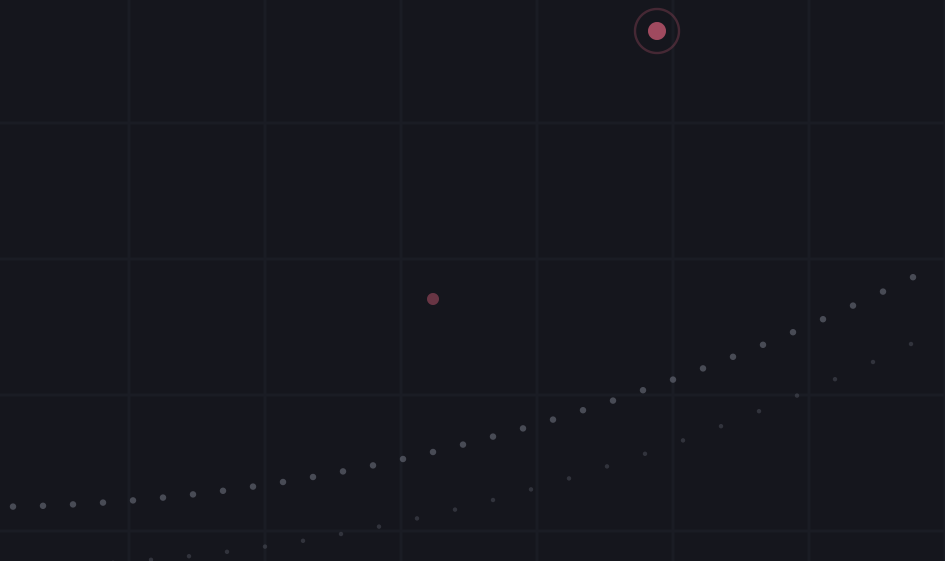
WHITE PAPER

Training Data for Cyber AI

Why models built for cyber operations are only as good as the adversary data underneath them.



Ask a team whose security model underperforms what went wrong and you will hear about architectures, features, and thresholds. Ask what the model was trained on and the room gets quieter. In cyber operations more than almost any other domain, the training corpus is the product: models learn to recognize adversaries from examples of adversaries, and the quality, provenance, and freshness of those examples set a ceiling no amount of architecture clears. This paper is about that ceiling, why security data is unusually hard to get right, and what a corpus curated from direct observation of threat actors looks like.

01 Why security data is different

Most machine learning domains enjoy a stable relationship between past and future. A model that learns cats from last year's photographs will recognize next year's cats, because cats are not trying to stop being recognized. Cyber operations break this assumption at the root. The phenomenon being modeled is an adversary who studies detection and adapts against it, so yesterday's distribution is not a sample of tomorrow's. Every model in this domain is in an arms race with its own training set.

The class imbalance is worse than practitioners expect, and it compounds. Malicious events are a vanishing fraction of observed traffic, which means a model can achieve superb accuracy by learning to say "benign," and many quietly do. Labels, meanwhile, are scarce and expensive precisely where they matter most: confirming that a domain, an IP, or a flow was genuinely malicious takes investigation, and investigation does not scale the way scraping does.

So the industry scrapes. The common training corpora for security models are assembled from public blocklists, shared feeds, and honeypot capture, sources that are abundant, cheap, and structurally compromised. Blocklists inherit each other's contents and each other's errors, so corpora built from them are correlated in ways their consumers never audited. Honeypots see the attacks that fall into honeypots, a biased slice of unsophisticated activity. Feeds lag the adversary by design, listing infrastructure discovered after use, which trains models to recognize last quarter's campaign. And all public sources share one vulnerability that should keep security ML teams awake: the adversary can read them too, and salt them. A model trained on data the adversary can influence has a supply chain problem before it has a modeling problem.

02 Ground truth from observation

The alternative to inherited labels is earned ones. Vigilocity's corpus comes from its own collection: a decade of tracking threat actors as they register domains, stand up command and control, and stage campaigns, followed by observation of that infrastructure in operation, including the victim networks in live communication with it.[1] A domain in this corpus is not labeled malicious because a list said so. It is labeled by what was watched: who registered it, in what acquisition pattern, wired to which infrastructure, serving which campaign, contacted by whom.

