

When Training and Test Sets are Different: Characterising Learning Transfer

Amos J Storkey

Presenter: Desh Raj

Outline

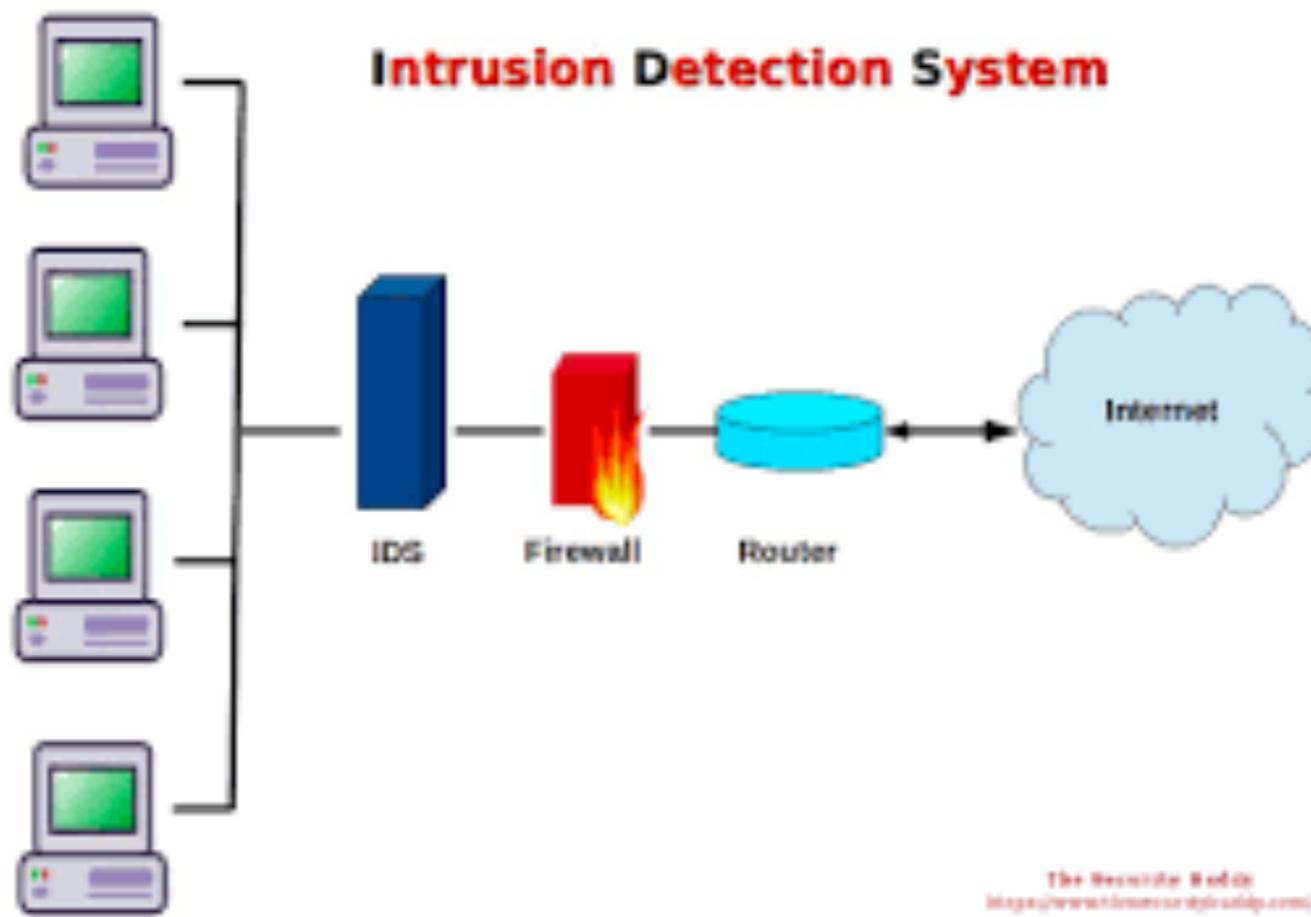
- Why is this important?
- Preliminaries - conditional and generative models
- Categories of dataset shift
- Practical considerations

Why is this important?



Use recognition algorithm developed for one device on another device.

Why is this important?



Will an old Network intrusion detection software still work?

Why is this important?



Spam filter trained on different users' data

Why is this important?

- Environments are non-stationary.
- Most predictive ML models work by ignoring the different training and test scenarios.
- How to address this?

But first, a small point

- Are dataset shift and transfer learning the same thing?
- **No!**
- *Transfer learning is more general*; transfer information from a variety of environments to help with learning and inference in a new environment.
- *Dataset shift is more specific*; make prediction in one environment given data in another closely related environment.

Preliminary: Conditional and Generative models

- Generative model: joint probability distribution over all variables of interest.

$$P(\mathbf{y}, \mathbf{x})$$

- Conditional model: distribution of some smaller set of variables is given for each possible known value of the other variables.

$$P(\mathbf{y}|\mathbf{x}, \Theta_y)$$

- Conditional models offer more flexibility in choosing model parameterizations.
- Fit of $P(\mathbf{y}|\mathbf{x})$ is never compromised in favor of a better fit for the unused model $P(\mathbf{x})$.
- If generative model accurately specifies a known generative process, then the choice of modeling structure may fit real constraints better than a conditional model.

- $P(\mathbf{y}, \mathbf{x}) = P(\mathbf{y}|\mathbf{x}) P(\mathbf{x})$
- “Causal graphical model”: structure is more than just a representation of a factorization
- May happen that factorization holds even if model is not causal.
- Dataset shift still changes $P(\mathbf{y}|\mathbf{x})$ and $P(\mathbf{x})$ by intervening on some possibly latent variable.
- So even non-causal conditional models are affected by dataset shift.

