Deep Learning with Data-Efficiency

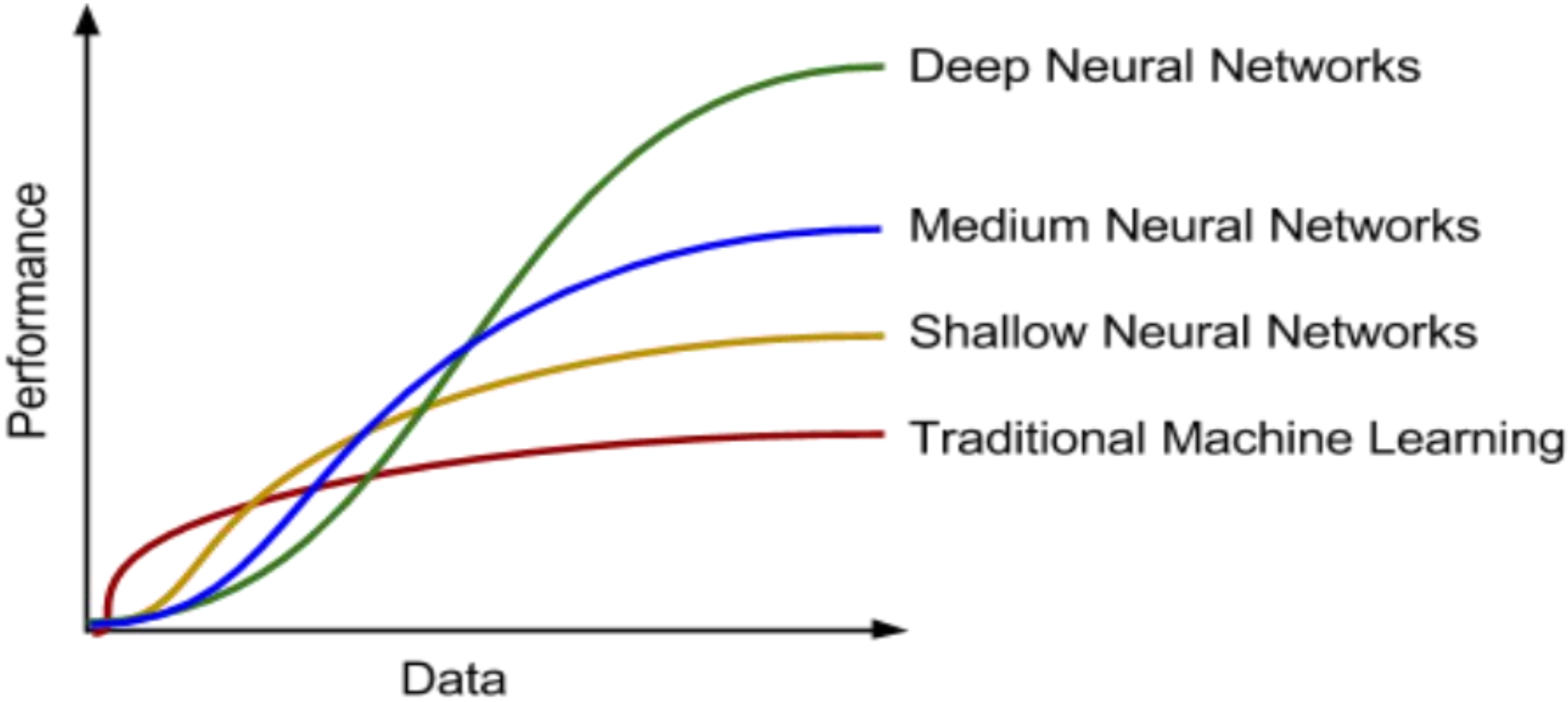
Prof. Yike Guo FREng

Director , Data Science Institute
Imperial College London

Distinguished Visiting Professor
Tsinghua University

Imperial College London

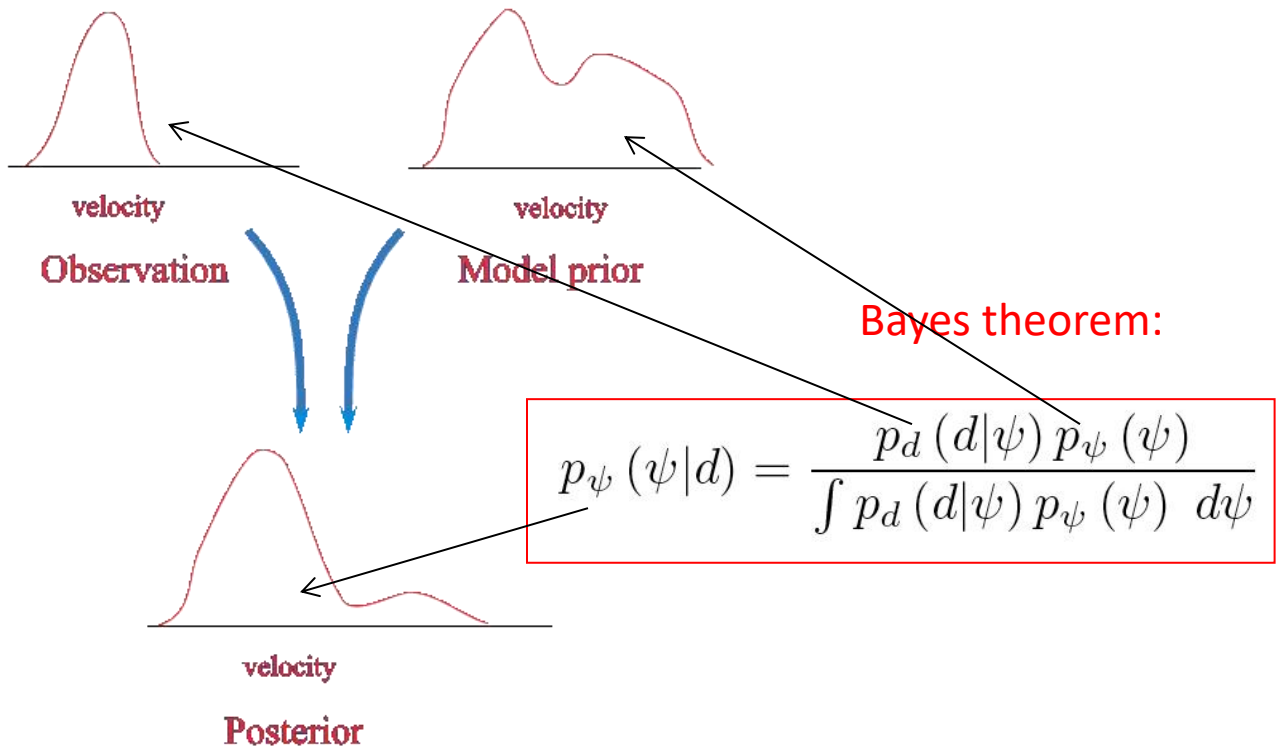
A Traditional View of Deep Learning



Do we really need huge data sets?

Bayesian View of Learning

Data Driven **Knowledge Driven**



Bayes theorem:

$$p_\psi(\psi|d) = \frac{p_d(d|\psi)\, p_\psi(\psi)}{\int p_d(d|\psi)\, p_\psi(\psi)\ d\psi}$$

What is Data-Efficiency

Example: Two students (***A*** and ***B***) are preparing for the same exam.

	Mock Test Questions (Teaching)	Final Scores
Student ***A***	100	95
Student ***B***	20	95

Can we propose a new framework for data-efficient training?

	Training Data Size	Validation Performance
Algorithm ***A***	10,000	95%
Algorithm ***B***	500	95%

same data or same model ??

Imperial College London

Why Data-Efficiency (DE)

Motivation

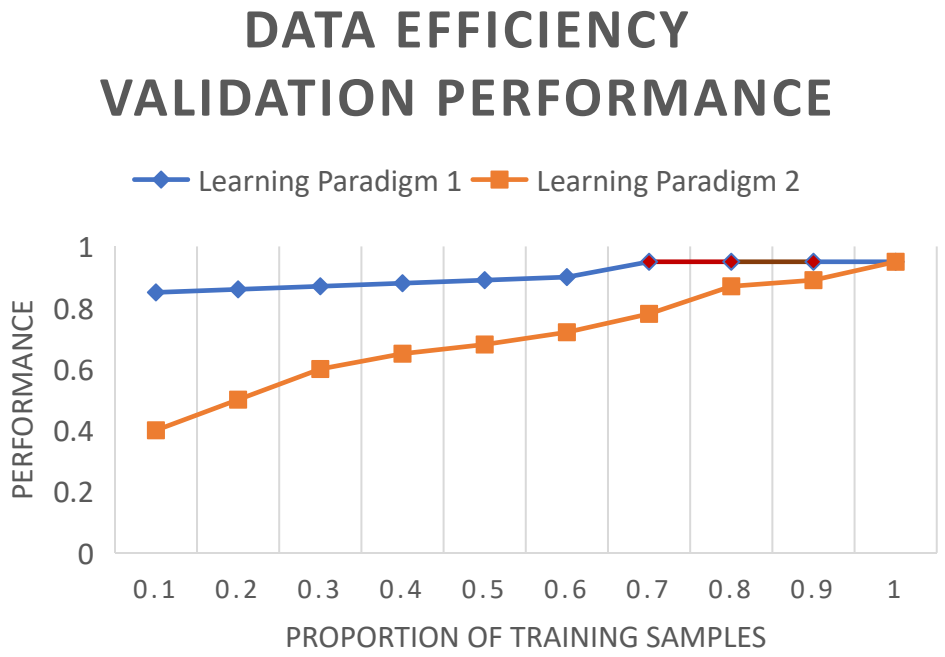
- Training a Deep neural network usually requires a large dataset with high-quality annotation—the consequence of complex model
- In many areas, lack of labels due to high cost of annotations e.g. medical images (expensive labelling), dynamic fluid field (time-consuming simulation) and etc.
- “Big data is a cure for overfitting”? Is it true?
- Is it really necessary to use such a big dataset to train neural networks? Using small dataset will significantly save computational resource. (e.g. XLNet training costs 60,000 USD!!!)

Imperial College London

Definition of Data-Efficiency

- Data efficiency can be considered as the **property** of a given learning paradigm. It describes how **efficient** the paradigm could use **training samples** to achieve a performance target.

- Namely, given a performance metric m (e.g. accuracy), data-efficiency is to find the minimum size s of training data used to achieve m.

