COMP 6611C: Advanced Topics in Embedded AI Systems

Lecture 2: Challenges in Embedded AI Systems

Xiaomin Ouyang

Assistant Professor

Department of Computer Science and Engineering, HKUST

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Recap

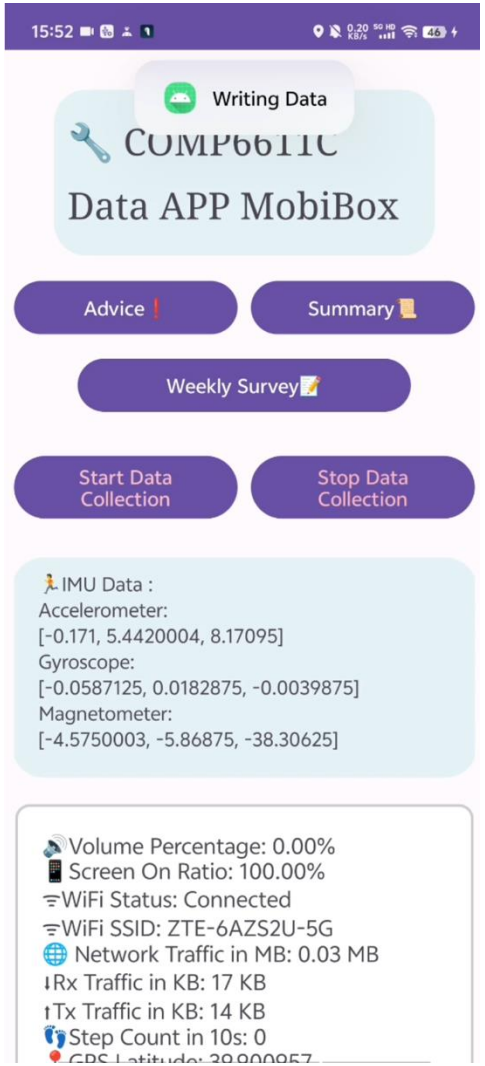
- Paper selection spreadsheet by end of today (Feb 11)
  - Auditing students are also **expected to present one paper**
- Team formulation spreadsheet by end of Friday (Feb 14)
  - Auditing students are encouraged (not mandatory) to join in a project team
  - About 7-8 teams (15-17 students now)
- Q&A recording spreadsheet

Outline

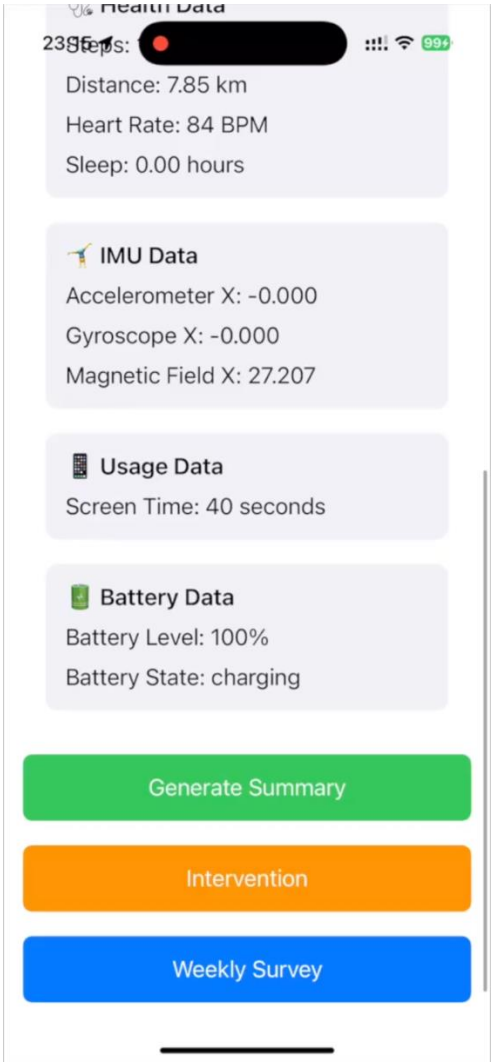
- **Course APP Introduction and Installation**
- Challenges in Embedded AI Systems

MobiBox APP and Dataset

- Data Collection
  - IMU (Accelerometer / Gyroscope / Magnetic), GPS, Screen & App Usage, Battery Status, Bluetooth connection, Network Traffic, Step Count, Wi-Fi Connections.
- Daily & Weekly Activity Summary
- Bump-up Activity Advice



