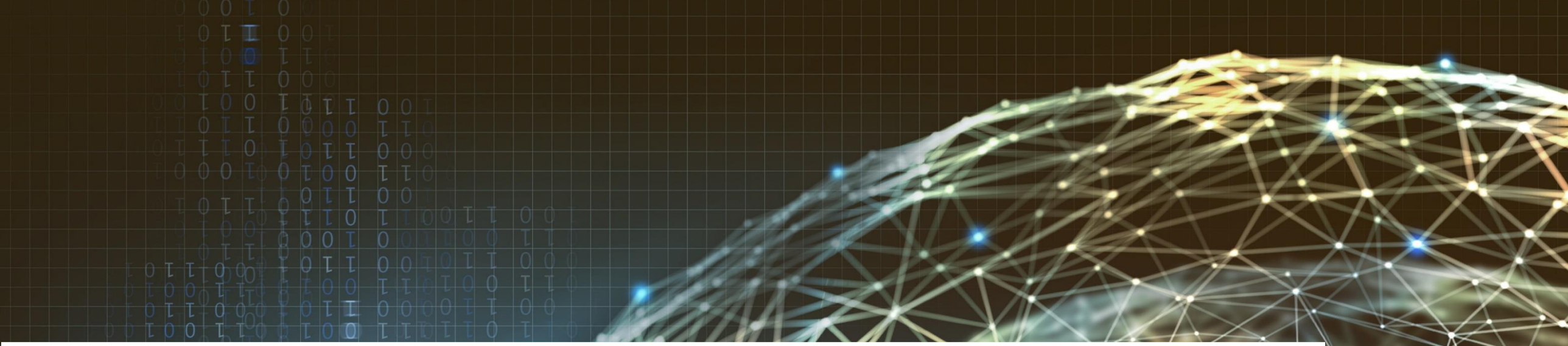
AI and machine learning: Why now?

Network optimization in the age of 5G

Monica Paolini, Senza Fili



Sponsored by

Conversations with

Table of contents

Analyst report 3

1. Introduction: coming of age of AI and ML 4
2. Why now? 5
3. Intelligence and learning 6
4. Complexity: a challenge, and an opportunity 7
5. The integration of AI/ML in wireless networks 8
6. What is the best place to learn in wireless networks? 11
7. Culture, expectations and fears 12
8. Takeaways: How will AI/ML change wireless networks? 14

Conversations 15

Intel | Establishing the AI and ML ecosystem 16

Getting machine learning off the ground 22

Uhana | Harnessing network complexity with AI 29

Glossary 36

References 37

Analyst report

1. Introduction: Coming of age of AI and ML

Artificial intelligence (AI) and machine learning (ML) have been around for quite some time, since the 1950s. They have generated a highly productive stream of work, created highly innovative technology tools, and, even more importantly, a new, far-reaching conceptual framework for learning, automation and optimization that is applicable to any process or activity. But for most of their history, AI/ML remained confined as research endeavors in academic circles and in R&D groups at companies such as IBM, with limited practical applications and hardly any economic or social impact.

In the last few years, however, AI has suddenly surged into awareness, and more recently it has been followed by ML, with the two terms often used interchangeably. Two main factors have enabled this. The first is that we now have the computational power to bring AI/ML out of the labs and apply them in the real world, across economic sectors, in different things we use or do, in a way that is affordable, reliable and secure. The second is that our world is becoming more complex, with things, people and processes increasingly connected to each other and dynamically reinforcing each other. There is not much you can do to improve a simple system. Complexity is challenging, but it also expands the scope for automation and optimization. AI/ML offer the tools and a platform to both manage and benefit from this complexity.

The promise is that AI/ML can improve what humans and machines do already but not well enough and could be done more efficiently. AI/ML can give us a better understanding of our environment, optimize performance, increase productivity and efficiency, reduce costs, and, ultimately, enhance our lives. These prospects, however, trigger two opposite reactions that may limit or slow AI/ML's impact: unrealistic expectations about AI/ML's ability to effortlessly solve all problems, and fears that they will reduce the relevance and role of humans, and possibly replace them eventually in the workplace. The success of AI/ML depends on our ability to have realistic expectations of what they can achieve and where they can help us.

In wireless, AI/ML have the potential to shake, deep down, the way we operate, and to become the foundation of the transformation that leads to the fourth industrial revolution. But to succeed at that, they require hard work, a long-term commitment, and a deep cultural change.

AI and ML: a new way to learn, automate and optimize

John McCarthy (MIT) is credited with first defining AI as "the science and engineering of making intelligent machines," during a workshop at Dartmouth College in 1956 that was notably attended also by fellow AI pioneers Allen Newell (CMU), Herbert Simon (CMU), Marvin Minsky (MIT) and Arthur Samuel (IBM).

In 1959, ML made its appearance, with Arthur Samuel quoted as defining it as the "field of study that gives computers the ability to learn without being explicitly programmed." Samuel saying that, but the attribution is widely accepted as accurate.

In 1998, Tom Mitchell further clarified the concept: with machine learning "a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

AI and ML fundamentally change the way we learn, automate and optimize. Humans mostly use explicit, declarative, deterministic knowledge and expertise to run processes or manage things, and to improve performance over time. We retain control of this process and visibility into it: we have reliable expectations about what will happen under different circumstances, and we know how things work.

AI/ML help us deal with complex problems for which we do not know the solution ahead of time. They do so by correlating data inputs with target outputs and trying to approximate the desired output, in a non-deterministic, statistical way. Because we do not prescribe how an AI or ML system has to behave, we do not have to know the solution in advance: the system learns to automate and optimize on its own, using large data sets that humans are not able to use systematically. This approach allows us to find new ways to solve problems or perform tasks in a dynamic, recurrent way, as AI and ML continue to learn from experience and automatically feed the learning into their recommendation process.

2. Why now?

In wireless, increased computational power and higher network complexity have converged to both enable and require smarter networks that need AI/ML.

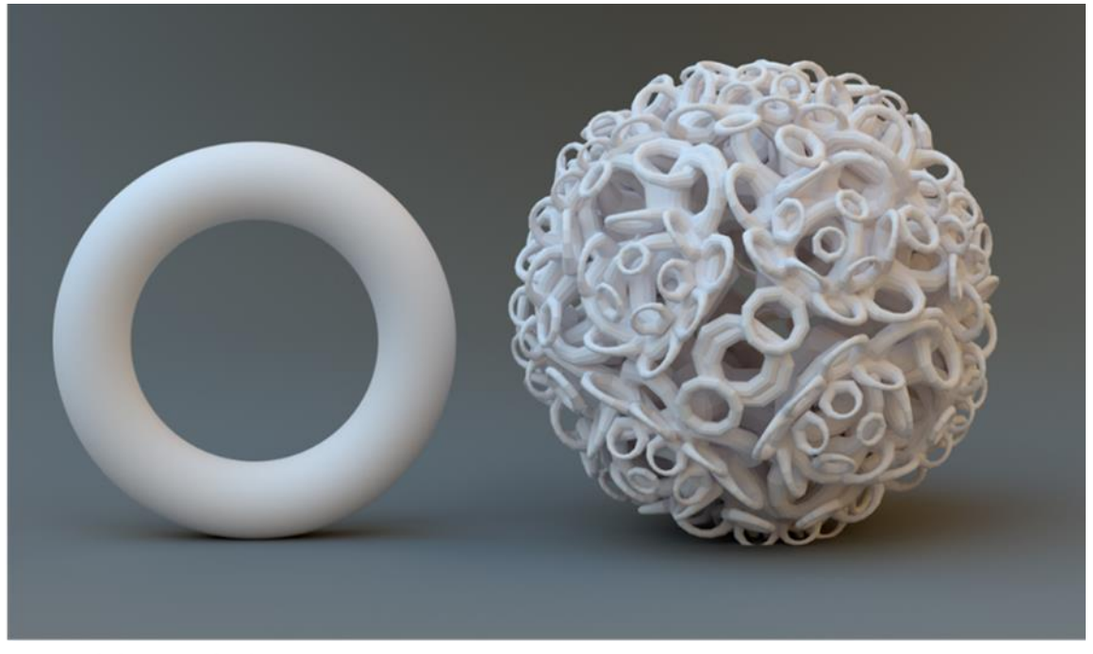
Network virtualization gives wireless operators the flexibility to choose the hardware, resource allocation and network topology they need for AI/ML platforms and big data sets, and that, at the same time, meet their compute, cost, reliability and security requirements.

The increase in network complexity that has started with 4G and that will accelerate with 5G and IoT creates a fertile environment for AI/ML. In most markets, wireless networks are still centralized, monolithic and homogeneous, and carry almost exclusively voice and best-efforts data. These networks have been optimized to some extent, but the impact of and gains from optimization have been limited, because most operators can maximize throughput and minimize latency only for the entire traffic load at a given location.

Wireless networks are evolving to include multiple access interfaces, more device types and use cases, and a wider range of applications. This creates complex networks with a much higher number of variables that can, and should, be used to optimize resource allocation and network performance. And as these variables change over time, optimization has to be done continuously, in real time.

The requirements of and gains from optimization will grow with the number of variables and the interactions among them. We are reaching a point where it is impossible or at least wasteful, to rely on humans to turn the control knobs for all these variables. This is when AI/ML, jointly with automation, become not only beneficial but also necessary to run wireless networks. Without them, operators will not be able to reap the financial and performance benefits from the investment in 5G and other new network infrastructure.

Past obstacles	Future drivers
Homogeneous, simple networks	Network diversity and transformation
Traffic mostly from voice, and best-efforts data and video	Proliferation of applications, human and IoT use cases
Limited opportunity to optimize performance	Ability to access and use massive network data sets
Limited gain from optimization	Real-time optimization
Limited computational power	Faster and cheaper processing



Growth in complexity

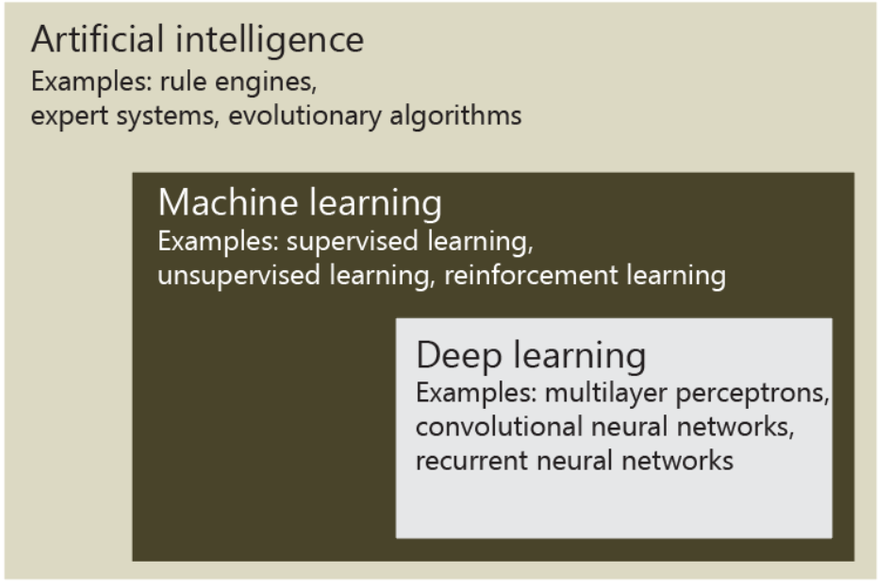
3. Intelligence and learning

In this report we use the AI/ML umbrella term to cover tools that enable operators to optimize wireless networks in an automated way, that can process large input data sets, and that can recommend the network or element configurations to meet set targets. We do this to follow industry usage and to avoid a contraposition between the two terms that does not serve any purpose. However, each term – AI and ML – on its own falls short of denoting the full scope of AI/ML as defined here.

AI historically includes models that try to replicate human intelligence or some aspect of it. This was a central goal when AI was established, and it continues to play an important role in the research community. In wireless, AI is used to complement human intelligence, rather than recreate it. We do not need human replicas, we need machines that are better than humans at performing mostly boring tasks that require computational capabilities that go beyond what humans can do on their own. In this context, "artificial" may be misleading, and "intelligence," if defined as "capacity of mind," as Wiktionary has it, is not appropriate either. The Oxford Dictionaries definition of intelligence as the "ability to acquire and apply knowledge and skills" is closer to the role of intelligence in wireless networks.

Arthur Samuel's definition of ML as learning that does not require explicit programming fits well the learning component of AI/ML. Yet it is too restrictive, because it is limited to the learning component, and AI/ML in wireless (and elsewhere) are used not only to learn, but also to make recommendations, often in real time, to optimize the network.

At the same time, while AI/ML can be used to optimize pretty much any component of a wireless network at any level of granularity, that does not mean they should be used everywhere. AI/ML require a financial investment, hard work, and a long-term commitment that are not justified for tasks sufficiently simple or well understood that they can be run by applying deterministic rules.



Source: Zhang et al, Senza Fili

Deep learning

Within machine learning, there has been growing interest and work in deep learning. The term "deep learning" was coined in 1986 by Rina Dechter, but deep learning is widely inspired by research on artificial neural networks, which started in the 1940s with the work of McCulloch and Pitts.

Deep learning is loosely inspired to the human neural architecture: deep neural networks are multi-layer networks with multiple units (neurons) connected by weighted links. The network learns by changing these weights. When the initial training is completed, the network can predict outcomes, make recommendations, recognize patterns, or identify anomalies, and, at the same time, continue to learn to fine tune its performance.

