Domain Generalization via Invariant Feature Representation

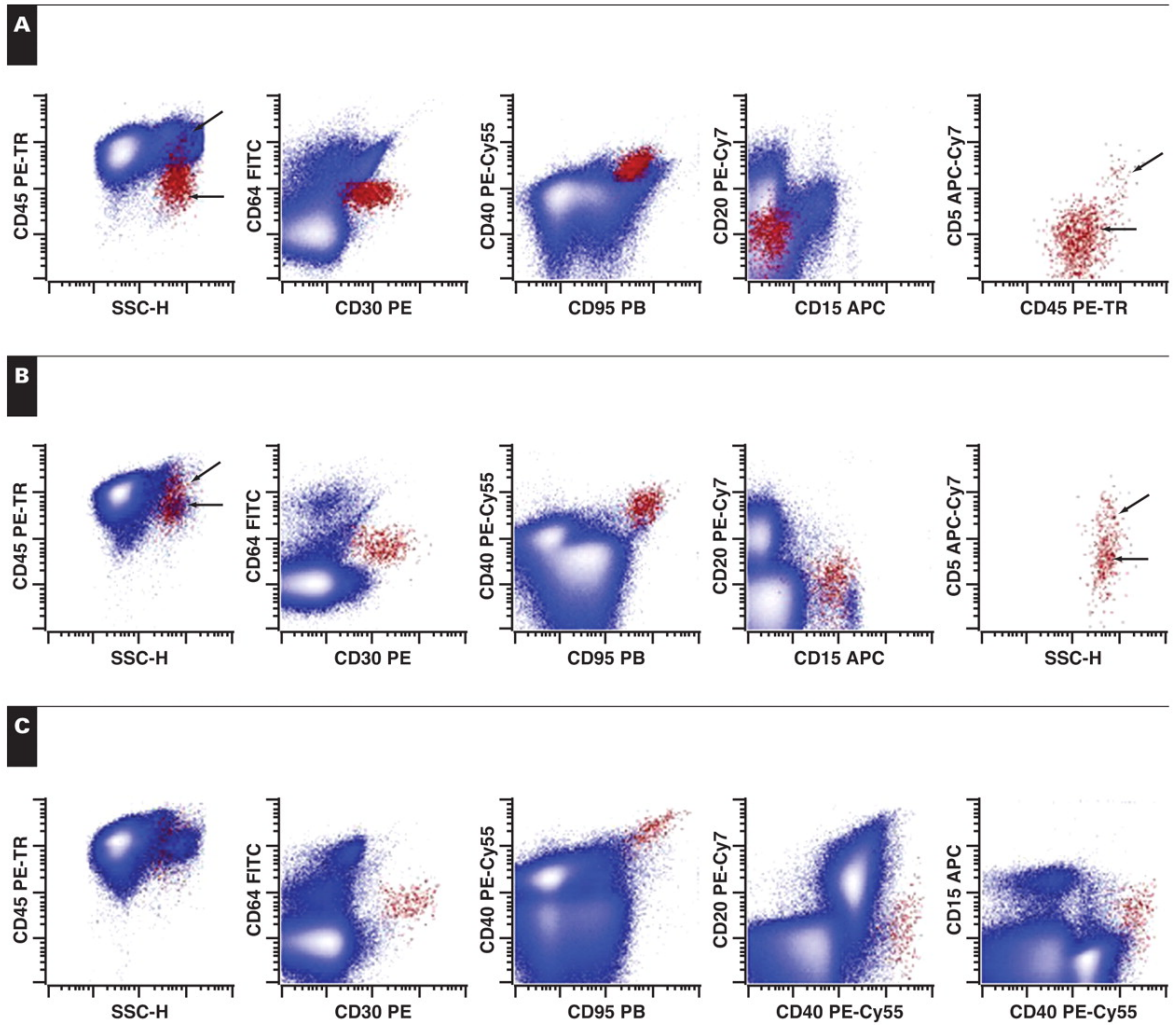
Krikamol Muandet[1], David Balduzzi[2], Bernhard Schölkopf[1]

[1]Empirical Inference Department, MPI for Intelligent Systems
[2]Machine Learning Laboratory, ETH Zurich

