How good is good enough?

What measurements and methods are useful in various clinical imaging contexts, and how to evaluate imaging algorithm performance

De câtă precizie e nevoie?

Măsurători şi metode utile în context medical, şi posibilităţi de evaluare a algoritmilor de imagistică medicală

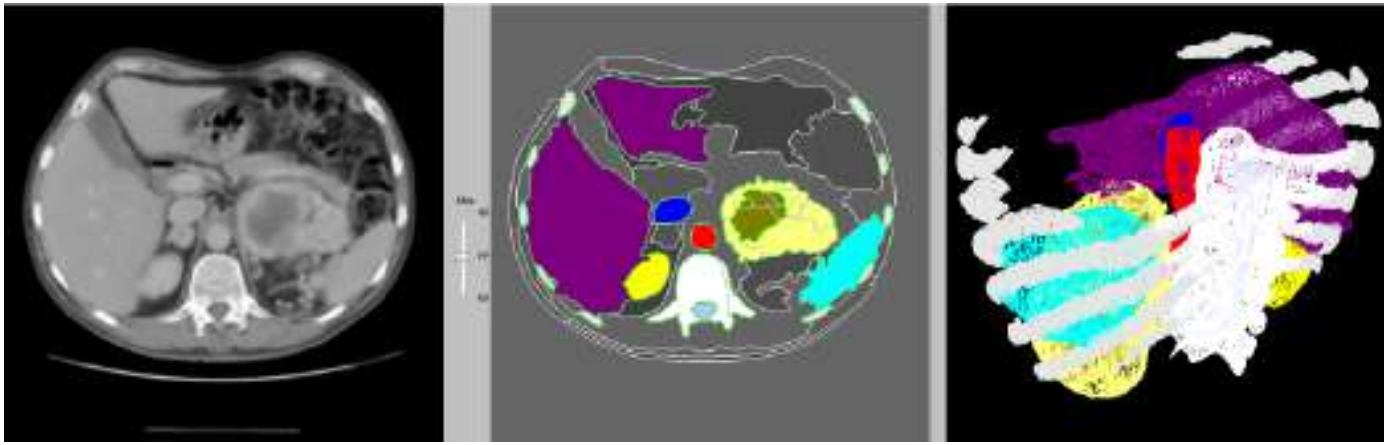
Abordări orientate către om pentru Inteligenţă Artificială de încredere

SMART DIASPORA, Timişoara 2023

Irina Voiculescu

Oxford Medimaging

Who we are and what we do



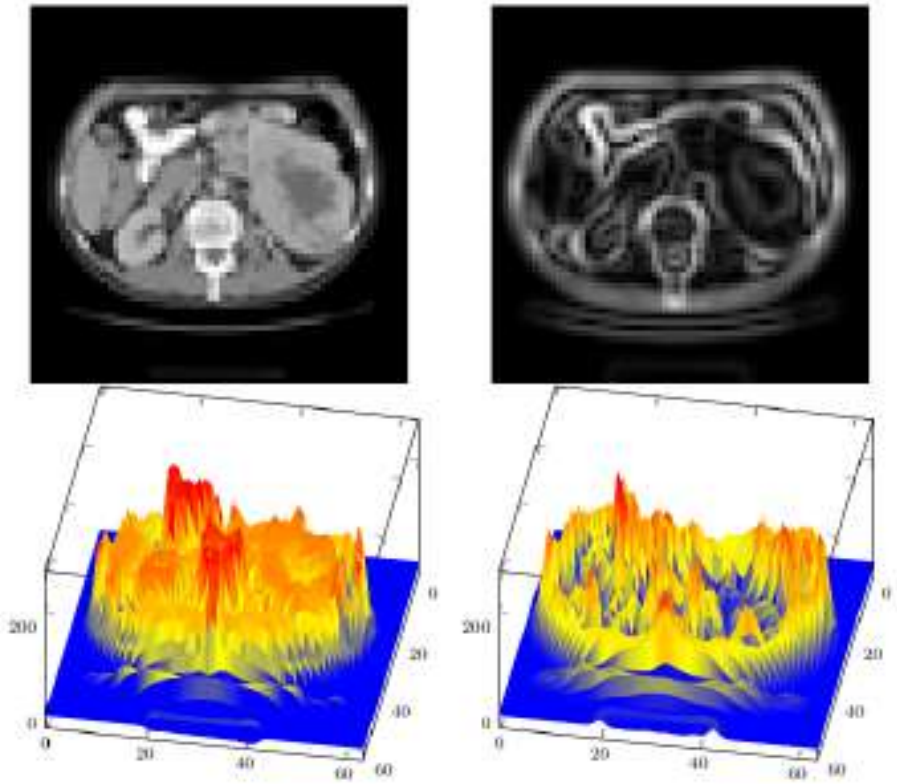
Clinical decision support through:

- semantic segmentation
- 3D reconstruction
- objective anatomical measurements

Segmentation

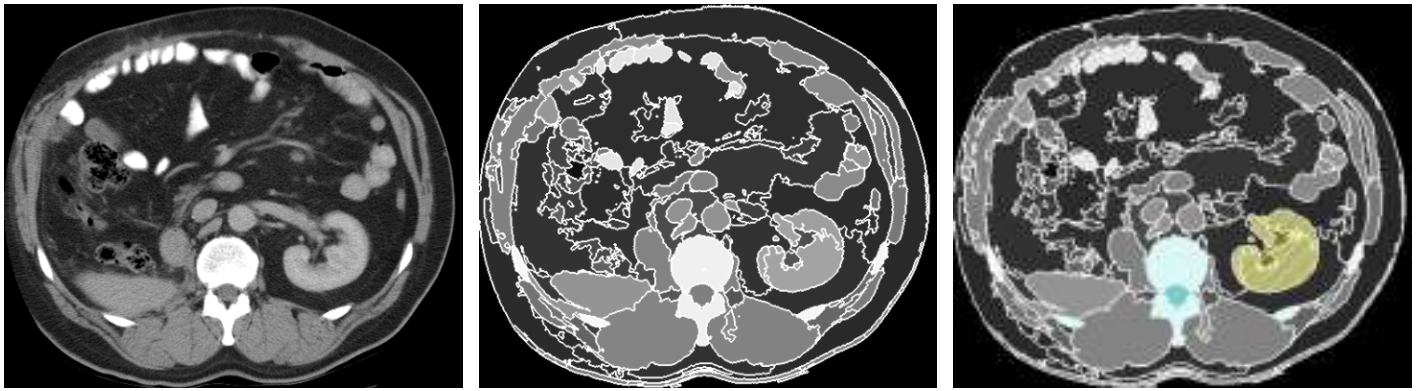
Deterministic methods

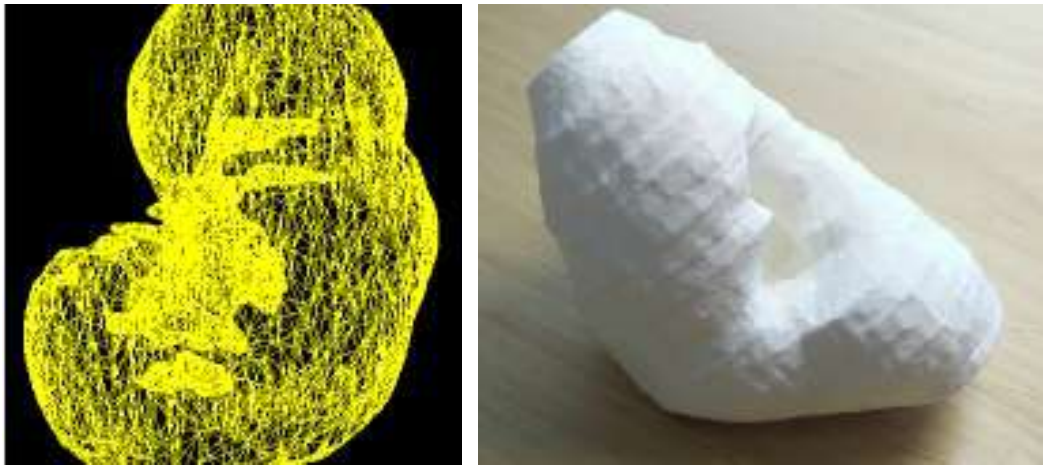
A 2D image can be viewed like a 3D terrain map



An nD image can be viewed like an $(n+1)$D terrain map, $n = 1, 2, 3, \ldots$

Deterministic methods – kidney segmentation





The volumetric calculation correlates with the clinical kidney function test after partial resection

