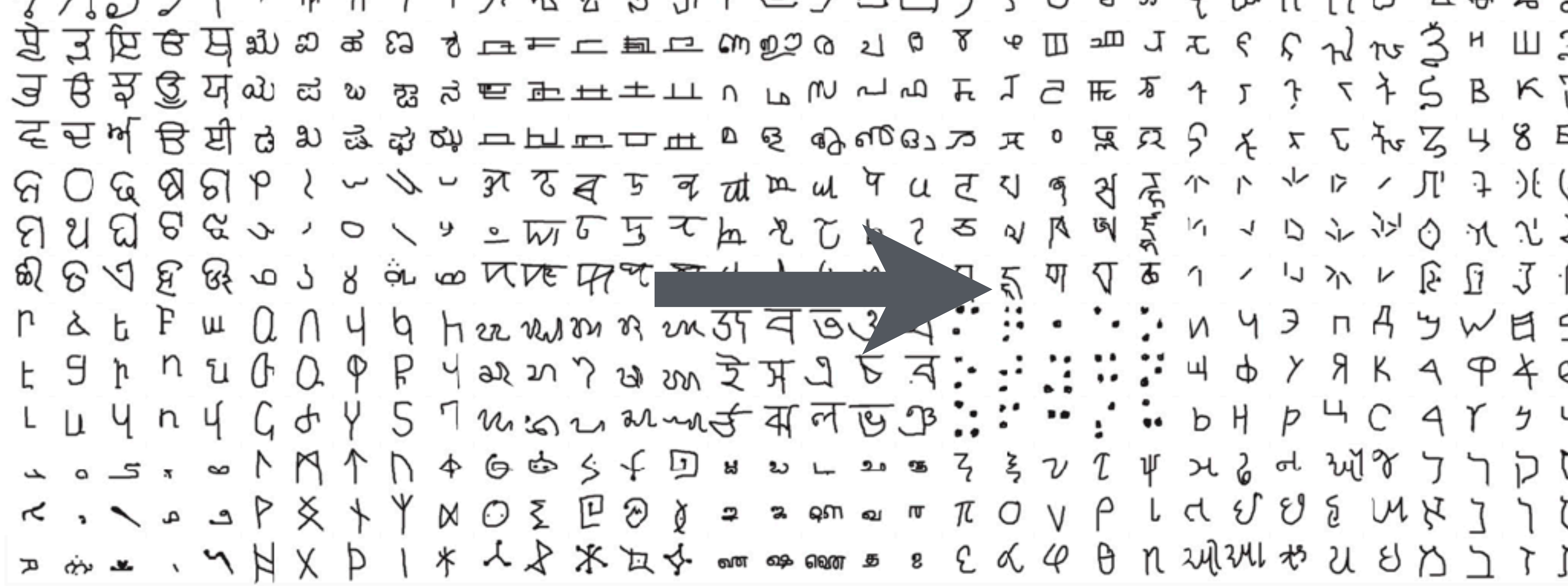
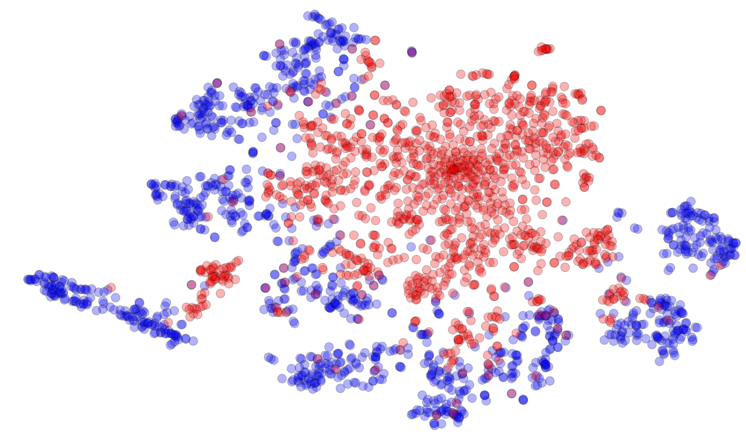


A Tutorial on Transfer Learning

Yang Li
2024/6/14

Outline

- What’s Transfer Learning
- Traditional transfer learning algorithms
  - Task transfer learning
  - Domain adaptation
  - Transfer bound on domain adaptation
- When to transfer?
  - Transferability estimation
- Research trends
  - *Transfer learning in the age of foundation models*

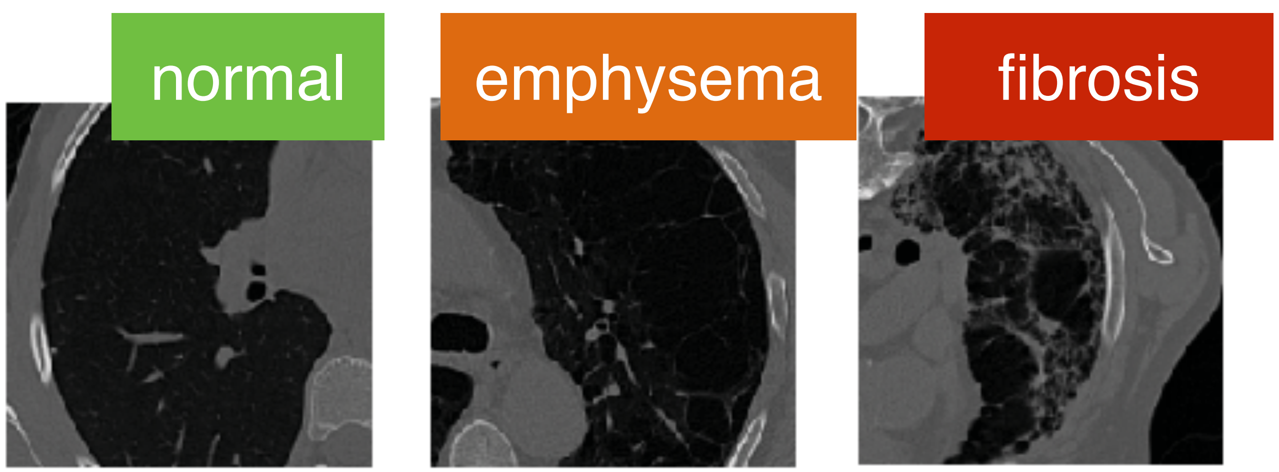


Why we need transfer learning?

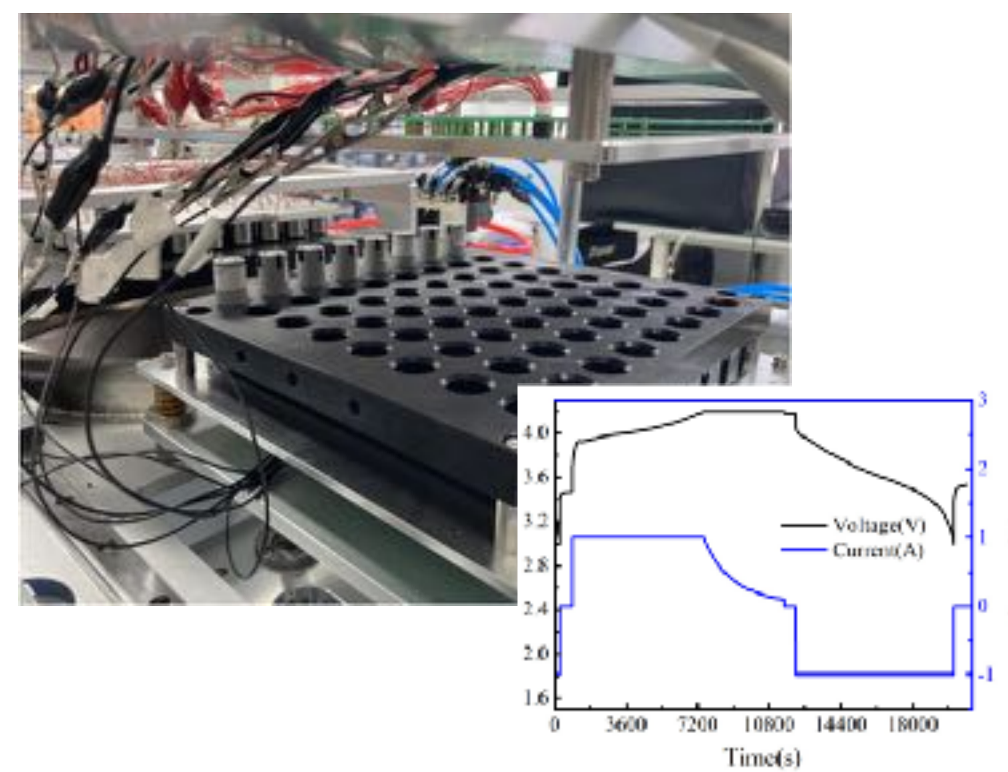
When facing a new learning task

- **Lack of annotations**: Training labels may be expensive to obtain
- **Limited training time or resource**: can't train from scratch every time

Medical image classification



Battery capacity estimation

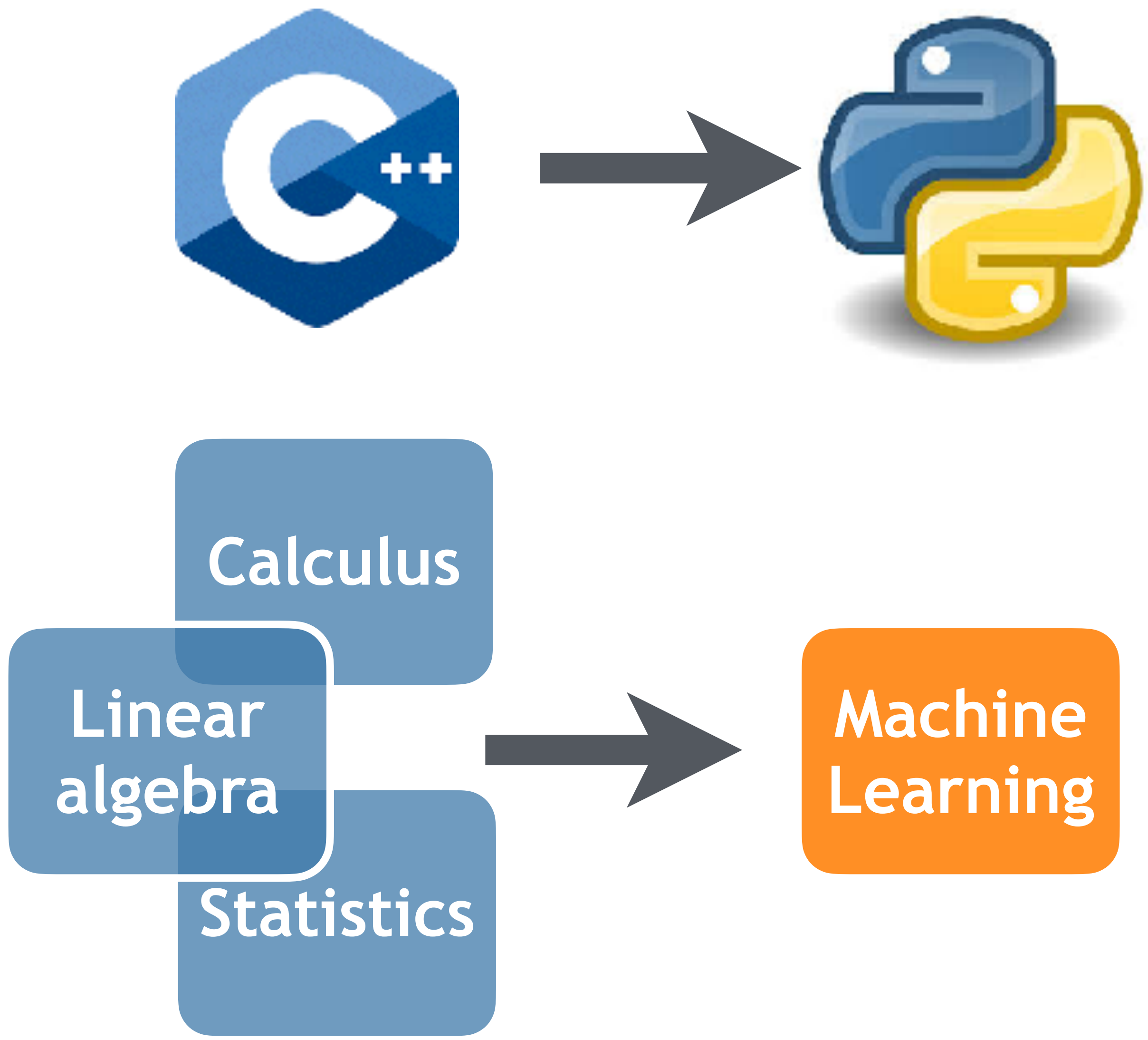


Transfer learning

- Human learners can inherently transfer knowledge between tasks

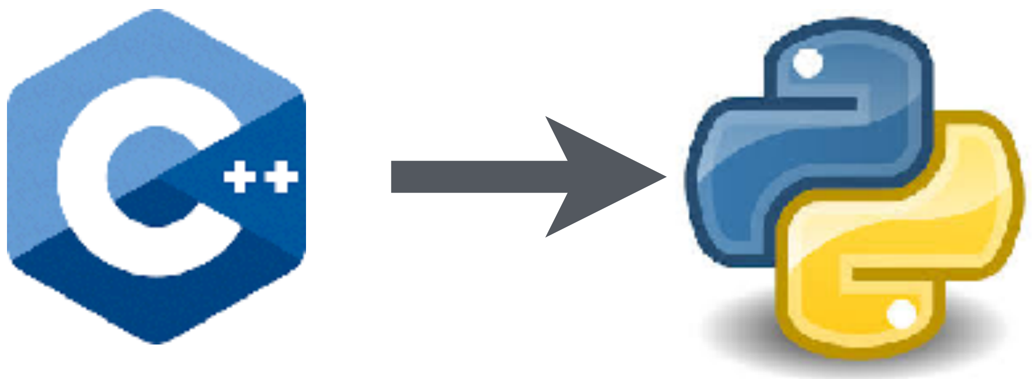
Transfer learning

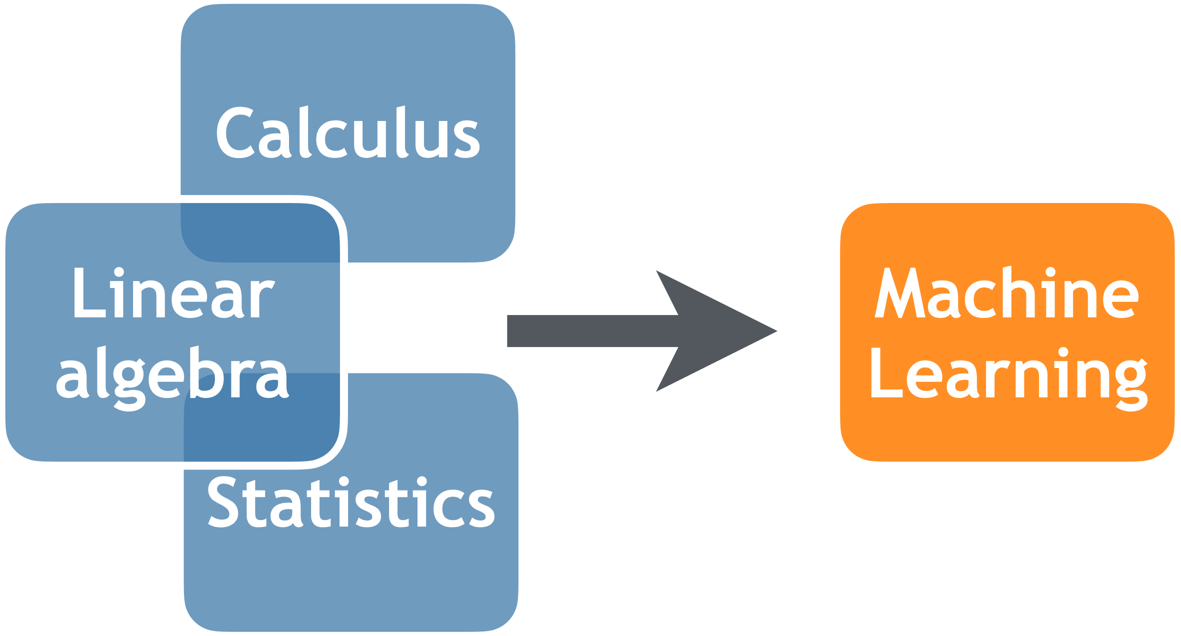
- Human learners can inherently transfer knowledge between tasks



Transfer learning

- Human learners can inherently transfer knowledge between tasks

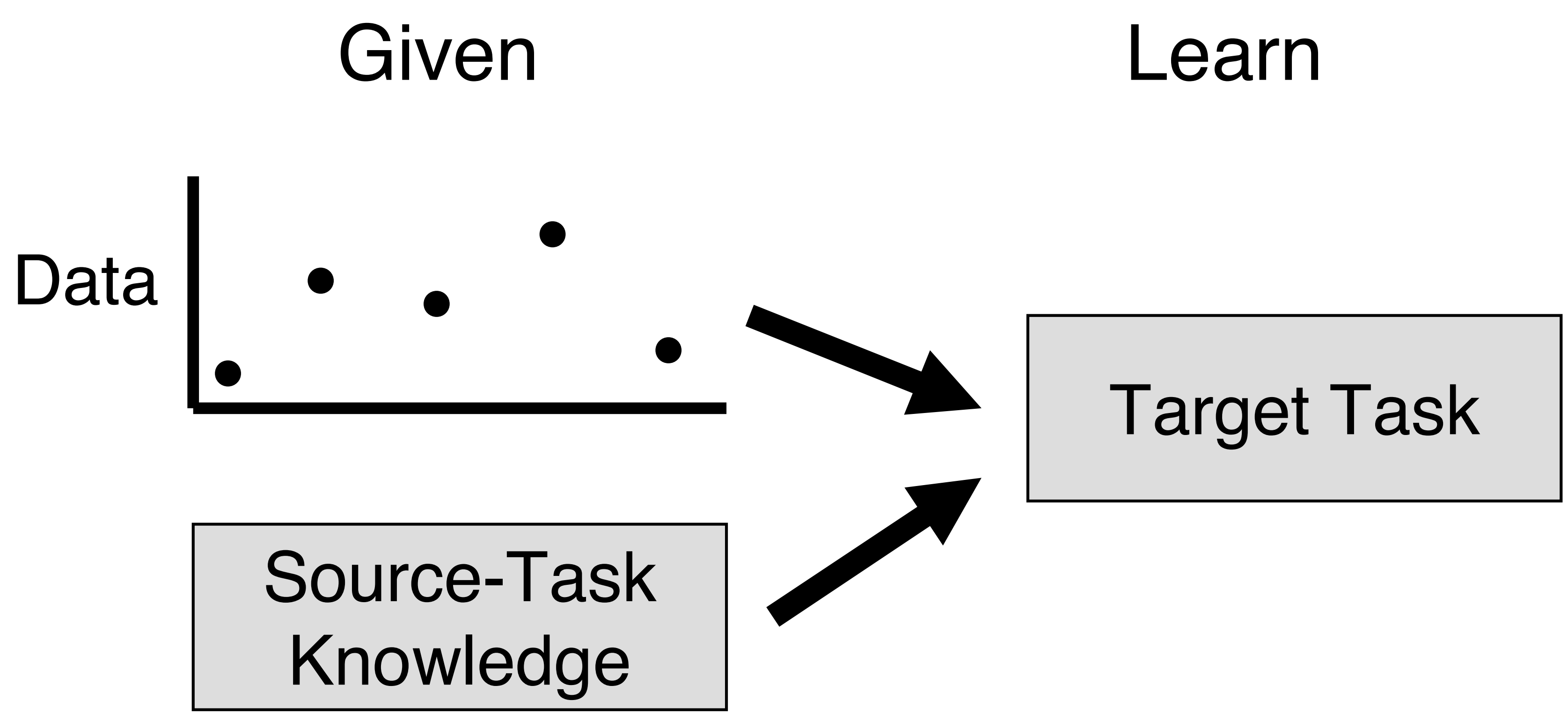




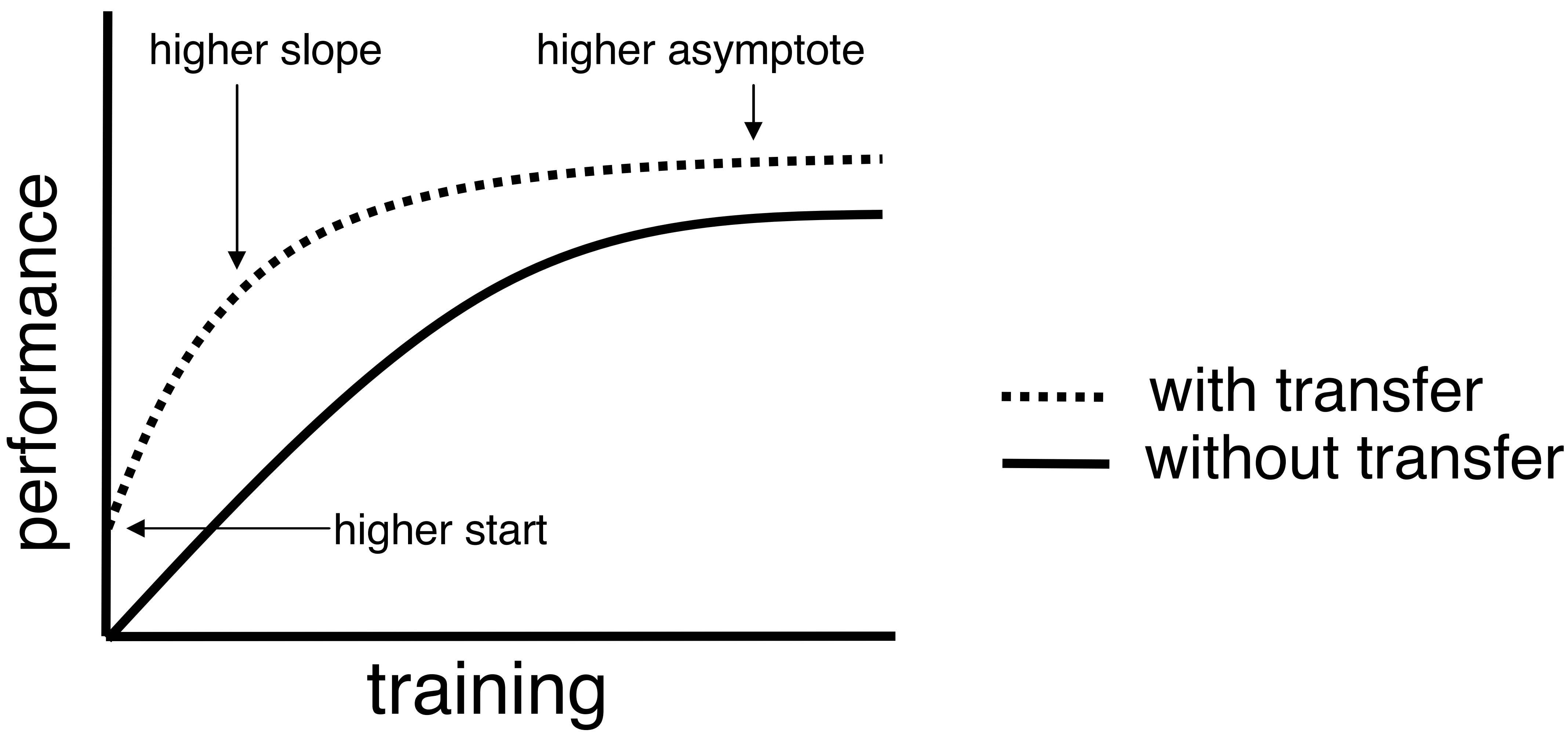
How can machines recognize and apply relavent knowledge from previous learning experience?

Transfer Learning at 1000 feet

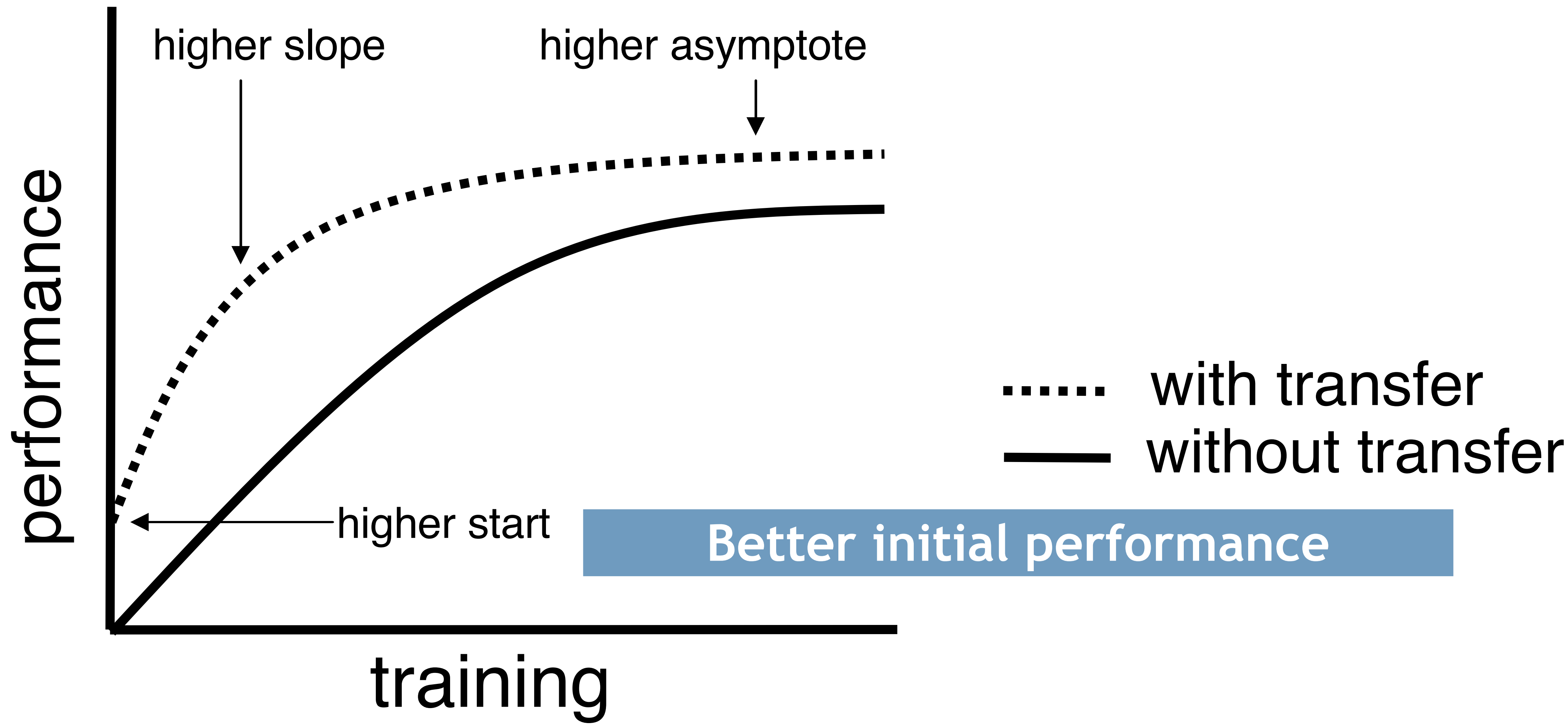
- Transfer knowledge from one or more source tasks or domains to a target domain or task.



How transfer might improve target learnring

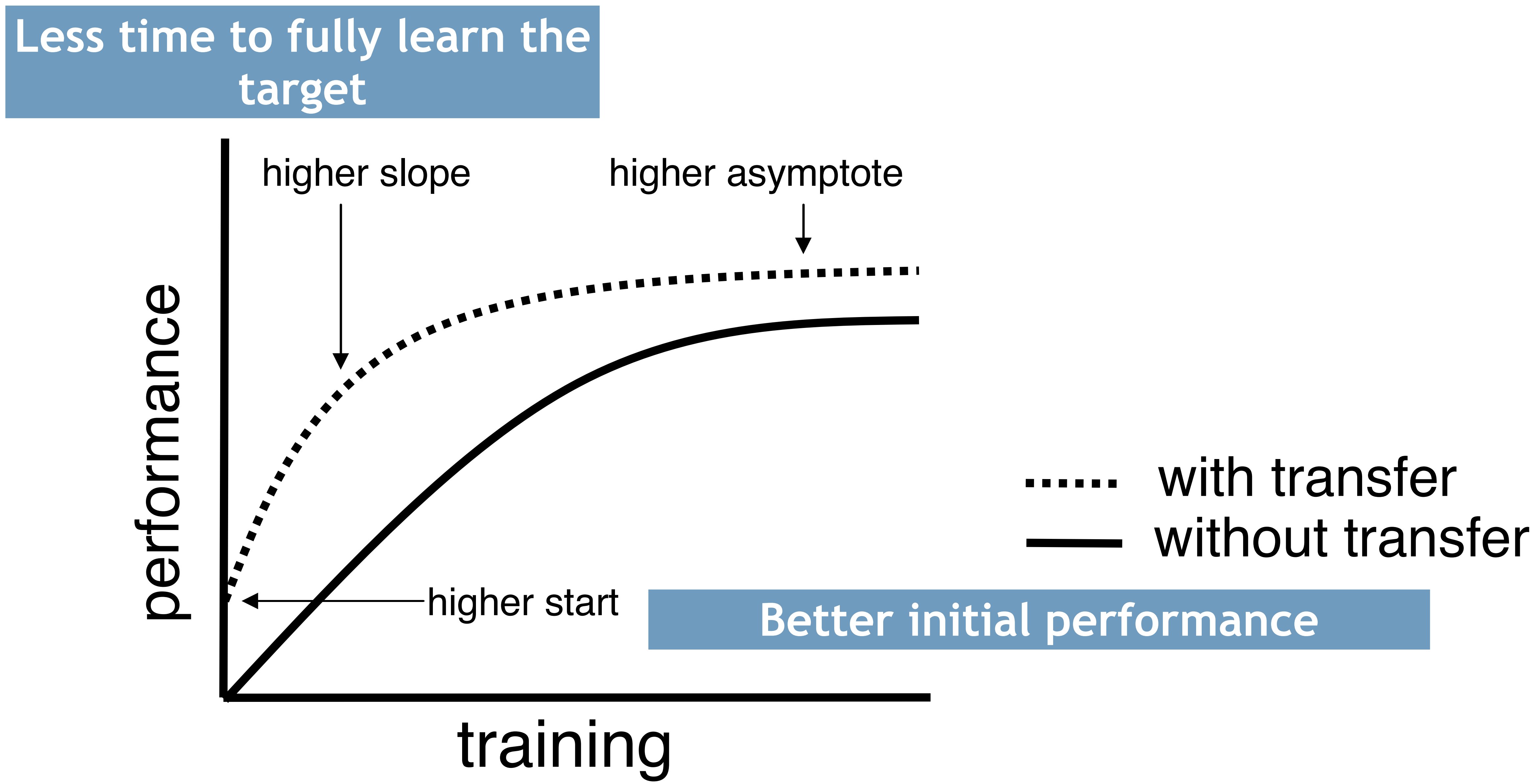


How transfer might improve target learnring



Better initial performance

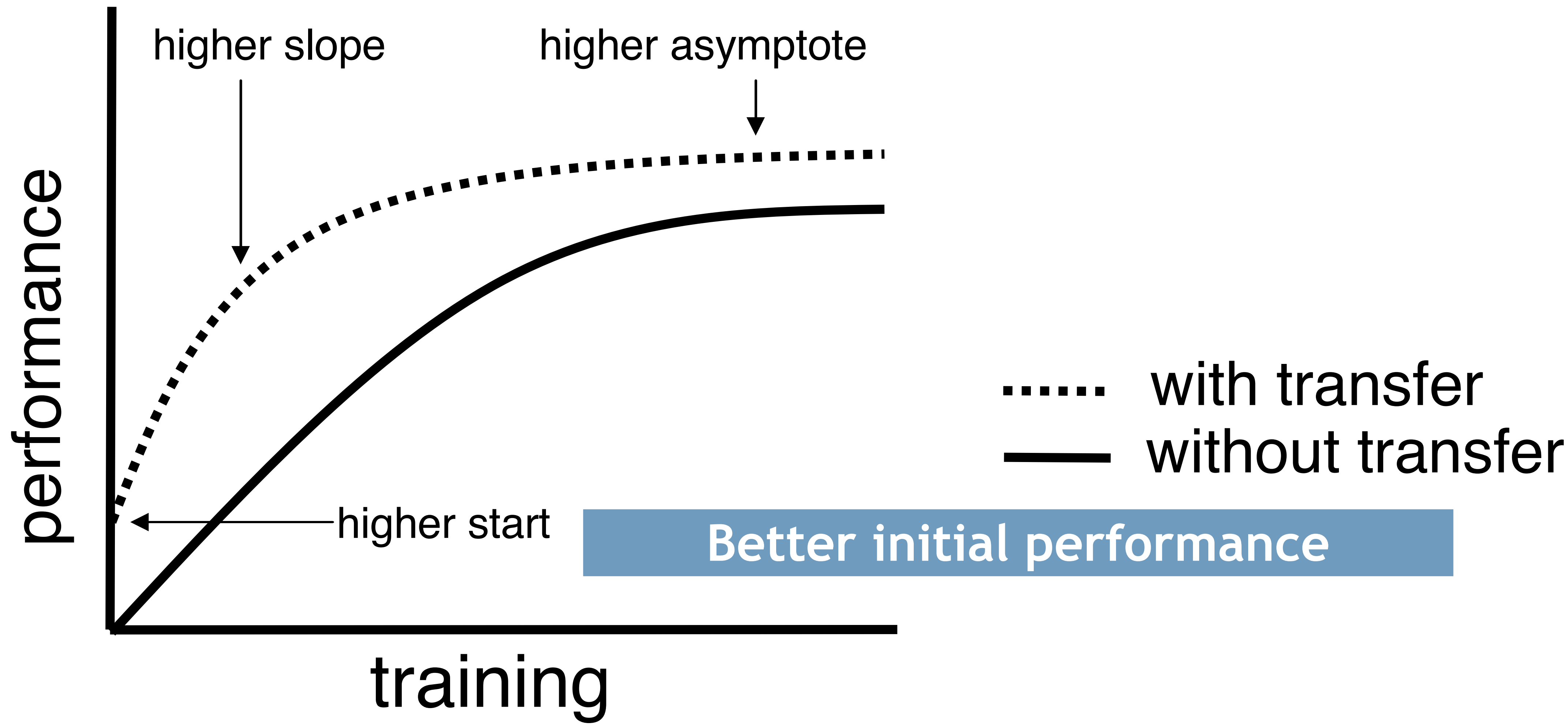
How transfer might improve target learnring



