SAND2024-06451C

Pacific Northwest NATIONAL LABORATORY

AIRES

Heterogenous Data Fusion with Variational Autoencoders

May 22, 2024

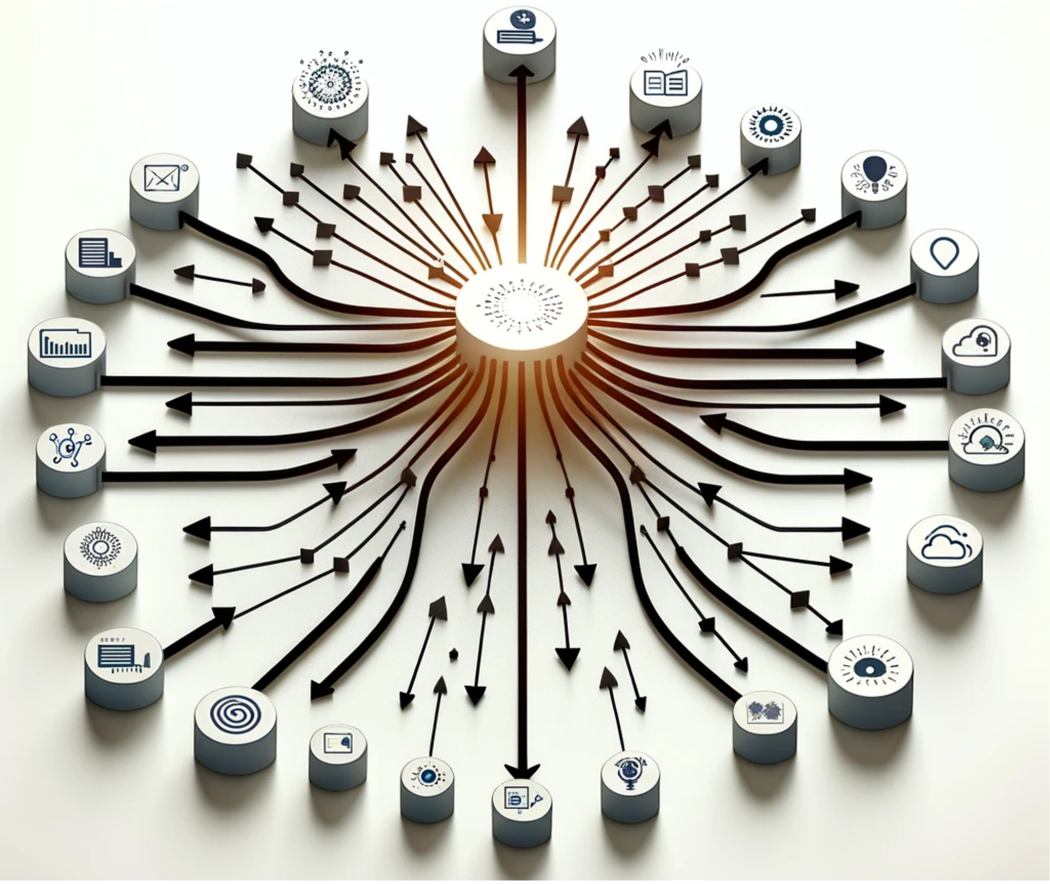
Lekha Patel
Computational Statistician
Scientific Machine Learning @ Sandia National Labs

U.S. DEPARTMENT OF ENERGY BATTELLE

What is data fusion?

Data fusion integrates multiple data sources (of potentially varying data structures) to produce **more consistent, accurate, and useful information** than that provided by any individual data source alone.

- Enhances decision-making capabilities.
- Reduces uncertainty by combining information from different sources.
- Enables comprehensive analysis across different data types.

