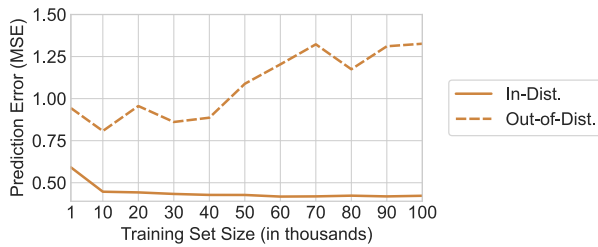


Observation

ML models in networking **fail to generalize** and naively training on **more data** does not help.

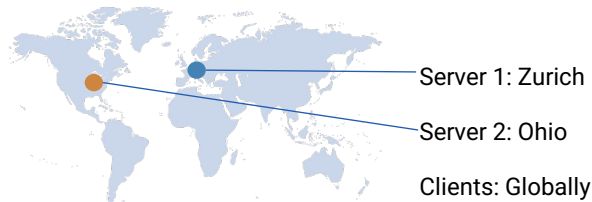


- **Diminishing returns** in In-Dist. performance
- **Decrease** in Out-of-Dist. performance
- ML models require **representative data**

Video-chunk Transmission Time Prediction

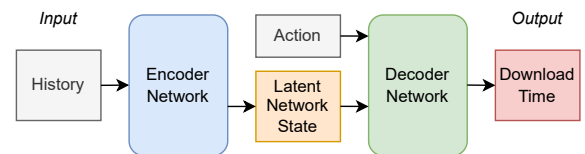
Strategic Data Collection

We collected more **representative** datasets by streaming to **real clients globally**.



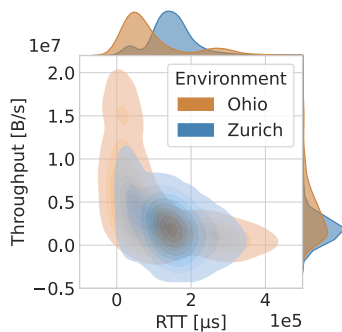
Model Architecture

We used an **Encoder-Decoder** architecture for *Transmission Time Prediction (TTP)*.

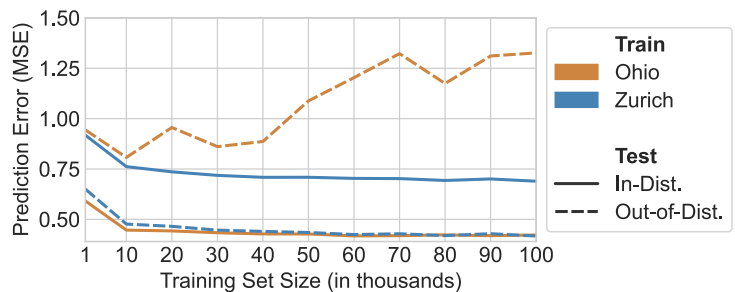


Preliminary Results

Datasets that broadly cover **higher RTT** and **lower throughput** areas improve **generalization**.



Kernel Density Estimates (KDE)



Video-chunk Transmission Time Prediction

Future Work

Where are the **limits**?

- Do our findings hold at **more extreme RTT** and **throughput** values?
- Do RTT and throughput **remain good proxy metrics** for other **prediction tasks**, **environments** and **architectures**?

How can we deepen our **insights**?

- Can we use our findings to generate **representative synthetic datasets** with equal **generalization** performance?
- Which **other metrics** can we explore as proxies for a dataset?

Extended Abstract:

