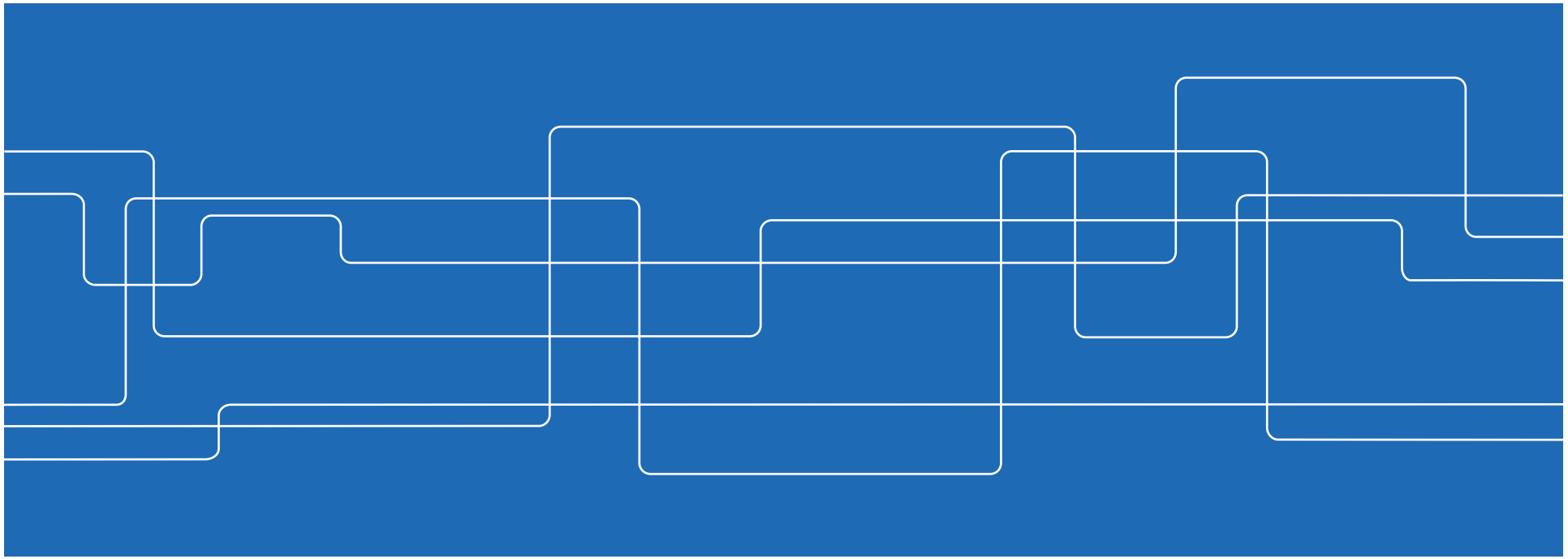
Internet of Reality: Challenges, Initial Results and the Road Ahead

Ericsson Research Lecture Series

James Gross

Dec 7th, 2020



James Gross

- Professor for Machine-to-Machine Communications
  - PhD from TU Berlin in 2006
  - Assistant Prof. at RWTH 2008 – 2012
  - Since 2012 at KTH

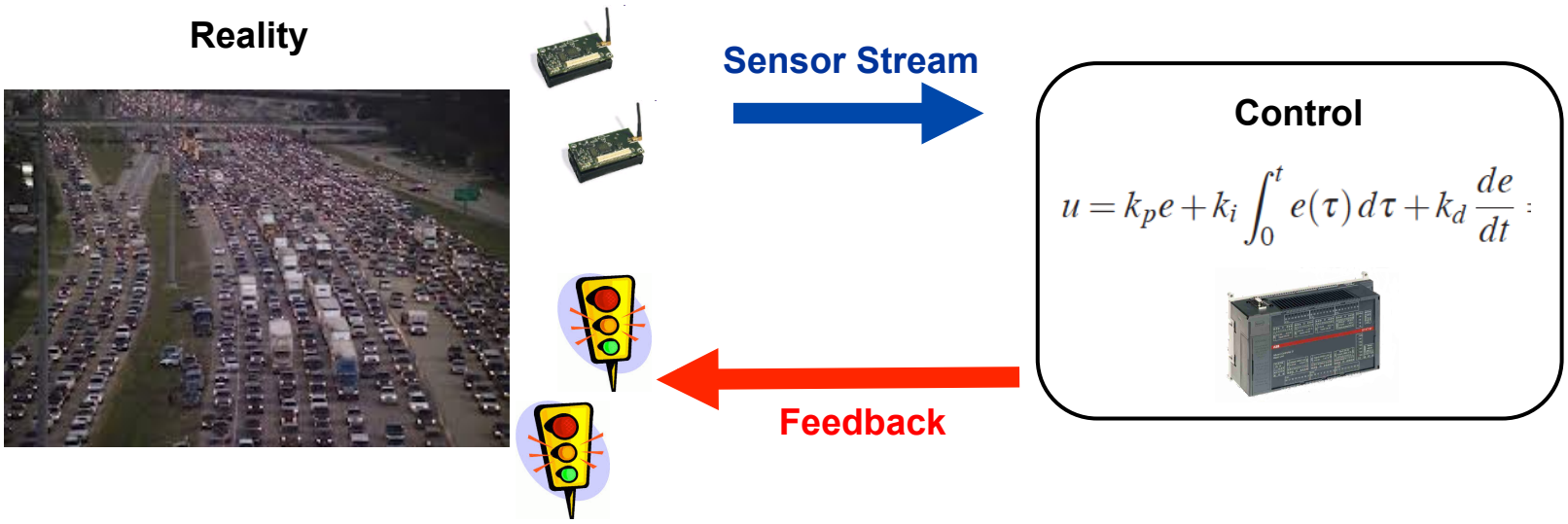
- Research focus:
  - Cellular networks, critical machine-to-machine communications
  - Network performance models
  - Edge computing for closed-loop applications

- Associate Director KTH Digital Futures
- Co-Director KTH Competene Center TECoSA
- Founder@R3 Communications (spin-off in industrial wireless)

Outline

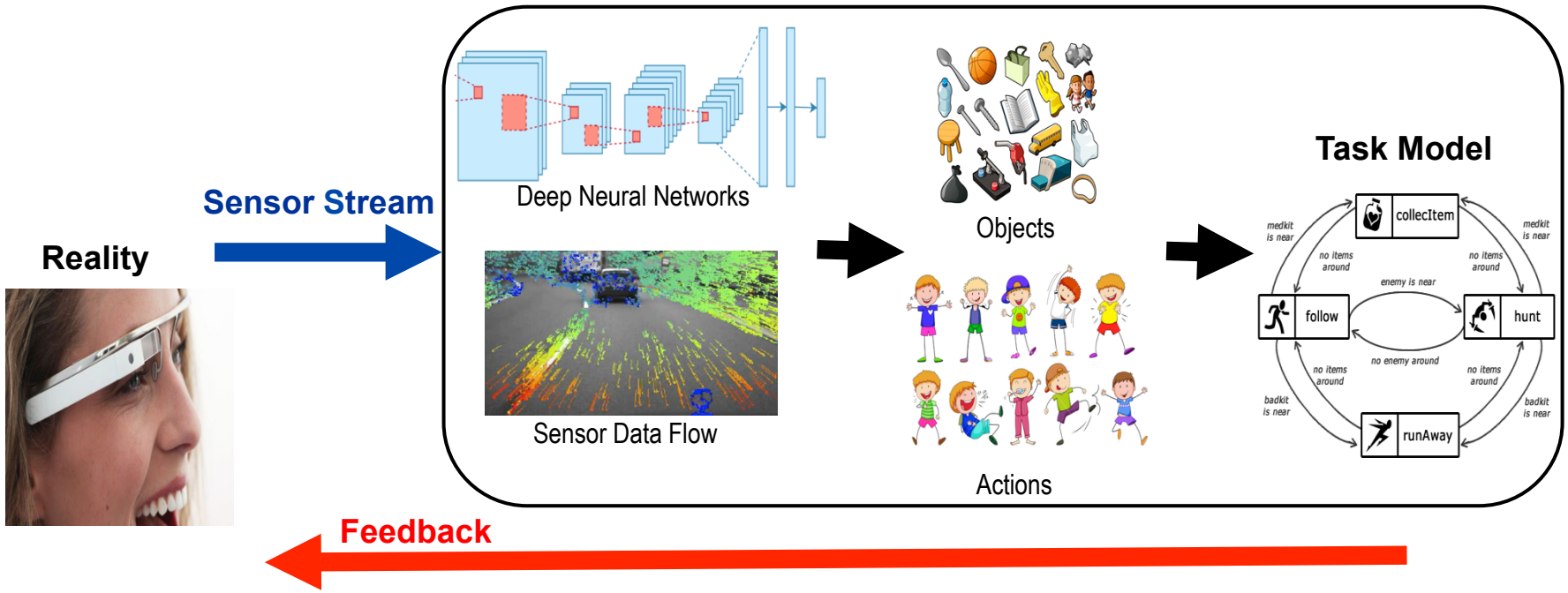
- **Towards an Internet of Reality?**
- Initial Challenges and Results
  - EdgeDroid and the quest for latency impact
  - Scheduling for closed-loop
- Upcoming Challenges & Outlook

Cyber-Physical System (CPS)



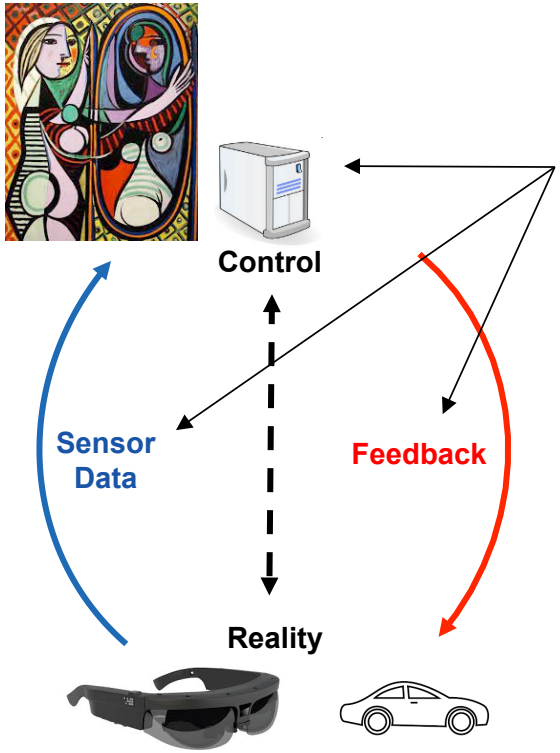
Traditionally in industrial automation, but broader use cases exist

Wearable Cognitive Assistant (HITL)



Generalizes to human-in-the-loop (HITL), many different scenarios

Underlying Principles

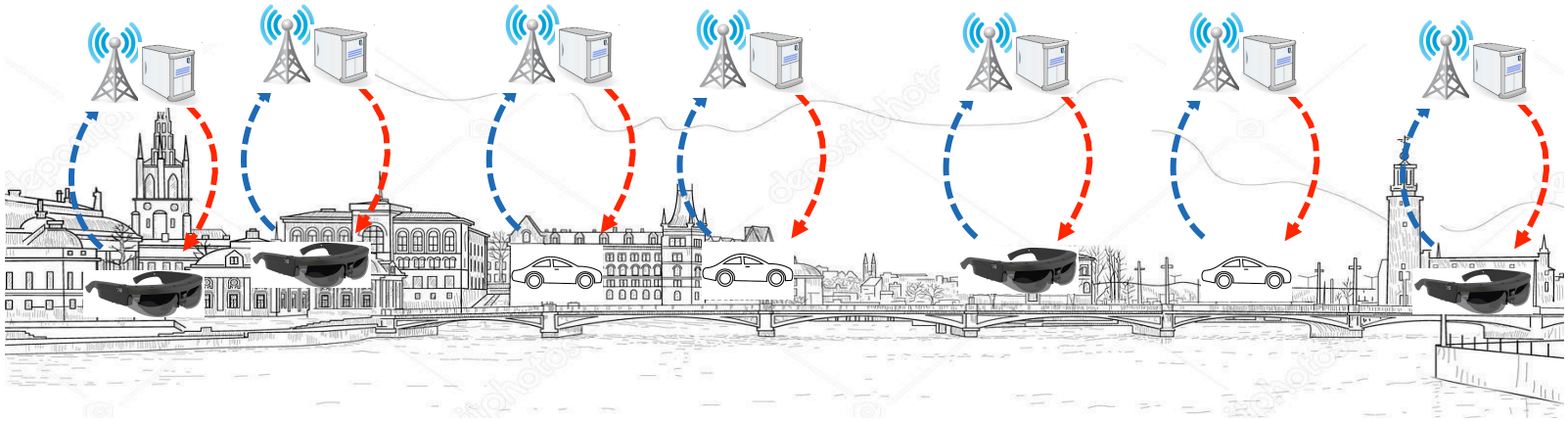


Representation of reality:
constant update of a specific context

Feedback system:
loop set-up, dependable, 1-to-1 load ratio

Substantial utility of applications:
Automation gain, knowledge transfer, automated assistance

Ubiquitous Provisioning?



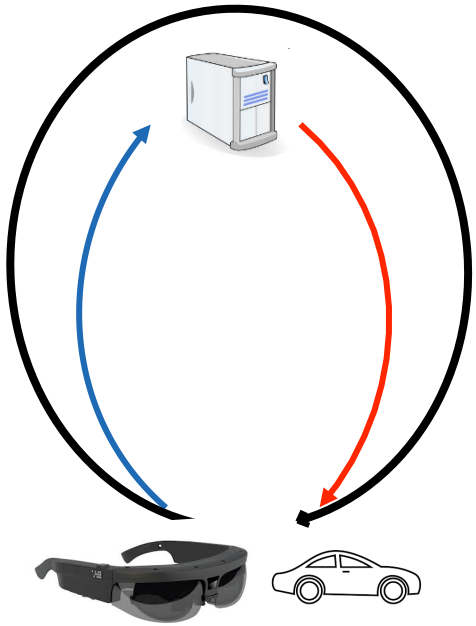
- Run over shared network infrastructure
- Efficient support of such applications?
- Interaction between applications and network?

Outline

- Towards an Internet of Reality?
- **Initial Challenges and Results**
  - **EdgeDroid and the quest for latency impact**
  - **Scheduling for closed-loop**
- Upcoming Challenges & Outlook

Application Characteristics

End-to-end latency over the loop is the central metric!



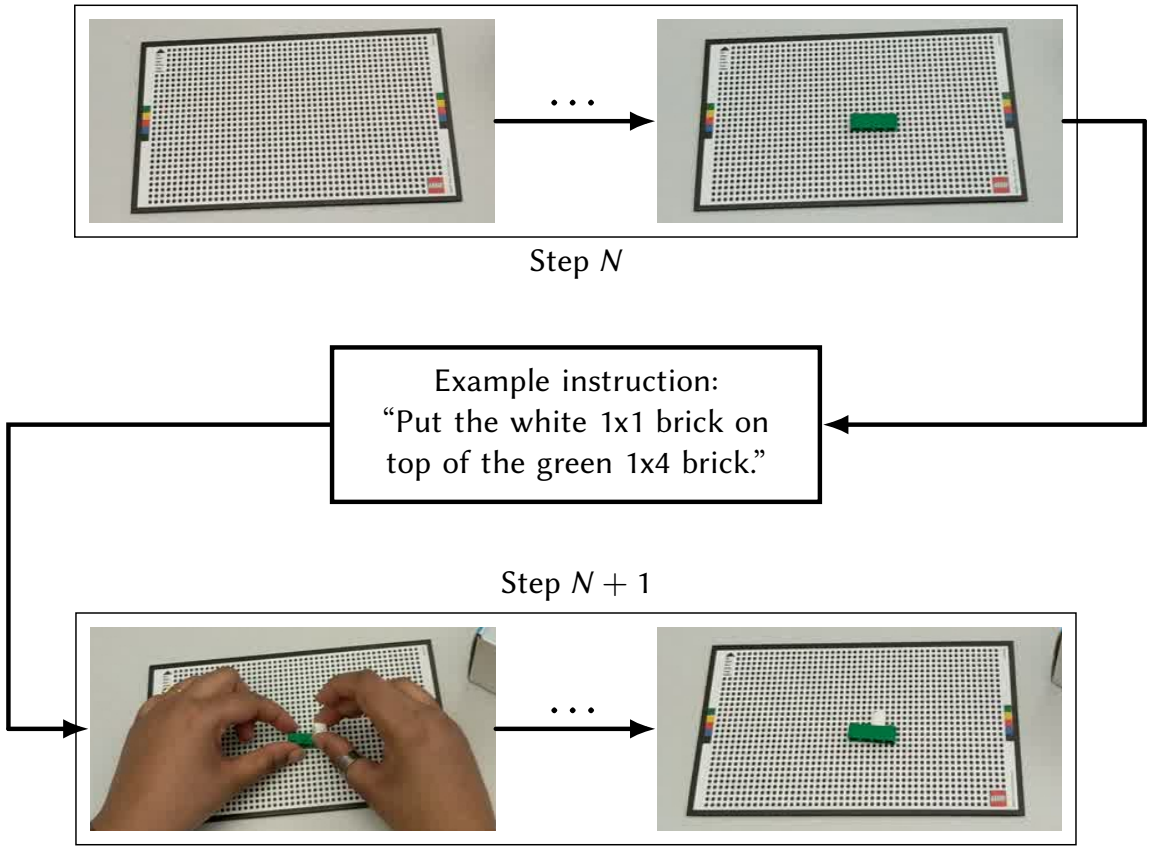
Diverse Footprints:

Uplink/Downlink
+ Compute

Diverse Requirements:

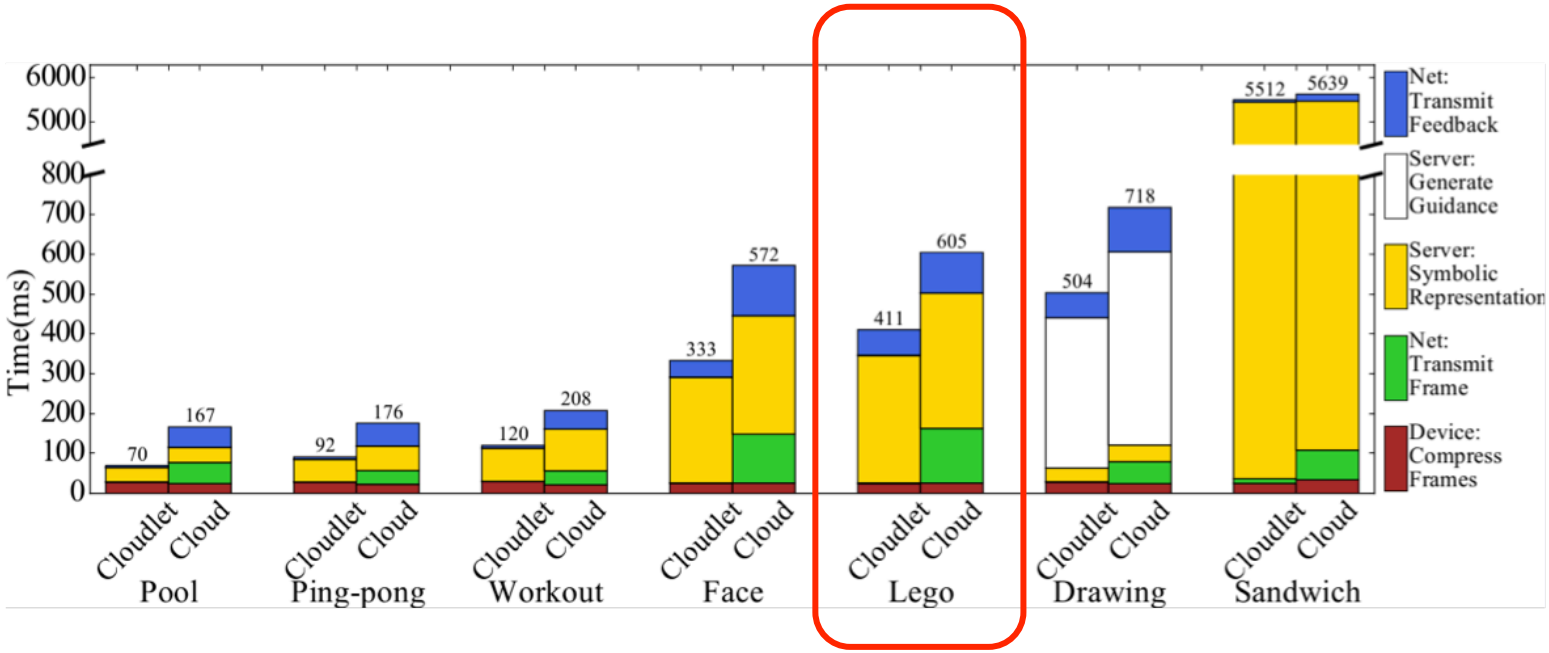
- HITL: ~ 800 ms
- CPS: ~10ms

Example HITL: LEGO Assistant



Ha et al. "Towards wearable cognitive assistance," *ACM Mobisys* 2014

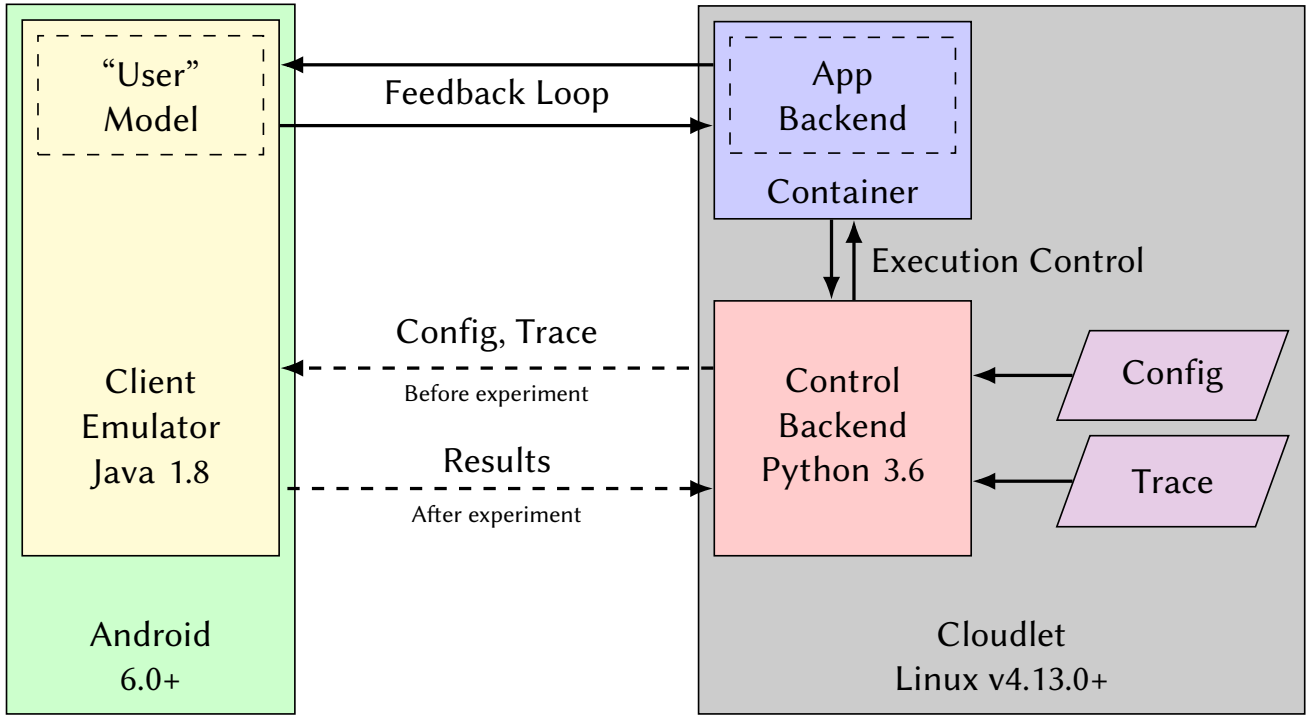
LEGO Footprint & Latency Requirement



QoE determined by latencies t_{low}=600 ms, t_{up} = 2.7 s !

Chen et al. “An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance,” *IEEE SEC* 2017.

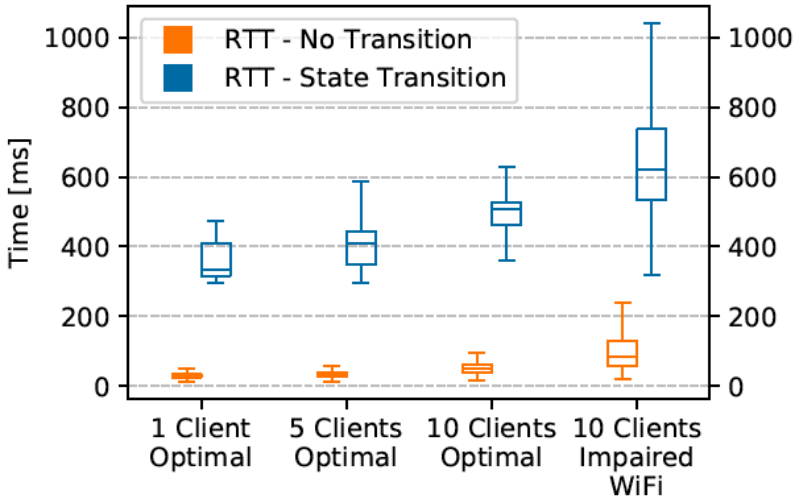
HITL Emulation: EdgeDroid

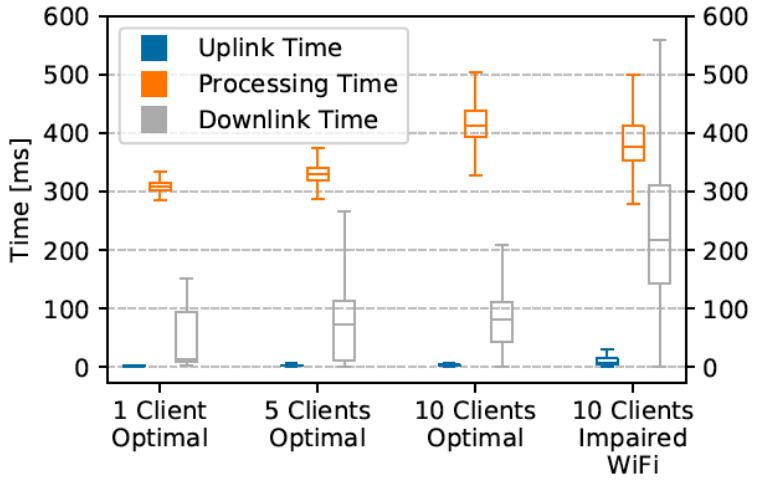


Muñoz et al. "EdgeDroid: An experimental approach to benchmarking human-in-the-loop applications," *ACM HotMobile* 2019.

Infrastructure Impact

IEEE 802.11n & simple cloudlet set-up, office environment

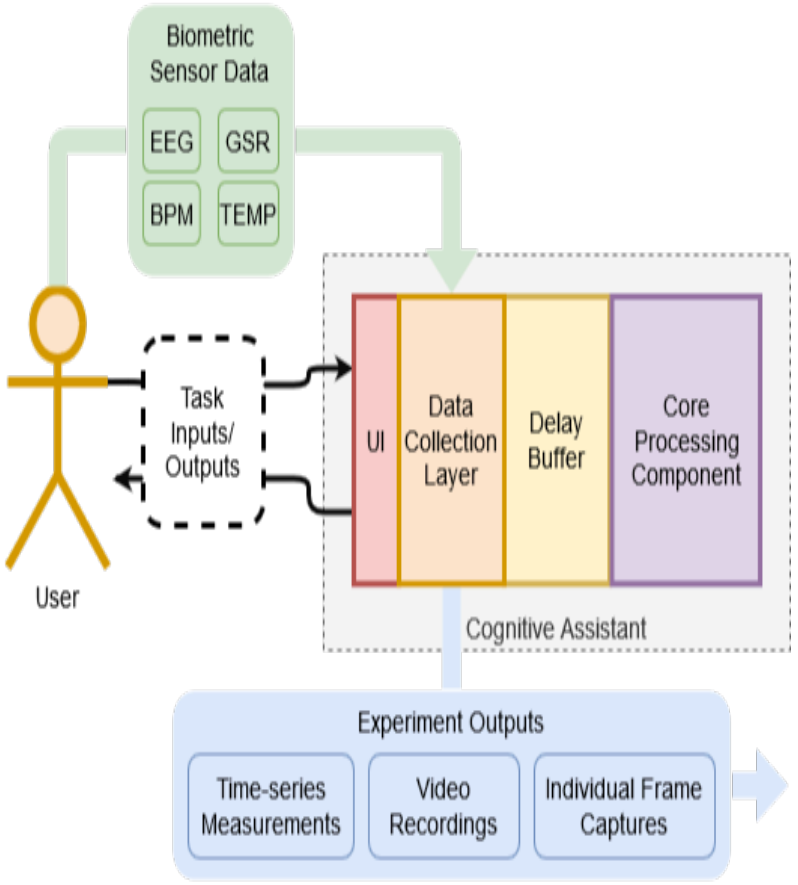




- Exponential latency scaling, various contributing factors
- Load easily pushes latencies beyond t_{down}, impact?

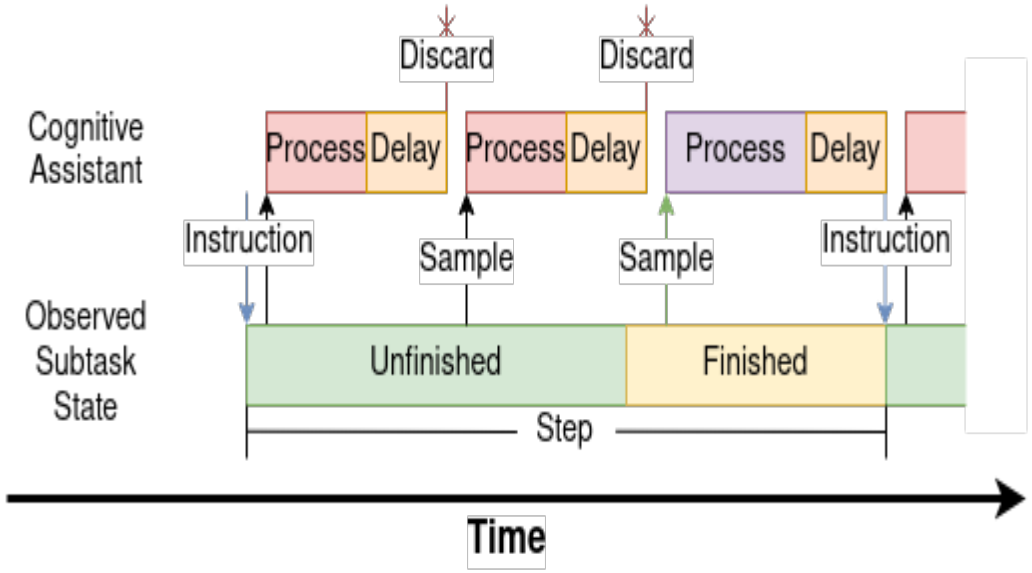
User Study on Delayed Feedback

- User study at CMU in fall 2019
- Modified assistant to control latencies
- Tracking of various outputs
- 40 participants, mostly students from CMU

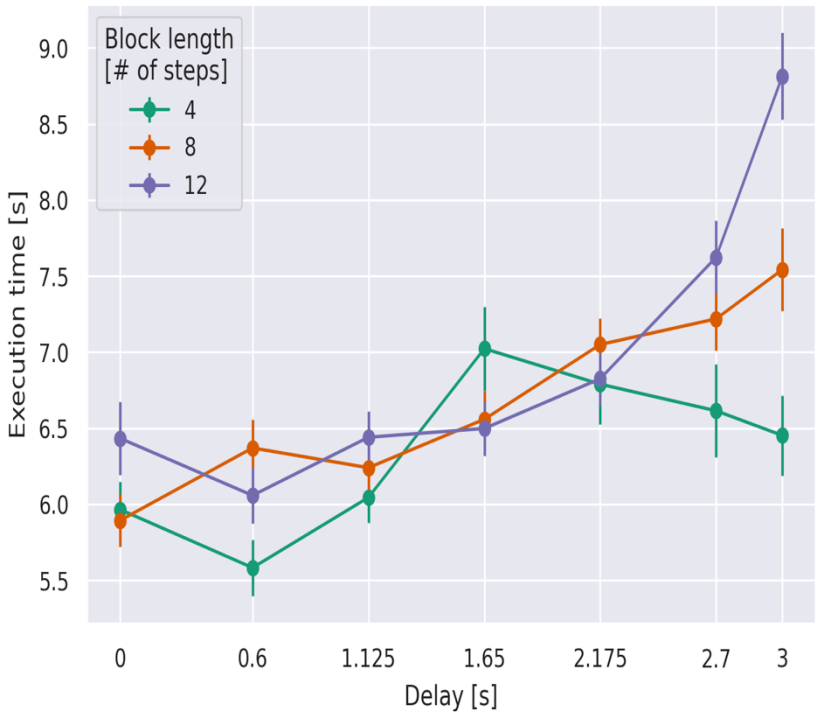


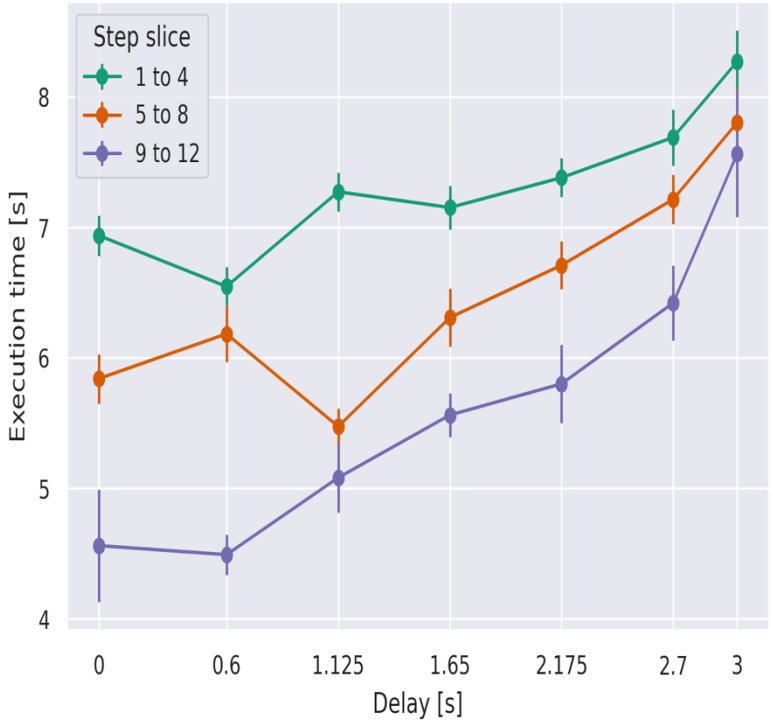
Muñoz et al. “Impact of delayed response on wearable cognitive assistance,” *ArXiv 2020*

Experimental Set-up



Key Results





Normalized execution delay increases with delay, delay also prevents task execution acceleration.

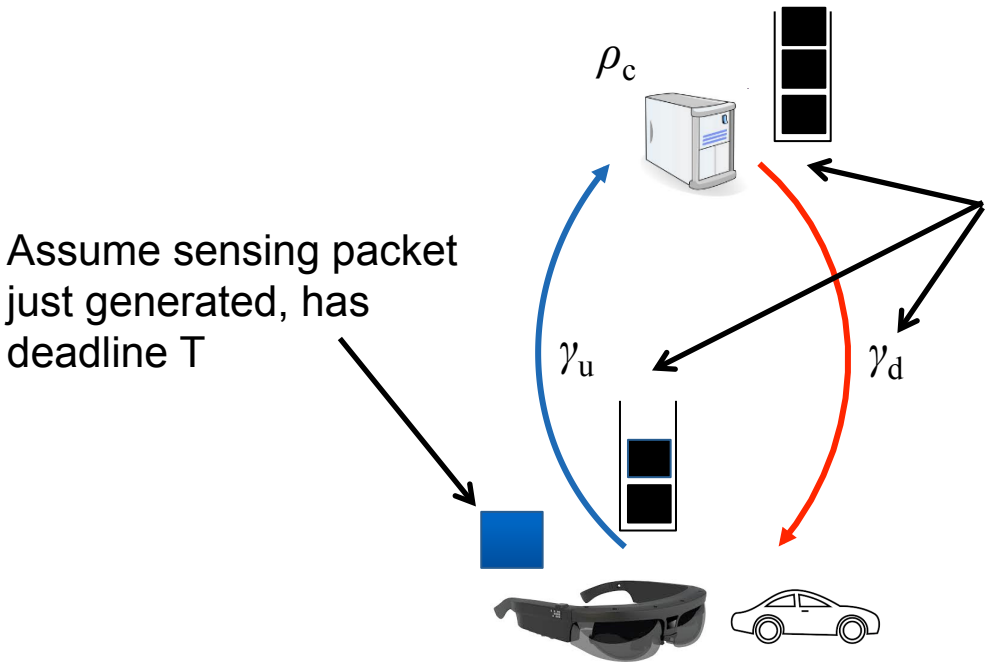
Conclusions & Interpretation

- Increased infrastructure delay leads to:
  - Slowing in execution, prevented from acceleration
  - Execution slow-down lingers even if delay recovers!

➔ **Delay causes disruption of cognitive task automation, while making it hard to re-automate one happened!**

- System consequence: Significantly longer application execution, higher load, higher resource consumption!

How to Prevent Delays?



Assume sensing packet just generated, has deadline T

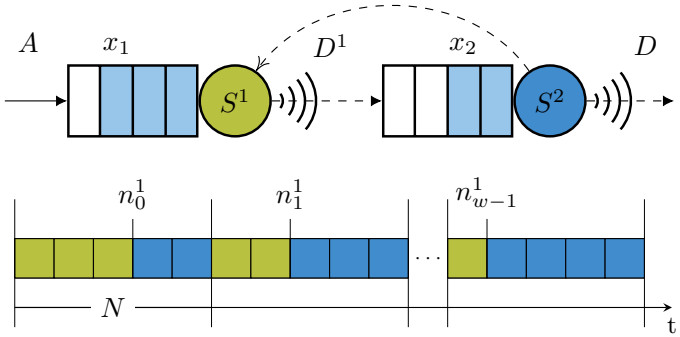
System is subject to particular conditions at packet generation time.

How to schedule the system to minimize likelihood of violating the deadline?

Zoppi et al. “Scheduling of Two-hop lossy wireless networks for time-critical feedback systems,” *ArXiv* 2020.

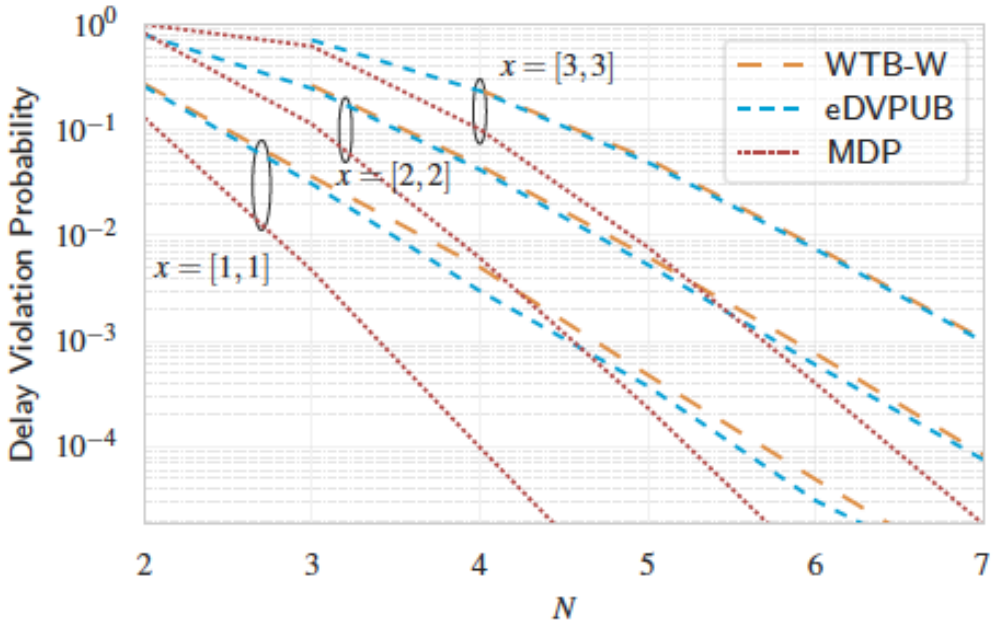
Model and Scheduling Choices

- Two-hop queuing model with joint slots for up-/down-link



- Scheduling variants:
  - Static allocation of slots
  - Allocate slots for the time until deadline, don't change
  - Constantly reallocate slots up until deadline

Results



- Substantial performance differences
- Fully adaptive approaches outperform all other schemes
- Load disbalance, underutilization most important factors

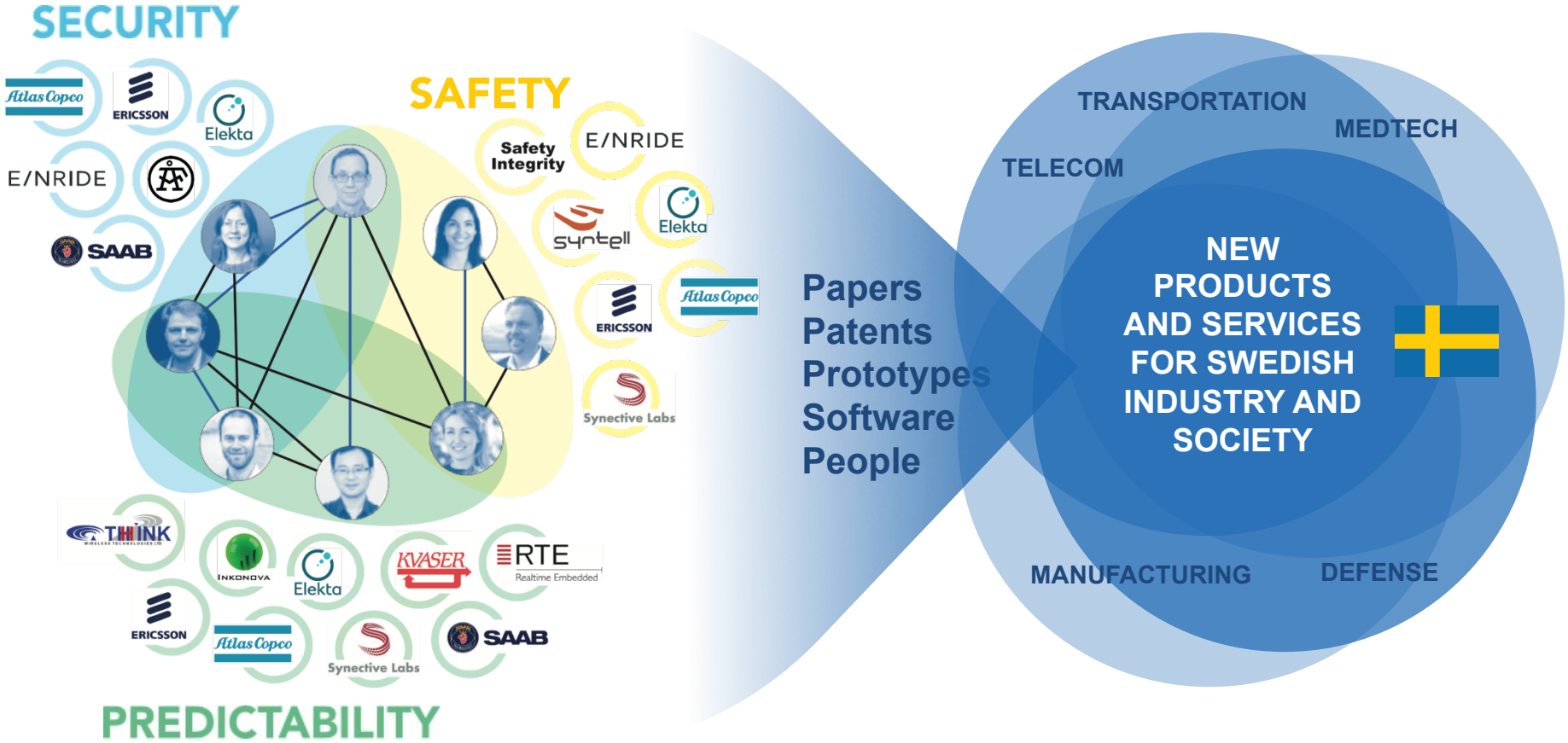
Outline

- Towards an Internet of Reality?
- Initial Challenges and Results
  - EdgeDroid and the quest for latency impact
  - Scheduling for closed-loop
- **Upcoming Challenges & Outlook**

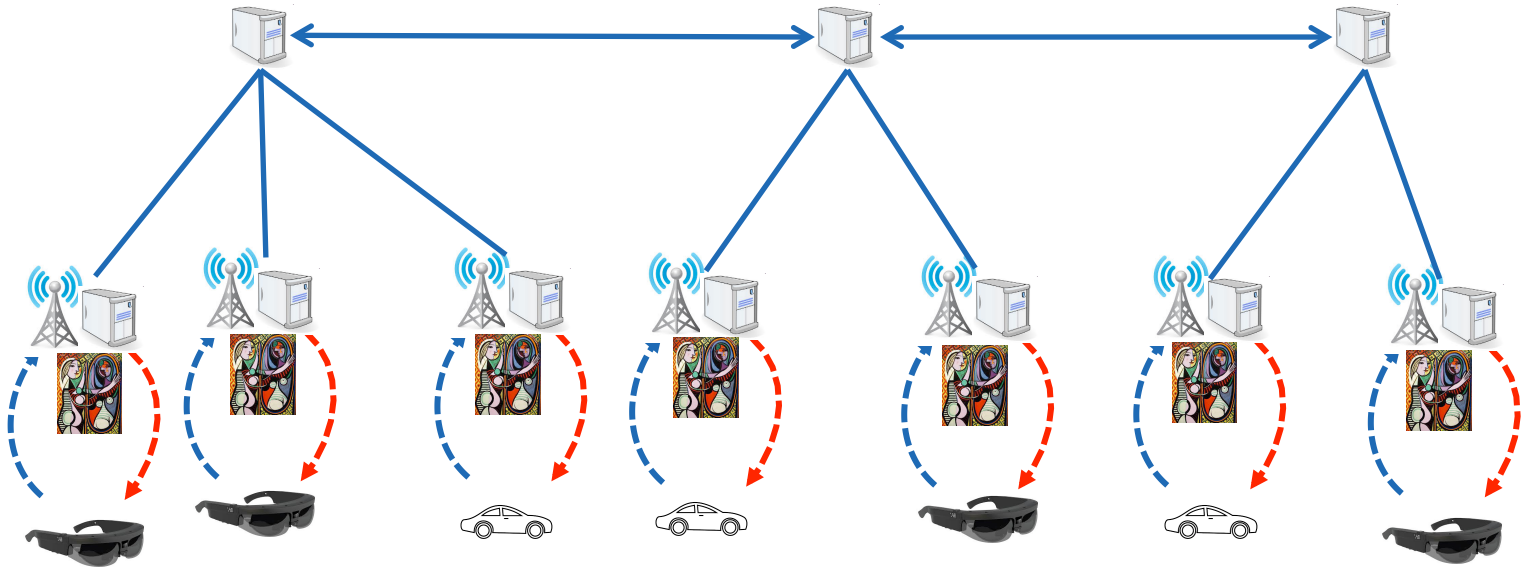
Current Activities

- Automated models of human reaction in HITL
- Latency & quality of control trade-offs: Cleave
  - https://github.com/KTH-EXPECA/CLEAVE
- Optimal sampling & semantics: When to sense reality?
- Predicting loop end-to-end latencies for real systems

VINNOVA Competence Center TECoSA



What to do with the Representations?



- Build a fabric that tracks the representations/ part of them?
- How to universally represent reality?
- Human footprint in such a fabric?

www.digitalfutures.kth.se

Summary & Conclusions

- Upcoming feedback systems that process reality
  - Powerful application class!
  - Novel footprints and requirements
  - End-to-end latency is key metric
- Mastering communication & compute interaction is key
  - Severe consequences if not
  - Still, complex scheduling task, mostly still open
- Towards an Internet of Reality? More research needed …