A Necessary Journey Towards an AI-driven Operation

Telecom Argentina perspective

A Technical Paper prepared for SCTE by

Claudio Righetti
Chief Scientist
Telecom Argentina S.A.
crighetti@teco.com.ar

Mariela Fiorenzo
Tech Expert Scientist
Telecom Argentina S.A.
mafiorenzo@teco.com.ar

Horacio Arrigo
Tech Scientist
Telecom Argentina S.A.
hgarrigo@teco.com.ar

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Abstract

More than a decade ago we began introducing analytics, machine learning, and finally artificial intelligence to our networks and services. Evolving our work teams from a data-driven culture to AI-driven. It was not an easy task, it involved great challenges and cultural changes, it really is an accelerated transformation process during this pandemic, and it continues.

In this technical paper, we go through the path we are transitioning in Telecom Argentina from data analysis and AI/ML perspective to achieve operational excellence. We present the challenges we went through along, difficulties, learned lessons, success stories and next steps.

Content

1. Introduction

In our networks and services, the Artificial Intelligence (AI) has the potential to change, the way we operate, and to become the foundation of the transformation that leads to the fourth industrial revolution. But this requires hard work, a long-term commitment, and a deep cultural change. That is why we present here our journey that we started to make our operations AI-driven.

In the industry, Analytics, AI, and Automation are often differentiated. Let us remember that in [5], we define:

- Data Analytics: monitoring data to look for patterns and anomalies (without applying intelligence) and applying those patterns towards effective decision making.

- Artificial Intelligence: the development of computer systems capable of performing tasks that normally require human intelligence; this includes visual perception, speech recognition, decision-making, and translation between languages.

In a survey regarding enterprise networks, automation was enquired. The poll results indicates that more than 65 percent of enterprise networking tasks are carried out manually (often referred to as "ClickOps"), indicating that own network automation underlies on servers' automation. Ansible, customized "DIY" scripts (usually based on Python), and single-vendor, network infrastructure-focused packages are among the most widely used network automation tools. It is worth noting that these scripts are totally deterministic, that is, they only perform repetitive tasks. AIOps is the term used when a decision-making process is automated using an AI algorithm.

In next sections, we introduce in more detail what we understand by AIOps, we may state:

AIOps = Analytics + AI + Automation

AIOps is part of the 5G ecosystem, since from its conception the knowledge plane has been included. That is, a layer within the architecture oriented to the operation and orchestration of networks and services [7]. Figure 1 shows the percentage of network activities that are automated according to Gartner.

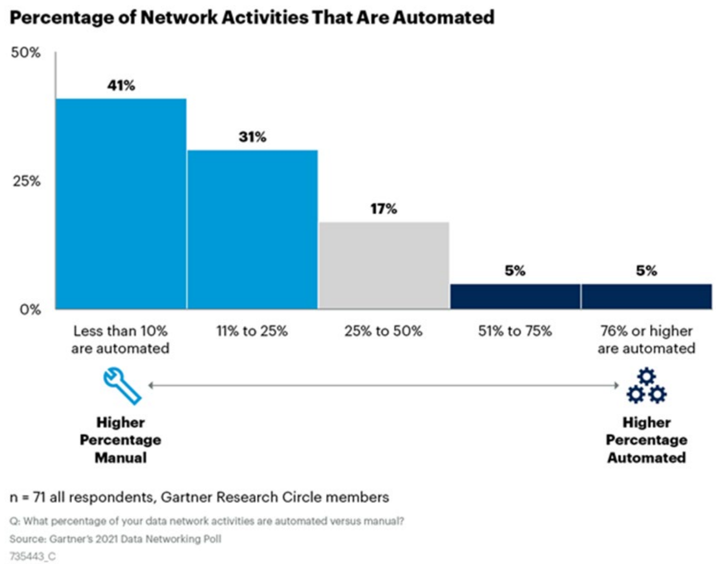

Figure 1 – Percentages of Network Activities that are automated, 2021.

Source: gartner.com

In this technical paper, we outline the path we've taken toward converting our AI-driven networks and services, along with a few use cases. The use cases development provided us with many lessons learned.

This document is organized as follows. After this section, we introduce Telecom Argentina and Financial impact of Analytics, AI, and Automation. At section 2, we expose AI-driven Operation. In section 3 we present the concept of the knowledge plane and our reference architecture. In section 4, we introduce the first use cases we did using AI and Big Data. On section 5 we present the AIOps use cases in Telecom Argentina and the state of the art in the evolution towards autonomous networks. In section 6 we outline the key lessons learned along this journey and lastly, at section 7 we describe the next steps we are considering.

1.1. About Telecom

Telecom Argentina is a company in constant evolution, which offers connectivity and entertainment experiences and technological solutions throughout the country. We boost the digital life of our over 30 million customers, with a flexible and dynamic service, in all their devices, through highspeed mobile and fixed access, and a live and on demand content platform that combines series, movies, gaming, music, and TV programs.

Our trademarks Telecom, Personal and Flow consolidate an ecosystem of platforms, and new businesses, a comprehensive and convergent experience for individuals, companies, and institutions across the country. We are present in Paraguay with mobile services, and in Uruguay, with cable television services.

Telecom Argentina has become a company that thrives in the digital world. It evolved from a traditional telecommunications company to consolidate itself as an ecosystem of apps and platforms that are based on connectivity as a differential quality value.

With the vision of going beyond connectivity, we are developing new 100% digital businesses, based on IoT, 5G, Fintech, entertainment, and Smart Home solutions, among others. With the most innovative technology, in alliance with world class technology partners, and an investment which over the last five years reached USD 5 Bn, the company focuses on enhancing its infrastructure and systems and providing more and better services.

Our fixed-mobile network is the most extensive of the country. With more than 75,000 kilometers of FTTH, HFC and ADSL technologies, we are present in 18 provinces with over 60% of coverage of households in the country. Our 4G+ mobile network is the fastest in the country, and it is available in 100% of our infrastructure. We reach more than 1,900 locations and have a coverage of 95% of the population. In 2021 we inaugurated the first 5G network in Argentina, with 20 sites in the City of Buenos Aires, Rosario, and Costa Atlántica. We also bring connectivity to numerous towns with less than 500 inhabitants in different provinces, and in many cases, we are the only link they have with the rest of the country and the world.

The convergent and comprehensive operation of the network is one of the key challenges of the evolution of Telecom Argentina. Automation, analytics, and artificial intelligence, among other innovative technologies, will undoubtedly mark the path of our completely transformation from a Communication Service Providers (CSPs) to a Digital Service Providers (DSPs), Figure 2.

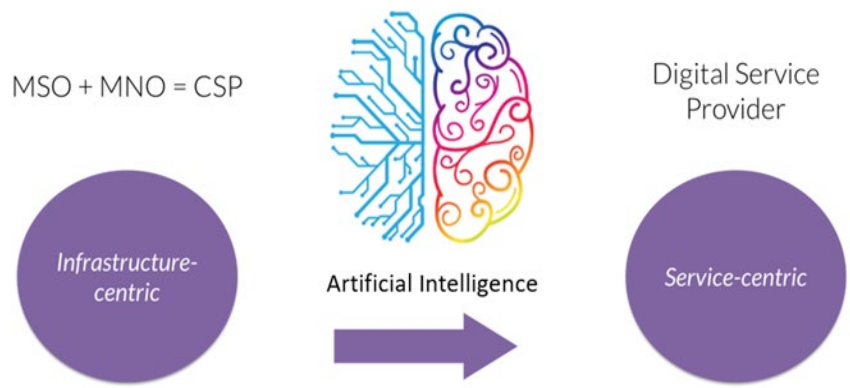

Figure 2 – From CSP to DSP.

1.2. AI-driven

Adoption of cloud computing, network function virtualization (NFV), and the development of software-defined networking (SDN) have all advanced considerably in recent years.

These developments have made it possible to create infrastructure that is more adaptable and to make storage and computing power more plentiful than before. We are compelled to use both artificial intelligence and machine learning techniques because of this progress and the growing requirement to enhance the management and administration of networks and services.

On the other hand, the next generation of 6G communications is already being investigated, where the devices, antennas and infrastructures are embedded with artificial intelligence software. Today networks cannot survive without artificial intelligence [1].

We are confident that these technologies will assist us in resolving and enhancing the current challenges with network efficiency so that our more than 30 million consumers have a better digital experience. Additionally, it will free up our specialized labor to work on more difficult tasks connected to emerging digital services and businesses.

These technologies will help us to:

- Reduce time to market for new products and services
- Predict and mitigate network and equipment service issues before they happen

Operating, managing, and provisioning future services with automation processes becomes essential to increase efficiency.

- Future Challenges: We are starting down the path towards the automation of our network operations. Using AI to efficiently manage the exponential traffic growth and the complexity and variety of new services that 5G will enable.

2. What it means to be AI-driven operation

When working on AI-driven initiatives, it is crucial to establish the minimum level of understanding inside our organization regarding the scope and restrictions of AI technology as it relates to the functioning of our networks and services.

There is no single definition for all AI technologies or framework. When we refer to technologies such as DOCSIS or 5G, there are no ambiguities since they are very mature in their standardization process.

AI-driven operations refer to the use of AI for the operation, planning and decision making in networks and services. To convert the operation of networks and services into an AI-driven operation, it is necessary to go through three stages:

Stage 1: Use of artificial intelligence techniques to develop applications and use cases.

Stage 2: Using AI to improve services, processes, or products.

Stage 3: Use of AI to help decision making.

In Figure 3 we present the Hype Cycle for artificial intelligence according to Gartner.