Typically, it is not clear to the human observer what these units encode – what their role or function is within the system – and this creates the perception that a neural network is a black box that absorbs inputs and spits out outputs in obscure ways. But the lack of interpretability of the nodes is exactly what makes neural networks powerful: they find new and efficient ways to learn, which are different from what humans would explicitly use.

4. Complexity: a challenge, and an opportunity

Network complexity compounds as technology advances and as wireless connectivity becomes more deeply embedded in our lives and our environment. Nokia estimates that a typical 5G node has over 2,000 parameters, and that complexity of operations will multiply by 50 as we move from 4G to 5G. The growth in complexity comes from multiple sources (see table) – in fact, from everywhere and everything in the network.

As a set of learning and optimization tools, AI/ML can be applied to any element or function in a wireless network (see table on next page). For example, AI/ML have been used to optimize mMIMO and beamforming in the RAN, where the number of beam patterns may be on the order of 10,000, when you consider all the combinations of parameters such as tilt, azimuth, and vertical and horizontal width. The beam selection depends on network conditions, so it has to be done in real time, although the initial learning can be done offline and remotely. This is a good example of a function that can no longer be performed simply by human intervention, without automation – at least not without missing out on the performance gains from mMIMO or beamforming.

In the learning phase, an AI/ML system can learn, for example, how to optimize throughput and latency at a given location. Then it can predict the best antenna pattern given the current network conditions and the desired performance (e.g., minimize latency for URLLC, or give priority to IoT and safety applications) in real time. As it generates predictions, the system continues to learn from its own behavior (i.e., the impact of the recommended action on the network performance).

AI/ML tools can be applied to improve any performance and financial metric, such as:

- Quality of experience
- Reliability
- Resource utilization
- Throughput (and throughput density – e.g., throughput per sq km)
- Latency (overall, or per slice, or for specific applications)
- Per-valuable-bit, per-bit costs

Where does the complexity come from?
Device types: smartphones and laptops; IoT devices from sensors and actuators, to security cameras and drones
Applications and use cases: voice and best-efforts data, but also URLLC and remote-control use cases, video surveillance, security and safety
RAN: coexistence and integration across access technologies (from 2G to 5G), including unlicensed bands; mMIMO and beamforming
Core: edge and cloud computing, virtualization, network slicing
Traffic management: application-level traffic management, policy, analytics
Testing and monitoring: real-time network troubleshooting

5. The integration of AI/ML in wireless networks

As operators start using – or planning to use – AI/ML tools for learning and optimization, they need to decide how to integrate them within their networks. It is a complex decision, because it depends on the readiness of the operator, the learning and optimization goals, the financial and human resources available, and the adoption of new technologies such as 5G, mMIMO or network slicing that require smarter networks.

Initially, AI/ML are most likely to be used to optimize and automate specific network elements or functions – for example, to improve mMIMO performance. A focused application of AI/ML is more manageable, and it is easier to quantify the required effort, as well as the financial and performance benefits. It also helps both the operators and the other ecosystem players to become familiar with the new tools and build confidence in them, and to start introducing the internal cultural changes that, as we will see later, AI/ML require.

In the longer term, however, as AI/ML become more widely deployed, we expect them to inform end-to-end network optimization, in addition to (or as an alternative to) single functions and elements. At that stage, AI/ML systems will look not only at the parameters for that specific function and element, but at how they interact. Optimization of each element and function in the network may then be done not with the goal of maximizing their individual performance, but to maximize a target – e.g., QoE or end-to-end latency – that depends on multiple network elements or functions.

This increases the complexity and difficulty of the AI/ML learning and optimization process. It also requires the operator and ecosystem to have more confidence and expertise in AI/ML. This evolution toward broader, more powerful AI/ML systems will expand the benefits of AI/ML, but we are still a few years away from it.

An example of a wider AI/ML adoption is the combination of edge computing and network slicing to improve support of multiple services and use cases and, more importantly, for their coexistence in the same location. For instance, to optimize QoE, the operator has to coordinate not only network slicing and edge computing, but also backhaul and fronthaul, plus take into account RAN conditions (e.g.,

Optimization and automation of wireless networks with AI/ML: examples of use cases
On-device applications, device management
QoE
RAN planning and deployment, scheduling
Energy consumption
Access technology integration, offload, carrier aggregation
mMIMO, beamforming
Backhaul/fronthaul management, functional splits
SON, interference management
Mobility management
Security, fraud detection
Traffic management and prioritization, load balancing
Edge computing, MEC
Network slicing
VoLTE and voice
Analytics, policy, service assurance, orchestration
Customer service
Testing and monitoring, predictive diagnostics
Enterprise services, IoT
Application management

congestion and interference, resource availability, traffic load from different applications), and, if possible, get performance data from the devices. And depending on the requirements and demand from active applications, AI/ML will be able to provide recommendations on how to optimize the network at the application level.

Even further, if they prove effective, AI/ML will become fully integrated within wireless networks. The ultimate sign of success will be that we will no longer need to talk about AI or ML or deep learning, because they will be part of the operators' basic toolkit used to deploy and run wireless networks.

The use of AI/ML will also evolve in other directions. Today, in the early applications of AI/ML, learning usually takes place offline; only when the learning phase is completed and tested does the deployment in commercial networks start. This is necessary to protect network performance and to ensure that the learning has progressed in the right direction. Because the entire ecosystem, from vendors to operators, is taking its first steps, extra caution is in order.

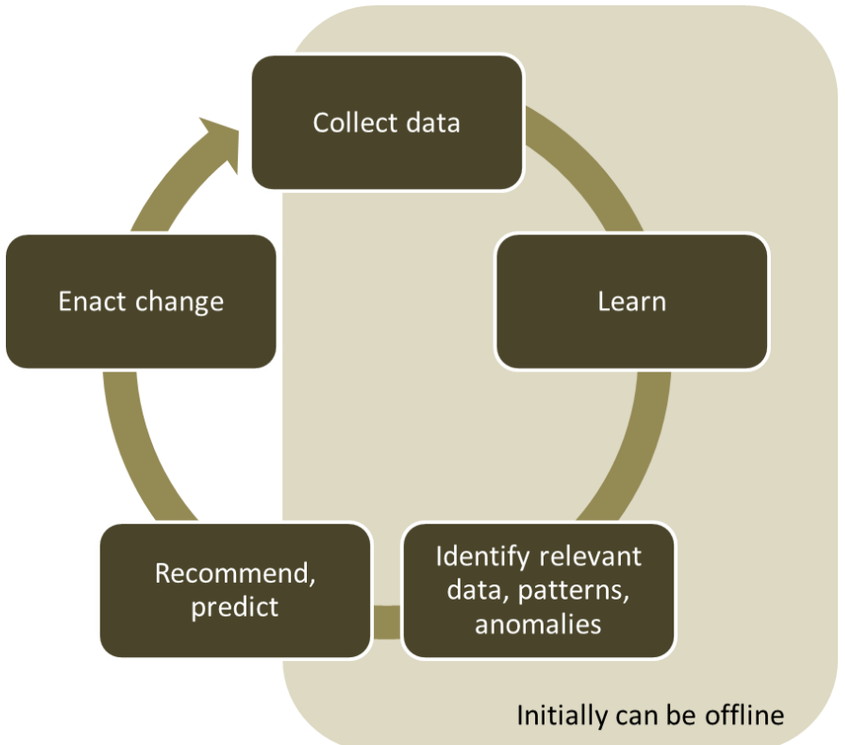
Eventually, however, we will move from today's open-loop systems to closed-loop systems in which learning and optimization will run jointly and continuously. After the initial learning, AI/ML systems will continue to refine their performance and continue to adapt to changes in the network. As wireless networks become more agile and dynamic, it is imperative for network optimization to continue after the initial learning phase, and to remain flexible and responsive to changes in network conditions.

The transition from open loop to closed loop will also bring us from supervised learning to unsupervised learning, and eventually of reinforcement learning:

- In supervised learning, the network is trained with both inputs and outputs. It learns to predict the expected output when presented with a specific input. With supervised learning, networks can learn only problems for which we know the solution. Regression models and Bayesian learning use supervised learning.
- In unsupervised learning, the network learns exclusively from the inputs, by identifying a pattern or structure in the data that enables it to reconstruct the underlying model that generated the input data. Unsupervised learning is useful to address problems that are not well understood and for which we do not have a desired output prior to the training. Clustering, principal component analysis and independent component analysis use supervised learning.

AI/ML today	AI/ML moving forward
Applications with a narrow focus	Applications with a wider focus, end-to-end use of AI/ML
Offline learning	Continuous learning, real-time operations
Open loop	Closed loop
Supervised learning	Unsupervised learning, reinforcement learning

Network optimization with AI/ML



- In reinforcement learning, the network uses agents to explore the environment defined by the input data set, and to maximize a cumulative reward. Unlike supervised and unsupervised learning, reinforcement learning does not need to reconstruct the model that generated the input data set, nor must it assume that there is one. Instead, it tries to find the optimal solution given the constraints imposed by the inputs. Reinforcement learning models the environment as a Markov Decision Process and is useful for tackling problems for which there may be no single right solution, but rather multiple optimal solutions, suited to specific conditions (e.g., optimizing performance in unpredictable or unusual network conditions).

As AI/ML get deployed beyond narrow and well-defined use cases – again such as the mMIMO use case – operators will have to deal with two new questions: What do they want to learn? What should they optimize?

In the mMIMO case, these are pretty straightforward questions: the goal may be to find the antenna pattern that maximizes the overall throughput of the cell site, although operators may have somewhat different performance thresholds, ways to define performance thresholds, or KPI prioritization. The variability in targets will translate into different models that will make different predictions or recommendations, which in turn may result in different RAN performance.

For use cases such as the edge computing and network slicing example mentioned before, the choice of what we should learn and optimize is more complex. An operator may use both edge computing and network slicing to lower latency, but it is unlikely to choose to minimize latency across the network to optimize performance. This approach would lower the average latency across the active users, but it may not provide the extremely low latency that URLL applications need, because it would treat them like non-URLL applications. The operator has to manage traffic at the application level and carefully define what priority to give different traffic flows. On the other hand, another operator may not support URLLC and would thus choose an entirely different way to manage application traffic – for instance maximizing QoE – or maybe choose not to manage traffic at the application level at all.

The decision of what the learning and optimization target should be has, in turn, implications for what is the most effective AI/ML system to use (e.g., supervised or unsupervised learning?), how it should be designed, what it will learn, and what recommendations it will generate.

Automation

It is not by accident that AI/ML and automation have become hot topics at about the same time. They all address in a fundamental way – i.e., changing the way we run and optimize wireless networks – the complexity we are starting to see in 4G networks and that will grow more in 5G networks and with the adoption of IoT.

AI/ML and automation are complementary and, to some extent, inter-dependent. It is difficult to see how AI/ML could succeed without a high level of automation in the network, especially as we move to real-time, closed-loop learning and optimization. At the same time, automation of complex, dynamic networks is not going to be effective if we continue to use traditional, deterministic optimization processes that have limited effectiveness.

AI/ML and automation enable and reinforce each other, to the point that they have started to converge to the same platforms and be used for the same use cases.

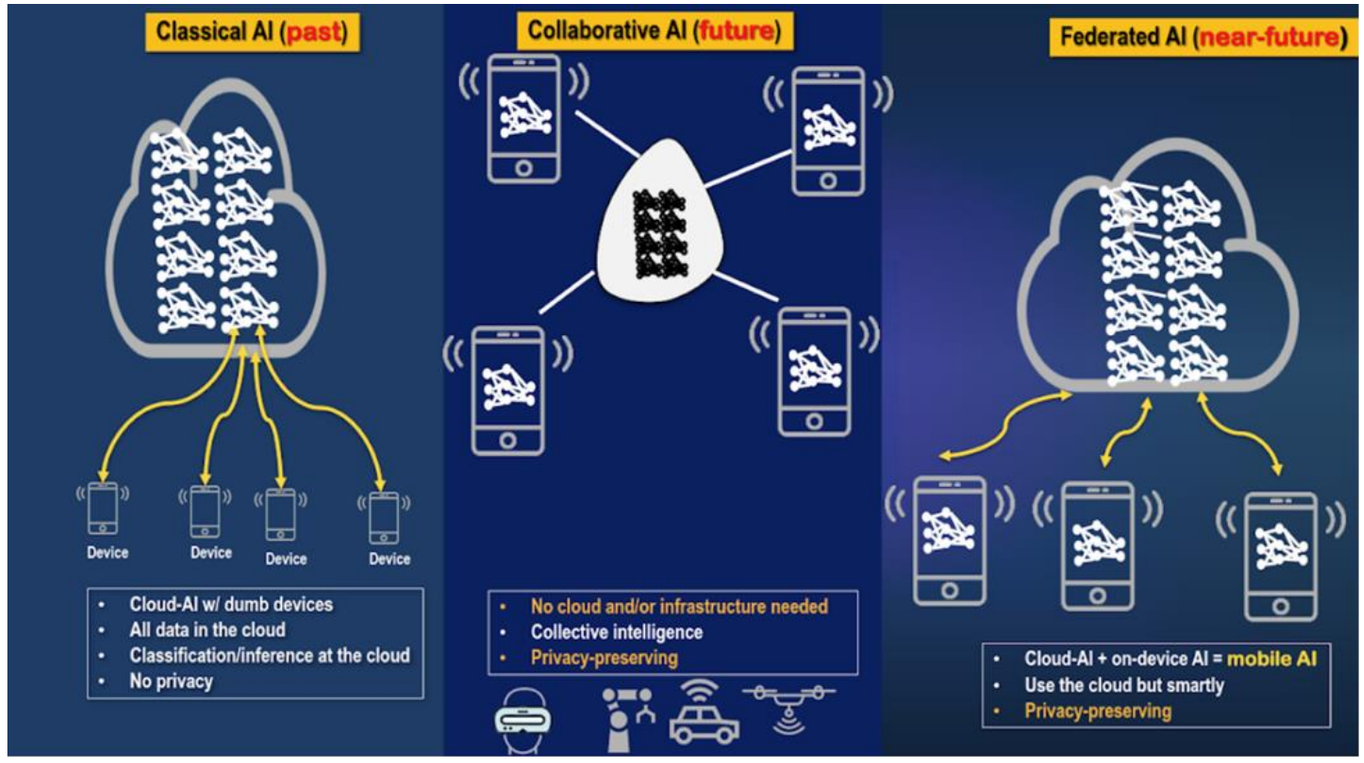
6. What is the best place to learn in wireless networks?

