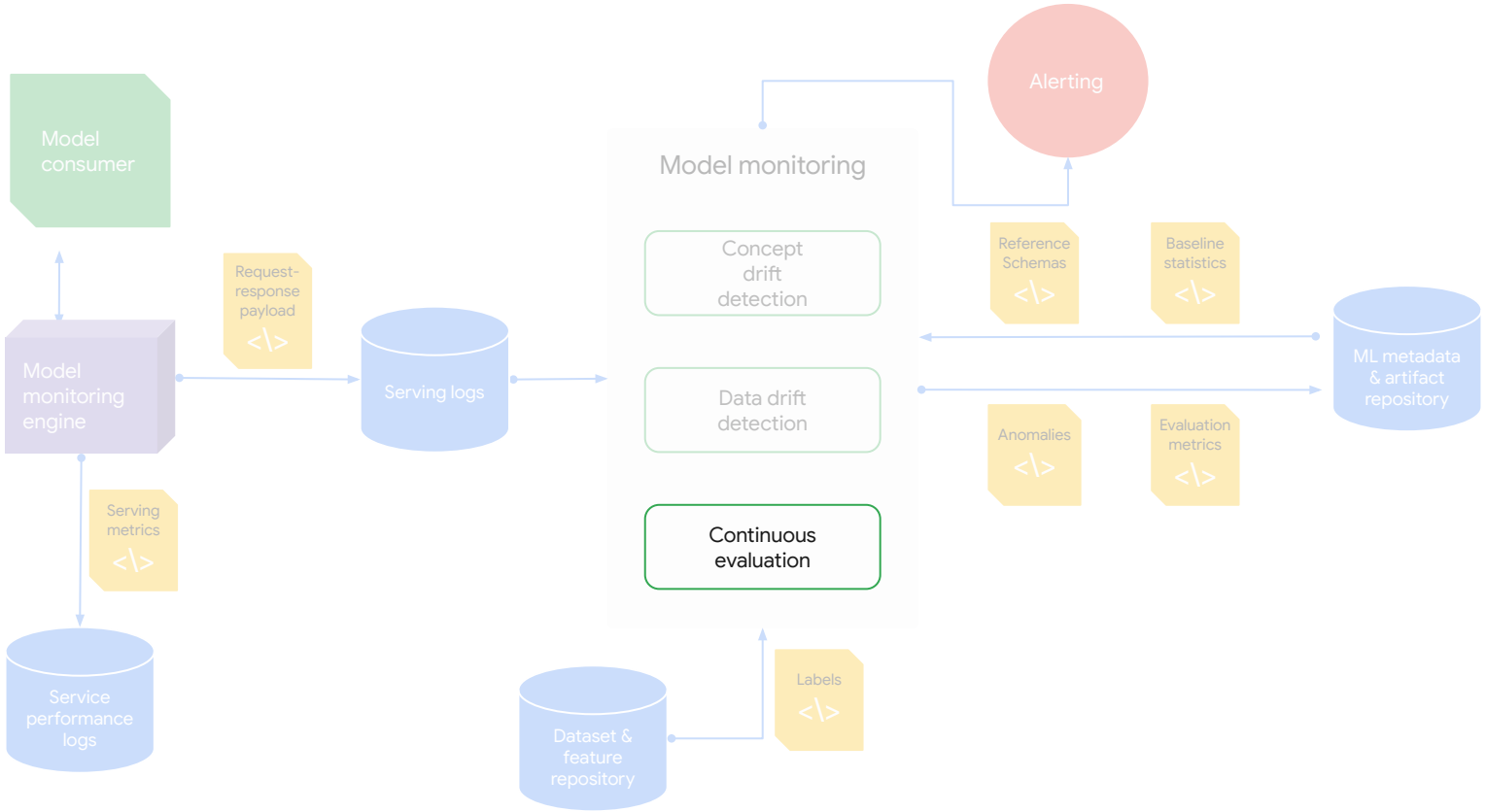
Continuous Evaluation Challenges for TinyML

MLOps: Continuous Monitoring



The MLOps **Personas**

ML Engineer

ML Researcher

Data Scientist

Data Engineer

Software Engineer

DevOps

Business Analyst

Continuous Monitoring for **TinyML**

- Monitoring may **not always** be a **feasible** option
  - Low power communication protocol
  - Device isn't wifi-enabled

- Monitoring opens up **security and privacy risks**

Continuous Monitoring for TinyML

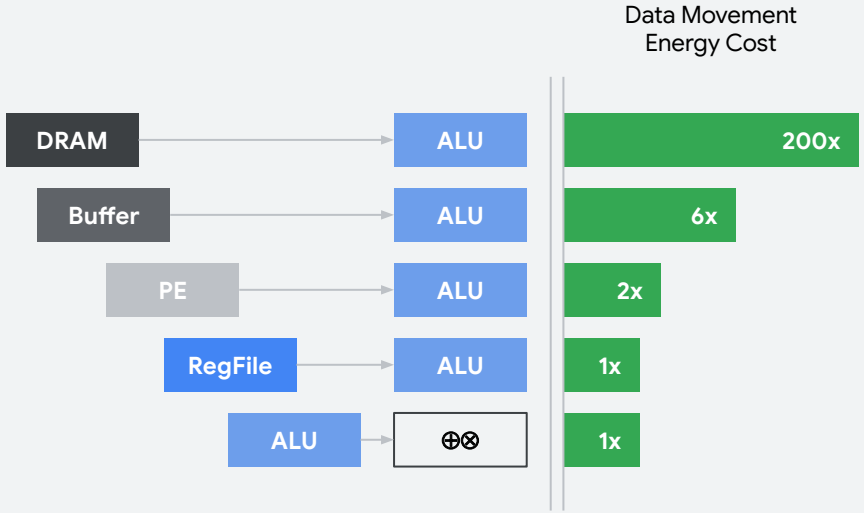
- Monitoring may **not always** be a **feasible** option
  - Low power communication protocol
  - Device isn't wifi-enabled

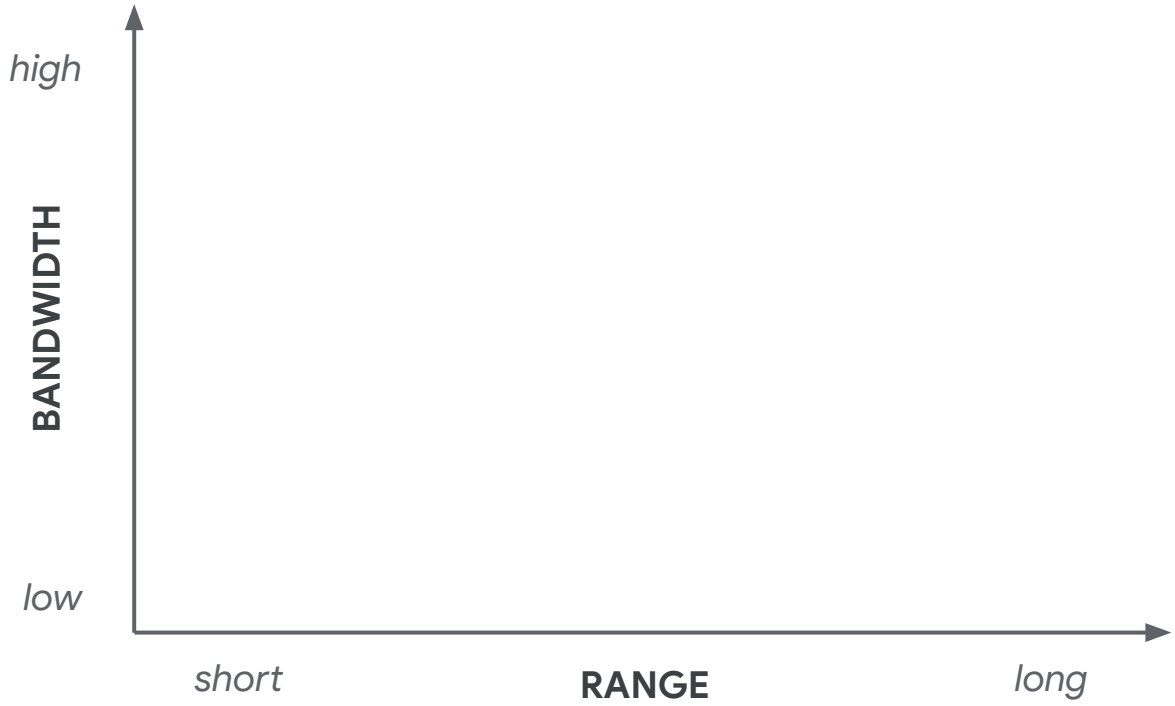
- Monitoring opens up **security and privacy risks**

