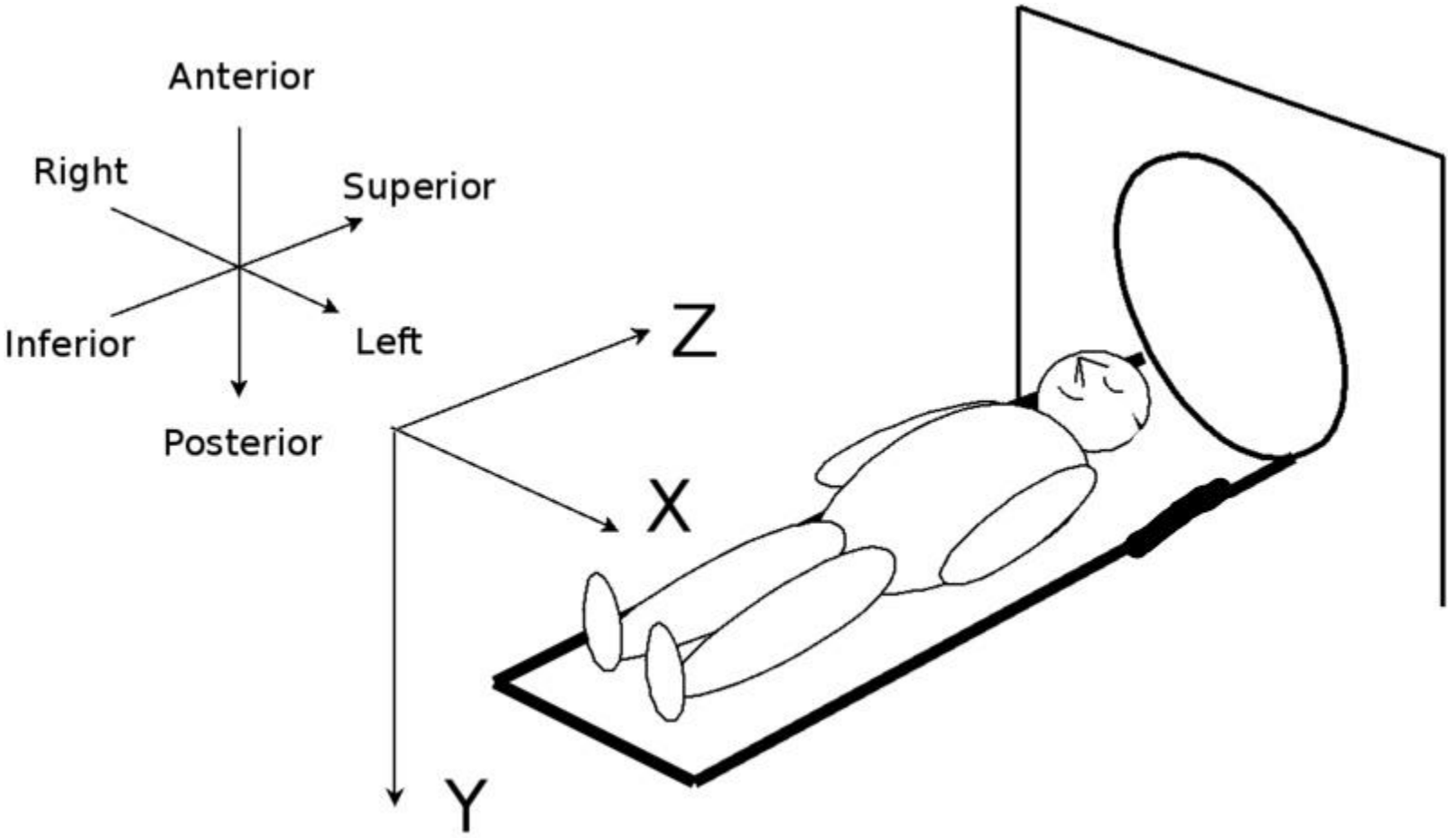
Different coordinate systems

CT scan data is expressed in voxel-address-based coordinate system (I,R,C)

