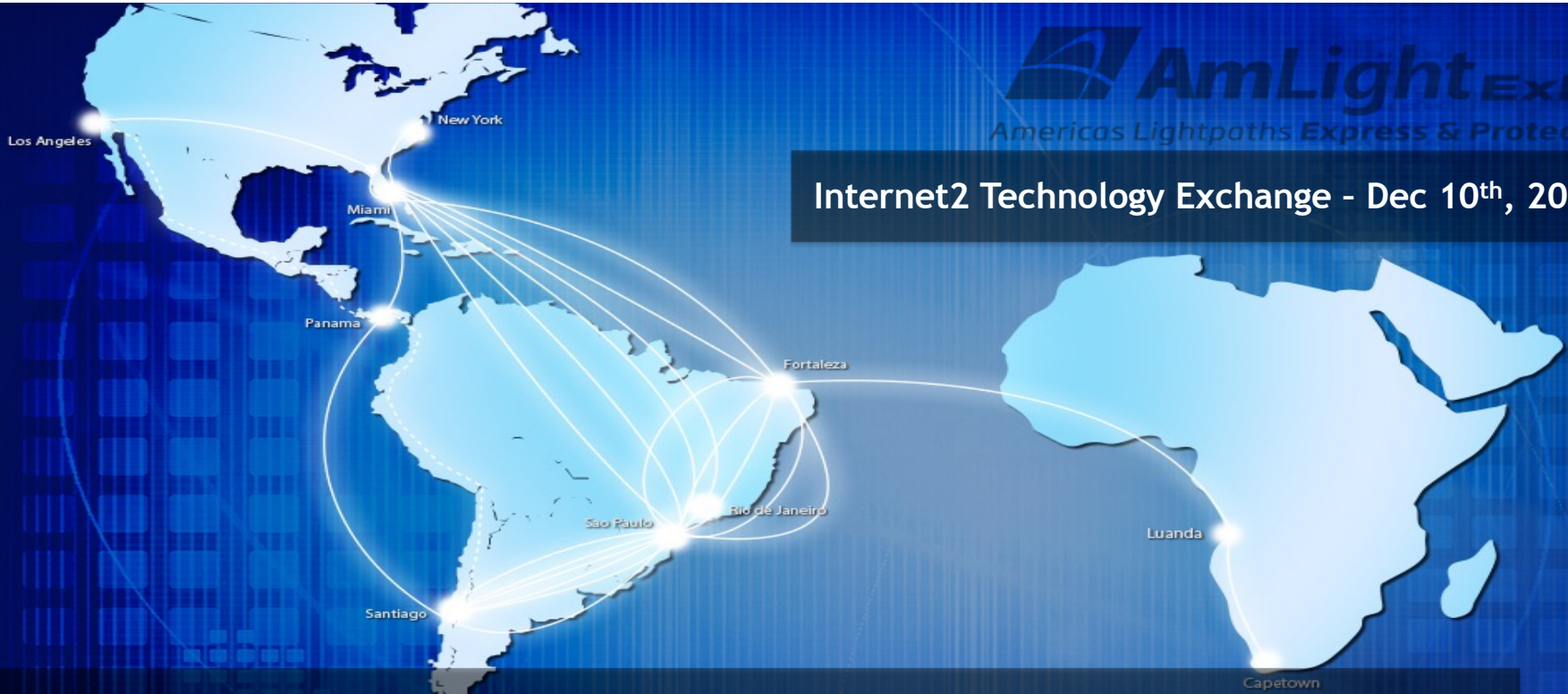




Internet2 Technology Exchange - Dec 10th, 2025



AmLight as an “AI-ready network”: First Steps and Challenges

Jeronimo Bezerra - Chief Network Engineer - FIU/AmLight

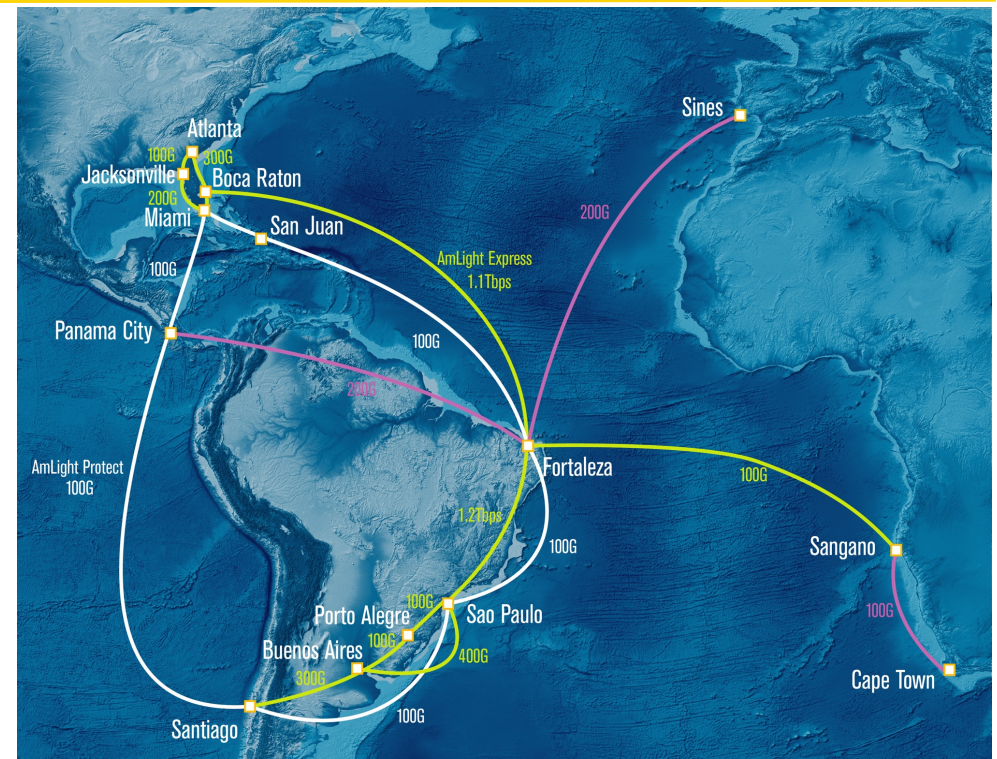
Renata Frez - Senior Network Engineer - RNP/AmLight

Outline

- What is AmLight?
- AmLight as a **Data Provider** to support ML and AI communities
 - Our journey and future
 - The new **AmLight: Next Frontier** project
 - The AmLight Data Pipeline
- AmLight as a **Data Consumer** to support its ML and AI initiatives
 - CICI: LLMDaL – *LLM-Driven Data Labeling* for Training Machine Learning Models
 - AI-Driven Network Operation
- Next Steps and Conclusion

Introducing the AmLight Network

- A distributed academic exchange point built to enable collaboration among Latin America, Africa, and the U.S.
 - Members: FIU, AURA, Vera Rubin Observatory, RNP, Rednosp, RedClara, REUNA, FLR, SANReN, TENET, and Internet2
- Supported by NSF and the IRNC program under award # OAC-2029283 for the 2021-2025
 - **Breaking news: AmLight: The Next Frontier for 2026-2030!**
- 4.9+ Tbps of total connectivity
 - A blend of optical spectrum, optical waves, and leased services
 - 1.7 Tbps of peering with global R&E networks
 - **2x 800Gbps** links activated in October 2025
- NAPs: Florida(3), Atlanta, Brazil(2), Chile, Puerto Rico, Argentina, Panama, and South Africa
- Infrastructure managed by a homemade SDN controller: Kytos-ng (github.com/kytos-ng)