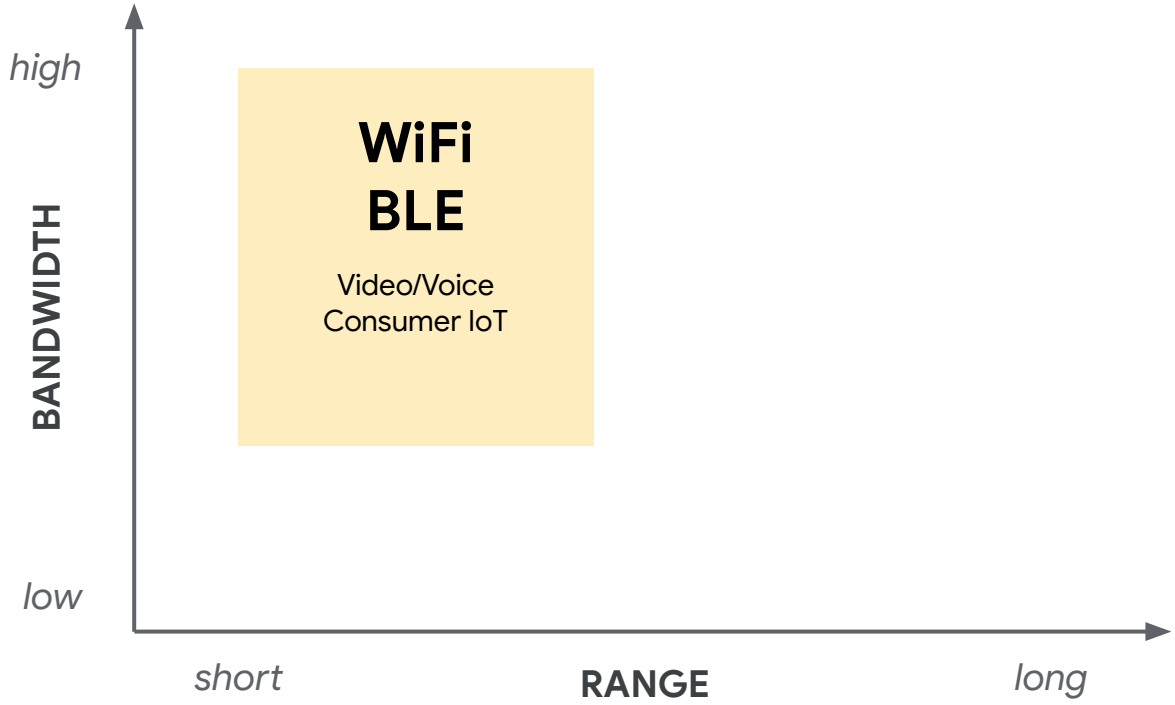
- How can we enable **Continuous Monitoring** to enable **Continuous Training** without moving the data off the endpoint tiny ML device?

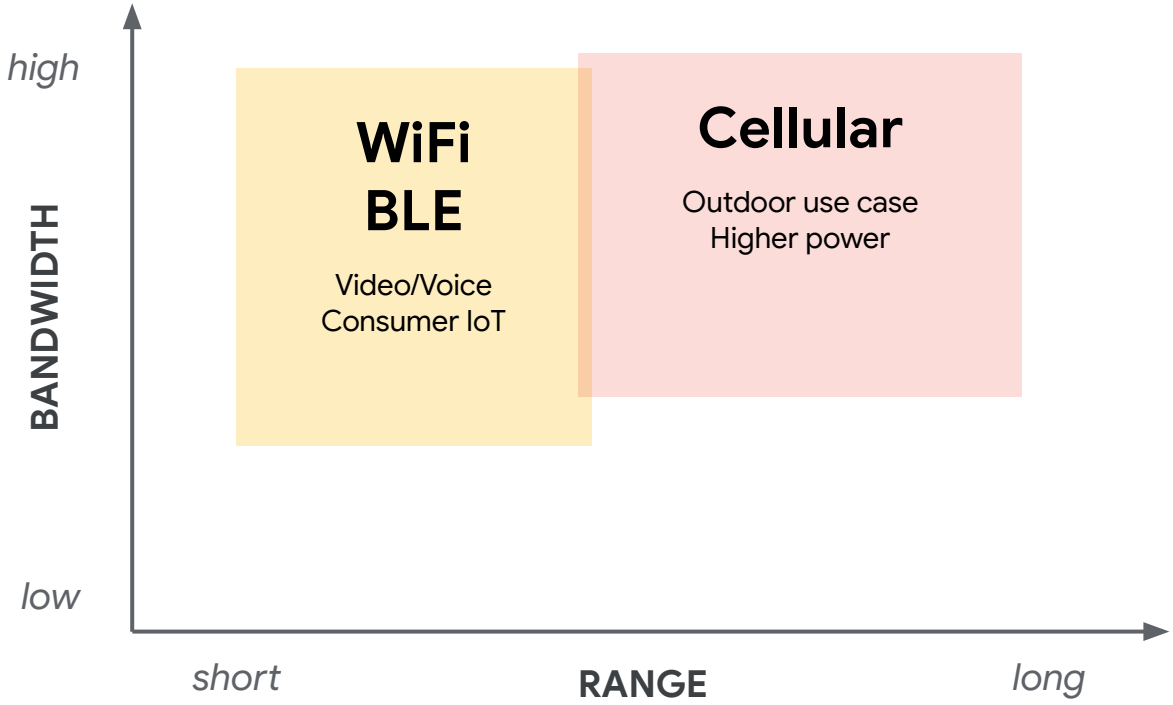
Computation vs. Communication

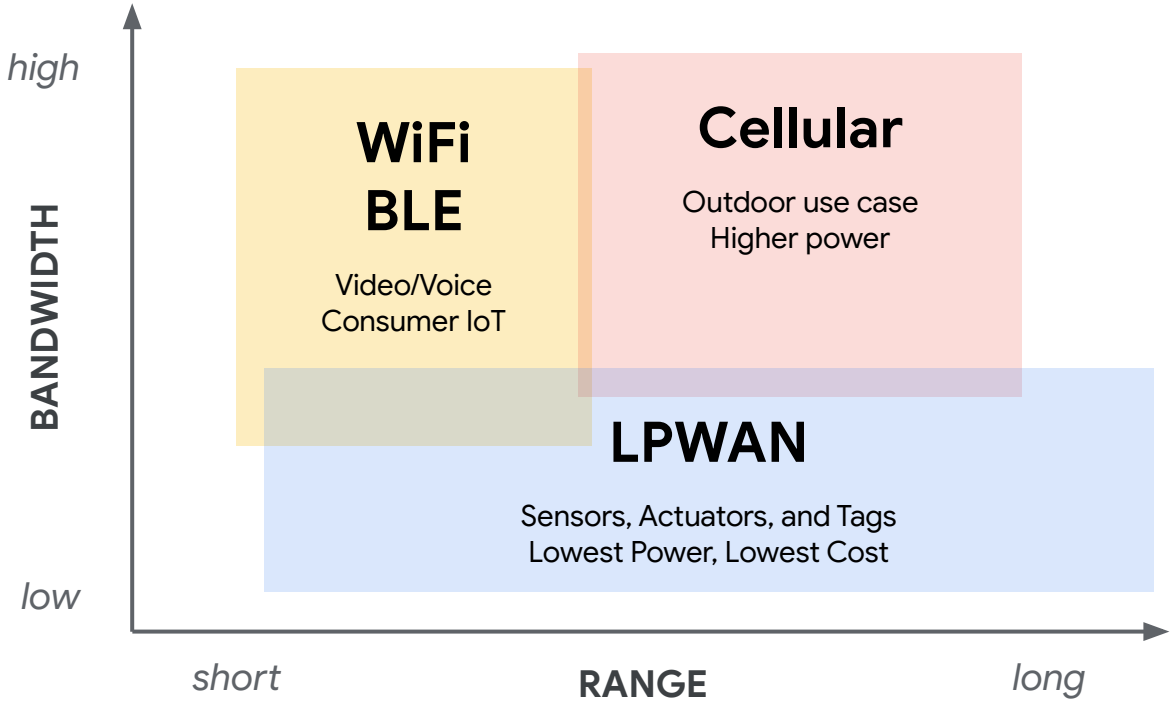
- Data movement is extremely important to keep in mind for efficient system engineering
- Data movement is more costly than computation itself











