AI/ML Data Readiness for Cognitive EW

Addressing the Data Wrangling Bottleneck in Military AI Development Through Canonical Schemas and Annotation Workflows

December 21, 2025

Executive Summary

The Electromagnetic Spectrum (EMS) has evolved from a supportive medium for communication into a primary domain of conflict, characterized by a complexity and velocity that far exceeds human cognitive processing speeds. As the Department of Defense (DoD) transitions from counter-insurgency operations to near-peer competition, the ability to dominate the EMS, Electronic Warfare (EW) has become a strategic imperative. The operational tempo of modern spectrum warfare, driven by software-defined radios (SDRs) and agile, programmable threats, demands a shift from static, rule-based systems to adaptive, Artificial Intelligence (AI) and Machine Learning (ML) driven capabilities, collectively known as Cognitive EW.

However, a critical disconnect exists between the aspirational doctrine of AI-enabled warfare and the material reality of defense data infrastructure. While the algorithms powering Cognitive EW, Deep Neural Networks (DNNs), and Reinforcement Learning (RL) are maturing rapidly, they are fundamentally constrained by a lack of “data readiness.” This white paper identifies the "data wrangling" bottleneck as the single most significant impediment to the deployment of operational military AI. Currently, highly skilled engineering and intelligence resources expend up to 80% of their operational cycles locating, converting, and cleaning disparate data formats rather than developing the countermeasures required for mission success.

This report provides an exhaustive, architecturally grounded roadmap for resolving this crisis. It advocates for the deployment of Enterprise RF Data Lakes that leverage **SigMF (Signal Metadata Format)** as a canonical schema to normalize the chaotic landscape of proprietary signal formats. By decoupling massive binary storage from lightweight, queryable metadata, and by implementing robust, security-compliant annotation workflows, the defense industrial base can transform its "dark data" into high-value training assets. This strategy is rigorously aligned with the DoD’s **Project Linchpin**, the **Joint All-Domain Command and Control (JADC2)** framework, and the **Modular Open Systems Approach (MOSA)**, ensuring that data infrastructure is not merely an IT concern but a decisive factor in future combat lethality.

1. The Strategic Imperative: The Cognitive Shift in Electronic Warfare

The history of Electronic Warfare is a history of cat-and-mouse games played in the frequency domain. For decades, this game was deterministic. Intelligence agencies would capture a signal, analyze it to extract its parameters Pulse Repetition Interval (PRI), Carrier Frequency, Pulse Width, and catalog it in a static threat library. When a Radar Warning Receiver (RWR) on an aircraft detected a signal matching those parameters, it would trigger a pre-programmed jamming response. This "lookup table" paradigm relied on the assumption that the adversary's signature was immutable, hard-wired into the physical circuitry of their emitters.

1.1 From Hardware-Defined to Software-Defined Threats

That assumption no longer holds. The proliferation of high-performance Software Defined Radios (SDRs) and commercially available programmable hardware has fundamentally altered the threat landscape. Modern adversaries utilize "mode-agile" or Wartime Reserve Modes (WARM) that allow radar and communications systems to modify their waveforms dynamically.[1] A hostile radar can now shift its frequency, modulation scheme, and timing characteristics on a pulse-by-pulse basis, effectively "disappearing" from a static threat library.[2]

In this fluid environment, the traditional cycle of "Record -> Analyze -> Update Library -> Redeploy" is dangerously slow. The latency of this cycle, often measured in months, is exploited by adversaries who can update their waveforms in milliseconds. The conflict in Ukraine has served as a grim proving ground for this reality, where drone telemetry links and jamming countermeasures evolve on a weekly basis, rendering static EW tools obsolete almost as soon as they are fielded.[3]

1.2 The Promise of Cognitive EW

Cognitive EW represents the necessary technological evolution to counter this agility. It moves beyond the lookup table, employing AI and ML to perceive the spectrum in real-time, characterize unknown signals based on learned features rather than exact matches, and synthesize novel countermeasures on the fly.[4]

This capability relies on two primary branches of Machine Learning:

