An Overview of Transfer Learning

with an emphasis on domain adaptation

Archer Gong Zhang
Department of Statistical Sciences
University of Toronto

@Learning Reading Group
Nov 9, 2022

Main references

This presentation is based on some popular survey papers in the literature.

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

A Survey on Transfer Learning

Sinno Jialin Pan and Qiang Yang, *Fellow*, *IEEE*

CONTRIBUTED PAPER

A Comprehensive Survey on Transfer Learning

This survey provides a comprehensive understanding of transfer learning from the perspectives of data and model.

By Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, *Senior Member IEEE*, Hui Xiong, *Fellow IEEE*, and Qing He

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 3, MARCH 2021

A Review of Domain Adaptation without Target Labels

Wouter M. Kouw and Marco Loog

Outline

- What is transfer learning?
- When can transfer learning be useful?
- How does transfer learning work?

What is transfer learning, in a rough sense?

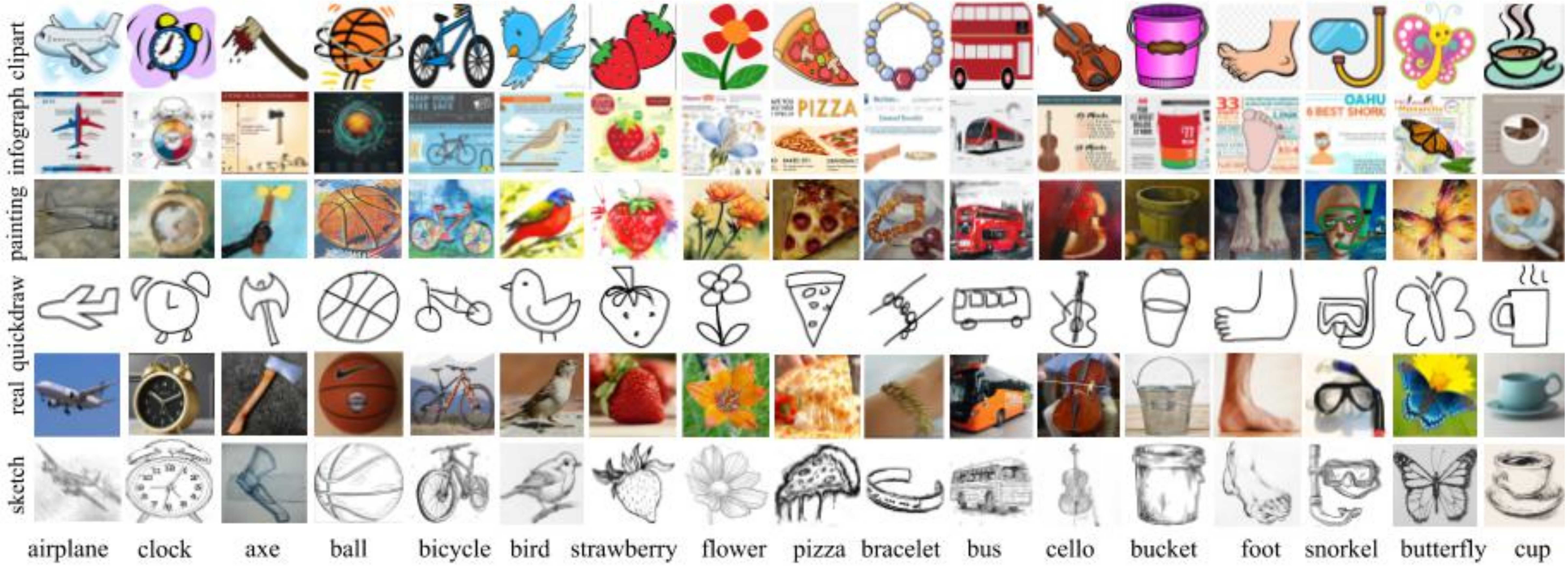
Transfer learning

In a rough sense

Wikipedia: “**Transfer learning** is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.”

Motivating examples

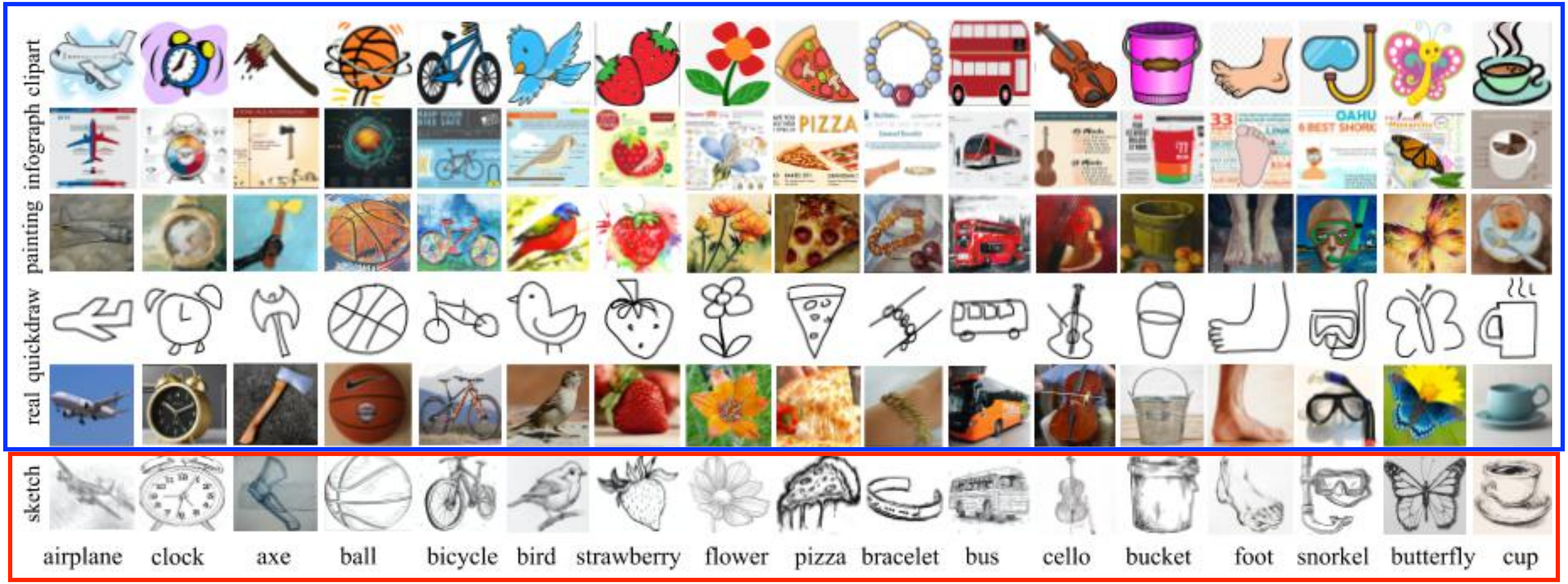
DomainNet: an image dataset of common objects in six different domains



Peng, Xingchao, et al. "Moment matching for multi-source domain adaptation." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

Motivating examples

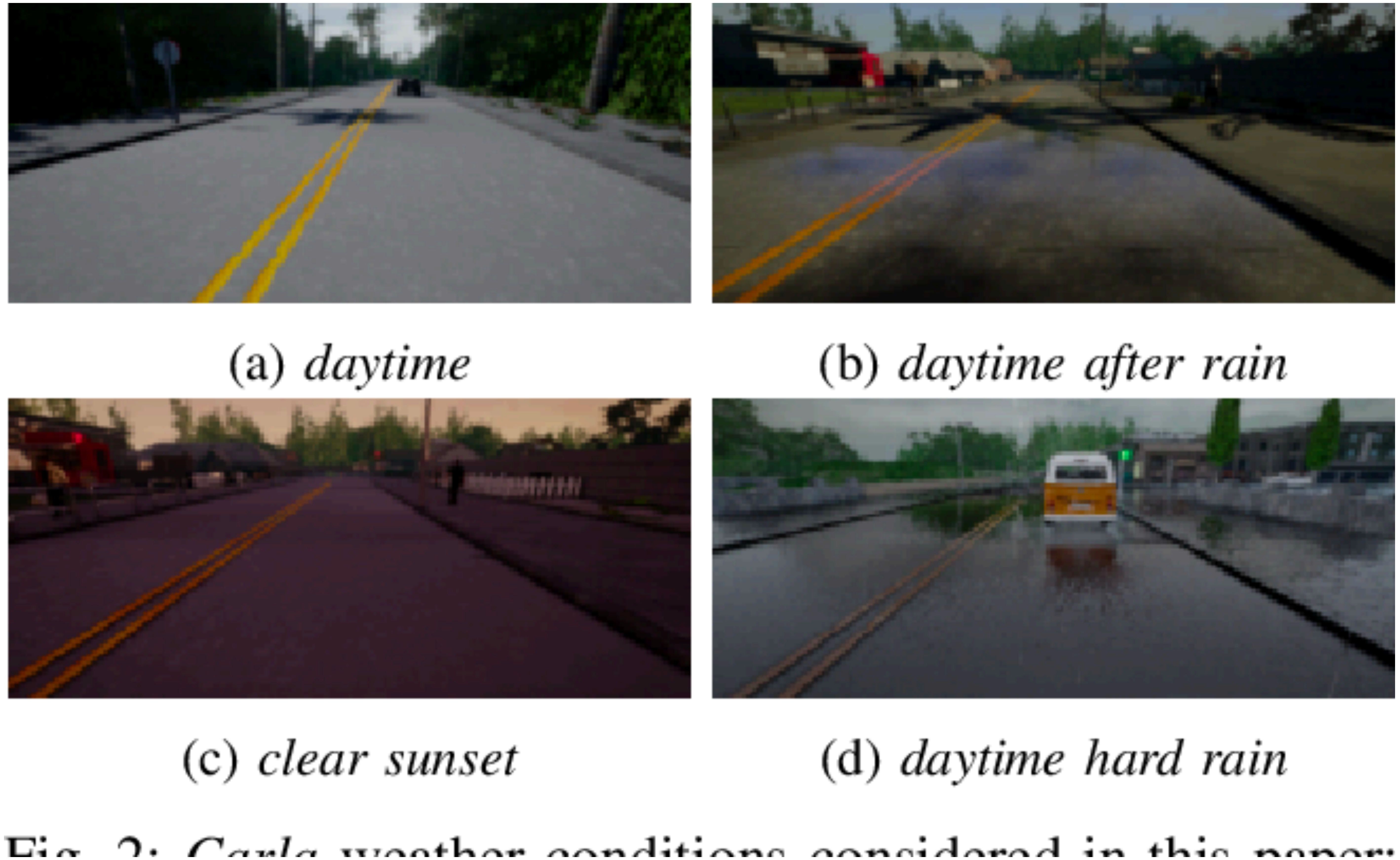
DomainNet: an image dataset of common objects in six different domains

