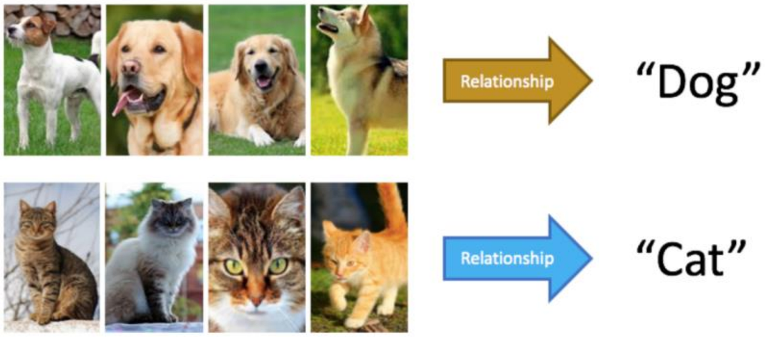
Supervised Machine Learning: “Learning By Example”

Background

Supervised ML learns relationships between known input data and a set of descriptors (e.g. labels) to classify future unknown data.



These relationships involve creating mathematical relationships between the label and a set of “features” that were extracted from the input data. Features represents a measurable property of the data such as the count of words in a file, number of connections to a machine over time, etc. which can be used to discriminate between the labels. A general rule of thumb is the more informative and independent (w.r.t a label type) the better the data can be separated and classified into the correct class.

Needs of a supervised system:

1. A dataset with labels that are relevant in some way to the data
2. Define a *sufficiently informative* feature set to describe the data samples to be classified

Machine Learning Network Behaviors at an Enterprise Scale

With our now highly connected world monitoring a company’s employee behavior for unproductive, suspicious, or malicious data is becoming an extremely difficult “big data” problem.   With a robust method to model the “normal” behavior of a network it is possible to detect behaviors that are not normal such as:

- Insider Threats
- Network Access Policy Violations
- Unhealthy Employee Relations and/or work ethic issues
- Advanced Persistent Threats (APT)

This field is known as “Anomaly Detection” and there exists numerous products and solutions available that leverage machine learning techniques to detect behaviors that deviate from the norm. However common criticism of these solutions is that they typically have a high false positive rate as rare but normal behaviors can be seen as anomalous.

Lack of Network Data Available for Testing

One of the shortcomings of supervised learning is the reliance on labeled data and in the case of network traffic, the amount of data is massive and unlabeled. Many anomaly detectors require a training phase to establish a "network normal" and then the administrator must refine the model by reviewing and correcting any false positives. As actual critical behaviors are quite rare, especially when compared to the amount of normal data available, these types of detectors are hampered by another set of learning issues: data imbalance and false negatives.

- Data Imbalance – A classification problem when the number of observations per label-type is not evenly distributed causing a bias in classification.
    - Ex) Count of normal activity >> Count of malicious behavior.
- False Negative- a test result which incorrectly indicates that a condition or attribute is absent.
    - Ex) System not alerting when a person extracts out customer database

In computer security data imbalance is a real challenge as the behaviors can be extremely complex, sensitive in terms of PII, and/or diverse. Each of these systems will need to be tuned for a specific company network, employee structure, and behavior meaning that there is no one data set available to test all networks, let alone the *thousands of examples needed* to train a sophisticated sensor.

To combat these issues, we believe synthetic data is the solution to combatting data imbalance and PII to help verify the correct feature sets and verifying critical behaviors are appropriately reported.

ExactData Synthetic Data with Ixia's Network Traffic

ExactData has the unique capability to model and simulate enterprise network interactions for any number of employees over the course months to years. This network behavior model generates behaviors that play out over time that is related to past events and relationships to others. This includes employees' emails, website visits, login/logouts, etc. whose interactions are based of the company's work structure, employee social relations, and employee individual personalities. Ixia, an industry leader in network security testing, has partnered with ExactData to bring the sophisticated network data to be played as if it was occurring over the wire.

Synthetic data provides the following benefits over real world network data:

1. No PII concerns
2. Customizable to specific networks, company sizes, and workflows
3. Unlimited generation of data allowing for variations in data and detection difficulty
4. Verifiable performance given the generated labels
5. Allows verification of the effectiveness of features and labels
6. Expandable to new behaviors and interactions

Leveraging Ixia's network packet generation provides:

1. The ability to directly test a network sensor's detection performance
2. Assess the sensors ability to process data at scale and speed
3. Leverages customer's pre-existing Ixia Breakingpoint integrations

Enhanced Use Cases of Synthetic Network Behaviors

As mentioned previously, a robust sensor has the capability to detect a wide range of behaviors, however we believe with our data models that sensors can be enhanced to detect more nuanced behaviors.

- Insider Threats **with early indicators** (such as change in tone when talking to others)
- Network Access Policy Violations **based off role in company**
- Unhealthy Employee Relations and/or work ethic issues **measure morale over time**
- Advanced Persistent Threats (APT) **showing initial breach and cyber kill chain progress**

ExactData partnering with Ixia provides the novel capability of generating unlimited quantities of network data that can be played over the wire with new or existing Ixia Breakingpoint solutions. This sort of data and testing solutions will pave the way to sophisticated, robust, and verifiable security and monitoring solutions.