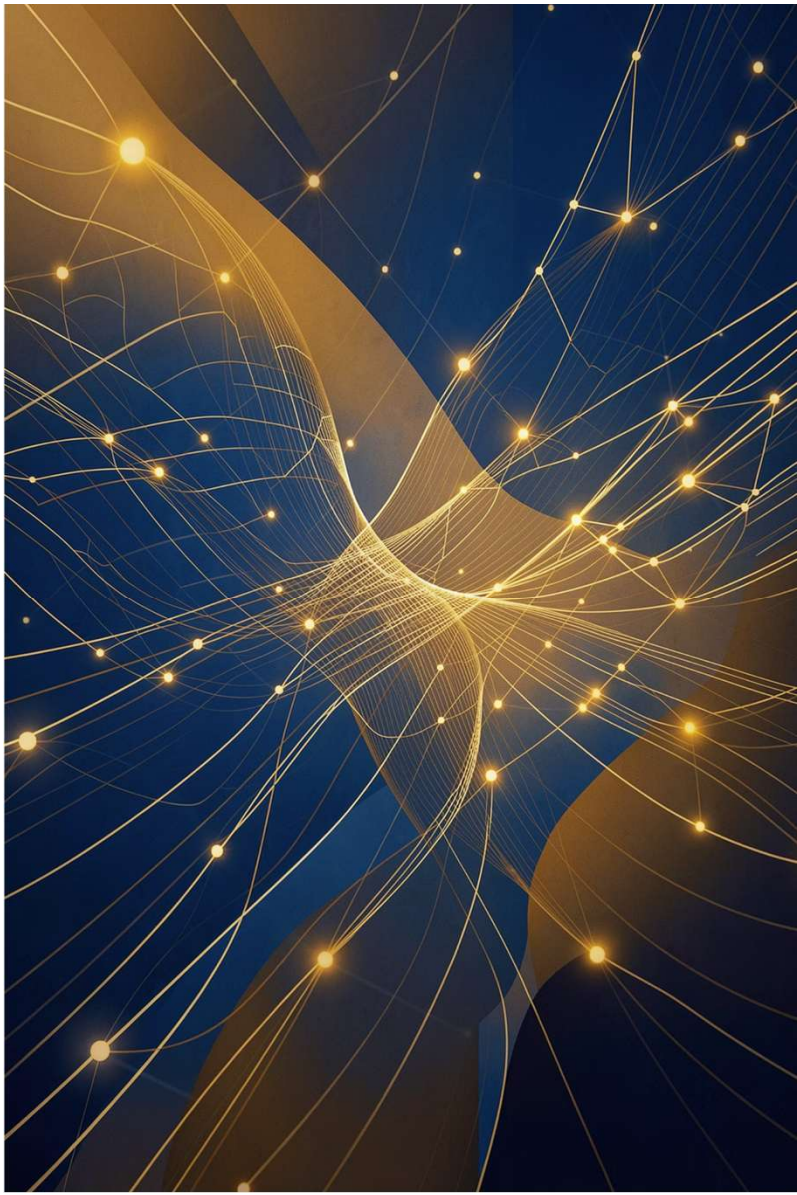
AI to deter cybercrime

An auditor's lens on evidence, federation, and compliance

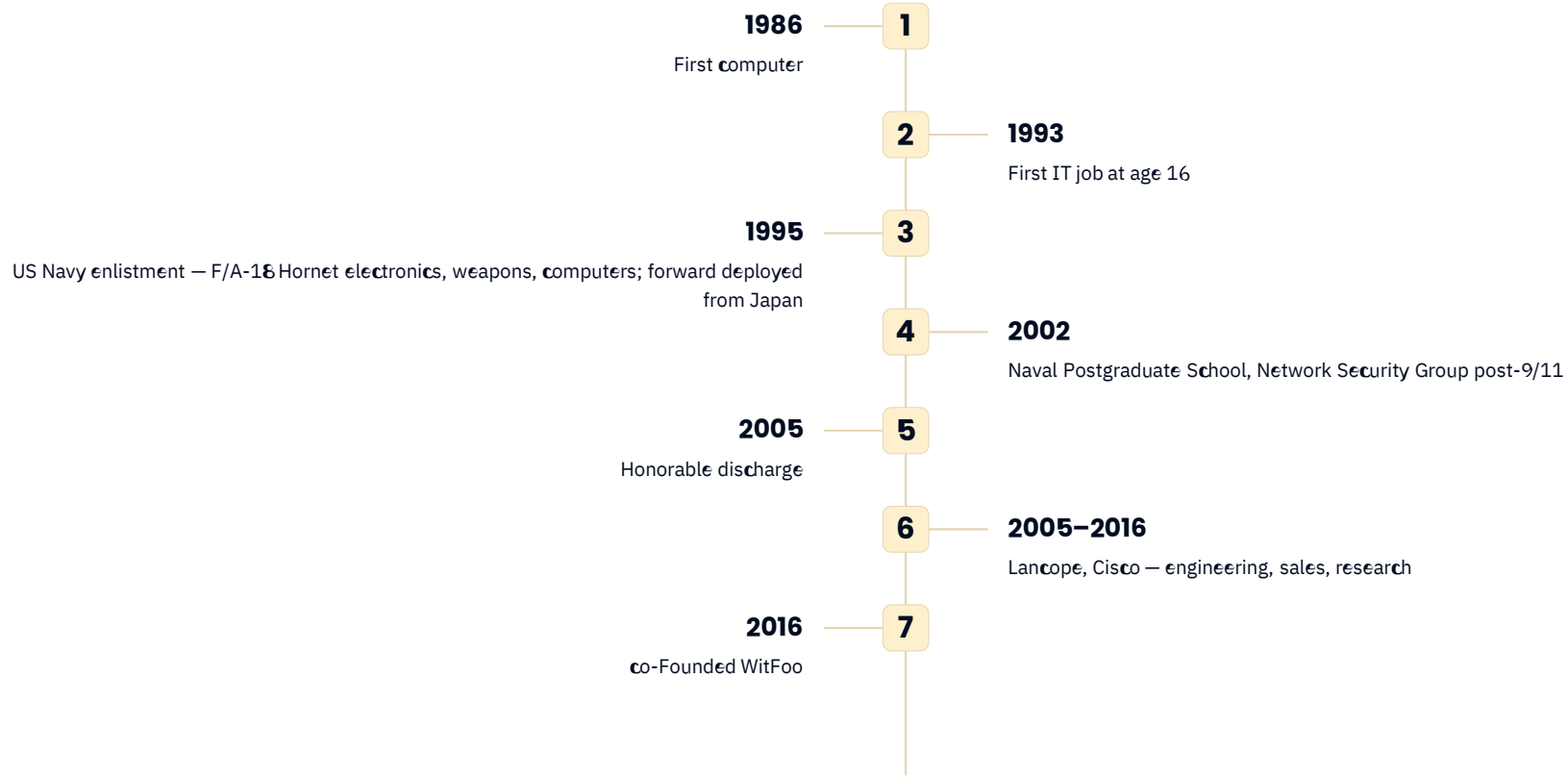
Charles D. Herring
WitFoo, Chairman and co-Founder

