SAND2021-9359C

Sandia National Laboratories

GAMVT: **G**enerative **A**lgorithm for **M**ulti**V**ariate **T**imeseries

Joint Statistical Meetings (JSM)

August 8-12, 2021

Presenting Author: Jamie Thorpe, jthorpe@sandia.gov

SAND number

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

Motivation

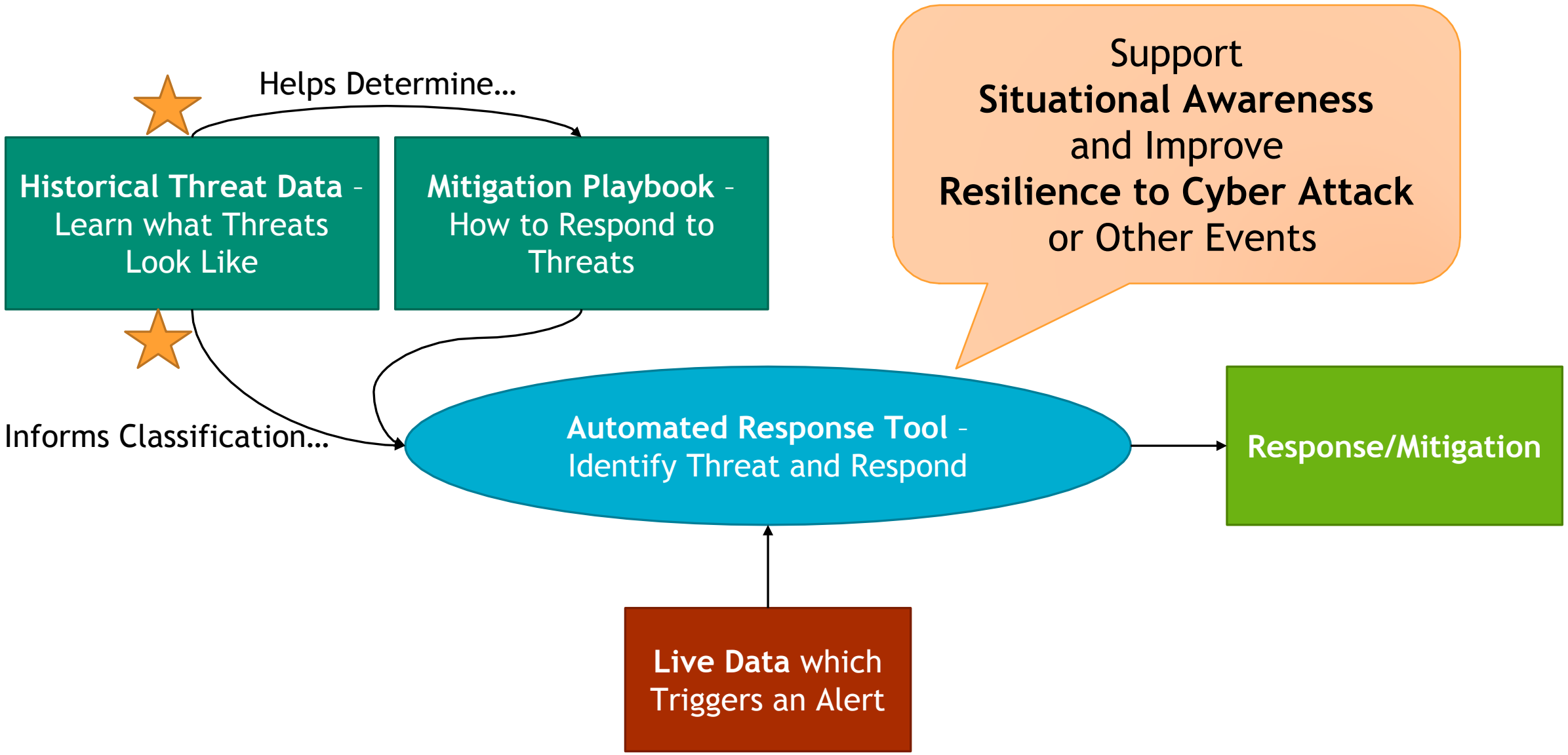
GAMVT

Use Case

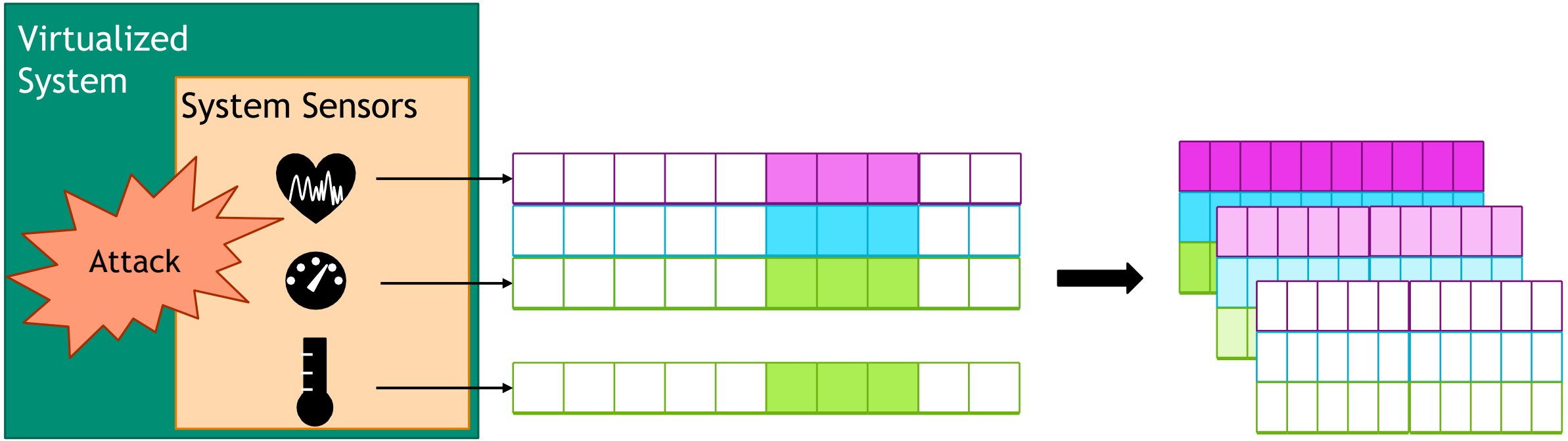
Results

Future Work

Motivation: Automated Response in a Cyber System



Motivation: System Emulation



Motivation: Generating Enough Data

Generate All Data Through Emulation?