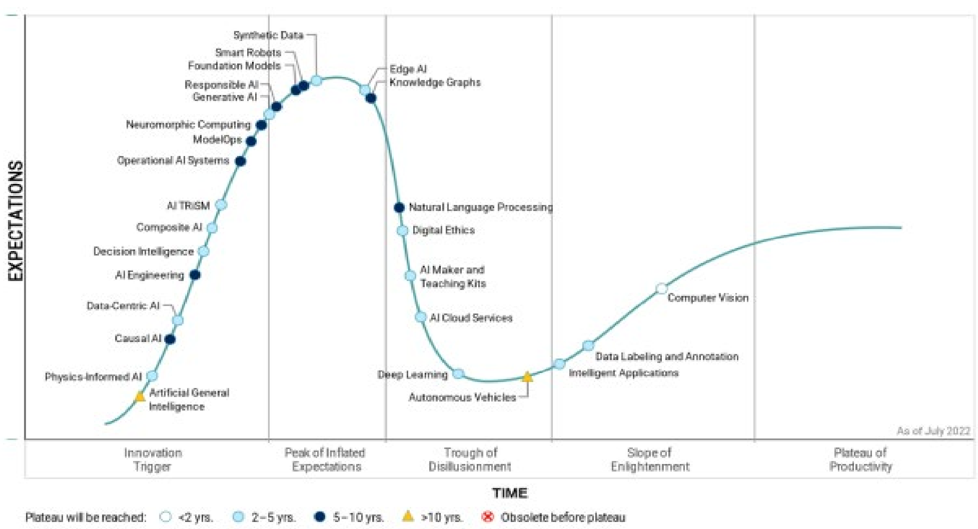

Figure 3 – Hype Cycle for Artificial Intelligence, 2022.

Source: gartner.com (July 2022)

3. Our first steps without AI using math, statistics and small data

In this section we present some data-driven decision-making use cases we performed as our first steps in this AI-driven operation journey.

3.1. Dimensioning of our VoD system

By the end of year 2012, Telecom Argentina launched its Video on Demand service. Due to certain recommendations from other operators, the original coverage was 900 HD set-top boxes (STBs) per service area (SA). However, when the service was launched it was not possible to achieve this (because of CAPEX restrictions), and many service areas were oversized up to 2,500 STBs. Hence, our customers experienced VoD service outages.

For the above, it was necessary to propose a statistical model that enables the resizing the SAs, establishing a balance between the system capacity, the number of customers and the *blocking probability* [2].

After studying and analyzing various models, we proposed a model based on the Queuing Theory applied to VoD traffic, the Erlang B formula, already used in telephony traffic [2]. In particular, the Erlang B formula allows us to relate three fundamental variables: the offered traffic, the number of available streams and the blocking probability, resulting in the appropriate number of STBs that the SAs must contain given a certain blocking probability that is considered acceptable.

Traffic definition is based on empirical assumptions, including peak service period, penetration per SA, average content duration, and average number of first-attempt requests. These last two depending on the quality of the content (SD or HD).

The data analysis was not only critical for the design of the model but, also contributed to establish an acceptable blocking probability for the service and to determine the phenomena that influence the performance of the service and their affectation degree.

3.1.1. Erlang B statistical model

The objective is to calculate the number of STBs per SA to obtain an acceptable blocking probability for our system.

For its formulation, certain empirical assumptions were considered after the characterization of the service. Such as, peak period, peak period duration, average duration per quality (SD/HD), average number of requests and penetration rate. We didn't consider blocked orders to be retried.

The traffic calculation is then formulated as follows:

$$A = \frac{h \cdot (\lambda_{SD} \cdot t_{SD} + \lambda_{HD} \cdot t_{HD}) \cdot p}{T}$$

Being,

- h: n° of STB per SA.
- λ_q: average number of first attempt requests, per STB and per period (q = SD or HD).
- t_q: average time spent on the system per period (min) (q = SD o HD).

- *p*: VoD service penetration per period.
- *T*: peak period duration (min).

Since we analyze the number of STBs needed for each SA, the goal is precisely h. It is obtained, using the Erlang B formula, based on the blocking probability (P_B) and the available streams during the peak period (N). They are dynamic depending on the bitrate of the content, which varies between 1.875 and 15 Mbps (depending on SD/HD quality and encoding).

The application of the formula is possible since the system satisfies the hypotheses of the model.

$$P_B = \frac{\frac{A^N}{N!}}{\sum_{i=0}^{N} \frac{A^i}{i!}}$$

Figure 4 shows graphically this formulation evaluate on different scenarios.

We sat a blocking probability (P_B) value of 3%, based on our observations and conditions of acceptance for quality of service. Then, from the model, for a MPEG2 content with a ratio of SD/HD of 70/30, our recommendation was 700 STBs per SA.

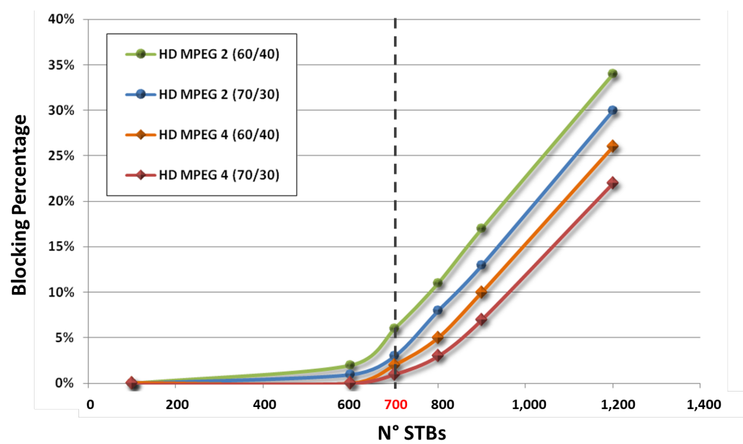

Figure 4 – Relationship between n° STB and Blocking (%) for different scenarios according streaming codification and quality.

The variables that most affect the model were also studied, these are:

- Very high rating content (peak time)
- Free content
- SD/HD service ratio
- HD Encoding (MPEG 2, MPEG 4)
- Holidays

Given the success of the model, we extended its use to assess the impact, in terms of probability of blocking, of different modifications to the service, such as: conversion to HD of CatchUp content, the addition of new services related to the 2014 World Cup, among others.

3.2. Characterization and impact of user behavior of OTT services

One of our major tasks was the study of applications' traffic. OTT video services represents more than 50% of our total downstream traffic. In terms of downstream traffic and client consumption, Netflix was the service with the highest utilization, especially during peak hours.

We found evidence that the link between subscriber count and traffic volume fits sub-exponential distributions. These long-tailed distributions can conveniently describe high variability. Historically, some long tail distributions have an origin in the income distribution, for example, Pareto and Log-Normal. The latter is one of the least understood and most widely used functions.

In Argentina, broadband users access streaming video content, mostly located in the USA. Thus, the impact of Round-Trip Time (RTT) in service performance is very important. A typical RRT in our service groups is around 120-150 msec. Therefore, CDN usage turned crucial.

Two server farms were set up in different company data centers to create a local cache of Netflix content. Each Open Connect Appliance is made up of three 12 Gbps-capable servers and were provided by Netflix. We had to connect to Netflix in Brazil via international peering since we were getting roughly 100 Gbps of Netflix traffic at its busiest each day. This allowed us to meet the demand for traffic that our CDN was unable to provide.

Since throughput is inversely related to RTT, which is smaller at shorter distances, an increase in throughput was obtained by connecting to Brazil instead of the USA. This was made possible by Netflix's adoption of the TCP-based DASH protocol.

3.2.1. Measuring performance for decision making

At the beginning of the Netflix CDN implementation, since it was only applied to the half of the service groups, we performed several tests where the traffic from Netflix users was compared pre- and post-implementation for each partition.

We observed the performance increase was noticeable for service groups using CDN, even more during the morning when workload was low. We have found the traffic generated by Netflix CDN users had doubled the traffic generated by those who were not using CDN.

Based on the above results we could estimate the total bandwidth traffic growth in the service groups. Assuming an increase of about 25% in traffic, since the implementation of the CDN, and maintaining the same number of active flows, the associated downstream Netflix traffic should have the same grow proportion. Let:

- N: *"Previous Netflix Traffic"*
- T: *"Previous Total Traffic"*
- N': *"Post Netflix Traffic"*
- T': *"Post Total Traffic"*
- Δ: *"Traffic Growth"*

So,

$$N' = N + \Delta N \ , \ T' = T + \Delta T \quad \text{and} \quad \Delta T = \Delta N \quad \rightarrow \quad T' = T + \Delta N$$

$$\frac{T'}{T} = 1 + \frac{\Delta N}{T} * \frac{T'}{T} = 1 + \frac{\Delta N}{N} * \frac{N}{T}$$

Thus, the percentage increase in total traffic is the product between the percentage increase in Netflix traffic and the percentage that Netflix represents of total traffic. Assuming a service group not using the CDN the relation N/T is 35% (calculated previously) and the percentage increase is 25%, we got:

$$\frac{T'}{T} = 1 + 0.23 * 0.35 = \mathbf{1.0805}$$

So, the total traffic growth of the service group will be 8.05%, supposing the number of concurrent subscribers remains invariant.

3.2.2. Forecasting subscribers and traffic

Having understood the Netflix's traffic importance on the capacity planning, we developed a time series model to forecast Netflix subscriptions, the amount of generated traffic and therefore, the impact on the access network.

