MICCAI 2022 Singapore

25th International Conference on Medical Image Computing and Computer Assisted Intervention

September 18–22, 2022

Resorts World Convention Centre Singapore

ÉTS Le génie pour l'industrie

Erasmus MC

Erasmus MC University Medical Center Rotterdam

UC SANTA CRUZ

Learning with Limited Supervision

Yuyin Zhou (Yan Wang)
Ismail Ben Ayed
Jose Dolz
Christian Desrosiers
Marleen de Bruijne
Hoel Kervadec

Erasmus MC
University Medical Center Rotterdam

UNIVERSITY OF COPENHAGEN

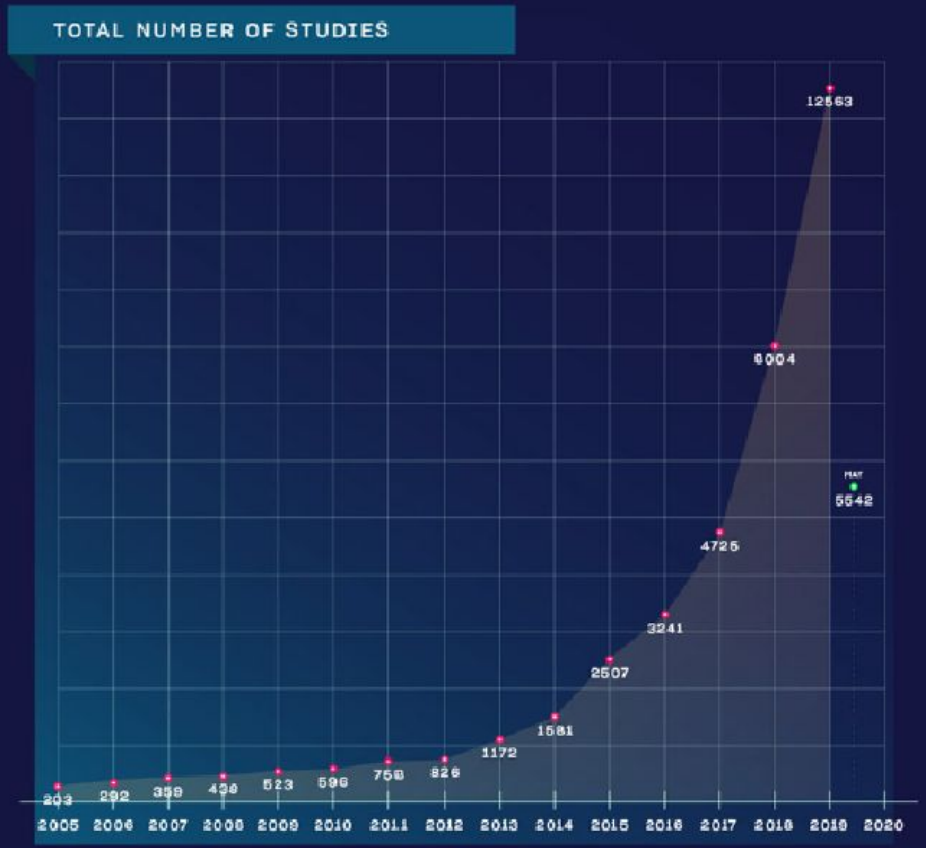
Learning with Limited Supervision: Applications in Radiology

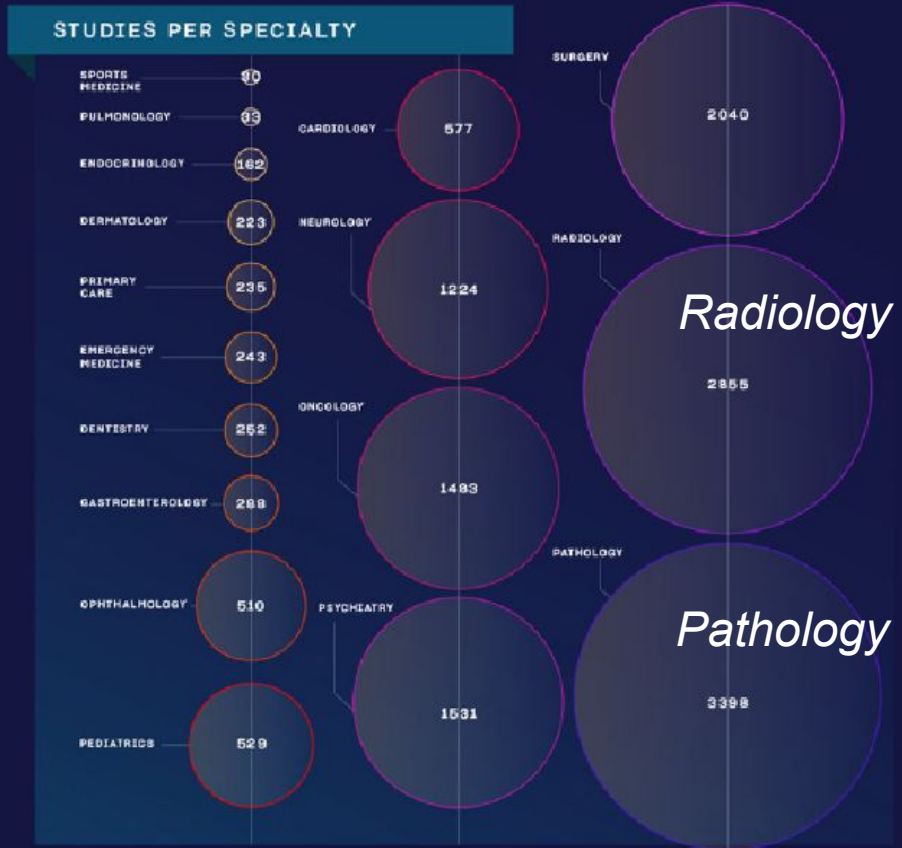
Marleen de Bruijne

a

b

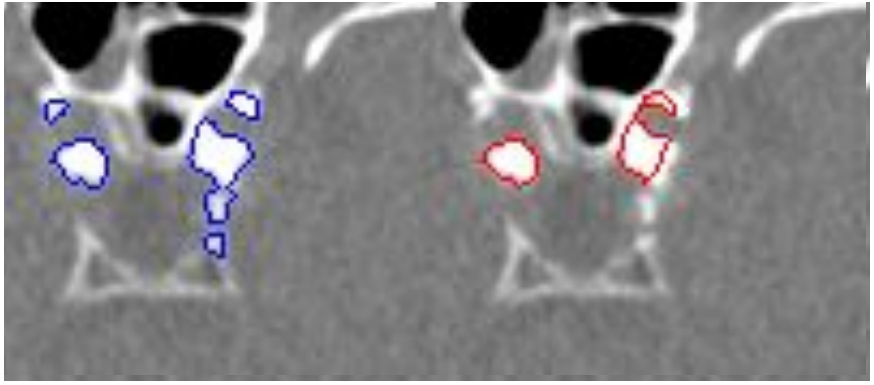
MACHINE AND DEEP LEARNING STUDIES ON PUBMED.COM

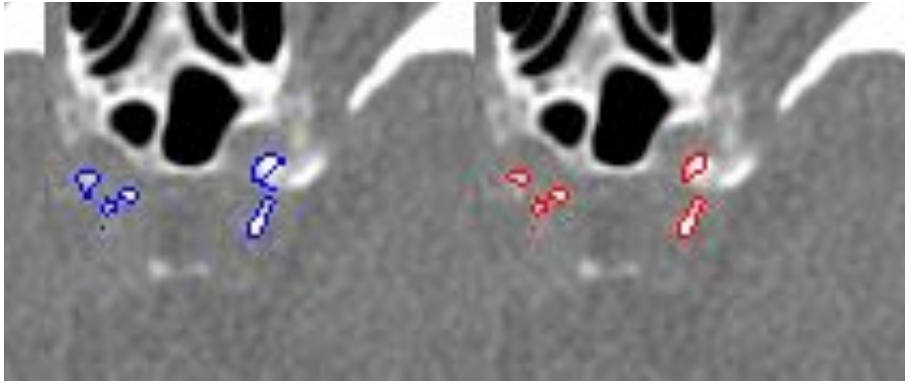




Meskó and Görög, NPJ Digital Medicine 2020

Deep networks can solve many challenging medical imaging tasks as good as humans can



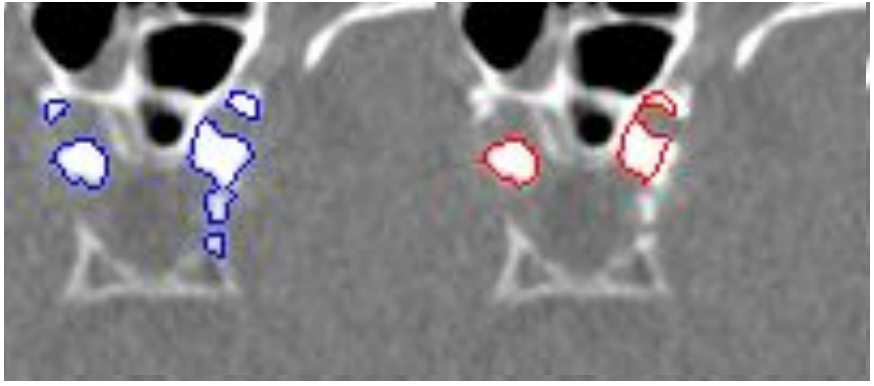


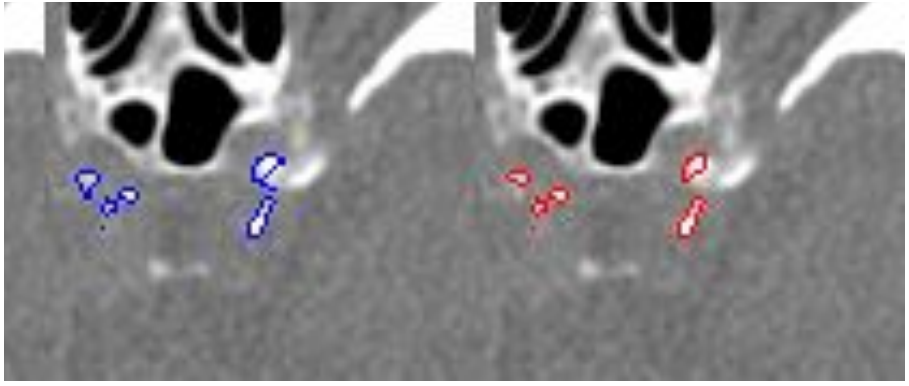
Manual Automatic

- Segmentation of intracranial calcifications in CT
- Agreement manual/automated scores ICC 0.98
- On average, automated segmentation slightly preferred

Bortsova et al 2021

Deep networks can solve many challenging medical imaging tasks as good as humans can





Manual Automatic

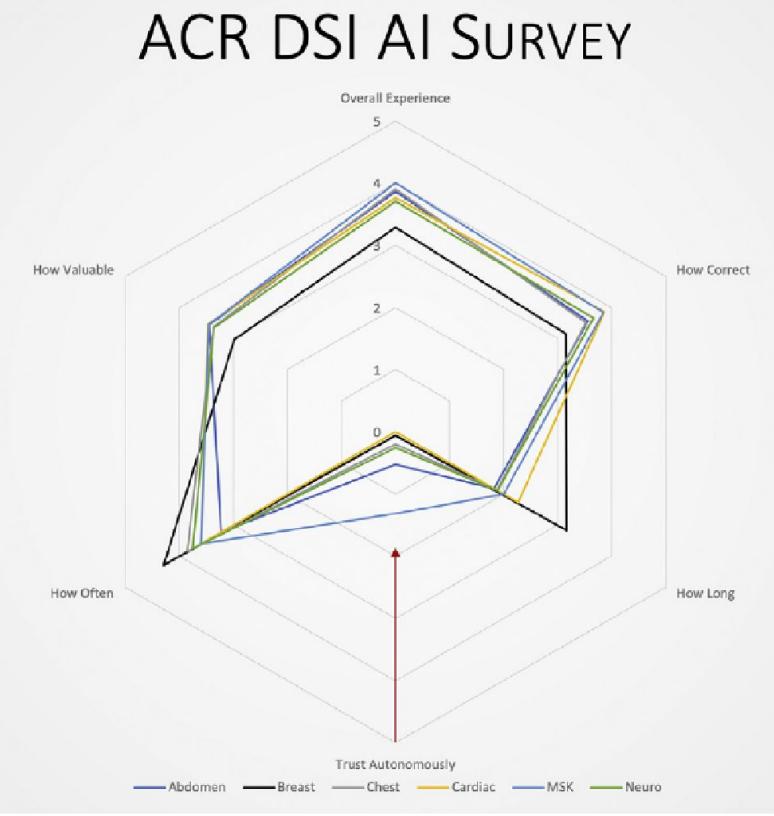
- This model used a large training set (1000)
- Good results, but worse performance for more rare cases
- Increasing the training set still improved performance

Bortsova et al 2021

Erasmus MC

MEDICAL IMAGING IS A SMALL SAMPLE SIZE DOMAIN

What do clinicians think about AI?



