University of Ljubljana
Faculty of Computer and Information Science

What Language Do Stocks Speak?

Marko Poženel and **Dejan Lavbič**

2nd July 2018

Baltic DB&IS 2018
13th International Baltic Conference on Databases and Information Systems

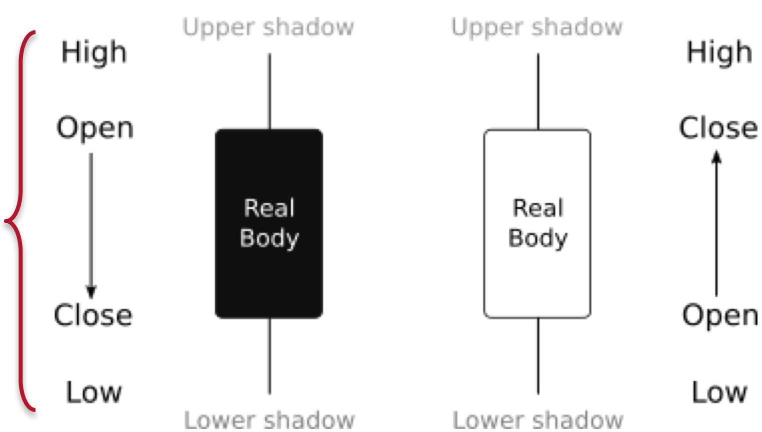
Agenda

- **Motivation** for our work,
- Proposed **Forecasting Model**,
  - Input OHLC Data, OHLC Data Normalization,
  - Japanese Candlestick Pattern Identification,
  - Sentence Construction,
  - Learning Context with Word2Vec,
  - Prediction,
- **Evaluation**,
  - Russell Top 50 Index,
  - Buy & Hold, Moving Average, MACD,
- **Conclusion** and **Future Work**.

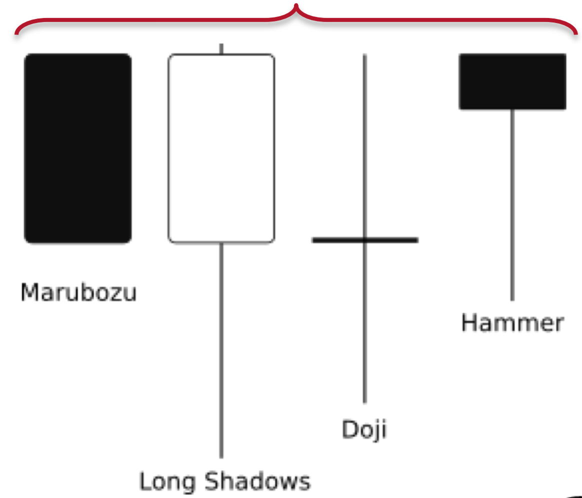
Motivation (1)

- Many approaches to forecasting future stock values:
  - **technical analysis** (historical price changes),
    - **Japanese Candlestick Trading Strategy** from 18th century used for trading rice,

OHLC values for given trading unit (e.g. day)



Popular Japanese Candlestick **patterns**



Motivation (2)