Machine learning for segmentation: fully supervised

Conventional annotated data: fully supervised learning

- plentiful
- reliably annotated

- publicly available
- clinically relevant

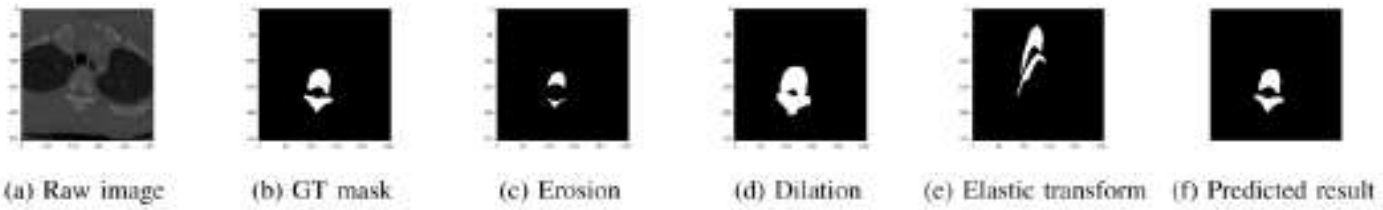


Machine learning for segmentation: partial labels

- Semi-supervised (cross-pseudo-supervision, multi-view learning, etc.)

  As little as 2% of the data is annotated

- Imprecise annotation (noise-robust learning)



(a) Raw image (b) GT mask (c) Erosion (d) Dilation (e) Elastic transform (f) Predicted result

- Weakly supervised (scribble supervision)

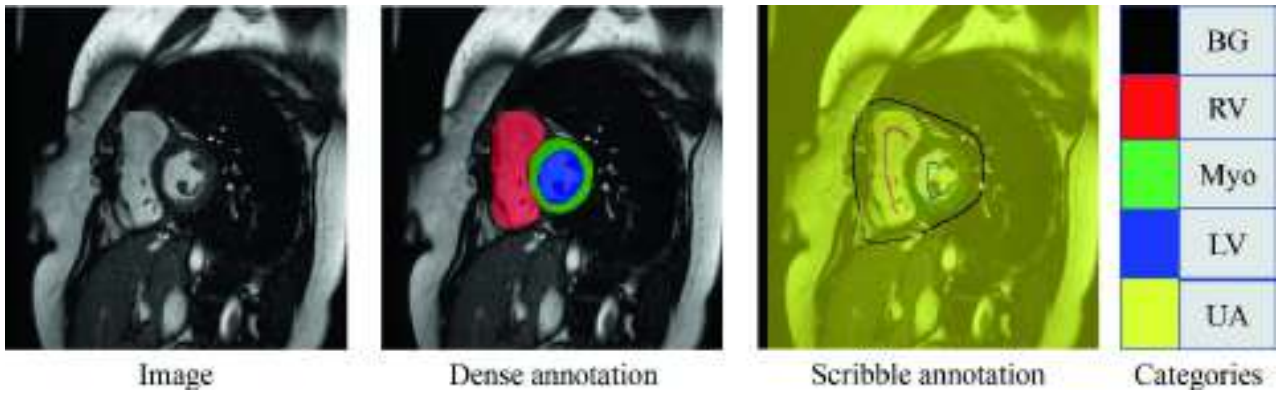
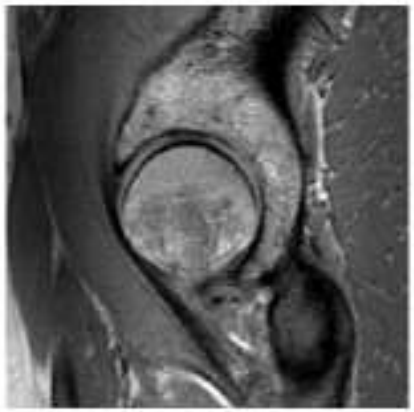


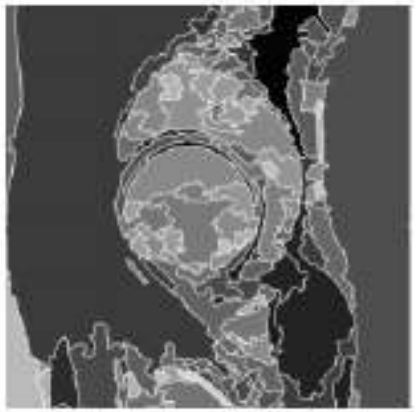
Image Dense annotation Scribble annotation Categories

Landmarks

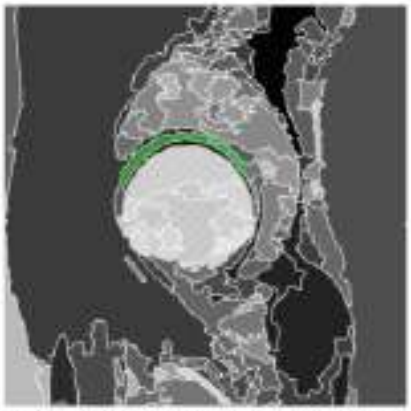
Do we always need near-prefect segmentation?



