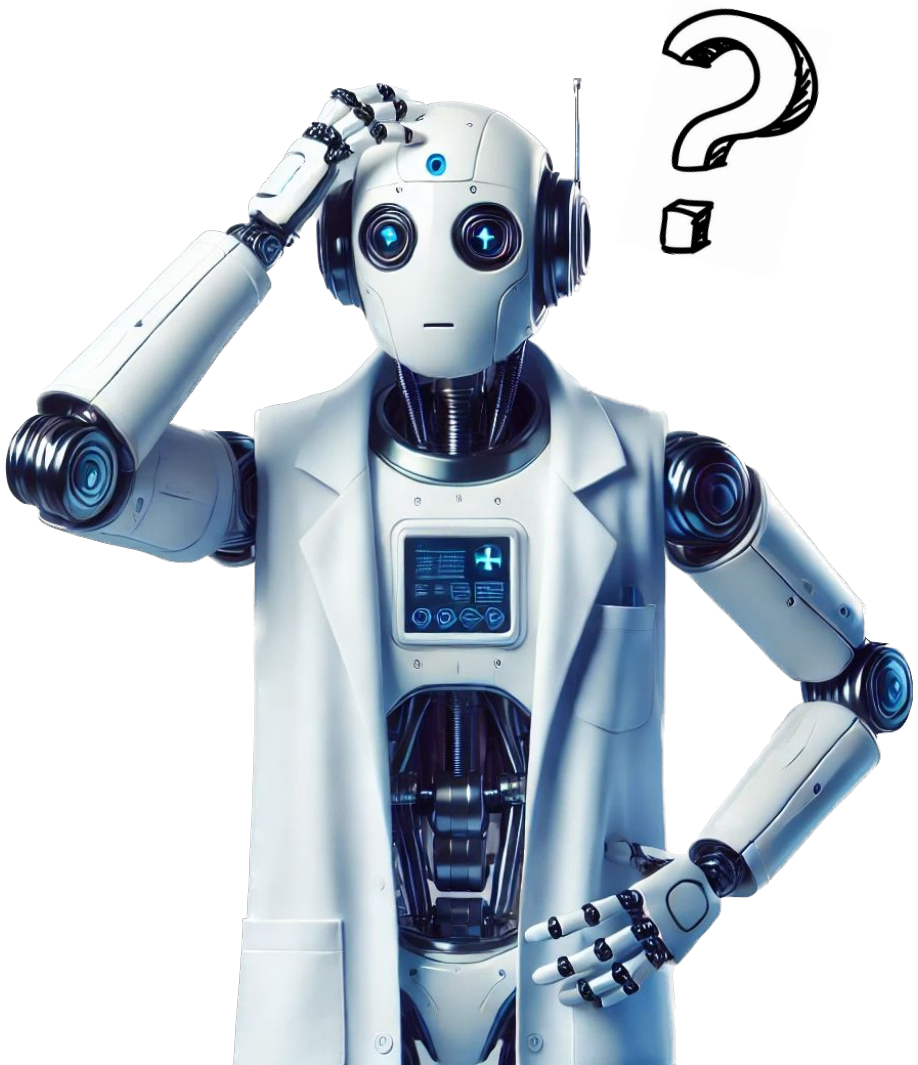
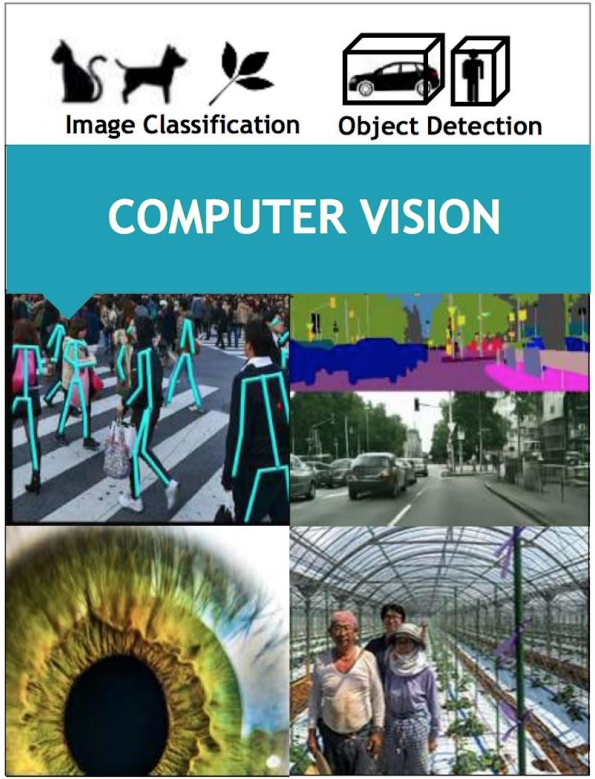
Applying Machine learning in the Wild: The Curious Case of Medical Imaging



Mohammad Havaei
mhavaei@google.com
July 2024

AI's success



AI's success

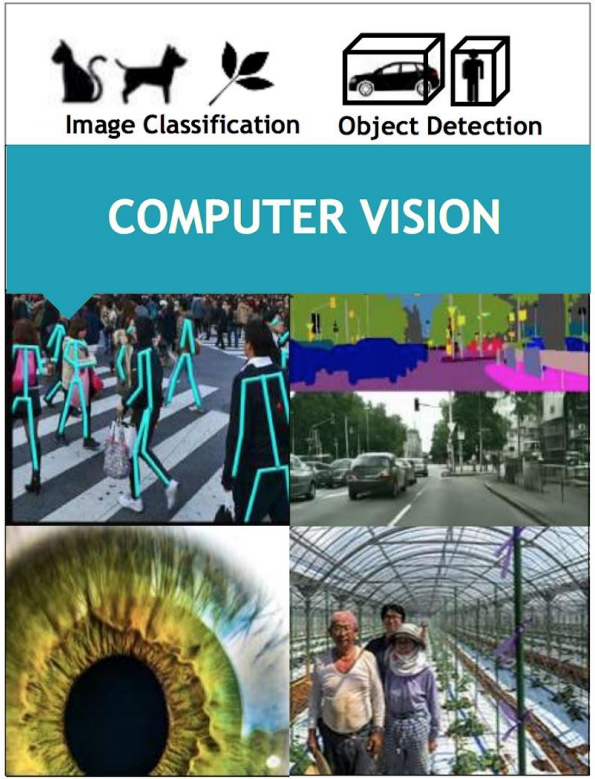
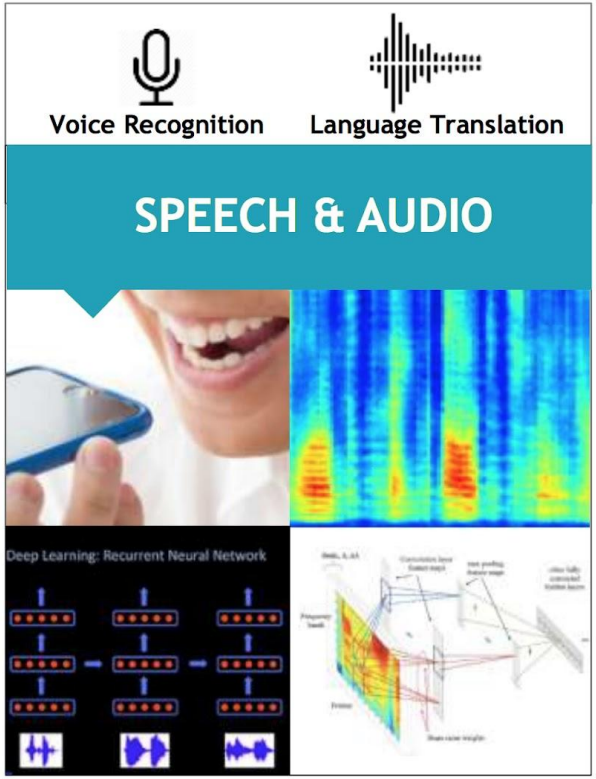


Image Classification
Object Detection
COMPUTER VISION


Voice Recognition
Language Translation
SPEECH & AUDIO
Deep Learning: Recurrent Neural Network

AI's success

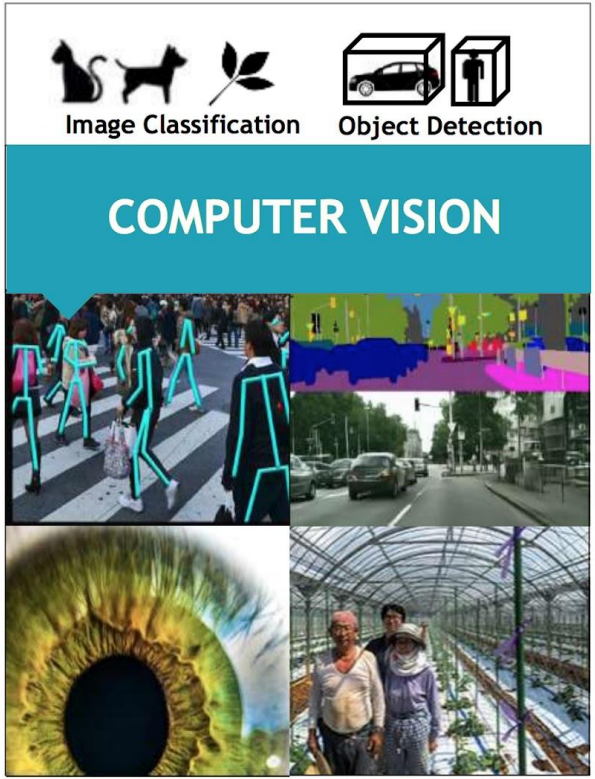
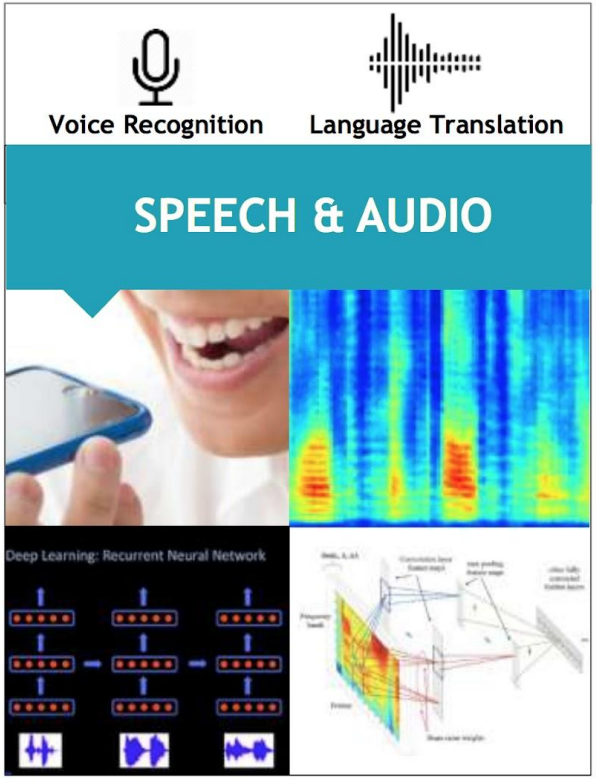
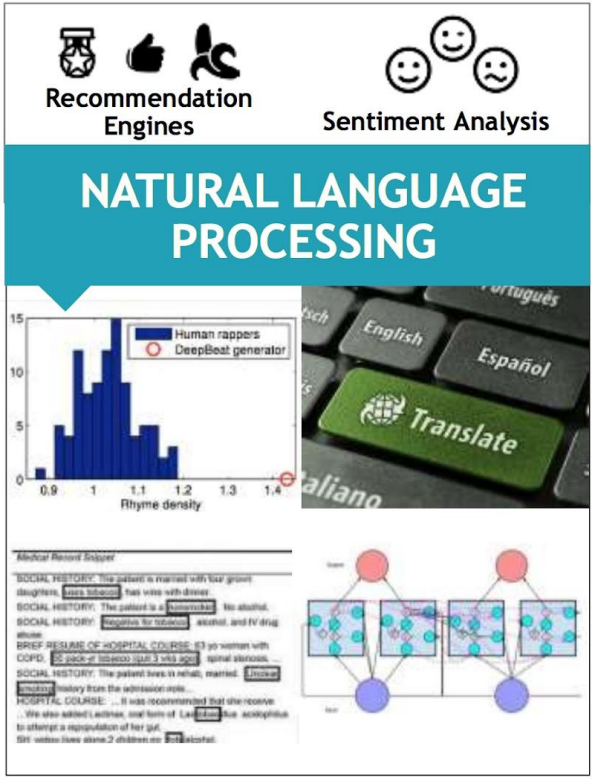
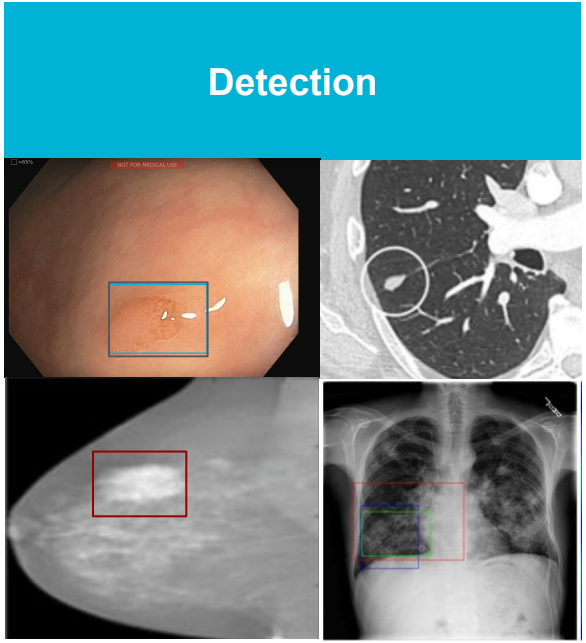


Image Classification
Object Detection
COMPUTER VISION


Voice Recognition
Language Translation
SPEECH & AUDIO
Deep Learning: Recurrent Neural Network

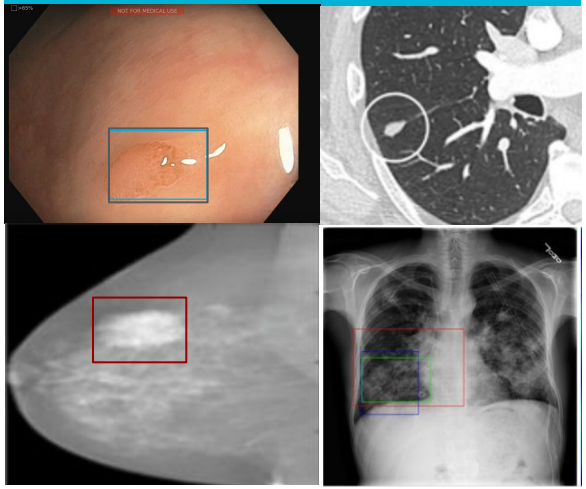

Recommendation Engines
Sentiment Analysis
NATURAL LANGUAGE PROCESSING
Human rappers
DeepBeat generator
Rhyme density
English
Español
Translate

AI applied to medical imaging

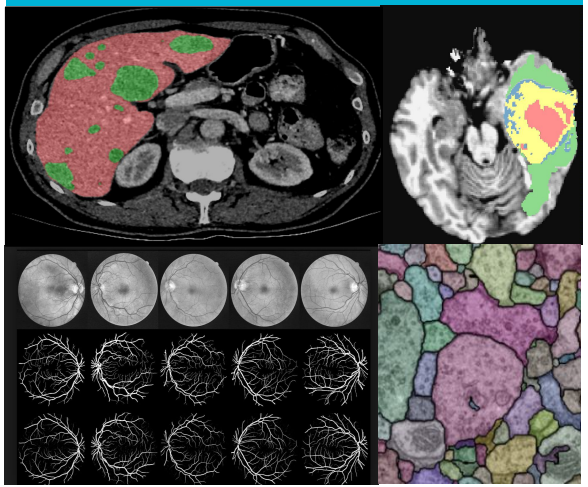


AI applied to medical imaging

Detection

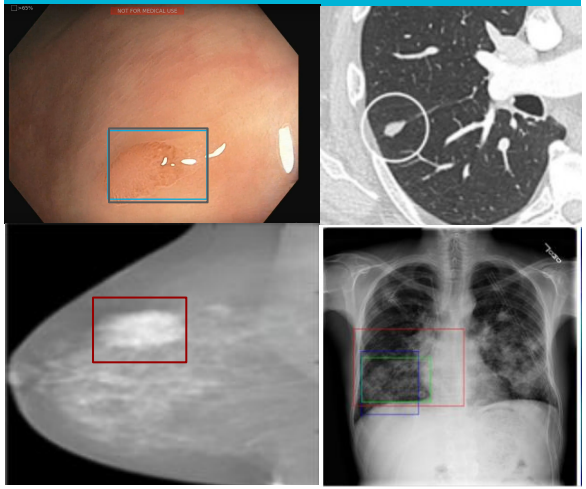


Segmentation

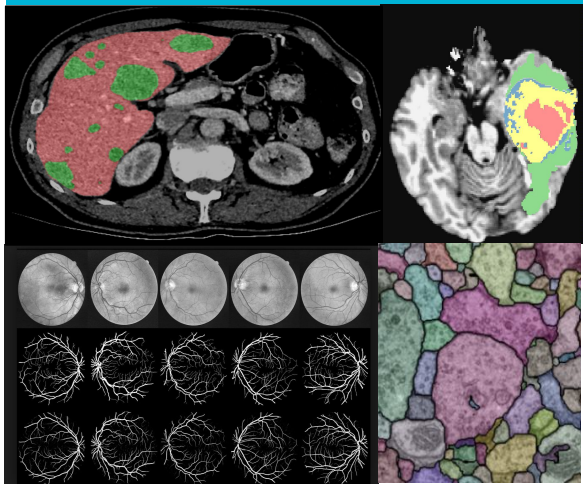


AI applied to medical imaging

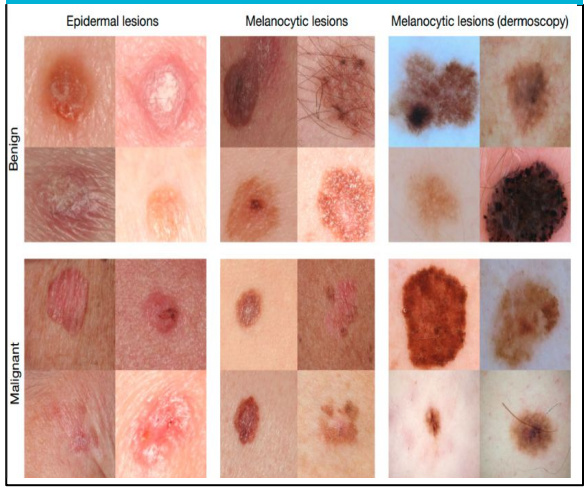
Detection



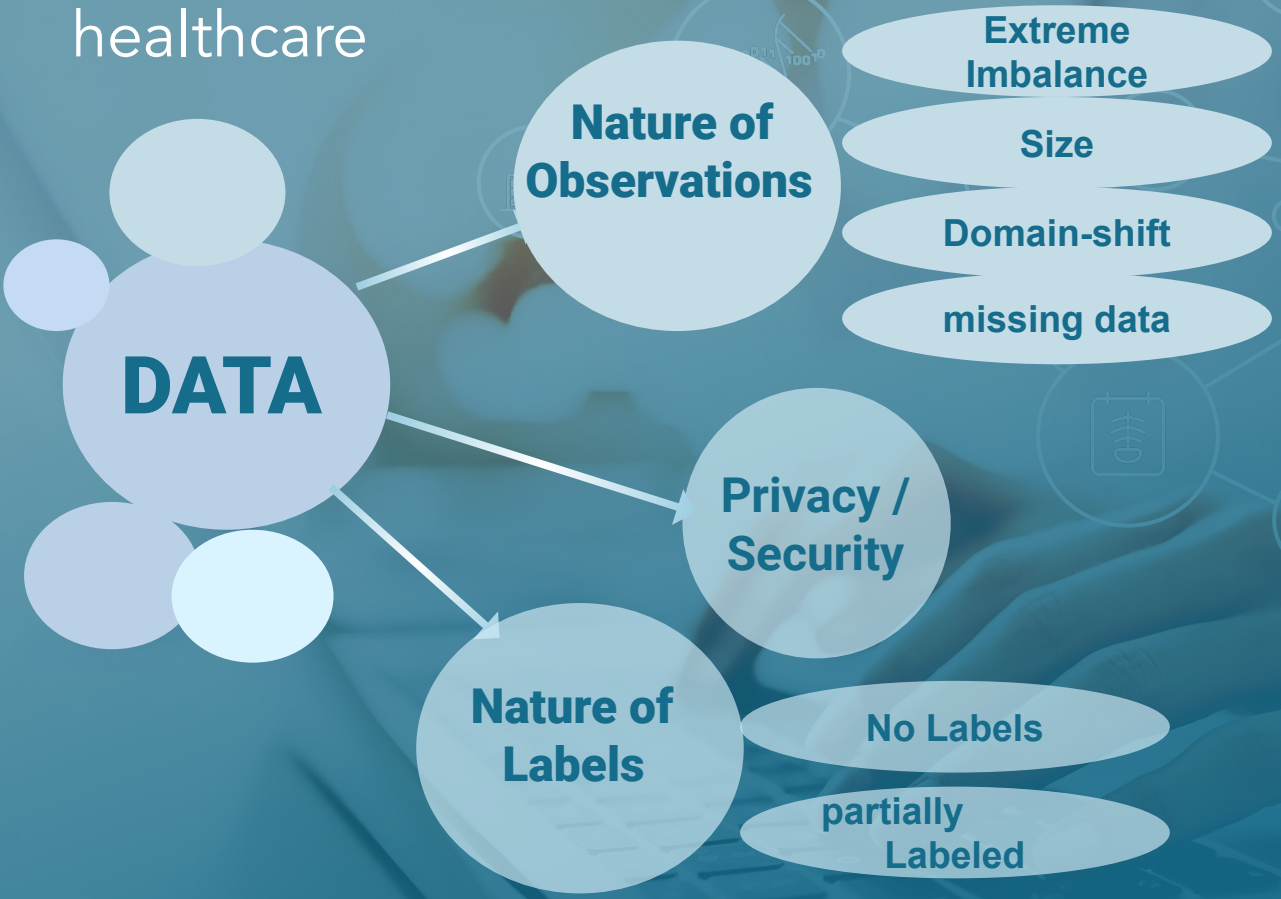
Segmentation



Diagnosis



Challenges of applying AI to healthcare



Nature of Observations

- Extreme imbalance w/r ROI

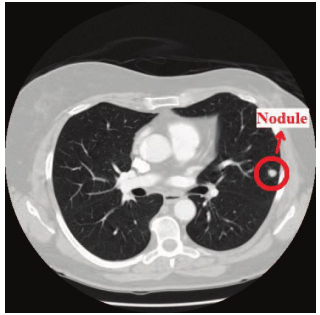
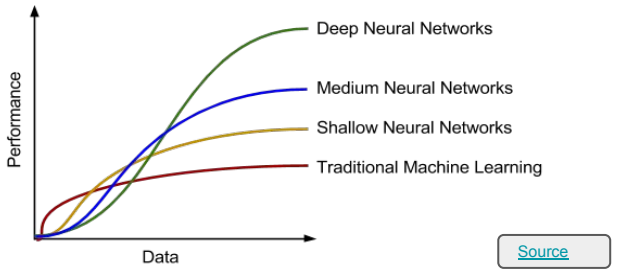


Figure from (Farahani et al, 2015)

Nature of Observations

- Extreme imbalance w/r ROI

- Small datasets
  - Medical data are usually orders of magnitude smaller than vision datasets

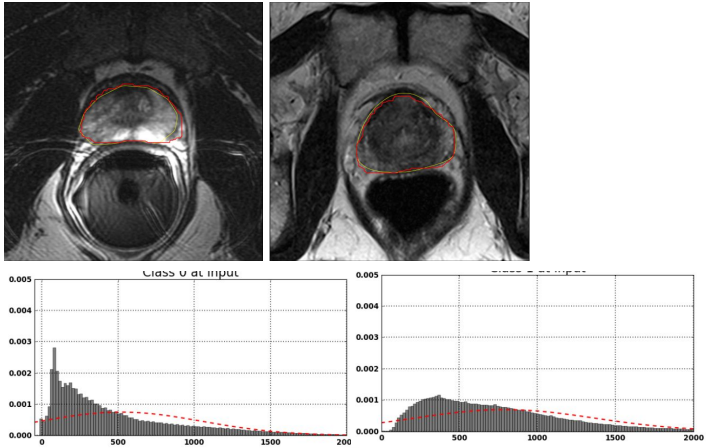


Nature of Observations

- Extreme imbalance w/r ROI

- Small datasets
  - Medical data are usually orders of magnitude smaller than vision datasets

- Distribution mismatch
  - Noise in data acquisition
  - Different image reconstruction algorithm
  - Data from multiple sources (different scanners)



Figures from (Drozdzal et al, 2017)

Nature of Observations

- Extreme imbalance w/r ROI

- Small datasets
  - Medical data are usually orders of magnitude smaller than vision datasets

- Distribution mismatch
  - Noise in data acquisition
  - Different image reconstruction algorithm
  - Data from multiple sources (different scanners)

- Missing data:
  - absence or corruption of pixel values or metadata, which can lead to biased models, reduced performance, and challenges in training robust and accurate diagnostic algorithms.

