



Study on 5G System Support for Application AI/ML-based Services (FS_5GAIML)

Prepared by OPPO, Samsung

Rel-18 SA1 Study on AI/ML Model Transmission (AMMT)

The study of the use cases and the potential performance requirements for 5G system support of Application Layer Artificial Intelligence (AI)/Machine Learning (ML) model distribution and transfer (download, upload, updates, etc.), and identifies traffic characteristics of AI/ML model distribution, transfer and training for various applications, e.g. video/speech recognition, robot control, automotive, other verticals.

The aspects addressed in SA1 study include:

- AI/ML operation splitting between AI/ML endpoints;
- AI/ML model/data distribution and sharing over 5G system;
- Distributed/Federated Learning over 5G system.

Proposed Rel-18 SA2 study is to focus on enabling the intelligent transmission support for Application Layer AI/ML Services via 5GS assistance for AI/ML model distribution, transfer, training and real-time control for various applications as defined by SA1 Rel-18 AMMT requirements, e.g. video/speech recognition, robot control, automotive etc.

***NOTE:** SA2 has been focused on AI based analytics work to support network operation automation, but not much on how the service providers could leverage 5GS as the platform to assist the implementation and transmission to support Application AI/ML operation.*

Widely used AI services

- Image recognition

The AI model has been widely used for image recognition. Different AI model applies to different scenarios.

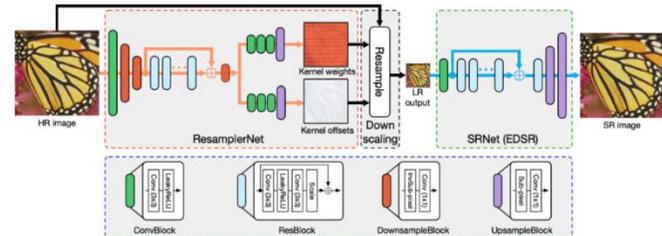
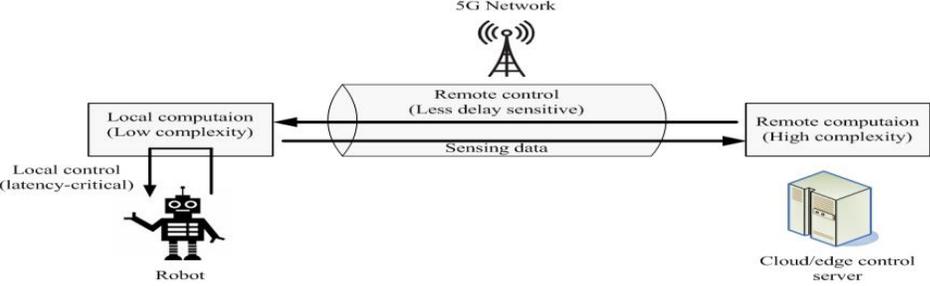
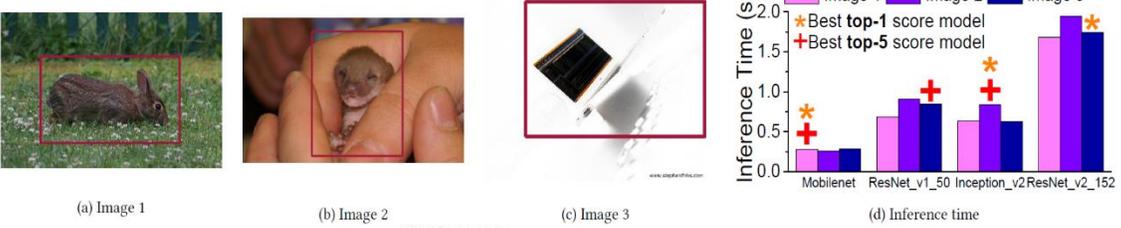
- Real-time control (e.g. robot)

The real-time control helps to realize the unmanned services like auto-driving, automatic industry

- Video streaming

AI model can be trained to realize/replace some functionalities of video/voice coding/encoding

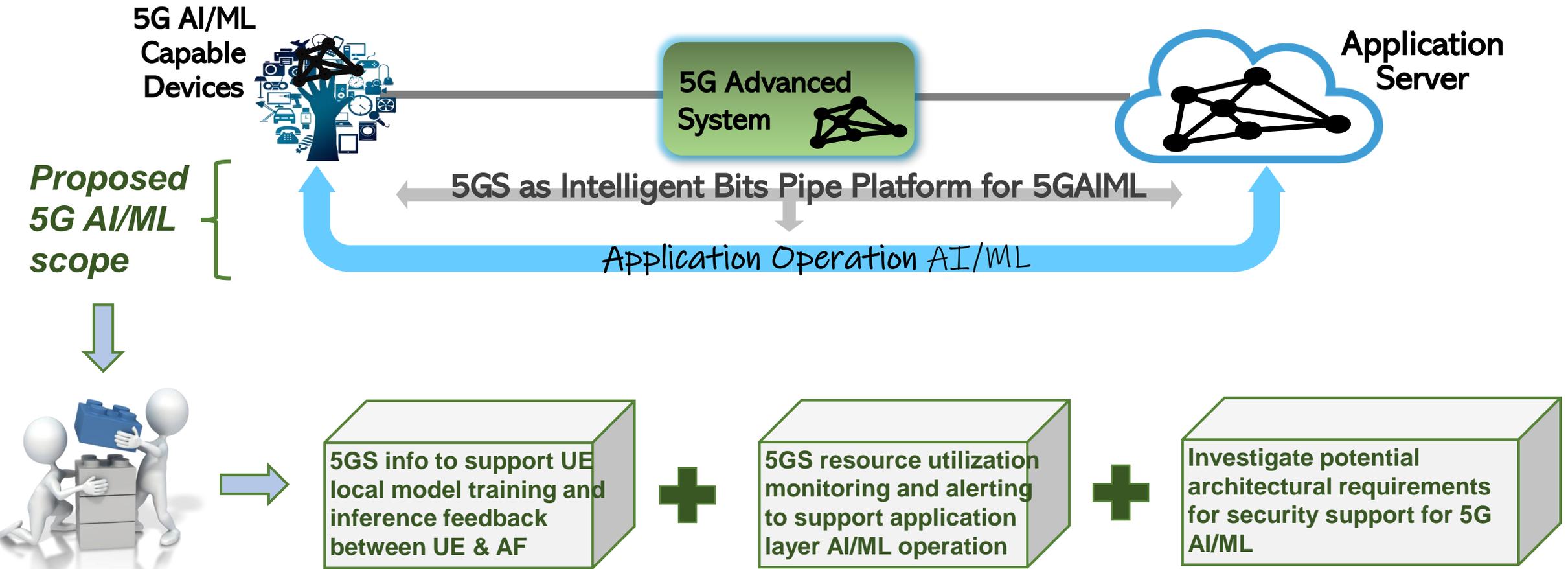
.....



Why eNA_PH3 and 5GAIML are different?

Functionality/Operation	5G Application Layer AI/ML	NWDAF-based Network Automation
Model Distribution and Control of the AI/ML Operation	Application Provider Control with the 5GS to provide the transmission support (e.g. via require API and possible middleware support)	Network Operator Control with the 5GS to provide the transmission support (i.e. no need for API or middleware support)
Training/Inference Functions	Application Layer (between UE and Application server) functionality and “not” part of the 5GC – i.e. communication support for AI/ML operation is different than eNA	Network Layer (between UE and 5GC) with functionality within the 5GC
Storage management support for model distribution and aggregation	Application Layer Control <u>NOTE:</u> Application provider may subscribe to 3GPP operator’s cloud services for storage support but <u>not</u> for controlling the operation for model distribution and aggregation	Network Layer Control
Security and UE Privacy Control	Application Provider’s responsibility to ensure UE’s privacy is protected from Network Operator even though the transport of the analytic data is over 5GS and requires coordinated support from 5GS for the transport protection aspect.	Network Layer full Control
Analytic AI/ML related information transfer	The contents of the application layer AI/ML related information transfer is transparent to Network layer. <u>NOTE:</u> However, the transport mechanism may require specific handling by the 5G system, e.g. QoS and bandwidth control for UL/DL.	Network layer is aware of the analytic information transfer
Charging/Accounting Consideration	Application layer AI/ML operation may introduce a new type of accounting model to 5G system to support its application service provider’s customers, e.g. rather than charging, it could be rewarding	Nothing changes
Impacts to 5GS	Application Provider leverages 5GS as the “intelligent transmission platform/pipe” to support the Application Layer AI/ML operation which requires 5GS to expose and to support UE’s status and network resource monitoring (see further details in the subsequent slides), however, no Application Layer intelligence built-in to the 5GS network functions.	NWDAF and its consumers (e.g. 5GC NFs) as well as UEs are enhanced to support network layer automation analytic.

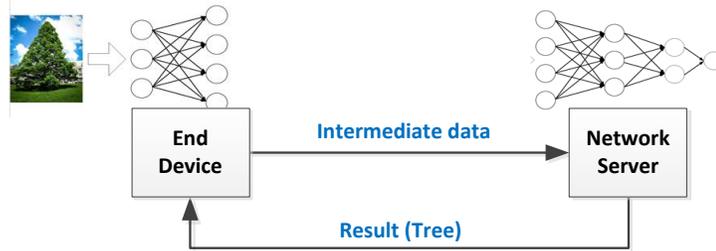
Objective#1: Study 5GS Architectural and Functional extensions to support Application 5G AI/ML according to SA1 Requirements



Objective#2: QoS and Policy enhancement to support AI/ML services & transmission with 5GS assistance

■ Enhancement to support asymmetric QoS

- For many AI enabled services, like Augmented Reality, the round-trip latency mostly matters. e.g. figure below shows the size of UL intermediate data can be much larger than DL.
- Asymmetric QoS helps to properly schedule the AN/CN resources for data transmission as long as the round trip latency performance is maintained



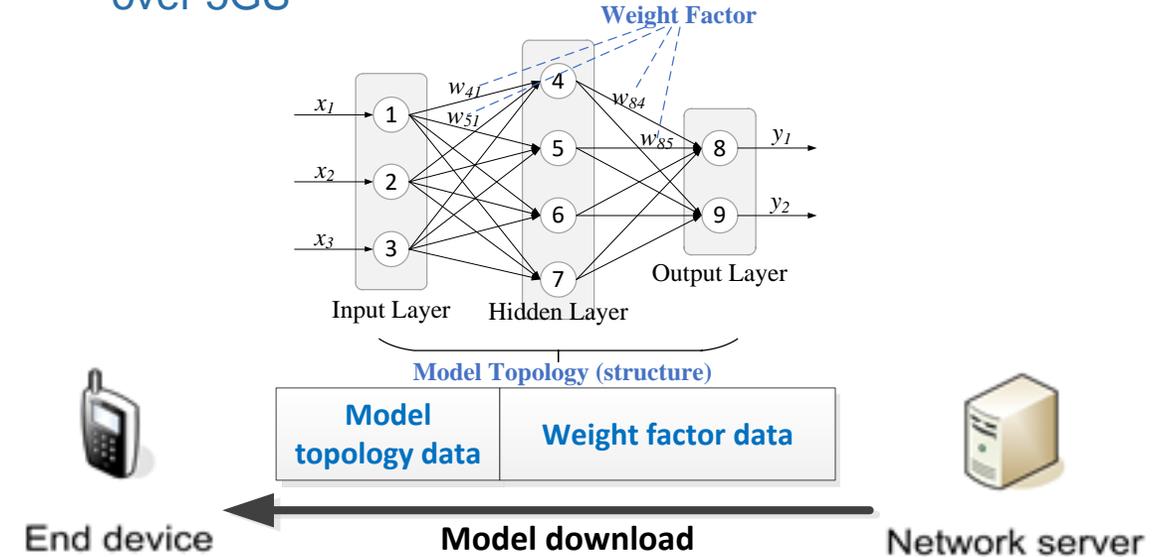
■ New 5QI for AI/ML service transmission

For different variants of AI/ML services that employ AI model splitting, model transmission, federated/distributed learning, new 5QI may be needed to provide asymmetric QoS support.

■ Application initiated QoS modification

Due to anticipated network event or environmental condition, application may need to adjust the QoS and/or data rate transmission to adapt to the changes in real-time – support real-time QoS modification.

■ Content-type aware QoS for model transmission support over 5GS



- AI/ML model is composed of **model topology data** and **weight factor data**, where the topology data needs higher transmission reliability for some use cases to ensure the model can be utilized accurately. AI/ML model data may be tagged to enable ML flow classification to enable differentiated QoS management.
- **ML flow classification** mechanism is to derive the characteristics of the ML flow (e.g. the size of the previous N packets, source IP address, protocols or flow arrival interval) rather than checking the payloads of the packets so that encrypted flows can be classified and therefore proper QoS can be applied.

■ Potential policy control and charging

Determine whether any impact to existing policy control and aspects (e.g. lower charging rate for AI/ML transmission)

Objective#3: 5GS information exposure extensions to support Application 5GAI/ML operation

■ The environments of communication (e.g. QoS, Location) could lead to a change of the AI/ML operation (~~e.g. bandwidth adjustment~~) and affect user experience. It is proposed to study what analytical information exposed by 5GC to assist AI/ML analytic operation

- The information exposed by 5GS can be **real time analytics/alert of situation of the communication, for example:**
 - prediction and alert the 3rd party of the GBR that cannot be guaranteed and how soon it could happen (e.g. congestion)
 - prediction and alert the 3rd party that UE is moving towards a certain area, e.g. in 30 seconds

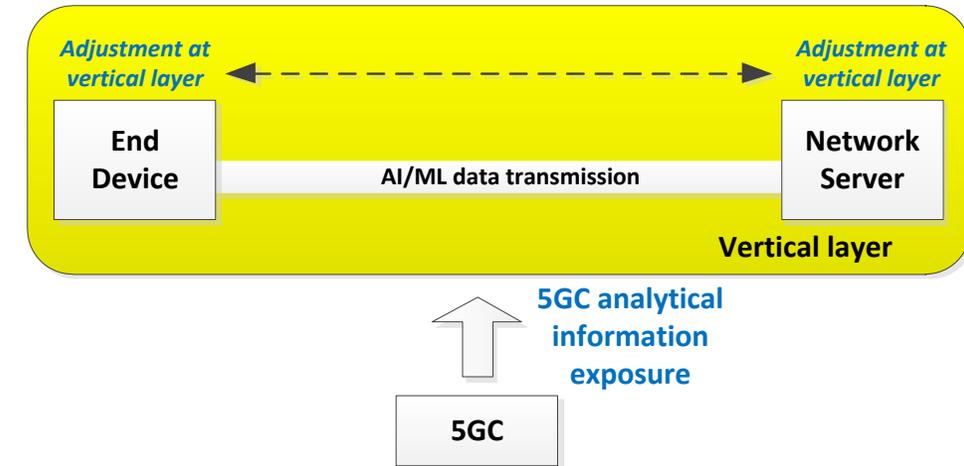
NOTE-1: For some applications, the forewarning needs to notify the 3rd party in realtime to react.

NOTE-2: As the geographic location of the UE could influence the proceed of AI/ML operation may require geo-fencing support

e.g. As shown in table, there is a mapping between splitting point (operational behavior) and UL data rate (communication quality). Splitting point could be changed when detecting the change of the UL data rate (For model spiting, see slide 11 for more background info)

Split point	Required UL data rate (Mbps)
Candidate split point 0 (Cloud-based inference)	36
Candidate split point 1 (after pool1 layer)	65
Candidate split point 2 (after pool2 layer)	41
Candidate split point 3 (after pool5 layer)	4.8
Candidate split point 4 (Device-based inference)	N/A

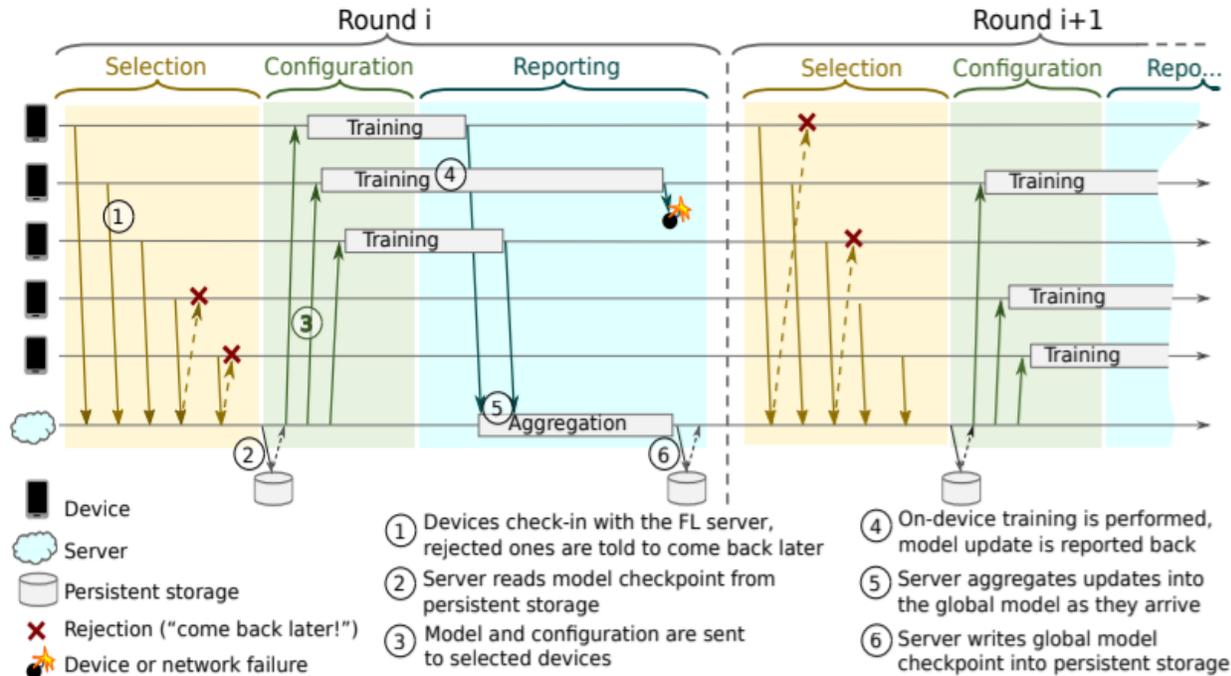
Excerpt from TR 22.874



NOTE: Operational splitting is a common and effective technique to support AI/ML analytic. The intent of operational splitting is to offload computation-intensive and energy-intensive training parts to Network Endpoints, whereas leave the privacy-sensitive and delay-sensitive training parts at the End Device.

Objective#4: Co-operative AAI/ML Operation via assistance from 5GS to support different models of Federated Learning

Federated Learning allows for collaborative training of a single global model by multiple local learners without exchanging or transporting data anywhere!



FL imposes additional challenges to be operated over the mobile network that need to be addressed, e.g. device selection and aggregation, training optimization, transmission scheduling, QoS exposure, etc.

$$\sum \text{Gradient uploading latency} + \text{Federated aggregation latency} + \text{global/local model downloading latency} \leq \text{GPU computing time at device}$$

- Support for various models of Federated Learning as described in SA1 Rel18 AMMT requirements.

Conclusions

- ✓ FS_5GAIML is a distinct AI/ML based feature study from the current eNA feature with distinct objectives as described in this presentation.
- ✓ 5GAIML feature enables 5G system to become an intelligent pipeline/platform to support Application Layer AI/ML based operation which could open up new business and revenue opportunity for 5G mobile operators' to provide services for their application service provider customers/partners.
- ✓ 5GAIML feature also enables more effective business operation for application service provider who leverages 5G system capability to support its Application AI/ML operations.

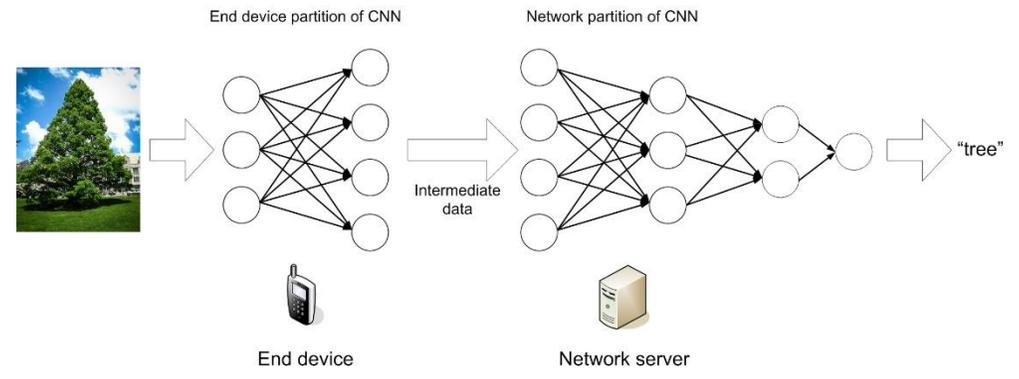


5G Application AI/ML support should be an important feature for the 5G Advanced System

Backup and Reference Slides

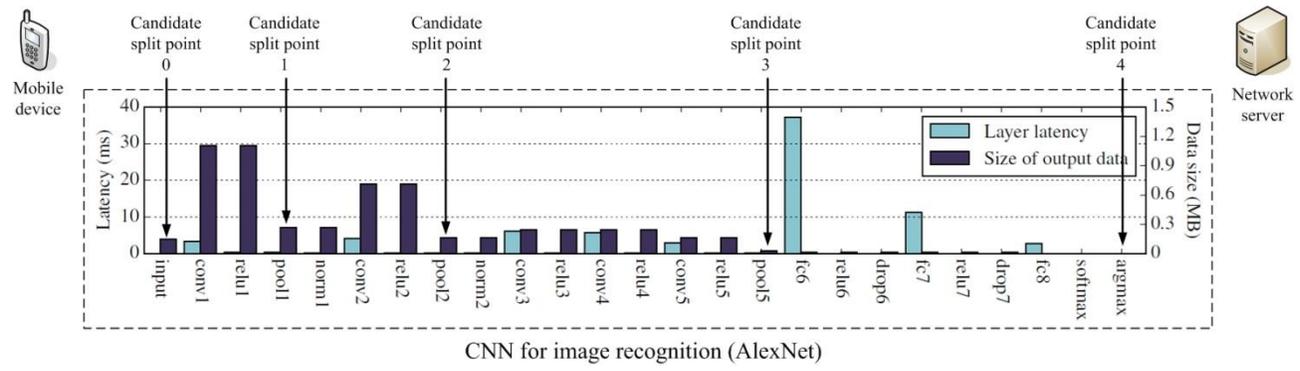
Model Splitting

✓ The device executes the inference up to a specific CNN layer and sends the intermediate data to the network server. The network server runs through the remaining CNN layers.



✓ The split AI/ML image recognition algorithms can be analyzed based on the computation and data characteristics of the layers in the CNN. The intermediate data size transferred from one CNN layer to the next depends on the location of the split point. Hence, the required UL data rate is related to the model split point and the frame rate for the image recognition.

Required UL data rate for different split points of AlexNet model for video recognition @30FPS



Split point	Approximate output data size (MByte)	Required UL data rate (Mbps)
Candidate split point 0 (Cloud-based inference)	0.15	36
Candidate split point 1 (after pool1 layer)	0.27	65
Candidate split point 2 (after pool2 layer)	0.17	41
Candidate split point 3 (after pool5 layer)	0.02	4.8
Candidate split point 4 (Device-based inference)	N/A	N/A

Main Use Cases on AI Imaging Recognition – Business Opportunity for Mobile Network Service Offerings for OTT Application

✓ **Logo Detection in Social Media Analytics**

- Social listing (via social media posts) and brand monitoring support for marketing research and competitive analysis
- Business are using logo detection to calculate ROI from sponsoring sport events or to define whether their logo was misused

✓ **Medical Image Analysis**

- Software powered by deep learning models help medical practitioners (e.g. radiologists) to deal with huge workload of interpreting various medical images: computed tomography (CT), ultrasound scans, magnetics resonance imaging (MRI) or x-rays etc.
- Medical images account for ~90%of all medical data.

✓ **Visual Product Sales & Marketing**

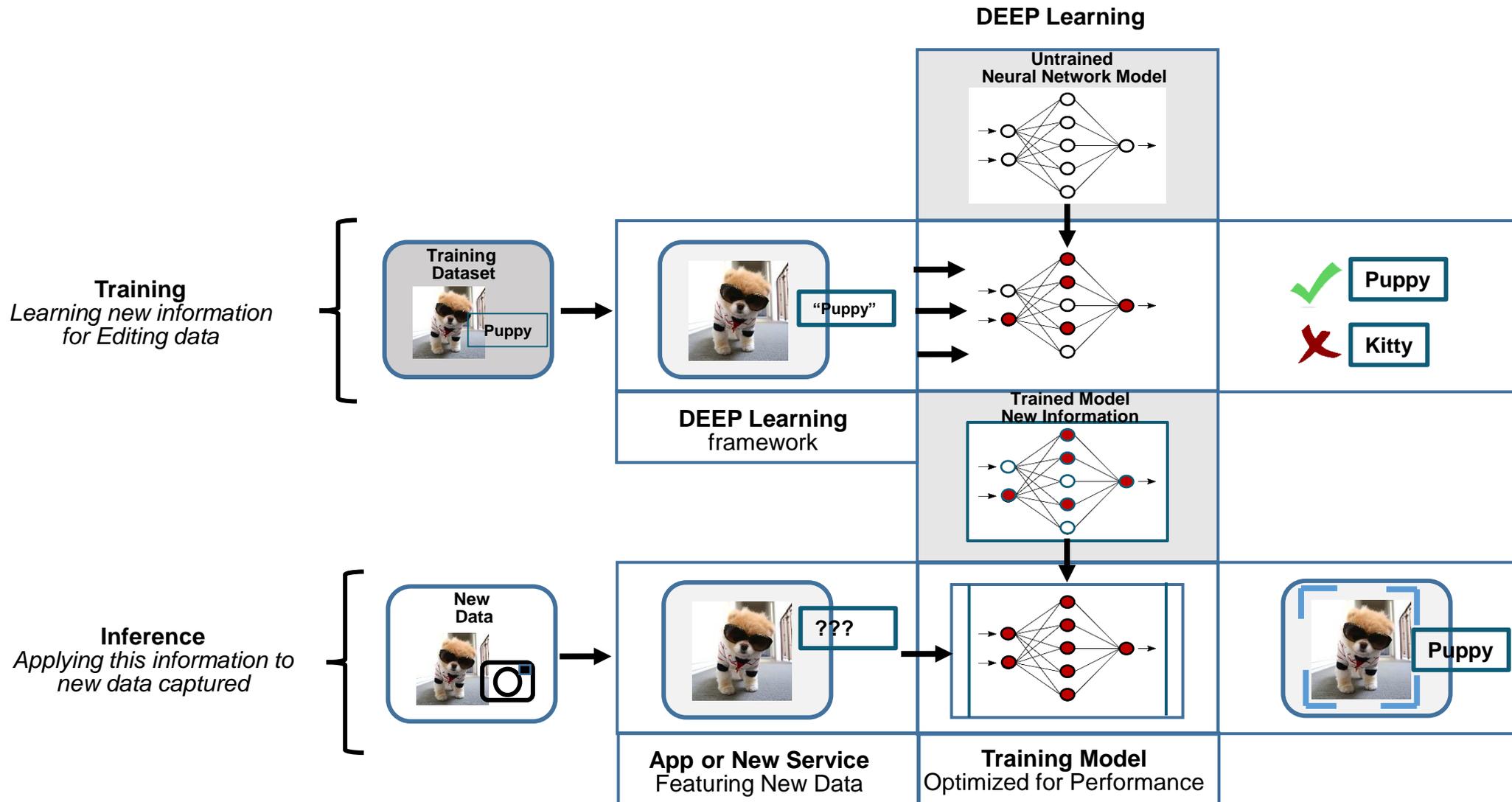
- Improvement for online shopping experience to get detailed descriptions which could lead to instant order.
- Empowering marketing research opportunity to identify customer's interests on fabric,style, color and to enable in-store personalization and product promotion.

✓ **Facial Recognition enabling Enterprise services and Private/Public Safety to streamline routine operations**

- Additional biometric technology to enhance efficiency and security on routine operations.

✓ Plus many more

AI Training vs. AI Inference



Machine Learning as a Process

- Define measurable and quantifiable goals
- Use this stage to learn about the problem

Define Objectives

Model Deployment

Data Preparation

- Normalization
- Transformation
- Missing Values
- Outliers

Model Evaluation

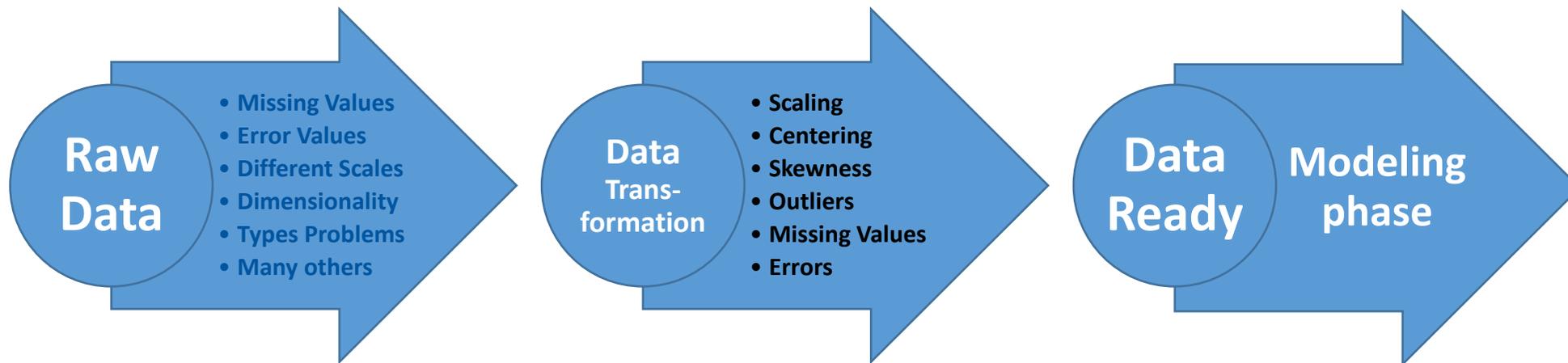
Model Building

- Data Splitting
- Features Engineering
- Estimating Performance
- Evaluation and Model Selection

- Study models accuracy
- Work better than the naïve approach or previous system
- Do the results make sense in the context of the problem

ML as a Process: Data Preparation

- Needed for several reasons
 - Some Models have strict data requirements
 - Scale of the data, data point intervals, etc
 - Some characteristics of the data may impact dramatically on the model performance
- Time on data preparation should not be underestimated

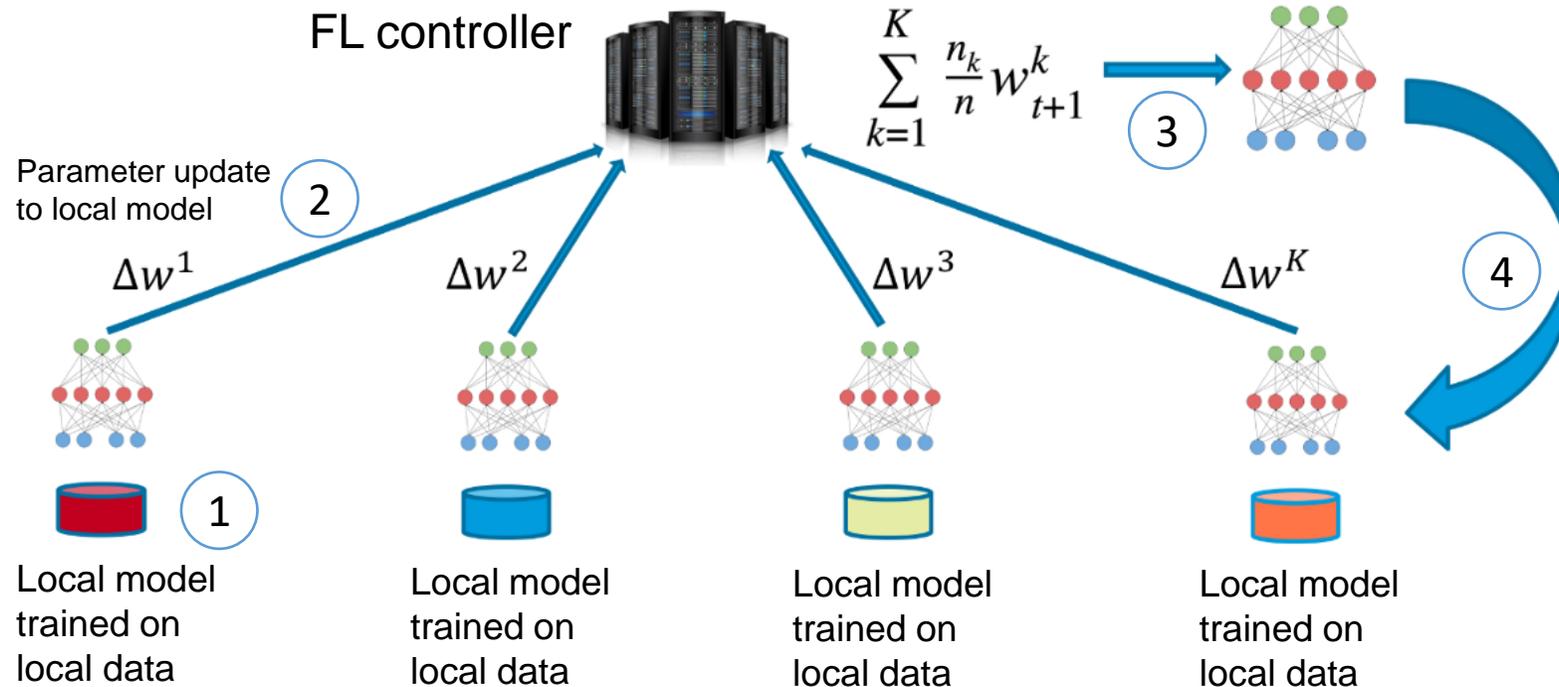


ML as a Process: Model Building

- Data Splitting
 - Allocate data to different tasks
 - model training
 - performance evaluation
 - Define Training, Validation and Test sets
- Feature Selection (Review the decision made previously)
- Estimating Performance
 - Visualization of results – discovery interesting areas of the problem space
 - Statistics and performance measures
- Evaluation and Model selection
 - The ‘no free lunch’ theorem no a priory assumptions can be made
 - Avoid use of favorite models if NEEDED

What is Federated Learning?

Source: https://miro.medium.com/max/1400/1*HaH611vAy2eB1e42vz3X4g.png



- The backhaul on a wireless network is very congested: we don't want to move data over it
- Federated Learning allows for collaborative training of a single global model by multiple local learners without exchanging or transporting data anywhere!

The localized (local data trained model) vs. global model tradeoff is an open research problem

In each *round* of training:

1. Each local learner trains a *local* version of the ML model on its *local* data
2. Only the local model parameter *updates* are sent up to the controller (this reduces the load on the links to the controller)
3. The controller performs a weighted average of the parameter updates to get the new set of parameters for the global model
4. The controller sends the new global model back down to each local learner as its initialized local model for the next round of training

Thanks!