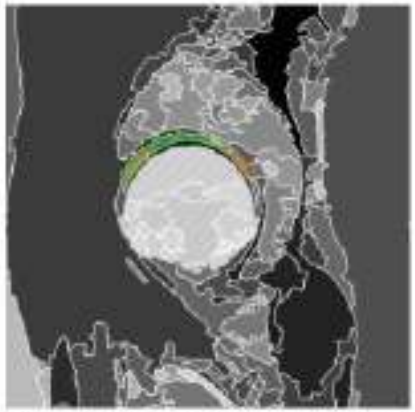
(a) Original image



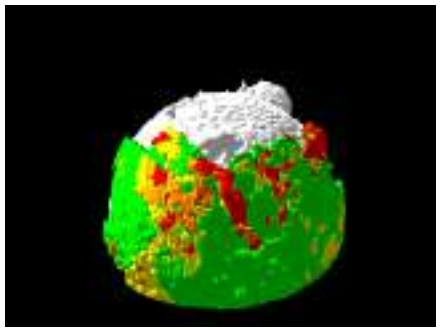
(b) After partitioning, at layer 3 of 6



(c) After user segmentation of the cartilage and femur

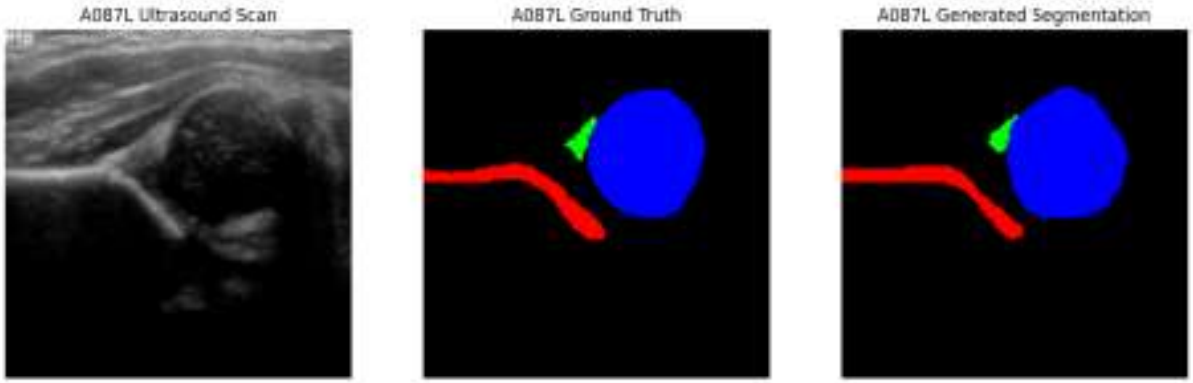


(d) After running the identifier, with yellow and orange regions clear on this scan

