TRAIL Brownbag
Covariate Shift Tutorial

Minghe Wang
Biostatistics MS student

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

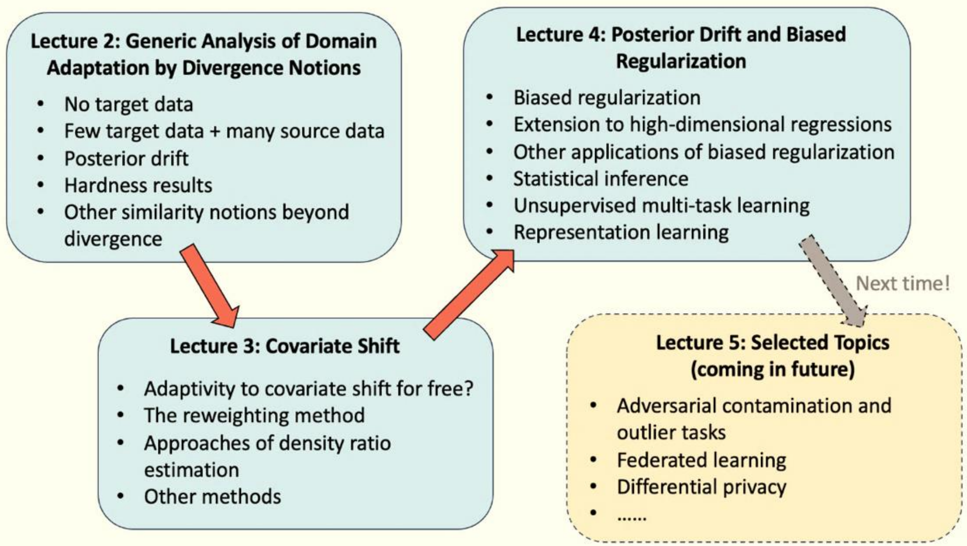
Acknowledgement

This tutorial is modified based on the Theoretical Covariate Shift Lecture from Ye Tian

A (Selective) Introduction to the Statistics Foundations of Transfer Learning

Time: 9AM-5PM, May 20, 2024

Schedule

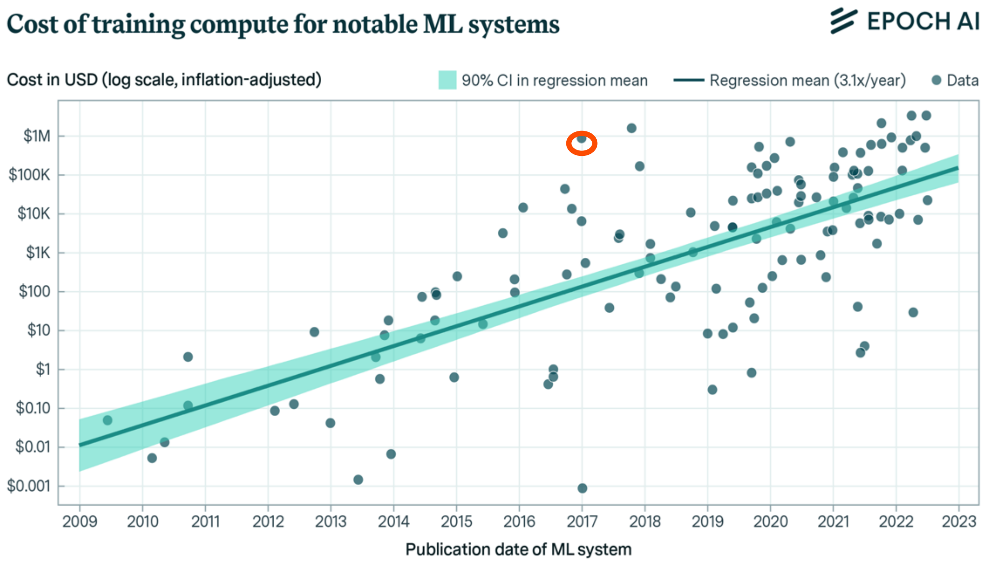


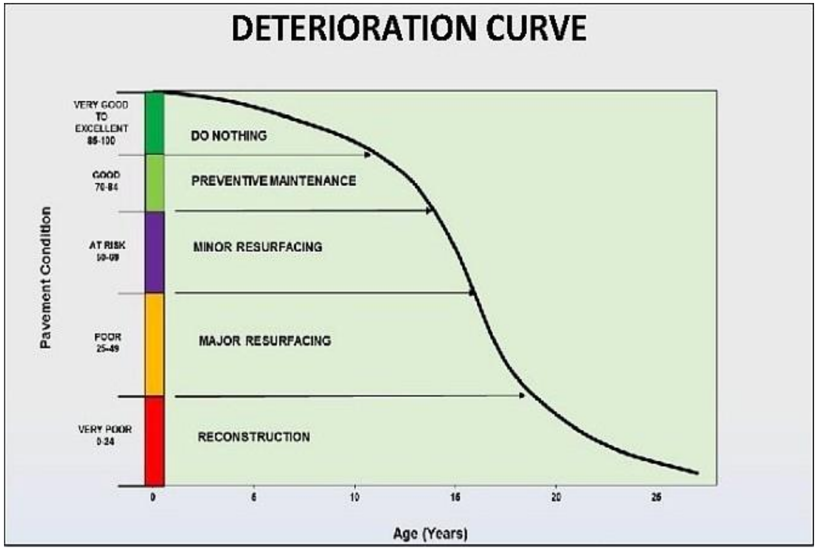
https://www.columbia.edu/~yt2661/STL.html

Machine Learning Dilemma

Machine learning is powerful but **expensive**, because it requires:

- large labeled datasets
- significant computing power and training time





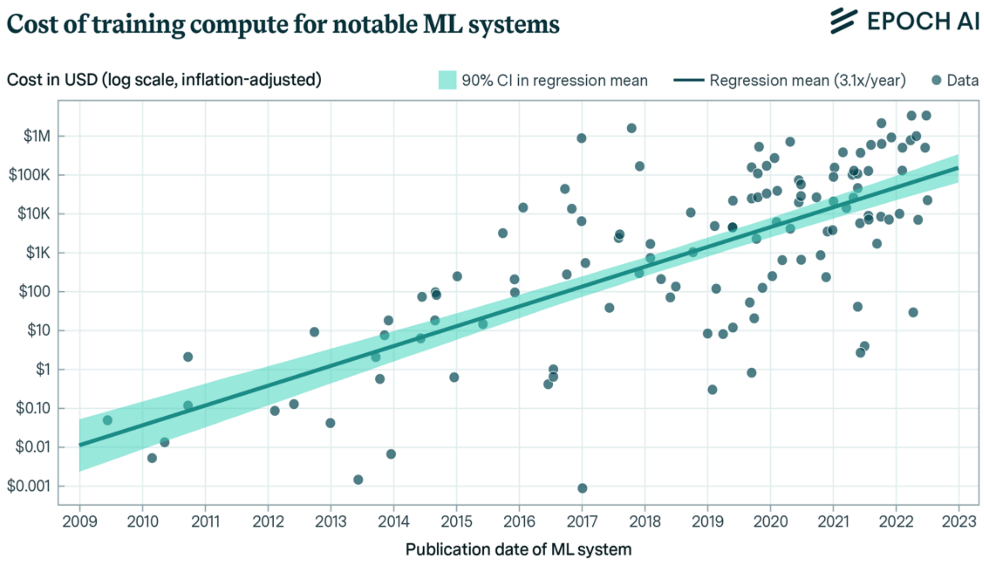
Ruiz, Victor M. et al.
The Journal of Thoracic and Cardiovascular Surgery, Volume 164, Issue 1, 211 - 222.e3

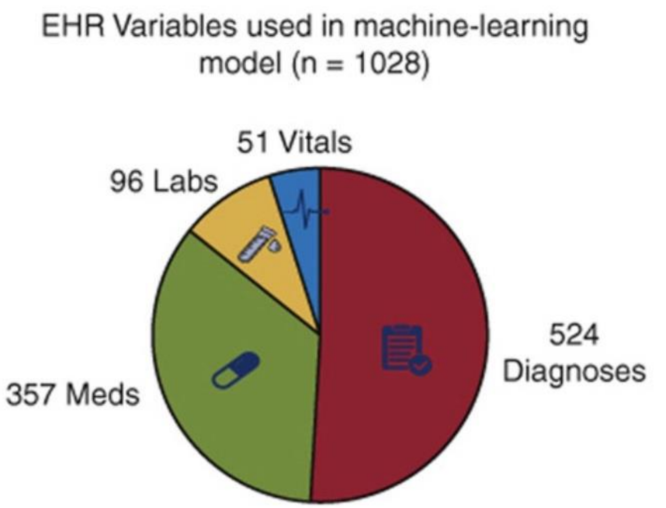
Machine Learning Dilemma

Machine learning is powerful but **expensive,** because it requires:

- large labeled datasets
- significant computing power and training time

Annotation (eg. chart review) by doctors is costly





Ruiz, Victor M. et al.
The Journal of Thoracic and Cardiovascular Surgery, Volume 164, Issue 1, 211 - 222.e3

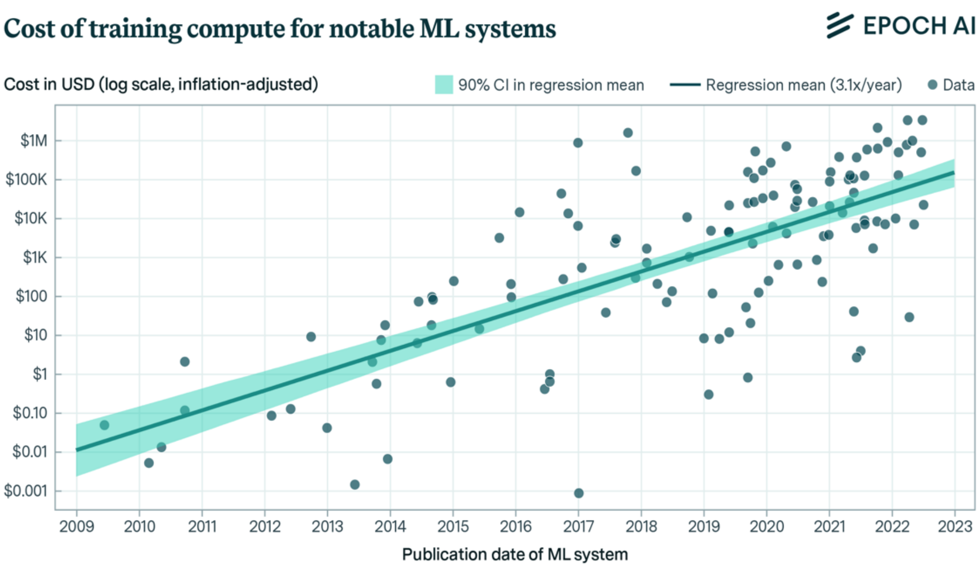
Machine Learning Dilemma

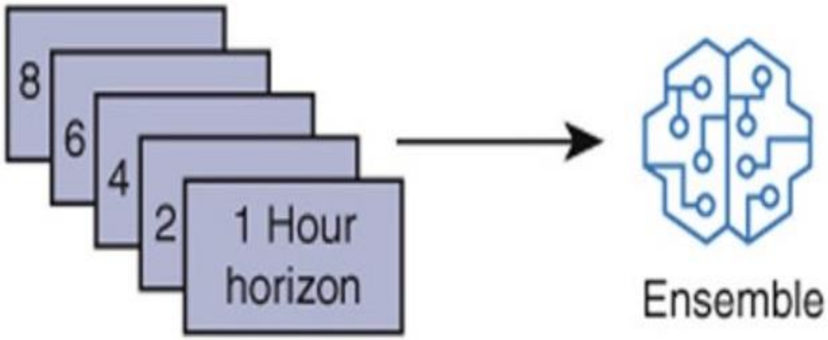
Machine learning is powerful but **expensive,** because it requires:

- large labeled datasets
- significant computing power and training time

Annotation (eg. chart review) by doctors is costly

Complex models trained on free text diagnosis

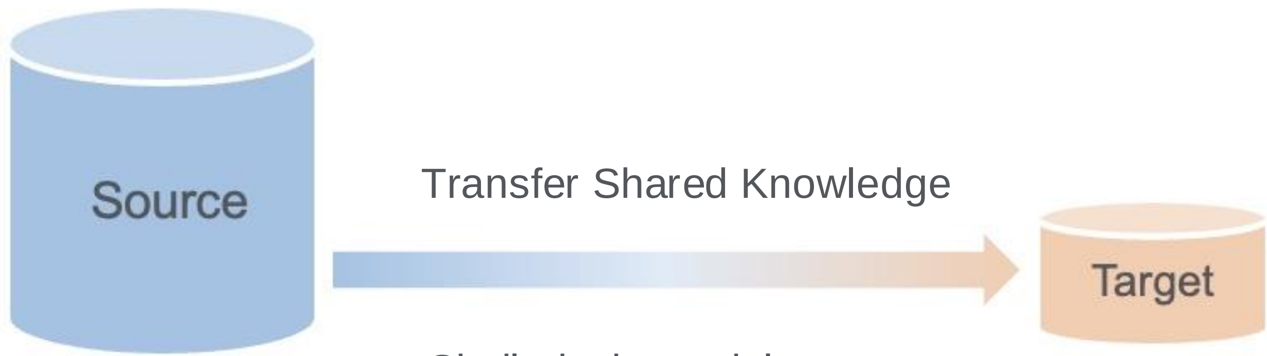




Ruiz, Victor M. et al.
The Journal of Thoracic and Cardiovascular Surgery, Volume 164, Issue 1, 211 - 222.e3

Transfer Learning

Leverage knowledge from a model trained for one task and reused to help with a similar, related task



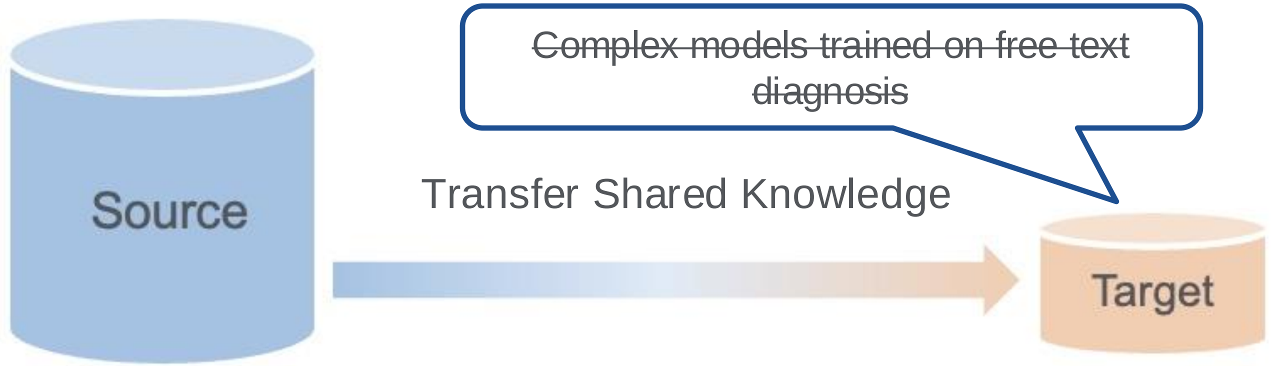
- Similarity in model parameters
- Shared model structure
- Shared embedding structure
- …

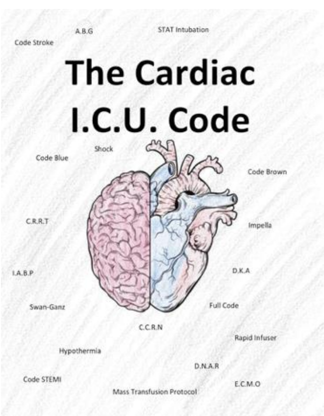
Transfer Learning

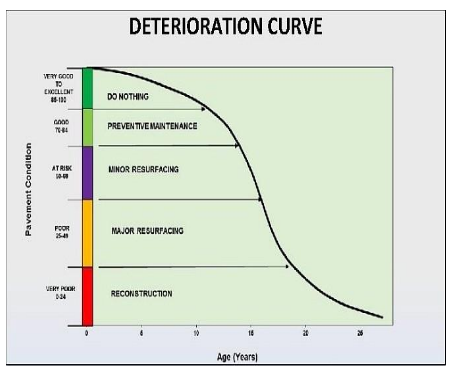
Leverage knowledge from a model trained for one task and reused to help with a similar, related task

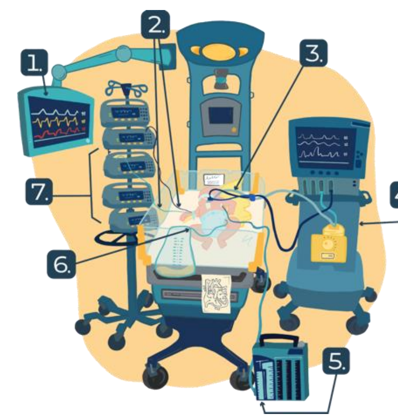
~~Annotation(eg. chart review) by doctor is costly~~

~~Complex models trained on free text diagnosis~~





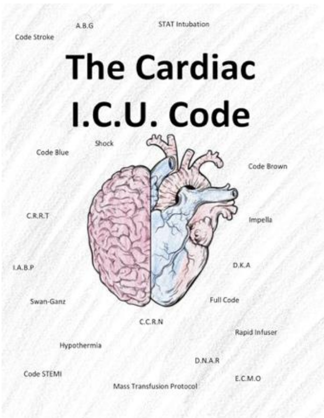


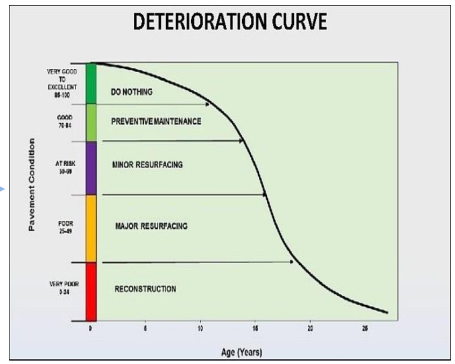


Target Deterioration Curve

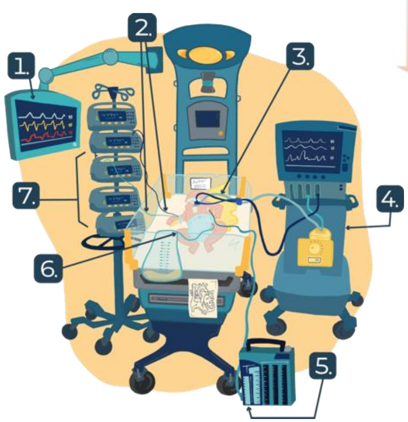
Transfer Learning

Difference between source and target dataset:



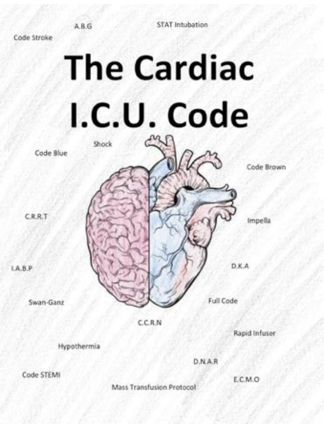


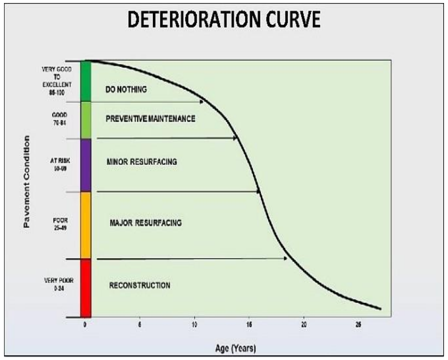
- Sample size
  Rich vs Scarce
- Variables/Features
  Young vs Old

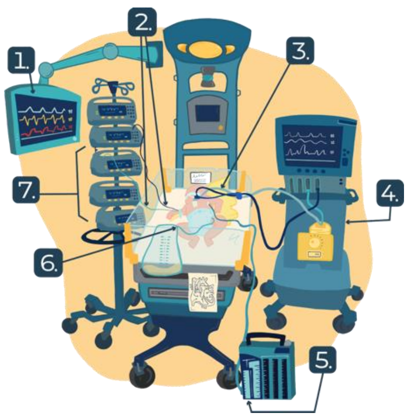


Target
Deterioration
Curve

Transfer Learning







Target Deterioration Curve

Domain Adaptation: address the challenge of training a model on one data distribution and applying it to a related but different data distribution

Domain Adaptation (a solution)

- Model finetuning
- Covariate shift adjustment
- Learning invariant representations
- …