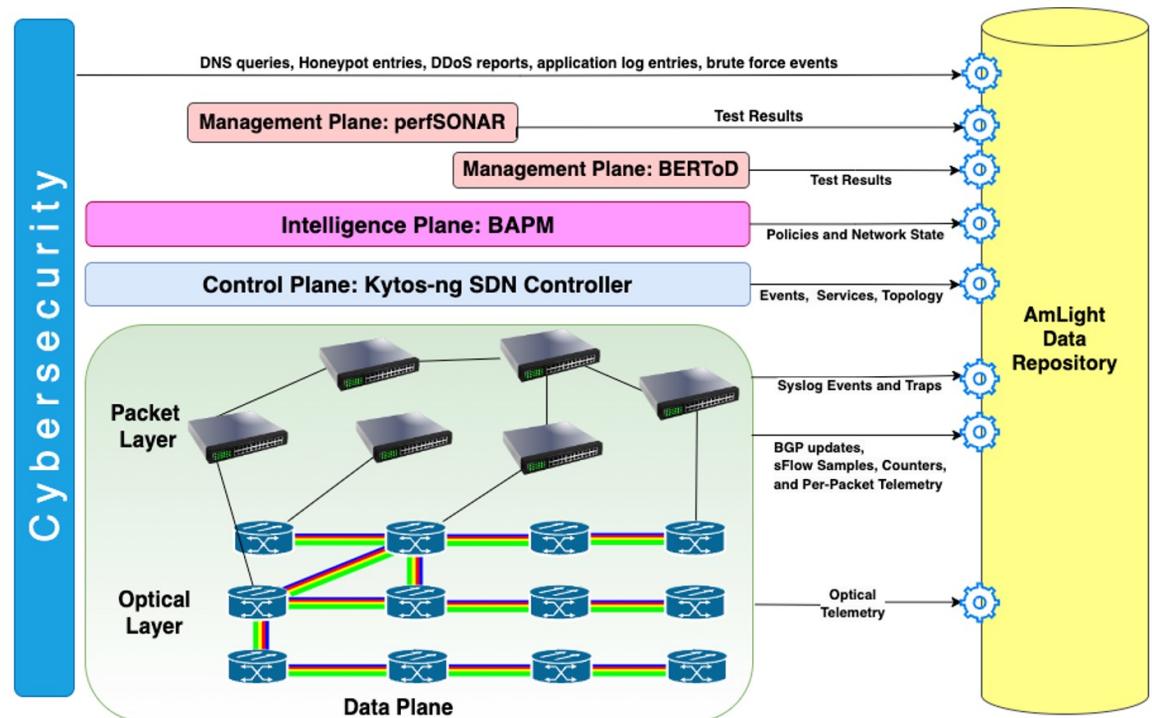
AmLight as a **Data Provider** to support ML and AI communities

How has our journey started?

- 2022, NSF creates a new program focused on creating network-related datasets
 - ML/AI communities were complaining about the lack of production datasets
- 2022, CICI LaSIC project is awarded to FIU and the University of Memphis
 - Goal was to create cybersecurity *labeled* datasets using the AmLight network telemetry.
- 2023, FIU recruited an AI engineer/researcher to support AmLight's AI plans
 - Focus was on enhancing our network and traffic optimization routines
- 2025, NSF motivates AmLight to share data through the Open Science Grid (OSG)'s Open Science Data Federation (OSDF) service (<https://osg-htc.org/services/osdf>)
 - The NSF 2026-2030 **AmLight: Next Frontier** award has data sharing as one of its goals

Sharing datasets with research communities [1]

- **AmLight: Next Frontier** is highly focused on supporting ML and AI communities, especially for **cybersecurity** and **environmental sensing**.
- 19 labeled data sources will become available openly.
- DDoS traces (Kentik reports and INT pcap files) will be the first datasets to be made available
 - Currently, 6TB of raw data.