Clinical Data Reality:

Nature of Observations: Extreme Imbalance

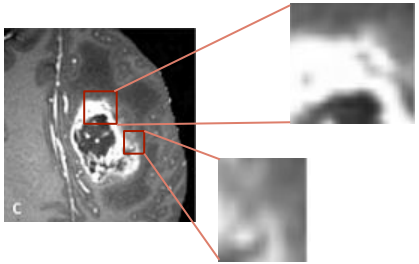
Some solutions:

- **Data intervention methods**
  - Resampling
  - Training on patches [Havaei 2015]
  - Mixup family augmentation: [Balanced-MixUp 2021]
  - Curriculum learning [Cased 2017]
  - Synthetic data augmentation: generated data for under-represented class

- **Loss intervention**
  - class weighted cross entropy
  - focal loss
  - instance weighted cross entropy for both segmentation and classification

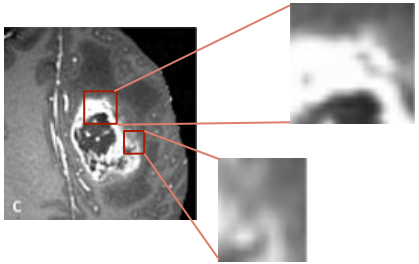
Nature of Observations: Extreme Imbalance

- **Training with patches**
  - Sample patches Instead of training full image
  - Allows us to create balanced batches for training

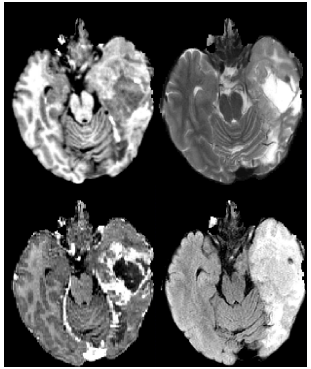


Nature of Observations: Extreme Imbalance

- **Training with patches**
  - Sample patches Instead of training full image
  - Allows us to create balanced batches for training



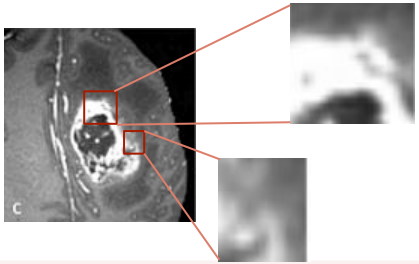
(Havaei et al 2015)



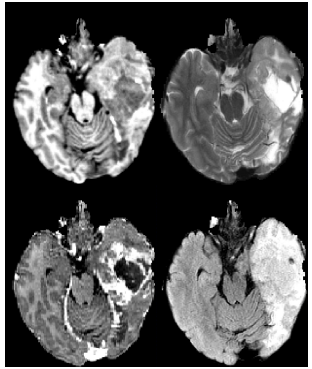
Train mostly on patches with **high entropy (multiple labels in a patch)** in the label spaces.

Nature of Observations: Extreme Imbalance

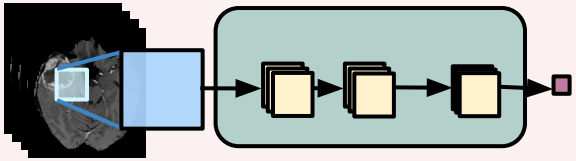
- **Training with patches**
  - Sample patches Instead of training full image
  - Allows us to create balanced batches for training



(Havaei et al 2015)



Train:

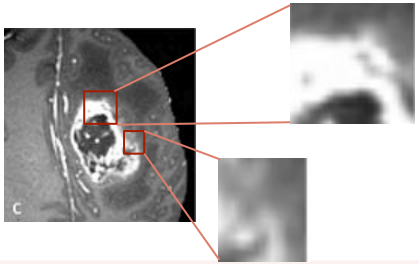


Training First phase

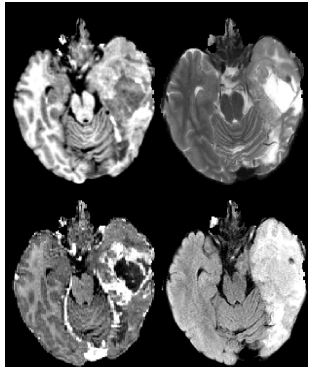
- Uniform distribution over classes
- patches

Nature of Observations: Extreme Imbalance

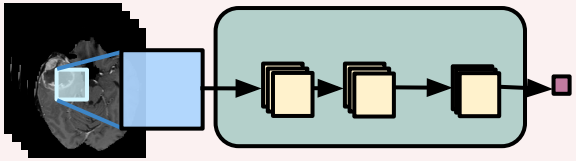
- **Training with patches**
  - Sample patches Instead of training full image
  - Allows us to create balanced batches for training

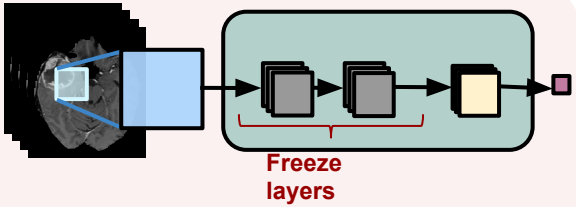


(Havaei et al 2015)



Train:

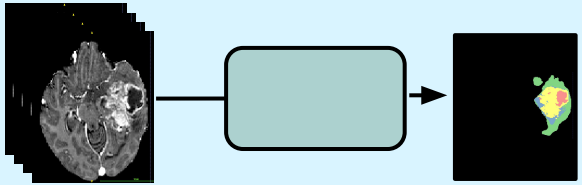


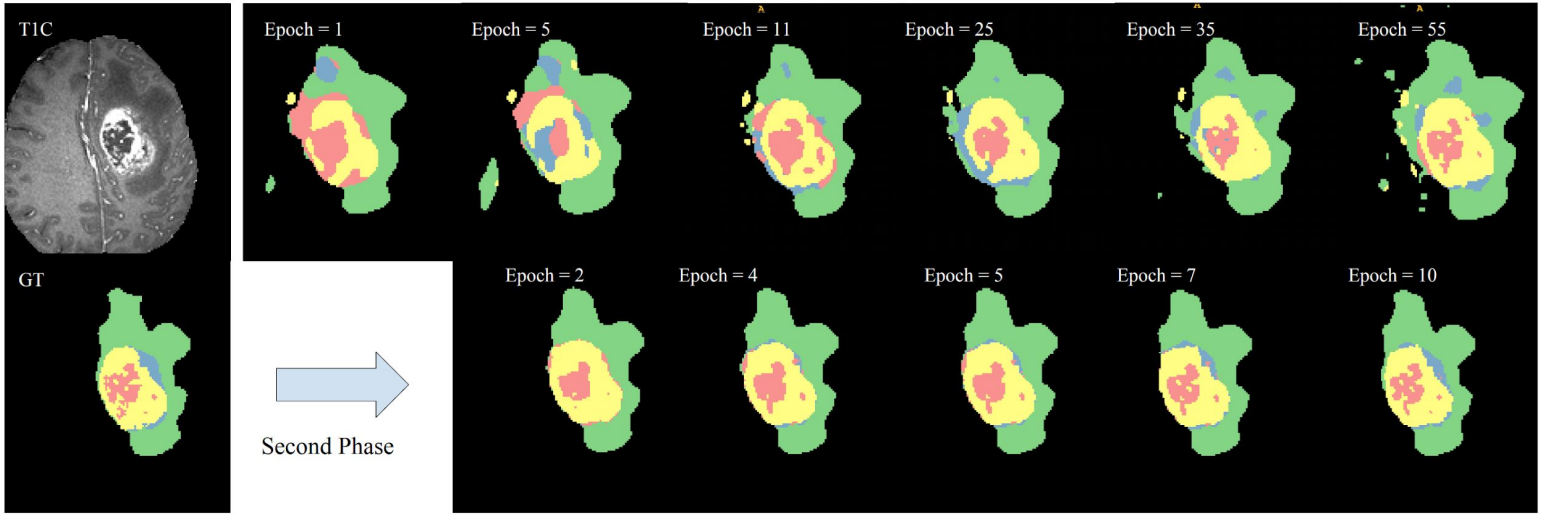


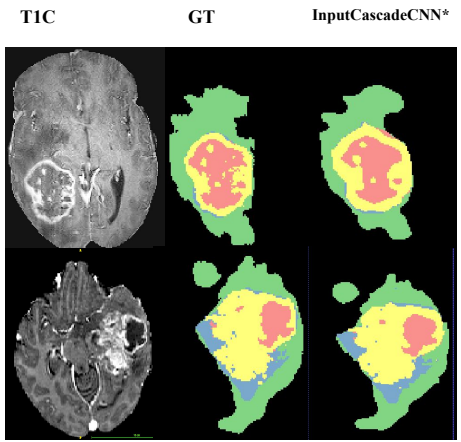
Training First phase
- Uniform distribution over classes
- patches

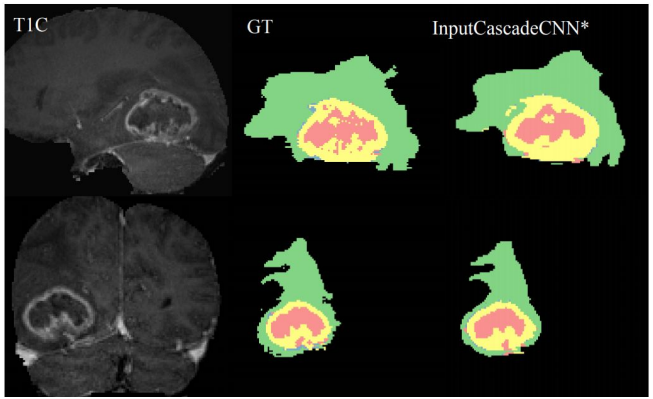
Training second phase
- Real distribution
- patches

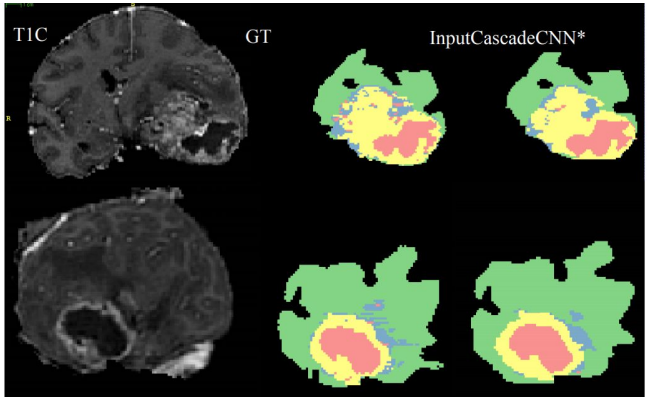
Test:




T1C
GT
Epoch = 1
Epoch = 5
Epoch = 11
Epoch = 25
Epoch = 35
Epoch = 55
Second Phase
Epoch = 2
Epoch = 4
Epoch = 5
Epoch = 7
Epoch = 10


T1C
GT
InputCascadeCNN*

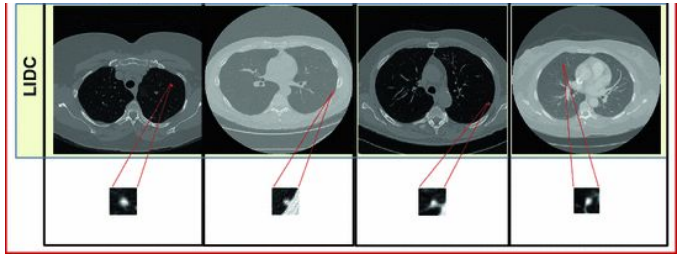

T1C
GT
InputCascadeCNN*

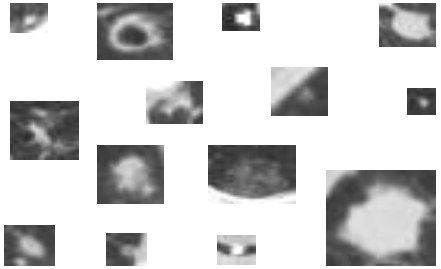

T1C
GT
InputCascadeCNN*

Nature of Observations: Extreme Imbalance

- **Curriculum learning :**
  - Start the training with easy patches,
  - move to more difficult examples as training progresses.

(Jesson et al 2017)

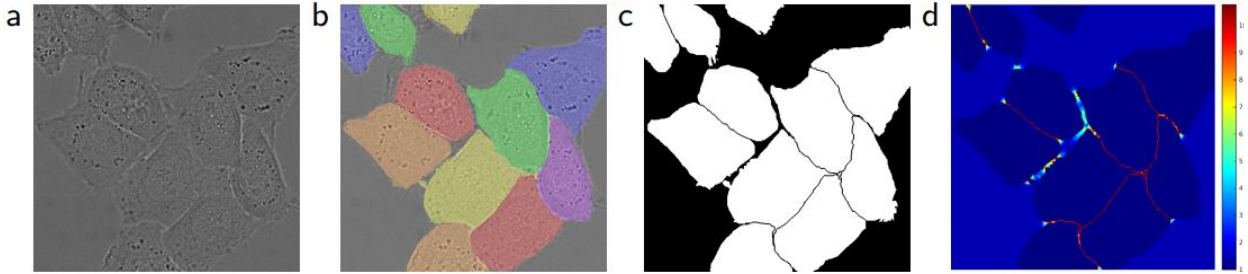




Method	False Positives Per Scan 0.125	0.25	0.5	1	2	4	8	**Average**
CASED	0.759	0.825	0.866	0.903	0.926	0.946	0.959	**0.883**
CASED-Sphere	0.701	0.762	0.813	0.862	0.897	0.923	0.939	0.842
AIDENCE*	0.739	0.788	0.828	0.879	0.910	0.938	0.963	0.864
ZNET [15]	0.661	0.724	0.779	0.831	0.872	0.892	0.915	0.811
ETROCAD [18,15]	0.250	0.522	0.651	0.752	0.811	0.856	0.887	0.676
M5LCAD [11,15]	0.306	0.360	0.540	0.691	0.762	0.797	0.798	0.608

Nature of Observations: Extreme Imbalance

- **Use pixel weighted loss function**
  - Give more importance to the ROI



Pixel wise weight

$$E = \sum_{\mathbf{x} \in \Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x}))$$

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right)$$

Used by UNET

(Ronneberger et al 2015)

Nature of Observations: Small Datasets

Why small datasets are a problem for Deep Learning Models?

Answer: Prone to overfitting

Some solutions:

a) Regularization methods: i) Dropout, ii) weight decay (L1,L2) , iii) Early stopping iv) Ensembles v) reducing number of parameters

b) Cross validation

c) Increase the size of the data: Expensive, sometimes not possible

d) Transfer learning

e) Synthetic data augmentation: very attractive with the advent of strong generative models

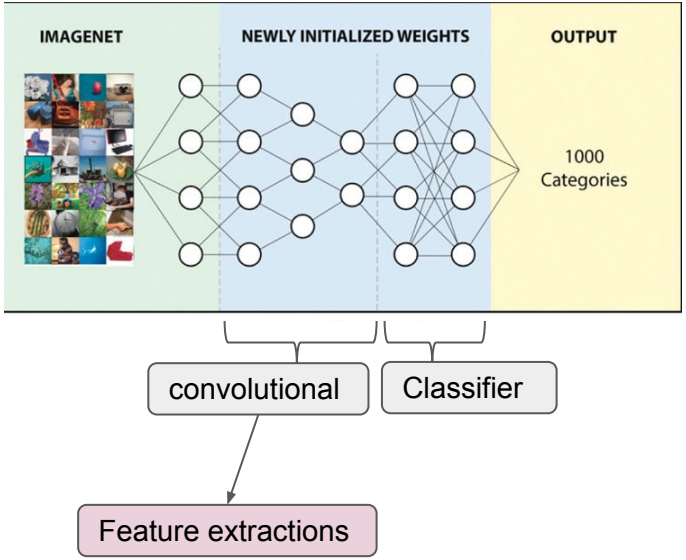
f) Federated Learning: when privacy is at play. will be discussed in the next section

Nature of Observations: Small Datasets

- **Transfer Learning:**

  - Once the training is complete, discard classifier or class specific layers.

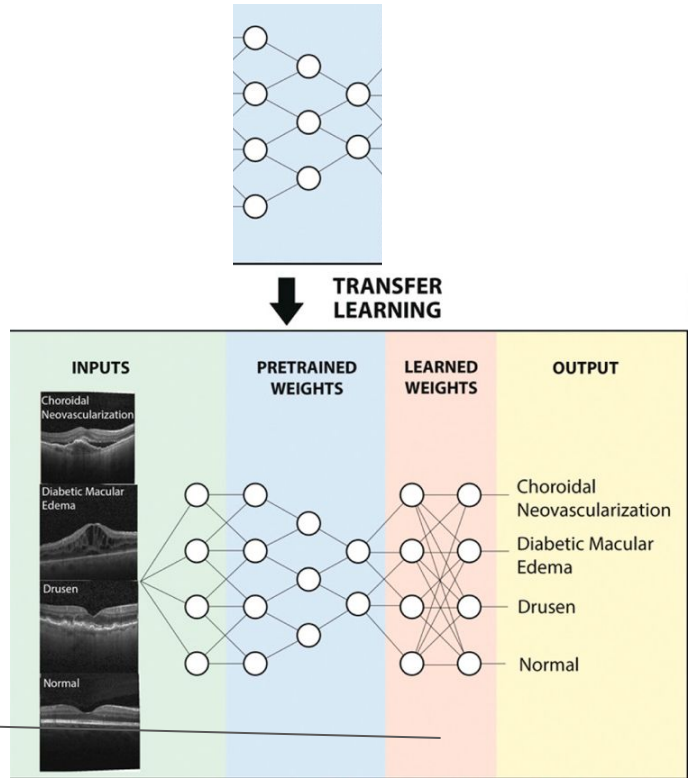
  - Use the feature extraction layers for the small dataset



We could also use **self-supervised learning (unsupervised learning)** to pre-train the weights

Nature of Observations: Small Datasets

- **Pre-training:**



Initialized weights for the new task

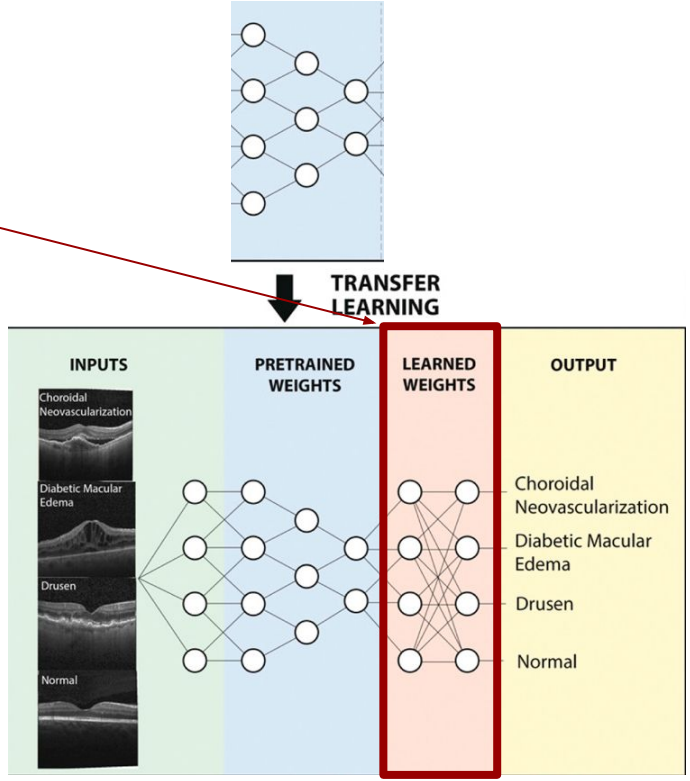
Nature of Observations: Small Datasets

- **Pre-training:**

Pretrained weights

Feature extraction

Fine-tune



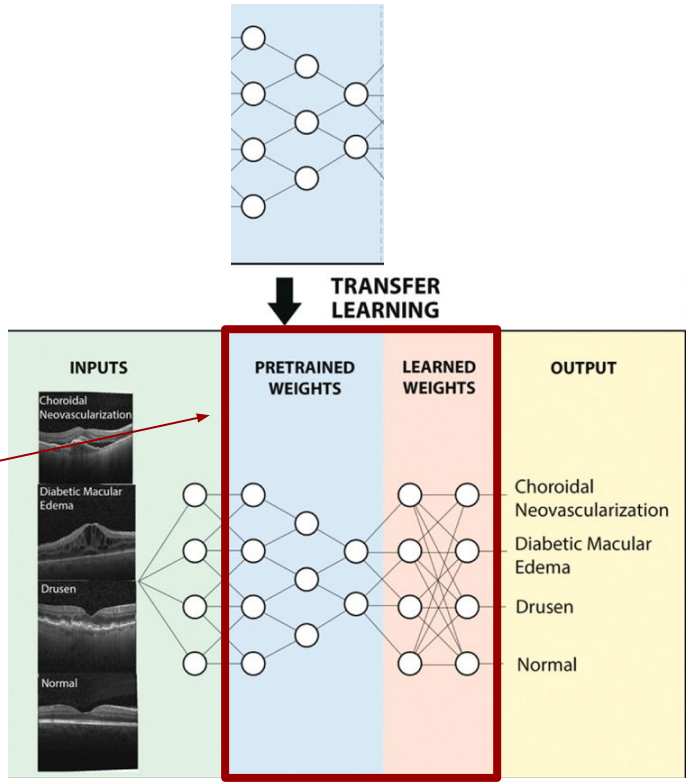
Nature of Observations: Small Datasets

- **Pre-training:**

Pretrained weights

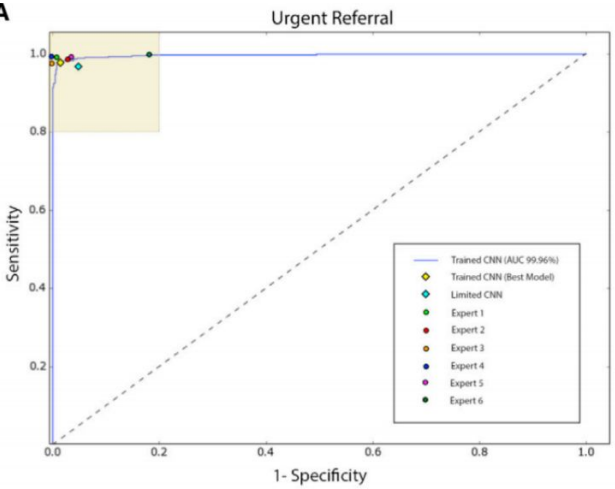
Feature extraction

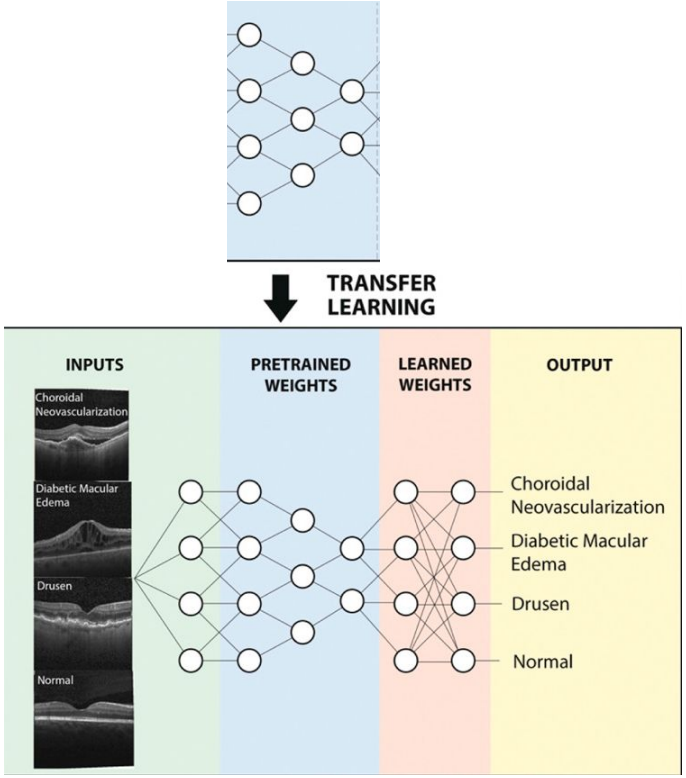
Fine-tune



Nature of Observations: Small Datasets

- **Pre-training :**





(Kermany et al, 2018)

Transfer learning in medical imaging

Mixed Results:

- Transfer learning for medical images using pretrained weights from natural images is inconsistent.
- Improvements are seen with small datasets and similar domains (Kermany et al., 2018; Matsoukas et al., 2022).
- Limited performance gains with significant domain shifts, even in small datasets (Transfusion 2019).