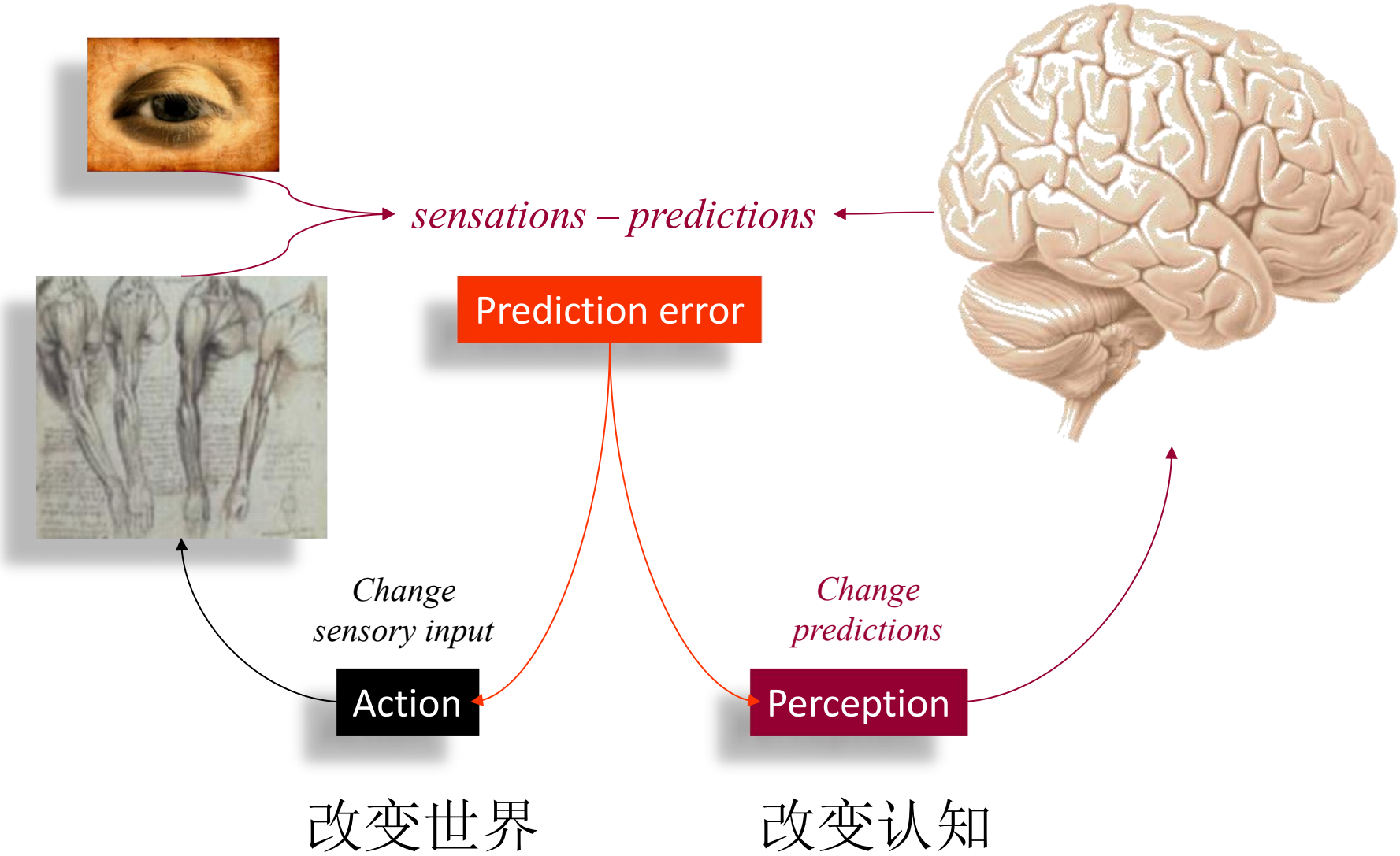


Paradigm 1 > Paradigm 2

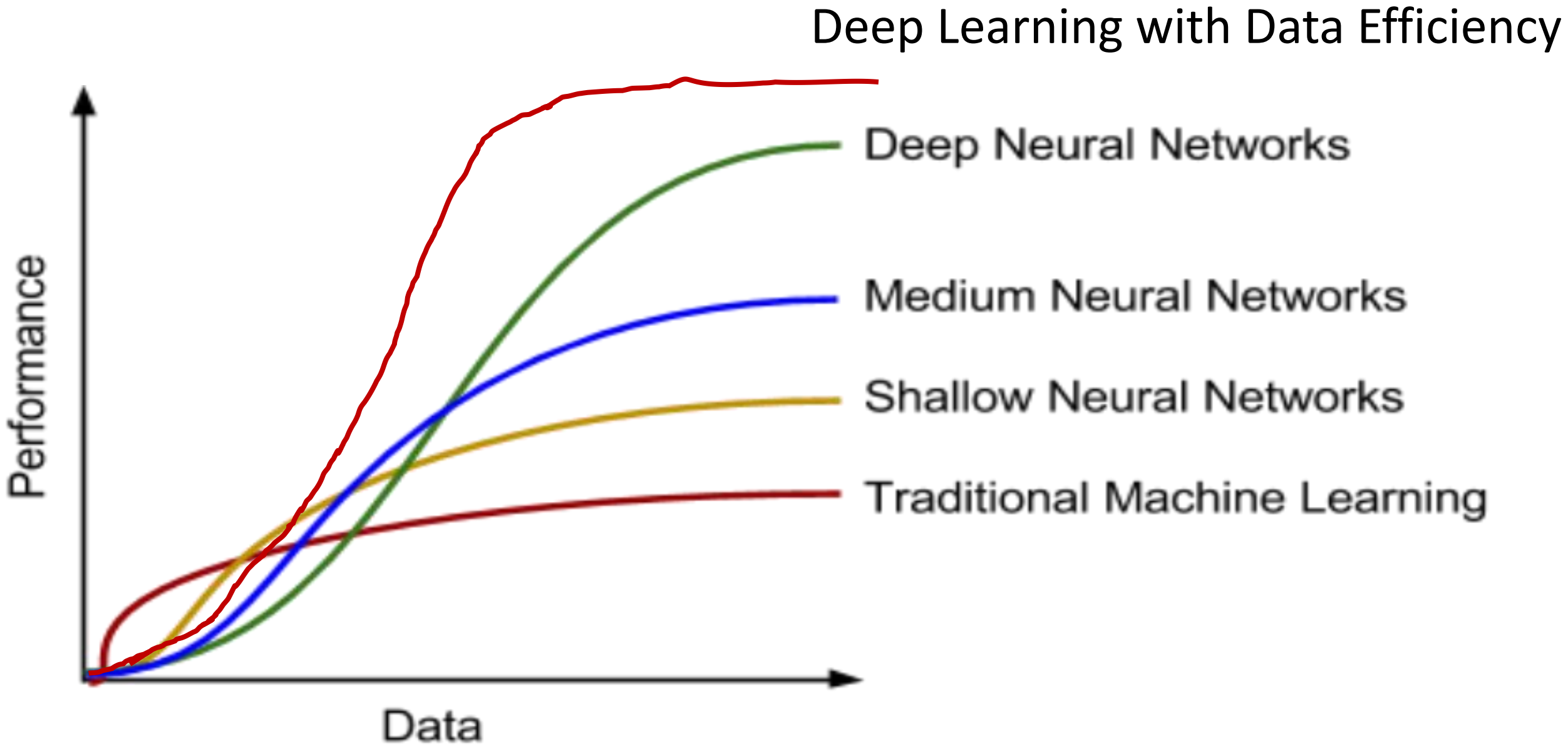
“>” means better DE

Imperial College London

Human Brain is a Data Efficient Learning Machine



Aim: Deep Learning with Data Efficiency



Current Research Trends on Data-Efficiency

General Idea: Introducing **prior knowledge** into the learning paradigm is the way to improve data-efficiency:

- External knowledge:
    1. Statistical prior
    2. Zero(few)-shot learning. (Knowledge graph as prior)
    3. Data Assimilation
    4. **Deep labelling : Transferring training data to generate label at the fine grain level**
- Internal knowledge:
    1. Sequential learning. (Boosting, bagging methods)
    2. **Transfer Learning**
    3. **Deep Boosting**. (Training status as prior for next training iteration)

Part I. Deep Labelling

Fine-grain labelling in Precise Medical Image Segmentation (Data Microscoping)

Imperial College London

Overview

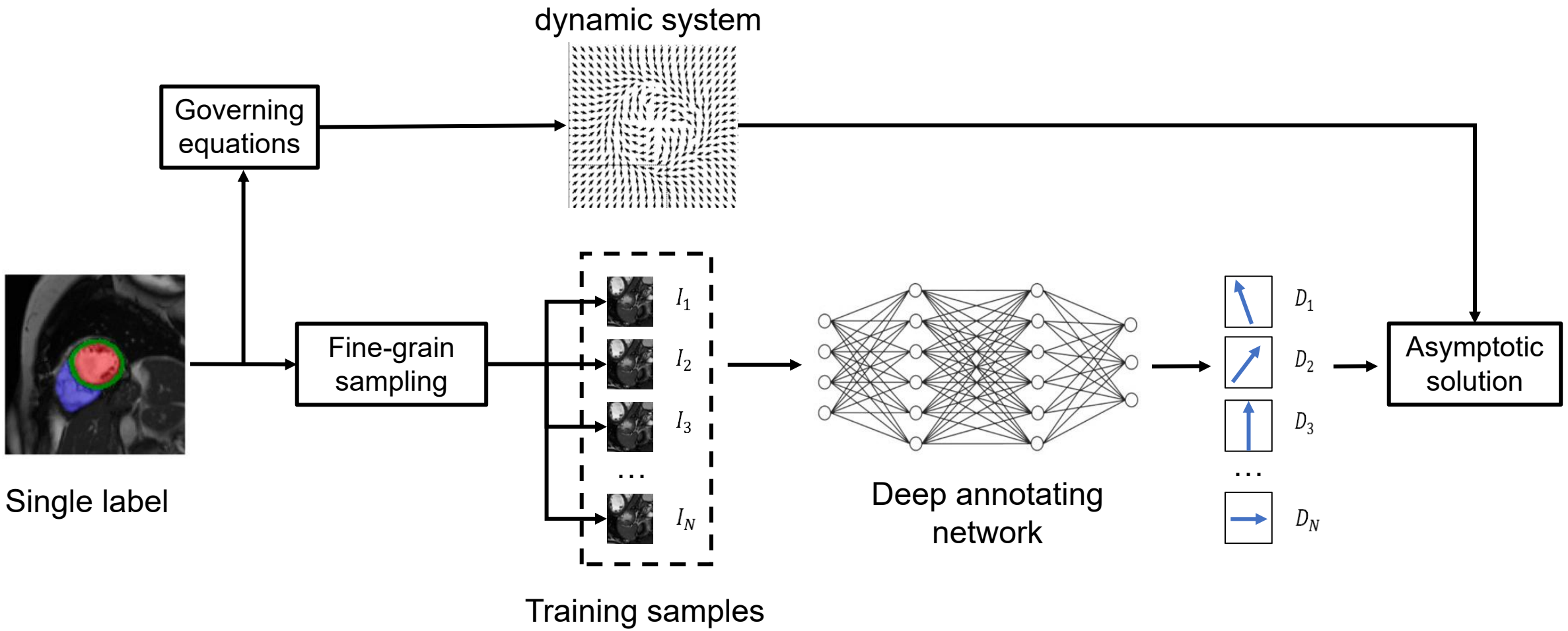
1. Medical image segmentation are usually presented as binary label maps which are lacking of some essential prior information (e.g. the drawing process of radiologists).