Scale matters here, but shape matters more. A three-year window of this collection spans over twelve million domain registrations by four million registrants, and the value is not the row count. It is that the rows are longitudinal, the same actors observed across years, through their bursts, their quiet periods, their operational security mistakes, and their reinventions.

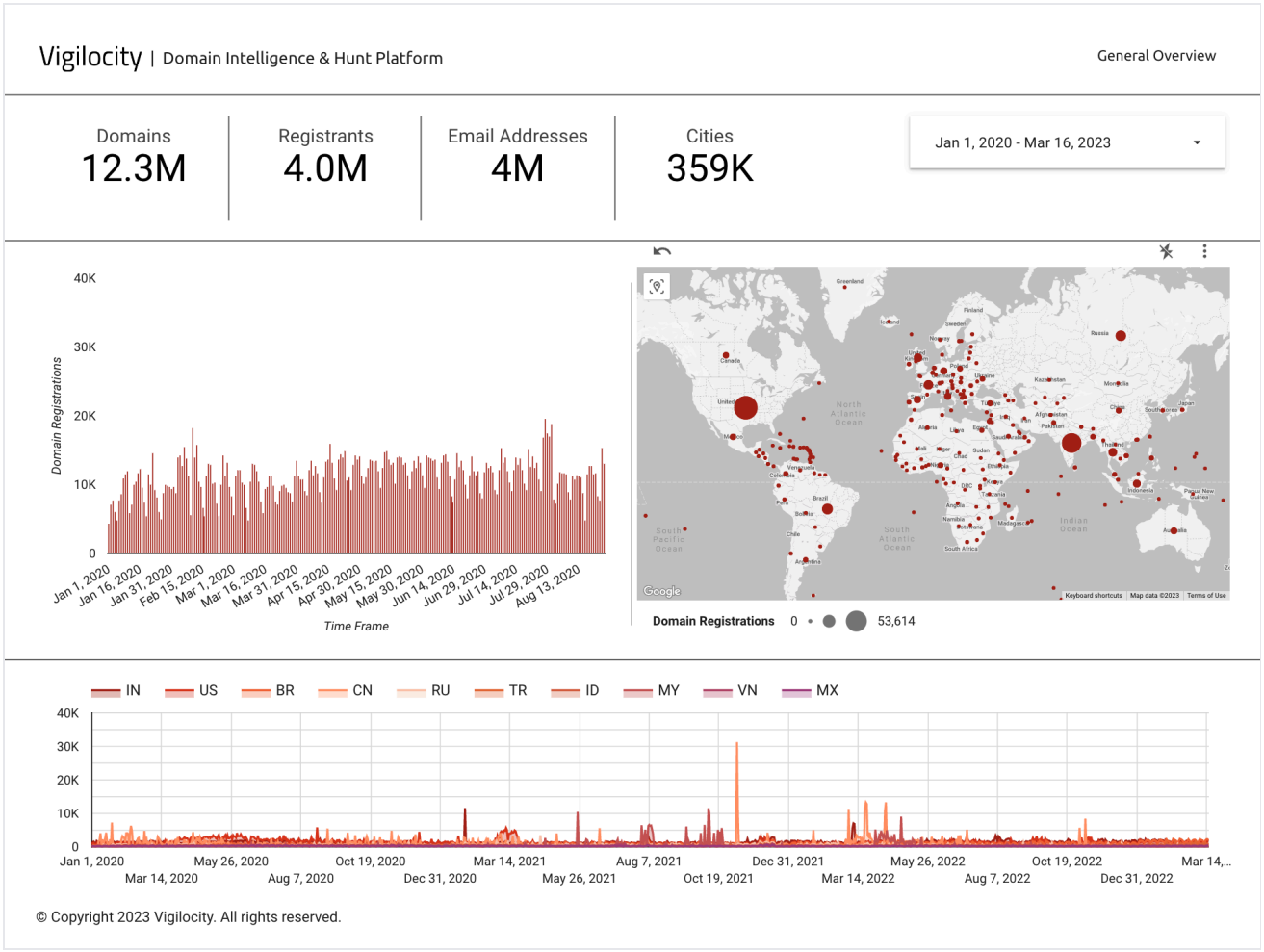


FIG. 1 Three years of global domain registration, colorized by registrant country: the raw material of a longitudinal adversary corpus.