- 2021 ACR survey of AI use among radiologists
- 30% use AI in some form
- 20% plan to purchase AI in the next 5 years
- Overall experience is positive
- Good accuracy, valuable
- Performance inconsistent
- Very low trust in autonomous AI

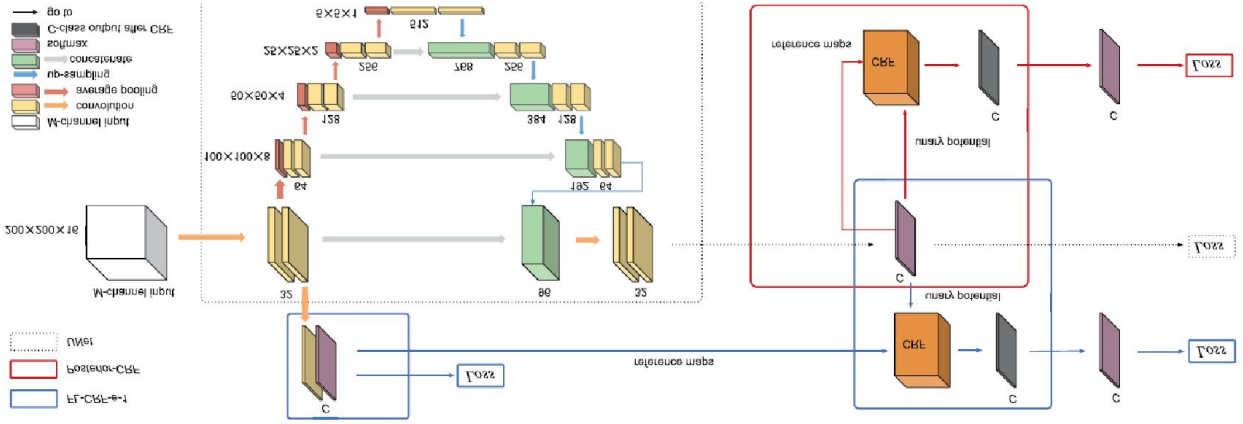
Allen et al 2021

Erasmus MC

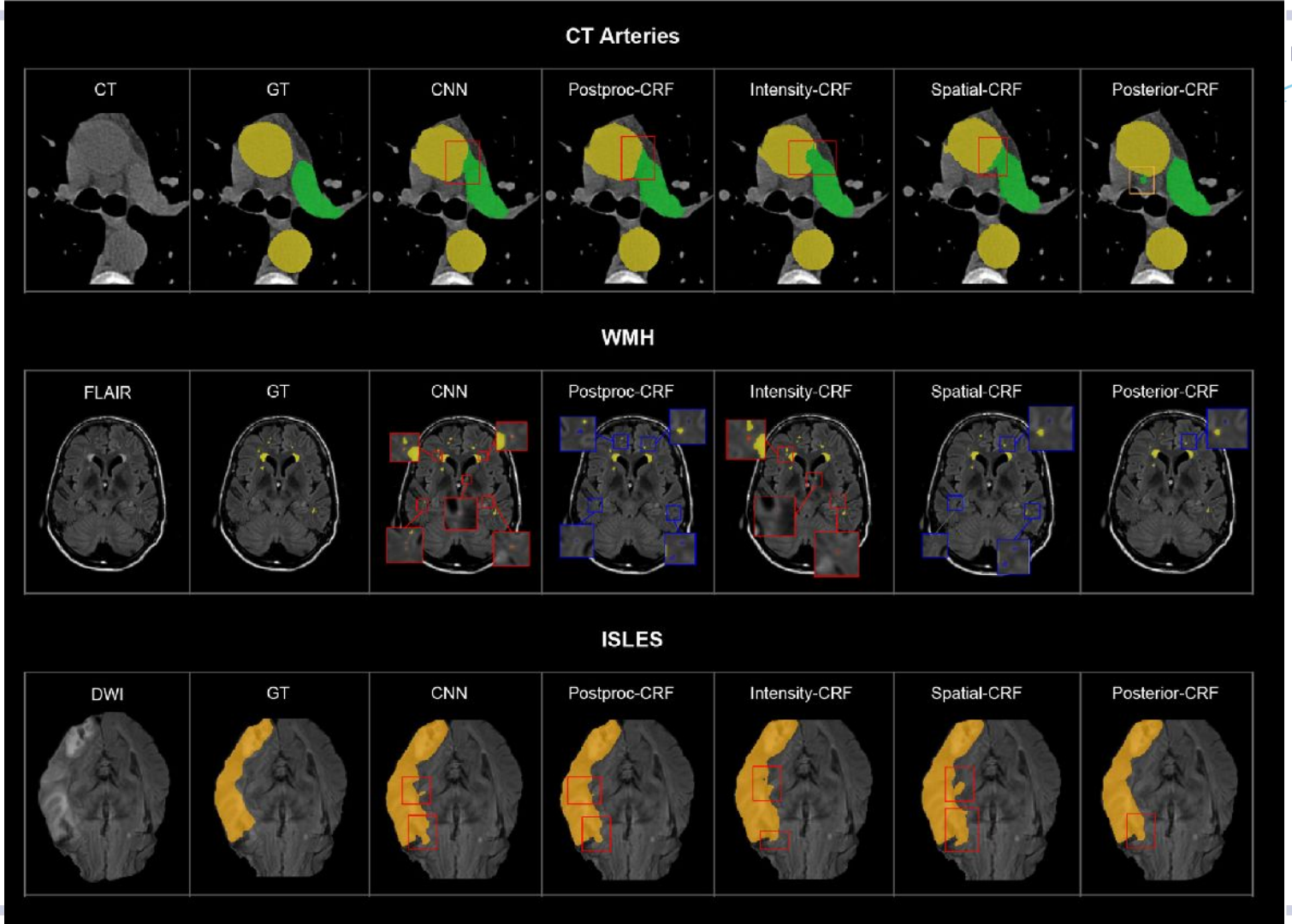
HOW CAN WE LEARN RELIABLE MODELS WITH LESS DATA, OR LESS ANNOTATIONS?

Smoothness and label consistency

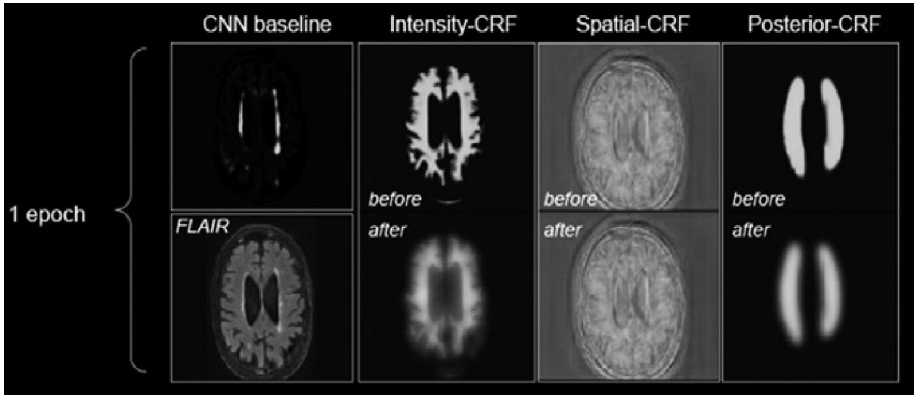
- CRF: neighboring voxels and/or voxels with similar appearance are likely to have the same label
- Optimize together with CNN (Zheng et al 2015, “CRF as RNN”)
- But also use learned features in pairwise potentials
- Significantly improves accuracy

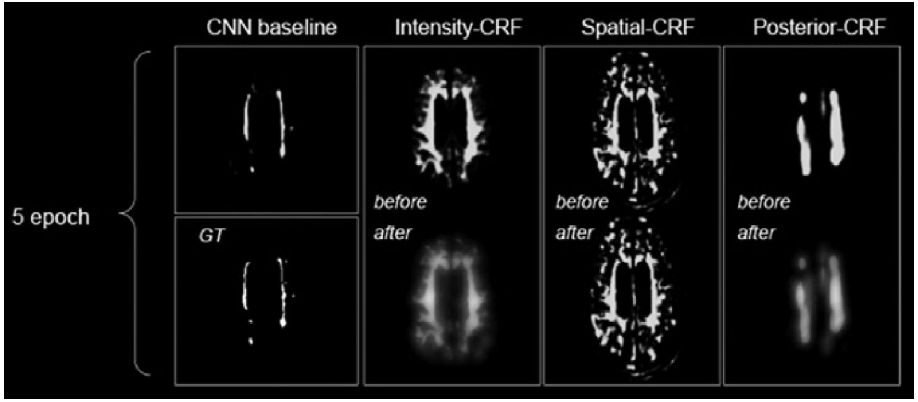


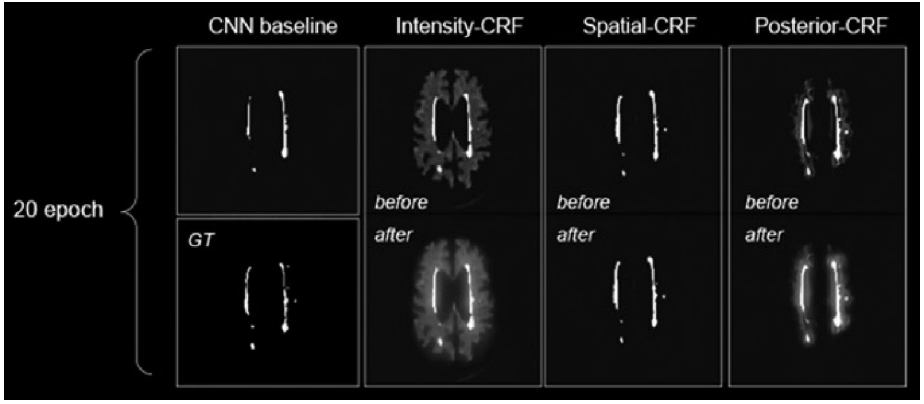
Chen et al, MEDIA 2021

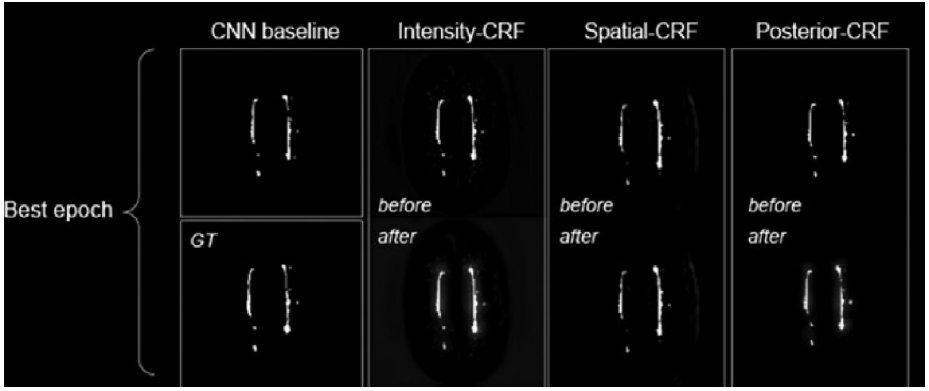

CT Arteries
CT
GT
CNN
Postproc-CRF
Intensity-CRF
Spatial-CRF
Posterior-CRF
WMH
FLAIR
GT
CNN
Postproc-CRF
Intensity-CRF
Spatial-CRF
Posterior-CRF
ISLES
DWI
GT
CNN
Postproc-CRF
Intensity-CRF
Spatial-CRF
Posterior-CRF