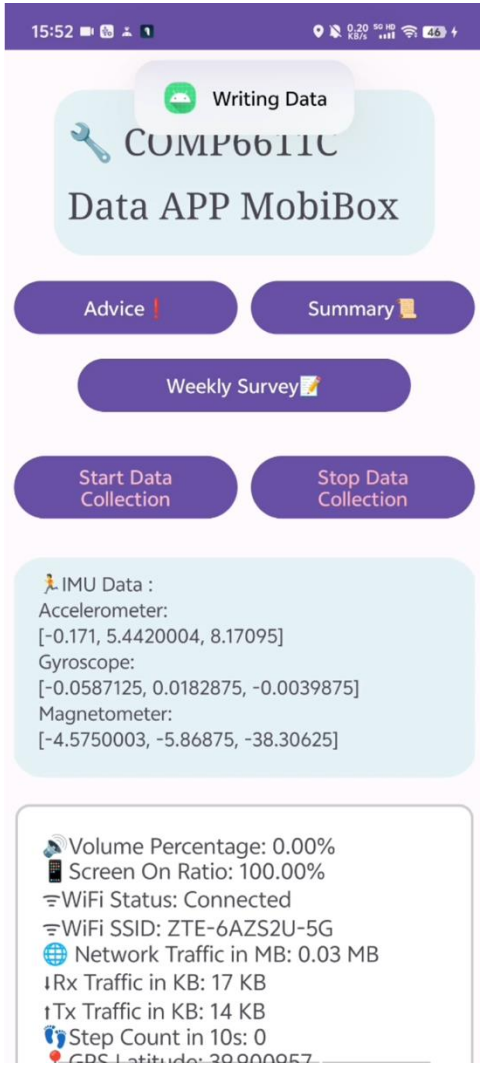
Android version



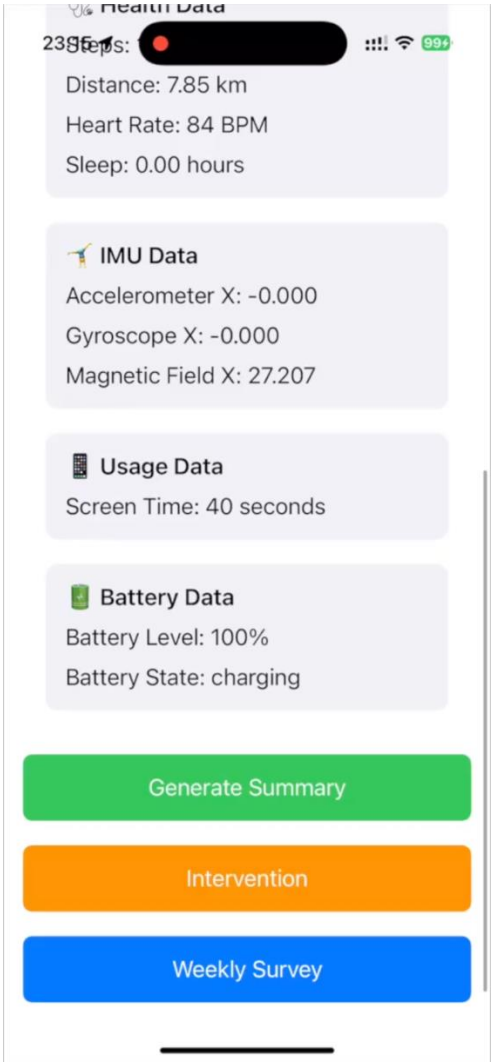
IOS version

MobiBox APP and Dataset

- Data Collecting Duration:
  - 8 a.m. – 10 p.m.
- Daily actions:
  - Start the APP at 8 a.m.
  - Check the APP at 10 p.m.
  - Pump-up advice every hour



Android version



IOS version

MobiBox APP Installment

- Android version
  - Directly install the App by APK package: MobiBox-App

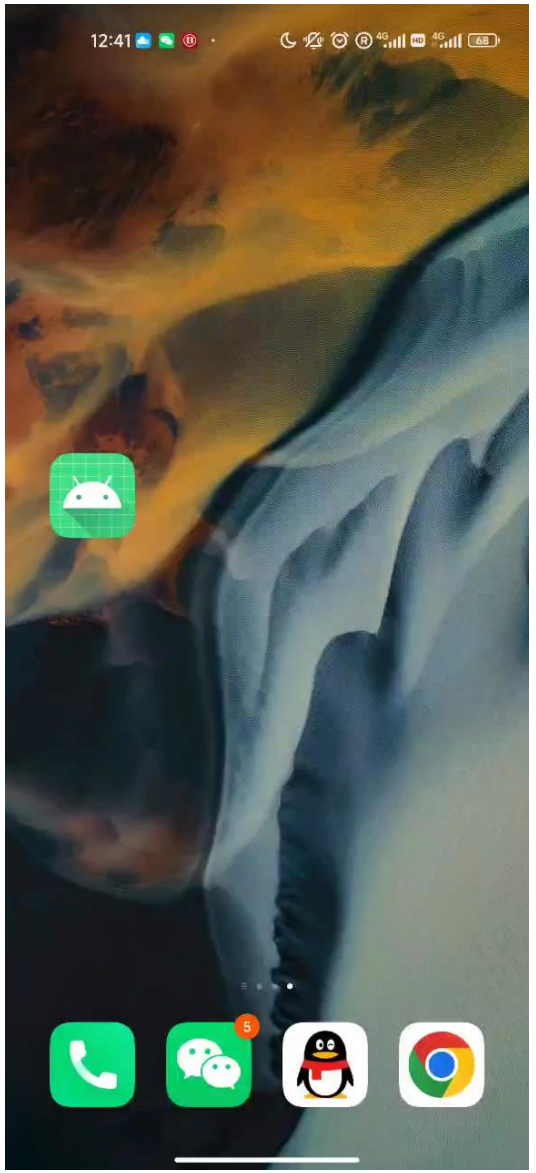
- IOS version:
  - Coming soon.

- Ensure that the necessary permissions are granted.

MobiBox APP Usage

- Should grant necessary permissions at first use.
  - Android: follow the right video instruction.
  - IOS: follow internal instruction

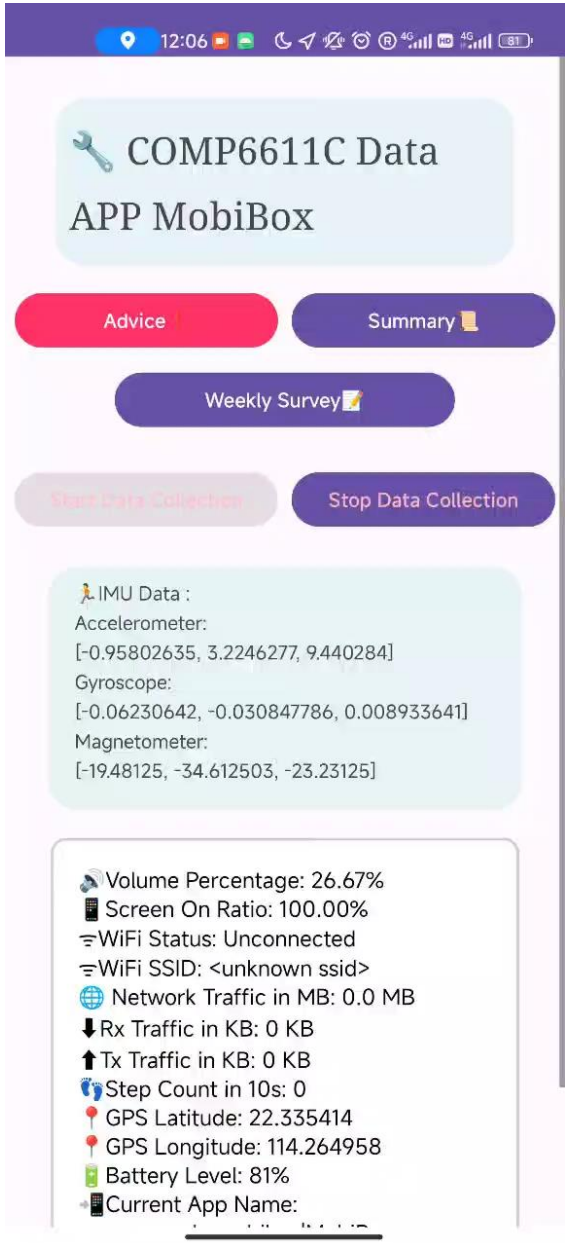
- The app continuously collects sensor data in the background. You can use your phone as usual, but it must remain active in the background.



