

Глубокие модели

День 3



Какие модели рассмотрим

MLP

MLP-based

DeepAR

RNN-based

TFT

Transformer-based

NBeats

TS-specific

DeepState

State space models based

TCN, Wavenet

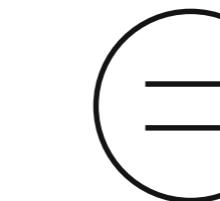
CNN-based

Многослойный перцептрон MLP

Подготовка
данных

Dates	Store ID	Product ID	Product name	# of Sales
02-12-2021	001	RP01	Almond milk	21
10-12-2021	005	RS21	Oat milk	15
18-01-2022	004	RK32	Hazelnut milk	9

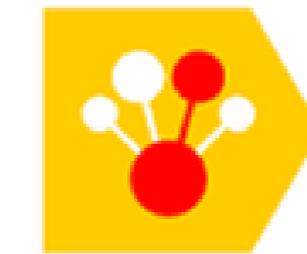
ML-модели



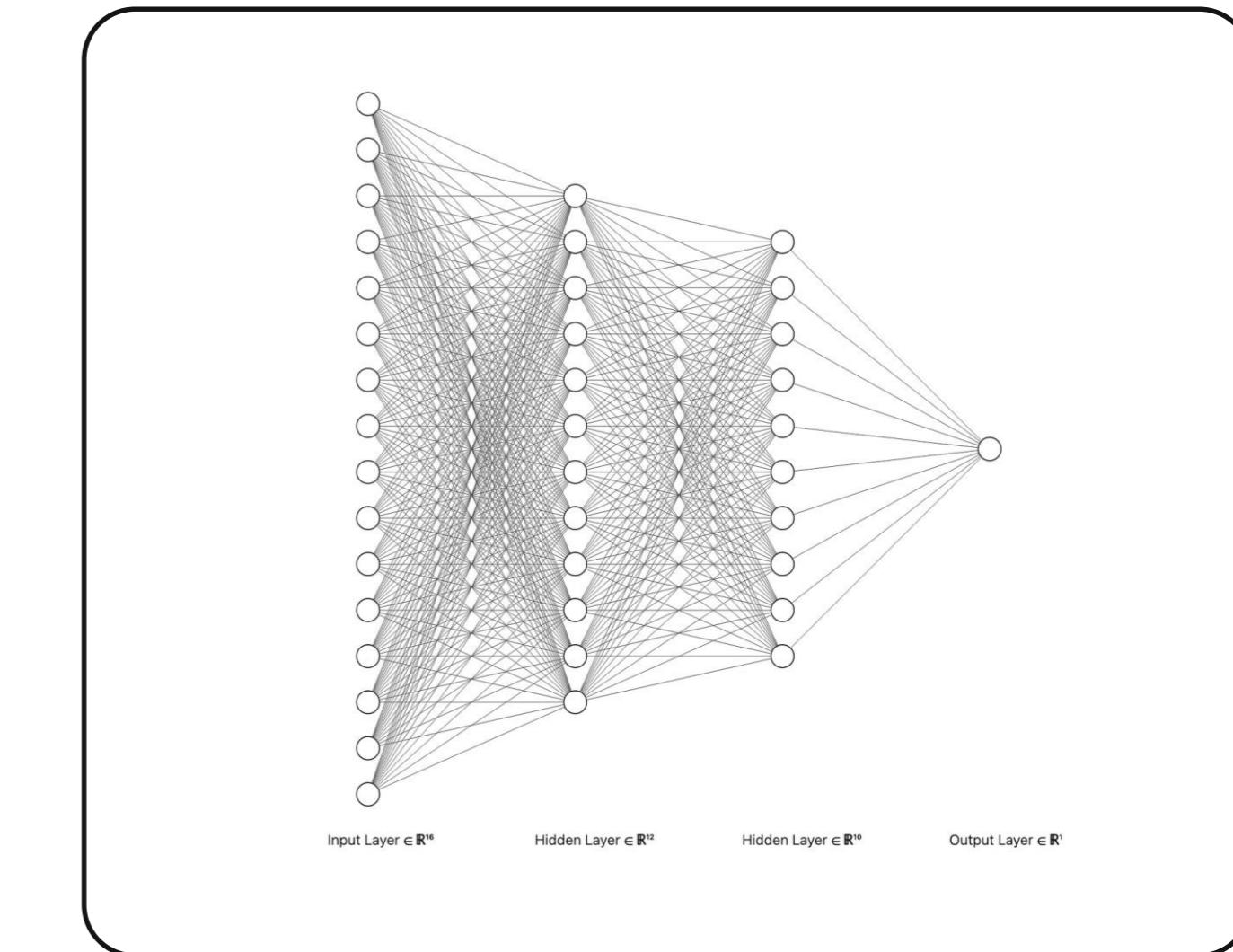
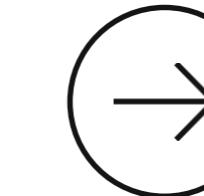
Dates	Store ID	Product ID	Product name	# of Sales
02-12-2021	001	RP01	Almond milk	21
10-12-2021	005	RS21	Oat milk	15
18-01-2022	004	RK32	Hazelnut milk	9

MLP

Модель



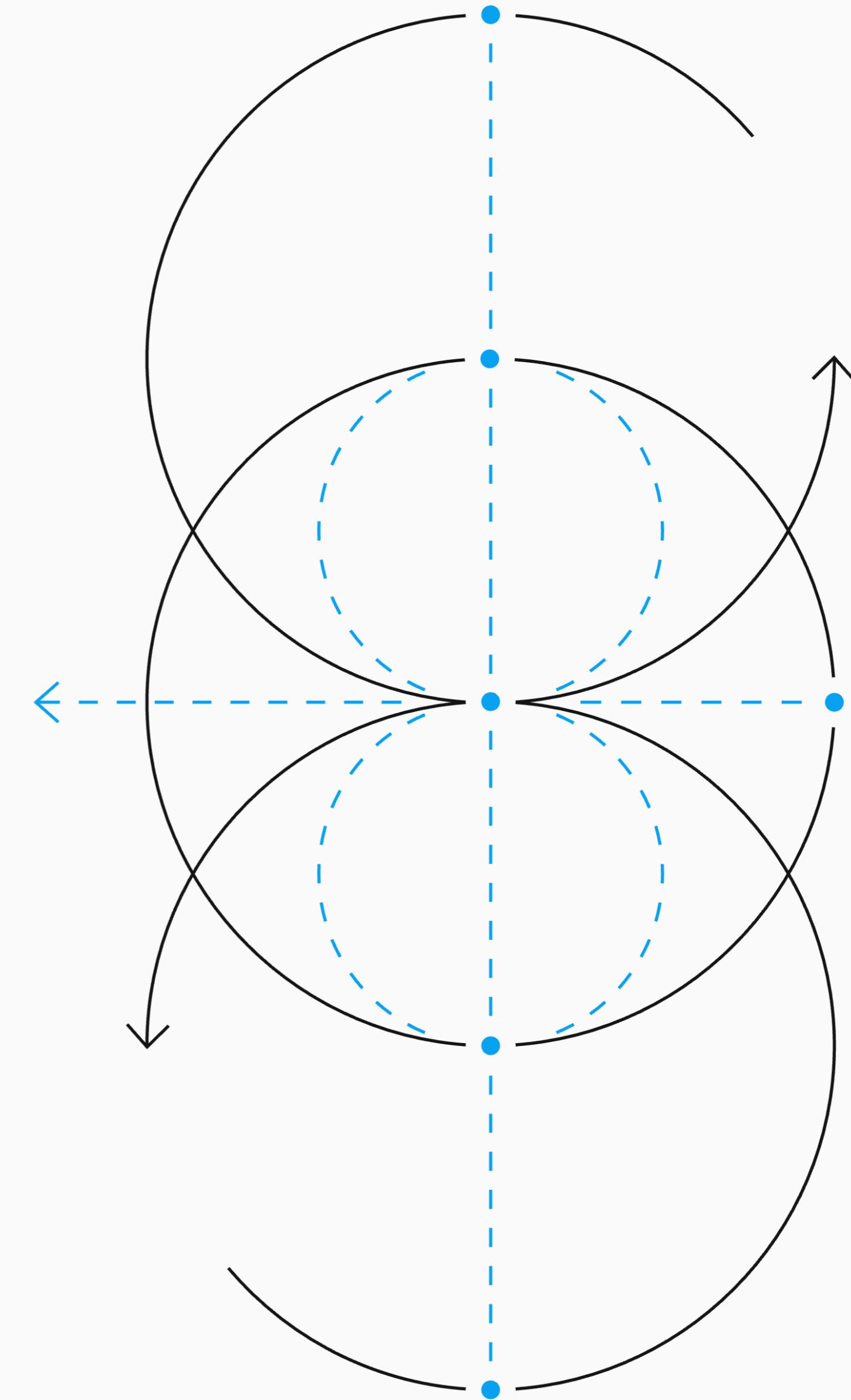
CatBoost



Многослойный перцептрон MLP

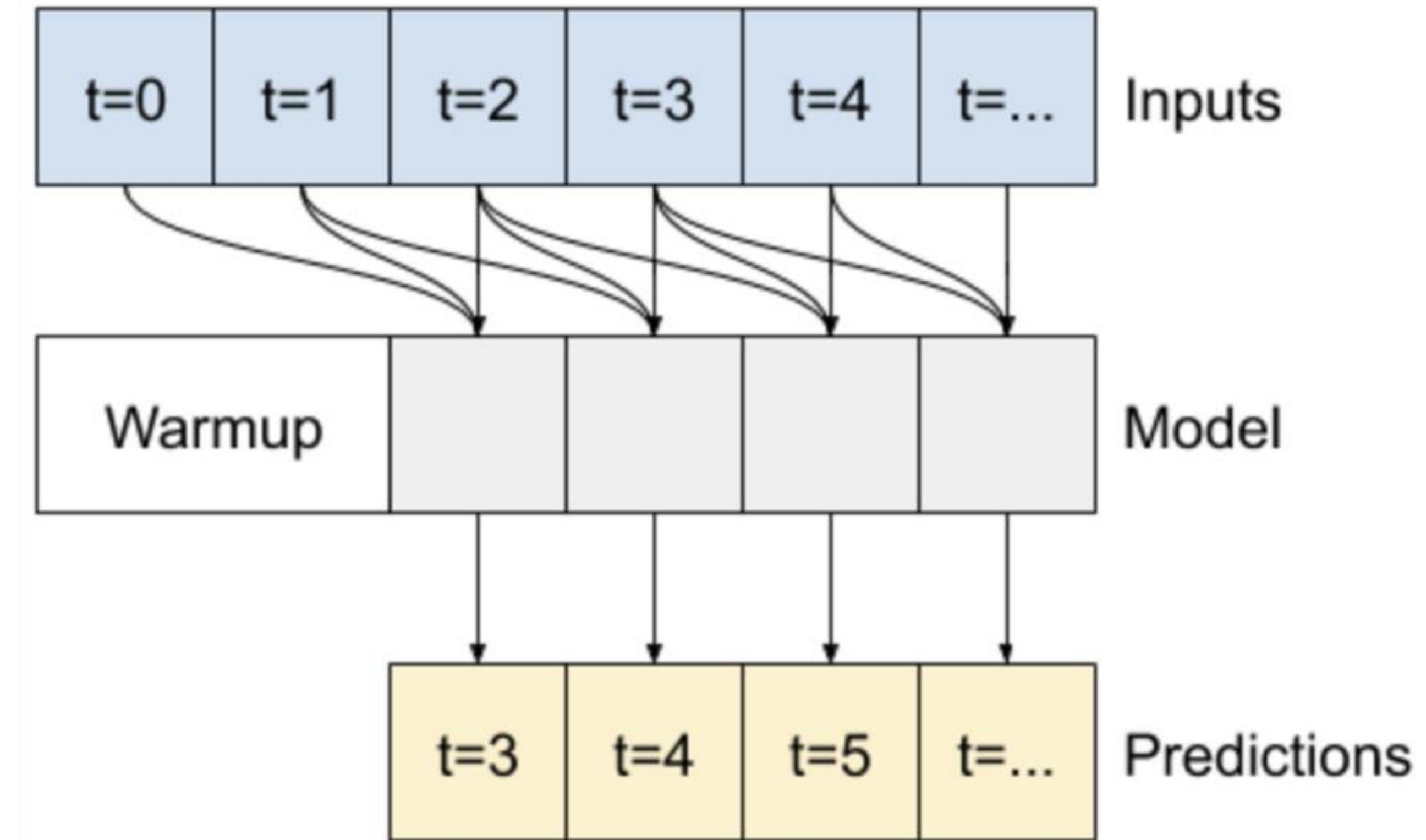
Почему вам могут понадобиться MLP-модели

- Меньше проблем с экстраполяцией, чем у деревьев, и в то же время модель сильнее, чем просто линейная регрессия.
- Проще интегрировать свою функцию потерь, чем в деревья или градиентный бустинг: иногда это может быть важно.
- Легко переделать в вероятностную модель, поменяв функцию потерь и последний слой. Обсудим это в контексте предсказательных интервалов.



CNN-based

- Это расширение классических подходов со скользящим окном, но теперь параметры обучаемые, и мы можем использовать несколько слоев для увеличения контекста (receptive field).
- В целом можно пробовать любые архитектуры из speech recognition, например WaveNet.



Рекуррентные нейронные сети RNN

Мы же используем сети — зачем нам строить признаки руками?

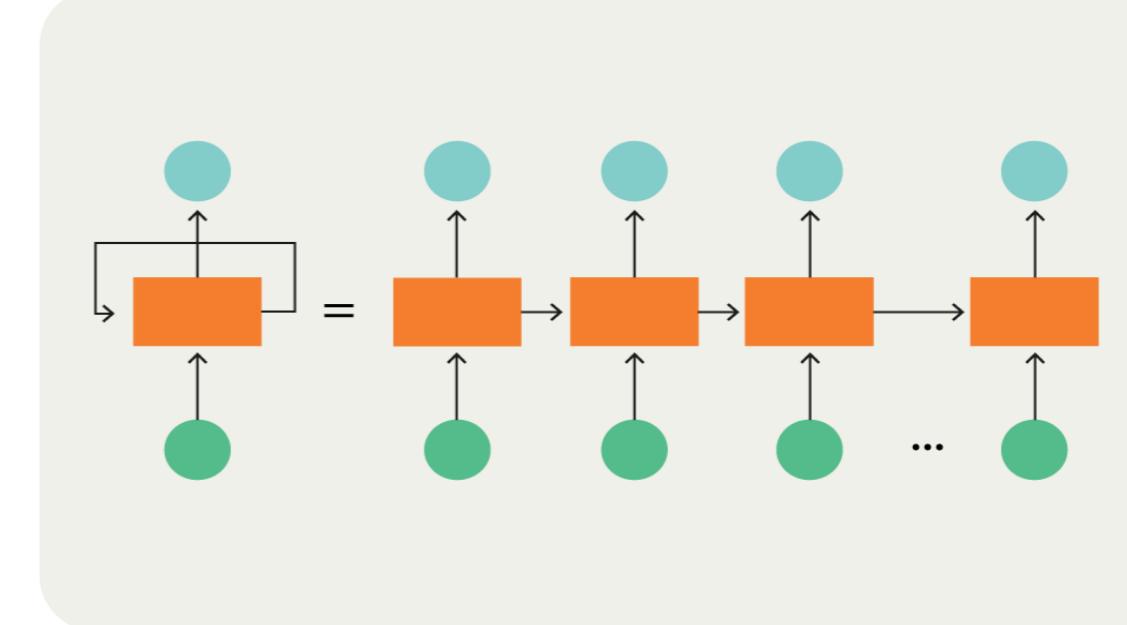
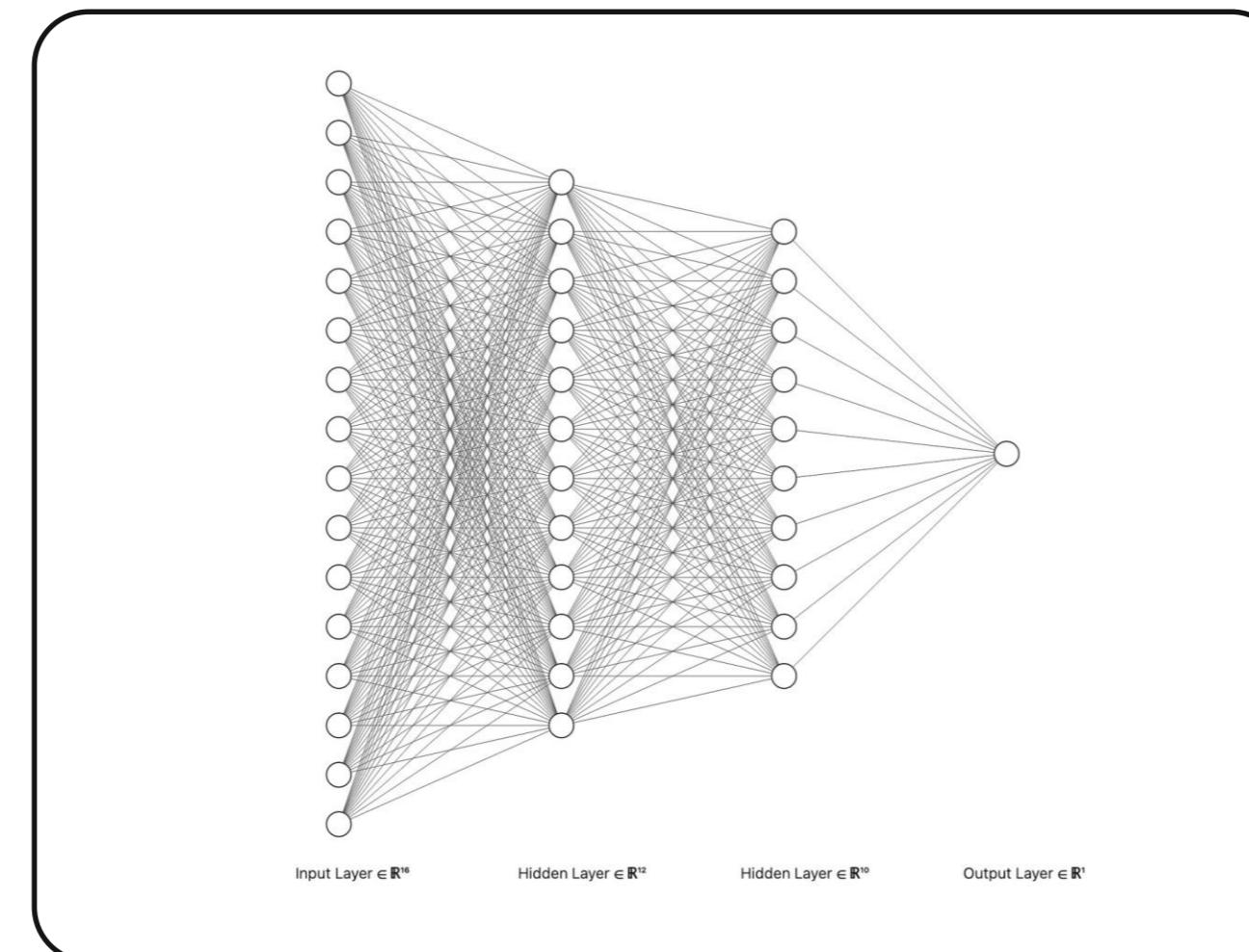
Подготовка
данных

