AmLight ExP
Americas Lightpaths Express & Protect

XVI WTR, PoP-BA, Sep 29th, 2025

Building a Data Pipeline for AI-Driven Network Operations:
Our Strategy to Support the AI/ML and Research Communities

Jeronimo Aguiar Bezerra
AmLight's Chief Network Engineer / Co-PI

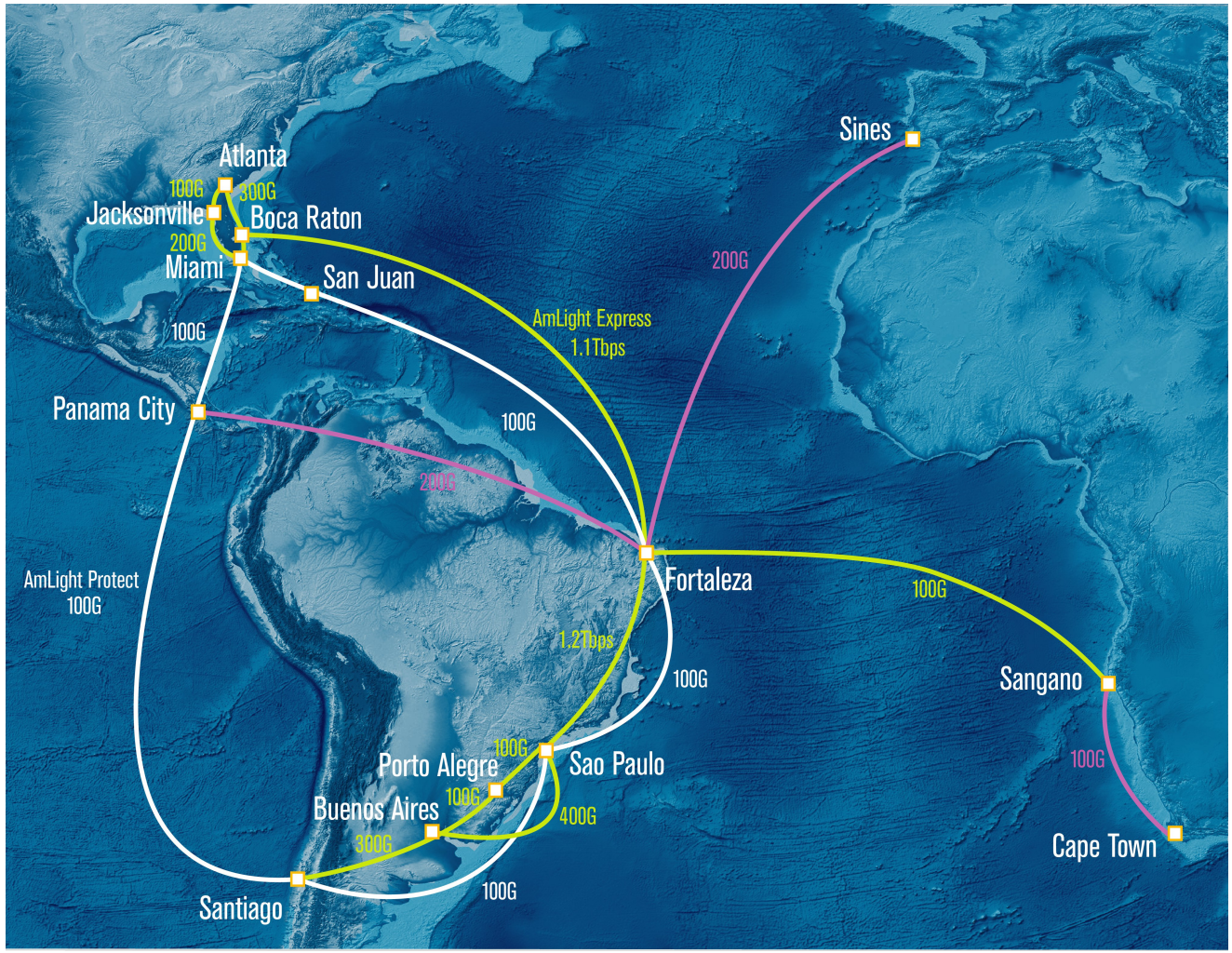
Outline

- 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
  - Enhancing our support for the Vera Rubin Observatory Long Haul Network

