Autonomous Driving SLICES

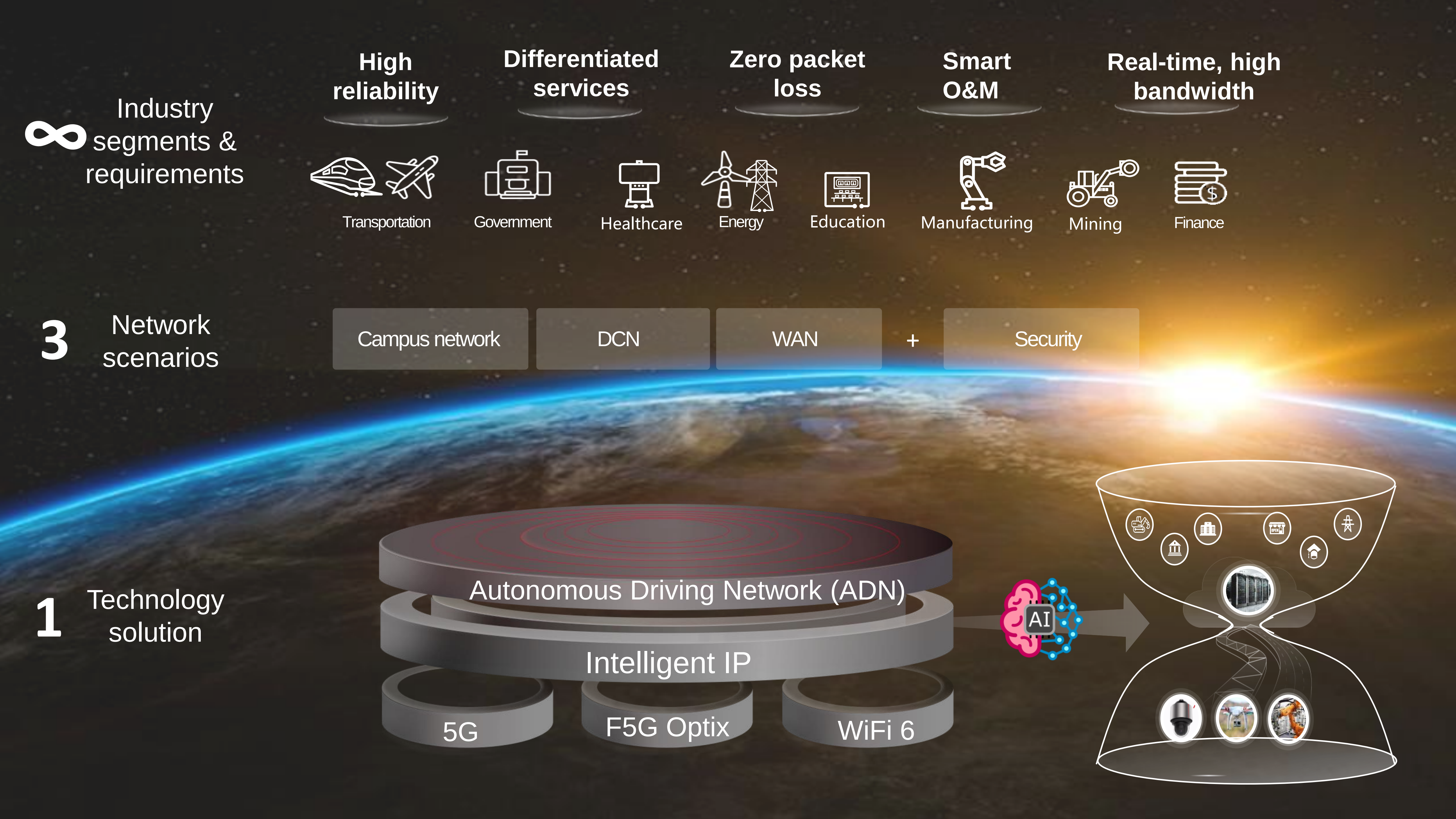
what SLICES can bring to Network AI research

Dario Rossi
dario.rossi@huawei.com

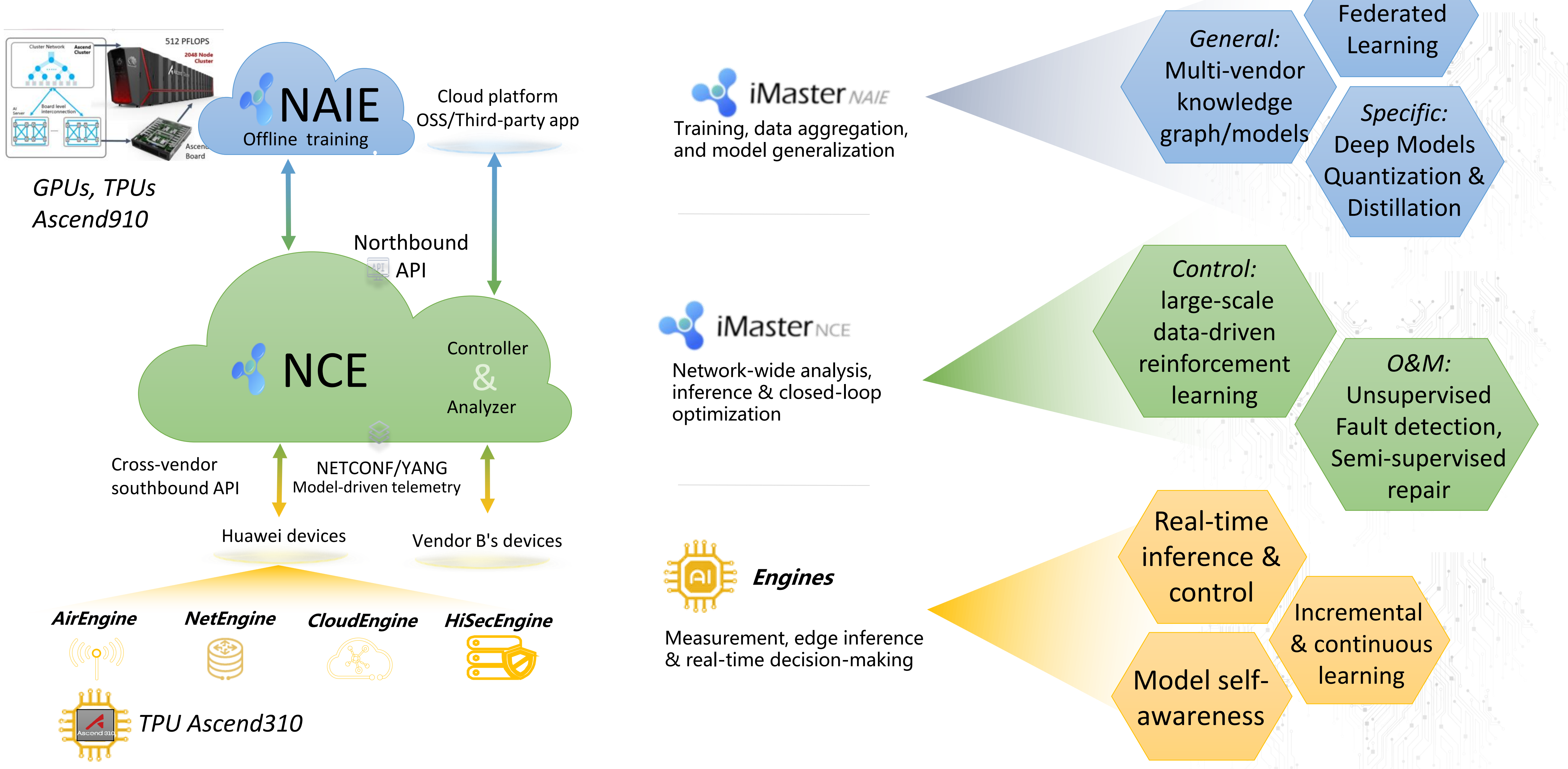
Huawei Network AI CTO
Director, DataCom Lab, Paris Research Center