How transfer might improve target learnring

Less time to fully learn the target

Better final (asymptotic) performance

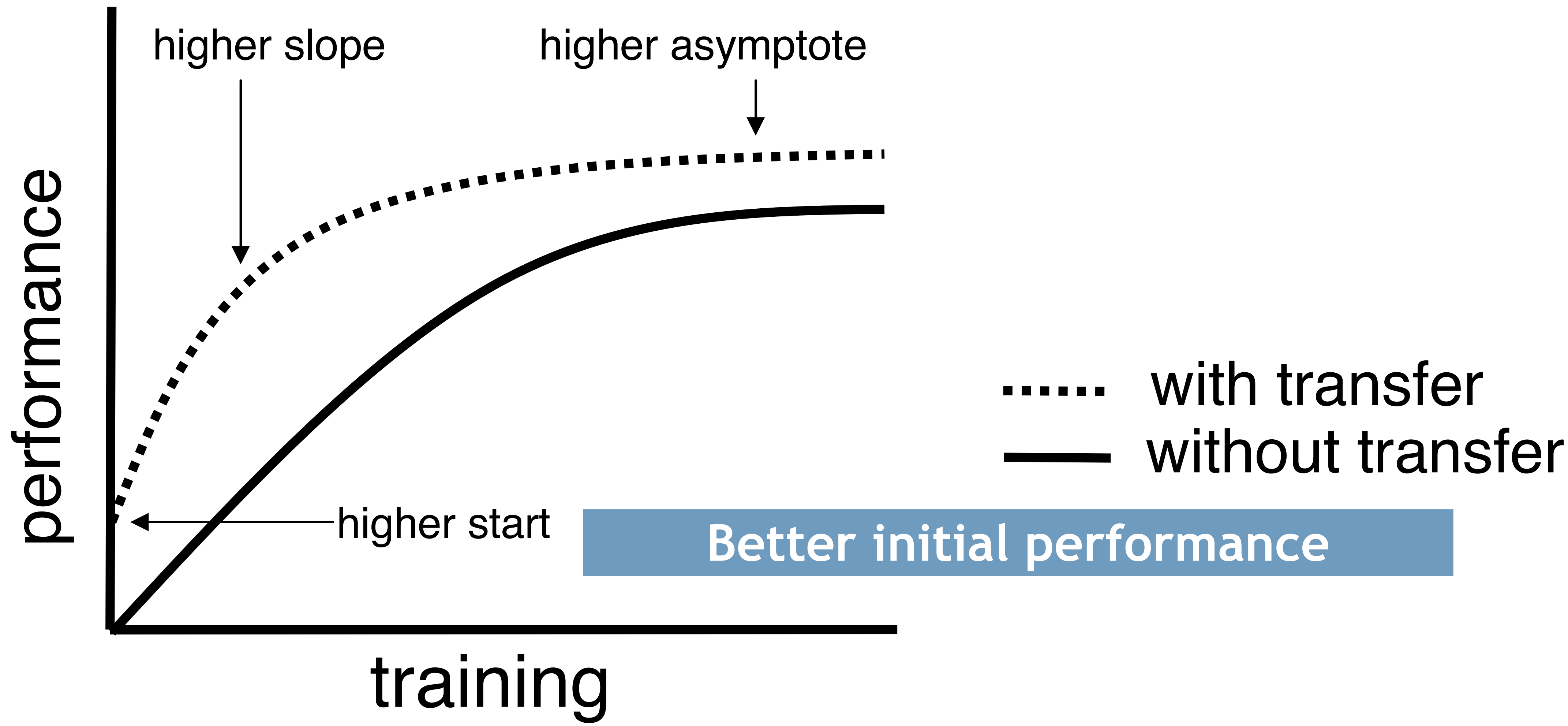


Better initial performance

How transfer might improve target learnring

Less time to fully learn the target

Better final (asymptotic) performance



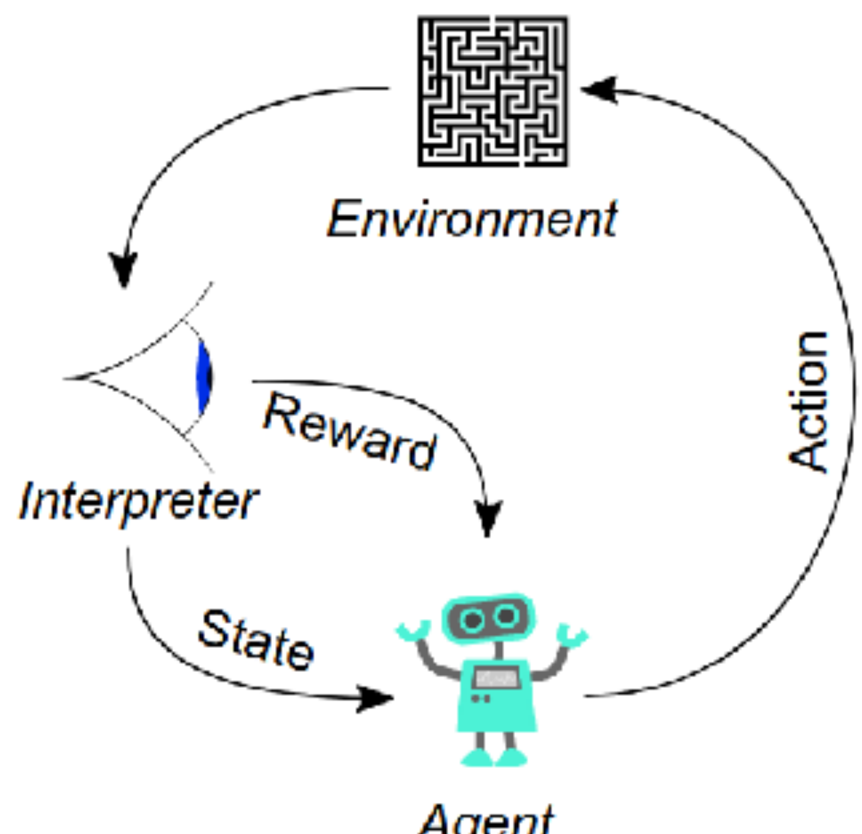
Better initial performance

Transfering might reduce target learning performance (negative transfer)

Two Branches of Transfer Learning Paradigms

Inductive Learning: Learn decision function f from training data, test on unseen data

Reinforcement Learning: sequential decision making problems



Inductive Transfer Learning Examples

- Domain-specific computer vision tasks
- Common to transfer pre-trained features from ImageNet

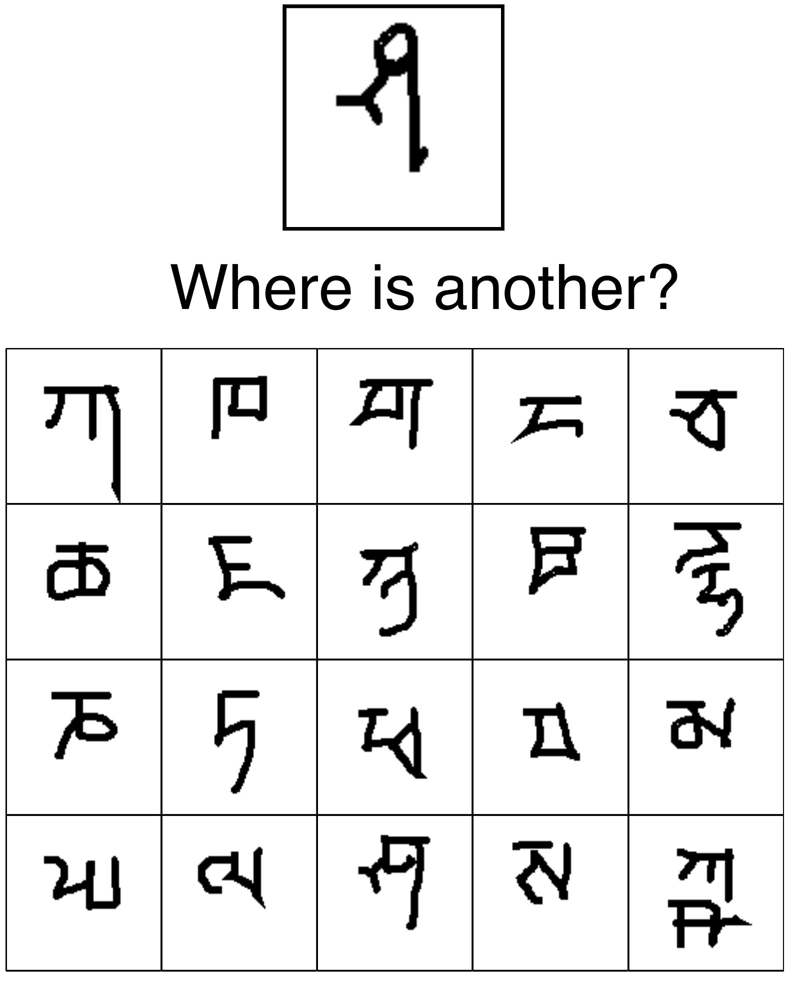
ImageNet 1000-class classification task

Structural Damage Detection

Yuqing Zhao et. al. Deep Transfer Learning for Image-Based Structural Damage Recognition

Learning with Small Samples: K-Shot Learning

- When the training set of a task only has k samples
- e.g. one-shot alphabet classification:



OMIGLOT dataset

Learning with Small Samples: K-Shot Learning

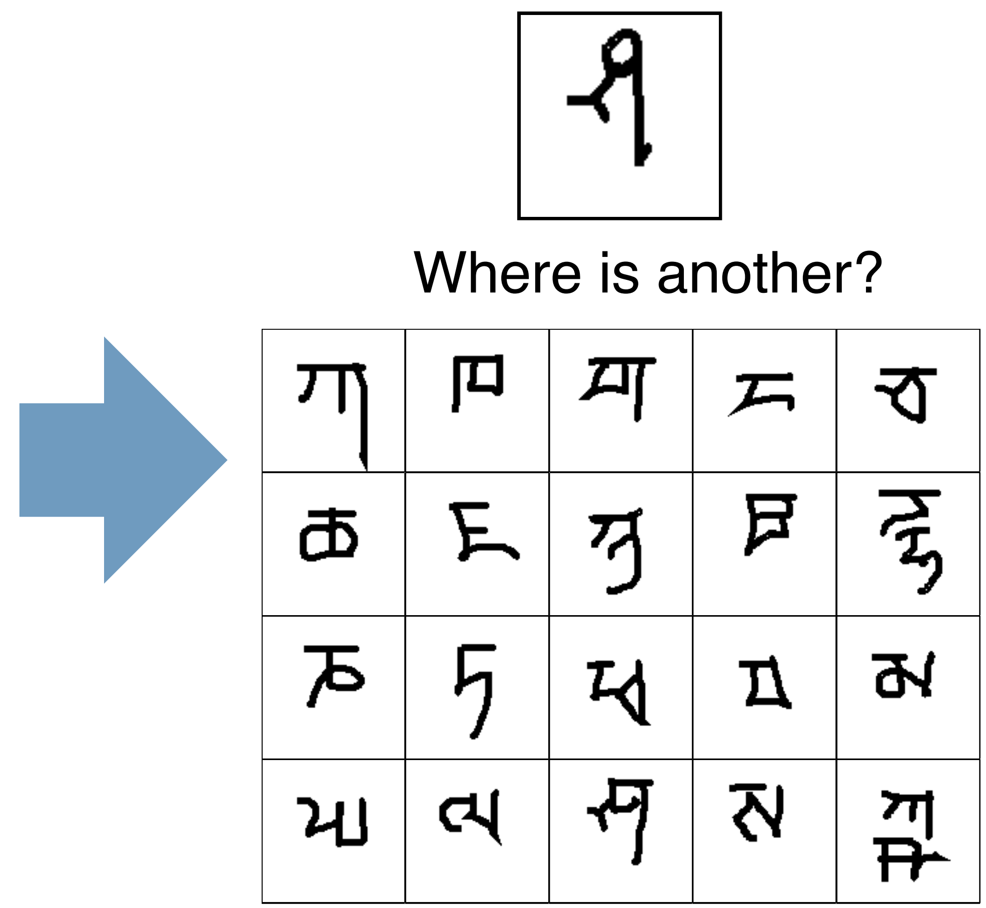
- When the training set of a task only has k samples
- e.g. one-shot alphabet classification:

OMIGLOT dataset

K-Shot Learning

- **Transfer latent knowledge of handwritten characters from other tasks**

50 classification tasks in different alphabets



K-Shot Learning

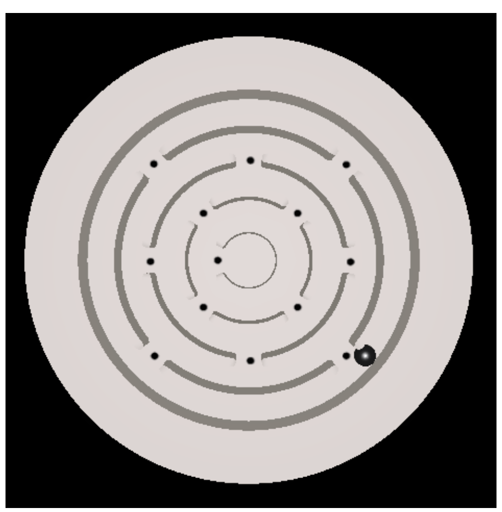
- One-shot person re-identification from video

Key Idea: Transfer knowledge from multiple domains (datasets)

Bak et. al. (2017) One-Shot Metric Learning for Person Re-identification

Reinforcement Transfer Learning Examples

- **Reinforcement learning for robotic control, e.g**
  - SIM2Real： transfer learned policy/value function from simulated robot to physical robot

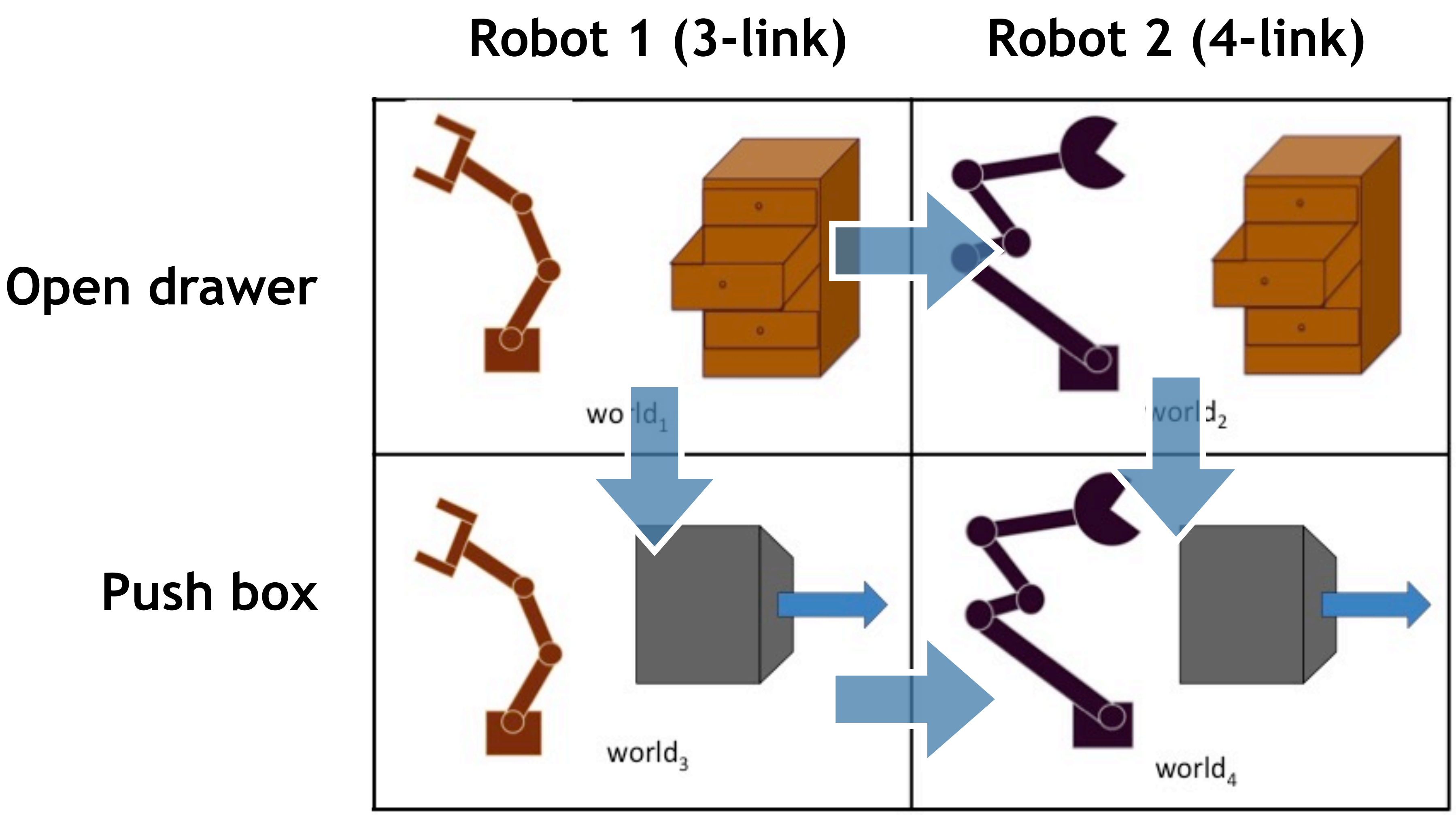


Simulated marble maze game

Real maze on robotic arm

Applications of Transfer Learning

- **Reinforcement learning for robotic control, e.g**
  - Transfer between robots and between tasks

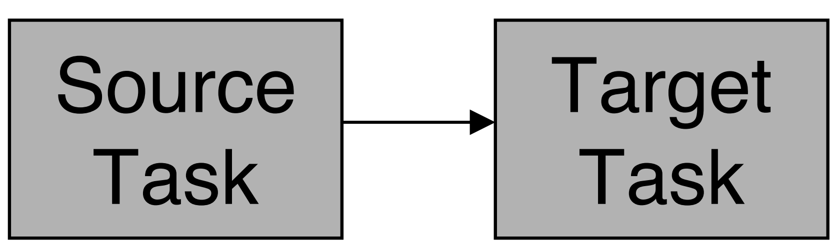


Devin (2016) Learning Modular Neural Network Policies for Multi-Task Multi-Robot Transfer

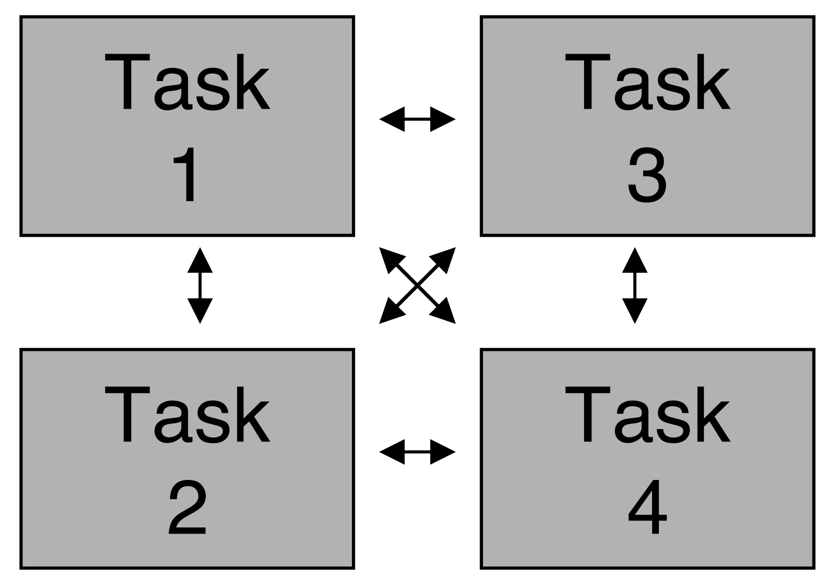
Transfer Learning vs Multi-Task Learning

TL is more likely to encounter in real world than MTL

Transfer Learning



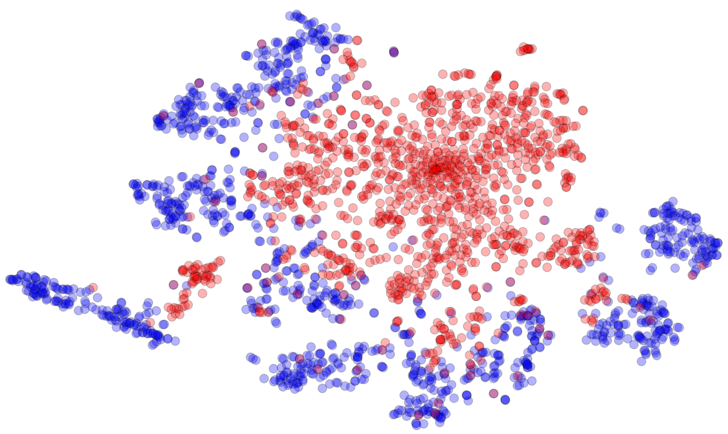
Multi-task Learning



TL: Source task is learned without knowledge of any target tasks

Outline

- What's Transfer Learning
- Traditional transfer learning algorithms
    - Task transfer learning
    - Domain adaptation
    - Transfer bound on domain adaptation
- When to transfer?
    - Transferability estimation
- Research trends



Transfer Learning Definition

Terminologies

- Domain: $D = \{X, P_X\}$
- Task: $T = \{Y, f\}$

Transfer Learning Definition

Terminologies

input features

- Domain: $D = \{X, P_X\}$
- Task: $T = \{Y, f\}$

Transfer Learning Definition

Terminologies

input features

- Domain: $D = \{X, P_X\}$

input distribution

- Task: $T = \{Y, f\}$

Transfer Learning Definition

Terminologies

input features

- Domain: $D = \{X, P_X\}$

input distribution

- Task: $T = \{Y, f\}$

labels

Transfer Learning Definition

Terminologies

- Domain: $D = \{X, P_X\}$
- Task: $T = \{Y, f\}$

input features

input distribution

labels

predictive function $P_{Y|X}$

Transfer Learning Definition

Terminologies

input features

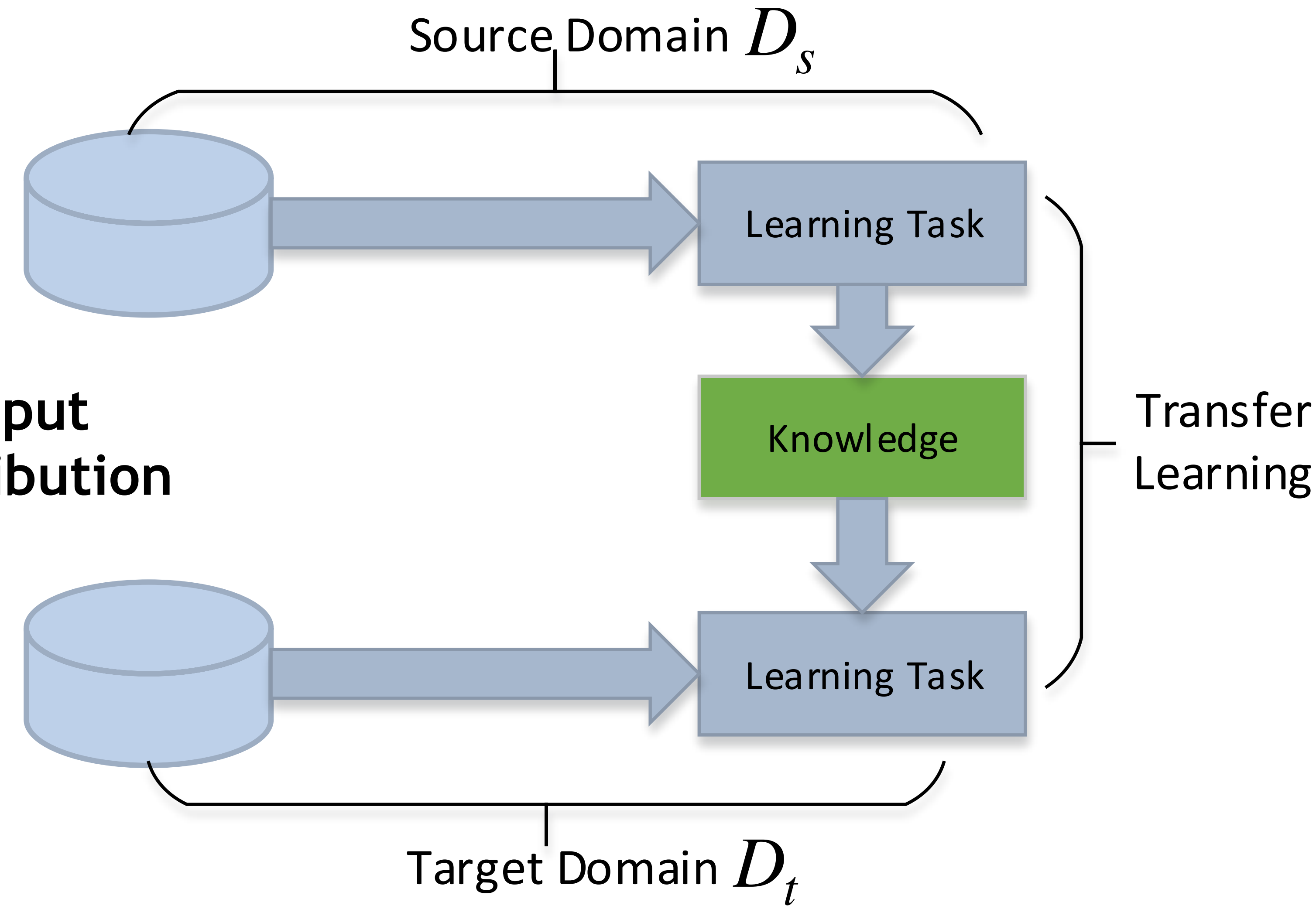
- Domain: $D = \{X, P_X\}$

input distribution

- Task: $T = \{Y, f\}$

labels

predictive function $P_{Y|X}$



Transfer Learning Definition

Terminologies

input features

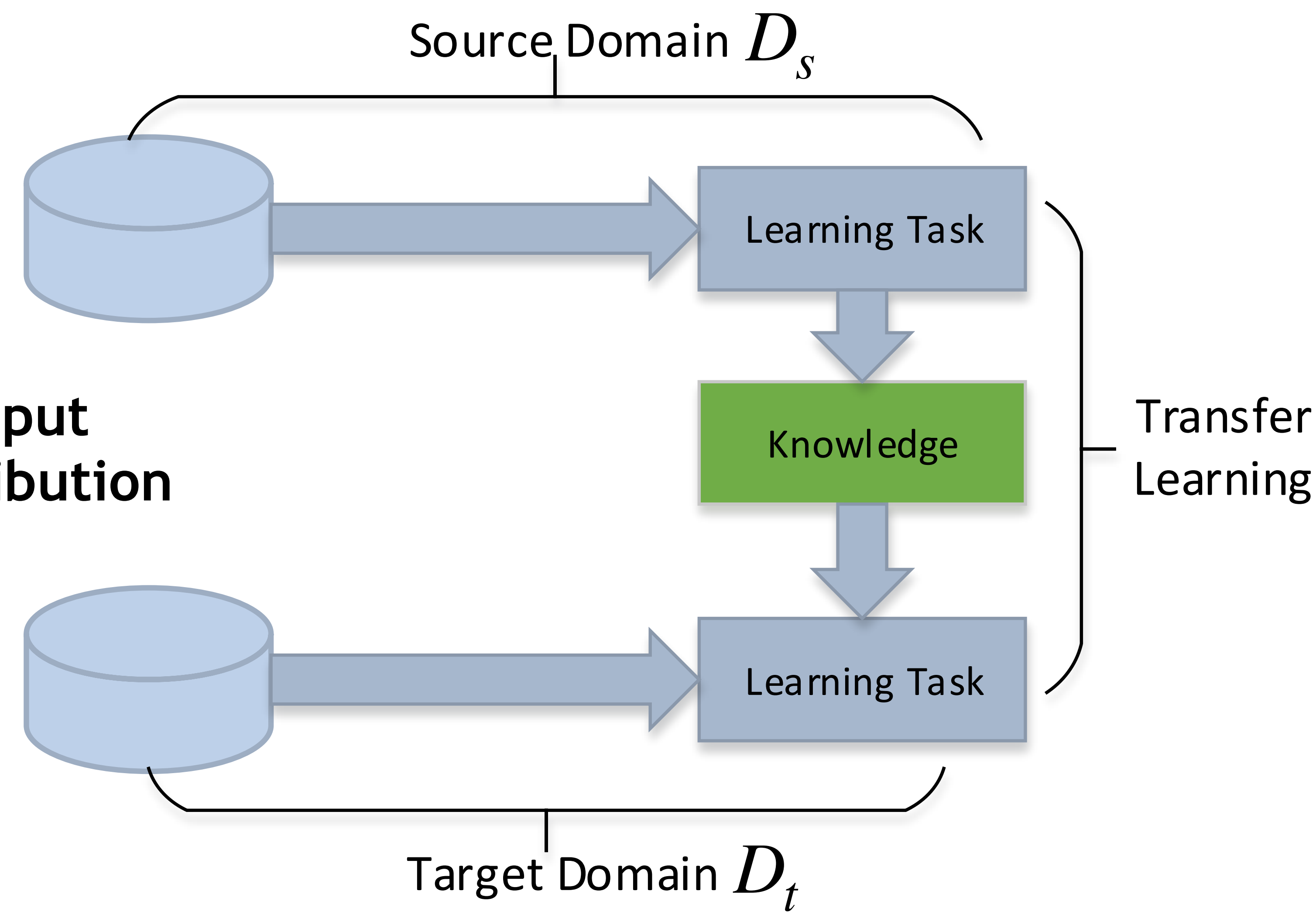
- Domain: $D = \{X, P_X\}$

input distribution

- Task: $T = \{Y, f\}$

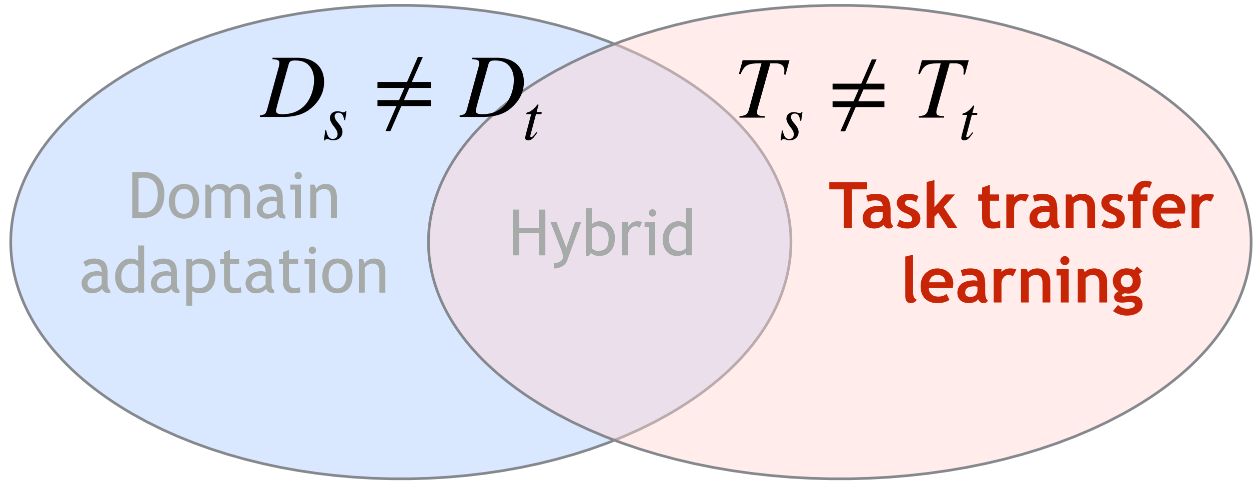
labels

predictive function $P_{Y|X}$



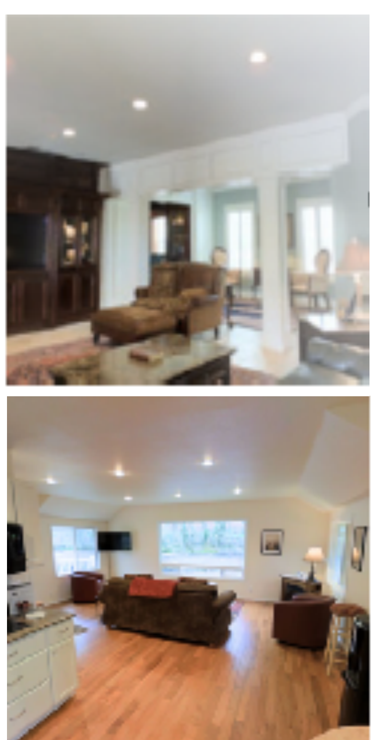
Transfer learning: improve the performance of predictive function f_t for T_t by discover and transfer latent knowledge from (D_s, T_s) , where $D_s \neq D_t$ and/or $T_s \neq T_t$

