AN INTRODUCTION TO FISHER SCORE-BASED CONTROL CHARTS FOR MONITORING ML MODELS

Leah Jones

MOTIVATION

Machine learning models are used in many high-stakes national security settings

Models can degrade over time as the underlying relationships in the process change.

If that change goes undetected, the model performance can quietly become unreliable.

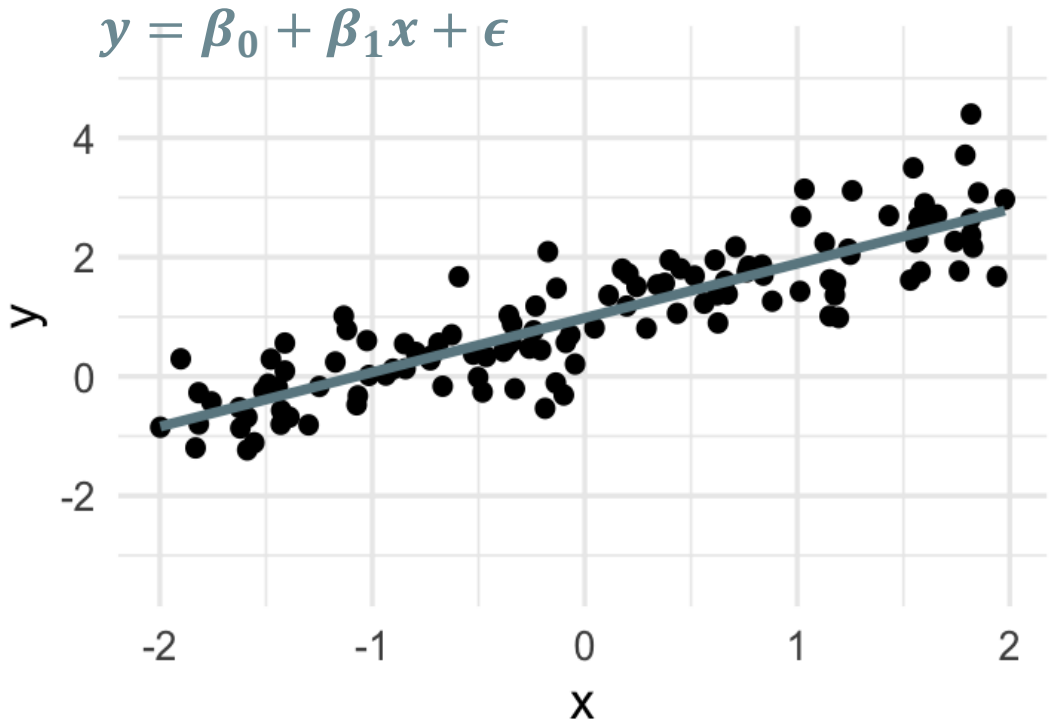
This threat demands clear signals when a model can no longer be trusted.

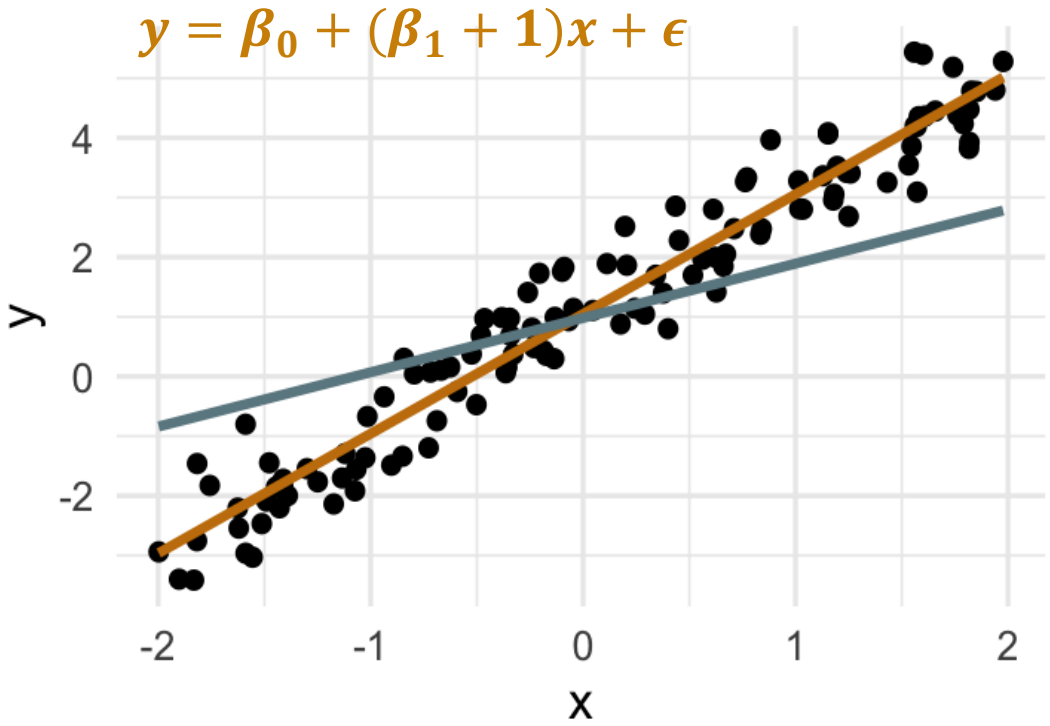
CONCEPT DRIFT

A change in the relationship between input features (X) and the output (Y), meaning $P_t(Y|X) \neq P_{t+1}(Y|X)$.

