FoldFormer: sequence folding and seasonal attention for fine-grained long-term FaaS forecasting

Luke Darlow
sirlab@huawei.com
Edinburgh Research Centre,
Central Software Institute, Huawei.
Edinburgh, United Kingdom

Artjom Joosen
Edinburgh Research Centre,
Central Software Institute, Huawei.
Edinburgh, United Kingdom

Martin Asenov
Edinburgh Research Centre,
Central Software Institute, Huawei.
Edinburgh, United Kingdom

Date: 08 May 2023

Qiwen Deng
Edinburgh Research Centre,
Central Software Institute, Huawei.
School of Informatics,
University of Edinburgh.
Edinburgh, United Kingdom

Jianfeng Wang
Hangzhou Research Centre,
Central Software Institute, Huawei.
Hangzhou City, China

Adam Barker
Edinburgh Research Centre,
Central Software Institute, Huawei.
School of Computer Science,
University of St Andrews.
United Kingdom

HUAWEI

FoldFormer Overview

- Problem space: FaaS forecasting
- Periodic assumptions and inductive bias
- Model overview
  - Time-to-latent-folding
  - FFT convolutions
  - Seasonal Attention
- Data and experimental setup
- Results
- Systemic challenges and future work
- Demonstration

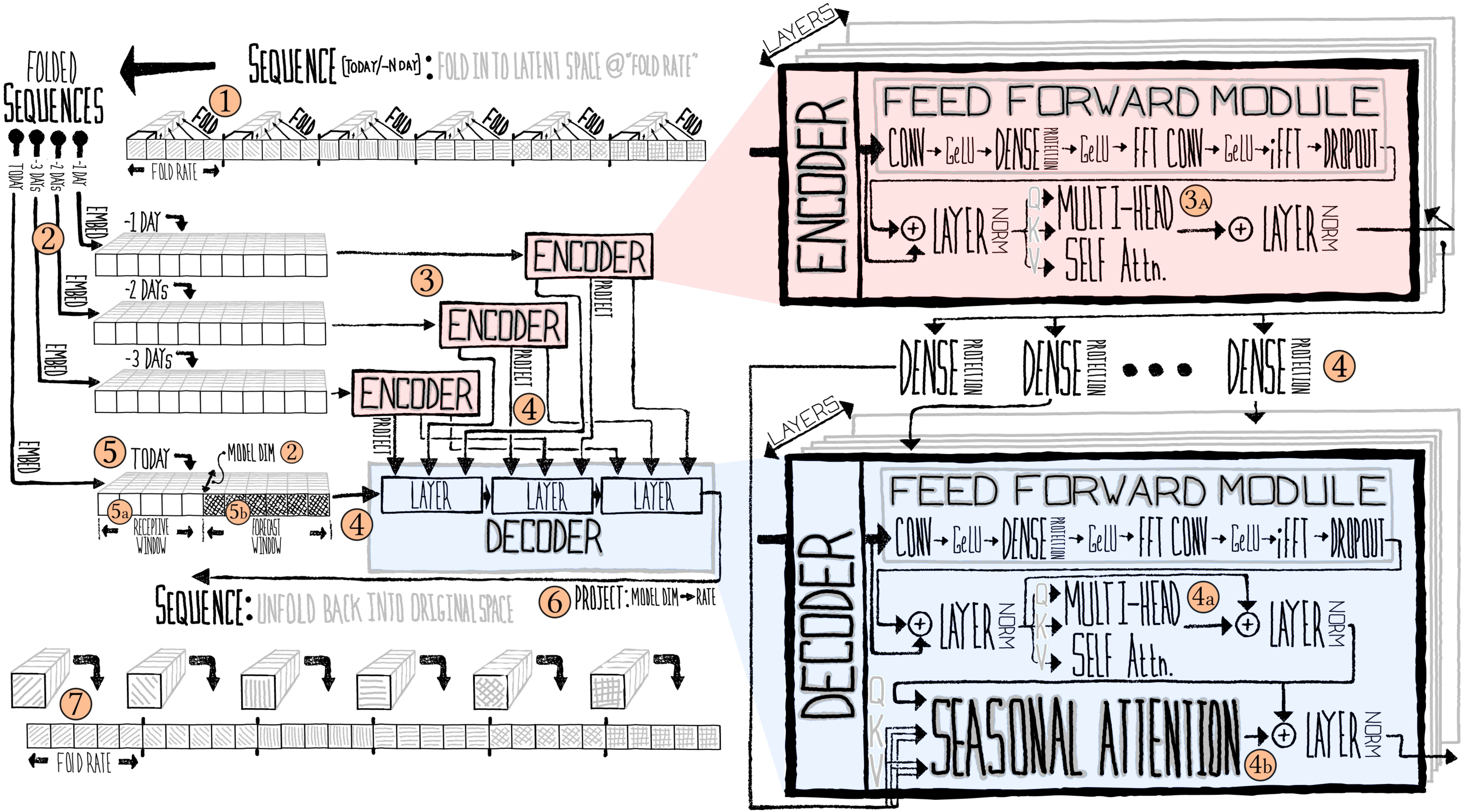
Problem space: FaaS forecasting

- Function as a service (FaaS) is a stateless, event driven platform
- Cold-start problem:
  - Resources not ready to receive function requests
- Over-commit problem:
  - Too many resources waiting for work
- Accurate forecasting of incoming function requests could remedy or alleviate both issues
- FaaS data tends to be fine-grained, long-term, and often periodic in nature