1. **Deep Learning (Supervised):** Utilizing Convolutional Neural Networks (CNNs) to classify signals by treating radio frequency spectrograms as images. These models can learn to identify the subtle "fingerprints" of a specific emitter type, even amidst noise and interference, provided they have been trained on vast labeled datasets.[5]
2. **Reinforcement Learning (RL):** Utilizing agents that interact with the electromagnetic environment, learning optimal jamming strategies through trial and error (reward functions) in simulated environments.[6]

As Dr. Michael Simon of Parallax Advanced Research notes, Cognitive EW "takes the person out of the loop" because the OODA loop (Observe-Orient-Decide-Act) has collapsed to microsecond timescales.[2] However, the efficacy of these "cognitive" systems is not determined by the sophistication of the code, but by the volume, diversity, and quality of the experience (data) they consume during training.

1.3 The Data Readiness Gap

Herein lies the crisis. The commercial AI revolution, exemplified by Large Language Models (LLMs) like GPT-4, was fueled by the internet, a practically infinite, publicly accessible corpus of text and code. There is no "internet of RF" for the defense sector. High-fidelity recordings of adversarial radar systems are rare, highly classified, and sequestered in isolated silos.[7]

The DoD's *Data, Analytics, and AI Adoption Strategy* emphasizes the VAULTIS framework, making data Visible, Accessible, Understandable, Linked, Trustworthy, Interoperable, and Secure.[8] Yet, the operational reality for a Machine Learning engineer in the defense sector is starkly different:

- **Visibility:** They cannot search for data because it resides on unlabeled hard drives.
- **Accessibility:** They cannot access the data because file sizes (5TB+) preclude network transfer.
- **Interoperability:** They cannot use the data because it is locked in a proprietary format (e.g., X-DAT, TMP) readable only by expensive, licensed software.[10]

This gap between the *need* for massive AI training sets and the *reality* of fragmented, inaccessible data is the "Data Readiness Gap." Bridging it requires a fundamental re-architecture of how the military collects, stores, and manages signal data.

2. The Anatomy of the Data Wrangling Bottleneck

The phrase "Data Wrangling" is often used euphemistically in data science to describe the pre-processing steps required to make data usable. In the context of Defense RF systems, however, "wrangling" is a literal description of a physically and computationally exhausting struggle against the laws of physics and the legacy of proprietary engineering.

2.1 The Physics of Data Gravity

"Data Gravity" is the concept that data, as it grows in mass, becomes increasingly difficult to move, forcing applications and services to gravitate toward it.[11] In the RF domain, data gravity is extreme.

Consider a modern wideband sensor recording 500 MHz of instantaneous bandwidth.

- **Sample Rate:** 500 Million samples per second (complex I/Q).
- **Bit Depth:** 16-bit resolution (2 bytes per sample, 4 bytes per I/Q pair).
- **Data Rate:** $500 \times 10^6 \times 4 \text{ bytes} = 2 \text{ Gigabytes per second (GB/s)}.$

A single hour of recording generates **7.2 Terabytes (TB)** of data. A typical mission might run for several hours. This volume fundamentally breaks standard IT infrastructure. Moving a 7TB file across a tactical network, or even a robust commercial fiber connection, is often unfeasible due to latency and timeout risks.

Consequently, the standard transport mechanism for RF data remains the "Sneakernet" , the physical shipment of hard drives via courier.[10] This creates a logistical bottleneck where data collected in the Pacific theater might take weeks to reach an analysis lab in the continental United States. During this transit time, the data is "dark" inaccessible to analysts and useless for retraining models to address immediate threats.

2.2 The 80/20 Inversion in Defense AI

It is a truism in commercial data science that 80% of a project's time is spent on data preparation, with only 20% dedicated to modeling and analysis.[13] In the defense sector, this ratio is often even more skewed, approaching 90/10 due to the complexity of the data and the lack of standardization.

Development Phase	Commercial (Image/Text)	Defense RF (SIGINT/EW)	The "Wrangling" Reality
Discovery	SQL/Elastic Search queries.	Physical search of lab shelves.	"I recall we saw a Type-055 radar last year, but which drive is it on?"
Ingestion	JSON/CSV API streams.	Proprietary Binary Parsing.	Engineers write custom C++ parsers for every new sensor, dealing with endianness and padding.
Cleaning	Automated outlier removal.	Manual Spectrogram Review.	RF data is plagued by DC offsets, IQ imbalance, and dropped packets that must be visually identified.
Labeling	Crowdsourced (Mechanical Turk).	Cleared Expert Only.	Only TS/SCI-cleared analysts can interpret the data; this labor cannot be outsourced or easily scaled.

