Data Science & Machine Learning in Cybersecurity

AT&T Business

Author: SaiGanesh Gopalakrishnan
Lead Product Manager, Cybersecurity Solutions, AT&T

Date: 05/22/2017

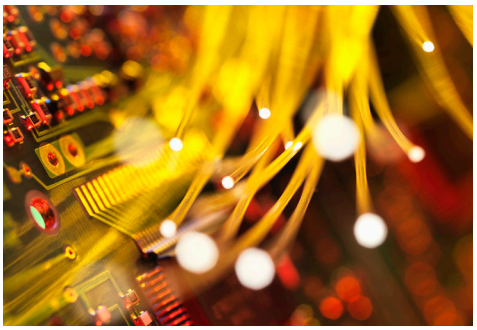


Table of Contents

Executive Summary

Welcome to this primer on Data Science and Machine Learning in the Cybersecurity space for product executives. This paper attempts to introduce "Data Science" and "Machine Learning" concepts at a higher level in terms of methodologies, core algorithms, technologies, and potential benefits to justify product features. Typical security products look for known challenges (infected files, network volume, authorized or unauthorized users, etc.). However, today, we need to look for "Slow and Strategic" threats, which require the power of Machine Learning.

Imagine you are a Chief Information Security / Compliance Officer for a Fortune 500 company. You see the following activity in your network and are puzzled with regard to expected versus malicious behavior:

- "One of your sales engineers logged into the production code box and ***downloaded 700 MB of data"***
- "One of your system administrators logged into the Corporate Finance application and ***downloaded financial PDF reports***"
- "One of your Analysts logged into ***a Customer database simultaneously from US and China at 11 PM***"
- "One of your development leads edited a set of code lines in production from London (where you don't have an office)"

These are complicated activities in a large corporation's environment and need sophisticated data, algorithms, and processes to help track and detect malicious intent at a large scale. This paper provides perspective on utilizing Data Science and Machine Learning to help identify and predict anomalies.

Imagine you are a Chief Information Security / Compliance Officer for a Fortune 500 company. You see the following activity in your network and are puzzled...

Key Concepts and Buzzwords in Machine Learning

Below is a set of key concepts to be aware of before we deep dive into this paper. Becoming familiar with these concepts, will help us to understand high level Machine Learning model results.

Terms / Concepts	
Machine Learning (ML)	Google's definition - Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.
Artificial Intelligence (AI)	AI is a sub-field of computer science and the goal is to enable computers to perform tasks that normally require human intelligence, such as speech recognition, visual perception, decision making and language translation.
Data Scientist	Mostly PhDs and experts in understanding and implementing AI techniques. They help Product Managers validate key use case hypotheses (e.g. Is a type of threat correlated to location?) by mining large data sets and finding patterns (if any) to model for clues in data.
Features	These are attributes that are extracted from data sources to analyze and form a key input to the predictive models. e.g. Time features include – Start / end access times, duration of logon, etc.
Models	Models take features as inputs and they apply simple or complicated mathematical algorithms to come up with a specific outcome for a use case. e.g. Is this normal behavior or an anomaly?
Supervised Learning	The Machine Learning task of inferring a function from labeled training data.
Unsupervised Learning	The Machine Learning task of inferring a function from un-labeled training data.
Deep Learning	A branch of Machine Learning utilizing a set of algorithms that attempt to model high level abstractions in data. e.g. Deep learning provides algorithms and concepts that have the potential to mimic the human brain!
Confusion Matrix	A matrix showing the predicted and actual classifications from the model results.
Accuracy Rate	The rate of correct (or incorrect) predictions made by the model over a data set. Accuracy is usually estimated by using an independent test set that was not used at any time during the learning process. More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with data sets containing a small number of instances.

Boxplot	A boxplot is a standardized way of displaying the distribution of data based on the five-number summary namely: minimum, first quartile, median, third quartile, and maximum.
Zscore	The absolute difference between a data value and it's mean normalized with the standard deviation. Usually, Zscores are used to detect outliers in data, when a data point is above or below some threshold (could be standard deviations). In short, this a measure of how far a data point is from the mean.
TPF	True Positive Rate - % of real malicious events correctly identified
FPR	False Positive Rate - % of legitimate events named as "malicious"
Neural Networks	A computer system modeled on the human brain and nervous system which can be utilized to detect IDS (Intrusion Detection System)
SIEM	Security Information and Events Management (SIEM), is an approach to provide real-time analysis of security alerts generated by applications and network hardware.
LDAP	LDAP (Lightweight Directory Access Protocol) is a software protocol for enabling anyone to locate organizations, individuals, and other resources such as files and devices in a network, whether on the public Internet or on a corporate intranet
PCA	Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

The State of Machine Learning

"Machine Learning" is the process of feeding data to complex algorithms that can, in-turn, automatically learn the patterns and help with decision making. E.g. In the case of cybersecurity, we could use classification algorithms to identify anomalies in the data based on thousands of "features". These would be very difficult for analysts to scan through in order to mine insights manually. Imagine a data set with petabytes of data records and thousands of attributes and the task of mining them to find anomalies in real time! This is where Machine Learning comes in handy. Machine Learning is not just for structured (databases, tables, files, etc.) information, but also for unstructured and streaming data (sensors, PDFs, web logs, etc.) as well. Machine Learning is broadly categorized into two segments; namely, Supervised and Unsupervised learning, which will be explained in forthcoming sections.