	LAN Bluetooth, ZigBee, WiFi	**Cellular** 2G, 3G, 4G	**LPWAN** LoRa, Sigfox, NB-IoT
Data Rate	~100kbps - 100mbps	~100kbps - 100mbps	10kbps
Range	Short	Long	Long Range (10 miles)
Battery Life	Varies	Medium	Long Battery Life (10 yrs)
Cost	Expensive	Very Expensive	Best Price
Use Cases	*Smart TV, WiFi Network, Bluetooth Speakers*	*Smart Grid, CCTV, Personal Communication (smartphones)*	*Monitoring, Metering, Temperature, Asset Tracking, Weather, Location*

Long Range

Deep indoor coverage (multi-floor buildings)

Star topology network design

Long Battery Life

Low-power optimized

Up to 10 year lifetime

Up to 10x versus Cellular M2M

High Capacity

Millions of messages per gateway

Multi-tenant interoperability

Public/Private network deployments

Low Cost

Minimal infrastructure

Low cost end-node

Open source software

Geolocation

Indoor/outdoor

Accurate without GPS

No battery life impact

Updates

Firmware updates over-the-air for apps and LoRaWAN stack

Roaming

Seamless handoffs from one network to another

Security

Embedded end-to-end AES-128 encryption

Unique ID

Application

Network

Example Use Cases

- Vending machines could alert distributors when a product requires **maintenance**

Example Use Cases

- Vending machines could alert distributors when a product requires **maintenance**
- Animal lovers can **track** pets or study migration patterns over longer distances

Example Use Cases

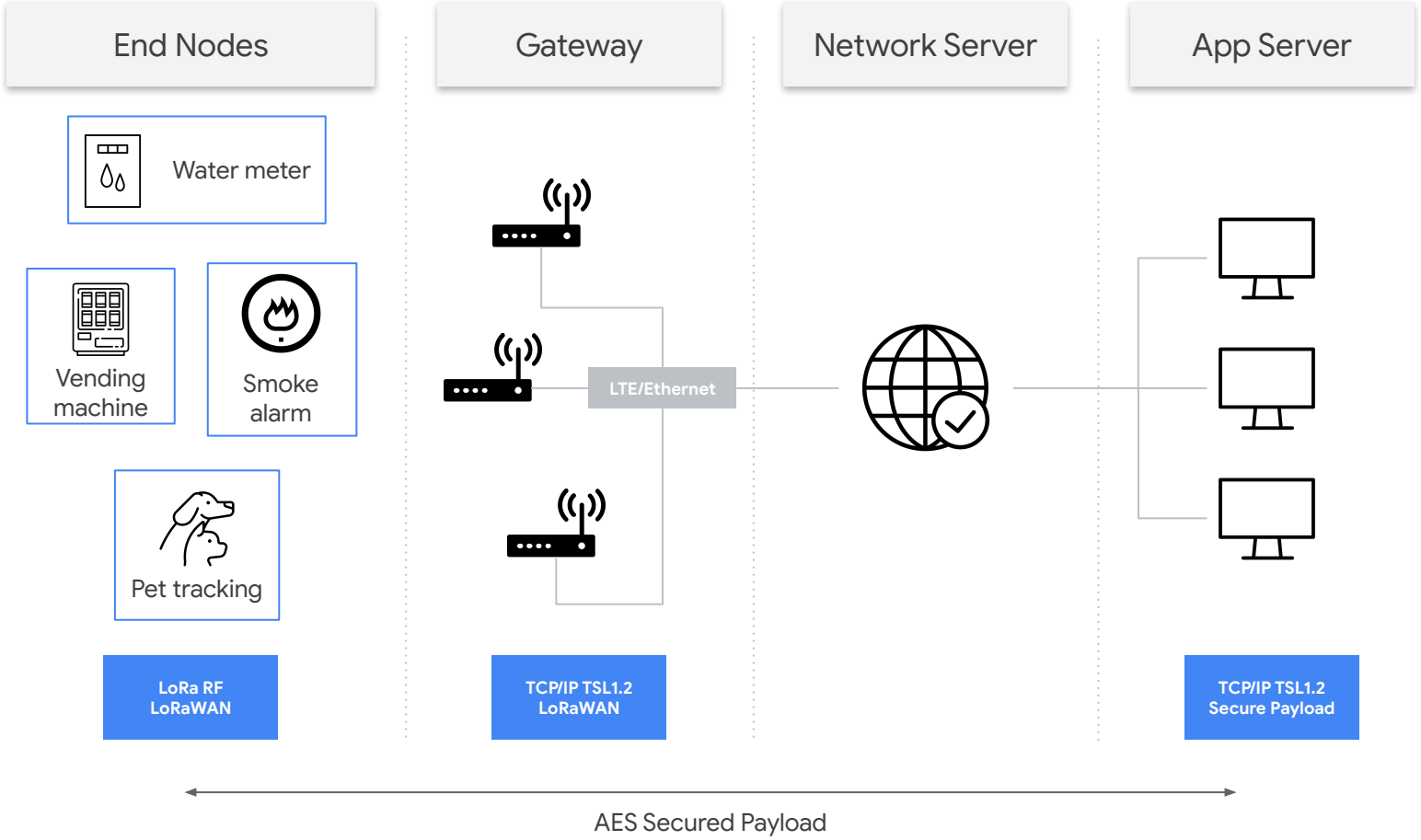
- Vending machines could alert distributors when a product requires **maintenance**
- Animal lovers can **track** pets or study migration patterns over longer distances
- Oil companies could **receive alerts** when home oil tanks are running low

Example Use Cases

- Vending machines could alert distributors when a product requires **maintenance**
- Animal lovers can **track** pets or study migration patterns over longer distances
- Oil companies could **receive alerts** when home oil tanks are running low
- Logistics providers could **track** cargo containers on trucks, ships and trains