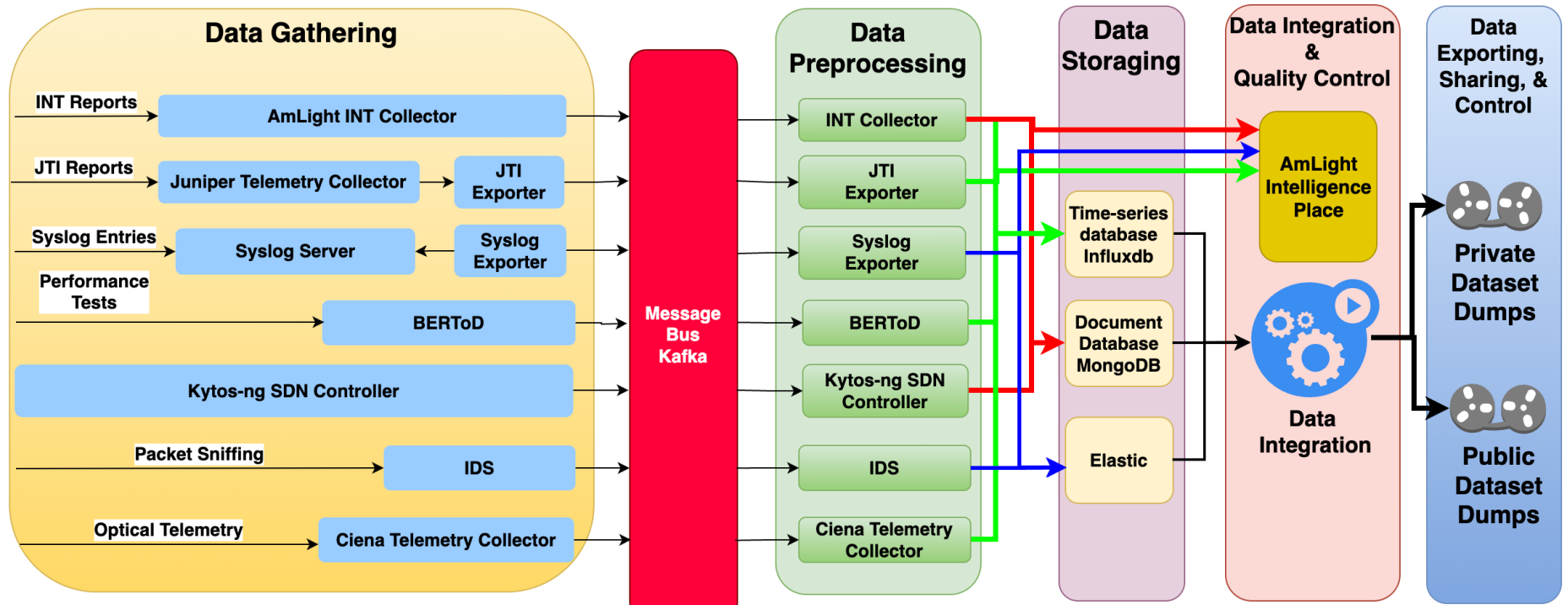


Sharing datasets with research communities [2]

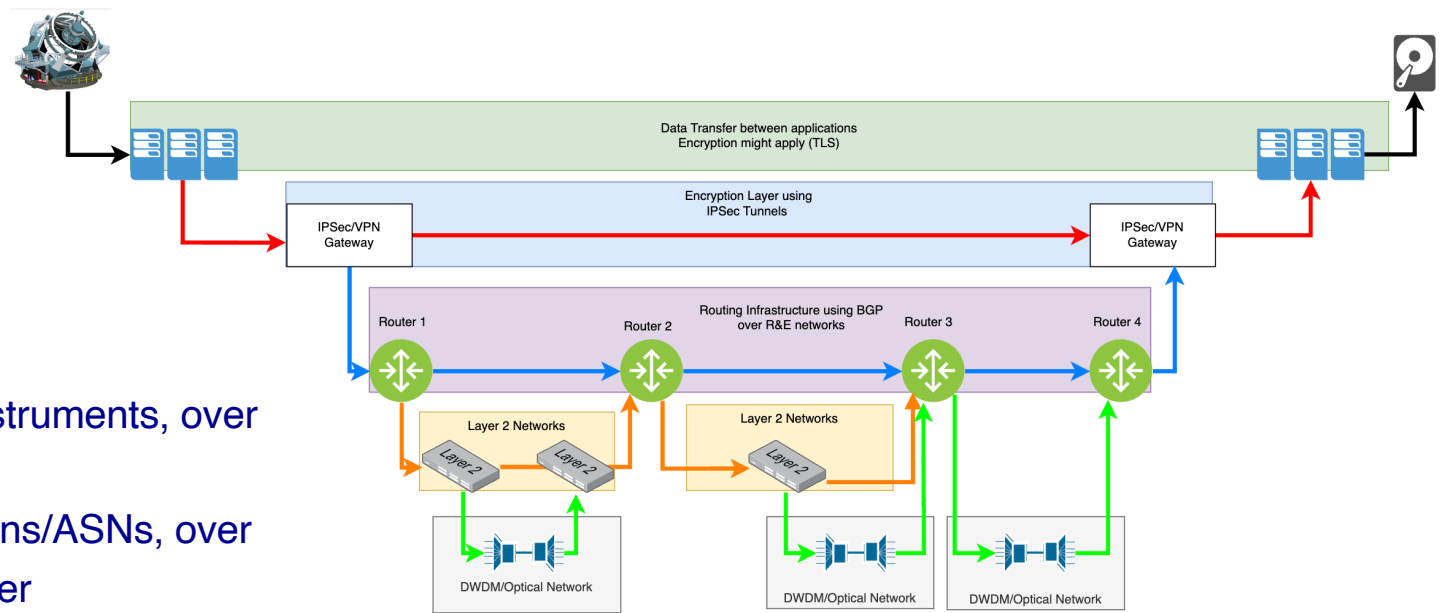
- Data exporting and sharing will be accomplished by leveraging several projects:
 - FIU's CICI **EnviStor** (200TB)
 - Open Science Data Federation (**OSDF**)
 - LaSIC: Labeled Security Information Capture (**LaSIC**)
 - Community Understanding of Network Datasets (**Comunda**)
- Findable, Accessible, Interoperable, and Reusable (**FAIR**) principles will guide our efforts
 - <https://www.go-fair.org/fair-principles>