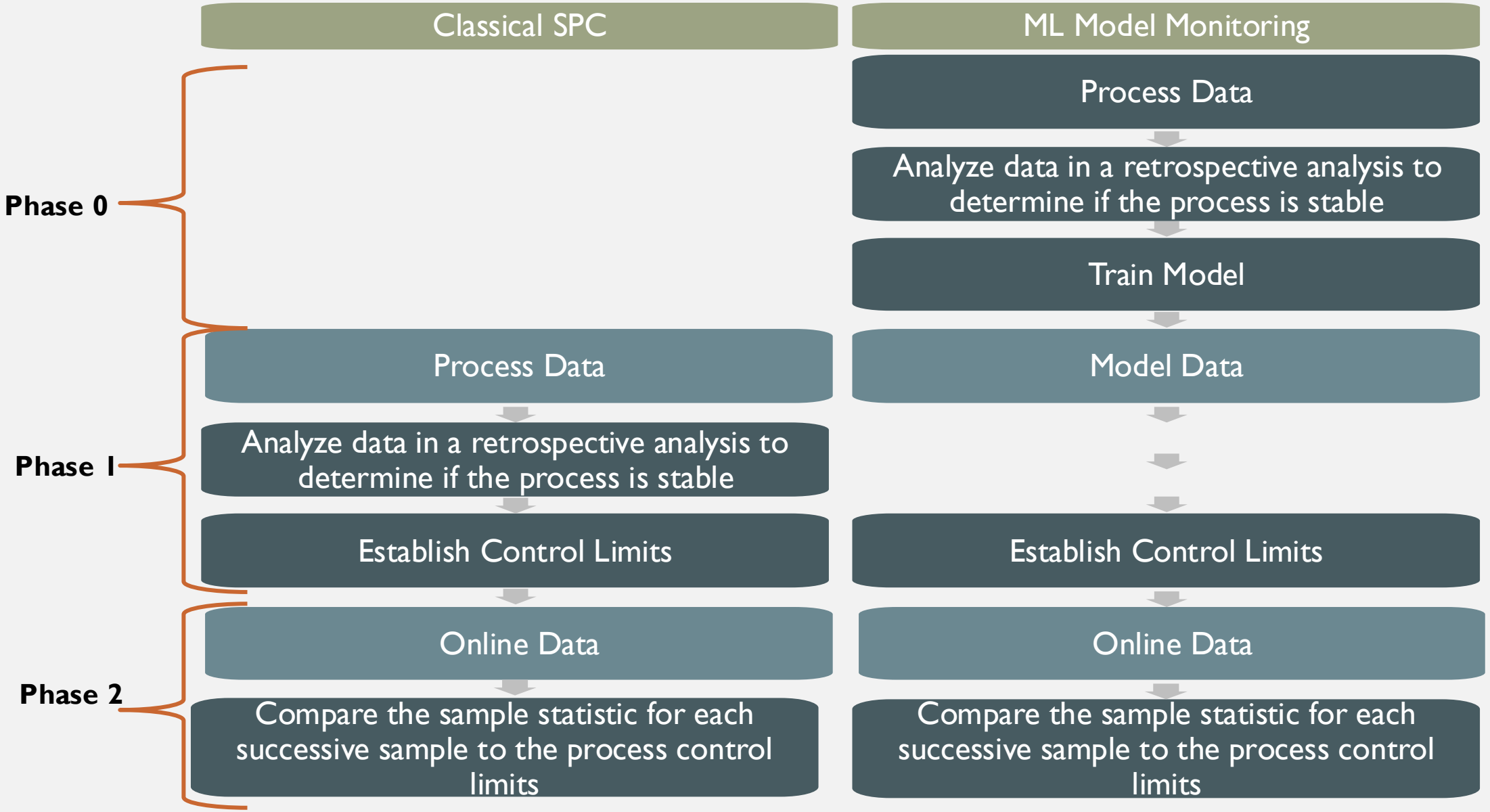
Before drift: the pre-drift model fits well

After drift: using the pre-drift model would mispredict the post-drift relationship



