NIST Special Publication 2100-04

Summary: Workshop on Machine Learning for Optical Communication Systems

Josh Gordon
Abdella Battou
Michael Majurski
Dan Kilper
Uiara Celine
Darko Zibar
Massimo Tornatore
Joao Pedro
Jesse Simsarian
Jim Westdorp

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.2100-04

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 2100-04

Summary: Workshop on Machine Learning for Optical Communication Systems

Josh Gordon
NIST, Communications Technology Lab (CTL)

Abdella Battou, Michael Majurski
NIST, Information Technology Lab (ITL)

Dan Kilper
University of Arizona, James C. Wyant College of Optical Sciences

Uiara Celine, Darko Zibar
Technical University Denmark

Massimo Tornatore
Politecnico di Milano, Italy

Joao Pedro
Infinera

Jesse Simsarian
Nokia-Bell Labs

Jim Westdorp
Ciena

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.2100-04

March 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Publications in the SP 2100 subseries are proceedings from conferences organized predominately by NIST scientific and technical staff. These proceedings are published as a single document that includes all abstracts or extended abstracts accepted by the conference organizers. This publication may include external perspectives from industry, academia, government, and others. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the United States Government.

National Institute of Standards and Technology Special Publication 2100-04
Natl. Inst. Stand. Technol. Spec. Publ. 2100-04, 22 pages (March 2020)
CODEN: NSPUE2

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.2100-04

Abstract

Optical communication systems are expected to find use in new applications that require more intelligent and automated functionality. Optical networks are needed to address the high speeds and low latency of 5G wireless networks. The analog nature of optical transmission and the complexity of operation and management remain an impediment to greater use of software controls. The optical community at large has proposed many possible applications and avenues for using and implementing artificial intelligence and machine learning to improve functionality of optical systems for communications. However, broad agreement has yet to be reached due to both technical and non-technical reasons. On August 2nd , 2019 The National Institute of Standards and Technology (NIST) Communications Technology Laboratory (CTL) hosted a Workshop on Machine Learning for Optical Communication Systems to bring together industry, academia and government in order to discuss the roll of AI and ML in optical communication systems. This document provides an overview and summary of the workshop.

Key words

Artificial Intelligence; Disaggregation; Machine Learning; Network Defragmentation Open Line; Optical Networks; Quality of Transmission (QoT); Reconfigurable Optical Add Drop Multiplexer (ROADM); Software Defined Networks (SDN); Transponder

Conference Organizing Committee Members

Josh Gordon (NIST-CTL-Chair Person)
Abdella Battou (NIST-ITL-Co-Chair Person)
Ari Feldman (NIST-CTL)
Nada Golmie (NIST-CTL)
Michael Garris (NIST-ITL)
Dan Kilper (University of Arizona)
Tod Sizer (Nokia-Bell Labs)
Martin Birk (AT&T)
Marc Lyonnais (Ciena)

Acknowledgments

Special thanks to NIST leadership in particular Marla L. Dowell the Director of the Communications Technology Laboratory (CTL) and Paul Hale (Acting Division Chief) for supporting this workshop as well as the CTL admin team including Cindy Kotary and Chris Carson, and Brian Copello. We would also like to thank the scribes Ari Feldman, Tasshi Dennis, and Alex Yuffa for their attentiveness and hard work as well as the workshop participants who provided invaluable input and contribution throughout the workshop and during the breakout sessions, a summary of which is given in this document.

Table of Contents

1. Overview

Optical communication systems are expected to find use in new applications that require more intelligent functionality. Optical networks are needed to address the high speeds and low latency of 5G wireless networks. The analog nature of optical transmission and the complexity of operation and management remain an impediment to greater use of software controls. Furthermore, optical systems are running up against spectral density limits that threaten traditional capacity-based scaling. New efficiency-based scaling methods are needed to further improve the cost/bit/s without relying on capacity improvements alone. Artificial intelligence (AI) and machine learning (ML) provide a new direction with the potential to both enable wider use of software controls and to further optimize the efficiency of optical systems across multiple dimensions. Reference data sets for ML would improve functionality and operability across industry further enabling scaling and efficiency.

On August 2nd 2019, the National Institute of Standards and Technology (NIST) Communications Technology Lab (CTL) in partnership with the Laboratory for Information Technologies (ITL) held a Workshop on Machine Learning for Optical Communication at the Boulder Colorado campus. The purpose of this workshop was to bring together industry, academia and government to discuss the role of ML in optical communication systems (MLOS). Topics discussed during the workshop ranged from identifying applications of AI and ML in the context of accelerating the use of software-based networking in optical systems for improved performance and scalability, paths to realizing reference training data sets for ML in optical communications systems and needs and rolls of metrology and telemetry.

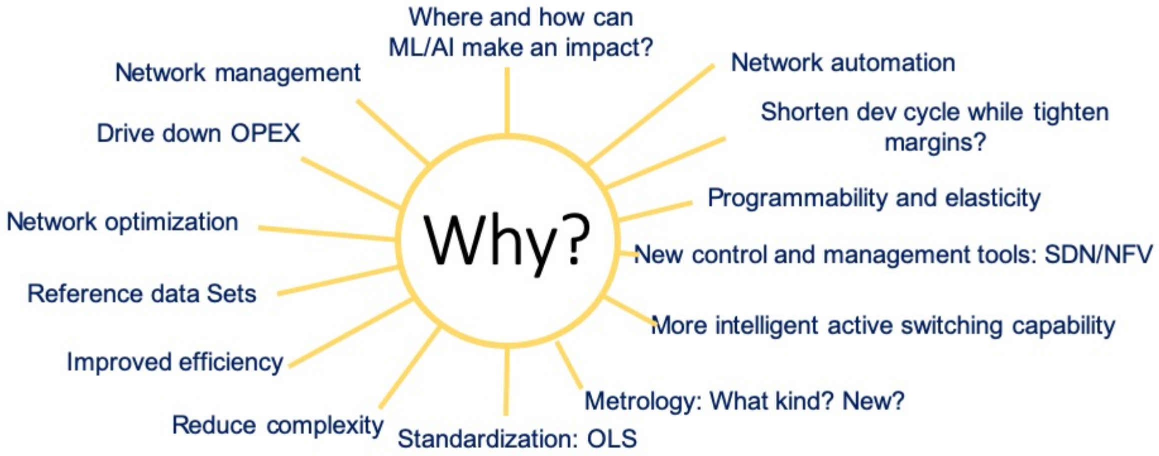
Why?
Where and how can ML/AI make an impact?
Network management
Network automation
Drive down OPEX
Shorten dev cycle while tighten margins?
Programmability and elasticity
Network optimization
New control and management tools: SDN/NFV
Reference data Sets
More intelligent active switching capability
Improved efficiency
Metrology: What kind? New?
Reduce complexity
Standardization: OLS

Figure 1. Why Machine Learning for Optical Communications? There are many proposed reasons "Why".

