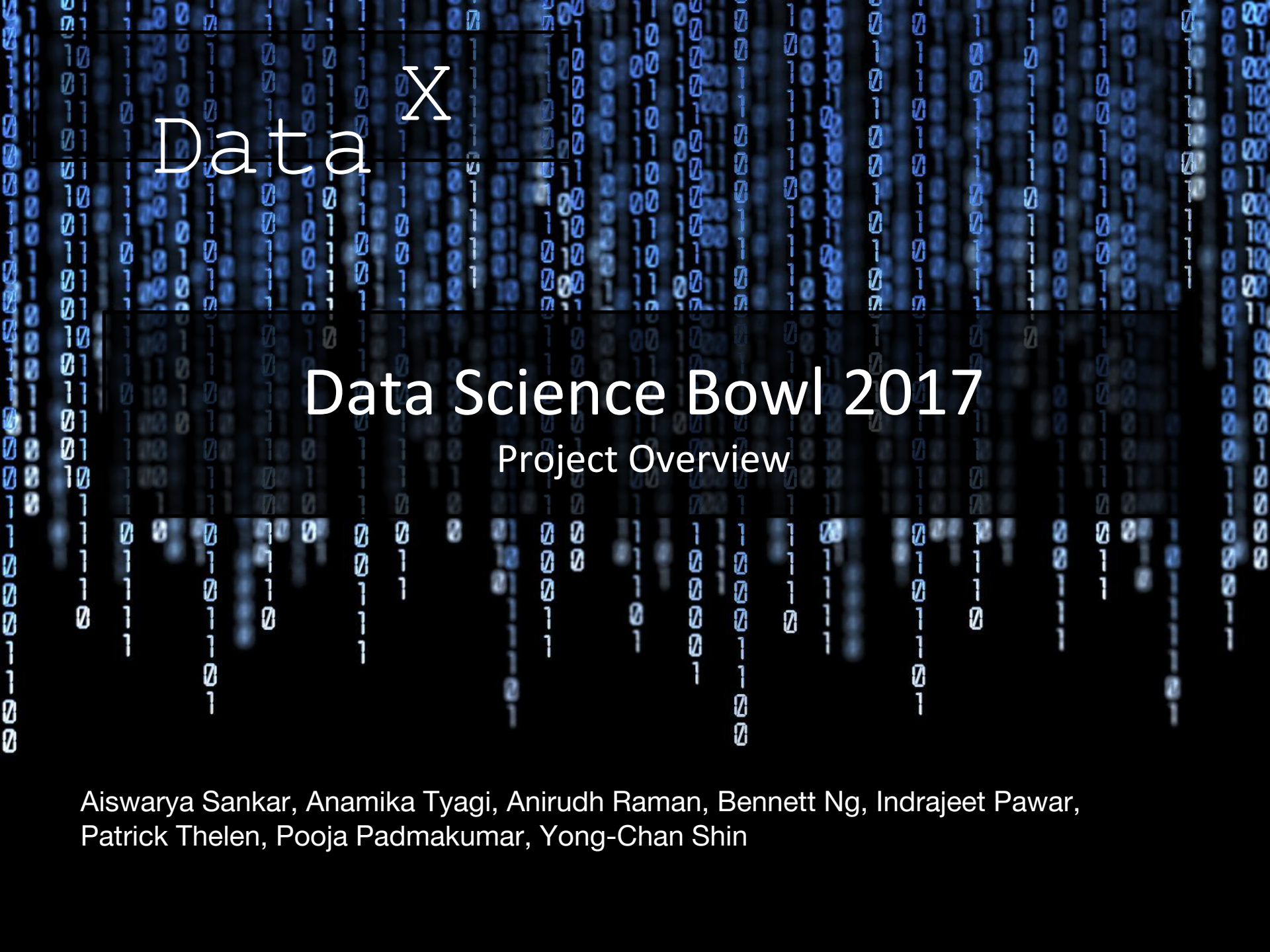


Data Science Bowl 2017

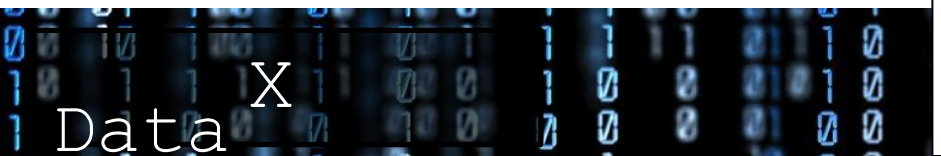
Project Overview

Aiswarya Sankar, Anamika Tyagi, Anirudh Raman, Bennett Ng, Indrajeet Pawar, Patrick Thelen, Pooja Padmakumar, Yong-Chan Shin

