COMP5703
CAPSTONE PROJECT

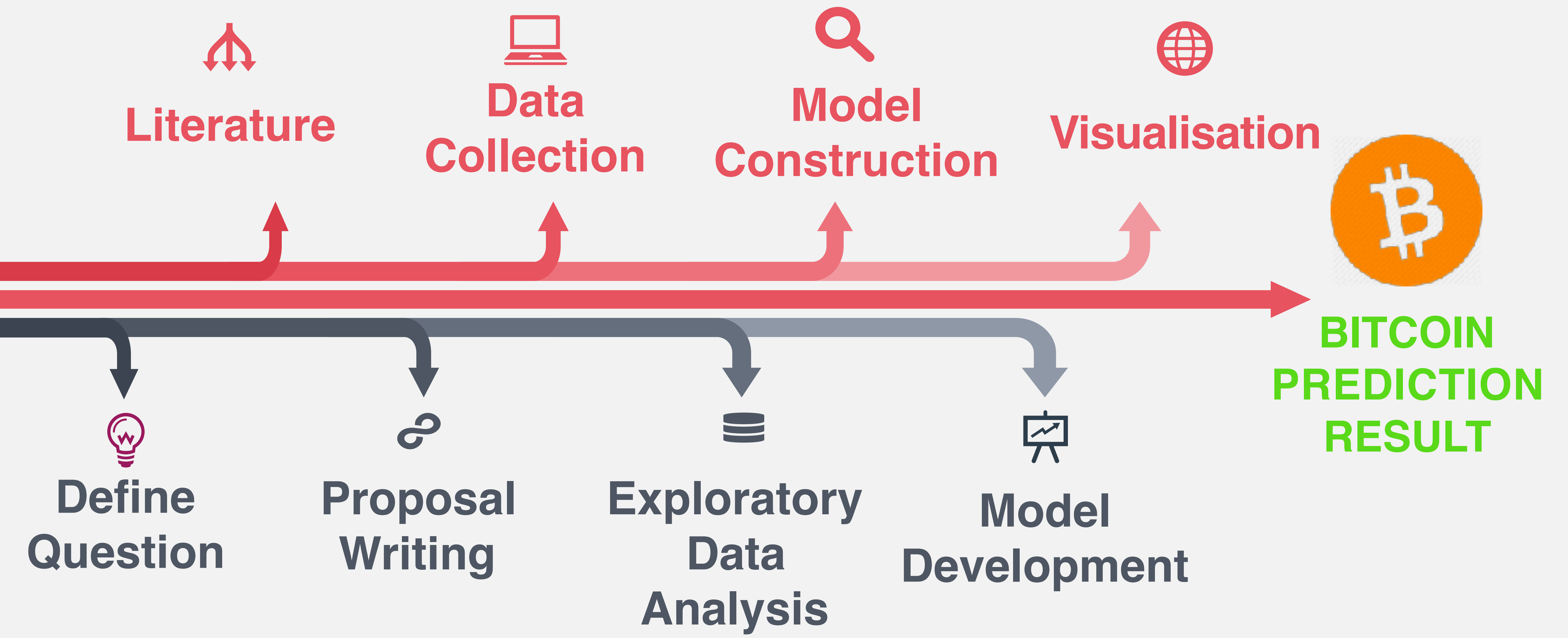
Introduction

Tutored by: Hamid Samani

Our project: “Visualisation and Prediction for Bitcoin-Exchanges”

Based on the data that we retrieve, and analyse from Bitcoin-exchange(s), we will conduct time-series analysis and machine learning techniques with Python scripting language to forecast, predict and visual the price movements

The Process of Our Project



Define Questions

Can we forecast the bitcoin price / trend in the short period / long period?

Can we predict the bitcoin price in the short period / long period?

What possible factors/features affect the bitcoin price?

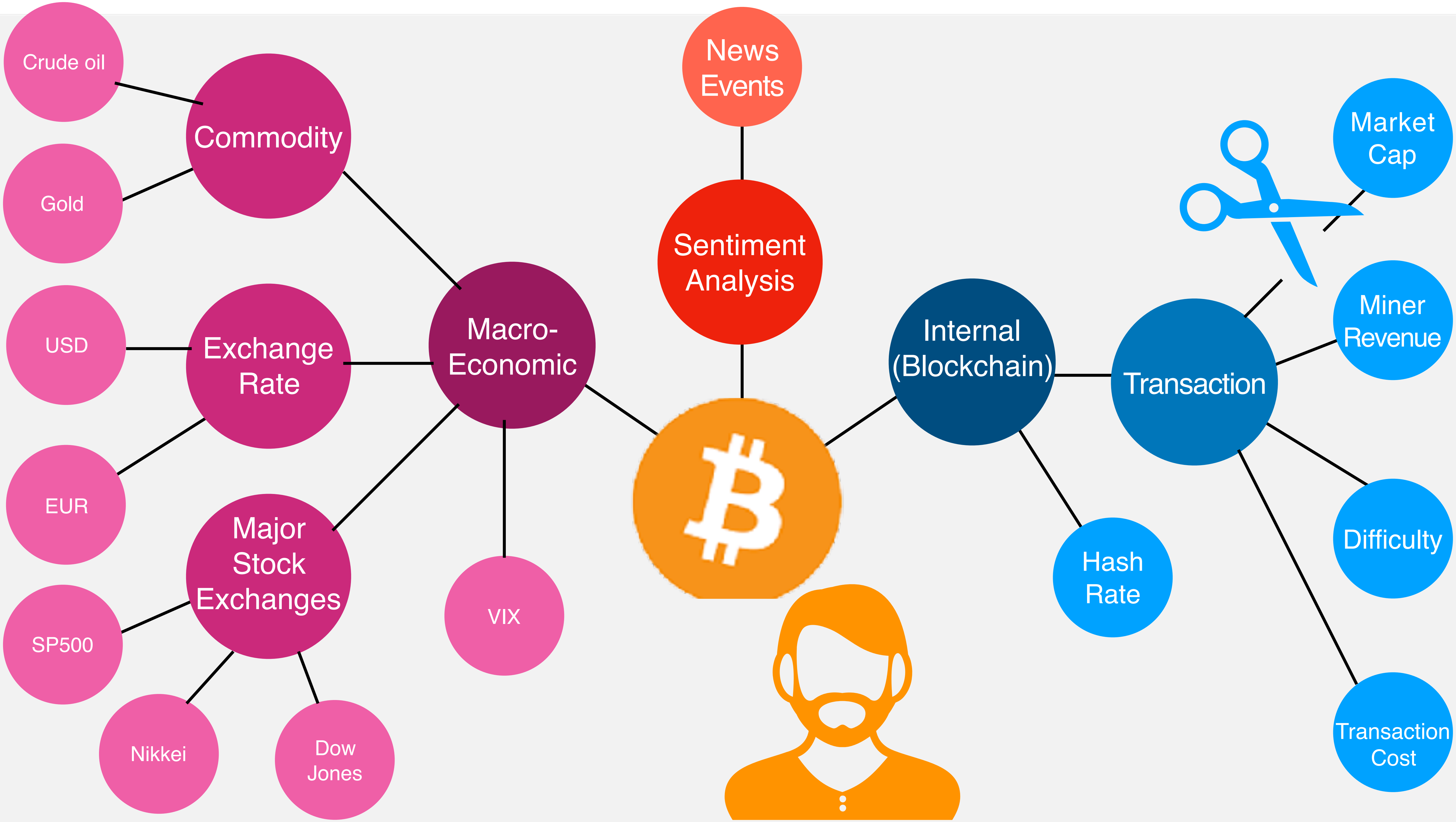
Background / Literature Review

Kristoufek (2014) pointed out bitcoin price is possible affect by economic, transaction and technical respectively