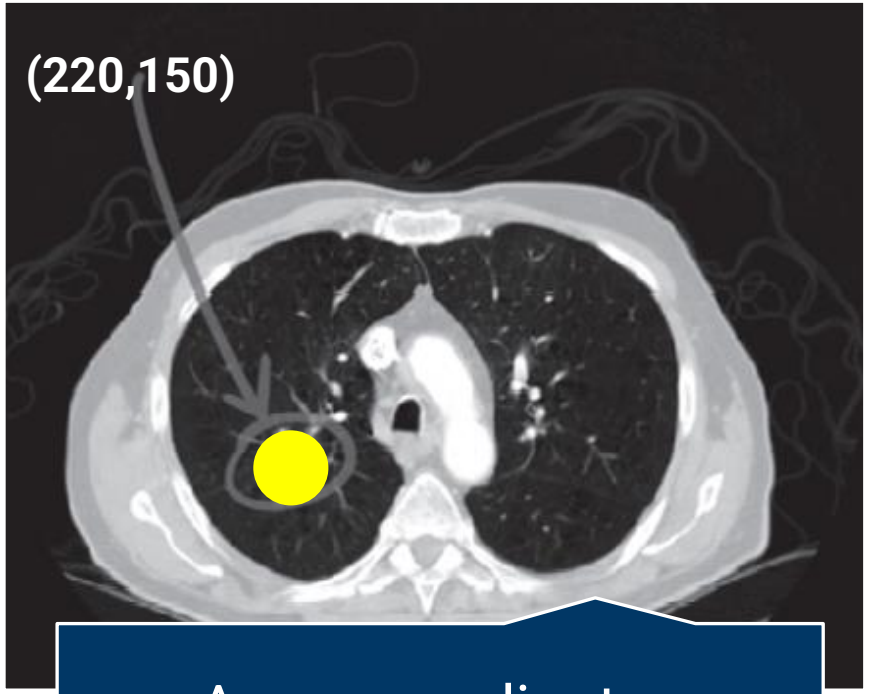
Our coordinates in the csv files are in millimiter-based coordinate system (X,Y,Z)



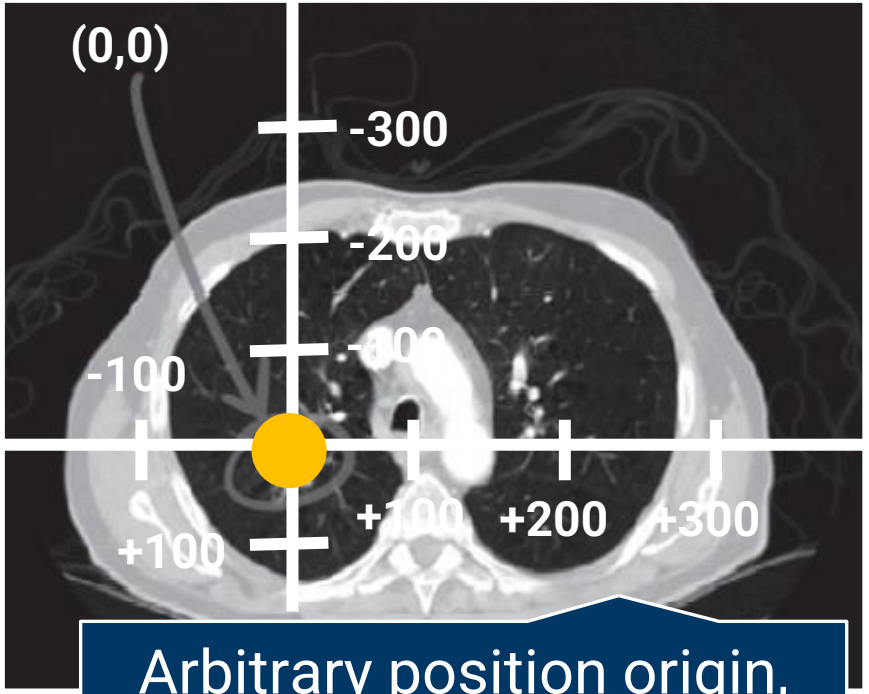
Patient coordinate system



Patient coordinate system



Array coordinates rapresentation



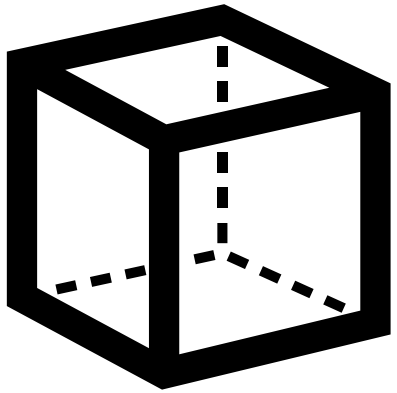
Arbitrary position origin, often in the location of interest

