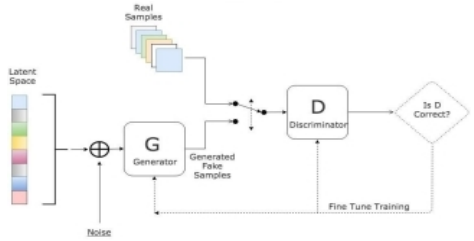
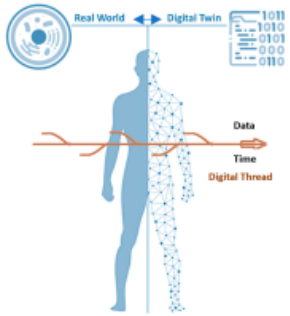
Synthetic Data Generation Using GenAI-Based WGAN





Uma Balakrishnan

Sandia National Laboratories

8th Annual Machine Learning/Deep Learning (MLDL) Workshop, 9-12 September, 2024.

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.

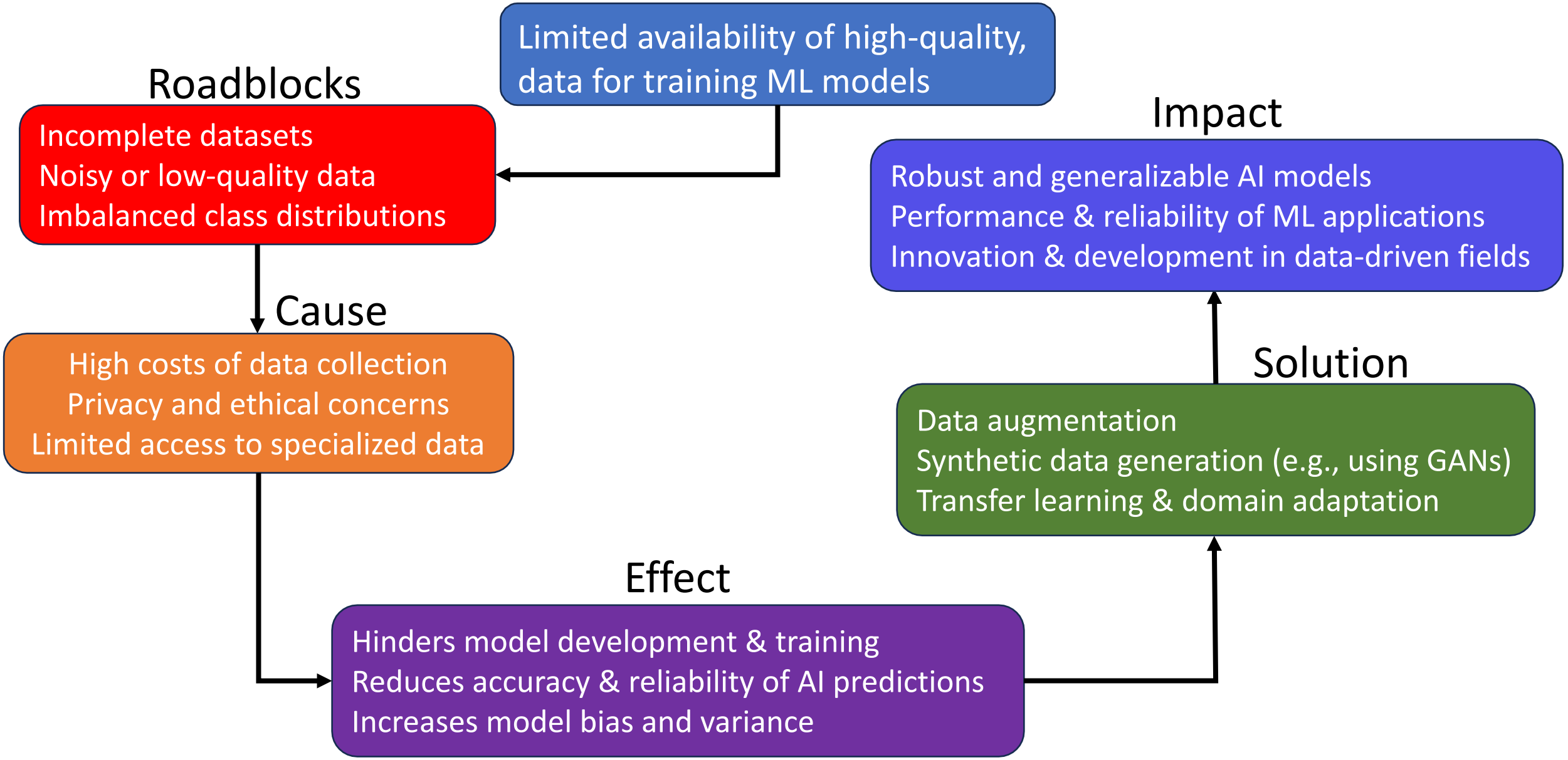
SAND2024-

Thanks to all Collaborators

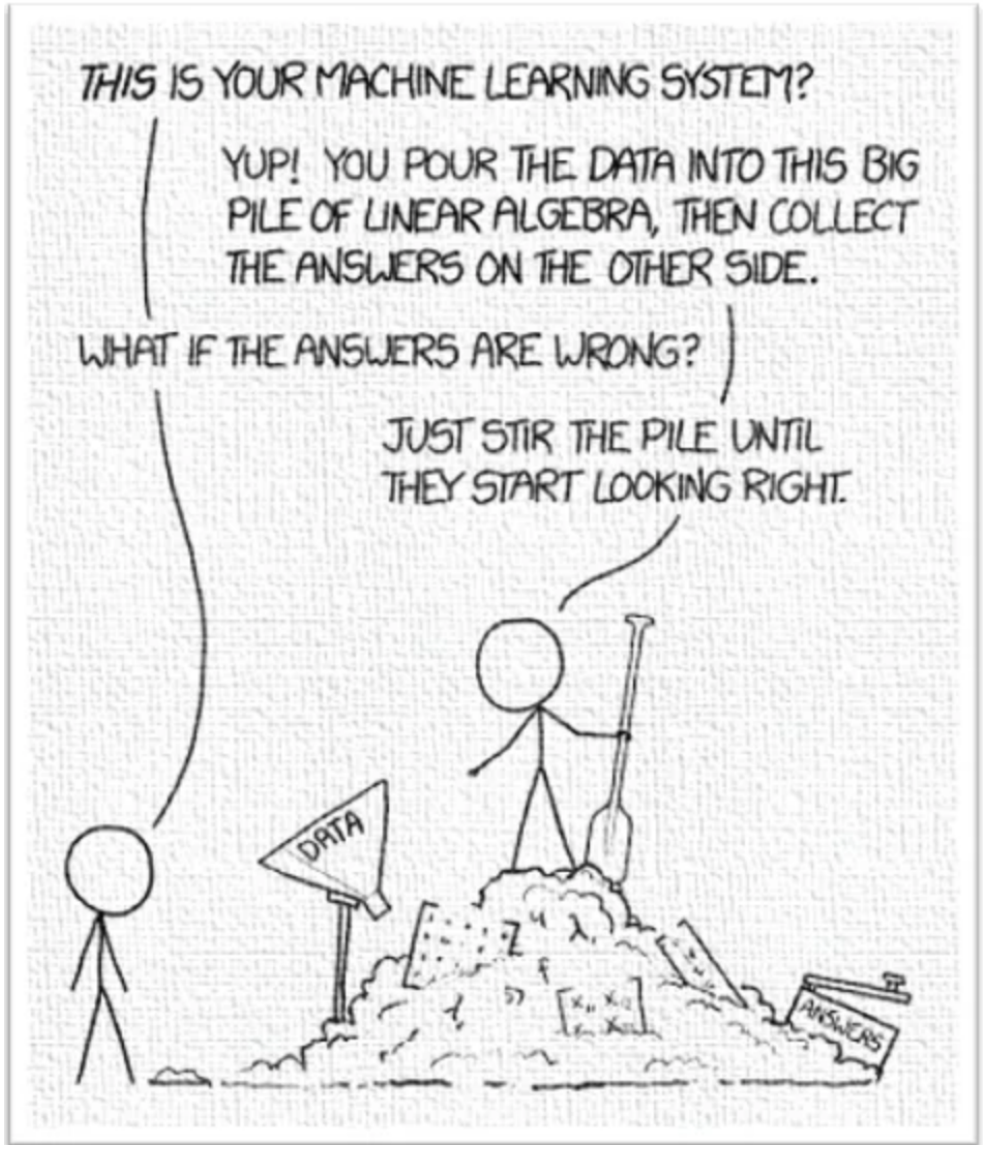
Bio-surveillance: Methun Kamruzzaman
Jorge Salinas
Kunal Poorey
Hemanth Kolla
Kenneth Sale

Climate Change: Jorge Salinas
Sagar Gautam
Changpeng Fan
Methun Kamruzzaman
Kunal Poorey
Umakant Mishra

Why we need to generate synthetic data? Data Scarcity



Impact of Synthetic Data Generation



?

Are we building the product right?

Verification

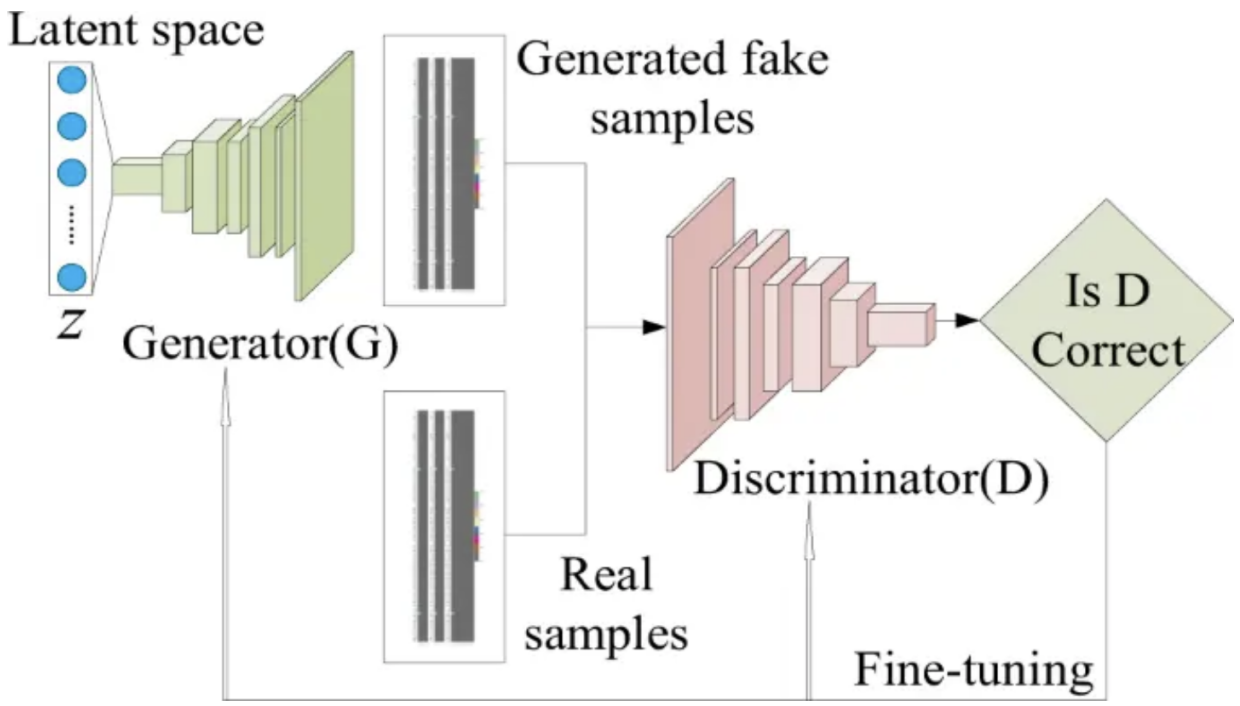
Are we building the right product?

Validation

Generative Adversarial Networks (GANs) – Data Generation

Loss function: cross-entropy loss

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[log D(\mathbf{x}^{(i)}) + log(1 - D(G(\mathbf{z}^{(i)}))) \right]$$



Vanishing Gradient: Logarithm in the loss function leads to vanishing gradients.

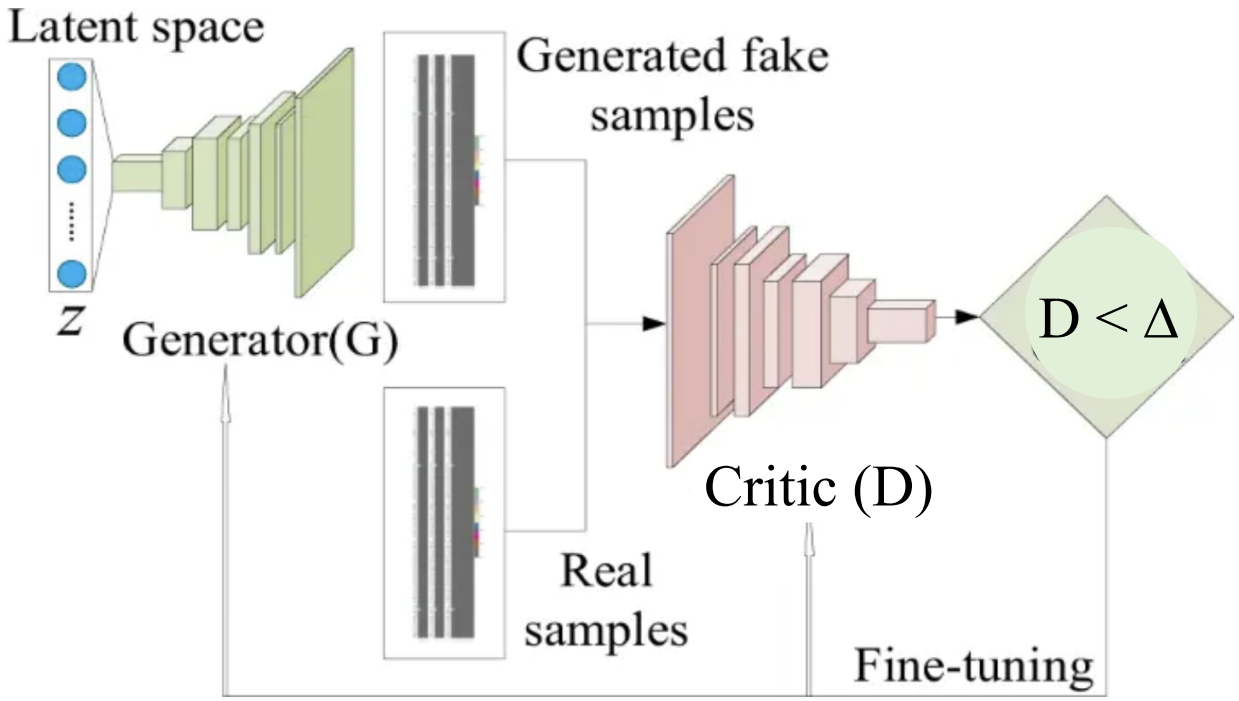
Mode Collapse: Generator produces a limited variety of samples, often focusing on a single mode of the data distribution.