Greaves & Au (2015) applied various machine learning algorithms on prediction, including regression-based and classification-based multilayers neural network

McNally (2016) investigated further on prediction by using advanced neural network such as recurrent neural network and long short-term memory (LSTM) network and acquires more accurate prediction

Explore the potential features



Data Preparation

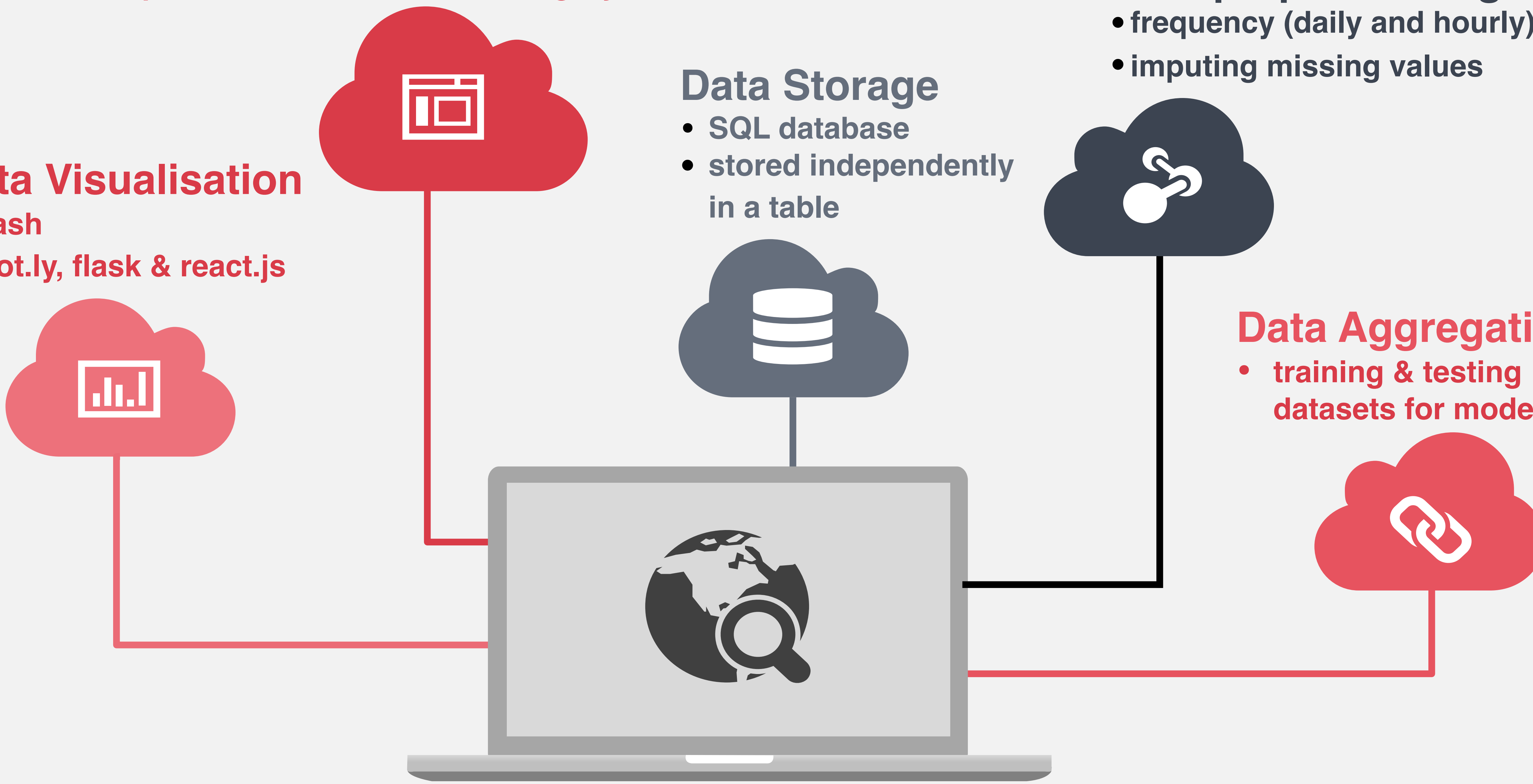
Data Scraping
- https://fred.stlouisfed.org
- https://www.coindesk.com/category/news/

Data preprocessing
- frequency (daily and hourly)
- imputing missing values

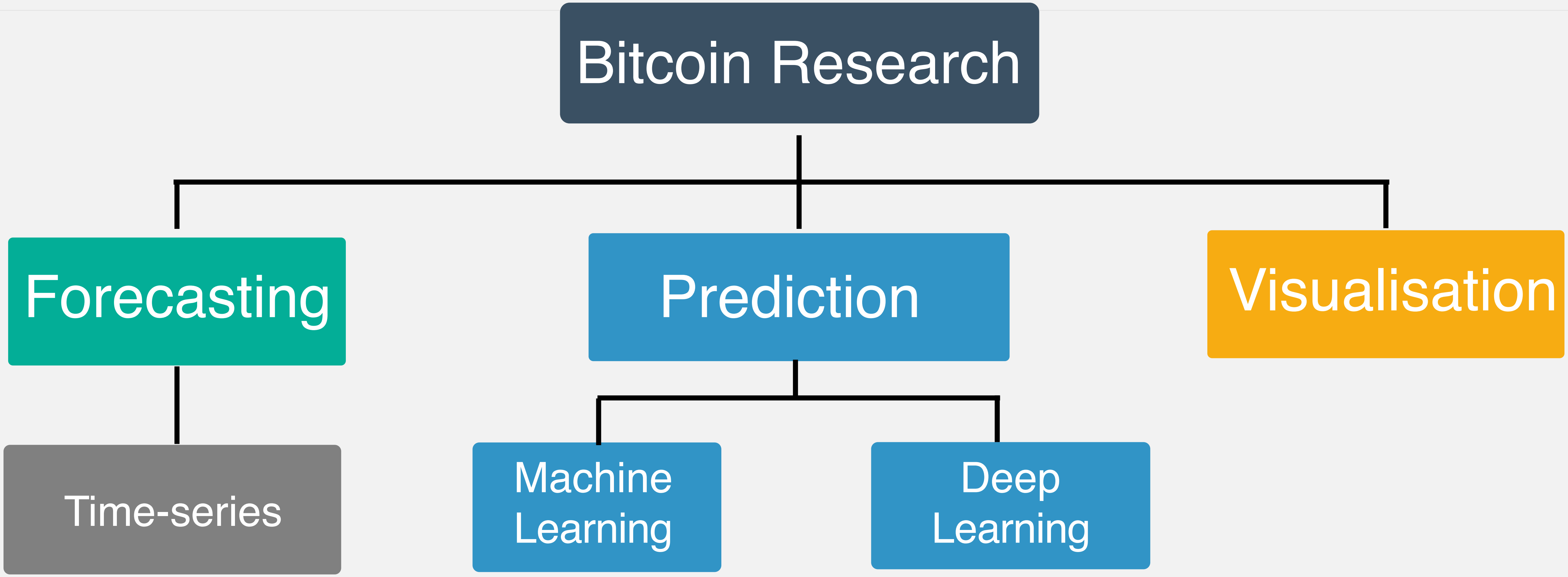
Data Storage
- SQL database
- stored independently in a table