Longitudinal observation is what produces the labels that scraped corpora cannot contain at all. Which of these ten thousand domains were registered by one identity wearing ninety-six names. Which registration burst preceded a campaign by six weeks, and which turned out to be benign bulk speculation, the negative example that keeps a model honest. Which actor's new infrastructure matches the construct of infrastructure they burned two years ago. These are relations across time and identity, invisible in any snapshot, and they are precisely the structure that separates a model that memorizes indicators from one that learns behavior.

03 What quality means, concretely

Teams evaluating training data for cyber AI can reduce "quality" to five questions, in roughly descending order of how often they are asked:

01 **Provenance.** Does every label trace to an observation someone can describe, or to an aggregation of aggregations? If the vendor cannot say how a row earned its label, the model inherits the mystery.

02 **Freshness.** What is the lag between an event occurring and appearing in the corpus? Feeds that lag by weeks train models for a world that has moved.

03 **Context.** Is a malicious domain a string, or a record with registration detail, actor linkage, infrastructure relationships, and campaign membership? Features live in context; strings train string-matchers.

04 **Negative space.** Does the corpus contain hard negatives, the benign bulk registrations and legitimate look-alikes that sit near the decision boundary, or only easy contrast between obvious evil and obvious good?

05 **Contamination resistance.** Could an adversary have influenced the corpus contents? Data the adversary can write to is not training data. It is attack surface.

Curation is the discipline that keeps those answers true, and in this domain curation is analyst work, not pipeline work. Deduplication across correlated sources, adjudication of conflicting labels, redaction of collateral personal data, and the judgment call on ambiguous cases all require people who have watched actual campaigns assemble. A corpus curated by adversary hunters encodes their judgment. That judgment is most of what a downstream model is really learning.

04 What the data trains

Teams put curated adversary infrastructure data under a familiar family of models, and the corpus properties above decide whether those models work outside the lab. Domain classifiers and DGA detectors need the hard negatives and naming-construct depth. Actor clustering and infrastructure attribution need the longitudinal identity structure, the reused email addresses and recycled habits that link "unrelated" campaigns. Early-warning models, the ones that score a registration burst by its resemblance to past pre-campaign staging, need labeled sequences of staging-then-launch, which only exist in a corpus that watched both halves happen. Triage and enrichment models need campaign context so an analyst's queue arrives ordered by consequence rather than recency.

A note on evaluation, learned the hard way and cheap to pass on. Random train-test splits flatter security models, because near-duplicate infrastructure lands on both sides of the split and the model grades itself on memorization. Split by time, train on the past and test on the later past, and split by actor where the data supports it, so the model is scored on adversaries it never saw. Expect the honest numbers to be lower. They are also the only numbers that predict field performance, and a fresh longitudinal corpus is what makes the honest evaluation possible at all.

05 Conclusion

The gold rush in AI is a gold rush for data, and the security corner of it has been panning in a crowded stream: everyone training on the same inherited lists, converging on the same blind spots, lagging the same adversary together. The way out is not more data but different data, examples that were observed rather than aggregated, labeled by investigation rather than inheritance, deep enough in time to expose behavior, and closed to adversary influence. That corpus exists where the observation happened. Vigilocity curates its training data from the same collection that powers Mythic, its breach intelligence platform, and makes it available to teams building AI and machine learning for cyber and counterintelligence operations. The models are yours. The ceiling does not have to be.

06 Sources

1. Vigilocity, "Reverse Attack Surface Analysis: Watching the breach from the attacker's side," 2026, and "Geopolitical Cyber Event Prediction," second edition, 2026. vigilocity.com/resources

ABOUT VIGILOCITY

Vigilocity builds technology for identifying and eliminating threat actor infrastructure, drawing on decades of counterintelligence operations inside adversary networks. Its platform, Mythic, confirms material breaches by observing them from the adversary's side.

BRIEFINGS

Request a briefing

vigilocity.com/contact