For Good Measure
Remember the Recall

DAN GEER AND MIKE ROYTMAN

Dan Geer is the CISO for In-Q-Tel and a security researcher with a quantitative bent. He has a long history with the USENIX Association, including officer positions, program committees, etc. dan@geer.org

Michael Roytman is the Chief Data Scientist at Kenna Security. His work there focuses on cybersecurity data science and Bayesian algorithms. He serves on the board of the Society of Information Risk Analysts as Program Director. He is also a technical advisor in the humanitarian space, having worked with Doctors Without Borders, The World Health Organization, and the UN, and is a member of Forbes Technology Council. He is the cofounder and board chair of Dharma.ai, for which he landed on the 2017 Forbes 30 under 30 list. He holds an MS in Operations Research from Georgia Tech, and his coffee roastery, Sputnik Coffee Company, offers dog treats to all visiting canines. mikeroytman@gmail.com

> When you have eliminated the impossible, whatever remains, however improbable, must be the truth.—*Sir Arthur Conan Doyle*

Sherlock's statement is most often quoted to imply that uncommon scenarios can all be explained away by reason and logic. This is missing the point. The quote's power is in the elimination of the impossible before engaging in such reasoning. The present authors seek to expose a similar misapplication of methodology as it exists throughout information security and offer a framework by which to elevate the common Watson.

There was a time when one might answer the question, "What do you do?" with "computer security," but even five years ago generalists were beginning to be scarce, or, as one of us wrote:

> While some people like to say, "Specialization is for insects," tell me that the security field itself is not specializing. We have people who are expert in forensics on specific operating system localizations, expert in setting up intrusion response, expert in analyzing large sets of firewall rules using non-trivial set theory, expert in designing egress filters for universities that have no ingress filters, expert in steganographically watermarking binaries, and so forth. Generalists are becoming rare, and they are being replaced by specialists. This is biologic speciation in action, and the narrowing of ecologic niches. In rough numbers, there are somewhere close to 5,000 various technical certifications you can get in the computer field, and the number of them is growing thus proving the conjecture of specialization and speciation is not just for insects and it will not stop. [1]

A year ago, Oppenheimer & Co CISO Henry Jiang offered a visual for the specialization state of the security world [2]; it has 86 specialties (and commenters asked for more). Yet among the many specialties that make up security, all recognize the inherent uncertainty created by a sentient opponent, and all are currently grappling with one of two formulations of the same problem: (1) there is too much noise and not enough signal, or (2) there is a shortage of qualified security professionals, a shortage arguably made more acute by specialization, however logical specialization is in the face of security pressure as it now is.

Specialization has not proved to be a panacea; we are still beset by errors. Gorovitz and MacIntyre [3] explored a similar phenomenon in medicine by categorizing errors doctors were making into three types: failures of ignorance, failures of ineptitude (failing to apply knowledge that already exists), and necessary fallibility, a kind of prehistoric black swan that we shall not concern ourselves with here.

In our pursuit of knowledge, we have generated ineptitude. This is not an asseveration about the security industry; it is rather an allusion to a law, an analytic relationship between precision and recall in search problems [4, 5], and it is our claim that "search problems" is what information security is all about.

Explanatory aside:

> Library scientists use *precision* to mean what fraction of search results are actually useful and *recall* to mean what fraction of potentially useful results that the search actually returned. Epidemiologists know precision as *predictive value positive*, what fraction of positive tests actually have disease, and recall as *sensitivity*, what fraction of those with disease will test positive. As may be obvious, making recall (sensitivity) rise makes false positives rise, too. Conversely, as false positives rise, the precision (predictive value of a positive result) falls.

Security tools mostly deal with answering some form of the question, "Does this matter?" In vulnerability management, that question is, "Does this vulnerability pose a risk?" In incident response, that question is, "Was this a malicious event or a false positive?" In threat intelligence, it can be said as, "Is this indicator malicious or not?"

Each of these questions by whatever name is a statistical test, a classifier which vendors have tried to answer since before the time of KDD '99—a contest for "a predictive model capable of distinguishing between legitimate and illegitimate connections in a computer network" [6].

