Domain Generalization

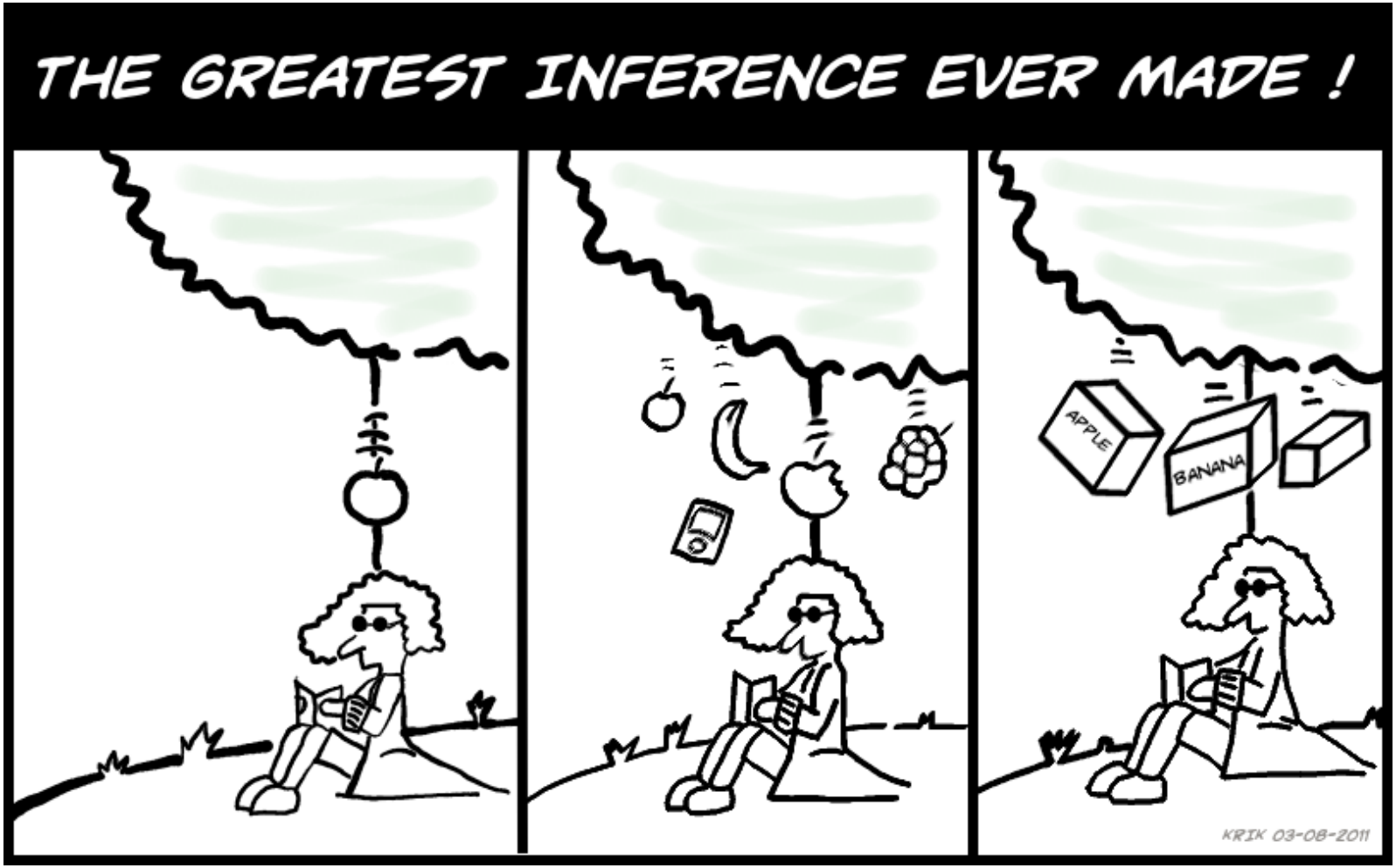
Krikamol Muandet

Max Planck Institute for Intelligent Systems

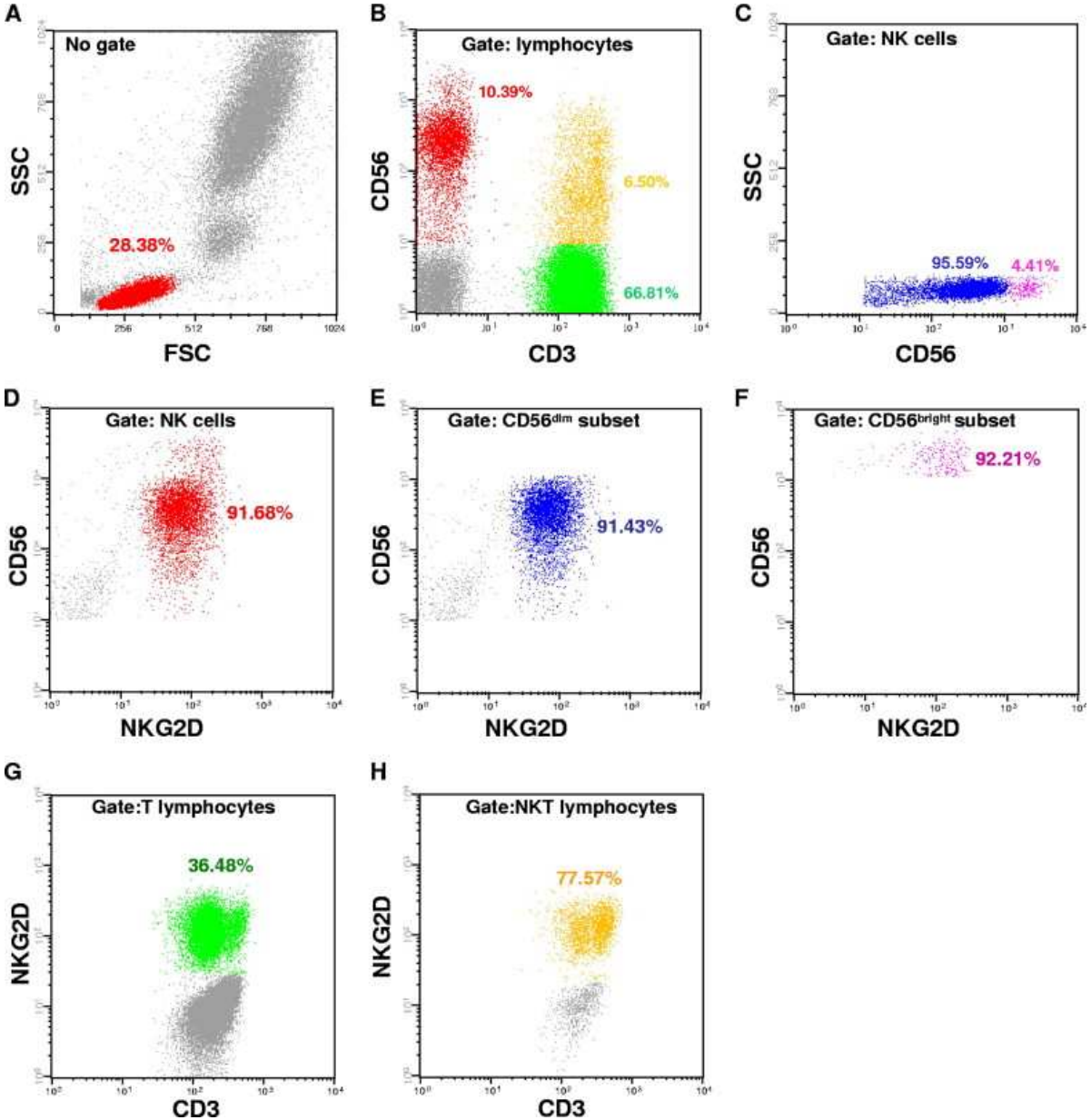
Tübingen, Germany

March 1, 2012

Motivation


THE GREATEST INFERENCE EVER MADE !
APPLE
BANANA
KRIK 03-08-2011

Gating of Flow Cytometry Data



Formal Setup

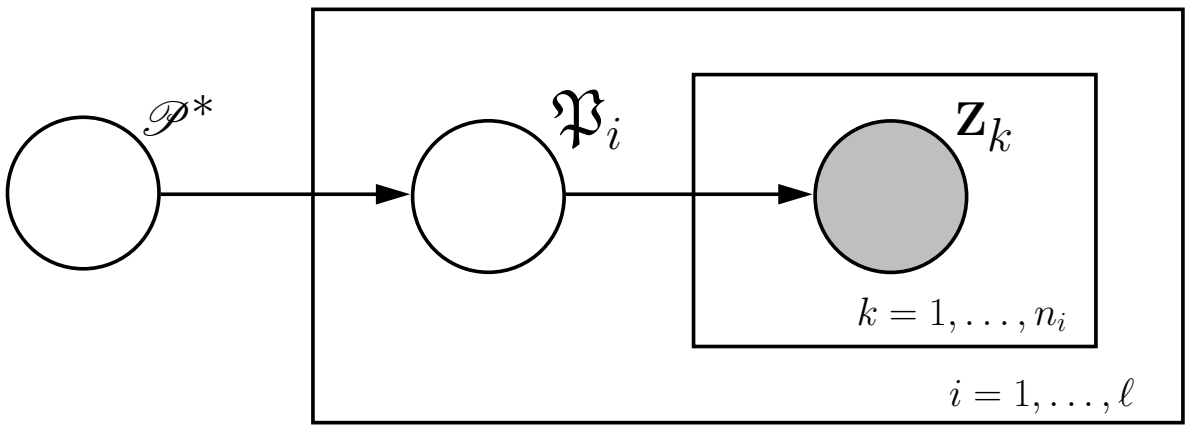
Domain Generalization

Given sample sets $\mathcal{S} = \{S_1, S_2, \ldots, S_\ell\}$. Each S_i consists of n_i i.i.d. realizations drawn according to $\mathbb{P}_{XY}^{(i)}$. Learn a function $f : \mathscr{P} \times \mathcal{X} \to \mathcal{Y}$ that generalizes well to an unseen sample set $S^T = \{x_1^T, x_2^T, \ldots, x_m^T\}$ drawn according to an unseen marginal distribution $\mathbb{P}^T$.

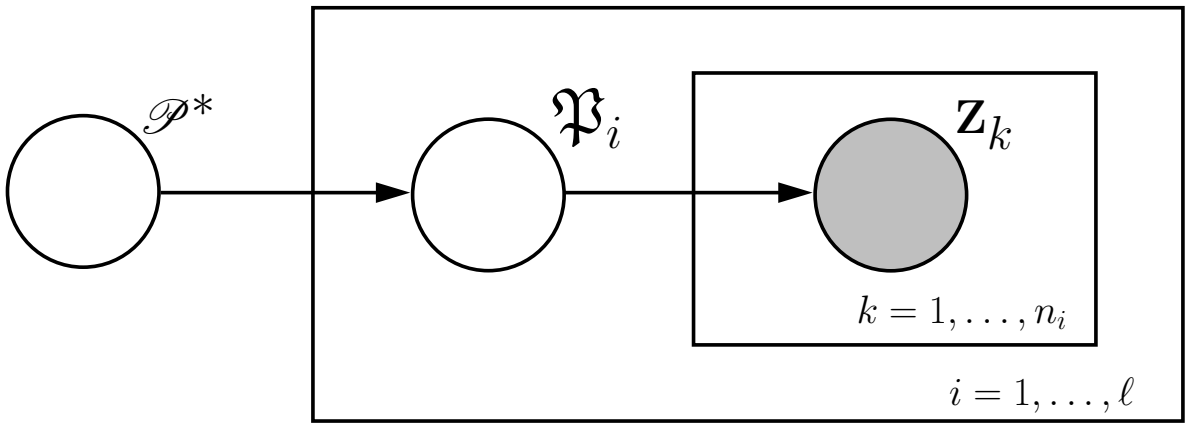
Formal Setup

Domain Generalization

Given sample sets $\mathcal{S} = \{S_1, S_2, \ldots, S_\ell\}$. Each S_i consists of n_i i.i.d. realizations drawn according to $\mathbb{P}_{XY}^{(i)}$. Learn a function $f : \mathscr{P} \times \mathcal{X} \to \mathcal{Y}$ that generalizes well to an unseen sample set $S^T = \{x_1^T, x_2^T, \ldots, x_m^T\}$ drawn according to an unseen marginal distribution $\mathbb{P}^T$.



Formal Setup



	Supervised Data		Unsupervised Data	
Supervised Domain	$\mathfrak{P} = (\mathbb{P}_{XY}, W)$	$\mathbf{Z} = (X, Y)$	$\mathfrak{P} = (\mathbb{P}_X, Z)$	$\mathbf{Z} = X$
Unsupervised Domain	$\mathfrak{P} = \mathbb{P}_{XY}$	$\mathbf{Z} = (X, Y)$	$\mathfrak{P} = \mathbb{P}_X$	$\mathbf{Z} = X$

Domain Generalization and Related Frameworks

Framework	Mismatch	Multiple Sources	Target Domain
Standard Setup	✗	✗	✗
Transfer Learning	✓	✗	✓
Multi-task Learning	✓	✓	✗
Domain Adaptation	✓	✓	✓
Domain Generalization	✓	✓	✗