Transfer Learning



Task Transfer Learning: adapt source hypothesis or feature to target task

D_s/D_t: Indoor Scene



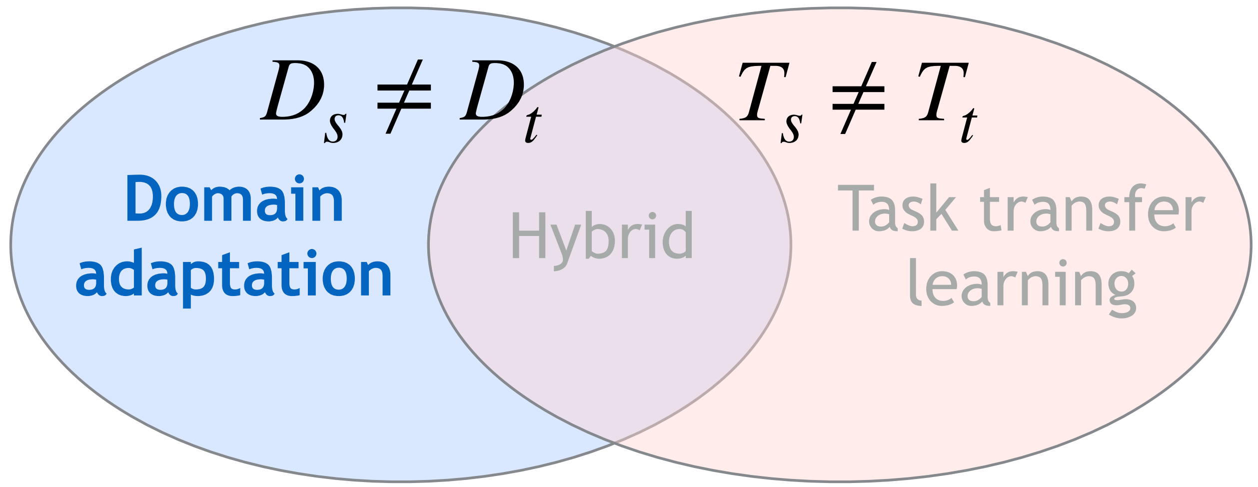
T_s: scene classification →

living room

T_t: object detection →

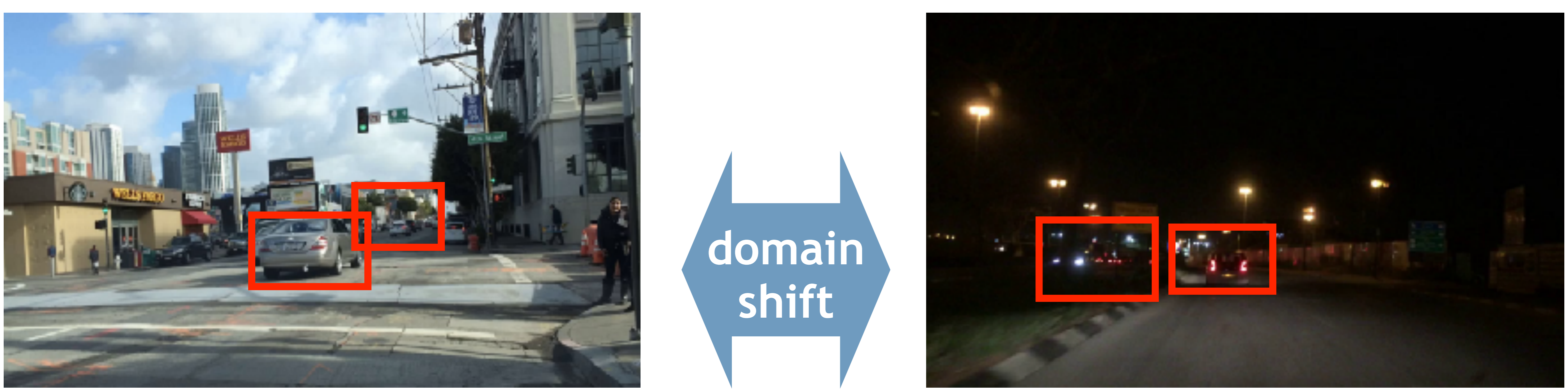
sofa, table, lamp, …

Transfer Learning



Domain adaptation: Learn domain agnostic representations

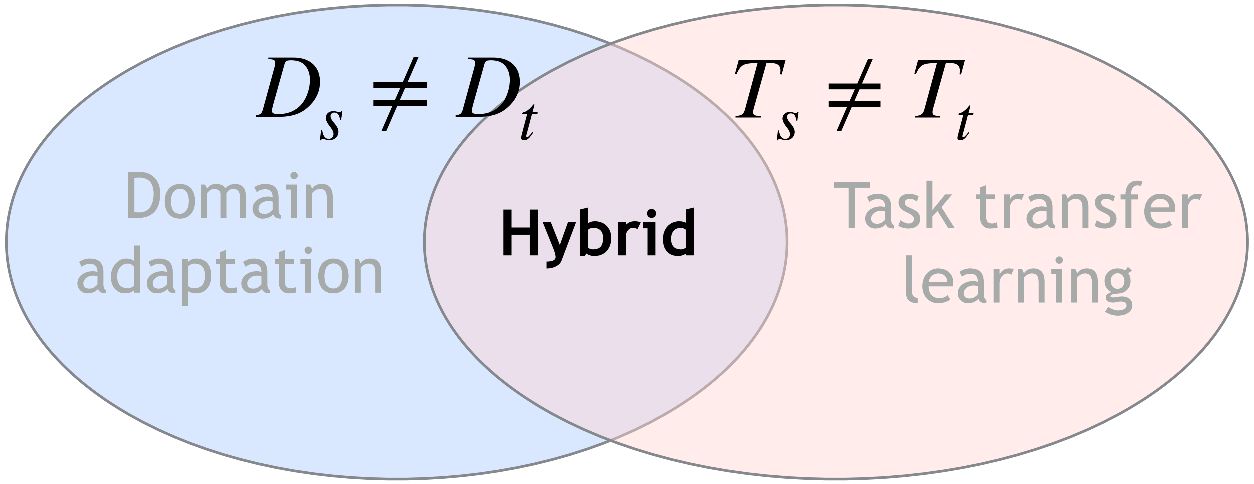
T_s/T_t: Vehicle Detection



D_s (day)

D_t (night)

Transfer Learning



Task Transfer Learning: adapt source hypothesis or feature to target task

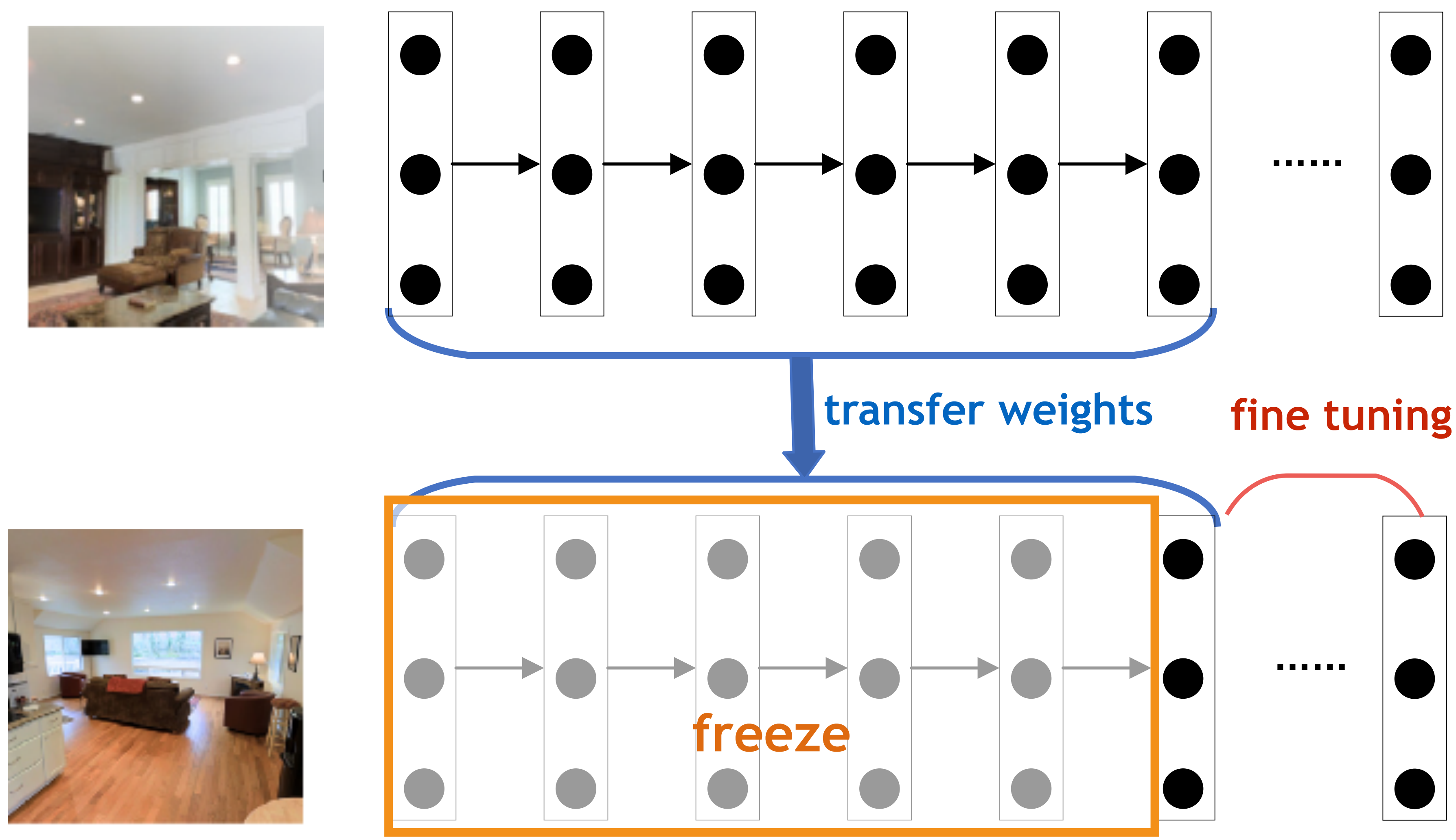
Domain adaptation: Learn domain agnostic representations

Most transfer learning problems in practice are hybrid!

Task Transfer Learning

- Pre-trained Model + Fine Tuning

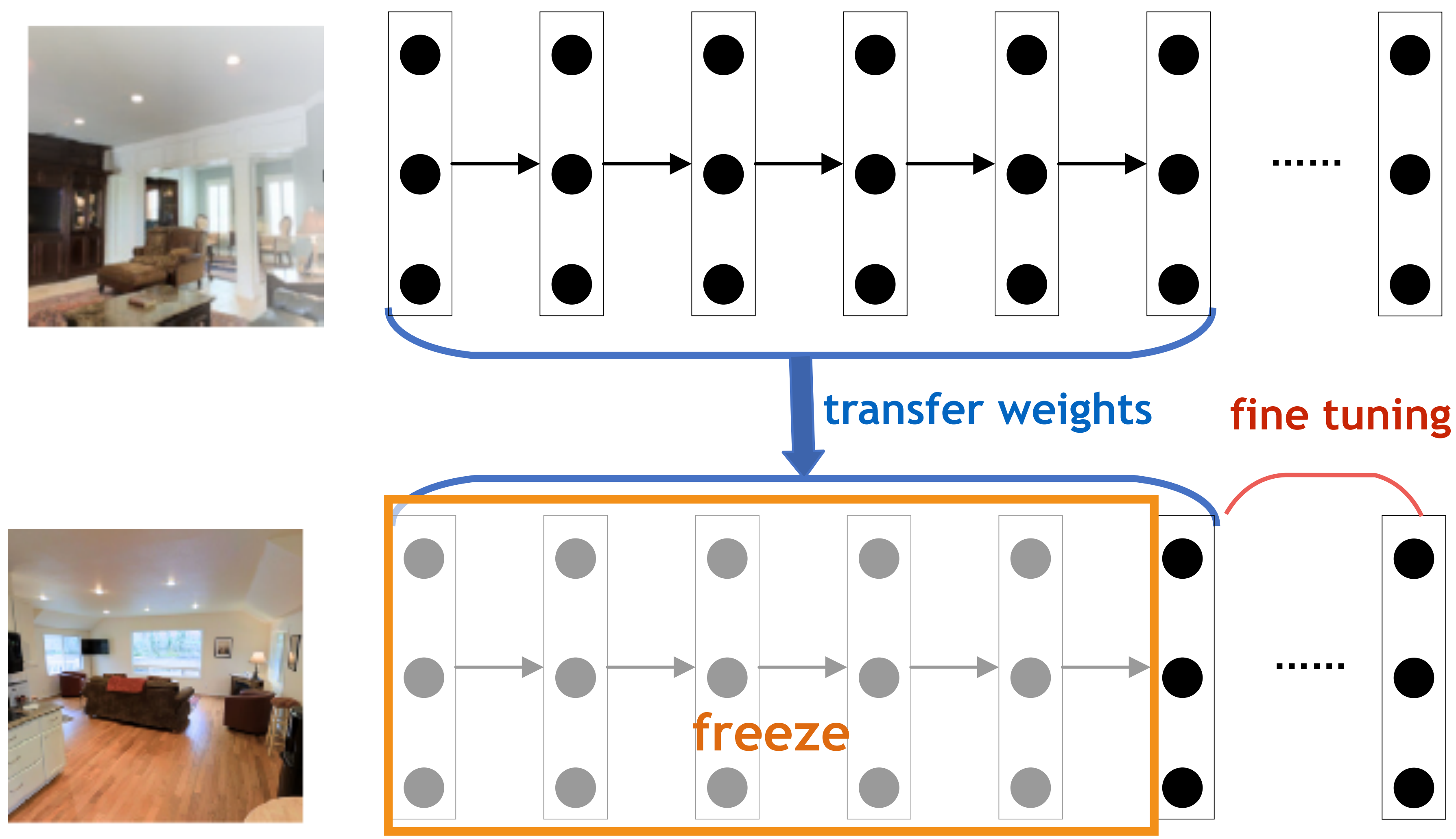
e.g object classification -> scene classification



Task Transfer Learning

- Pre-trained Model + Fine Tuning

e.g object classification -> scene classification

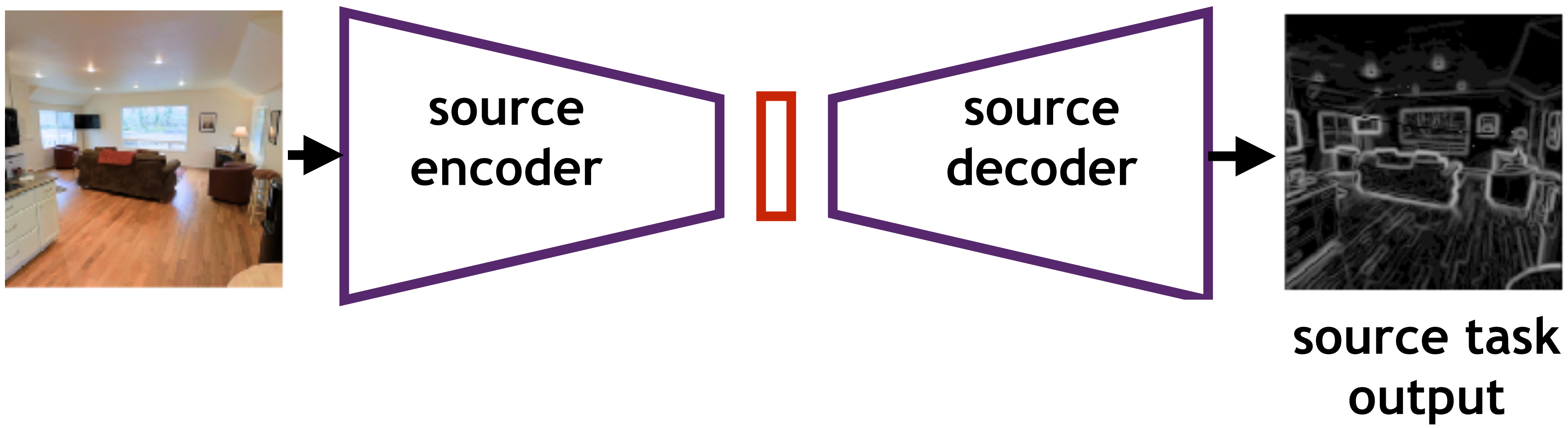


intuition: low level features are shared across most vision tasks

Heterogeneous Task Transfer Learning

- Heterogeneous task transfer learning using **encoder-decoder** network

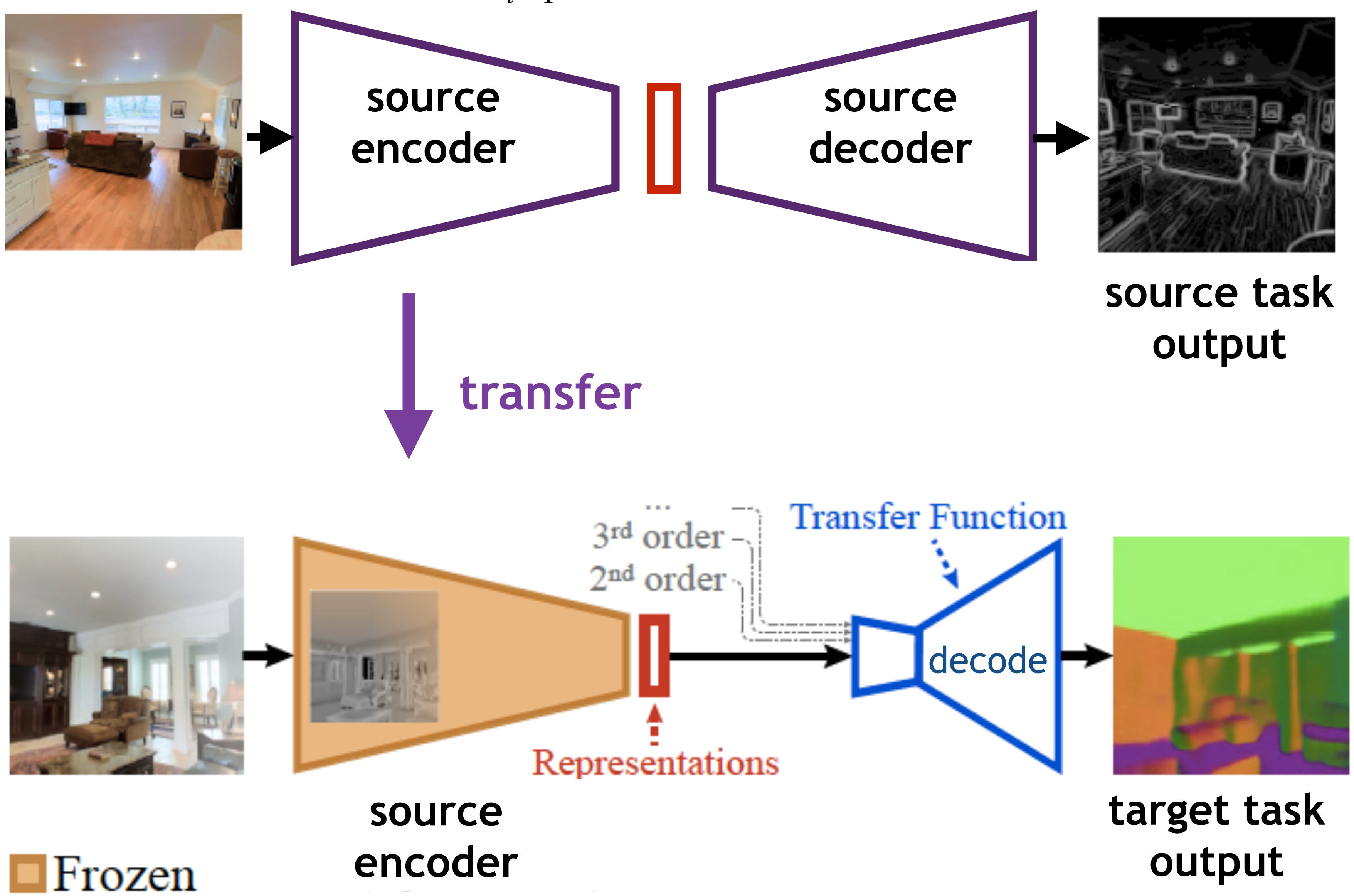
$$L = \sum_{i=1}^{N} ||y_s^i - D_s(E_s(x_s^i))||^2$$



Heterogeneous Task Transfer Learning

- Heterogeneous task transfer learning using **encoder-decoder** network

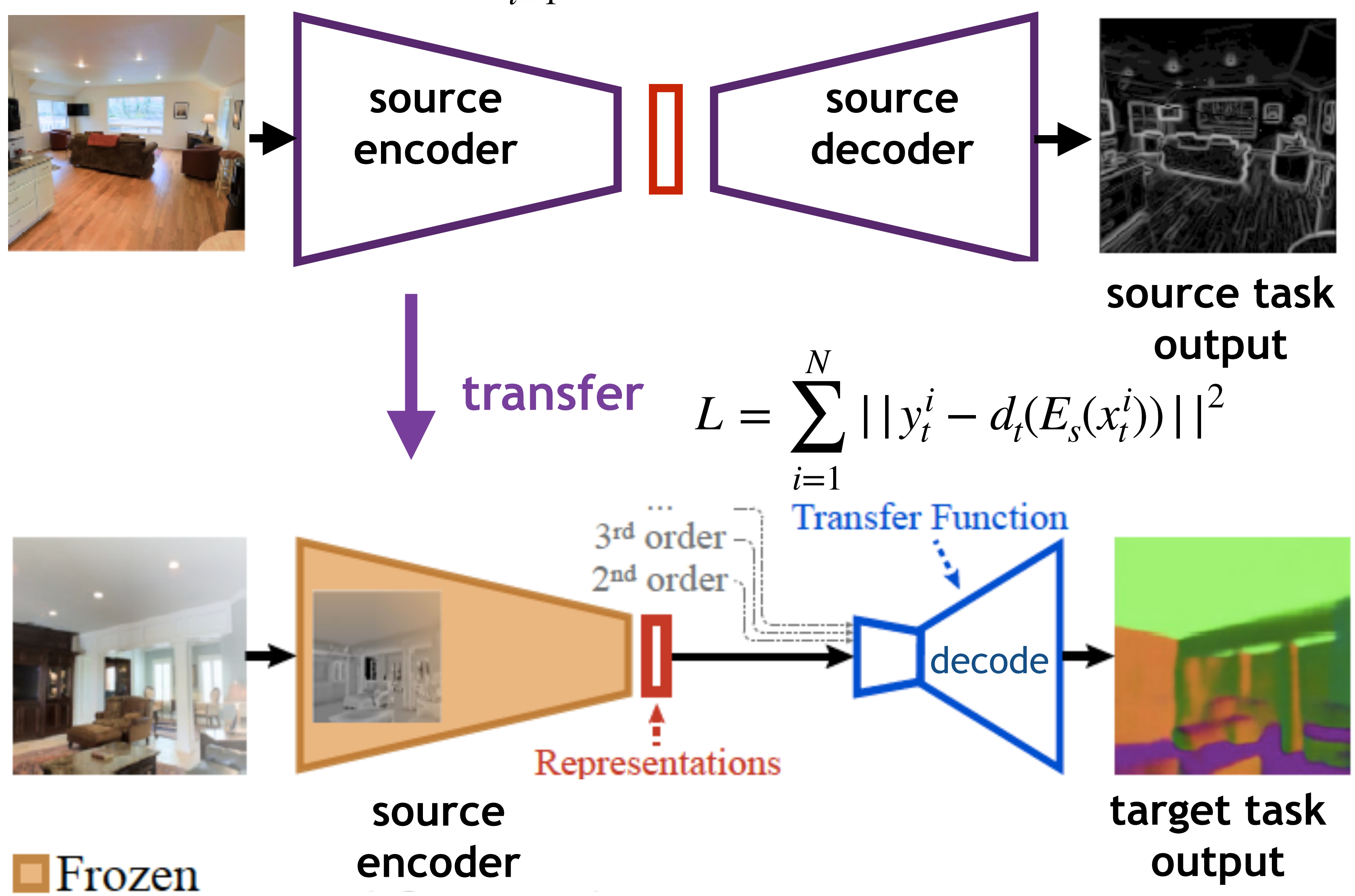
$$L = \sum_{i=1}^{N} ||y_s^i - D_s(E_s(x_s^i))||^2$$



Heterogeneous Task Transfer Learning

- Heterogeneous task transfer learning using **encoder-decoder** network

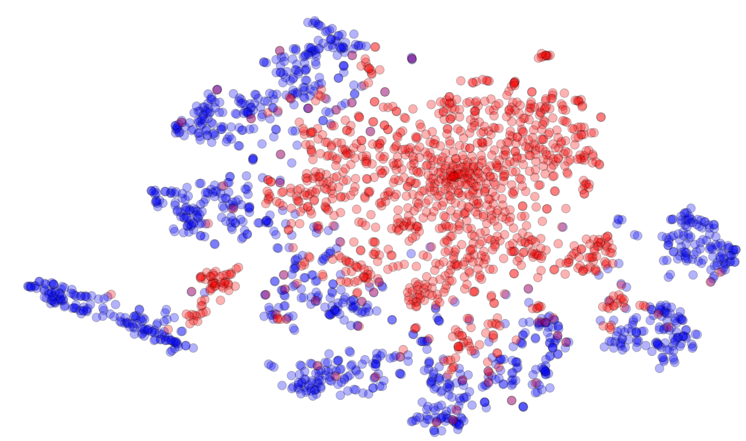
$$L = \sum_{i=1}^{N} ||y_s^i - D_s(E_s(x_s^i))||^2$$



$$L = \sum_{i=1}^{N} ||y_t^i - d_t(E_s(x_t^i))||^2$$

Outline

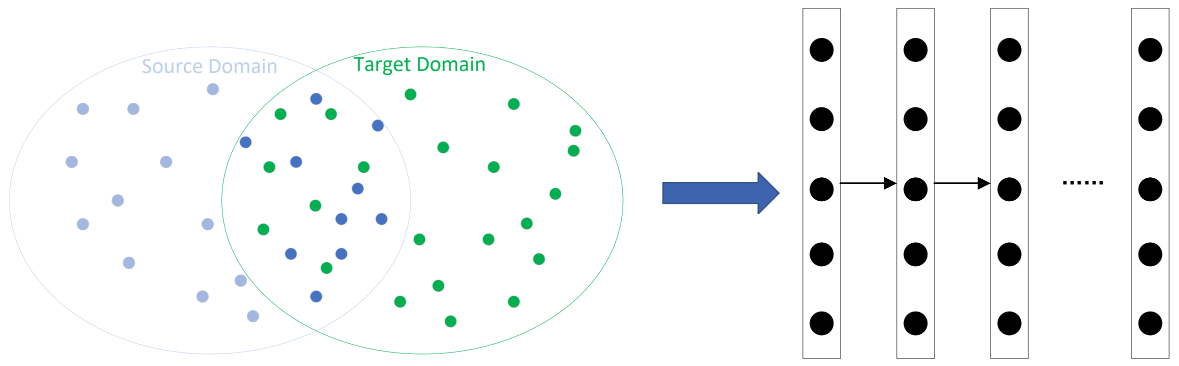
- What's Transfer Learning
- Traditional transfer learning algorithms
  - Task transfer learning
  - Domain adaptation
  - Transfer bound on domain adaptation



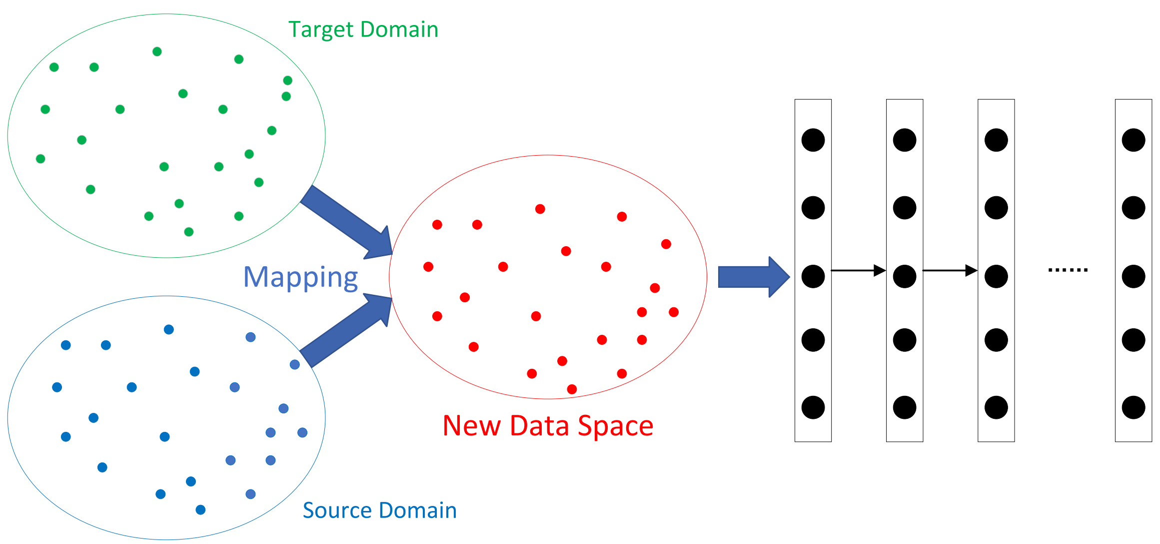
- When to transfer?
  - Transferability estimation
- Research trends

Domain Adaptation Techniques

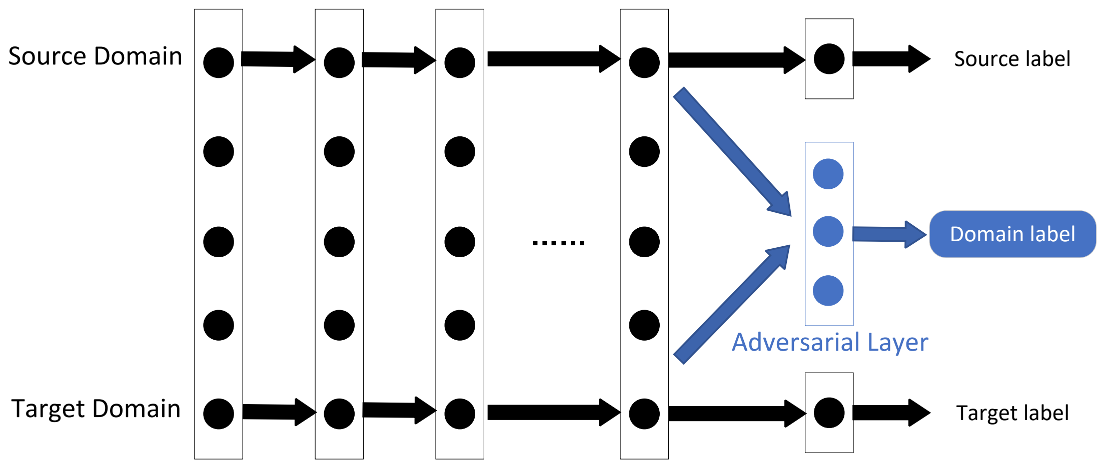
- Instance-based approach



- Mapping-based approach

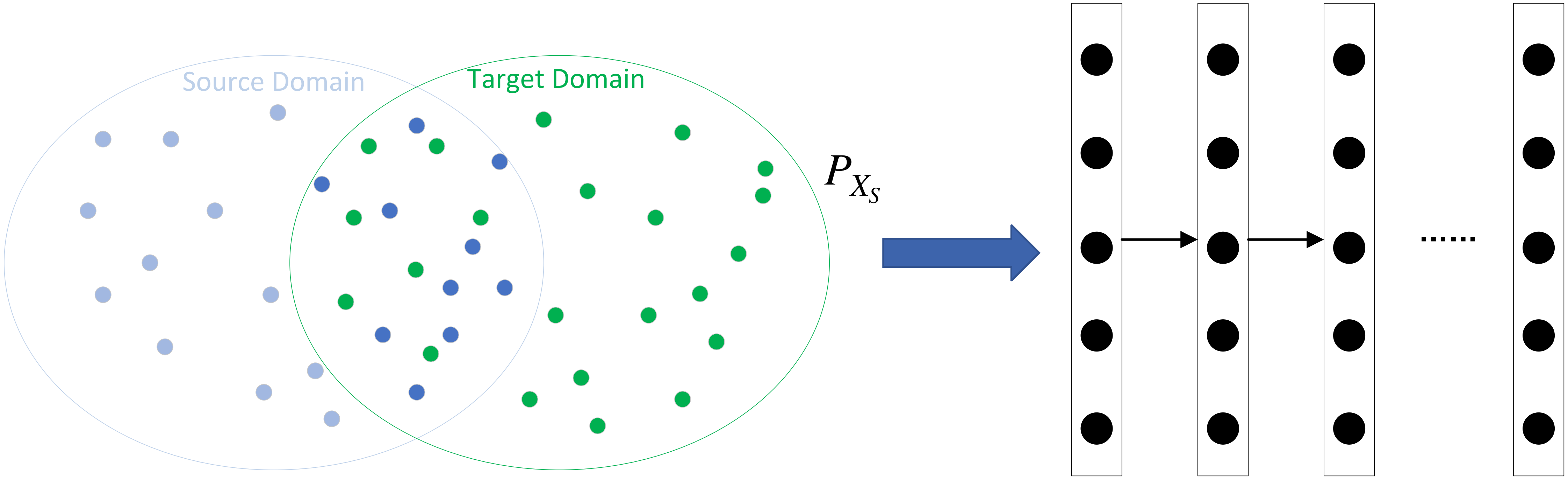


- Adversarial-based approach



Instance-based approaches

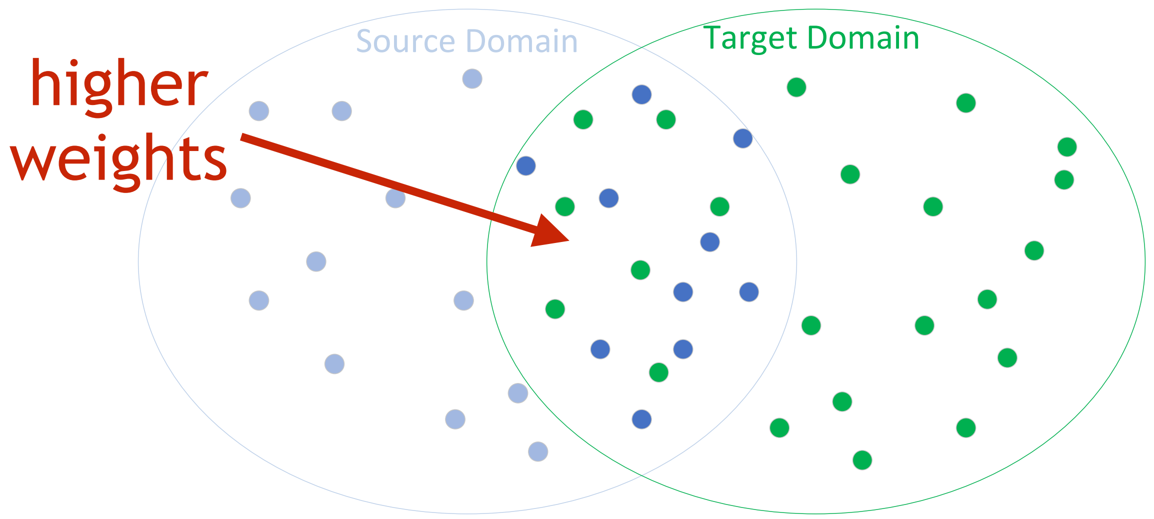
- select **partial instances** from the source domain as supplements to the training set in the target domain



Partial instances in the source domain can be utilized by the target domain with appropriate weights

Boosting for instance-based transfer

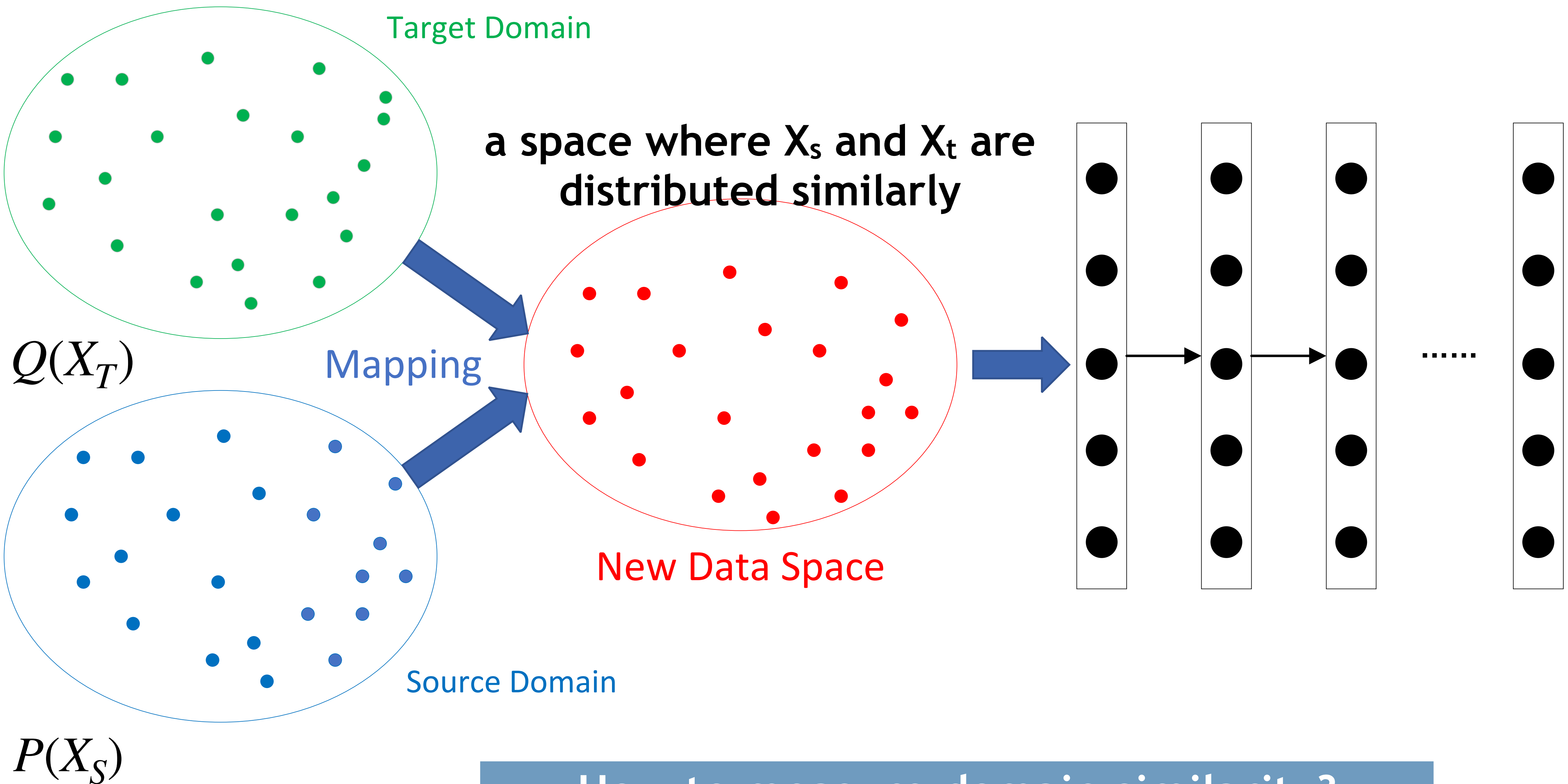
- TrAdaBoost (Dai 2007)
  - Use AdaBoost to filter out source domain instances that are dissimilar to target domain
  - Reweight source domain instances to resemble target domain distribution
  - Train model with reweighted source + target domain instances



- TaskTrAdaBoost (2010): a boosting technique for transferring from multiple sources

Mapping-based approach

- Mapping instances from the source domain and target domain into a new data space



How to measure domain similarity?