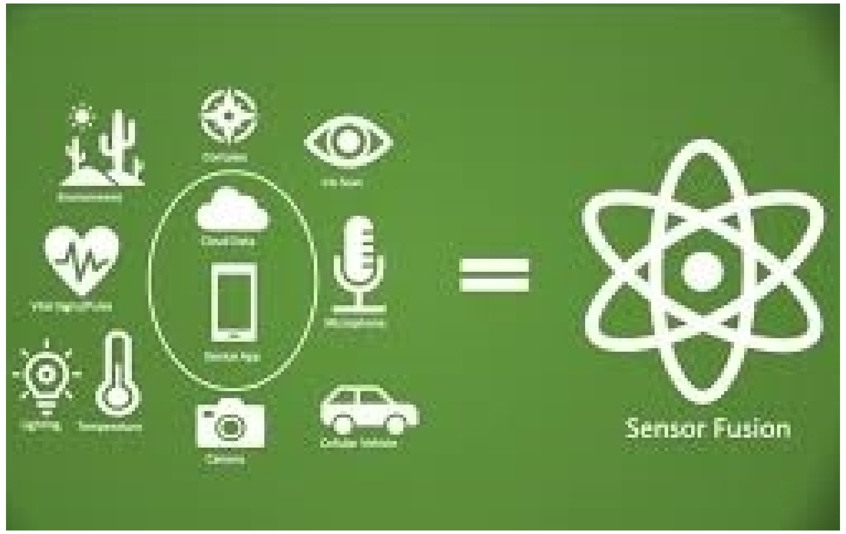


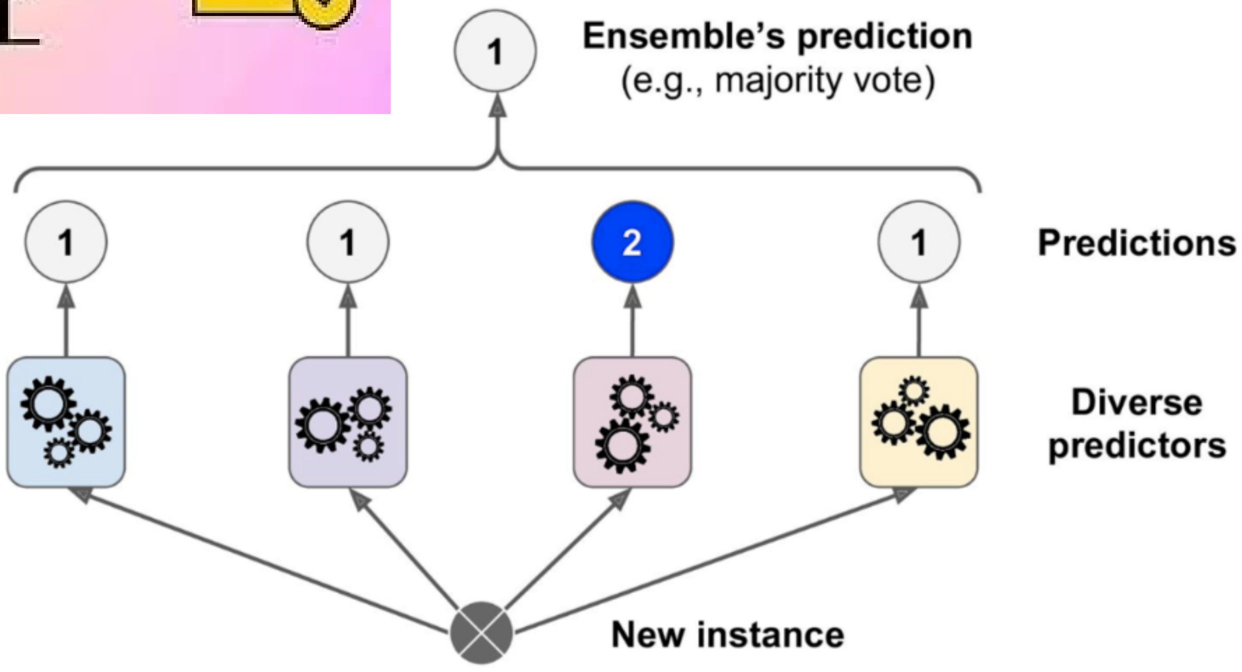
Classical approaches

- **Data Fusion:** Combining raw data from different sources.

- **Feature Fusion:** Extracting features from each data source and then combining them.

- **Decision Fusion:** Combining decisions from multiple models or algorithms.





Limitations

Data Heterogeneity: Difficulty in handling different types of data (e.g., text, images, numerical data) due to structural and statistical differences.

Feature Incompatibility: Challenges in combining features from different modalities. How can they be compared?

Loss of Information: Potential loss of important information during the fusion process.

Scalability Issues: Difficulty in scaling with increasing data volume and variety, particularly for large-scale multimodal fusion.

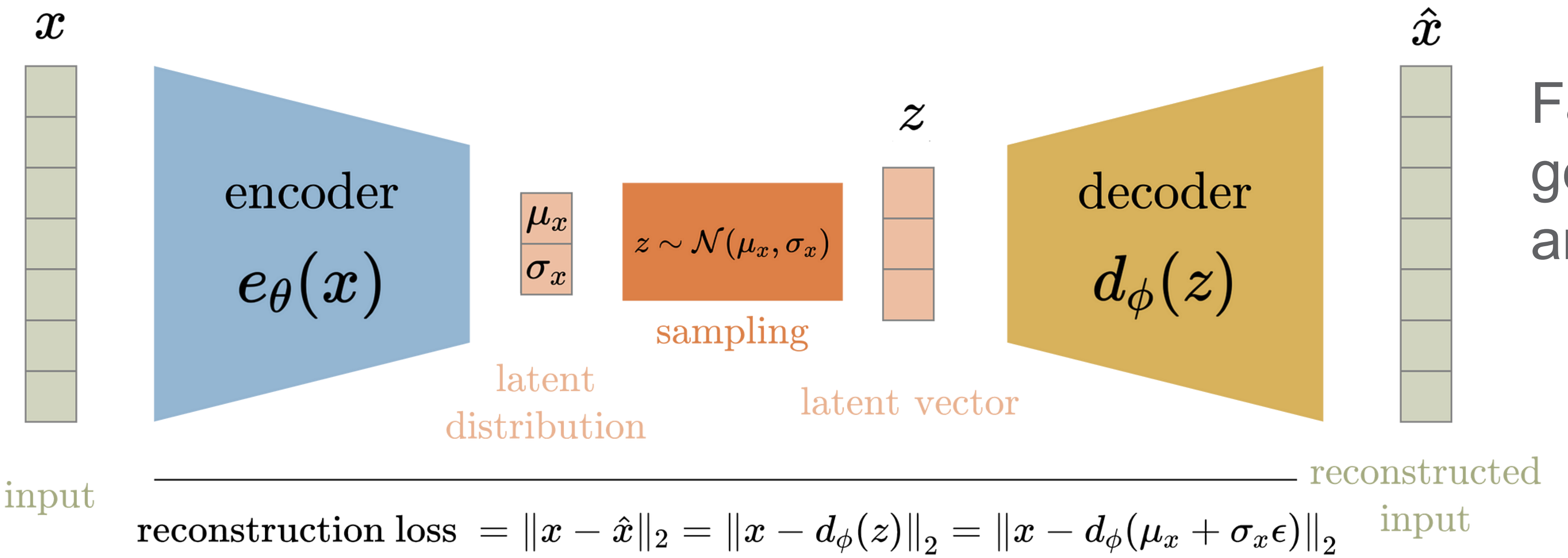
Traditional methods often struggle with the complexities of multimodal data. Generative ML models (e.g. Variational Autoencoders (VAEs)) offer a powerful and interpretable solution by providing a unified framework for encoding and decoding diverse data types.

Pacific Northwest
NATIONAL LABORATORY

Variational Autoencoders (VAEs)

VAEs are a *generative* ML model that learn to *encode* data into a latent space and then *decode* to enable the novel generation or *sampling* of new data from the approximated statistical distribution.

