Deep Factors for Forecasting

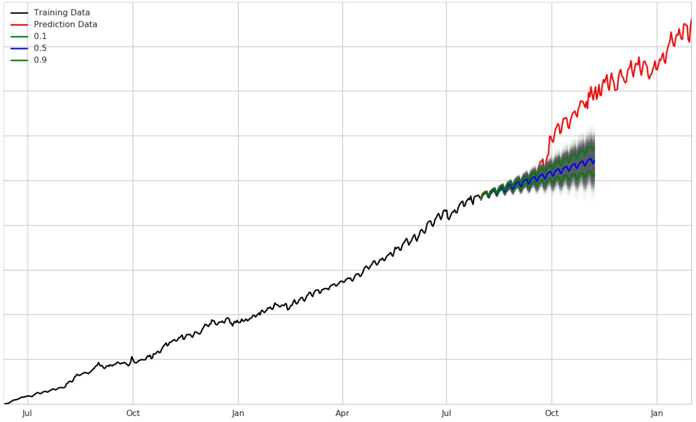
Yuyang Wang, Alex Smola, Danielle C. Maddix,
Jan Gasthaus, Dean Foster, Tim Januschowski

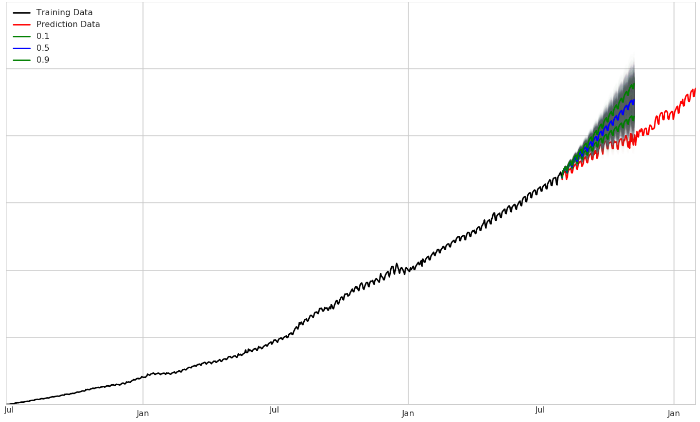
Amazon AI

aws

Time Series Prediction … at Amazon

weekly units
shipped
and **forecast**
years ahead





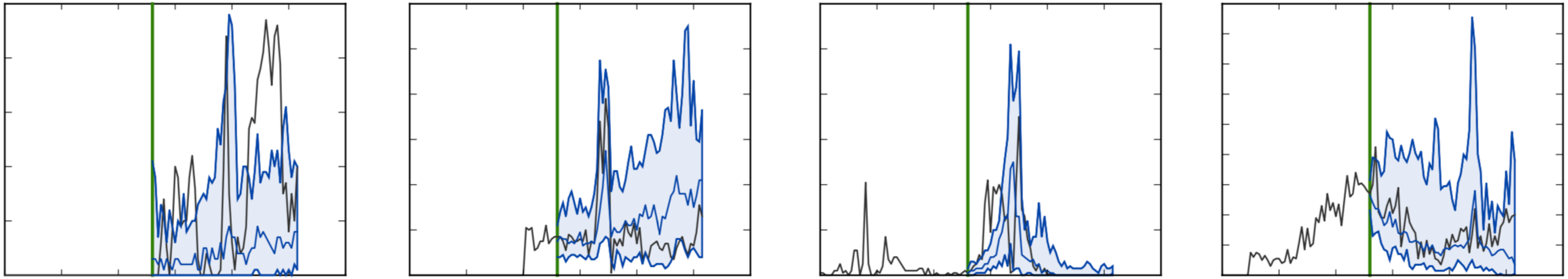