The evolution of "Neural Networks" has been key in teaching computers to think and understand the world and objects as we see them while also giving us the speed and scalability. This is a system that helps classify objects and information (Cat vs. Dog) in a very similar manner to our brains (for example, what is a cat versus what is a dog?). Imagine using fingerprints and keyboard patterns to classify anomalies in the cyber space which will feed millions of data points on the user behavior and his/her interactions with the system. Typically, Neural Networks are used for extracting deep hidden patterns in the data not visible to ordinary algorithms. For example, consider a data set containing 99.9% of non-anomalous and 0.1% of anomalous data (which is the practicality in cyberspace today) and

imagine sifting through massive volumes to identify anomaly patterns. This include images, user behavioral characteristics, sensor data, and streaming information to integrate and mine for insights. This will be months of analysis effort before we even identify true positives which leads to the need for deep learning aspects of Machine Learning.

What is Cybersecurity & Anomaly Detection?

Cybersecurity is a set of processes and technologies that enable us to help protect the data and integrity of an enterprise in the context of computing and networks. The purpose of cybersecurity is to help prevent threat attacks from happening by using data, network, user behavior, and policies through continuous monitoring of the computing assets and even better, automatically.

Anomaly detection is a set of technology processes and Machine Learning models to build expected behavior profiles for entities (like users) to form a baseline. Once the baseline is established, the models look for deviations (e.g. Zscore would be an appropriate statistical tool to detect deviations) in entity behaviors to detect possible anomalies. Anomaly detection is amongst the various techniques that are being implemented in Machine Learning to help predict threats with higher accuracy before the threats actually occur.

Key Data Sources in Cybersecurity for Machine Learning

The following are some of the key data sources to be utilized for Machine Learning in anomaly detection and UEBA (User Entity and Behavior Analytics, being discussed later in the document) and forms the core product nucleus.

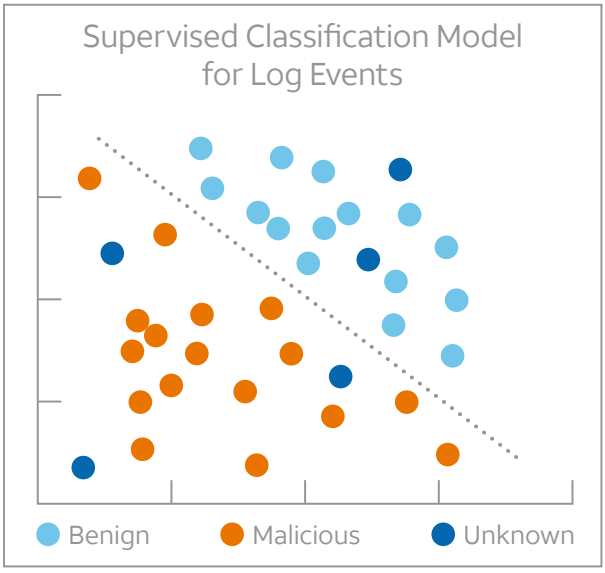
SIEM	This is a great repository for collection of information from server logs, directories and tools. The advantage of using SIEM for security analytics is that, it can help analyze malware and anomalies quickly using unsupervised algorithms utilizing user entity events.
VPN, Proxy, Flow	Utilizing this source helps to analyze large volume transactions by similar users, locations, devices used in the communication. Using machine learning to detect internal attacks in the long run is going to be a key strategy for organizations.
Endpoint	Collects user activity for application, network and cloud based applications distinguishing between users, devices and files.
LDAP	Information on user roles, privileges, org hierarchy, access rights and authentication rules form a basis for developing a user baseline. Supervised machine learning algorithms, can potentially use the baselines to identify anomalies separating it from false positives.

Utilizing Machine Learning for Cybersecurity

The objective of utilizing Machine Learning is to completely move to an automated state where, we don't fine tune rules, thresholds and metrics which are automatically addressed as business changes happen. At AT&T, we utilize machine learning to generate probabilistic scores and baselines for anomalies (rather than a threshold) to reduce false positives. The section below will help understand the top Machine Learning algorithms utilized in Cybersecurity today.

