ETSI AI Conference 2025

Overview of AI/ML related work in 3GPP

Presented by:

Dr. Juan Montojo

Qualcomm

11/02/2025

Outline

- Scope of standardization of AI/ML
- High level assessment of AI/ML in Cellular Networks
- Former & Ongoing AI/ML standards activities in 3GPP
- Scope of AI/ML for NR Air Interface
- Scope for other 3GPP related AI/ML projects
  - AI/ML in NG-RAN
  - Enhance Network Automation
- Conclusions

Standardization of AI/ML: Scope

- **Network** and **Device** implementations have been using AI/ML models to solve certain problems as an implementation choice complementing / replacing conventional methods
  - 3GPP embarked in AI/ML projects to:
    - improve performance with the understanding that certain implementations would be using AI/ML for solving certain problems,
    - provide operator/infrastructure control
  - As a matter of principle, the **AI/ML models** themselves are **not standardized**
- Most of the standardization work revolves around the following aspects:
  - Infrastructure vendor or mobile operator **control** and **performance monitoring** of device-side models
    - AI/ML model activation, deactivation, swapping
  - Air interface **extensions** or new device measurements enabling AI/ML model training/inference and their generalization
  - Standards-based approaches for **Data collection** (for training) and **Model transfer/delivery** (e.g., improved models)
- **Training**: Online vs. offline
  - 3GPP has been assuming thus far offline training of models
    - AI/ML models are not deployed (commercially) until they have been fully trained & tested (in the field)
  - Computation advances in the future may enable online training
    - Model fine-tuning / refinement may be the 1st step
- **Testability** and **consistent device behavior** is a key concern

AI/ML in Cellular Networks

High Level Assessment

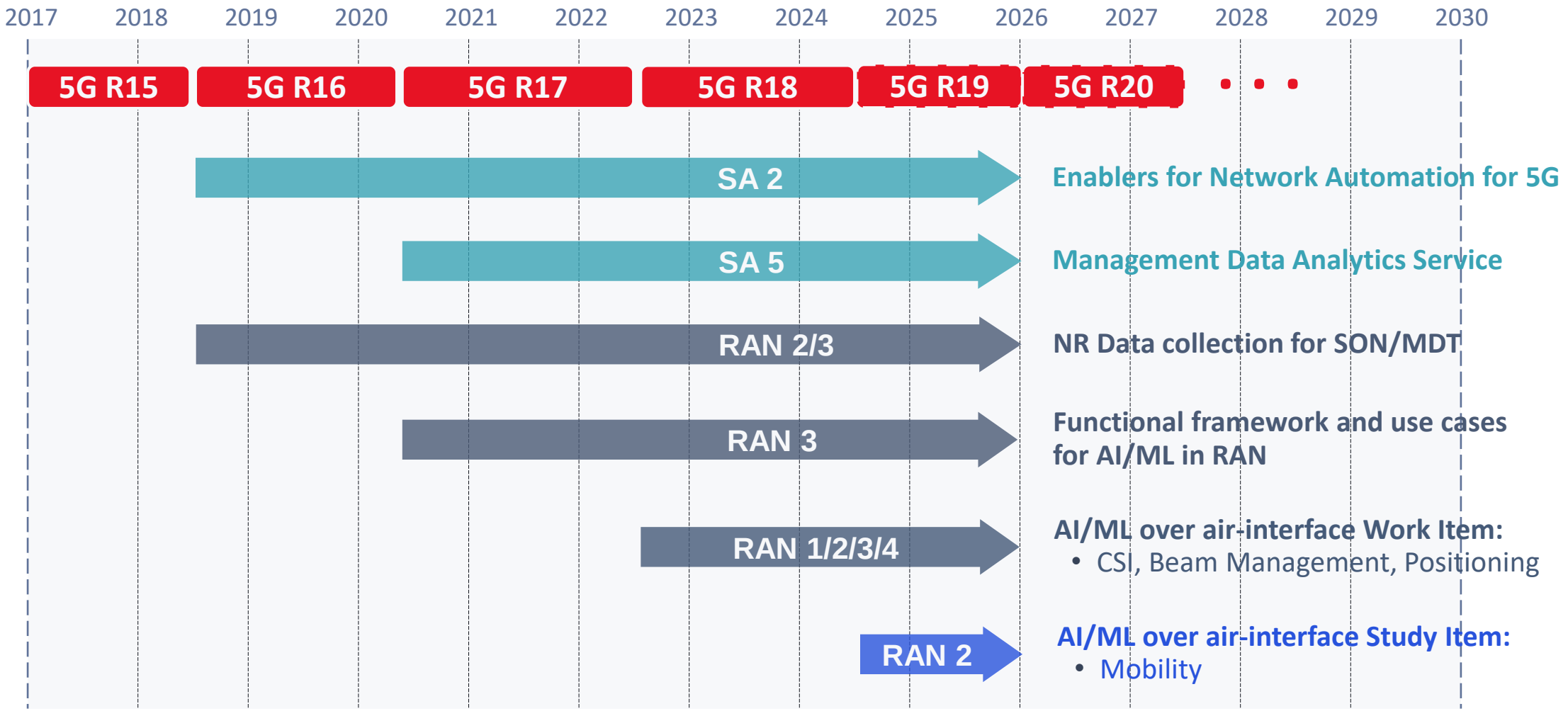
- **Good for:**
  - Data Analytics for, e.g., Network optimization
    - Including putting together data coming from different sources
  - Solving complex problems with no analytical solution
    - Enabling data-driven solutions not necessarily constrained to single solution for all cases or complex NW optimizations
  - Prediction based on data and context awareness
    - Optimal Transmit Beam, Channel State Information, Mobility events…

