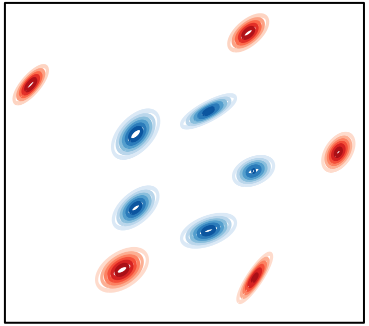
Towards Explaining Distribution Shifts

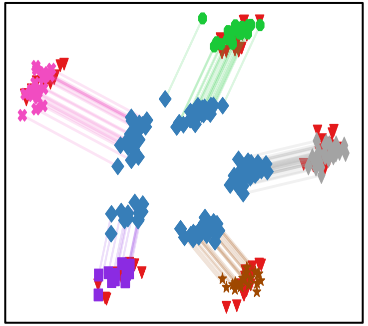
Sean Kulinski

David Inouye

Oracle Shift from P_{src} to P_{tgt}



Proposed Distribution Shift Explanation



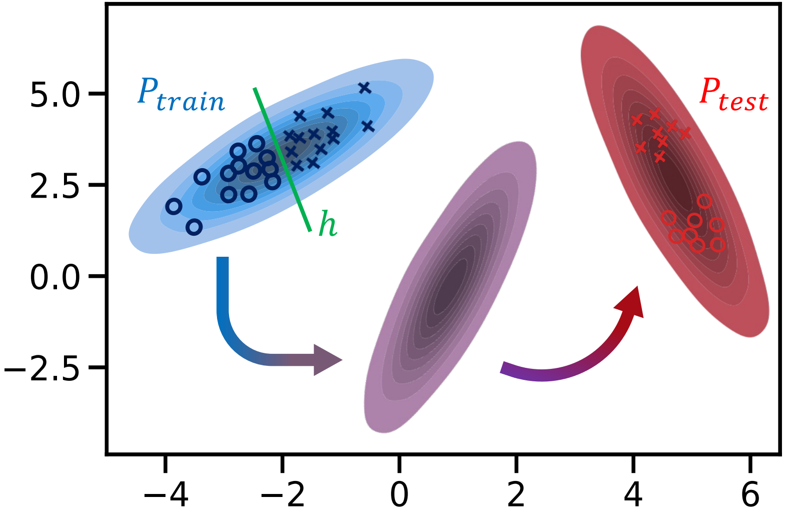
40 Years ICML International Conference On Machine Learning

A distribution shift is when a data distribution changes from what is expected

- In machine learning, a distribution shift is when a **testing distribution** no longer matches the **training distribution**

$$P_{test}(x) \neq P_{train}(x)$$

- Under distribution shift, the patterns learned by **a model** might not be present in P_{test}

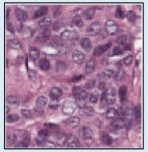
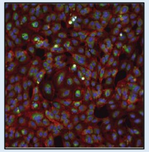
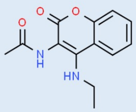


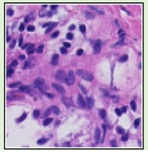
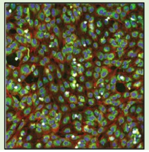
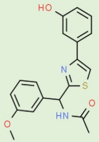
Distribution shifts are ubiquitous

- Any changes in a current data generating environment can cause shifts

- Applying a model to a new domain is almost always a shift

Dataset	iWildCam	Camelyon17	RxRx1	FMoW	PovertyMap	GlobalWheat	OGB-MolPCBA	CivilComments	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	satellite image	satellite image	wheat image	molecular graph	online comment	product review	code
Prediction (y)	animal species	tumor	perturbed gene	land use	asset wealth	wheat head bbox	bioassays	toxicity	sentiment	autocomplete
Domain (d)	camera	hospital	batch	time, region	country, ru/ur	location, time	scaffold	demographic	user	git repo

Source example								What do Black and LGBT people have to do with bicycle licensing?	Overall a solid package that has a good quality of construction for the price.	`import numpy as np`
										…
										`norm=np.___`