Feature Reuse:

- Effective feature reuse occurs primarily in very low layers (Transfusion 2019).
- ViTs benefit more from early layer feature reuse, while CNNs benefit throughout the network (Matsoukas et al., 2022).
- Transfer learning on natural images shows limited practical use in few-shot learning setups.

Domain Similarity:

- The closer the pretraining domain to the target domain, the better the results (Matsoukas et al., 2022).
- Smaller, simpler models can perform comparably to larger, pretrained models (Transfusion 2019).

This has motivated the research community in investing in foundation models for medical image analysis.

Transfer learning in medical imaging

Big Self-Supervised Models Advance Medical Image Classifications

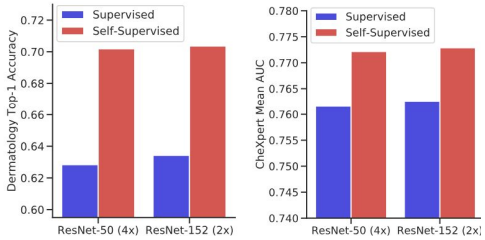


Figure 2: Comparison of supervised and self-supervised pretrai ing, followed by supervised fine-tuning using two architectures c dermatology and chest X-ray classification. Self-supervised lear ing utilizes unlabeled domain-specific medical images and signi icantly outperforms supervised ImageNet pretraining.

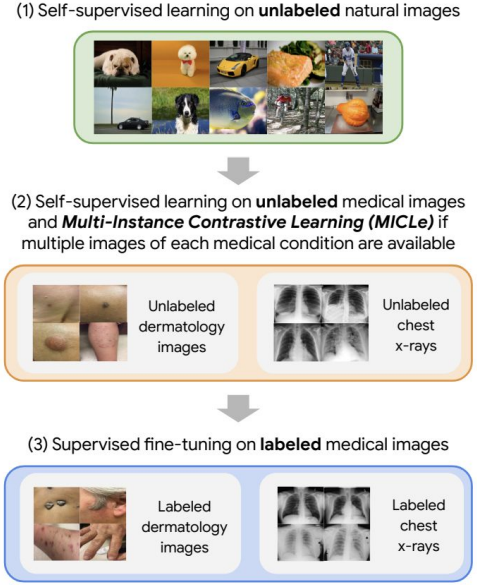


Table 3: Comparison of best self-supervised models *vs.* supervised pretraining baselines on dermatology classification.

Architecture	Method	Pretraining Dataset	Top-1 Accuracy
ResNet-152 (2×)	Supervised	ImageNet	63.36 ± 0.12
ResNet-101 (3×)	BiT [23]	ImageNet-21k	68.45 ± 0.29
ResNet-152 (2×)	SimCLR	ImageNet	66.38 ± 0.03
ResNet-152 (2×)	SimCLR	ImageNet→Derm	69.43 ± 0.43
ResNet-152 (2×)	MICLe	ImageNet→Derm	**70.02** ± 0.22

Table 4: Comparison of best self-supervised models *vs.* supervised pretraining baselines on chest X-ray classification.

Architecture	Method	Pretraining Dataset	Mean AUC
ResNet-152 (2×)	Supervised	ImageNet	0.7625 ± 0.001
ResNet-101 (3×)	BiT [23]	ImageNet-21k	0.7720 ± 0.002
ResNet-152 (2×)	SimCLR	ImageNet	0.7671 ± 0.008
ResNet-152 (2×)	SimCLR	CheXpert	0.7702 ± 0.001
ResNet-152 (2×)	SimCLR	ImageNet→CheXpert	**0.7729** ± 0.001

Transfer learning in medical imaging

Self-supervised Learning from 100 Million Medical Images

Big Self-Supervised Models are Strong Semi-Supervised Learners

Synthetic data augmentation

Another way to increase size of the dataset is by using synthetic data points.

- Generated from noise
  - conditional generation
    - Control on not only the class
    - but also various style of the generations.
    - Generate the class of interest in various context. This helps generalization
    - generate counterfactual examples
    - a lot of research on generative models with focus on disentanglement e.g. style and content and counterfactual image generations
  - two papers SGAN and MIDL PAPER

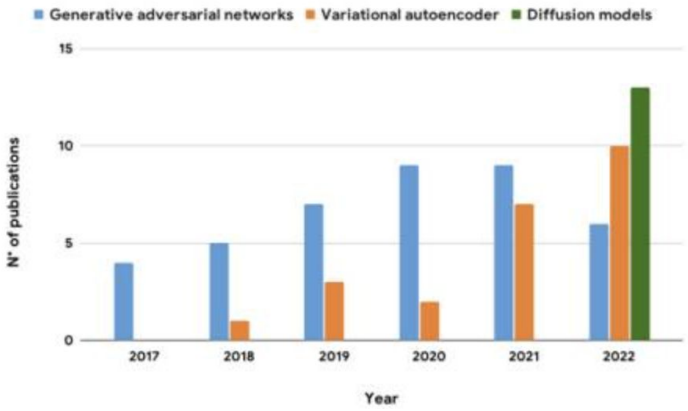
- via Domain Translation (will be covered in the next section)

Synthetic data augmentation

Another way to increase size of the dataset is by using synthetic data points.

- Generated from noise
  - conditional generation
    - Control on not only the class
    - but also various style of the generations.
    - Generate the class of interest in various context. This helps generalization
    - generate counterfactual examples
    - a lot of research on generative models with focus on disentanglement e.g. style and content and counterfactual image generations
  - two papers SGAN and MIDL PAPER

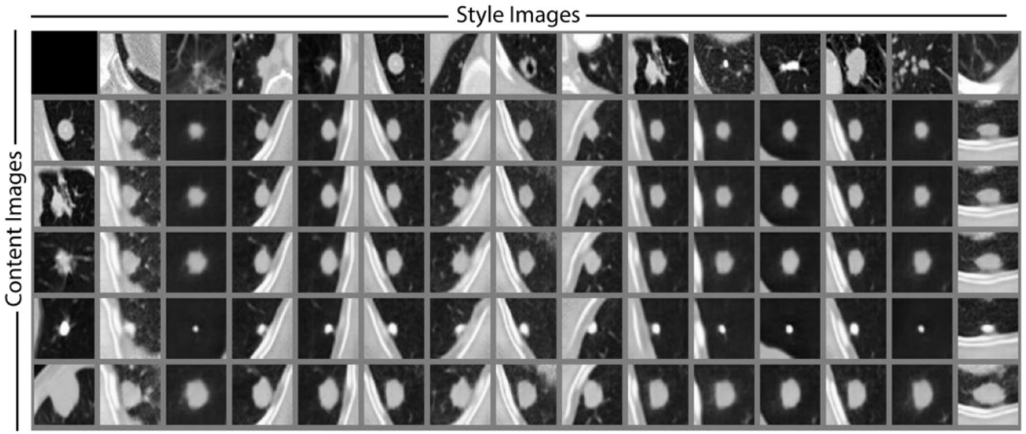
- via Domain Translation (will be covered in the next section)



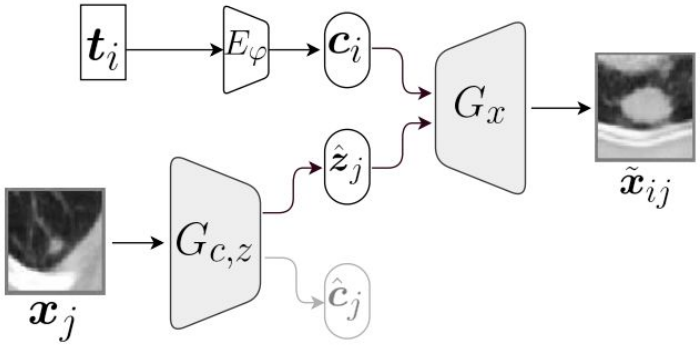
Synthetic data augmentation

DRAI (Havaei et al 2021)

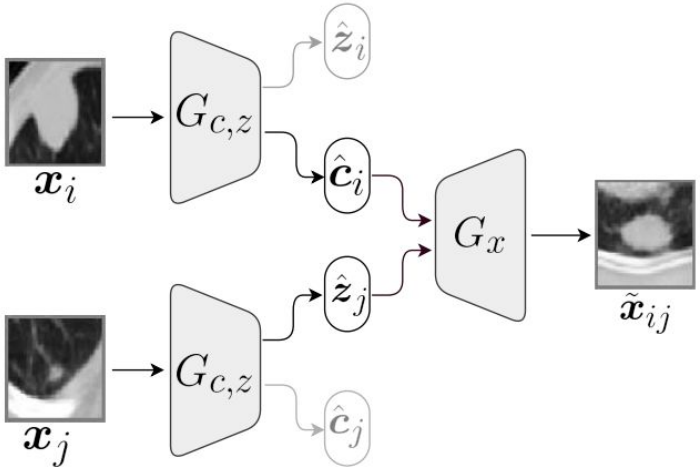
- One trained it has interesting capabilities.
- In addition to conditional generation using attributes t, we can control the generation with style images or content images.



DRAI

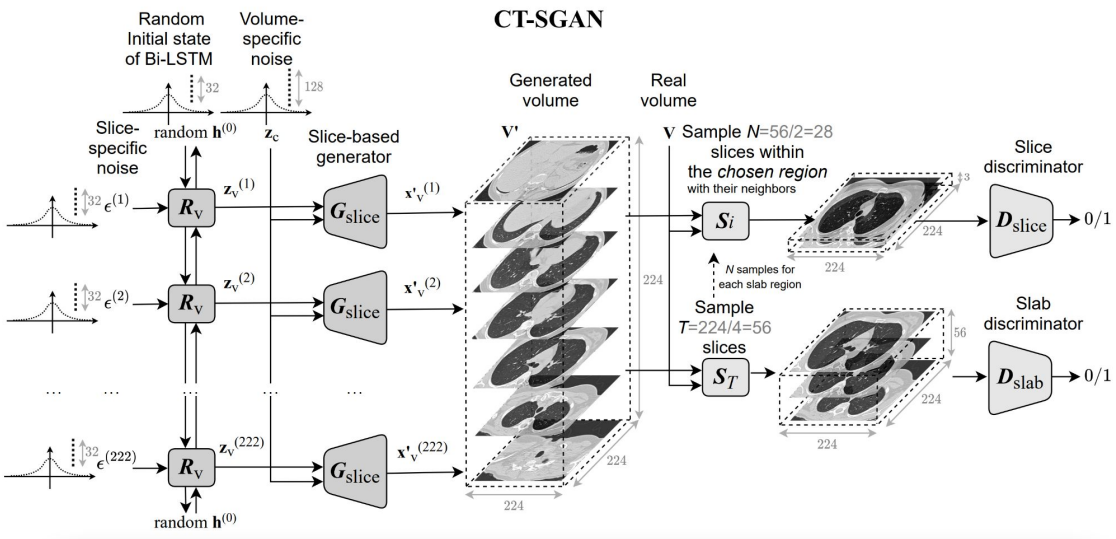


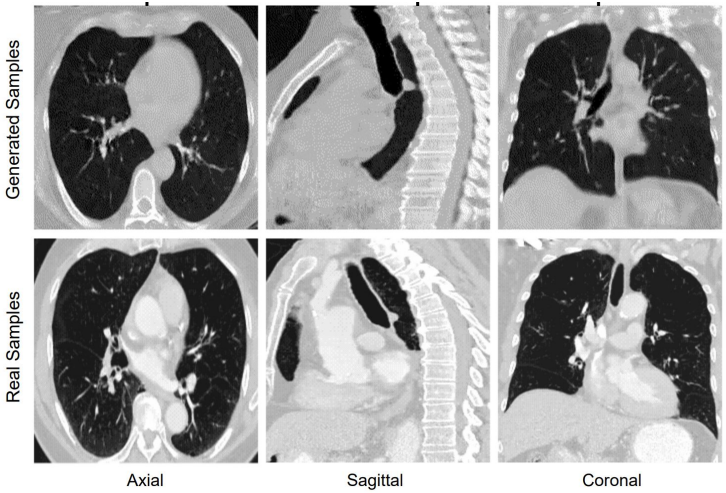
(a)



CTSGAN (Pesaranghader et al Havaei 2021)

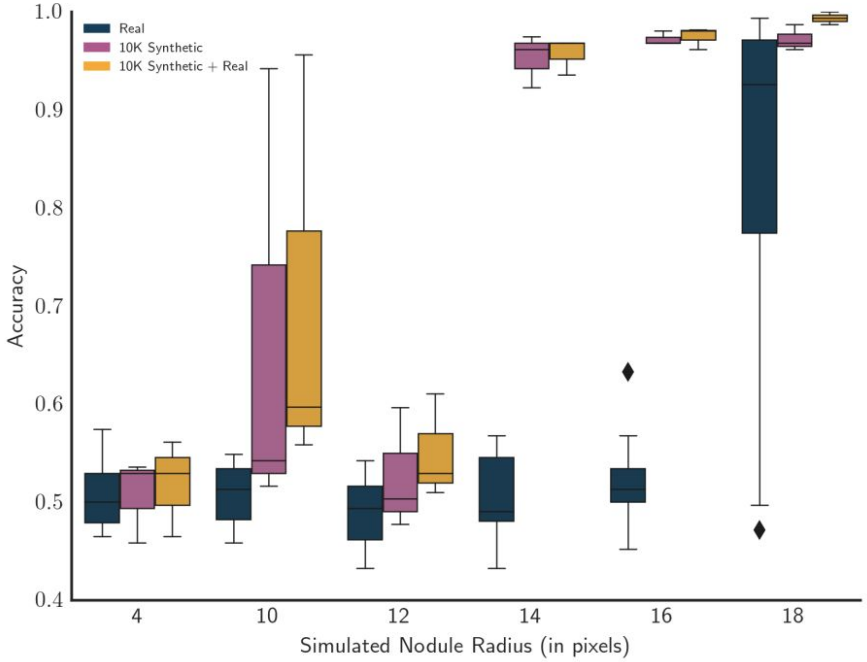
Generate high resolution (224x224x224) lung CT volumes using Bi-LSTM and GAN

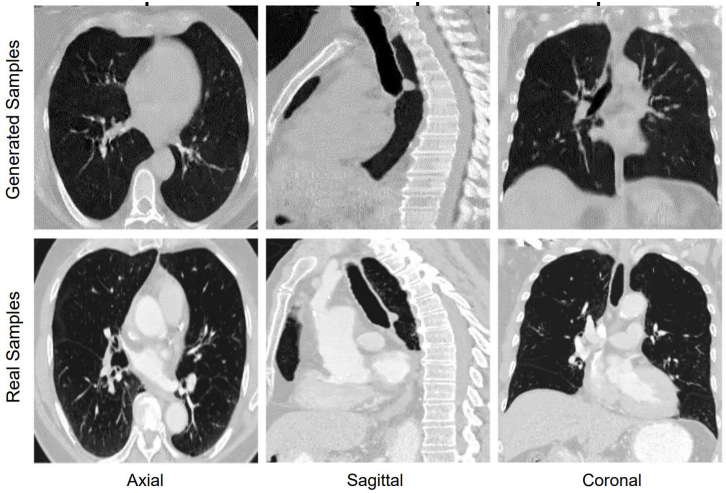




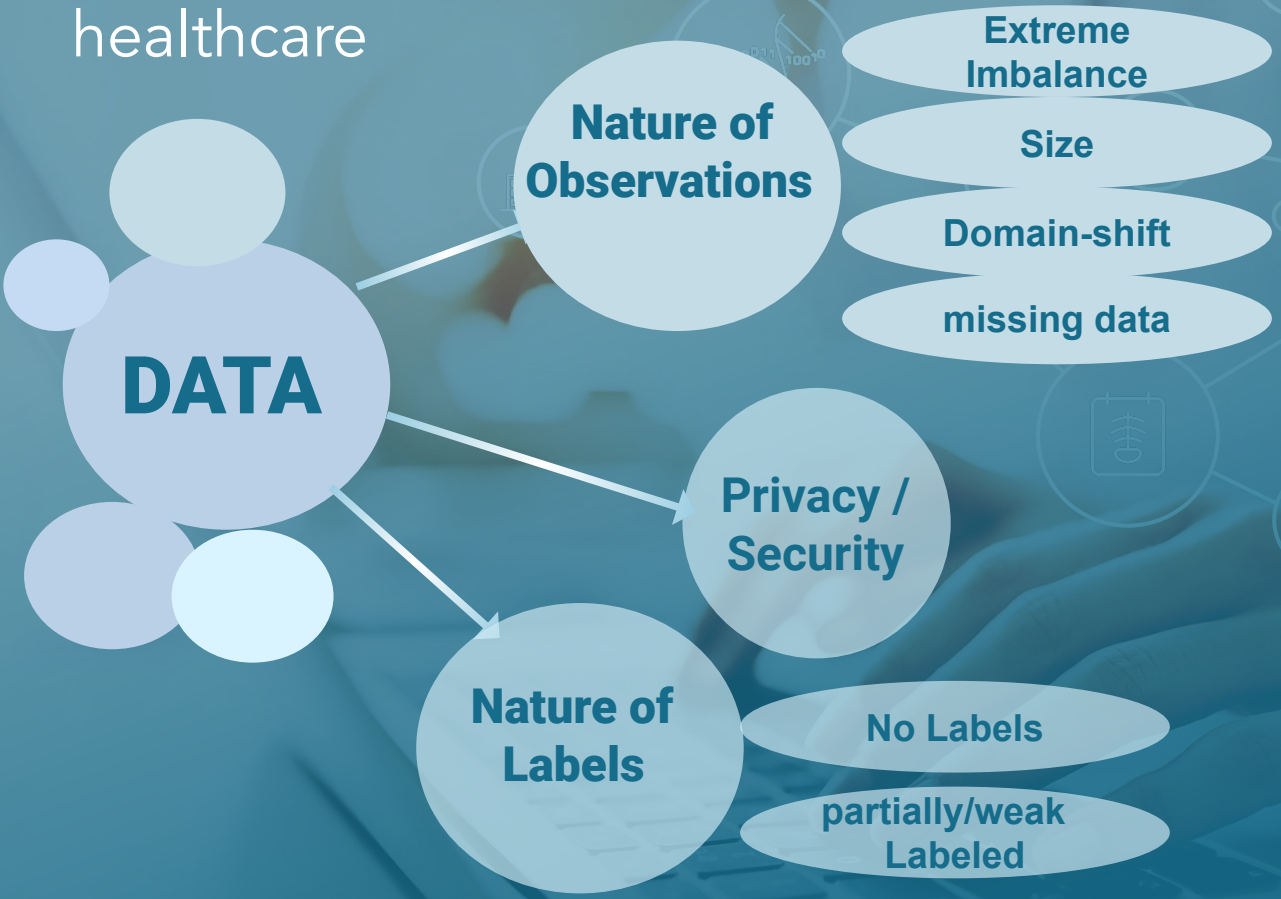
CTSGAN (Pesaranghader et al Havaei 2021)

Evaluate whether pre-training on synthetic data can improve nodule detection task.





Challenges of applying AI to healthcare



Domain Shift in Medical Imaging

- Hardware Acquisition Noise: Differences in imaging equipment (e.g., MRI machines, CT scanners).
  - apply data normalization methods
  - Domain adaptation methods
- Sample Bias:Differences in patient demographics, disease prevalence, and clinical settings.
  - Include data from various groups
  - causal representation learning to avoid counterfactual shortcuts
- Image Processing Differences:
- Protocol Variations: Different standards for imaging in different countries or states
- Temporal Shift: Distribution drift.
- Geographic Differences:
- Annotation Variability:

Domain Shift in Medical Imaging

- Hardware Aquisition Noise: Differences in imaging equipment (e.g., MRI machines, CT scanners).
  - apply data normalization methods
  - Domain adaptation methods

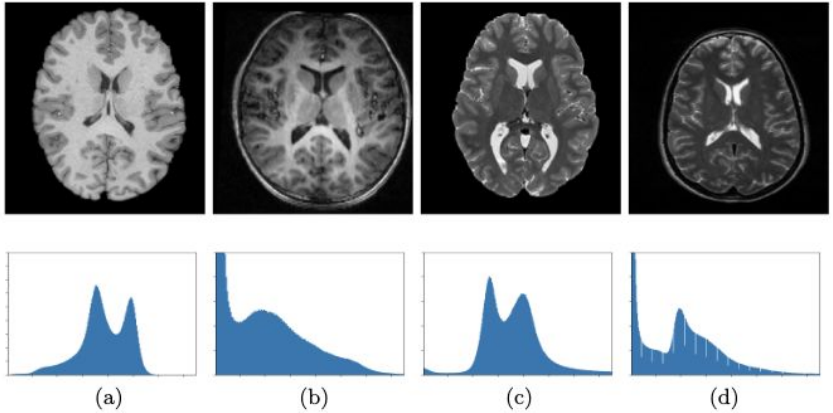


Fig. 2. Image slices (top) and corresponding intensity distribution (bottom) of normalized T1-weighted (a, b) and T2-weighted (c, d) MRIs from different scanners. Image courtesy to N. Karani [29].