- **Concerns**:
  - Testability and Consistent device behavior
  - Feasibility of two-sided AI/ML models
  - “Finding problems for solutions” because of the AI hype

- **Not so good for**:
  - Solving (complex) problems with analytical solution
  - Complexity and Energy consumption (NW & UE)

- **Controversies**
  - Data ownership and data sharing
  - Standards mechanism for data collection, model delivery…

Former & Ongoing AI/ML standards activities in 3GPP

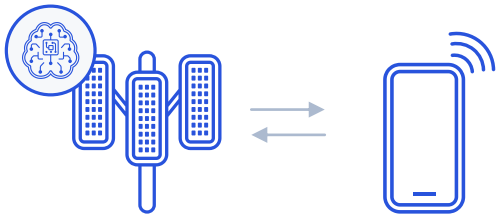


SON: Self Organizing Networks
MDT: Minimization of Driving Tests

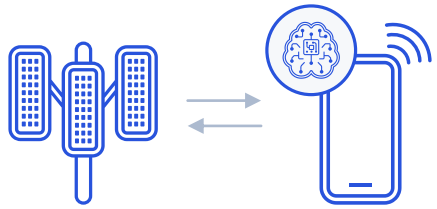
RAN: Radio Access Network
CSI: Channel State Information

AI/ML for NR Air Interface: Scope

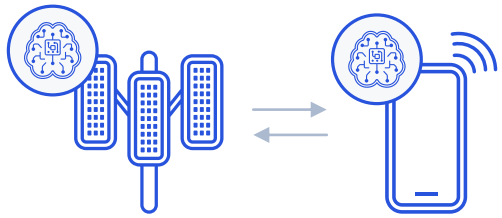
- General framework for Life Cycle Management (LCM) of **one-sided** AI/ML models