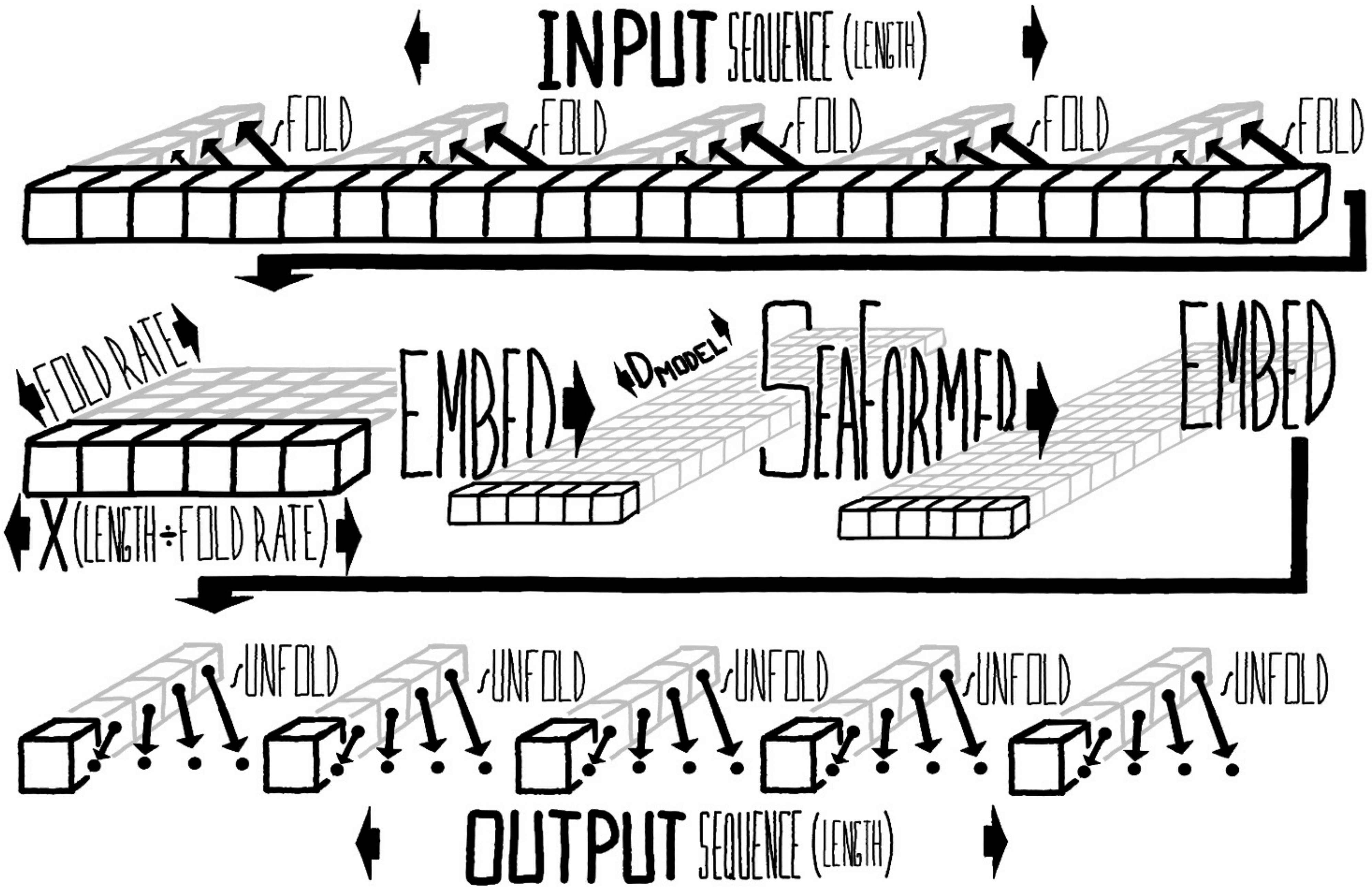
Periodic assumptions and inductive bias

- FaaS requests are a function of human users, subject to the human diurnal cycle
- We found that most high-demand functions tended to have strong periodicity
- FoldFormer was designed around this assumption
- Data sampling regime:
  - Let the model see only ‘snippets’ of data from successive periods (see video at the end)
  - Enables longer term ingesting (less data to process)
  - Enables longer term forecasting (autoregressive process has minimal divergent impact)