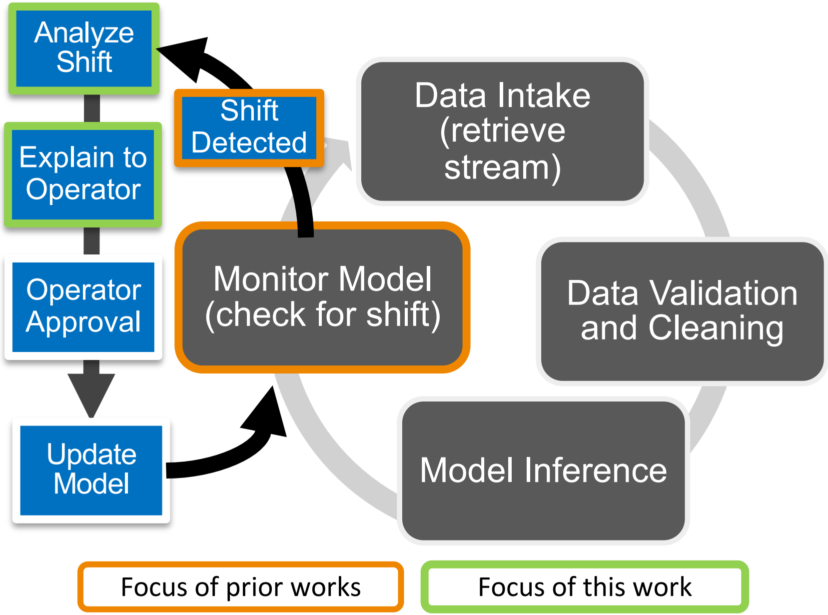
Target example								As a Christian, I will not be patronizing any of those businesses.	I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp`
										`p=sp.Popen()`
										`stdout=p.___`

Exemplar Real-World Distribution Shift datasets from Stanford WILDS benchmarks overview

PURDUE UNIVERSITY

Knowing what has changed under a shift allows us to more **effectively** respond to mitigate the shift

- **Problem**: Most prior works focus on only *detecting* a shift and do not help with "How should I respond?"
- To most effectively mitigate the shift, an operator needs to know what changed
  - E.g, "Preferences of 18-25 year-olds changed" or "X feature of the data intake pipeline is broken"
- **Our goal**: Aid the operator by **explaining** how P_{src} shifted to P_{tgt}

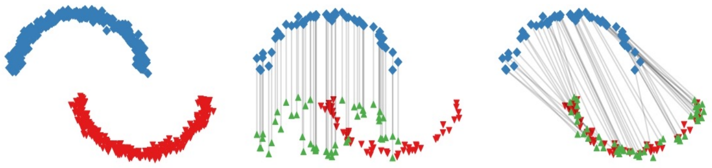


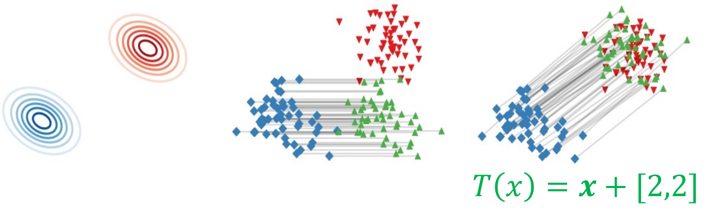
A typical ML deployment cycle

PURDUE UNIVERSITY

Distribution shifts can be explained by hypothesizing how to map P_{src} to P_{tgt}

- Given two distributions P_{src}, P_{tgt}:
  - a transport map $T(\cdot)$, is a function which moves a point from P_{src} to P_{tgt}, such that $P_{T(P_{src})} \approx P_{tgt}$
- If T is interpretable, it can explain how P_{src} shifted to P_{tgt}



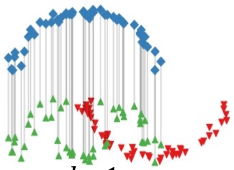


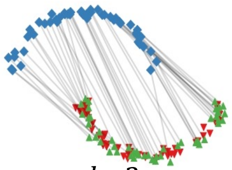
We can leverage prior Optimal Transport work to find **good** interpretable mappings

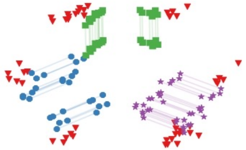
- By relaxing alignment in Optimal Transport and restricting our possible mappings to be interpretable we get *Intrinsically Interpretable Transport*:

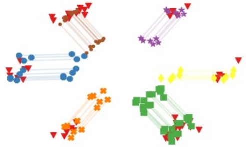
$$T_{IIT} \coloneqq \operatorname{argmin}_{\underbrace{T \in \Omega_{int}}_{\text{Ω_{int}: A set of interpretable mappings}}} \mathbb{E}_{P_{train}} \Big[\underbrace{c\big(\boldsymbol{x}, T(\boldsymbol{x})\big)}_{\text{Cost function: T should retain as much of the original point as possible}}\Big] + \lambda\, \underbrace{\phi\big(P_{T(X)}, P_{test}\big)}_{\text{Divergence: T should align $P_{T(X)}$ and P_{test} as much as possible}}$$

- Ω_{int} can be defined based on context, or one can use our pre-defined mappings: k-sparse feature mappings or k-cluster mappings

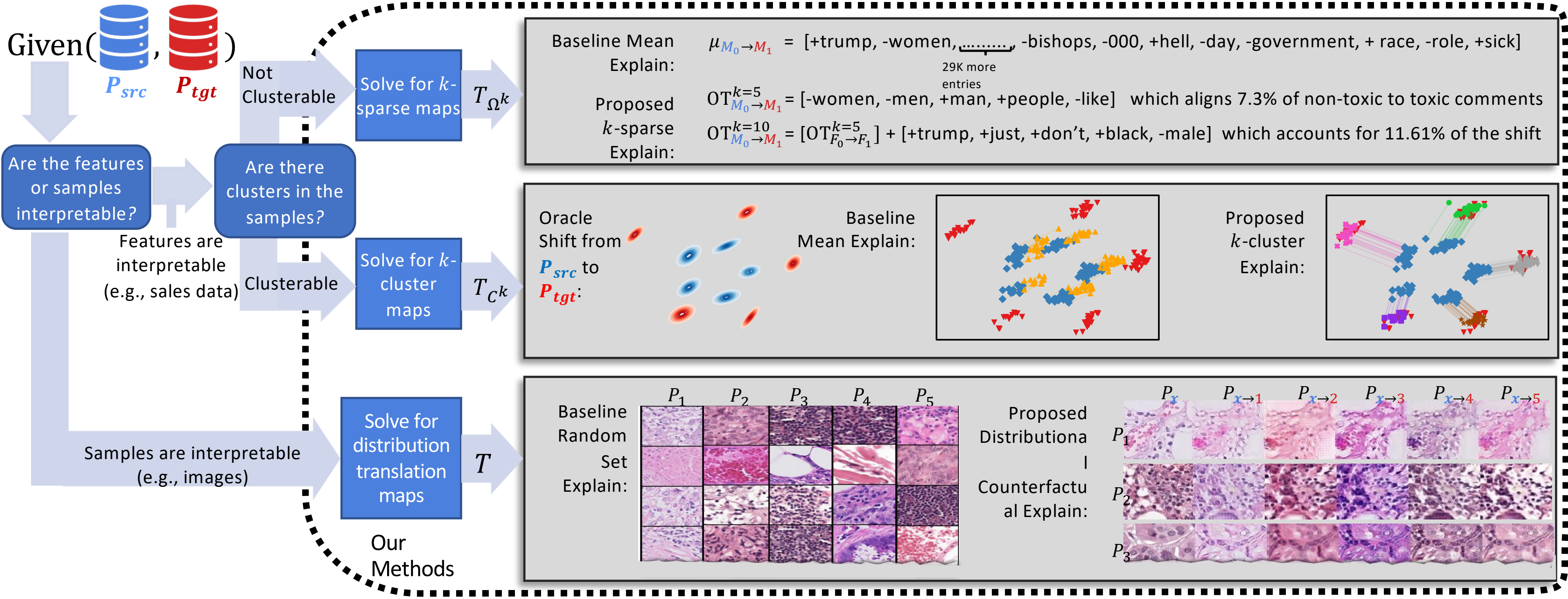

$T^{k=1}_{\text{sparse}}$


$T^{k=2}_{\text{sparse}}$


$T^{k=3}_{\text{cluster}}$


$T^{k=6}_{\text{cluster}}$

Methodology for solving for a shift explanation

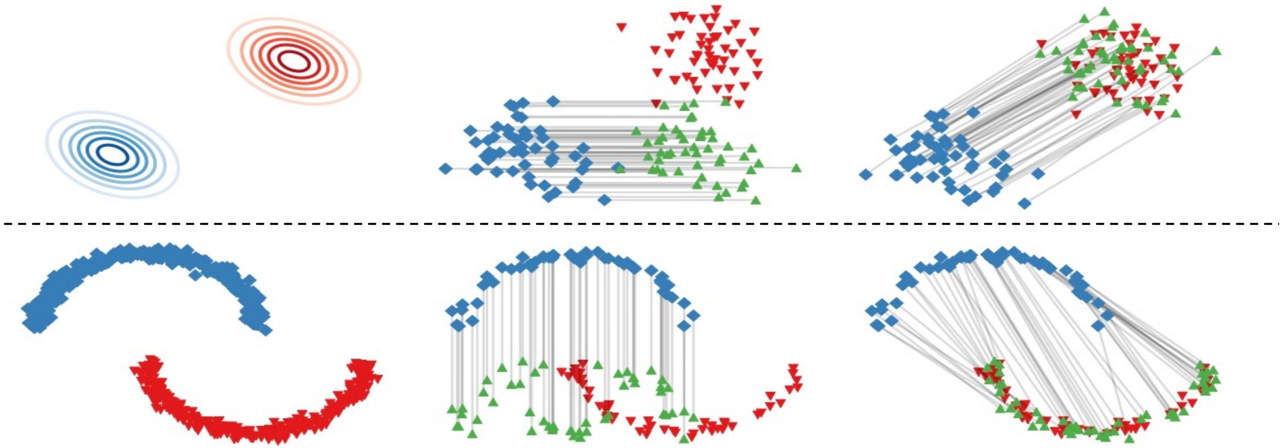


k-Sparse Feature Mappings can show how features moved along defined axes

- Ω_{sparse}^{k}: Find a T which yields the best alignment, while only moving points along k dimensions

