



High Level Overview of AI/ML work in 3GPP

Dr. Juan Montojo

Singapore

June 12, 2026

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.



Outline

- Scope of standardization of AI/ML
- Personal high-level assessment of AI/ML in Cellular Networks
- Former & Ongoing AI/ML standards activities in 3GPP
- Lifecycle of UE models
- Data collection and Model transfer/delivery architectures considered
- 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)
 - Role of **AI agents** in the 6G core
- **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 behavior** is a key concern

AI/ML in Cellular Networks

Personal 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...
- Procedures simplification & interworking via agents

• Concerns:



- Testability and interoperability
- Consistent behavior
- Feasibility of two-sided AI/ML models
- Agentic protocols not deterministic & tend to be “chatty”
- “Finding problems for solutions” because of the AI hype

• Not so good for:



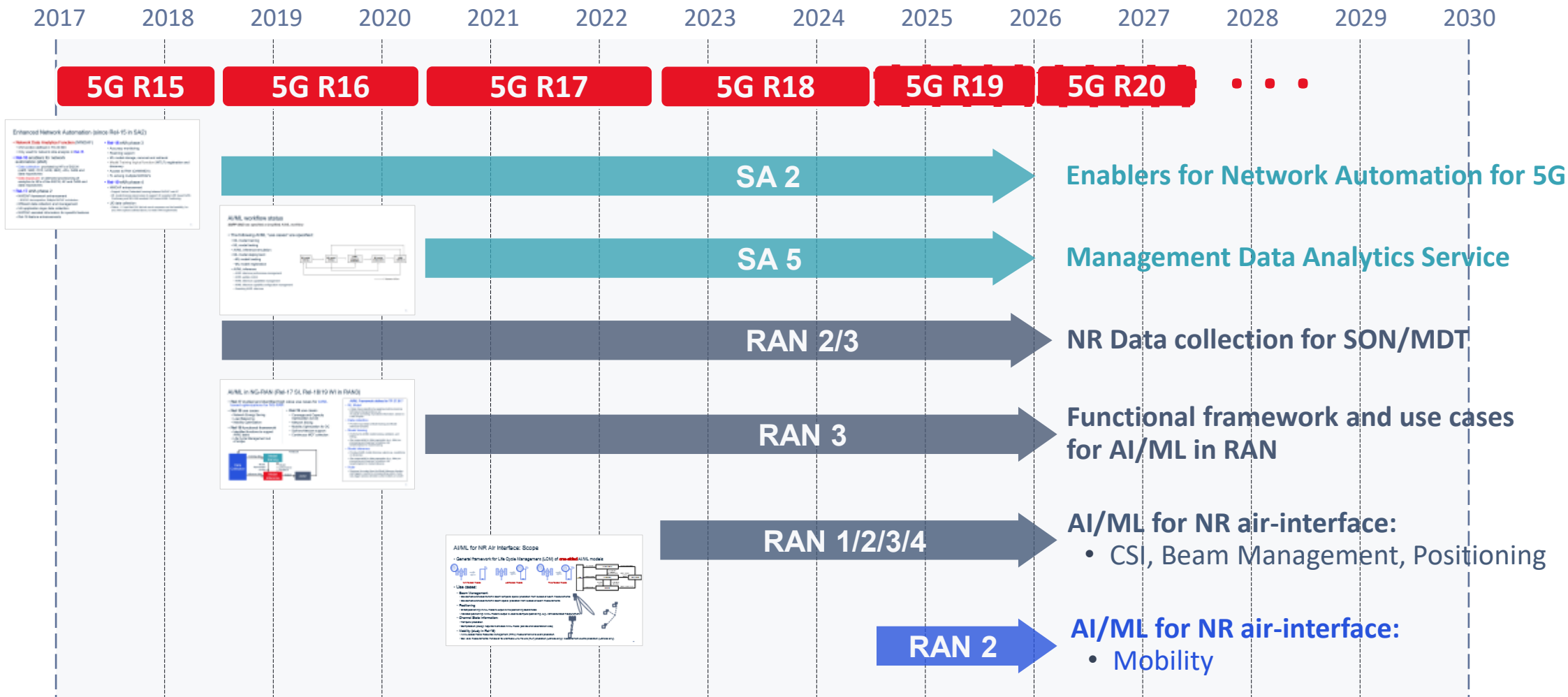
- Solving (complex) problems with analytical solution
- Complexity and Energy consumption (NW & UE)

