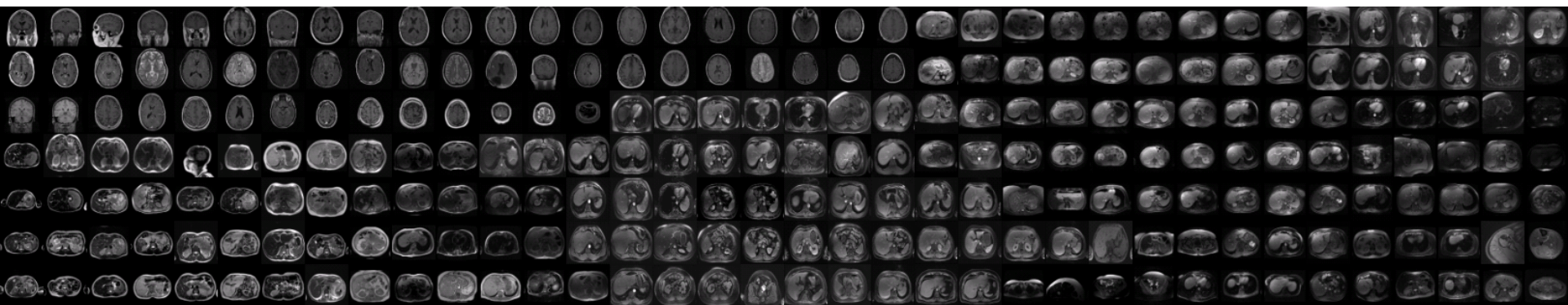


# Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition

Xiaosong Wang, Le Lu, Hoo-chang Shin, Lauren Kim, Mohammadhadi Bagheri, Isabella Nogues, Jianhua Yao and Ronald M. Summers

*Imaging Biomarkers and Computer-Aided Diagnosis Laboratory,  
Department of Radiology and Imaging Sciences,  
National Institutes of Health Clinical Center, Bethesda, MD 20892*



# Motivation


- The availability of well-labeled data is the key for large scale machine learning, e.g. deep learning
- Labels for large medical imaging database are NOT available
- Conventional ways for collecting image labels are NOT applicable, e.g.
  - ❑ Google search followed by crowd-sourcing
  - ❑ Annotation on medical images requires professionals with clinical training

## *Large scale natural image datasets*

 **VISUALGENOME**

 **IMAGENET**

 **PASCAL2**  
Pattern Analysis, Statistical Modelling and  
Computational Learning

 **COCO**  
Common Objects in Context

**places**   
THE SCENE RECOGNITION DATABASE

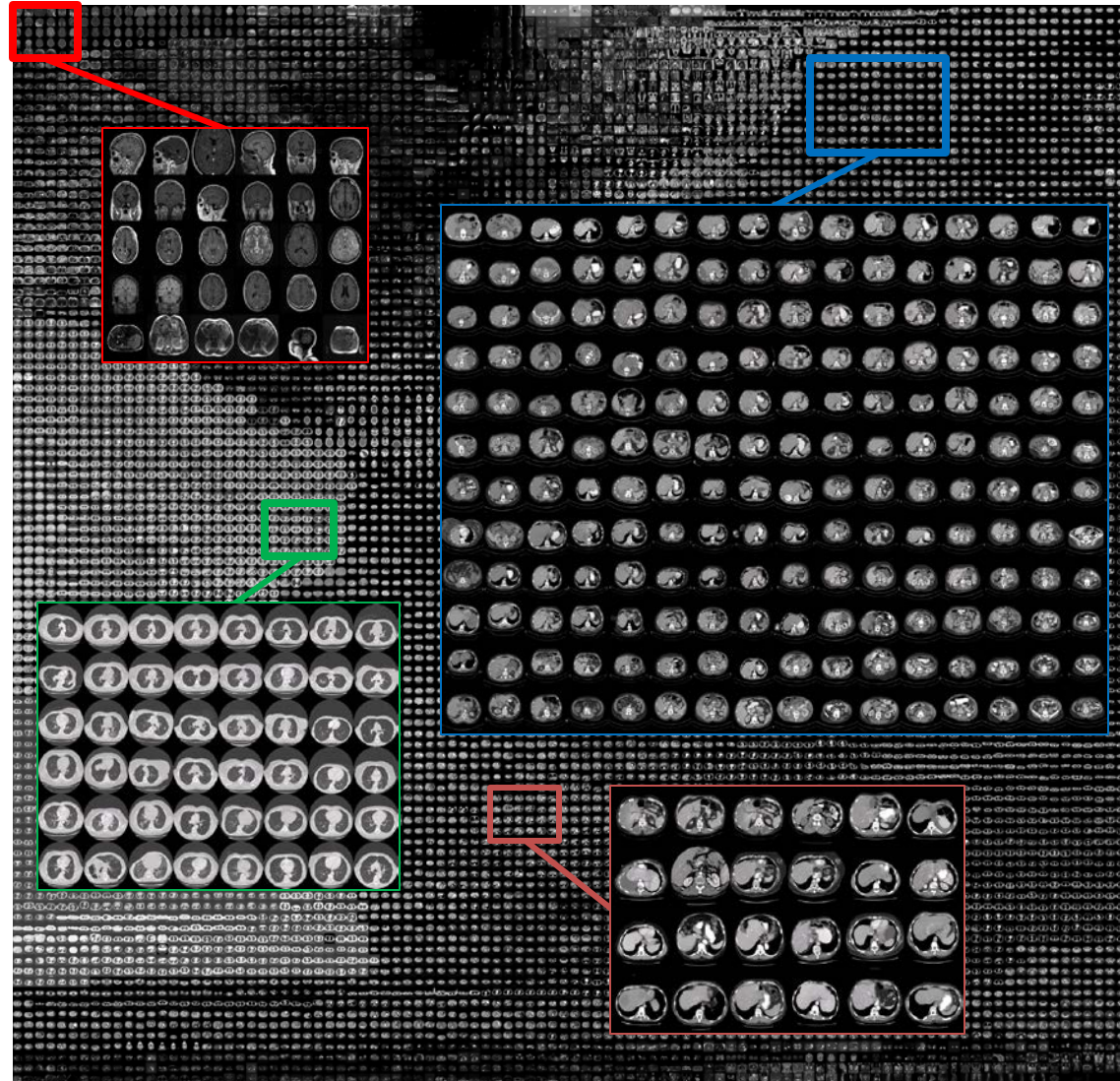
 **SUN**

## *Large scale Medical Image dataset*



\* Dataset logos shown here are from respective public dataset websites.

- A great treasure has been stored in our PACS system, i.e. images together with radiological reports.
- “Keyimage” dataset: 215,786 key images from 61,845 unique patients.
- Key images are significant one or more images in a study referenced in the linked radiological report.
- Key images are directly extracted from the DICOM file and resized as 256\*256 bitmap images (.png).
- Their intensity ranges are rescaled using the default window settings stored in the DICOM header files