Evolution of CNN and CRF output during training





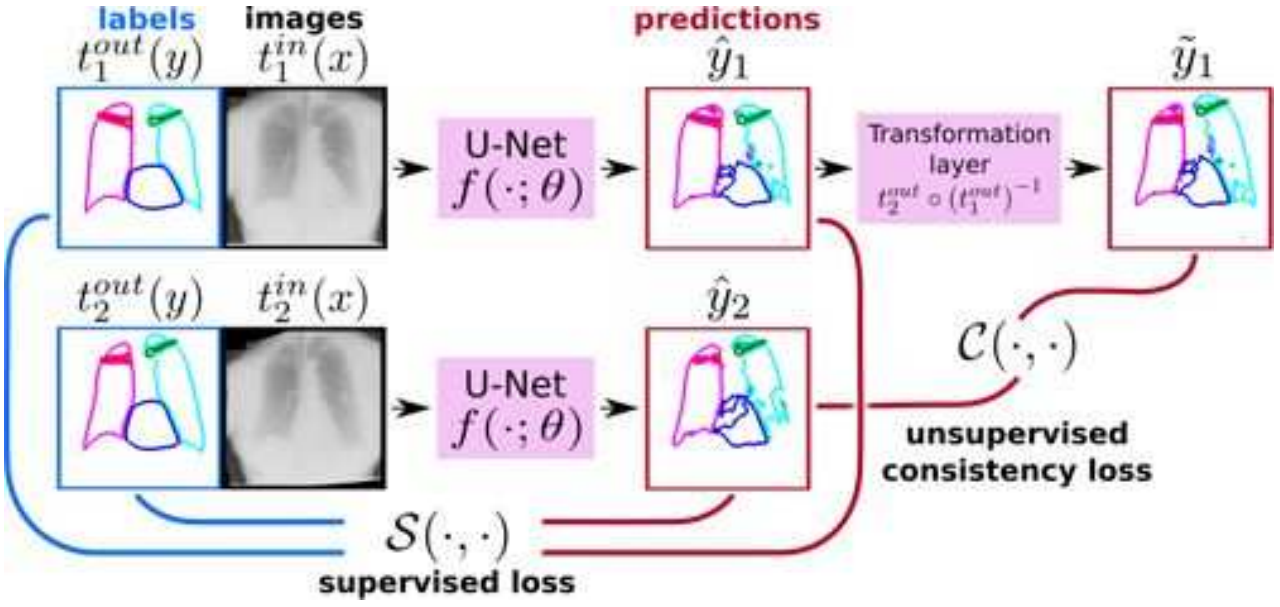




Label consistency under image deformation

Estimated segmentations may be wrong, but should be consistent under elastic deformations of the input

Useful in semi-supervised learning



Bortsova et al, MICCAI2019

Label consistency loss improves segmentation

Erasmus MC

Segmentation overlap for chest Xray segmentation:

Loss	$\mathcal{X}_u$	5	10	25	50	100
$\mathcal{L}_{sup}^{\mathcal{T}}$	$\varnothing$	74.2 ± 3.8	82.8 ± 1.3	87.5 ± 0.4	89.0 ± 0.3	90.6 ± 0.2
$\mathcal{L}_{sup}^{\mathcal{T}} + \mathcal{L}_{cons}^{\mathcal{T}}$	$\varnothing$	76.4 ± 3.8	83.6 ± 1.4	87.8 ± 0.4	89.5 ± 0.2	90.9 ± 0.3

- Only labeled data: Improves performance when training set is small
- Labeled+unlabeled data: Much larger improvements, especially when training set is small

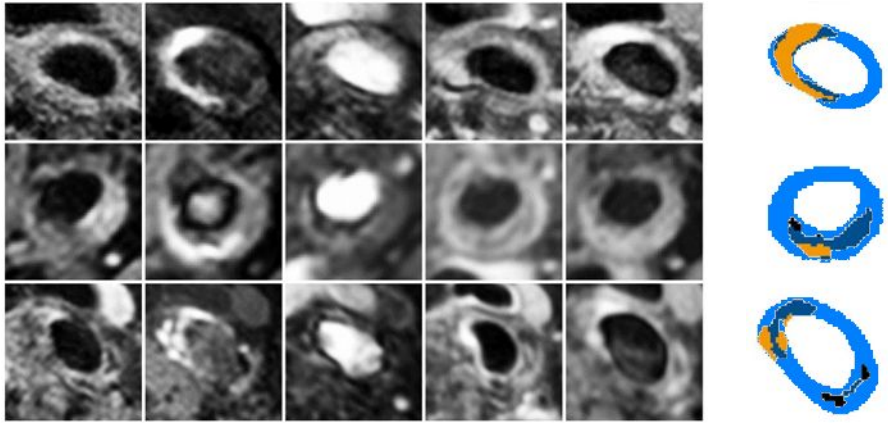
Bortsova et al, MICCAI2019

Erasmus MC

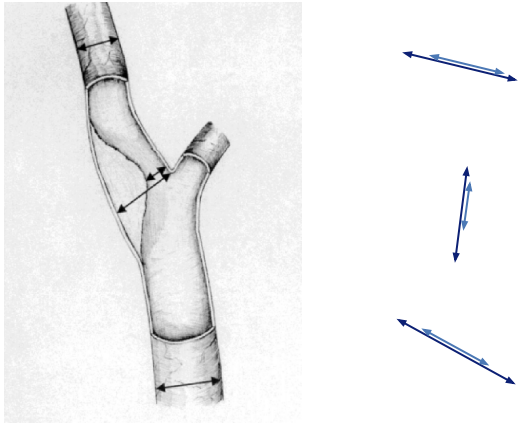
LEARNING FROM WEAK LABELS

EXAMPLE: ASSESSING THE RISK OF STROKE

Segment carotid artery

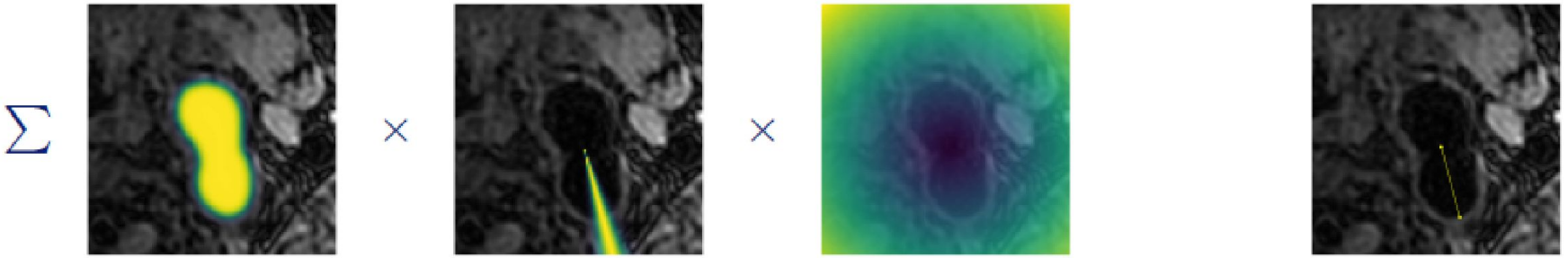


Annotations for model training



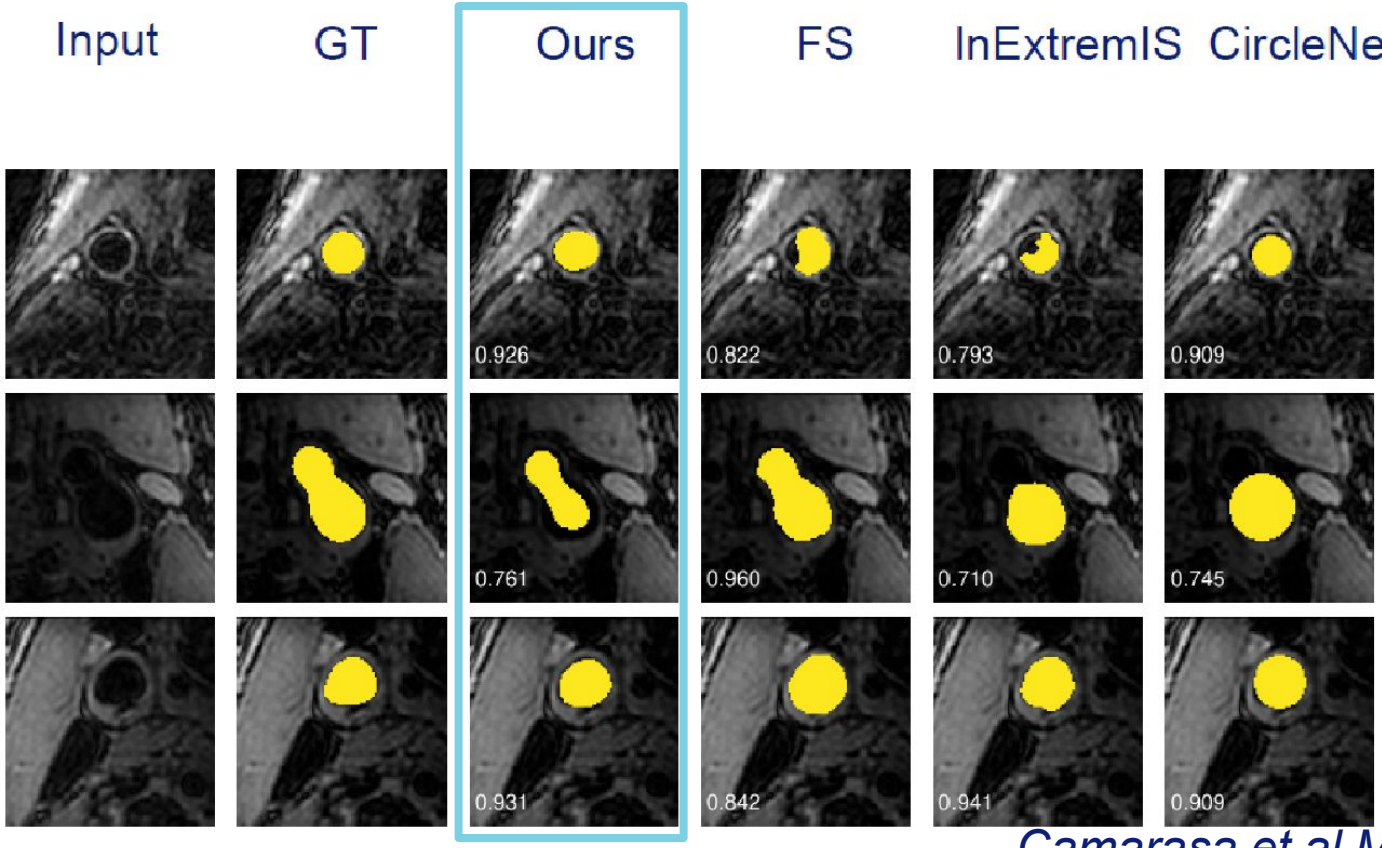
Clinical annotations

Can we train model based on diameters?



