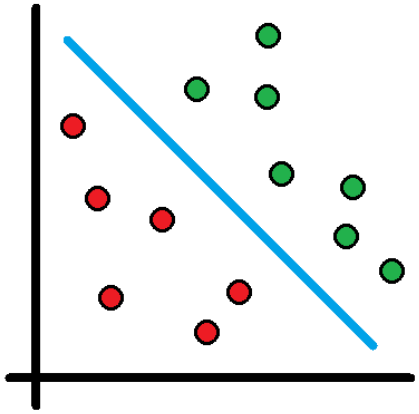
Cyber-Security and Machine Learning

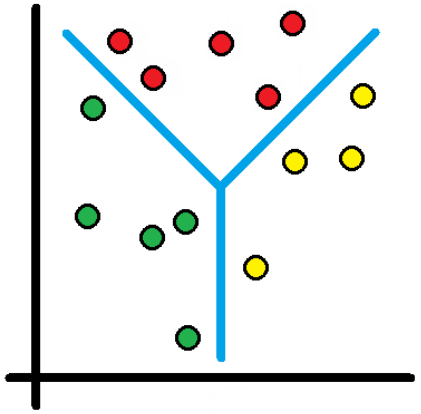
Purushottam Kar

Indian Institute of Technology Kanpur

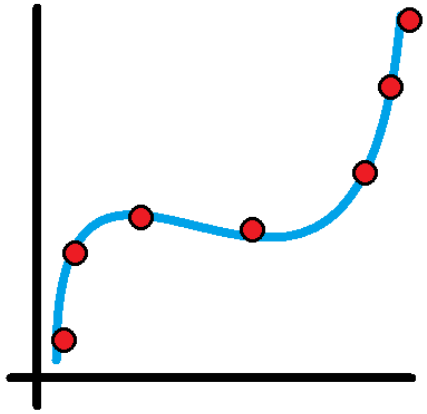
Machine Learning Primitives



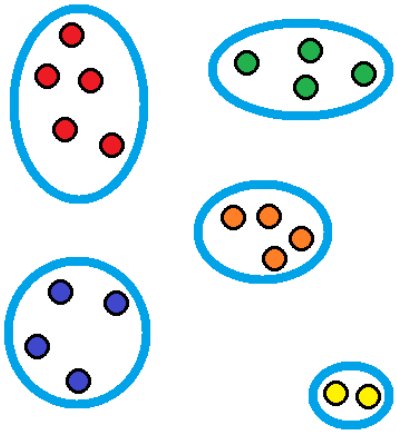
Binary Classification



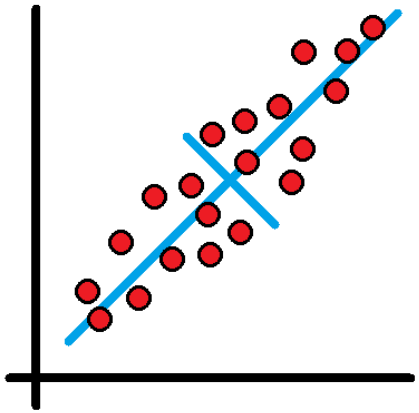
Multi Classification



Regression

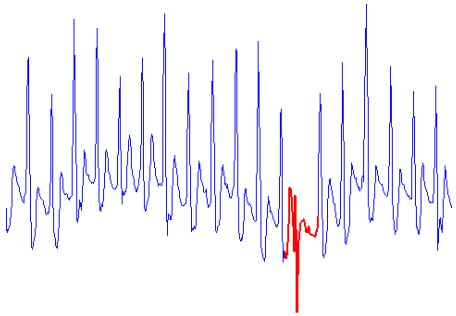


Clustering



Component Analysis

Cyber Security Applications



Anomaly Detection

Intrusion Detection

Entity Profiling

Hardware Verification

Data Analytics

anomalydetectionresearch.com, tomnichols.net, 4shadowapp.com, brinet.com, yellowfinbi.com

Can Machine Learning Help?