Training instability: Difficulty in achieving a stable equilibrium between the generator and discriminator, leading to oscillations in the loss function and poor quality of generated samples.

Wasserstein Generative Adversarial Networks (WGANs)

Loss function: Wasserstein distance
(Earth movers distance)

$$\nabla_w \frac{1}{m} \sum_{i=1}^{m} \left[f(\mathbf{x}^{(i)}) - f(G(\mathbf{z}^{(i)})) \right]$$



Arjovsky, M., Chintala, S., and Bottou, L., "Wasserstein Generative Adversarial Networks". International Conference on Machine Learning, PMLR (2017) 214–223.

Wet Clipping: The Lipschitz constraint on the critic (discriminator). The weights of the critic are clipped within a specified range.

No Log in Loss: No logarithm in the loss function helps avoiding vanishing gradients.

Stable Training: More stable training and improved convergence.

Applications

Material Science: New materials for nuclear weapons and defense systems.

Cybersecurity: Cybersecurity systems to recognize and respond to potential cyber threats.

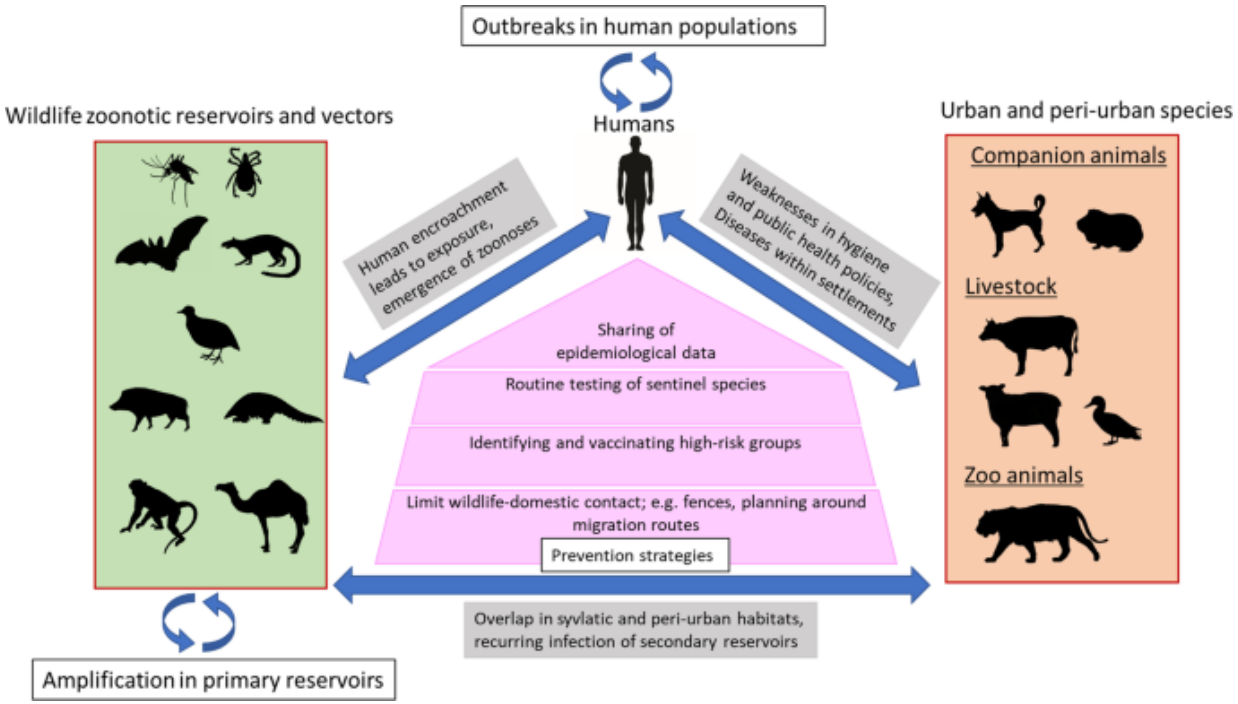
Intelligence and Surveillance: Train models to detect anomalies in surveillance data;
Augment satellite imagery and other surveillance data detect potential threats.

Healthcare: Early detection of biothreats;
Privacy-preserving data for patient records.

Climate Change: Filling gaps in soil organic carbon data;
Better climate prediction models.

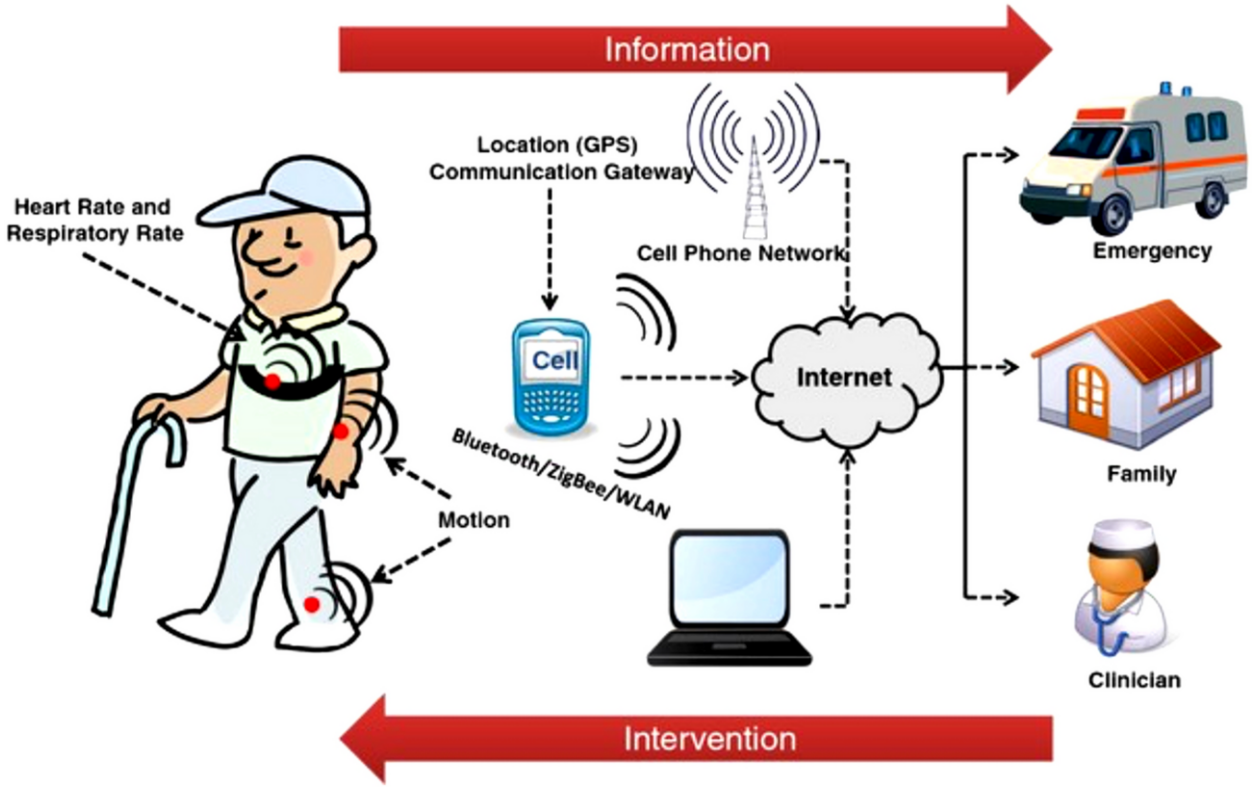
Finance: Fraud detection;
Risk assessment

Early Detection of Biothreats



- Measure for surveillance of population at different resolutions
- Measure to quantify risk or threat
- Algorithms that can raise an alarm when anomaly arises

Bio-surveillance using Wearable Sensors



- Wearable sensors give us an opportunity to identify health problems at an early stage, allowing for timely intervention.

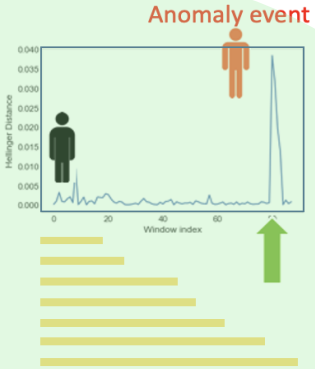
- Transferable technology for monitoring other threats.

Bridging the Gap

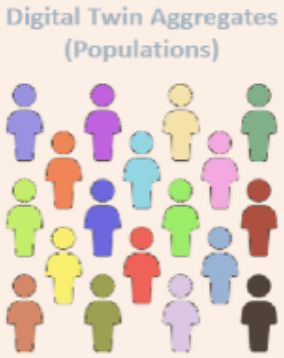
- Scarcity of real world population data (120 real users data from Fitbit / Garmin / Apple watch).
  (Most work has been carried out with limited datasets)
  - Collecting more real-time data are constrained by cost, logistics, privacy reasons, etc.
  - GenAI to augment synthetic population (Digital Twins) for method development.
  - Generated data needs to be verified by comparing its statistical signature with real data.
- Algorithms for anomaly detection based on the field of application.
  (Challenges exist in evaluating true and false positives)
  - Fourth order standardized moment (co-kurtosis) based algorithm.
  - Applicable to wide range of datasets from univariate to multivariate features.
  - Optimize the algorithm for the large population.