Supervised Machine Learning

Supervised Machine Learning is the process of creating a model where the data it contains is labeled, making it a bit easier for algorithms to learn from the labels. For example, in the context of cybersecurity, log data may be categorized as "Malicious" or "Benign" based on the characteristics they exhibit. By training a classification model on the labeled data, we can feed in unlabeled log records to the model to help us predict the threat category as depicted below.

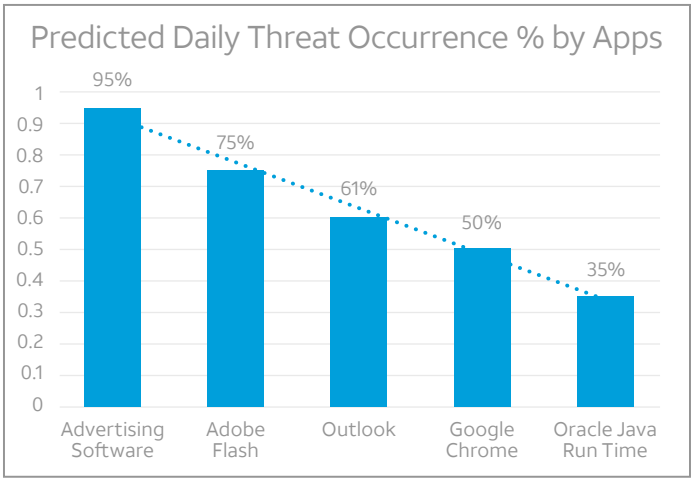


In the above model, we see a third category named "Unknown". This means the input data did not have a label for this category and the model could not recognize whether it is an actual threat. Some of the top supervised learning algorithms that we utilize for classification are listed below. We evaluate these algorithms based on the classification accuracy by a few metrics namely "Confusion Matrix" and "Out-of-bag errors".

- Random Forest – An ensemble learning method for classification, regression which utilizes multitudes of decision trees and outputting of the mode of the classes (classification) and mean prediction (regression).
- Support Vector Machines – Another powerful set of Machine Learning algorithms used for classification of anomalies.
- K-nearest neighbor – Used heavily in pattern recognition and classification, they use the majority votes to align a data point to a specific class.

Apart from classification, Supervised Learning also lets us perform "Regression" analyses that help us estimate relationships among features. The fundamental idea here is to identify a relationship between a dependent variable (what we are predicting e.g. number of malwares or anomalies) and independent variables (e.g. traffic, location, IP, etc.).

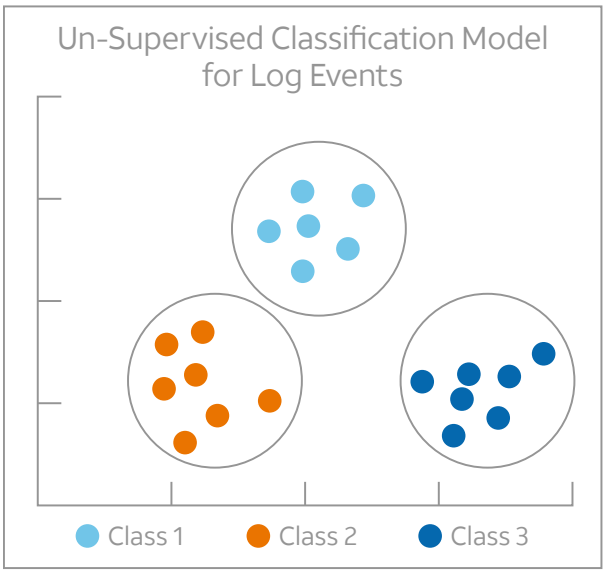
Let's say we have hundreds of applications running in our networks with thousands of servers and nodes in different geographies. Regression analysis can help us predict "Threat Occurrence %" by application which would help us organize our threat mitigation strategy. Here is a hypothetical example of a chart depicting this information to help us strategize our priorities.



Based on the above analysis, we could infer that "Advertising Software" has the potential today to be vulnerable. Drilling down to the various software components and transactions, we could potentially narrow our focus to help predict and mitigate threats.

Unsupervised Learning

Unsupervised Machine Learning is a process of creating a model where the data **does not have labels**, but scans through the data and its attributes to come up with a set of classes the model deems relevant. This is a great technique for anomaly detection, especially when new threat patters arise. The algorithm can learn from what it perceives as normal behavior and can classify characteristics outside of the norm as "Malicious". This is one of the key use cases for Machine Learning in cybersecurity which can provide greater automation to known threats and continuously learn to detect new threats as depicted below.