Distributional Shift

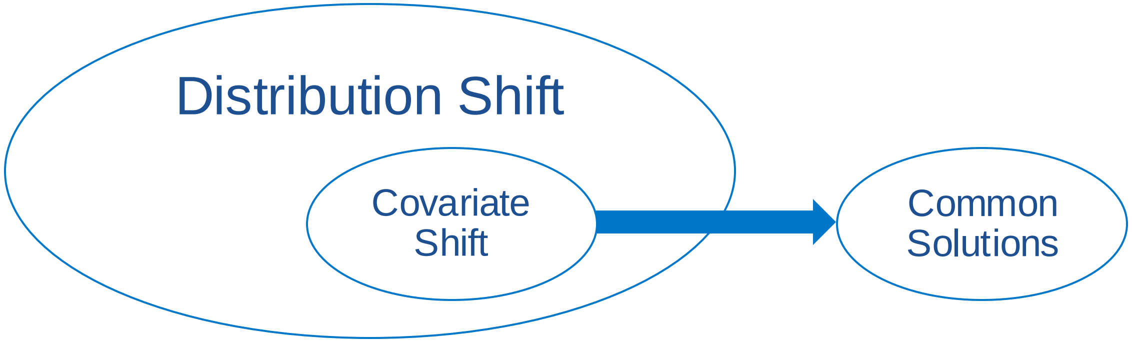
- Target T, $(X,Y)_T$
  Source S, $(X,Y)_S$
- Probability Chain Rule: $\boxed{P(X,Y)} = P(Y|X)\ P(X)$
- Distribution Shift types:
  - Joint Shift: $P_S(X,Y) \neq P_T(X,Y)$

Distributional Shift

- Target T, $(X, Y)_T$
  Source S, $(X, Y)_S$
- Probability Chain Rule: $P(X, Y) = \boxed{P(Y|X)}\, P(X)$
- Distribution Shift types:
  - Joint Shift: $P_S(X, Y) \neq P_T(X, Y)$
  - Model Shift: $P_S(Y|X) \neq P_T(Y|X)$

Distributional Shift

- Target T, $(X, Y)_T$
  Source S, $(X, Y)_S$
- Probability Chain Rule: $P(X, Y) = P(Y|X)\,\boxed{P(X)}$
- Distribution Shift types:
  - Joint Shift: $P_S(X, Y) \neq P_T(X, Y)$
  - Model Shift: $P_S(Y|X) \neq P_T(Y|X)$
  - **Covariate Shift:** $P_S(X) \neq P_T(X)$



Tutorial Plan

- Separate Density Estimation
    - Kernel Density Estimation
    - Histogram-based Method

April 18th

- Estimating Weight as a whole
    - Kernel Mean Matching
    - Least Square Method
    - Kullback-Leibler Method
    - Discriminative Learning
    - Profile Likelihood Method
- Beyond Reweighting Technique and Covariate Shift
    - Marginal Transfer Learning
    - Domain Invariant Method
    - Optimal Transformation

Tutorial Plan

- Separate Density Estimation
  - Kernel Density Estimation
  - Histogram-based Method
- Estimating Weight as a whole
  - Kernel Mean Matching
  - Least Square Method
  - Kullback-Leibler Method
  - Discriminative Learning
  - Profile Likelihood Method
- Beyond Reweighting Technique and Covariate Shift
  - Marginal Transfer Learning
  - Domain Invariant Method
  - Optimal Transformation

April 25th

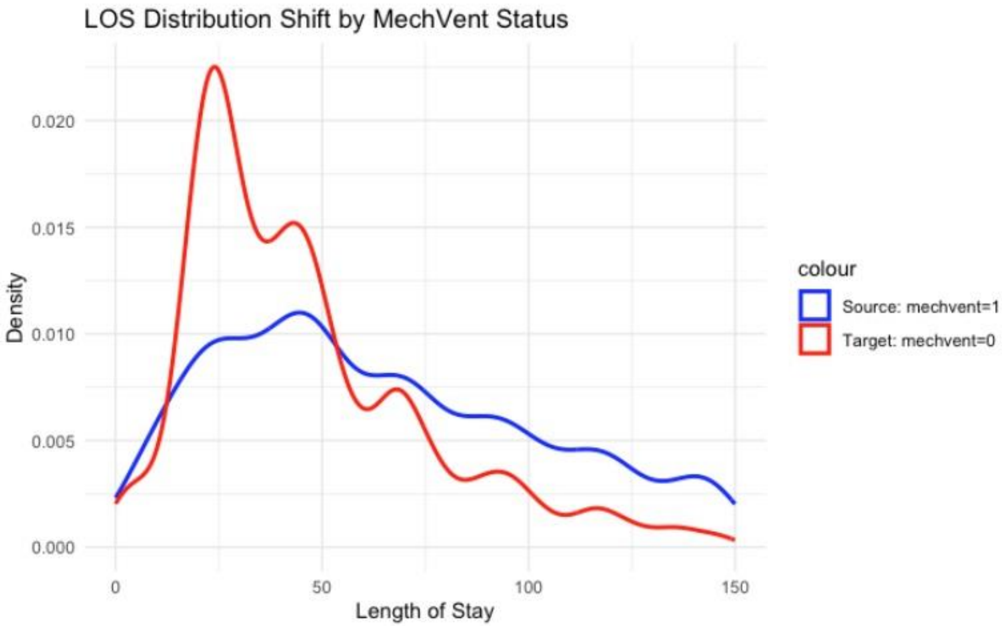
Tutorial Plan

- Separate Density Estimation
  - Kernel Density Estimation
  - Histogram-based Method
- Estimating Weight as a whole
  - Kernel Mean Matching
  - Least Square Method
  - Kullback-Leibler Method
  - Discriminative Learning

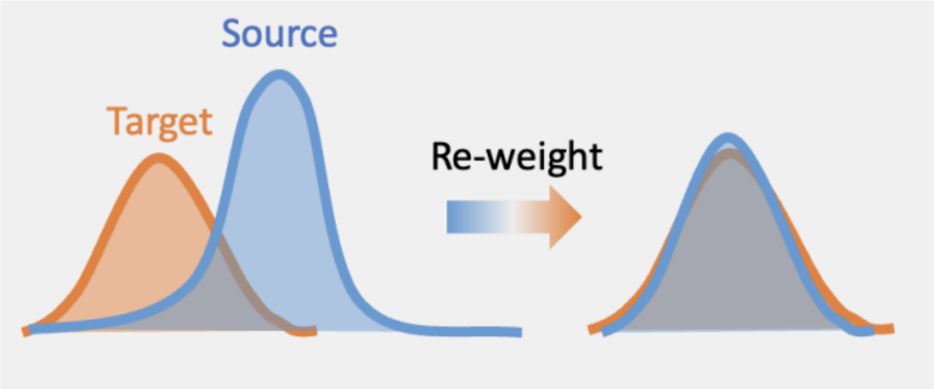
Fall 2025

  - Profile Likelihood Method
- Beyond Reweighting Technique and Covariate Shift
  - Marginal Transfer Learning
  - Domain Invariant Method
  - Optimal Transformation

Covariate Shift



Importance Weighting Framework



Density ratio model: $\omega(\boldsymbol{X}) = \frac{P^{\mathcal{T}}(\boldsymbol{X})}{P^{\mathcal{S}}(\boldsymbol{X})}$

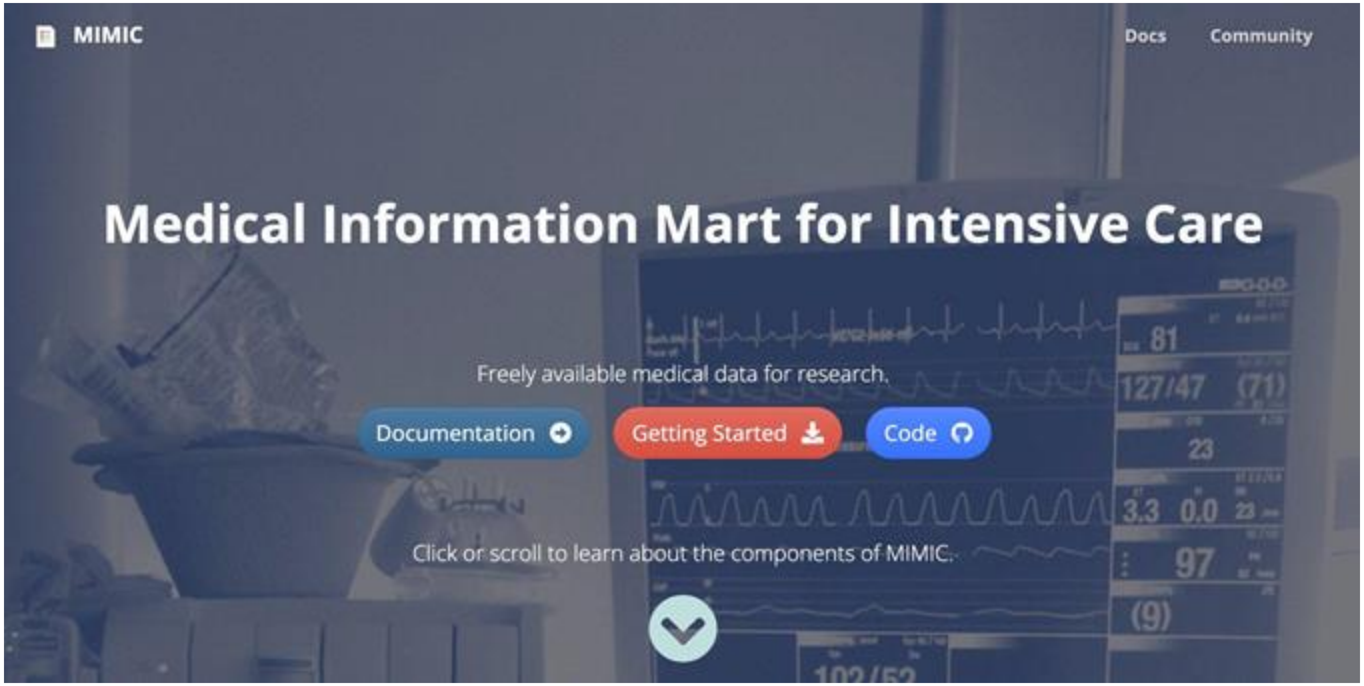
$$\mathrm{E}^{\mathcal{T}}[f(\boldsymbol{X})] = \mathrm{E}^{\mathcal{S}}[\omega(\boldsymbol{X})f(\boldsymbol{X})]$$

holds for any $f(\boldsymbol{X})$

$$\int f(\boldsymbol{X})\, P^{\mathcal{T}}(\boldsymbol{X})\, dX = \int \frac{P^{\mathcal{T}}(\boldsymbol{X})}{\cancel{P^{\mathcal{S}}(\boldsymbol{X})}} f(\boldsymbol{X})\, \cancel{P^{\mathcal{S}}(\boldsymbol{X})}\, dX$$