2. Generate/Integrate prior information by constructing a dynamic annotating system for each label.

3. This enables us to extract ‘deep’ information from a single training sample such that the data-efficiency is improved.

Imperial College London

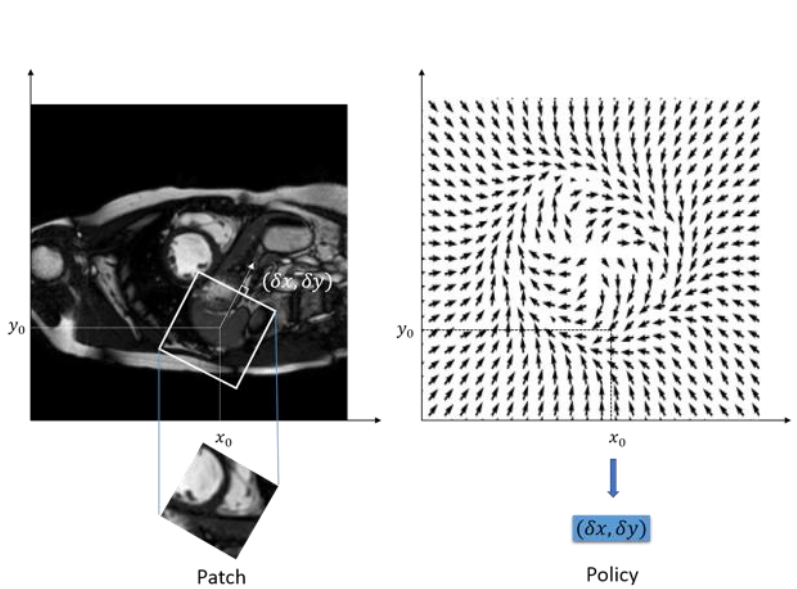
Framework

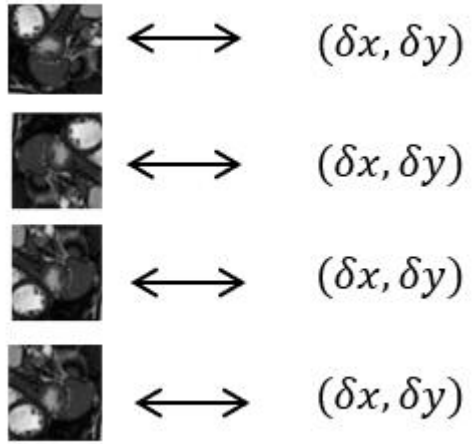


Imperial College London

Generate Patch-Displacement Pairs

Fine-grain sampling





Patch-Displacement Pairs

This is where efficiency comes from. For each sample, such generative process could produce thousands of patch-displacement pairs from one sample which will be used for training the network.

Imperial College London

Generate Dynamics for Labels

Using customized governing equations

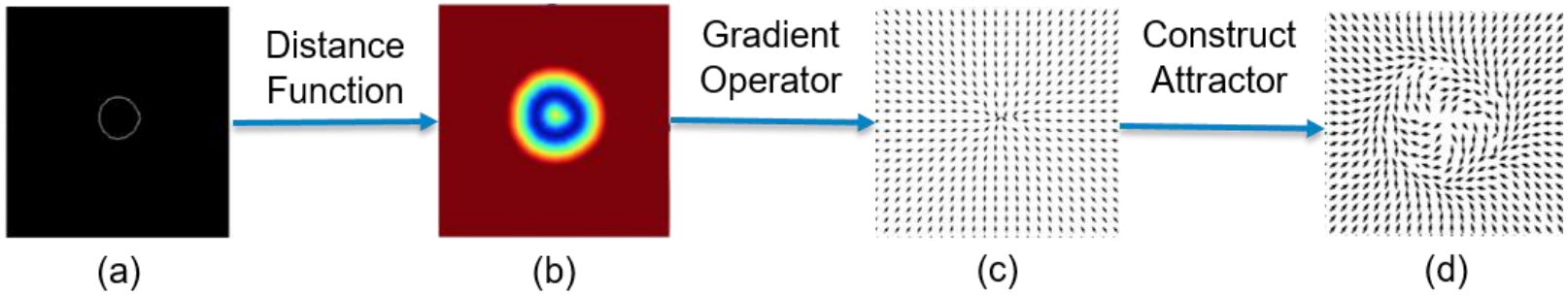
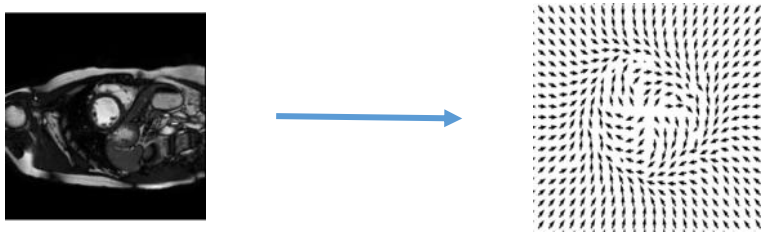


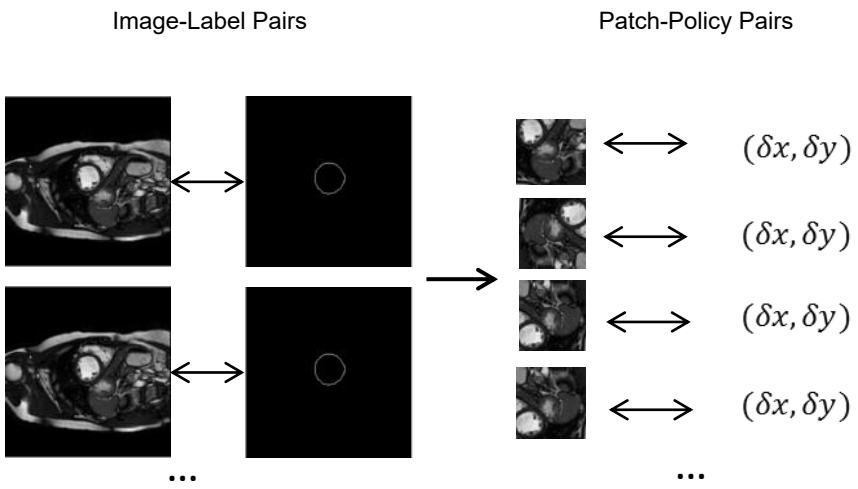
Image-to-Dynamic Pairs

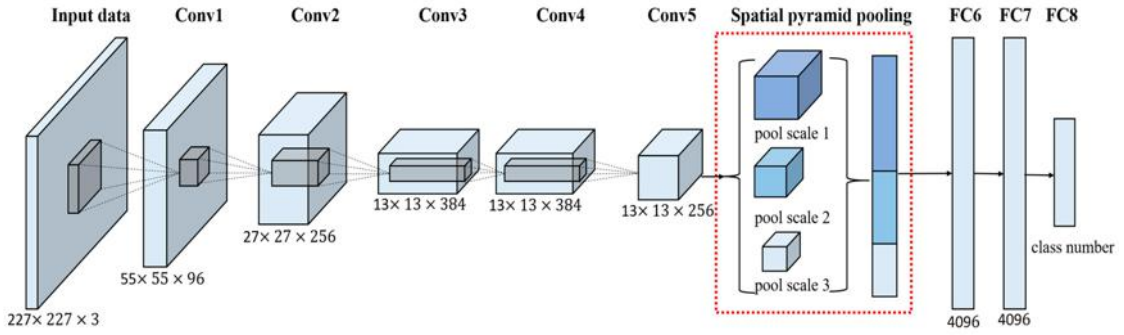


For each label, a distance function is applied to transfer a binary map into a scalar field. A gradient system can be constructed based on the given scalar field. Finally, an attractor is added to the gradient system providing a periodic solution. This step integrates the prior knowledge into each label.