We proposed a model based on time series, which provides powerful statistics. Commonly used in business and economics where data occurs in the form of successive values of a variable, in an ordered sequence in an equally spaced time interval. A stochastic model for a time series will generally reflect the fact that nearest observations in time, more closely relation than observation further apart.

Time series models allows us to:

- Understand underlying characteristics and structure that produced the observed data, through different analysis methods.

- Fit a model and forecast future values based on previously observed values for monitoring or even feedback and feedforward control.

Through statistical software, we observed that ARIMA was the best-fitting model. For Netflix users forecasting, we selected a confidence interval (CI) of 95% that, calculated from a given set of sample data, returns an estimated range of values which is likely to include an unknown population parameter. The results are shown in Figure 5.

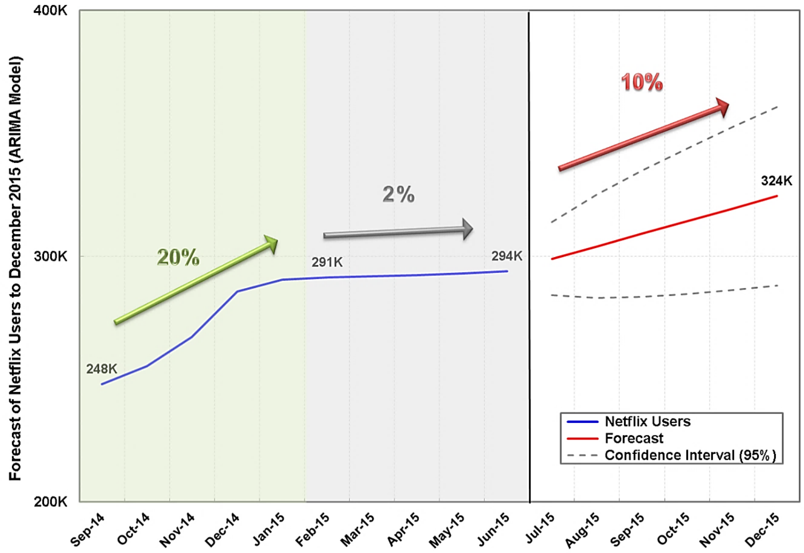

Figure 5 - Monthly evolution and forecast of Netflix users among our subscribers.

Remarkably this model has been successfully applied to forecast the whole broadband subscribers' population and particularly for Netflix subscribers, by fitting properly ARIMA model parameters.

In following sections, we describe how these first forecast models evolved into a tool with much more features and characteristics.

3.3. IP traffic dimensioning based on patterns

In this work we presented a forecasting model for the Average Bandwidth per Subscriber parameter, as a novelty we have made clusters by Service Tier Levels and by subscriber's location.

Exploiting our DPI tool, we analyzed trends and Internet users' preferences. Through mining our data at different periods, we found common daily patterns among our subscribers. This allows us to focus on a single day to have a whole understanding of the network. Considering the different Service Tier Levels, we also found common patterns related to the most used applications. Only Video Streaming services showed a variation among the lowest tiers. Thus, increasing download access speed drives consumption much more in video applications.

The resulting statistical parameters of our subscriber's characterization are inputs for a network dimensioning tool we have developed to analyze traffic impact over the service group's QAM carriers and simulate different scenarios.

For the last years, we have been collecting and analyzing various statistical measurements including the average residential bandwidth (BW) traffic considering the access speed and we found there is a mathematical relationship between them.

Some technology vendors have different proposals to estimate bandwidth capacity. Cloonan et.al proposed in [3], a technic based on the average amount of bandwidth per subscriber during the busy hour (T_{Avg}) calculated as follows:

$$T_{Avg} = 8 \cdot \frac{B}{W * N_{sub}}$$

Where N_{sub} is the number of subscribers within a typical service group, B is the number of bytes passed into the service group within a given window of time (W) measured in seconds.

We introduce a different methodology, we collect 5 minutes of bandwidth traffic data from the whole network (global, per service and per tier) and the four service tier levels, using a DPI technology. Monthly, we take the maximum bandwidth traffic per service tier level and divide it by the total number of subscribers of each. Then, we model this data with Linear Regression algorithms.

We found at least three different applications of this methodology:

- Adjust current BW traffic values per Tier
- Estimate BW traffic values for future offered access speed
- Forecast BW traffic values for the following two years

We define

- Avg_BW_subs: Average bandwidth per subscriber per Tier.
- $Speed$: Max access speed for the Tier.
- j: Tier number.
- t: Number of months used to forecast.
- β, γ: Parameters to estimate.

To adjust and to estimate, our model takes the form:

$$Avg\_BW\_subs_j = \beta \cdot \sqrt{Speed_j} + \varepsilon, \quad j \in \{1, \cdots, 5\}$$

To forecast:

$$Avg\_BW\_subs_j = \gamma_j \cdot t + \varepsilon, \quad j \in \{1, \cdots, 5\}$$

Figure 6 shows the Linear Regression of the average bandwidth per subscriber. Blue line represents the adjusted and estimated current values. Green and orange lines, the forecast for the next two years.

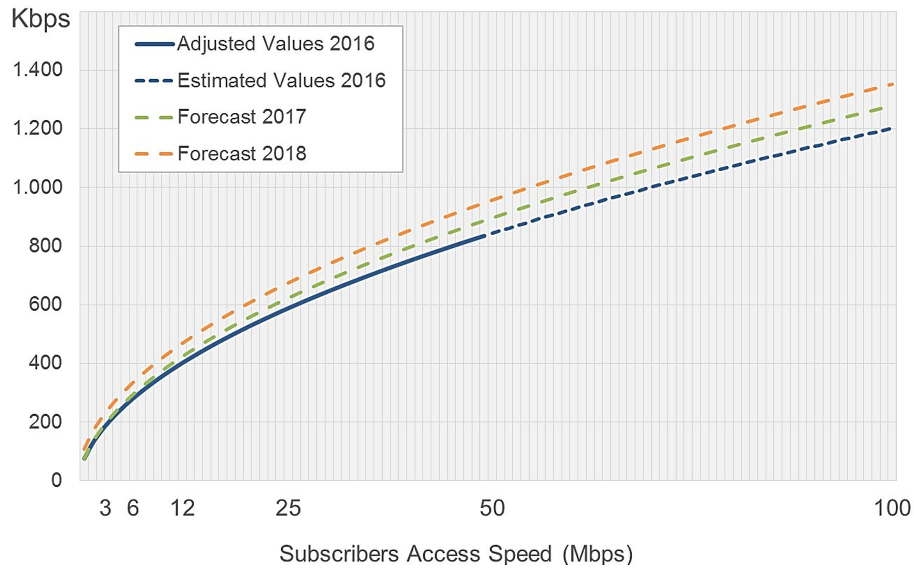

Figure 6 - Linear Regression of the average bandwidth per subscriber. Blue line represents the adjusted and estimated current values. Green and orange lines, the forecast for the next two years.

3.3.1. Dimensioning tool based on simulations

We developed a tool that aims to obtain a set of simulations to evaluate the use of bandwidth by service group, traffic consumption, customer portfolio and its evolution over time. The tool outcomes allow us to visualize measurements, make estimates and use projected values to simulate consumption.

The traffic sizing criteria adopted suggests taking the maximum 95th percentile, within a 15-day period, to predict the estimated growth at a given time. We take these statistics as sample of the worst hour. Simulations are performed in 24x15 scheme to compare measurements and simulations using the same criteria.

Portfolio update scenario simulation entails some essential input parameters: service tier levels distribution per service group and Avg_BW_subs values obtained through linear regression. The latter allows us to analyze service group's load, considering the number of subscribers and their respective service tier. It also allows to estimate the impact on the network by adding new tiers to the customer base as well as a massive portfolio upgrade.

3.3.2. COVID-19 and HFC traffic growth

Since the lockdown started in 2020 because of the COVID-19 pandemic, totally disruptive changes in clients' behavior began to be observed, affecting network performance. Therefore, we began to analyze the evolution of downstream and upstream traffic on a weekly basis with the aim of finding new patterns of use, that allow the growth prediction in a completely uncertain scenario and take proactive actions in the network to alleviate the impact on the service.

We presented these changes graphically to briefly understand the suffered impact after lockdown, Figure 7; and numerically to enable the comparison between past situations and predict future ones.

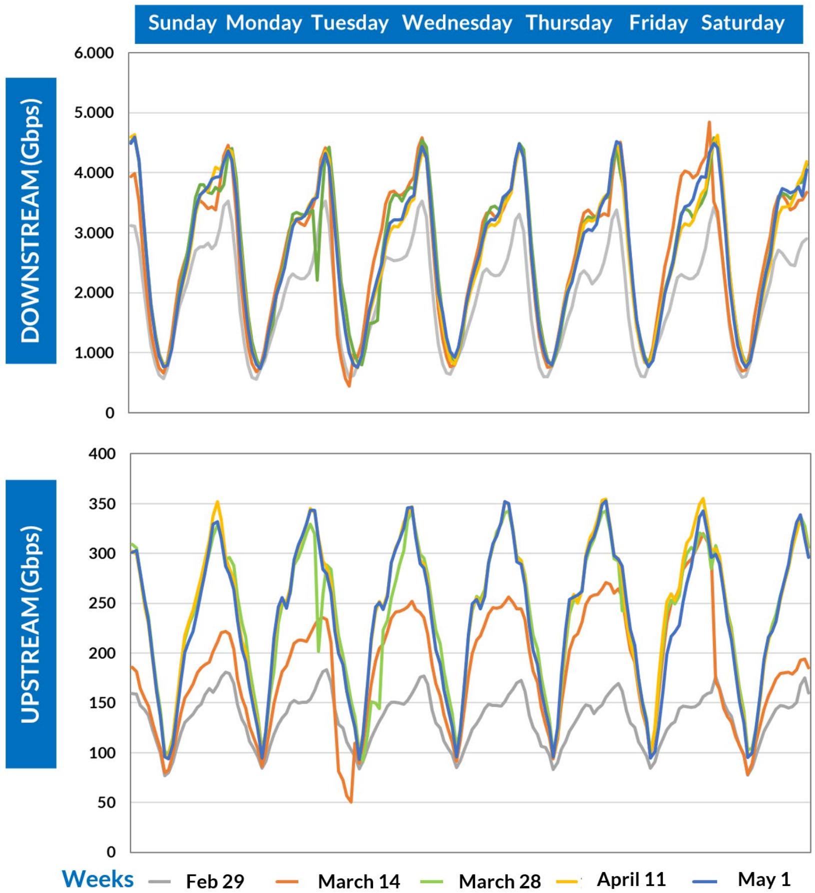

Figure 7 – Change in daily HFC traffic patterns.

