Disentangled representation learning in medical imaging

Prof. Sotirios A. Tsaftaris
(Sotos)

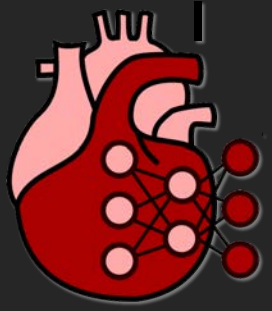
Canon Medical / Royal Academy of Engineering Research Chair in Healthcare AI

http://tsaftaris.com

THE UNIVERSITY of EDINBURGH

ROYAL ACADEMY OF ENGINEERING

CARDIAC AI



The Alan Turing Institute

Canon

CANON MEDICAL RESEARCH EUROPE LTD.

NO conflicts of interest

- We receive financial support from Canon Medical

- Part of our work is funded by: NIH, EPSRC, BBSRC, Innovate UK, European Commission

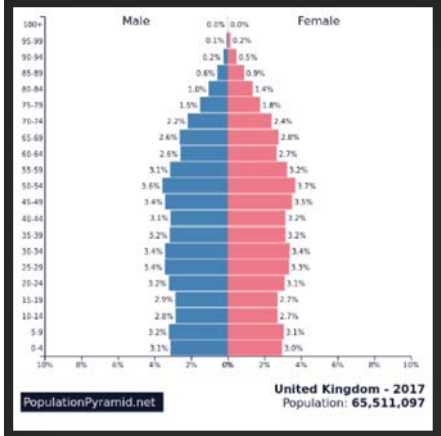
- The views presented herein represent the views of the speaker and not of any of the sponsoring bodies

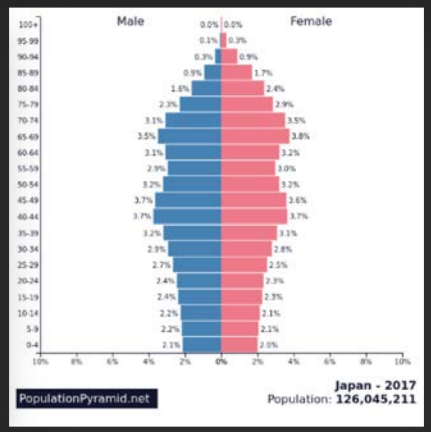
A short bio

- PhD (MSc) from Northwestern (2006)
- Faculty at EECS/Radiology
- IMT Sept. 2011-2015, Director of PRIAn
- University of Edinburgh, Sept 2015
  - **Turing fellow (since 2017)**

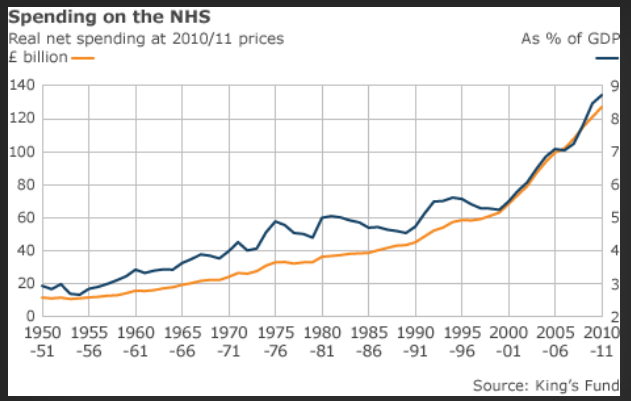
The problem: **healthcare under a perfect storm**

Ageing population





Increasing costs

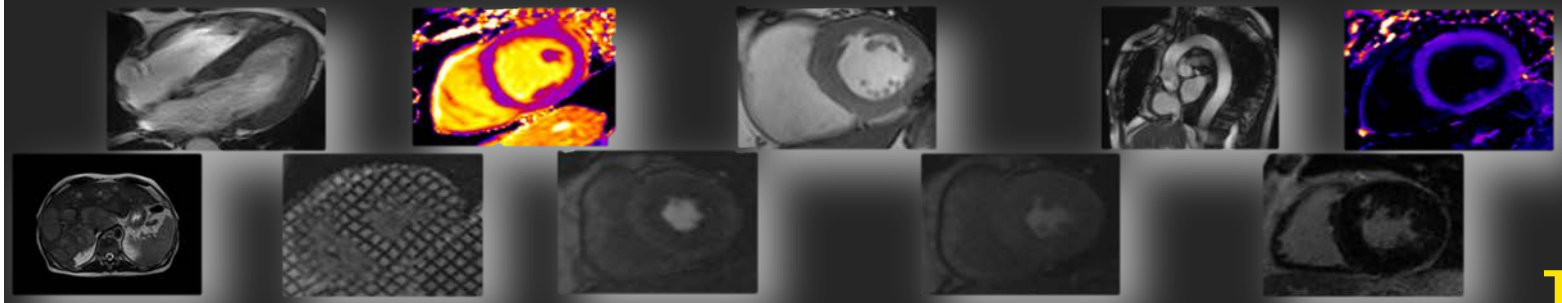


Shortage of experts

Better care for **less**

Sources: Royal college of Cardiologists; King's Fund

Multimodal data: information explosion



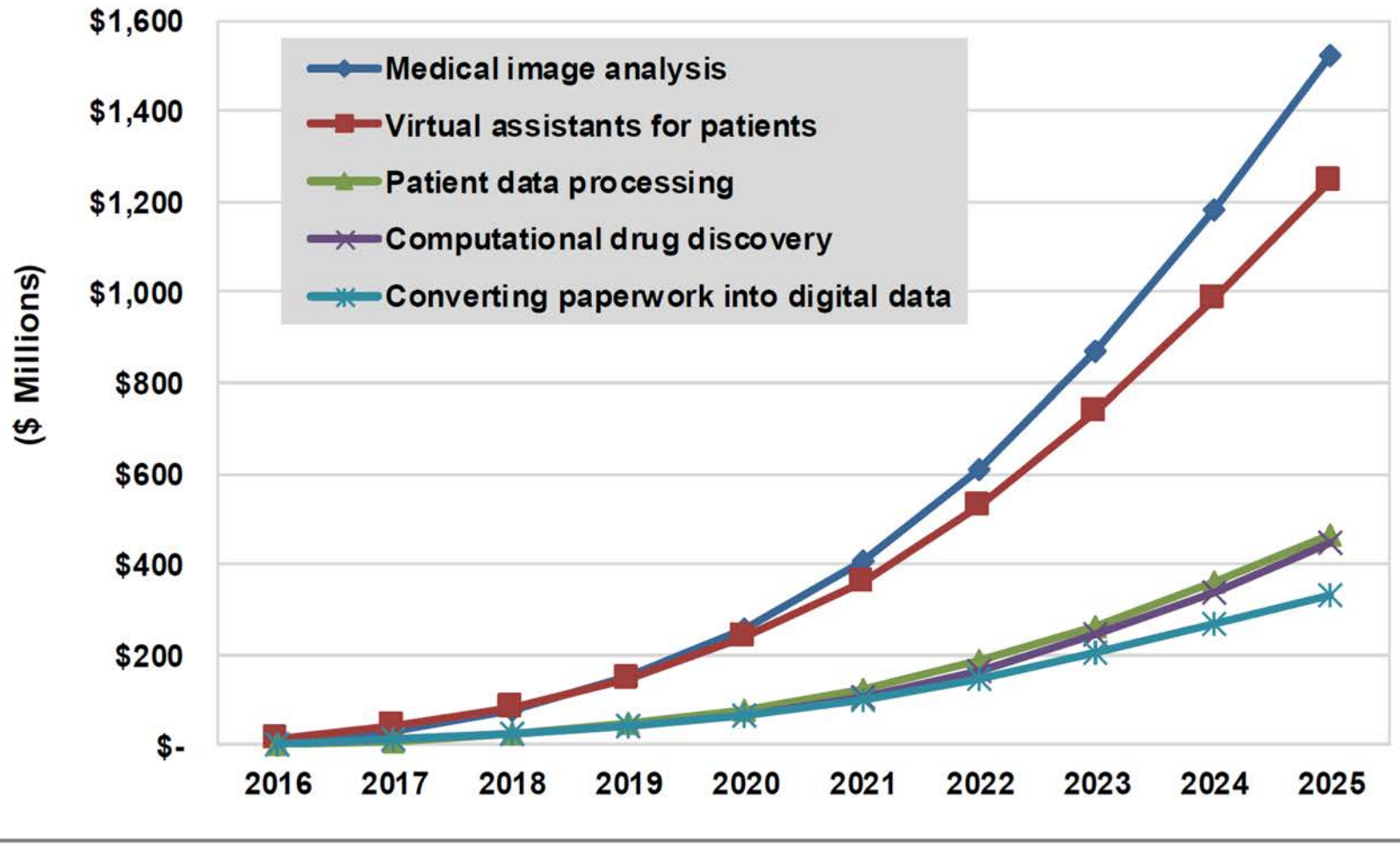
Thousands of images in one exam

Unstructured

Pt is 40yo mother, software engineer
HPI : Sleeping trouble on present dosage of Clonidine. Severe Rash on face and leg, slightly itchy
Meds : Vyvanse 50 mgs po at breakfast daily, Clonidine 0.2 mgs -- 1 and 1 / 2 tabs po qhs
HEENT : Boggy inferior turbinates, No oropharyngeal lesion. Lungs : clear. Heart : Regular rhythm. Skin : Papular mild erythematous eruption to hairline
Follow-up as scheduled

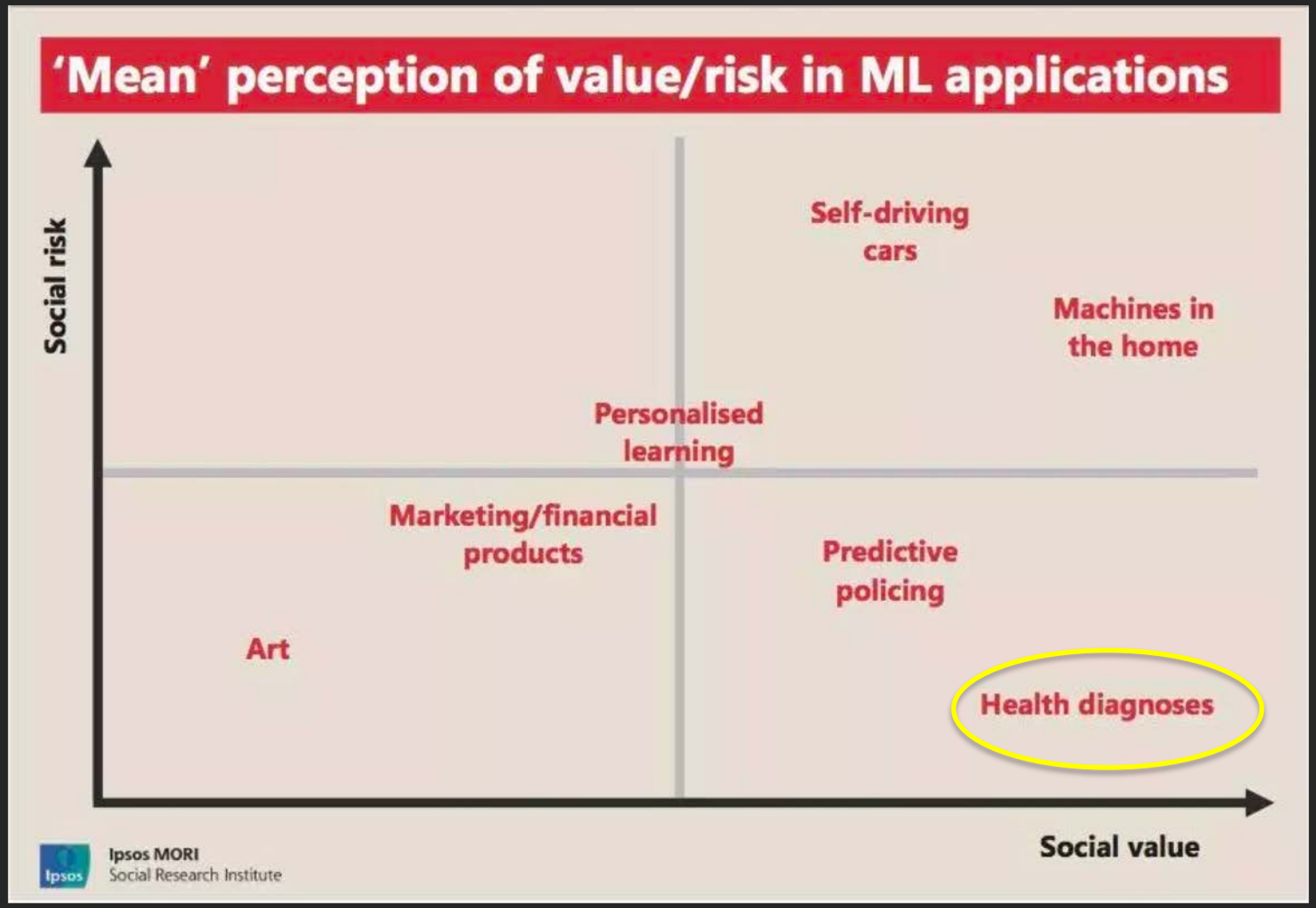
Opportunity: market view

Chart 1.3 Top Five Healthcare Artificial Intelligence Use Cases Revenue, World Markets: 2016-2025



(Source: Tractica)

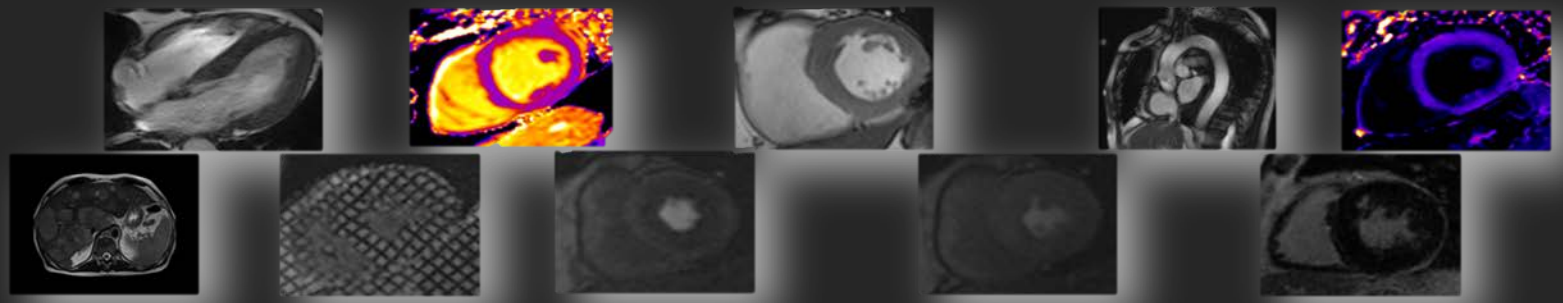
Opportunity: patient view



- Patients say it is … "*a **necessary** development*"

Source: page 29, Ipsos MORI report on behalf of The Royal Society: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

Medical Image Analysis



- Lots of images in 2D, 2D+time, 3D, 3D+time
- Various intensity profiles ➔ multiple image modalities
- Mix of quantitative vs qualitative imaging
- Goal extract biomarkers:
  - Segment anatomy
  - Segment pathology
  - Register anatomy/pathology across images in time or across image type (modality)
- Be robust …

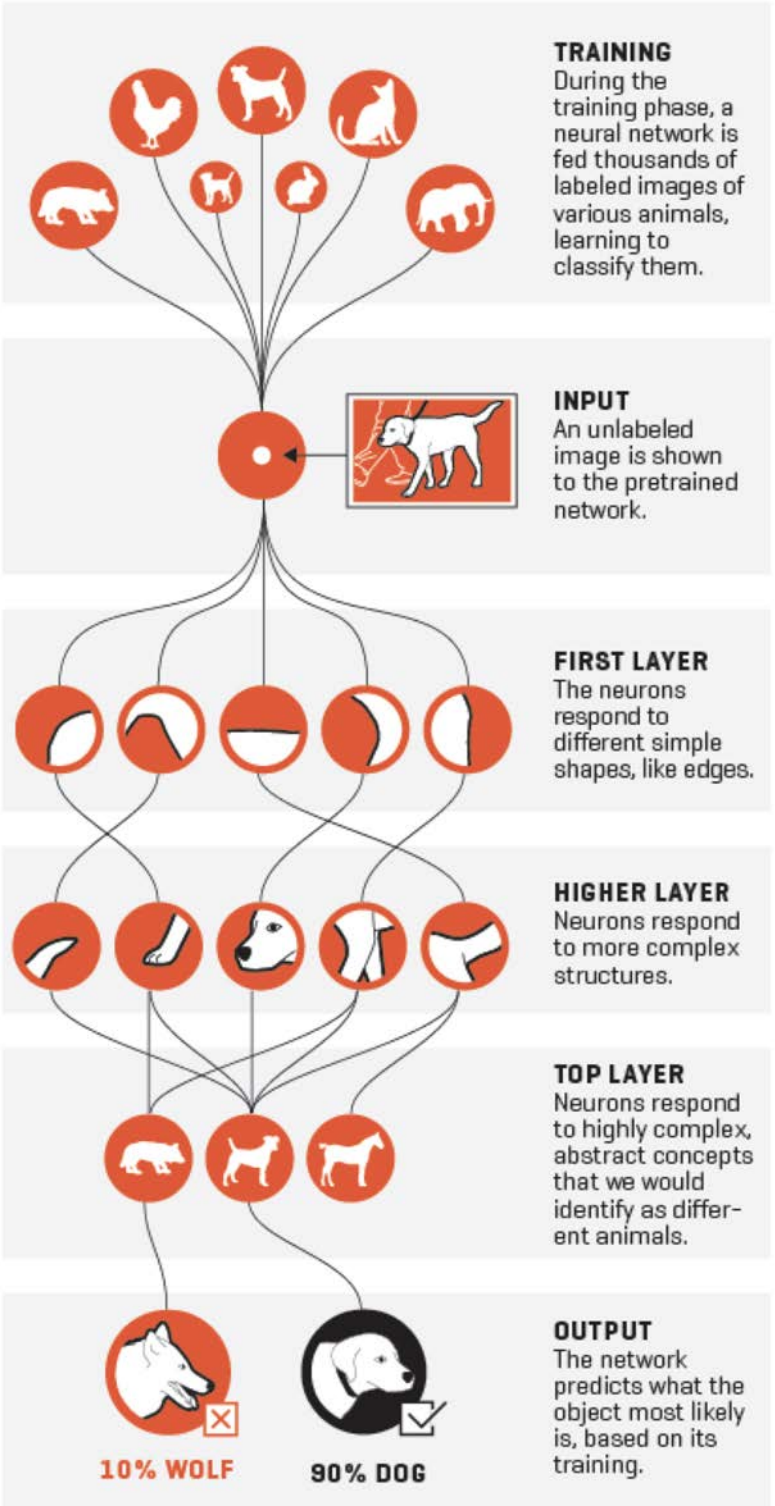
Broad goal: algorithms that work well…

- Across different
  - **tasks**?
  - **hospitals/scanners**?
  - **populations**?