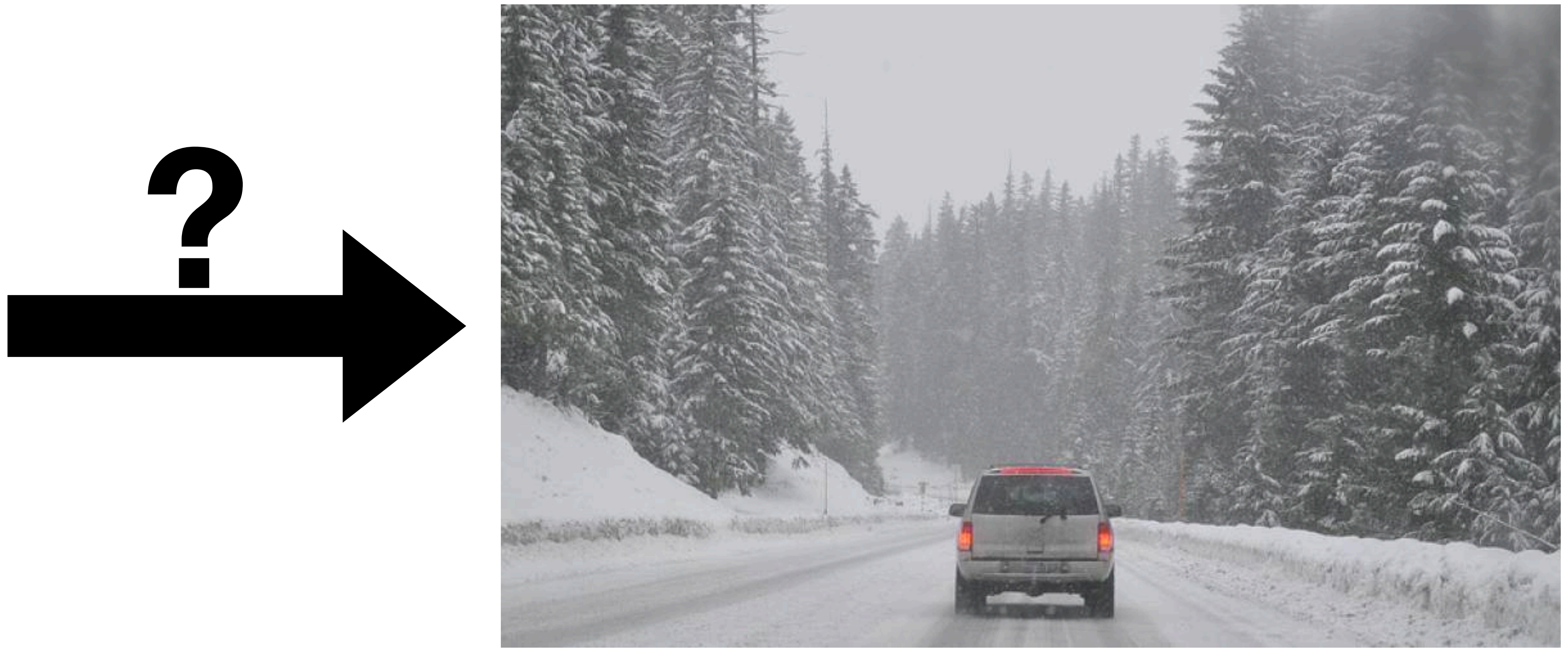


Peng, Xingchao, et al. "Moment matching for multi-source domain adaptation." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

Motivating examples

Visual-based autonomous driving development in different training conditions.

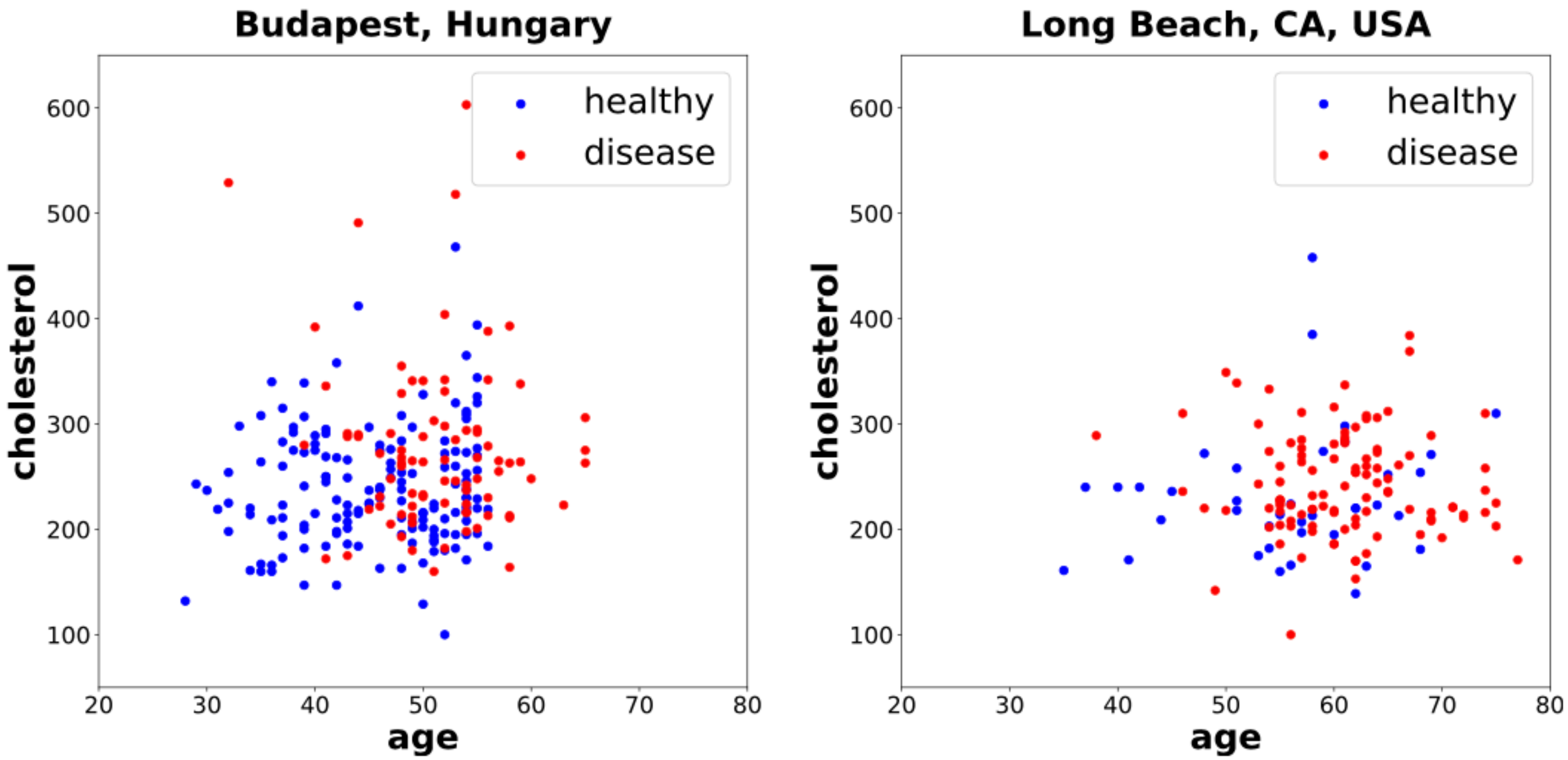




Tai, Lei, et al. "Visual-based autonomous driving deployment from a stochastic and uncertainty-aware perspective." *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