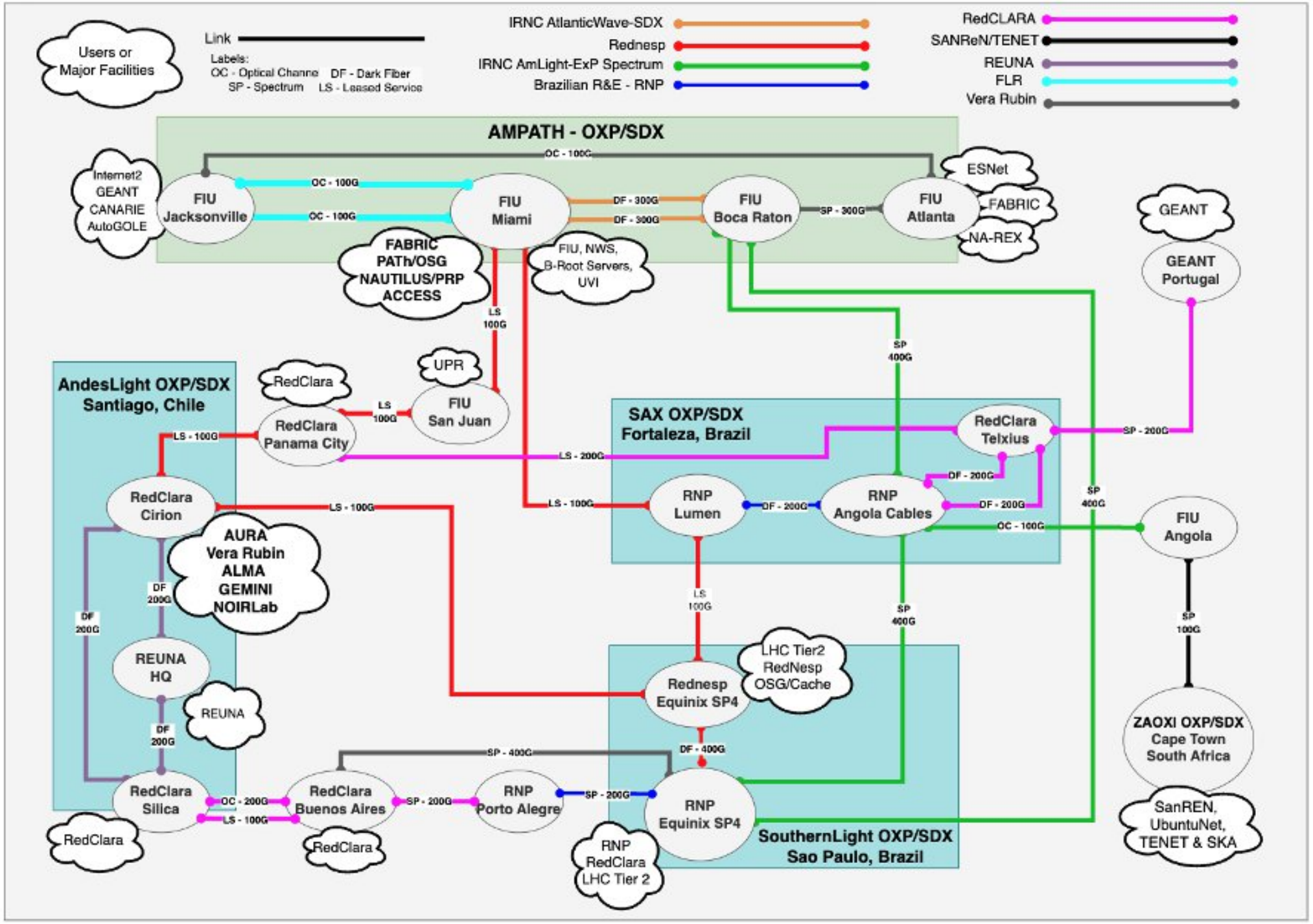
Introducing the AmLight Network

- A distributed academic exchange point built to enable collaboration among Latin America, Africa, and the U.S.
  - Members: FIU, AURA, Vera Rubin Observatory, RNP, Rednesp, 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
  - 1.7Tbps of peerings with global R&E networks
  - **2x 800Gbps** link to be 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 ExP
Americas Lightpaths Express & Protect

Intr



.ight ExP