- Too Slow

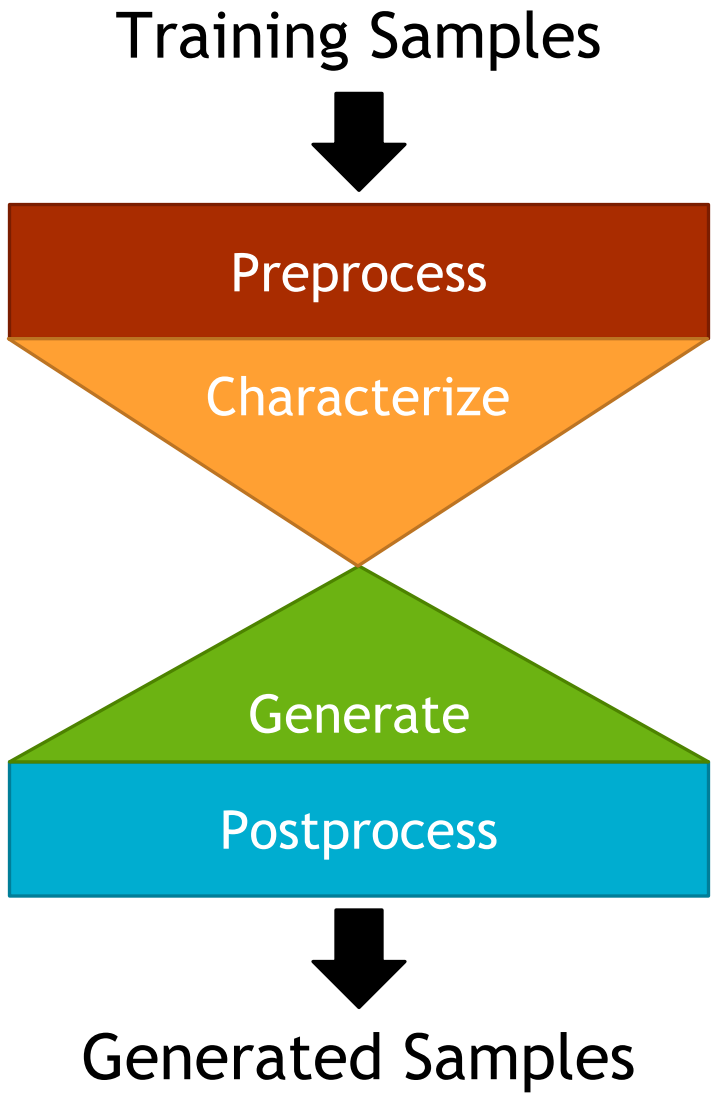
Perturb the Existing Data?

- Results in Samples with Inverted Timeline or Missing/Unordered Variables

Use Existing Deep Learning Methods for Data Generation?

- Image Data ≠ Multivariate Timeseries Data
- Large Training Dataset Requirements

GAMVT: Generative Algorithm for MultiVariate Timeseries



GAMVT: Preprocessing

Maintain Certain Data Artifacts

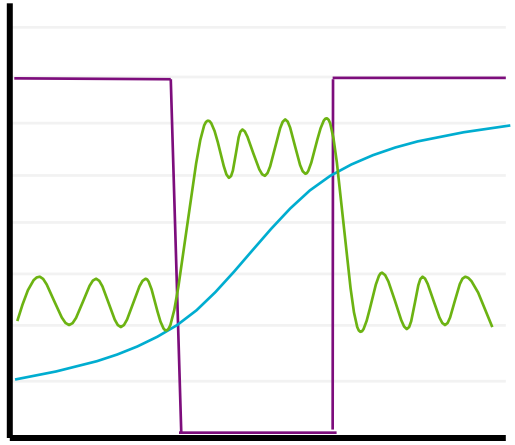
- Data Types
- Sample Shape

Simplify the Data

- Better Captures Trends
- Improves Results of First Step of Characterization: Timeseries Clustering

