



EUROPEAN CENTRAL BANK

EUROSYSTEM

# Variational Autoencoders for Multivariate Time- Series Outlier Detection

---

Tamara Fajt Mayer

Giovanni Raimondo Quaratino

Juan Francisco Javier Cordero Romero



# Overview

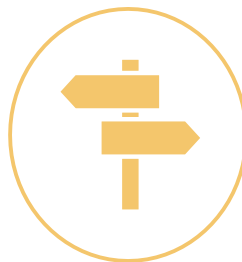
## Context

The Eurosystem – i.e. the ECB and national central banks of the euro area – uses bank-level statistical information to deepen analysis of monetary and economic developments



## Objective

Explore the feasibility of using deep generative learning techniques for outlier detection on bank balance sheet data



## Next steps

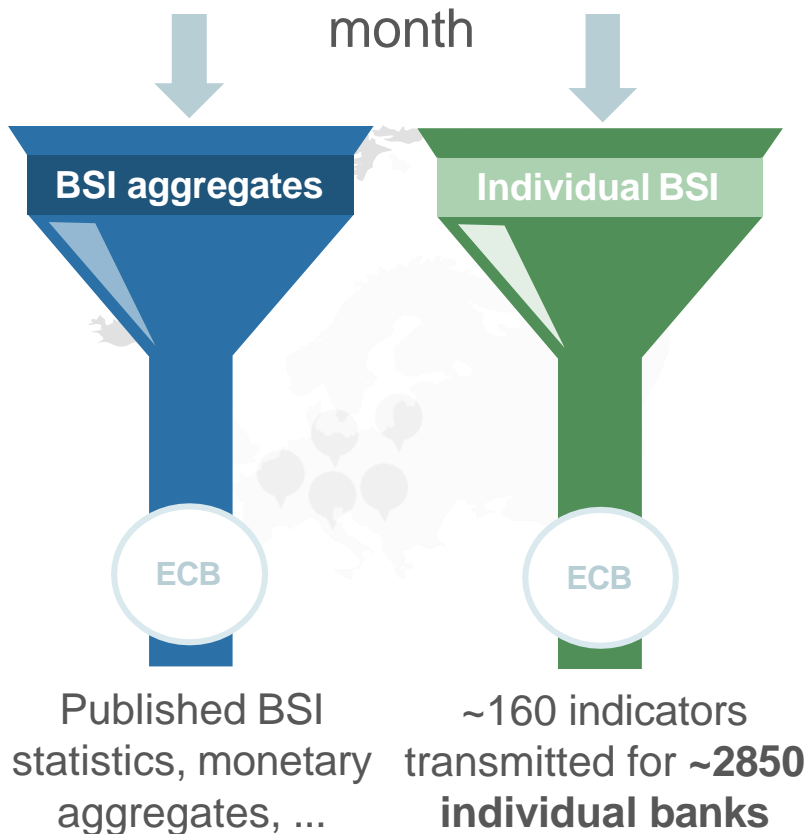
The results indicate very robust and efficient detection of outliers, but these techniques would need to be integrated into production pipelines



# The IBSI Dataset

## Context

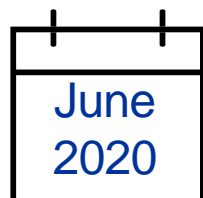
**National central banks** collect balance sheet items (**BSI**) data from euro area banks each month



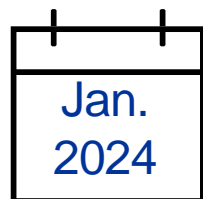
**Individual BSI** transmission to the ECB began in **2012** to provide policy-makers and network of users with vital bank-level data



Start of back data (initially for ~250 banks)



Most recent significant expansion in coverage



Cut-off date for model training

IBSI indicators cover the main needs for the analysis of **monetary aggregates** and **credit**

<u>Assets</u>	<u>Liabilities</u>
Cash	Deposits
Loans	Capital and reserves
Debt securities held	Debt securities issued
External assets	External liabilities

- End-month **outstanding amounts**
- For key series, also adjustments to allow derivation of **transactions**

### The IBSI dataset includes:

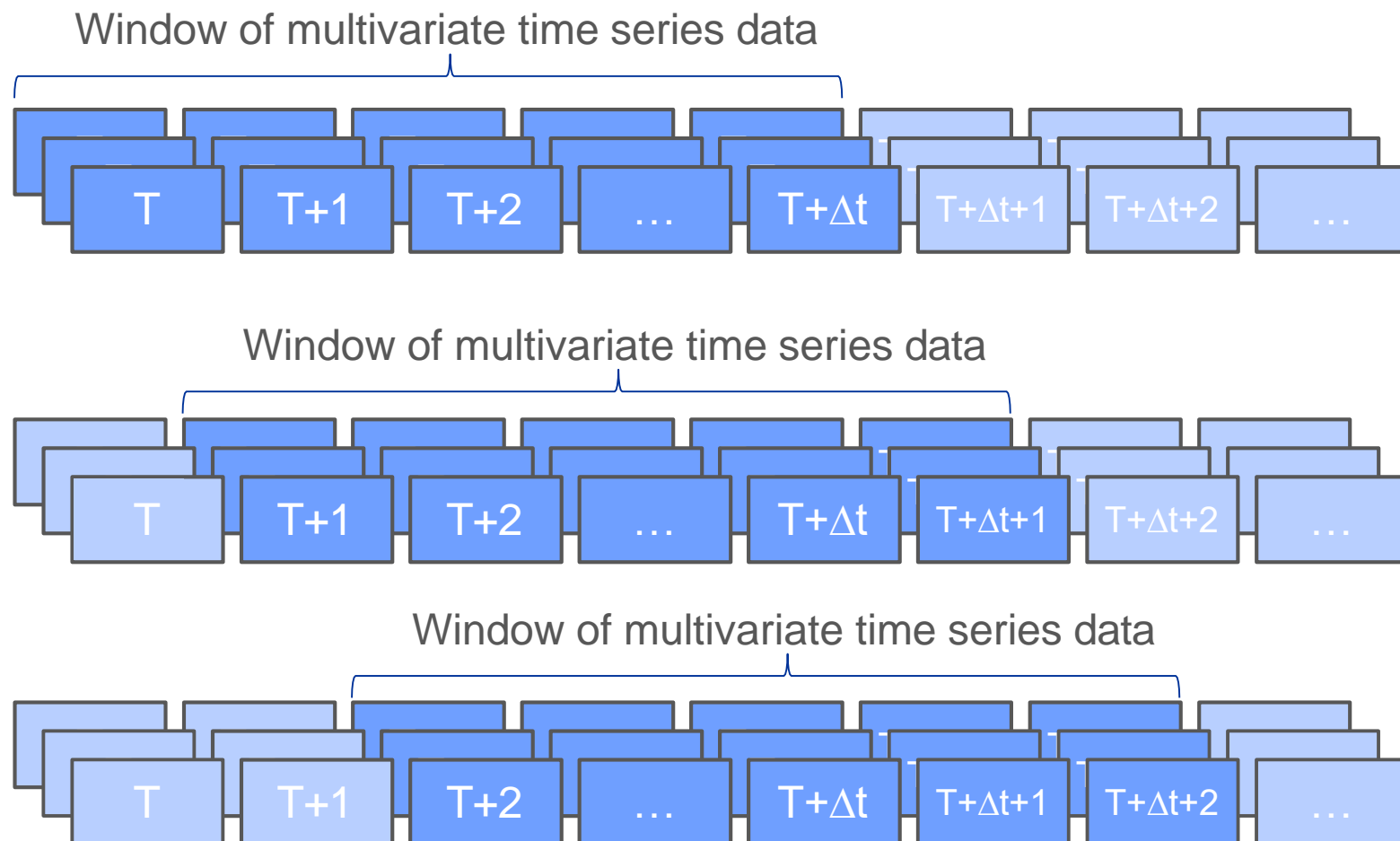
- Bank-specific characteristics
- Temporal dependencies

### Data reshaping into training sequences:

- By window size (12) in steps of 1 (stacked)
- By MFI (158 series each)

### Ensuring temporal consistency:

- Pad zeros to sequences that are too short
- Apply Z-score normalization



**SDMX categorical variables** from IBSI series keys were incorporated into the training data to improve performance

Accomplished with **Multi Hot Encoding** algorithm:

- Works by converting all unique values in a categorical column to **binary digits**, then encoding each value into its binary format, creating new columns
- **Advantages:** saves computational resources and avoids the high dimensionality and sparsity of one-hot encoding

Dataset (untransformed)									
Index	Balance Sheet Item	...							
0	A2C	...							
1	A20	...							
2	L22		Dataset (Multi Hot Encoded)						
3	L40		Index	Balance Sheet Item	Balance Sheet Item_0	Balance Sheet Item_1	Balance Sheet Item_2	Balance Sheet Item_3	...
4	A51		0	A2C	0	0	0	0	...
5	L21		1	A20	0	0	0	1	...
6	L7C		2	L22	0	0	1	0	...
7	A10		3	L40	0	0	1	1	...
8	L2A		4	A51	0	1	0	0	...
9	L7A		5	L21	0	1	0	1	...
			6	L7C	0	1	1	0	...
			7	A10	0	1	1	1	...
			8	L2A	1	0	0	0	...
			9	L7A	1	0	0	1	...

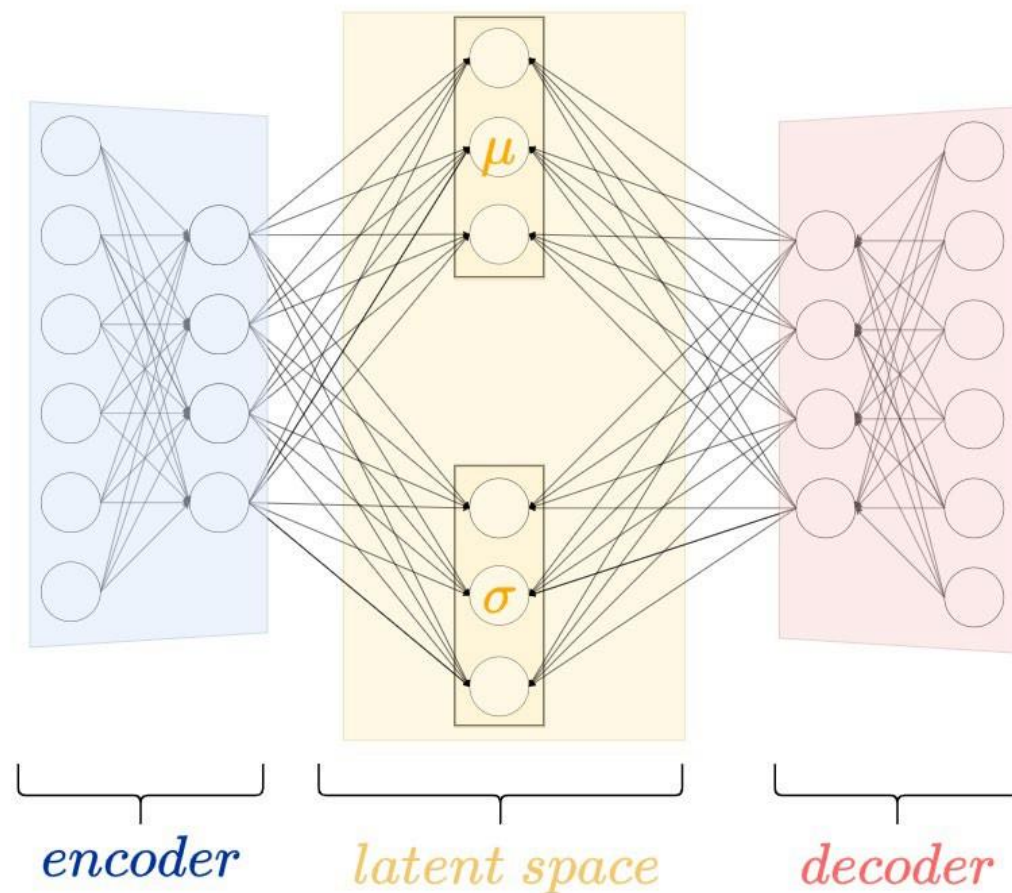


## Variational Autoencoders (VAEs)

- **Probabilistic deep generative learning models** for anomaly detection
- If trained on sufficient data, the model can **accurately represent the IBSI dataset distribution**
- VAEs will learn to reconstruct the training data well but **fail to reconstruct anomalies** effectively