• Controversies



- Data ownership and data sharing
- Standards mechanism for data collection, model transfer/delivery...
- Agentic Core vis-à-vis traditional Core

AI/ML standards activities in 3GPP before Rel-20



SON: Self Organizing Networks
MDT: Minimization of Driving Tests

RAN: Radio Access Network
CSI: Channel State Information

AI/ML Activities on AI/ML in Rel-20 (1/2)

5G-Advanced projects

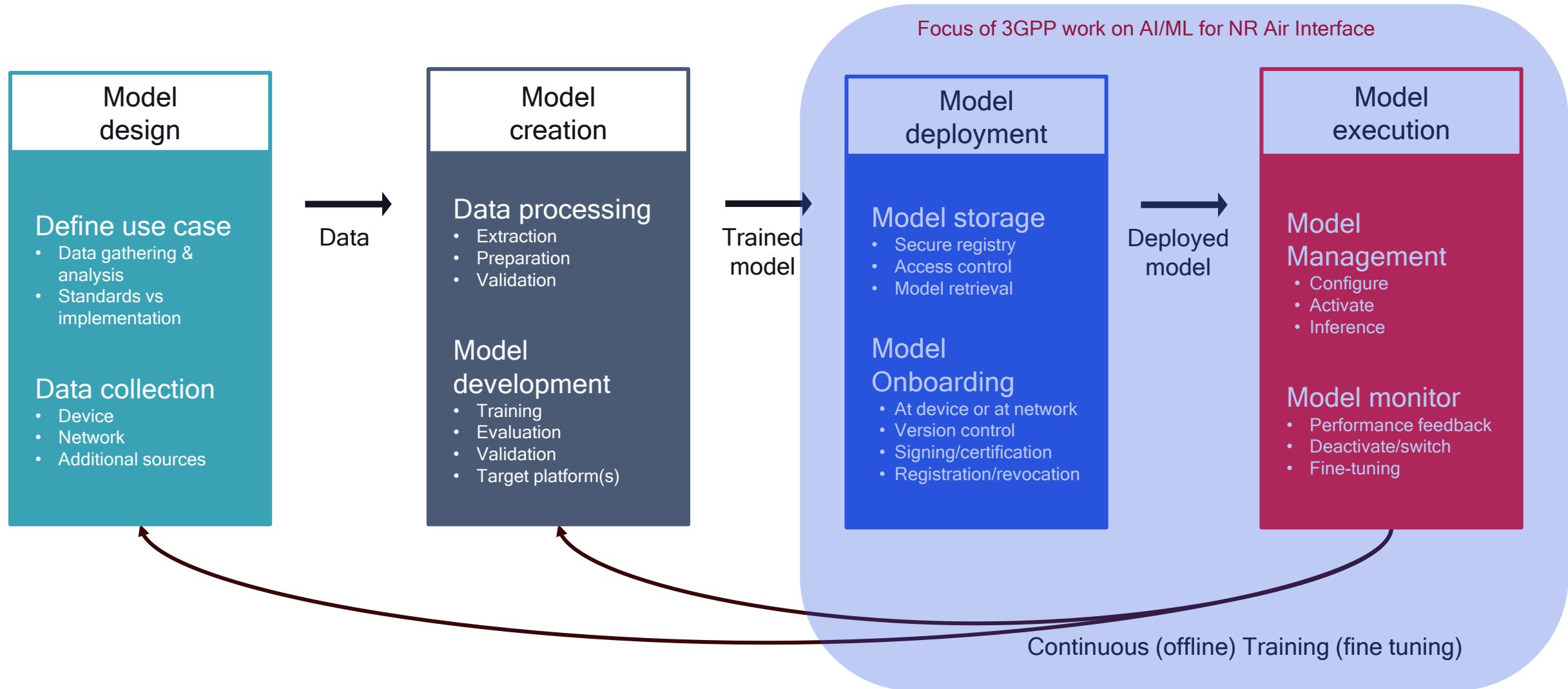
- **AI/ML for NR Air Interface Phase 2** [RAN1]:
 - Specification of AI/ML-based CSI compression (two-sided model); Related RAN4 “trial” to enable cross-node training
- **AI/ML for Mobility** [RAN2]:
 - AI-aided RRM measurement & event predictions based on cell-level prediction; Associated LCM procedures;
- **Mobile AI/ML traffic Characteristics** [RAN2]:
 - As part of “XR for NR Phase 4” project and in close collaboration with SA4
- **AI/ML for NG-RAN Phase 3** [RAN3]:
 - Multi-hop UE Trajectory, Intra/Inter-CU Lower-layer Triggered Mobility (LTM)
- **AI/ML enhancements** [SA2]:
 - Study of transfer of data over UP for UE data collection for AI/ML for NR air interface operation with UE-side model training. AI/ML-assisted User Plane (UP) traffic pattern and behavior analysis to support efficient performance of UP
- **Ultra-low-rate voice codec** [SA4]:
 - Codec itself likely AI/ML-based. Ongoing discussion on Dataset for training and model comparison/selection procedure