Traffic increases were calculated between peak of April 30th, 2020 (post-lockdown), and the peak of the reference week, Feb 29th to March 6th (pre-lockdown), for midday and peak hour. Midday is defined as 12 p.m. to 2 p.m., and peak hours are defined as 8 p.m. to 2 a.m. Peak hours remain to be when DS and US traffic peaks happen rather than midday. For US traffic, the post-lockdown midday spike outpaced the corresponding spike pre-lockdown peak-hour by 41%. Figure 8 and Figure 9, display the evolution of US and DS. SARIMA model was implemented to describe and predict the US and DS behaviors.

The Home Office, school online courses, and Internet-related recreation are the main drivers behind this transformation. The most popular categories are video meetings (Zoom, Webex, etc.), streaming (Netflix, Youtube) and gaming; file sharing also increased. The former calls for a heavy reliance on upstream, which has not been in great demand for years.

Traffic Increase April 30, 2020	DS	US
Mid-day	4%	85%
Peak Hour	28%	93%

Max Traffic Increase	DS	US
Mid-day	37%	90%
Peak Hour	48%	95%

Table 1 – Traffic increase for a particular date and maximum traffic increase registered.

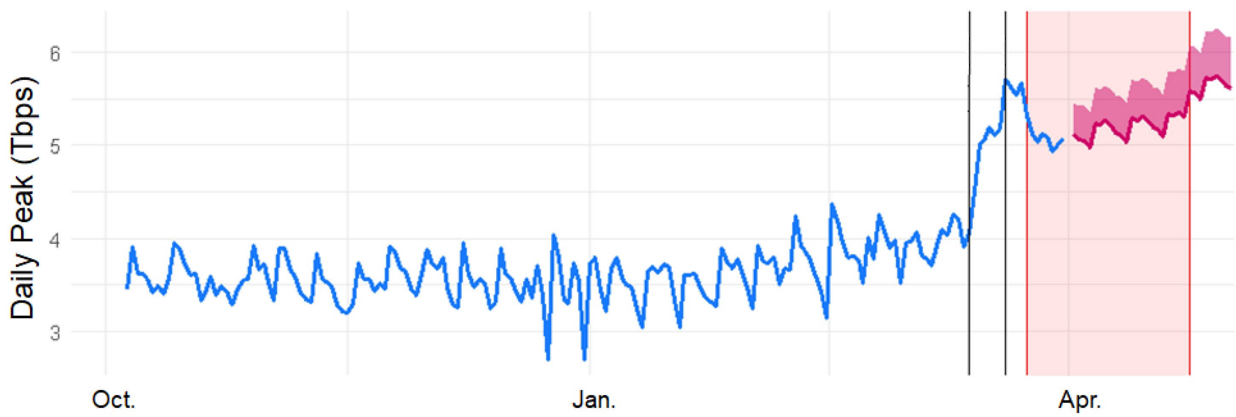

Figure 8 - Evolution and forecast for Downstream traffic, 2020.

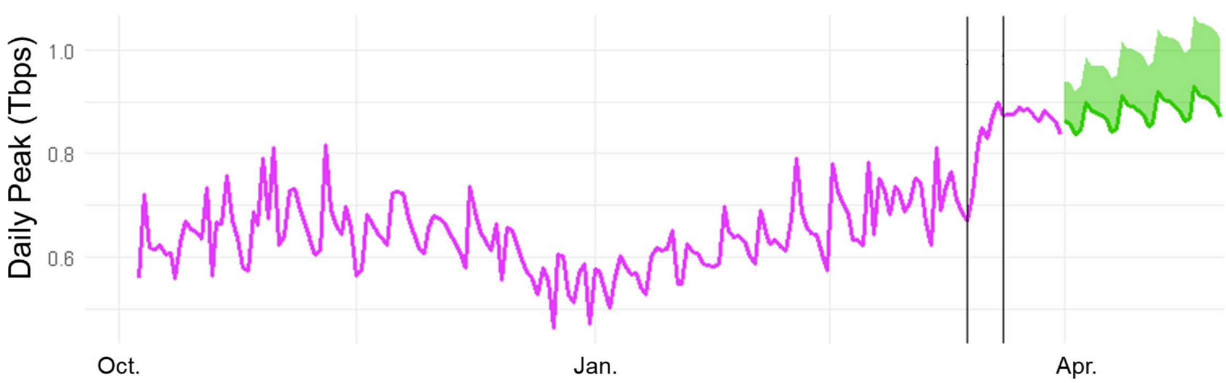

Figure 9 – Evolution and forecast for Upstream traffic, 2020.

3.4. Real-time analytics for IP video multicast

To understand the impact of multicast implementation, it was necessary to collect data on key indicators such as the number of concurrent streams, average bitrate, and average bandwidth to estimate bandwidth gain [13].

3.4.1. Multicast gain formula

We evaluated the multicast gain at service group levels as a percentage of the capacity needed under a 100% Unicast scheme, which we define as follows:

$$Capacity\ 100\%\ Unicast = \sum_{\substack{All \\ channels}} Concurrence \cdot Avg\ bitrate$$

When working with data from the Legacy system, we used the access frequency to approximate the concurrence, and we assumed that the average (Avg.) bitrate is 4 Mbps.

We examined different scenarios defined as follows: *Top "X"*, the *"X"* most popular channels are delivered to Multicast, and the rest remain Unicast.

The capacity needed is calculated as:

$$Capacity\ Top\,\text{"X"}\ Scenario = \sum_{\substack{Top\ \text{"X"} \\ channels}} Avg\ bitrate + \sum_{\substack{Other \\ channels}} Concurrence\ \cdot Avg\ bitrate$$

Finally, the multicast gain is:

$$Multicast\ gain\ for\ Top\ \text{"X"}\ Scenario = \frac{Capacity\ 100\%\ Unicast - Capacity\ Top\ \text{"X"}\ Scenario}{Capacity\ 100\%\ Unicast}$$

Gain is calculated using the overall concurrence for each channel. Additionally for service group level gain, internal service group concurrence is used.

In addition, a theoretical scenario is proposed, in which all the channels are transmitted via multicast to estimate the maximum multicast gain. This is helpful to determine whether the gain in other cases is nearly at its maximum or not.

$$Capacity\ 100\%\ Multicast = \sum_{\substack{All \\ channels}} Avg\ bitrate$$

The maximum gain is estimated as:

$$Maximum\ Multicast\ gain = \frac{Capacity\ 100\%\ Unicast - Capacity\ 100\%\ Multicast}{Capacity\ 100\%\ Unicast}$$

When there are approximately 10 channels delivered via multicast, it has been shown that the gain is approximately 50% during peak hours. The marginal gain tends to decline as more channels are multicast supplied. It was discovered that the benefit almost reaches its maximum with 25 multicast channels.

3.4.2. K-means clustering applied to the selection of multicast channels

K-means algorithm was used on ranking data to investigate the channel count that would be provided using multicast if it were an autonomous and unsupervised procedure, Figure 10. We discovered this technique categorized between 4 and 9 channels as the most popular after evaluating six months of worth data.

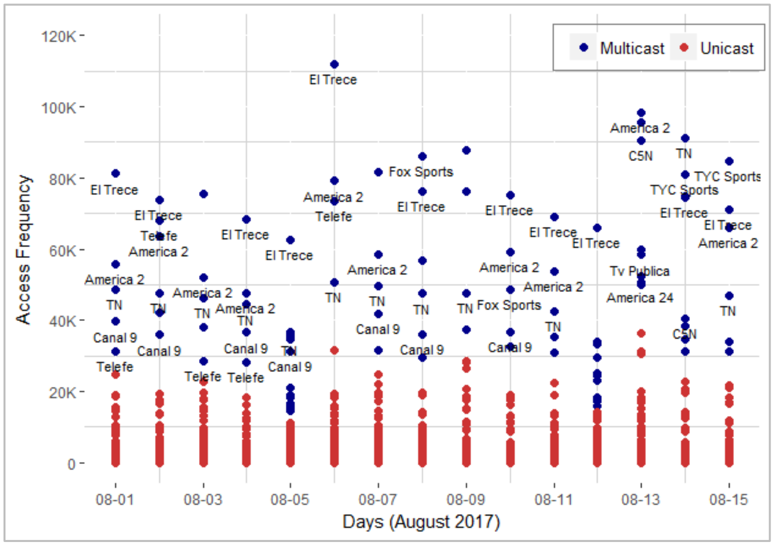

Figure 10 - K-means clustering applied to the access frequency per channel by day for the Legacy system. Algorithm used to classify the signals between multicast and unicast.

