
Contribution ID: **72** Type: **not specified**

# In network ML: Inference at the Speed of Data

*Wednesday 27 September 2023 09:45 (45 minutes)*

How fast should your machine learning be? ideally, as fast as you can stream data to it.
In this presentation I will discuss the role of computing infrastructure in machine learning, and argue that to face the growing volume of data and support latency constraints, the best place for inference is within the network. I will introduce in-network machine learning, the offloading of machine learning models to run within programmable network devices, and explain the technology and methodologies that enable innovation in the field, as well as existing tools. Finally, I will explore the use of in-network machine learning for a range of applications, ranging from security and finance to edge computing and smart environments.

**Presenter:** ZILBERMAN, Noa

**Session Classification:** Invited Talks

**Track Classification:** Invited Talks