June 18, 2013

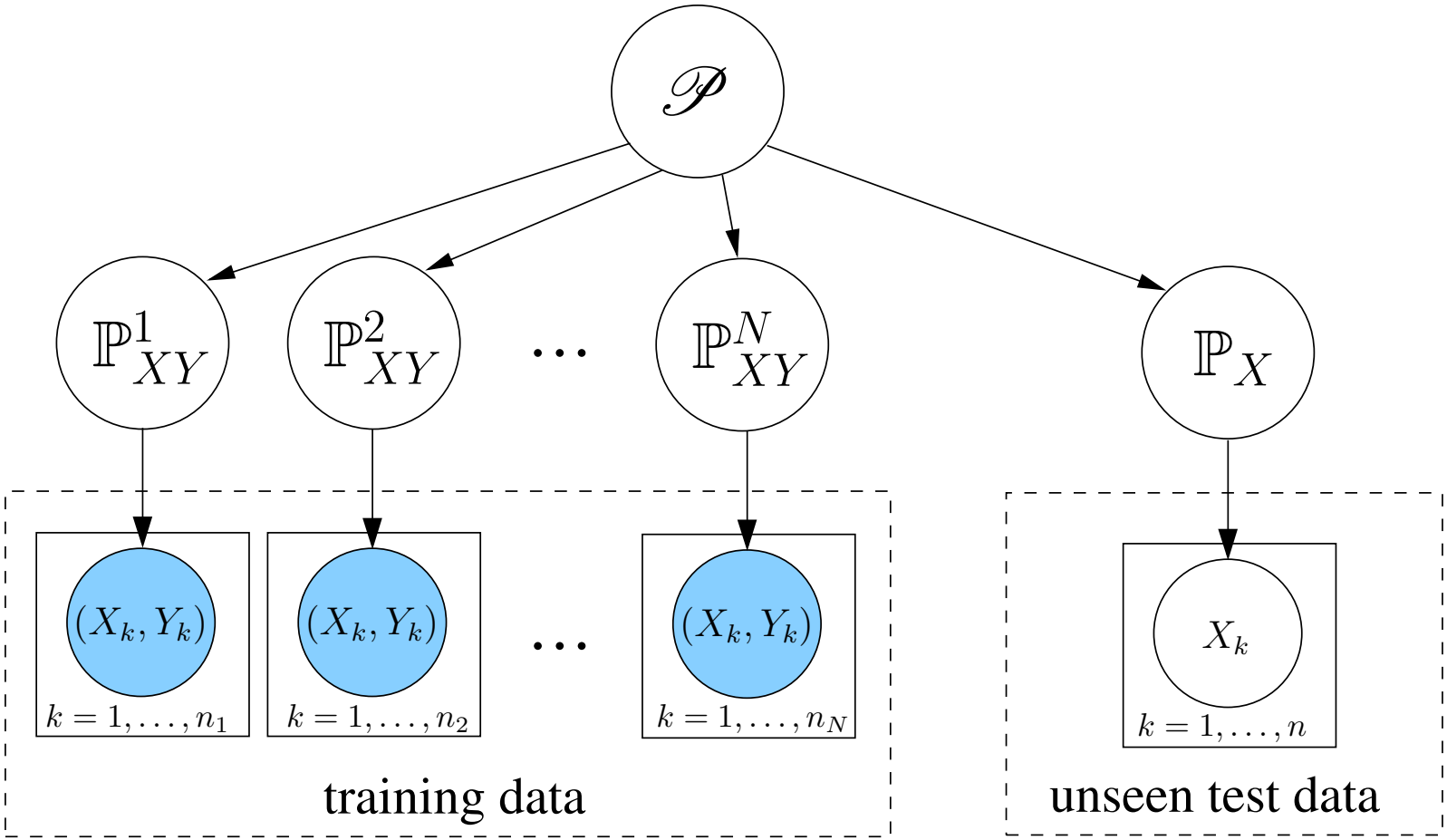
Flow Cytometry

Image courtesy of American Journal of Clinical Pathology.



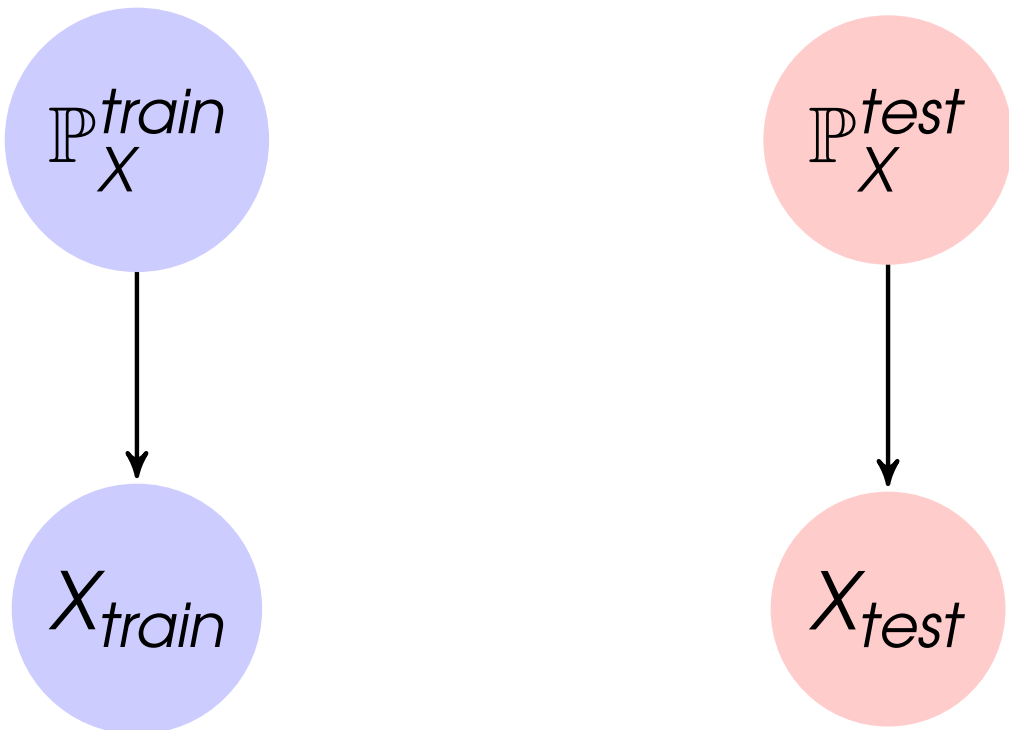
Domains = Patients ($\mathbb{P}_{XY}$), Train Data $\{X_j^{(i)}, Y_j^{(i)}\}_{j=1}^{n_i}$.