Artificial Intelligence (AI) & Machine Learning (ML)

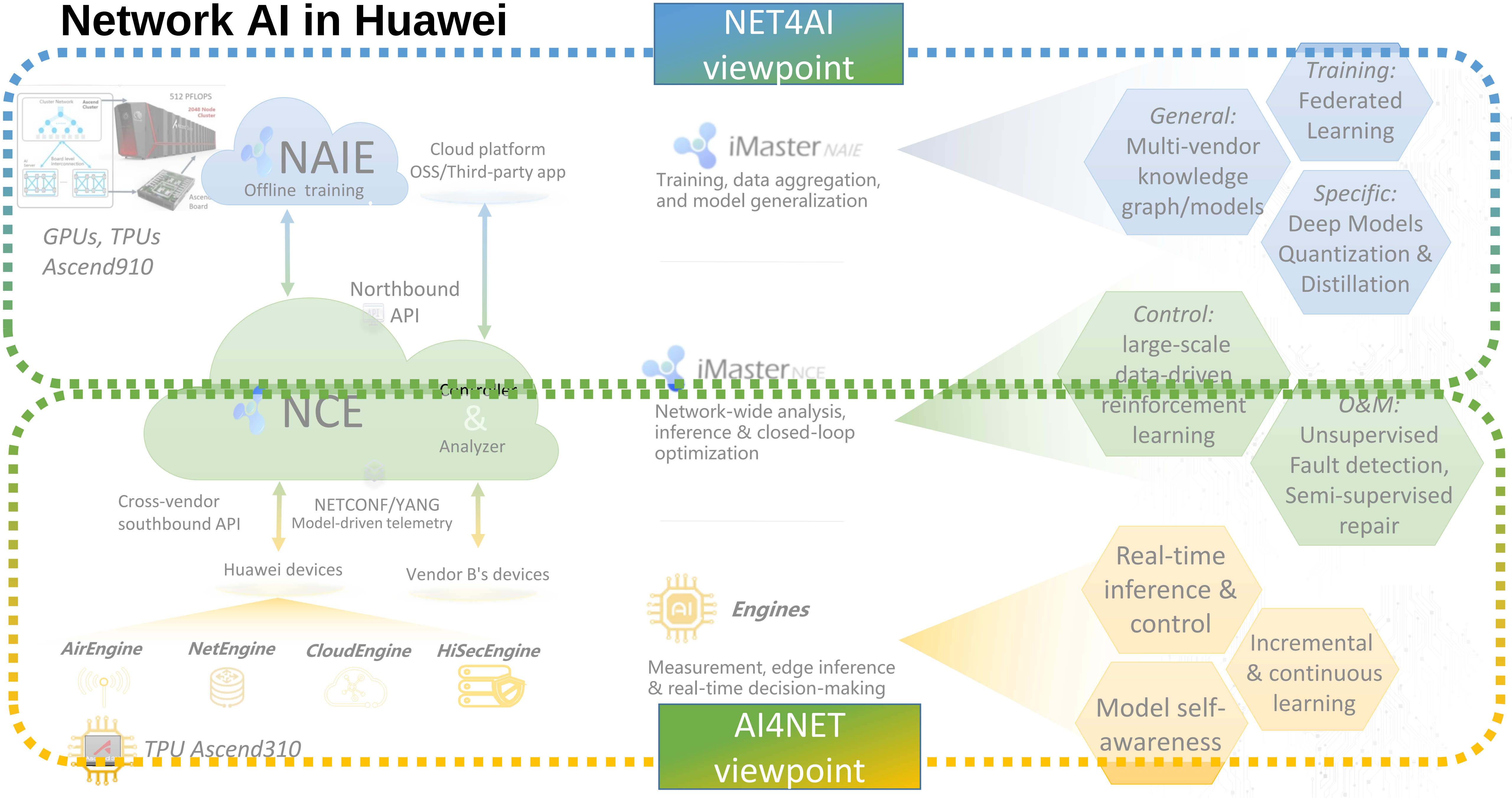

A new dawn
AI
Autonomous driving
network


High reliability
Differentiated services
Zero packet loss
Smart O&M
Real-time, high bandwidth
∞ Industry segments & requirements
Transportation
Government
Healthcare
Energy
Education
Manufacturing
Mining
Finance
3 Network scenarios
Campus network
DCN
WAN
+
Security
1 Technology solution
Autonomous Driving Network (ADN)
Intelligent IP
5G
F5G Optix
WiFi 6
AI