Valve detection vs myocardial segmentation

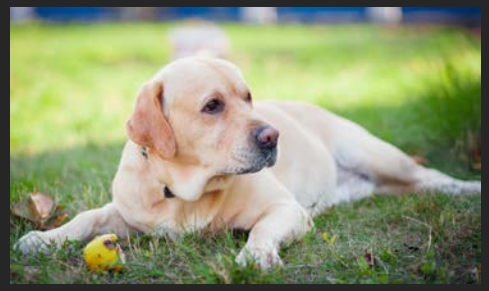
1T vs 3T

- Despite not having **enough and perfect** training data?
  - Collecting data not easy [privacy, cost]
  - Annotations are costly (experts)
  - Carry bias



Deep learning

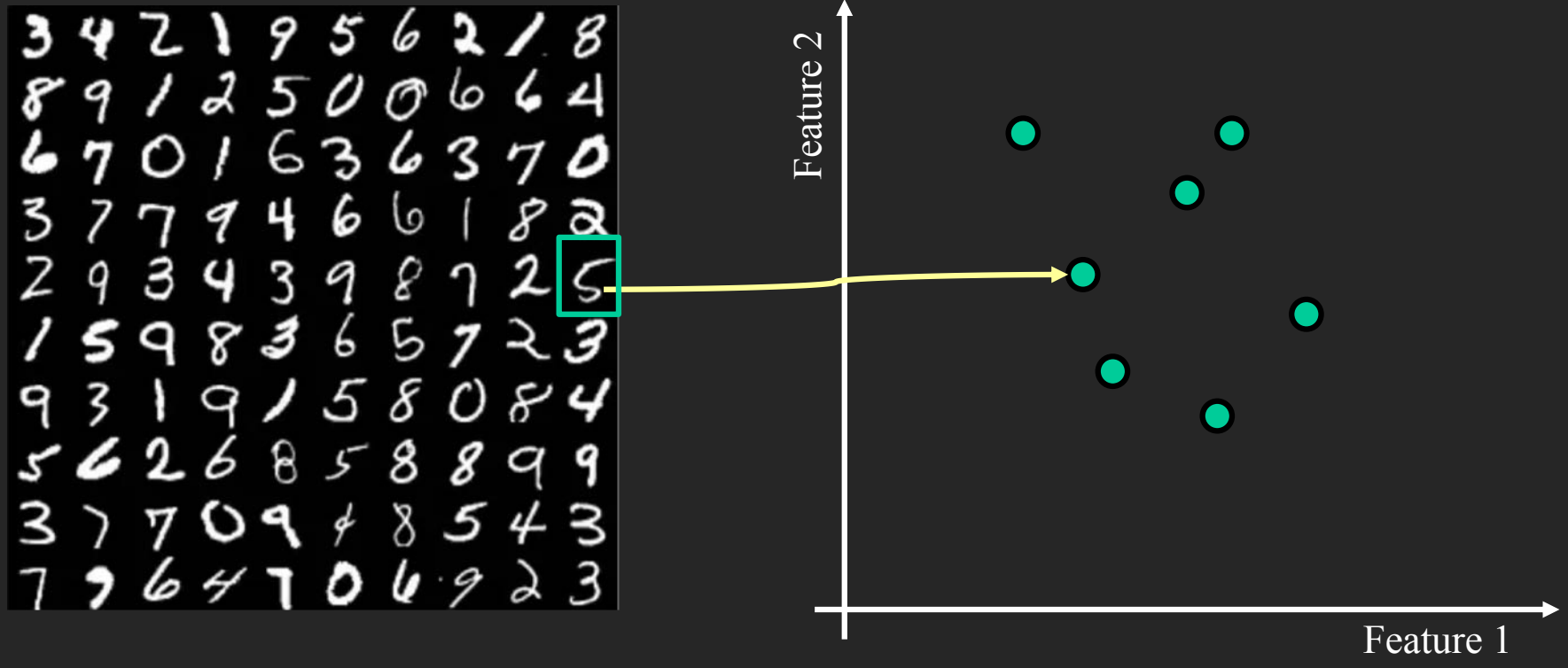
- Where is the representation space?
- What makes a good space?



http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

What is a disentangled space?

Consider a mapping from digit pictures into a space

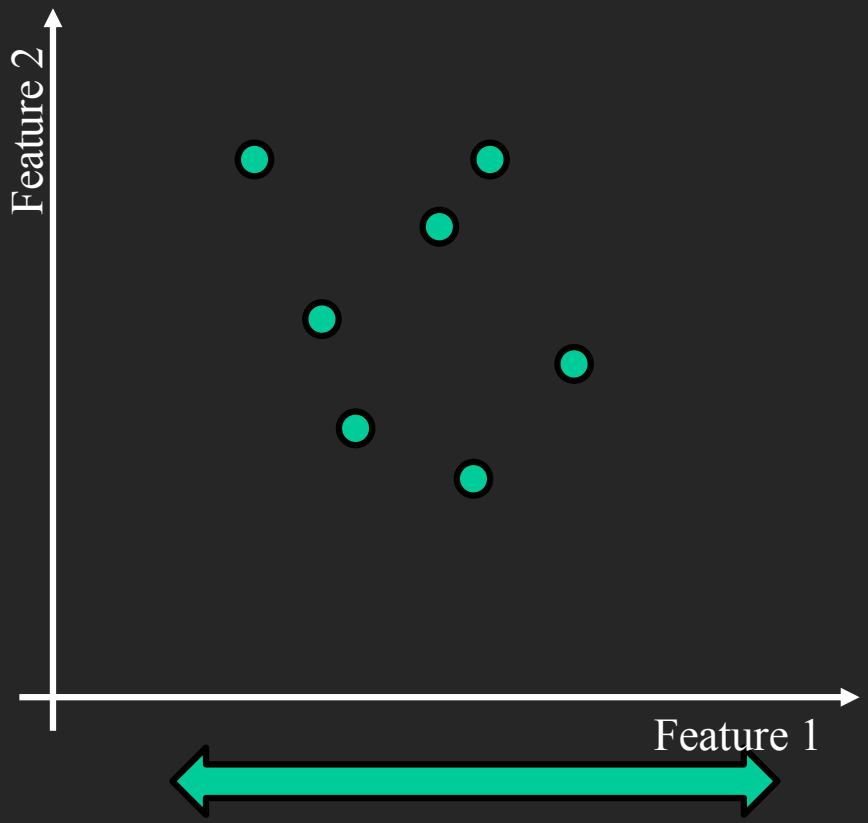


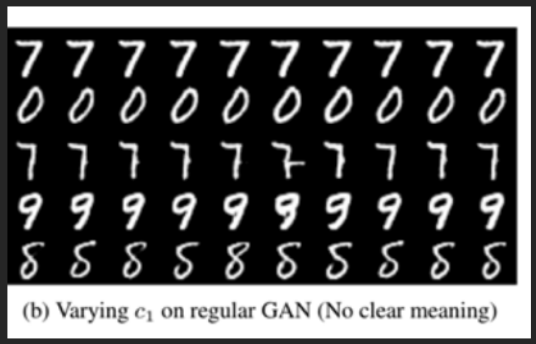
e.g. similar digits to be close together

Source: Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.

What is a disentangled space?

Walking along 1 feature dimension



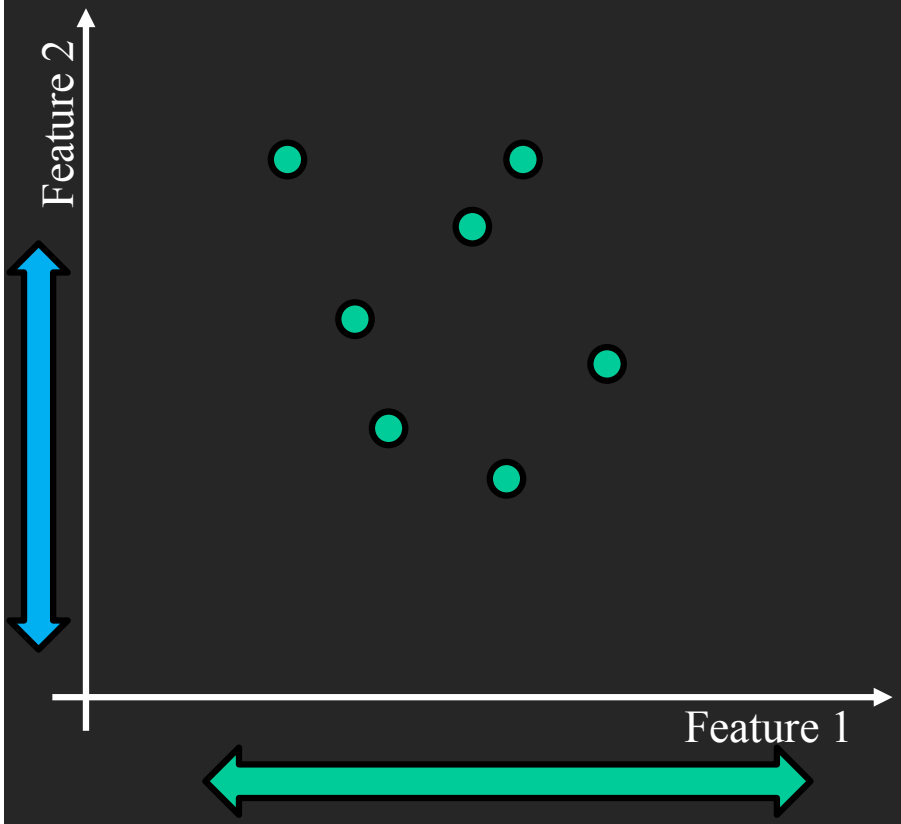


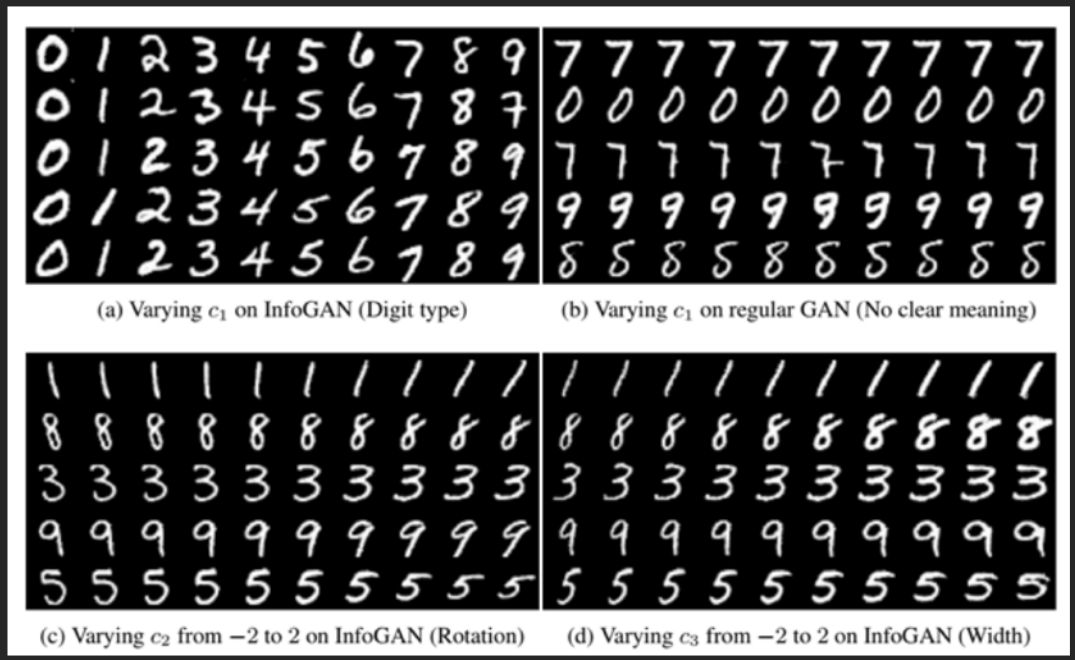
(b) Varying c_1 on regular GAN (No clear meaning)

this is what is captured

What is a disentangled space?

Variables unrelated from each other that best explain input data and its structure





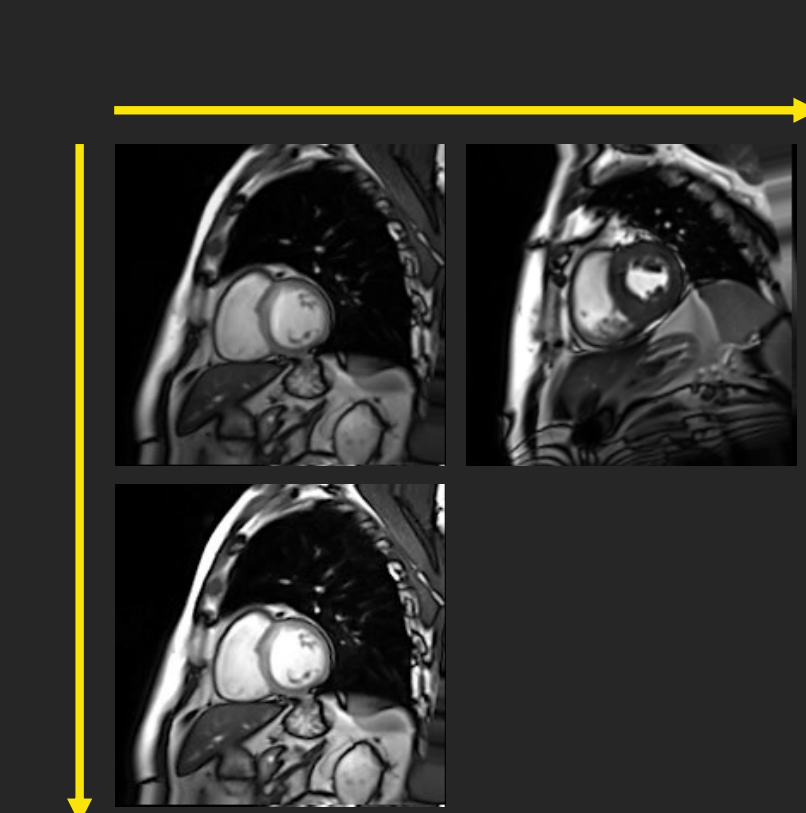
(a) Varying c_1 on InfoGAN (Digit type)

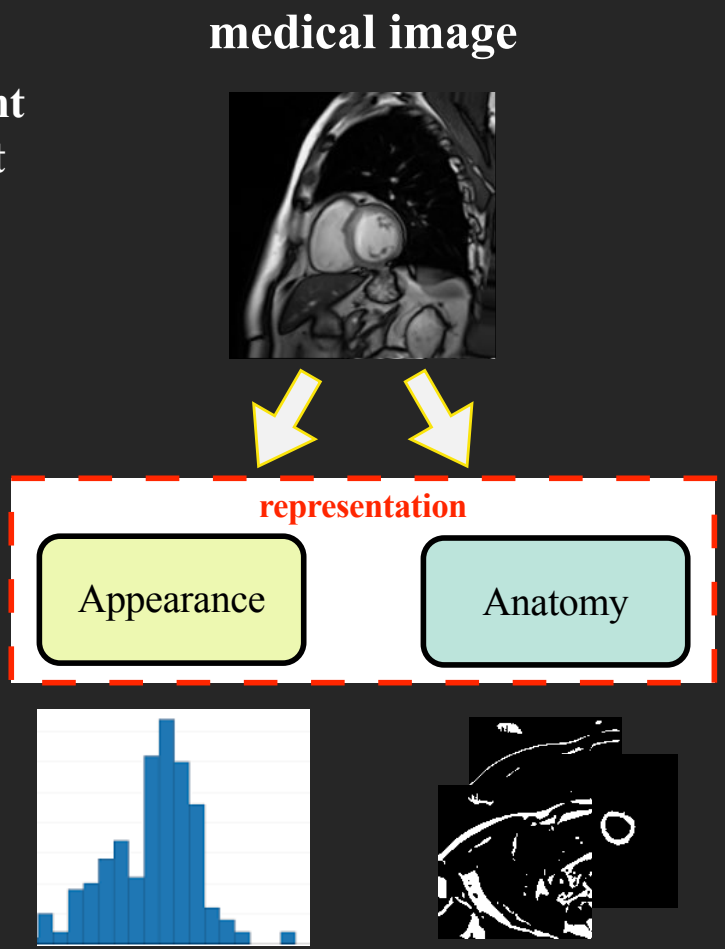
(b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from −2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from −2 to 2 on InfoGAN (Width)

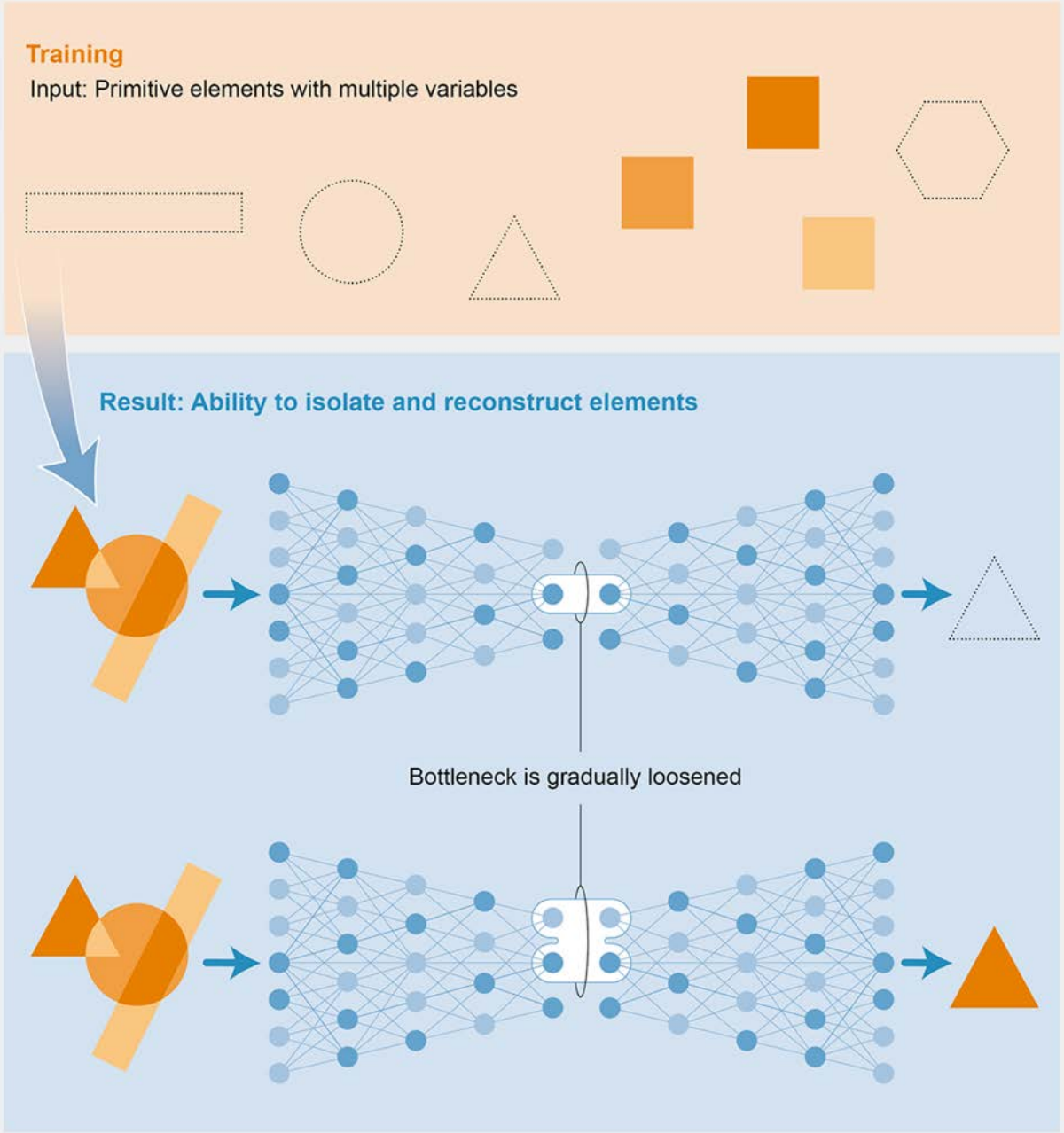
An example in our context





Disentanglement

A machine can learn to pick apart a scene into the objects that constitute it. One network compresses the input data; the other expands them again. By constricting the link between the two, the system is forced to find the most parsimonious description. That is usually the description a human would use, too, thereby making the network more transparent in its operation.