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

MIMIC-III Dataset

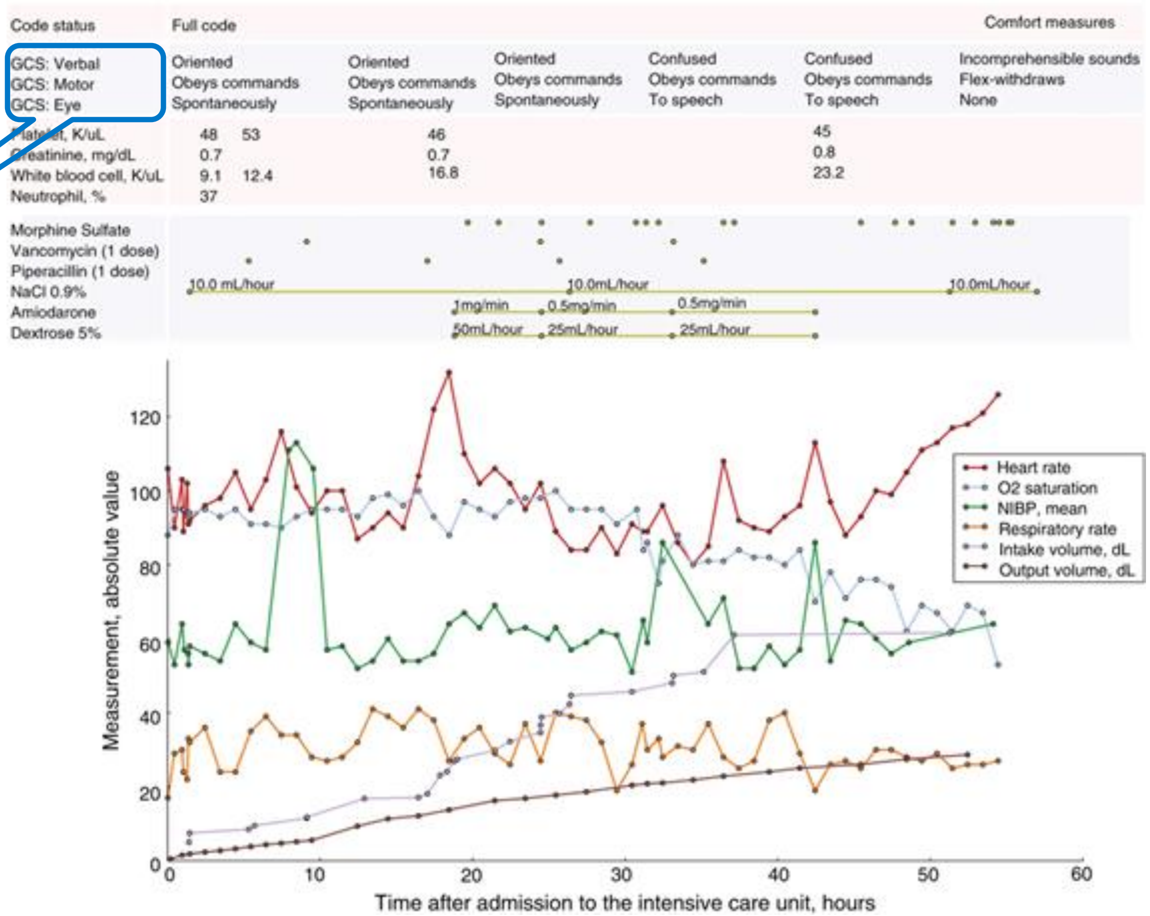


- PhysioNet: https://physionet.org/
  - The Research Resource for Complex Physiologic Signals
  - Need to get credentialed and trained

MIMIC-III Dataset

Sample data for a single patient stay in a medical intensive care Unit including:

- Consciousness
- Blood test result
- Drug
- Vital sign measurement

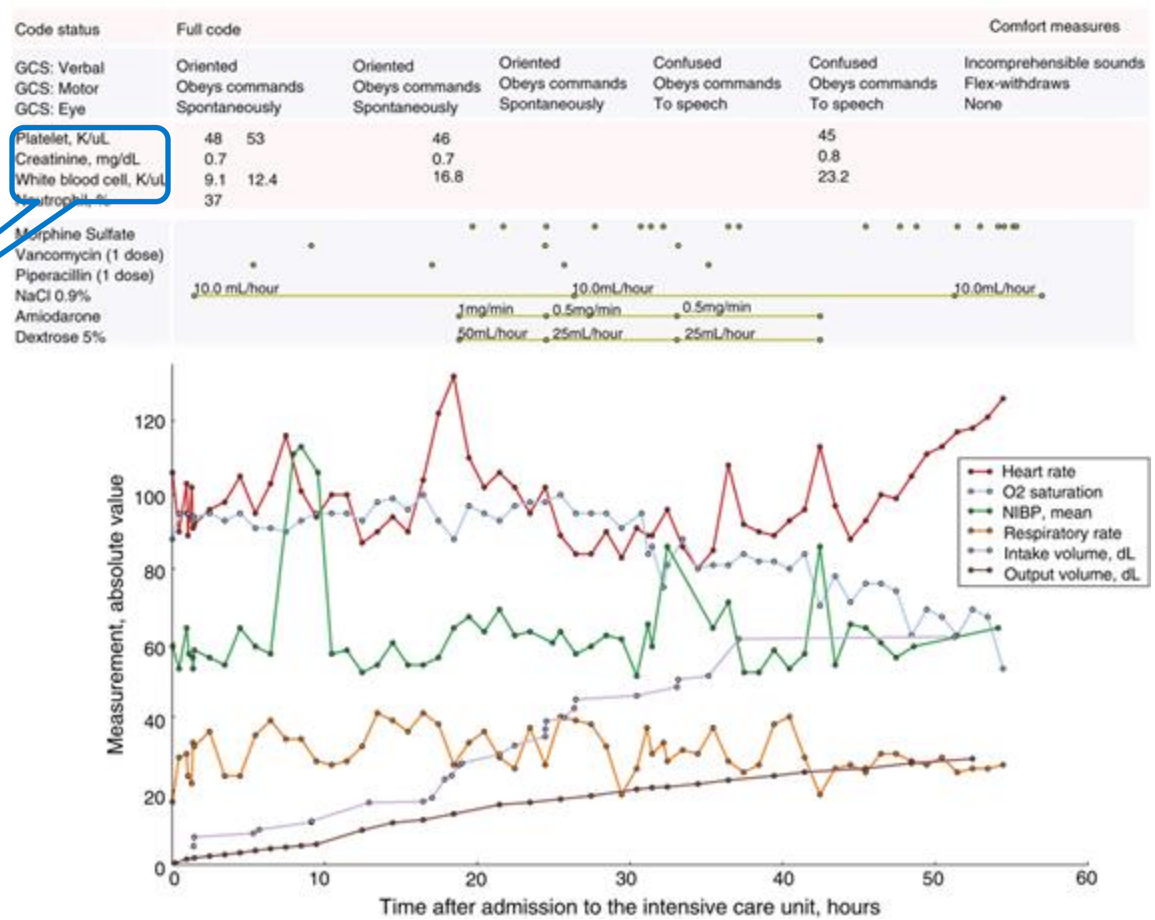


https://www.nature.com/articles/sdata201635#Fig2

MIMIC-III Dataset

Sample data for a single patient stay in a medical intensive care Unit including:

- Consciousness
- Blood test result
- Drug
- Vital sign measurement

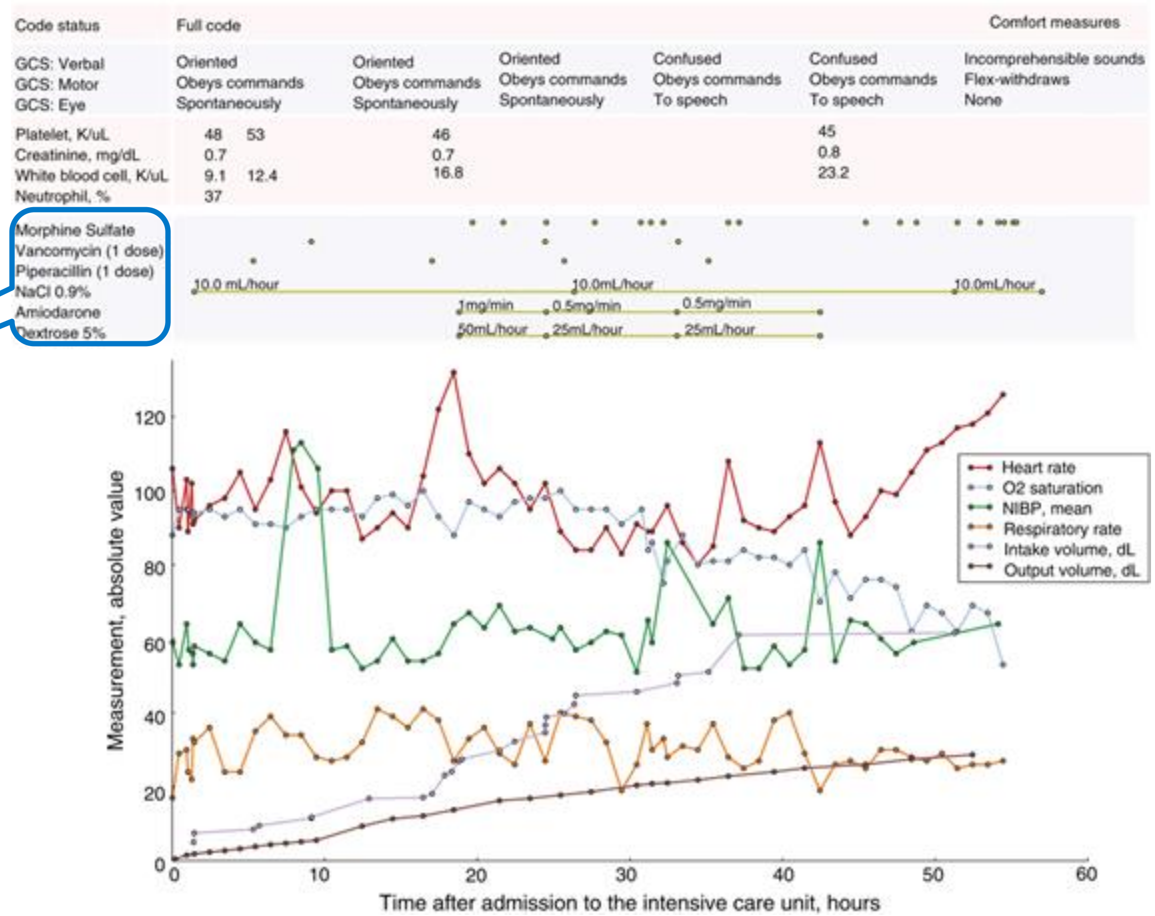


https://www.nature.com/articles/sdata201635#Fig2

MIMIC-III Dataset

Sample data for a single patient stay in a medical intensive care Unit including:

- Consciousness
- Blood test result
- Drug
- Vital sign measurement

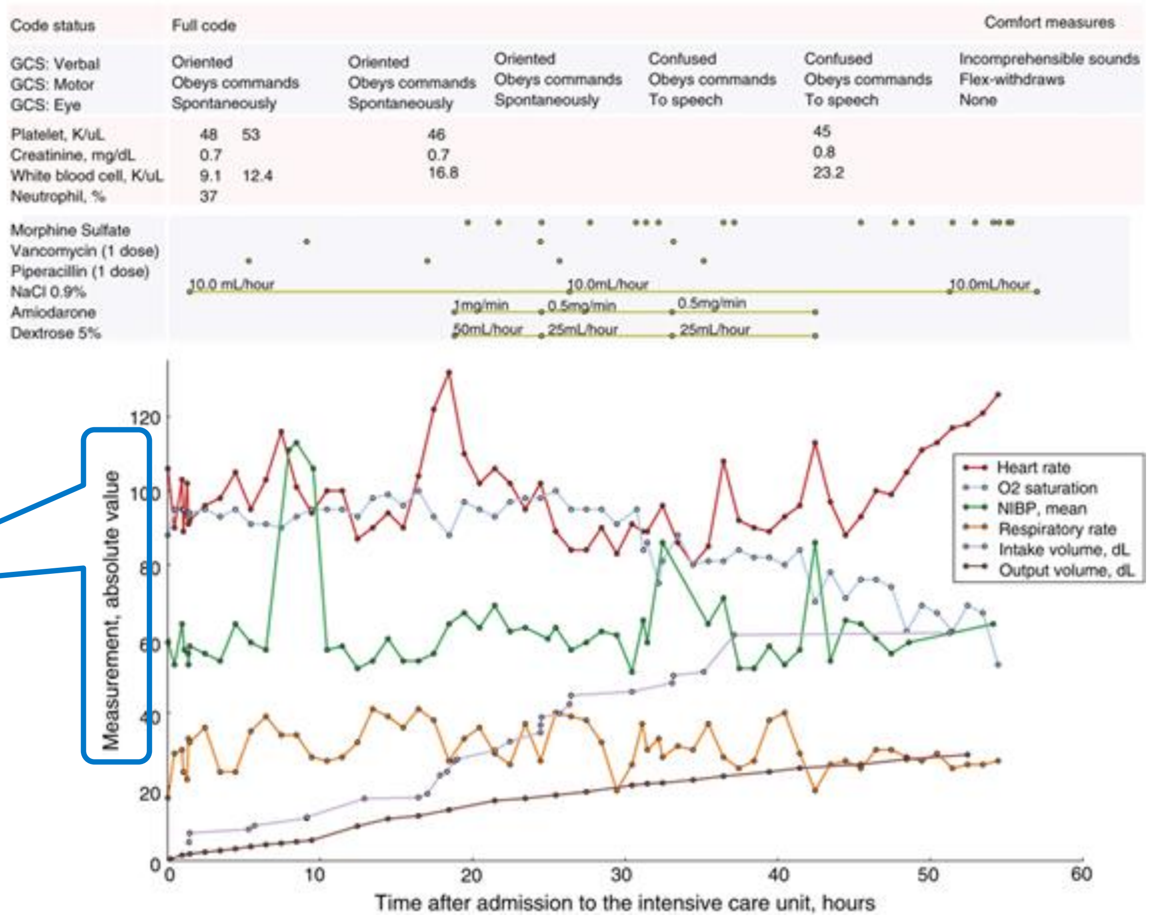


https://www.nature.com/articles/sdata201635#Fig2

MIMIC-III Dataset

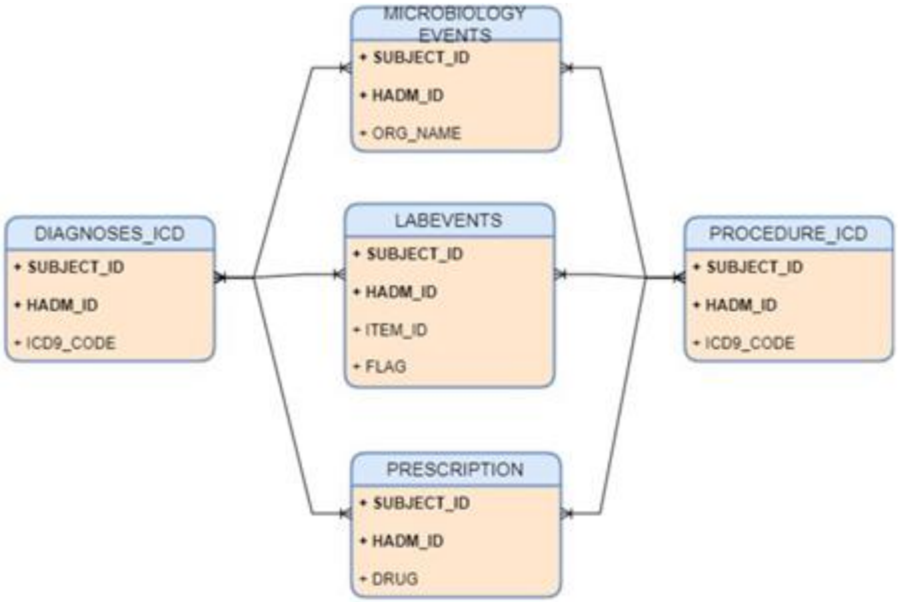
Sample data for a single patient stay in a medical intensive care Unit including:

- Consciousness
- Blood test result
- Drug
- Vital sign measurement



https://www.nature.com/articles/sdata201635#Fig2

MIMIC-III Dataset



'Transfer': Physical locations for patients throughout their hospital stay

'service'(med_service_only): Lists services that a patient was admitted/transferred under.

'Callout': Provides information when a patient was READY for discharge from the ICU, and when the patient was actually discharged from the ICU.

'Patients': Defines each SUBJECT_ID in the database, i.e. defines a single patient.(demographic)

'Elixhauser': Elixhauser Comorbidity Index (ECI) is a tool used to assess the severity and number of chronic health conditions in a patient

'Diagnoses_icd': Definition table for ICD diagnoses.

'Drgcodes': Contains diagnosis related groups (DRG) codes for patients.

https://physionet.org/content/mimiciii-demo/1.4/

Separate Density Estimation

- Separate Density Estimation
  - Kernel Density Estimation
  - Histogram-based Method

April 18th

- Estimating Weight as a whole
  - Kernel Mean Matching
  - Least Square Method
  - Kullback-Leibler Method

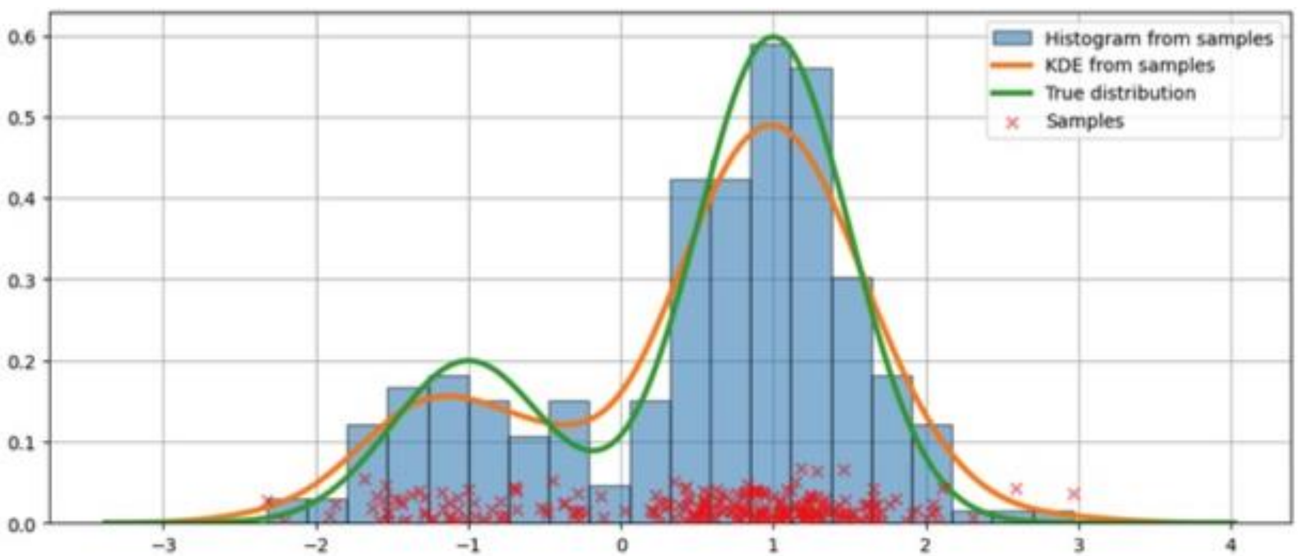
d

ue and Covariate Shift

  - Marginal Transfer Learning
  - Domain Invariant Method
  - Optimal Transformation

$$\omega(\boldsymbol{X}) = \frac{P^{\mathcal{T}}(\boldsymbol{X})}{P^{\mathcal{S}}(\boldsymbol{X})} = \frac{f_T(\boldsymbol{X})}{f_S(\boldsymbol{X})}$$