Android version

Activity Advice

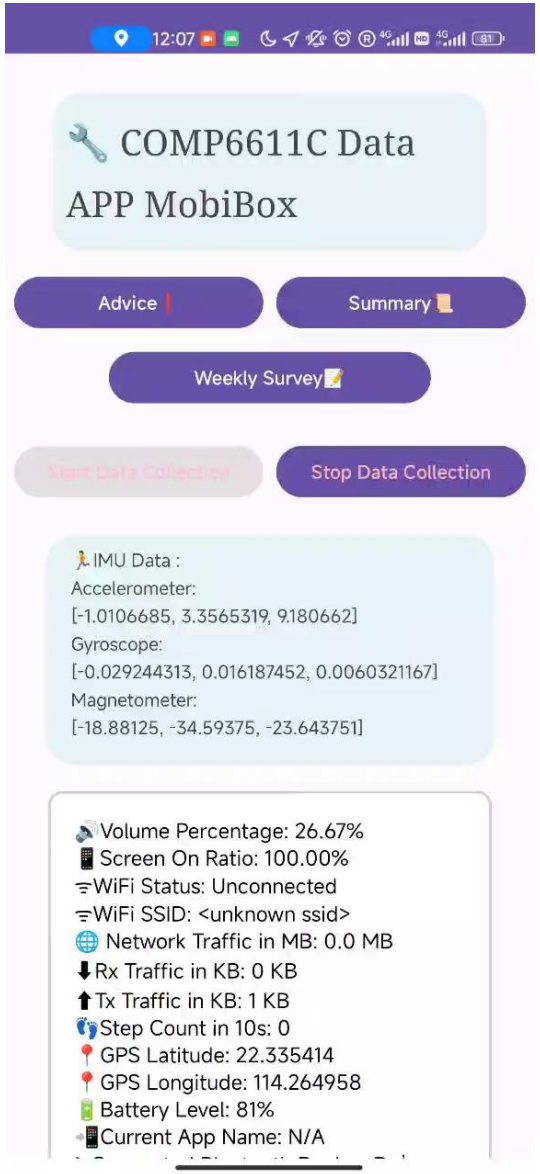
- Pump-up every hour, or pro-active click
- After receiving the **activity advice,** Please provide a quick feedback in the App.



Activity Summary

- Daily Summary Notification:
  - Every day at 10:00 p.m.
  - Please select the best option of summary and write the feedback

- Weekly Mental Health Survey:
  - Every Sunday at 10:00 p.m.
  - Please complete the Weekly Survey manually.



Android version

Group Chat for Feedback and Discussions

- To ensure a smooth and convenient data collection process, please join the WhatsApp/Wechat group chat.
- This will help us stay organized and keep you updated with important reminders. Thank you!

COMP6611C APP Group

WhatsApp 群组

WhatsApp Group

群聊: COMP6611C Group

Wechat Group

Bonus for the Dataset Collection

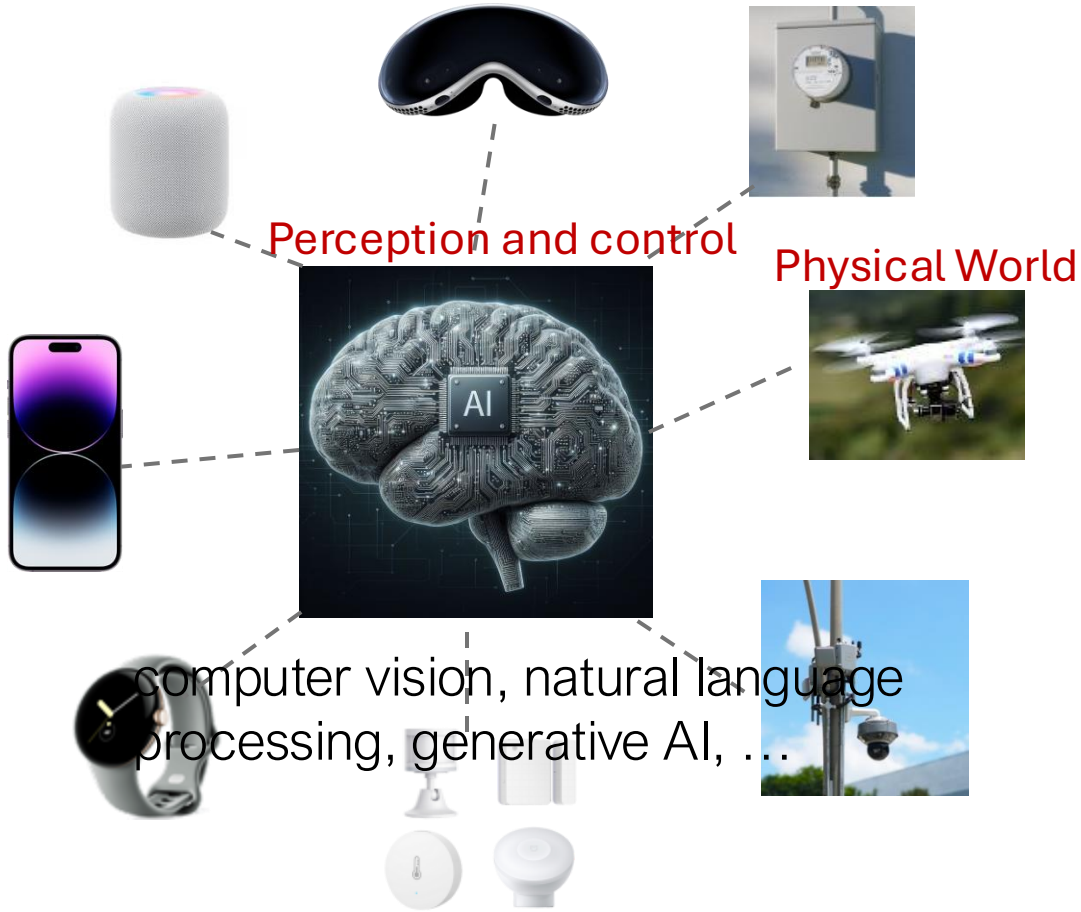
- Complete Data collection can obtain up to 10 extra points.
- Account in Course Project.

Outline

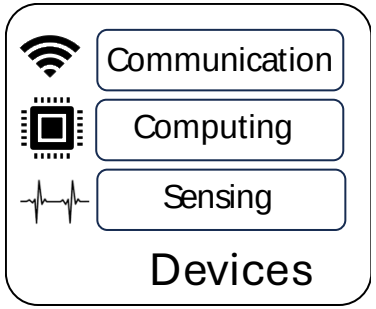
- Course APP Introduction and Installation
- **Challenges in Embedded AI Systems**
  - You may refer to them for deciding your project topics

What is Embedded AI

- AI on network edge, physical devices & “Things”
  - Mobiles, wearables, vehicles, robots, sensors