Imperial College London

Training with fine-grain samples

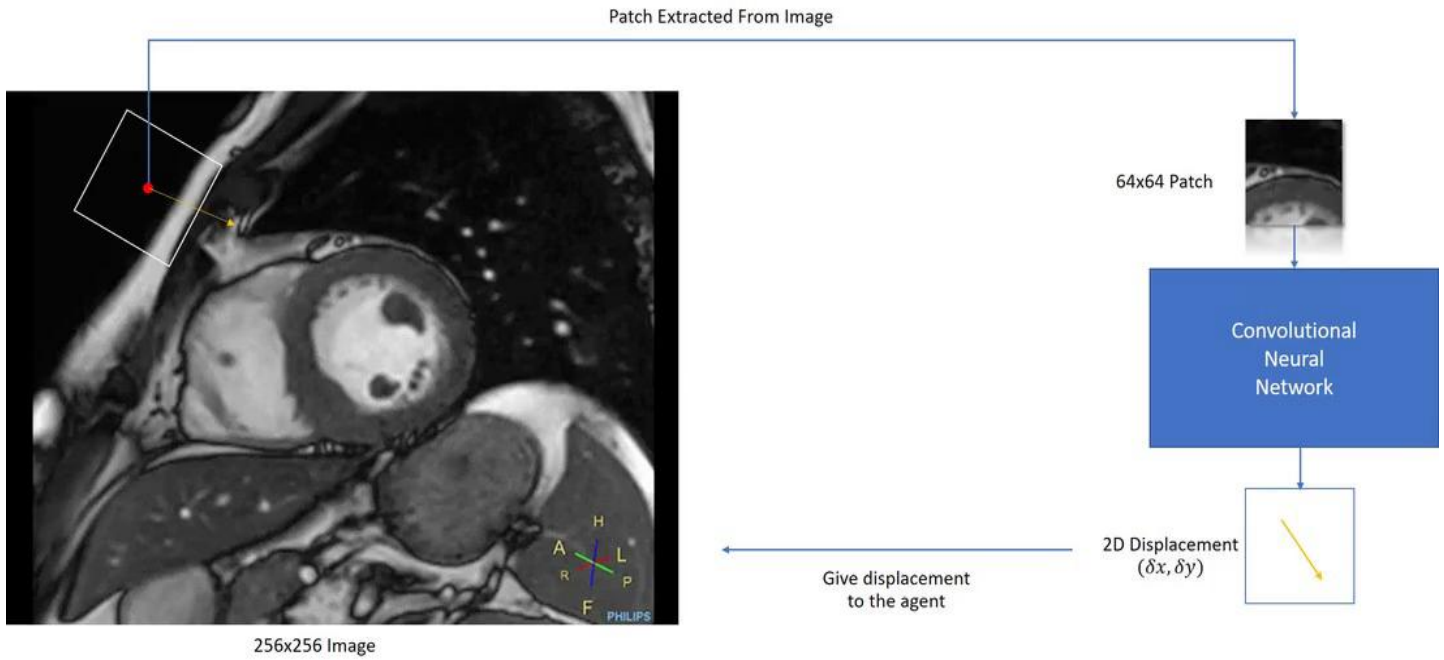




The displacement is output of the trained network, allowing an iterative annotating process.

Imperial College London

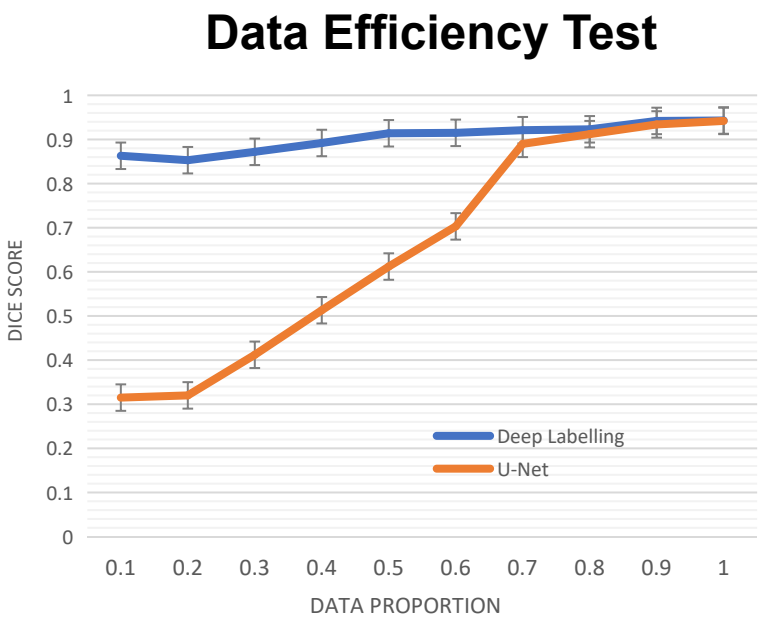
Inference stage

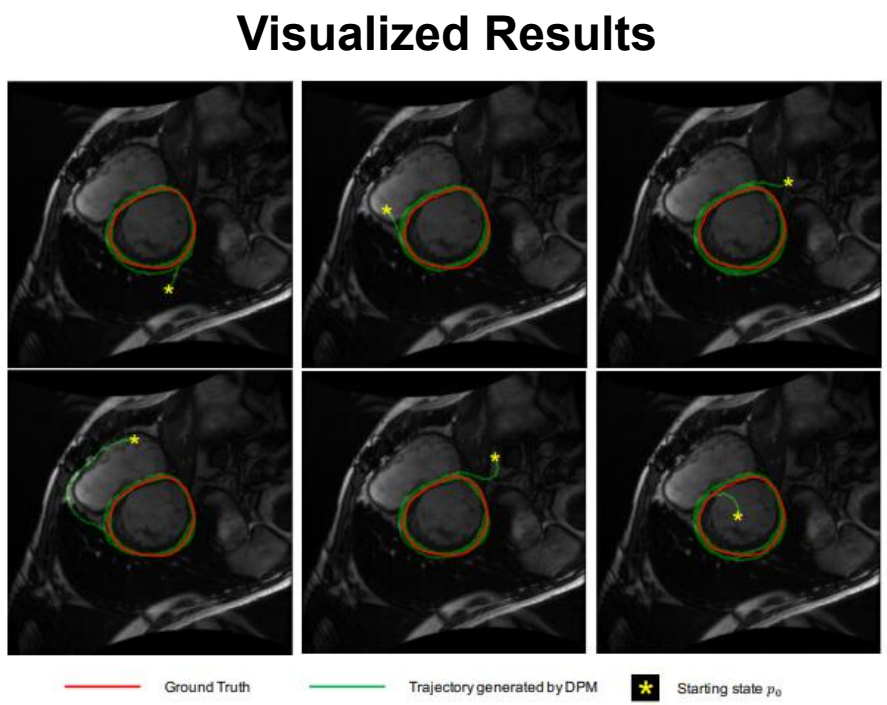


- The method operates on a local patch.
- Stopping criterion is based on Poincaré map.
- After finite times of iteration, the trajectory will converge to the boundary of ROI.

Imperial College London

Results





Imperial College London

Part II. Transfer Learning

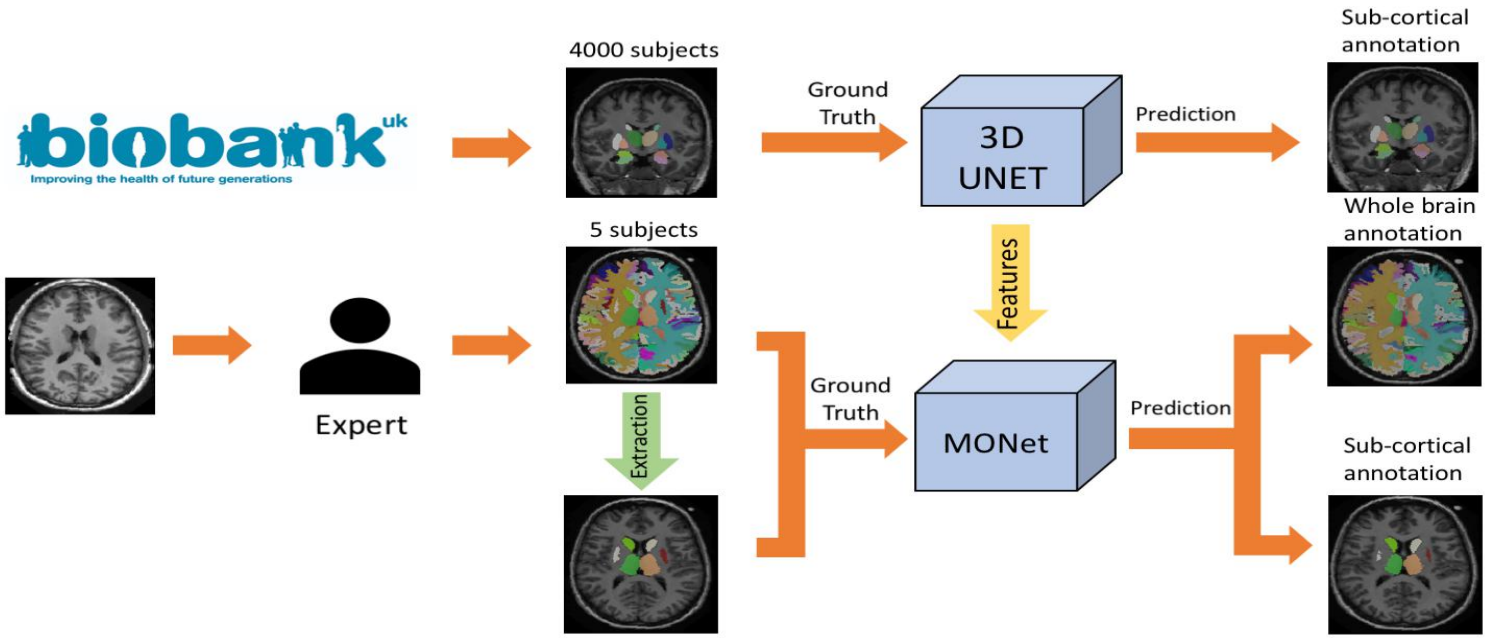
From Partial Annotations for Whole Brain Segmentation

Imperial College London

Anatomical Prior in Partial Annotations

1. Fine segmentations for anatomical structures from the medical image can be difficult to acquire. There are not enough of them to train a robust machine-learning model.
2. Partial annotations are easier to generate (e.g. manual or semi-automatic) and the availability is much better.
3. The anatomical prior learnt from the partial annotation can be transferred for more detailed segmentation.