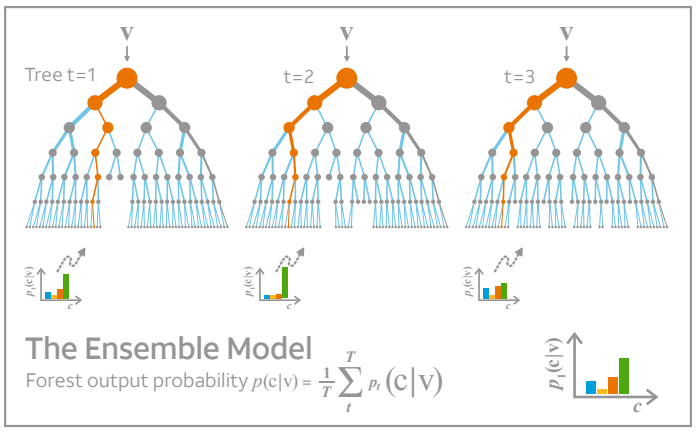
Fundamentally, unsupervised learning helps to identify outliers by baselining normal behavior and finding patterns deviating from the norm. The following are some of the top use cases utilizing unsupervised learning:

- Do we see traffic / login patterns from unusual geographic locations (e.g. from locations where you don't operate or it's not technically possibly to operate?)
- Do we see abnormal amount of network traffic from a single or multiple host?
- Are users accessing resources he or she isn't supposed to access or is it that deviating a lot from peer group?
- Do we know, if we are having compromised account login attempts?

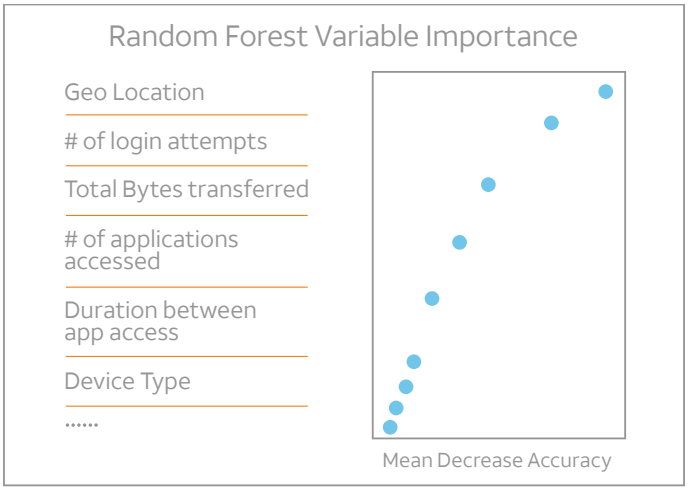
Classification techniques

Let's look at a couple of classification techniques that could be potentially used for Cybersecurity for log analysis and malware detection.

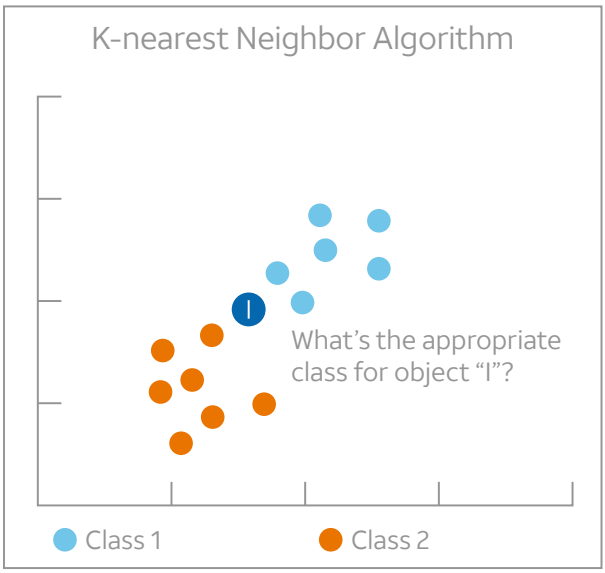
- **RandomForest** – This is an ensemble learning model used either for classification, or regression. This model constructs a multitude of decision trees using a technique called "Bootstrap Aggregation" that produces a mean prediction by combining the results from all trees.



Beyond this, RandomForest also gives us a very useful chart of some of the important features that were used for classification. These charts can be used by SOC analysts to determine the top features that drive anomalies and help mitigate. Below is sample chart that highlights the key features driving anomalies in Cybersecurity with "Geo Location" being the topmost feature followed by "# of login attempts" and so forth.



- **K-nearest neighbors** – This algorithm is primarily used for classification tasks (like RandomForest) in which we identify the nearest data points (neighbors) of the data point in question to determine its class regardless of labels, as shown below.



In the above example, the object "I" is closer to both Class 1 & Class 2 and hence where does it belong? In this case, Class 2 contributes 2 votes and Class 1 just one vote. Based on these findings, it decides to classify the object "I" towards Class 2. This is done by calculating the distance between the objects closer and the one in question using the technique "Euclidean Distance" shown below.

Euclidean distance measuring	$d_E(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$

Role of UEBA in Cybersecurity

What's UEBA?

