Concise ML Explanations

*Project Number (SL22-Concise ML Explanations-PD3SZ)*
Vitus J. Leung, Nathan Marshall, Mandy C. Nevins, and Michael D. Clemenson

U.S. DEPARTMENT OF ENERGY

NNSA
National Nuclear Security Administration

Overview & Goals

- Some nonproliferation machine learning (ML) classifiers on time series data require extensive manual labor to validate the classifications.
- Recently, a new black box **explainer for ML classifiers** on time series data, CoMTE [Ates et al.], was open source licensed. **CoMTE identifies the key time series data** for a classification.
- The goal is to apply CoMTE to one of these nonproliferation classifiers.
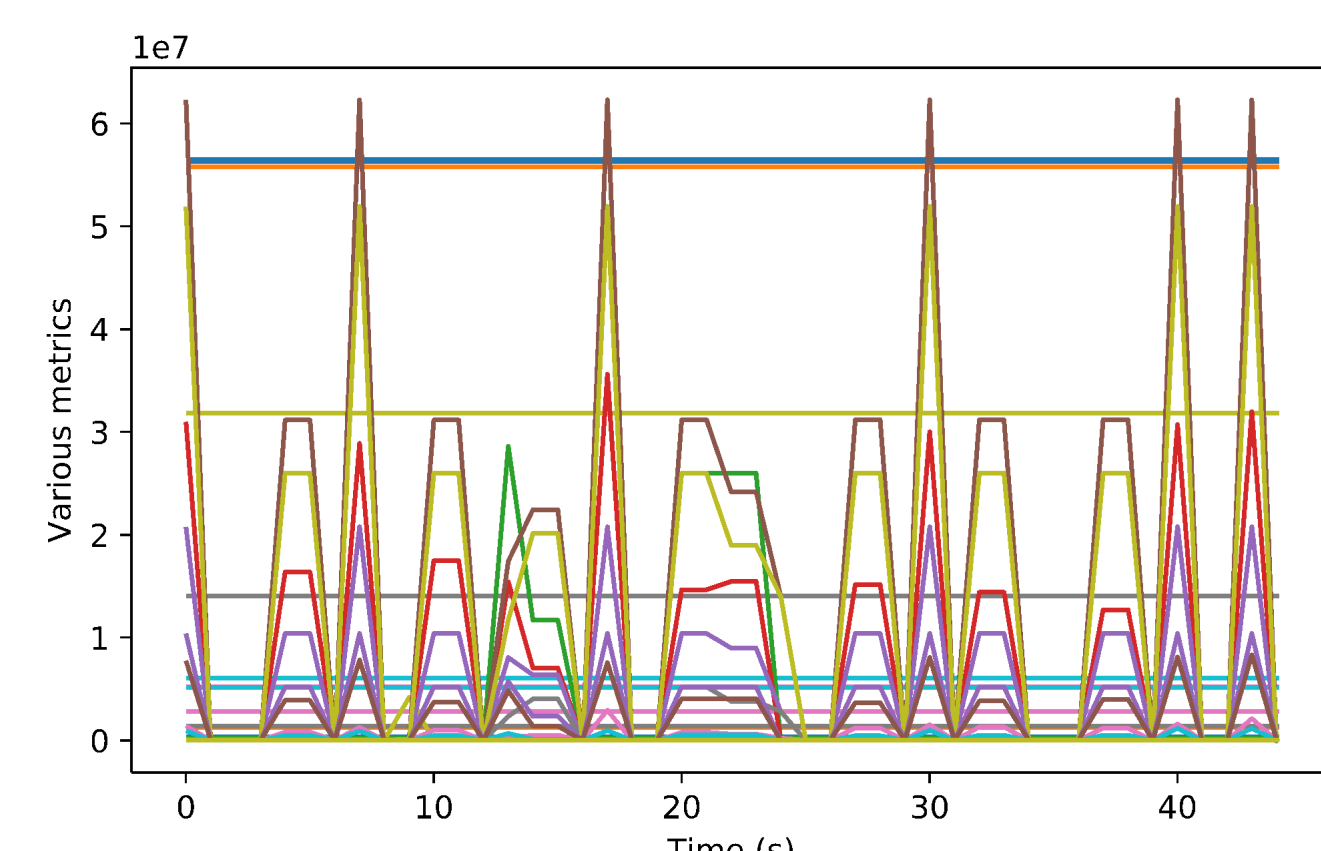
Telemetry data collected from a large-scale computing system
199 metrics are shown out of 2000

Why is this work important?