Express & Protect

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

How has our journey started?

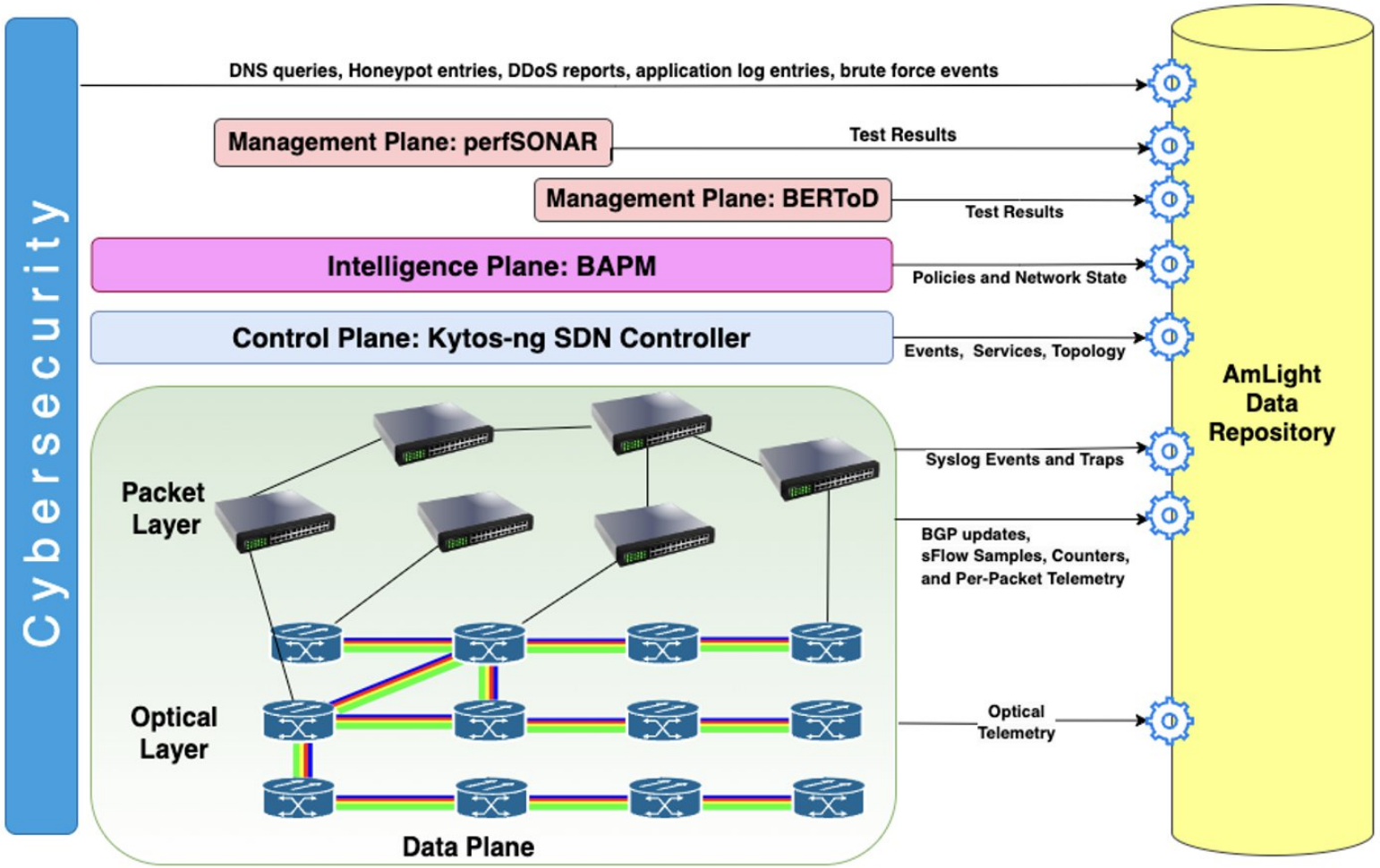
- 2022, NSF creates a new programs focused on creating network-related datasets
  - ML/AI communities constantly complaining about the lack of production datasets
- 2022, CICI LaSIC project is awarded to FIU and University of Memphis
  - Goal was to create cybersecurity *labeled* datasets using the AmLight 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.