Technical Requirements

1. Modular design with separate components for loading, preprocessing, and visualizing data, as well as isolated classifiers.
2. Preprocessing - determine best resolution to use as input to classifiers (balance information content vs. runtime).
3. Candidate ML algorithms: Linear Discriminant Analysis, Quadratic Discriminant Analysis (Gaussian Discriminant Analysis), Convolutional Neural Nets, Residual Neural Nets.
4. Investigate resources for large-data processing such as Apache Spark.

5. Packaged app for end users should have minimal dependencies - ideally a compiled binary.
6. App should be deployable on Picture Archiving and Communication System (PACS) server for use in clinical setting.

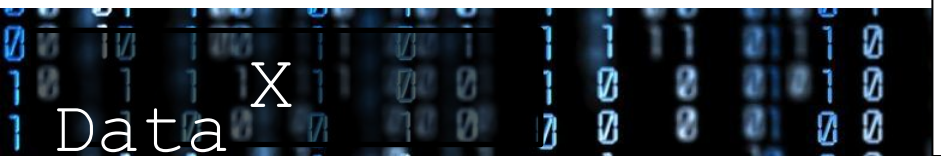


Priority
Blue = Certain, need to execute / code
Orange = Uncertainties, need to validate / learn

User Requirements

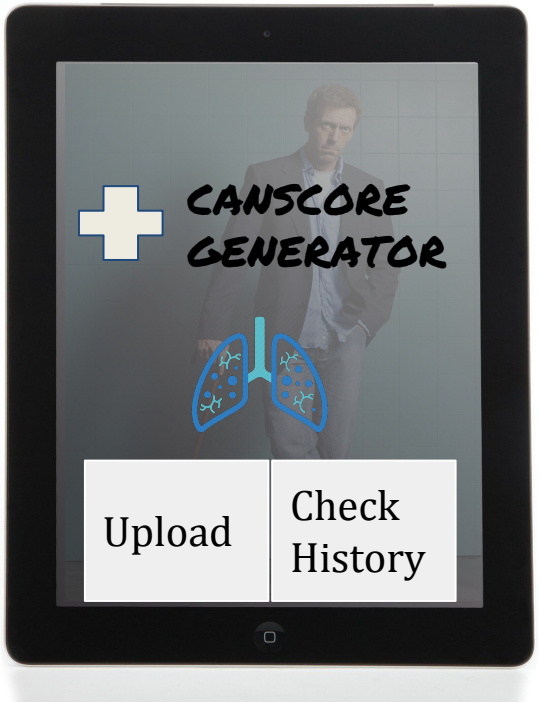
1. Organized, well structured, and documented code (Python format)
2. Concise and informative version control history (Git/GitHub)

3. Lightweight script/app with user-friendly documentation to use the app on new lung CT scans.
4. App should provide a cancer risk estimate (percentage).
5. App should show the named, versioned classifier used to make predictions.
6. App should show the image and highlight the probable cancerous region (stretch goal)



Priority
Blue = Certain, need to execute / code
Orange = Uncertainties, need to validate / learn