servers
forecast
and **used**

Capacity planning

Market entry

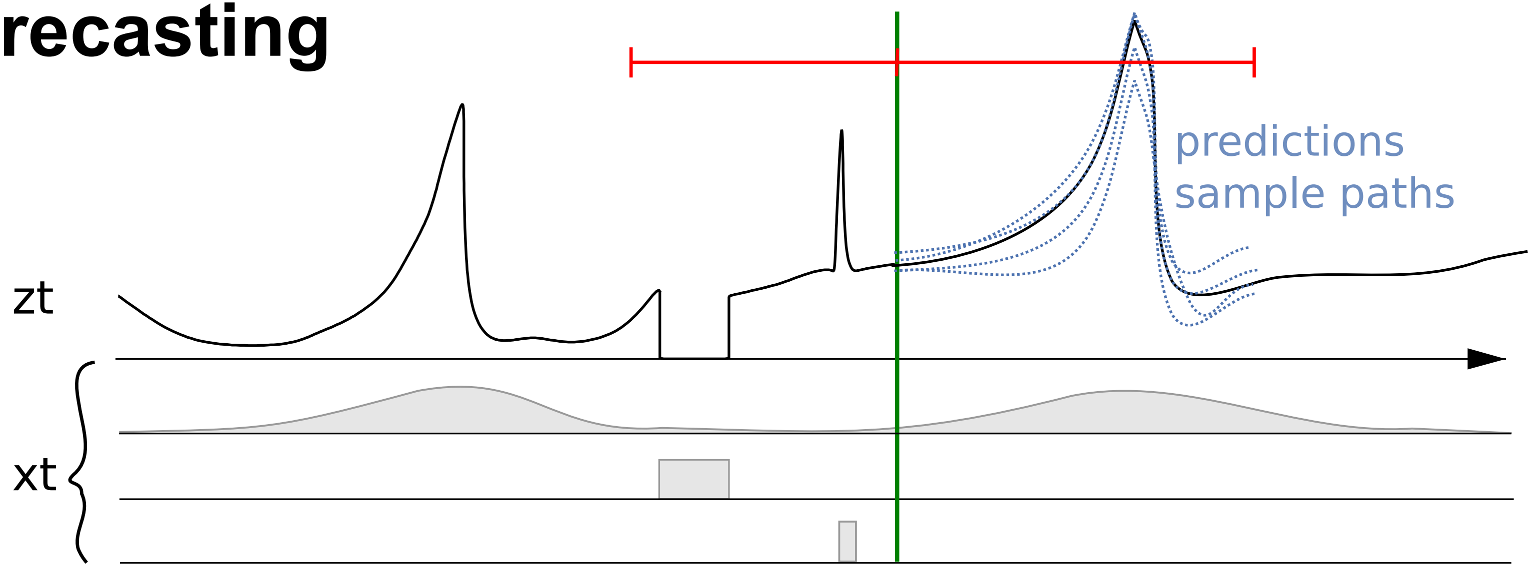
Topology Planning

Time Series Prediction … at Amazon



- Predict demand for each product available at Amazon
- Problem
  - How many items to order
  - Where to order
  - When to mark down (ugly sweaters after Christmas)