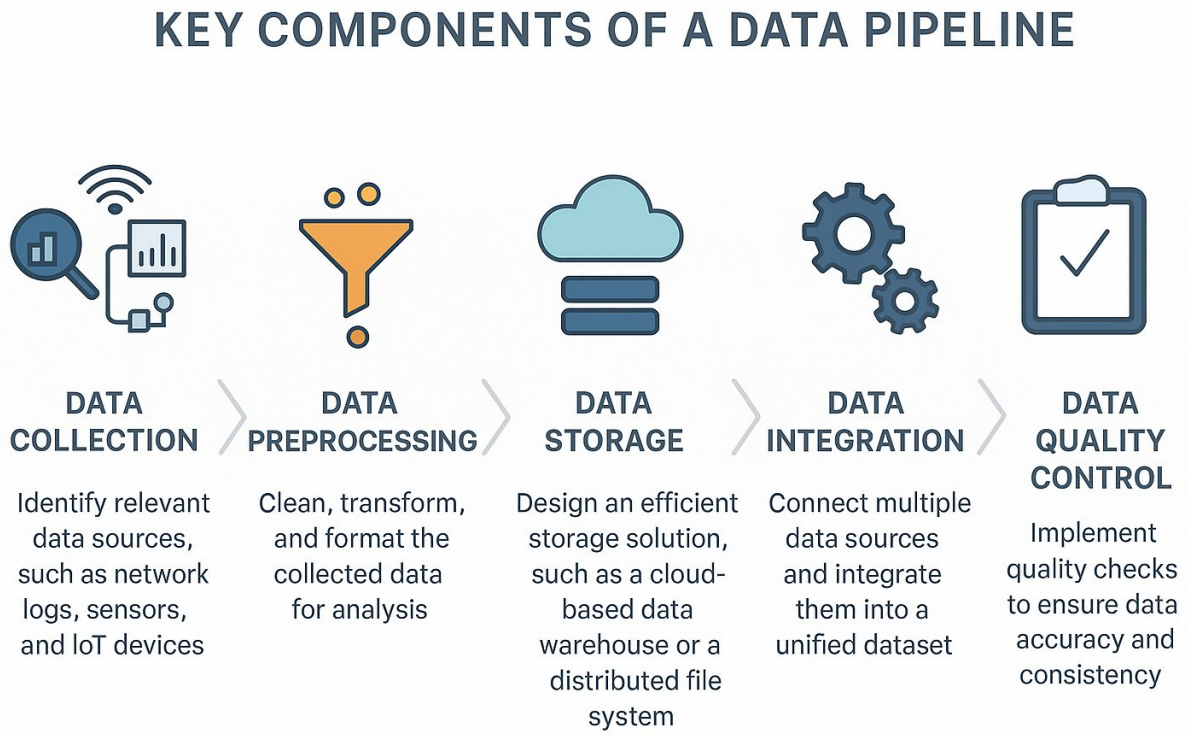
Sharing datasets with research communities [2]