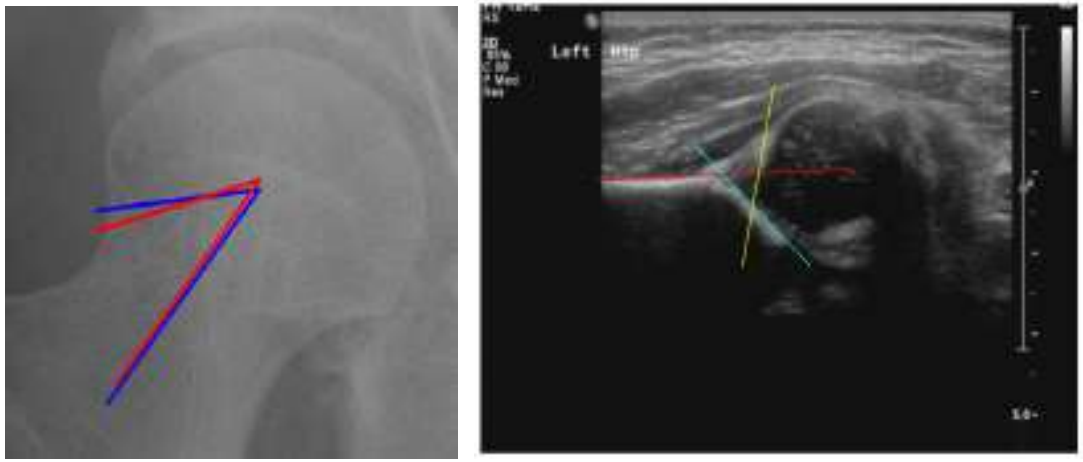


Distance, angle or alignment measurements

The clinical problem should dictate what we measure



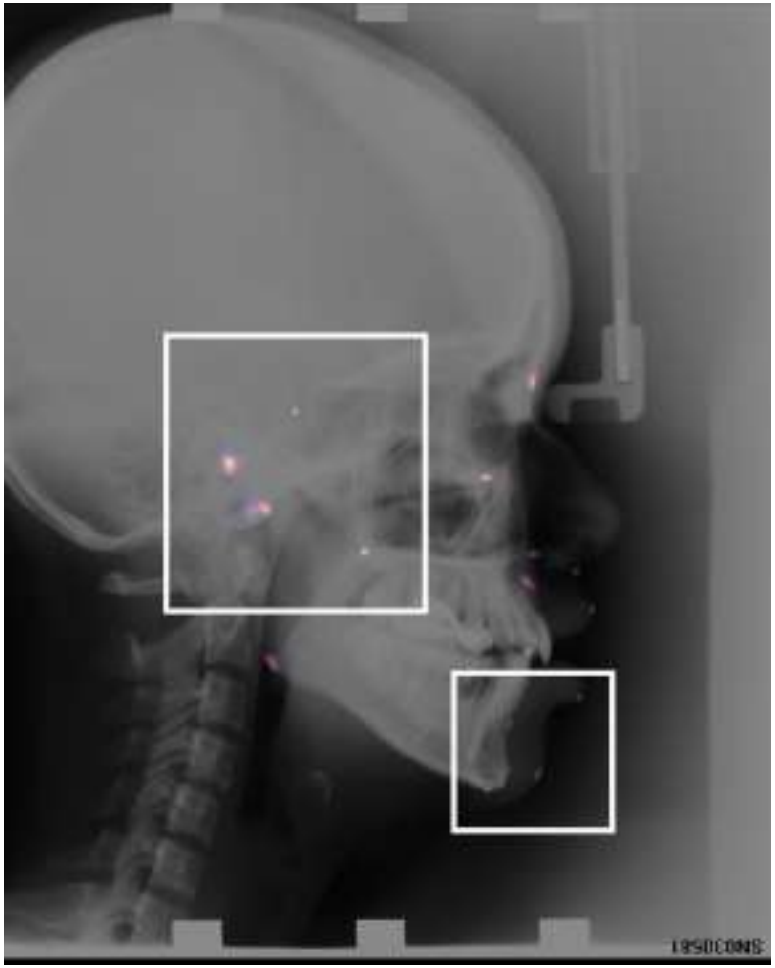
Angles or relative positions — no need for masks