Forecasting



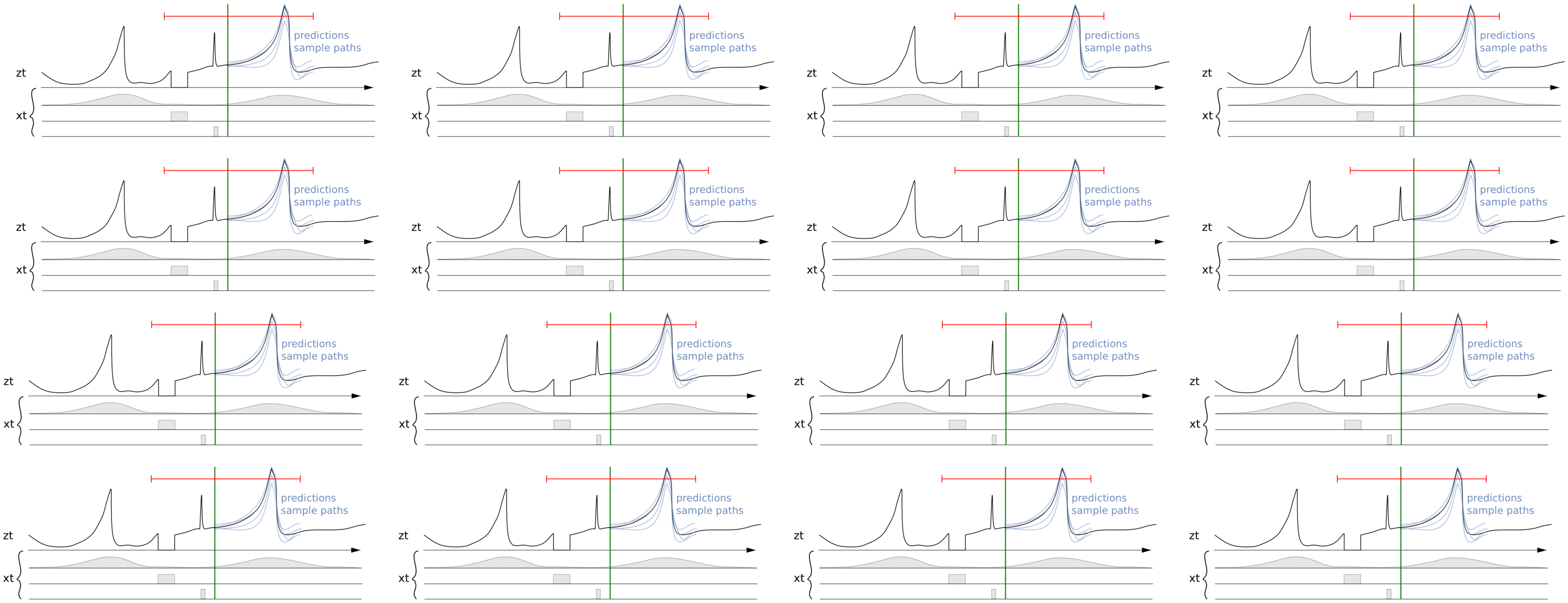
Estimate future observations (univariate case)

$$p(z_{t+1} \,|\, (x_1, z_1), \ldots (x_t, z_t), x_{t+1})$$

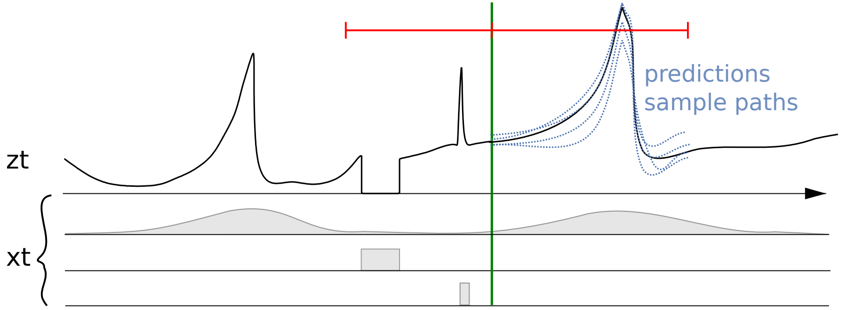
Make optimal decisions

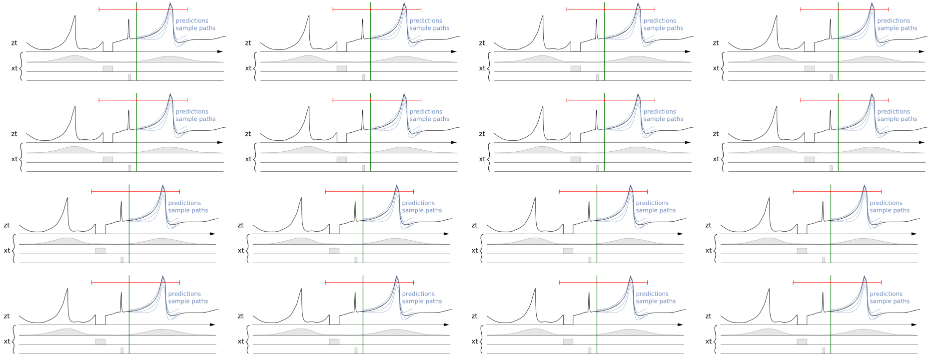
$$\mathrm{argmin}_a \mathbb{E}_{z_{t+1}|\mathrm{past}}[l(a, z_{t+1}, \mathrm{past})]$$