This bottleneck has severe financial implications. Research suggests that data collection and preparation account for **15-25% of the total cost** of AI projects.[15] In the context of multi-billion dollar defense programs, this represents hundreds of millions of dollars wasted on manual labor that could be automated.

2.3 Format Chaos and the "Bitrot" of Intelligence

The most pernicious aspect of the wrangling bottleneck is "Format Chaos." For decades, the defense electronics market has been dominated by hardware vendors (e.g., Keysight, Rohde & Schwarz, Tektronix) whose business models relied on locking customers into their ecosystem. A recording made on a specific spectrum analyzer would be saved in a proprietary binary format that could only be opened by that vendor's software.[10]

This leads to two critical failures:

1. **Interoperability Failure:** A dataset collected by the Navy using one vendor's gear cannot be easily combined with Army data collected by another vendor to train a joint AI model.
2. **Metadata Loss (Bitrot):** Often, the critical context of Center Frequency, Sample Rate, GPS Location, and Antenna Gain is not embedded in the file headers but exists in separate "sidecar" text files, Excel spreadsheets, or even handwritten logbooks. When the binary file is separated from this context, the recording becomes "digital trash." An AI model cannot learn from a signal if it doesn't know the frequency or time scale.[16]

Solving the wrangling bottleneck therefore requires a technical solution that addresses both the physical weight of the data and the semantic chaos of its formatting.

3. The Enterprise RF Data Lake: A Strategic Architecture

To operationalize AI at the scale required for cognitive warfare, defense organizations must transition from file-centric workflows to data-centric architectures. The solution is the implementation of an **Enterprise RF Data Lake**, exemplified by the **SigDrive** system design.[10]

3.1 Decoupling Compute from Storage

The foundational principle of the RF Data Lake is the separation of the "heavy" binary data (the IQ samples) from the "light" metadata (the context).

- **The Object Store (The Vault):** The raw binary files (WAV, DAT, BIN) are stored in immutable, high-throughput Object Storage systems (e.g., S3-compatible architecture like MinIO). This layer handles the "Data Gravity," optimized for massive ingest speeds (resumable uploads of 5TB+ files) and long-term durability.[10]
- **The Metadata Store (The Brain):** Upon ingestion, the system extracts all technical and contextual metadata and stores it in a high-performance relational database (PostgreSQL). This index is lightweight, allowing users to query petabytes of data in milliseconds.

This architecture enables "Zero-Download Analysis." An analyst can execute a crafted query to *"Show me all X-Band radar pulses recorded within 50km of Guam in 2024"* and receive an immediate list of results without the system needing to retrieve or move the underlying Terabytes of binary data.

3.2 Automated Ingestion and Normalization

The "Ingestion Service" serves as the universal translator for the Data Lake. It is designed to ingest the chaotic array of legacy formats and normalize them into a single, queryable standard.

The Ingestion Workflow:

1. **Identification:** The service inspects file headers (magic bytes) to identify the format (e.g., Midas Blue, WAV, SigMF).
2. **Extraction:** A plugin-based architecture invokes specific parsers:
    - *Midas Blue:* Parses the 512-byte binary header to extract frequency and timing.[10]
    - *WAV/RF64:* Extracts sample rates from the RIFF chunk.
    - *Raw IQ:* Prompts the user for manual entry of missing parameters.
3. **Normalization:** All extracted metadata is mapped to a **Canonical Schema** (discussed in Section 4). This ensures that Center Frequency is always stored as core:frequency (in Hz), regardless of whether the source file called it freq, fc, or center_freq.
4. **Indexing:** The normalized metadata, including Geospatial coordinates, is indexed in PostGIS, enabling complex spatial searches (e.g., "Find recordings inside this polygon").[10]

3.3 The Air-Gap Constraint: Architecture for the Disconnected Edge

Defense AI development rarely happens on the open cloud. It occurs in Sensitive Compartmented Information Facilities (SCIFs) on networks like SIPRNet and JWICS that are physically air-gapped from the internet. Commercial "SaaS" (Software as a Service) models fail in this environment because they rely on continuous connectivity for licensing and updates.