COMPUTER SCIENCE

ARTIFICIAL IMAGINATION

How machines could learn creativity and common sense, among other human qualities

By George Musser

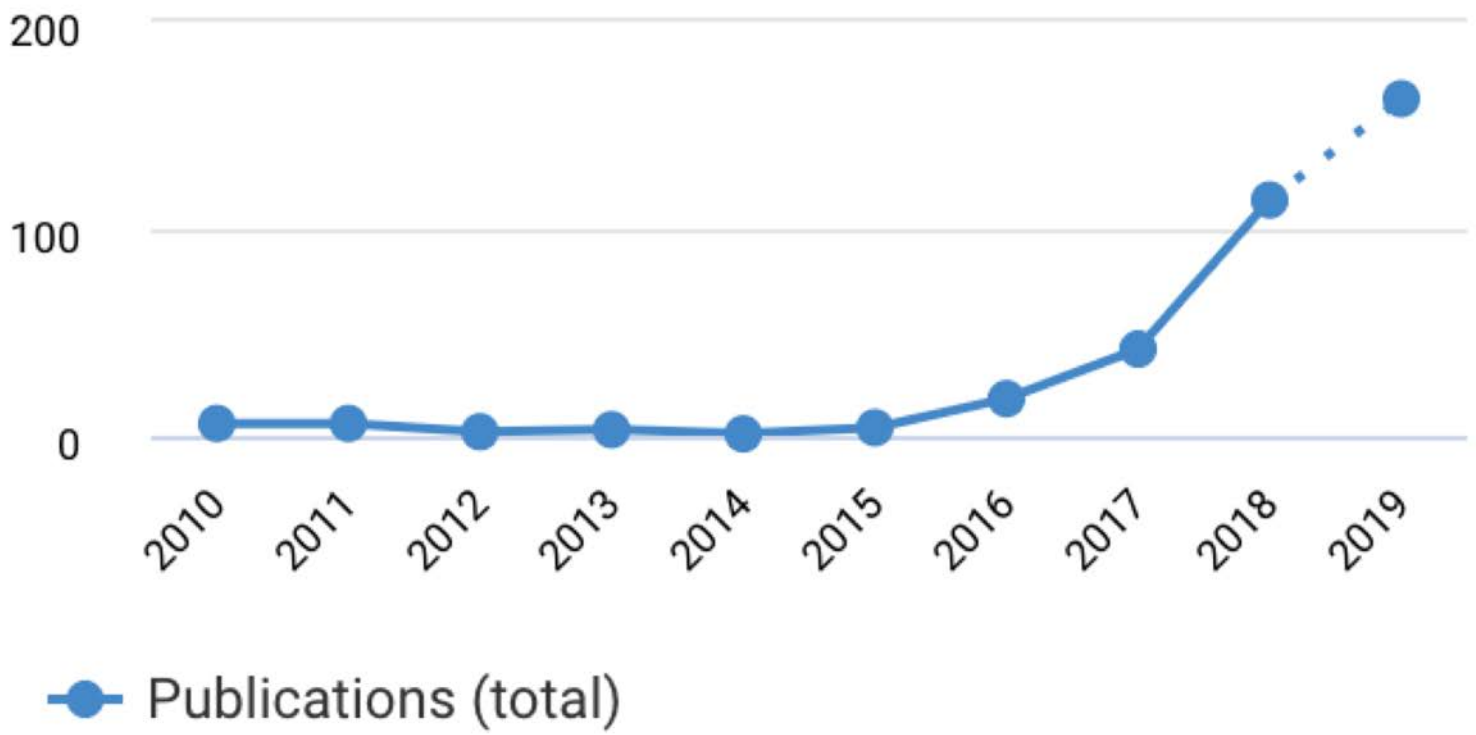
IF YOU EVER FEEL CYNICAL ABOUT HUMAN BEINGS, A GOOD ANTIDOTE IS TO TALK to artificial-intelligence researchers. You might expect them to be triumphalist now that AI systems match or beat humans at recognizing faces, translating languages, playing board and arcade games, and remembering to use the turn signal. To the contrary, they're always talking about how marvelous the human brain is, how adaptable, how efficient, how infinite in faculty. Machines still lack these qualities. They're inflexible, they're opaque and they're slow learners, requiring extensive training. Even their well-publicized successes are very narrow.

IN BRIEF

Several emerging methods endow artificial-intelligence systems, such as neural networks, with features that were once considered to be quintessentially human. **Meta-learning primes a network** to adapt quickly so that it can pick up new tasks without requiring reams of data. **So-called generative adversarial networks** provide a form of imagination, letting machines reproduce the statistical features of data sets. **Disentanglement sensitizes** neural networks to the underlying structure of data, making their inner workings more understandable in human terms.

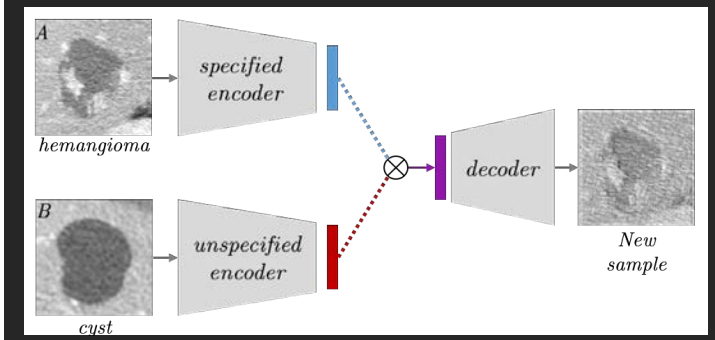
Illustration by Gabriel Silveira

Increasing popularity

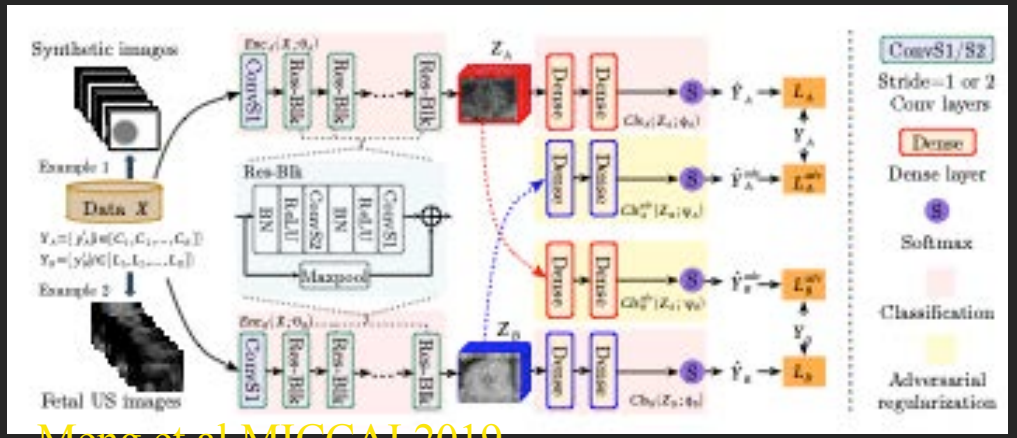


Source: Google Scholar 2019

…and in our community



Ben-Cohen et al Arxiv 2018



Meng et al MICCAI 2019

10 papers in MICCAI 2019 alone

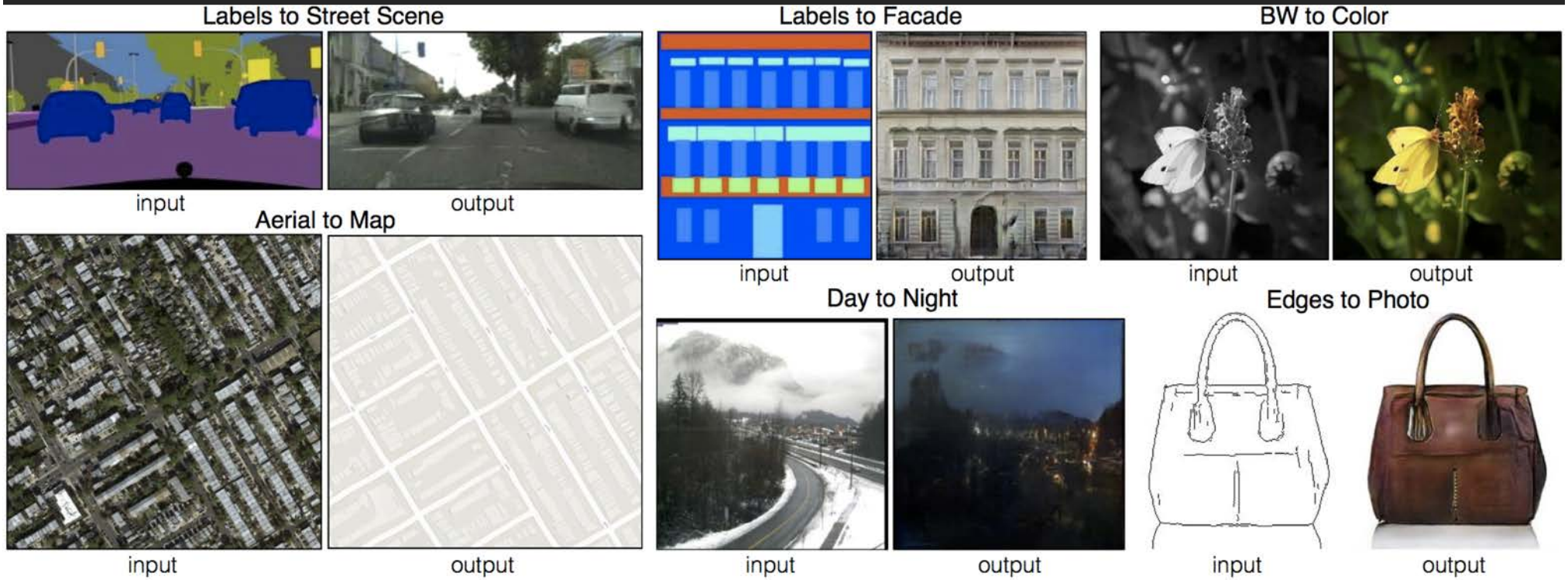
DISENTANGLEMENT IS EVERYWHERE

Few recent examples from our work

Image translation

- A mapping of one image (modality) to another
  - a segmentation map
  - the same image (reconstruction)
  - another image (e.g. from T1 to T2)

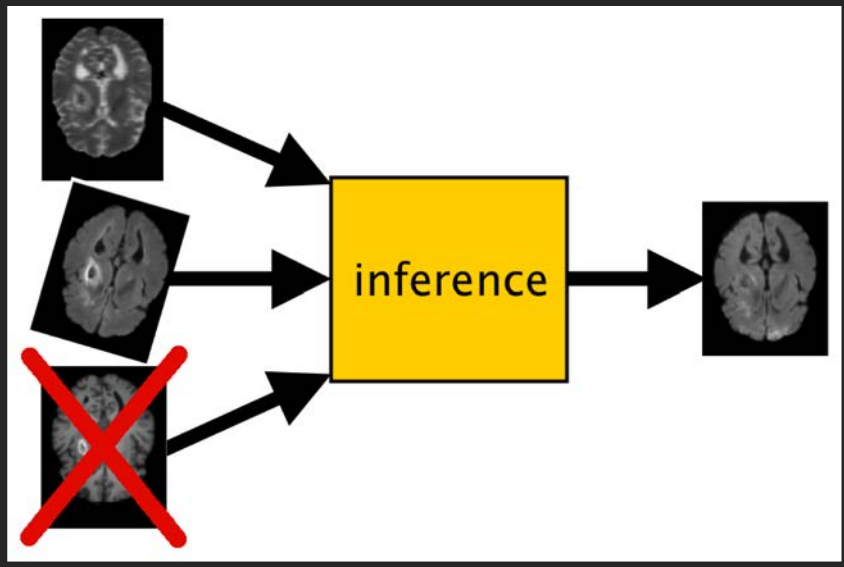
} Self-supervision



Example taken from Zhu et al, CycleGAN 2017

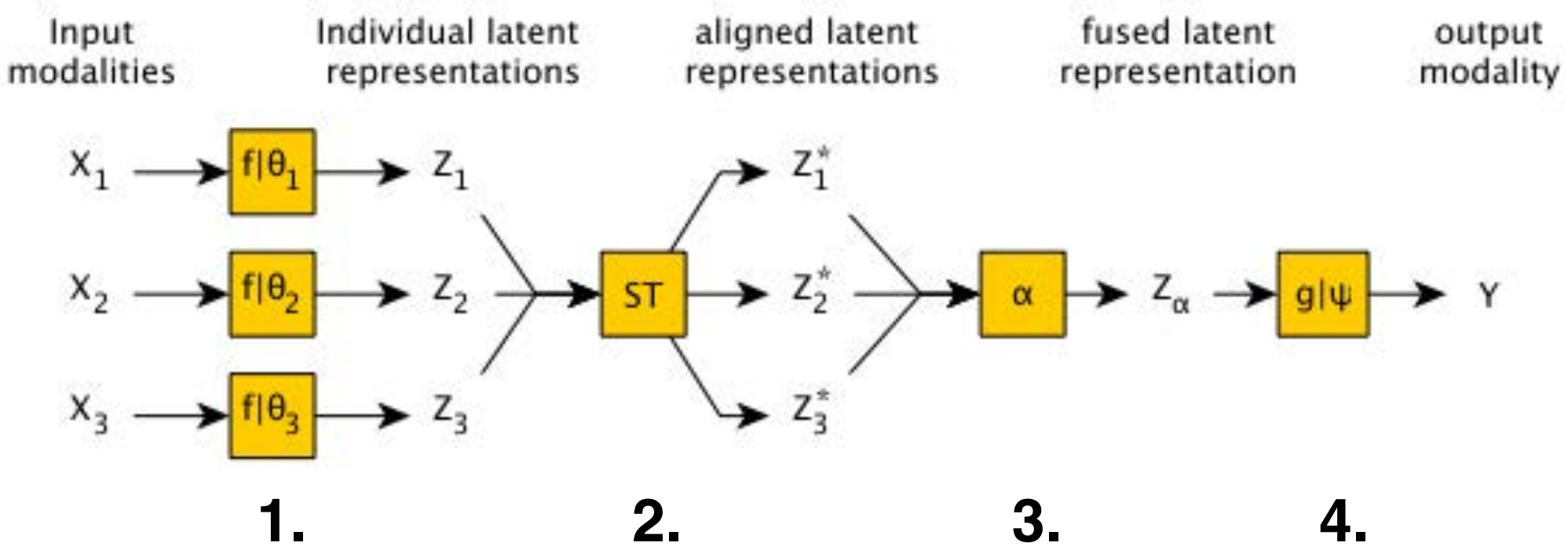
Disentangling the modality [Brain]

- A robust multi input fully convolutional neural network:
  - a) multiple input modalities *when* they are available
  - *b)* *not requiring* any specific input modalities
  - *c)* *overcomes* small registration errors between inputs
  - d) learns from a *variety* of sources



- We achieve this by learning a **modality invariant** latent representation

Chartsias et al IEEE TMI 2017; Joyce et al MICCAI 2017

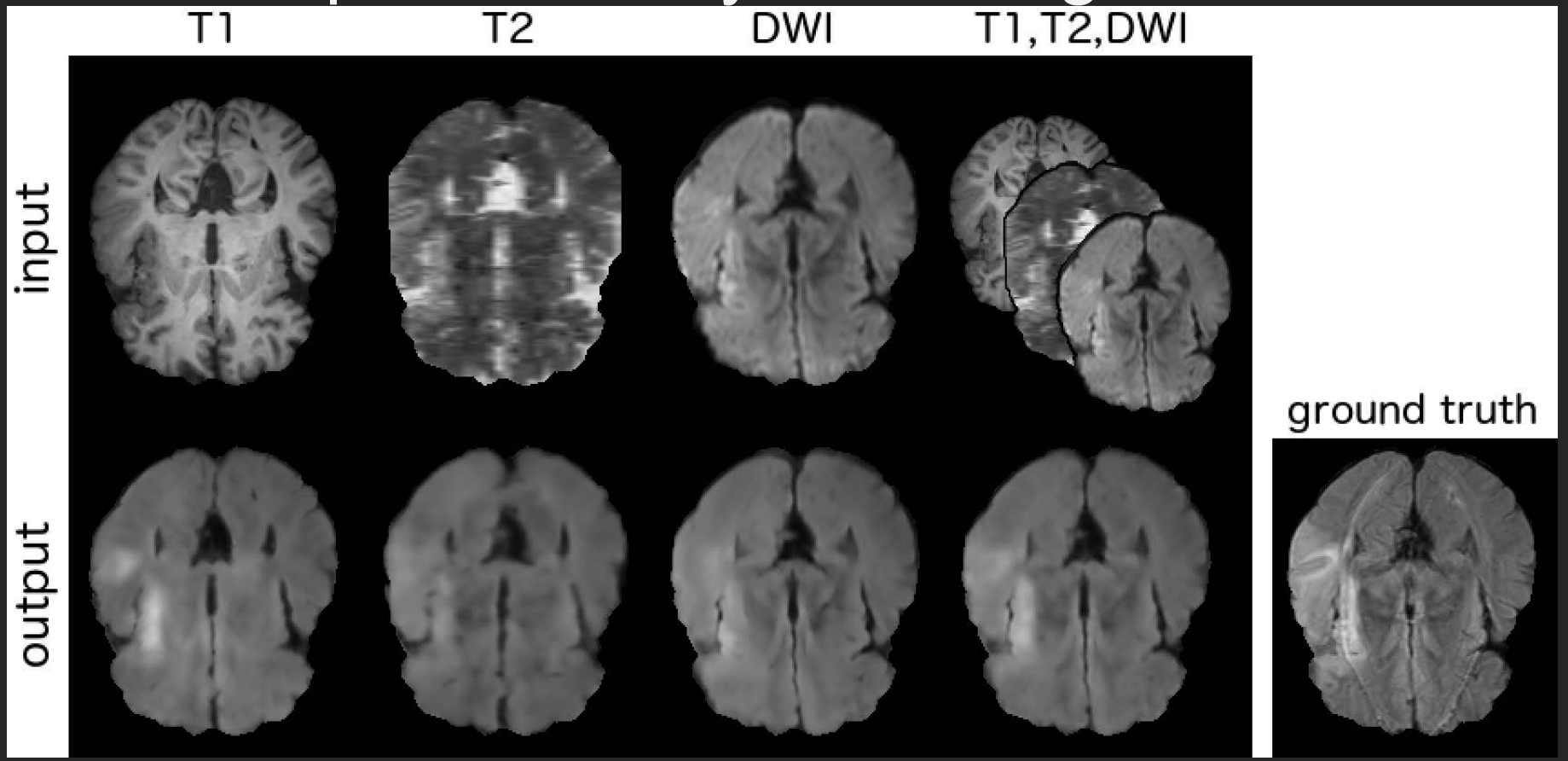


During inference, we process the inputs in four stages:

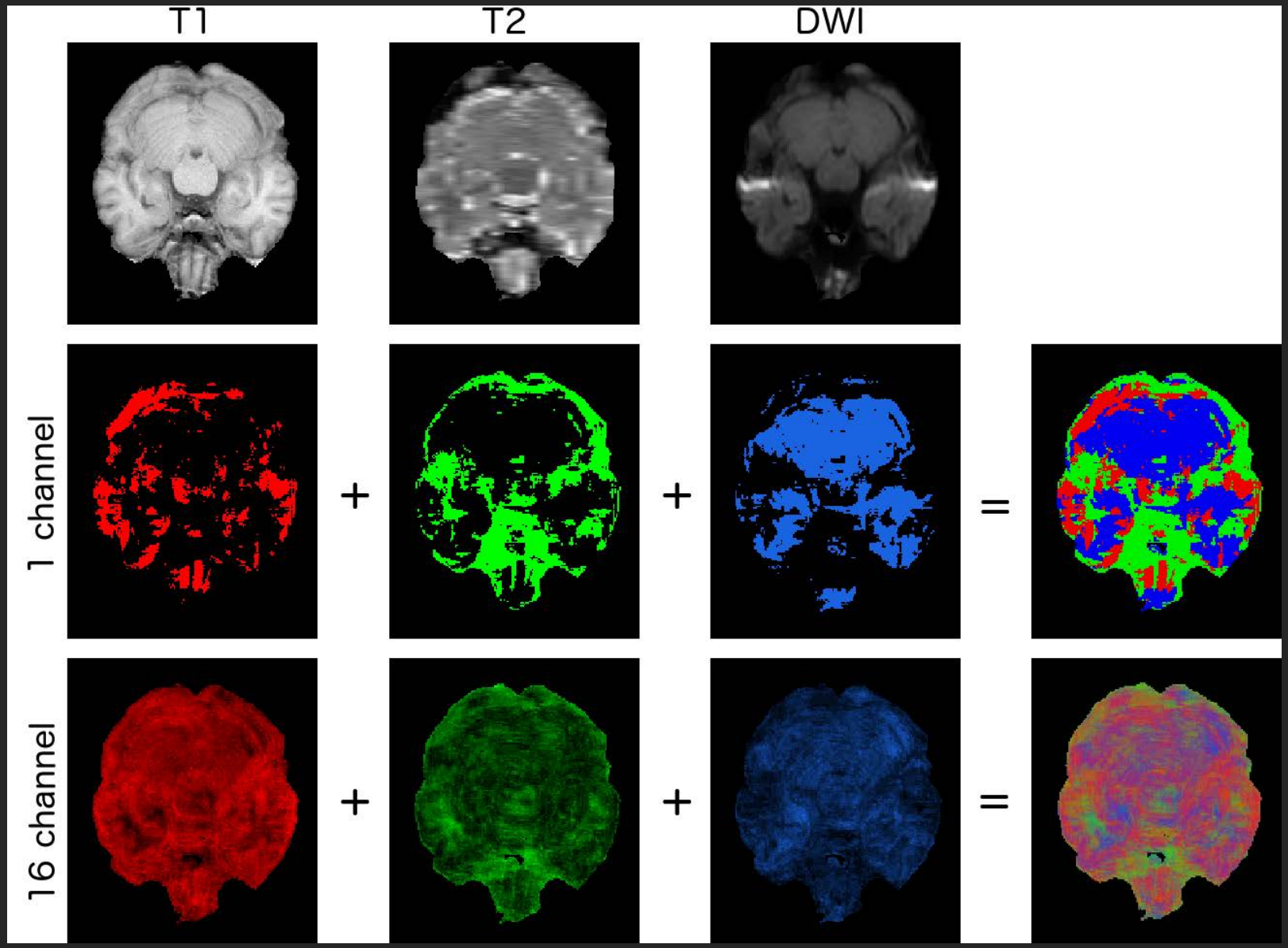
1. **Encode** inputs to latent representations (16-channel “images”)
2. Spatially **align** latent representations ➔ robust to misregistrations
3. **Fuse** latent representations into a fused representation
   ➔ combines information and makes robust to missing inputs
4. **Decodes** fused latent representation to target output modality