Domain Generalization



Related Works

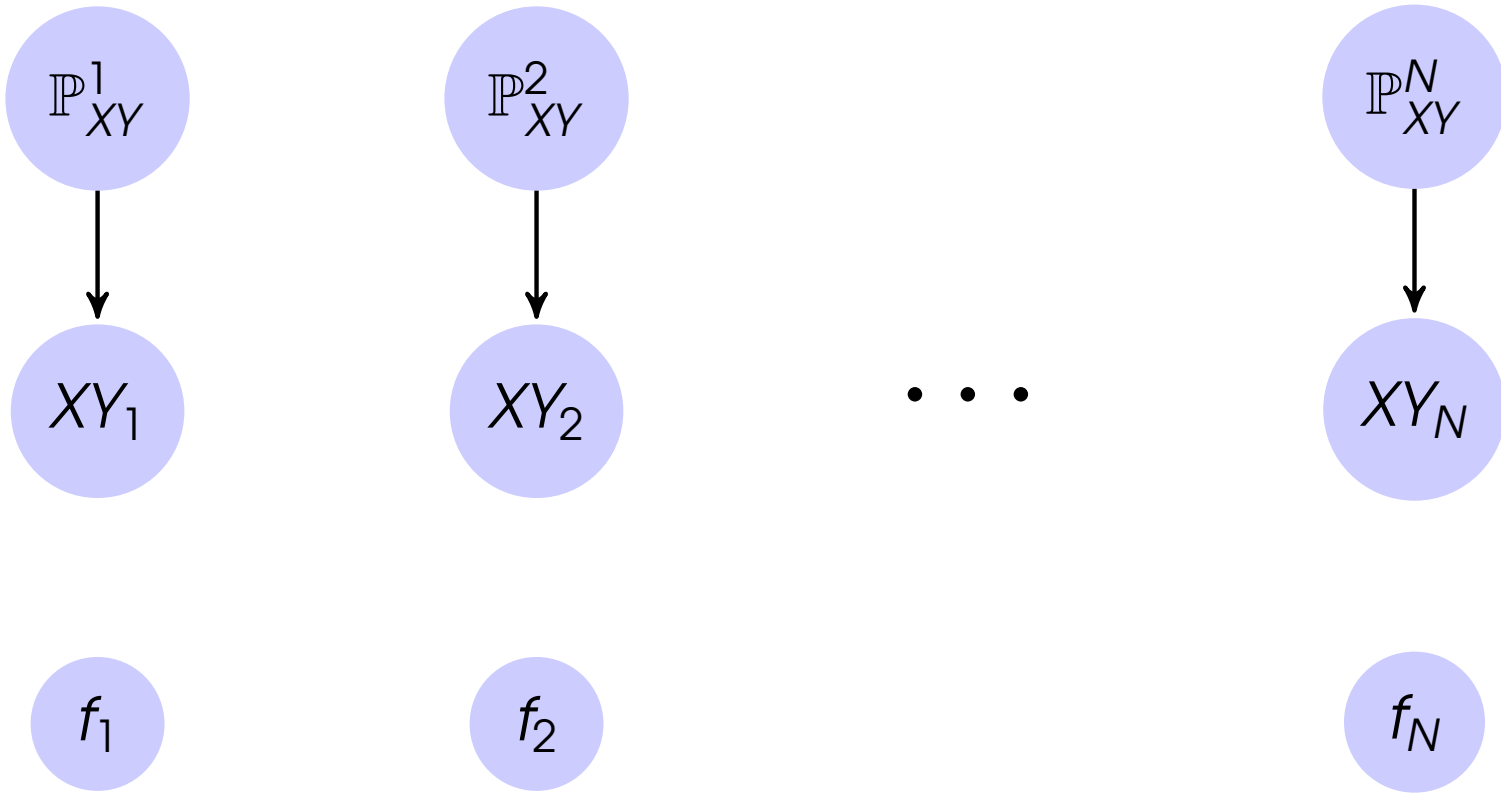
Domain Adaptation (Bickel, Brückner, and Scheffer 2009)



Deal with a mismatch between training and test distributions.