But in reality …



Two old ideas





- Model each time series locally and individually
  - Easy to add more
  - Simple models
  - Doesn't work so well
- Model all time series jointly and globally
  - Works better
  - Impossible to add more
  - Complex model

… make a good one

- **Local model**
  - Reads from global model
  - Updates local state
- **Global model**
  - Nonlinear backbone
  - Nonparametric
- **Theorem (deFinetti for time series)**
  For an **exchangeable** distribution over time series the joint distribution can be written as a local/global model.

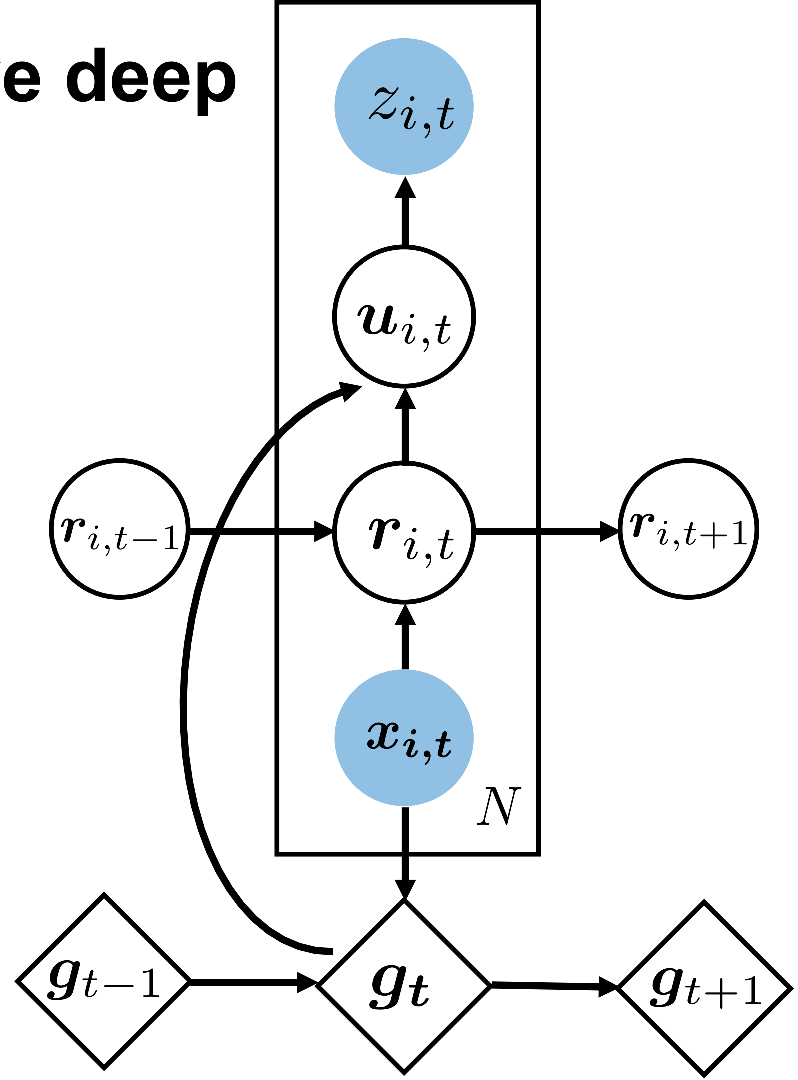
$$p(x^i_{1,\ldots,T} \text{ for } i \in \{1,\ldots N\}) = \int dg \prod_{t=1}^{T} p(g_t \,|\, g_{t-1}, \ldots g_1) \prod_{i=1}^{N} p(x^i_t \,|\, x^i_{t-1}, \ldots, x^i_t, g_t, \ldots g_1)$$