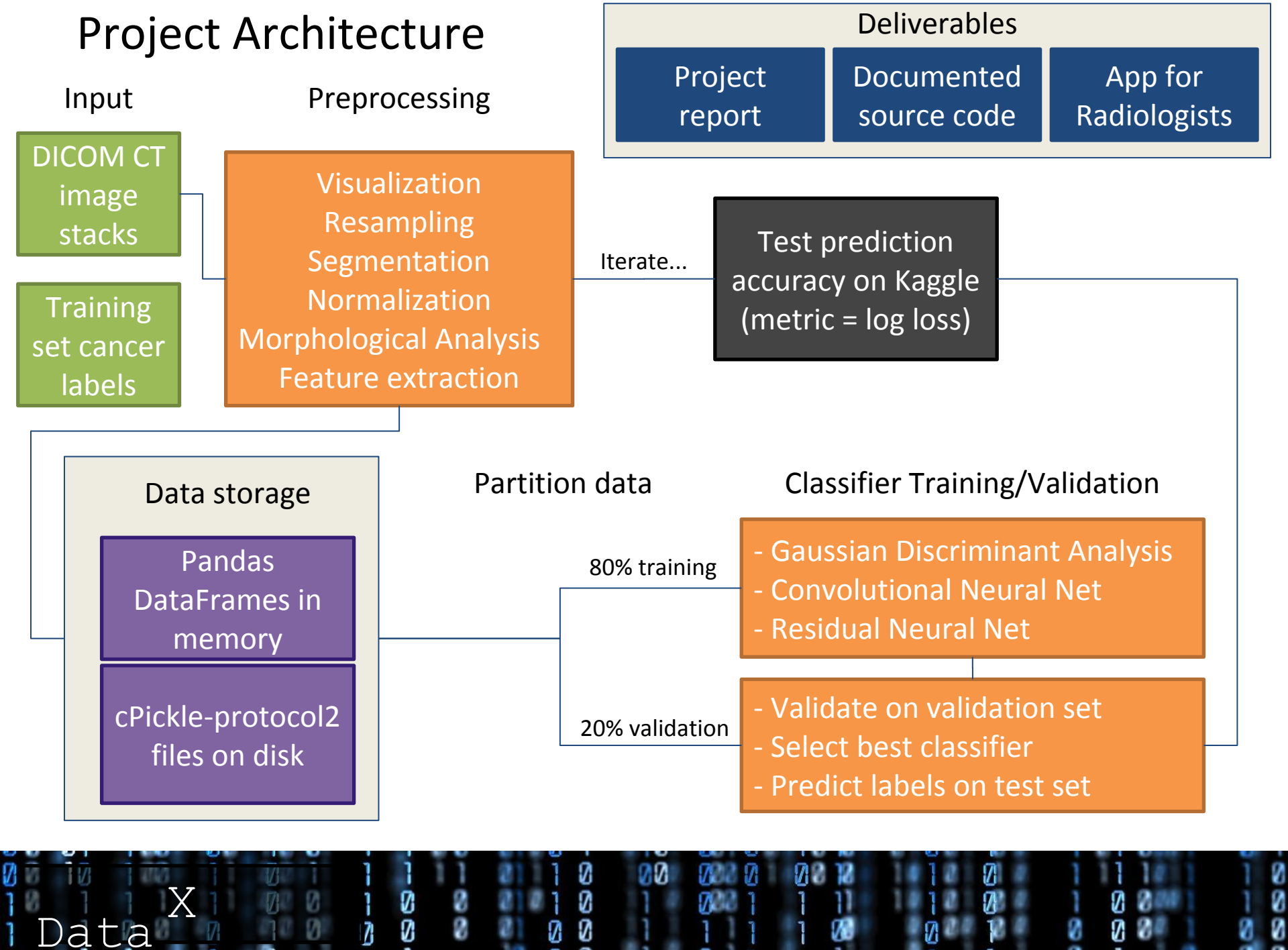
Place an illustration of your Final User Interface below



Three Options indicate:

- Choose the advanced analytics option to explore further- outcome of patients with similar percentages, location of tumor, etc.
- Choose the Diagnostics option for choosing treatment path- treatment history of patients with similar Canscore
- Choose the settings option to toggle through various algorithms for for CanScore generation

Project Architecture



Week 7 Agile Sprint Cycle (3/3-3/9)