User and Entity Behavior Analytics, a term coined by Gartner in 2015, helps us with a set of processes to detect targeted attacks, insider threats, and financial fraud utilizing advanced analytics (e.g. Machine Learning). In UEBA, we look for patterns in users, devices, servers, applications, DLP (data loss prevention), IAM (Identity and Access Management), network flow data, etc.), to calculate risk to identify anomalies against the baseline. With the use of Big Data, UEBA is able to look for anomalies across diverse data sets at scale.

In a typical web application environment, we look at the following checkpoints for user profiling in UEBA related information namely:

- **Registration** - Process for signing up new credit cards or new checking accounts.
- **Logins** - For account authentication, time of day, # of attempts, and devices used
- **Transactions** - completing a financial transfer or completing a purchase online
- **Logouts** – User behavior during logouts (do they typically logoff, or close the browser or leave it to expire)

Why should we care about UEBA now?

SIEMs are powerful tools that can aggregate logs to detect anomalies, but they still rely on a simple correlation rules-based approach that is unable to spot lineage of advanced threats. There is a greater possibility for threats to go unnoticed when SIEMs do not have rules to cover all sort of threats. These rules can detect threats in near real–time, but advanced threats can occur over months or even years, which are very complicated to analyze. The power of UEBA lies not in utilizing signatures or rules, but in utilizing advanced machine learning algorithms and risk scoring methods to correlate these events over a long period of time. So, by utilizing a combination of SIEM and UEBA, organizations can be better prepared to improve their threat detection capabilities.

Instead of waiting for a threat to occur and doing a post-mortem on the breach, UEBA guides us to develop an understanding of a user's risk profile and vulnerability when they first engage in a business transaction. In UEBA, we build user risk profiles and customer behavior biometrics (Human vs Bot), as part of the customer journey to become extremely predictive on their risk scores. ***The customer biometric is used to help safeguard the security of customer information, such as to confirm the customer's identity and to help us to know if someone has stolen the information to take over their identity, which is the complex part of the puzzle that UEBA helps us to resolve.***

Per a 2015 Gartner study, "UEBA successfully detects malicious and abusive activity that otherwise goes unnoticed, and effectively consolidates and prioritizes security alerts sent from other systems…organizations need to develop or acquire statistical analysis and machine learning capabilities to incorporate into their security monitoring platforms or services. Rule-based detection technology alone is unable to keep pace with the increasingly complex demands of threat and breach detection."

The UEBA and the Machine Learning Context

UEBA is bit different from a rules-based approach since it uses Machine Learning and advanced analytics for the system to learn and automatically detect anomalies from peer group user profiles in a very short time period. With the right data and Machine Learning models, we can significantly help to reduce the "False Positives" and accelerate the investigation proactively.