GAMVT: Characterization

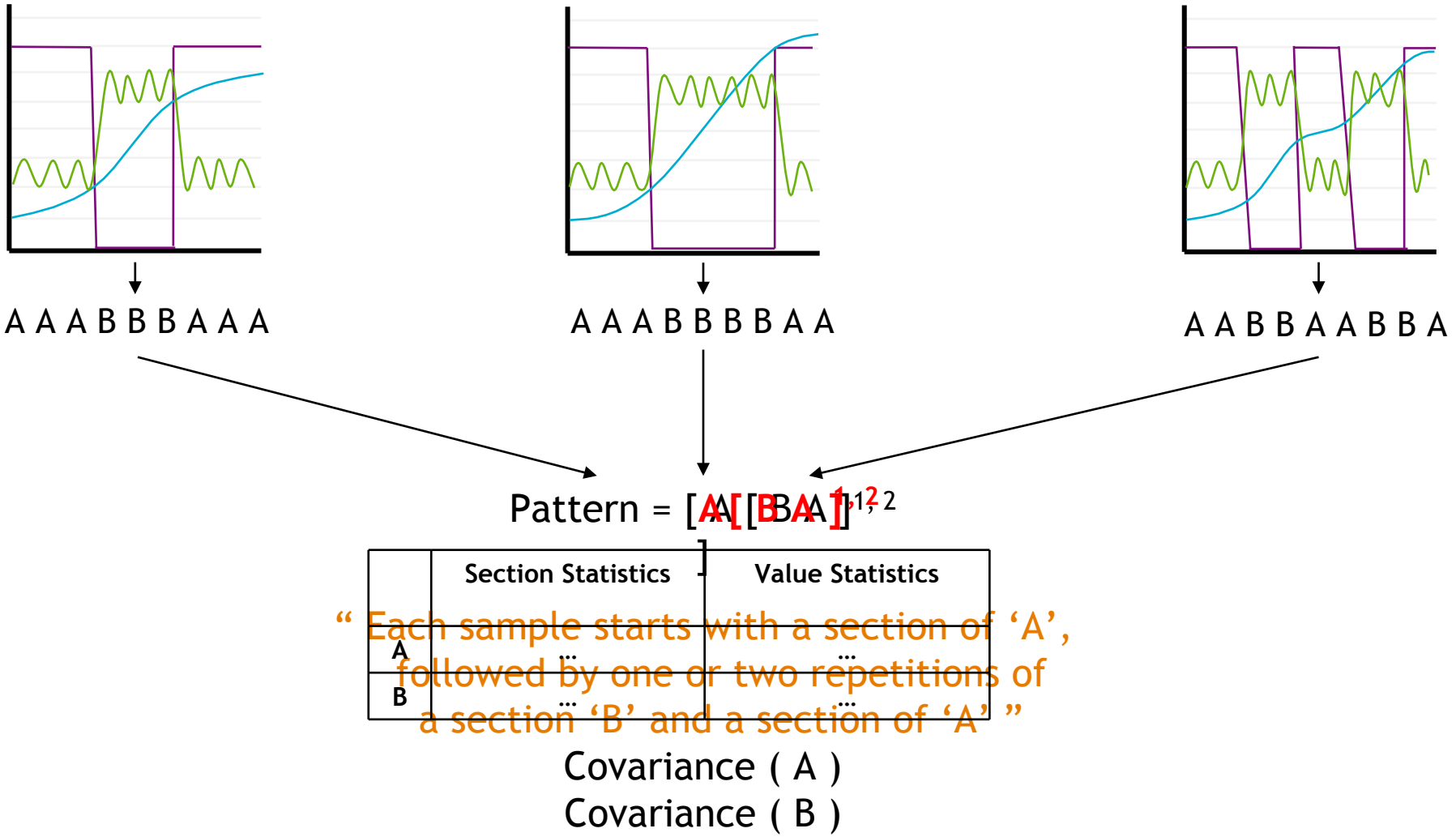
1. Timestep Clustering with TICC*
   - How do variables relate to each other in a given cluster?
   - Additional Subdivision of Samples by Cluster Label
2. Pattern Inference and Section Statistics
   - What order do sections appear in?
   - How long are those sections?
3. Value Statistics
   - What trends exist in the preprocessed data, per-class and per-cluster?
   - Mean, Minimum, Maximum for each variable



Call “Value Statistics” something more specific

* D. Hallac, S. Vare, S. Boyd, and J. Leskovec, “Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data”, 2017. [Online]. Available: arXiv.org, https://arxiv.org/abs/1706.03161.

Class Characterization



GAMVT: Generation

Class 3

9 timesteps

- Class 3 Pattern
- Class 3 Section Statistics
- Class 3 Value Statistics
- Covariance Matrices

Pattern = [A [B A]^{1,2}]
Section Stats = { A : [1, 3] ; B : [2, 4] }
Length = 9 timesteps

A B B B A A B B A

A B B B A A B B A

Value Stats: Mean, Range for Variables in Each Cluster in Class 3
Covariance: How Do Variables Relate to Each Other?