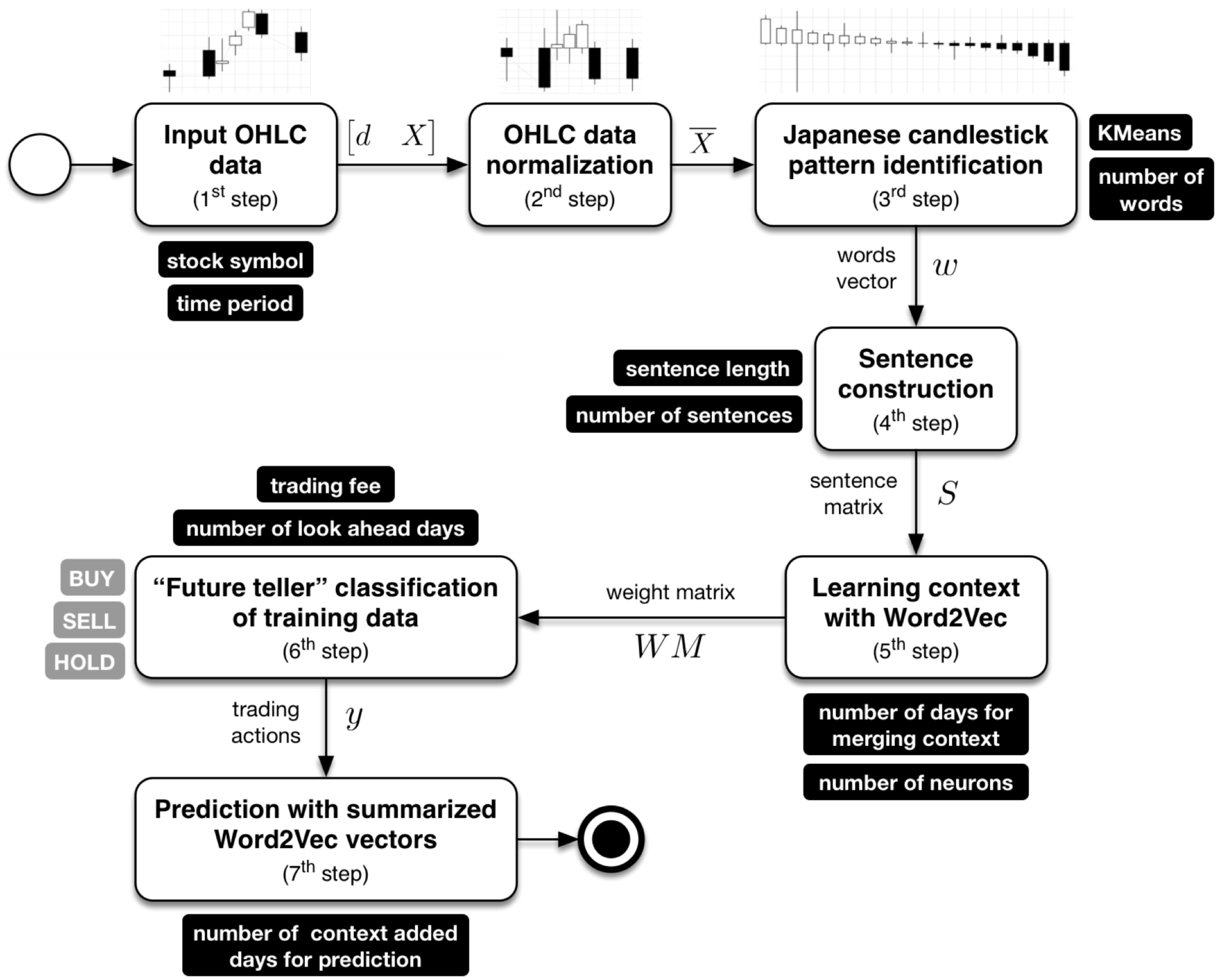
- Many approaches to forecasting future stock values:
  - **fundamental analysis** (company's business, news etc.),
    - some approaches are based on NLP algorithms,
      - **Word2Vec** is a group of models (e.g. CBOW, skip-gram) that is a 2-layer neural network used to extract linguistic contexts of words,

vector("**Madrid**") + vector("**Spain**") – vector("**Paris**") → vector("**France**")

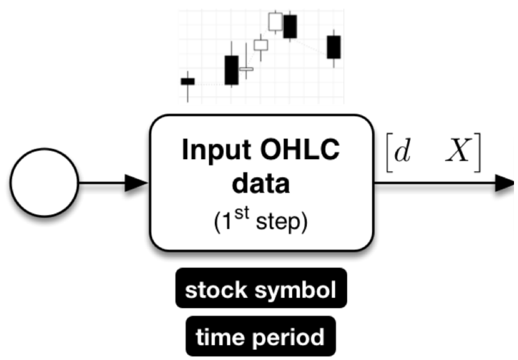
Proposed Forecasting Model (1)

- combine **Japanese Candlestick Trading Strategy** and **Word2Vec** approach:
  - create a **simplified OHLC language**, used for input to Word2Vec,
  - **learn** rules and **patterns with Word2Vec** and use this knowledge to predict future trends in stock value.

Proposed Forecasting Model (2)



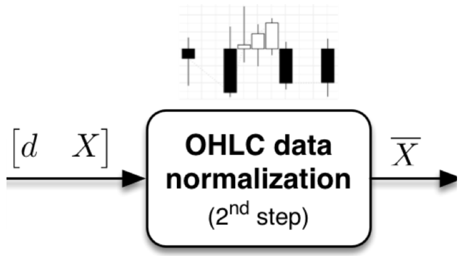
Input OHLC data (1st step)



- we observe **input data** on a trading day basis for ***n_d trading days*** and Japanese Candlesticks are represented as ***OHLC tuples***,

$$\left[d_{(1\ \times\ n_d)}\ X_{(4\ \times\ n_d)}\right] = \left[\begin{array}{c|cccc} d_1 & O_1 & H_1 & L_1 & C_1 \\ d_2 & O_2 & H_2 & L_2 & C_2 \\ \dots & \dots & \dots & \dots & \dots \\ d_{n_d} & O_{n_d} & H_{n_d} & L_{n_d} & C_{n_d} \end{array}\right]$$