Related Works

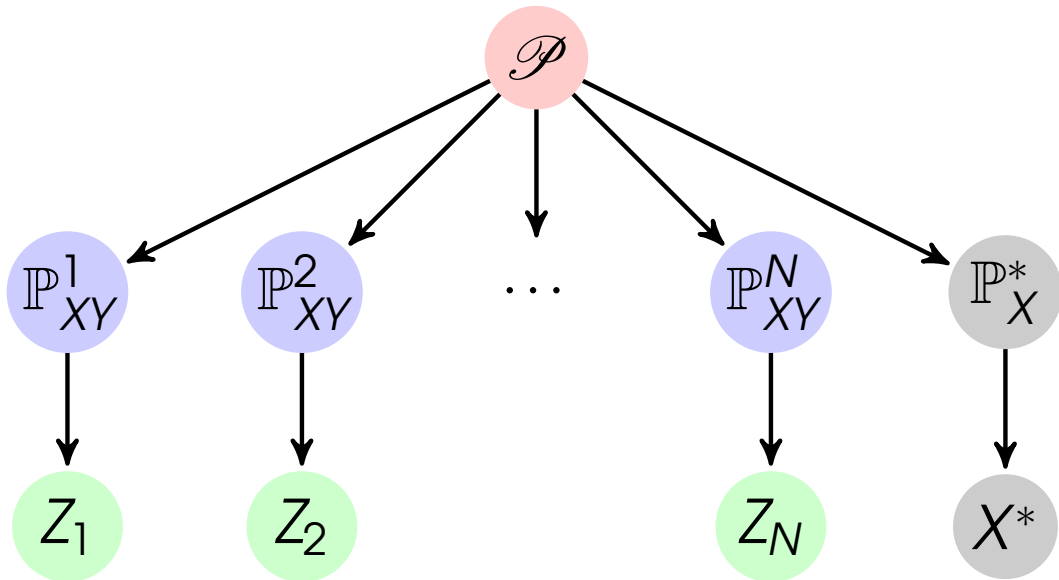
Multitask Learning (Caruana 1997)



Learn multiple tasks simultaneously.

Domain Generalization

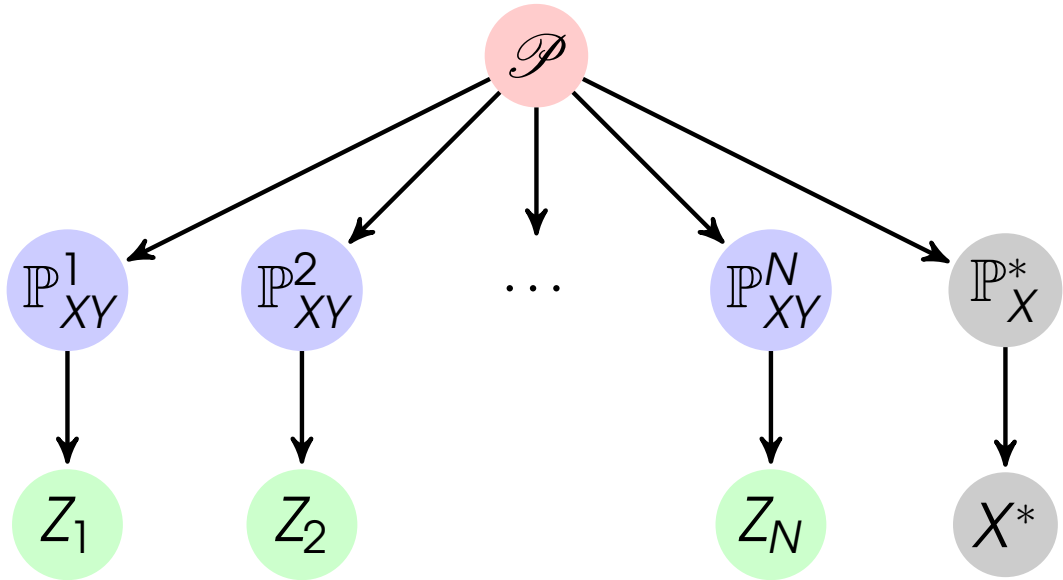
Blanchard, Lee, and Scott 2011



Generalize from multiple source domains to previously unseen domains.

Domain Generalization

Problem Setting



Train: The joint distributions $\mathbb{P}^1_{XY}, \mathbb{P}^2_{XY}, \ldots, \mathbb{P}^N_{XY} \sim \mathscr{P}$.