- 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

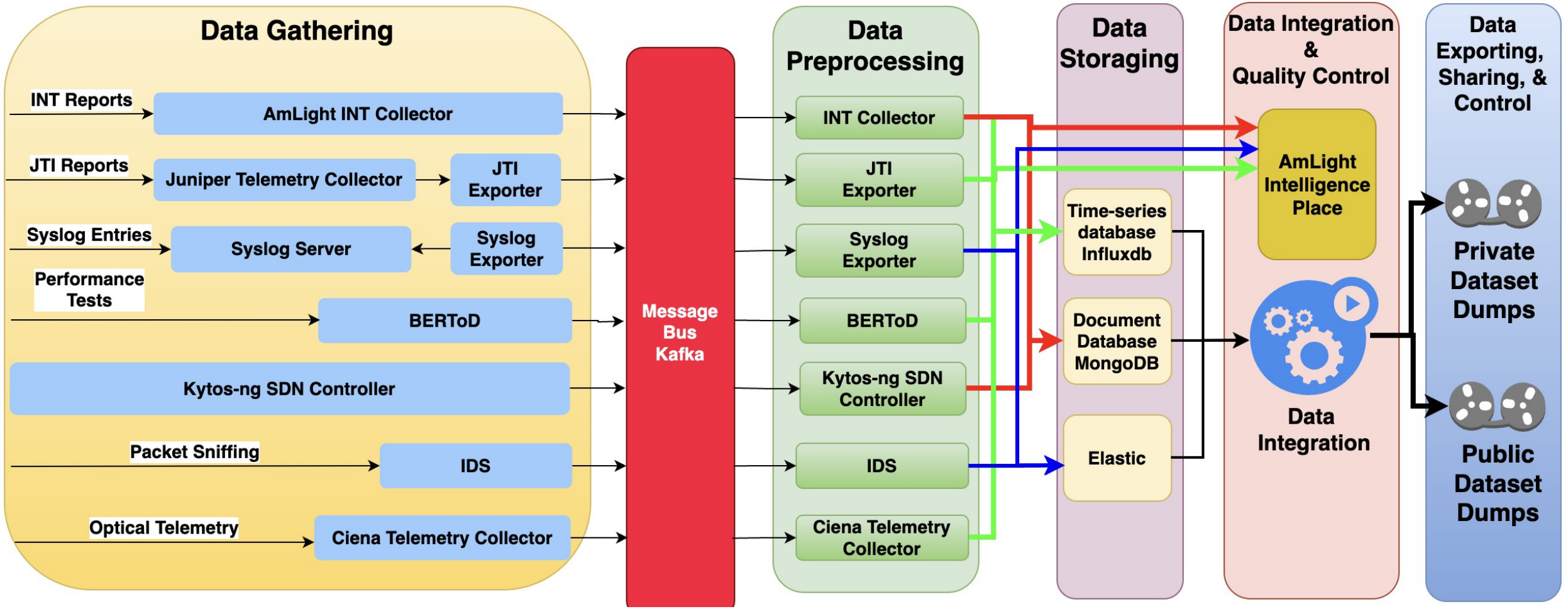
Data Source	Main Fields of Research	Description	Data Type	Data Format	Dataset Size	Requires Anonymization?
DNS queries	Cybersecurity	DNS queries transported over AmLight	text	RFC 5424	MB	Yes
Honeypot entries		Attemps to attack AmLight honeyport	text	RFC 5424	MB	No
DDoS reports		DDoS reports	text	RFC 5424	KB	Yes
Application log entries		Attemps to attack AmLight applications	text	RFC 5424	KB	No
Brute force event logs		Attemps to brute force AmLight applications	text	RFC 5424	MB	No
perfSONAR test results	Performance Evaluation	perfSONAR test results	text	JSON	MB	No
BERToD test results		BERToD 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*

- Data Collection:
  - Identify the specific data requirements for AI/ML research based on the sources available at AmLight, such as network logs, INT, and sensors.
- Data Preprocessing:
  - Clean, transform, and format the data collected and anonymize fields when necessary.
- Data Storage:
  - Identify efficient storage solutions that support easy public data access as well as real-time consumption by AmLight's Intelligence Plane.
- Data Integration:
  - Augment the dataset with other data sources, for instance, network topology and services, to create a unified dataset to be easily consumed.
- Data Quality Control:
  - Perform quality checks to ensure data accuracy, consistency, and relevance.