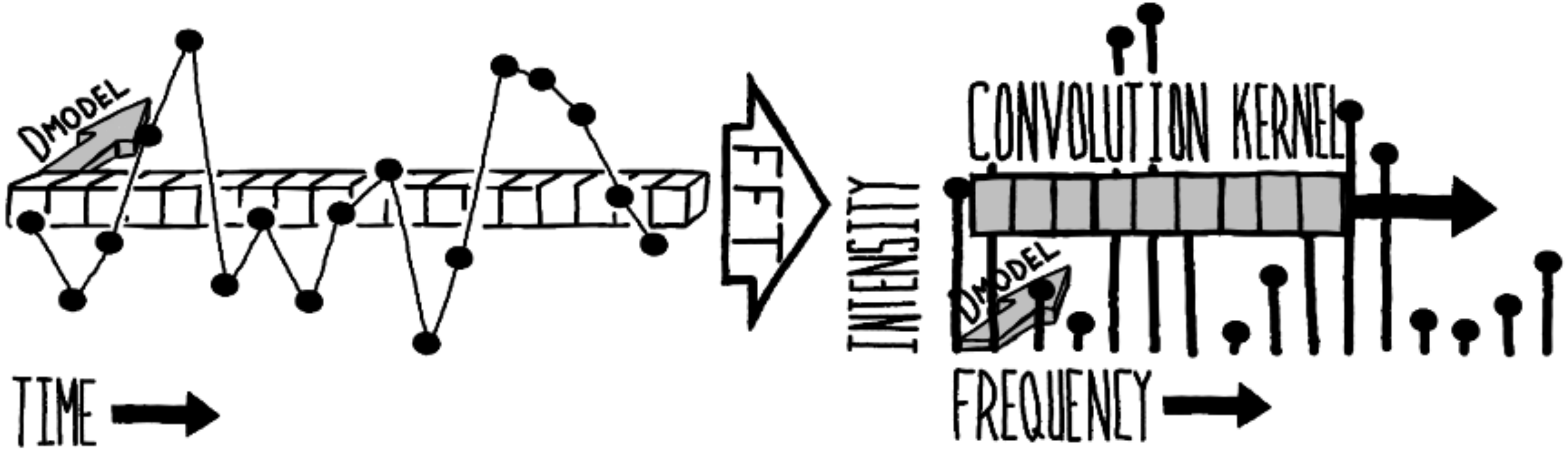
FoldFormer overview



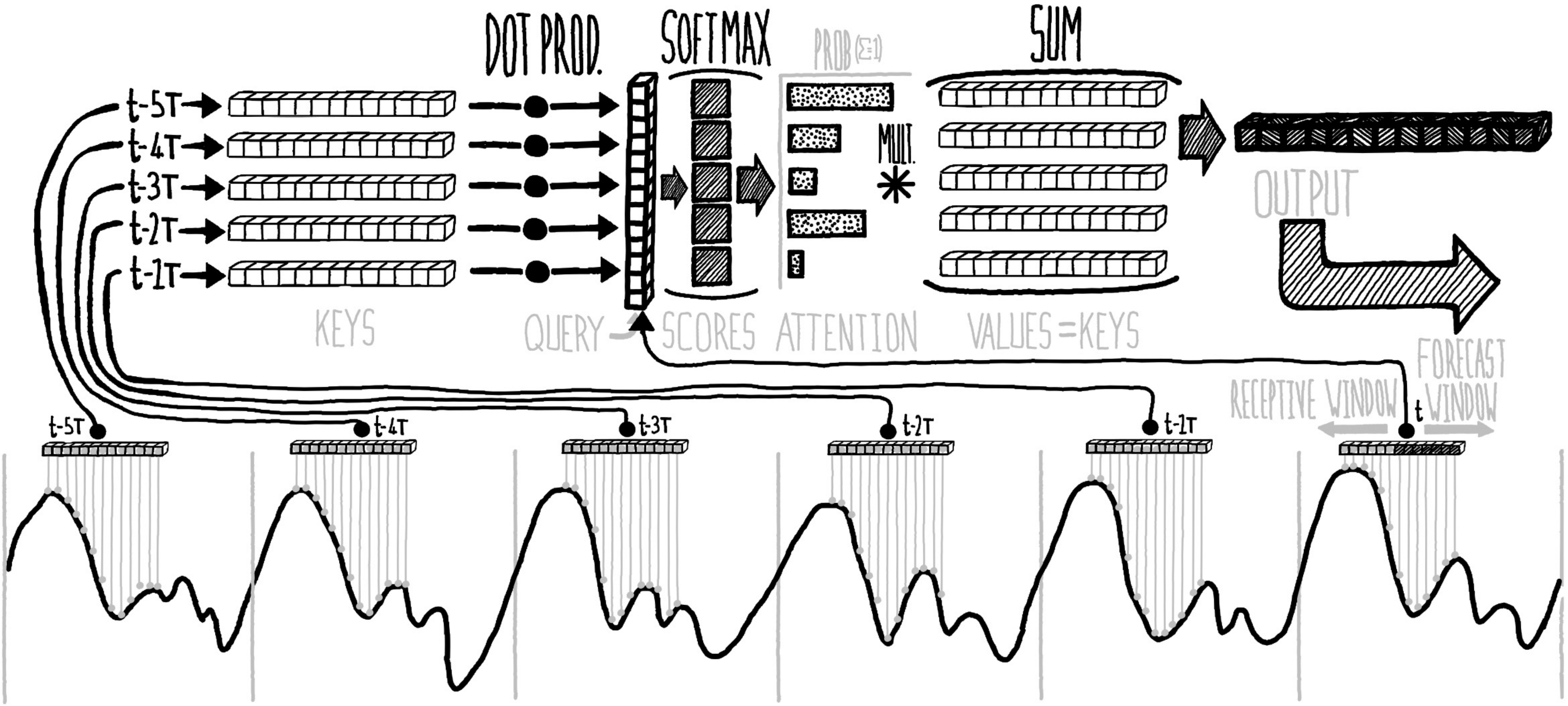
Time-to-latent folding



FFT convolutions



Seasonal attention



Data and experimental setup

- Three data sources
  - Azure function traces (AZ) – per-minute
  - FunctionGraph product (FG) – per-minute
  - An internal serverless platform (SP) – per-second and per-minute
- Train global models wherever possible
  - This is a better solution to scale-up
- Compare to:
  - N-BEATS
  - N-HiTS
  - Linear regression
  - Basic Transformer
  - Prophet
  - Autoformer
  - FFT solution