The workshop attempted to spur open public discussion of the many perspectives from industry, government and academia as to "Why" machine learning might be a good idea to implement in optical communication systems and transport networks (see Fig. 1). The role and potential impact of ML on optical communications has been discussed at large from many points of view and with regards to many applications. However, a road map to implementation, realizing positive change, and actional paths toward progress have yet to be agreed upon. As there are many, many facets to implementing ML on such a large scale, bringing together the many points of view in an open forum to share information is beneficial in identifying actional paths forward. The workshop was designed to stimulate questions, discuss, and collaboration, culminating in three breakout sessions covering three topics relevant to ML and optical communications systems: 1) Data from reconfigurable optical add drop multiplexer (ROADM) and optical layer, 2) Possible data sets from coherent transponders, 3) Cross layer end-to-end networking. These areas were chosen as they address overarching architectures present in optical communication networks including: core optical network, transponders, cross layer networking. The discussions which arose from these breakout sessions are summarized below.

The workshop format was setup up to guide discussion through organized talks which lead to two panel discussions. The panel discussions then lead to three breakout sessions where all workshop attendees participated in discussion. Talks given by industry, academia and government speakers covered a wide range of topics including: an overview of Machine Learning Applications for Optical Transport Networks, Machine Learning Models, Data relevance-what Data Matters, Data starved systems, data from ROADM and the optical layer, cross layer and multi-vendor end-to-end networking. A summary of the breakout sessions is given below. The agenda, and slides from for the talks given at the workshop can also be found at the website: https://www.nist.gov/news-events/events/2019/08/machine-learning-optical-communication-systems

2. Talk Summaries

2.1. Keynote: "Machine Learning for Optical Communication Systems"

Speaker: Massimo Tornatore (Politecnico di Milano, Italy and University of California, Davis)

Modern optical networks generate an extremely high and diverse set of data (e.g., signal quality indicators, network alarms, etc.). Machine Learning (ML) is being regarded as a promising solution to extract useful information from these large datasets and possibly enable advanced forms of network automation, e.g., network self-reconfiguration and cognitive fault management. The keynote talk was intended to provide an introductory reference for researchers and practitioners in the audience. It started by providing a high-level introduction to some important theoretical concepts in machine learning (supervised vs unsupervised learning, basic algorithms), assisted by some examples coming from the field of optical communications. Then, the most relevant applications of ML to optical communications and networking have been overviewed, with a specific focus on QoT estimation and failure management. As several use cases can benefit from the application of ML techniques, in the talk these use cases were divided in i) physical layer and ii) network layer use cases, as graphically shown in Fig. 2. A good number of research papers appeared in the past years were covered. Finally, the talk discussed new possible research directions in the field. The talk was based on a recent survey [1] and a recent tutorial, [2] published by the speaker.

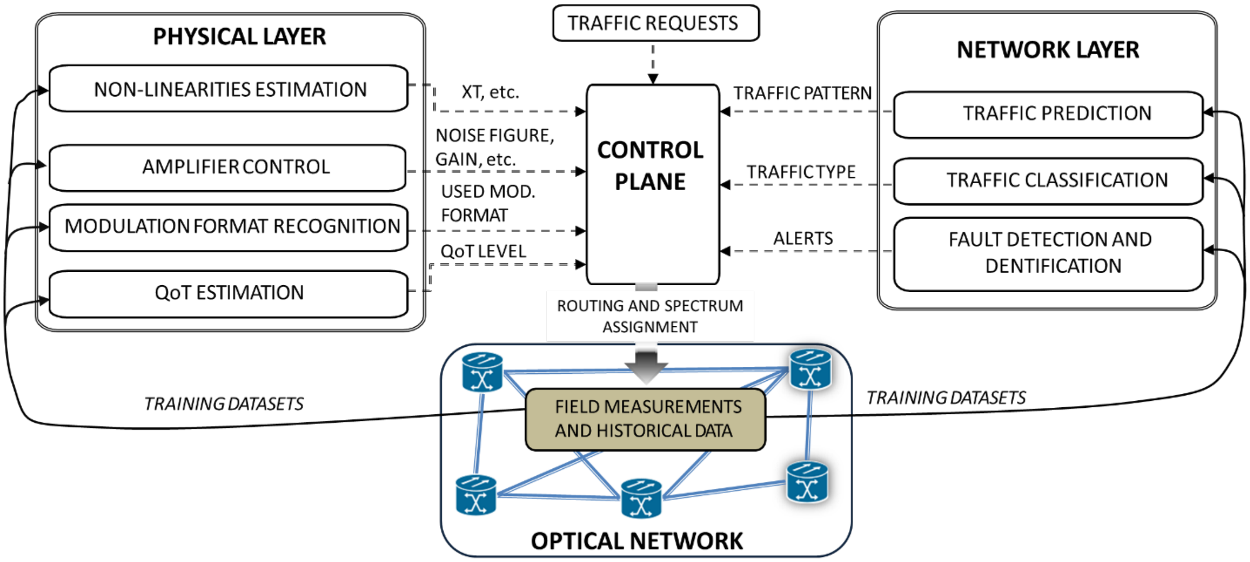

Figure 2. ML applied to the use cases of the physical layer and network layer.