The SigDrive architecture addresses this via an **"Air-Gap Ready"** design strategy [10]:

- **Containerization:** The entire system is packaged as Docker containers, orchestratable via Kubernetes.
- **Offline Dependencies:** All libraries and dependencies are bundled within the images; the system makes no external API calls to the internet.
- **Sneakernet Updates:** Updates are delivered via secure physical media (DVDs/Hard Drives) containing signed container images and migration scripts.
- **Local Licensing:** License validation is performed against a local, cryptographic key rather than a cloud server.

This architectural rigidity is a prerequisite for deployment on classified networks, ensuring that the Data Lake can serve as the backbone for sensitive Cognitive EW programs.

4. The Canonical Schema: SigMF as the Foundation of Interoperability

If the Data Lake is the warehouse, the **Canonical Schema** is the inventory system. Without a unified language to describe RF data, the lake becomes a "Data Swamp." The defense industry is coalescing around **SigMF (Signal Metadata Format)** as this lingua franca, driven by its adoption in open-source tools (GNU Radio) and its alignment with DoD interoperability mandates.[16]

4.1 SigMF: The Structure of Signal Intelligence

SigMF was designed specifically to solve the portability issues inherent in RF data. A SigMF recording consists of a pair of files:

1. **.sigmf-data:** The binary file containing the raw digital samples (IQ data).
2. **.sigmf-meta:** A plain-text JSON file containing the metadata.[17]

The power of SigMF lies in its JSON structure, which is human-readable and machine-parseable. It organizes metadata into three primary scopes:

- **Global:** Information that applies to the entire recording (e.g., core:sample_rate, core:author, core:hw for hardware description).
- **Captures:** Information about specific segments of time (e.g., core:frequency and core:datetime). This handles cases where the radio retuned to a different frequency halfway through a recording.
- **Annotations:** Metadata describing specific features *within* the recording (e.g., "This 5ms segment is a P4 code radar pulse"). This is the critical layer for AI labeling.[17]

4.2 The Transport vs. Storage Debate: VITA 49 vs. SigMF

A frequent point of confusion in military procurement is the relationship between SigMF and **VITA 49 (VRT - VITA Radio Transport)**. Clarifying this distinction is vital for system architects.

Feature	VITA 49 (VRT)	SigMF
Primary Purpose	**Transport:** Streaming data over a wire/network in real-time.	**Storage:** Describing data at rest on a disk/cloud.
Structure	Packetized Binary (Headers interleaved with data).	Decoupled: Binary Data + JSON Metadata.
Readability	Opaque (Requires complex packet parsing software).	Transparent (Human-readable text).
Overhead	High (Header overhead per packet).	Low (One metadata file per recording).
AI Suitability	Poor (Must be decoded/depacketized before training).	Excellent (Native format for tools like TorchSig).

Table 1: Comparison of VITA 49 and SigMF Standards.[19]

The optimal workflow, as implemented in SigDrive, acknowledges the strengths of both: **Stream in VITA 49, Store in SigMF.** The ingestion layer acts as the transcoder, stripping the VRT packet headers and consolidating the metadata into a SigMF JSON file. This preserves the fidelity of the capture while making the data immediately accessible for AI training without the need for packet-level decoding during every training epoch.

4.3 Mapping Legacy Formats to the Canonical Standard

To operationalize the "Data Readiness" strategy, the system must map legacy headers to the SigMF canonical schema. This mapping logic is the "secret sauce" of the ingestion engine, ensuring that data from a 1990s-era Midas Blue file is semantically equivalent to data from a 2024 SDR.

- **Midas Blue:** The system parses the 512-byte binary header. The bandwidth field at byte offset 0x120 is mapped to core:sample_rate; the timecode is converted to ISO-8601 for core:datetime.[10]
- **WAV/RF64:** The fmt chunk provides the sample rate and bit depth. Since WAV files lack center frequency metadata, the ingestion workflow must prompt the user or parse the filename (if standardized) to populate the core:frequency field.

