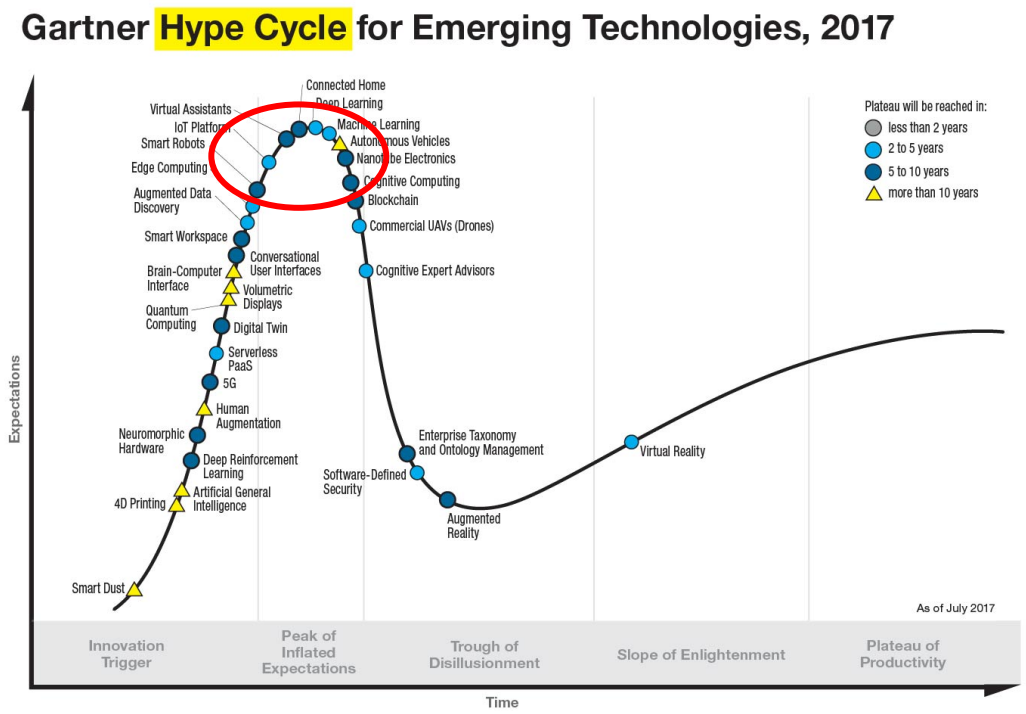
Artificial Intelligence use cases in 5G Networks

Presenter:
Bendek Kovács, PhD
Senior Specialist, Network Service Performance,
Head of Technology and Innovation ETH R&D, Ericsson

AI, IoT and Digitalization Today



GOOGLE, FACEBOOK, AND MICROSOFT ARE REMAKING THEMSELVES AROUND AI

Wired.com

At Sundar Pichai's Google, AI Is Everything—And Everywhere

Fastcompany.com

GAFA: AI Is the Future

Guardian.com

2017 Is The Year Of Artificial Intelligence

forbes.com

McKinsey&Company

"We estimate a potential economic impact of as much as $11.1 trillion per year in 2025 for IoT applications."

pwc

"AI could contribute up to $15.7 trillion1 to the global economy in 2030, more than the current output of China and India combined."

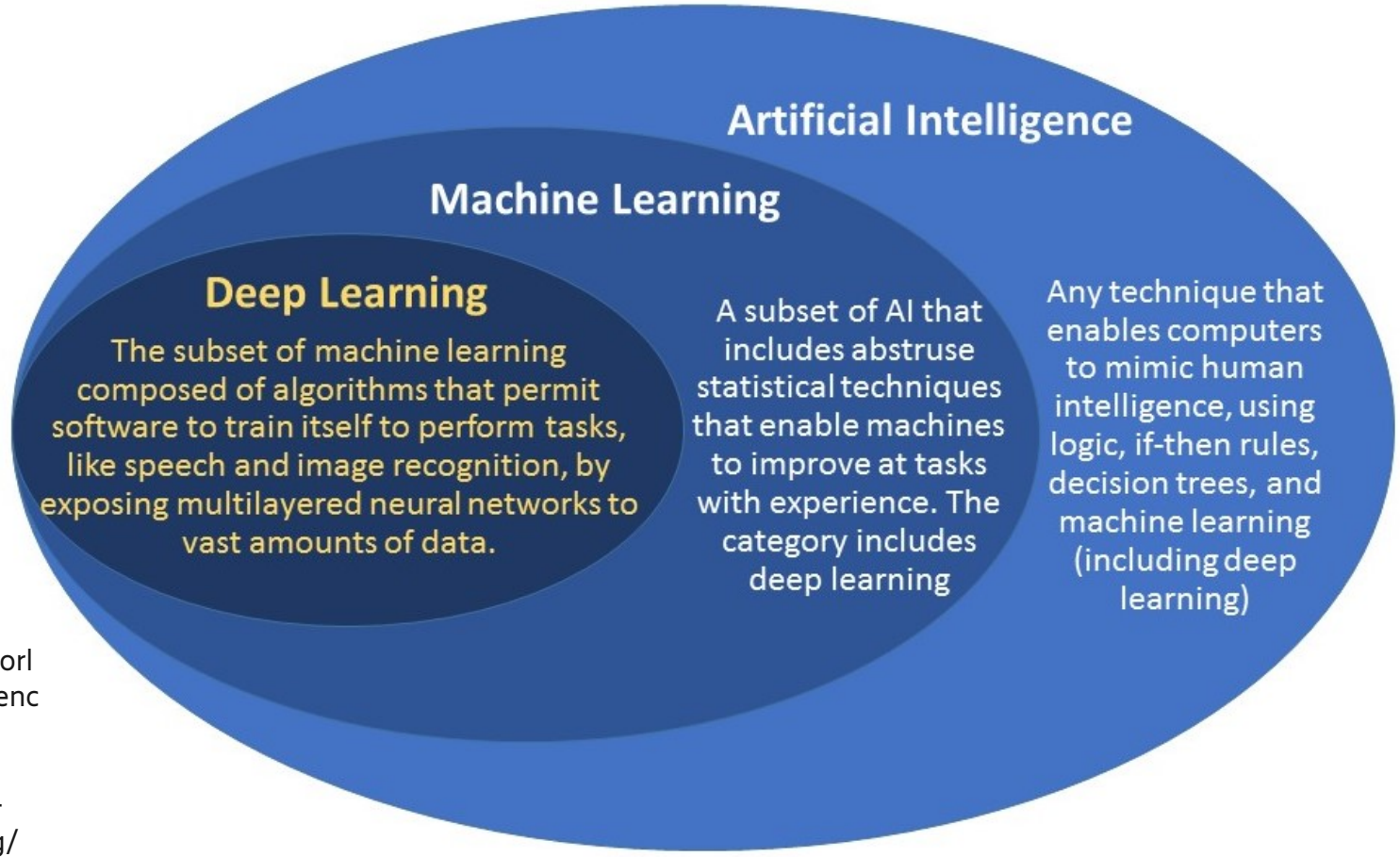
Why Machine Learning Now?

Why Machine Learning Now?

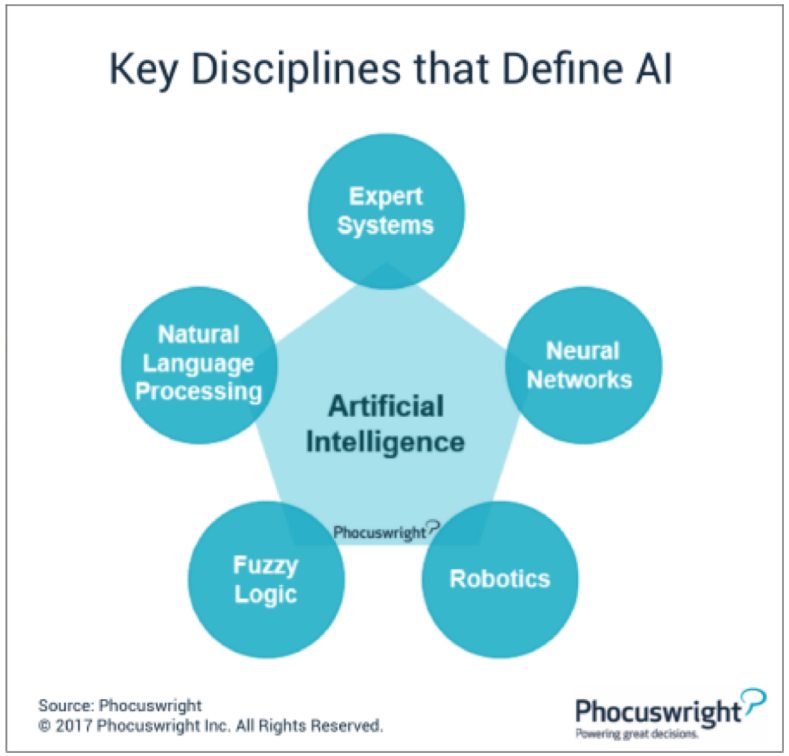
Science

Neural Networks has been researched for a long time
=> Requires compute and data

AI and Machine Learning



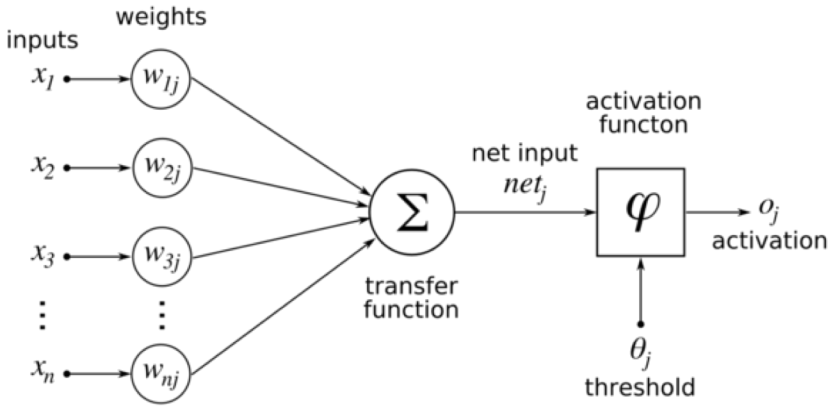
Source: www.geospatialworld.net/blogs/difference-between-ai%EF%BB%BF-machine-learning-and-deep-learning/



