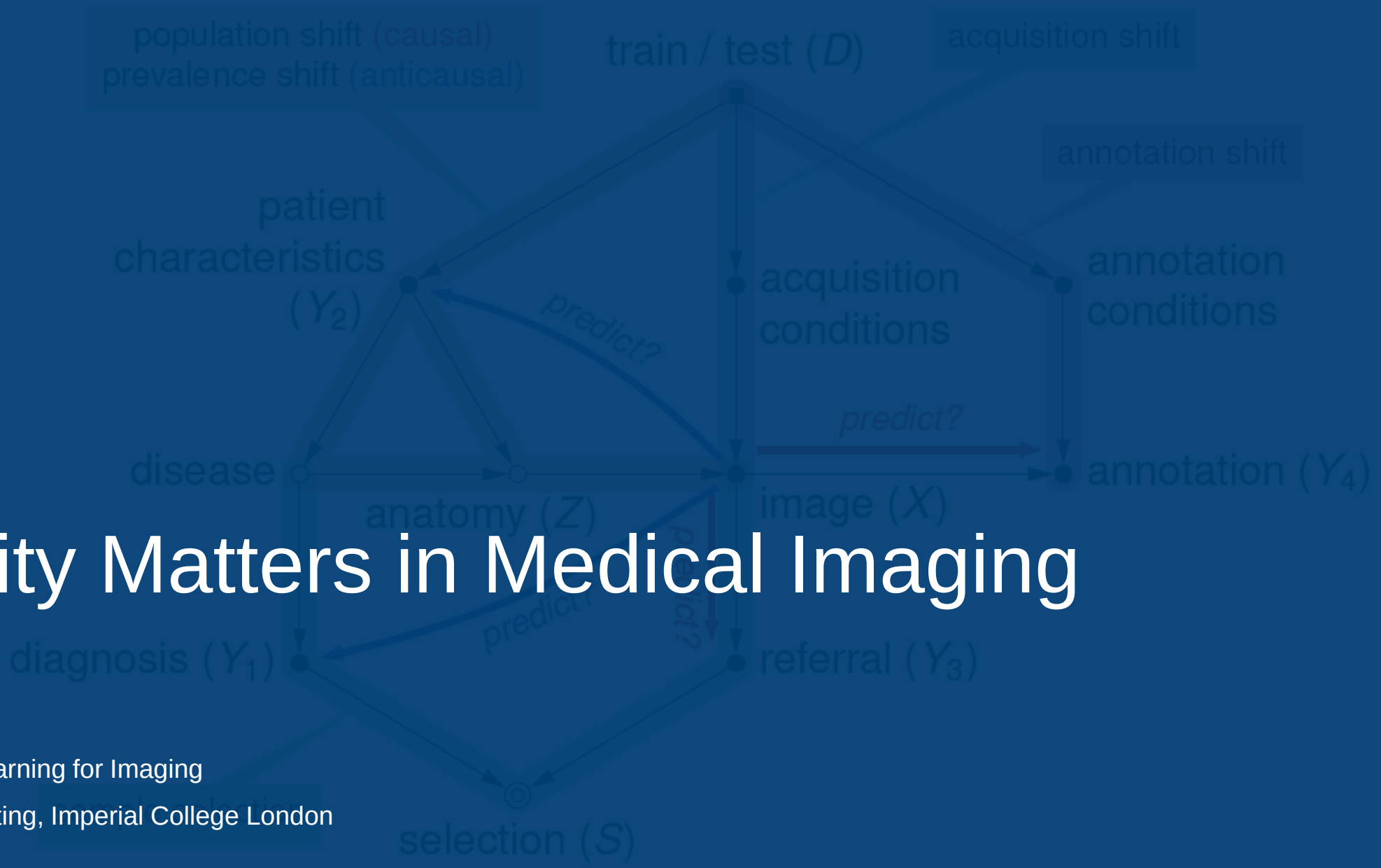
Imperial College London



Causality Matters in Medical Imaging

Ben Glocker

Reader in Machine Learning for Imaging

Department of Computing, Imperial College London

Predictive Modelling

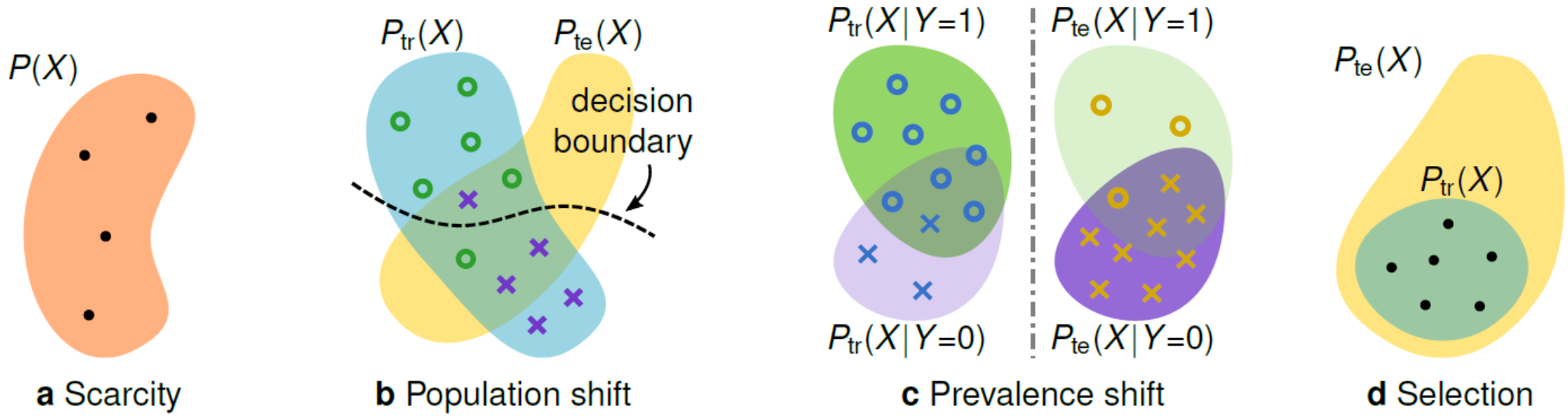
Given an **image** X, train a **model** to **predict** some annotation Y

$$P(Y|X)$$

Assumptions

- Sufficient training data (X, Y) is available
- Training and test data come from the same distribution

Challenges: Data Scarcity & Mismatch



A Causal Perspective

Background on causal reasoning

$$\textbf{cause } \frac{A \rightarrow B}{\textbf{mechanism}} \textbf{ effect}$$

Independence of cause and mechanism

A Causal Perspective

Background on causal reasoning

Independence of cause and mechanism

A Causal Perspective

What is the relationship between image X and annotation Y?

$$X \rightarrow Y$$

causal
(predict effect from cause)

$$P(Y|X)$$

A Causal Perspective

What is the relationship between image X and annotation Y?

$$Y \rightarrow X$$

anti-causal
(predict cause from effect)

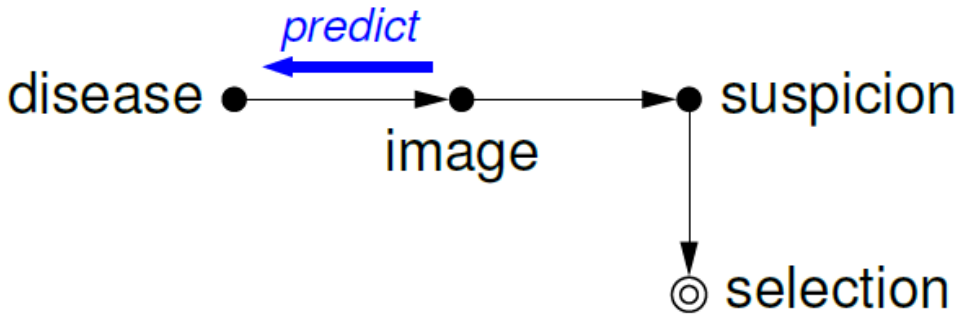
$$P(Y|X)$$

Example: Skin Lesion Classification

X – dermascopic image

Y – biopsy-derived diagnosis

$Y \rightarrow X$ **anti-causal**
(predict cause from effect)



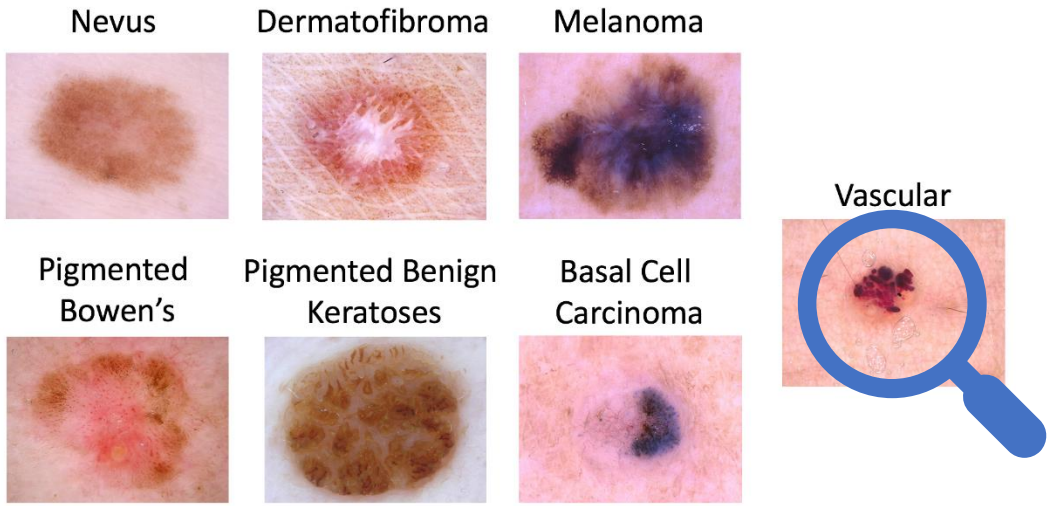


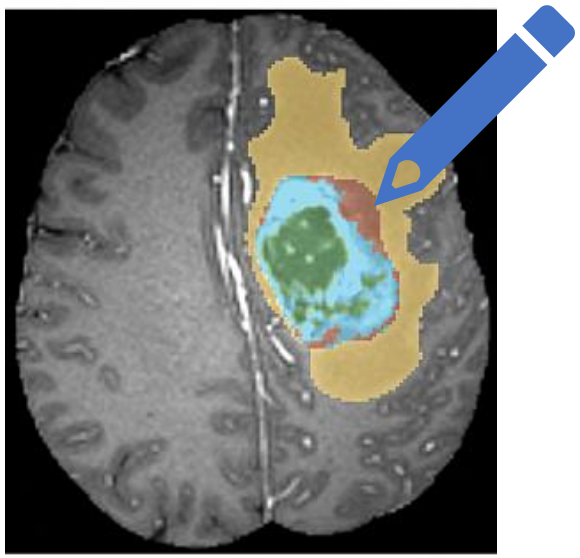
Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Example: Brain Tumour Segmentation