- **Predictive** modeling
  - Use system behavior to predict failure
  - Use system descriptions to predict anomalous behavior
  - Use access patterns to assess threat levels

- **Analytic** modeling
  - Building models of anomalies, attacks, failures
  - Identify points of failure in a system
  - Differentiate stochasticity from anomaly

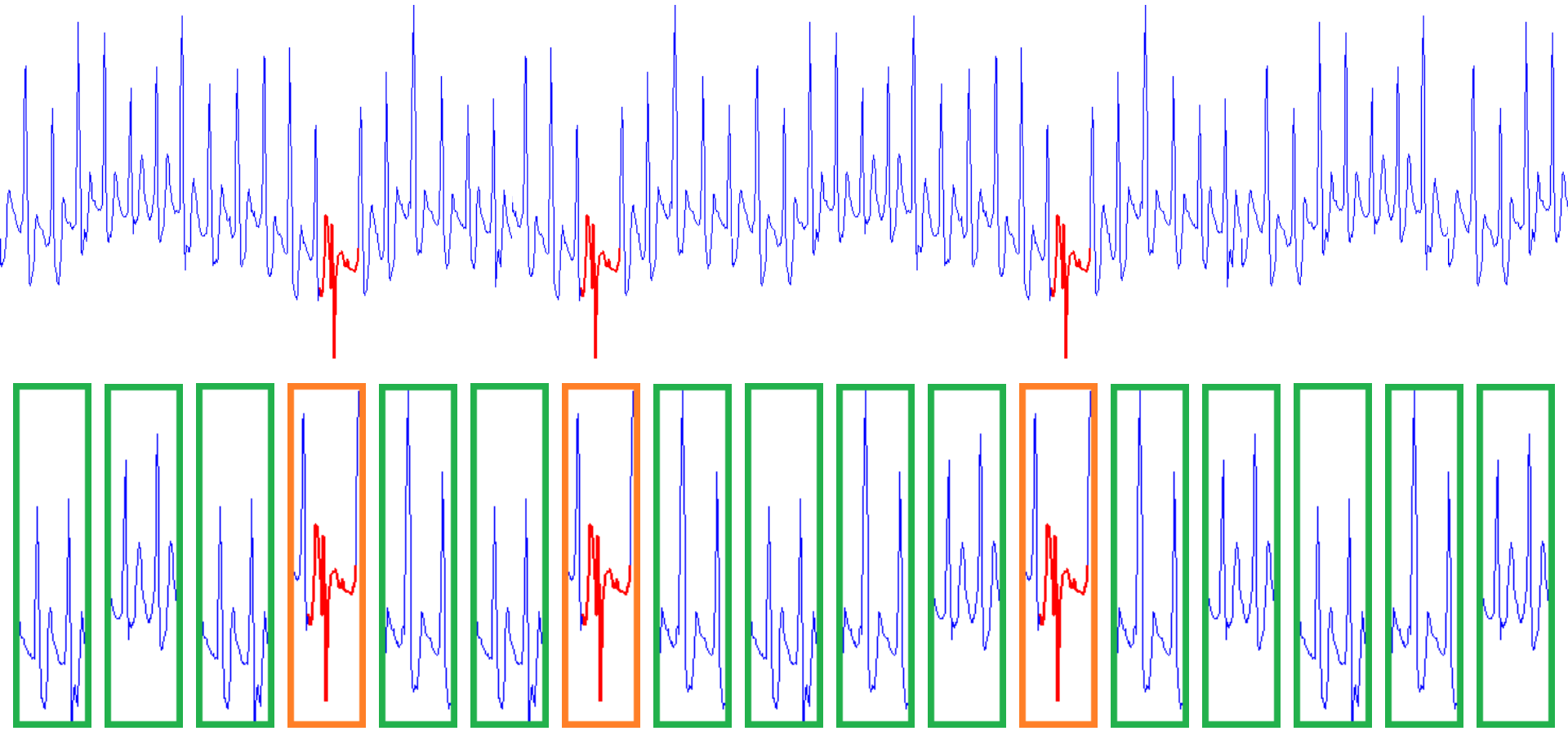
So business as usual? Nope!

