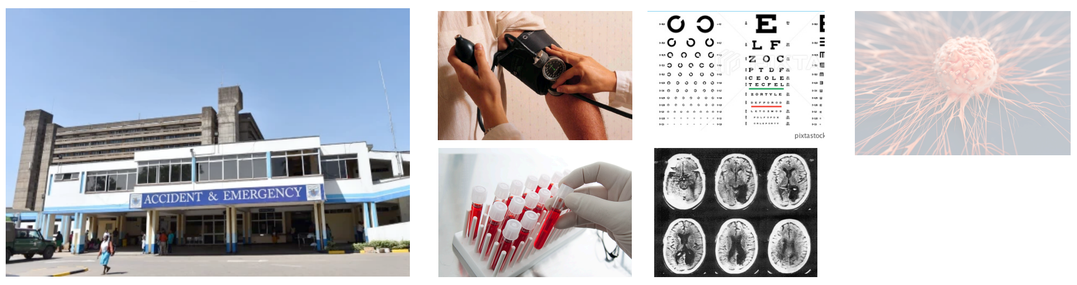
Adapting to Shifts in Latent Confounders via Observed Concepts and Proxies

Matt J. Kusner[1,2], Ibrahim Alabdulmohsin[1], Stephen Pfohl[1], Olawale Salaudeen[1], Arthur Gretton[2],
Sanmi Koyejo[1*], Jessica Schrouff[1*], Alexander D'Amour[1*]
[1]Google Research
[2]University College London
[*]equal contribution

Adapting to New Domains

hospital P — **expensive features** $x \in X$ — **label** $y \in Y$

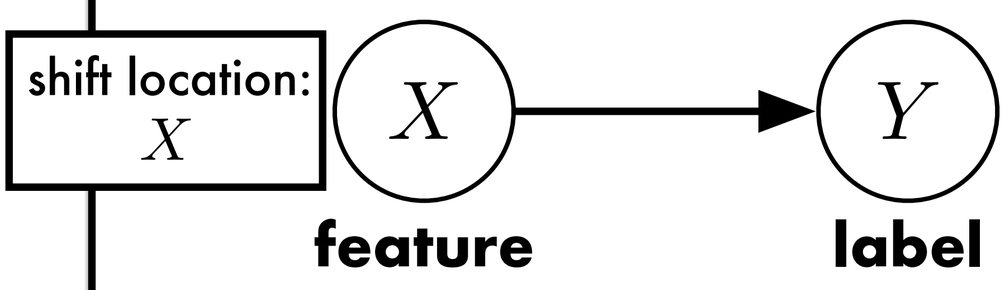
hospital Q — **cheap features** $x' \in X$ — **label (unobserved)** $y' \in Y$



Goal: learn $q(Y|X)$ using $\{x'\} \in Q$
without assumptions, this is impossible!

Popular Assumption #1: **Covariate Shift**

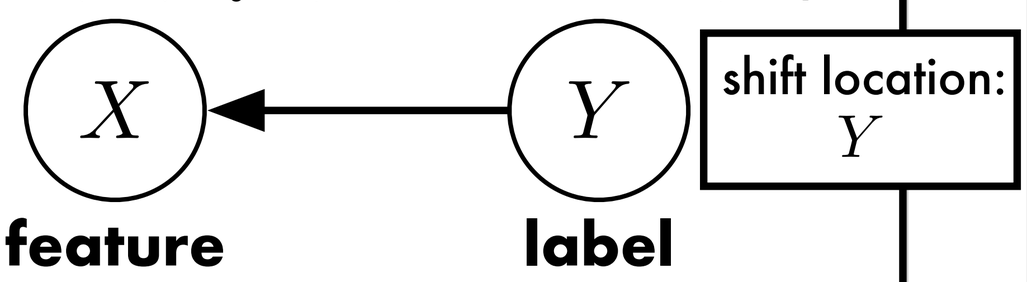
[Shimodaira, 2000; Zadrozny, 2004; Huang et al., 2006; Gretton et al., 2009; Bickel et al., 2009; Chen et al., 2016; Schneider et al., 2020]



$$p(X) \neq q(X) \Rightarrow p(Y|X) = q(Y|X)$$

Popular Assumption #2: **Label Shift**

[Gart & Buck, 1966; Manski & Lerman, 1977; Rosenbaum & Rubin, 1983; Saerens et al., 2002; Forman, 2008; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Garg et al., 2020; Tachet des Combes et al., 2020]

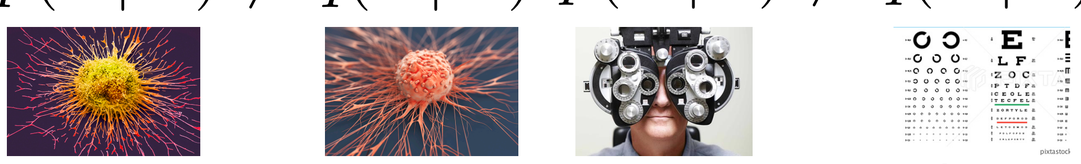


$$p(Y) \neq q(Y) \Rightarrow p(X|Y) = q(X|Y)$$

What about our example?