Results

Metric	Method	AZ 1	AZ 2	AZ 3	AZ 4	AZ 5	FG 1	FG 2	FG 3	FG 4	FG 5	SP 1	SP 2	SP 3	SP 4	SP 5	SP_s 1	SP_s 2	SP_s 3	SP_s 4	SP_s 5
RMSE	FoldFormer	0.528	0.518	0.422	1.464	0.435	0.929	**0.564**	**0.511**	1.011	**0.536**	0.215	**0.406**	0.166	0.226	0.091	**0.247**	**0.392**	**0.186**	**0.251**	**0.318**
	N-BEATS	0.446	0.371	0.527	1.102	0.521	1.013	0.873	0.952	1.017	0.586	0.203	0.581	0.169	0.238	0.086					
	N-HiTS	0.369	0.335	0.604	1.010	0.594	1.091	0.765	0.773	2.436	0.833	**0.197**	0.764	0.225	0.274	0.230	1.217	4.095	1.323	1.330	1.418
	Regression	0.350	0.301	0.512	1.003	0.498	**0.842**	0.734	0.795	0.990	0.600	0.230	0.690	**0.157**	**0.218**	**0.082**					
	Transformer	0.489	0.432	0.467	0.926	0.465	0.944	0.957	1.033	1.709	0.666	0.516	0.706	0.212	0.279	0.261					
	Prophet	**0.291**	**0.278**	0.572	**0.572**	0.558	1.063	1.069	1.164	1.017	1.129	0.581	0.650	0.258	0.315	0.226					
	Autoformer	1.269	1.408	0.723	1.884	1.622	1.080	1.937	2.186	1.101	0.997	1.619	1.149	1.086	0.869	1.081					
	FFT	0.526	0.486	**0.399**	0.659	**0.406**	1.053	1.915	1.932	**1.002**	3.058	0.307	1.903	0.277	0.300	0.316	1.256	4.176	0.511	0.525	0.366
MAPE	FoldFormer	0.146	0.127	0.110	1.140	0.108	0.677	**0.230**	**0.216**	2.042	**0.183**	0.027	**0.081**	0.047	0.060	0.066	**0.120**	**0.070**	**0.054**	**0.058**	**0.100**
	N-BEATS	0.123	0.096	0.158	1.004	0.156	0.787	0.465	0.496	2.819	0.222	0.026	0.104	0.043	0.055	**0.042**					
	N-HiTS	0.095	0.079	0.165	0.745	0.162	0.587	0.398	0.433	16.804	0.264	**0.024**	0.163	0.055	0.066	0.106	0.256	1.002	0.604	0.602	2.547
	Regression	0.094	0.075	0.144	0.708	0.143	**0.544**	0.650	0.699	2.590	0.324	0.031	0.172	**0.036**	**0.046**	0.085					
	Transformer	0.135	0.109	0.124	0.676	0.127	0.863	0.496	0.533	10.838	0.258	0.050	0.135	0.062	0.073	0.129					
	Prophet	**0.090**	**0.074**	0.172	**0.479**	0.172	0.821	0.566	0.606	2.936	0.515	0.077	0.160	0.069	0.084	0.107					
	Autoformer	0.448	0.358	0.263	2.486	0.545	5.673	2.179	3.247	1.031	0.463	0.219	0.276	0.487	0.289	1.103					
	FFT	0.144	0.127	**0.106**	0.542	**0.106**	3.344	0.814	0.822	2.402	0.926	0.036	0.430	0.077	0.083	0.203	0.127	1.069	0.121	0.122	0.150
MAE	FoldFormer	0.416	0.403	0.310	1.104	0.310	0.649	**0.414**	**0.359**	**0.204**	**0.397**	0.172	**0.317**	0.089	0.099	0.069	**0.192**	**0.290**	**0.100**	**0.115**	**0.130**
	N-BEATS	0.365	0.310	0.444	0.864	0.445	0.737	0.661	0.710	0.283	0.449	0.173	0.418	0.095	0.105	**0.055**					
	N-HiTS	0.285	0.260	0.487	0.801	0.482	0.772	0.554	0.580	0.745	0.662	**0.165**	0.545	0.134	0.143	0.156	1.013	3.707	0.924	0.901	1.103
	Regression	0.272	0.237	0.423	0.782	0.421	**0.626**	0.598	0.637	0.247	0.491	0.205	0.642	**0.075**	**0.082**	0.064					
	Transformer	0.388	0.345	0.350	0.701	0.360	0.673	0.744	0.812	1.346	0.524	0.353	0.575	0.141	0.149	0.150					
	Prophet	**0.237**	**0.221**	0.487	**0.480**	0.483	0.781	0.919	1.000	0.289	1.018	0.515	0.443	0.178	0.203	0.121					
	Autoformer	1.114	1.171	0.567	1.602	1.216	0.885	1.459	1.825	0.439	0.841	1.354	1.061	0.853	0.614	0.832					
	FFT	0.428	0.405	**0.292**	0.528	**0.294**	0.858	1.324	1.322	0.215	2.014	0.244	1.398	0.189	0.186	0.236	0.645	3.680	0.318	0.324	0.155