Maximal Mean Discrepency (MMD)

- Maximal Mean Discrepency : a kernel-based 2 sample test for the null hypothesis P=Q (Fortet and Mourier, 1953)

$$D_{MMD}[P,Q] \triangleq \sup_{\phi \in \mathscr{F}} \left(\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(Y)]\right)$$

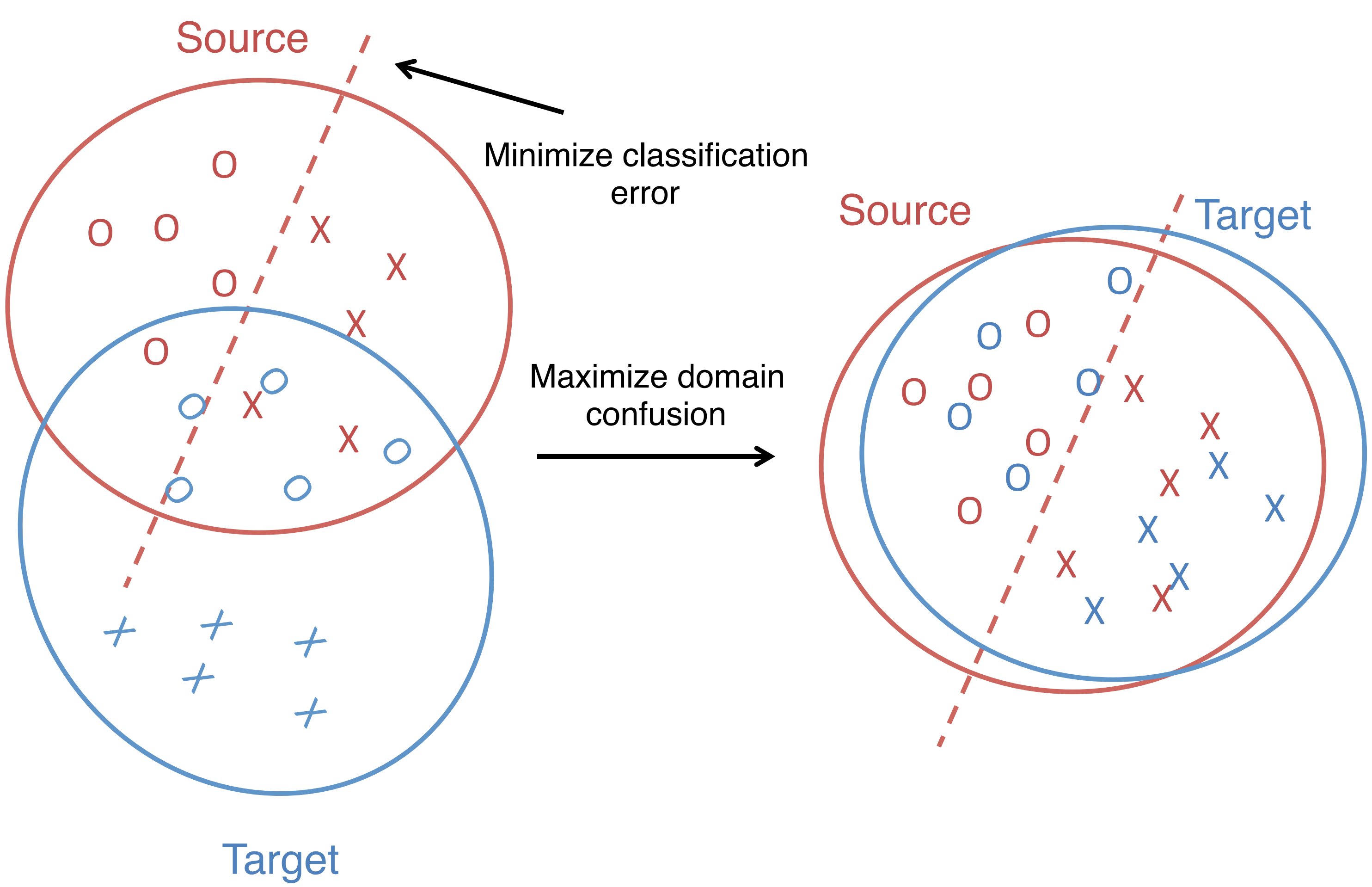
  - where $X \sim P, Y \sim Q$
  - feature map $\phi(\cdot)$
- Used in Transfer Component Analysis (TCA) (Yang, 2018) to correct domain shift

$$D_{MMD}(X_S, X_T) = \left\| \frac{1}{N_S} \sum_{x_s \in X_S} \phi(x_s) - \frac{1}{N_T} \sum_{x_t \in X_T} \phi(x_t) \right\|_{\mathscr{H}}$$

Use MMD as a Domain Regularization Term

- Given pre-trained source model, train an adpation network that minimizes classification error and domain MMD

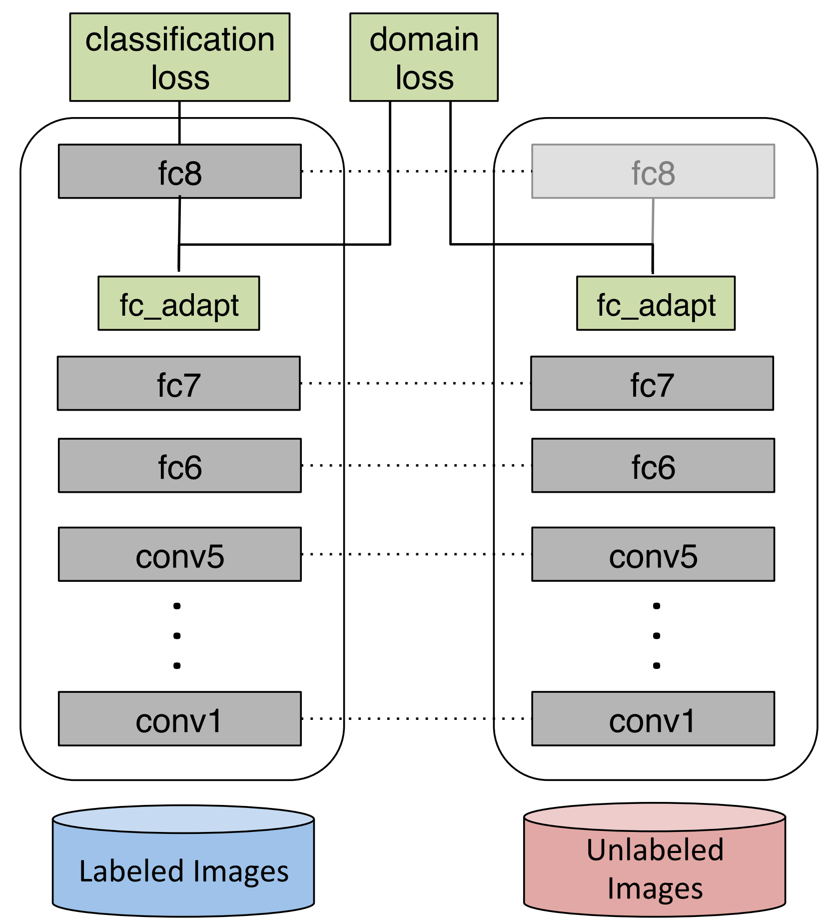
$$L = L_C(X_L, y) + \lambda D^2_{MMD}(X_S, X_T)$$



Tzeng et. al. Deep Domain Confusion: Maximizing for Domain Invariance

Use MMD as a Domain Regularization Term

- Training step:
  - 1. Select the layer to transfer from using MMD metric
  - 2. Train an adaptation layer f_a on source and target data using MMD as a regularizer
- Testing step:
  - Transform target input by $f_a(X_T)$

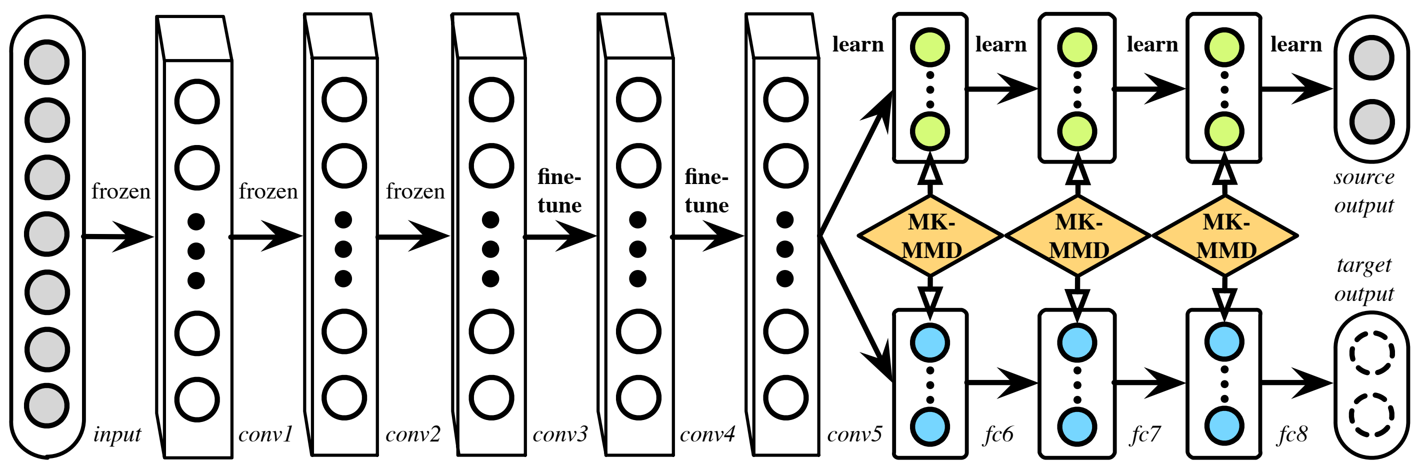


$$L = L_C(X_L, y) + \lambda D^2_{MMD}(X_S, X_T)$$

Tzeng et. al. Deep Domain Confusion: Maximizing for Domain Invariance

Variations with MMD-based domain adaptation

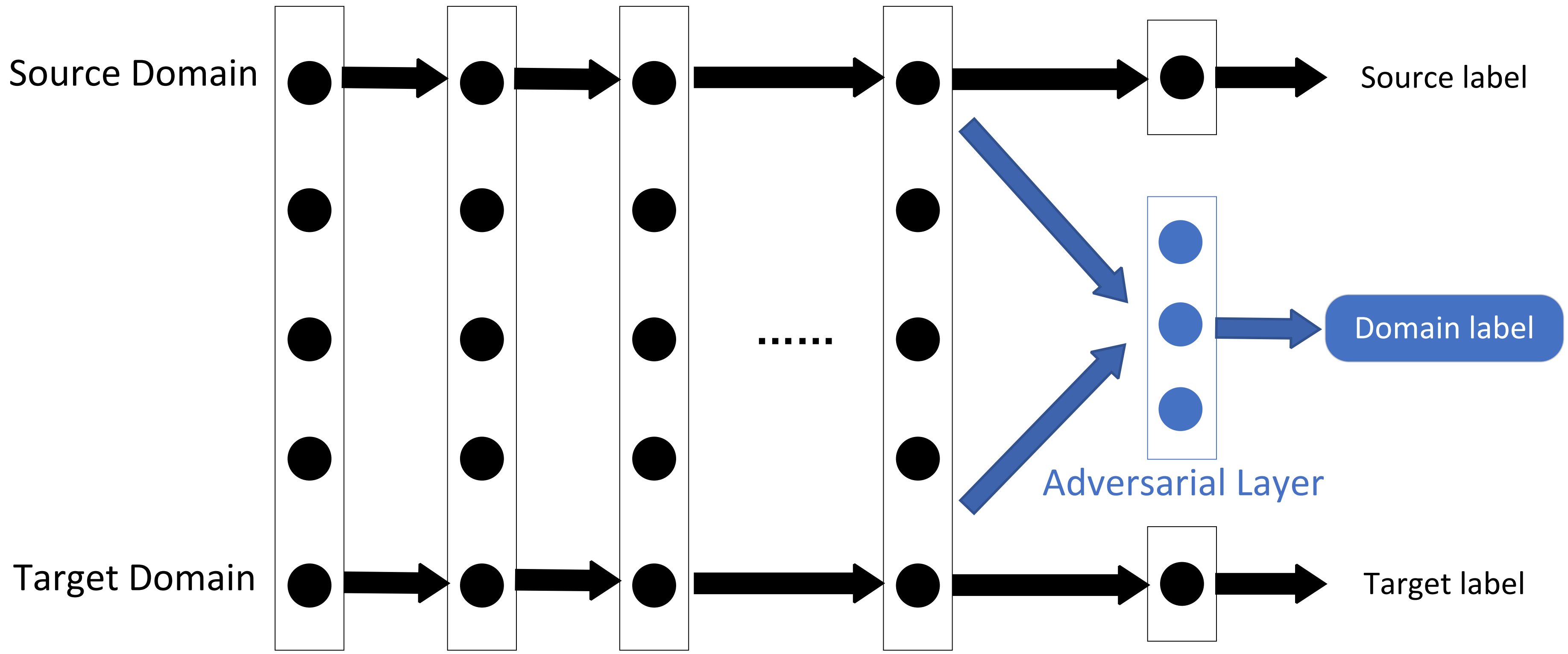
- Deep Adaptation Network (Long et.al. 2015):
  - Use multi-kernel MMD (MK-MMD)
    $$D_{MMD}[P, Q, K] \triangleq \|(\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(Y)])\|_{\mathscr{H}_K}$$
  - Fine-tune source task jointly with MMD constraints on multiple layers



- Joint Adaptation (2018): adapt joint distributions instead of $P(X_s)$, $Q(X_t)$

Adversarial-based approach

- Adopt adversarial training in learning transferable representation.

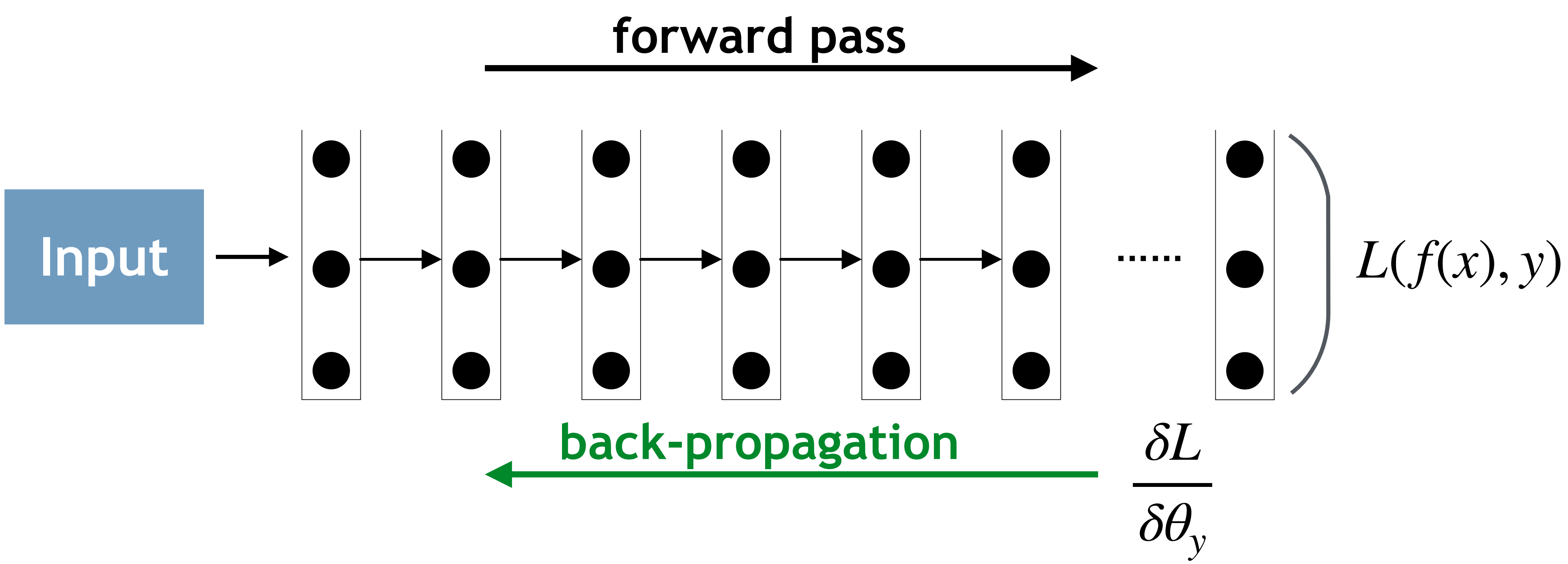


Effective features should be discriminative for the main learning task and indiscriminative between the source domain and target domain.

Adversarial-based approach

Ajakan et al. (2014) Domain-adversarial neural networks.

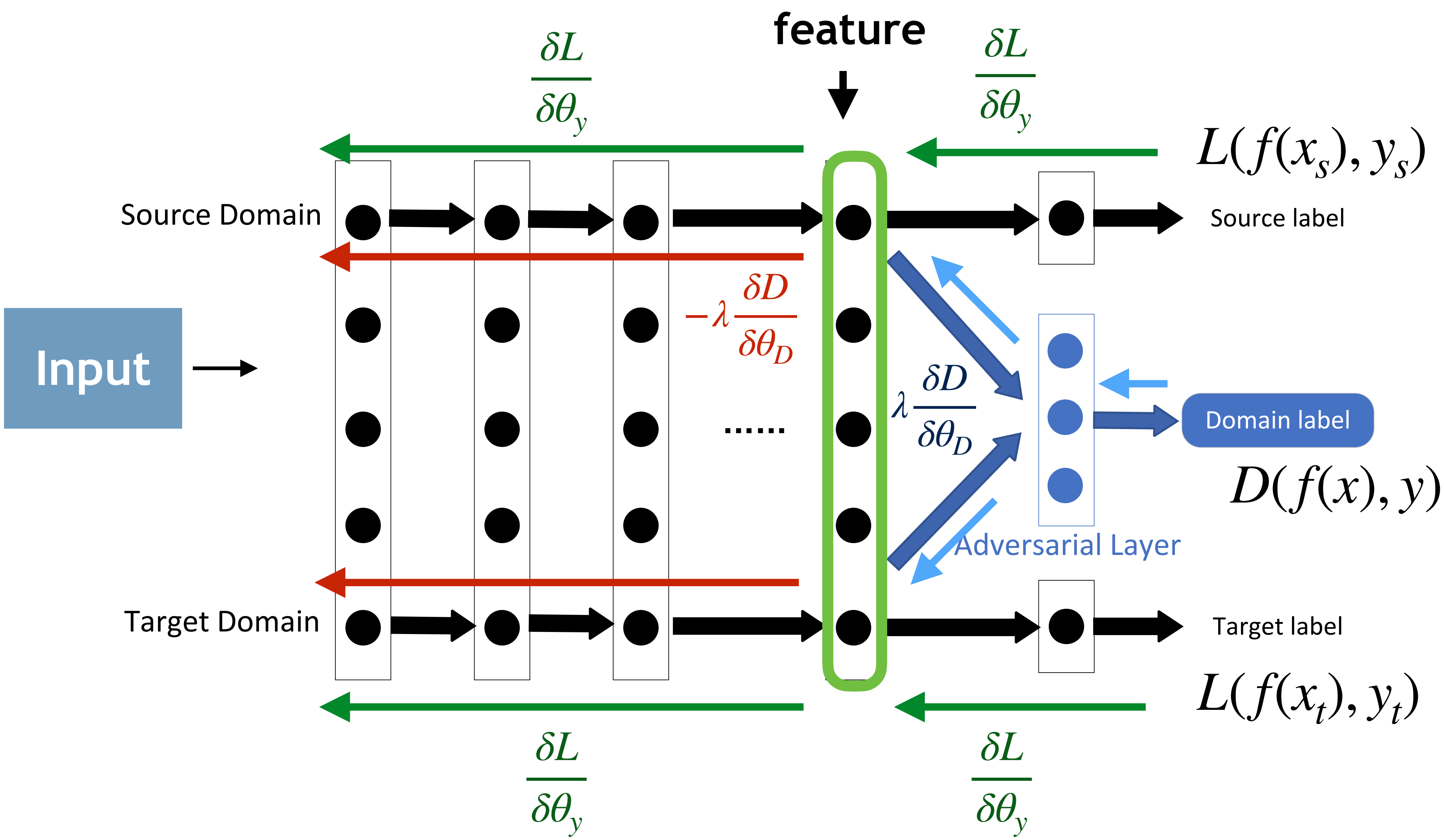
- Standard deep neural network training



Domain Adversarial Neural Networks

Ajakan et al. (2014) Domain-adversarial neural networks.

- Gradient Reversal



Domain Adversarial Neural Networks (DANN)

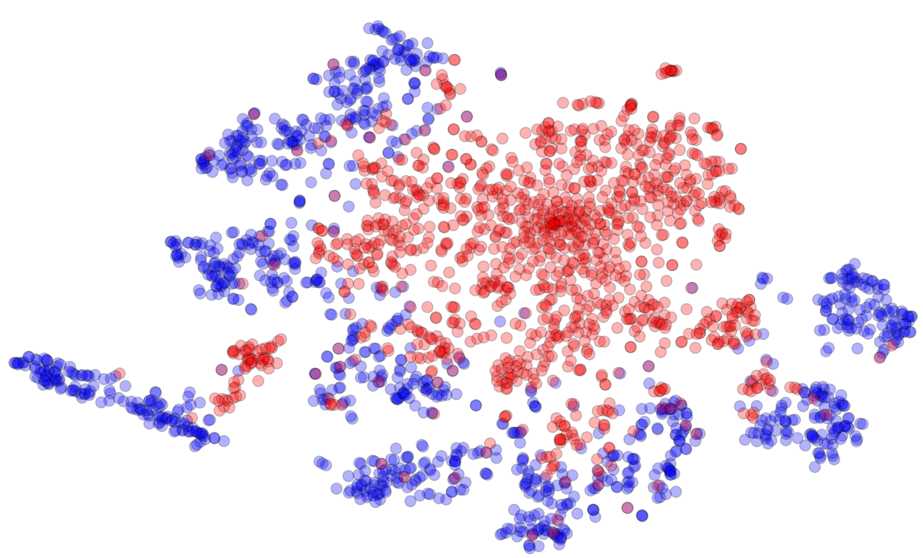
Ajakan et al. (2014) Domain-adversarial neural networks.

- DNN adapted feature distribution

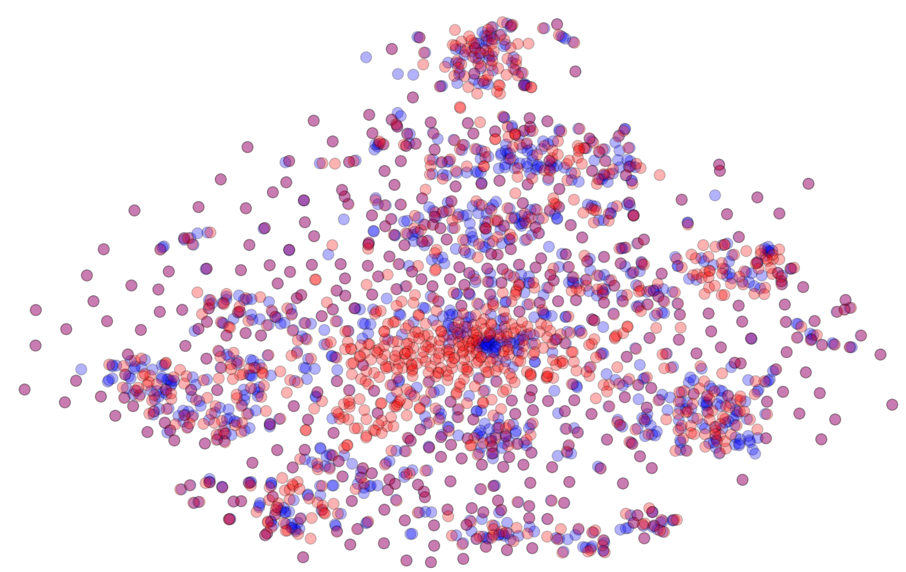
source domain (MINIST)

target domain (MNIST-M)

TSNE visualization of CNN extracted features



Non-Adapted



Adapted

Domain Adaptation Discussion

- Instance-based approach: select and reweight instances in the source domain to be similar to the target distribution
- Mapping-based approach: map source and target data to latent space where source and target domains are similar
- Adversarial-based approach: find features that are indiscriminative between source and target domains

Domain Adaptation Discussion

- Instance-based approach: select and reweight instances in the source domain to be similar to the target distribution

  easy to implement, work with any base classifiers

- Mapping-based approach: map source and target data to latent space where source and target domains are similar

- Adversarial-based approach: find features that are indiscriminative between source and target domains

Domain Adaptation Discussion

- Instance-based approach: select and reweight instances in the source domain to be similar to the target distribution

  easy to implement, work with any base classifiers

- Mapping-based approach: map source and target data to latent space where source and target domains are similar

  easy to incorporate to neural network training

- Adversarial-based approach: find features that are indiscriminative between source and target domains

Domain Adaptation Discussion

- Instance-based approach: select and reweight instances in the source domain to be similar to the target distribution

  easy to implement, work with any base classifiers

- Mapping-based approach: map source and target data to latent space where source and target domains are similar

  easy to incorporate to neural network training

- Adversarial-based approach: find features that are indiscriminative between source and target domains

  good performance in computer vision

Domain Adaptation Discussion

- **Instance-based** approach: select and reweight instances in the source domain to be similar to the target distribution

  easy to implement, work with any base classifiers

- **Mapping-based** approach: map source and target data to latent space where source and target domains are similar

  easy to incorporate to neural network training

- **Adversarial-based** approach: find features that are indiscriminative between source and target domains

  good performance in computer vision

Why does such methods work?

Domain Adaptation Discussion

- Instance-based approach: select and reweight instances in the source domain to be similar to the target distribution

  easy to implement, work with any base classifiers

- Mapping-based approach: map source and target data to latent space where source and target domains are similar

  easy to incorporate to neural network training

- Adversarial-based approach: find features that are indiscriminative between source and target domains

  good performance in computer vision

Why does such methods work?

A detour to learning theory

Transfer Bounds for Domain Adaptation

- Given input $x \sim D$ with discrete alphabet $\mathcal{X}$ and label $y \in \{0,1\}$
- A hypothesis is a function $h : \mathcal{X} \rightarrow \{0,1\}$
- Generalization error (risk) of hypothesis h :

$$\epsilon(h) = \mathbb{E}_{x \sim D}[\,|\,h(x) - y\,|\,]$$

- Empirical risk of hypothesis h given N samples (x_i, y_i) drawn i.i.d. from D:

$$\hat{\epsilon}(h) = \frac{1}{N}\sum_{i=1}^{N} |\,h(x_i) - y_i\,|$$

- Source risk: $\epsilon_S(h) = \mathbb{E}_{x_S \sim P}[\,|\,h(x_S) - y_S\,|\,]$
- Target risk: $\epsilon_T(h) = \mathbb{E}_{x_T \sim Q}[\,|\,h(x_T) - y_T\,|\,]$

Transfer Bounds for Domain Adaptation

Ben-David et.al. (2010). A theory of learning from different domains

Theorem. Let $h \in \mathscr{H}$ be a hypothesis, $\epsilon_S(h)$ and $\epsilon_T(h)$ be risks of source and target respectively, then

$$\epsilon_T(h) \leq \epsilon_S(h) + \boxed{d_{\mathscr{H}}(P, Q)} + C_0$$

C_0: a constant for the complexity of $\mathcal{H}$

where

$$d_{\mathscr{H}}(P, Q) \triangleq 2 \sup_{\eta \in \mathscr{H}} \left| \Pr_P[\eta(x_S) = 1] - \Pr_Q[\eta(x_T) = 1] \right|$$

is the H-divergence between P and Q.

Transfer Bounds for Domain Adaptation

Ben-David et.al. (2010). A theory of learning from different domains

Theorem. Let $h \in \mathscr{H}$ be a hypothesis, $\epsilon_S(h)$ and $\epsilon_T(h)$ be risks of source and target respectively, then

$$\epsilon_T(h) \le \epsilon_S(h) + \boxed{d_{\mathscr{H}}(P, Q)} + C_0 \longleftarrow$$

C_0: a constant for the complexity of $\mathcal{H}$

where

$$d_{\mathscr{H}}(P, Q) \triangleq 2 \sup_{\eta \in \mathscr{H}} \left| \Pr_P[\eta(x_S) = 1] - \Pr_Q[\eta(x_T) = 1] \right|$$

is the **H-divergence** between P and Q.

Lemma. The H-divergence can be bounded by the empirical estimate:

$$d_{\mathscr{H}}(P, Q) \le \boxed{\hat{d}_{\mathscr{H}}(P, Q)} + C_1$$

Transfer Bounds for Domain Adaptation

Ben-David et.al. (2010). A theory of learning from different domains

Theorem. Let $h \in \mathcal{H}$ be a hypothesis, $\epsilon_S(h)$ and $\epsilon_T(h)$ be risks of source and target respectively, then

$$\epsilon_T(h) \leq \epsilon_S(h) + \boxed{d_{\mathcal{H}}(P,Q)} + C_0$$

$\leftarrow$ C_0: a constant for the complexity of $\mathcal{H}$

where

$$d_{\mathcal{H}}(P,Q) \triangleq 2 \sup_{\eta \in \mathcal{H}} \left| \Pr_P[\eta(x_S) = 1] - \Pr_Q[\eta(x_T) = 1] \right|$$

is the **H-divergence** between P and Q.