Types of Dataset Shift

Dataset
Shift

Simple Covariate Shift

Prior Probability Shift

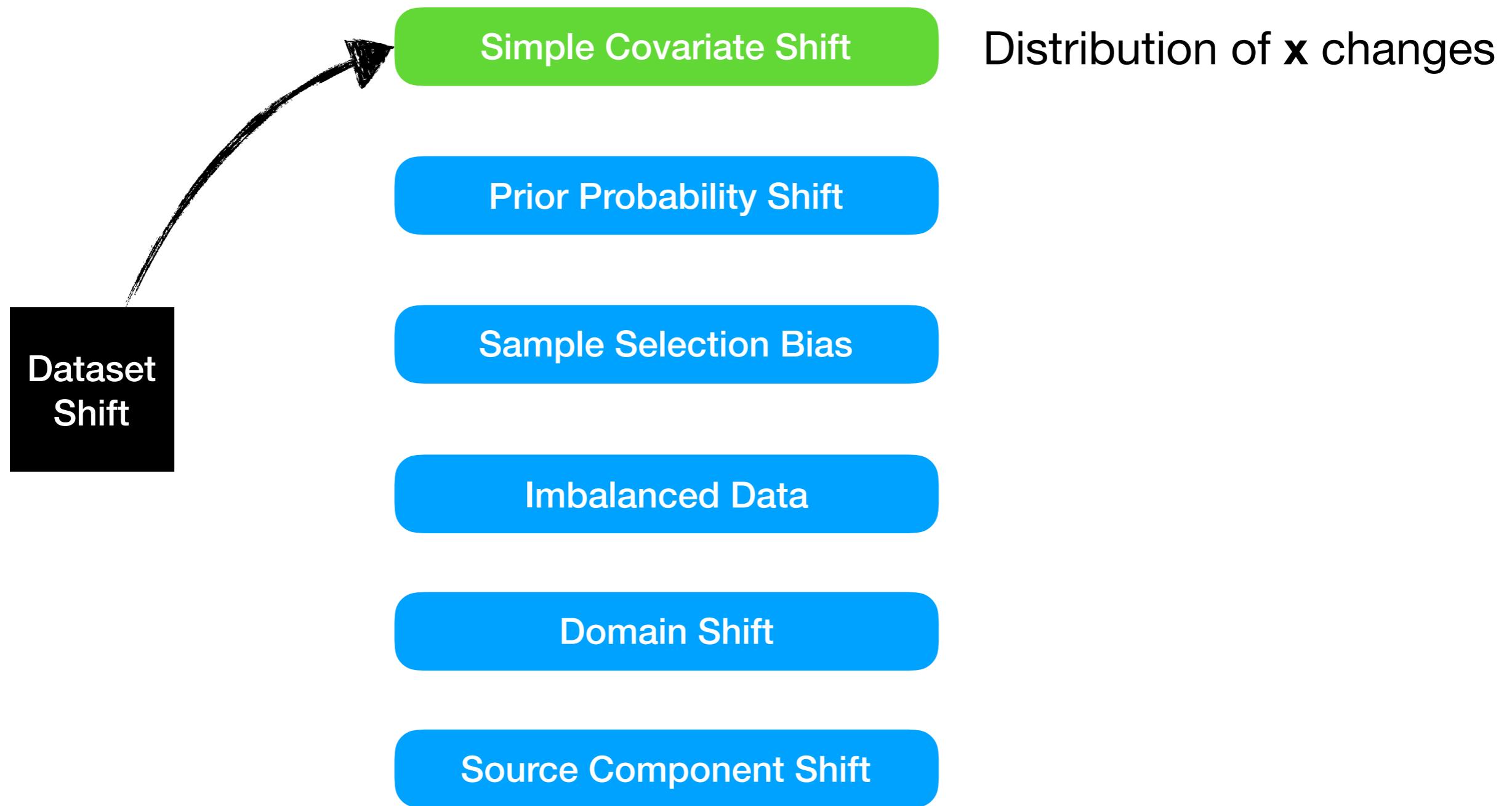
Sample Selection Bias

Imbalanced Data

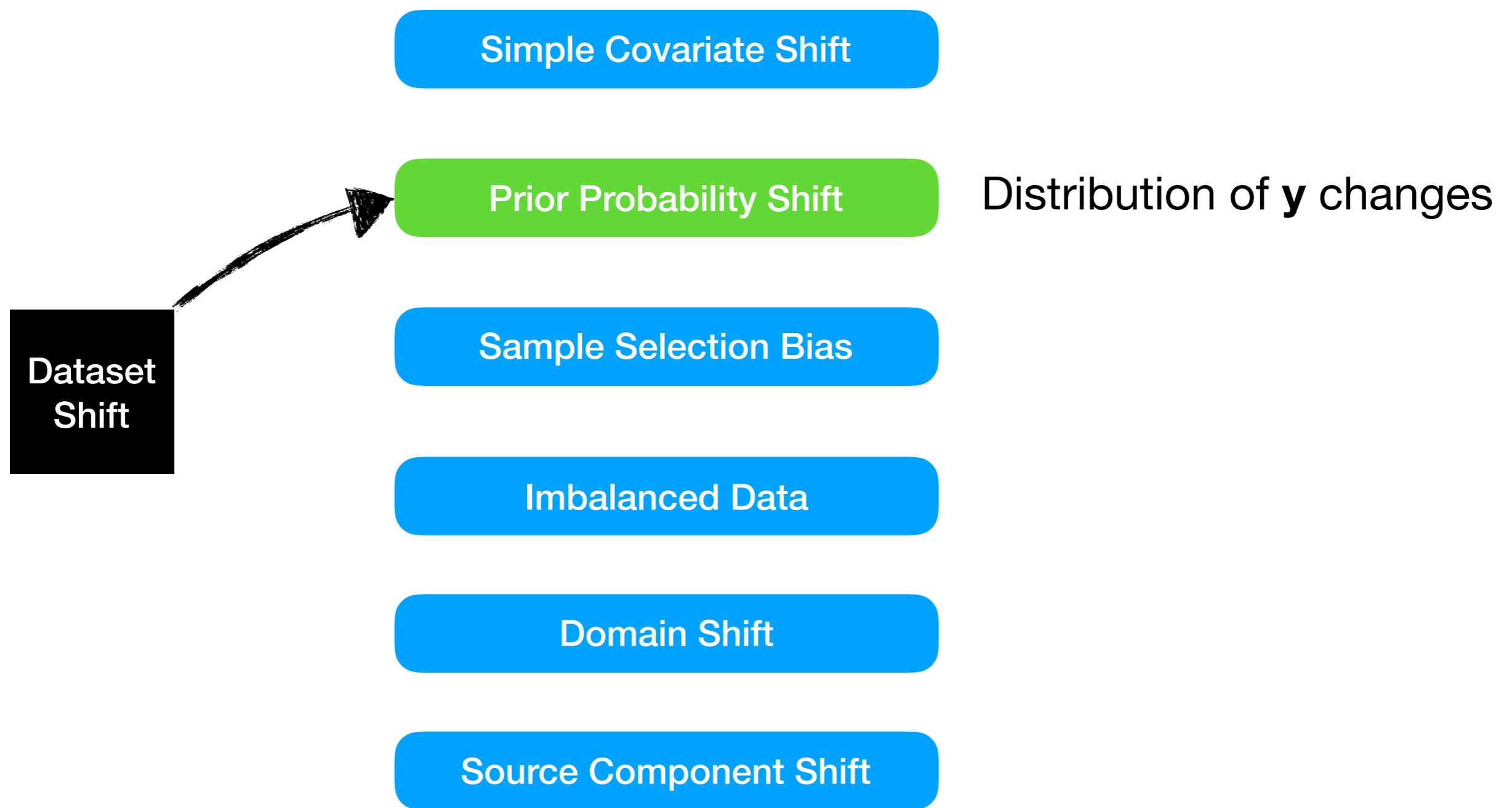
Domain Shift

Source Component Shift

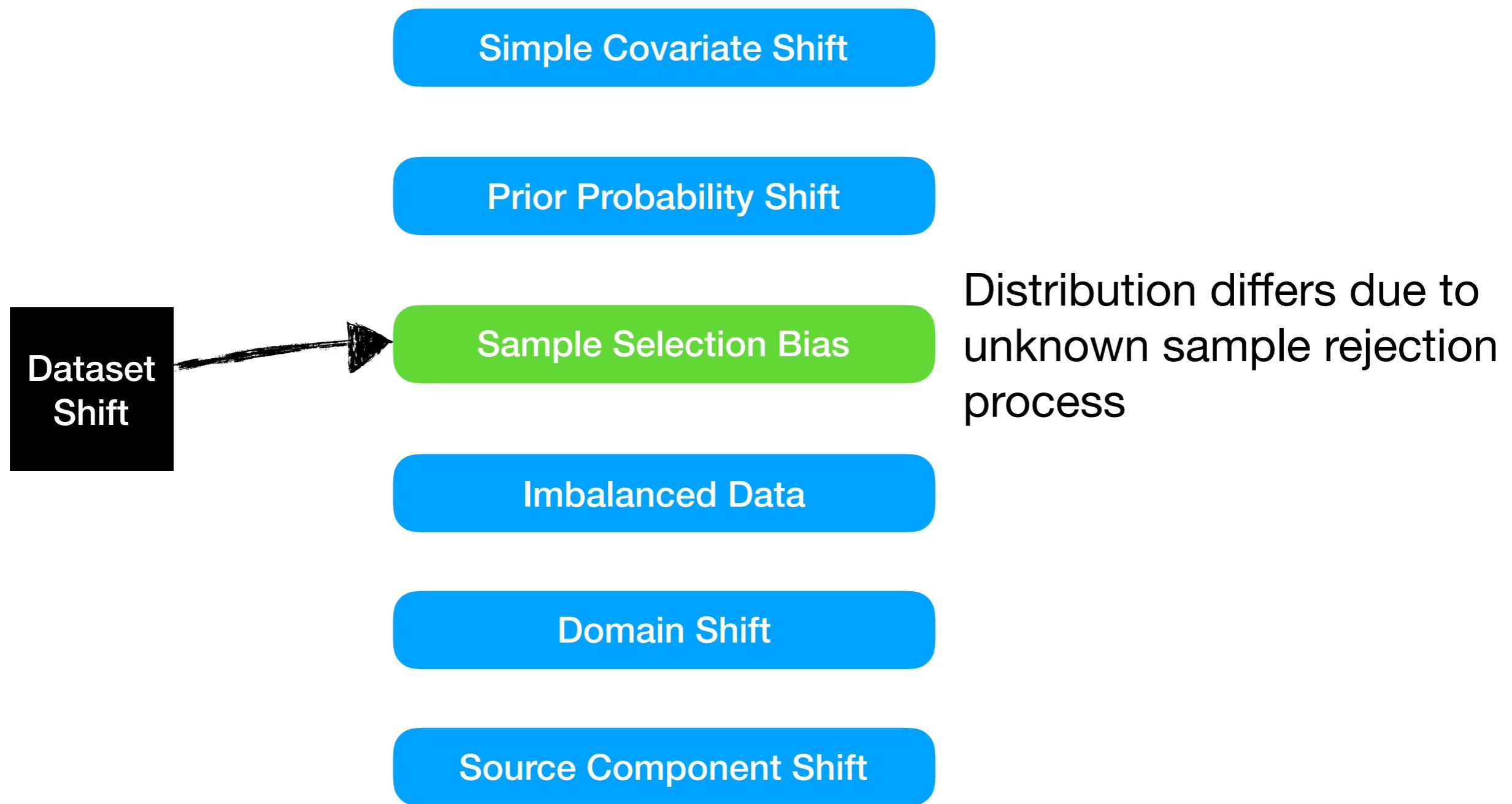
Types of Dataset Shift



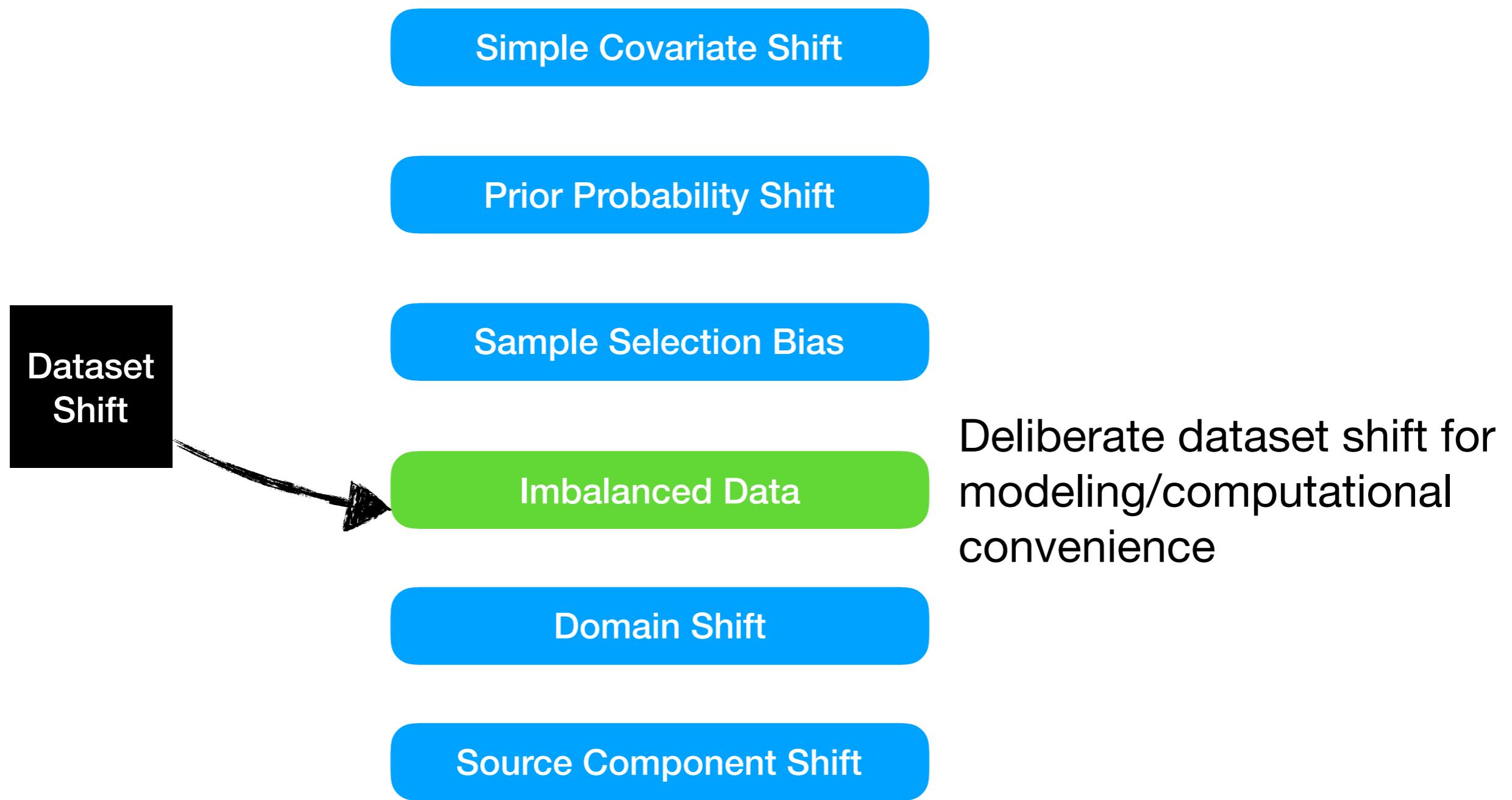
Types of Dataset Shift



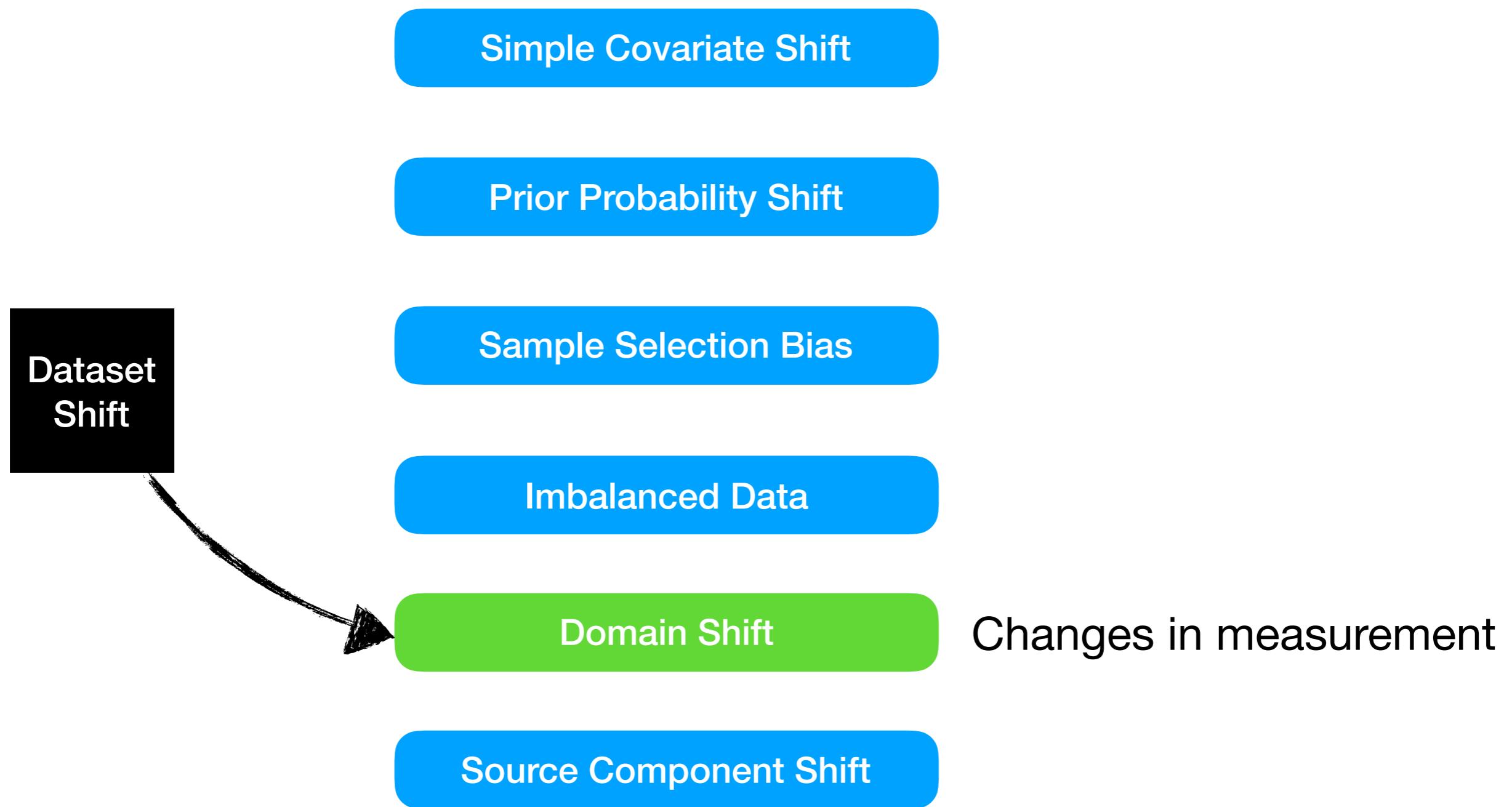
Types of Dataset Shift



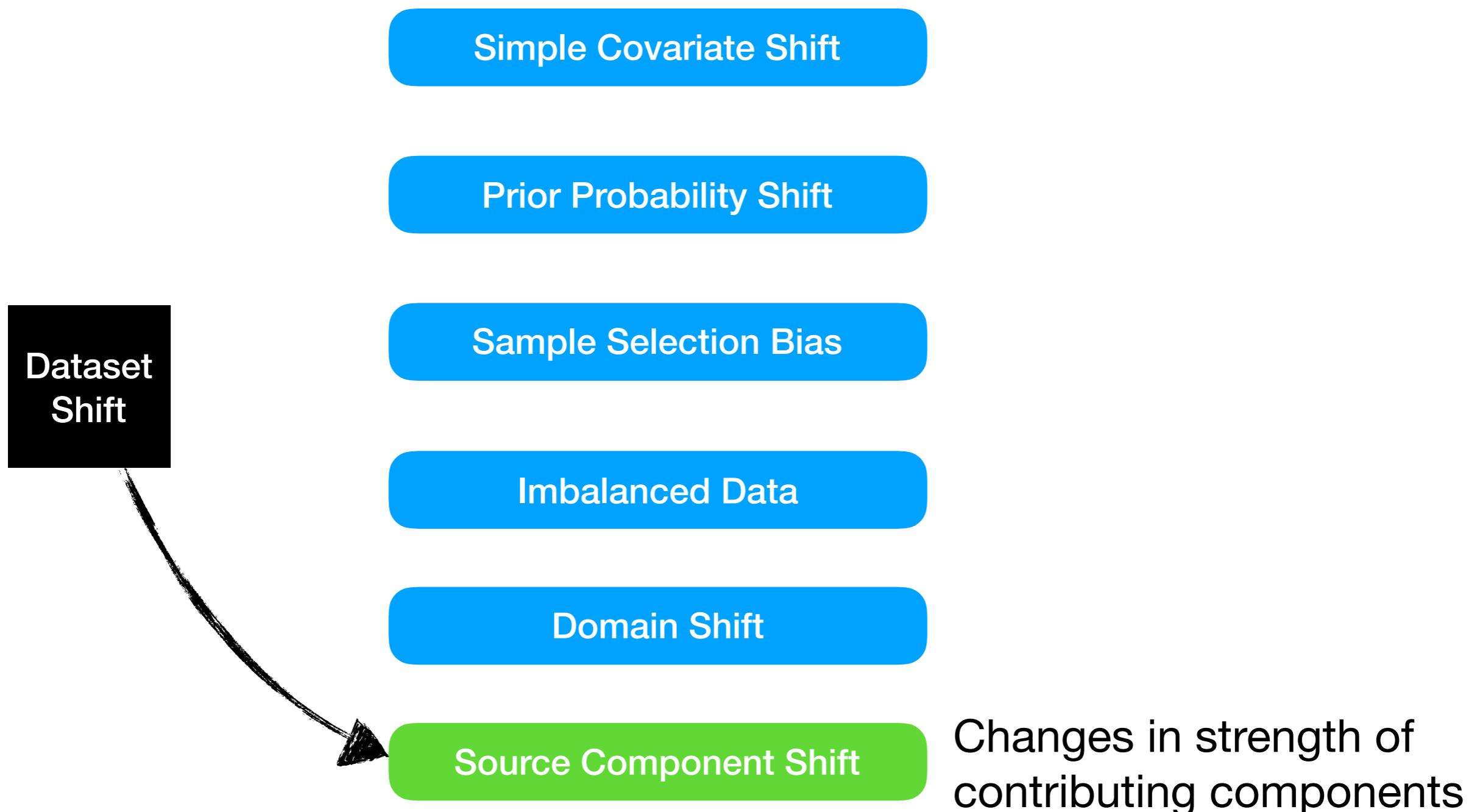
Types of Dataset Shift



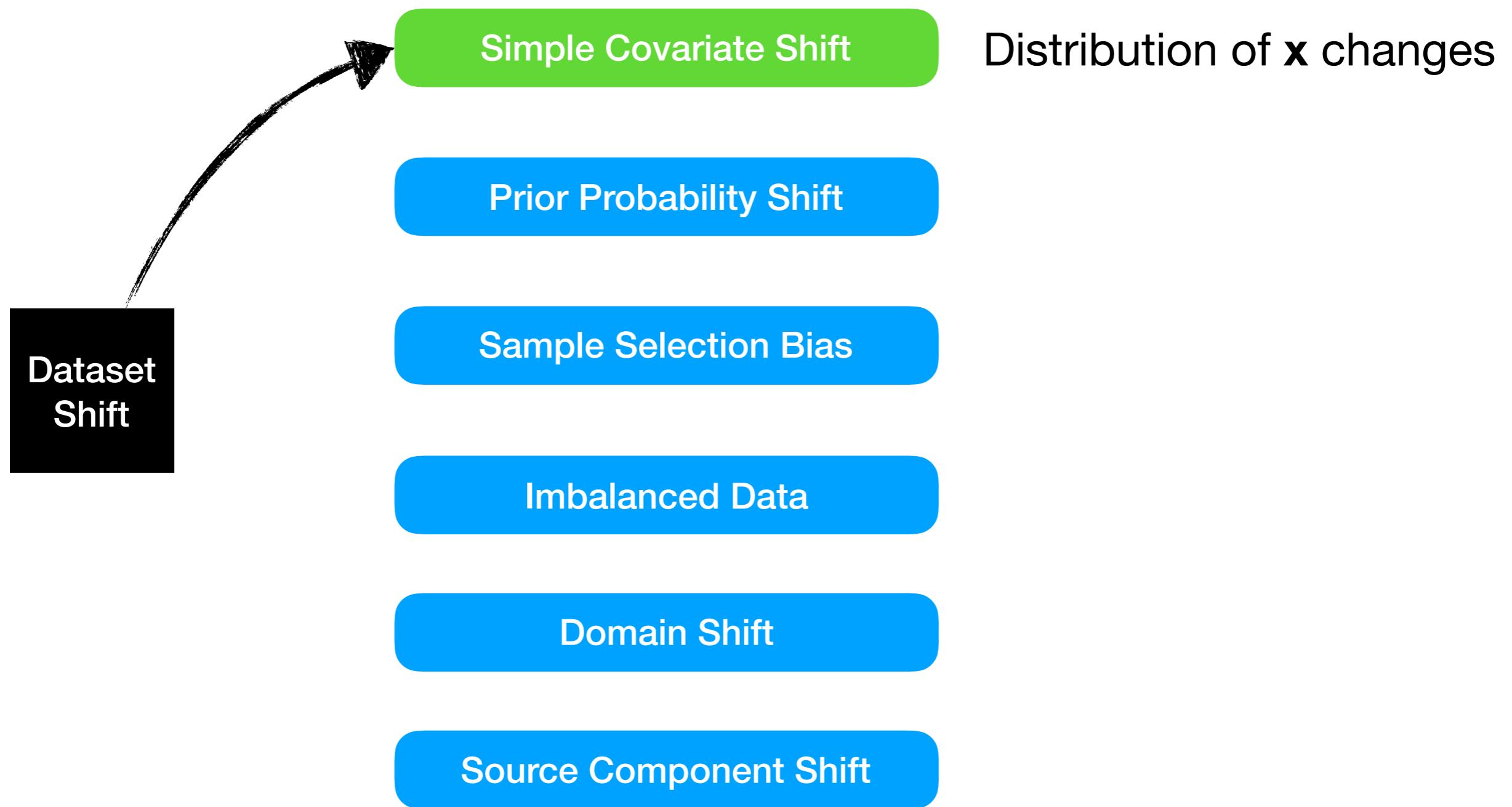
Types of Dataset Shift