Simple mean shift:

Complex conditional shift:



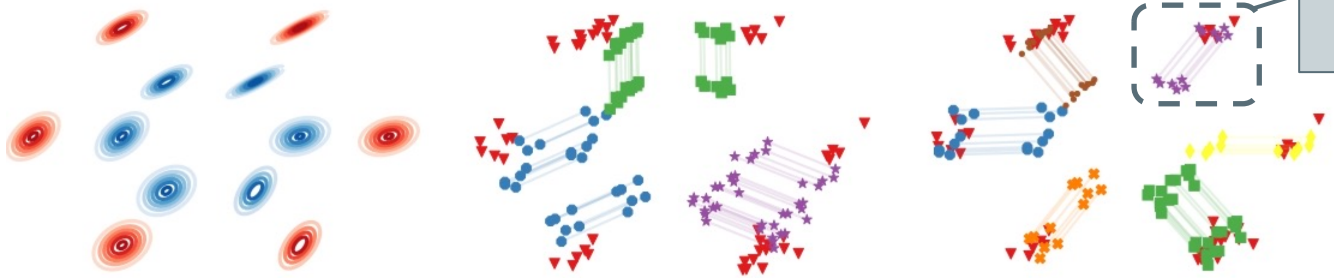
Oracle Shifts from P_{src} to P_{tgt} $\quad T_{\text{sparse}}^{k=1}$ $\quad T_{\text{sparse}}^{k=d}$

PURDUE UNIVERSITY

k-Cluster Mappings can show how heterogenous subgroups have shifted

- $\Omega^k_{\text{cluster}}$: Find k-cluster-specific transport maps which maximizes alignment between $\boldsymbol{P_{T(P_{tgt})}}$ and $\boldsymbol{P_{tgt}}$
  - We can restrict per cluster transport maps to a specific class of transport functions

Moving points by cluster-specific vector: $T^{(\boldsymbol{C})} = \boldsymbol{x} + \boldsymbol{\delta}^{(\boldsymbol{C})}$ if $\boldsymbol{x} \in \boldsymbol{C}$



Oracle Shift from $\boldsymbol{P_{src}}$ to $\boldsymbol{P_{tgt}}$

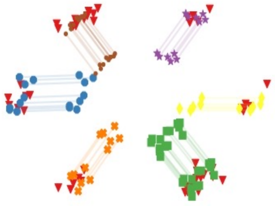
k-cluster mapping $k = 3$

k-cluster mapping $k = 6$

PURDUE UNIVERSITY

T_{IIT} can be used to gain actionable insights from explanations of complex shifts