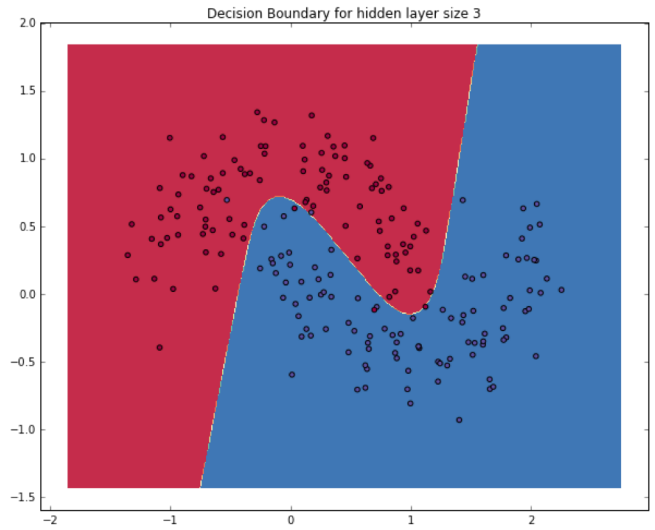
Narrow AI, artificial general intelligence (AGI), and superintelligent AI.

Machine learning

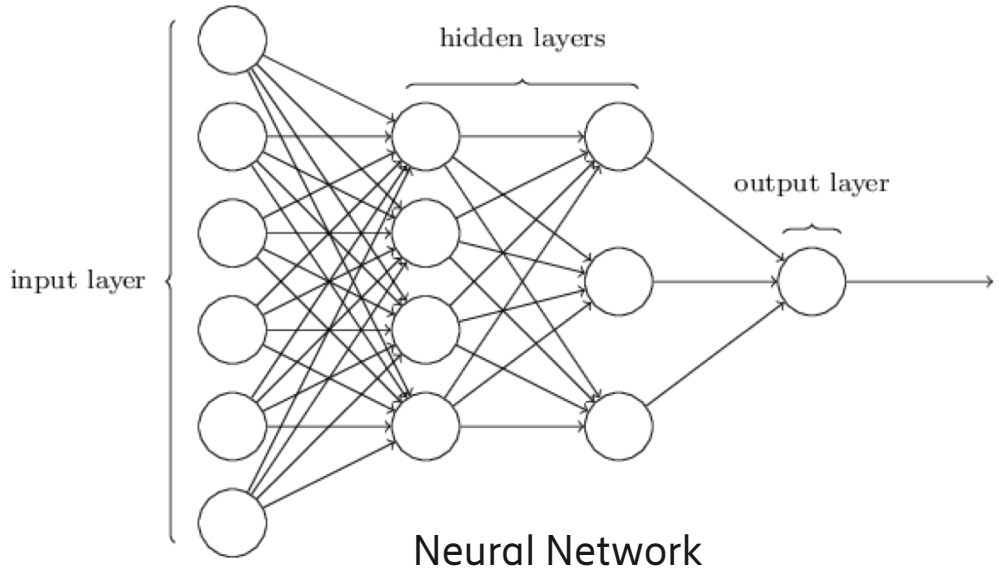
- Machine Learning
  - Field of study that gives computers the ability to learn without being explicitly programmed
- Neural networks
  - Nonlinear function approximator
- Computationally intensive, lots of linear algebra
  - Multiplying matrices and vectors



Perceptron

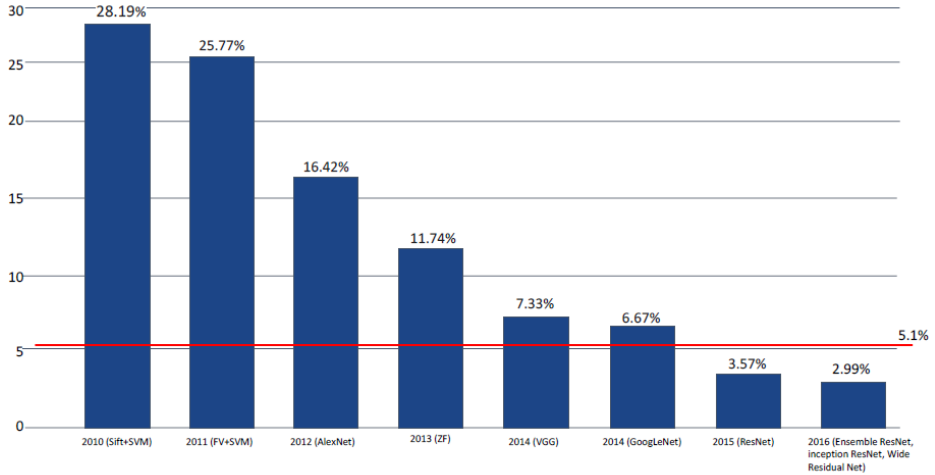


Classification



Neural Network

Science: Deep Learning



Object Detection of the ImageNet Large Scale Visual Recognition challenge.
Red line is human performance on the same benchmark (Andrej Karpathy).

Computer Vision

Acoustic model	Recog WER	RT03S FSH	Hub5 SWB
Traditional features	1-pass –adapt	**27.4**	**23.6**
Deep Learning	1-pass –adapt	**18.5** (–33%)	**16.1** (–32%)

Language Processing

- Significant developments in the recent years
  - Large set of data
  - HW (GPU-s)
  - Algorithms including
    - Learning refinements
    - CNNs
    - Word vector representations
- Resulting
  - Larger models
  - Trained faster
  - On more data
- **Feature learning**
  - **E2E Deep Learning**

Why Machine Learning Now?

Science

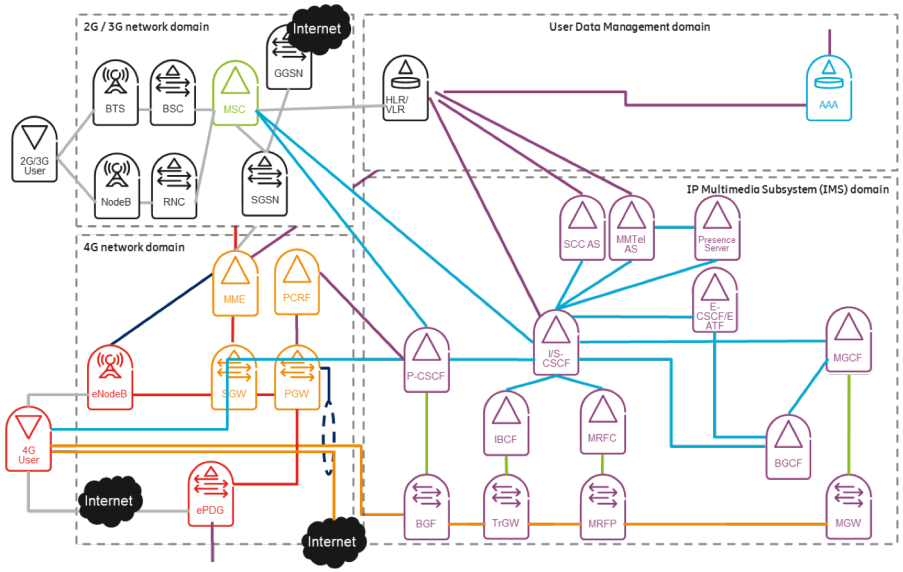
Neural Networks has been researched for a long time