charlesherring.com |
Charles@witfoo.com |
@charlesherring

IIA/ISACA Chicago
11th Annual Hacking Conference
11 May 2026



About Charles



> Two decades chasing situational awareness an investigator can stand behind.

Why this room, why now

AI is being inserted into security operations across every sector you audit. Most AI-in-security claims are not auditable.

> Miyagi-Do is not about looking defensive. It is about controls that work when struck. Auditors decide whether what looks like defence actually is.

The IIA Three Lines, with AI inserted:

First line	Second line	Third line
Builds it	Risk-rates it	Audits it

Today's questions sharpen all three.

Three goals of cybersecurity AI analysis

1

Comprehend signals at scale

Fully comprehend thousands to billions of signals per second — in real time, without loss.

2

Find and organise evidence

Find evidence of crime, organise it into work units, and analyse those units with precision.

3

Explain to lay audiences

Translate technical work for judges, lawyers, jurors, and other non-technical audiences.

⚠ The work product must serve investigators AND non-technical audiences. If any of the three goals fail, the case fails.

Four functions called "AI" — only some are auditable

The auditability of an AI workflow is determined by how much of it is deterministic. The less the model has to guess, the more an auditor can stand behind the result.

Function	What it is	Reproducible?	Auditor's lens
GenAI (LLMs)	Neural network + training set	No — probabilistic	Commentary, not primary evidence
NLP / semantic framing	Coded parsers and grammars	Yes — deterministic	Same input, same output. Defensible
Graph theory	Code + JSON state	Yes — deterministic	Auditable state machine
Narrow AI (ANI)	Coded expertise	Yes — deterministic	Same standards as any rule engine

The analytic pipeline through an audit lens

1

Comprehend signals

NLP parsing and normalisation

2

Maintain state

Graph theory — JSON state machine

3

Create work units

ANI — coded expertise

4

Analyse

ANI plus bounded GenAI

5

Produce commentary

GenAI — final 1–5%

The first 95–99% is deterministic. GenAI is the final 1–5%. Everything before the last step is what makes the last step safe.

Empathetic processing — the deterministic substrate

WitFoo's deterministic, security-aware pre-AI data layer. The "empathy" is machine-to-human — the system models how the operator thinks, not the other way around.

Empathetic listening

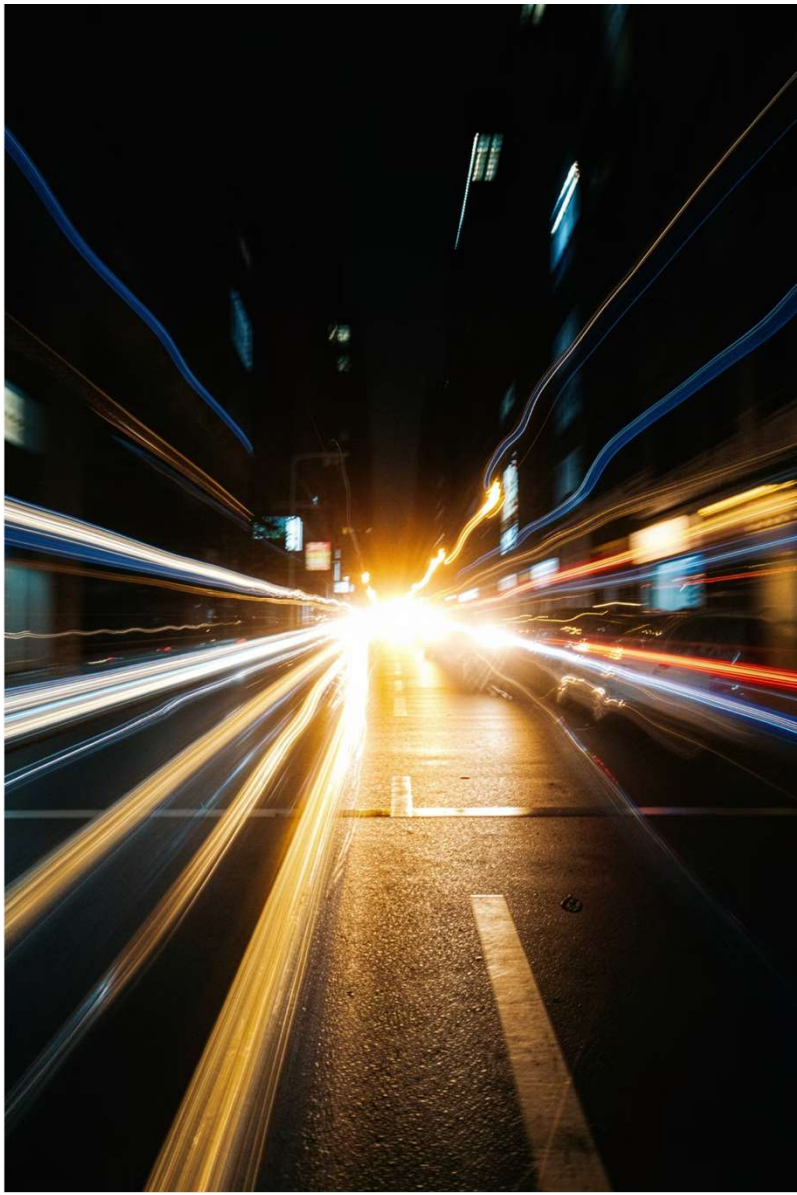
Context-aware parsing and normalisation, no regex to maintain. The system understands the meaning of events, not just their syntax.