(Guan et al 2021)

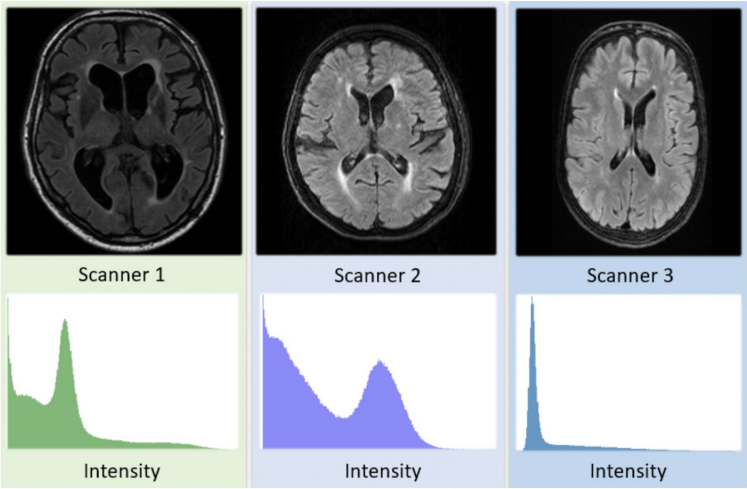
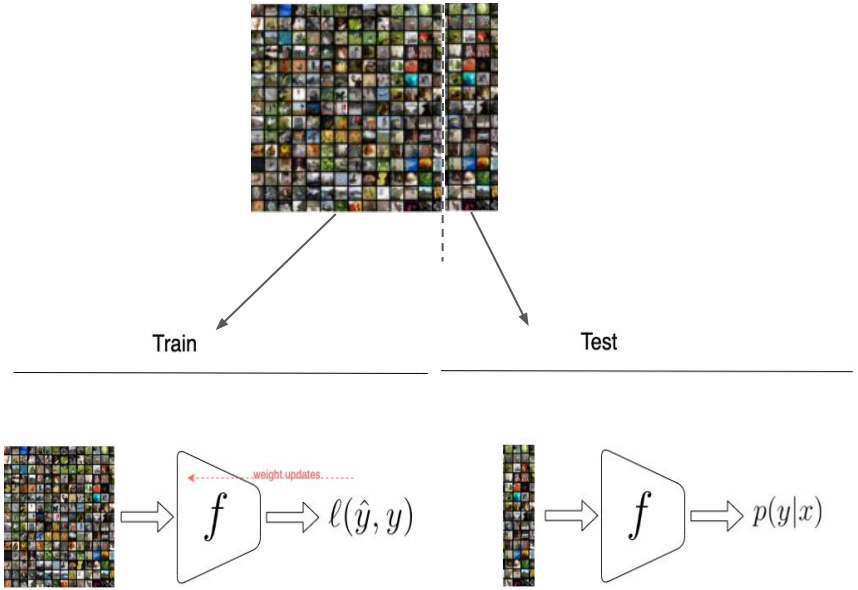


Illustration of domain shift in intensity distribution, contrast level and noise level from a MRI FLAIR sequence acquired at different centres.

(Li et al 2020)

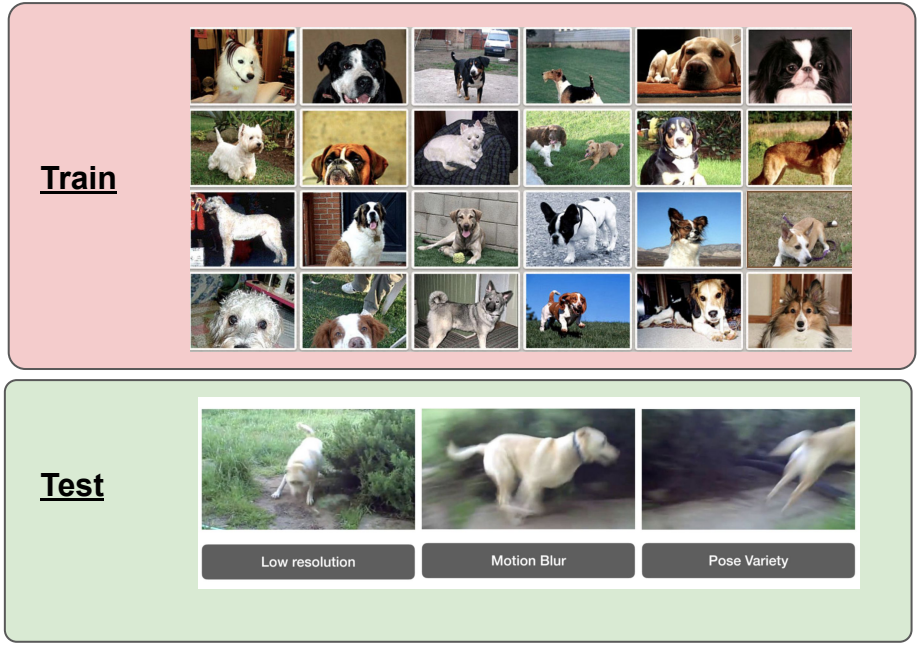
IID assumption in training models



We assume that “Train” and “Test” data are drawn from the same distribution.

$$P_{source}(X) = P_{target}(X)$$

Real world applications:



$$P_{source}(X) \neq P_{target}(X)$$

source:https://drive.google.com/file/d/1ZKEfwgbIEQ7LaPileg1VhPxxLTvls1zG/view

Domain Adaptation

Train on **source** , test on **Target**

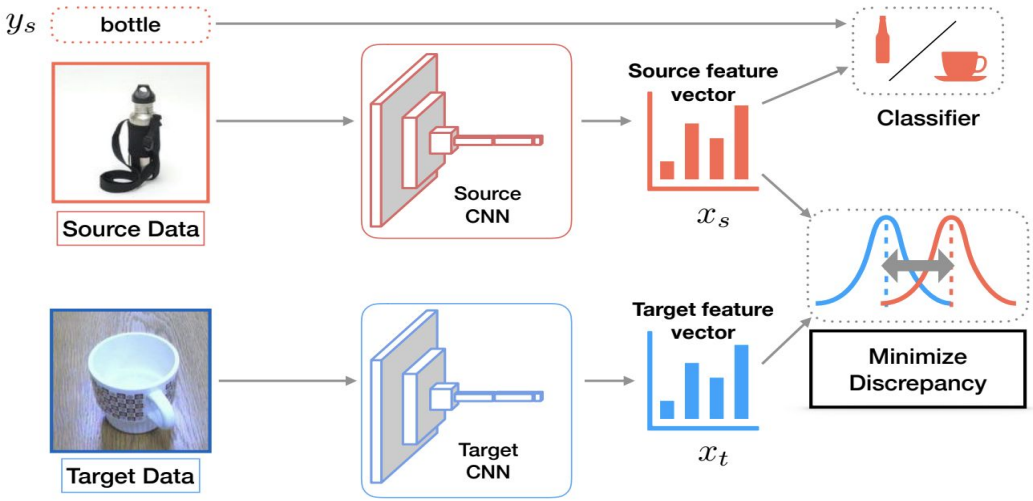
- $\mathcal{D}_s = \{(x_i^S, y_i^S)\}_{i=1}^{N_s}$ is the labeled source data
- $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$ is the unlabeled target data
- $P_S(Y|X) = P_T(Y|X)$ label distribution between source and target remains the same and only the marginals are different $P_S(X) \neq P_T(X)$

General approach to Domain Adaptation

We are interested in having low $R_T(h)$. We can define an upper bound and minimize it.

$$\min_h R_T(h) =$$
$$\min_h R_S(h) + \mathrm{disp}_h(\hat{P}_S, \hat{P}_T)$$

(Ben-David et al., 2010)
(Zhang et al., 2019)

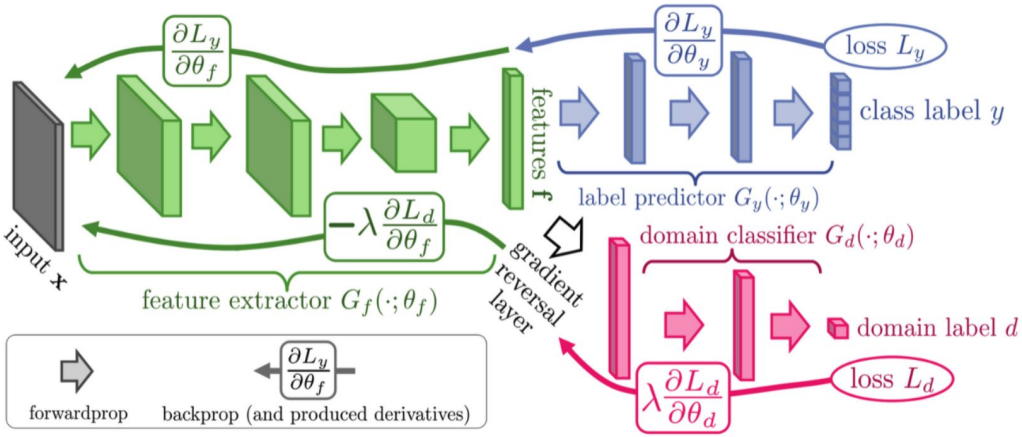


$\hat{P}_S$ Empirical Source distribution

$\hat{P}_T$ Empirical Target distribution

Domain Adversarial Nets (DAN) (Ganin et al 2016)

- $\mathcal{D}_s = \{(x_i^S, y_i^S)\}_{i=1}^{N_s}$ is the labeled source data
- $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$ is the unlabeled target data
- Task on source and target is the same



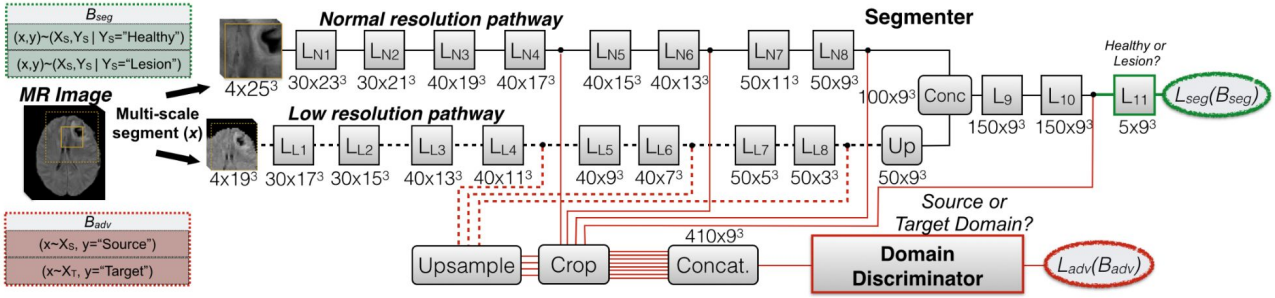
The goal is to learn domain invariant features.

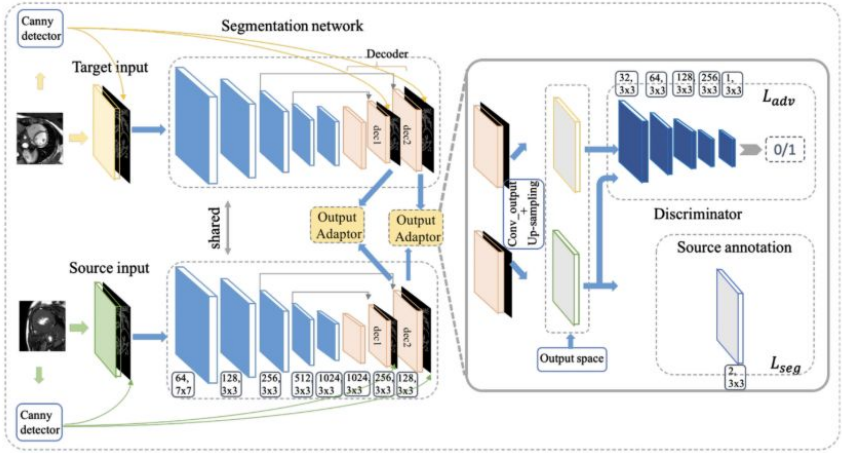
- The domain classifier (discriminator) tries to classify source and target domains
- The feature extractor tries to fool

Domain Adversarial Nets

Application on Segmer

(Kamnitsas et al 2016)





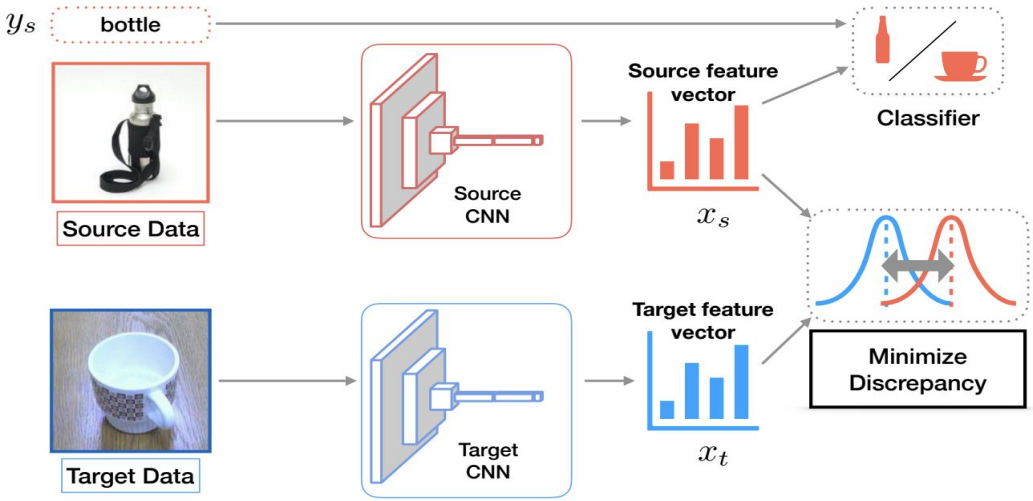
General approach to Domain Adaptation

$$\min_h R_T(h) =$$
$$\min_h R_S(h) + \mathrm{disp}_h(\hat{P}_S, \hat{P}_T)$$

simultaneous access to source and target

$\hat{P}_S$ Empirical Source distribution

$\hat{P}_T$ Empirical Target distribution



- **Privacy concerns**
- **It requires centralized training**

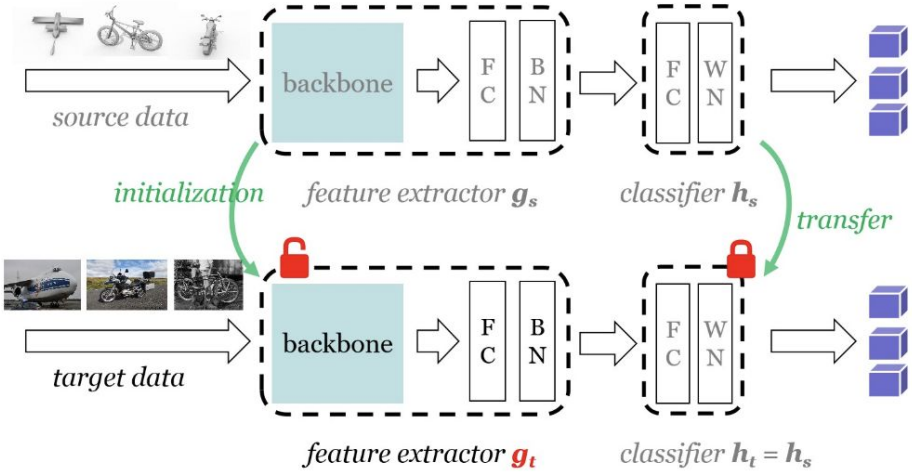
Test time adaptation for UDA

Source training:

$$f_s : \mathcal{X}_s \rightarrow \mathcal{Y}_s$$

$$\mathcal{L}_{src}(f_s; \mathcal{X}_s, \mathcal{Y}_s) = -\mathbb{E}_{(x_s, y_s) \in \mathcal{X}_s \times \mathcal{Y}_s} \sum_{k=1}^{K} q_k \log \delta_k(f_s(x_s))$$

$$f_s(x) = h_s(g_s(x))$$

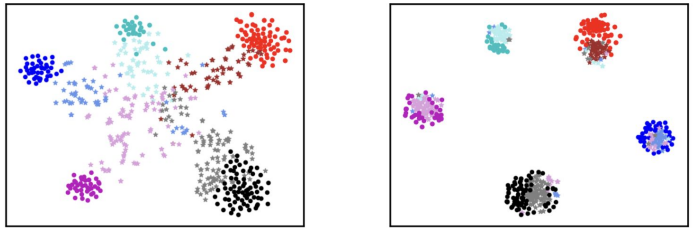


Target Adaptation

$$\mathcal{L}_{ent}(f_t; \mathcal{X}_t) = -\mathbb{E}_{x_t \in \mathcal{X}_t} \sum_{k=1}^{K} \delta_k(f_t(x_t)) \log \delta_k(f_t(x_t))$$

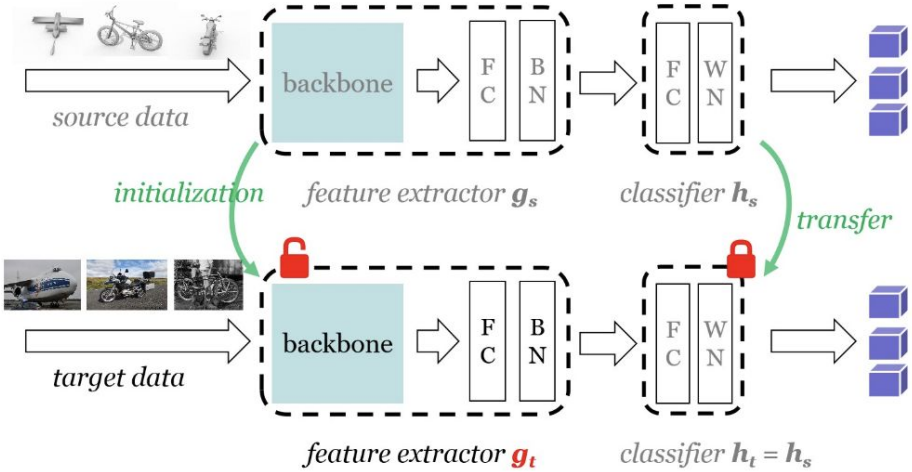
$$\mathcal{L}_{div}(f_t; \mathcal{X}_t) = \sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k$$

Test time adaptation for UDA



(a) Source model only (b) SHOT-IM

Figure 1. t-SNE visualizations for a 5-way classification task. Circles in dark colors denote unseen source data and stars in light denote target data. Different colors represent different classes.



Hypothesis Disparity Regularized Mutual Information Maximization (Lao et al, Havaei AAAI 2021)

Transfer Objective: [Model Fitness]:

The question is how do we share the knowledge among target hypotheses $\{h_i^T \in \mathcal{H}^T\}_{i=1}^M$

$$\mathrm{HD}_{h_i, h_{\text{anchor}} \in \mathcal{H}^T, i \neq \text{anchor}}(h_i, h_{\text{anchor}}) = \sum_{x \in \mathcal{X}} d(h_i(x), h_{\text{anchor}}(x))$$

Any distance measure.
Used CrossEntropy

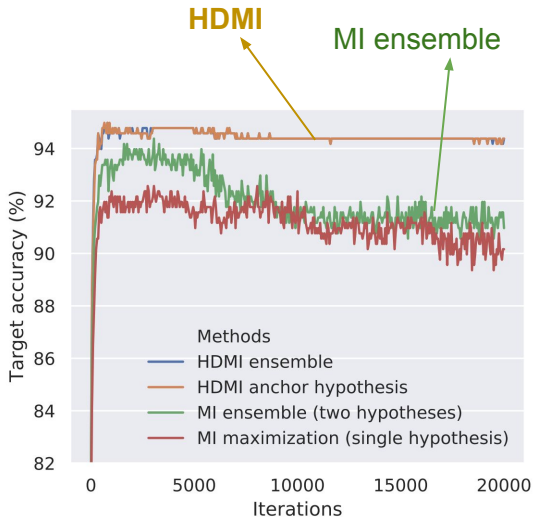


Figure 3: The hypothesis disparity regularization stabilizes the optimization for MI maximization (on A→D, Office-31).

Final target objective to minimize:

$$\underbrace{\mathbb{E}_{h \in \mathcal{H}^T}\left[-I(X^T; h(X^T))\right]}_{\text{Data Fitness term}} + \lambda \underbrace{\mathbb{E}_{h_{i,j} \in \mathcal{H}^T, i \neq j}\left[\mathrm{HD}(h_i, h_j)\right]}_{\text{Model Fitness term}}$$

Penalized Diversity HDMI (Rafiee et al, Havaei 2022)

Medical imaging dataset: LIDC

- CT scans of lungs, The task is to identify if there is any nodules in the scan. Binary classification
- Contains data from 4 manufacturers: Toshiba(T), Siemens(S), General Electric(G), Philips(P)

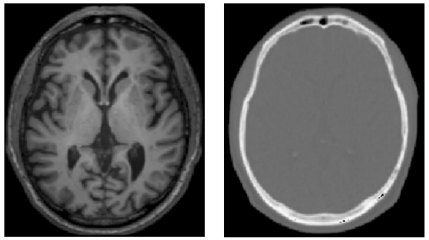
Table 1: Target accuracy (%) on Office-31 datasets (source → target).