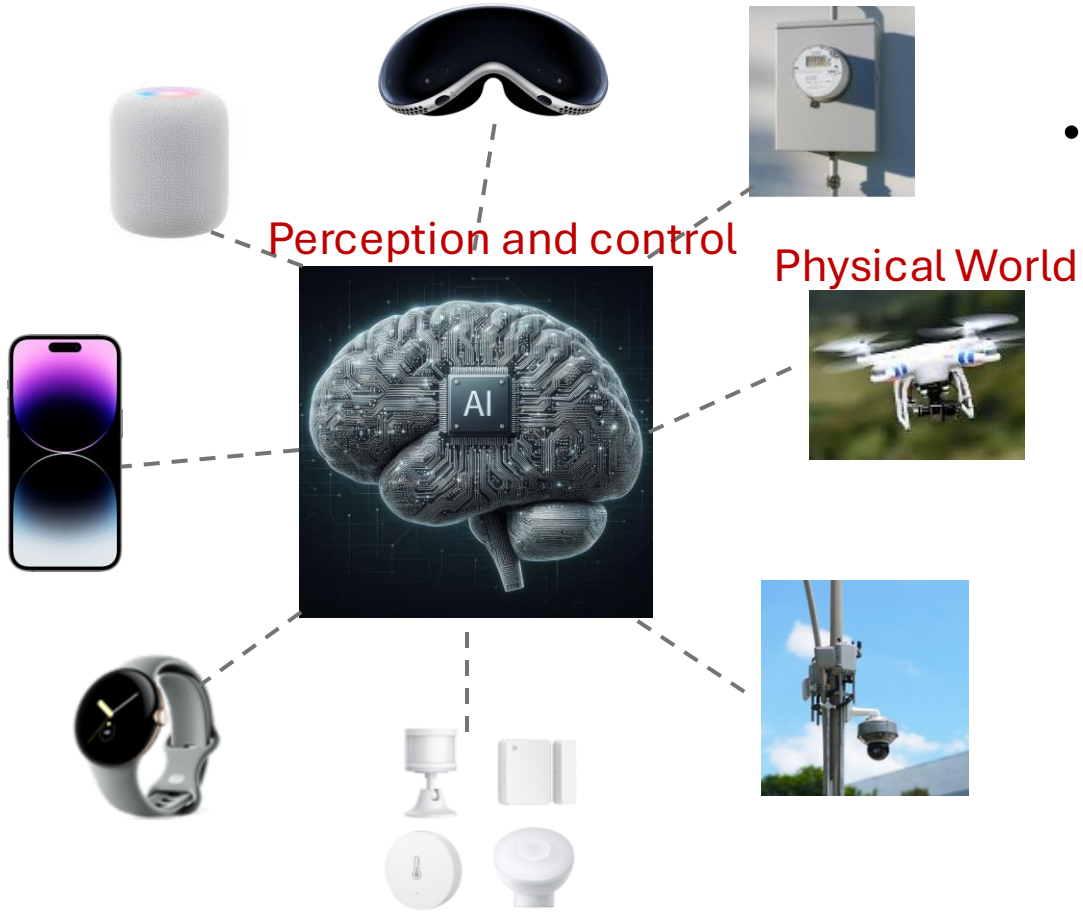
Environment with **Ambient Intelligence**



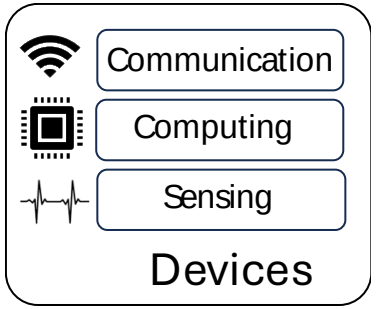
- In-situ sensing
- On-device computing
- Networked computing

What is Embedded AI

- AI on network edge, physical devices & “Things”
  - Mobiles, wearables, vehicles, robots, sensors



- **AI on resource-constrained devices**



- Real-time, efficient, and reliable intelligence at the edge.

Challenges in Embedded AI Systems

- **Data Challenges**
- **System Challenges**
- Challenges related to Specific Sensor Modalities
- Challenges related to Specific Applications

Data Challenges

➢ How to harness **distributed and imperfect data**?

- Distributed: Data siloed across devices/users/locations.
- Imperfect: Data not ready for naïve (supervised) deep learning.

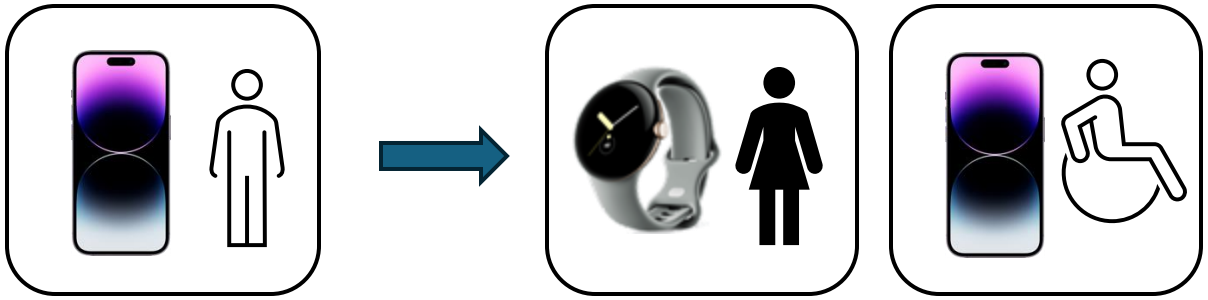
Data Challenges

- How to harness **distributed and imperfect data**?
  - **Distributed**
    - Domain shift
    - Security and privacy
  - **Imperfect**
    - Limited labeled data
    - Missing data
    - Data heterogeneity and noise
    - Data skewness
  - **Others**
    - Complex event detection
    - Long-term data analysis

Data Challenges

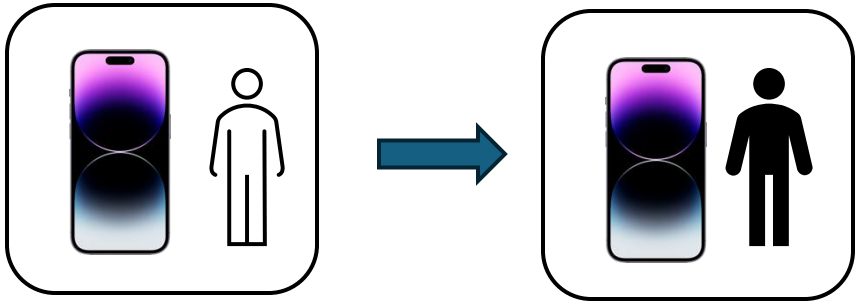
➢ Domain shift:

- **Spatial:** adapt to new subjects/environments



- **Techniques:**
  - Domain generalization
  - Domain adaptation
  - Meta learning