AI/ML Activities on AI/ML in Rel-20 (2/2)

6G Studies

- **AI/ML for 6G Air Interface:**
 - Many Use cases being considered in addition to those considered for 5G-Advanced, e.g.,
 - Relaxation of protocols parameterization to enable AI/ML-based selection based on contextual awareness
 - Low-overhead reference signal with AI/ML receiver; AI/ML-based sequence design for control blocks (3-11bits), random access preambles, sounding signals
 - AI/ML for sounding reference signals: low-overhead, power-imbalance compensation; AI/ML for early contention resolution in random access
 - Power control related: pathloss prediction, UL closed-loop power control
 - AI/ML based non-linearity handling at transmitter or receiver
 - Digital Twin construction related for e.g., ISAC
- **Agentic AI in 6G** about using Agentic protocols for:
 - Operators to offer a secure platform for AI agents
 - Simplified OAM enabling natural language requests into command line interfaces for different vendors
 - Simple access to NW services for 3rd party LLMs or AI agents using agentic protocols
 - User/Application to access MNO/NW services (e.g., QoS, sensing)
- **“Data plane”** being discussed as part of 6G studies
 - Potential concept on top of the existing Control Plane (CP) and User Plane (UP)
 - Visibility, termination point and “data consumer” being discussed

ML Workflow - Life cycle for UE models



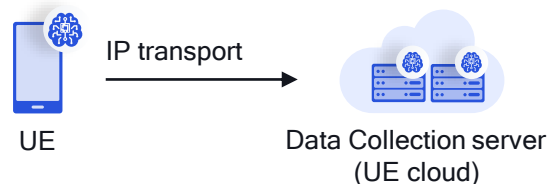
Data: Considered Options for data collection

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

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 rotating, 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



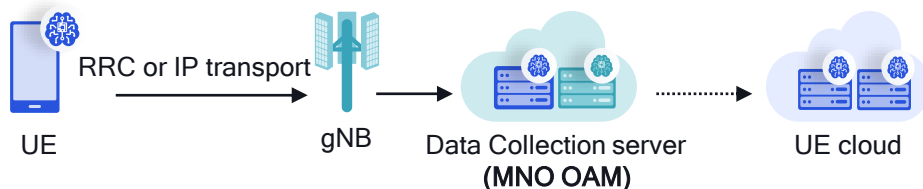
- **Option 1a - Data collection server OTT (outside the MNO)**
 - User plane
 - Proprietary solution (no standards impact)



- **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



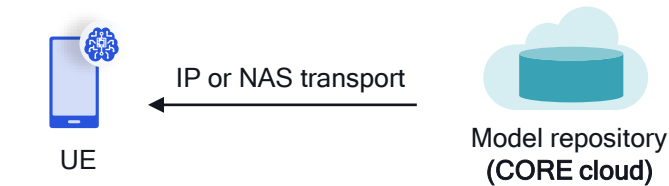
- **Option 3 - Data collection server in the RAN (OAM)**
 - Control plane (RRC) or user plane [FFS]
 - UE collects training data and transfers it to the OAM
 - OAM transfers the data to a UE-side model training/OTT server

Models: Considered Options for model transfer/delivery

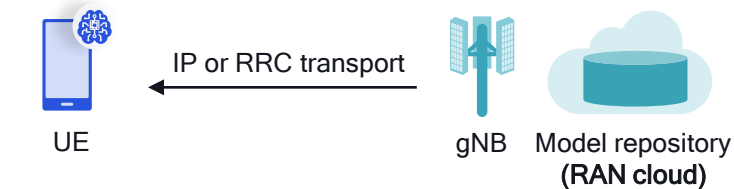
3GPP Rel-19 Work Item status - AI/ML for Air Interface