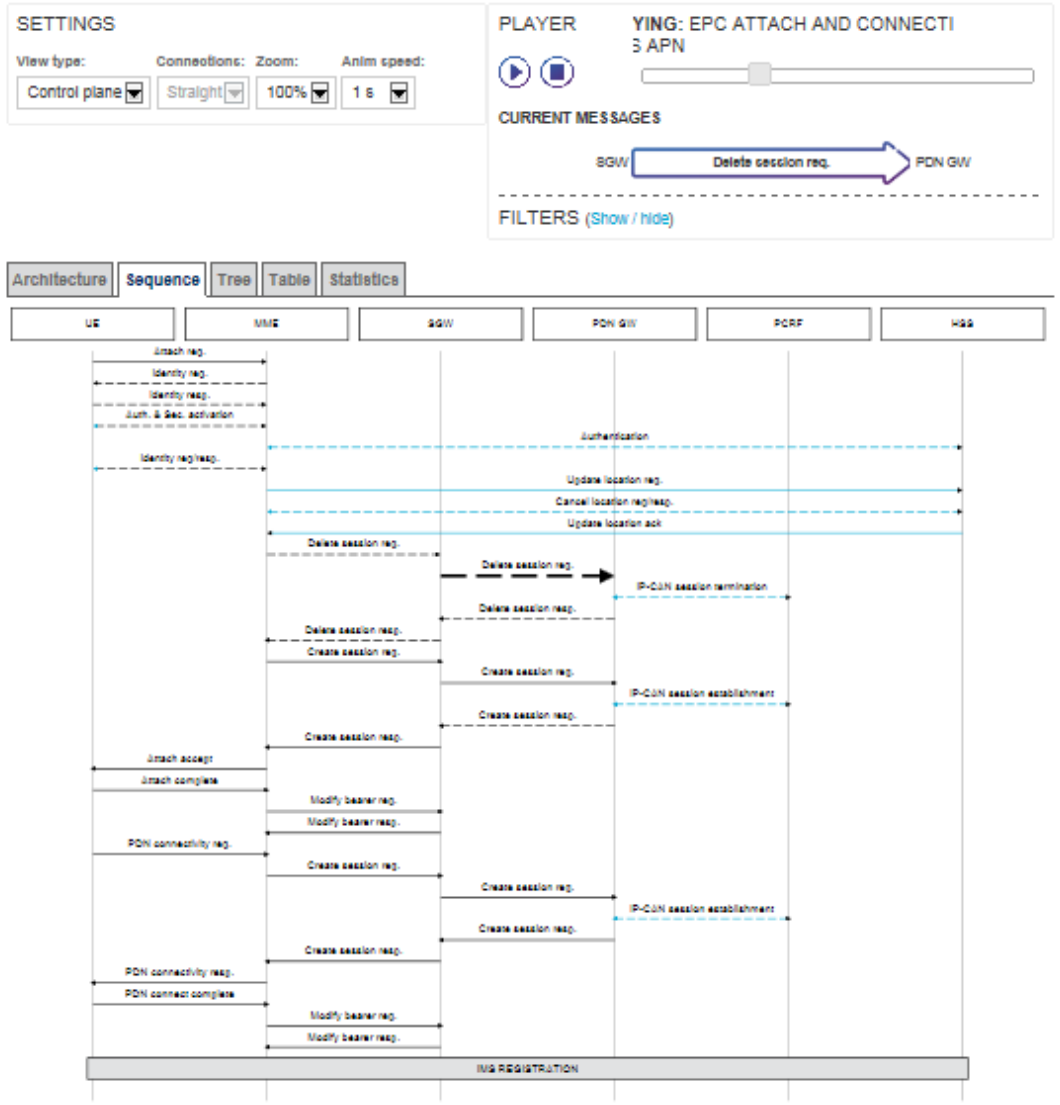
=> Requires compute and data

Use case: Network Signaling Analytics

Signaling Analytics: big data

- Call setup: 400 signaling
  => 3 million subscribers 1call/hour
  => 1.2 milliárd üzenet (1Gbyte/s)
- ~10 3GPP protocols
- Real-time analytics





Representation Learning: Clustering

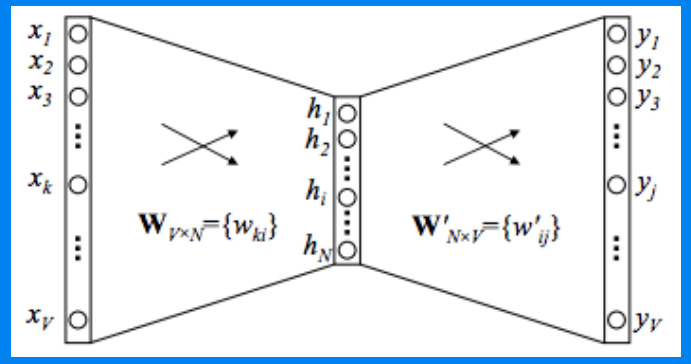
Data Cleaning

Representativity Analysis

Pre-filtering

Training set selection

Representation Learning



Entity embedding

Signaling Analysis

Similarity-based Sequence analytics

Malicious activity detection

Error detection

Network traces

Mobility sequences

Why Machine Learning Now?

Science

Neural Networks has been researched for a long time
=> Requires compute and data

Data

Big data systems producing data for analytics

Why Machine Learning Now?

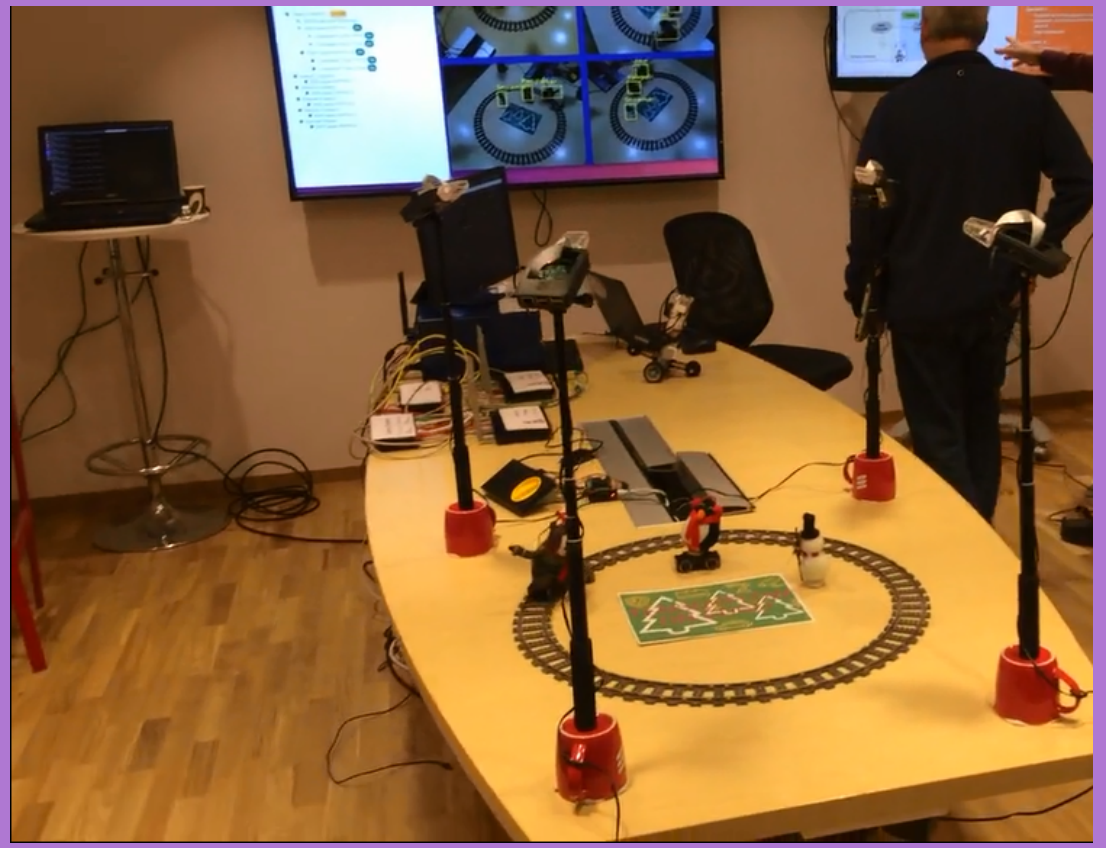
Science

Neural Networks has been researched for a long time
=> Requires compute and data

Data

Big data systems producing data for analytics
=> Requires global storage places, use cases

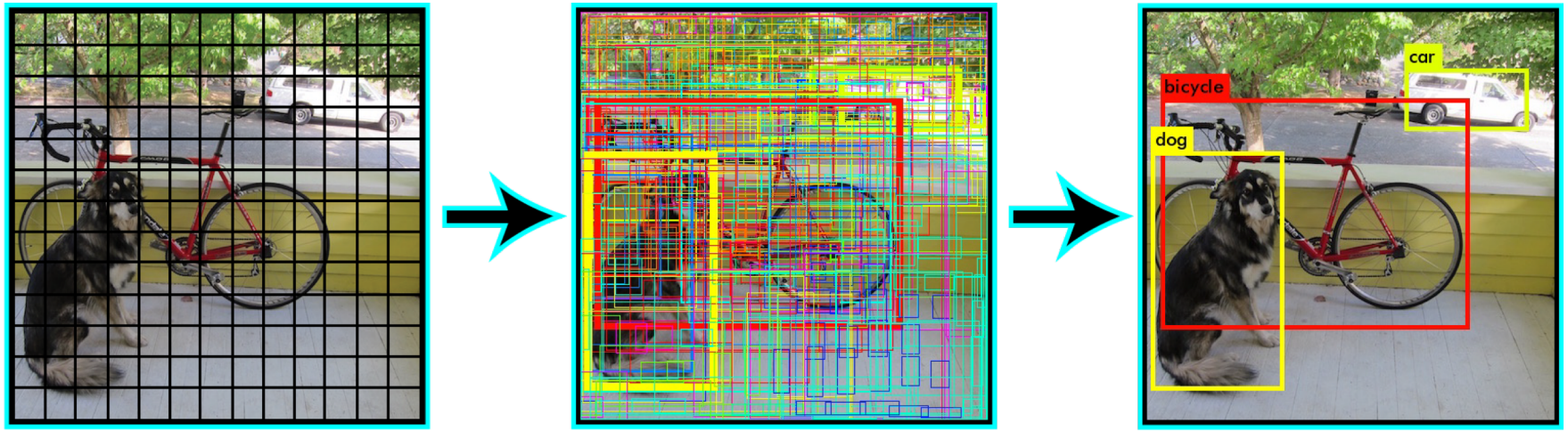
Internet of Eyes



Low latency & high reliability applications: Using deep learning at the network edge

Video

You Only Look Once (YOLO)



- Whole image goes through the network once (no sliding window, no region proposals)
- Image divided into grids -> network predicts category for each grid
- Predictions informed by global context
- More than 1000x faster than R-CNN and 100x faster than Faster R-CNN

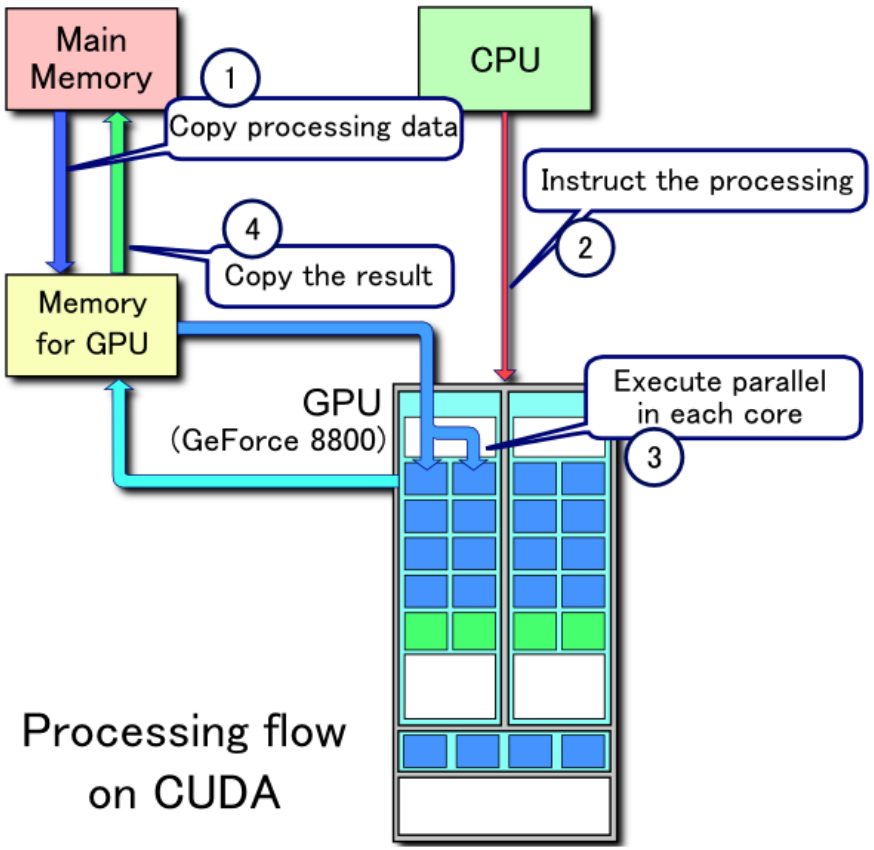
Why Machine Learning Now?