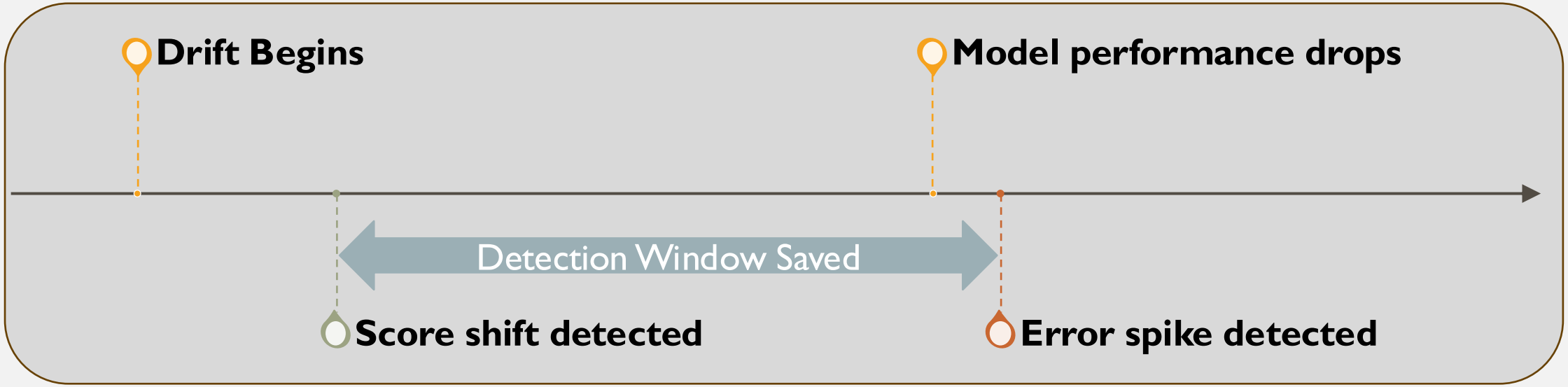


ML MONITORING EXTENDS CLASSICAL SPC



DETECTING CONCEPT DRIFT: TWO APPROACHES

	Error-based Monitoring	Score-based Monitoring
Monitors	The model's mistakes	The model itself
Metrics	Residuals, accuracy, precision, recall, etc.	Scores
When can detect drift	Only after performance drops	Possibly before errors appear



MONITORING THE FISHER SCORE MATRIX

Fisher Score: the gradient of the log-likelihood with respect to the parameters.

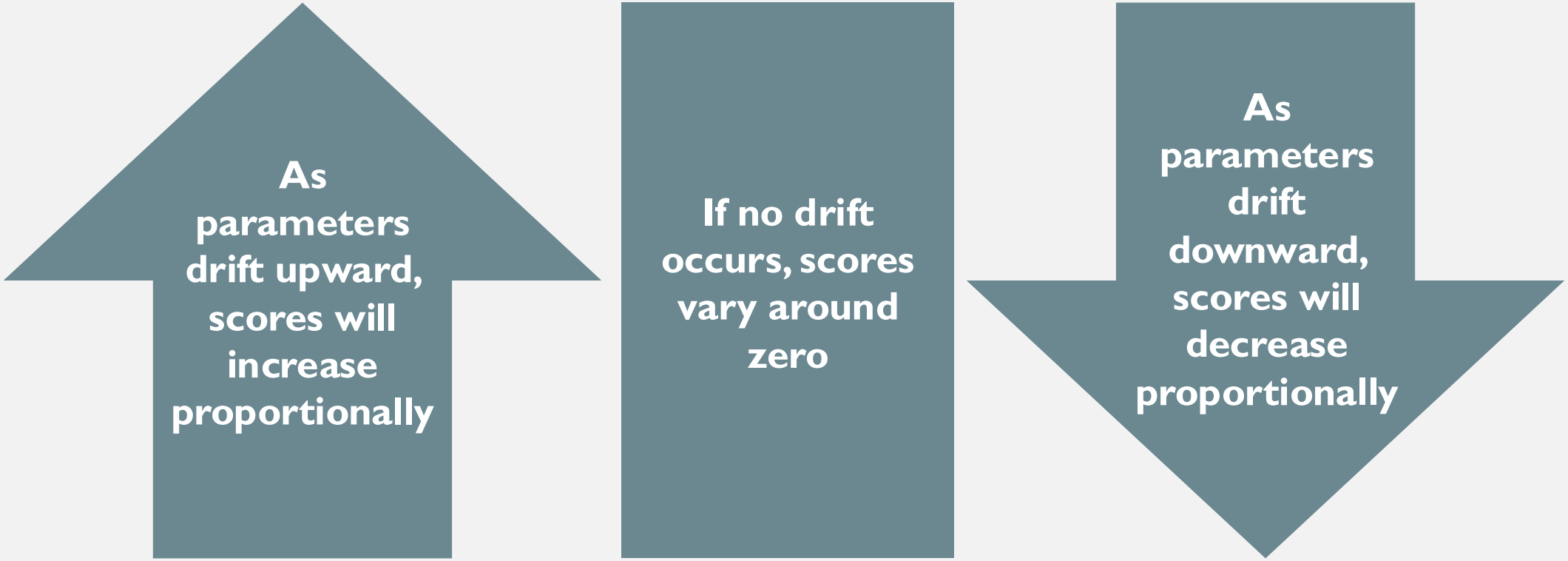
Linear Regression Case: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$

$$s\big(\boldsymbol{\beta}; (X, \boldsymbol{y})\big) = \frac{1}{\sigma^2} X^T (\boldsymbol{y} - X\boldsymbol{\beta})$$