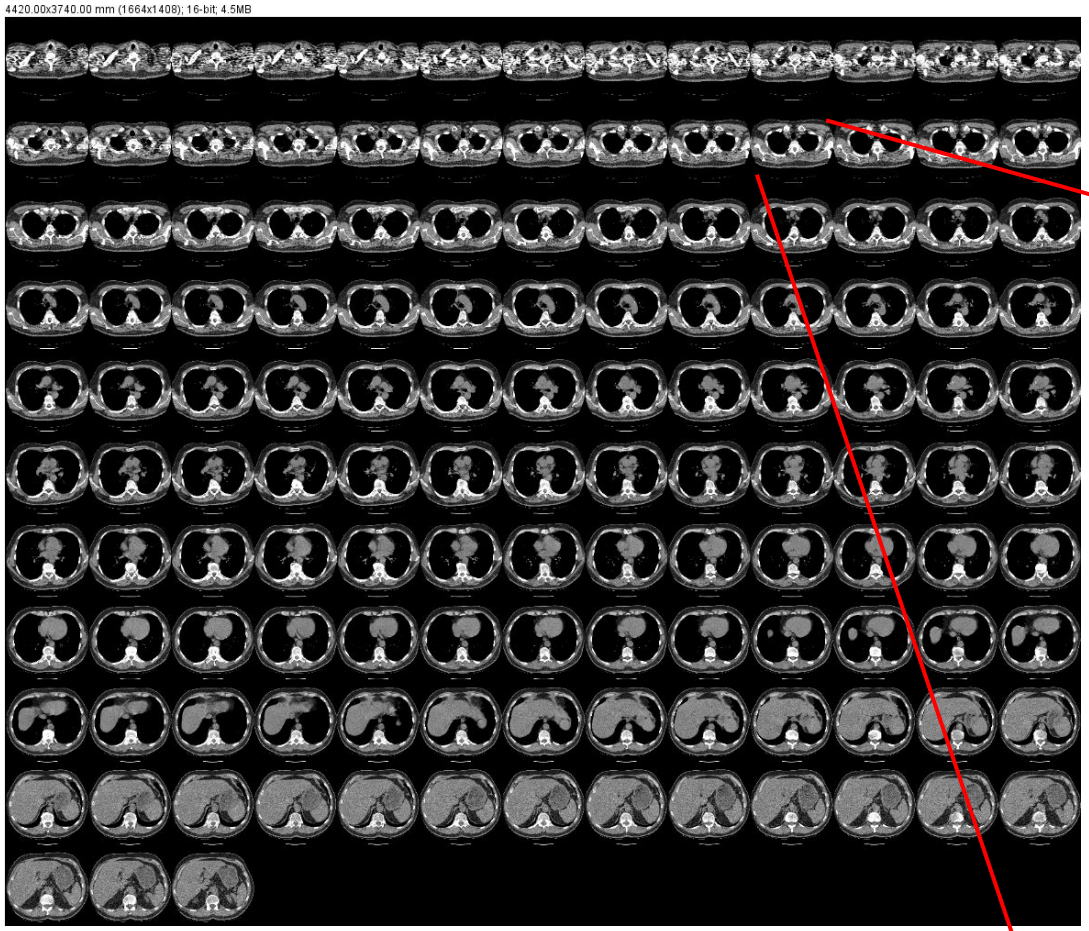
Task	Category	Assignee	Completed?
Implement Gaussian Discriminant Classifier	To code	Yong-Chan (Liner) Shin	
Segment lungs from all images. Resample, normalize, downsample. Post to group bDrive.	To code	Bennett Ng	
Download complete data and check AWS server credits availability for better computation strength	Infra	Indrajeet Pawar	
Form a team on Kaggle and competition enrolment	Infra	Aiswarya Sankar	

