# Results of the simulation, validation and discussion

Gent Rexha, Ilir Osmanaj & Princ Mullatahiri

194.049 Energy-efficient Distributed Systems

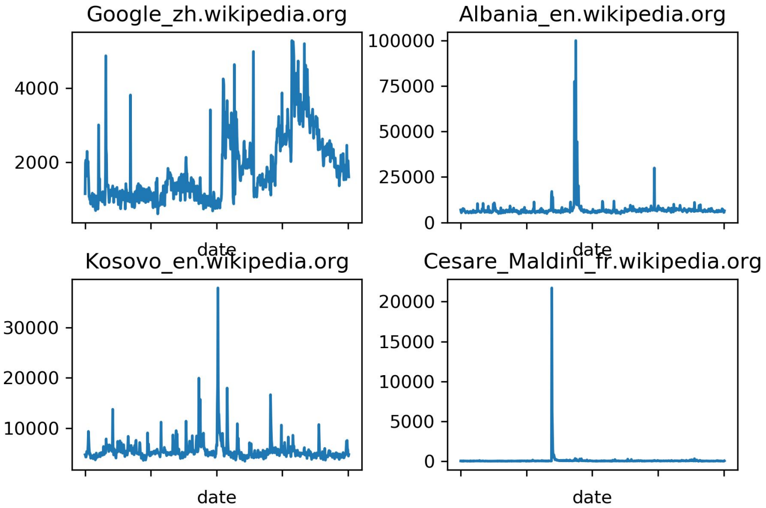
# Presentation outline

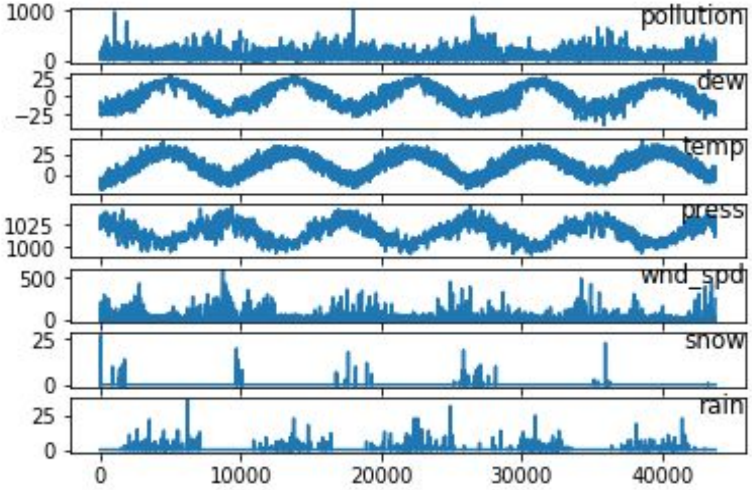
- Datasets
- Forecasting Methods
- Performance measures
- Pre-processing
- Experiments and results
- Conclusion

# Datasets

- Web Traffic Time Series Forecasting:
  - The training dataset consists of approximately 145k time series.
  - We've reduced the dataset to a smaller version with 6 different articles.
  - Each of these time series represent a number of daily views of a different Wikipedia article.
  - Dataset contains name of the article as well as the type of traffic (all, mobile, desktop, spider).
- Air Pollution Forecasting
  - Daily values for pollution, dew, temperature, pressure, wind direction, wind, speed, snow, and rain.
  - Starting from January, 1st, 2010 up until December 31th, 2014.
- Both datasets are from Kaggle