Handles high-dimensional and complex data



Facilitates data generation and anomaly detection

$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

$$loss = reconstruction\ loss + similarity\ loss$$

Represents data probabilistically

Pacific Northwest NATIONAL LABORATORY

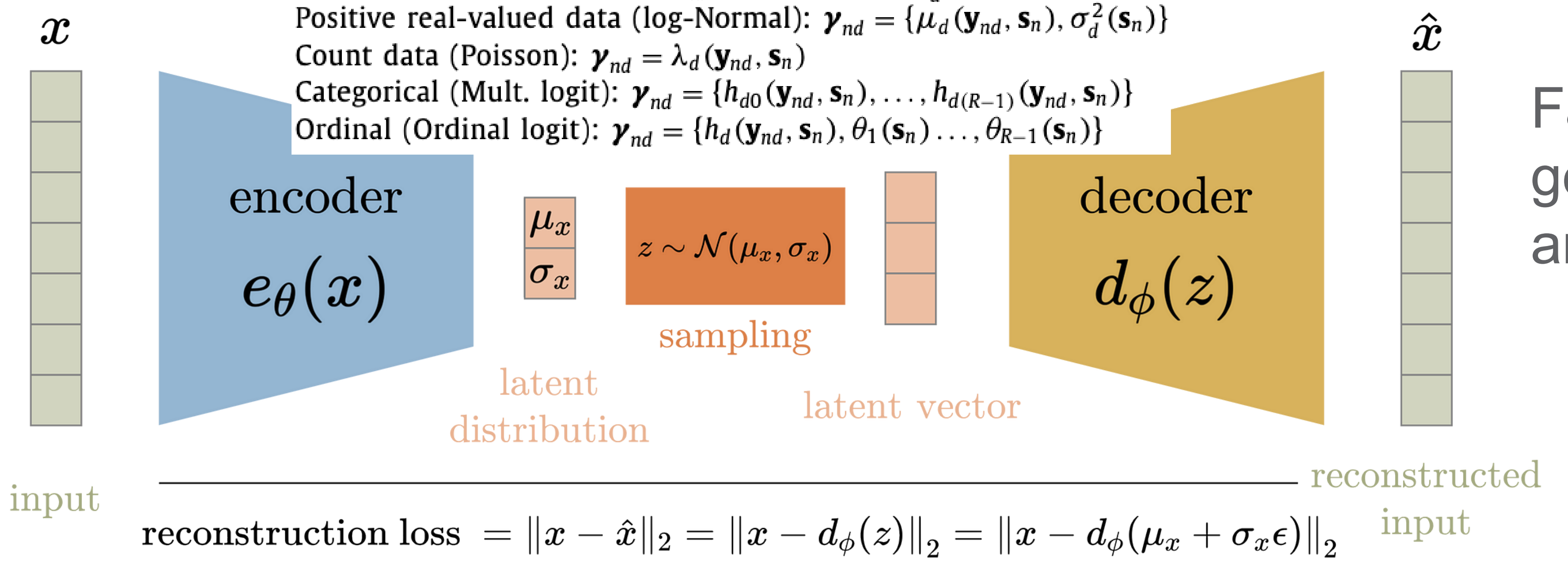
Fusing Heterogeneous Data with VAEs

VAEs can be used to fuse heterogeneous data by encoding different data types through suitable prior distributions into a shared latent space, capturing their common structure.

See e.g. Nazábal et al., 2020

Real-valued data (Normal): $\boldsymbol{\gamma}_{nd} = \{\mu_d(\mathbf{y}_{nd}, \mathbf{s}_n), \sigma_d^2(\mathbf{s}_n)\}$
Positive real-valued data (log-Normal): $\boldsymbol{\gamma}_{nd} = \{\mu_d(\mathbf{y}_{nd}, \mathbf{s}_n), \sigma_d^2(\mathbf{s}_n)\}$
Count data (Poisson): $\boldsymbol{\gamma}_{nd} = \lambda_d(\mathbf{y}_{nd}, \mathbf{s}_n)$
Categorical (Mult. logit): $\boldsymbol{\gamma}_{nd} = \{h_{d0}(\mathbf{y}_{nd}, \mathbf{s}_n), \ldots, h_{d(R-1)}(\mathbf{y}_{nd}, \mathbf{s}_n)\}$
Ordinal (Ordinal logit): $\boldsymbol{\gamma}_{nd} = \{h_d(\mathbf{y}_{nd}, \mathbf{s}_n), \theta_1(\mathbf{s}_n) \ldots, \theta_{R-1}(\mathbf{s}_n)\}$

Need to balance reconstruction accuracy with latent space regularization



Facilitates data generation and anomaly detection

$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\textit{similarity loss} = \textit{KL Divergence} = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

Change parameters of interest wrt to latent distributions

$$\textit{loss} = \textit{reconstruction loss} + \textit{similarity loss}$$

Represents data probabilistically

Pacific Northwest NATIONAL LABORATORY

Fusing Heterogeneous Data with VAEs

VAEs can be used to fuse heterogeneous data by encoding different data types through suitable prior distributions into a shared latent space, capturing their common structure.

