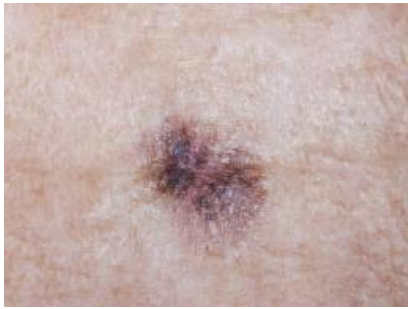
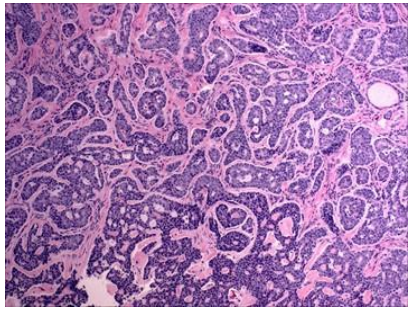
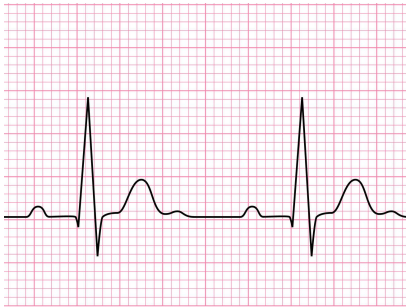
AI For Healthcare

Pranav Rajpurkar
PhD Candidate, Computer Science
Stanford ML Group

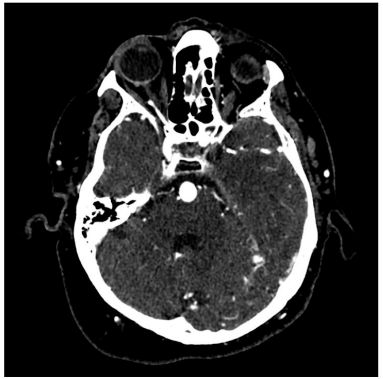





Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)
Lymph node metastases from H&E slides (Bejnordi et al., 2018)
Arrhythmias from ECG signals (Hannun & Rajpurkar et al., 2019)

CheXNeXt

MURA

MRNet



HeadXNet

Rajpurkar & Irvin et al., PLOS Medicine, 2018
Rajpurkar & Irvin et al., MIDL, 2018
Bien & Rajpurkar et al., PLOS Medicine, 2018
Park & Chute & Rajpurkar et al., JAMA Network Open. 2019

What are **challenges** for building and deploying AI for medical image interpretation?

What are **opportunities** for making AI medical image interpretation routine part of clinical care**?**

Challenges for AI in Medical Image Interpretation

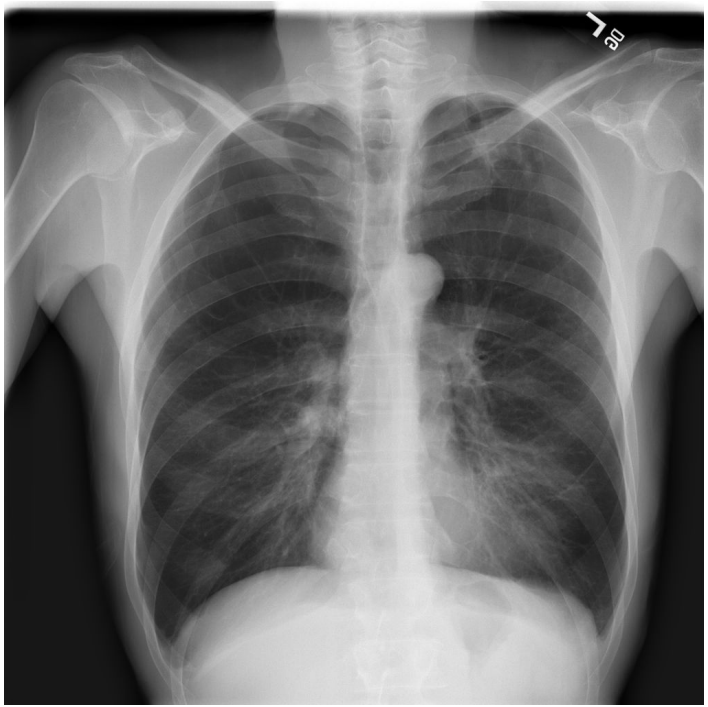
Data

Deployment

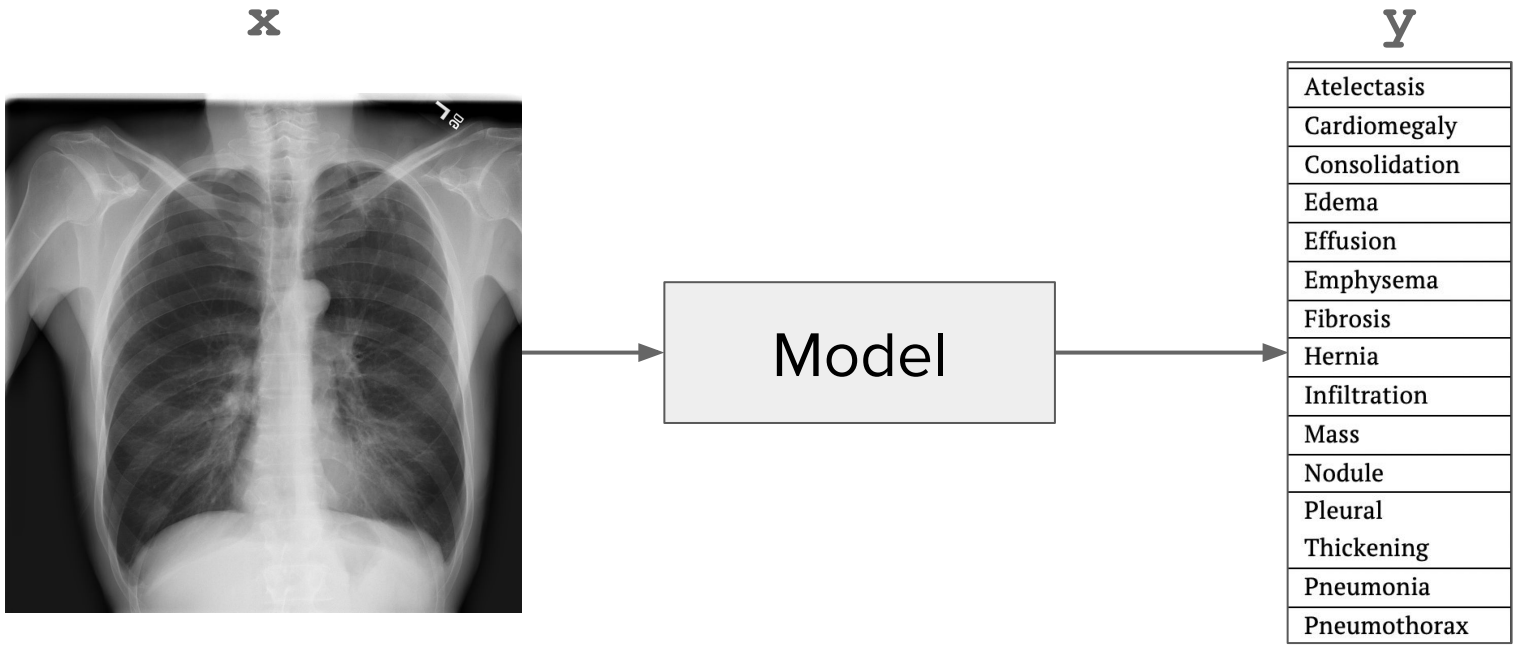
Data Challenges

- *Prevalence*
- *Noisy Labels*
- *Evaluation*

CheXNeXt



Rajpurkar & Irvin et al., PLOS Medicine, 2018

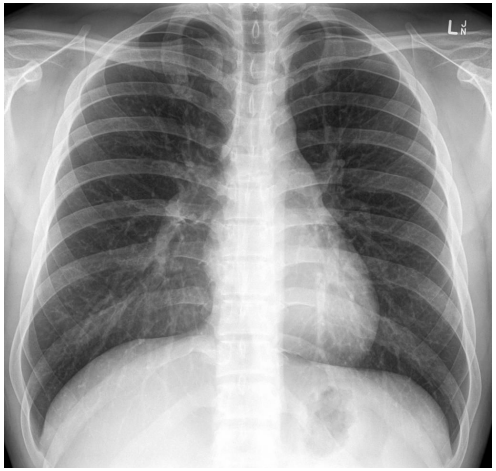


Rajpurkar & Irvin et al., PLOS Medicine, 2018

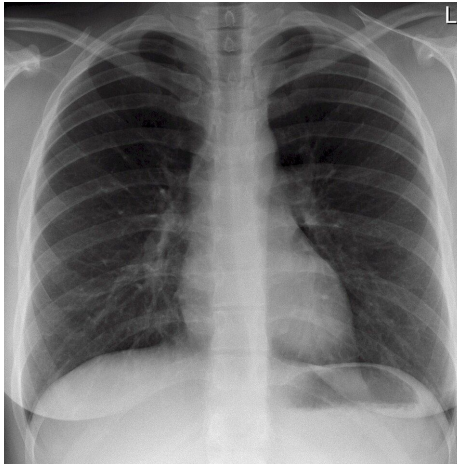
Data Challenges

- ***Prevalence***
- *Noisy Labels*
- *Evaluation*

Prevalence Challenge

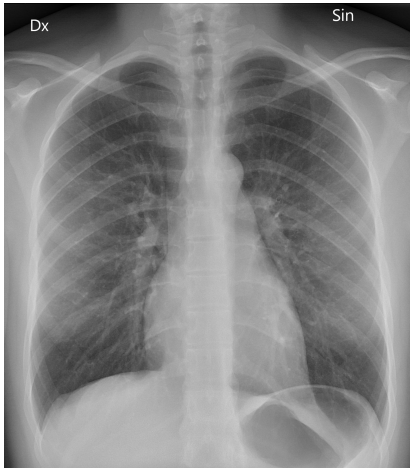


Normal

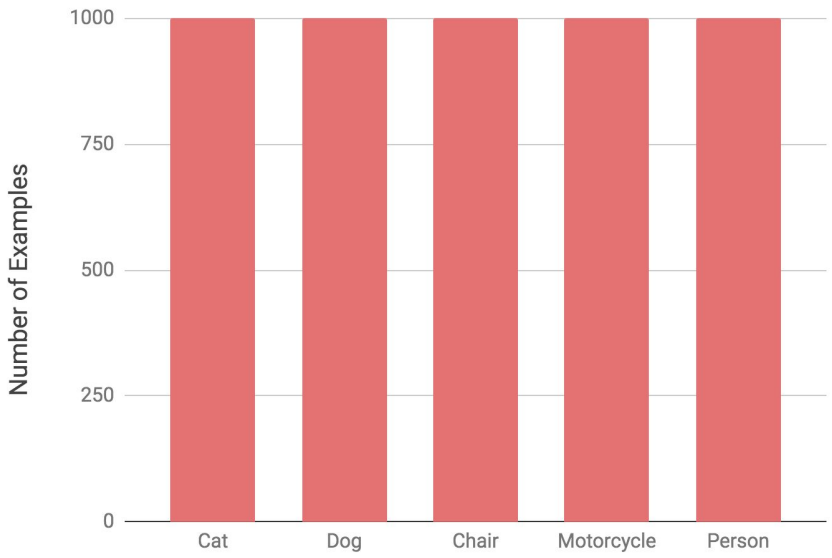


Normal

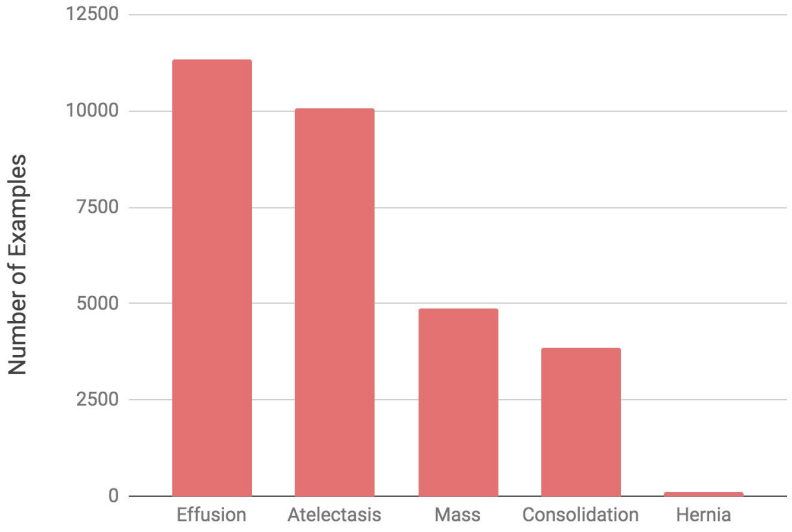
Nodule



Normal



ImageNet



ChestX-Ray14

Training Data as seen in real world

P1 Normal

P2 Normal

P3 Normal

P4 Nodule

P5 Normal

P6 Normal

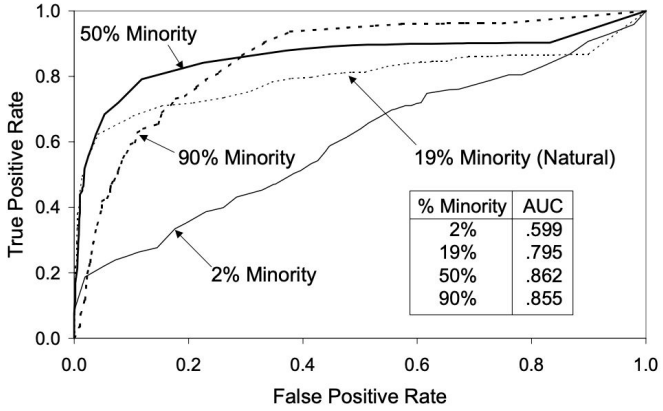
P7 Nodule

P8 Normal

→ **Learning Algorithm**

↓

Poor Test Set Performance



Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Training Data as seen in real world