Lemma. The H-divergence can be bounded by the empirical estimate:

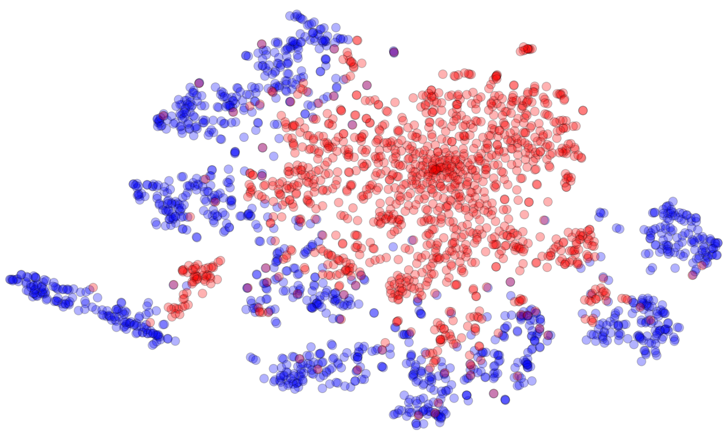
$$d_{\mathcal{H}}(P,Q) \leq \boxed{\hat{d}_{\mathcal{H}}(P,Q)} + C_1$$

Make P and Q as indistinguishable as possible $\Leftrightarrow$ **Decrease the upper bound on target risk !**

e.g. minimize MMD, MK-MMD, domain discriminative loss, etc

Outline

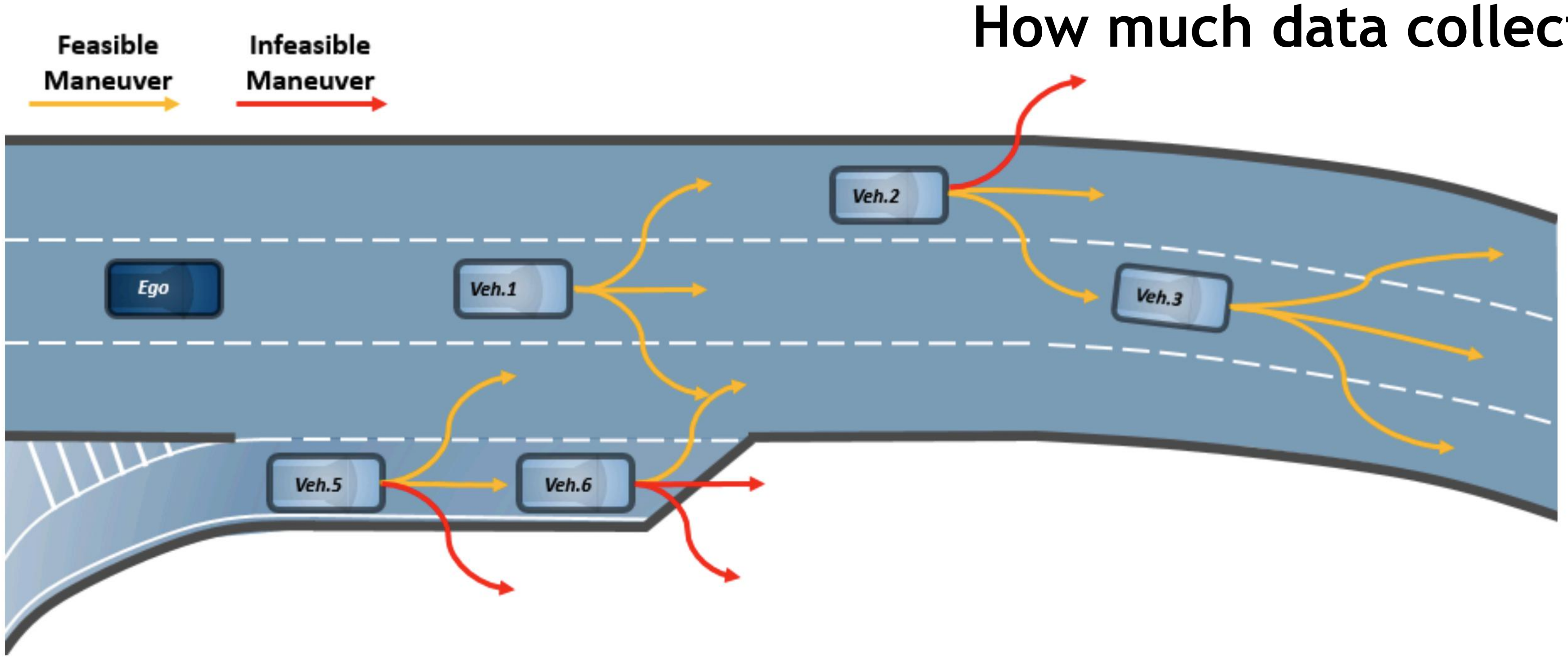
- What's Transfer Learning
- Traditional transfer learning algorithms
  - Task transfer learning
  - Domain adaptation
  - Transfer bound on domain adaptation
- When to transfer?
  - Transferability estimation
- Research trends



Motivating Example 1: driving trajectory prediction

Can we *transfer* the San Francisco model to San Diego?

How much data collection $$ can be saved?



Apoorv Singh, TRAJECTORY-PREDICTION WITH VISION: A SURVEY

Motivating Example 2: Model selection for few-shot tasks

307M

ViT-Large

23M

ResNet50

Task "Sketch"

Task "Real"

Is Bigger model better?

Motivating Example 2: Model selection for few-shot tasks

307M

ViT-Large

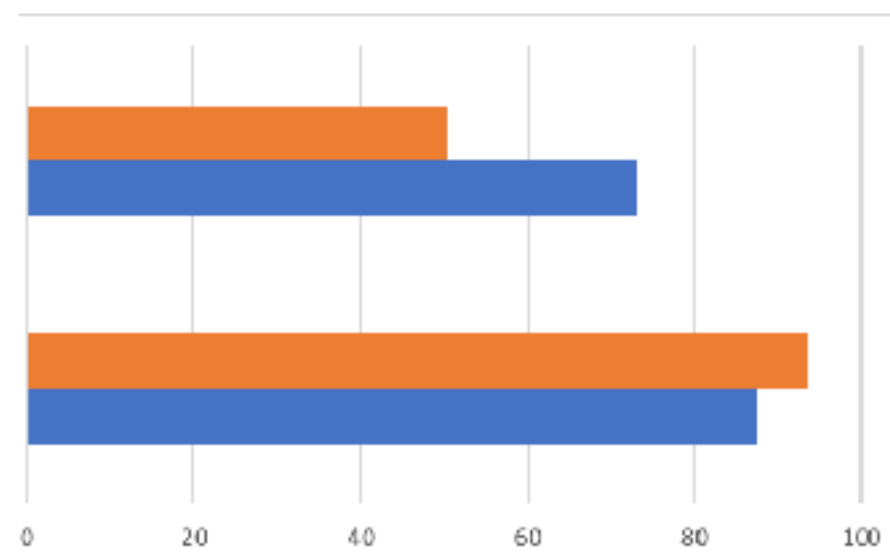
23M

ResNet50

Task “Sketch”

Task “Real”

testing accuracy on target data (n=50)



Motivating Example 2: Model selection for few-shot tasks

307M

ViT-Large

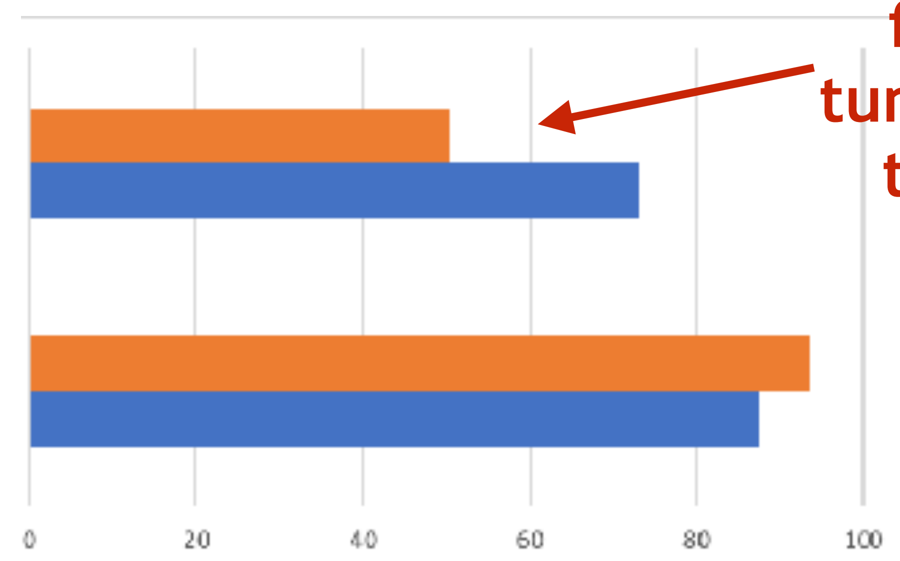
23M

ResNet50

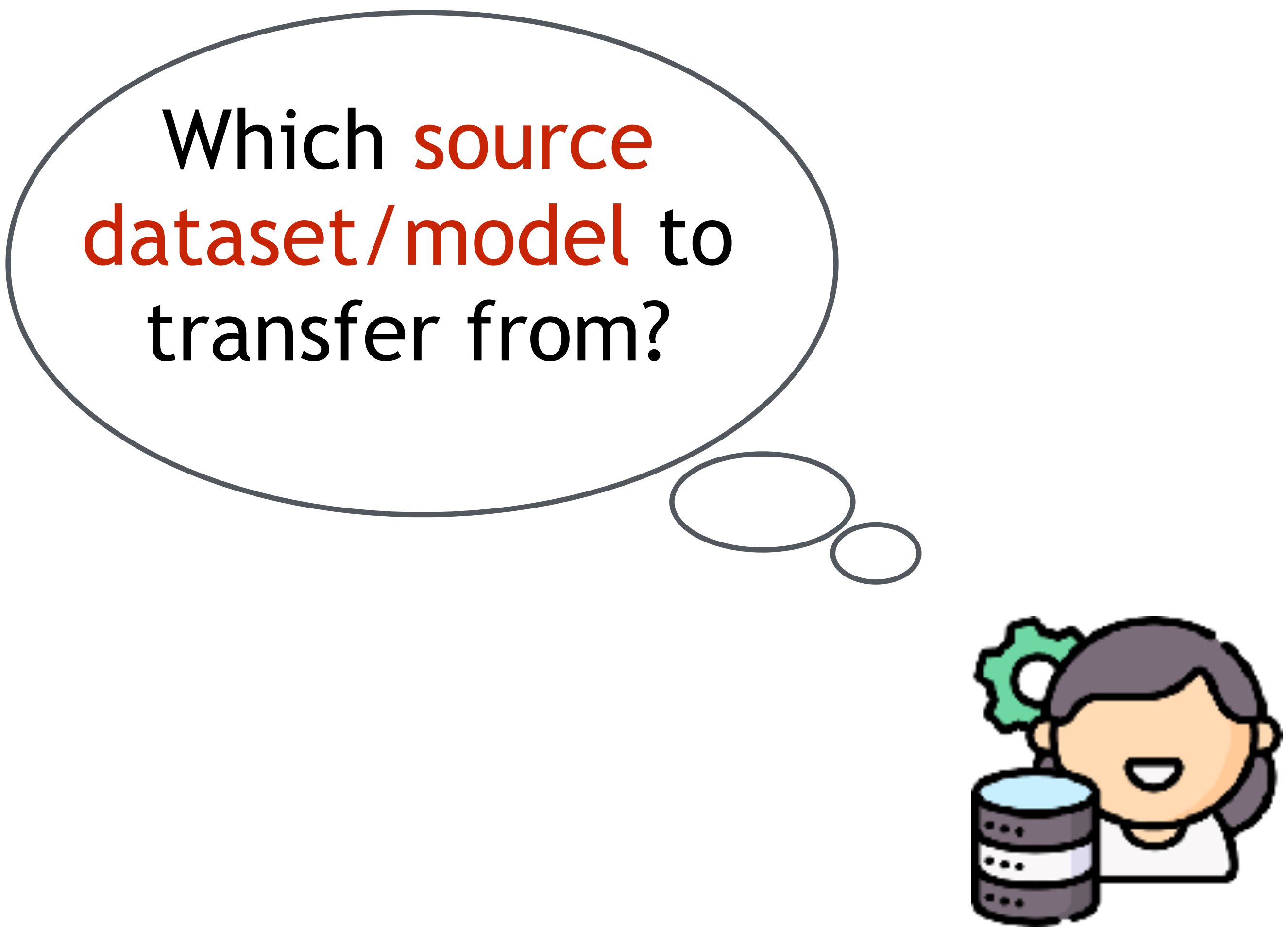
Task “Sketch”

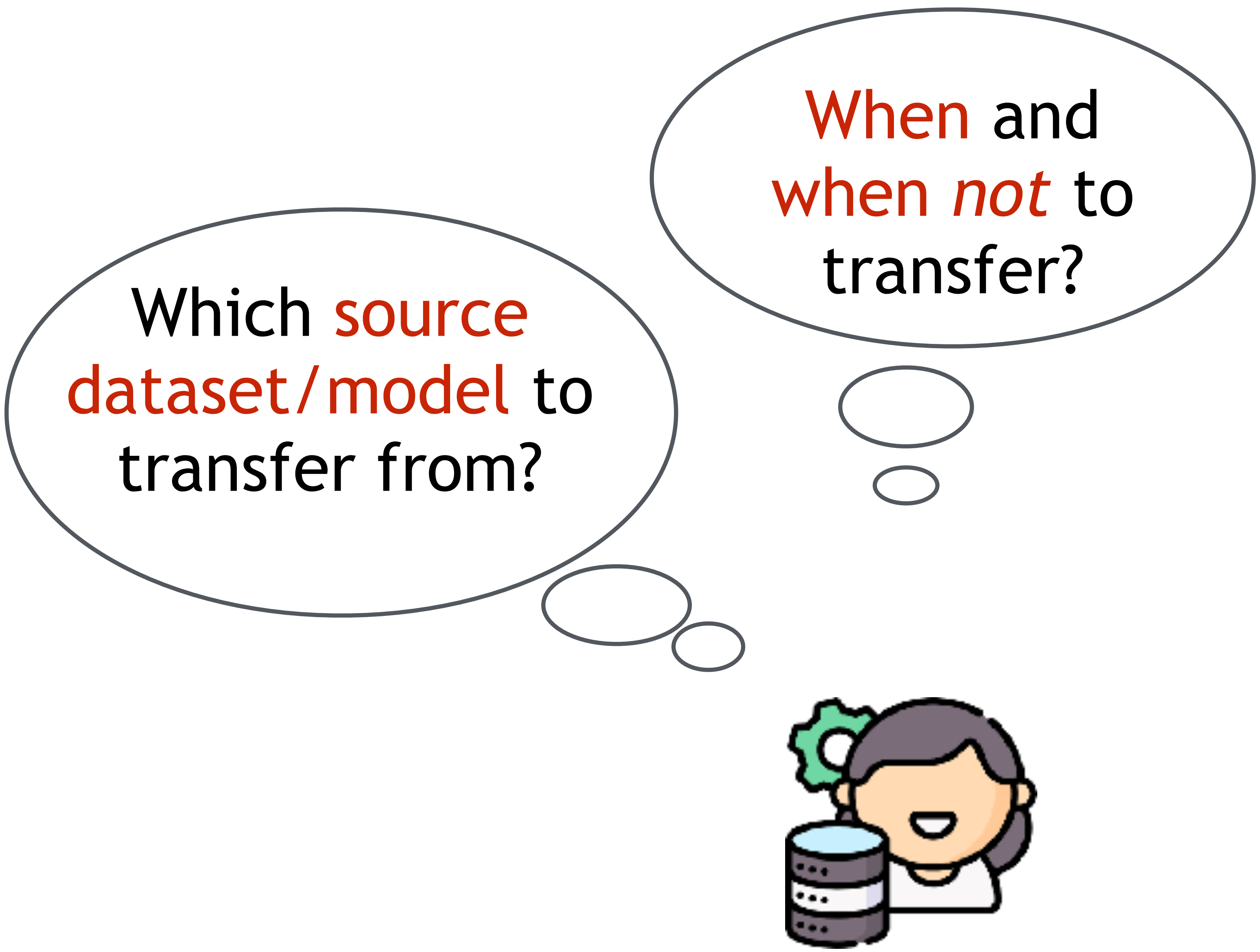
Task “Real”

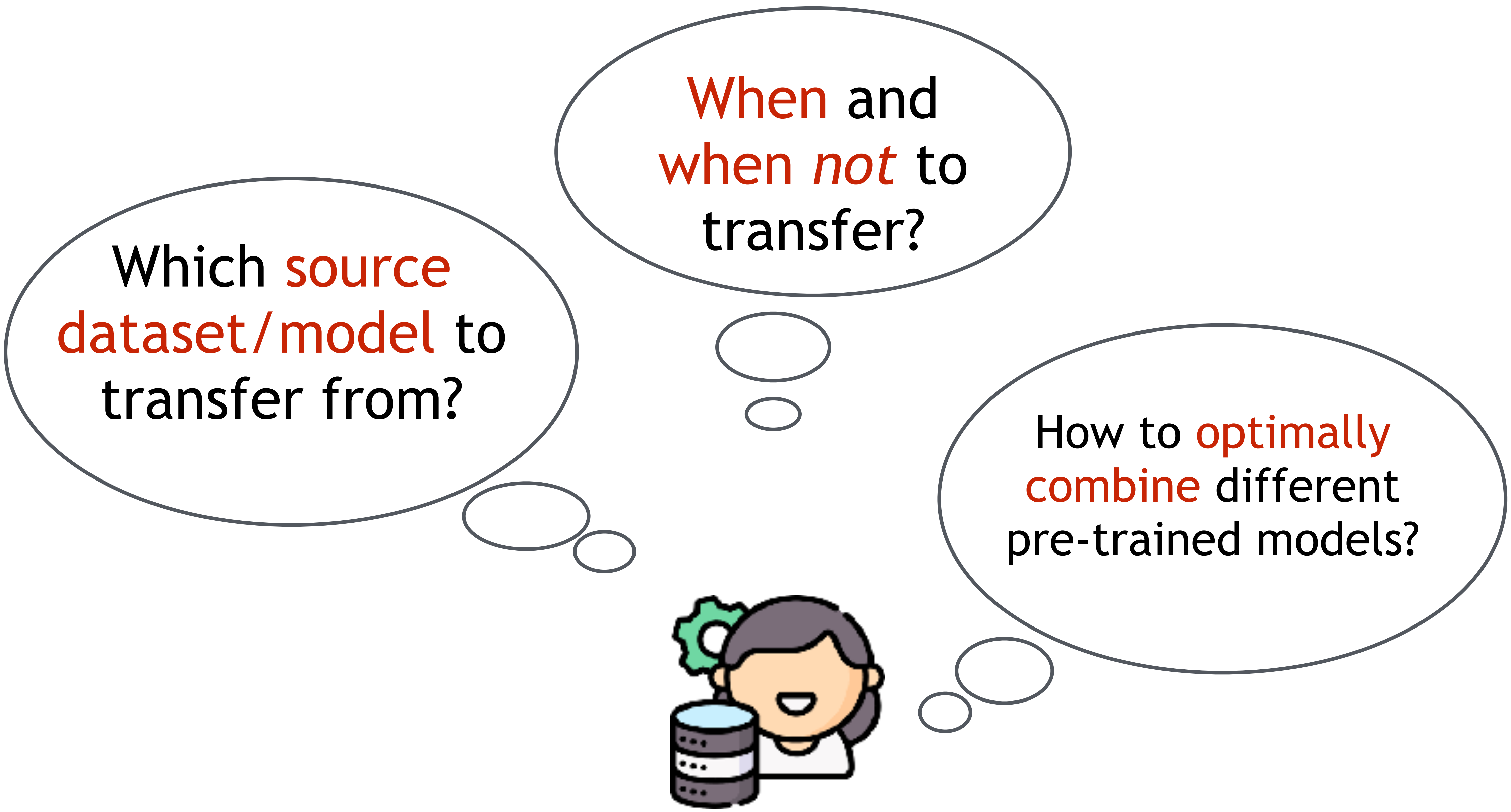
testing accuracy on target data (n=50)



Without full fine-tuning, ViT transfers poorly


Which source dataset/model to transfer from?


When and when *not* to transfer?
Which source dataset/model to transfer from?


Which source dataset/model to transfer from?
When and when *not* to transfer?
How to optimally combine different pre-trained models?

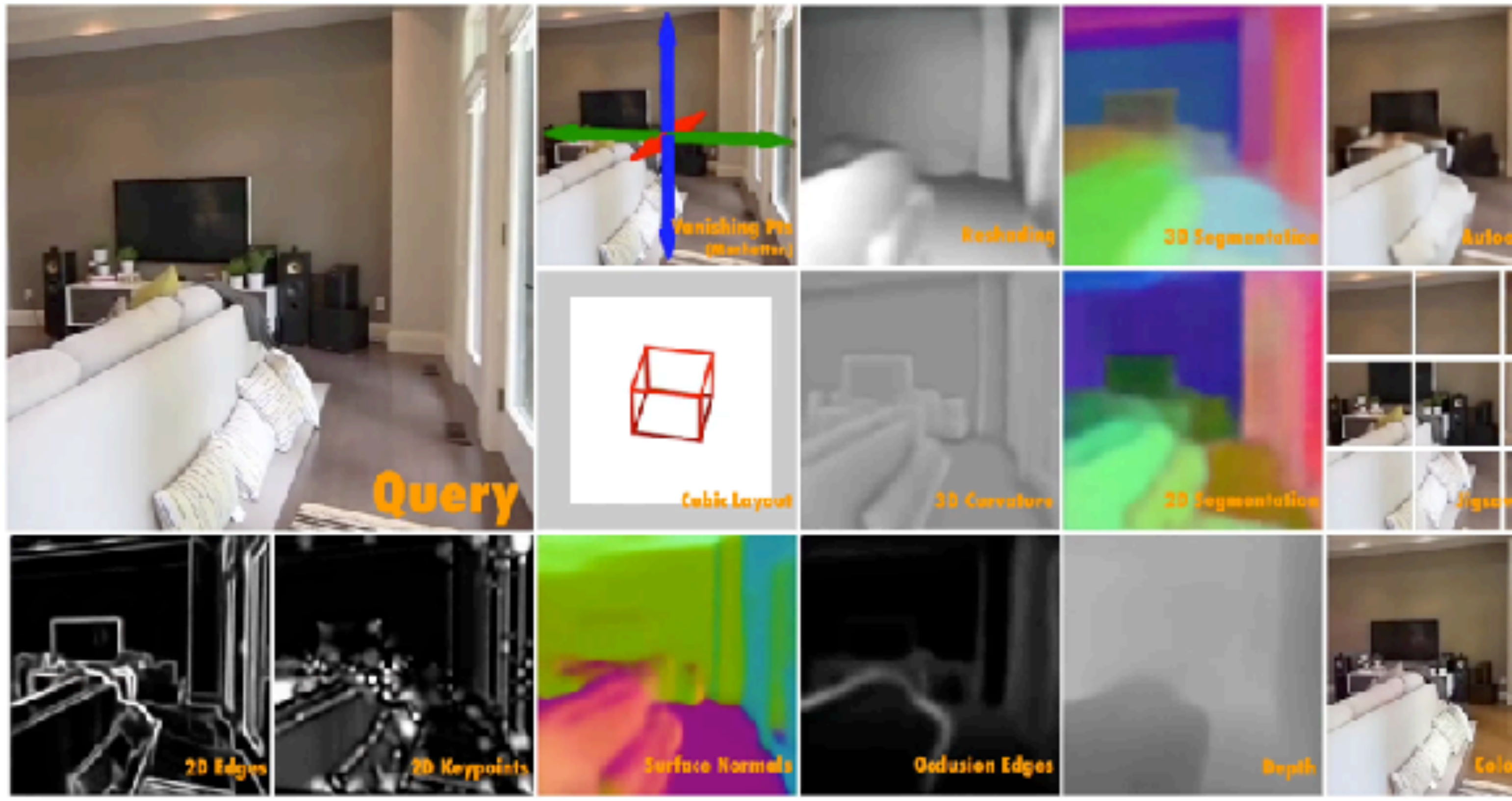
Large-scale studies on empirical transferability has attracted huge attention

Taskonomy (2018): investigated the transferability among 26 image-based indoor scene understanding tasks on low-data scenario

26 Task-Specific Networks

3000 Transfer Networks (include high-order relations)

47,829 GPU hours



A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik and S. Savarese, "Taskonomy: Disentangling Task Transfer Learning," CVPR 2018

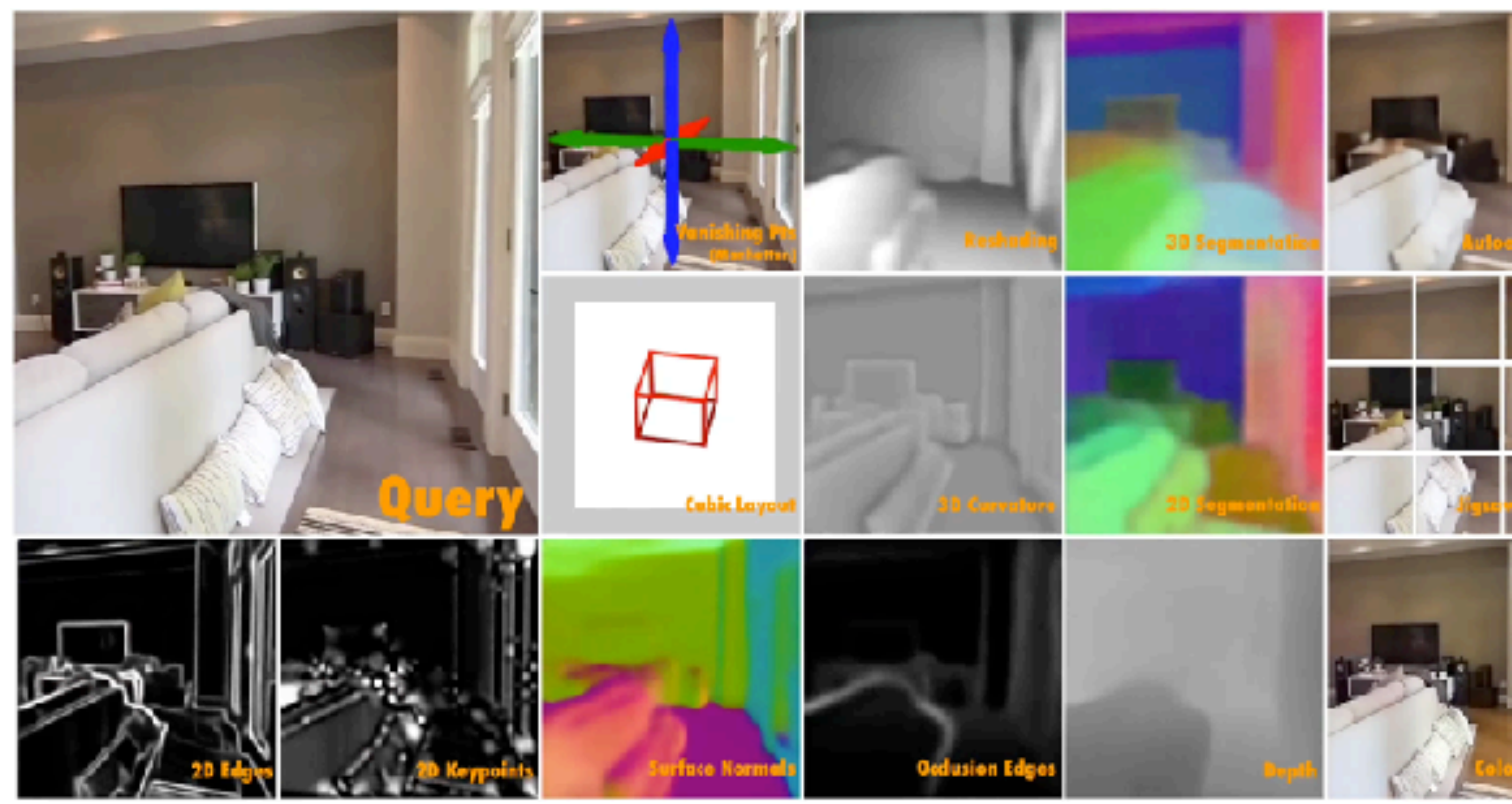
Large-scale studies on empirical transferability has attracted huge attention

Taskonomy (2018): investigated the transferability among 26 image-based indoor scene understanding tasks on low-data scenario

26 Task-Specific Networks

3000 Transfer Networks (include high-order relations)

47,829 GPU hours



A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik and S. Savarese, "Taskonomy: Disentangling Task Transfer Learning," CVPR 2018

Large-scale studies on empirical transferability has attracted huge attention

VTAB (2020): Visual Task Adaptation Benchmark

18 Models pertained on ImageNet

19 target tasks from various domains

Different transfer algorithms tested

	Mea...	Cif...	Cal...	Ca...	Cle...	Cle...	DM...	DTD	Eur...	Flo...	KIT...	Pets	Resi...	Reti...	SVHN	Sun...	dSp...
Sup-Rotation-100%	90.2	84.8	94.6	85.9	99.8	92.5	76.5	75.9	98.8	94.7	82.3	91.5	94.9	79.5	97.0	70.2	100
Sup-Exemplar-100%	90.1	84.1	94.4	86.7	99.8	92.7	76.8	74.5	98.6	93.4	84.0	91.8	95.1	79.5	97.1	69.4	100
Sup-100%	89.7	83.8	94.1	83.9	99.8	92.1	76.4	74.0	98.8	93.2	80.7	91.9	95.3	79.3	97.0	70.7	100
Semi-Exemplar-10%	88.8	82.7	85.3	86.0	99.8	93.1	76.8	70.5	98.6	92.2	81.5	89.0	94.7	78.8	97.0	67.4	100
Semi-Rotation-10%	88.6	82.4	88.1	78.6	99.0	93.2	76.1	72.4	98.7	93.2	81.0	87.8	94.9	79.0	96.9	63.7	100
Rotation	86.4	73.6	88.3	86.4	99.8	93.3	76.8	63.3	98.3	83.4	82.6	71.8	93.4	78.6	96.9	60.5	100
Exemplar	84.8	70.7	81.8	84.7	99.8	93.3	74.7	61.1	98.6	79.3	78.2	67.8	93.5	79.0	96.7	63.2	100
Rel.Pat.Loc	83.1	65.7	79.9	85.3	99.5	87.7	71.5	65.2	97.8	78.3	75.0	66.8	91.5	79.8	93.7	53.0	100
Jigsaw	83.0	65.3	79.1	83.0	99.6	88.6	72.0	63.9	97.9	77.9	74.7	65.4	92.0	80.1	93.9	59.2	100
From-Scratch	75.4	64.4	55.9	81.2	99.7	89.4	71.5	31.3	96.2	50.3	68.4	23.8	83.3	78.0	96.3	52.7	100
Uncond-BigGAN	68.2	58.1	73.6	82.2	47.6	54.9	54.8	44.9	89.8	63.5	57.4	30.9	75.4	75.9	93.0	45.3	86.1
VAE	66.8	44.2	48.4	81.3	98.4	90.1	59.7	16.0	92.5	18.4	67.0	14.0	65.0	74.2	93.1	29.3	100
WAE-MMD	64.9	38.8	50.6	80.6	98.1	89.3	52.6	11.0	94.1	20.3	61.6	16.2	64.9	73.8	90.9	31.6	100
Cond-BigGAN	51.4	56.3	0.148	81.3	12.4	24.5	51.4	44.8	94.5	68.8	49.7	31.6	76.5	75.3	91.4	44.3	6.16
WAE-GAN	43.5	24.8	42.0	77.1	52.2	70.2	37.3	8.67	81.5	15.5	62.3	13.1	33.4	73.6	78.5	12.8	97.7
WAE-UKL	45.8	23.2	41.7	76.4	44.5	67.8	36.7	12.3	76.1	17.2	55.1	12.3	35.9	73.6	65.5	12.0	98.1

Zhai et. al. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, 2020

Theoretical solution is intractable in practice

Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

**Shai Ben-David · John Blitzer · Koby Crammer ·
Alex Kulesza · Fernando Pereira ·
Jennifer Wortman Vaughan**

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 Septer
Published online: 23 October 2009

Domain Adaptation: Learning Bounds and Algorithms

Yishay Mansour
Google Research and
Tel Aviv Univ.
mansour@tau.ac.il

Mehryar Mohri
Courant Institute and
Google Research
mohri@cims.nyu.edu

Afshin Rostamizadeh
Courant Institute
New York University
rostami@cs.nyu.edu

Abstract

This paper addresses the general problem of do-
main adaptation which arises in a variety of appli-

many other areas. Quite often, little or no labeled dat
available from the *target domain*, but labeled data fro
source domain somewhat similar to the target as well as l
amounts of unlabeled data from the target domain are at o

Bridging Theory and Algorithm for Domain Ac

Yuchen Zhang [*1 2] **Tianle Liu** [*2 3] **Mingsheng Long** [1 2] **Michael I. Jordan** [4]

Abstract

This paper addresses the problem of unsupervised domain adaption from theoretical and algorithmic perspectives. Existing domain adaptation theories

Remarkable theoretical advances have been achieved in domain adaptation. Mansour et al. (2009c); Ben-David et al. (2010) provided rigorous learning bounds for unsupervised domain adaptation, a most challenging scenario in this field. These earliest theories have later been extended in many

Theoretical solution is intractable in practice

- Focus on the *theoretical optimal performance*

Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer ·
Alex Kulesza · Fernando Pereira ·
Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 Septe
Published online: 23 October 2009

Domain Adaptation: Learning Bounds and Algorithms

Yishay Mansour
Google Research and
Tel Aviv Univ.
mansour@tau.ac.il

Mehryar Mohri
Courant Institute and
Google Research
mohri@cims.nyu.edu

Afshin Rostamizadeh
Courant Institute
New York University
rostami@cs.nyu.edu

Abstract

This paper addresses the general problem of do-
main adaptation which arises in a variety of appli-

many other areas. Quite often, little or no labeled dat
available from the *target domain*, but labeled data fro
source domain somewhat similar to the target as well as l
amounts of unlabeled data from the target domain are at o

Bridging Theory and Algorithm for Domain Ac

Yuchen Zhang [*1 2] Tianle Liu [*2 3] Mingsheng Long [1 2] Michael I. Jordan [4]

Abstract

This paper addresses the problem of unsupervised domain adaption from theoretical and algorithmic perspectives. Existing domain adaptation theories

Remarkable theoretical advances have been achieved in domain adaptation. Mansour et al. (2009c); Ben-David et al. (2010) provided rigorous learning bounds for unsupervised domain adaptation, a most challenging scenario in this field. These earliest theories have later been extended in many

Theoretical solution is intractable in practice

- Focus on the *theoretical optimal performance*
- Strict model assumptions

Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 Septe
Published online: 23 October 2009