Real time analytics provides an efficient alternative for monitoring a policy-driven multicast strategy, due the advent of unconsidered special event which could drastically shift the ranking for a few hours.

It is proposed a continuous process, based on k-means clustering algorithm, executed every 10 minutes. The application searches for the channels with the highest viewing rates and determines if they are on the list of multicast channels. It sends an alert and reports the top list when it detects one or more channels that are being accessed massively and are not in the multicast list.

4. Evolving with AI and Big Data

Machine Learning has made significant advances in the telecommunications business, delivering numerous benefits. Given the tendency of strong virtualization, it simplifies operations.

However, we must not forget that the success of the application is dependent on the duties performed by people. In general, the tasks must be established to locate the learning models are as follows:

- Data: Split data into development and validation. Define instances, classes and attributes.
- Experimentation: Attributes selection. Performance metrics. Cross-validation.
- Model validation: Processes intended to verify that models are performing as expected, in line with their design objectives and business use case. It's the most important step in the model building sequence.

At Telecom, there are a few initiatives that we have considered within the AIOps framework. At STEM team one of our missions is to lead AIOps in our current and future networks. Our recipe is diversity, work in cells, agile mindset, and self-learning. In the following sections we present two of the initiatives we have been working on.

4.1. Network capacity and Machine Learning

We proposed the use of machine learning techniques such as Principal Components Analysis (PCA) and Artificial Neural Networks (ANN) to characterize the optical nodes that integrate our network and define the strategies the company will use to meet short and long-term demand [4].

We conducted different analysis at node level based on variables such as monthly consumption, households passed, traffic per port and downstream channels distributions, among others. We used a huge volume of data from different sources to obtain examples for the training sets used in the algorithms. Due the produced results are needed for a significant number of cases and on a regular basis, the process must be automated by applying machine learning and multivariate analysis techniques.

This analysis used traffic-data collected every Sunday during prime time, from all network's ports.
One variable in the datasets contains the maximum traffic (Kbps) registered in each port. Our approach consisted in analyzing two key indicators:

- Average bandwidth traffic per residential subscriber at peak time.
- Ports usage.

The first one will provide information about the zones where there is a need for higher bandwidth, and the second will help us find the optical nodes where ports are operating at almost their full capacity, conditioning the Quality of Service (QoS) and limiting the demand.

To assess the average bandwidth traffic per subscriber at peak time, the metric was defined:

$$Average\ BandWidth\ per\ Subscriber\ [Kbps] = \frac{Port\ Traffic}{\#\ Subscribers\ connected} \qquad (1)$$

For measuring ports usage, the maximum utilization was defined:

$$Max\ utilization\ [\%] = \frac{Max\ Port\ Traffic}{Port\ capacity} \qquad (2)$$

To gather data about how the two key indicators relates to other variables, we also included in our analysis: the count of segments or zones connected to the port, CMTS model, optical node classification according to the region of location, investment plan status, network capacity (1GHz or other), DOCSIS version, cable modems count, HHP, network extension (in Km) and total monthly downstream consumption.

4.1.1. Principal Component Analysis (PCA)

Principal Components are the underlying structure in the data. Their interpretation is based on the weights obtained from the original variables. PCA is a way of identifying patterns in data and expressing it with fewer variables.

We made a PCA for the ports database. The variables included were:

- Maximum traffic per port for each Sunday
- Count of downstream channels in use per port (Channels_used)
- Count of areas connected to each port (Areas_port)
- Households passed (HHP)
- Residential subscribers per port (Subscribers)

- Traffic Management

We concluded there are two main PC, which can be explained as follows:

- PC1: It takes higher values for those ports that registered more traffic during the time surveyed. Channels in use, areas, amount of cable modem and HHP per port also have a positive yet lower impact.
- PC2: This component takes higher values as the number of areas, cable modems and HHP per port increase, as well as traffic management. On the other hand, it takes lower values as the number of downstream channels in use increases. This variable informs about a port's incapacity to provide a good service in highly populated areas.

PCA highlighting there are two port groups, Figure 11. One where the aggregation of more subscribers draws a substantial increment in the traffic, and another where the impact of adding subscribers is lower.

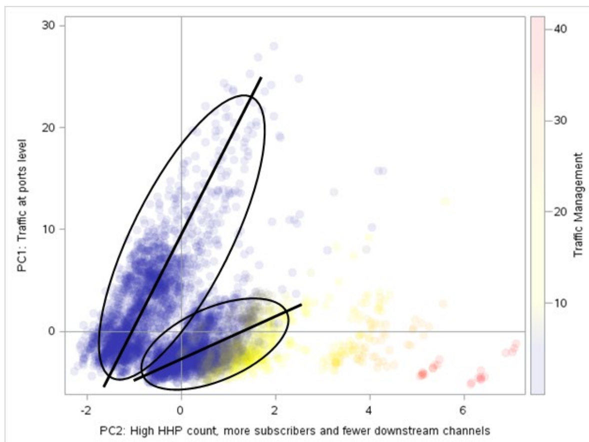

Figure 11 - PC1 vs PC2 and its relation with traffic management.

4.1.2. Artificial Neural Network (ANN)

We based our ANN on the following principles:

- Parsimony Principle: the simplest model that fits the data is also the most plausible.
- Sampling Bias Principle: if the data is sampled in a biased way, then learning will produce a similarly biased outcome.
- Data Snooping Principle: if a dataset has affected any step of the learning process, its ability to assess the outcome has been compromised.

We determine one of these four options to increase network capacity and, as a result, access speed: chassis upgrade, recombination, node segmentation, and node division

For the neural network training we first needed a sample of nodes to be classified by the expert team. Sample characterization was made by classifying nodes into three strata: the first one has the nodes in which their ports have a mean utilization below 50%; the second stratum contains nodes where mean utilization lies below or equal to 80%, the third groups the nodes with mean utilization above 80%. It is in our interest to have a faithful representation of the HHP variable in the sample.

For overfitting avoidance, sample data was split in training (60%), cross-validation (20%), and testing (20%). Training set was used to estimate the weights in the neural network. Cross-validation helped validating the model in terms of variables and optimization of the selected parameters. The testing set wasn't used in the construction of the neural network but to check whether there was overfitting or not, by measuring network classification performance with 'new observations. Quadratics terms were included to search for higher accuracy. The ANN scheme got is shown in Figure 12.

Network's first layer contains eight inputs, the four variables mentioned (represented as x) and the same variables at square (x^2). Second layer, also called hidden layer, contains eight data points too ($a^{(1)}$), and the output layer contains five classes ($a^{(2)}$), which refer to the four strategies already detailed and the fifth option 'no action needed', for the nodes where no investment was needed at the time.

The optimal weights solution in the network threw an accuracy level of 96% with the training set, and an accuracy around 90% with testing set.

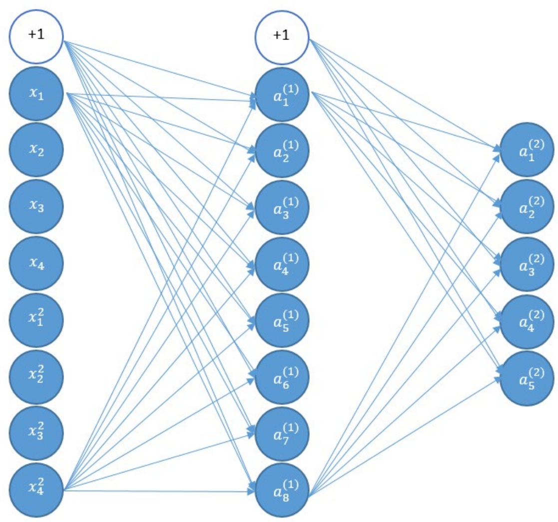

Figure 12 – Our ANN scheme.

4.2. Optimizing video customer experience with Machine Learning

For video service providers it is important to evaluate the processes that affect the quality of video considering the perception of customers. The subjective evaluation of the quality of the video measuring the opinion of human users is expensive and slow, although there are public databases with standardized results. To automate the evaluation of quality of video objective models that try to approach the subjective evaluation human are used. Recently, objective models emerged using Machine Learning (ML) algorithms which are trained using databases with subjective evaluations, to combine a variety of classical metrics. Classical metrics are much simpler to implement and at a lower cost but, they produce worse results that do not always fit the human perception. On the contrary, the metrics based on ML produces results very close to the subjective opinion of the customers, but they are more complex to implement and provide development opportunities.

Within the objective metrics based on ML there is a method named Video Multimethod Assessment Fusion (VMAF), an open-source method proposed by Netflix in 2016 and is a video quality metric that combines human vision modeling with ML to provide a great viewing experience to their members.

Telecom has its own IPTV platform, called Flow, that is based on unmanaged (second screens) and on managed devices (set top boxes). It provides different types of advance video services, including Linear TV, various flavors of On Demand services (VoD, CuTV, Reverse EPG, StartOver, network DVR, Pause Live TV, and Trick Modes), using different streaming technologies, Search and Recommendations. Thus, it's very important for the Company to develop VMAF as a tool for optimizing Flow customer experience and to equalize video quality with other existing video platforms.

In order to train VMAF for optimized Flow customer experience we defined a dataset following Netflix recommendations regarding the type of content. We selected 35 videos, each 10-sec long from Flow catalog. To make the distortions, each source video was encoded with 6 resolutions up to 1080p. Video characteristics were variable, and used a selection of videos with fire, water, nature, animation, close-up, action, crowd, among others.