Data Visualisation
- Dash
- plot.ly, flask & react.js

Data Aggregation
- training & testing datasets for modelling



Outline


Bitcoin Research
Forecasting
Prediction
Visualisation
Time-series
Machine Learning
Deep Learning

FORECASTING

TIME-SERIES ANALYSIS

Forecasting: Steps

STEP 1
Visualise the time-series data

STEP 2
Stationarize time series

STEP 3
Plot ACF/PACF & grid search to find Optimal Parameters

STEP 4
Build ARIMA/ SARIMA Model

STEP 5
Forecasting

Forecasting: Explore the time-series data

Daily price movement from
2017-01-01 to 2018-05-16

Hourly price movement from
2018-01-01 00:00 to 2018-05-15 23:00

Forecasting: Stationarized time series & FInd optimal parameters

Apply Dickey-Fuller Test

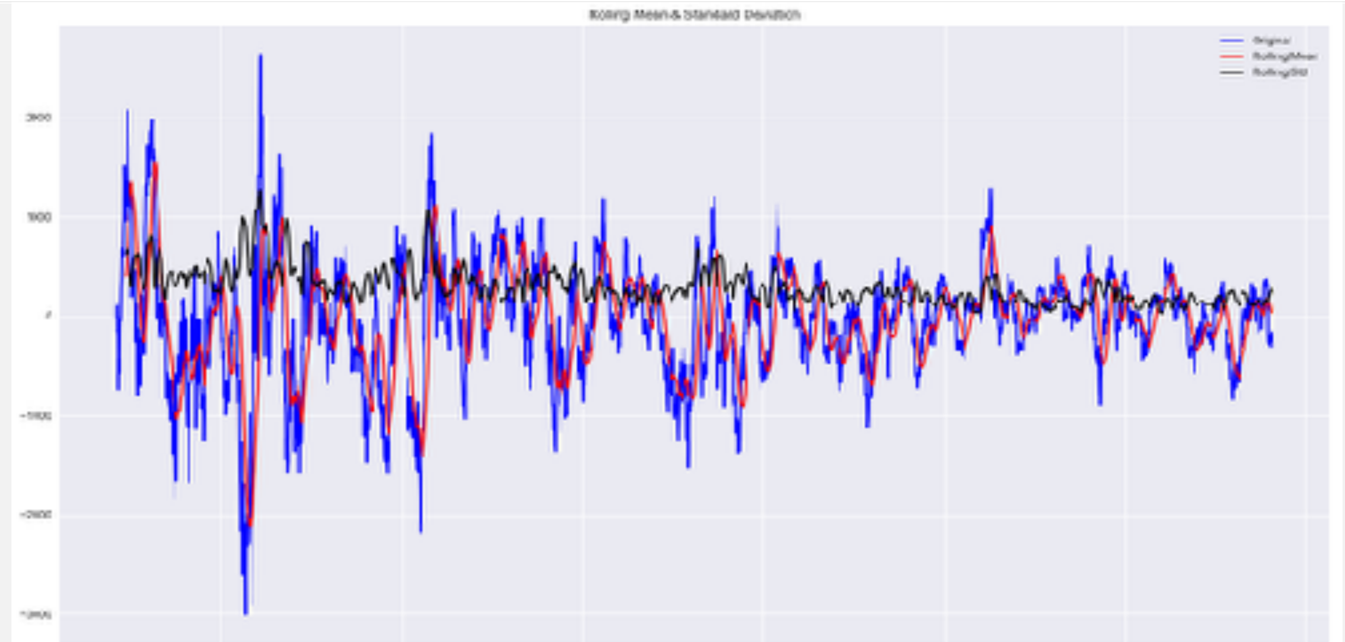
- Original data: non-stationary

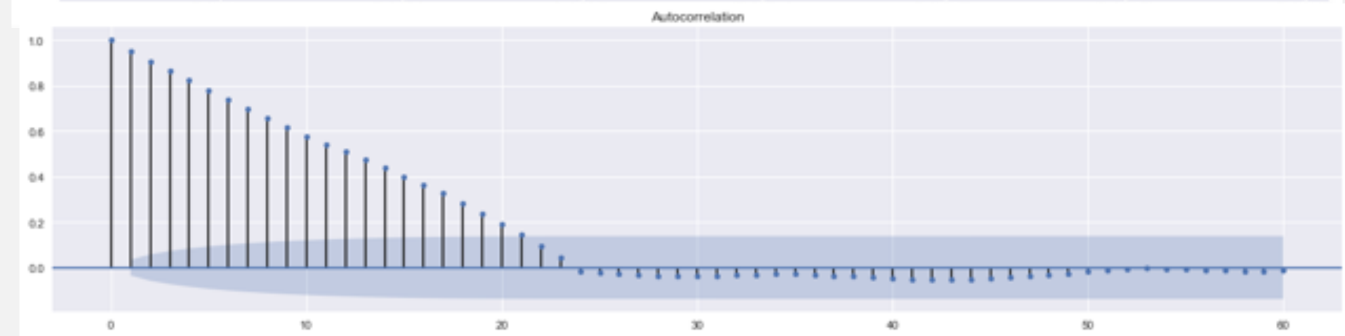
Data transformation

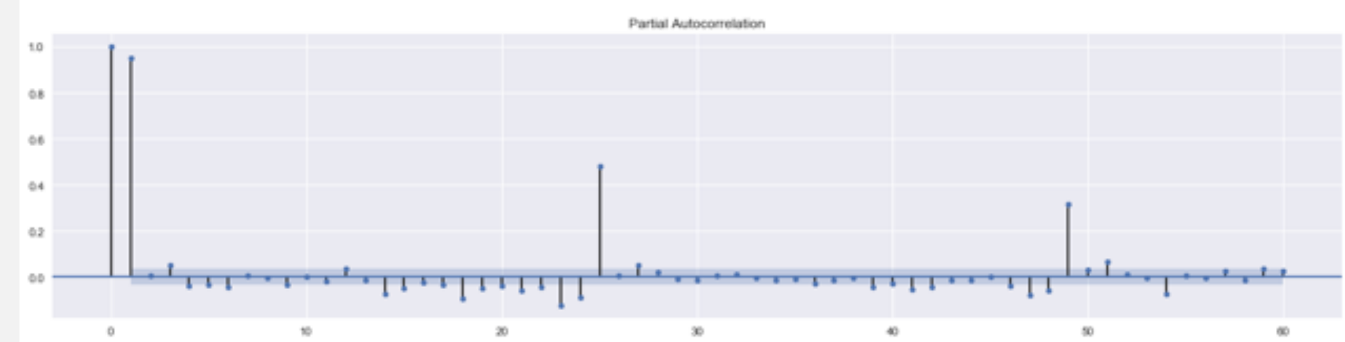
- Take Seasonal differential of original data
- Stationary;Pass Significantly P-value