- Using our k-cluster mappings $\Omega^k_{\text{cluster}}$, we can see how heterogenous groups (clusters) moved differently under a distribution shift

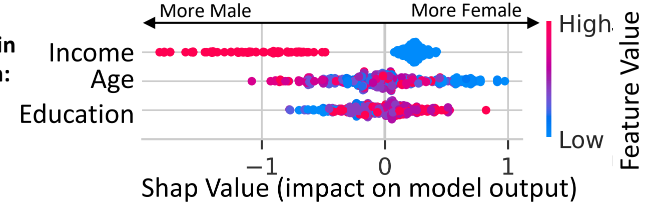


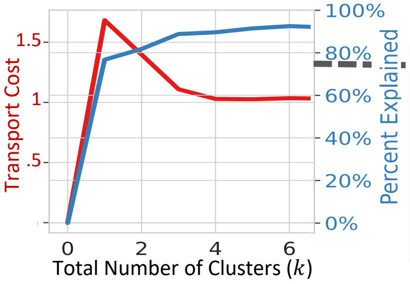
Example of 6-cluster mapping

Baseline Mean Shift Explanation:

μ_M : [Age: 39.7, Edu: 10.2, Inc: 0.29]
μ_F : [Age: 36.6, Edu: 10.1, Inc: 0.10]

Baseline Feature Importance of Domain Classifier Explanation:





k - Cluster Explanation (ours), $k = 4$:

$\mu_{C_M^1}$: [Age: 35.6, Edu: 12.9, Inc: 0.00]
$\mu_{C_M^1 \to C_F^1}$: [Age: 32.1, Edu: 12.5, Inc: 0.01]

$\mu_{C_M^2}$: [Age: 29.7, Edu: 8.9, Inc: 0.03]
$\mu_{C_M^2 \to C_F^2}$: [Age: 26.3, Edu: 9.0, Inc: 0.00]

$\mu_{C_M^3}$: [Age: 56.3, Edu: 8.4, Inc: 0.13]
$\mu_{C_M^3 \to C_F^3}$: [Age: 53.7, Edu: 8.7, Inc: 0.01]

$\mu_{C_M^4}$: [Age: 43.5, Edu: 12.1, Inc: 1.00]
$\mu_{C_M^4 \to C_F^4}$: [Age: 40.2, Edu: 11.9, Inc: 0.38]

Insight 1: Income is largest predictor between M and F

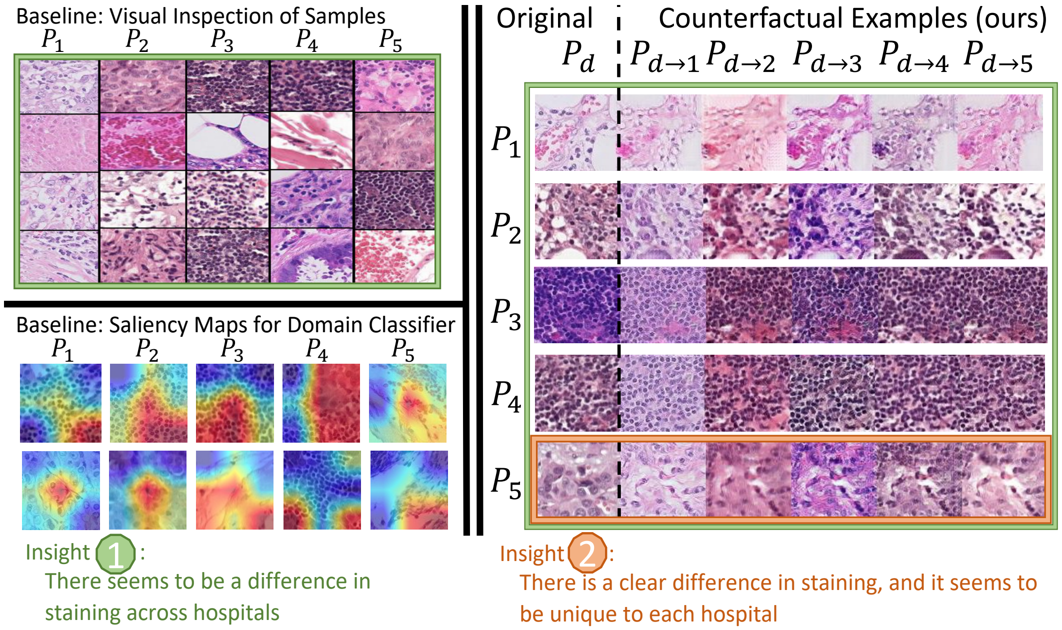
Insight 2: The income difference is largest in M_{C^4}, middle-aged adults with a bachelor's degree