Camarasa et al MIDL 2022

Good segmentations, close to full supervision



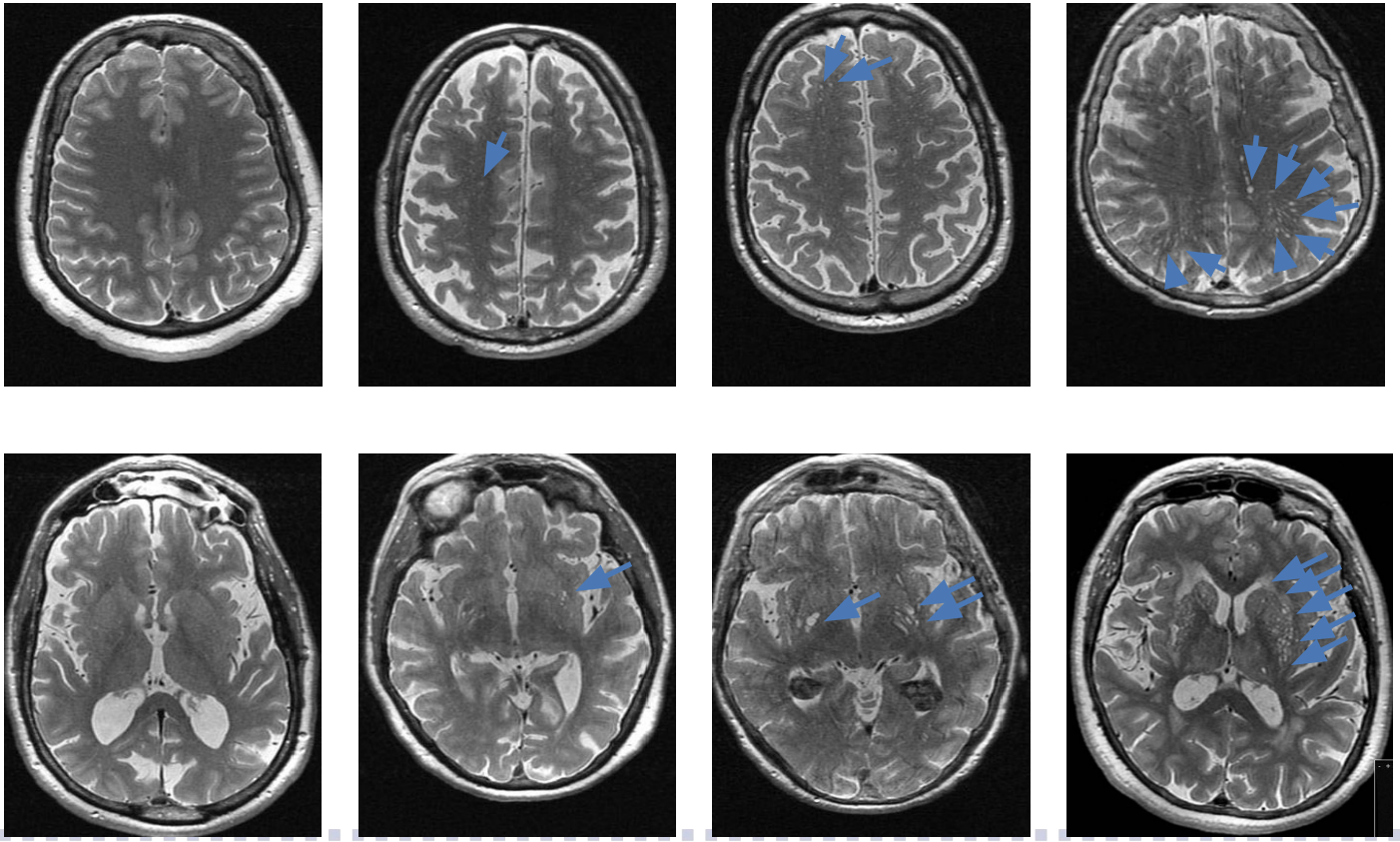
Camarasa et al MIDL 2022

EXAMPLE: LEARNING IMAGING BIOMARKERS OF DEMENTIA

Quantifying enlarged perivascular spaces (PVS) -- A marker of small vessel disease (vascular dementia)

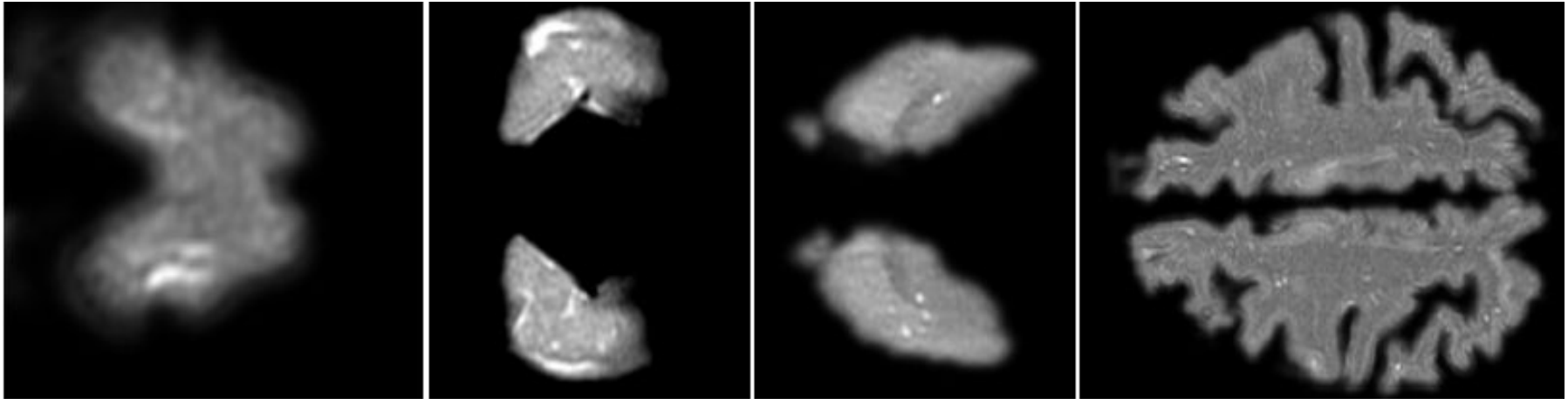
Erasmus MC

No lesions



Many lesions

Current assessment: visual scoring



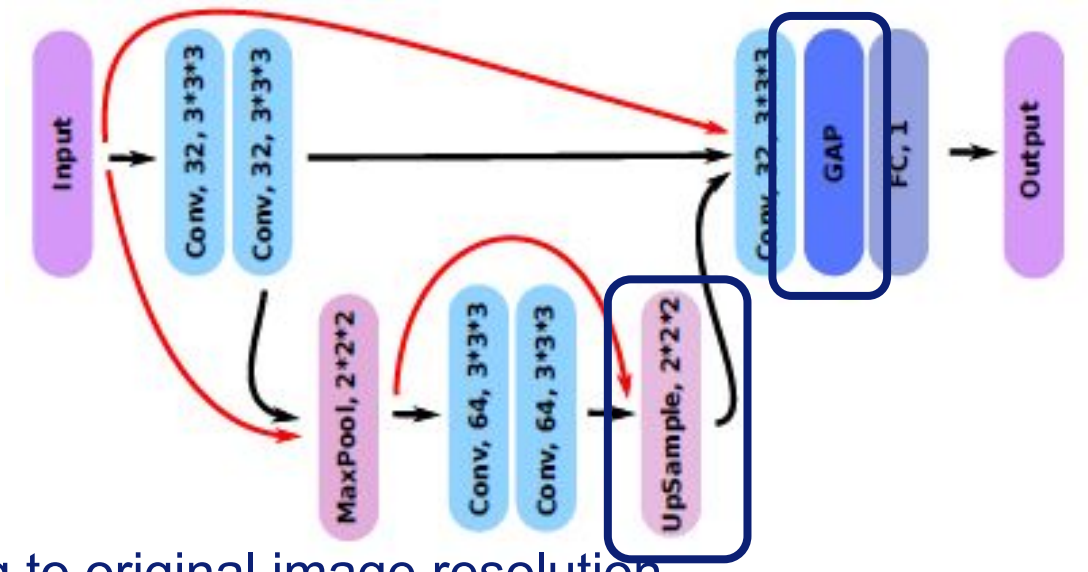
midbrain *hippocampus* *basal ganglia* *white matter*

- Count number of lesions in certain regions in certain slices

- Automated quantification: 3D CNN regression to predict this number

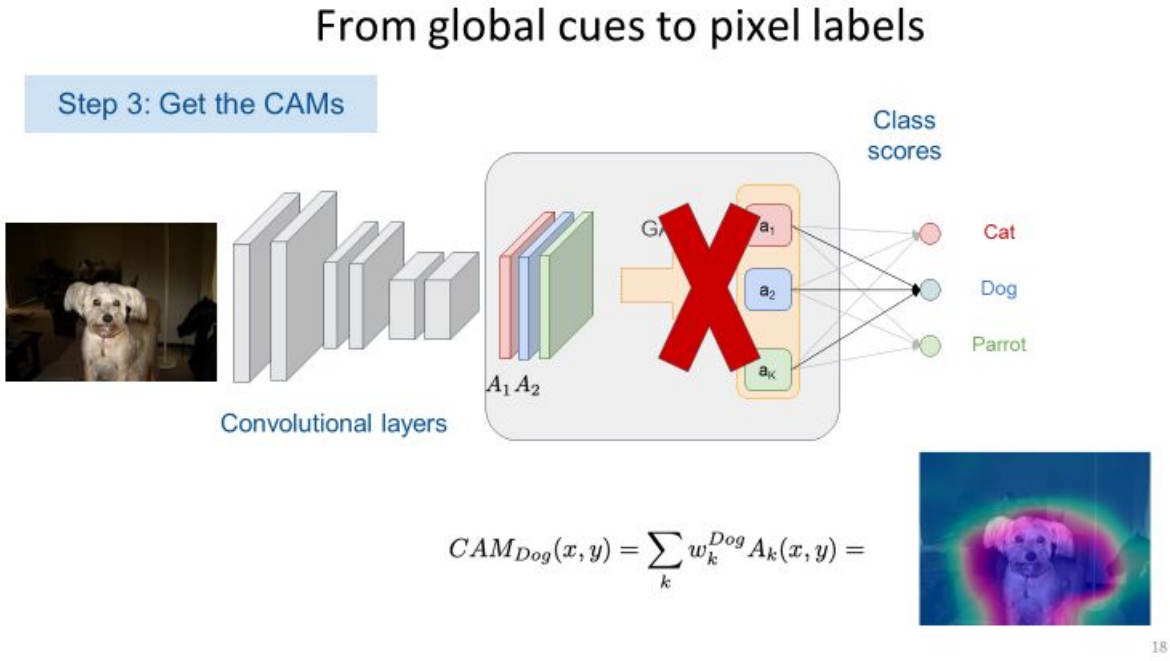
Inspecting network attention

- Use slightly different network architecture:



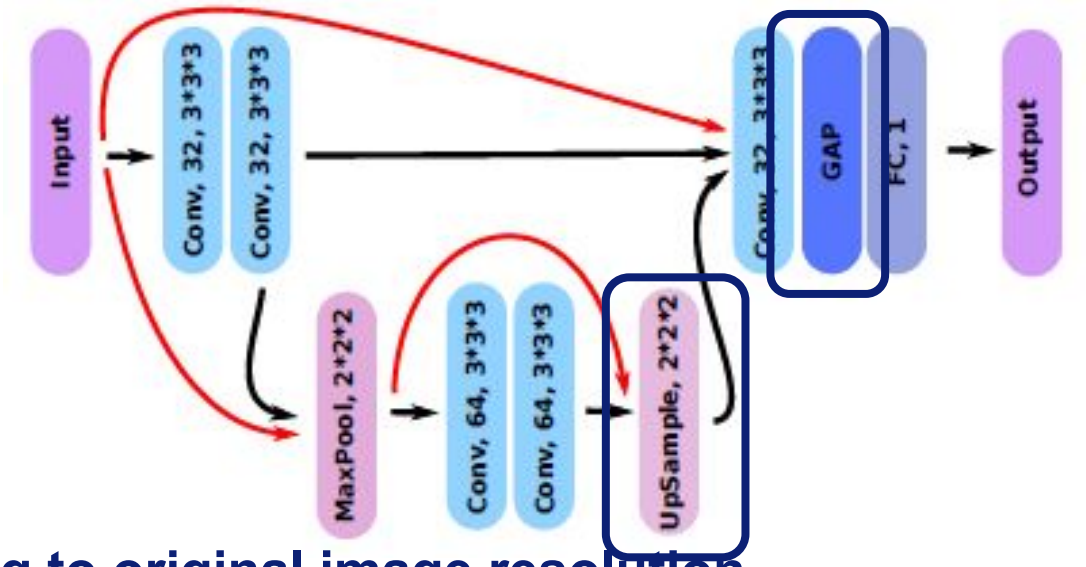
- Upsampling to original image resolution
- With global (average) pooling → prediction
- Without global pooling → attention maps