- Parametric density estimation
- Non-parametric density estimation

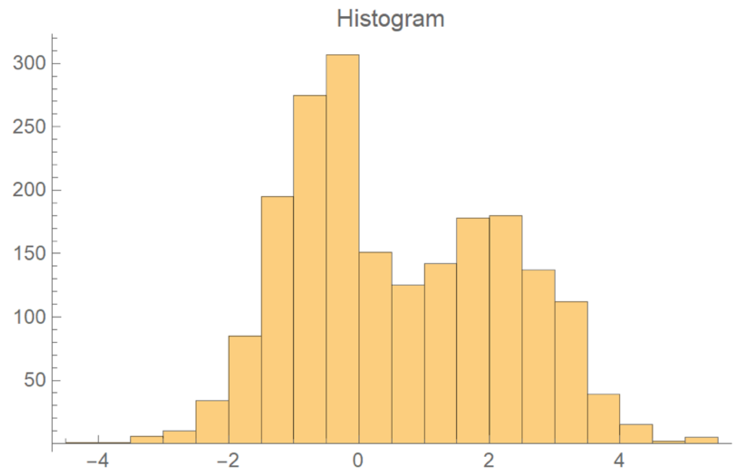


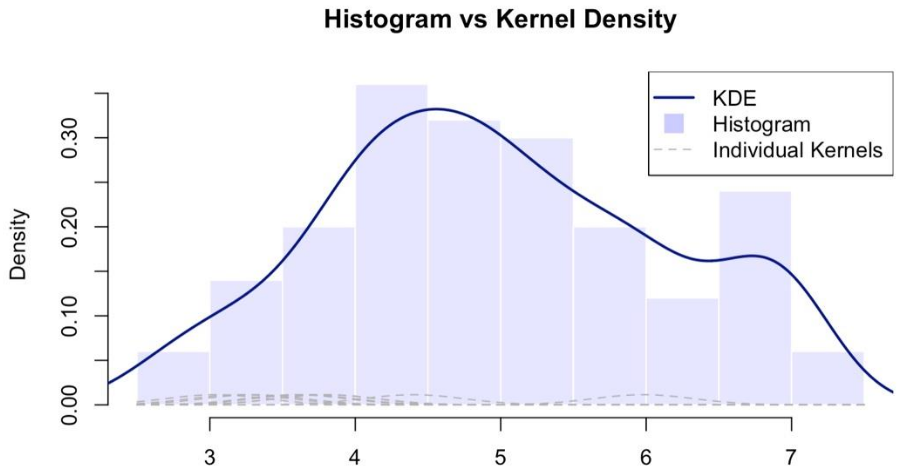
https://www.columbia.edu/~yt2661/STL/slides/Lecture-3.pdf

https://minghe4419.github.io/covariate_shift.github.io/separate_density_estimation.html

Kernel Density Estimation (KDE)

A non-parametric probability density estimation.

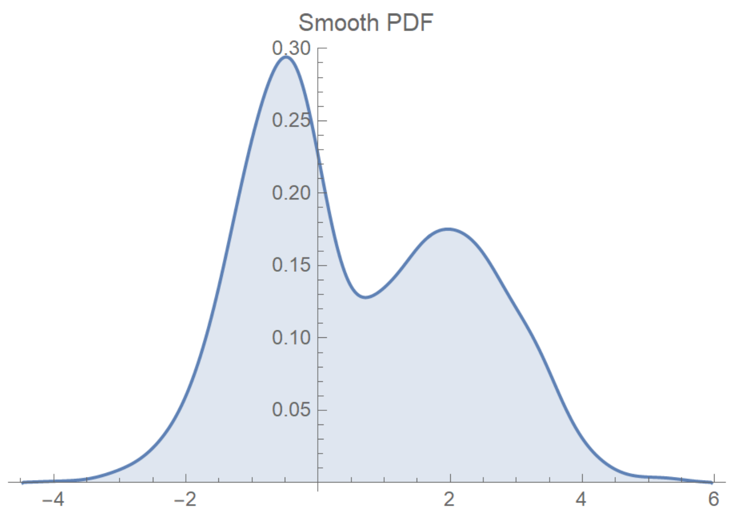


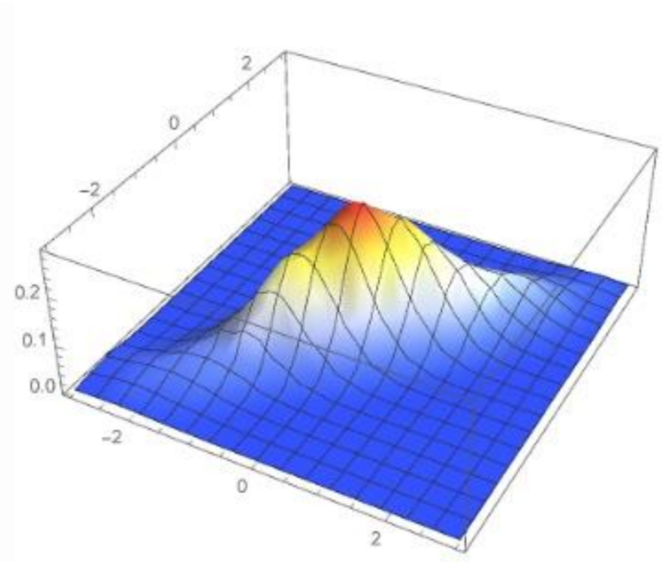


https://minghe4419.github.io/covariate_shift.github.io/separate_density_estimation.html

Kernel Density Estimation (KDE)

A non-parametric probability density estimation.





library(kdensity)

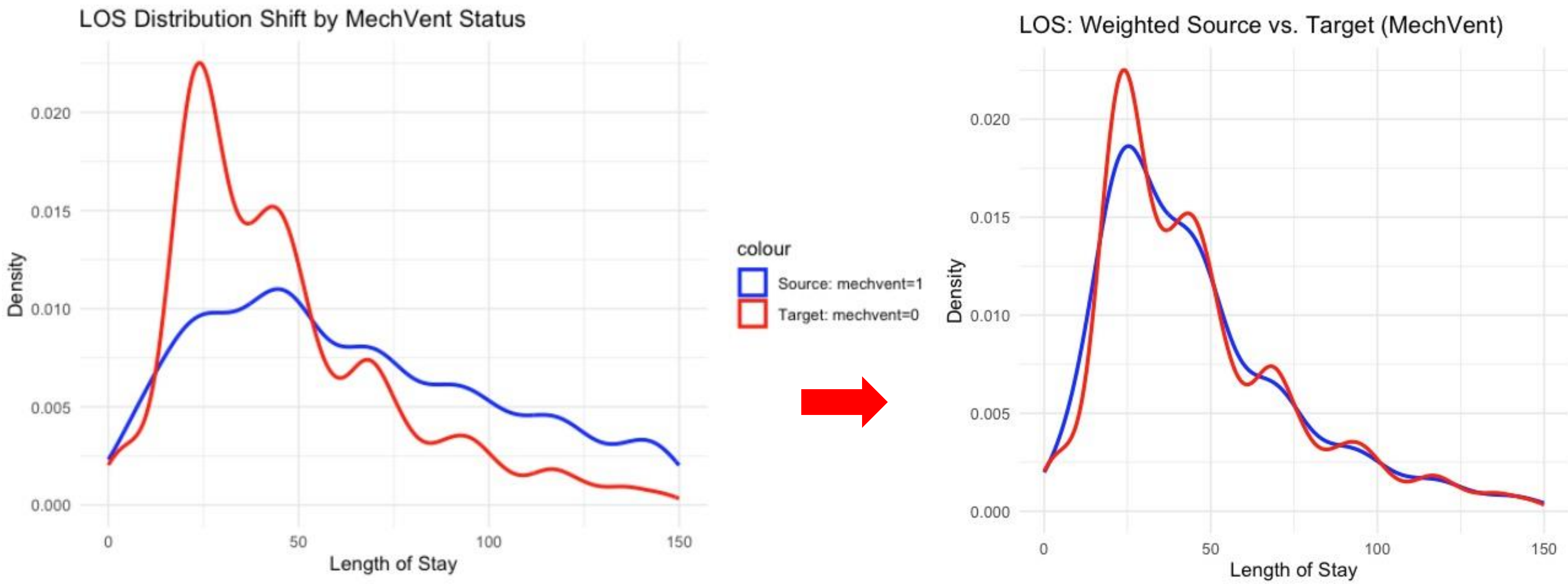
library(ks)

https://minghe4419.github.io/covariate_shift.github.io/separate_density_estimation.html

Separate KDE

Univariate Density Estimation
'library(kdensity)'

Source population: patients experienced Mechanical Ventilation
Target population: patients did NOT experience Mechanical Ventilation



https://minghe4419.github.io/covariate_shift.github.io/separate_density_estimation.html

Separate KDE

$$\omega(\boldsymbol{X}) = \frac{P^{\mathcal{T}}(\boldsymbol{X})}{P^{\mathcal{S}}(\boldsymbol{X})} = \frac{f_T(\boldsymbol{X})}{f_S(\boldsymbol{X})}$$

Univariate Density Estimation
'library(kdensity)'

Source population: patients experienced Mechanical Ventilation
Target population: patients did NOT experience Mechanical Ventilation

```
kde_source_hd <- kde(x = as.matrix(source_hd))
kde_target_hd <- kde(x = as.matrix(target_hd))

# 3. Evaluate densities at the source data points
p_source_hd <- predict(kde_source_hd, x = as.matrix(source_hd))
p_target_hd <- predict(kde_target_hd, x = as.matrix(source_hd))

# 4. Compute the density ratio weights: w(x) = P_T(x) / (P_S(x) + epsilon)
epsilon_val <- 1e-10
weights_hd <- p_target_hd / (p_source_hd + epsilon_val)
```

Separate KDE

$$\omega(\boldsymbol{X}) = \frac{P^{\mathcal{T}}(\boldsymbol{X})}{P^{\mathcal{S}}(\boldsymbol{X})} = \frac{f_T(\boldsymbol{X})}{f_S(\boldsymbol{X})}$$

Univariate Density Estimation
'library(kdensity)'

Source population: patients experienced Mechanical Ventilation
Target population: patients did NOT experience Mechanical Ventilation

```
kde_source_hd <- kde(x = as.matrix(source_hd))
kde_target_hd <- kde(x = as.matrix(target_hd))

# 3. Evaluate densities at the source data points
p_source_hd <- predict(kde_source_hd, x = as.matrix(source_hd))
p_target_hd <- predict(kde_target_hd, x = as.matrix(source_hd))

# 4. Compute the density ratio weights: w(x) = P_T(x) / (P_S(x) + epsilon)
epsilon_val <- 1e-10
weights_hd <- p_target_hd / (p_source_hd + epsilon_val)
```

Separate KDE

$$\omega(\boldsymbol{X}) = \frac{P^{\mathcal{T}}(\boldsymbol{X})}{P^{\mathcal{S}}(\boldsymbol{X})} = \frac{f_T(\boldsymbol{X})}{f_S(\boldsymbol{X})}$$

Univariate Density Estimation
'library(kdensity)'