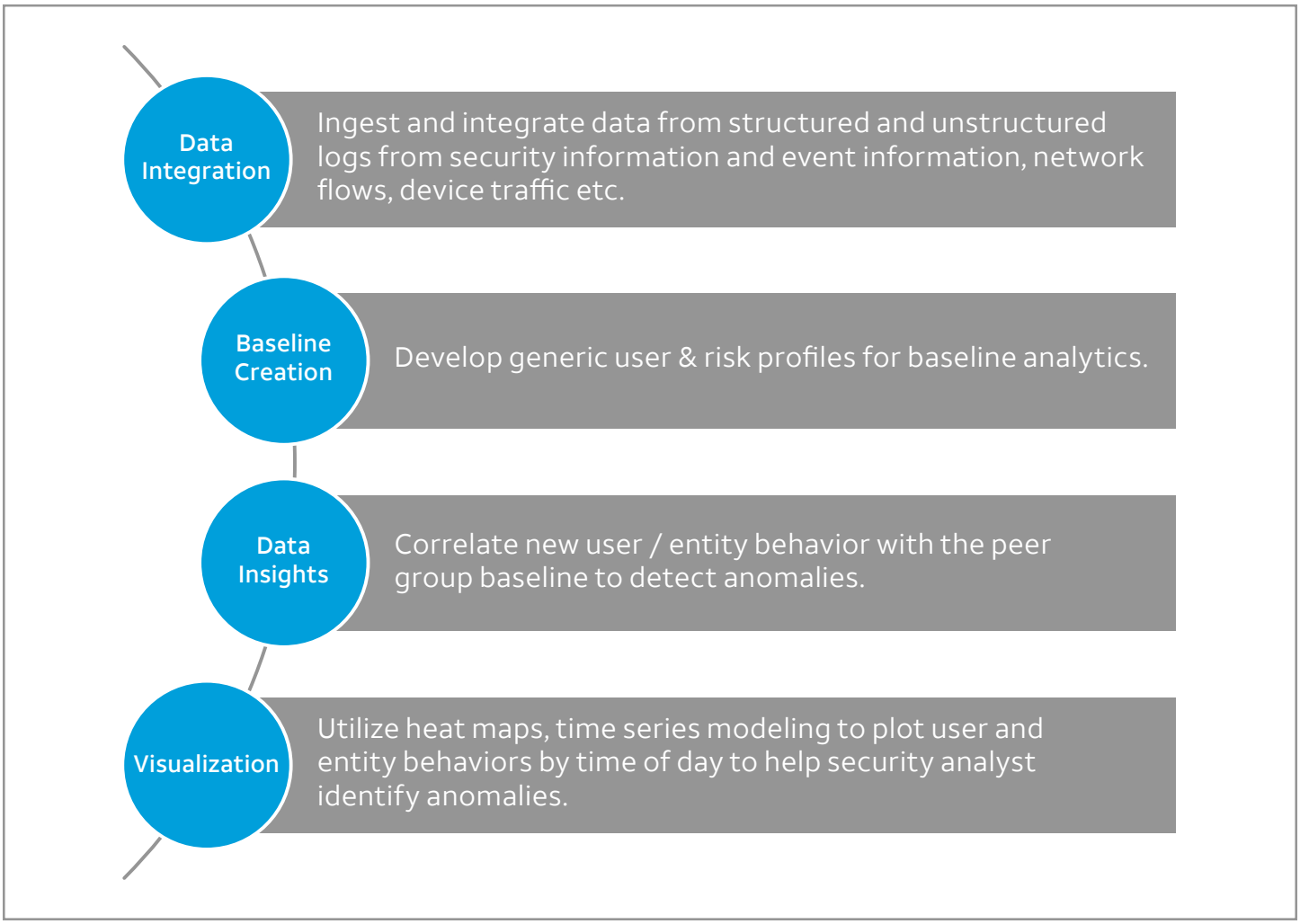
UEBA uses the following key components that utilize Machine Learning techniques to continuously learn and build real-time generic user profiles to detect anomalies. From a Machine Learning context, we could utilize simple SVD (singular value decomposition) algorithms to classify account types (user, service, bots, etc.) and build account behaviors from log files. Typical attributes include: maximum number of connections during peak load, number of connected end-points, applications accessed, location, permission requested (mostly read), etc. We can build initial baseline profiles based on the aforementioned attributes. Apart from account types, we also need Machine Learning to help us classify other entity types, namely between a server node and a user desktop using behavioral attributes based on the activities performed on these assets.

The use of recommendation algorithms in identifying a peer group of users based on their prior activities is a great example of utilizing Machine Learning in the context of UEBA. Imagine if a user accesses a laptop or desktop for the first time with lots of activity (seems suspicious at the face of it), we could potentially correlate this with a peer group (baseline) profile to help reduce false positives significantly.

Here's a set of key use cases (currently being researched and tested), we have been focusing on to help identify anomalies using Machine Learning and Deep Learning. Each of these use cases, are extremely difficult to be detected by normal supervised algorithm and hence we apply deep learning and SVD methodologies to be ahead of the curve.

Use Case	Data Sources	ML Methods	Example
Geolocation Anomalies	Log files, Packet inspection and Active directory	Singular Value, Decomposition, PCA. Auto encoders etc.	Detecting Pass-the-Hash techniques from different locations at odd times.
Data Exfiltration / DLP	Flows data and network session (layer 7), payload	Supervised ML, clustering, network modeling etc.	Identify data exfiltration, based on traffic volume per protocol per use model.
VPN Anomalies	VPN logs, Active directory, Packet Inspection etc.	Singular Value, Decomposition, PCA, network modeling, baselining	Detect VPN user behavior anomalies by location to look for compromised accounts and Pass-the-Hash gaps.