- **Temporal:** distribution shifts over time after deployment

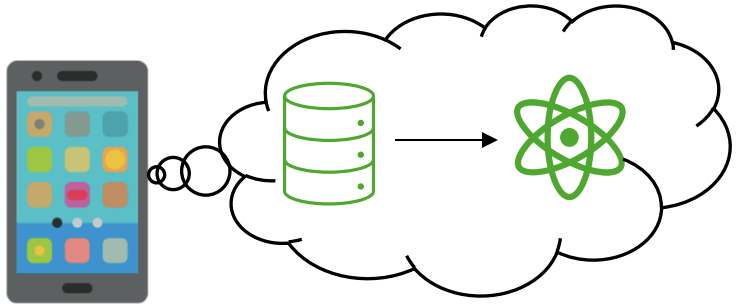


- **Techniques:**
  - Continuous/online learning
  - Test-time adaptation

Data Challenges

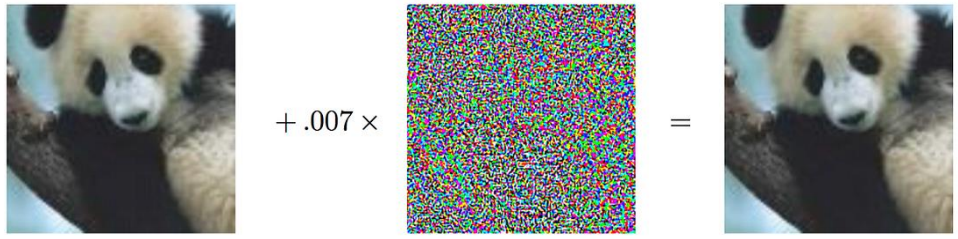
- Security and Privacy:

  - **Data privacy:** cannot be uploaded to cloud



  - **Techniques:**
    - Federated Learning
    - Edge computing
    - Cloud-edge cooperation

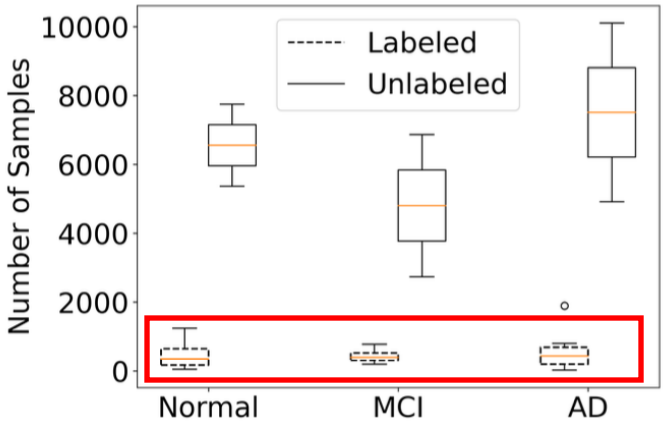
  - **Data security:** physical/adversarial/backdoor attacks

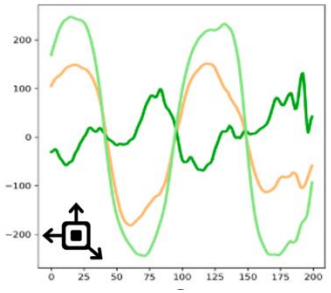


  - **Techniques:**
    - Encryption
    - Adversarial Training
    - Anomaly Detection

Data Challenges

➢ Label scarcity

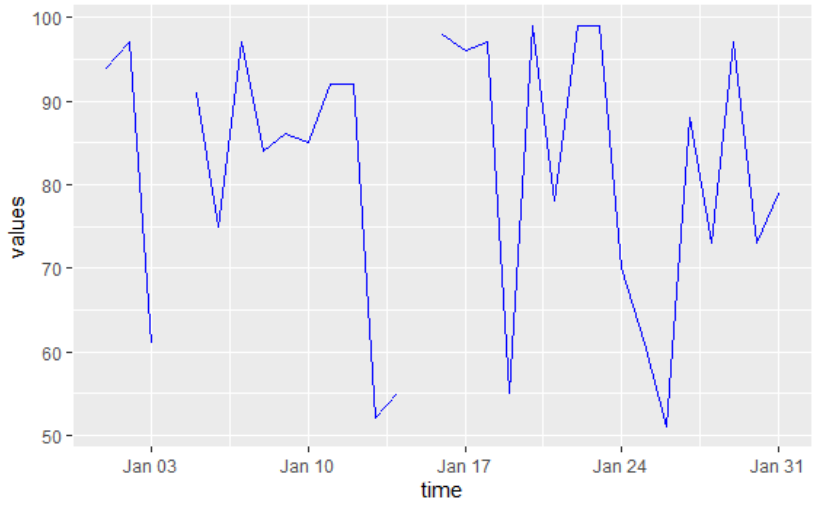


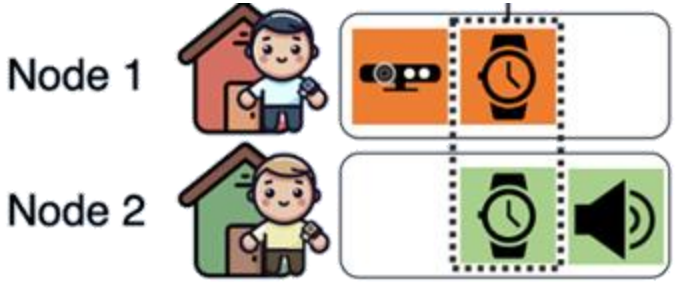


- **Techniques:**
  - Unsupervised/Semi-supervised learning
  - Active learning
  - Reinforcement learning
  - Weakly supervised learning