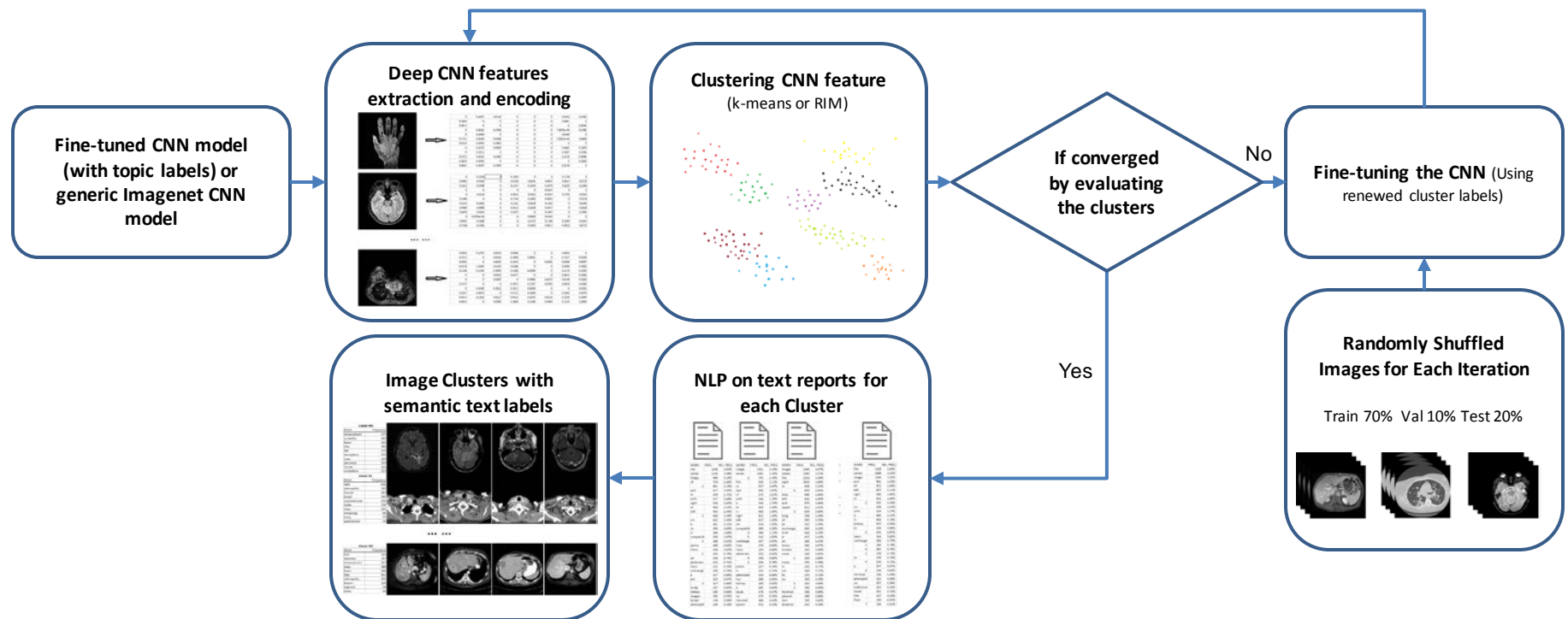


\* 10000 random images from the dataset, using CNN FC7 features of images embedded with t-SNE



# Unsupervised Categorization

The proposed framework is designed towards automatic medical image annotation

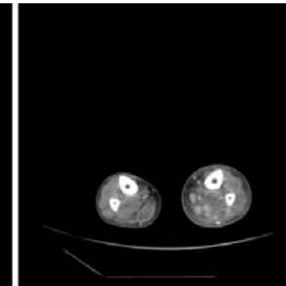
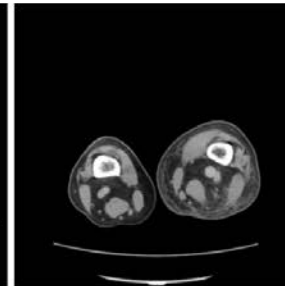
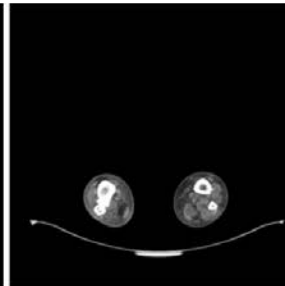
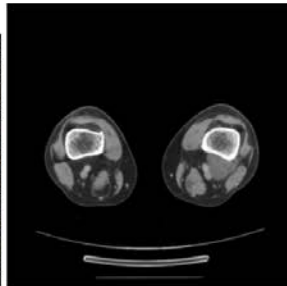


- Hypothesized “convergence”: better labels lead to better trained Convolutional Neural Network (CNN) models which consequently feed more effective deep image features to facilitate more meaningful clustering/labels.