Dates	Store ID	Product ID	Product name	# of Sales
02-12-2021	001	RP01	Almond milk	21
10-12-2021	005	RS21	Oat milk	15
18-01-2022	004	RK32	Hazelnut milk	9

MLP

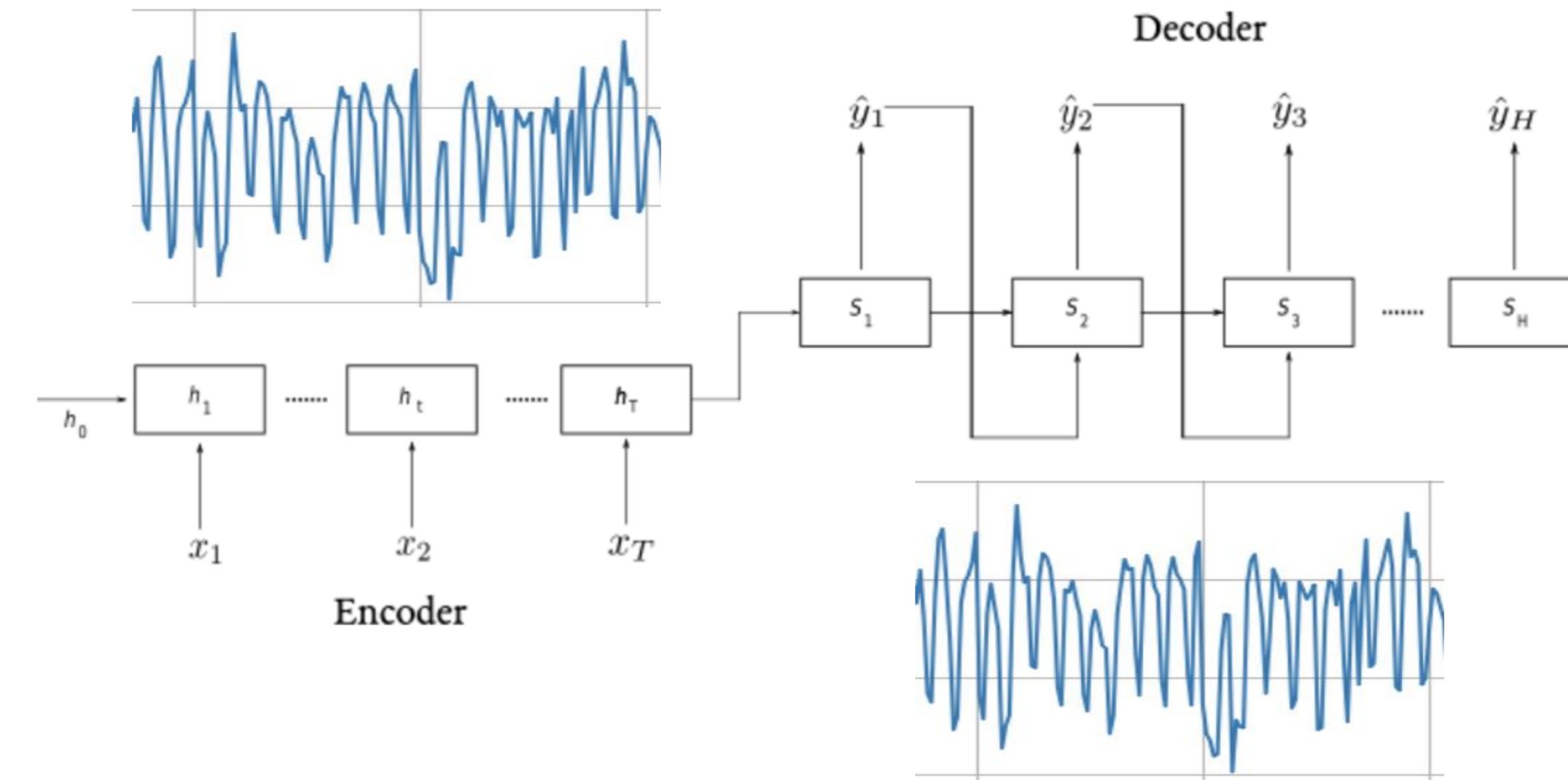
RNN

Модель



Рекуррентные нейронные сети RNN

Обучение: teacher forcing.



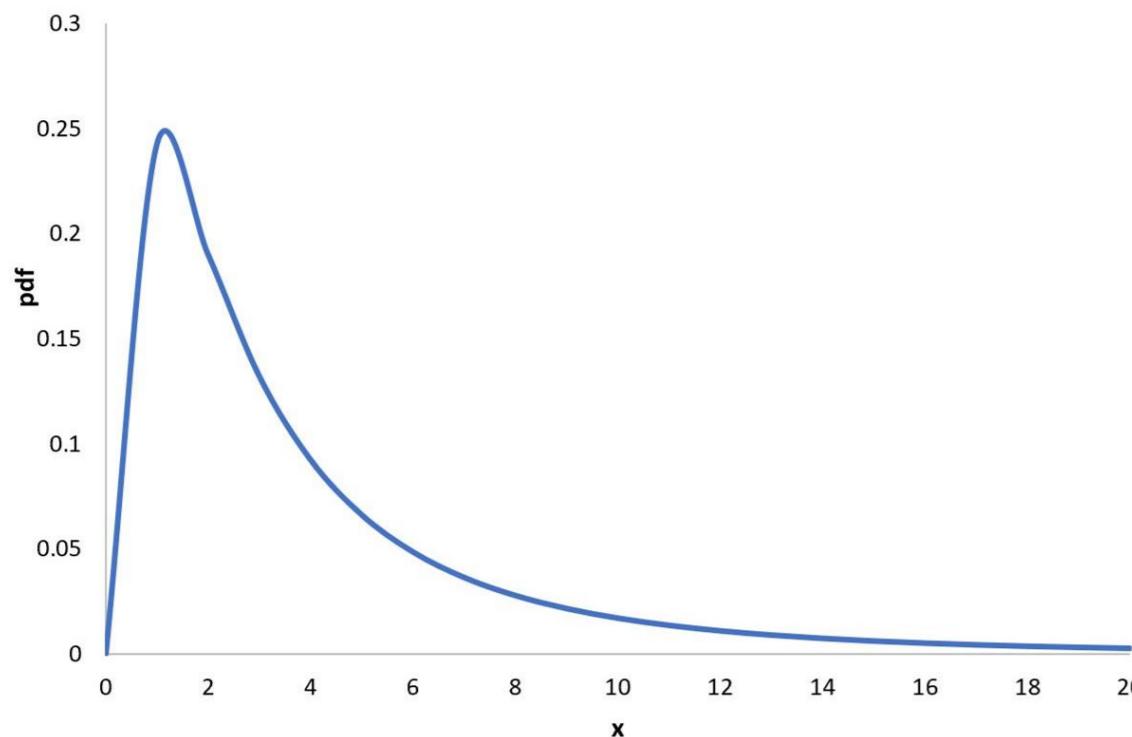
Инференс: авторегрессионно
предсказываем значения.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

DeepAR

Отличия от стандартного RNN

1. **Sampling:** боремся со скошенностью распределения.



2. **Scaling:** боремся с разницей в масштабах.

Прямое преобразование (таргет)

$$\nu_i = 1 + \frac{1}{t_0} \sum_{t=1}^{t_0} z_{i,t},$$

Обратное преобразование (параметры)

$$\mu = \nu_i \log(1 + \exp(o_\mu))$$

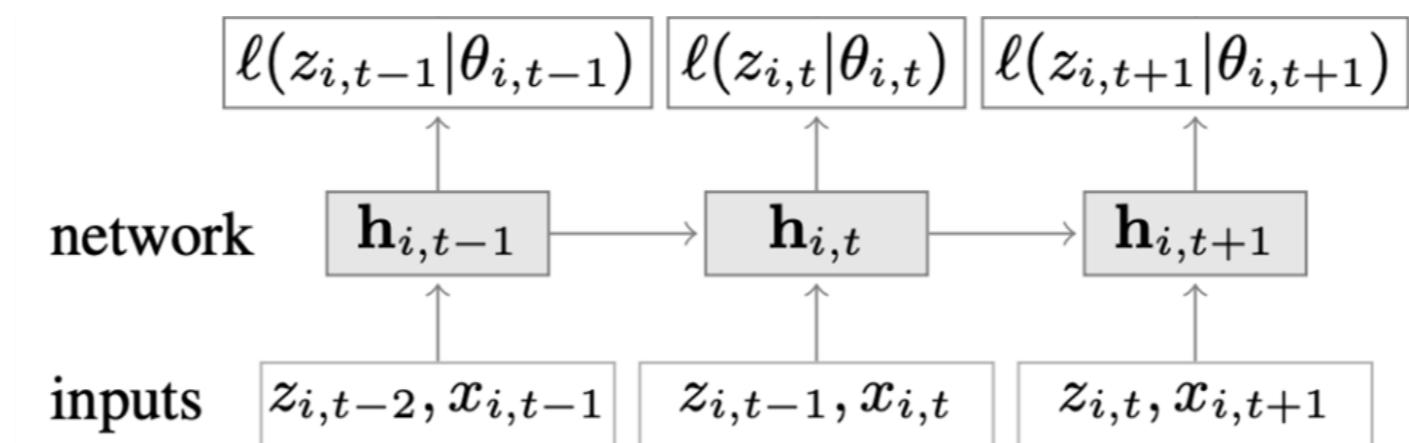
3. **Probabilistic:** предсказываем параметры распределения, а не значения ряда.

$$\ell(z_{i,t} | \theta(\mathbf{h}_{i,t}, \Theta))$$

DeepAR

Обучение

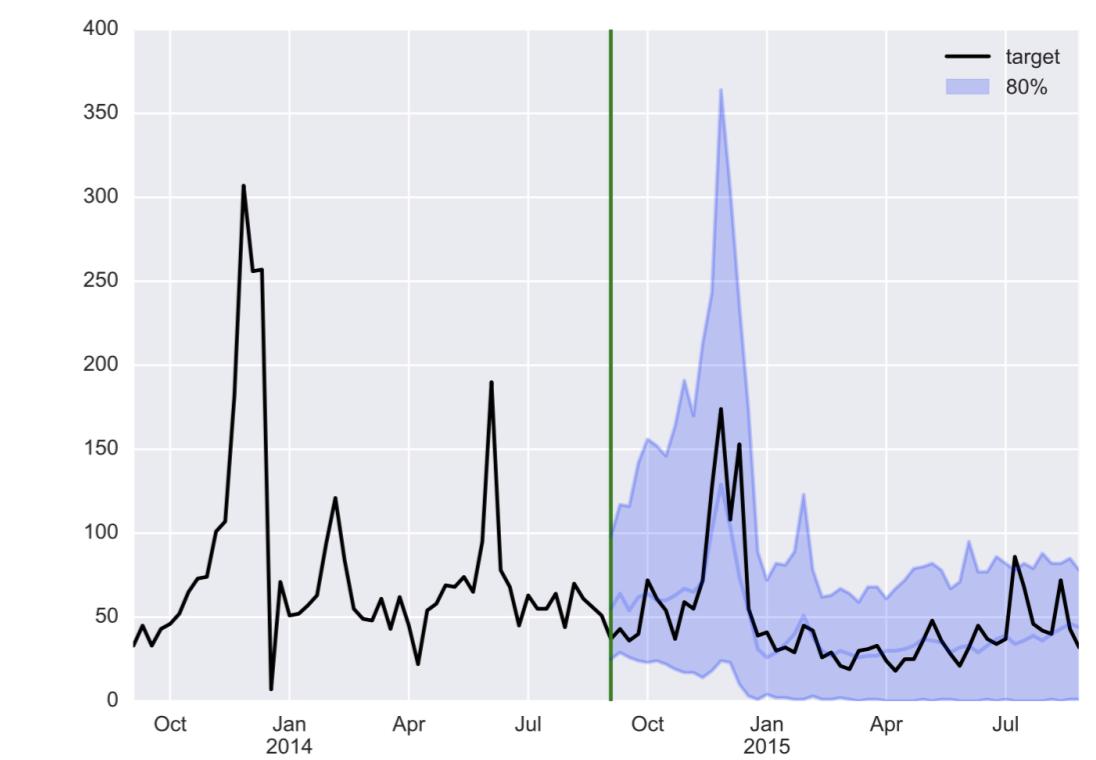
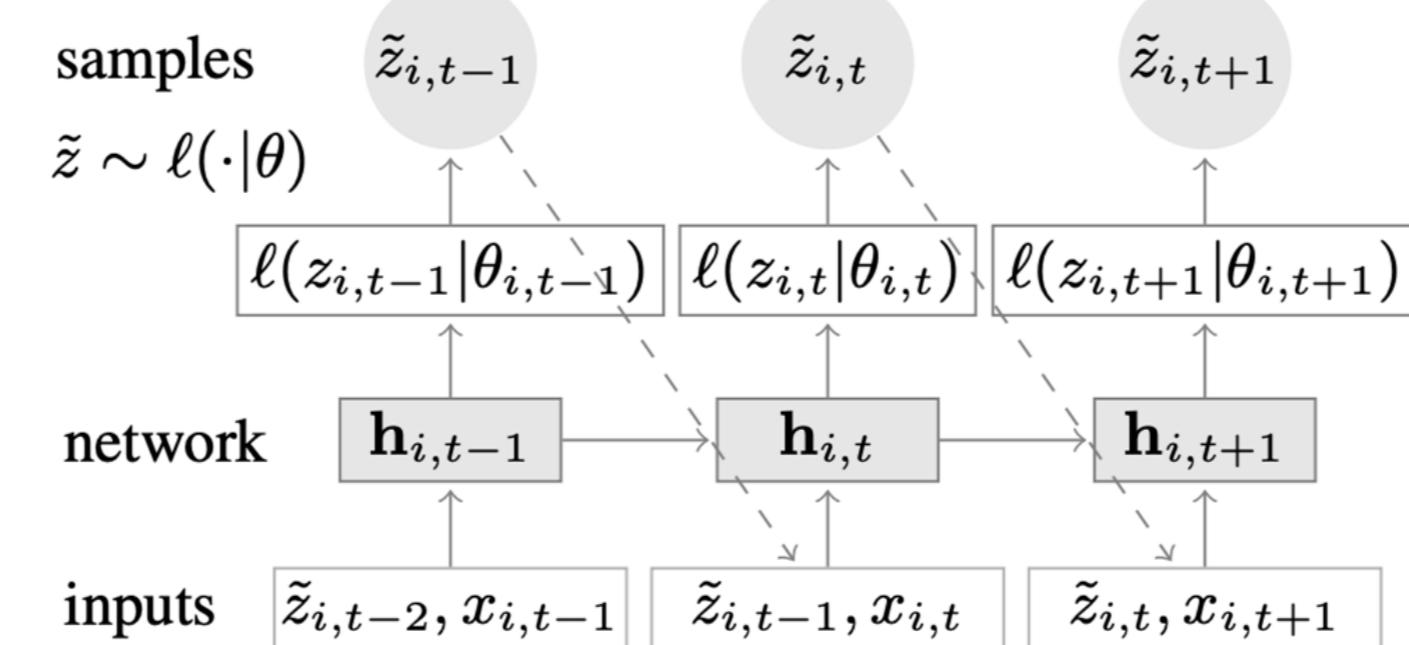
Предсказываем
параметры
распределения.



$$\mathcal{L} = \sum_{i=1}^N \sum_{t=t_0}^T \log \ell(z_{i,t} | \theta(\mathbf{h}_{i,t}))$$

Инференс

Авторегрессионно
семплируем
из распределения.