A	B	B	B	A	A	B	B	A
10	0	0	0	10	10	0	0	10
2.0	2.5	4.0	5.5	7.0	7.5	8.0	11.0	14.0
3.5	7.5	8.0	7.0	3.0	3.5	7.0	7.5	4.0

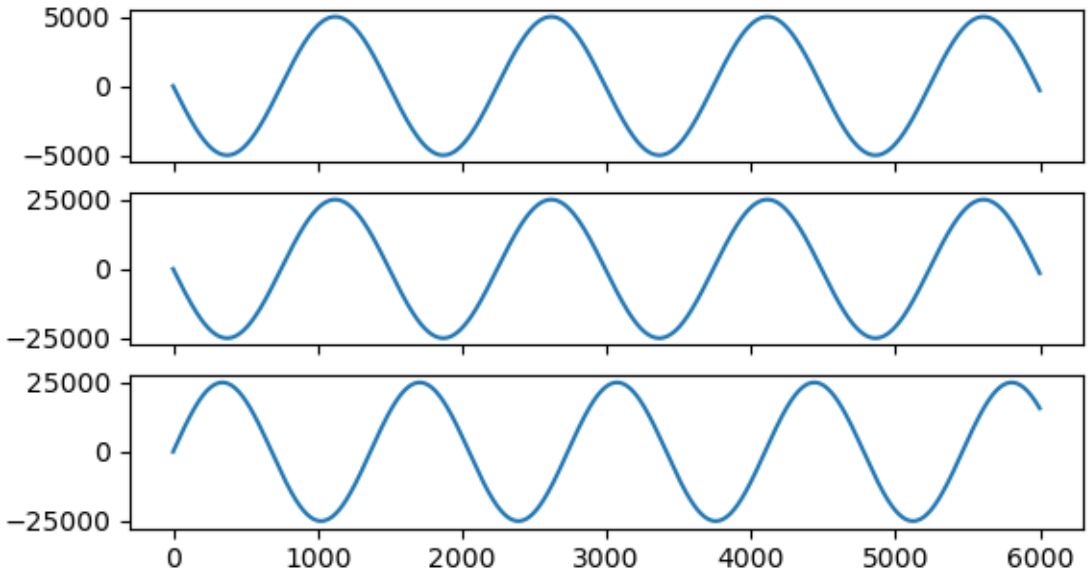
* Notional Example

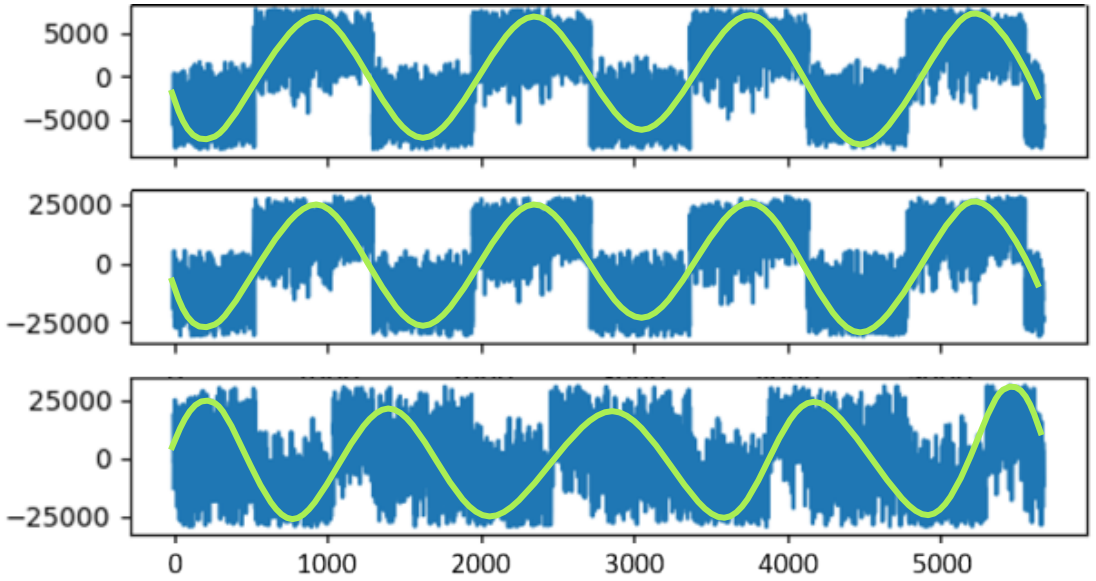
GAMVT: Postprocessing

Undo Preprocessing

- Undo Simplifications
- Reinforce Data Types

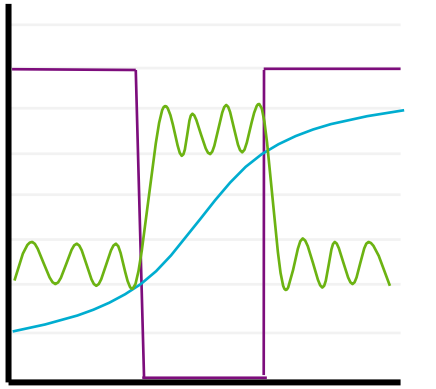
Smooth Generated Results



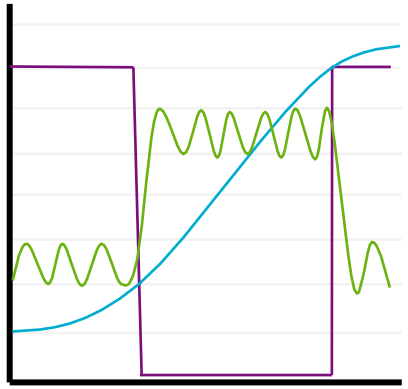


GAMVT-Generated Sample

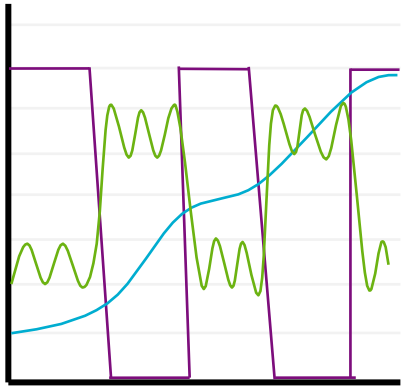
Training Set:



A A A B B B A A A

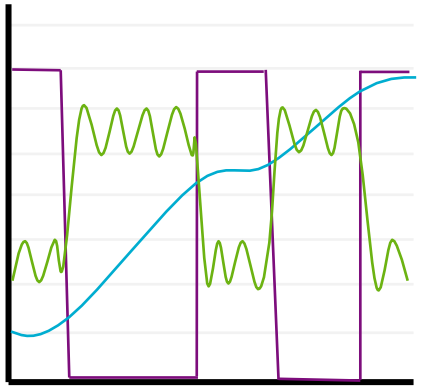


A A A B B B B A A



A A B B A A B B A

Generated Sample:



A B B B A A B B A

Use Case: Space-Cyber System

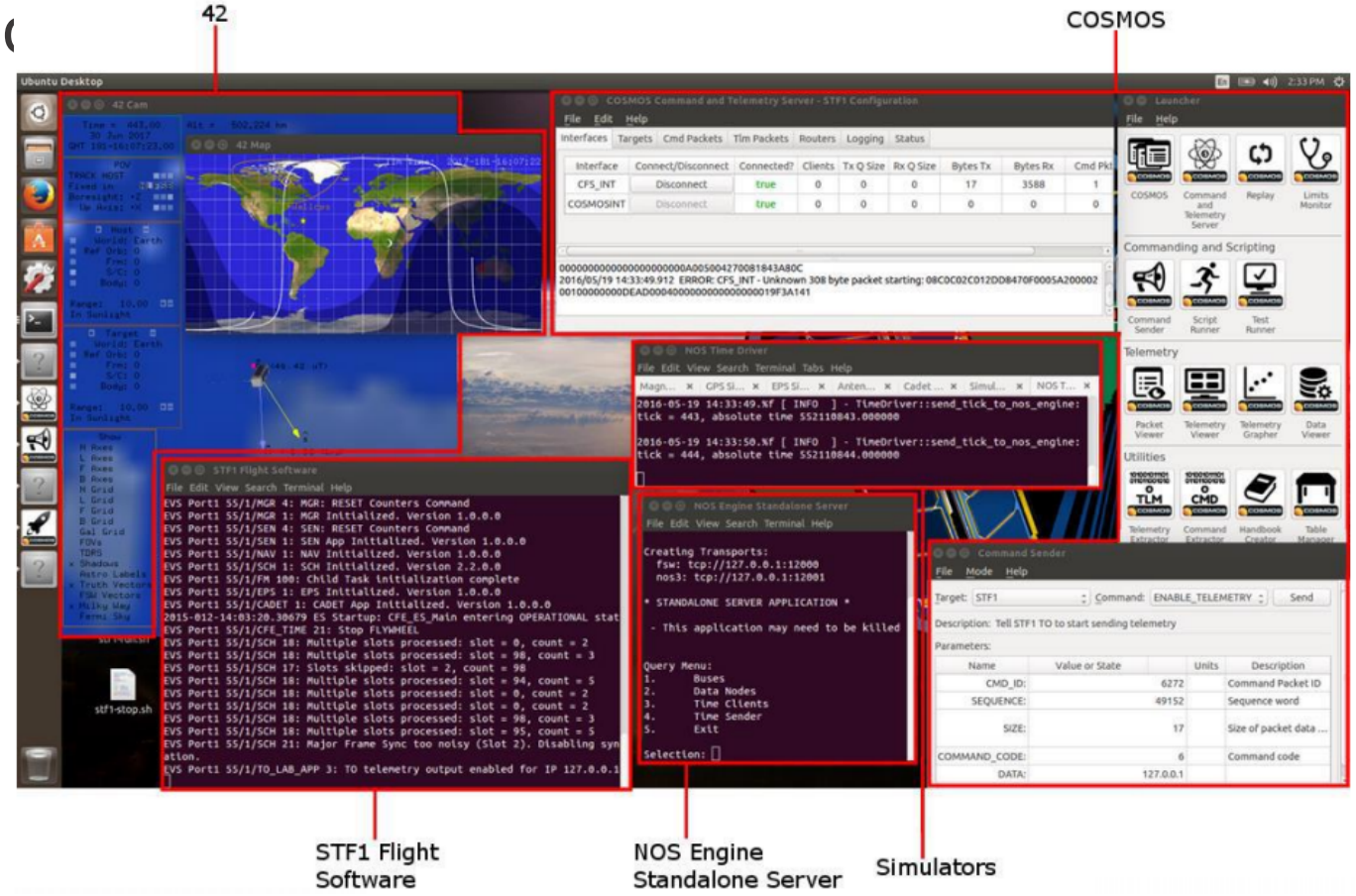
NASA-Developed satellite emulation – NOS3

Variables from System Sens

- GPS
  - X, Y, Z position components
  - X, Y, Z velocity components
- Camera
  - On/Off state
  - Memory usage