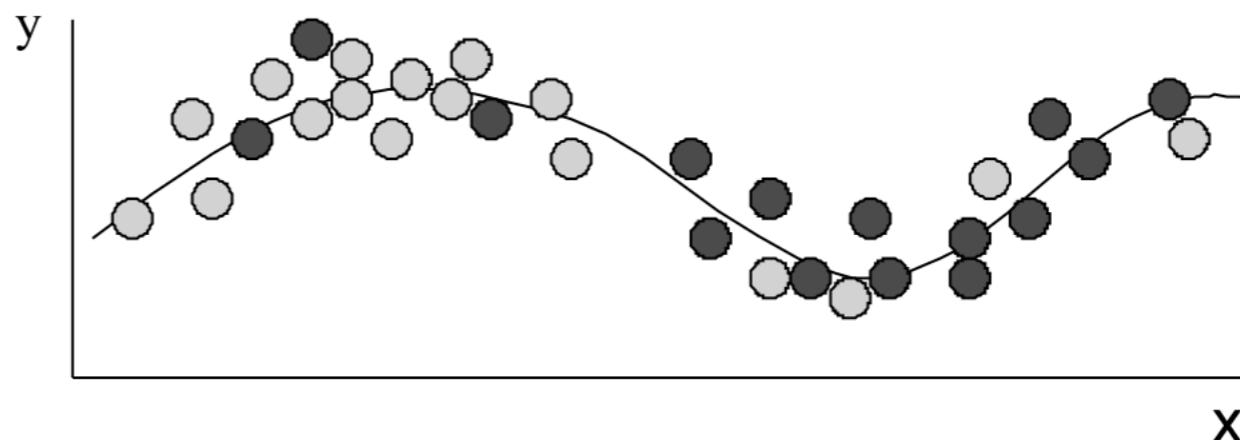
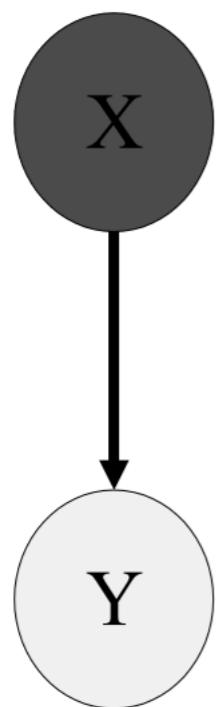
Types of Dataset Shift



Types of Dataset Shift



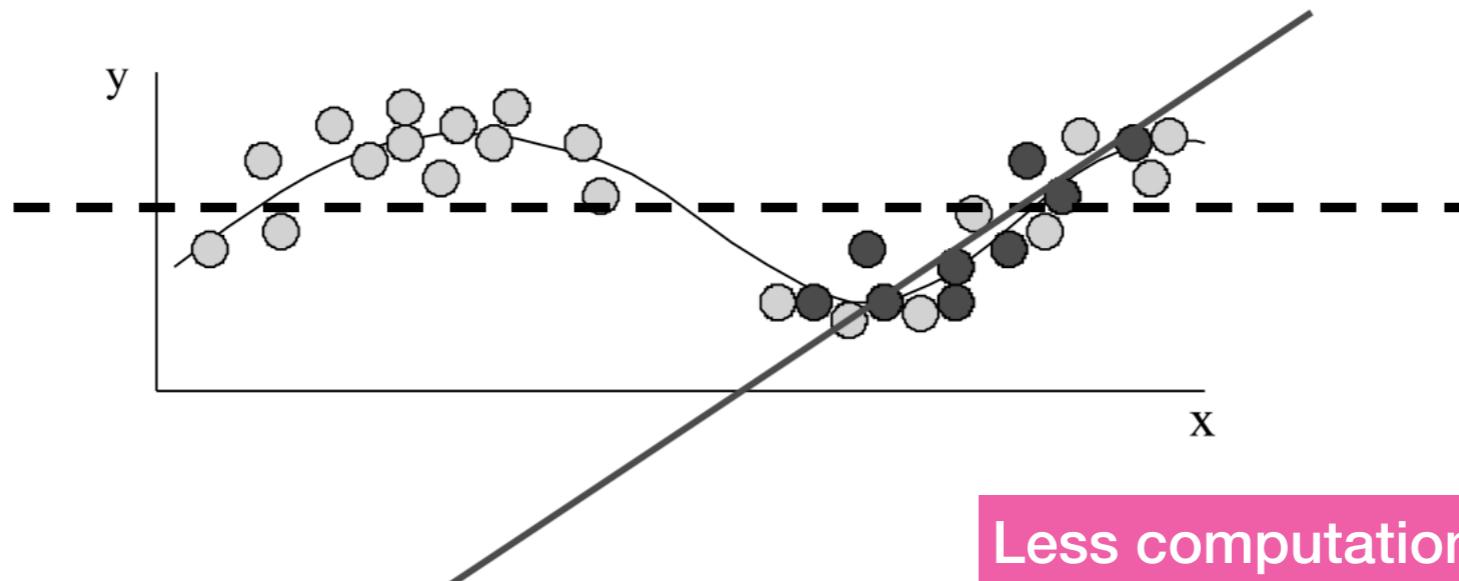
Simple Covariate Shift



Simple Covariate Shift

- y depends on that particular x , so change in distribution should not affect a prediction.
- But still some work claiming $P(y|x)$ needs to be changes to deal with covariate shift. Why?
- If the class of $P(y|x)$ does not contain true conditional model, then it is not actually true distribution.

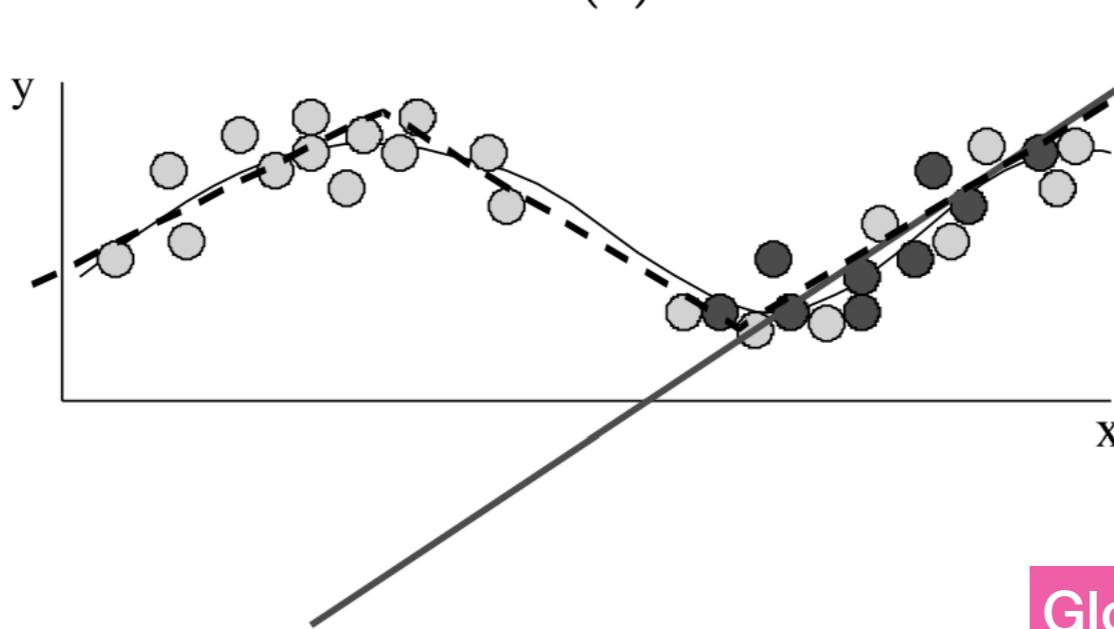
**Linear with
importance
weighting**



Less computation but not global

(a)

Local linear model



Global but wasteful

(b)

Figure 2: Covariate shift for mis-specified models: (a) The linear model is a poor fit to the global data (dashed line). However by focussing on the local region associated with the test data distribution the fit (full line) is much better as a local linear model is more appropriate. (b) The global fit for a local linear model is more reasonable, but involves the computation of many parameters that are never used in the prediction.