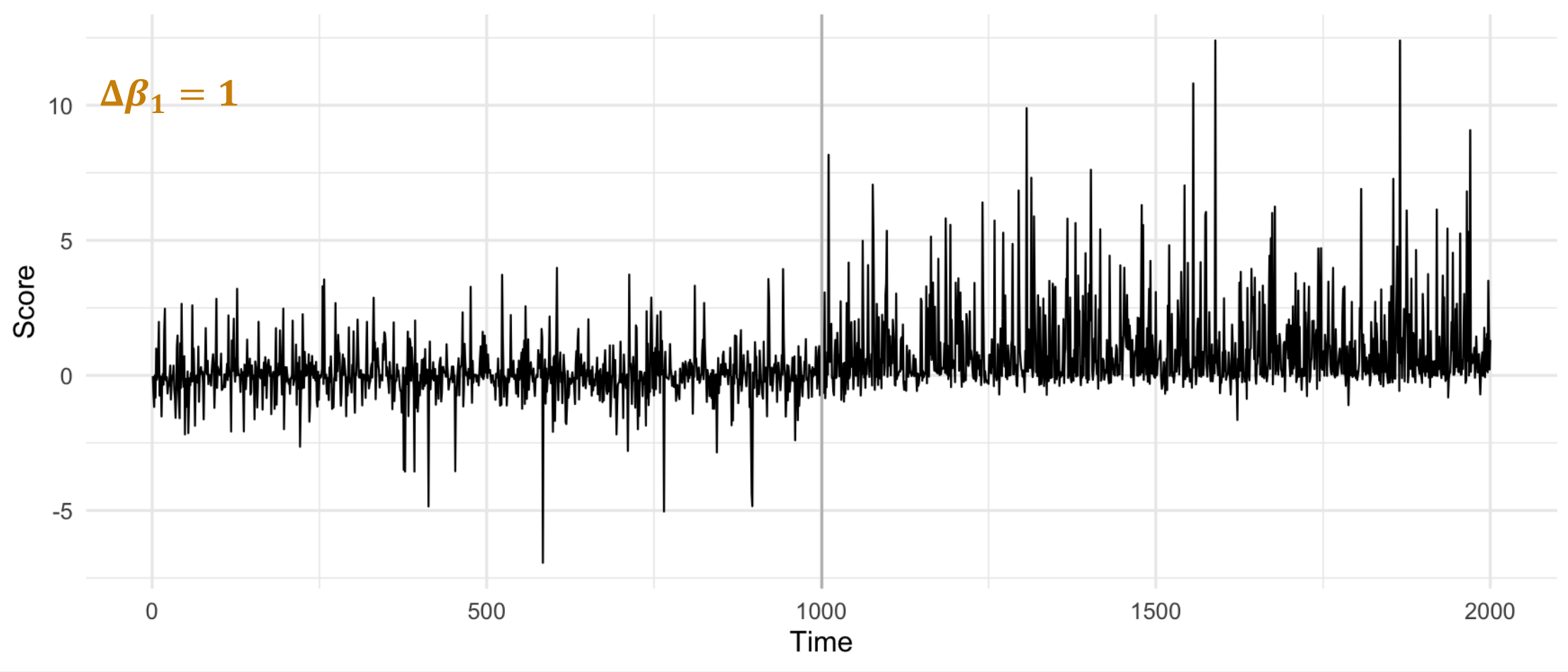
$$= \begin{pmatrix} \frac{1}{\sigma^2}(y_1 - \boldsymbol{x_1\beta})x_{11} + & \frac{1}{\sigma^2}(y_2 - \boldsymbol{x_2\beta})x_{21} + & \cdots & \frac{1}{\sigma^2}(y_n - \boldsymbol{x_n\beta})x_{n1} \\ \frac{1}{\sigma^2}(y_1 - \boldsymbol{x_1\beta})x_{12} + & \frac{1}{\sigma^2}(y_2 - \boldsymbol{x_2\beta})x_{22} + & & \frac{1}{\sigma^2}(y_n - \boldsymbol{x_n\beta})x_{n2} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\sigma^2}(y_1 - \boldsymbol{x_1\beta})x_{1p} + & \frac{1}{\sigma^2}(y_2 - \boldsymbol{x_2\beta})x_{2p} + & & \frac{1}{\sigma^2}(y_n - \boldsymbol{x_n\beta})x_{np} \end{pmatrix} \begin{matrix} \leftarrow \beta_1 \\ \leftarrow \beta_2 \\ \\ \leftarrow \beta_p \end{matrix}$$