A visual example

- A multi-input model synthesizing FLAIR

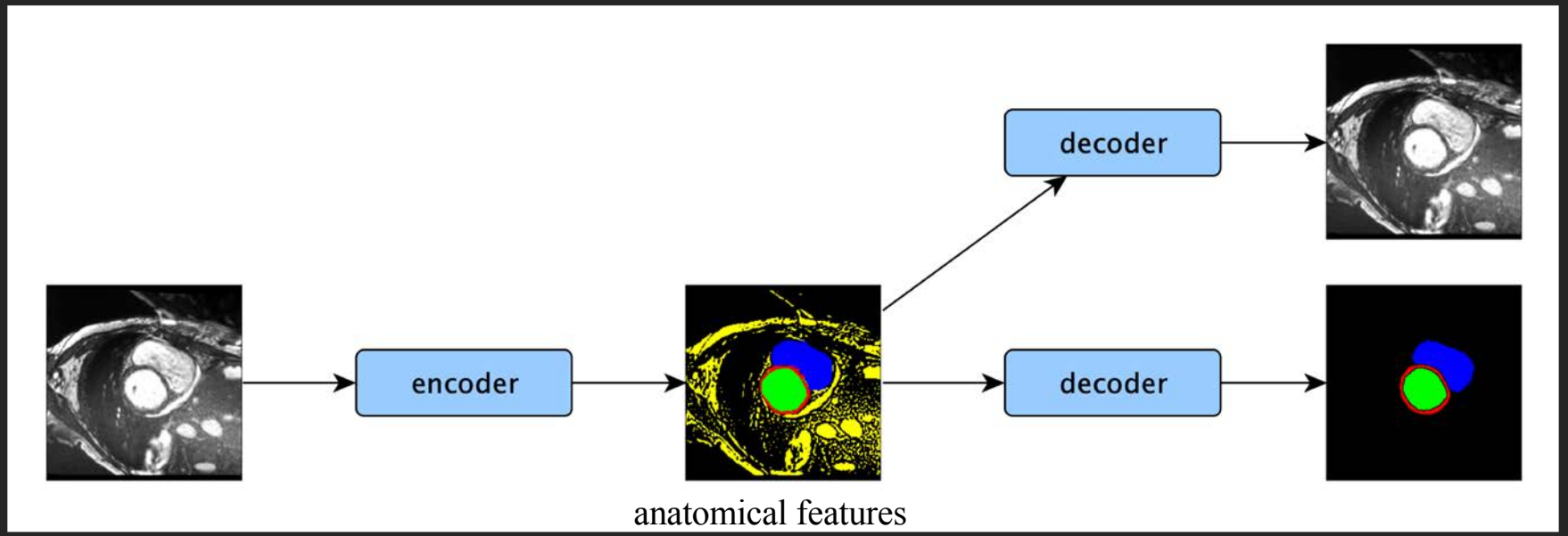


What uses from where?



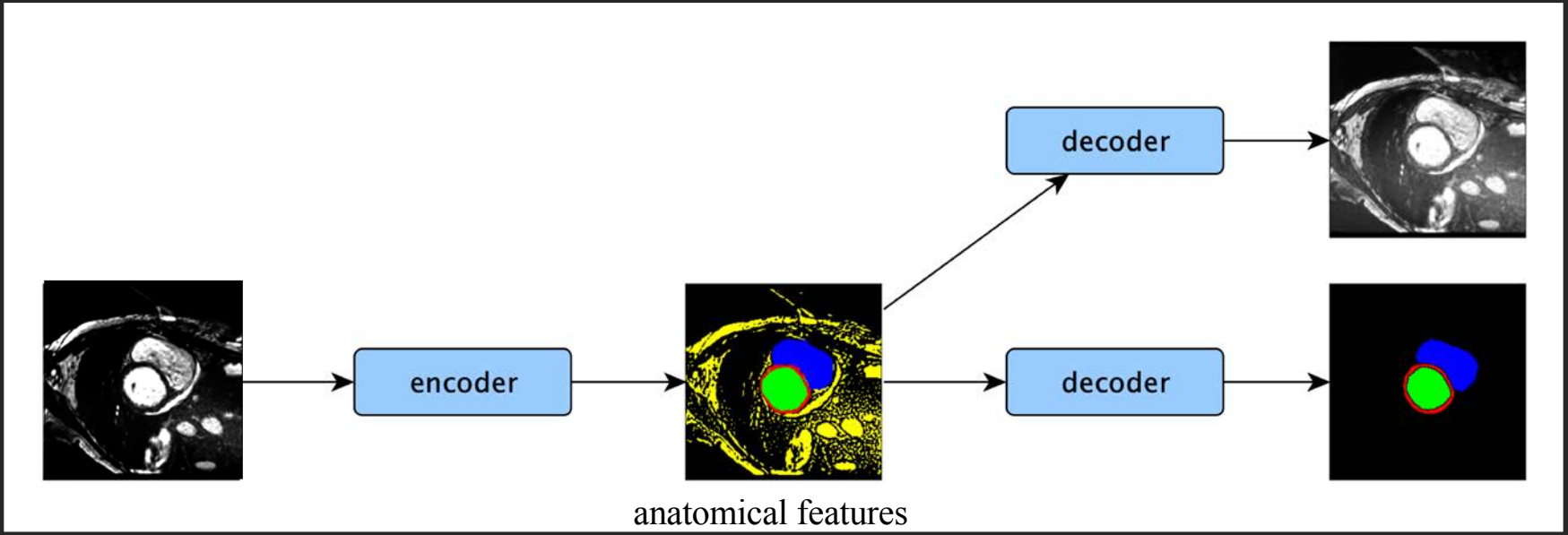
Representing images

- Representations could be a multiclass image of anatomy:
  - semantically meaningful
  - can be decoded into segmentation masks
- reconstruction ➔ use of non-annotated data (semi-supervised)



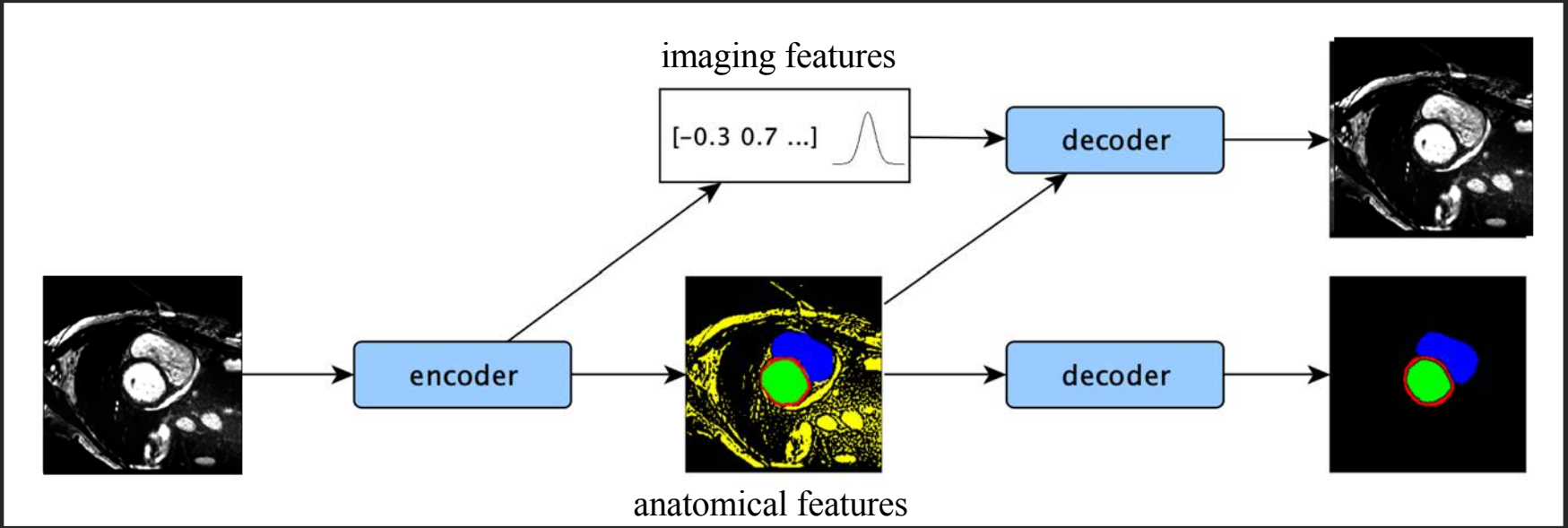
Start of a problem

But…
many images may have the same anatomical representation



Solving the problem…

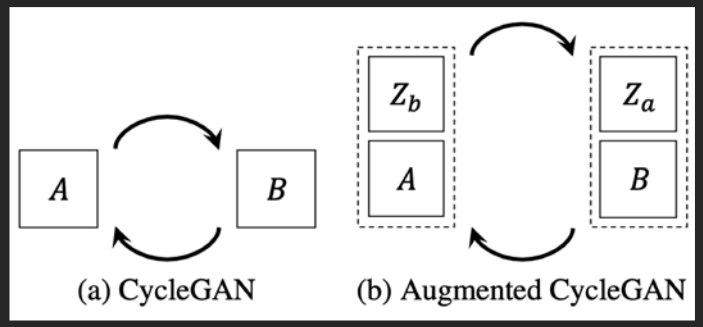
Add something that describes the image stats



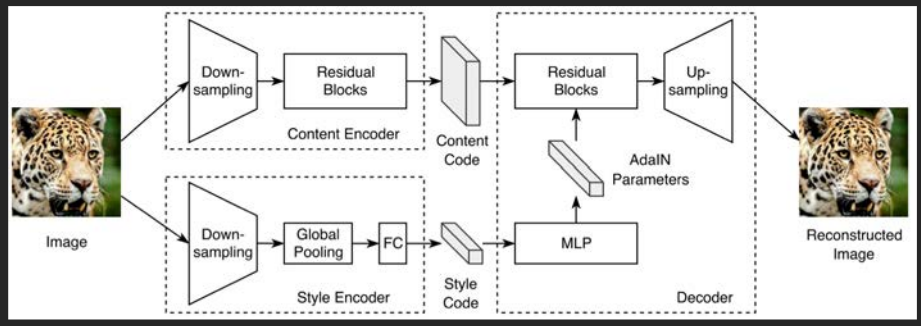
Chartsias et al MICCAI 2018; concurrent work with Almahairi et al ICML 2018; Huang et al ECCV 2018

In CV: content vs style disentanglement

- Concurrently with Agis [MICCAI 2018]

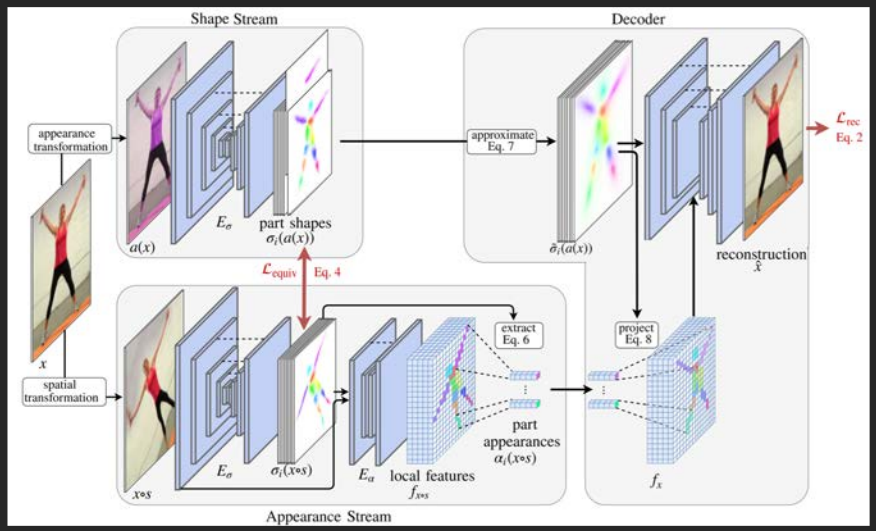


Almahairi et al ICML 2018



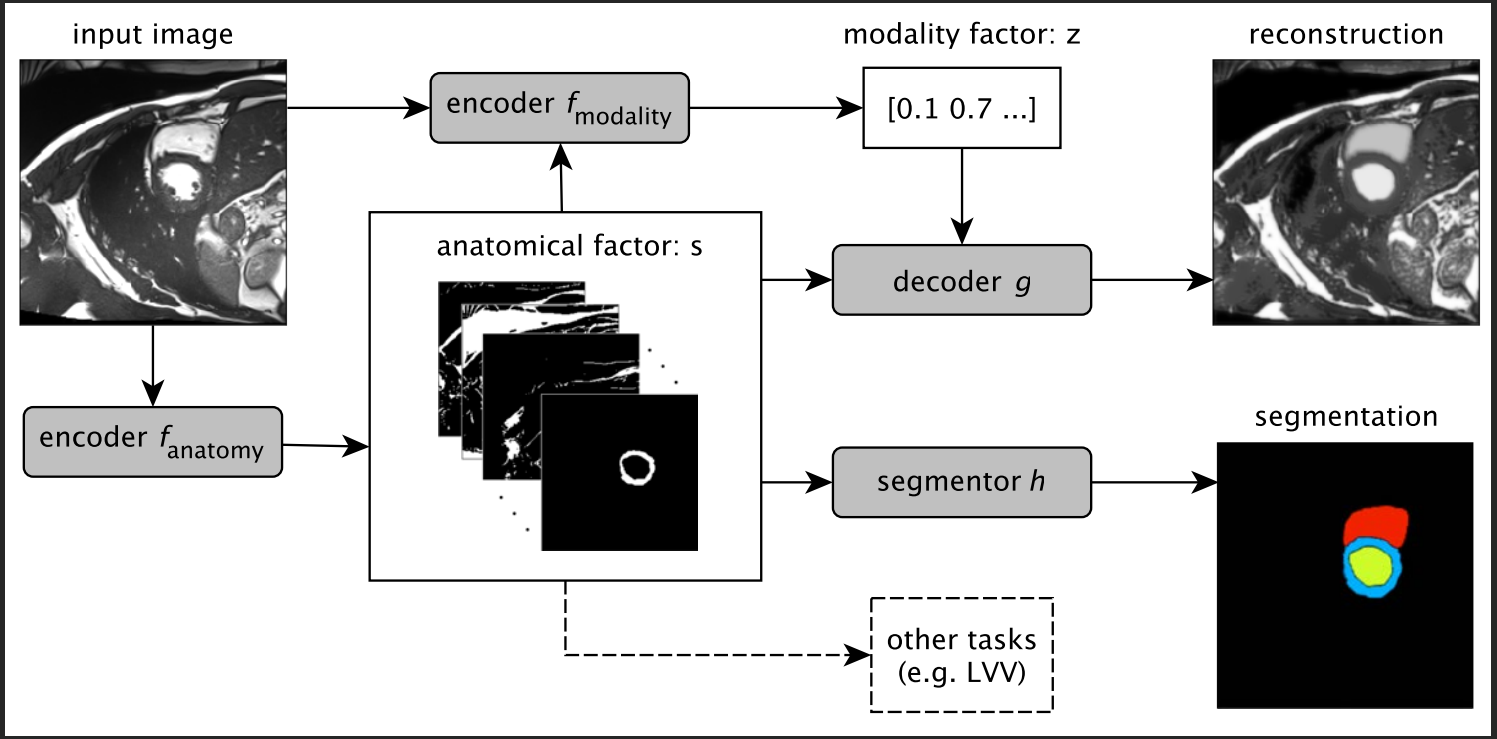
Huang et al ECCV 2018

- Since then, an explosion of approaches
- However, in our domain (medical) content must
  - have semantic meaning
  - serve quantitative purpose



Lorenz et al CVPR 2019

Disentangling the learning of anatomy

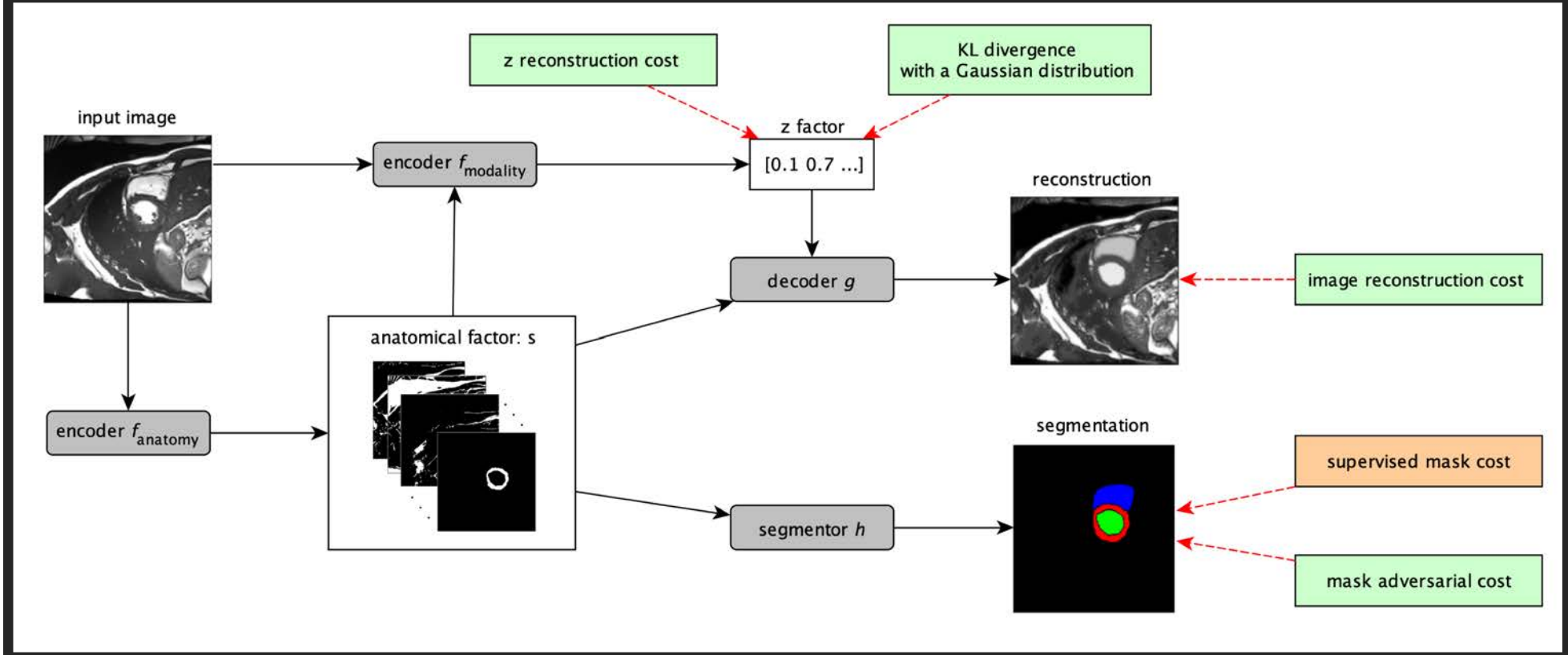


Images are decomposed into **spatial** and **vector** representations.