Enters the *AmLight Data Pipeline [2]*



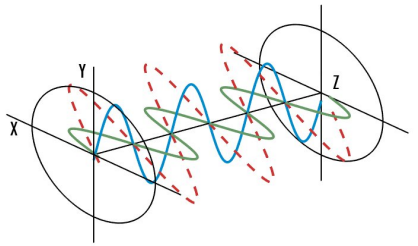
AmLight ExP
Americas Lightpaths Express & Protect

A Use Case: Environmental Sensing

- Existing fiber networks and coherent transceivers can support discovery of events around the fiber:
  - Earthquake, construction/digging, winds, and other vibration-related events.
- Evaluating the State-of-Polarization (SOP) metrics exported by coherent transceivers in submarine and land-based fiber-optic cable using Machine Learning has led to transformative discoveries.
- AmLight: Next Frontier will support the environmental sensing community by:
  - Installing new transponders capable of collecting and exporting SoP data
  - Sharing the measurements with the community

Using Global Existing Fiber Networks for Environmental Sensing

Localization of seismic waves with submarine fiber optics using polarization-only measurements



Fiber Optic Cables Detect and Characterize Earthquakes

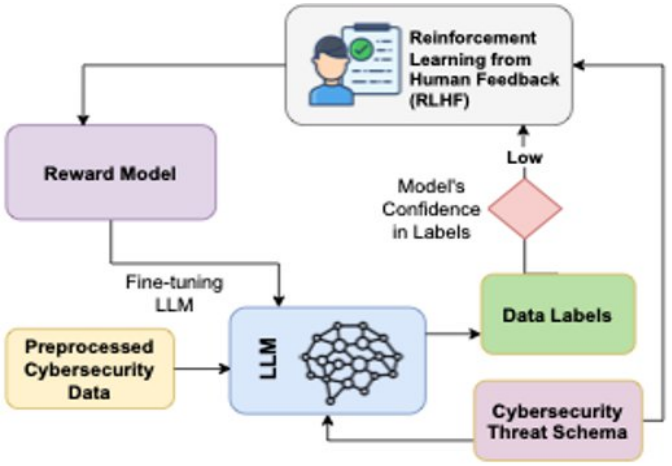
Fiber-Optic Cables Are Natural Earthquake Detectors › Optical cables could give early warnings through dense, low-cost arrays

Internet Sensor Network Testbeds

AmLight ExP
Americas Lightpaths Express & Protect

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

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

- Cybersecurity innovation
  - Fine-tunes open-source LLM with AmLight network data and cybersecurity knowledge
  - Builds LLM-powered Labeling Agent to automatically and accurately label network data for AI-based network solutions
  - Earlier methods include manual review simulations, controlled setups, and honeypots
    - These methods are time consuming could not scale with evolving threats, and lack real world-complexity
- Approach for Transitioning the Innovation
  - The project employs a new LLM-based methodology to generate reliable labels for cybersecurity data by combining:
    - Retrieval-Augmented Self-Refinement, Expert Verification, and LLM Ensemble

- Benefits to Scientific Cyberinfrastructure
  - The project produces labeled datasets from AmLight, a real scientific network maintained at FIU
    - Relevant to the unique traffic and threat patterns of scientific cyberinfrastructures, which are typically not captured in simulated or generic datasets
  - An LLM-based automated labeling approach that reduces cost and manual effort
    - Lowers one of the main barriers to creating such datasets
    - Enables us to augment the data in line with the evolving threat landscape
  - As a dataset with real-world production data will helps us test AI based algorithms for automated network security and management solutions
  - Provides valuable experience in developing and operationalizing scientific data products.