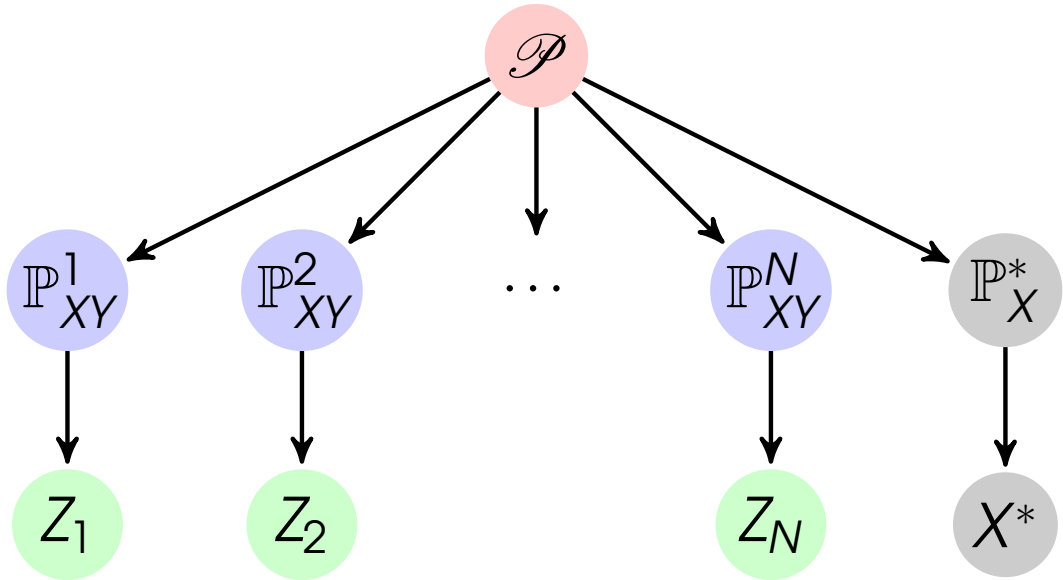
Prediction: An unseen distribution $\mathbb{P}^*_X \sim \mathscr{P}$.

Goal: Learn $f : \mathfrak{P} \times \mathscr{X} \to \mathscr{Y}$.

Assume: $\mathbb{P}^1_{Y|X} \approx \mathbb{P}^2_{Y|X} \approx \cdots \approx \mathbb{P}^N_{Y|X}$.
i.e. functional relationship is stable

Domain Generalization

Problem Setting



Train: The joint distributions $\mathbb{P}^1_{XY}, \mathbb{P}^2_{XY}, \ldots, \mathbb{P}^N_{XY} \sim \mathscr{P}$.

Prediction: An unseen distribution $\mathbb{P}^*_X \sim \mathscr{P}$.

Goal: Learn $f : \mathfrak{P} \times \mathscr{X} \to \mathscr{Y}$.

Assume: $\mathbb{P}^1_{Y|X} \approx \mathbb{P}^2_{Y|X} \approx \cdots \approx \mathbb{P}^N_{Y|X}$.
i.e. functional relationship is stable

Domain Adaptation under Target and Conditional Shift
K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang (ICML2013)

Objective

Find feature representation, $\mathscr{B}(X)$ that is *invariant* across domains.

❶ minimize the distance between empirical distributions $\widehat{\mathbb{P}}^1_X, \widehat{\mathbb{P}}^2_X, \ldots, \widehat{\mathbb{P}}^N_X$ of the transformed samples $\mathscr{B}(X)$.

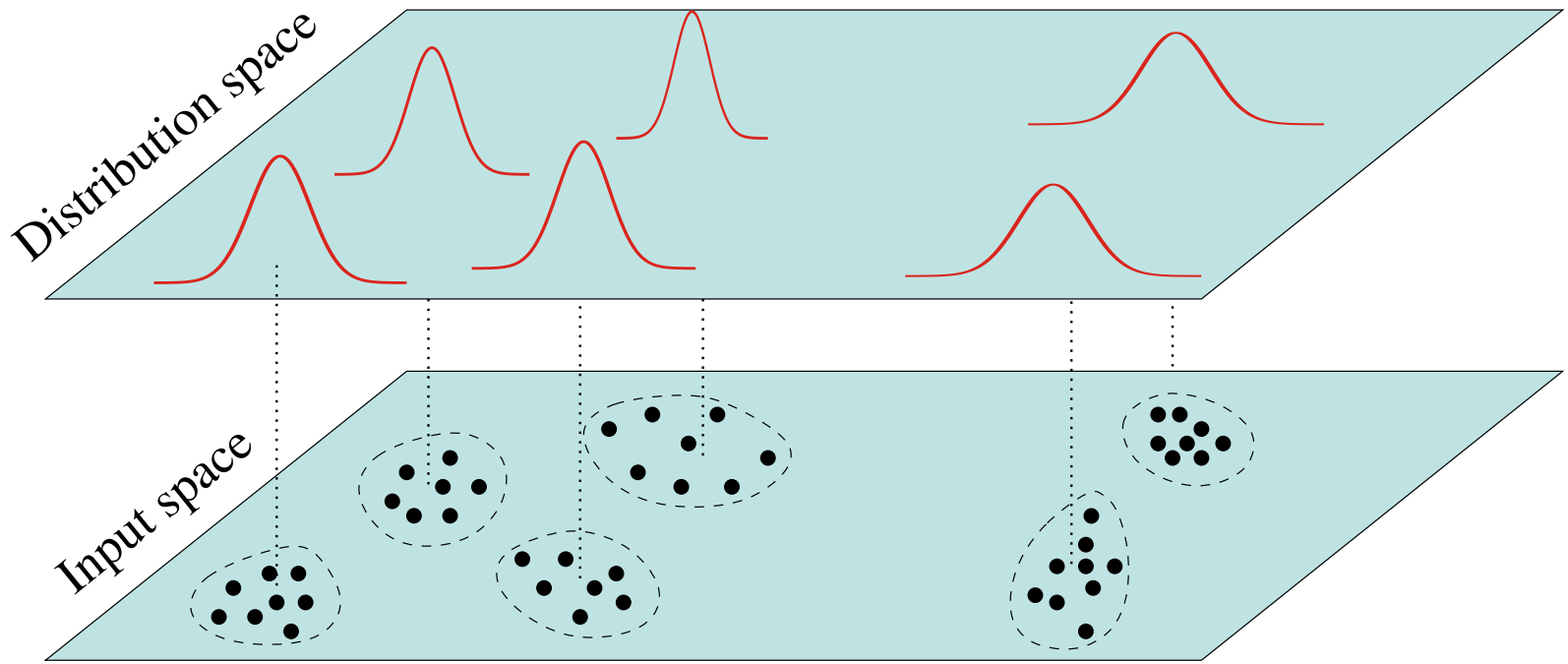
$$\mathbb{P}^1_{Y|X} \cdot \mathbb{P}^1_X \qquad \mathbb{P}^2_{Y|X} \cdot \mathbb{P}^2_X \qquad \cdots \qquad \mathbb{P}^N_{Y|X} \cdot \mathbb{P}^N_X$$