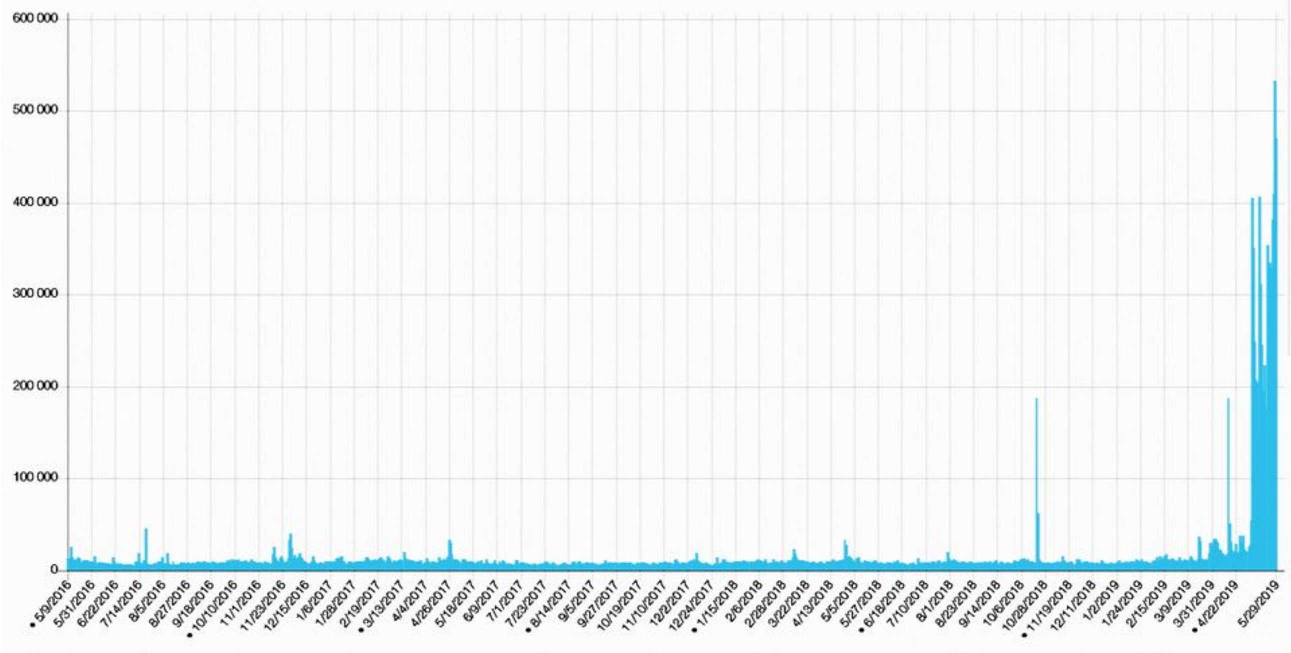
# Datasets cont.



# Datasets cont.

- Air temperature a little bit more predictable because of seasonality
- Web-traffic usually follows trends (e.g. Chernobyl searches)



# Forecasting Methods

- LSTM :
  - Artificial recurrent neural network architecture used in the field of deep learning.
  - It can process complete information sequences like speech or video.
  - These algorithms take time and sequence into account, they have a temporal dimension.
  - LSTMs have been created to cope with the explosive and disappearing gradient issues that can be experienced in traditional RNN training.
- Prophet:
  - Prophet is open source software released by Facebook’s Core Data Science team.
  - It is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
  - It works best with time series that have strong seasonal effects and several seasons of historical data.

# Performance measures

- Root-mean-square error (RMSE) is a frequently used measure of the variations between the expected values of a model or estimator and the observed values.
- In other words, RMSE tells you how concentrated the data is around the line of best fit.