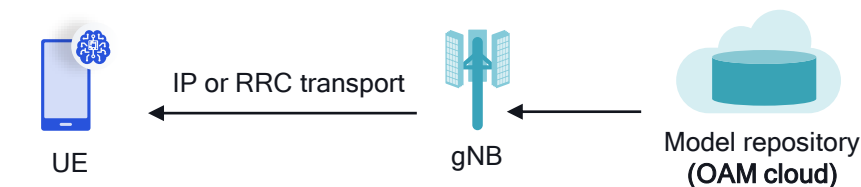
- Model repository OTT (outside the MNO)
 - User plane
 - Proprietary solution (no standards impact)



- 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

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
 - Technical Report on 3GPP AI/ML Consistency Alignment ([TR 22.850](#))
- **Standardization impact** of AI/ML is somewhat limited [not bad as over-specifying would slow-down AI]
 - **AI/ML models** themselves are **not standardized**
 - **Model governance, performance monitoring and requirements** are the main drivers for AI/ML standardization
 - How agentic AI makes its way into the 6G Core is currently debated
- AI/ML is expected to play a key role in **6G**

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated.
Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

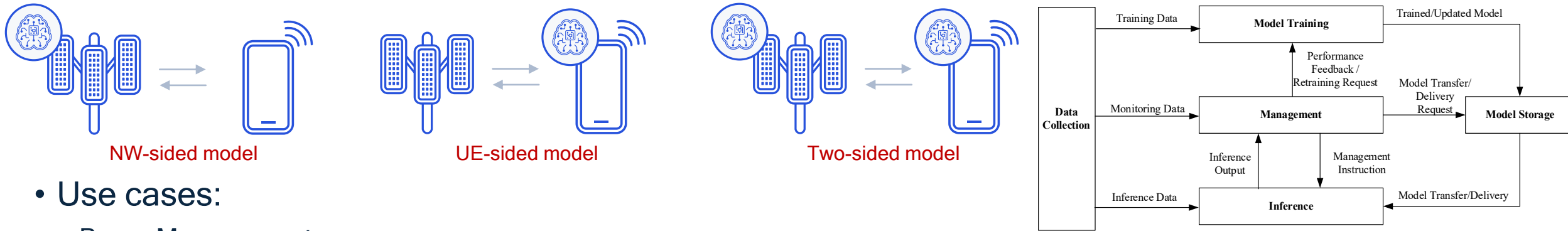
Follow us on: [in](#) [X](#) [@](#) [v](#) [f](#)

For more information, visit us at [qualcomm.com](https://www.qualcomm.com) & [qualcomm.com/blog](https://www.qualcomm.com/blog)



AI/ML for NR Air Interface: Scope

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



- 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

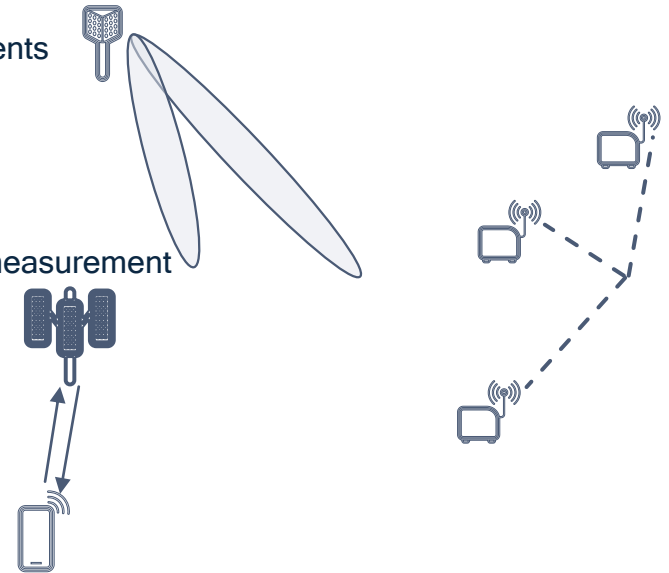
- Direct positioning: AI/ML model's output is the positioning coordinates
- Assisted positioning: AI/ML model's output is used to compute positioning, e.g., refined/curated measurement