This normalization allows the Data Lake to present a unified API to the AI/ML pipeline. A query for core:frequency > 2.4 GHz will return relevant files regardless of their original format.

5. The Annotation Workflow: Generating Ground Truth

Ingestion and normalization prepare the data, but they do not make it "intelligent." For Supervised Learning models to function, the data must be **labeled**. The annotation workflow is the process of generating "Ground Truth" the definitive reference that tells the AI "This signal is an SU-35 Radar" or "This is 5G Interference."

5.1 The Region of Interest (ROI) Concept

In Computer Vision, annotators draw bounding boxes around cars or pedestrians. In RF Machine Learning, the equivalent concept is the **Region of Interest (ROI)**.[10] An ROI is defined by a specific bounding box in time and frequency:

- **Time Range:** Start Sample (t_{start}) to End Sample (t_{end}).
- **Frequency Range:** Lower Frequency (f_{min}) to Upper Frequency (f_{max}).

This ROI is visualized on a **Spectrogram**, a 2D heat map where the X-axis is time, the Y-axis is frequency, and color represents signal power (amplitude).

5.2 The Spectrogram Labeling Challenge

Labeling RF data is significantly more challenging than labeling images.

1. **Ambiguity:** RF signals often overlap in time and frequency. Identifying a faint radar pulse buried beneath strong LTE interference requires high Signal-to-Noise Ratio (SNR) visualization and expert judgment.[22]
2. **Domain Expertise:** Unlike labeling images of "cats" vs. "dogs," which can be outsourced to generalist crowdsourcing platforms, identifying a "Linear Frequency Modulated (LFM) Chirp" requires a physicist or a highly trained EW officer.

3. **Visual Cues:** Experts look for subtle cues. For example, in speech processing, the "velar pinch" (a convergence of formants) indicates a specific consonant.[23] Similarly, in EW, the "sidelobes" of a radar pulse or the specific "ramp-up" time of a transmitter provide clues to its identity.

To support this, the SigDrive architecture mandates a **WebGL-based Spectrogram Viewer**.[10] This tool allows analysts to stream "tiles" of the spectrogram (similar to Google Maps) from the server to the browser, enabling them to zoom in on microsecond-level details within a 5TB file without downloading the entire dataset.

5.3 AI-Assisted Annotation and Synthetic Data

Given the scarcity of expert analysts, manual labeling is the tightest bottleneck in the pipeline. To scale, the workflow must leverage **AI-Assisted Labeling** and **Synthetic Data**.

AI-Assisted Labeling (Human-in-the-Loop):
Instead of drawing every box manually, the analyst runs a "pre-labeling" model (e.g., a simple energy detector or a legacy classifier). This model proposes ROIs. The analyst then reviews, corrects, and approves them. This "verify-only" workflow can speed up annotation by 10x-100x compared to manual drawing.22
Synthetic Data Generation:
For threats where real data is nonexistent (e.g., a new adversary radar that has not yet been observed), the system must rely on synthetic data. Using tools like DeepRFSoC or Project Linchpin's synthetic data generators, engineers can mathematically simulate the waveform of the threat, inject it into real recorded background noise, and generate fully labeled SigMF files.25 This approach, known as "Direct to Phase II" in SBIR terminology, is critical for training models on "Black Swan" events.

6. Integrating with the AI Ecosystem: The Feature Store

The ultimate consumer of the Data Lake is the AI training pipeline. To facilitate this, the Data Lake must function as a **Feature Store**, serving curated data directly to ML frameworks like PyTorch and TensorFlow.

6.1 TorchSig and the Python Ecosystem

The adoption of SigMF has enabled the development of standardized data loaders. **TorchSig** is a prime example: an open-source library that serves as a bridge between SigMF files and PyTorch.[27]

The TorchSig Workflow:

1. **Dataset Definition:** TorchSig allows researchers to define a dataset that points to a collection of SigMF files.
2. **On-the-Fly Loading:** It reads the SigMF metadata to identify signal locations.