Sample Classes

- 1 baseline, 3 attacks
- 50 samples total



Visualized Data Sample

X, Y, Z Coordinate Position Data

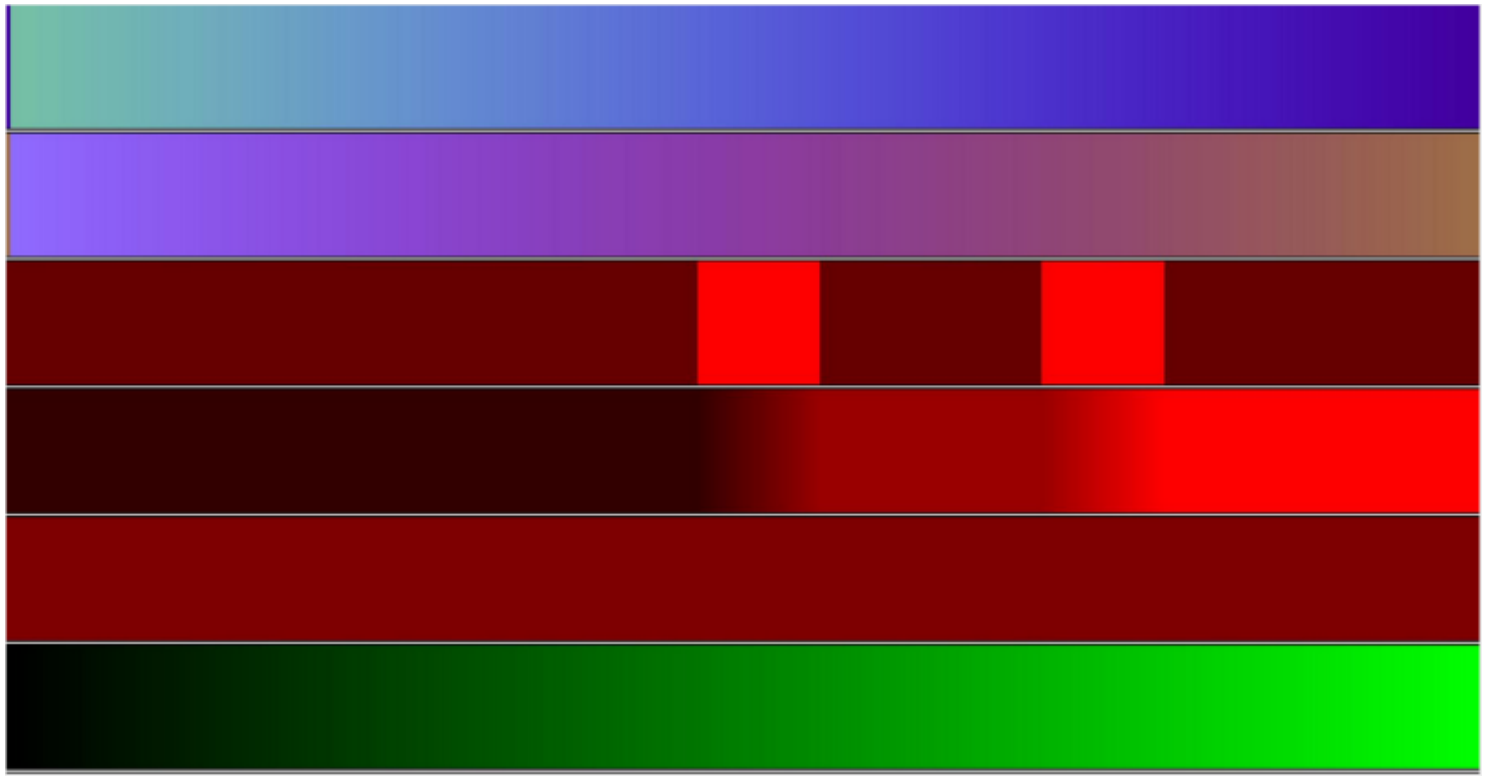
X, Y, Z Coordinate Velocity Data

Camera On/Off State

Camera Memory Usage

(Reserved for Future Usage)

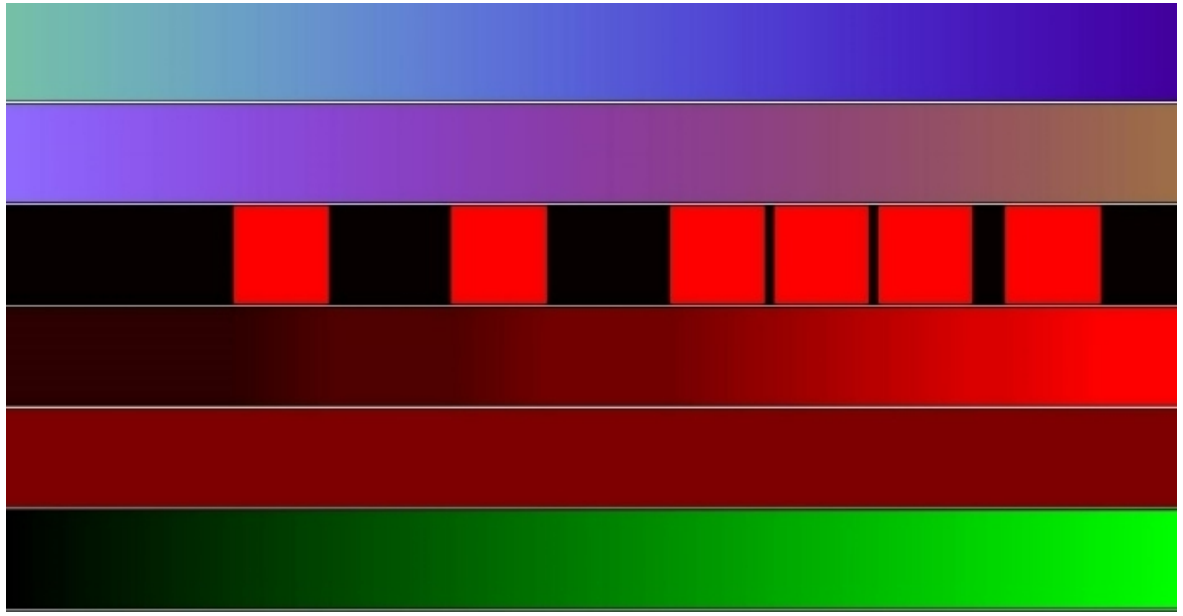
Timestamps



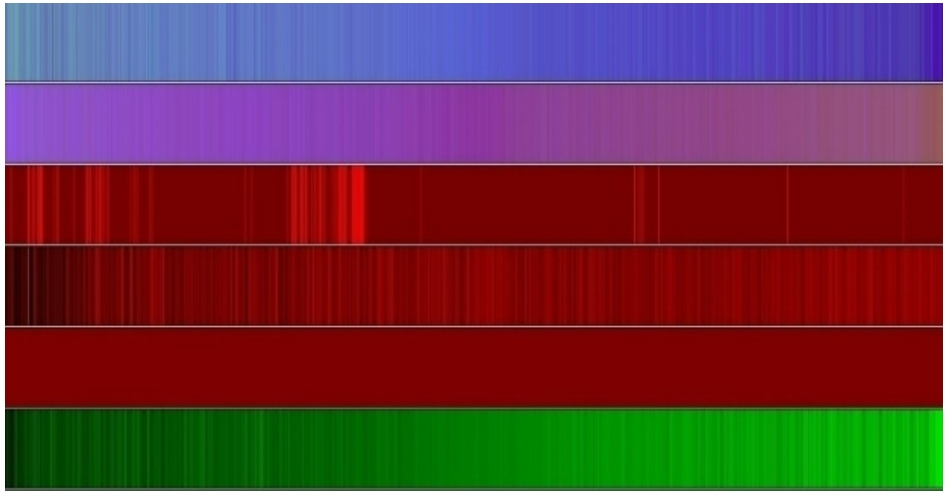
Generation Objective: To generate samples that differ from the real data, but which are similarly smooth and contain similar features to the real data.