Gain Prior Knowledge via Partial Annoation

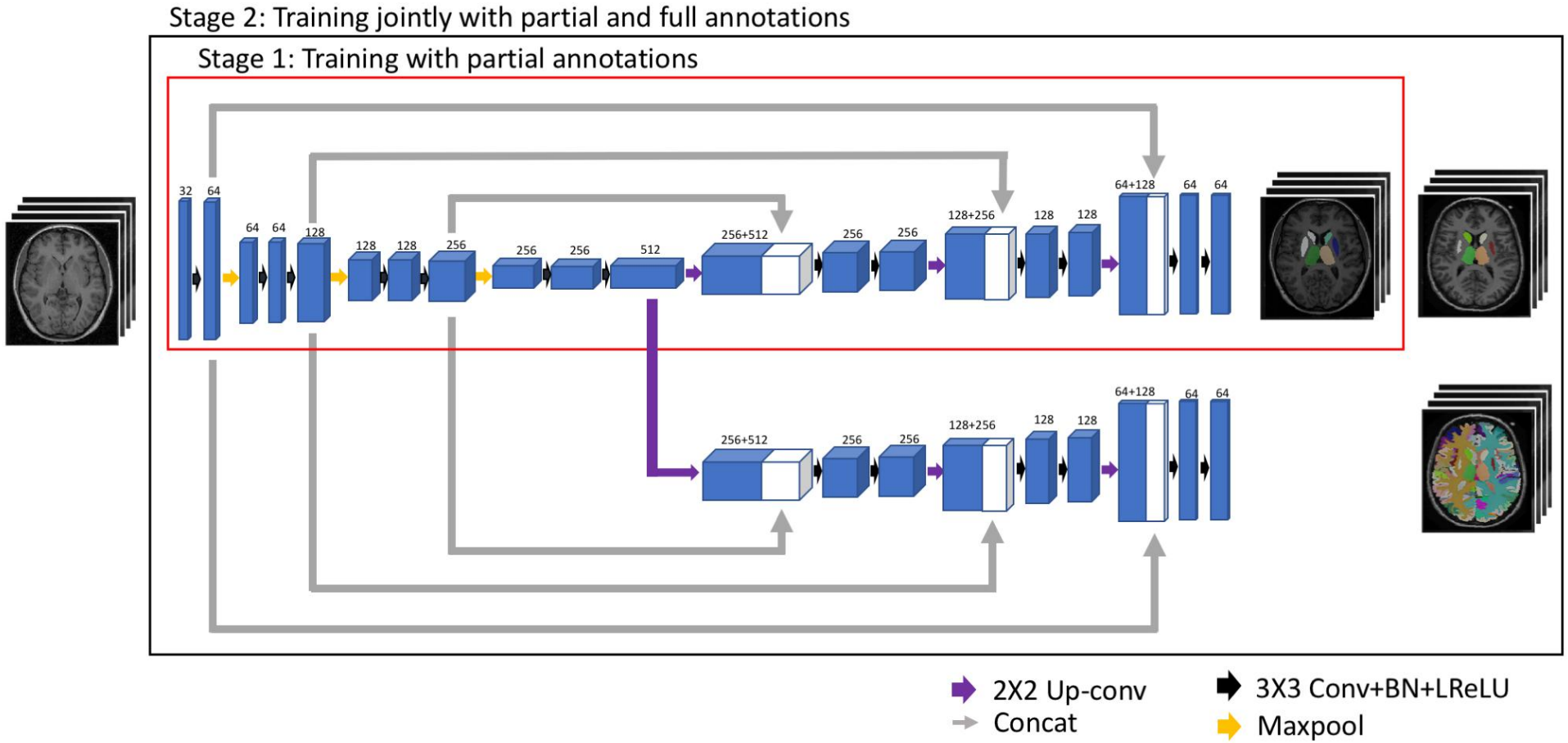


- Stage one: learning partial annotation with UNET.
- Stage two: jointly learning whole and partial segmentation from features leant in stage one.

Partial annotation refers to segmentation that only covers part of the brain structures. In our case, it refers to segmentation of 15 sub-cortical structures automatically generated by FSL.

Full annotation refers to segmentation of whole brain structures manually annotated by human experts, which is a superset of partial annotation and

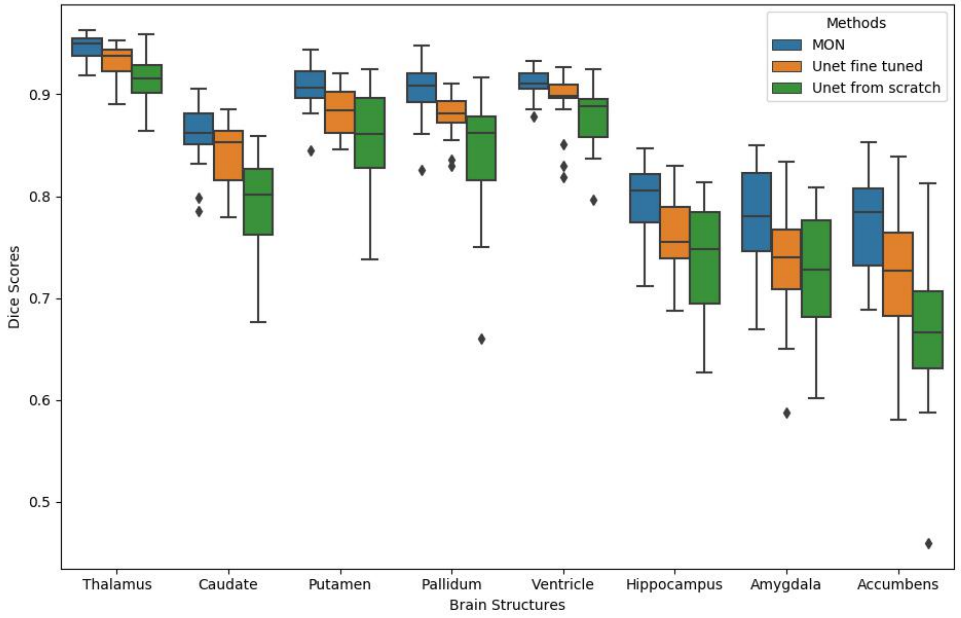
Neural Network for Multi-task Training



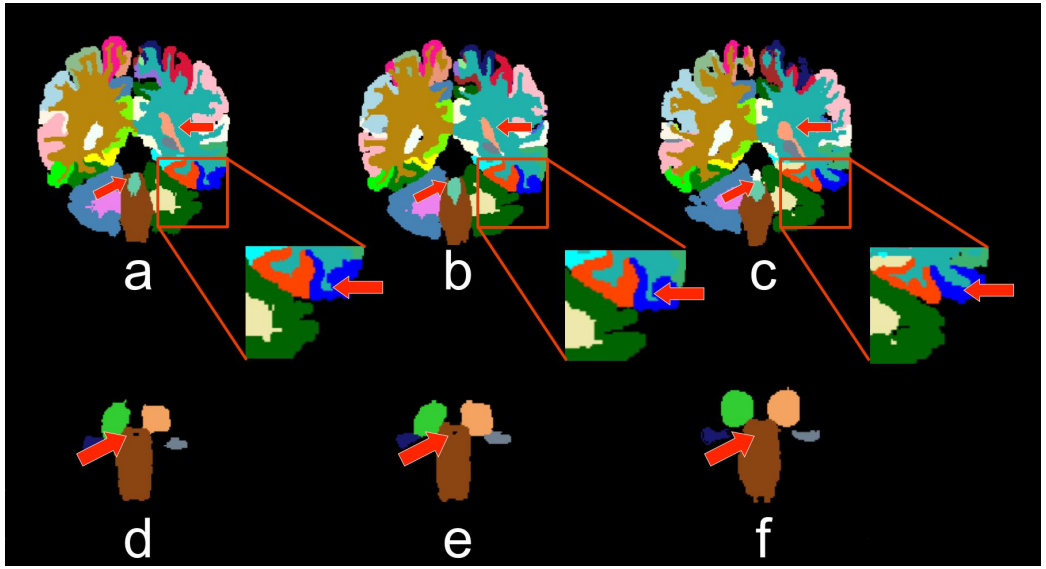
The multi-output network (MO-Net) where encoder and both decoders are loaded with the pre-trained parameters. Multi-output design encourages the encoder to learn shared features for partial segmentation and full segmentation

Results

Box-plot of Dice scores of models with and without incorporating anatomical prior



Visual inspection of segmentation result



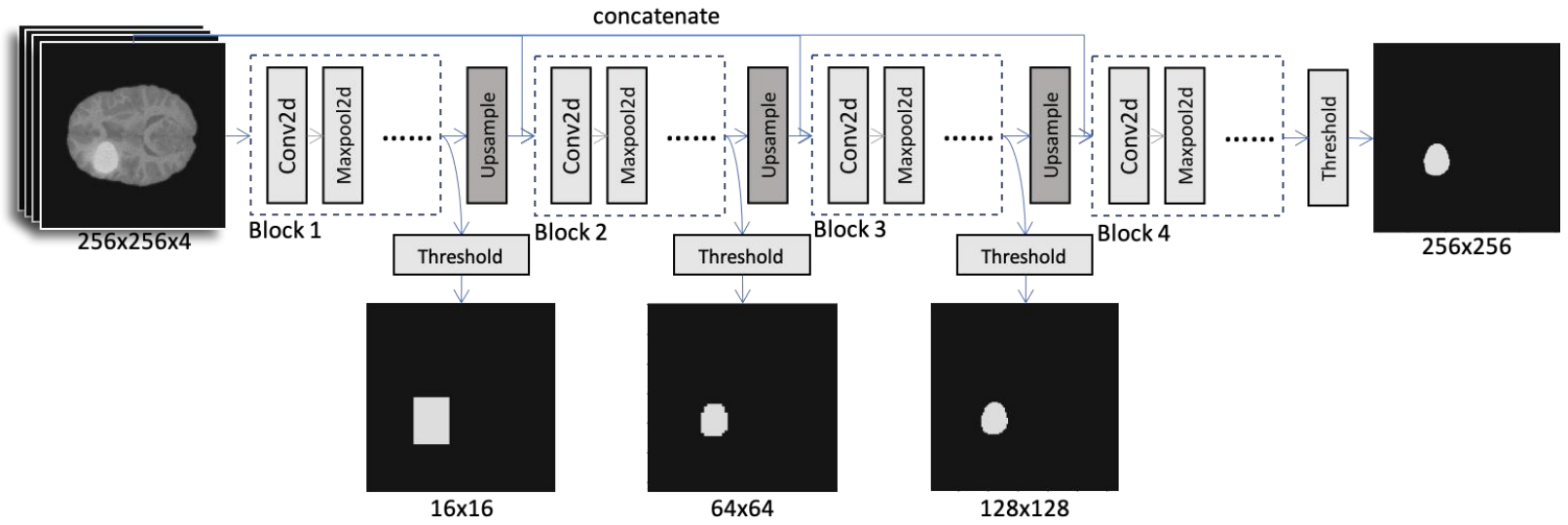
Ground truth of full (a) and partial (d) brain segmentation from the expert, full (b) and partial (e) brain segmentation from MO-Net full (c) segmentation from fine-tuned U-Net, and sub-cortical (f) segmentation from FSL. Red arrows indict regions where MO-Net looks consistent with manual annotations and outperforms other methods