Distributive Variance

Based on the notion of positive definite kernel between distributions, the variance between distributions can be evaluated as

$$\mathbb{V}_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_\ell) = \mathrm{tr}(KL)$$

where

$$K = \begin{pmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,\ell} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,\ell} \\ \vdots & \vdots & \ddots & \vdots \\ K_{\ell,1} & K_{\ell,2} & \cdots & K_{\ell,\ell} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

$$L_{kl} = \begin{cases} \dfrac{1}{n_i^2} - \dfrac{1}{\ell \cdot n_i^2} & \text{if } x_k, x_l \in S_i \\ -\dfrac{1}{\ell \cdot n_i \cdot n_j} & \text{if } x_k \in S_i \text{ and } x_l \in S_j, i \neq j \end{cases}$$

Stationary Component Analysis (SCA)

Unsupervised SCA

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \operatorname{tr}(V^{\mathsf{T}} KLKV) + \lambda \operatorname{tr}(V^{\mathsf{T}} V) \\ \text{subject to} \quad & V^{\mathsf{T}} KHKV = I \end{aligned} \tag{1}$$

Then, the projection V are the m leading eigenvectors of $(KLK + \lambda I)^{-1} KHK$

Stationary Component Analysis (SCA)

Unsupervised SCA

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \operatorname{tr}(V^{\mathsf{T}} KLKV) + \lambda \operatorname{tr}(V^{\mathsf{T}} V) \\ \text{subject to} \quad & V^{\mathsf{T}} KHKV = I \end{aligned} \tag{1}$$

Then, the projection V are the m leading eigenvectors of $(KLK + \lambda I)^{-1} KHK$

Supervised SCA

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \operatorname{tr}(V^{\mathsf{T}} K_X L K_X V) + \lambda \operatorname{tr}(V^{\mathsf{T}} V) \\ \text{subject to} \quad & \frac{1}{n} V^{\mathsf{T}} K_X K_Y (K_Y + n\varepsilon I_n)^{-1} K_X V = I_m \end{aligned} \tag{2}$$

Thus, the solution V of are the m leading eigenvectors of $(1/n)(K_X L K_X + \lambda I)^{-1} K_X K_Y (K_Y + n\varepsilon I_n)^{-1} K_X$.

Multiple Kernel Feature Learning

We consider learning a feature representation and the functional relationship among several sources simultaneously. That is, the goal is to learn the function

$$f(\mathbb{P}, x) = \mathbf{w}^{\mathsf{T}} \phi(\mathbb{P}, x) + b = \sum_{i=1}^{\ell} \sum_{k=1}^{n_i} \alpha_i \mathcal{K}((\mathbb{P}_i, x_k^{(i)}), (\mathbb{P}, x)) \tag{3}$$

We consider the following optimization problem:

$$\arg\min_{\mathcal{K}, f} \mathcal{L}(\mathcal{K}, f, \mathcal{P}) + \lambda \Omega(\mathbb{V}(\mathcal{P})) \tag{4}$$

where Ω is any monotonic increasing function and $\lambda > 0$ is the regularization parameter.

Multiple Kernel Feature Learning

In this work, we consider the hinge loss with the kernel function

$$\mathcal{K}((\mathbb{P}_i, x_k^{(i)}), (\mathbb{P}_j, x_l^{(j)})) = K(\mathbb{P}_i, \mathbb{P}_j) \cdot k(x_k^{(i)}, x_l^{(j)}) \quad (5)$$

$$= \sum_{m=1}^{M} d_m K(\mathbb{P}_i, \mathbb{P}_j) \cdot k_m(x_k, x_l) \quad (6)$$

where $d_m \geq 0$ for all m. By employing the p-norm regularization on $\mathbf{d}$, the primal can therefore be formulated as

$$\underset{\mathbf{w},b,\boldsymbol{\xi}\geq 0,\mathbf{d}\geq 0}{\text{minimize}} \quad \frac{1}{2}\sum_m \mathbf{w}_m^{\mathsf{T}}\mathbf{w}_m + C\sum_i\sum_k \xi_k^{(i)} + \theta\left(\sum_m d_m p_m\right)^2 + \frac{\lambda}{2}\left(\sum_m d_m^p\right)^{\frac{2}{p}} \quad (7)$$

$$\text{subject to} \quad y_k^{(i)}\left(\sum_m \sqrt{d_m}\mathbf{w}^{\mathsf{T}}\phi_m(\mathbb{P}_i, x_k^{(i)}) + b\right) \geq 1 - \xi_k^{(i)} \quad (8)$$

Questions?