2.2. “What Data Matters”

Speaker: Uiara de Moura (Technical University of Denmark)

Starting the morning talks section, Uiara de Moura, from the Technical University of Denmark (DTU), introduced the vision of the Machine Learning in Photonic Systems (M-LiPS) group regarding what kind of data matters for optical communications and specifically for machine learning applications. She started introducing the M-LiPS group at DTU Fotonik department, a new group led by Assoc. Prof. Darko Zibar. The group was created at the beginning of 2019 with the main goal of concentrate efforts in developing and applying machine learning techniques to photonic systems in general. More specifically, M-LiPS acts on the fields of advanced photonic classical and quantum measurements, communication and sensing systems, and device and subsystem designs.

Following the short introduction, she presented the results of the group’s recent activities in the area of phase noise characterization for lasers and frequency combs [3,4], Raman amplifier inverse design and modelling [5-7], and auto-encoders for optical communication systems [8-10]. Additionally, to have a comprehensive view of the main applications of machine learning in optical communications and the current state-of-the-art, she discussed the key point of a few relevant surveys on the topic [11-14].

She also highlighted the areas that will benefit from applying machine learning, such as: noise characterization of lasers at the quantum limit (low signal-to-noise ratio); (inverse) design of optical subsystems, components and devices for traditional and spatial division multiplexing systems; and techniques to increase the transmission rate through the nonlinear fiber optical channel. Moreover, as requested by Prof. Dan Kilper, she also pointed out the problems in optical communications that may not require the power machine learning to be addressed. As examples she singled out linear impairment compensation (e.g. chromatic dispersion) in coherent systems and erbium doped fiber amplifier design for traditional systems.

Before reaching the main topic of the presentation, she gave an overview of the most popular use cases of machine learning in optical communications, focusing on physical-layer applications. These use cases, retrieved from [12-14], are summarized in Table 1. They consider a machine learning model trained using a set of artificial (numerical) or experimental input and output data sets. Thus, Table 1 also highlights what data sets are relevant for each use case and how they are normally generated.

Table 1. Most popular use cases of machine learning in optical communications and their data sets [12-14].

Use case	Input data sets	Output data sets	Data set generation
Nonlinear mitigation	Received constellations Received symbols	Decoded symbols with impairment estimated or mitigated Nonlinearity mitigated constellation points Symbol decision boundaries	Random sequence of bits
Optical performance monitoring	Amplitude histograms Constellations Eye diagrams	OSNR[1] PMD[2] CD[3] Q-factor[4]	Vary noise, CD, PMD, or a combination, and random sequence of bits
Modulation format recognition	Stokes space parameters Received symbols Amplitude histograms	Modulation format	Vary the modulation format, and random sequence of bits

1. OSNR – optical signal-to-noise ratio, 2. PMD – polarization mode dispersion, 2. CD – chromatic dispersion, 4. Q-factor – quality factor.

From the examples in Table 1, she concluded that the relevant input/output data sets for optical communications depend on the specific use case. However, that mainly applies for the output data set. For most of the cases reported on the literature and related to the physical-layer, the input data can be represented by the received waveforms.

Then, she summarized in Fig. 3 some general relevant data in optical communications for machine learning from the M-LiPS point of view. The red box presents the adjustable parameters at transmitter and receiver (Tx/Rx) and on the optical channel. Most of them are degrees of freedom that can be wisely adjusted by an artificial-intelligence approach. The pink box shows the different ways to represent the signal. These different signal representations can be retrieved from the waveforms (digital signals before digital to analog converter on the transmitter or after analog to digital converter on the receiver). Table 1 shows that they are extensively used as input data sets for machine learning models. The orange box shows the figures of merit of the system, obtained from the received signals. They can also be estimated/optimized by adjusting the parameters on the red box. And the purple box shows the penalties that can be mitigated, compensated and/or estimated from the received signals. Most of these penalties are introduced by the optical channel.