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
DAN	78.6±0.2	80.5±0.4	63.6±0.3	97.1±0.2	62.8±0.2	99.6±0.1	80.4
DANN	79.7±0.4	82.0±0.4	68.2±0.4	96.9±0.2	67.4±0.5	99.1±0.1	82.2
SAFN+ENT	90.7	90.1	73.0	98.6	70.2	99.8	87.1
rRevGrad+CAT	90.8	94.4	72.2	98.0	70.2	**100.**	87.6
MDD	93.5±0.2	**94.5**±0.5	74.6±0.3	98.4±0.1	72.2±0.1	**100.**±0.0	88.9
TENT	94.1±0.2	90.5±0.3	73.5±0.4	98.0±0.1	74.0±0.2	99.8±0.0	88.3
SHOT	93.1±0.3	90.9±0.2	74.5±0.3	98.8±0.2	74.8±0.3	99.9±0.0	88.7
Source-DBA ensemble	84.3±0.4	78.2±0.7	63.2±0.6	96.0±0.3	64.7±0.6	98.4±0.1	80.8
Source-SEB ensemble	82.1±0.2	77.8±0.4	62.3±0.3	96.2±0.2	61.3±0.2	98.4±0.1	79.63
HDMI	94.4±0.1	94.0±0.2	73.7±0.4	**98.9**±0.1	75.9±0.3	99.8±0.0	89.5
PD	**95.6**±0.1	94.3±0.1	**75.3**±0.3	98.7±0.1	**76.4**±0.2	99.8±0.0	**90.0**

Nature of Observations: Distribution mis-match

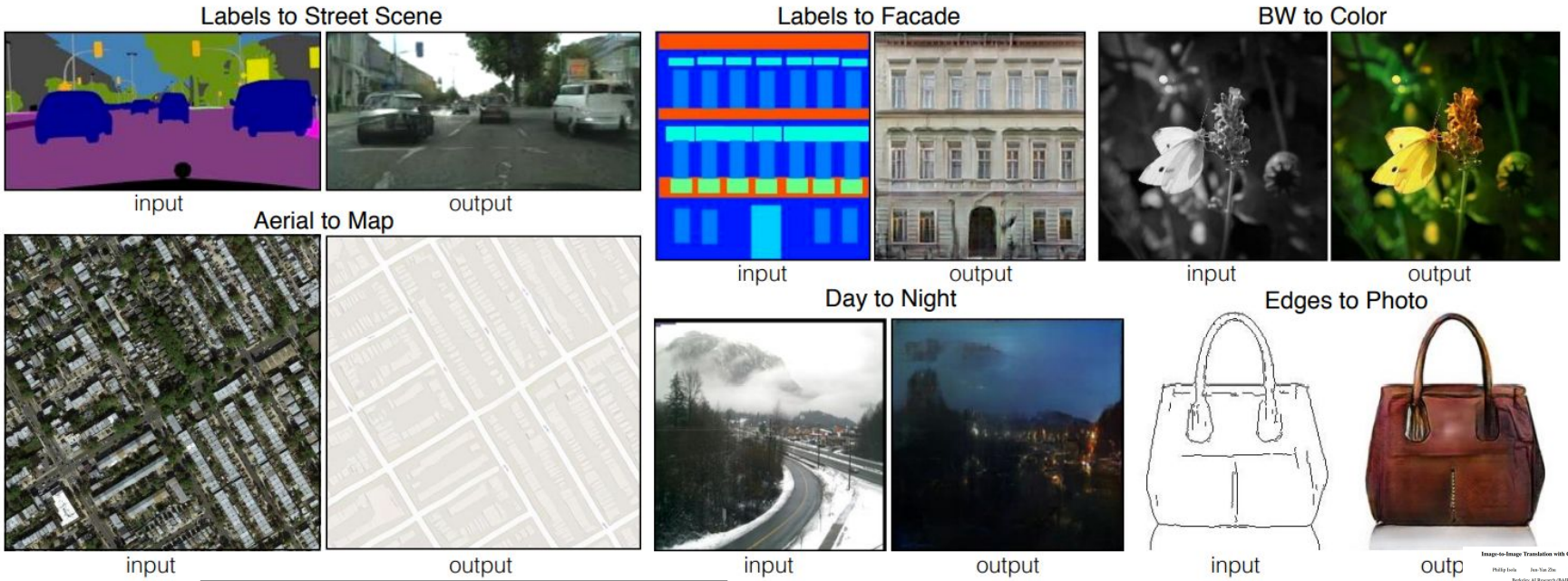
- **Domain transfer**

We have a dataset that contains data from multiple domains (modalities).



One solution: Train a domain transfer model that transfers MRI into CT.
(MRI -> CT or CT -> MRI)

Domain Translation:



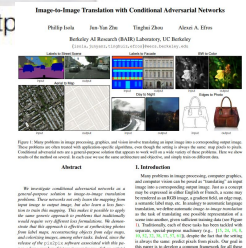


Image-to-Image Translation with Conditional Adversarial Networks

Philip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory, UC Berkeley

Figure 1: Many problems in image processing, graphics, and vision involve translating an input image into a corresponding output image. These problems are often treated with application-specific algorithms, even though the setting is always the same: map pixels to pixels. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Here we show results of the method on several. In each case we use the same architecture and objective, and simply train on different data.

Abstract

We investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. We demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. Moreover, since the release of the pix2pix software associated with this pa-

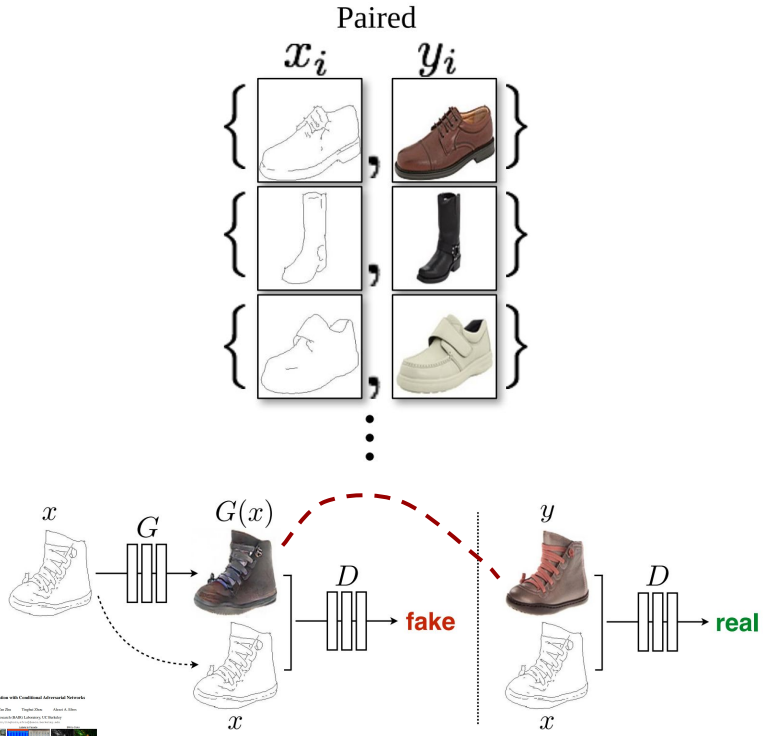
1. Introduction

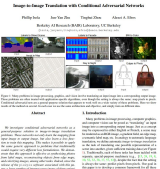
Many problems in image processing, computer graphics, and computer vision can be posed as "translating" an input image into a corresponding output image. Just as a concept may be expressed in either English or French, a scene may be rendered as an RGB image, a gradient field, an edge map, a semantic label map, etc. In analogy to automatic language translation, we define automatic image-to-image translation as the task of translating one possible representation of a scene into another, given sufficient training data (see Figure 1 for several tasks we study in this paper). Language, like images, typically requires special purpose machinery (e.g., [15, 32, 19, 9, 37, 12, 36, 7, 24, 25]), despite the fact that the setting is always the same: predict pixels from pixels. Our goal in this paper is to develop a common framework for all these

Source Domain → **Target Domain**

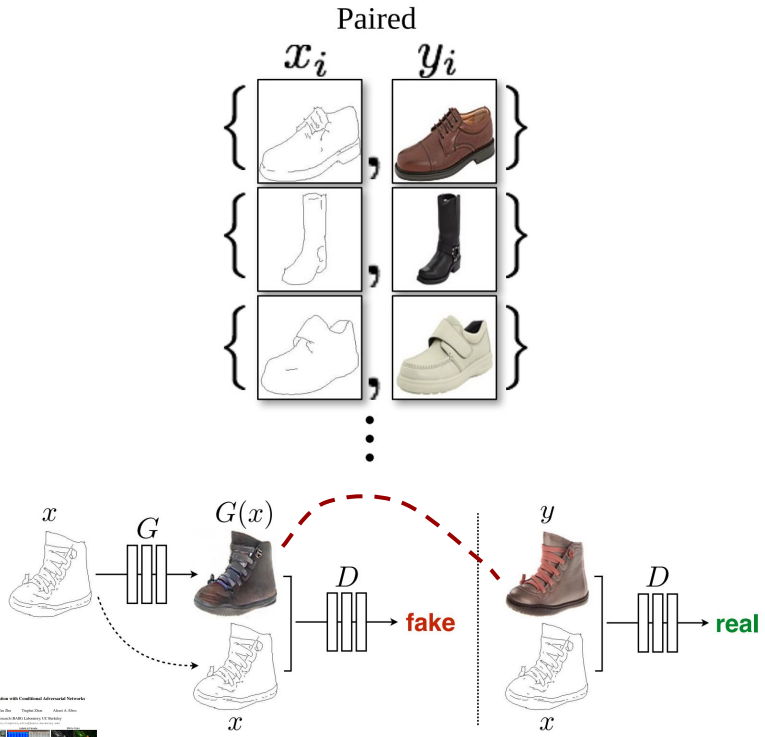
f(.)

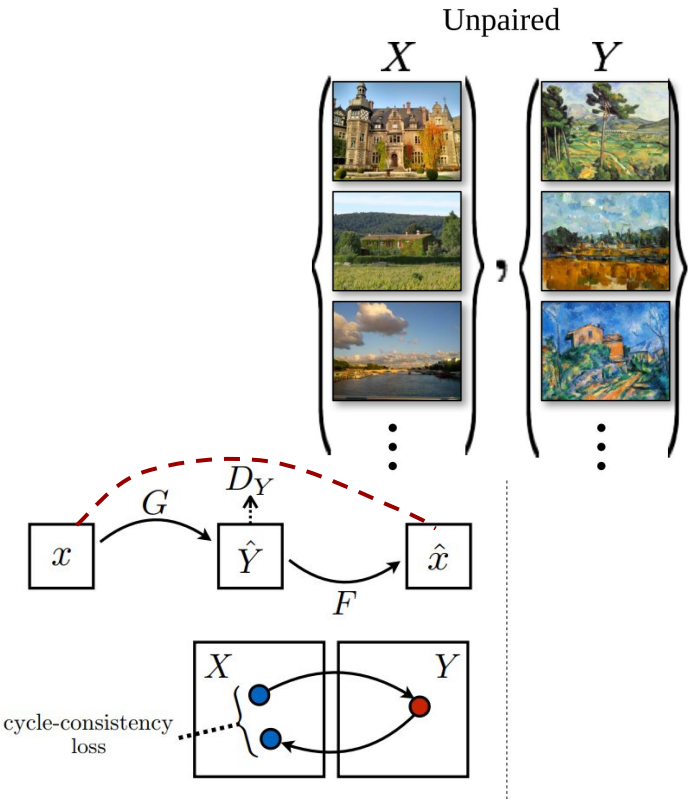
Domain Translation:

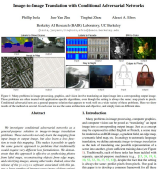




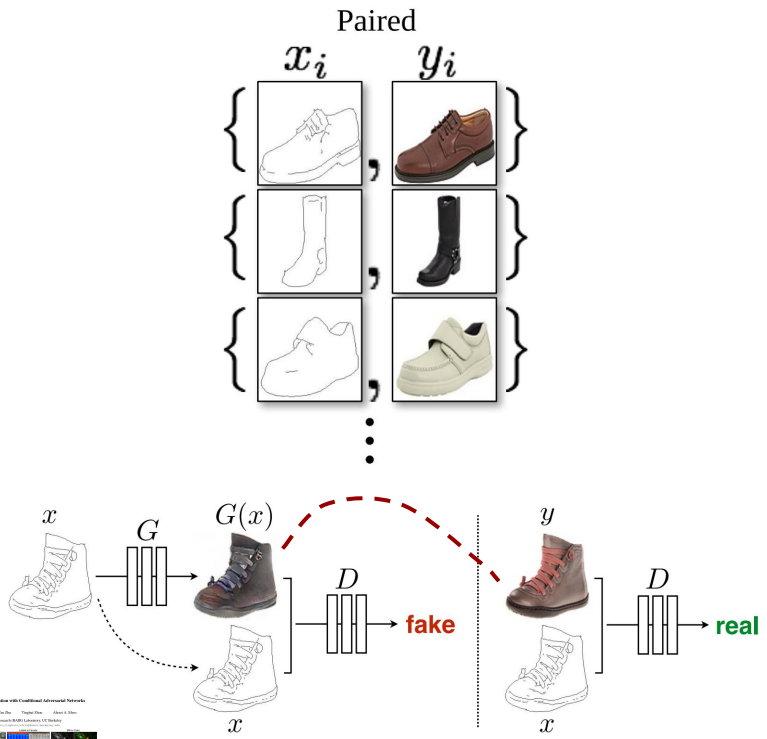
Domain Translation:

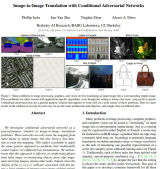


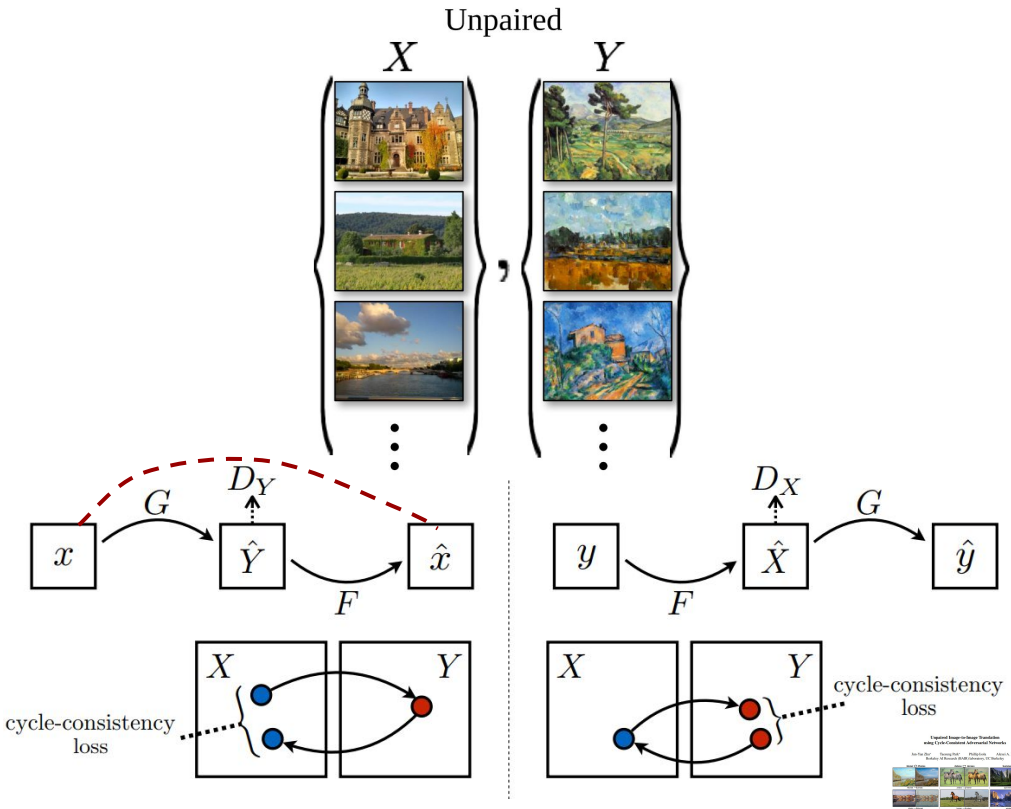




Domain Translation:



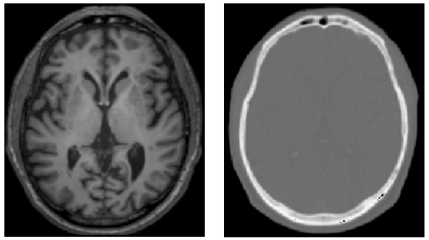




Nature of Observations: Distribution mis-match

Applications in medical imaging:

- (Nie et al 2017)



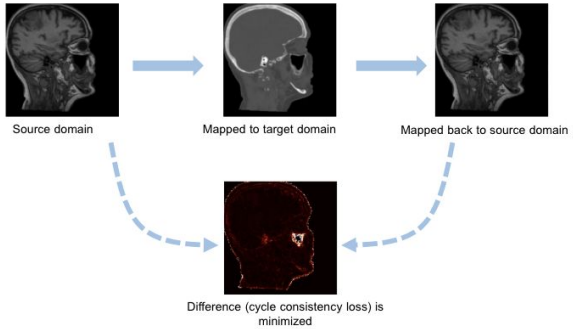


Fig. 7. Image-to-image translation via cycle consistency loss (CycleGAN) [118], [119]. The source domain MR image is mapped to the target domain CT image, and then mapped back to the source domain. The difference between the input MR image and the reconstructed MR image is minimized.

- (Wolterink et al 2017)

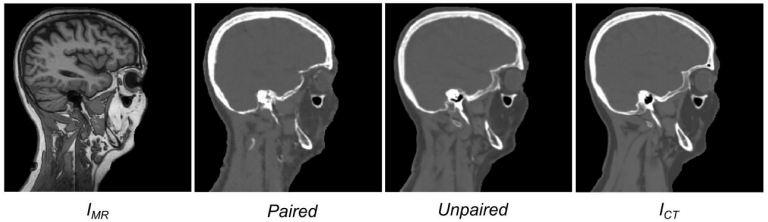
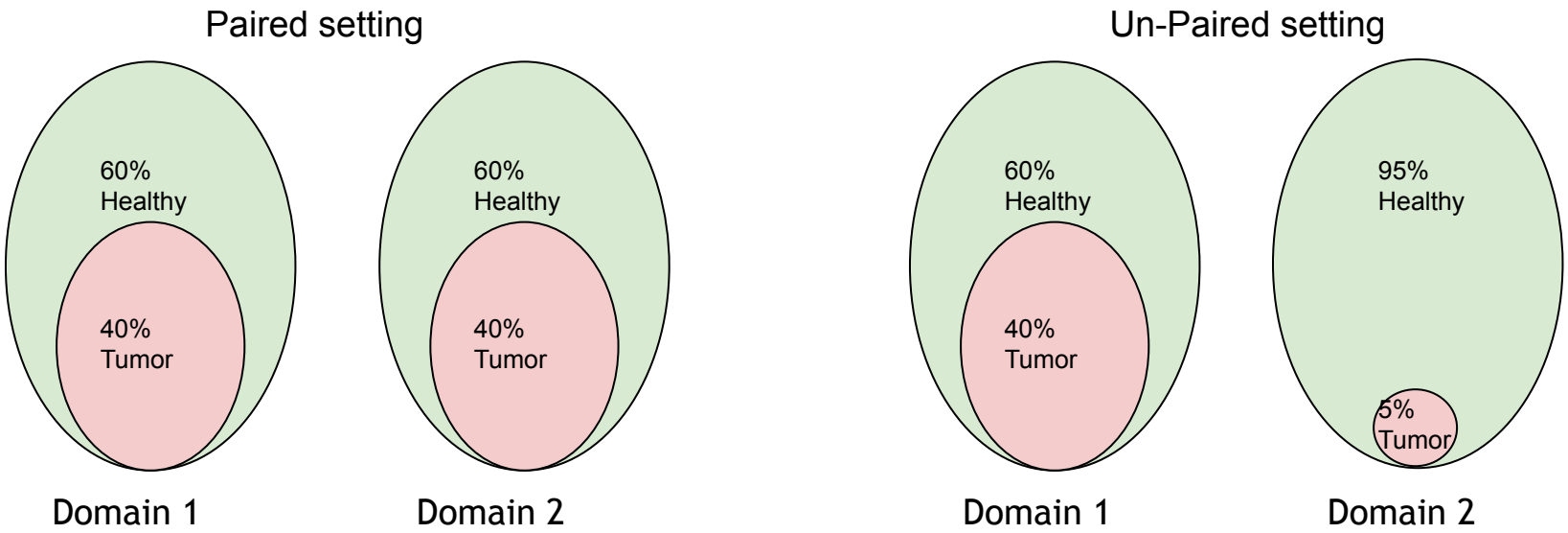
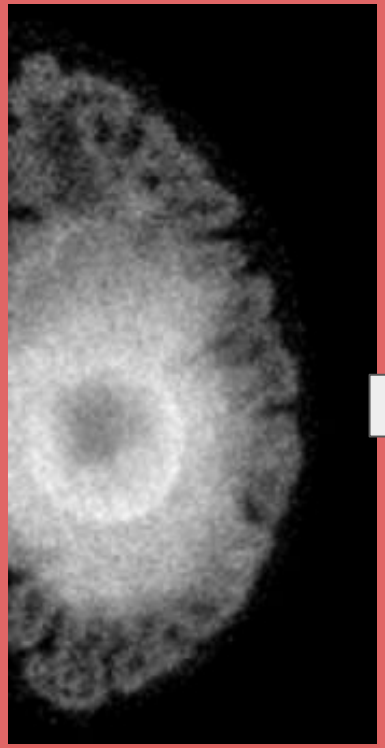


Fig. 5: *From left to right* Input MR image, synthesized CT image with paired training, synthesized CT image with unpaired training, reference real CT image.