Accuracy, in the technical sense, has little meaning when searching for rare events. If one in a hundred machines is infected, I am 99% accurate when I routinely guess that "none of our machines are infected." Hence, we turn to measures of recall and precision when evaluating how "good" we are as an industry at answering these questions. The hope for all analytic tools is to reduce the number of errors the user makes. Are present-day security tools reducing failures of ignorance and failures of ineptitude sufficiently? We illustrate that the answer is a resounding "no." But there is hope. We also show that by measuring existing models through the lens of precision and recall, small changes to the models can have outsized impact on error reduction.

The perhaps foreseeable consequence of Ian Grigg's 2008 article "The Market for Silver Bullets" [7] has been a maddening proliferation of vendors, paralleling the proliferation of cybersecurity practice specialties. In 2016 alone, venture capital firms laid out $3.1 billion in funding 279 new security vendors [8]. We've "enjoyed" a 25% compound annual growth rate in venture money for cybersecurity over the past 13 years, with over $800 million placed in 2017 Q4 alone, but when a typical organization uses 50+ vendor products at once, the output of that instrumentation means an overload in the volume, velocity, and variability of the data that describes the ground truth we seek to classify. It is no surprise that most alerts are never examined. In the course of developing security tools, defenses, and processes, we as an industry have made one simple miscalculation—we have attempted to output truth, aka results, but instead have output a vast amount of noise and have overloaded our most precious resource—the security professional's time.

Generally speaking, security instrumentation seeks to improve models on its own grounds—by reducing false positives or by increasing the search space. While either approach is logical, it means that vendors always start their analyses at the theoretical level, move to the population level, and progress steadily towards the customer's environment. The result is overwhelming, and every additional product installed adds to the problem. It would be better to optimize for the capacity an organization has at hand and construct models with the feasible in mind. So far as the present authors know, every CISO survey ever taken has essentially found cybersecurity to be a lemons market, one where the buyer can't readily tell a low-quality product from a high-quality one. Investopedia then reminds us that "Ironically, [a lemons market] creates a disadvantage for the seller of a premium [product], since the potential buyer's asymmetric information, and the resulting fear of getting stuck with a lemon, means that he is not willing to offer a premium price even though the [product] is of superior value" [9].

Eliminating the Impossible

We offer a solution widely deployed in practice in other fields and talked about in the parlor rooms of security. If security tooling were to focus on analyst enablement, the approach to testing might be altered to resemble something more akin to medical practice—cost-effective multi-stage testing and process termination (see [10] for our prior work on testing)—multi-stage so as to be able to optimize test performance without incurring side-effects, and process termination when no therapeutic difference would follow from sharper diagnostic detail even if that detail were available for free.

Rare does not mean malicious, and building models specifically for very low base rate maliciousness means there is very little chance your positive test results are true positives. If the base rate of a non-malicious event, vulnerability, or indicator is high, we can be fairly certain that our test will categorize a non-malicious event as such—whatever remains after that is a new search space, where the base rate of "malicious" to "benign" is more evenly balanced and hence lends itself better to a second test, one where precision is the goal. In short, the first tests you must apply are not the ones that identify the malicious but the ones that identify the benign. It is conceivable that in a specialist-heavy field, recall is always > precision, perhaps because specialization increases hourly rates.

A Worked Example

Consider ACME, Inc., a fictitious organization constructed by sampling the data set we have at hand: Kenna Security's vulnerability scan data [11]. It contains 8,551,837 assets, 293 organizations, a median vulnerability count per asset of 116, a median monthly close rate of 25 vulnerabilities per asset, and a median monthly open rate of 20 vulnerabilities per asset (hence a net reduction of five vulnerabilities per asset per month).