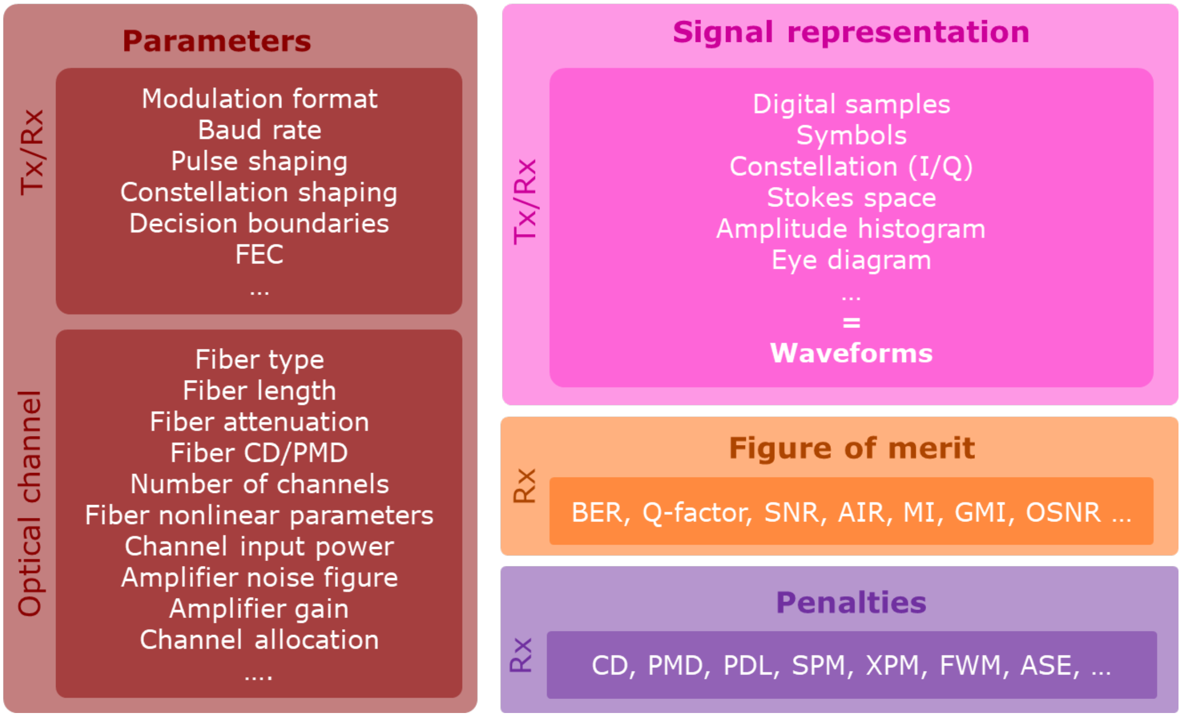

Figure 3. Not comprehensive summary of the relevant data that can be used in machine learning applications. Acronyms not defined yet: FEC – forward error correction, I/Q – In-phase and quadrature, BER – bit error rate, SNR – signal-to-noise ratio, AIR – achievable information rate, MI – mutual information, GMI – generalized MI, PDL – polarization dependent loss, SPM – self-phase modulation, XPM – cross-phase modulation, FWM – four wave mixing, ASE – amplified spontaneous emission.

As an overall conclusion of the talk, from the most popular use cases applying machine learning on the physical layer shown in Table 1, all data-sets used to train the machine learning model are associated to the received waveforms. The waveforms are directly impacted by the 'penalties' and 'parameter' changes/adjustments shown in Fig. 3 and from them it is possible to retrieve the 'figure of merits'. Thus, having access to standardized waveforms (and labeled for most use cases) can provide fair comparison between different machine learning approaches.

2.3. "Machine Learning Models for Optical Communication Systems & Networks"

Speaker: Joao Pedro (Infinera)

This talk introduced the scenarios where ML models are expected to find application. These include cases where (1) conventional approaches are not applicable or not efficient as a result of model-deficit (e.g. no physics-based mathematical model for the problem exists due to insufficient domain knowledge) and/or algorithm-deficit (e.g. mathematical model exists but algorithms to run it are too complex); (2) details of how the task is solved are not relevant (e.g. not necessary to explain how decisions are made); (3) phenomenon or function being learned is stationary for a sufficiently long period of time; and (4) sufficiently large labelled training data set exists or can be created. In addition, it emphasized that since ML models rely on data, the cost of obtaining data in quantity and quality determines to a great extent the motivation to adopt ML models.

Some of the key challenges faced by transport network operators were overviewed, such as the requirement to increase capacity while keeping capital and operational expenditures low. This demands not only reducing the cost per bit transported but also simplifying network operation, facilitating network expansion, fostering introduction of new features and reducing the time-to-market of new services. A vision in line with these objectives is that of realizing an autonomous, self-learning and self-driving network (Fig. 4), which would be characterized by the ability to (1) auto-configure (e.g. routing / spectrum / modulation format adapted based on real-time physical layer data); (2) self-heal (e.g. route cause analysis and failure prediction complemented with automatic repair and preventive maintenance); and (3) predict traffic and self-optimize (e.g. avoid congestion, suggest network augmentation). ML models are well positioned to be a key ingredient in materializing this vision. Moreover, several developments in optical networks are helping to realize this vision and, in the process, to start employing ML models: (1) flexible coherent line interfaces, enabling to collect a rich set of performance data at the receiver end of an optical channel; (2) SDN control, whose centralized nature facilitates using virtually limitless CPU and storage; (3) real-time performance planning avoiding traditional margin stacking and overprovisioning, which is data intensive and demands estimating performance degradation and trend analysis; and (4) disaggregated transport platforms and line systems, which on one hand require open standards and protocols to ensure interoperability, while on the other hand, raise challenges related to end-to-end performance estimation that may need to be addressed via learning from the devices, systems or optical channels.

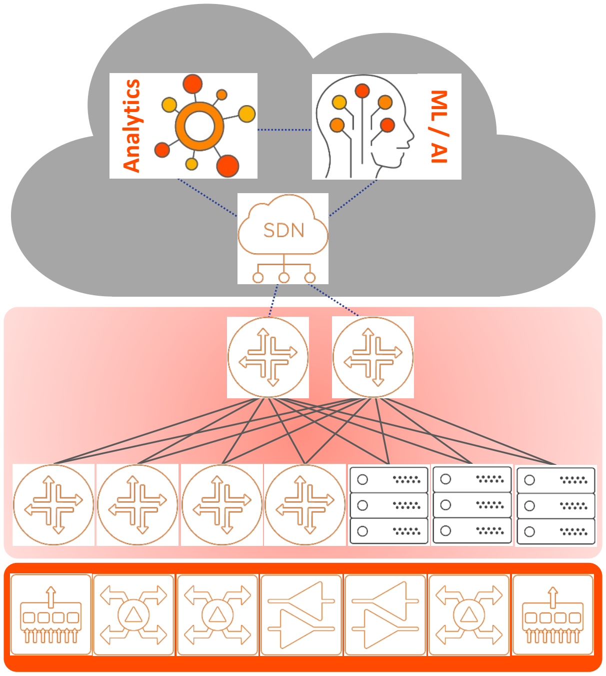

Figure 4. Setting the vision: the autonomous, self-learning and self-driving network.

Despite the enablers described above, it was highlighted that there are some specifics of optical networks that can delay, if not impede, the exploitation of ML in production scenarios. For instance, optical networks comprise a mature ecosystem and these networks have been successfully deployed and operated for decades. Particularly, (1) significant investments have been done in non-ML models and processes to operate them, which means the introduction of ML models will only occur if improvements in key metrics can be shown; (2) established practices and labor force are accustomed to deterministic approaches, i.e. a lengthy and possible costly shift in mindset is needed; and (3) existing network infrastructure may not be ready to exploit ML (e.g. data collection, storage and processing need to enhanced).

The possible issues with access to data were also discussed. Different communities, namely the research community, system vendors and network operators, will have different privileges to directly access every source of data, as summarized in Table 2. In view of the importance of data to successfully apply ML, this highlights the need to establish partnerships within and between different communities as playing a central role in the creation and exploitation of meaningful and rich data sets. Nonetheless, attaining this requires addressing a wide array of concerns, such as identifying clear incentives for sharing data and finding trustworthy ways to handle confidential data.