Motivating examples

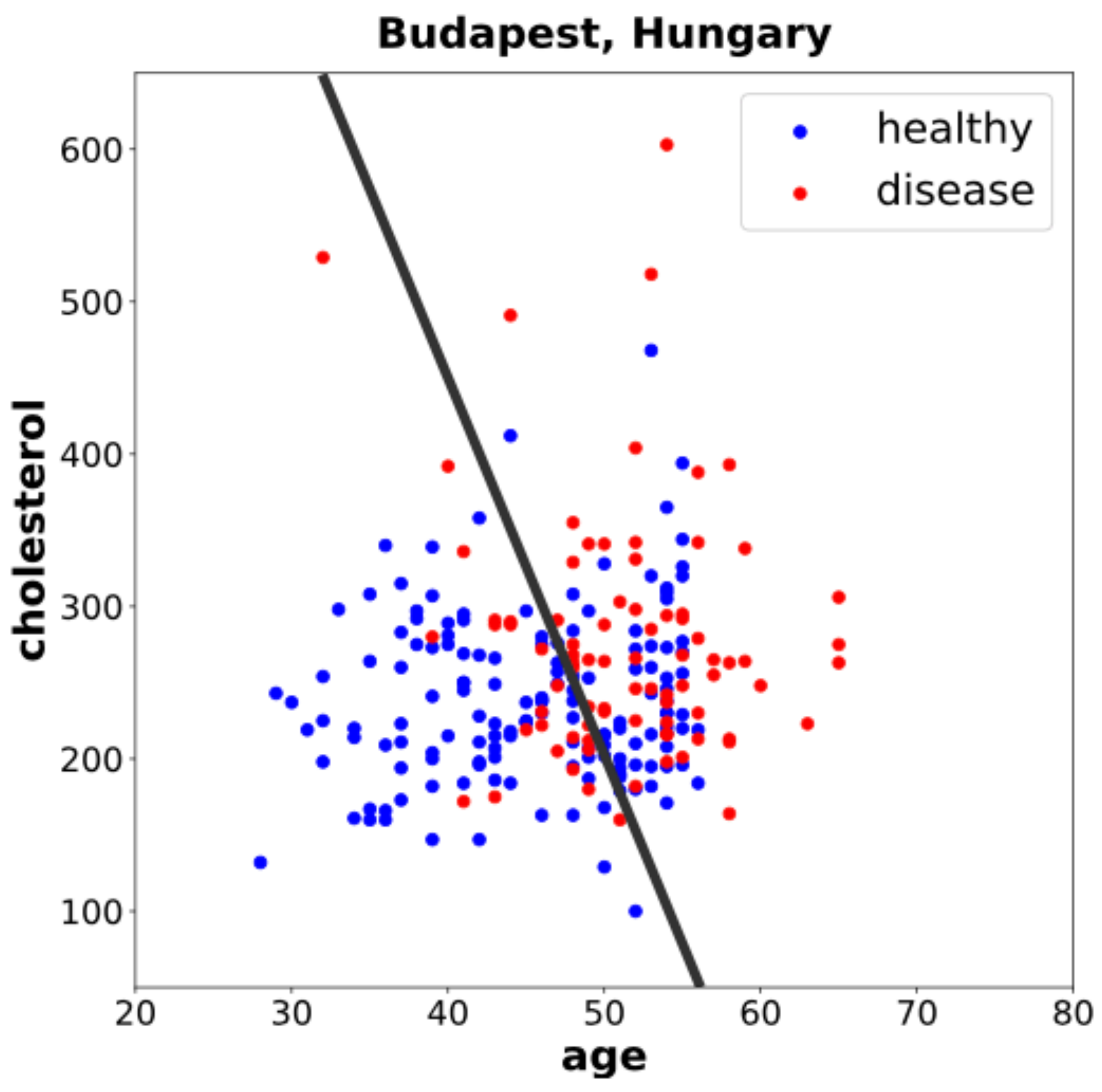
Heart disease diagnosis based on age & cholesterol

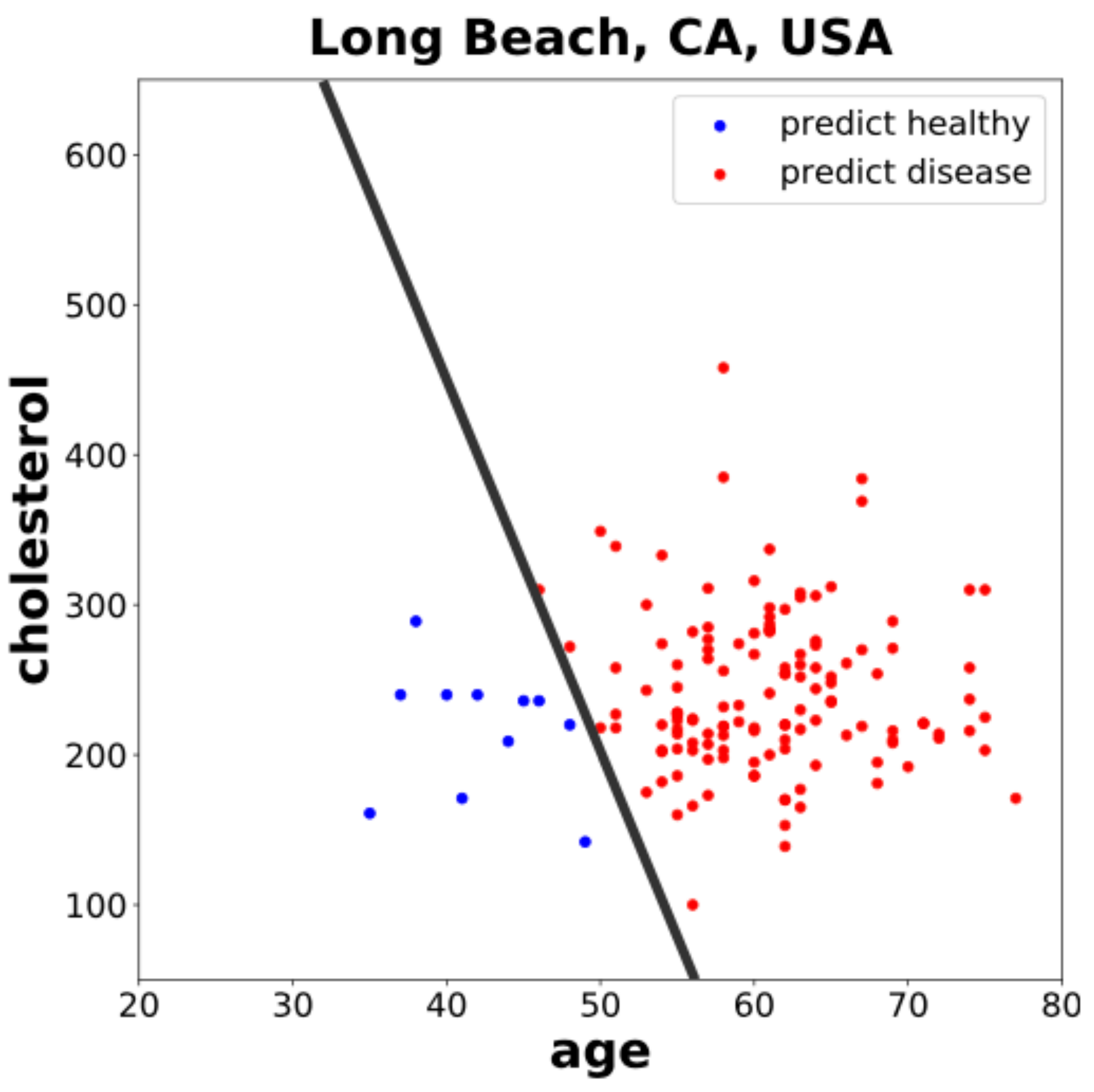


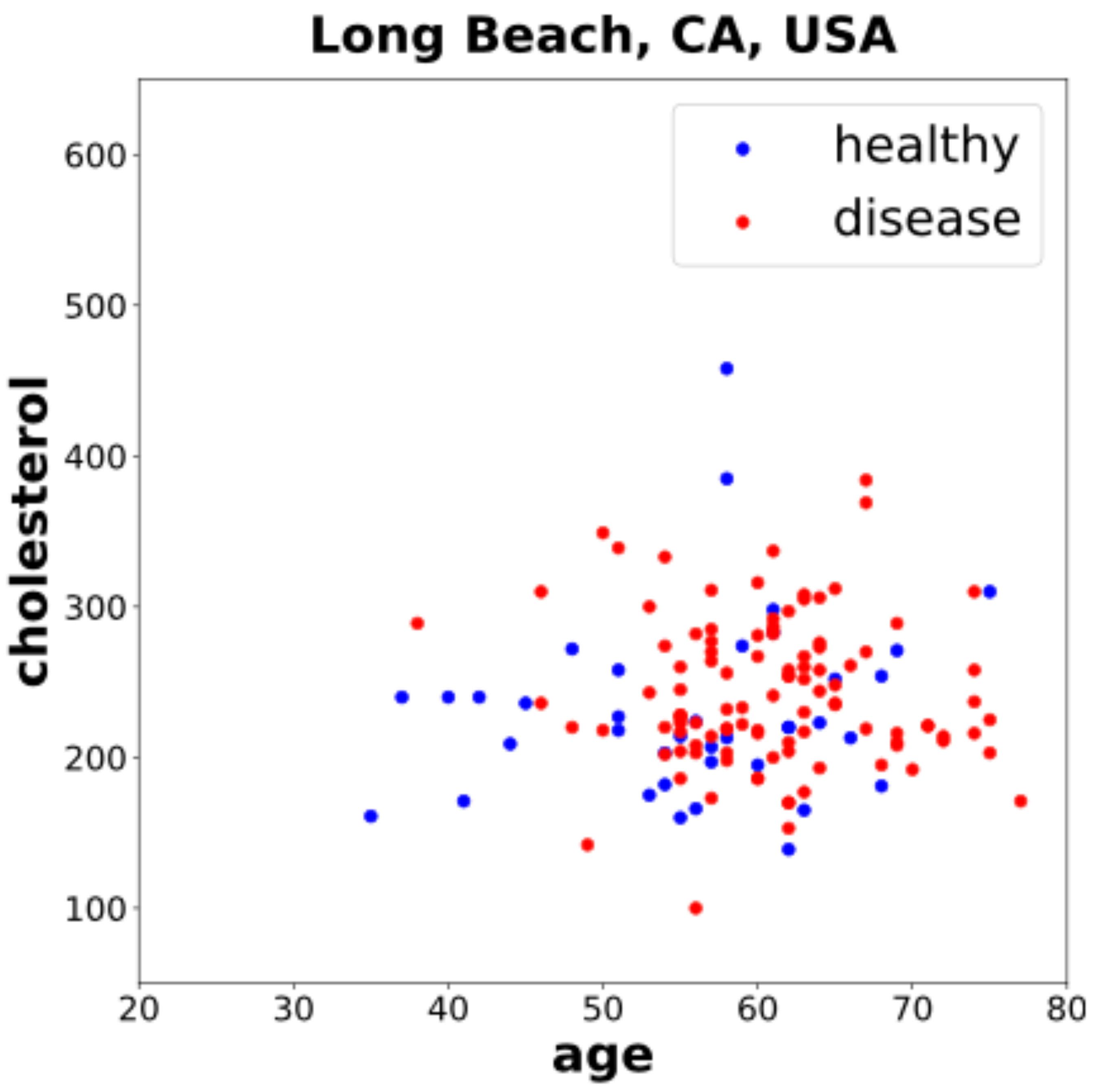
Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Motivating examples