- **Spatial** representation is a segmentation mask of the myocardium.
- **Vector** representation contains intensity (appearance) information, and residual anatomical information.

which are then combined

Training



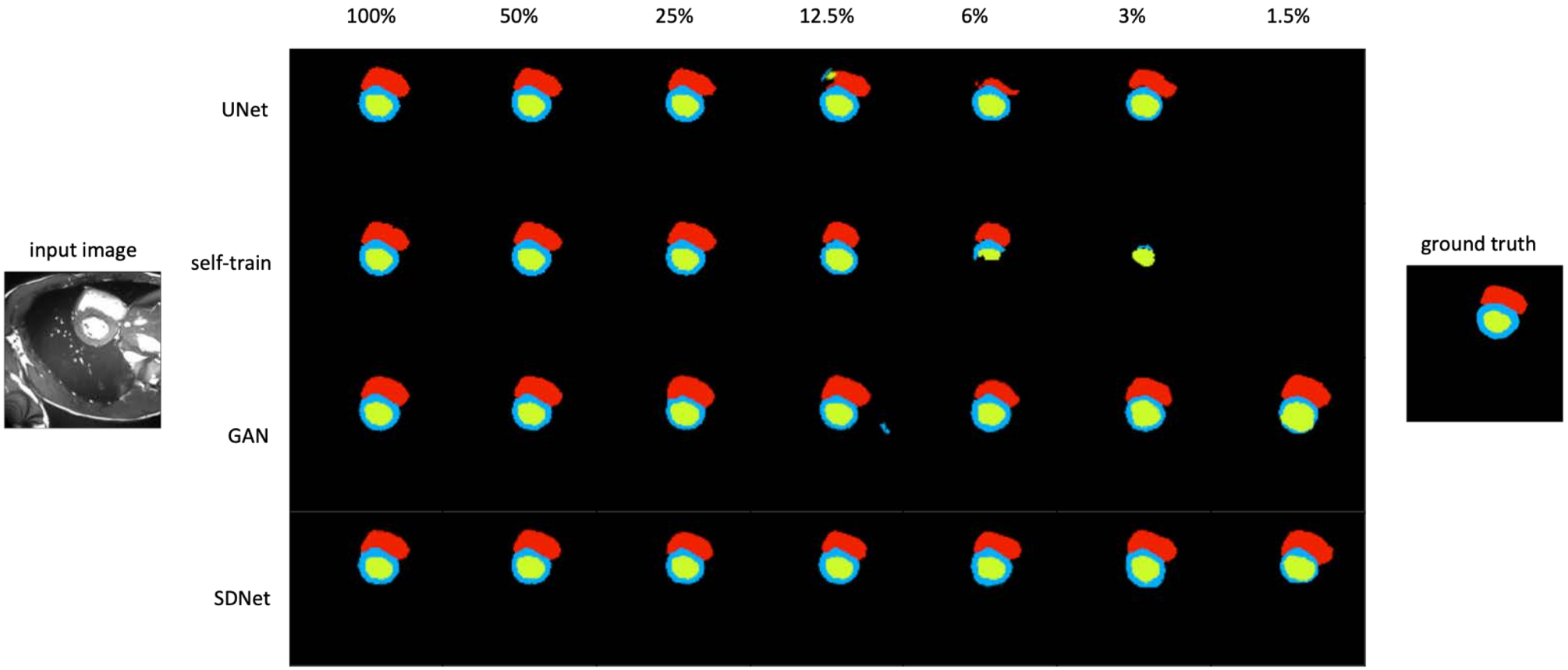
Unsupervised costs:

- Self-supervised image reconstruction
- Adversarial for mask discriminator
- Self-supervised z reconstruction
- z smoothness cost with KL divergence

Supervised costs:

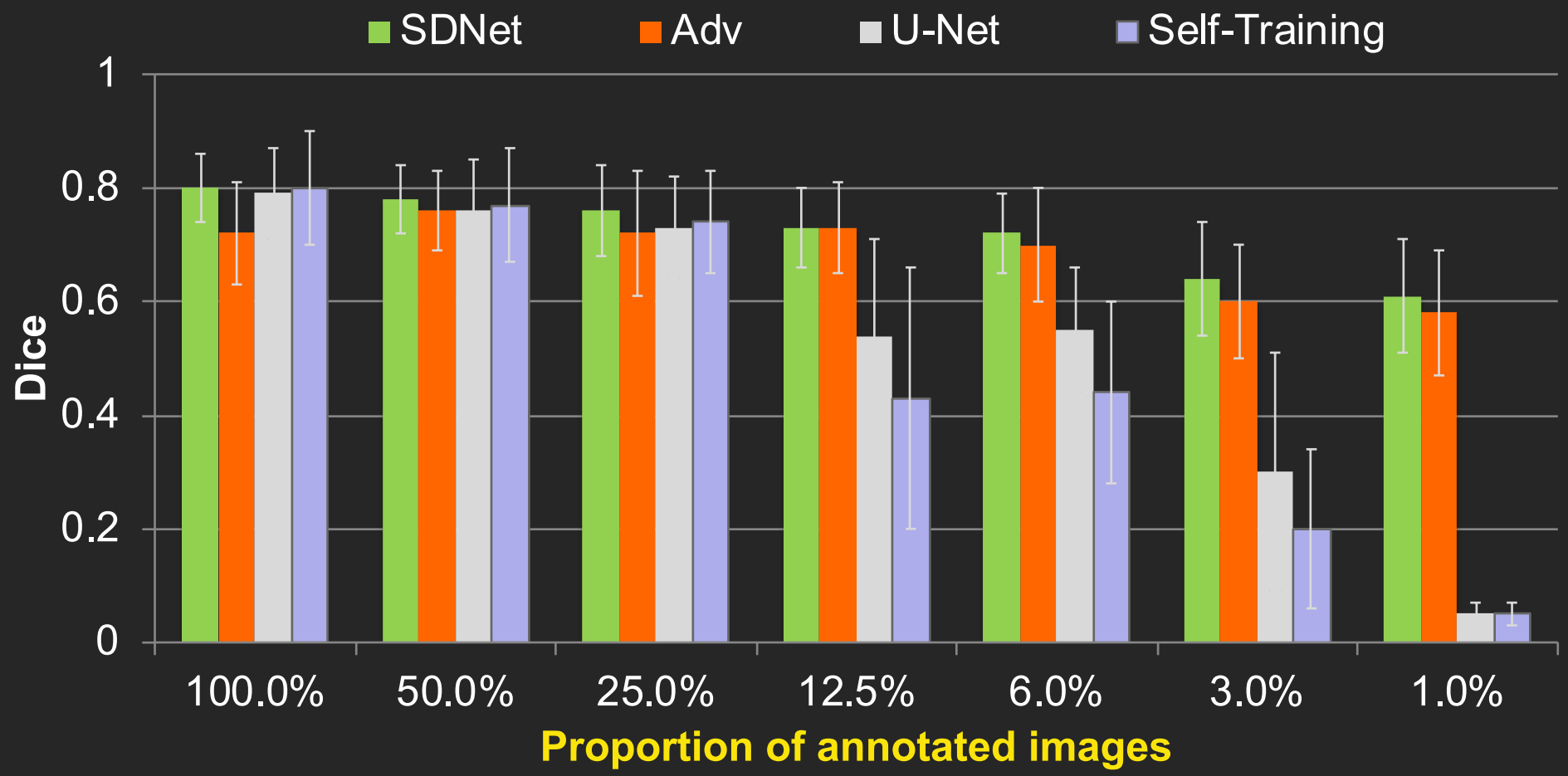
- Supervised cost between output mask and ground truth mask

A visual result



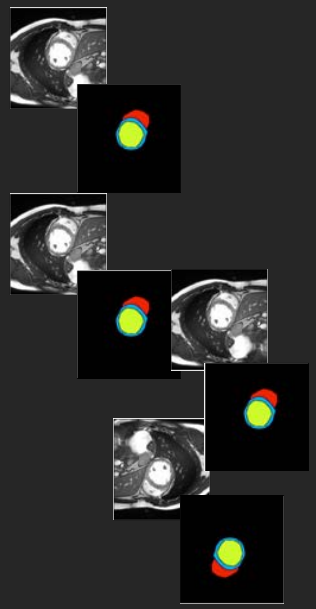
Source code: https://github.com/agis85/anatomy_modality_decomposition

Quantitative analysis (Dice, test set, N=400)

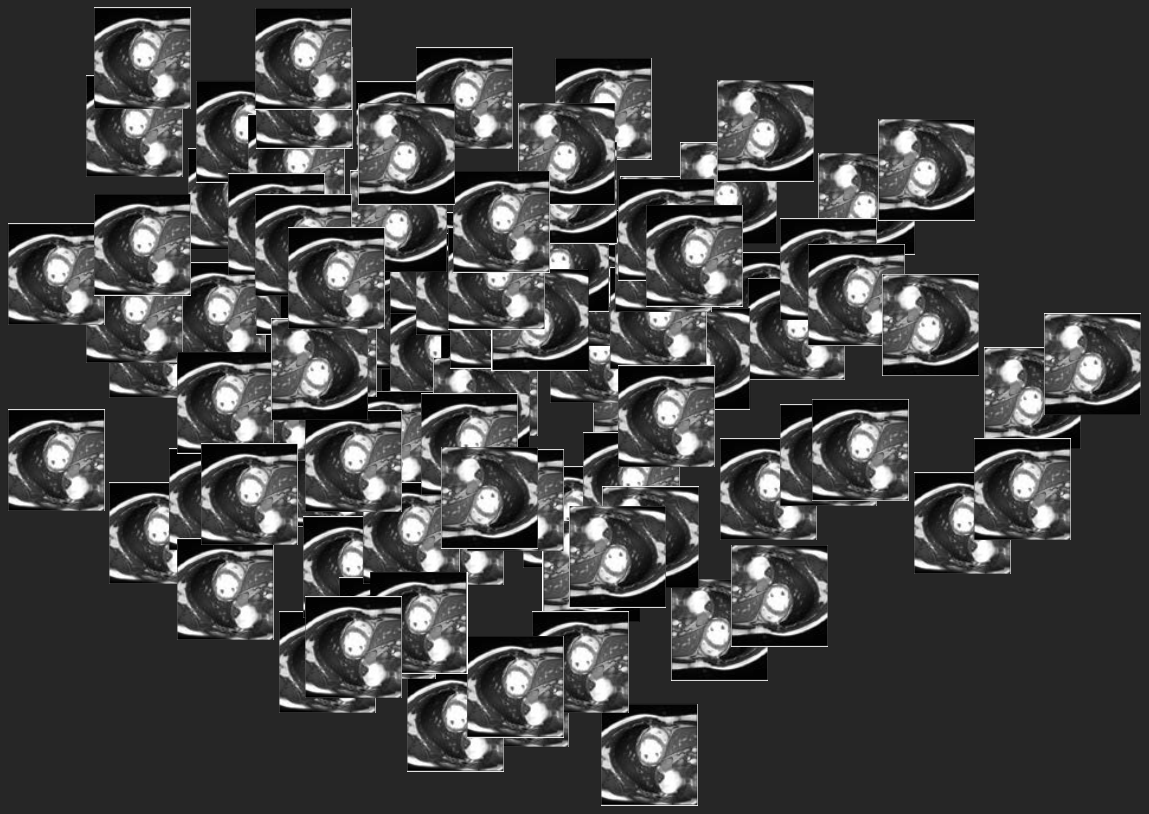


- labelled (% varies) ; 1200 unlabeled images.
- Good performance and low variance with a fraction of labelled images.

Doing more with less: benefit of using EHR



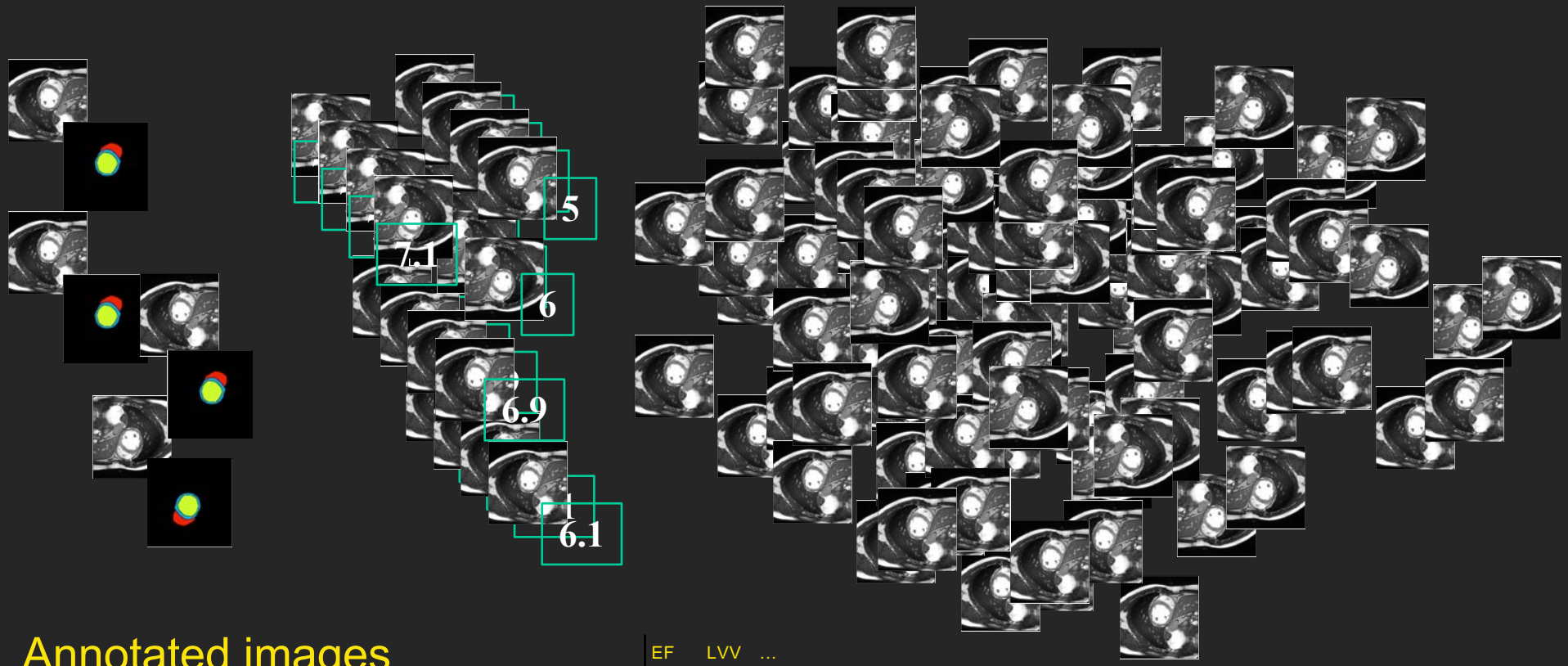
Annotated images
and masks



Other images

- Using 6% of annotated data: 0.756 (Dice)

Doing more with less: benefit of using EHR



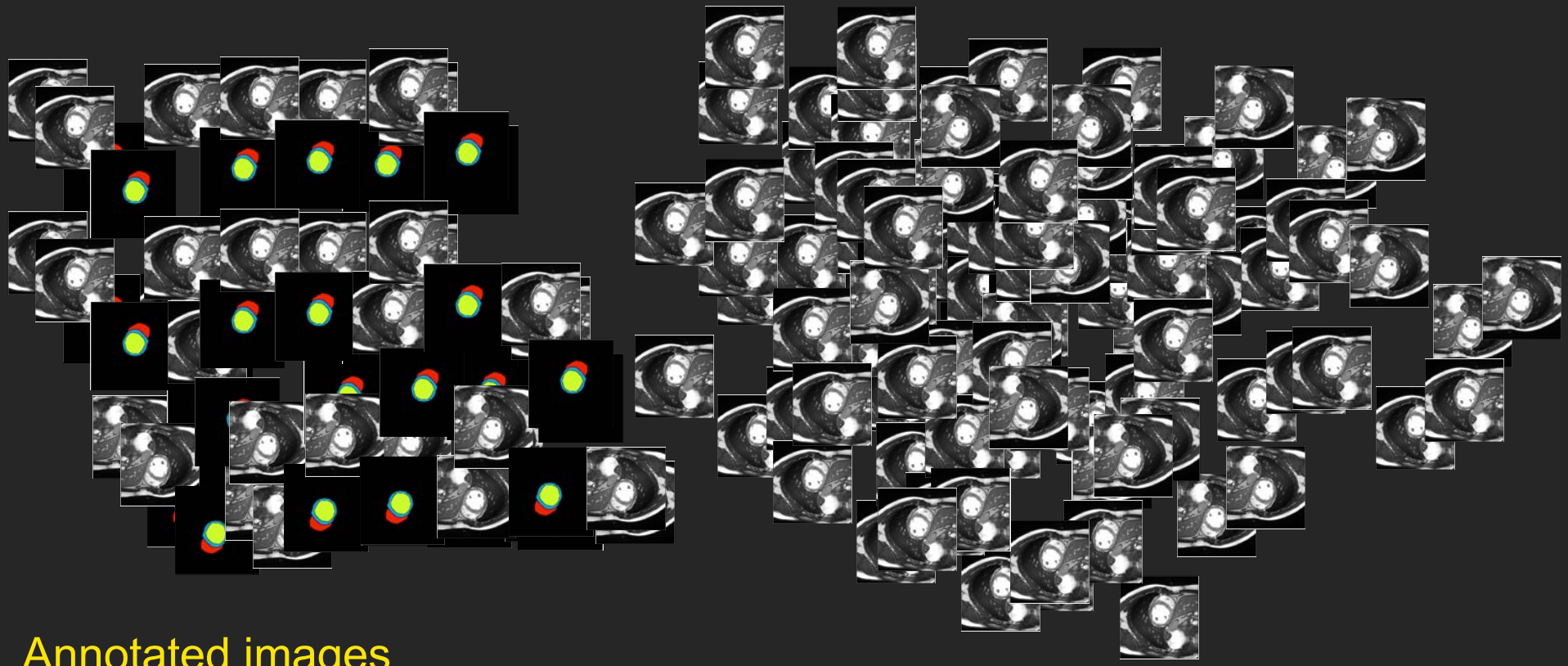
Annotated images and masks

	EF	LVV	...
image1	63%	6.1	8.5
image2	70%	6.8	9
...			