See e.g. Nazábal et al., 2020

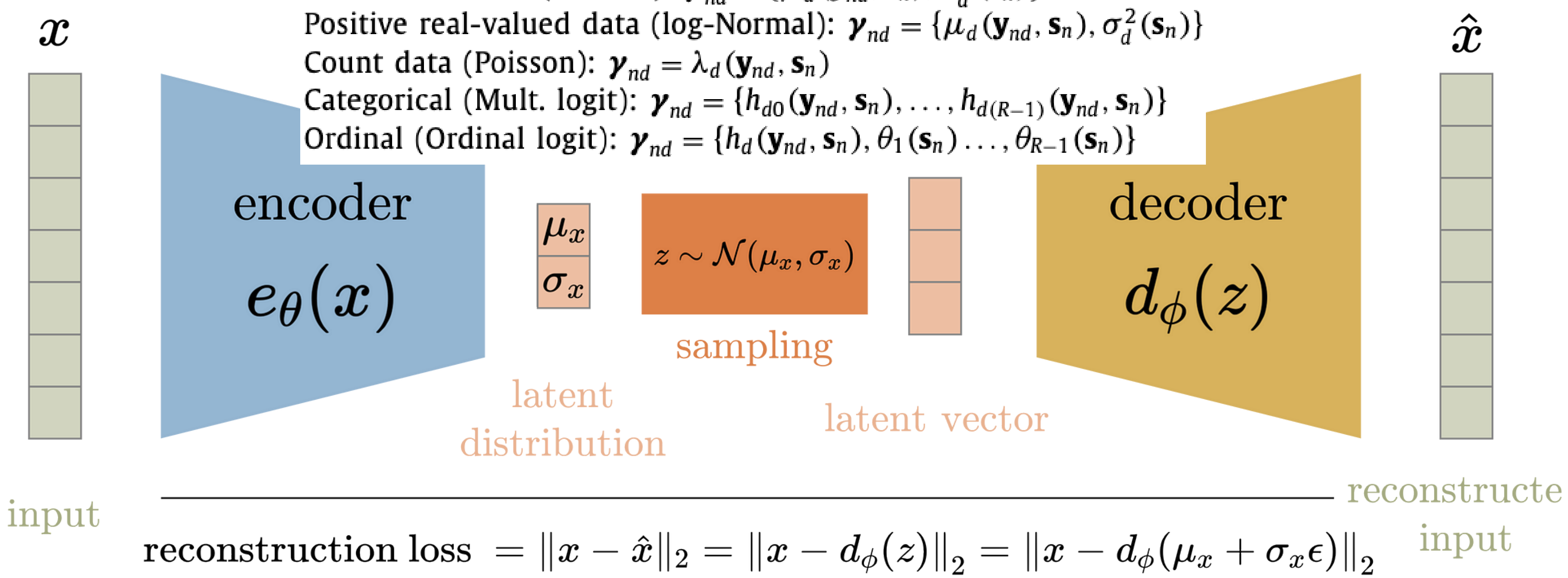
Real-valued data (Normal): $\boldsymbol{\gamma}_{nd} = \{\mu_d(\mathbf{y}_{nd}, \mathbf{s}_n), \sigma_d^2(\mathbf{s}_n)\}$

Positive real-valued data (log-Normal): $\boldsymbol{\gamma}_{nd} = \{\mu_d(\mathbf{y}_{nd}, \mathbf{s}_n), \sigma_d^2(\mathbf{s}_n)\}$

Count data (Poisson): $\boldsymbol{\gamma}_{nd} = \lambda_d(\mathbf{y}_{nd}, \mathbf{s}_n)$

Categorical (Mult. logit): $\boldsymbol{\gamma}_{nd} = \{h_{d0}(\mathbf{y}_{nd}, \mathbf{s}_n), \ldots, h_{d(R-1)}(\mathbf{y}_{nd}, \mathbf{s}_n)\}$

Ordinal (Ordinal logit): $\boldsymbol{\gamma}_{nd} = \{h_d(\mathbf{y}_{nd}, \mathbf{s}_n), \theta_1(\mathbf{s}_n) \ldots, \theta_{R-1}(\mathbf{s}_n)\}$



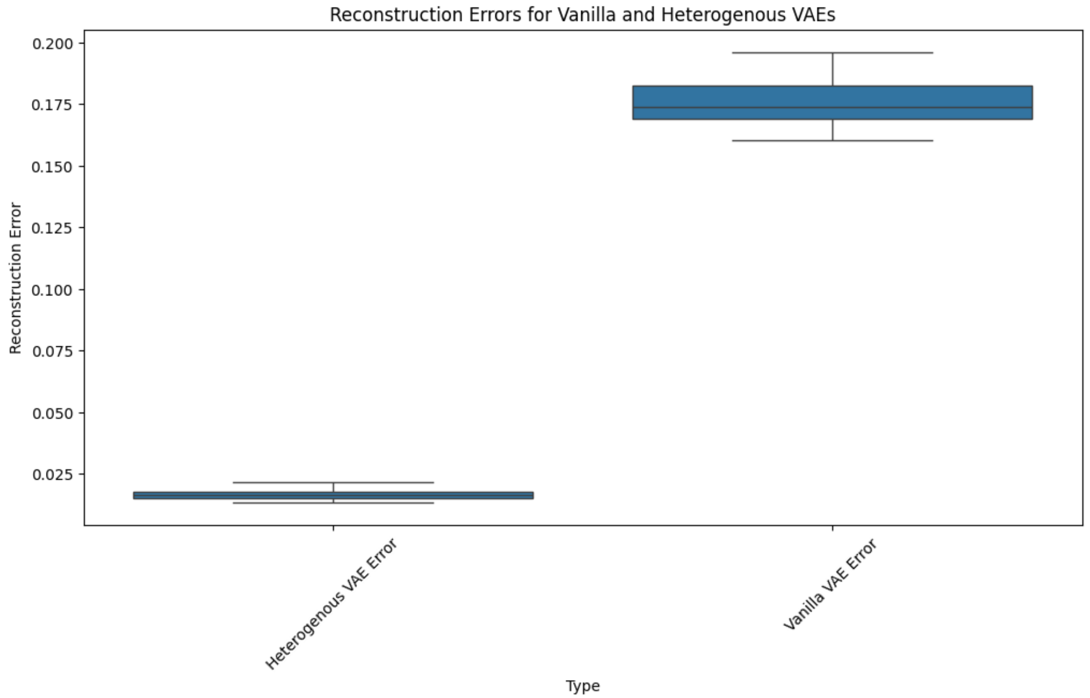
$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

$$loss = reconstruction\ loss + similarity\ loss$$

Optimization in the latent space can be done in the *usual way* (i.e. with reparameterization tricks that are not unique to Gaussians)



Cybersecurity Applications

Example 1: Intrusion Detection Systems (IDSs)

- Heterogenous VAEs can integrate network traffic data, system logs, and user behavior for anomaly detection.
- **Benefits**: Improved detection accuracy (without loss of underlying network structure) with potentially reduced false positives.

