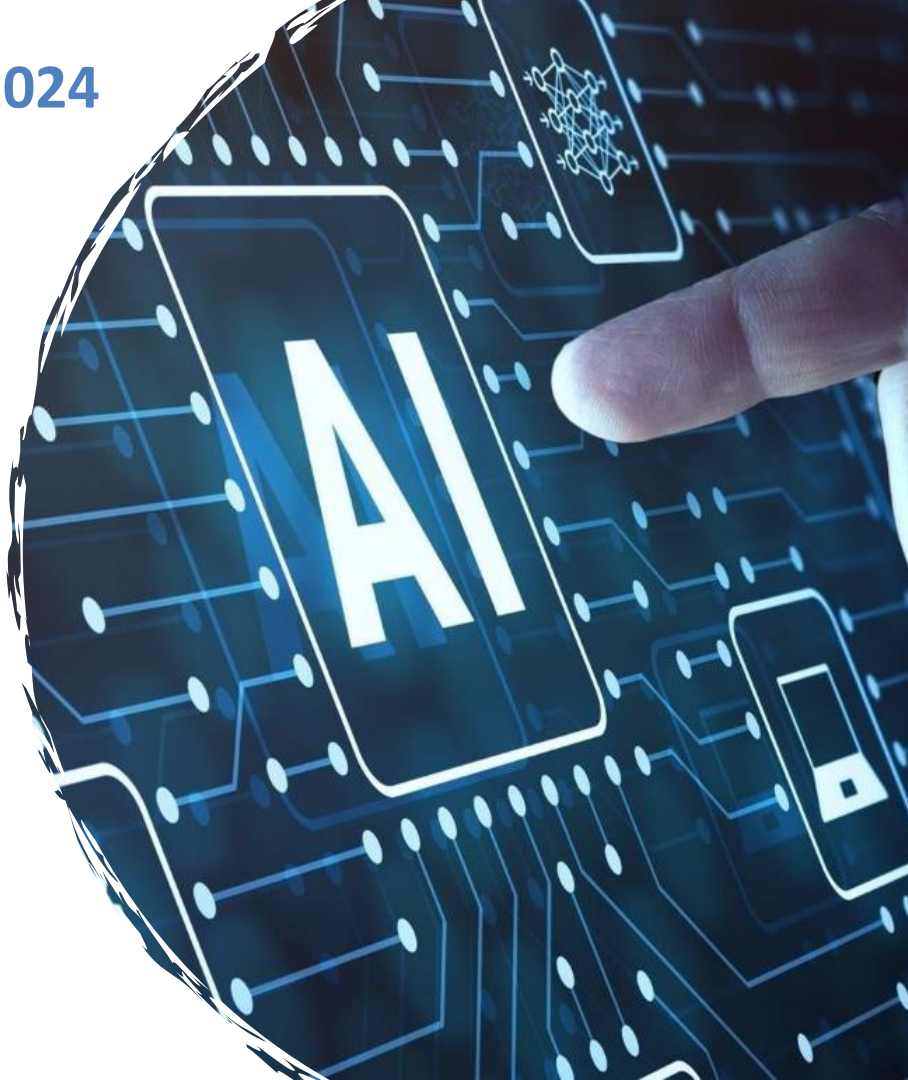
ETSI AI Conference 2024

Overview of AI/ML support related work in 3GPP 5G/5G-Adv Systems



Presented by: **Puneet Jain**

3GPP SA Chair

Sr. Director, Intel Corporation

Feb 6, 2024

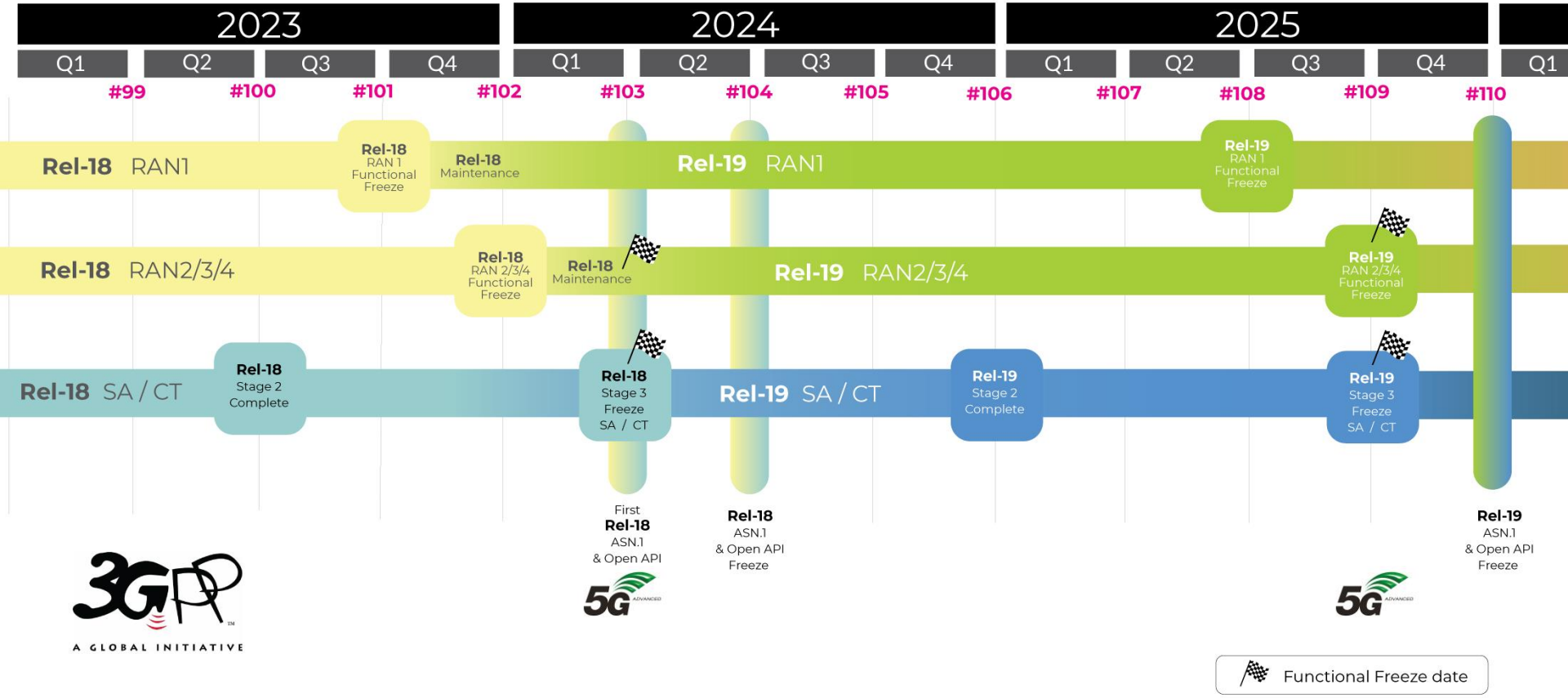
Overview of AI/ML support related work in 3GPP 5G/5G-Adv System