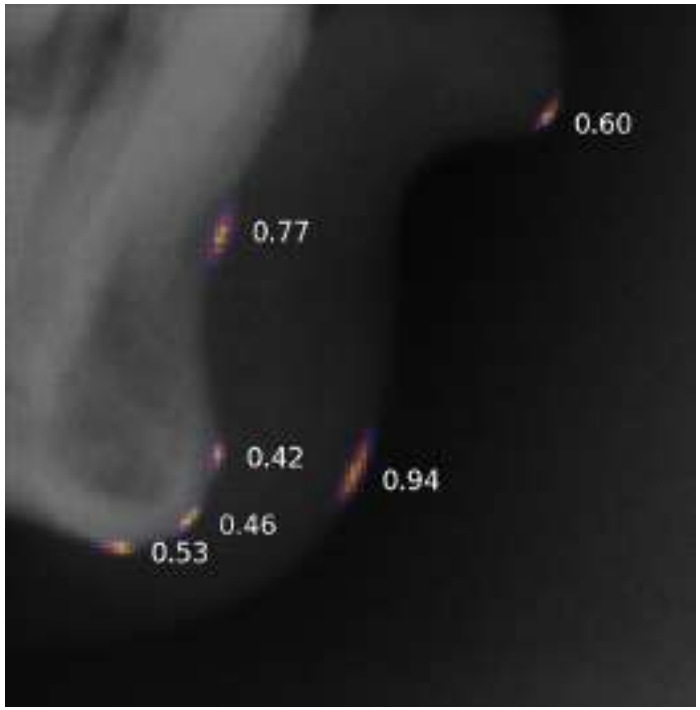


Classification (screening) task need not measure pixels

Landmarks

Landmarks lead to angles and distances



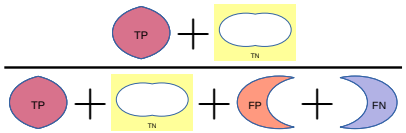


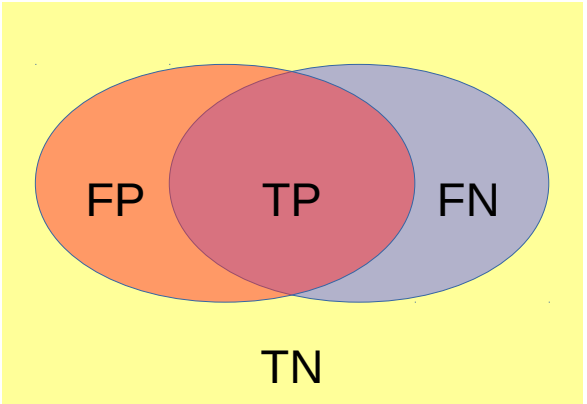
Landmark detection can incorporate a measure of uncertainty

Robustness

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} =$$





People like to hear ‘99% accuracy’

No more relevant than other metrics

Easy to achieve if the feature is small relative to the overall image size

Evaluation measures (full dense masks)

Pipeline: humans draw contours, turn those into masks, generate other masks automatically, and then measure overlap or difference

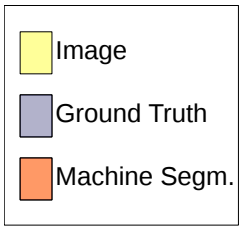
Dice similarity coefficient (DSC) $\frac{2\times TP}{FP + TP + TP + FN}$

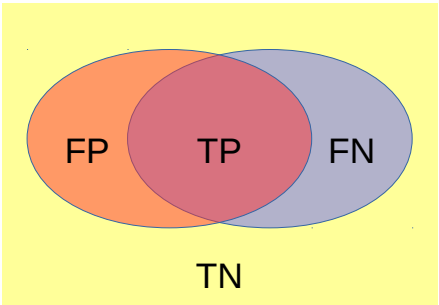
Jaccard similarity coefficient (JSC) $\frac{TP}{FP + TP + FN}$

true positive vol fract (recall, TPVF) $\frac{TP}{TP + FN}$

true negative vol fract (TNVF) $\frac{TN}{TN}$

precision (Prec) $\frac{TP}{FP + TP}$





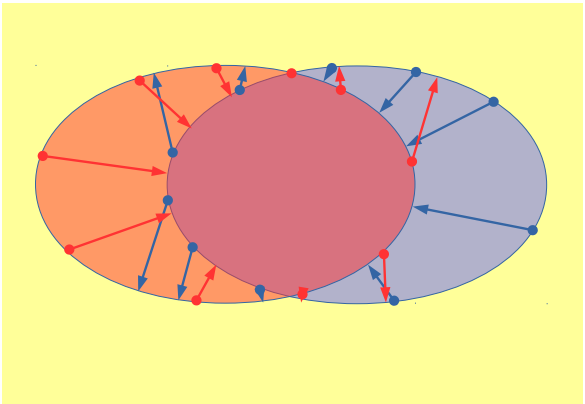
Distance measures

- Distance between two (point) landmarks
- Distance between landmark one-hot-points
- Distance between contours

Define $dist(x, A)$ as **minimum** of $dist(x, y)$ where $y \in A$

maximum symmetric surface distance (Hausdorff, HD)

average symmetric surface distance (ASSD)

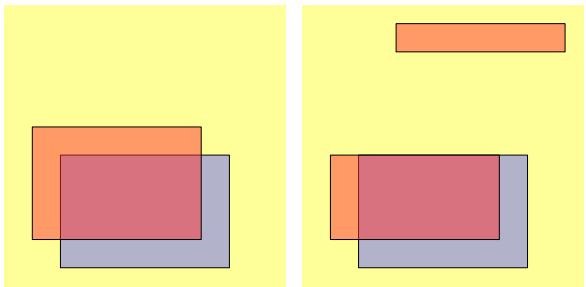


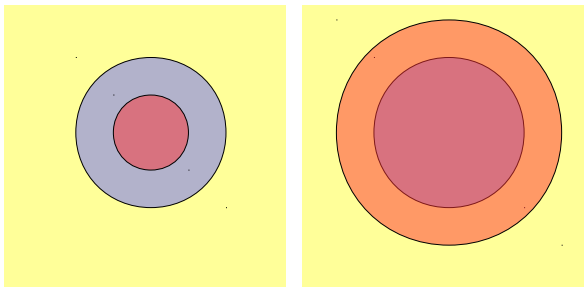
Is this metric suitable?