Source population: patients experienced Mechanical Ventilation
Target population: patients did NOT experience Mechanical Ventilation

```
kde_source_hd <- kde(x = as.matrix(source_hd))
kde_target_hd <- kde(x = as.matrix(target_hd))

# 3. Evaluate densities at the source data points
p_source_hd <- predict(kde_source_hd, x = as.matrix(source_hd))
p_target_hd <- predict(kde_target_hd, x = as.matrix(source_hd))

# 4. Compute the density ratio weights: w(x) = P_T(x) / (P_S(x) + epsilon)
epsilon_val <- 1e-10
weights_hd <- p_target_hd / (p_source_hd + epsilon_val)
```

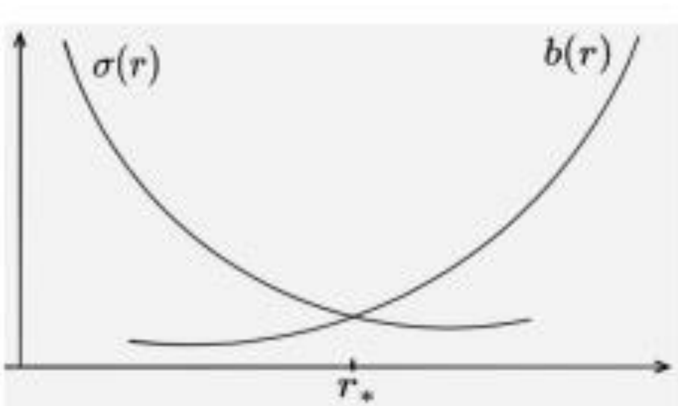
Separate Density Estimation

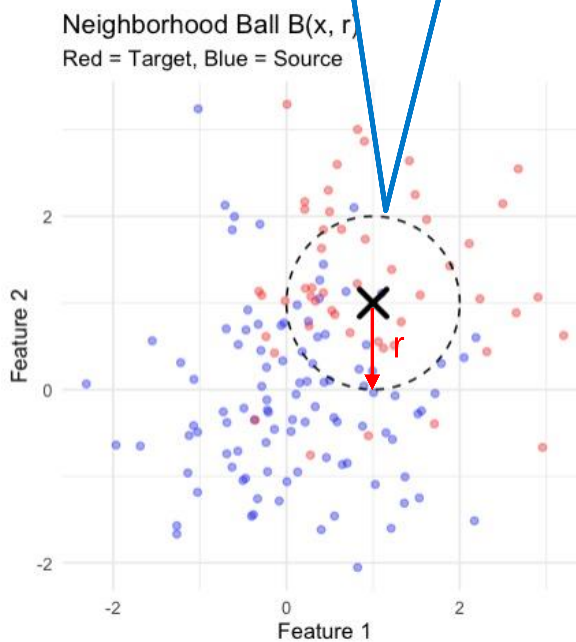
$$\omega(x) = \frac{number\ of\ target\ point\ in\ circle}{number\ of\ source\ point\ in\ circle}$$

- Histogram-based Method

$$\widetilde{\omega}(x) = \frac{P_T(B(x,r))}{P_S(B(x,r))} = \frac{n_T^{-1}\sum_{i=1}^{n_T} I(\|x - x_{T,i}\| \le r)}{n_S^{-1}\sum_{i=1}^{n_S} I(\|x - x_{S,i}\|_2 \le r)}$$

$$\widehat{\omega}(x) = \widetilde{\omega}(x) I(P_T(B(x,r)) \ge \alpha)$$

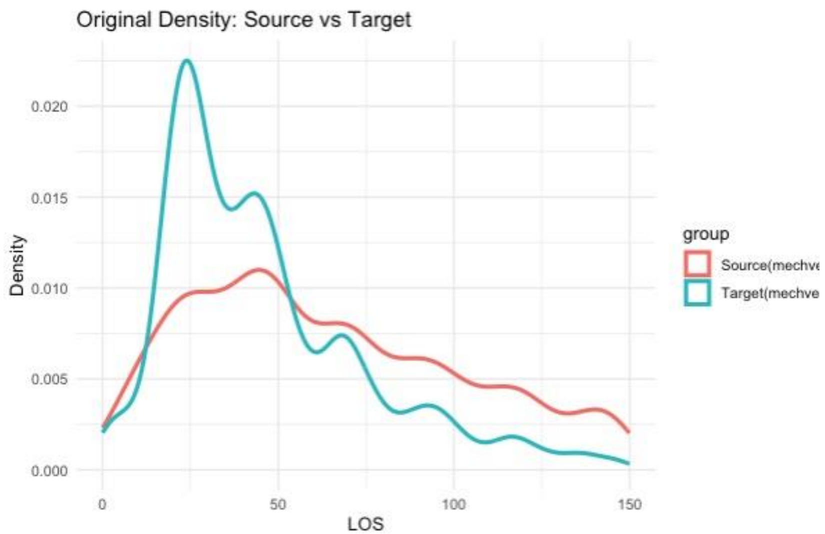


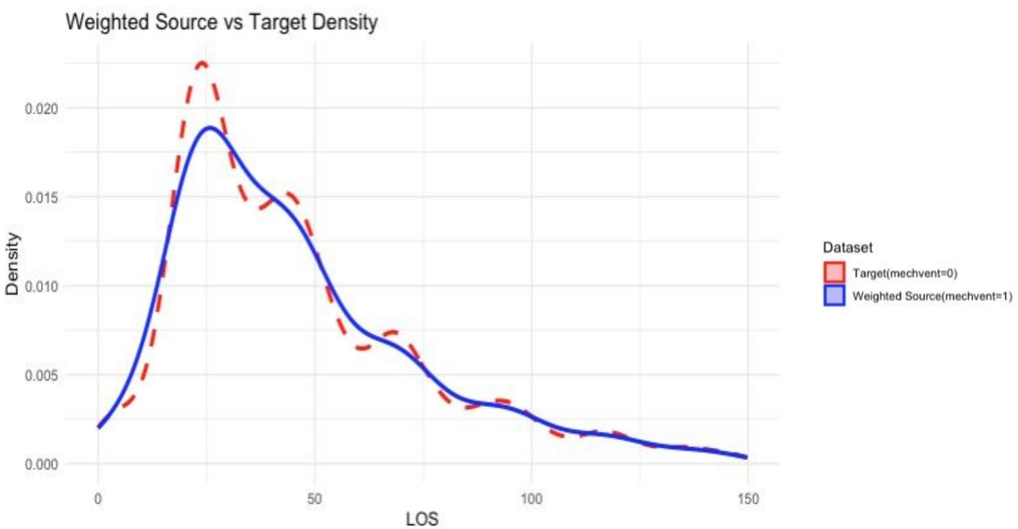


Separate Density Estimation

- Histogram-based Method

$$\widetilde{\omega}(x) = \frac{P_T(B(x,r))}{P_S(B(x,r))} = \frac{n_T^{-1}\sum_{i=1}^{n_T} I(\left\|x - x_{T,i}\right\| \le r)}{n_S^{-1}\sum_{i=1}^{n_S} I(\left\|x - x_{S,i}\right\|_2 \le r)} \qquad \widehat{\omega}(x) = \widetilde{\omega}(x) I(P_T(B(x,r)) \ge \alpha$$

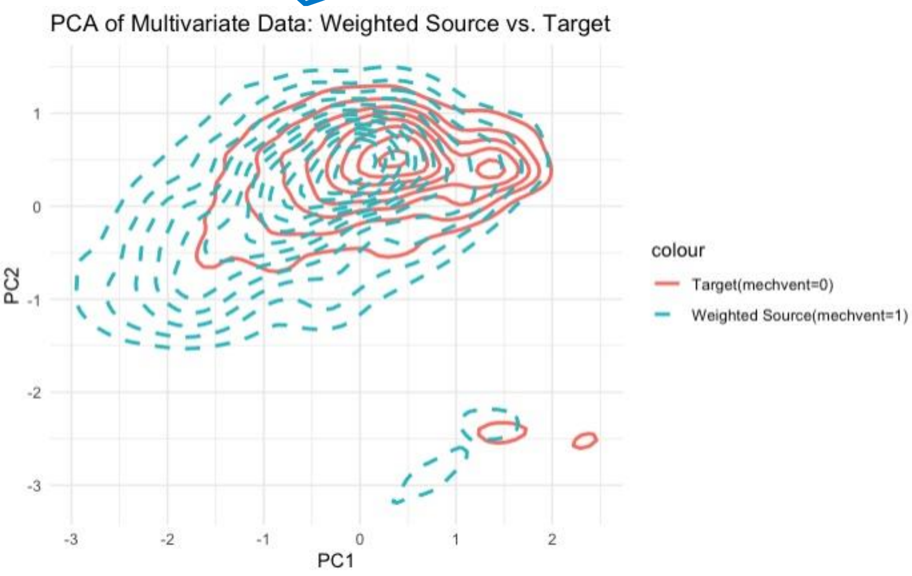




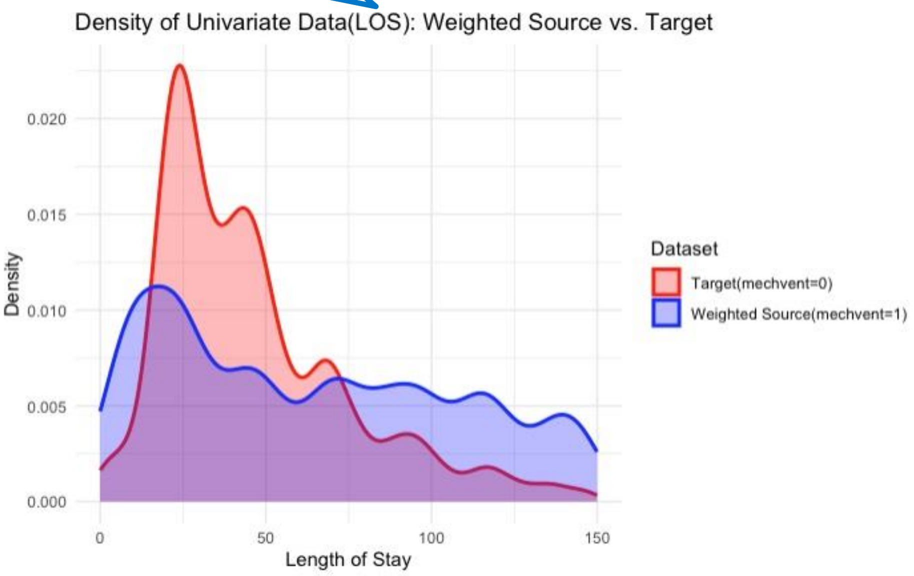
Separate Density Estimation

Multivariate Density Estimation
'library(ks)'