Example Use Cases

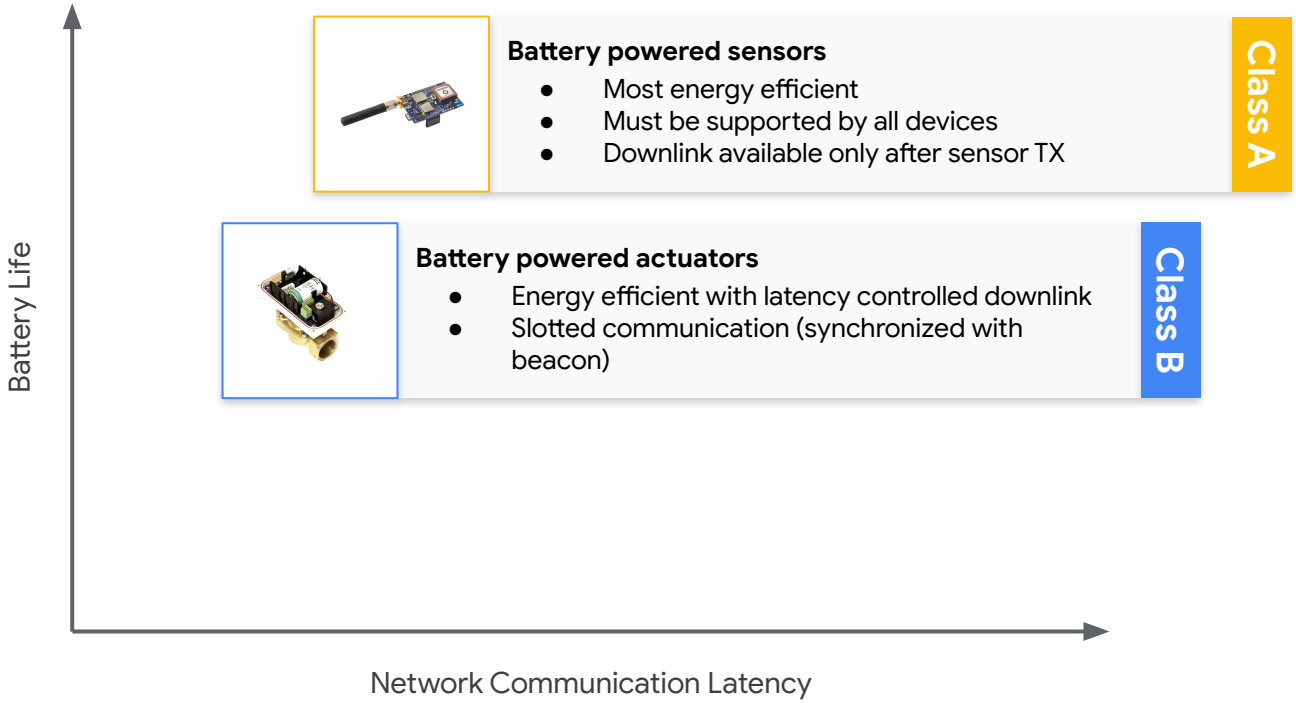
- Natural **Disaster Prevention**
- Smart **Agriculture Monitoring**
- Animal **Production Monitoring**
- Endangered Species **Protection**
- **Smart** Industry **Control**
- **Smart** Cities, Homes, Buildings & Offices
- Supply Chain **Logistics**, **Asset Tracking** & **Quality Management**
- **Smart Metering** Facilities (Water, Gas, Electricity)

