# Bridging the Gap: A Linear Algebra Unit for Critical Infrastructure Defense

Donna Beers
Department of Mathematics
Simmons University
Boston, Massachusetts
e-mail: donna.beers@simmons.edu
Clifton P. Morrow
Taylor Business Institute
Chicago, Illinois
e-mail: clifton.morrow@tbiil.edu

***Abstract*—The critical infrastructure sector faces a compounding crisis: a widening supply-demand gap for skilled cyber-defenders coinciding with a paradigm shift in offensive tradecraft. Adversarial campaigns like Volt Typhoon have moved beyond deploying malicious code to living off the land – abusing legitimate system tools to evade detection. This widening offense-defense gap renders traditional signature-based tools insufficient, necessitating a workforce capable of behavioral anomaly detection. However, current training often limits analysts to tool identification, leaving them ill-equipped to analyze novel threats. This industry report introduces a modular training unit designed to bridge these strategic gaps. We propose a two-fold, complementary approach aligning linear algebra with the National Institute of Standards and Technology (NIST) Cybersecurity Framework: first, the Geometric Approach (Detection), where trainees use k-Nearest Neighbors (k-NN) on a dataset to map benign traffic topology, modeling the geometry of automated alerting; second, the Algebraic Approach (Hunting), where trainees apply Singular Value Decomposition (SVD) to identify pattern of life deviations, modeling the proactive hunting required for advanced persistent threats. By grounding Artificial Intelligence (AI) concepts in their mathematical roots, this architecture aims to produce a workforce capable of dissecting and trusting the algorithms protecting critical infrastructure.**

***Keywords-cybersecurity education; workforce development; anomaly detection; threat hunting; linear algebra; critical infrastructure defense.***

## I. Introduction

State-sponsored campaigns like Volt Typhoon [1] represent a fundamental shift in threats and countermeasures in critical infrastructure systems, moving beyond malware to living off the land techniques that evade signature detection. To counter this threat, our work leverages Artificial Intelligence and Security, specifically applying Artificial Intelligence (AI) for threat and anomaly detection through the mathematical lenses of k-Nearest Neighbors (k-NN) [2] and Singular Value Decomposition (SVD) [3]. The nature of the threat and the mathematical content of the tools create a significant workforce capability gap where the complexity of modern threats outpaces the cognitive tools available to junior analysts, a problem only partially addressed by existing certifications. Furthermore, by elevating analyst capability from rote memorization to first-principles understanding, this unit directly addresses usability and awareness in secure systems, ensuring the workforce can effectively operate these advanced defense architectures.

To address these capability gaps, we propose a modular training unit that satisfies three core operational criteria. First, the solution is understandable to early-career analysts by utilizing a geometric-first pedagogy that visualizes high-dimensional anomalies before introducing algebraic formalism. Second, the approach is defendable as a necessary evolution of tradecraft; relying on black-box tools is a liability against living off the land attacks, whereas first-principles mathematical analysis provides a robust audit trail for detection logic. Finally, the framework is feasible for rapid institutional adoption, relying on open-source tools (Python, Jupyter) and privacy-neutral synthetic datasets that require no proprietary infrastructure.

In alignment with the conference targets, this work presents an architectural solution to the workforce gap, supported by a practical implementation of a modular curriculum. While the theoretical foundations of k-NN and SVD are well-established, our contribution lies in the novel architectural synthesis of these methods for critical infrastructure defense.

In Section I, we note gaps in the preparation of early-career cybersecurity analysts. The goal of our module is to bridge those gaps.

The remainder of the paper is organized as follows: In Section II, we note that the Computing Technology Industry Association Cybersecurity Analyst (CySA+) [4] and the SANS Institute GIAC Certified Intrusion Analyst (GCIA) [5] certifications require arithmetic, and existing curricular modules abstract away the mathematics. In Section III, we present a case study demonstrating our approach. In Section IV, we evaluate the viability of this instructional architecture against key industry metrics. In Section V, we compare our approach with prior art and propose future work.