For ML classifiers to be effective they must require a minimum of labor, let alone expert labor, to verify the classifications. CoMTE may be able to **reduce the expertise and manual labor required** to verify classifications.
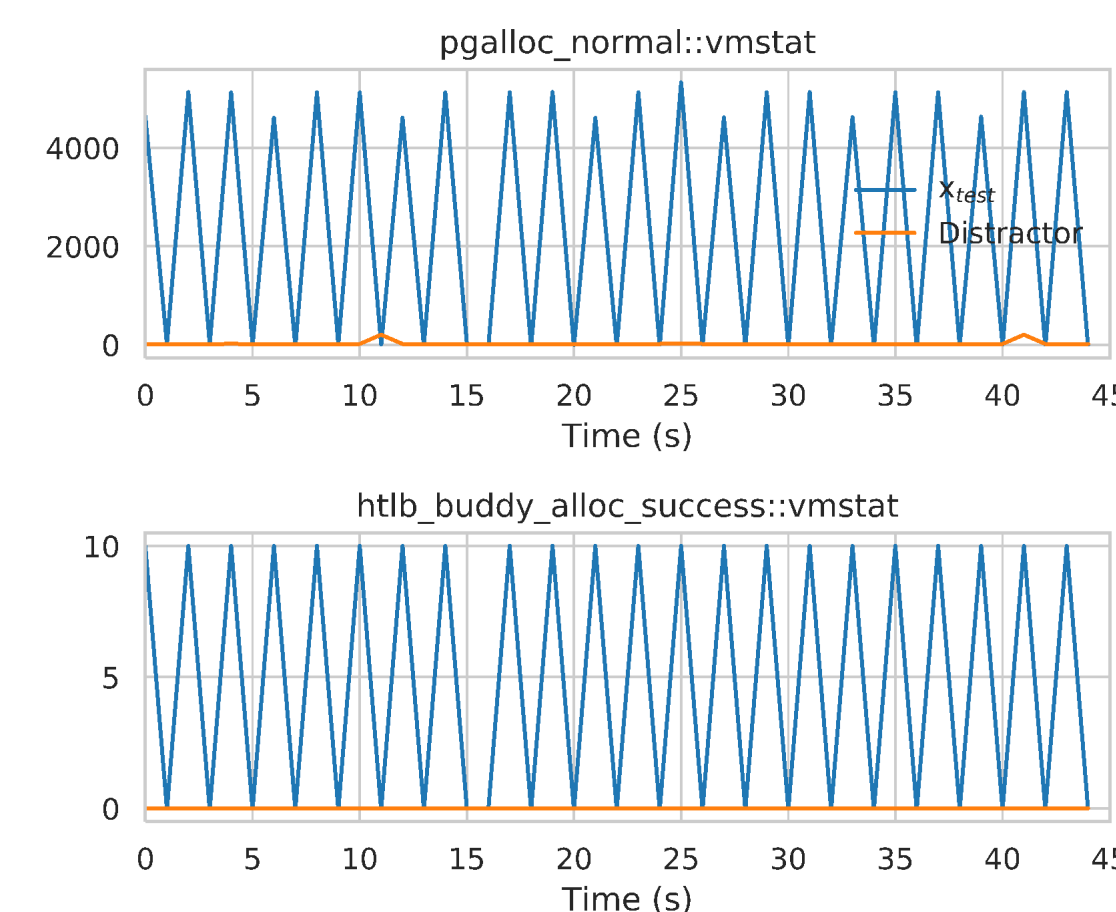
Key time series for memory leakage
X_test: Correctly classified sample with memory leakage anomaly
Distractor: Correctly classified sample without memory leakage anomaly

Technical Approach

CoMTE: Counterfactual explanations for multivariate time series.

- For a given **sample** and a **class of interest**, CoMTE finds a minimum set of substitutions from a **distractor***, such that the resulting sample is predicted as the class of interest
- E.g., "If feature X was not exceeding Y over time, test sample wouldn't be classified as Z"

*Another sample with class of interest's label.

