WitFoo Releases Largest Structured Open-Source Cybersecurity Dataset

100 Million Records Built from Live Attack Traffic Released to Advance Cybersecurity Research at the University of Canterbury and Beyond

CHRISTCHRUCH, NEW ZEALAND, April 20, 2026 /EINPresswire.com/ -- WitFoo, Inc., a cybersecurity company specialising in security operations analytics, today announced the open-source release of the Precinct 6 Cybersecurity Dataset, containing 100 million structured, labelled security event records derived from live attack traffic observed between July and August 2024. Unlike lab-generated or synthetic alternatives, the dataset captures real adversary behaviour as it unfolded against production environments. The data has been sanitised to protect participating organisations but preserves the patterns, timing, and structure of actual attacks. The dataset is freely available on Hugging Face under an Apache 2.0 licence and was created in partnership with the University of Canterbury (UC) | Te Whare Wānanga o Waitaha to support academic research in cybersecurity, artificial intelligence, and data science.

The release expands WitFoo's initial 2 million record dataset (published earlier in 2026) by 50x, making it the largest structured, labelled, multi-source cybersecurity operations dataset available as open source. Critically, the data originates from real-world security operations, not controlled laboratory environments or traffic generators. The attack patterns, lateral movement sequences, and adversary tactics reflected in the dataset are those of actual threat actors, captured and processed through WitFoo's Empathetic Processing pipeline. The Precinct 6 dataset provides parsed, normalised, and enriched security signals ready for machine learning consumption, complete with MITRE ATT&CK mappings, provenance graphs, incident correlation, and security orchestration lifecycle metadata.

What the Dataset Contains

The dataset contains four integrated subsets: 100 million normalised Signals records covering syslog, Windows Security Auditing, VPC flow logs, and endpoint telemetry, with network metadata, hostnames, usernames, severity levels, and sanitised message content; Graph Edges and Graph Nodes providing provenance graph structures that map relationships between hosts, users, processes, and network connections; and Incidents, which hold correlated security events with binary classification labels, confidence scores, MITRE ATT&CK technique and tactic mappings, suspicion scores, and Security Orchestration, Automation and Response (SOAR) lifecycle metadata. Participant identities have been fully sanitised while preserving the statistical properties, temporal relationships, and behavioural patterns that make the data valuable for

research. The complete sanitisation codebase has also been released as open source, allowing researchers to inspect exactly how the dataset was generated and how participant data was protected.

Research Use Cases

The dataset's scale, structure, and labelling support a wide range of research applications: Intrusion Detection and Anomaly Detection: With 100 million labelled events across multiple log sources, researchers can train and benchmark supervised and unsupervised models for network intrusion detection, host-based anomaly detection, and multi-source event correlation.

Graph-Based Threat Detection: The provenance graph subsets enable research into graph neural networks, temporal graph analysis, and lateral movement detection, including graph-based approaches to Advanced Persistent Threat (APT) detection.

Large Language Models for Security Operations: The sanitised message content and structured metadata provide training and evaluation data for LLMs applied to security log analysis, automated triage, alert summarisation, and natural language querying of security data.

Benchmarking and Reproducibility: The dataset provides a common, large-scale benchmark for comparing detection algorithms, feature engineering approaches, and model architectures, with Apache 2.0 licensing removing barriers to reproducible research.

Supporting the Next Generation of Cyber Defence

"For a decade, WitFoo ran over 4,000 experiments with Fortune 500 companies, universities, and government agencies to develop Empathetic Processing. This dataset is the product of that research, and we believe it belongs in the hands of the academic community. Most publicly available cybersecurity datasets were generated in lab environments with scripted attacks and synthetic traffic. That's useful for basic benchmarking, but it doesn't teach you what real adversaries actually look like in a production network. We've sanitised the data to protect the organisations involved, and we've published the sanitisation code itself as open source so researchers can verify exactly how we did it. Cybersecurity's biggest bottleneck isn't compute or clever algorithms. It's the lack of realistic data that researchers can actually train against."
— Charles Herring, Chairman and Co-Founder, WitFoo, Inc.

"One of the persistent challenges in cybersecurity research is that most available datasets are either synthetic or derived from controlled laboratory exercises, which limits how well models trained on them generalise to real-world conditions. A dataset of this scale built from live attack traffic is genuinely rare. It opens up research pathways that simply weren't feasible before, from graph-based threat modelling to evaluating AI-driven detection systems against authentic adversary behaviour. We look forward to incorporating this resource into our research and teaching programmes at Canterbury."
— Dr Etienne Borde, Associate Professor, Computer Science & Software Engineering, University of Canterbury | Te Whare Wānanga o Waitaha

Availability
The Precinct 6 Cybersecurity Dataset is available immediately on Hugging Face at huggingface.co/datasets/witfoo/precinct6-cybersecurity-100m under an Apache 2.0 licence, free for academic, commercial, and government use. The complete sanitisation codebase is available on GitHub at github.com/witfoo/dataset-from-precinct6.

About WitFoo
Founded in 2016 and headquartered in Dunwoody, Georgia, WitFoo builds security operations platforms powered by its proprietary Empathetic Processing methodology, which achieves 98% reductions in compute, storage, and energy consumption for security analytics. WitFoo's products, Conductor and Precinct, serve enterprise, government, and defence customers through channel partners including Accenture Federal Services. WitFoo operates internationally through WitFoo Limited, based in Christchurch, New Zealand. For more information, visit witfoo.com.

About the University of Canterbury
The University of Canterbury (UC) | Te Whare Wānanga o Waitaha, established in 1873, is a leading New Zealand research university located in Christchurch. The Department of Computer Science and Software Engineering conducts research across cybersecurity, artificial intelligence, software engineering, and human-computer interaction.

WitFoo Press Office
WitFoo, Inc.
+1 678-203-9800
email us here
Visit us on social media:
LinkedIn
Facebook
X