Network AI in Huawei



Network AI in Huawei



Network 4 AI viewpoint

❑ Model training

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

❑ Model-driven telemetry (MDT)

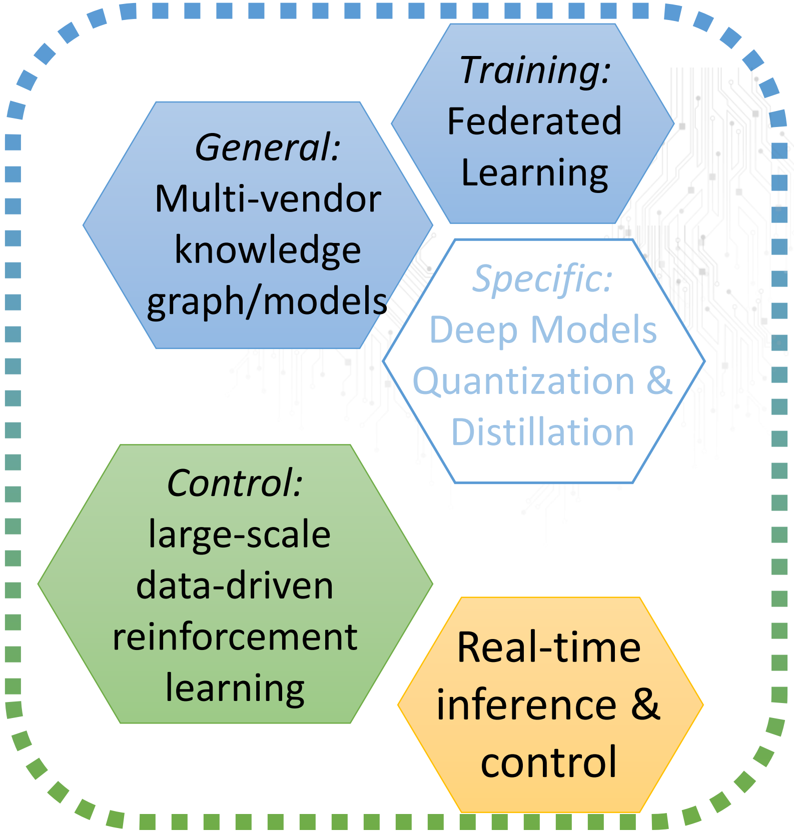
- Heterogenity in the input data: multi-vendor (good to have “dirty data” AI problem)

❑ 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

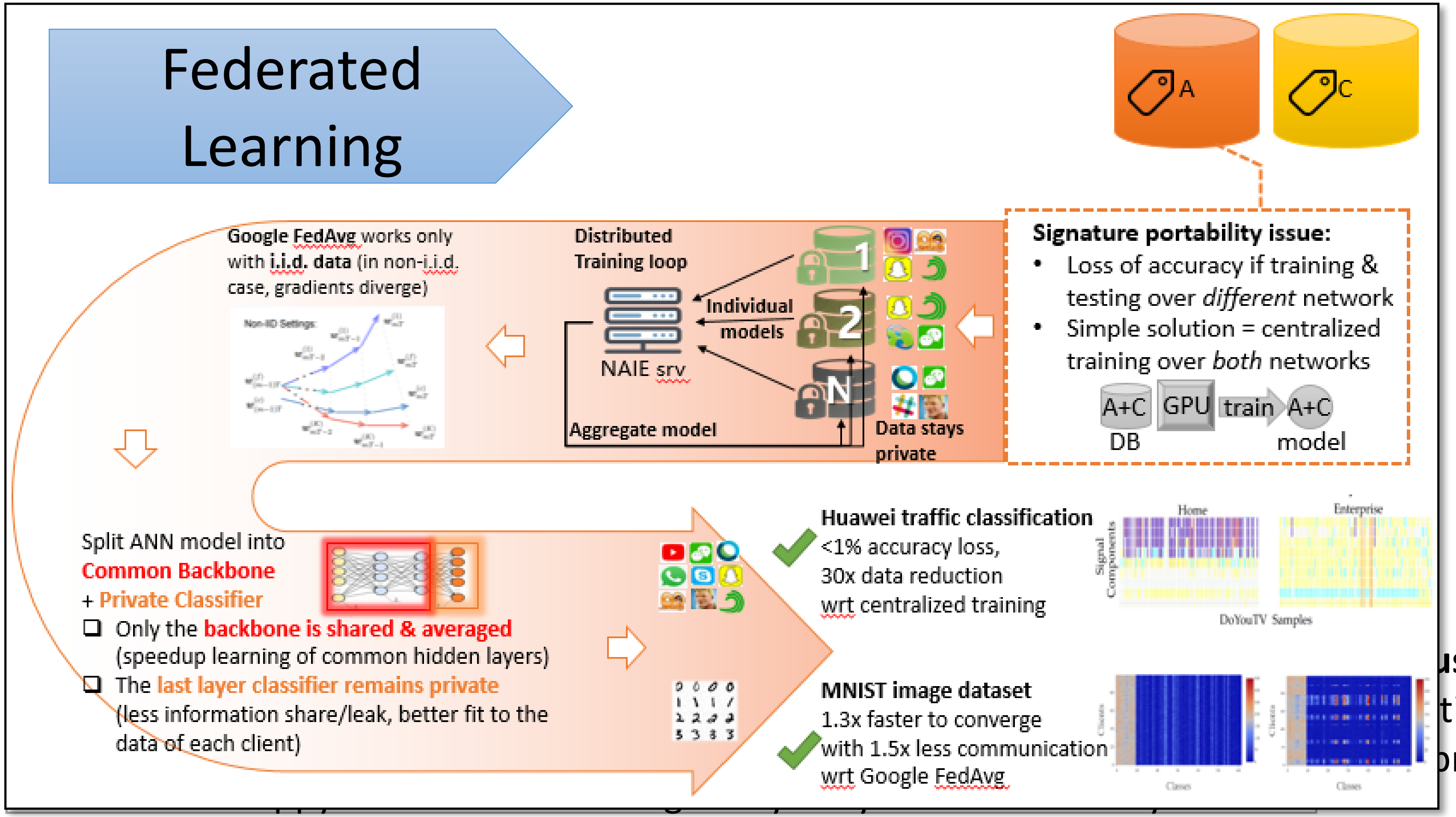
Large user community
=> critical mass to push reproducibility standards