Part III. Greedy Block-wise Training

For Brain Tumors Segmentation

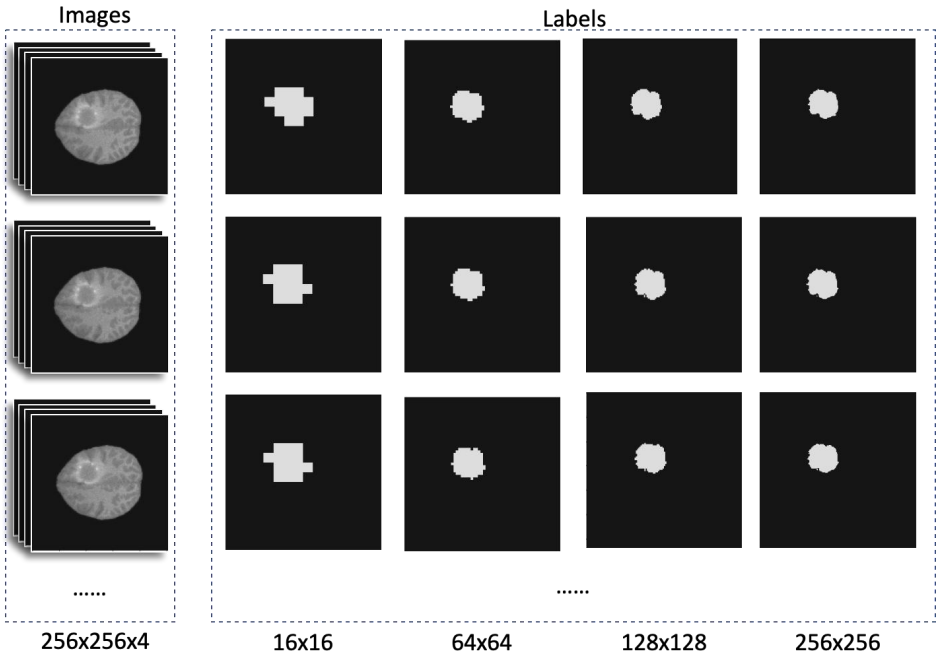
Imperial College London

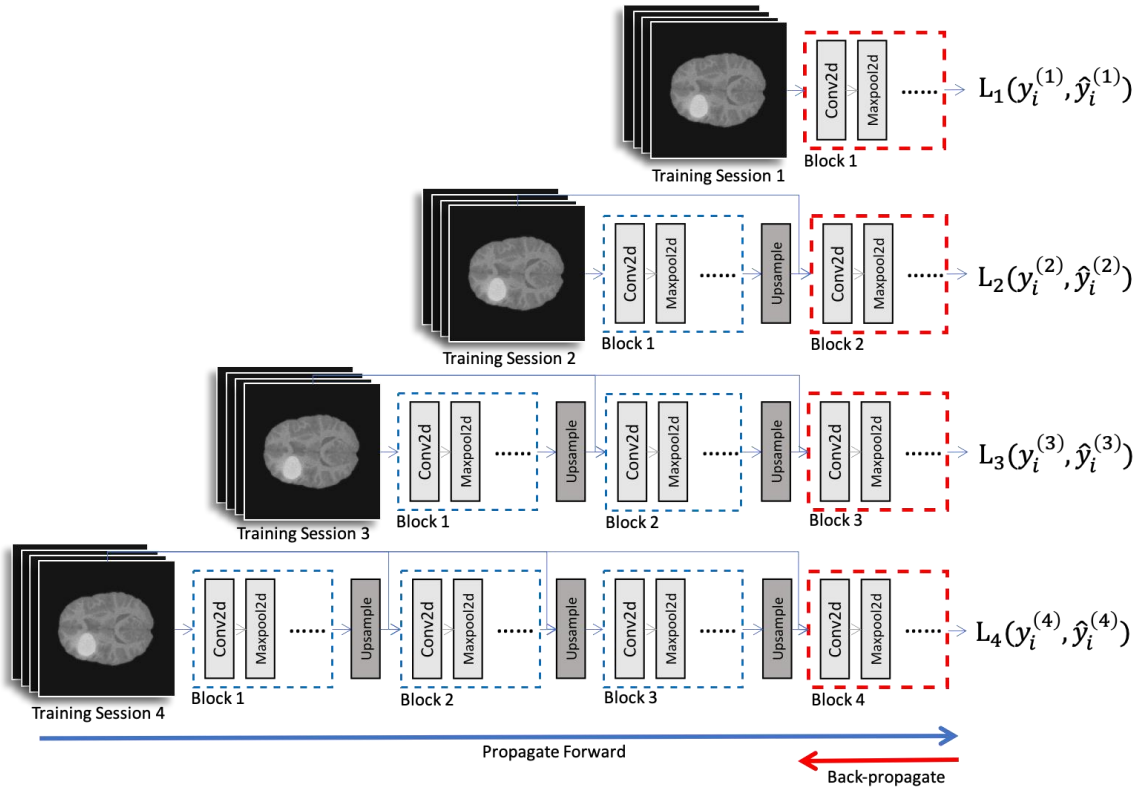
Proposed Network



- The proposed architecture is a sequence of computational blocks containing a number of convolutional layers in which each block provides its successive block with a coarser segmentation map as a reference.

Neural Network for Block-wise Training





- Before training stage, the proposed method requires a data transformation step which converts each label into different resolutions for the training of each blocks.
- Then, the proposed training scheme will be adopted to train each block according to a feedforward order. In so doing, each trained block is able to provide its successive block with a meaningful and coarse segmentation results, which gradually increases training difficulty instead of starting with the hardest samples.

Results

Table 1: Comparison of brain tumour segmentation performance on BRATS13.

Methodology	Dice			Sensitivity		
	WT	CT	ET	WT	CT	ET
Havei et al.[13]	0.88	0.79	0.73	0.87	0.79	0.80
Urban et al. [14]	0.87	0.77	0.73	0.92	0.79	0.70
Pereira et al.[15]	0.88	0.83	0.77	0.89	0.83	0.81
T. HNL et al. [16]	0.89	0.79	0.74	0.90	0.89	0.93
Proposed Method	0.91	0.80	0.71	0.95	0.90	0.92

Table 2: Comparison of brain tumour segmentation performance on BRATS15.

Methodology	Dice			Sensitivity		
	WT	CT	ET	WT	CT	ET
Chang et al. [17]	0.87	0.81	0.72	-	-	-
Deep Medic [18]	0.896	0.754	0.718	0.903	0.73	0.73
DMRes [18]	0.896	0.763	0.724	0.92	0.754	0.763
T. HNL et al. [16]	0.88	0.82	0.73	0.91	0.76	0.78
Proposed Method	0.89	0.80	0.73	0.94	0.81	0.80

- We trained our network on the training sets of BRATS 2015 and BRATS 2013 respectively.
- We report the Dice metric and sensitivity on three tasks of brain tumour segmentation. They are:
  - the whole tumour (WT) region
  - the core tumour (CT) region
  - the enhancing tumour (ET) region

Part IV. Deep Boosting

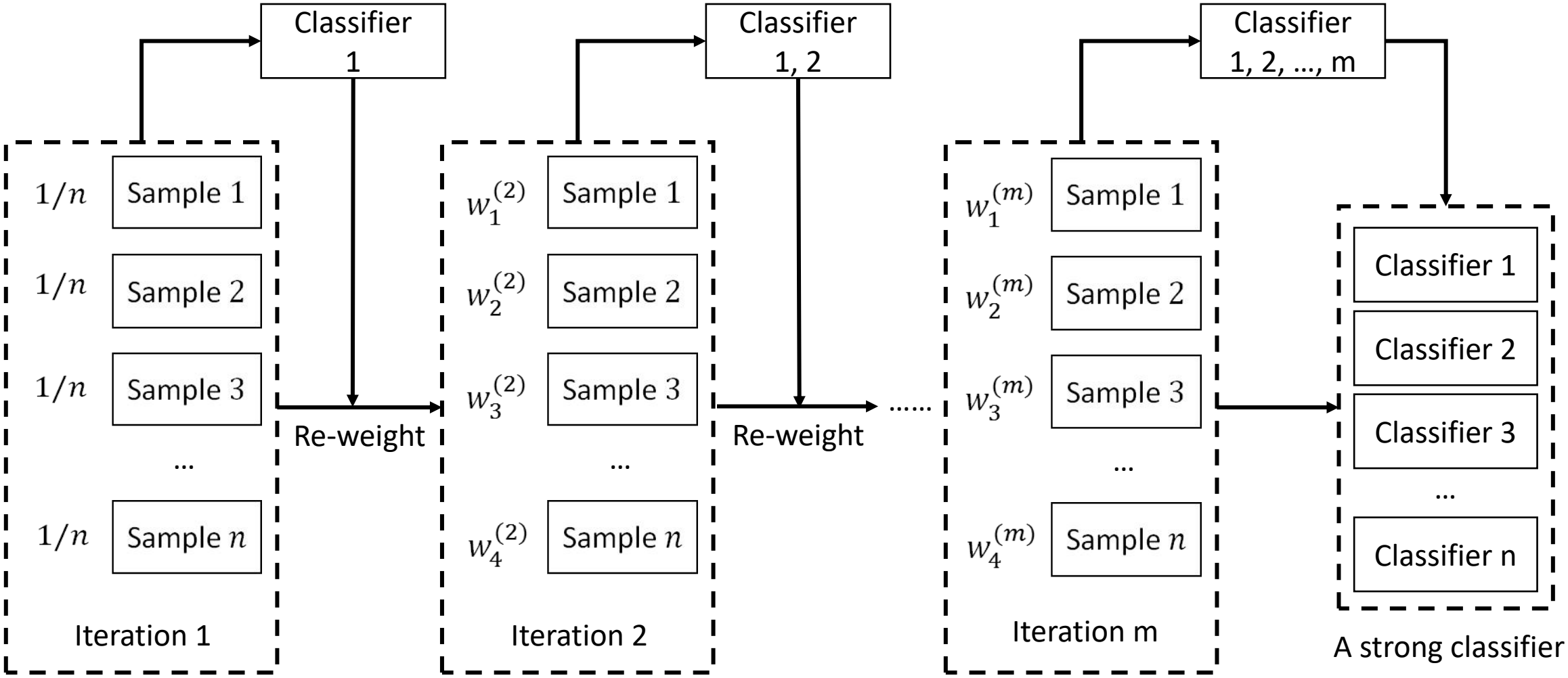
Sequential Generation of Efficient Training Samples

Imperial College London

What is Boosting

1. A boosting algorithm is usually an **iterative** process that progressively learn a strong classifier. At each iteration, an **auxiliary** classifier will be created using the training samples which is **re-weighted** according to the classification results of last iteration. Finally, the **auxiliary** classifiers are ensembled as a strong classifier.
2. The re-weighted training samples of current training iteration are the prior for the next training iteration.