Find optimal parameters

1. ACF/PACF of 24 hours seasonal differential
2. Grid Search



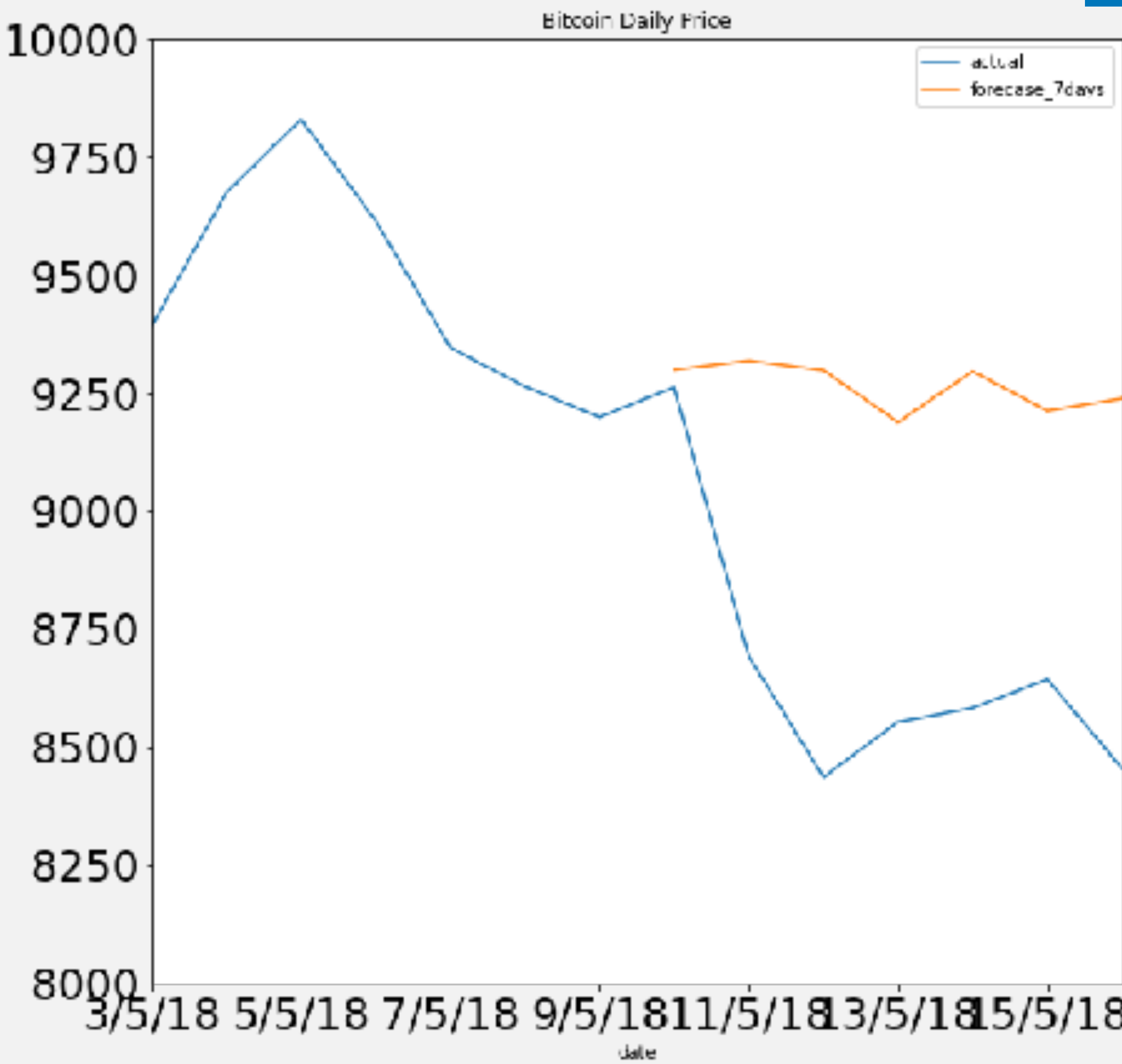


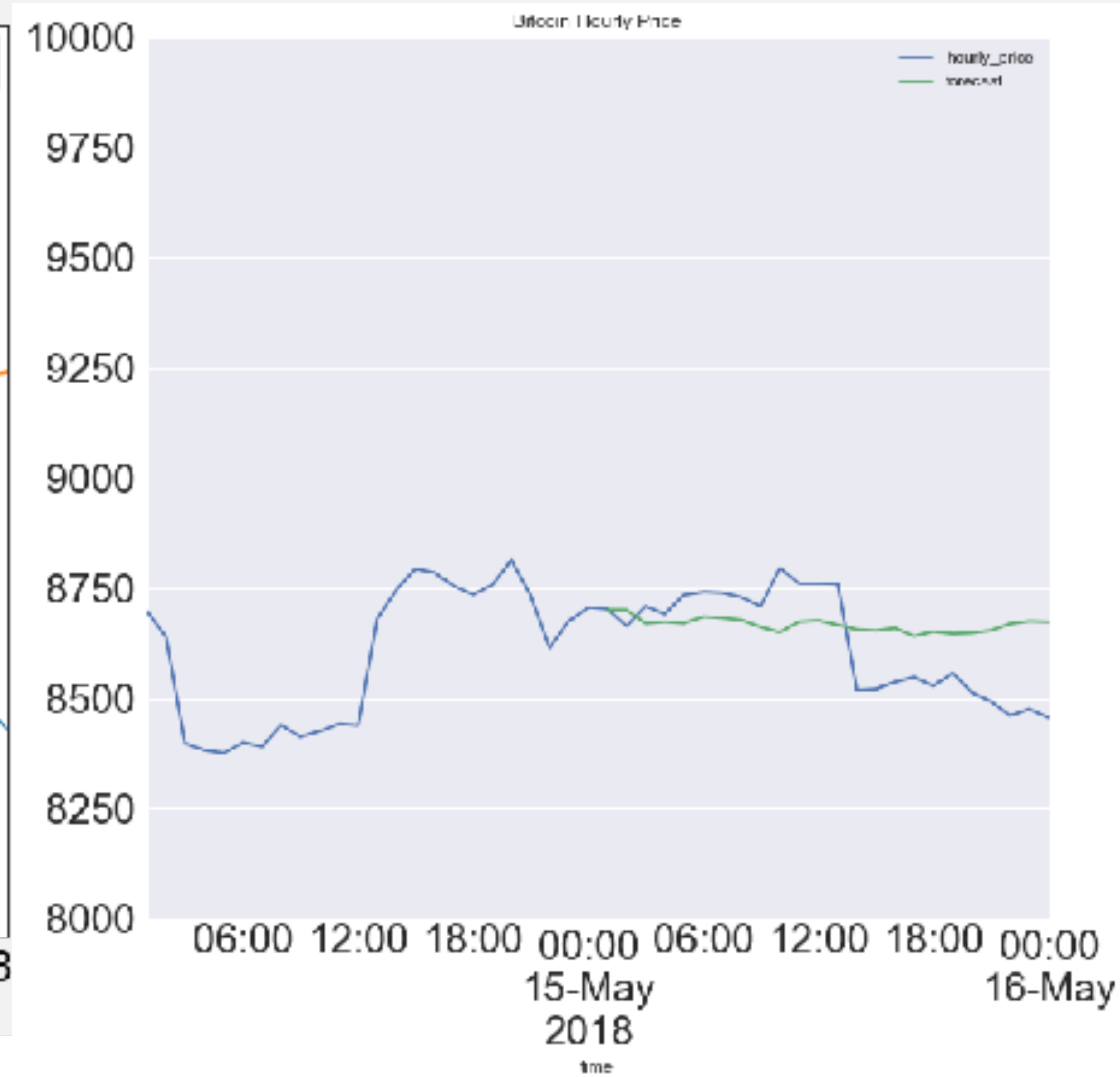


Parameters	AIC	Select or not
SARIMA(1, 1, 1) x (0, 1, 1, 24)	40079.2903287091	
SARIMA(1, 1, 1) x (1, 0, 0, 24)	40308.7366242086	
SARIMA(1, 1, 1) x (1, 0, 1, 24)	40291.5978456418	
SARIMA(1, 1, 1) x (1, 1, 0, 24)	41322.9050864317	
SARIMA(1, 1, 1) x (1, 1, 1, 24)	**40080.8210003745**	☑

Forecasting: Evaluation

Model	RMSE	MAE
ARIMA (daily price)	652.666	603.285
SARIMA (hourly price)	115.786	100.290