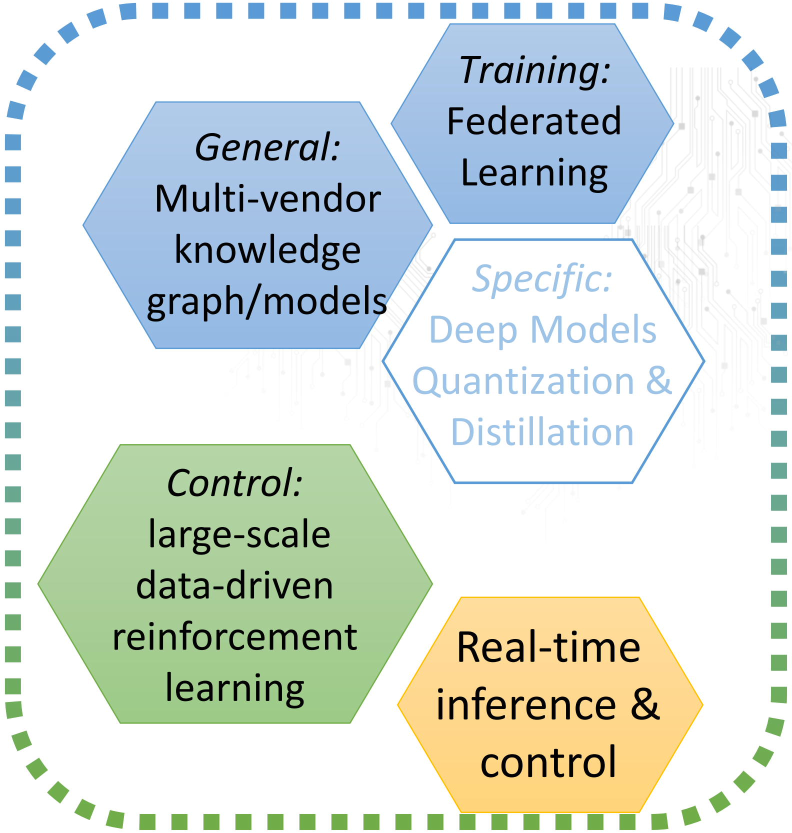
slices

Network 4 AI viewpoint

❑ Model training

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





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

user community
tical mass to push
producibility standards

slices

Network 4 AI viewpoint

❑ Model training

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

❑ Model-driven telemetry (MDT)

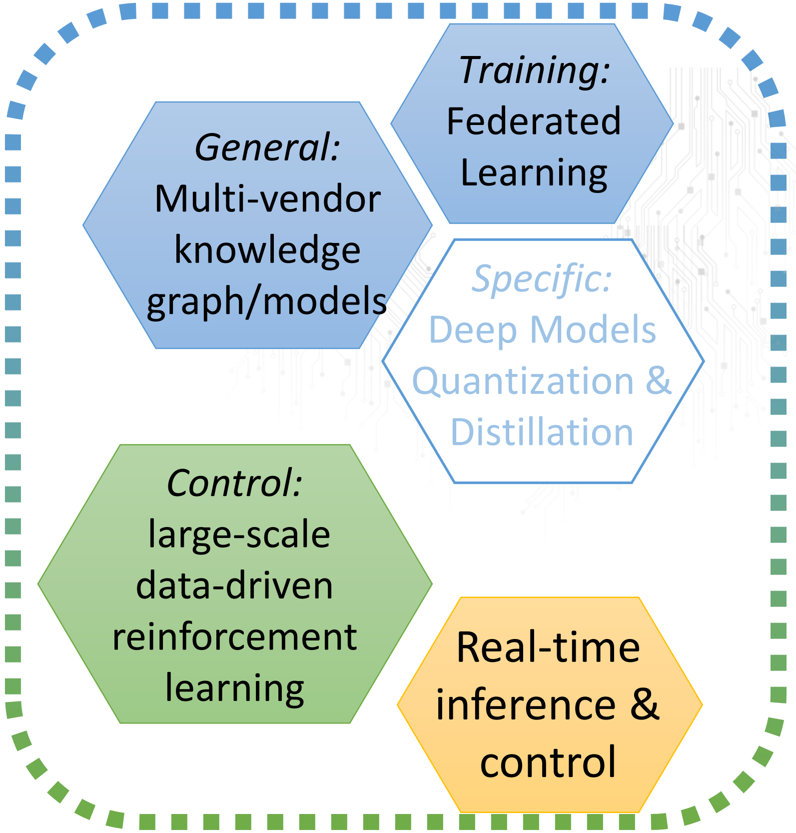
- Heterogenity in the input data: multi-vendor (good to have “dirty data” AI problem)