Puneet Jain
3GPP SA Chair

Acknowledgement

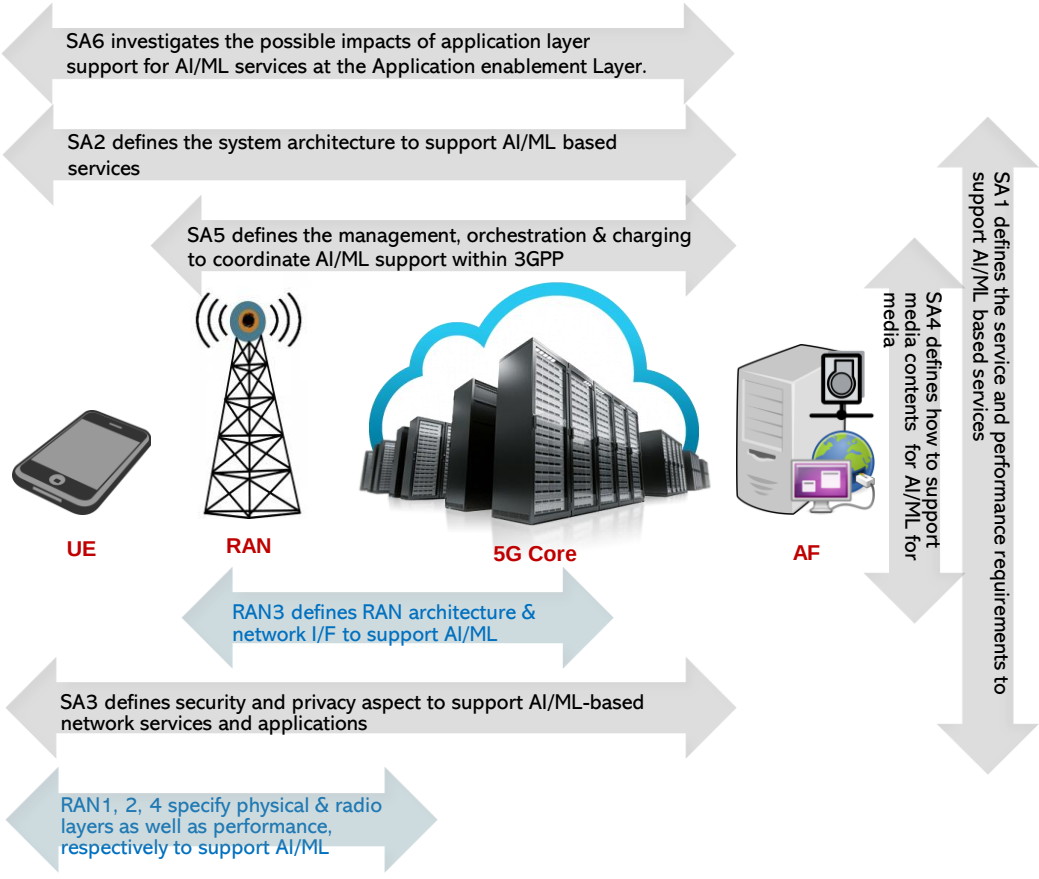
- Thanks to 3GPP delegates, especially SA2 Rapporteurs Tricci So and Xiaobo Wu, for their valuable support in preparing the content of this presentation.

3GPP Rel-18 and Rel-19 timelines



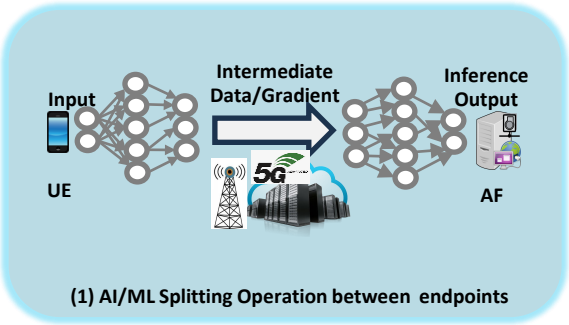
AI/ML work in different 3GPP Working Groups

- ✓ **SA WG-1 (SA1):** Responsible for identifying service and performance requirements for 3GPP systems, in Rel-18, SA1 focused on defining the AI/ML model transfer in 5G.
- ✓ **SA WG-2 (SA2):** Responsible for developing system architecture, in Rel-18, SA2 worked on 5G system support for intelligent transport for the AI/ML-based services.
- ✓ **SA WG-3 (SA3):** Responsible for security and privacy aspects. For AI/ML, SA3 examined and determined the system security and privacy impacts towards 5G Core when supporting AI/ML-based network services and applications.
- ✓ **SA WG-4 (SA4):** Responsible for defining media codec for the system and delivery aspects of the media contents, inRel-18, SA4 defined the AI/ML for media.
- ✓ **SA WG-5 (SA5):** Responsible for management, orchestration, and charging for 3GPP systems, in Rel-18, SA5 defined AI/ML based management functions and the AI/ML management operations to coordinate AI/ML functions across 5G system.
- ✓ **SA WG-6 (SA6):** SA6 is looking for AI/ML services at the Application enablement Layer. SA6 investigates the possible impacts of application layer support for AI/ML services for different deployments and business models.
- ✓ **RAN WG-3 (RAN3):** Responsible for the overall RAN architecture and the specification of protocols for the related network interfaces, in Rel-17 and 18, RAN3 defined the initial support for AI/ML for next-generation RAN (NG-RAN).
- ✓ **RAN WG-1, 2, and 4 (RAN1, RAN2, and RNA4):** Responsible for the physical layer, radio layer, and performance of the radio Interfaces for UE, Evolved UTRAN, NG-RAN, and beyond, respectively, in Rel-18, these WGs define AI/ML for new radio (NR) air interface which is led by RAN1.

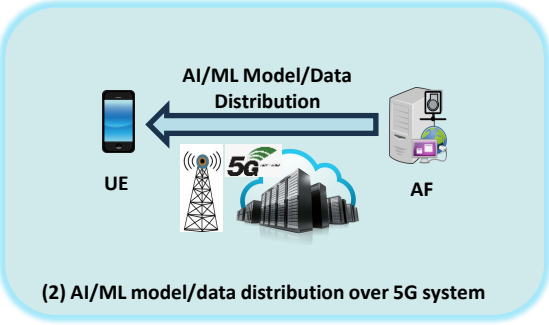


SA1 Services & Performance Definitions & Requirements

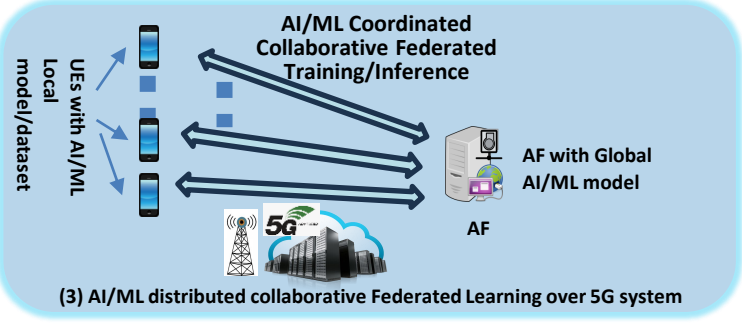
❑ Defining 3 AI/ML Model Transfer use cases:



(1) AI/ML Splitting Operation between endpoints



(2) AI/ML model/data distribution over 5G system



(3) AI/ML distributed collaborative Federated Learning over 5G system

❑ Defining AI/ML Service Requirements:

- ✓ Identify the AIML related key requirements to Uu interface, including
  - *Candidate member selection for Federated Learning (FL)*
  - *Aggregated QoS management for Federated Learning*
  - *In-time exposure of Network status, Event alerting (e.g. QoS prediction) to the authorized AIML application*
  - *Network resource monitoring for an authorized AIML application*

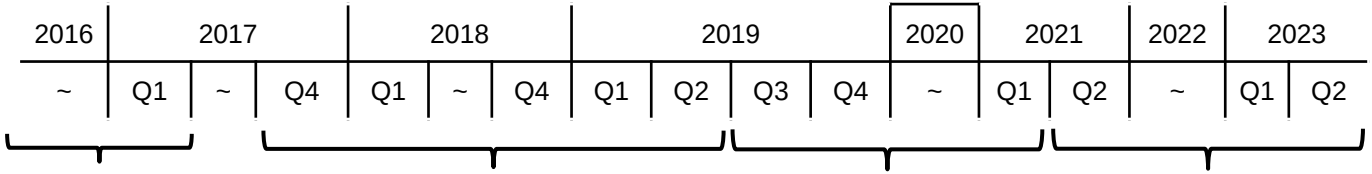
NOTE: The applicability of the requirements is subject to operator policy, user consent, and regulatory requirements

❑ Defining AI/ML Performance Requirements:

- ✓ Specify KPIs for AI/ML model transfer in 5G system, including end-to-end latency, experienced data rate, reliability, and communication service availability, among others.

NOTE: 3GPP SA1 Requirements for AI/ML are specified in TS 22.261.

SA2 enablers for Network Automation (eNA) evolution



R15: Analytics Concept

1. Introduce the concept of NWDAF(Network Data Analytics Function) to produce data analytics for network automation.
2. Only one network data analytics is supported i.e. slice specific network data analytics

R16: monolithic Arch

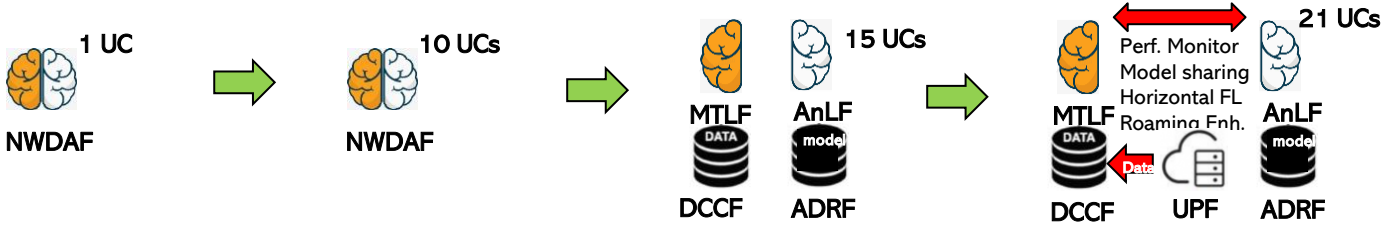
1. Define general framework for data collection and data analytics exposure.
2. NWDAF consist of data collection/model training and model inference

R17: distributed Arch

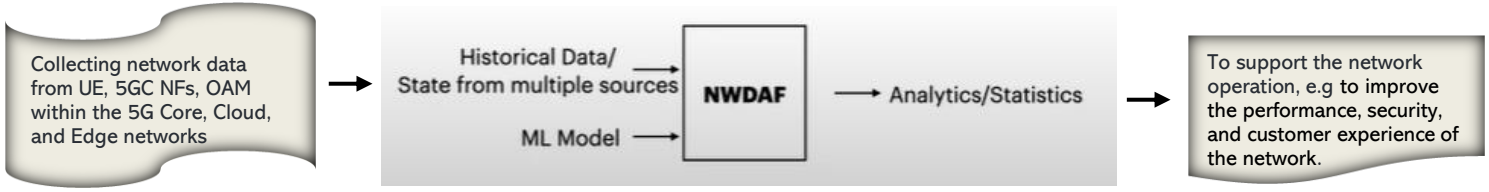
1. Decompose monolithic NWDAF into Model Training logical function (MTLF) and Analytics logical function (AnLF)
2. Define new Analytics and Data Repository Function (ADRF) to support NWDAF functionality with modular design.
3. Data collection framework enhancement via DCCF (Data collection coordination Function)

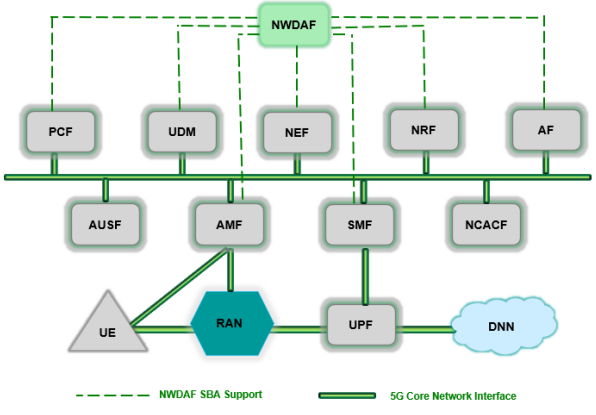
R18: Arch enhancement

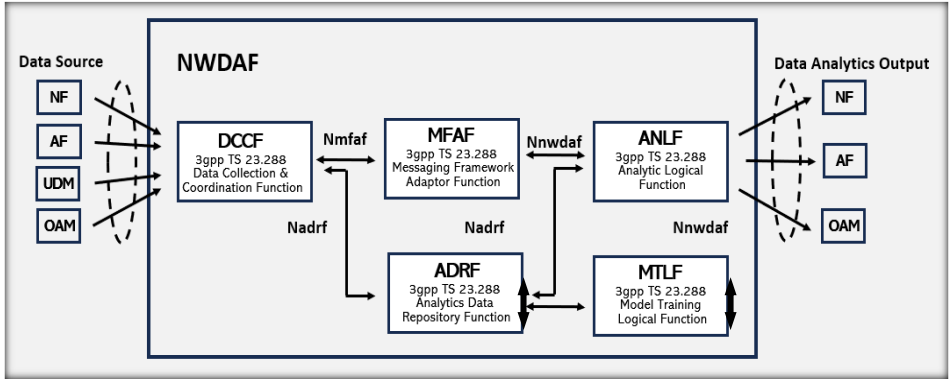
1. Architecture enhancement e.g. model performance monitoring, multi-Vendor ML Model sharing supporting interoperability, Horizontal FL, data analytics in roaming Data collection framework enhancement via UPF data exposure/DCCF relocation.
2. 18 network analytics defined so far



SA2 architecture enhancement for network AI/ML operation (eNA)







What Is Network Data Analytics Function (NWDAF)?

NWDAF as defined in 3GPP TSs 23.288 & 29.520 incorporates standard interfaces from the service-based architecture to collect data, provide analytics, trained ML models by subscription or request model from other network functions. This is to deliver analytics functions in the network for automation or reporting, solving major custom interface or format challenges.