Other images

- Adding EHR data: 0.832 (vs. 0.756)

Doing more with less: use the EHR

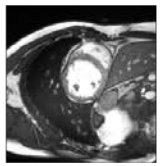
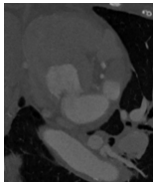
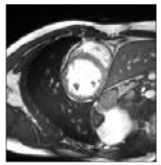
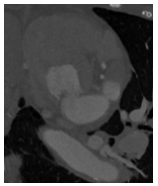


Annotated images
and masks

Other images

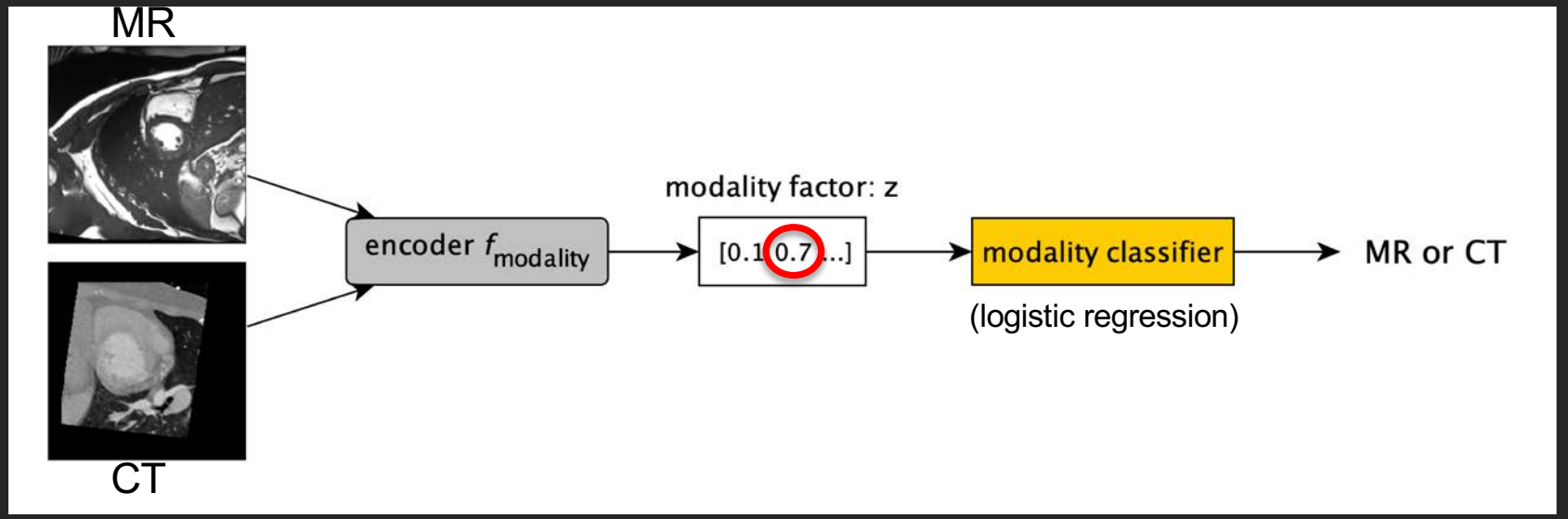
- Adding EHR data: 0.832
  (same gain as adding 8x more pixel annotation)

Doing more with less: pool imaging data

MR data	CT data	MR Dice	CT Dice
		0.78	0.80
		0.74	-
		-	0.77

Chartsias, Joyce, Papanastasiou, Semple, Williams, Newby, Dharmakumar, Tsaftaris, Disentangled representation learning in cardiac image analysis, Medical Image Analysis, 2019.

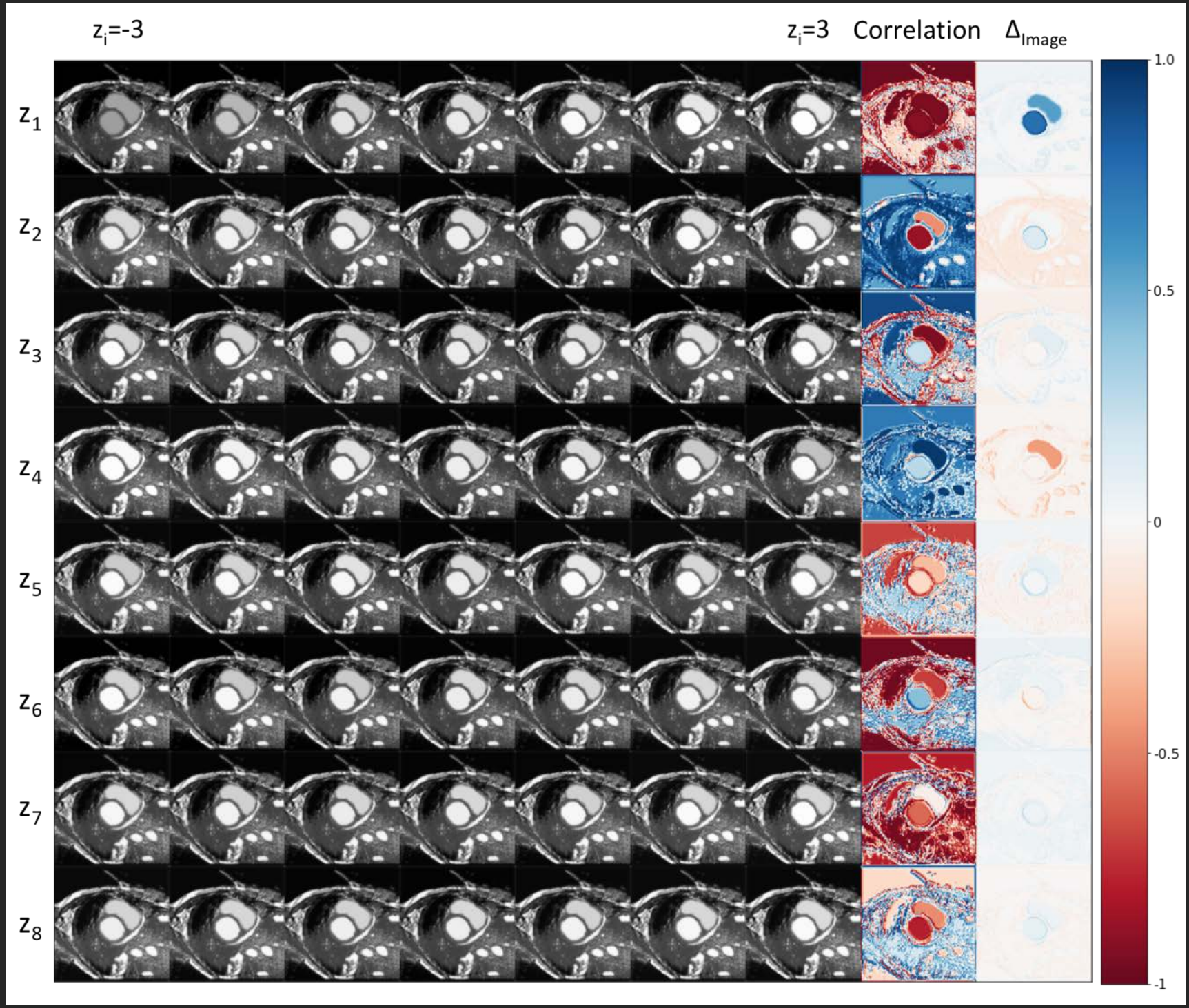
What we learn?



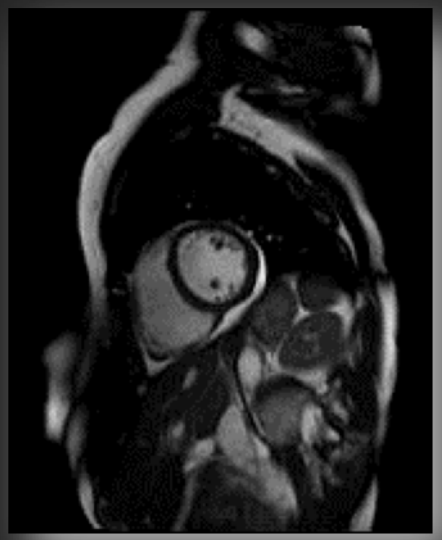
MR / CT: global intensity changes

➔ single z dimension captures most differences

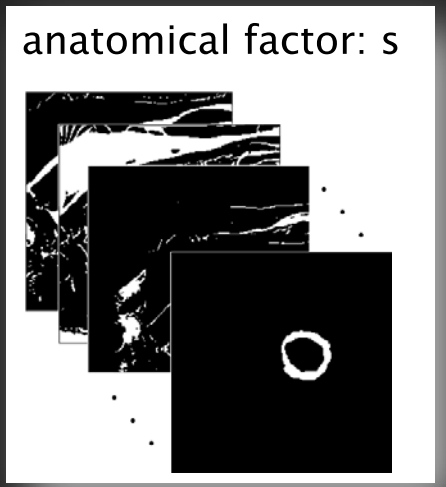
Factor dimensions used	Accuracy
All z vector	92%
Single element of z	82%


z_i=-3
z_i=3
Correlation
Δ_{Image}
z_1
z_2
z_3
z_4
z_5
z_6
z_7
z_8
1.0
0.5
0
-0.5
-1

Doing even more with less: use temporal correlation

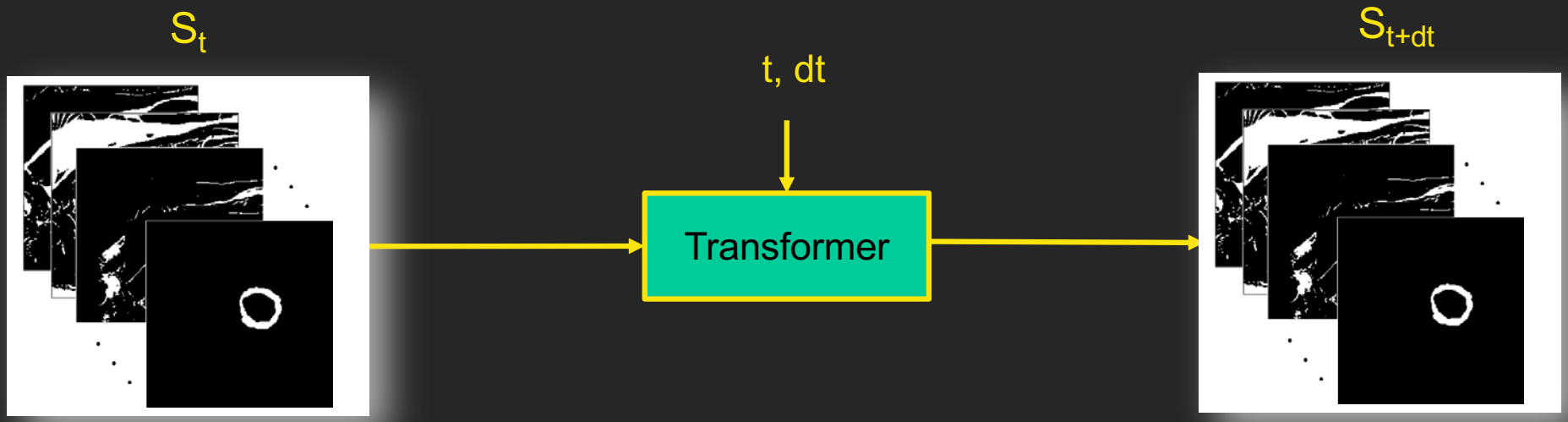


- Images close in time should have similar s
- Parts in s should move the same



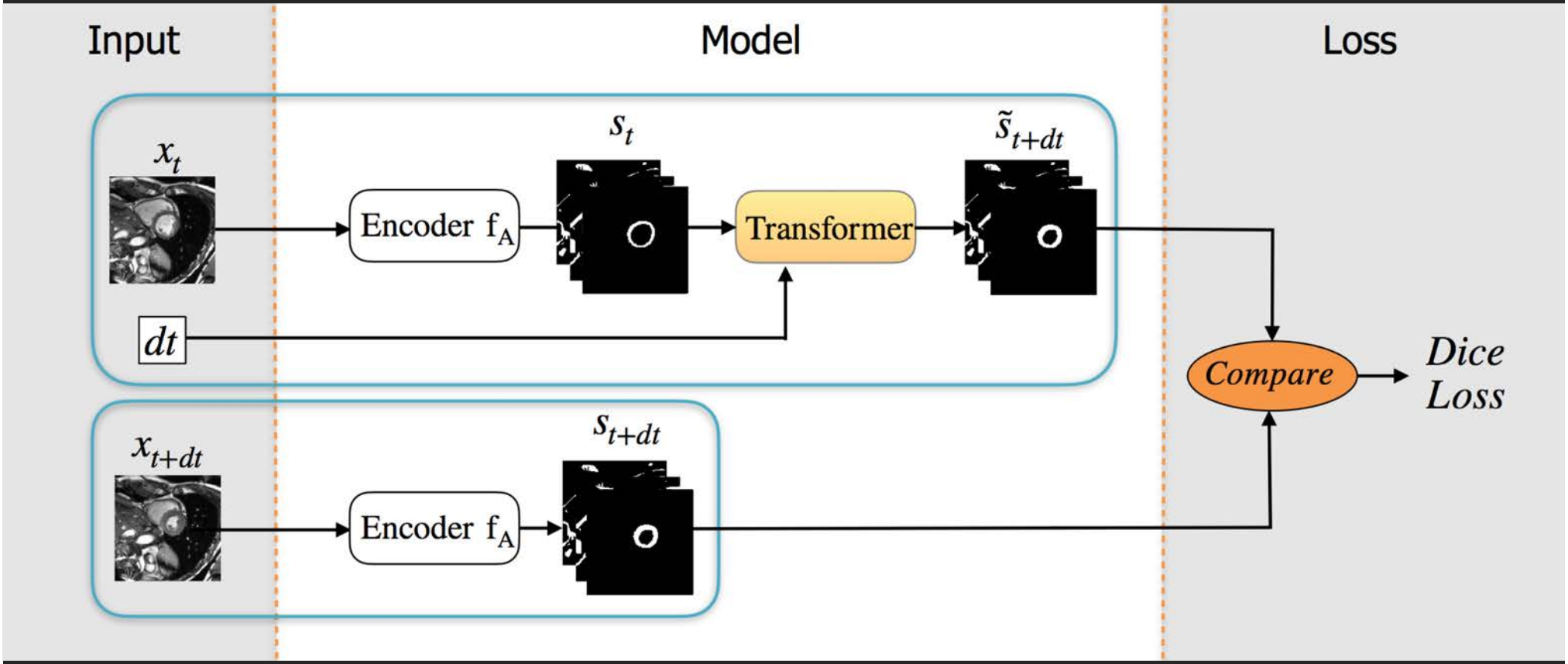
- Completely free → no annotations needed

Doing even more with less: use temporal correlation



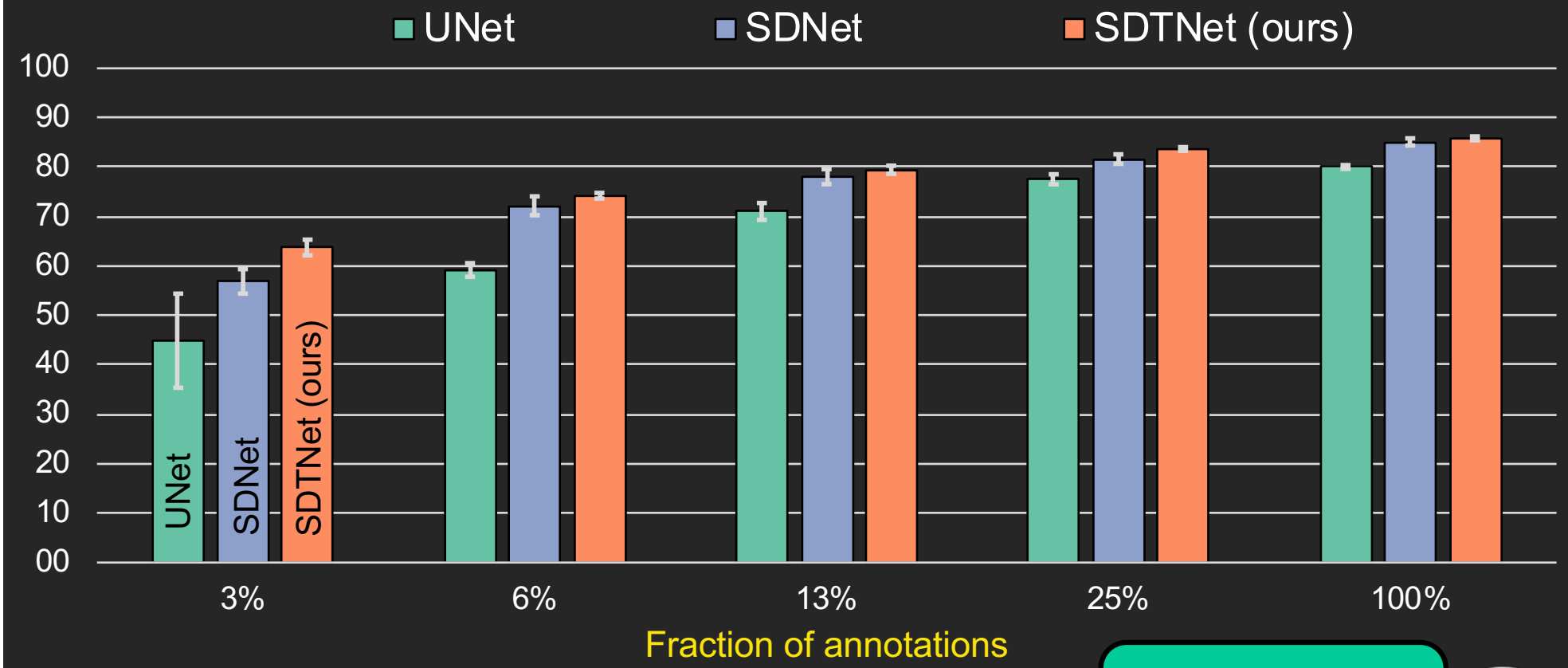
Valvano et al DART@MICCAI 2019, TODAY after 6pm

The model

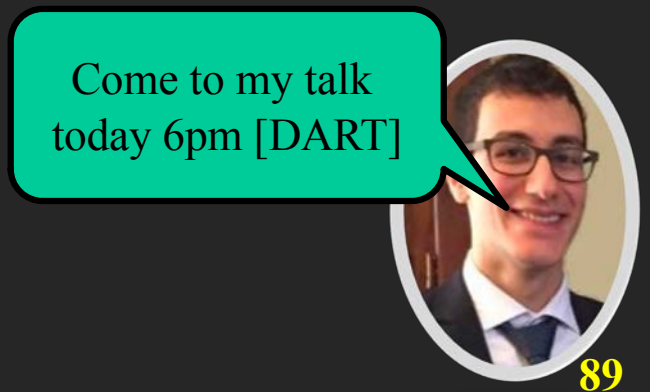


Source code: https://github.com/gvalvano/sdtnet

Results (Dice (%), test-set, ACDC)



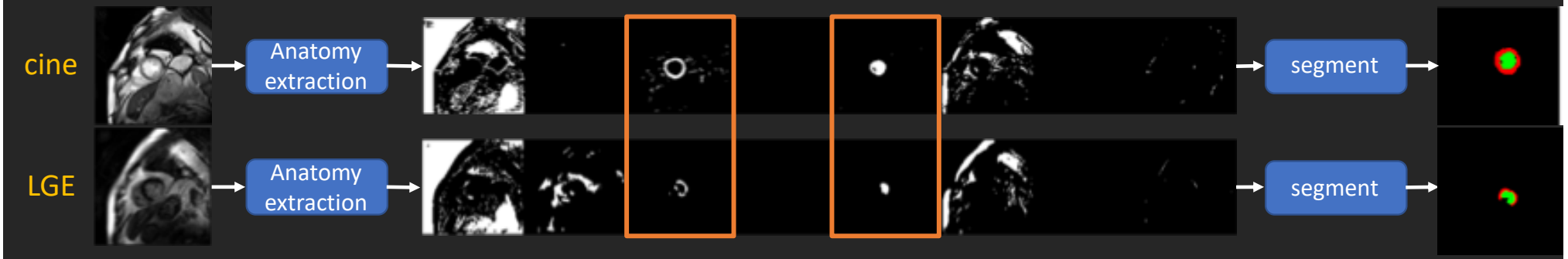
- Up to 7% gain using temporal information
- Compared to classical ML [Unet at 100%] same performance [SDTnet] at 12.5%

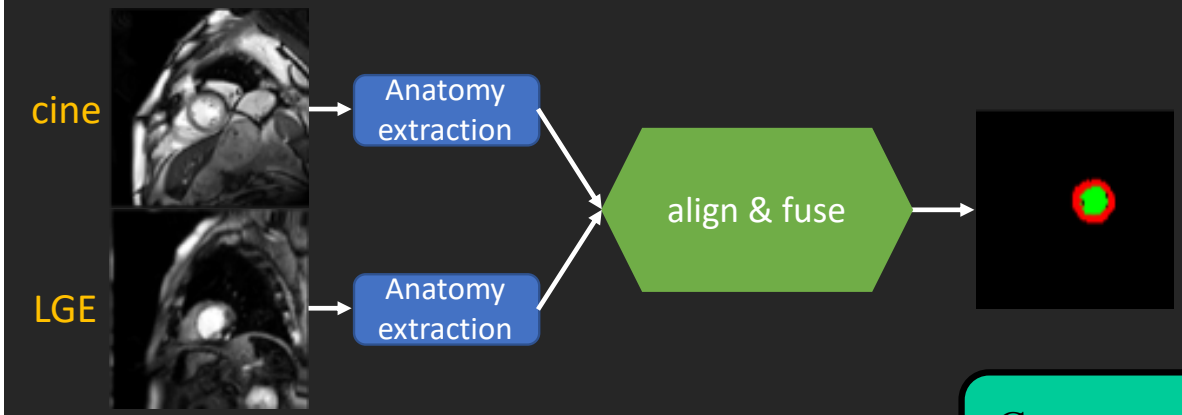


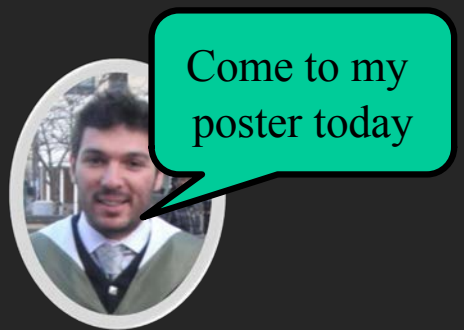
How radiologists review

https://www.healthcareittoday.com/wp-content/uploads/2018/12/1197119827-radiology-imaging-command-center-scans-screens.jpg

Multimodal inference







- Understand and relate different imaging inputs
- Exploit correlation across inputs and combine information
- We can train models with 0% annotations for LGE.