❑ Real-time

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

❑ Control

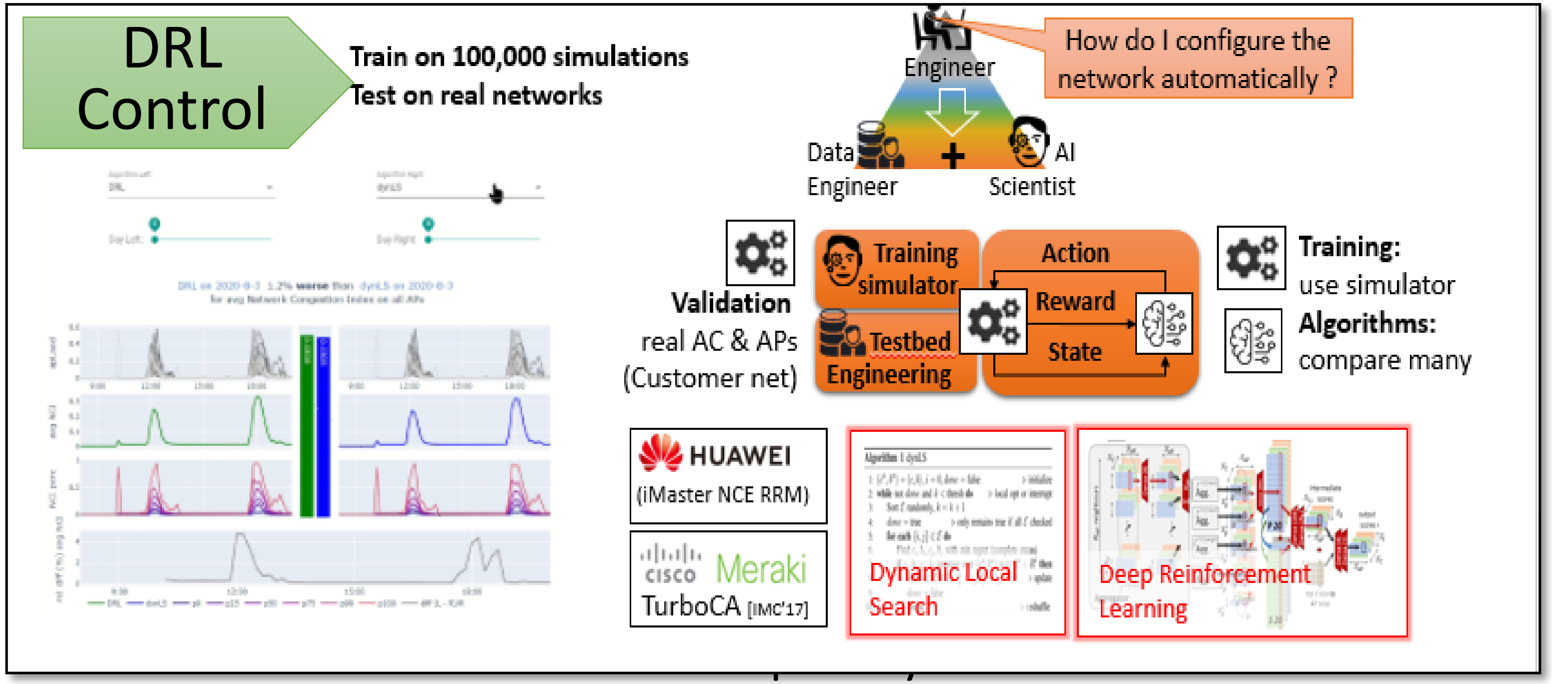
- Delay+noise of MDT data streams: controllable/reproducible AI experiment in more challenging environment
- Train on simulation (e.g., DRL takes lifetimes, cannot learn from real network) refine & validate on SLICES

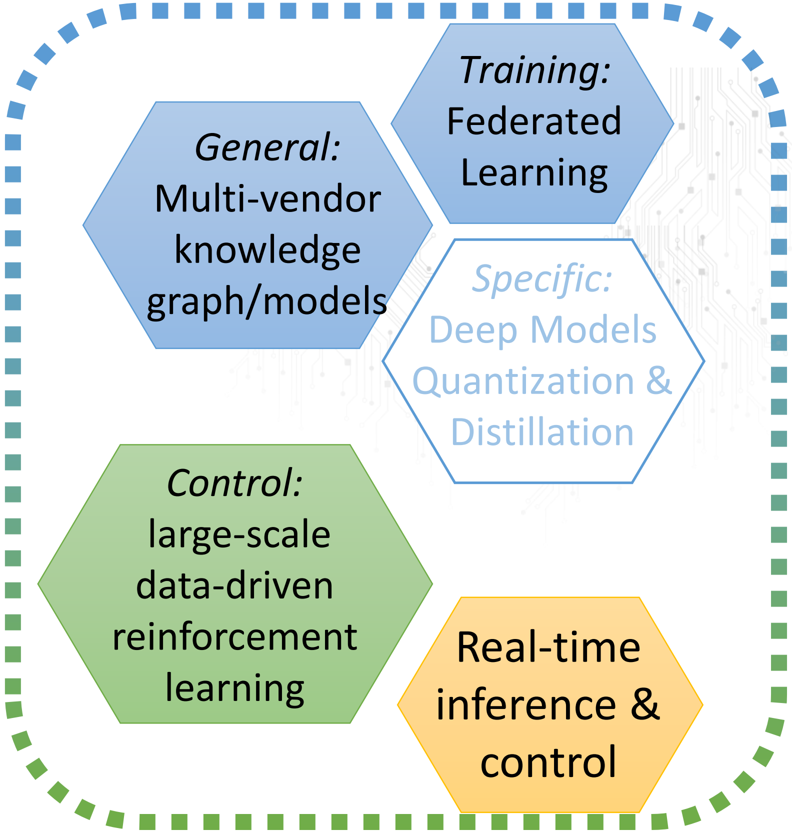


Large scale, heterogeneous RI
=> lower access barrier to experimental study & more realistic challenges

Large user community
=> critical mass to push reproducibility standards

Network 4 AI viewpoint





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

❑ 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 user community
=> critical mass to push reproducibility standards

slices

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/)?