3. **Augmentation:** As data is streamed into the GPU, TorchSig applies randomized augmentations, adding noise, simulating fading channels, introducing frequency offsets, to make the model robust.[29]
4. **Tensor Generation:** It converts the IQ samples into complex tensors ready for ingestion by a neural network.

This seamless integration eliminates the need for data scientists to write custom file parsers, allowing them to focus entirely on model architecture and hyperparameter tuning.

6.2 Pulse Deinterleaving and Complex Scenarios

One of the most complex tasks for Cognitive EW is Pulse Deinterleaving. In a dense combat environment, a sensor receives pulses from dozens of different emitters simultaneously. The AI must separate (deinterleave) these interleaved pulse trains to identify individual radars.
This requires datasets where every single pulse is labeled not just with "Radar" but with a "Track ID" (e.g., "Emitter #1", "Emitter #2"). The SigDrive annotation schema supports this by allowing "Relationship" links between annotations, effectively creating a "Track" entity that binds multiple ROIs together across time.30

7. DoD Doctrine and Compliance: Project Linchpin and MOSA

The transition to this data-centric architecture is not just a technical optimization; it is a compliance requirement for future DoD programs.

7.1 Project Linchpin: The AI Ecosystem

Project Linchpin is the US Army's flagship initiative to operationalize AI for sensor modernization.32 It recognizes that the "AI Pipeline" is too expensive for individual programs to build alone. Linchpin aims to provide a centralized "AI/MLOps" environment.
Crucially, Linchpin emphasizes "Data Readiness" as a deliverable. Contractors are no longer just delivering the "Black Box" model; they must deliver the curated, labeled data used to train it, formatted in open standards like SigMF.34 The SigDrive architecture is effectively a "Linchpin-compliant" node, designed to plug into this broader ecosystem.

7.2 MOSA and the Open Standard Mandate

The Modular Open Systems Approach (MOSA) mandates that defense systems utilize modular interfaces to prevent vendor lock-in.35 In the context of data, MOSA means that the government must own the "Data Rights" and the data must be in a format that allows a different vendor to train a new model on it.
By adopting SigMF, the SigDrive architecture ensures MOSA compliance. It breaks the dependency on proprietary tools from legacy hardware vendors, allowing the DoD to mix and match sensors and analysis software.

7.3 The VAULTIS Framework

The DoD Data Strategy utilizes the **VAULTIS** acronym to benchmark data readiness. The Enterprise RF Data Lake directly addresses each pillar [8]:

VAULTIS Pillar	Problem	Solution (SigDrive/SigMF)
Visible	Data hidden on drives.	Metadata Index & Search Engine makes all data discoverable.
Accessible	5TB files hard to move.	Zero-Download Analysis via Web Spectrograms.
Understandable	Binary blobs, no context.	Canonical Schema (SigMF) provides human-readable context.
Linked	Isolated recordings.	Relationship schema links signals to missions and tracks.
Trustworthy	Unknown provenance.	Immutable Audit Logs track every modification.
Interoperable	Vendor formats.	SigMF Standardization.
Secure	Air-gap challenges.	RBAC, ITAR tagging, and Local Licensing.

Table 2: Alignment of Enterprise RF Data Lake with VAULTIS Strategy.

8. Financial and Operational Analysis

8.1 The Cost of Inaction

The cost of maintaining the status quo is staggering. With the global Electronic Warfare market projected to reach **$35.4 billion by 2030** [1], the inefficiency of manual data wrangling represents a massive financial leak. If 20% of a program's budget is data preparation [15], and that process is 50% inefficient due to format chaos, the DoD is losing billions annually to "digital friction."

8.2 Operational Impact: The Speed of Relevance

More critical than cost is the "Speed of Relevance." In a conflict with a near-peer adversary, the time to adapt to a new waveform is the difference between survival and defeat.

- **Current State:** Adversary introduces new radar mode -> 2 weeks to ship drives -> 2 weeks to clean data -> 1 week to train. **Total: 5 weeks.**
- **Target State (Data Lake):** Adversary introduces new radar mode -> Data uploaded to local lake -> Annotated in 2 hours -> Retrained overnight. **Total: 24 hours.**

This acceleration of the adaptation cycle is the strategic objective of Cognitive EW.