PREDICTION

MACHINE LEARNING TECHNIQUES

Prediction: Steps

STEP 1
Split the training & validation dataset

STEP 2
EDA / Pre-processing Data

STEP 3
Construct the Models

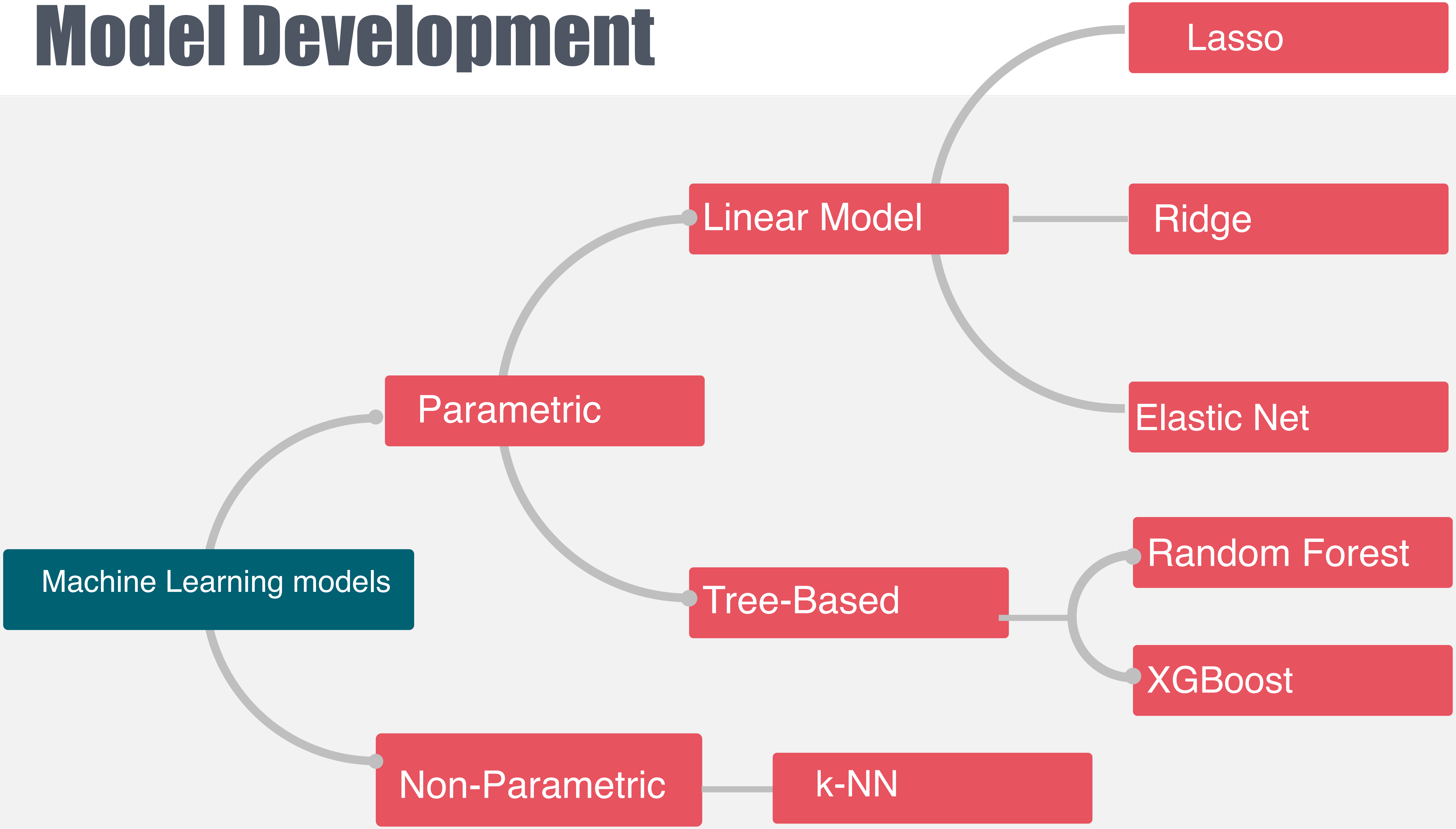
STEP 4
Parameter Optimization by CV

STEP 5
Use Test Dataset to Validate

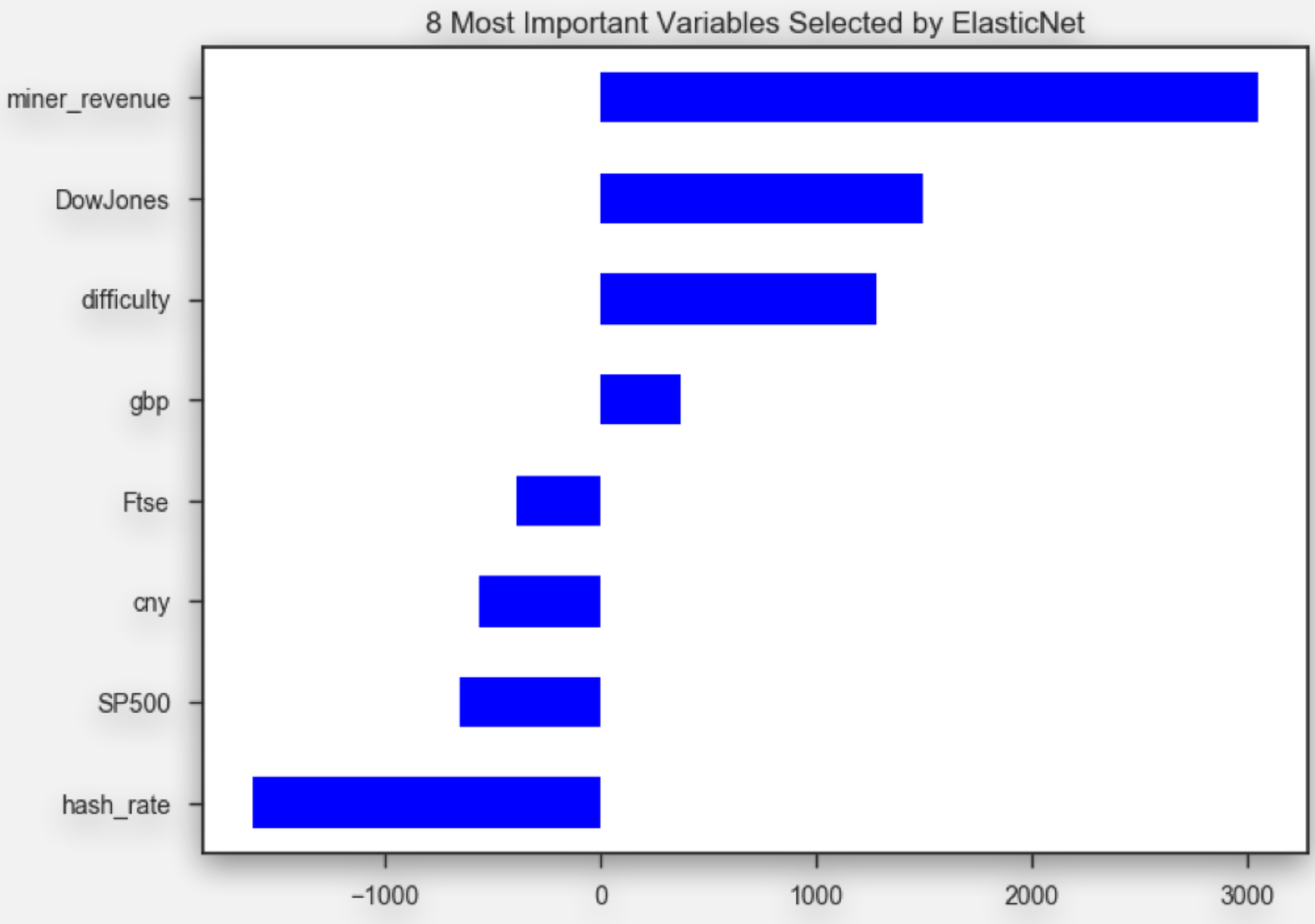
STEP 6
MODEL SELECTION

STEP 7