❷ preserve functional relationship between X and Y.

Minimizing Distributional Variance

Hilbert space embedding

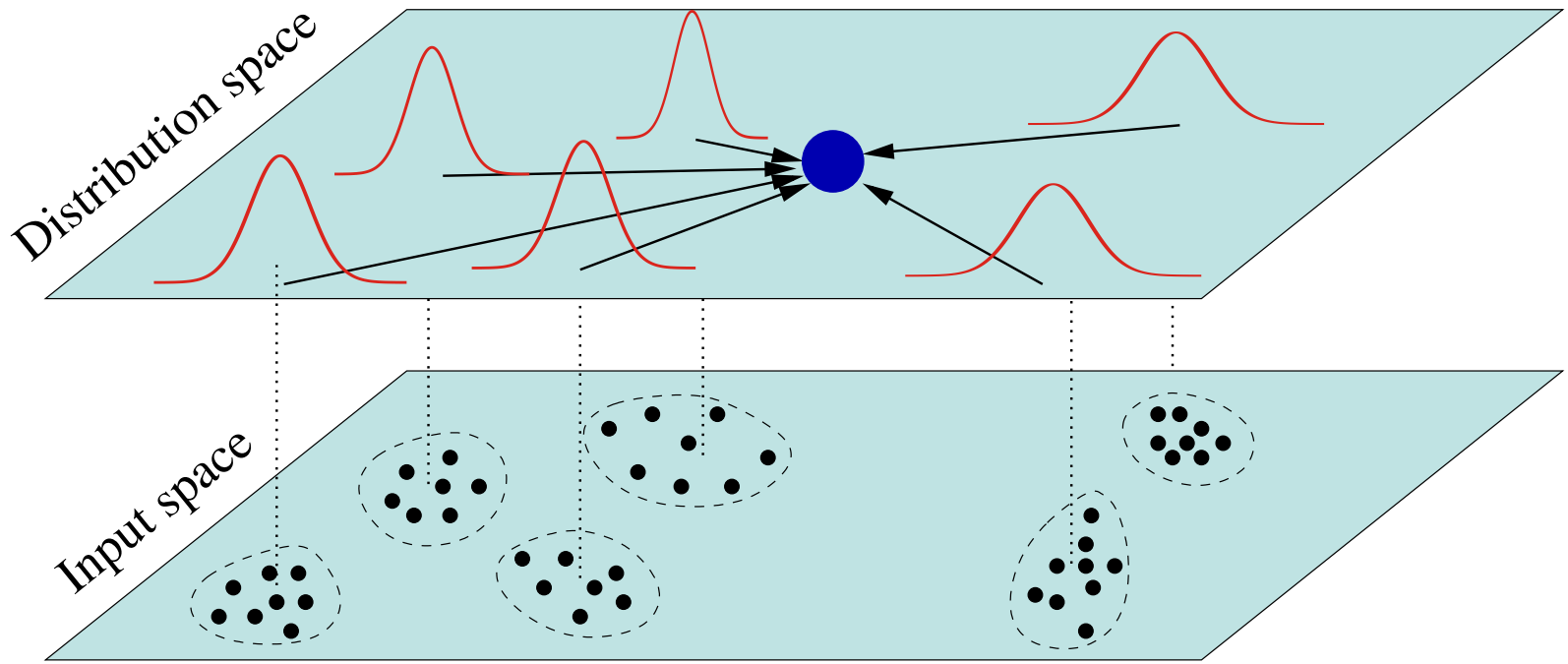
$$\mu : \mathfrak{P}_{\mathscr{X}} \to \mathscr{H}, \quad \mathbb{P} \mapsto \int_{\mathscr{X}} k(x, \cdot)\, \mathrm{d}\mathbb{P}(x) =: \mu_{\mathbb{P}}.$$