Observation 1 Observation 2 Observation n

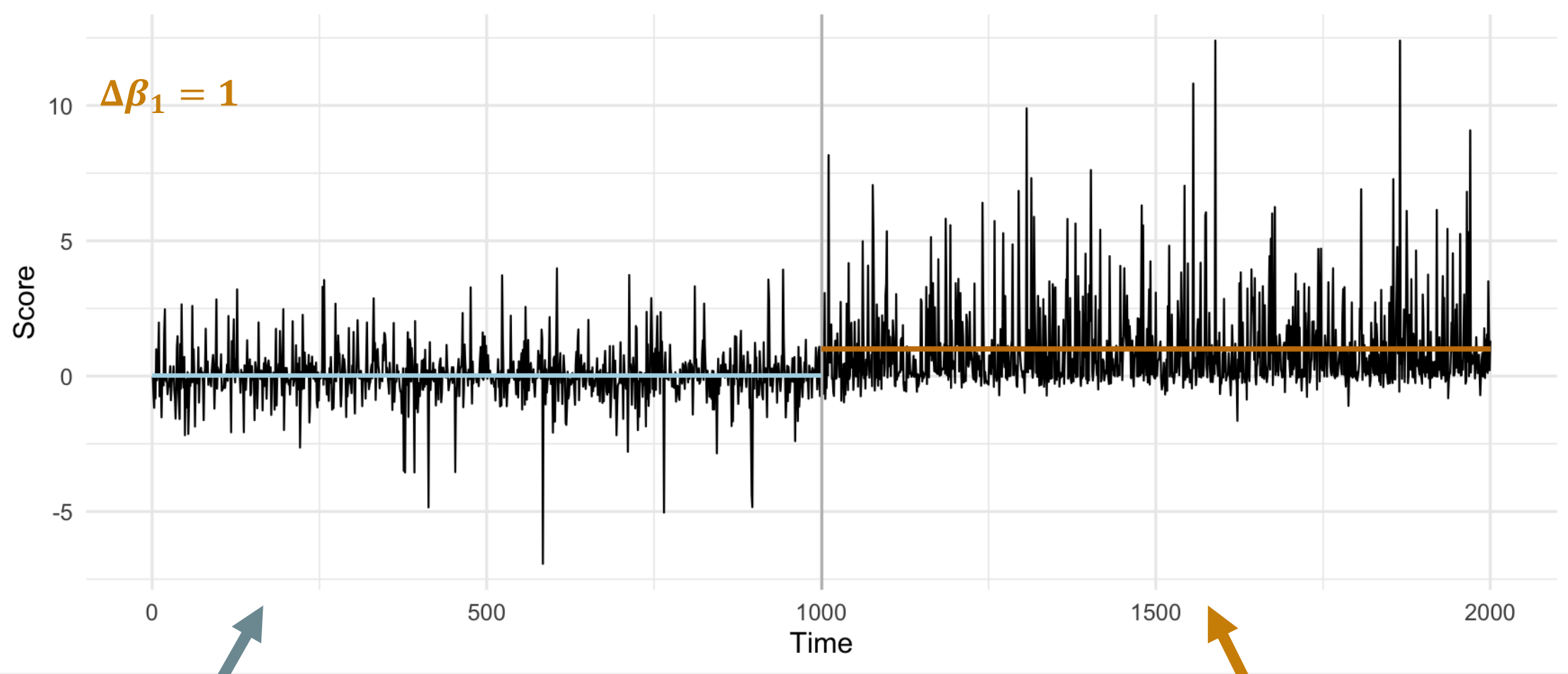
WHAT DO SCORES TELL US ABOUT CONCEPT DRIFT?