Domain Adaptation: Learning Bounds and Algorithms

Yishay Mansour
Google Research and
Tel Aviv Univ.
mansour@tau.ac.il

Mehryar Mohri
Courant Institute and
Google Research
mohri@cims.nyu.edu

Afshin Rostamizadeh
Courant Institute
New York University
rostami@cs.nyu.edu

Abstract

This paper addresses the general problem of do-
main adaptation which arises in a variety of appli-

many other areas. Quite often, little or no labeled dat
available from the *target domain*, but labeled data fro
source domain somewhat similar to the target as well as l
amounts of unlabeled data from the target domain are at o

Bridging Theory and Algorithm for Domain Ac

Yuchen Zhang [*1 2] **Tianle Liu** [*2 3] **Mingsheng Long** [1 2] **Michael I. Jordan** [4]

Abstract

This paper addresses the problem of unsupervised domain adaption from theoretical and algorithmic perspectives. Existing domain adaptation theories

Remarkable theoretical advances have been achieved in domain adaptation. Mansour et al. (2009c); Ben-David et al. (2010) provided rigorous learning bounds for unsupervised domain adaptation, a most challenging scenario in this field. These earliest theories have later been extended in many

Theoretical solution is intractable in practice

- Focus on the *theoretical optimal performance*
- Strict model assumptions
- Intractable model complexity measures

Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 Septe
Published online: 23 October 2009

Domain Adaptation: Learning Bounds and Algorithms

Yishay Mansour
Google Research and
Tel Aviv Univ.
mansour@tau.ac.il

Mehryar Mohri
Courant Institute and
Google Research
mohri@cims.nyu.edu

Afshin Rostamizadeh
Courant Institute
New York University
rostami@cs.nyu.edu

Abstract

This paper addresses the general problem of domain adaptation which arises in a variety of appli-

many other areas. Quite often, little or no labeled dat available from the *target domain*, but labeled data fro *source domain* somewhat similar to the target as well as l amounts of unlabeled data from the target domain are at o

Bridging Theory and Algorithm for Domain Ad

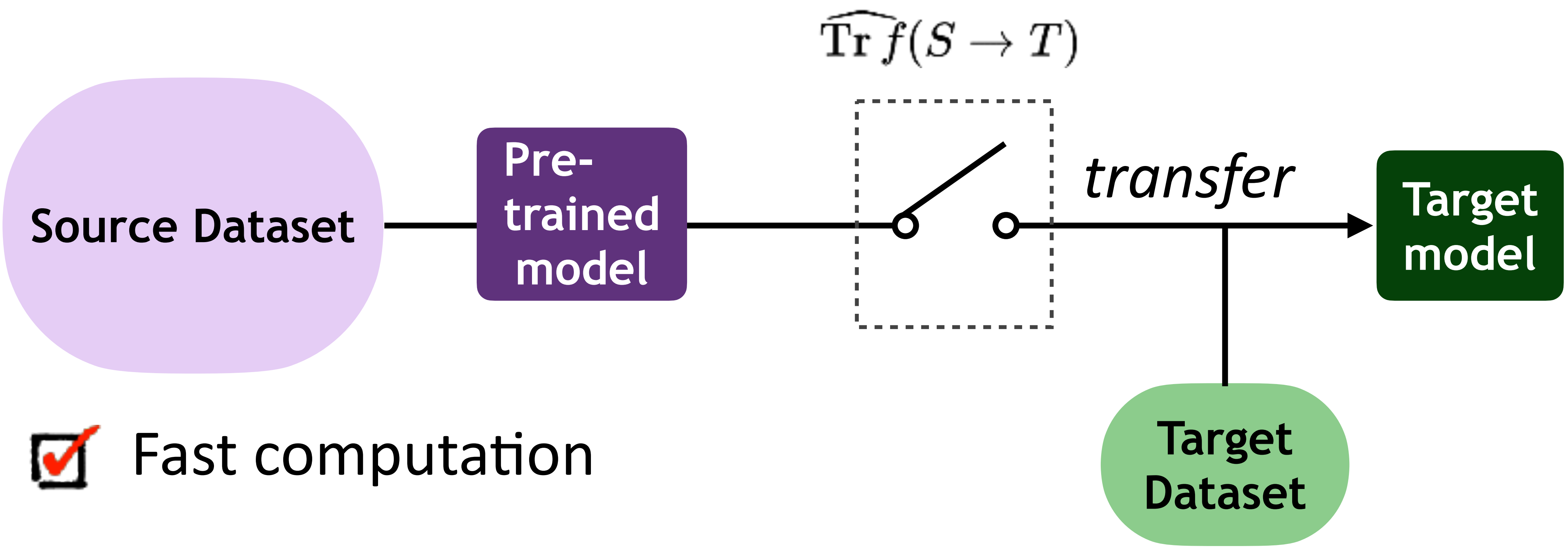
Yuchen Zhang [*1,2] **Tianle Liu** [*2,3] **Mingsheng Long** [1,2] **Michael I. Jordan** [4]

Abstract

This paper addresses the problem of unsupervised domain adaption from theoretical and algorithmic perspectives. Existing domain adaptation theories

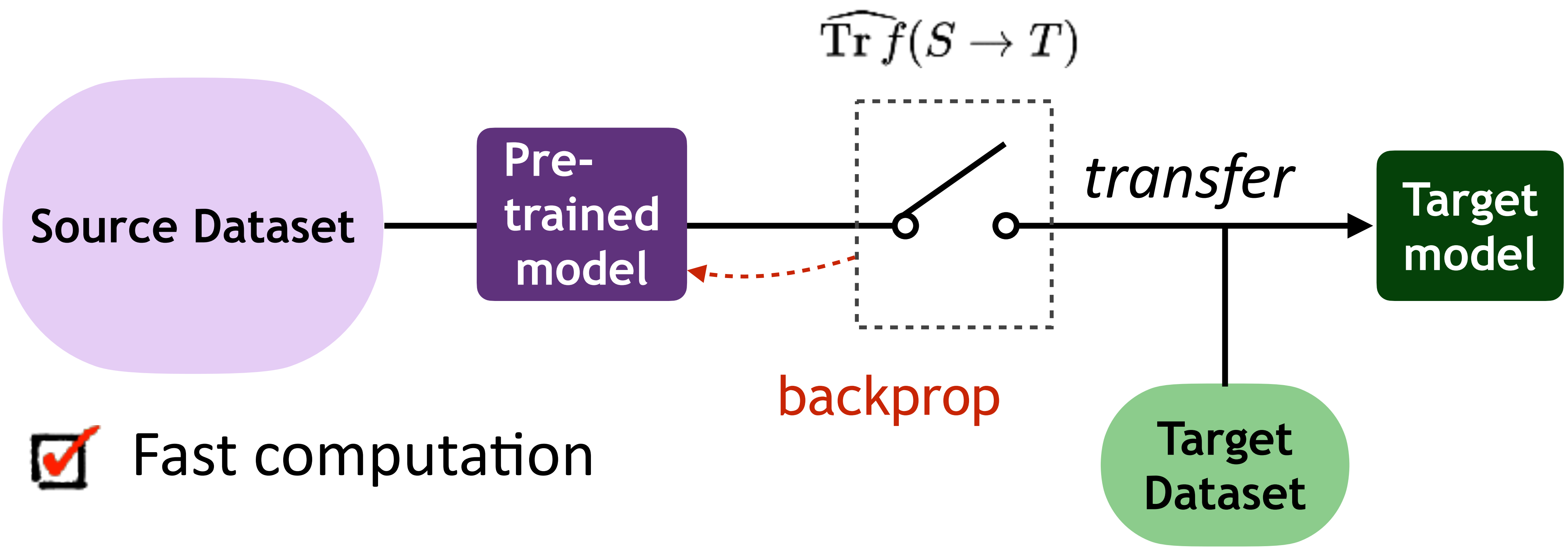
Remarkable theoretical advances have been achieved in domain adaptation. Mansour et al. (2009c); Ben-David et al. (2010) provided rigorous learning bounds for unsupervised domain adaptation, a most challenging scenario in this field. These earliest theories have later been extended in many

How to quantify **transferability** from data ?



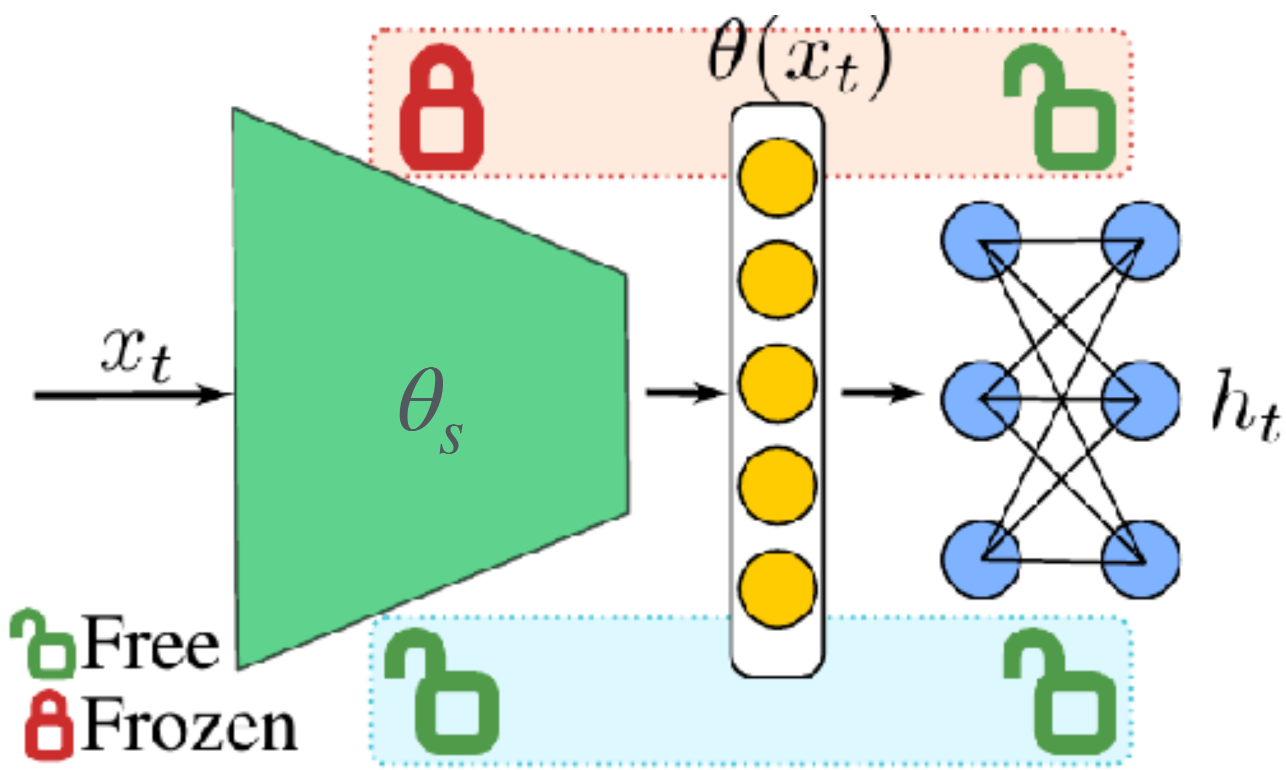
- ☑ Fast computation
- ☑ Theoretically meaningful
- ☑ Differentiable

How to quantify **transferability** from data ?



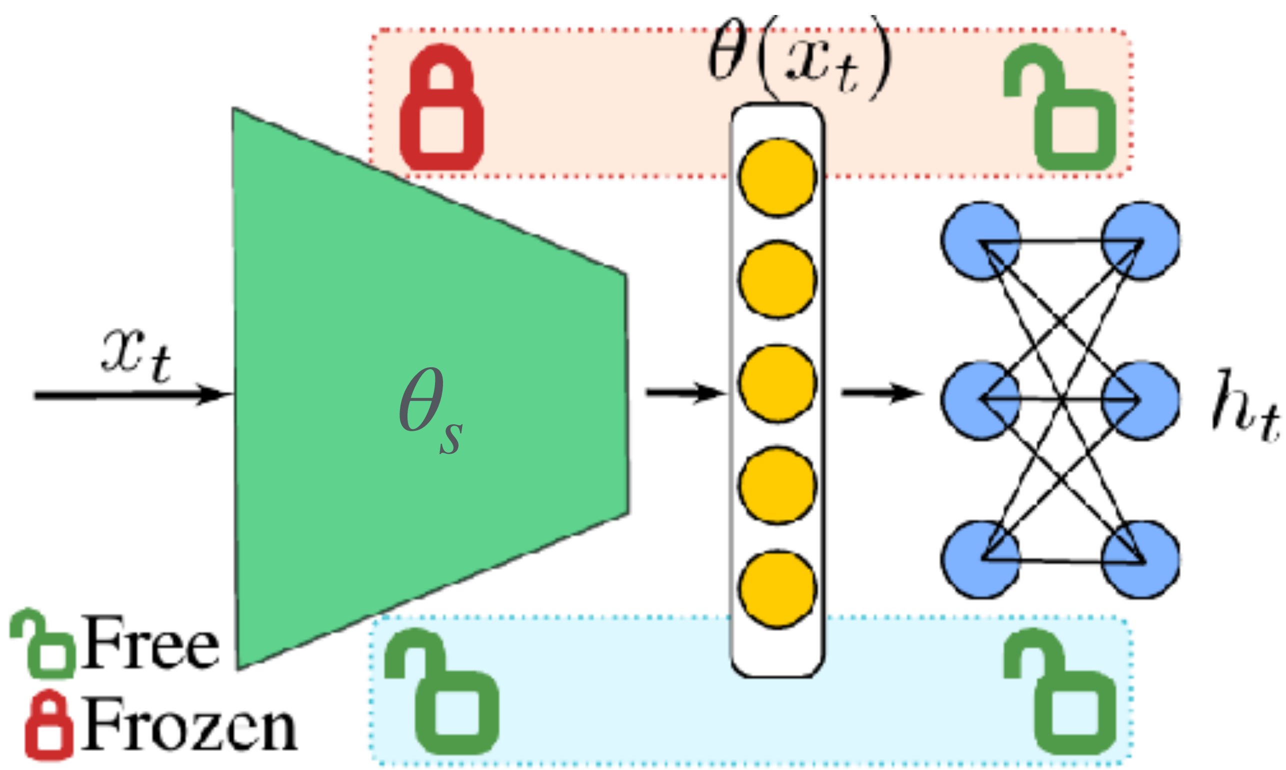
- ☑ Fast computation
- ☑ Theoretically meaningful
- ☑ Differentiable

Empirical transferability is the likelihood of the target training data (X_t, Y_t) using source feature extractor θ_s



$$Trf(S \to T) = \begin{cases} \hat{\mathbb{E}}\left[\log P(Y_t \mid X_t; \theta_s, h_t)\right] & \textbf{(retrain head)} \\ \hat{\mathbb{E}}\left[\log P(Y_t \mid X_t; \theta_t : \theta_t^{(0)} = \theta_s, h_t)\right] & \textbf{(fine-tune)} \end{cases}$$

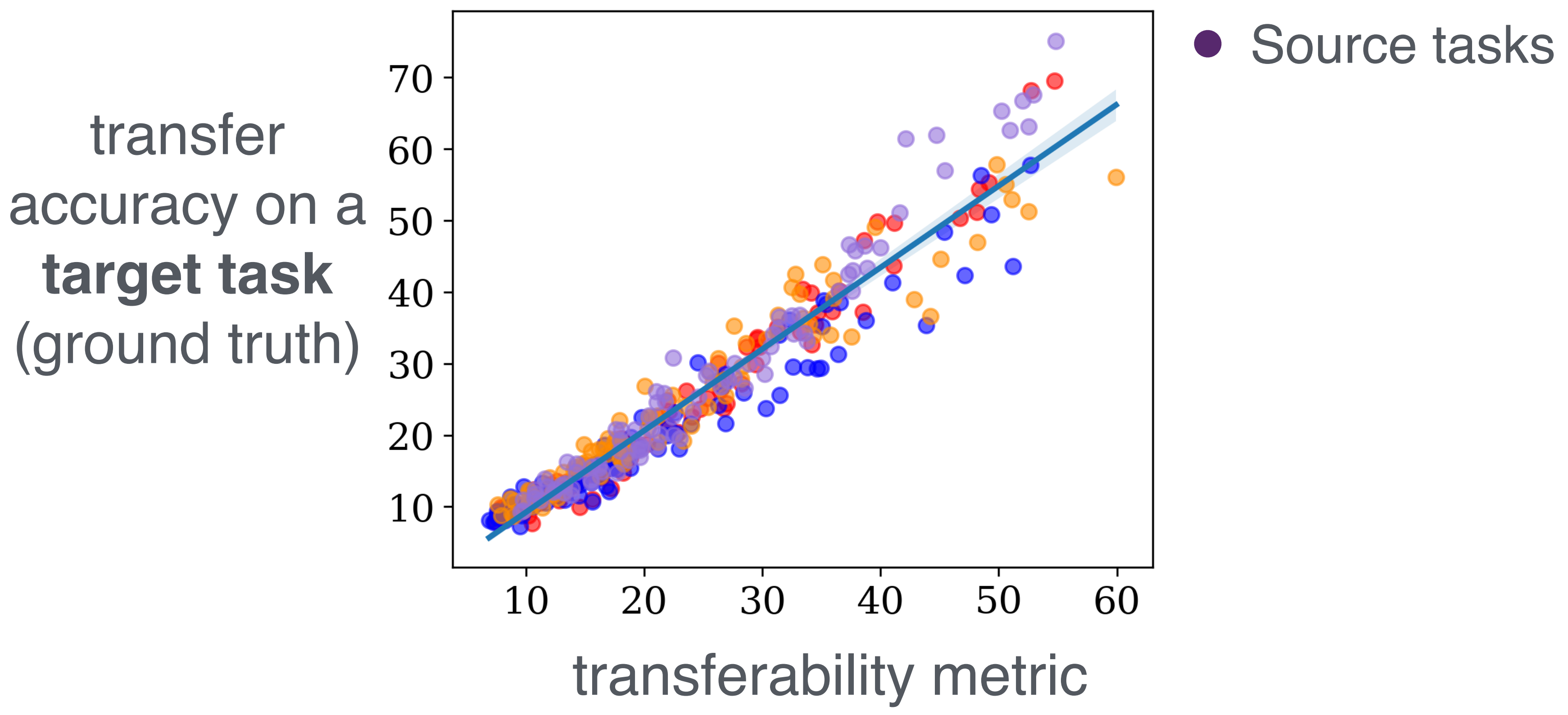
Empirical transferability is the likelihood of the target training data (X_t, Y_t) using source feature extractor θ_s



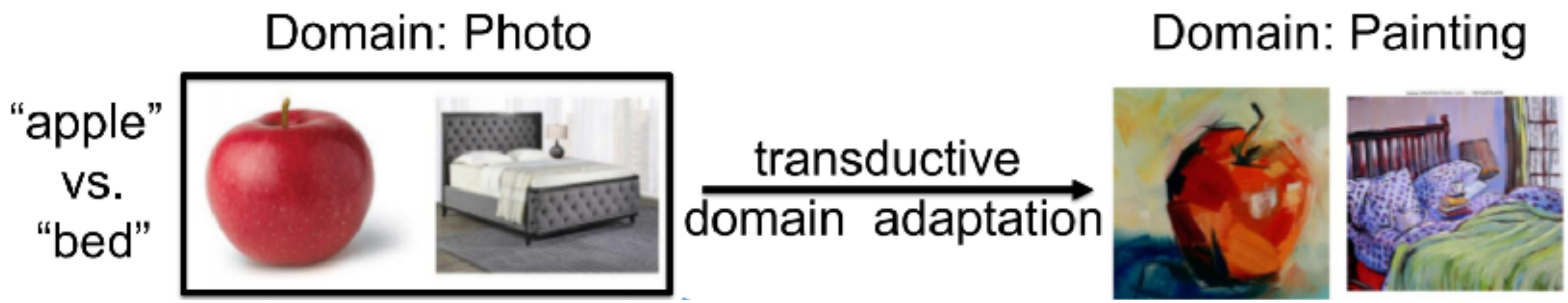
$$Trf(S \to T) = \begin{cases} \hat{\mathbb{E}}\left[\log P(Y_t \mid X_t; \theta_s, h_t)\right] & \textbf{(retrain head)} \\ \hat{\mathbb{E}}\left[\log P(Y_t \mid X_t; \theta_t : \theta_t^{(0)} = \theta_s, h_t)\right] & \textbf{(fine-tune)} \end{cases}$$

Lower bound

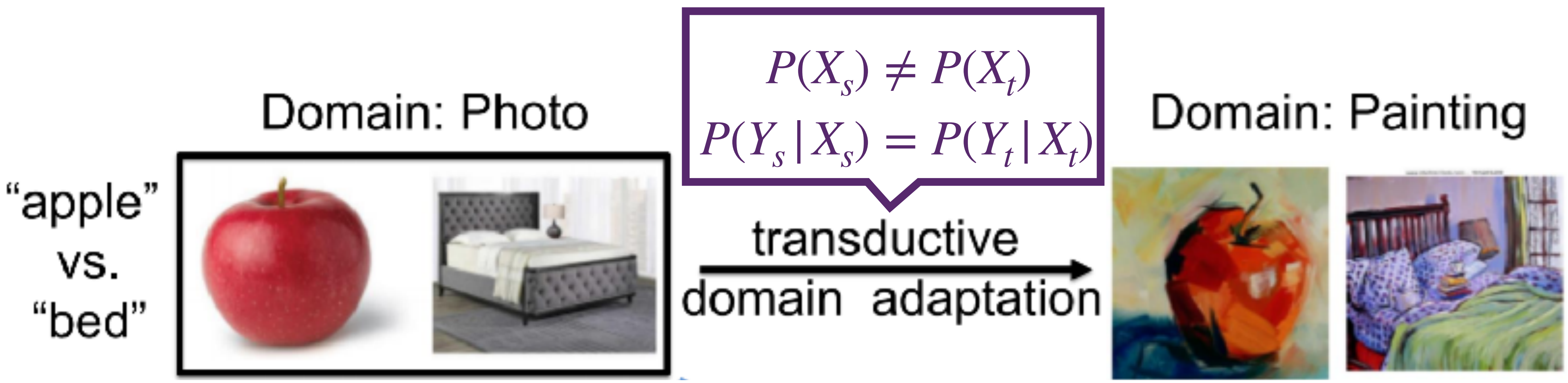
Analytical transferability: estimate transferability w/o training the target network



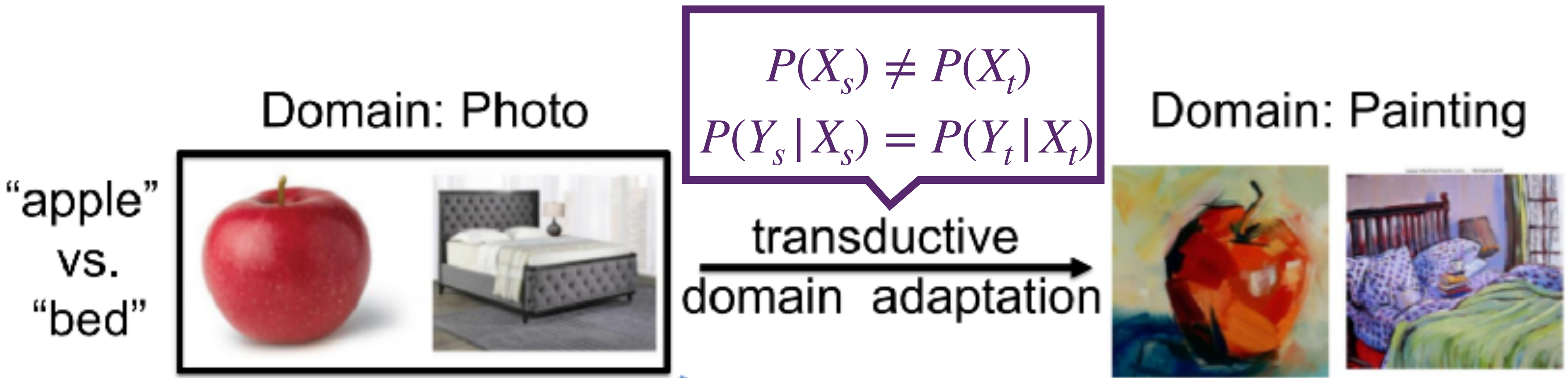
Cross-Domain Transferability are measured using domain divergence $\mathrm{Dist}(P(X_s), P(X_t))$



Cross-Domain Transferability are measured using domain divergence $\mathrm{Dist}(P(X_s), P(X_t))$

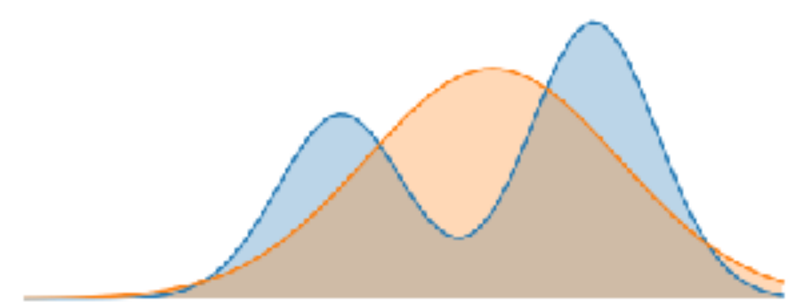


Cross-Domain Transferability are measured using domain divergence $\mathrm{Dist}(P(X_s), P(X_t))$

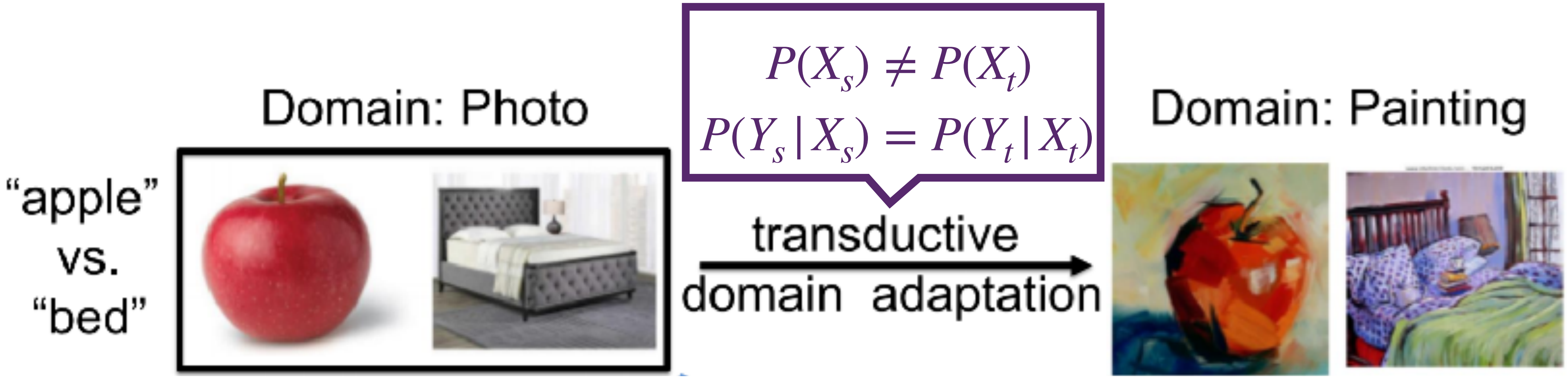


e.g.

- Proxy A-Distance (Ben-David 2006)
- Wasserstein distance (Kantorovich 1942)

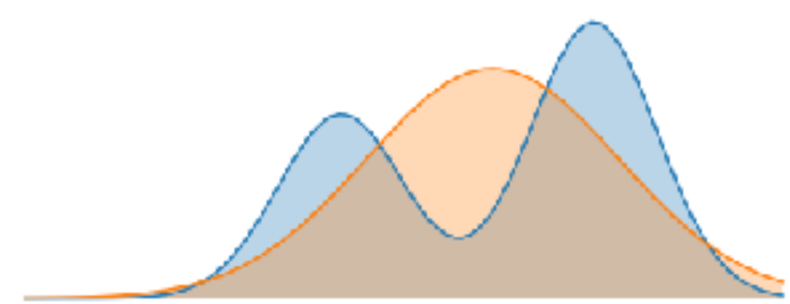


Cross-Domain Transferability are measured using domain divergence $\mathrm{Dist}(P(X_s), P(X_t))$



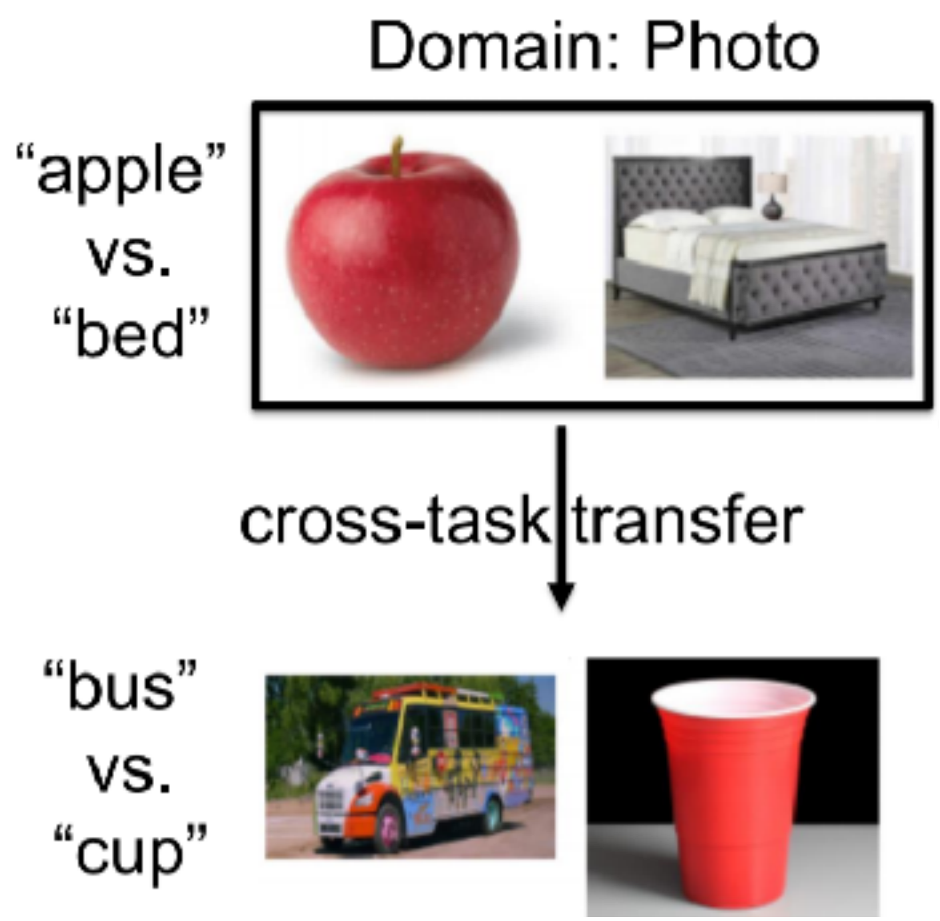
e.g.

- Proxy A-Distance (Ben-David 2006)
- Wasserstein distance (Kantorovich 1942)

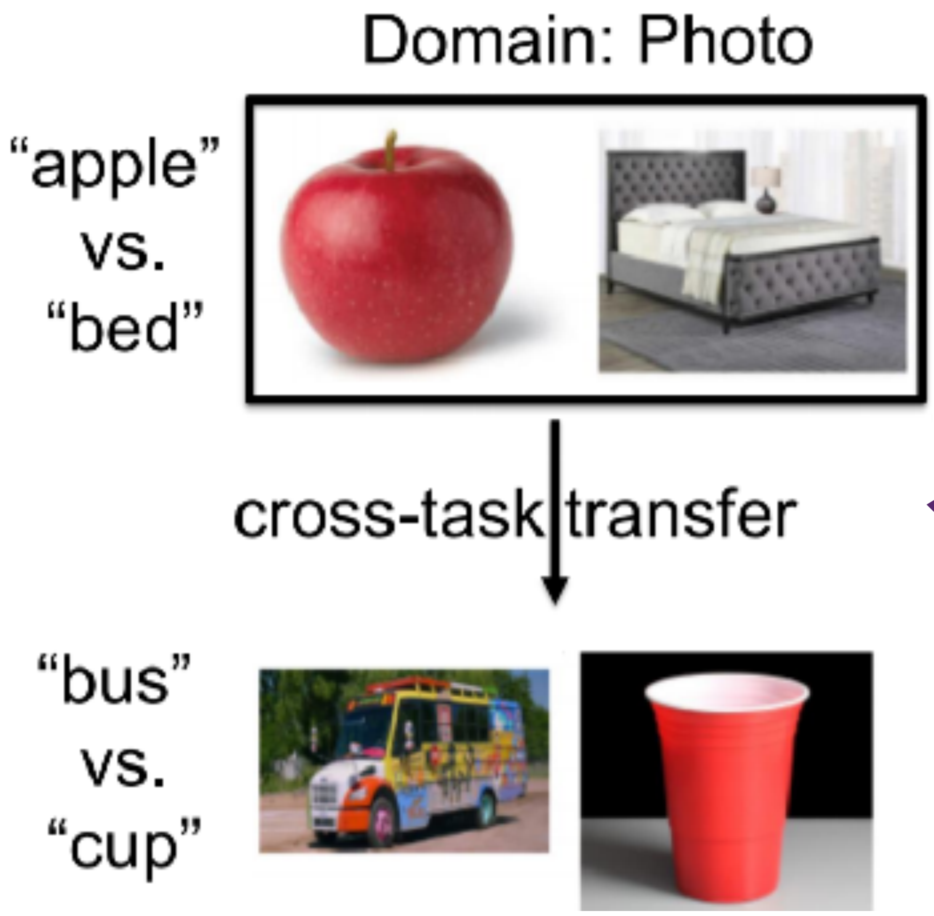


Does not apply to different label space or label distribution!