Minimizing Distributional Variance

Find transformation $\mathscr{B}$ that minimizes

$$\mathbb{V}_{\mathscr{H}}(\mathscr{P}) = \frac{1}{N}\sum_{i=1}^{N} \|\mu_i\mathscr{B} - \bar{\mu}\mathscr{B}\|^2_{\mathscr{H}}$$



Minimizing Distributional Variance

- Minimizing distributional variance **alone** does not necessarily help with generalization!
  - Setting $\mathscr{B} = \mathbf{0}$ gives zero distributional variance!
- We **also** need to preserve the functional relationship between X and Y encoded in $\mathbb{P}_{Y|X}$.

Preserving Functional Relationship

Central Subspace

The central subspace C is the minimal subspace that captures the functional relationship between X and Y, i.e. $Y \perp\!\!\!\perp X | C^\top X$.

Theorem (Li 1991;Kim and Pavlovic 2011;Muandet 2013)

If B maximizes

$$\mathbf{b}_k^\top \Sigma_{\mathrm{xx}}^{-1} \mathbb{V}(\mathbb{E}[X|Y]) \Sigma_{\mathrm{xx}} \mathbf{b}_k$$

then $Y \perp\!\!\!\perp X | B^\top X$.

Domain-Invariant Component Analysis

preserve functional relationship.

$$\max_{B \in \mathbb{R}^{n \times m}} \frac{f(B)}{g(B)}$$

minimize distributional variance.

Domain-Invariant Component Analysis

preserve functional relationship.

$$\mathbf{b}_k^\top \widehat{\Sigma}_{xx}^{-1} \widehat{\mathbb{V}}(\mathbb{E}[X|Y]) \widehat{\Sigma}_{xx} \mathbf{b}_k$$

$$\max_{B \in \mathbb{R}^{n \times m}} \frac{f(B)}{g(B)}$$

minimize distributional variance.

$$\mathbb{V}_{\mathscr{H}}(\mathscr{P}) = \sum_{i=1}^{N} \|\mu_i \mathscr{B} - \bar{\mu} \mathscr{B}\|_{\mathscr{H}}^2$$

Domain-Invariant Component Analysis

preserve functional relationship.

$\mathbf{b}_k^\top \widehat{\Sigma}_{xx}^{-1} \widehat{\mathbb{V}}(\mathbb{E}[X|Y]) \widehat{\Sigma}_{xx} \mathbf{b}_k$

$$\max_{B \in \mathbb{R}^{n \times m}} \frac{\frac{1}{n}\mathrm{tr}\left(B^\top L(L+n\varepsilon I_n)^{-1} K^2 B\right)}{\mathrm{tr}\left(B^\top KQKB + BKB\right)}$$

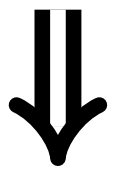
minimize distributional variance.

$\mathbb{V}_{\mathscr{H}}(\mathscr{P}) = \sum_{i=1}^{N} \|\mu_i \mathscr{B} - \bar{\mu}\mathscr{B}\|_{\mathscr{H}}^2$

Domain-Invariant Component Analysis

Maximization Problem

$$\max_{B \in \mathbb{R}^{n \times m}} \frac{\frac{1}{n}\mathrm{tr}\left(B^\top L(L+n\varepsilon I_n)^{-1}K^2 B\right)}{\mathrm{tr}\left(B^\top KQKB + BKB\right)}$$



Generalized Eigenvalue Problem

$$\frac{1}{n}L(L+n\varepsilon I)^{-1}K^2 B = (KQK + K + \lambda I)B\Gamma$$

Learning guarantee

Theorem

Under reasonable assumptions, it holds with probability at least $1-\delta$ *that,*

$$\mathbb{E}[\mathit{error}] \leq c_1 \mathbb{V}_{\mathscr{H}}(\mathscr{P} \cdot \mathscr{B}) + L(n, N).$$

