① INTRODUCTION TO EDGE AI

- Edge ≈ local processing
  ≠ in the cloud

- It is needed where <u>low latency</u> is required, or
  where network may not be available.

- It is mainly used for <u>real-time decision-making</u>

Application of AI at the Edge

① Network communications can be expensive and at times impossible (remote locations / natural disasters)
  — edge needs no (or little) ~~intere~~ network.

② Real-time processing is needed in many applications, like self ~~dri~~ driving cars

③ Edge application can be used for sensitive data (like health data) as they are not sent to the cloud.

④ Optimization software can achieve great efficiency w/ edge AI models.

Reasons for development at the Edge

① Proliferation of devices

② Need for low-latency compute

③ Need for disconnected devices

Structure of the Course

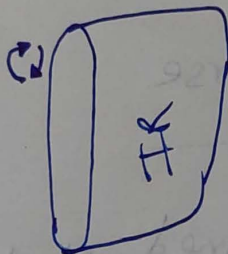① Leveraging Pre-trained Models

② The Model Optimizer

③ The Inference Engine

④ Deploying an Edge App

```
┌──────────┐          ┌──────────────┐
│ Train a  │  ──────► │  Run Model   │
│  Model   │          │  Optimizer   │
└──────────┘          └──────────────┘
                            │
                            ▼
                      ┌──────────┐        ┌──────────┐
                      │   IR     │ ─────► │Inference │
                      │  .xml    │        │ Engine   │
                      │  .bin    │        └──────────┘
                      └──────────┘             │
                                               ▼
                                        ┌──────────────┐
                                        │    Edge      │
                                        │ Application  │
                                        └──────────────┘
```

flowchart of the course