Example 2: Threat Intelligence

- Heterogeneous VAEs can combine structured and unstructured data (e.g., threat reports, IP addresses, malware signatures) for comprehensive threat analysis.
- **Benefits**: Enhanced situational awareness and proactive threat mitigation.

Cybersecurity: What features exist?

Real-Valued Features (communication statistics):

- Examples: "Flow Duration", "Total Length of Fwd Packet", "Fwd Packet Length Mean"
- Characteristics: Continuous values capturing metrics such as duration, lengths, and means.

Positive-Valued Features (communication times) :

- Examples: "Flow Bytes/s", "Flow Packets/s", "Fwd IAT Mean", "Bwd IAT Std"
- Characteristics: Non-negative continuous values representing rates, inter-arrival times, and statistical measures.

Count Features (communication types):

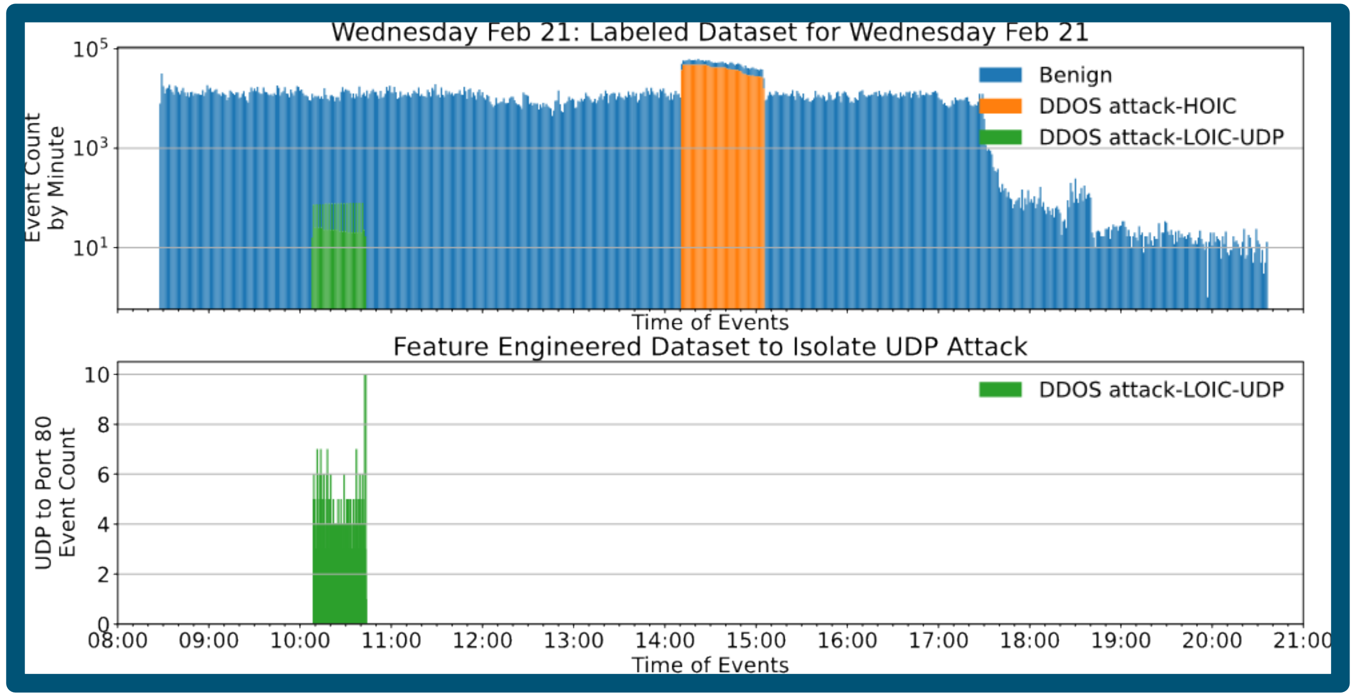
- Examples: "Total Fwd Packet", "Total Bwd packets", "FIN Flag Count", "SYN Flag Count"
- Characteristics: Integer values indicating counts of packets, flags, and other discrete events.