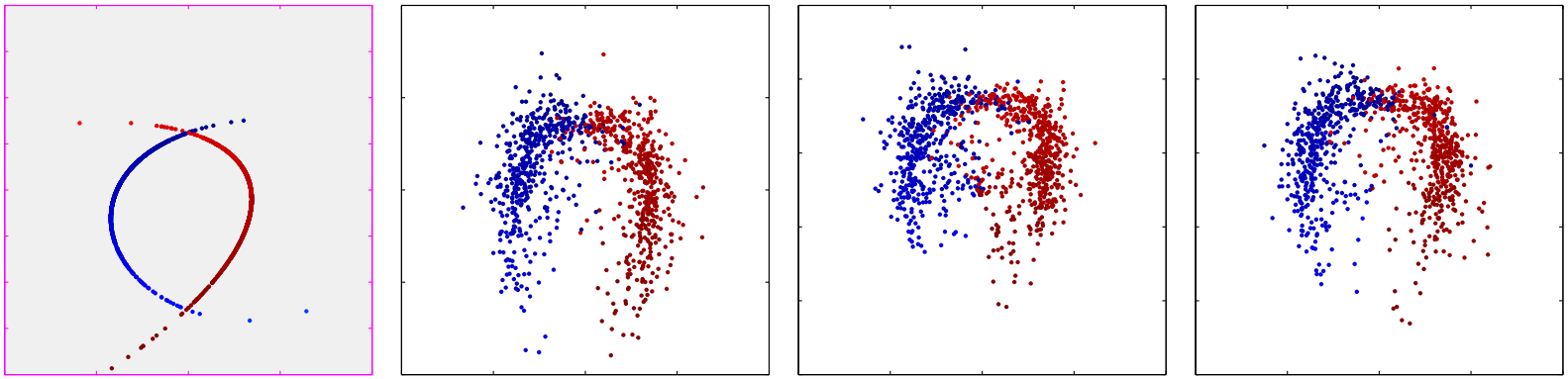
- Bound depends on the distributional variance.
- $L(n, N) \to 0$ as samples n and domains N go to infinity.

Experimental Results

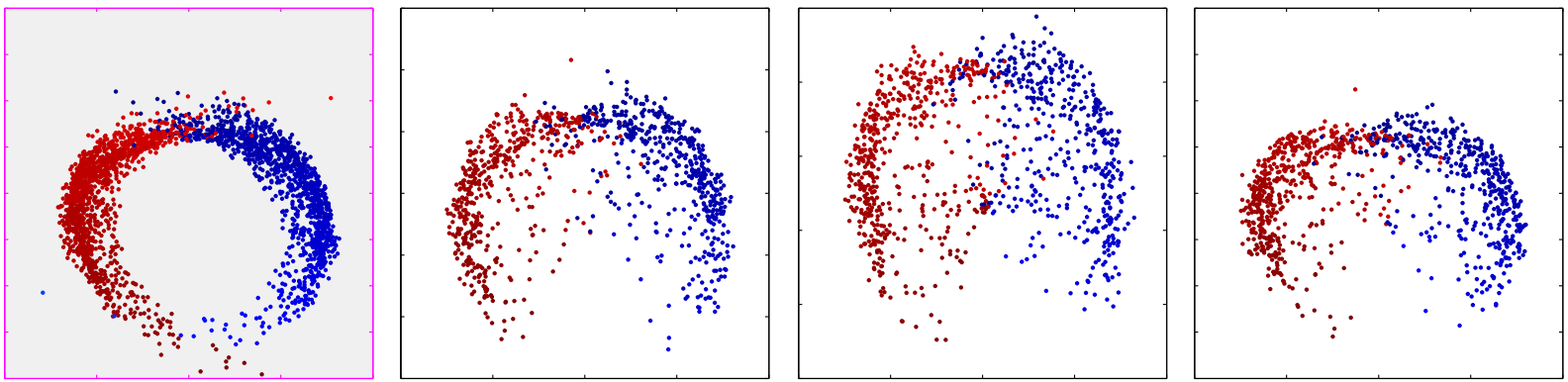
Synthetic Data

- Generate 10 collections of $n_i \sim \texttt{Poisson}(200)$ data points.
- For each collection, $x \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$ where $\Sigma_i \sim \mathcal{W}(0.2 \times I_5, 10)$.
- The output value is $y = \text{sign}(b_1^\top x + \varepsilon_1) \cdot \log(|b_2^\top x + c + \varepsilon_2|)$, where $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, 1)$.

Experimental Results: synthetic data



COIR



DICA

Experimental Results

Real-world Data

- Flow cytometry dataset (classification).
- Parkinson's telemonitoring dataset (regression).

Learning algorithms

- Pooling SVM : pool data from all domains and apply standard SVM.
- Distributional SVM : apply the kernel

$$\kappa((\mathbb{P}^i, x_k^i), (\mathbb{P}^j, x_l^j)) = K(\mathbb{P}^i, \mathbb{P}^j) \cdot k(x_i^k, x_l^j)$$

(Blanchard, Lee, and Scott 2011).

Experimental Results: Flow cytometry

Methods	**Pooling SVM**	**Distributional SVM**
Input	92.03±8.21	93.19±7.20
KPCA	91.99±9.02	93.11±6.83
COIR	92.40±8.63	92.92±8.20
UDICA	92.51±5.09	92.74±5.01
DICA	**92.72±6.41**	**94.80±3.81**

Similar results for Parkinson's telemonitoring dataset.

Conclusion

Domain-Invariance Component Analysis (DICA) finds an **invariant representation** that

- minimizes "differences" between domains
- while preserving discriminative information.

To learn more, please come to our poster!

Thank you!