Virtualization, edge computing and, eventually, 5G have a transformative effect on the architecture and topology of wireless networks. Monolithic, centralized, static networks are giving way to a new breed of agile, dynamic, distributed networks, which operators can much more flexibly shape to meet their needs and preferences. The distinction between a distributed RAN and a centralized core is fading away, with hardware and functionality becoming more distributed across the network.

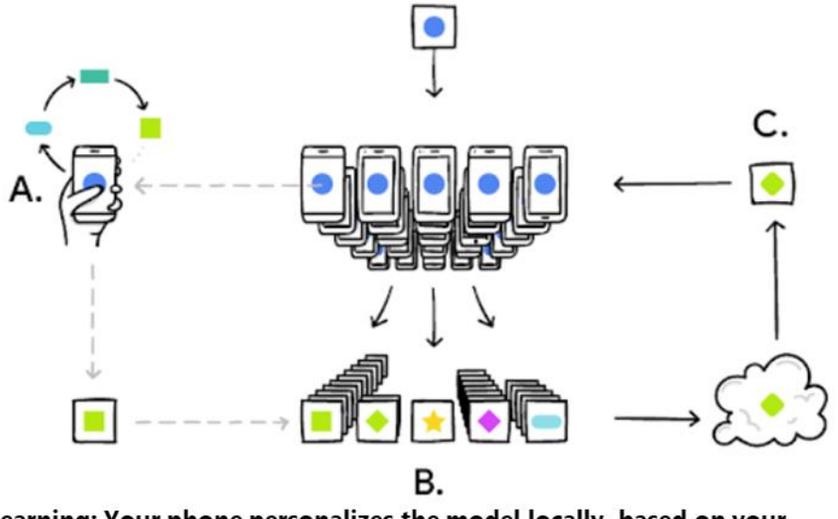
As operators choose what to do where in the network, they also have to choose what the best place to learn and to optimize network performance is. Initial AI/ML systems are centralized and often hosted in the cloud. Because today most of the learning is done offline and is supervised, location does not have a large impact and the choice of location mostly depends on cost and availability. With the move to unsupervised closed-loop learning and real-time automation, location will become prominent, especially for traffic that requires low latency and for tasks that depend on location (e.g., managing interference, offload, or congestion).

The choice of location depends on multiple factors. The primary one is the use case (e.g., an edge location is valuable for SON, a centralized location is better for customer support) and its requirements (e.g., latency-sensitive use cases benefit from edge processing). But cost, availability, operational constraints, and security are other factors that have to be taken into account. In a virtualized, dynamic network, location of functionality can change through time, and this gives operators the flexibility to adapt to changes in network capabilities, conditions and demand.

The ability to choose where learning and optimization take place introduces a big change over today's AI/ML systems and greatly expands their potential. Work is underway to explore how to deploy edge AI/ML while preserving the scalability, reliability and security of centralized AI/ML and without imposing overhead or duplication in the networks. As in general for edge computing, we are still



Different approaches of AI to wireless networks. ***Source: Mehdi Bennis***



Federated learning: Your phone personalizes the model locally, based on your usage (A). Many users' updates are aggregated (B) to form a consensus change (C) to the shared model, after which the procedure is repeated. ***Source: Google***

working on what the optimal edge location is and what the requirements, performance and costs tradeoffs are for different use cases: these are new and complex issues that will be gradually resolved as we learn from trials and early deployments.

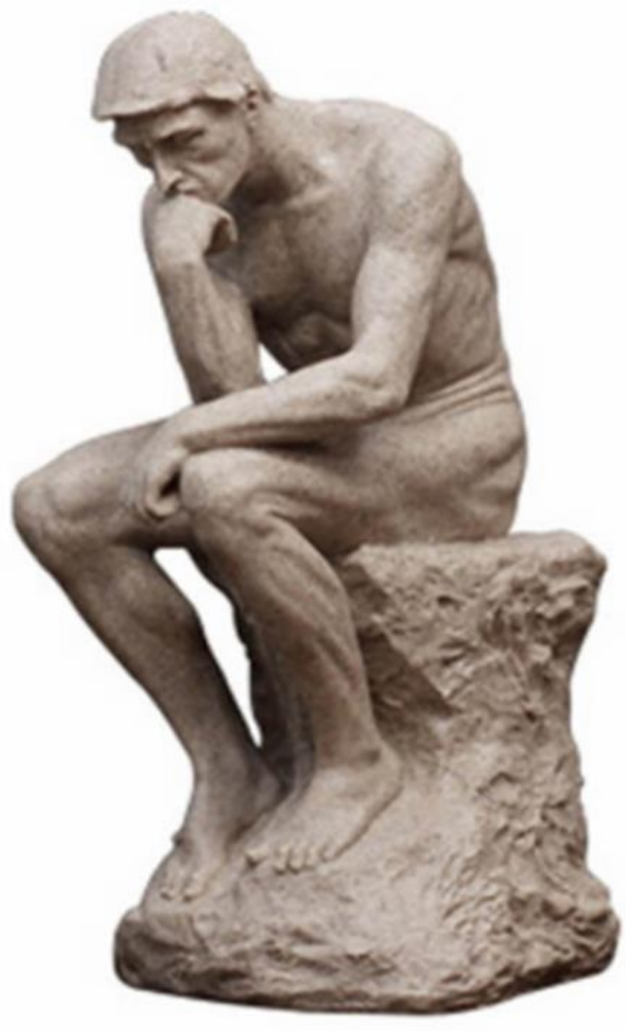
Edge AI/ML is not the last frontier: beyond the edge, on-device AI/ML is gaining traction. With the increase in number and diversity of devices, end-to-end network optimization can benefit greatly from the contribution of devices, and devices can become smarter and more efficient by participating in the optimization of wireless networks. For instance, QoE can be inferred by the network, but it is defined at the device end: devices can have a big role in guiding the network to maximize QoE. The initial learning can be done centrally because it is not specific to the individual device, but eventually, stripped-down AI/ML networks have to move to the devices. Cloud AI/ML, edge AI/ML and on-device AI/ML will not only exist side by side, but will work in concert and reinforce each other to establish a collaborative AI or federated AI environment where intelligence is distributed across the network and, eventually, there is no need for a centralized control infrastructure.

7. Culture, expectations and fears

The biggest challenge to the adoption of AI/ML may be tied not to the technology, but to the cultural change required to succeed, and to the need to move past the combination of hype and fear that AI/ML elicit, in wireless and in society at large.

AI/ML take us away from a deterministic, rule-based and static approach to deploying and operating wireless networks. While that approach is becoming obsolete, inadequate as it is for dealing with the network transformation underway, it has served the wireless industry well until now. It is an approach that both operators and vendors know well and trust, and that gives them a high level of control over the network infrastructure. The adoption of AI/ML requires operators, along with the rest of the wireless ecosystem, to trust a non-deterministic environment in which statistical and stochastic methods complement – and eventually may overshadow – more-traditional logic, a knowledge-based system, and expert insight.

The first cultural challenge is to understand how to use AI/ML in wireless and avoid the magic-box approach that some of the hype is encouraging. It may be tempting to see AI/ML as a cure-all, off-the-shelf solution that can effortlessly be applied across the board, but acting on this impression will likely backfire and impede progress, as discussed earlier in the report.



Rodin, The thinker

A second, equally important challenge is to become comfortable with a learning and optimization process that is going to be less smooth than we are accustomed to with today's deterministic models. With AI/ML, learning becomes a continuous process with incremental but non-linear progress that is likely to hit local minima, where performance may get locally worse before it continues to improve.

A third challenge comes from the black-box perception of AI/ML networks, which comb through large data sets and generate predictions or recommendations without following the logical path human observers are accustomed to.

An operator or vendor needs to have sufficient confidence in the system to accept the higher variability and lower transparency of AI/ML, with the anticipation that the outcome will be superior to that of a rule-based, deterministic network. To achieve this level of confidence, most operators and vendors choose the gradual approach described earlier in the report: start with smaller AI/ML systems with supervised learning, and gradually expand to more ambitious ones.

At the same time, we should not underestimate the role that humans retain in defining, overseeing and providing the essential contribution that only experience can provide. Throwing an AI/ML system at a problem without human guidance is a losing proposition that stems from unrealistic expectations of what the technology can do.

These challenges are difficult to face in the current wireless ecosystem because the assessment of network performance and the rewards to the people running the networks are based on the established deterministic, command-and-control framework.

Moving to AI/ML is not going to replace humans (at least not in the foreseeable future), but it is going to unfairly expose them to uncertainty and risk in their jobs unless it is accompanied by a cultural change in the organization and a new way to assess human performance.

At the same time, the skills that employees need will change. Although workers will be relieved of boring and repetitive tasks, they will have to oversee and refine AI/ML networks. To do so, they will have to shift from being able to fix problems to being able to predict them and prevent them from arising. This will require new skills, and – as with cultural change – top-down, organization-wide changes are crucial to support the acquisition of new skills among the workforce.

8. Takeaways: How will AI/ML change wireless networks?

Initially, AI/ML's role in improving network performance will be narrowly focused on well-contained use cases, which are more manageable, require less expertise, and promise easily quantifiable financial and performance benefits. In the long term, however, AI/ML will have a much wider and transformative impact on the end-to-end network, transcending the technological change and sweeping in an ecosystem-wide cultural change.

While new technologies such as 5G NR, URLLC or mMIMO have a well-defined role in the network transformation, and AI/ML brings an essential contribution to a horizontal paradigm shift across the entire wireless infrastructure and all the ecosystem participants introduced by virtualization, distributed architectures and automation.

The convergence of these new approaches to deploying and operating wireless networks is not accidental: all are empowered by the recent availability of affordable, reliable and secure computational power; all are driven by the challenges of increased network complexity; and all aim at improving both the utilization of network resources and the user experience. In turn, they will jointly make wireless networks more dynamic, more agile and more capable of supporting new use cases.

This shift, however, will require time and hard work. The performance impact will be incremental and, because it affects the entire network, it will be difficult to quantify with the KPIs and TCO models we use today.

As we change how we run wireless networks, we also need to change how we assess their performance and value. For instance, static, network-wide KPIs are no longer sufficient: they need to be complemented by performance metrics that are monitored in real time and at the application level. Similarly, tightly focused monetization efforts and cost savings may divert attention from the longer-term, but more substantial and sustainable, financial benefits that the horizontal paradigm shift can bring with a smarter utilization of network resources and more efficient revenue generation.

Do the hard work first	Reap the benefits
Leverage complexity as an opportunity to optimize the network	Shift to more agile and dynamic networks that autonomously learn to optimize their performance
Start on simple tasks, gain confidence	Use network resources more efficiently
View learning as a continuous process in a dynamic system	Optimize networks in real time
Carefully choose where to use AI/ML – and where to avoid them	Unlock the value of 5G innovation
Tread lightly, keep expectations realistic	Work closely with enterprises to support their AI/ML efforts
Get used to being outside the comfort zone	Improve support for IoT
Spearhead cultural change	Strengthen security
Adopt a new approach to monitoring network performance	Generate revenues more efficiently

Conversations

Intel | Establishing the AI and ML ecosystem

A conversation with Caroline Chan, VP and GM, Network Business Incubator in the Data Center Group, Intel

For AI and ML to succeed, we need a strong ecosystem that extends beyond operators and vendors to include application developers, enterprises, venue owners and smart cities.

In this conversation with Caroline Chan, VP and GM Network Business Incubator in the Data Center Group at Intel, we talked about how this ecosystem is developing and about taking AI and ML beyond proof of concept to extract tangible benefits.

Monica: Caroline, AI has become a hot topic. Why?

Caroline: AI and machine learning have been around for a long time. On the network side, it has become a hot topic today, and we have started seeing the proliferation of AI because we have made dramatic improvements in processing power.

Intel has invested in making this possible and developing an ecosystem around AI. Finally, we have brought down what it takes to process AI and machine-learning algorithms, and we put accelerators in our roadmap that will also help adoption of AI.

The increased processing power and pricing efficiency are more favorable for the adoption of AI. In the future, AI and machine learning will start to pop up more in network discussions.

Monica: With this additional power, operators can use their vast amounts of data to optimize network performance, using tools such as AI and ML. How quick will the change be?

Caroline: The speed of change is pretty dramatic. When I started talking about 5G with you and with our customers, people did not talk much about AI. And people still are working on the NFV/SDN – a topic you and I have talked about quite a few times.