# Sample Categories

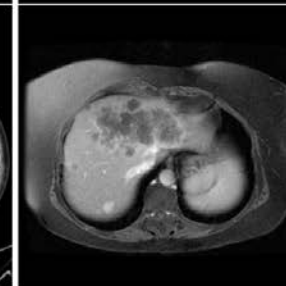
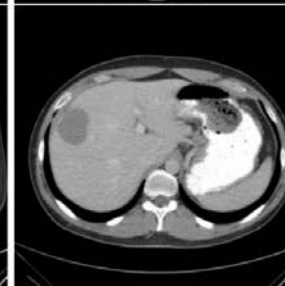
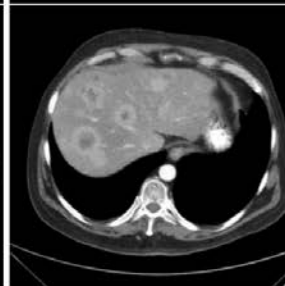
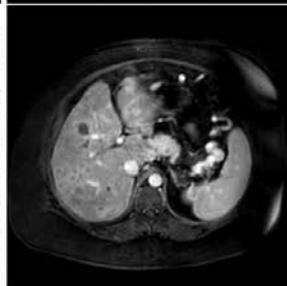
**Cluster #14**

Word	Frequency
calf	369
mass	263
subcutaneous	205
thigh	204
lesion	127
lower	124
enhancing	111
bone	105
fossa	92
nerve	88



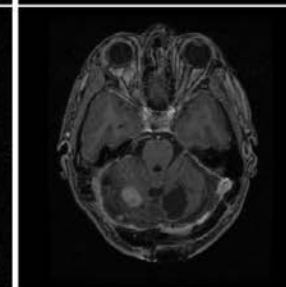
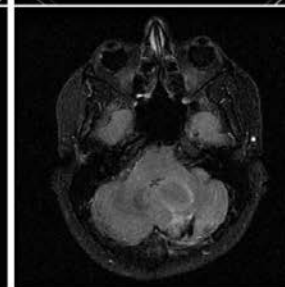
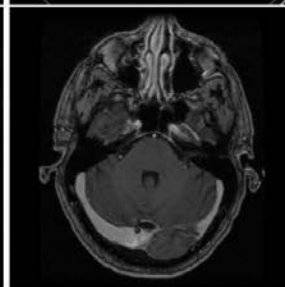
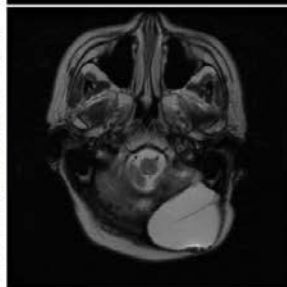
**Cluster #23**

Word	Frequency
liver	524
abdomen	337
enhancement	217
mass	198
lesion	168
lobe	161
adenopathy	119
lesions	109
segment	58
bulky	45



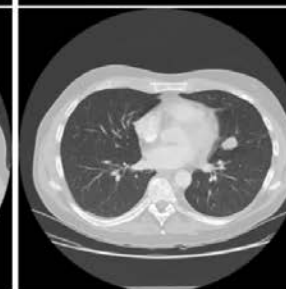
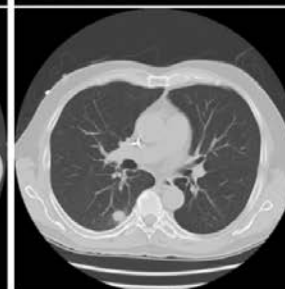
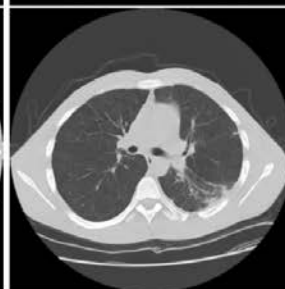
**Cluster #64**

Word	Frequency
enhancement	277
cerebellar	193
lesion	192
lobe	186
flair	173
hemisphere	155
mass	134
abnormal	119
frontal	115
cerebellum	113



**Cluster #224**

Word	Frequency
lung	637
lobe	450
chest	361
mass	215
nodule	160
pleural	158
adenopathy	128
granulomata	111
atelectasis	86
pericardial	81