- Corresponding result for trees, too (via Tree-deFinetti)

Latent variable autoregressive deep

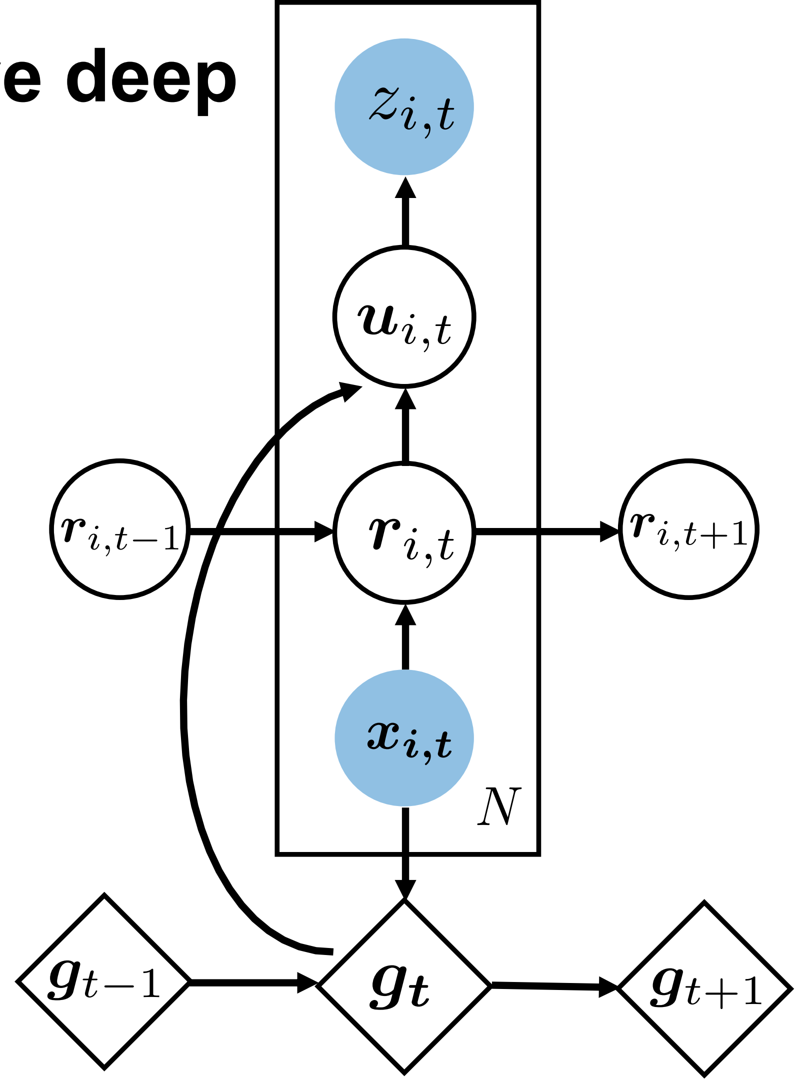
- **Global model**
  - Nonlinear backbone
  - Nonparametric
- **Local model**
  - Reads from global model
  - Updates local state

- Works well in practice



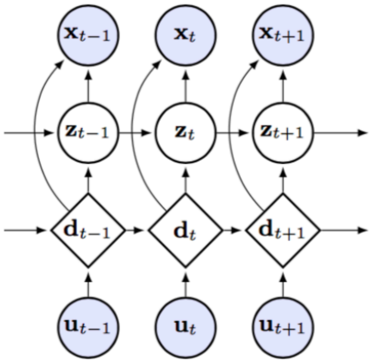
Latent variable autoregressive deep

- **Global model**
  - Nonlinear backbone
  - Nonparametric (e.g. LSTM) to handle complex dependence
- **Local model**
  - Reads from global model
  - Use LDS / Gaussian Process for speed and uncertainty characterization
- **Inference**
  - Exact for Gaussian
  - VAE for other likelihoods