EXAMPLE: $y = \beta_0 + \beta_1 x \rightarrow y = \beta_0 + (\beta_1 + 1)x$



EXAMPLE: $y = \beta_0 + \beta_1 x \rightarrow y = \beta_0 + (\beta_1 + 1)x$

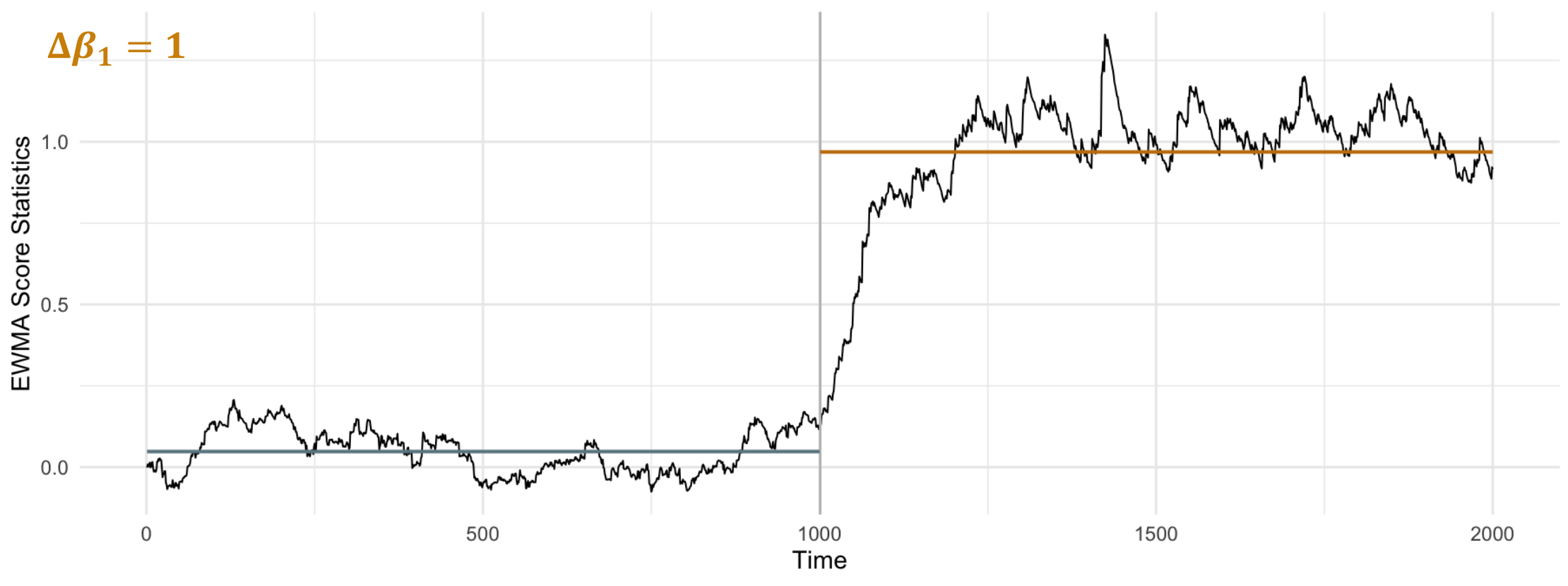


Pre-drift scores vary around 0

Post-drift scores vary around $\Delta\beta$

CAN AVOID FALSE ALARMS DUE TO NOISE BY MONITORING THE EWMA STATISTICS VIA:

$$Z_t = \lambda S_t + (1 - \lambda) Z_{t-1}$$

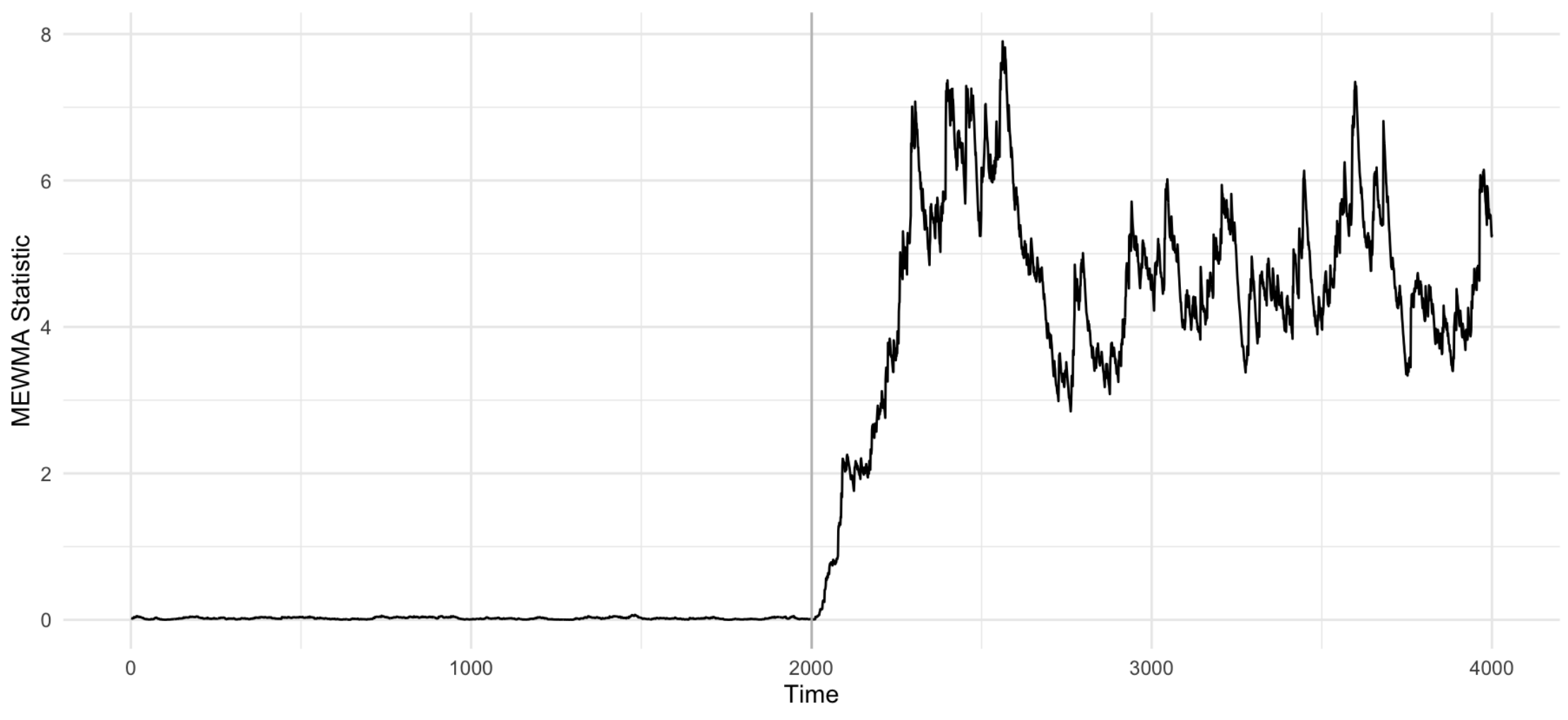


SCORE-BASED MONITORING NOT ONLY DETECTS DRIFT, BUT CAN DIAGNOSE WHICH PARAMETERS ARE DRIFTING

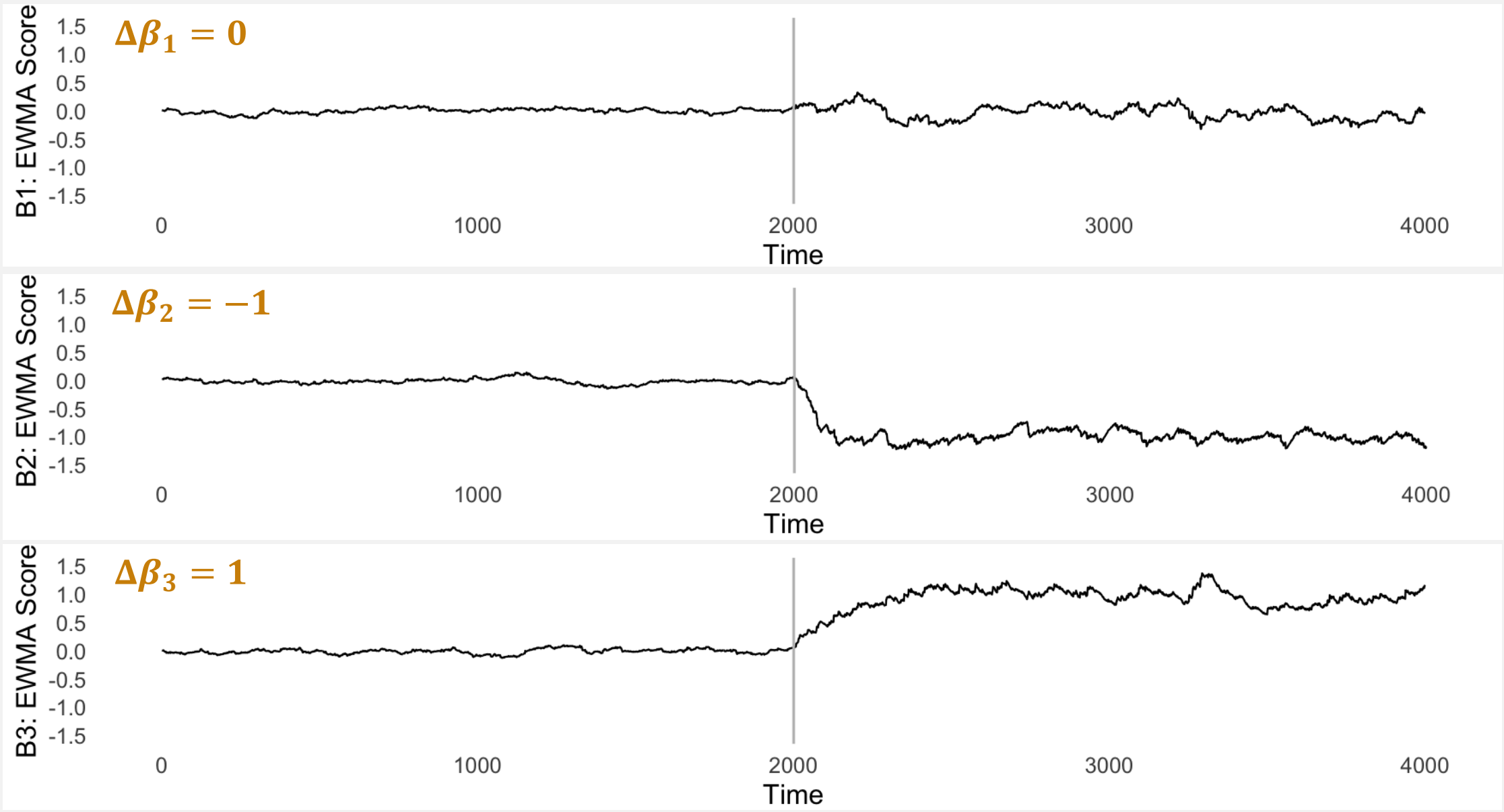
Consider a multivariate model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

1. Monitor the multivariate EWMA value of all parameters:
$$T_t^2 = (z_t - \bar{s})^T \widehat{\Sigma^{-1}} (z_t - \bar{s})$$
2. Once the multivariate statistic flags, study individual parameter scores
3. Update the system or model to address the changing relationship