Data Source	Main Fields of Research	Description	Data Type	Data Format	Dataset Size	Requires Anonymization?
DNS queries	Cybersecurity	DNS queries transported over AmLight	text	RFC 5424	MB	Yes
Honeypot entries		Attempts to attack AmLight honeypot	text	RFC 5424	MB	No
DDoS reports		DDoS reports	text	RFC 5424	KB	Yes
Application log entries		Attempts to attack AmLight applications	text	RFC 5424	KB	No
Brute force event logs		Attempts to brute force AmLight applications	text	RFC 5424	MB	No
perfSONAR test results	Performance Evaluation	perfSONAR test results	text	JSON	MB	No
BERTO D test results		BERTO D test results	text	CSV	MB	No
Implemented policies	Capacity Planning, Network Management	TE and Security policies implemented	text	YAML	KB	No
Network state		State of the network and topology	text	YAML	KB	No
Network events			text	JSON	MB	No
Network services			text	JSON	MB	Yes
Network topology			text	JSON	KB	Yes
Syslog interface flap events	Performance Optimization, Capacity Planning, Network Management, Performance Evaluation	Counters and events observed by the data plane	text	RFC 5424	MB	No
Syslog BFD flap events			text	RFC 5424	MB	No
Syslog BGP flap events			text	RFC 5424	MB	No
sFlow samples			binary	libpcap	GB	Yes
Interface counters			text	csv	KB	No
per-packet telemetry			binary	libpcap	GB+	Yes
optical measurements			text	JSON	MB	No