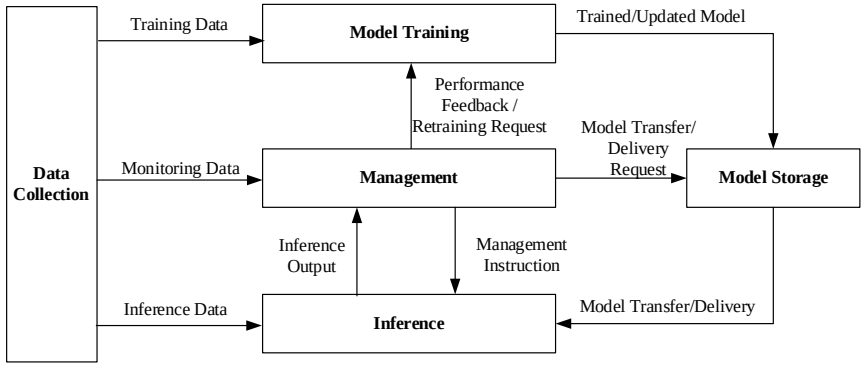
NW-sided model



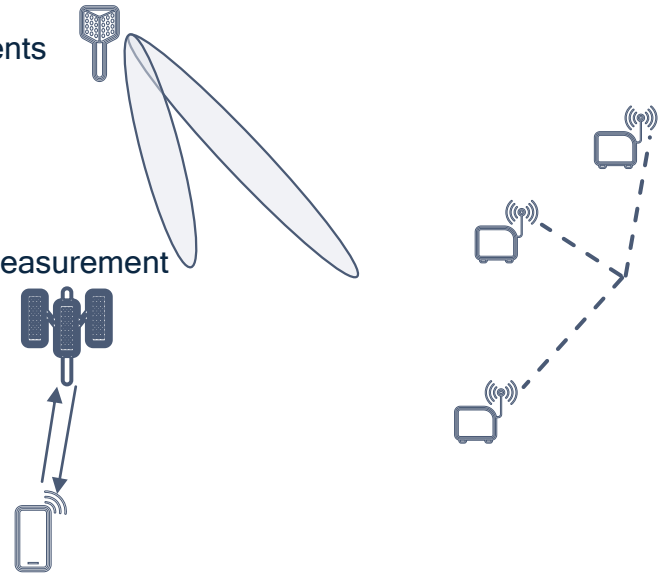
UE-sided model



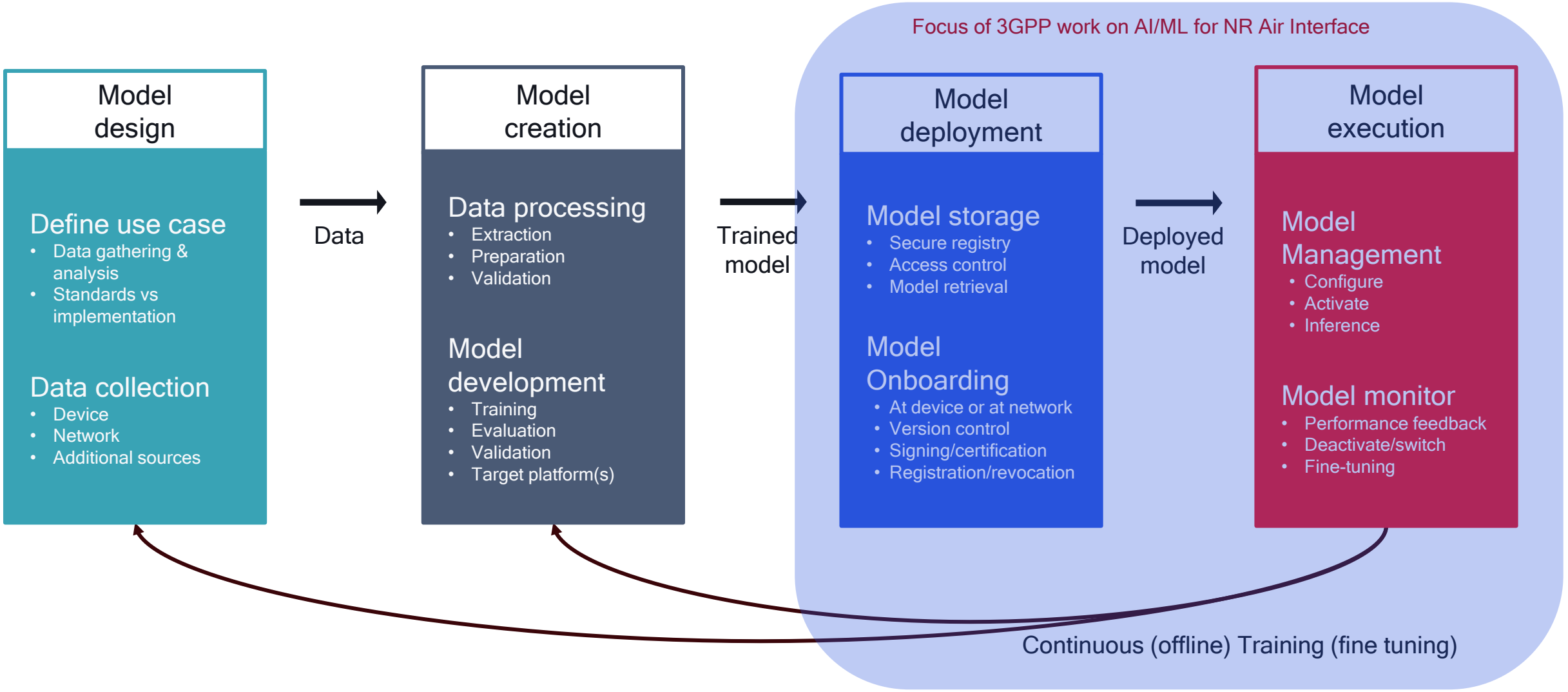
Two-sided model



- Use cases:
  - Beam Management
    - Device/network-sided transmit Beam temporal spatial prediction from subset of beam measurements
    - Device/network-sided transmit Beam spatial prediction from subset of beam measurements
  - Positioning
    - Direct positioning: AI/ML model's output is the positioning coordinates
    - Assisted positioning: AI/ML model's output is used to compute positioning, e.g., refined/curated measurement
  - Channel State Information:
    - Temporal prediction
    - Compression (study): requires two-sided AI/ML model (device and base-station side)
  - Mobility (study in Rel-19)
    - AI/ML based Radio Resource Management (RRM) measurement and event prediction