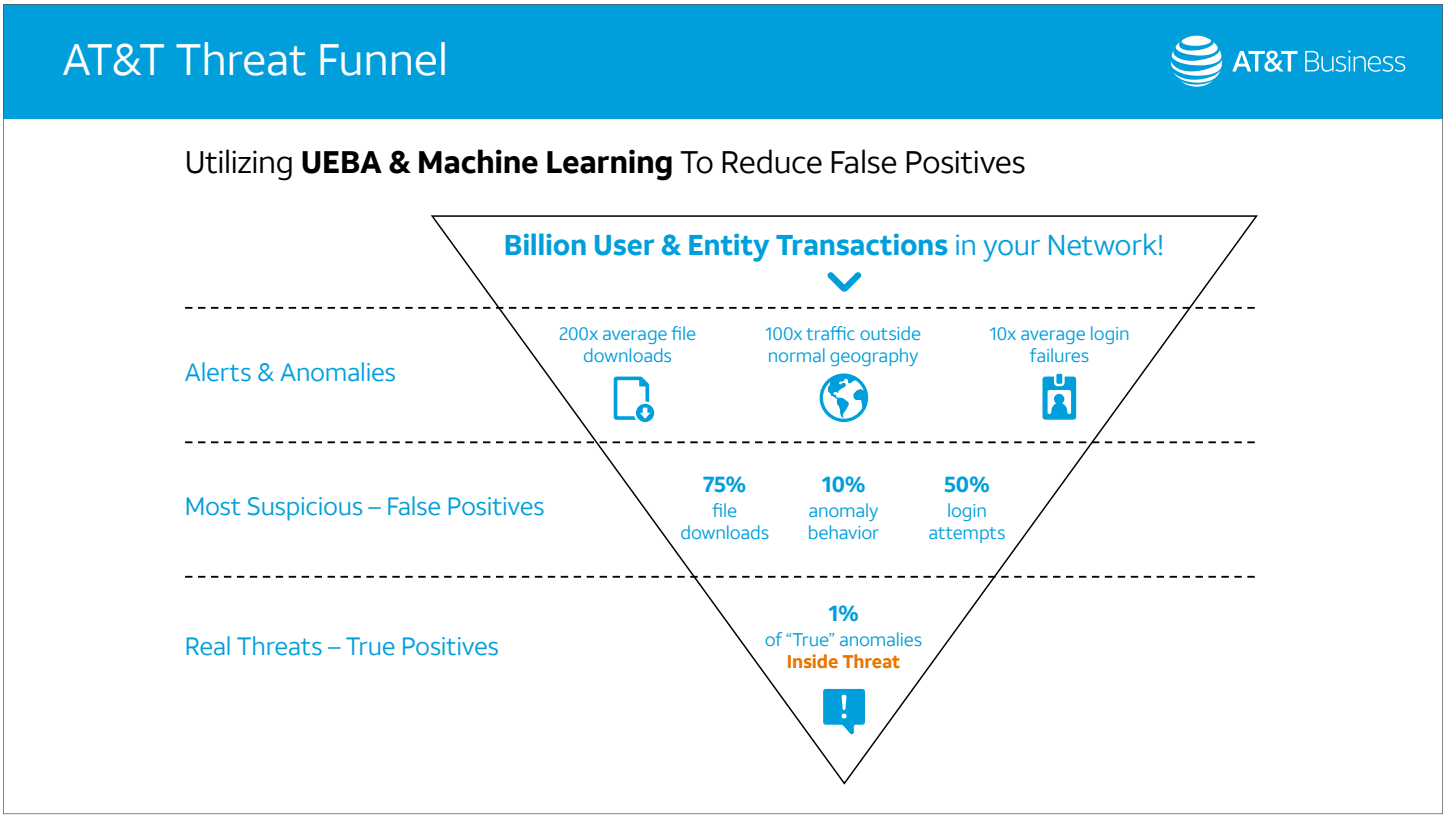
Below is a thoughtful framework for using UEBA and Machine Learning for Cybersecurity and let's apply this for a sample case to help reduce false positives.



Reducing False Positives through UEBA and Machine Learning

Imagine, we build a system with false positive rate of 0.1% and with a true positive rate of 100%, which is great at the face of it. However, let's say that, you see 1 malicious transaction per 20,000 in an hour / day which means, you will have 21 alarms (20,000*0.001+1) out of which one is the real malicious event and 99% of them are false. In this context, we need a robust unsupervised machine learning and UEBA to help analyze all transactions to identify the 0.01% of malicious events with precise accuracy.

Reducing significant false positives has been the top priority for most Cybersecurity enterprises and here is AT&T's approach to utilizing Machine Learning from the "Threat Analytics Funnel".



Let's examine a "DLP" alert with the help of Machine Learning and identify "real insights" to look for "True Positives" using the framework above.

What's DLP? – Data loss prevention (DLP) is a strategy used to see to it that end users (e.g. employees, third party entities) do not send sensitive or critical information outside the corporate network. DLP allows organizations to help detect potential data breaches / data ex-filtration by monitoring, detecting and blocking sensitive data both in-motion and at rest. The following picture helps us understand some of the key "**Data types**" we need to protect as an origination utilizing DLP.

Corporate Data	Financial Data	Customer Data	Personally Identifiable Data
✓ Patents & Designs ✓ Product & Service Price ✓ Intellectual Property ✓ Un-announced merger plans ✓ Legal & Privacy documents ✓ Assets & Valuations	✓ Performa's & Forecasts ✓ Bank Accounts & Payments ✓ Sales Projections ✓ Management Reporting ✓ Vendor Data ✓ Sales Volumes ✓ Customer Volumes	✓ Customer List ✓ Contact List ✓ Buying Patterns ✓ Customer Billing Data ✓ Account Balances ✓ Contract Terms ✓ Location	✓ Name, DOB, Address, SSN ✓ Biometric Data ✓ Credit Card Information ✓ Driver License Number ✓ Health Information

Data Integration - The DLP (Data loss prevention) typically requires Alerts, user identify context, access logs, network traffic, data volumes, location etc. as typical data sources to start with. Collecting and integrating this data in the lake is the first step upon which we perform exploratory analysis to gain data understanding and initial directions.