❑ Heterogenity (again)

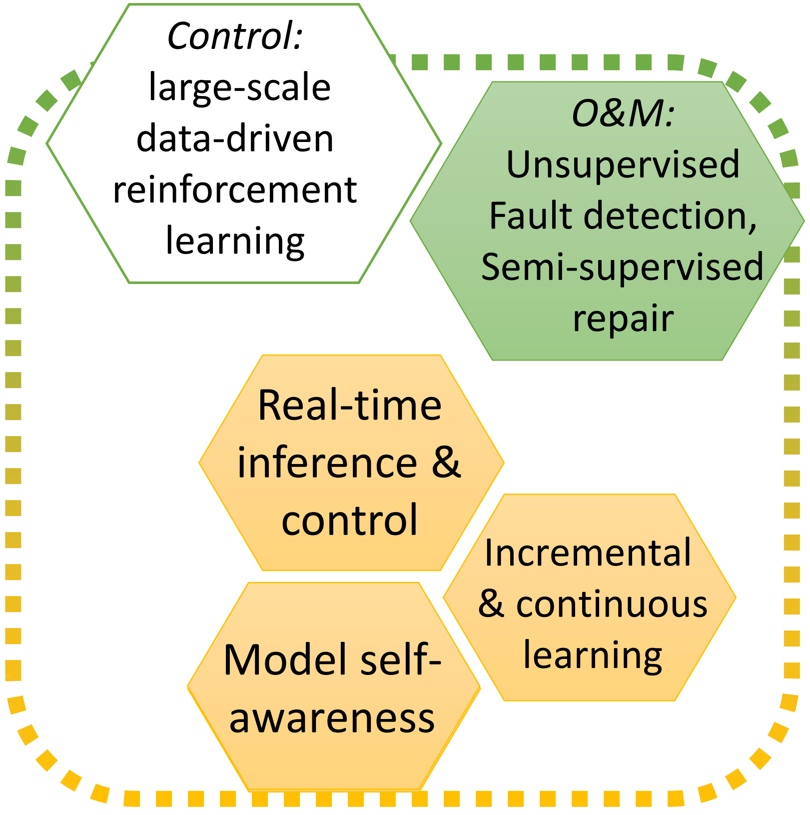
- Model ages and data drifts: study ageing of models imperative for deployment in a full AI lifecycle

❑ Incremental training

- Incremental training: system-level problems bring algorithmic challenges

❑ Real-time inference

- Inference: real-time low cost accurate inference



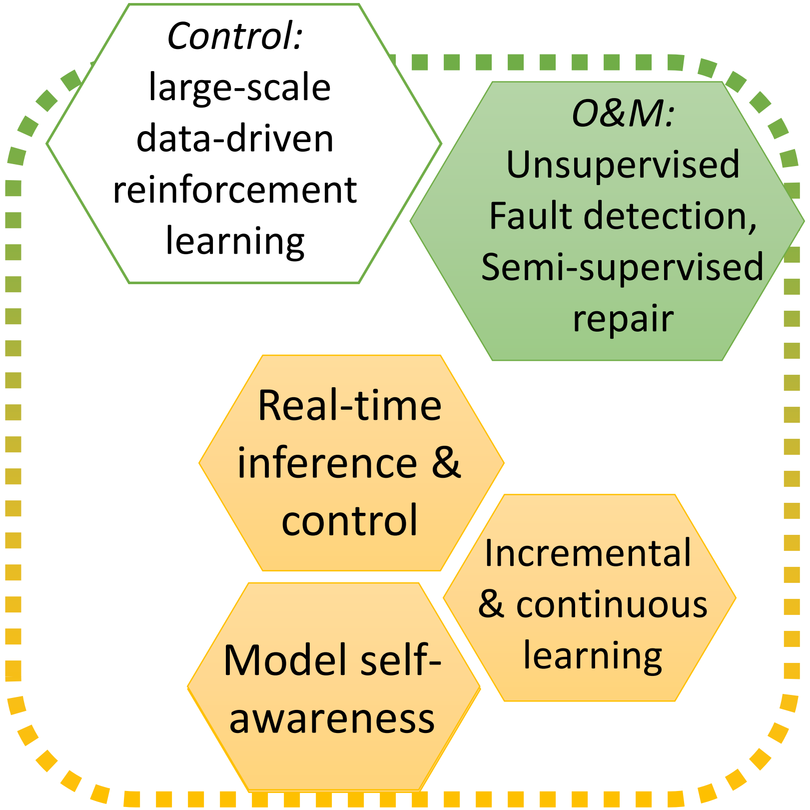
Large scale, heterogeneous RI
=> critical piece to stress test generalization & transfer

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

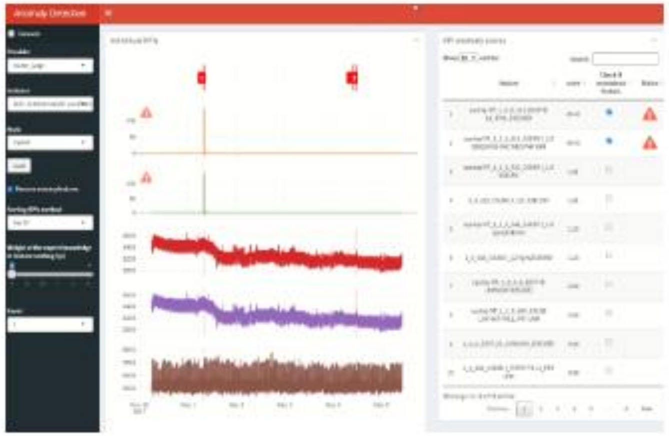
- ❑ Anomalies are rare -> unsupervised learning
- ❑ Need to correlate faults -> multi-variate methods
- ❑ Need to correlate KPI and logs -> multi-mode methods
- ❑ Cannot store all data -> stream-based learning
- ❑ Interact with operator -> explainability (XAI)

❑ Algorithm benchmarking

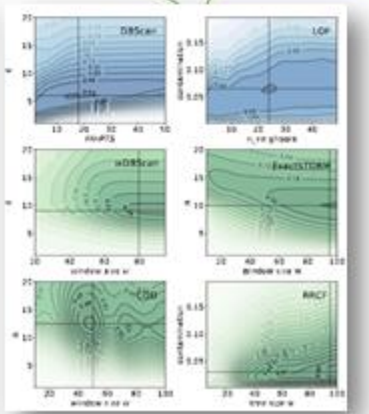
- ❑ Still requires labels !

❑ Algorithm selection & tuning

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



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



Large scale, heterogeneous RI

=> critical piece to stress test generalization & transfer

Large user community

=> critical mass for crowd-sourcing labeling expertise

slices

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/)?