Enters the AmLight Data Pipeline

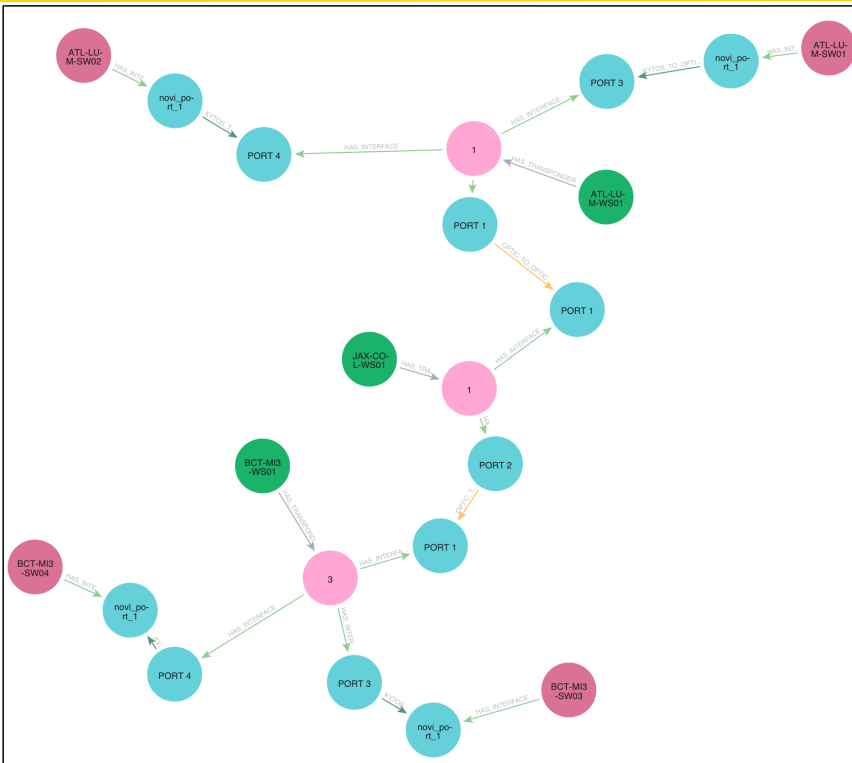


Challenge: How to represent the multi-layer topology? [1]

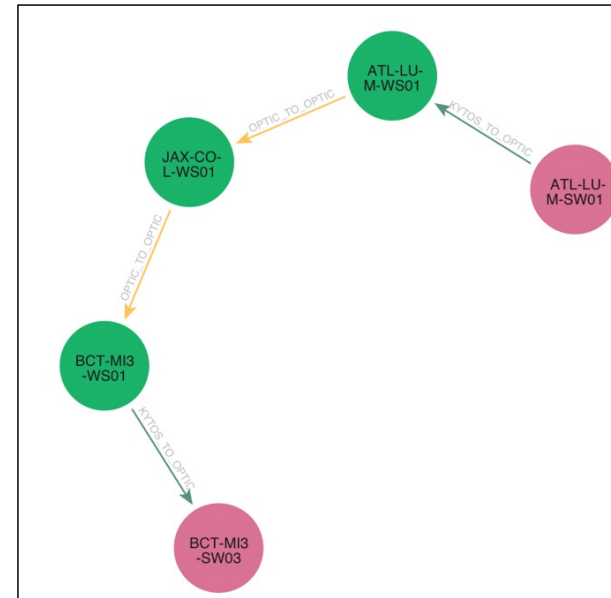


- User and Scientific Instruments, over
- VPN tunnel, over
- Multiple routing domains/ASNs, over
- Many L2 networks, over
- Many Open Line Systems, DCIs, and transceivers, over
- Many cross-connects, patch cords and panels, tie cables, etc.

Challenge: How to represent the multi-layer topology? [2]



```
MATCH path = (a {hostname: "ATL-LUM-SW01"})-
[:KYTOS_TO_OPTIC|OPTIC_TO_OPTIC*1..10]-> (b
{hostname: "BCT-MI3-SW03"}) RETURN path
```

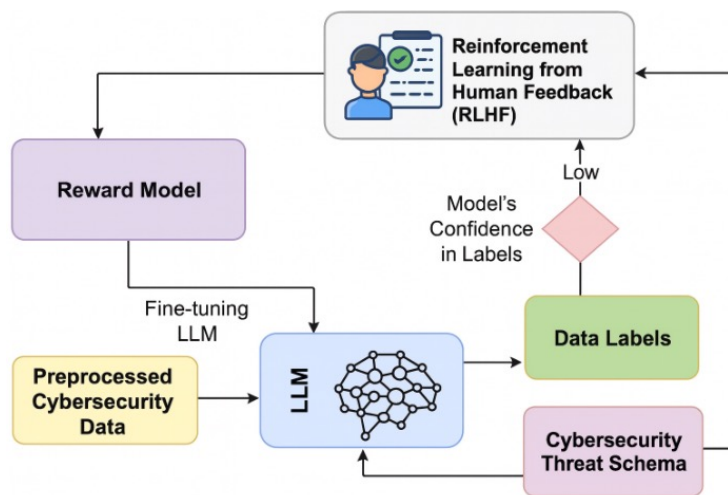


Neo4j



AmLight as a **Data Consumer** to support its ML and AI initiatives

LLMDaL: LLM-Driven Data Labeling for Training ML Models



The project uses Large Language Models (LLMs) to convert unstructured network data into high-quality, labeled cybersecurity datasets.