- Changes in the **nature of data**
  - Volume of data – large to huge
  - Data access – online or streaming
  - Data distributions – heavy tailed, skewed
  - Noise levels – high, malicious corruptions

- Changes in **application requirements**
  - Extreme precision
  - Cost sensitivity – allowing an attack vs. false alarm
  - Scalability, ease of use and modification

Machine Learning for Critical Applications

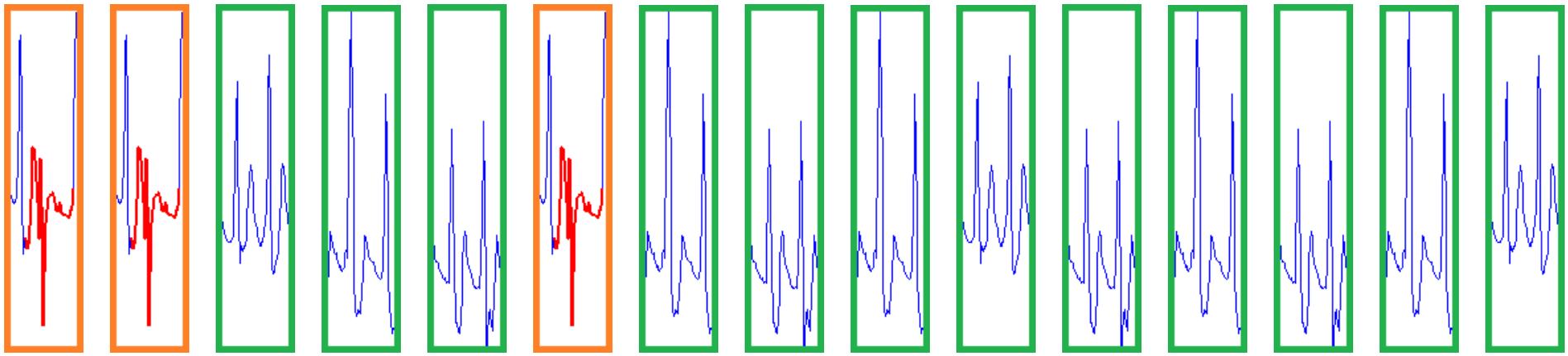
Learning on Streams



Online learning, Stochastic Optimization

Machine Learning for Critical Applications

Learning with imbalanced, heavy tailed data

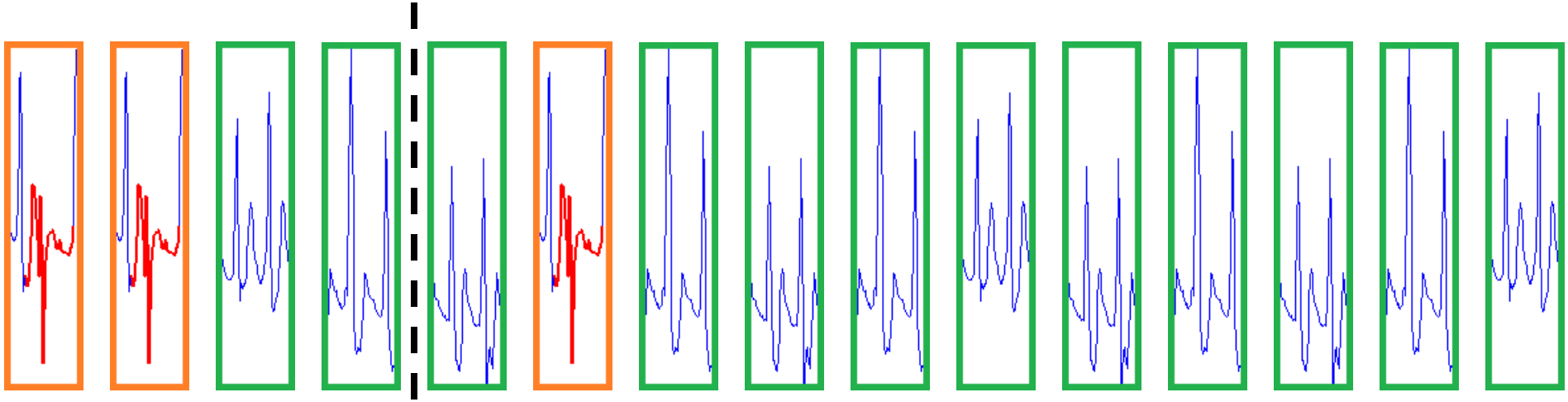


Precision at the Top, Area under the ROC Curve, Precision-Recall Break-even Point, partial AUC, concentrated AUC