Simple Covariate Shift

- Might suggest that only model known form of model for test region and ignore the others.
- But what if test region is also contaminated with several other source.
- More in **source component shift**.

Two examples

- **Gaussian process model** is a conditional model.

$$\begin{aligned} P(\{y_i\} | \{\mathbf{x}_i\}, \{\mathbf{x}^k, y^k\}) &= \int dy^* P(\{y_i\}, y^* | \{\mathbf{x}_i\}, \mathbf{x}^*, \{\mathbf{x}^k, y^k\}) \\ &= P(\{y_i\} | \{\mathbf{x}_i\}, \mathbf{x}^*, \{\mathbf{x}^k, y^k\}) \end{aligned}$$

- Introducing new covariate point \mathbf{x}^* has no predictive effect. (Kolmogorov consistency)

Two examples

- **Support Vector Machine (SVM)** is NOT a conditional model.

$$\exp\left(-\sum_{i=1}^N(1-y_i(\mathbf{w}^T \cdot \mathbf{x}_i)_+)\right) \exp\left(-\frac{1}{2C}|\mathbf{w}|^2\right).$$

Two examples

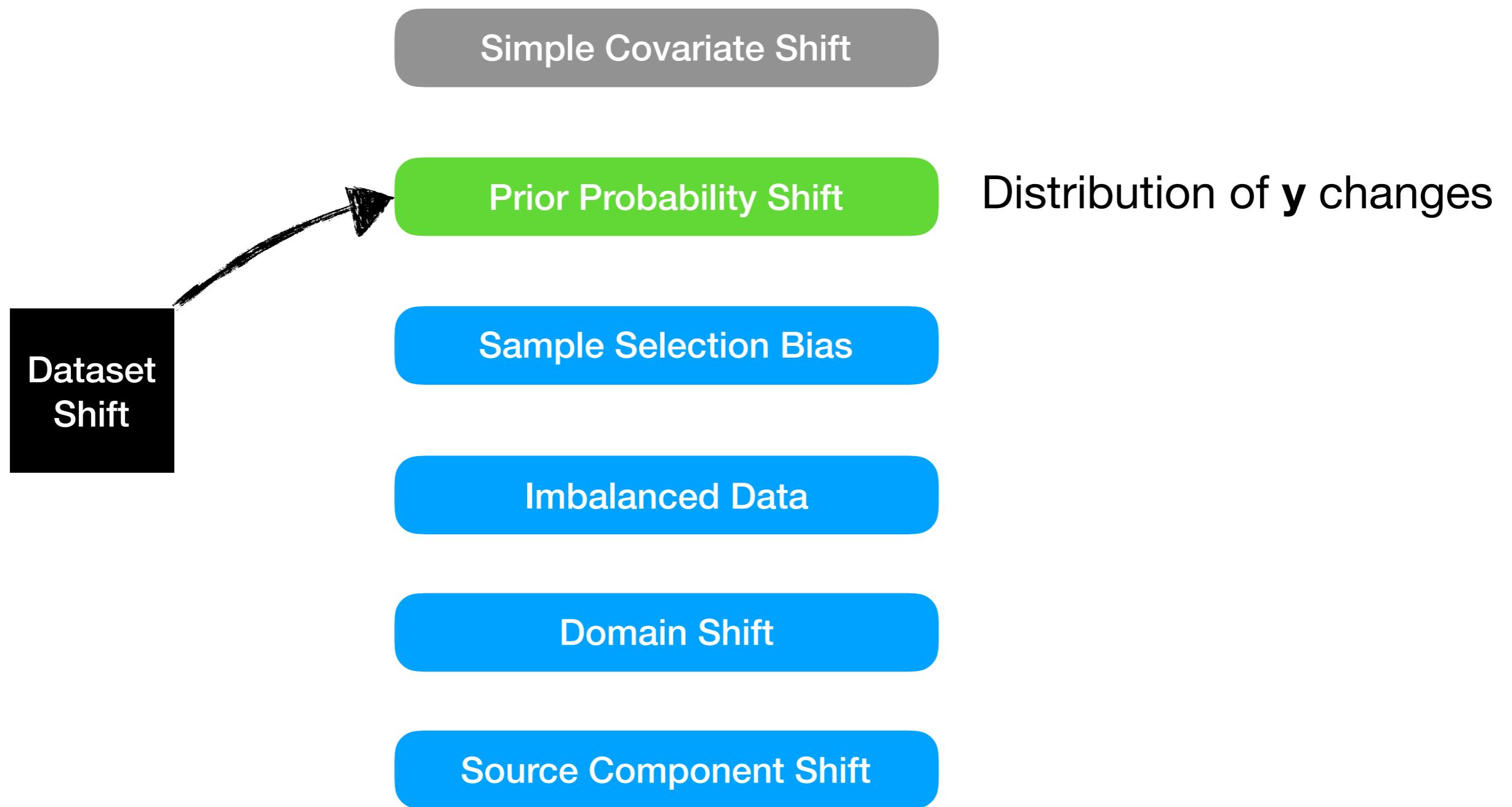
- **Support Vector Machine (SVM)** is NOT a conditional model.

$$\exp \left(- \sum_{i=1}^N (1 - y_i (\mathbf{w}^T \cdot \mathbf{x}_i)_+) \right) \exp \left(- \frac{1}{2C} |\mathbf{w}|^2 \right).$$

↓

$$\left[\prod_{i=1}^N P(y_i | \mathbf{w}) \right] Z_N(\mathbf{w}) \exp \left(- \frac{1}{2C} |\mathbf{w}|^2 \right).$$

Types of Dataset Shift



Prior Probability Shift

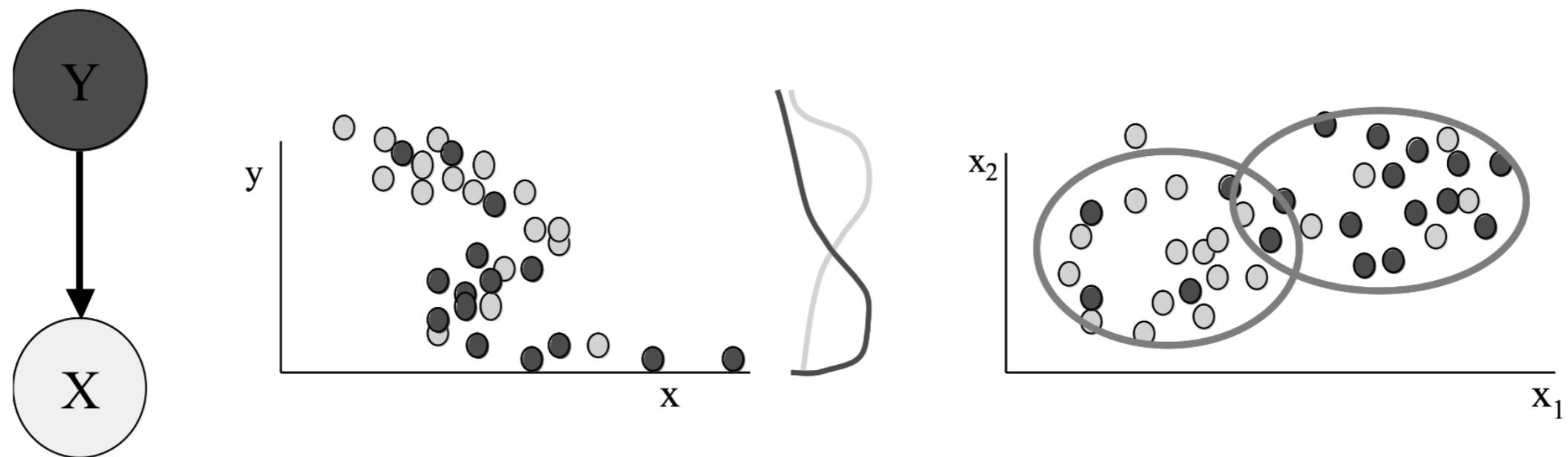
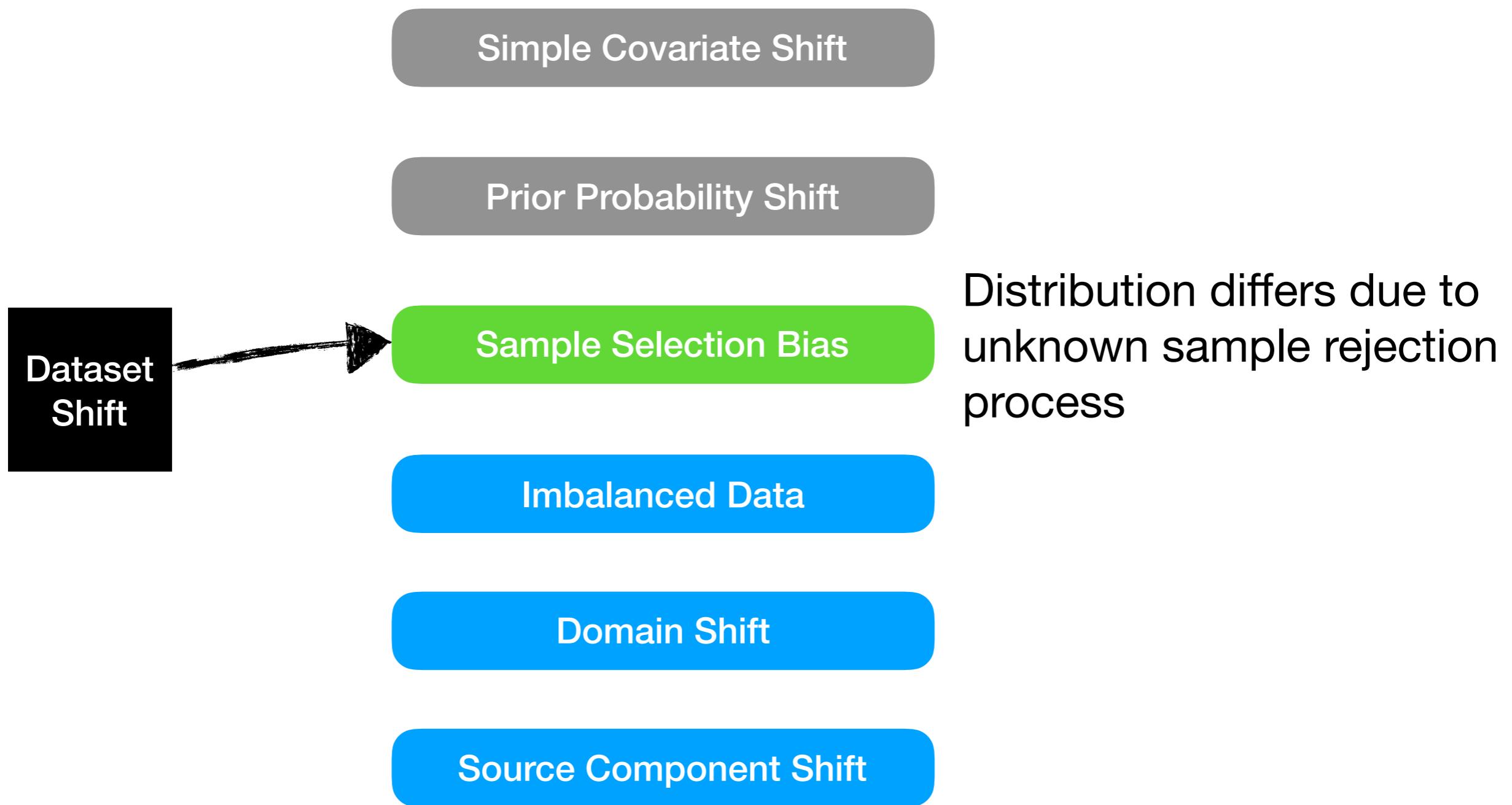