Why it matters:

- Addresses cybersecurity dataset scarcity
- Reduces manual labeling costs
- Enhances ML-based scientific cyberinfrastructures

AI-Driven Network Operation [1]

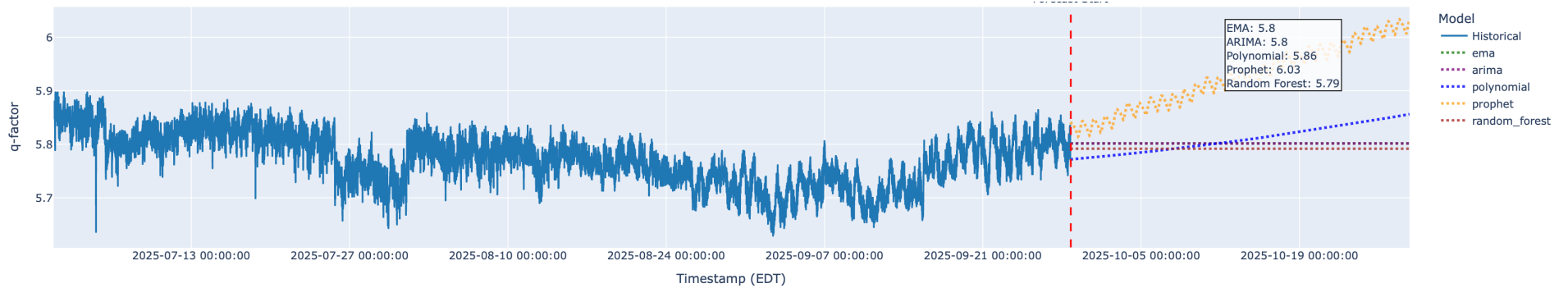
- AmLight has a **wide range of telemetry tools** and systems (Slide 7)
- For the Net Ops team, to **go through tool by tool** to look for issues, it's inefficient.
- Our next goal is to *build an AI environment that reads the primary telemetry sources, correlates events, and creates RT tickets* with enough content to expedite our manual troubleshooting.
 - One day, we expect the AI environment to handle some of those issues.
- For instance:
 - Anticipating packet loss caused by degradation on the photonics substrate:
 - Proactive load balancing of services across several assets and links, respecting user requirements.
- So, what next?

AI-Driven Network Operation [2] – How to Start?

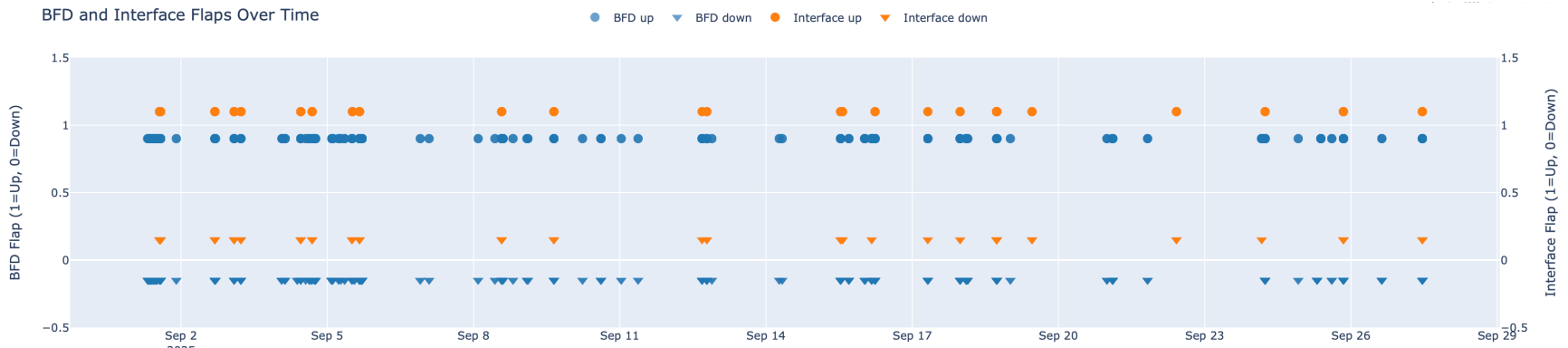
- *First*, find a clear problem where AI could be useful
 - Identifying the root cause of BFD (Bidirectional Forwarding Detection) flaps
- *Second*, document how we currently troubleshoot those BFD flaps:
 - Check for (1) interface flaps, (2) topology changes, (3) packet drop, (4) packet loss, (5) photonic issues, (6) routers' CPU utilization, (7) router alarms
- *Third*, locate the troubleshooting data sources:
 - Syslog events, In-band Network Telemetry (INT) counters, Optical telemetry counters, Kytos-ng SDN controller topology and service logs
- *Fourth*, cleaning data sources, syncing time across sources, and evaluating source's reliability
 - Time to correlate events that happen on the same path “around” the same time frame
 - BFD flaps only happen after 3 lost probes (2.25 seconds)