Data Challenges

➢ Missing data

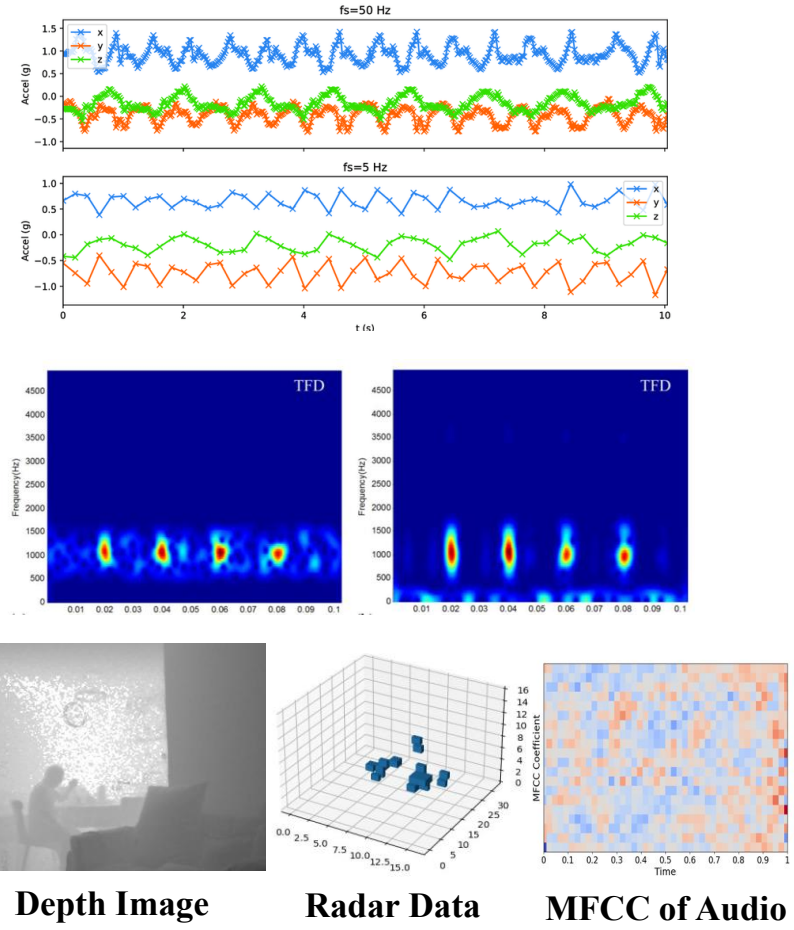




- **Techniques:**
  - Data imputation/reconstruction/generation
  - Learning with missing data
    - masked training

Data Challenges

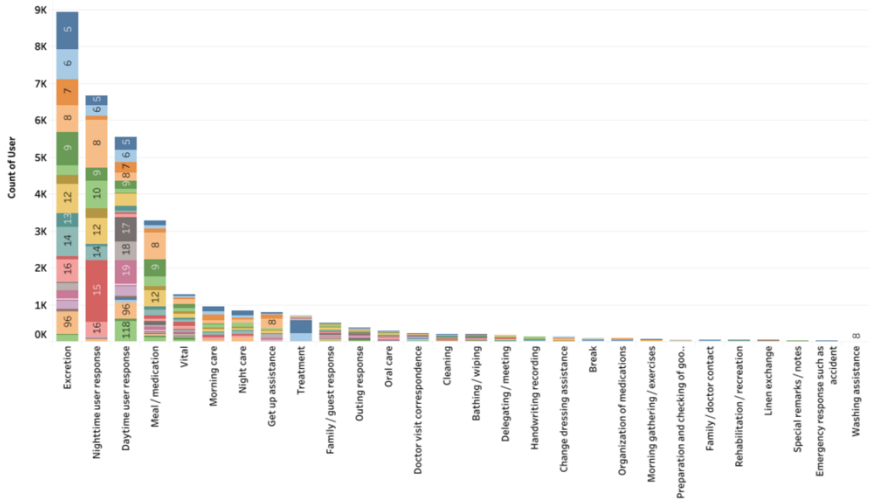
- Data heterogeneity and noise

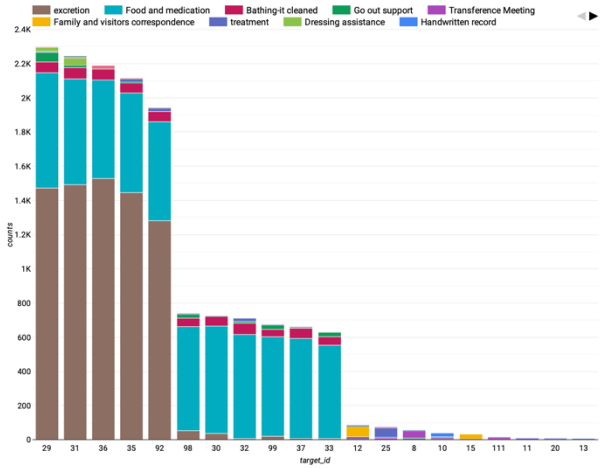


- **Techniques:**
  - Data preprocessing: sampling/denoise
  - Robust representation learning
  - Multimodal Fusion

Data Challenges

- Data skewness/imbalance

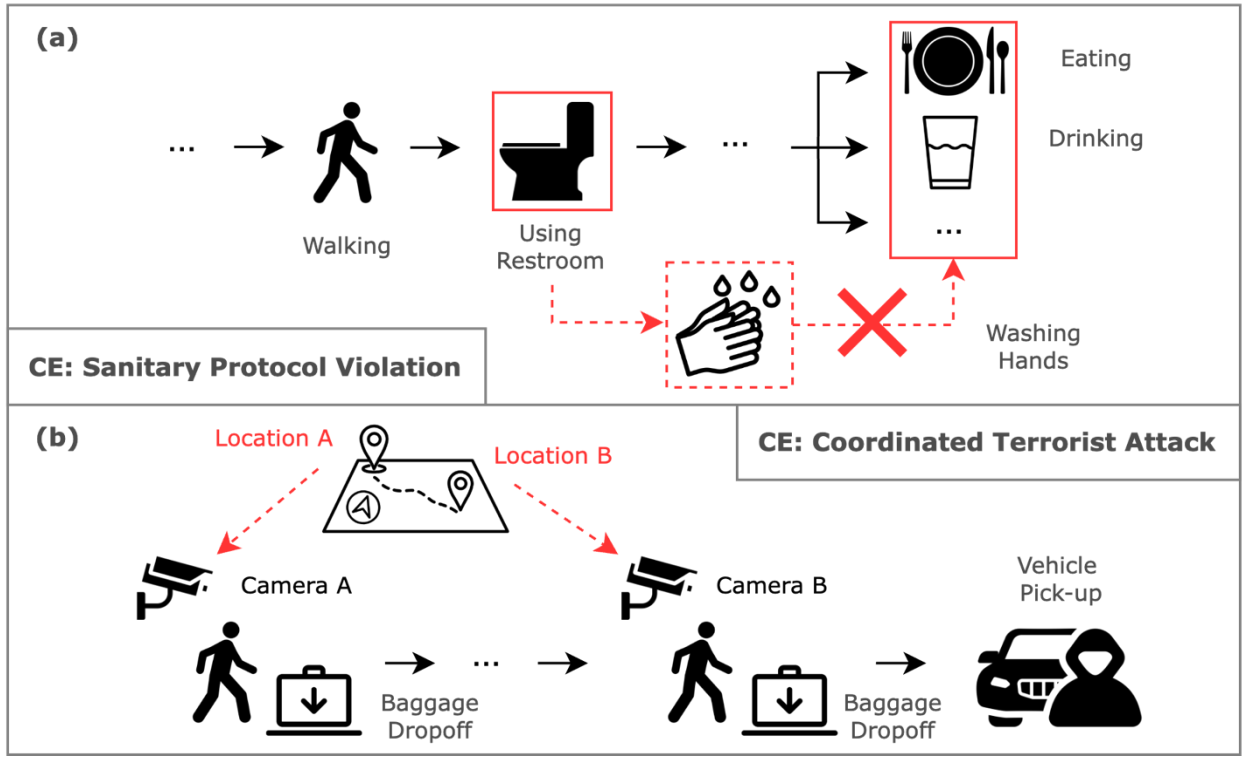




- **Techniques:**
  - Resampling
  - Data augmentation
  - Penalize training loss functions