Figure 3: Prior Probability Shift. Here the causal model indicated the covariates \mathbf{x} are directly dependent on the predictors \mathbf{y} . The distribution over \mathbf{y} can change, and this effects the predictions in both the continuous case (left) and the class conditional case (right).

Prior Probability Shift

- If $P_{te}(\mathbf{y})$ is known, it is easy to correct for the new joint probability.
- What to do if it is not known?
- Given $P(\mathbf{x}|\mathbf{y})$ and the covariates of test data, certain distributions of \mathbf{y} are more or less likely.
- Specify a prior over valid $P(\mathbf{y})$ and compute posterior based on test covariates.

Types of Dataset Shift



Sample Selection Bias

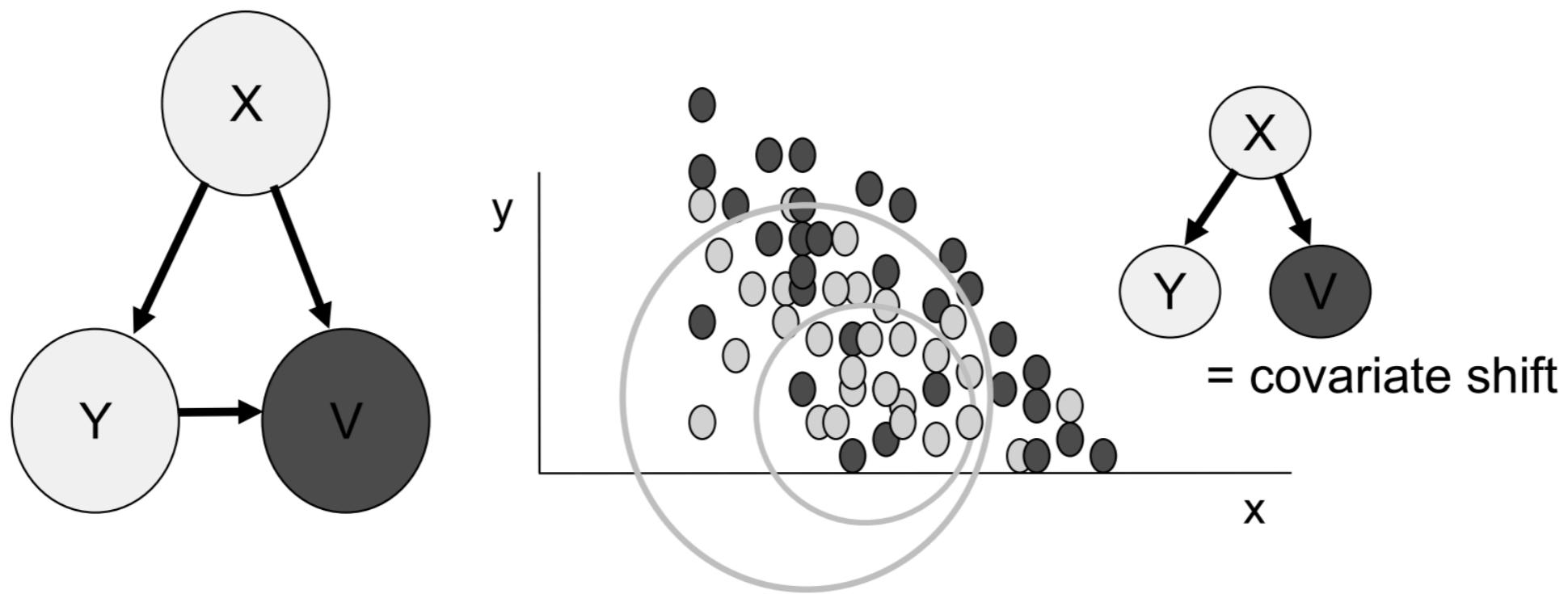
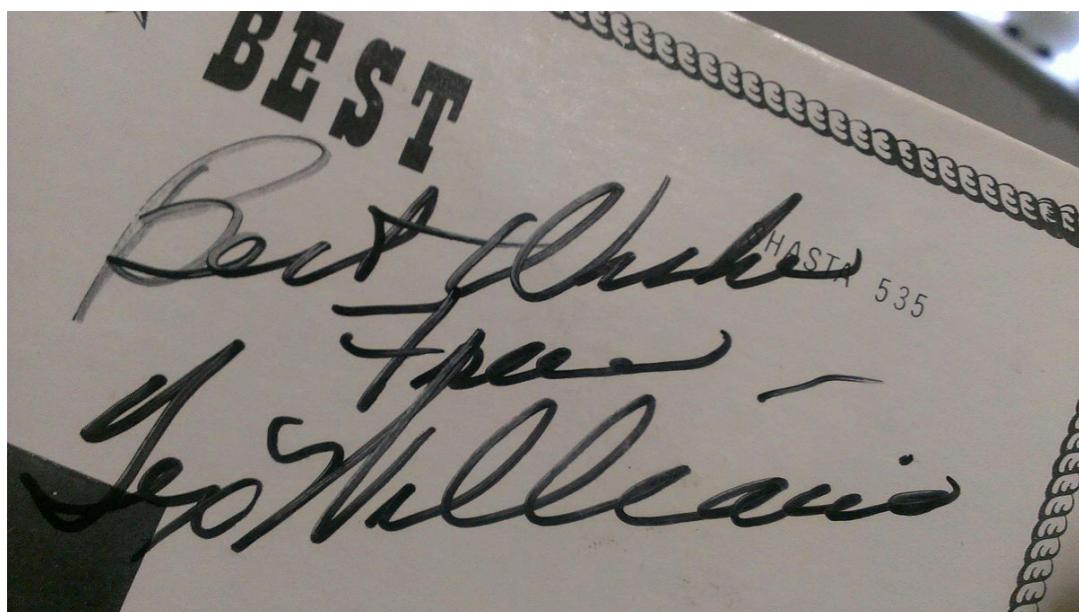


Figure 4: Sample Selection Bias. The actual observed training data is different from the test data because some of the data is more likely to be excluded from the sample. Here v denotes the selection variable, and an example selection function is given by the equiprobable contours. The dependence on y is crucial as without it there is no bias and this becomes a case of simple covariate shift.



Surveys: Some people may be less willing to participate in, say, election polls.



Handwriting recognition: Some unintelligible (but important) characters may be discarded.

Sample Selection Bias

- “Regression to the mean”
- Example:
 - Measure rate of illness X in several districts.
 - Choose those with highest rate to try new drug.
 - May improve just because of random fluctuations; incorrectly attributed to the new drug.

Some approaches

- Training:

$$P_{\text{tr}}(\mathbf{y}, \mathbf{x}) = P(\mathbf{y}, \mathbf{x}, v = 1) = P(v = 1 | \mathbf{y}, \mathbf{x}) P(\mathbf{y} | \mathbf{x}) P(\mathbf{x})$$

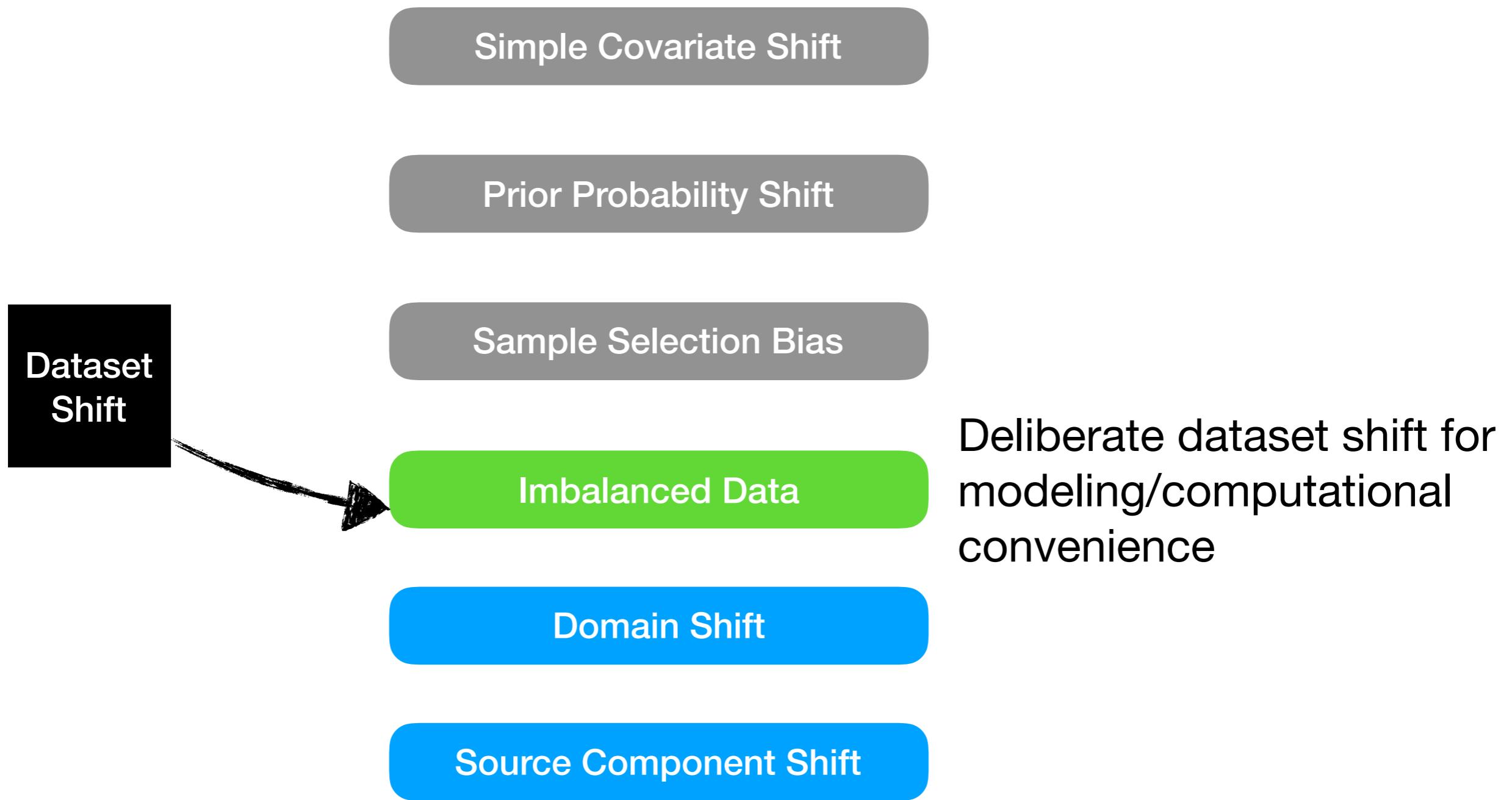
- Test:

$$P_{\text{te}}(\mathbf{y}, \mathbf{x}) = P(\mathbf{y}, \mathbf{x}) = P(\mathbf{y} | \mathbf{x}) P(\mathbf{x}).$$

$$P(\mathbf{y} | \mathbf{x}) = P(\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{f}(\mathbf{x})) \text{ and}$$

$$P(v = 1 | \mathbf{y}, \mathbf{x}) = P(\nu > g(\mathbf{x}) | \mathbf{y} - f(\mathbf{x})) = P(\nu > g(\mathbf{x}) | \boldsymbol{\epsilon})$$

Types of Dataset Shift



Imbalanced Data

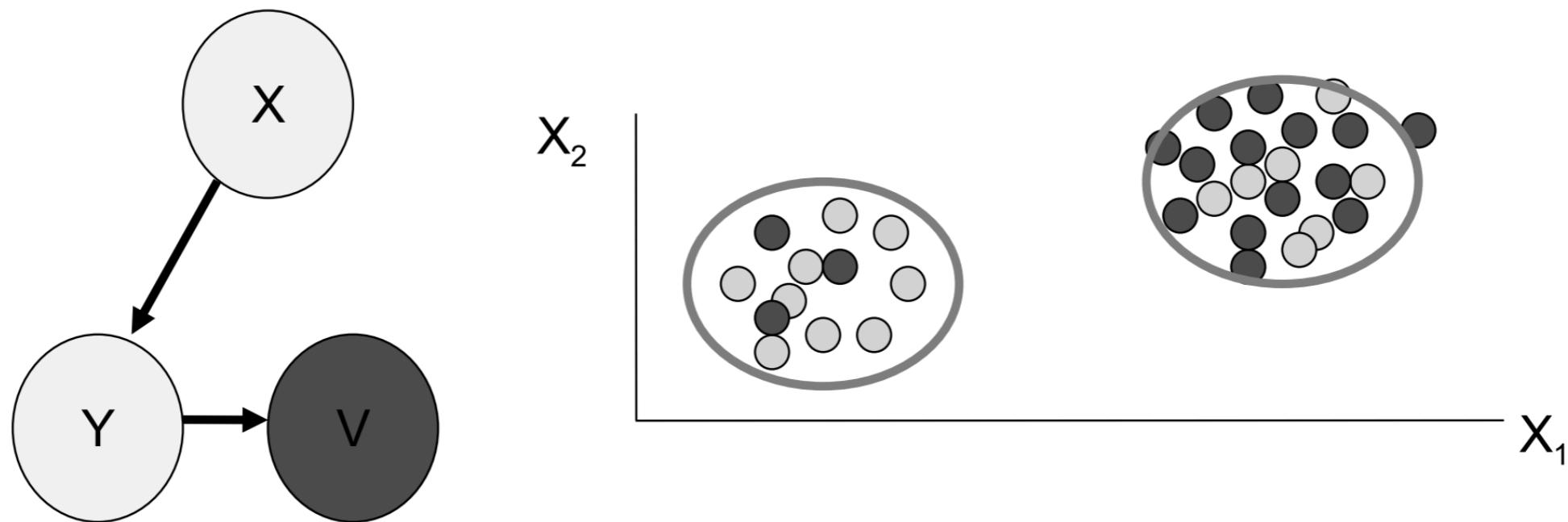


Figure 5: Imbalanced Data: imbalanced data is sample selection bias with a designed known bias that is dependent on only the class label. Data from more common classes is more likely to be rejected in the training set in order to balance out the number of cases of each class.

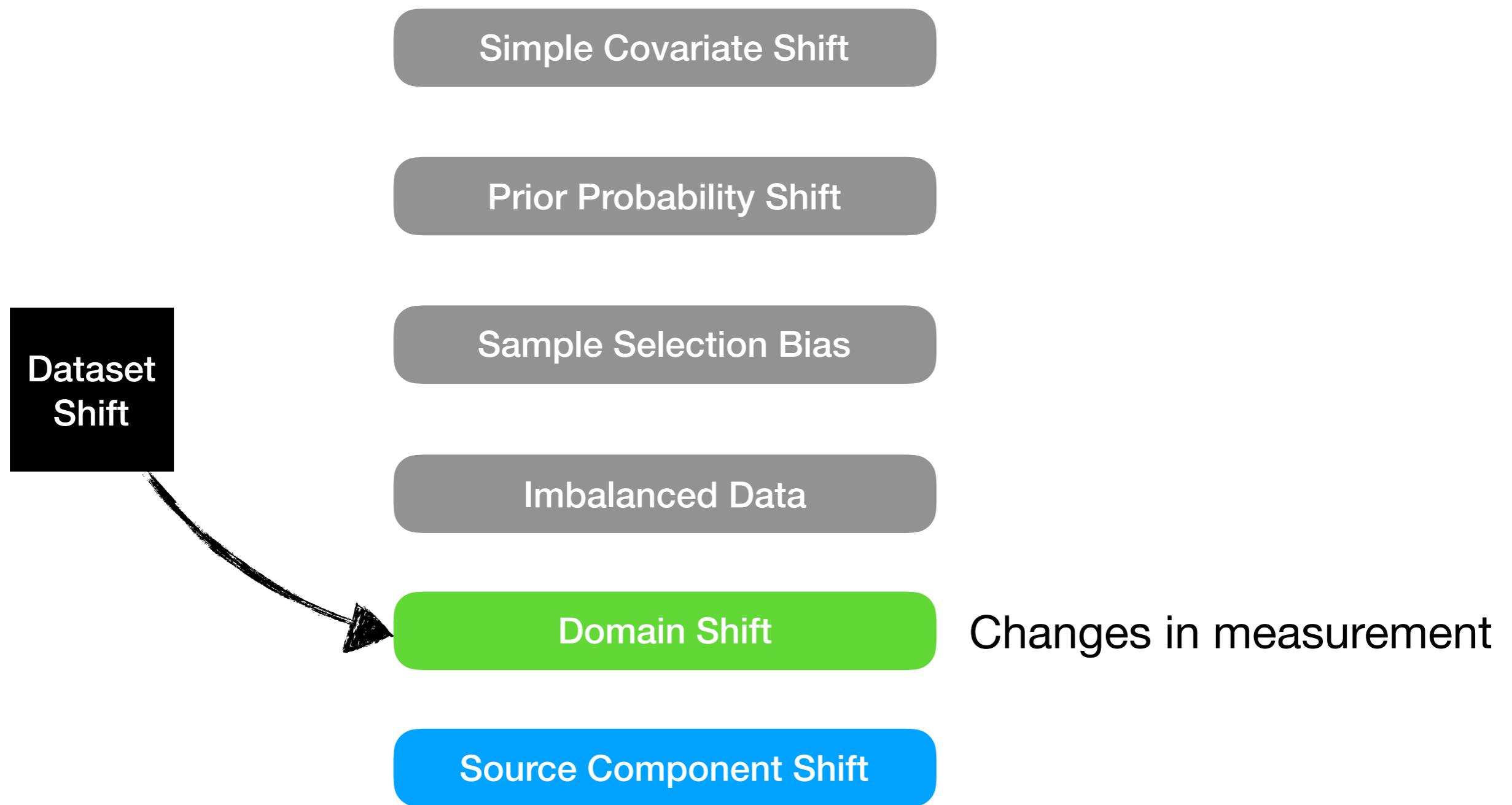
Imbalanced Data

- **Example:** prediction of rare events like *loan defaulting*.
- Throw away common class to make the data balanced.
This is okay since common class is already easy to characterize (large amount of data).
- But this creates **imbalanced training and test distributions!**

Imbalanced Data

- **Generative models** may be better at handling this issue.
- $P(\mathbf{y}, \mathbf{x}) = P(\mathbf{x}|\mathbf{y}) P(\mathbf{y})$
- How? The problem is just **change in $P(\mathbf{y})$ with a known shift.**
- Imbalance is decoupled from modeling.