9. Conclusion

The "Data Wrangling Bottleneck" is not merely an IT nuisance; it is a strategic vulnerability in the US military's ability to wage spectrum warfare. The sophistication of Cognitive EW algorithms is rendered irrelevant if the data required to train them is trapped in proprietary silos, divorced from its metadata, and inaccessible to the engineering workforce.

To achieve AI readiness, the defense community must embrace a **Data-Centric** architecture. This requires:

1. **Standardization:** The universal adoption of **SigMF** as the canonical schema for RF data storage, ensuring interoperability and preserving metadata.
2. **Infrastructure:** The deployment of **Enterprise RF Data Lakes** (like SigDrive) that decouple storage from compute and solve the physics of Data Gravity.
3. **Workflow Modernization:** The implementation of browser-based, AI-assisted **Annotation Workflows** that allow experts to generate Ground Truth efficiently.
4. **Compliance:** Strict adherence to **MOSA** and **Project Linchpin** standards to ensure that data is treated as a long-term strategic asset.

By transforming raw signal data from a logistical burden into a curated, queryable, and actionable intelligence asset, the DoD can unlock the true potential of Artificial Intelligence, ensuring dominance in the electromagnetic spectrum for the decades to come.

10. Glossary of Key Terms

- **Cognitive EW:** Electronic Warfare systems that use AI/ML to perceive and adapt to the spectrum in real-time.
- **Data Gravity:** The tendency of massive datasets to be difficult to move, requiring compute to move to the data.
- **IQ Data:** In-Phase and Quadrature samples; the raw digital representation of an RF signal.
- **MOSA:** Modular Open Systems Approach; a DoD strategy for open interfaces.
- **Project Linchpin:** US Army initiative to build a trusted AI/ML pipeline for sensors.
- **SigMF:** Signal Metadata Format; the open-source JSON standard for RF recordings.
- **Spectrogram:** A visual representation of signal strength over time and frequency.
- **VITA 49 (VRT):** A packet-based standard for streaming RF data (distinct from storage).
- **VAULTIS:** DoD Data Strategy framework (Visible, Accessible, Understandable, Linked, Trustworthy, Interoperable, Secure).

References

1. Testing Cognitive Radar Systems - Journal of Electromagnetic Dominance, https://www.jedonline.com/2022/06/01/testing-cognitive-radar-systems/
2. Cognitive Electronic Warfare and the Fight for Spectrum Superiority, https://parallaxresearch.org/news/blog/cognitive-electronic-warfare-and-fight-spectrum-superiority
3. AI Bubble and Military Bottleneck: A Systemic Crisis - IRIS, https://www.iris-france.org/en/ai-bubble-and-military-bottleneck-a-systemic-crisis/
4. Cognitive Electronic Warfare: An Artificial Intelligence Approach, Second Edition, https://us.artechhouse.com/Cognitive-Electronic-Warfare-An-Artificial-Intelligence-Approach-Second-Edition-P2445.aspx
5. What Is Cognitive Electronic Warfare (CEW)? - BAE Systems, https://www.baesystems.com/en-us/definition/what-is-cognitive-electronic-warfare
6. Implement AI in Electromagnetic Spectrum Operations | Proceedings - U.S. Naval Institute, https://www.usni.org/magazines/proceedings/2023/august/implement-ai-electromagnetic-spectrum-operations
7. August/September 2018 - Cognitive Electronic Warfare: Radio Frequency Spectrum Meets Machine Learning | Avionics Digital Edition - Aviation Today, https://interactive.aviationtoday.com/avionicsmagazine/august-september-2018/cognitive-electronic-warfare-radio-frequency-spectrum-meets-machine-learning/
8. DOD Data Strategy, https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF
9. Secure data is superior data: A security-first approach to the DoD Data Strategy - Elastic, https://www.elastic.co/blog/secure-data-dod-data-strategy
10. Strategic Business Plan_ SigDrive Enterprise RF Data Lake (1).pdf
11. What is Data Gravity? | Talend, https://www.talend.com/resources/what-is-data-gravity/
12. Data Gravity vs. Data Velocity - Interconnections - The Equinix Blog, https://blog.equinix.com/blog/2024/01/11/data-gravity-vs-data-velocity/
13. What is Data Wrangling? | Altair Data Analytics, https://altair.com/what-is-data-wrangling
14. What is Data Wrangling and Why Does it Take So Long? - Elder Research, https://www.elderresearch.com/blog/what-is-data-wrangling-and-why-does-it-take-so-long/
15. AI Project Cost Estimation: 2026 Pricing Breakdown for Manufacturing Leaders, https://usmsystems.com/ai-project-cost-estimation/
16. SigMF, https://sigmf.org/
17. sigmf/SigMF: The Signal Metadata Format Specification - GitHub, https://github.com/sigmf/SigMF
18. SigMF: The Signal Metadata Format - Proceedings of the GNU Radio Conference, https://pubs.gnuradio.org/index.php/grcon/article/download/52/38/