# Experiment - CNN Setting

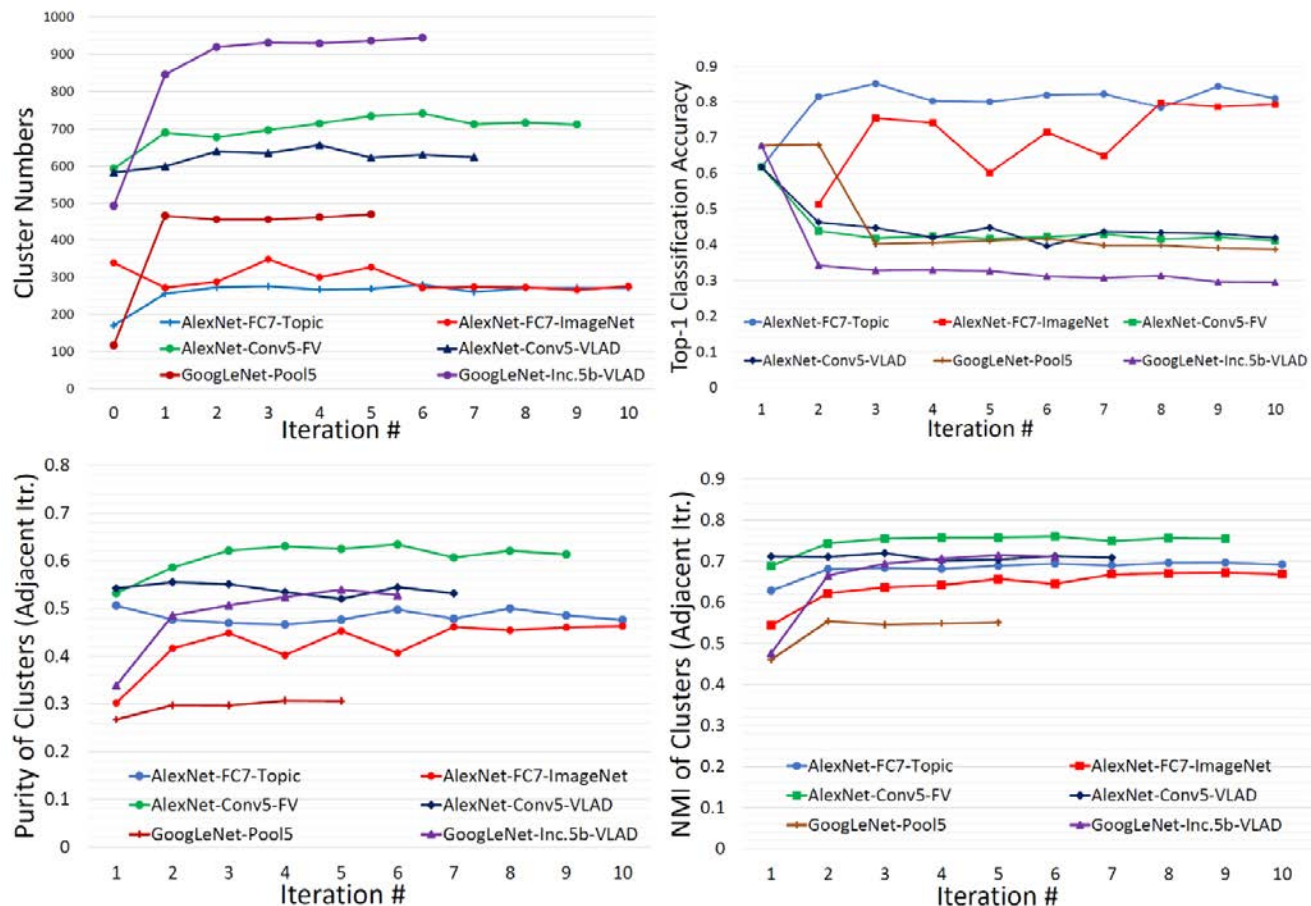
- The proposed framework is applicable to a variety of CNN models, by analyzing the CNN activations from layers of different depths.
- Encode the convolutional layer outputs in a form of dense pooling via Fisher Vector (FV) and Vector Locally Aggregated Descriptor (VLAD)
- Principal Component Analysis (PCA) is performed to reduce the dimensionality to 4096.