Cross-Task Transferability

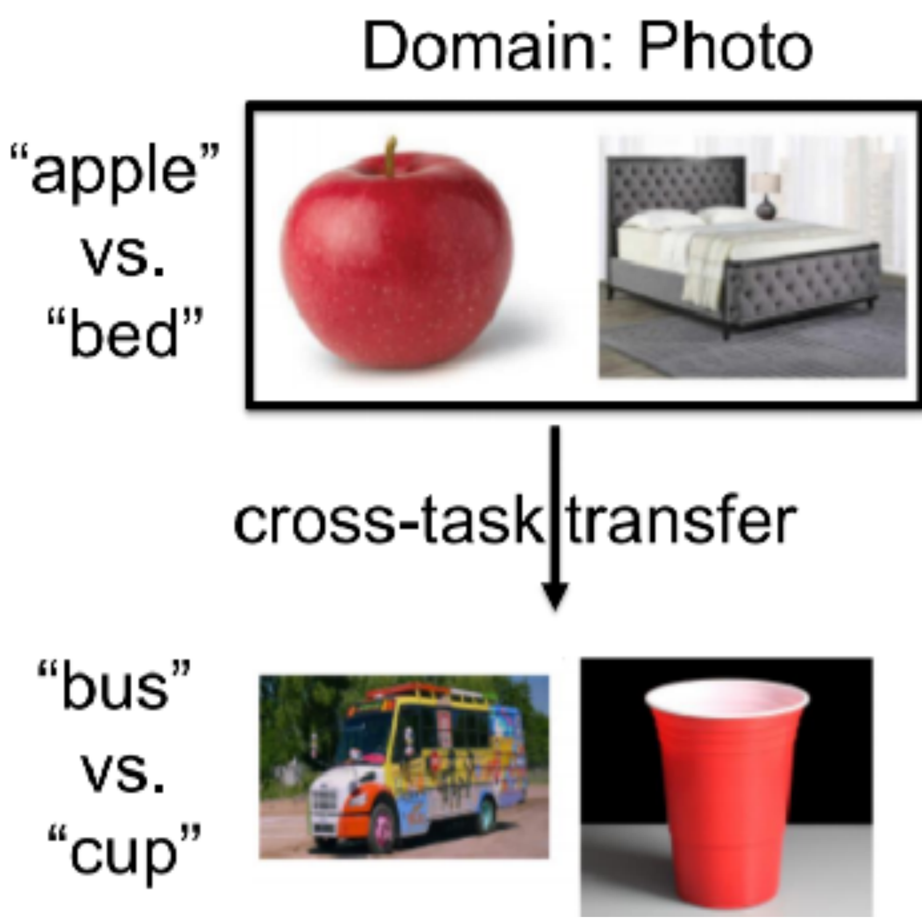


Cross-Task Transferability



$$P(X_s) = P(X_t)$$
$$P(Y_s \mid X_s) \neq P(Y_t \mid X_t)$$

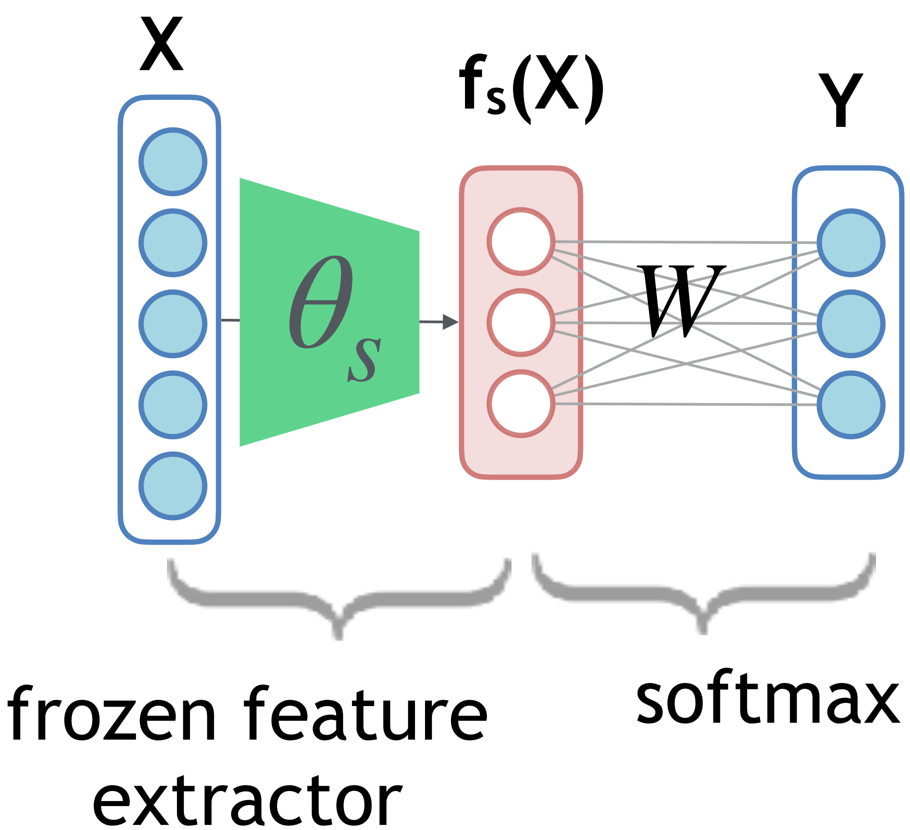
Cross-Task Transferability is estimated via optimal target loss in a simplified transfer model (e.g softmax regression)



$$P(X_s) = P(X_t)$$
$$P(Y_s \mid X_s) \neq P(Y_t \mid X_t)$$

Analytical Task Transferability Metric: H-Score

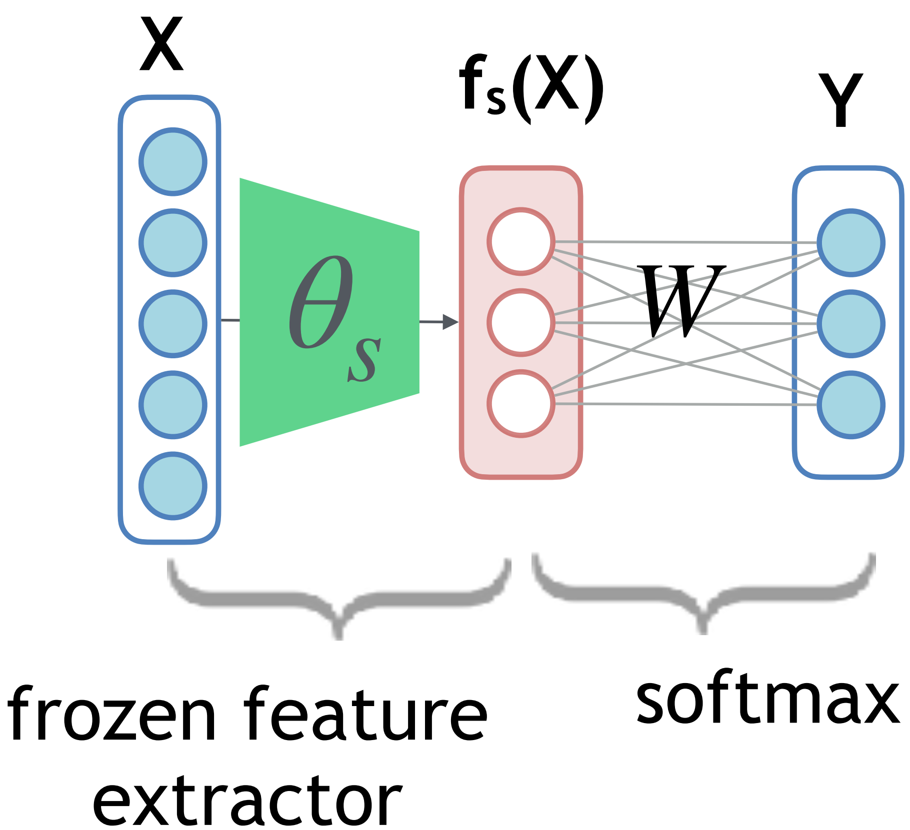
Yajie Bao, Yang Li, et. al. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE ICIP*, pp. 2309-2313. 2019.



Only need to compute $\mathscr{H}(f_s)$ for source selection

Analytical Task Transferability Metric: H-Score

Yajie Bao, Yang Li, et. al. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE ICIP*, pp. 2309-2313. 2019.



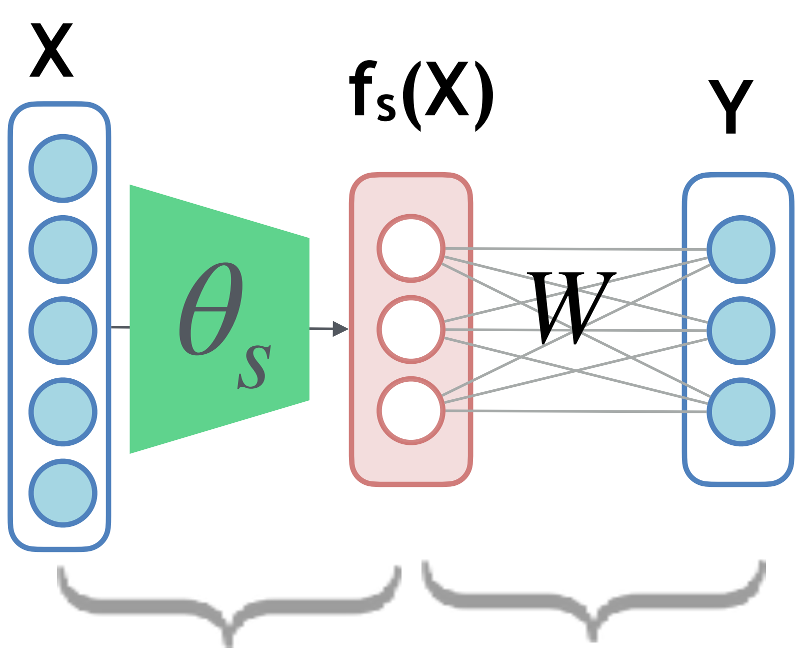
- By local information geometry, given zero-mean feature $f(x)$, the optimal target loss is

$$L(f_s, W^\star) = Const(X, Y) - \mathscr{H}(f_s) + o(\epsilon^2)$$

Only need to compute $\mathscr{H}(f_s)$ for source selection

Analytical Task Transferability Metric: H-Score

Yajie Bao, Yang Li, et. al. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE ICIP*, pp. 2309-2313. 2019.



- By local information geometry, given zero-mean feature $f(x)$, the optimal target loss is

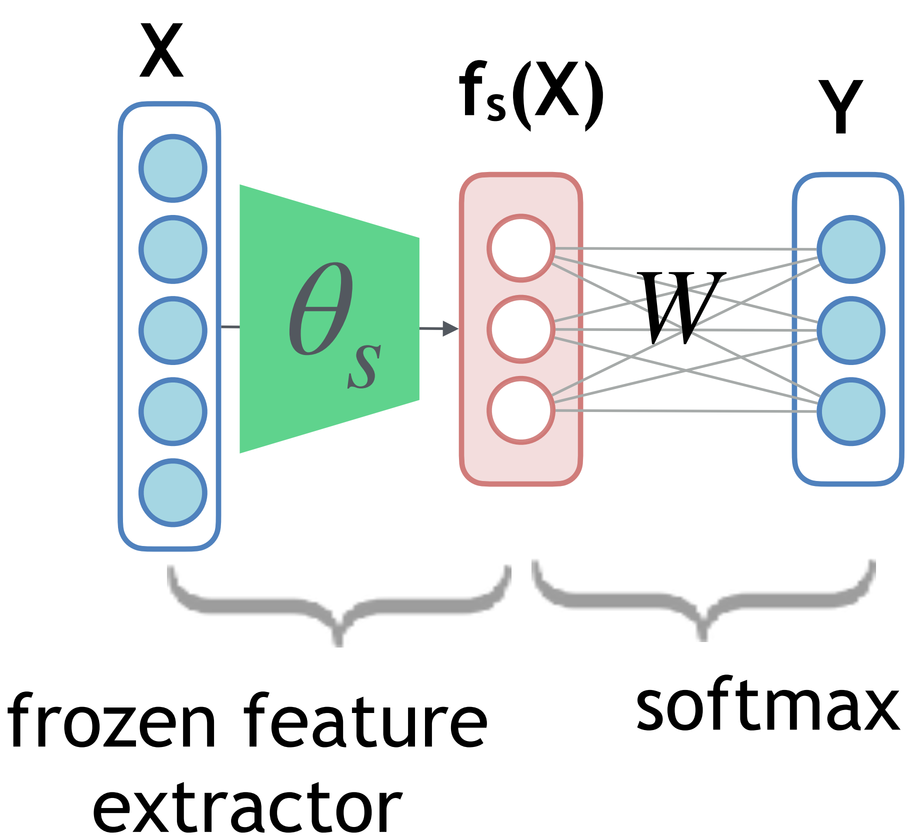
$$L(f_s, W^\star) = Const(X, Y) - \underbrace{\mathscr{H}(f_s)}_{\text{H-score of } f(x)} + o(\epsilon^2)$$

$$\boxed{\mathscr{H}(f_s) = \mathrm{tr}\left(\mathrm{cov}(f_s(X))^{-1}\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f_s(X) \mid Y])\right)}$$

Only need to compute $\mathscr{H}(f_s)$ for source selection

Analytical Task Transferability Metric: H-Score

Yajie Bao, Yang Li, et. al. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE ICIP*, pp. 2309-2313. 2019.



- By local information geometry, given zero-mean feature $f(x)$, the optimal target loss is

$$L(f_s, W^\star) = Const(X, Y) - \underbrace{\mathcal{H}(f_s)}_{\text{H-score of } f(x)} + o(\epsilon^2)$$

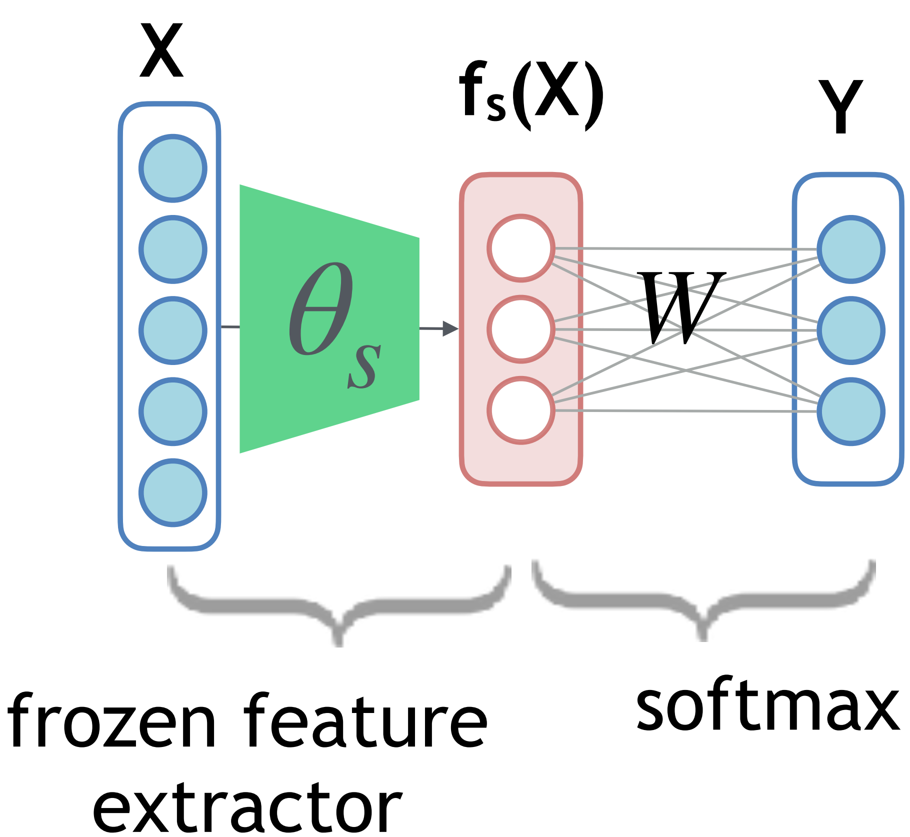
$$\boxed{\mathcal{H}(f_s) = \mathrm{tr}\left(\mathrm{cov}(f_s(X))^{-1}\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f_s(X)\,|\,Y])\right)}$$

Higher H-score $\leftrightarrow$ Better Performance

Only need to compute $\mathcal{H}(f_s)$ for source selection

Analytical Task Transferability Metric: H-Score

Yajie Bao, Yang Li, et. al. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE ICIP*, pp. 2309-2313. 2019.



- By local information geometry, given zero-mean feature $f(x)$, the optimal target loss is

$$L(f_s, W^\star) = Const(X, Y) - \underbrace{\mathscr{H}(f_s)}_{\text{H-score of } f(x)} + o(\epsilon^2)$$

$$\boxed{\mathscr{H}(f_s) = \mathrm{tr}\left(\mathrm{cov}(f_s(X))^{-1}\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f_s(X) \mid Y])\right)}$$

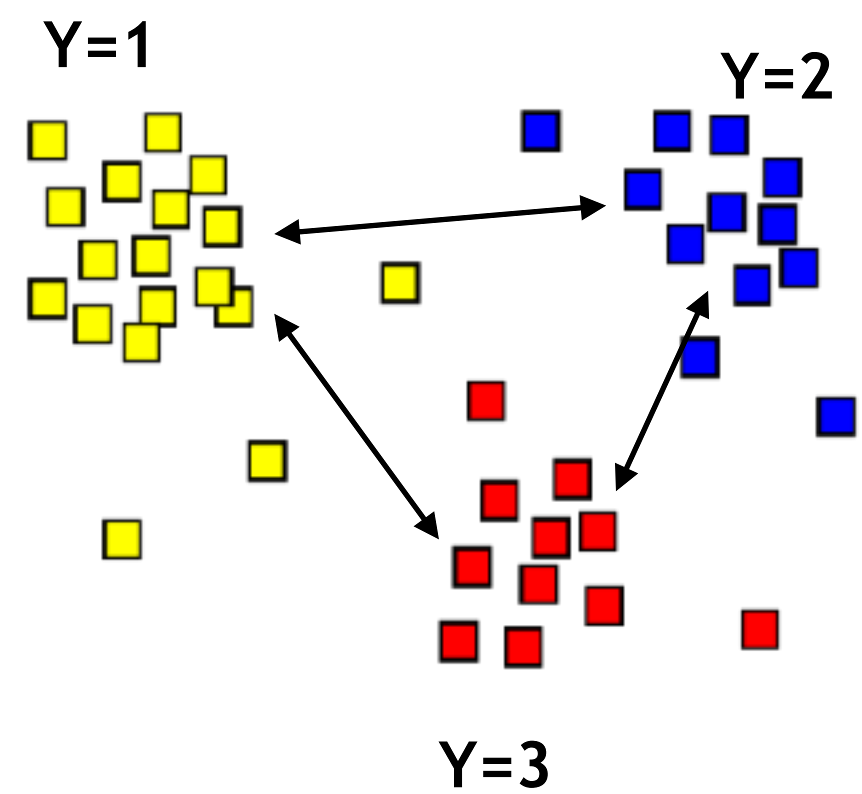
- Normalized H-score $\dfrac{\mathscr{H}(f_s)}{\mathscr{H}(f_t^\star)}$

Higher H-score $\leftrightarrow$ Better Performance

Only need to compute $\mathscr{H}(f_s)$ for source selection

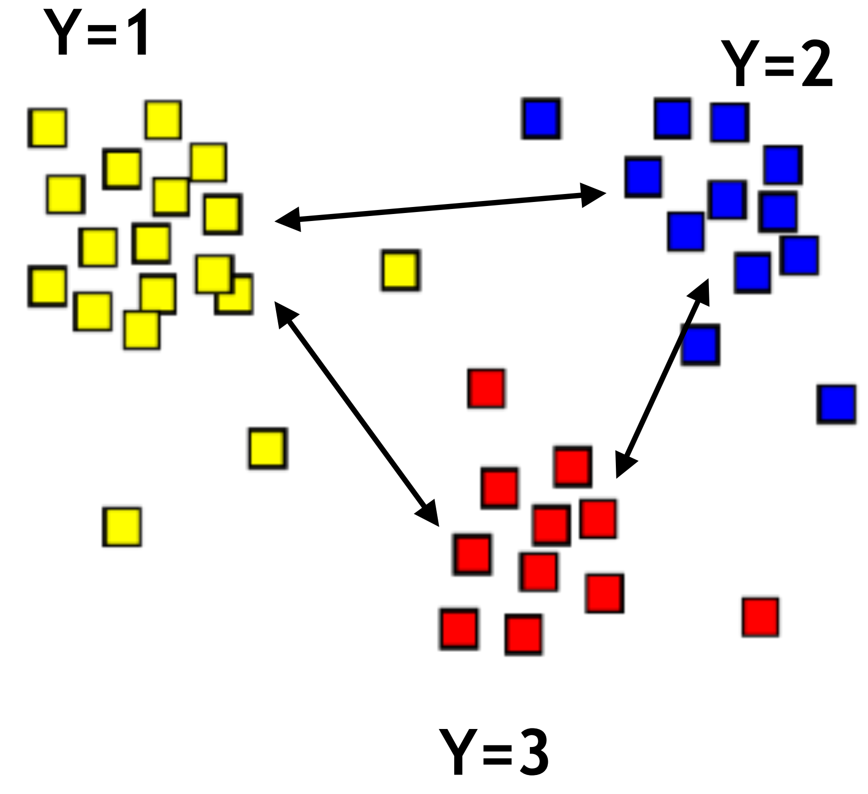
Intuitively, H-score minimizes feature redundancy and maximize intra-class distance.

$$\mathscr{H}(f) = \mathrm{tr}(\mathrm{cov}(f(X))^{-1}\mathrm{cov}(\mathbb{E}_{X|Y}[f(X) \mid Y]))$$



Intuitively, H-score minimizes feature redundancy and maximize intra-class distance.

$$\mathscr{H}(f) = \mathrm{tr}(\underbrace{\boxed{\mathrm{cov}(f(X))}}_{\text{feature redundancy}\downarrow}{}^{-1}\mathrm{cov}(\mathbb{E}_{X|Y}[f(X)\,|\,Y]))$$

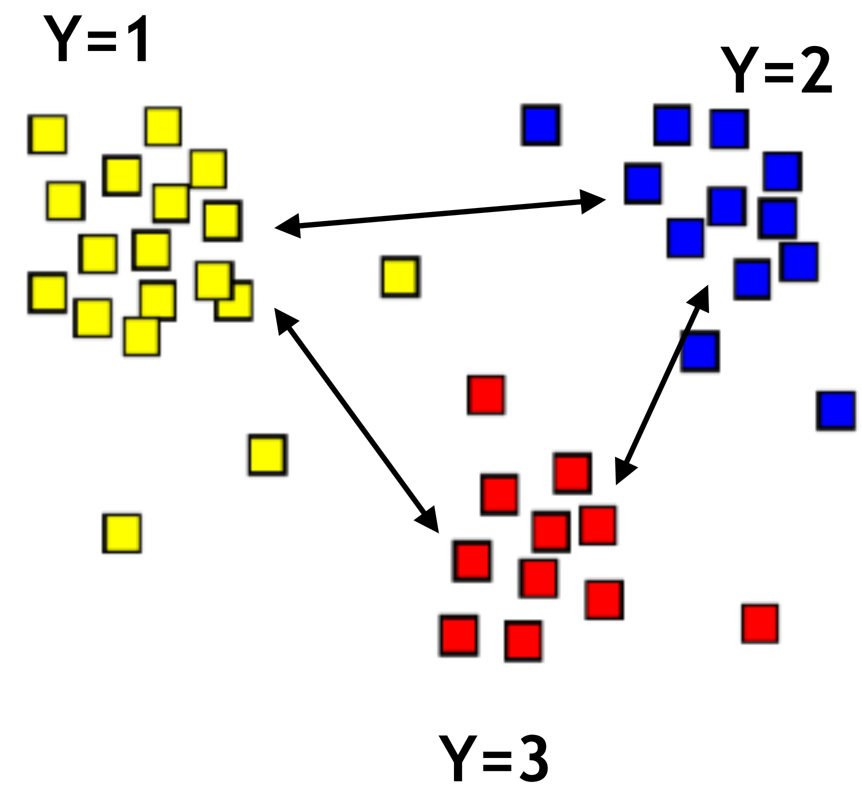


Intuitively, H-score minimizes feature redundancy and maximize intra-class distance.

$$\mathscr{H}(f) = \mathrm{tr}(\underbrace{\boxed{\mathrm{cov}(f(X))}^{-1}}_{\text{feature redundancy}\downarrow}\boxed{\mathrm{cov}(\mathbb{E}_{X|Y}[f(X)\,|\,Y])})$$

average inter-class distance↑

$$\mathbb{E}[\|\mathbb{E}[f(X)\,|\,Y]\|^2]$$



Intuitively, H-score minimizes feature redundancy and maximize intra-class distance.

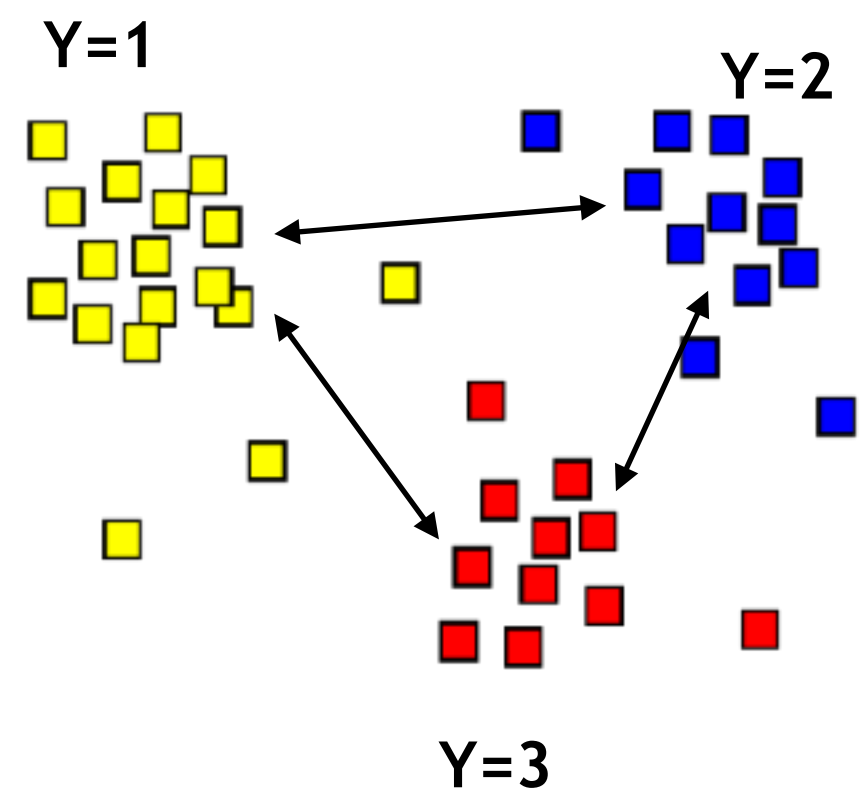
$$\mathscr{H}(f) = \mathrm{tr}\left(\boxed{\mathrm{cov}(f(X))}^{-1}\boxed{\mathrm{cov}(\mathbb{E}_{X|Y}[f(X) \mid Y])}\right)$$

H-score ↑

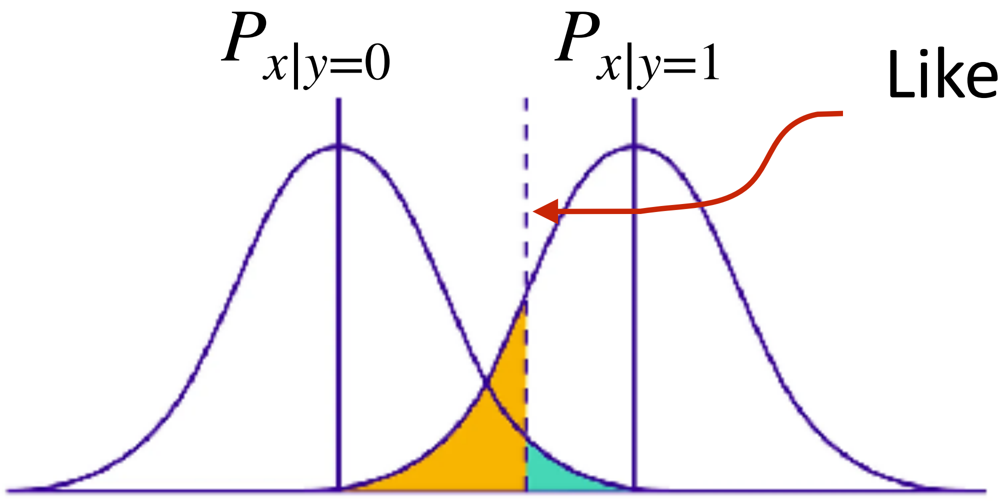
feature redundancy↓

average inter-class distance↑

$$\mathbb{E}[\|\mathbb{E}[f(X) \mid Y]\|^2]$$



Statistically, H-score characterizes the asymptotic error probability of the test statistics based on $f(X)$



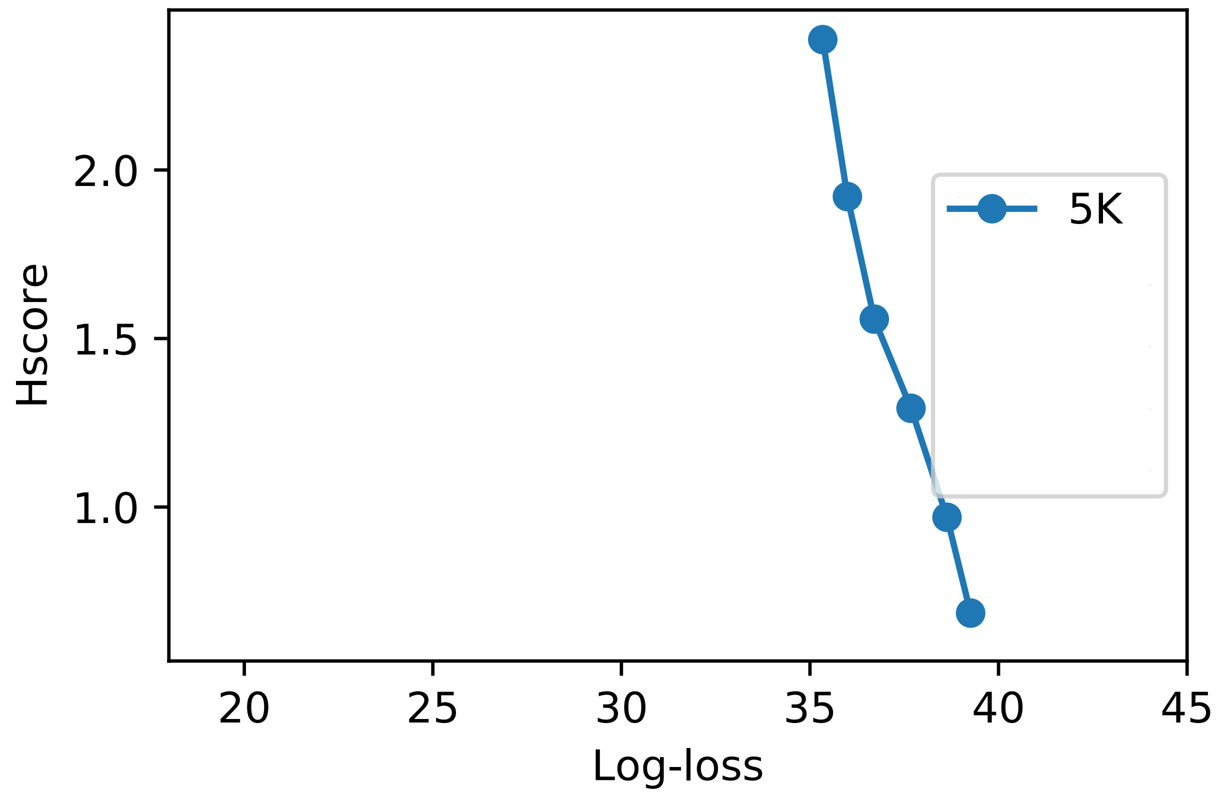
Likelihood ratio predictor
$\mathbb{E}[f(X)]$

Expected error probability

$P_e = \alpha\{\text{Type I Error}\} + \beta\{\text{Type II Error}\}$

Higher H-score $\leftrightarrow$ Faster error decay with increasing sample size

H-Score is **negatively correlated with target log-loss** on ImageNet (Resnet50) -> Cifar100, under different training size

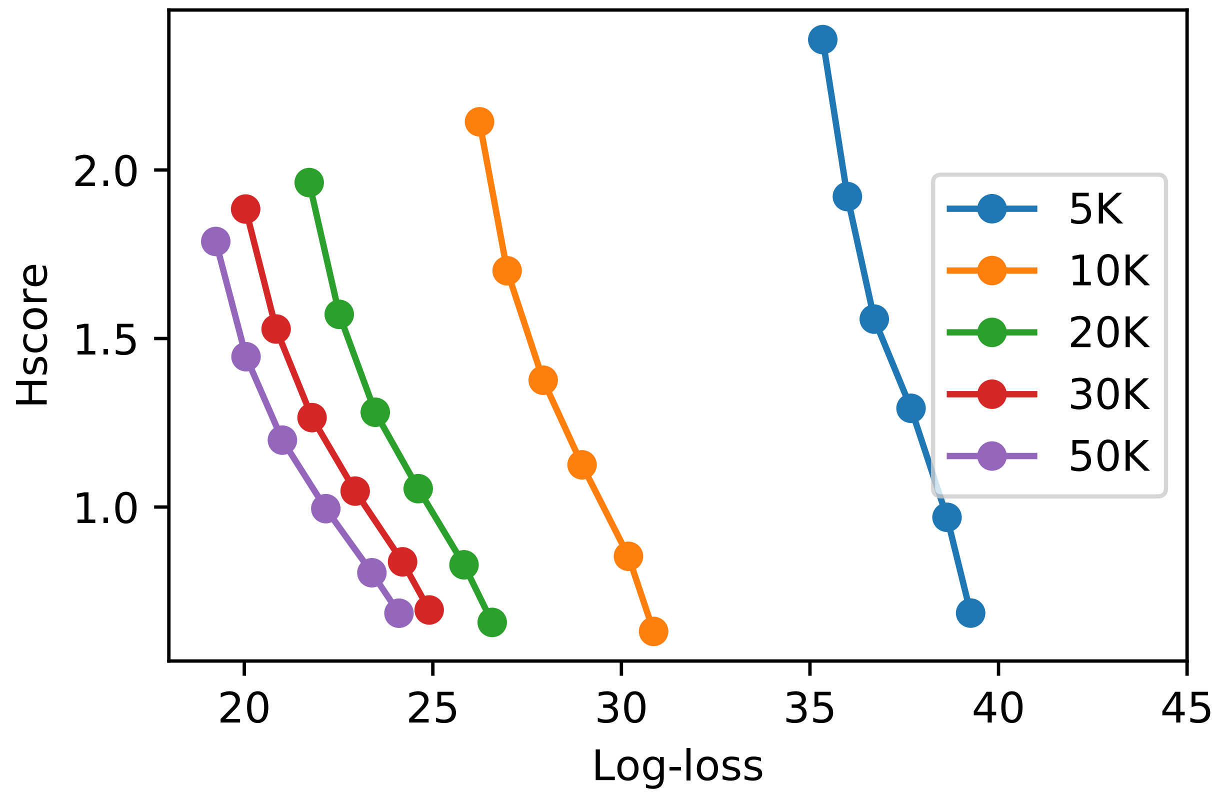


- 6 Source models: Layers 4a - 5f in ResNet50
- Target dataset: Cifar 100-class classification on 5K, 10K, …, 50K images