X – structural brain MRI

Y – manually drawn contour

$X \rightarrow Y$ **causal**
(predict effect from cause)



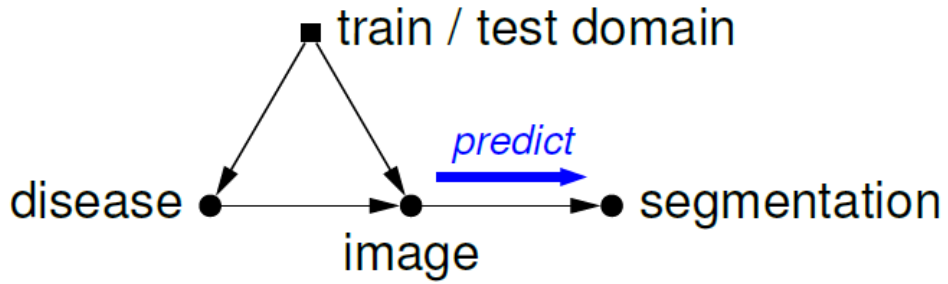


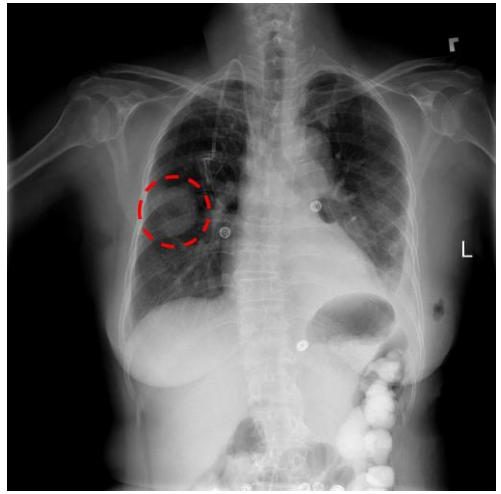
Image: BraTS 2018 challenge (http://braintumorsegmentation.org/)

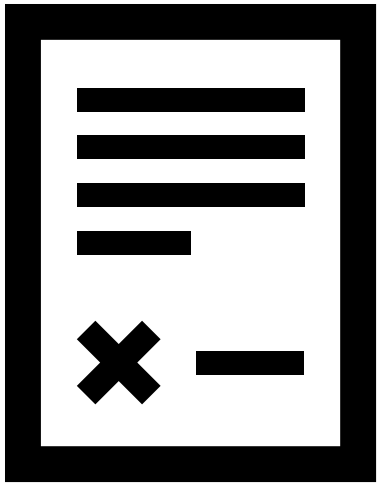
Example: Radiology Reports

X – chest X-ray

Y – diagnosis extracted from report

$X \overset{?}{\leftrightarrow} Y$ **causal or anti-causal?**





Meta-information is essential to establish causal relationships

Image: National Institutes of Health (https://bit.ly/2GsqXIw)

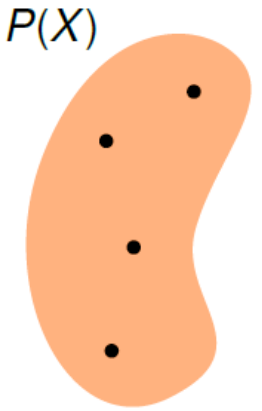
Relevant Meta-Information

Attribute	Examples
Causal direction (predict effect from cause or cause from effect)	
field of application	diagnosis / screening / prognosis / exploratory research
task category	segmentation / classification / regression / detection
annotation method	manual / (semi-)automatic / clinical tests; annotation policy
nature of annotations	image-wide label / pixel-wise segmentation / spatial coordinates
annotation reliability	image noise, acquisition artefacts, low contrast; user or software errors; signal-to-noise ratio, inter- and intra-observer variability

Causality Matters?

Data Scarcity

Labelled data is scarce, **unlabelled** data is abundant

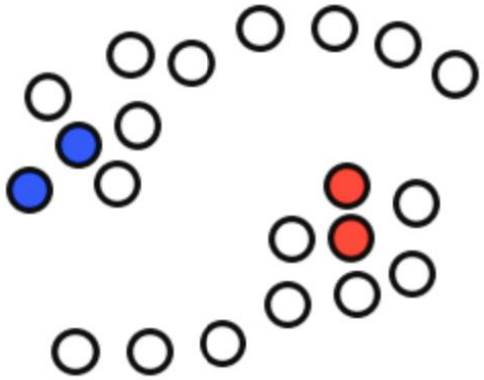


a Scarcity

Can we leverage unlabelled data?

(semi-supervised learning)

Semi-supervised learning



Labelled data: $P(X, Y)$

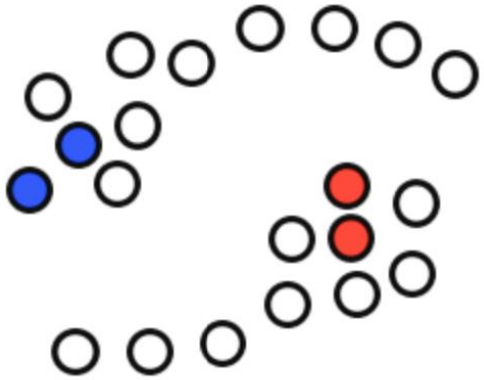
Unlabelled data: $P(X)$

$X \rightarrow Y$ **causal**
(predict effect from cause)

$Y \rightarrow X$ **anti-causal**
(predict cause from effect)

$$P(Y|X)$$

Semi-supervised learning



Labelled data: $P(X, Y)$

Unlabelled data: $P(X)$ **cause**

$X \rightarrow Y$ **causal** (predict effect from cause)

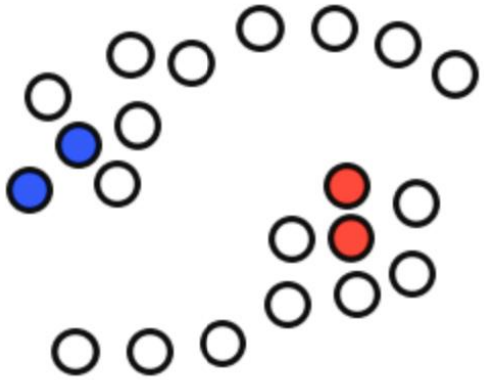
$Y \rightarrow X$ **anti-causal** (predict cause from effect)

$$\frac{P(Y|X)}{\textbf{mechanism}}$$

Independence of cause and mechanism

Schölkopf et al. 2019. Semi-supervised learning in causal and anticausal settings

Semi-supervised learning



Labelled data: $P(X, Y)$

Unlabelled data: $P(X)$ **cause**

SSL may be futile in image segmentation

$Y \rightarrow X$ anti-causal (predict cause from effect)

$P(Y|X)$ mechanism

Independence of cause and mechanism

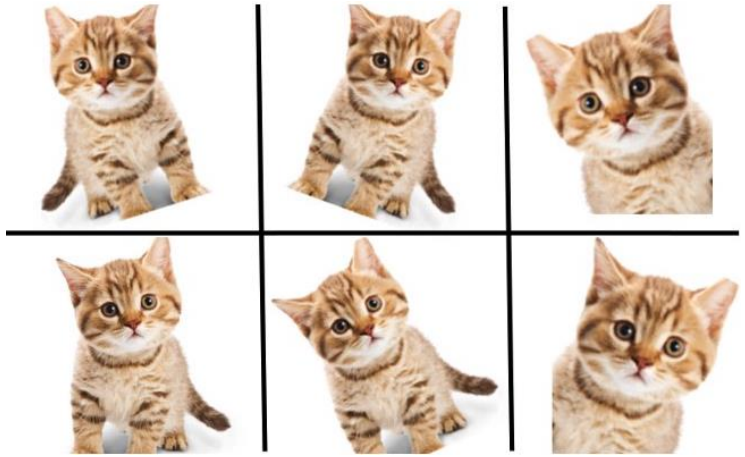
Schölkopf et al. 2019. Semi-supervised learning in causal and anticausal settings

Data Augmentation

Labelled data: $P(X, Y)$

Generate additional (X, Y) pairs via realistic perturbations

Learn augmentations...
...from unlabelled data!