CNN model	Layer	Activations	Encoding
AlexNet	Conv5	(13, 13, 256)	FV+PCA
AlexNet	Conv5	(13, 13, 256)	VLAD+PCA
AlexNet	FC7	4096	—
GoogLeNet	Inception5b	(7, 7, 1024)	VLAD+PCA
GoogLeNet	Pool5	1024	—

# Experiment - Convergence

- Clustering via K-means only or over-fragmented K-means followed by Regularized Information Maximization (as an effective model selection method), are extensively explored and empirically evaluated.
- Two convergence measurements have been adopted, i.e., Clustering Purity and Normalized Mutual Information (NMI).
- Newly generated clusters are better in terms of
  - ☐ Visually more coherent and discriminative from instances from other clusters
  - ☐ Balanced classes with approximately equivalent images per cluster
  - ☐ The number of clusters is self-adaptive according to the nature of data

# Quantitative Results

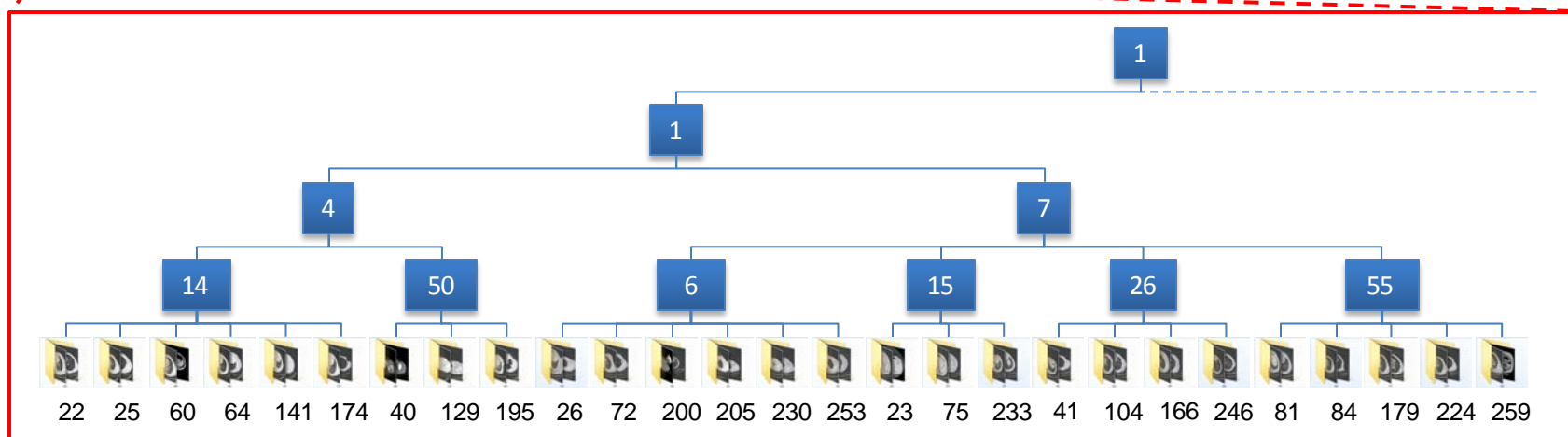
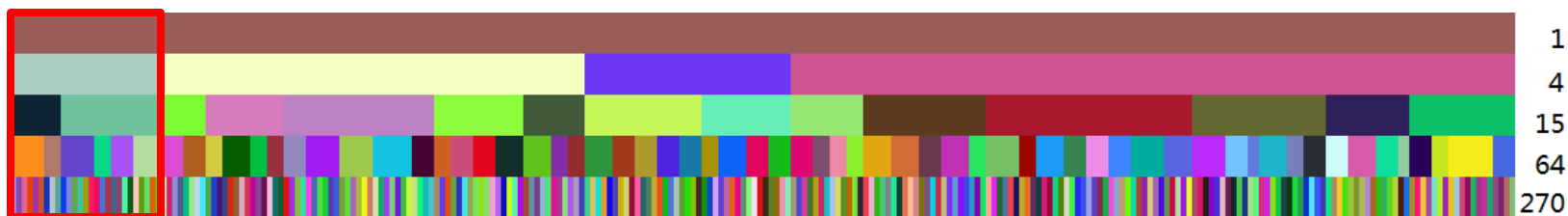


- The convergence of our categorization framework is measured and observed in the cluster-similarity measures, the CNN training classification accuracies and the self-adapted cluster number.
- AlexNet-FC7-Topic is preferred by two radiologists, which results in total 270 categories. The adopted FC7 feature is able to preserve the layout information of images.



# Hierarchical Category Relationship

- Hierarchical category relationships in a tree-like structure can be naturally formulated and computed from the final pairwise CNN classification confusion measures. The resulting category tree has (270, 64, 15, 4, 1) different class labels from bottom (leaf) to top (root). The random color coded category tree is shown below.



# Application on Scene Recognition

- Images from the same scene category may share similar object patches but are different in overall setting, e.g. buildings all have windows but in different style.
- Integrate patch mining as a form of image encoding into our LDPO framework and perform the categorization and patch mining iteratively.

## MIT Indoor-67 (I-67)

indoor scenes | 67 classes  
15620 images



Airport

## Building-25 (B-25)

Architecture Style | 25 classes  
4794 images



American Craftsman

## Scene-15 (S-15)

Indoor & outdoor | 15 classes  
4485 images



Bedroom

# Evaluation on Clustering Accuracy

- The purity and NMI measurements are computed between the final LDPO clusters and GT scene classes ( purity becomes the classification accuracy against GT).
- We compare the LDPO scene recognition performance to those of several popular clustering methods.
- The state-of-the-art fully-supervised scene Classification Accuracies(CA) for each dataset are also provided.

Dataset	KM [57]	LSC [4]	AC [22]	EP [10]	MDPM [34]	LDPO-A-FC	LDPO-A-PM	LDPO-V-PM	Supervised
	Clustering Accuracy (%)								CA(%)
<b>I-67 [44]</b>	35.6	30.3	34.6	37.2	53.0	37.9	63.2	<b>75.3</b>	<b>81.0[8]</b>
<b>B-25 [62]</b>	42.1	42.6	43.2	43.8	43.1	44.1	59.2	<b>59.5</b>	<b>59.1 [42]</b>
<b>S-15 [32]</b>	65.0	76.5	65.2	73.6	63.4	73.1	<b>90.1</b>	84.0	<b>91.6 [66]</b>
	Normalized Mutual Information								
<b>I-67 [44]</b>	.386	.335	.359	-	.558	.389	.621	<b>.759</b>	-
<b>B-25 [62]</b>	.401	.403	.404	-	.424	.407	<b>.588</b>	.546	-
<b>S-15 [32]</b>	.659	.625	.653	-	.596	.705	<b>.861</b>	.831	-

\* KM: k-means; AC: agglomerative clustering ; LSC: large-scale spectral clustering ; EP: ensemble projection + k-means;  
MDPM: mid-level discriminative patch mining + k-means

# Evaluation on Learned Image Features and Initialization Settings

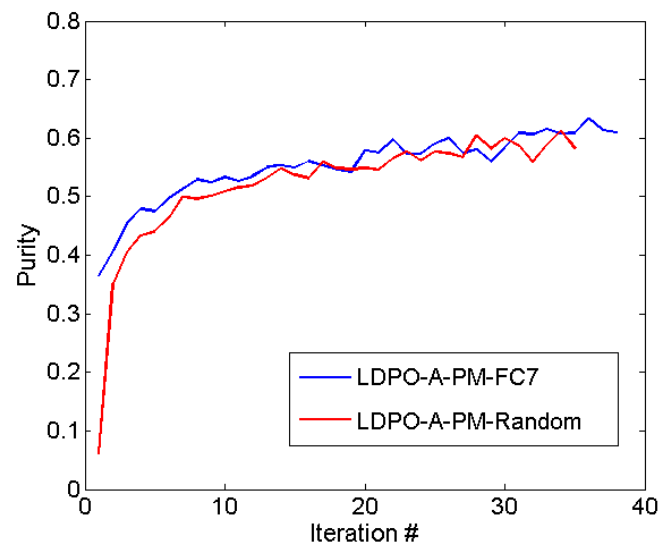
## \* Learned image representation:

1. Classification task on MIT-67, standard partition [44]
  2. One-versus-all Liblinear classification on image features
- × LDPO-V-PM-LL does not improve upon purely unsupervised LDPO-V-PM. This may indicate that LDPO-PM image representation is sufficient to separate images from different classes.

Method	Accuracy (%)
<b>D-patch [53]</b>	38.1
<b>D-parts [54]</b>	51.4
<b>DMS [13]</b>	64.0
<b>MDPM-Alex [34]</b>	64.1
<b>MDPM-VGG [34]</b>	77.0
<b>MetaObject [60]</b>	78.9
<b>FC (VGG)</b>	68.87
<b>CONV-FV (VGG) [8]</b>	<b>81.0</b>
<b>LDPO-V-PM-LL</b>	72.5
<b>LDPO-V-PM</b>	<b>75.3</b>

## \* Different initialization settings:

1. Random initialization
  2. Image labels obtained from k-means clustering on FC7 features of an ImageNet pretrained AlexNet
- ✓ Both schemes ultimately converge to similar performance levels and it suggests that LDPO convergence is insensitive to the chosen initialization.

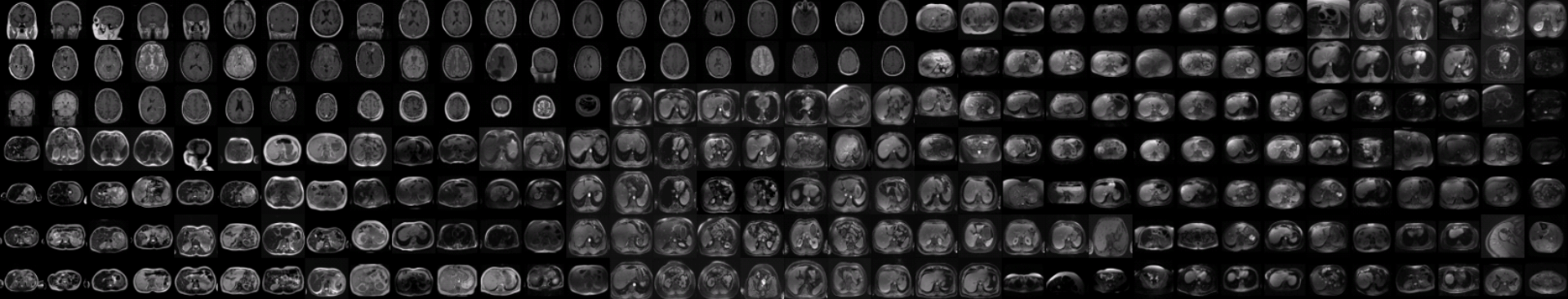


\*Both results are computed using MIT Indoor-67 dataset.



# Conclusion

- To paraphrase Professor Kunio Doi, it is time to wake up the data sleeping in the PACS.
- A novel looped deep pseudo-task optimization framework is presented for category discovery from a large-scale medical image database.
- Extracted categories are visually more coherent and semantically meaningful.
- We systematically and extensively conduct experiments under different settings of the proposed framework to validate and evaluate its quantitative and qualitative performance on two different types of dataset.
- The measurable “convergence” makes the ill-posed auto-annotation problem well constrained, at no human labeling costs.



# THANK YOU!

## Acknowledgement

- This work is supported by the Intramural Research Program of the National Institutes of Health Clinical Center.
- This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>)
- We thank Nvidia corporation for the GPU donation.



Scan to contact