P1 Normal
P2 Normal
P3 Normal
P4 Nodule
P5 Normal
P6 Normal
P7 Nodule
P8 Normal

$$L(X, y) = -\ y \log p(Y = 1|X) - (1 - y) \log p(Y = 0|X)$$

Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Cost Sensitive Learning

P1 Normal	1
P2 Normal	1
P3 Normal	1
P4 Nodule	1
P5 Normal	1
P6 Normal	1
P7 Nodule	1
P8 Normal	1

Total Contribution

Normal	Nodule
6	2

Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Cost Sensitive Learning

P1 Normal	2/8
P2 Normal	2/8
P3 Normal	2/8
P4 Nodule	6/8
P5 Normal	2/8
P6 Normal	2/8
P7 Nodule	6/8
P8 Normal	2/8

$$L(X, y) = -w_+ \cdot y \log p(Y = 1|X) - w_- \cdot (1 - y) \log p(Y = 0|X)$$

Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Cost Sensitive Learning

P1 Normal	2/8
P2 Normal	2/8
P3 Normal	2/8
P4 Nodule	6/8
P5 Normal	2/8
P6 Normal	2/8
P7 Nodule	6/8
P8 Normal	2/8

Total Contribution

Normal	Nodule
6 x 2/8 = 12/8	2 x 6/8 = 12/8

Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Cost Sensitive Learning

P1 Normal	2/8
P2 Normal	2/8
P3 Normal	2/8
P4 Nodule	6/8
P5 Normal	2/8
P6 Normal	2/8
P7 Nodule	6/8
P8 Normal	2/8

→ **Learning Algorithm** → **Good** Test Set Performance

Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.

Cost Sensitive Learning

P1 Normal	2/8
P2 Normal	2/8
P3 Normal	2/8
P4 Nodule	6/8
P5 Normal	2/8
P6 Normal	2/8
P7 Nodule	6/8
P8 Normal	2/8

→ **Learning Algorithm** → **Good** Test Set Performance

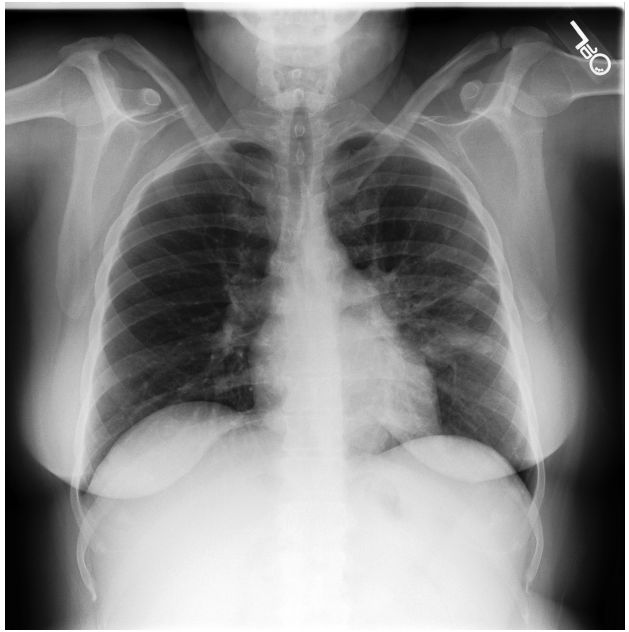
Pathology	AUC of algorithm trained on original training set (95% CI)
Atelectasis	0.843 (0.803,0.878)
Cardiomegaly	0.827 (0.784,0.867)
Consolidation	0.860 (0.819,0.896)
Edema	0.933 (0.901,0.96)
Effusion	0.901 (0.869,0.929)
Emphysema	0.790 (0.709,0.862)
Fibrosis	0.809 (0.705,0.9)
Hernia	0.782 (0.704,0.852)
Infiltration	0.710 (0.636,0.778)
Mass	0.880 (0.828,0.926)
Nodule	0.882 (0.842,0.92)
Pleural Thickening	0.783 (0.731,0.834)
Pneumonia	0.774 (0.678,0.858)
Pneumothorax	0.917 (0.868,0.956)

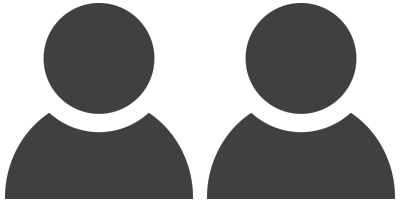
Weiss et al. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. 2003
Buda et al. A systematic study of the class imbalance problem in convolutional neural networks. 2018.
Rajpurkar & Irvin et al., PLOS Medicine, 2018

Data Challenges

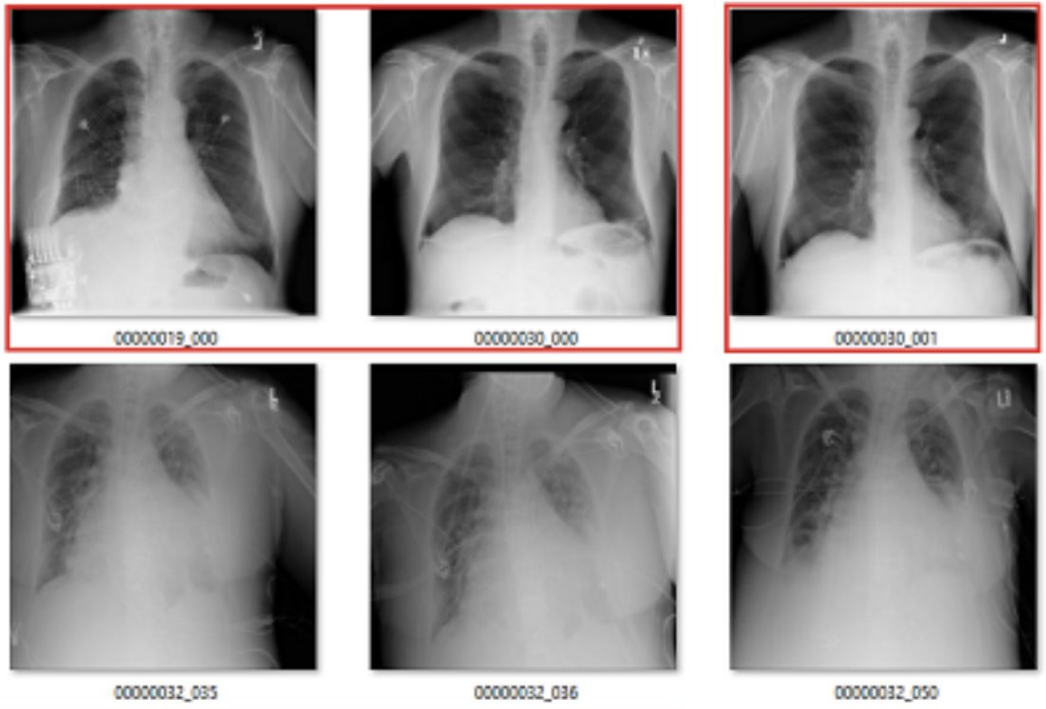
- *Prevalence*
- ***Noisy Labels***
- *Evaluation*

Noisy Labels Challenge



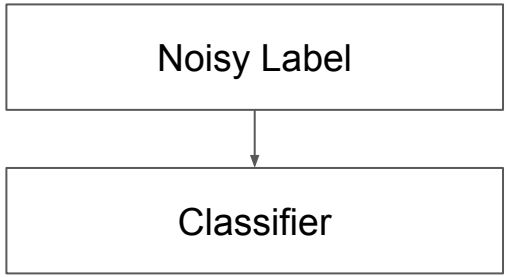


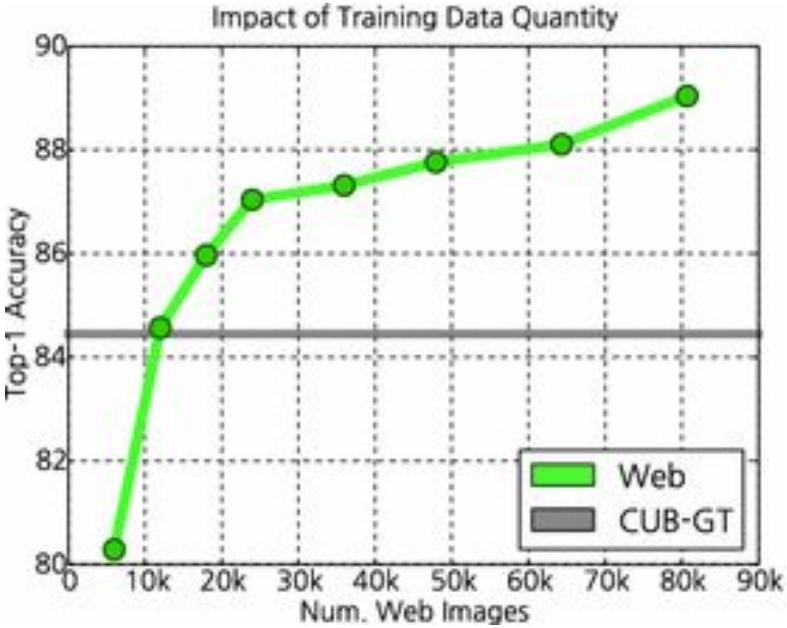
Noisy Labels



Oakden-Rayner. Exploring the ChestXray14 dataset: problems. 2017

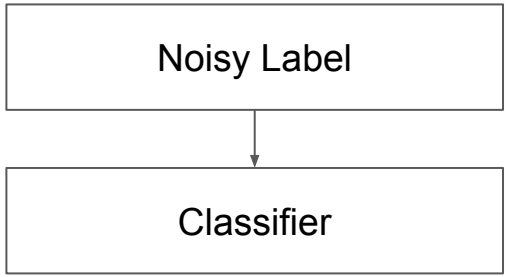
Keeping the Noise





Joulin et al. Learning Visual Features from Large Weakly Supervised Data. ECCV 2016.
Krause et al. The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition. ECCV 2016.