OHLC data normalization (2nd step) (1)



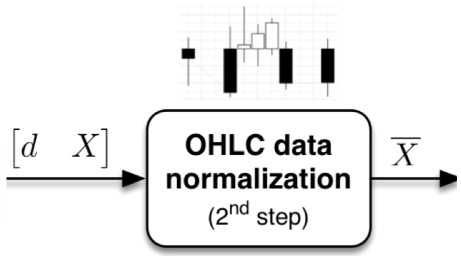
- we're interested in the **shape of Candlestick** and not the absolute value, so we ***normalize with Open (O) value***

$$norm(\langle O, H, L, C \rangle) = \langle 1, \frac{H}{O}, \frac{L}{O}, \frac{C}{O} \rangle : X \rightarrow \overline{X}$$

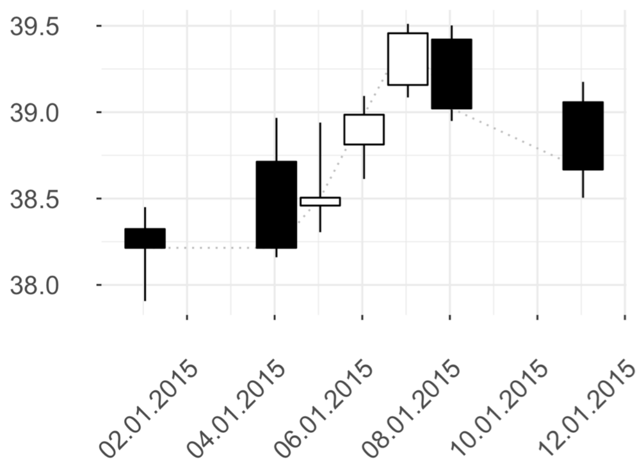
- which results in

$$\overline{X}_{(4 \times n_d)} = \begin{bmatrix} 1 & \frac{H_1}{O_1} & \frac{L_1}{O_1} & \frac{C_1}{O_1} \\ 1 & \frac{H_2}{O_2} & \frac{L_2}{O_2} & \frac{C_2}{O_2} \\ \dots & \dots & \dots & \dots \\ 1 & \frac{H_{n_d}}{O_{n_d}} & \frac{L_{n_d}}{O_{n_d}} & \frac{C_{n_d}}{O_{n_d}} \end{bmatrix}$$

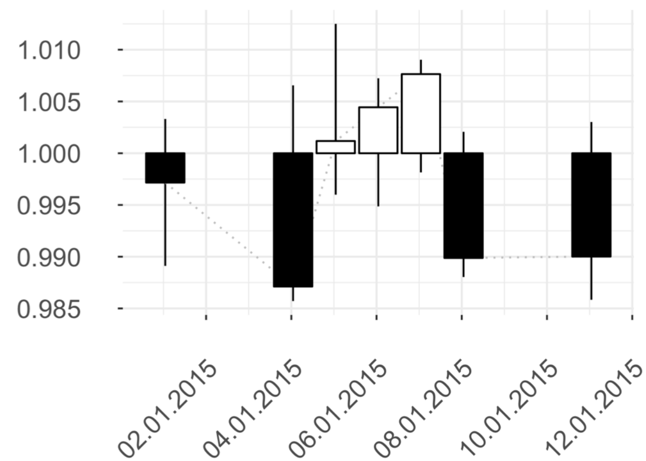
OHLC data normalization (2nd step) (2)



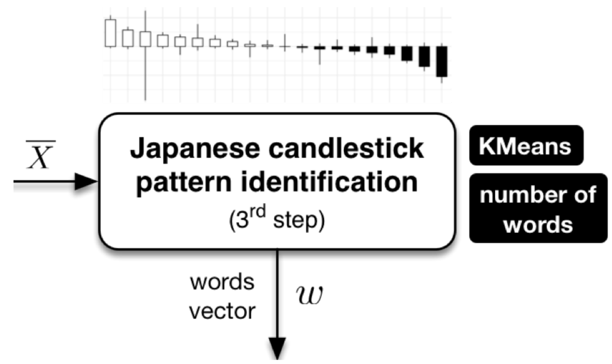
- we're interested in the **shape of Candlestick** and not the absolute value, so we ***normalize with Open (O) value***



→



Pattern identification (3rd step) (1)