Group of standard functions that are defined by 3GPP for data analytics to support 5G Network Operation:

- ❑ NWDAF-ANLF – Analytical Logical Function
- ❑ NWDAF-MTLF – Model Training Logical Function
- ❑ DCCF – Data Collection Coordination (& Delivery) Function
- ❑ ADRF – Analytical Data Repository Function
- ❑ MFAF – Messaging Framework Adaptor Function

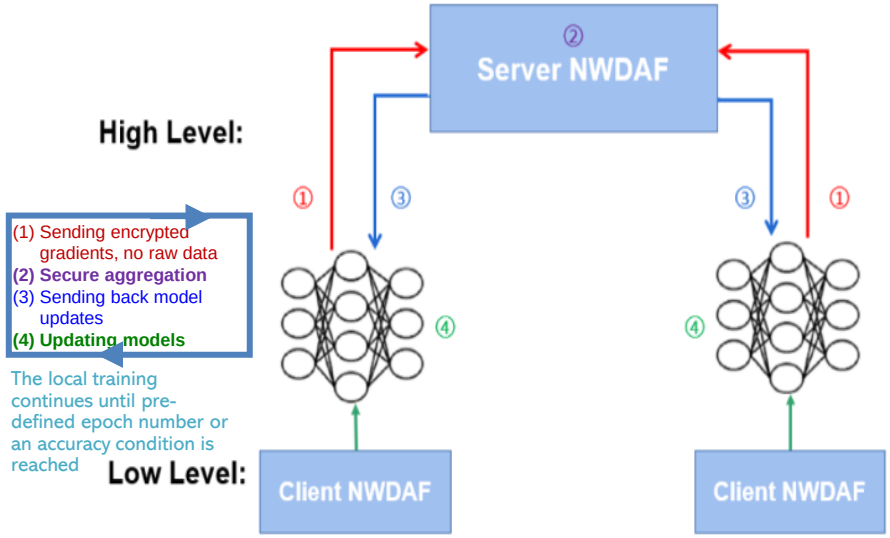
SA2 Architecture Enhancement for Federated Learning (eNA)

Referring to 3GPP TS 23.288, clause 5.3 ….

Federated learning among multiple NWDAFs is a machine learning technique in core network that trains an ML Model across multiple decentralized entities holding local data set, without exchanging/sharing local data set.

This approach stands in contrast to traditional centralized machine learning techniques where all the local datasets are uploaded to one server, thus allowing to address critical issues such as data privacy, data security, data access rights.

When starting an FL procedure, the FL server NWDAF is to provide an initial model to each FL client NWDAF, and then each FL client NWDAF is to perform local model training using their local data set.



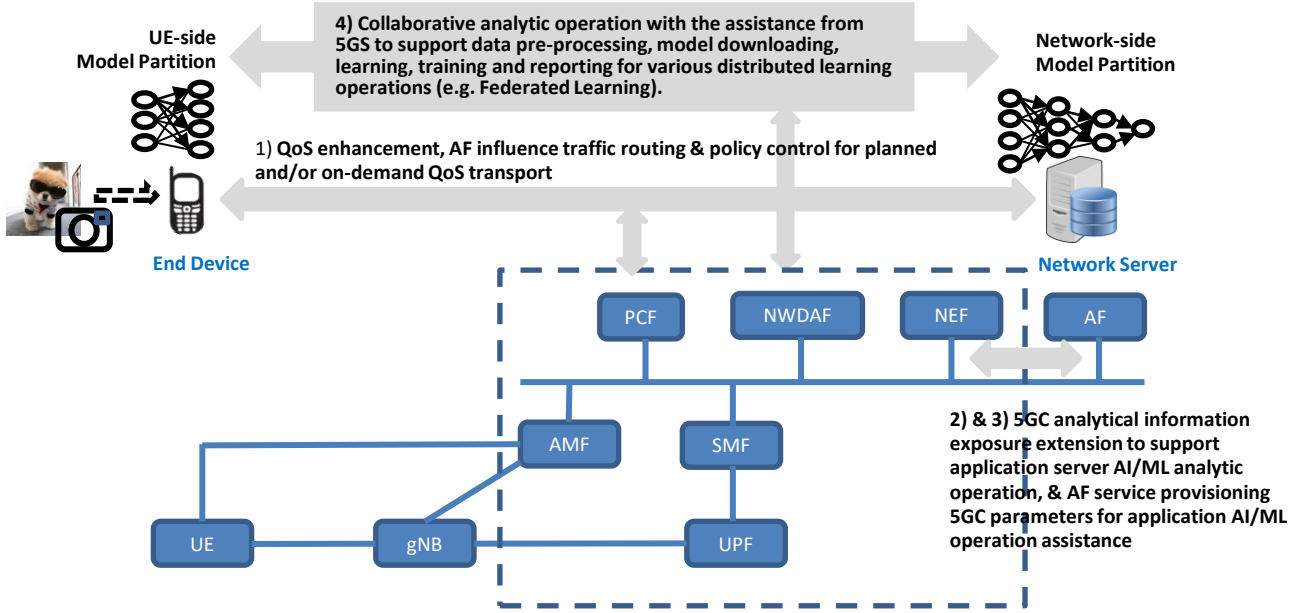
Basic Architecture Framework for Federated Learning is supported in TODAY 5G Core

NOTE: Rel-18 defined horizontal federated learning (HFL). Vertical federated learning (VFL) uses different datasets of different feature space to jointly train a global model (Rel-19 study focus).

SA2 Architecture Enhancement for Application AI/ML Operation (AIMLsys)

In Rel-18, 5G Core is extended to assist Application AI/ML operation. AF remains to control the logic of the application layer AI/ML operation while 5GC:

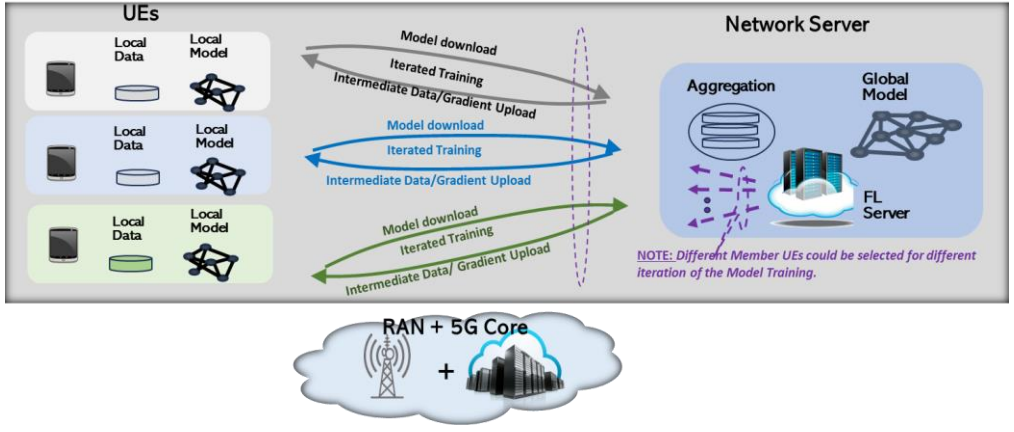
1) Enabling application influence on traffic routing and policy control to provide planned or on demand QoS transport.
2) Extending the network exposure function (NEF) in 5GC to support monitoring and configuration capability for detection and/or reporting of monitoring events to authorized external party
3) Enhancing provisioning capability to allow the external party to provision information to 5GC to facilitate the support of application layer AI/ML operation in 5G system.
4) Enabling 5G system assistance to assist application layer federated learning operation (see next slide for more info).