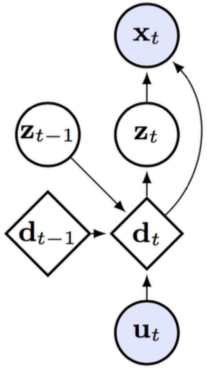


We are not the first to realize this …

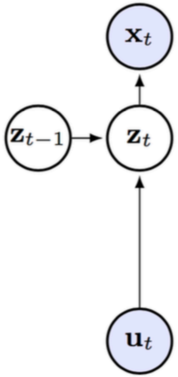
SRNN [Fraccaro et al., 2016]



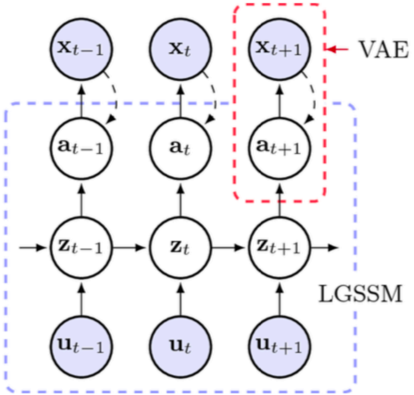
VRNN/SSL/LSTM-LDA [Chung et al., 2015; Zaheer et al., 2017; Zheng et al., 2017]



DMM [Krishnan et al., 2017, 2015]