Remember the Class Activation Maps (CAM) (Dolz lecture)



Inspecting network attention

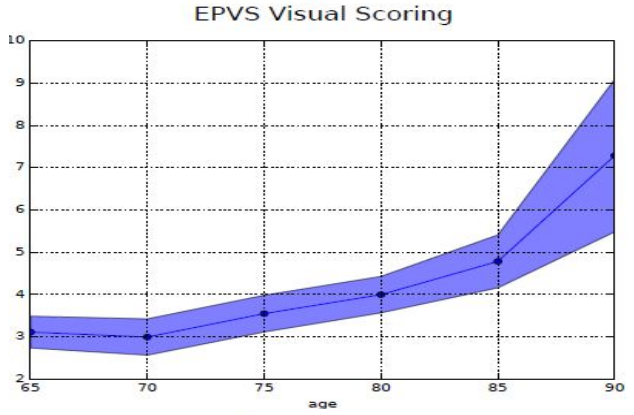
- Use slightly different network architecture:

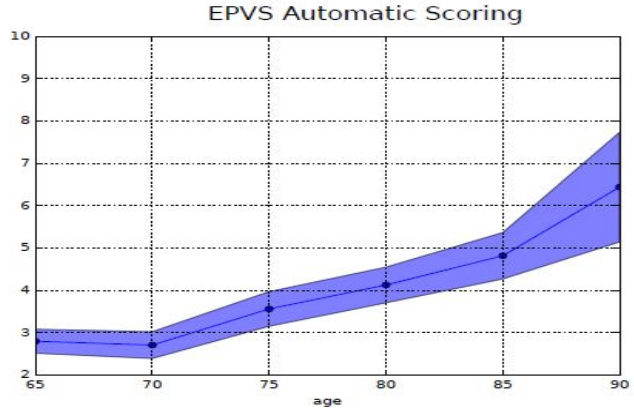


- **Upsampling to original image resolution**
- With global (average) pooling → prediction
- Without global pooling → attention maps

Good agreement with expert scores

Region	Inter-observer Agreement	Trained on 1600 scans
Midbrain	0.75	0.75
Hippocampi	0.82	0.88
Basal Ganglia	0.62	0.82
Centrum Semiovale	0.80	0.86



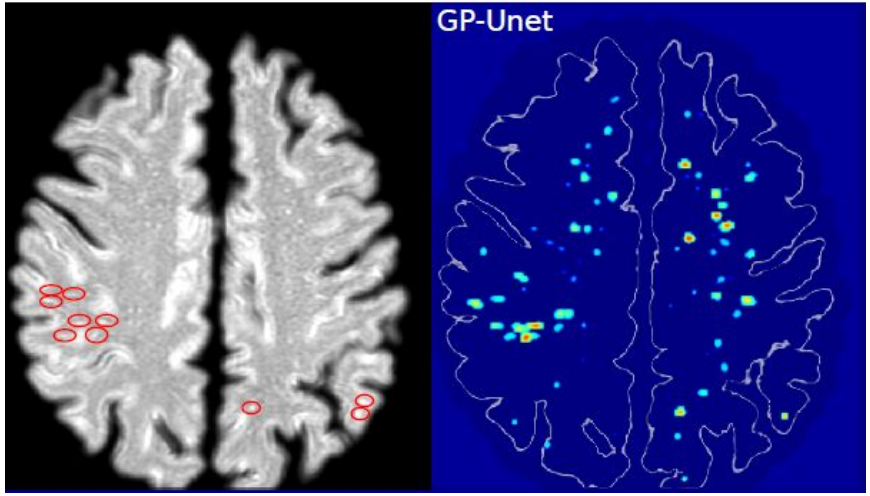


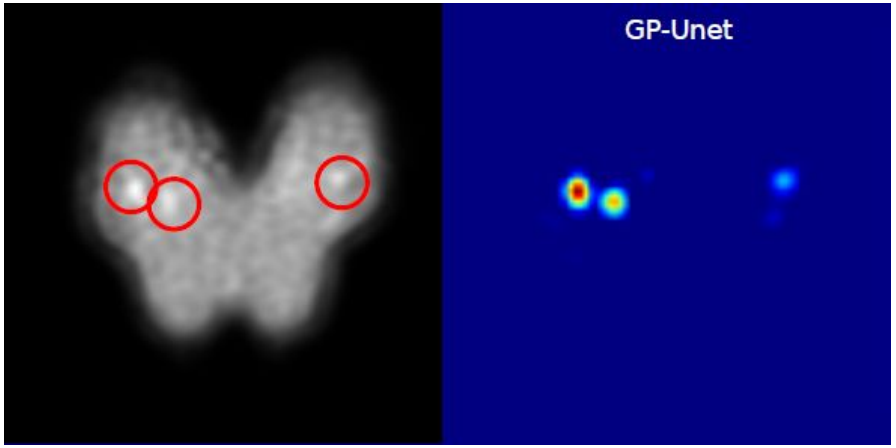
F. Dubost et al, Neuroimage 2019, MEDIA 2020

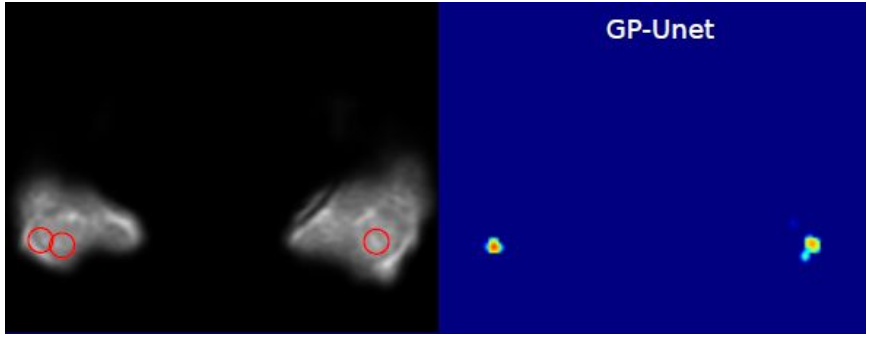
What has the network learned?

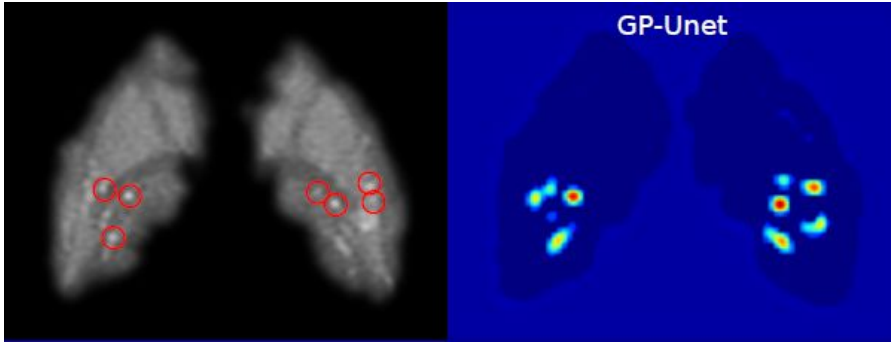
- Something that correlates very well with visual scoring

Attention maps reveal focus on PVS







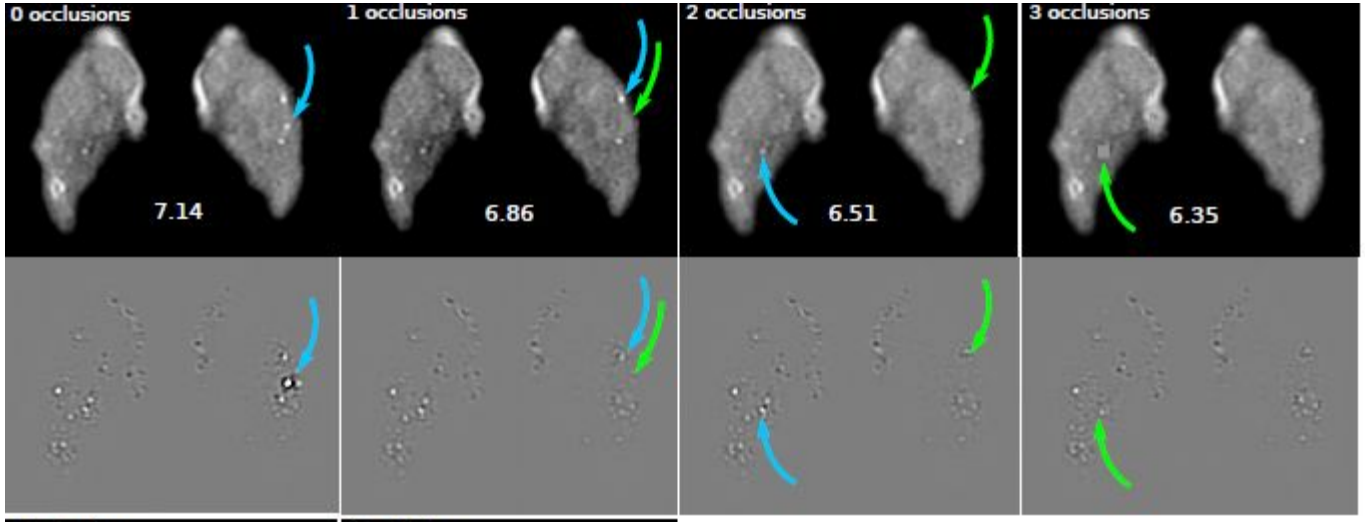


What has the network learned?

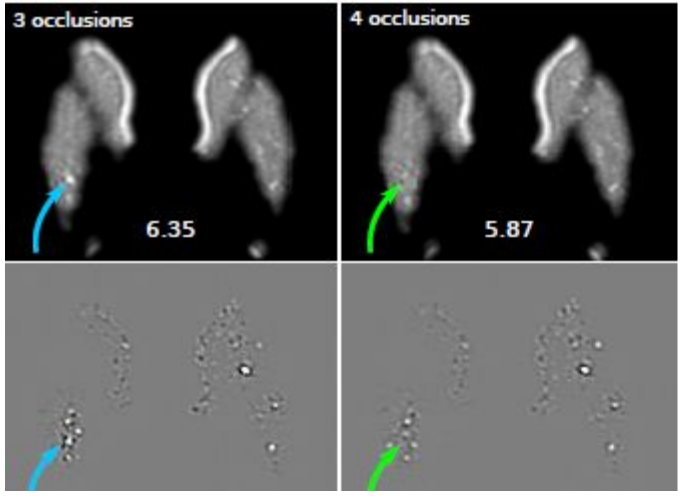
- Something that correlates very well with visual scoring

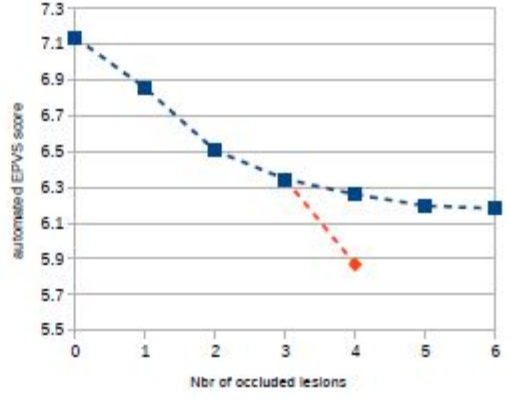
- Seems to focus on the lesions

Annotated slice



Other slice



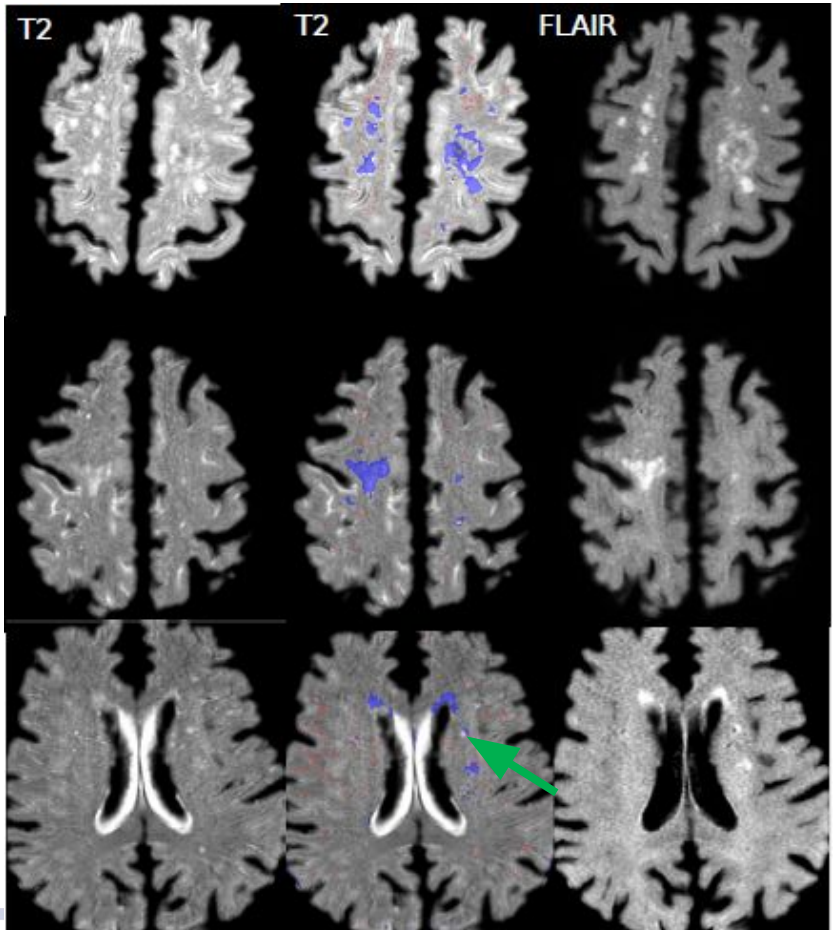


What has the network learned?