Table 2. User communities and the data sources they have access to.

User Community	Data Sources
"Open Access" (e.g. broad research community)	▪ Analytical and simulation-based models and databases that are public ▪ Academic lab setups (usually limited in size and with limited access to device and sub-systems internal details)
System Vendors	▪ Same as "Open Access", plus ▪ Detailed characterization of (their) devices and sub-systems properties (but with limited visibility of phenomena arising in a complex, heterogeneous, live network)
Network Operators	▪ Same as "Open Access", plus ▪ Specific deployment instances (with visibility over multiple planes, layers, domains, vendors, enforced policies, with impact of time-varying effects)

Finally, it was stressed that an important aspect to consider when deploying ML models in production is the need to guarantee ML model lifecycle management. Lifecycle management can include all or some of the following aspects: (1) fairness; (2) robustness; (3) assurance; and (4) explainability. Achieving fairness can require additional logic to help detect and remove bias from the ML models. Robustness focuses on granting the capability to detect and prevent data contamination and or tampering, whereas assurance intends to clarify questions such as why a dataset is created, who funded its creation, what preprocessing/cleaning was done, and will it be updated. While explainability remains a weak point of ML models, solutions for the other three aspects start to be readily available. For example, concerning assurance, recent efforts target the specification of datasheets for datasets [15].

The talk also included a discussion on possible use cases for ML models in optical networks. It was acknowledged that a wide array of use cases have started to be investigated. For instance, from modelling individual devices (e.g. EDFA) extending to capturing network-wide behaviors (e.g. optical channel QoT), as well as from modelling physical layer impairments (e.g. BER, OSNR) up to predicting traffic flows / patterns. Use cases also differ in their potential impact and complexity of use. The talked concluded by exemplifying two use cases where ML models can be applied to optical networks.

2.4. "Small Data Deep Learning: AI Applied to Domain Datasets"

Speaker: Michael Majurski (NIST, Information Technology Laboratory)

The recent growth in deep learning methods has been fueled by large, publicly available research datasets like ImageNet which consists of 1.2 million images labeled with 1 of 1000 categories. However, when operating within scientific domains, one cannot put forth those sorts of large-scale human annotation efforts. This talk focuses on how to take advantage of deep learning methods for computer vision when limited annotated data are available. We are motivated in domain applications to utilize deep learning methods, despite the dearth of annotations, because there are problems where it is easier to specify the desirable behavior than to explicitly write a program and because deep learning methods can deliver highly accurate results.

In machine learning, the practitioner has two main goals: make the error on the training data small and make the gap between the training error and the test error as small as possible. Underfitting happens when a model cannot reach an acceptable training error. Overfitting happens when a model has too large a gap between the training and test errors. Modern high capacity of deep learning models have the ability to memorize a training dataset which is too small. Right now, there are two common approaches to mitigating the problems of building deep learning models with small datasets: 1) data augmentation, and 2) transfer learning.

Data augmentation consists of expanding the dataset by performing label preserving transformations. Intuitively this adds invariances to the trained model. For example, a picture of a dog is still a dog if you flip the image left-right. Common data augmentation transformations include rotation, reflection, jitter, cropping, noise, photometric distortions, and zooming. Data augmentation provides a relatively time-efficient method for domain experts to specify what invariance should exist within the trained model. The state of the art in data augmentation is learning the set of transformations and the sequence in which they should be applied to a specific dataset as part of the model training process.

Transfer learning is when you build a model using a large available dataset and then refine the trained model on the small data domain application. You can perform transfer learning from large research datasets like ImageNet or from representation learning utilizing unannotated data via Generative Adversarial Networks (GANs).

The case study presented involves performing non-destructive quality assurance for Induced Pluripotent Stem Cell therapies for Age Related Macular Degeneration. Destructive testing was used to provide 1000 ground truth for an image-based measurement to predict stem cell implant quality based upon cell morphology. Morphology was extracted from single cell segmentation performed using an encoder-decoder convolutional neural network architecture called UNet. For this application transfer learning from a large research dataset Common Object is Context (COCO) outperformed both data augmentation and representation learning via GAN when quantifying single cell segmentation accuracy via the Adjusted Rand Index metric. Data augmentation using domain expert provided invariances was optimal when considering finding the edges between cells.

3. Breakout Sessions

The three breakout sessions broadly covered three main areas relevant to ML in optical communications systems. These areas were chosen as they address overarching architectures present in optical communication networks including: core optical network, transponders, cross layer networking. More specifically, breakout session topics were:

- Data from reconfigurable optical add drop multiplexer (ROADM) and optical layer
- Possible data sets from coherent transponders
- Cross Layer End-to-end networking

3.1. ROADM and Optical Layer

Optical line systems within the optical layer of core networks were considered as a focus area of breakout group discussion. Included in this scope are the ROADM switching nodes, optical amplifiers, and the network of fibers connecting them, but not the transceivers. Core optical networks were taken to be all metro and long-haul networks in which ROADM-based systems are used.

Three aspects of optical line systems were identified as potential application focus areas for AI and ML techniques: 1) quality of transmission (QoT) estimation, 2) fault identification, and 3) failure prediction. Of these QoT estimation was considered the most promising application. Fault identification was considered to be difficult in terms of collecting data because of the sensitivity of this data to the network operator. Failure prediction would also be difficult for similar reasons but could be based on data related to the overall health of the network and therefore data might be more accessible.