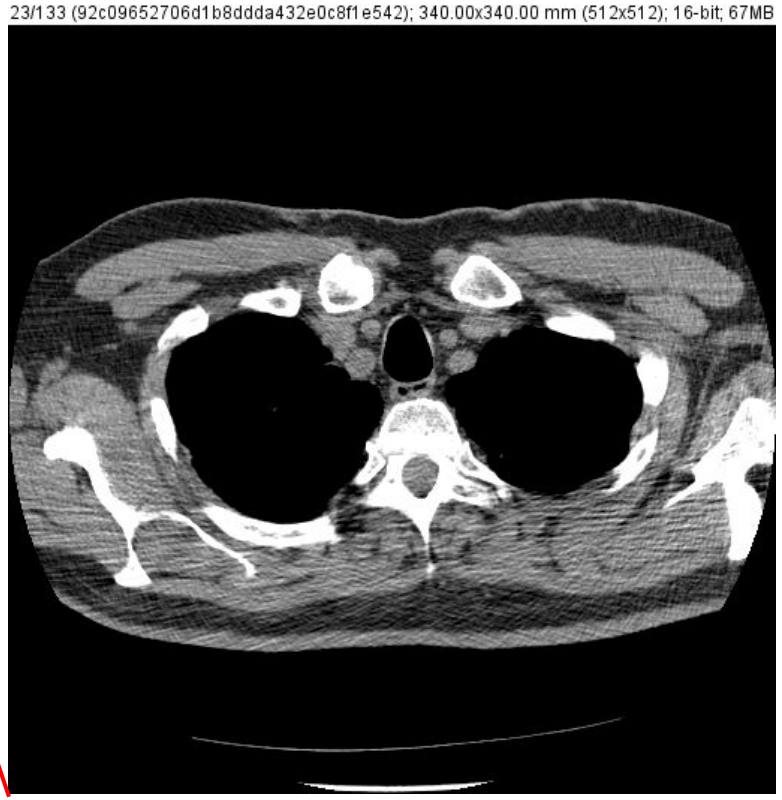
Switch to code / MVP demonstration

https://github.com/bkng/dsb/blob/develop/project_YCLS.ipynb

Data Exploration

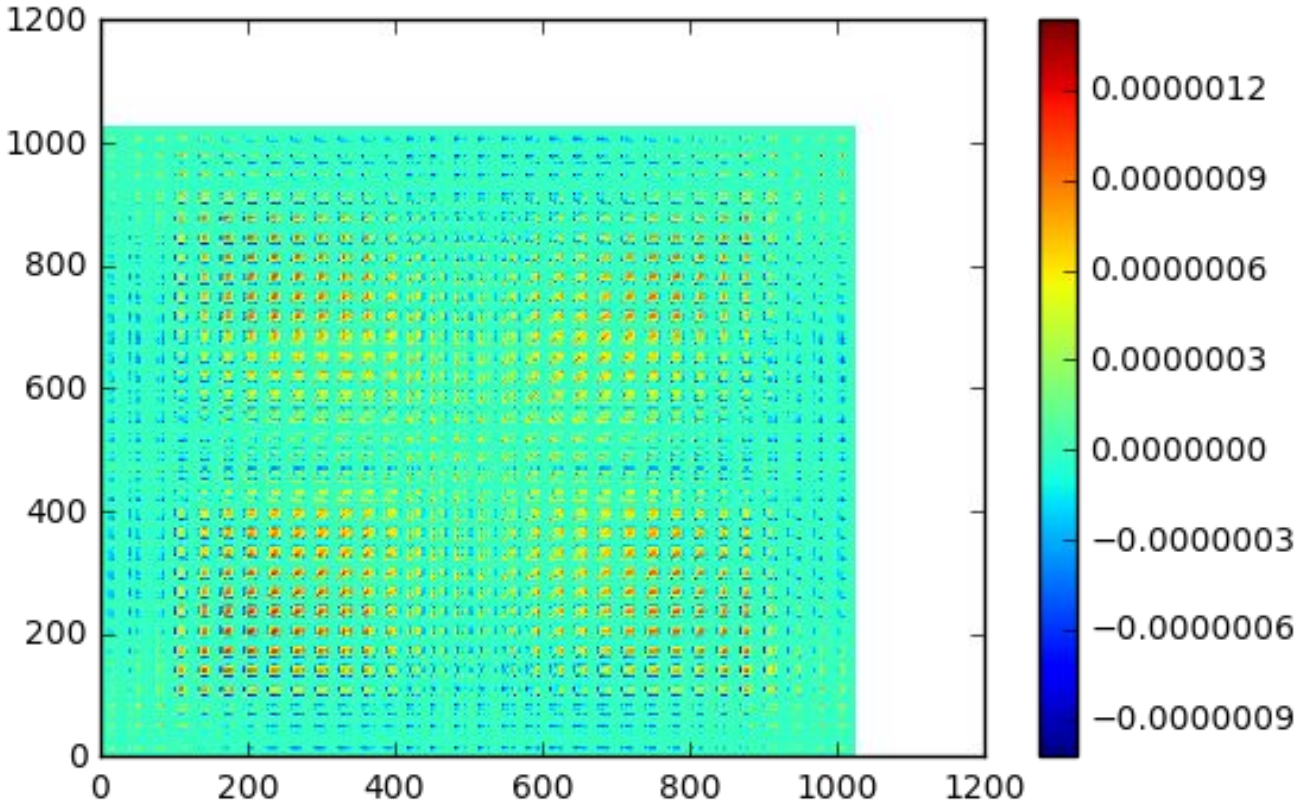
1. DSB dataset includes 1595 abdominal CT scan sets, totaling 143 GB
2. Each set has the following specifications:
   - Average number of slices: 179
   - 512 x 512 transverse resolution
   - 16-bit grayscale depth
3. Ground truth labels (cancer/no cancer) are provided for 1397 image stacks (87.5%)
4. Header information is mostly technical acquisition details.
   - Slice spacing varies across stacks
   - No details on lesion location/qualities are included.