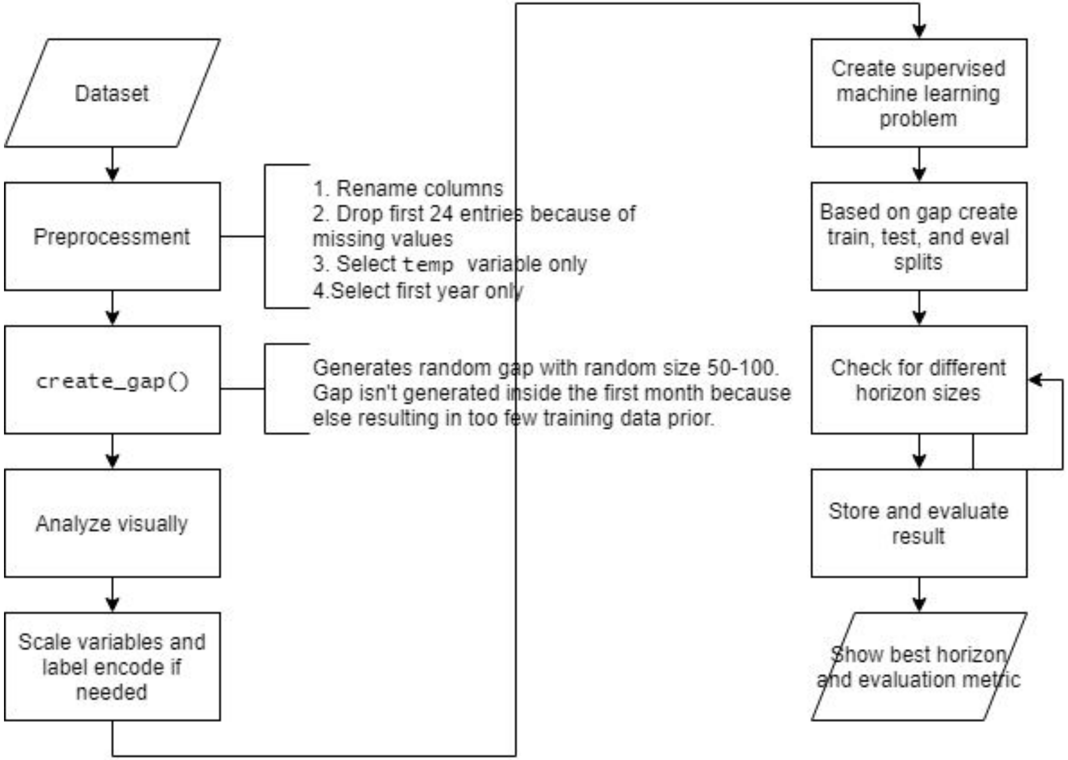
$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

# Pre-processing

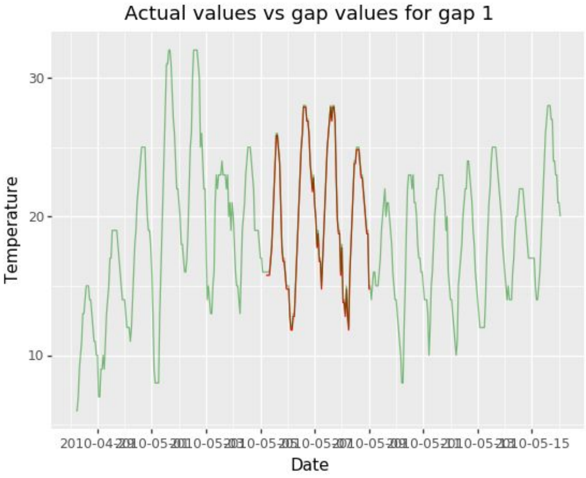
To handle the missing value we've used multiple approaches:

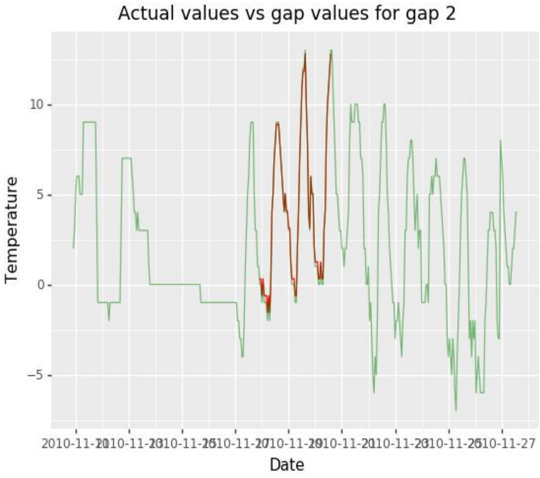
- Manually specify column names
- Drop the first 24 hours because all of them have missing values
- Mark all NA values with 0

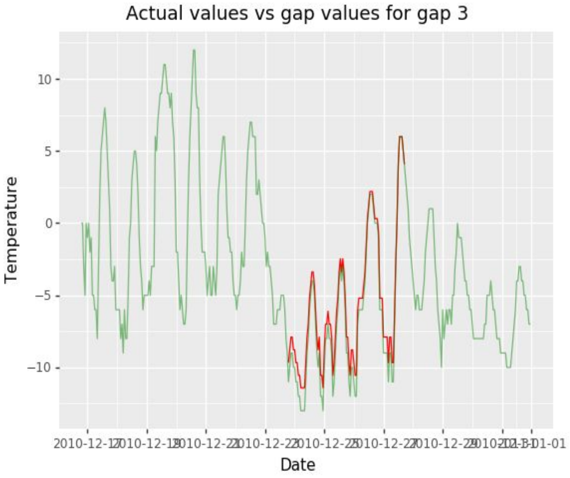
# Experiment Setting



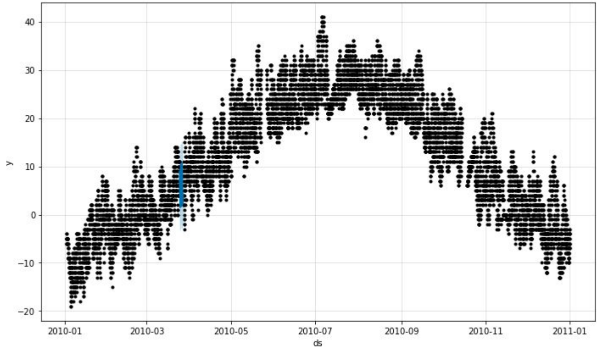
# Experiments and results: Pollution LSTM