LUNA data to PyTorch

99.9% of voxels in a CT scan won't be part of the actual nodule

To simplify an area around each candidate was extracted

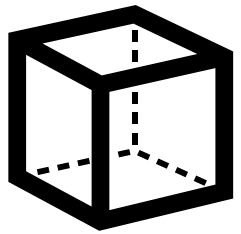
CT Array



Candidate location

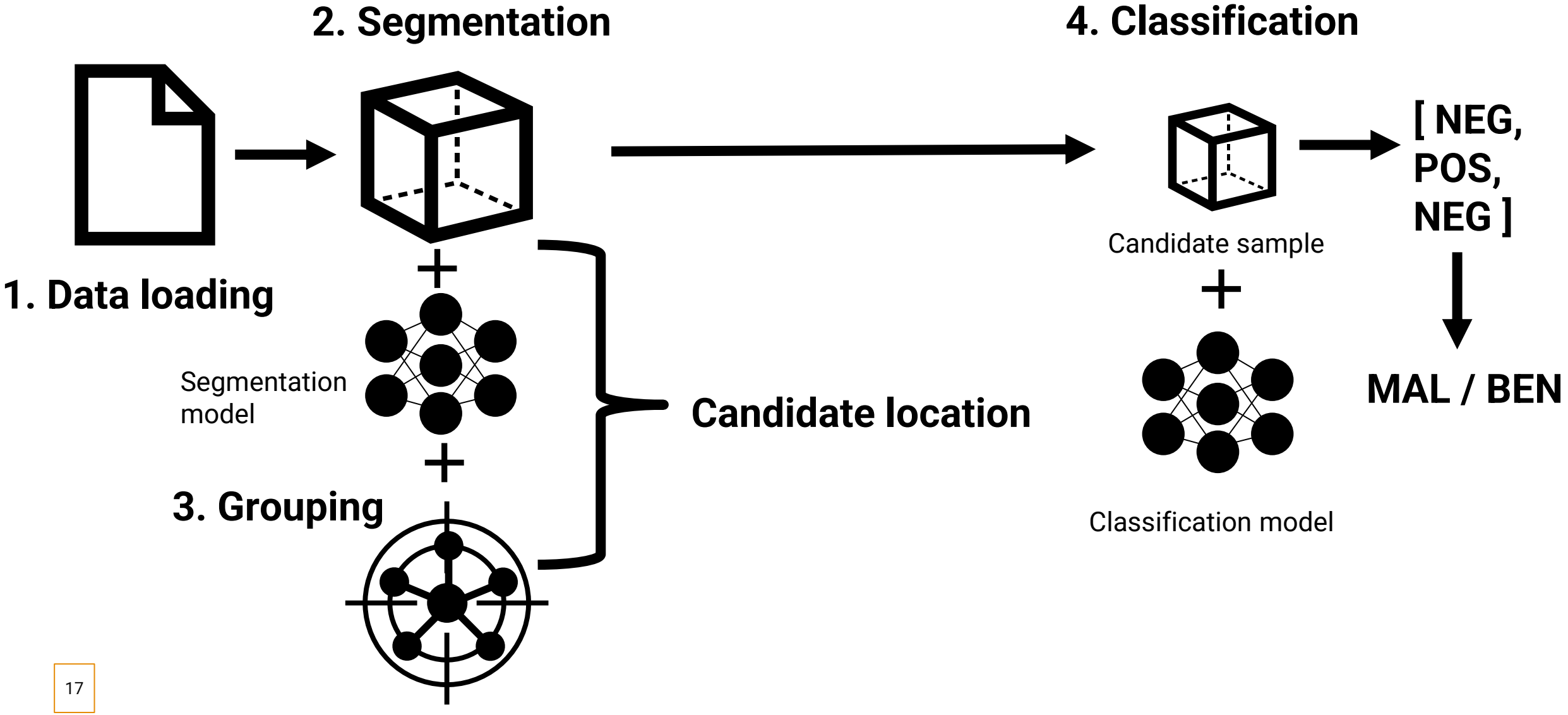
[(I,R,C),
(I,R,C),
(I,R,C)]

Sample tuple



SEGMENTATION

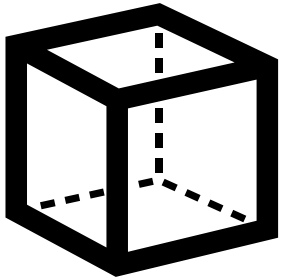
End-to-end detector for lung cancer



Segmentation model

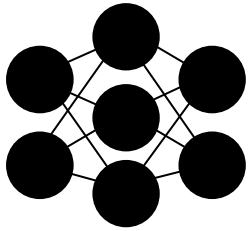
We need to tell our classifier where to look