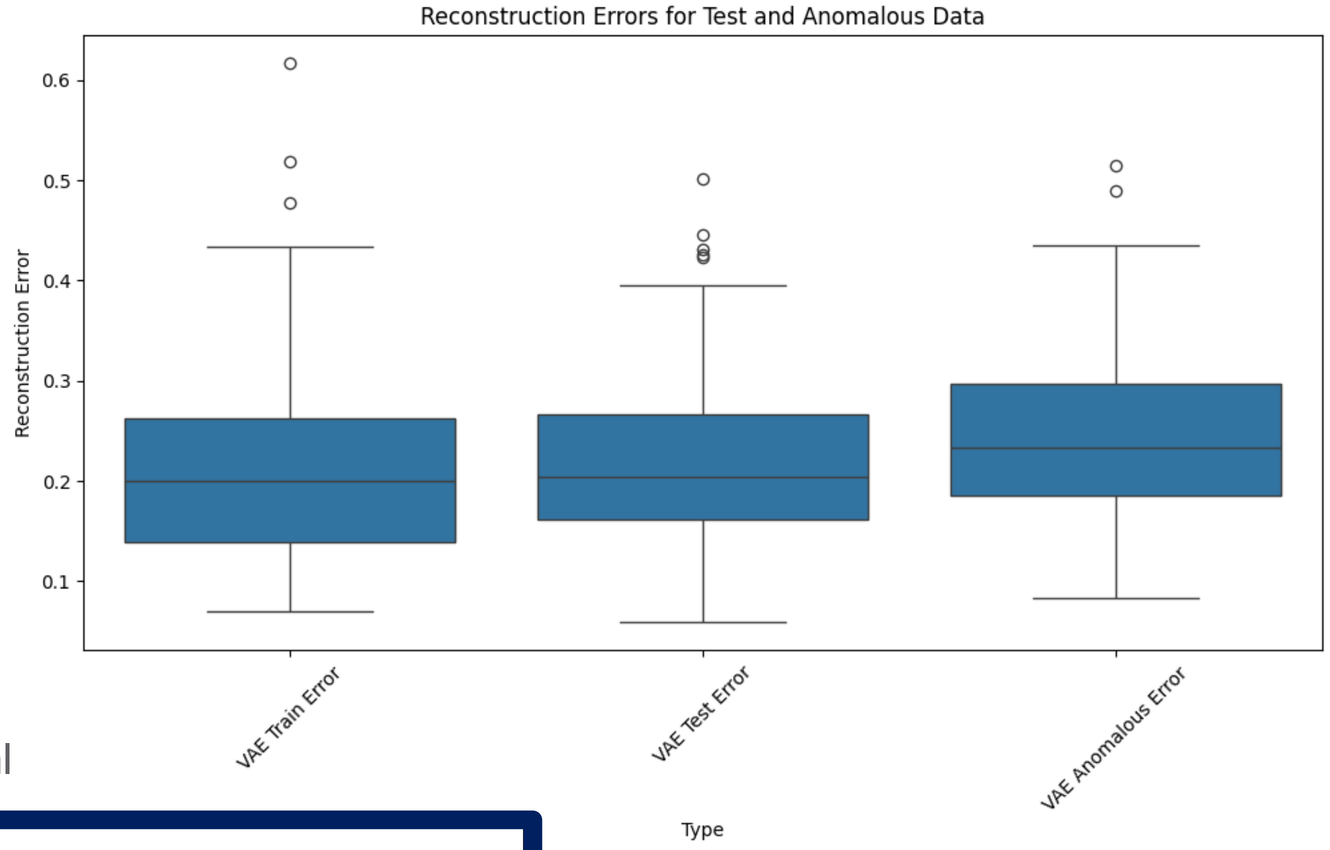
Categorical and Ordinal Features (particularly as network characteristics):

- Examples: "Protocol", "ICMP Type", "Src Port", "Dst Port"
- Characteristics: Discrete values representing categories or ranks, such as protocol types and port numbers.

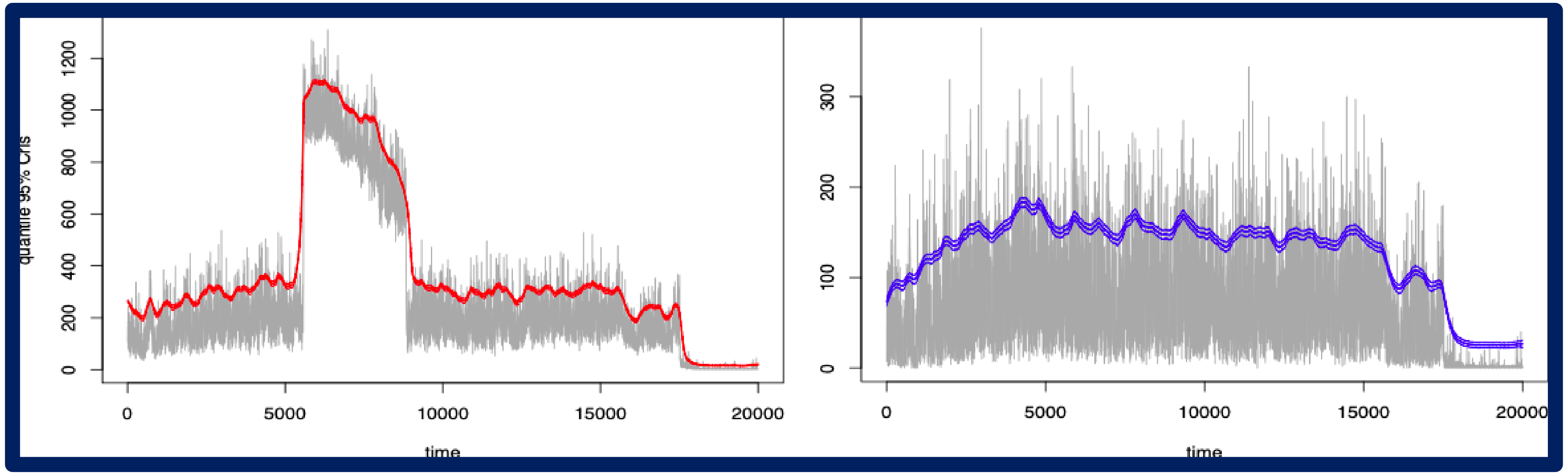
Robust detection accuracy of cyber threats is becoming increasingly important in the era of big data.