Using $\Omega^k_{\text{cluster}}$ to compare male and female response to the US 1994 Census

Transport Maps can also explain distribution shifts in high-dimensional regimes (images)

- When raw features are not semantically meaningful, but samples are (e.g., images), we can use *domain counterfactuals* to understand a complicated T
- Distributional-Counterfactuals $:=$

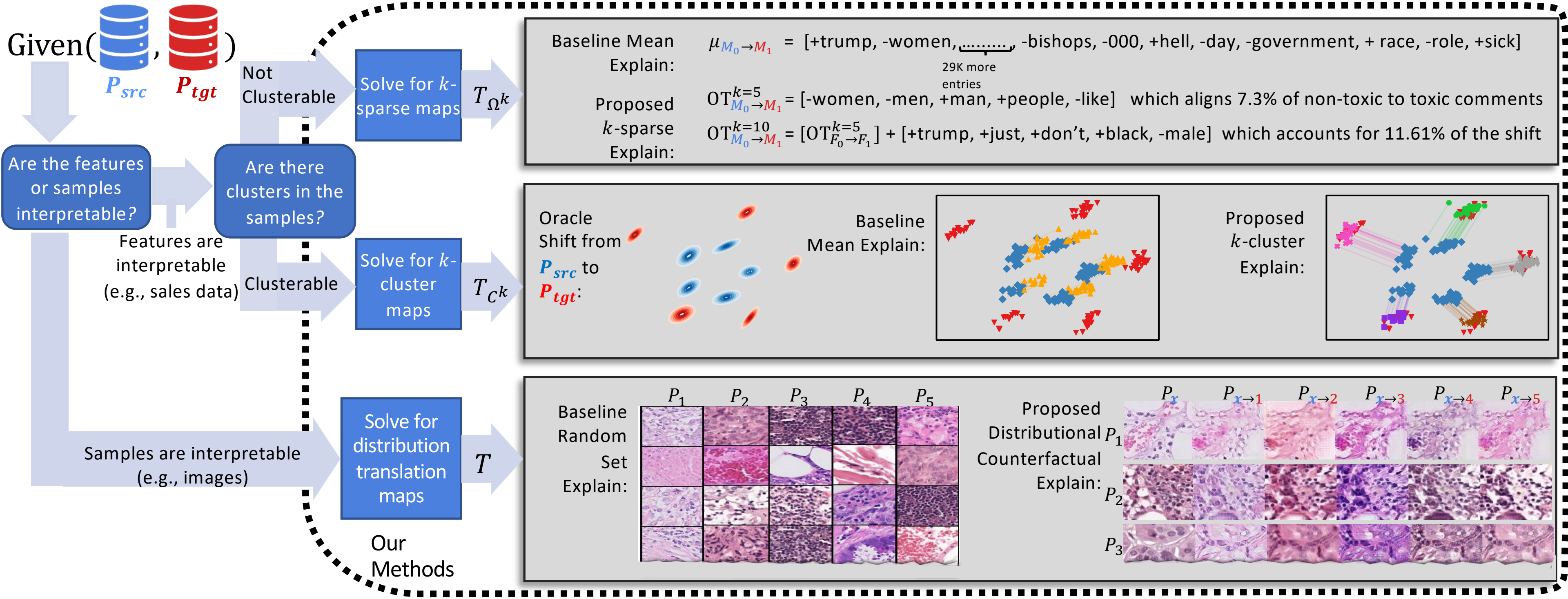
$$\{\boldsymbol{x}, T(\boldsymbol{x}) : x \sim {\color{blue}P_{src}}, T(\boldsymbol{x}) \sim {\color{red}P_{tgt}}\}$$



Using StarGAN to show the difference between tissue samples across 5 hospitals

PURDUE UNIVERSITY

Methodology for solving for a shift explanation



PURDUE UNIVERSITY

Thank you for listening!

Code

Towards Explaining Distribution Shifts

Paper


Sean Kulinski


David Inouye