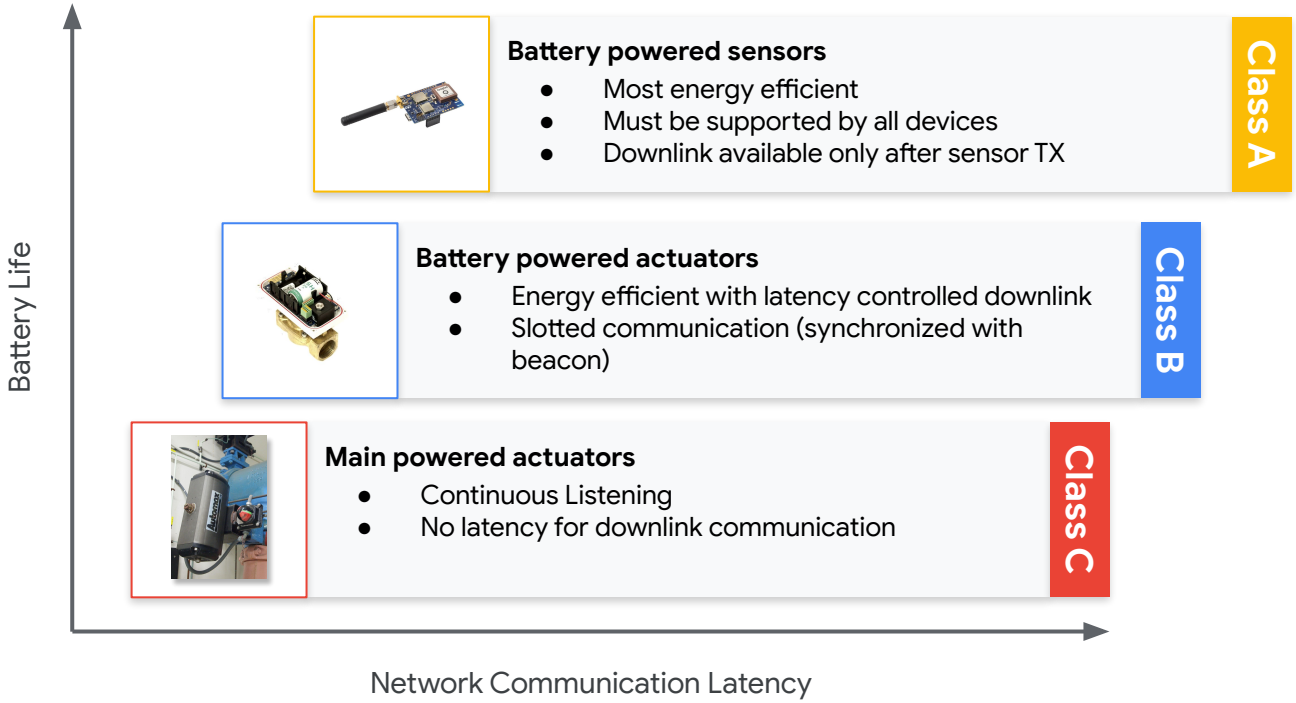


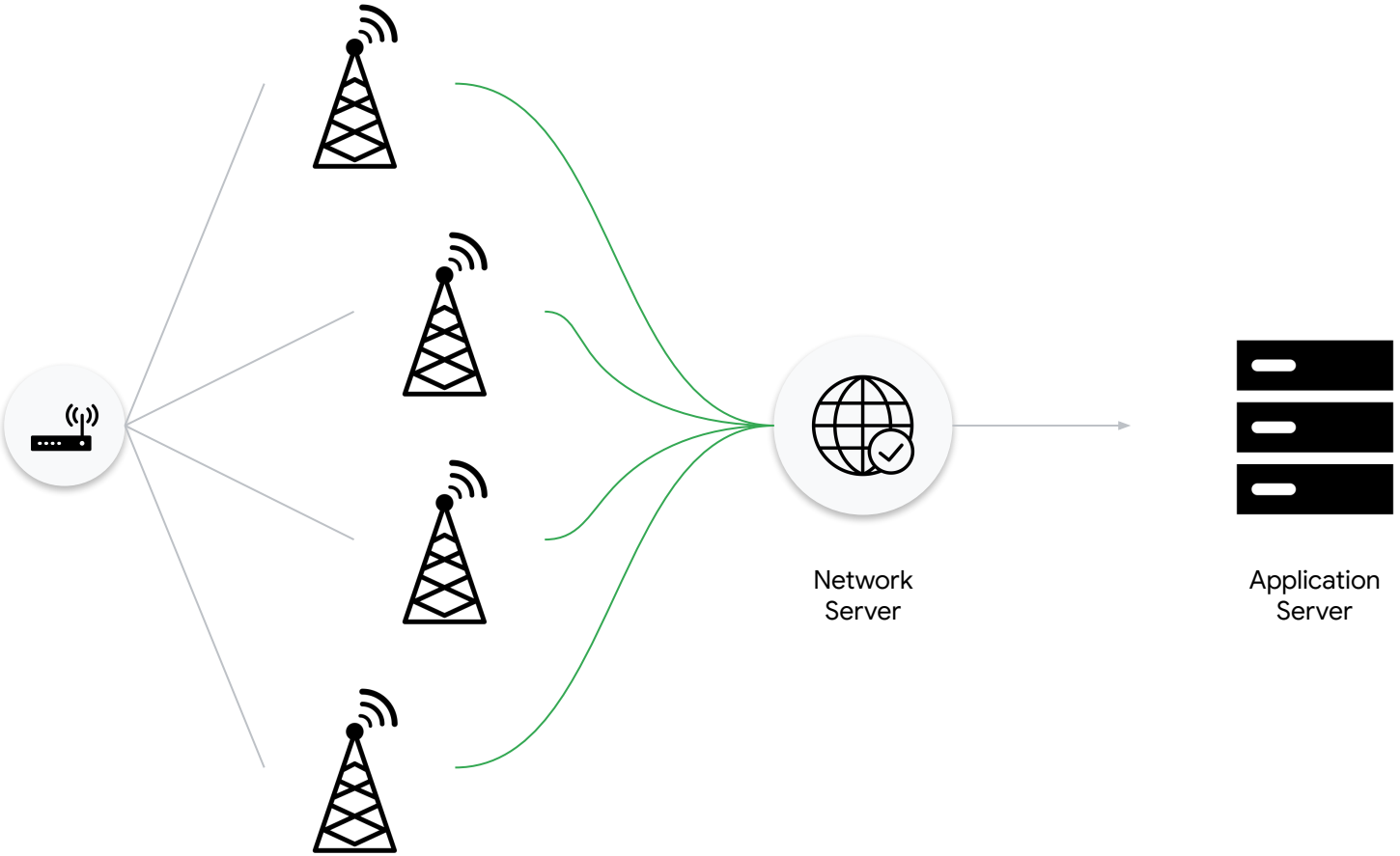
Battery powered sensors

- Most energy efficient
- Must be supported by all devices
- Downlink available only after sensor TX