## II. Related work | Methods

*1) The Analytical Gap in Technical Certifications:* While the National Institute of Standards and Technology (NIST) Cybersecurity Framework [6] emphasizes the continuous function of detection, standard industry curricula for security practitioners leave a critical gap in algebraic fluency. Highly technical, analyst-focused certifications—such as CySA+ and GCIA –

represent the industry standard for training Security Operations Center personnel. However, their curricula predominantly focus on rule-based heuristics, packet-level signature matching, and the operational usage of Security Information and Event Management dashboards.

In these technical tracks, behavioral anomaly detection and machine learning are frequently introduced only conceptually, treating algorithmic defense as a proprietary black-box appliance. Consequently, trainees learn to interpret alerts but lack the linear algebra required to understand the underlying manifold geometry [2]. As adversaries increasingly deploy adversarial machine learning and protocol mimicry [7] to evade standard signatures, defending critical infrastructure requires moving analysts from operational monitoring to algorithmic engineering (Bloom's Taxonomy Levels 4 and 5: Analyzing and Evaluating [8]). By forcing students to manually calculate eigendecompositions and geometric thresholds, this instructional architecture provides direct pedagogical insight into the mechanics of false positives and false negatives, empowering the future workforce to actively tune detection models rather than blindly trusting opaque vendor alerts.

*2) The Mathematical Gap in Curricula:* Existing curricular modules in the CLARK repository, such as Serra's comprehensive guide to anomaly detection [9], focus on the implementation of high-level algorithms like isolation forests and Support Vector Machines (SVM). While excellent for application, these modules often abstract away the underlying mathematical machinery. Our work differentiates itself by focusing specifically on the linear algebra foundations (SVD and geometric topology) that precede these advanced algorithms, filling the pedagogical gap between basic mathematics and black-box AI application.

*3) The Risk of Black-Box Models:* Engineers dislike black boxes they cannot fix. From an engineering perspective, black-box models introduce unacceptable liability in safety-critical systems. As noted in NIST Interagency Report (IR) 8312 [10], trust in AI requires that outputs be "meaningful" and "explainable" to the operator. Engineers do not reject AI out of prejudice; they reject it out of adherence to these rigorous risk management principles. Many current AI security tools are black boxes. When they throw a false positive, the analyst cannot fix it. Linear algebra is the screwdriver. Beyond serving as a mathematical theorem, SVD provides a mechanism to 'tune the noise filter' (finding $\sigma$ thresholds). Similarly, k-NN acts as more than a mere topology; it provides a framework to 'calibrate the sensitivity' (by choosing k and the distance metric). We frame linear algebra not as abstract theory, but as the component engineering of the detection engine. Just as a mechanical engineer must understand thermodynamics to tune an engine, a cyberdefender must understand the SVD factorization to tune the signal-to-noise ratio of an anomaly detector.

## III. Experimental Design and Case Study

"Stop whatever you're doing. Look at it from the other side" [11]. We designate our synthetic dataset the Janus Matrix, named after the Roman god of beginnings and transitions who possessed two faces looking in opposite directions [12]. Just as the mythological Janus held the key to distinct realities, our dataset is designed to be unlocked via two distinct mathematical perspectives: the geometric topology of k-NN and the algebraic variance of SVD.

### A. Experimental Design

In alignment with the taxonomy established by Chandola et al. [13], our architecture adopts a hybrid detection strategy. We utilize k-NN to detect proximal anomalies (points distant from local neighborhoods) and SVD to detect spectral anomalies (points violating the global correlation subspace). This ensures the analyst can identify threats that might be invisible to a single mathematical modality.