Heart disease diagnosis based on age & cholesterol







Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

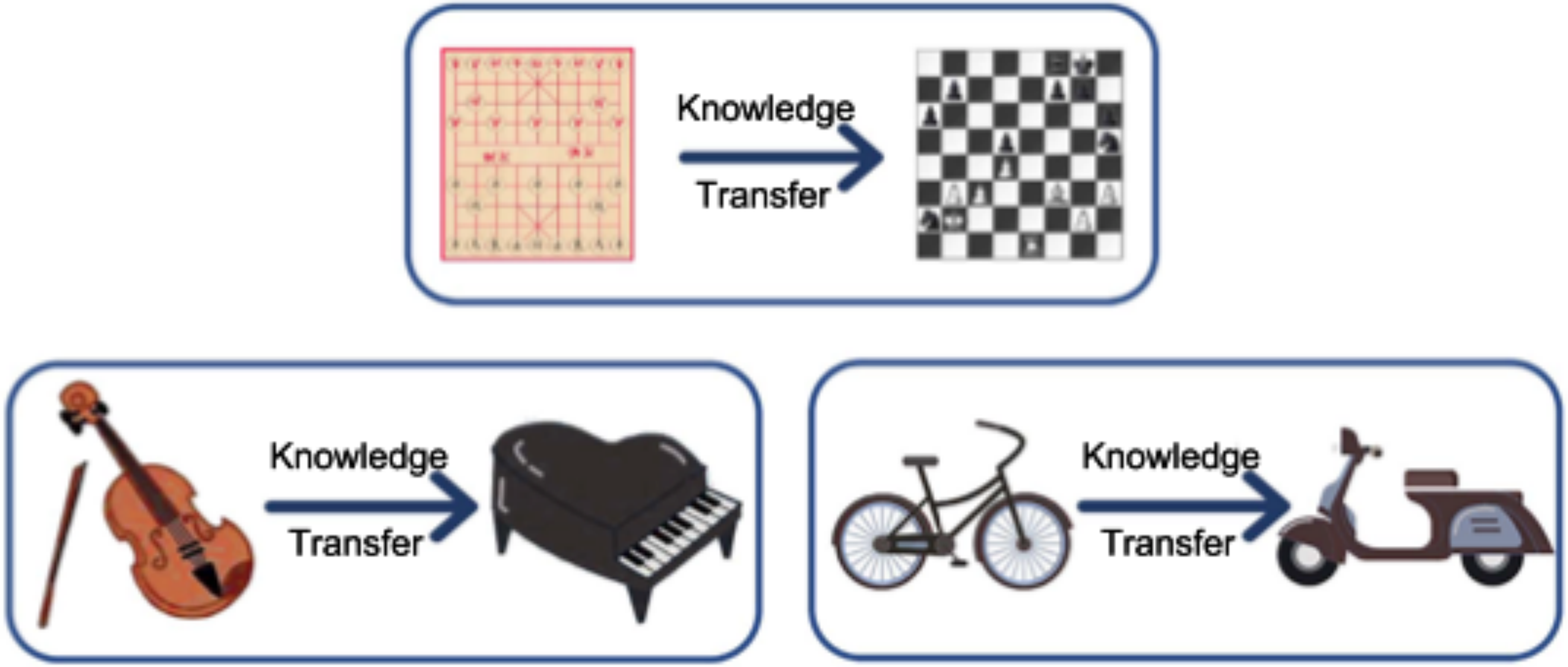
When can transfer learning be useful?

A "successful" transfer

- The concept of transfer learning may initially come from educational psychology.
- A psychologist C. H. Judd: *learning to transfer is the result of the generalization of experience*. It is possible to realize the transfer from one situation to another, as long as a person generalizes his experience.
- According to this theory, the prerequisite of transfer is that there needs to be a "connection" between two learning activities.

Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Examples of successful transfer



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

~~When can transfer learning be useful?~~

Can/when can transfer learning be harmful?

Negative transfer

An “unsuccessful” transfer

- Happens when the knowledge of source domain contributes to the reduced performance of learning in the target domain.

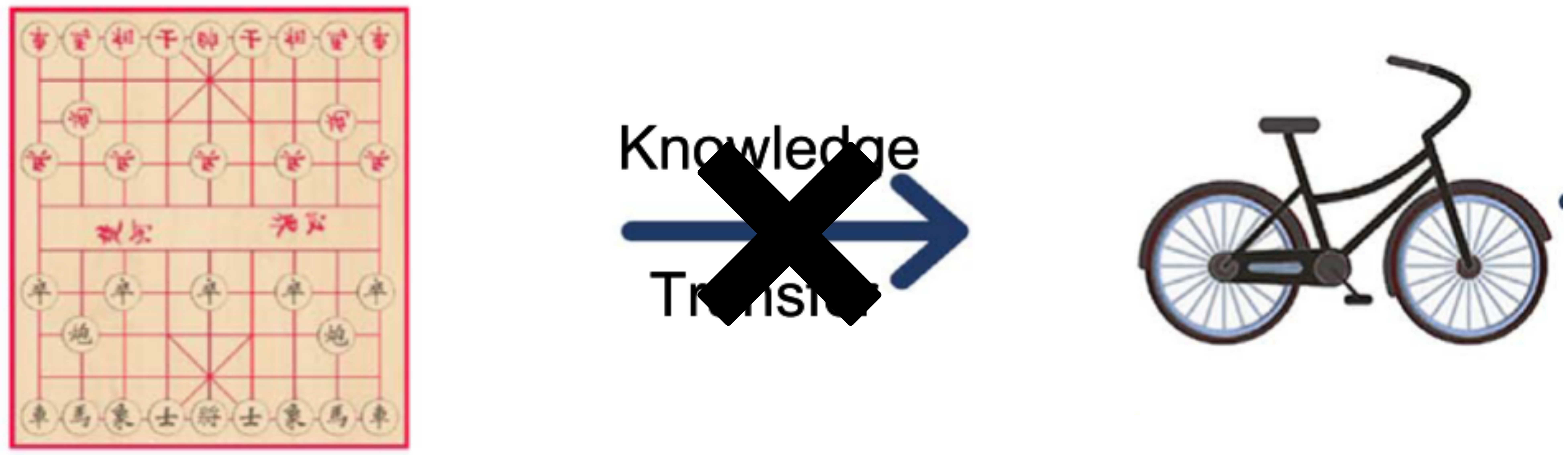
- This could probably happen when
  - two domains/tasks are too dissimilar: brute-force transfer may even hurt
  - the similarities between domains do not always facilitate learning: sometimes the similarities may be misleading

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Examples of unsuccessful transfer

Dissimilar domains/tasks



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Examples of negative transfer

Misleading similarities

Previous successful experience in Spanish can interfere with learning the word formation, usage, pronunciation, conjugation, and so on, in French.

Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

(Pause for questions.)

How does transfer learning work?

Basic settings

- Input/feature space $\mathcal{X} \subseteq \mathbb{R}^D$, with data $X = \{x_i \in \mathcal{X} : i = 1,\ldots,n\}$
- Marginal distribution $P(X)$
- Domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$

$\mathcal{D}_S$ for source domain

$\mathcal{D}_T$ for target domain

- Label space $\mathcal{Y}$: either binary or multi-class, with data $\{y_i \in \mathcal{Y} : i = 1,\ldots,n\}$
- Decision/predictive function $f(\cdot) : \mathcal{X} \to \mathbb{R}$; usually viewed as $P(y \mid \cdot)$ for $y \in \mathcal{Y}$
- Task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

$\mathcal{T}_S$ for source domain

$\mathcal{T}_T$ for target domain

“Definition” of Transfer Learning

Definition 1 (Transfer Learning). *Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$,* transfer learning *aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.*

Key 1: rather than learning all of the source and target tasks simultaneously (i.e., multitask learning), transfer learning cares most about the target task.

The roles of the source and target tasks are no longer symmetric in transfer learning.

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

"Definition" of Transfer Learning

Definition 1 (Transfer Learning). *Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$,* transfer learning *aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.*

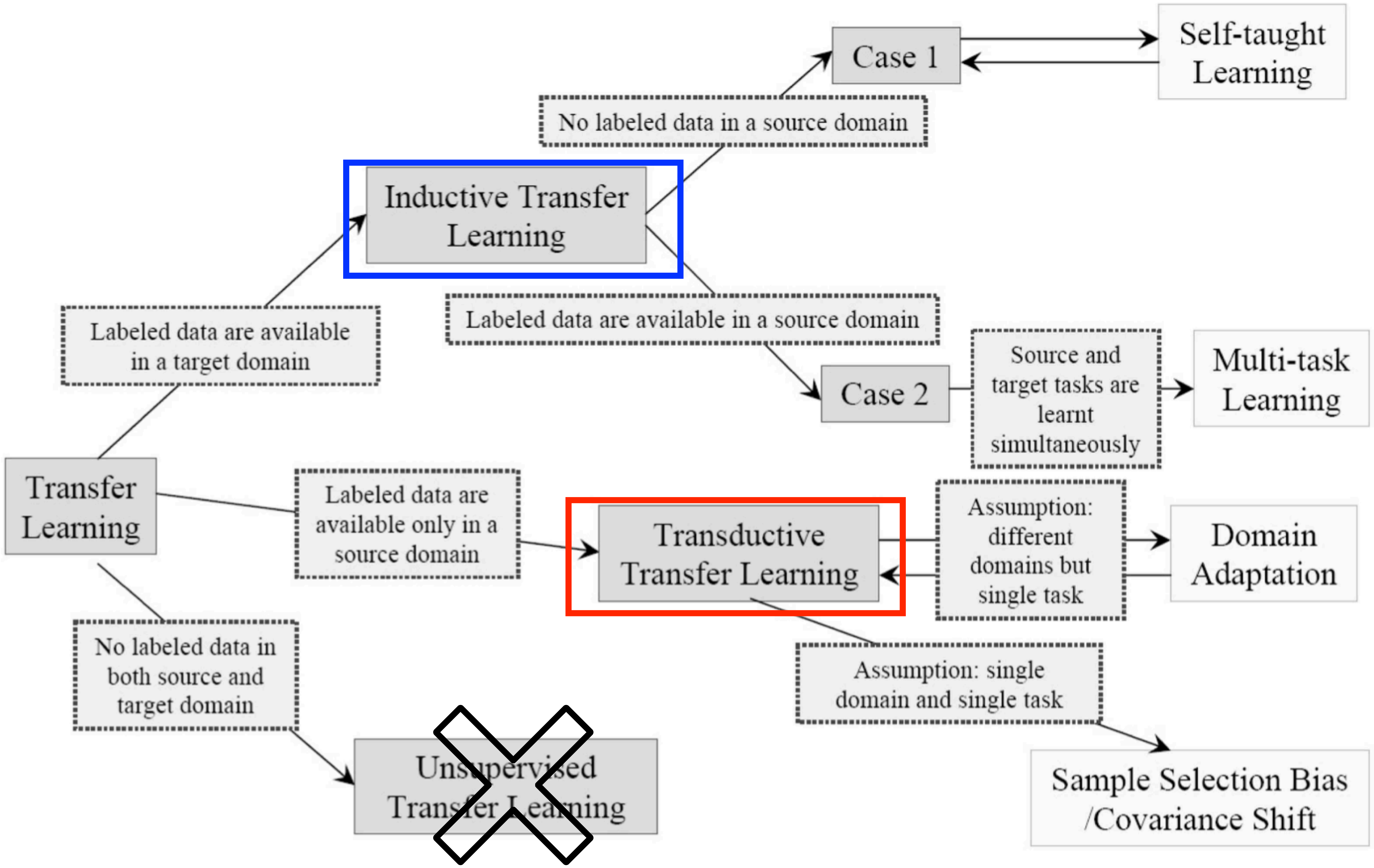
Key 2: at least one of the two pairs $(\mathcal{D}_S, \mathcal{D}_T)$ and $(\mathcal{T}_S, \mathcal{T}_T)$ differs!