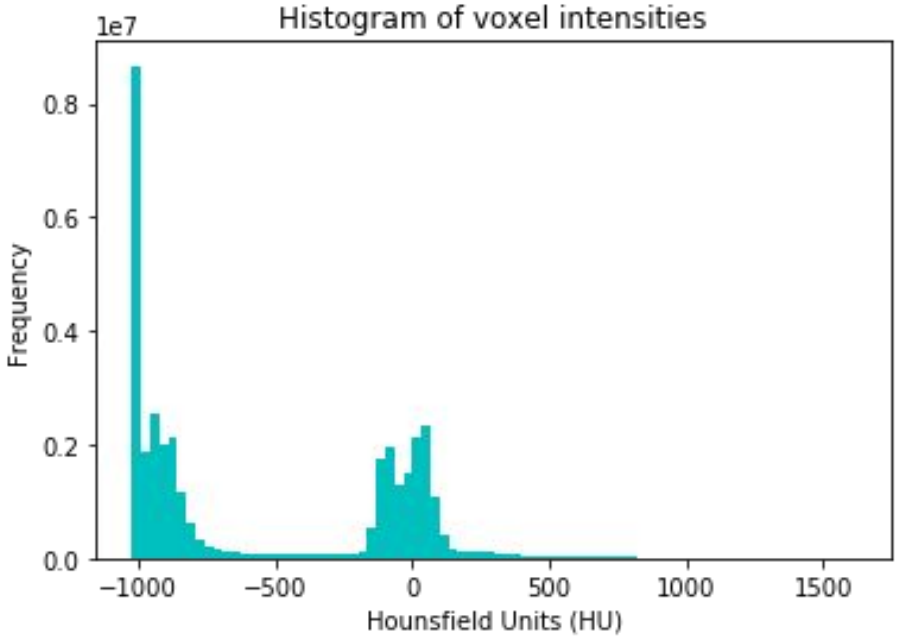

4420.00x3740.00 mm (1664x1408); 16-bit; 4.5MB

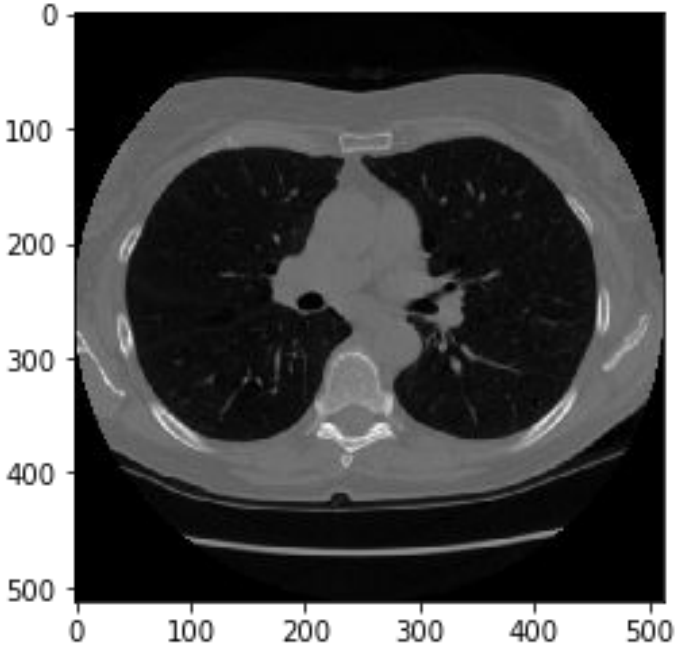

23/133 (92c09652706d1b8ddda432e0c8f1e542); 340.00x340.00 mm (512x512); 16-bit; 67MB