We ran a subjective test through 6 different focus groups. Each group of about 15 subjects. Each subject sits in a living room-like environment and was instructed to watch an unimpaired reference video, then the same video impaired and give a rating on a continuous scale from "bad" to "excellent" (ACR methodology), then we translated the scale to a range from 1 to 5 and calculated the MOS.

Then, we trained several models with different sets of parameters for the Support Vector Regressor, the ML model used to perform VMAF (Figure 13), to avoid the underfitting and the overfitting.

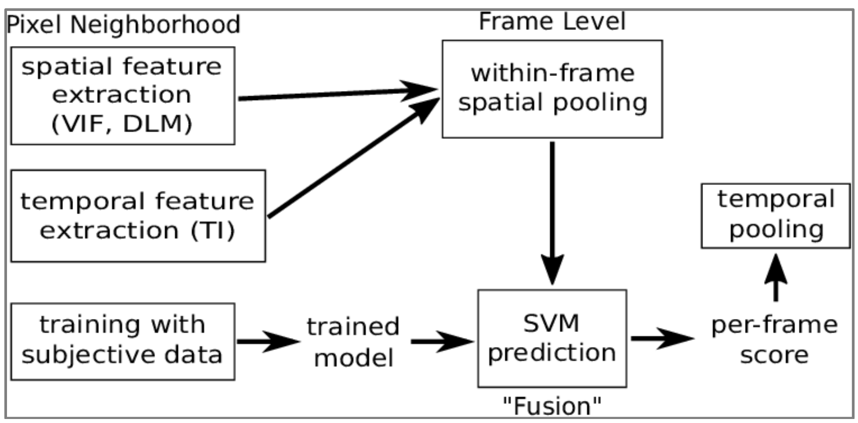

Figure 13 – VMAF Algorithm.

Once we had our model trained, we tested it with the testing dataset previously defined. After the testing, we calculated the performance metrics to measure the prediction accuracy (60%). Based on the results we obtained, we understood that we must continue improving the model with larger training and testing datasets or with other elementary metrics combination. In any case, the accuracy of machine learning models that use subjective variables does not usually exceed 70%.

5. AIOps

The term AIOps was coined by Gartner in 2016 and have pushed the concept into the marketplace. According to Gartner "AIOps combines big data and machine learning to automate IT operations processes, including event correlation, anomaly detection and causality determination". And, according to Figure 14 has three main elements: observe, engage, and act.

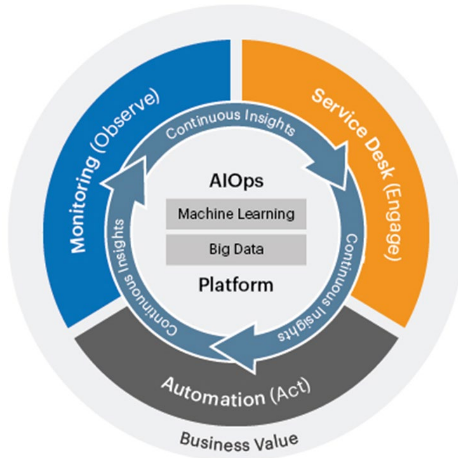

Figure 14 – AIOps cycle by Gartner.

While AIOps was developed to give scalability to the management of IT systems, given their increasing complexity and framework, we can extend its applications to Telecommunications operations.

A first definition of AIOps in the telecommunications industry is the use of Artificial Intelligence for the operation of networks and services. The long-term goal is to achieve autonomous networks (AN), which in some ways, at the beginning, entails automating many of the operational procedures [5].

Automation in AIOps is understood to range from automating network capacity planning through some ML algorithm, automatically detecting anomalies in traffic flows through ARIMA time series, or even automating a process of adjusting a modulation profile in the form autonomous (closed loop) (Figure 15).

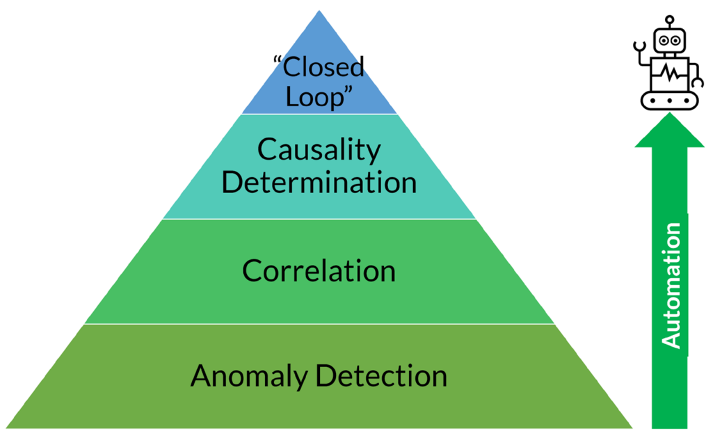

Figure 15 – AIOps automation path.

Gartner also defines four levels of Network AIOps, that we can see in Figure 16.

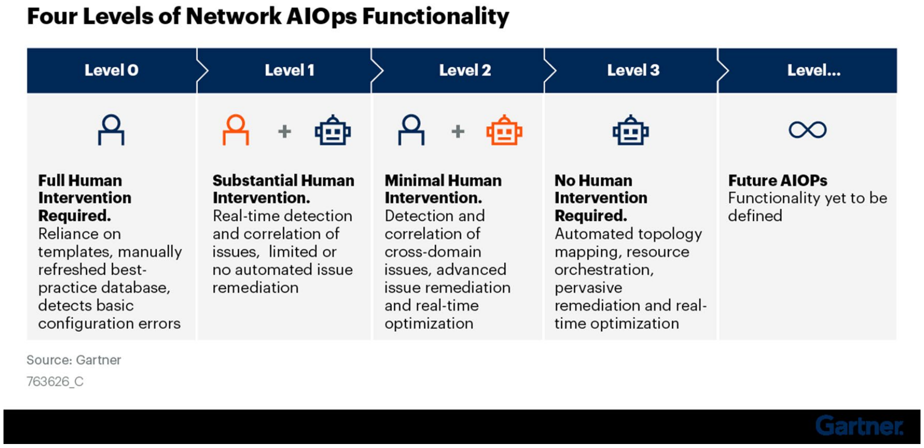

Figure 16 – Levels of Network AIOps functionality.

We adopt AIOps as a framework to continue our journey towards an AI-driven operation. While autonomous networking is our goal, we will only make a brief reference below.

In summary, AIOps refers to the application of AI technology to automate business processes and operations at CSPs.

5.1. Ada

At Telecom Argentina, we have been working on a network dimensioning solution since a long time ago. In 2017 we developed a tool that assisted the decision-making process for HFC network dimensioning, mentioned in a previous section. Later, we incorporate data from the radio-access network (RAN) and relaunched the project under the name Ada. The goal is to have a tool that combines machine learning and subject matter experts (SMEs) input to assess investment decisions on CAPEX and OPEX.

Historical data is collected on every HFC node, as well as for every cell on the mobile network. The concept is that for short-term decisions (two years forward), ML-based forecasts are offered, and the duty of experts is to identify priorities and act on them. Long-term decisions (5-10 years periods), on the other hand, ML-based forecasts would require a longitude of history that hasn't yet occurred. In our experience, however, it is possible to provide reasonable approximations if forecasts are supported not only by historical data but also by experts' knowledge. Hence, a ML clustering model has been added, which characterizes the nodes according to their traffic and other variables. Based on the short-term forecast, nodes characterization, and feasible tasks that can be carried out on the network (such as migration from HFC to FTTH), possible long-term scenarios are obtained.

Planning engineers used to research on potential scenarios and make a series of calculations to approximate what they thought was going to happen in the long-term, based on their knowledge of the average client and use cases. Working together, we built profiles for the operation area, so they can refine what to expect from a variety of use cases. For example, consider two cases where a household group mostly uses the service for social media and the income is low, with another group with heavy streamers and high income. We provide better information to planning teams at the beginning. An example of the Ada application is displayed on Figure 17.

We continue working on how to enable engineers to pass information to a model about what they expect the CAGR would be at different kinds of operating sectors, in the next 5-10 years. This will involve a combination of simulation and forecasting techniques. It will also require a certain level of automation that is not achievable outside a ML framework.

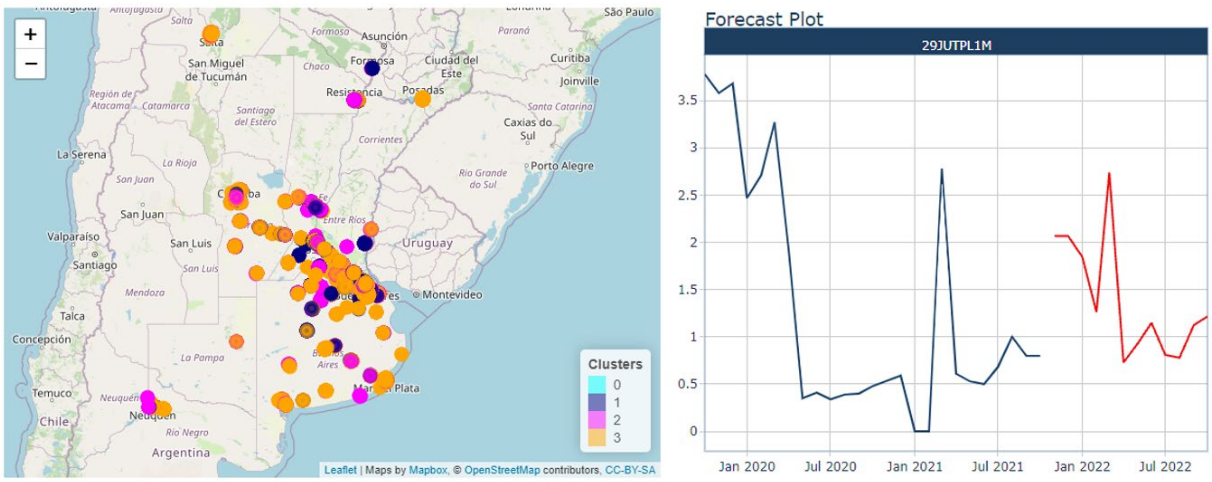

Figure 17 – Clustering and Forecast performed with Ada.