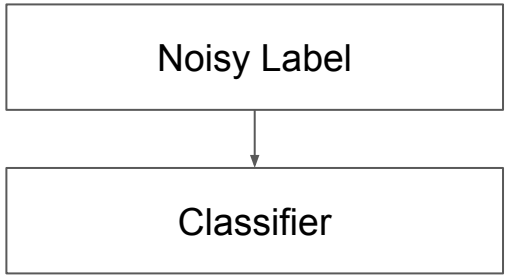
Keeping the Noise



Pathology	AUC of algorithm trained on original training set (95% CI)
Atelectasis	0.843 (0.803,0.878)
Cardiomegaly	0.827 (0.784,0.867)
Consolidation	0.860 (0.819,0.896)
Edema	0.933 (0.901,0.96)
Effusion	0.901 (0.869,0.929)
Emphysema	0.790 (0.709,0.862)
Fibrosis	0.809 (0.705,0.9)
Hernia	0.782 (0.704,0.852)
Infiltration	0.710 (0.636,0.778)
Mass	0.880 (0.828,0.926)
Nodule	0.882 (0.842,0.92)
Pleural Thickening	0.783 (0.731,0.834)
Pneumonia	0.774 (0.678,0.858)
Pneumothorax	0.917 (0.868,0.956)

Rajpurkar & Irvin et al., PLOS Medicine, 2018

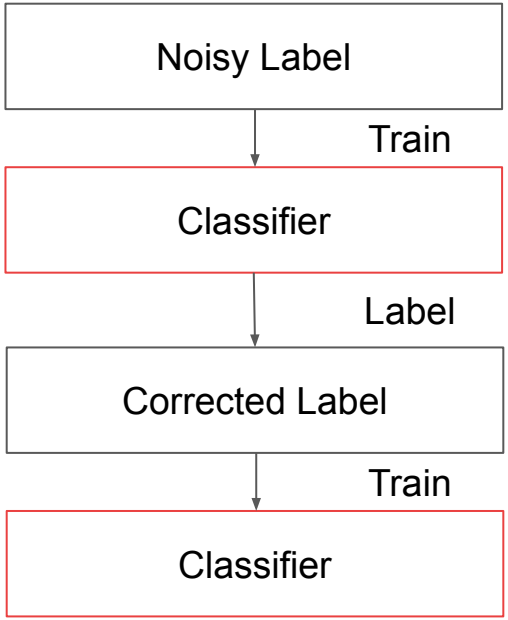
Keeping the Noise



Pathology	Prevalence (examples)	Algorithms
Effusion	**11,000**	0.901
Hernia	**110**	0.851

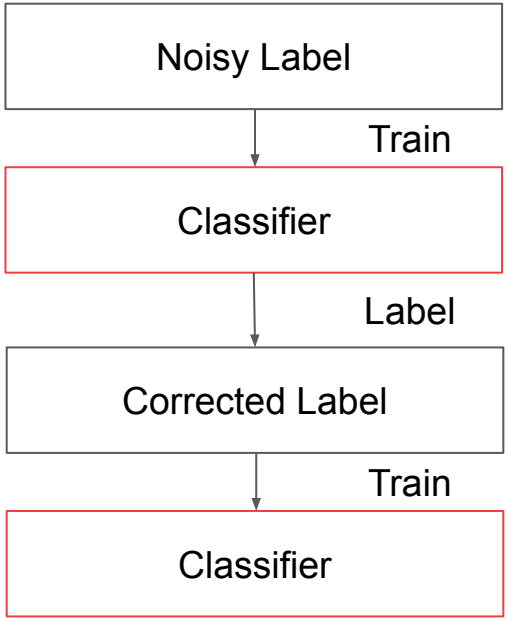
Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)
Joulin et al. Learning Visual Features from Large Weakly Supervised Data. ECCV 2016.
Rajpurkar & Irvin et al., PLOS Medicine, 2018

Correcting Noise via Self-Training



Yarowsky et al. Unsupervised word sense disambiguation rivaling supervised methods. ACL. 1995.
Xie et al. Self-training with Noisy Student improves ImageNet classification. 2019.
Rajpurkar & Irvin et al., PLOS Medicine, 2018

Correcting Noise



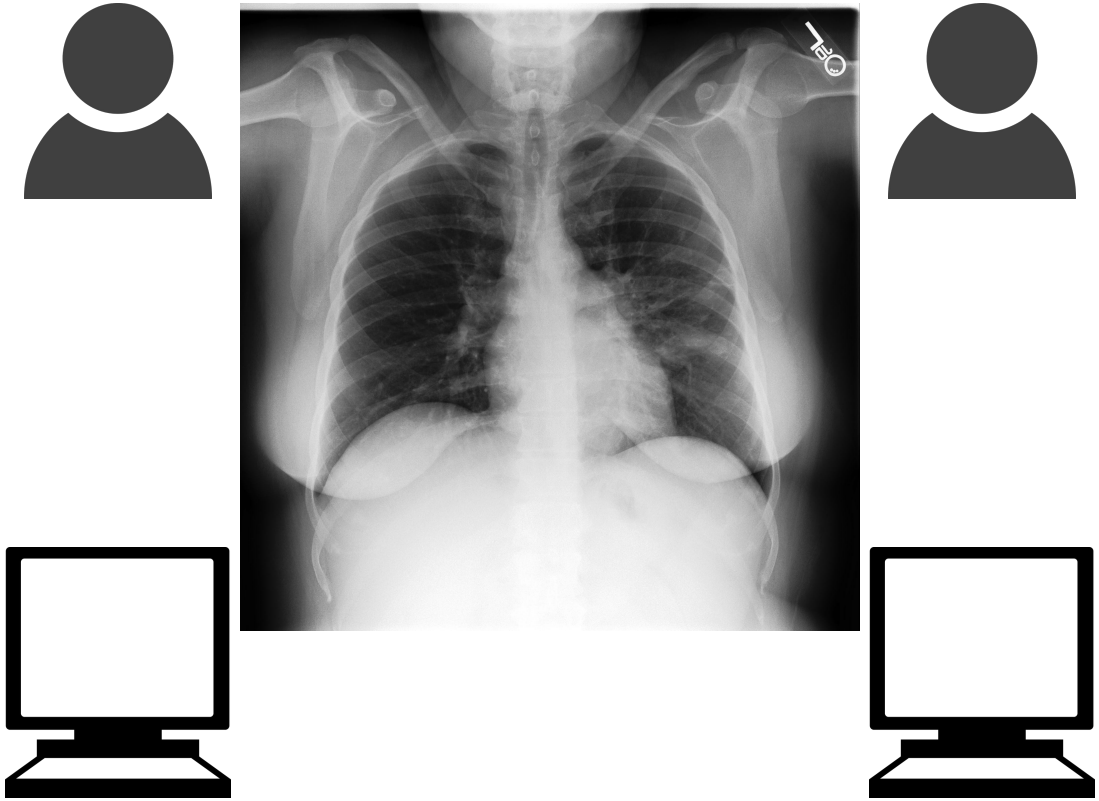
Pathology	AUC of algorithm trained on original training set (95% CI)	AUC of algorithm trained on relabeled training set (95% CI)
Atelectasis	0.843 (0.803,0.878)	0.862 (0.825,0.895)
Cardiomegaly	0.827 (0.784,0.867)	0.831 (0.79,0.87)
Consolidation	0.860 (0.819,0.896)	0.893 (0.859,0.924)
Edema	0.933 (0.901,0.96)	0.924 (0.886,0.955)
Effusion	0.901 (0.869,0.929)	0.901 (0.868,0.93)
Emphysema	0.790 (0.709,0.862)	0.704 (0.567,0.833)
Fibrosis	0.809 (0.705,0.9)	0.806 (0.719,0.884)
Hernia	0.782 (0.704,0.852)	0.851 (0.785,0.909)
Infiltration	0.710 (0.636,0.778)	0.721 (0.651,0.786)
Mass	0.880 (0.828,0.926)	0.909 (0.864,0.948)
Nodule	0.882 (0.842,0.92)	0.894 (0.853,0.93)
Pleural Thickening	0.783 (0.731,0.834)	0.798 (0.744,0.849)
Pneumonia	0.774 (0.678,0.858)	0.851 (0.781,0.911)
Pneumothorax	0.917 (0.868,0.956)	0.944 (0.915,0.969)

Rajpurkar & Irvin et al., PLOS Medicine, 2018

Data Challenges

- *Prevalence*
- *Noisy Labels*
- ***Evaluation***

Evaluation Challenge

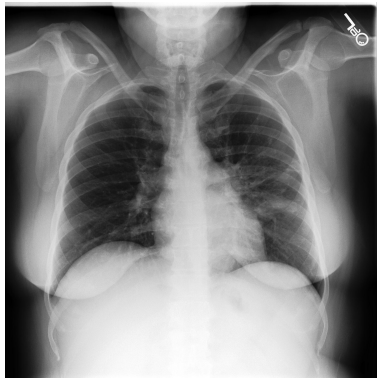


Evaluation Challenge

Training Set

Noisy Ground Truth Label

Test Set



Strong Ground Truth

Train

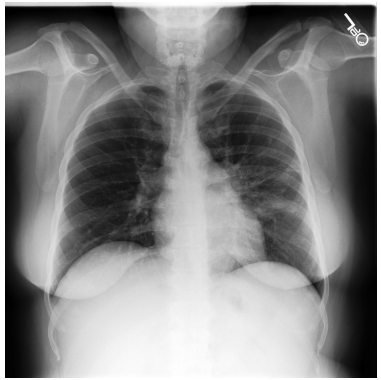
Classifier

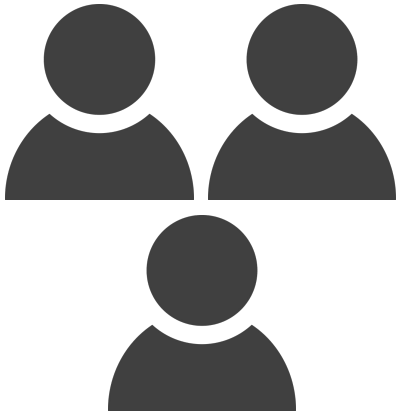
Evaluation Challenge

Training Set

Noisy Ground Truth Label

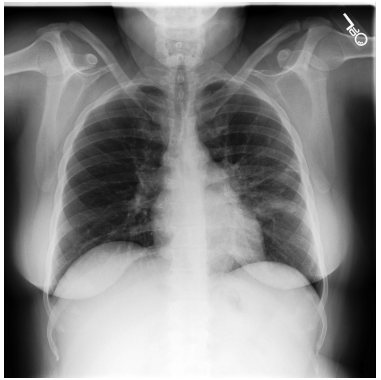
Test Set





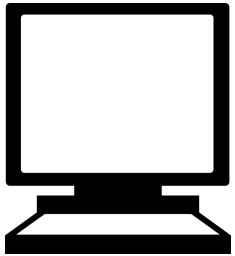
Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)
Joulin et al. Learning Visual Features from Large Weakly Supervised Data. ECCV 2016.
Rajpurkar & Irvin et al., PLOS Medicine, 2018

Test Set



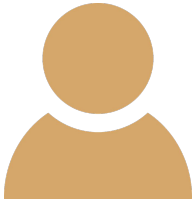
Ground Truth

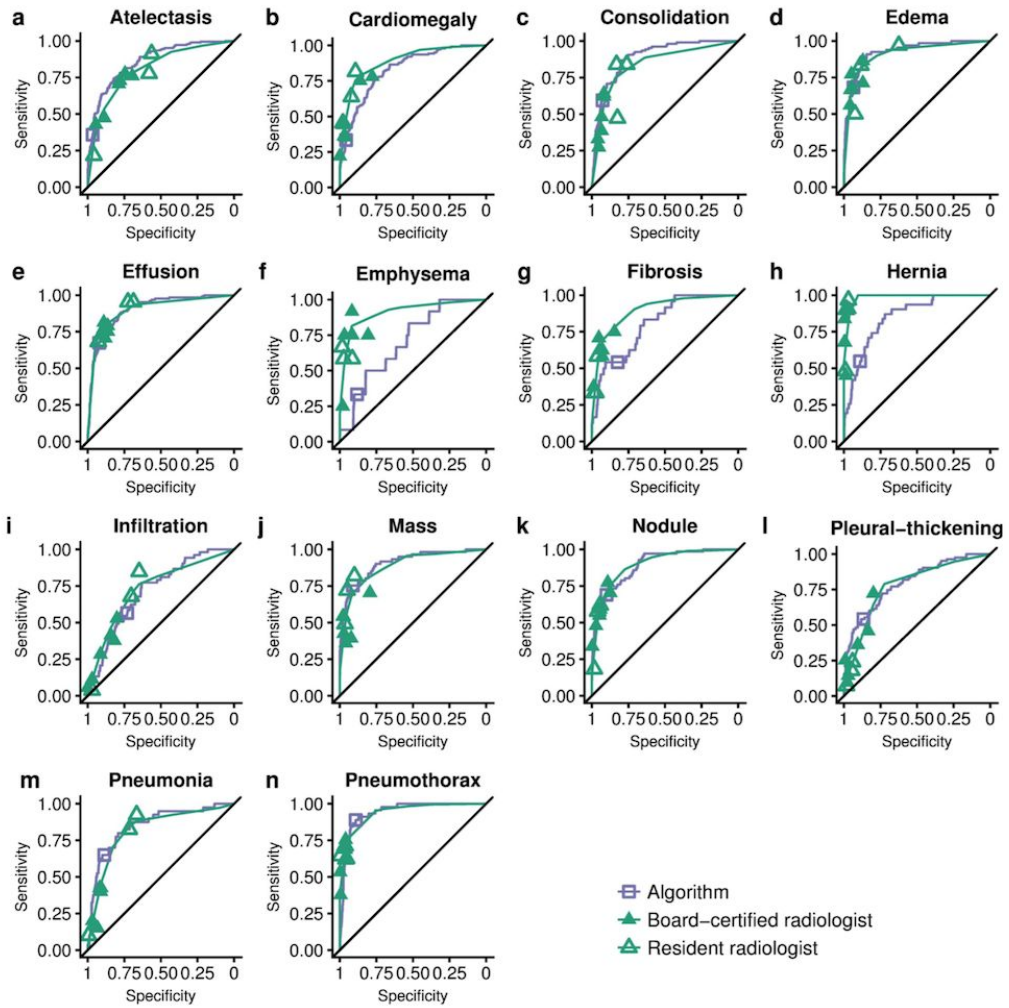
AI



v.s.