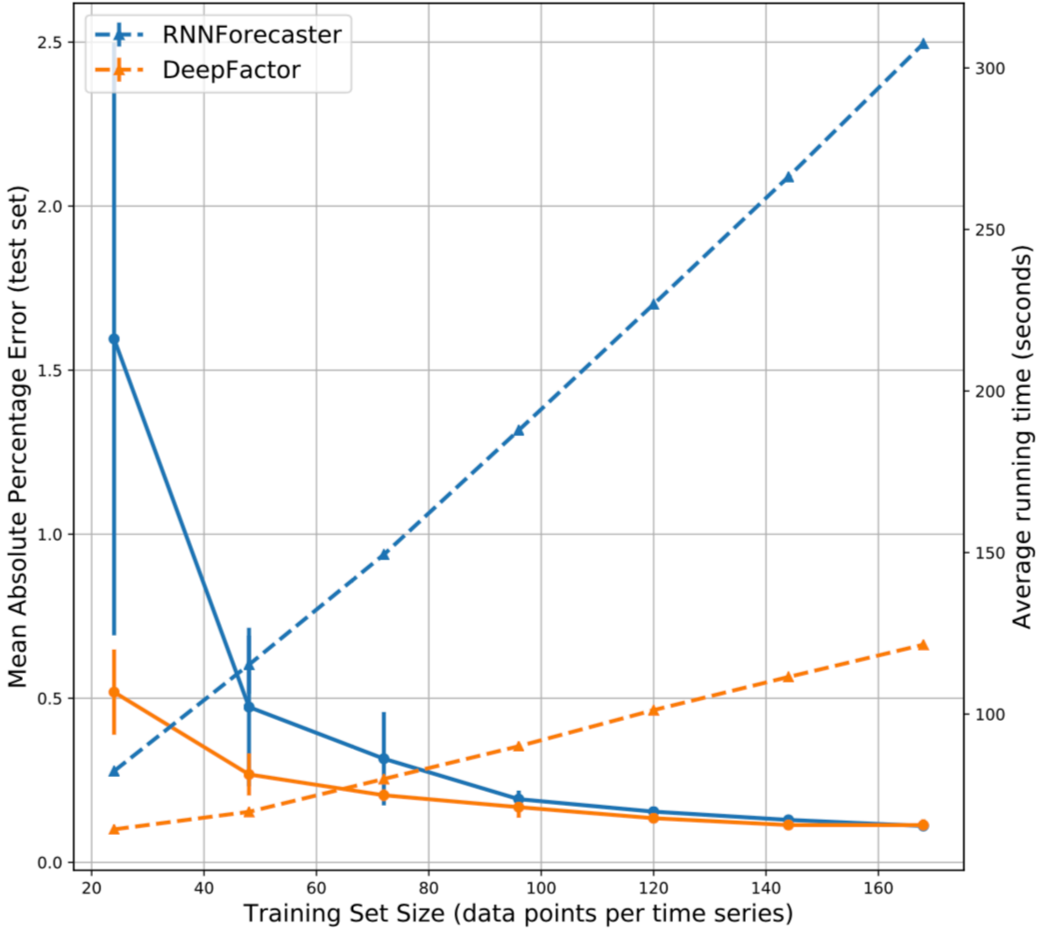


KVAE/DVBF [Fraccaro et al., 2017; Karl et al., 2017]



aws

Faster, stronger, better (than previous work)