Cybersecurity: Case study





Open source data available at: https://www.unb.ca/cic/datasets/index.html



Pacific Northwest
NATIONAL LABORATORY

Physics-based Applications

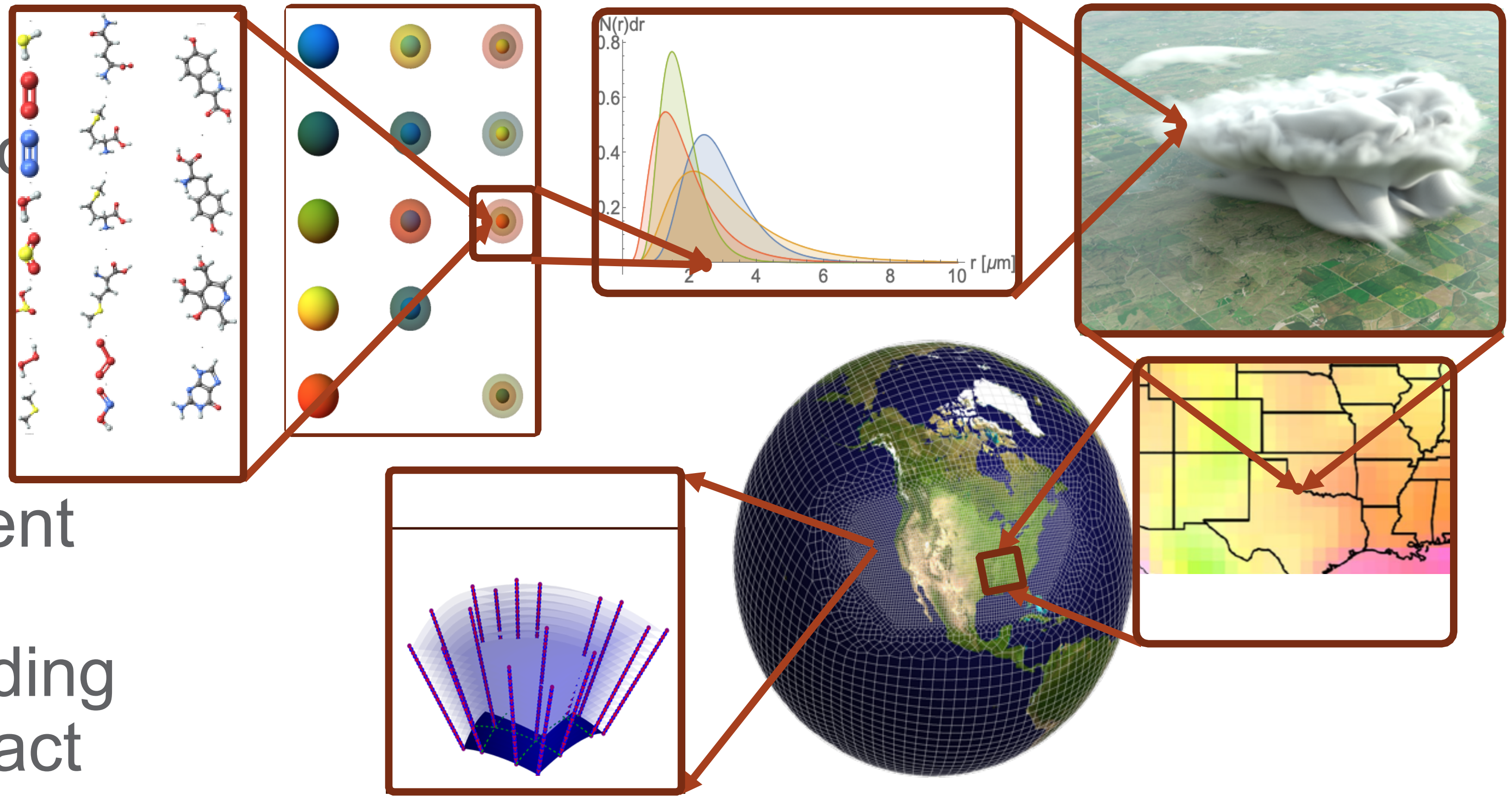
Example 1: Climate Modeling

Explanation: Using VAEs to combine observational data, satellite imagery, and simulation outputs for better climate predicti

Benefits: Improved model accuracy and predictions.

Example 2: Aerosol Dynamics

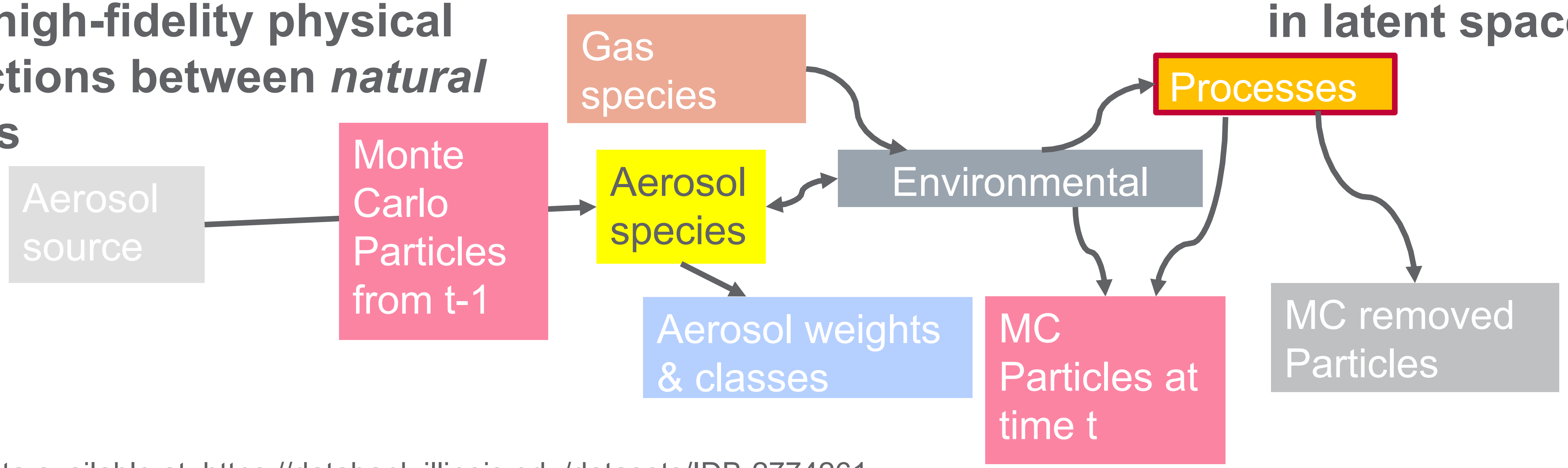
- **Explanation**: Integrating different types of aerosol measurement data to enhance the understanding of aerosol behavior and its impact on climate.
- **Benefits**: More accurate simulations for aerosol processes.



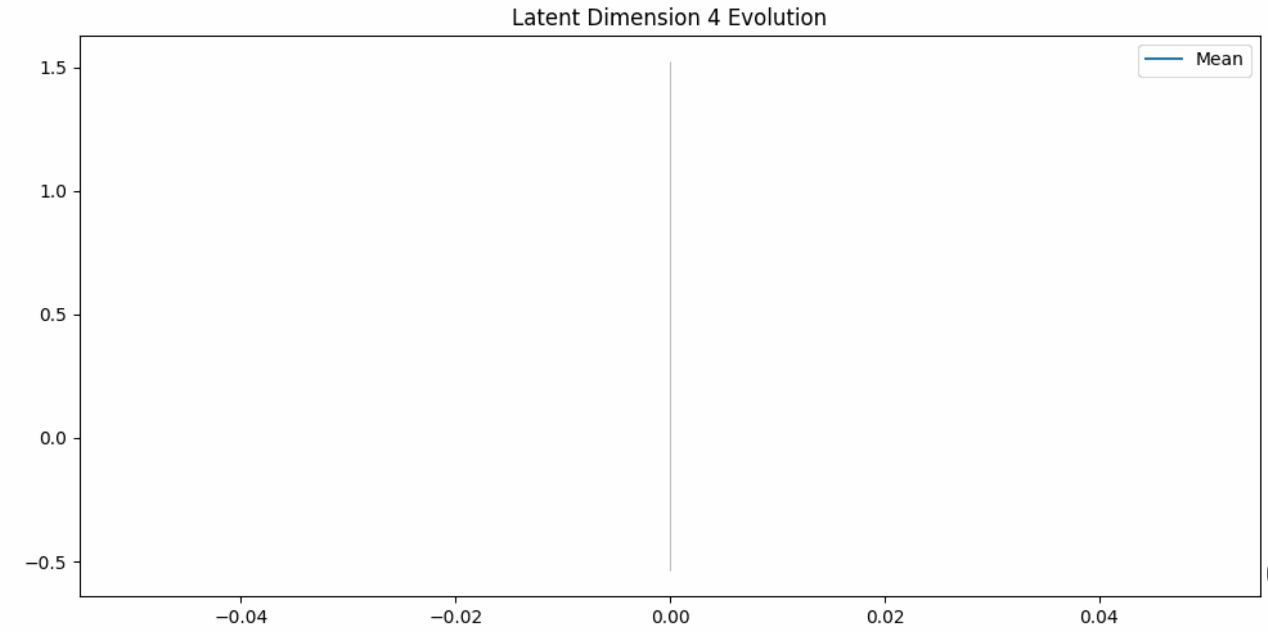
Aerosol dynamics: Case study

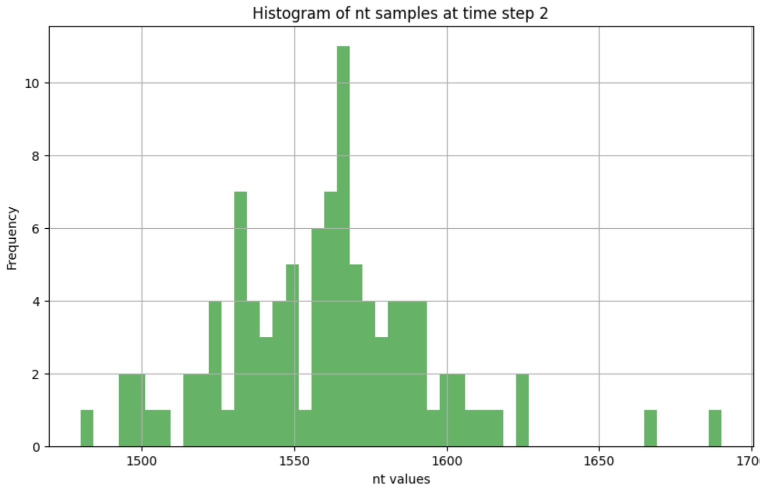
Match high-fidelity physical connections between *natural* clusters

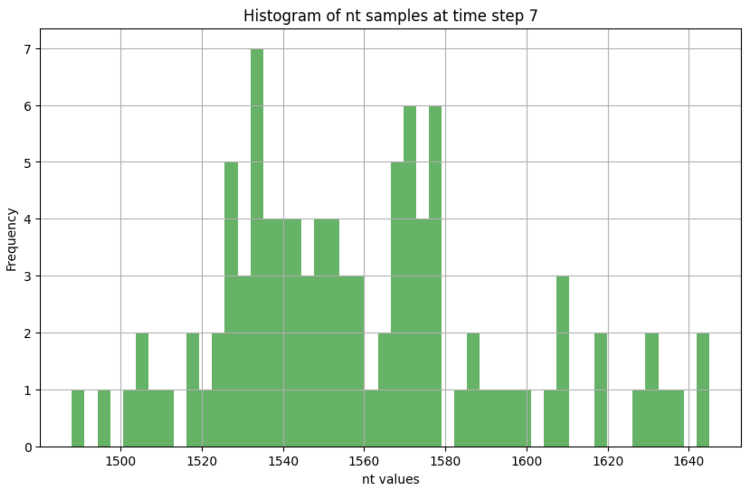
Capture physical processes in latent space



Open source data available at: https://databank.illinois.edu/datasets/IDB-2774261







CVAE enables efficient MC sampling with an understanding of the posterior *effective sample* size distribution

Conclusion

- Data fusion has a long history in Statistics and Machine Learning.
- State-of-the-art methods typically directly model or transform data/features into a similar space or continuous topology.
- Doing so may lose inherent structure in the data, particularly for categorical/ordinal variables whose meaning should not be changed.
- Careful handling of multimodal and mixed type (heterogenous) data can be critical when dealing with highly complex features such as in cyber-security and aerosol science.
- Heterogenous methods can be devised, particularly when utilized within popular methods such as VAEs and Diffusion Models (DMs).
- We study heterogenous VAEs with flexible latent structures and test them on reconstruction and anomaly detection of cybersecurity and aerosol representation data.

Pacific Northwest
NATIONAL LABORATORY

AIRES

Thank you