Predict on Test Data parts

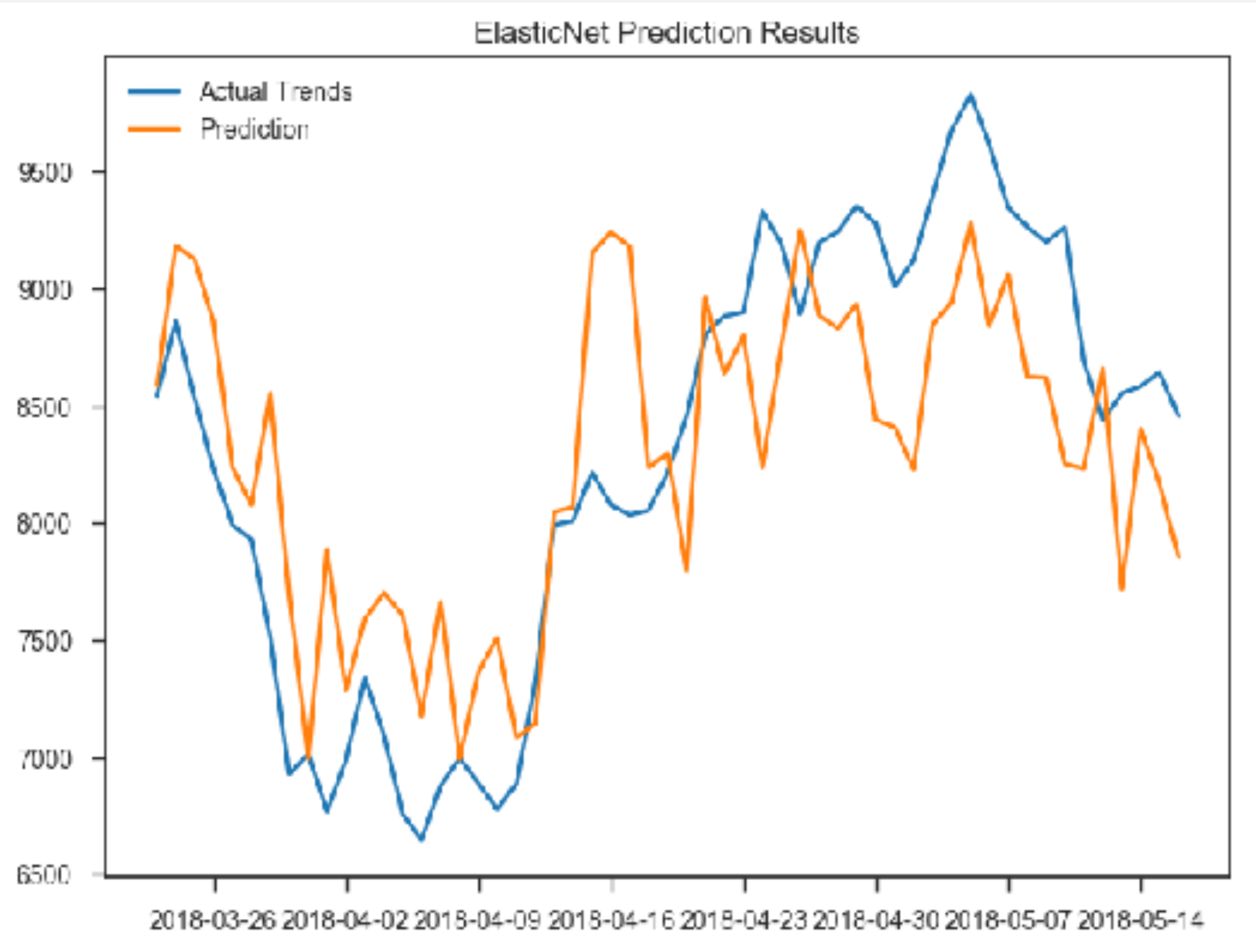
Model Development



Prediction: Features Selected by Elastic Net



Prediction: Evaluating results



Model	Training Error (CV RMSE)	Testing Error (RMSE)
Ridge	583.35	616.47
Lasso	589.40	602.86
ElasticNet	587.46	601.59
Random Forest	1065.75	1022.56
XGBoost	667.1	1042.74
K-NN	960.69	1350.47