What is Boosting



Adaptive sample re-weighting scheme

Imperial College London

What is Deep Boosting

1. Real-world data is distributed in high-dimensional space but concentrate on manifolds $\mathcal{M}$ (contains redundant information).
2. Learn a low-dimensional data structure $\mathcal{D}$ for efficient representation.
3. Select or generate the most 'efficient' samples from $\mathcal{D}$ to be annotated and then used as training data.

Imperial College London

Efficient Data Representation

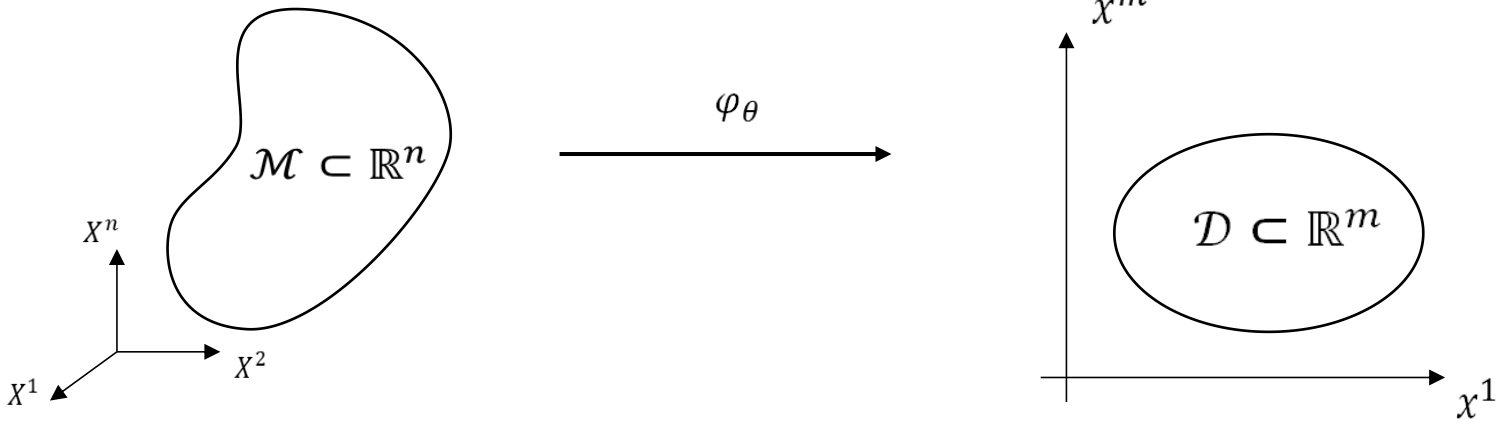
Raw Data

Manifold learning

Data structure

High-dimensional manifold

Low-dimensional latent space

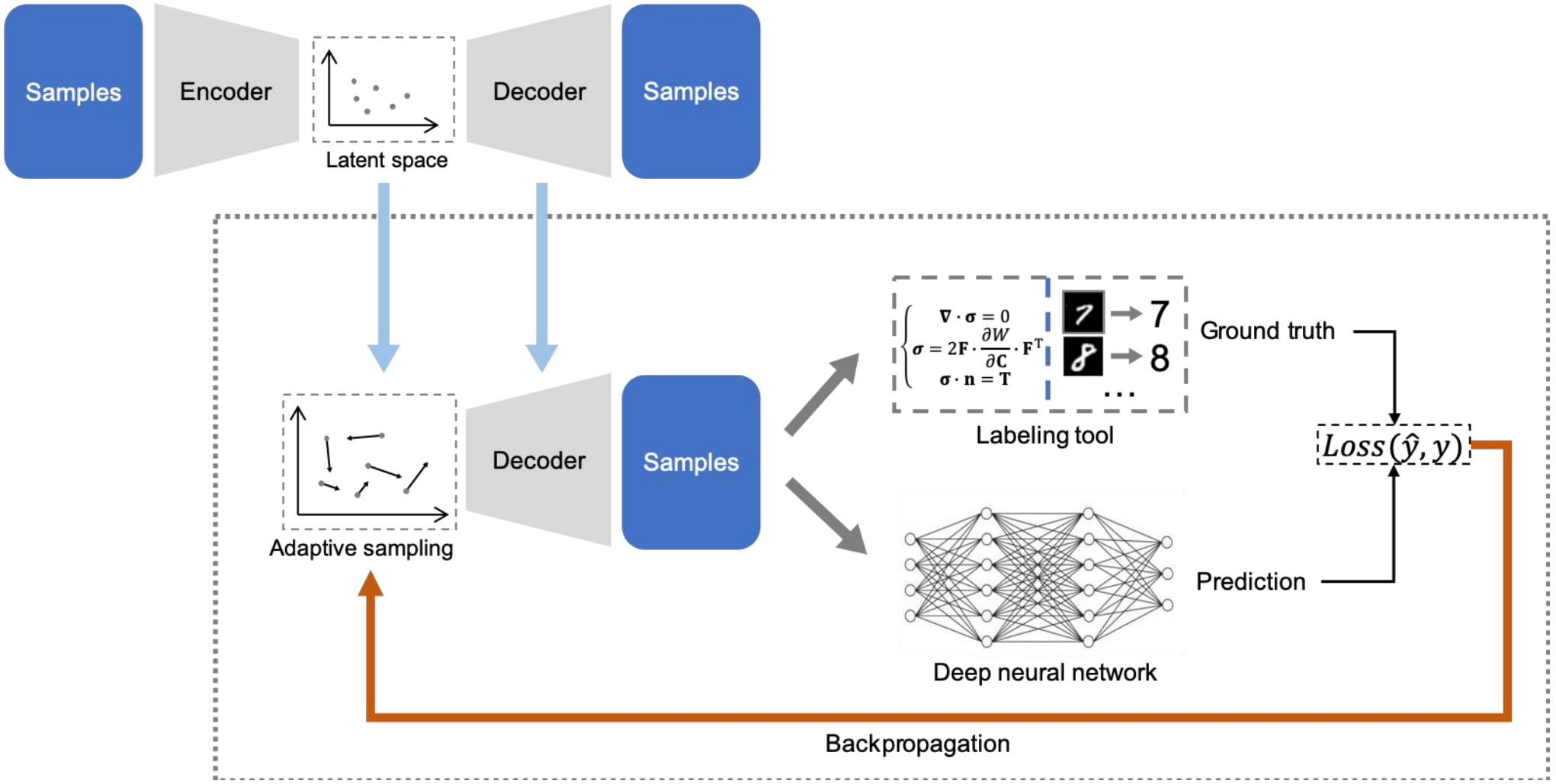


Learning the intrinsic data structure through manifold learning:

- Kernel tricks
- PCA, ICA, LDA, t-SNE …
- Auto-encoder

…

Proposed Framework

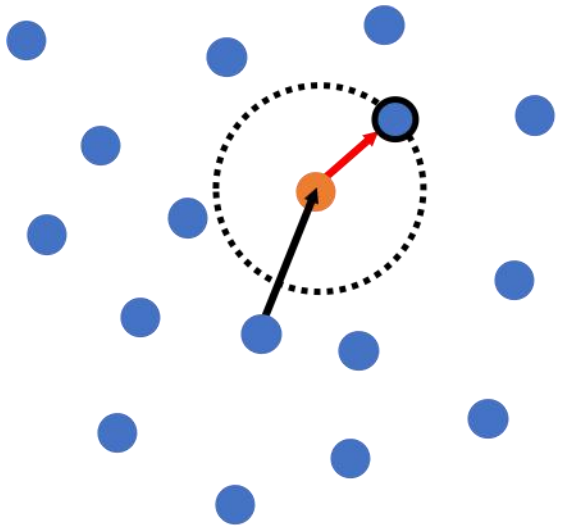


Framework

1. The preparation stage represents the training of a variational auto-encoder (VAE) using unannotated samples.
2. In the main stage, the decoder (generator) as well as its latent space are used for mining hard training samples according to the error information propagated backward via the target model and decoder (generator).
3. Each proposed sample will be annotated by the labeling tool

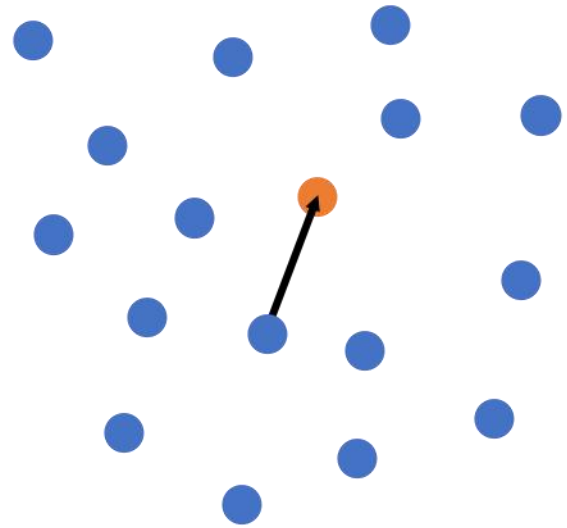
Imperial College London

Two Sampling Schemes in Latent Space



SNN

When there is no labelling tools available. We can use the nearest neighbours to train the target model.

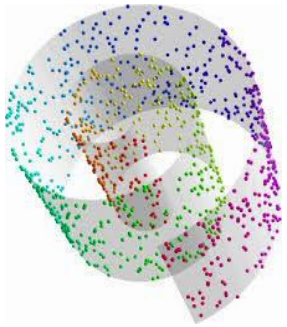


SI

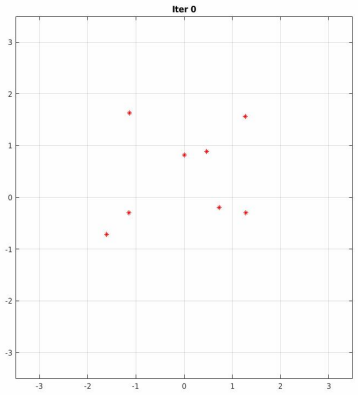
When there is a labelling tool available. We can use the interpolated point to train the target model.