Baseline Creation – In this step, we analyze and mine data sources by user to understand their behaviors and norms in accessing the above data types (including the purpose) to build generic user profiles. The baselines typically aggregates the frequencies, devices, data types, user rights and sender information giving them a risk score at various time intervals.

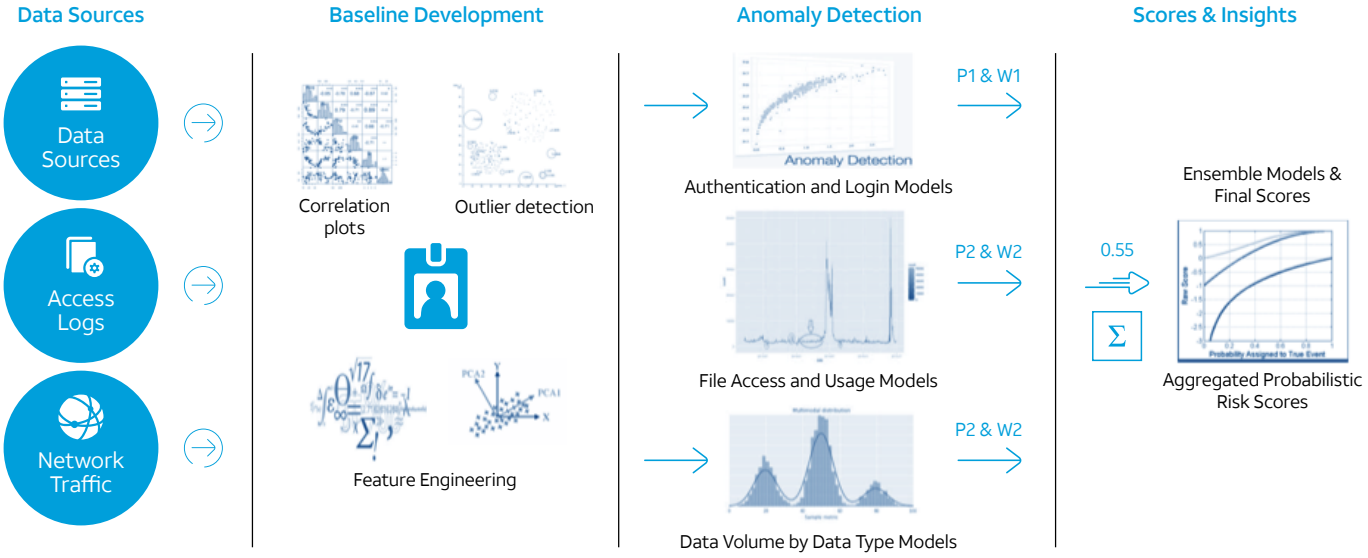
For e.g. the baseline risk score helps us **not to trigger alerts / anomalies** if the data transfer threshold is beyond a certain limit (thereby reducing false positives) and instead the scores act intelligently based on the levels. The risk scores do have a built in confidence interval and based on the range, we could invoke alert API's to the respective system administrators appropriately.

Utilizing the AT&T TMLA Machine Learning framework, below we could see the flow for DLP and the use of Machine Learning to help detect anomalies and finally arrive at a risk score. Our risk scores are a key differentiator in the market which includes the confidence level and assembling based on various anomaly detection models which help to significantly reduce false positives.

AT&T – TMLA Machine Learning Framework

AT&T Business

Use Case: DLP Detection in Action



AT&T – TMLA Machine Learning Framework

AT&T Business

What data is analyzed?	DLP alerts, access logs, Network traffic, User identity context
What Machine Learning algorithms are being used?	Robust z-scores, k-means, PCA to identify which dimensions exhibit the most variance to detect threat.
What's a Risk Score?	A risk score is an ensemble of probabilities and weights coming out of different models and finally giving us the aggregated one.
What insights are being explored?	Different models are being run on different data types (classifications to clustering) and the models helps us to identify which alerts / threats are most important to focus based on risk scores.

Closing Comments

With the introduction of UEBA, there has been a lot of focus on Machine Learning related intelligence, and entities need to decide where to utilize these techniques in Cybersecurity to help reap the benefits. Moreover, entities need to train their analyst teams to interpret Machine Learning results to take actions when buying these products.

I wanted to caution the readers that, you will not become a Data Scientist reading this paper, but it can guide meaningful conversations with data scientists to identify great product features. Good luck!