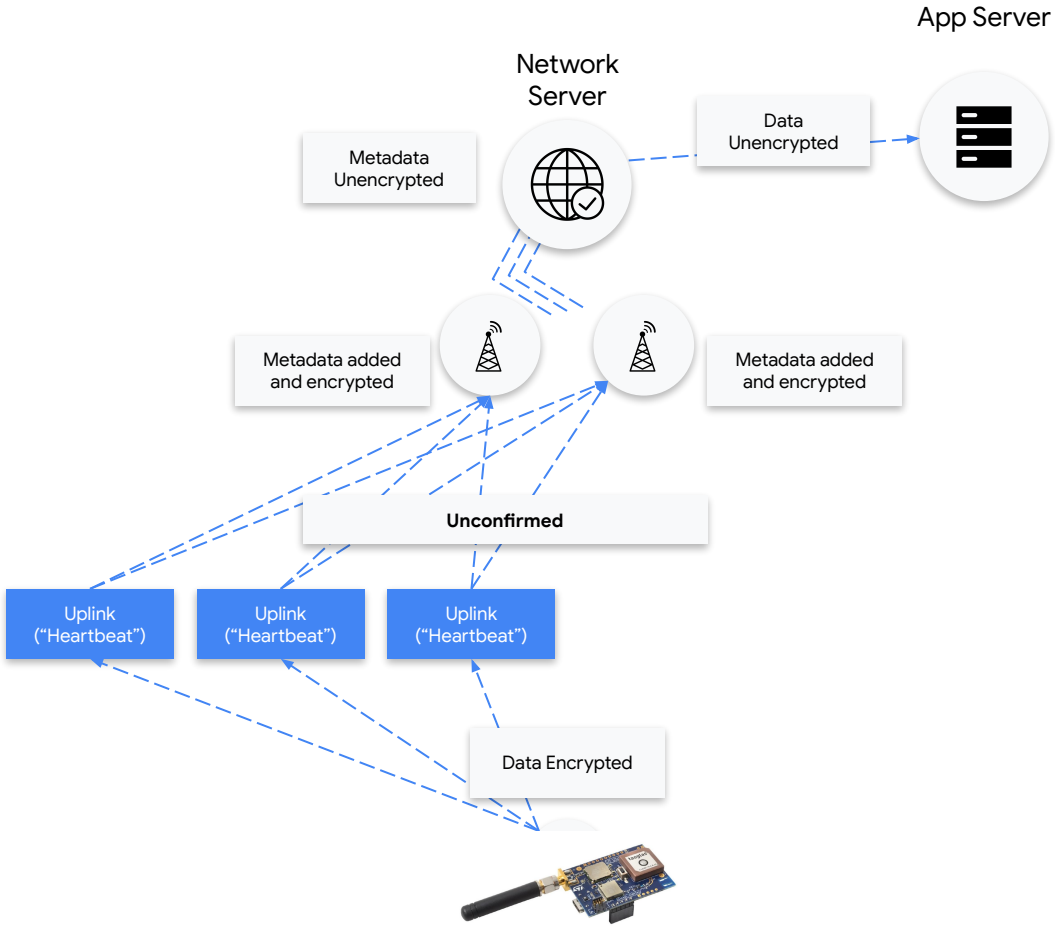
Class A



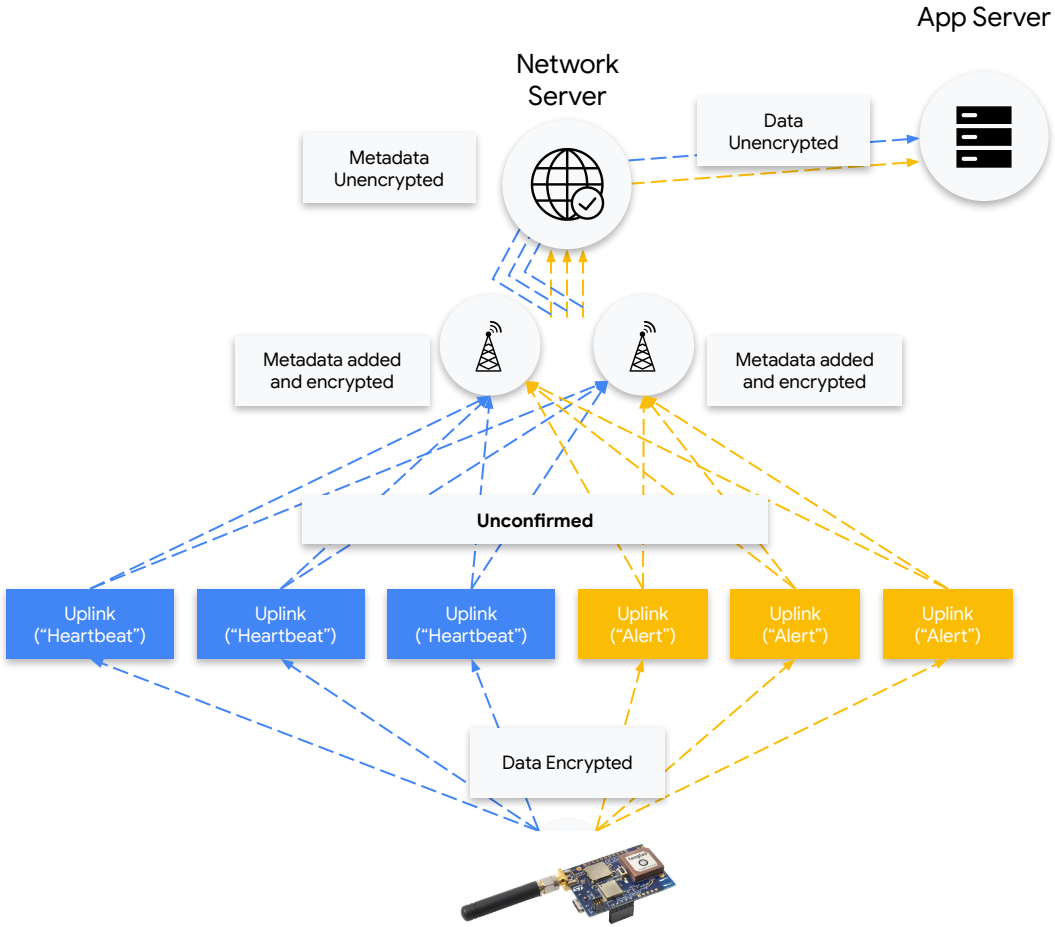




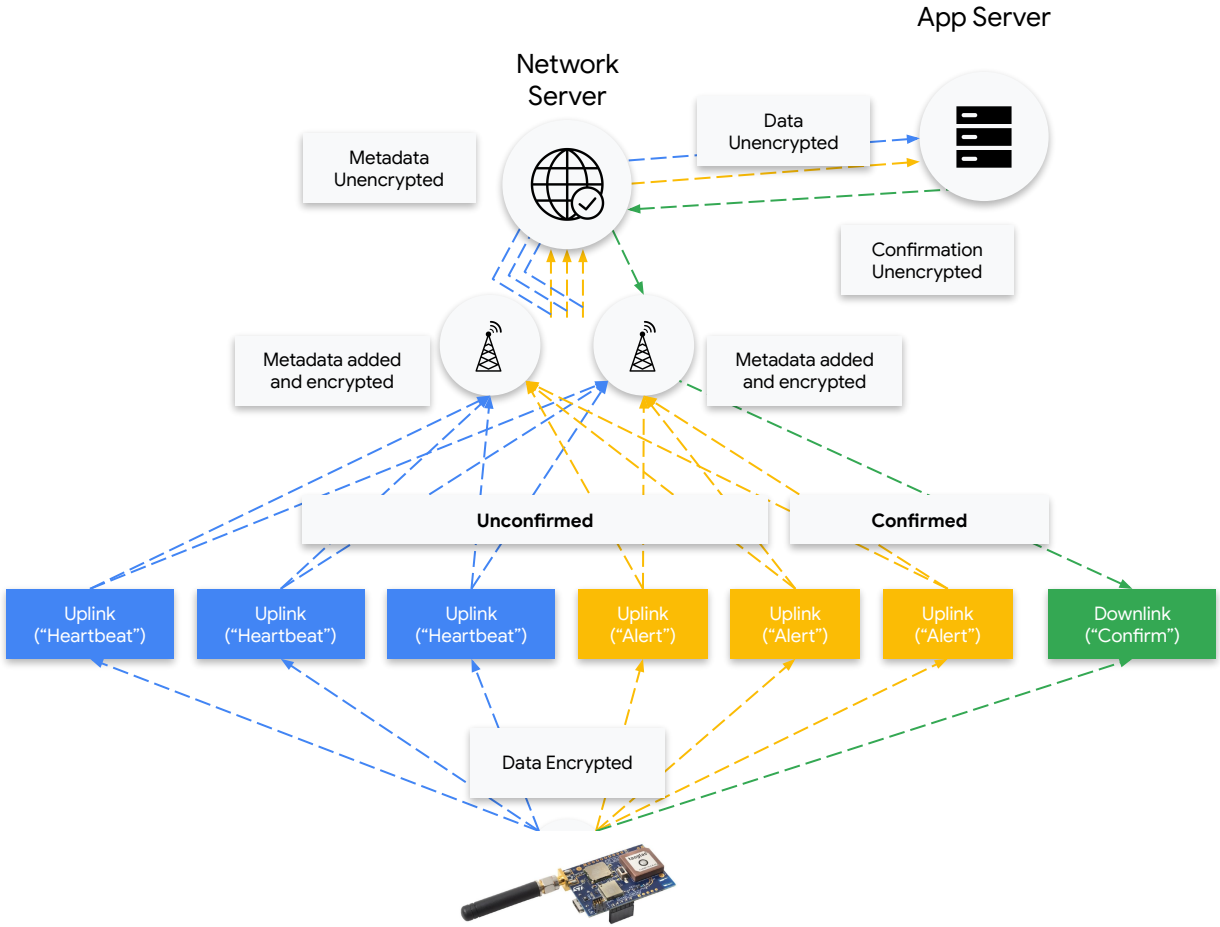
Example Use Case



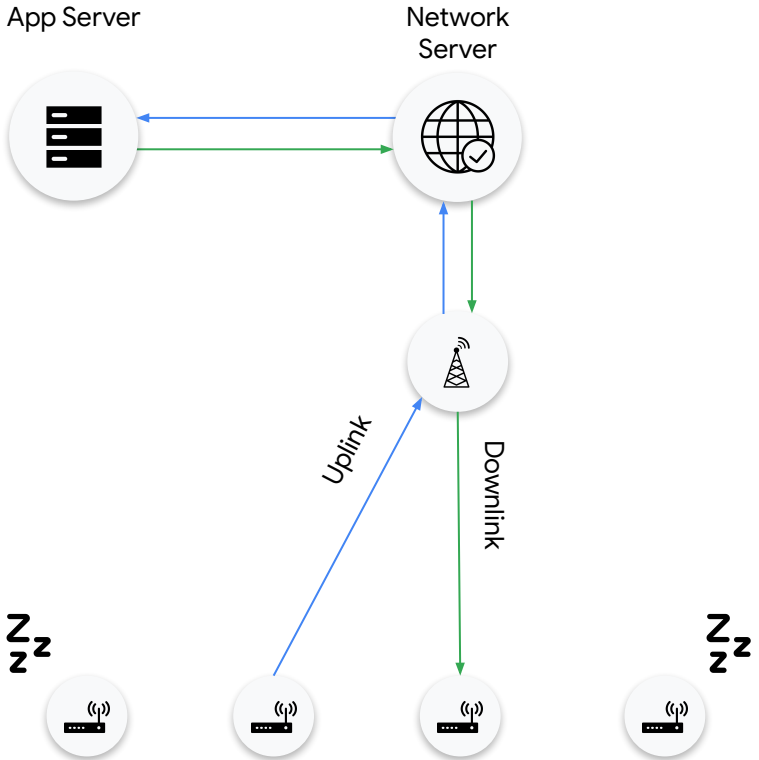
Example Use Case



Example Use Case

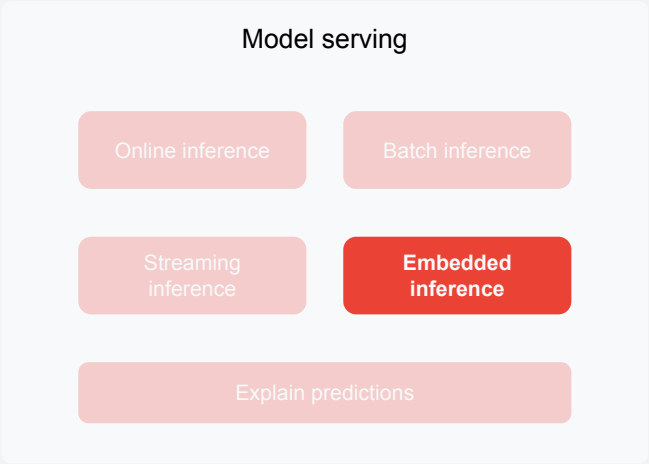


Example Use Case



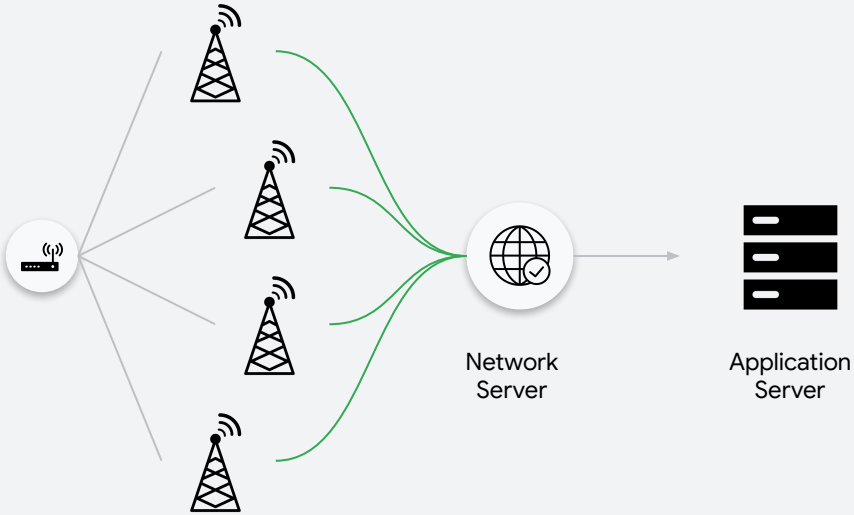
Embedded Inference

- computation v. communication **trade-off**
- different communication **protocols**