Science	Neural Networks has been researched for a long time => Requires compute and data
Data	Big data systems producing data for analytics => Requires global storage places, use cases
HW	GPU enables fast computation for Neural Networks

Breakthrough: GPU to make it faster

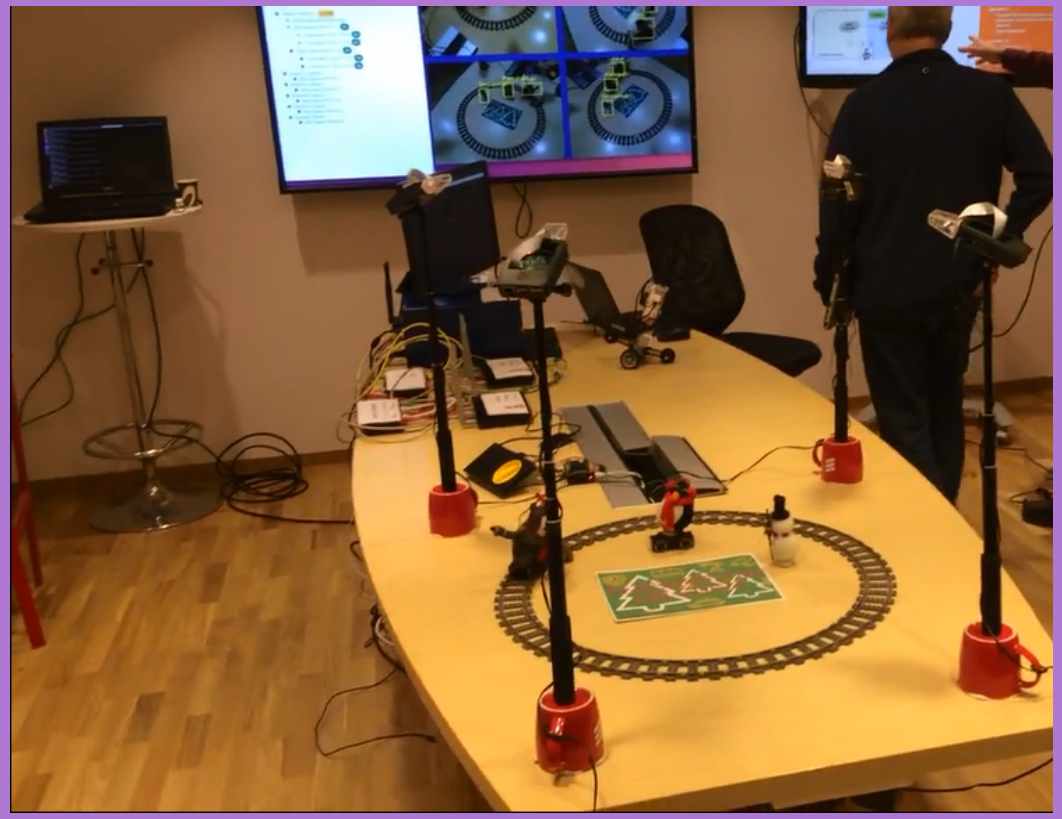
- Training a Deep Neural Network requires lots of parallel computations
- Nvidia provides a framework that makes it possible to use GPUs for Deep Learning
- Only for Nvidia GPUs
- Can make computation many times, sometimes even orders of magnitude faster (with expensive GPUs)
- Provides support for any machine learning task

Device	Speed of training, examples/sec
2 x AMD Opteron 6168	440
i7-7500U	415
GeForce 940MX	1190
GeForce 1070	6500