    - Cell level measurements; Handover failure/Radio Link Failure (RLF) prediction (UE-side only); Measurement events prediction (UE-side only)



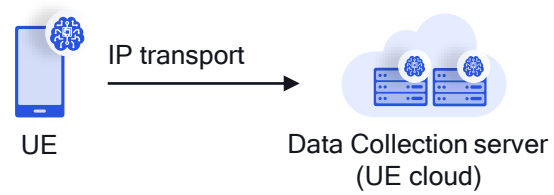
ML Workflow - Life cycle for UE models



Life cycle is for illustrative purposes to help understand the various steps in AI/ML model development and management

4 3 options for data collection under RAN2 / SA2 / SA5 discussion

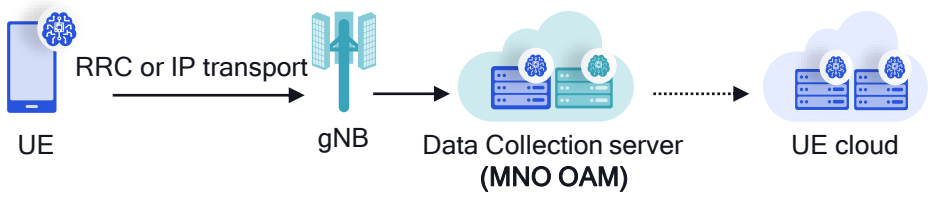
3GPP Rel-19 Work Item status - AI/ML for Air Interface



- Option 1a – Data collection server OTT (outside the MNO)
  - User plane
  - Proprietary solution (no standards impact)

- Option 1b – Data collection server inside the MNO
  - User plane
  - UE collects training data and transfers it to the server
    - Optionally, the server transfers the data to an OTT server (outside the MNO)