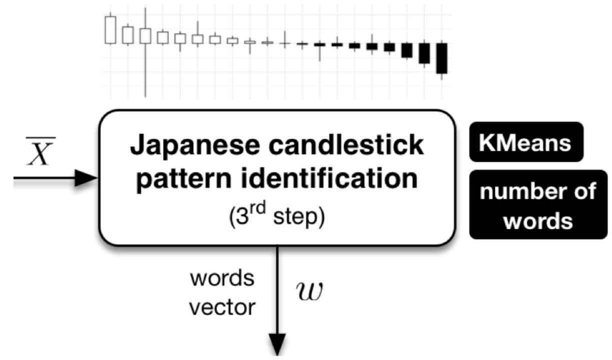


- automatically detect candlestick clusters by employing KMeans,
- limit the ***number of*** possible ***words*** of stocks' language ***n_w*** (OHLC shapes),

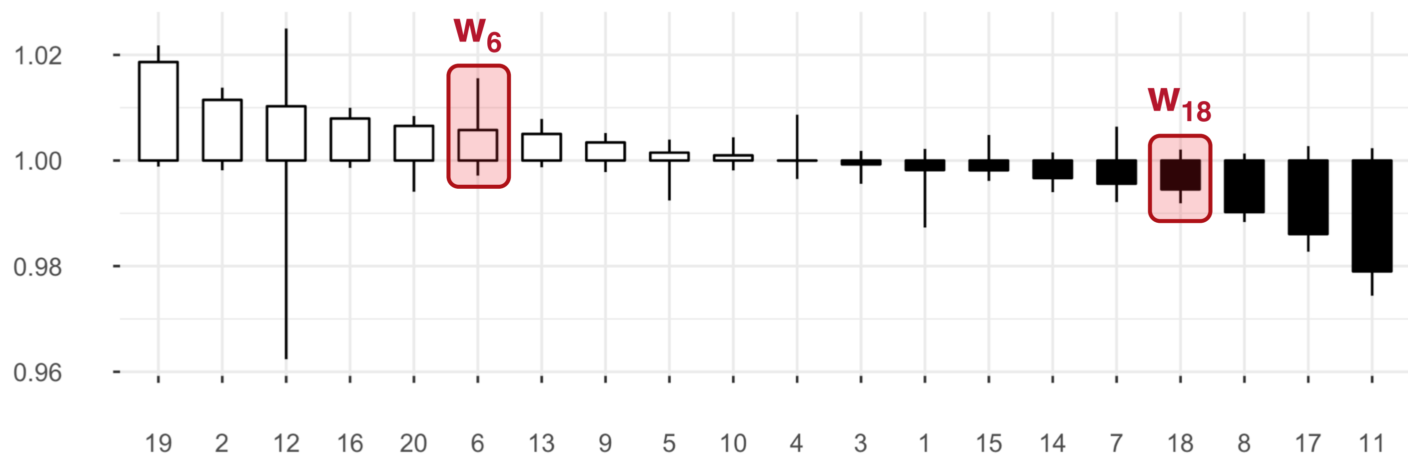
$$KMeans(n_w) : \overline{X} \rightarrow w \qquad\qquad w_{(1 \times n_d)} = \left[w_1 \; w_2 \; \ldots \; w_{n_d} \right]^T$$

  – the value of n_w is based on the Silhouette measure,
  – **word** is **an individual trading day** and is a representation of a specific Japanese candlestick

Pattern identification (3rd step) (2)

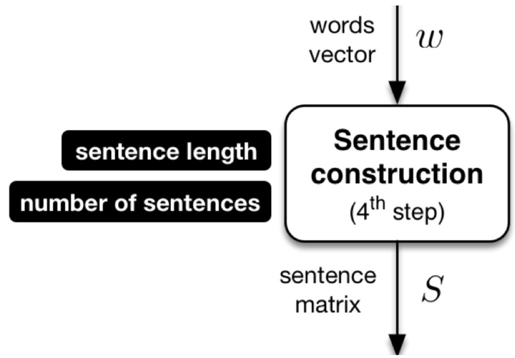


- example of $n_w = 20$ identified words for **Coca Cola (KO) stock**,



Sentence construction (4th step)