Ask yourself: are there other more relevant metrics?

Popular evaluation measures based on region overlap or boundary distance

- mostly sensitive to one or another type of segmentation error (size, location, shape)
- as a result, produce contradicting rankings of segmentation results

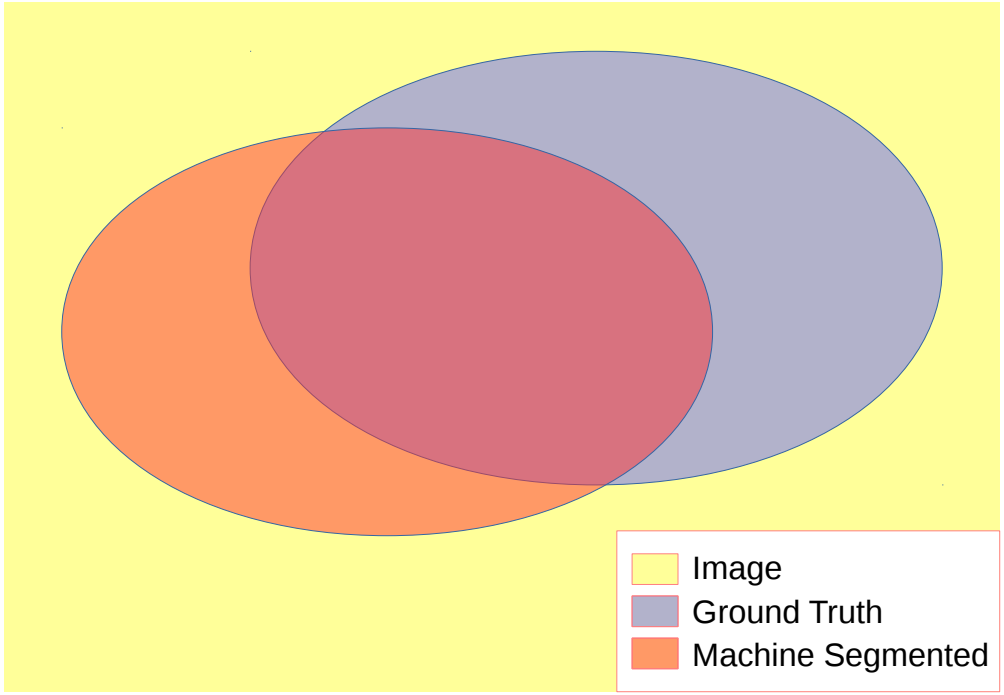




Boundary overlap

Alternative: boundary match

Symmetric Boundary Dice (SBD): Dice similarity coefficient in a small neighbourhood N_x of each point x, x on **first region boundary or second region boundary**

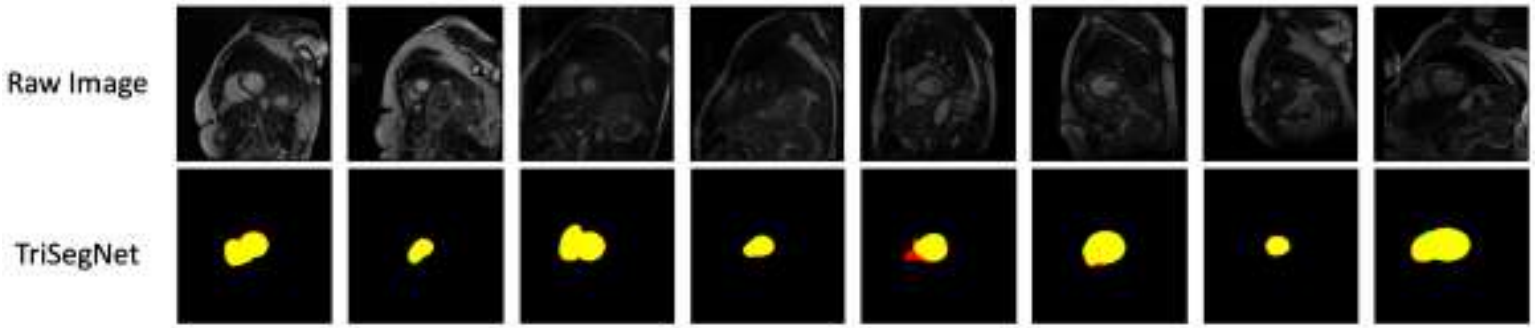


65% is a pretty good match but people don't like it...