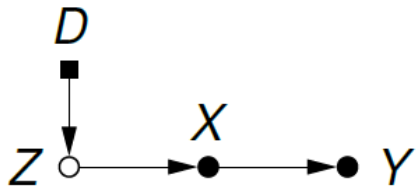


Chaitanya et al. 2019: Semi-supervised and task-driven data augmentation (arXiv:1902.05396)

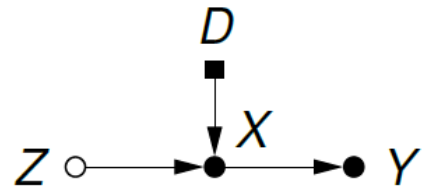
Image: https://medium.com/nanonets/how-to-use-deep-learning-when-you-have-limited-data-part-2-data-augmentation-c26971dc8ced

Dataset Shift

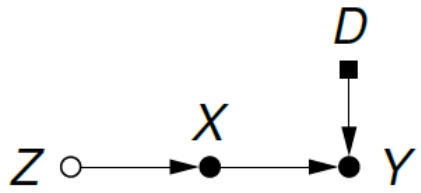
D is a domain indicator, Z is the unobserved true anatomy



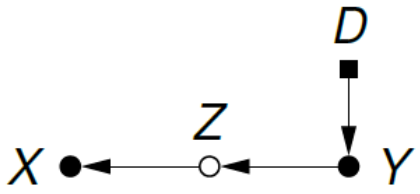
a Population shift:
$\boldsymbol{P_D}(\boldsymbol{Z})P(X|Z)P(Y|X)$



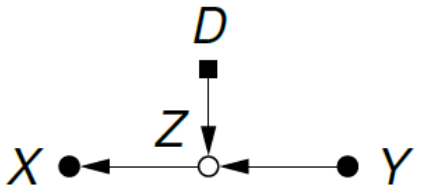
b (Causal) Acquisition shift:
$P(Z)\boldsymbol{P_D}(\boldsymbol{X}|\boldsymbol{Z})P(Y|X)$



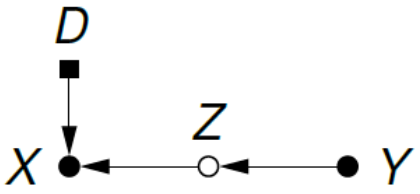
c Annotation shift:
$P(Z)P(X|Z)\boldsymbol{P_D}(\boldsymbol{Y}|\boldsymbol{X})$



d Prevalence shift:
$P(X|Z)P(Z|Y)\boldsymbol{P_D}(\boldsymbol{Y})$

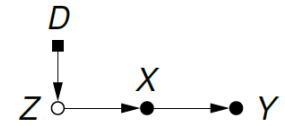


e Manifestation shift:
$P(X|Z)\boldsymbol{P_D}(\boldsymbol{Z}|\boldsymbol{Y})P(Y)$

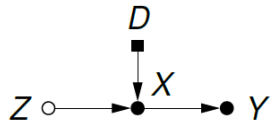


f (Anticausal) Acquisition shift:
$\boldsymbol{P_D}(\boldsymbol{X}|\boldsymbol{Z})P(Z|Y)P(Y)$

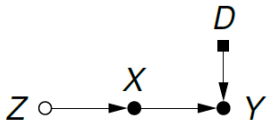
Dataset Shift



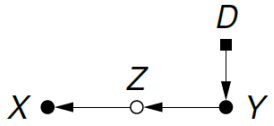
a Population shift:
$\boldsymbol{P_D}(\boldsymbol{Z})P(X|Z)P(Y|X)$



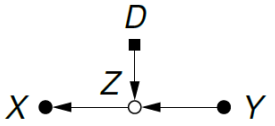
b (Causal) Acquisition shift:
$P(Z)\boldsymbol{P_D}(\boldsymbol{X}|\boldsymbol{Z})P(Y|X)$



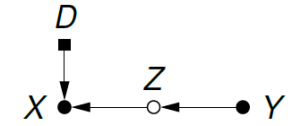
c Annotation shift:
$P(Z)P(X|Z)\boldsymbol{P_D}(\boldsymbol{Y}|\boldsymbol{X})$



d Prevalence shift:
$P(X|Z)P(Z|Y)\boldsymbol{P_D}(\boldsymbol{Y})$



e Manifestation shift:
$P(X|Z)\boldsymbol{P_D}(\boldsymbol{Z}|\boldsymbol{Y})P(Y)$



f (Anticausal) Acquisition shift:
$\boldsymbol{P_D}(\boldsymbol{X}|\boldsymbol{Z})P(Z|Y)P(Y)$

Type	Direction	Change	Examples of differences
Population shift	causal	$P_D(Z)$	ages, sexes, diets, habits, ethnicities, genetics
Annotation shift	causal	$P_D(Y\|X)$	annotation policy, annotator experience
Prevalence shift	anticausal	$P_D(Y)$	case-control balance, target selection
Manifestation shift	anticausal	$P_D(Z\|Y)$	anatomical manifestation of the target disease or trait
Acquisition shift	either	$P_D(X\|Z)$	scanner, resolution, contrast, modality, protocol

Acquisition Shift: A Little Experiment

- 3T T1w brain MRI from two studies (Cam-CAN[1], UK Biobank[2])
- n=592 age- and sex-matched subjects (296 per site)

Goal: Simulate an *ideal* multi-site neuroimaging study

[1] **Cam-CAN:** 3T Siemens TIM Trio scanner with a 32-channel receiver head coil. 3D MPRAGE, TR=2250ms, TE=2.99ms, TI=900ms; FA=9 deg; FOV=256x240x192mm; 1mm isotropic; GRAPPA=2; TA=4mins 32s

[2] **UKBB:** 3T Siemens Skyra scanner with a 32-channel receiver head coil. 3D MPRAGE, R=2, TR=2000ms, TE=385ms, TI=880ms; FOV=208x256x256mm; 1mm isotropic; Duration 4mins 54s

Glocker et al. 2019. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects

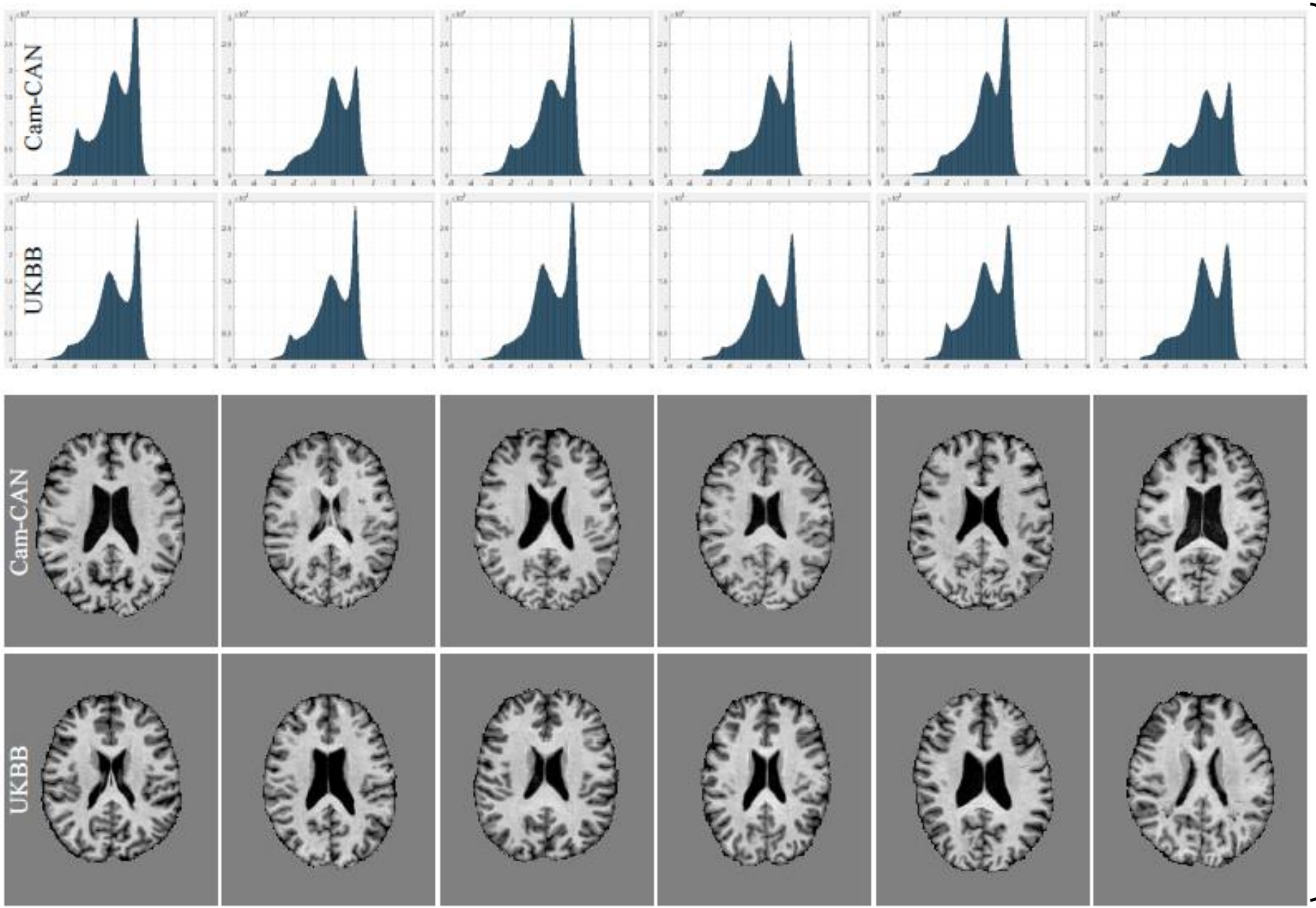
Acquisition Shift: A Little Experiment

Pre-processing:

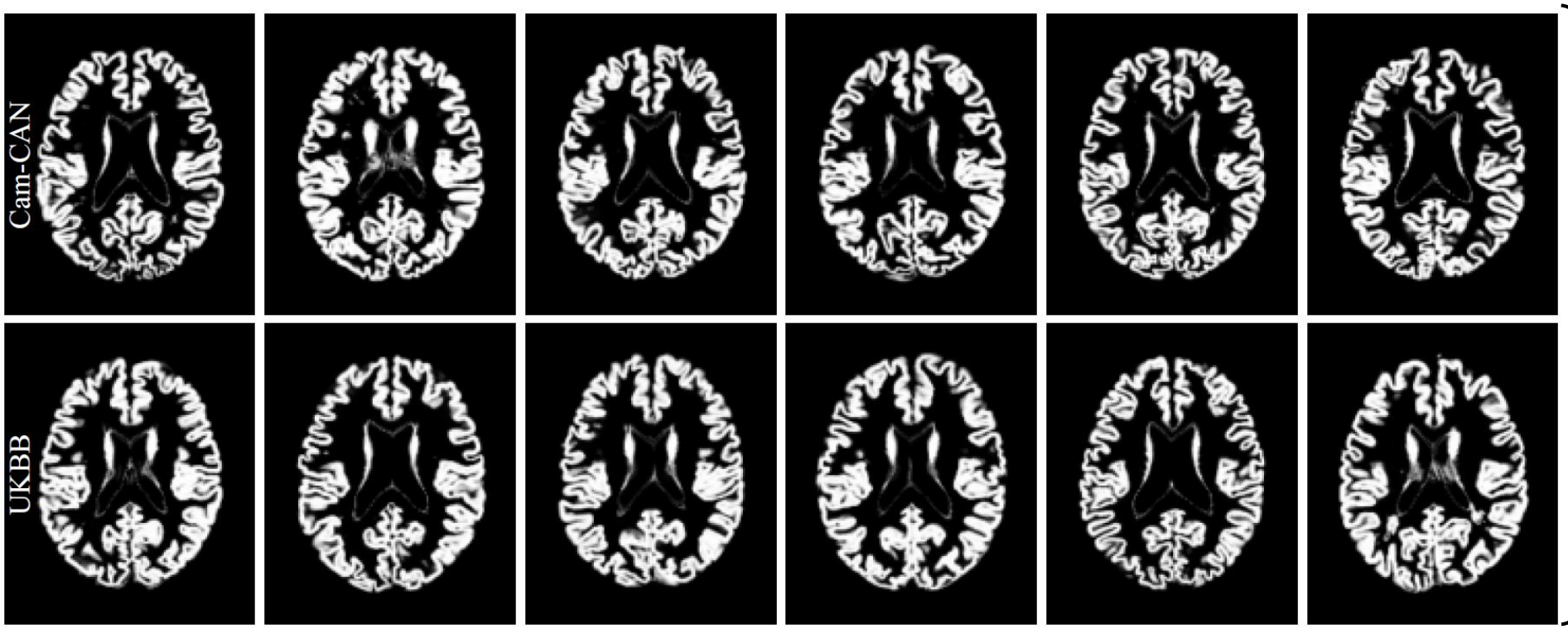
1. Lossless reorientation to standard view (left, posterior, superior)
2. Skull stripping with ROBEX v1.2
3. Bias field correction with N4ITK
4. Linear registration to MNI ICBM 152 2009a Nonlinear Symmetric
5. Intensity normalisation

We also run SPM12 and FSL's FAST v4.0 to extract tissue maps

Glocker et al. 2019. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects



No obvious differences



No obvious differences

Acquisition Shift: A Little Experiment

A site classifier can tell where the data is from with very high accuracy

Stripped	Bias Field	Aligned	Intensities	Accuracy	Avg. Entropy	Avg. Prob.
✓	✓	rigid	whitening	96.96%	0.4039	0.8296
✓	✓	affine	whitening	98.82%	0.3876	0.8397
SPM12 – Segment				**Accuracy**	**Avg. Entropy**	**Avg. Prob.**
✗	✓	rigid	graymatter	80.24%	0.6363	0.6399
✗	✓	non-linear	graymatter	96.62%	0.5675	0.7234
FSL – FAST				**Accuracy**	**Avg. Entropy**	**Avg. Prob.**
✓	✓	rigid	graymatter	93.24%	0.4542	0.7968

Glocker et al. 2019. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects

Acquisition Shift: A Little Experiment

Predictive accuracy on a proxy task: sex classification

Data Arrangement	Aligned	Accuracy	Avg. Entropy	Avg. Prob.
Multi-site age/sex-matched	rigid	82.60%	0.5304	0.7388
Single-site (Cam-CAN)	rigid	81.42%	0.5592	0.7179
Single-site (UKBB)	rigid	84.46%	0.5049	0.7572
Cam-CAN females / UKBB males	rigid	94.59%	0.4036	0.8311
Cam-CAN 80/20% / UKBB 20/80%	rigid	85.87%	0.5038	0.7616
Cam-CAN train / UKBB test	rigid	81.42%	0.5617	0.7124
UKBB train / Cam-CAN test	rigid	78.04%	0.5284	0.7419

Glocker et al. 2019. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects

Acquisition Shift: A Little Experiment

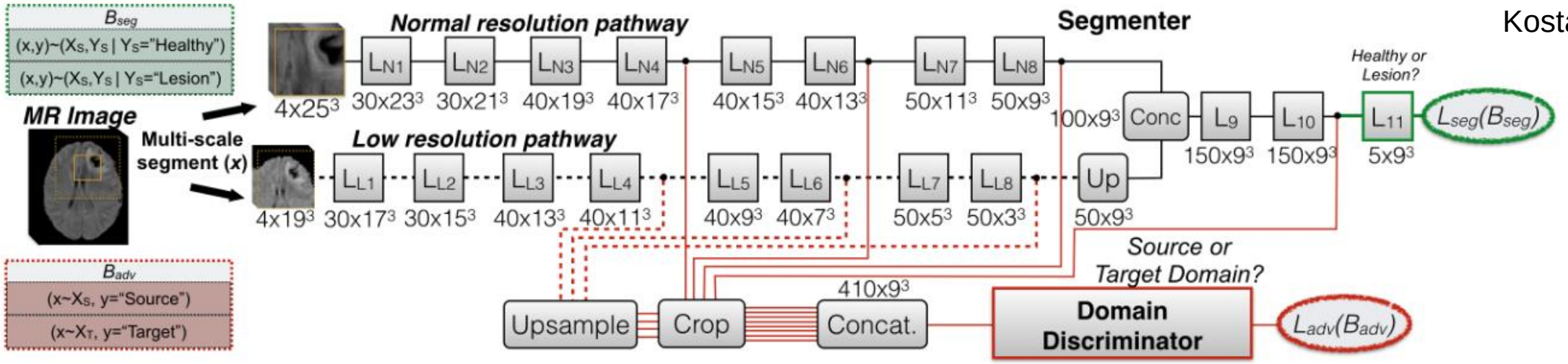
Predictive accuracy on a proxy task: sex classification

Data Arrangement	Aligned	Accuracy	Avg. Entropy	Avg. Prob.
Multi-site age/sex-matched	affine	79.73%	0.6345	0.6389
Single-site (Cam-CAN)	affine	77.70%	0.6439	0.6269
Single-site (UKBB)	affine	81.08%	0.6393	0.6316
Cam-CAN females / UKBB males	affine	98.99%	0.4641	0.8013
Cam-CAN 80/20% / UKBB 20/80%	affine	84.78%	0.5713	0.7125
Cam-CAN train / UKBB test	affine	73.65%	0.6462	0.6245
UKBB train / Cam-CAN test	affine	62.16%	0.6075	0.6769