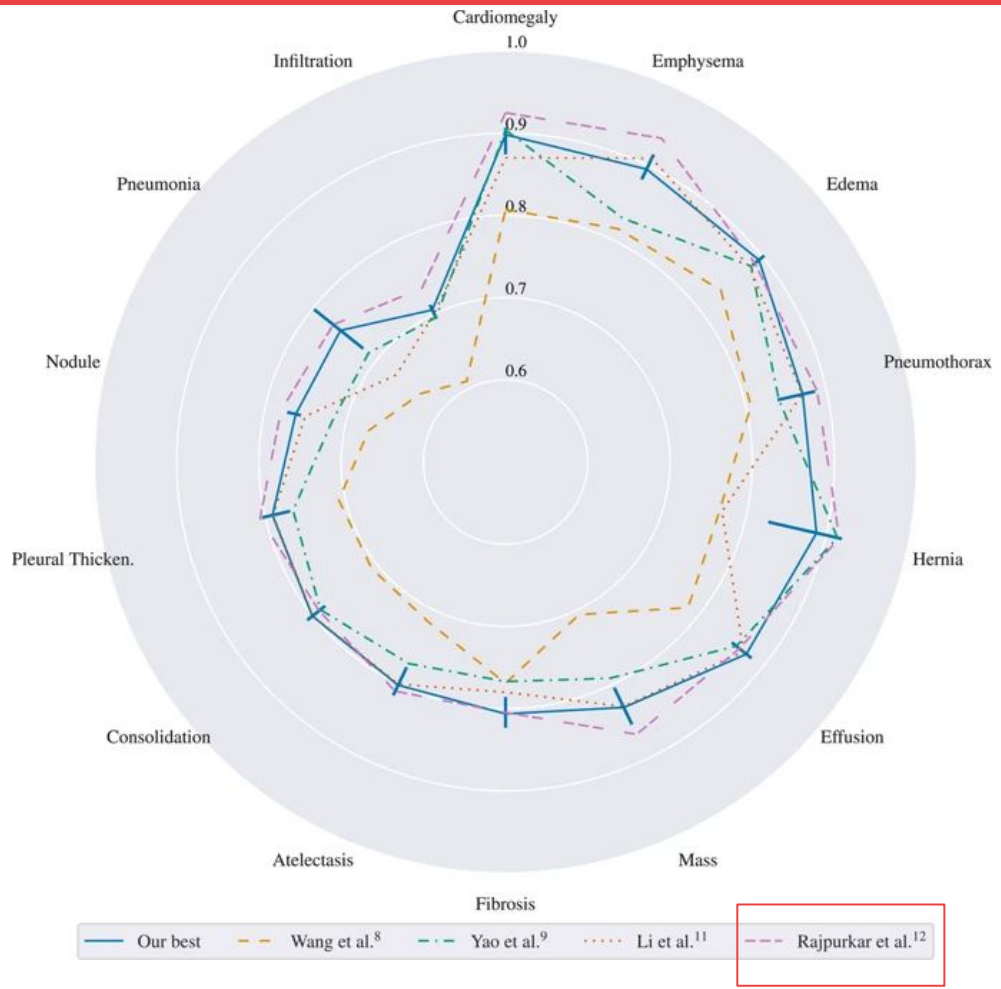
Reader





Model Comparable to Radiologists

Except when small number of examples (e.g. Hernia)



Baltruschat et al., Scientific Reports, 2019

Recent study comparing several chest x-ray models

CheXNeXt training strategy competitive

Data Challenges

- *Prevalence*
- *Noisy Labels*
- *Evaluation*

→

Cost-Sensitive Learning

Weak-Supervision at Scale

Strong Ground Truth Collection

Data Challenges

- *Prevalence*
- *Noisy Labels*
- *Evaluation*

→ *Design Datasets*

Design datasets that solve data challenges

CheXpert

200k Chest X-Rays

MRnet DATASET

1,370 Knee MRI exams

MURA

40k Bone x-rays

Work with Dr. Matt Lungren, Dr. Curt Langlotz and many others at the Stanford Med School

Irvin & Rajpurkar et al. 2019, AAAI; Bien & Rajpurkar et al, PLOS Medicine; Rajpurkar & Irvin et al. 2018, MIDL

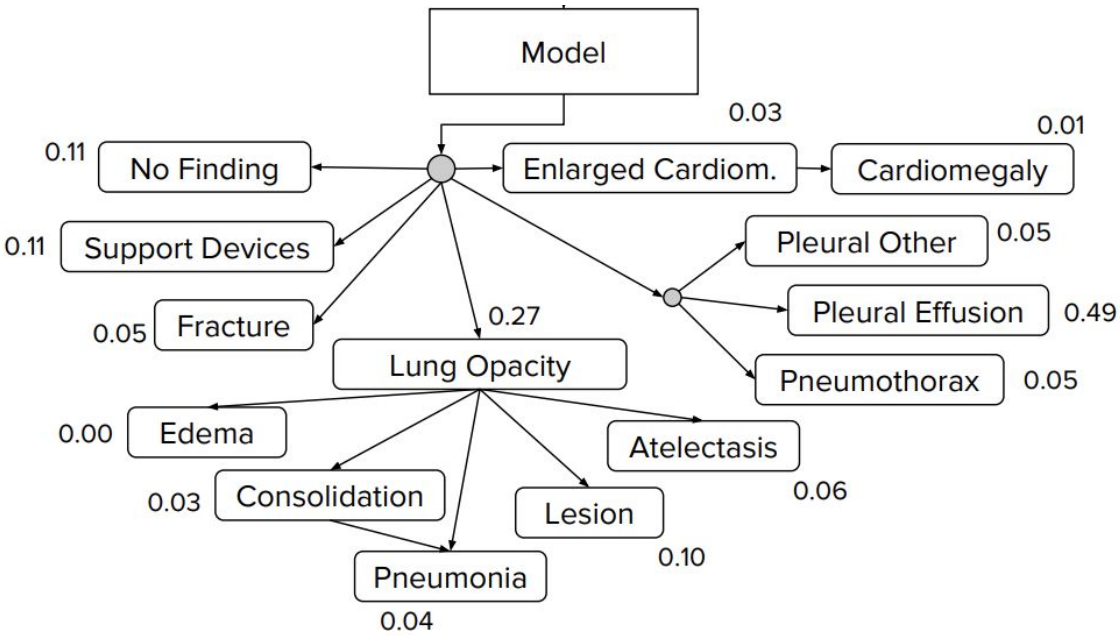
CheXpert

Irvin & Rajpurkar et al. 2019, AAAI

Designing to Tackle Prevalence

CheXPert

200k Chest X-Rays



Irvin & Rajpurkar et al. 2019, AAAI

Designing to Tackle Noisy Labels

1. *unremarkable* cardiomediastinal silhouette

2. diffuse reticular pattern, which can be seen with an atypical infection **or** chronic fibrotic change. *no* focal consolidation.

3. *no* pleural effusion or pneumothorax

4. mild degenerative changes in the lumbar spine and old right rib fractures.

Observation	Labeler Output
No Finding	
Enlarged Cardiom.	0
Cardiomegaly	
Lung Opacity	1
Lung Lesion	
Edema	
Consolidation	0
Pneumonia	u
Atelectasis	
Pneumothorax	0
Pleural Effusion	0
Pleural Other	
Fracture	1
Support Devices	

Designing to Tackle Noisy Labels

Open-Sourced Report Labeler

Category	Mention F1		Negation F1		Uncertain F1	
	NIH	Ours	NIH	Ours	NIH	Ours
Atelectasis	0.976	**0.998**	0.526	**0.833**	0.661	**0.936**
Cardiomegaly	0.647	**0.973**	0.000	**0.909**	0.211	**0.727**
Consolidation	0.996	**0.999**	0.879	**0.981**	0.438	**0.924**
Edema	0.978	**0.993**	0.873	**0.962**	0.535	**0.796**
Pleural Effusion	0.985	**0.996**	0.951	**0.971**	0.553	**0.707**
Pneumonia	0.660	**0.992**	0.703	**0.750**	0.250	**0.817**
Pneumothorax	0.993	**1.000**	0.971	**0.977**	0.167	**0.762**

Designing to Tackle Noisy Labels

MIMIC-CXR Database

	Mention		Negation		Uncertainty	
Category	NegBio	CheXpert	NegBio	CheXpert	NegBio	CheXpert
Atelectasis	0.930	**0.998**	**0.727**	0.400	0.379	**0.835**
Cardiomegaly	0.596	**0.954**	0.043	**0.830**	0.000	**0.333**
Consolidation	0.966	**0.986**	0.917	**0.958**	0.235	**0.486**
Edema	0.855	**0.996**	0.701	**0.878**	0.214	**0.742**
Pleural Effusion	0.971	**0.987**	0.873	**0.947**	0.368	**0.500**
Pneumonia	0.836	**0.981**	0.750	**0.785**	0.388	**0.674**
Pneumothorax	0.983	**0.998**	**0.951**	0.948	0.182	**0.286**

Johnson et al. 2019. MIMIC-CXR. 2019

Designing to Tackle Noisy Labels

NegBio User Guide

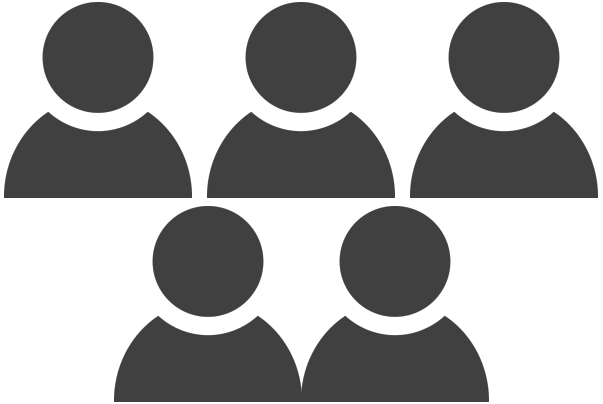
Run the pipeline step-by-step

The step-by-step pipeline generates all intermediate documents. You can easily rerun one step if it makes errors. The whole steps are

1. `text2bioc` combines text into a BioC XML file.
2. `normalize` removes noisy text such as `[**Patterns**]`.
3. `section_split` splits the report into sections based on titles at `patterns/section_titles.txt`
4. `ssplit` splits text into sentences.
5. Named entity recognition
   a. `dner_mm` detects UMLS concepts using MetaMap.
   b. `dner_chexpert` detects concepts using the CheXpert vocabularies at `negbio/chexpert/phrases`.
6. `parse` parses sentence using the Bllip parser.
7. `ptb2ud` converts the parse tree to universal dependencies using Stanford converter.
8. Negation detection
   a. `neg` detects negative and uncertain findings.
   b. `neg_chexpert` detects positive, negative and uncertain findings (recommended)

Peng Y et al. NegBio. AMIA 2018 Informatics Summit. 2018.
Wang X et al. ChestX-ray8. CVPR 2017.