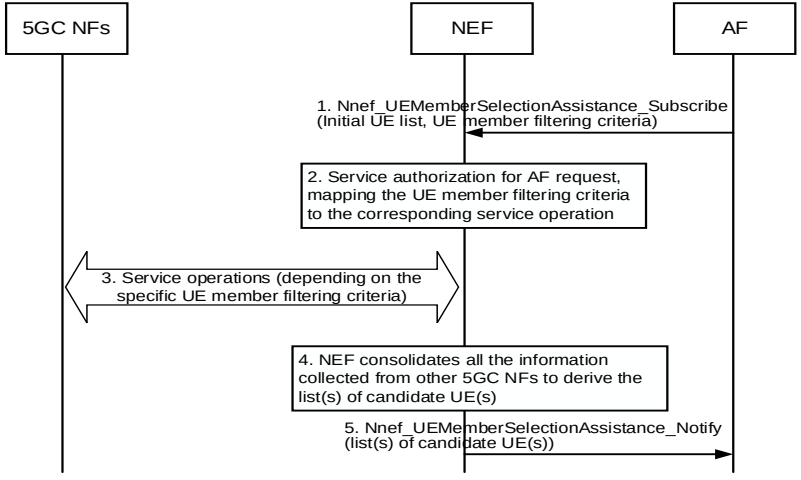


SA2 Architecture Enhancement for Application AI/ML Operation (AIMLsys) – Application Layer Horizontal Federated Learning (AL-HFL) Support

5G Core provides assistant to support Application layer Federated Learning operation, including

1) Candidate FL member selection according to specific set of selection criteria (e.g. UE performance, location and trajectory, network resource availability etc.)
2) Real time Aggregated QoS monitoring to monitor the QoS usage for the FL task
3) Proper time window negotiation with required QoS in order to perform FL and other AIML model transfer service

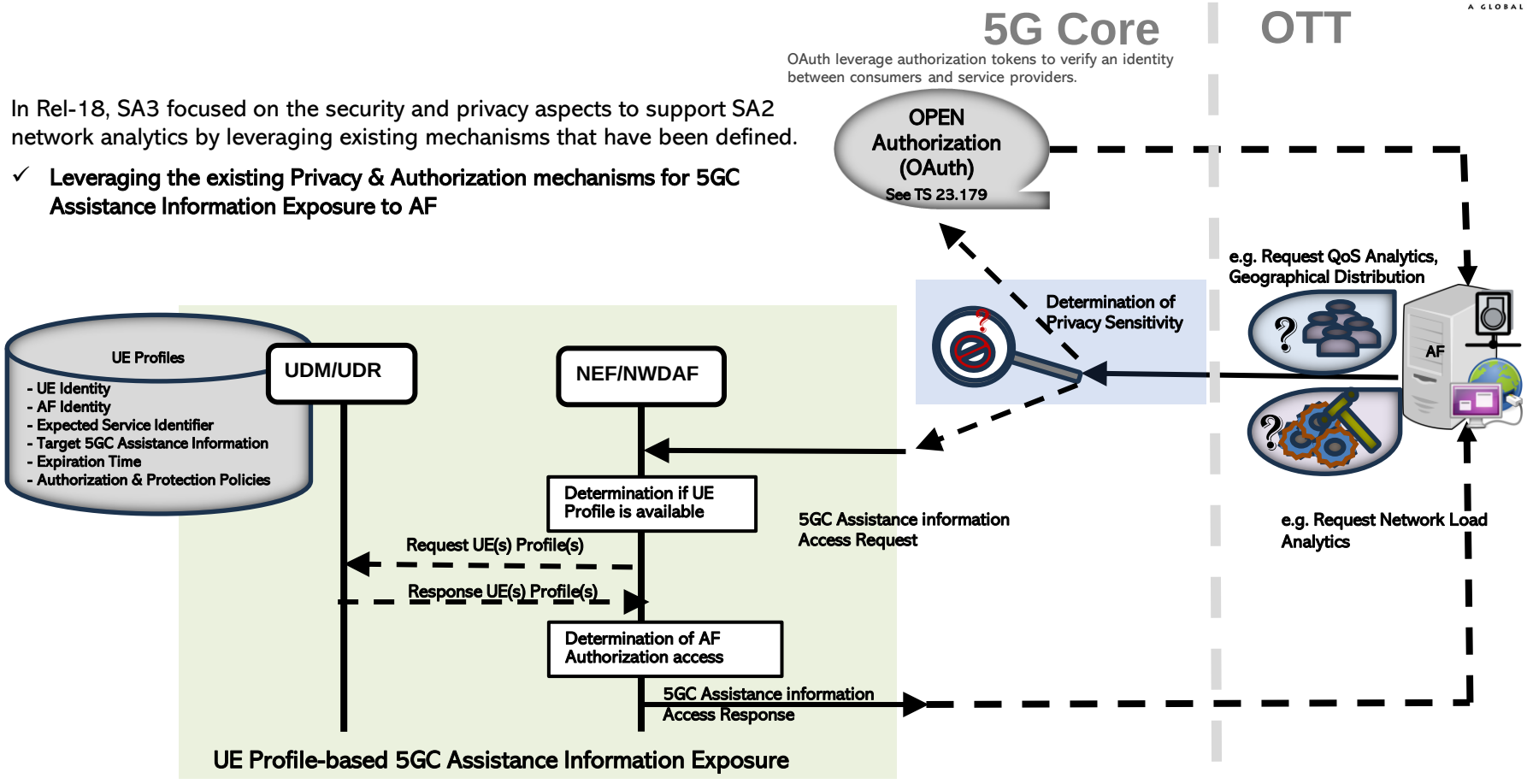




SA3 Security & Privacy support for Network Analytics (eNA_Sec)

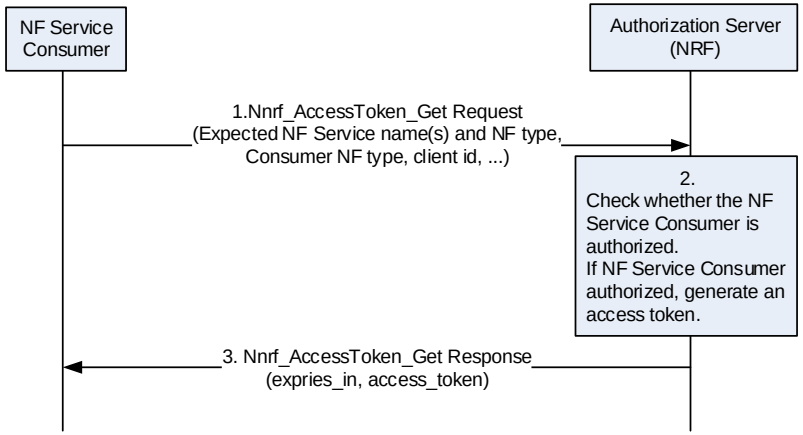
In Rel-18, SA3 focused on the security and privacy aspects to support SA2 network analytics by leveraging existing mechanisms that have been defined.