❑ Heterogenity (again)

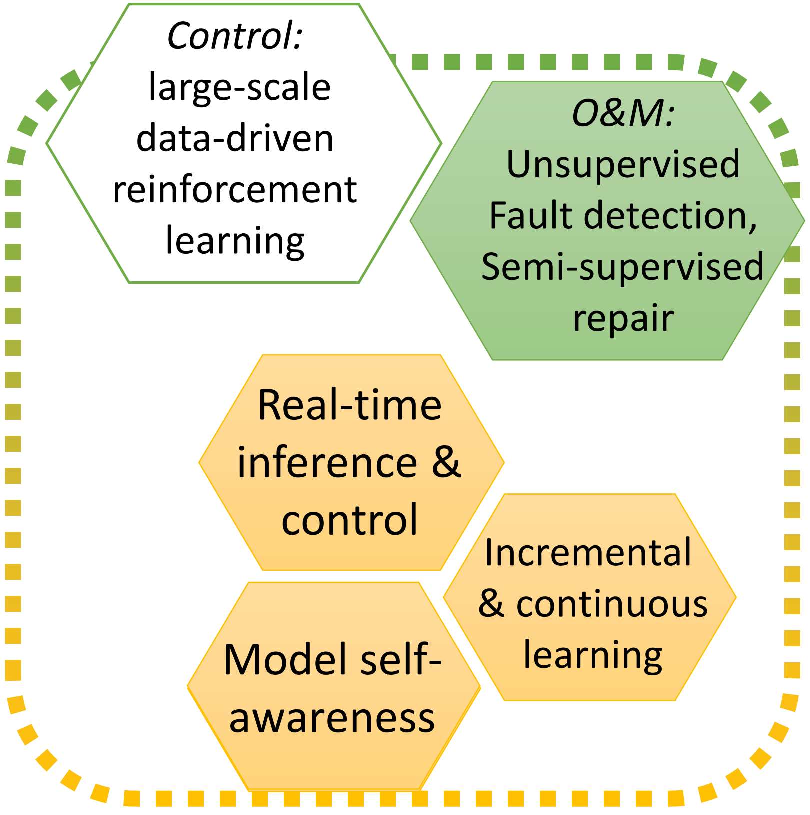
- Model ages and data drifts: study ageing of models imperative for deployment in a full AI lifecycle

❑ Incremental training

- Incremental training: system-level problems bring algorithmic challenges

❑ Real-time inference

- Inference: real-time low cost accurate inference



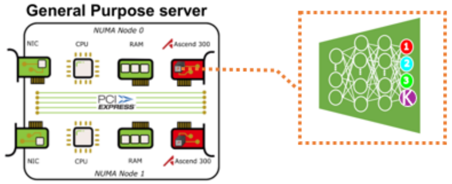
Large scale, heterogeneous RI
=> critical piece to stress test generalization & transfer

Large user community
=> critical mass for crowd-sourcing labeling expertise

slices

AI 4 Network viewpoint

Real-time Inference

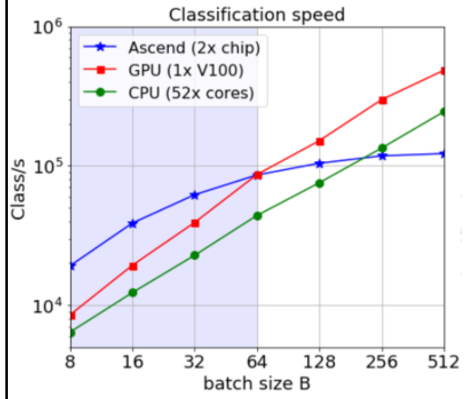


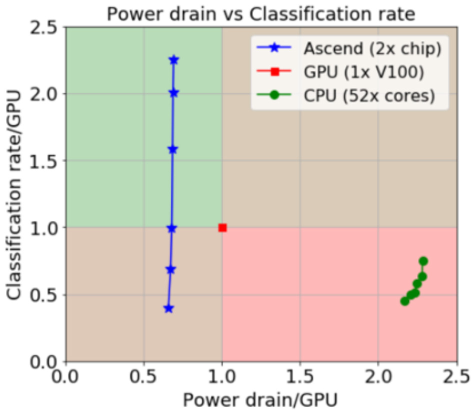
Experimental setup

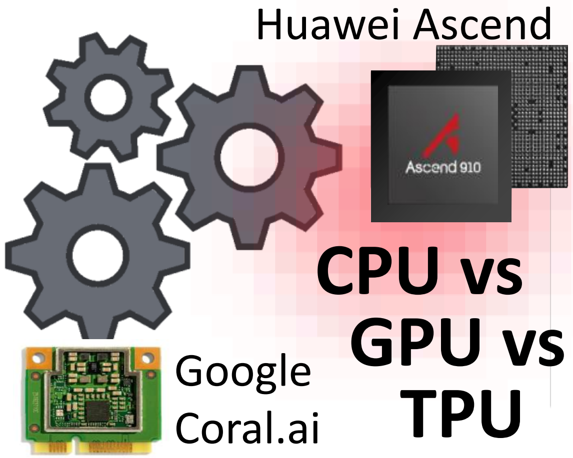
- ❑ 2 Servers Intel Xeon Platinum 8164 CPUs @2.00GHz (L1/L2/L3 caches: 32 data+32 instruction/1MB/36MB)
- ❑ 1.5TB RAM (64GBx24 DDR4 @ 2666MT)
- ❑ 100 Gbps NICs (Mellanox MCX515A-CCAT ConnectX-5)
- ❑ Huawei Atlas 300I:3010 Inference Card (4x Ascend 310)