…

Validates our claim $L(f, \theta^\star) = Const(X, Y) - \mathscr{H}(f) + o(\epsilon^2)$

H-Score is **negatively correlated with target log-loss** on ImageNet (Resnet50) -> Cifar100, under different training size

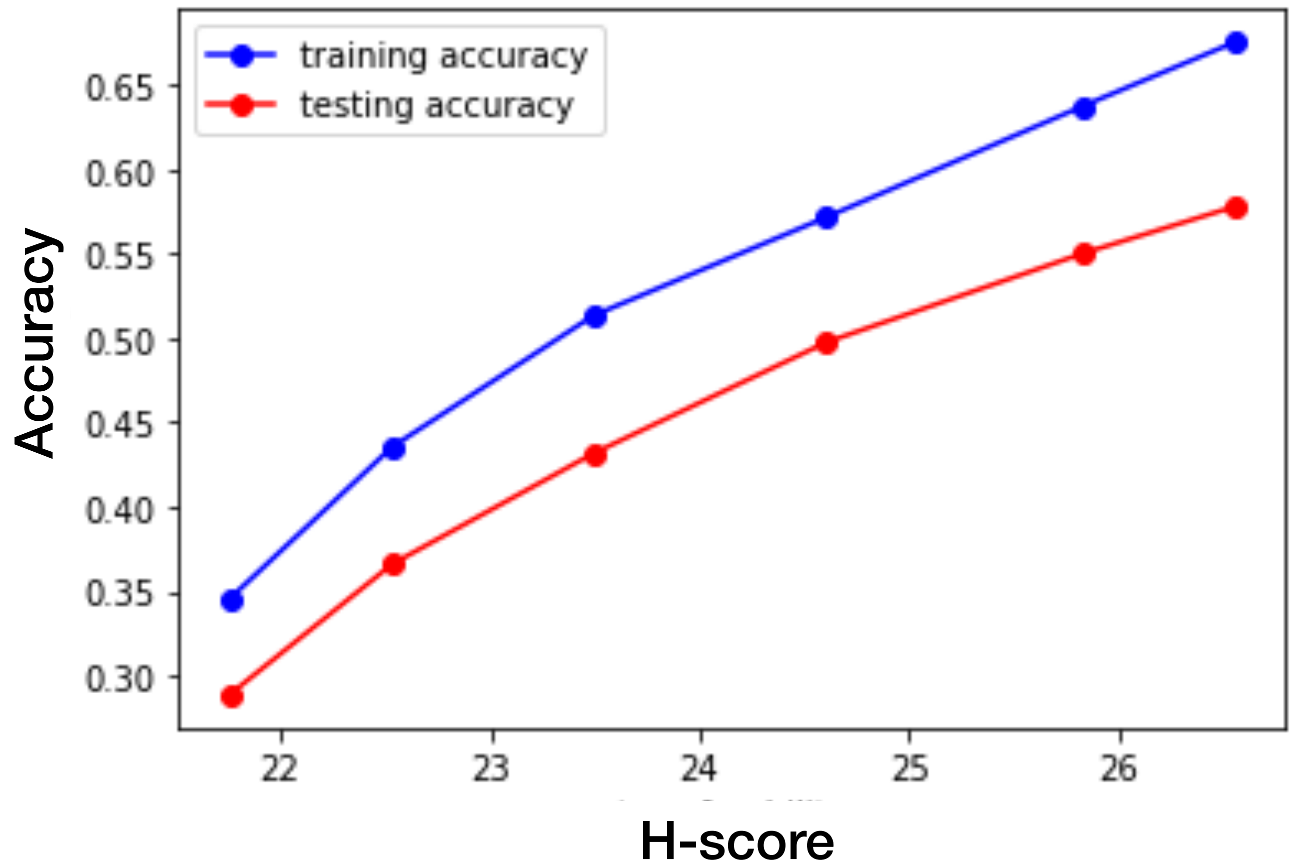


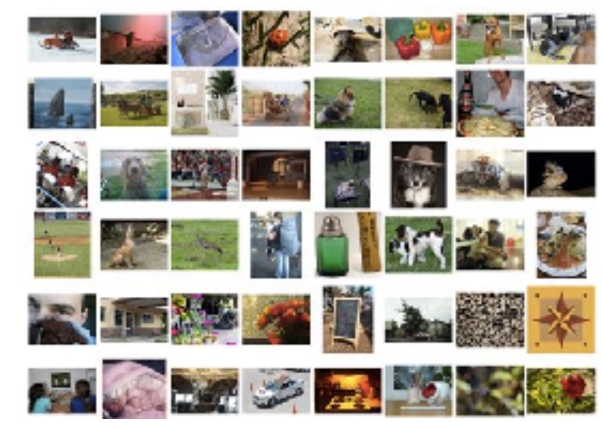
- 6 Source models: Layers 4a - 5f in ResNet50
- Target dataset: Cifar 100-class classification on 5K, 10K, …, 50K images

…

Validates our claim $L(f,\theta^{\star}) = Const(X,Y) - \mathscr{H}(f) + o(\epsilon^2)$

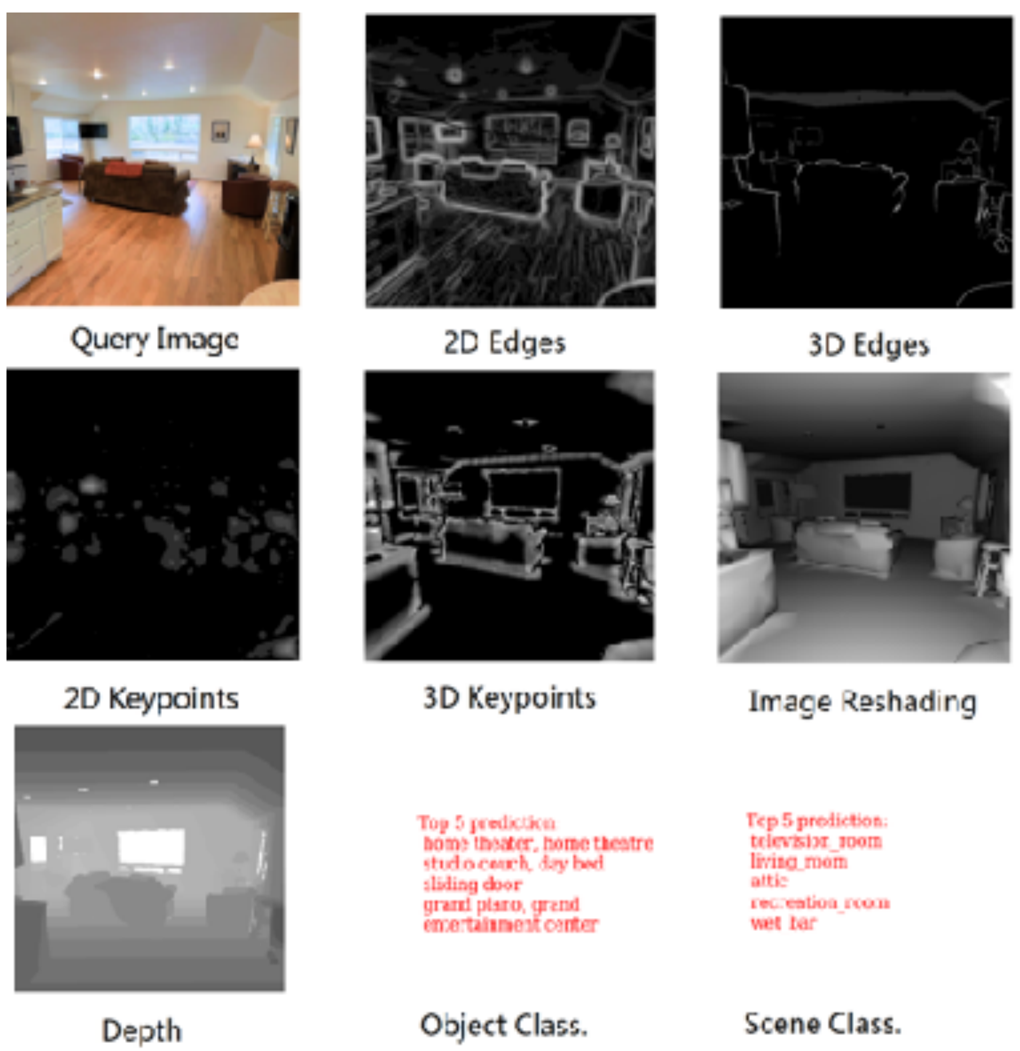
H-Score is **positively correlated with target training & testing accuracy**



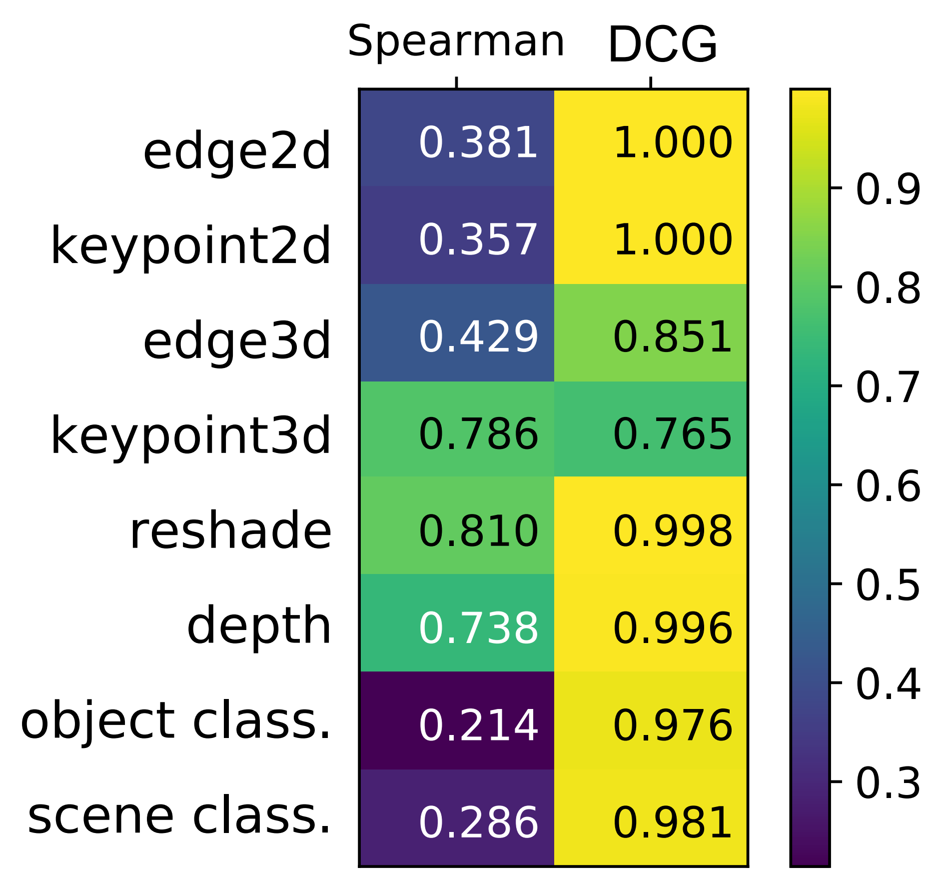


Target sample size: 20,000

On Taskonomy Benchmark, H-Score is positively correlated with empirical-based transferability with **~6x speedup**

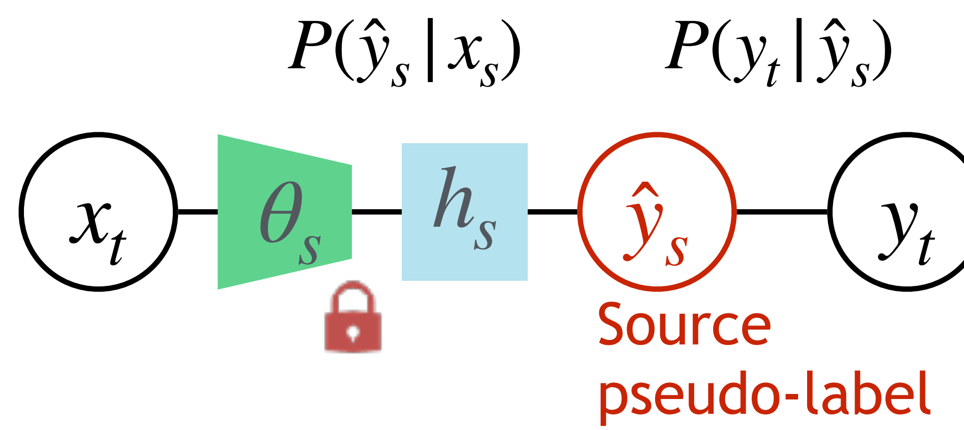


Ranking correlation with Task Affinity
(Zamir 2018)



H-Score also has known limitations,

- numerical instability
- regression tasks (LEEP by Nguyen et. al. 2020)
- same-domain assumption



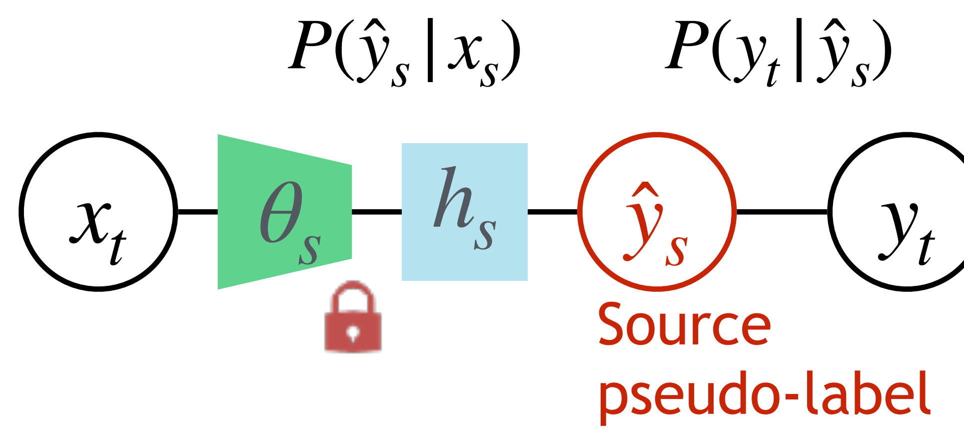
$$LEEP = \frac{1}{n}\sum_{i=1}^{n} \log P(y_t^i \mid x_t^i)$$

Task transferability (H-Score, LEEP..) assumes source and target task has the same input distribution $P_s(x) = P_t(x)$

Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and 952 Matthias Seeger. Leep: A new measure to evaluate transferability of learned representations.ICML, 2020.

Cross-domain Cross-task Transferability

- numerical instability
- regression tasks (LEEP by Nguyen et. al. 2020)
- same-domain assumption



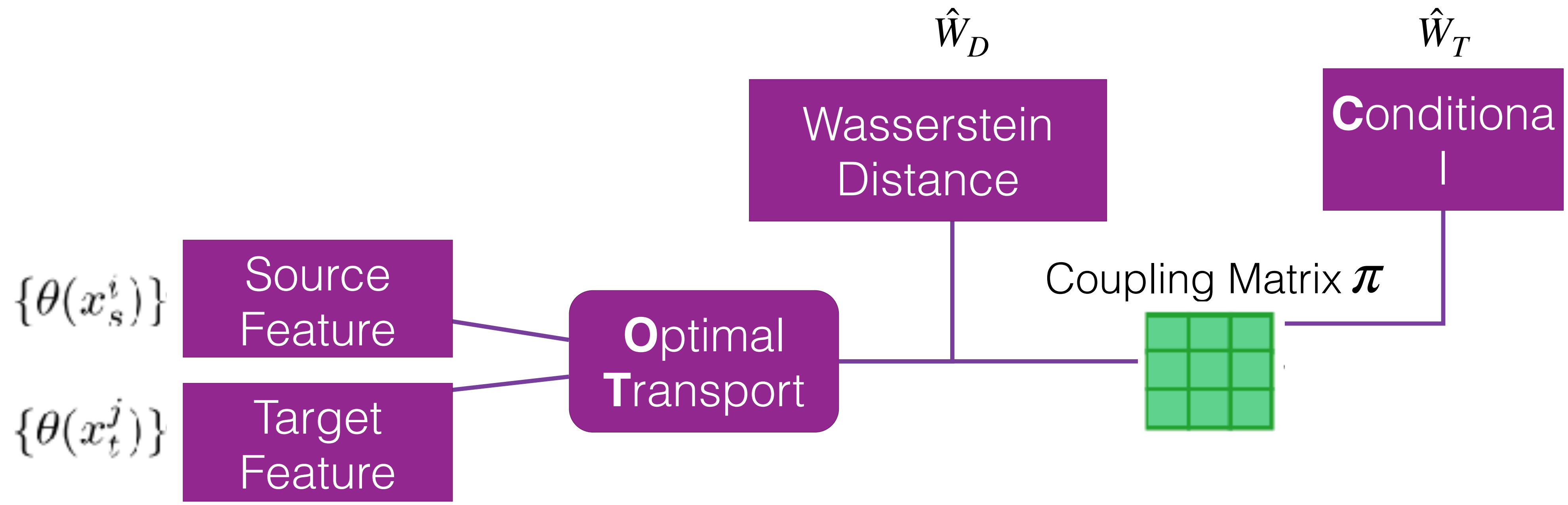
$$LEEP = \frac{1}{n} \sum_{i=1}^{n} \log P(y_t^i | x_t^i)$$

Task transferability (H-Score, LEEP..) assumes source and target task has the same input distribution $P_s(x) = P_t(x)$

Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and 952 Matthias Seeger. Leep: A new measure to evaluate transferability of learned representations.ICML, 2020.

OTCE： Cross-domain Cross-task Transferability

Decompose “transfer hardness” into domain difference and task difference

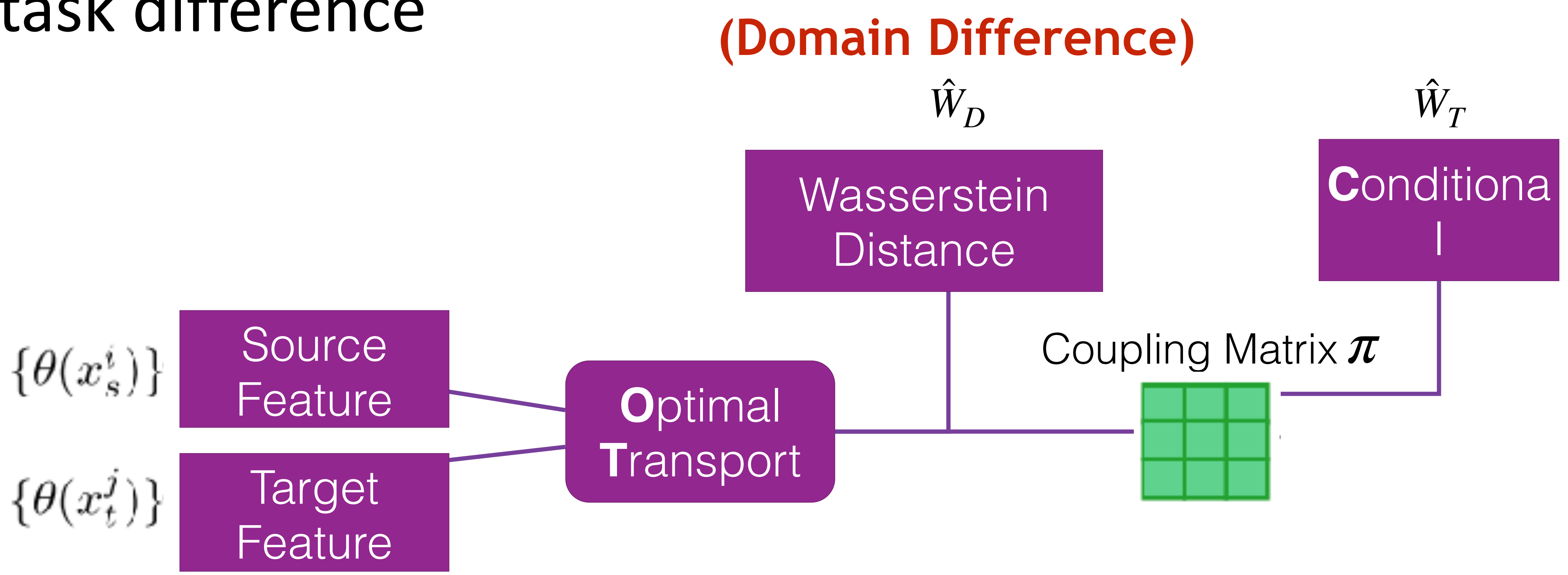


$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b$$

Yang Tan, Yang Li*, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." CVPR， 2021

OTCE： Cross-domain Cross-task Transferability

Decompose “transfer hardness” into domain difference and task difference

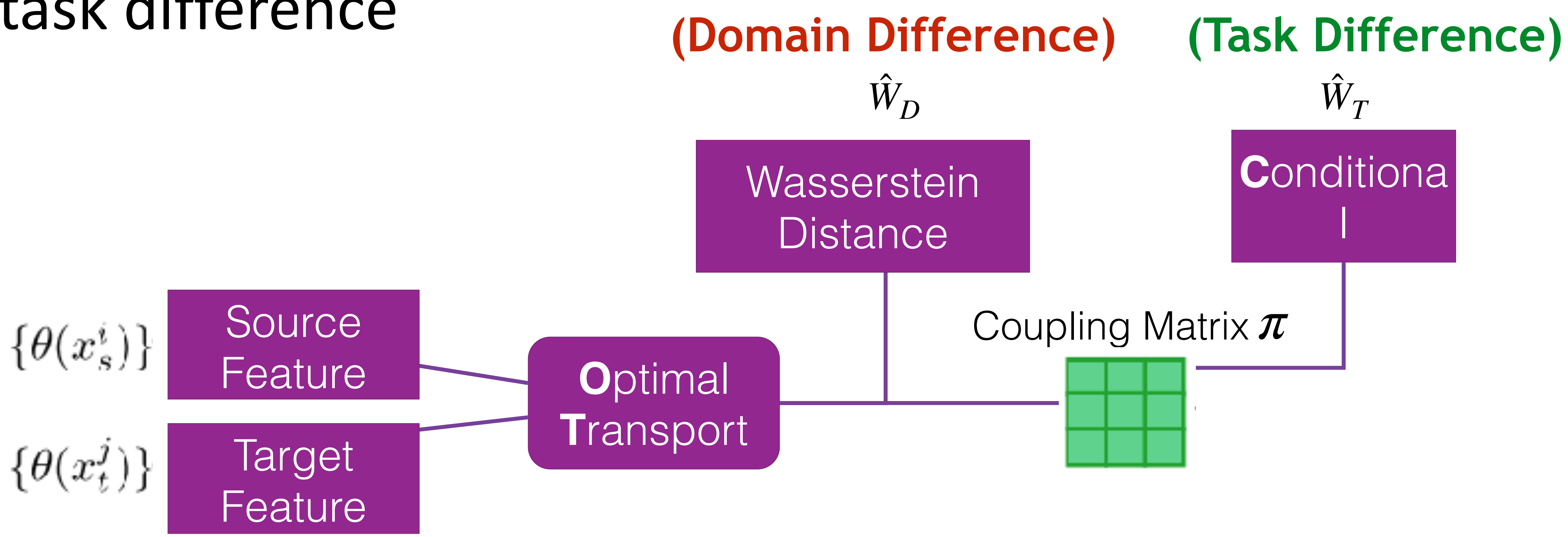


$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b$$

Yang Tan, Yang Li*, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." CVPR， 2021

OTCE： Cross-domain Cross-task Transferability

Decompose “transfer hardness” into domain difference and task difference

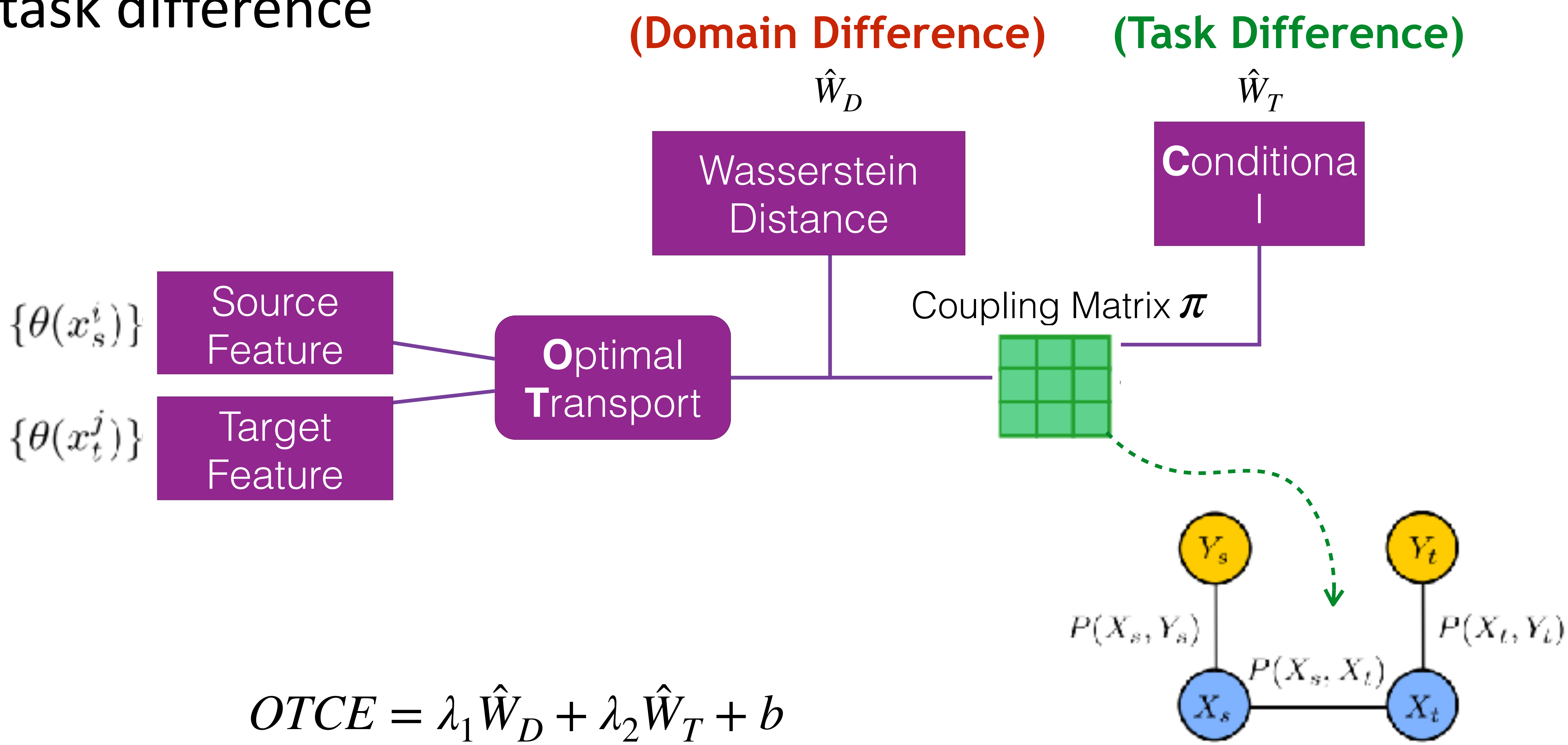


$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b$$

Yang Tan, Yang Li*, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." CVPR，2021

OTCE： Cross-domain Cross-task Transferability

Decompose “transfer hardness” into domain difference and task difference

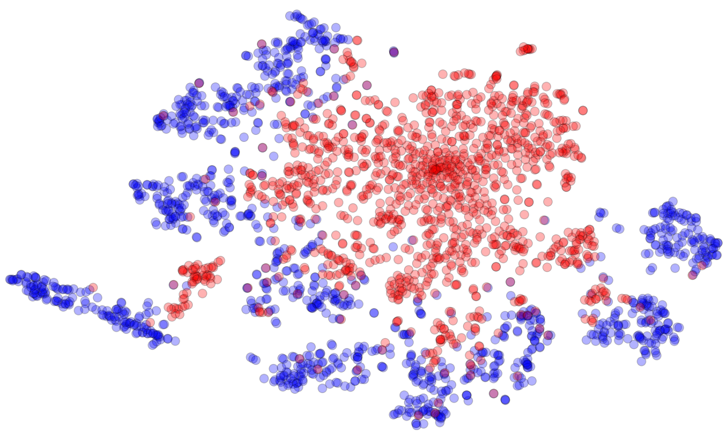


$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b$$

Yang Tan, Yang Li*, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." CVPR， 2021

Outline

- What's Transfer Learning
- Traditional transfer learning algorithms
  - Task transfer learning
  - Domain adaptation
  - Transfer bound on domain adaptation
- When to transfer?
  - Transferability estimation
- Research trends



Beyond Transfer Learning

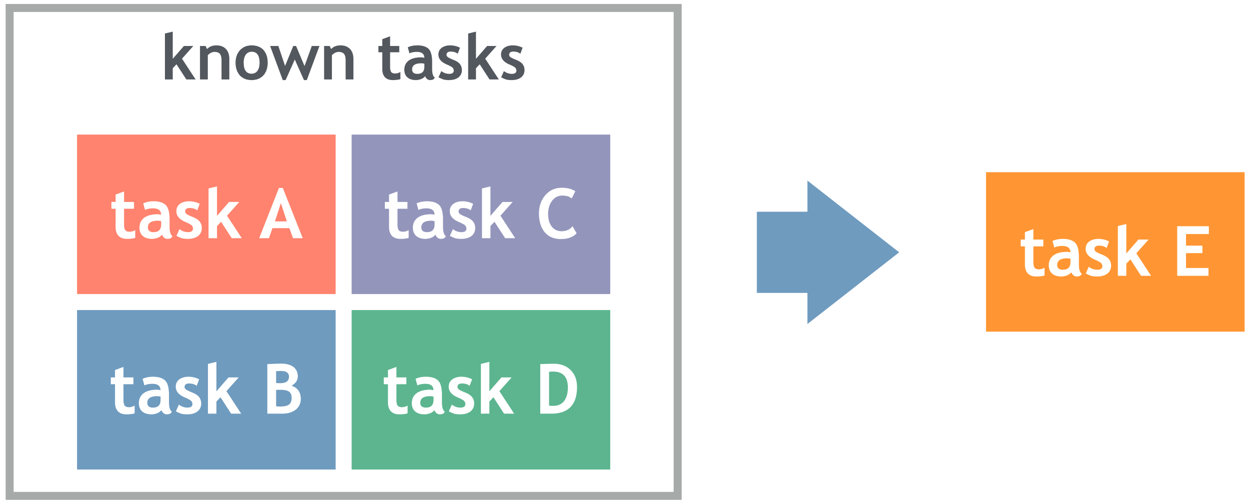
- **Multi-source transfer learning**: how to efficiently, adaptively combine features from multiple source tasks in transfer learning?
- **Continuous domain adaptation**: leverage intermediate domains to adapt model to distant target tasks

Beyond Transfer Learning

- **Multi-source transfer learning**: how to efficiently, adaptively combine features from multiple source tasks in transfer learning?

- **Continuous domain adaptation**: leverage intermediate domains to adapt model to distant target tasks

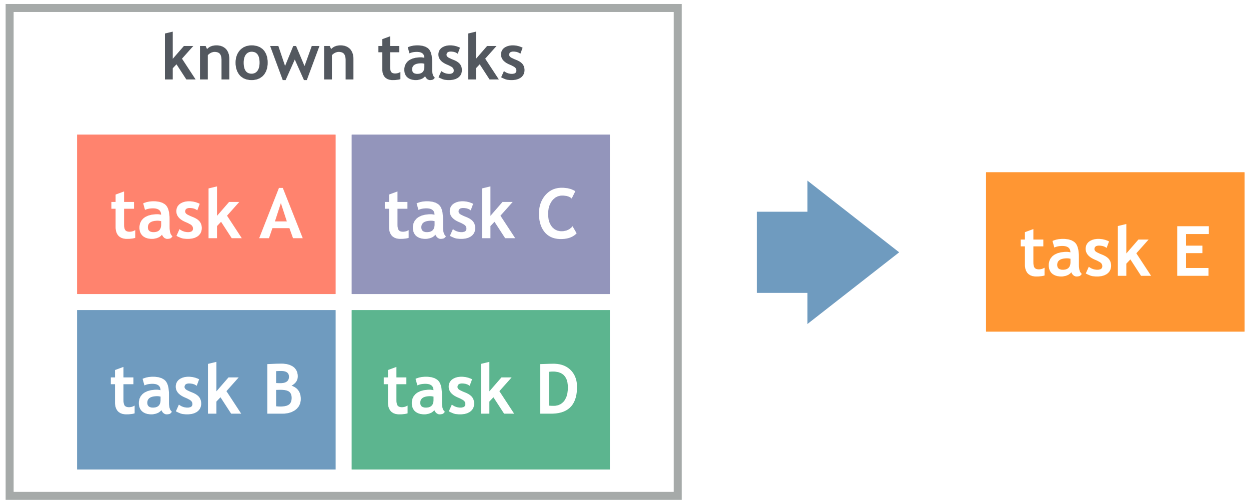
- **Meta learning/Domain generalization:** given data/experience on previous tasks/domains, learn a *generalizable* model for a new task/domain



Beyond Transfer Learning

- **Multi-source transfer learning**: how to efficiently, adaptively combine features from multiple source tasks in transfer learning?
- **Continuous domain adaptation**: leverage intermediate domains to adapt model to distant target tasks

- **Meta learning/Domain generalization:** given data/experience on previous tasks/domains, learn a *generalizable* model for a new task/domain



Transfer learning using foundation models

New Challenges for transferring from foundation models

- **Zero-shot/Few-shot** adaptation
- Full update is too slow: parameter-efficient model adaptation
- No access to source data: Source data free model selection
- New transfer paradigms
    - Transfer attention-maps for Vision Transformer
    - Prompt tuning