BEYOND STYLE-CONTENT

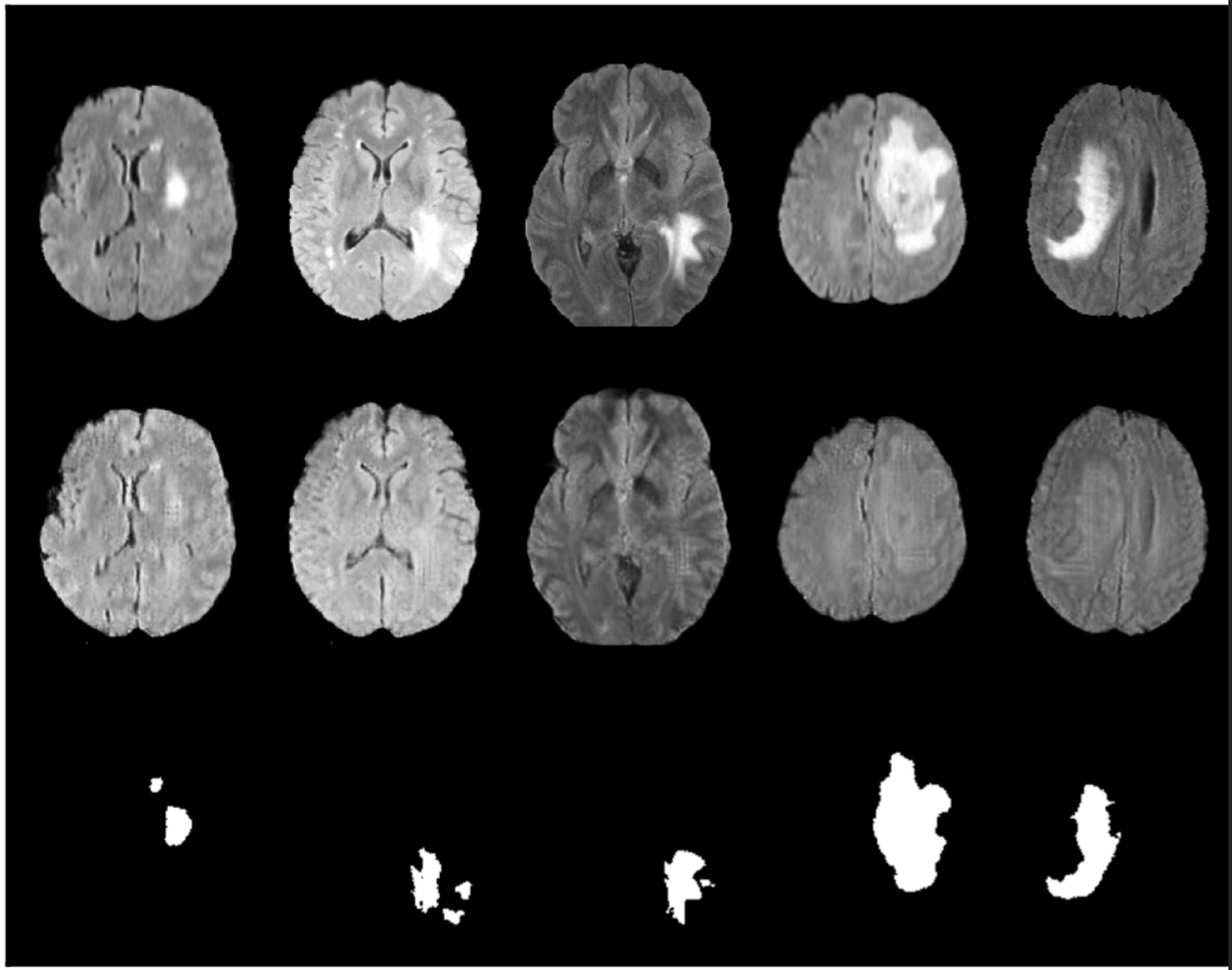
Disentangling pathology and age

What happens if we have ”pathology” somewhere?

Pathological images

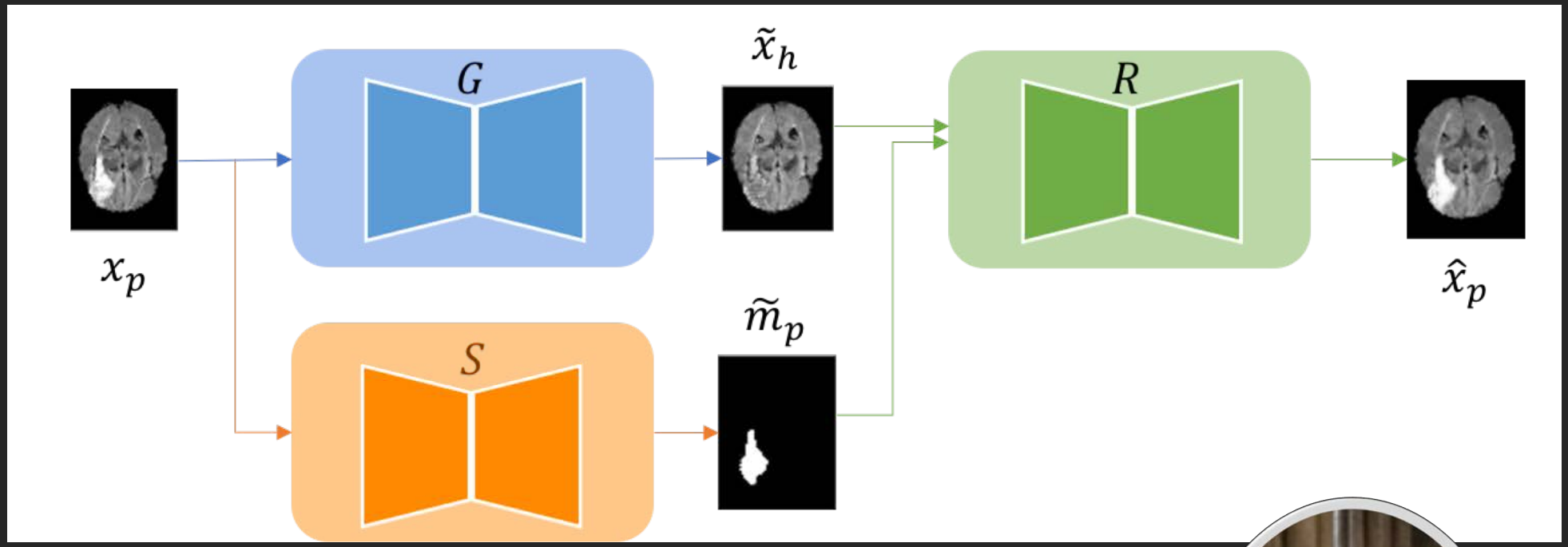
Pseudo healthy images

Pathology mask



Xia et al MIDL 2019 (Best paper runner up); Xia et al MedIA submitted.

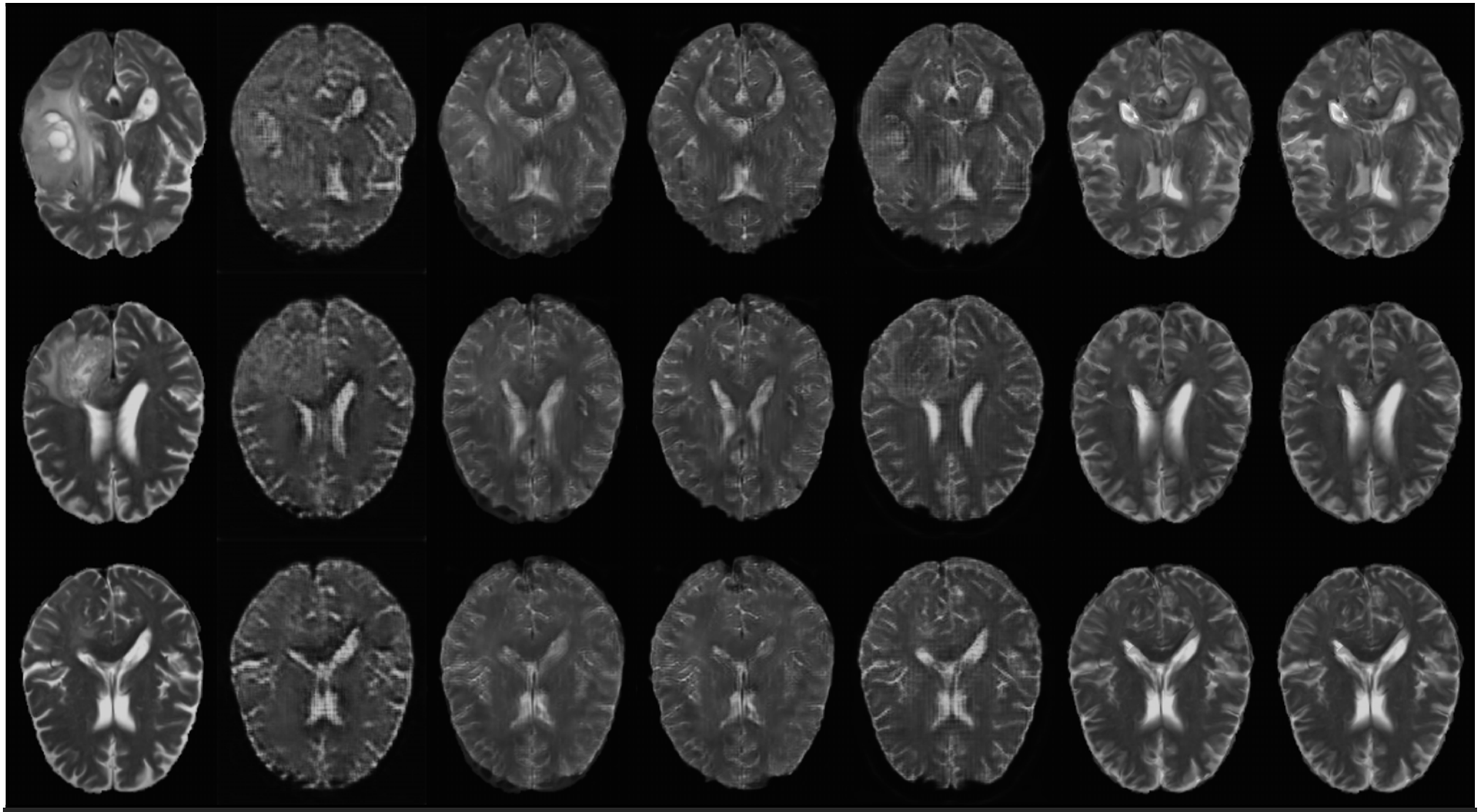
Disentangle pathology



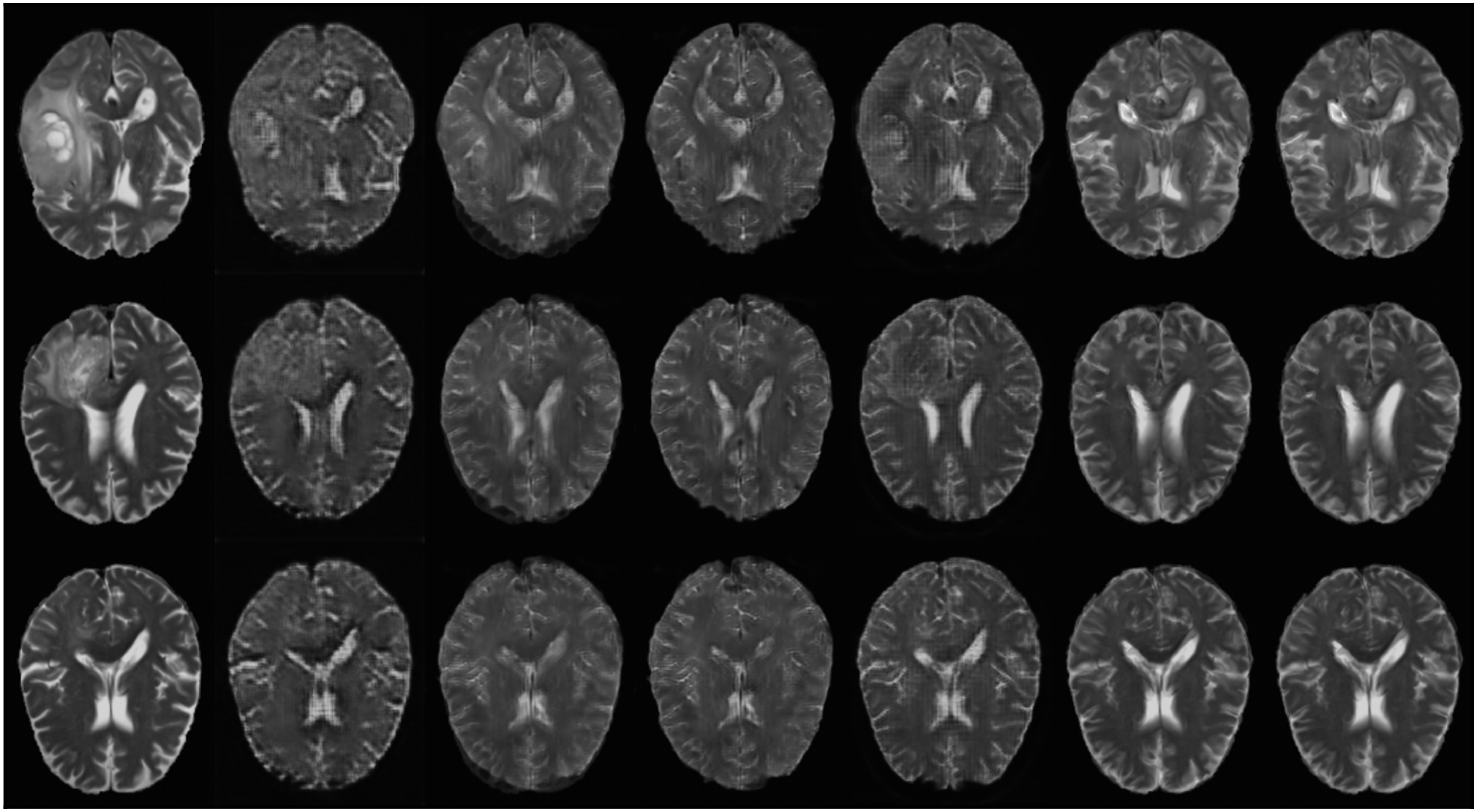
- Can we learn to separate pathology?

Xia et al MIDL 2019 (Best paper runner up); Xia et al MedIA submitted.

Visual results



Visual results



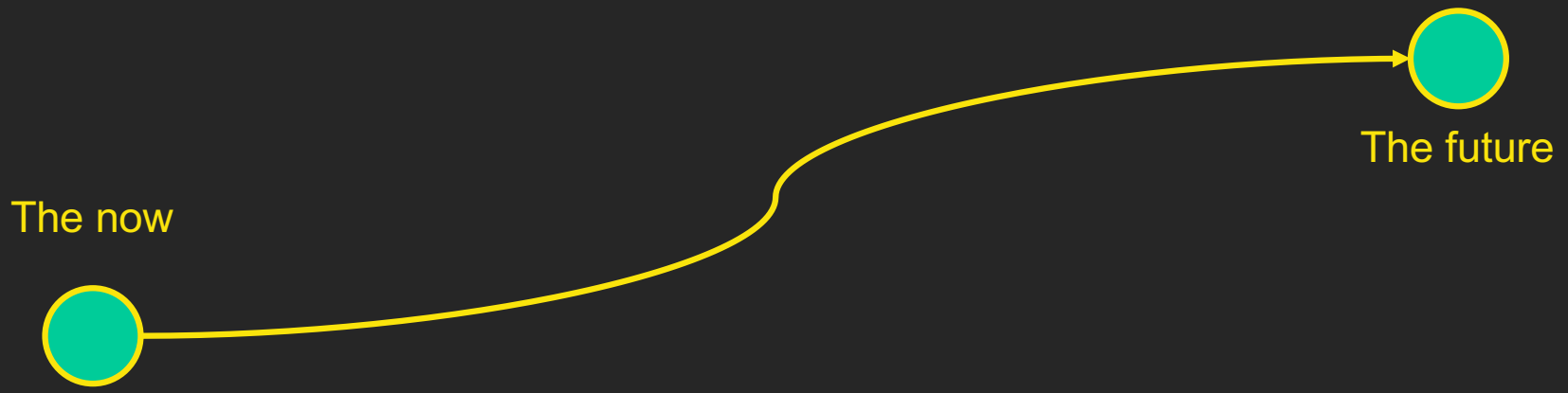
Pathological images	AAE	vaGAN	Conditional GAN	CycleGAN	Ours (unpaired)	Ours (paired)

Preserve identity and make it look healthy

Method	T1			T2			T2 (human evaluation)		
	iD	*h*	*DeC*	*iD*	*h*	*DeC*	'identity'	'healthiness'	'def. corr.'
AAE	$0.65_{0.12}$	$0.79_{0.12}$	$0.73_{0.05}$	$0.63_{0.12}$	$0.81_{0.11}$	$0.78_{0.04}$	$0.54_{0.12}$	$0.43_{0.13}$	$0.38_{0.11}$
vaGAN	$0.72_{0.11}$	$0.88_{0.07}$	$0.86_{0.06}$	$0.74_{0.10}$	$0.88_{0.09}$	$0.84_{0.05}$	$0.57_{0.12}$	$0.65_{0.14}$	$0.61_{0.12}$
conditional GAN	$0.70_{0.14}$	$0.81_{0.11}$	$0.85_{0.04}$	$0.69_{0.09}$	$0.84_{0.10}$	$0.86_{0.04}$	$0.44_{0.12}$	$0.55_{0.11}$	$0.58_{0.10}$
CycleGAN	$0.82_{0.08}$	$0.88_{0.08}$	$0.75_{0.09}$	$0.81_{0.07}$	$0.86_{0.07}$	$0.77_{0.06}$	$0.58_{0.11}$	$0.49_{0.13}$	$0.40_{0.09}$
Ours (unpaired)	$\mathbf{0.84}_{0.08}$	$0.90_{0.07}$	$\mathbf{0.93}^{*}_{0.04}$	$0.83_{0.06}$	$0.96^{*}_{0.03}$	$0.90^{*}_{0.05}$	$0.68^{*}_{0.14}$	$0.79^{*}_{0.12}$	$0.70^{*}_{0.11}$
Ours (paired)	$0.83_{0.06}$	$\mathbf{0.95}^{*}_{0.06}$	$0.92^{*}_{0.05}$	$\mathbf{0.85}^{*}_{0.04}$	$\mathbf{0.97}^{*}_{0.04}$	$\mathbf{0.91}^{*}_{0.04}$	$\mathbf{0.72}^{*}_{0.13}$	$\mathbf{0.80}^{*}_{0.14}$	$\mathbf{0.72}^{*}_{0.13}$

- iD: Identity is the "pseudo-healthy" image of the same subject?
  - A masked SSIM loss
- *h*: Healthiness how healthy is the pseudo-healthy?
  - An off-the shelf disease segmentor
- *DeC*: does it correct deformations from pathology?
- Beats baselines even without input-mask (paired) annotations!