Domain translation

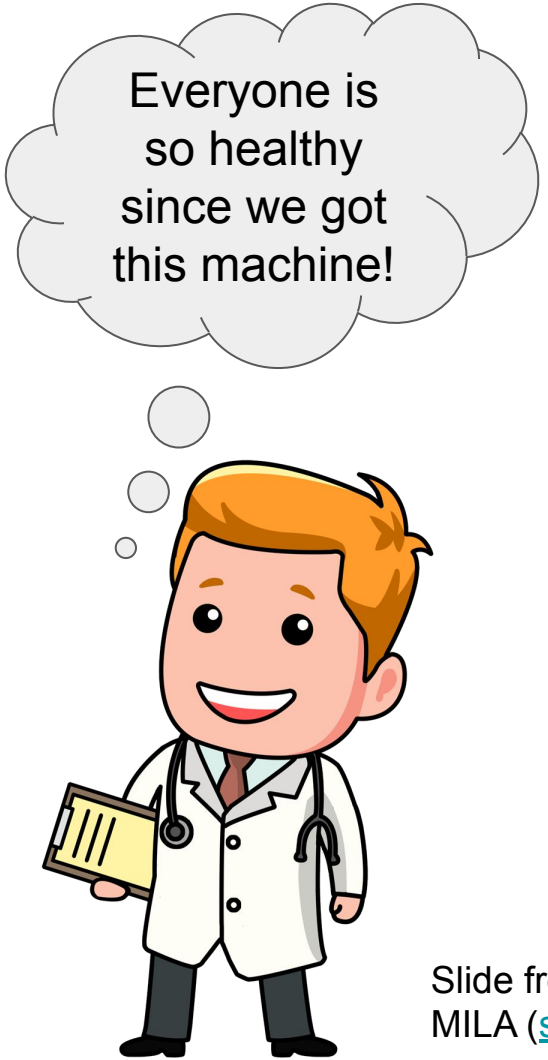
Implicit distribution matching may cause feature hallucination





Flair Real

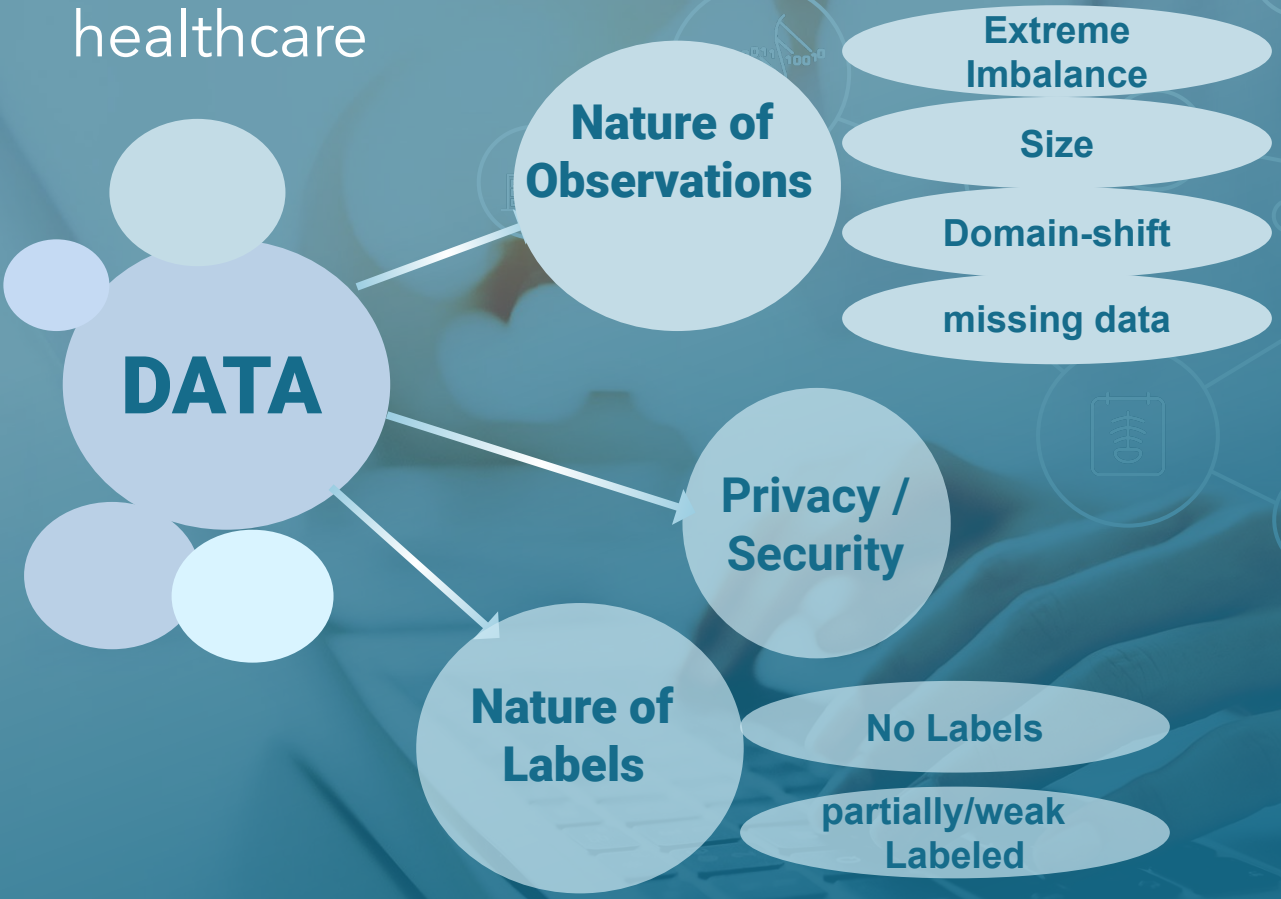
T1 Transformed



T1 Real

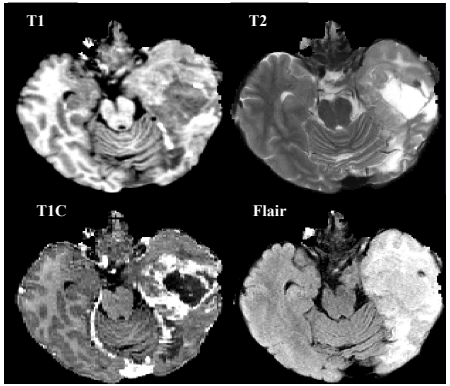
Slide from Joseph Cohen, MILA (source)

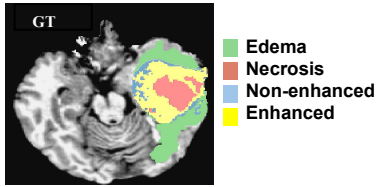
Challenges of applying AI to healthcare

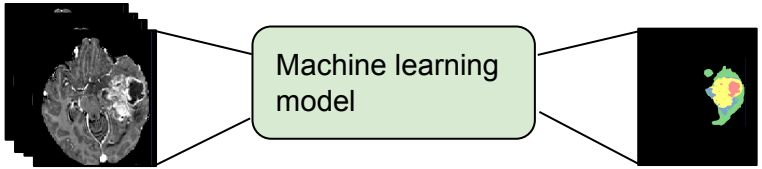


Partial Observations

- **Case study** : Brain tumor segmentation missing modalities.







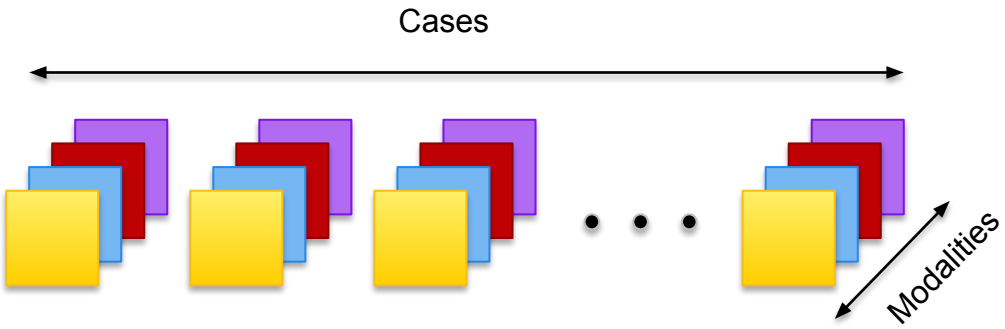
(Havaei et al 2016)

Partial Observations

- **Case study** : Brain tumor segmentation missing modalities.

Neural networks assume fixed number of input modalities at every instance

Model Assumption:

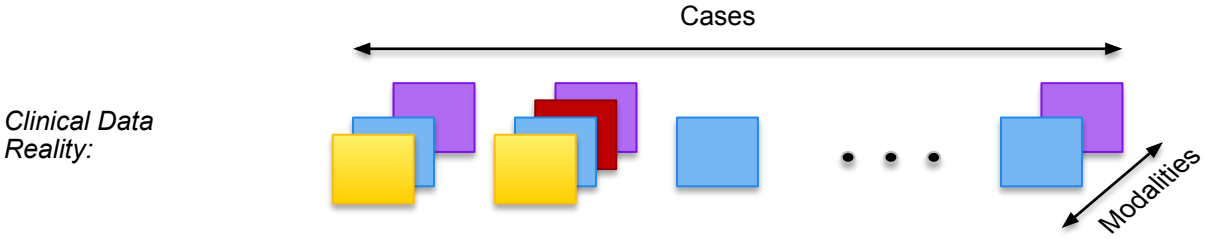


(Havaei et al 2016)

Partial Observations

- **Case study** : Brain tumor segmentation missing modalities.

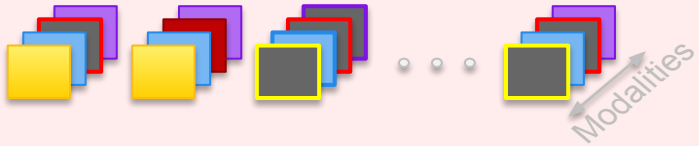
  Large clinical datasets rarely provide full modality coverage in every case



Naive solution

Least common *denominator modality:*

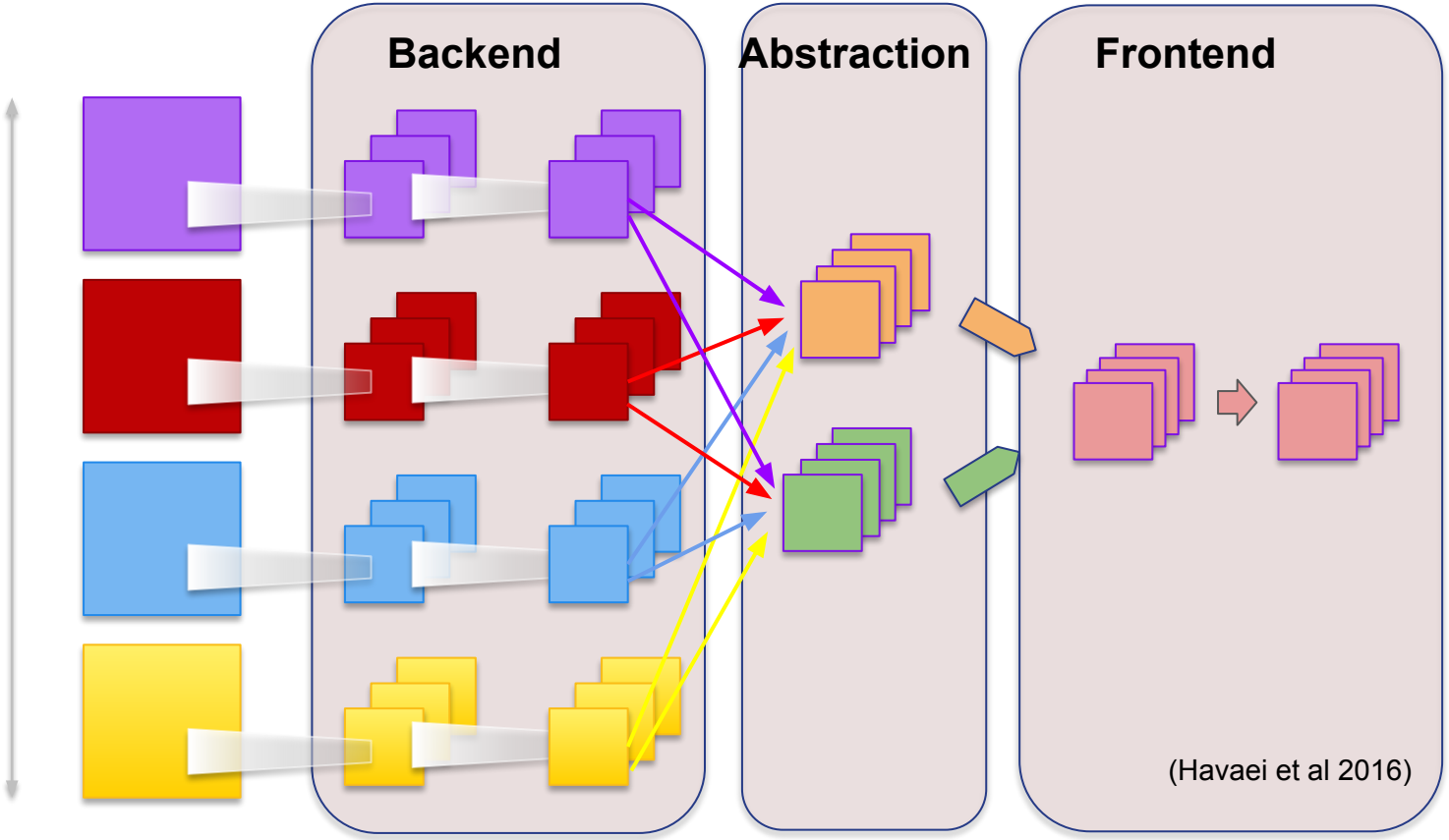
*Replace the missing ones by **zero** or **mean** Or **imputation**:*



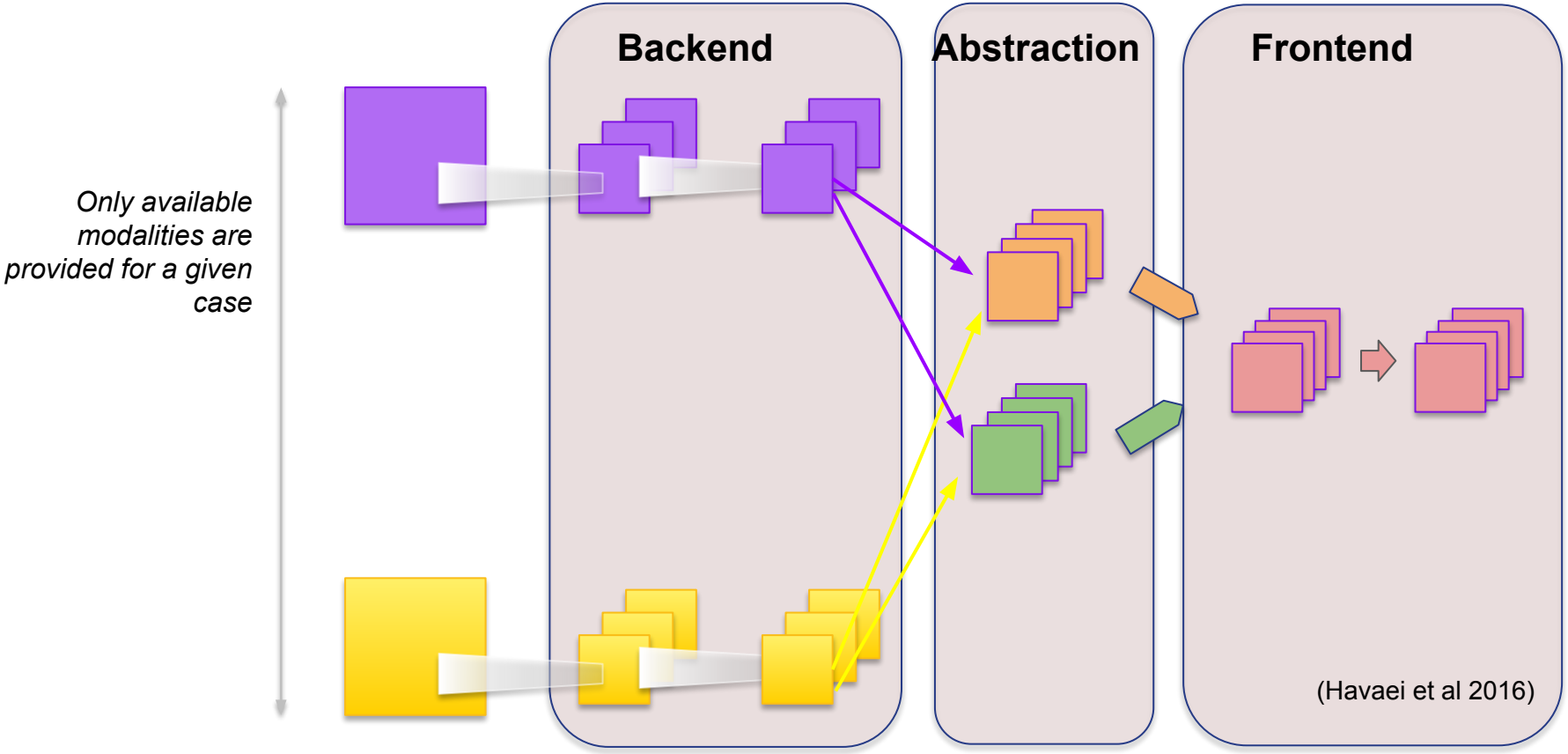
(Havaei et al 2016)

Partial Observations

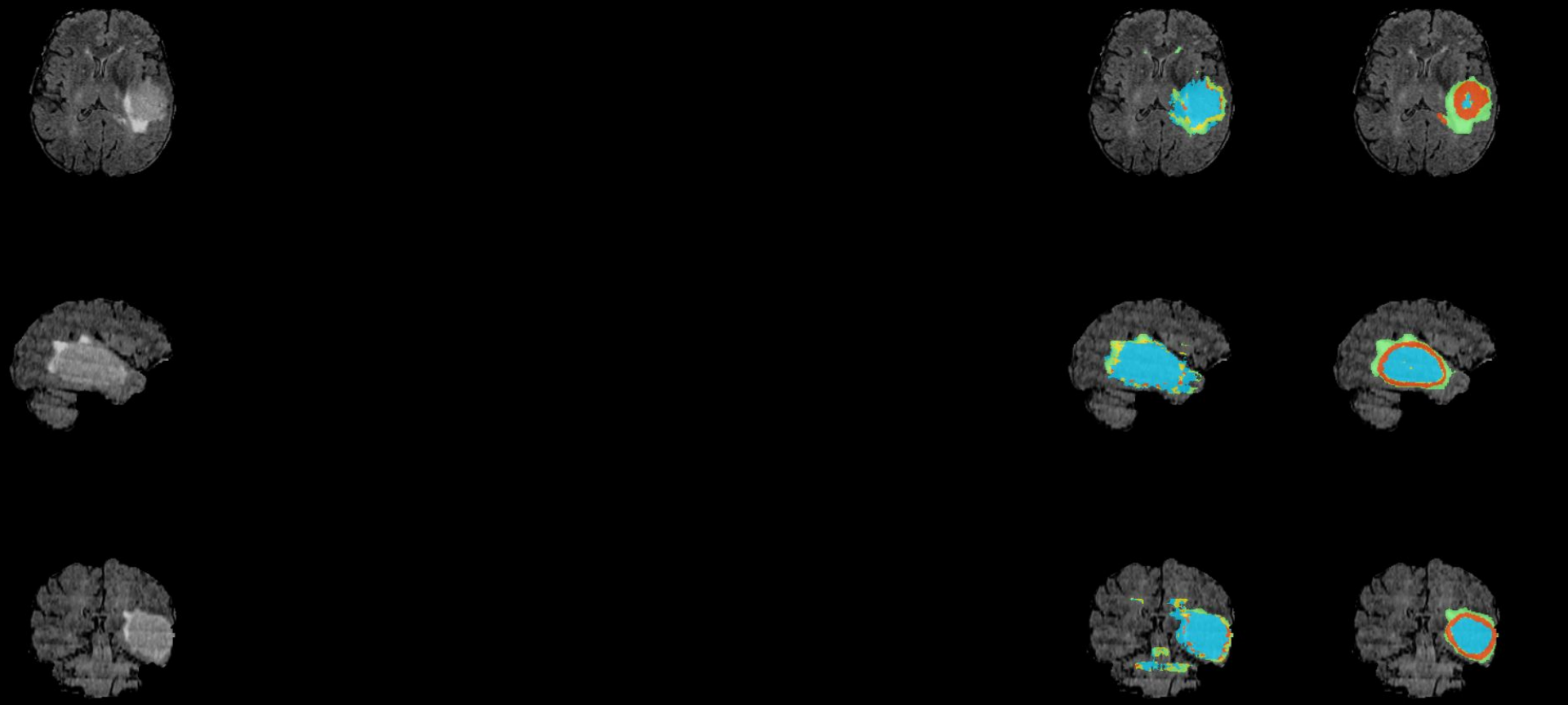
Only available modalities are provided for a given case



Partial Observations

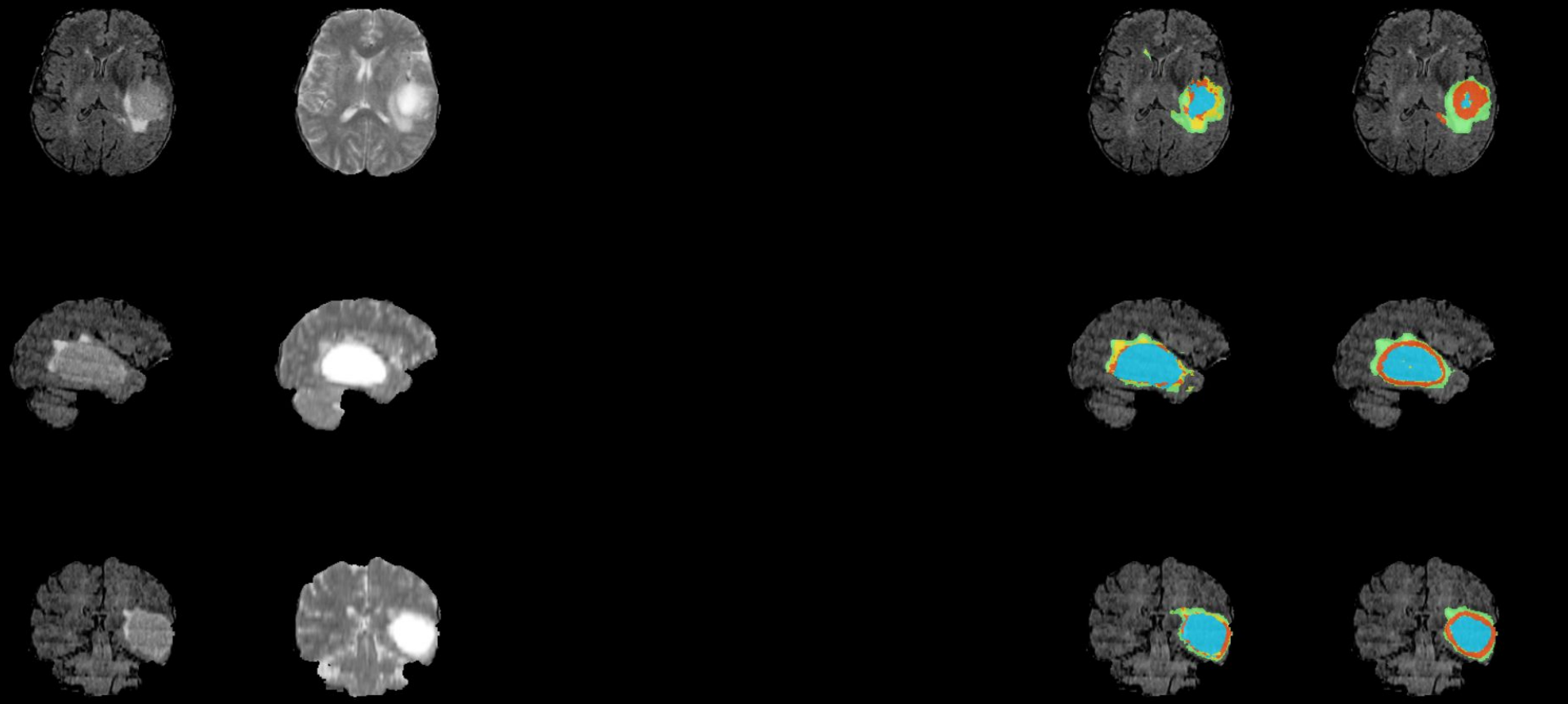


Results (BRATS)