We have to flag voxels that look like a nodule



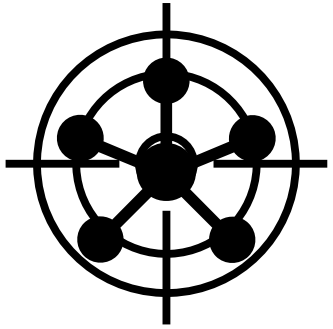
Segmentation

+



Segmentation model

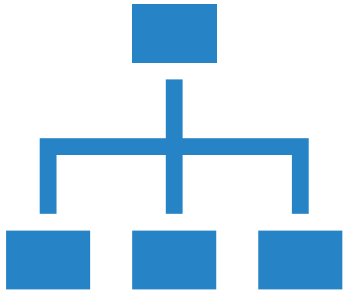
+



Grouping

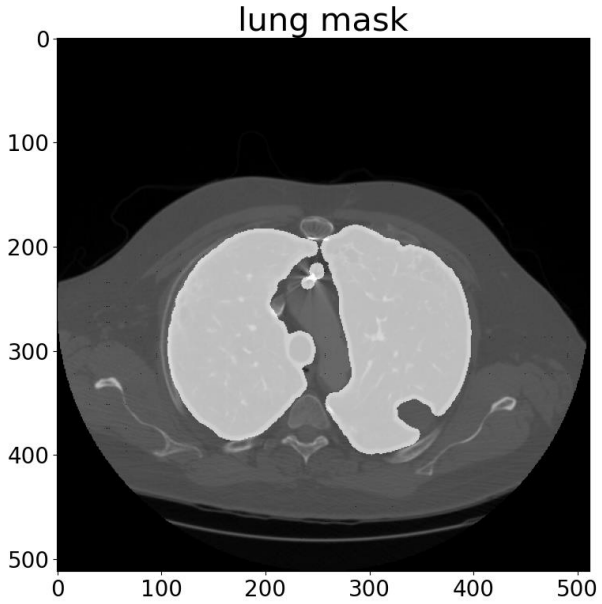
Proposed solution

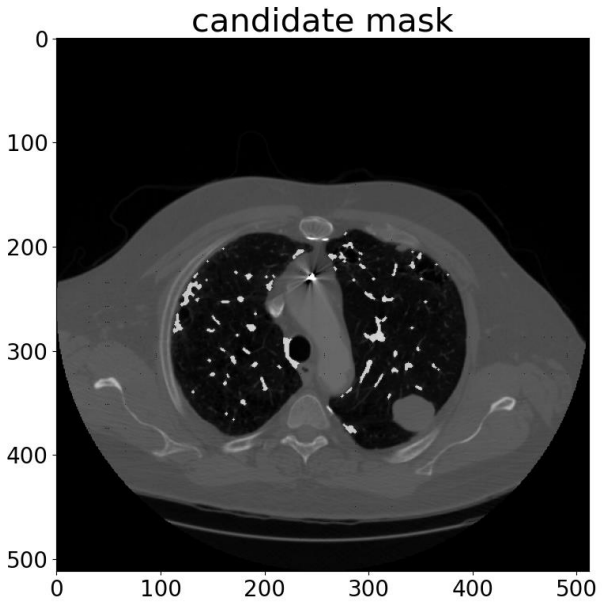
Updating the dataset adding a mask for segmentation