### B. Case Study

To demonstrate the mathematical mechanics of anomaly detection, we construct a Janus Matrix ($X \in R^{7\times 2}$) using a deliberately constrained micro-slice of the Knowledge Discovery and Data Mining (KDD) Cup 1999 dataset [14]. While modern datasets like UNSW-NB15 [15] are necessary for benchmarking production AI, their high dimensionality makes them pedagogically opaque. We specifically select the KDD '99 dataset because its HTTP traffic features (src_bytes vs. dst_bytes) provide an intuitive, rank-1 linear correlation (Request $\propto$ Response). By constraining the matrix to $7 \times 2$, learners are able to perform the k-NN distance calculations and the SVD eigendecomposition by hand, successfully converting a black-box machine learning algorithm into a glass-box mathematical exercise [16].

Furthermore, these specific features establish a clear adversary threat model: a **Signature-Proof Data Exfiltration** attack. In this scenario, the adversary tunnels data out of the network by strictly mimicking valid minimum and maximum byte bounds to evade standard firewall heuristics. However, in doing so, they break the fundamental structural asymmetry of normal web browsing, allowing algebraic detection methods to succeed where geometric thresholds fail.

$$X = \begin{bmatrix} 220 & 1200 \\ 240 & 1350 \\ 260 & 1500 \\ 280 & 1650 \\ 300 & 1800 \\ 320 & 1950 \\ \mathbf{290} & \mathbf{1300} \end{bmatrix} \begin{matrix} \leftarrow \text{A} \\ \leftarrow \text{B} \\ \leftarrow \text{C} \\ \leftarrow \text{D} \\ \leftarrow \text{E} \\ \leftarrow \text{F} \\ \leftarrow \text{Q = query point to analyze} \end{matrix}$$

Glancing at the numbers (like a signature-based tool), $Q$ seems benign, because a src_bytes value of 290 resides in the benign range ([220,320]), and a dst_bytes value of 1300 resides in the benign range ([1200,1950]).

*1) Geometric Analysis (k-NN):* In the geometric modality, we visualize the data in $R^2$ (Figure 1). We calculate the distances in Table I. With $k = 3$, the nearest neighbors are $\{B, A, C\}$. A naïve voting algorithm would classify $Q$ as "Benign".

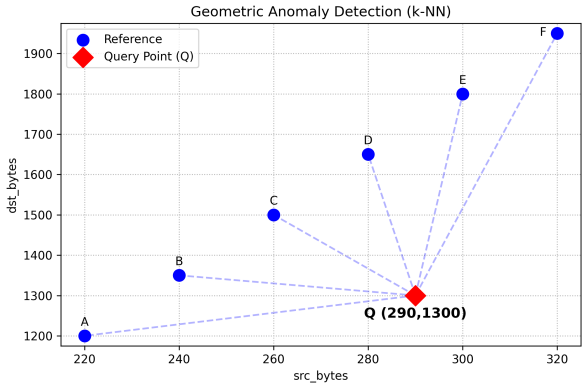


Figure 1. Visualization of the 7-point dataset. The query point $Q$ is geometrically distant from the reference points even though its coordinates reside in the benign range of both dimensions.

TABLE I. EUCLIDEAN DISTANCES (EXCERPT: CLOSEST, MEDIAN, FURTHEST)

| Point | Coord | Distance $d(Q, P_i)$ |
|---|---|---|
| **B** | (240, 1350) | $\sqrt{50^2 + 50^2} \approx \mathbf{70.7}$ |
| **D** | (280, 1650) | $\sqrt{10^2 + 350^2} \approx \mathbf{350.1}$ |
| **F** | (320, 1950) | $\sqrt{30^2 + 650^2} \approx \mathbf{650.7}$ |

A more thorough application of k-NN also fails on this dataset. Table II shows that the average internal cohesion of the reference set is 353.10, while the query point $Q$ resides at an average distance of 315.99. Hence we have $\bar{d}_Q < \bar{d}_{ref} \longrightarrow$ false negative.