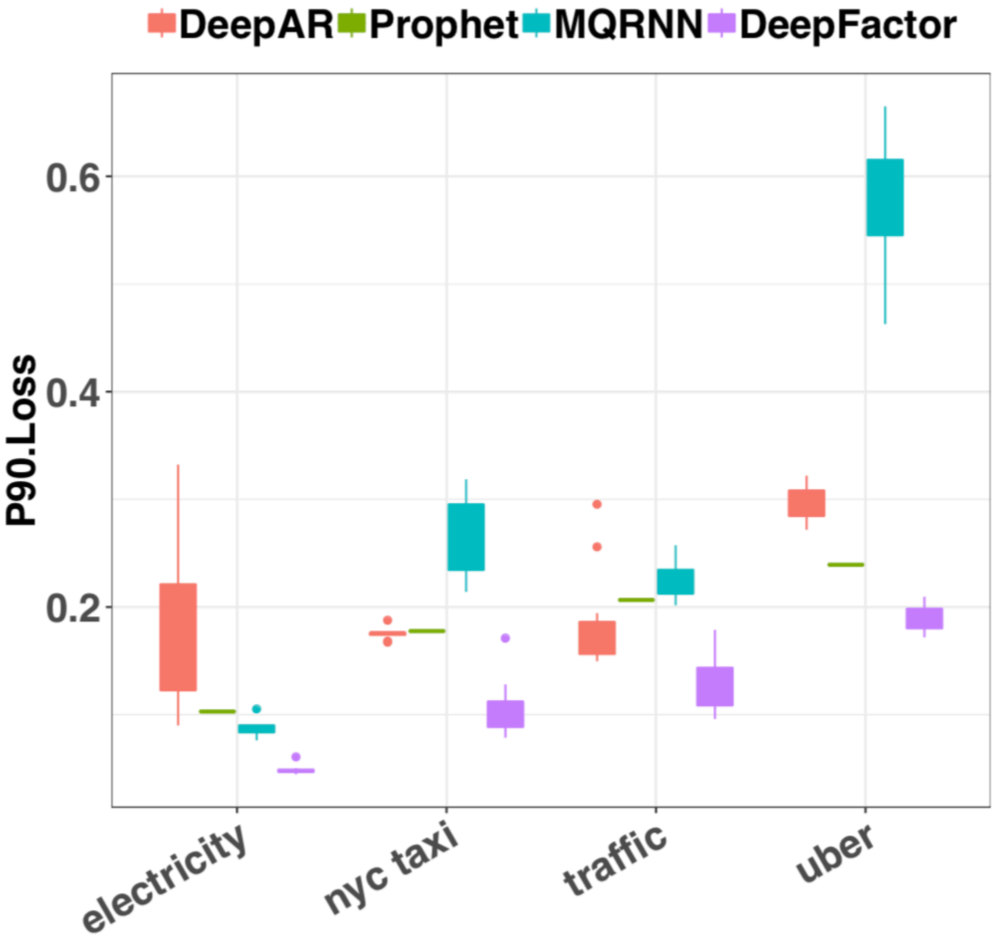


When compared to pure RNN forecaster

- Lower uncertainty
- Faster
- More accurate

aws

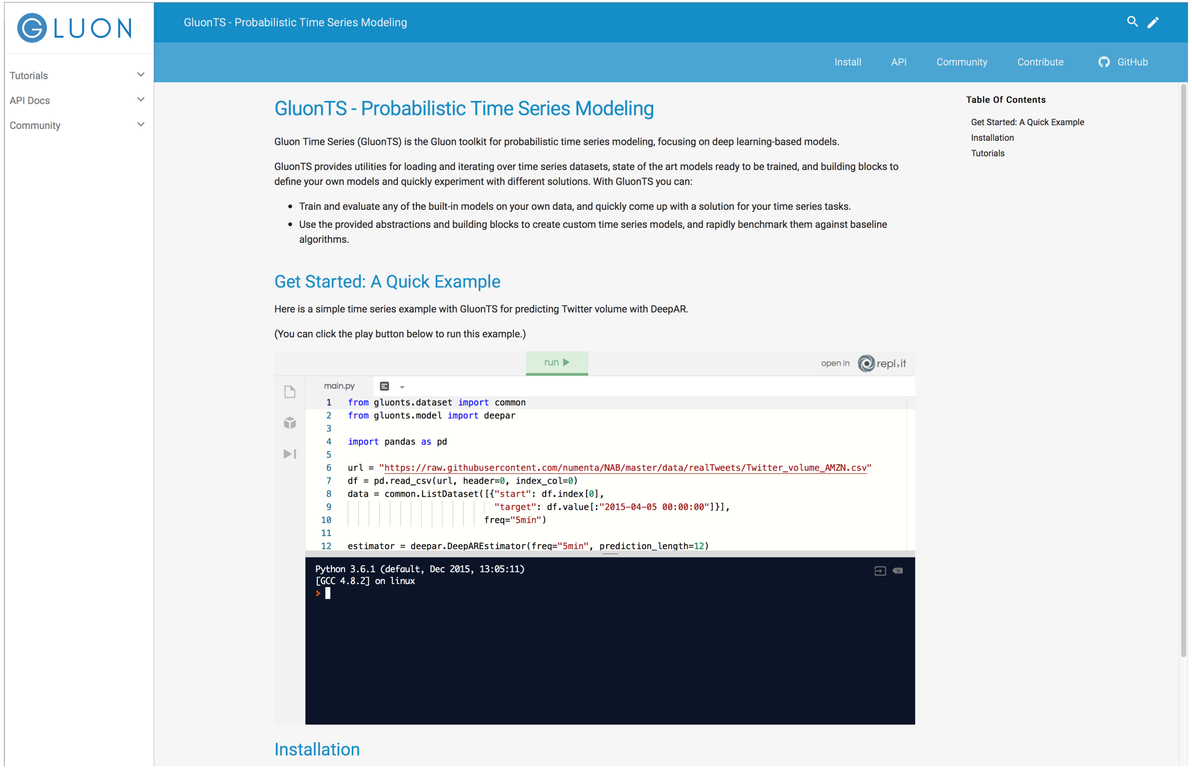
Faster, stronger, better (than previous work)



When compared to SOTA (DeepAR and MQ-RNN)

- Less variance
- More accurate

To be available in GluonTS: gluon-ts.mxnet.io



Featuring NN-based Forecasting models

- DeepAR [Valentin et. al., 2017]
- MQ-DNN [Wen et. al., 2017]
- Deep State Space Models [Rangapuram et. al., 2018]
- Spline Regression RNN [Gasthaus et. al., 2019]
- GPs, KF, LDS
- More …

Come to our talk at the Time Series Workshop on Friday!

gluon-ts.mxnet.io

aws