Table 1: Results for autoregressive two-day prediction for top 5 functions per dataset. Before computing results, requests were standardised using training data statistics. Each dataset is shown in 5 columns and listed in order from most to least popular function. Best results for each dataset are emphasised. Per-second results are shown only for SP and denote as SP_s and only on the models that could ingest this granularity of data.

Results

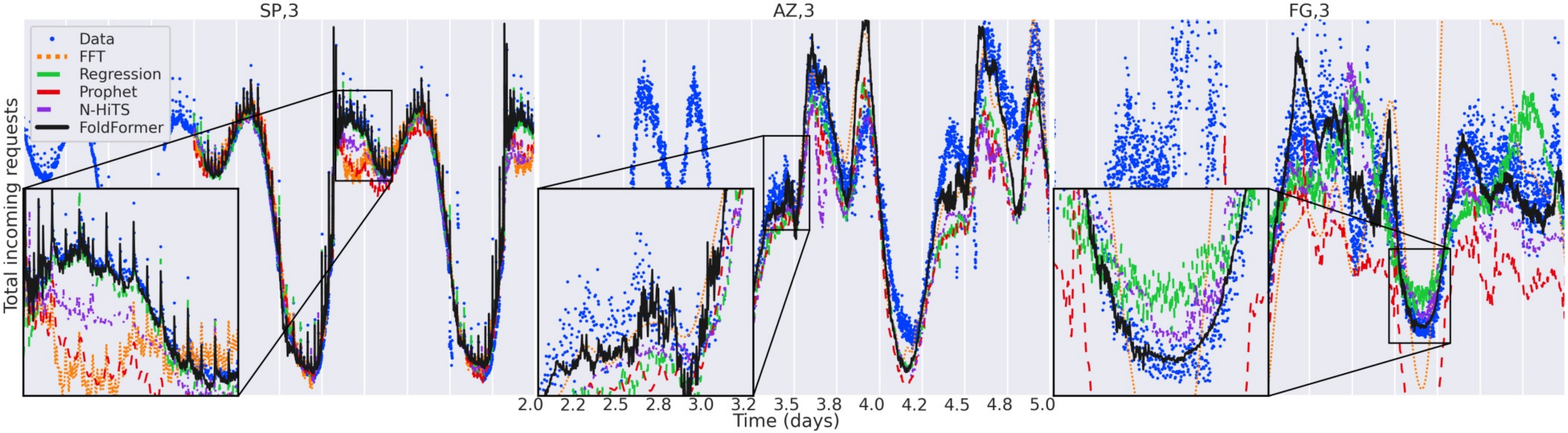


Figure 4: Third most popular functions from the top 5, showing ground-truth data and only the top-performing forecasts for optimal visualisation clarity. The final context day is shown along with both forecast days. The scale of the y-axis is removed to obscure sensitive traffic information.