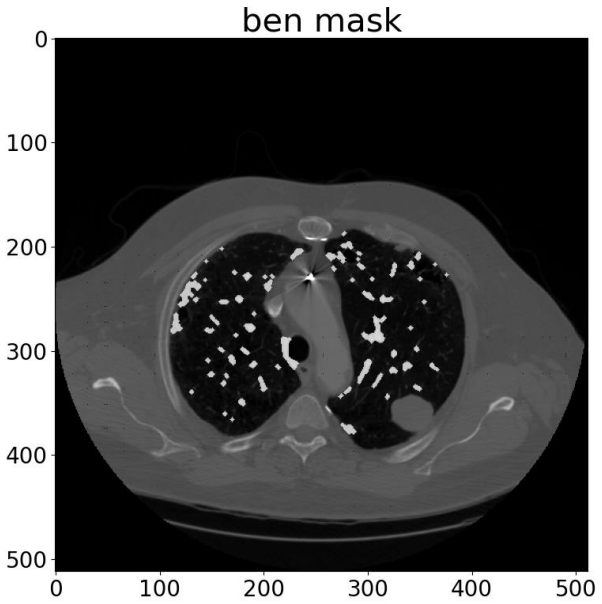


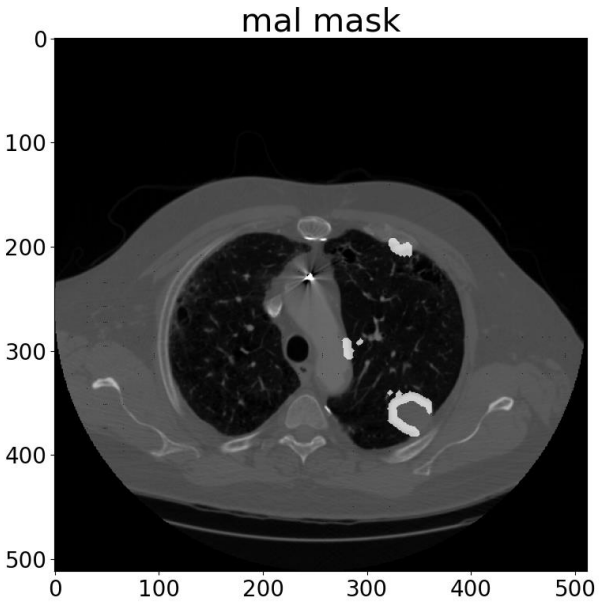
U-Net architecture for segmentation

Building mask example

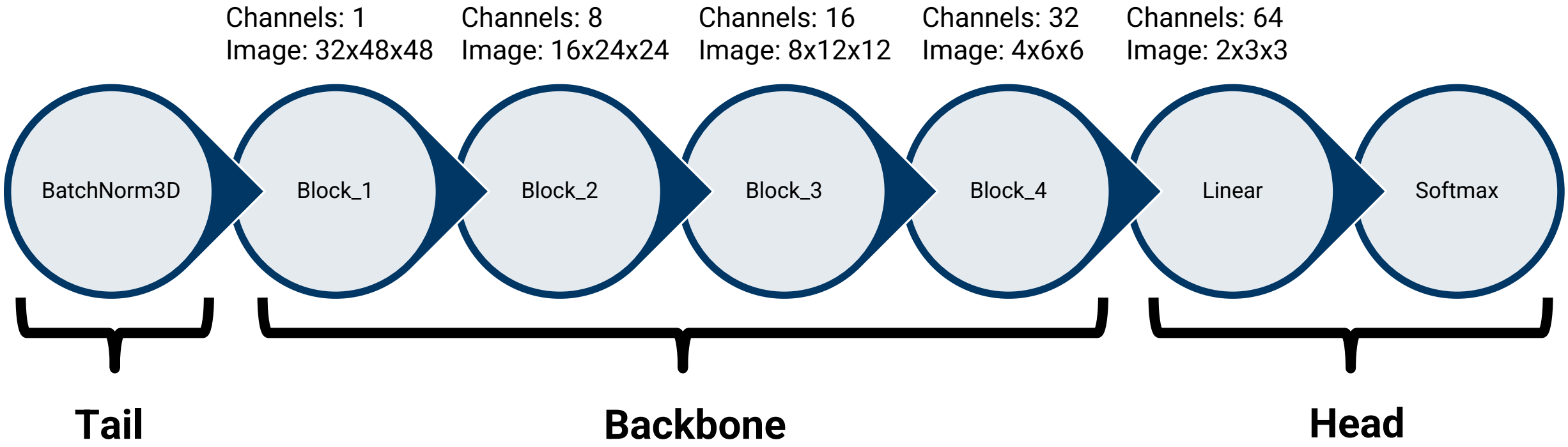




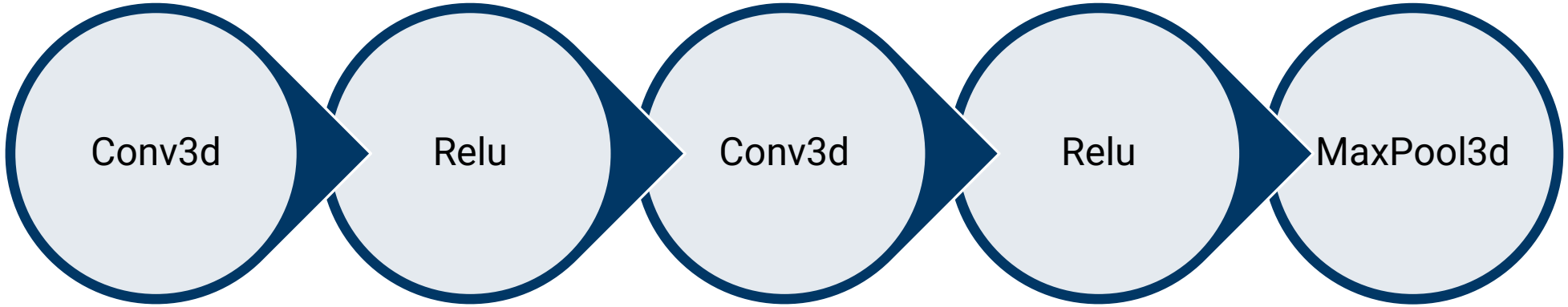


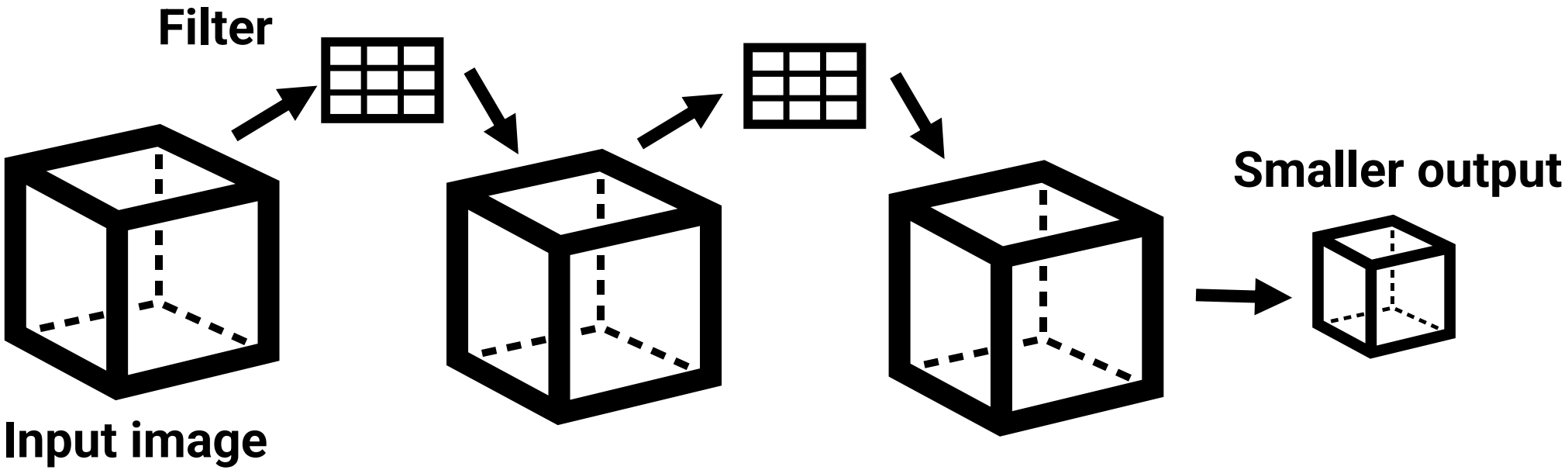


Classification model



Block


Conv3d
Relu
Conv3d
Relu
MaxPool3d

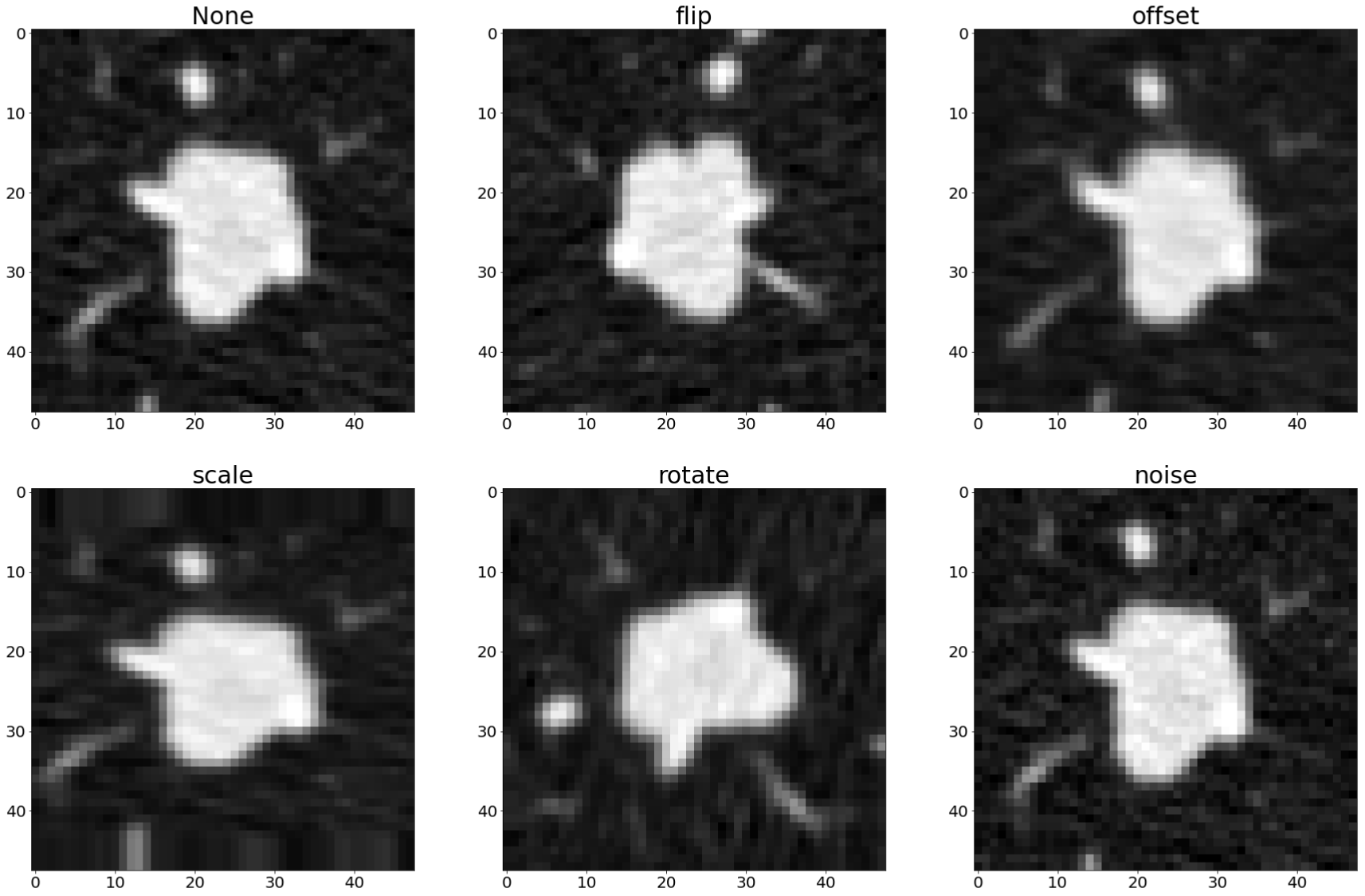

Filter
Input image
Smaller output

Improving training with augmentation

DATA SCIENCE & SCIENTIFIC COMPUTING