Designing to Tackle Evaluation



Ground Truth

Readers

Irvin & Rajpurkar et al. 2019, AAAI

Designing to Tackle Evaluation

2400 Downloads
84 Teams Competing

Rank	Date	Model	AUC	Num Rads Below Curve
1	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute*	**0.930**	2.6
2	Oct 10, 2019	yww	0.929	2.8
3	Sep 09, 2019	Multi-Stage-Learning-CNN-V3 (ensemble) *VINBDI Medical Imaging Team*	0.928	2.6
4	Sep 19, 2019	SenseXDR *ensemble*	0.927	2.6
5	Sep 18, 2019	ihil (ensemble) *UESTC*	0.927	2.6

Irvin & Rajpurkar et al. 2019, AAAI

Data Challenges

- *Prevalence*
- *Noisy Labels*
- *Evaluation*

→ *Design Datasets*

Challenges for AI in Medical Image Interpretation

Data

Deployment

Deployment Challenges

- *Algorithm Translation*
- *Clinical Impact*

Deployment Challenges

- ***Algorithm Translation***
- *Clinical Impact*

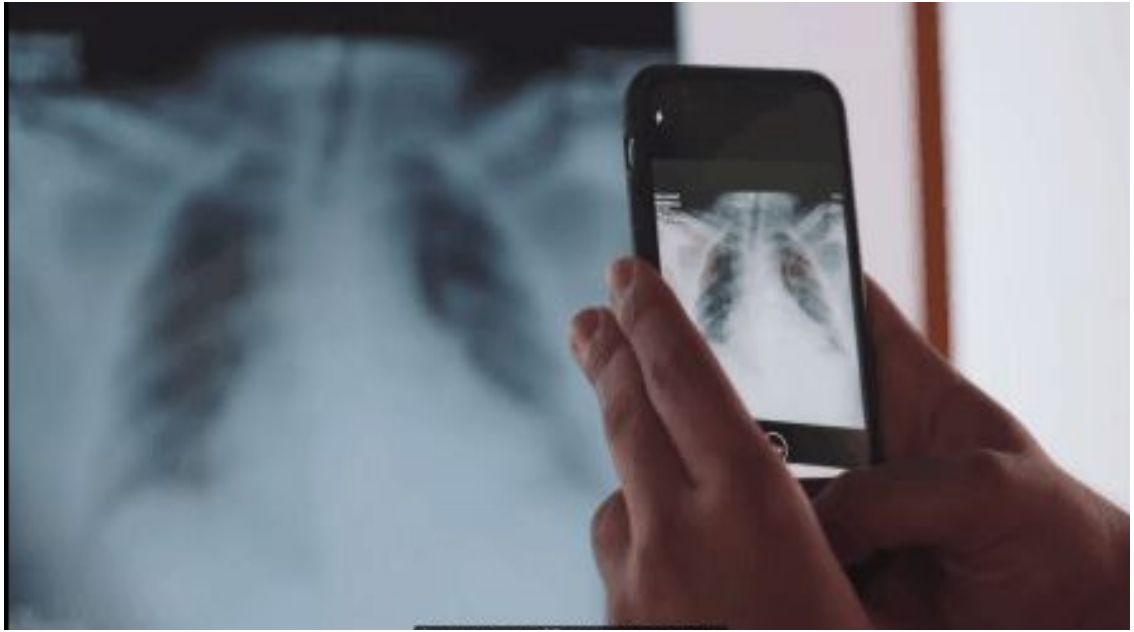
Algorithm Translation Challenge

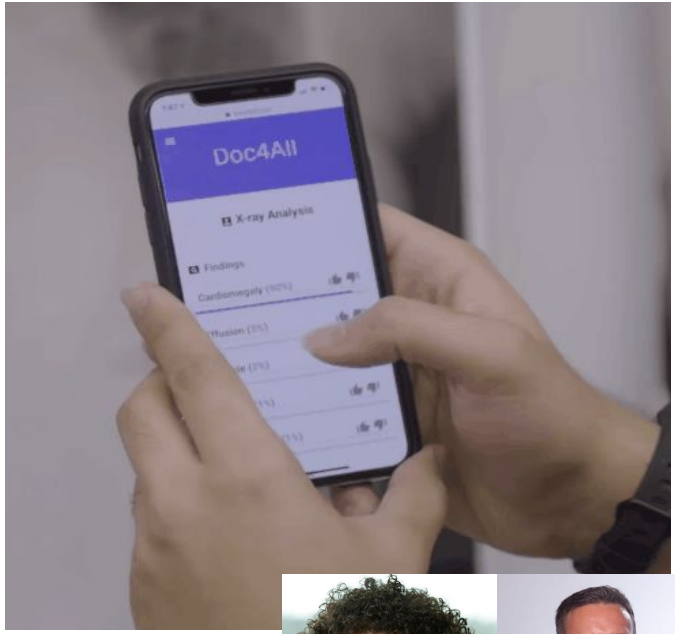
Translation Engineering

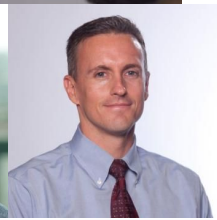
Model

Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
Nodule
Pleural Thickening
Pneumonia
Pneumothorax

Engineering for Translation: XRay4All

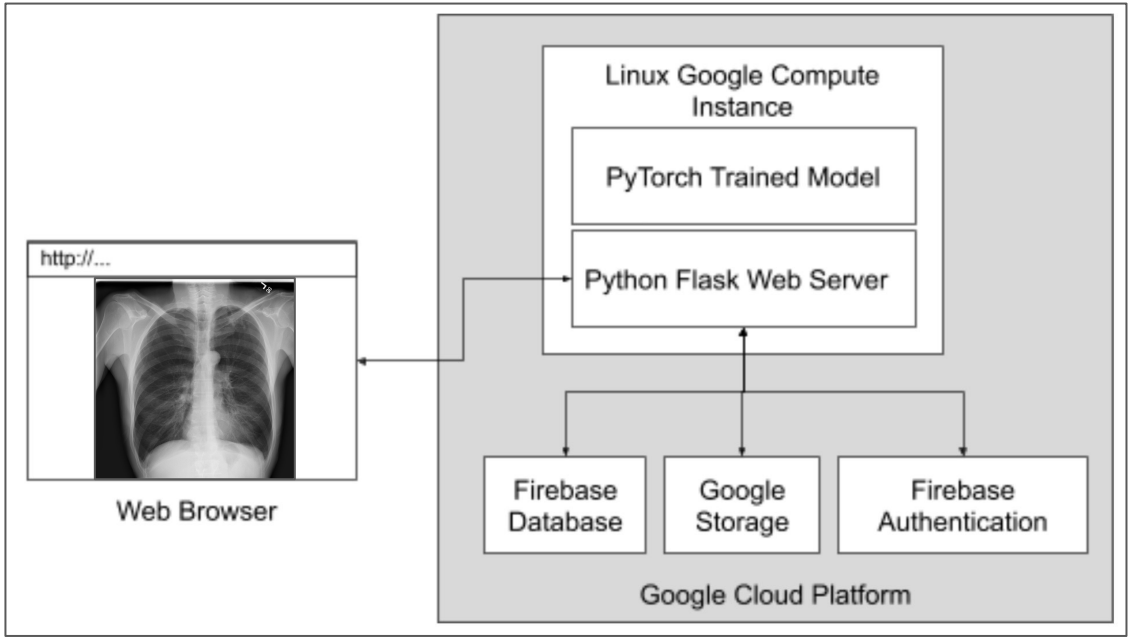


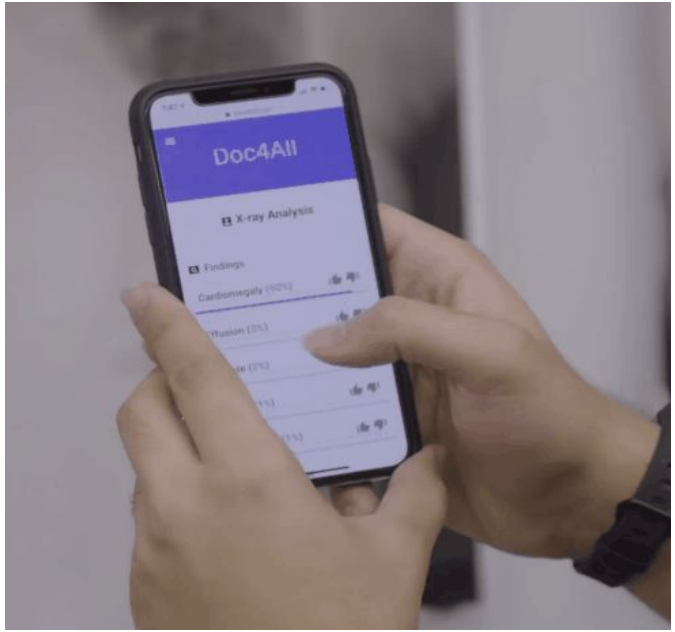




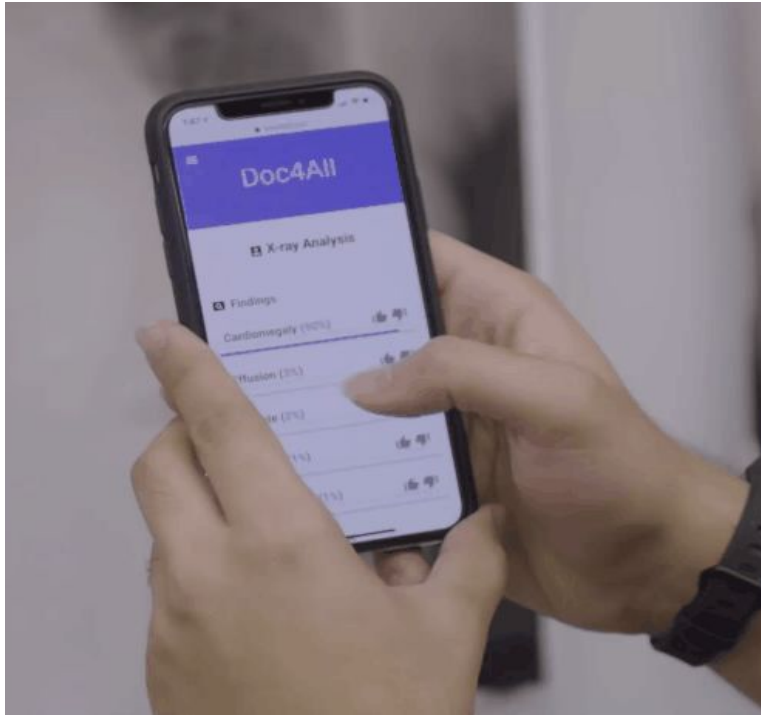
With Amir Kiani, Dr. Matt Lungren, and others, Stanford ML Group