PREDICTION

DEEP LEARNING TECHNIQUES

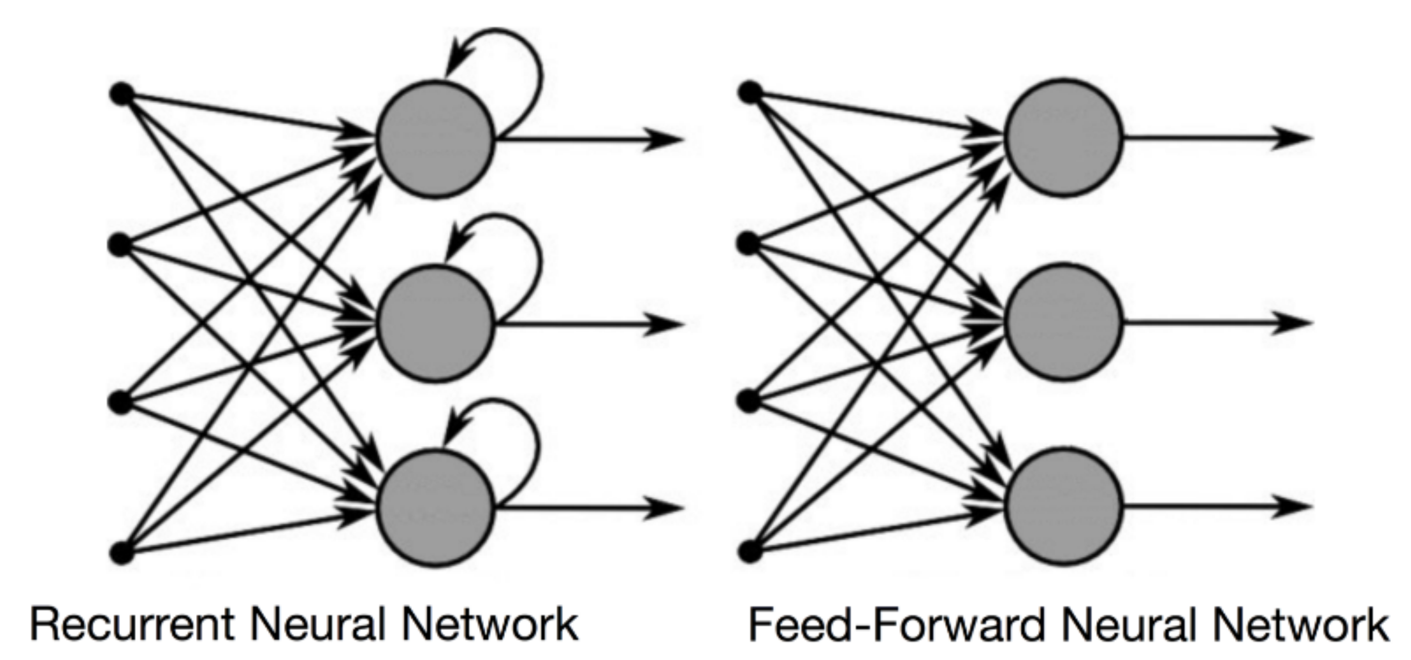
Deep Learning: RNN Overview

DEFINITION

- One special neural network that remembers its input due to an internal memory.

HOW IT WORKS

- In a RNN model, the information cycles through a loop.
- When it makes a decision, it takes into consideration the current inputs and also what it has learned from the inputs it received previously.



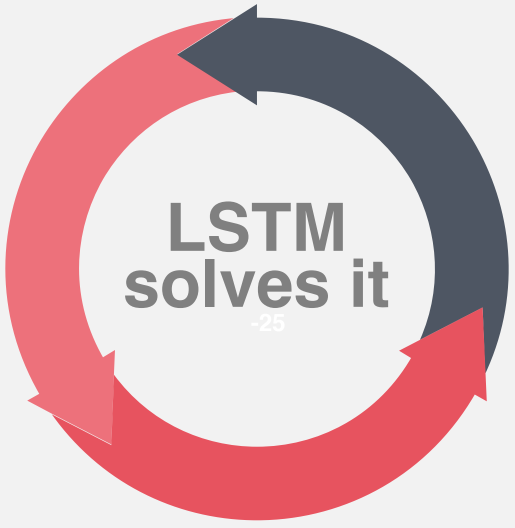
Deep Learning: RNN Overview

PROS

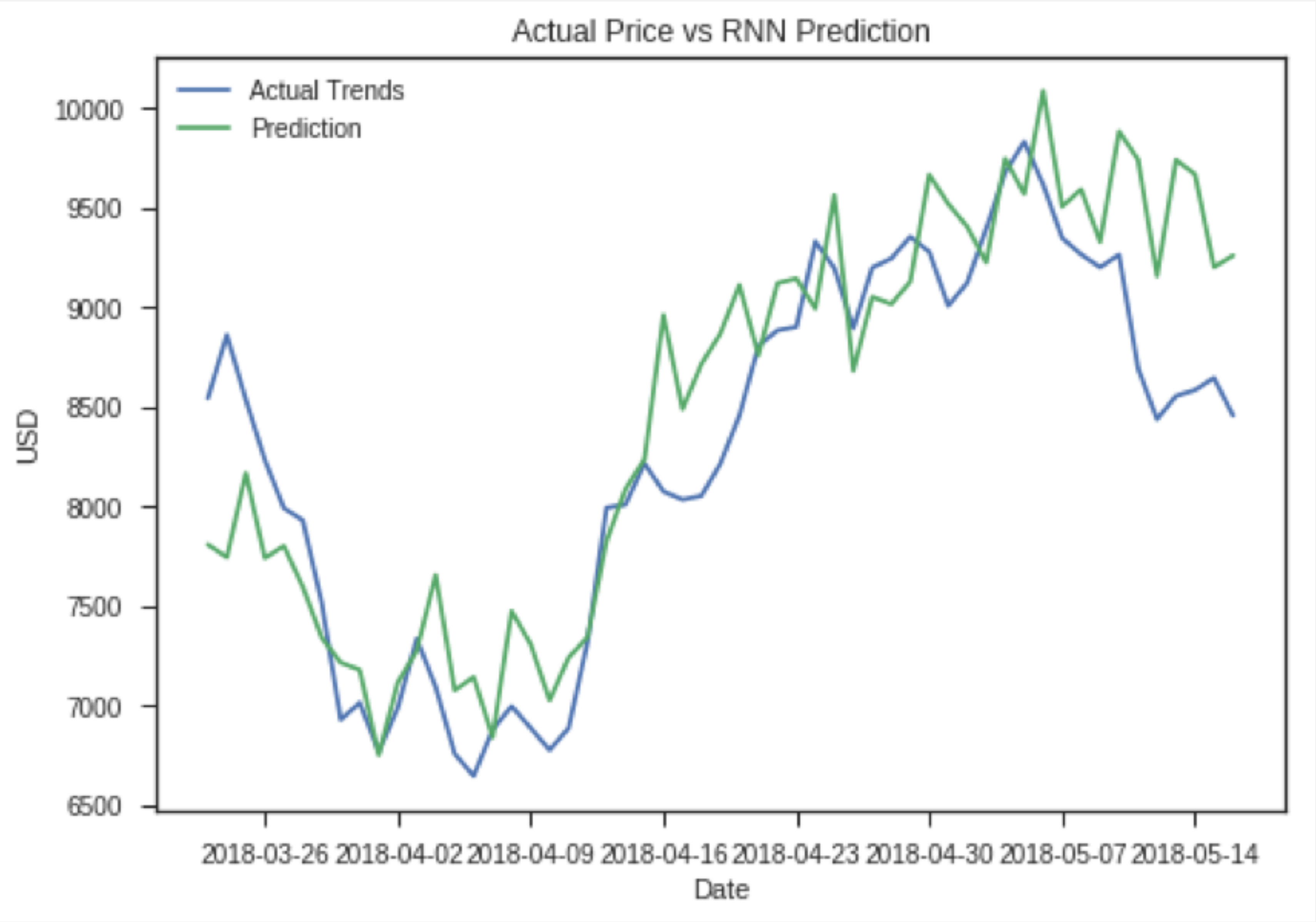
- Suited for ML problems that involve sequential data that temporal dynamics that connects the data is more important than the spatial content of each individual frame
- AN usual RNN has a short-term memory.
- In combination with a LSTM , they also have a long-term memory.