Glocker et al. 2019. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects

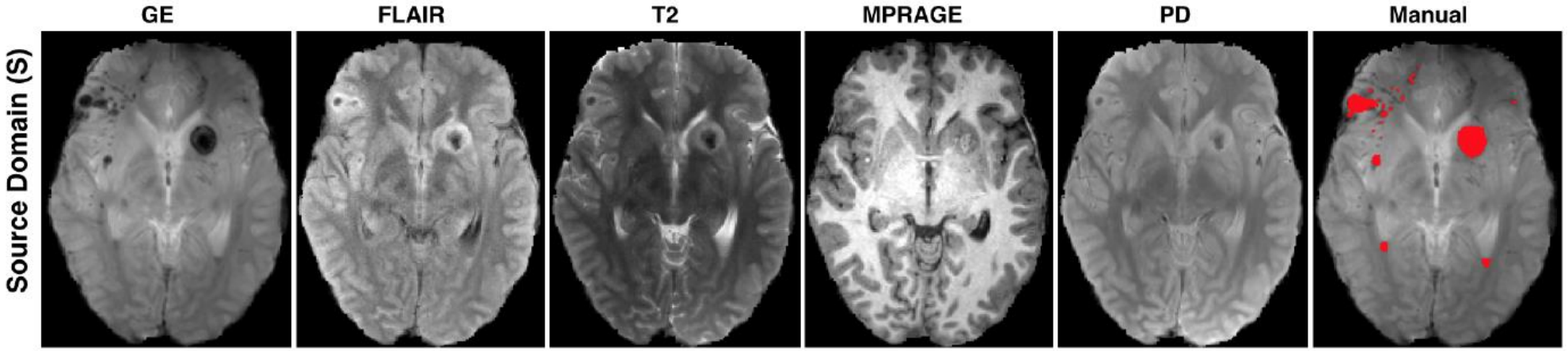
Domain Adaptation

Kostas

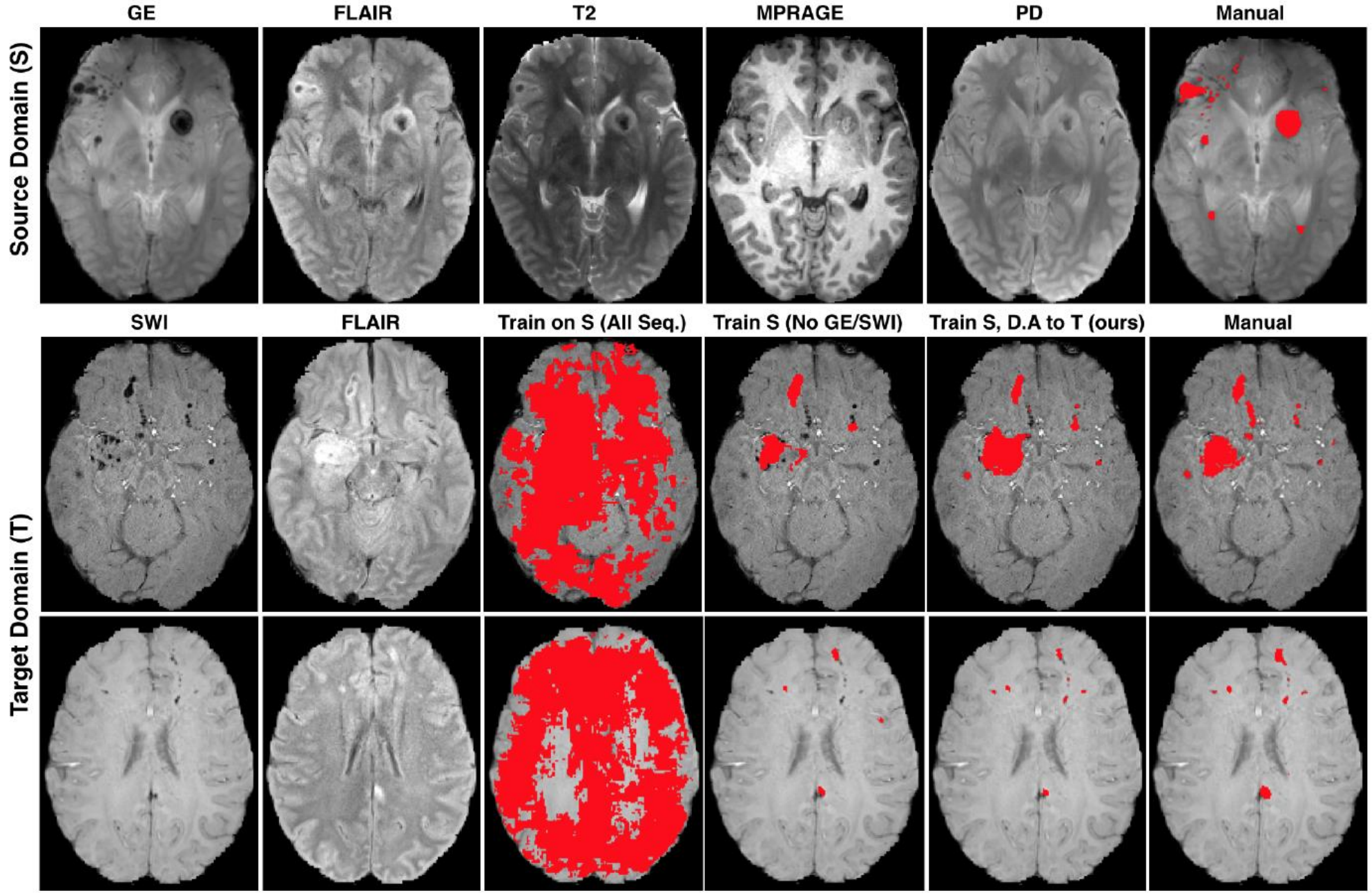


- Adversarial domain discriminator encourages domain invariant features
- No task labels required for test domain
- Need to retrain for each new test domain

Kamnitsas et al. 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks

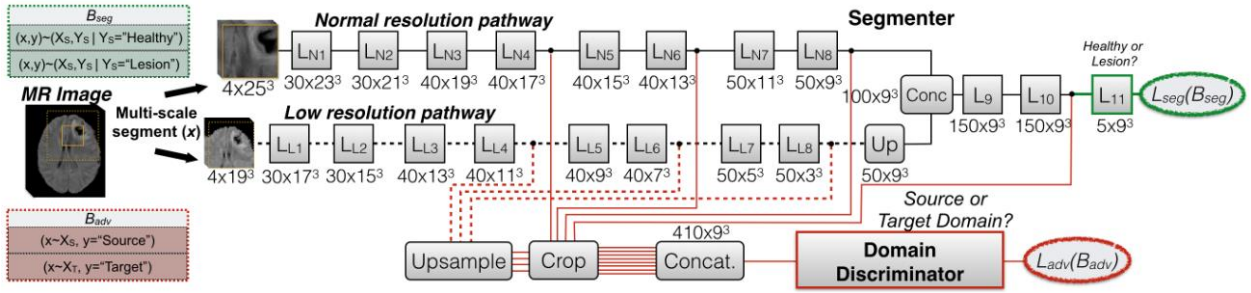


Kamnitsas et al. 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks

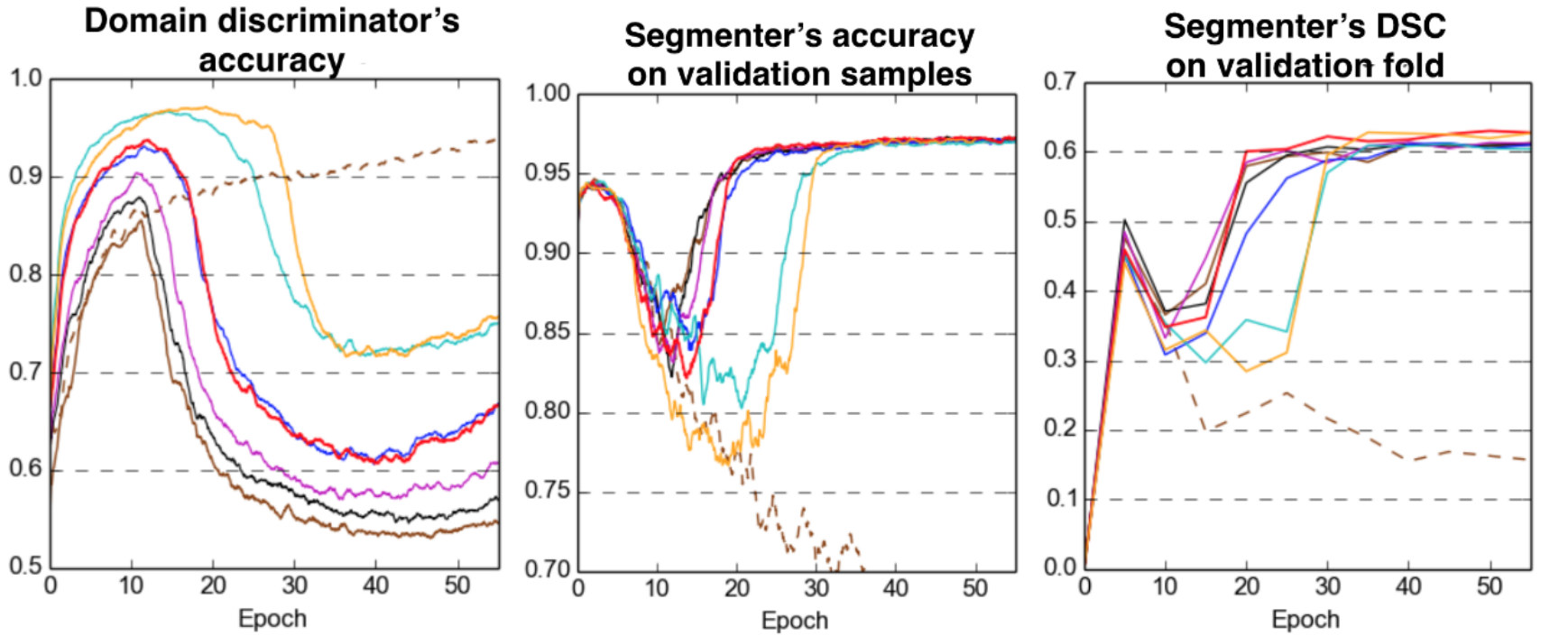


Kamnitsas et al. 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks

Domain Adaptation

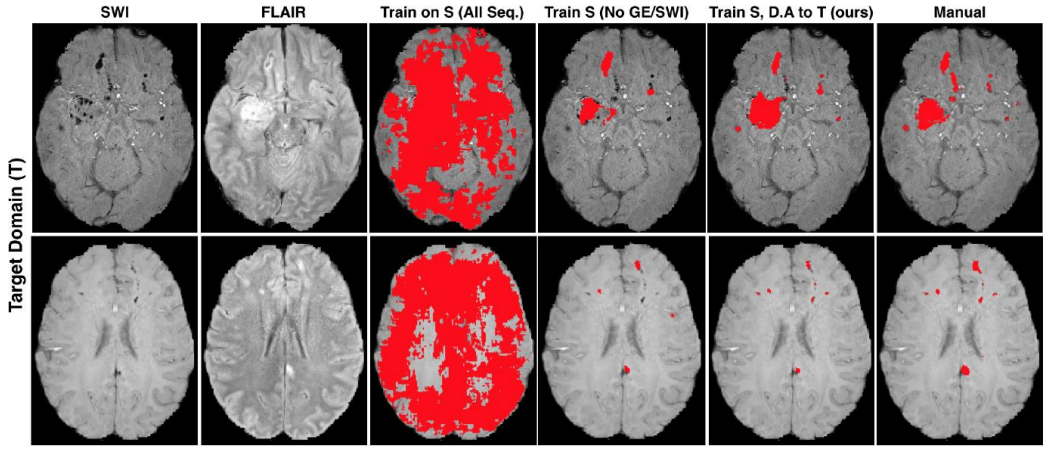


Analysis of training behaviour



Kamnitsas et al. 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks

Domain Adaptation



Segmentation performance and baselines

	DSC	Recall	Precision
Train on S	15.7(13.5)	80.4(12.3)	09.5(09.0)
Train on S (No GE/SWI)	59.7(22.1)	55.7(22.6)	69.7(21.5)
Train on S → UDA to T (ours)	62.7(19.8)	58.9(21.2)	71.6(18.4)
Train on T	63.5(20.2)	60.6(21.1)	71.5(19.8)
Train on S+T	66.5(17.7)	66.6(19.1)	69.4(19.0)
Train on S+T (GE/SWI diff chan.)	64.7(19.2)	65.7(20.2)	67.0(20.8)

Kamnitsas et al. 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks

Domain Generalisation

Qi

How can we generalise to new domains?