Data Challenges

➢ Complex Event Detection

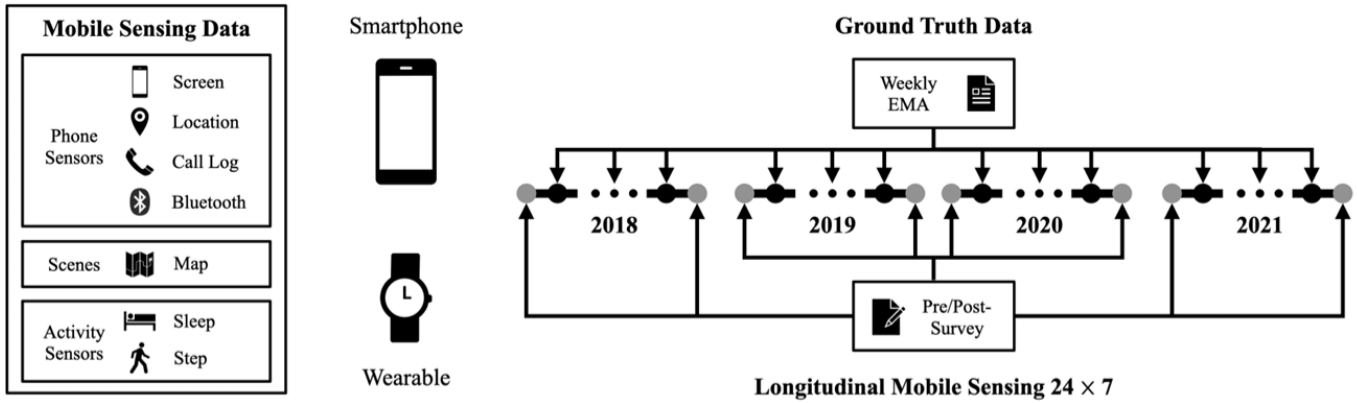


Complex Event: A series of Atom Events

- **Techniques:**
  - Neuro-Symbolic Learning
  - Spatio-Temporal Analysis
  - LLMs

Data Challenges

➢ Long-term Data Analysis

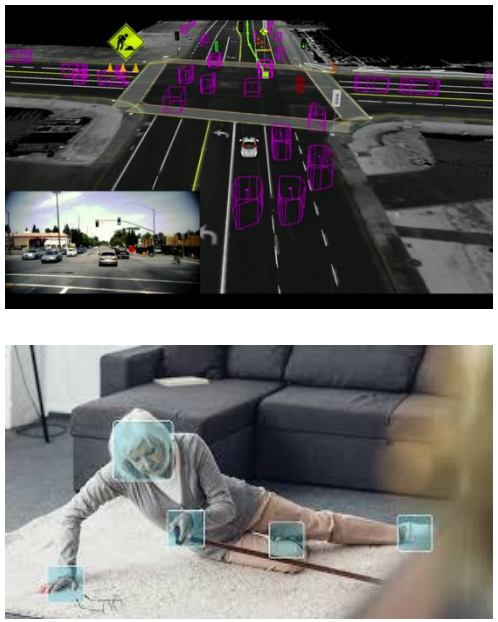


- **Techniques:**
  - Hierarchical analysis
  - Time-series data analysis
  - Memory-based Model (Mamba)

System Challenges

- How can we make the system more **resource-efficient, real-time, robust, and scalable**?

- Resource constraints



- Task requirements

System Challenges

- How can we make the system more **resource-efficient, real-time, robust, and scalable**?
  - **Limited resources**
  - **Real-time Performance**
  - **Robustness**
  - **Scalability**

System Challenges

➢ Limited resources – memory usage

To give some examples of how much VRAM it roughly takes to load a model in bfloat16:

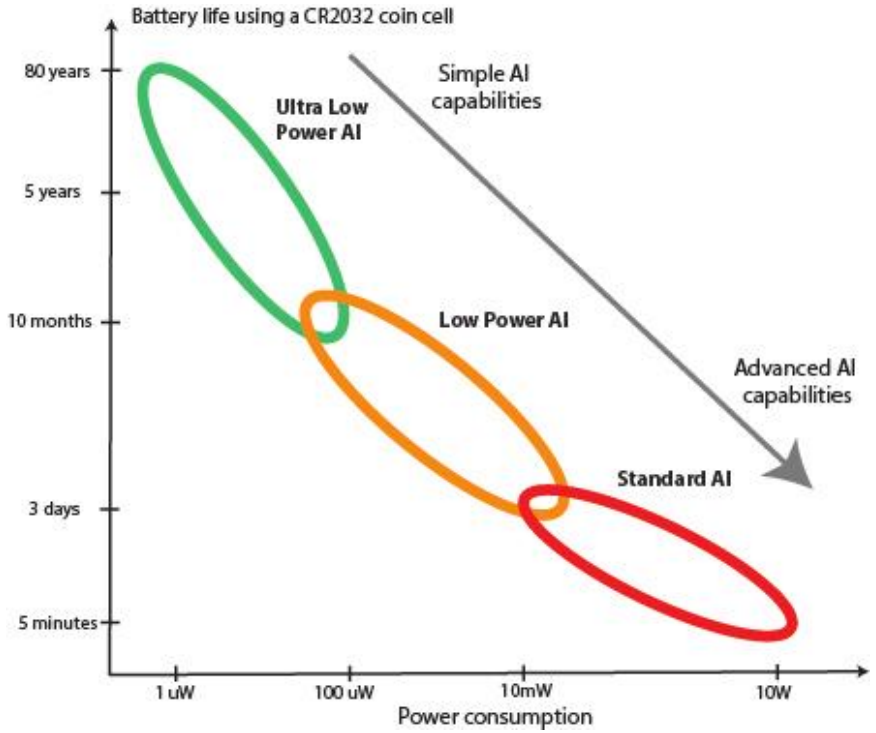
- **GPT3** requires 2 * 175 GB = **350 GB** VRAM
- **Bloom** requires 2 * 176 GB = **352 GB** VRAM
- **Llama-2-70b** requires 2 * 70 GB = **140 GB** VRAM
- **Falcon-40b** requires 2 * 40 GB = **80 GB** VRAM
- **MPT-30b** requires 2 * 30 GB = **60 GB** VRAM
- **bigcode/starcoder** requires 2 * 15.5 = **31 GB** VRAM