The fictitious ACME has 10,000 assets; they vary from load balancers to Linux boxes to printers, and so forth. ACME has 1,160,000 vulnerabilities (116 x 10,000). ACME can remediate 250,000 vulnerabilities in a month (25 x 10,000) during which time another 200,000 vulnerabilities will be released (20 x 10,000), i.e., ACME has the capacity to reduce total vulnerabilities by 50,000 per month (5 x 10,000). ACME's goal is to remove the riskiest vulnerabilities from the organization's environment, so they turn to the 28% of the in-the-wild detected vulnerabilities that are ranked "critical" by CVSS, meaning 324,800 (1,160,000 x 0.28) vulnerabilities are ACME's first concern.

But 87.8% of those CVSS criticals are false positives (prior work at [12]), so ACME's meaningful effort towards security is limited to 39,626 (324,000 x (1 – 0.878)) vulnerabilities that are CVSS true positives. This number (39,626) is well within ACME's remediation capacity (39,626 < 50,000), but only if the remediation is somehow aimed only at true-positive vulnerabilities. However, if the level of effort required to remediate 50,000 vulnerabilities is spread across all those 324,800 vulnerabilities marked as critical, 85% (1 – 50,000/324,800) of ACME's investment will yield no useful result. Not only that, there will still be 33,525 ((324,800 – 50,000) x (1 – 0.878)) unremediated true positive vulnerabilities extant. They go on next year's budget…

Turning to the threat and incident use cases (and using the base rates in BalaBit's contextual security intelligence report [18]), ACME would collect about 6.78 billion raw log entries per month, process about 34% of those, getting it down to 2.26 billion processed log entries, and receive 17,300 alerts per month—one alert for each 130,635 processed log entries. ACME has the capacity to investigate 34% of those, that is to say 5,900 (17,300 x 0.34) alerts. They incur a false-positive rate of 18% while taking an average of seven minutes to decide whether an alert is or is not malicious, 688 (5,900 x 7/60) hours of work of which 124 (688 x 0.18) hours is wasted. ACME will correctly classify and investigate 4,800 (5,900 x (1 – 0.18)) of the 17,300 events a month, neglecting 9,363 ((17,300 – 5,900) x (1 – 0.18)) actually malicious events because that would require an additional 1,330 ((17,300 – 5,900) x 7/60) hours of labor to get to. If the organization could handle all 17,300 alerts, 363 (17,300 x 0.18 x 7/60) hours of their labor would be spent on false positives.

Capacity Optimality

The following should be treated as axioms:

1. We are data rich and signal poor.
2. Multi-stage testing cost-effectively increases both precision and recall.
3. Analyst time is the capacity constraint for most security problems (and Cybersecurity Ventures predicts 3.5 million unfilled cybersecurity positions in 2021).

When we say "Remember the Recall" in the alert scenario, "recall" means the percentage of alerts investigated that are in fact malicious. In the vulnerability problem, "recall" means the percentage of vulnerabilities that are identified as worth actually fixing. In both alerting and in vulnerability remission, false positives burn analyst time, our most precious resource. But suppressing false positives is not good enough to be "the" answer [13]. We have to go multi-stage.

The first-stage test has to find and dismiss absolutely the maximum number of benign markers be they alerts, vulnerability notifications, or whatever. This test has to be cheap, which is to say automated. It has to have no false negatives, that is, whatever it says is benign has to be benign. Epidemiologists call this "specificity." In our ACME example, one alert from 130,635 processed log entries illustrates strongly reduced search space—discarding the benign as fast as is possible, and there's a lot of benign to discard.

Where the first-stage test exists to throw out every datum it can so long as there are no false negatives, the second-stage test exists to select every datum it can so long as there are no false positives. The second stage is the analyst, the person with that seven-minute budget for selecting true, not false, positives. His or her tools can be much better if, and only if, the analyst plus tool combo is presented with a search space with the benign removed, that is the second stage can really be focused on recall (sensitivity). Medicine is riddled with this technique [14]. Legal document review [15] and payment fraud [16] are already there, too. And for those who want academic backup, see [17].

This framework is necessary to understanding the current state of security. We exist in a dual-stage testing regime. We are subject to a low prevalence (rare event) environment. To act rationally in this scenario, the first test must remove as many false negatives as it can. This necessarily implies automation in hopes of increasing the analyst's seven minutes to a more reasonable figure. To act with real foresight is to look to methods that automate the second test as well, saving analyst time for the highest quality, pre-cleaned data we can provide.