- Channel State Information:

- Temporal prediction
- Compression (study): requires two-sided AI/ML model (device and base-station side)

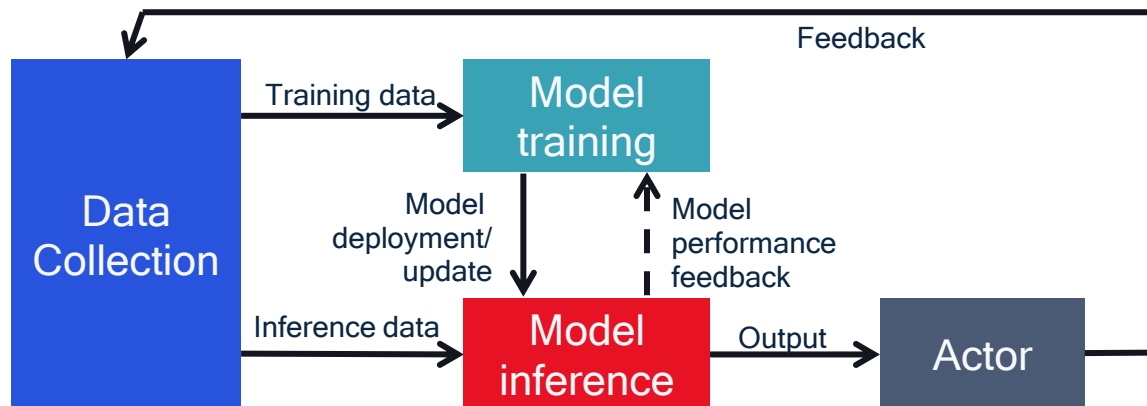
- Mobility (study in Rel-19)

- 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)



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

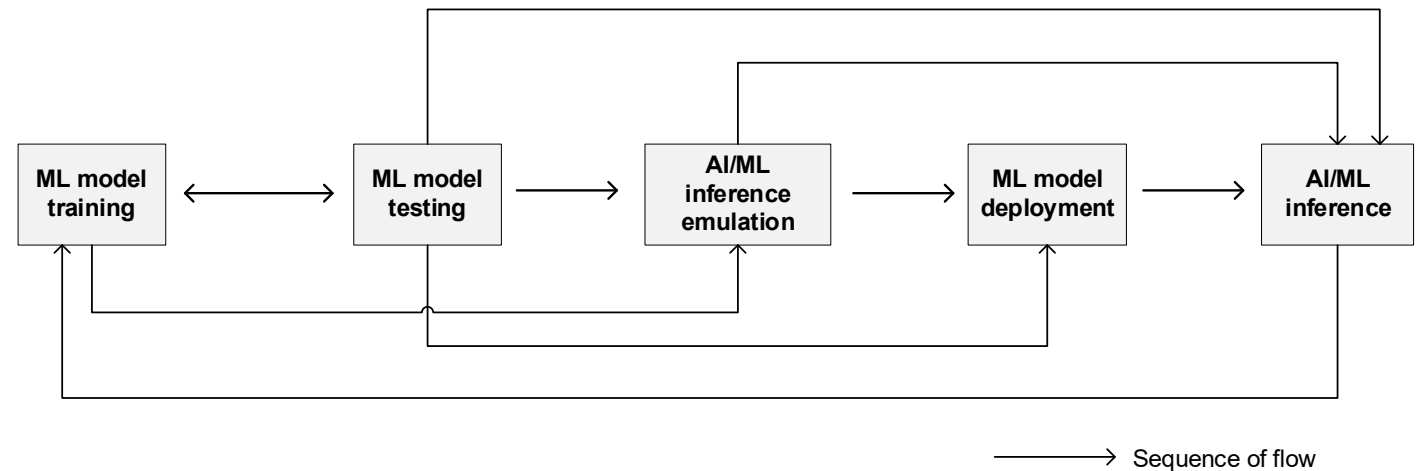
- NWDAF enhancement
 - Support Vertical Federated Learning between NWDAF and AF
 - ML model training and provision to support UE assisted LMF-based AI/ML Positioning and NG-RAN assisted LMF-based AI/ML Positioning.
- UE data collection
 - Status - LS sent that SA2 did not reach consensus on the feasibility, 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



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