Important to mention: AmLight is just one side of the BFD session, and we don't have logs from the user side

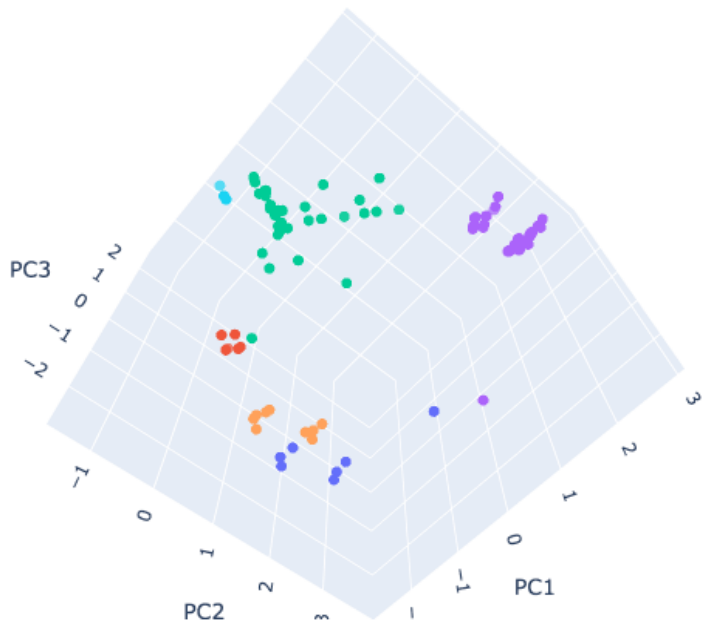
AI-Driven Network Operation [3] – First Results



BFD and Interface Flaps Over Time



AI-Driven Network Operation [4] – First Results



Clustering BFD flaps' events:

- Our goal: Correlate network events with BFD flaps and identify common patterns using AI clustering models (K-Means, DBSCAN, etc.).
- Our objective is to organize the cases into 3 groups:
 - Cluster 1: Event found anywhere on the path used by the BFD session.
 - Meaning: we have a perfect match.
 - Cluster 2: Event found outside of the path used by the BFD session.
 - Meaning: maybe the topology was not updated.
 - Cluster 3: No events found anywhere.
 - Meaning: the issue is within the user's network, or an unknown issue is occurring.
- With this grouping, we could automate the initial problem analysis and save the Engineers' time.

Analysis is still in progress.

Next Steps and Conclusion

- AmLight as a **Data Provider** to support ML and AI communities:
 - **AmLight: Next Frontier** should be making data available by **March 2026**.
 - Public data will be available initially via OSDF and Comunda.
 - We encourage the community to try our datasets and provide feedback regarding the FAIR principles.
- AmLight as a **Data Consumer** to support its ML and AI initiatives
 - **LLMDaL** will be the first step for AmLight engineers to engage with AI
 - We expect to build an AI system to focus on the BFD issue after learning with FIU's AI engineers and researchers.
 - We would love to collaborate with those interested in the field of AI for networking.



Thank You! / Questions? / Comments?



AmLight as an “AI-ready network”: First Steps and Challenges

Jeronimo Bezerra <jbezerra@fiu.edu>

Renata Frez <renata@amlight.net>