One Local Maxima   Two Local Maxima

$f(x,y)=e^{-(x^2+y^2)}$ $f(x,y)=e^{-(x^2+y^2)} + 2e^{-((x-1.7)^2+(y-1.7)^2)}$
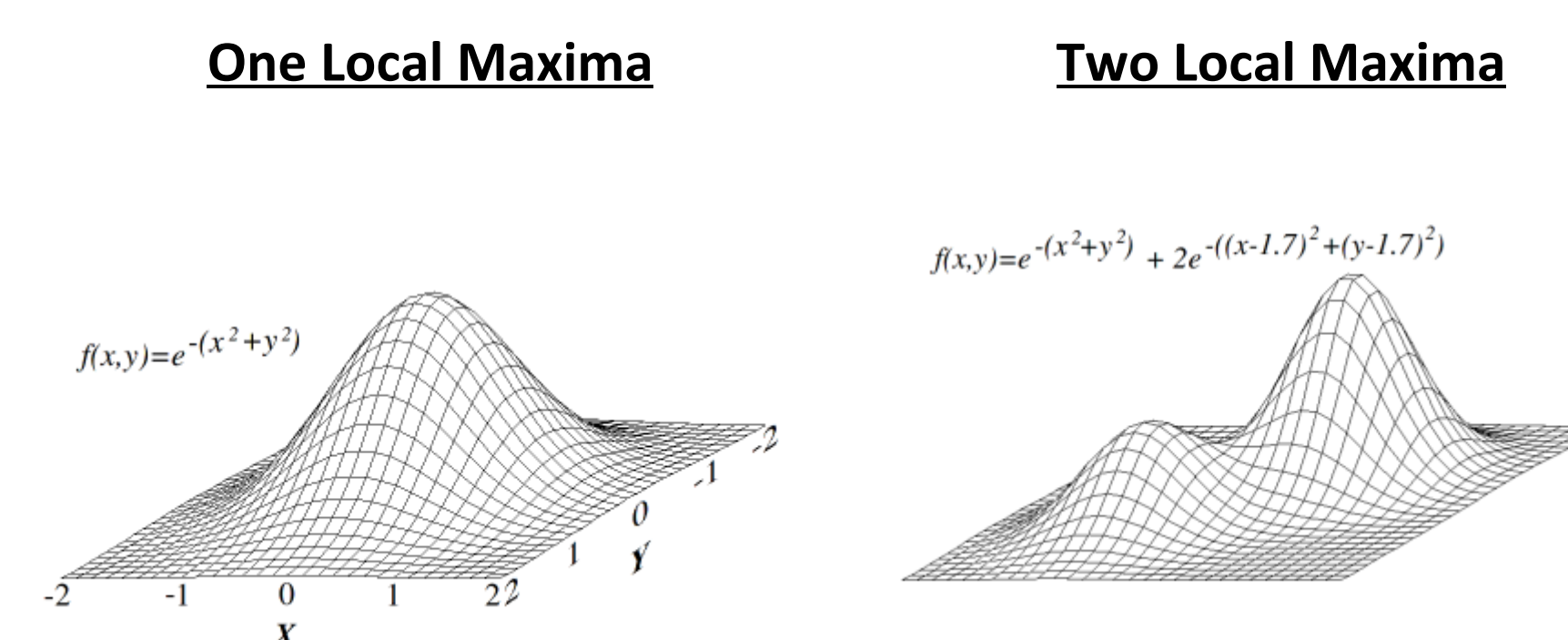
Hill climbing algorithms optimize over a surface
Figures: https://en.wikipedia.org/wiki/Hill_climbing

Algorithm summary

☐ **User provides test sample**

☐ **CoMTE chooses distractor**
  ▪ "Close" to the sample
  ▪ All training set instances are stored in *KD-Tree* per class

☐ **CoMTE solves optimization**
  ▪ Random-Restart Hill Climbing
  ▪ Sequential Greedy Approach

☐ **CoMTE provides explanation**

Expected Technical Challenges

Connecting to someone else's research code

We have already encountered **customizations in the ML** that the developers don't even think of as custom.