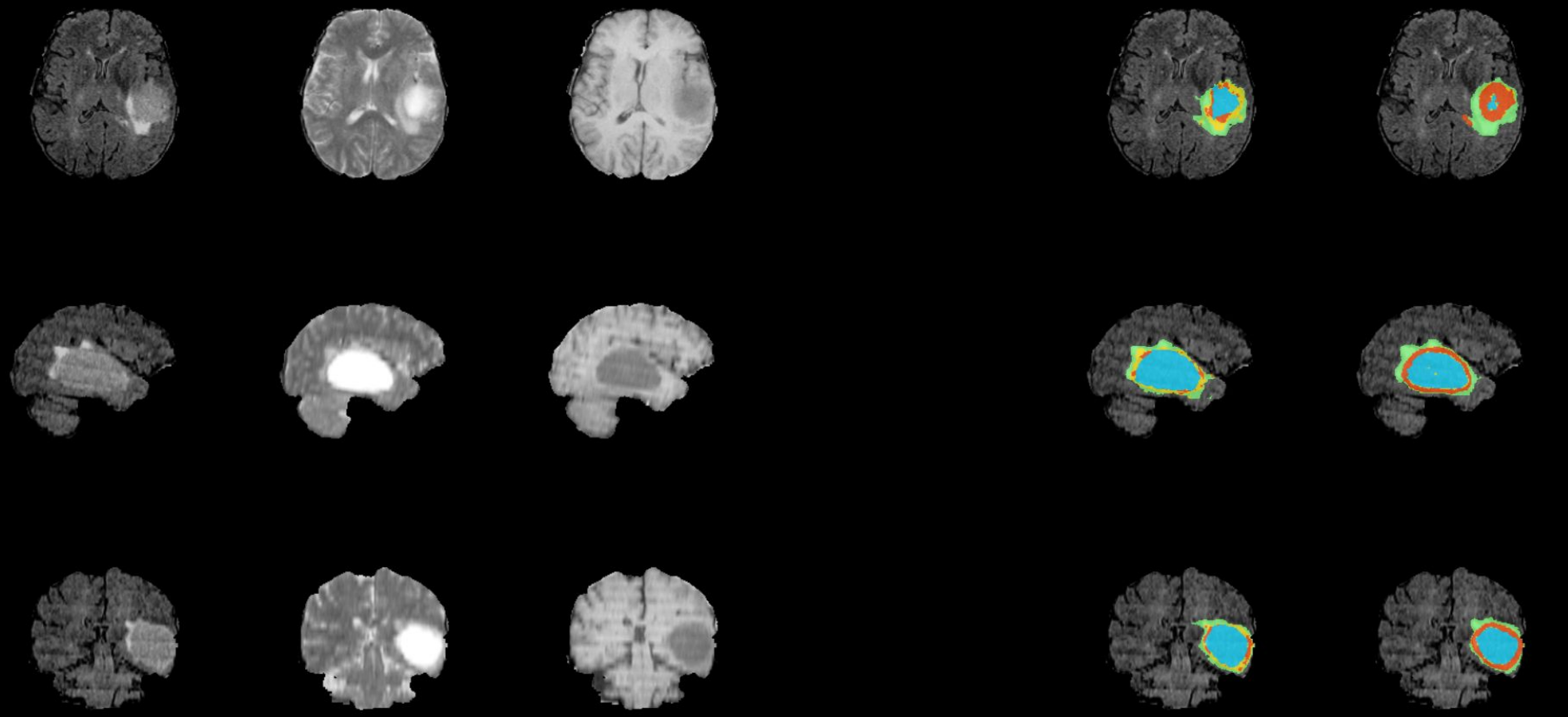
FLAIR T2W T1W T1c HeMIS Truth

Results (BRATS)



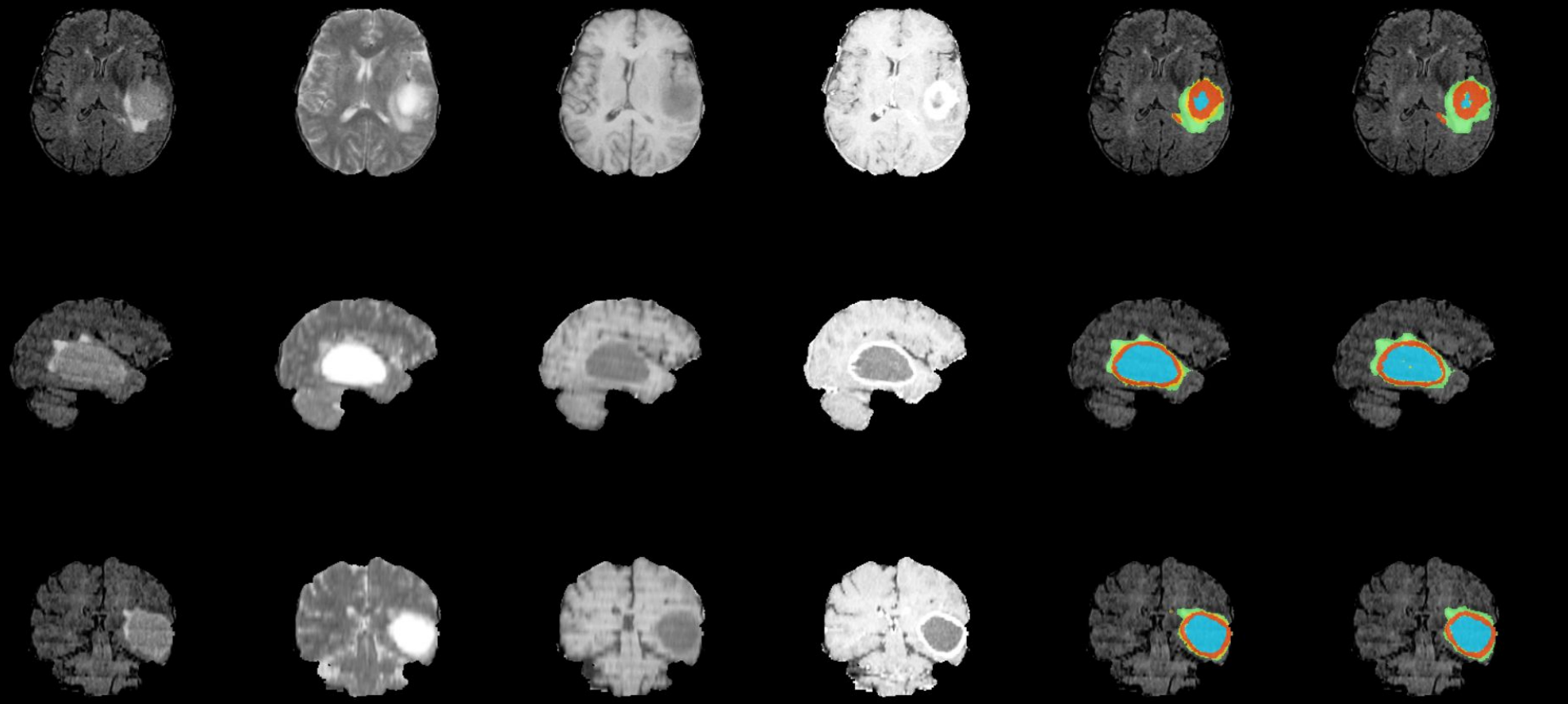
FLAIR T2W T1W T1c HeMIS Truth

Results (BRATS)



FLAIR T2W T1W T1c HeMIS Truth

Results (BRATS)

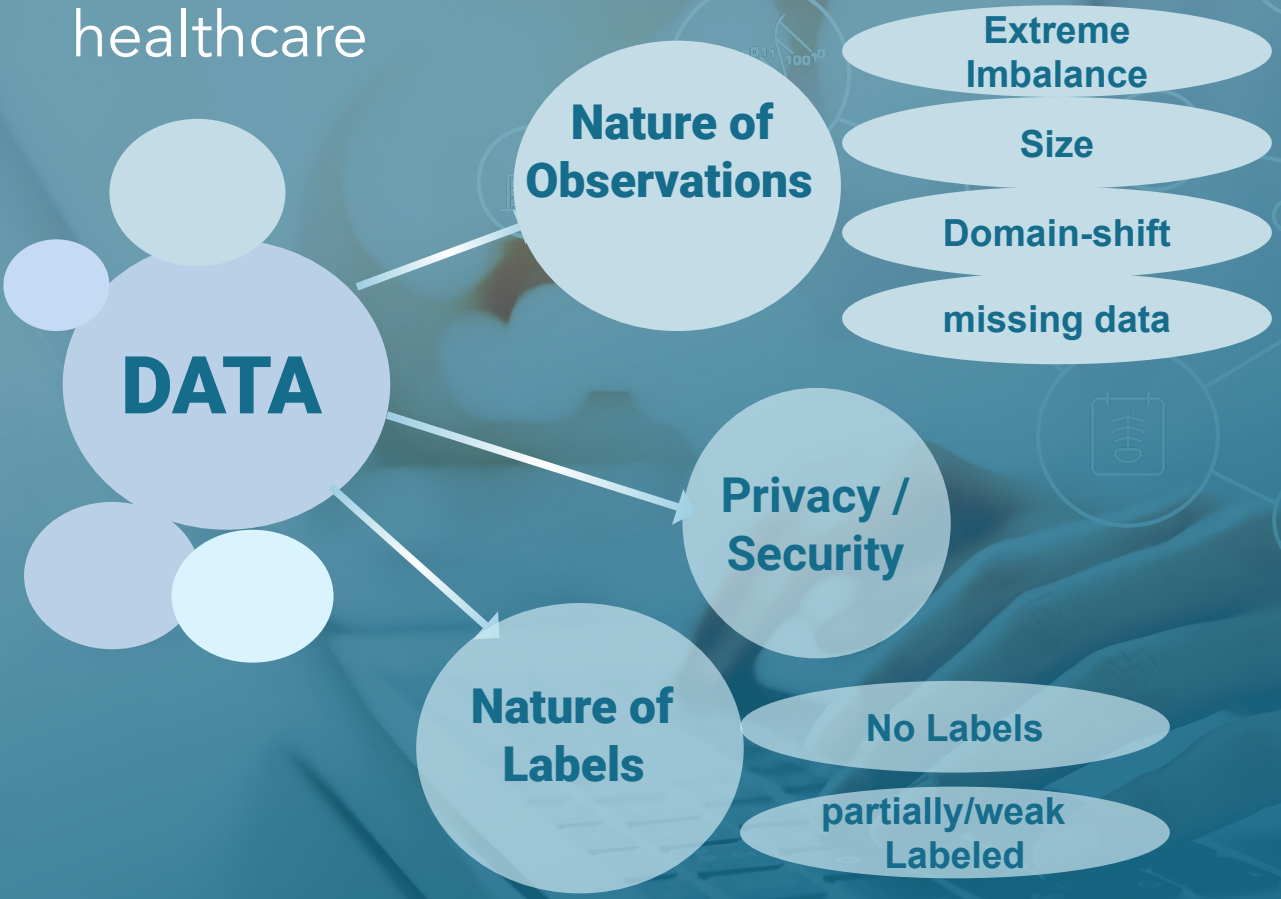


Partial Observations

Modalities				RRMS			BRATS								
				Lesion			Complete			Core			Enhancing		
F	T_1	T_1c	T_2	HeMIS	Mean	MLP	HeMIS	Mean	MLP	HeMIS	Mean	MLP	HeMIS	Mean	MLP
○	○	○	●	1.74	2.66	**12.77**	58.48	2.70	**61.50**	**40.18**	4.00	37.32	**20.31**	6.25	18.62
○	○	●	○	2.67	0.00	**3.51**	**33.46**	23.11	2.04	**44.55**	23.90	17.70	**49.93**	30.02	32.92
○	●	○	○	3.89	0.00	**6.64**	**33.22**	0.00	2.07	**17.42**	0.00	10.52	4.67	6.25	**10.78**
●	○	○	○	34.48	9.77	**38.46**	71.26	**72.30**	63.81	**37.45**	0.00	34.26	5.57	6.25	**15.90**
○	○	●	●	**27.52**	4.31	25.83	**67.59**	35.01	64.97	**63.39**	30.92	49.38	**65.38**	39.00	60.30
○	●	●	○	8.21	0.00	**8.26**	**45.93**	23.63	1.99	**55.06**	41.89	26.55	**62.40**	43.80	40.93
●	●	○	○	38.81	11.62	**39.15**	**80.28**	75.58	78.13	**49.52**	0.00	48.97	22.26	6.25	**25.18**
○	●	○	●	**31.25**	8.31	29.39	**69.56**	1.77	66.88	**47.26**	2.63	43.66	23.56	6.25	**26.37**
●	○	○	●	**39.64**	33.31	38.55	**82.1**	81.01	81.35	**53.42**	25.94	52.41	23.19	6.25	**25.01**
●	○	●	○	**41.38**	6.42	39.33	79.8	45.97	**81.13**	**66.12**	29.85	65.51	**67.12**	35.14	66.19
●	●	●	○	**41.97**	9.00	40.63	80.88	81.57	**82.19**	69.26	62.13	**69.34**	**71.30**	67.13	70.93
●	●	○	●	**46.6**	41.12	41.83	**83.87**	77.84	80.40	**57.76**	20.66	53.46	**28.46**	6.25	28.34
●	○	●	●	**41.90**	38.95	41.47	82.78	64.19	**83.37**	**70.62**	42.36	70.45	70.52	49.62	**70.56**
○	●	●	●	**34.98**	5.78	29.46	**70.98**	30.86	67.85	**66.60**	45.79	55.40	**67.84**	50.21	64.81
●	●	●	●	**48.66**	43.48	43.48	**83.15**	82.43	82.43	**72.5**	71.46	71.46	**75.37**	72.08	72.08
# Wins / 15				**9**	0	6	**10**	1	4	**14**	0	1	**9**	0	6

(Havaei et al 2016)

Challenges of applying AI to healthcare



Privacy in Medical Imaging

Ensuring the privacy of patient data in medical imaging is a critical yet complex challenge.

- **Sensitive Patient Data**:
  - Medical images often contain highly sensitive personal information, making privacy protection crucial.
- **Data Breaches**:
  - Unauthorized access to medical images can lead to data breaches, with severe consequences for patients and healthcare providers.
- **Regulatory Compliance**:
  - Adhering to strict regulations such as GDPR and HIPAA is complex and requires robust data protection measures.
- **Data Sharing Dilemmas**:
  - Sharing medical images across institutions can enhance research but increases the risk of privacy violations.
- **Anonymization Limitations**:
  - Simply removing identifiable information may not be enough to fully anonymize medical images, posing risks of re-identification.
- **Security Vulnerabilities**:
  - Medical imaging systems can be targets for cyberattacks, necessitating strong security protocols.
- **Inconsistent Privacy Standards**:
  - Varied privacy practices across institutions complicate the standardization and enforcement of privacy protections.
- **Trust Issues**:
  - Patients may be reluctant to share their medical images if they fear their privacy will not be adequately protected.

Federated learning

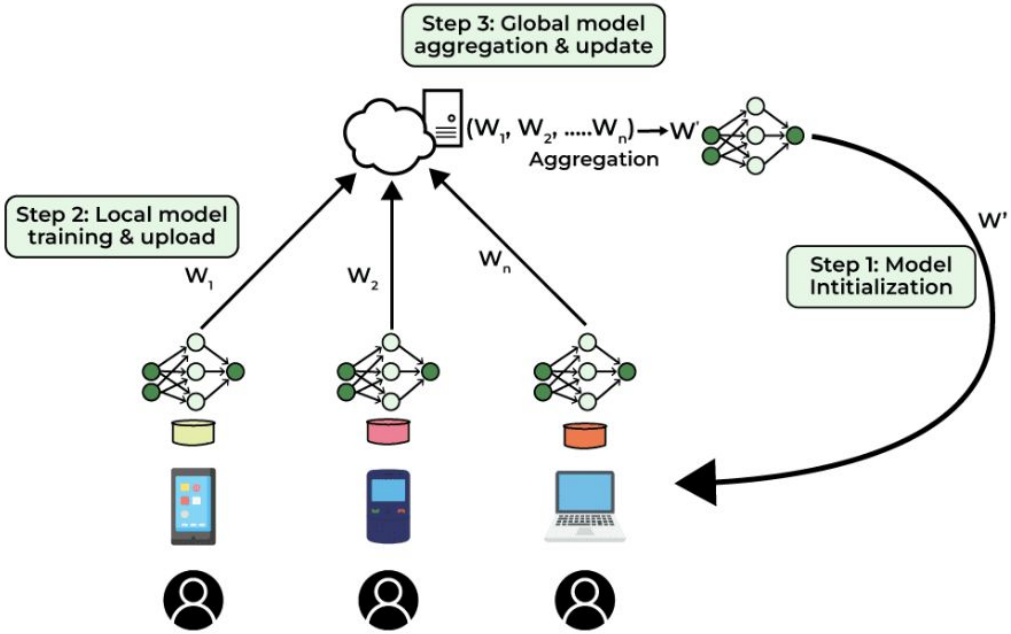
SGD:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; w)$$

FedAvg:

$$\min_{w} \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(w)$$

$$\mathcal{L}_k(w) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(x_i, y_i; w)$$



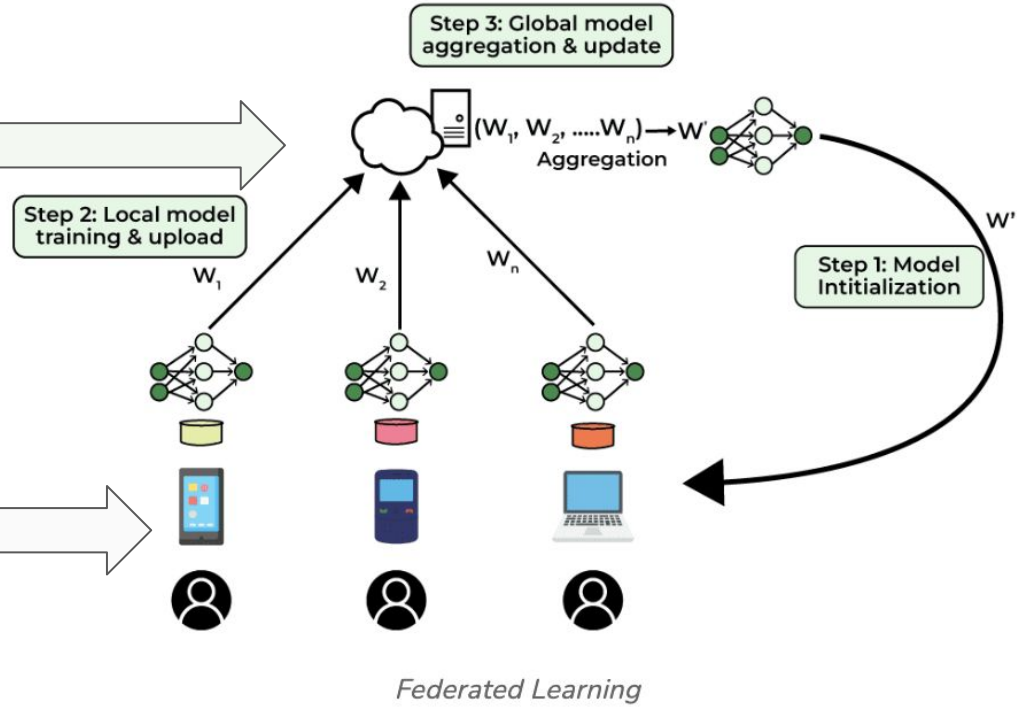
Federated Learning

Federated learning

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
initialize w_0
for each round $t = 1, 2, \ldots$ **do**
  $m \leftarrow \max(C \cdot K, 1)$
  $S_t \leftarrow$ (random set of m clients)
  for each client $k \in S_t$ **in parallel do**
    $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
  $m_t \leftarrow \sum_{k \in S_t} n_k$
  $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4]

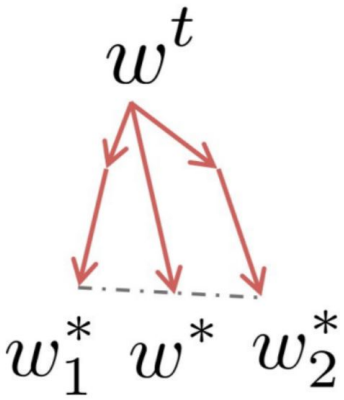
ClientUpdate(k, w): *// Run on client k*
$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch i from 1 to E **do**
  for batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w - \eta \nabla \ell(w; b)$
return w to server



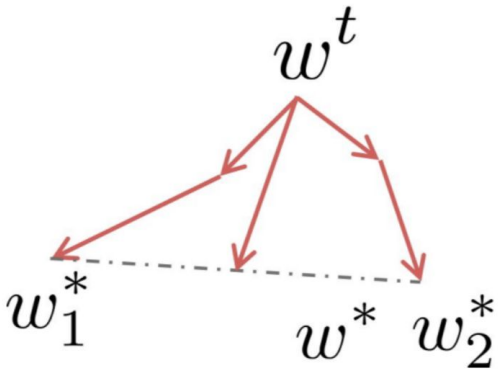
Heterogeneity in Federated learning

A lot of the research in Federated learning is to address heterogeneity of devices.

Fed averaging works well with domain homogeneous clients



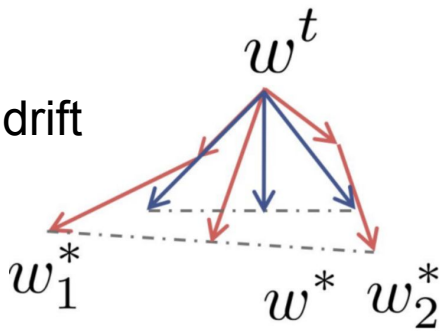
The case of heterogeneity



Heterogeneity in Federated learning

- FedProx adds a regularization on the client side to prevent drift

$$\min_{w} \mathcal{L}_k(w) + \beta \left\| w - w^t \right\|^2$$



Federated learning in medical imaging

TABLE II

CLASSIFICATION RESULTS (MEAN±STANDARD DEVIATION) OF DIFFERENT FEDERATED LEARNING SETTINGS IN TERMS OF FOUR METRICS. ADNI1-TR: ADNI-1 IS ADOPTED AS THE TRAINING SET. ADNI2-TR: ADNI-2 IS USED AS THE TRAINING SET.