Because $Q$ resides in the bounding box (hiding in the gap between $D$ and $E$), Euclidean distance algorithms classify it as a benign inlier. This confirms that for sophisticated tunneling attacks that respect min/max boundaries, geometry is insufficient. Detection requires the algebraic covariance check provided by SVD, which detects the structural violation despite the geometric proximity.

*2) Algebraic Analysis (SVD):* While k-NN fails to distinguish the anomaly $Q$ from the reference cluster (due to $\bar{d}_Q < \bar{d}_{ref}$), the SVD reveals the structural violation. We define the reference matrix $X$ using the mean-centered KDD micro-slice.

*3) Algorithmic Synthesis: Spheres vs. Cylinders:* To address the comparative efficacy of detection algorithms, it is critical to understand why neither k-NN nor SVD universally dominates. We selected k-NN as our baseline because it intuitively models *radial geometric bounds* (a convex hull or sphere around normal behavior [2] ). SVD, conversely, models *orthogonal variance limits* (a cylinder projected along the principal component).

TABLE II. INTERNAL COHESION OF REFERENCE SET (EXCERPT)

| Point $i$ | Point $j$ | Distance $d(P_i, P_j)$ |
|---|---|---|
| **A** | **B** | 151.33 |
| **B** | **C** | 151.33 |
| ... | ... | ... |
| **A** | **E** | 605.31 |
| **A** | **F** | 756.64 |

As demonstrated in Table II, a Protocol Tunneling attack easily defeats k-NN by hiding within the spatial gap of the reference sphere. However, SVD detects the anomaly because the point deviates from the cylindrical axis.

Conversely, consider the *Opposite Case*: a massive data exfiltration attack that perfectly maintains the valid HTTP Request-to-Response ratio (e.g., $(3000, 22500)$). Because this query point lies exactly on the $v_1$ subspace, its projection error onto the $v_2$ noise axis is near zero. SVD yields a false negative, failing to flag the massive volume. Here, k-NN successfully triggers an alert, as the point's magnitude places it thousands of units outside the reference sphere.

Therefore, robust critical infrastructure defense requires a hybrid detector that enforces the intersection of both constraints: the algebraic correlation of the SVD cylinder, and the geometric magnitude bounds of the k-NN sphere.

*a) Subspace Decomposition:* The covariance structure of the HTTP traffic is decomposed into its principal components ($X_{cov} = V\Lambda V^T$) [2]. Unlike the geometric proximity check of k-NN, SVD identifies the invariant subspace (the linear correlation between src_bytes and dst_bytes).

$$V^T \approx \begin{bmatrix} 0.13 & 0.99 \\ -0.99 & 0.13 \end{bmatrix} \begin{matrix} \leftarrow v_1 \text{ (Traffic Pattern)} \\ \leftarrow v_2 \text{ (noise axis)} \end{matrix}$$

Figure 2 illustrates how the first vector $v_1$ captures the valid traffic law (Request $\propto$ Response). The second vector $v_2$ represents the forbidden orthogonal variance.

*b) The Projection Test:* We project the centered query point $Q_c$ onto the noise axis $v_2$.

$$\text{Score} = |Q_c \cdot v_2| = |(20)(-0.99) + (-275)(0.13)| \approx \mathbf{55.6}$$

This high projection score (55.6) contrasts sharply with the reference set, which has near-zero projection on $v_2$. Thus, SVD mathematically reveals the anomaly by detecting the variance violation, succeeding where geometric distance failed.

*4) Complexity Analysis (Defense Against Heuristics):* A common critique in low-dimensional detection is the Ratio Heuristic – why not simply calculate $y/x$? While effective in $R^2$, heuristics fail in high-dimensional conservation tasks (e.g., Equal-Cost Multi-Path Routing [17] in $R^3$, where $x + y + z = C$).