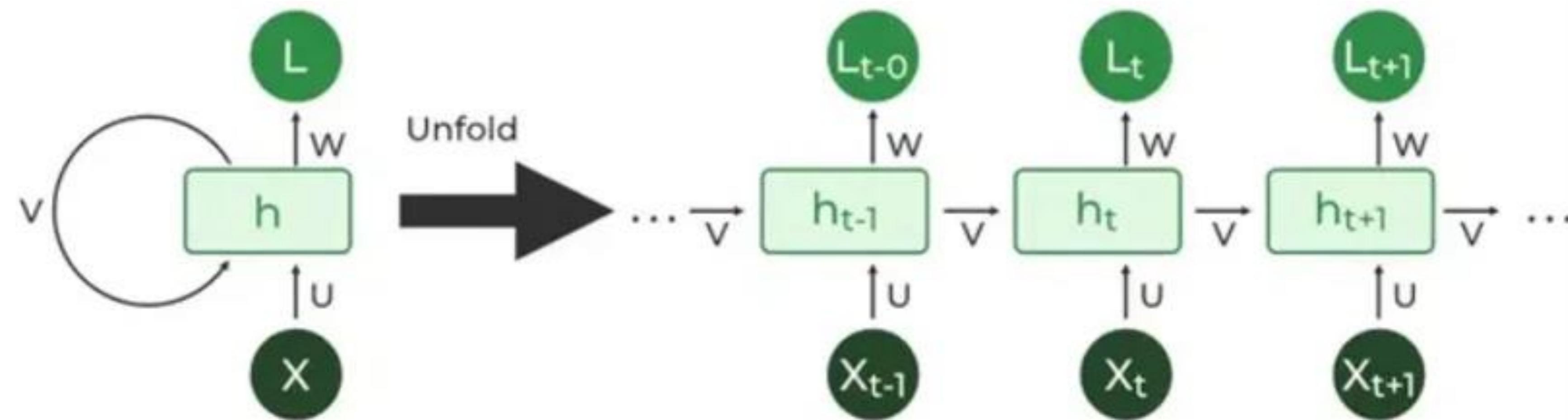


RNN, GRU, LSTM, а может быть, что-то ещё



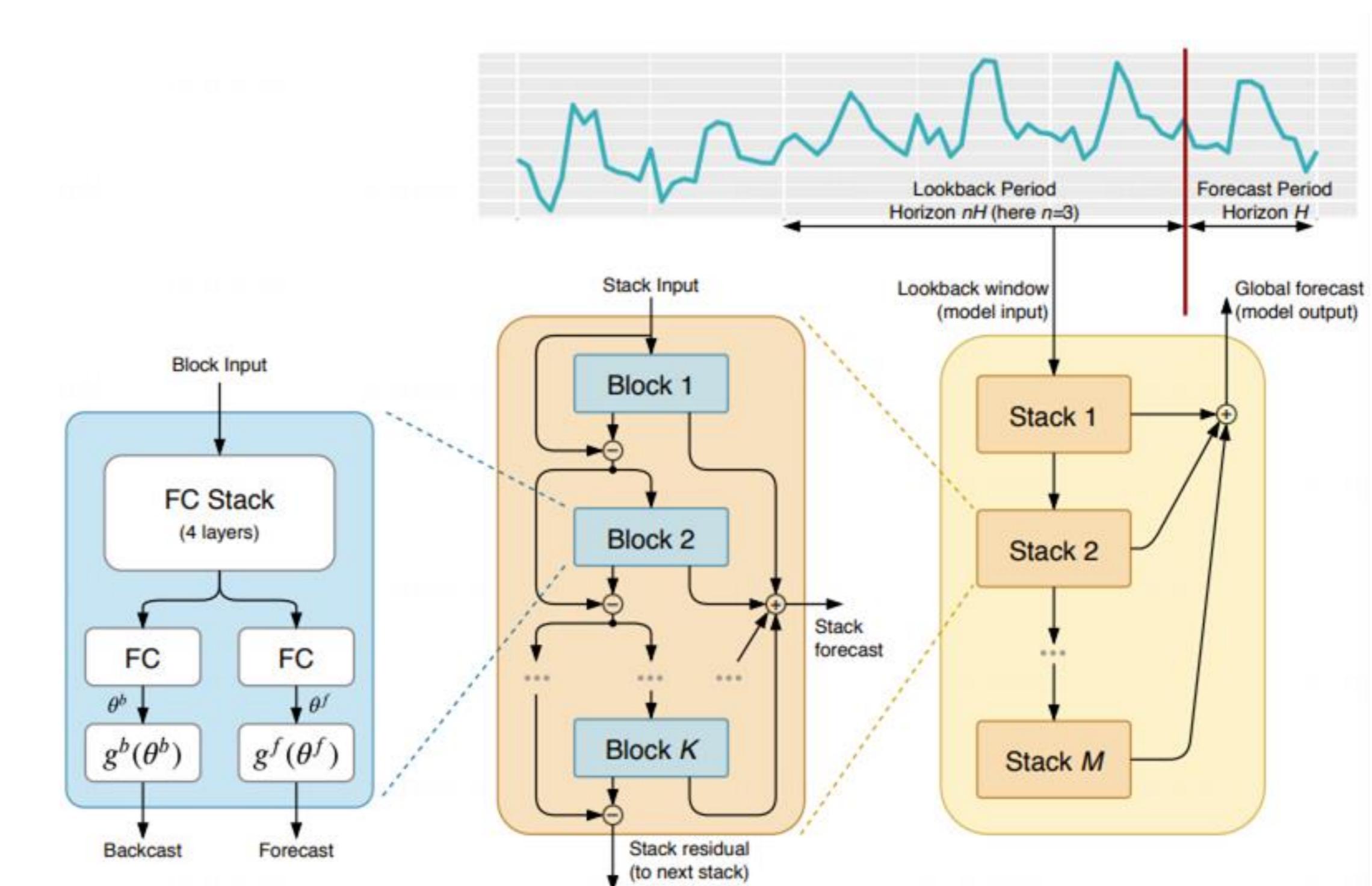
Классические RNN — линейные в смысле сложности, в отличие от трансформеров, но медленные, так как содержат нелинейность в скрытом состоянии, и мы не можем легко посчитать выход: выход считается рекуррентно — в этом контексте говорят, что RNN не параллелируется.

Mamba глобально решает проблему: она линейна, в отличие от Transformer, и параллелируется, в отличие от RNN. Идея красавая, но это не работает в реальных кейсах.



N-BEATS

- Архитектура основана на остатках прогнозирования и residual-связях сети.
- Основным блоком является полно связанный слой.
- Использование базисов для построения backcast/forecast.
 - generic $\hat{y}_\ell = \mathbf{V}_\ell^f \theta_\ell^f + \mathbf{b}_\ell^f, \quad \hat{\mathbf{x}}_\ell = \mathbf{V}_\ell^b \theta_\ell^b + \mathbf{b}_\ell^b$
 - trend $\hat{y}_{s,\ell} = \sum_{i=0}^p \theta_{s,\ell,i} t^i$
 - seasonality $\hat{y}_{s,\ell} = \sum_{i=0}^{[H/2-1]} \theta_{s,\ell,i}^f \cos(2\pi i t) + \theta_{s,\ell,i+[H/2]}^f \sin(2\pi i t)$
- Интерпретируемость прогнозов (в случае интерпретируемых базисов).



N-BEATS: интерпретируемость

- Использование базисов для построения backcast/forecast.
 - generic $\hat{\mathbf{y}}_\ell = \mathbf{V}_\ell^f \theta_\ell^f + \mathbf{b}_\ell^f, \quad \hat{\mathbf{x}}_\ell = \mathbf{V}_\ell^b \theta_\ell^b + \mathbf{b}_\ell^b$
 - trend $\hat{\mathbf{y}}_{s,\ell} = \sum_{i=0}^p \theta_{s,\ell,i} t^i$
 - seasonality $\hat{\mathbf{y}}_{s,\ell} = \sum_{i=0}^{\lfloor H/2-1 \rfloor} \theta_{s,\ell,i}^f \cos(2\pi i t) + \theta_{s,\ell,i+\lfloor H/2 \rfloor}^f \sin(2\pi i t)$

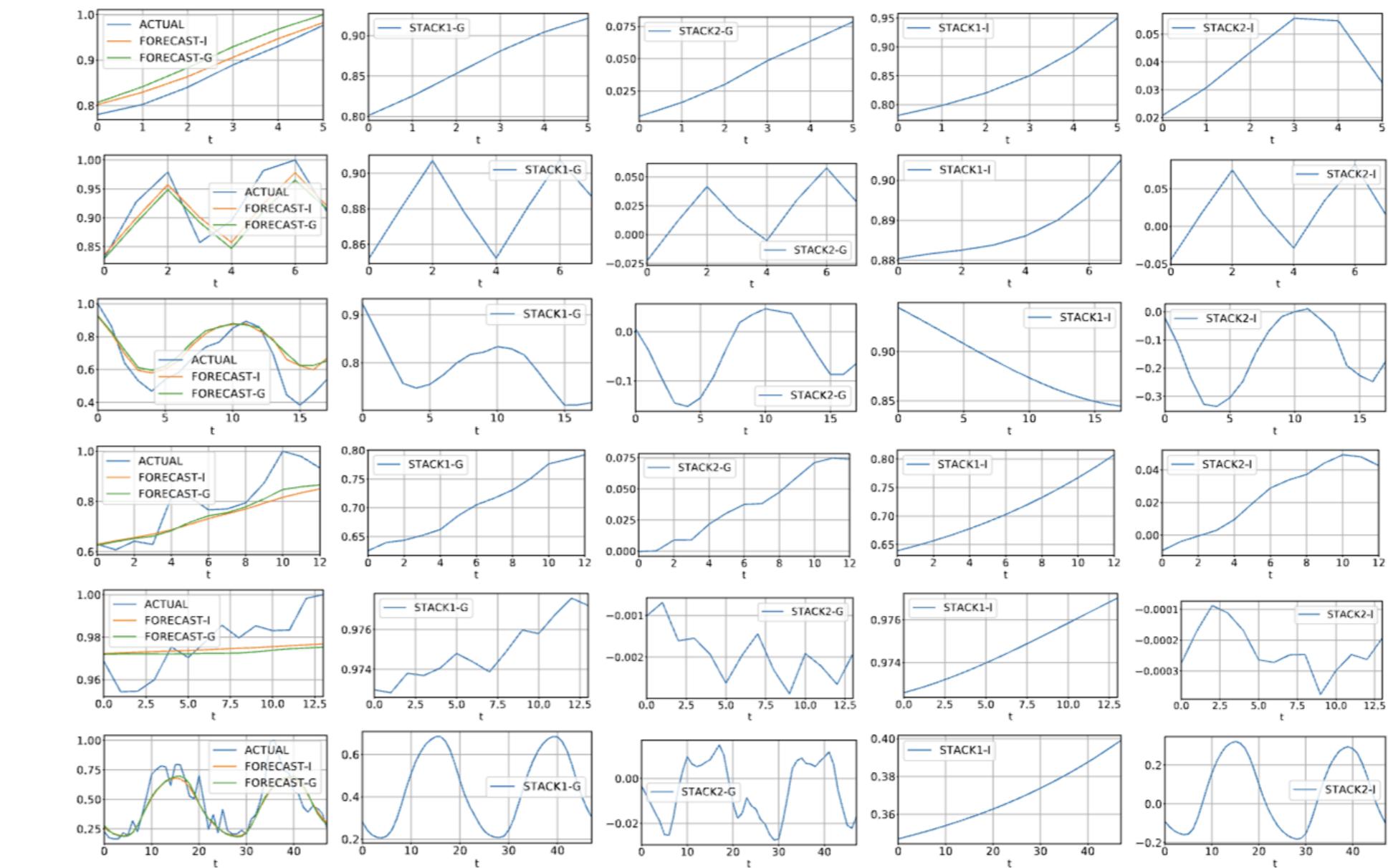
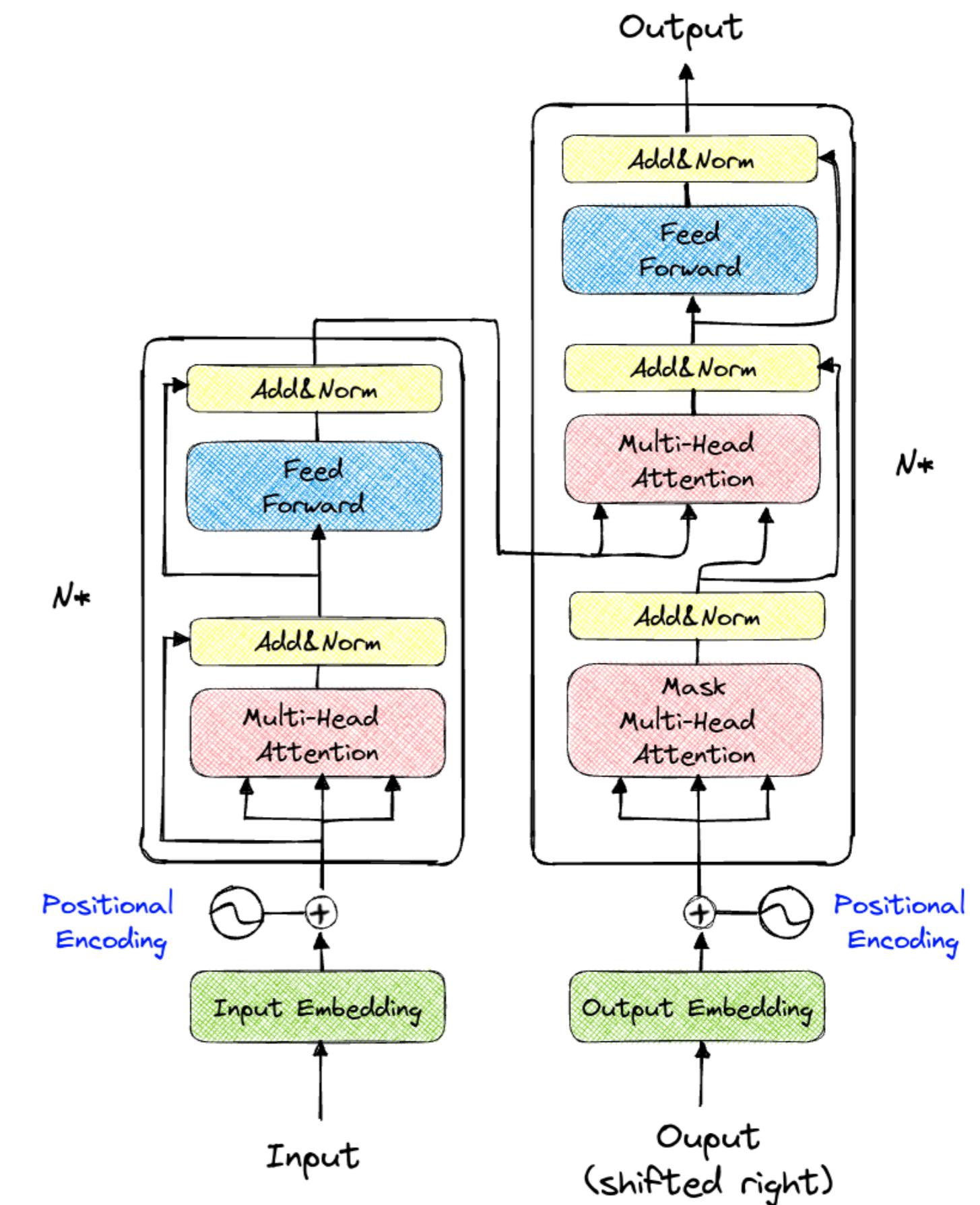


Figure 5: The outputs of generic and the interpretable configurations, M4 dataset. Each row is one time series example per data frequency, top to bottom (Yearly: id Y3974, Quarterly: id Q11588, Monthly: id M19006, Weekly: id W246, Daily: id D404, Hourly: id H344). The magnitudes in a row are normalized by the maximal value of the actual time series for convenience. Column (a) shows the actual values (ACTUAL), the generic model forecast (FORECAST-G) and the interpretable model forecast (FORECAST-I). Columns (b) and (c) show the outputs of stacks 1 and 2 of the generic model, respectively; FORECAST-G is their summation. Columns (d) and (e) show the output of the Trend and the Seasonality stacks of the interpretable model, respectively; FORECAST-I is their summation.

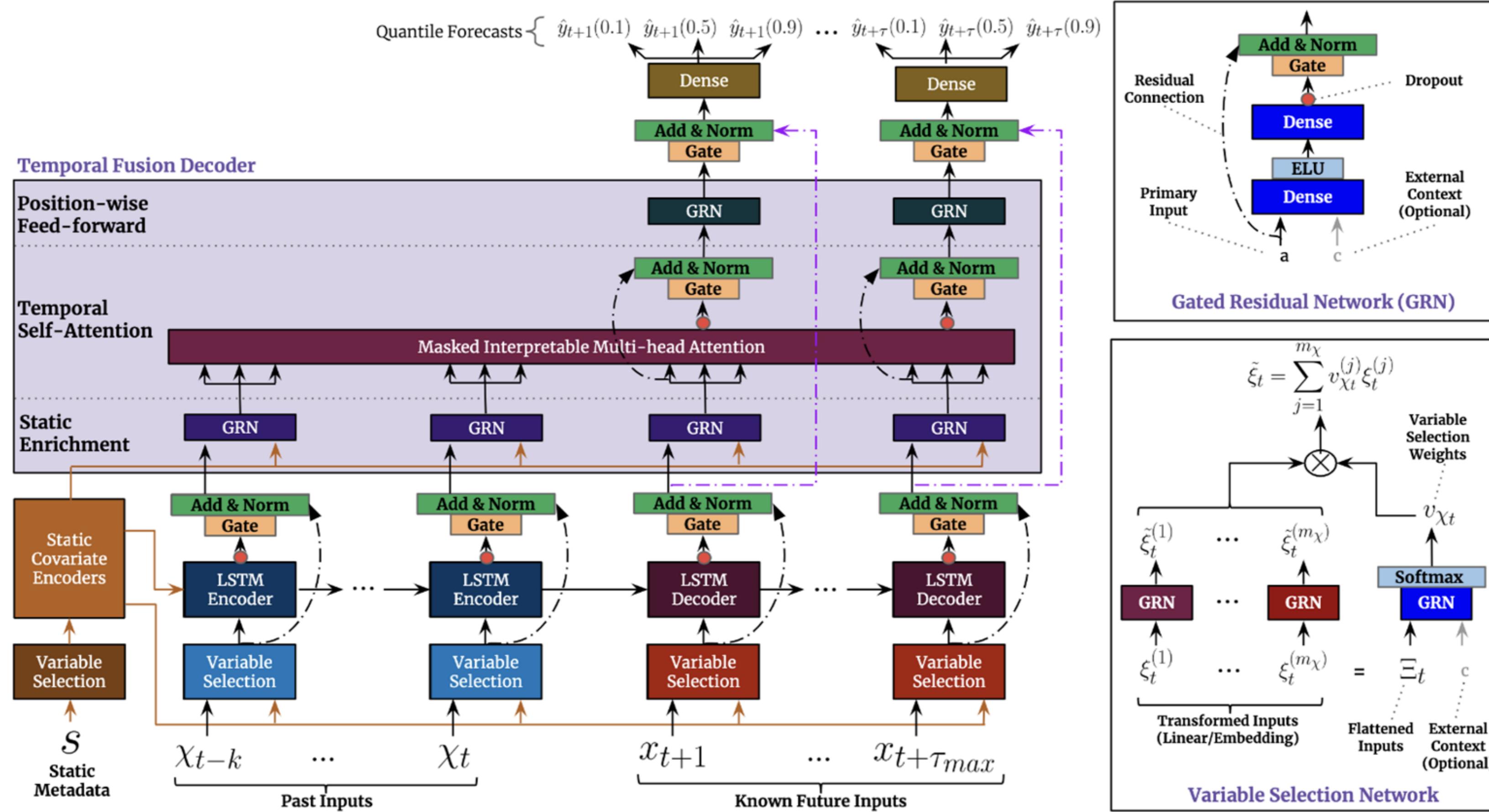
Transformers



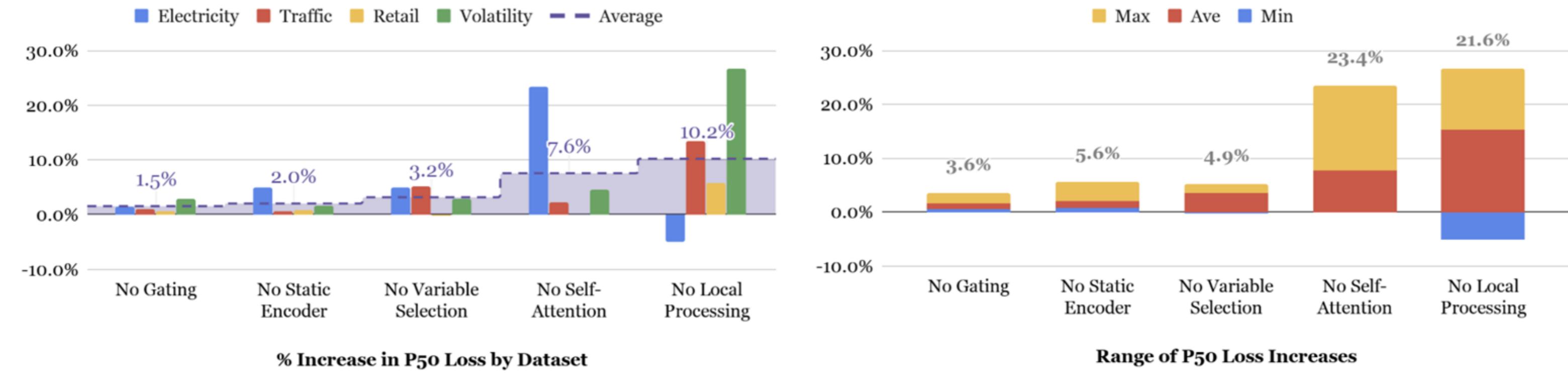
- Позволяет использовать большой контекст.
- В основном используется для предсказания на шаг вперёд в авторегрессионном режиме, но можно и предсказывать вектор сразу.
- Attention можно использовать для интерпретации модели.
- Пример: Temporal Fusion Transformer.