- Something that correlates very well with visual scoring

- Seems to focus on the lesions
- Whole region information instead of single slice

Can discriminate between types of lesions



Very little overlap between EPVS (red) and other types of lesions - white matter hyperintensities (blue) or lacune (green arrow)

What has the network learned?

Erasmus MC

- Something that correlates very well with visual scoring

- Seems to focus on the lesions
- Whole region information instead of single slice
- Can discriminate between types of lesions

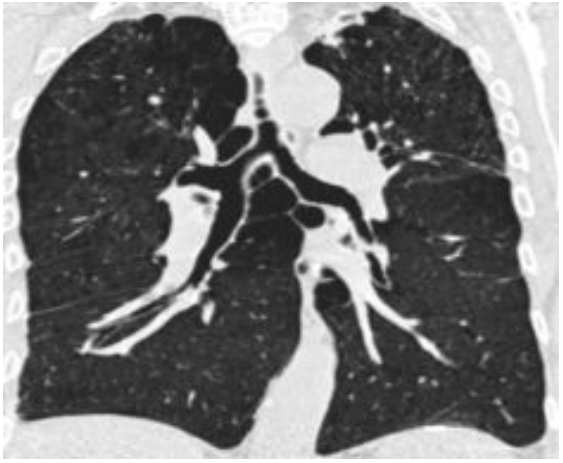
- → Could replace visual assessment in large studies
- Potential to improve on visual assessment by using full 3D information and quantifying shape, volume of lesions from activation maps

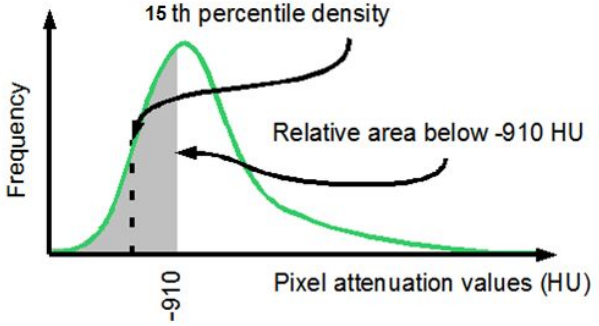
EXAMPLE: LEARNING IMAGING BIOMARKERS OF COPD

Chronic Obstructive Pulmonary Disease (COPD)

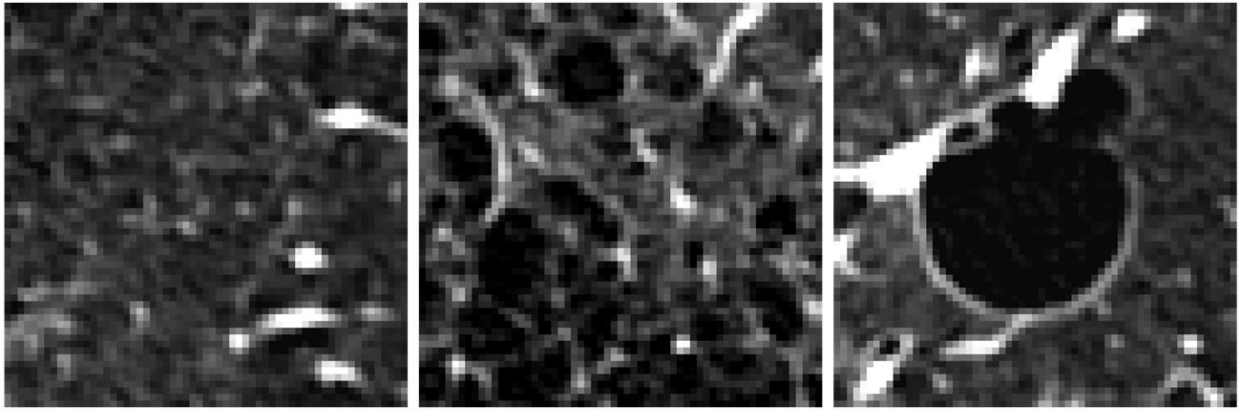
- Major cause of death worldwide
- Early stages severely underdiagnosed
- Poorly understood
- Different phenotypes may require different treatment

Quantifying COPD and emphysema



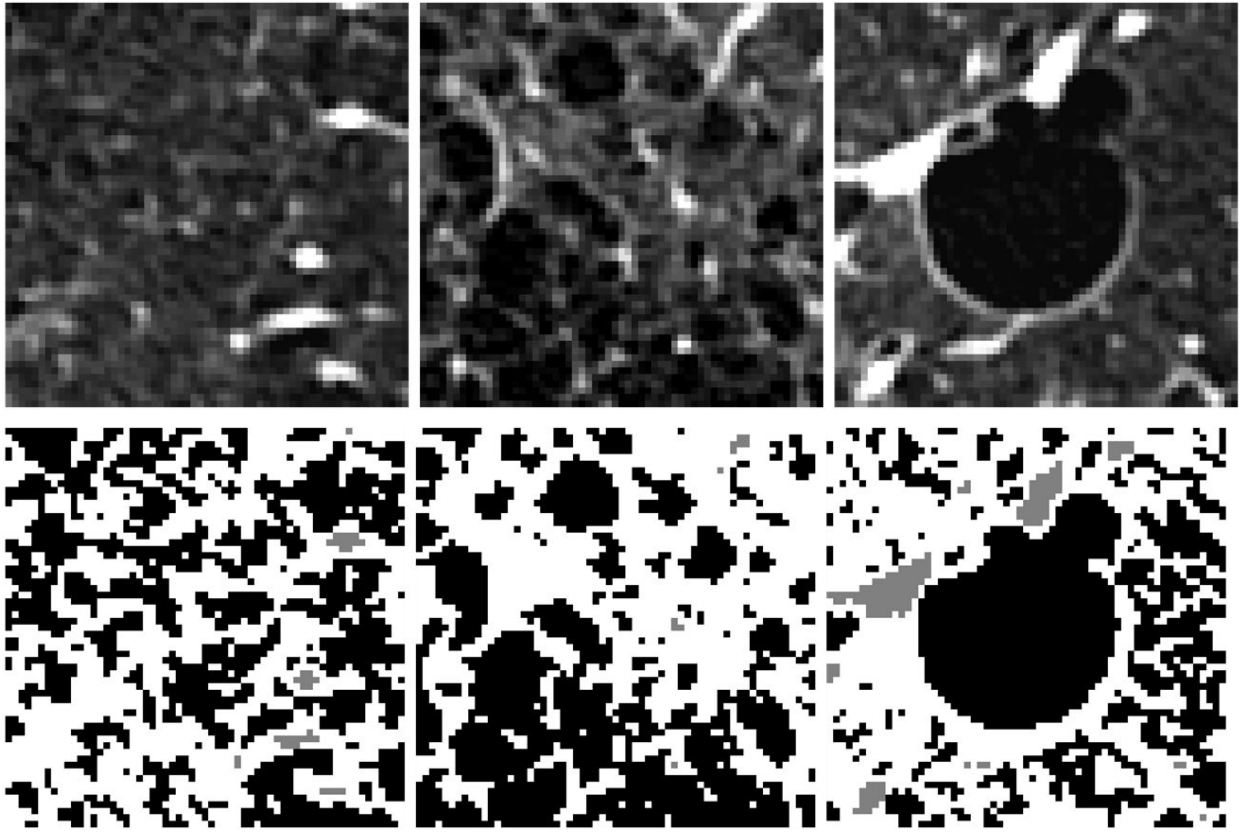


Beyond density: can we do better?

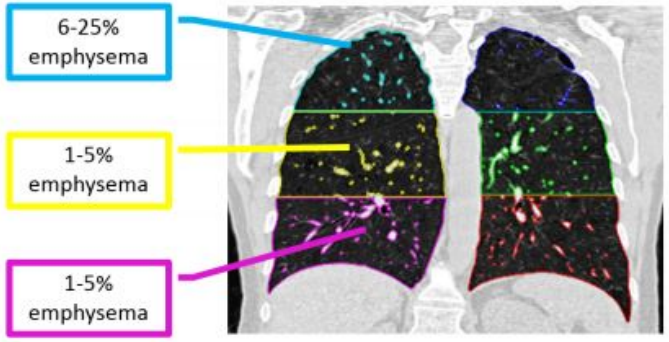


Emphysema has various disease subtypes, with different patterns in CT

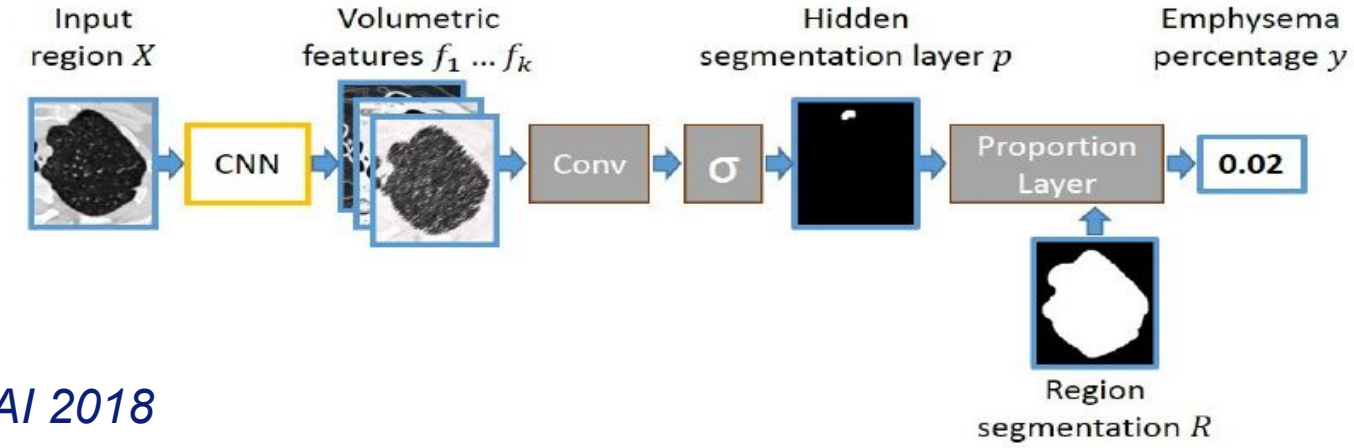
Different patterns may have identical densitometry values…