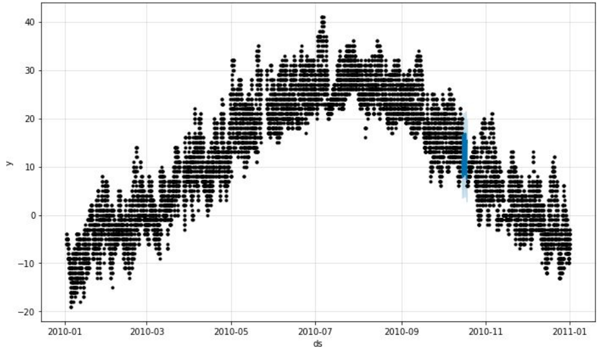


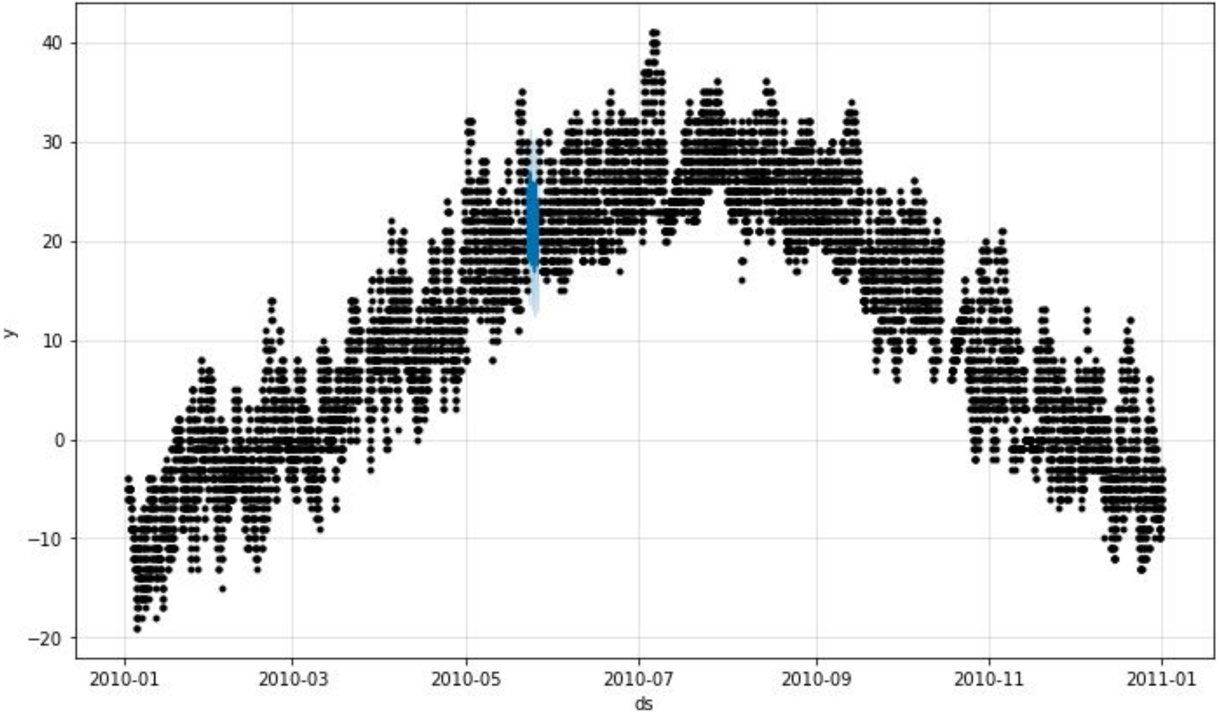



# Experiments and results: Pollution Prophet

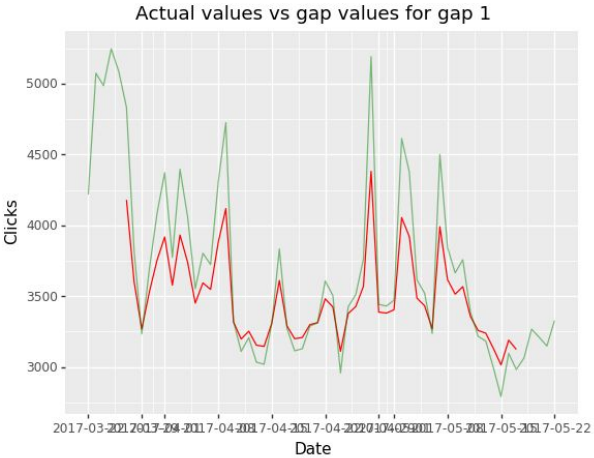


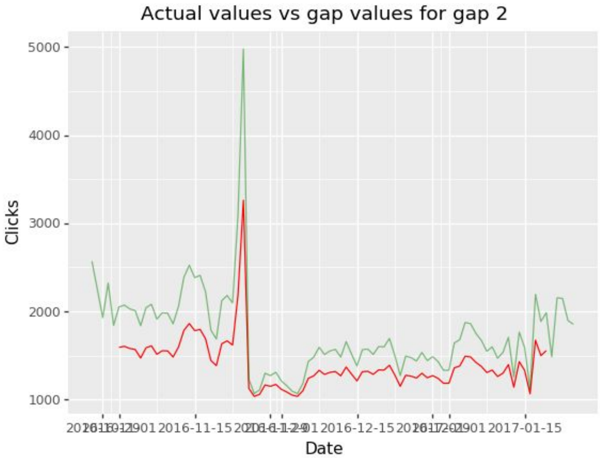



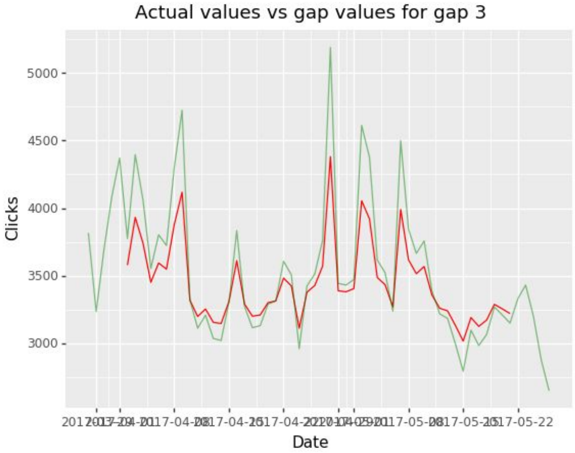



# Experiments and results: RMSE Comparison

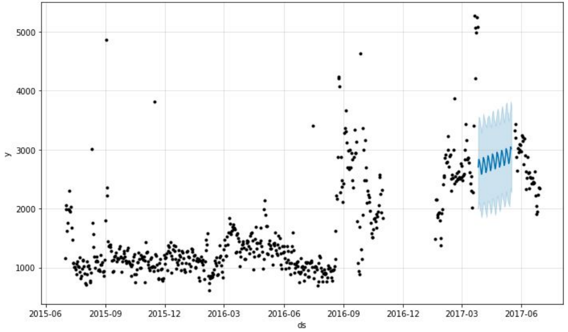
# Experiments and results: Web Traffic LSTM