Flip	Offset	Scale	Rotate	Noise
• Randomly flipping the data left-right, up-down and front-back	• Randomly offsetting the data slightly along the X and Y axes	• Randomly increasing or decreasing the size of the candidate	• Randomly rotating the data around the head-foot axis	• Randomly adding noise to the data

Augmentation result

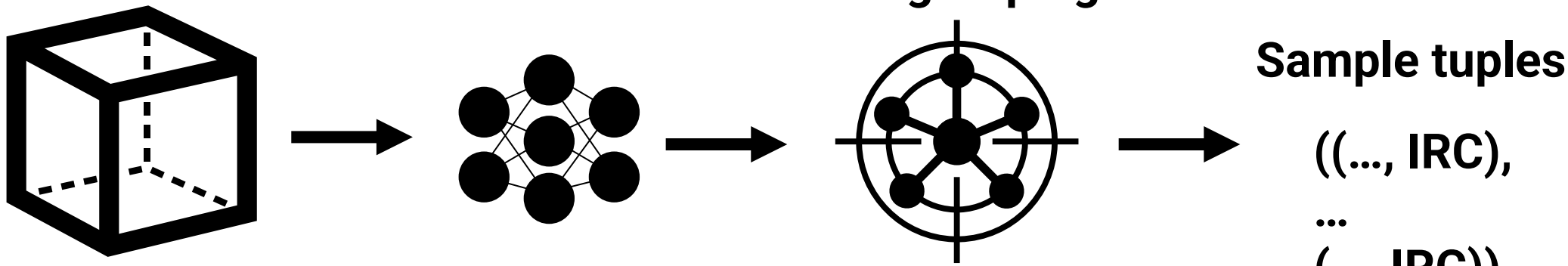


Nodule candidate generation

The segmentation model will predict if a given pixel is of intereset.

Group voxels into nodule candidates by applying a threshold

Each identified nodule candidate will be used to construct a sample tuple for classification



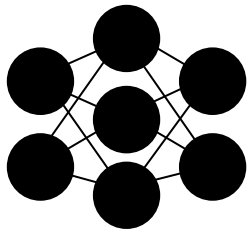
Nodule and malignancy classification

Each nodule candidate from grouping will be classified as nodule or not

Define metrics for examining performance

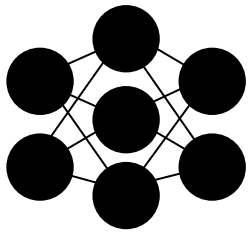
Model specifically for classifying benign and malignant nodules