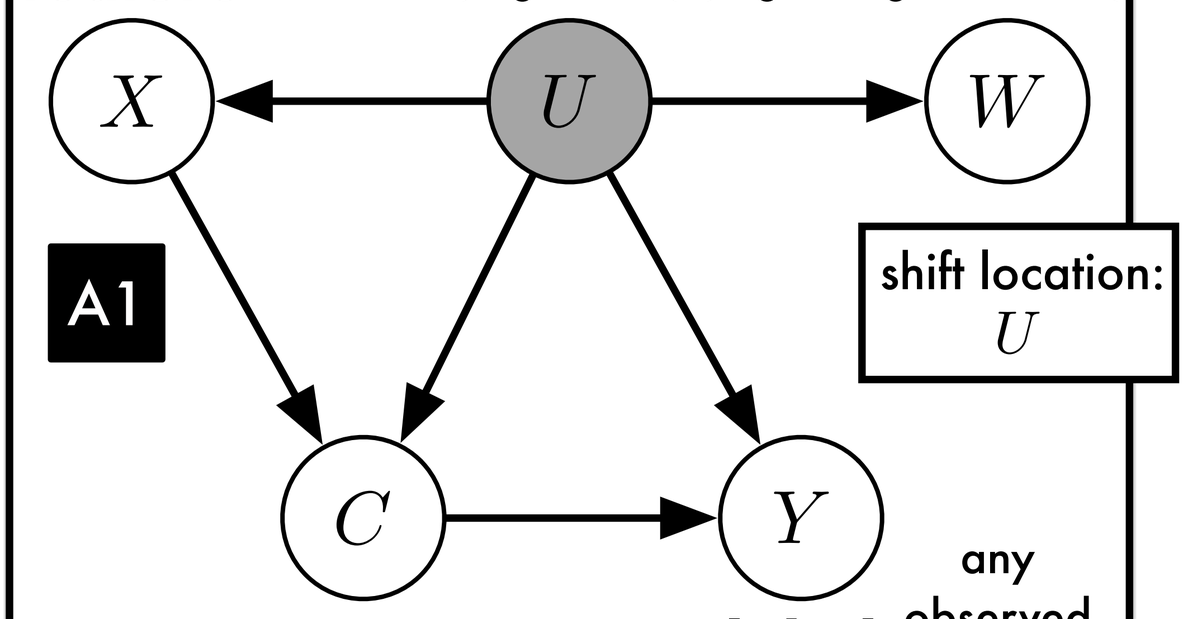
Social Determinants of Health (SHD)
(income, education, discrimination, etc...)

changes diagnoses	changes frequency of hospital visits
$p(Y\|X) \neq q(Y\|X)$	$p(X\|Y) \neq q(X\|Y)$



"cancer!" — "just a mole..." — "severe symptoms!" — "nothing to worry about..."

Solution: New Assumptions



A2 $p(U) \neq q(U) \Rightarrow p(V|V', U) = q(V|V', U)$

A3 **all data is discrete,** $\dim(X), \dim(W) \geq \dim(U)$

A4 **conditional probs. are unique,** $\forall i \in U, p(i) > 0 \text{ if } q(i) > 0$

given the above assumptions, all p.m.f.s p of P over $W, X, C, Y, \tilde{U}$ are identifiable!

Decomposition of $q(Y|X)$

unknown sorting of U

$$q(Y|X) = \sum_{i=1}^{k_U} q(Y|X, U=i) q(U=i|X)$$

$$\overset{\text{A2}}{=} \sum_{i=1}^{k_U} p(Y|X, \tilde{U}=i) \frac{q(X|\tilde{U}=i) q(\tilde{U}=i)}{q(X)}$$

$$\overset{\text{A2}}{=} \sum_{i=1}^{k_U} p(Y|X, \tilde{U}=i) \frac{p(\tilde{U}=i|X) p(X) q(\tilde{U}=i)}{p(\tilde{U}=i) q(X)}$$

$$\propto \sum_{i=1}^{k_U} p(Y|X, \tilde{U}=i) p(\tilde{U}=i|X) \frac{q(\tilde{U}=i)}{p(\tilde{U}=i)}$$

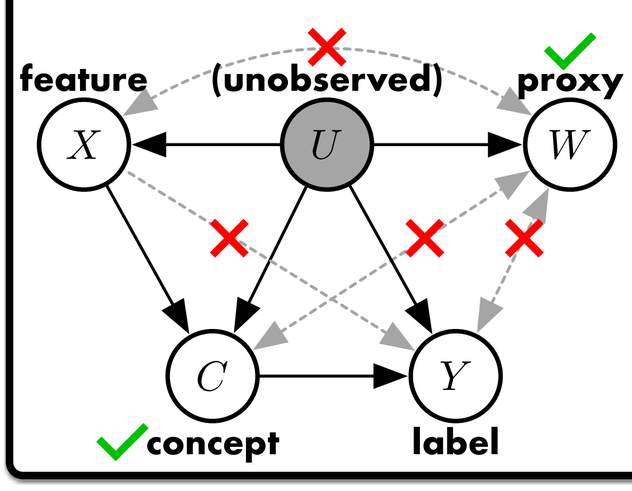
identify via label shift methods!

can estimate all p.m.f.s involving $\tilde{U}$ via eigendecomposition $\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}^{-1}\Delta\mathbf{S}$

entries are p.m.f.s over observed vars.

entries are used to recover all p.m.f.s involving $\tilde{U}$

Uniqueness



Simulation