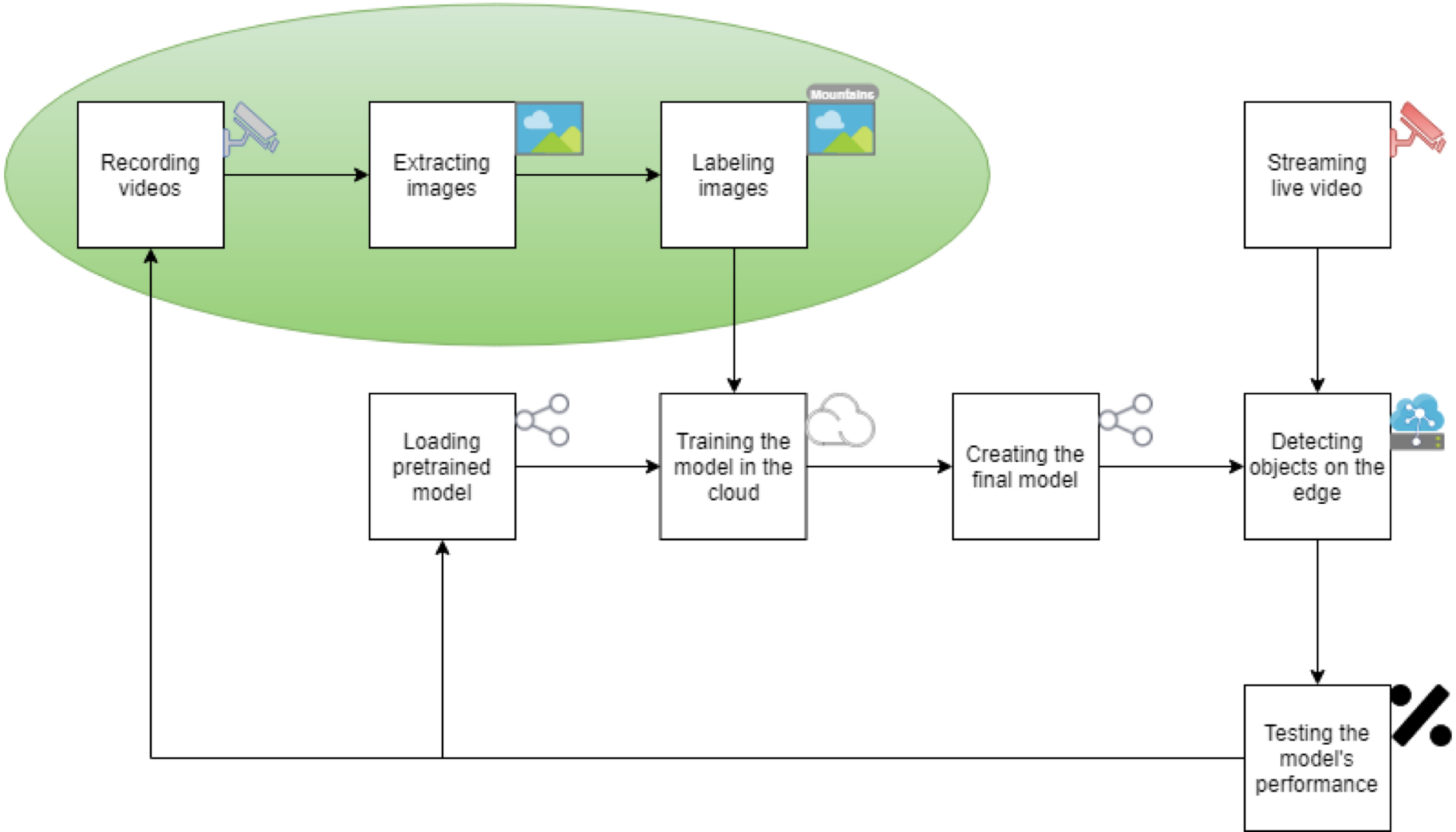


Internet of Eyes use case

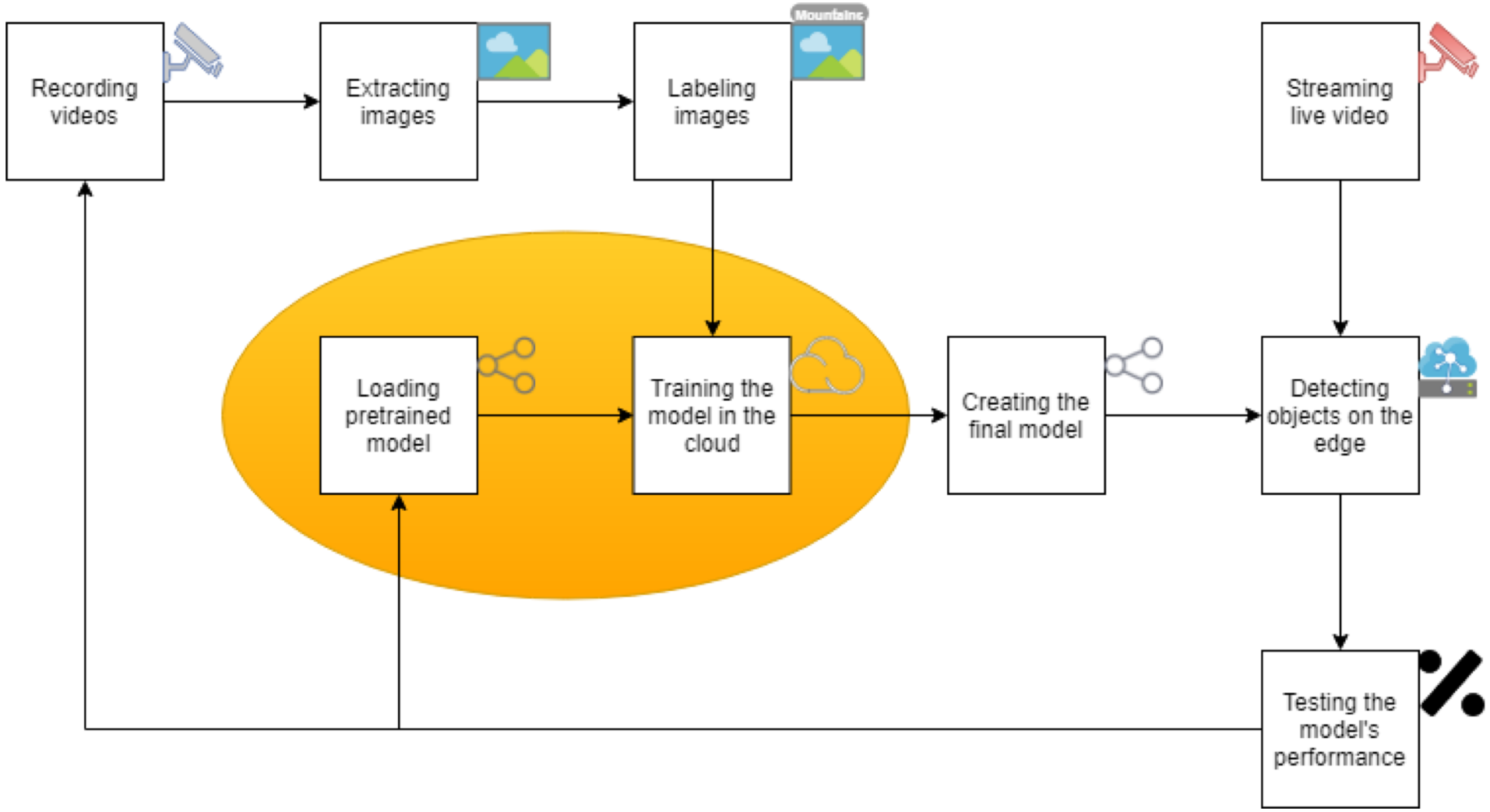
Outline of the process



Getting and labeling data

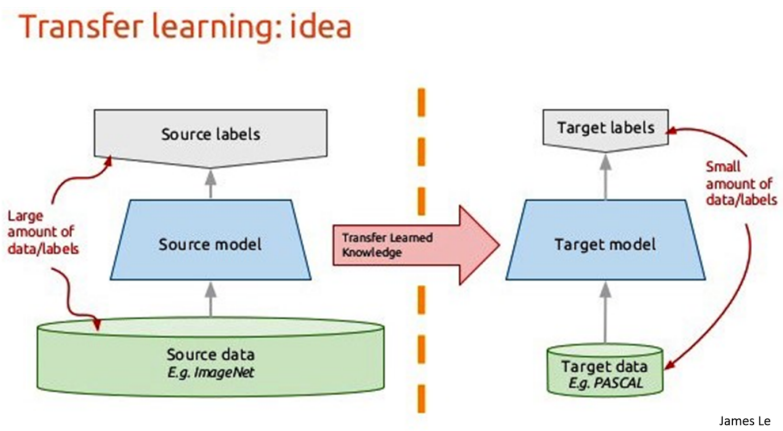


Training the network (in the central cloud)



Training the network (in the central cloud)
Transfer learning

- Normally we randomly initialize a neural network's weights (free parameters), then iteratively change them during training
- In transfer learning, we use the weights of a network that was already trained on some dataset



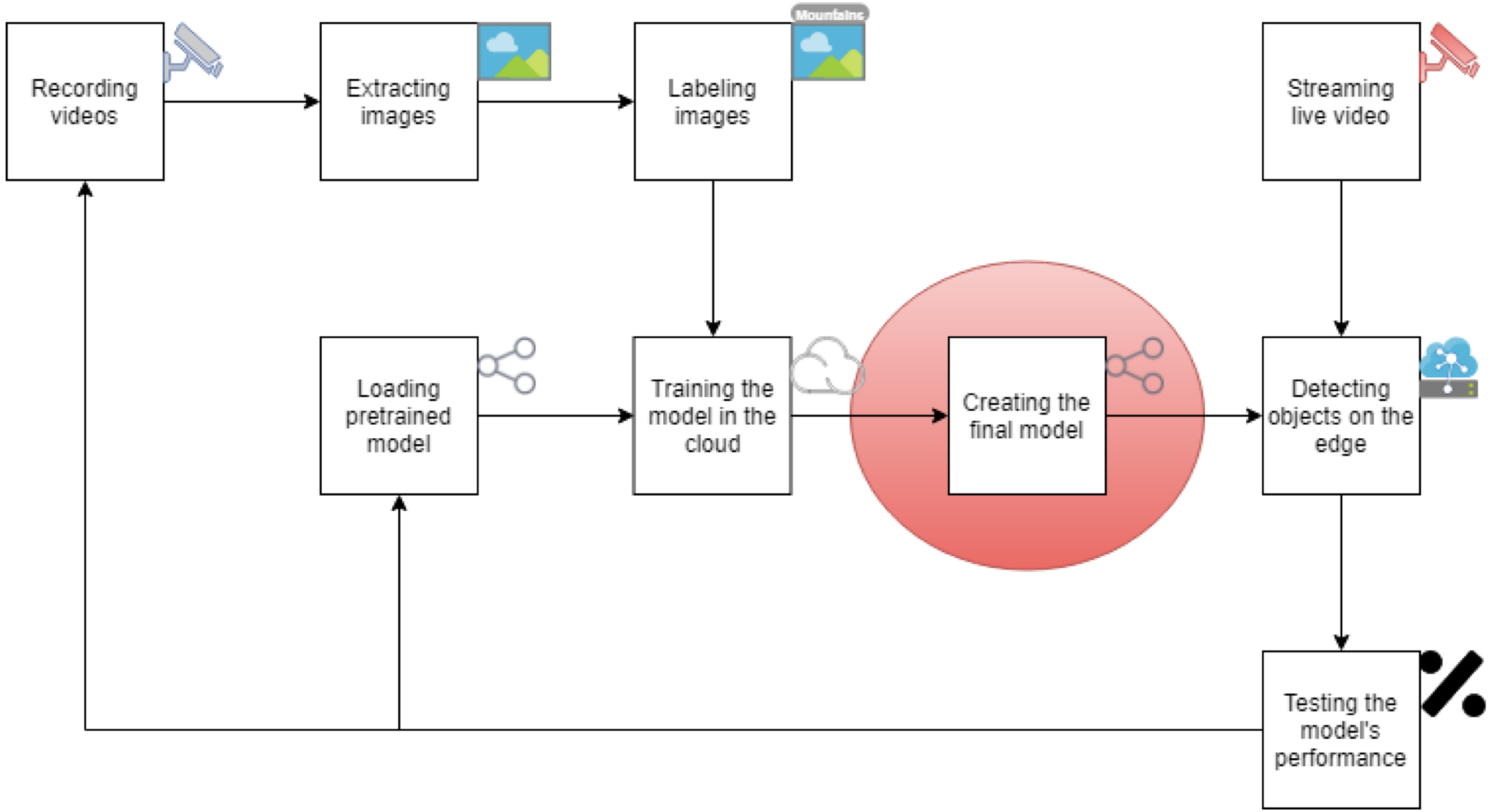
- This reduces the training time, because only few layers of the network are tuned
- The network we used was trained on the COCO dataset

- Training of a network requires extensive computational capacity therefore it is assumed to be performed in cloud datacenters

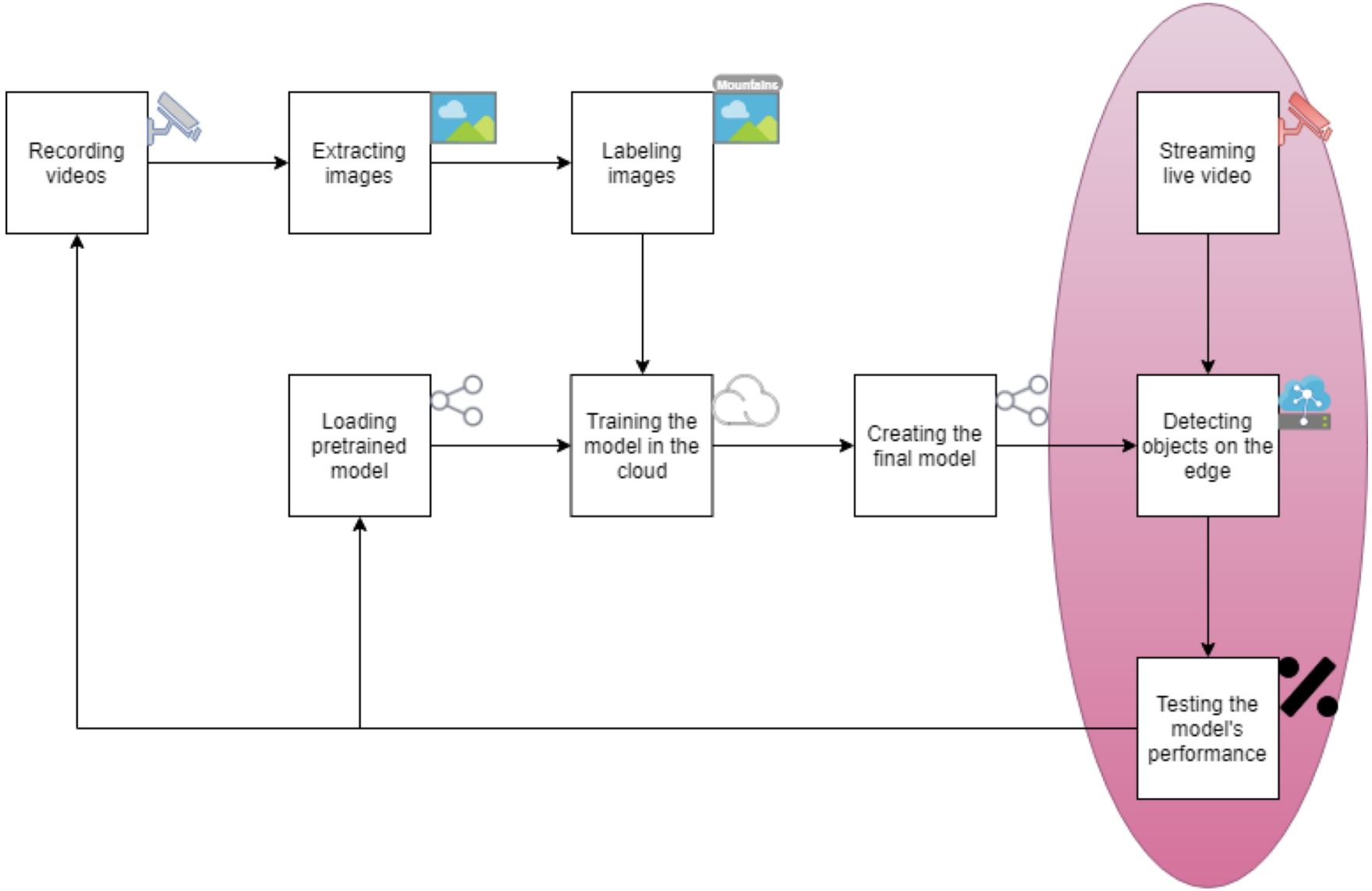
Why Machine Learning Now?