The three application areas can be further aligned to consider specific network use cases. The use cases identified by the group include system disaggregation, network defragmentation, and faster dynamic operation (i.e. faster add/drop and switching of optical signals). Disaggregated systems involve different components such as transceivers, amplifiers, and switches being supplied by different vendors, rather than being a part of a single proprietary or closed system designed, integrated, and performance guaranteed by a single vendor. There was consensus that disaggregation creates an environment in which there is a greater need for and benefit from machine learning approaches since the system has not been engineered end to end. The potential to use third party control and management systems through an SDN control plane also opens up the potential for innovation around machine learning based control. There might also be more opportunity to collect data from disaggregated systems since the systems are open. Defragmentation refers to re-organizing the wavelength routes in a system when there is a large amount of stranded or fragmented bandwidth due to a series of wavelength provisioning events that were not optimized for efficient use of the bandwidth (for example when these occur over a period of time and cannot be optimized altogether). Given the very complex nature of network defragmentation it was suggested that machine learning might provide some benefit. Faster dynamic operation, such as faster provisioning of channels, is also very complex and could benefit from machine learning. Dynamic operation could also make use of reinforcement learning as data is collected from each change applied to the system.

In addition to these applications, the group identified four types of data that are relevant to these applications: 1) component data, 2) path/network data, 3) operating data, and 4) fault management data. Component data refers to parameters related to the state and health of individual components such as the temperature of an amplifier pump laser. The path or network data include relevant line system and signal characteristics that influence the performance of the optical signals over the path from add to drop location. Operating data and fault management data go together in terms providing information about the system health and operating performance. Examples might be the ROADM alarm status or data regarding the frequency or root cause of failures.

Discussion around the QoT application considered different sources and formats of data as well as the complications in obtaining such data. The sources of data included live network data from operators, field testbeds, and laboratory testbeds. The first two sources are preferred although laboratory testbed data could be useful for correlating with field data and understanding how training and data from the lab might be applied to systems in the field. Operator data is clearly the most difficult to obtain due to privacy and proprietary considerations. Research and education networks such as ESnet are likely easier to obtain data from and in fact already provide much data online. In some cases, there might be network segments that are not live or that are used for research purposes and these could be used to collect data. The COSMOS testbed is an example of a research testbed in the field that can be used to collect data. A number of academic, national lab, and industry lab testbeds might be good sources of data as well. Concerns were raised that industry labs might provide biased data sets that favor their products.

For QoT estimation, the electrical signal to noise ratio and an associated bit error rate or Q-factor are often available from the transceivers or standard performance monitoring telemetry. This information can be used for supervised learning based on other data collected from the system. Other data might include a wide variety of parameters the most important being the system parameters (fiber type, distances, etc.), optical power levels, and optical signal to noise ratio. An important resource of developing machine learning algorithms for QoT estimation would be the analytical models and their performance as a reference for comparison. The GNpy model is a recent public and open source QoT estimation tool that can be used for this purpose.

Path data would be the most important type of data for QoT estimation. Key question includes the format of the data and what models are relevant. Wavelength dependent path data is particularly important and difficult to obtain. Filtering effects are also very important, but it might be difficult to obtain data about the filtering effects or parameters.

This breakout session also considered the potential for launching a machine learning challenge for optical networks. QoT estimation was again considered the most appropriate application area. Many groups are already working on developing QoT estimation algorithms and could easily compete in such a challenge. Several sources of data could potentially be used for this: 1) field trials such as SCinet at the SC conference, 2) field testbeds such as COSMOS, and 3) R&D labs. The breakout group agreed to investigate whether these data sources could be used to produce data for a challenge during the upcoming year. Data could potentially be collected during the turn up of the systems for SCinet and then a challenge run over the winter. Winners could be announced at OFC. The goal of algorithm development in the challenge would be to predict the received bit error rate, Q-factor, or generalized optical signal to noise ratio (OSNR).

3.2. Possible Data Sets from Coherent Transponders

Coherent modems by their very nature generate a number of different optical parameters which historically have been difficult to measure in real time. This breakout session group reviewed the types of data that is now available or soon to be available from coherent modems in modern optical systems. We considered both directly measured parameters as well as computed ones derived from the observables. A list of the kinds of observables that are achievable from coherent modems follows, although the specific list will vary by vendor.

Parameter	Units	Description
Total Pwr	mdBm	Displays the total power (mdBm).
Chan Pwr	mdBm	Displays the channel power (mdBm).
FE Tx CD	ps/nm	Displays Far-End Tx Dispersion Pre-Compensation (ps/nm)
Rx CD Comp	ps/nm	Displays the receive chromatic dispersion compensation (ps/nm).
BER	decimal	BER Displays the Bit Error Rate (decimal).
Fast BER	decimal	Displays the fast BER (decimal).
FER	decimal	Displays the Frame Error Rate (decimal).
Uncorrected Blocks	blocks	Displays the number of Uncorrected Blocks.
Cycle Slip	slips	Displays the Cycle Slip count.
PM Tick Count	count	Displays the performance monitoring tick count.
PMD	ps	Displays the (instantaneous) Polarization Mode Dispersion (ps).
PDL	dB, peak	Displays the Polarization Dependent Loss (dB).
SOPs1	decimal	Displays the State of Polarization s1 (decimal).
SOPs2	decimal	Displays the State of Polarization s2 (decimal).
SOPs3	decimal	Displays the State of Polarization s3 (decimal).
FreqOffset	MHz	Displays the frequency offset (MHz).
MERx	dB	Displays the Modulation Error Ratio for value X (dB).
MERy	dB	Displays the Modulation Error Ratio for value Y (dB).
SPM	dB	Displays the Self Phase Modulation (dB). SPM is a nonlinear optical effect of
TNLE	dB	Displays the Total Non-Linear Noise (dB).
ESNR	dB	Displays the Electrical Signal to Noise Ratio (dB).
OSNR	dB/0.1nm	Displays the Optical Signal to Noise Ratio (dB).

In this breakout session group the kinds of applications which could benefit from applying AI/ML techniques to datasets containing observables such as these were discussed. Candidate applications could include the following:

- A predictive assessment of the health of the network equipment itself.
- Analytics on the health and future viability of the fiber.

- Estimation of network performance at different operating points and different physical locations.
- Remote sensing: fiber type determination, fiber stress, intrusion detection, etc.

There is also a potential tie-in with the ROADM and optical layers which can produce power level, spectral analysis, and OTDR data at each amp site that could then be used to augment the end-end observables that the modems produce.

Key questions outstanding are the identification of additional problem sets that could use this data, and how to select the needed data and export it from the system in a way that is standardized and thus could be used cross platform.

3.3. Machine Learning Data Sets for Cross-Layer and End-to-End Networks

The emergence of cloud computing has had a profound impact, enabling for example, scalable compute and storage for enterprises, and cloud-based personal digital assistants based on artificial intelligence (AI) in homes. While considerable investment has gone into creating on-demand cloud-compute resources, the flexibility of the underlying network, which is crucial to universal access to the cloud, has been largely under realized. This is despite steady improvements in network hardware to support such flexibility at the optical layer with the introduction of contentionless, directionless, flexgrid reconfigurable add-drop multiplexers (CDCF ROADMs) and wavelength, modulation format, and symbol-rate flexible optical transponders. Two recent technological advances have the potential to create a truly dynamic network infrastructure – the development of software-defined networks (SDN) and the application of AI and machine learning (ML) to the transport network. By combining the control and virtualization capabilities of SDN with network intelligence gained through streaming telemetry and ML, we can achieve a flexible and intelligent network infrastructure.

One of the missing ingredients in the above formulation is the lack of available datasets that will allow for the development of AI and ML algorithms for intelligent network control. The development of reliable algorithms requires training of models, for example artificial neural networks, with large data sets. Of particular interest is network performance data across domains (e.g., access, cloud, metro, core) and layers (e.g., IP and optical) of the network. Unfortunately, data sets such as this are mostly not publicly available if at all, which is likely due to 2 main reasons: 1) network performance data is often considered proprietary by commercial network service providers. 2) Technologies and tools for easily collecting performance data from different network domains and layers have recently been developed and the data may not yet be widely collected and stored. Some examples from a survey of available datasets include optical backbone performance data from Microsoft [16], IP traffic traces from Internet exchange points [17], and optical network topology databases [18,19]. There are several datasets from large cloud data center operators that give cluster server usage data [20] and traffic characteristics inside data centers [21].

At the Machine Learning for Optical Communication Systems Workshop, a breakout group addressed the topic of whose purpose is to make more data sets available for cross-layer and end-to-end (CLE2E) applications. The breakout group chose to focus on 2 main use cases: IP/optical cross-layer interaction, and end-to-end services for emerging 5G wireless networks.

3.3.1. IP/optical cross-layer interaction

Historically, IP and optical networks have been managed separately, even by different teams of engineers at the larger service providers. While IP traffic is carried by the underlying optical transport network, there is little visibility or coordination between them. However, two trends are working towards breaking this separation: increased dynamism in connectivity driven by the cloud era and 5G mobility; and the new flexibility in modulation format and symbol rate at the optical layer allowing adaptation to the network conditions to maximize capacity. Managing the IP and optical services jointly under these conditions is needed to achieve the desired efficiency and performance.

Bell Labs, has explored cross-layer aware applications by introducing error awareness in a multi-layer network operating system [22]. Services at the packet layer can be automatically rerouted based on postFEC errors measured at the optical layer depending on the error tolerance of the packet service. In that work, optical errors were introduced by degrading the optical signal to noise ratio or by other intentional means. The development of multi-layer applications such as this would greatly benefit from field data collected from operational networks. For example, error data from the packet and optical layers that could be time correlated would be invaluable. Also, traffic traces at the granularity of packet flow-level would allow for the impact of optical errors on packet flows to be studied. This IP/optical interaction is arguably one of the most basic performance metrics of a wide-area transport network, and yet it is a metric on which the research community has no readily available data.

Exploring possible sources of data that would allow the application of machine learning algorithms to tasks such as error correlation, the probability and distribution of errors, and the prediction of the impact of errors on the network performance was identified as a necessary step by the CLE2E breakout group. In this breakout group possible sources of data were also discussed. One promising source could be to work with operators of government-funded networks such as ESnet [23] and Internet2 [24], who may be willing to make such datasets available.

3.3.2. End-to-End 5G Mobile Networks

The migration to 5G mobile networks will have a major impact on networks, requiring increased capacity, ultra-reliability and low latency, and massively scalable connectivity for internet of things (IoT) devices. The impact of 5G mobility will reverberate across the network in ways that go beyond driving higher capacity. In particular, there will be a need to dynamically create virtual network slices that cross network domains and layers. Slicing creates partitions of the physical network resources to meet the service requirements of each application such as throughput, latency, and availability. Devices with demanding network service requirements, for example, cloud-controlled industrial robots, can be onboarded to slices that meet their needs. A high degree of network automation will be necessary to create and use end-to-end slices that cross network domains and layers. ML will be crucial to providing the intelligence for this automation.

Some problems that can be addressed by ML include network performance prediction for slice creation, the optimization of the physical layer split processing available in eCPRI [25], the onboarding of IoT devices to network slices, and the placement of virtual network functions at distributed cloud data centers for the 5G mobile core [26]. Regarding data set availability for ML, we find ourselves in a particularly open space as we are at the cusp of 5G mobile deployments that will be ramping up over the next few years. However, there are research-oriented field deployments of 5G technologies that we may look to for data collection, specifically, the COSMOS project, which is a 5G field research network in upper Manhattan [27], and the Berlin 5G testbed [28].

3.3.3. Other CLE2E Breakout Group Topics

Other topics that are worthy of being addressed and related to CLE2E are:

- Further investigation of testbeds or sources of data that would aid in ML applications for the above two use cases. Also, identification of what types of data would be most useful. The possibility of getting data from the Utah Internet2 point of presence was also discussed.
- The format of data sets will need to be discussed.
- Suitable ML formulations, e.g., prediction, correlation, and classification should be discussed further to help define the requirements for the data sets.
- Engaging other 5G testbeds, e.g., ADRENALINE in Spain [29].
- Other open challenges to data set availability should be identified, e.g., legal, business, practical.

4. Themes

Themes relevant to data and machine learning arose as a result of the workshop presentations and breakout sessions. In general, these themes address data in the context of use cases, reliability and access. These themes are summarized in the following bullets:

- More access to representative data sets as opposed to relying on simulated data or over-controlled laboratory data
    - Data that is obtained from real world test beds.
    - Data that is obtained in-the-field.