This scenario distinguishes transfer learning from traditional machine learning.

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Categorization of Transfer Learning

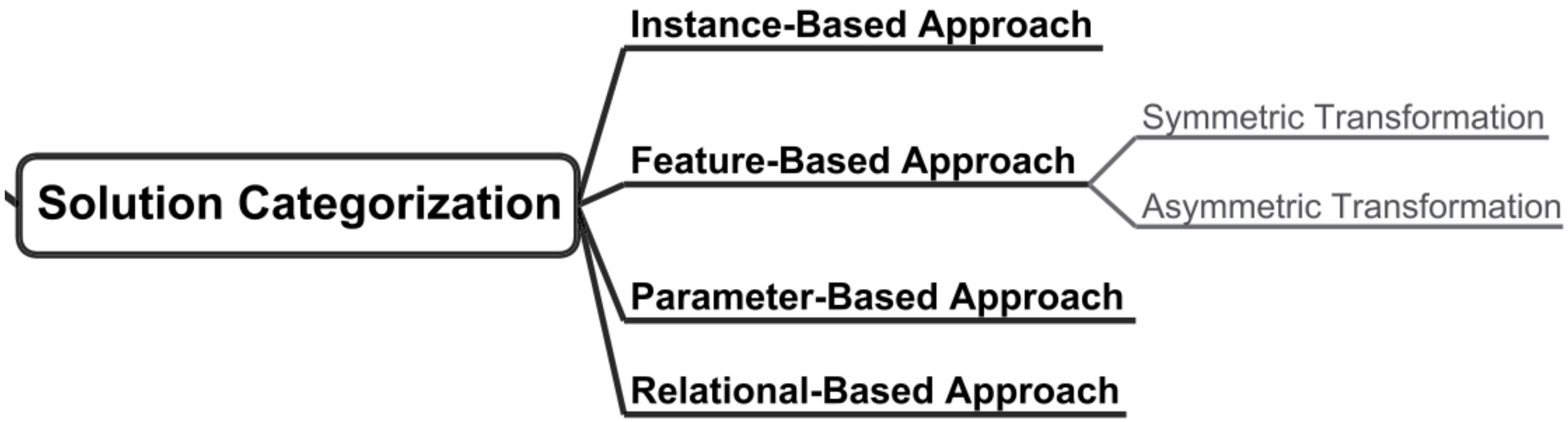
By types of problems/settings



Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Categorization of Transfer Learning

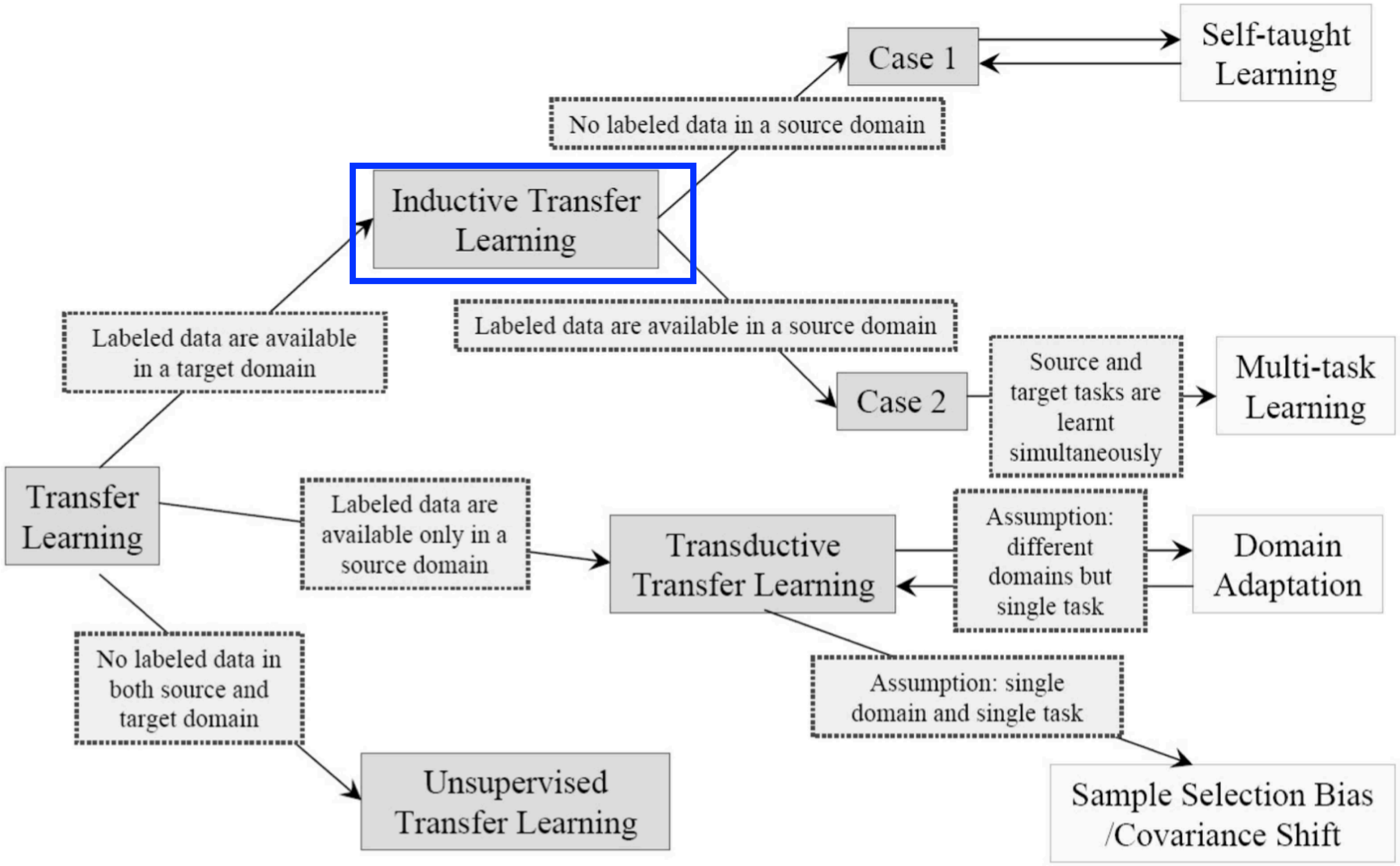
By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Categorization of Transfer Learning

By types of problems/settings



Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Inductive Transfer Learning

When $\mathcal{T}_S \neq \mathcal{T}_T$

- Mostly used when some labeled data are available in a target domain
- Two cases:

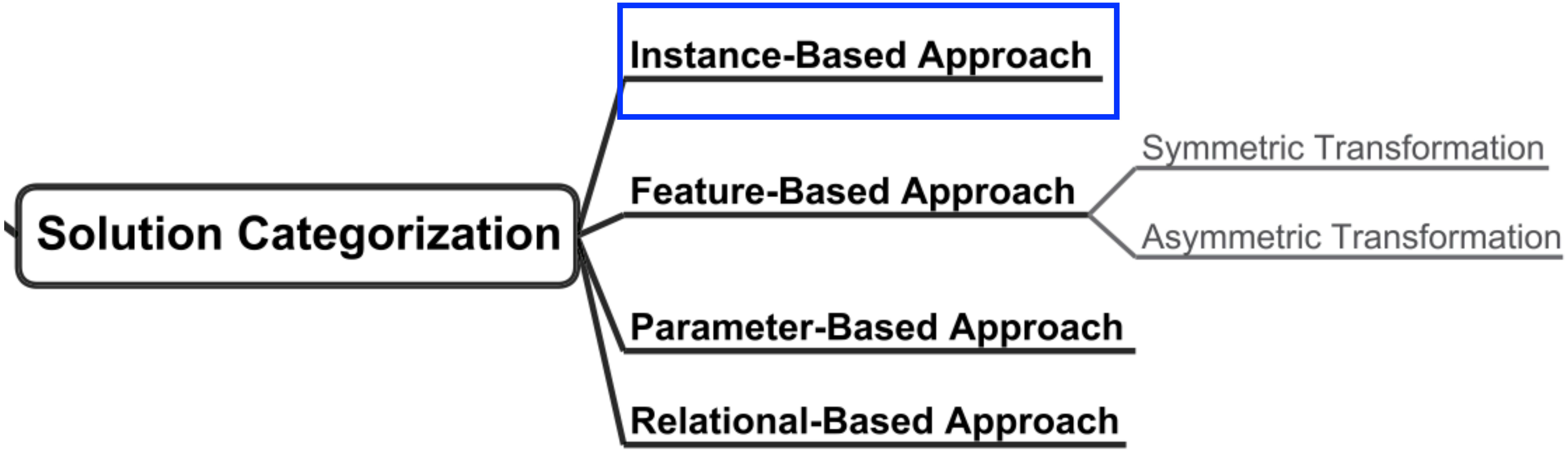
only cover this

1. $\boxed{\text{Labeled data in the source domain are available}}$
2. Only unlabeled data in the source domain are available