Science	Neural Networks has been researched for a long time => Requires compute and data
Data	Big data systems producing data for analytics => Requires global storage places, use cases
HW	GPU enables fast computation for Neural Networks => Requires cloud compute

Creating the model to be used for object detection



Object detection (on the Edge Cloud)



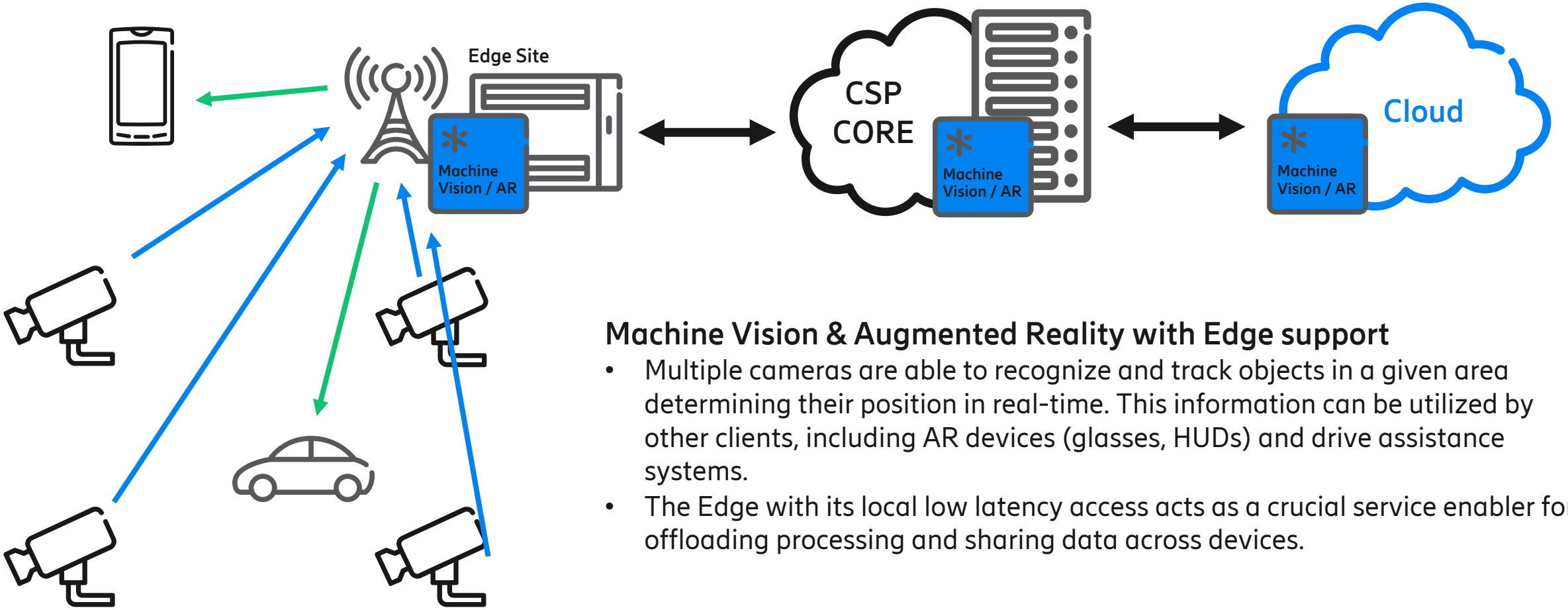
PoC conclusions

- Image recognition on trained network:
  - Still require heavy computations
  - Should be available close to the user
  - To be performed on the network edge

HW

GPU enables fast computation for Neural Networks
=> Requires cloud compute and everywhere availability

Reference Edge Use Case



Machine Vision & Augmented Reality with Edge support

- Multiple cameras are able to recognize and track objects in a given area determining their position in real-time. This information can be utilized by other clients, including AR devices (glasses, HUDs) and drive assistance systems.
- The Edge with its local low latency access acts as a crucial service enabler for offloading processing and sharing data across devices.

PoC conclusions

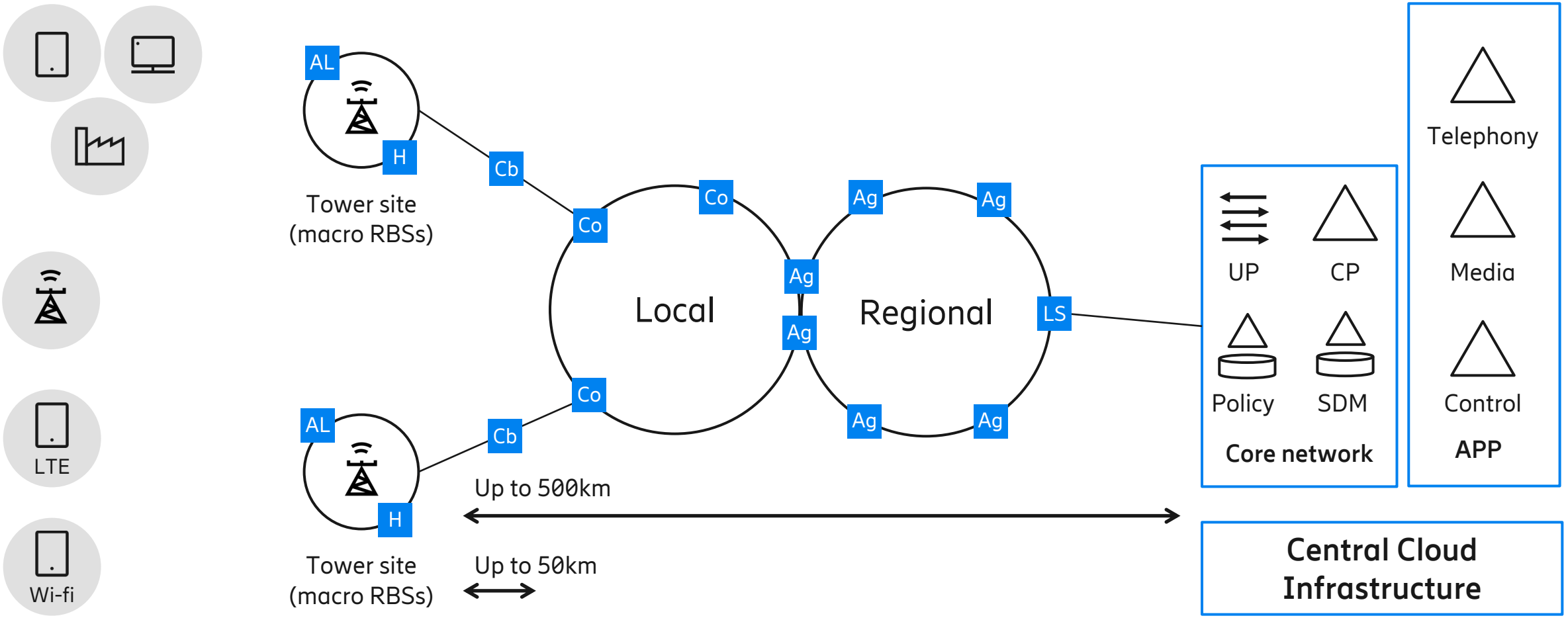
- Image recognition on trained network:
  - Still require heavy computations
  - Should be available close to the user
  - To be performed on the network edge

HW	GPU enables fast computation for Neural Networks => Requires cloud compute and everywhere availability
Network	Provides low latency communication, connectivity