Predicting future health status

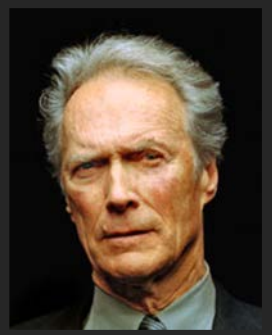
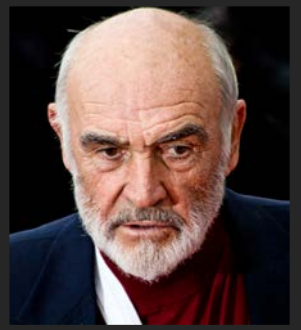


- Given my state now and some knowledge

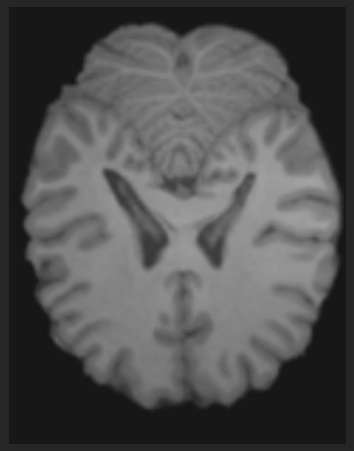
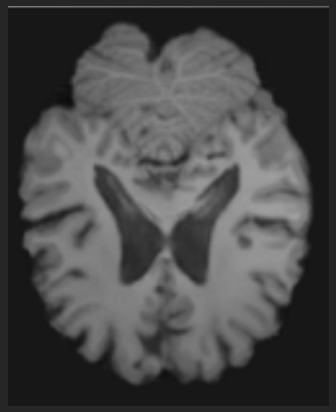
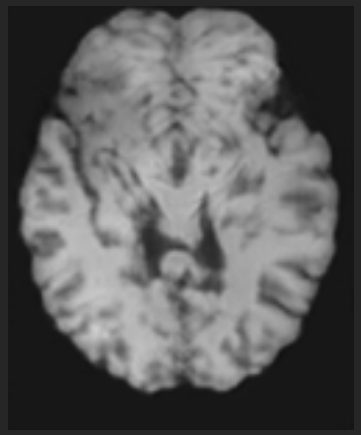
… How will I be in the future?

A simpler proxy: Learning to age

- How would I look in 30 years?

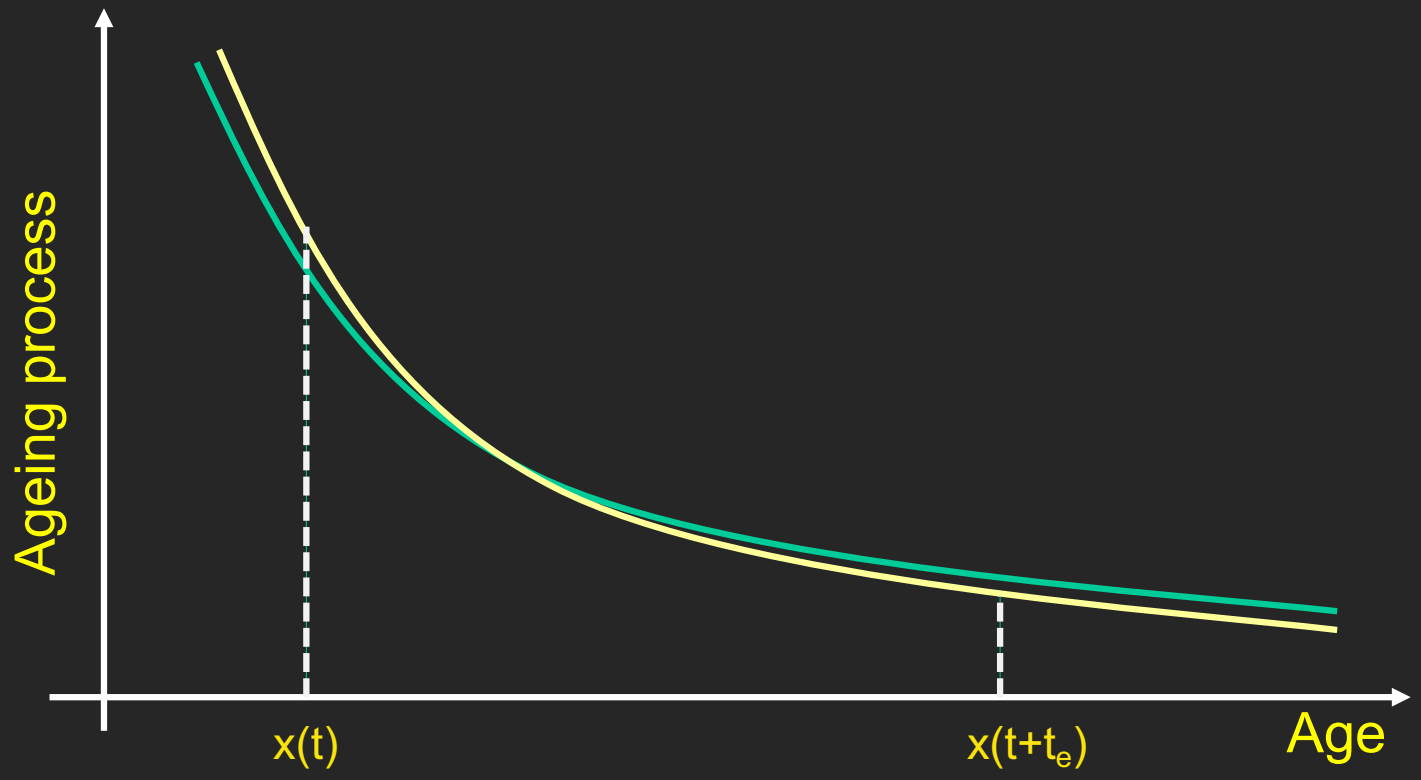



- A much harder task is how would my brain look?

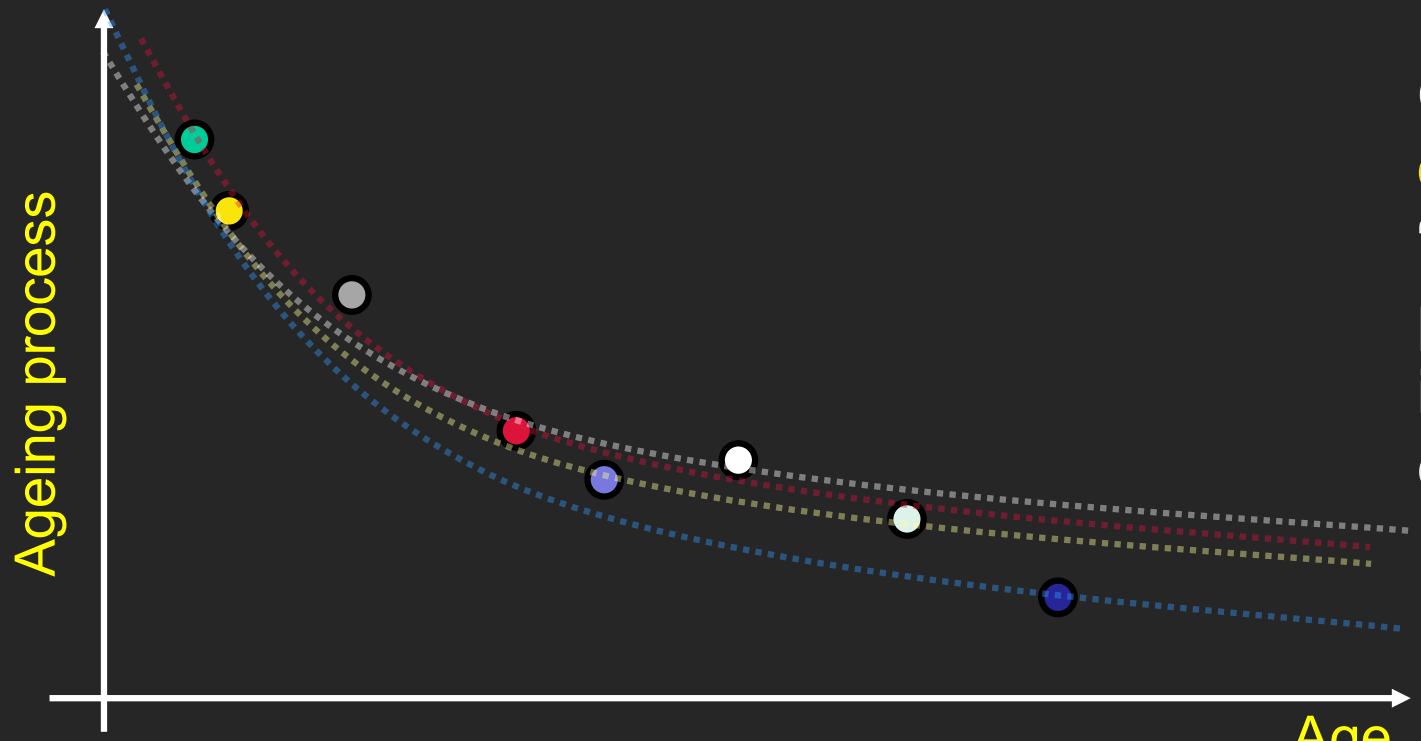
What we need to solve…

Learn auto-regressive functions $x(t+t_e) = \mathbf{f}(x(t), t_e)$ but …



What we need to solve…

Learn auto-regressive functions $x(t+t_e) = \mathbf{f}(x(t), t_e)$ but … we don't have longitudinal observations

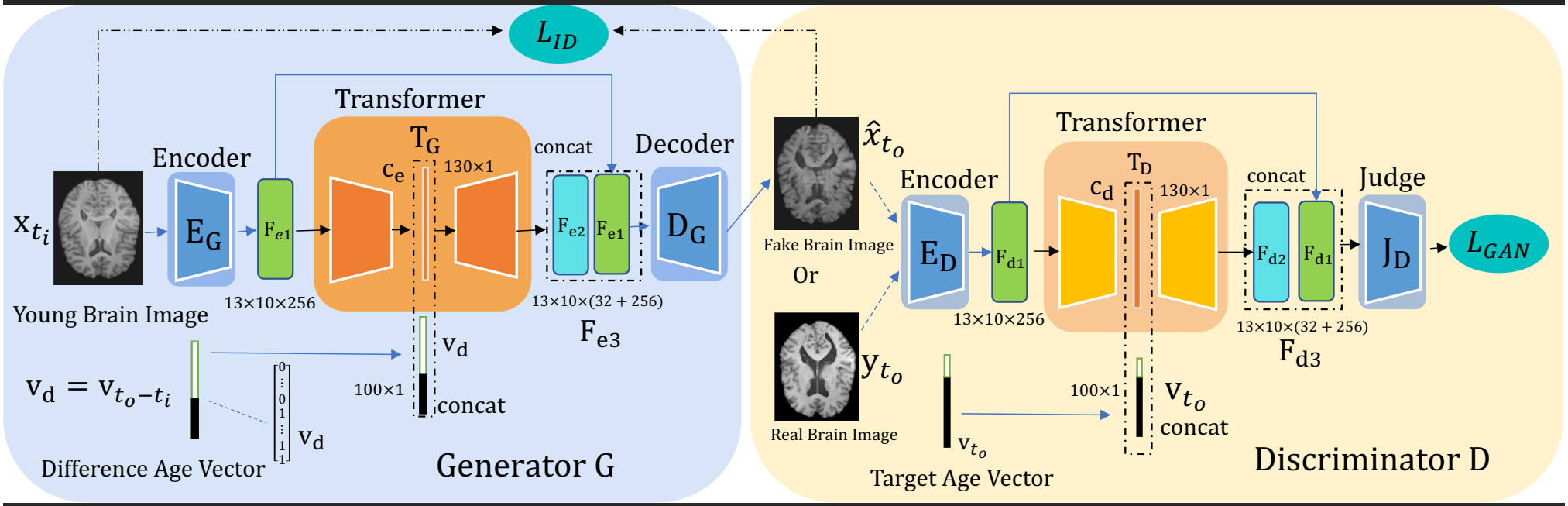


Can we disentangle "the ageing process" from individual characteristics?

Come to my poster MON 1pm [M-2-M-183]

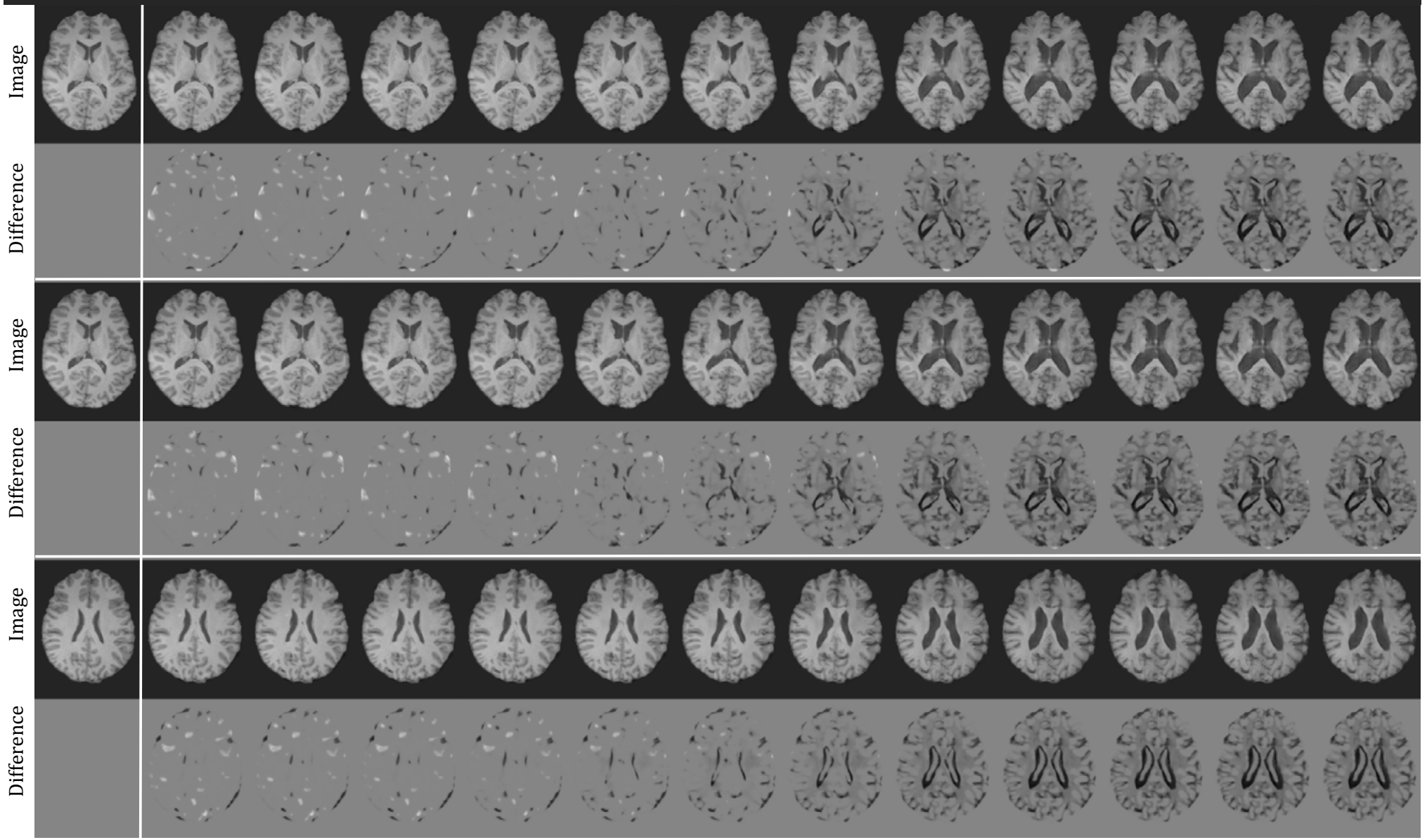
Xia et al MICCAI 2019 Monday 1-2pm M-2-M-183

The model



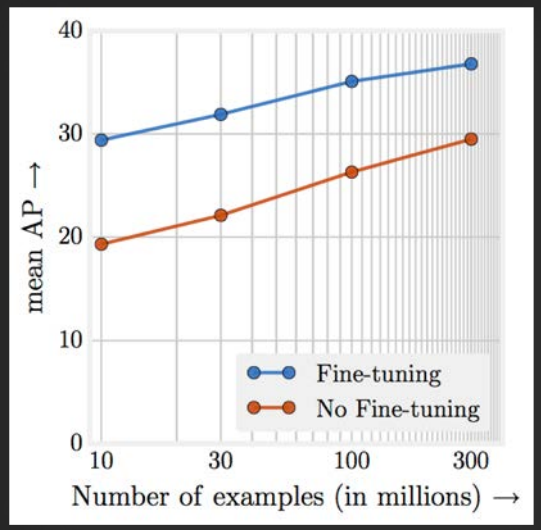
- Key ingredients:
  - conditional GANs learn joint distribution of brain appearance and age
  - ordinal encoding for conditioning
  - age-modulated identity regularization
    - Output and input differ more if age difference is larger

Visual results



CHALLENGES

Supervised learning has limits



We may **never** have enough data

Sun et al ICCV 2017

Mahajan et al ECCV 2018

The curse of compositionality

Camera Pose(4):
azimuth
elevation
tilt
distance

Lighting(4):
# of light sources
type
position
color

Texture(1)

Material(1)

Scene Layout(3):
background
foreground
position

Fig. 8 An illustration of combinatorial explosion. We consider the (already simplified) rendering process of one object. It involves choosing the camera pose, lighting condition, object texture, etc: a total of (merely) 13 parameters. If we allow 1,000 different values for each parameter, then we obtain a total of 10^{39} different images. This is way beyond the size of any dataset, as well as the number of images humans see per year.

10^{39}

Disentanglement will help address this

Yuille & Yu 2019

Where is disentanglement going…

- Universal unsupervised disentanglement
  ➔ does not exist
- Needs ”restrictions”:
  - annotated data or
  - inductive biases
- Usually involves many costs
- Models on VAE are popular
- Style-content disentanglement is harder

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Francesco Locatello [1 2] Stefan Bauer [2] Mario Lucic [3] Gunnar Rätsch [1] Sylvain Gelly [3] Bernhard Schölkopf [2] Olivier Bachem [3]

In summary

- Medicine is full of multimodal information
  - Could provide useful training signal
  - Complementary information
- But we still treat decision making as a narrow task
- Representation learning is key to:
  - Combining and disentangling information
  - Reducing supervision
  - Embedding knowledge
  - Interpretable and explainable decisions
- Still several challenges lie ahead

Acknowledgements

UK

Lab:

- A. Dobrescu
- A. Chartsias
- T. Xia
- V. Giuffrida
- N. Dionelis
- G. Jacenkow
- G. Valvano
- X. Liu

Collaborators:

- D. Newby
- S. Semple
- A. Smout
- A. O'Neil
- G. Papanastasiou
- M. Williams

USA-EU

- R Dharmakumar
- A. Frangi
- J. Prince
- X. Papademetris
- N. Merchant
- H. Scharr
- P. Perata

Acknowledgements

Funding:

NIH 4R01HL091989
NIH 1R01HL136578

EP/N510129/1

EP/P022928/1

BB/N02334X/1
BB/P023487/1

Sotirios A. Tsaftaris, PhD

Email: S.Tsaftaris@ed.ac.uk

URL: http://tsaftaris.com