Idea: Simulate domain shift during training

Assumption: Two or more domains available for training

PACS Benchmark

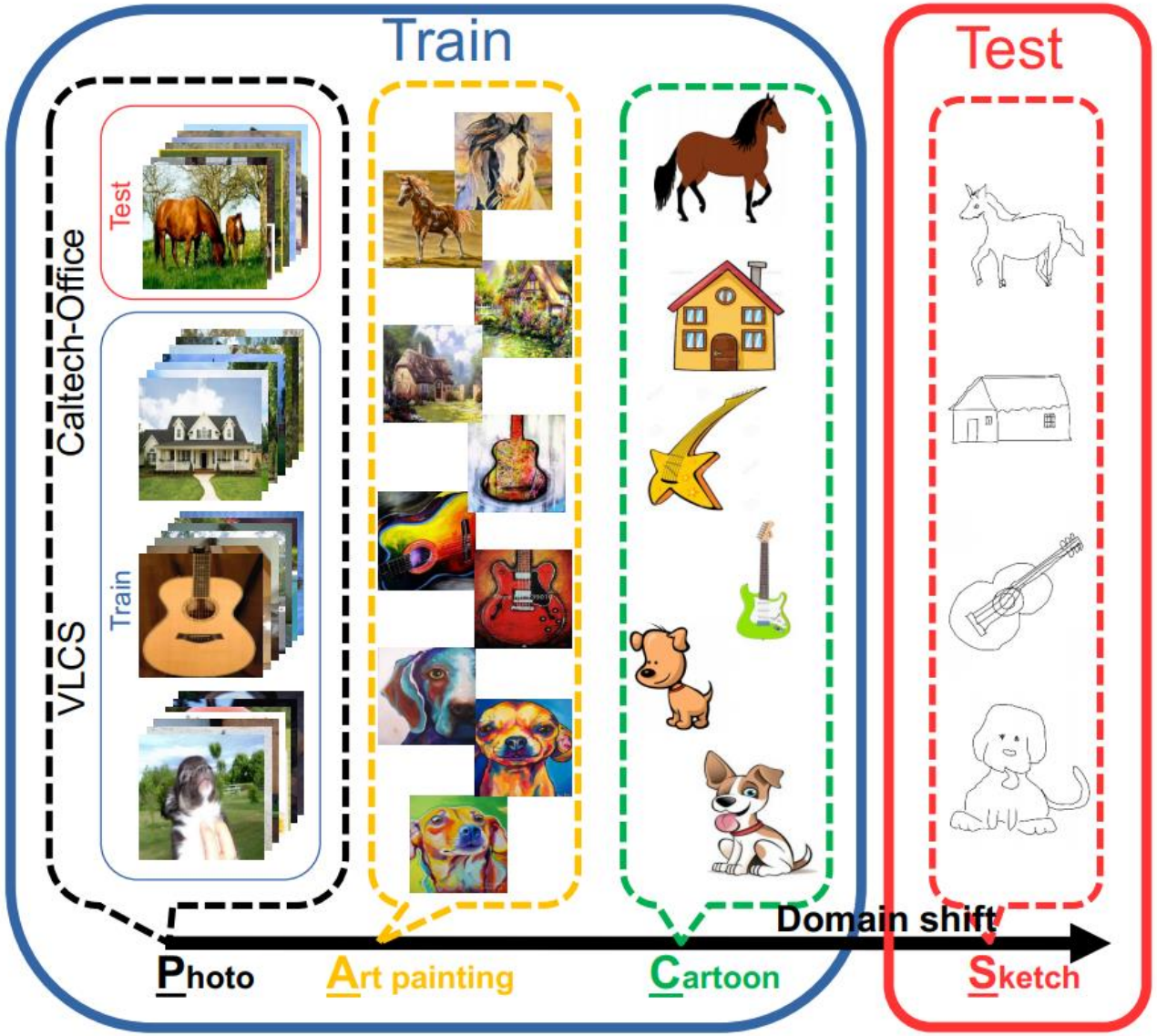
Four domains

- photo, art, cartoon, sketch

Seven categories

- dog, elephant, giraffe, guitar, horse, house, person

In total 9991 images



Li et al. 2017. Deeper, broader and artier domain generalization

Learning of Semantic Features for DG

1. Episodic training procedure to simulate domain shift
2. New loss function with global and local constraints

Let's decompose our model

$$F_\psi : \mathcal{X} \rightarrow \mathcal{Z}$$

Feature extractor

$$T_\theta : \mathcal{Z} \rightarrow \mathbb{R}^C$$

Task network

Dou et al. 2019. Domain Generalization via Model-Agnostic Learning of Semantic Features

Episodic Training

$$F_\psi : \mathcal{X} \to \mathcal{Z}$$

Feature extractor

$$T_\theta : \mathcal{Z} \to \mathbb{R}^C$$

Task network

Split training domains $\mathcal{D}$ into meta-train $\mathcal{D}_{\text{tr}}$ and meta-test $\mathcal{D}_{\text{te}}$

$$(\psi', \theta') = (\psi, \theta) - \alpha \nabla_{\psi,\theta} \mathcal{L}_{\text{task}}(\mathcal{D}_{\text{tr}}; \psi, \theta)$$

$$F_{\psi'} : \mathcal{X} \to \mathcal{Z}$$

$$T_{\theta'} : \mathcal{Z} \to \mathbb{R}^C$$

Finn et al. 2017. Model-agnostic meta-learning for fast adaptation of deep networks

Global Class Alignment

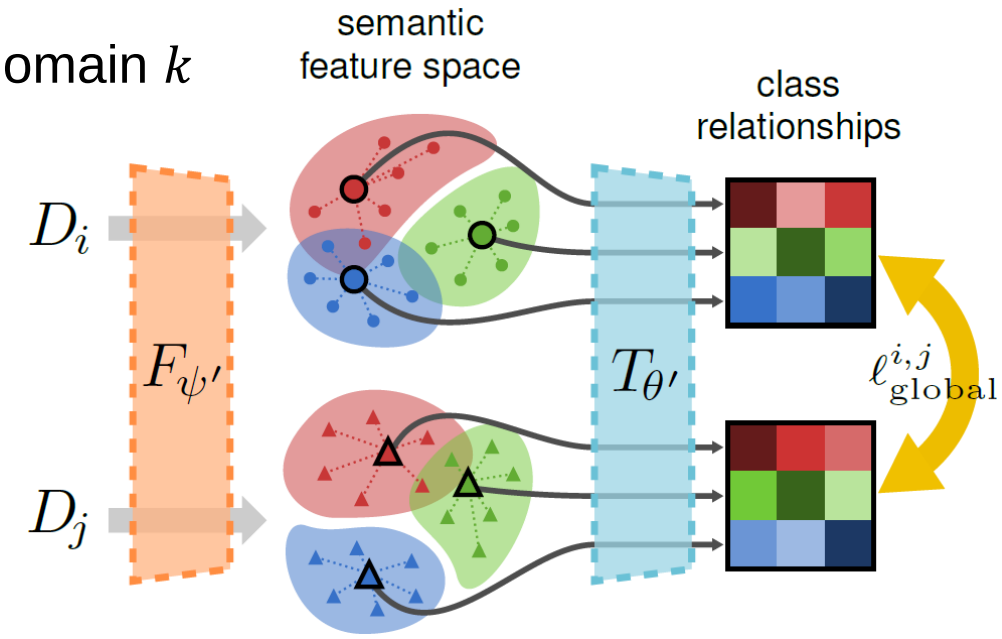
Preserve inter-class relationships

- Compute class-specific mean feature vector for each domain k

$$\bar{\mathbf{z}}_c^{(k)} = \frac{1}{N_k^{(c)}} \sum_{n:y_n^{(k)}=c} F_{\psi'}(\mathbf{x}_n^{(k)}) \approx \mathbb{E}_{D_k}[F_{\psi'}(\mathbf{x}) \mid y = c]$$

- Compute soft label distributions

$$\mathbf{s}_c^{(k)} = \mathrm{softmax}(T_{\theta'}(\bar{\mathbf{z}}_c^{(k)})/\tau)$$



- Class alignment loss

$$\ell_{\text{global}}(D_i, D_j; \psi', \theta') = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2}[D_{\text{KL}}(\mathbf{s}_c^{(i)} \,\|\, \mathbf{s}_c^{(j)}) + D_{\text{KL}}(\mathbf{s}_c^{(j)} \,\|\, \mathbf{s}_c^{(i)})]$$

Local Sample Clustering

Domain-invariant discriminative features

- Use an embedding network M_ϕ that takes $\mathbf{z} = F_{\psi'}(\mathbf{x})$ as input

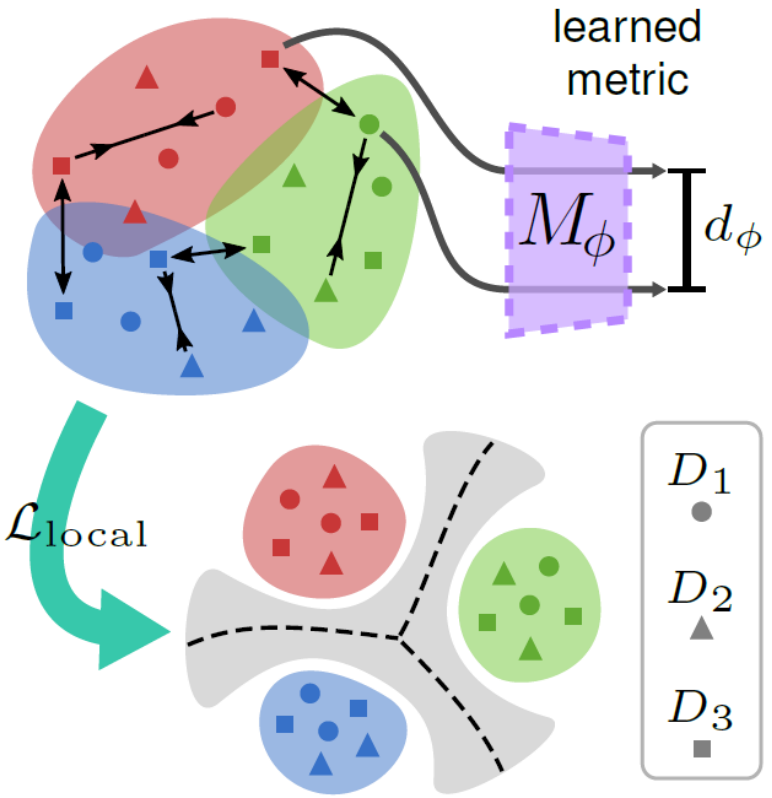
$$d_\phi(\mathbf{z}_n, \mathbf{z}_m) = \|\mathbf{e}_n - \mathbf{e}_m\|_2 = \|M_\phi(\mathbf{z}_n) - M_\phi(\mathbf{z}_m)\|_2$$

- Option 1: Contrastive loss (mild domain shift)

$$\ell_{\text{con}}^{n,m} = \begin{cases} d_\phi(\mathbf{z}_n, \mathbf{z}_m)^2, & \text{if } y_n = y_m \\ (\max\{0,\ \xi - d_\phi(\mathbf{z}_n, \mathbf{z}_m)\})^2, & \text{if } y_n \neq y_m \end{cases}$$

- Option 2: Triplet loss (severe domain shift)

$$\ell_{\text{tri}}^{a,p,n} = \max\{0,\ d_\phi(\mathbf{z}_a, \mathbf{z}_p)^2 - d_\phi(\mathbf{z}_a, \mathbf{z}_n)^2 + \xi\}$$



Results

VLCS benchmark

- Four domains, all photos, five classes (bird, car, chair, dog, person)

Table 1: Domain generalization results on VLCS dataset with object recognition accuracy (%).

