Harnessing Deep Learning for Mobile Service Traffic Decomposition to Support Network Slicing

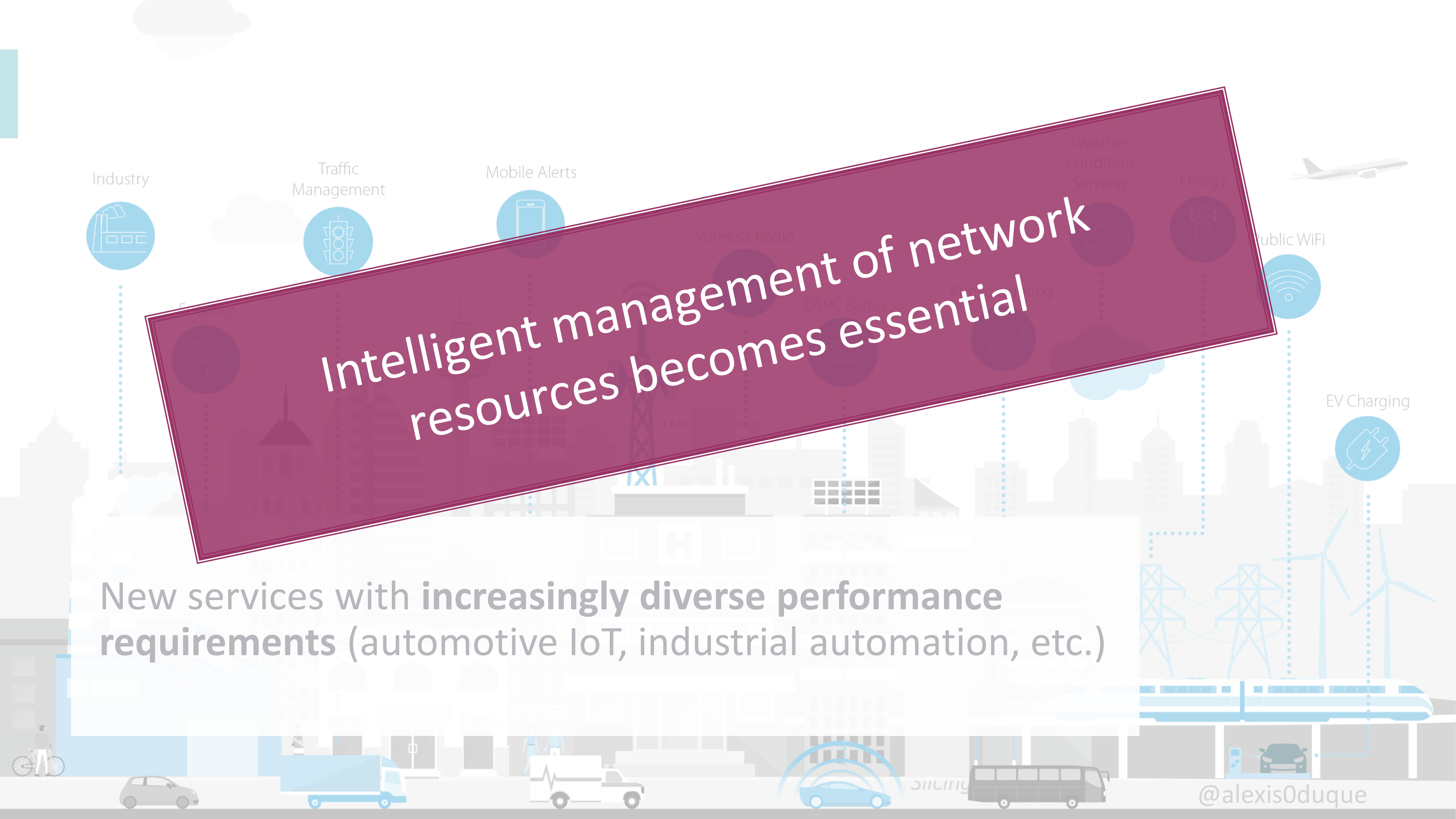
Alexis Duque, work with **Paul Patras**,
Chaoyun Zhang, and Marco Fiore

THE UNIVERSITY *of* EDINBURGH
informatics

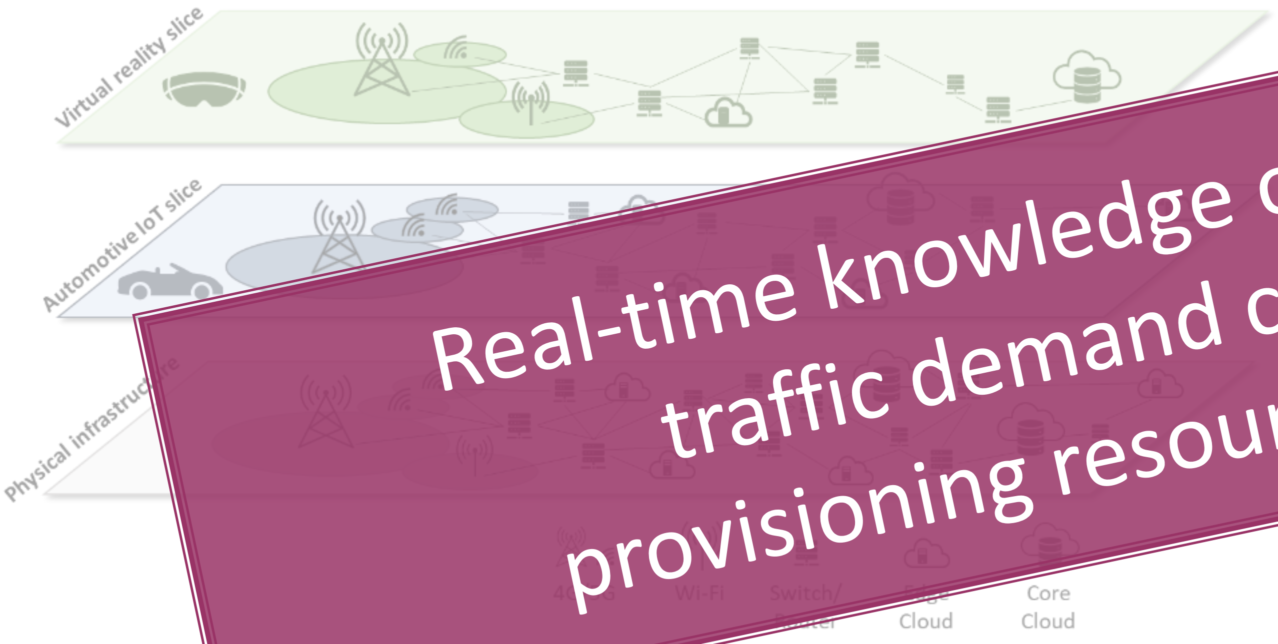
@alexis0duque

Intelligent management of network resources becomes essential

New services with **increasingly diverse performance requirements** (automotive IoT, industrial automation, etc.)



Network slicing:
Key to effectively managing and monetizing 5G



- Logical partitioning of physical infrastructure into virtual networks customized for specific services
- 20-50% Total Cost of Ownership (TCO) savings [Futurithmic/Nokia]

Real-time knowledge of per-service traffic demand critical to provisioning resources to slices

Current approach: DPI-based traffic classification

Image: arubanetworks.com

Hardware-based:
- Expensive (FPGA)
- Not scalable (impossible to update)

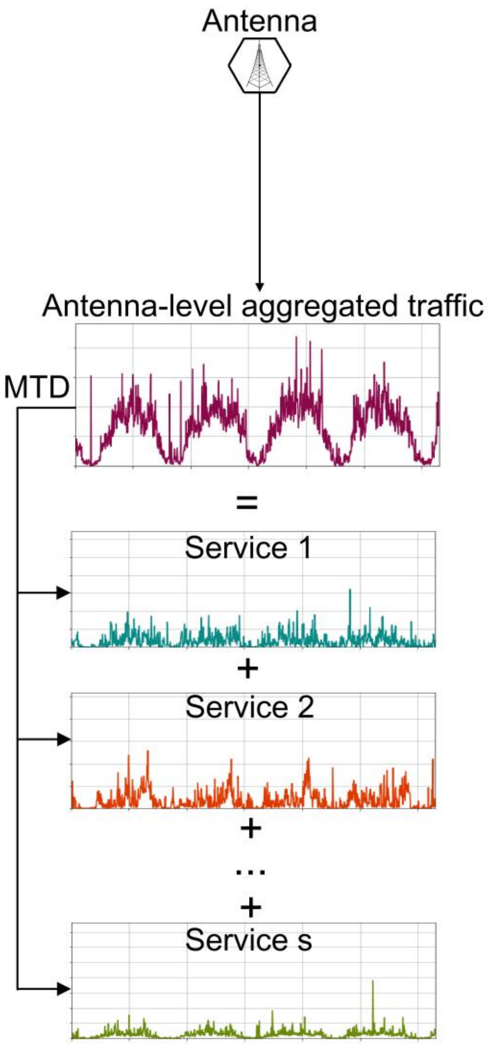
Software-based:
- Slow (packet capture, OS scheduling, buffering, …)
- Prone to packet loss

All:
- Complicated by encryption

Proposed alternative: Mobile Traffic Decomposition

- Breaking down time series of traffic aggregates into separate time series corresponding to individual services.

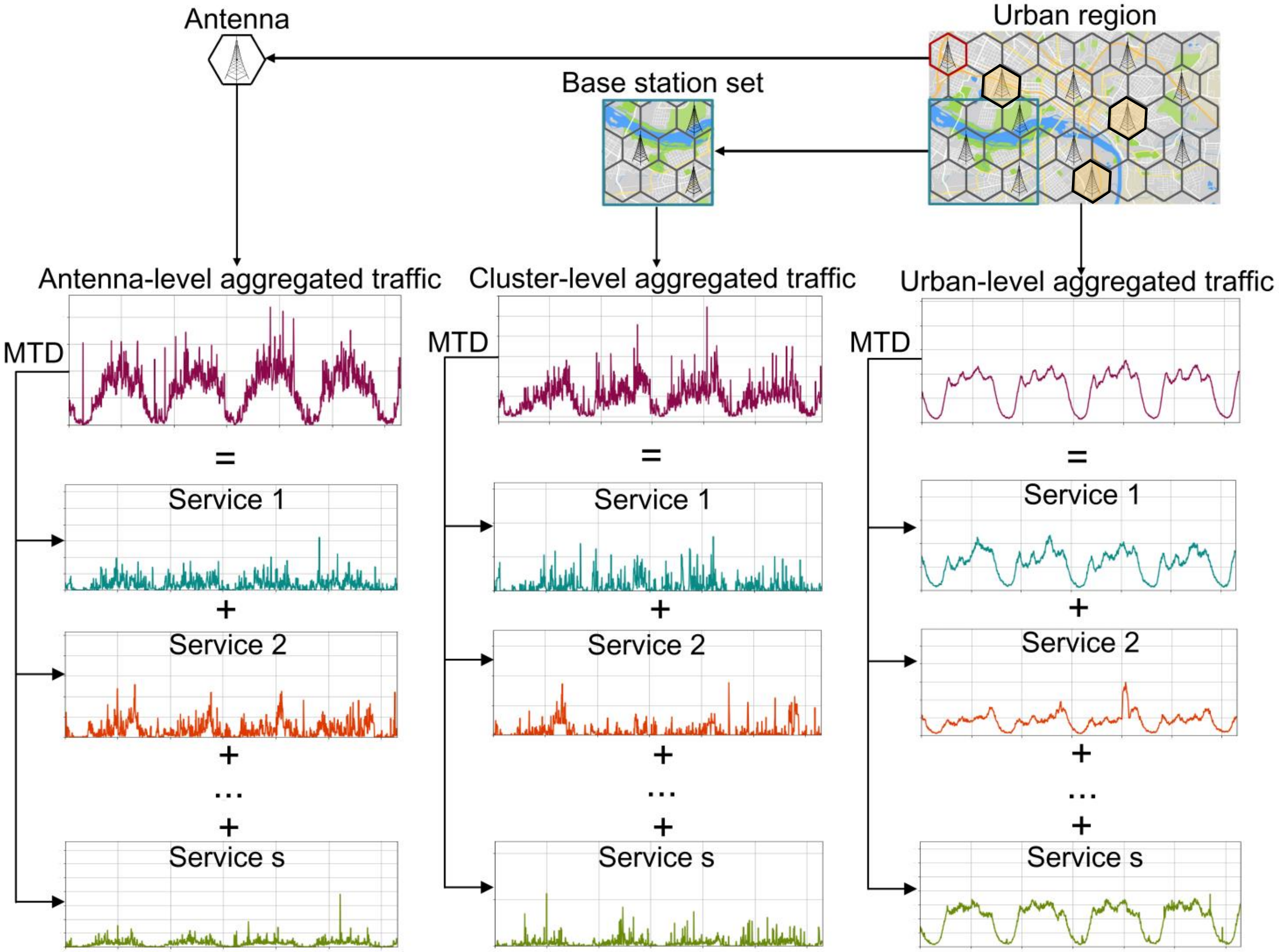
Proposed alternative: Mobile Traffic Decomposition



- Breaking down time series of traffic aggregates into separate time series corresponding to individual services.

- Operating at various levels, as required by different application scenarios.

Proposed alternative: Mobile Traffic Decomposition



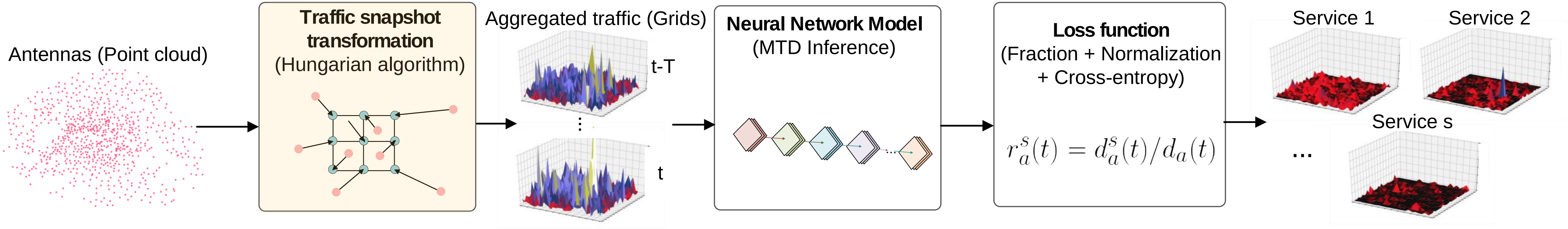
- Breaking down time series of traffic aggregates into separate time series corresponding to individual services.
- Operating at various levels, as required by different application scenarios.
- Exploiting spatiotemporal correlations characteristic to mobile network traffic.

Challenges of decomposition

1) Decomposing a single signal into multiple time series may have multiple solutions
2) Capturing complex spatial and temporal correlations to resolve the ambiguity is not trivial
3) Techniques used in other domains, e.g., factorial hidden Markov models work on single time series

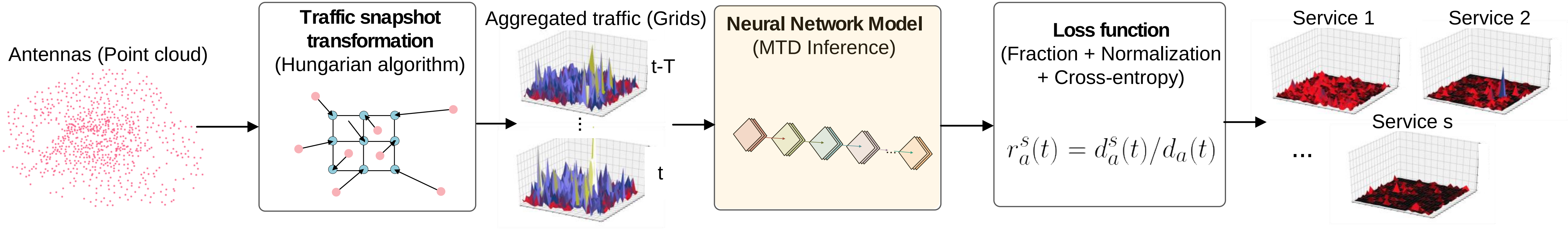
Our goal: decompose *multiple input time series* concurrently at different network locations

Microscope: Dedicated deep learning-based framework for Mobile Traffic Decomposition



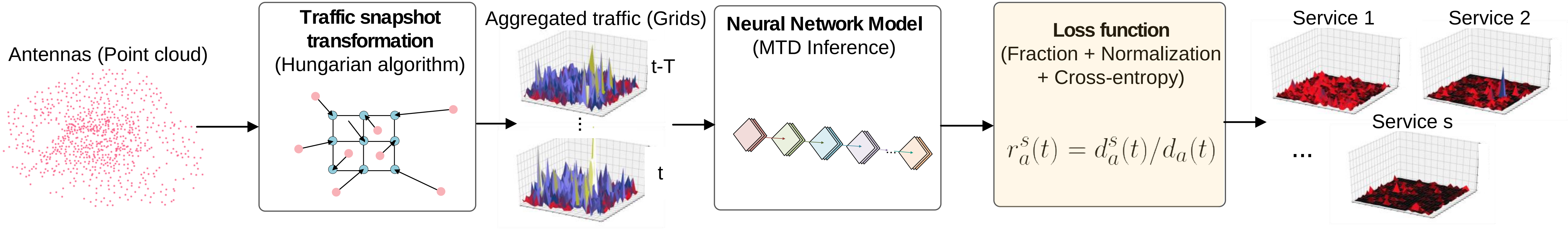
Converts traffic measurements into a format suitable for analysis with minimum loss of geographic information

Microscope: Dedicated deep learning-based framework for Mobile Traffic Decomposition



Deep neural model that learns abstract spatiotemporal correlations of mobile traffic to solve the MTD problem.

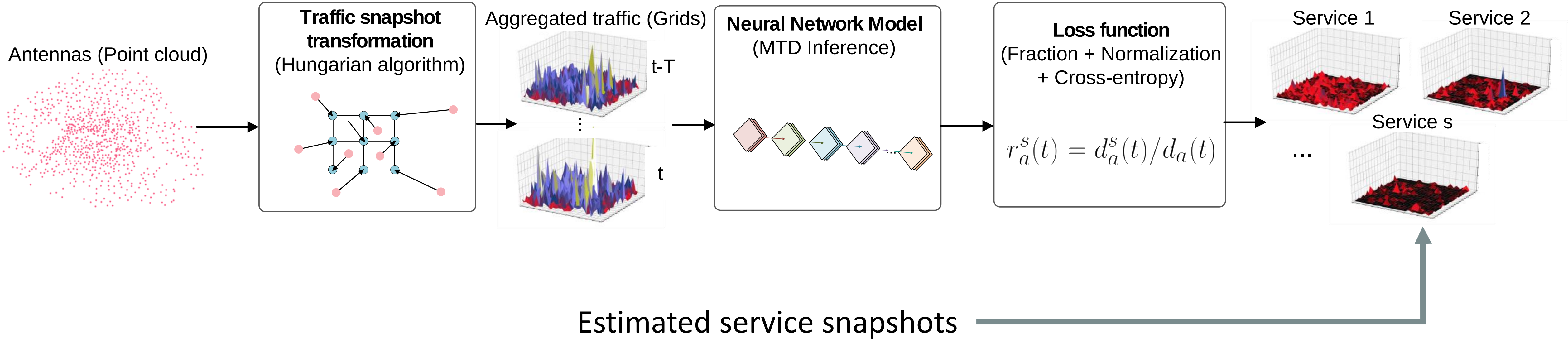
Microscope: Dedicated deep learning-based framework for Mobile Traffic Decomposition



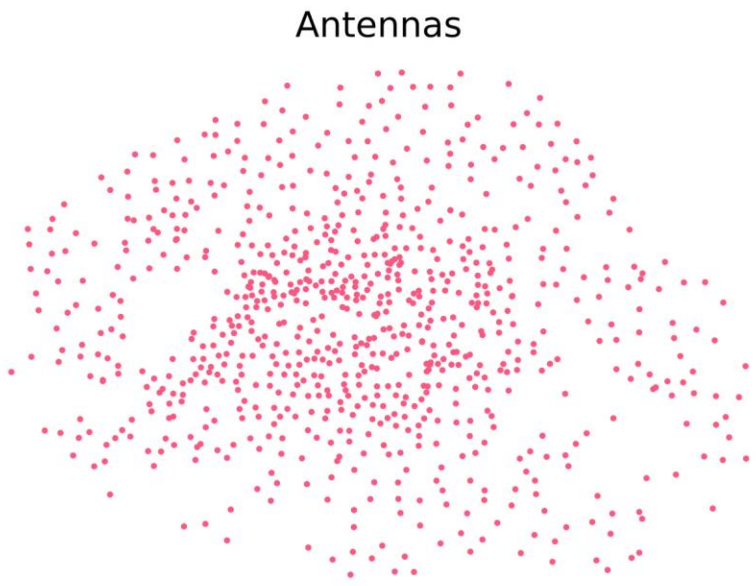
Loss function to drive the training process

Output normalization

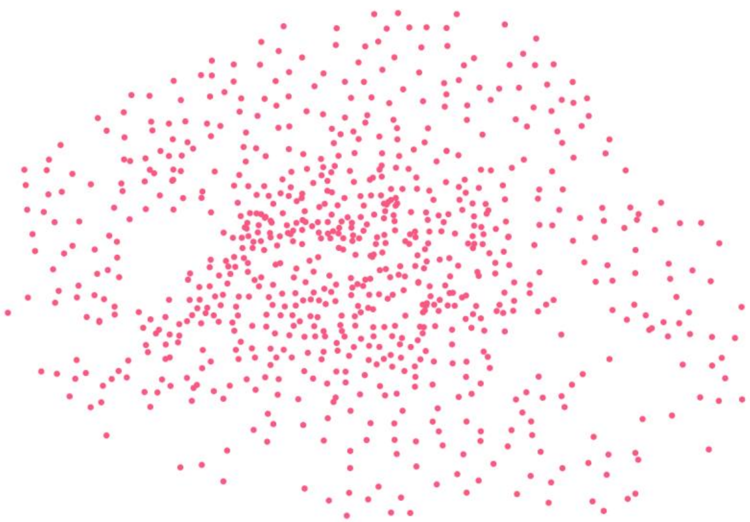
Microscope: Dedicated deep learning-based framework for Mobile Traffic Decomposition



Point-cloud to grid transformation

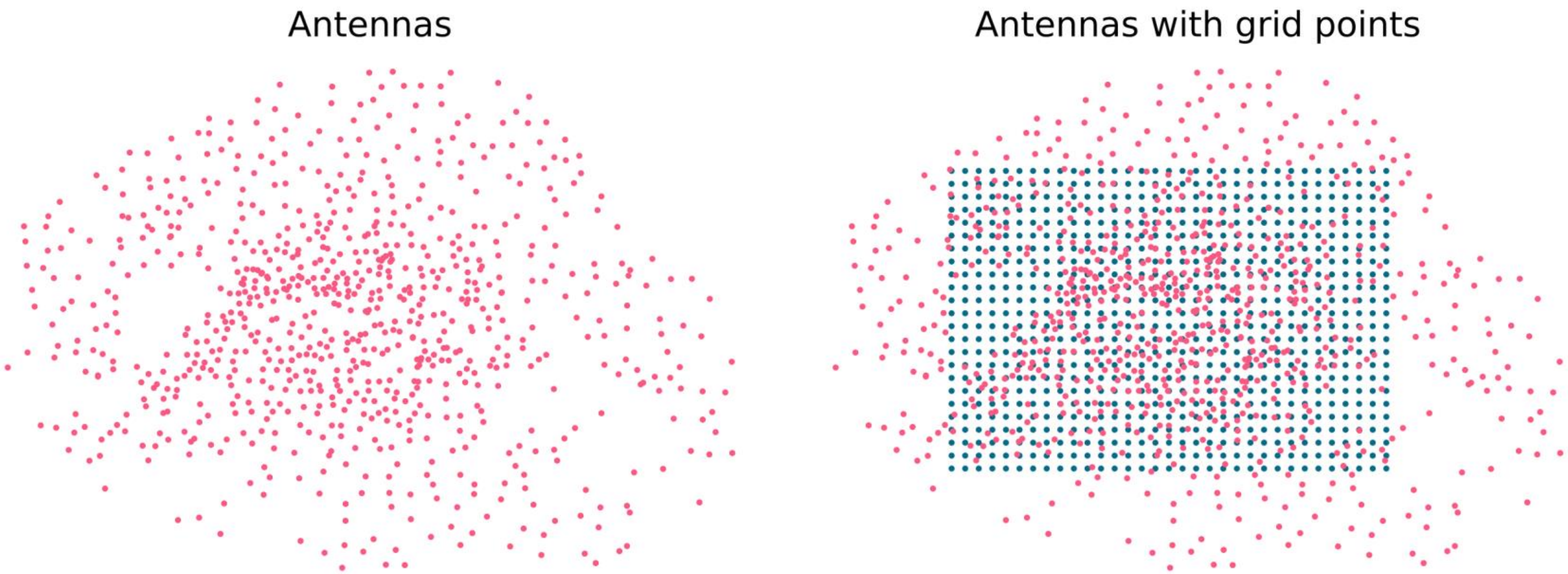


Point-cloud to grid transformation



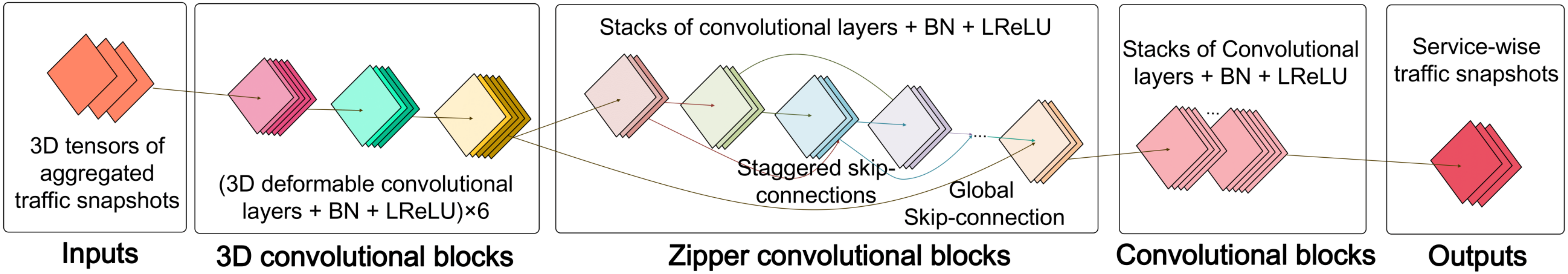
- Construct regular grid with same number of points as the number of antennas (suitable for convolution)

Point-cloud to grid transformation



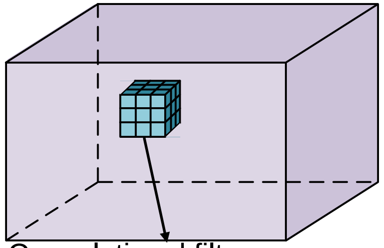
- Construct regular grid with same number of points as the number of antennas (suitable for convolution)
- Perform one-to-one association that minimizes displacement of original locations (preserve spatial correlations in traffic that can be exploited)
  → Hungarian algorithm (polynomial time)

3D-Deformable Convolutional Neural Net (3D DefCNN)



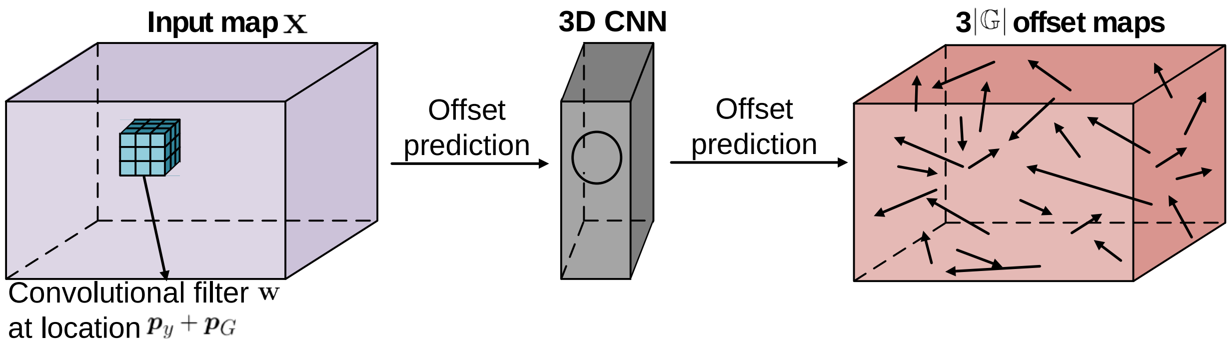
- New class of convolutional NNs specifically designed for decomposition
- Input: sequences of *T* aggregate traffic snapshots
- Output: traffic snapshots for individual mobile services

3D-Deformable CNN



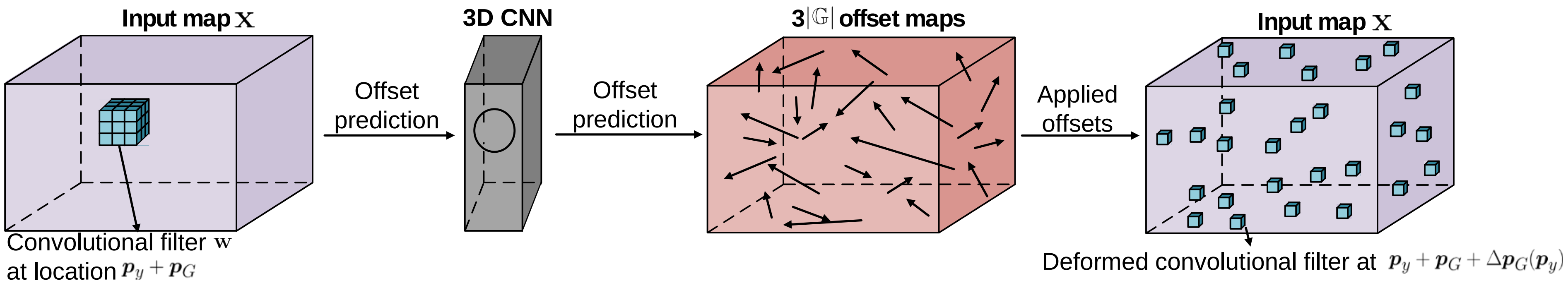
- Start from a compact 3D filter (cube) scanning the input

3D-Deformable CNN



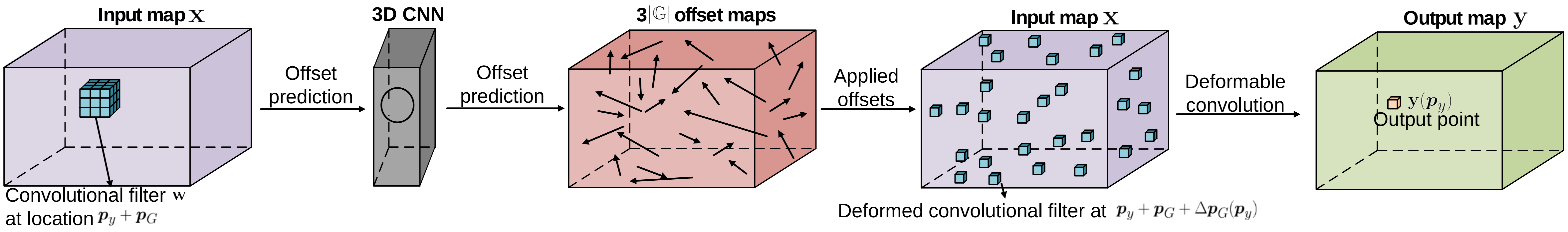
- Start from a compact 3D filter (cube) scanning the input
- Learn offsets to be applied to filter, defining extent of spatiotemporal correlation between different locations in the input

3D-Deformable CNN



- Start from a compact 3D filter (cube) scanning the input
- Learn offsets to be applied to filter, defining extent of spatiotemporal correlation between different locations in the input
- Obtain *deformed* convolution filter scanning not necessarily adjacent locations

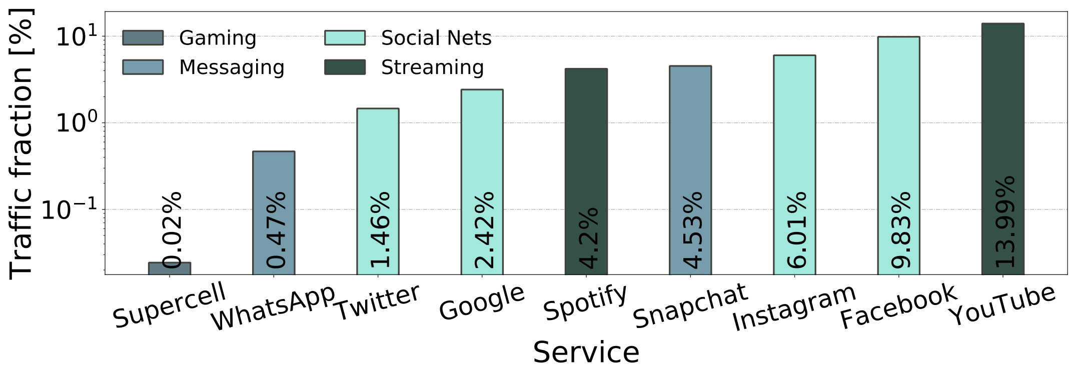
3D-Deformable CNN



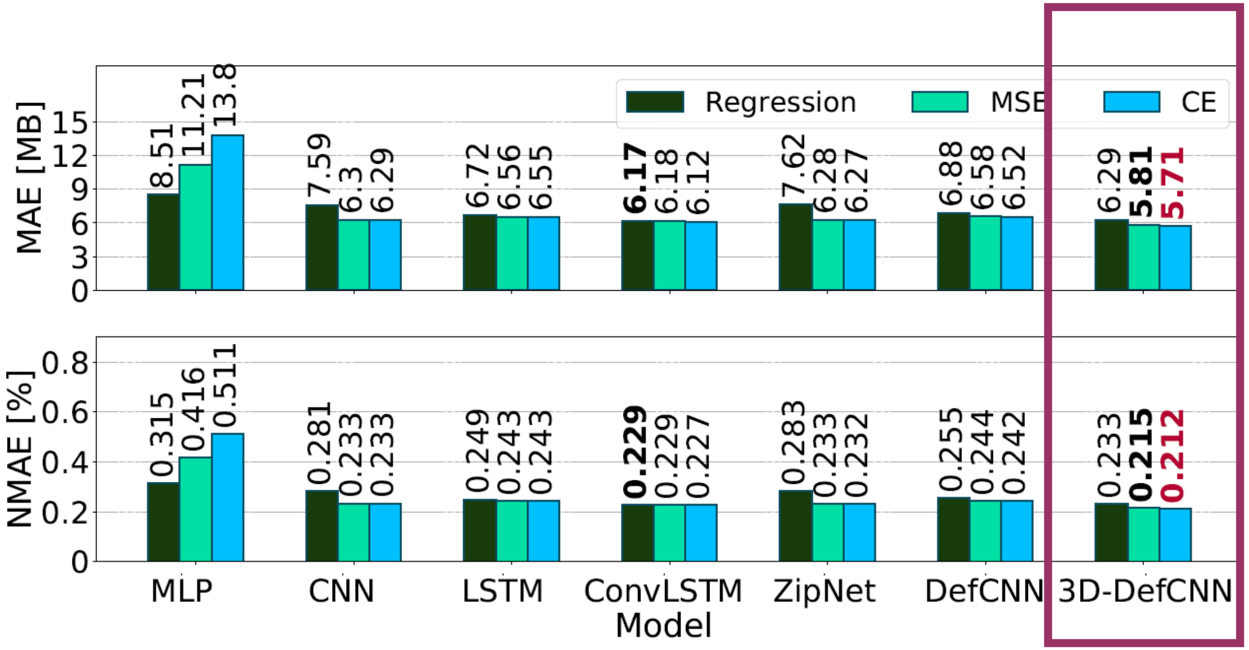
- Start from a compact 3D filter (cube) scanning the input
- Learn offsets to be applied to filter, defining extent of spatiotemporal correlation between different locations in the input
- Obtain *deformed* convolution filter scanning not necessarily adjacent locations
- Output: abstract map corresponding to different services

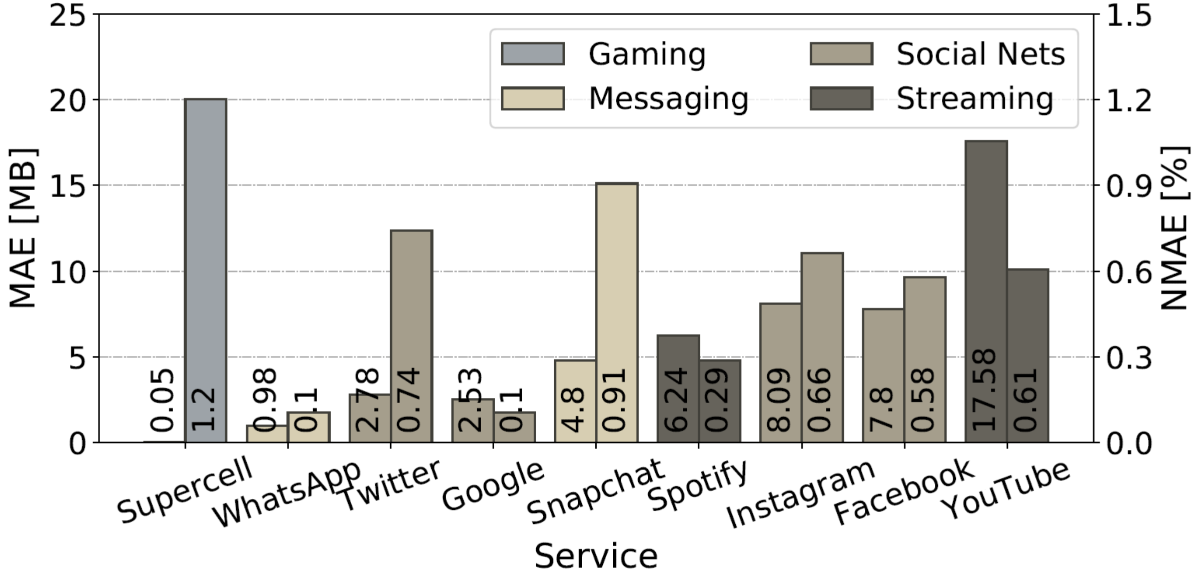
Experiments

- Implemented Microscope using TensorFlow and TensorLayer
- Trained using different loss functions and Adam optimizer
- HPC cluster with Nvidia Tesla K40M GPUs
- Data collected in a large city over 85 days; focused on 9 most demanding services
- Performance evaluation at different network levels (RAN, MEC facility, core datacenter)



Performance evaluation

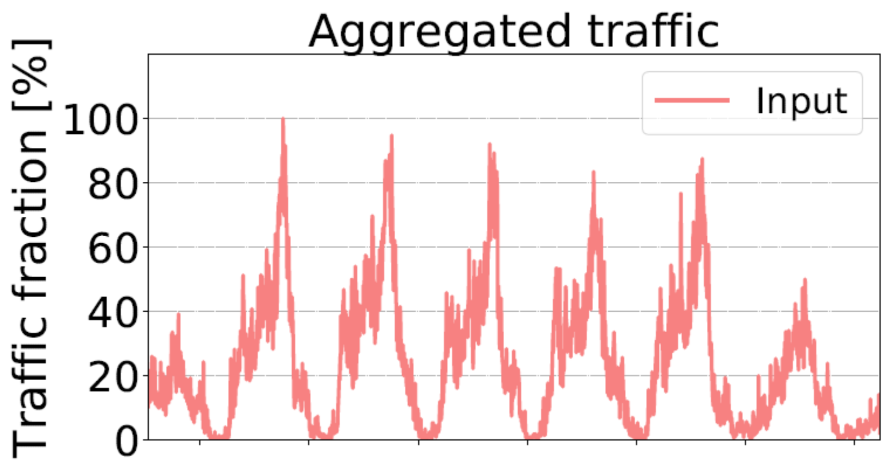




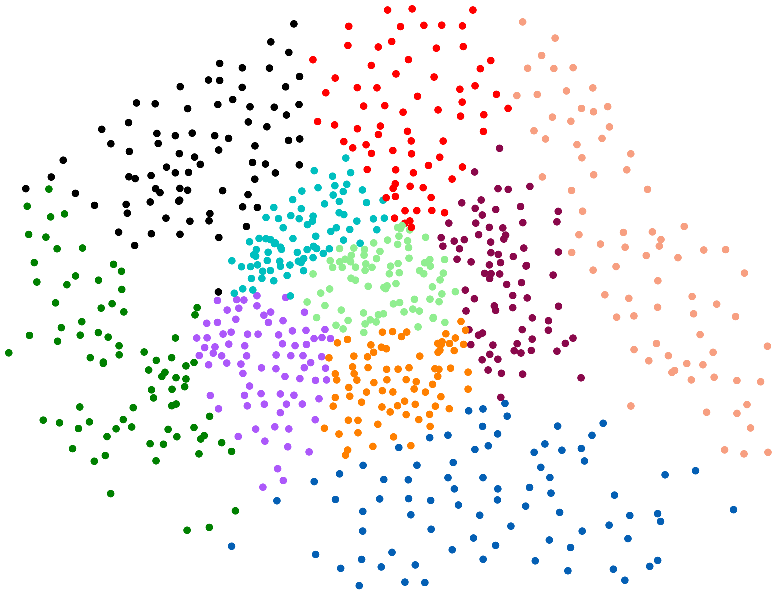
3D-DefCNN + CE performs the best

Achieves NMAE below 1.2%

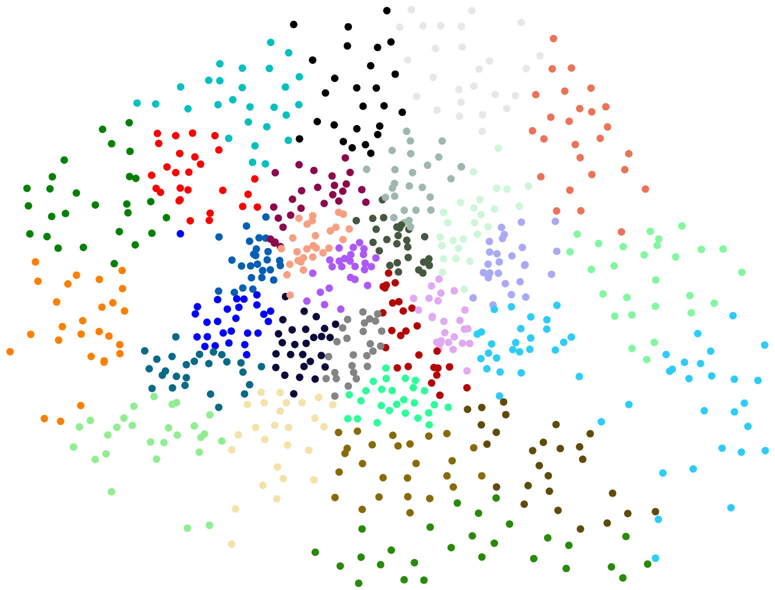
Service-level performance



Traffic decomposition at decenter level



Assignment to 10 core datacenters



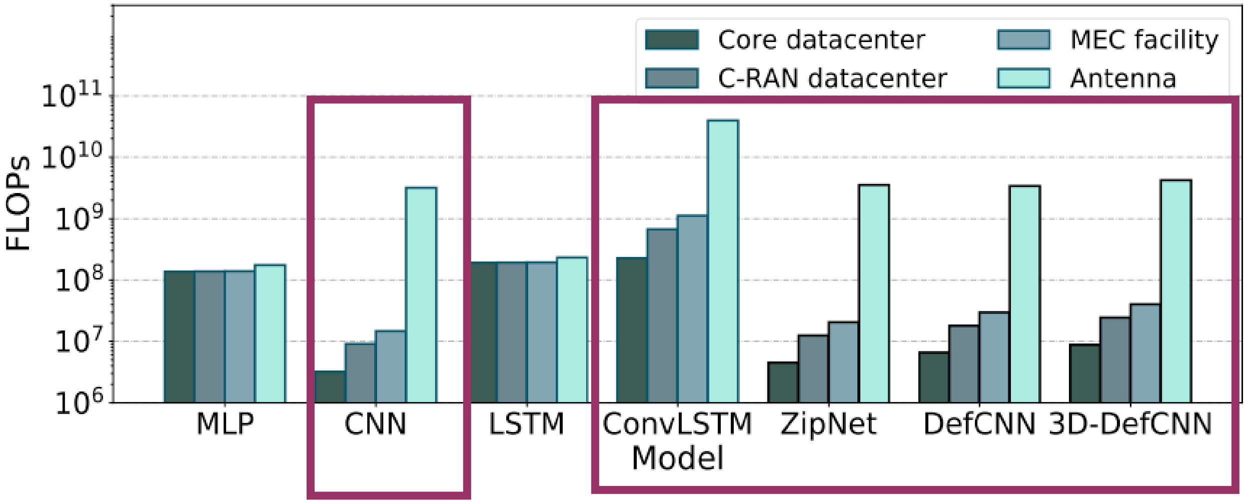
Assignment to 30 C-RAN datacenters

- Antenna clusters serving comparable traffic loads, while minimizing latency (i.e the distance)
- Obtained via Karlsruhe Fast Flow Partitioning (KaFFPa) heuristic

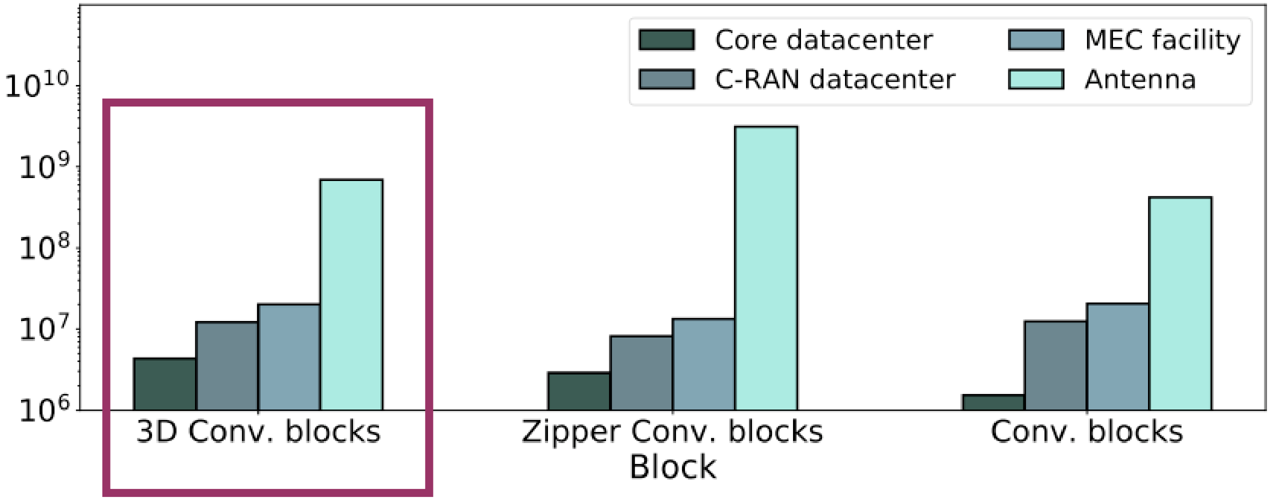
Performance with different resource orchestration intervals

- LSTM sufficient for estimation per-service traffic consumption at core datacenter level, irrespective of temporal granularity
- 3D-DefCNN works best in allocating resources at C-RAN datacenter and MEC facility level
- Infrequent resource management (e.g., every 1h) based on decomposed traffic can be served with low-complexity LSTMs

Complexity analysis



Complexity (measured in FLOPs) of all evaluated neural network models across different network levels.



Complexity (measured in FLOPs) of each block in the 3D-DefCNN model across different network levels.

- Computational requirements of CNN-based models surpass those of LSTM only for antenna-level MTD
- Marginal cost introduced by deformable convolution operation

Evaluation of SLA violation or overprovisioning due to MTD errors

- **At C-RAN and core datacenters**, Microscope carries percent costs in the range from 8% to 58%, computed with respect to the true demand
- **At antenna level**, just 6 Mbps of additional throughput are needed per antenna leading to 7.5% additional CPU time vs where perfect knowledge of service traffic is available

MTD can be a viable low-cost approach to service-level demand estimation in practical NSaaS management

Microscope in a nutshell

Patent Pending

Microscope – dedicated framework for Mobile Traffic Decomposition, supporting resource allocation to network slices

Can incorporate different neural models and adapt to different management location or timescale

Experimental results with metropolitan-scale network measurements shows that it infers per-service traffic demands with 99% accuracy

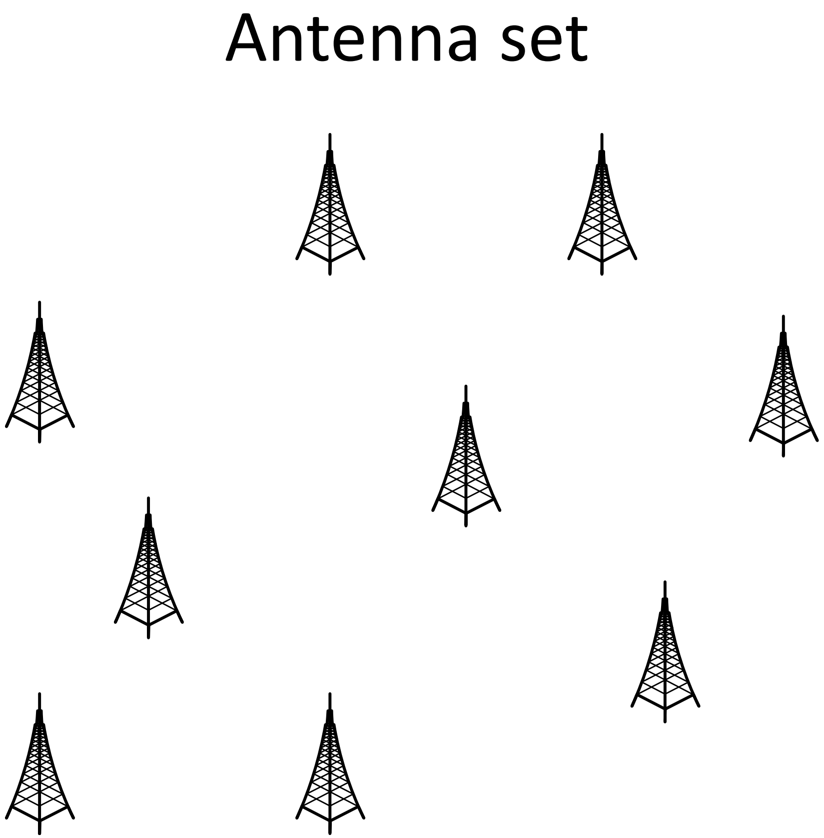
Solves computationally intensive traffic analytics essential to agile resource provisioning in 5G

Learning on Point Cloud?

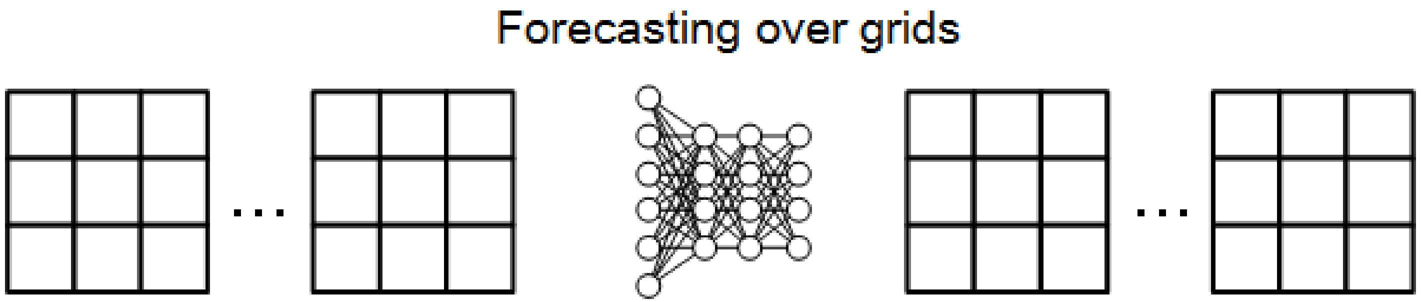
1) Mobile traffic analysis needs the spatio-temporal correlations and the configuration of the data points over time to be preserved
2) Existing spatio-temporal inference models require **grid-structural data**
3) Data preprocessing is so required, like the point-cloud to grid transformation we have proposed before.

Our goal: eliminates the need for the data preprocessing without losing spatio-temporal correlations

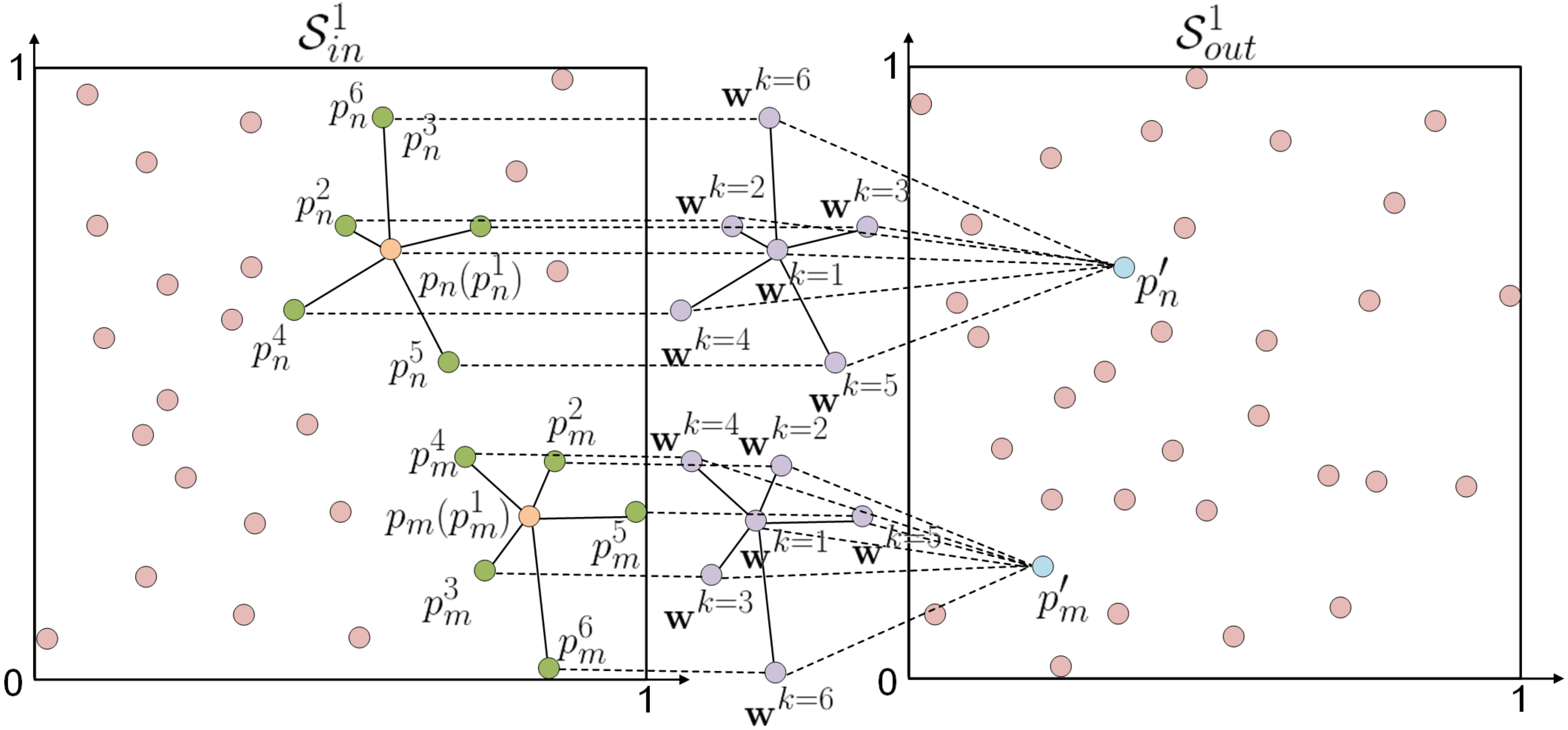
Forecasting on Scattered Antennas



Forecasting on Scattered Antennas

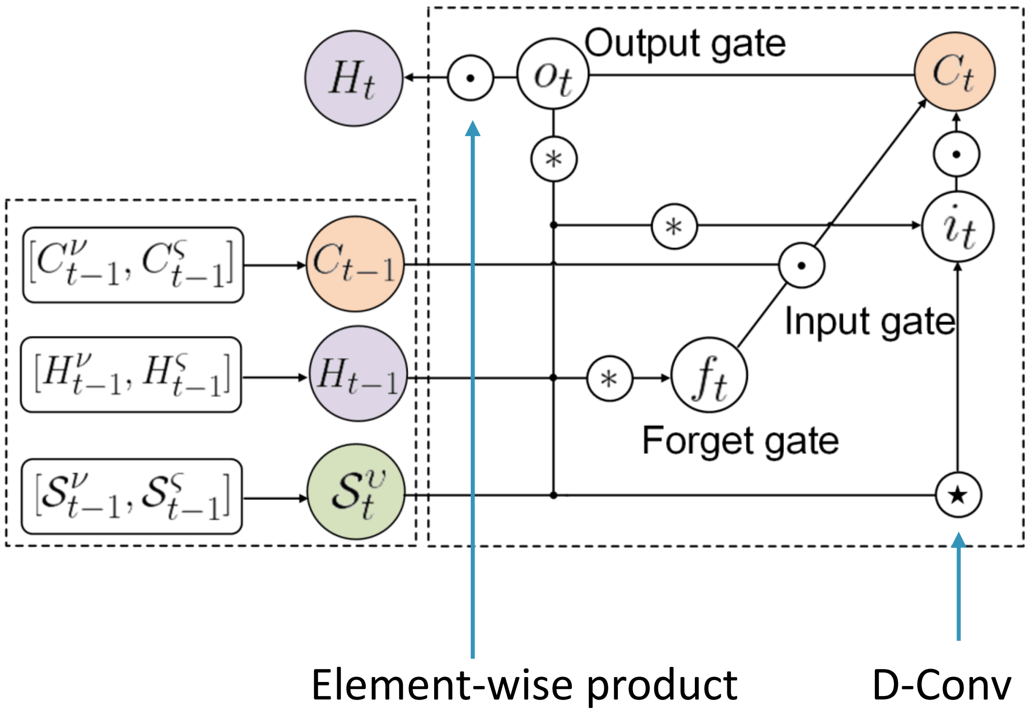


Dynamic Point Cloud Convolution (D-Conv)



Convolutional Point Cloud LSTM (CloudLSTM)

CloudLSTM cell



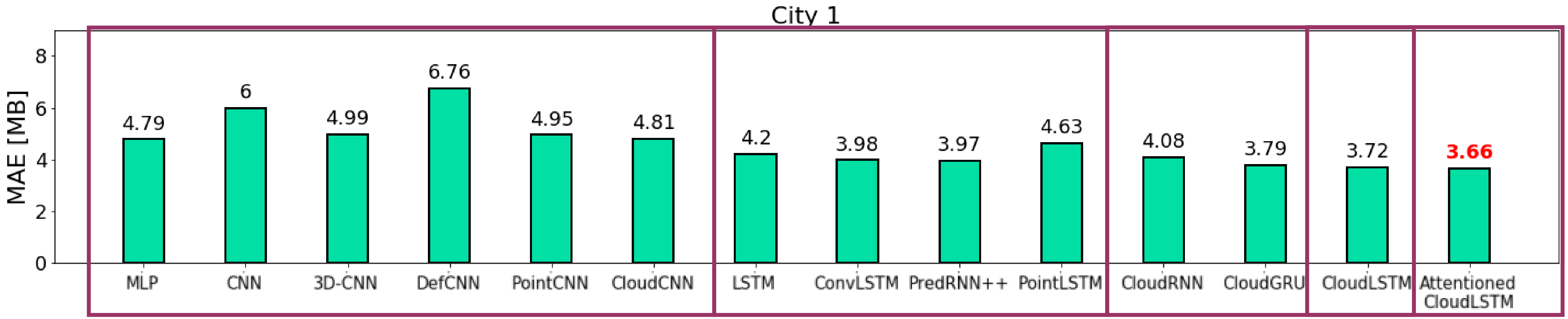
Dataset

- A large-scale mobile traffic dataset collected by a major operator in **two large European metropolitan areas** for **approximately 3 months.**

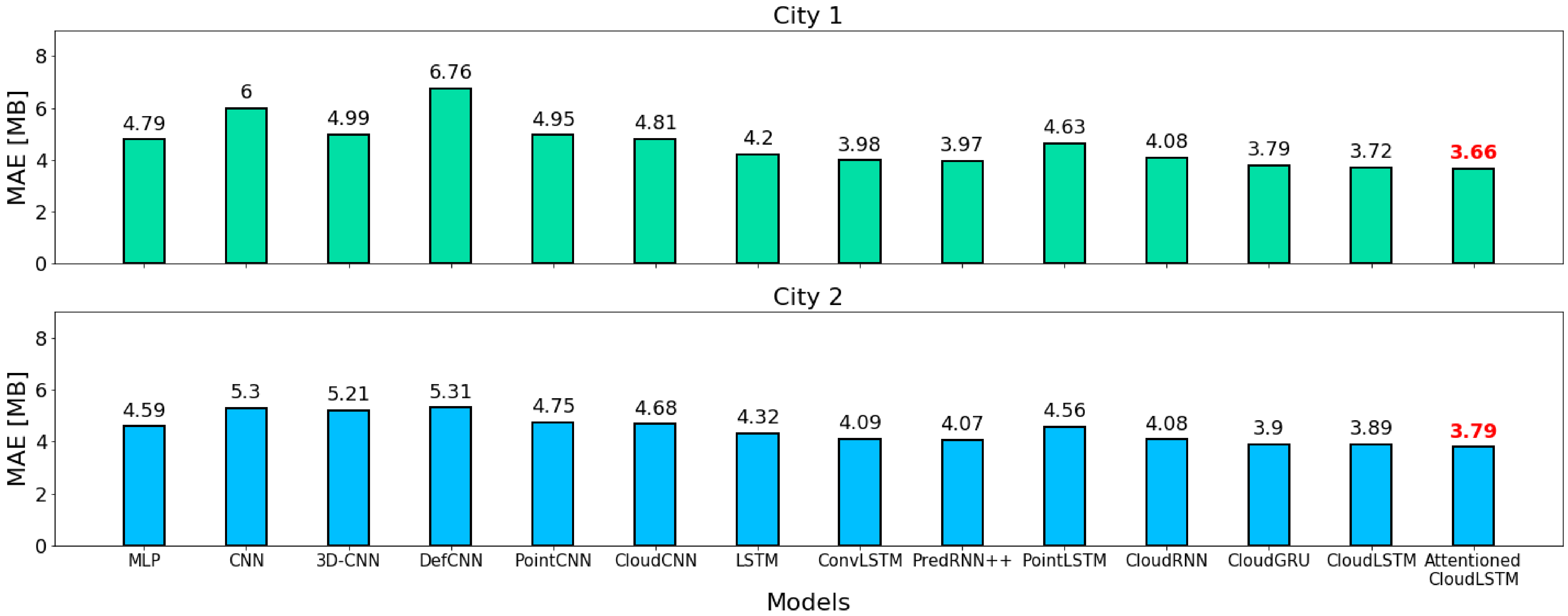
- Collection of **traffic measurement for 36 distinct services** (including YouTube, Netflix, Snapchat, Instagram, Facebook, Pokemon Go, Spotify, etc.)

- **Input 6 snapshots (30 min),** and **forecast the following 6 snapshots** (30 min) for all mobile services.

Performance Evaluation

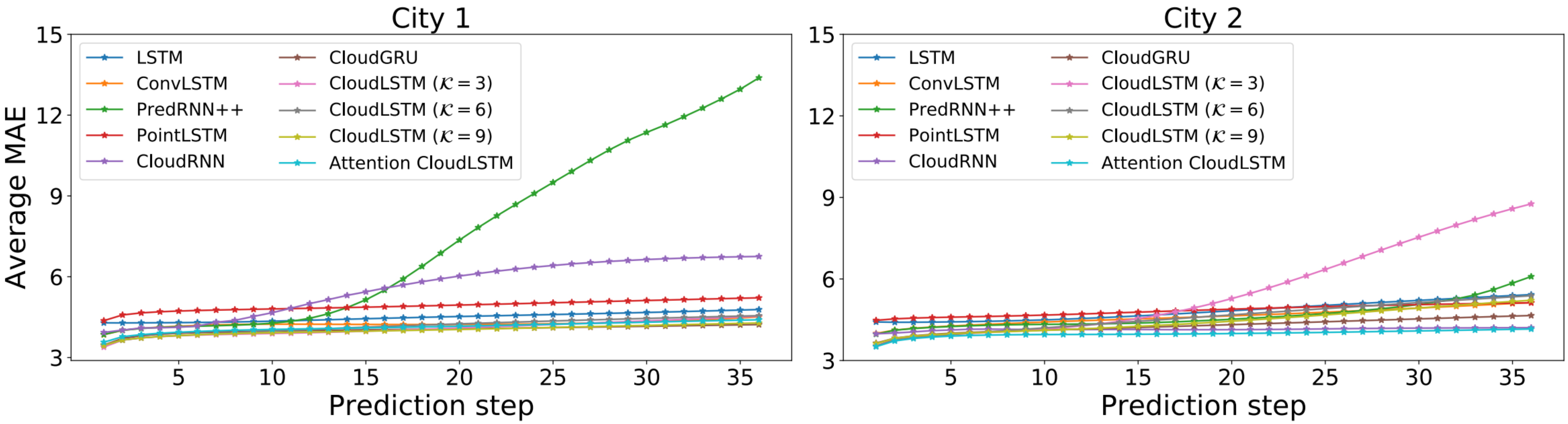


Performance Evaluation



Attentioned CloudLSTM achieves up to 45.9% lower prediction error

Performance Evaluation



MAE evolution wrt. prediction horizon achieved by RNN-based models. Input length is unchanged.

- These models are reliable in terms of long-term forecasting
- Low K may lead to poorer long term performance for CloudLSTM

D-Conv and Cloud LSTM in a nutshell

D-Conv - operator, which performs convolution over point-clouds to learn spatial features while maintaining permutation invariance

Can be easily combined with various RNN models (i.e., RNN, GRU, and LSTM), Seq2seq learning, and attention mechanisms

CloudLSTM - a dedicated neural model for spatiotemporal forecasting tailored to point-cloud data streams built upon D-Conv operator

Experimental results with metropolitan-scale network measurements show CloudLSTM outperforms state of the art models for mobile traffic forecasting

Want to know more?

- **Microscope: Mobile Service Traffic Decomposition for Network Slicing as a Service**, C. Zhang, M. Fiore, C. Ziemlicki, and P. Patras. ACM MobiCom 2020. https://dl.acm.org/doi/10.1145/3372224.3419195

- **CloudLSTM: A Recurrent Neural Model for Spatiotemporal Point-cloud Stream Forecasting,** C. Zhang, M. Fiore, I. Murray, and P. Patras. https://arxiv.org/abs/1907.12410

Mobile Intelligence Lab
https://mi.inf.ed.ac.uk/

Get in touch

- Microscope now a patent pending technology
- At the core of Net AI, a University of Edinburgh spinout
- Inquiries about partnerships and investments welcome at contact@netai.tech

net AI THE UNIVERSITY of EDINBURGH informatics