- ✓ **Leveraging the existing Privacy & Authorization mechanisms for 5GC Assistance Information Exposure to AF**



SA3 Security & Privacy support for network AI/ML operation

Security Goal	Procedure	References
Authorization of NF Service Consumers	Procedure for NF Service Consumer authorization to access data via DCCF, including token generation and service request initiation.	Annex X. TS 33.501
Data Security in Messaging Framework	Focus on confidentiality, integrity, and replay protection in data transfer between 3GPP entities and MFAF.	Annex X. TS 33.501
Protection of Data Between AF and NWDAF	Secure transfer of UE data over SBA interface, ensuring integrity and confidentiality in data exchange.	Annex X. TS 33.501
User Consent Requirements	Compliance with regulatory requirements and document standards for network automation.	Annex V, TS 23.288

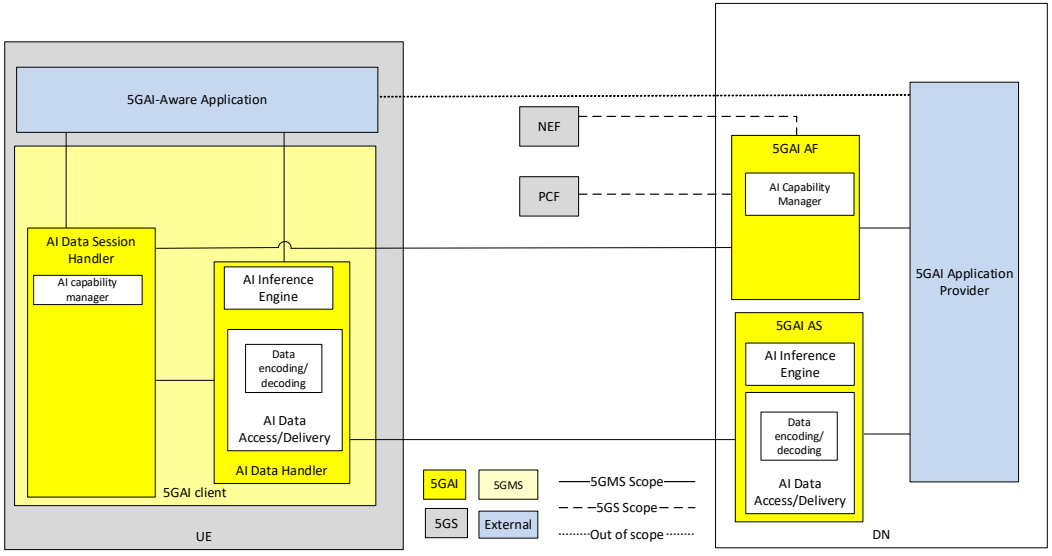


- In Rel-18
  - **eNA_Sec**, SA3 has identified and provided security requirements and procedures for the Network Automation features.
  - **AIMLsys_Sec**, SA3 had a study on the Security and Privacy of AI/ML-based Services and Applications in 5G. No normative work was pursued.
  - **FS_NR_AIML_NGRAN_SEC**, SA3 also focused on the security and privacy aspects to support the RAN3 Rel-18 AI/ML Framework (see later slide on RAN3 reporting). The study concluded that there are no new potential security threats and existing security methods are sufficient. No normative work was pursued.

SA4 AI/ML for Multi-media (AI4Media)

Main Objectives – Defining media service architecture for AI/ML and relevant service flows; in addition, determining the data formats and protocols for various types of data components for AI/ML-based media services, traffic characteristics of the data components delivered over 5G and the respective KPIs.

When applying AI/ML for media, one main consideration is ***the splitting the AI/ML inference between network and UE***. Split points can depend on a number of factors including UE capabilities (e.g., memory, compute, energy consumption, and inference latency), network conditions (e.g., capacity, load, and latency), model characteristics, and user/task specific requirements (e.g., delay and privacy).



AI data delivery general architecture

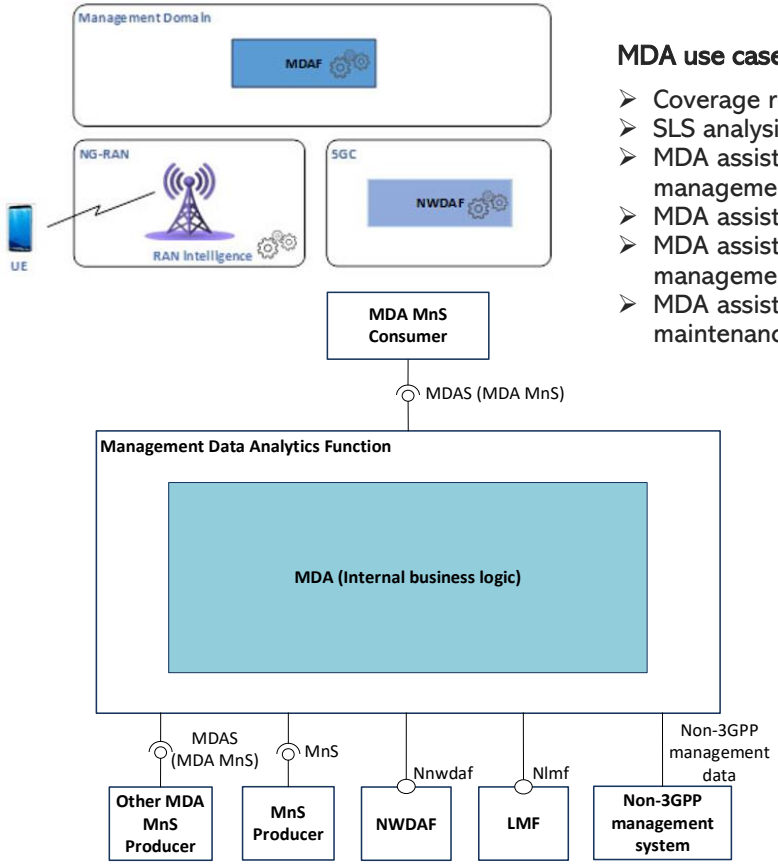
Source: 3GPP TR 26.927

SA5 AI/ML based Management and Orchestration In 5G System

SA5 started the **Management Data Analytics (MDA)** in Rel-17 and continues the AI/ML management specifications development in Rel-18 on the concepts and operational workflows, as well as to address a wide range of use cases (for MDA capabilities) along with the corresponding potential requirements and solutions for the management capabilities and services required for AI/ML **training** & **inference** phases.

Management Data Analytics Service (MDAS), the services exposed by the MDA, can be consumed by various consumers, including for instance MnFs (i.e. MnS producers/consumers for network and service management), NFs (e.g. NWDAF), SON functions, network and service optimization tools/functions, SLS assurance functions, human operators, and AFs, etc.

A management function (MDAF) may play the roles of MDA MnS producer, MDA MnS consumer, other MnS consumer, NWDAF consumer and LMF service consumer, and may also interact with other non-3GPP management systems.