- Configuration Cost
  Heuristics require analysts to manually derive conservation rules ($O(1)$ compute, $O(\text{Human})$ cost). SVD learns the null space automatically.
- Combinatorial Explosion
  To replicate SVD's coverage in $R^{100}$, a heuristic system would need to evaluate $\binom{100}{2} = 4,950$ pairwise ratios. Our SVD architecture evaluates the entire feature space simultaneously, evaluating a new event in $O(n)$ time without human configuration, making it the superior choice for complex network defense.

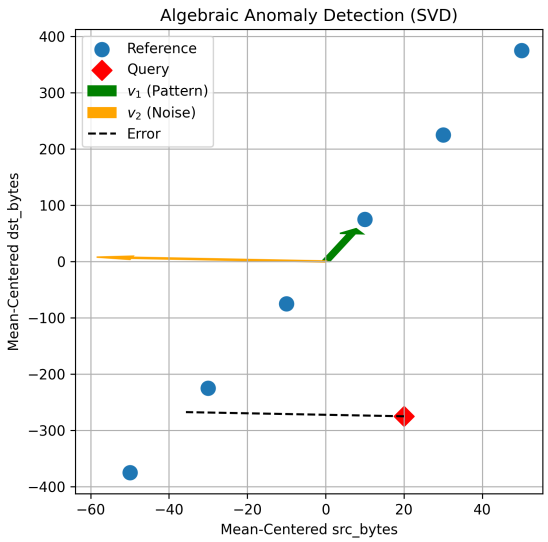


Figure 2. Visualization of the mean-centered data and the eigenvectors (scaled by $50\times$ for visualization)

## IV. Discussion | Evaluation

To assess the viability of this instructional architecture, we evaluate the proposed solution against three key industry metrics: implementation practicality, computational scalability, and workforce alignment.

### A. Practicality

The Zero Infrastructure Metric. Unlike traditional cybersecurity training that requires expensive cyber ranges or proprietary virtualization hardware [18], our modular architecture relies entirely on open-source scientific computing tools (Python, NumPy, Jupyter). This ensures high practicality for under-resourced institutions, as the curriculum can be deployed on standard student laptops or free cloud environments (e.g., Google Colab) with zero licensing costs.

### B. Scalability

Algorithmic Efficiency. From a technical perspective, the Janus Matrix approach demonstrates superior scalability compared to deep learning alternatives. Traditional deep neural networks often require massive computational overhead to train. In contrast, our architecture utilizes SVD to ensure minimal resource consumption. The algorithmic complexity is strictly bounded by $O(\min(mn^2, m^2n))$, where $m$ is the number of network events and $n$ is the number of features [3]. This allows the module to be scaled across thousands of endpoints or trainees without necessitating GPU clusters, ensuring the barrier to entry remains low.

### C. Strategic Benefits

Alignment with the National Institute of Standards and Technology (NIST) National Initiative for Cybersecurity Education (NICE). The primary benefit of this contribution is its direct mapping to the NIST NICE Workforce Framework for Cybersecurity (Special Publication 800-181) [19]. Specifically, it bridges the gap for the cyber defense analyst (PR-CDA-001) work role by moving beyond the defined Knowledge, Skills, and Abilities (KSAs) of tool operation into the implicit requirement for mathematical reasoning.

### D. Future Validation

Future work will validate the pedagogical efficacy through a pre-test/post-test instrument measuring student self-efficacy in interpreting false positives. This study will be conducted under institutional review to quantify the shift in Bloom's Taxonomy [8] levels among participants.

## V. Conclusion and Future Work

We have presented a modular instructional architecture that addresses the critical workforce gap in behavioral anomaly detection. By anchoring the curriculum in the Janus Matrix dataset and the SVD factorization, we provide a mathematical foundation often absent in standard industry certifications.

### A. Comparison with Prior Art

When comparing this architecture to existing curricular modules, such as Serra's work in the CLARK repository, a distinct divergence in pedagogical outcomes is evident. The prior art results in a workforce proficient in the application of high-level algorithms (e.g., isolation forests, SVMs) via Python libraries. While valuable for rapid deployment, this approach treats the detection engine as a black box, creating liability when facing adversarial evasion.

