Synthetic Mobile Network Data Generation

Mahesh K. Marina

Networked Systems (NetSys) Group

Joint work with Kai Xu, Rajkarn Singh, Marco Fiore, Howard Benn, et al.

THE UNIVERSITY of EDINBURGH
informatics

icsa | Institute for Computing Systems Architecture

Recent Work

1. **Mobile networking systems design**
   - Nervion cloud-native RAN emulator [MobiCom'21] ➔ **Jon's talk**
   - WhiteHaul white space spectrum aggregation system [MobiSys'20]
2. **Data-driven mobile network automation and optimization**
   - Anomaly detection and troubleshooting
     - Automated jammer detection with JADE [INFOCOM'22] ➔ **Caner's talk**
     - Network slice performance monitoring [TNSM'20]
   - Energy efficient virtualized RANs [INFOCOM'21]
   - Spectrum sharing
     - Learning driven spectrum sharing in neutral-host small cells [JSAC'19]
     - Communication-free inter-operator interference management [TCCN'19]
3. **Mobile network data generation and analysis**
   - City-scale traffic snapshot generation with CartaGenie [PerCom'22]
   - City-scale spatiotemporal traffic synthesis with SpectraGAN [CoNEXT'21]
   - National scale mobile service usage diversity analysis [WWW'19]

Barriers to Accessing Real-World Network Data

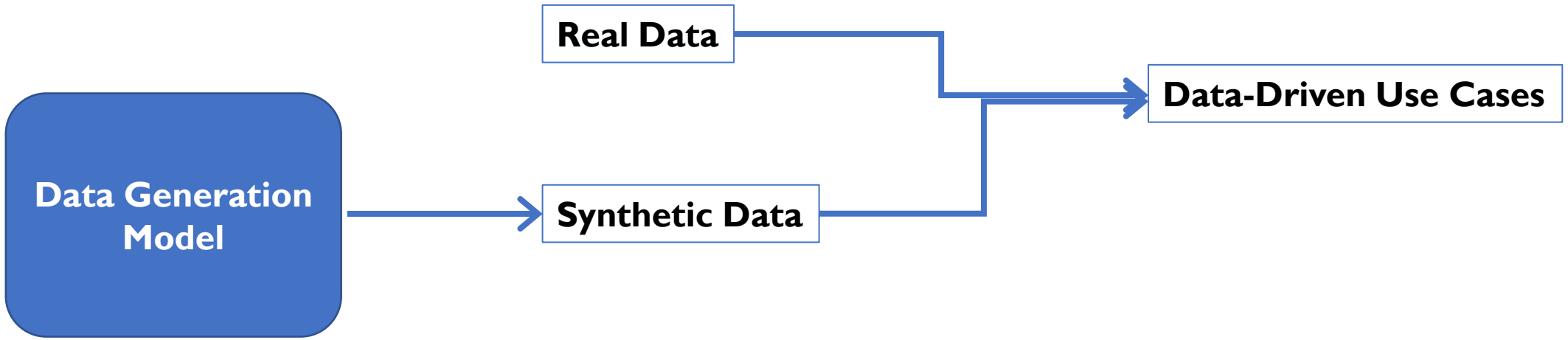
- Operators / service provider concerns about:
  - revealing commercially sensitive info
  - compromising subscriber privacy

- Result: Only few have access to data (through restrictive NDAs) ➔ Limits innovation and reproducibility

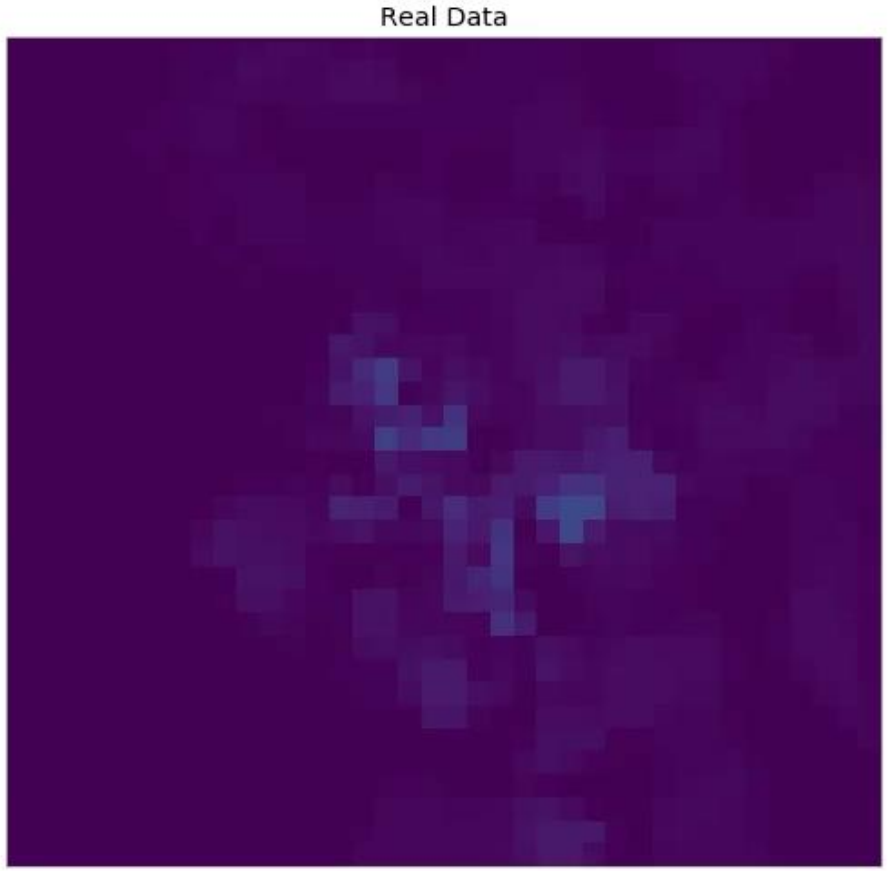
- Measurement data collection is costly and time-consuming

Synthetic Data Generation as a Remedy

- Leverage access to limited amount of real data for designing models that can then generate unlimited amount of "like-real" data

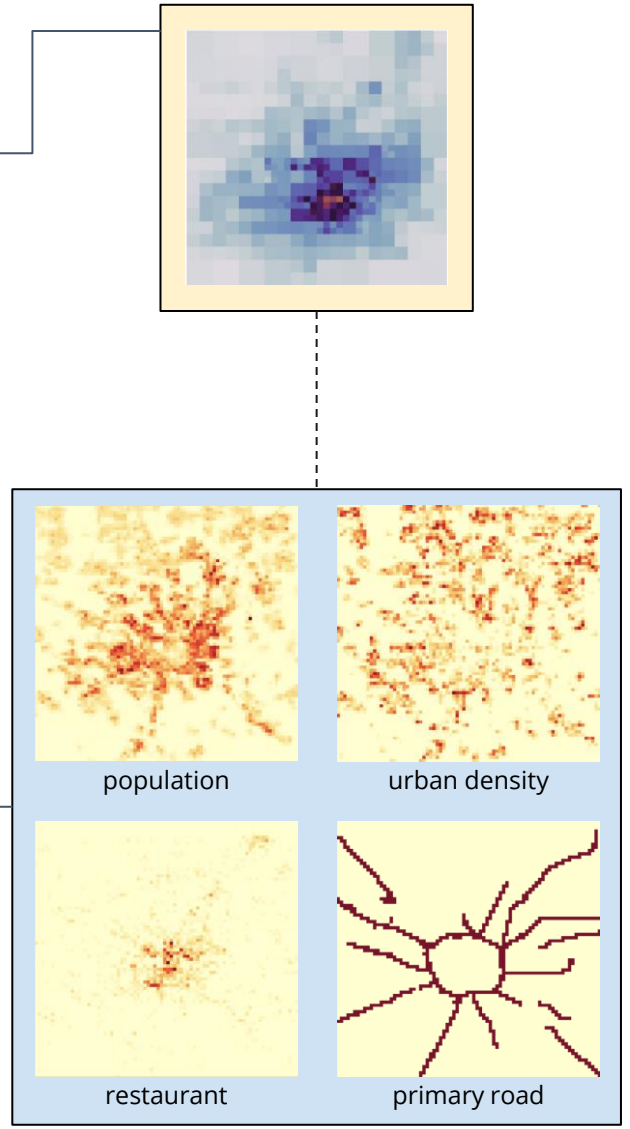


Mobile Network Traffic Data



- Lots of applications within networking:
  - Resource management
  - Mobile network infrastructure planning
  - Network energy efficiency optimization
  - Network monitoring
  - …
- And beyond:
  - Urban sensing & computing
  - Inference of commuting patterns & segregation
  - Monitoring demographic patterns
  - Detection of land use & its dynamics
  - Transportation engineering, urban planning, road traffic surveillance
  - …

A Key Insight: mobile traffic data is correlated with publicly available "context" info

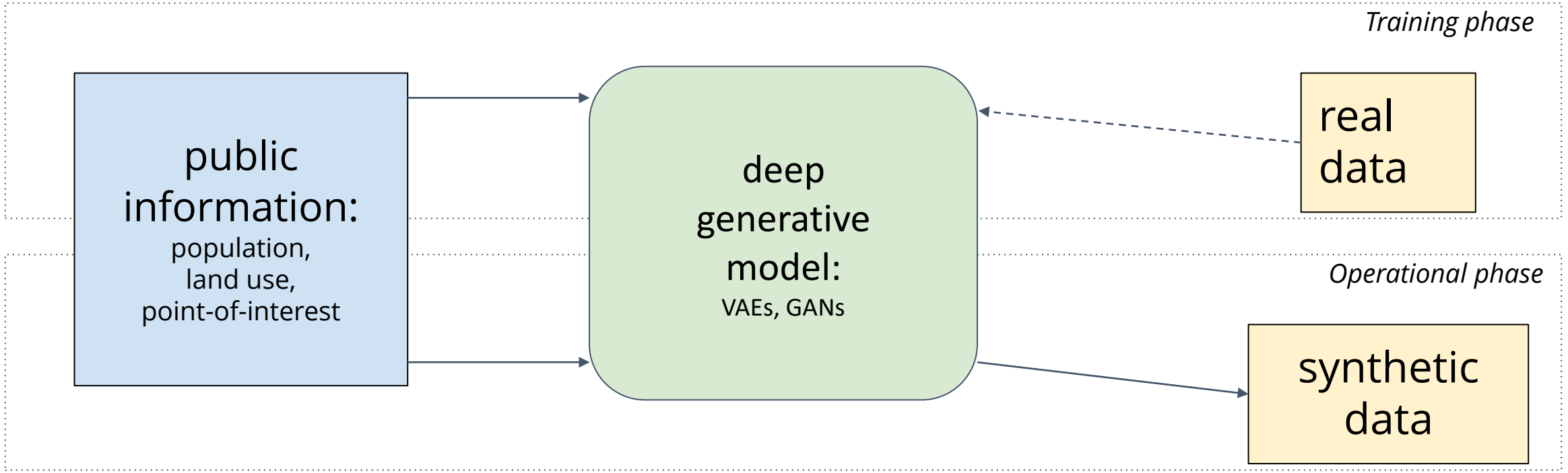


… but in a *complex*, *non-deterministic* manner

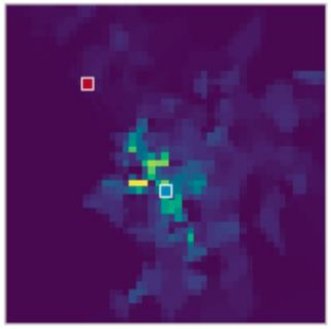
→ something *deep generative models* can capture

So: Learn p(mobile traffic data | public information)

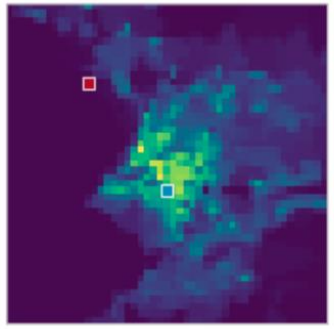
… using deep generative models.



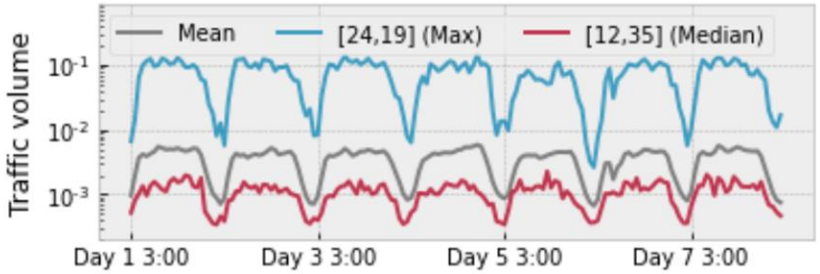
Another Key Insight



A snapshot of spatiotemporal mobile traffic data

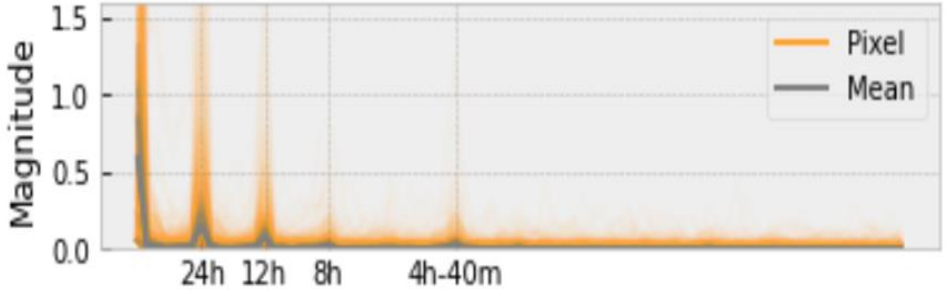


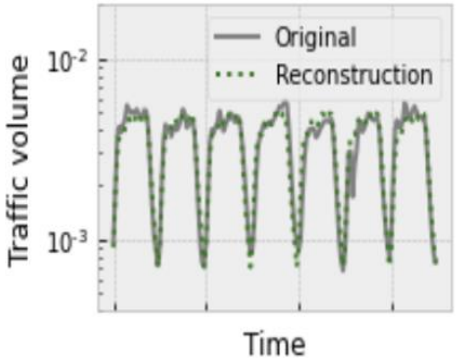
Census context (public)



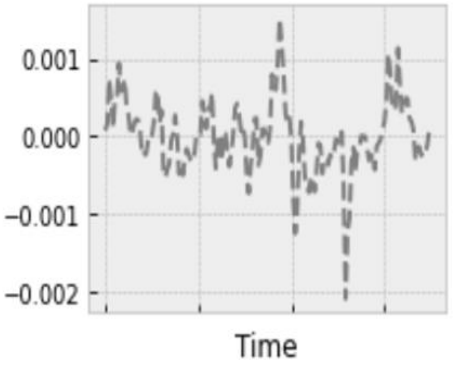
Weekly traffic: spatial-averaged (gray) and at two locations in the above traffic and context data (red/blue)

Frequency analysis of the mobile traffic



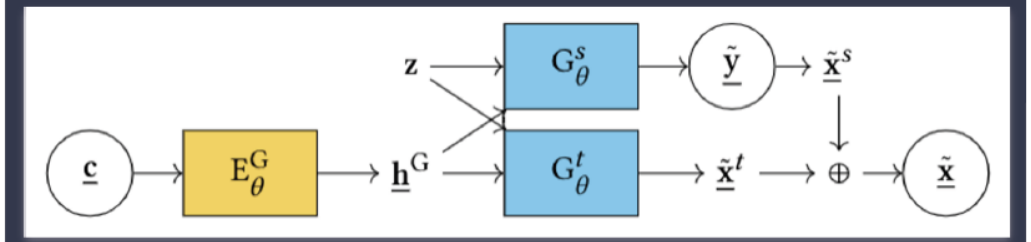


Reconstruction by significant components



Residual signal

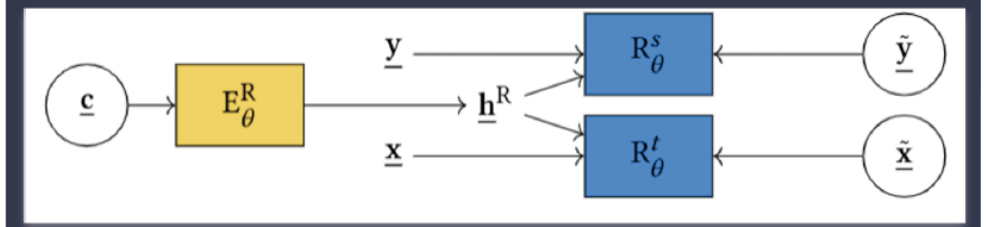
SpectraGAN: Model Design



Conditioned on the context information $\underline{c}$, SpectraGAN generates the significant frequency components $\tilde{\underline{y}}$ and residual time series $\tilde{\underline{x}}^t$ separately

- The context input is encoded into a hidden representation $\underline{h}^G$ by the context encoder
- Both frequency and time generators take this representation $\underline{h}^G$ as well as a noise vector z
- The frequency generator outputs $\tilde{\underline{y}}$
- The generated frequency components $\tilde{\underline{y}}$ converted back to the time domain as $\tilde{\underline{x}}^s$
- The time generator outputs $\tilde{\underline{x}}^t$
- The overall output is the sum of $\tilde{\underline{x}}^s$ and $\tilde{\underline{x}}^t$

SpectraGAN: Training



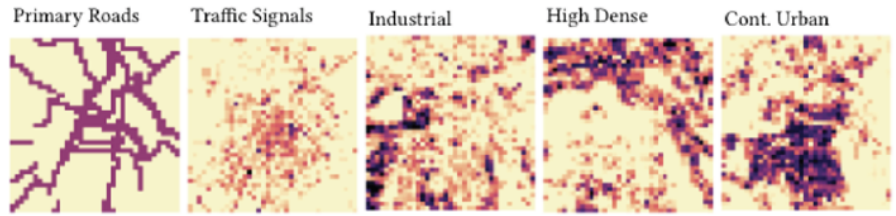
SpectraGAN uses adversarial training for learning the conditional distributions to model, i.e. using discriminators to distinguish real and generated data

- One discriminator for data in frequency domain $\underline{\mathbf{y}}$ $\underline{\tilde{\mathbf{y}}}$
- One discriminator for data in time domain $\underline{\mathbf{x}}$ $\underline{\tilde{\mathbf{x}}}$
- The context input is encoded into a hidden representation $\underline{\mathbf{h}}^{\mathrm{R}}$ by the context encoder
- Both discriminators takes a form of real or synthetic data and this hidden representation
- Extra L1 loss to improve training
- Loss function is designed to encourage the spectrum generator to attain significant components

Evaluation: Datasets, Metrics & Baselines

Datasets of real mobile traffic from two major European countries

- Country 1: CITY A-CITY I (9 cities)
- Country 2: CITY 1-CITY 4 (4 cities)
- 27 publicly available context attributes



Metrics for different aspects of the data

- Marginal: Total variation for marginals (M-TV)
- Spatial: SSIM on time-averaged data (SSIM)
- Temporal: Auto-correlation differences (AC-L1)
- Spatiotemporal: train-synthetic-test-real (TSTR), Fréchet video distance (FVD)

Baselines to benchmark the generation quality

- FDaS [1]: models the marginal
- Pix2Pix [2]: image translation
- DoppelGANger [3]: time-series GAN
- Conv3D+LSTM [4]: spatio-temporal GAN

[1] P. Di Francesco, F. Malandrino, and L. A. DaSilva. 2018. Assembling and Using a Cellular Dataset for Mobile Network Analysis and Planning. IEEE Transactions on Big Data 4, 4 (2018), 614–620.

[2] P. Isola et al. 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1125–1134.

[3] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar. 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In Proceedings of the ACM Internet Measurement Conference. 464–483.

[4] D. Saxena and J. Cao. 2019. D-GAN: Deep generative adversarial nets for spatio-temporal prediction. arXiv preprint arXiv:1907.08556 (2019).

Results: Quantitative Generation Quality

Average testing performance in Country 1

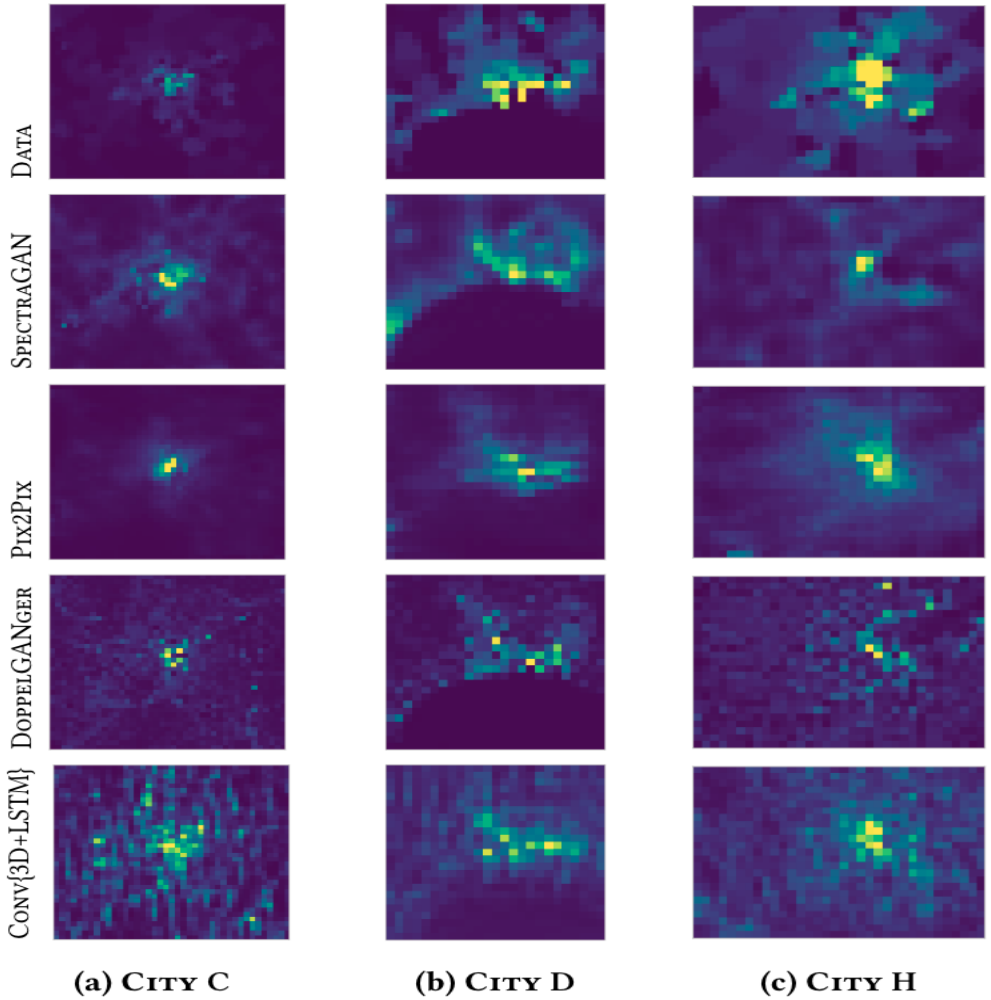
Method	M-TV ↓	SSIM ↑	AC-L_1 ↓	TSTR ↑	FVD ↓
SpectraGAN	0.0362	0.787	46.8	0.893	205
Pix2Pix	0.0522	0.800	84.4	0.557	214
DoppelGANger	0.0498	0.744	54.8	0.890	247
Conv{3D+LSTM}	0.0460	0.750	60.2	0.895	281
Data	0.00359	0.999	25.2	0.903	128

Average testing performance in Country 2

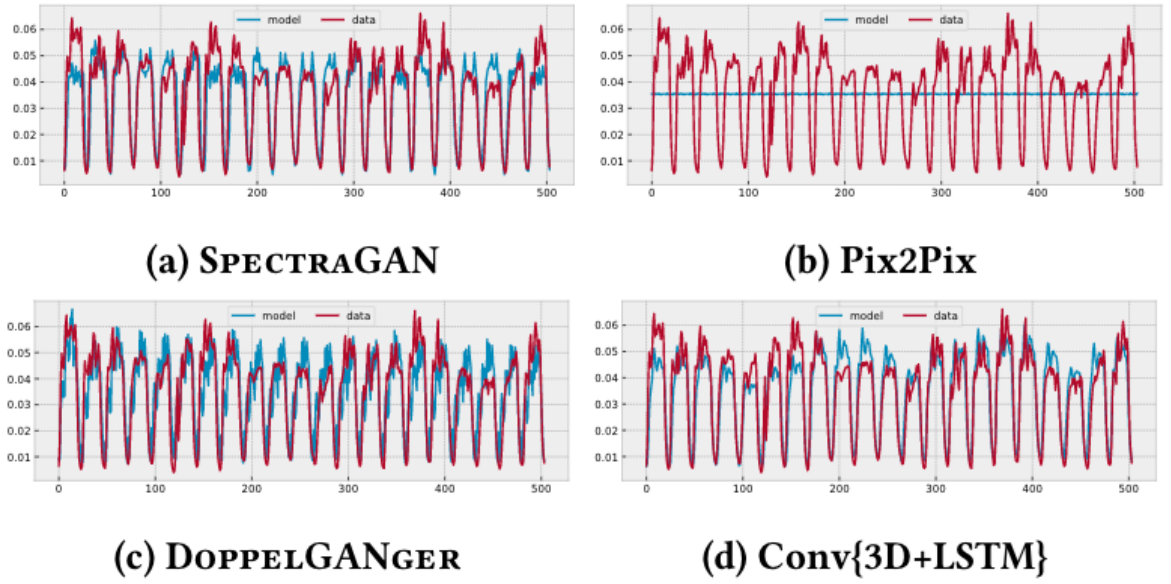
Method	M-TV ↓	SSIM ↑	AC-L_1 ↓	TSTR ↑
SpectraGAN	0.0607	0.686	34.8	0.977
Pix2Pix	0.121	0.564	117	0.653
DoppelGANger	0.0521	0.472	40.9	0.964
Conv{3D+LSTM}	0.0514	0.613	99.5	0.946
Data	0.0076	0.996	22.8	0.978

SpectraGAN is the best model considering all metrics.

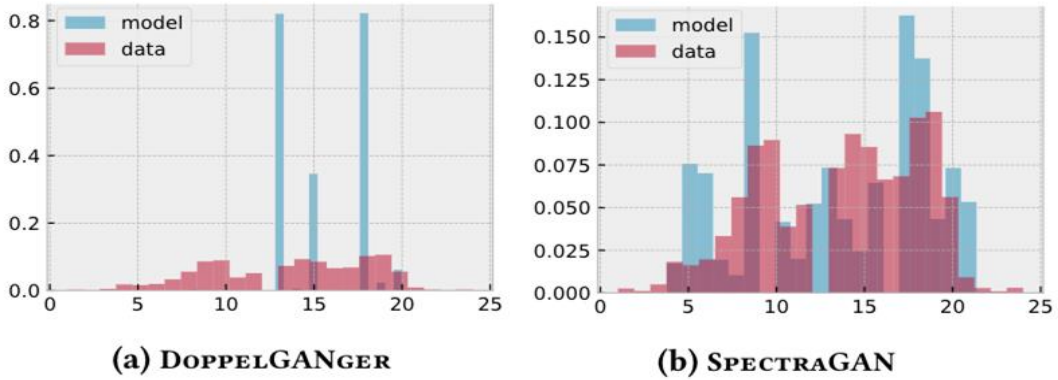
Qualitative Results



Time-averaged mobile traffic (all models)

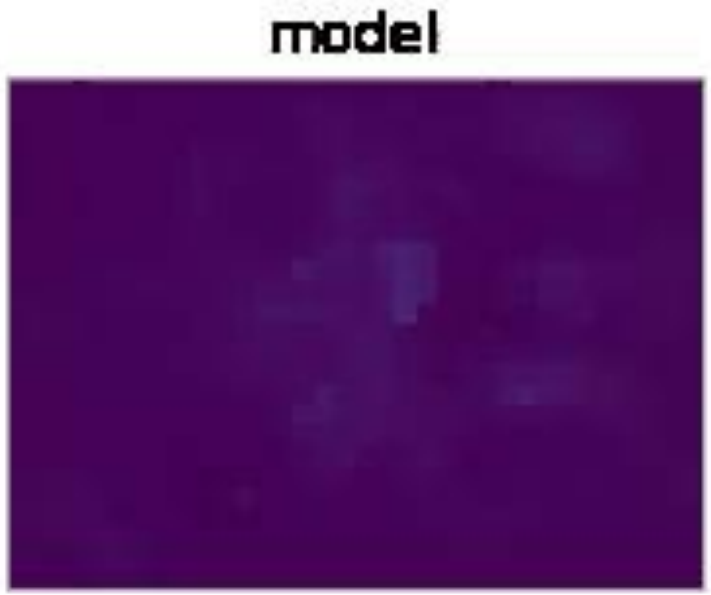


Spatial-averaged city-wide mobile traffic (City B)



Peak distributions (City B)

SpectraGAN Demo

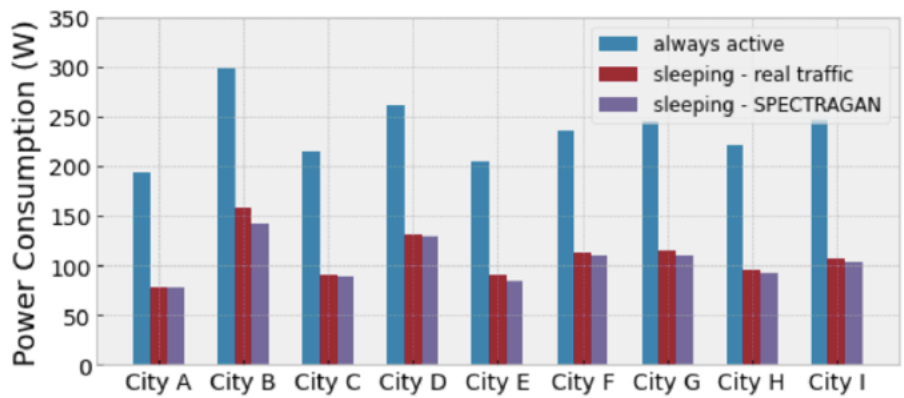


Use Cases: Data-Driven Micro Base Station Sleeping

Synthetic data allows researchers to evaluate performance of a new data-driven solution for network management and beyond

Dynamically switch base stations on/off based on traffic

- The substantial operating expense due to energy consumption at base stations (BSs) ⇒ a number of solutions for saving power in the RAN, e.g. [1]
- Heterogeneous RAN deployment:
  - Micro BS: each pixel (i.e. 1x1 grid cells)
  - Macro BS: an umbrella coverage of 5×5 grid cells
- Sleeping reduces power consumption by 47–62%
- Inline with what can be achieved by real data



[1] G. Vallero, D. Renga, M. Meo, and M. A. Marsan. 2019. Greener RAN Operation Through Machine Learning. IEEE Transactions on Network and Service Management 16, 3 (2019), 896–908.

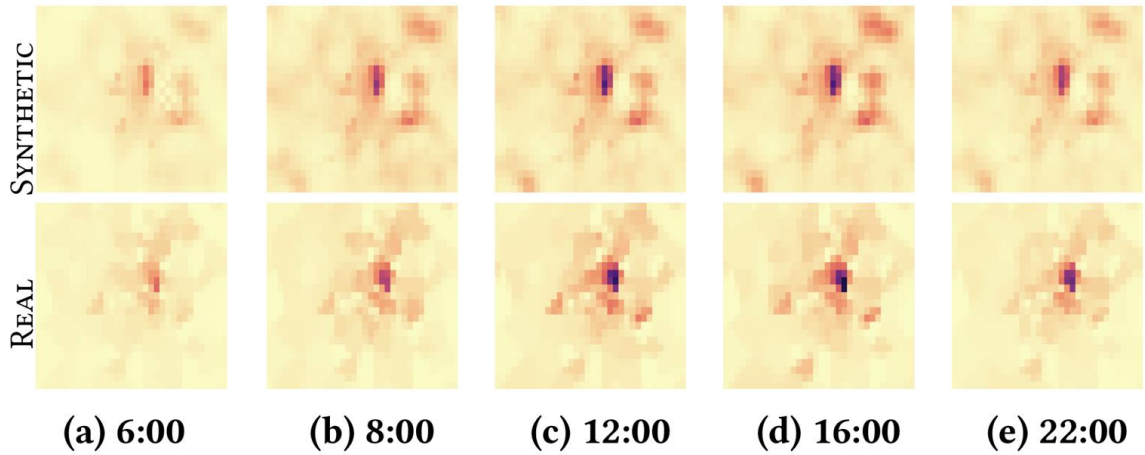
Use Cases: Dynamic Human Presence Mapping

Tracking of population density in real-time using mobile traffic data

- Multivariate regression model [1] to track population $p_i(t)$ at grid cell i and time t from network traffic $x_i(t)$

$$p_i(t) = e^{k_1 \lambda_i(t) + k_2} x_i(t)^{k_3 \lambda_i(t) + k_4}$$

  - Estimated according to [1]
- Comparing the outputs the model
  - Peak signal-to-noise ratio (PSNR) > 25
  - Values of PSNR > 20 are acceptable [2]



Dynamic people presence estimated at five different times of the day for a sample city

[1] G. Khodabandelou et al. 2019. Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata. IEEE Transactions on Mobile Computing 18, 9 (2019), 2034–2047. https://doi.org/10.1109/TMC.2018.2871156
[2] N. Thomos, N. V. Boulgouris, and M. G. Strintzis. 2006. Optimized transmission of JPEG2000 streams over wireless channels. IEEE Transactions on Image Processing 15, 1 (2006), 54–67. https://doi.org/10.1109/TIP.2005.860338

Summary

- Readily available context + generative models = solution to data accessibility
- Domain specific insights necessary for high-fidelity and generalizable data synthesis
- Developed a suite of deep generative models for mobile traffic and beyond:
  - SpectraGAN [CoNEXT'21], CartaGenie [PerCom'22], …
- Synthesized mobile traffic data for multiple cities publicly released via https://github.com/netsys-edinburgh/
- Several further challenges to investigate on:
  - Mobile network data generation
  - Foundational research in generative modelling, informed by our domain specific data synthesis experience