Nodule classification



Metrics

Malignancy model

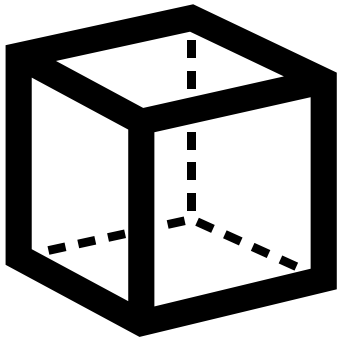


End-to-end detection

Segment our CT to get nodule candidates samples to classify

Perform nodule classification

Perform malignancy classification model

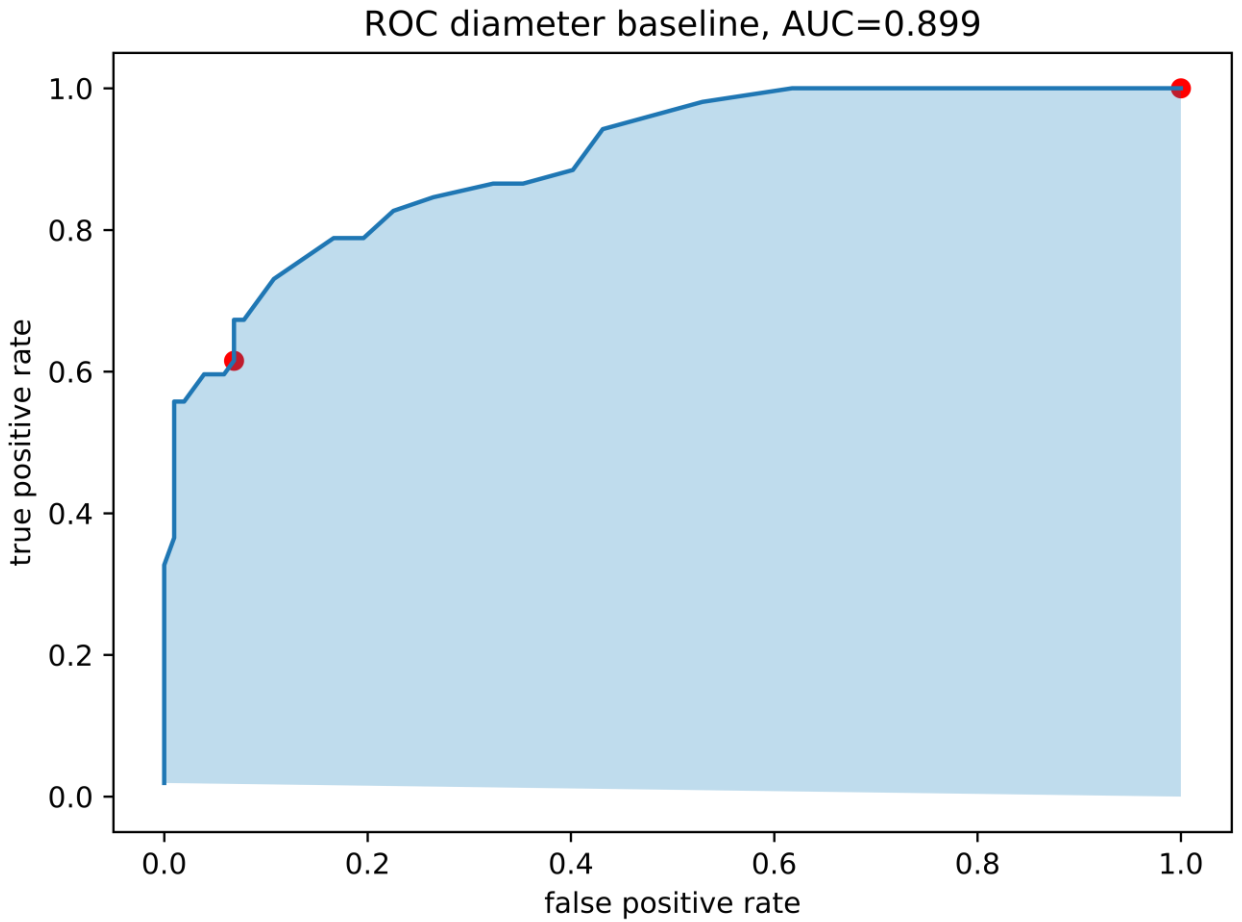


(..., IRC) → **Is nodule?** → **Is malignant?**

RESULTS AND CONCLUSION

Classifying by diameter

- To detect malignancy we use the diameter size
- Picking the right threshold is key
- We can use ROC curve for this purpose
- A good threshold is between 0.6mm and 10.55mm



Quantitative validation

- Running the following command we optain:

```
Python3 -m nodule_analysis --run-validation
```

Total

	Complete Miss	Filtered Out	Pred. Benign	Pred. Malignant
Non-Nodules		177539	1461	512
Benign	12	4	71	15
Malignant	5	5	10	32

Complete miss

- Segmentation didn't find a nodule

Filtered out

- Classifier's work

Predicted nodules

- Those it marked as nodules

Result

DATA SCIENCE & SCIENTIFIC COMPUTING

01

128 detected of the 154 nodules, or 83%

02

Of the 26 missed, 17 were not considered candidates

03

93,52% of the detected nodules are false positive

04

Correctly flag about 70% of malignant ones

Conclusion

This results are not in a level for medical implementation

However this results can be useful at least to indicate potential scan to look at

A possible improvement could be using a better segmentation technique or preventing overfitting for example using dropout

THANK YOU FOR YOUR ATTENTION

DATA SCIENCE & SCIENTIFIC COMPUTING