- Option 2 – Data collection server in the Core
  - Control plane (NAS) or user plane [FFS]
  - UE collects training data and transfers it to the CN
    - CN transfers the data to a UE-side model training/OTT server



- Option 3 – Data collection server in the RAN (OAM)
  - Control plane (RRC) or user plane [FFS]
  - UE collects training data and transfers it to the OAM
    - OAM transfers the data to a UE-side model training/OTT server

High-level Requirements for Data collection

Agreed principles

1. The data collected is secured and data integrity and confidentiality for that data is ensured.
2. User data privacy, anonymity and user consent is respected.
3. The MNO has full control of the standardized data collection transfer process and can manage data transfer to the server for UE-side data collection, without the need of SLA for this purpose. This includes initiating, terminating, and fully managing data transfer.
4. MNO has full visibility for standardized data.
5. The design is futureproof and extendable.

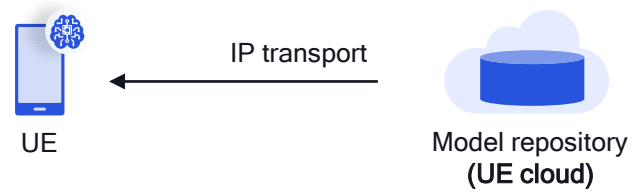
FFS/study if and how to handle non-standardized data (i.e. partial visibility).

FFS controllability on data collection

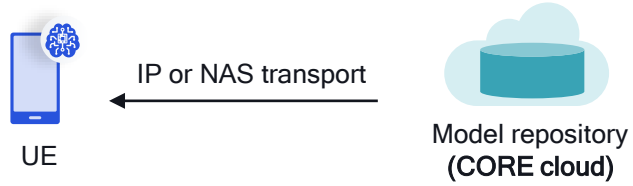
Standardized Solutions should follow the principle of aiming to minimize air interface overhead and impact to NW operation

4 options for model transfer/delivery under RAN2 discussion

3GPP Rel-19 Work Item status - AI/ML for Air Interface

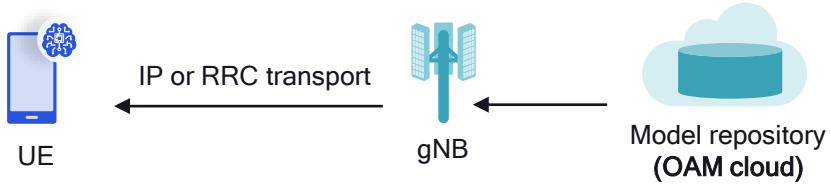


- Model repository OTT (outside the MNO)
  - User plane
  - Proprietary solution (no standards impact)



- Model repository inside the MNO
  - User plane or control plane (NAS signaling)
  - Trained, tested, and validated models are stored at the CN

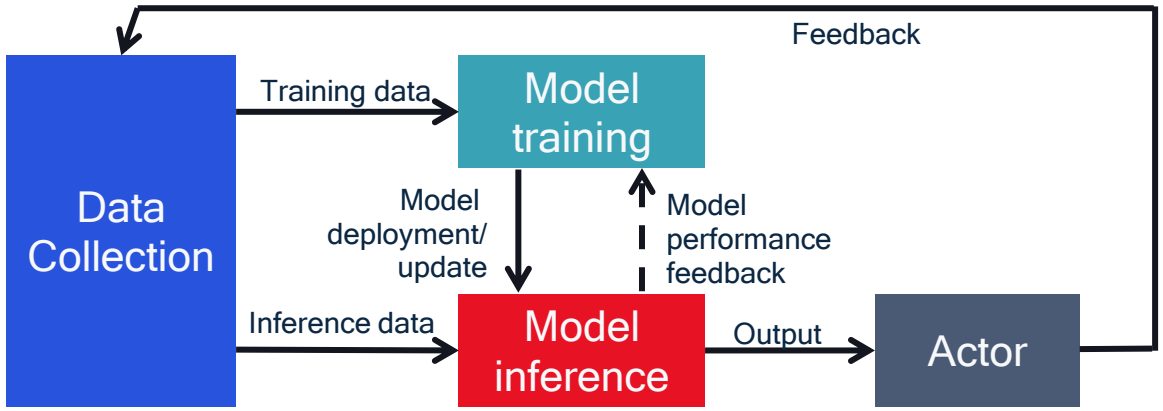
- Model repository in the RAN
  - Control plane (RRC) or user plane
  - Trained, tested, and validated models are stored at the gNB
    - Alternatively, the gNB can transfer parameters sets of standardized/known models (primarily for two-side models) [no need for testing and validation of parameter sets]