5.2. Customer claim prediction for HFC network

Within the domain of AIOps one of the most popular use cases is customer claim or tickets prediction. The main idea is to use information derived from different elements of the network, such as: CMTS, cablemodems, electronic devices, etc. to estimate the probability of a customer to generate a complaint.

Information is generally collected through OSS systems and ingested in a machine learning pipeline where preprocessing, analysis and ML model testing is performed. We are currently developing this kind of solution to ultimately increase customer satisfaction.

Using hourly collected information from over 3.5 million DOCSIS 3.0/3.1 cablemodems we are trying to anticipate customer complaints two days in advance. To handle this amount of data we have partnered with Google to develop and deploy this project using Google Cloud Platform (GCP) services. A high-level view of GCP implementation is shown in Figure 18.

Potentially relevant variables (e.g., signal to noise ratio, consumed bytes, average Rx, t3 and t4 time outs, etc.) have been identified and our Service Assurance team have been able to efficiently transfer these data, collected by our OSS systems, to GCP.

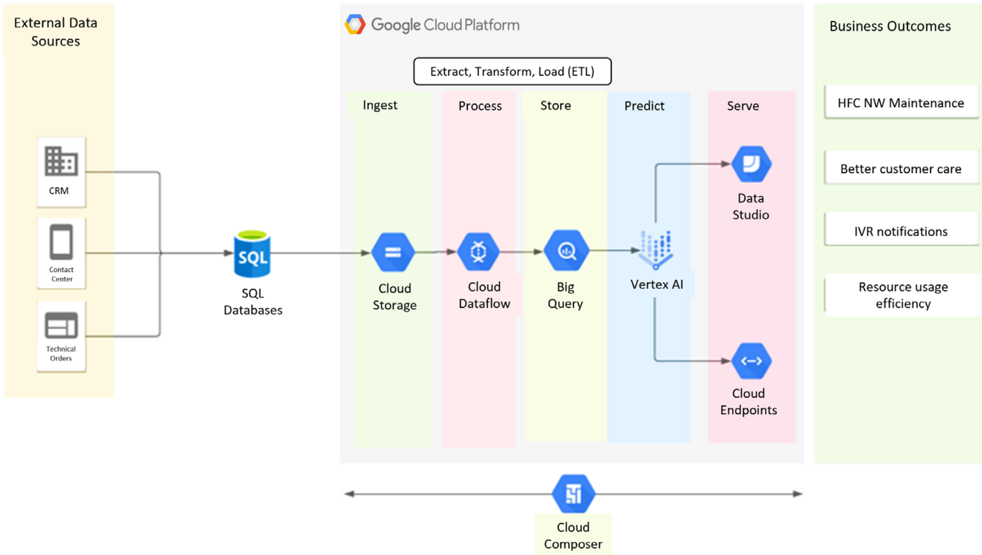

Figure 18 - High-level view of our project: inputs, outputs, outcomes and architecture.

Regarding technical challenges, two of the most difficult tasks have been dealing with an extremely unbalanced dataset and label noise. Although from a business perspective having a low proportion of customer claims is a good indicator, some ML models can struggle to learn from these kind of data sets. In our case, after filtering the target population, the positive class dropped to 0.1%, which added more complexity to the problem. To reduce the impact of this issue on model performance we applied hyperparameter tunning and a technique called SMOTE combined with an under sampling of the majority class. Both approaches led to improvements in model performance.

We have been able to overcome many challenges and finally the ML model selected was XGBoost. As a result, we obtain a daily list of customers with high probability to claim. From this list, with CX and Field Service teams, we can make proactive calls to solve customer problems remotely or to send a technician to their home if necessary. Then, we measured through surveys and NPS how was the experience of our customers and we found that this proactive action has a positive impact on their satisfaction.

Although we are going through an early stage of this project, we understand that the future result will be to increase the number of promoters of the company and consequently avoid churn. In addition, the truck roll and the costs associated with it would be reduced.

5.3. Intelligent agents and Autonomous Networks

Peter Norvig and Stuart Russell present in [6], eight definitions of AI, from which they define Intelligent Agents as “agents that receive precepts from the environment and take actions that affect the environment.” Agents’ actions change the world. Thinking (Cognition): Interpret sensory data. Then updates its environment (model of the world) and, decides on next best action. The autonomous networks that are being defined in the TM Forum and ITU are based on this definition.

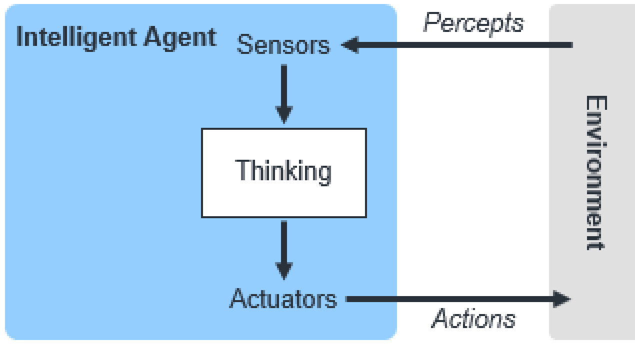

Figure 19 – Intelligent agent.

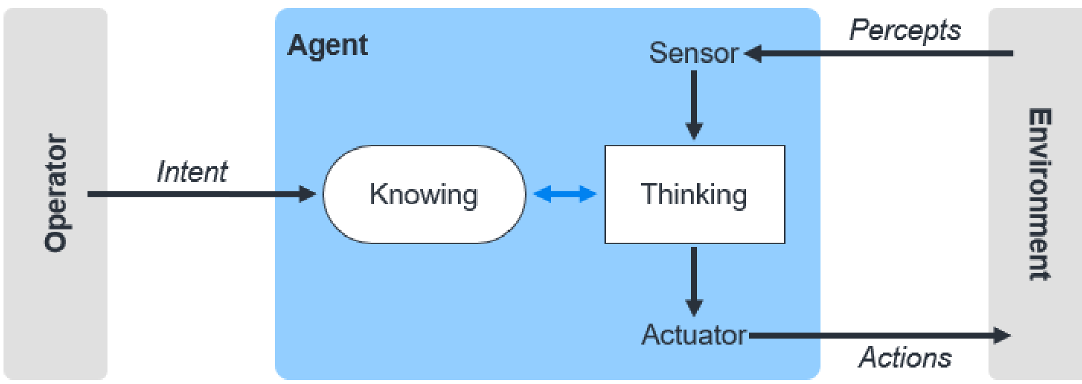

Figure 20 – Agent in AN source: Autonomous Networks Technical Architecture (IG1230). Source: TM Forum

There are two concepts we have been working on with our stakeholders and team in the CTO office.

- Autonomous Networks make decisions without human intervention (decisions made by Agent).
- Automation operates without human control (can be implemented without AI technology, as we have seen).

Another related idea is "intent-driven automation". Intent-based networks enable service providers to define the behavior they expect from their network from service and business perspective. Intent was first introduced around 2015 in the context of SDN controllers. "Intent is the formal specification of all expectations including requirements, goals, and constraints given to a technical system".

This definition is inspired by and compatible with the definition IETF has published in 2020 [10]. The definition associate's intent with goals, requirements and constraints provided in a declarative way. Intent constitutes and expresses knowledge about these concerns and enables sharing this knowledge between the originator and receiver of the intent.

"Autonomous networks are those that possess the ability to monitor, operate, recover, heal, protect, optimize, and reconfigure themselves; these are commonly known as the self-properties" [11].

The impact of autonomy on the network will be in all areas including planning, security, audit, inventory, optimization, orchestration, and quality of experience. At the same time, autonomy raises questions about accountability for non-human decisions that affect customers.

The growing virtualization and cloudification (Telco-Cloud) of our networks make possible the evolution towards AN. At Telecom, we use the AIOps framework since we know that it will lead to AN through automation [5].

6. Learned lessons

In this part, we'll highlight some of the lessons we've learnt from our experiences or those of other CSPs with whom we've collaborated on AI working groups for network and service operations. Over the years, adopting an AIOps framework and making our operations AI-driven have required significant culture change and teamwork, all in line with the company's technology strategy. The fact that AI is used in the process does not imply that we are AI-driven.

$$\text{Operations} + \text{AI technology} \neq \text{AI-driven}$$

We are training our engineers to focus on the operations-related tasks that are more challenging and where they can bring the most value rather than the less interesting and repetitive ones that can be handled by the systems. As indicated by the AIOps TM Forum framework, engineers working on AI-systems exception management and continuous optimization.

6.1. Define and communicate AI-driven operations

A lot of expectations and skepticism occur when AI is introduced because of the numerous movies that have been made over the past few decades. Because of this, it is crucial to carry out an internal communication and evangelizing work on what we mean by AI-driven operations. The various stages of the analytical component must be clearly discussed.

- Descriptive – What is happening now?
- Diagnostic – What happened and why?
- Predictive – What might happen?
- Prescriptive – What actions should be taken?