Learning from visual scoring



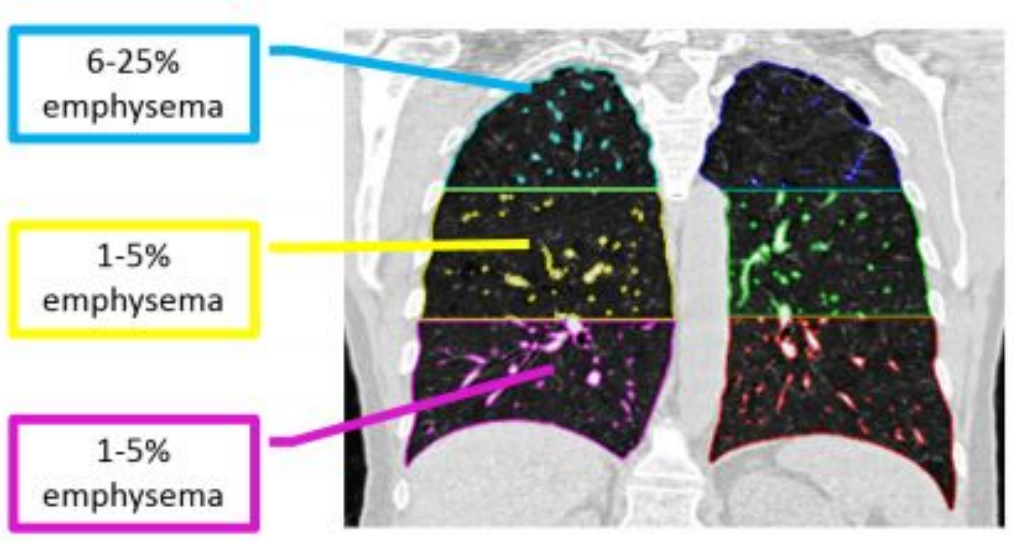
- Train a network to reproduce these scores
- Via an intermediate (segmentation) image with the right proportions



Bortsova et al MICCAI 2018

The scoring system

- Categories:
  - 0: 0% (healthy)
  - 1: <5%
  - 2: 6-25%
  - 3: 26-50%
  - 4: 51-75%
  - 5: 76-100%

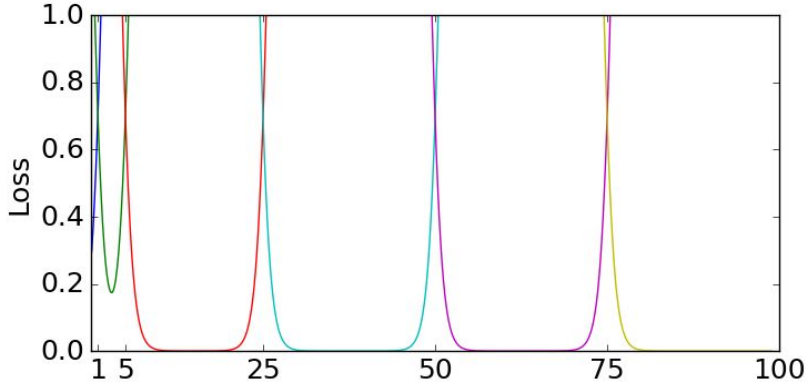


Loss function for Learning from Label Proportions

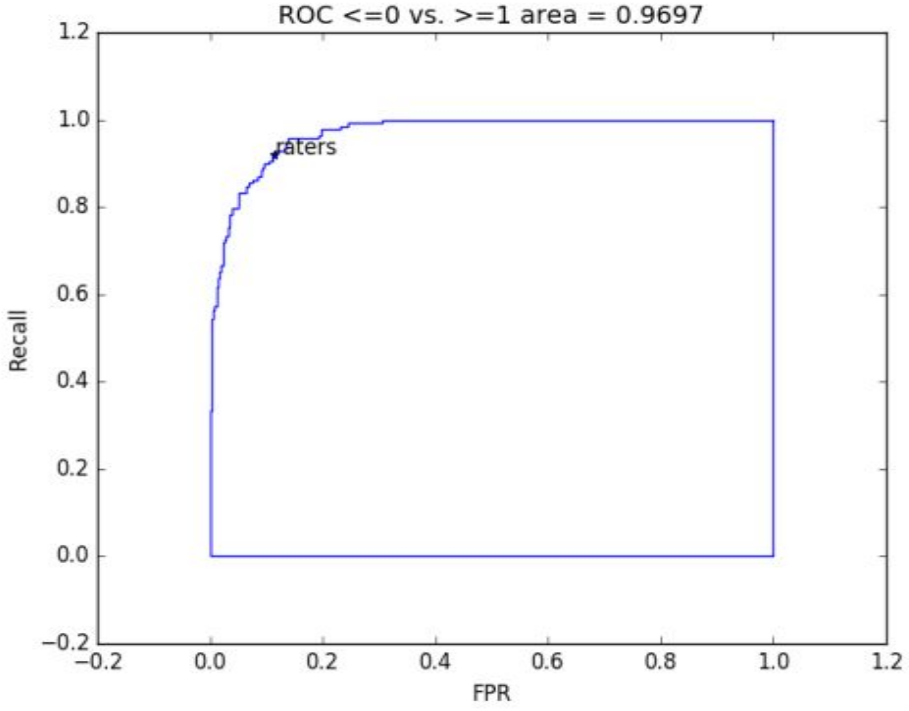
- weighted Sum of five cross-entropy losses:

$$\mathcal{L}(y, \hat{y}) = \sum_{c=1}^{5} w_c \text{CrossEntropy}(\sigma^*(y - \hat{y}^{min}), \mathbb{I}(\hat{y}^{cat} \geq c))$$

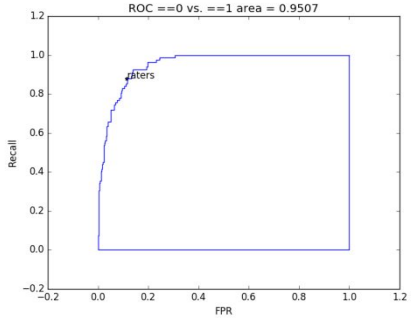
- Loss for examples of different grades:



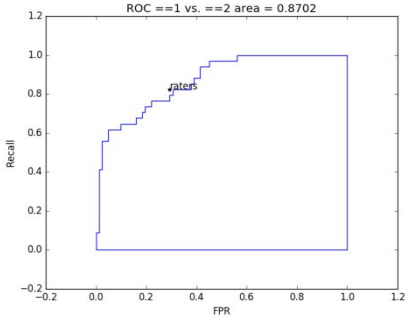
Performs similar to trained observers



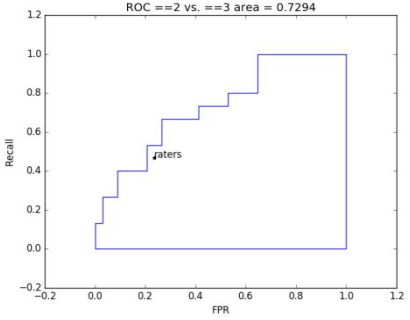
Emphysema detection



Healthy vs. mild (grade 1)



Grade 1 vs. 2

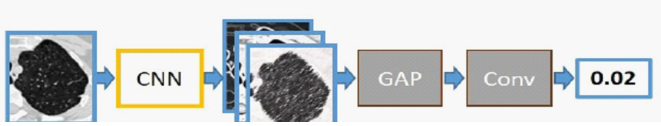


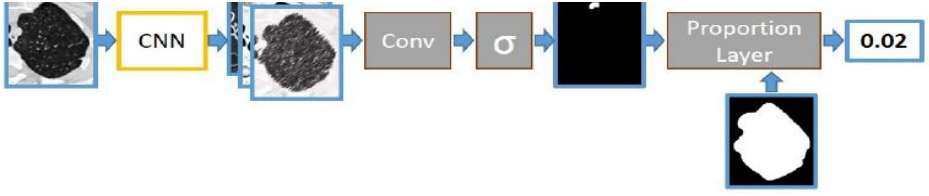
Grade 2 vs. 3

Gerda Bortsova, ECR 2018, MICCAI2018

ProportionNet consistently better than regression network -- especially in smaller training sets

Architecture:	GAPNet		ProportionNet	
Training set size\Task:	Presence	Extent	Presence	Extent
small sets (50, 75, 100)	0.90 ± 0.04	0.74 ± 0.06	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.79 \pm 0.02}$
medium sets (150, 200, 300)	$\mathbf{0.96 \pm 0.01}$	0.80 ± 0.02	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$
large set (700)	0.96	0.79	**0.97**	**0.86**

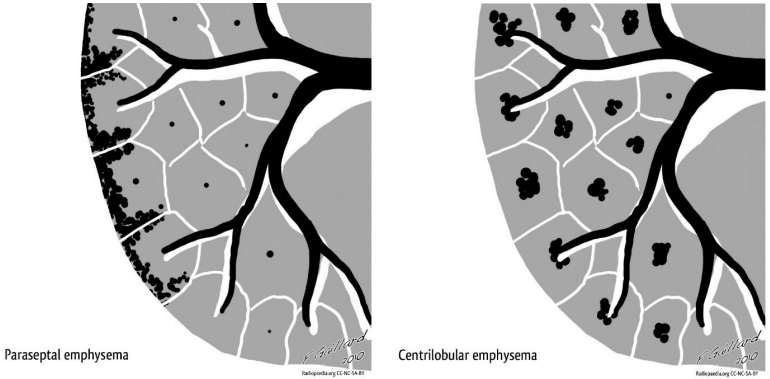




Bortsova et al MICCAI 2018

Evaluation of emphysema segmentations produced with ProportionNet

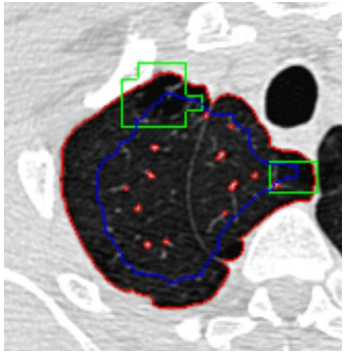
- Radiologists labeled emphysema type
- Paraseptal
- Centrilobular

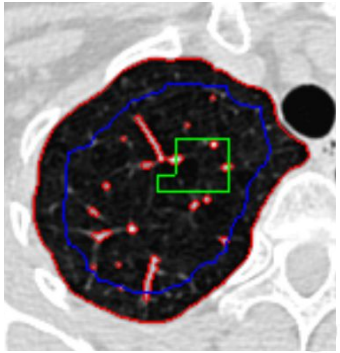


Evaluation of emphysema segmentations produced with ProportionNet

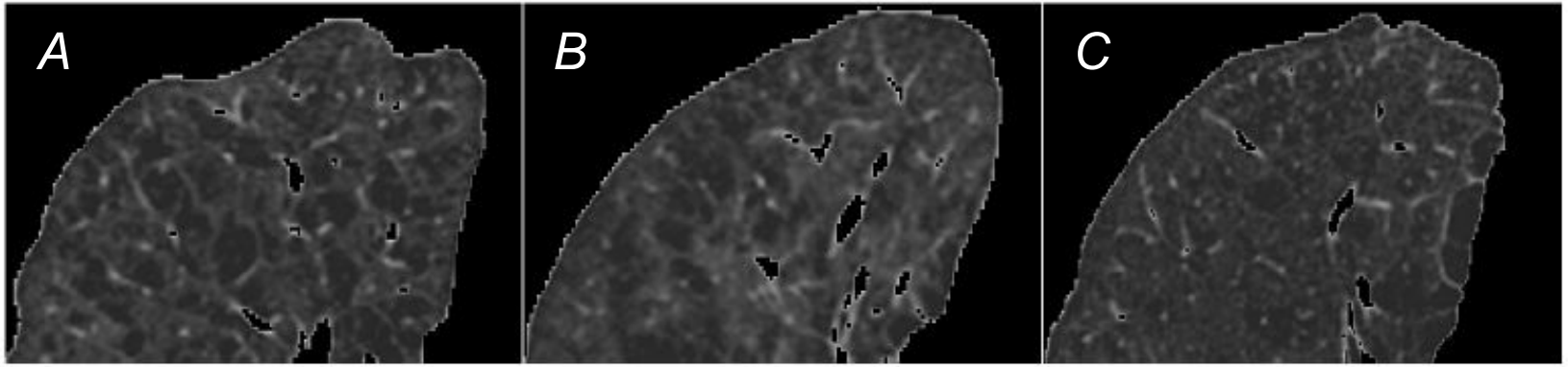
- Paraseptal emphysema pattern score = volume of emphysema around lung border / volume of emphysema elsewhere
- **ROC AUC: 0.894 (rater leveL performance)**

Paraseptal *Centrilobular*





Do we need expert annotations?