Assuming that our first test has, as we suggest, high specificity, it is then safe(r) to automate and bias the second test towards recall—meaning we work to solve failures of ignorance. But if we can automate the second test, we can then increase the amount of time the analyst can spend deciding—meaning we are working to solve failures of ineptitude. Perhaps then, and only then, will we get enough minutes back to spend those minutes chasing those rare birds, the black swans.

References

[1] D. Geer, "Trends in Cyber Security," Nov. 6, 2013: geer.tinho.net/geer.nro.6xi13.txt.

[2] Cybersecurity domains: www.linkedin.com/pulse/map-cybersecurity-domains-version-20-henry-jiang-ciso-cissp.

[3] S. Gorovitz and A. MacIntyre, "Toward a Theory of Medical Fallibility," *The Hastings Center Report*, vol. 5, no. 6 (December 1975), pp. 13–23: www.jstor.org/stable/3560992.

[4] L. Egghe, "The Measures Precision, Recall, Fallout and Miss in Function of the Number of Retrieved Documents and Their Mutual Interrelations," *Information Processing & Management*, vol. 44, no. 2 (2008), pp. 856–876: doclib.uhasselt.be/dspace/retrieve/22360/measures%202.pdf.

[5] S. A. Alvarez, "An Exact Analytical Relation among Recall, Precision, and Classification Accuracy in Information Retrieval," Boston College, Technical Report BCCS-02-01, 2002: pdfs.semanticscholar.org/d8ff/71a903a73880599fdd2c7be12de1f3730d29.pdf.

[6] "KDD Cup 1999: Computer Network Intruder Detection": www.kdd.org/kdd-cup/view/kdd-cup-1999.

[7] iang.org/papers/market_for_silver_bullets.html.

[8] Cybersecurity Ventures: cybersecurityventures.com/cybersecurity-market-report.

[9] "Lemons Problem," Investopedia: https://www.investopedia.com/terms/l/lemons-problem.asp.

[10] D. Geer, "Testing," *;login:*, vol. 39, no. 5 (October 2014): www.usenix.org/system/files/login/articles/login_1410_12_geer.pdf.

[11] Kenna Security, "How the Rise in Non-Targeted Attacks Has Widened the Remediation Gap," September 2015: www.kennasecurity.com/wp-content/uploads/Kenna-NonTargetedAttacksReport.pdf.

[12] D. Geer and M. Roytman, "Measuring vs. Modeling," *;login:*, vol. 38, no. 6 (December 2013): www.usenix.org/system/files/login/articles/14_geer-online_0.pdf.

[13] A. Mokarian, A. Faraahi, A. G. Delavar: "False Positives Reduction Techniques in Intrusion Detection Systems-A Review," *International Journal of Computer Science and Network Security*, vol. 13, no. 10 (October 2013): paper.ijcsns.org/07_book/201310/20131020.pdf.

[14] J. Liu, F. Chen, H. Yu, P. Zeng, L. Liu, "A Two-Stage Bayesian Method for Estimating Accuracy and Disease Prevalence for Two Dependent Dichotomous Screening Tests When the Status of Individuals Who Are Negative on Both Tests Is Unverified," *BMC Medical Research Methodology* (September 2014): bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-110).

[15] http://www.nytimes.com/2011/03/05/science/05legal.html.

[16] J. Markoff, "Armies of Expensive Lawyers Replaced by Cheap Software," *New York Times,* March 4, 2011: https://pdfs.semanticscholar.org/2259/447c4ee67017999cbe0a539b86185e263eec.pdf.

[17] M. K. Buckland and F. C. Gey, "The Relationship between Recall and Precision," *Journal of the American Society for Information Science*, vol. 45, no. 1 (January 1994), pp. 12–19: www.researchgate.net/publication/220433788_The_Relationship_between_Recall_and_Precision.

[18] https://pages.balabit.com/rs/balabititsecurity/images/BalaBit-eCSI-magazine-201503-10-13.pdf.