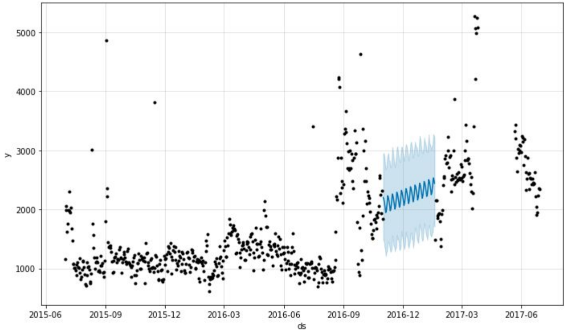


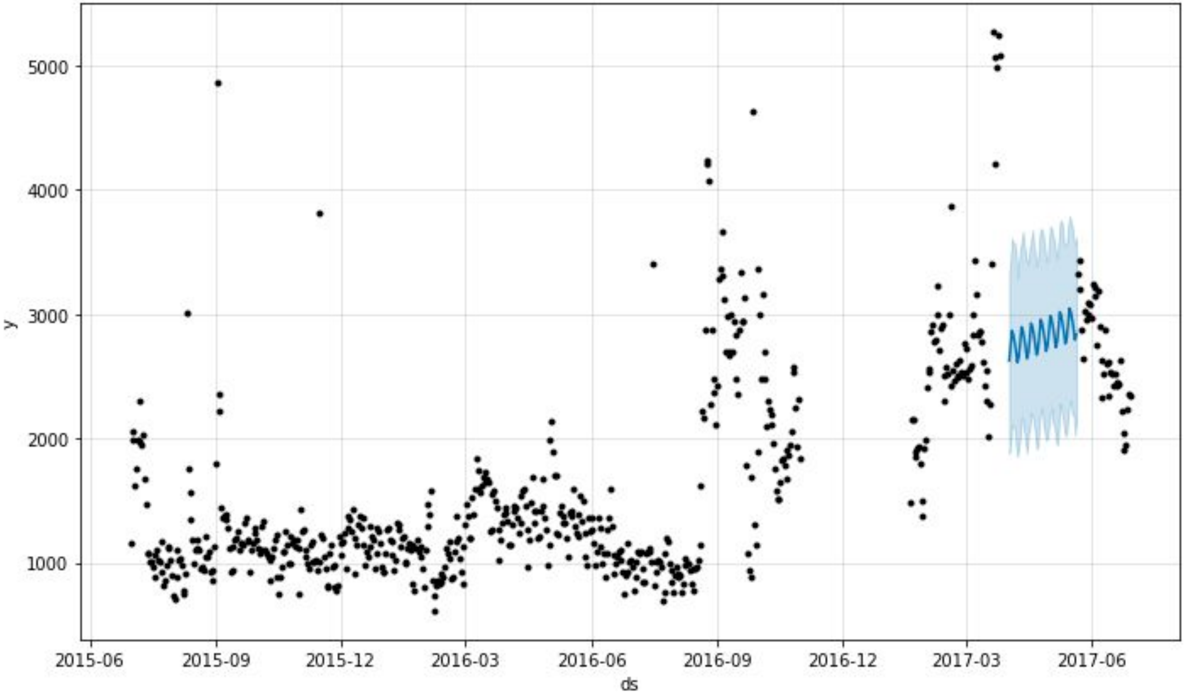



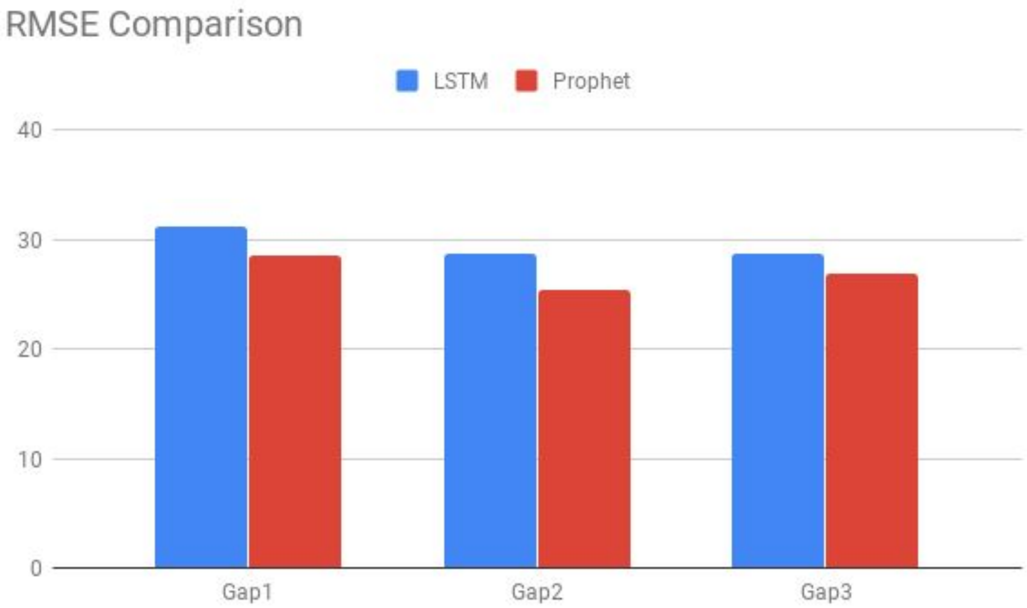
# Experiments and results: Web Traffic Prophet



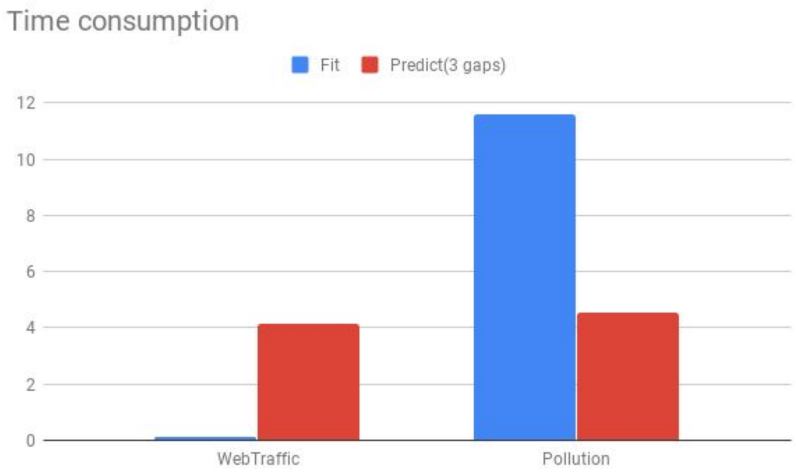



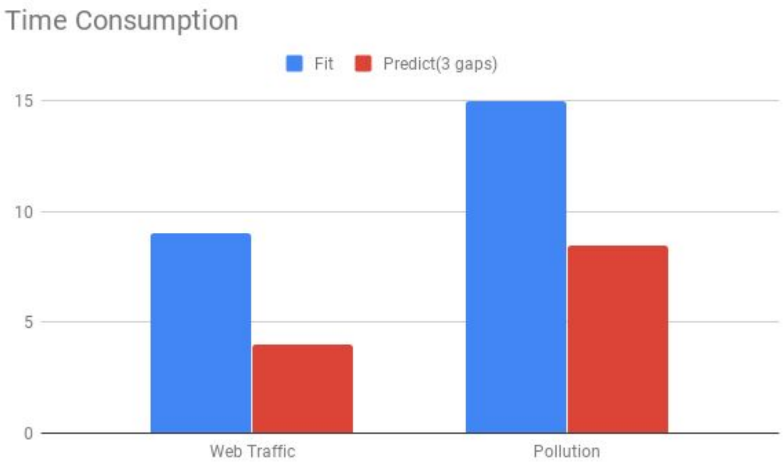



# Experiments and results: RMSE Comparison



# Record (approximate) runtimes of the forecasting methods




* LEFT: Prophet Runtime. RIGHT: LSTM Runtime.

** LSTM is dependent on how many epochs you train your neural network for

# Conclusion

- Implementation ease: LSTM vs Prophet
  - Prophet handles boilerplate stuff for you
- Gaps
  - Creating artificial gaps not always reflects real world scenario
  - Different levels of accuracies depending on horizon size
- Datasets
  - Air pollution dataset easier to predict (does not change with spikes)
  - Web traffic harder to predict (consider trends)

# Our project repository

https://github.com/gentrexha/energy_efficient_ds

# Q&A

Feel free to ask us anything!