Engineering for Translation: XRay4All





XRay4All: Live demo

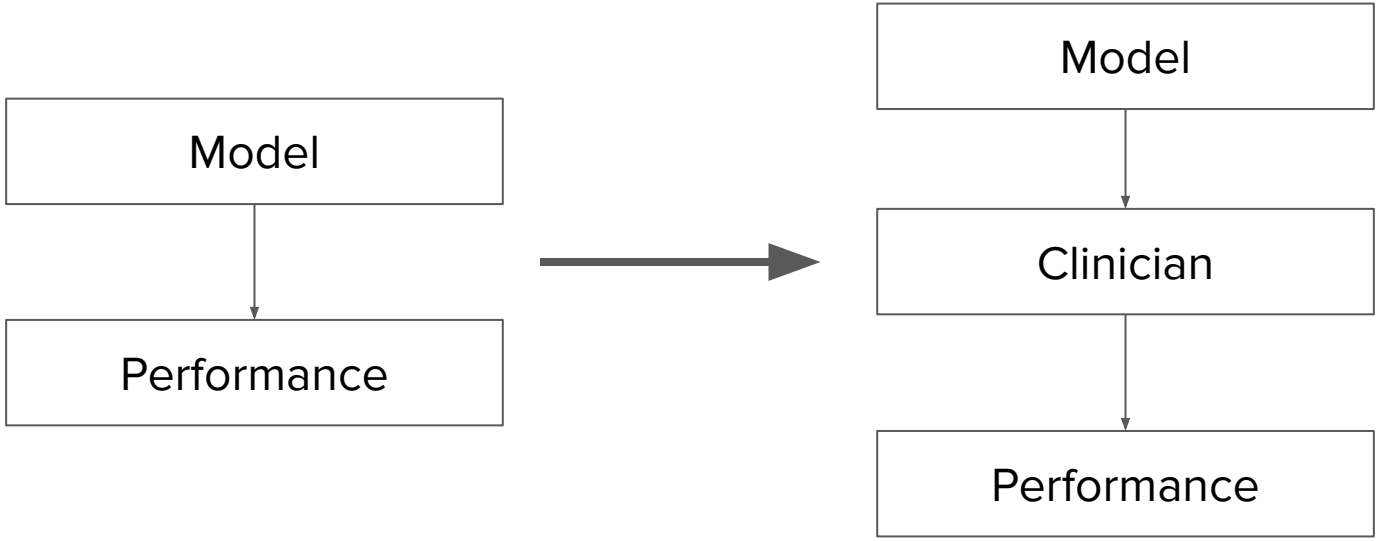


Link1 Link2 Link3

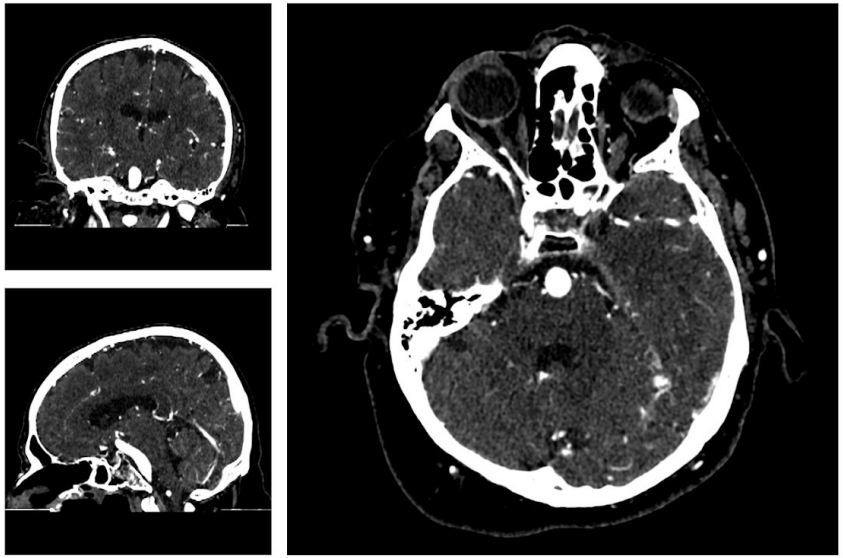
Deployment Challenges

- *Algorithm Translation*
- ***Clinical Impact***

Clinical Impact Challenge



HeadXNet



Park & Chute & Rajpurkar et al., JAMA Network Open. 2019

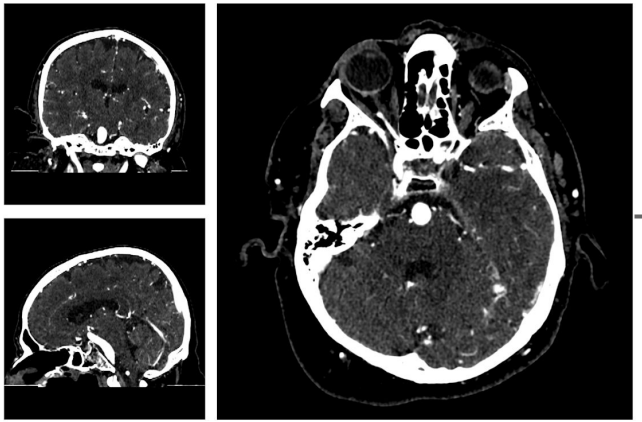
x

2D CNN

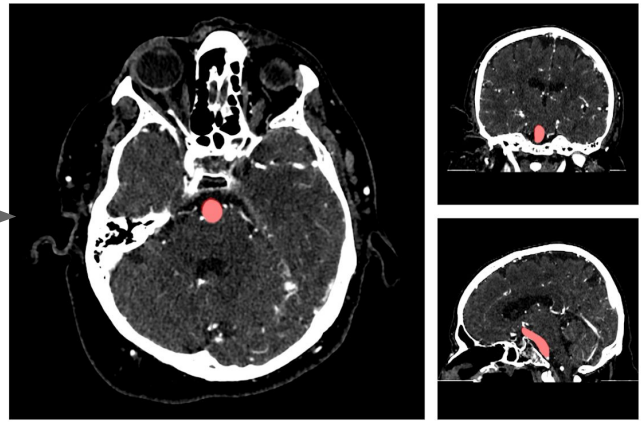
Pneumonia

y

x

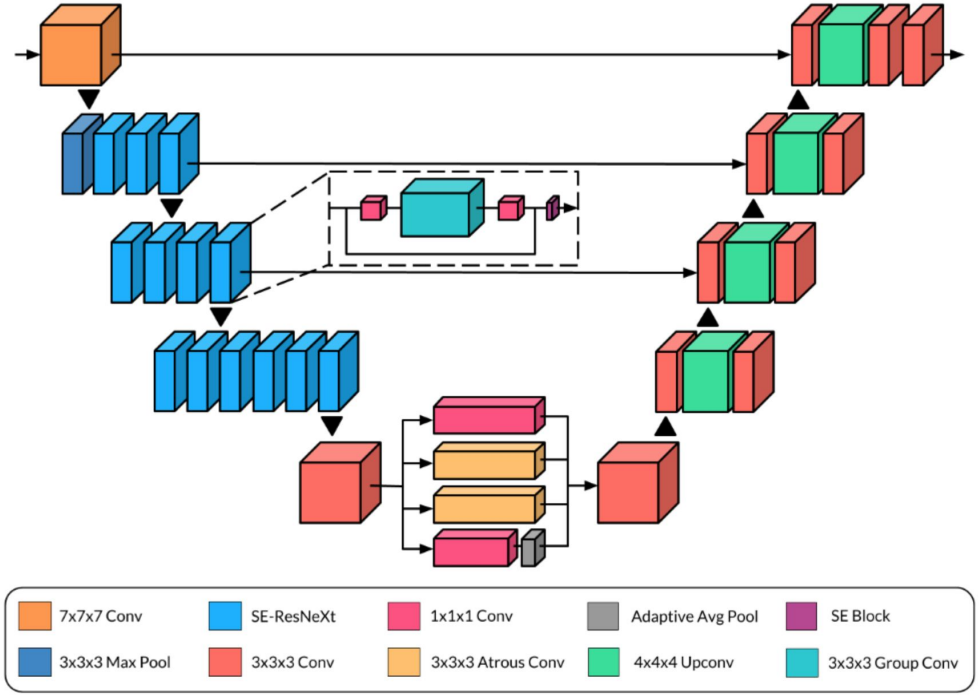


3D CNN



y

HeadXNet takes in a series of CTA images and outputs segmentation for each voxel.



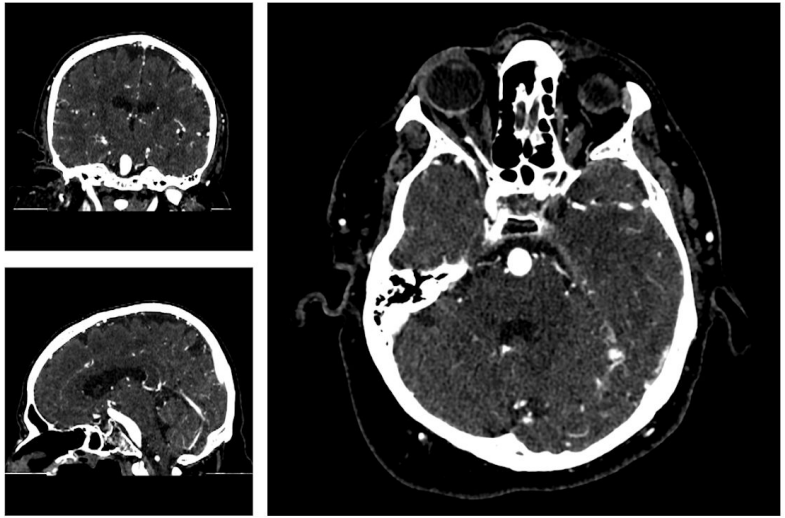
Park & Chute & Rajpurkar et al, JAMA Network Open, 2019

High Sensitivity Model for Detecting Aneurysms

	Sensitivity (95% CI)	*Specificity (95% CI)*	*Accuracy (95% CI)*
Clinicians	0.831 (0.794, 0.862)	0.960 (0.937, 0.974)	0.893 (0.782, 0.912)
Model	0.949 (0.860, 0.963)	0.661 (0.530, 0.771)	0.809 (0.727, 0.870)

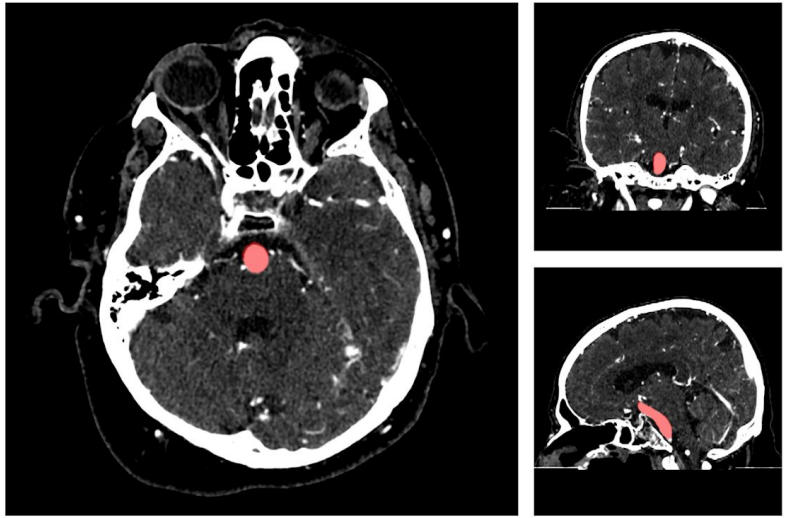
Park & Chute & Rajpurkar et al, JAMA Network Open, 2019

B. Original CTA Scan (Unaugmented Read)



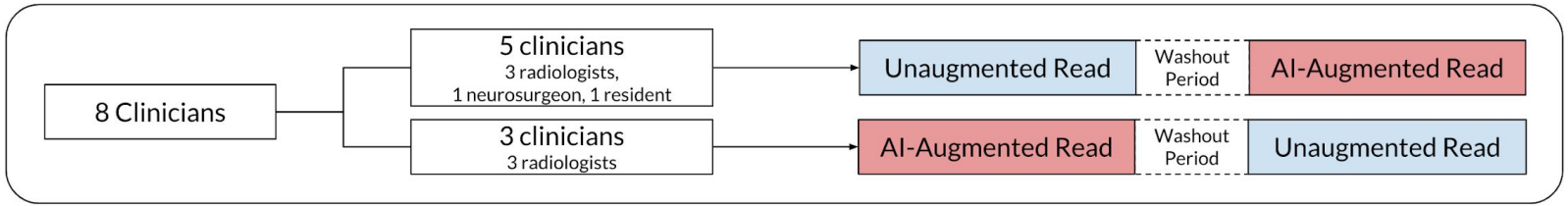
Unaugmented CTA Aneurysm Interpretation

C. Model Segmentation Overlay (AI-Augmented Read)



AI-Augmented CTA Aneurysm Interpretation

A. Crossover Study Design



Park & Chute & Rajpurkar et al, Jama Network Open, 2019

0.06 increase in **sensitivity** of detecting aneurysms.

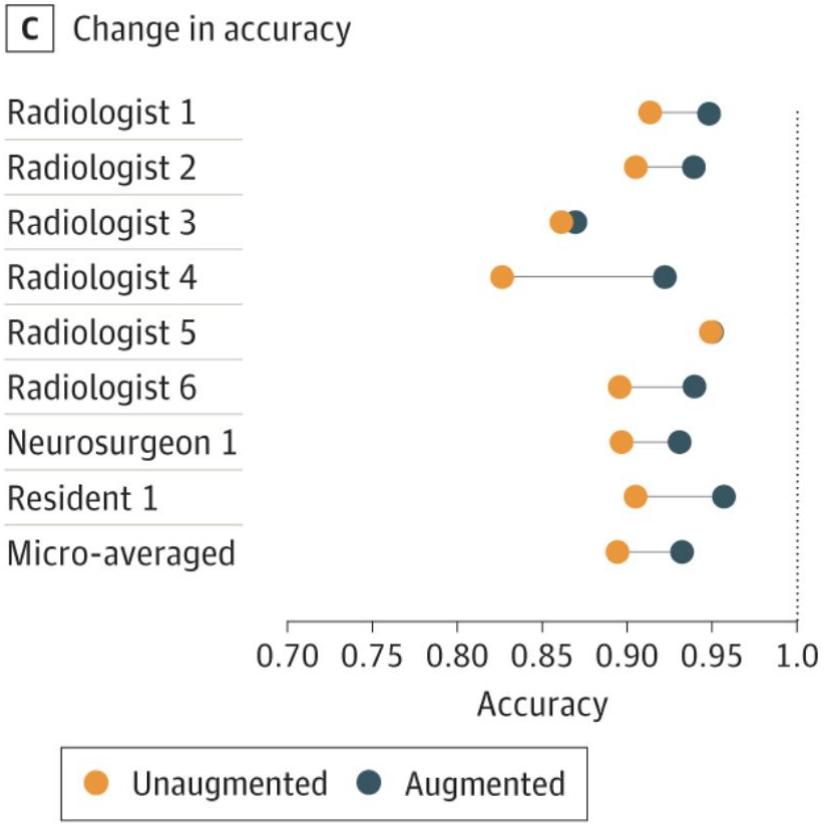
Specificity difference is positive, but not significant.

Metric	Without Augmentation (95% CI)	With Augmentation (95% CI)	P-value
Sensitivity	**0.831 (0.794, 0.862)**	**0.890 (0.858, 0.915)**	**0.01**
Specificity	0.960 (0.937, 0.974)	0.975 (0.957, 0.986)	0.16
Accuracy	0.893 (0.782, 0.912)	0.932 (0.913, 0.946)	0.02

Park & Chute & Rajpurkar et al, JAMA Network Open. 2019

Greatest improvement for clinician with lowest unaugmented score.

Smallest improvement for clinician with highest unaugmented score.