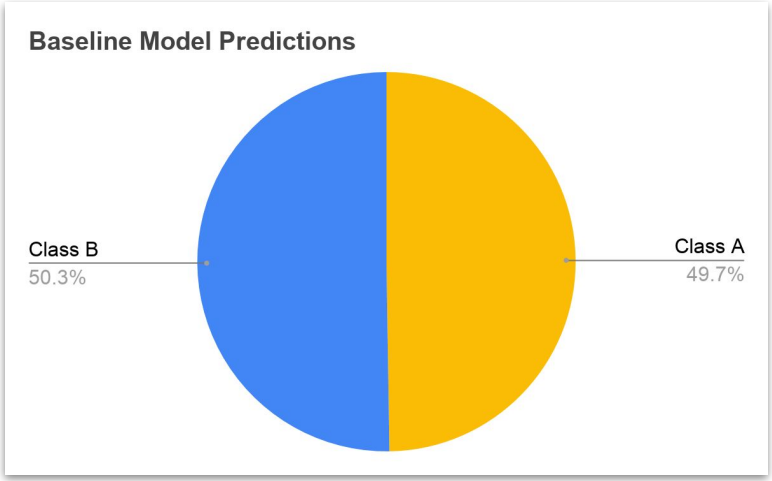


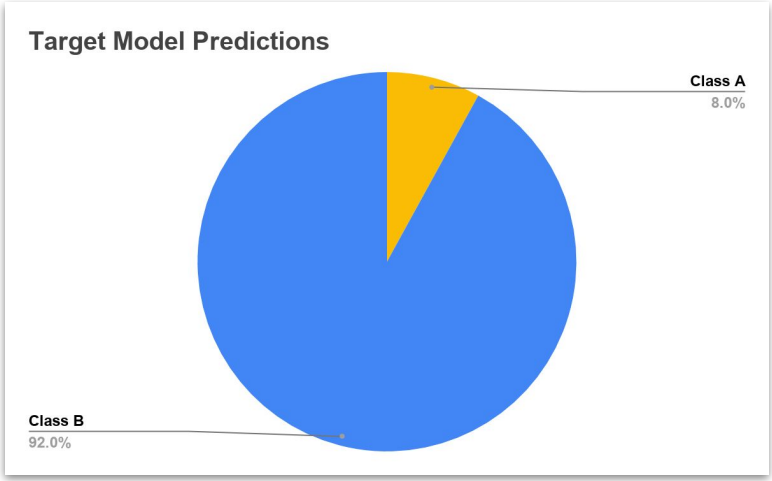
Network Advantages for Class A Devices

- **No complicated network** planning is required. Gateways can be added anywhere at any time
- **Accurate message delivery is robust**, since multiple gateways receive the same data packet during each uplink. This is called uplink spatial diversity
- There is **no need to plan for different frequencies for each gateway**, or to reallocate frequencies when the number of gateways change. All gateways are constantly listening to all frequencies of the network
- Mobile devices can **operate at ultra low power** thanks to the fact that any gateway can receive messages from any device at any time



Detecting Drift—Predictions



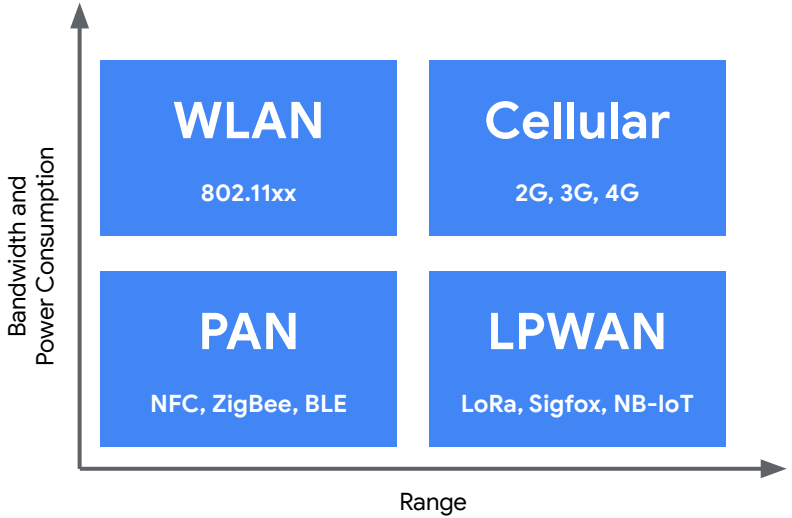


Detecting Drift—Data Distributions

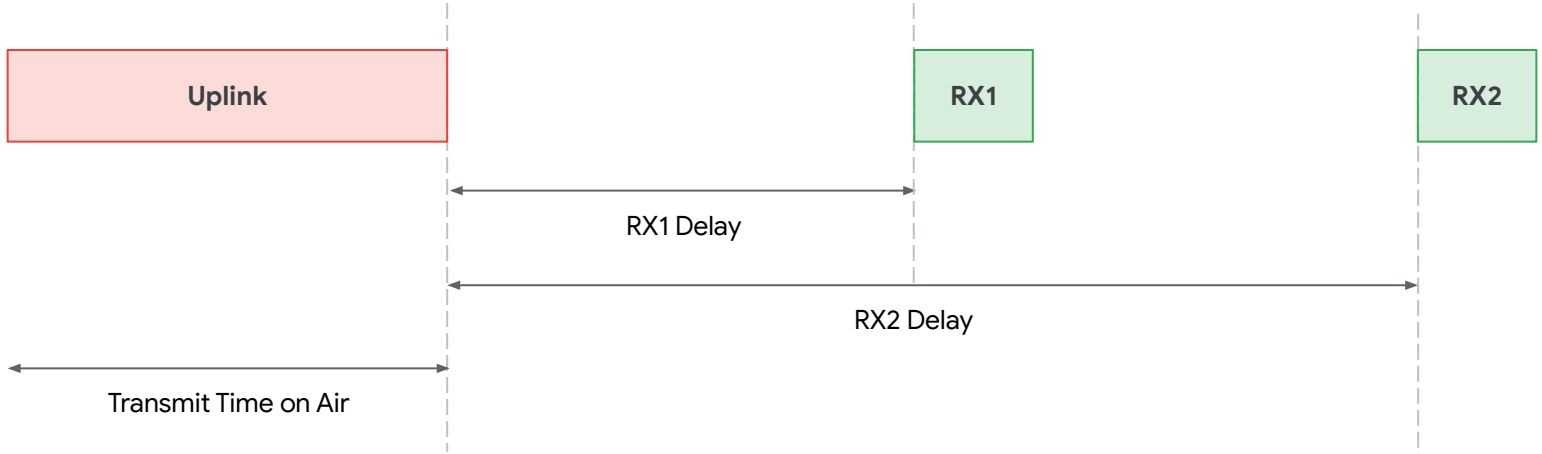
BASELINE	TARGET
60% male, income 50k, etc.	80% male, income 60k, etc.