X = [age, length of stay, heart rate]



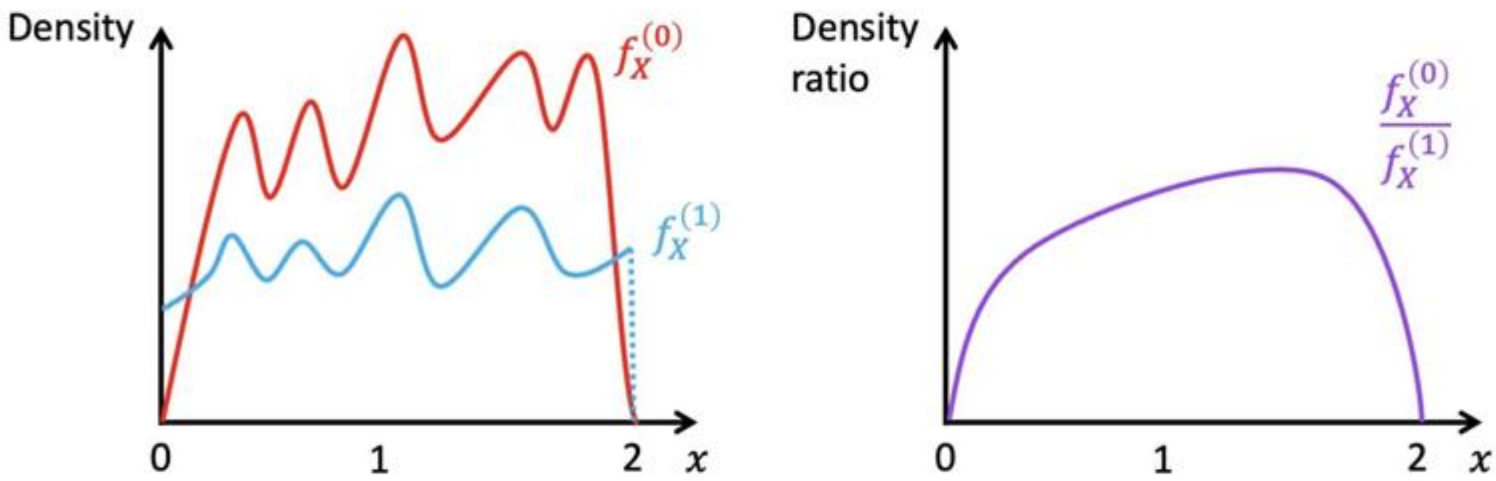
X = [age, length of stay, heart rate]



Separate Density Estimation

Issue: Worse smoothness in density ratio

- Densities are rough while the density ratio is smooth
- Resulting large variance in weights



Summary

- Transfer learning and domain adaptation
- Importance weighting framework
- Separate density estimation
  - KDE
  - Histogram-based
- Helpful sources
  - Our website about Covariate Shift: https://minghe4419.github.io/covariate_shift.github.io/index.html
    - Summary of methods to interact with more helpful resources
  - PhysioNet: https://physionet.org/
    - The Research Resource for Complex Physiologic Signals
    - Need to get credentialed and trained

- Separate Density Estimation
  - Kernel Density Estimation
  - Histogram-based Method

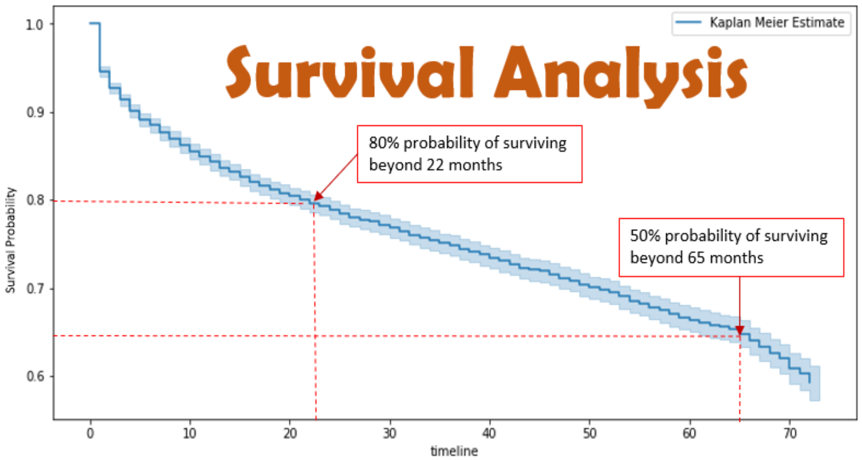
April 18th

- Estimating Weight as a whole
  - Kernel Mean Matching
  - Least Square Method
  - Kullback-Leibler Method
  - Discriminative Learning
  - Profile Likelihood Method
- Beyond Reweighting Technique and Covariate Shift
  - Marginal Transfer Learning
  - Domain Invariant Method
  - Optimal Transformation

Example: Survival Analysis & Cox Model

Survival Analysis: model the time until an event of interest occurs

- Medical Research
- Finance
- Social Sciences



https://medium.com/data-science/survival-analysis-intuition-implementation-in-python-504fde4fcf8e

Example: Survival Analysis & Cox Model

$$\underbrace{\mathrm{H}(t|\boldsymbol{X})}_{\text{Hazard function at time t}} = \underbrace{h_0(t)}_{\text{Baseline Hazard}} \; \underbrace{\exp\{\boldsymbol{X}^T \overset{\text{Target parameter}}{\boldsymbol{\beta}}\}}_{\text{Scaling Factor}}$$

01 **Proportional Hazard Assumption**: The model assumes that the hazard rate is proportional across different levels of covariates

02 **Semi-parametric Nature**: The Cox model is a semi-parametric model, allowing for flexible baseline hazard functions without imposing rigid assumptions

03 **Handling Time-Dependent Covariates**: Can handle time-dependent covariates, which means that the effect of a variable on the hazard rate can change over time

04 **Identify Key Factors**: A powerful tool for identifying factors that affect the hazard rate of an event

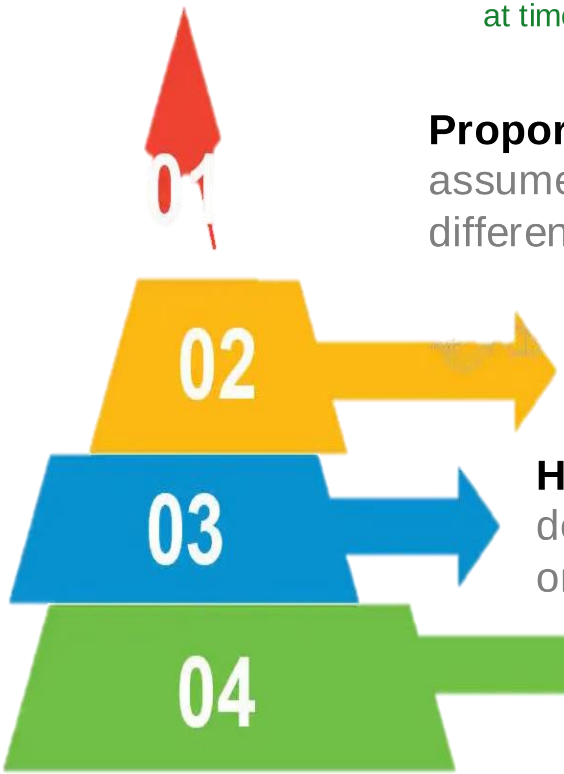
Example: Survival Analysis & Cox Model

Target parame

$$\underbrace{\mathrm{H}(t|\boldsymbol{X})}_{\text{Hazard function at time t}} = \underbrace{h_0(t)}_{\text{Baseline Hazard}} \ \underbrace{\exp\{\boldsymbol{X}^T\boldsymbol{\beta}\}}_{\text{Scaling Factor}}$$

Challenging in small-sample settings

- Rare diseases
- Subpopulations with limited representation
- Specialized clinical trials



Proportional Hazard Assumption: T model assumes that the hazard rate is pr tional across different levels of covariates

Semi-parametric Natu parametric model, allow functions without imposi

Handling Time-Depende dependent covariates, whi on the hazard rate can cha

Identify K identifying of an even

All of Us RESEARCH PROGRAM

Event of interest: end-stage-renal disease (ESRD) among high-risk patients (impaired kidney function)

→ Over ~127,000 participants
→ 203 high-risk Hispanic
→ 33 developed ESRD

https://fastercapital.com/content/Cox-proportional-hazards- tml

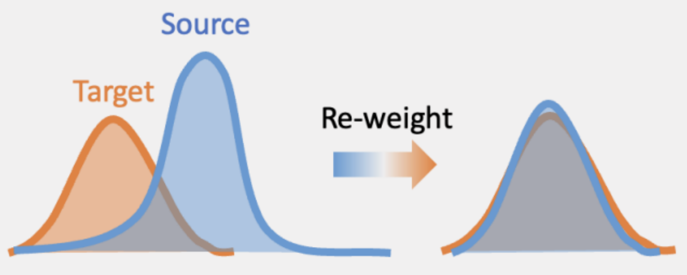
Transfer Learning in Survival Analysis
What are the gaps

- **Covariate (X) Shift:** Systematic differences between source & target datasets, which will introduce biases, reducing model robustness, especially in the Cox model when model misspecification exists
  - Cox model is used as a “working model” to analyze the relationship between covariates and the hazard rate, even when its assumptions—such as proportional hazards—do not strictly hold
- **Limited Adaptation:** Existing methods do not fully address cross-population heterogeneity, including baseline hazard variation and coefficient difference

🔍 **Our contribution**: A robust transfer learning framework that accounts for multi-level data shifts in Cox model

Proposed 2-step CoxTL

Step 1: Density Ratio Estimation $\hat{\omega}$



Step 2: Calibrate Log-likelihood for Joint Estimation

$$\boldsymbol{l}^{TL}(\boldsymbol{\beta}) = \underbrace{\boldsymbol{l}^{T}(\boldsymbol{\beta}; X_{T}, Y_{T})}_{\text{Target}} + \underbrace{\boldsymbol{\nu}}_{\text{Prevent Negative Transfer}} \hat{\boldsymbol{\omega}}\, \underbrace{\boldsymbol{l}^{S}(\boldsymbol{\beta}; X_{S}, Y_{S})}_{\text{Calibrated Source}}$$

Obtain $\boldsymbol{\beta}^{TL}$ by maximizing $\boldsymbol{l}^{TL}(\boldsymbol{\beta})$

Adjust the difference in covariate X

- Adjust the difference in parameter $\boldsymbol{\beta}$
- Prevent negative effects from heterogeneous sources