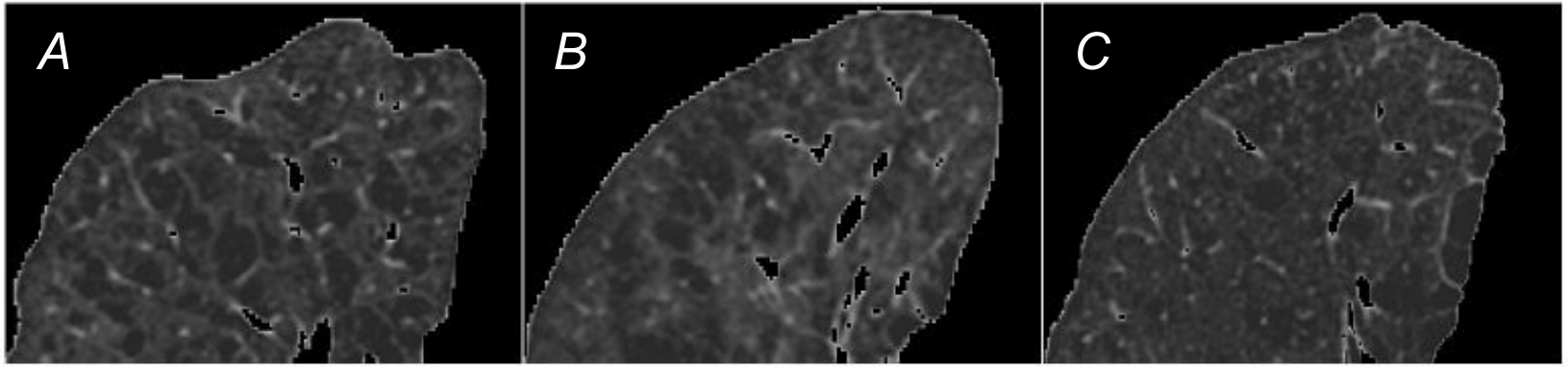


Is image A more similar to B or C?

- Defining emphysema subtype and severity is difficult even for experts. Defining visual similarity may be easier.
- Triplet embeddings based on visual similarity scores
- Could be obtained by crowdsourcing

Work by Silas Orting

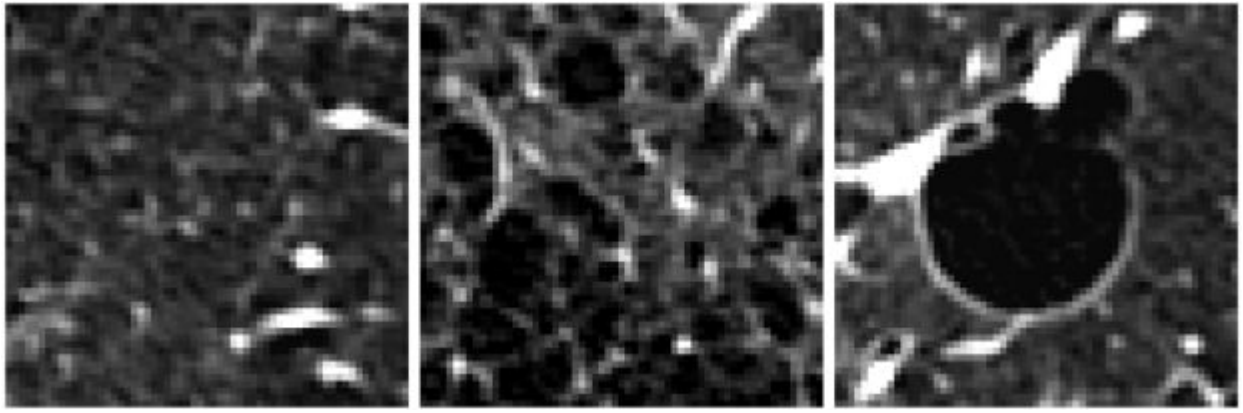
Do we need expert annotations?



Crowd sourcing not as good as experts, but promising

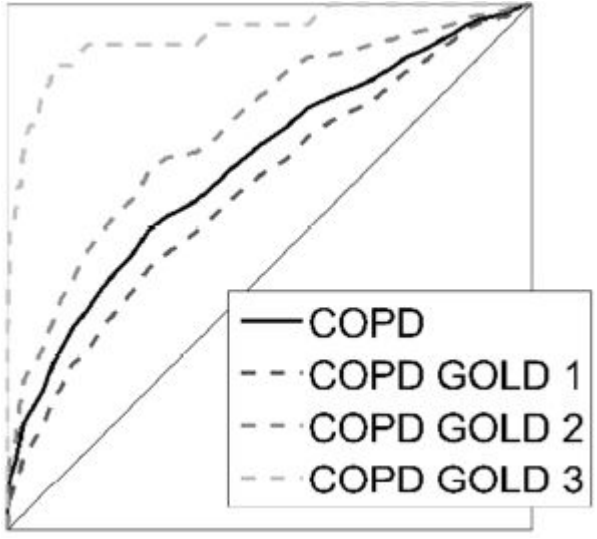
Work by Silas Orting

AI to improve diagnosis: uncovering new patterns

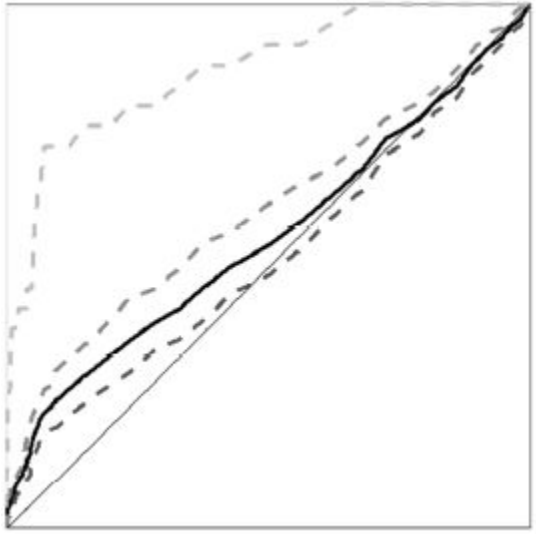


- Various disease subtypes, with different patterns in CT
- Radiologists describe 3 patterns, could there be more?
- Can CT pattern predict disease?

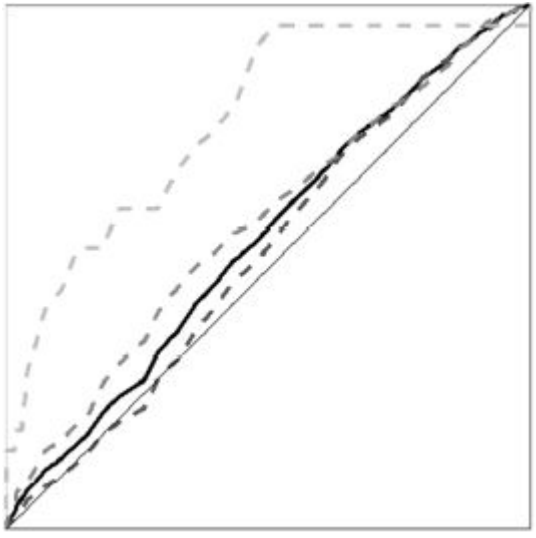
Machine learning detects COPD much better than conventional density measures



Machine learning



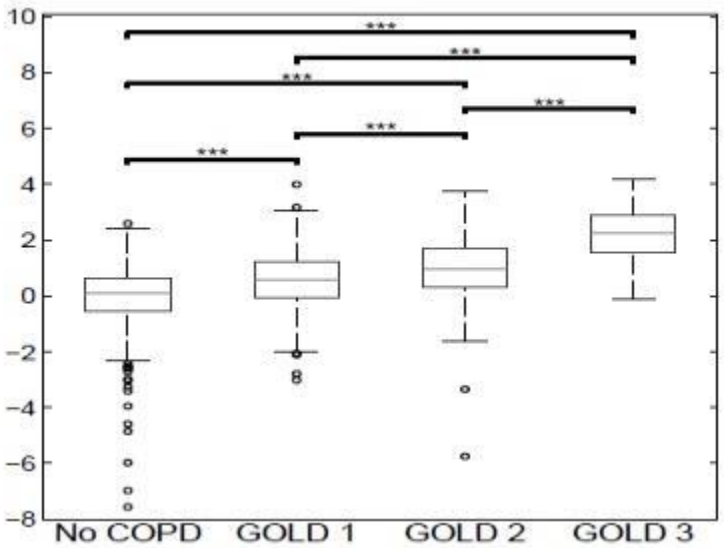
Area below -950 HU



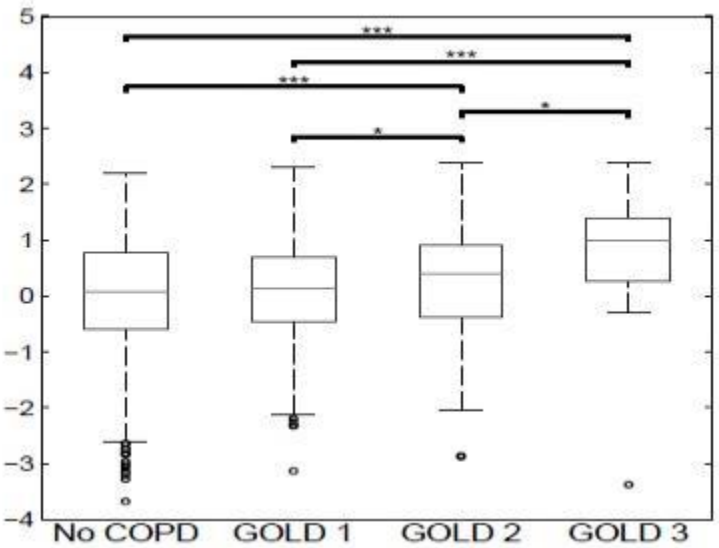
15th percentile

Sorensen et al '19

COPD Severity staging:
Machine learning can detect early stages



Machine learning



Densitometry (RA 950)

ML could also predict future development of COPD (AUC=0.6, $p<0.001$) whereas densitometry could not (AUC~0.5, $p>0.05$)

Sørensen et al '19

To conclude

- Many possible ways to reduce annotation workload
- Many possible ways of introducing prior knowledge in learned models
- Brute force learning probably not the best
- This is a very active area of research

- In some cases, direct prediction of “weak” labels may be better than first estimating “strong” labels as intermediate step

Thanks…

Adria Perez	Aasa Feragen	Netherlands Organization for Scientific Research (NWO)
Andres Arias	Jon Sporring	The Netherlands Organisation for Health Research and Development (ZonMW)
Annegreet van Opbroek	Mads Nielsen	Center for Translational Molecular Medicine (CTMM)
Antonio Garcia-Uceda Juarez	Marco Loog	Danish Strategic Research Council
Arna van Engelen	Stefan Klein	Danish Council for Independent Research (DFF)
Deep Kayal	Theo van Walsum	Innovative Medicines Initiative (IMI)
Florian Dubost	Wiro Niessen	Astra Zeneca
Gerda Bortsova		COSMONiO
Gijs van Tulder	Aad van der Lugt	Quantib
Hakim Achterberg	Arfan Ikram	Vertex
Hoel Kervadec	Asger Dirksen	
Jens Petersen	Harm Tiddens	
Kim van Wijnen	Haseem Ashraf	
Lauge Sørensen	Hieab Adams	
Nora Baka	Jolien Roos	
Oliver Werner	Jesper Pedersen	
Pechin Lo	Klaus Kofoed	
Raghavendra Selvan	Mathilde Winkler Wille	
Robin Camarasa	Meike Vernooij	
Sepp de Raedt	Saher Shaker	
Shuai Chen	Zaigham Saghir	
Silas Ørting		
Veronika Cheplygina		
Vladlena Gorbunova		
Zahra Sedghi Gamechi		