Client	Method	ACC	SEN	SPE	AUC
ADNI-1	ADNI1-tr	–	–	–	–
	ADNI2-tr	0.818	0.809	0.825	0.886
	Single	0.844±0.013	0.786±0.045	0.895±0.040	0.889±0.018
	Mix	0.870±0.017	0.823±0.045	0.923±0.050	0.901±0.018
	FedAvg	0.860±0.028	0.783±0.025	0.901±0.049	0.897±0.013
	FedSGD	0.823±0.030	0.752±0.047	0.882±0.011	0.880±0.034
	FedProx	0.858±0.031	0.815±0.077	0.896±0.034	0.900±0.044
ADNI-2	ADNI1-tr	0.811	0.623	0.960	0.885
	ADNI2-tr	–	–	–	–
	Single	0.828±0.016	0.750±0.045	0.890±0.046	0.863±0.043
	Mix	0.872±0.012	0.843±0.048	0.898±0.038	0.910±0.013
	FedAvg	0.842±0.021	0.823±0.018	0.864±0.038	0.907±0.017
	FedSGD	0.844±0.039	0.819±0.066	0.871±0.056	0.908±0.040
	FedProx	0.856±0.045	0.845±0.072	0.861±0.047	0.908±0.036

Differential Privacy

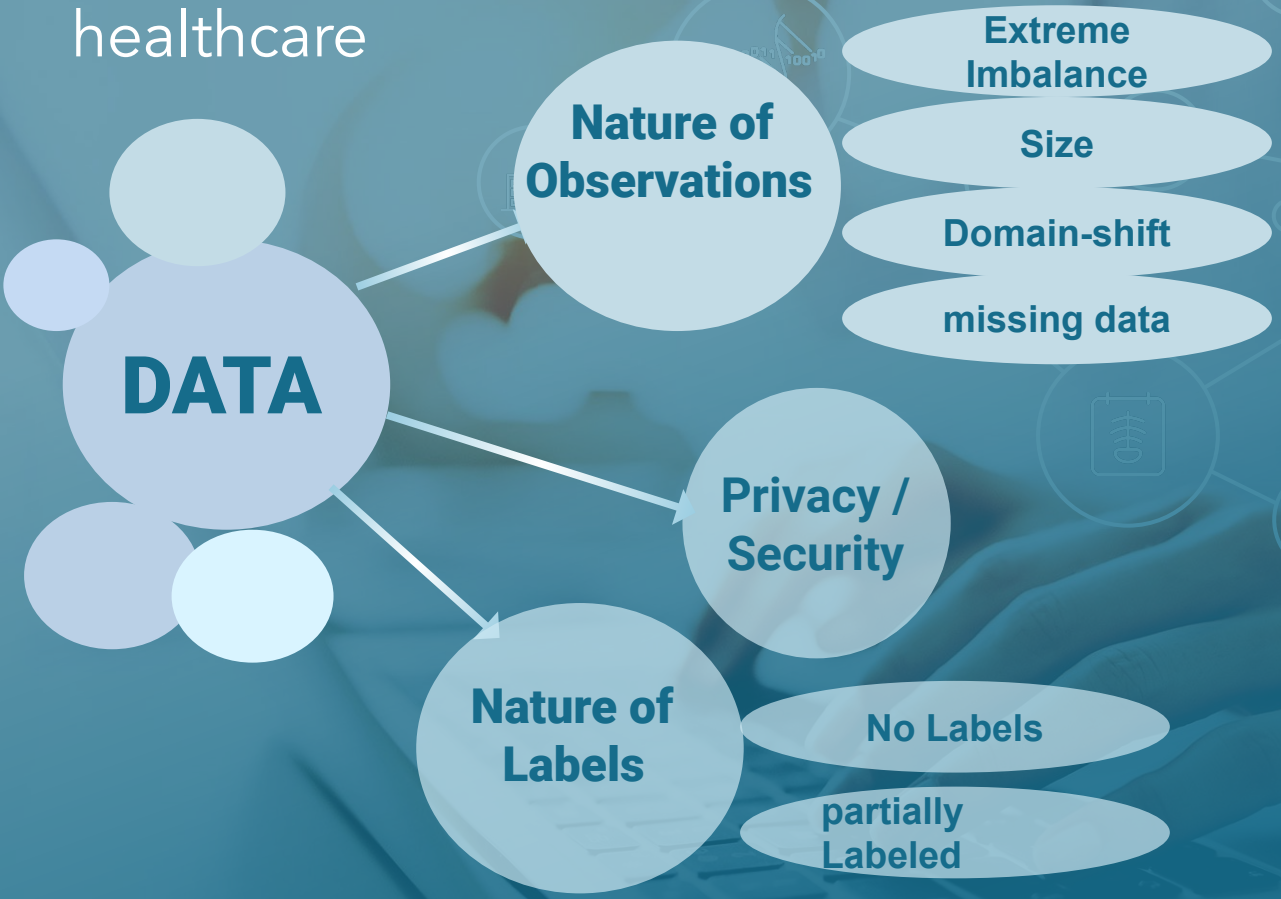
- Differential Privacy is a technique that helps protect individual data in a dataset.
- It ensures that the output of a data analysis or a machine learning model does not reveal information about any single individual in the dataset.
- It achieves this by adding a bit of randomness

Differential Privacy in Federated Learning

For each example in a training batch on each **CLIENT:**

- Compute gradient $$g(x_i) \leftarrow \nabla \mathcal{L}(w_k, x_i)$$
- Clip $$\overline{g}(x_i) \leftarrow g(x_i) / \max\left(1, \frac{\|g(x_i)\|_2}{C}\right)$$
- Add noise $$\tilde{g} \leftarrow \frac{1}{B}\left(\sum_{i=0}^{B} \overline{g}(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)\right)$$
- Update weights $$w_k \leftarrow w_k - \eta \tilde{g}$$

Challenges of applying AI to healthcare



Nature of Labels

- **No/Partial labels**
  - Huge amount of unlabeled medical data available.
  - For a given task a lot of unlabeled data and some labeled data.

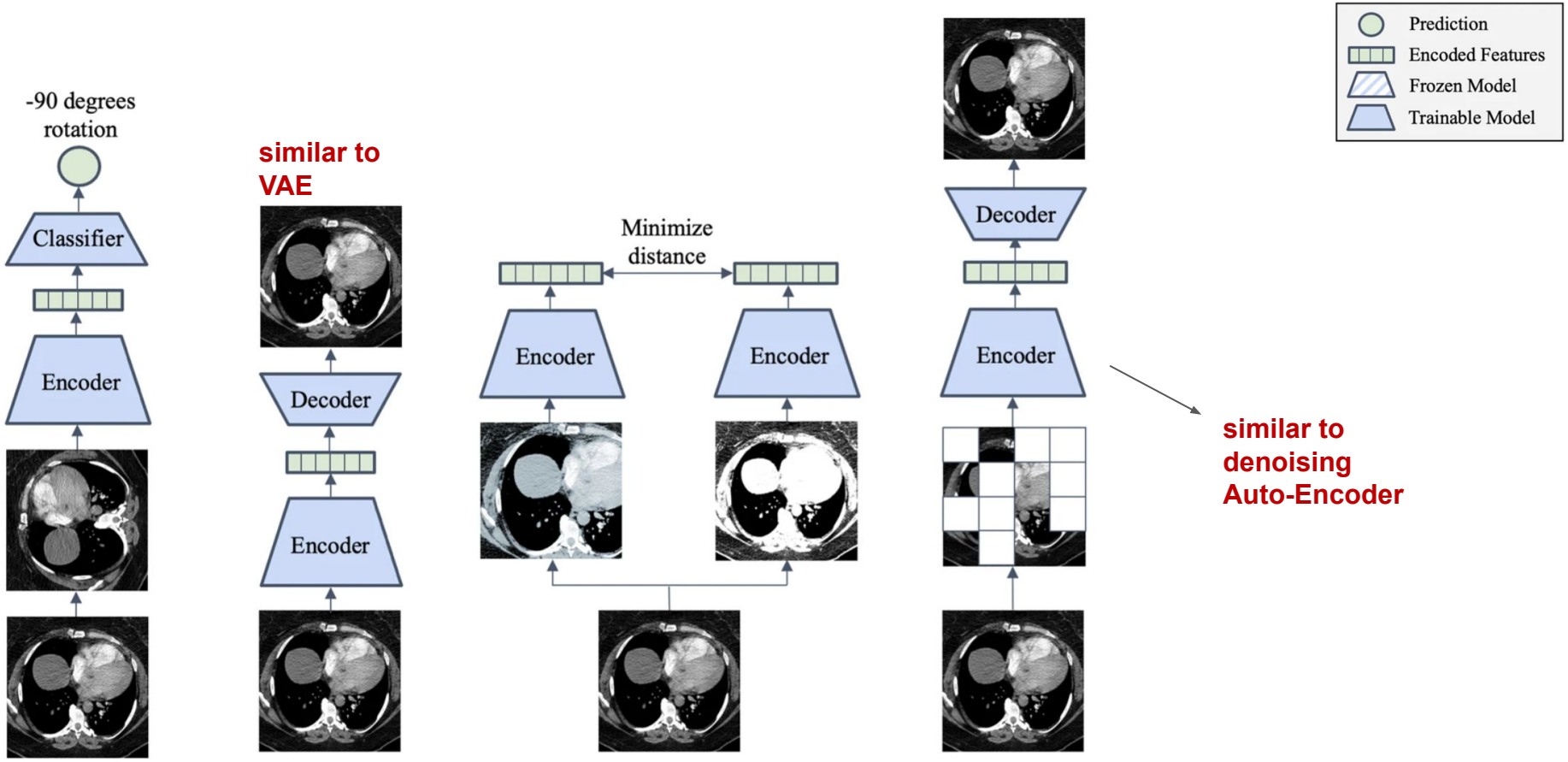
Nature of Labels: Not/Partially labeled

- **No labels** : How can we leverage unlabeled data?
  One solution: *Unsupervised / self-supervised pre-training*

Unsupervised/self supervised learning is when we learn from the input data it self!

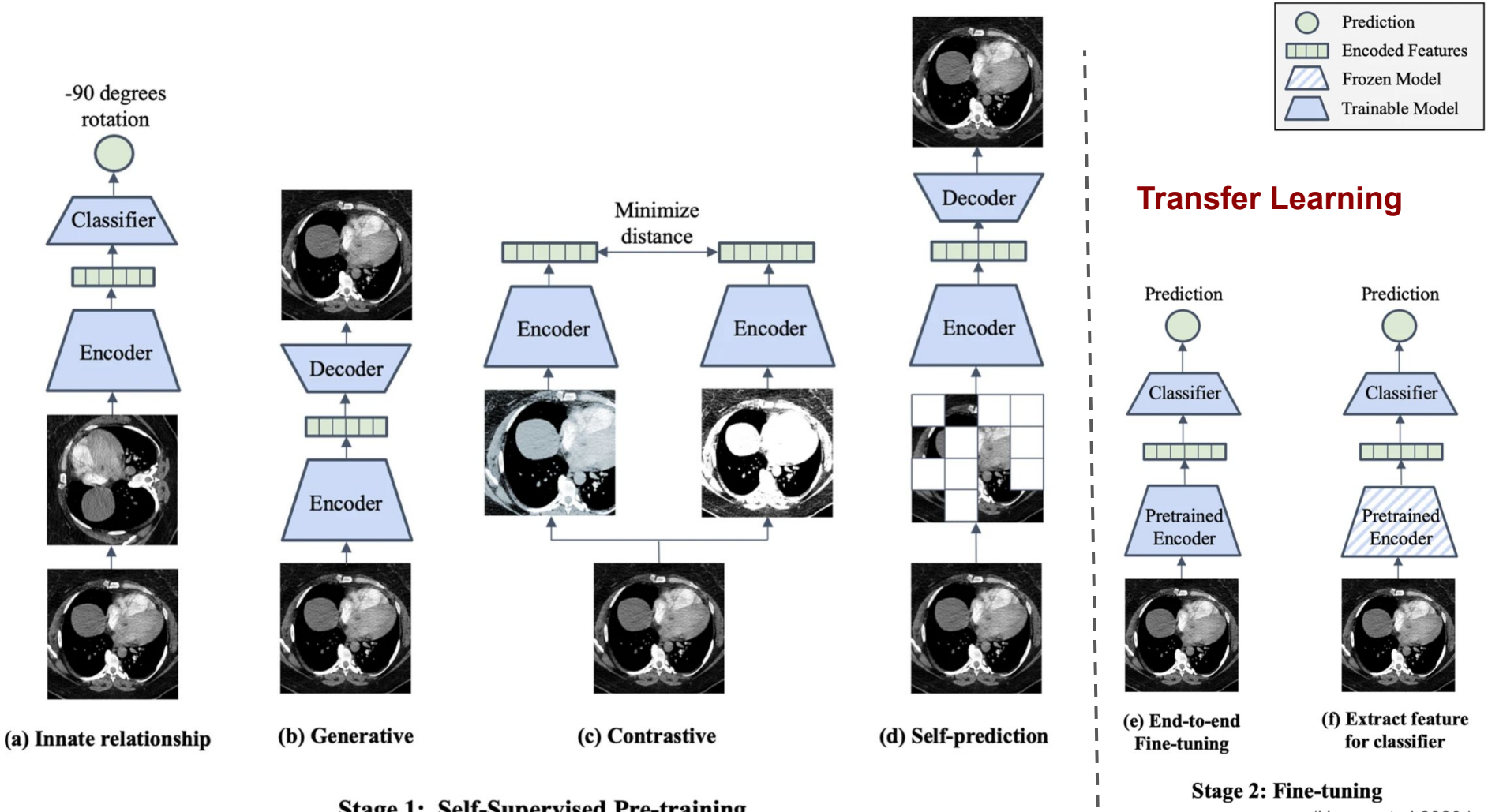
step 1: use the unlabeled data to train a “feature extractor”

step 2: collect little training data



Stage 1: Self-Supervised Pre-training

(Huang et al 2023)



(Huang et al 2023)

Self-supervised learning on medical data

Table 1: Performance of dermatology skin condition and Chest X-ray classification model measured by top-1 accuracy (%) and area under the curve (AUC) across different architectures. Each model is fine-tuned using transfer learning from pretrained model on ImageNet, only unlabeled medical data, or pretrained using medical data initialized from ImageNet pretrained model (e.g. ImageNet $\rightarrow$ Derm). Bigger models yield better performance. pretraining on ImageNet is complementary to pretraining on unlabeled medical images.

	Dermatology Classification			Chest X-ray Classifcation	
Architecture	Pretraining Dataset	Top-1 Accuracy(%)	AUC	Pretraining Dataset	Mean AUC
ResNet-50 (1×)	ImageNet	62.58 ± 0.84	0.9480 ± 0.0014	ImageNet	0.7630 ± 0.0013
	Derm	63.66 ± 0.24	0.9490 ± 0.0011	CheXpert	0.7647 ± 0.0007
	ImageNet→Derm	63.44 ± 0.13	0.9511 ± 0.0037	ImageNet→CheXpert	0.7670 ± 0.0007
ResNet-50 (4×)	ImageNet	64.62 ± 0.76	0.9545 ± 0.0007	ImageNet	0.7681 ± 0.0008
	Derm	66.93 ± 0.92	0.9576 ± 0.0015	CheXpert	0.7668 ± 0.0011
	ImageNet→Derm	67.63 ± 0.32	0.9592 ± 0.0004	ImageNet→CheXpert	0.7687 ± 0.0016
ResNet-152 (2×)	ImageNet	66.38 ± 0.03	0.9573 ± 0.0023	ImageNet	0.7671 ± 0.0008
	Derm	66.43 ± 0.62	0.9558 ± 0.0007	CheXpert	0.7683 ± 0.0009
	ImageNet→Derm	68.30 ± 0.19	0.9620 ± 0.0007	ImageNet→CheXpert	0.7689 ± 0.0010

(Azizi et al 2021)

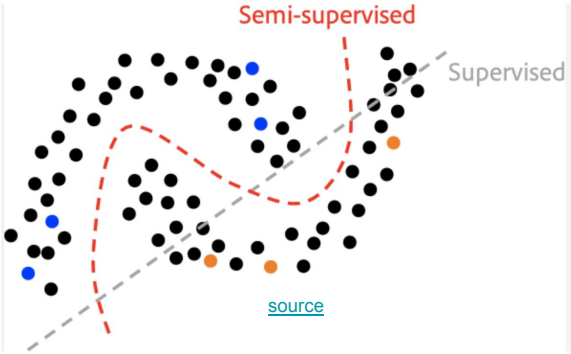
Partially labeled dataset

1) Semi-supervised learning

2) Active learning : identify most informative unlabeled data points and label them

(Azizi et al 2021)

Semi-Supervised Learning



source

Assumptions:

- **Cluster assumption :** When two samples $x1$ and $x2$ are similar or belong to the same cluster, their corresponding outputs outputs $y1$ and $y2$ should also be similar or belong to the same category, and vice versa. This assumption implies that the samples in a single class tend to form a cluster.

- **Low-density Separation:** Decision boundary should fall in a low density regions in data space

- **The Manifold Assumption.** If two samples $x1$ and $x2$ reside in a local neighborhood within a low-dimensional manifold, they are likely to possess similar class labels.

See tomorrow's lecture for more details and examples

Active learning Example: MedAL

Algorithm :

1. Train model on training data

$$f = \arg\min_w \mathcal{L}(w; D^{\text{labeled}})$$

$$f(x) = h(g(x))$$

2. apply the trained model on unlabeled data and select those with high entropy according to a threshold θ

$$D^{\text{high_entropy}} = \{x \in D^{\text{unlabeled}} \mid H(f(x)) \geq \theta\}$$

3. select examples that are different from training data

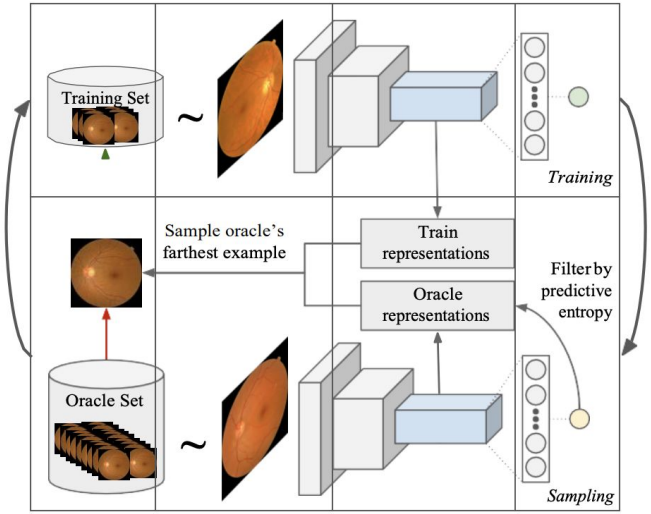
$$D^{\text{selected}} = \left\{x \in D^{\text{high_entropy}} \mid \min_{x' \in D^{\text{labeled}}} d(g(x), g(x')) \geq \delta\right\}$$

4. Label those examples

$$D^{\text{labeled_selected}} = \{(x, \text{label}(x)) \mid x \in D^{\text{selected}}\}$$

5. Augment training data

$$D^{\text{labeled}} \leftarrow D^{\text{labeled}} \cup D^{\text{labeled_selected}}$$



Active learning Example: MedAL

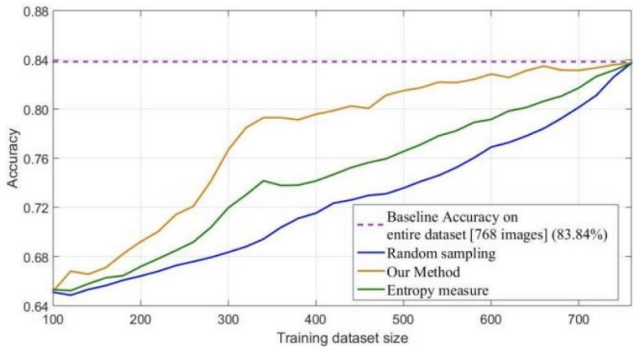


Fig. 5. **Messidor Results.** Test set accuracy as a function of the number of labeled training images. We can see that our method achieves 80% accuracy with 200 less labeled images in the training set than uncertainty sampling. Using our method, it is possible to obtain comparable results to a model trained on the full dataset of 768 images using only 650 labeled images.

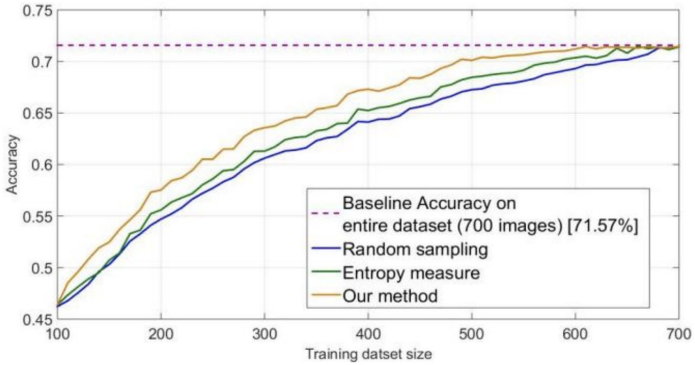
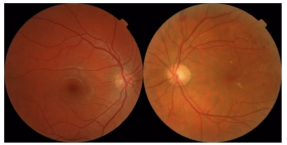
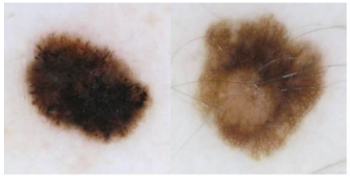


Fig. 7. **Skin Cancer Results.** Our method also works on an unbalanced dataset. Using our sampling technique, we achieve the baseline results using 610 labeled images, 90 images less than the full dataset.



No DR DR

Messidor Dataset contains 1200 eye fundus images from 654 diabetic and 546 healthy patients. This dataset was labeled for Diabetic Retinopathy (DR) grading and for Risk of Macular Edema. In this work, we are using Messidor to classify eye fundus images as healthy (DR grade = 0) or as having DR (DR grade > 0). We used 768 images as the combined oracle and training sets, 192 for testing and the remaining 240 as a validation set where we tuned our parameters.



Benign Malignant

Skin Cancer dataset contains 900 benign and malignant cell tissue images [20]. We used 700 images as the combined training and oracle sets and 200 images for testing. The class distribution of this dataset was highly biased towards the negative (benign) class (80%) vs positive (malignant) class (20%). To balance the dataset, in each iteration of the AL algorithm, we augmented the positive class samples of the training set using techniques such as cropping, flipping, translating, etc.

References

- Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.
- Hwang, Sangheum, and Hyo-Eun Kim. "Self-transfer learning for fully weakly supervised object localization." *arXiv preprint arXiv:1602.01625* (2016).
- Kohl, Simon AA, et al. "A Probabilistic U-Net for Segmentation of Ambiguous Images." *arXiv preprint arXiv:1806.05034*(2018).
- Kallenberg, Michiel, et al. "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring." IEEE transactions on medical imaging 35.5 (2016): 16322-1331.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
- Jesson, Andrew, et al. "Cased: Curriculum adaptive sampling for extreme data imbalance." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.
- Havaei, Mohammad, et al. "Brain tumor segmentation with deep neural networks." *Medical image analysis* 35 (2017): 18-31.
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Dong, J. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122-1131.
- Nie, Dong, et al. "Medical image synthesis with context-aware generative adversarial networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.
- Wolterink, Jelmer M., et al. "Deep MR to CT synthesis using unpaired data." *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, Cham, 2017.
- Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint*(2017).
- Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *arXiv preprint* (2017).
- Drozdzal, Michal, et al. "Learning normalized inputs for iterative estimation in medical image segmentation." Medical image analysis 44 (2018): 1-13.
- Farahani, Farzad Vasheghani, Abbas Ahmadi, and MH Fazel Zarandi. "Lung nodule diagnosis from CT images based on ensemble learning." *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*. IEEE, 2015
- Havaei, Mohammad, et al. "HeMIS: Hetero-modal image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2016.
- Loghmani, Mohammad Reza, Barbara Caputo, and Markus Vincze. "Recognizing Objects In-the-wild: Where Do We Stand?." 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018.
- Grm, Klemen, et al. "Strengths and weaknesses of deep learning models for face recognition against image degradations." IET Biometrics 7.1 (2017): 81-89.
- He, Kaiming, et al. "Mask r-cnn." *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017.
- Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).
- Drozdzal, Michal, et al. "The importance of skip connections in biomedical image segmentation." *Deep Learning and Data Labeling for Medical Applications*. Springer, Cham, 2016. 179-187.

References

- Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S. and Shao, L., 2019. Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2079-2088).
- Jiang, X., Ding, L., Havaei, M., Jesson, A., Matwin, T., "Task Adaptive Metric Space for Medium-Shot Medical Image Classification" International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 [To appear on arxiv soon]
- Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).
- BenTaieb, A. and Hamarneh, G., 2017. Uncertainty driven multi-loss fully convolutional networks for histopathology. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (pp. 155-163). Springer, Cham.
- Guan, H. and Liu, M., 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, *69*(3), pp.1173-1185.
- Li, H., Loehr, T., Sekuboyina, A., Zhang, J., Wiestler, B. and Menze, B., 2020. Domain adaptive medical image segmentation via adversarial learning of disease-specific spatial patterns. arXiv preprint arXiv:2001.09313.
- Dorent, R., Joutard, S., Modat, M., Ourselin, S. and Vercauteren, T., 2019. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22 (pp. 74-82). Springer International Publishing.
- Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S. and Chaudhari, A.S., 2023. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine, 6(1), p.74.
- Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T. and Natarajan, V., 2021. Big self-supervised models advance medical image classification. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3478-3488).

Thank you

Mohammad Havaei
mhavaei@google.com