Boundary overlap example

Mean results from one of our segmentation algorithms

Dice	Acc	Pre	Rec/Sen	Spe	SBD
0.932	**0.995**	**0.934**	**0.930**	**0.997**	**0.657**



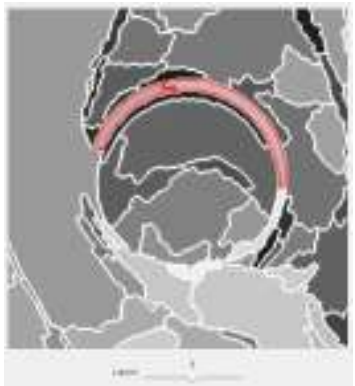
Machine segmentation against ground truth

yellow=TP, green=FN, red=FP, black=TN

Robustness

What other ‘ingredients’ could make this work robust?

Inter–operator and intra–operator variability



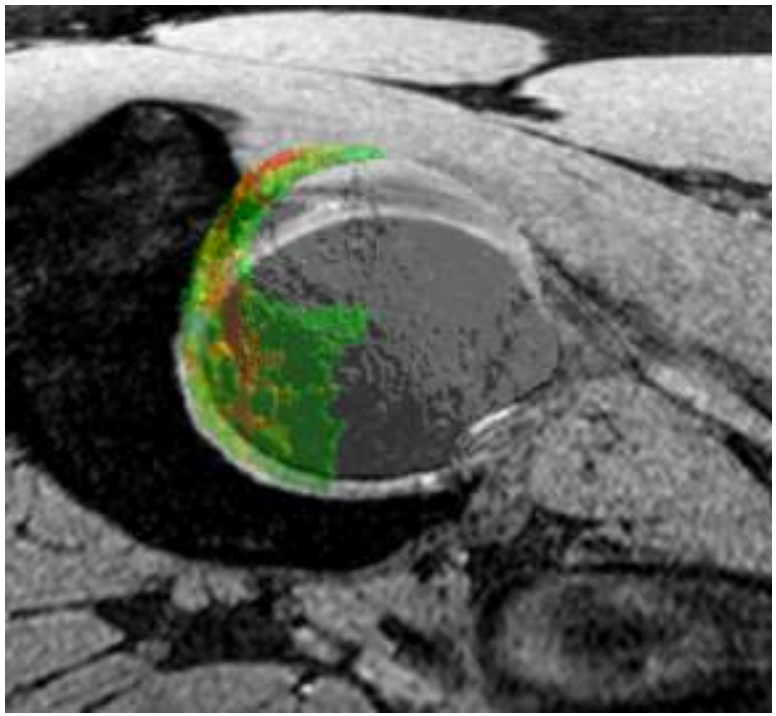
	Dice Similarity Coefficient (DSC)		Accuracy		Sensitivity		Specificity	
Segmentation	Average	SD	Average	SD	Average	SD	Average	SD
Manual vs. semi-automated	0.8803	0.0211	0.9886	0.0315	0.9418	0.0232	0.9984	0.0015
Semi-automated vs. semi-automated								
Intra-observer	0.9726	0.0093	0.9997	0.0009	0.9808	0.0183	0.9996	0.0003
Inter-observer	0.9354	0.0231	0.9991	0.0004	0.9009	0.0551	0.9998	0.0003
Manual vs. manual								
Intra-observer	0.9410	0.0142	0.9992	0.0001	0.9796	0.0115	0.9993	0.0001
Inter-observer	0.9036	0.0141	0.9987	0.0002	0.9660	0.0204	0.9990	0.0002

Is machine result within the difference between humans?

Conclusion

Conclusion

Robust clinical AI applications need



- intuitive visualisation
- appropriate evaluation
- measure of (un)certainty
- explainability

Thanks to the team

Colleagues, research assistants, graduate and undergraduate students: Stephen Cameron, Varduhi Yeghiazaryan, Stuart Golodetz, James McCouat, Abhinav Singh, Sophie Fischer, Andrew Stamper, Cara Higgins, Ziyang Wang, Avraham Sherman, Thaïs Rahoul, Jolyon Shah, Edoardo Pirovano, Chaoqing Tang, Mokrane Gaci, Marija Marčan, Clarice Poon, Ioana Ivan, Chris Nicholls, Jess Pumphrey, Samuel Littley, Tom McDonald, Élise Pegg

Clinicians: Zoe Traill, David Cranston, Andrew Protheroe, Mark Sullivan, Hemant Pandit, Tom Hamilton, David Murray, Scott Fernquest, Daniel Park, Siôn Glyn–Jones, Simon Newman, Daniel Parry

Radiographers: Anthony McIntyre and many others