Networked AI Learning

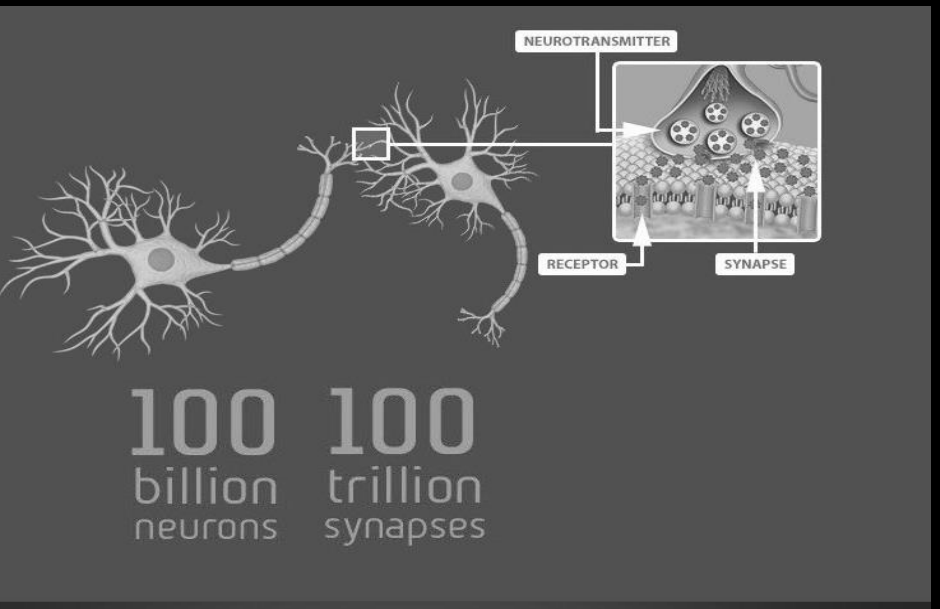
A non-federated learning approach

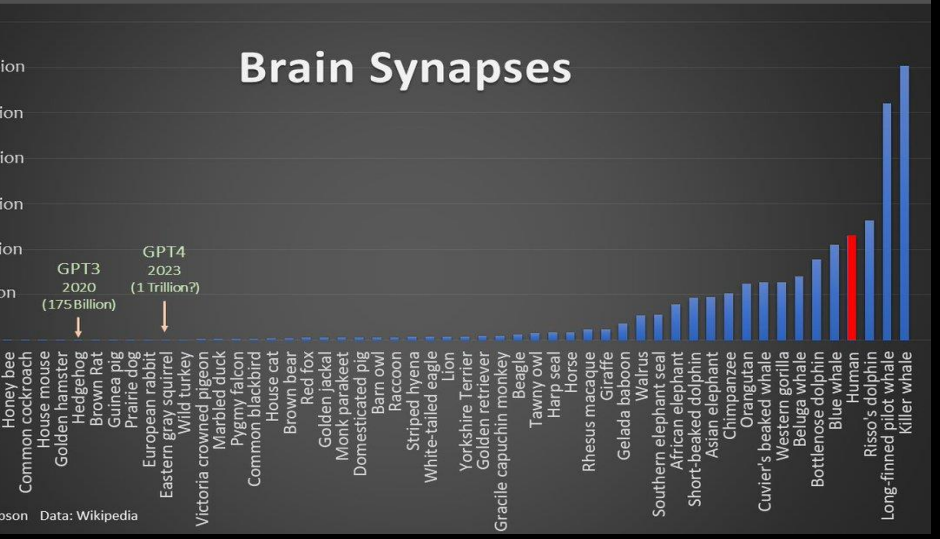
Dr. Wen Tong

CTO, Huawei Wireless

International Workshop on Federated Learning and Foundation Models for Multi-Media
Keynote
July 15th 2024
Niagara Falls, CANADA

AI Model Complexity (1)



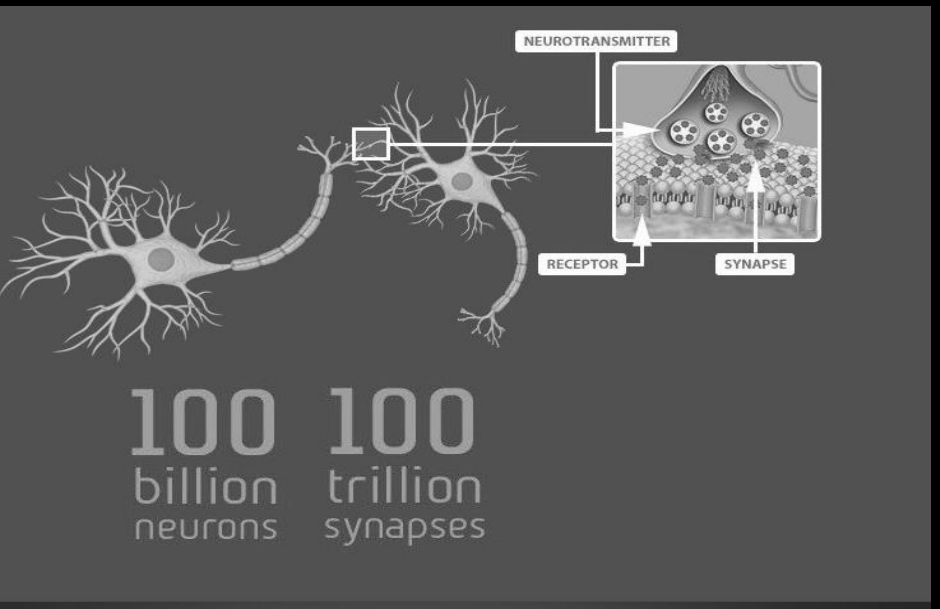


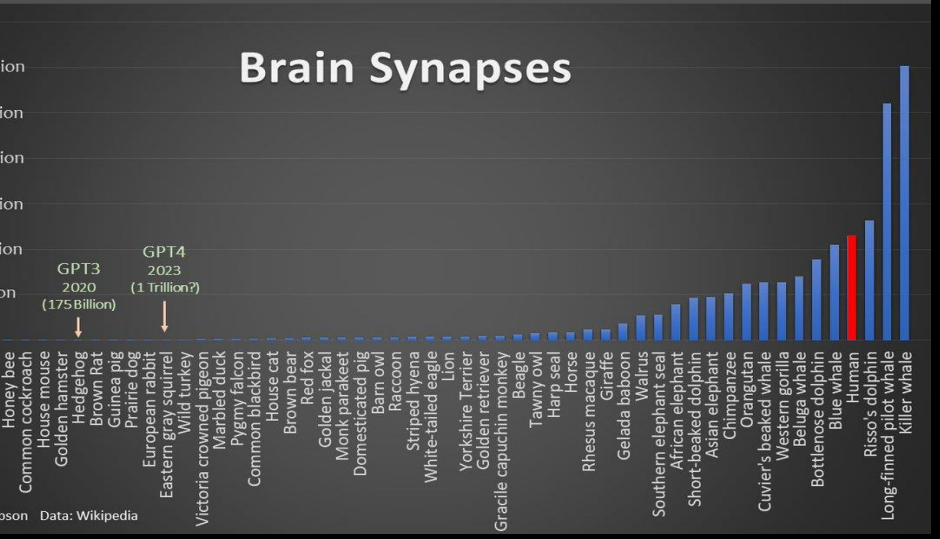
Today

Training	Inferencing		
1T 100B	1T 500B		
10B 100M	200B 100B		
10M	10B 100M		

AI Model Complexity (2)

2030

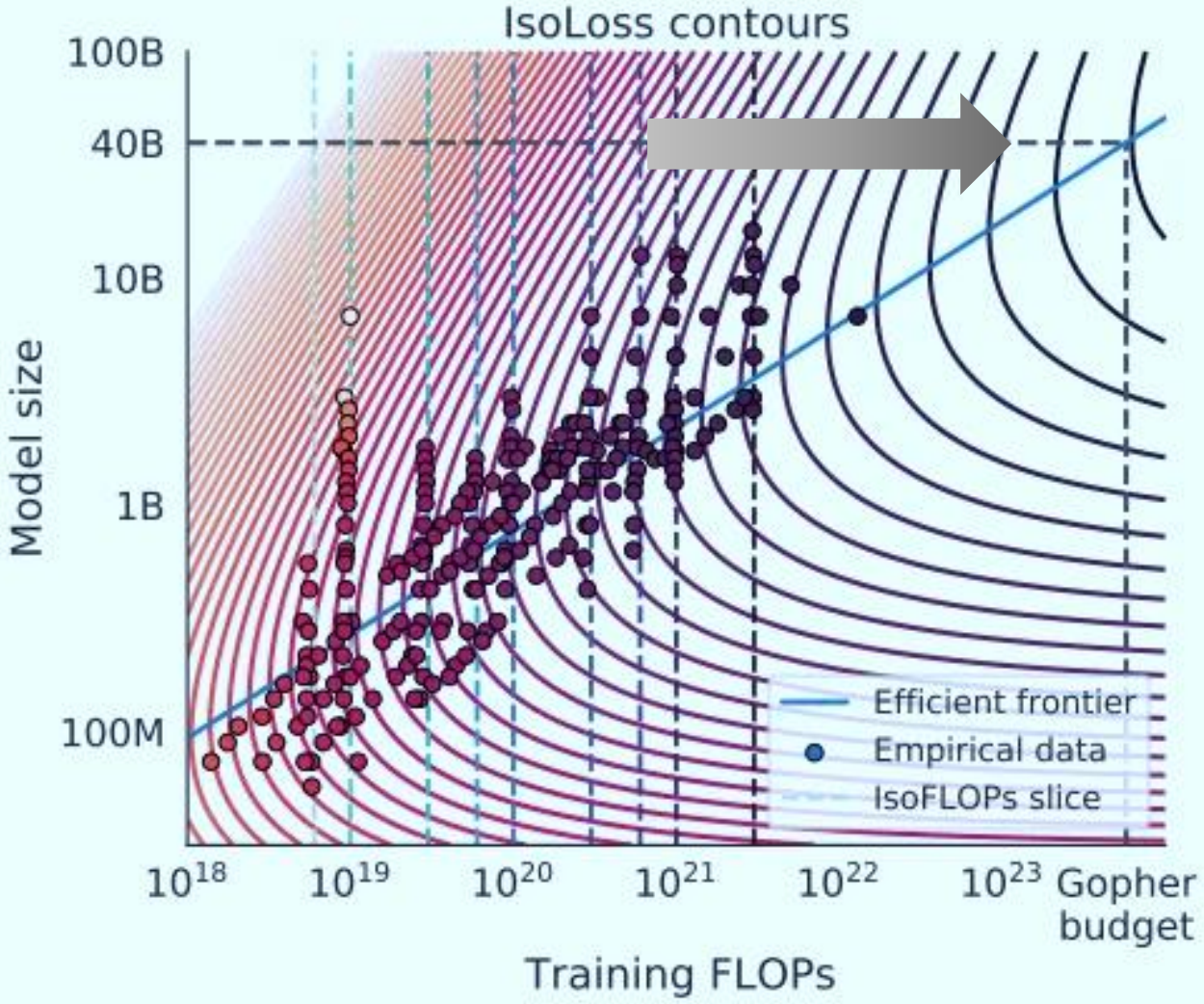


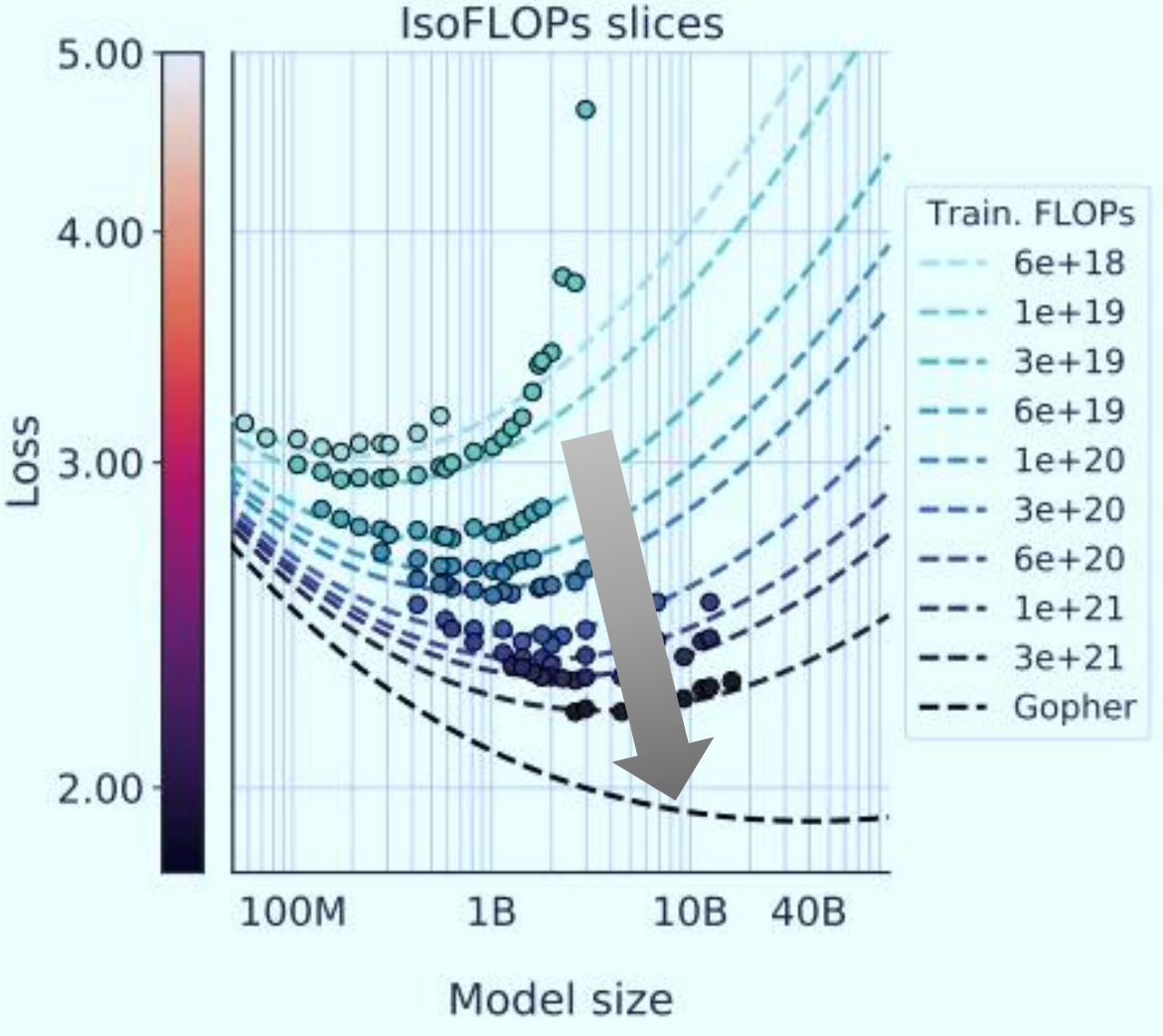


Training	Inferencing		
100T 10+T	100+T 10+T		
500B 50B	1T 500B		
10B	100B 10B		

AI Model Scaling Law

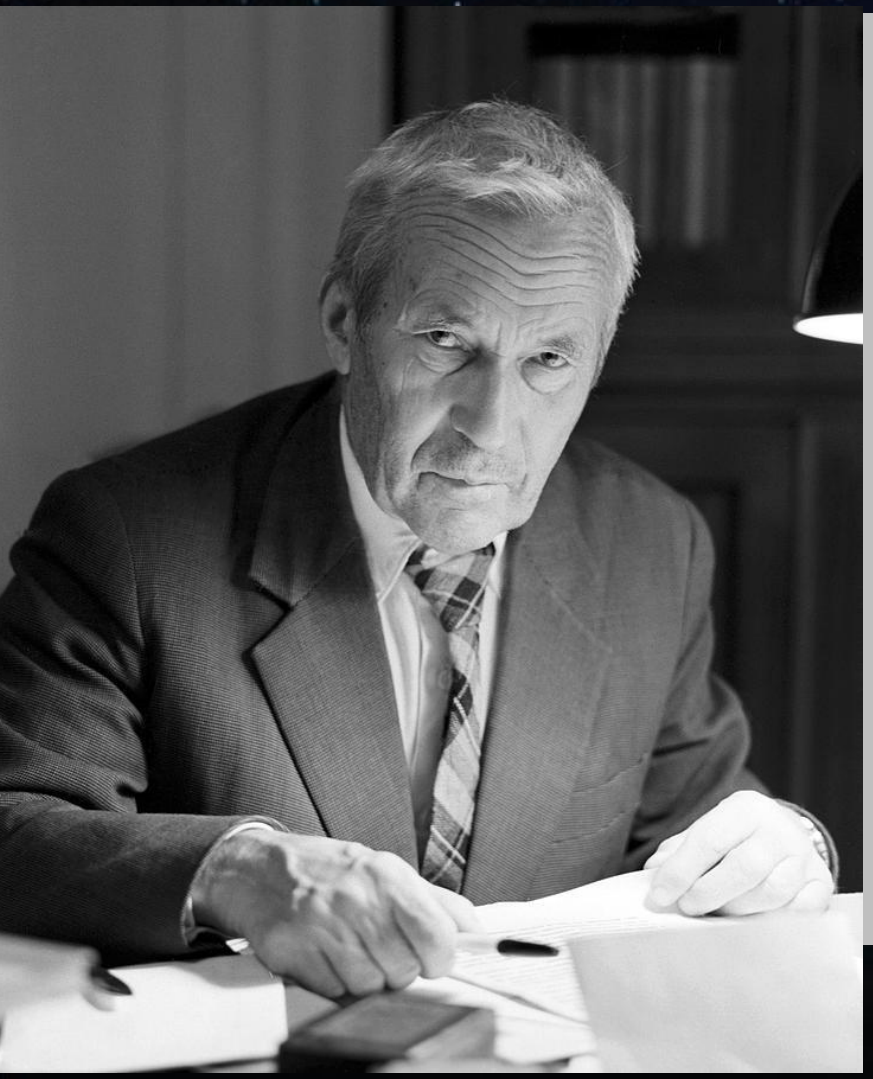
More intelligence "for free" by Scaling



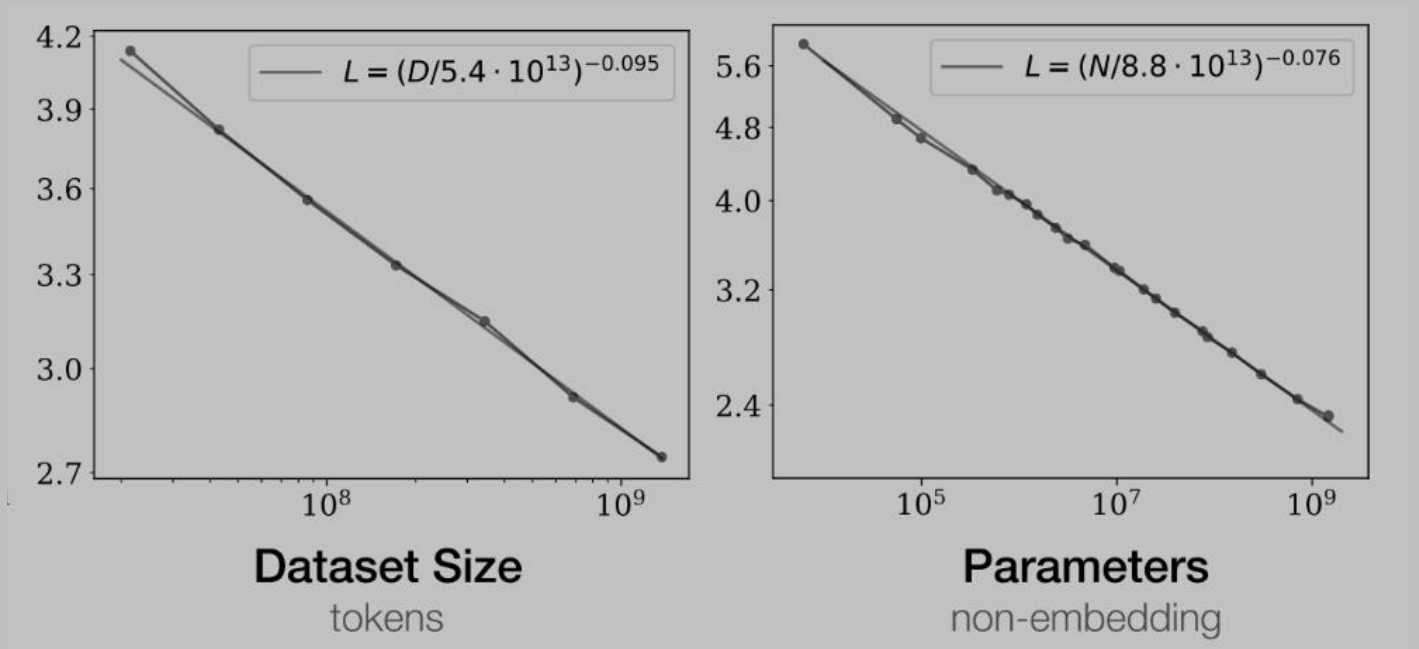


Training Dataset Scaling

Kolmogorov Compression



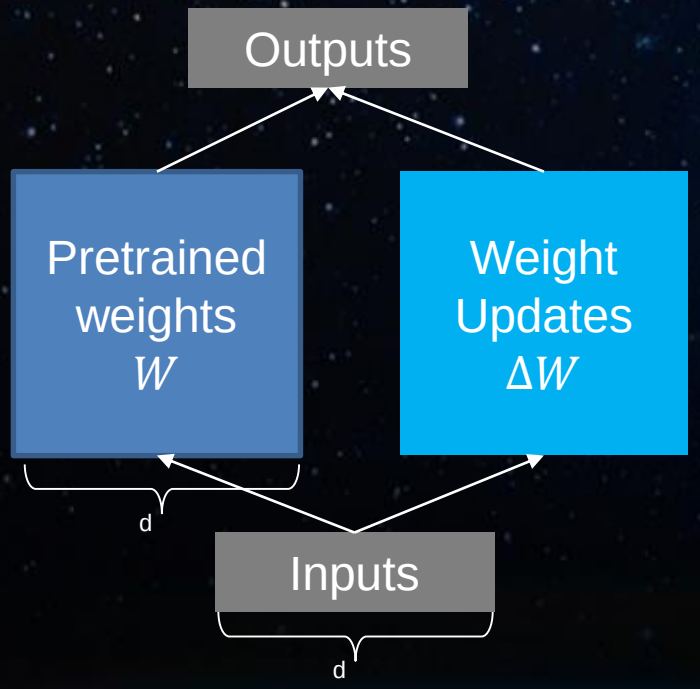
The more data the more intelligence

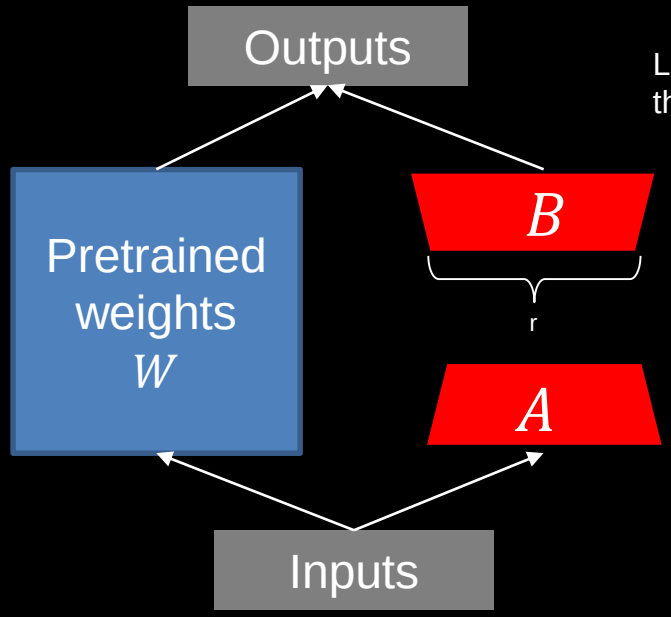


Wen Tong and Yiqun Ge "Information Theory and Learning-AN INFORMATION THEORY VIEW ON LEARNING PROBLEMS" Cambridge University Press 2022

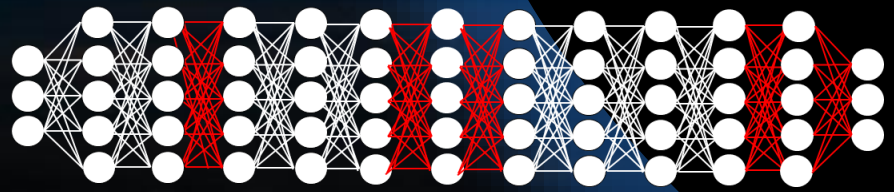
Model Fine-tuning

In-Network LoRA





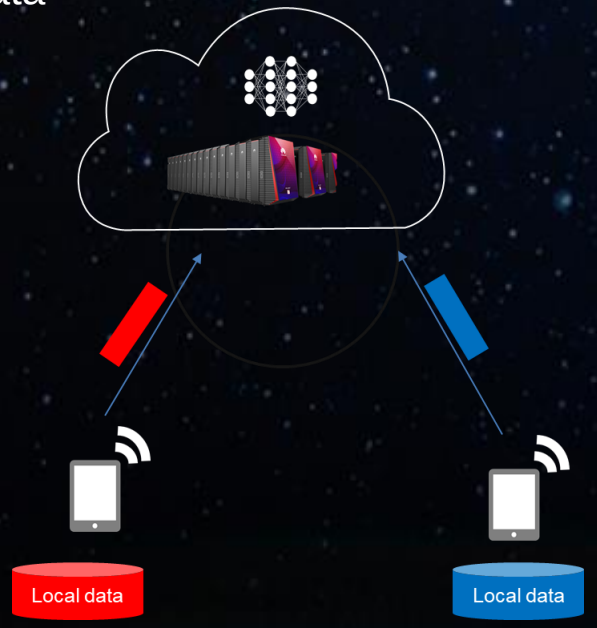
LoRA matrix A and B approximate the weight update matrix ΔW



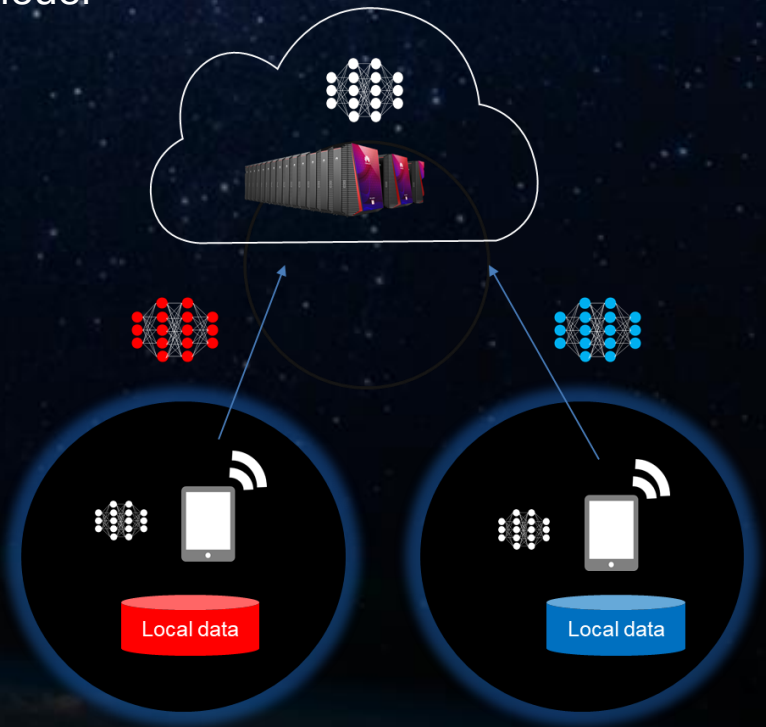
Federated Learning

Share model not the Data

Share Data



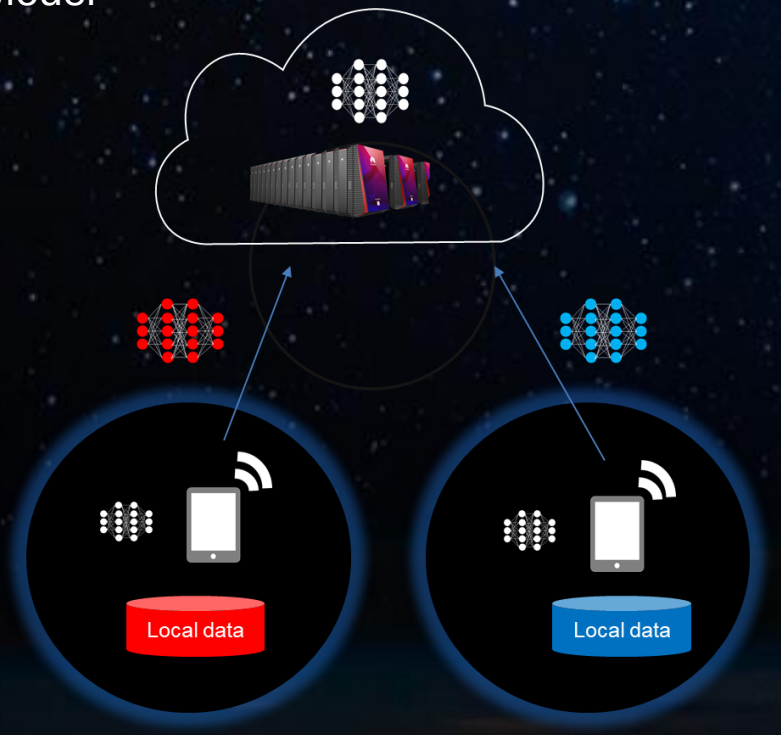
Share Model



Issue of Federated Learning (1)

The Need of a Trusted Host

Share Model



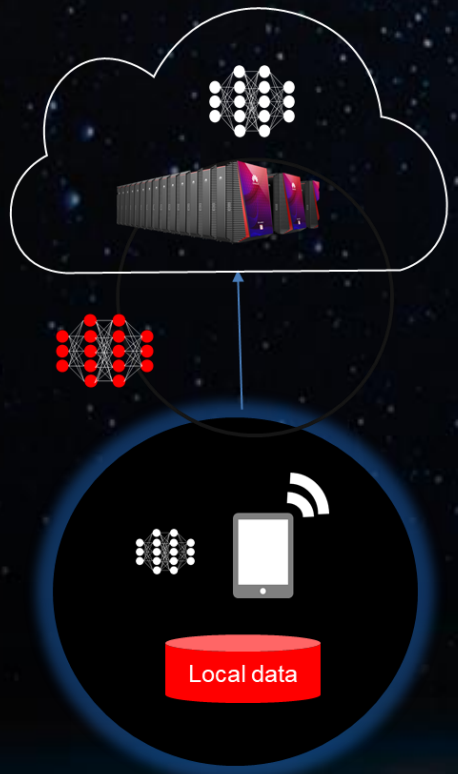
Federated:

Set up a single controlled center within which each state division keeps some internal autonomy

The trusted host to share the model

Issue of Federated Learning (2)

The Communications Cost



Constant

$$\text{Communication Cost} = n_{parameters} * n_{bits/param} * n_{users} * n_{rounds}$$

Reduce model size
- E.g.: using smaller models
- Reduces the performance

Quantization for transmitting
- E.g.: Mixed precision training
- Up to 4X times reduction

Reduce participants
- E.g.: Partial participation or scheduling
- Require additional information
- Up to 10X times reduction

Reduce the number of rounds
- E.g.: Fast-convergence methods or higher local training
- Additional computation on clients
- Up to 10X times reduction

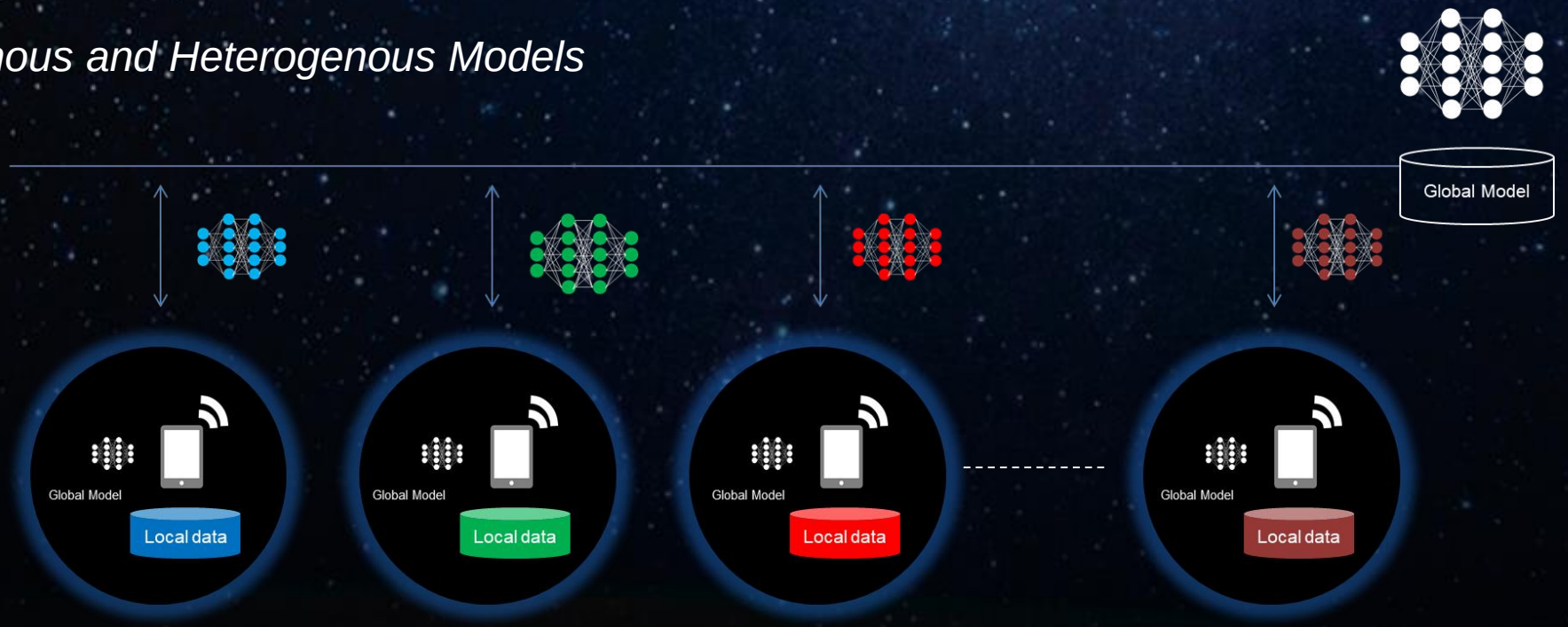
The total communication cost of *10 participants* sending *Llama2-7B* to a server in *100 rounds* of FL is:

"14 GB $*$ 10 $*$ 100 $*$ 2 = 28 TB"

Distributed Learning

Model-Follow-Data Concept

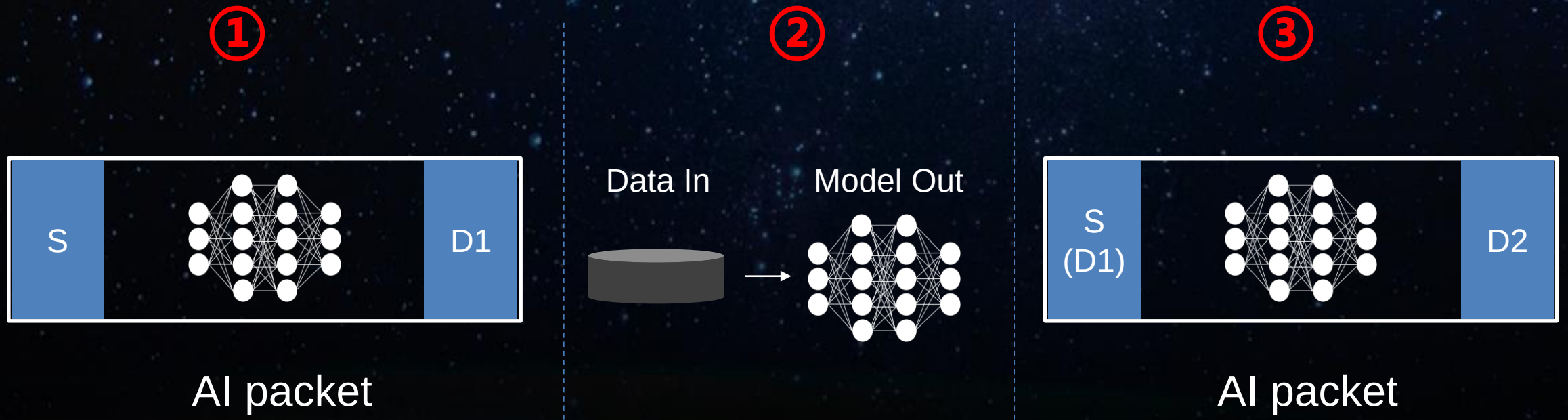
Homogenous and Heterogenous Models



Wen Tong: "IEEE SPAWC-2022" July, 2022

Model Communications (1)

Model-Follow-Data Concept



S: Source
D: Destination

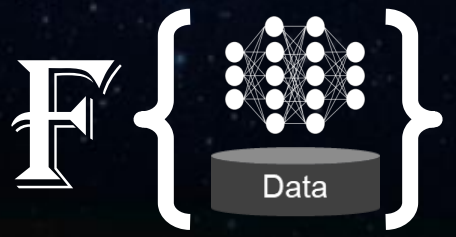
Model Communications (2)

Model-Follow-Data Concept

① AI Packet Routing

- Publish the data availability
- IP protocol for model forwarding

② Compute Data Interest



- Self training data cleaning
- Data quality criteria
- Option of distributed compute power

③ Update Routing Algorithm

Internet Routing Table

Destination	Subnet Mask	Gateway
xxxx	xxxx	xxxx

- Express of interests
- OBSF flooding

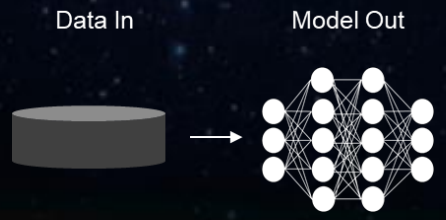
Model Communications (3)

Model-Follow-Data Concept

① Model Version Management

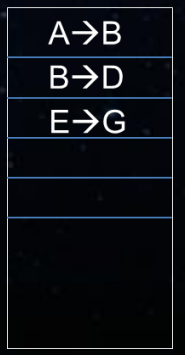
- Base Model Version
- Incremental Update Version

② Autonomous Model Training



- Discovery of data
- Discovery of compute
- Model verification

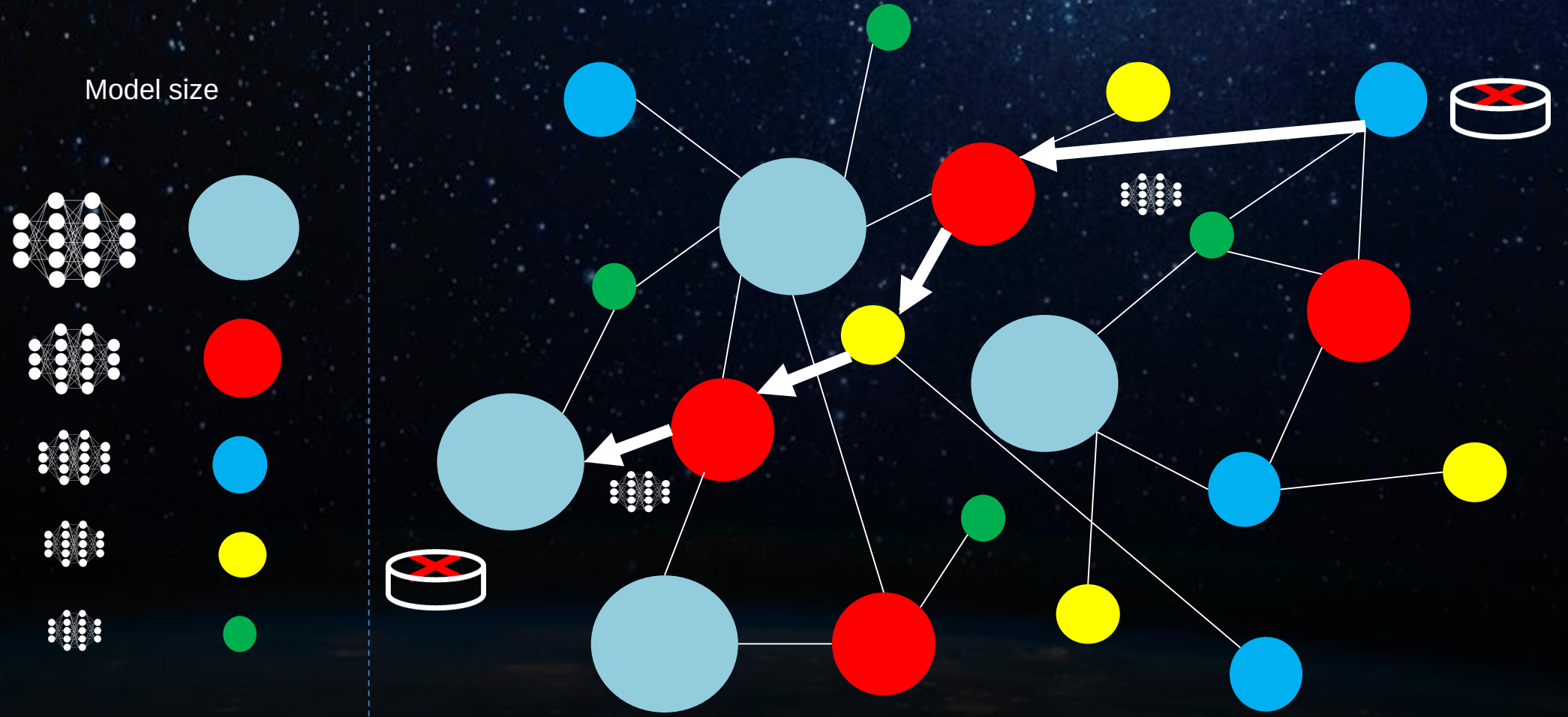
③ Model Distributed Ledger System



- Version validation
- Version audit

Heterogenous Model Communications

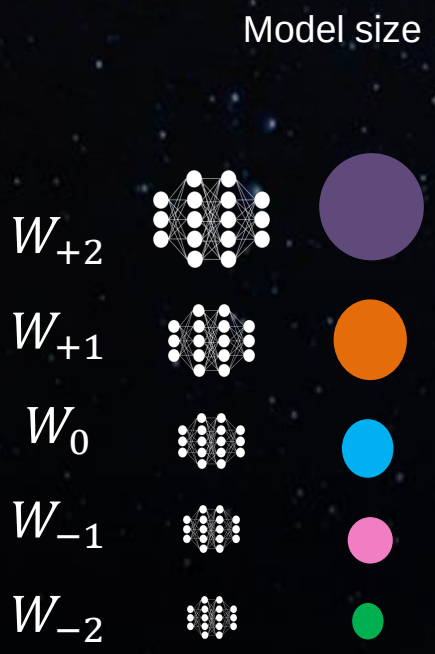
Model-Follow-Data Concept

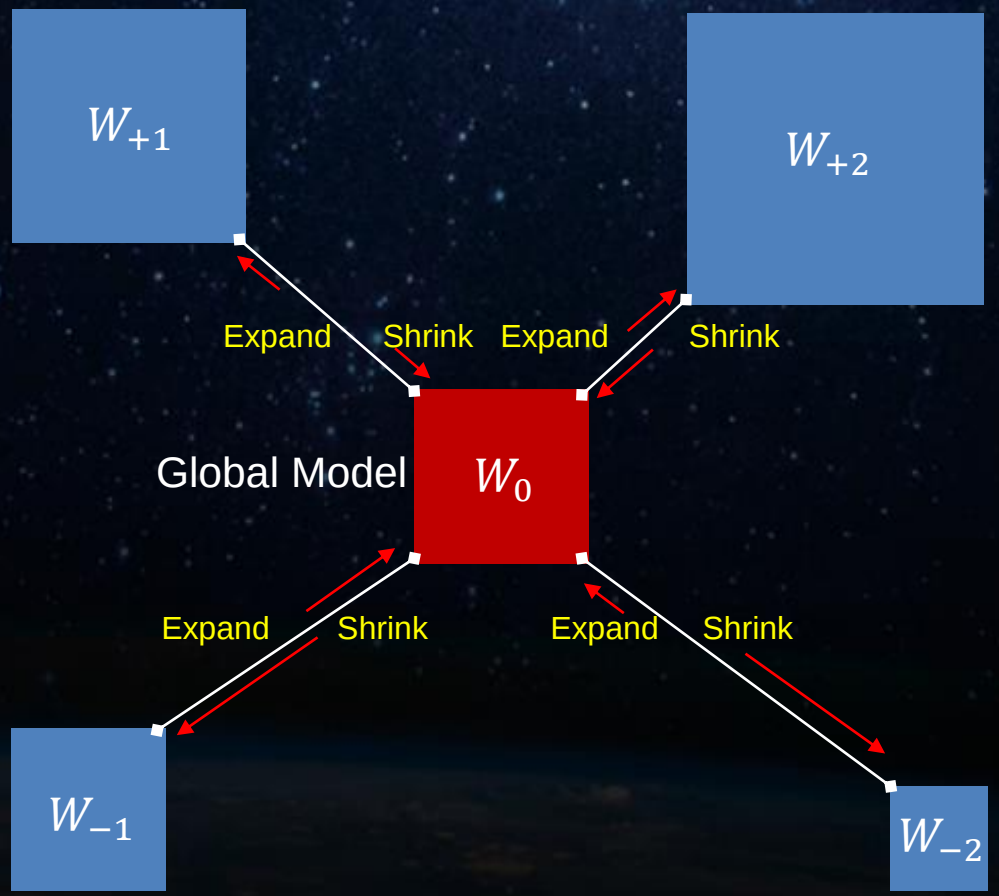


Wen Tong: "IEEE SPAWC-2022" July, 2022

Standard Model - Global Model

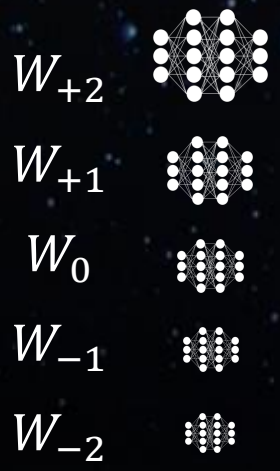
Heterogenous Model Scaling with Local Data Training

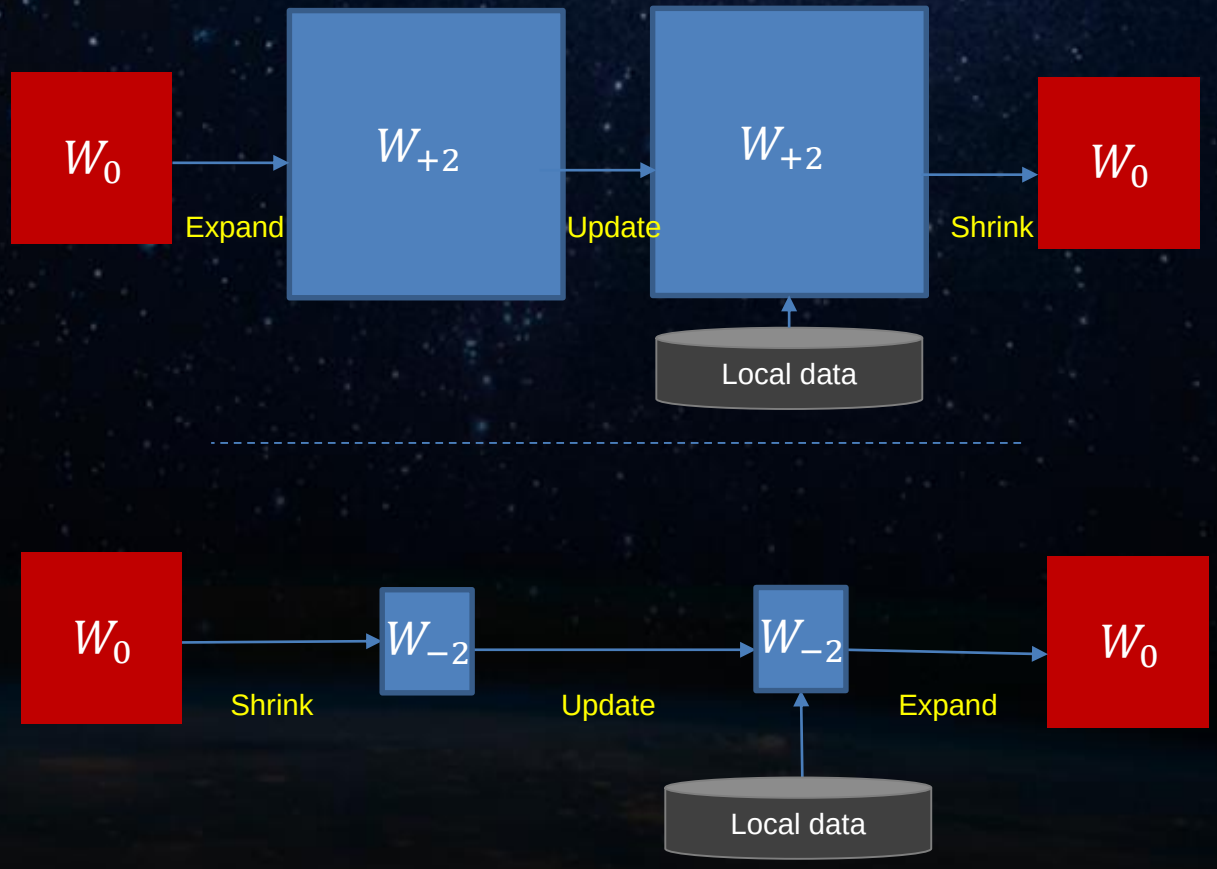




In-Network Model Processing

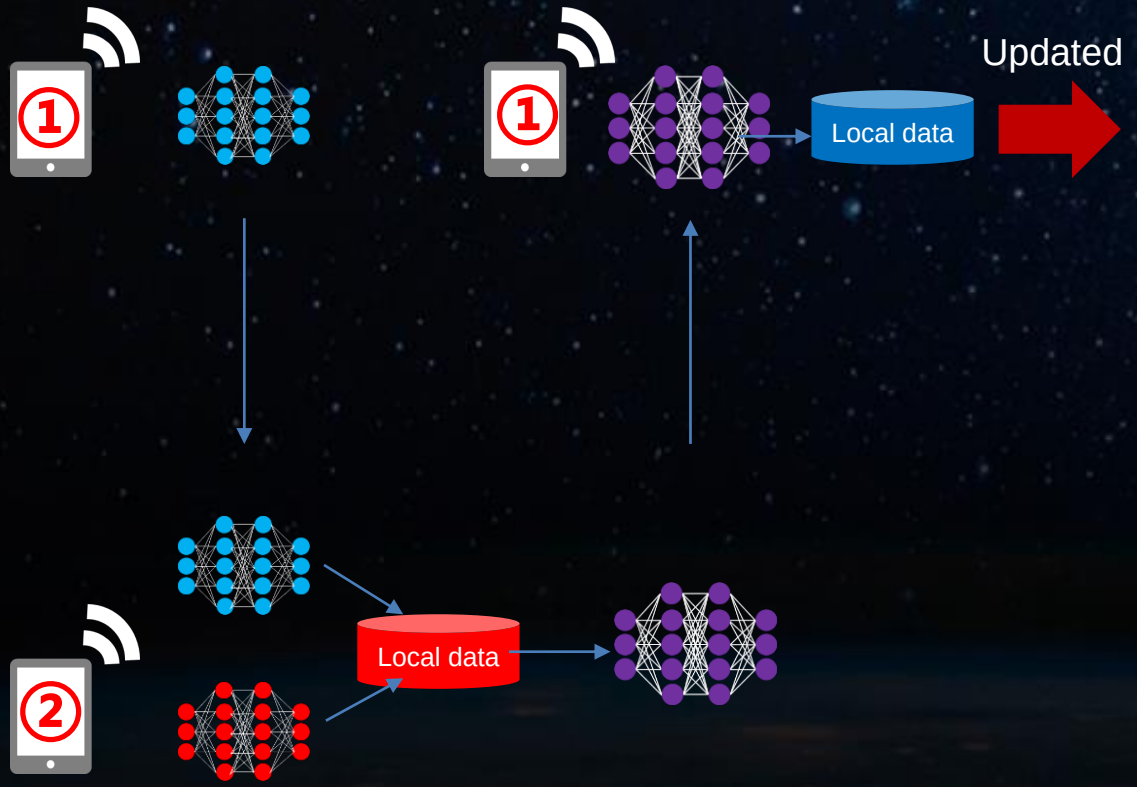
Heterogenous Model-Re-Normalization and Updates

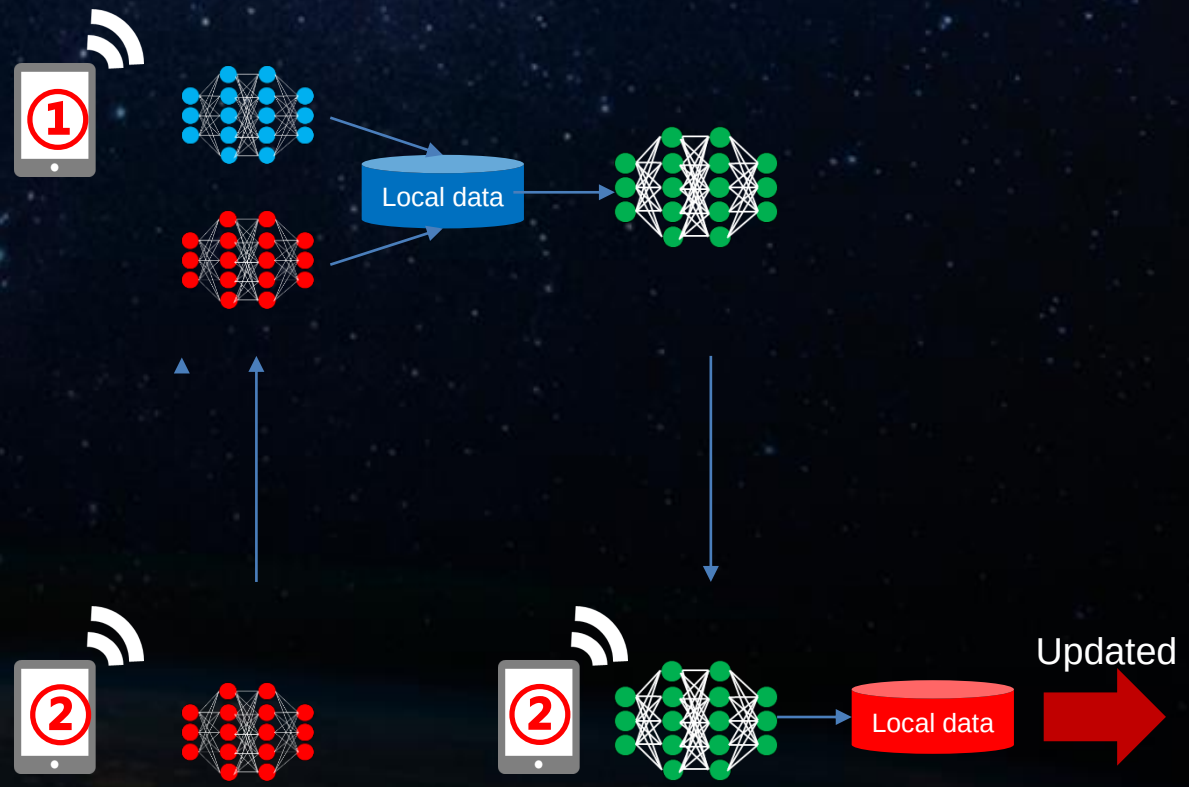




Distributed Learning Algorithm – (Knowledge Distilling)

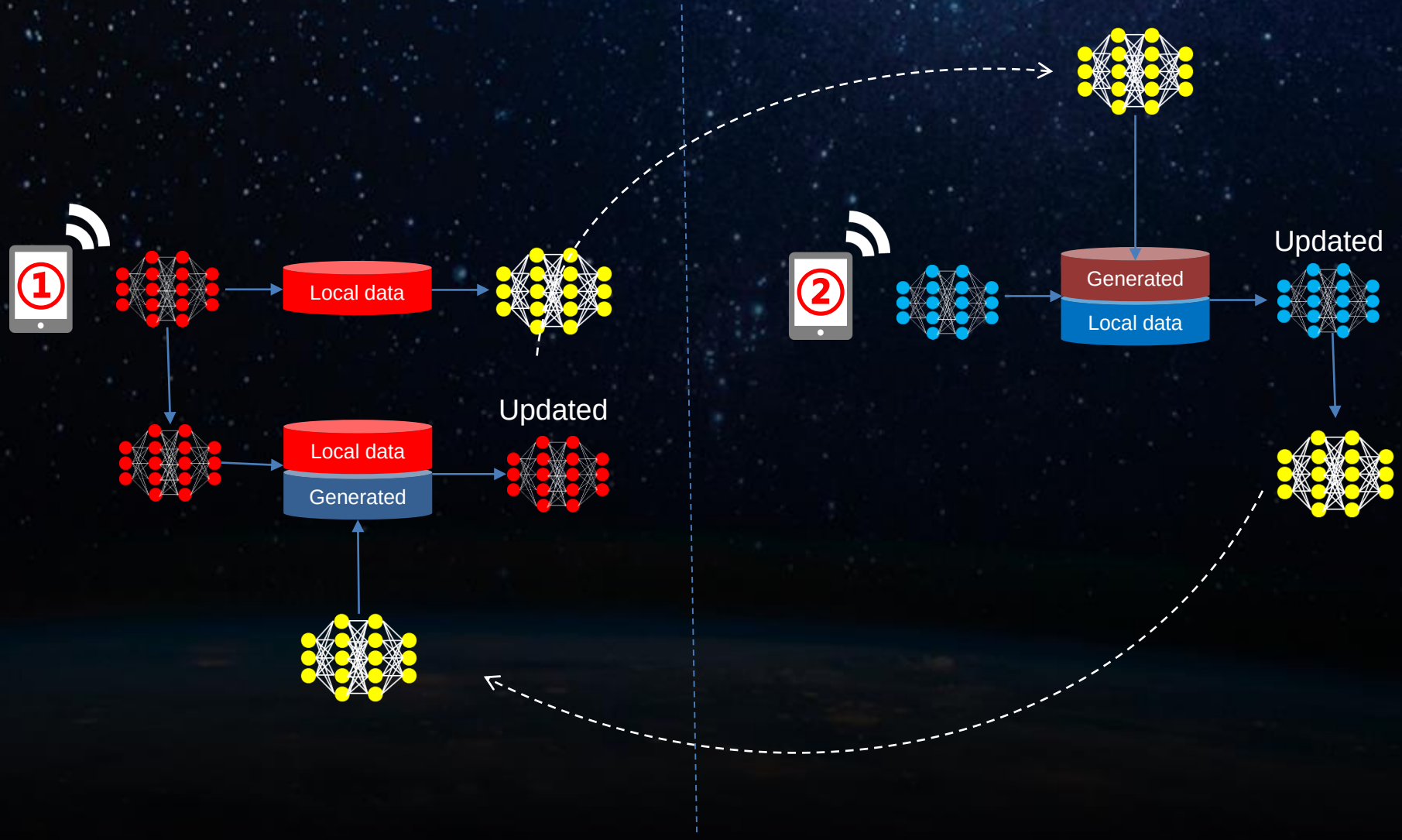
Heterogenous Model





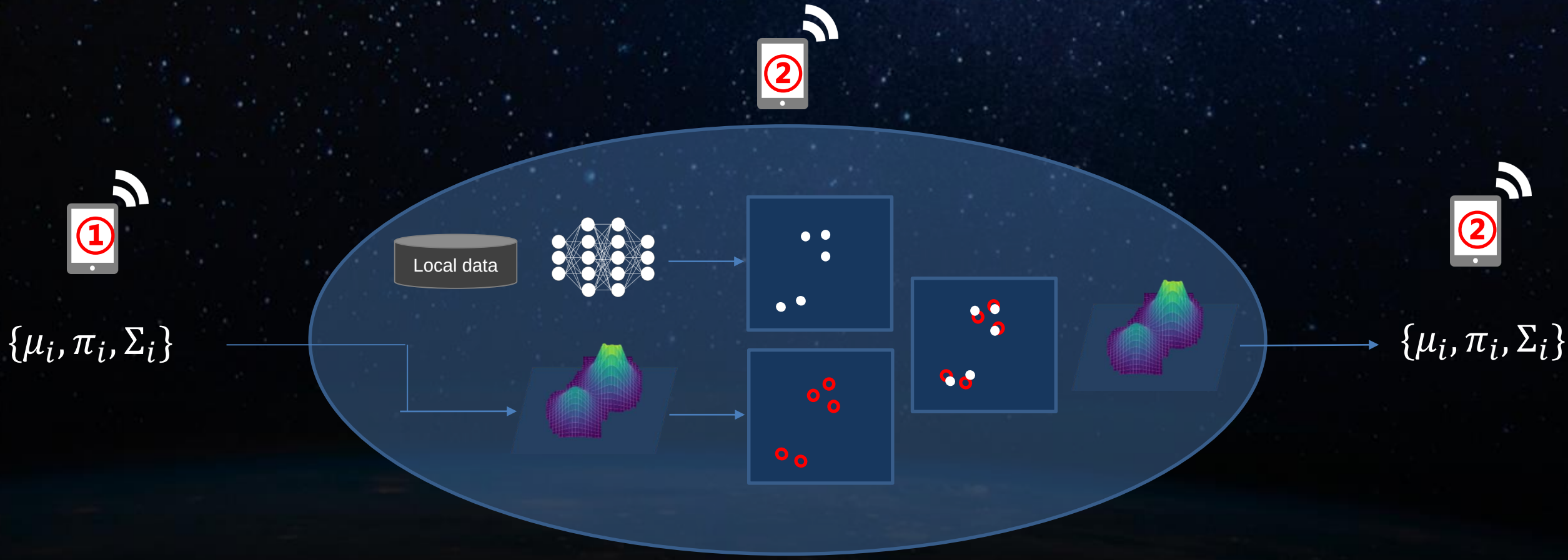
Distributed Learning Algorithm – (Generative Model)

Heterogenous Model



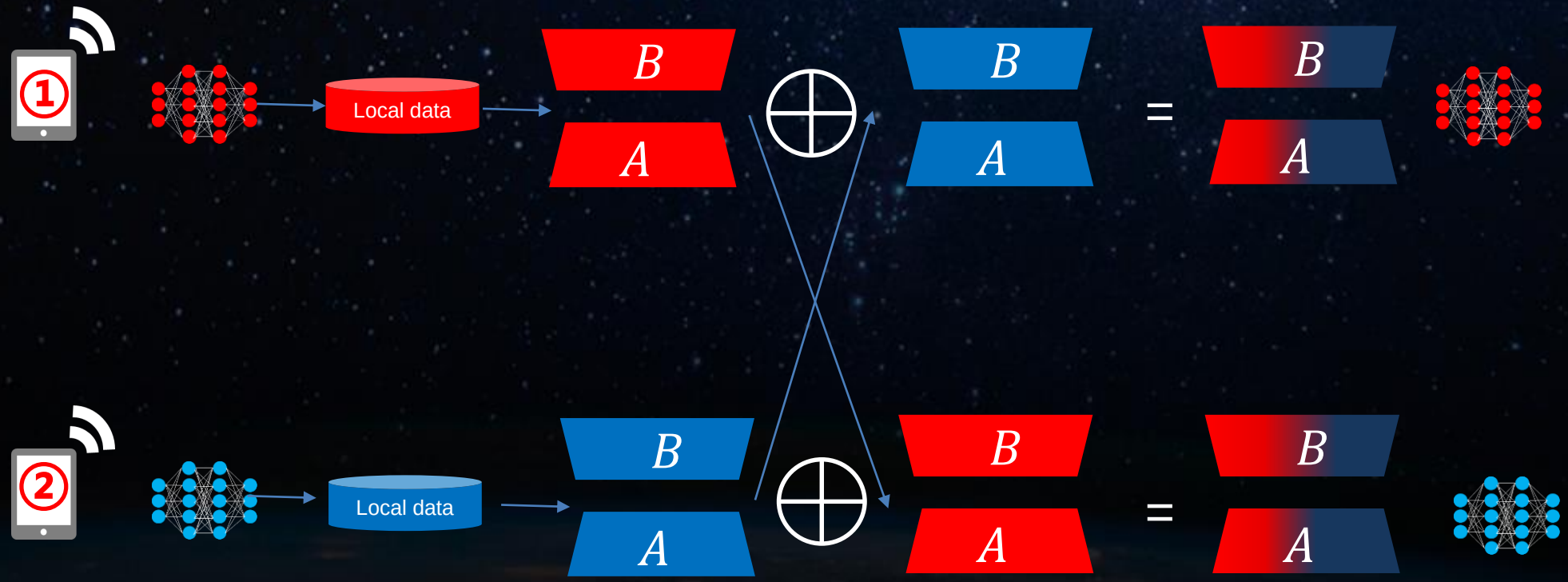
Distributed Learning Algorithm – (Information Bottleneck)

Heterogenous Model



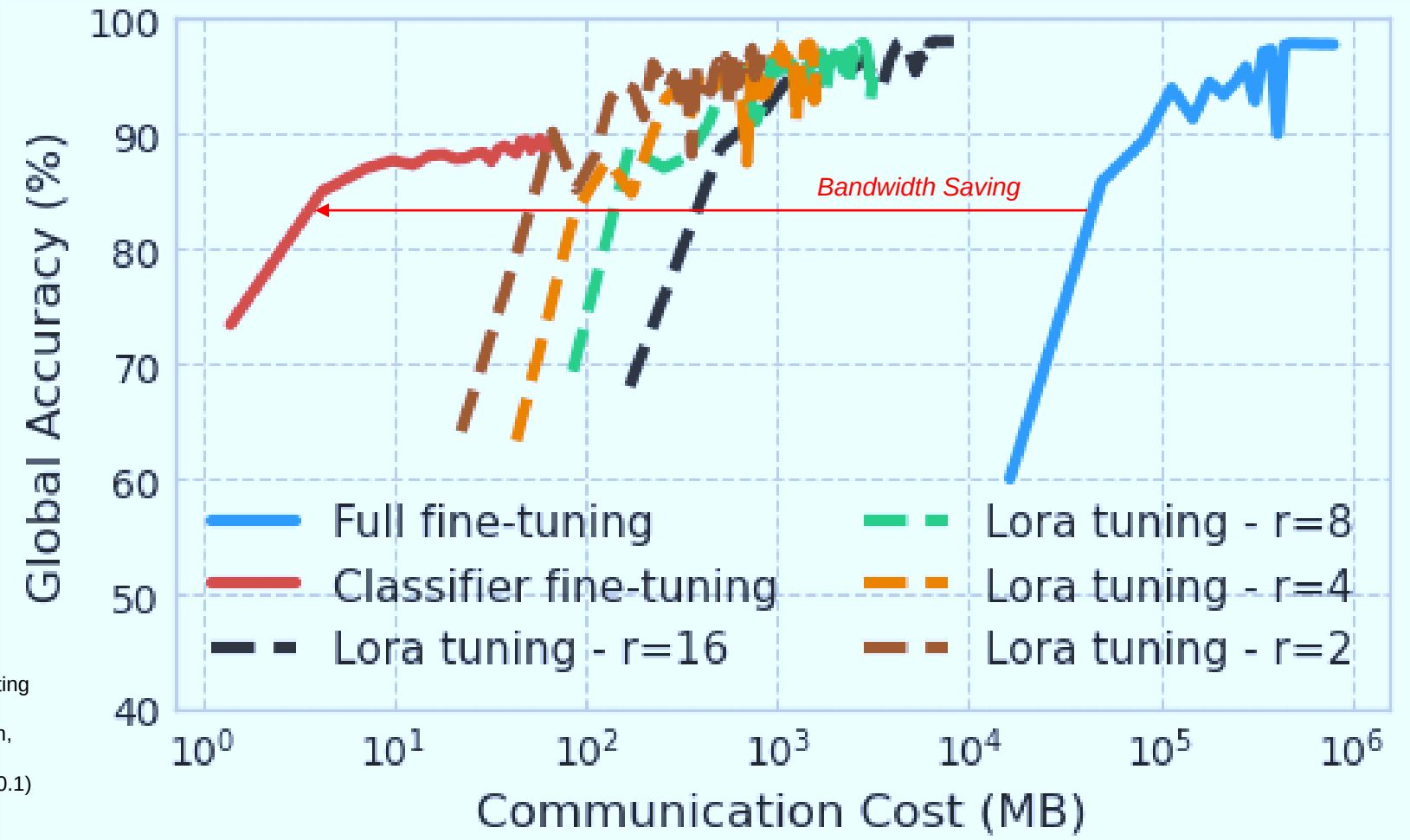
Distributed Learning Algorithm – (LoRA)

Heterogenous Model



Distributed Learning Algorithm – (LoRA)

Heterogenous Model



Topology: 10 Clients, 5 clients participating in each round
Dataset: CIFAR10 – 25k images as train, 25k as test
Distribution: Heavy label shift, Dirichlet(0.1)
One base model: ViT-base (86M), pre-trained on ImageNet21K images

Federated and Distributed Comparison

Reduced Communications Cost

▪ Federated Learning	100n
▪ Distributed Learning	
▪ Knowledge Distilling	4n
▪ Generative	2n
▪ Information Bottleneck	2n
▪ LoRA	2n

Federated and Distributed Algorithms

Remove Inferior Model Contributions--how to un-learn

- Federated Learning

$$\sum\left(\ldots\ldots\left(\sum \widetilde{W}_{n,i}\right)\sum \Delta\widetilde{W}_{n,i}\right)$$

- Distributed Learning

$$\left(\left((W_0 + A_1B_1) + A_2B_2\right) + A_3B_3\right) + \cdots$$

Data and Computing Aware Model Routing

Model-Follow-Data Concept

Conventional:

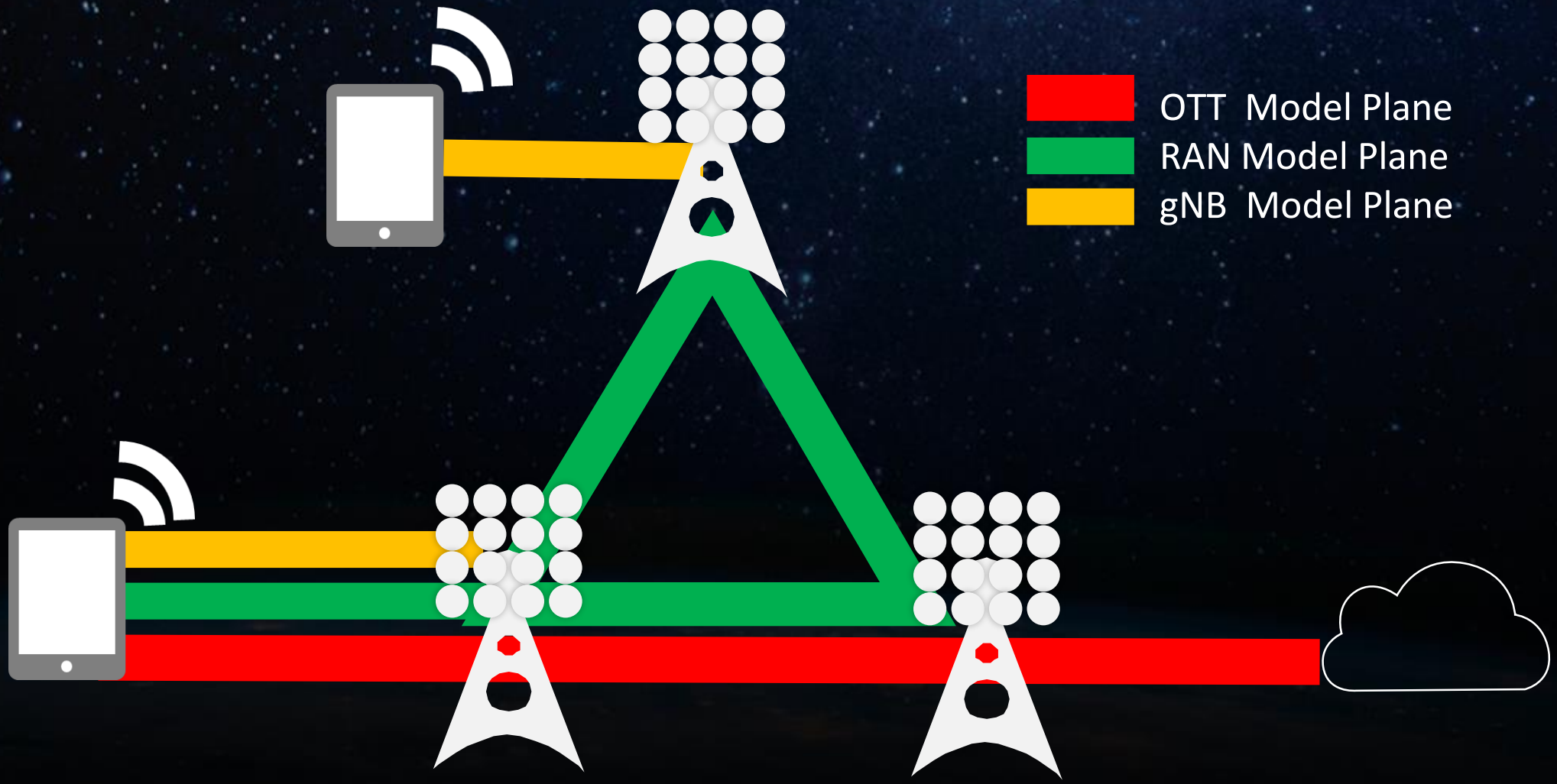
- Randomly chosen next hop
- Neither the model architecture nor the data characteristics have any impact on the decision

Ours:

- Next hop selection based on new metric
- The model node obtains random minibatches from network nodes
- The model node computes a metric for each node based on its minibatch and current state of the model.

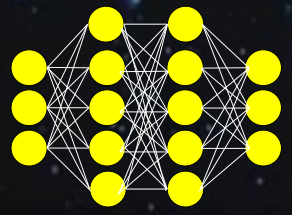
Distributed Training for RAN

3GPP Architecture



Model Based Communications

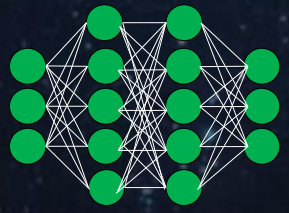
Model for Efficient Communications



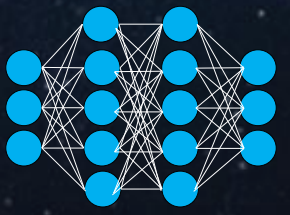
1/1000

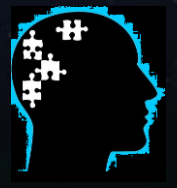
Model for Reliable Communications

Model for Secure Data Translator

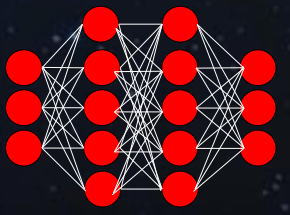


Model for Data Memory



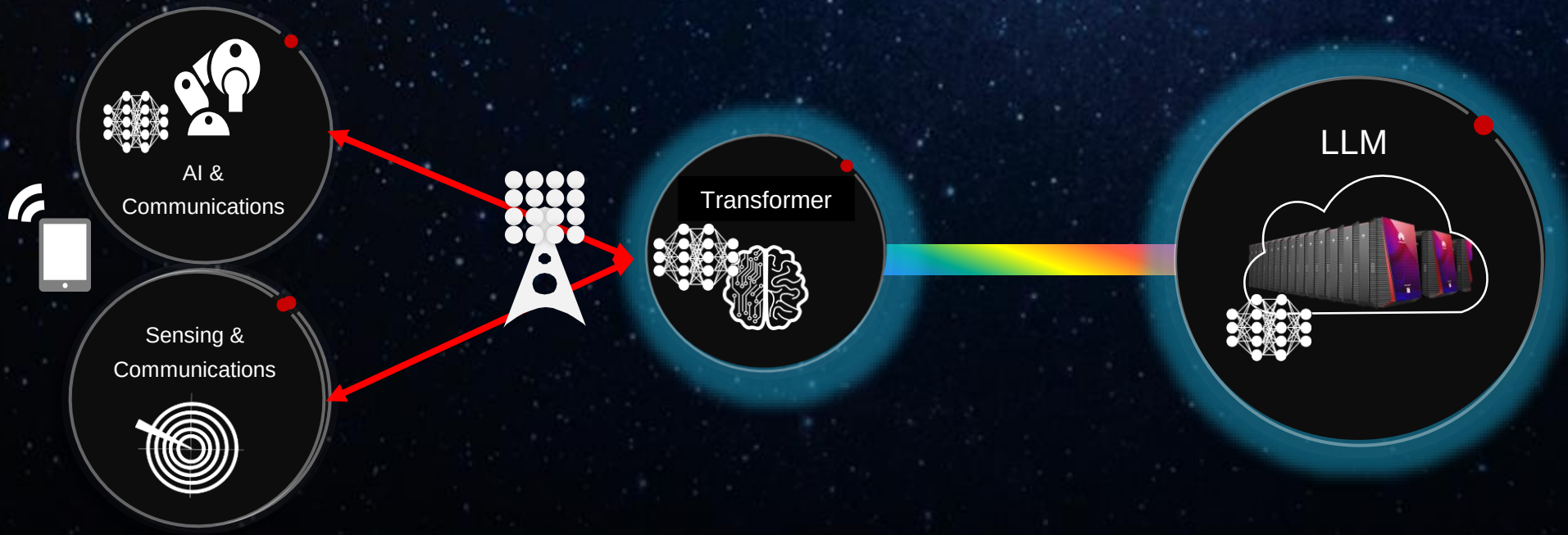


Model as Trustworthy User ID



(Wen Tong, Keynote IEEE-CTW-2021)

A-RAN™



Neuron Agent (10B)

Neuron Edge (100B)

Neuron Center (1T)

(Wen Tong, Keynote 6G SUMMIT Abu Dhabi Nov.16th, 2023)

Thank You.