Source	Target	D-MTAE [12]	CIDDG [30]	CCSA [34]	DBADG [25]	MMD-AAE [28]	MLDG [26]	Epi-FCR [27]	JiGen [3]	DeepAll (Baseline)	MASF (Ours)
L,C,S	V	63.90	64.38	67.10	69.99	67.70	67.7	67.1	70.62	68.67±0.09	69.14±0.19
V,C,S	L	60.13	63.06	62.10	63.49	62.60	61.3	64.3	60.90	63.10±0.11	64.90±0.08
V,L,S	C	89.05	88.83	92.30	93.63	94.40	94.4	94.1	96.93	92.86±0.13	94.78±0.16
V,L,C	S	61.33	62.10	59.10	61.32	64.40	65.9	65.9	64.30	64.11±0.17	67.64±0.12
Average		68.60	69.59	70.15	72.11	72.28	72.3	72.9	73.19	72.19	74.11

Results

PACS benchmark

- Four domains, seven classes

Table 2: Domain generalization results on PACS dataset with recognition accuracy (%) using AlexNet.

Source	Target	D-MTAE [12]	CIDDG [30]	DBADG [25]	MLDG [26]	Epi-FCR [27]	MetaReg [1]	JiGen [3]	DeepAll (Baseline)	MASF (Ours)
C,P,S	Art painting	60.27	62.70	62.86	66.23	64.7	69.82	67.63	67.60±0.21	70.35±0.33
A,P,S	Cartoon	58.65	69.73	66.97	66.88	72.3	70.35	71.71	68.87±0.22	72.46±0.19
A,C,S	Photo	91.12	78.65	89.50	88.00	86.1	91.07	89.00	89.20±0.24	90.68±0.12
A,C,P	Sketch	47.68	64.45	57.51	58.96	65.0	59.26	65.18	61.13±0.30	67.33±0.12
Average		64.48	68.88	69.21	70.01	72.0	72.62	73.38	71.70	75.21

Results

Deep residual networks

Table 4: PACS results with deep residual network architectures (accuracy, %).

Source	Target	ResNet-18		ResNet-50	
		DeepAll	MASF (ours)	DeepAll	MASF (ours)
C,P,S	Art-painting	77.38 ± 0.15	80.29 ± 0.18	81.41 ± 0.16	82.89 ± 0.16
A,P,S	Cartoon	75.65 ± 0.11	77.17 ± 0.08	78.61 ± 0.17	80.49 ± 0.21
A,C,S	Photo	94.25 ± 0.09	94.99 ± 0.09	94.83 ± 0.06	95.01 ± 0.10
A,C,P	Sketch	69.64 ± 0.25	71.69 ± 0.22	69.69 ± 0.11	72.29 ± 0.15

Results

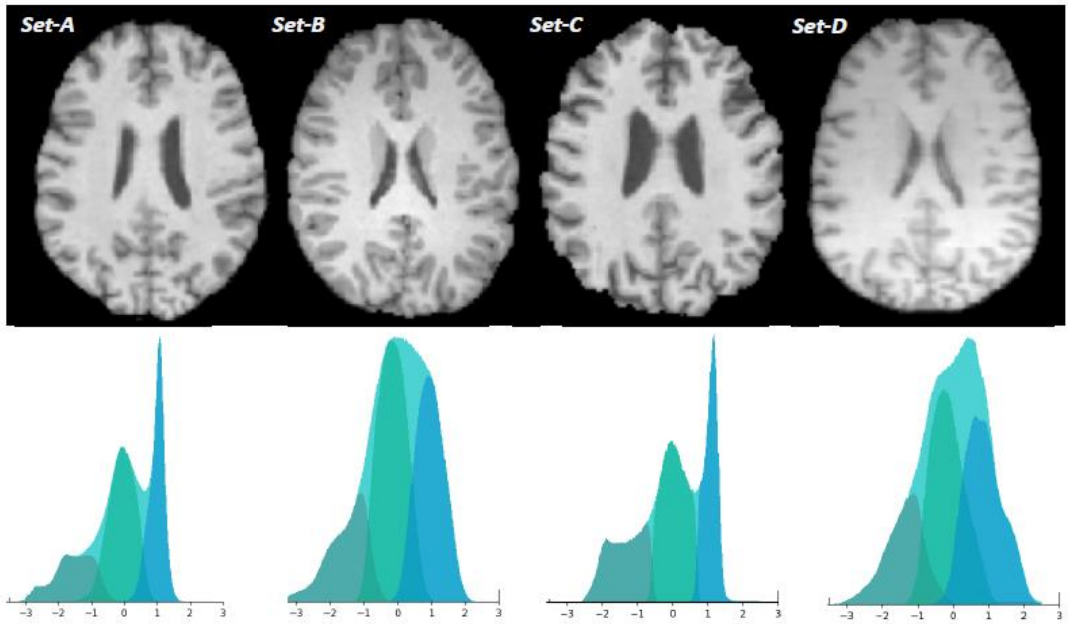
Ablation study

Table 3: Ablation study on key components of our method with the PACS dataset (accuracy, %).

Episodic	$\mathcal{L}_{\text{global}}$	$\mathcal{L}_{\text{local}}$	Art	Cartoon	Photo	Sketch	Average
-	-	-	67.60±0.21	68.87±0.22	89.20±0.24	61.13±0.30	71.70
✓	-	-	69.19±0.10	70.66±0.37	90.36±0.18	59.89±0.26	72.52
-	✓	-	69.43±0.29	70.22±0.21	90.64±0.15	60.11±0.17	72.60
-	-	✓	69.50±0.15	70.25±0.13	90.12±0.12	63.02±0.12	73.22
-	✓	✓	69.48±0.20	71.15±0.16	90.16±0.15	64.73±0.34	73.88
✓	✓	-	69.94±0.15	72.16±0.28	90.10±0.12	63.54±0.13	73.93
✓	-	✓	69.50±0.20	71.44±0.34	90.16±0.15	64.97±0.28	74.02
✓	✓	✓	70.35±0.33	72.46±0.19	90.68±0.12	67.33±0.12	75.21

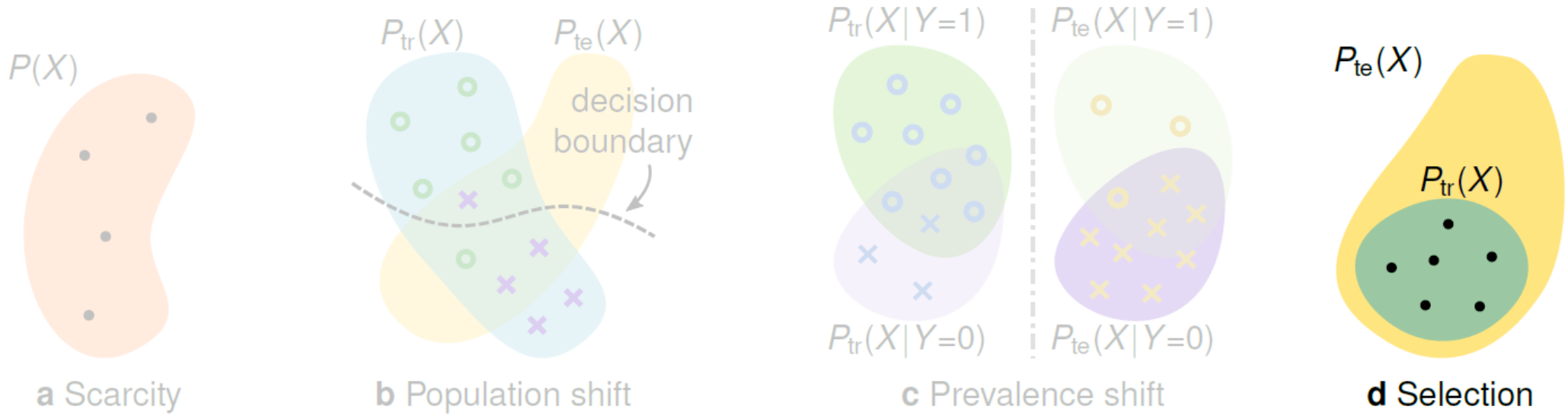
Results

Tissue segmentation on brain MRI from four different sites



Test \ Train	Set-A	Set-B	Set-C	Set-D	DeepAll	MASF
Set-A	90.62	88.91	88.81	85.03	89.09	89.82
Set-B	85.03	94.22	81.38	88.31	90.41	91.71
Set-C	93.14	92.80	95.40	88.68	94.30	94.50
Set-D	76.32	88.39	73.50	94.29	88.62	89.51

Back to Causality



a Scarcity **b** Population shift **c** Prevalence shift **d** Selection

Sample Selection

S is a selection variable (accepted/rejected)

random image-dependent target-dependent joint

Sample Selection

Type	Causation	Examples of selection processes	Resulting bias
Random	none	uniform subsampling, randomized trial	none
Image	$X \rightarrow S$	visual phenotype selection (e.g. anatomical traits, lesions)	population shift
		image quality control (QC; e.g. noise, low contrast, artefacts)	acquisition shift
Target	$Y \rightarrow S$	hospital admission, filtering by disease, annotation QC, learning strategies (e.g. class balancing, patch selection)	prevalence shift
Joint	$X \rightarrow S \leftarrow Y$	combination of the above (e.g. curated benchmark dataset)	spurious assoc.

Causality Matters!

Guidelines and Regulation

POLICY FORUM | TECHNOLOGY AND REGULATION

Regulation of predictive analytics in medicine

Ravi B. Parikh[1], Ziad Obermeyer[2], Amol S. Navathe[1,3]
+ See all authors and affiliations

Science 22 Feb 2019:
Vol. 363, Issue 6429, pp. 810-812
DOI: 10.1126/science.aaw0029

Evidence standards framework for digital health technologies

As digital health technologies develop at an increasing pace, we've worked with partners to develop standards that ensure new technologies are clinically effective and offer economic value.

The aim of the standards is to make it easier for innovators and commissioners to understand what good levels of evidence for digital healthcare technologies look like. Digital healthcare technologies must also meet the needs of the health and care system, patients, and users.

NHS England

NICE National Institute for Health and Care Excellence

Public Health England

THE LANCET
Volume 394, Issue 10192, 6–12 July 2019, Pages 9-11

Comment

WHO and ITU establish benchmarking process for artificial intelligence in health

Thomas Wiegand [a] ✉, Ramesh Krishnamurthy [b], Monique Kuglitsch [a], Naomi Lee [c], Sameer Pujari [b], Marcel Salathé [d], Markus Wenzel [a], Shan Xu [e]

Department of Health & Social Care

Guidance

Code of conduct for data-driven health and care technology

Updated 18 July 2019

THE LANCET
Volume 393, Issue 10181, 20–26 April 2019, Pages 1577-1579

Comment

Reporting of artificial intelligence prediction models

Gary S Collins [a] ✉, Karel G M Moons [b]

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

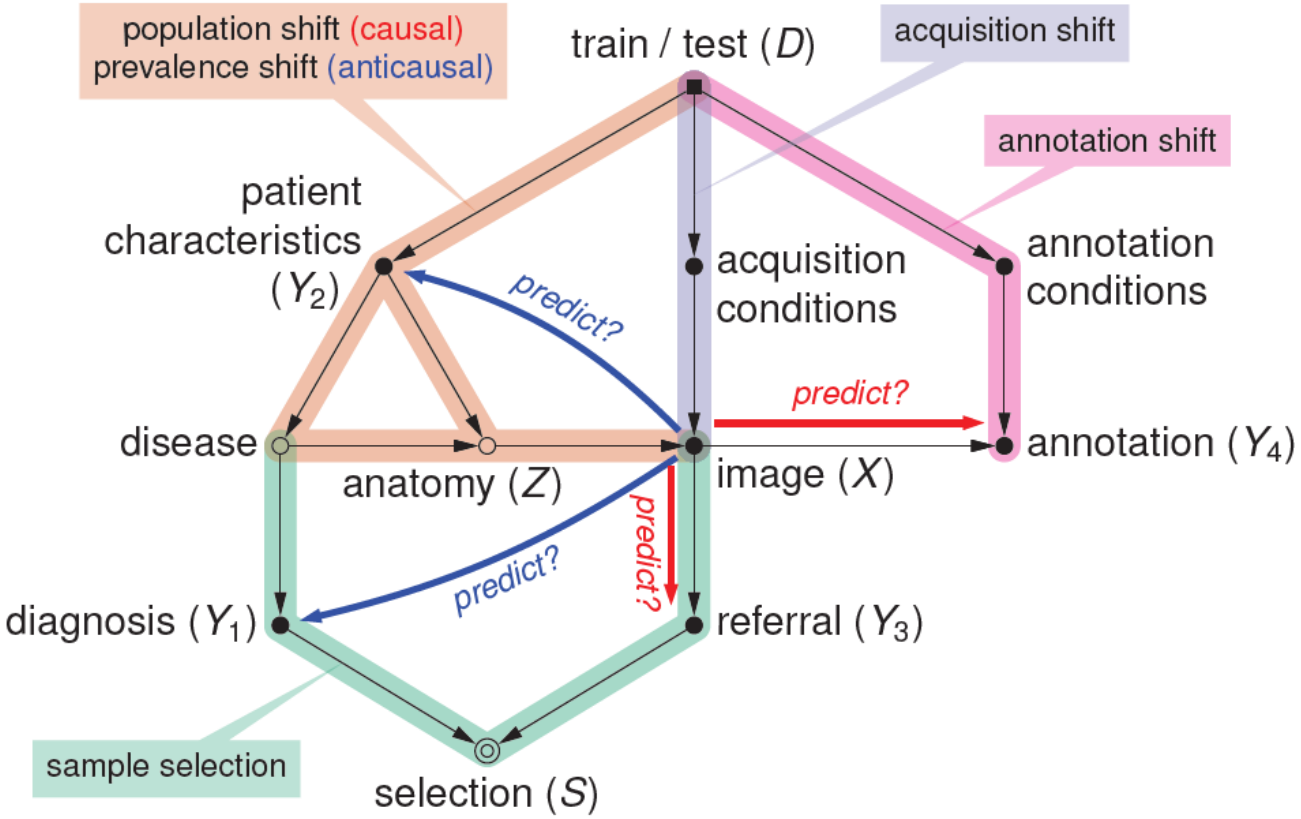
Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Recommendations

1. Gather meta-information about the data collection and annotation
2. Establish the predictive causal direction: $X \rightarrow Y$ vs. $Y \rightarrow X$
3. Identify evidence of mismatch between development and deployment
   - population shift, annotation shift, prevalence shift, manifestation shift
4. Verify what type of acquisition shift is expected
5. Determine whether the data collection was biased (sample selection)
6. Draw the full causal diagram and report it with your results

Castro, Walker and Glocker 2019. Causality Matters in Medical Imaging (arXiv:1912.08142)

Recommendations



Castro, Walker and Glocker 2019. Causality Matters in Medical Imaging (arXiv:1912.08142)

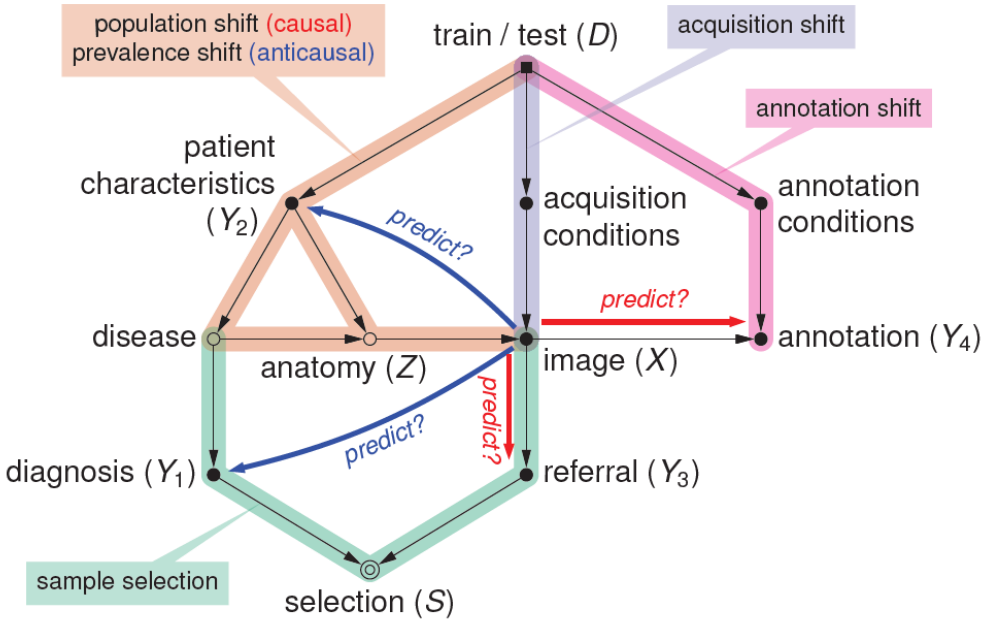
More on Causality

https://arxiv.org/abs/1912.08142

Causality matters in medical imaging

Daniel C. Castro*, Ian Walker, & Ben Glocker
Biomedical Image Analysis Group, Imperial College London, London SW7 2AZ, UK

This article discusses how the language of causality can shed new light on the major challenges in machine learning for medical imaging: 1) data scarcity, which is the limited availability of high-quality annotations, and 2) data mismatch, whereby a trained algorithm may fail to generalize in clinical practice. Looking at these challenges through the lens of causality allows decisions about data collection, annotation procedures, and learning strategies to be made (and scrutinized) more transparently. We discuss how causal relationships between images and annotations can not only have profound effects on the performance of predictive models, but may even dictate which learning strategies should be considered in the first place. For example, we conclude that semi-supervision may be unsuitable for image segmentation—one of the possibly surprising insights from our causal analysis, which is illustrated with representative real-world examples of computer-aided diagnosis (skin lesion classification in dermatology) and radiotherapy (automated contouring of tumours). We highlight that being aware of and accounting for the ... in medical imaging data is important for the safe develop... ...tial for regulation and responsible reporting. ... for future studies.



Daniel

Ian

Conclusions

- **Causal reasoning** is a useful tool to communicate (and scrutinize) our assumptions about the **data generating processes**
- Insights about key challenges such as **data scarcity** and **data mismatch**

→ Causal Representation Learning

Schölkopf 2019. Causality for Machine Learning (arXiv:1911.10500)

Acknowledgements

European Research Council
Established by the European Commission
erc
EPSRC
Engineering and Physical Sciences Research Council
Innovate UK
eHealth
SEVENTH FRAMEWORK PROGRAMME
European Commission Information Society and Media
MRC Medical Research Council
NHS National Institute for Health Research
HeartFlow®
KHEIRON MEDICAL TECHNOLOGIES