X
Data

Pixel covariance across slices in a stack



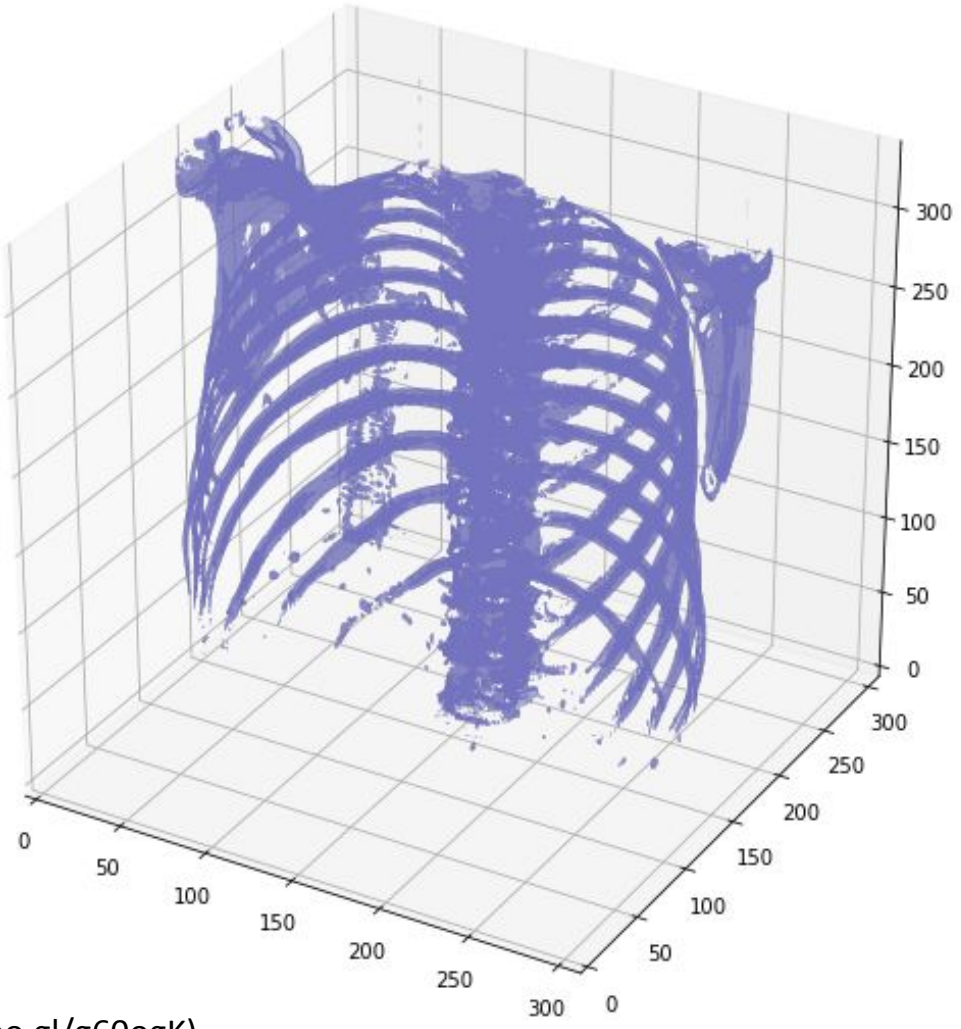




Credit to Guido Zuidhof for his useful tutorial (https://goo.gl/g60ogK)