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Categorization of Transfer Learning

By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Inductive TL: $\mathcal{T}_S \neq \mathcal{T}_T$; both domains have labeled data

Transferring knowledge of instances X

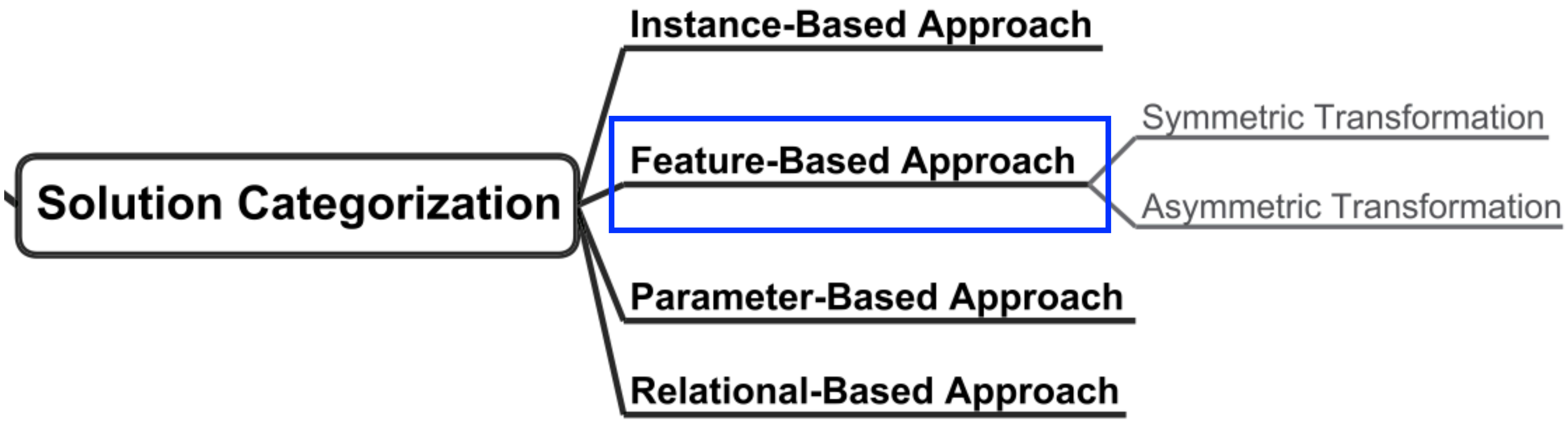
TrAdaBoost

- Combine the labeled source-domain and labeled target-domain instances as a whole training set. Attempt to iteratively re-weight the combined training data to
  - reduce the effect of the "bad" instances
  - encourage the "good" instances to contribute more
- In each iteration, TrAdaBoost trains a weak classifier on the re-weighted data and updates the weights based on the classification error.
- Ensemble the weak classifiers to form a final strong classifier.
- TrAdaBoost extends AdaBoost by using different strategies for updating the weights for source-domain instances and for target-domain instances.

W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for Transfer Learning," Proc. 24th Int'l Conf. Machine Learning, pp. 193-200, June 2007.

Categorization of Transfer Learning

By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Inductive TL: $\mathcal{T}_S \neq \mathcal{T}_T$; both domains have labeled data

Transferring knowledge of features

Supervised feature construction

1. Learn a low-dim feature representation that is shared across tasks
2. Common features learned by solving:

parameters

$$\arg\min_{A,U} \sum_{t \in \{T,S\}} \sum_{i=1}^{n_t} L(y_{t_i}, \langle a_t, U^T x_{t_i} \rangle) + \gamma \|A\|_{2,1}^2$$
$$s.t. \quad U \in \mathbf{O}^d.$$

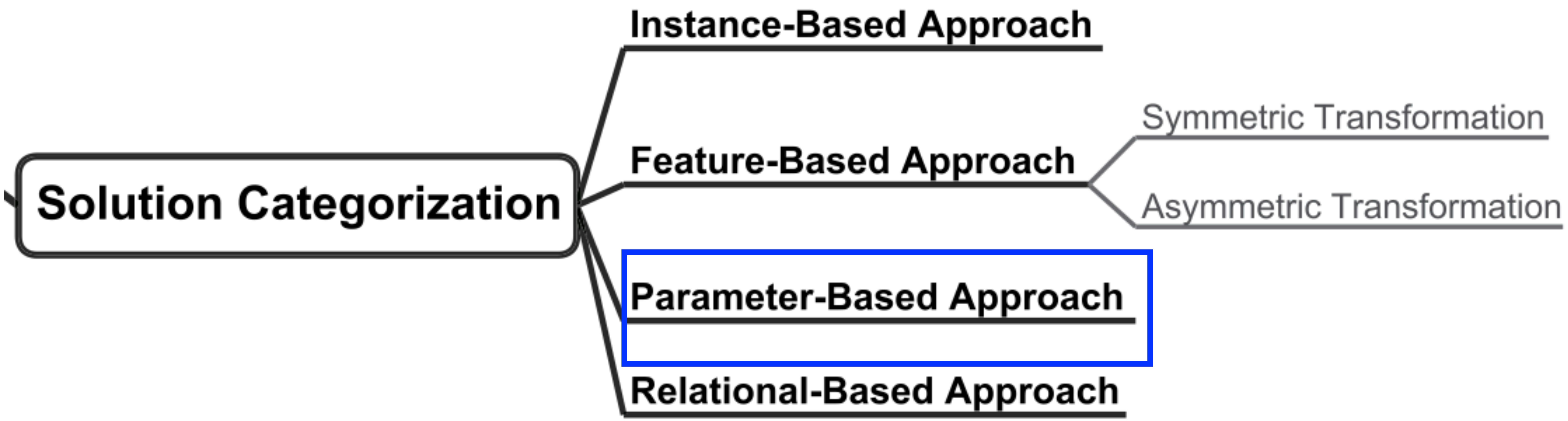
feature map

3. Does not work well for "non-linear" decision/predictive function.

A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-Task Feature Learning," Proc. 19th Ann. Conf. Neural Information Processing Systems, pp. 41-48, Dec. 2007.

Categorization of Transfer Learning

By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Inductive TL: $\mathcal{T}_S \neq \mathcal{T}_T$, $\mathcal{D}_S$ has labeled data

Transferring knowledge of parameters

Assume individual models for related tasks share some parameters or prior distributions of hyper-parameters

1. Gaussian Process (GP): transfer the GP prior

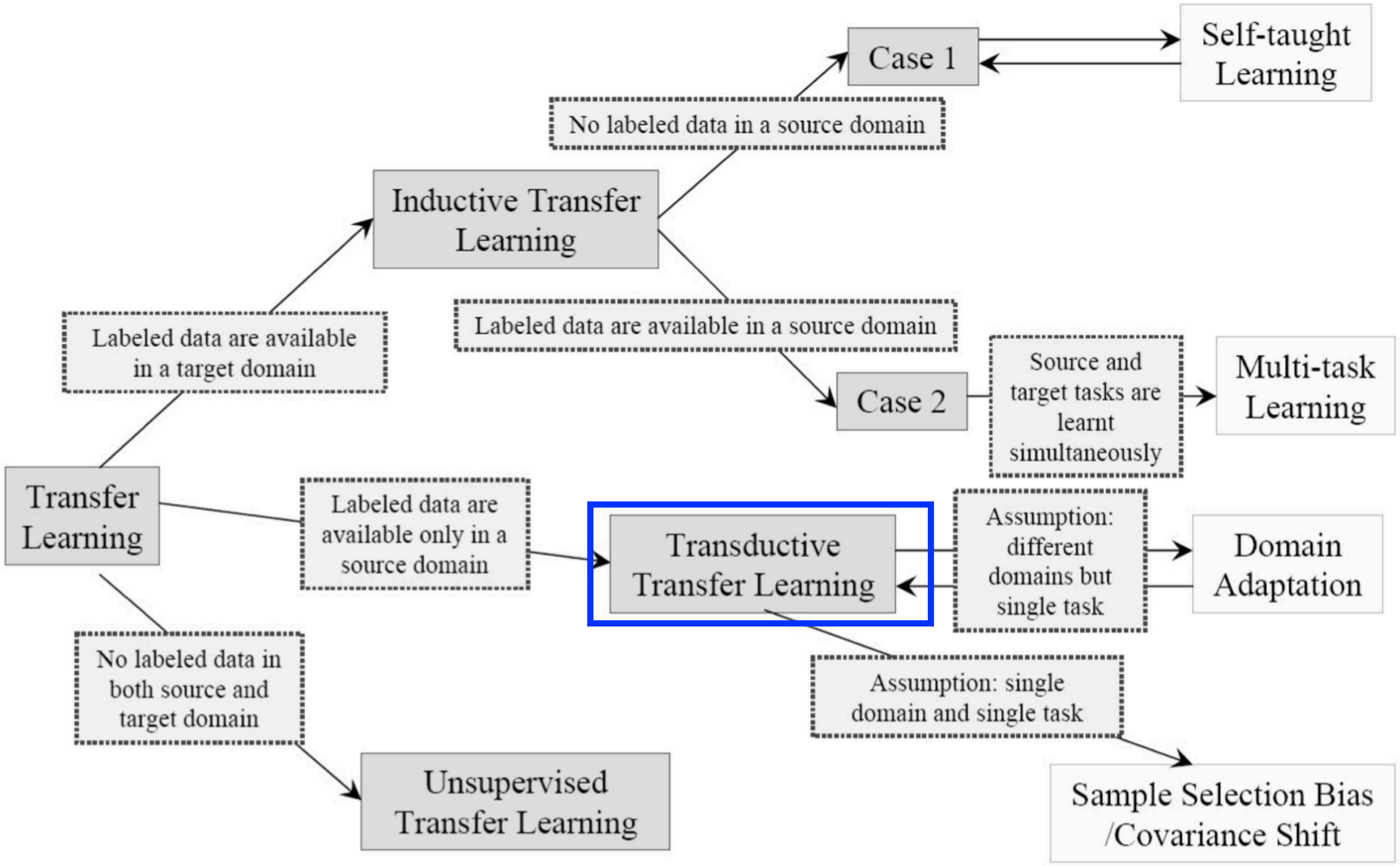
2. Support Vector Machine (SVM): transfer parameters of SVMs

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

(Pause for questions.)