We've been focusing mostly on NFV, SDN and MEC, and the combination of them. Everybody is starting to realize that, as we move to 5G, the usage model has dramatically grown beyond managing voice and internet browsing. Now we have applications in different verticals and with different enterprises. I may spend all day in a workshop with our industrial automation group in our IoT division and with our reach-out group. You start realizing the network has dramatically improved in terms of how we deal with the demand and how we work with the output.

The amount of data we need to extract and analyze has dramatically increased from a few years ago.

It's almost a perfect storm, or maybe a perfect intersection, of more data and more compute on the network. Virtualization has started to resemble more to a cloud and a data center. The edge data center has become reality, and it is no longer just a PowerPoint slide.

As a result, the use of AI for network optimization has increased, along with data analytics, to serve the end user better – whether it's a consumer, a hotel or an industrial plant.

Monica: Absolutely. It is the convergence of multiple technology drivers that makes this possible. There are many more dimensions along which you can optimize. If you only have

voice, there's not a whole lot of optimization you can do. You either have the capacity or not.

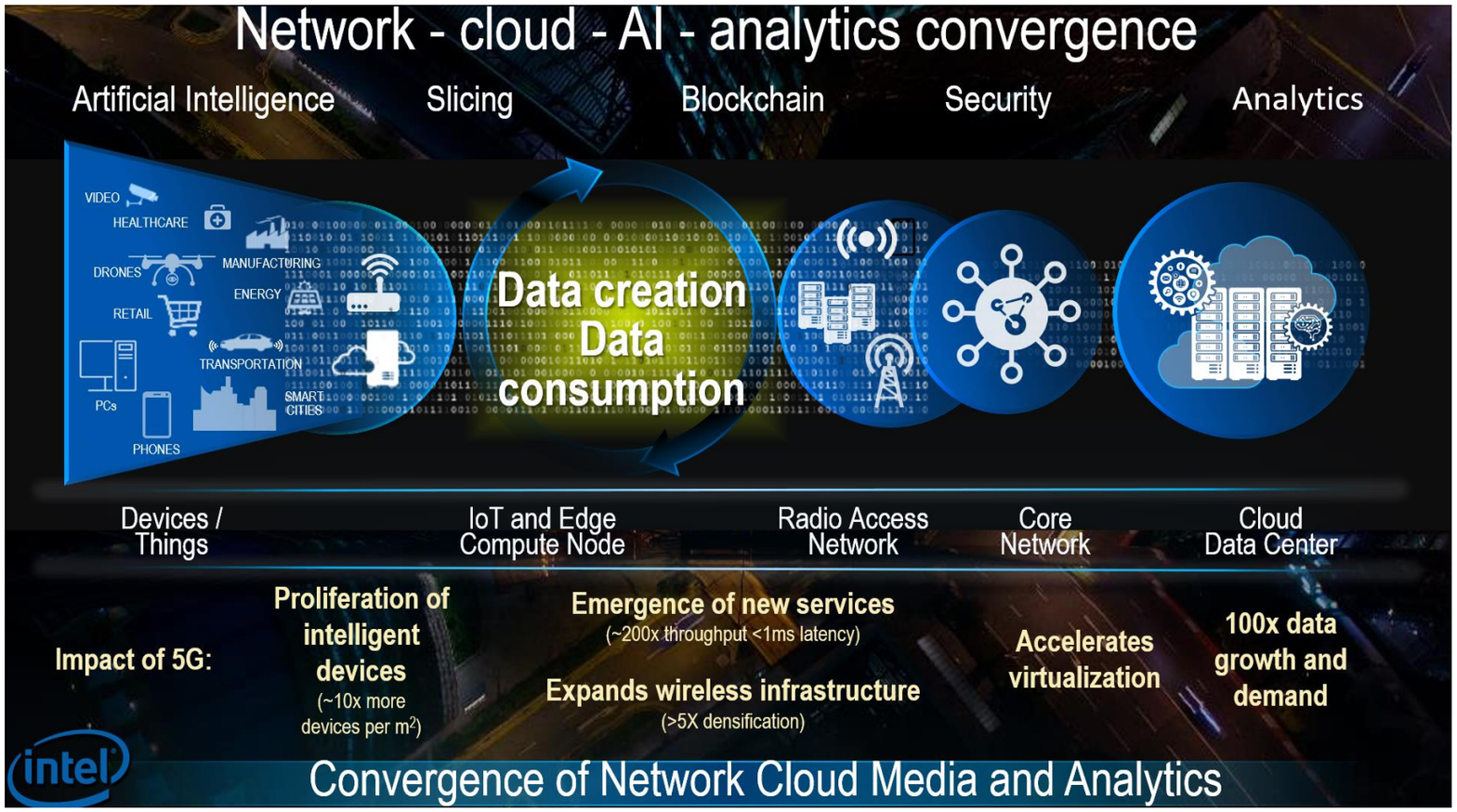
With MEC, edge computing and network slicing, you gain more opportunities to optimize network performance.

Caroline: I'll give an example.

We've been working in different aspects of incubating different use cases to utilize AI. When we looked at using the network serving beyond the B2C and B2B, we had many conversations with 5G-ACIA, which focuses on industrial automation. We were discussing the needs and requirements with European industrialists. For us, it was natural to say, "Hey, we can tune our network to your needs."

Factory owners would come back to us and tell us that a network outage does not mean only that people cannot make calls. There are more stringent requirements that are not met during an outage. The production line or the robots may stop functioning. You are talking about a lot more dollars and cents. Maybe even safety concerns. It is not just, "Oh, you have the network outage." There's a liability discussion.

We're supporting industrial applications with end-to-end network slicing. People may say: "You're just simply slicing the core network or the radio network." But it's a lot more than that. It's stringing together a slice or maybe even a particular tube, and pairing it with the right type of spectrum, pairing the right outflow from the device to the network to serve particular use cases.



The convergence of cloud, AI and analytics in wireless networks with 5G ***Source: Intel***

In partnership with our IoT group, we have a couple of projects with some leading industry partners. These projects aim at taking some of the control elements of a factory and putting them on a 5G network that uses edge computing. When you have all of that together, you are not talking just about a regular network, the network of the old days.

You have to apply some of the learnings and the network needs to be smarter. It's way beyond just adding another channel card to meet the capacity requirements. It's really about putting the right processing power to the right latency and SLA requirements – and maybe even the right spectrum. You need to have the right connectivity onto a particular use case that happens to be in demand at that moment.

You hear a lot about X as a service: function as a service, infrastructure as a service, platform as a service. The management of the network is much more complex. It's beyond having an engineer monitoring the capacity and usage in a network.

Monica: As you say, you need to optimize the network to manage performance and support specific use cases. The optimization revolves

around both the wireless network and the enterprise applications, at the same time. And we use AI on both sides.

Caroline: My first job out of college was working for a major telco equipment manufacturer. My job was to monitor the network of one of its customers. I was looking at the alarms and making decisions about whether we should dispatch a new capacity channel card. I had to determine what to order, send the order to a warehouse and dispatch it to the customer side.

Some of my decisions were based on someone who was mentoring me and showing me what to do. But other decisions were judgment calls. For instance, there were alarms, such as 65% usage of the system that prompted the addition of more capacity.

It was mono-dimensional, and it relied a lot on human intelligence, or rather experience.

With 4G, MEC and eventually 5G we will have multi-dimensional networks which are more complex. Humans are not really good at dealing with the multi-dimensional problems. That's what we are facing now with 4G and MEC, and what we will face with 5G and MEC moving forward – and also what we are facing with different types of applications already.

It's not just entertainment. It's not just having a Netflix running. It's about dealing with hospitals, telemedicine, V2X and factories. The dimensions of the inputs coming in are going to be so much bigger than when I was sitting in front of a screen, monitoring this network. I just can't fathom how a human can do this. You have to add machine learning to this, and you have to run in some neural network to read all the inputs. It should be a closed-loop activity.

I'm excited to be in this arena facing this dilemma, facing lots of issues, and I believe we will be able to solve them with AI, with machine learning, with MEC, and with the processing power that we have today.

Monica: It's all very exciting and a great opportunity, but there is also a need. Network optimization is not an optional thing: it is crucial to making sure all the applications – for instance, healthcare applications – are secure and safe.

Caroline: Yes, you are absolutely right. At the very beginning, when we first started talking about MEC, a major hotel chain which owns a large casino asked us to help them. Typically, casinos have frequent-shopper or reward programs. Casinos and hotels know a lot about the people who stay there. For instance, I'm a frequent guest at Marriott properties, so they know a lot about me, they know my preferences.

They want to apply that personalized experience for their guests, and they want to achieve that because it creates a better user experience as you stay at the hotel.

For example, you may have guests going to different concerts – one may be country music, another one hip hop. People attending these concerts may have different spending habits and preferences. As they come out of the concert and go back to the hotel, the hotel wants to treat them in different ways – e.g., send them different offers or content. The hotel wants to give them a much more personalized experience, and it can do so with edge computing.

At the time the major hotel chain approached us, we were at an early stage of edge computing. We did not pursue that request, but today we would take on that sort of project, because we have all this information at our fingertips.

As long as you have that information, you can act upon it: you know the guests in your hotel because they are on your network, and you can match their SIM card to their preferences. How can you deliver a better personalized experience to them? That is a very hot topic today.

I've been talking with different hospitality owners about that. How do we learn to do this? How do we know that a subscriber is in the network? The operator knows who you are. How do you apply that knowledge to enrich a venue so that they personalize the treatment for their guests?

Monica: There are many opportunities, but you need an ecosystem that includes more and extends beyond the operators and the vendors. What kind of ecosystem do we need? And how do we create that?

Caroline: The operators have started recognizing that the partnerships they need

span way beyond the traditional ones they have. This is also true for device and infrastructure vendors.

Within Intel, we built what we call Network Builders, and now we have extended it to Edge Builders. This is a partnership with long-time developers who write code and applications for verticals. As part of the infrastructure ecosystem, our job is to provide a horizontal platform that has virtualized capabilities, similar to those we provide today on the cloud side. Cloud providers provide a very efficient way to secure onboard devices and applications, allowing a flexible use of compute function and service.

We need to create the same type of capabilities through the network ecosystem so we can leverage all the things that 5G promises.

Monica: Security is another critical aspect as you get a bigger ecosystem and more players get involved.

Caroline: Very much so. Cloud security is always at the top of mind. We have goals for different security checkpoint information. Especially, enterprises have proprietary information that they want to hold on-site and they don't necessarily want to have in the cloud or in the network.

Many enterprises still want a private cloud. We are looking at applying network security using a blockchain-like technology applied to the network. We're looking at secure onboard of devices and applications. We are working on isolation of applications so that if one application goes wrong, it does not impact the rest of the network or other enterprises.

All of that comes into play. For many of the vertical use cases, security is the number one requirement. If we don't solve this, we can't get into the next step. AI and machine learning are beneficial, but security is the fundamental that we'll have to provide.

Monica: A few years ago, both operators and enterprises would say they would not know whether to trust AI. Has that changed? Who is moving ahead more aggressively, the enterprises or the operators?

Caroline: They are both moving forward. We've been talking with operators about using AI and machine learning to improve network efficiency and serve customers better.

Enterprises are keen to use AI and machine learning to gather more data and run real-time data analytics. They can use data analytics both in the short term and in the longer term to improve their operational efficiency.

Both operators and enterprises are marching ahead. Especially with 5G, operators recognize that they need to use AI for efficient network management.

Monica: In terms of challenges, what might slow down the adoption of AI?

Caroline: Regulations that limit the ability to acquire data and to learn from it may slow down adoption. That's why, when we look at regional differences, we see China a bit more advanced, because there's less awareness of data privacy there.

In China, enterprises and operators have access to a massive amount of data. There are smart cameras everywhere. It's about learning. AI can learn from this data and improve the network, using a closed-loop process. Major cloud providers like Alibaba, Baidu and Tencent are focusing on acquiring data and using it to improve operations. They are all using data to improve how they serve the customers.

From a global perspective, I've seen a lot more improvement in the advancement in China because of regulatory efforts. The Western world is also working on regulations, but most of the work is still in an incipient stage.

Monica: Can you give us an example of initiatives you are working on in China?

Caroline: A couple of months ago, we announced a collaborative smart transportation initiative with Alibaba on a smart roadside in China. But it is not just in China. Many car companies and their partners are involved.

Roadside units today do not have a lot of compute to advance and to provide an architecture where you don't just take the data, you also analyze it. Data such as live updates and camera information can also be used to improve car safety.

Knowing where the roads are being repaired, or maybe knowing where the roads need

repair, and having traffic congestion data can help smart city initiatives to manage traffic pattern flows.

In China, road congestion and pollution are horrible issues. The government has rolled out initiatives to resolve these issues and help improve road safety, as well as to make cities more intelligent.

We're involved in the early proof of concept phase of these projects. We wanted to take the compute power and use it to make V2X more rewarding. It is more than just saying, "Hey, I'm here. Here's a pole to hang the equipment on." We also want to collaborate to collect, analyze and ingest the data, and to come back with a better service and management of road traffic.

Monica: We've been talking about AI in the network. What about the devices? Do you see any role for AI in the devices to drive end-to-end network optimization?

Caroline: I believe so. Our device team is looking at using some of the AI learnings to understand when the devices can offload some workload onto the network.

I expect that with AI we will make as much progress on smart offload as we made with MEC, especially in gaming and in sophisticated AR and VR. We want to understand how to offload in a way that the device continues to be very reasonably priced and to have low power consumption, yet leverages all these cloudlets or the edge cloud near you. Those are interesting topics that I hope the ecosystem will focus on.