- More available training data sets to address specific use cases such as:
    - 5G
    - IoT
    - Cross later-end-to-end use cases
    - Physical layer/optical layer use cases
    - Hardware and components

- Considerations on proper data normalization and guaranteeing reliability of data sets.
    - Can metrology play a role for normalization?
    - How does one guarantee data set reliability?

5. References

[1] F. Musumeci et al., "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, Second Quarter 2019.

[2] F. Musumeci, C. Rottondi, G. Corani, S. Shahkarami, F. Cugini and M. Tornatore, "A Tutorial on Machine Learning for Failure Management in Optical Networks," in Journal of Lightwave Technology, vol. 37, no. 16, pp. 4125-4139, 15 Aug.15, 2019.

[3] Brajato, Giovanni, et al. "Optical Frequency Comb Noise Characterization Using Machine Learning." 2019 European Conference on Optical Communication (ECOC). IEEE, 2019.

[4] Chin, Hou-Man, et al. "Phase Compensation for Continuous Variable Quantum Key Distribution." CLEO: Applications and Technology. Optical Society of America, 2019.

[5] D. Zibar et. al., Machine Learning-Based Raman Amplifier Design, OFC *2019*, San Diego, CA, USA.

[6] D. Zibar et. al., Inverse System Design using Machine Learning: the Raman Amplifier Case, under revisions in Journal of Lightwave Technology.

[7] Brusin, Ann Margareth Rosa, et al. "An ultra-fast method for gain and noise prediction of Raman amplifiers." 2019 European Conference on Optical Communication (ECOC). IEEE, 2019.

[8] Jones, Rasmus T., et al. "Deep learning of geometric constellation shaping including fiber nonlinearities." 2018 European Conference on Optical Communication (ECOC). IEEE, 2018.

[9] Jones, Rasmus T., et al. "Geometric constellation shaping for fiber optic communication systems via end-to-end learning." arXiv preprint arXiv:1810.00774 (2018).

[10] Jones, Rasmus T., et al. " End-to-end Learning for GMI Optimized Geometric Constellation Shape." 2019 European Conference on Optical Communication (ECOC). IEEE, 2019.

[11] D. Zibar et al., Machine learning under the spotlight, Nature Photonics, (11) 749-751, 2017.

[12] J. Mata, I. de Miguel, R. J. Durán, N. Merayo, S. K. Singh, A. Jukan, M. Chamania, Artificial intelligence (AI) methods in optical networks: A comprehensive survey, Optical Switching and Networking, 2018.

[13] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, M. Tornatore, "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, 2019.

[14] F. N. Khan, Q. Fan, C. Lu and A. P. T. Lau, "An Optical Communication's Perspective on Machine Learning and Its Applications," in JLT, vol. 37, no. 2, pp. 493-516, 2019.

[15] https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf

[16] https://www.microsoft.com/en-us/research/project/microsofts-wide-area-optical-backbone/

[17] https://portal.linx.net/stats/lans

[18] http://www.topology-zoo.org/

[19] “Internet Atlas: A Geographic Database of the Internet,” Ramakrishnan Durairajan et al., HotPlanet '13 Proceedings of the 5th ACM workshop on HotPlanet, Pages 15-20 2013.

[20] https://github.com/google/cluster-data

[21] http://pages.cs.wisc.edu/~tbenson/IMC10_Data.html

[22] “Error Awareness in a Multi-Layer Transport Network Operating System,” Jesse E. Simsarian, Young-Jin Kim, Nakjung Choi, Catello Di Martino, Nisho k N. Mohanasamy, Peter J. Winzer and Marina Thottan, J. OPT. COMMUN. NETW., vol. 10, no. 3, pp. 152 – 161, March 2018.

[23] http://es.net/

[24] https://www.internet2.edu/

[25] “eCPRI Overview,” Tero Mustala and Olivier Klein, Nokia.

[26] ”View on 5G Architecture,” 5GPPP Architecture Working Group, Version 2.0, December 2017.

[27] “COSMOS: Optical Architecture and Prototyping,” Jiakai Yu, Tingjun Chen, Craig Gutterman, Shengxiang Zhu, Gil Zussman, Ivan Seskar, and Daniel Kilper, OFC 2019, M3G.3.

[28] “Photonics-Supported 5G Test Facilities for Low Latency Applications,” Behnam Shariati, Kai Habel, Volker Jungnickel, Johannes Fischer, and Ronald Freund, ICTON 2019, Fr.D3.3.

[29] “The ADRENALINE testbed: An SDN/NFV packet/optical transport network and edge/core cloud platform for end-to-end 5G and IoT services,” Muñoz, Raul, et al., in Proc. IEEE EuCNC’17, 2017.

6. Acronym List

AI: Artificial Intelligence
BER: Bit Error Rate
CPU: Central Processing Unit
EDFA: Erbium Doped Fiber Amplifier
ESNR: Electrical Signal to Noise
ML: Machine Learning
NFV: Network Functions Virtualization
OLS: Open Line Systems
OPEX: Operating Expenditure
OSNR: Optical Signal to Noise
QoT: Quality of Transmission
ROADM: Reconfigurable Optical Add Drop Multiplexer
Rx: Receiver
SDN: Software Defined Network
Tx: Transmitter

7. Appendix: NIST Resources

The NIST mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

Communications Technology Laboratory

The National Institute of Standards and Technology's Communications Technology Laboratory (NIST CTL) was established in 2014 to unite NIST's many wireless communications efforts into a unified research and development organization. Through metrology and research in physical phenomena, materials capabilities and complex high-speed communications systems, we are establishing the technological basis upon which the ongoing wireless revolution depends. https://www.nist.gov/ctl

Information Technology Laboratory

The Information Technology Laboratory (ITL), one of six research laboratories within the National Institute of Standards and Technology (NIST), is a globally recognized and trusted source of high-quality, independent, and unbiased research and data. As a world-class measurement and testing laboratory encompassing a wide range of areas of computer science, mathematics, statistics, and systems engineering, ITL's research program supports NIST's mission to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and related technology. https://www.nist.gov/itl