Results

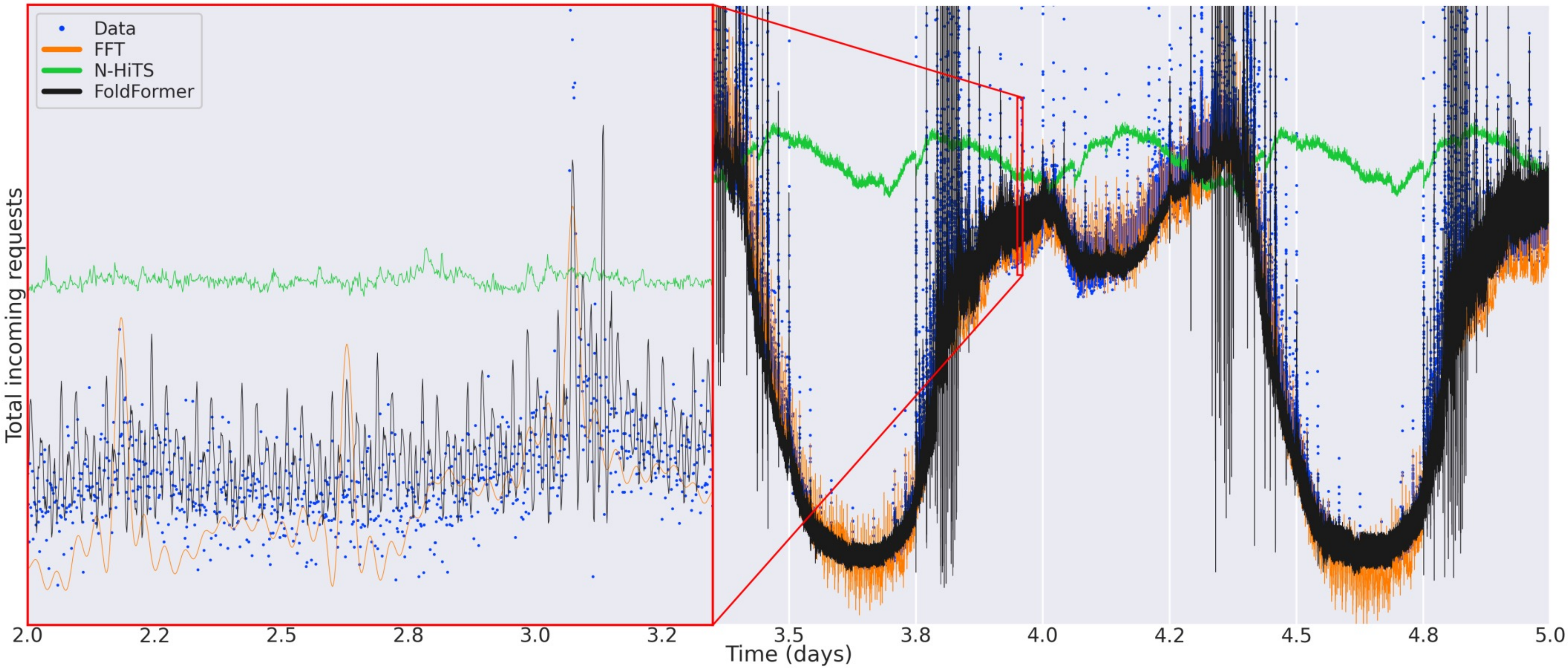
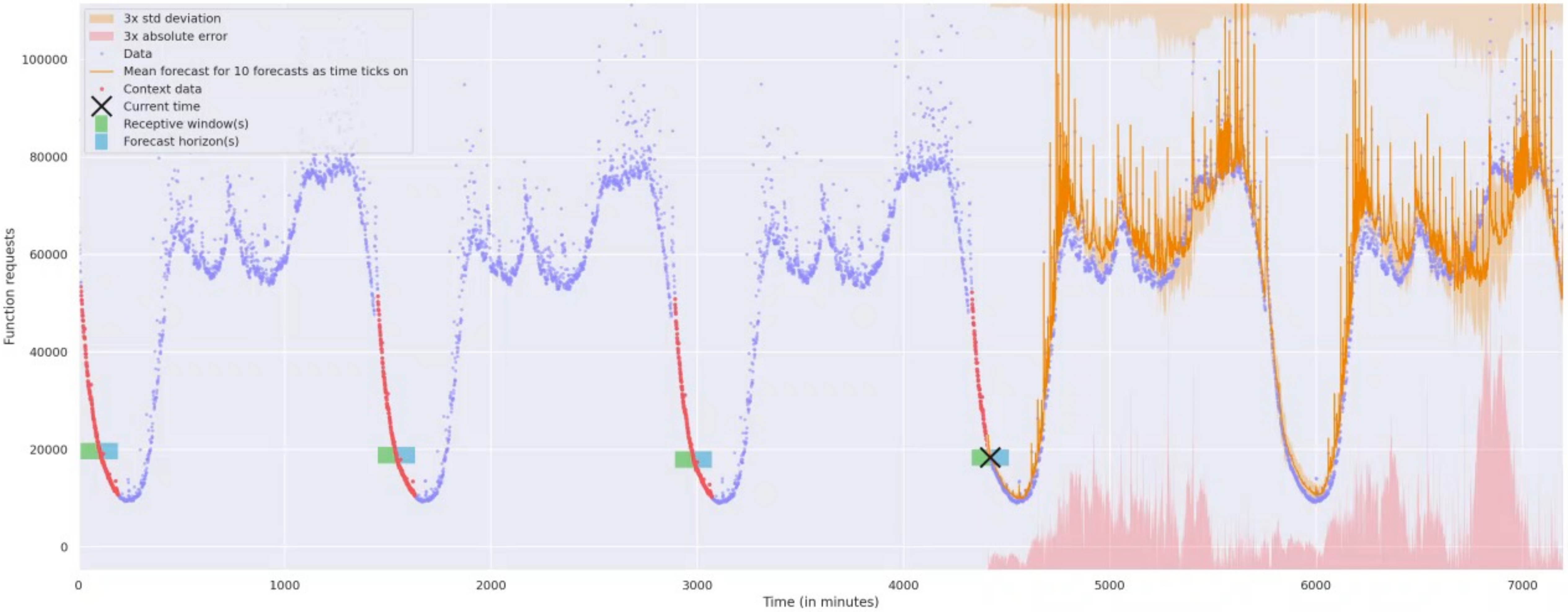


Figure 5: Per-second data and forecasts. The zoom inset covers a region of only 15 minutes. The scale of the y-axis is removed to obscure sensitive traffic information.

Systemic challenges and future work

- Implication of periodic data sampling regime is that FoldFormer is better suited (but not only suited) to periodic data
- Still uses a transformer, which can be expensive
- Weekly periodicity not accounted for in current version: we need to use more context days
  - Data challenge: e.g., Azure only has 2 weeks of data
- Future work:
  - Very long-term ingestion of context data
  - Very long-term forecasting
  - Architecture updates
  - Incorporate seq2seq approach within FoldFormer


3x std deviation
3x absolute error
Data
Mean forecast for 10 forecasts as time ticks on
Context data
Current time
Receptive window(s)
Forecast horizon(s)
Function requests
Time (in minutes)
0
20000
40000
60000
80000
100000
0
1000
2000
3000
4000
5000
6000
7000