



Why is Data Security Information so Noisy?

We're always hoping for an easy score. But network traffic is the manifestation of intents; the traffic is there because someone or something has a goal. It might be exchanging email, sharing data or hacking you. In almost all cases, determining the goal by looking at the traffic requires a priori knowledge or assumptions about what the traffic means; it requires information that isn't found in the traffic itself.

The object of a Security Event Manager (SEM) or an IDS/IPS is to derive knowledge from traffic data, and then reduce it to a score. The less knowledge in the process, the less valuable the score will be, which is the reason that administrators have to investigate false positives from network intelligence devices. Heuristics and signatures are two approaches to draw knowledge from data. Anomaly detection pulls patterns from data with little increase in knowledge.

Heuristics

Let's say that you own a restaurant, and it has a security system. You get notified that the front door is open. What can you reasonably infer?

Knowledge	Heuristic	Observation
Door is open.	<none>	Door is open
Door is open.	<ul style="list-style-type: none"> If the door is open then someone has opened the door. 	Someone has opened the door.
Door is open.	<ul style="list-style-type: none"> If the door is open then someone has opened the door. If someone has opened the door then someone has entered the restaurant. 	Someone has entered the restaurant.
Door is open. It is 3 AM.	<ul style="list-style-type: none"> If the door is open then someone has opened the door. If someone has opened the door then someone has entered the restaurant. If someone has entered the restaurant outside normal hours then someone is probably breaking in. 	Someone is <i>probably</i> breaking in.

To make the simple observation that there might be a break-in in progress, we apply heuristics to data; we have derived knowledge. In the end, we can say that someone is *probably* breaking it.

Signatures

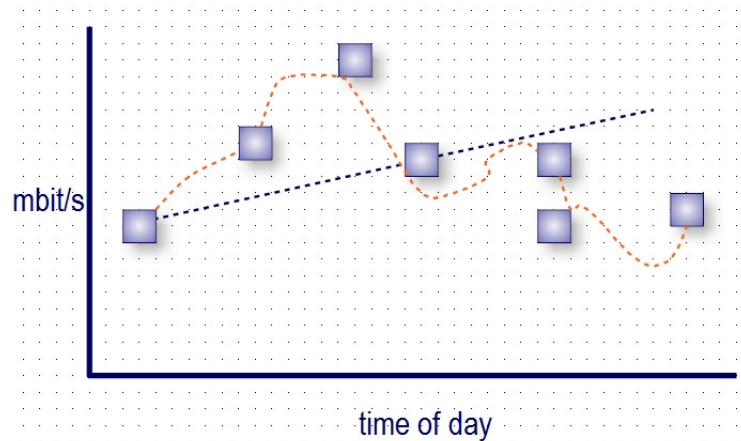
Signatures simplify event recognition for known patterns. They condense multivariate input—essentially the meat of heuristics—into a decision that creates the score. For the restaurant, a signature that said:

“the door is open and it is after hours”

would provide the same result as heuristics. Heuristics are more flexible, but signatures are efficient to process; they’re quick. Of all the possible ways to apply knowledge to complex events, signature recognition probably has the most going for it. But if a signature isn’t sufficiently specific, it can generate noise. Too specific and it might not fire.

Anomalies

Anomaly detectors watch patterns in traffic to see if they look different than training data. The more complex the input, the more complex the model, and the less sanguine its approximation will be when presented with a novel situation; higher order curve-fitting is prone to false positives by its nature.



The Semantic Gap

Semantic derivation is the process of increasing knowledge about the event. *Semantic reduction* is how we produce the score. When we combine anomaly detection, signatures and heuristics and semantically reduce them, the worst of the uncertainty in each comes out. This suggests that the more semantic reduction system that takes place in a SEM, the noisier the results will be.

What *does* provide reliable results? Simple metrics such as checksums on files and recognition of unplanned reboots unambiguously tell you something significant has happened, albeit late. A SEM will highlight activity you would have otherwise missed. But one can never eliminate the noise; there are semantic gaps between what is happening on the network, what the SEM understands, and the indication you receive on back-end; you’re much smarter than your network intelligence tools can ever be.