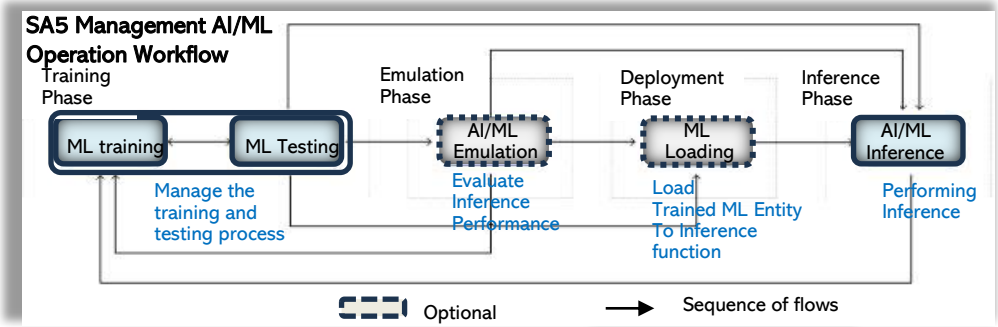
MDA functional overview and service framework

MDA use cases:

- Coverage related analytics
- SLS analysis
- MDA assisted fault management
- MDA assisted Energy Saving
- MDA assisted mobility management
- MDA assisted critical maintenance management

Source: TS 28.104

SA5 AI/ML Management Operations



Category	Use cases
Management Capabilities for ML training phase	
Event data for ML training	Pre-processed event data for ML training
ML entity validation	ML entity validation performance reporting
ML entity testing	Consumer-requested ML entity testing
	Control of ML entity testing
	Multiple ML entities joint testing
ML entity re-training	Producer-initiated threshold-based ML Retraining
	Efficient ML entity re-training
	ML entities updating initiated by producer
ML entity joint training	Support for ML entity modularity – joint training of ML entities
Training data effectiveness	Training data effectiveness reporting
	Training data effectiveness analytics
	Measurement data correlation analytics for ML training
ML context management	ML context monitoring and reporting
	Mobility of ML Context
	Standby mode for ML entity
ML entity capability discovery and mapping	Identifying capabilities of ML entities
	Mapping of the capabilities of ML entities
Performance evaluation for ML training	Performance indicator selection for ML model training
	Monitoring and control of AI/ML behavior
	ML entity performance indicators query and selection for ML training
	ML entity performance indicators selection based on MnS consumer policy for ML training
Configuration management for ML training	Control of producer-initiated ML training
ML Knowledge Transfer Learning	Discovering sharable Knowledge
	Knowledge sharing and transfer learning
Management Capabilities for ML emulation phase	
ML Inference emulation	AI/ML Inference emulation
	Orchestrating ML Inference emulation

Category	Use cases
Management Capabilities for ML entity deployment phase	
ML entity loading	ML entity loading control and monitoring
Management Capabilities for AI/ML inference phase	
AI/ML Inference History	Tracking AI/ML inference decisions and context
Orchestrating AI/ML Inference	Knowledge sharing on executed actions
	Knowledge sharing on impacts of executed actions
	Abstract information on impacts of executed actions
	Triggering execution of AI/ML inference functions or ML entities
	Orchestrating decisions of AI/ML inference functions or ML entities
Coordination between the ML capabilities	Alignment of the ML capability between 5GC/RAN and 3GPP management system
Performance evaluation for AI/ML inference	AI/ML performance evaluation in inference phase
	ML entity performance indicators query and selection for AI/ML inference
	ML entity performance indicators selection based on MnS consumer policy for AI/ML inference
	AI/ML abstract performance
Configuration management for AI/ML inference	ML entity configuration for RAN domain ES initiated by consumer
	ML entity configuration for RAN domain ES initiated by producer
	Partial activation of AI/ML inference capabilities
	Configuration for AI/ML inference initiated by MnS consumer
	Configuration for AI/ML inference initiated by producer
	Enabling policy-based activation of AI/ML capabilities
AI/ML update control	Availability of new capabilities or ML entities
	Triggering ML entity update
Common management capabilities for ML training and AI/ML inference phase	
Trustworthy Machine Learning	AI/ML trustworthiness indicators
	AI/ML data trustworthiness
	ML training trustworthiness
	AI/ML inference trustworthiness
	Assessment of AI/ML trustworthiness

Trustworthiness is identified as a **common** management capability for both the training phase and the inference phase.

- Trustworthiness = AI/ML models {robust, explainable, and fair}.
- Trustworthiness Indicator – configurable and be monitored/evaluated according to Risk & Use Case(s).
  - Preprocessing of training/testing/inference data may be needed according to the desired trustworthiness measure of the corresponding AI/ML model.
- AI/ML MnS producer should allow the consumer to query the AI/ML training producer, inference producer, and/or assessment producer about the supported trustworthiness capabilities and request the configuration, measurement, and reporting of a selected set of trustworthiness characteristics.

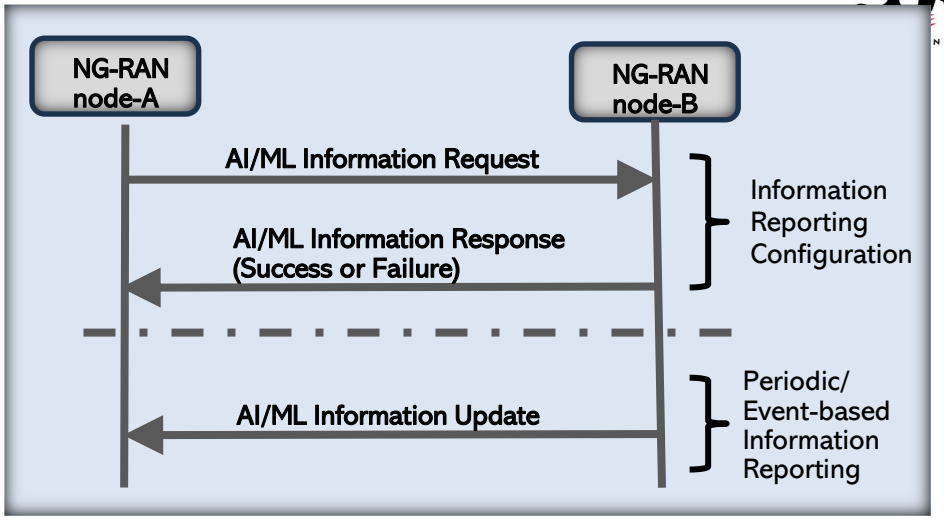
Source: TR 28.908

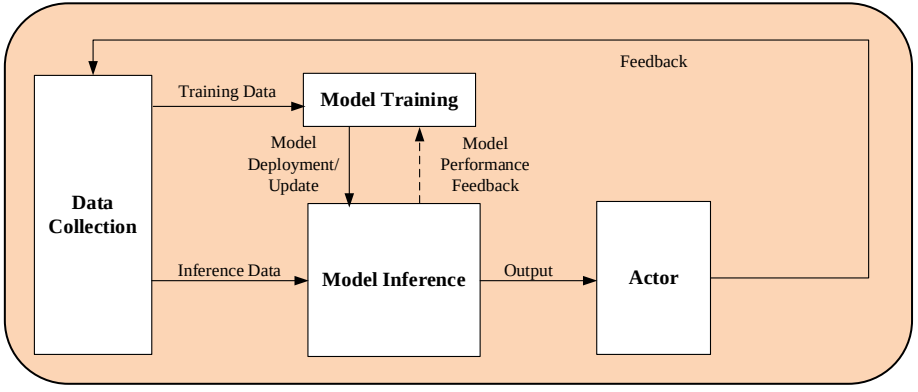
RAN3 AI/ML-enabled NG-RAN

Objective: Improving network performance and user experience, through analyzing the data collected and autonomously process by the NG-RAN with signaling support for: (1) AI/ML based network energy saving, (2) Load Balancing, and (3) Mobility Optimization.