Framework

Raw data



latent representation

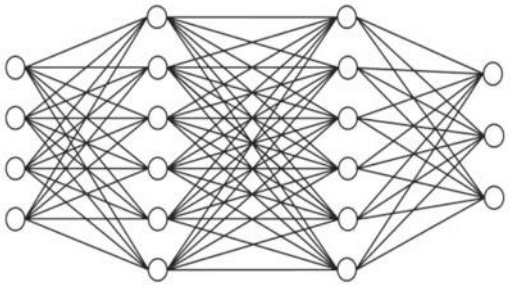


Active sampling
for specific task

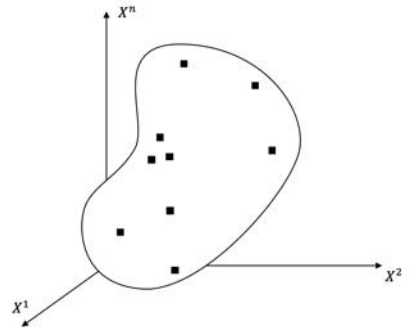
Deep boosting engine

$$\begin{bmatrix}(S_1,P_1)\\(S_2,P_2)\\\cdot\\\cdot\\\cdot\\\cdot\\\cdot\\(S_n,P_n)\end{bmatrix}^{(1)} \quad \dots \quad \begin{bmatrix}(S_1,P_1)\\(S_2,P_2)\\\cdot\\\cdot\\\cdot\\\cdot\\\cdot\\(S_n,P_n)\end{bmatrix}^{(m)}$$

Adaptive training batches



Deep neural network



Efficient training set

Imperial College London

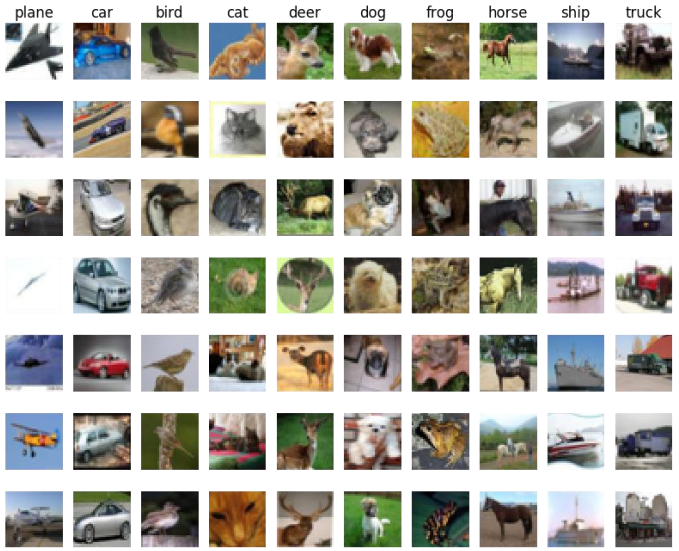
Results

- Deep boosting (DB) progressively extend the training data, which means, at certain iteration, DB could achieve same loss with less size of training samples.
- With more training samples coming, the loss maintains the decreasing trend which demonstrates the robustness of DB.

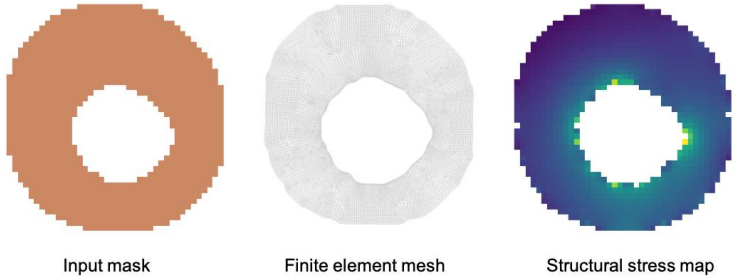
Imperial College London

Dataset --- MNIST, Cifar10, IVUS

MNIST



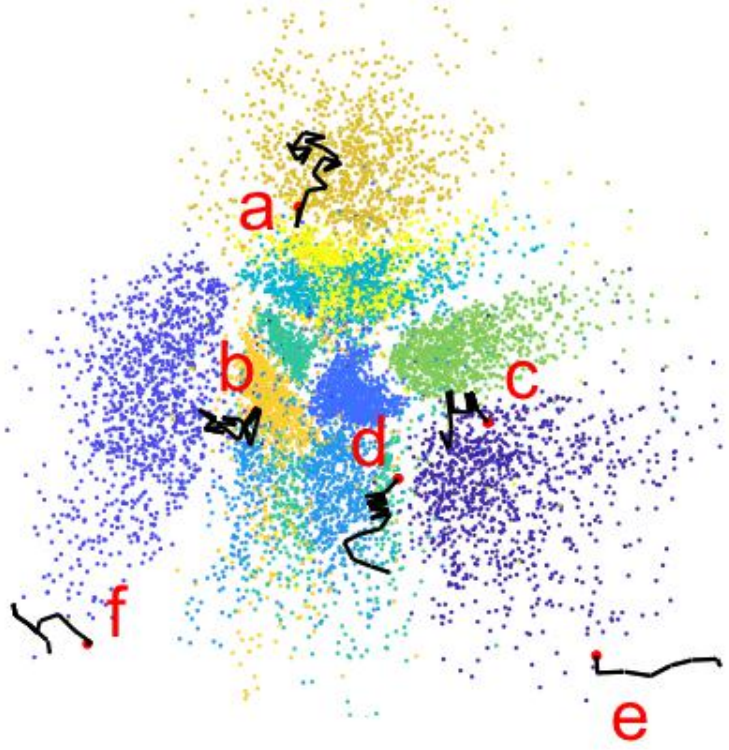
Cifar10



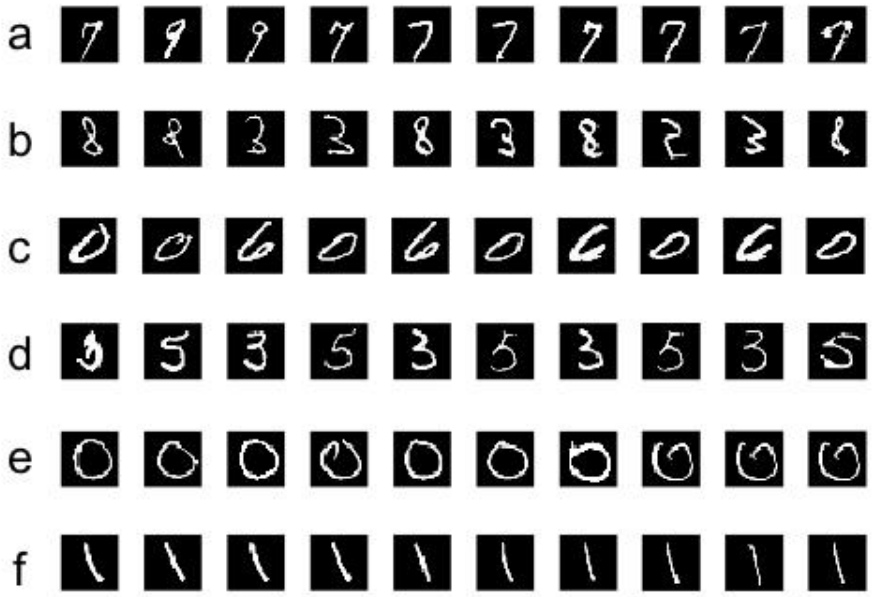
IVUS

Imperial College London

Visualization on the MNIST



2D latent space

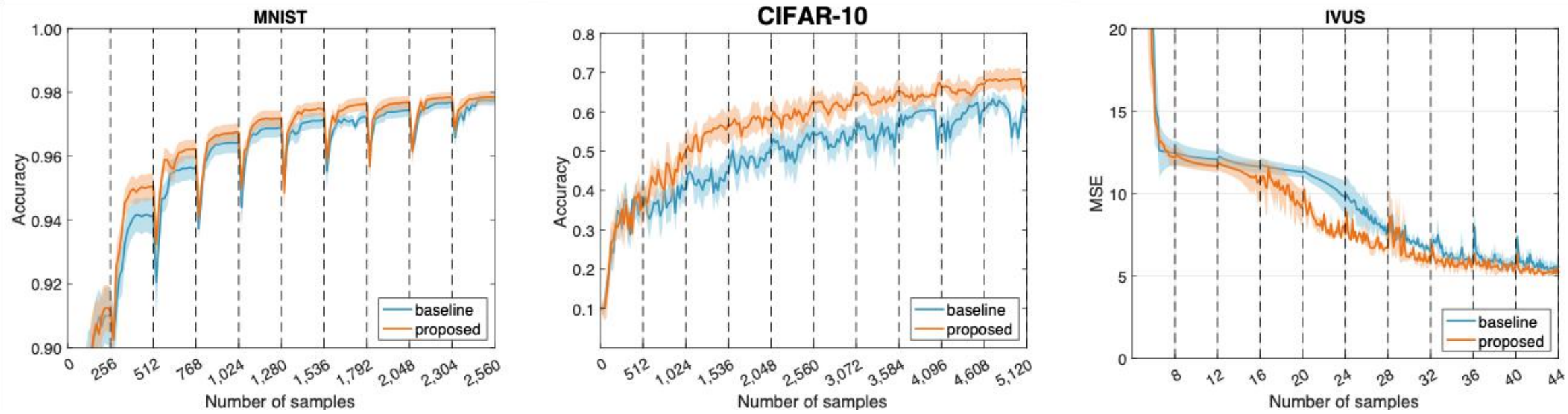


Sampling snapshots

- a couple of initial sampling point are evolved using rules of sampling by nearest neighbour
- 6 typical trajectories are selected to be visualized in the 2D MNIST latent space.
- It can be observed that trajectories like **a**, **b**, **c, d** are the most desired exploration strategy as they walk around the boundary between classes, where

Imperial College London

Results



We progressively increased the size of train set and reported the accuracy and means square error (MSE) on the independent test sets respectively.

Imperial College London

Data Efficiency Will Not Compromise Performance

Conclusion

- Our efficient learning algorithms provide a small-sample solution for image learning system.
- The cost of medical imaging analysis can be reduced significantly.
- The proposed frameworks can be easily generalized to other areas where high-quality annotation is difficult to obtain, e.g. material optimization, drug discovery and medical text analysis (HER).
- Hope the similar approach can be developed in NLP