CONS

- Vanishing Gradients: model stops learning
- Long-Term Dependencies: Long memory required



Prediction: RNN Result

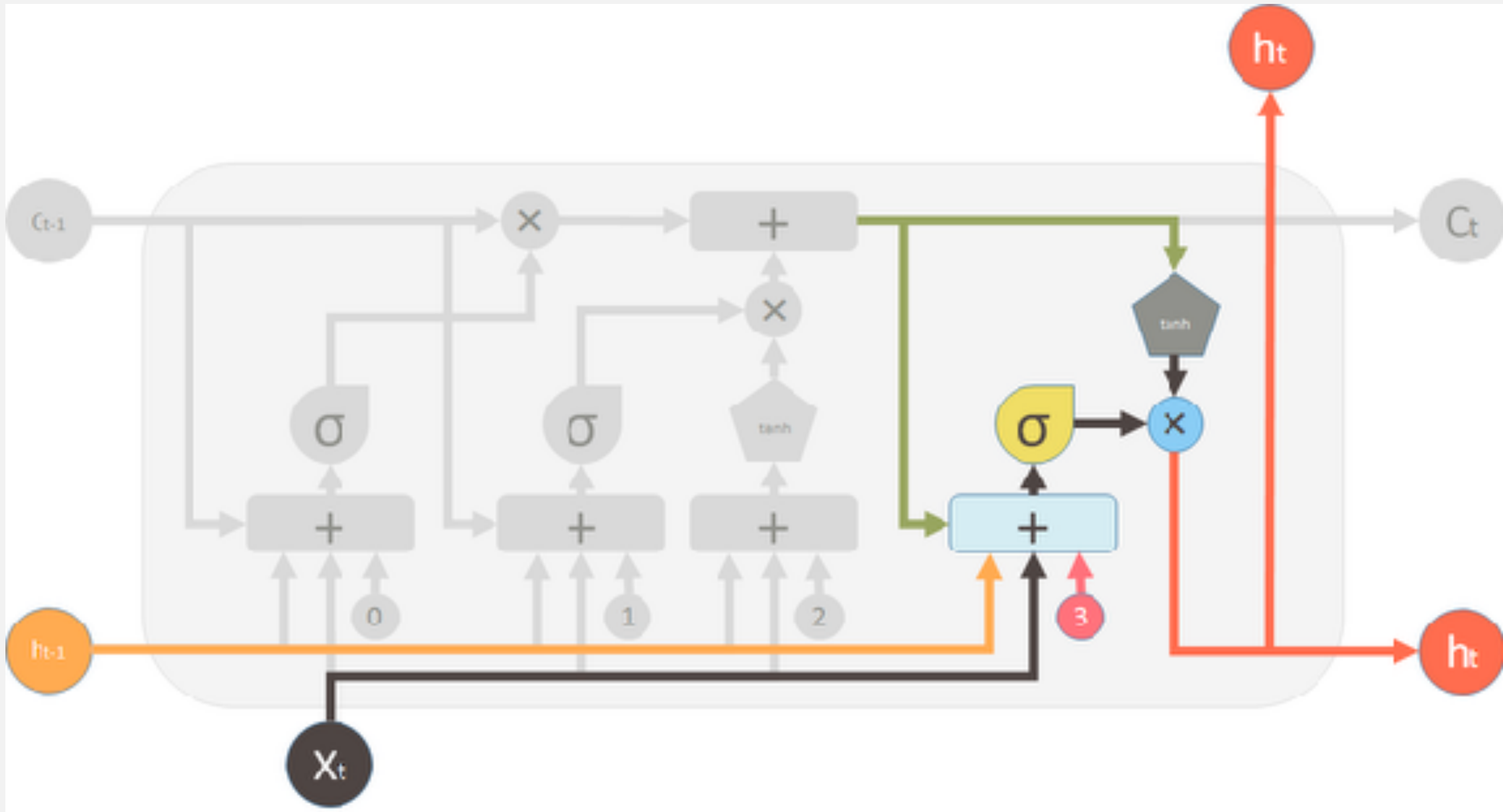


Model	Testing Error (RMSE)
Elastic	601.59
RNNs	487.260

Final model: RNNs

Why RNN Works Better ?

- For Time Series problem, the recurrence operation allows response to depend on a time-evolving state.
- It is able to store information from arbitrarily long time ago for further forecasting



Model Evaluation

	Time-series	Machine Learning	Deep Learning
RMSE	Low	Medium	Low
Model Complexity	Low	Medium - Hight	High
Computational Cost (Training Time)	Low	Medium - Hight	Medium - Hight
Forecasting/ Prediction	Forecasting	Prediction	**Both**
Input Variables	Price & Time	Price & Predictor Variables	Price & Predictor Variables
Long-run Prediction & Guarantee Performance	✗	✓	✓

Visualisation

Dash

Use Dash as a visualisation framework

Built on top of plot.ly, flask & react.js

User Interfaces

Allow us to build visualisation apps

Customisable user interfaces using python.

Hide & Focus

Hide abstraction & technologies in the background

Focus on building and customising our visualisations.

Layout

Built-in support for layout tools

Facilitating in communicating results & analysis to the audience.

Web Browser

Rendered in the web browser

Efficient for testing layouts before the final visualisation framework

Conclusion

Bicoin Price
is hard to predict
by quantitative
methods

The limitaiton of
linear models

The cost of
collecting some
data is high

The result of time-
series methods
heavily rely on the
actual trend of the
price movement

Future works

QUALITATIVE METHODS

NON-LINEAR METHODS

FEATURE ENGINEERING

SENTIMENT ANALYSIS