Dissonance resolution

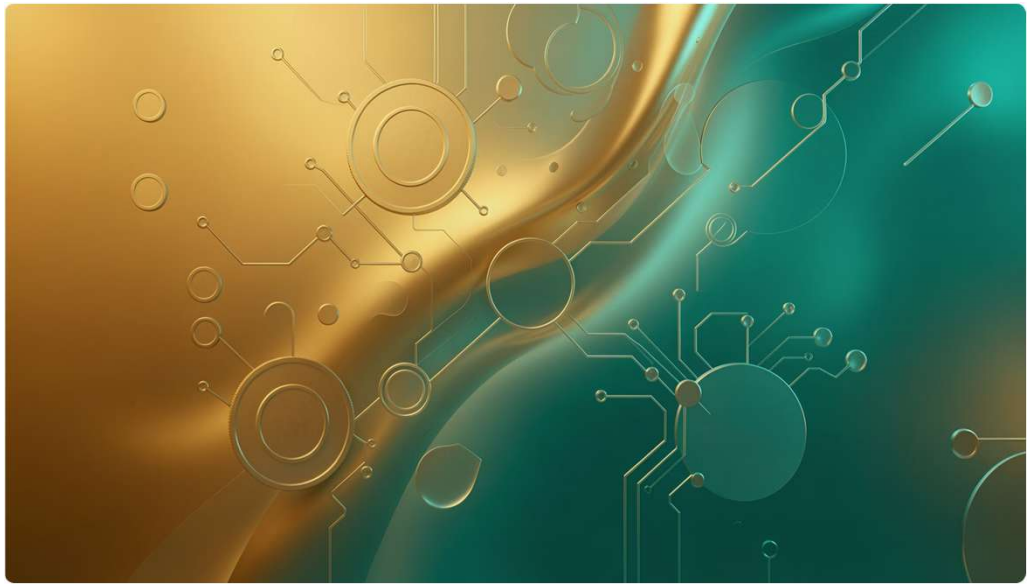
6 to 10 fold redundancy collapses to one enriched artifact, with every contributing source preserved for chain-of-custody purposes.

Empathetic speaking

Same evidence, four shapes: analyst, AI agent, SIEM, and auditor. The output adapts to the audience without altering the underlying facts.



Why EP matters for AI defensibility



Move cheap work out of the AI

AI is most expensive when it does work that deterministic logic could do faster, cheaper, and more defensibly. Empathetic Processing moves that work out of the AI entirely.

Chain of custody is the architecture

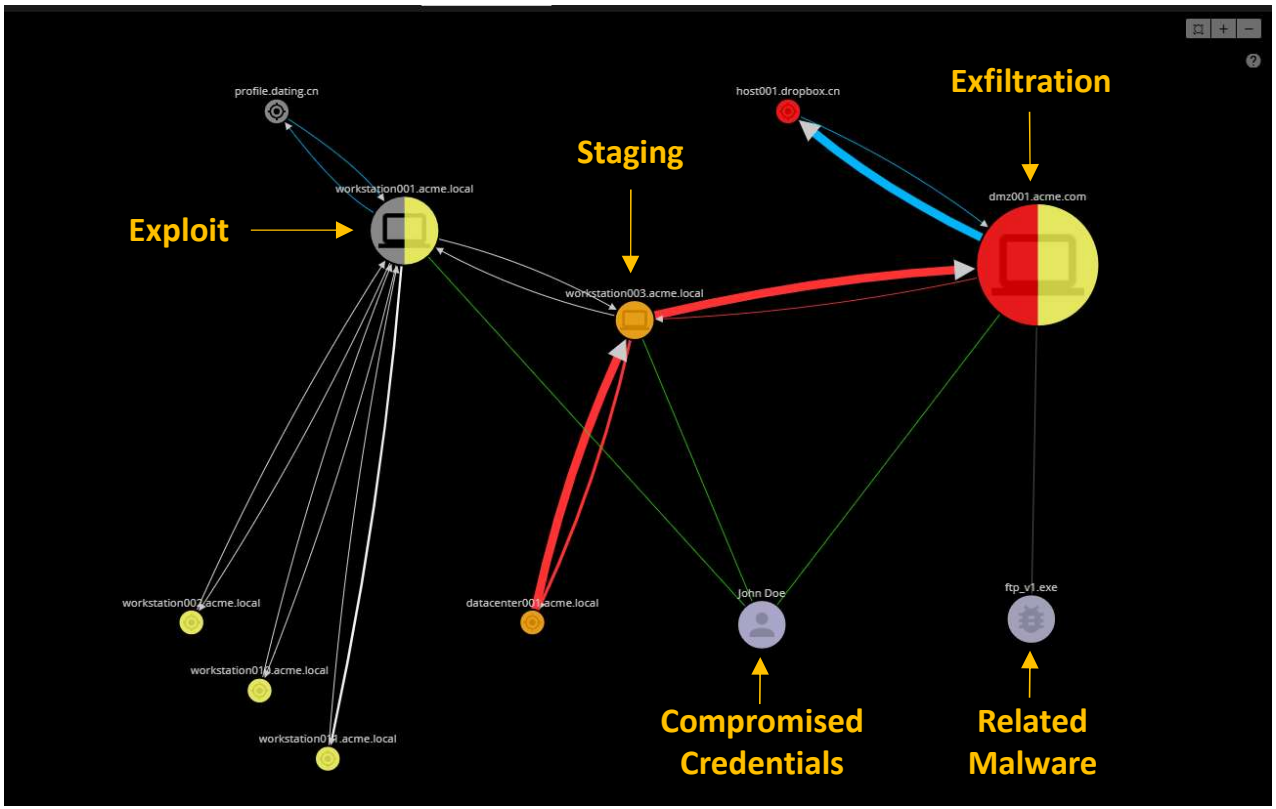
Every claim traces back to the raw event that produced it. Chain of custody is not a bolt-on — it is baked into the design from the ground up.