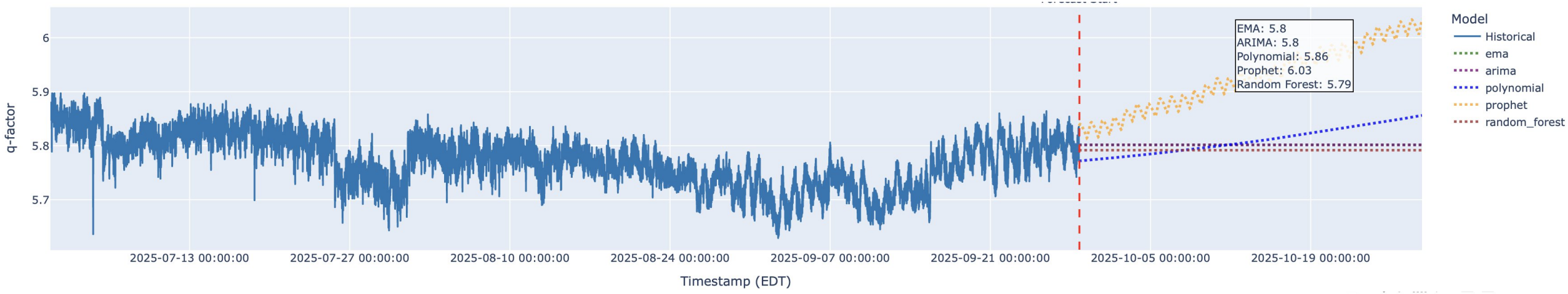
- Evaluating and Demonstrating Transition
  - Main metric Accuracy
    - Comparing LLM-generated labels against expert annotations on real world incidents and benchmark datasets
  - Other success metrics:
    - Ability to emerging threat landscapes
    - Reduction in time and human effort for data labeling
- The project will make its code, fine-tuned LLMs, and Labeling Agent publicly available under open-source license, along with the labeled datasets
  - Researchers and organizations can easily replicate, build on, and improve the work.

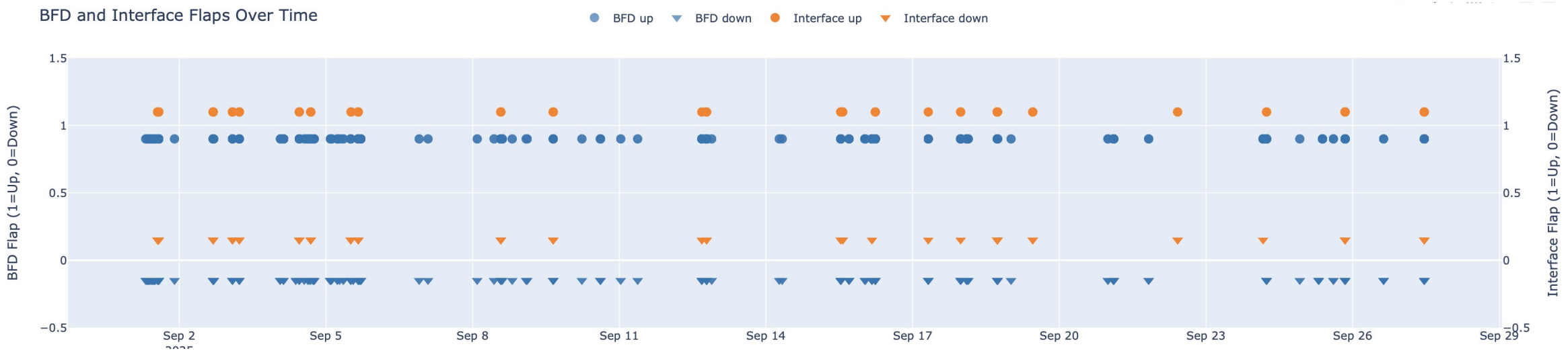
AI-Driven Network Operation

- *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 troubleshoot 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, sync time across sources, evaluate each 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 of side of the BFD session and we don't have logs from the user side