Categorization of Transfer Learning

By types of problems/settings



Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Transductive Transfer Learning

When $\mathscr{D}_S \neq \mathscr{D}_T, \mathscr{T}_S = \mathscr{T}_T$

- Mostly used when labeled data are ONLY available in a source domain.
- Usually require that all/part of the unlabeled data in the target domain are available at training time.
- Famously known as Domain Adaptation (DA).
- Can be further split to:

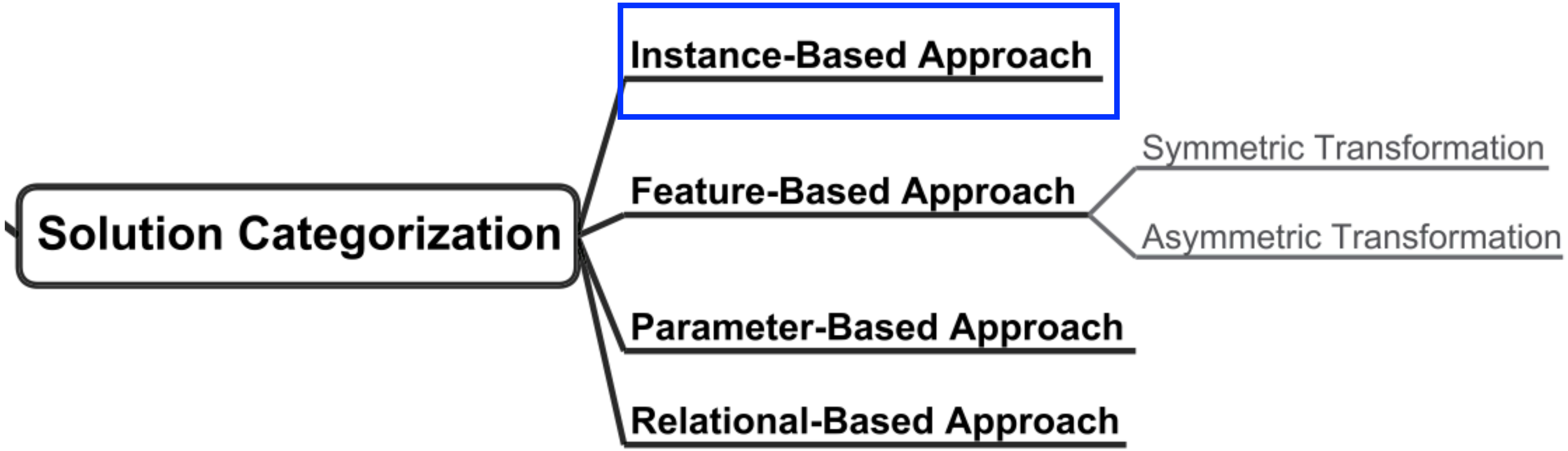
1. $\mathscr{X}_S \neq \mathscr{X}_T$: heterogenous transfer learning
2. $\boxed{\mathscr{X}_S = \mathscr{X}_T, P(X_S) \neq P(X_T)}$

only cover this

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

Categorization of Transfer Learning

By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Instance/Sample-based approaches in DA

Data importance-weighting

- Train a classifier $h(\cdot)$ for the target domain that minimizes the target risk.
- Make use of the source domain information via importance sampling

$$
\begin{aligned}
R_{\mathcal{T}}(h) &= \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y)\; p_{\mathcal{T}}(x, y)\; \mathrm{d}x \\
&= \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y)\; \frac{p_{\mathcal{T}}(x, y)}{p_{\mathcal{S}}(x, y)}\; p_{\mathcal{S}}(x, y)\; \mathrm{d}x\,.
\end{aligned}
$$

- Under covariate shift: $p_S(y \mid x) = p_T(y \mid x)$, the above equals

$$
\sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y) \frac{\cancel{p_{\mathcal{T}}(y \mid x)}\, p_{\mathcal{T}}(x)}{\cancel{p_{\mathcal{S}}(y \mid x)}\, p_{\mathcal{S}}(x)} p_{\mathcal{S}}(x, y) \mathrm{d}x
$$

Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Instance/Sample-based approaches in DA

Data importance-weighting

- Train a classifier $h(\cdot)$ that minimizes $R_T(h)$ under covariate shift:

$$R_T(h) = \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y)\ p_T(x, y)\ \mathrm{d}x = \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y) \frac{\cancel{p_T(y \mid x)}\, p_T(x)}{\cancel{p_S(y \mid x)}\, p_S(x)} p_S(x, y) \mathrm{d}x$$

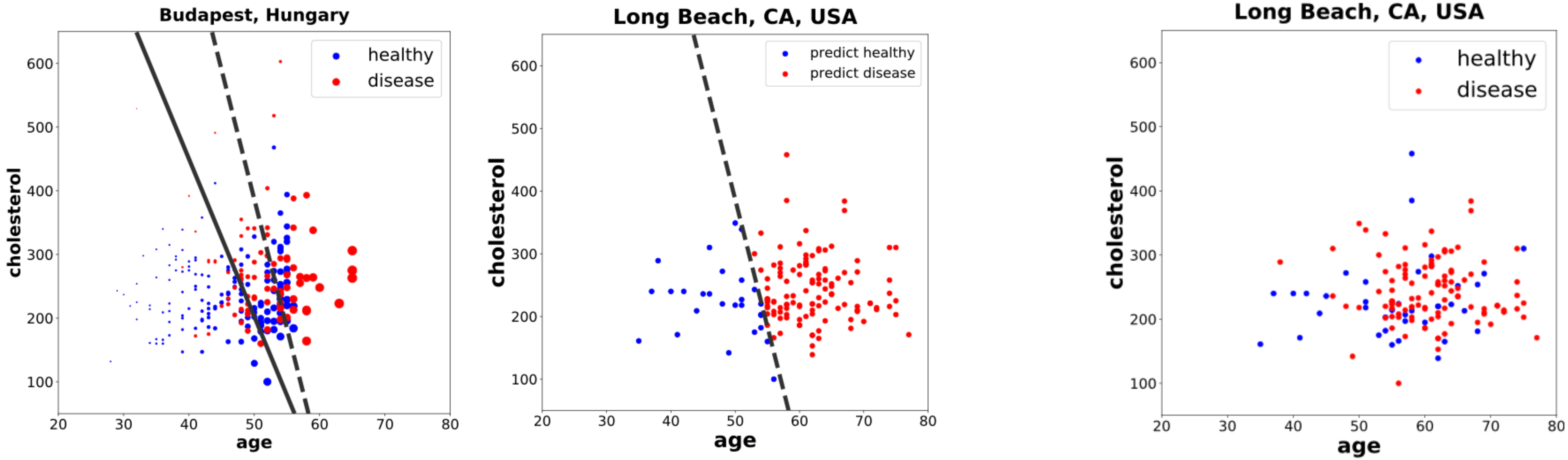
- Now the question becomes: how to well estimate $w(x) := p_T(x)/p_S(x)$?

1. Parametrically: $\hat{w}(x_i) = \dfrac{\mathcal{N}(x_i \mid \hat{\mu}_{\mathcal{T}}, \hat{\Sigma}_{\mathcal{T}})}{\mathcal{N}(x_i \mid \hat{\mu}_{\mathcal{S}}, \hat{\Sigma}_{\mathcal{S}})}$
2. Non-parametrically: $\hat{w}(x_i) = \dfrac{m^{-1} \sum_{j=1}^{m} \kappa_{\sigma_{\mathcal{T}}}(x_i - z_j)}{n^{-1} \sum_{i'=1}^{n} \kappa_{\sigma_{\mathcal{S}}}(x_i - x_{i'})}$
3. Directly estimate $w(x)$ as an independent parameter via optimization using some discrepancy measures:
4. Others: e.g., logistic regression to discriminate between samples from each domain

Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Heart disease diagnosis based on age & cholesterol

Using data importance-weighting



Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Instance/Sample-based approaches in DA

Data importance-weighting

- Directly estimate $w(x) := p_T(x)/p_S(x)$ as an independent parameter via optimization using some discrepancy measures:

  1. Kernel Mean Matching (KMM): matching the means between the source-domain and the target-domain instances in a reproducing kernel Hilbert space (RKHS)

$$\| \mathbb{E}_S[w\phi(x)] - \mathbb{E}_T[\phi(x)] \|_{\mathcal{H}} \approx \frac{1}{n^2}\sum_{i,i'}^{n} w_i\kappa(x_i, x_{i'})w_{i'} - \frac{2}{mn}\sum_{i}^{n}\sum_{j}^{m} w_i\kappa(x_i, z_j)$$

(dropped some "constants")

Minimize w.r.t. w_i s.t. normalization constraints

J. Huang, A. Gretton, K. M. Borgwardt, B. Scho€lkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in Proc. 19th Int. Conf. Neural Inf. Process. Syst., 2007, pp. 601–608.

