Introduction and motivation
- We developed classifiers to find any potential lung diseases from chest X-ray.
- Built independent binary classifiers for each of the lung diseases.
- Trained Logistic Regression and SVM models on top of SIFT and HoG features extracted from X-ray image.

Dataset
- Published by National Institutes of Health (NIH) Clinical Center.
- 100,000+ frontal-view X-ray images.
- 32,717 unique patients, 14 lung diseases.
- Each image has multi-label.
- Images are gray scale of size 1024 x 1024.

Features
- Scaled 1024 x 1024 image to 224 x 224.
- Cropped image to make lungs a focal point, resulting image size 180 x 200.
- Applied histogram equalizer to increase contrast of the image.
- Derived SIFT and HoG features. Both techniques help in finding robust key-points in the image.

Models
- Logistic Regression:
  - Cost function: $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left( \frac{1}{1 + e^{-\theta^T z}} - y_i \right)^2$
- SVM with Radial basis function:
  - Cost function: $J_\epsilon(\alpha) = \frac{1}{m} \sum_{i=1}^{m} L_\epsilon(K_{OPT}(y_i, y_i)) + \frac{1}{2} \alpha^T K \alpha$
  - $L_\epsilon(z, y) = \max \{0, 1 - y \}$
  - $K(x, z) = \exp \left( -\frac{1}{2\sigma^2} ||x - z||^2 \right)$

Discussion
- Logistic regression seems to be generalizing much better than SVM(Kernel) for SIFT. SVM is overfitting training data.
- SIFT was able to detect various descriptors on the lung in the image, that proved useful as a features.

Future
- Pre-trained CNN models for feature extraction.
- Min-hashing to generate feature.
- HoG with PCA.
- Pre-processing to make lungs more focal.

References:
1. http://scikit-learn.org/
3. https://docs.opencv.org/3.1.0/da/df5/tutorial_py_sift_intro.html