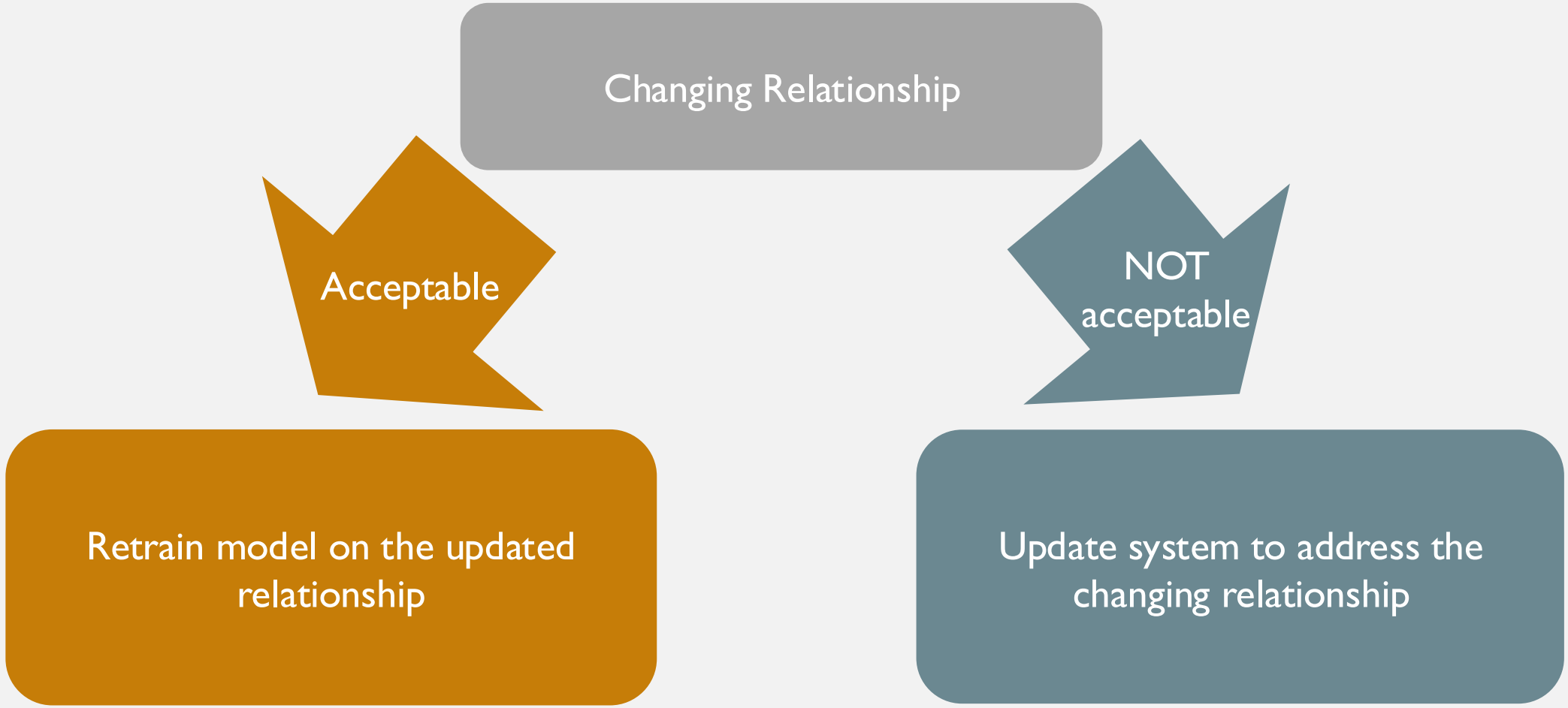
STEP 1) MONITOR THE MULTIVARIATE EWMA VALUE OF ALL PARAMETERS: $T_t^2 = (z_t - \bar{s})^T \widehat{\Sigma^{-1}} (z_t - \bar{s})$



STEP 2) ONCE THE MULTIVARIATE STATISTIC FLAGS, STUDY INDIVIDUAL PARAMETER SCORES



STEP 3) ADDRESS SYSTEM



SUMMARIZING SCORE-BASED MONITORING

- Concept drift occurs when the relationship $P(Y|X)$ changes over time, degrading model performance.
- Score-based monitoring tracks the gradient of the log-likelihood, whose expected value is zero when no drift has occurred.
- Changes in the average score vector indicate concept drift.
- Individual parameter scores can be monitored to diagnose how the process is changing
- Scores can detect drift even if errors don't increase

ADDITIONAL CONSIDERATIONS

Poster session:
4:45-6:30PM

- How do we set control limits?
- How does the distribution of X, $P(X)$, affect score monitoring?
- How does correlation amongst predictors affect score monitoring?
- What if a new variable, previously unmonitored, enters the relationship?
- How do we best use Phase-0 and Phase-I data for estimating monitoring statistics and control limits?
- What about different types of drift?

THANK YOU!