In contrast, our results demonstrate that a glass-box approach – starting with the geometry of k-NN and the linear algebra of SVD – equips the analyst to audit and explain the model's decisions. This alignment with NIST IR 8312 explainability principles suggests that while the prior art optimizes for implementation speed, our architecture optimizes for defensive resilience.

### B. Future Work

Future phases will deploy this module across the partner institutions to capture quantitative data on cybersecurity practitioner self-efficacy. However, the immediate architectural result is a scalable, open-source framework that demystifies the mathematics of critical infrastructure defense.

## References

[1] CISA, NSA, and FBI, "PRC state-sponsored actors compromise and maintain persistent access to U.S. critical infrastructure", Cybersecurity Advisory AA24-038A, 2024, [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-038a (visited on 01/23/2026).

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Johns Hopkins University Press, 2013.

[4] CompTIA, "CySA+ (V3) exam objectives summary", 2023, [Online]. Available: https://www.comptia.org/en-us/certifications/cybersecurity-analyst/#objectives (visited on 03/06/2026).

[5] GIAC, "GIAC Certified Intrusion Analyst Certification (GCIA)", 2026, [Online]. Available: https://www.giac.org/certifications/certified-intrusion-analyst-gcia/ (visited on 03/06/2026).

[6] National Institute of Standards and Technology, "Cybersecurity framework", Version 2.0, National Institute of Standards and Technology, 2024, [Online]. Available: https://www.nist.gov/cyberframework (visited on 01/23/2026).

[7] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems", in *Proceedings of the 9th ACM Conference on Computer and Communications Security*, ACM, 2002, pp. 255–264.

[8] L. W. Anderson and D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001, ISBN: 978-0801319037.

[9] E. Serra, "Anomaly and novelty detection", CLARK Curriculum Module, 2024, [Online]. Available: https://clark.center/details/edoardoserra/aa4b123a-d0e5-4e7a-93c3-edd8e57562f5/0 (visited on 01/23/2026).

[10] P. J. Phillips *et al.*, "Four principles of explainable artificial intelligence", 2021, [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf (visited on 01/23/2026).

[11] M. Aurelius, *Meditations*, Koine Greek, trans. Koine Greek by M. Hammond. New York: Penguin Classics, 2014.

[12] P. Ovidius Naso, *Fasti* (Oxford World's Classics), A. Wiseman and P. Wiseman, Eds. Oxford: Oxford University Press, 2004, See Book I for Janus.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey", 2009, [Online]. Available: https://dl.acm.org/doi/10.1145/1541880.1541882 (visited on 01/24/2026).

[14] UCI Machine Learning Repository, "Kdd cup 1999 data", Subset: HTTP src_bytes vs dst_bytes, 1999, [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (visited on 01/24/2026).

[15] N. Moustafa and J. Slay, "Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)", in *2015 Military Communications and Information Systems Conference (MilCIS)*, IEEE, 2015, pp. 1–6.

[16] X. Liu *et al.*, "From black box to glass box: A practical review of explainable artificial intelligence (xai)", 2025, [Online]. Available: https://www.mdpi.com/2673-2688/6/11/285 (visited on 03/06/2026).

[17] C. Hopps, "Analysis of an Equal-Cost Multi-Path Algorithm", 2000, [Online]. Available: https://www.rfc-editor.org/info/rfc2992 (visited on 01/27/2026).

[18] NIST and NICE, "The cyber range: A guide", 2023, [Online]. Available: https://www.nist.gov/system/files/documents/2023/09/29/The%20Cyber%20Range_A%20Guide.pdf (visited on 01/23/2026).

[19] National Initiative for Cybersecurity Education (NICE), "Workforce framework for cybersecurity (nice framework)", Nov. 2020, [Online]. Available: https://csrc.nist.gov/pubs/sp/800/181/r1/final (visited on 01/23/2026).