Monica: Let me ask you a final question on 5G. You said 5G and AI need each other. Do you see people waiting for 5G before they roll out AI, or vice versa?

Caroline: I don't think we're waiting for 5G to do AI. 5G does demand AI, and 5G enhances the performance of AI. But we are already doing that today with a combination of LTE and MEC. We are involved in some of the projects already.

I don't think people are waiting, but 5G and AI definitely go hand in hand and will benefit each other.

About Intel

Intel expands the boundaries of technology to make the most amazing experiences possible. Information about Intel can be found at newsroom.intel.com and intel.com. Intel and the Intel logo are trademarks of Intel Corporation in the United States and other countries. *Other names and brands may be claimed as the property of others.

About Caroline Chan

Caroline Y. Chan is VP and GM, Network Business Incubator in the Data Center Group, at Intel Corporation. She has overall responsibility for Intel's global network infrastructure strategy and solution development related to 5th-generation wireless technology. Chan and her team identify and develop use cases that incorporate the Internet of Things, innovation in wireless technologies and deployment models, such as mobile edge computing and alternative spectrums that will enable new service providers and enterprise networks.

Nokia | Getting machine learning off the ground

A conversation with Tero Rissa, Chief Architect, Machine Learning, Nokia Mobile Networks

Machine learning promises to improve the performance and cost efficiency of wireless networks. But it is also a hyped hot topic these days, with some expecting machine learning to solve all problems, and others fearing the inscrutable black box that makes it so powerful.

In this conversation with Tero Rissa, Chief Architect, Machine Learning at Nokia Mobile Networks, we talked about how machine learning can optimize performance and investment, and about what we can realistically expect machine learning to deliver.

Monica: Tero, what are you personally working on at Nokia in the AI and ML area?

Tero: I'm leading the AIML trials at Nokia Mobile Networks. We coordinate the machine learning activities across the organization. We have over 20,000 people in Nokia Mobile Networks, and we are ramping up in machine learning at an accelerated pace.

I have four main work streams.

The first is the AIML expertise development. This is a new corporate-wide skill set and mindset that we have in the organization. We have a machine learning lead in about 200 internal organizations at Nokia Mobile Networks. We coordinate the activities in these organizations through the machine learning center of excellence that I'm governing. We also develop the training material and curriculum for the whole organization.

Second, we are responsible for the AIML infrastructure, management and availability – from evaluating different kinds of acceleration platforms to hosting the acceleration learning services and the capacity for Nokia.

Third, we are responsible for the AIML data strategy for the radio access network data across the business units. Data strategies define how to capture, store, and retrieve the data that is required. I have to say, it's much more complex under the hood than that.

Last, but definitely not least, we coordinate AIML use cases, development, and engagements with operators. Doing these joint trials with operators, we develop new business cases.

Monica: It requires a big cultural change within the organization itself. It's not just a new product. It's a pervasive change that you have across the company.

Tero: That's absolutely right. When I do these trainings, for example, I often say that the hardest part to understand or learn in machine learning is the actual mindset – what machine learning can solve, how it can solve it, and what it means when you do something with machine learning.

The hardest part is not deriving the partial derivative in gradient descent in backpropagation. It is to understand what you can do and how you can approach this systematically, and how you can gather the data and the training material for the machine learning algorithms.

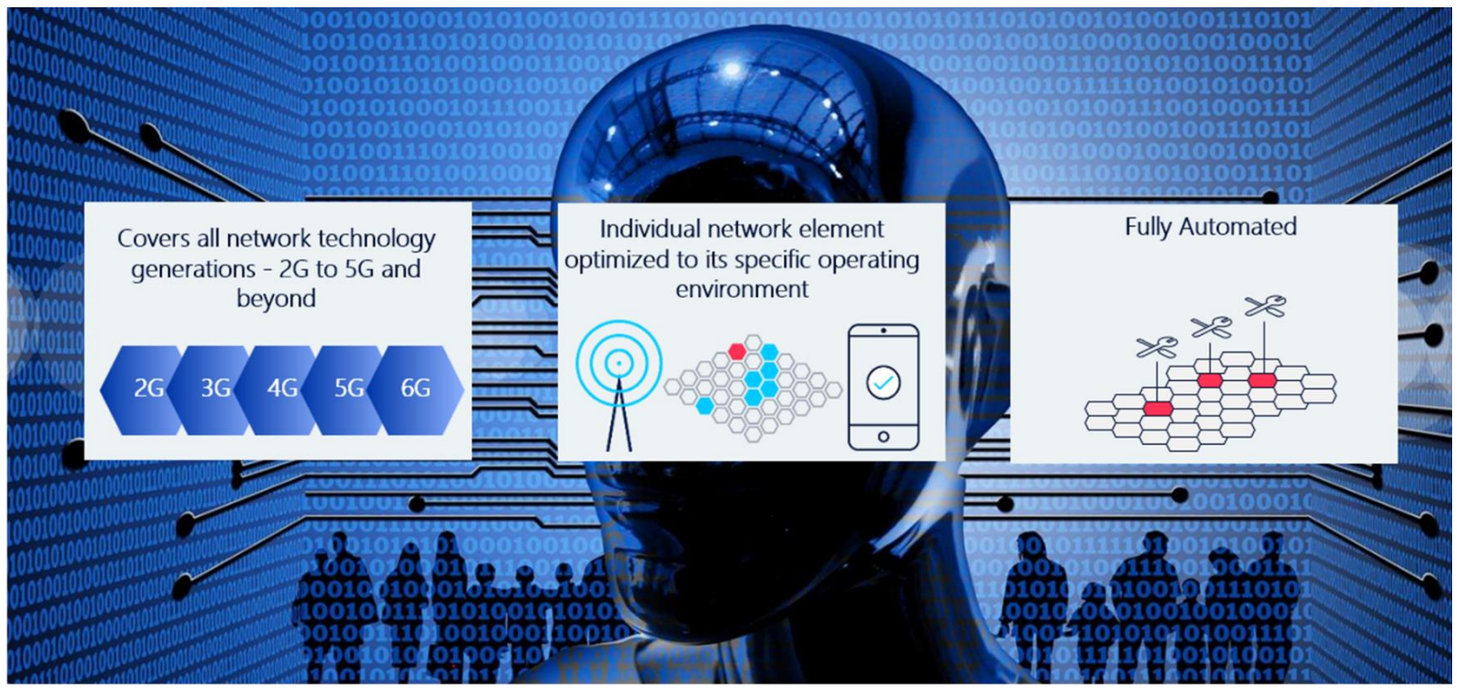
Monica: AI and machine learning have different meanings to different people. What is the difference between AI and machine learning in your view?

Tero: I try to use the term "machine learning" as often as possible. It is a reasonably useful term that can be used to refer to certain

algorithms and deep learning methods that are commonly understood, such as supervised learning, reinforcement learning, and unsupervised learning.

AI is used to refer to these as well, but it is also often used for something that I don't believe exists: general machine intelligence, self-aware systems, superhuman intelligence, or whatever that could be. I really try to move the lingo from "artificial intelligence" to "machine learning," because this is something much more tangible.

There is a lot of hype. And because of the hype and the lack of understanding of machine learning or AI, people incorrectly assume that it is magic.



Nokia's vision for AI-powered networks ***Source: Nokia***

They may say, "I have this problem that I can't even understand, could you use AI to solve it? My life would become easier." AI is definitely not that. My common answer is that automation could solve the problem. And machine learning is the automation of decisions.

Human power can't be used to make millions of consistent decisions in real-time. We can use machines to make those decisions in an automated fashion. This narrows down the scope of AI to a point that it helps people understand the problem.

Monica: It is refreshing to hear this. We don't want our expectations to be so high that we fail because we try to do more than we realistically can. AI and ML are not magic, they are hard work.

Tero: Actually, the title one of my training materials is "Machine learning is not magic – Rational and scientific application of ML in an industrial context."

Monica: There is definitely hype now, but why now? Why not ten years ago?

Tero: There was hype ten years ago. At that time, I was actually doing machine learning with artificial neural networks, as they were called at that point. Back then, I decided that AI was not going anywhere, and changed direction in my career. But I have come back. The question is why now?

Three distinct things happened in the last ten years.

The first one is algorithms that are the solid foundation of deep supervised learning. For example, ReLUs and stochastic gradient descent have become more practical. Although they are very simple mathematical applied methods, they enable much deeper networks to be trained.

On the academic front, we now have vast amount of research compared what was there ten years ago. So, if you want to wander further from supervised learning, there's a really good academic community, and there is a lot we can pull from this research – more than any single institution could invest into research by itself.

The frameworks have improved vastly. If you had to do something with artificial intelligence

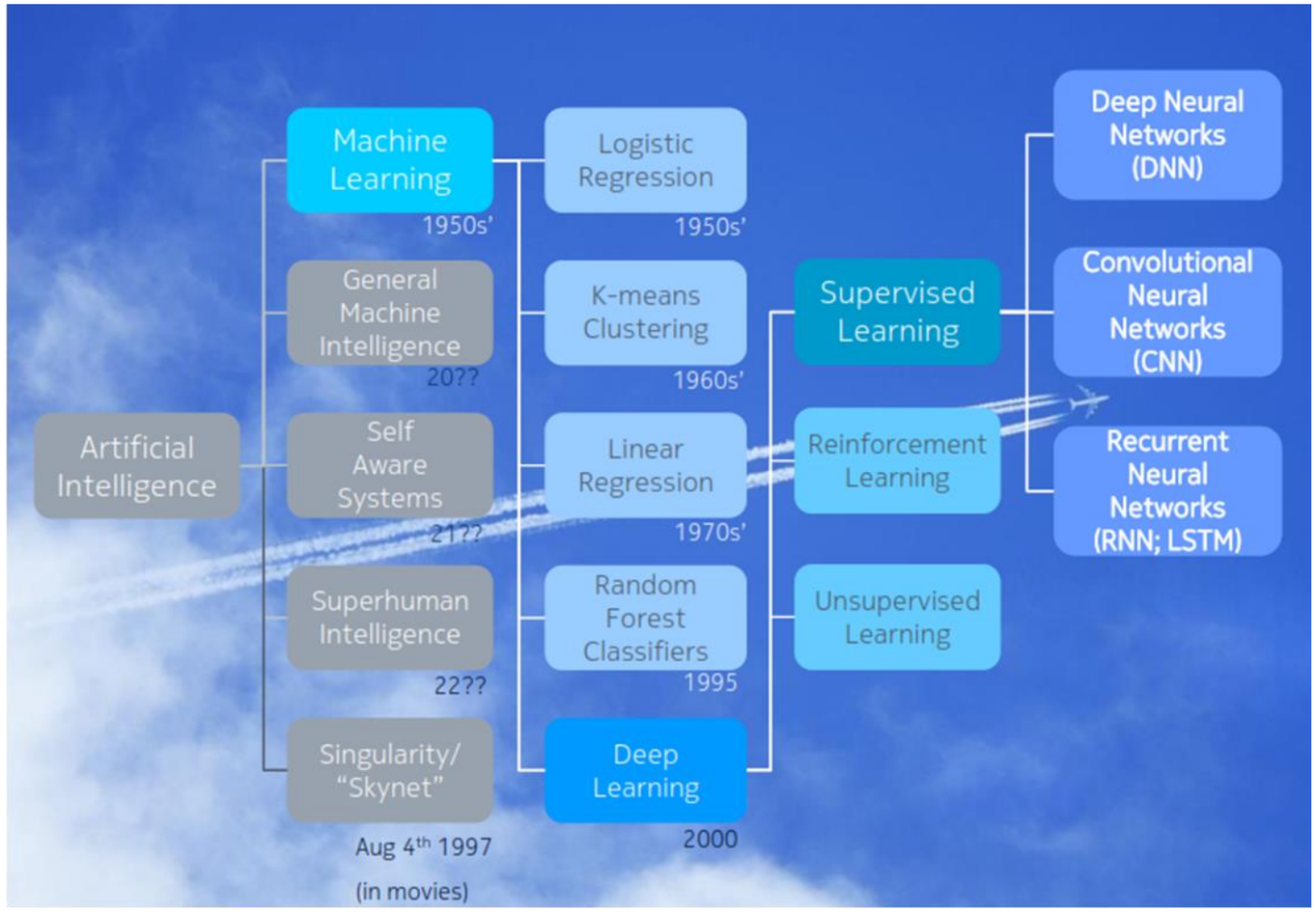
on neural networks ten years ago, it would have taken several months to achieve the same results that you can have today in an afternoon while you have two coffee breaks.

There are also frameworks such as Keras, TensorFLow and PyTorch, utilities like Docker, and acceleration libraries from the vendors that have dramatically increased productivity. You can do much more. You can achieve much more in a shorter time.

But the final nail in the coffin was the computational capacity. I was doing mobile phone image processing at the time. The performance and memory requirements of mobile devices were off by 10,000 times. Now we have moved from a few giga operations per second to hundreds of tera operations per second. Most importantly, we have moved to an energy efficiency of multiple tera OPS per watt.

Also, what is really important and a big part of the work we do here is the reduced bit-width architectures that are dedicated for machine learning. So far, machine learning has been done with devices that are capable of matrix multiplication in accelerated fashion, but now we are seeing an emergence of devices with special hardware architectures that address more specifically the needs of machine learning.

Monica: What should we realistically expect from machine learning? What are realistic goals that we can achieve?