We need to plan out how we can effectively inform the various business teams about these AI-driven projects.

It's also important to mention that operations engineers view AI as a tool for augmented intelligence (AgI).

While AI builds machines that behave and function like people, AgI uses the same machines but takes a different tack to enhance human skills. AgI actually entails a collaborative effort between humans and machines that makes use of each party's advantages to boost overall commercial value. In other words, AgI's main objective is to enable people to work more effectively and smarter. [7].

Finally, it is also important to communicate what problems can we solve, and which ones cannot, their scope and limitations.

6.2. Workforce

Believing that a pair of data scientists can resolve a network operating or planning problem is one of the most frequent errors we have observed.

Teams using AI should be set up with professionals in data sciences, operations engineering, field service, customer experience and also with technicians to tackle operational challenges. We need to build internal AI teams.

Like many businesses throughout the world, our main challenge is finding experts who are knowledgeable in programming, telecommunications, and AI. We have junior and semi-senior data scientists on the STEM team because of this. Tech Scientist is the next level of seniority. The tech scientist has expertise in the technology that he uses in collaboration with the operations and field service engineers, in addition to data science and programming skills.

In several talks with colleagues from other companies, they have even raised the difficulty of recruiting data specialists. Because of this, digital training and recruitment are a priority in Telecom. Along with learning about communication technologies, our engineers at the CTO office continually train in data sciences. Our aim is to keep our work teams AI-driven learning continuously so they may continue to advance their careers and bring value to our clients.

Our recipe is diversity, work in cells (with the experts of service assurance, field service, customer experience, etc.), agile mindset, and self-learning.

To ensure that these people choose to work at Telecom despite the rising demand for these profiles in this hyperconnected world, we are creating a reskilling strategy for our employees as well as loyalty and retention programs.

Automation will also make it possible for operational departments to integrate technology to streamline procedures and improve our customers' digital experiences, which will further raise business efficiency.

To sum up, we are establishing a strategic talent plan for transforming our current teams into an AI-enabled workforce.

6.3. Models, algorithms and Explanaible AI

Always keep in mind that "All models are wrong; some models are useful," and that the key is for the work team to agree on which model to adopt.

Once the useful model has been adopted to solve an operation or decision-making problem, we must somehow make the models transparent, understandable, and interpretable. Experts must comprehend how the findings were obtained and the degree of confidence that the model has before they can deciding to adopt an AI tools.

The model's output shouldn't be what matters, but rather understanding why an algorithm produces a particular output. Because of this, explainable AI (XAI) is a new growing field. The goal of enabling Explainability in AI/ML, as stated "is to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms" [7].

Explainability sits at the intersection of transparency (consumers have the right to have decisions affecting them explained in understandable terms), causality (it is expected of the algorithms to provide not only inferences but also explanations), bias (the absence of bias should be guaranteed), fairness (it should be verified that decisions made by AI are fair) and safety (reliability of AI systems) [8].

We know that many machine learning algorithms have been labeled "black box" models because of their inscrutable inner workings. What makes these models accurate is what makes their results difficult to interpret and understand. They are very complex. The discussion about audit AI is still open [7].

6.4. Cloud and Data/AI platforms

The democratization of data by hybrid clouds is another lesson learned. “No amount of AI algorithmic sophistication will overcome a lack of data (architecture), Data collection & preparation is the most time consuming and difficult part of AI” [12].

There is a project in our OSS systems evolution program that is solely responsible for gathering and organizing data. In complicated hybrid multi-cloud systems, we are employing platform services to tackle problems using data and AI. but before using platform services, one should give them a serious evaluation.

Platform services are increasingly being used by CSPs to accelerate the use of AI in business operations. We should take a hybrid approach to deploying platform services.

7. Next steps

Our next step is to define long-term roadmap towards Autonomous Networks. We anticipate finishing the surveys, definitions, frameworks, and scope for the strategic definition toward the autonomous networks paradigm by the end of the year. Virtualization and softwarization of networks are clearly included in this strategic framework. This ecosystem includes 5G and IoT.

In the AIOps framework, we began by identifying the claim root cause and its causality[1] using our model of customer claim prediction for the HFC network.

Conclusion

We offer the AI-driven initiatives that were significant turning points in terms of lessons learned since we started our journey. Right now, our focus is on developing within the AIOps framework.

AIOps' objective is to advance from automation to autonomous networks, but we must remember to see it from the standpoint of augmented intelligence (AgI). Together, people and robots may maximize the value of their own capabilities for the benefit of the organization.

Include operations engineers and technicians early in the use case, and keep in mind how crucial Explainability of AI models and their results are.

To reach full AN it is necessary to advance to the "Telco Cloud".

Cultural and process change is required. **"Think big, start small".**

Abbreviations

4G	fourth generation wireless
5G	5th generation mobile network
6G	sixth-generation wireless
ACR	Absolute Category Rating
ADSL	Asymmetric Digital Subscriber Line

[1] We are trying to apply AI models with causal inference methods

AgI	Augmented Intelligence
AI	Artificial Intelligence
AIOps	Artificial Intelligence for IT Operations
ANN	Artificial Neural Network
AP	Access Point
ARIMA	Auto Regressive Integrated Moving Average
AVG	Average
bps	bits per second
BW	Bandwidth
CAGR	Compound Annual Growth Rate
CAPEX	capital expenditure
CDN	Content Delivery Network
CI	Confidence Interval
CMTS	cable modem termination system
COVID-19	Coronavirus disease of 2019
CSP	Communication Service Provider
CTO	Chief Technology Officer
CX	Customer Experience
DASH	Dynamic Adaptive Streaming over HTTP
DIY	Do It Yourself
DOCSIS	Data Over Cable Services Interface Specification
DPI	Deep packet inspection
DS	Downstream
DSP	Digital Service Providers
DVR	Digital Video Recording
EPG	Electronic programme guide
ETL	Extract Transform Load
ETSI	European Telecommunications Standards Institute
FEC	forward error correction
FTTH	Fiber to the home
Gbps	Gigabits per second
GCP	Google Cloud Platform
GHz	GigaHertz
HD	high definition
HD	High Definition
HFC	Hybrid Fiber-Coaxial
HHP	household passed
Hz	hertz
IETF	Internet Engineering Task Force

IPTV	Internet Protocol television
IT	Information technology
ITU	International Telecommunication Union
K	kelvin
Kbps	Kilobits per second
Mbps	Megabits per second
ML	Machine Learning
MPEG	Moving Picture Experts Group
NFV	Network functions virtualization
OKR	Objectives and key results
OPEX	Operational expenditure
OSS	Operational Support System
OTT	Over the top
PCA	Principal component Analysis
QAM	Quadrature Amplitude Modulation
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RTT	Round trip time
Rx	Reception
SA	Service Area
SCTE	Society of Cable Telecommunications Engineers
SD	Standard Definition
SDN	Software-Defined Networking
SME	subject matter expert
SMOTE	Synthetic Minority Oversampling Technique
SQL	Structured Query Language
STB	Set-top Box
SVM	Support Vector Machine
Tbps	Terabits per second
TCP	Transmission Control Protocol
TM Forum	TeleManagement Forum
Tx	Transmission
US	Upstream
USA	United States of America
VMAF	Video Multimethod Assessment Fusion
VoD	Video on Demand
W	Window time
XAI	Explainable AI

Bibliography & References

[1] C. Righetti, E. Gibellini, C. Carreño and G. Carro, Can Future Networks Survive Without Artificial Intelligence? SCTE Cable Tec Expo 2019, 2019.

[2] L. Takacs, "On Erlang's Formula", The Annals of Mathematical Statistics, vol. 40, no. 1, pp. 71-78, 1969.

[3] T. Cloonan, J. Allen, M. Emmendorfer. Simulating the Impact of QoE on Per-Service Group HSD Bandwidth Capacity Requirements. SCTE Expo 2014.

[4] C. Righetti, E. Gibellini, F. De Arca, C. Carreño, M. Fiorenzo, G. Carro, F. Ochoa, Network Capacity and Machine Learning. SCTE Cable Tec Expo 2017, 2017.

[5] Claudio Righetti, Mariela Fiorenzo, Emilia Gibellini & Martin Juiz The Evolution Towards Autonomous Networks - 2021 FALL TECHNICAL FORUM PROCEEDINGS -SCTE-NCTA-Cablelabs

[6] Stuart J. Russel, Peter Norvig 2003: “Artificial Intelligence, A Modern Approach“ Third Edition (2016)

[7] C. Righetti, M. Fiorenzo, et.al., Augmented Intelligence: Next Level Network and Services Intelligence, SCTE NCTA CableLabs 2020 Fall Technical Forum, 2020.

[8] S. F. M. H. J. K. S. V.-T. a. H. W. S. Barocas, "The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning," 20 07 2022. [Online]. Available: http://www.fatml.org

[9] H. Hagras, "Toward Human-Understandable, Explainable AI," IEEE Computer, vol. 51, no. 9, pp. 28-36, 2018

[10] Intent-Based Networking - Concepts and Definitions March 9, 2020 https://tools.ietf.org/id/draft-irtf-nmrg-ibn-concepts-definitions-01.html

[11] ITU-T (FG-AN), Available: https://www.itu.int/en/ITU-T/focusgroups/an/Documents/FG-AN_Terms_of_Reference.pdf

[12] 2018 MITSloan ”Reshaping business with AI” https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/

[13] C. Righetti, E. Gibellini, F. De Arca, M. Fiorenzo, G. Carro, Telecom Argentina, Real-Time Analytics for IP Video Multicast, NCTA, 2018