VAE training optimizes the **Evidence Lower Bound (ELBO)**, composed of two terms:

- **Reconstruction Term:** Measures how well the model reconstructs the observed data (**MSE**)
- **Regularization Term (KL divergence):** Ensures the variational distribution stays close to the prior distribution, preventing overfitting



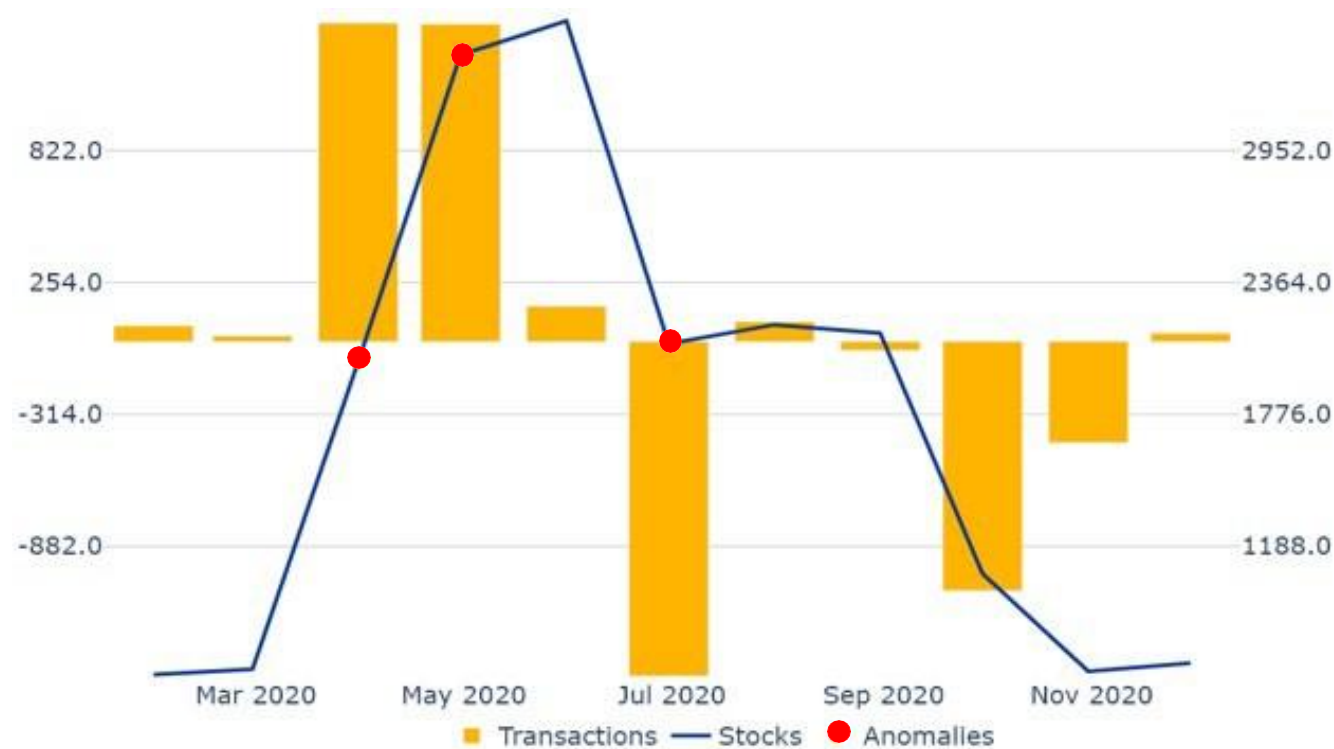
## Reconstruction error

- The difference between input and output
- Serves as the **indicator of anomaly**