Park & Chute & Rajpurkar et al, JAMA Network Open. 2019

Deployment Challenges

- *Algorithm Translation*
- *Clinical Impact*

→

Engineering For Translation

Clinician Studies

Opportunities for AI in Medical Image Interpretation

Data

Deployment

Data Opportunities

- *Fundamental Advances in Small Data and Transfer Learning*

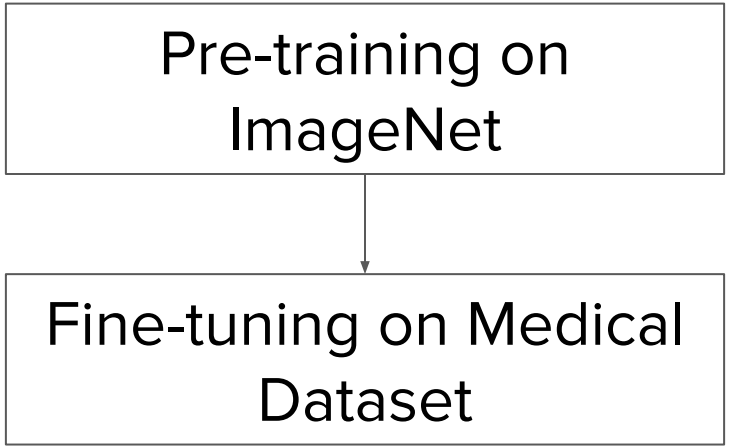
Deployment Opportunities

- *ML Generalizability and Clinical Impact Studies*

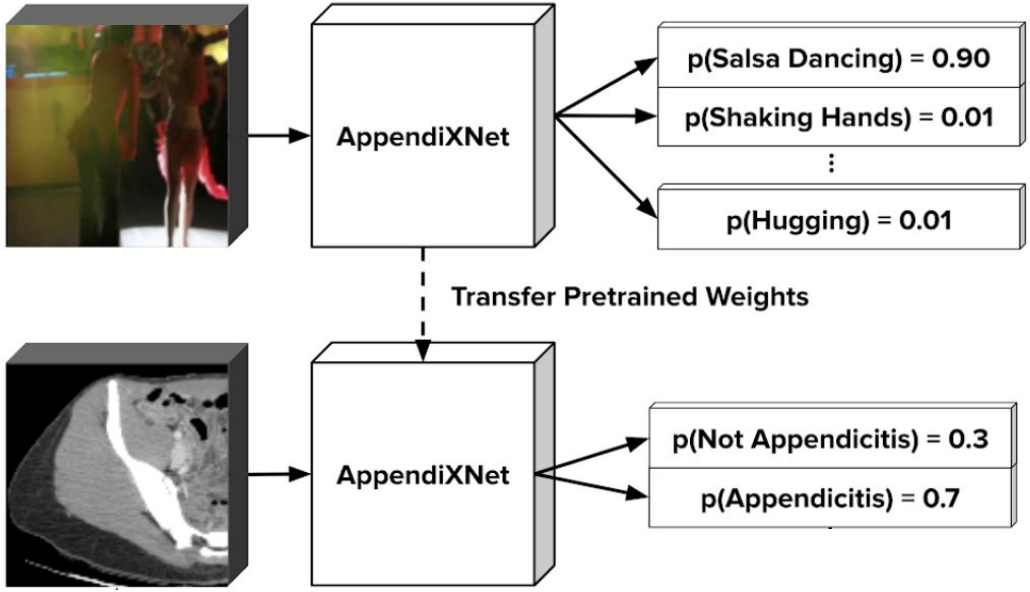
Data Opportunities

- *Fundamental Advances in Small Data and Transfer Learning*

Transfer learning on images effective



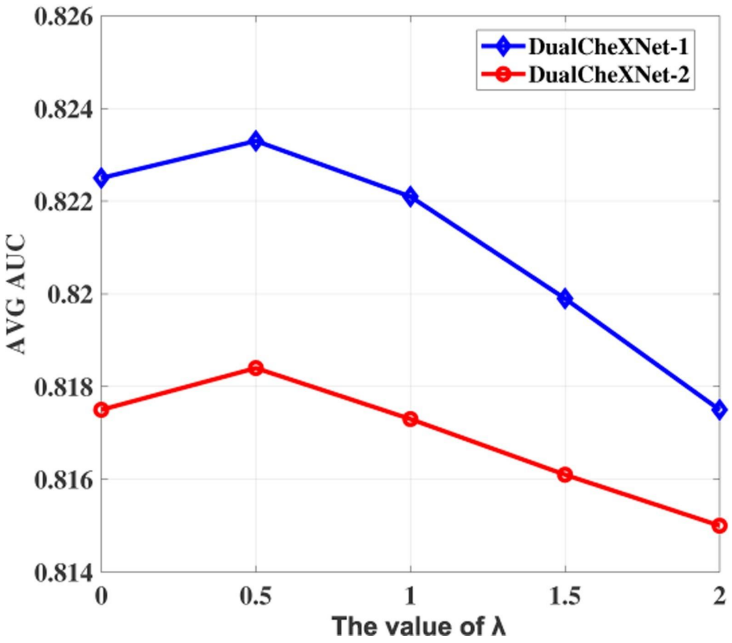
Transfer learning on video effective for small datasets



Rajpurkar & Irvin & Park et al. In Review.

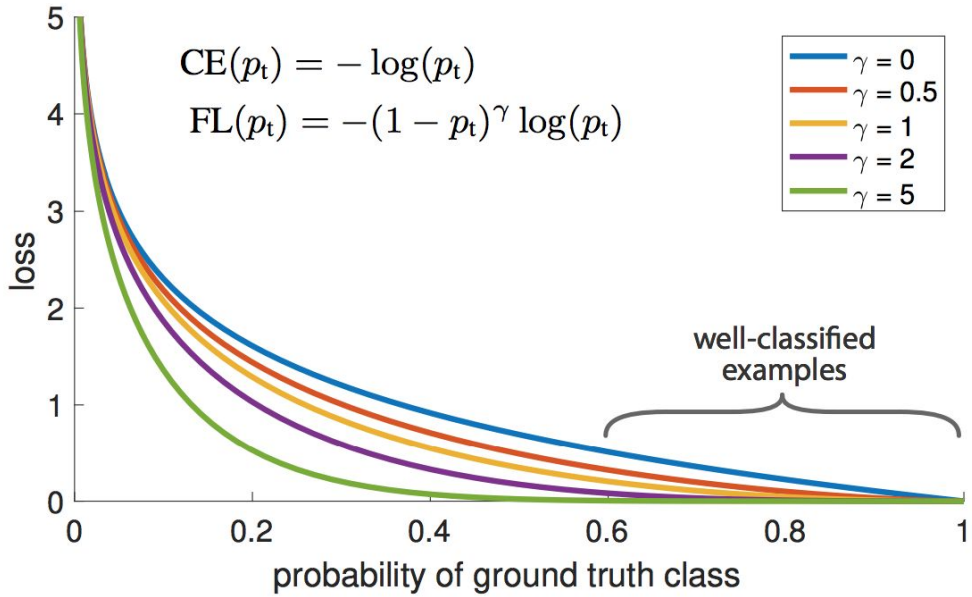
Class imbalance addressed to focus more on the hard-classified samples

$$\boldsymbol{L}(p(i|\boldsymbol{I}), \boldsymbol{x}) = -\frac{1}{|P|}\sum_{x_i=1}(1-p(i|\boldsymbol{I}))^{\lambda}\ln(p(i|\boldsymbol{I})) - -\frac{1}{|N|}\sum_{x_i=0}p(i|\boldsymbol{I})^{\lambda}\ln(1-p(i|\boldsymbol{I}))$$



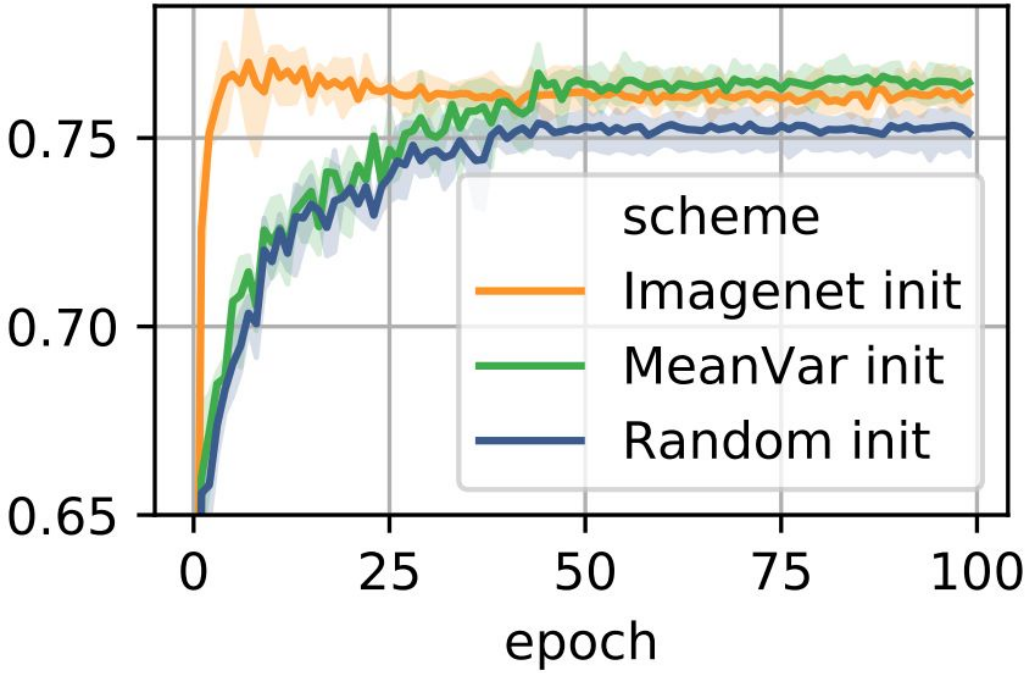
DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays, Chen et al., 2019

Idea comes from the Focal Loss Paper



Focal Loss for Dense Object Detection, Lin et al., 2018

Transfer learning on CheXNet has little effect on performance



Transfusion: Understanding Transfer Learning with Applications to Medical Imaging, Raghu et al., 2019

Pre-Training is an effective transfer learning Strategy

Strategy	**AUC (95% CI)**
Performance on TB task **with CheXpert pre-training**	0.831 (0.751, 0.910)
Performance on TB task **without CheXpert pre-training**	0.714 (0.618, 0.810)

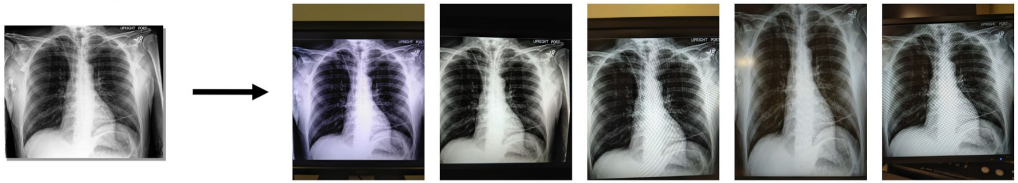
Rajpurkar & O'Connell & Schechter et al., *in submission*

Deployment Opportunities

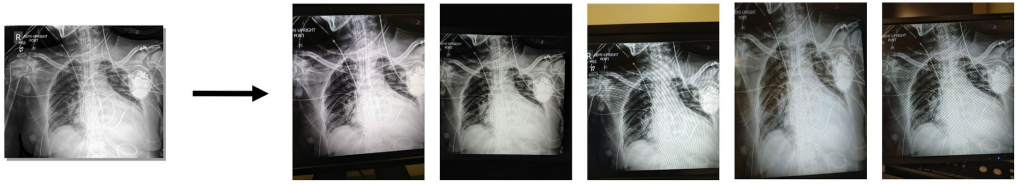
- *ML Generalizability and Clinical Impact Studies*