AI terminology ***Source: Nokia***

Tero: We have a pragmatic approach to machine learning in industry, but not everyone agrees with it.

I believe that every unit's first machine learning project should be about supervised learning. It restricts the scope a lot, and it's something that we have had heated debate on.

For instance, with reinforcement learning, you wouldn't need much training data, but it's a much harder paradigm. And it's much more fragile.

If you get supervised learning working in your system, you also have to bite the bullet and acquire the data to train your networks. You have to do the legwork.

Our approach is to initially deploy supervised machine learning systems that are very practical. They make a meaningful difference, and they improve the system, but they don't

change the whole system. They don't take over the world.

We first do a simple and manageably sized task with machine learning. It does not matter so much what the application or the use case is, as long as it improves the system and we can show a cost-benefit. There are so many other things that you have to take into account, including the actual machine learning algorithm.

If you start with too big of a target, the whole thing becomes too big. If you start with something that is simple and a slam dunk from a machine learning point of view, you may very well find out that there are other very demanding things you need to do, such as software integration, data acquisition, data engineering, and the infrastructure.

The best approach is to make sure your first machine learning project is successful, not that it would change the world.

Monica: Networks are becoming complex. How do you deal with the increase in complexity? Is the complexity going to be a challenge?

Tero: No. They go hand in hand. We have to cover all the technology generations. 2G is still there, and we are introducing and deploying now 5G. New technologies emerge in a surprisingly small amount of time. We may have 6G in ten years. Each generation is more complex than the previous.

True complexity actually arises from the requirement that we have to support multiple generations at the same time. We have to have all the Gs. The different configuration parameters on these systems alone are just becoming unreal for humans to manage, so we need machine learning.

We need to aspire to full automation of the networks. The main business benefit for the networks comes from the operational cost of devices and networks. With machine learning, we can optimize and customize individual network elements to be specific to the operational environment.

This is also sometimes misunderstood. Machine learning doesn't bring optimization and customization automatically. You have to work for it.

The environment itself can be used to train the network to be specific to the operational environment where the actual radio element is situated.

We can achieve significant benefits in the network. We can have a lot of leverage. From a business point of view, the optimization of operational expenses through advanced automation is one of the main things we can achieve.

In addition, we can deliver a better user experience in a cost-effective manner. We reduce cost, but we also increase the benefit and deploy the radio spectrum in a more efficient way. And we can also have energy savings, increased security, and resilience for intrusions.

Monica: You can optimize the network in real time, in a location-specific way. You don't just learn something for the entire network. You learn something for a specific location, and that allows you to dynamically react to local changes in network conditions and demand.

Within the network, where does learning take place? Is it going to be in a centralized location or in a distributed location, at the edge? And if it's towards the edge, where is the edge?

Tero: It's a good question, but we don't have a final and ultimate answer yet. At the beginning, we will use centralized training because we know that this approach works.

We will start from the training. This is also because we do require different hardware acceleration for training and for inference. Even if you do the training in the base stations, we still need different hardware for that. Distributed training just wouldn't make sense, and it wouldn't be cost effective.

In some cases, we can do millions of inferences per second. But the training would rarely be more frequent than a second, and even a day cycle is very common. The utilization rate is vastly different for training and inference.

In inference, the non-floating-point architectures will dominate – or at least we will require them. There's a hundred-times difference between an 8-bit add and a

floating-point add, for example, in the silicon real estate.

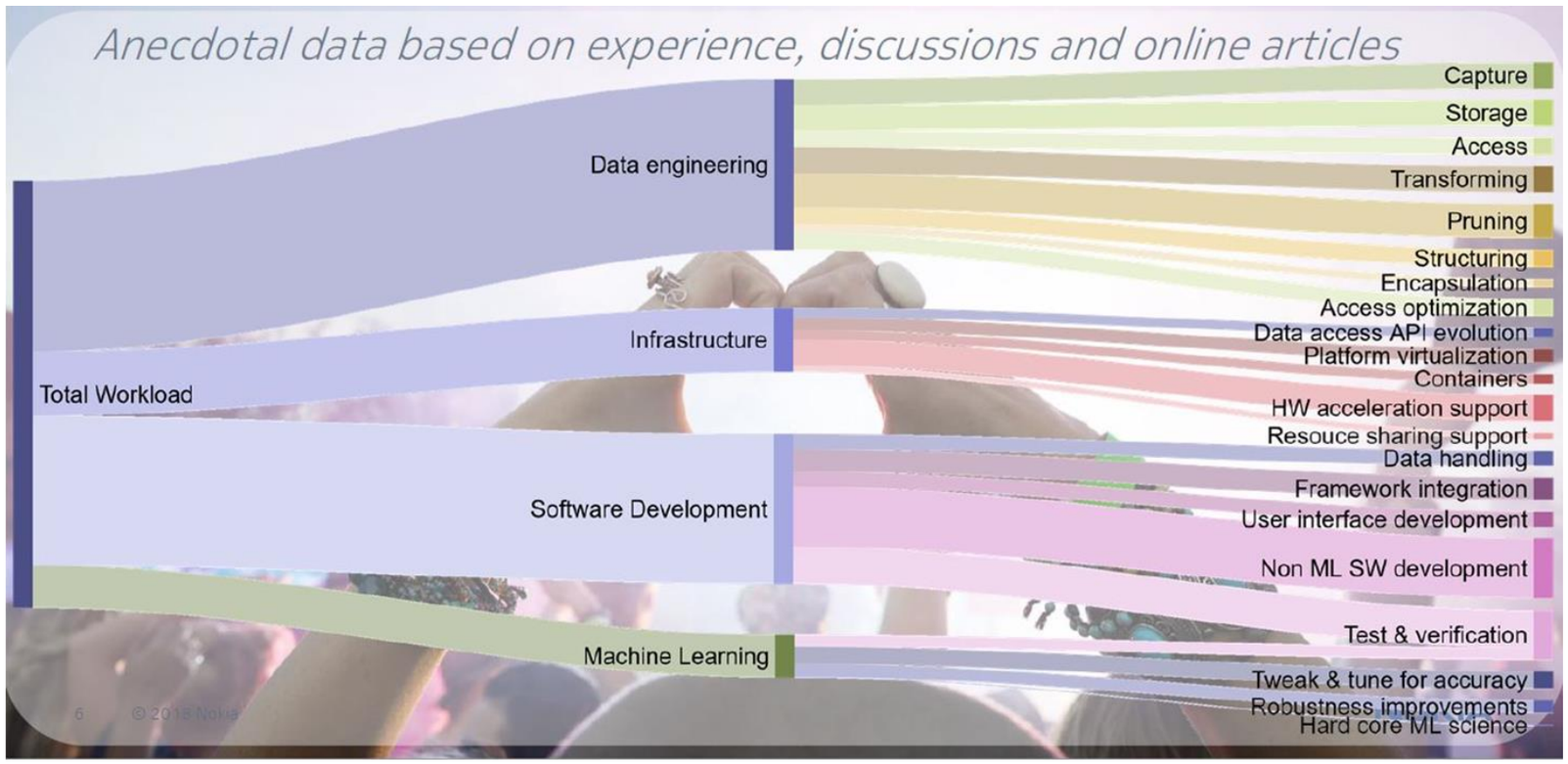
Also, we see that floating-point arithmetic is still required on the training side. The hardware is different, the cycle is different, and the bottlenecks are different. It's a better approach to start with centralized training facilities, and then we can move to distributed training as required.

Monica: Will you move from a centralized to a distributed approach as you move from training to inference?

Tero: You gather the data in the locality, but the training itself, even if the data has to be transported, happens in a centralized environment. Of course, it does not have to be a single centralized location, but in general training is going to be less distributed than inference.

Monica: Can you benefit from the centralized learning and then apply it to different environments?

Tero: This might not be typical for machine learning projects. Most companies can use ready-made networks – for example, for image and voice/speech recognition and processing. For radio networks, unfortunately, the situation is different, as there are no ready-made networks. We are accustomed to setting up and training our own networks. Also, even with centralized training, the network topology or the architecture can be customized to deal with the actual problem or actual data.



Example of machine learning software project workload ***Source: Nokia***

That is something that at Nokia we might do differently than in the rest of the industry. I speak about training, but at the same time refer also to architecture exploration. Even if we have a centralized system, we can still deploy completely individual networks and completely individual training for the system's elements.

Monica: What are the best use cases for machine learning in the short to medium term?

Tero: At Nokia, we are looking into four different categories.

The first one is to use machine learning in the making of the radio networks, as a way to increase the productivity of developing of radio networks – for example, parsing the computational SoC tool flow logs, and having a code review automation and automated testing, or software robotics. Machine learning can improve the production of radio networks – the infrastructure, the chips, the deployment and so forth.

The second category is the radio technology itself and the radio algorithms. We have a lot of use cases. We can improve radio resource management or massive MIMO scheduling, or have energy-aware scheduling.

An obvious category – the one I would pick as the most exciting, and the one there is the most to gain from – is the running of the network, SONs, i.e. self-organizing networks.

Opex and capex savings will come into play from predictive maintenance, automated configuration, optimization, fault prediction and recovery, and security management and resilience. Network operations will change a lot.

The radio frequency use cases are very good and exciting. However, you never see them. You never feel them. They work better, but you don't see them. If the network actually works and recovers itself, you can feel it.

The last category is user-plane edge computing. We already have the most distributed computational platform with our networks. The base stations are becoming more and more like data center servers, and we can deploy machine learning capabilities on them.

In the edge computing case or in the cloud computing case, we can also utilize the machine learning capabilities for third-party applications. This happens because the latency of the inference pushes the applications towards the edge cloud servers.

Then energy efficiency requirements push applications out from the end terminals, so they naturally converge in this edge cloud to be most efficient to be implemented.

Monica: Machine learning can do a lot. While some get excited, others fear machine learning is a black box. People may say, "Machine learning may work, but how do we know when to trust it? We don't really understand it."

Tero: This is one of the most important things we are working on: how to increase the understanding, and how to make the machine learning systems more of an engineering methodology and less of an art.

If you have to just try stuff and see how it works, how you can actually deduce what went wrong, and then take a logical step to make it better? In the future, this is something we have to look into.

Currently, this is one of the biggest obstacles preventing the use of machine learning in some use cases. For instance, you cannot use machine learning in use cases where you can't have any errors.

In the first step, before we have this engineering-based approach, we should not assume that a machine learning system is 98% right. We should start from the assumption that a machine learning system is 2% wrong, take that into account in the system and build the resiliency into the application. Then the interoperability and explainability of the machine learning systems will catch up. How to advance in this area is still a research topic today.

The engineering method is the one that works best to break down your problem. You don't need a huge end-to-end network – you can use separate paths for different networks. You need to decide how you divide your task, and how to have different networks do different things. The network of networks is actually more predictable, and you can have some kind of error bounds so that you know what can happen in the system.

Monica: In order to get the machines to learn, we have to learn to do things differently as well. It's a mutual process. There's a lot.

Tero: Definitely, absolutely.

About Nokia

Nokia is shaping the technologies at the heart of our connected world, to transform the human experience. Powered by the research and innovation of Nokia Bell Labs, we serve communications service providers, governments, large enterprises and consumers, with the industry's most complete, end-to-end portfolio of products, services and licensing. We adhere to the highest ethical business standards as we create technology with social purpose, quality and integrity. Nokia is enabling the infrastructure for 5G and the Internet of Things to transform the human experience.

Digital technologies are changing our world. Nokia is driving innovation and the future of technology to power this digital age and transform how people live, work and communicate. www.nokia.com

About Tero Rissa

Tero is Chief Architect, Machine Learning at Nokia Mobile Networks. He has 20 years of industrial experience of scaling new technologies from research and concept engineering into commercial products. He started working with artificial neural networks in 2009, and has since held senior positions in both engineering and management. Most recently before joining Nokia Networks, he was a director of R&D at Microsoft and the lead architect of OZO in Nokia Technologies.

Tero holds MSc in CS/EE, Digital and Computer Systems from TUT Finland, and PhD in Computing from Imperial College London UK. He is also an emeritus member of the Nokia CEO Technology Council and Bell Labs Distinguished Member of Technical Staff.

Uhana | Harnessing network complexity with AI

A conversation with Sachin Katti, Founder and President, Uhana

There is a lot of excitement about what AI can do to optimizing network performance. But to benefit from AI, we need to understand what it can and cannot do, how it works, and what data it requires to make wireless networks smarter.

In this conversation with Sachin Katti, the Founder and President of Uhana, we talked about the role of AI in identifying network anomalies and making recommendations and predictions. AI is a tool to manage growing network complexity, not a substitute for human expertise.

Monica: Sachin, you've been working on Uhana for quite some time, but we still know little about what Uhana is doing. Can you tell us more?

Sachin: Uhana is a little bit more than two years old now. Uhana is building an AI-enabled control-plane engine for mobile networks. We apply principles of real-time AI to learn and then tune the control plane for network optimization, automation and application acceleration in mobile networks.

Monica: Before we talk about network optimization, let me ask you about the different roles of AI, machine learning and neural networks in improving network performance. Which are you using and why?

Sachin: There is a lot more programmability coming to the network. With the advent of SDN and 5G, we are seeing a lot more control knobs inside the networks that we can dynamically tune for a variety of applications.

All these control knobs add complexity. They all have dependencies on each other. It's hard for humans to figure out how to program them to achieve a particular optimization objective, whether it is the network or an application they want to optimize.

Uhana is building a real-time engine that learns the control plane of the network. We are building an engine that can tune all of these knobs based on real-time data, independent of human intervention.