Types of Dataset Shift



Domain Shift

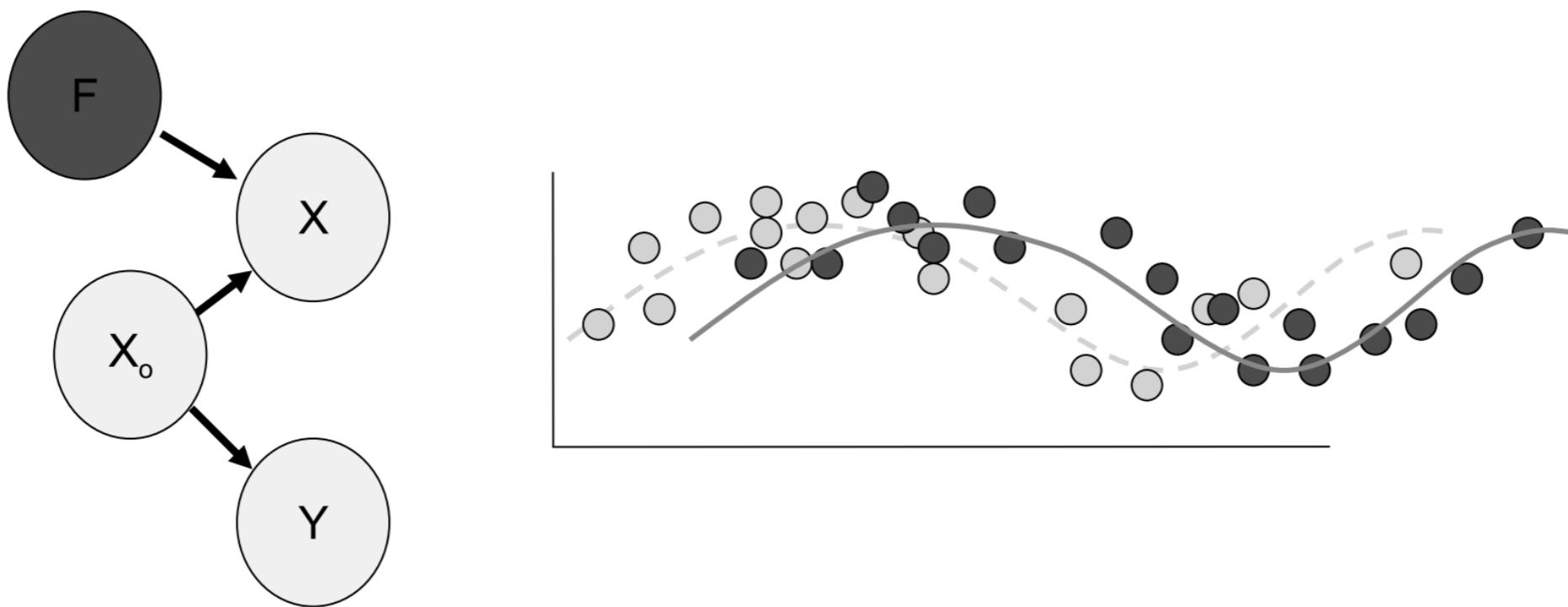
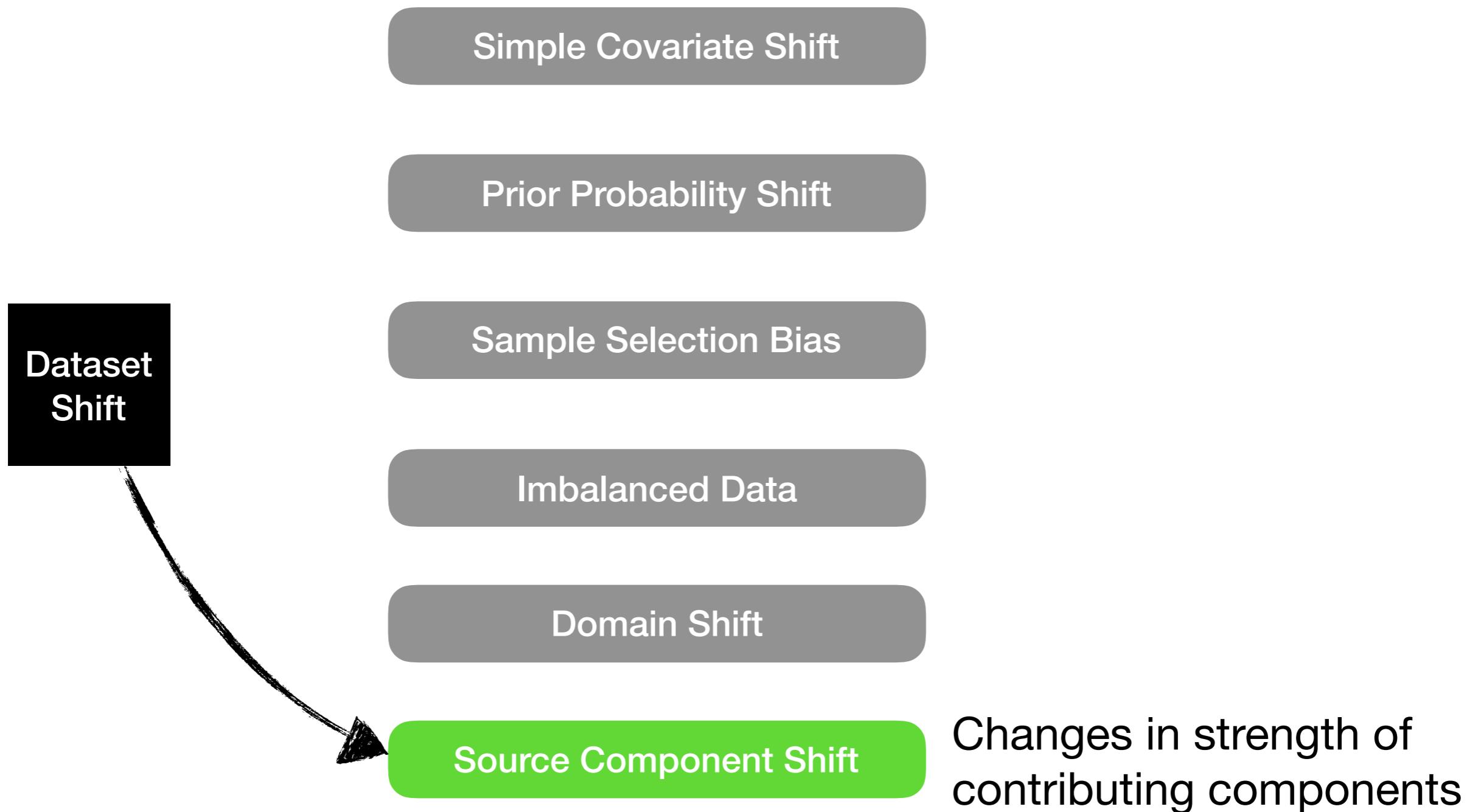


Figure 6: Domain Shift: The observed covariates \mathbf{x} are transformed from some idealised covariates \mathbf{x}_0 via some transformation F , which is allowed to vary between datasets. The target distribution $P(\mathbf{y}|\mathbf{x}_0)$ is unchanged between test and training datasets, but of course the distribution $P(\mathbf{y}|\mathbf{x}_0)$ does change if F changes.

Domain Shift

- **Example:** two photographs of the same scene from different cameras or in different lighting may look different.
- We never observe \mathbf{x}_0 . We only observe $\mathbf{x} = \mathbf{f}(\mathbf{x}_0)$.
- Modeling involves estimating \mathbf{f} using the distributional information. Example: gamma correction for photographs

Types of Dataset Shift



Source Component Shift

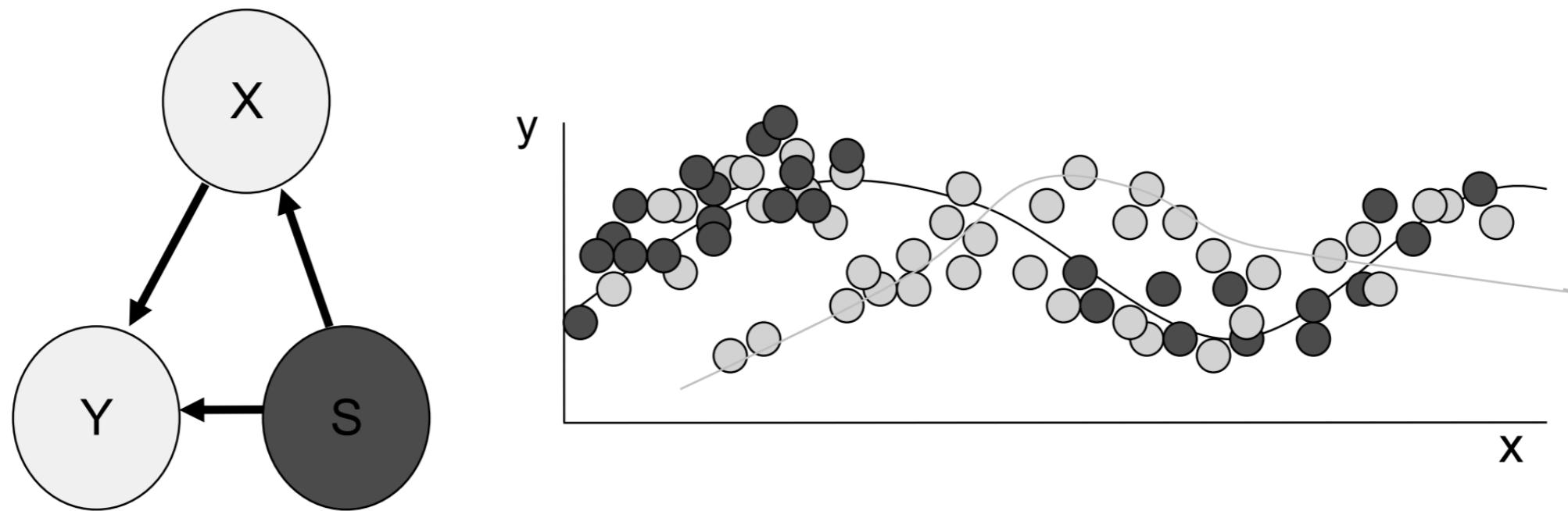


Figure 7: Source component shift. A number of different sources of data are represented in the dataset, each with its own characteristics. Here S denotes the source proportions and these can vary between test and training scenarios. In mixture component shift, these sources are mixed together in the observed data, resulting in two or more confounded components.

Source Component Shift

- Observed data is made up of a number of different sources with different characteristics.
- Proportions of these sources can vary between training and test scenarios.
- How is it different from sample selection bias? In S.S.B, change is a result of measurement process.

Shift or No Shift?

- Using a model designed to consider covariate shift may perform poorly on data where there is no shift.
- Must check model on real environment before rolling out.
- In the same way as semi-supervised learning can provide major benefits over unsupervised learning.

Thank you!