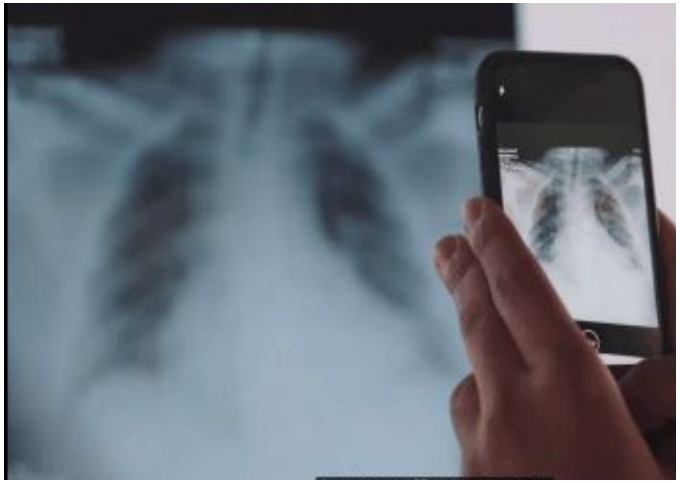
Generalizability to Photo Deployment

patient00049/study1/view1



patient00410/study1/view1





Black Box Assistant

Unassisted

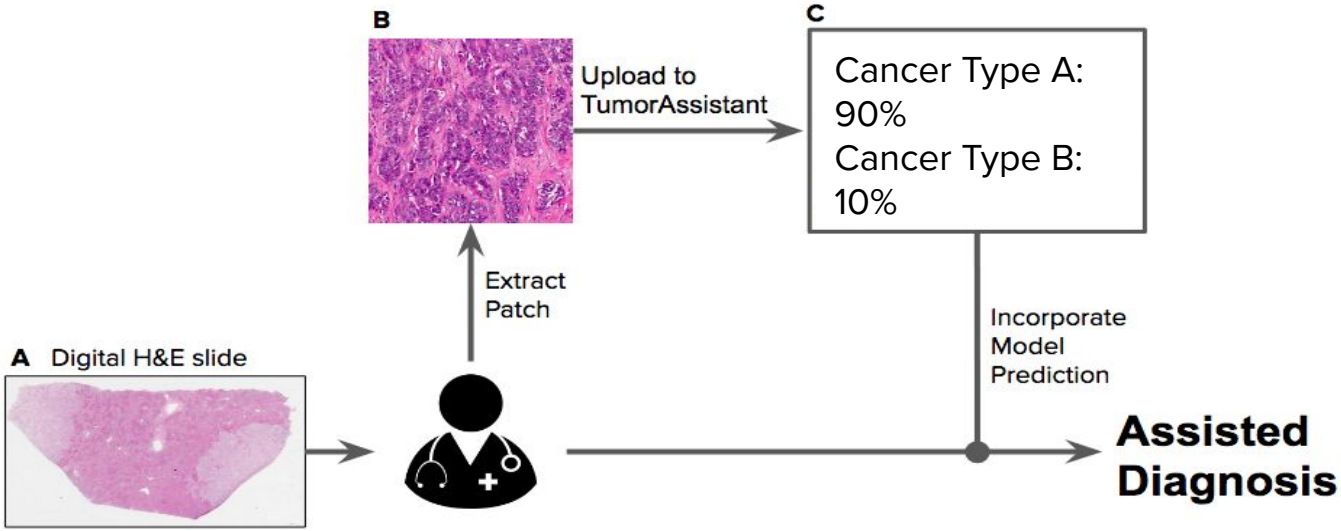
Assisted

DL: 0.8 Abnormal, 0.8 ACL tear, 0.1 MCL tear

Probability of Abnormality, ACL tear and MCL tear.

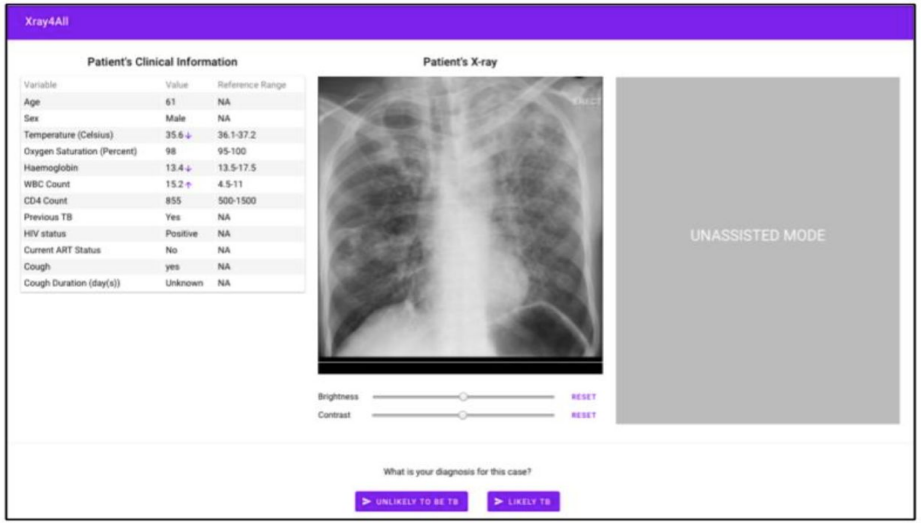
Bien & Rajpurkar et al, PLOS Medicine, 2018

Interactive AI Assistant

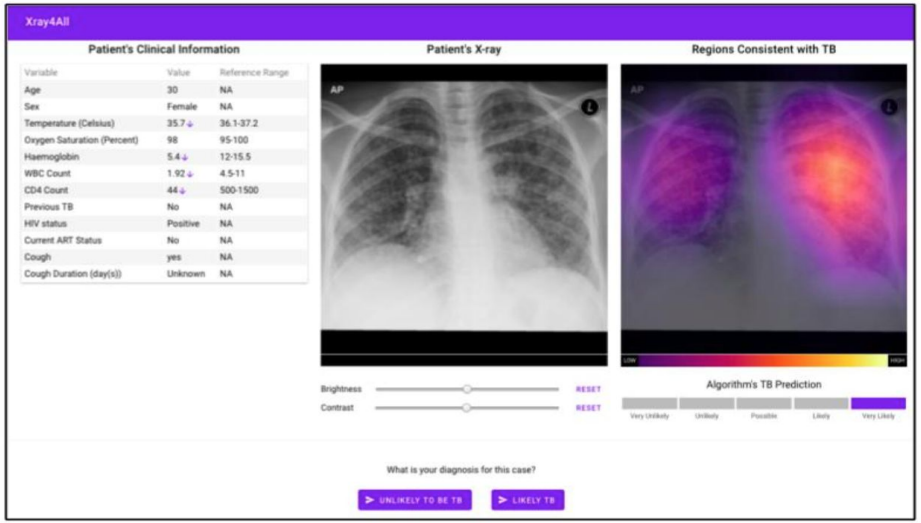


Rajpurkar & Uyumazturk & Kiani et al, *2019*

Incorporation of Clinical Variables in Field Simulation



Variable	Value	Reference Range
Age	61	NA
Sex	Male	NA
Temperature (Celsius)	35.6 ↓	36.1-37.2
Oxygen Saturation (Percent)	98	95-100
Haemoglobin	13.4 ↓	13.5-17.5
WBC Count	15.2 ↑	4.5-11
CD4 Count	855	500-1500
Previous TB	Yes	NA
HIV status	Positive	NA
Current ART Status	No	NA
Cough	yes	NA
Cough Duration (day(s))	Unknown	NA



Variable	Value	Reference Range
Age	30	NA
Sex	Female	NA
Temperature (Celsius)	35.7 ↓	36.1-37.2
Oxygen Saturation (Percent)	98	95-100
Haemoglobin	5.4 ↓	12-15.5
WBC Count	1.92 ↓	4.5-11
CD4 Count	44 ↓	500-1500
Previous TB	No	NA
HIV status	Positive	NA
Current ART Status	No	NA
Cough	yes	NA
Cough Duration (day(s))	Unknown	NA

Rajpurkar & O'Connell & Schechter et al., *in submission*

Opportunities for AI + Medicine

Community

Open invite to the world to participate in competitions w/ hidden test set

CheXpert

2400 Users
80+ Teams Competing

MRnet DATASET

1213 Users

MURA

3600 Users
70+ Teams Competing

Irvin & Rajpurkar et al. 2019, AAAI; Bien & Rajpurkar et al, PLOS Medicine; Rajpurkar & Irvin et al. 2018, MIDL

Weekly Newsletter on latest AI For Healthcare Research

1000+ Regular Weekly Readers

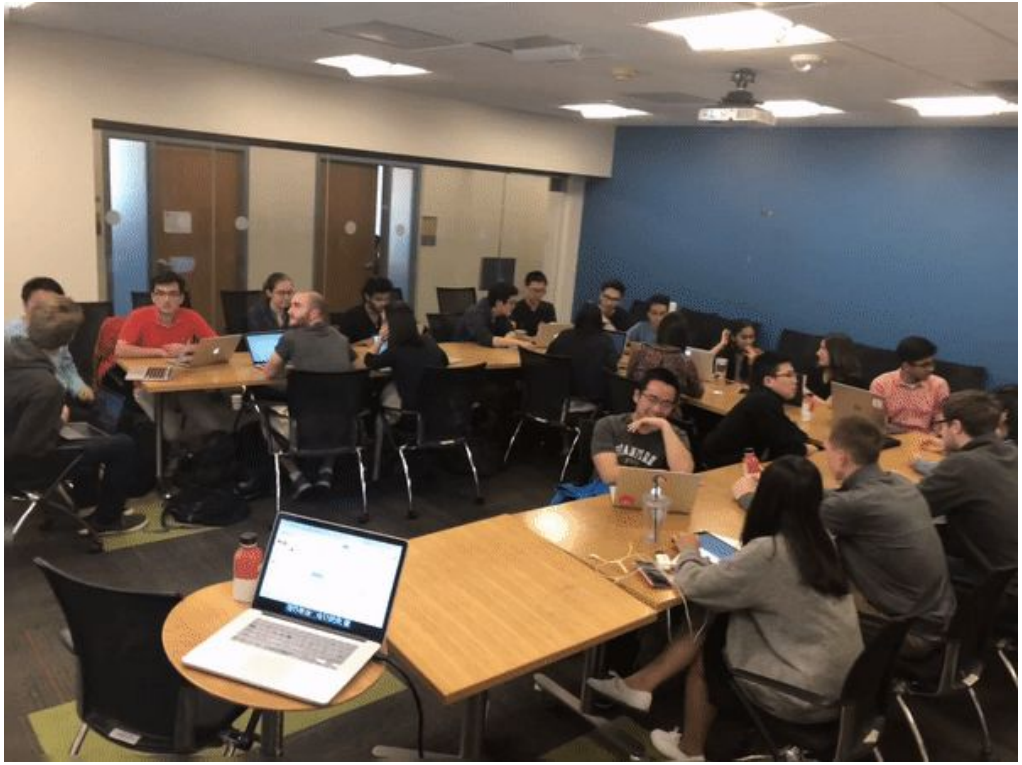
http://doctorpenguin.com

AI For Healthcare Bootcamp

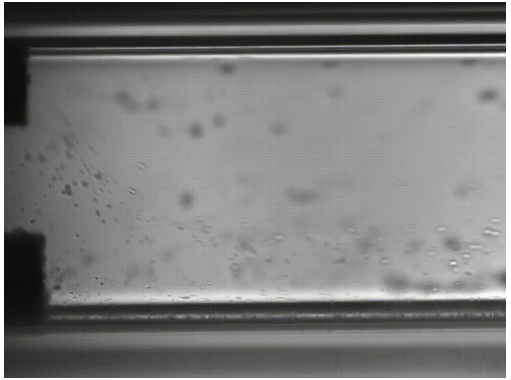
AI For Healthcare Bootcamp

2-quarter program that provides Stanford students an opportunity to do cutting-edge research at the intersection of AI and healthcare.

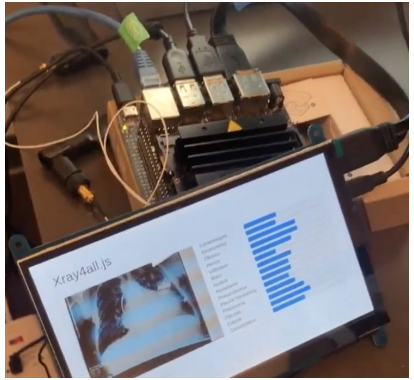
AI For Healthcare Bootcamp



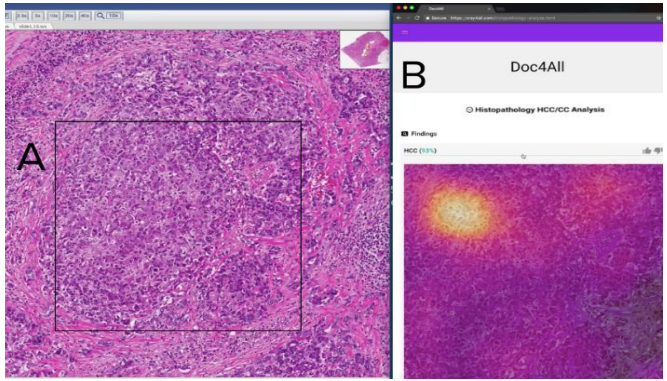
Training from PhD students and faculty in the medical school to work on high-impact research in small interdisciplinary teams.



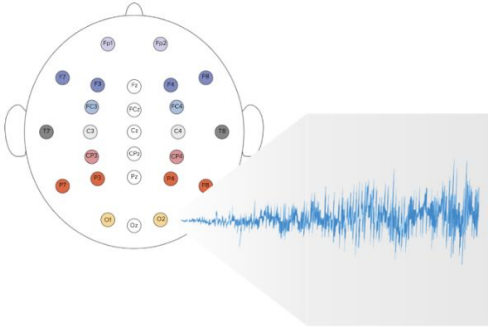
Automated Cell Sorting



Embedded Deployment



Histopathology Assistance



Depression Outcome Prediction

https://stanfordmlgroup.github.io/programs/aihc-bootcamp/

or Google “AI For Healthcare Bootcamp Stanford”