We use deep-reinforcement learning, a technique that gained fame after Google used it to beat the best Go human player with its AlphaGo. If you think about what is happening in that game, the moves you make on the game board are like control knobs. With deep reinforcement learning you find the optimal set of moves to win the game.

The network control problem is similar. We figure out what are the right sets of control knobs to change for achieving any particular objective. This is one of the fundamental techniques that we use to build an intelligent control plane.

Monica: As you mentioned, network complexity is growing, because the number of knobs is growing. Humans are not capable of dealing with all of them at once. We need a continuous, closed-loop learning process – and this is an ongoing process.

Sachin: It is a dynamic system. The network keeps changing and it's hard to keep track of all the changes. New applications keep coming, at a rate that is continuously increasing.

It's not a set of static applications like in the past where you had just voice and maybe some limited form of data. Nowadays, the network is being used for a mix of applications that changes continuously.

Because these things are so dynamic, your system and control plane also need to continuously adapt, learn what changes are taking place, and tune the network. It's not enough to statically analyze something offline

and then deploy the change. You always have to monitor and continuously learn to make sure that the network runs efficiently.

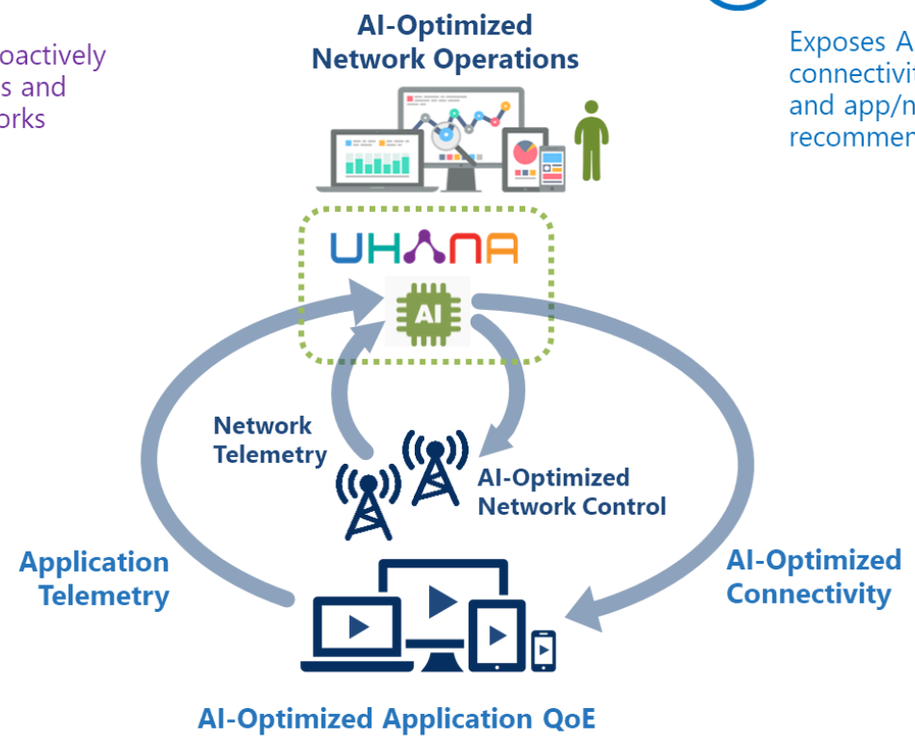
Monica: What do you need to learn in a network to optimize it? Can you walk us through how you're tackling network optimization?

Sachin: The control plane has both network-layer control knobs and application-layer control knobs. Our engine learns how to control both layers from the start. First, I will talk about the network layer and then about the application layer.

1 AI-Optimized Network Control & Operations

Neural network pipeline proactively identifies network problems and enables autonomous networks

Use Cases:
- Intelligent Load Balancing
- P3 Score Optimization
- Interference Detection
- Root Cause Analysis
- Full Stack Monitoring



2 AI-Optimized Application Control

Exposes APIs to provide predicted user connectivity quality (throughput & latency) and app/network layer control recommendation (e.g., video bitrate, fps)

Use Cases:
- Video Entertainment
- Connected Car
- Video Conferencing
- Interactive Gaming
- Video Surveillance
- Real-time Industrial IoT

Uhana's real-time AI pipeline for network and application control ***Source: Uhana***

On the network layer, we monitor real-time network performance. This could be at a variety of granularities, ranging from the subscriber to the regional level.

We not only monitor it, we also proactively identify problems that might arise. We cannot afford to only react, because it's too late to react when the subscriber experience has been hurt already. We want to predict anomalies and proactively identify them.

Once we have an anomaly, we also want to know how many people it has impacted. Inevitably a network is going to have a lot of problems, a lot of anomalies. If we try to chase every problem, we just don't have enough people to look at all the issues.

After detecting anomalies, the first half of learning is to assess the impact of each problem on the network or users. Then for each high-impact anomaly, we classify the root causes and find the likely explanation.

At this point, the system acts as a human augmentation system. It uses its learning to make humans more productive. We still expect humans to take corrective actions based on the root-cause mitigation recommendations.

Once there is enough trust and confidence in the system, we enter a final stage in which the system will do closed-loop control.

From the root-cause classification, the system also learns to change the configuration or control log in the network to automatically fix the problem. We have already seen some instances of this, and this will happen more and more over time, because operators need to optimize dynamically and in real time.

A human in the loop inevitably adds latency before control is applied. For future applications, networks and applications need real-time control. Inevitably, there'll be closed-loop control on the network side.

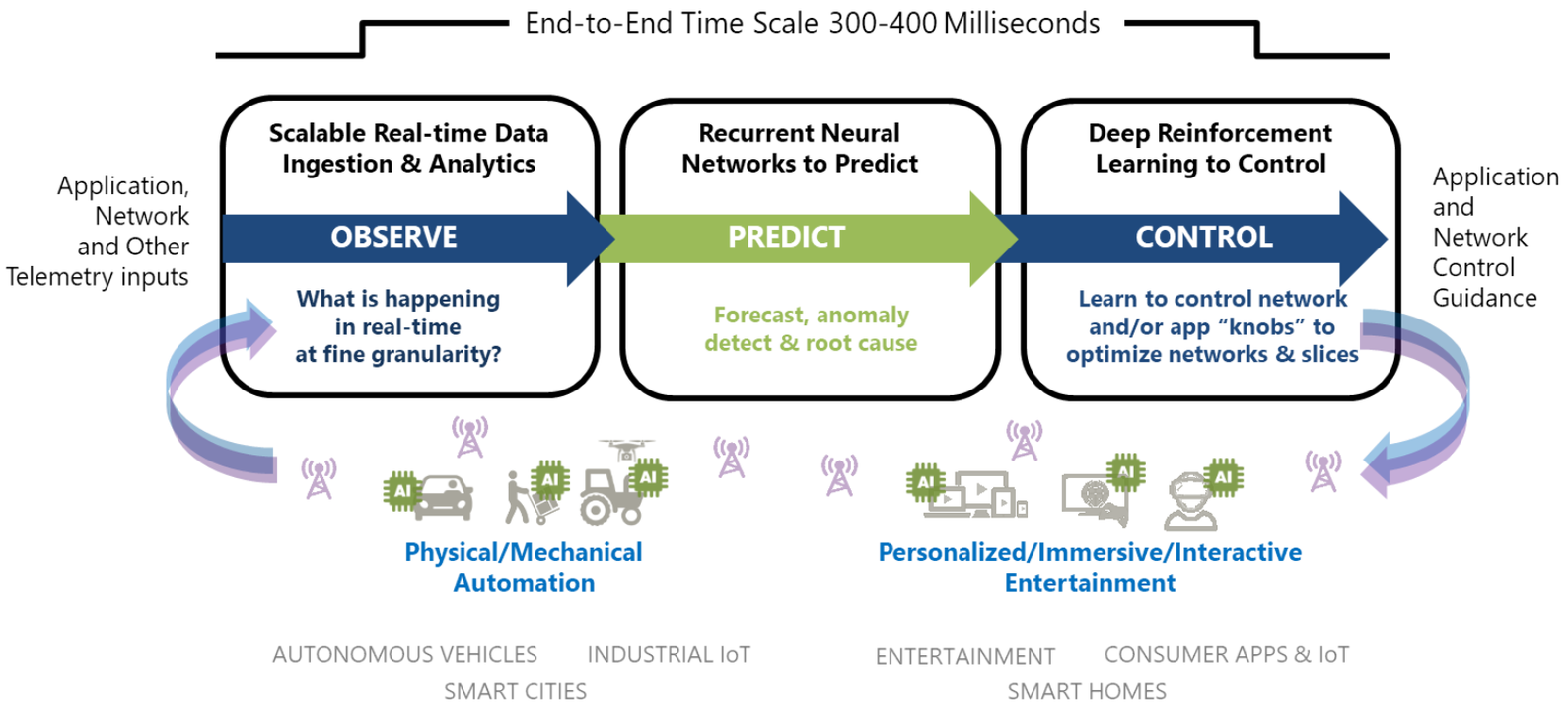
On the application layer, we also have a variety of control knobs to adjust. For example, in video streaming, we can continuously change the bitrate in response to network conditions and other factors.

Similarly, when we're taking in data from the network and the application, we're predicting or forecasting the connectivity that will be available from the network and required by the application. By that I mean the throughput and latency that a particular device will get in the future. Future, in this context, is near real time – the next 30 seconds or so.

Based on the prediction, we learn which control knobs to change in the application. In a streaming video use case, we may want to change the video bitrate, buffering, or other metrics to optimize the video experience over the next four seconds of video.

We are monitoring in real time. We are predicting anomalies and connectivity conditions. We use these predictions to continuously learn what control knobs to tune – and that pipeline applies both to applications and to networks.

Monica: For applications, both the network provider and the application provider have a role in the learning and optimization. Do you work with both? What are their respective roles?



AI is a requirement for next-generation 5G mobile networks and applications ***Source: Uhana***

Sachin: We need to work with both the applications and the network.

The network provides a predictive connectivity API to applications, based on data that only network providers have access to. Network providers are the only ones who know what is happening in the network with every single user. Therefore, they can precisely predict what is going to happen to each user's connectivity.

Application providers use these predictions to make informed control-plane decisions in the applications to optimize metrics of interest that are specific to the application – for instance, the quality of experience for streaming video, video conferencing and interactive gaming.

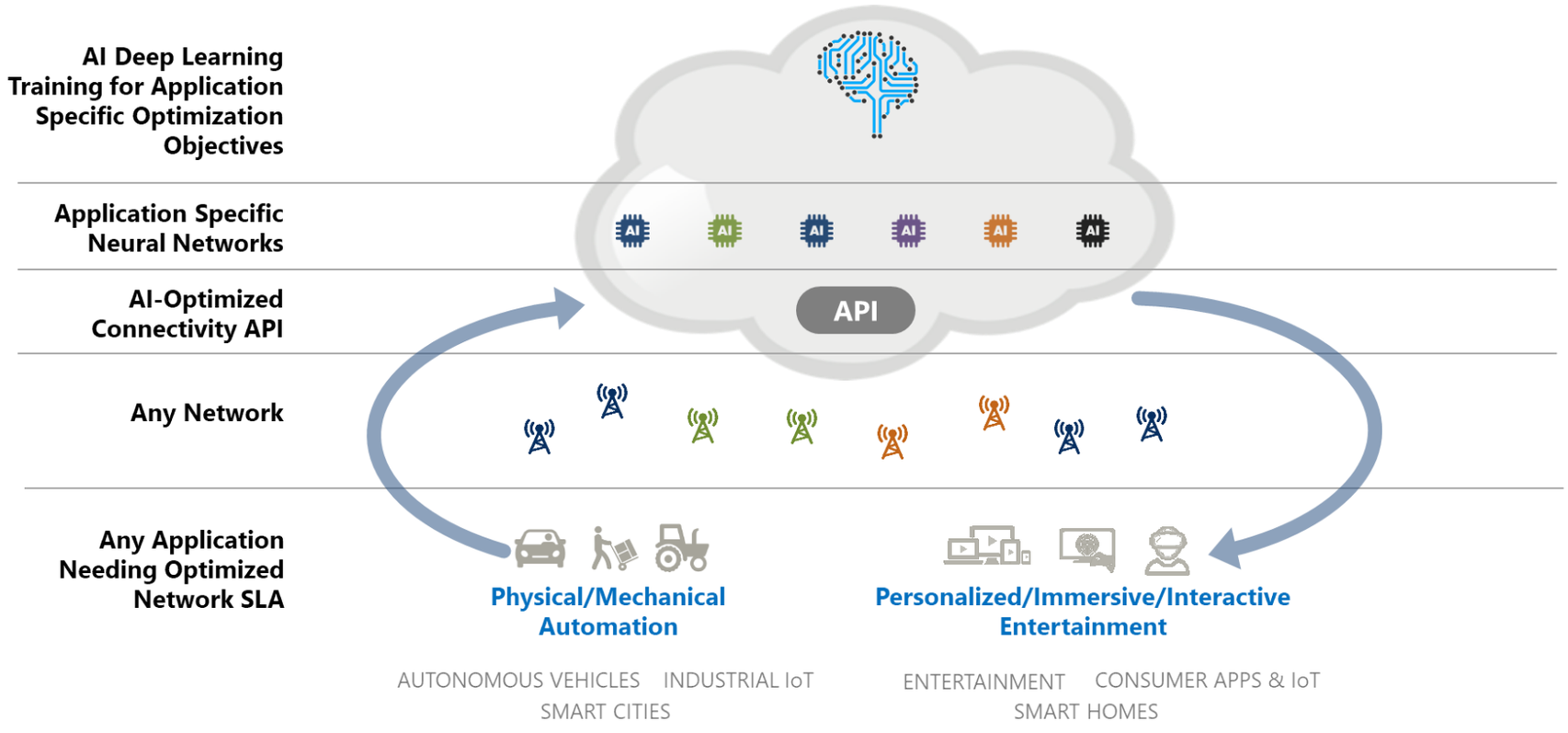
Network providers are becoming more than the cloud. They're providing real-time APIs to a variety of applications that run over their network. Applications work collaboratively with the network and the AI control plane, taking advantage of the predictions to optimize their quality of experience.

In return, the network operators also benefit because the applications can be optimized to reduce the stress on the network. Instead of pushing very high-quality video in a congested

network, the application uses the API to take corrective action to reduce network congestion. Ultimately, both the application and the network controls are optimally tuned simultaneously.

Monica: This is a win-win for everybody. From the operator perspective, applications use less transmission resources in the network. Application providers get a better user experience. And the user is obviously going to be happy.

You mentioned real time. What is real time for you? If you look at real time in the RAN, it's going to be much longer than in other parts of the network. What counts for you as real time?



Open API to provide applications AI-optimized mobile connectivity ***Source: Uhana***

Sachin: For us, real time is on the order of 100 ms or more, because the system we're building is completely deployed in the operator's cloud, and it ingests data from different points throughout the network and uses that data to build this model. In most cases, the control decisions we make – especially those at the application layer – are longer-term control decisions.

For example, what should the video bitrate be for the next four seconds? The application has to make a control decision in near real time. For that, the time horizon you need is on the order of hundreds of milliseconds rather than very small timescales like one millisecond that the RAN deals with.

At timescales on the order of one millisecond, the decision is best left to the infrastructure, because that time is too short for a human to apply learning across the available data. There's not enough time for a human to focus on all the data, and we have deterministic algorithms that run inside the infrastructure at these short timescales.

We avoid conflicts with lower-level control loops. We run relatively longer-timescale control loops in the cloud to build API to application setup.

Monica: You mentioned anomaly detection. This is a complicated process, because you don't know what an anomaly is until you see it.

This is especially the case in complex wireless networks, where you have many variables and parameters. And anomaly prediction is even harder. How do you detect or predict anomalies?

Sachin: The technical term we use for this is multivariate anomaly detection. Anomalies are a function of the granularity average you're looking at.

You could define anomalies at the whole network-wide level or at a single-user level. You can do that at different dimensions and different timescales. We built our system in a way that we do not make any assumptions about the granularity of the anomaly detection.

Our system learns what is normal behavior along every dimension. What is the normal user experience, given that particular user's location in the network? What kind of link does the user have to the network? What is the user's mobility, and how is it expected to change?

We are statistically and continuously computing what the current conditions and behavior for that user or that subset of the network are. We then compare this data to what we have trained our models to predict as normal behavior for that region given the current condition.

Instead of humans predefining what is normal, the system continuously learns along every dimension in time and space what is normal behavior. Then, in real time, it computes whether what it is seeing right now is statistically significantly different from what it is predicting to be normal behavior.

If the difference is beyond a particular threshold, our system will regard the difference as an anomaly. This gives us a set of anomalies to work with. It basically prioritizes human attention at that point. In the next stage, the system decides whether we should care about this anomaly.

If the anomaly affects only 10 users in a network of a million users, maybe we don't care much about it. But if it affects 100,000 users, we care about it a lot more. If it is an anomaly that we care about a lot, the system evaluates possible root causes.

Training models have been created for known root causes, and the system estimates the likelihood that a particular root cause explains the detected anomaly. Is it a fiber that has been cut somewhere? Is it a misconfigured antenna somewhere? Is it a misconfigured new device? Was a new iPhone model introduced and is there some bug in it? All of these are potential root causes that may be the likely explanation for a detected anomaly.

Monica: Some perceive AI systems as black boxes because you don't have control over them. This is not a deterministic model where you say if A then B. You look at complex situations and the system gets back to you with a recommendation, but you don't understand where it's coming from.

How can you address this? Do we have to think in a different way? We have to accept that our traditional deterministic mindset is no longer valid, and that networks work in a different way because they are dynamic and multidimensional.

Sachin: There are several aspects to this question. We definitely need to start thinking about network performance and anomaly detection more statistically. Inevitably, the systems will produce predictions that sometimes are hard to understand. You have to look at them in a holistic manner to understand statistically.

We designed our system so that when we identify an anomaly or give a root-cause recommendation, we also collect all the relevant data that was used to make that prediction, and we present it to the human.

We do so not because we want to repeat the analysis. It's just to give comfort and confidence to the human about what inputs were used to make this decision. That allows the human to decide whether this is a good recommendation or not.

But that also acts as input to the system itself. If the human expert operator, who has potentially decades of experience, sees that the prediction or recommendation is wrong, that's feedback to the system. That feedback helps us learn better and make a better recommendation in the next stage.

It's a continuous learning process. We cannot just freeze it. Part of this process is the humans continuously giving feedback to make sure we evolve and adapt the system to do better and better.

Monica: It sounds as if it is meta learning, where the humans learn from the black box, and the black box learns from the humans. Eventually the system becomes less of a black box, and you get more visibility into it.

Your recommendations are going to be only as good as the input data you have. If the input data is not reliable, or you do not have the

right data, then it doesn't matter how good your AI neural network is. You're not going to get the right recommendation. How can you ensure that you have the right inputs for your system?

Sachin: A lot of the hard work in using AI in networks goes to scrubbing the data, sanitizing the data, filtering the data, and processing the data to make sure that you give the model the right inputs. A lot of our effort goes into this.

Even within the data we get, there is a lot of data loss, and there's a lot of data noise. Our engineers spend much of their time using domain knowledge to scrub that data to make sure we don't learn the wrong things from noisy or missing data. There's a dependency on the quality of the data that comes from the underlying infrastructure.

Some initial work has been done. Bodies like 3GPP have specified what data should be exposed, but they have not specified how the data should be computed. How should we compute a particular KPI? What should the quality of that KPI computation be? Every vendor has its own interpretation of how to compute that data.

An AI system can learn over time what are the patterns for different vendors' infrastructure, and learn to appropriately use specific vendor data in a different manner.

But it's important for our whole industry to get serious about data quality and to start thinking about how we specify what monitoring and telemetry applications we want to embed on every layer of the infrastructure, and what quality we expect from this. But there has to be no variation across vendors for AI tools to be effectively applied across multi-vendor networks.

Standards bodies and associations such as the O-RAN Alliance have a big role to play in making sure that happens.

Monica: What about 3GPP? Does it have a role too?

Sachin: Absolutely. I think 3GPP is where it all starts, because that's the body that's already defining all of these standards and interfaces. 3GPP needs to start thinking about specifying beyond just control and messaging interfaces. 3GPP should also consider what telemetry data should be exposed by each element and at which level of network granularity.

The larger philosophical shift is that infrastructure specification has to have AI as a first-principle requirement. Today, AI and any of these learning approaches are an afterthought – you specify the infrastructure and then you figure out how to use AI on top.

If from day one you decide to use AI to run your networks, you have to specify your underlying infrastructure, data quality and what data is being reported. 3GPP is front and center in this whole conversation.

Monica: What is the response you get from operators? AI is something that challenges their networks and their internal culture. It's a big change from a qualitative point of view, and using AI requires a lot of effort. What kind of feedback do you get from them? Do they see AI as a challenge or as an opportunity?

Sachin: It's a bit of both. Operators have tremendous interest. As an industry, we are at the inflection point, because we're going from 4G to 5G. People are recognizing the increased complexity that's coming. They realize they need help. They need modern approaches to deal with this additional complexity. There's tremendous interest in solving these problems.

At the same, there are cultural issues. There is the case of culture built on how you run networks today. Even things like people's compensation in operational roles are tied to specific KPIs. Suddenly having software make decisions that impact KPIs that affect people's compensation is a big, big cultural shift.

All these incentives need to be aligned to make sure that AI can be used effectively. Our approach is not to replace human. We intend to make AI a human augmentation tool. The system is too complex to let AI run the whole show.

We'll get there at some point, but at this point it's a human augmentation system for network operations and management. We are trying to make sure by using AI experts can prioritize attention, and they have all the relevant information they need to make quick decisions.

That's how you start building confidence, ahead of cultural changes and before you start trusting closed-loop systems.

About Uhana

Uhana is an exciting start up founded to commercialize groundbreaking Stanford research on AI-based mobile network optimization. The company has built a real-time deep learning engine that is being used to optimize network operations and application quality of experience by multiple tier one network operators around the world.

The Uhana AI control plane is deployed in operator private clouds or public cloud infrastructure. The AI control plane pipeline begins by ingesting real-time telemetry from many sources, including the mobile network infrastructure and applications. The real-time telemetry data is combined with operator policies and joined with other inputs, then processed through application specific neural networks. The neural networks deliver real-time, predictive guidance that is used to optimize application QoE and network operations.

About Sachin Katti

Sachin Katti is the President and Founder of Uhana and a Professor of Electrical Engineering and Computer Science at Stanford University. His industry leading roles include helping to lead the O-RAN Alliance in developing next generation open and intelligent RAN architectures. Previously Sachin founded Kumu Networks which commercialized his research on full duplex radios.

Glossary

3GPP	3rd Generation Partnership Project
ACIA	Alliance for Connected Industries and Automation
AI	Artificial intelligence
AIML	Artificial intelligence, machine learning
API	Application programming interface
AR	Augmented reality
B2B	Business to business
B2C	Business to consumer
IoT	Internet of things
LTE	Long Term Evolution
KPI	Key performance indicator
MEC	Multi-access Edge Computing
MIMO	Multiple input, multiple output
ML	Machine learning
mMIMO	Massive MIMO
NFV	Network Functions Virtualization
NR	New Radio
OPS	Operations per second
O-RAN	Open Radio Access Network
QoE	Quality of experience
RAN	Radio access network
ReLU	Rectified linear unit
SDN	Software-defined networking
SoC	System on a chip
SON	Self-organizing network
URLL	Ultra-Reliable Low-Latency
URLLC	Ultra-Reliable Low-Latency Communications
TCO	Total cost of ownership
VoLTE	Voice over LTE
VR	Virtual reality

References

[1] Bennis, Mehdi, Smartphones Will Get Even Smarter With On-Device Machine Learning, IEEE Spectrum, 2018.
[2] Bennis, Mehdi, EDGE AI/ML, 2019.
[3] J. Chunxiao, Z. Haijun, R. Yong, H. Zhu, C. Kwang-Cheng, and H. Lajos, Machine Learning Paradigms for Next-Generation Wireless Networks, IEEE Wireless Communication, 2016.
[4] K. M. Golam, N. Kien, V. G. Porto, Z. Ou, I. Kentaro, and K. Fumihide, Big Data Analytics, Machine Learning and Artificial Intelligence in Next-Generation Wireless Networks, 2018.
[5] M. Brendan and R. Daniel, Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google AI Blog, 2017.
[6] M. David, Machine Intelligence and Networking Challenges, Opportunities and Realities, IETF, 2016.
[7] P. Jihong, S. Sumudu, B. Mehdi, and D. Merouane, Wireless Network Intelligence at the Edge, 2018.
[8] Paolini, Monica, Getting edgy. Optimizing performance and user experience with edge computing, Senza Fili, 2018.
[9] Paolini, Monica, Going deeper with automation, Senza Fili, 2018.
[10] Paolini, Monica, Know your network. Extracting the benefits of virtualization, Senza Fili, 2018.
[11] Paolini, Monica, Power at the edge. Processing and storage move from the central core to the network edge, Senza Fili, 2017.
[12] Rissa, Tero, Machine Learning – What's in it for communication networks, Nokia Mobile Networks, 2018.
[13] T. Pablo, P. Enrique, N. Laurent, and H. Petri, Implementing Operational AI in Telecom Environments, Tupl Inc., 2018.
[14] Z. Chaoyun, P. Paul, and H. Hamed, Deep Learning in Mobile and Wireless Networking: A Survey, IEEE Communications Surveys & Tutorials, 2019.

About Senza Fili

Senza Fili provides advisory support on wireless technologies and services. At Senza Fili we have in-depth expertise in financial modeling, market forecasts and research, strategy, business plan support, and due diligence. Our client base is international and spans the entire value chain: clients include wireline, fixed wireless, and mobile operators, enterprises and other vertical players, vendors, system integrators, investors, regulators, and industry associations. We provide a bridge between technologies and services, helping our clients assess established and emerging technologies, use these technologies to support new or existing services, and build solid, profitable business models. Independent advice, a strong quantitative orientation, and an international perspective are the hallmarks of our work. For additional information, visit www.senzafiliconsulting.com, or contact us at info@senzafiliconsulting.com.

About Monica Paolini

Monica Paolini, PhD, founded Senza Fili in 2003. She is an expert in wireless technologies, and has helped clients worldwide to understand technology and customer requirements, evaluate business plan opportunities, market their services and products, and estimate the market size and revenue opportunity of new and established wireless technologies. She frequently gives presentations at conferences, and she has written many reports and articles on wireless technologies and services. She has a PhD in cognitive science from the University of California, San Diego (US), an MBA from the University of Oxford (UK), and a BA/MA in philosophy from the University of Bologna (Italy).