19. SigMF - Wikipedia, https://en.wikipedia.org/wiki/SigMF
20. (PDF) The VITA 49 analog RF-digital interface - ResearchGate, https://www.researchgate.net/publication/234591876_The_VITA_49_analog_RF-digital_interface
21. VITA 49 enhances capabilities and interoperability for transporting SDR data, https://vita.militaryembedded.com/2879-vita-enhances-capabilities-interoperability-transporting-sdr-data/
22. Automated Labeling of Time-Frequency Regions for AI-Based Spectrum Sensing Applications - MATLAB & Simulink - MathWorks, https://www.mathworks.com/help/signal/ug/automated-labeling-of-time-frequency-regions-for-ai-based-spectrum-sensing-applications.html
23. How to Label Spectrograms for AI Models, https://labelstud.io/blog/how-to-label-spectrograms-for-ai-models/
24. Video Annotation Made Simple: Best Tools & Techniques - Roboflow Blog, https://blog.roboflow.com/video-annotation/
25. Artificial Intelligence/Machine Learning (AI/ML) Ready Synthetic Radio Frequency (RF) Data, https://armysbir.army.mil/topics/ai-ml-ready-synthetic-radio-frequency-rf-data/
26. DeepRFSoC: Dataset for Modulation Classification - University of Strathclyde, https://pureportal.strath.ac.uk/en/datasets/deeprfsoc-dataset-for-modulation-classification/
27. A PyTorch Signals Toolkit - TorchSig, https://torchsig.com/dist/about.html
28. TorchSig is an open-source signal processing machine learning toolkit based on the PyTorch data handling pipeline. - GitHub, https://github.com/TorchDSP/torchsig
29. Team members: Garrett Vanhoy, Luke Boegner, Manbir Gulati, Phil Vallance, Rob Miller, Bradley Comar, Silvija Kokalj- Filipovic, - GNU Radio Events, https://events.gnuradio.org/event/18/contributions/268/attachments/103/211/TorchSigWorkshop_GRCon.pdf
30. hugodrak/deinterleaving_ew_signal_intelligence: Deinterleaving of rf signals into separate emitters - GitHub, https://github.com/hugodrak/deinterleaving_ew_signal_intelligence
31. Case-Study Analysis for Deinterleaving of X-Band RADAR Scans - DTIC, https://apps.dtic.mil/sti/trecms/pdf/AD1228224.pdf
32. PM IS&A - CPE IEW&S, https://peoiews.army.mil/pm-isa/
33. Developing an AI/ML Operations Pipeline: Project Linchpin - PEO IEW&S, https://peoiews.army.mil/wp-content/uploads/2023/09/Project-Linchpin-Approved-for-Release-1.pdf
34. The Pentagon's new 'Project Linchpin' could shake things up - Euro-sd, https://euro-sd.com/2025/11/articles/exclusive/47645/the-pentagons-new-project-linchpin/
35. Implementing a Modular Open Systems Approach in Department of Defense Programs - USD(R&E), https://www.cto.mil/wp-content/uploads/2025/03/MOSA-Implementation-Guideb

ook-27Feb2025-Cleared.pdf
36. Modular Open Systems Approach - MOSA - Curtiss-Wright Defense Solutions, https://defense-solutions.curtisswright.com/system/files/2023-10/DSAll_Infographic_What-is-MOSA_0.pdf