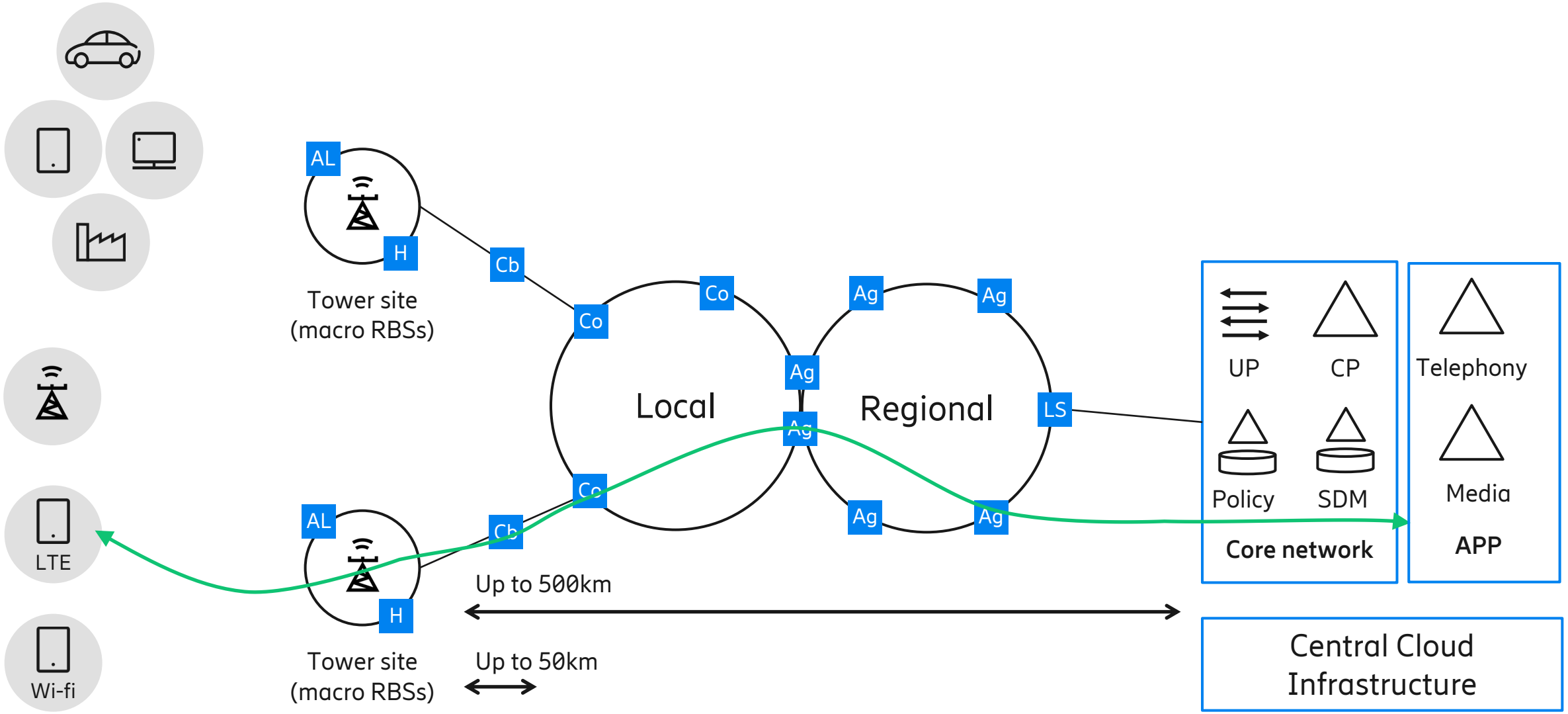
How such use case is implemented in the 5G network?

Any guess on what is this?