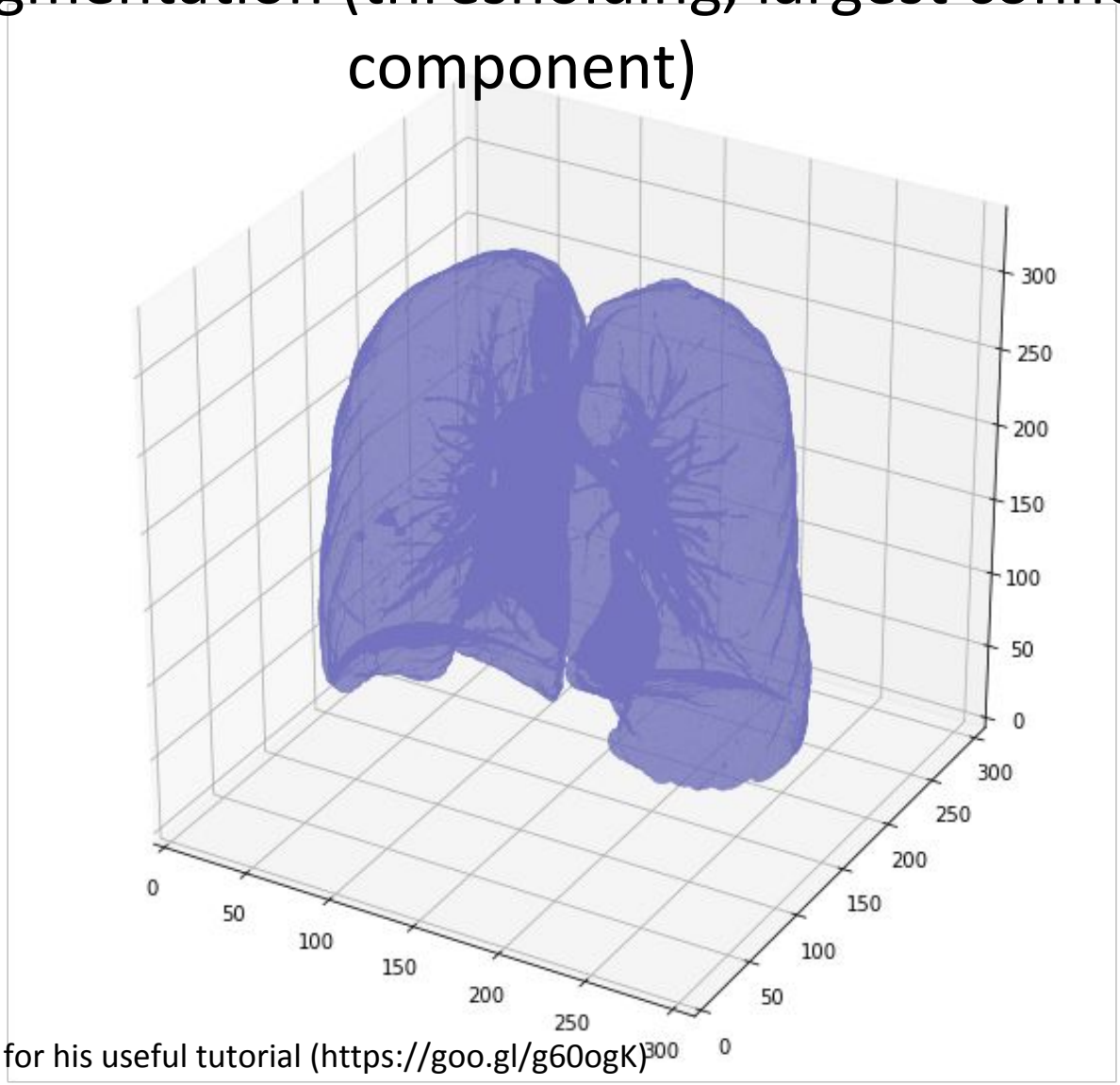
3D plot of bone (>400 HU threshold)

Full resolution, note image takes
> 1 minute to render on 3.60 GHz
Sandy Bridge Xeon-E5 CPU



Credit to Guido Zuidhof for his useful tutorial (https://goo.gl/g60ogK)

Lung segmentation (thresholding, largest connected component)



Credit to Guido Zuidhof for his useful tutorial (https://goo.gl/g60ogK)

Data preprocessing techniques of interest

- Morphological dilation/shrinkage
- Normalization/whitening
  - Subtract global voxel means across all images
- Downsampling (critical for runtime…)
- 3D filtering