Deep Learning Approaches

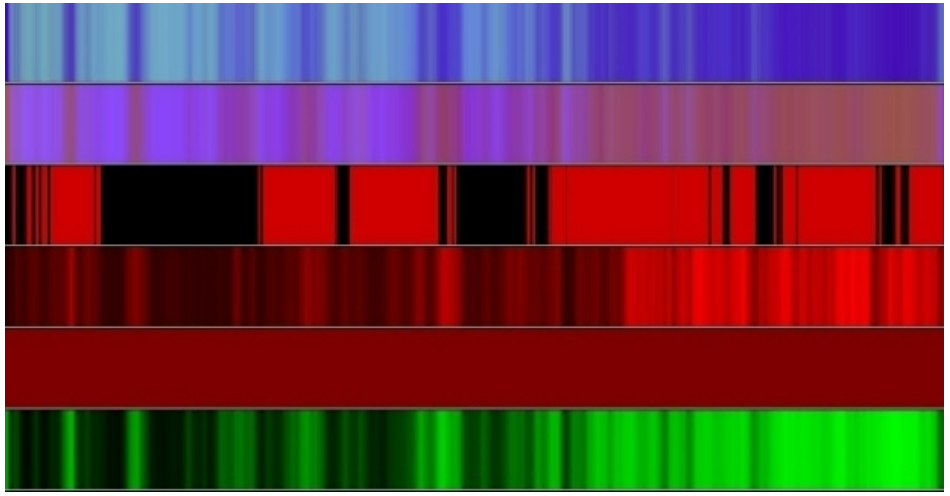
Real data



Generative Adversarial Network

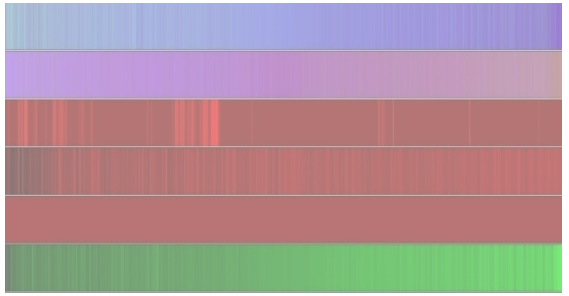


Variational Autoencoder

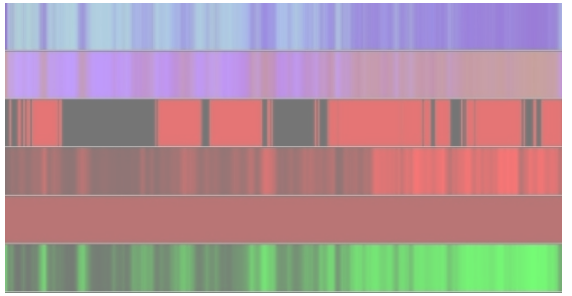


GAMVT Results

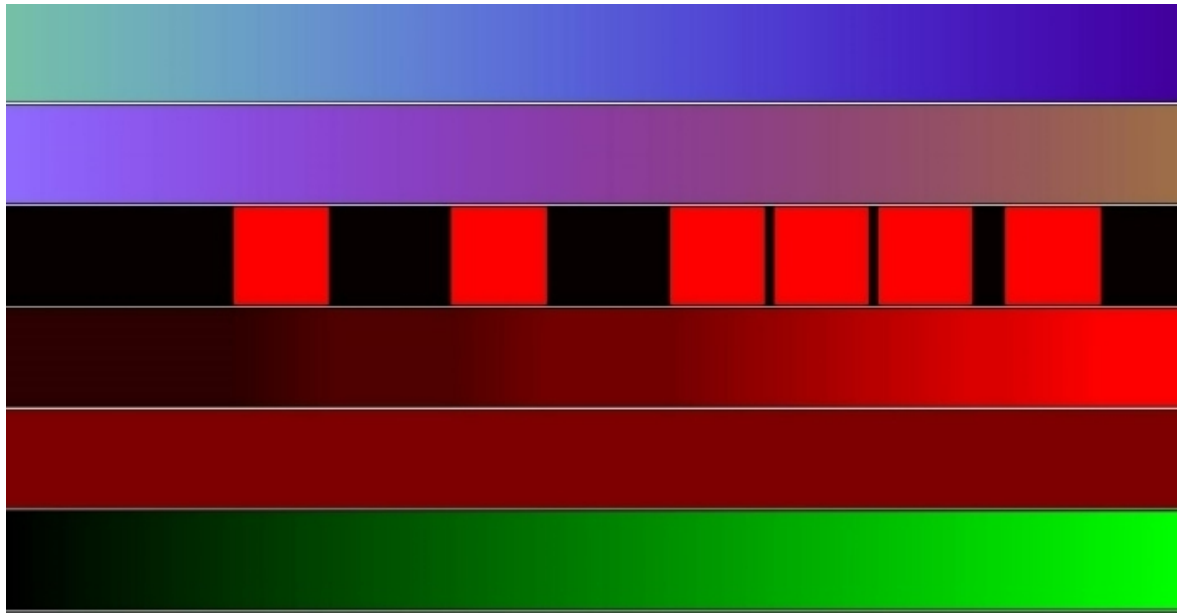
GAN



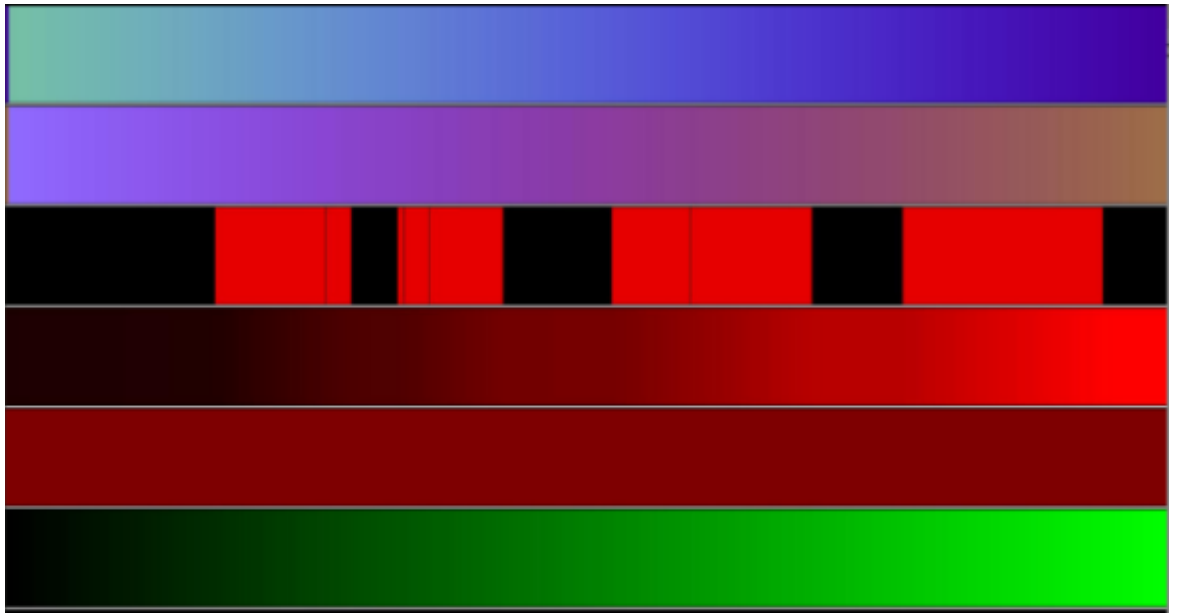
VAE



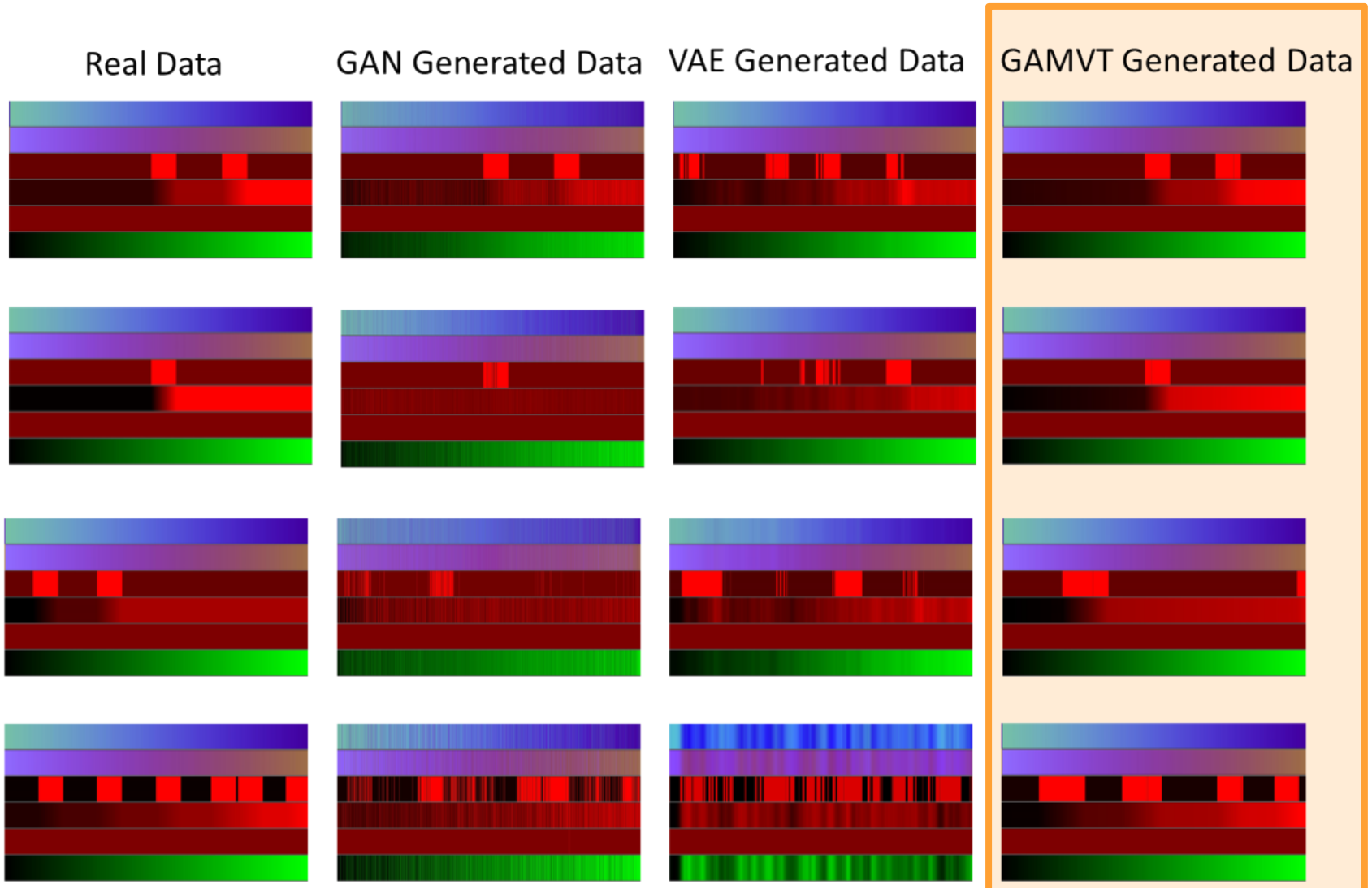
Real data



GAMVT



Results



	Real Data	GAN Generated Data	VAE Generated Data	GAMVT Generated Data
Quality	---	**0.7875**	**0.4775**	**0.7513**
Diversity	---	**23.2840**	**24.1815**	**1.4388**

Results from “Threat Data Generation for Space Systems”, Space Computing Conference, 2021.

Future Work

Test Generated Data with Downstream Process

Support More Complex Datasets

- Improve handling of oscillatory data
- Test and Develop with more complex relationships
- Expand support for a variety of generated distributions

Develop Methodology for Finding the Best Set of Pre/Post-Processing Functions

Comments and Questions to jthorpe@sandia.gov