Three Components

GenAI-based Digital Twin of real user wearable data

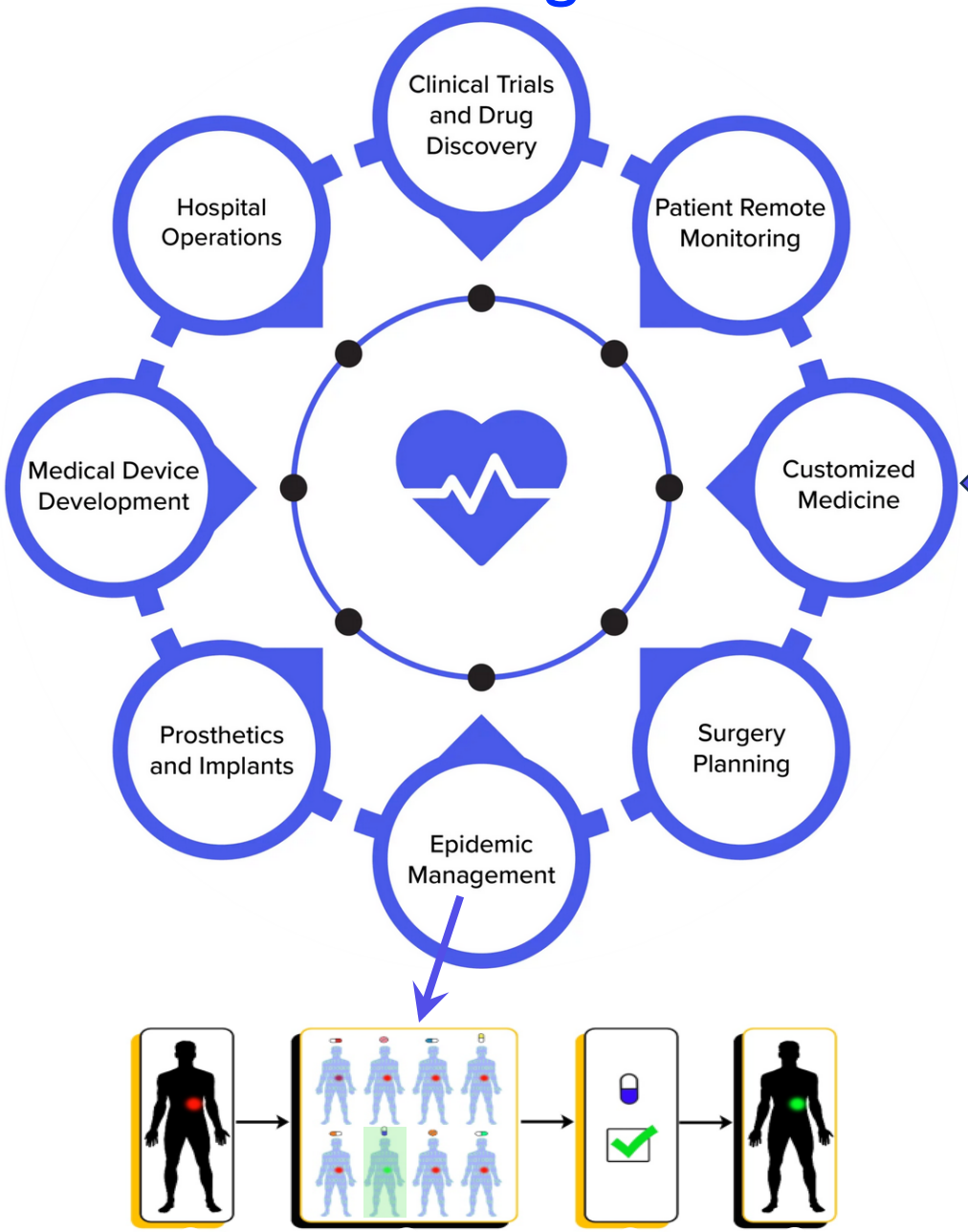


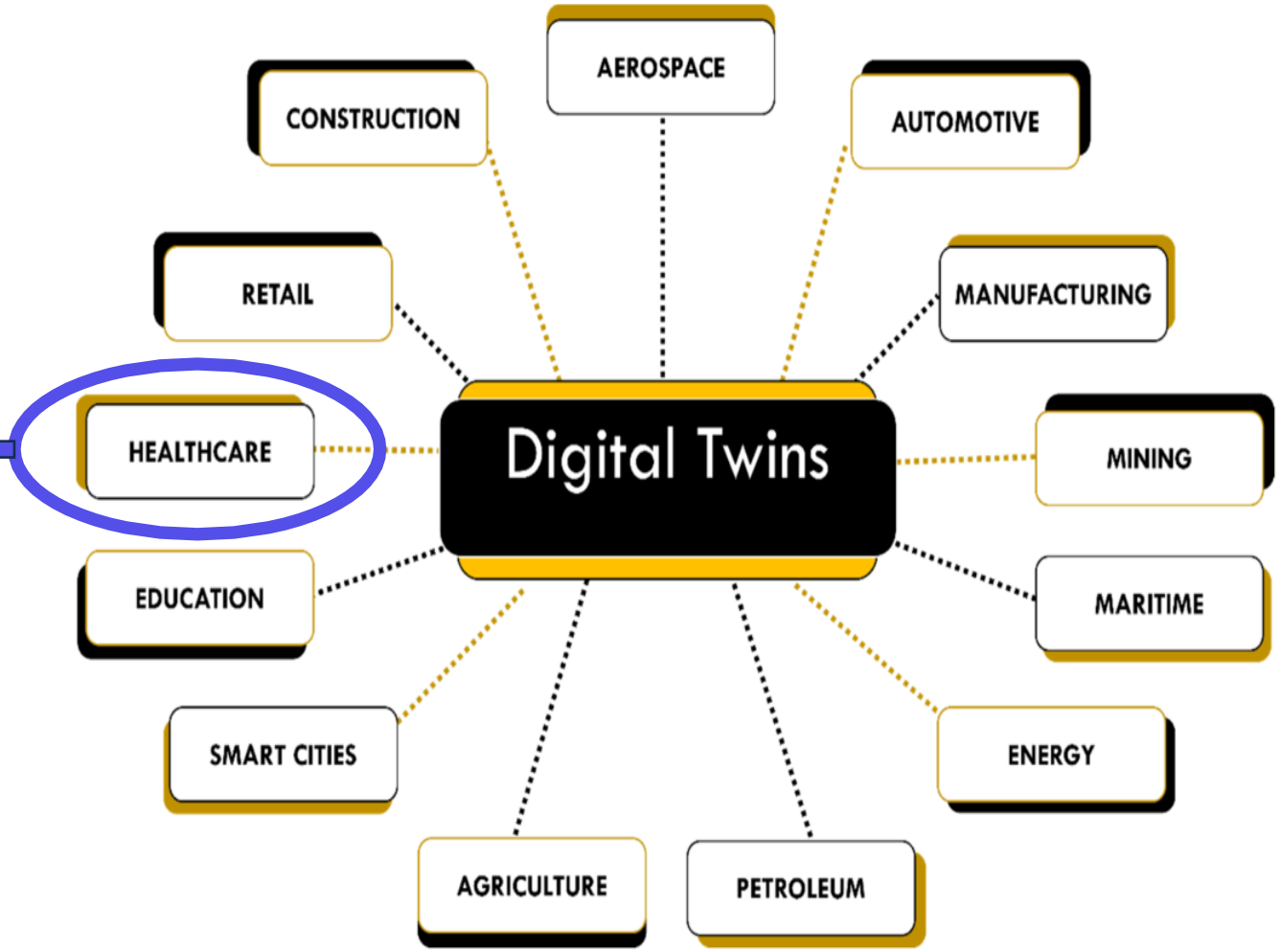
Fourth-order based anomaly detection



Anomalies in population generated by DTs

GenAI based Digital Twins



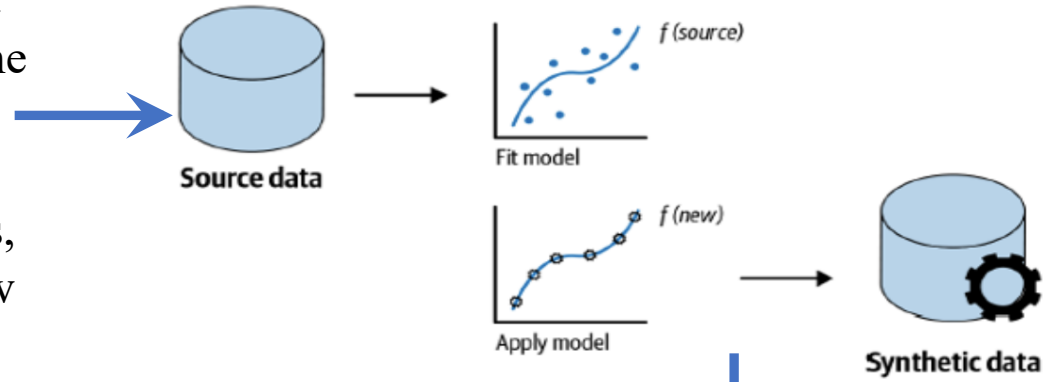


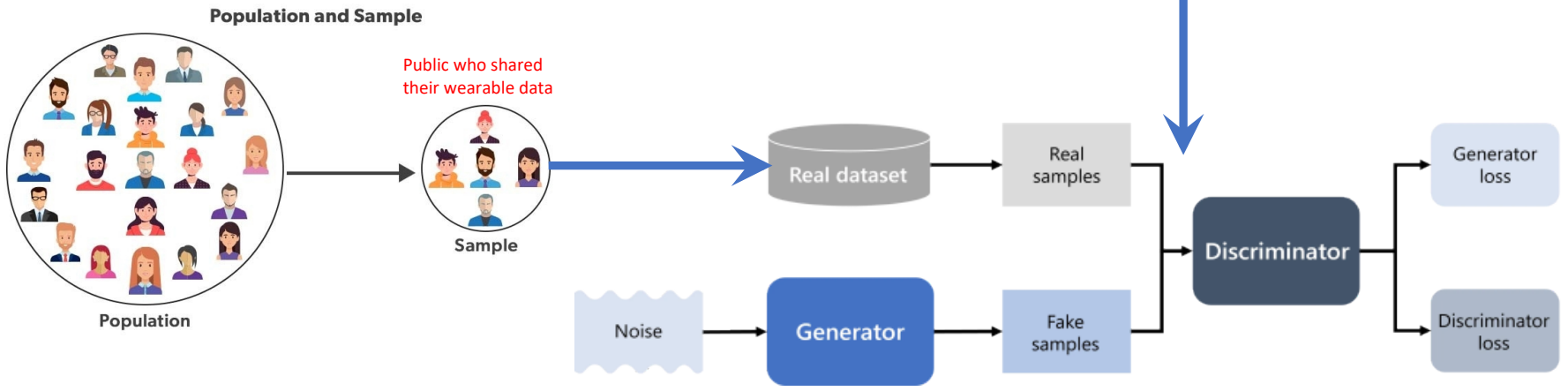
A digital twin is a virtual replica of a physical object, system, or process used to simulate, analyze, and optimize its real-world counterpart.

Synthetic Time-Series Data Generation

- Generating synthetic data (**Digital Twins**) mimicking original data from deep learning algorithms is an inexpensive way to increase the cohort size.

- AI algorithm should be capable of automatically detecting patterns, structures, correlations, etc. within real data, and then generate new dataset with the same patterns.



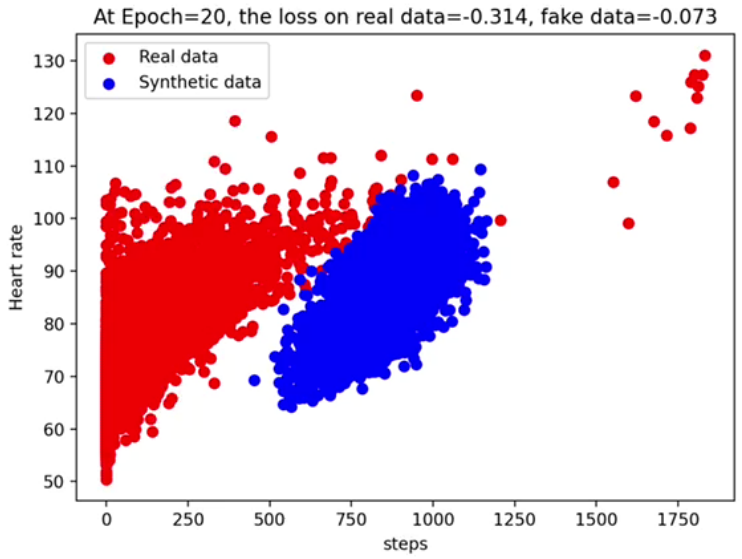


Wasserstein Generative Adversarial Network (WGAN)

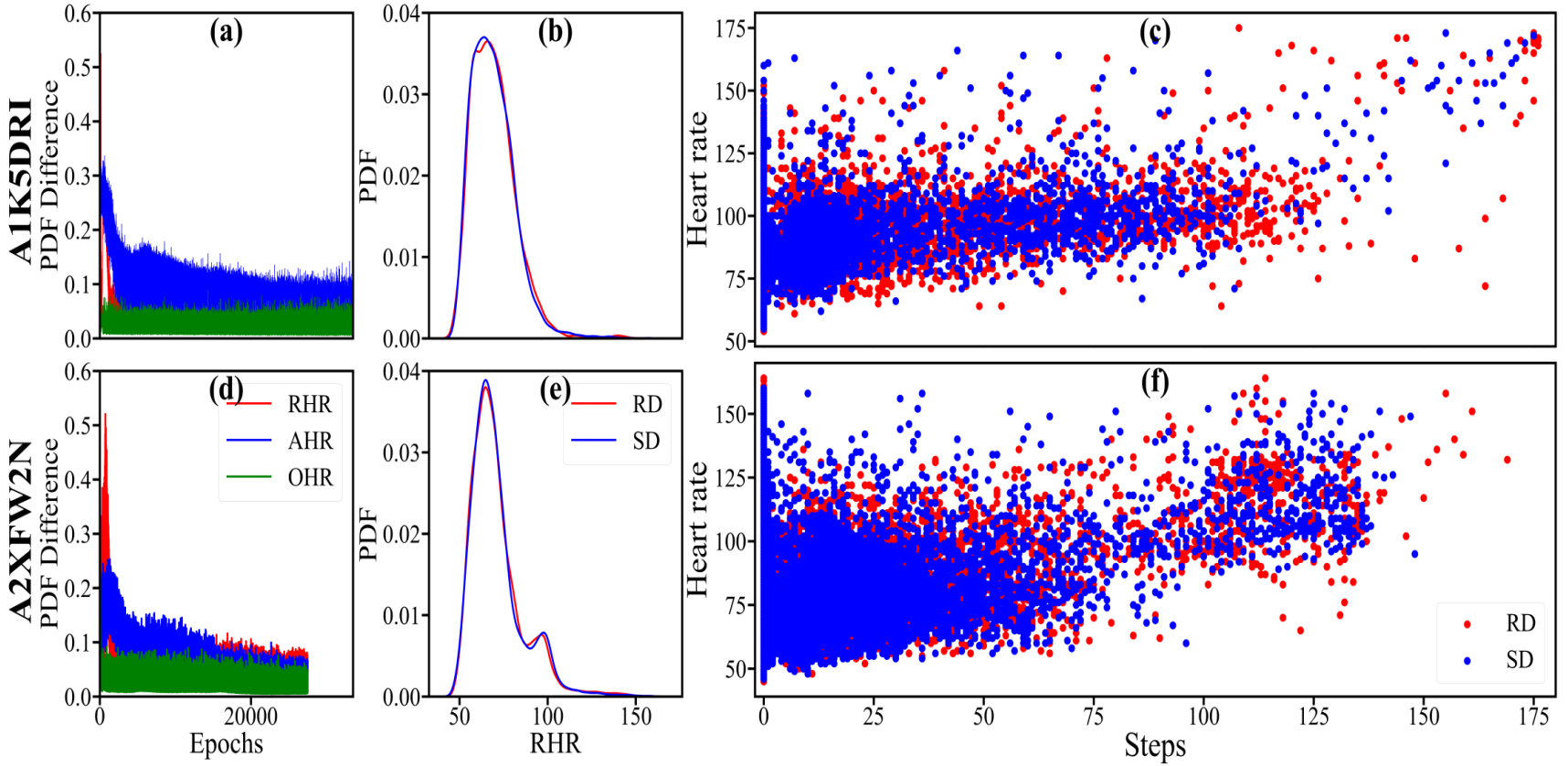
Steps for Data Generation

1. Load the original time series wearable data into a pandas DataFrame.
2. Split data into training and testing sets.
3. Use statistical techniques to learn underlying patterns and characteristics of the training data.
4. Generate synthetic data that mimics the original data by sampling from the learned statistical model.
5. Validate the synthetic data by comparing it with the testing data using statistical metrics such as MAE or RMSE.

Synthetic Twin of Real-Time Wearable Data



Update generator and discriminator till synthetic data matches the real data (satisfies Wasserstein distance).



Digital twin of two users are shown here. First column ((a) & (d)) shows the difference in PDF against the epoch and middle column ((b) & (e)) shows the density plot of RHR and last column ((c) & (f)) shows both the real and synthetic data of each user.

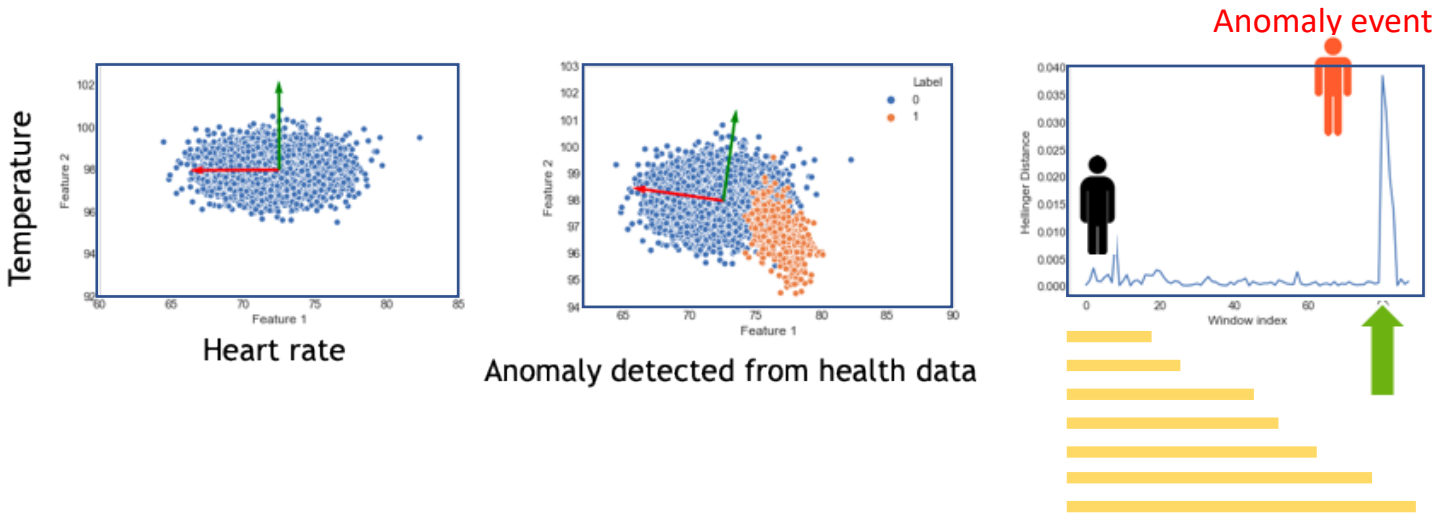
Anomaly Detection using Fourth-Order Moment

- Real time data from the wearable device is decomposed into several spatial sub-domains and time steps.
- Compute Hellinger distance with each increment timeline to detect anomaly.