Communication Protocol Features

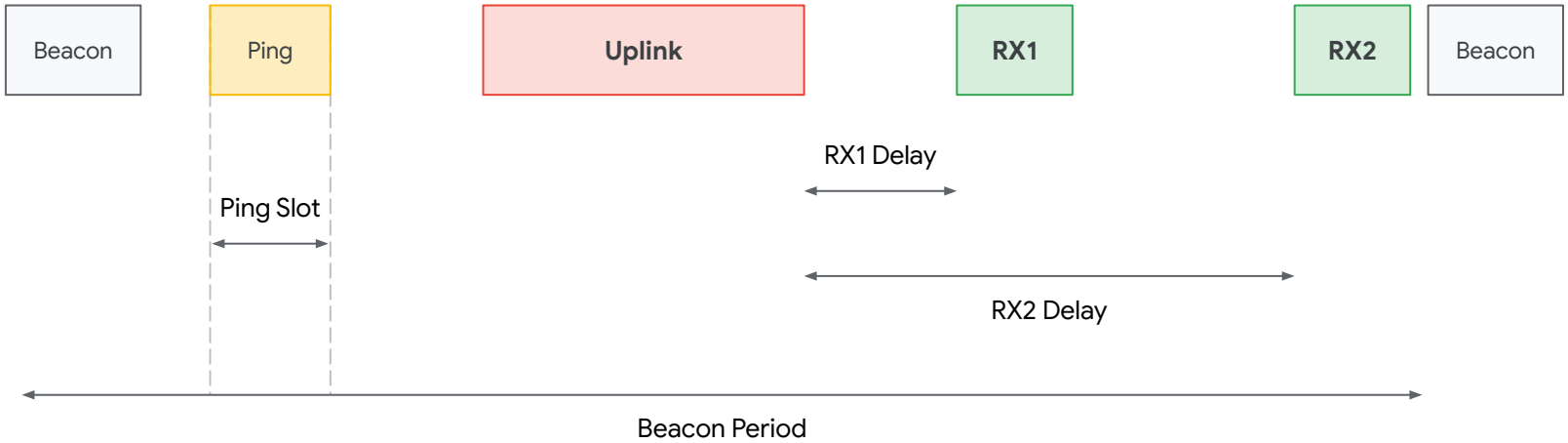
- Long/Short range
- Power consumption
- Data rate



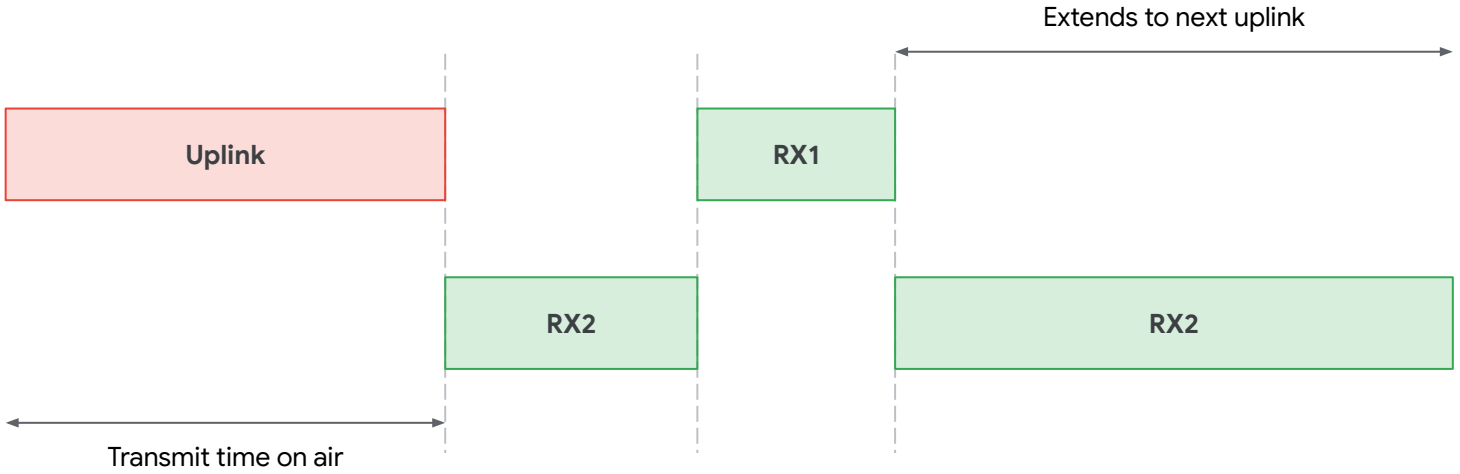
Class A



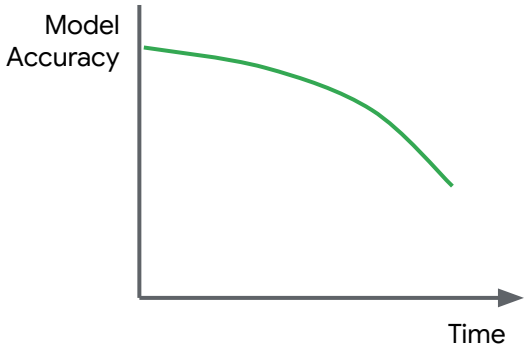
Class B



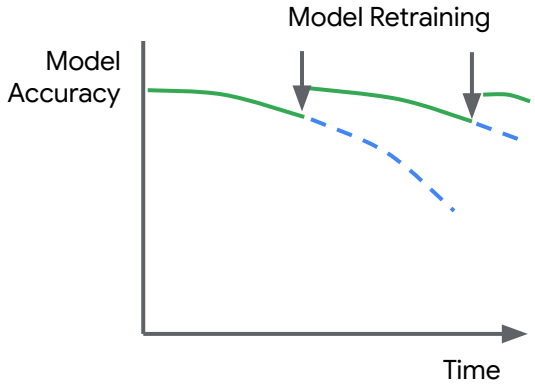
Class C



Different Communication Protocols



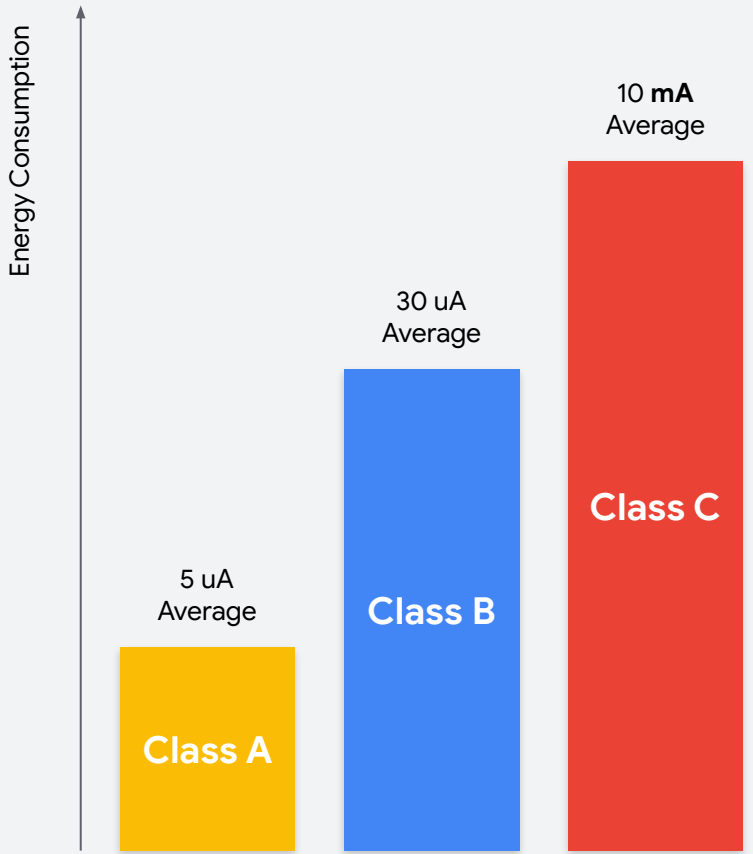
Model Decay over time



Regularly updated model

Cross Comparison

- **Class A** device communication is initiated only by the end device
- **Class B** devices have regularly-scheduled windows, in addition to those that open when a Class A-style uplink is sent to server
- **Class C** devices achieve the lowest latency among all operating modes



Cross Comparison

- **Class A** device communication is initiated only by the end device
- **Class B** devices have regularly-scheduled windows, in addition to those that open when a Class A-style uplink is sent to server
- **Class C** devices achieve the lowest latency among all operating modes