Instance/Sample-based approaches in DA

Data importance-weighting

- Directly estimate $w(x) := p_T(x)/p_S(x)$ as an independent parameter via optimization using some discrepancy measures:

  2. Kullback-Leibler Importance Estimation Procedure (KLIEP): minimize the KL-divergence between the importance-weighted source distribution $w(x)p_S(x)$ and the true target distribution $p_T(x)$

$$
\begin{aligned}
\mathrm{D_{KL}}[p_T(x), w(x)p_S(x)] &= \int_{\mathcal{X}} p_T(x) \log \frac{p_T(x)}{p_S(x)} \mathrm{d}x - \int_{\mathcal{X}} p_T(x) \log w(x) \mathrm{d}x \\
&\approx -\frac{1}{m} \sum_{j}^{m} \log w(z_j) \, .
\end{aligned}
$$

(dropped some "constants")

Minimize w.r.t. w_j s.t. normalization constraints

M. Sugiyama, S. Nakajima, H. Kashima, P.V. Buenau, and M. Kawanabe, "Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation," Proc. 20th Ann. Conf. Neural Information Processing Systems, Dec. 2008.

Instance/Sample-based approaches in DA

Data importance-weighting

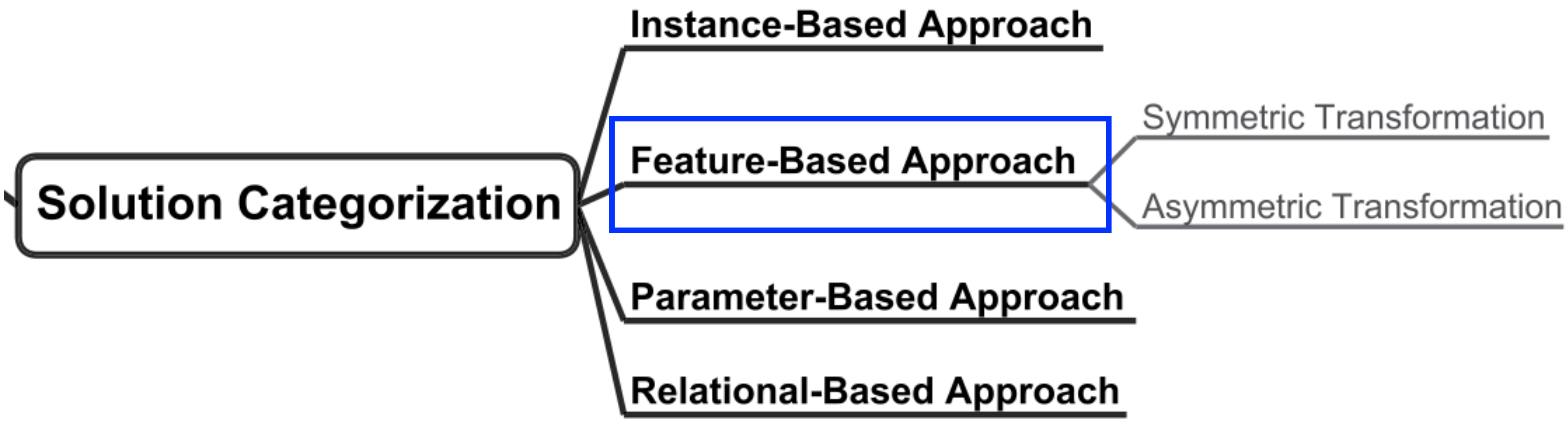
- Directly estimate $w(x) := p_T(x)/p_S(x)$ as an independent parameter via optimization using some discrepancy measures:

    3. L2-norm between the weights and the ratio of data distributions

T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," J. Mach. Learn. Res., vol. 10, pp. 1391–1445, 2009.

Categorization of Transfer Learning

By types of solutions/approaches



Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." *Proceedings of the IEEE* 109.1 (2020): 43-76.

Feature-based approaches in DA

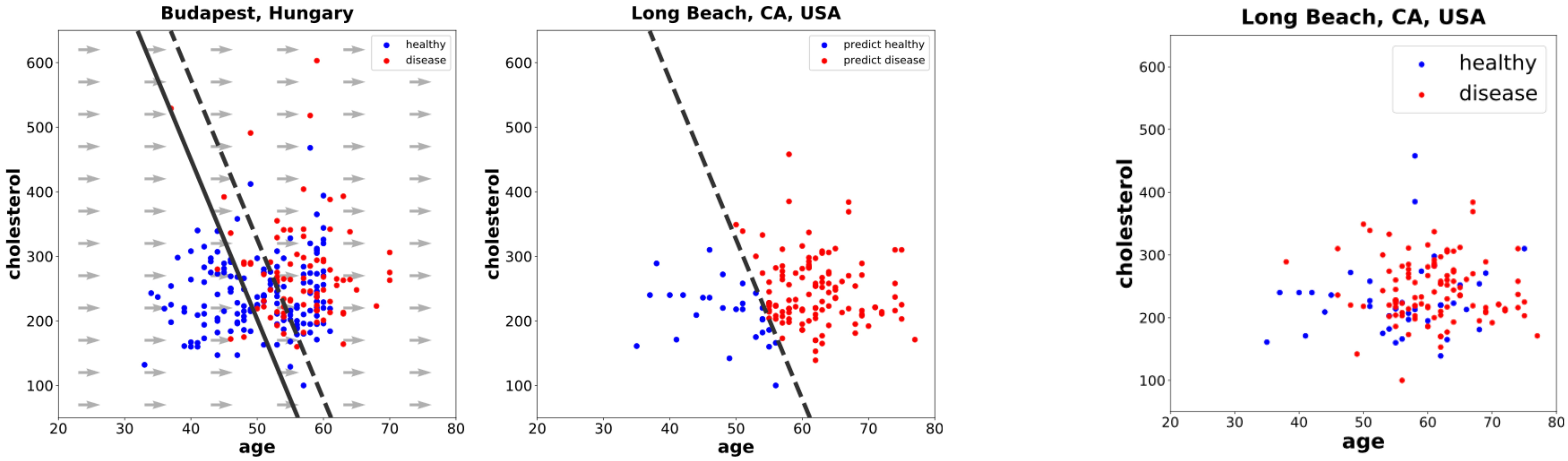
- Asymmetric: learn a transformation that maps source data onto target data.
- Symmetric: find a common latent feature space so as to transform both source & target domain data into new features for knowledge transfer.
- Objectives of constructing feature transformation:
  - Minimize the difference between marginal and conditional distributions
  - Preserve the properties/structures of the data
  - Find the correspondence between features

Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Heart disease diagnosis based on age & cholesterol

Using asymmetric feature-based method: data shifting

- Sometimes, there potentially exists a transformation that maps source data onto target data such that the two domains are brought closer.

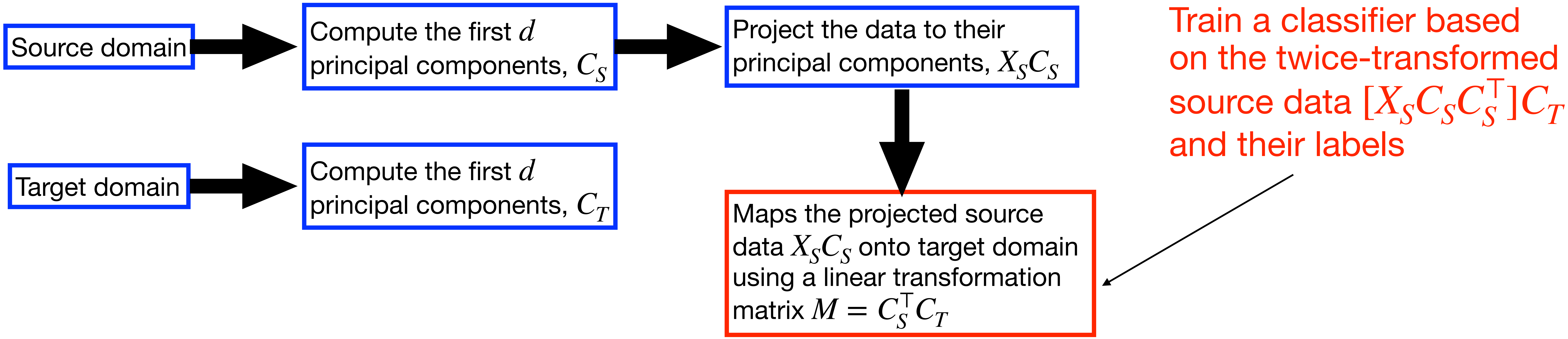


Kouw, Wouter M., and Marco Loog. "A review of domain adaptation without target labels." *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019): 766-785.

Feature-based approaches in DA

Subspace mappings

- Domains could contain domain-specific noise but common subspaces.
- Approach: find these subspaces and map the data onto these subspaces
- Subspace alignment:



- The space can be extended from the linear sense to graph, manifold, etc…

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsuper- vised visual domain adaptation using subspace alignment," in Proc. Int. Conf. Comput. Vis., 2013, pp. 2960–2967.

Feature-based approaches in DA

Optimal Transport

- Find a transportation map $t(\cdot)$ such that $p_T(y \mid t(x)) = p_S(y \mid x)$.
- Train a classifier on the labelled transformed source data.
- Finding such a $t(\cdot)$ among the set of all possible transformations is intractable.
- Instead, find a coupling γ of $p_S(x), p_T(x)$ to minimize the Wasserstein distance

$$\mathrm{D}_{\mathcal{W}}[p_{\mathcal{S}}(x), p_{\mathcal{T}}(x)] = \inf_{\gamma \in \Gamma} \int_{\mathcal{X} \times \mathcal{X}} d(x, z) \mathrm{d}\gamma(x, z)$$

- Sample version of the minimizer γ^* is not hard to find via linear algebra.
- Transform the source sample $\tilde{x}_i = \arg\min_x \sum_j \gamma^*(x_i, z_j) d(x, z_j)$.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

Thank you! :-)