AI-Driven Network Operation: Some Outcomes





AmLight ExP
Americas Lightpaths Express & Protect

AI-Driven Network Operation [3]

- AmLight has a wide range of telemetry tools and systems, from optical substrate to applications, from real-time to historical, from applications events to active performance measurement results.
- 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 main telemetry sources, correlates events, and create 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
    - Optical issues can be "predicted" based on existing telemetry and AI could automate the steering of services.
  - Proactive load balancing of services across several assets and links, always respecting user requirements.
    - Most of the 19 data sources previously listed have to be used to support this boal.

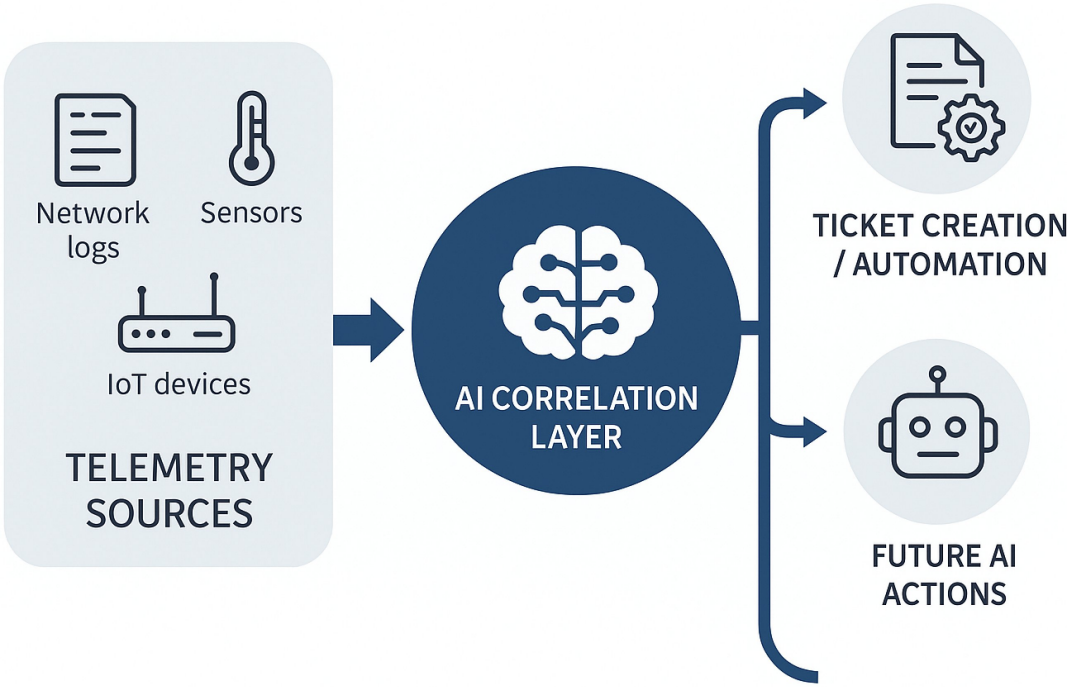
- There is a long road ahead, but we are on the right track.

AI-Driven Network Operation [3]

- **Next Steps with AI in AmLight:**
  - AmLight already collects telemetry across all layers: optical to applications, real-time to historical.
  - Manual “tool-by-tool” troubleshooting is inefficient for the NetOps team.
  - **Next goal:** build an AI environment to read main telemetry sources, correlate events, and auto-create RT tickets with rich context.
  - **Future vision:** AI handling some issues end-to-end.

- **Examples:**
  - Anticipate packet loss from photonic degradation.
  - Predict optical issues and automate service steering.
  - Proactively load balance across assets and links while honoring user requirements.
  - Most of the 19 data sources will feed this AI environment.

Long road ahead, but we are on the right track.



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 in regards 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.

Building a Data Pipeline for AI-Driven Network Operations:

Our Strategy to Support the AI/ML and Research Communities