Device Name	CPUs	RAM (GB)
Google Pixel 6 Pro	2x2.80 GHz Cortex-X1 2x2.25 GHz Cortex-A76 4x1.80 GHz Cortex-A55	12
Xiaomi Mi Mix 2S	4x2.8 GHz Kryo 385 Gold 4x1.8 GHz Kryo 385 Silver	6
Raspberry Pi 4B	4x1.8 GHz Cortex-A72	8

- **Techniques:**
  - Model compression
  - Neural architecture search
  - Online memory management (model slicing/chunks, communication scheduling)

System Challenges

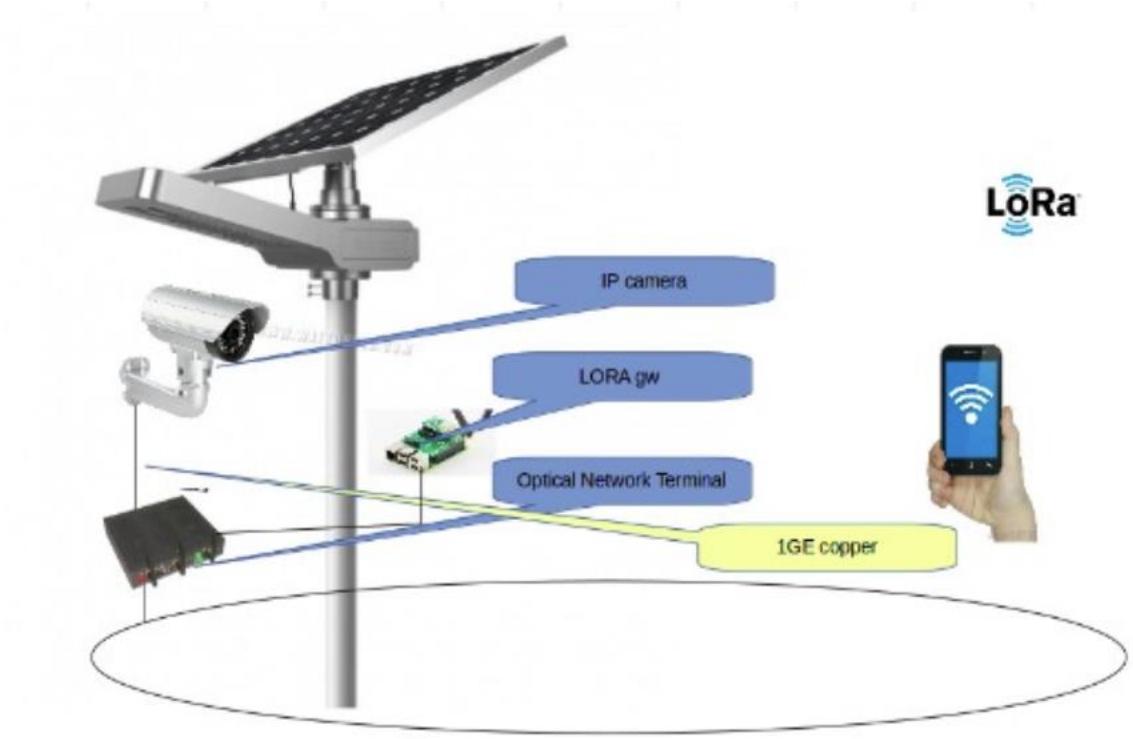
- Limited resources – power efficiency



- **Techniques:**
  - Model compression
  - Power management (e.g., adaptive voltage scaling, energy-aware scheduling)
  - Data filtering/duty cycling

System Challenges

- Limited resources – communication bandwidth



Video streaming, autonomous driving, 3D model rendering

- **Techniques:**
  - (Context-aware) Data compression/ filtering
  - Feature extraction
  - Edge computing

System Challenges

➢ Real-time performance

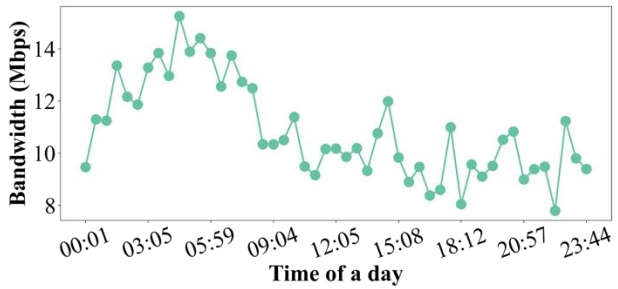
- **Techniques:**
  - Caching (space -> time)
  - Progressive inference or speculative decoding in LLMs
  - Context-aware inference
  - Multi-task scheduling

Inference latency no less than xx seconds/ milliseconds.

System Challenges

➢ Robustness

System Dynamics

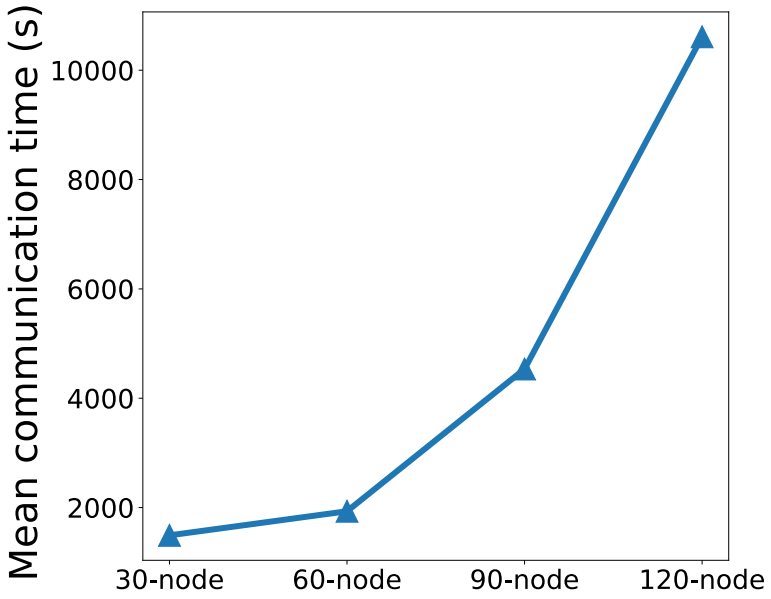


Heterogeneous Platforms

- **Techniques:**
  - RL-based tasks scheduling
  - Dynamic Resource Management
  - CPU-GPU co-scheduling

System Challenges

➢ Scalability



Ability to maintain performance, efficiency, and reliability as the system grows.

- **Techniques:**
  - Node selection/dropout/scheduling
  - Synchronization -> Asynchronization
  - Split learning between server and devices

Break

- **Next lecture: Challenges in Embedded AI Systems - Cont'd**
- **Reminder:**
  - **IOS version of Course APP to be released soon.**
  - **First paper presentations on next Thursday (Feb 20)**
- **Any questions?**