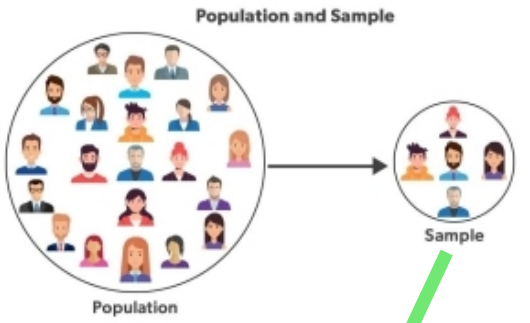
Tensor decomposition

$$\mathcal{X} \approx \mathbf{c}_1 \mathbf{b}_1 \mathbf{a}_1 + \ldots + \mathbf{c}_R \mathbf{b}_R \mathbf{a}_R$$

Hellinger distance
(Measures change in orientation of singular vectors weighted by singular values)

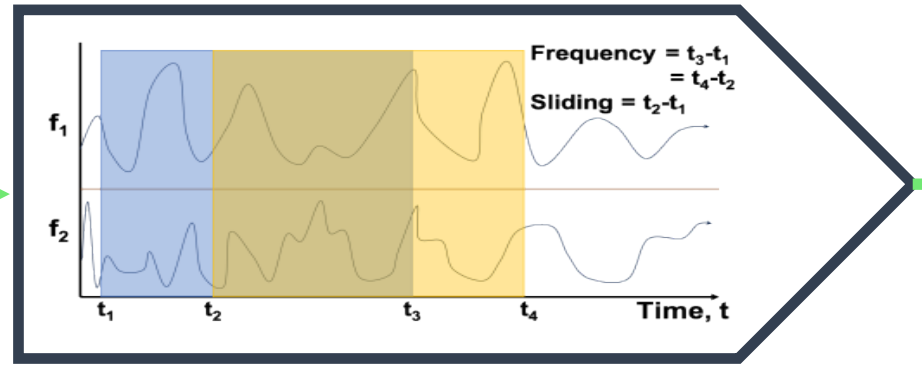


Anomaly Detection of Real-Time User



Real data

Data



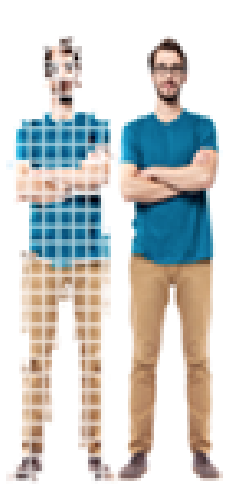
Synthetic data

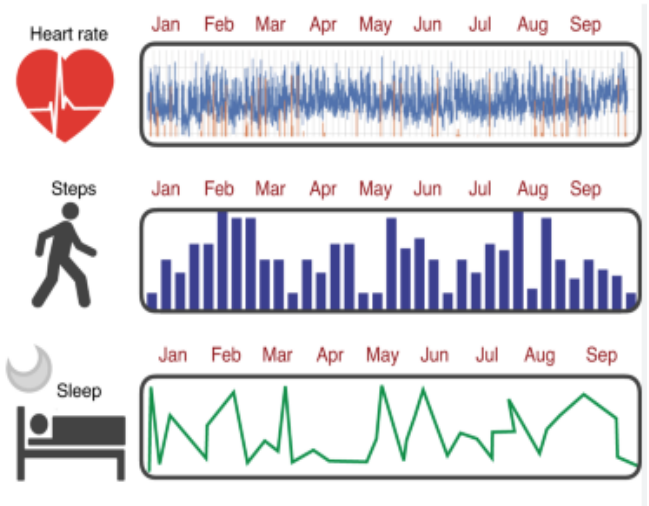
WGAN

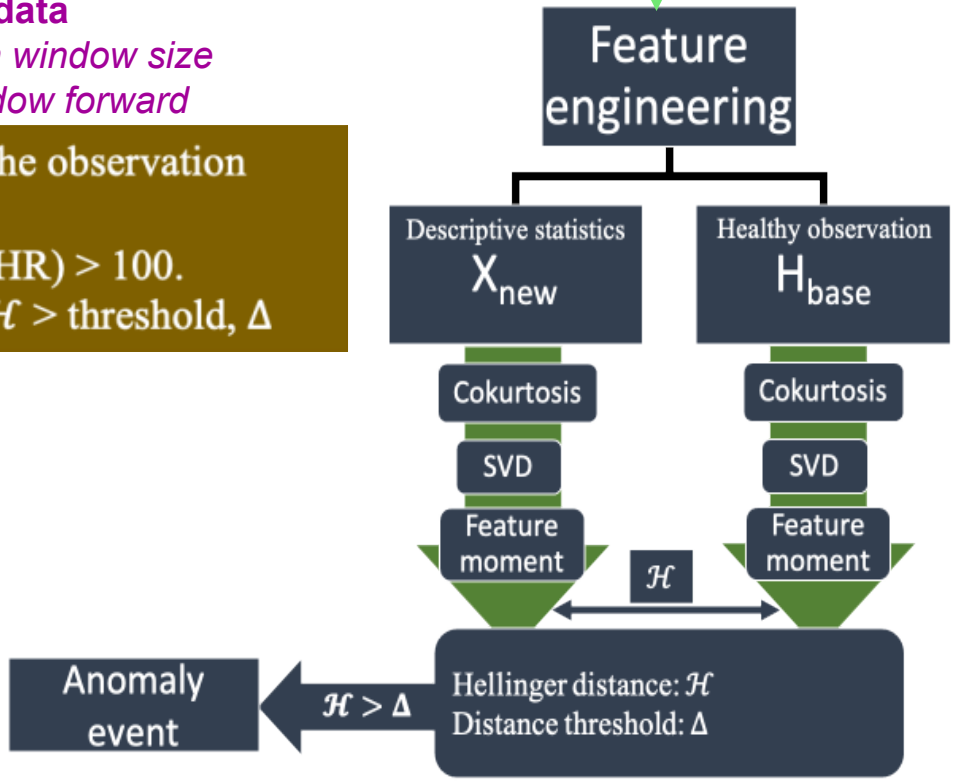
Processing streaming data
Frequency: *Observation window size*
Sliding: *Moving the window forward*

Sick: Given a patient $\mathcal{P}$ with the observation period from T_1 to T_2,

1. Resting heartrate (RHR) > 100.
2. Hellinger distance, $\mathcal{H}$ > threshold, Δ







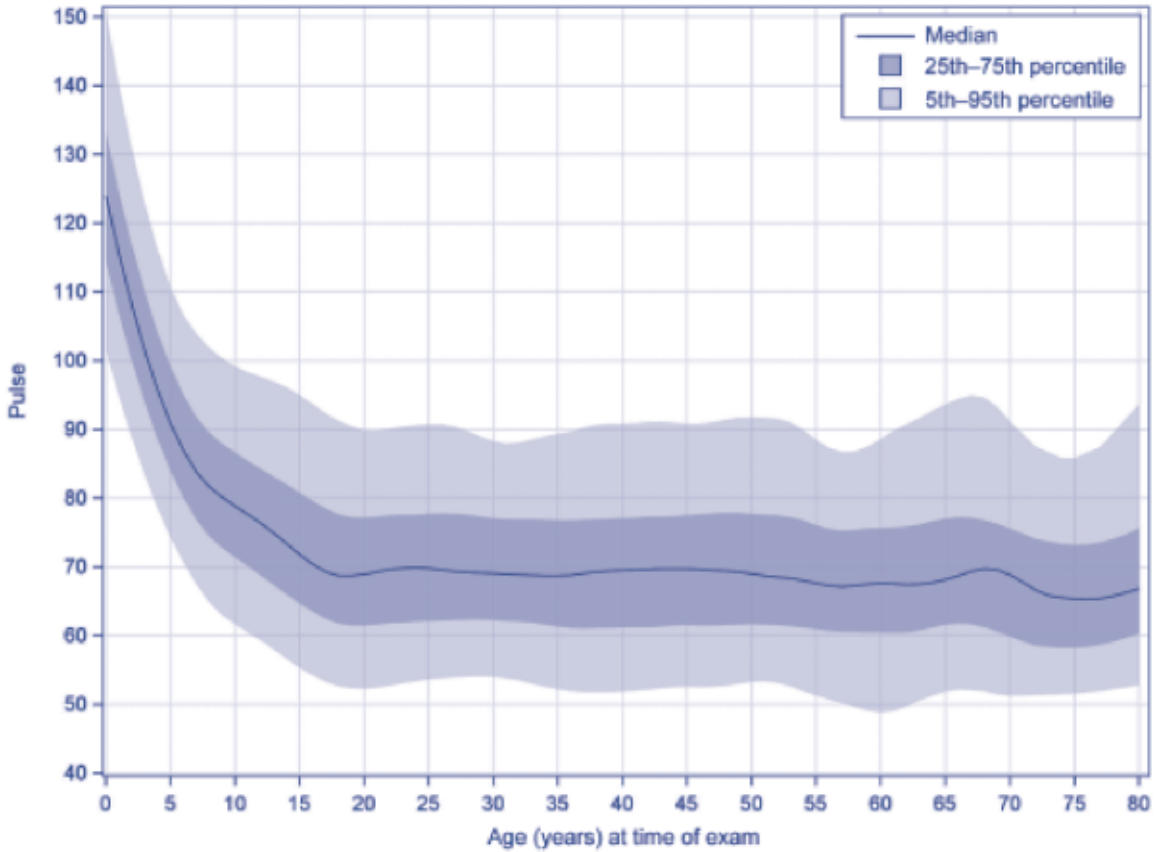
Performance Metric

TP – True Positive (RHR $> \nabla$ & $\mathcal{H} > \Delta$)
TN – True Negative (RHR $\leq \nabla$ & $\mathcal{H} \leq \Delta$)
FP – False Positive (RHR $\leq \nabla$ & $\mathcal{H} > \Delta$)
FN – False Negative (RHR $> \nabla$ & $\mathcal{H} \leq \Delta$)

$$Precision = \frac{TP}{TP + FP}; \; Recall = \frac{TP}{TP + FN}$$

;

$$\mathbf{FNR} = \frac{FP}{FP + TN}$$



CDC: Centers for Disease Control and Prevention
(https://www.cdc.gov/nchs/data/nhsr/nhsr041.pdf)

Population level		
Sick (RHR>100)	TP	FN
	FP	TN

F_1 Score	FNR
$\frac{2 * Precision * Recall}{Precision + Recall}$	$\frac{FN}{TP + FN}$

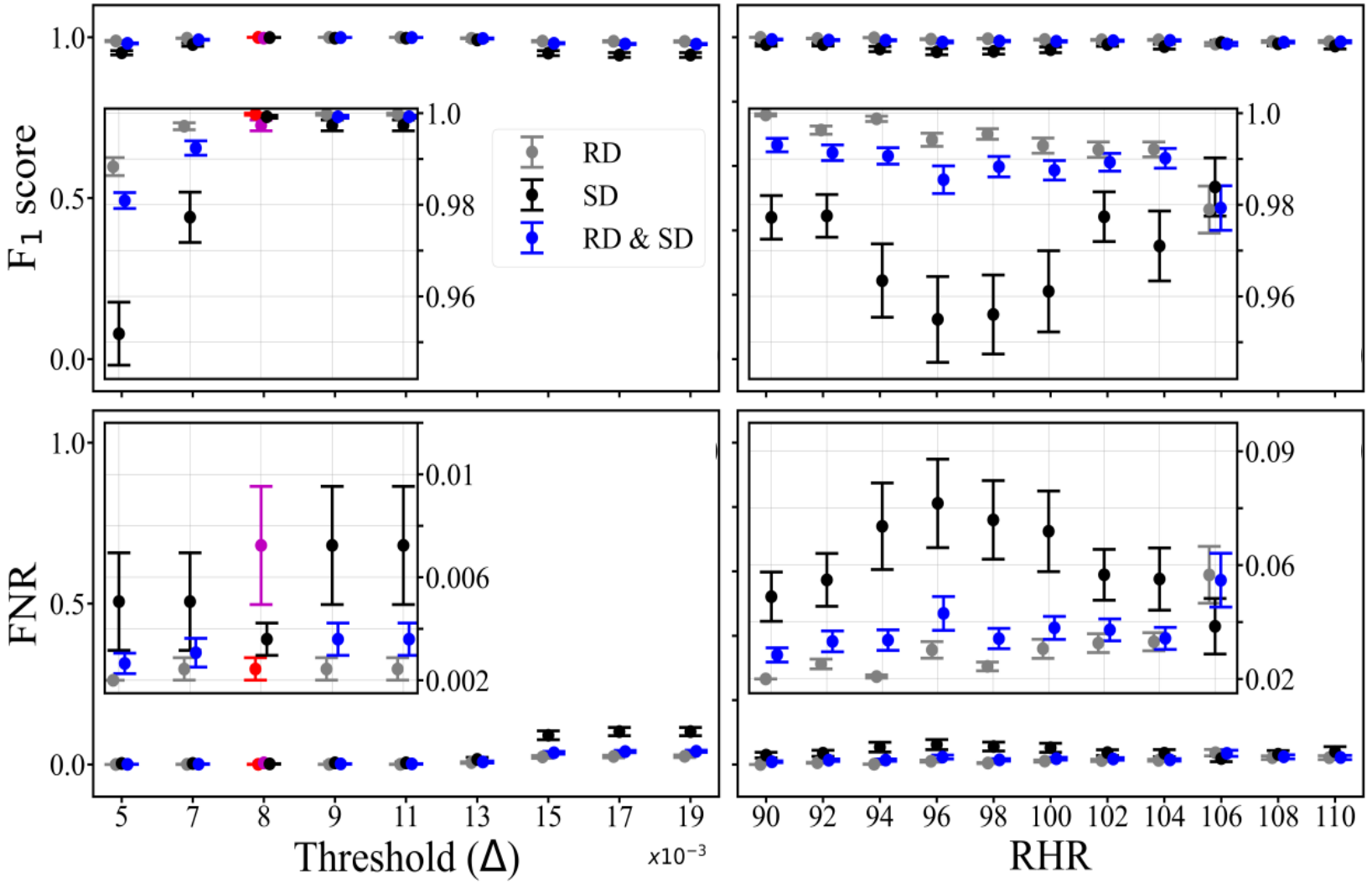
FNR – False Negative Rate

Uncertainties & Challenges

- RHR varies at the individual level due to factors such as age, gender and comorbidities.

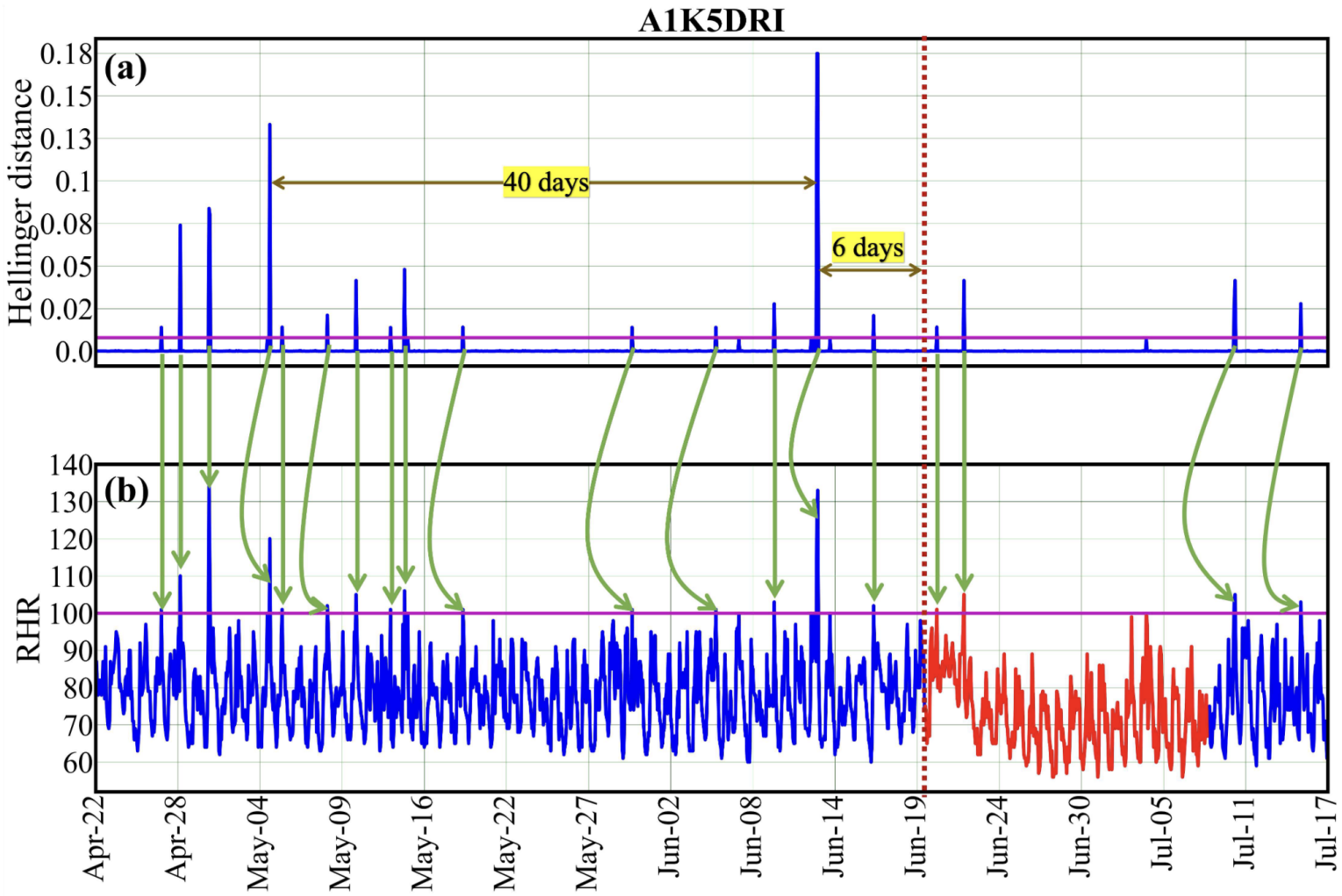
- What threshold value of Hellinger distance & RHR should be considered as anomalous?

Uncertainty Analysis using Real and Digital Twins



- Uncertainty in F_1 score and FNR with 90% confidence interval.
- A slight decline in model performance for thresholds $\Delta < 0.008$.
- Model perform consistently for $\Delta \geq 0.008$ & RHR values above 100 (CDC /AHA general RHR range: 60 -100 beats per minute).

Testing the Threshold on Real Data from Sick Individuals



Three-month RHR of sick user. When the user was sick due to a respiratory disease is shown in red.

- The purple horizontal bar represents the threshold (Δ = 0.008) for the Hellinger distance & RHR = 100.
- Any spikes in Hellinger distance that cross this threshold trigger an alert for potential sickness.
- Our anomaly detection method generates a series of alerts before the onset of actual sickness, validating the effectiveness of the chosen threshold value.

Take Home Message

- The integration of cutting-edge wearables, GenAI, and anomaly detection in our novel methodology provides a nuanced and accurate understanding of potential health concerns while significantly reducing the FNR.

- Augmenting the real dataset with GenAI-based digital twins to enhance population size resulted in strong concordance in uncertainty analysis compared to performing the same analysis solely on real data.

- Validation of the synthetic users (digital twins) by comparing their statistical signatures with those of real datasets showed excellent agreement, validating the effectiveness of our approach.

- We are in the process of (i) implementing a multilayered approach using the developed anomaly detection technique for large population-level analysis & (ii) developing a light weight app for wearables (edge computing).

Paper & Open-Source Software

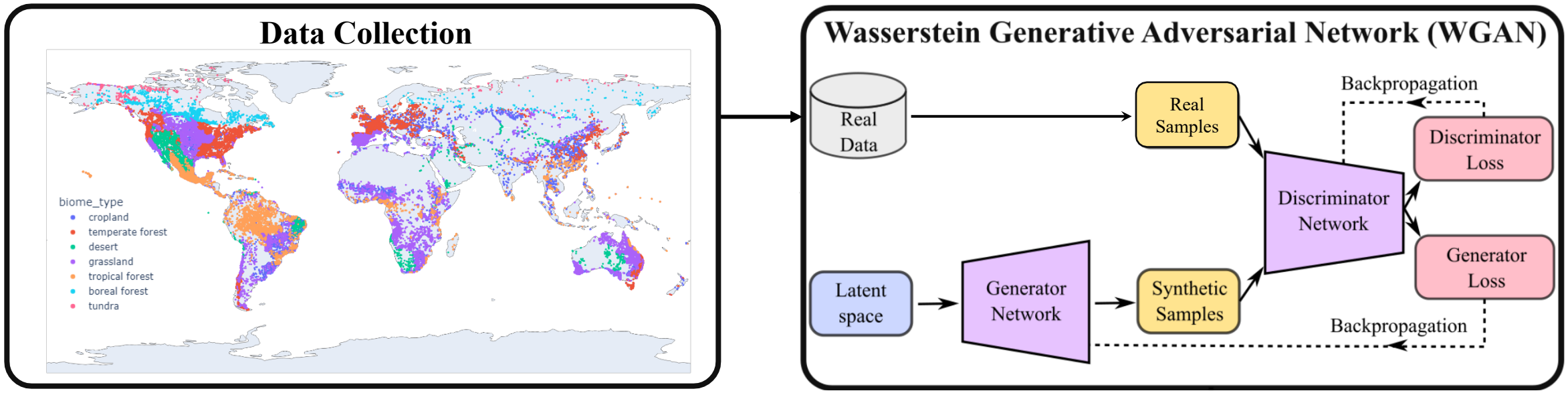
- TwinMe4AD: WGAN-based Digital Twins for Anomaly Detection (SCR#:3056.0) https://github.com/sandialabs/TwinMe4AD; https://doi.org/10.11578/dc.20240703.2.

- M. Kamruzzaman, Jorge Salinas, Hemanth Kolla, Kenneth Sale, Uma Balakrishnan, and Kunal Poorey, *GenAI based digital twins aided data augmentation increases accuracy in real-time cokurtosis based anomaly detection of wearable dataset*, *Under Review Nature Scientific Reports* (https://doi.org/10.21203/rs.3.rs-4427255/v1).

Climate Change: Filling gaps in Soil Organic Carbon data

- Limited availability of data in high latitude regions such as boreal forests and tundra.
- Difficult and often inaccessible regions for obtaining soil samples.
- Climate warming at high latitudes is causing widespread degradation and thawing of permafrost soils, leading to the release of greenhouse gases like CO_2 and CH_4.
- A significant portion of permafrost region SOC stocks could be emitted as greenhouse gases under changing climate conditions, making them a vulnerable component of the global carbon cycle.
- Reliable estimates of the magnitude and spatial variation of permafrost region SOC stocks are essential to understand environmental controls and reduce uncertainty in predicting carbon-climate feedbacks.

Soil Organic Carbon data Generation



We used GenAI-based WGAN to generate synthetic SOC data to mimic the real data were samples are often inaccessible.

Data Collection and Sampling

- What are the environmental variables of high importance?

Wang et al., *"Global soil profiles indicate depth-dependent soil carbon losses under a warmer climate"*, Nat. Comm., 2022

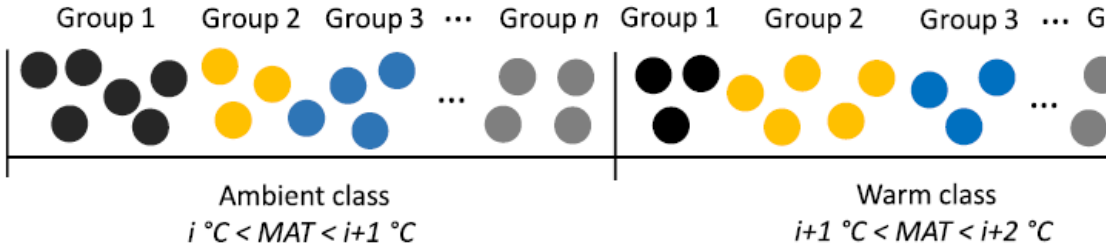


Fig. 1: Schematic representation of the approach used to quantify the response of soil organic carbon (SOC) to warming

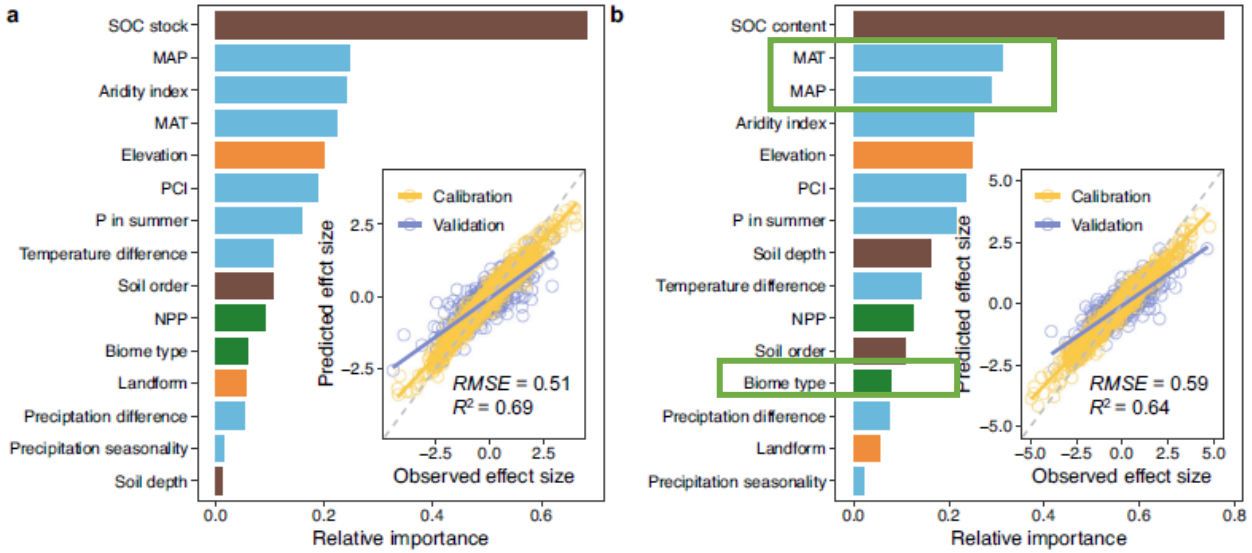


Fig. 3: The relative importance of environmental variables in influencing the response of soil organic carbon (SOC) to warming

*"The responses of SOC to **warming vary significantly across biome types**"*

*"The responses of SOC to **warming** are also **regulated by precipitation** including its seasonal pattern, but are **less influenced by soil type and landform**"*

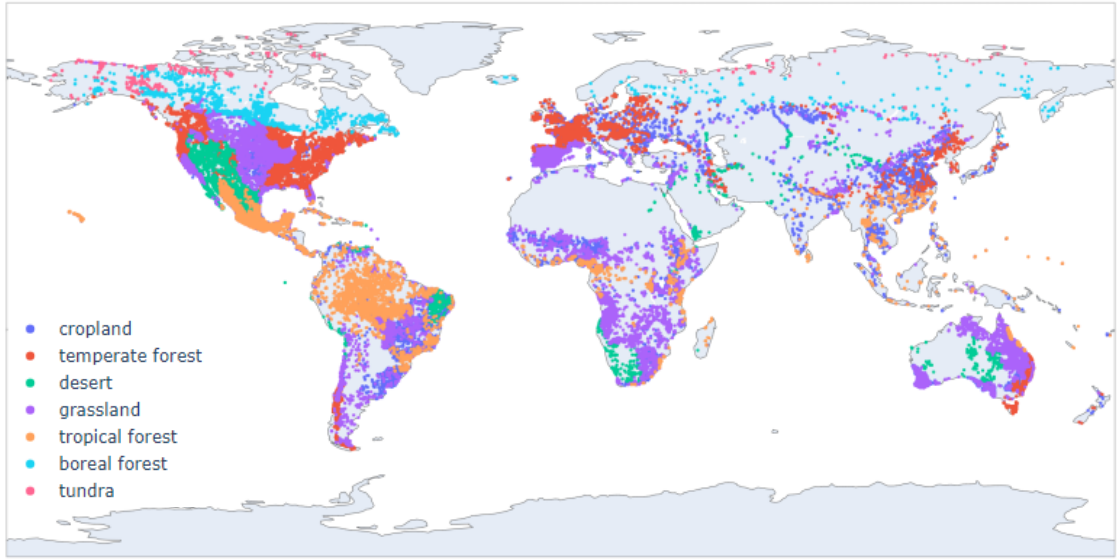
*"…our assessment does not support that soil type plays a significant role in regulating overall SOC balance under climate warming at the global scale. **At finer scales, we acknowledge that heterogeneous soil properties may be important**"*

Soil profiles sorted by:

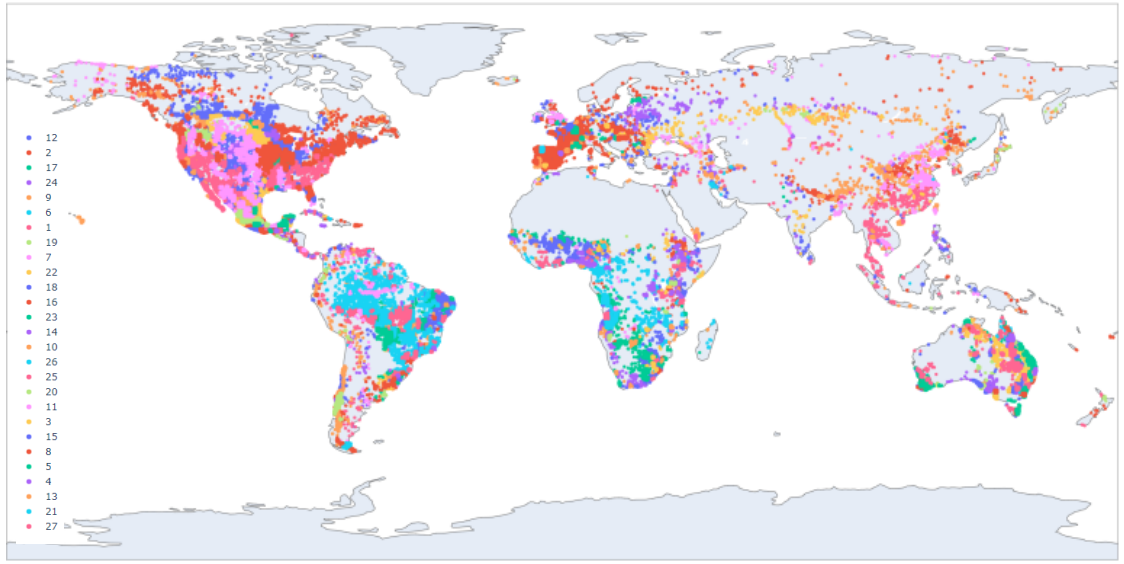
- *Biome Type*
- Soil type
- Mean annual temperature (MAT)
- Mean annual precipitation (MAP)

Data Collection and Sampling

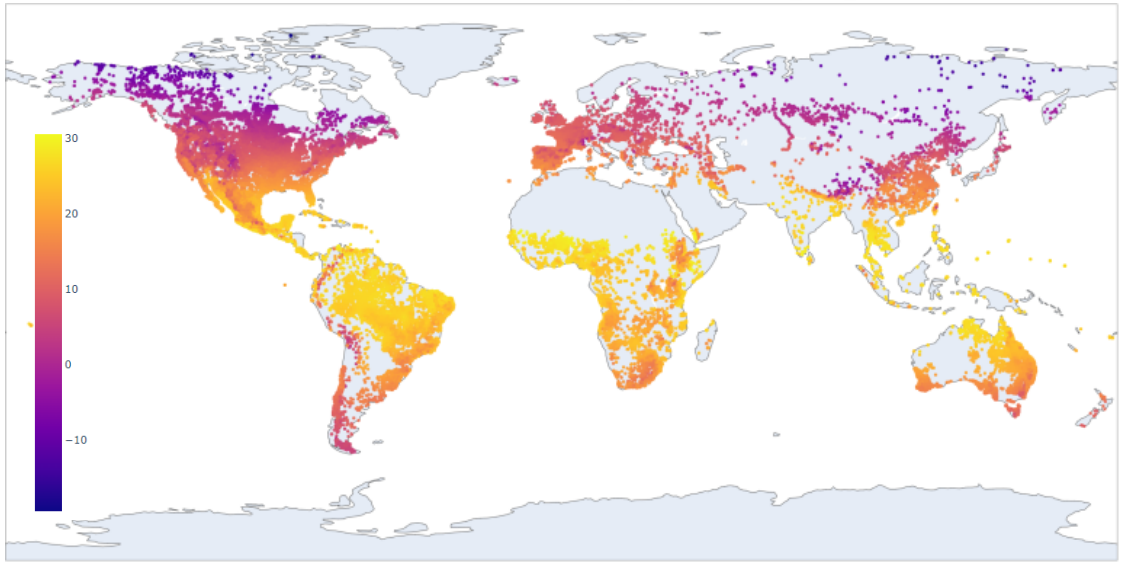
Biome Type



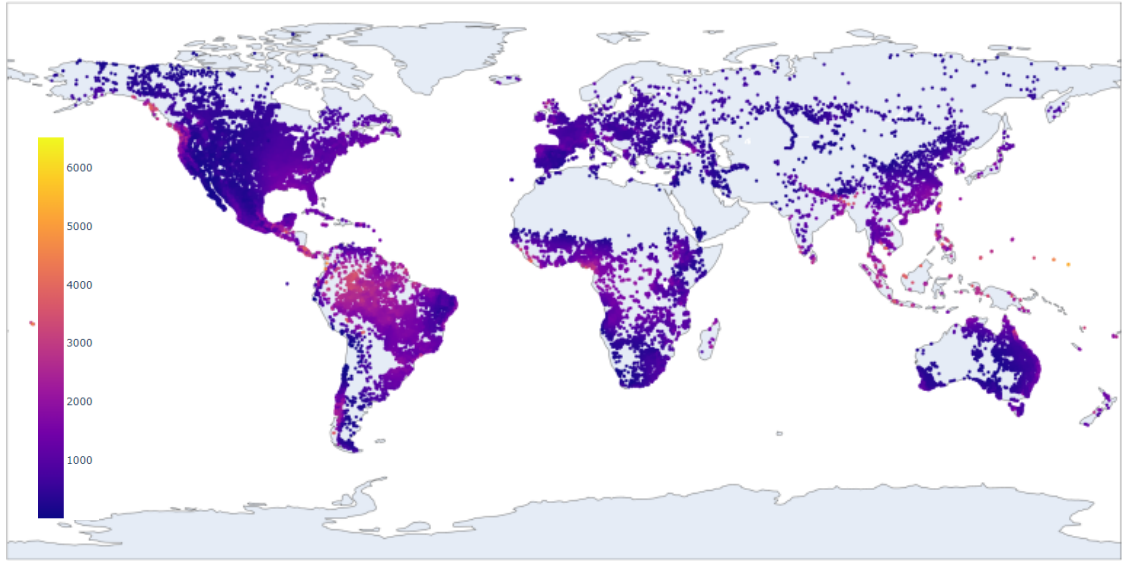
Soil Type



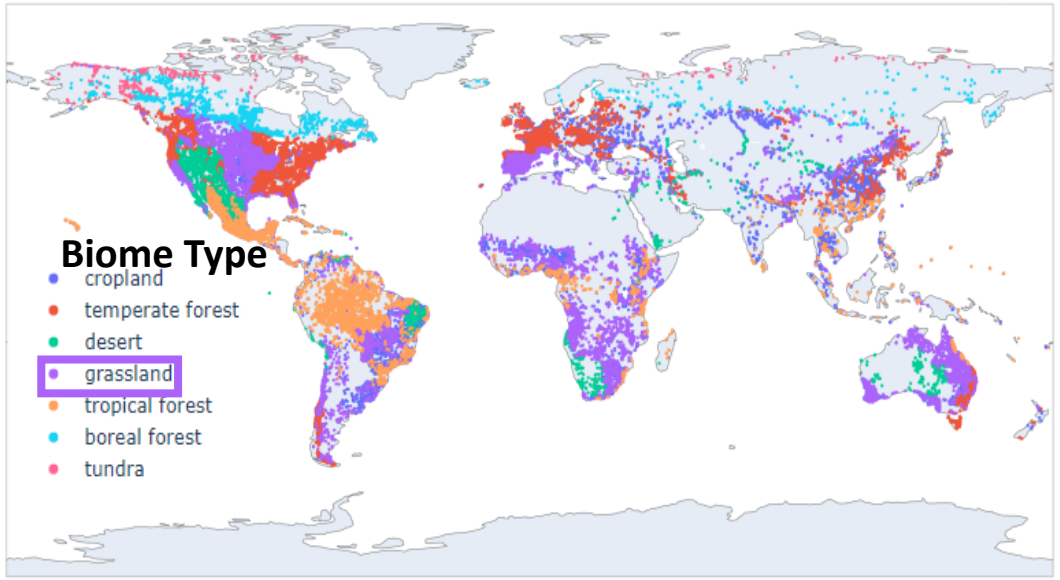
Mean Annual Temperature $[°C]$

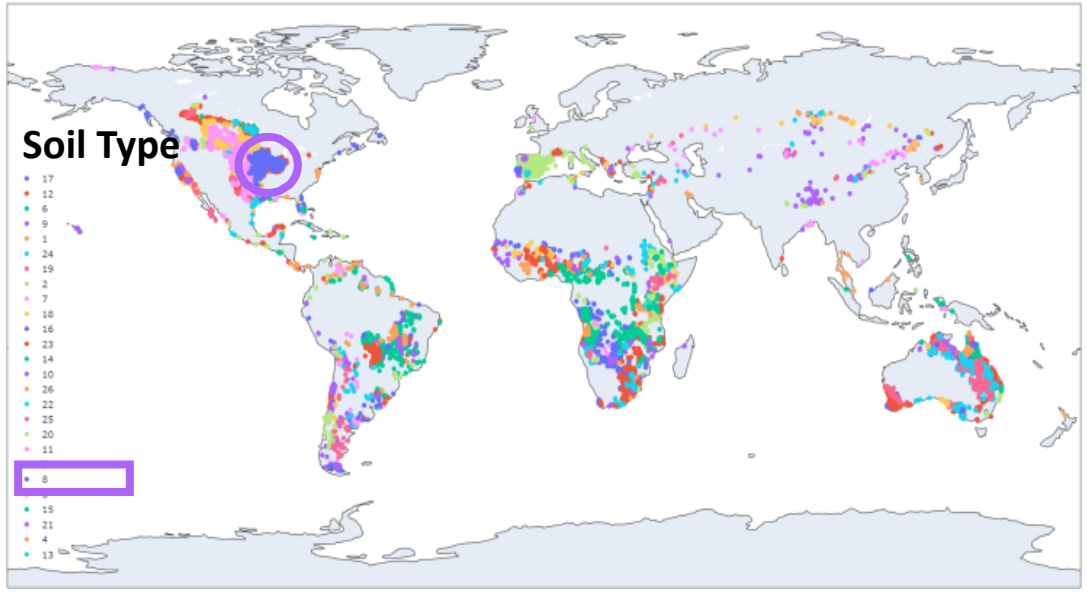


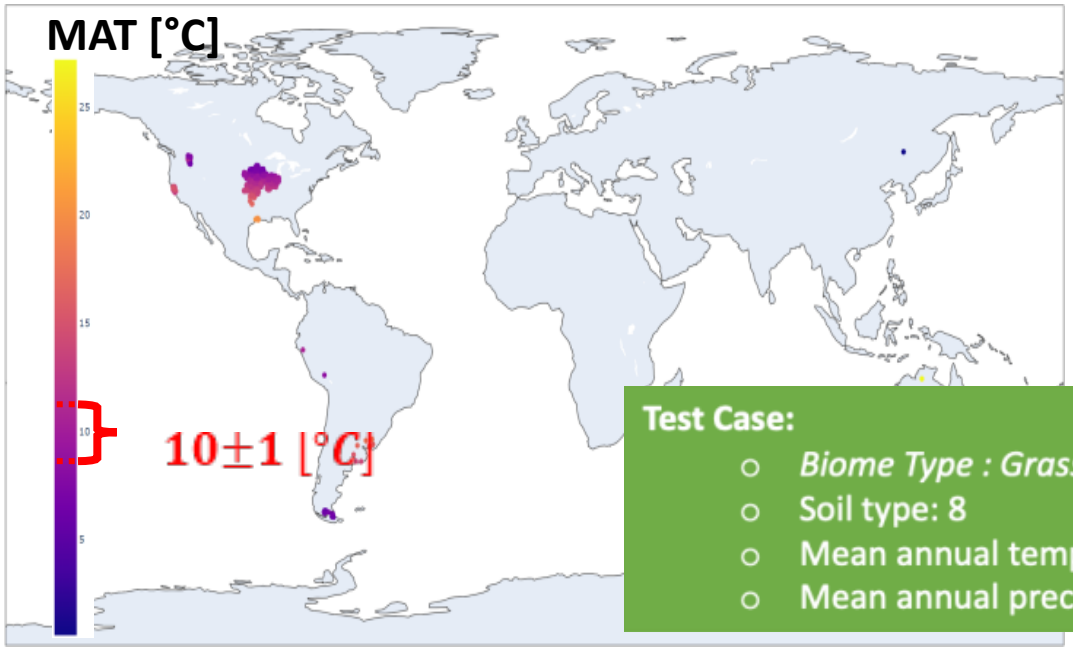
Mean Annual Precipitation $[mm]$

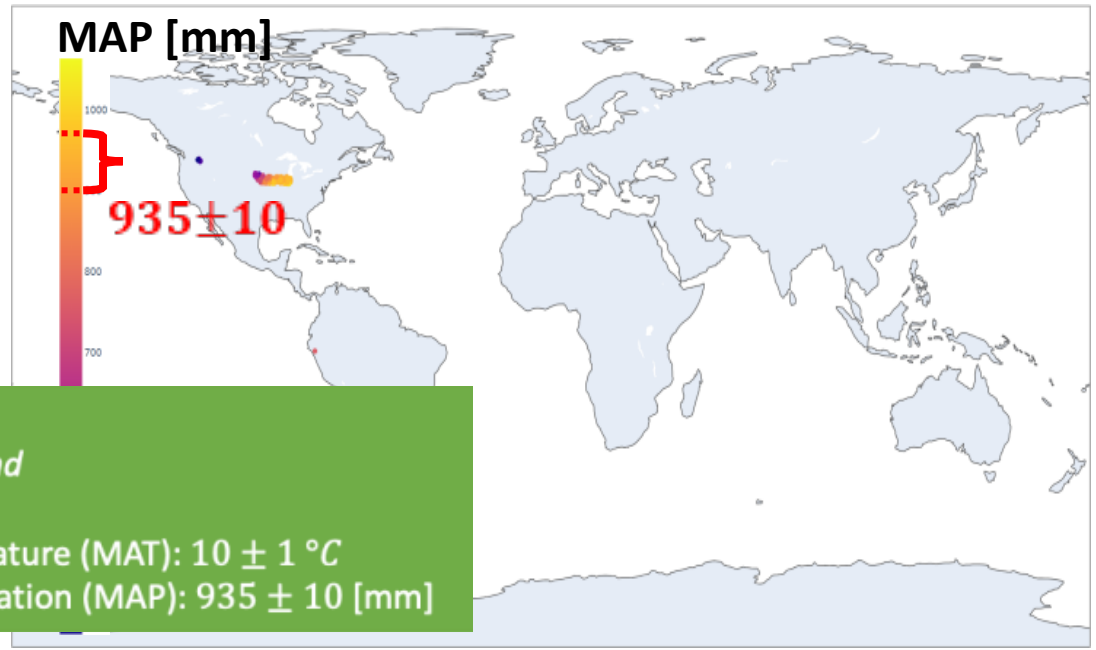


Test Case for Filling the Gaps in SOC









Test Case:

- *Biome Type : Grassland*
- Soil type: 8
- Mean annual temperature (MAT): $10 \pm 1\ °C$
- Mean annual precipitation (MAP): 935 ± 10 [mm]

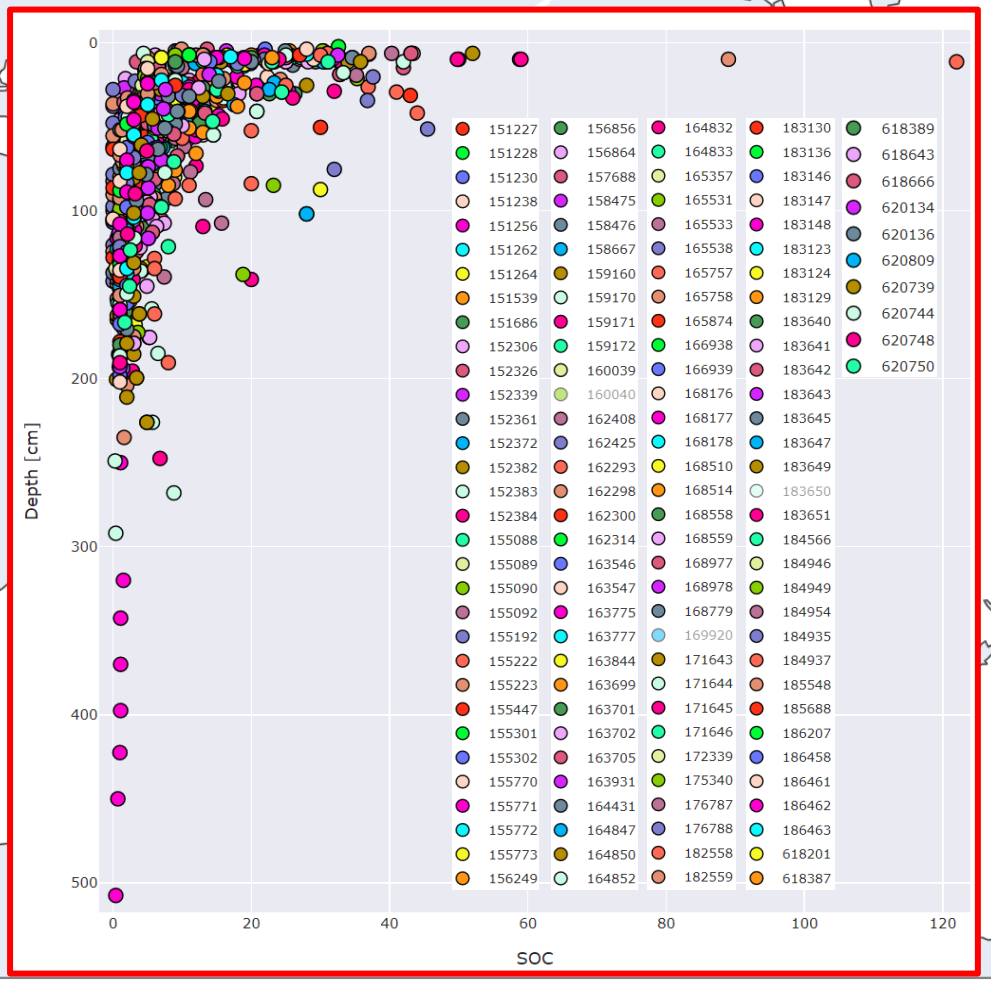
Synthetic Data Generation for Grasslands in the Midwest

Profile tested: ID 152339

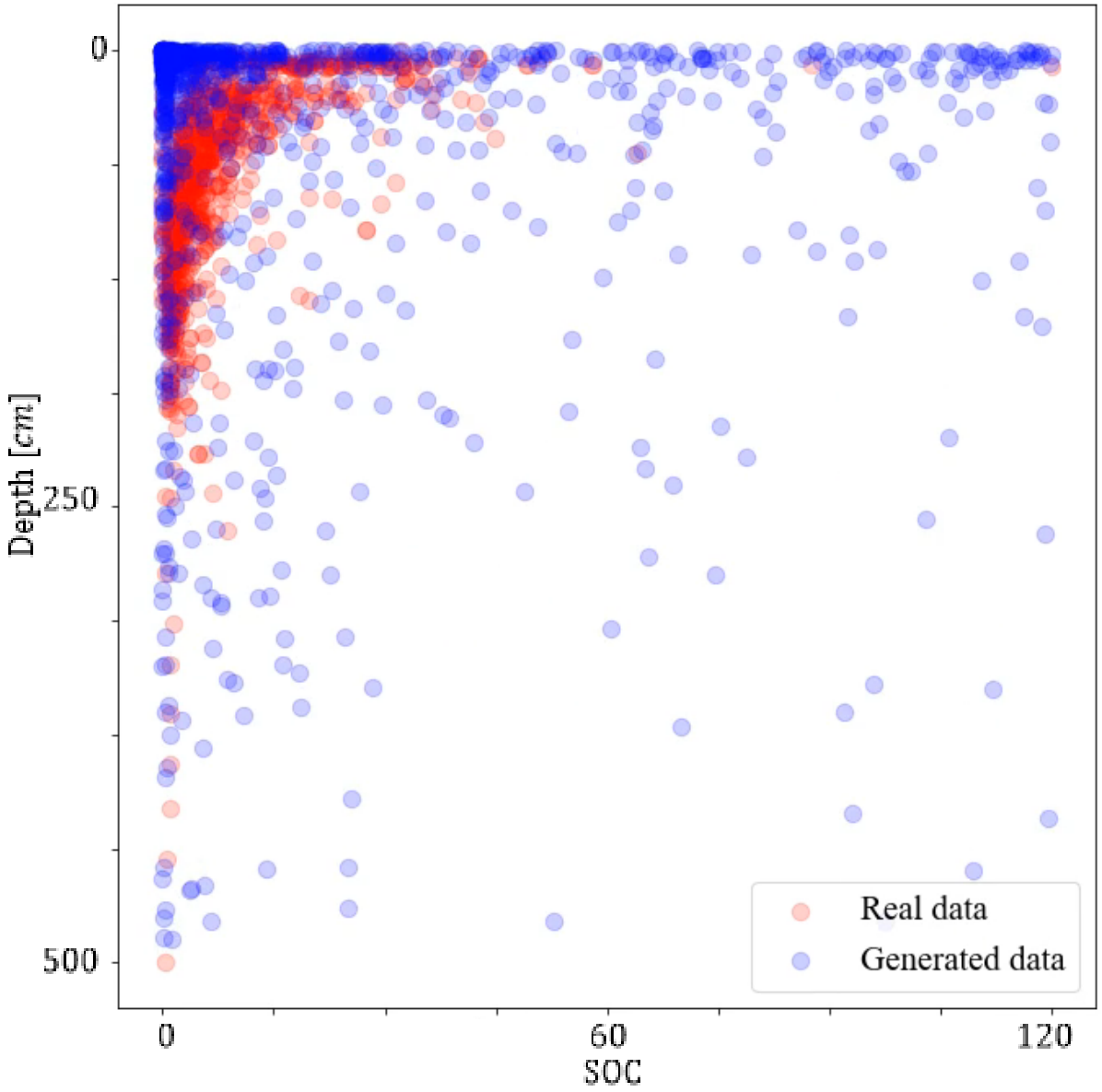
- *Taken out of sampled data*
- We will try to predict its profile

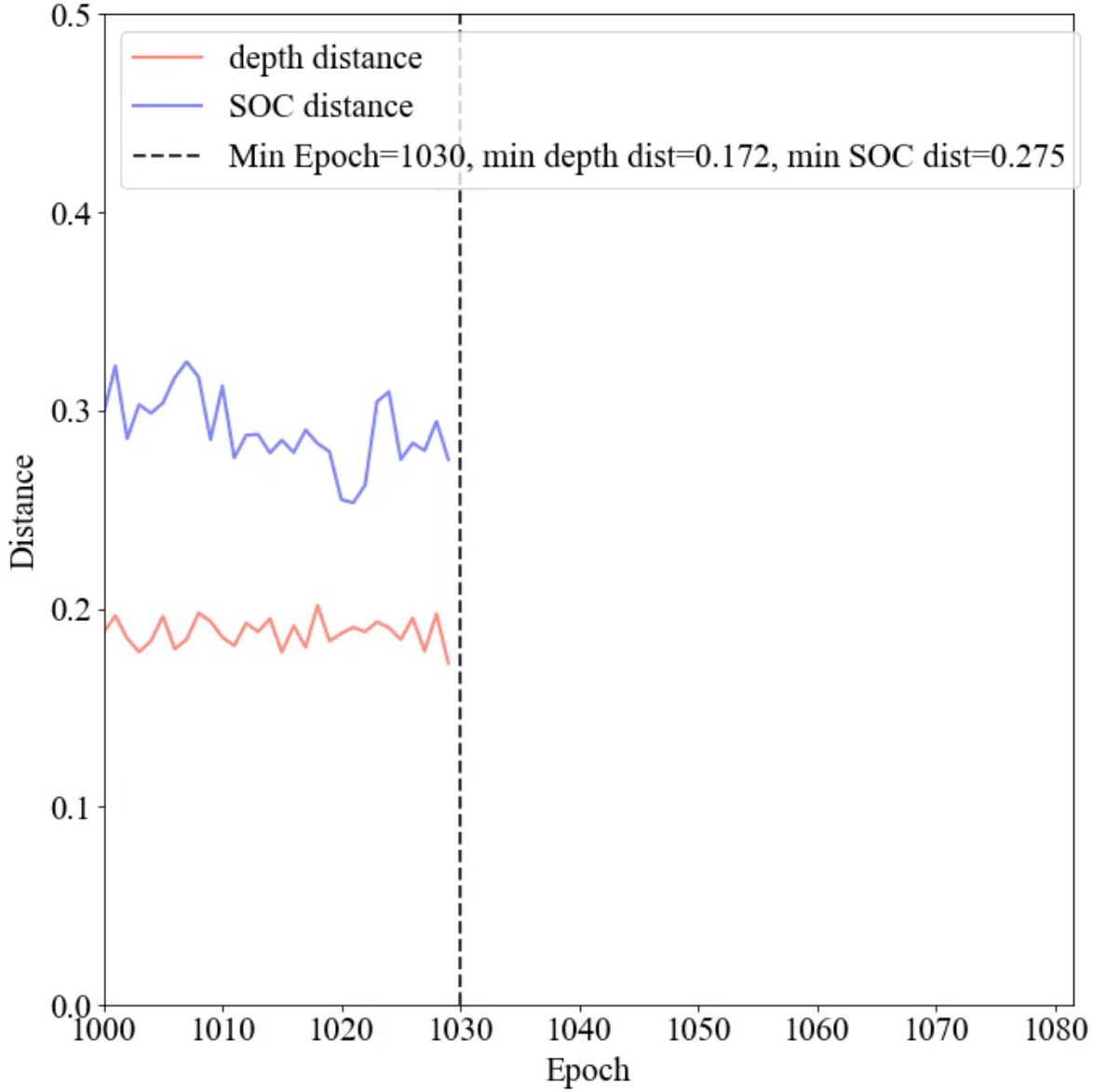
Profile ID: 152339

Synthetic data will be selected by its proximity in the (MAT, MAP) parameter space.



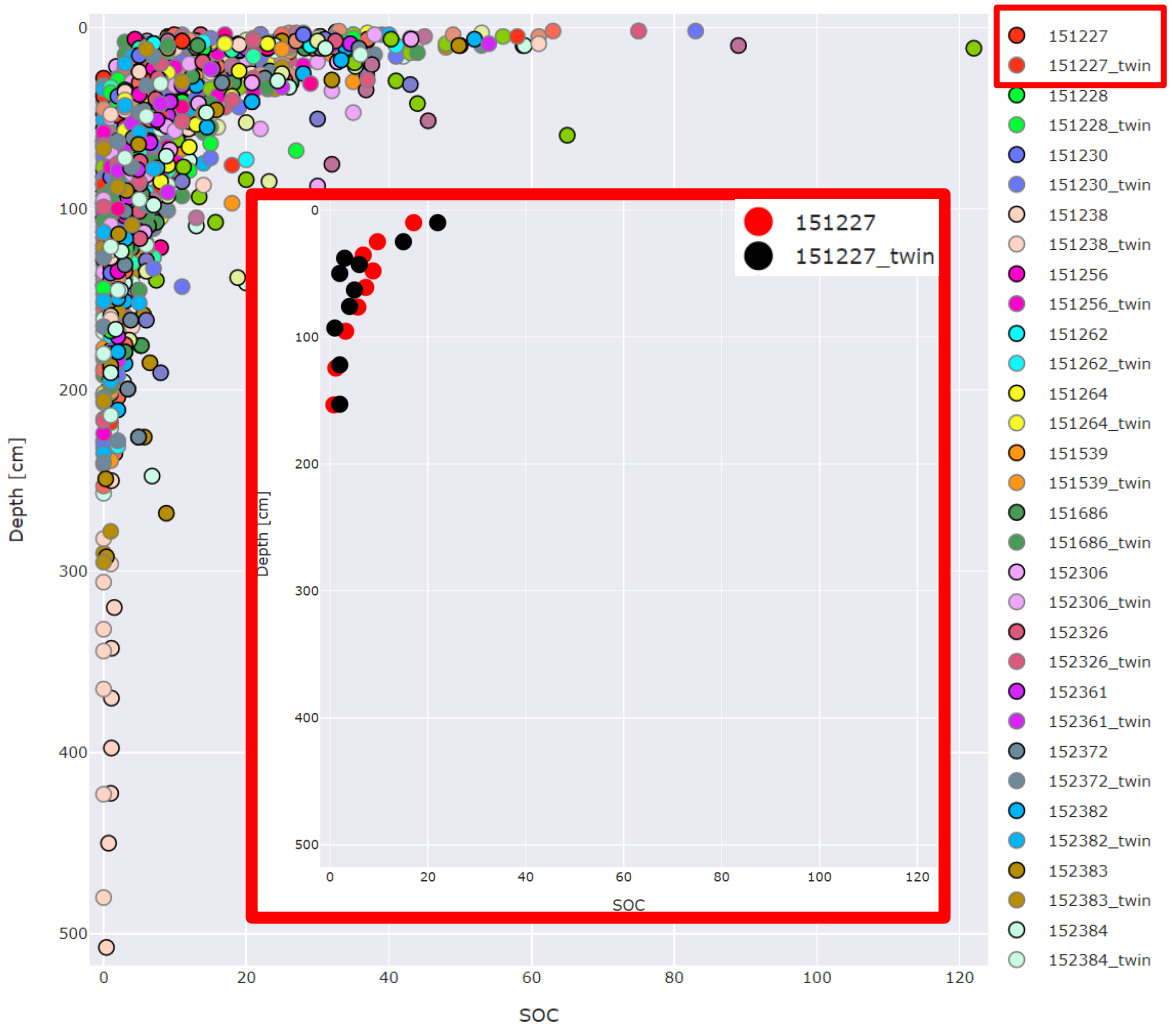
Synthetic Twin of Real-Time Soil Organic Carbon

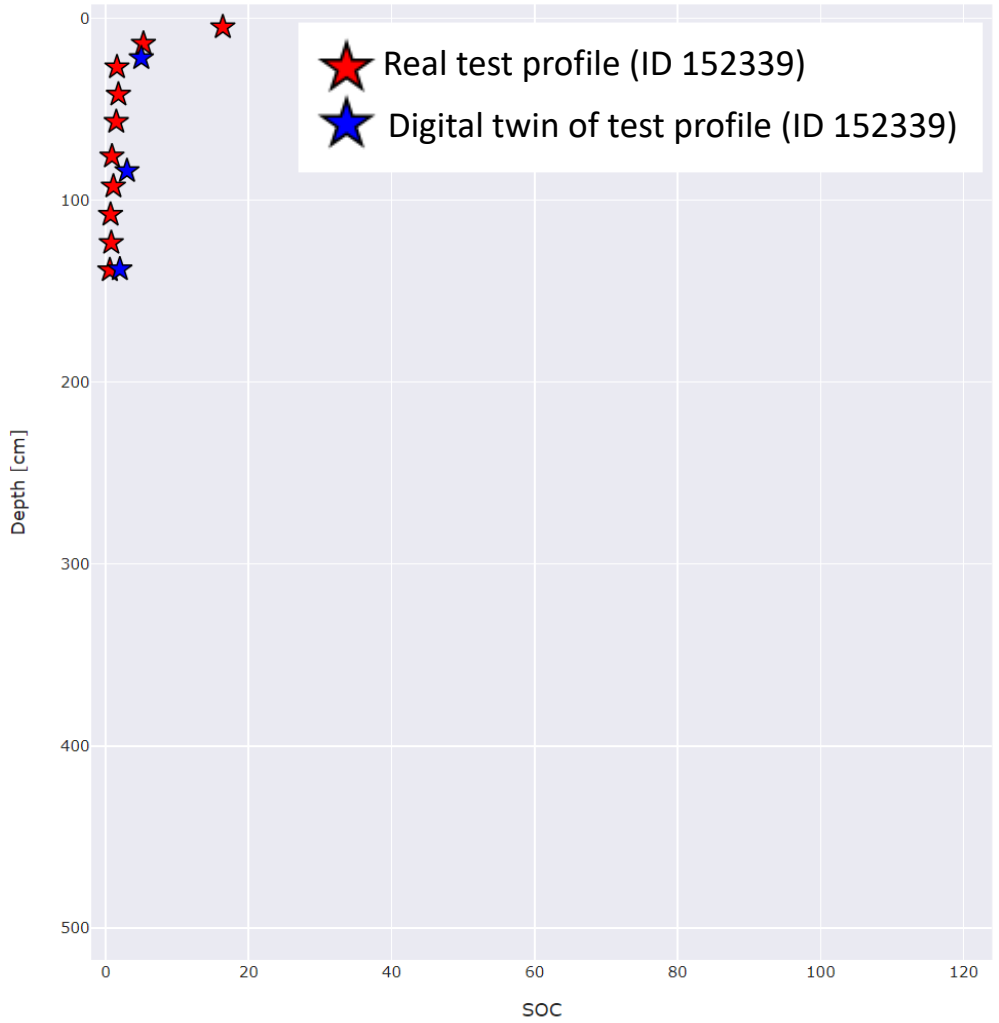




Low Wasserstein distance = Highly similar sample distributions

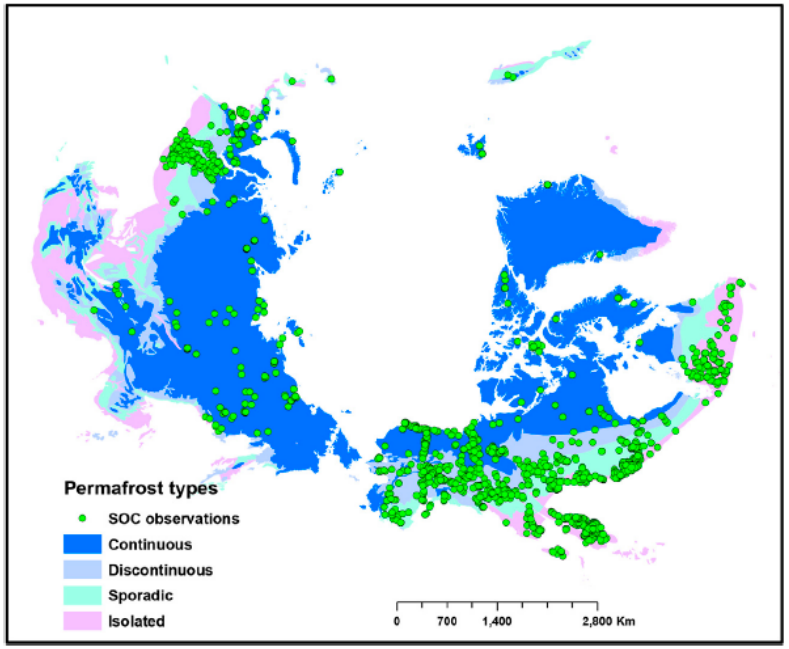
Selection of Synthetic Twin



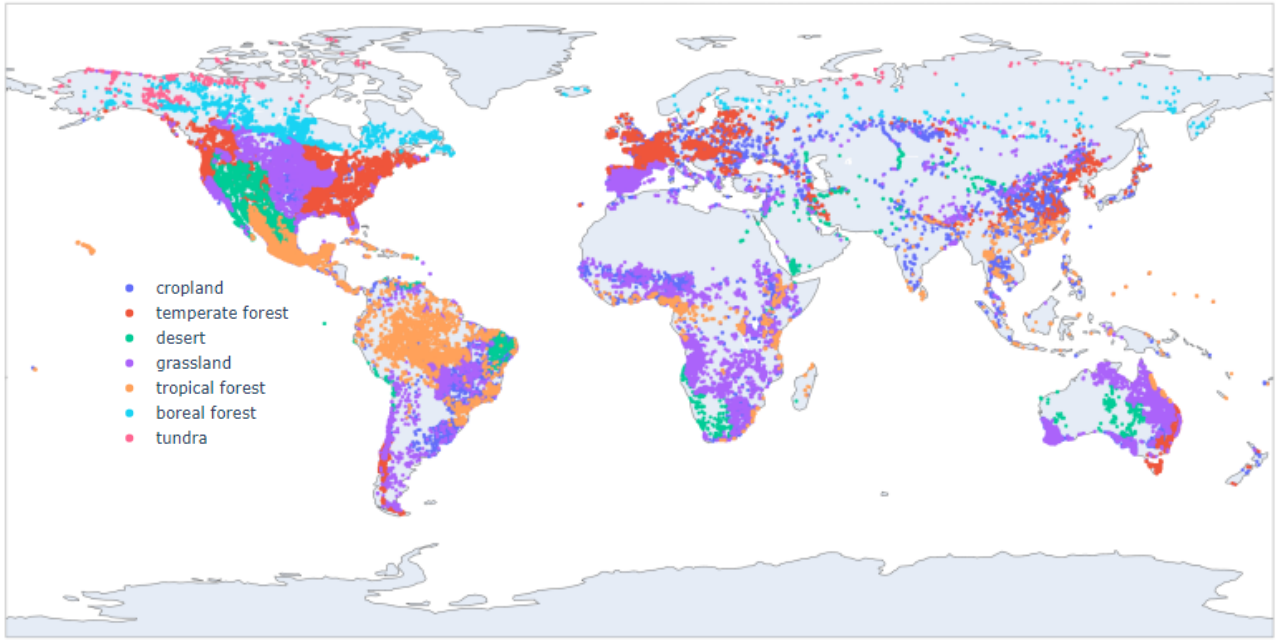


Work in Progress

- Systematic testing of digital twin generation by dividing known a-priori profiles between training and testing datasets
- Test digital twin generation in biome types of interest (i.e. boreal forest and tundra).
- Generate data on locations of interest



Biome Type



Mishra et al., *"Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region"*, Front. Big Data 3:528441., 2020

Scope of Digital Twins: Interested in Possible Collaboration

- Digital twins can simulate ecosystems and environmental processes, aiding in the study of climate change, pollution, and conservation efforts.
  - We are engaged in filling the gaps in soil carbon data across permafrost regions to enhance our understanding of soil carbon dynamics and climate change impacts for improved environmental monitoring and assessment.
- Digital twins can be utilized in ND space (Ex: predicting the performance of aging weapons, assessing the reliability of the stockpile), allowing for safer and more efficient research and development.
- Digital twins of biological systems can help in understanding disease mechanisms, developing new treatments, and personalizing medicine.
- Integration of Digital twins and ABM simulations offer improved decision-making.

and more …

Thank You for Your Time and Attention!

For questions or follow-up discussions:

Uma Balakrishnan, ubalakr@sandia.gov
(Mathematical Modeling, GenAI-WGAN, VVUQ)

Kunal Poorey, kpoorey@sandia.gov
(GenAI-WGAN, Bio-surveillance domain expert)

Umakant Mishra umishra@sandia.gov
(Soil Organic Carbon domain expert)

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

Anomaly detection of original wearable dataset

F: Frequency, S: Sliding, C: $\mathcal{S}_t$ [1: #features], Δ: Threshold on Hellinger distance (min Hellinger distance for sick person)

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.538
60	30	2	Actual Sick	55 (TP)	7 (FN)	Precision	0.534
			Actual healthy	48 (FP)	9 (TN)	Recall	0.887
						F1 score	0.667
						MCC score	0.066

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.429
60	60	2	Actual Sick	49 (TP)	7 (FN)	Precision	0.445
			Actual healthy	61 (FP)	2 (TN)	Recall	0.875
						F1 score	0.590
						MCC score	-0.176

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.555
60	30	3	Actual Sick	55 (TP)	14 (FN)	Precision	0.585
			Actual healthy	39 (FP)	11 (TN)	Recall	0.797
						F1 score	0.675
						MCC score	0.021

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.471
60	60	3	Actual Sick	49 (TP)	14 (FN)	Precision	0.500
			Actual healthy	49 (FP)	7 (TN)	Recall	0.778
						F1 score	0.609
						MCC score	-0.127

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.563
60	30	4	Actual Sick	55 (TP)	24 (FN)	Precision	0.663
			Actual healthy	28 (FP)	12 (TN)	Recall	0.696
						F1 score	0.679
						MCC score	-0.004

F	S	C	Confusion matrix				
				Sick (>Δ)	Healthy (<Δ)	Accuracy	0.513
60	60	4	Actual Sick	49 (TP)	46 (FN)	Precision	0.803
			Actual healthy	12 (FP)	12 (TN)	Recall	0.516
						F1 score	0.628
						MCC score	0.013

Accuracy, Precision, Recall: Close to 1 is better

F1 score: higher is better

MCC score: Ranges from -1 to 1. Close to -1 or close to 1 is better.

Anomaly detection of original wearable dataset

F: Frequency, S: Sliding, C: #columns, Δ: Threshold on Hellinger distance (min Hellinger distance for sick person)

F	S	C	Confusion matrix			Accuracy	0.429
				Sick (>Δ)	Healthy (<Δ)	Precision	0.500
120	60	2	Actual Sick	35 (TP)	33 (FN)	Recall	0.515
			Actual healthy	35 (FP)	16 (TN)	F1 score	0.507
						MCC score	-0.173

F	S	C	Confusion matrix			Accuracy	0.353
				Sick (>Δ)	Healthy (<Δ)	Precision	0.348
120	120	2	Actual Sick	32 (TP)	17 (FN)	Recall	0.653
			Actual healthy	60 (FP)	10 (TN)	F1 score	0.454
						MCC score	-0.240

F	S	C	Confusion matrix			Accuracy	0.345
				Sick (>Δ)	Healthy (<Δ)	Precision	0.333
120	60	3	Actual Sick	35 (TP)	8 (FN)	Recall	0.814
			Actual healthy	70 (FP)	6 (TN)	F1 score	0.473
						MCC score	-0.160

F	S	C	Confusion matrix			Accuracy	0.345
				Sick (>Δ)	Healthy (<Δ)	Precision	0.311
120	120	3	Actual Sick	32 (TP)	7 (FN)	Recall	0.821
			Actual healthy	71 (FP)	9 (TN)	F1 score	0.451
						MCC score	-0.092

F	S	C	Confusion matrix			Accuracy	0.412
				Sick (>Δ)	Healthy (<Δ)	Precision	0.365
120	60	4	Actual Sick	35 (TP)	9 (FN)	Recall	0.795
			Actual healthy	61 (FP)	14 (TN)	F1 score	0.500
						MCC score	-0.022

F	S	C	Confusion matrix			Accuracy	0.412
				Sick (>Δ)	Healthy (<Δ)	Precision	0.432
120	120	4	Actual Sick	32 (TP)	28 (FN)	Recall	0.533
			Actual healthy	42 (FP)	17 (TN)	F1 score	0.478
						MCC score	-0.184

Accuracy, Precision, Recall: Close to 1 is better

F1 score: higher is better

MCC score: Ranges from -1 to 1. Close to -1 or close to 1 is better.

Structure of Data Science Toolkit for Multi-layered Anomaly Detection in Pandemic