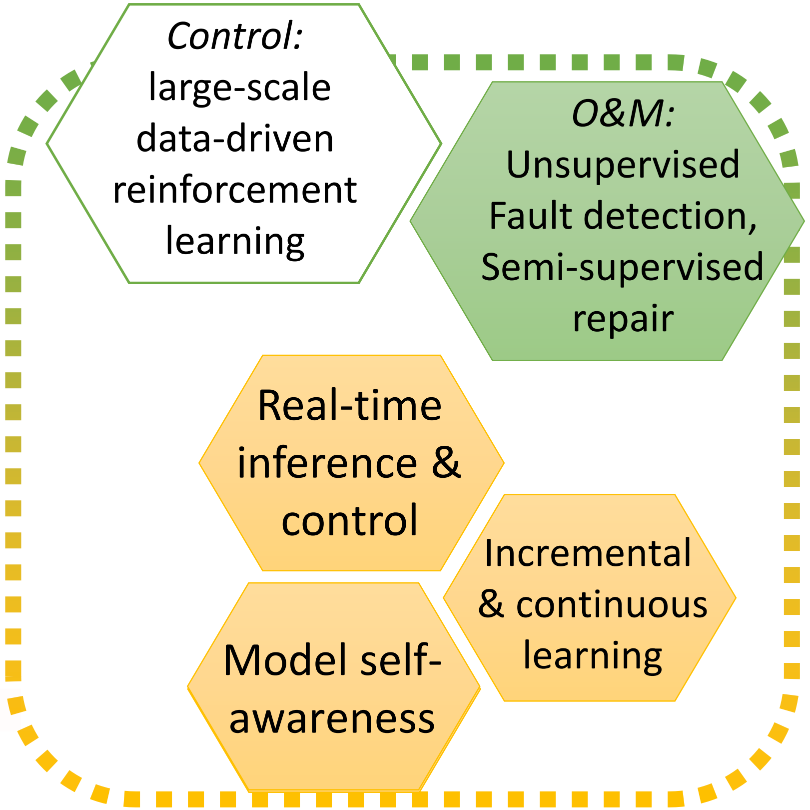
Input traffic

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









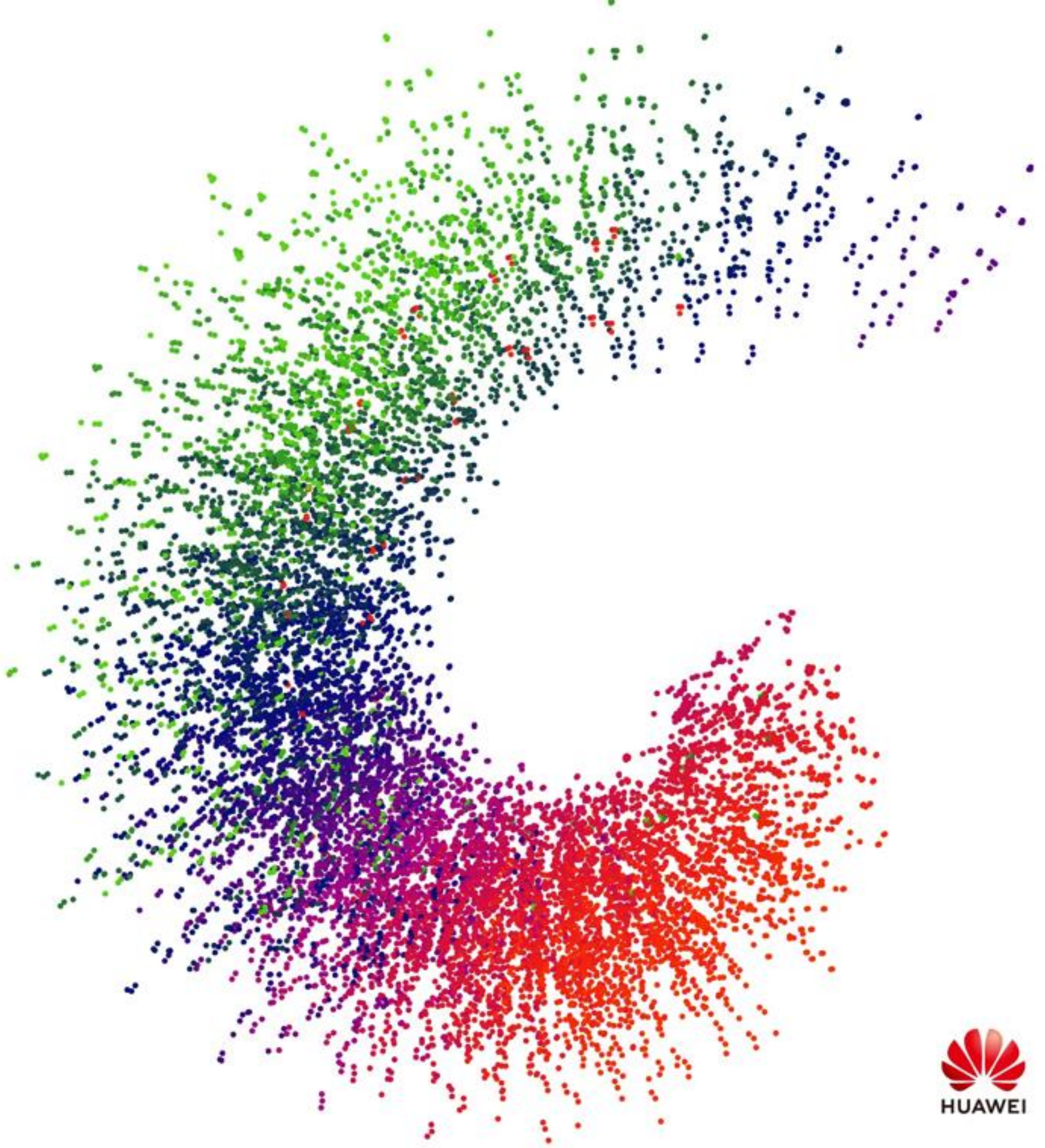
Large scale, heterogeneous RI
=> critical piece to stress test generalization & transfer

Large user community
=> critical mass for crowd-sourcing labeling expertise

slices

- ❑ **Real-time inference**
  - Inference: real-time low cost accurate inference

Thanks



Dario Rossi
dario.rossi@huawei.com
public research resources
https://nonsns.github.io

HUAWEI