Network topology

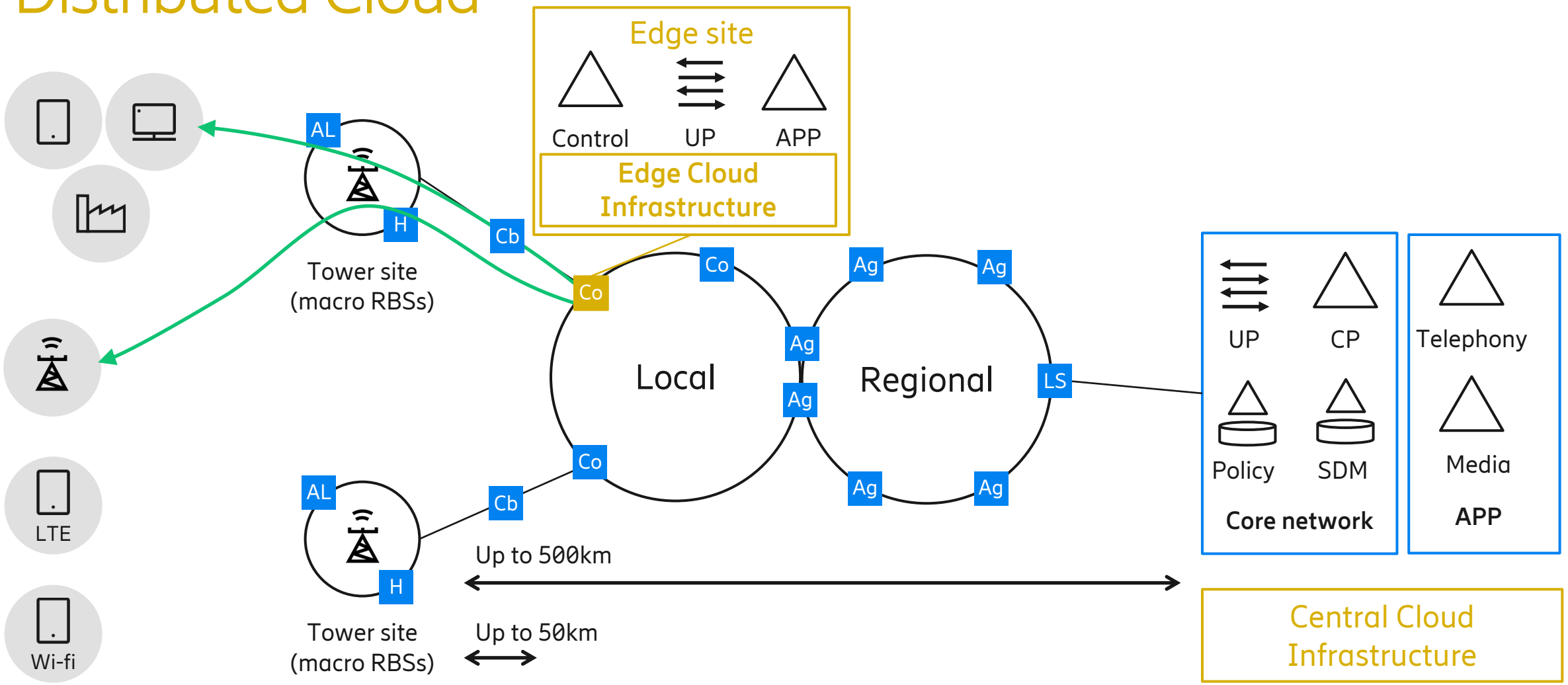


Central Cloud Topology Model

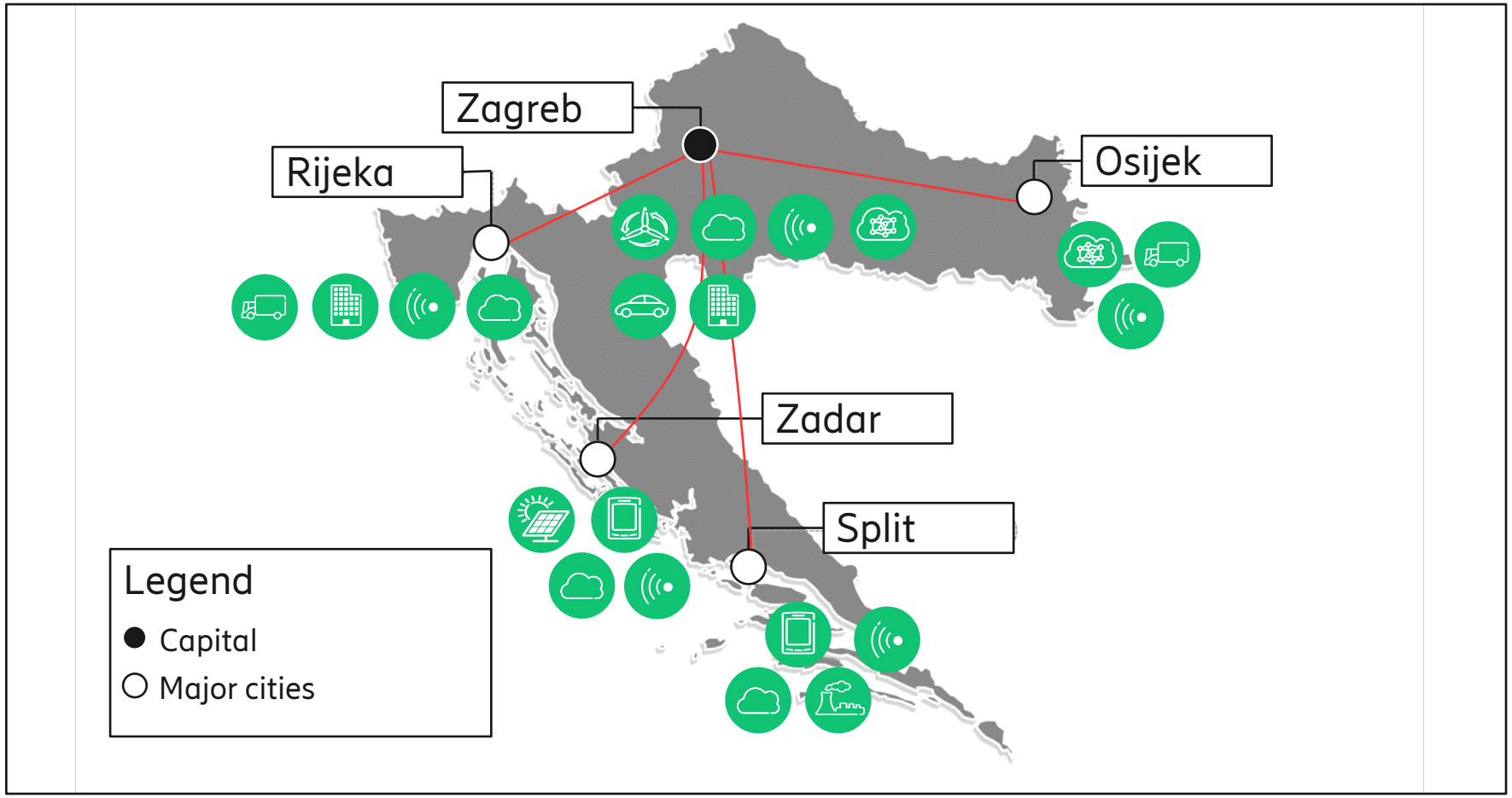


Distributed Infrastructure

Distributed Cloud

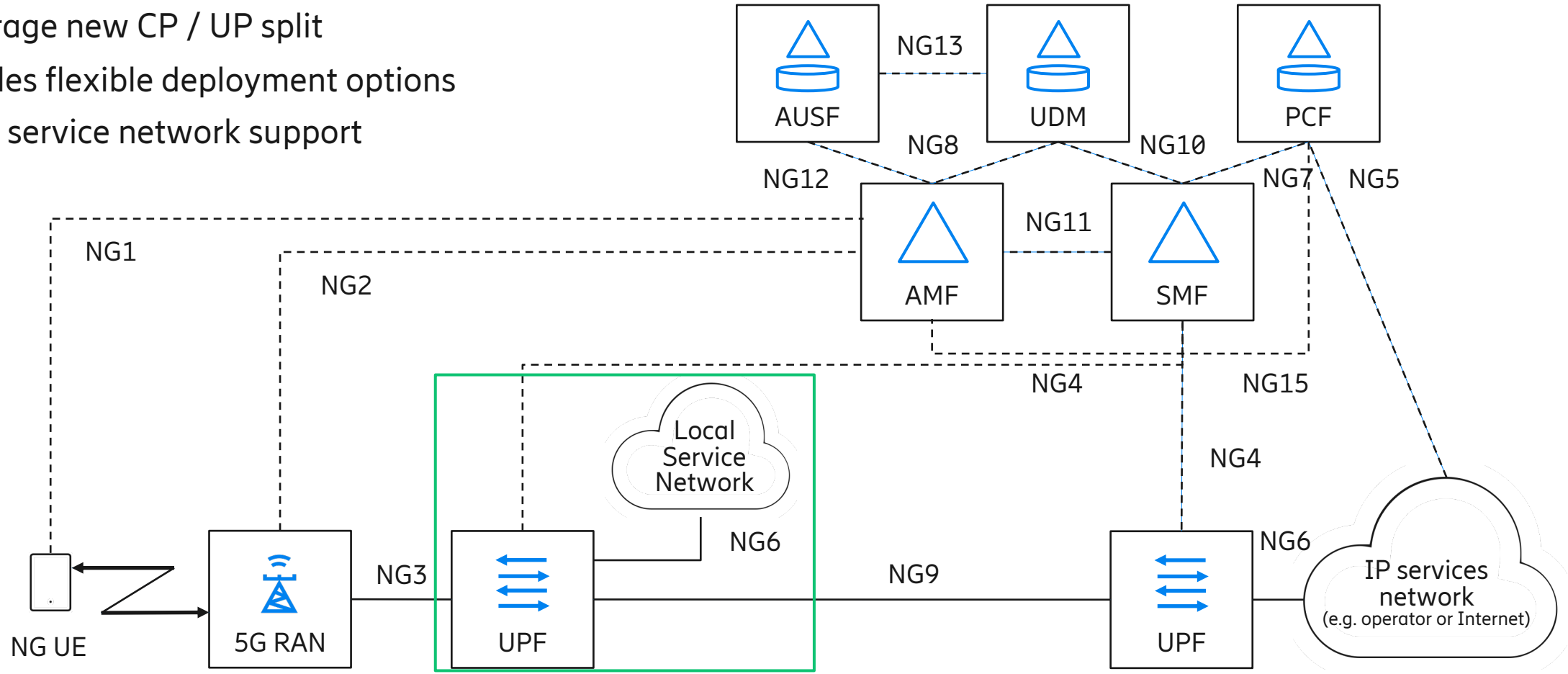


Geographical distribution



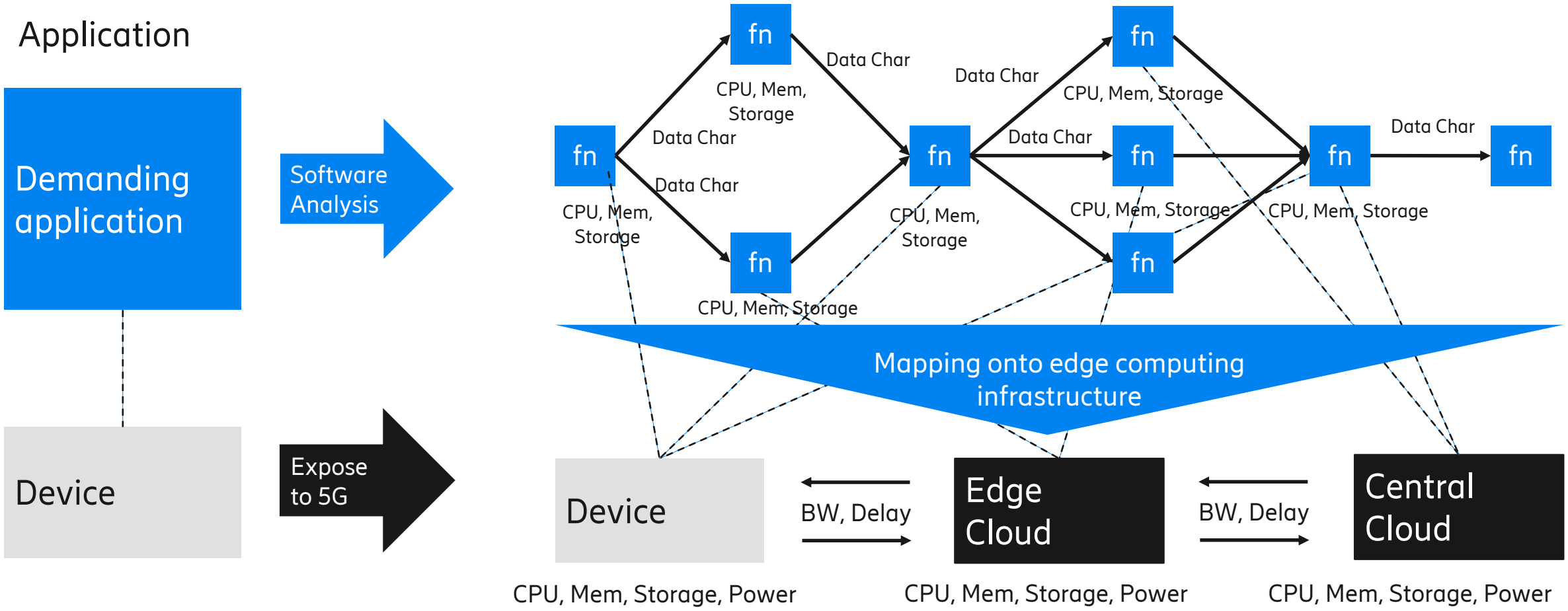
5G Core architecture Overview
Distributed User Plane Function

— Leverage new CP / UP split
— Enables flexible deployment options
— Local service network support

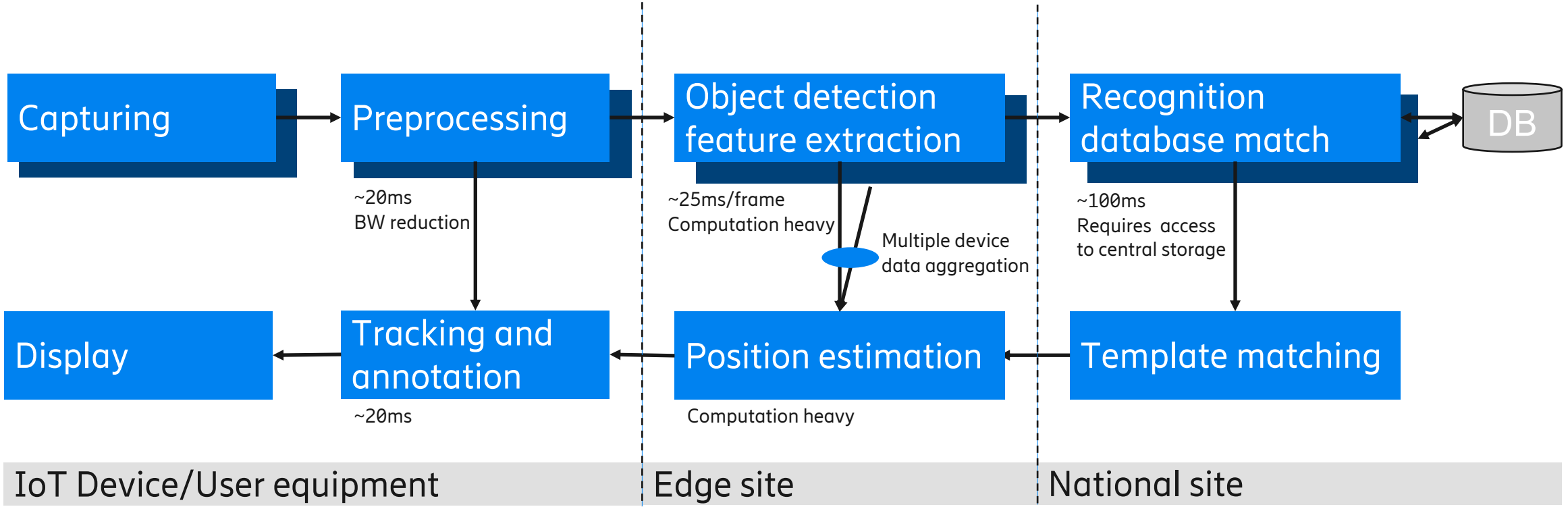


Demanding applications require edge computing

Mapping application components between Device, edge and central Cloud



AR/VR application with object recognition



Why Machine Learning Now?

Science	Neural Networks has been researched for a long time => Requires compute and data
Data	Big data systems producing data for analytics => Requires global storage places, use cases
HW	GPU enables fast computation for Neural Networks => Requires cloud compute and ubiquitous availability
Network	Provides low latency communication, connectivity => Enabling edge computing as a Distributed Cloud