Low Latency and AI/ML

- What is low latency?
  - Different use cases have varying latency requirements
  - We are talking about 1ms - 30ms, guaranteed
- What kind of latency we are talking about?
  - End-2-End? Layer-2-Layer?
  - Impact of lower layers on latency
    - For example 4G/5G networks and there protocols (PDCP, RLC, MAC, PHY)
- Data awareness could help to optimize latency
  - Duplication, Discarding, Prioritization
  - Data awareness requires that protocols can “understand” the payload
  - Encryption makes it hard to optimize latency
  - How to provide meta-data, context to the lower layers? Especially in case of cellular networks
- How can AI/ML help?
  - Semantic Communication

Low Latency and AI/ML

- Controversial AI/ML discussion
  - LLMs are a hype?
  - How useful are LLMs for network optimizations?
    - Congestion control —> seems mostly to be not accepted
- Instead of LLMs, do we need domain specific ML?
  - Federated learning
  - Small Language Models (SLM)