Principles:

- The AI/ML function requires inputs from neighbor NG-RAN nodes over Xn (e.g. predicted information such as cell-granularity UE trajectory, number of active UEs, RRC connections and radio resources, feedback information such as UE's UL/DL throughput performance, packet delay, PER, measurements such as energy efficiency metric etc.)
- Signaling procedures used for the exchange of AI/ML related information are use case and data type agnostic and not dependent on the input, output and feedback
- AI/ML algorithm and models as well as required performance are out of 3GPP scope
- Deployment options for RAN AI intelligence could be:
  - AI/ML model training is located in OAM and inference in gNB, or
  - both can be located in gNB





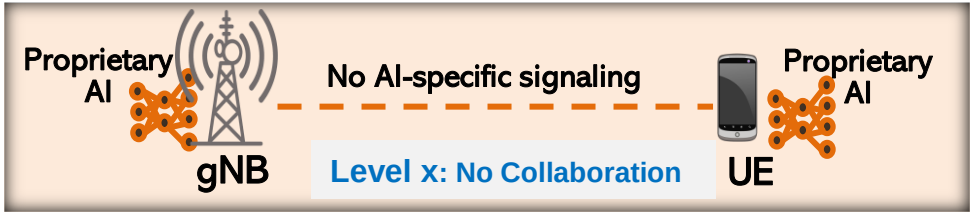
Functional Framework for RAN Intelligence (Source: TR 37.817)

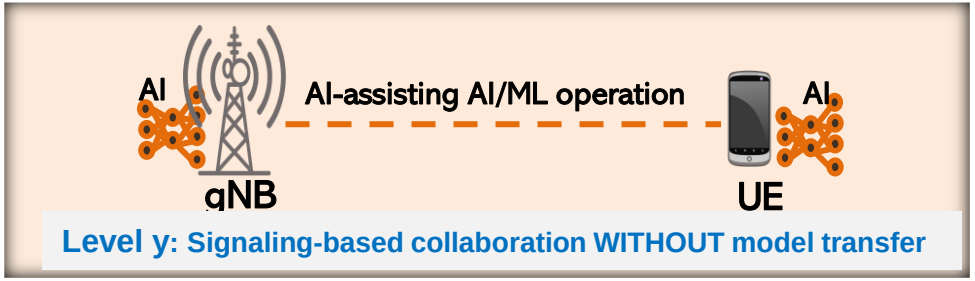
RAN1&2 AI/ML for Air Interface (pave the way to 6G)

Objective: Establishing a general framework for enhancing the air interface using AI/ML – stages of AI/ML algorithms, collaboration levels between gNB and UE, required datasets for AI/ML model training, validation and testing, and life cycle management of AI/ML models.

Three training collaboration models under investigated:

- ❑ **Level x:** No collaboration
- ❑ **Level y:** Signaling-based collaboration without model transfer
- ❑ **Level z:** Signaling-based collaboration with model transfer





Focusing on 3 use cases:

- ❑ **Channel state information (CSI) feedback Enhancement** – leveraging AI/ML techniques to improve CSI compression which includes an AI/ML-based CSI encoder at the UE and decoder at the gNB as well CSI Prediction.
- ❑ **Beam management** – leveraging AI/ML techniques to reduce beam management overhead and latency, as well as improving beam selection accuracy via spatial & temporal prediction.
- ❑ **Positioning** – leveraging AI/ML techniques to improve Direct AI/ML and AI/ML assisted positioning accuracy for different scenarios including those with heavy Non-line-of-sign (NLOS).

3GPP Rel-18 AI/ML Related Study/Work Items

3GPP Rel-18 AI/ML Related Study/Work Items	Acronym	Release	Working Group	% Completion	SID/WID
Study on Enablers for Network Automation for 5G - phase 3	FS_eNA_Ph3	Rel-18	SA2	100%	SP-220678
Enablers for Network Automation for 5G phase 3	eNA_Ph3	Rel-18	SA2	100%	SP-230110
Study on security aspects of enablers for Network Automation for 5G - phase 3	FS_eNA_SEC_Ph3	Rel-18	SA3	95%	SP-220199
Security aspects of enablers for Network Automation for 5G - phase 3	eNA_Ph3_SEC	Rel-18	SA3	65%	SP-230155
CT3 aspects of eNA_Ph3	eNA_Ph3	Rel-18	CT3	90%	CP-230119
CT4 aspects of eNA_Ph3	eNA_Ph3	Rel-18	CT4	90%	CP-230119
Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS	FS_AIML_MT	Rel-18	SA1	100%	SP-220441
AI/ML model transfer in 5GS	AIML_MT	Rel-18	SA1	100%	SP-220440
Study on 5G System Support for AI/ML-based Services	FS_AIMLsys	Rel-18	SA2	100%	SP-220071
System Support for AI/ML-based Services	AIMLsys	Rel-18	SA2	100%	SP-231278
Study on Security and Privacy of AI/ML-based Services and Applications in 5G	FS_AIML	Rel-18	SA3	100%	SP-220687
CT3 aspects of AIML	AIMLsys	Rel-18	CT3	92%	CP-230329
CT4 aspects of AIML	AIMLsys	Rel-18	CT4	90%	CP-230329
Study on AI/ML management	FS_AIML_MGMT	Rel-18	SA5	100%	SP-211443
AI/ML management	AIML_MGT	Rel-18	SA5	80%	SP-230335
Management Data Analytics	EMDAS_Ph2	Rel-18	SA5	87%	SP-220981
NEF Charging enhancement to support AI/ML in 5GS	AIMLsysNEF_CH	Rel-18	SA5	0%	SP-231706
Study on Artificial Intelligence (AI) and Machine Learning (ML) for Media	FS_AI4Media	Rel-18	SA4	50%	SP-220328
Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface	FS_NR_AIML_air	Rel-18	RAN1	100%	RP-221348
Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN	NR_AIML_NGRAN-Core	Rel-18	RAN3	80%	RP-233441
Study on the security aspects of Artificial Intelligence (AI)/Machine Learning (ML) for the NG-RAN	FS_NR_AIML_NGRAN_SEC	Rel-18	SA3	100%	SP-220529

3GPP Rel-19 AI/ML Related Study/Work Items

3GPP Rel-18 AI/ML Related Study/Work Items	Acronym	Release	Working Group	% Completion	SID/WID
Study on AI/ML Model Transfer Phase2	FS_AIML_MT_Ph2	Rel-19	SA1	100%	SP-220439
AI/ML Model Transfer Phase 2	AIML_MT_Ph2	Rel-19	SA1	100%	SP-230514
Study on Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML)	FS_AIML_CN	Rel-19	SA2	0%	SP-231800
Study on AI/ML management - phase 2	FS_AIML_MGT_Ph2	Rel-19	SA5	0%	SP-231780
Study on application layer support for AI/ML services	FS_AIMLAPP	Rel-19	SA6	35%	SP-231182
Study on enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN	FS_NR_AIML_NGRAN_enh	Rel-19	RAN3	0%	RP-234054
Study on Artificial Intelligence (AI)/Machine Learning (ML) for mobility in NR	FS_NR_AIML_Mob	Rel-19	RAN2	0%	RP-234055
Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface	NR_AIML_air	Rel-19			
Core part: Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface	NR_AIML_air-Core	Rel-19	RAN1	0%	RP-234039
Perf. part: Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface	NR_AIML_air-Perf	Rel-19	RAN4	0%	RP-234039

Thank You !

Puneet Jain
Chair of 3GPP SA
puneet.jain@intel.com
www.3gpp.org