- Model repository in the OAM
  - Control plane (RRC) or user plane
  - Trained, tested, and validated models are stored at the OAM
    - Delivered via the gNB as in model repository in the RAN

AI/ML in NG-RAN (Rel-17 SI, Rel-18/19 WI in RAN3)

- **Rel-17** studied and identified high value use cases for AI/ML based optimizations for NG-RAN
- **Rel-18** use cases
  - Network Energy Saving
  - Load Balancing
  - Mobility Optimization
- **Rel-18** functional framework
  - Identified functions to support AI/ML tasks
  - Life Cycle Management out of scope
- **Rel-19** use cases
  - Coverage and Capacity Optimization (CCO)
  - Network Slicing
  - Mobility Optimization for DC
  - Split architecture support
  - Continuous MDT collection



AI/ML Framework defined in TR 37.817

- ML Model
  - A data driven algorithm by applying machine learning techniques that generates a set of outputs consisting of predicted information, based on a set of inputs
- Data collection
  - Provides input data to Model training and Model inference functions
- Model training
  - Performs the AI/ML model training, validation, and testing.
  - Also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) for model training
- Model inference
  - Provides AI/ML model inference output (e.g., predictions or decisions)
  - Also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) for model inference
- Actor
  - Receives the output from the Model Inference function and triggers or performs corresponding actions. Actor may trigger actions directed to other entities or to itself

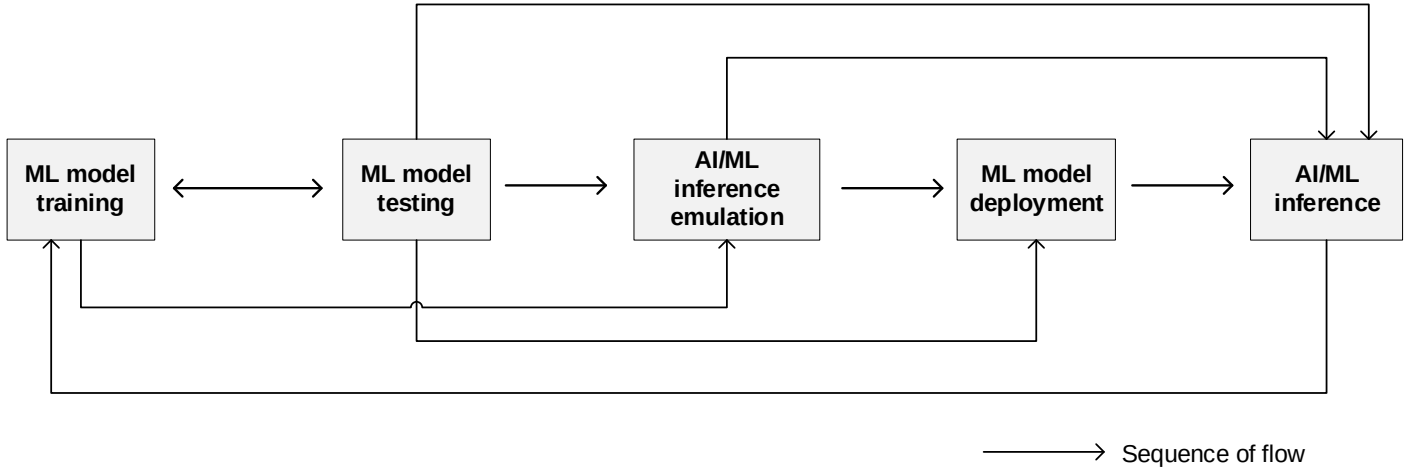
Enhanced Network Automation (since Rel-15 in SA2)