The economics

AI budget goes up in value per dollar by orders of magnitude because the AI is no longer doing cheap work at expensive prices. The deterministic layer does the heavy lifting; the AI does what only AI can do.

Fully Informed Graph (Data Theft)

Network Defense



DLP & Flow

Identity Defense

Endpoint Defense

TRIAGE TRAP — FULL EVIDENCE

A successful data theft, fully observed

This is what completeness looks like. Reconstructed from network defence, endpoint defence, identity defence, network flow, and DLP signals.

11 nodes

Full attack surface mapped

12 edge relationships

Every connection documented

Multi-stage attack

Exploit → scan → credential theft → lateral movement → staging → exfiltration

5 signal sources

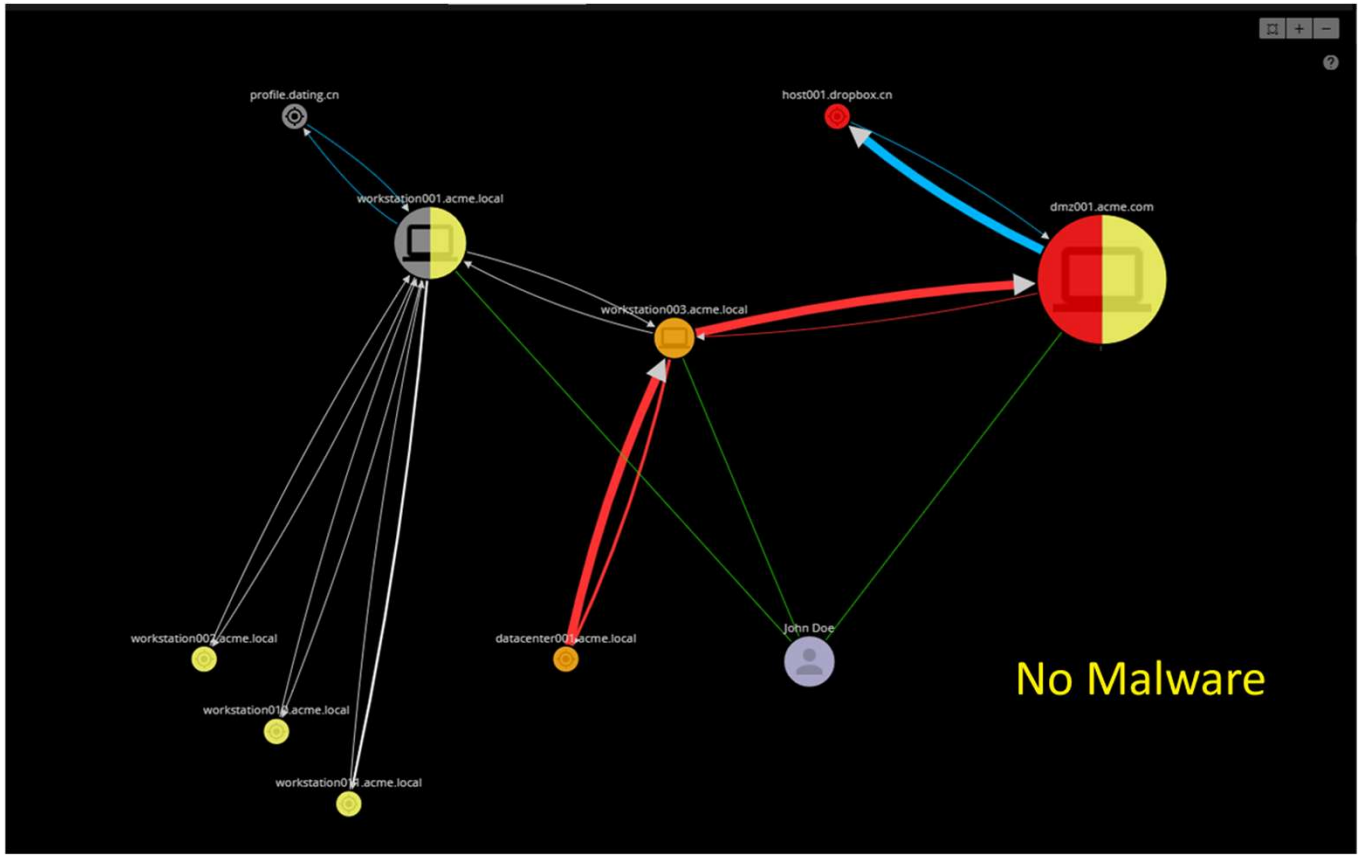
Network, endpoint, identity, flow, DLP — all present

All five signal classes present. Case is prosecutable. Evidence is complete.

No Endpoint Defense Signal

Network Defense

Identity Defense



DLP & Flow

Endpoint Defense

TRIAGE TRAP — REMOVE ENDPOINT SIGNALS

Remove endpoint signals

Investigation still operable. Data theft can be proven beyond reasonable doubt. Endpoint forensics can recover additional artifacts after the fact.

11 nodes

Attack surface still mapped

12 edge relationships

Connections still documented

Multi-stage attack

Exploit → scan → credential theft → lateral movement → staging → exfiltration

Endpoint signals

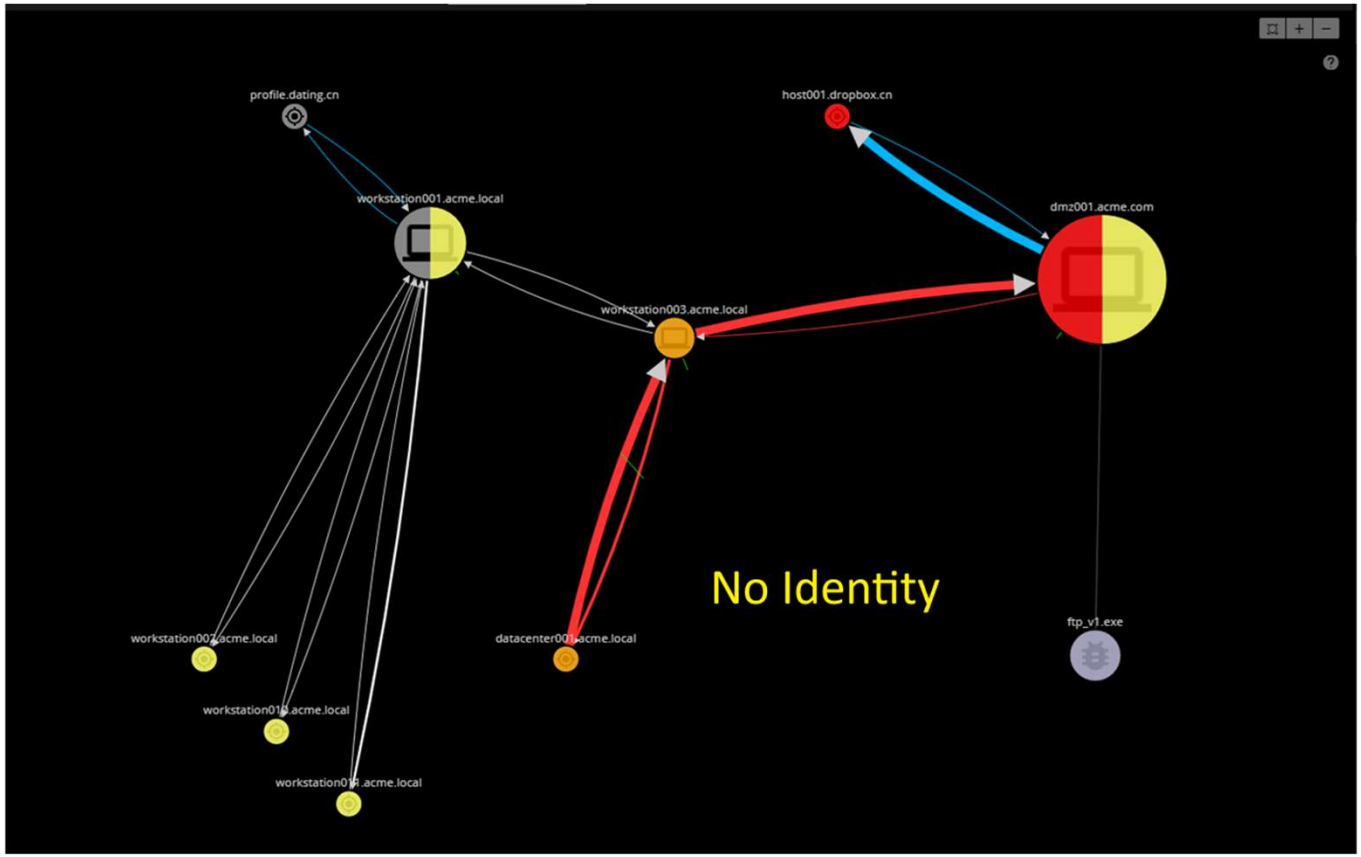
⚠ Removed — some detail lost

Cost: some lost detail. Outcome preserved. Case remains prosecutable.

No Identity Signal

Network Defense

Identity Defense



DLP & Flow

Endpoint Defense

TRIAGE TRAP — REMOVE IDENTITY SIGNALS

Remove identity signals

Lateral movement still visible. We see staging and exfiltration. But we lose the *how* — which credential, which user, which path.

Partial nodes

Attack surface partially mapped

Reduced edges

Some relationships now missing

Endpoint signals

⚠ Removed

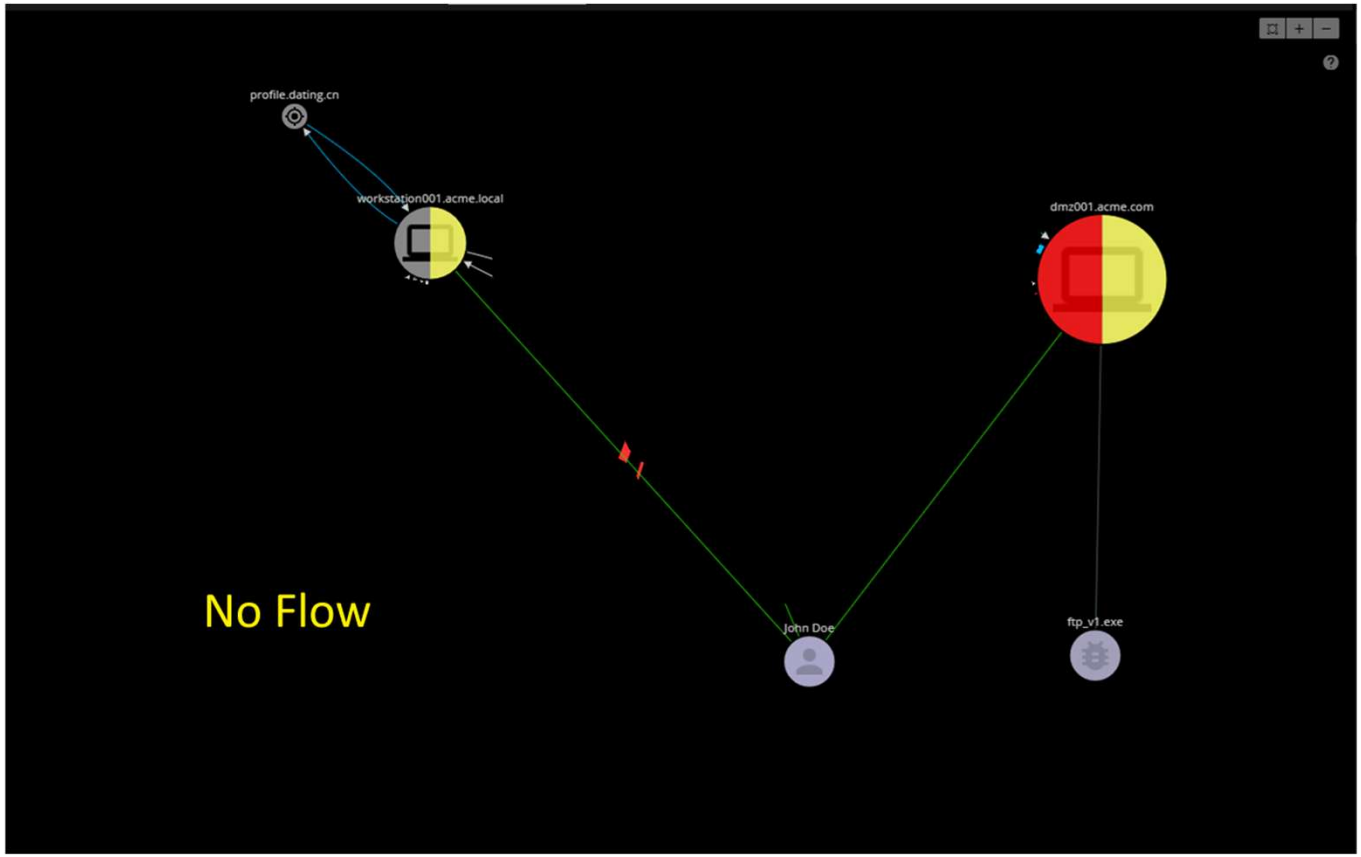
Identity signals

⚠ Removed — credential path unknown

⚠ Vulnerable to follow-on attacks using the same credential. The attacker's path is now invisible.

No Flow or DLP Signal

Network Defense



DLP & Flow

Identity Defense

Endpoint Defense

TRIAGE TRAP — REMOVE FLOW LOGS AND DLP

Remove flow logs and DLP

Graph becomes puzzle, not picture. Visible: 2 infected machines, 1 compromised credential. Most mature playbook: reimage and reset password.

2 infected machines

⚠ All that remains visible

1 compromised credential

⚠ Partial picture only

Endpoint signals

⚠ Removed

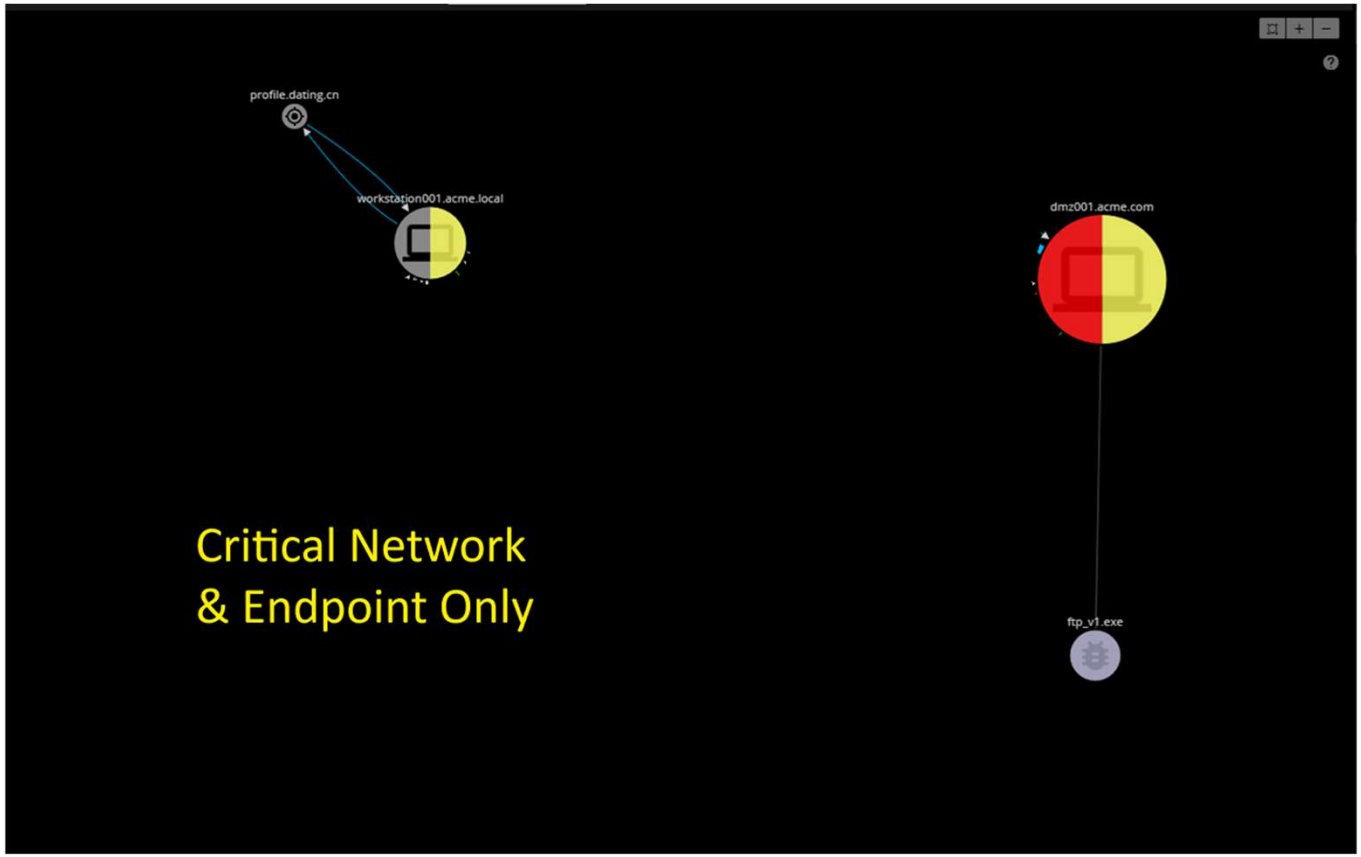
Identity, flow, DLP

⚠ All removed

⊗ No indication data theft occurred. By triaging one signal class, the case has been downgraded from prosecutable data theft to a help-desk ticket.

Triage Signals Only

Network Defense



DLP & Flow

Identity Defense

Endpoint Defense

TRIAGE ONLY — WHAT IS ACTUALLY LEFT

What triage-only actually leaves behind

2 machines to clean

Reimaged and returned to service

Compromised credentials

Still hidden — attacker retains access

Stolen data

Unknown — no evidence it occurred

The audit-completeness logic

If your tools triage signals, you cannot prove what didn't happen. This is the same logic an auditor applies to a transaction population.

Incomplete population → unsupported conclusion. The attacker wins not by evading detection, but by ensuring the evidence is never collected.

The last mile — GenAI doing what it is actually good at

Once the data below it is deterministic, GenAI's job becomes bounded:

> "Given this structured graph, write a narrative for this audience."

Narrative is constrained

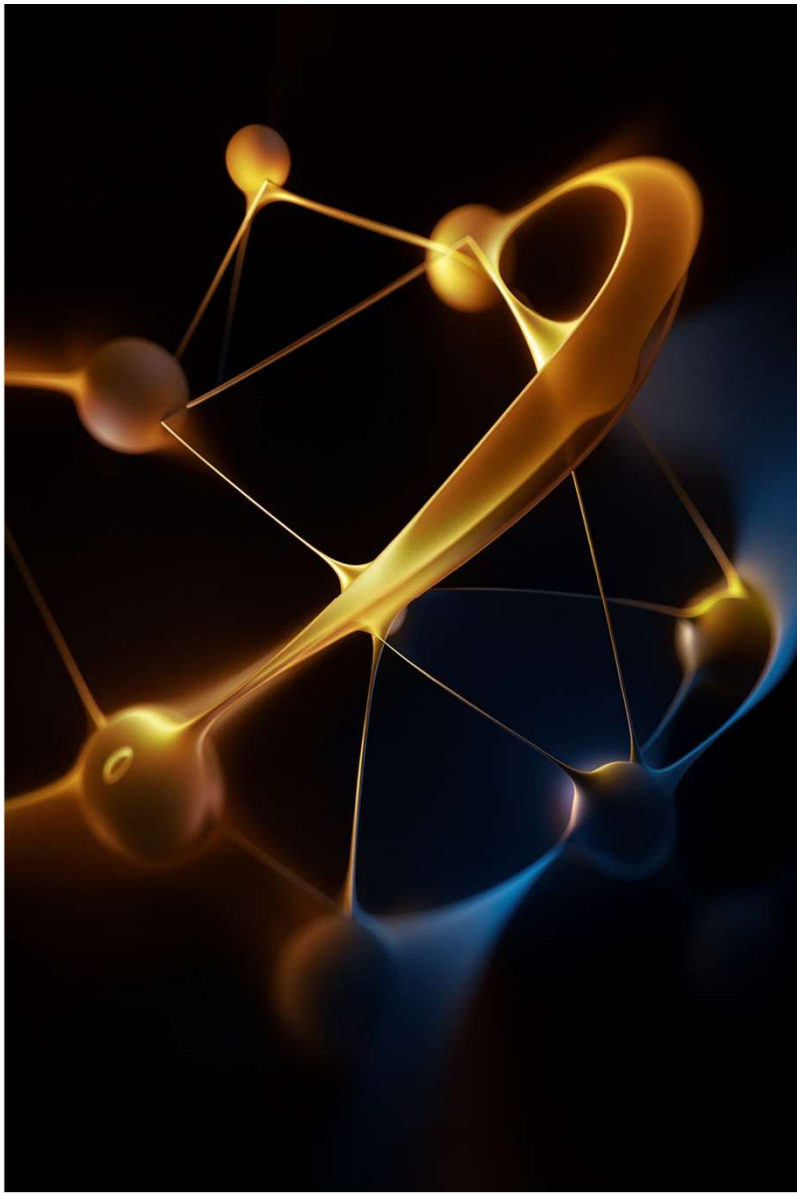
Structured evidence below the model limits what the model can say. It evaluates, not discovers.

Model is evaluating

The AI is not finding facts — it is describing facts that have already been established deterministically.

Auditable configuration

For audit reports, board materials, and regulatory filings — this is the only configuration where AI-generated narrative belongs near official artifacts.



CyberGrid — federated evidence, sovereign control

The deterrence argument rests on a simple chain of logic:

1

Deter cybercrime

Deterrence requires successful prosecution

2

Prosecution requires evidence

Court-grade evidence, not just logs

3

Evidence at scale

Requires federation across organisations

4

CyberGrid

Federated evidence with cryptographic chain of custody and sovereign control

CyberGrid — four EP properties make federation work

1

Normalisation to a shared schema

A login is a login regardless of source product. Shared vocabulary enables cross-organisation correlation.

2

Deduplication across sources

Collapses redundancy without losing evidence. The same event seen by multiple sensors becomes one enriched artifact.

3

Evidence-grade provenance

Cryptographic chain back to raw events. An indicator shared across the grid is not just data — it is evidence with a documented chain of custody.

4

Field-level redactability

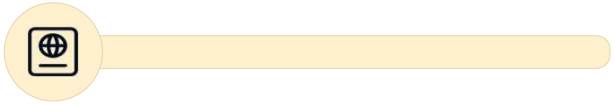
Sanitise victim details, preserve attacker indicators. Sovereignty preserved by architecture, not policy.

CyberGrid — three-tier law enforcement contribution

Default is "no contribution." Federation is opt-in by incident. The architecture does not compel disclosure — it makes disclosure useful when the participant chooses to provide it.

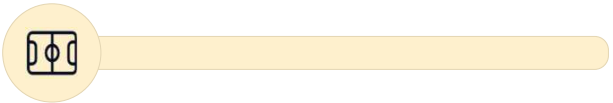
Anonymous tips

Sanitised indicators, attribution removed. Maps to long-standing anonymous-tip practice in law enforcement. No organisational exposure.



Structured indicator bundles

Named, attributed, restricted to law-enforcement group. Enables coordinated investigation across jurisdictions.



Digital affidavits

Full chain-of-custody evidence package, court-ready. The highest tier — full attribution, full provenance, ready for prosecution.

Reporter — compliance as continuous control test

Reporter is purpose-built for the conversations auditors have all week.

Tool-effectiveness reporting

Proves controls are *operating*, not just installed. The difference between a checkbox and a working control.

Compliance readiness

Continuous, not point-in-time. Every day is audit day, not just the day the auditor arrives.

Cost and savings analysis

Supports the ROI conversation with data, not assertions. Quantifies the value of security investment.

The shift Reporter encodes: from *"did this control exist on the audit date"* to *"has this control been working every day since the last audit."* This is the difference between a SOC 2 Type I and a Type II finding, automated.

Reporter — 12+ frameworks supported

Frameworks covered

- CIS Critical Security Controls v8 (CSC8)
- NIST Cybersecurity Framework 2.0
- NIST SP 800-53 and 800-171
- SOC 2 (Type I and II)
- ISO/IEC 27001
- HIPAA Security Rule
- PCI DSS 4.0
- GDPR Article 32
- NZ Privacy Act 2020 / Health Information Privacy Code
- Essential Eight (ACSC)
- CMMC Level 2 and Level 3
- FedRAMP moderate baseline

Three lines mapping

First line

Operates Conductor — the detection and collection layer

Second line

Consumes Reporter — continuous risk and compliance view

Third line

Tests Reporter output against framework definitions — independent assurance

Five questions to ask any AI security system

1 Where does deterministic processing end and generative inference begin?

The boundary between rule-based logic and probabilistic output must be explicit and documented.

2 Can any AI output be traced back to specific raw events?

If the answer is no, the output cannot be used as evidence. Full stop.

3 What is redacted vs preserved at the federation boundary, and who decides?

Sovereignty must be architectural, not a matter of policy or trust.

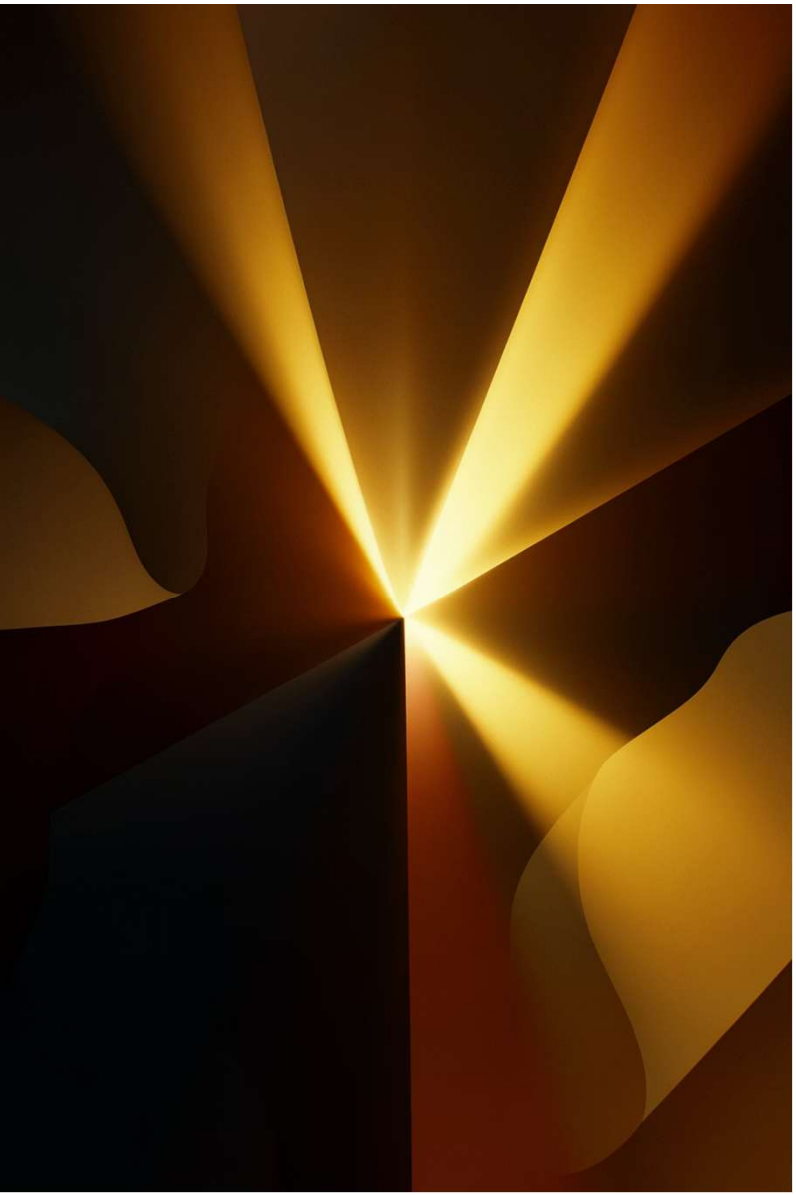
4 Which compliance frameworks does the system map to — and is the mapping continuous or point-in-time?

Point-in-time mapping is a snapshot. Continuous mapping is a control.

5 Could a defence attorney cross-examine the system's output without the training data?

If the answer is no, the output does not belong in a courtroom or a regulatory filing.

⊗ If everything is generative — run.



Thank you

Talk and references

charlesherring.com — slides and references will be posted after the conference

Open inbox

Charles@witfoo.com — reach out with questions, follow-ups, or to continue the conversation

Open-source AI toolkit

ArtiFish.dev — freely available tools for the community

Open dataset

Hugging Face: WitFoo Precinct 6 Cybersecurity Dataset — 114M labelled records, Apache 2.0

> **Bring this lens to every AI evaluation in your organisation.**

Q&A offline — out of respect for the next speaker.