## Normalization of reconstruction errors

- Assessing the significance of flagged anomalies relative to their volume enables compilers to focus efforts on most meaningful outliers
- For example, reconstruction errors (E) can be weighted by a normalized share of volume ( $\alpha$ ) to provide a **balanced number of outliers across countries**:

$$E_{\text{normalized}} = E(1 + 100\alpha^2)$$



MFI deposits - stocks (rhs) and flows (lhs) - with anomalies flagged by VAE

### Aim of the benchmarks:

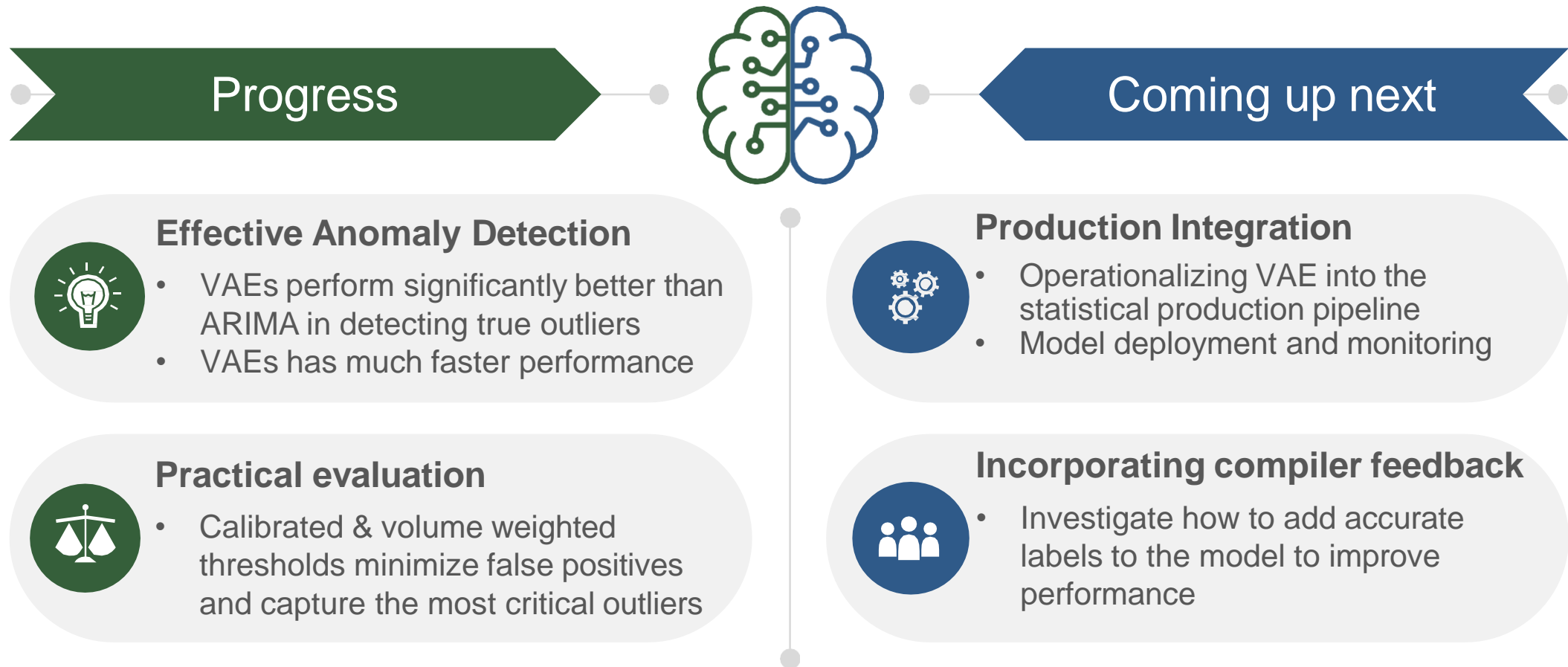
- Compare the performance of the currently used ARIMA model and VAE on synthetic outliers
- Measured for **one reference period**, containing **529,932 data points**

### Key findings:

- Both models performed comparably in detecting non-outlier points
- VAE **significantly outperformed** ARIMA in detecting true outlier points
- VAE demonstrated **superior time efficiency and performance metrics**
- ARIMA model struggled with the dataset's scale

	ARIMA			VAE		
Label	Precision	Recall	F1	Precision	Recall	F1
False	1.00	0.99	0.99	1.00	1.00	1.00
True	0.15	0.57	0.24	0.95	0.89	0.92
Duration	16 hours 38 minutes			10 minutes		





# Thank you!