GOOD CODE IS...
DON'T TOUCH!!!
... LIKE A MING VASE: BEAUTIFUL BUT FRAGILE
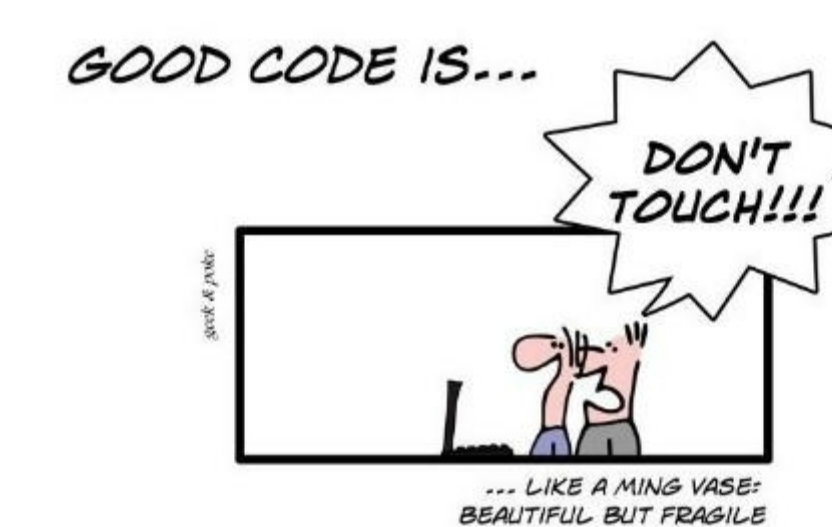
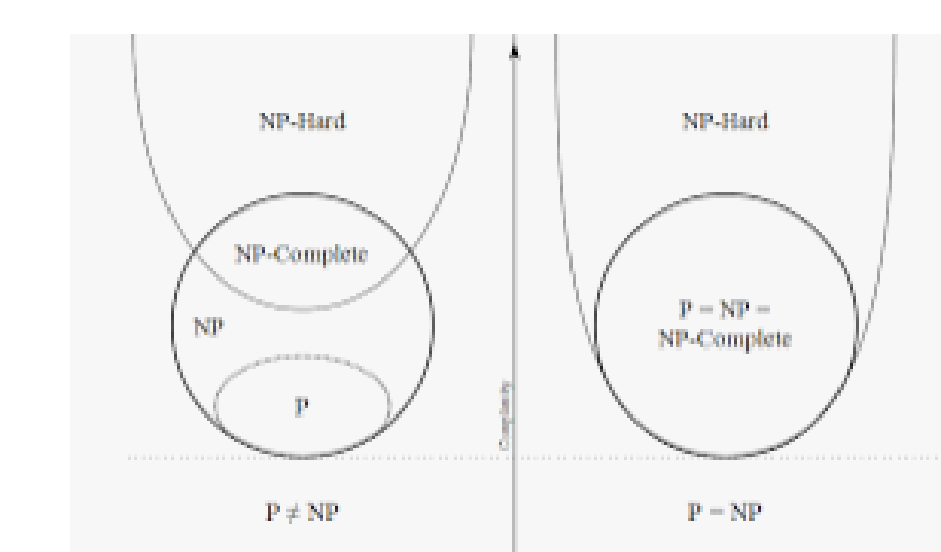### Connecting to someone else's private data

We have already encountered **permissions required** that the regular users of the data don't even think of as separate but necessary.

Free Icon | Private network
freepik.com

NP-hard search

The search in CoMTE has been proven to be NP-hard. The **heuristics** in CoMTE may prove inadequate for our nonproliferation application.

NP-Complete problems be also NP-Hard ...
stackoverflow.com

Progress to Date

We have been able to manually run the nonproliferation ML and are close to completing a **bare bones connection to CoMTE**.

Project plan

Task	By
Project baseline review	12/21
Data integration	06/22
Annual program review	06/22
ML integration	09/22
Search optimization	03/23
Independent assessment	06/23
Annual program review	06/23
Explanation evaluation	09/23
Final out brief	09/23

Task	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Project Baseline Review	■							
Task 1 Data Integration	■	■						
Task 2 Machine Learning Integration		■	■					
Task 3 Search Optimization				■	■			
Independent Assessment						■		
Task 4 Evaluation						■	■	
Annual Program Review		■				■		
Final Out Brief								■

Future Work

Currently, CoMTE only explains supervised classification.

Areas for future work include extending explanations to **both unsupervised learning and supervised learning** other than classification.

Reference

E. Ates, B. Aksar, V. J. Leung and A. K. Coskun, "Counterfactual Explanations for Multivariate Time Series," 2021 ICAPAI, 2021, pp. 1-8, doi: 10.1109/ICAPAI49758.2021.9462056.

CONTACT
Vitus J. Leung
505.844.1896
vjleung@sandia.gov

Sandia National Laboratories

ENTER THE SAND NUMBER HERE