Learning to Rank

Machine Learning for Critical Applications

Learning with imbalanced, heavy tailed data



Anomalous

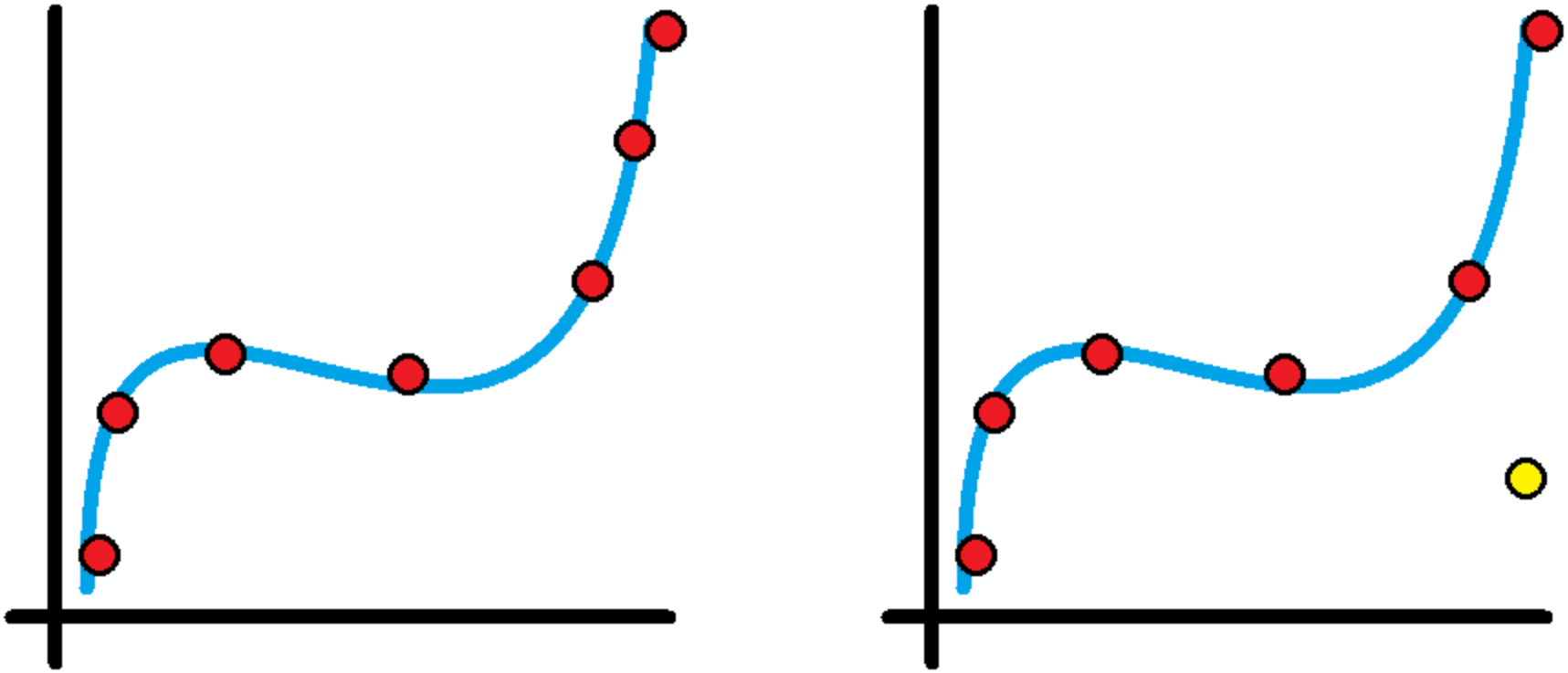
Normal

$$\Psi\left(\mathrm{TPR}, \mathrm{TNR}\right)$$

Multivariate Optimization

Machine Learning for Critical Applications

Learning with (adversarially) Corrupted Data



Robust Classification, Regression, Ranking

Practical Applications

- **Reality**: long stream of corrupted, imbalanced data
- **Critical**
  - proper modeling of data, feature
  - appropriate choice of performance measure
- **Desirable**
  - balance between scalability and accuracy
  - modularity, extendibility

Works in Progress

- Online optimization for ranking tasks
  - **Learn to rank** objects in a stream
  - NIPS 2014 ICML 2015
- Online optimization for learning with imbalanced data
  - **Learn to identify needles** in a stream of hay
  - ICML 2015
- Scalable robust optimization
  - **Learn to regress** in the presence of an adversary
  - NIPS 2015
- Scalable optimization for extreme classification
  - **Multi-label classification** with a million labels
  - NIPS 2015

Thank you

Up-next: **Prof. Piyush Rai**

Learning from Relational Data