

Autonomous Driving SLICES

*what SLICES can bring to
Network AI research*

Dario Rossi

dario.rossi@huawei.com



Huawei Network AI CTO
Director, DataCom Lab, Paris Research Center



Artificial Intelligence (AI) & Machine Learning (ML)

AI & ML



95% of network changes
involve manual operation



70% network faults are
caused by manual error



Remove humans
from the fast loop



Keep human in
the slow loop

A new dawn



Autonomous driving



network



∞ Industry segments & requirements

High reliability



Transportation

Differentiated services



Government



Healthcare

Zero packet loss



Energy



Education

Smart O&M



Manufacturing

Real-time, high bandwidth



Mining



Finance

3 Network scenarios

Campus network

DCN

WAN

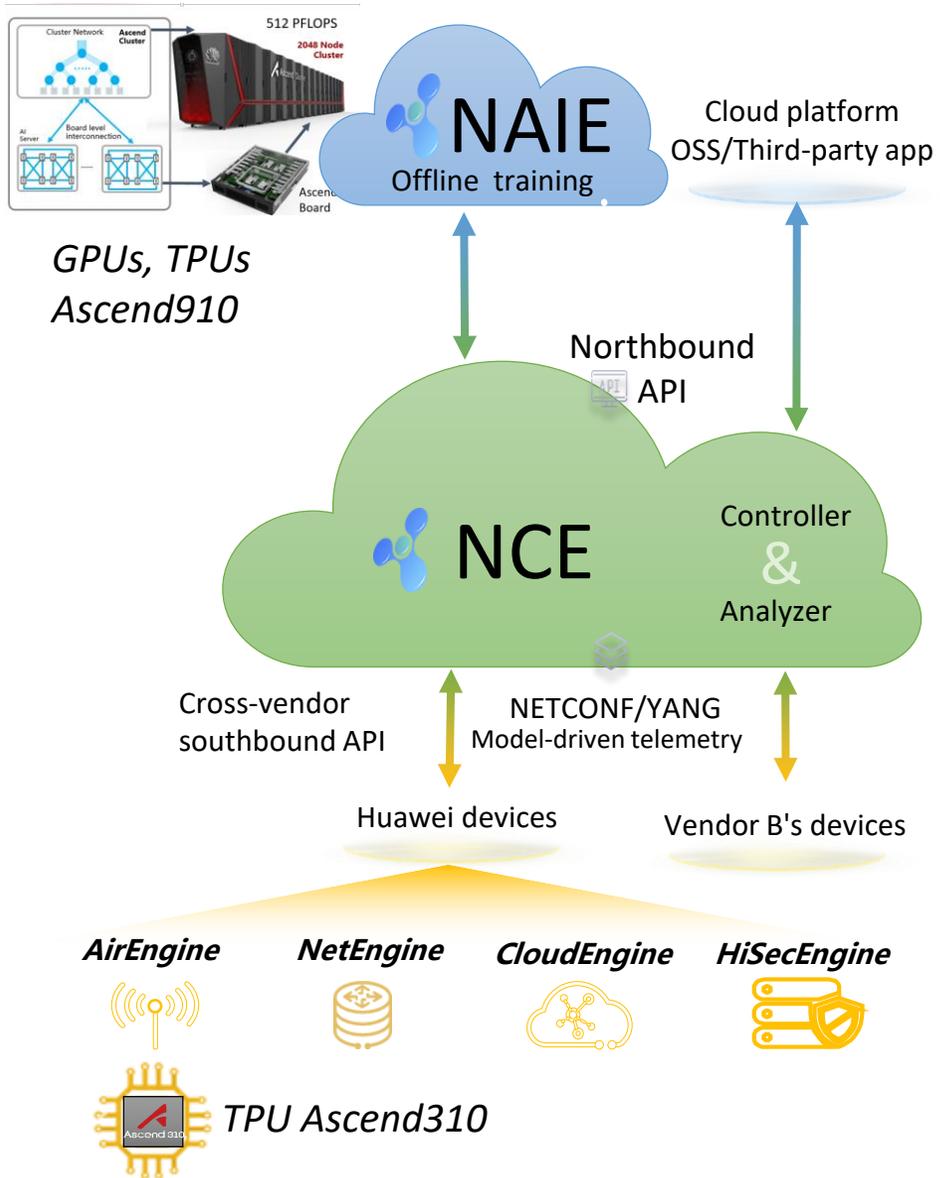
+

Security

1 Technology solution



Network AI in Huawei



iMaster NAIE
Training, data aggregation, and model generalization

iMaster NCE
Network-wide analysis, inference & closed-loop optimization

Engines
Measurement, edge inference & real-time decision-making

General:
Multi-vendor knowledge graph/models

Training:
Federated Learning

Specific:
Deep Models Quantization & Distillation

Control:
large-scale data-driven reinforcement learning

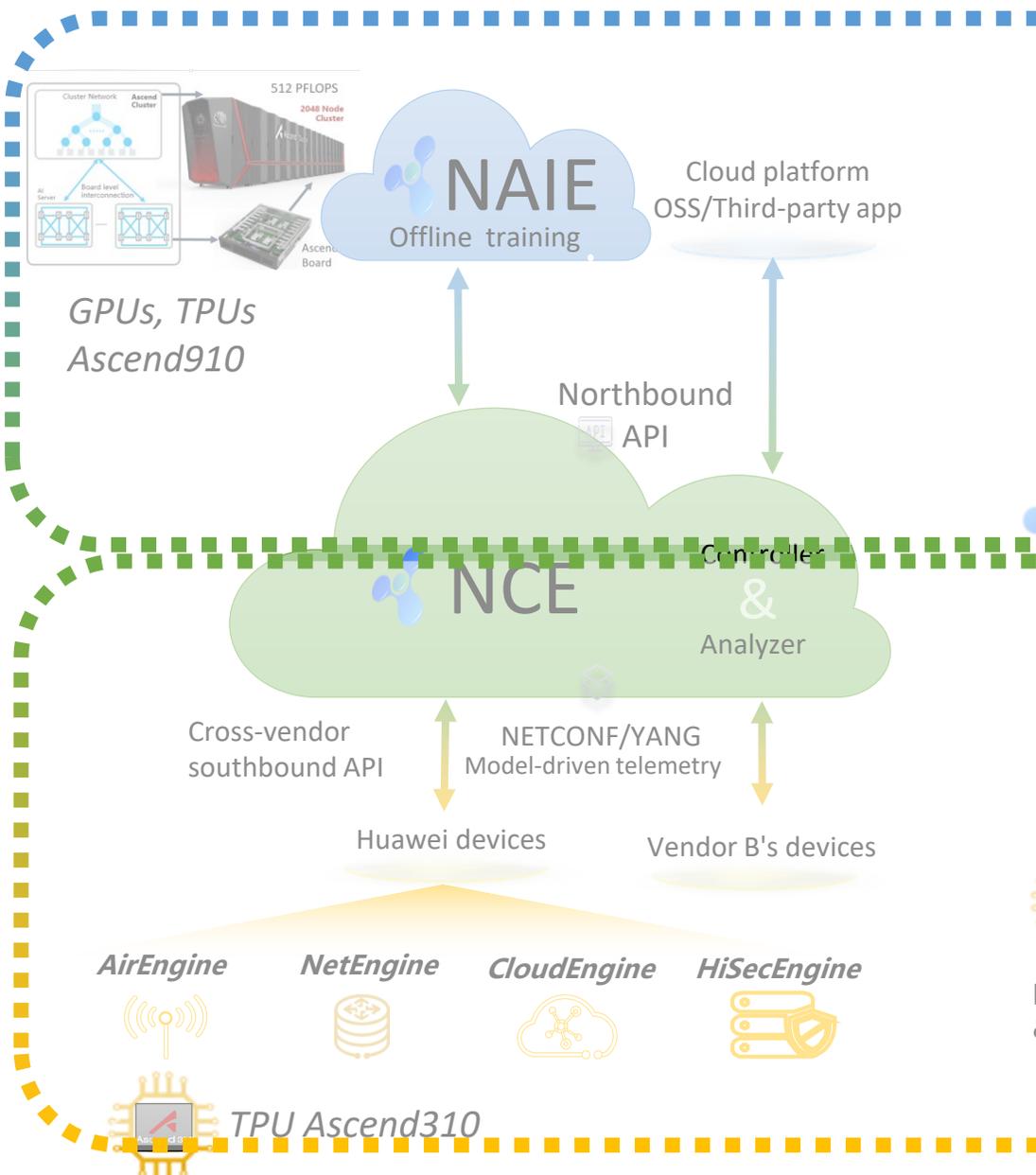
O&M:
Unsupervised Fault detection, Semi-supervised repair

Real-time inference & control

Model self-awareness

Incremental & continuous learning

Network AI in Huawei



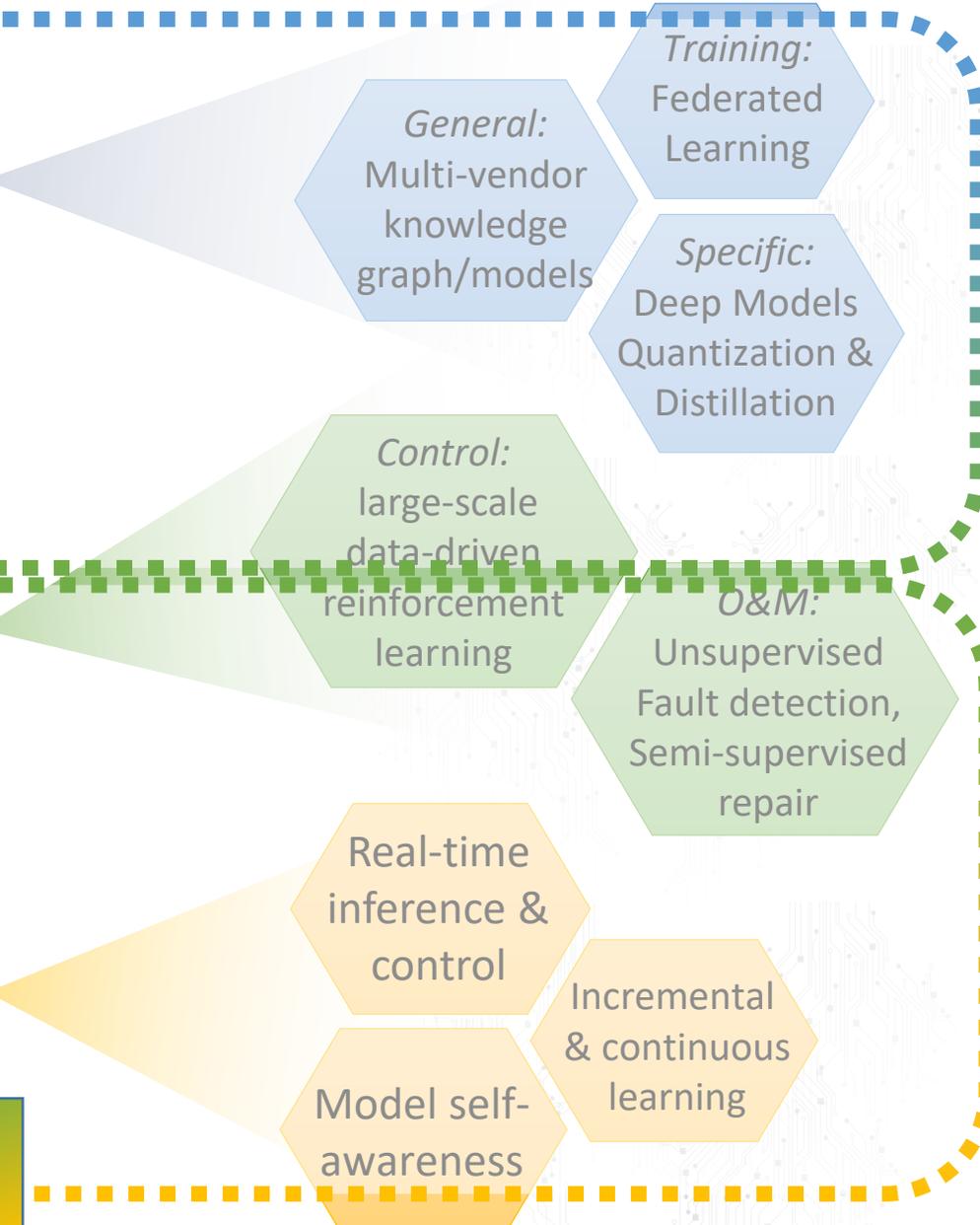
NET4AI viewpoint

iMaster^{NAIE}
 Training, data aggregation, and model generalization

iMaster^{NCE}
 Network-wide analysis, inference & closed-loop optimization

Engines
 Measurement, edge inference & real-time decision-making

AI4NET viewpoint



Network 4 AI viewpoint

❑ Model training

- E.g., realism in federated learning from heterogeneous deployments (practical system-level AI challenge)

❑ Model-driven telemetry (MDT)

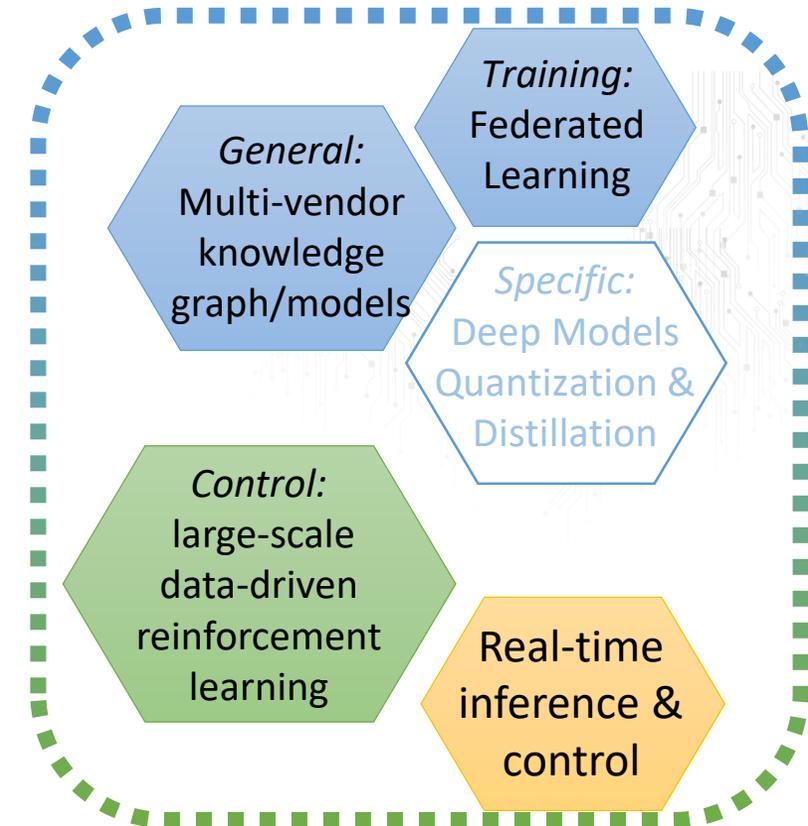
- Heterogeneity in the input data: multi-vendor (good to have “dirty data” AI problem)

❑ Real-time

- Where (Cloud vs Fog vs Edge) to allocate AI resources: architectural tradeoffs of privacy vs cost vs ...

❑ Control

- Delay+noise of MDT data streams: controllable/reproducible AI experiment in more challenging environment
- Train on simulation (e.g., DRL takes lifetimes, cannot learn from real network) refine & validate on SLICES



Large scale, heterogeneous RI

=> lower access barrier to experimental study & more realistic challenges

Large user community

=> critical mass to push reproducibility standards



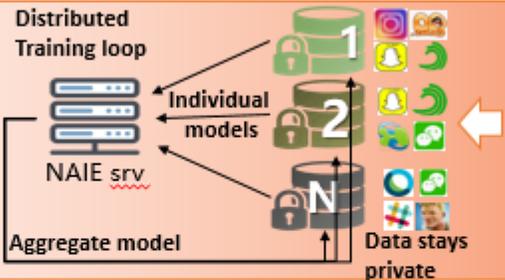
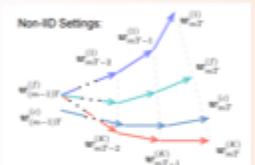
Network 4 AI viewpoint

□ Model training

- E.g., realism in federated learning from heterogeneous deployments (practical system-level AI challenge)

Federated Learning

Google FedAvg works only with i.i.d. data (in non-i.i.d. case, gradients diverge)

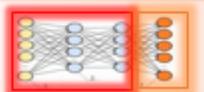


Signature portability issue:

- Loss of accuracy if training & testing over *different* network
- Simple solution = centralized training over *both* networks



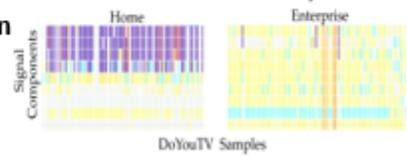
Split ANN model into **Common Backbone** + **Private Classifier**



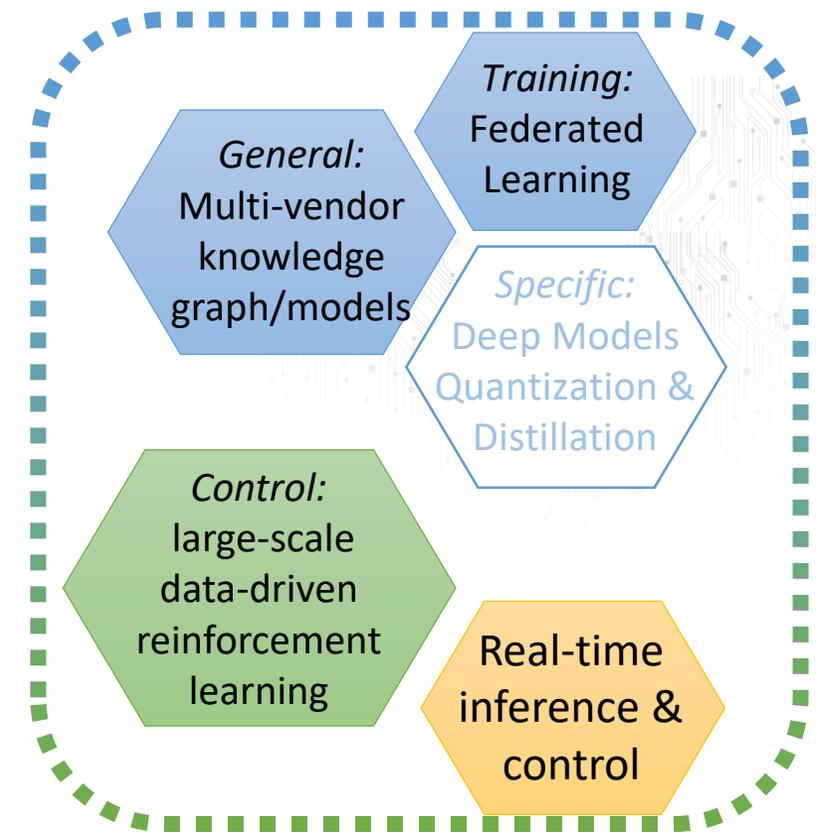
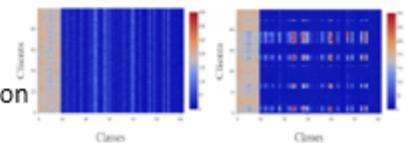
- Only the **backbone is shared & averaged** (speedup learning of common hidden layers)
- The **last layer classifier remains private** (less information share/leak, better fit to the data of each client)



✓ **Huawei traffic classification**
 <1% accuracy loss,
 30x data reduction
 wrt centralized training



✓ **MNIST image dataset**
 1.3x faster to converge
 with 1.5x less communication
 wrt Google FedAvg



Large scale, heterogeneous RL
 => lower access barrier to experimental study & more realistic challenges

user community
 critical mass to push
 reproducibility standards



Network 4 AI viewpoint

❑ Model training

- E.g., realism in federated learning from heterogeneous deployments (practical system-level AI challenge)

❑ Model-driven telemetry (MDT)

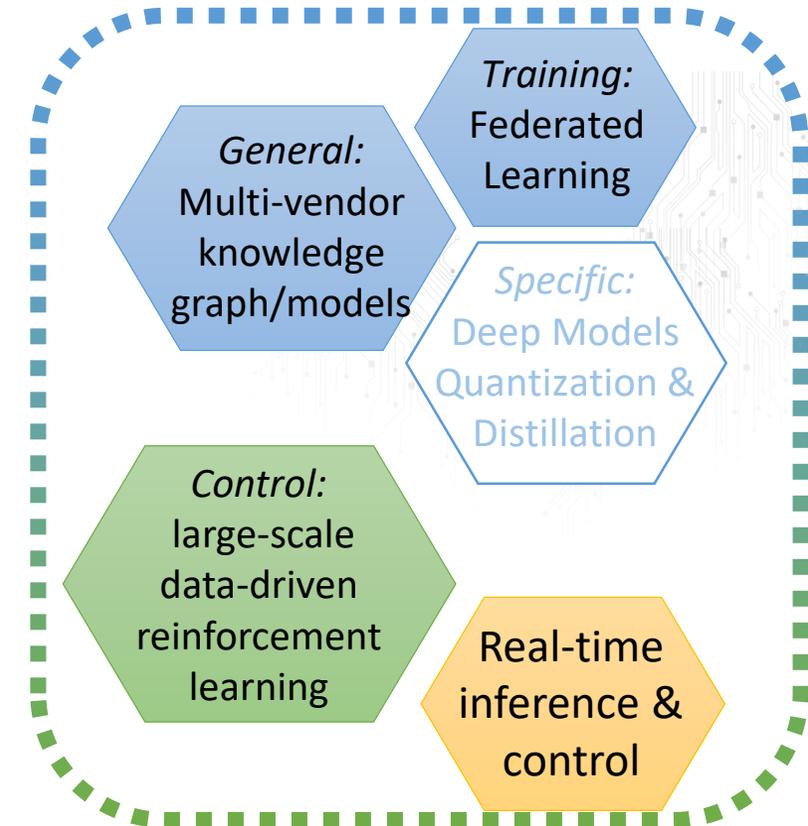
- Heterogeneity in the input data: multi-vendor (good to have “dirty data” AI problem)

❑ Real-time

- Where (Cloud vs Fog vs Edge) to allocate AI resources: architectural tradeoffs of privacy vs cost vs ...

❑ Control

- Delay+noise of MDT data streams: controllable/reproducible AI experiment in more challenging environment
- Train on simulation (e.g., DRL takes lifetimes, cannot learn from real network) refine & validate on SLICES



Large scale, heterogeneous RI

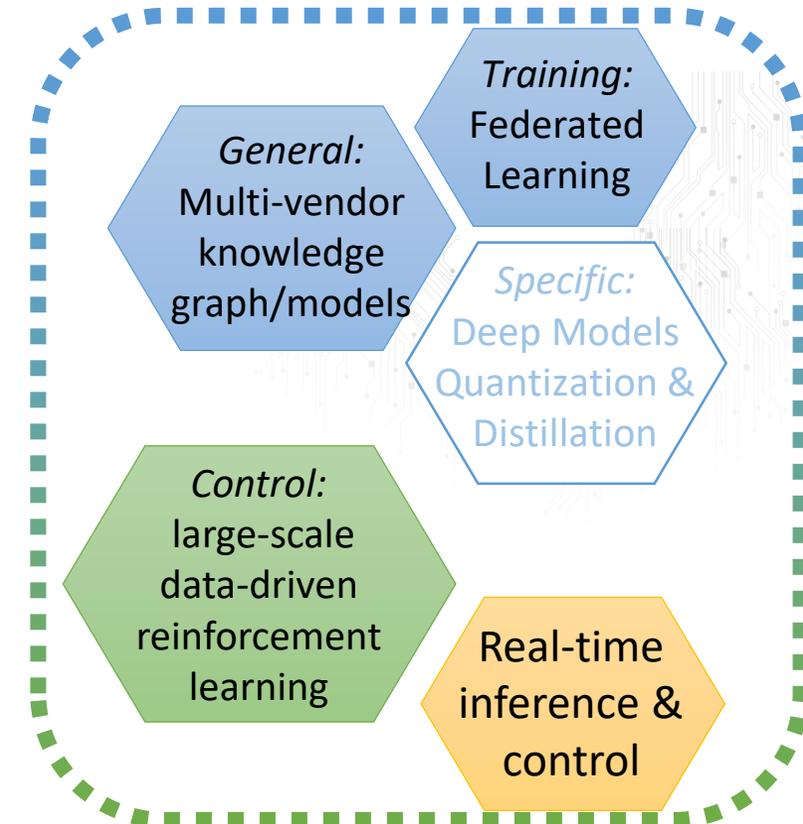
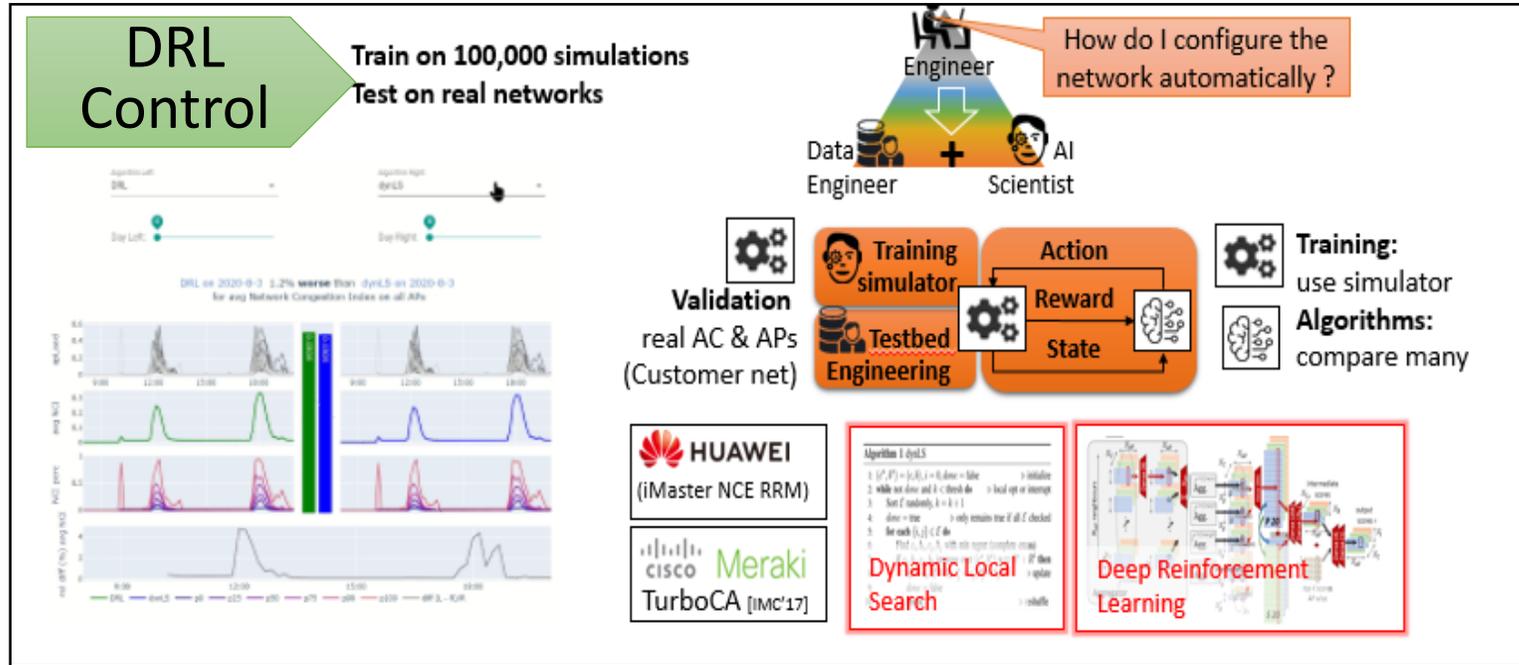
=> lower access barrier to experimental study & more realistic challenges

Large user community

=> critical mass to push reproducibility standards



Network 4 AI viewpoint



Control

- Delay+noise of MDT data streams: controllable/reproducible AI experiment in more challenging environment
- Train on simulation (e.g., DRL takes lifetimes, cannot learn from real network) refine & validate on SLICES

Large scale, heterogeneous RI

=> lower access barrier to experimental study & more realistic challenges

Large user community

=> critical mass to push reproducibility standards



Note: I was happy to find the "heterogeneity" keyword in Jon's keynote 😊

AI 4 Network viewpoint

❑ Model-driven O&M

- Unsupervised algorithms still need ground truth for benchmark
- Large SLICES crowd: can the community crowdsource anomaly detection database beyond KDD99 (s/ImageNet/AnomalyNet/)?

❑ Heterogeneity (again)

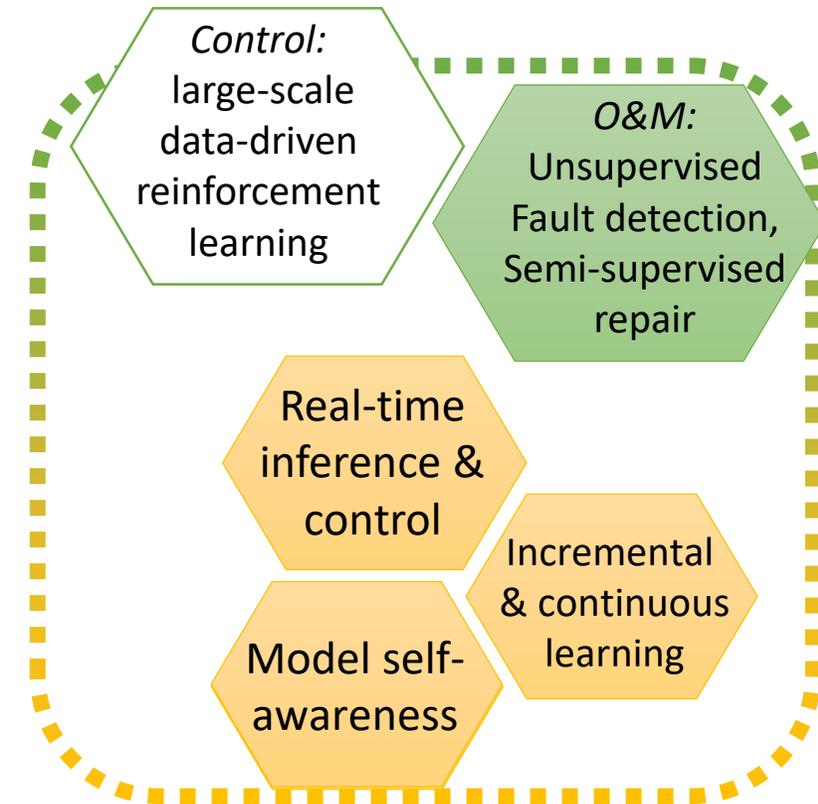
- Model ages and data drifts: study ageing of models imperative for deployment in a full AI lifecycle

❑ Incremental training

- Incremental training: system-level problems bring algorithmic challenges

❑ Real-time inference

- Inference: real-time low cost accurate inference



Large scale, heterogeneous RI

=> critical piece to stress test generalization & transfer

Large user community

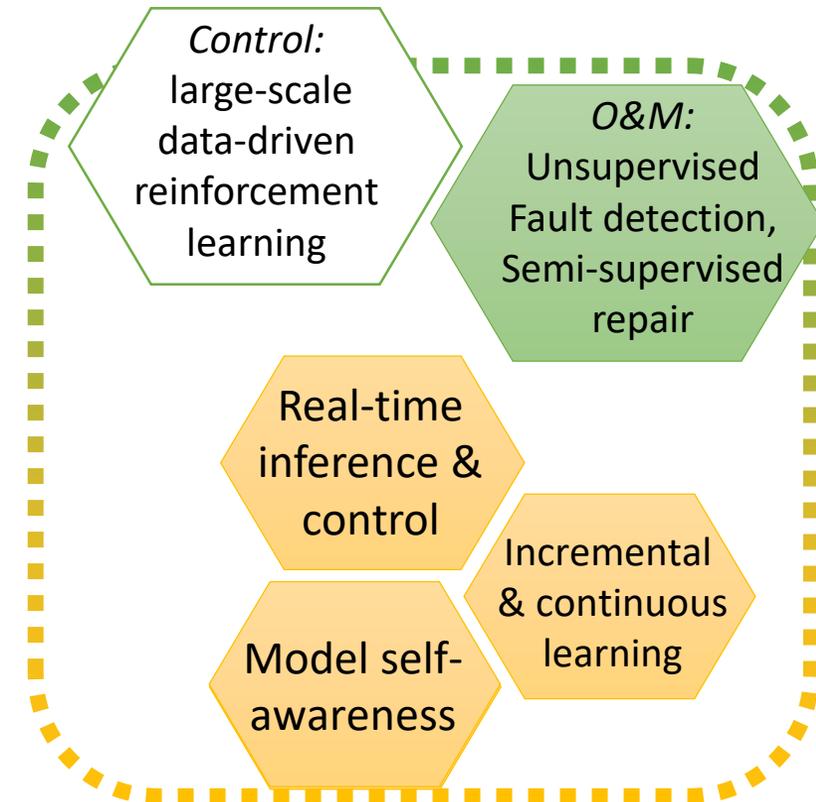
=> critical mass for crowd-sourcing labeling expertise



AI 4 Network viewpoint

❑ Model-driven O&M

- Unsupervised algorithms still need ground truth for benchmark
- Large SLICES crowd: can the community crowdsource anomaly detection database beyond KDD99 (s/ImageNet/AnomalyNet/)?



Unsupervised MDT O&M

Anomaly detection & root-cause analysis

Challenges

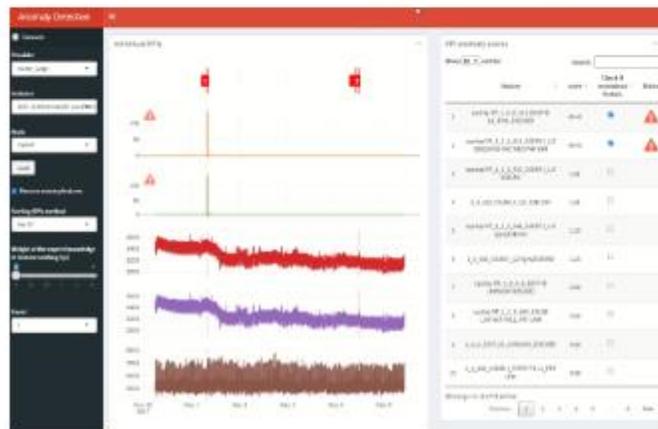
- ❑ Anomalies are rare -> unsupervised learning
- ❑ Need to correlate faults -> multi-variate methods
- ❑ Need to correlate KPI and logs -> multi-mode methods
- ❑ Cannot store all data -> stream-based learning
- ❑ Interact with operator -> explainability (XAI)

❑ Algorithm benchmarking

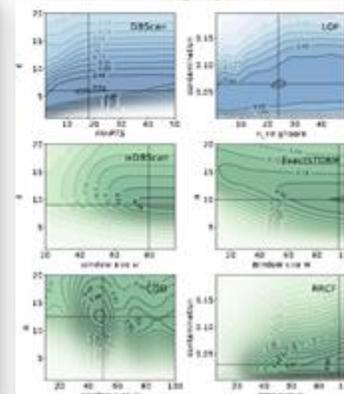
- ❑ Still requires labels!

❑ Algorithm selection & tuning

- ❑ Requires skills and time
- ❑ Lot of algorithms exist
- ❑ Each algorithm has many hyper-parameters



Algorithm	Stream/Batch
<i>Proximity-based using distance</i>	
KNN [20], [21]	Batch
DB-outlier [22]	Batch
STORM [23]	Stream
<i>Proximity-based using density</i>	
DBSCAN [25]	Batch
LOF [43]	Batch
LOCI [28]	Batch
HC3 [14]	Batch
ABOD [33]	Batch
<i>Proximity-based using clustering</i>	
CBLOF [34]	Batch
MCOD [35]	Stream
CAEDS [36]	Stream
<i>Ensemble-based using tree</i>	
IF [38]	Batch
RHF [39]	Batch
RRCP [40]	Stream
HST [41]	Stream
<i>Ensemble-based using subspaces</i>	
Feature bagging [42]	Batch
RS-Hash [19]	Stream/Batch
Loda [17]	Stream/Batch
sStream [18]	Stream



Large scale, heterogeneous RI

=> critical piece to stress test generalization & transfer

Large user community

=> critical mass for crowd-sourcing labeling expertise



AI 4 Network viewpoint

❑ Model-driven O&M

- Unsupervised algorithms still need ground truth for benchmark
- Large SLICES crowd: can the community crowdsource anomaly detection database beyond KDD99 (s/ImageNet/AnomalyNet/)?

❑ Heterogeneity (again)

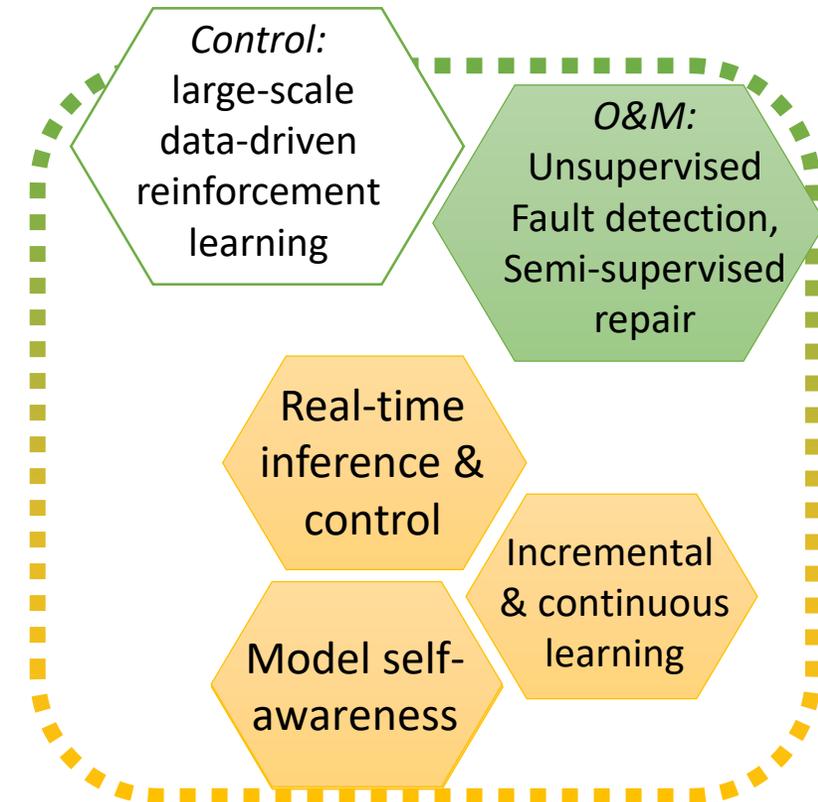
- Model ages and data drifts: study ageing of models imperative for deployment in a full AI lifecycle

❑ Incremental training

- Incremental training: system-level problems bring algorithmic challenges

❑ Real-time inference

- Inference: real-time low cost accurate inference



Large scale, heterogeneous RI

=> critical piece to stress test generalization & transfer

Large user community

=> critical mass for crowd-sourcing labeling expertise



AI 4 Network viewpoint

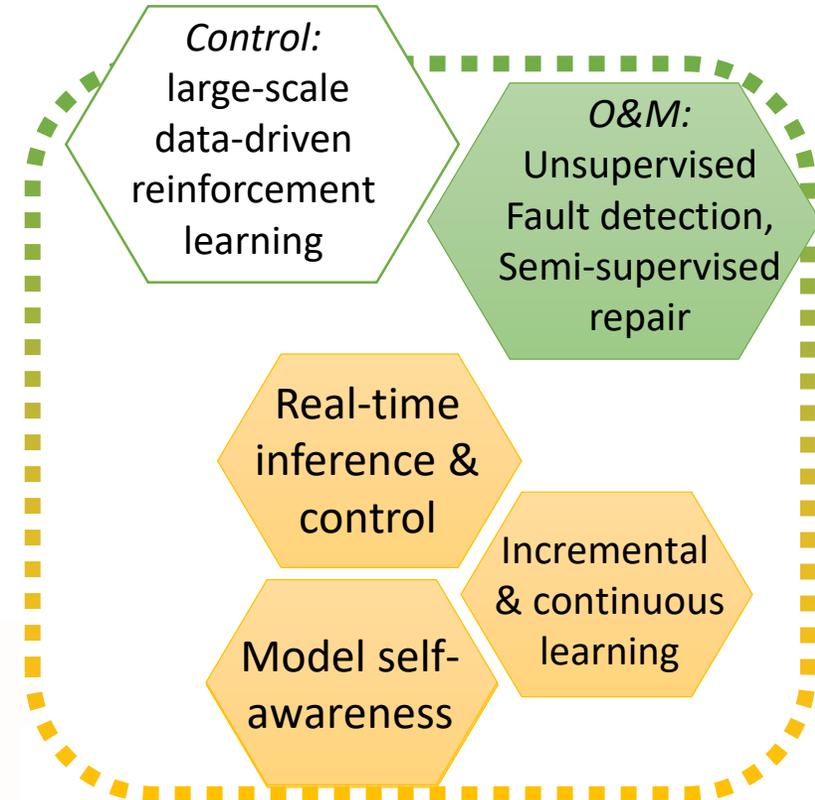
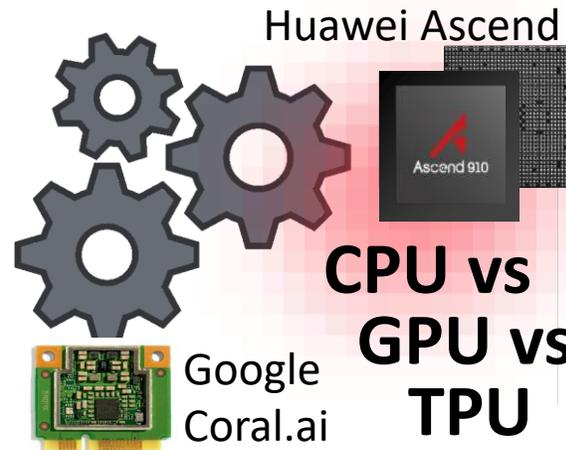
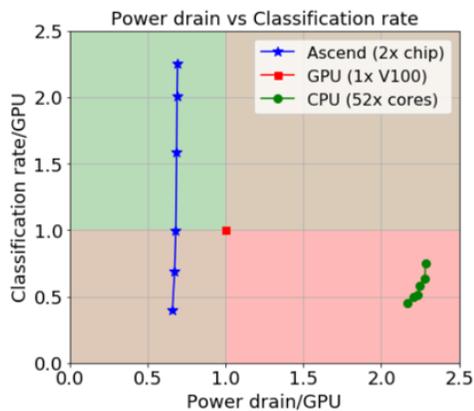
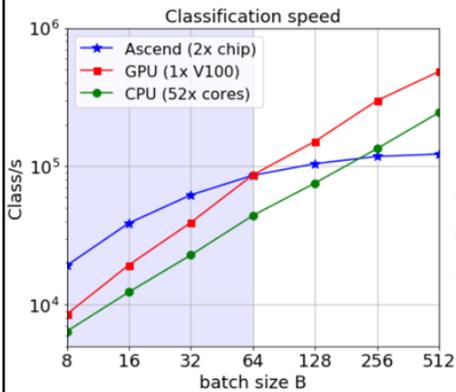
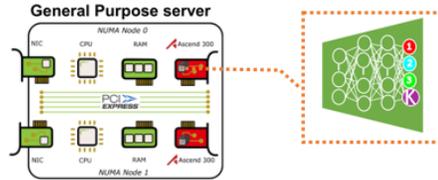
Real-time Inference

Experimental setup

- 2 Servers Intel Xeon Platinum 8164 CPUs @2.00GHz (L1/L2/L3 caches: 32 data+32 instruction/1MB/36MB)
- 1.5TB RAM (64GBx24 DDR4 @ 2666MT)
- 100 Gbps NICs (Mellanox MCX515A-CCAT ConnectX-5)
- Huawei Atlas 300I:3010 Inference Card (4x Ascend 310)

Input traffic

- 3 datasets (2 internal and 1 publicly available)
- Adversarial analysis at 100 Gbps (speedup traces): 30-50kclass/sec depending on scenario



Large scale, heterogeneous RI
=> critical piece to stress test generalization & transfer

Real-time inference

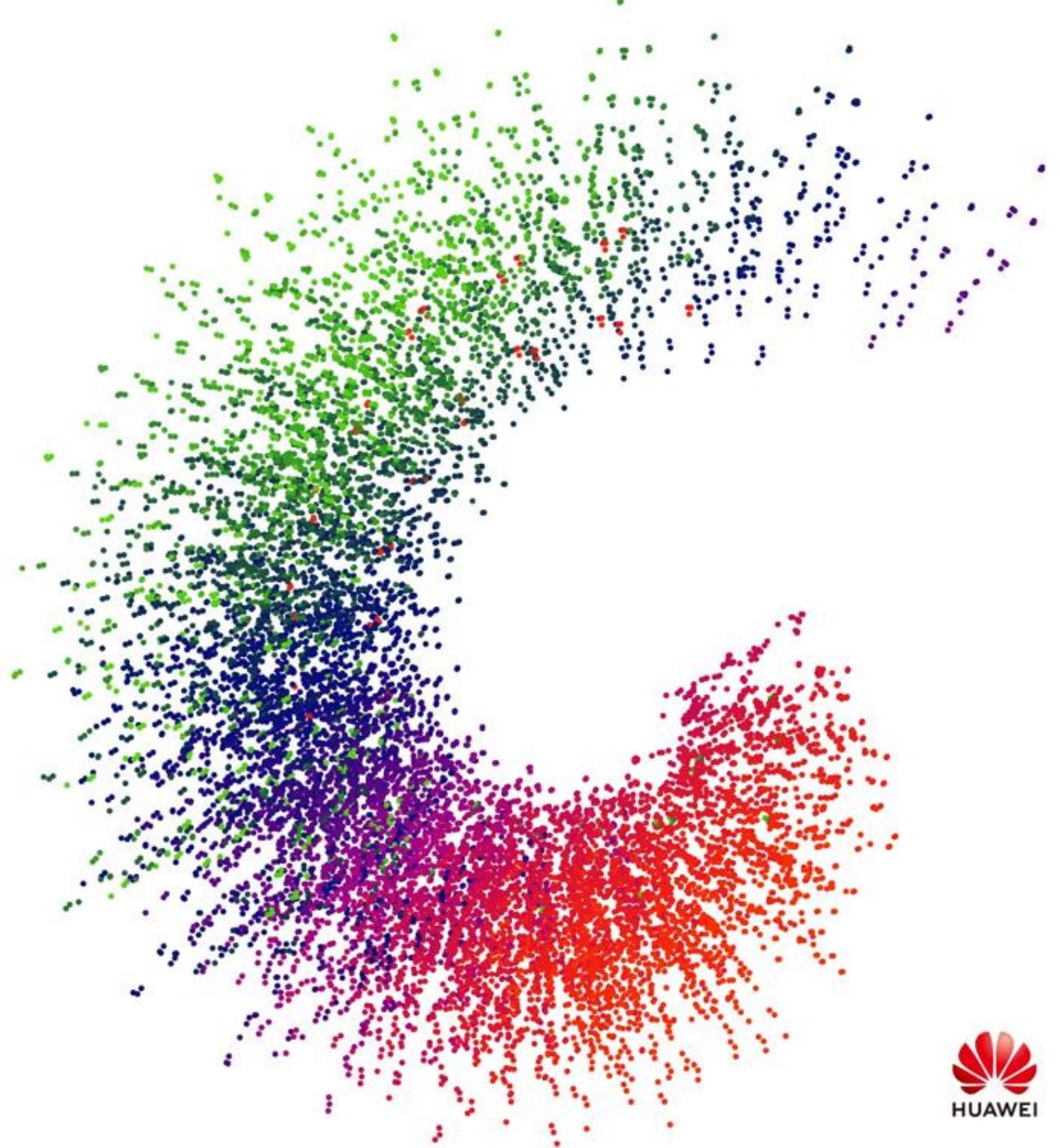
- Inference: real-time low cost accurate inference

Large user community

=> critical mass for crowd-sourcing labeling expertise



Thanks



Dario Rossi

dario.rossi@huawei.com

public research resources

<https://nonsns.github.io>