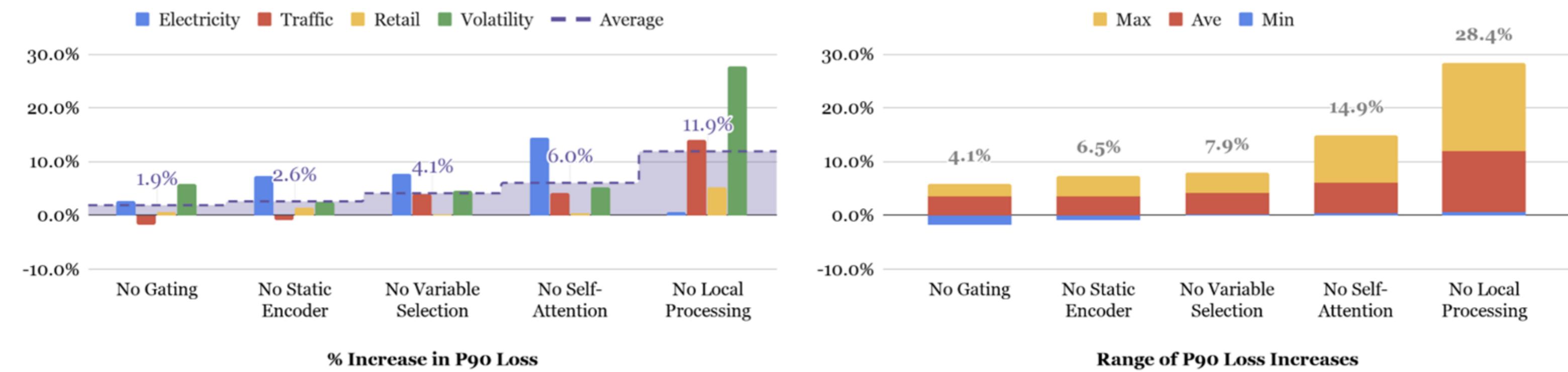
Temporal Fusion Transformer



Temporal Fusion Transformer



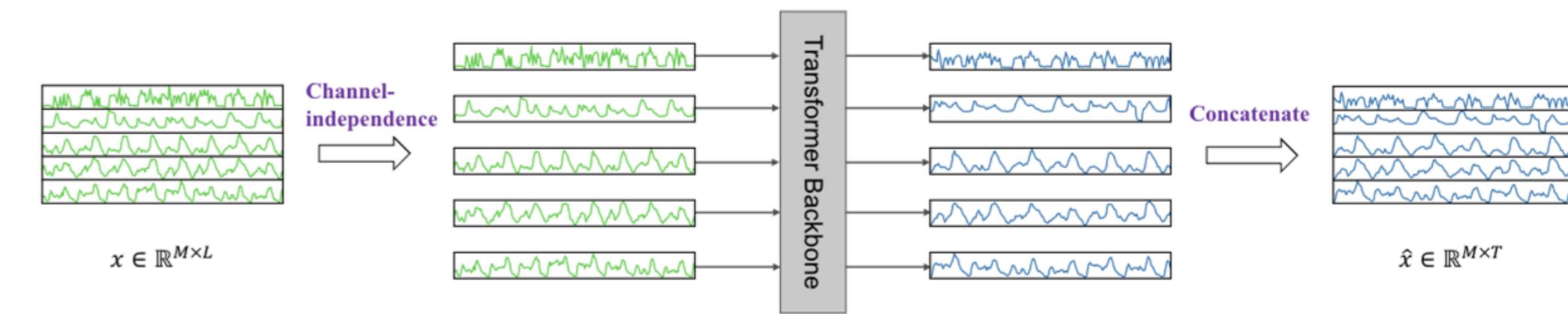
(a) Changes in P50 losses across ablation tests



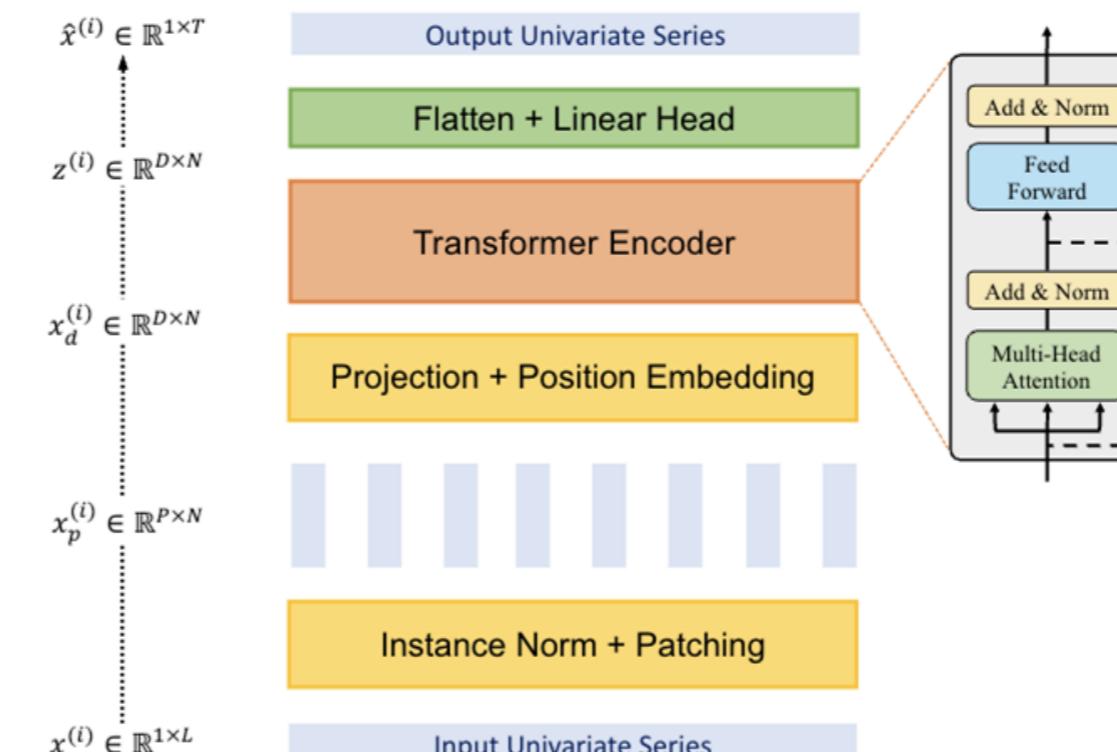
(b) Changes in P90 losses across ablation tests

PatchTST

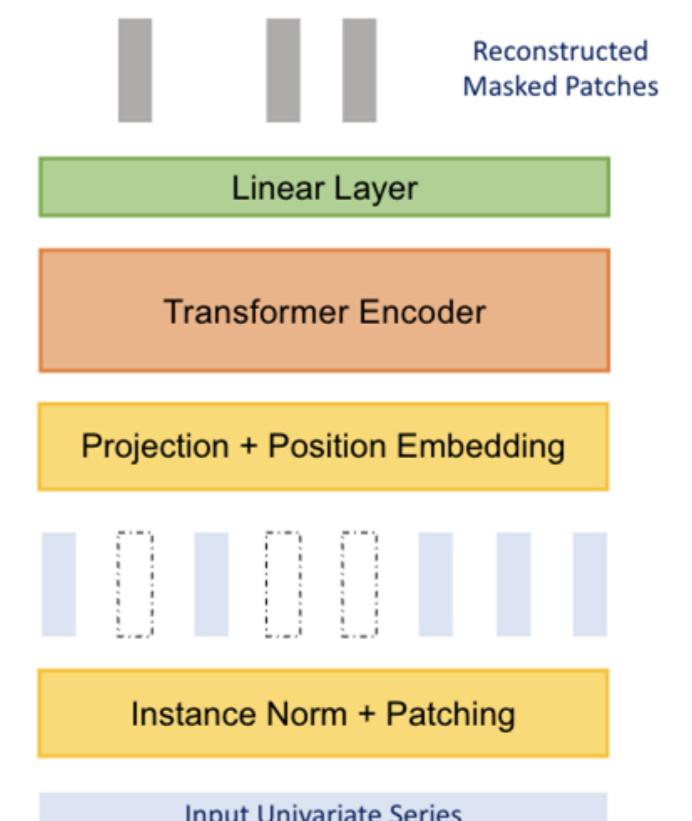
- Patching: мы с ним ещё встретимся.
- Каналы рассматриваются независимо.
- Patching позволяет работать с большими контекстами и более стабильно предсказывать на большие горизонты.



(a) PatchTST Model Overview



(b) Transformer Backbone (Supervised)



(c) Transformer Backbone (Self-supervised)



Foundation Models

Transformers и его друзья

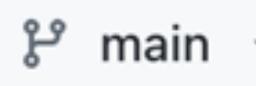
Много статей о применении разных вариаций трансформеров на 5.5-датасетах (Informer, Autoformer, Reformer, FEDFormer, ETSFormer, Crossformer и т. д.).

Models	LSTMa		LSTnet		MTGNN		Transformer		Informer		Autoformer		Pyraformer		FEDformer		Crossformer		
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
ETTh1	24	0.650	0.624	1.293	0.901	0.336	0.393	0.620	0.577	0.577	0.549	0.439	0.440	0.493	0.507	0.318	0.384	0.305	0.367
	48	0.720	0.675	1.456	0.960	0.386	0.429	0.692	0.671	0.685	0.625	0.429	0.442	0.554	0.544	0.342	0.396	<u>0.352</u>	0.394
	168	1.212	0.867	1.997	1.214	0.466	0.474	0.947	0.797	0.931	0.752	0.493	0.479	0.781	0.675	<u>0.412</u>	0.449	0.410	0.441
	336	1.424	0.994	2.655	1.369	0.736	0.643	1.094	0.813	1.128	0.873	0.509	0.492	0.912	0.747	<u>0.456</u>	0.474	0.440	0.461
	720	1.960	1.322	2.143	1.380	0.916	0.750	1.241	0.917	1.215	0.896	0.539	0.537	0.993	0.792	<u>0.521</u>	0.515	<u>0.519</u>	<u>0.524</u>
ETTm1	24	0.621	0.629	1.968	1.170	<u>0.260</u>	<u>0.324</u>	0.306	0.371	0.323	0.369	0.410	0.428	0.310	0.371	0.290	0.364	0.211	0.293
	48	1.392	0.939	1.999	1.215	<u>0.386</u>	<u>0.408</u>	0.465	0.470	0.494	0.503	0.485	0.464	0.465	0.464	<u>0.342</u>	0.396	0.300	0.352
	96	1.339	0.913	2.762	1.542	0.428	0.446	0.681	0.612	0.678	0.614	0.502	0.476	0.520	0.504	<u>0.366</u>	0.412	0.320	0.373
	288	1.740	1.124	1.257	2.076	0.469	0.488	1.162	0.879	1.056	0.786	0.604	0.522	0.729	0.657	0.398	0.433	<u>0.404</u>	0.427
	672	2.736	1.555	1.917	2.941	0.620	0.571	1.231	1.103	1.192	0.926	0.607	0.530	0.980	0.678	0.455	0.464	<u>0.569</u>	<u>0.528</u>
WTH	24	0.546	0.570	0.615	0.545	0.307	<u>0.356</u>	0.349	0.397	0.335	0.381	0.363	0.396	<u>0.301</u>	0.359	0.357	0.412	0.294	0.343
	48	0.829	0.677	0.660	0.589	0.388	0.422	0.386	0.433	0.395	0.459	0.456	0.462	<u>0.376</u>	<u>0.421</u>	0.428	0.458	0.370	0.411
	168	1.038	0.835	0.748	0.647	<u>0.498</u>	<u>0.512</u>	0.613	0.582	0.608	0.567	0.574	0.548	<u>0.519</u>	<u>0.521</u>	0.564	0.541	0.473	0.494
	336	1.657	1.059	0.782	0.683	<u>0.506</u>	<u>0.523</u>	0.707	0.634	0.702	0.620	0.600	0.571	<u>0.539</u>	0.543	0.533	0.536	0.495	0.515
	720	1.536	1.109	0.851	0.757	0.510	0.527	0.834	0.741	0.831	0.731	0.587	0.570	0.547	0.553	0.562	0.557	<u>0.526</u>	<u>0.542</u>
ECL	48	0.486	0.572	0.369	0.445	<u>0.173</u>	<u>0.280</u>	0.334	0.399	0.344	0.393	0.241	0.351	0.478	0.471	0.229	0.338	0.156	0.255
	168	0.574	0.602	0.394	0.476	<u>0.236</u>	<u>0.320</u>	0.353	0.420	0.368	0.424	0.299	0.387	0.452	0.455	0.263	0.361	0.231	0.309
	336	0.886	0.795	0.419	0.477	0.328	<u>0.373</u>	0.381	0.439	0.381	0.431	0.375	0.428	0.463	0.456	0.305	0.386	<u>0.323</u>	0.369
	720	1.676	1.095	0.556	0.565	0.422	0.410	0.391	0.438	0.406	0.443	<u>0.377</u>	0.434	0.480	0.461	0.372	0.434	<u>0.404</u>	<u>0.423</u>
	960	1.591	1.128	0.605	0.599	0.471	0.451	0.492	0.550	0.460	0.548	0.366	0.426	0.550	0.489	<u>0.393</u>	0.449	0.433	<u>0.438</u>
ILI	24	4.220	1.335	4.975	1.660	4.265	1.387	3.954	1.323	4.588	1.462	3.101	1.238	3.970	1.338	2.687	1.147	<u>3.041</u>	<u>1.186</u>
	36	4.771	1.427	5.322	1.659	<u>4.777</u>	<u>1.496</u>	4.167	1.360	4.845	1.496	<u>3.397</u>	1.270	4.377	1.410	2.887	1.160	<u>3.406</u>	<u>1.232</u>
	48	4.945	1.462	5.425	1.632	5.333	1.592	4.746	1.463	4.865	1.516	<u>2.947</u>	<u>1.203</u>	4.811	1.503	2.797	1.155	3.459	1.221
	60	5.176	1.504	5.477	1.675	5.070	1.552	5.219	1.553	5.212	1.576	<u>3.019</u>	<u>1.202</u>	5.204	1.588	2.809	1.163	3.640	1.305
	720	4.668	0.378	0.648	0.403	<u>0.506</u>	<u>0.278</u>	0.597	0.332	0.608	0.334	0.550	0.363	0.606	0.338	0.562	0.375	0.491	0.274
Traffic	48	0.709	0.400	0.709	0.425	0.512	<u>0.298</u>	0.658	0.369	0.644	0.359	0.595	0.376	0.619	0.346	0.567	0.374	<u>0.519</u>	0.295
	168	0.900	0.523	0.713	0.435	<u>0.521</u>	<u>0.319</u>	0.664	0.363	0.660	0.391	0.649	0.407	0.635	0.347	0.607	0.385	0.513	0.289
	336	1.067	0.599	0.741	0.451	<u>0.540</u>	<u>0.335</u>	0.654	0.358	0.747	0.405	0.624	0.388	0.641	0.347	0.624	0.389	0.530	0.300
	720	1.461	0.787	0.768	0.474	0.557	<u>0.343</u>	0.685	0.370	0.792	0.430	0.674	0.417	0.670	0.364	0.623	0.378	<u>0.573</u>	<u>0.313</u>



Transformers и их скептики

 **Transformers_Are_What_You_Dont_Need** Public

 main  1 Branch  0 Tags

 Go to file  Add file  Code

 Watch 11  Fork 15  Star 361

About

The best repository showing why transformers might not be the answer for time series forecasting and showcasing the best SOTA non transformer models.

 Readme
 Activity
 361 stars
 11 watching
 15 forks
Report repository

Releases

Transformers_Are_What_You_Dont_Need

The best repository showing why transformers don't work in time series forecasting

Transfer Learning

→ **Небольшие компании**

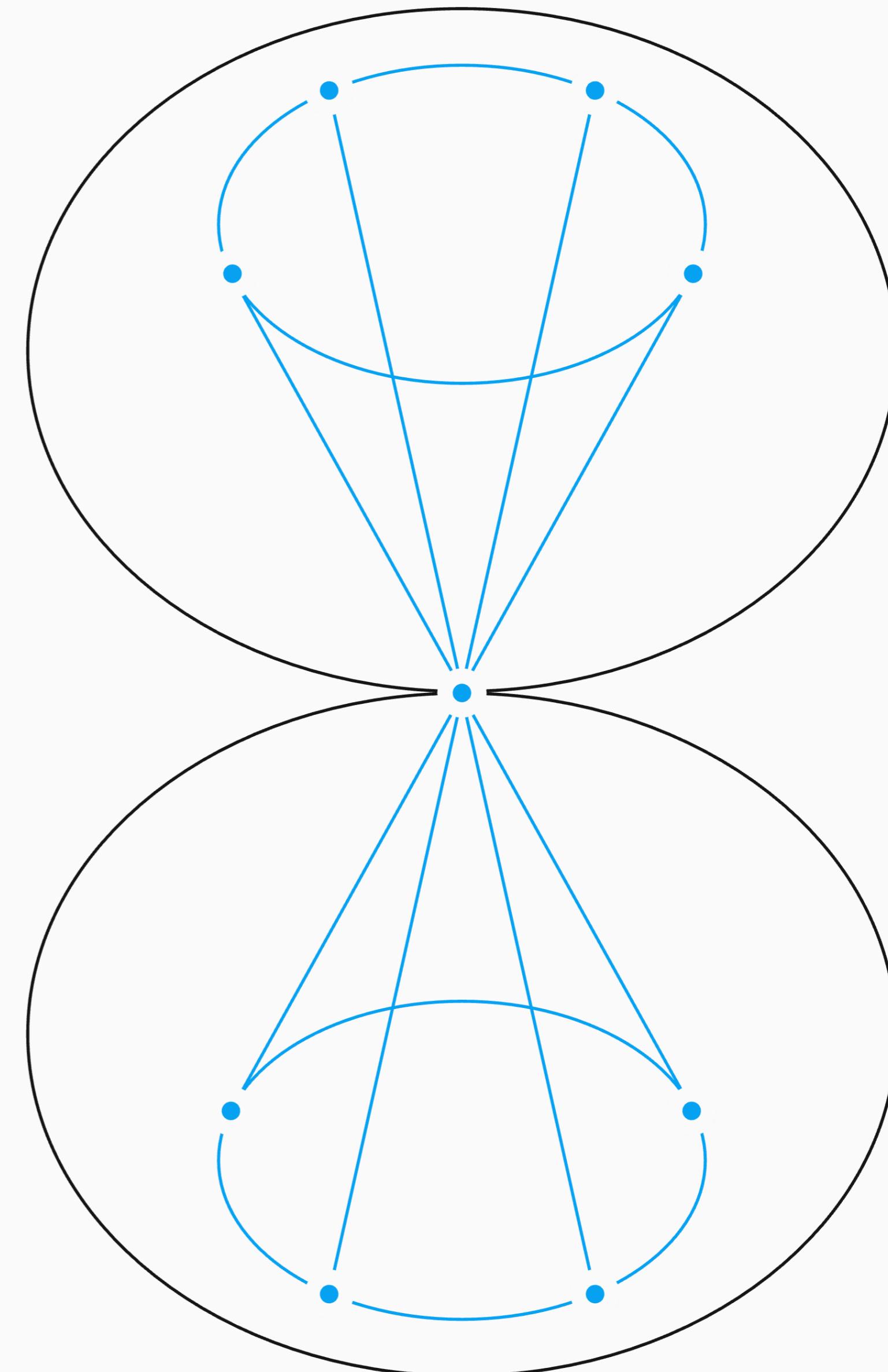
Данных мало. Бизнес-метрики имеют месячную/квартальную частотность. Как правило, всё решается правилом большого пальца.

→ **Большие компании**

Данных может быть очень много: метрики бизнес-процессов, логи, метрики сервисов и перформанса.

→ **Взаимосвязь процессов**

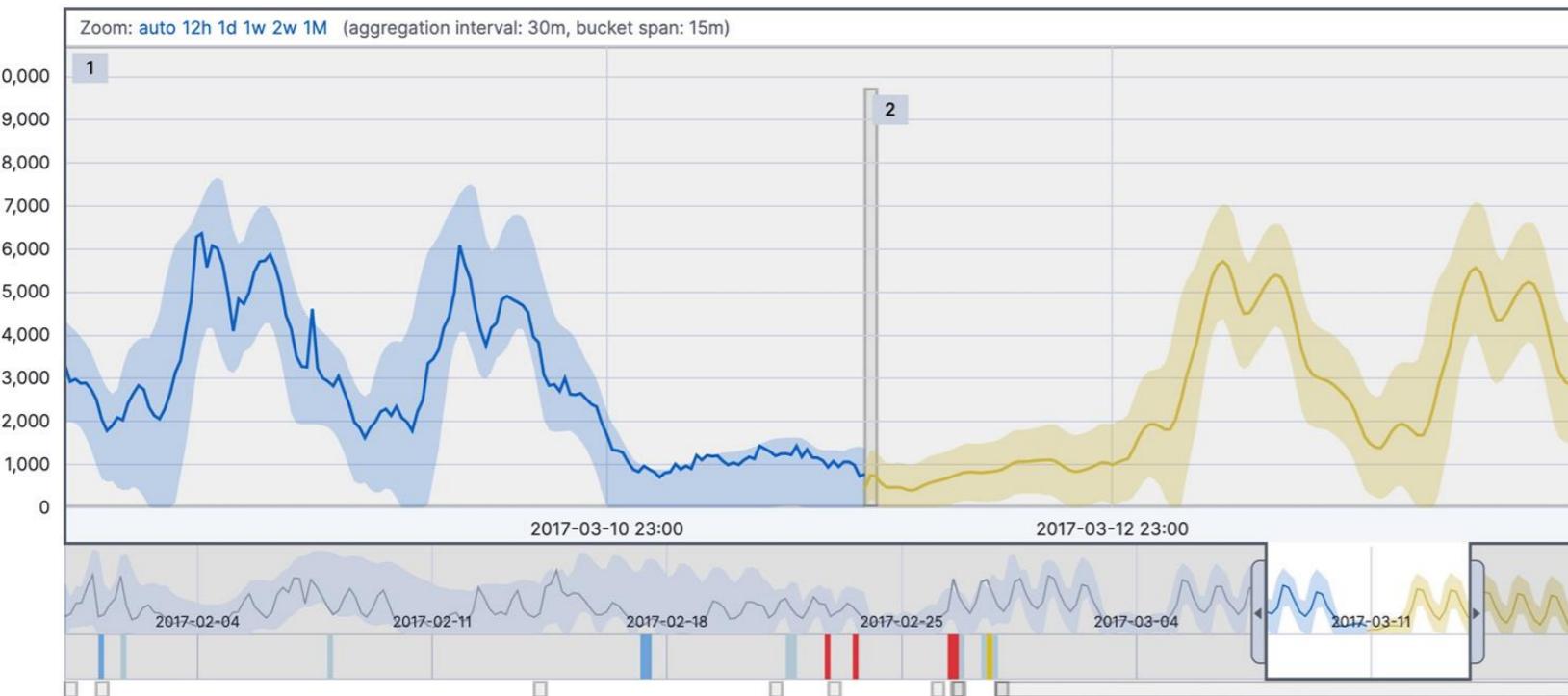
Бизнес-процессы могут быть сильно связаны между собой: акции, рекламные компании и т. д. Всё это должно повышать качество прогнозов.



Zero-shot-предсказания

BI-тулы и репортинг

Бизнес-метрики рендерятся с готовыми прогнозами либо предсказательными интервалами.



Системы мониторинга

Системы как Kibana, New Relic, DataDog и другие идут в API, и для них не нужно поддерживать фоновые джобы для обучения моделей.

Ad-hoc-аналитика

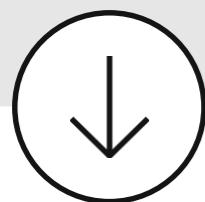
Делаем прогнозирование на уровне SQL-запроса в вашем Jupyter Notebook — вам не нужны ML-инженеры.



Toto: Time Series Optimized
Transformer for Observability

MLOps

Одна модель для всех задач



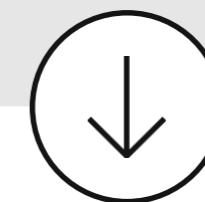
Мониторинг

У вас один сервис: вам не нужно следить за зоопарком моделей. Следите только за ключевыми метриками вашего продукта.



Процесс разработки модели

У вас разработкой занимается одна команда, либо вы вообще используете сторонние API.



Внедрение моделей

Вы вызываете один запрос: вам не нужно думать про переобучение, завозить Airflow и т. д.

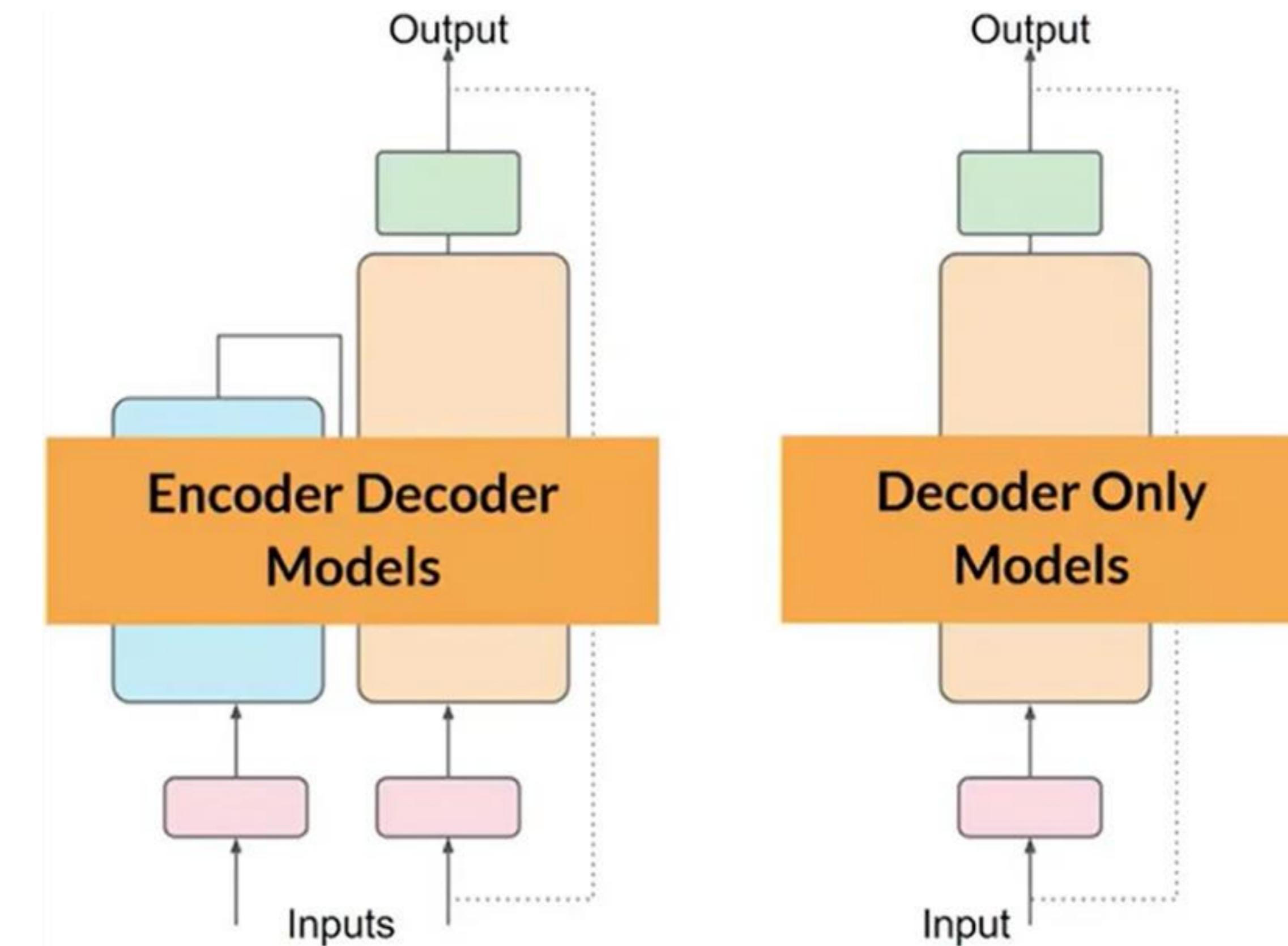


Timeline

-
- The diagram illustrates a timeline of AI model developments, numbered 01 through 09, connected by a vertical blue line with downward-pointing arrows.
- 01 TimeGPT-1 – Сентябрь 2023 – Nixtla – Closed
 - 02 LagLlama – Октябрь 2023 – Opened
 - 03 PreDct (TimesFM) – Октябрь 2023 – Google – Opened
 - 04 LLMTTime – Октябрь 2023 – NYU – Mixed
 - 05 Chronos – Март 2024 – Amazon – Opened
 - 06 Moirai – Март 2024 – Salesforce – Opened
 - 07 TimesFM релиз весов – Май 2024
 - 08 YINGLONG – Май 2025 – Alibaba
 - 09 TiRex – Май 2025 – NXAI

Архитектура

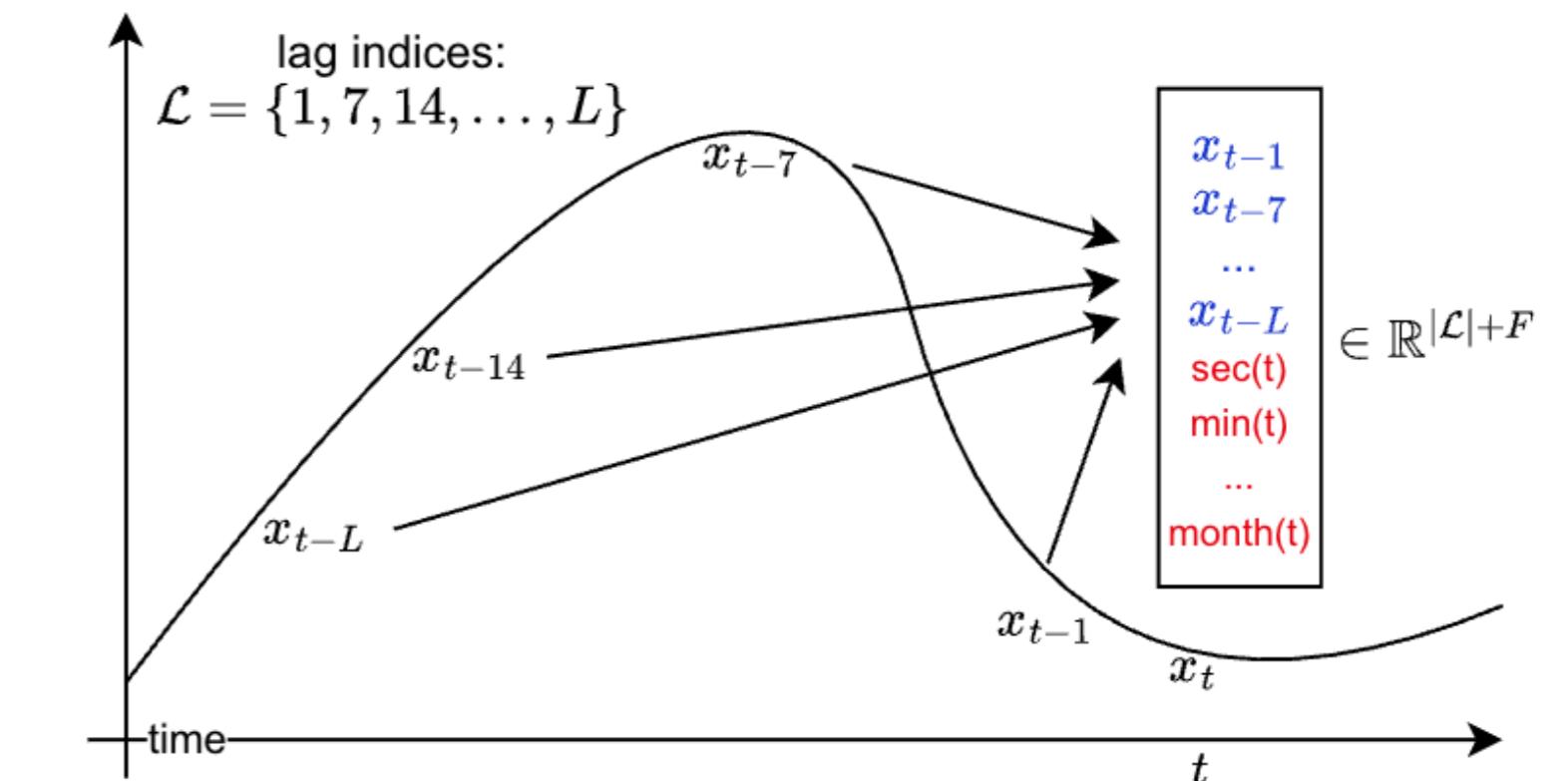
- **Decoder only Transformer**
TimesFM, LLMTIME, Lag-Llamma
- **Encoder-Decoder Transformer**
Chronos, TimeGPT-1
- **Encoder only Transformer**
Moirai, YINGLONG



Токенизация

→ **Лаги**

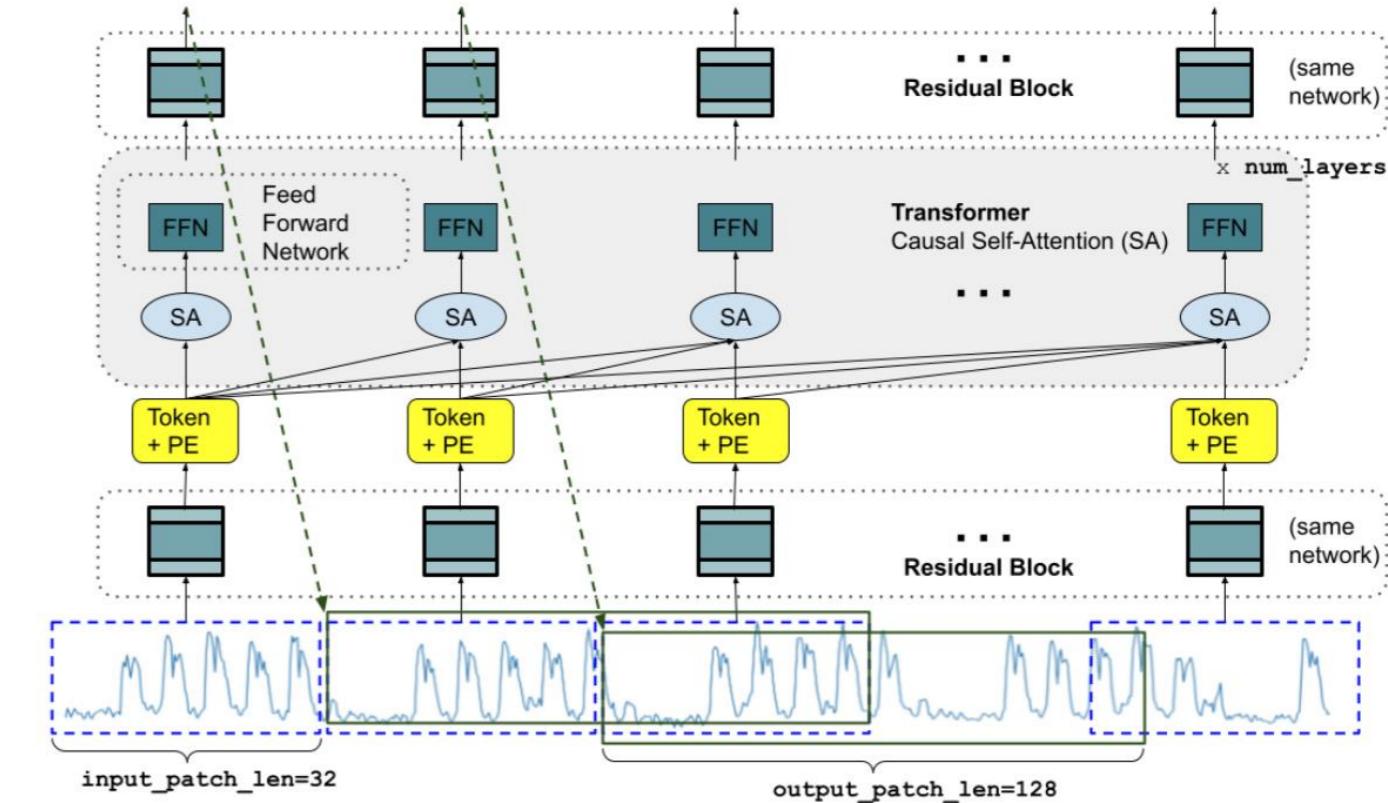
Lag-Lamma



→ **Патчи**

Moirai: длина патча динамически зависит от частоты.

TimesFM: большие патчи лучше для длинных горизонтов.



Токенизация LLMTime

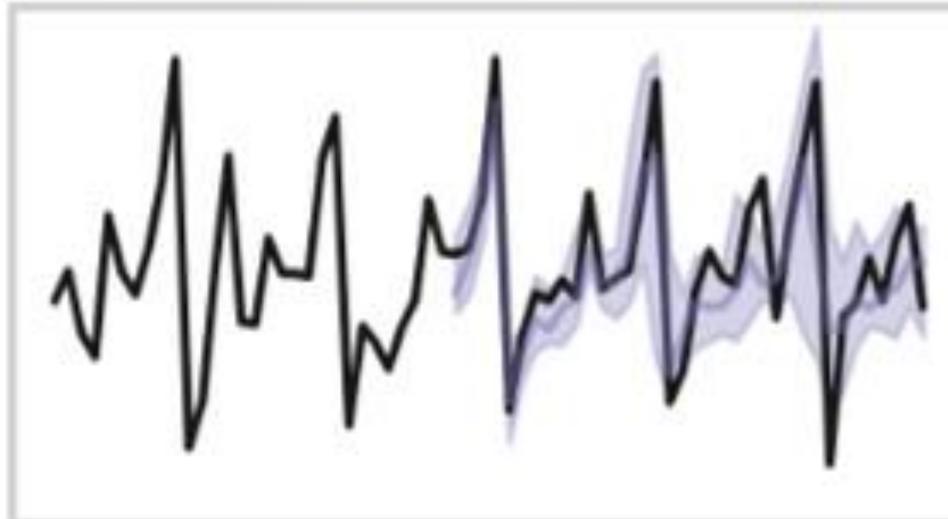
Careful tokenization is all you need.

Масштабирование на квантили позволяет избежать избыточного количества токенов.

GPT-3.5 лучше GPT-4.

"151,167,...,267"

"151,167,...,267"



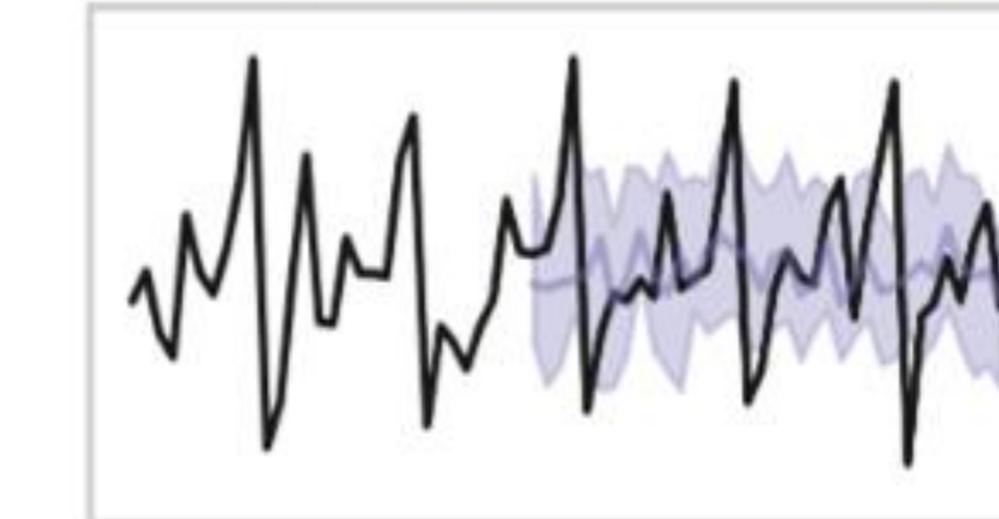
"151,167,...,267"

"151,167,...,267"



"151,167,...,267"

"151,167,...,267"



"151,167,...,267"

"151,167,...,267"



GPT-3 spaces

GPT-3 no spaces

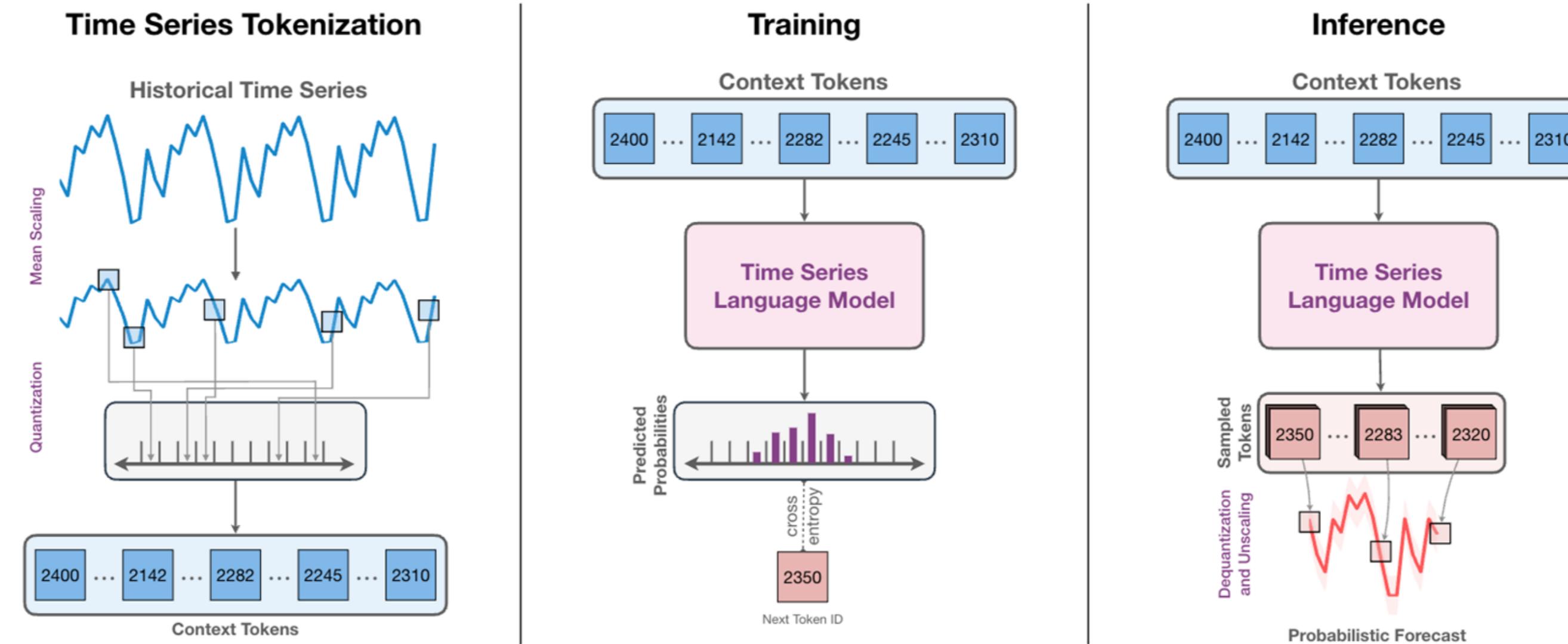
LLaMA spaces

LLaMA no spaces

Токенизация Chronos

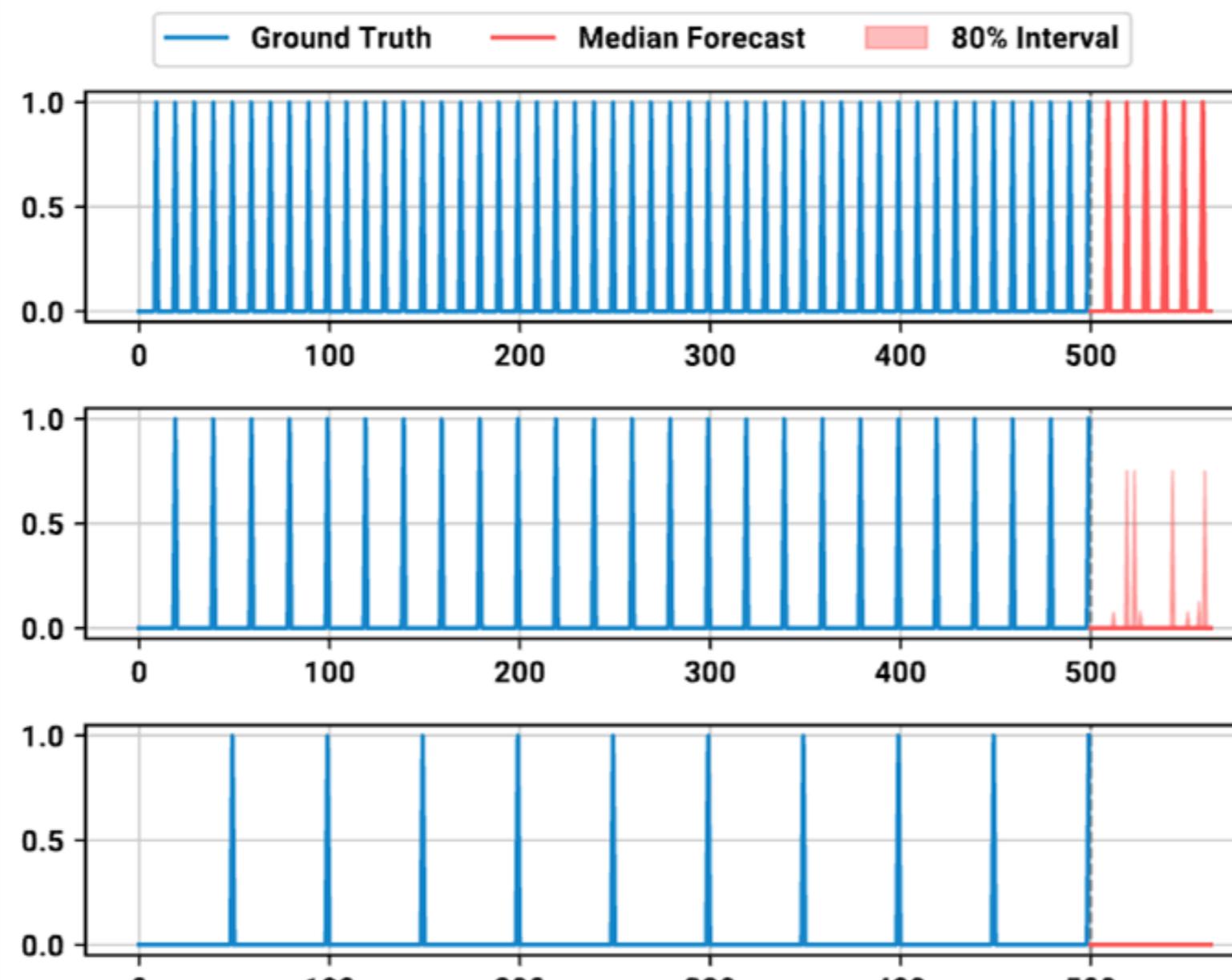
Ничего не будем делать: возьмём T5 и сделаем словарь из области значений — uniform binning.

В обновлённой версии — Chronos Bolt — сделали патчи.

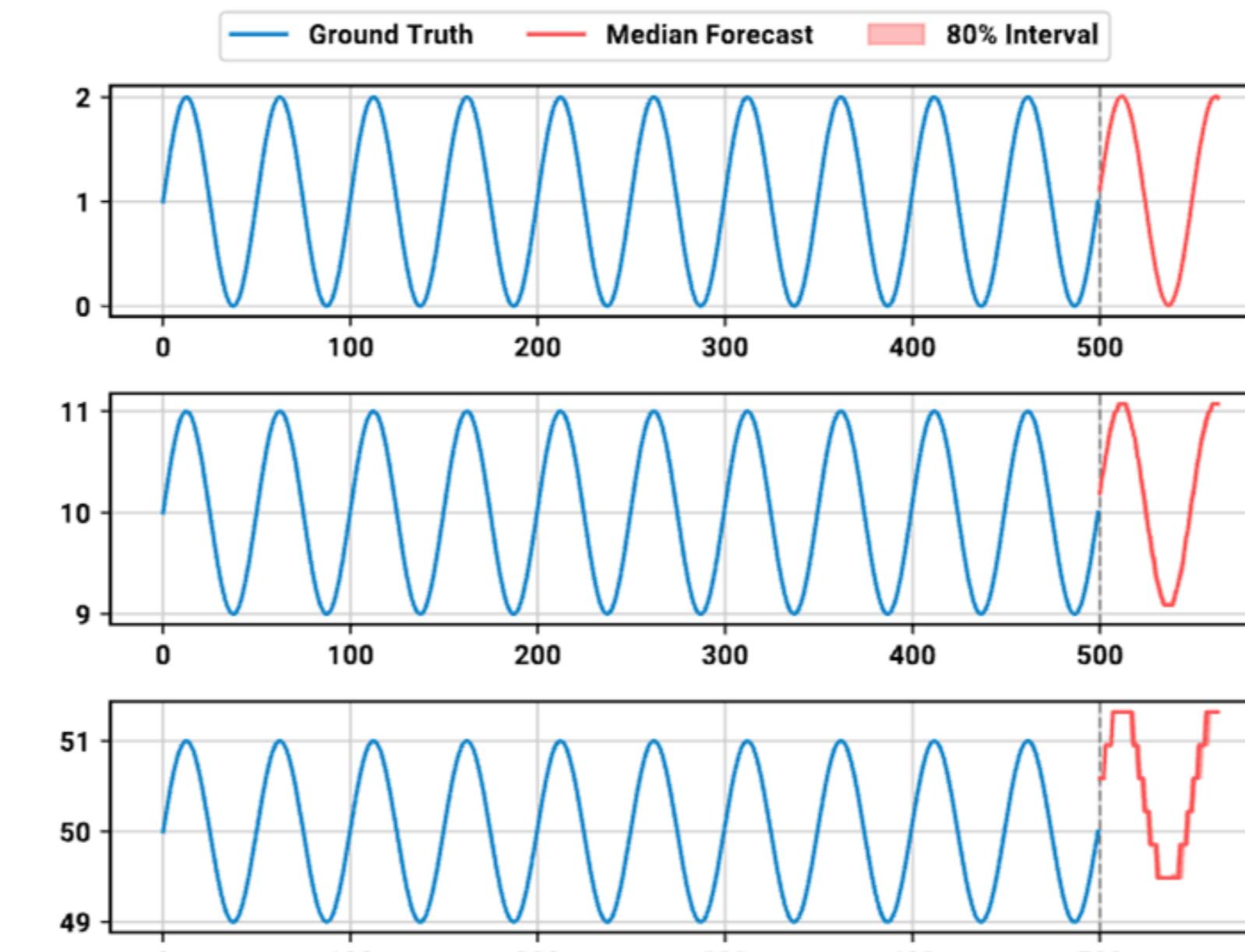


Токенизация Chronos

Синтетика и эффекты от квантизации



(a)



(b)

Moirai. Особенности

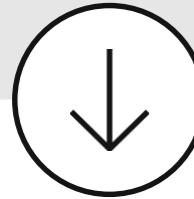
Any-Variate Attention: в один общий вход подаются и целевые данные, и регрессоры.

Длина патча динамическая: зависит от частотности ряда.

Используется packing: короткие ряды можно упаковать в один сэмпл.

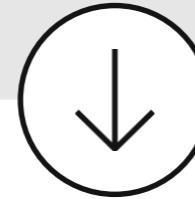


Позиционный энкодинг



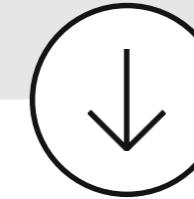
Relative Positional Encoding

Chronos переиспользует Т5 как есть, ничего не меняя и не добавляя.



RoPE + признаки времени

Lag-Llama переиспользует Llama* как есть, но в качестве признаков на вход подаются не только лаги, но и всевозможные признаки времени.



Признаки времени

Lag-Lamma, Moirai: год, месяц, день, час, минуты, секунды и т. д.

Аугментация

→ **FreqMix, FreqMask**

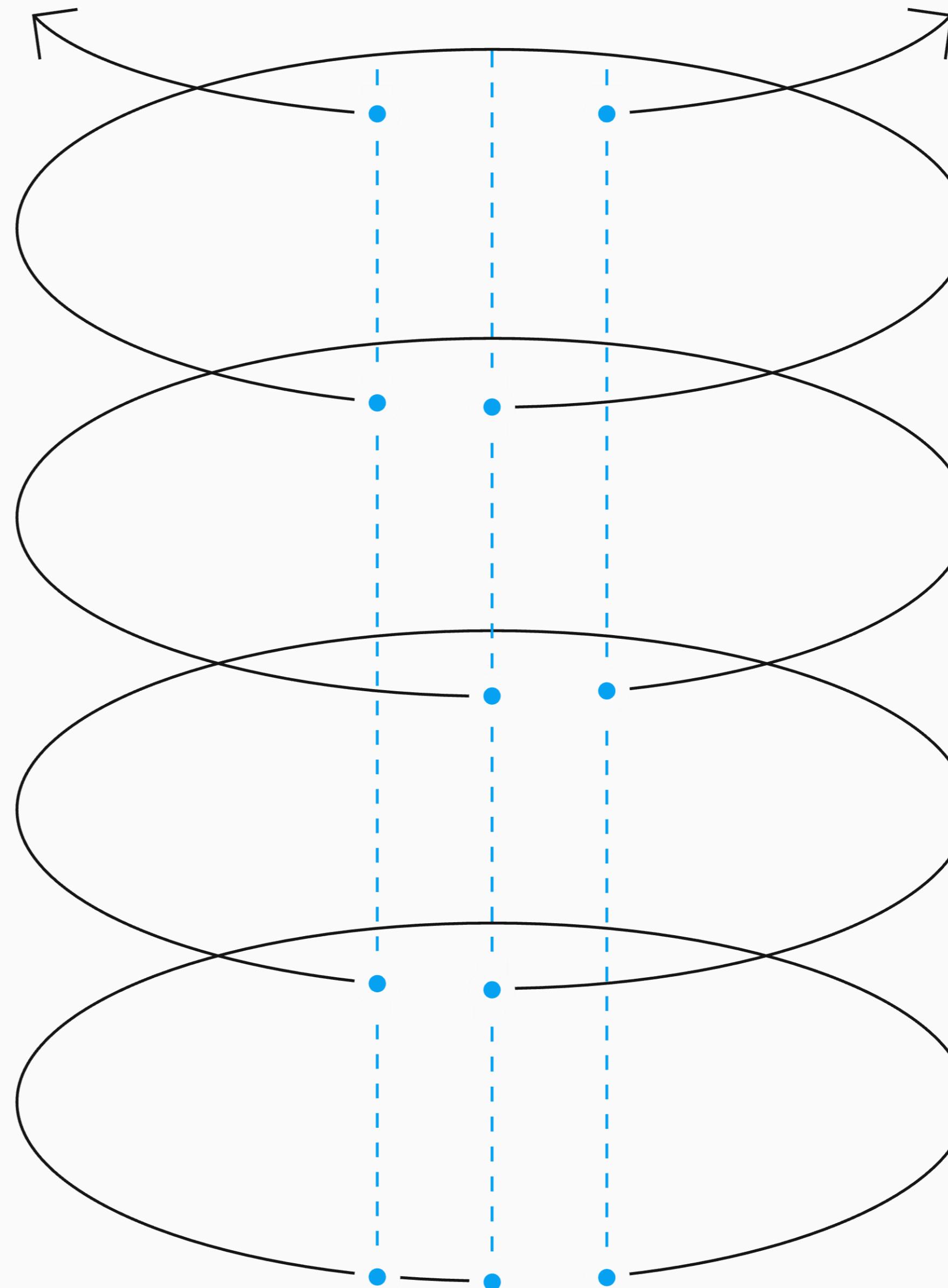
Преобразование Фурье для получения спектрального представления ряда. Часть гармоник смешиаем либо маскируем.
Проводим обратное преобразование Фурье.

→ **TSMixup**

Смешивание рядов с помощью весов из распределения Дирихле.

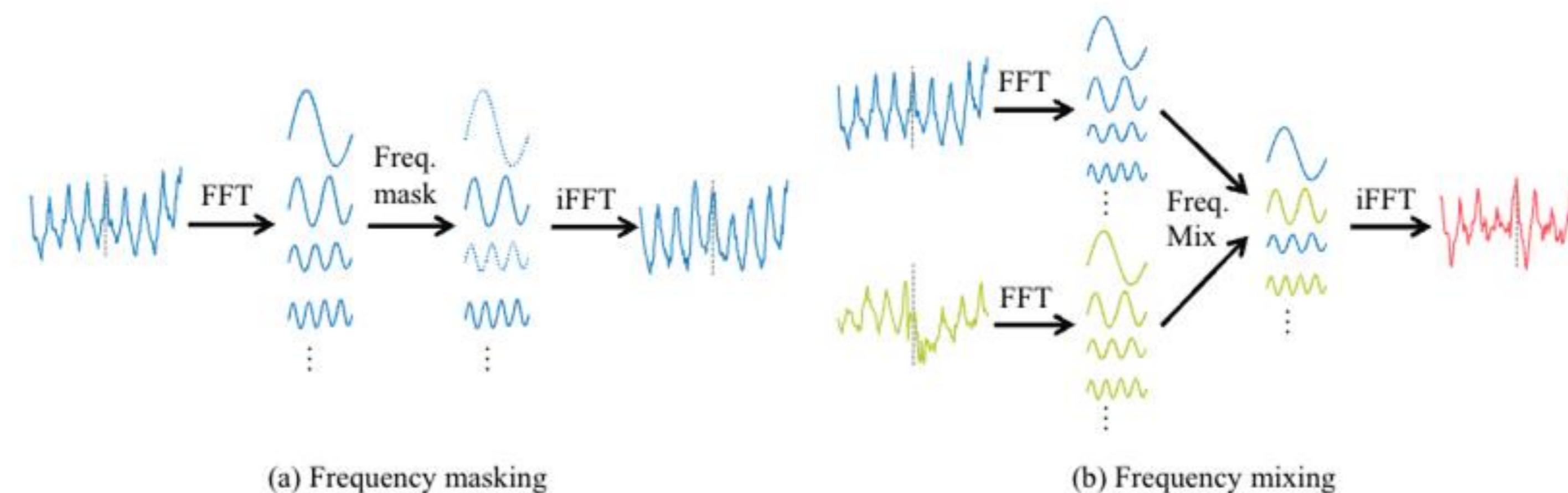
→ **Смешивание**

Смешивание целевого ряда и регрессоров на примере Moirai.

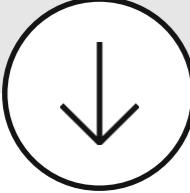


Аугментация Lag-Lamma FreqMix FreqMask

Используем преобразование Фурье для получения спектрального представления ряда. Часть гармоник смешиаем либо маскируем. Проводим обратное преобразование Фурье. Вы великолепны!

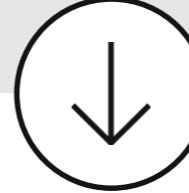


Голова и функции потерь



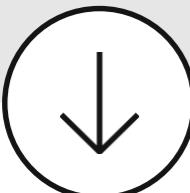
MseLoss

TimesFM и, скорее всего, TimeGPT-1 использовали обычную регрессионную функцию потерь.



t-Student distribution

Lag-Llamma позволяет делать вероятностные прогнозы из коробки.



CrossEntropyLoss

Chronos переиспользует T5 как есть и ничего не меняет и для функции потерь: мы просто учим языковую модель.



Смесь распределений

Uni2TS использовала смесь Lognormal, Negative Binomial, Low Variance Gaussian, чтобы закрыть все домены и их особенности.

Данные

Данные

→ Синтетика

ARMA-процессы, ряды с сезонностью и паттернами,
физические процессы, генеративные модели

→ Соревнования Mx

Порядка 80 миллионов временных наблюдений

→ Lotsa

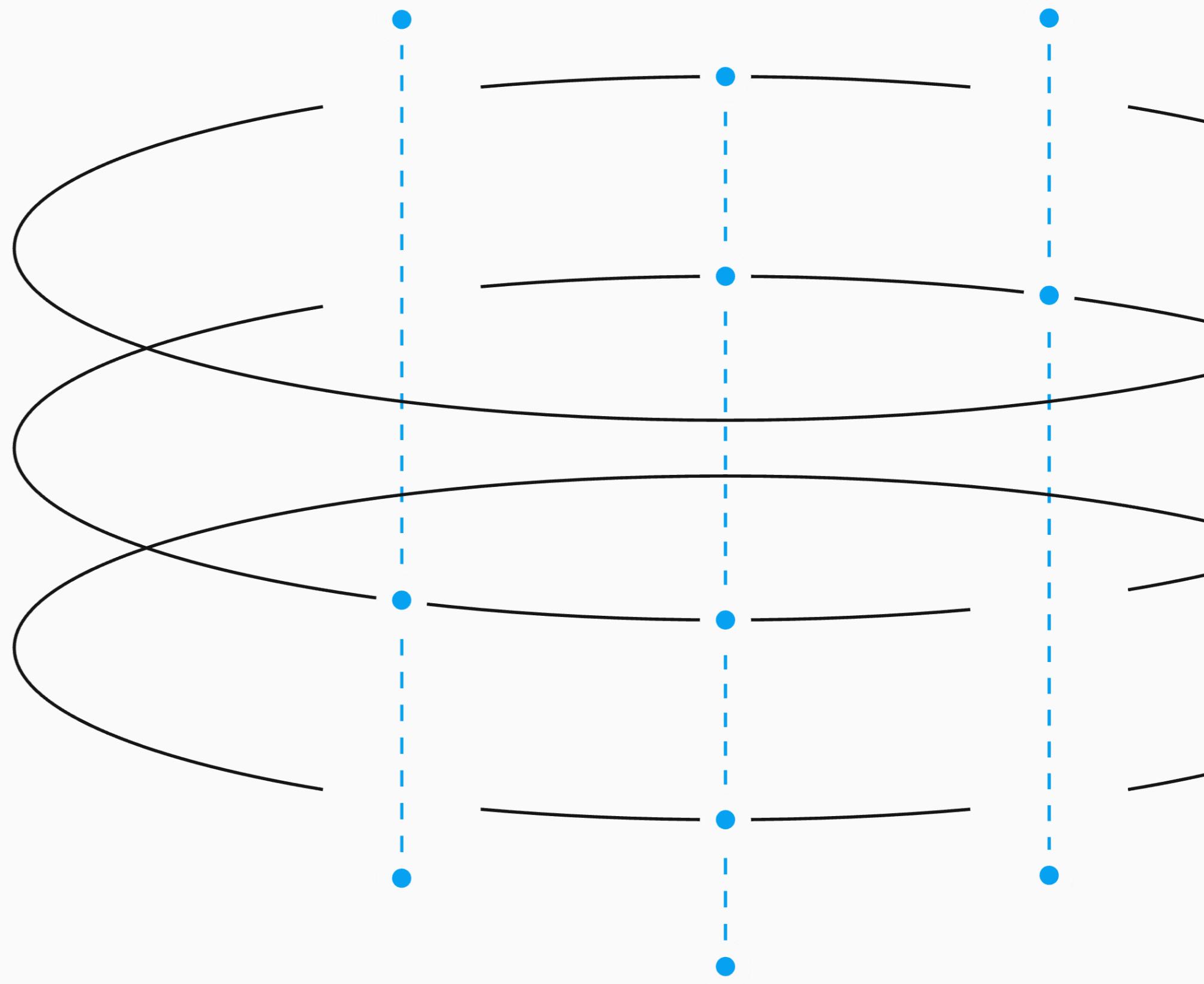
27 миллиардов временных наблюдений

→ Monash Forecasting Repository

800 миллионов наблюдений

→ Wiki Traffic

370 миллиардов наблюдений



https://forecastingdata.org

Home Aim **Datasets** Results Links About Us Contributions Acknowledgement Software

DATASETS

The following table shows a list of time series datasets that are currently available in our archive. The datasets are available in .tsf format which is a new format we propose to store time series data pioneered by sktime .ts format. The wrappers to load data into R and Python environments are available in our [github repository](#).

Dataset	Domain	No. of Series	Min. Length	Max. Length	Competition	Multivariate	Download	Source
M1	Multiple	1001	15	150	Yes	No	Yearly Quarterly Monthly	Makridakis et al., 1982
M3	Multiple	3003	20	144	Yes	No	Yearly Quarterly Monthly Other	Makridakis and Hibon, 2000
M4	Multiple	100000	19	9933	Yes	No	Yearly Quarterly Monthly Weekly Daily Hourly	Makridakis et al., 2020
Tourism	Tourism	1311	11	333	Yes	No	Yearly Quarterly	Athanasopoulos et al., 2011



https://huggingface.co/datasets?task_categories=task_categories:time-series-forecasting&sort=downloads

Hugging Face Search models, datasets, users... Models Datasets Spaces Posts Docs Enterprise Pricing

Main Tasks 1 Libraries Languages Licenses Other Filter Tasks by name Reset Tasks Multimodal Visual Question Answering Video-Text-to-Text Computer Vision Depth Estimation Image Classification Object Detection Image Segmentation Text-to-Image Image-to-Text Image-to-Image Image-to-Video Unconditional Image Generation Video Classification Text-to-Video Zero-Shot Image Classification Mask Generation

Datasets 92 Filter by name Full-text search Sort: Most downloads

- autogluon/chronos_datasets
- Monash-University/monash_ts
- Maple728/Time-300B
- thuml/UTSD
- AutonLab/Timeseries-PILE
- misikoff/zillow
- Salesforce/GiftEval
- raeidasaqr/NIFTY
- autogluon/chronos_datasets_extra
- PetraAI/PetraAI
- mikeboss/FIP1
- patrickfleith/GOCE-satellite-telemetry



Сколько данных использовали для обучения

Для сравнения:

Llama-2* — 2×10^{12} токенов
и 70×10^9 параметров — × 28

→ **TimeGPT-1**

Около 100×10^9 токенов

→ **Lag-Llama**

Около 360×10^9 токенов и 2.5×10^6 параметров — × 144

→ **TimesFM**

Около 300×10^9 токенов и 200×10^6 параметров — × 1500

→ **Moirai**

Около 27×10^9 токенов и 310×10^6 параметров — × 90

→ **Chronos**

Около 85×10^9 токенов и 700×10^6 параметров — × 170

→ **YINGLONG**

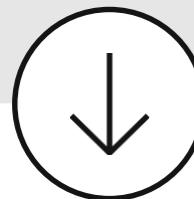
Около 78×10^9 токенов и 300×10^6 параметров — × 260

* Продукт компании Meta. Признана экстремистской организацией и запрещена на территории Российской Федерации.

Метрики и бенчмарки



Scaled MAE Mean – Monash – Moirai TimesFM



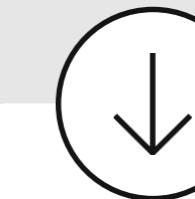
3 победы

TimesFM



2 победы

DeepAR, Moirai Large,
N-BEATS, TBATS



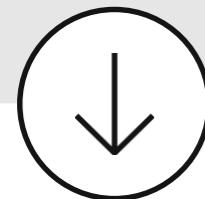
1 победа

Theta, PR, ETS, Catboost

n-beats	timesfm(zs)	moirai _{large}	moirai _{base}	tbats	catboost	fnn	ets	naive
0.789	0.790	0.830	0.834	0.889	0.893	0.915	0.941	1.000

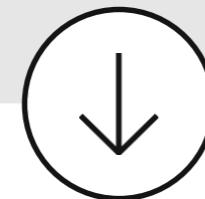
llmftime(zs)	pr	deepar	theta	dhr-arima	ses	transformer	wavenet	patchtst(zs)
1.019	1.072	1.268	1.274	1.296	1.363	1.599	1.916	2.221

Scaled MASE Mean – Monash – Nixtla Benchmark



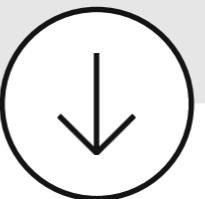
13 побед

StatisticalEnsemble –
AutoArima, AutoETS,
AutoCES, Theta



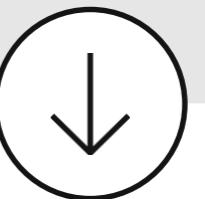
6 побед

TimesFM



5 побед

Chronos Mini

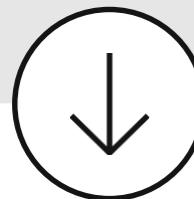


3 победы

Chronos Large

timesfm	chronos_{large}	statisticalensemble	chronos_{mini}	seasonalnaive
0.796	0.813	0.836	0.840	1.0

NMAE и SMAPE – внутренние данные



Chronos

Large- и Mini-версии
выигрывают у остальных
подходов.



Наивные подходы

Меньше проблем
с хвостом распределения.



TimesFM

Третье место.

Model	chronos _{large}	chronos _{base}	timesfm	snaive	naive	catboost	llmtime	moirai _{base}	moirai _{large}
NMAE _{median}	0.221	<u>0.229</u>	0.26	0.311	0.32	0.3258	0.39	0.398	0.51
NMAE _{q75}	0.555	0.563	0.604	0.741	0.757	0.725	0.815	0.997	1.465
SMAPE _{median}	30.068	<u>30.732</u>	33.053	33.578	32.722	38.122	46.289	41.457	52.273
SMAPE _{q75}	95.625	96.458	90.548	71.706	<u>74.863</u>	113.435	122.132	112.443	121.571

2024, октябрь



https://huggingface.co/spaces/Salesforce/GIFT-Eval

Overall By Domain By Frequency By Term Length By Variate Type About

Search
Separate multiple queries with ';'.

Select Columns to Display:
 MAPE CRPS Rank

Model types
 pretrained deep-learning statistical

Show Rows with the Following Values

T	model	MAPE	CRPS	Rank
◆	PatchTST	0.860	0.563	5.897
●	Moirai_large	0.856	0.585	6.072
●	Moirai_base	1.101	0.604	6.155
◆	TFT	1.116	0.596	7.031
●	Moirai_small	0.909	0.636	8.423
●	Chronos_large	0.930	0.629	8.814
●	Chronos_base	0.940	0.637	8.835
●	TimesFM	1.246	0.683	9.052
●	Chronos_small	0.928	0.638	9.763
◆	TIDE	1.172	0.766	12.433
◆	DeepAR	1.213	0.969	12.948

2025, июнь



G https://huggingface.co/spaces/Salesforce/GIFT-Eval

Overall By Domain By Frequency By Term Length By Variate Type About

Search
Separate multiple queries with ','.

Select Columns to Display:
 MASE CRPS Rank

Model types
 pretrained fine-tuned deep-learning statistical

Show Rows with the Following Values

T	model	MASE	CRPS	Rank
●	TiRex	0.650	0.421	5.155
●	Toto_Open_Base_1.0	0.673	0.437	7.577
●	YingLong_300m	0.716	0.463	10.113
●	YingLong_110m	0.726	0.471	10.897
●	TabPFN-TS	0.692	0.46	11.144
●	chronos_bolt_base (code)	0.725	0.485	11.371
●	timesfm_2_0_500m (code)	0.680	0.465	11.526
◆	TEMPO_ensemble	0.773	0.434	11.711
●	YingLong_50m	0.738	0.479	11.866
●	chronos_bolt_small (code)	0.738	0.487	12.423

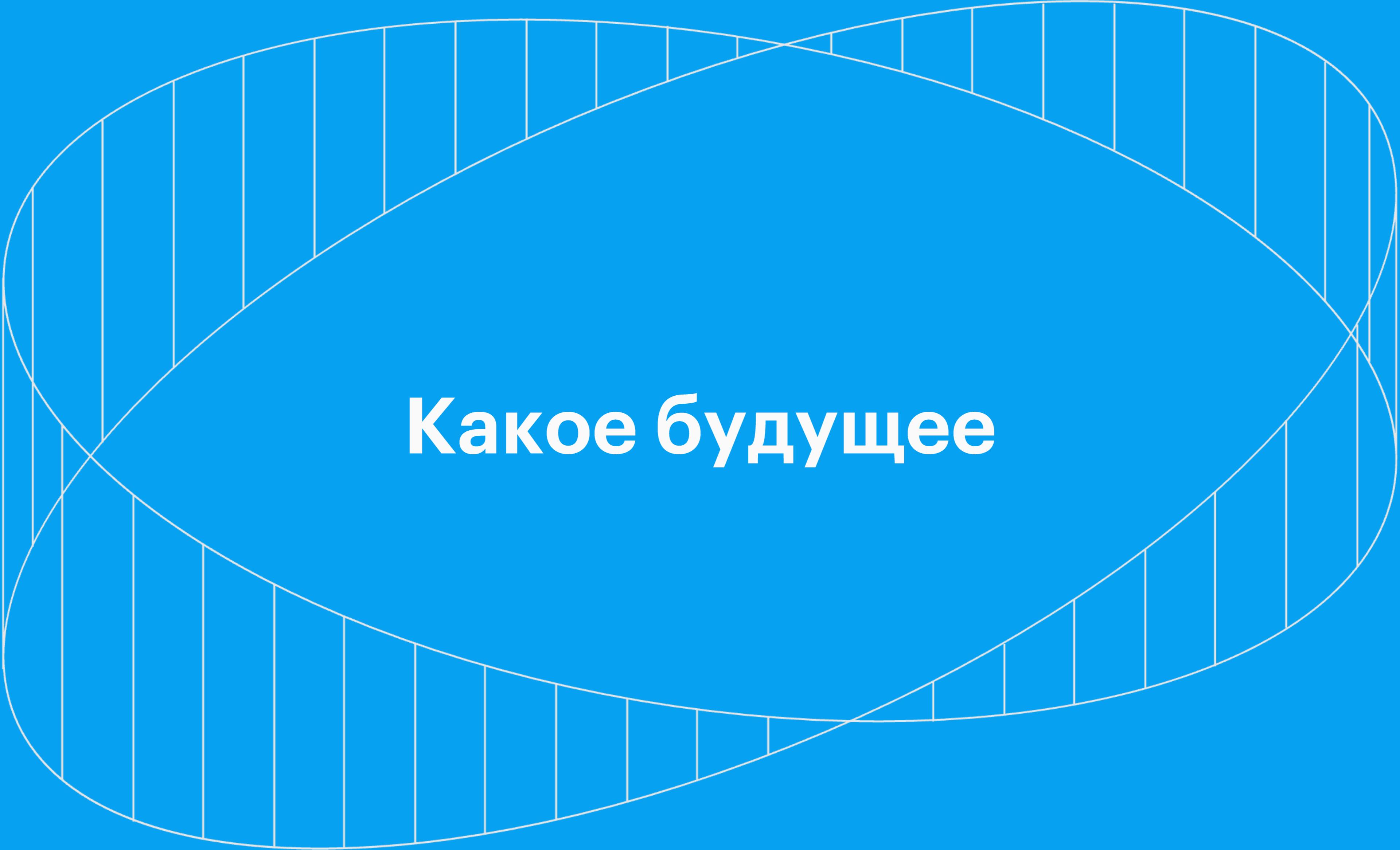


Table 4: **RMSSE results of zero-shot FMs forecasts for cloud data compared to the baselines of an online linear model and a naive seasonal forecaster.** The table shows, as for MASE, that the baseline methods perform the best across all datasets and forecast horizons (H), demonstrating that currently zero-shot FMs struggle to perform well for cloud data.

		Zero-Shot FMs							
Dataset	H	Linear	Seasonal	VisionTS	TTM	TimesFM	Chronos	Moirai	Mamba4Cast
D1	30	1.061	1.577	1.756	1.596	3.020	2.457	1.972	2.837
	96	1.583	1.557	1.596	1.814	4.664	4.551	2.188	4.419
	336	2.627	2.433	2.463	2.905	5.990	6.685	3.145	5.950
D2	30	0.977	1.017	1.599	2.001	2.722	2.509	2.486	3.006
	96	1.358	1.258	1.927	2.782	4.819	4.398	3.782	6.137
	336	2.233	1.908	2.789	3.927	6.416	6.114	4.699	10.569
D3	30	1.012	1.098	1.380	1.608	1.968	1.841	1.995	2.108
	96	1.200	1.237	1.473	1.971	3.232	2.757	2.626	3.481
	336	1.484	1.457	1.660	2.374	4.003	3.602	2.908	4.526
D4	30	1.081	1.053	1.181	1.816	2.015	2.095	2.019	2.166
	96	1.194	1.175	1.233	1.978	2.410	2.474	2.321	2.654
	336	1.400	1.328	1.414	2.188	2.693	2.837	2.595	3.185

Бенчмарки. Проблемы





Какое будущее

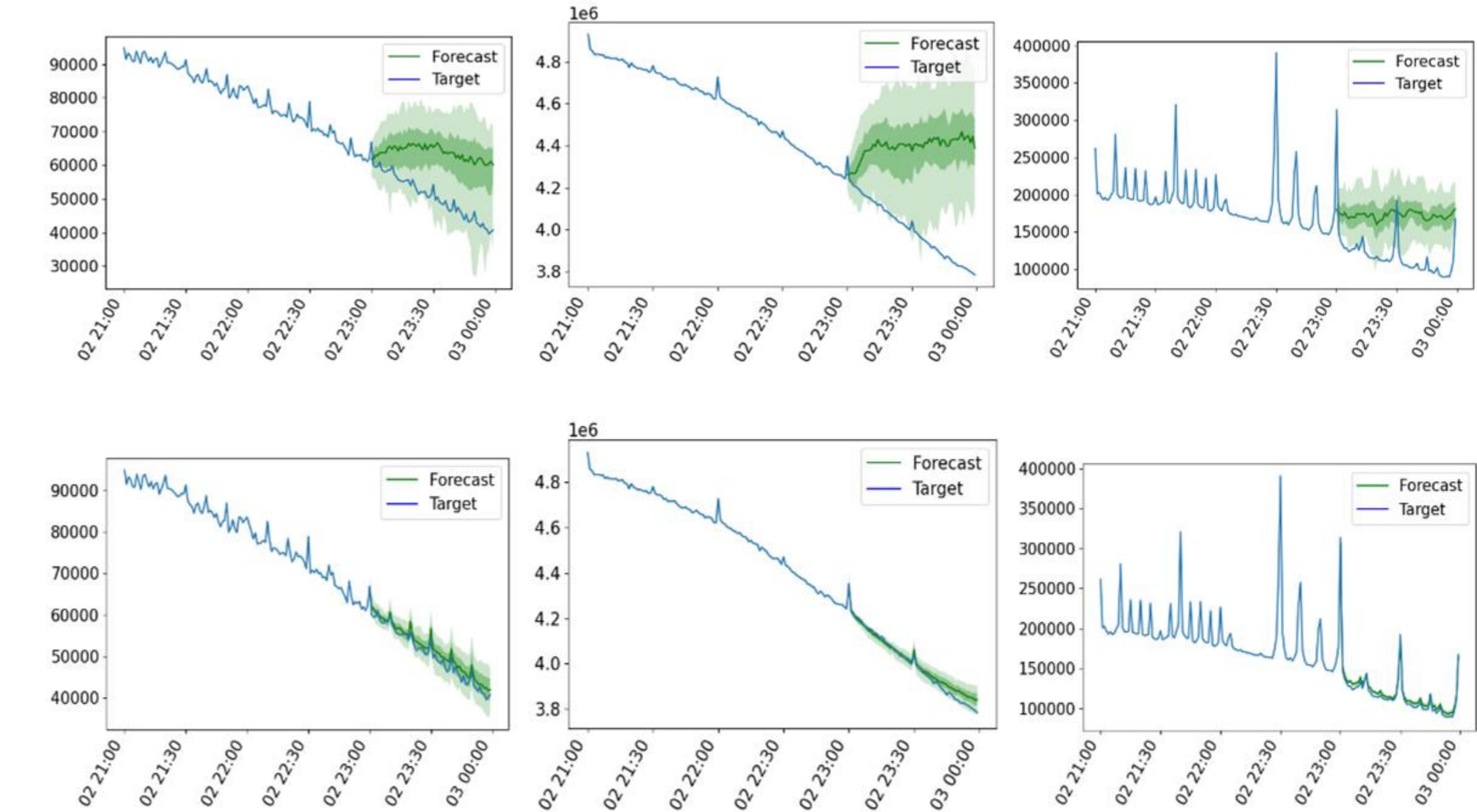
Zero-shot или Fine-tuning

Fine-tuning работает лучше

Дообученные модели во всех случаях показывают стабильно лучшие метрики.

Few-shot Learning

Возможно, чтобы как-то обойти дообучение модели, можно будет использовать Few-shot.



Проблема доп. данных

→ Метод из Moirai

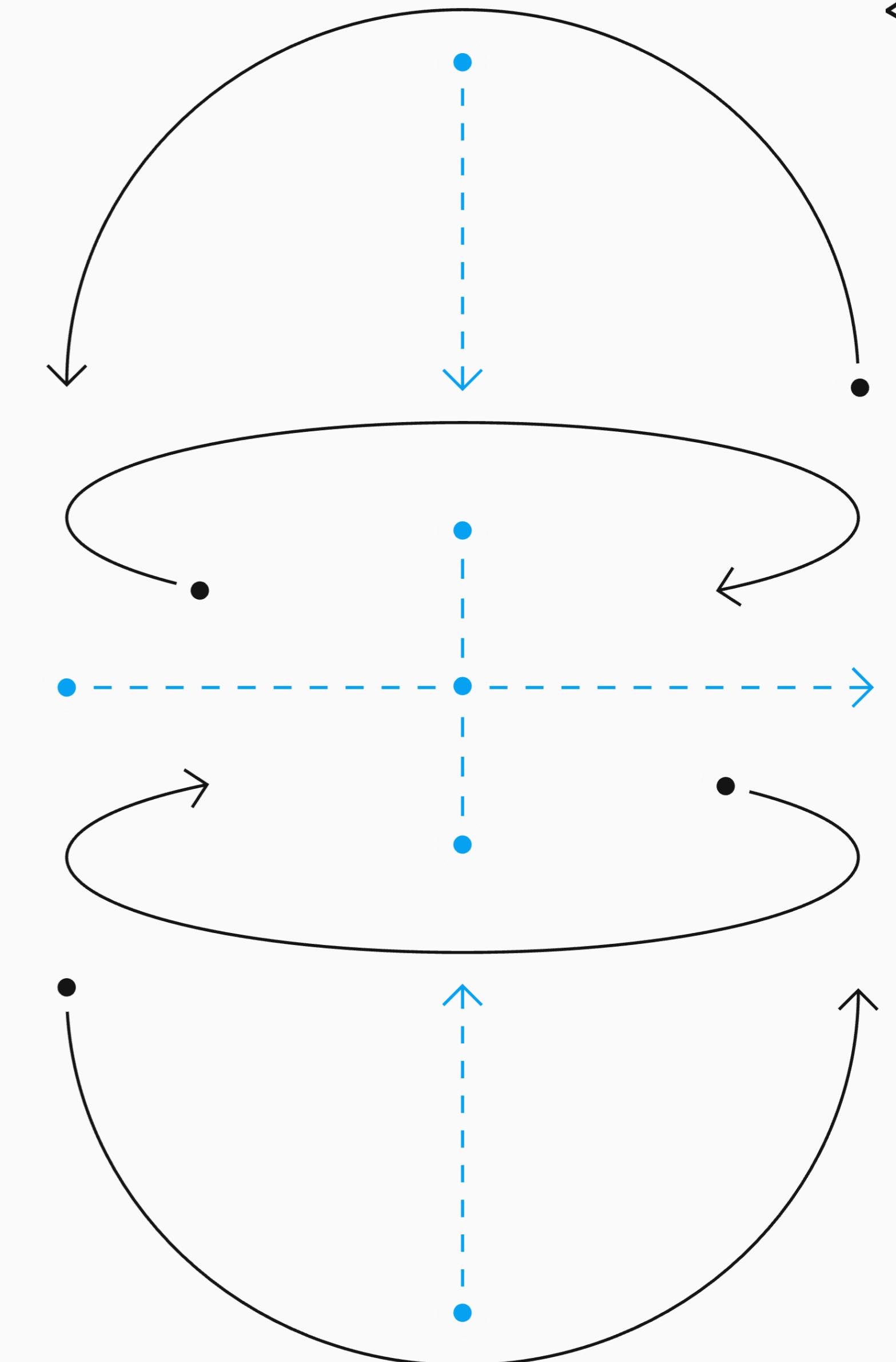
Потенциально он раздувает контекст, что в случае трансформеров может быть фатально.

→ Time Series Embeddings

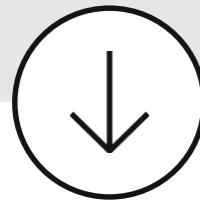
Почему бы не использовать модели для эмбеддингов, такие как TS2Vec, например.

→ Другие модальности

Мы могли бы добавлять информацию на естественном языке, например, как в PromptCast.



Deep Learning Tabular ML Classical Models



Ансамбли классических моделей

Мы видели, что ансамбли
могут показывать лучшие
метрики.



Добавление признаков

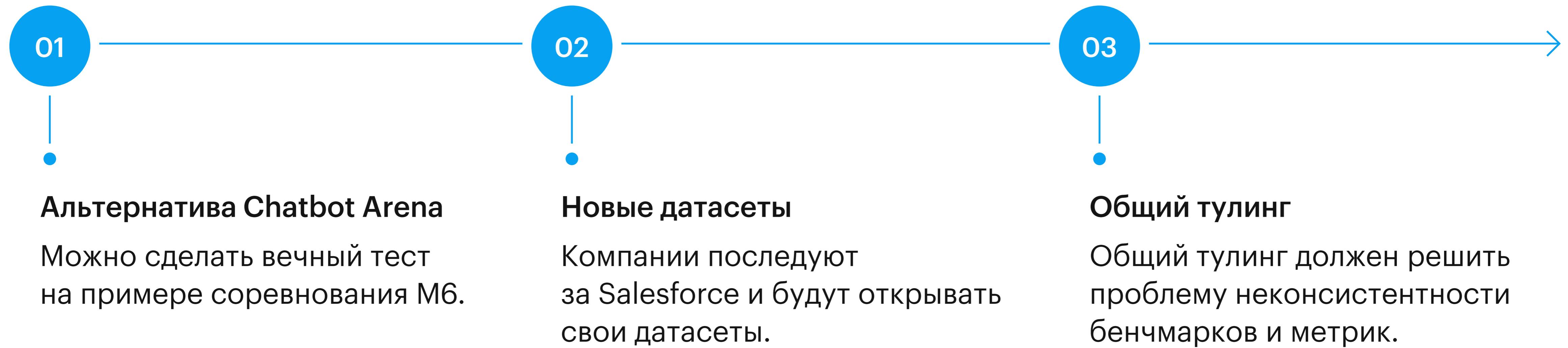
В зависимости от доменной
области, важность генерации
признаков может возрастать.



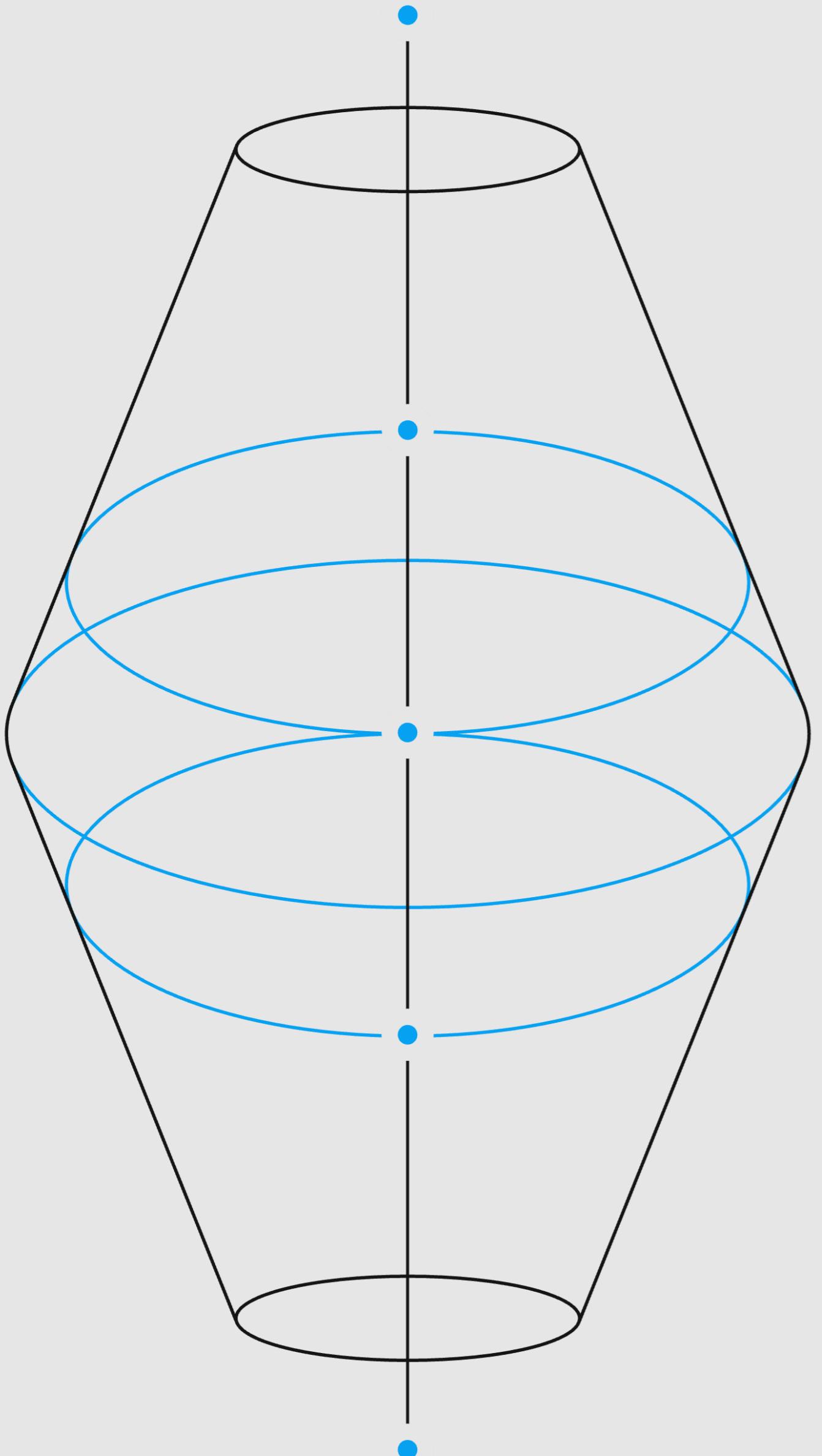
Проблема табличного ML

Всё ещё рулят бустинги
и генерация признаков —
явных сигналов, что это
не так, ещё нет.

Данные бенчмарки



Выводы



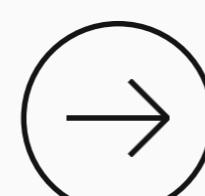
Не SOTA

Zero-shot Foundation Models — это не SOTA в Forecasting, но хорошие бейзлайны, которые стоит пробовать.



Classical ML

Не стоит забывать про классические модели и бустинги.



Вам не нужно много железа

На данный момент есть много пространства для работы, и вам не нужно много железа, чтобы попробовать что-то новое.