- **Network Data Analytics Function** (NWDAF)
  - CN function defined in TS 23.503
  - Only used for network slice analysis in **Rel-15**
- **Rel-16** enablers for network automation (eNA)
  - Data collection: provided by NFs of 5GCN (AMF, SMF, PCF, UDR, NEF), AFs, OAM and data repositories
  - Data exposure: on demand provisioning of analytics to NFs of the 5GCN, AF and OAM and data repositories
- **Rel-17** eNA phase 2
  - NWDAF framework enhancement
    - NWDAF decomposition; Multiple NWDAF architecture
  - Efficient data collection and management
  - UE application layer data collection
  - NWDAF assisted information for specific features
  - Rel-16 feature enhancements
- **Rel-18** eNA phase 3
  - Accuracy monitoring
  - Roaming support
  - ML model storage, removal and retrieval
  - Model Training logical function (MTLF) registration and discovery
  - Access to RAN (OAM/MDA)
  - FL among multiple NWDAFs
- **Rel-19** eNA phase 4
  - NWDAF enhancement
    - Support Vertical Federated Learning between NWDAF and AF
    - ML model training and provision to support UE assisted LMF-based AI/ML Positioning and NG-RAN assisted LMF-based AI/ML Positioning.
  - UE data collection
    - Status – LS sent that SA2 did not reach consensus on the feasibility, for any of the options outlined above, to meet RAN requirements

AI/ML workflow status

3GPP SA5 has specified a simplified AI/ML workflow

- The following AI/ML “use cases” are specified:
  - ML model training
  - ML model testing
  - AI/ML inference emulation
  - ML model deployment
    - ML model loading
    - ML model registration
  - AI/ML inference
    - AI/ML inference performance management
    - AI/ML update control
    - AI/ML inference capabilities management
    - AI/ML inference capability configuration management
    - Executing AI/ML inference



Conclusions

Main take-aways for AI/ML in 3GPP

- **AI/ML** is **pervasive** and its reach in 3GPP is ever increasing
  - There are related projects in 3GPP across Working Groups (WGs) in RAN and SA
- **3GPP** currently has **no single, unified framework** to deal with AI/ML
  - Part of the problem is that AI/ML has flourished bottom-up in 3GPP
    - Each WG developed its own understanding of AI/ML and studied how it can serve its purpose in the best possible way
  - AI/ML was not in scope at the inception of 5G
    - Ongoing study in SA (led by Deutsche Telekom) on 3GPP AI/ML Consistency Alignment (TR 22.850)
- **Standardization impact** of AI/ML is somewhat limited
  - **AI/ML models** themselves are **not standardized**
  - **Governance**, **performance monitoring and requirements** are the main drivers for AI/ML standardization
- AI/ML is expected to play an important role in **6G**
  - More in next slide…

What does 6G being **AI/ML native** really mean?

A personal view

- Formal definition of the meaning of "**AI/ML native**" not agreed
  - Although everybody talks about it…
- Some possible interpretations:
  - AI/ML will be available from **Day 1 of 6G**
    - Application of AI/ML based implementations possibly on top of conventional, non-AI/ML based implementations
    - Specific air interface extensions may be required to facilitate the inference process
  - Specifications could leave some leeway for **AI/ML based implementations**
    - Instead of having, e.g., a rigid set of parameter values, and corresponding, e.g., rigid UE behavior…
    - Let AI/ML engines come up with a best data-driven fit-to-purpose parameterization for certain functionalities
- 5G-Advanced has assumed offline training
  - Should 6G consider the possibility for online training of AI/ML models?
  - Federated learning has also been considered in the context of eNA and may further propagate elsewhere

Thank you

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog