Forecasting approaches and challenges

Milen Chechev

26/06/2024

Agenda

- Intro presenter
- Intro to Forecasting
- What is the experiment setup for forecasting? How we evaluate the results?
- What are the most popular forecasting approaches? Pros and cons?
- What is the current state of the art?
- XGBoost / Trasnsformers

Head of Data Science at Fourth

- 10+ years experience at ML, Leading ML/DS projects and teams
- 10+ years experience at Software Engineering

Education:

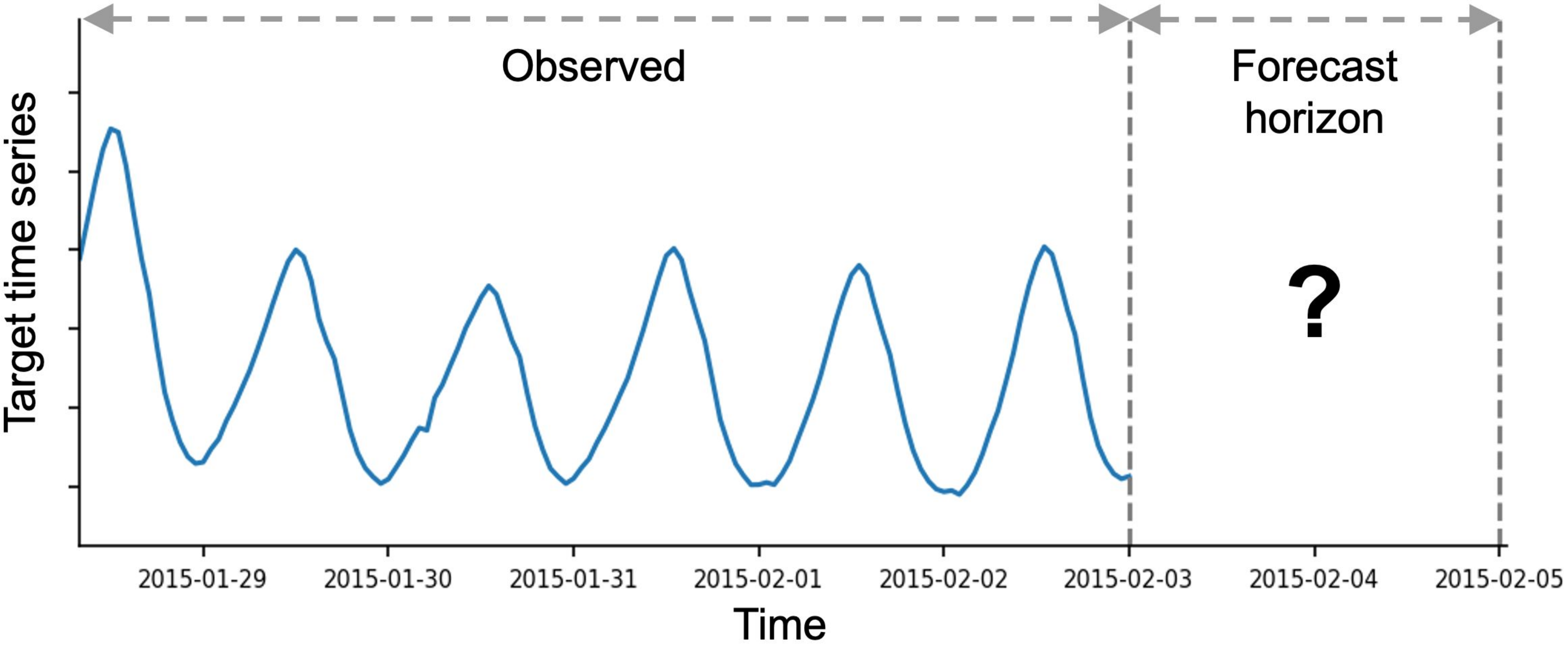
- PhD in Recommendation Systems, Sofia University
- Specialization in “Machine Learning” at Aalto University

Teaching Experience (15+ years, FMI, SU):

- Recommendation Systems
- Machine Learning
- Artificial Intelligence
- Data Structure and Algorithms

LinkedIn profile

What is timeseries forecasting ?



$$Y = levels + trends + seasonality + noise$$

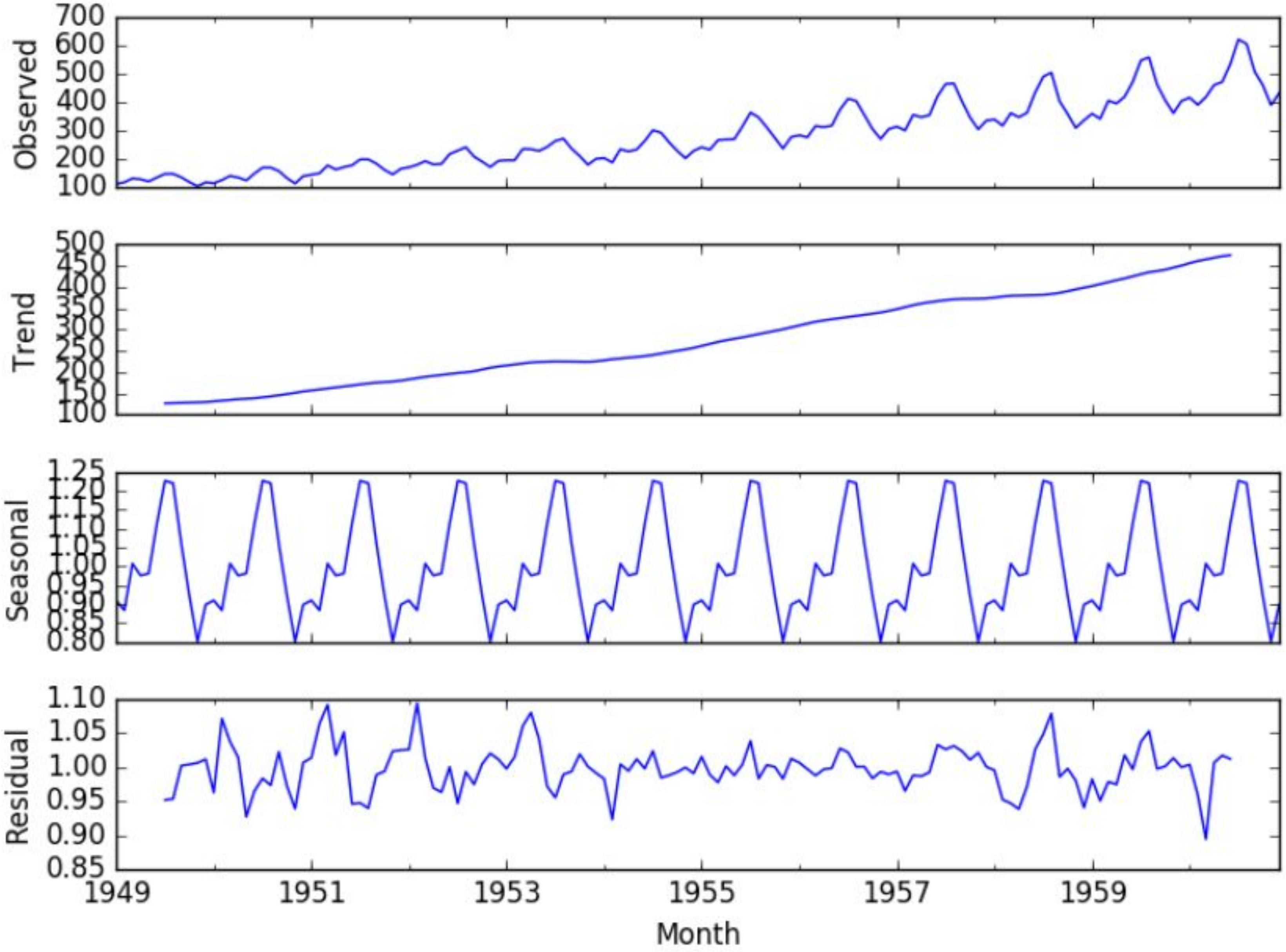
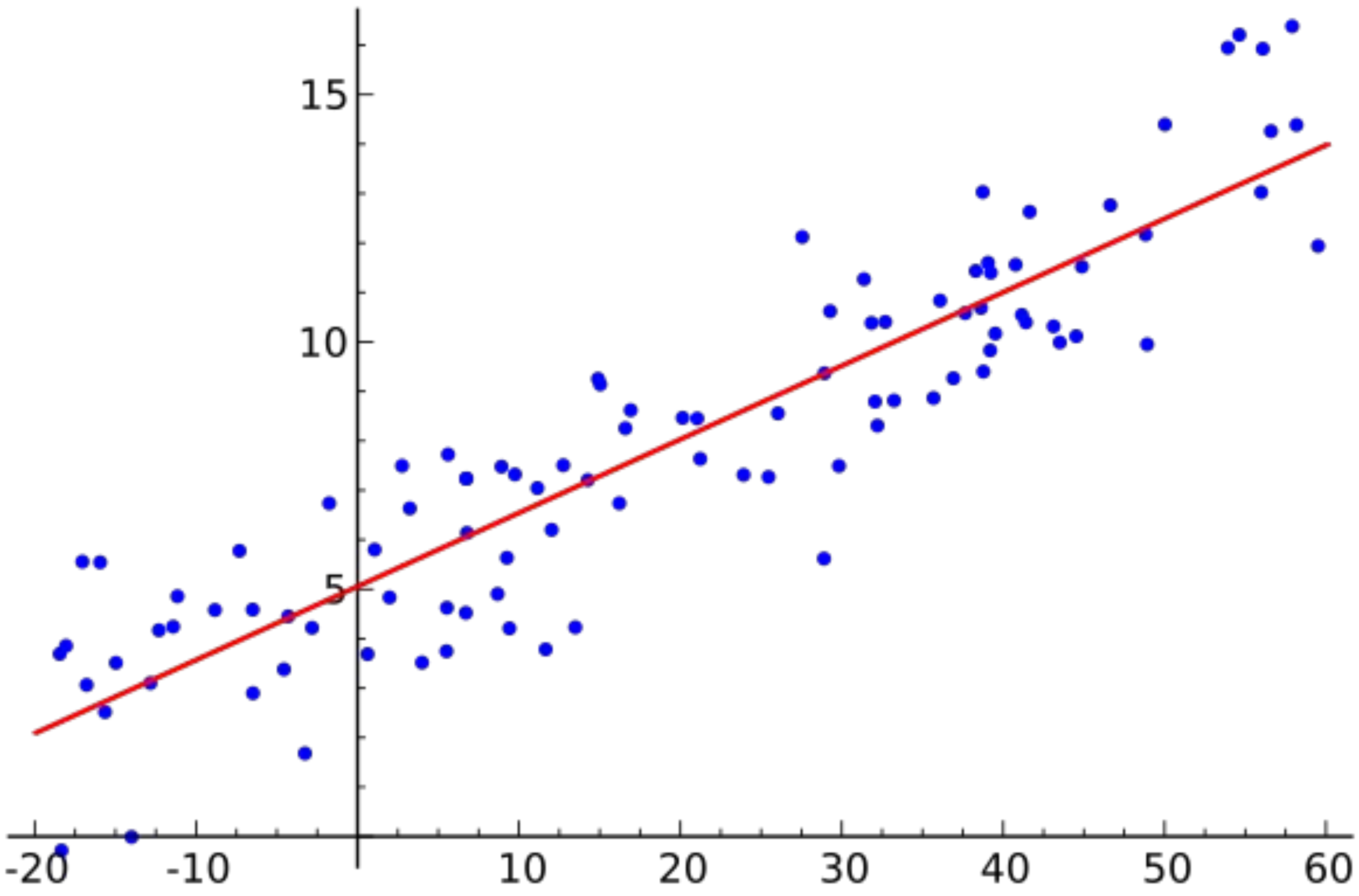


Image Courtesy: Machine Learning Mastery

Classical Forecasting Algorithms

- Autoregression (1927)
- Exponential Smoothing (1950)
- ARIMA (1970)
- ARIMAX (ARIMA with eXogenous Predictors) — allows for exogenous data
- SARIMA (Seasonal ARIMA) — accounts for seasonality patterns
- SARIMAX (SARIMA with eXogenous Predictors) — allows for exogenous data

Supervised Machine Learning: Regression



Forecasting vs Regression?

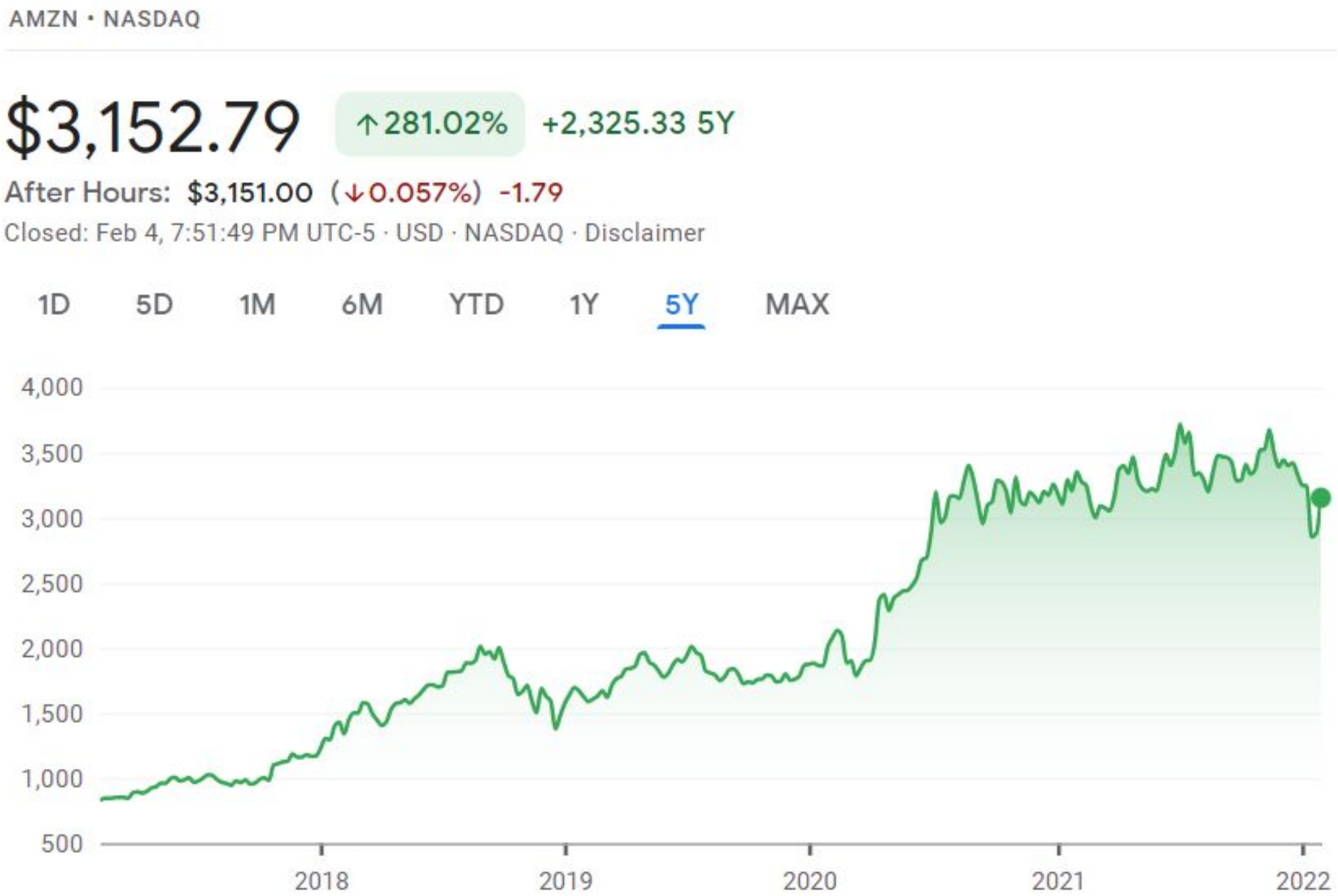
Similarities:

- Both are using historical data
- Both are giving numeric values as an output

Differences:

- Forecasting is working on Timeseries data
- The time is one directional, so each date is seen only once and never repeated
- Timeseries data have an explicit order defined from the time

Real world example of timeseries



Why to use ML instead of classical statistical methods

- Events!!!
  - How to integrate events at the forecasting algorithm?
  - Local/Global events
  - Internal/External events
  - Events from dependencies from other timeseries
  - …
- Could solve more complex domains

How to integrate different kind of events at classical statistical methods?

- Not an easy solution!
- Lots of expert knowledge and rules
- Expert systems vs Machine Learning?

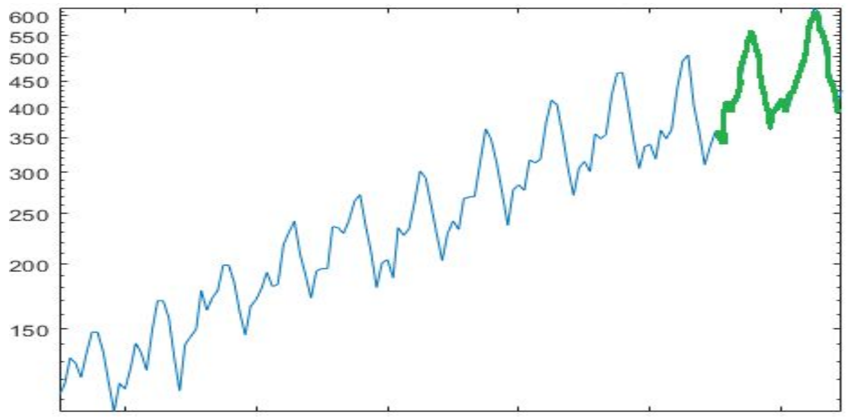
The Machine Learning solution

- Model the problem as regression problem
- Solve it as regression problem
- Evaluate it as a forecasting problem

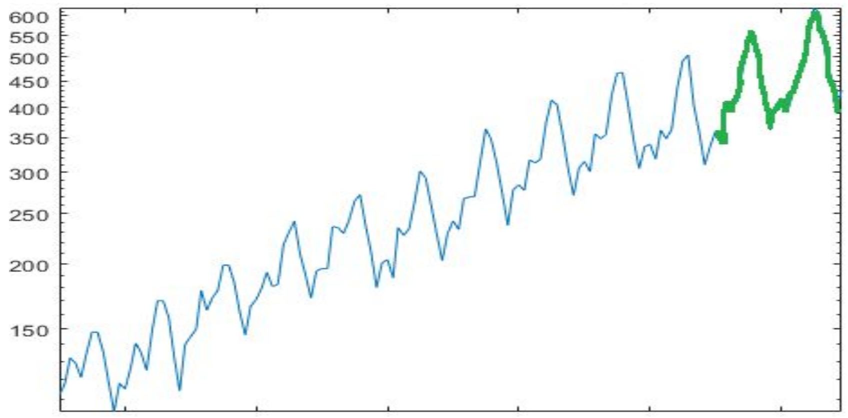
Model the Forecasting problem as regression problem

- We have only one feature - date
- Is this enough for the regression?
- Not, really…

- How to get more features?
  - Lag features – Lag1,2,3,…
  - Window mean
  - Window trend



Model the Forecasting problem as regression problem



- How to get more features?
  - Lag features – Lag1,2,3,…
  - Window mean
  - Window trend
  - Window min/max
  - …
  - Is there some automatic way to extract timeseries features?
  - Of course, for example the TsFresh library

How to evaluate the results?

- Cross validation for regression problems

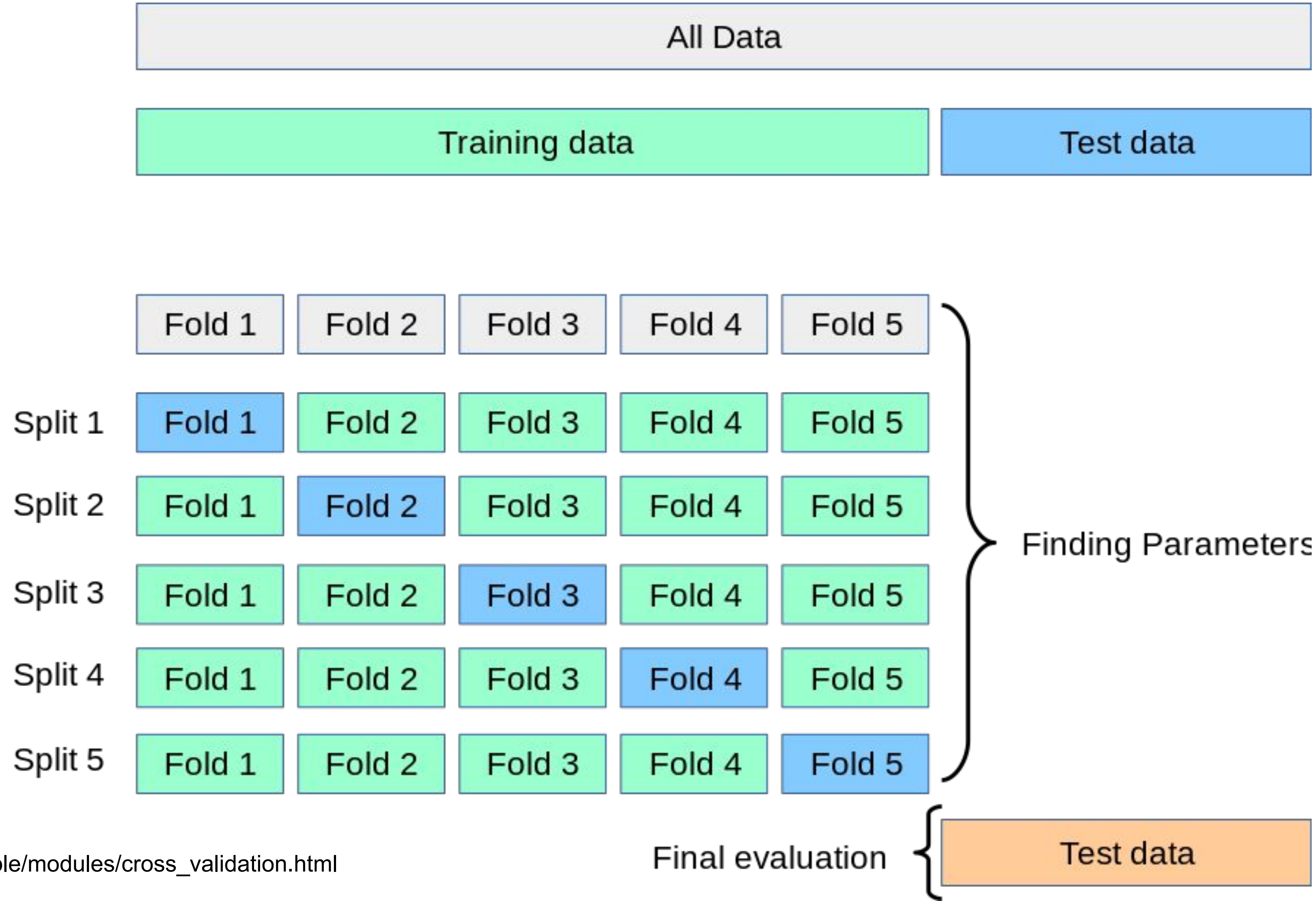


Image from: https://scikit-learn.org/stable/modules/cross_validation.html

What is the experiment setup for forecasting?

- Sliding window vs Expanding window

Diagrams source: https://www.kaggle.com/cworsnup/backtesting-cross-validation-for-timeseries/notebook

Metrics?

- RMSE

$$\text{RMSE} = \frac{\sum (\text{Ai-Fi})^2}{\text{n}}$$

- MAE

$$\text{MAE} = \frac{\sum |\text{Ai-Fi}|}{\text{n}}$$

- MAPE

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

- WAPE

$$\text{WAPE} = \frac{\sum |A - F|}{\sum A}$$

- …

What is the current state of the art?

- How we could easily check what is the current state of the art?

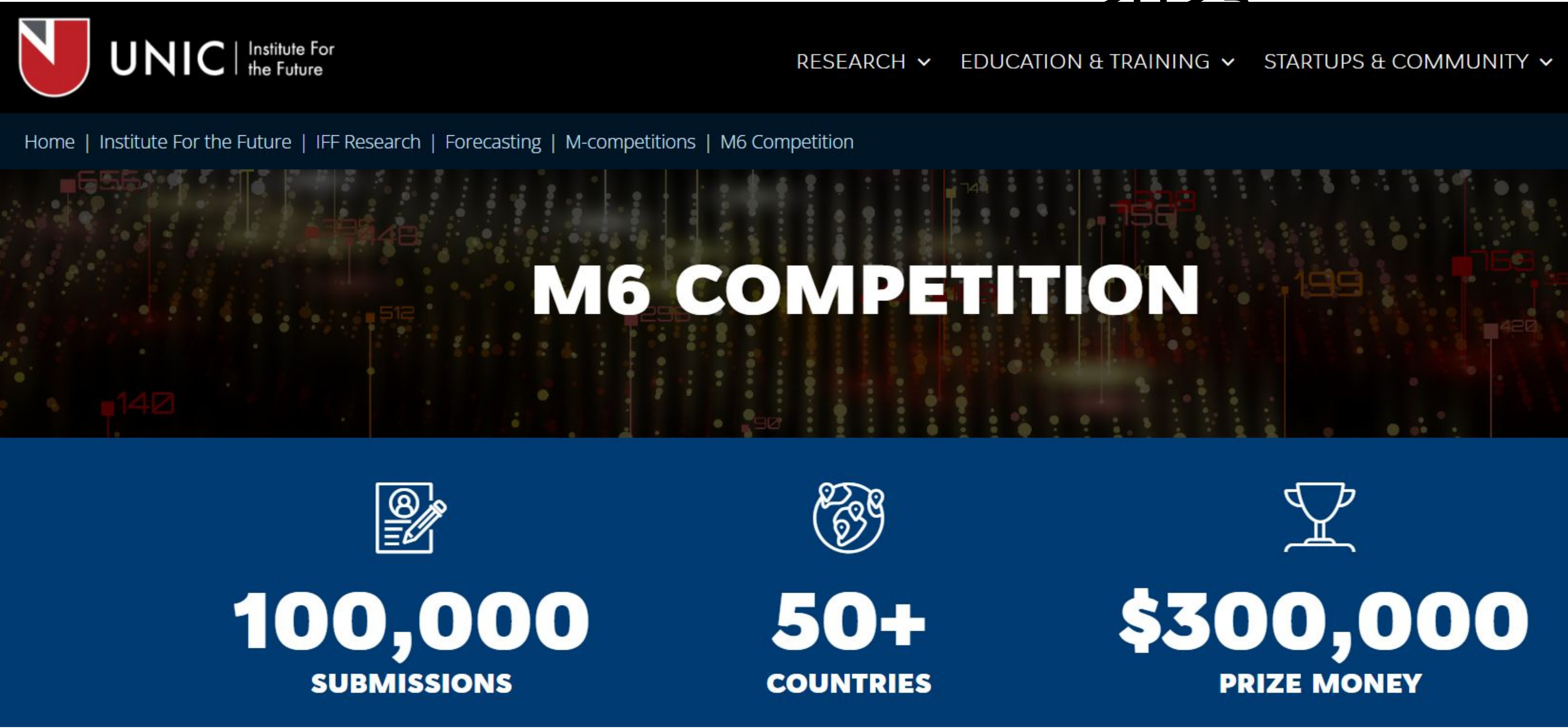
kaggle

30 June 2020

State Of the Art algorithms for M5

- Gradient Boosting Algorithms (XGBoost, LightGBM) are showed to be the most effective on the M5 competition

February 2023



State Of the Art algorithms for M6

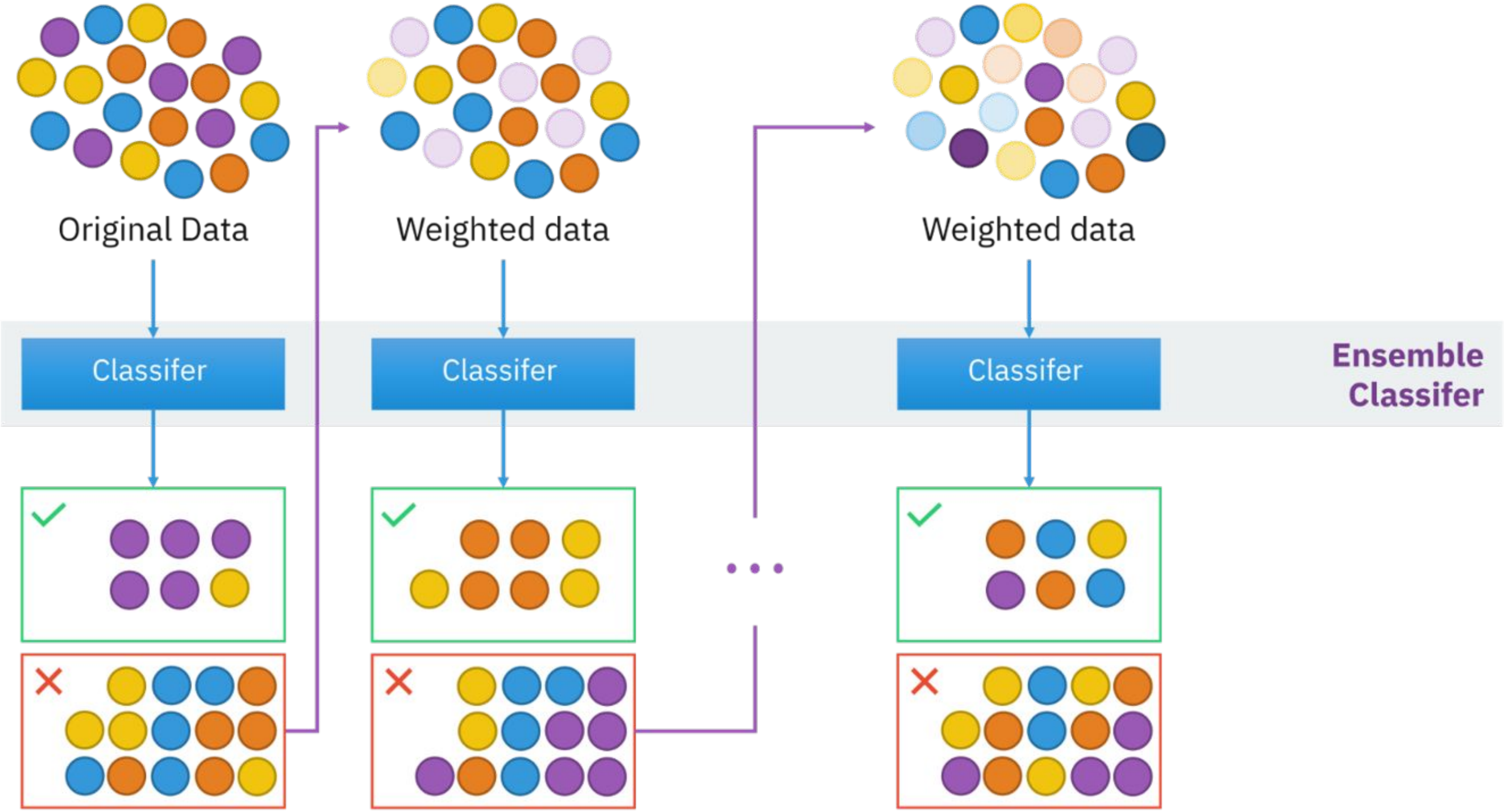
- Gradient Boosting Algorithms
- Probabilistic forecasting

Why Gradient Boosting Algorithms are so good?

- What is boosting?

Train models one after another as every new consider with higher weight the items which was misclassified and train on them. Weighted ensemble – based on model evaluation metric.

Boosting



Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer

ML Models Challenges

- Feature generation
- Complexity of the model – overfitting
- Hyperparameter tuning
- Explainability

Forecasting Challenges

- Short, medium and long term forecasting
- Multiseries forecasting
- Hierarchical forecasting
- Promotions and A/B tests
- Reinforcement Learning for Timeseries

Weather Forecast state of the art?

DeepMind

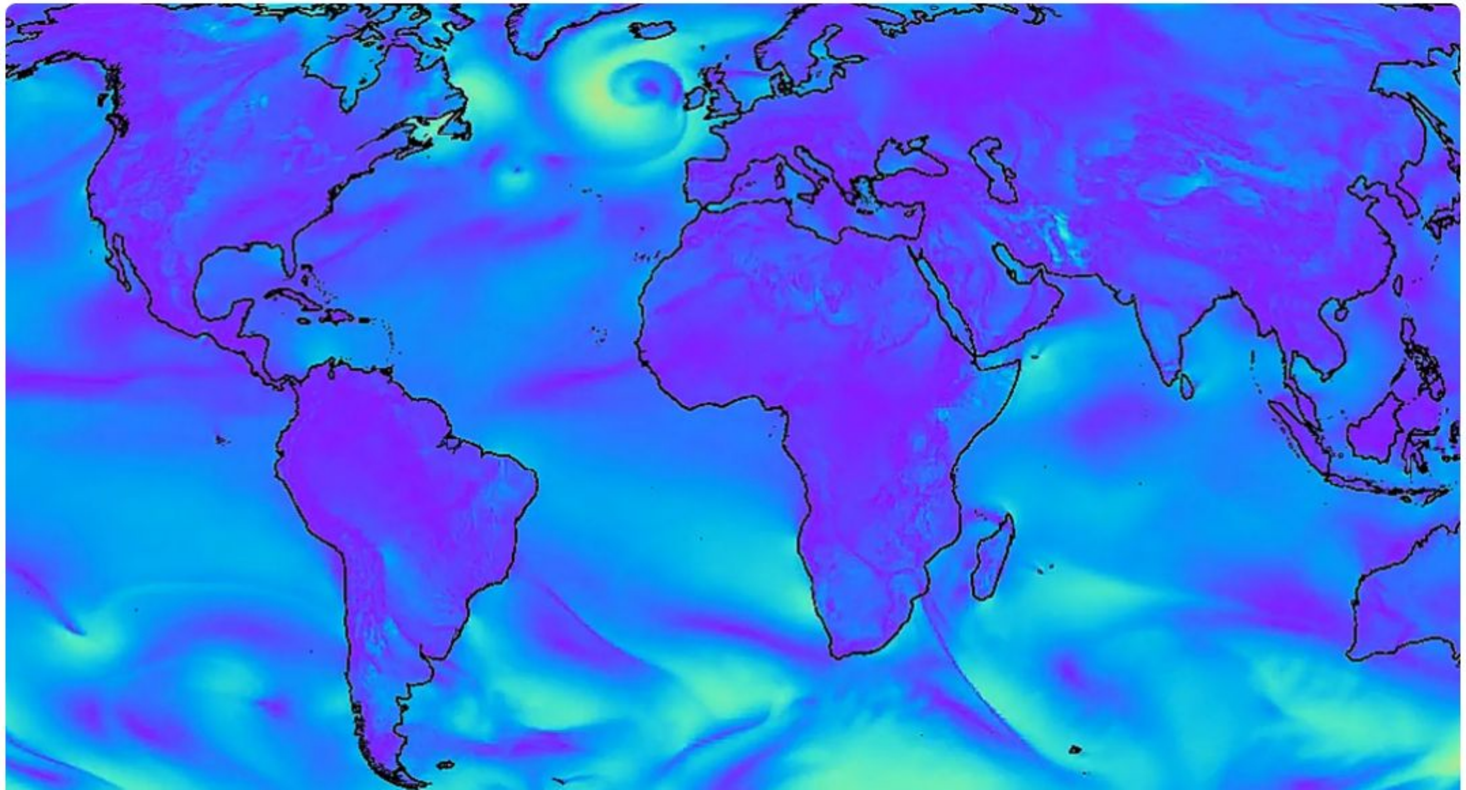
GraphCast: AI model for faster and more accurate global weather forecasting

14 NOVEMBER 2023

Remi Lam on behalf of the GraphCast team

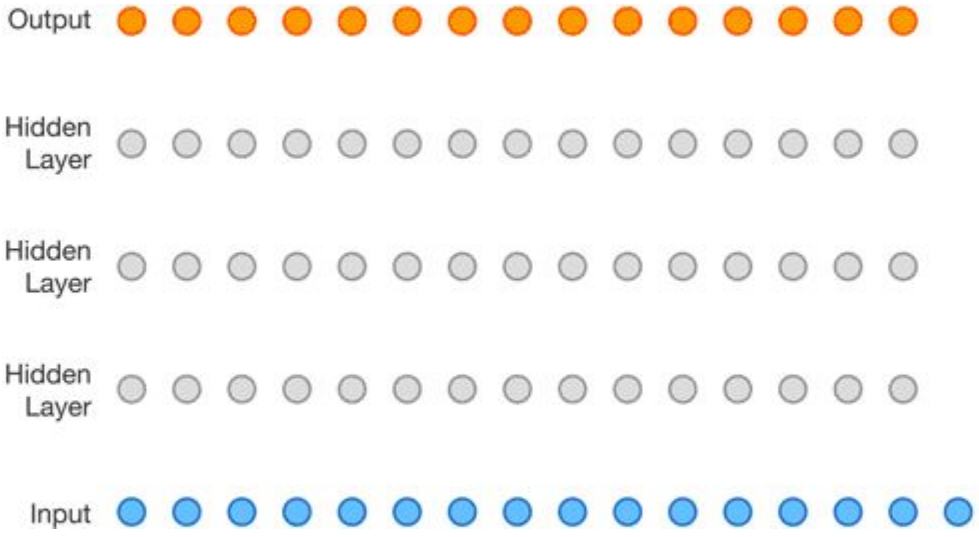
Share

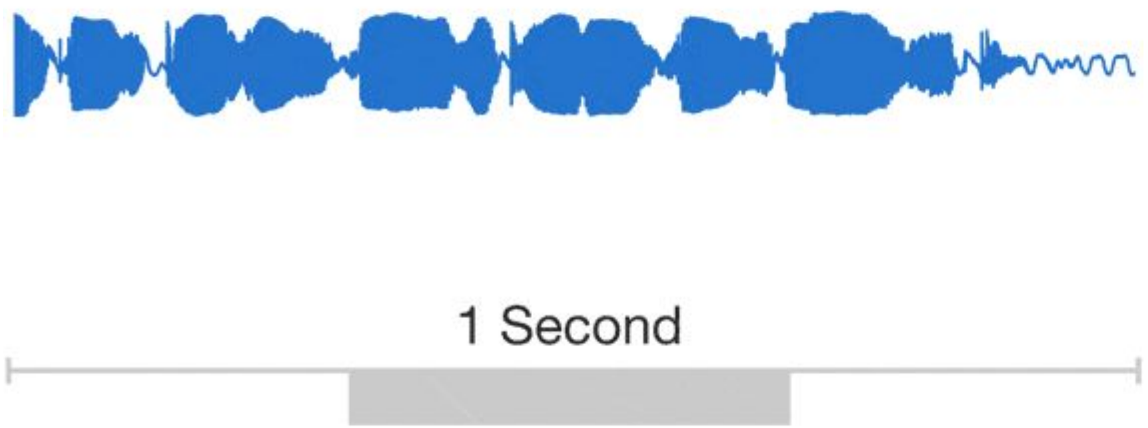
graph neural network



Could we use Deep Learning/Transformers for Forecasting?

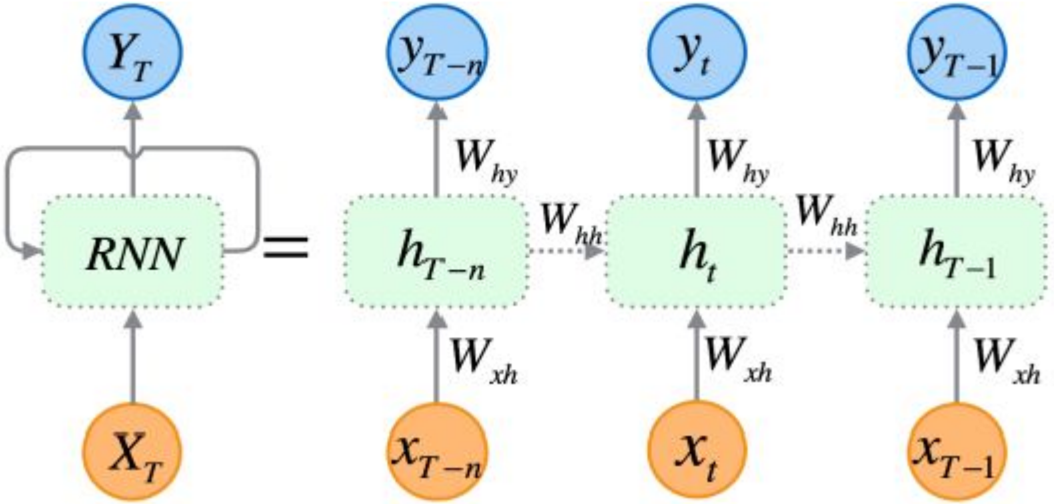
CNN

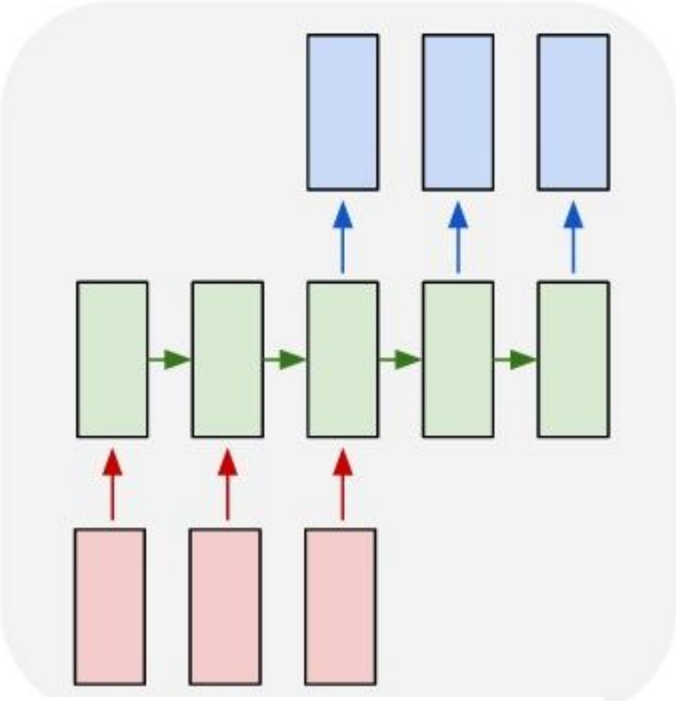


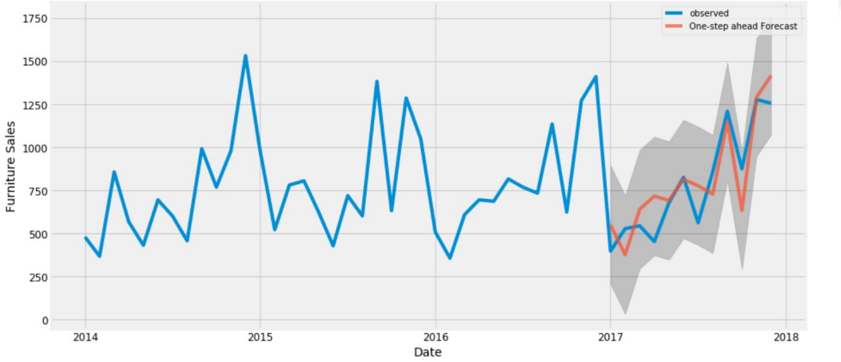


Wavenet - 2016

RNN


Y_T
y_{T-n}
y_t
y_{T-1}
W_{hy}
W_{hh}
RNN
=
h_{T-n}
h_t
h_{T-1}
W_{xh}
X_T
x_{T-n}
x_t
x_{T-1}




observed
One-step ahead Forecast
1750
1500
1250
1000
750
500
250
0
Furniture Sales
2014
2015
2016
2017
2018
Date

Intro to Transformer

2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

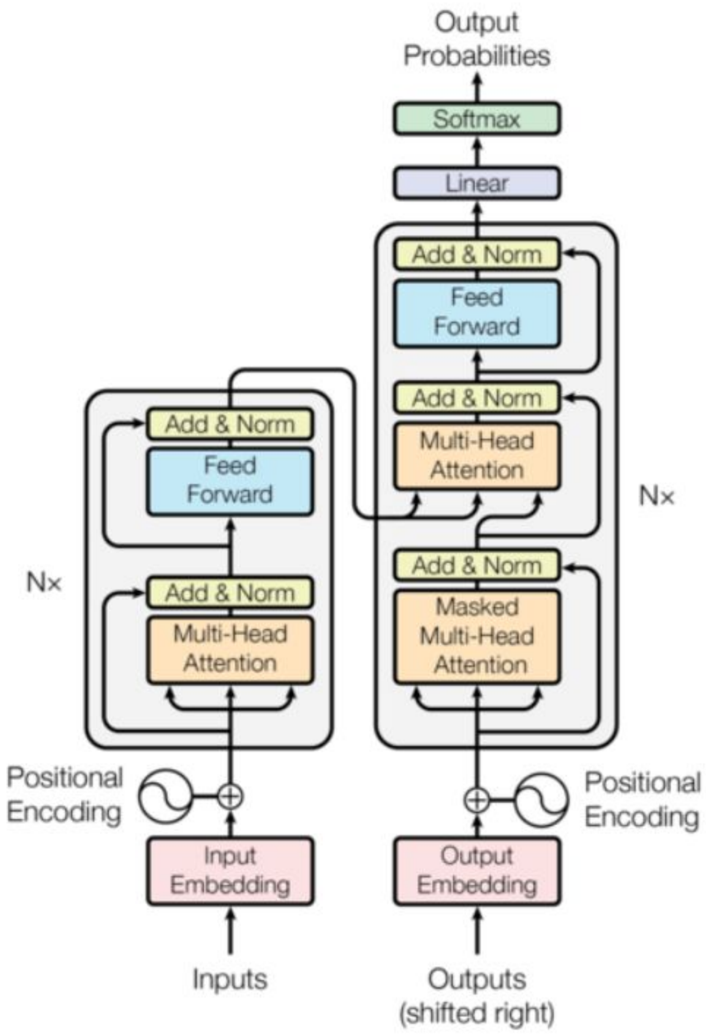
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Transformer architecture

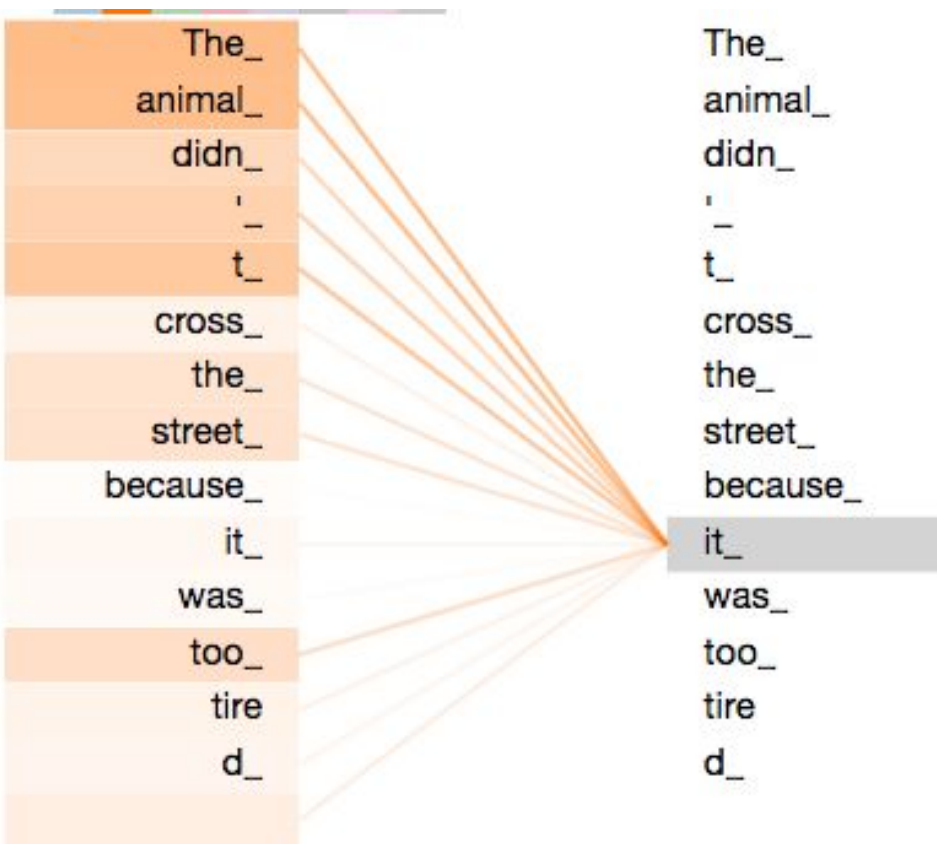


Transformer architecture

- Introduced first for Translation problem(sequence to sequence)
- Main advantage from RNN - could easily be parallelized
- To understand it we need to understand:
  - Self Attention
  - Multi-head attention
  - Masked multi-head attention
  - Embeddings and positional encoding

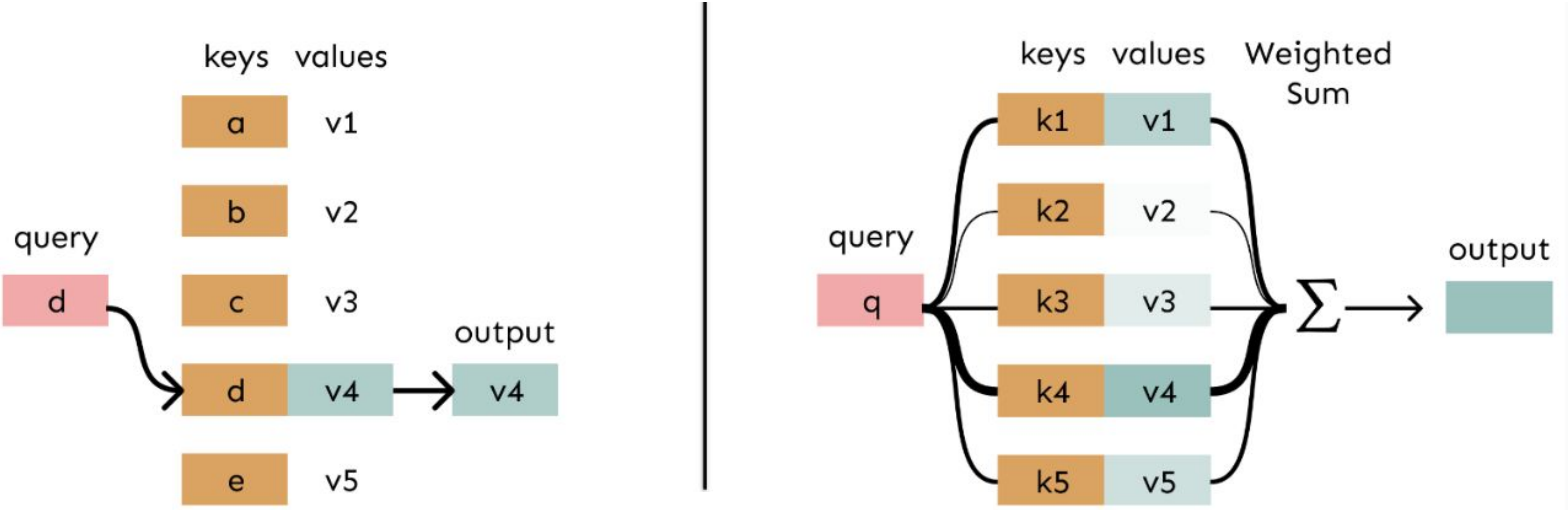
Why we need attention?

The animal didn’t cross the street because it was too tired.



Intuition for the attention

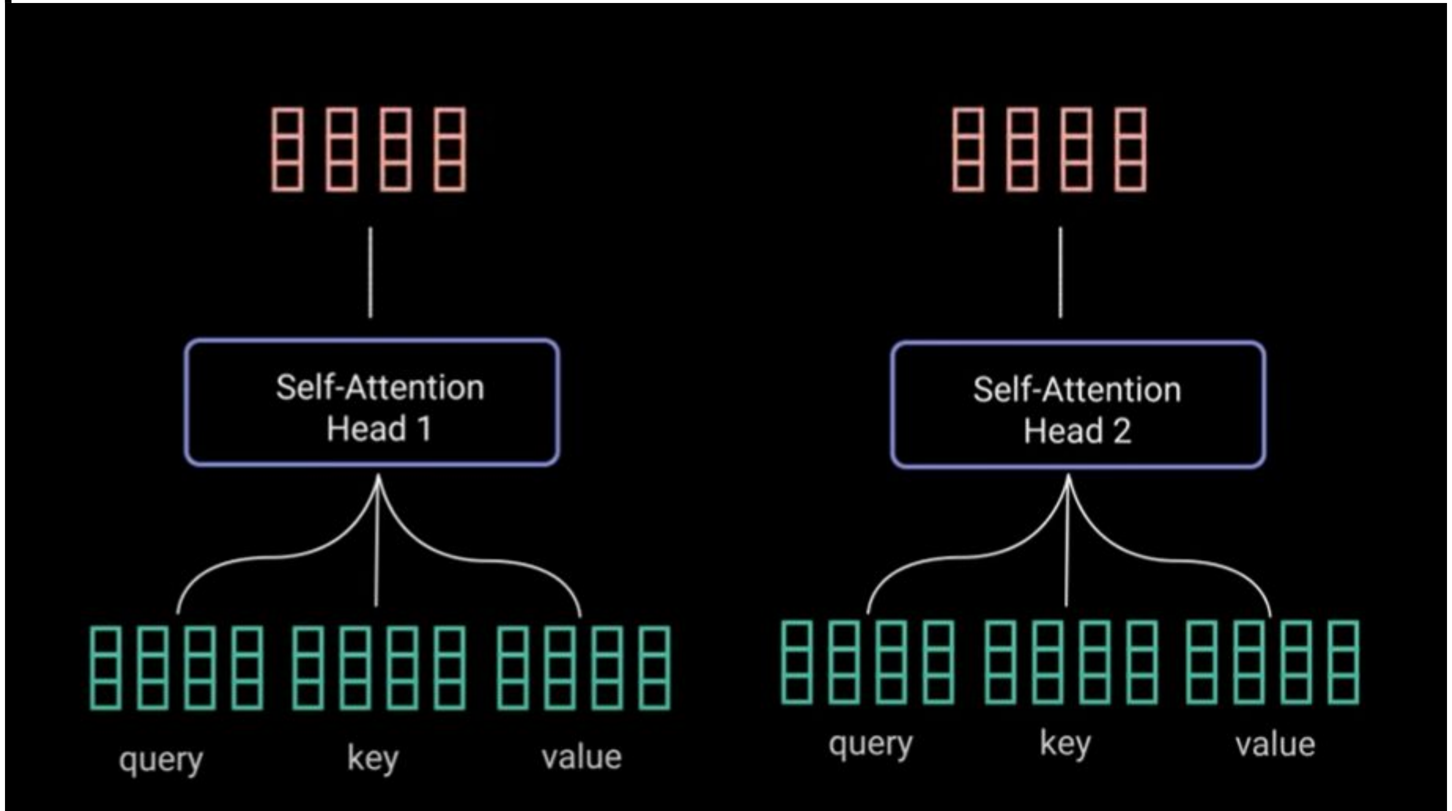
Dictionary vs Attention

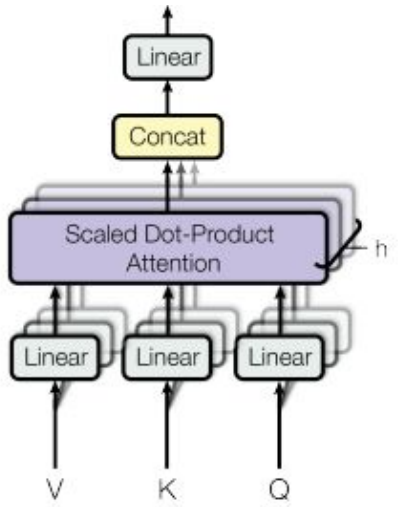


$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-head Self-Attention

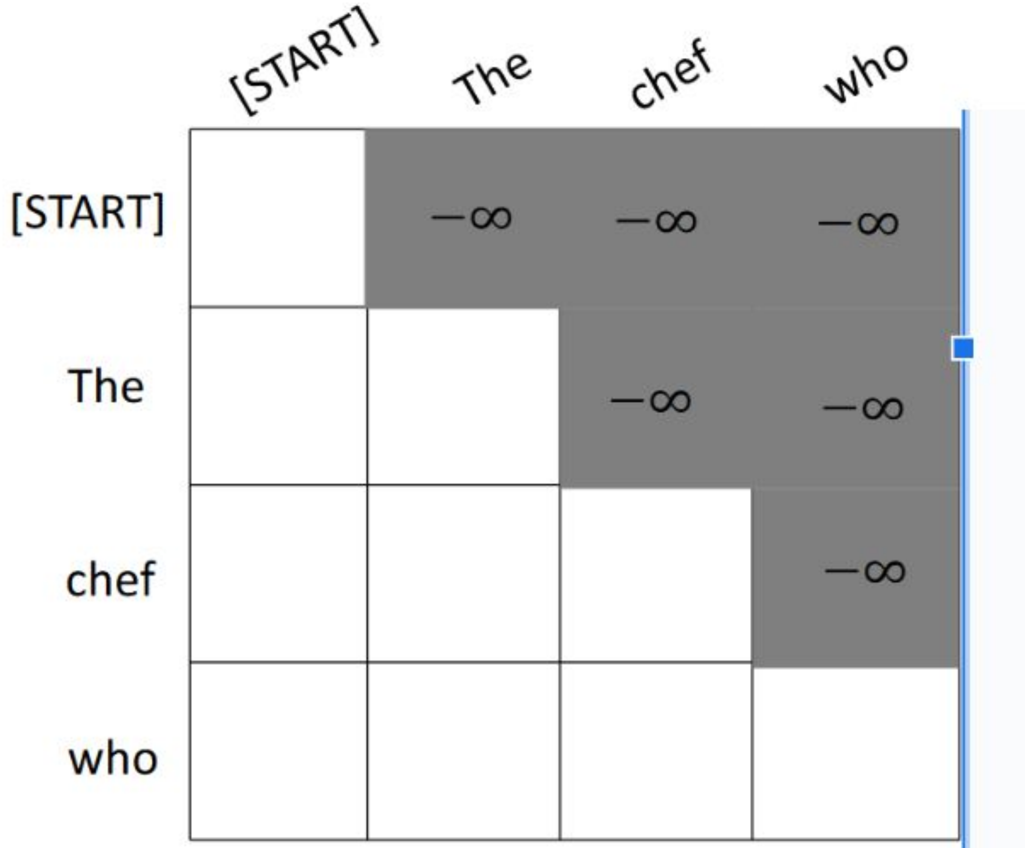
- Instead of one tupple of Q, K, V matrixes we train multiple as the intuition is that each head will look for different matching patterns





Masked Multi-head Attention

- During the training we don't want to give to the model the future values, so we masked the attention to the future values to be minus infinity
- Why not just hiding the future words?
  - It's less efficient



How to work with words?

- One hot encoding

  dog = [1,0,0,0, ………,0] (dimension D(the size of the dict))

  cat = [0,1,.................,0]

- Embeddings

  dog = [0.5,0.14,2.5,...1.7] (dim K(the size of the latent space)

  cat = [0.1,0.30,2.3,...1.2]

-

How to work with words?

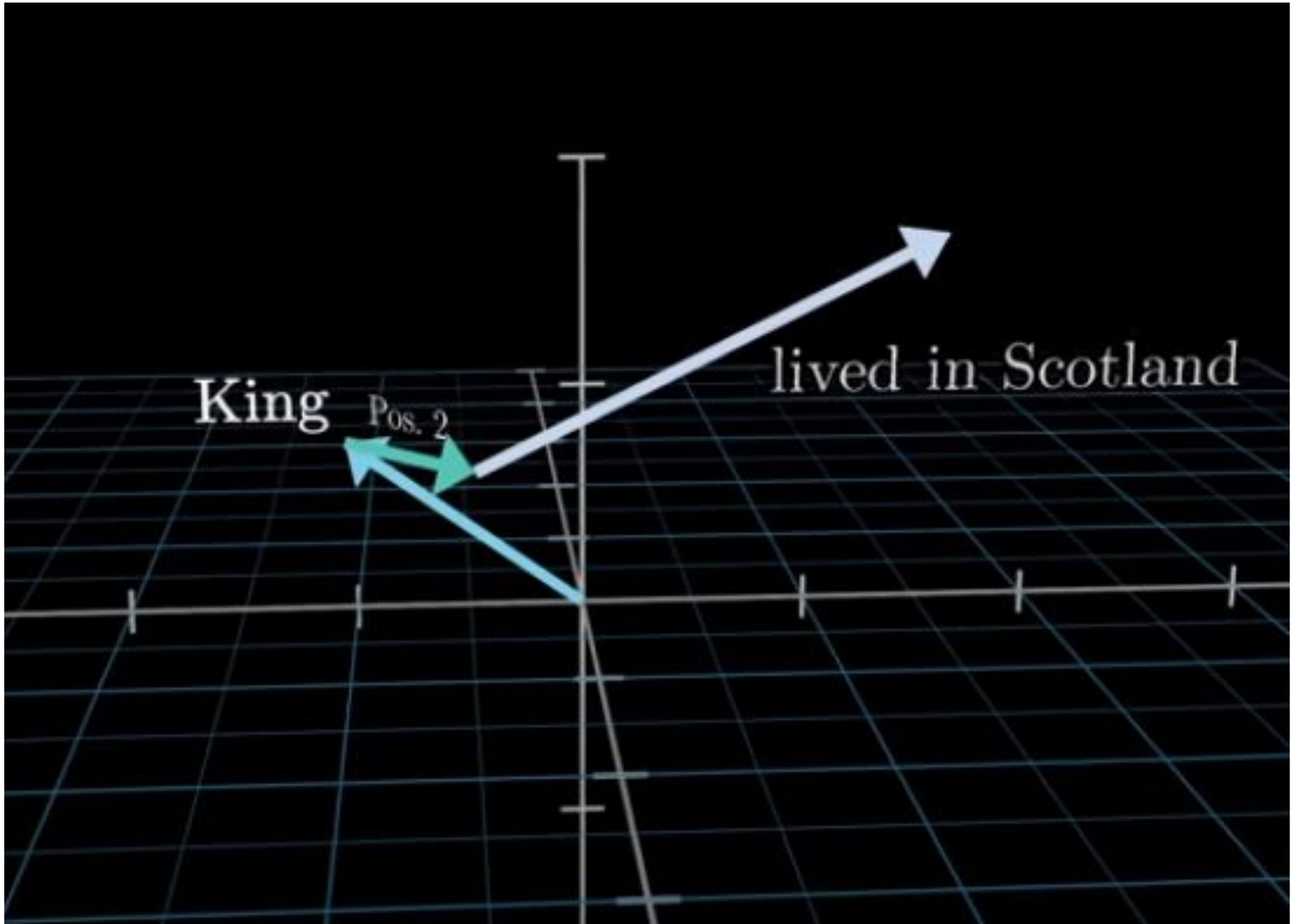
- What about the world order?
  - The order of the word in the sentence is very important as there could be even exactly opposite meaning if we shuffle the words.
  - we could think of it as again being encoded as 1 hot encoding and after that added to the embedding of the word

Dog bark the cat

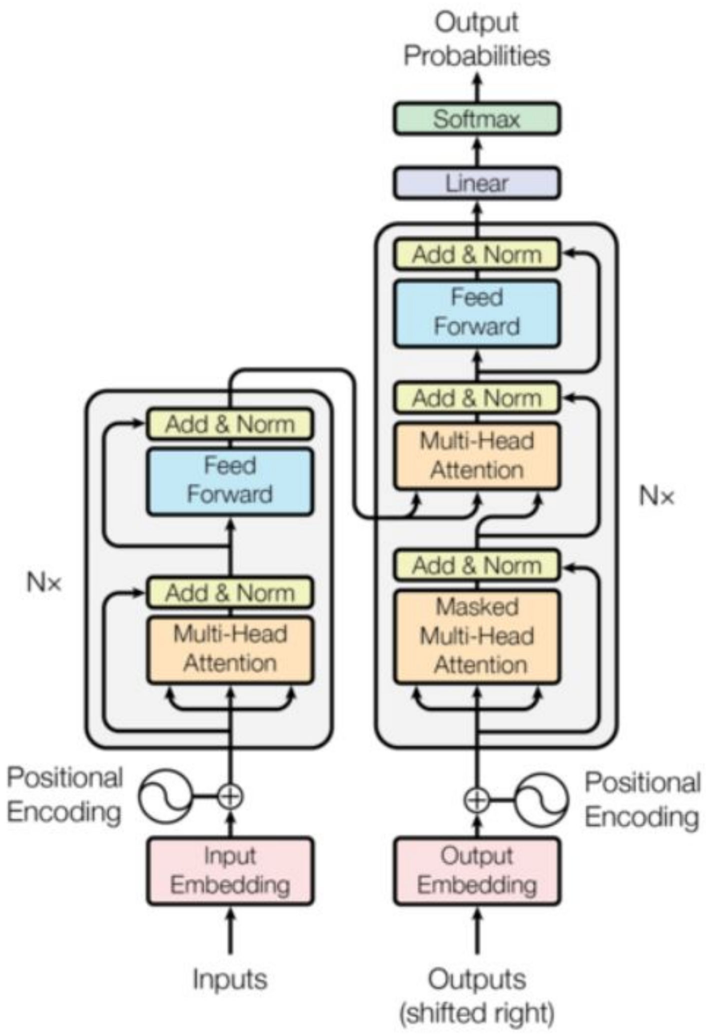
dog = [1,0,0….,0][1,0,0,0]

bark =[1,0,1,.....0][0,1,0,0]

Intuition of using word and positional embedding



Transformer architecture



https://nlp.seas.harvard.edu/2018/04/03/attention.html

Could we use Transformers for Forecasting?

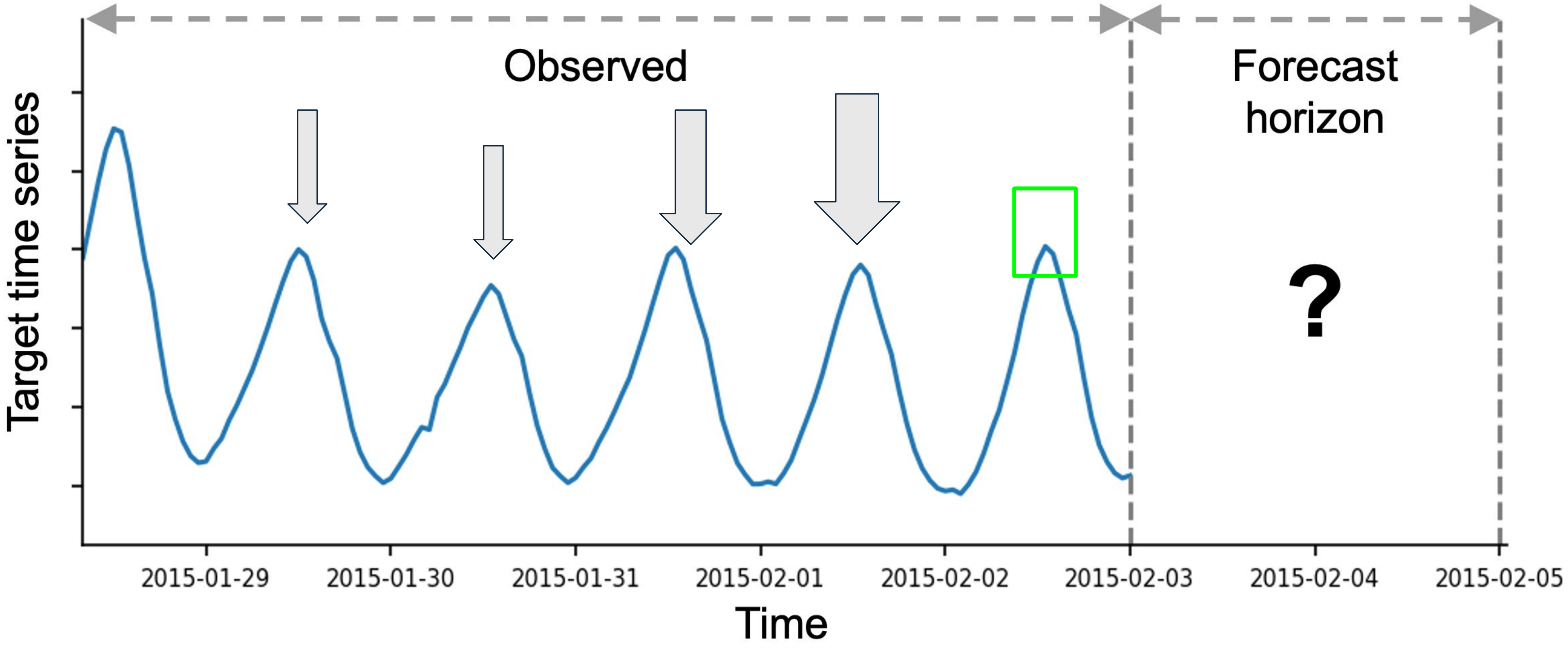
Transformers for forecasting?

- Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 12. 2021.
- Lim, Bryan, et al. "Temporal fusion transformers for interpretable multi-horizon time series forecasting." *International Journal of Forecasting* 37.4 (2021)
- Wu, Haixu, et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." Advances in neural information processing systems 34 (2021)
- Woo, Gerald, et al. "Etsformer: Exponential smoothing transformers for time-series forecasting." arXiv preprint arXiv:2202.01381 (2022).
- Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." International conference on machine learning. PMLR, 2022.

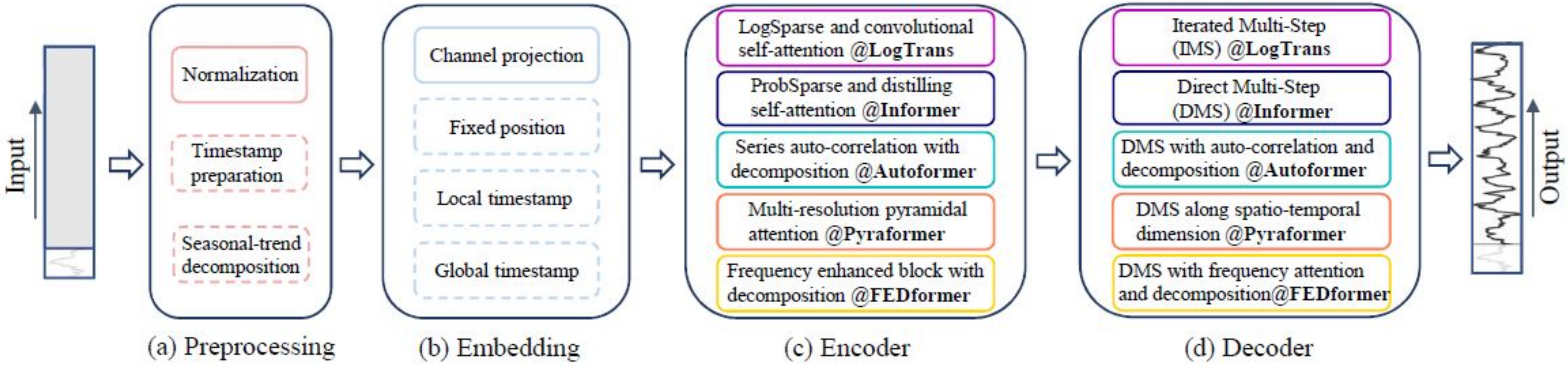
Transformers for forecasting?

- PatchTST - Nie, Yuqi, et al. "A time series is worth 64 words: Long-term forecasting with transformers." *arXiv preprint arXiv:2211.14730*.ICLR,2023
- QuatFormer - Chen, Weiqi, et al. "Learning to rotate: Quaternion transformer for complicated periodical time series forecasting." Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022.
- Zhang, Yunhao, and Junchi Yan. "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting." The eleventh international conference on learning representations. 2022.
- Liu, Yong, et al. "itransformer: Inverted transformers are effective for time series forecasting." arXiv preprint arXiv:2310.06625 (2023).ICLR2024
- Wu, Haixu, et al. "Timesnet: Temporal 2d-variation modeling for general time series analysis." The eleventh international conference on learning representations. 2023
- Zeng, Ailing, et al. **"Are transformers effective for time series forecasting?."** Proceedings of the AAAI conference on artificial intelligence. Vol. 37. No. 9. 2023.

Attention at Timeseries Forecasting?



Modeling the problem with Transformer

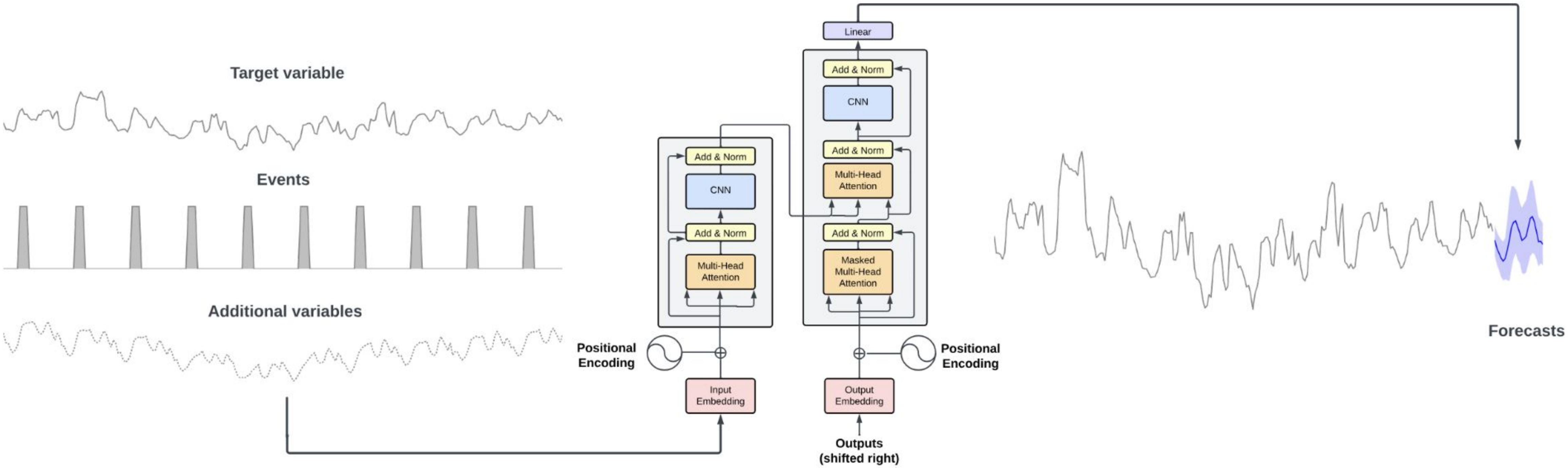


Universal Forecasting Models

Pretrained Models

- TimeGPT - Nextla Oct 2023
- Lag-Llama – Monthreal University Oct.2023
- Moment – Carnegie Mellon Univesity Feb.2024
- Moirai - Salesforce AI Research Feb.2024
- TimesFM - Google Research Feb.2024
- Chronos - Amazon Research Mar.2024
- …

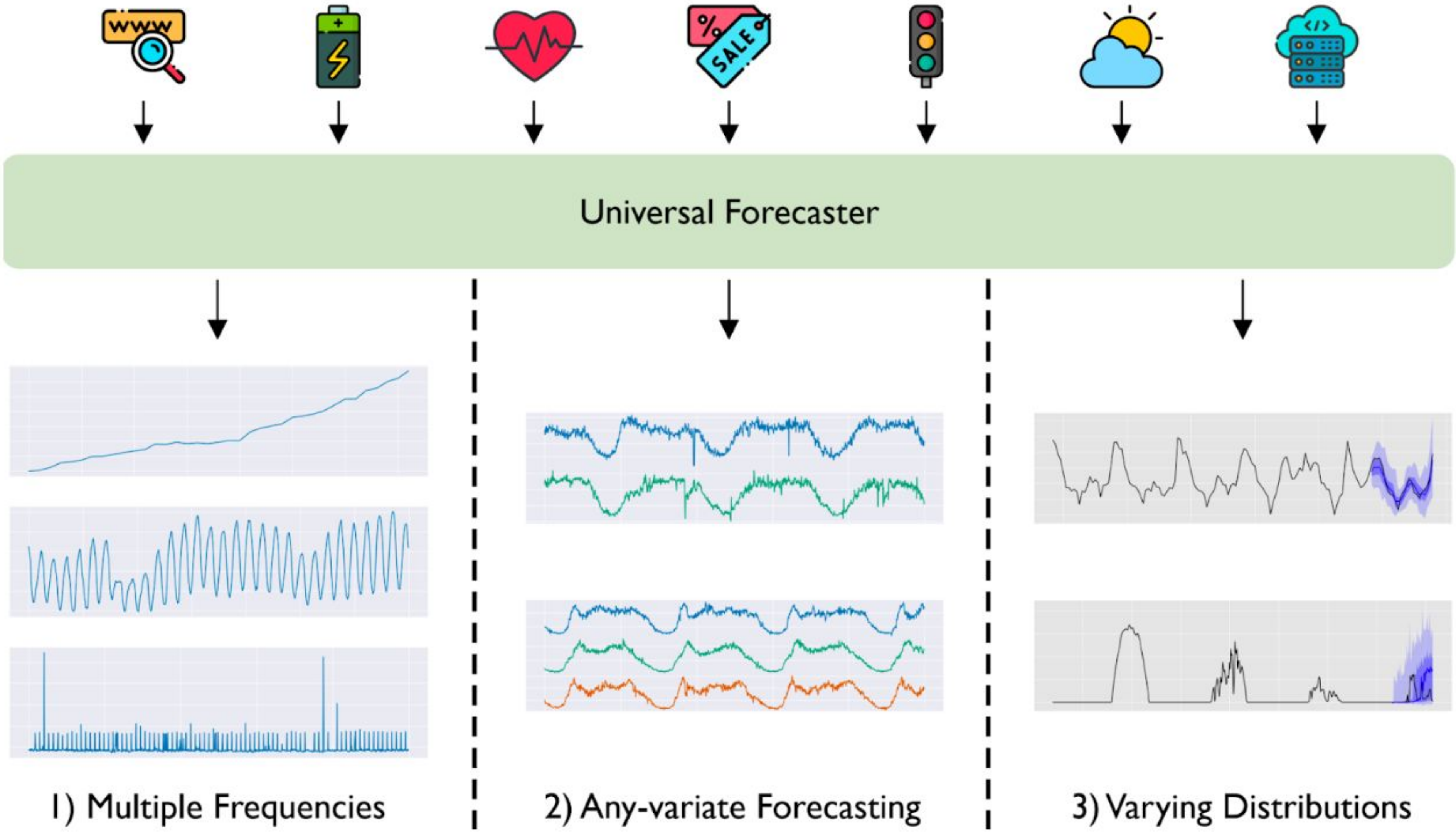
TimeGPT


Target variable
Events
Additional variables
Positional Encoding
Input Embedding
Multi-Head Attention
Add & Norm
CNN
Add & Norm
Output Embedding
Outputs (shifted right)
Positional Encoding
Masked Multi-Head Attention
Add & Norm
Multi-Head Attention
Add & Norm
CNN
Add & Norm
Linear
Forecasts

TimeGPT

	Monthly		Weekly		Daily		Hourly	
	rMAE	rRMSE	rMAE	rRMSE	rMAE	rRMSE	rMAE	rRMSE
ZeroModel	2.045	1.568	6.075	6.075	2.989	2.395	10.255	8.183
HistoricAverage	1.349	1.106	4.188	4.188	2.509	2.057	2.216	1.964
SeasonalNaive	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Theta	0.839	0.764	1.061	1.061	0.841	0.811	1.163	1.175
DOTheta	0.799	0.734	1.056	1.056	0.837	0.806	1.157	1.169
ETS	0.942	0.960	1.079	1.079	0.944	0.970	0.998	1.009
CES	1.024	0.946	1.002	1.002	0.919	0.899	0.878	0.896
ADIDA	0.852	0.769	1.364	1.364	0.908	0.868	2.307	2.207
IMAPA	0.852	0.769	1.364	1.364	0.908	0.868	2.307	2.207
CrostonClassic	0.989	0.857	1.805	1.805	0.995	0.933	2.157	2.043
LGBM	1.050	0.913	0.993	0.993	2.506	2.054	**0.733**	**0.709**
LSTM	0.836	0.778	1.002	1.002	0.852	0.832	0.974	0.955
DeepAR	0.988	0.878	0.987	0.987	0.853	0.826	1.028	1.028
TFT	0.752	0.700	0.954	0.954	0.817	0.791	1.120	1.112
NHITS	0.738	0.694	0.883	0.883	**0.788**	**0.771**	0.829	0.860
TimeGPT	**0.727**	**0.685**	**0.878**	**0.878**	0.804	0.780	0.852	0.878

Moirai



WWW
SALE
%
Universal Forecaster
1) Multiple Frequencies
2) Any-variate Forecasting
3) Varying Distributions

Moirai

Mixture Distribution
Multi Patch Size Output Projection
Patch Size 8
Patch Size 16
Patch Size 32
Patch Size 64
Patch Size 128
Transformer (Full Self-Attention)
Time ID
0 1 2 0 1 2 0 1 2
Variate ID
0 0 0 1 1 1 2 2 2
Patch Embedding
[mask]
[mask]
Multi Patch Size Input Projection
Patch Size 8
Patch Size 16
Patch Size 32
Patch Size 64
Patch Size 128
Variate 0
Variate 1
Variate 2

Moirai

https://blog.salesforceairesearch.com/moirai/

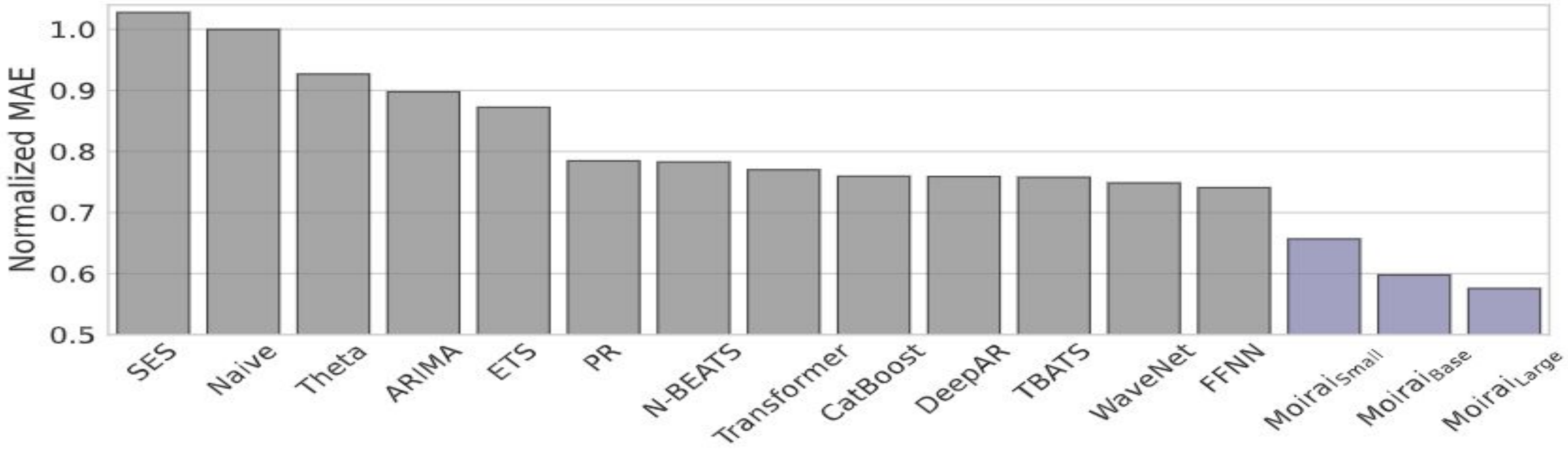


Figure 3. Aggregate results of the Monash Time Series Forecasting Benchmark. The normalized MAE is reported, which normalizes the MAE of each dataset by the naive forecast's MAE, and aggregated by taking the geometric mean across datasets.

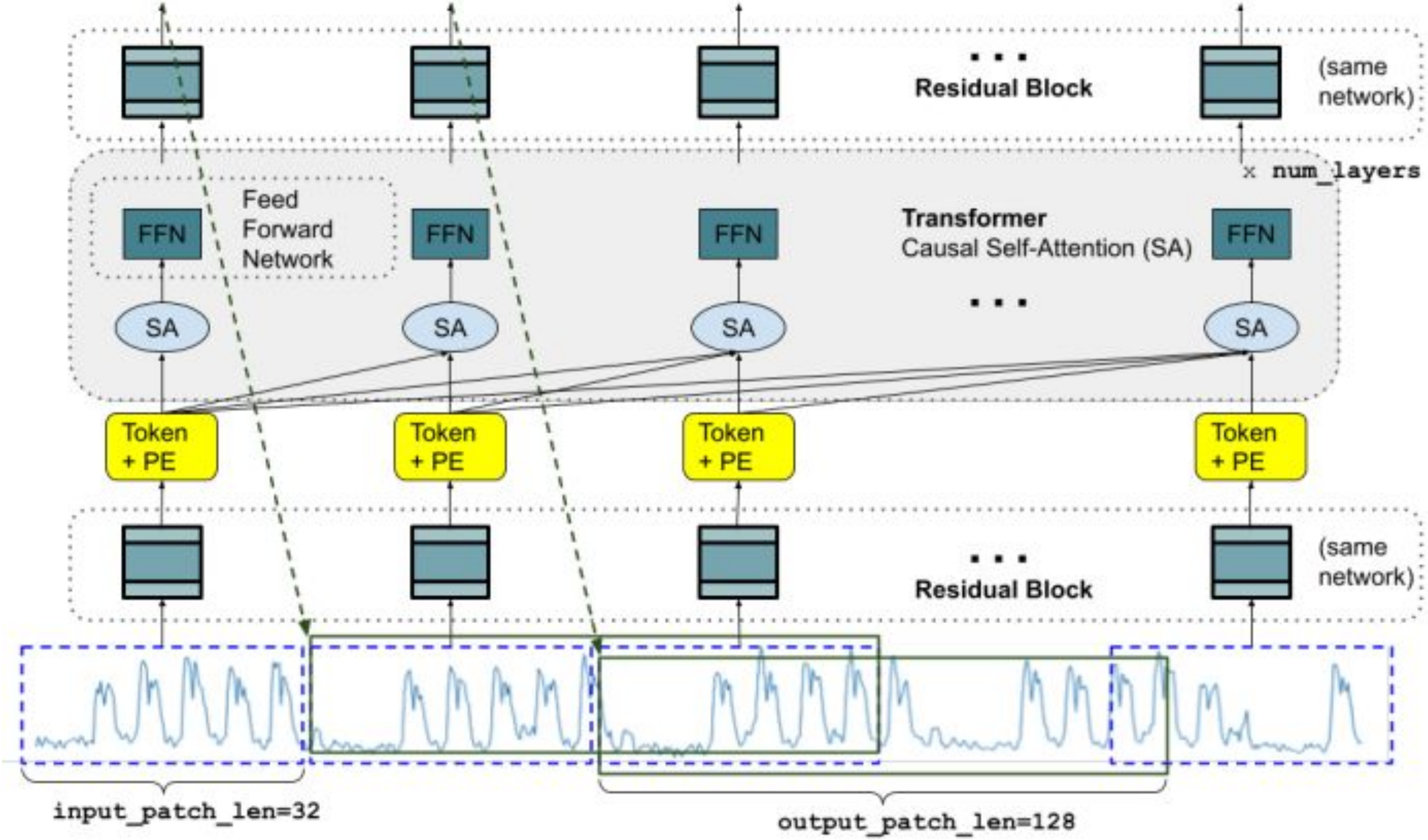
Available at huggungface: https://huggingface.co/Salesforce/moirai-1.0-R-large

TimesFM – google research

- TimesFM (Time Series Foundation Model) is a pretrained time-series foundation model developed by Google Research for time-series forecasting.
- It performs univariate time series forecasting for context lengths up to 512 time points and any horizon lengths, with an optional frequency indicator.
- It focuses on point forecasts and does not support probabilistic forecasts

Model available at huggingface:
https://huggingface.co/google/timesfm-1.0-200m

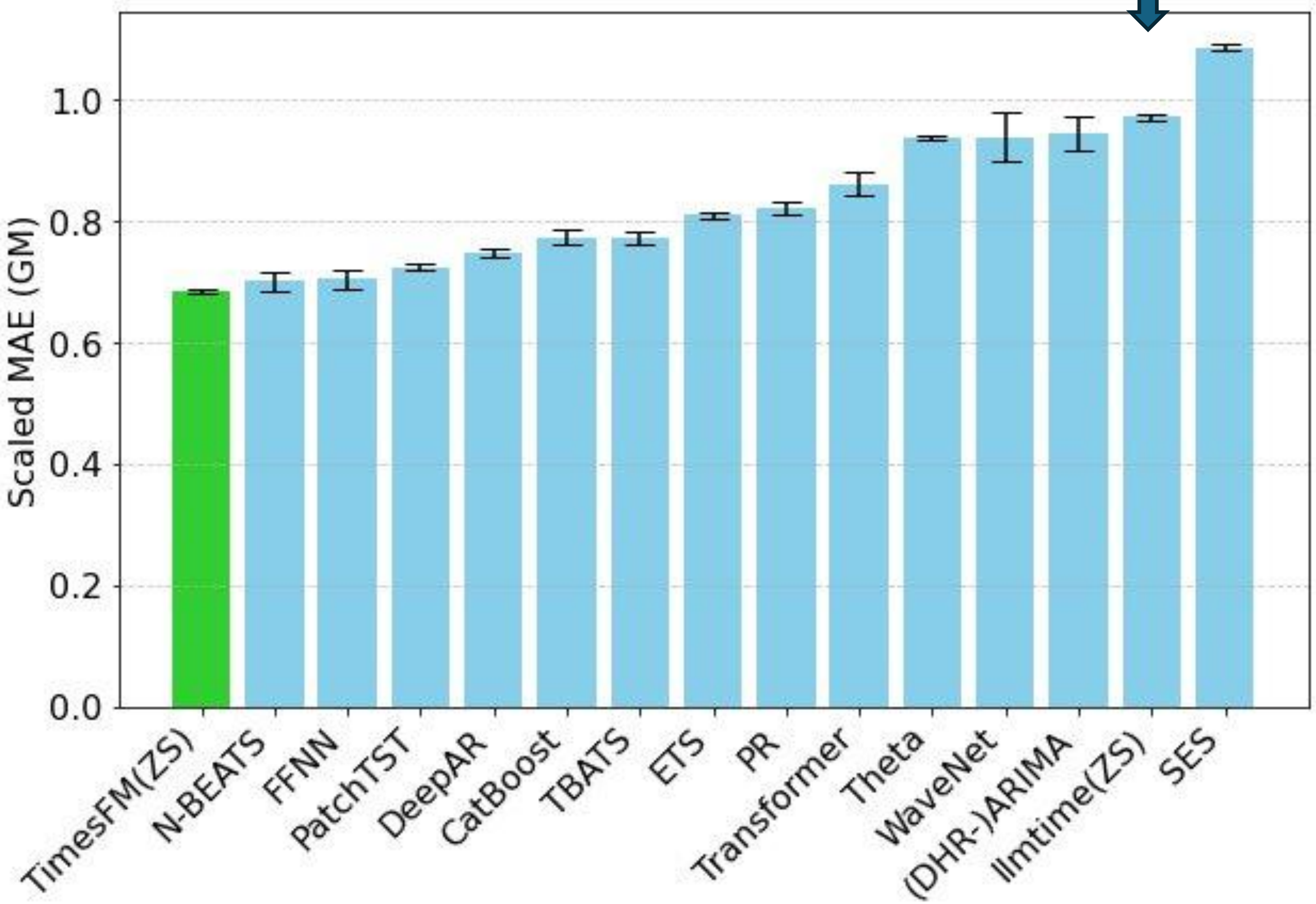
TimesFM



Residual Block
(same network)
× num_layers
Feed Forward Network
FFN
FFN
FFN
FFN
Transformer
Causal Self-Attention (SA)
SA
SA
SA
SA
Token + PE
Token + PE
Token + PE
Token + PE
Residual Block
(same network)
input_patch_len=32
output_patch_len=128

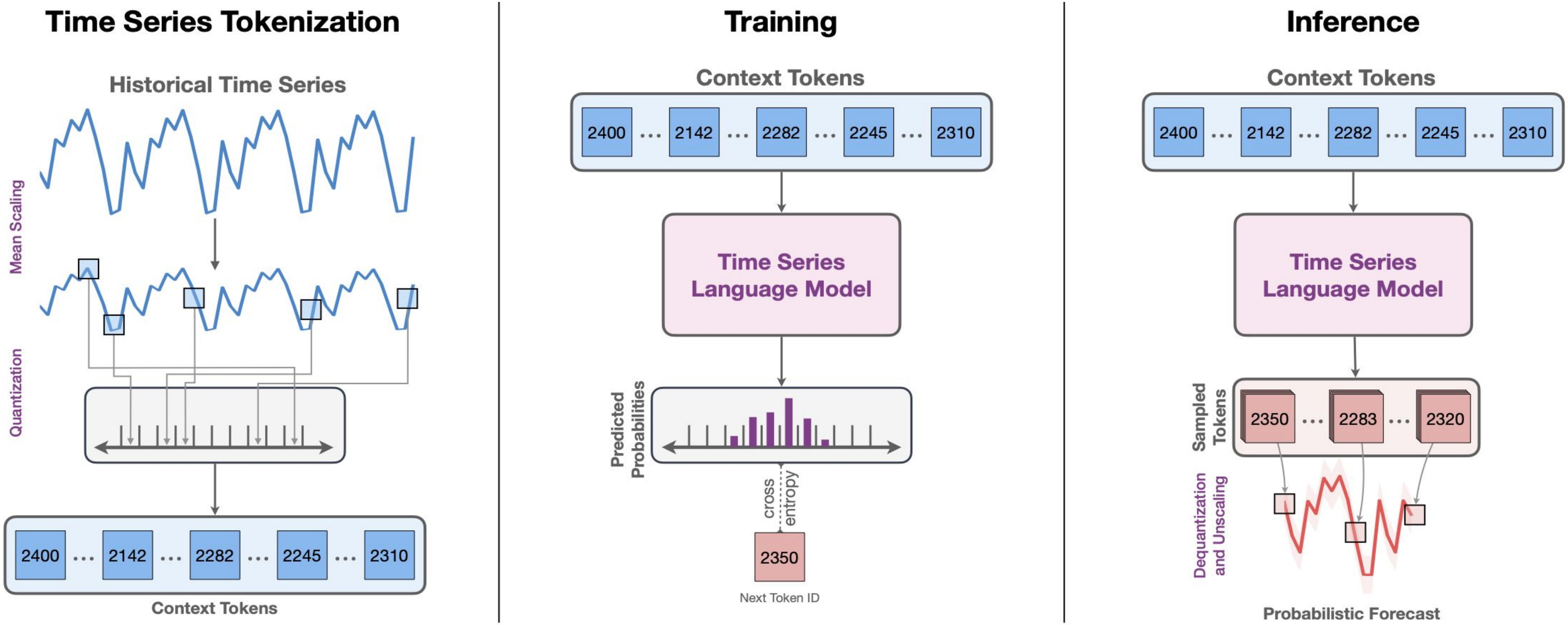
TimesFM

Using GPT 3.5

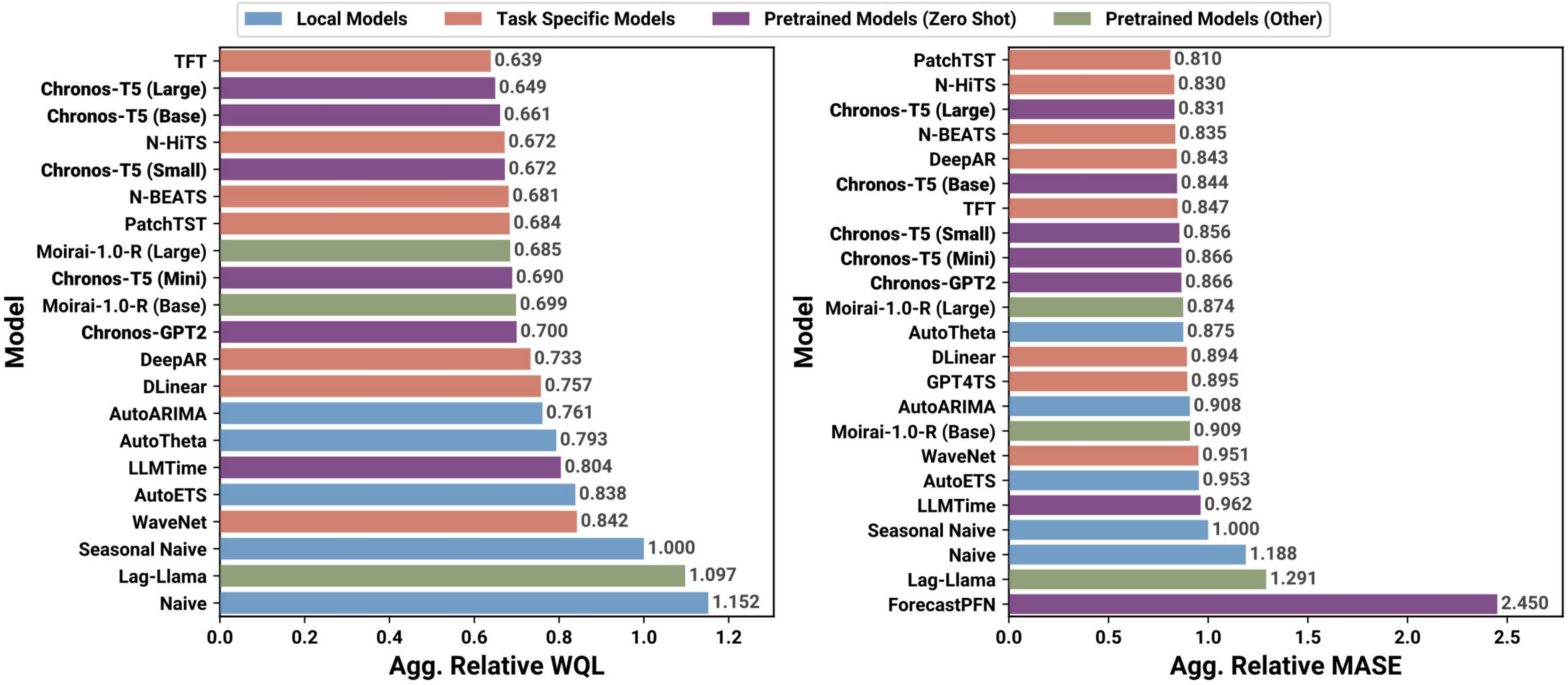


Chronos – amazon research

The models are based on the T5 architecture



Chronos



Does Chronos work with covariates or features?¶

The current iteration of Chronos does not support covariates or features, however we will provide this functionality in later versions.

Summary and conclusions

- The usage of Transformer for forecasting is one of the current hottest topics to for research
- At the last months multiple pretrained models have been released as all they are claiming strong zero shot performance
- Events are tricky to be used with pretrained models as they could have a lot different behaviour compared with the datasets on which the models was trained.

Transformers for forecasting?

- Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 12. 2021.
- Lim, Bryan, et al. "Temporal fusion transformers for interpretable multi-horizon time series forecasting." *International Journal of Forecasting* 37.4 (2021)
- Wu, Haixu, et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." Advances in neural information processing systems 34 (2021)
- Woo, Gerald, et al. "Etsformer: Exponential smoothing transformers for time-series forecasting." arXiv preprint arXiv:2202.01381 (2022).
- Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." International conference on machine learning. PMLR, 2022.

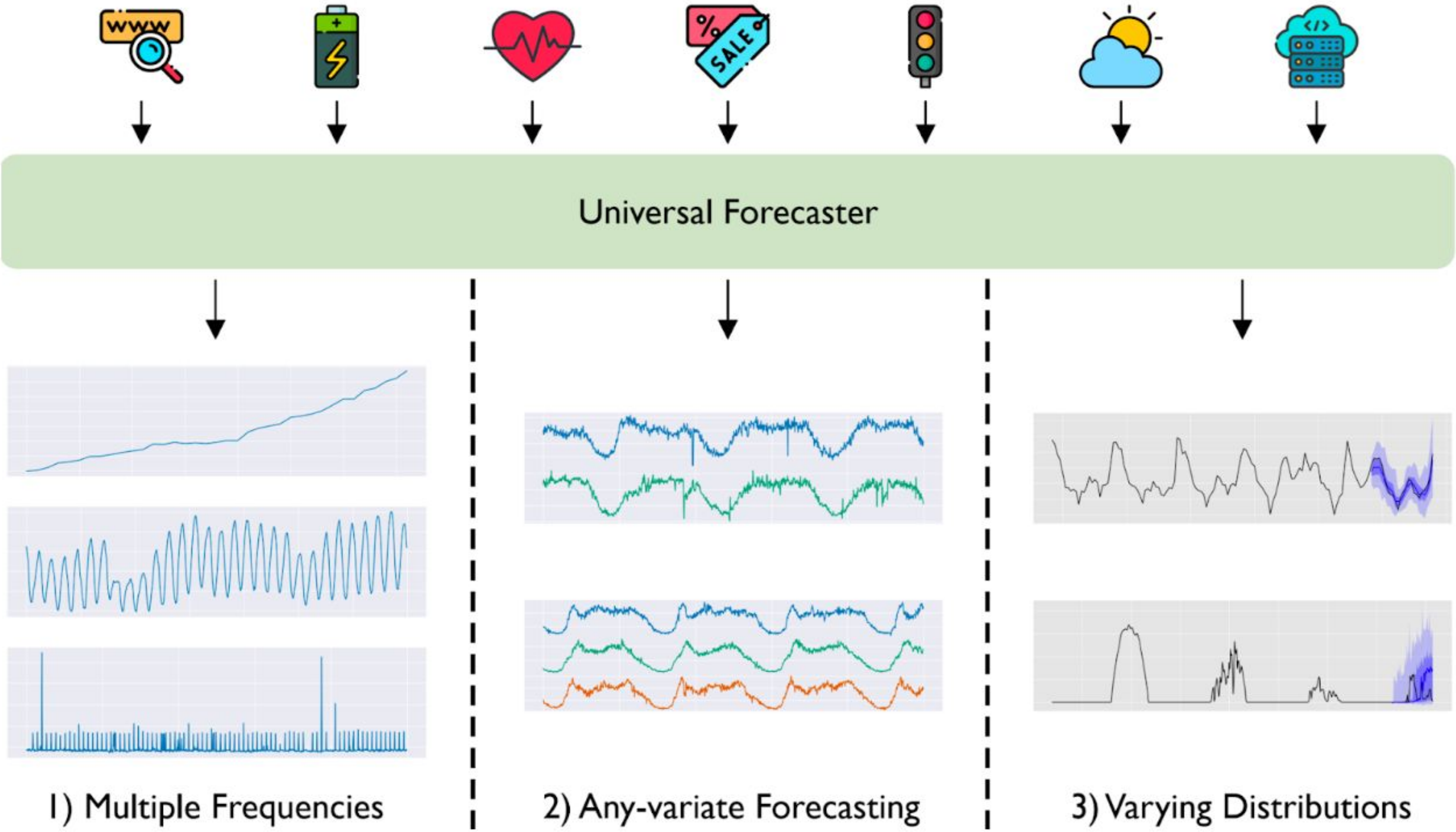
Transformers for forecasting?

- PatchTST - Nie, Yuqi, et al. "A time series is worth 64 words: Long-term forecasting with transformers." *arXiv preprint arXiv:2211.14730*.ICLR,2023
- QuatFormer - Chen, Weiqi, et al. "Learning to rotate: Quaternion transformer for complicated periodical time series forecasting." Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022.
- Zhang, Yunhao, and Junchi Yan. "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting." The eleventh international conference on learning representations. 2022.
- Liu, Yong, et al. "itransformer: Inverted transformers are effective for time series forecasting." arXiv preprint arXiv:2310.06625 (2023).ICLR2024
- Wu, Haixu, et al. "Timesnet: Temporal 2d-variation modeling for general time series analysis." The eleventh international conference on learning representations. 2023
- Zeng, Ailing, et al. **"Are transformers effective for time series forecasting?."** Proceedings of the AAAI conference on artificial intelligence. Vol. 37. No. 9. 2023.

Pretrained Models

- Lag-Llama – Monthreal University Oct.2023
- Moment – Carnegie Mellon Univesity Feb.2024
- Moirai - Salesforce AI Research Feb.2024
- TimesFM - Google Research Feb.2024
- Chronos - Amazon Research Mar.2024

Moirai


WWW
SALE
Universal Forecaster
1) Multiple Frequencies
2) Any-variate Forecasting
3) Varying Distributions

Moirai


Mixture Distribution
Multi Patch Size Output Projection
Patch Size 8
Patch Size 16
Patch Size 32
Patch Size 64
Patch Size 128
Transformer (Full Self-Attention)
Time ID
0
1
2
0
1
2
0
1
2
Variate ID
0
0
0
1
1
1
2
2
2
Patch Embedding
[mask]
[mask]
Multi Patch Size Input Projection
Patch Size 8
Patch Size 16
Patch Size 32
Patch Size 64
Patch Size 128
Variate 0
Variate 1
Variate 2

Moirai

https://blog.salesforceairesearch.com/moirai/

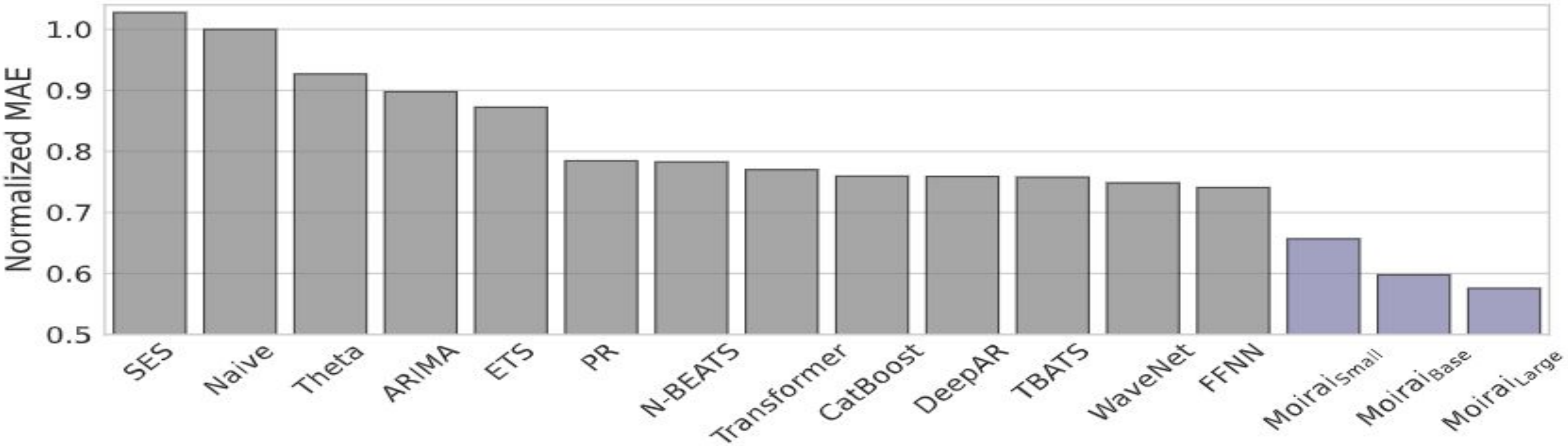


Figure 3. Aggregate results of the Monash Time Series Forecasting Benchmark. The normalized MAE is reported, which normalizes the MAE of each dataset by the naive forecast's MAE, and aggregated by taking the geometric mean across datasets.

Available at huggungface: https://huggingface.co/Salesforce/moirai-1.0-R-large

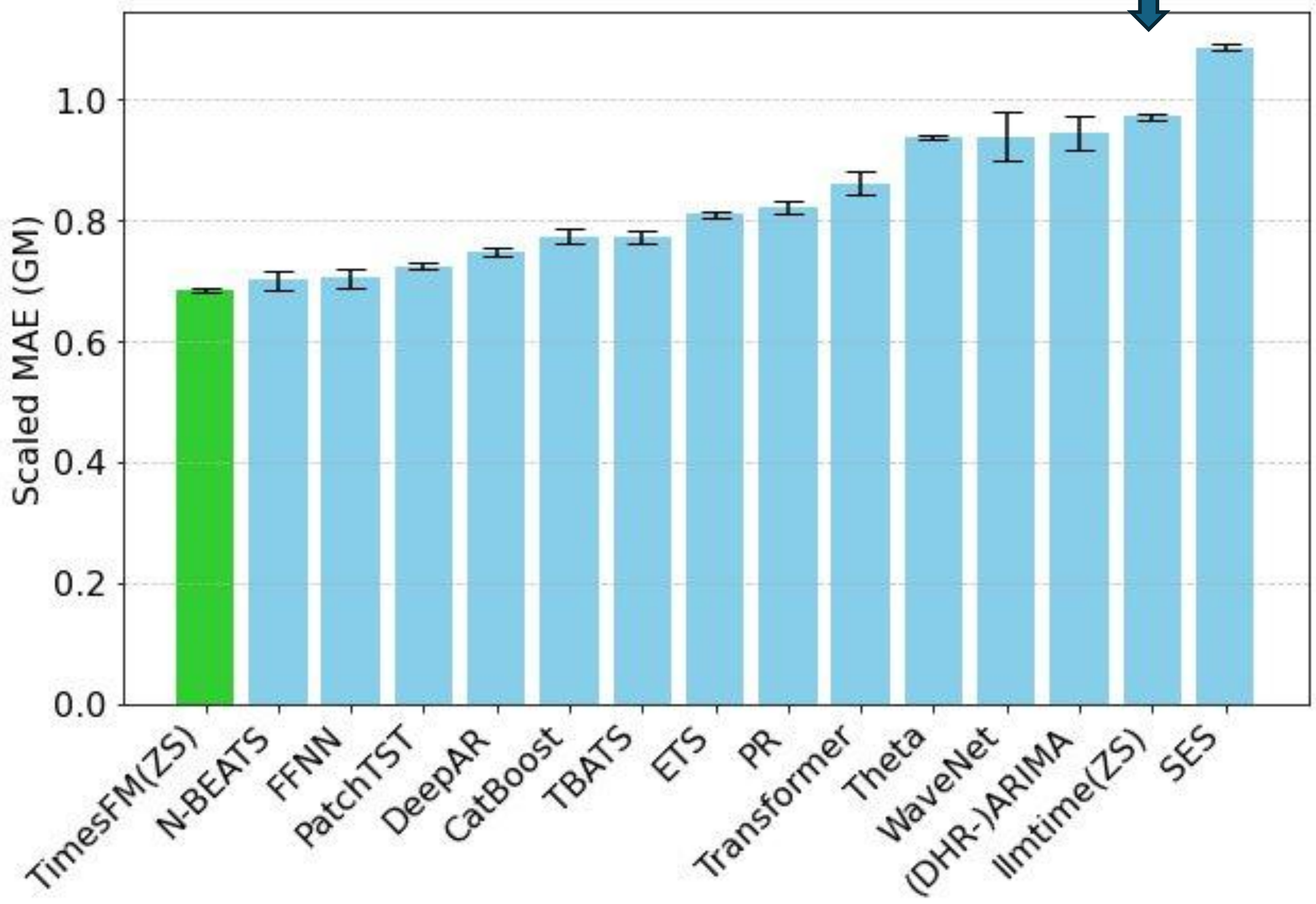
TimesFM – google research

- **TimesFM (Time Series Foundation Model) is a pretrained time-series foundation model developed by Google Research for time-series forecasting.**
- **It performs univariate time series forecasting for context lengths up to 512 time points and any horizon lengths, with an optional frequency indicator.**
- **It focuses on point forecasts and does not support probabilistic forecasts**

Model available at huggingface:
https://huggingface.co/google/timesfm-1.0-200m

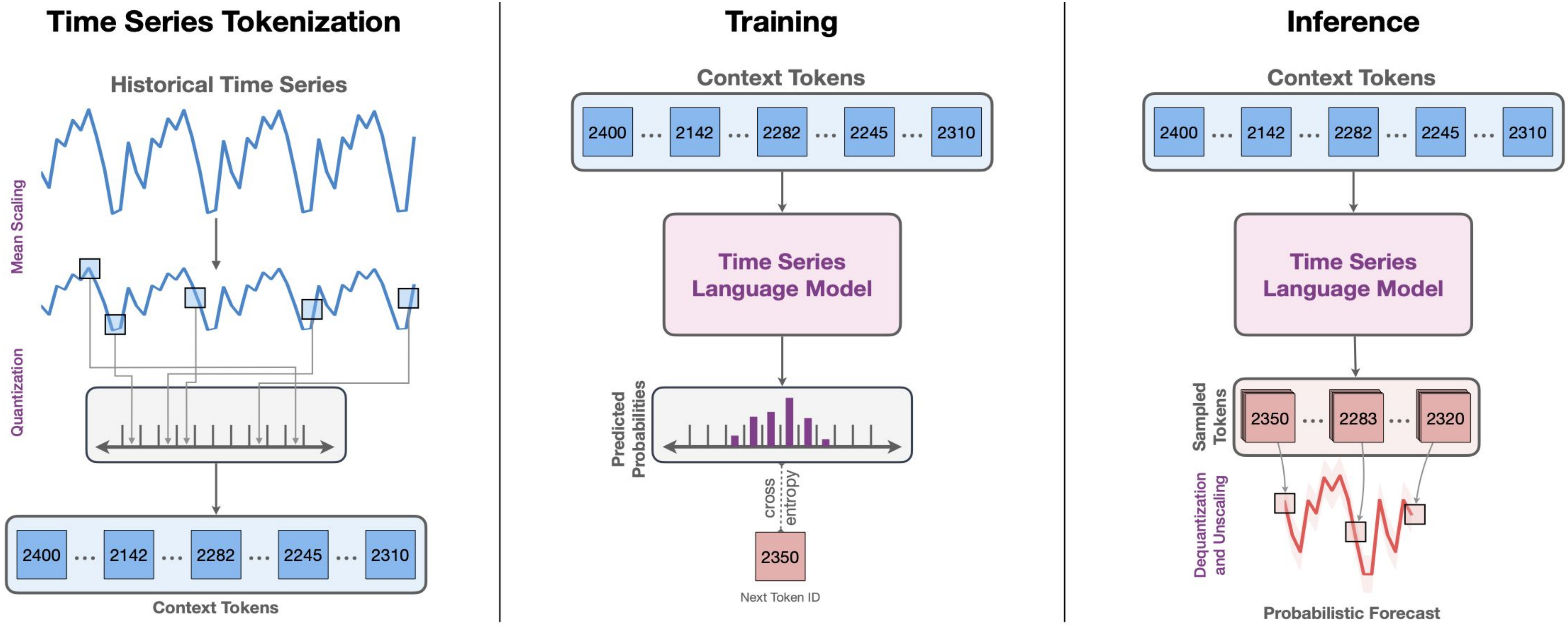
TimesFM

Using GPT 3.5

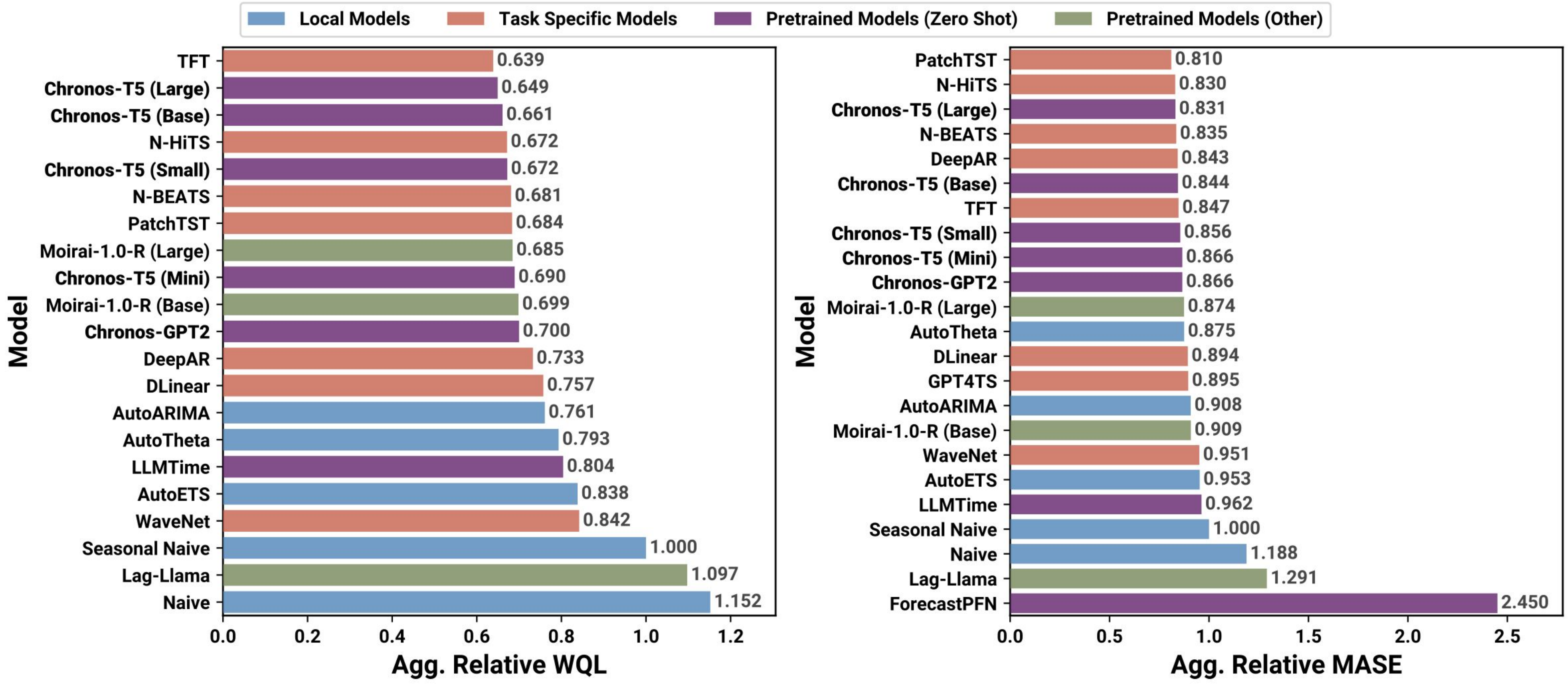


Chronos – amazon research

The models are based on the T5 architecture



Chronos



Does Chronos work with covariates or features?¶

The current iteration of Chronos does not support covariates or features, however we will provide this functionality in later versions.

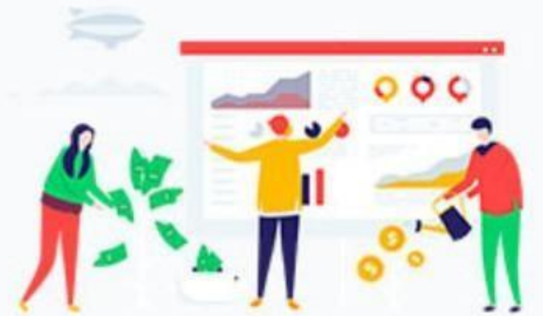
Summary and conclusions

- The current state of the art is to solve forecasting problems with ML algorithms (due to the complexity of the data the)
- Transformer models are heavily researched for forecasting
- At the last 6 months there are multiple pretrained models released as this is the fastest growing direction right now
- A combination of pretrained models in ensemble with another models for utilizing more features is a possible strategy to introduce events

PyData SOFIA

Data Science Society

MEETUP



Q&A

Modern Forecasting approaches and challenges

Milen Chechev

Head of Data science

Fourth

When? 15/05/2024, 18:00 EET

Where? Tsarigradsko shose 125, Faculty of Economics and Business Administration, Sofia University, hall 200

Next Event:

27.05.2024 "Apache Arrow – Exploring the technology that powers the modern data (science) stack",   Uwe Korn, CTO QuantCo ,