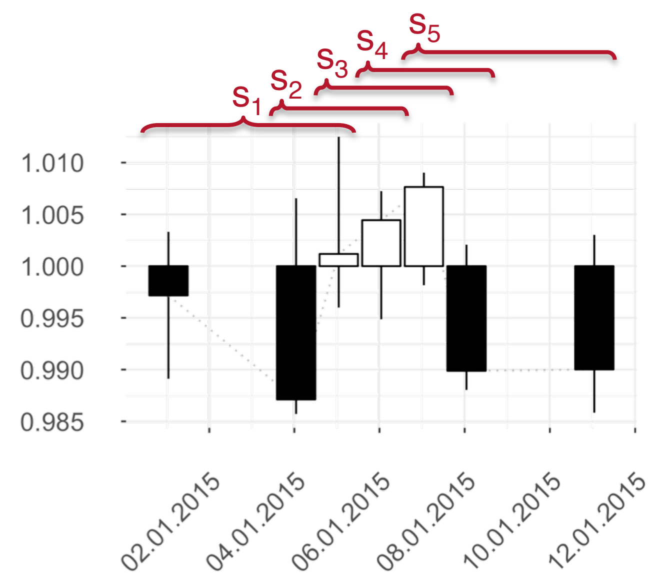
- we specify a ***sentence length l_s*** that defines a number of consecutive words (i.e. trading days) grouped into sentences,
- the ***number of sentences n_s*** is therefore $n_s = n_d - (l_s - 1)$



- the result is a ***sentence matrix S*** of rolling window of trading data

$$S_{(l_s \times n_s)} = \begin{bmatrix} w'_1 & w'_2 & \dots & w'_{l_s} \\ w'_2 & w'_3 & \dots & w'_{l_s+1} \\ \dots & \dots & \dots & \dots \\ w'_{n_s} & w'_{n_s+1} & \dots & w'_{n_d} \end{bmatrix}$$

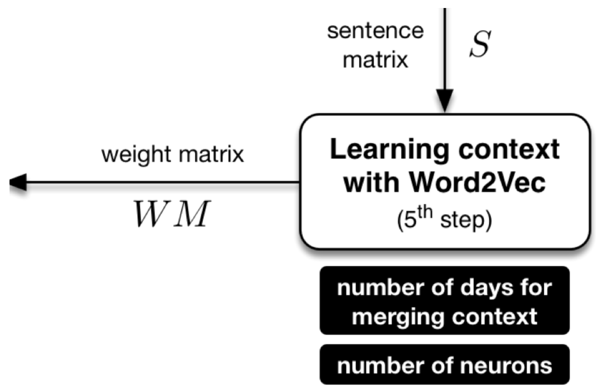
Simple example with ***$n_d = 7$***, ***$l_s = 3$*** and ***$n_s = 5$***.



Learning with Wor2Vec (5th step) (1)

- Word2Vec acquires vectors for words (i.e. trading days) that explicitly contain various rules and patterns,
- we define the ***number of days for merging context n_{ww}*** and the ***number of neurons n_v*** in hidden layer weight matrix,
- Word2Vec algorithm **performs** the following **transformation**

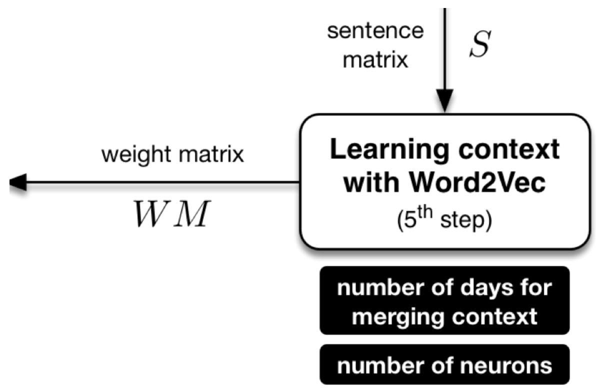
$$W2V\left(S, n_{ww}, n_v\right) : S \rightarrow WM$$



Learning with Wor2Vec (5th step) (2)

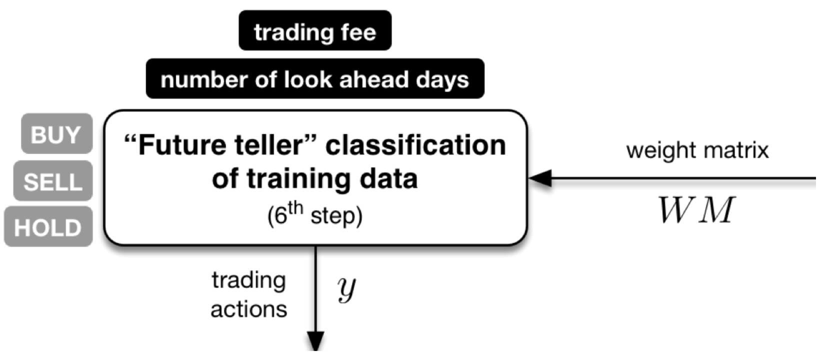
- the result of Word2Vec learning phase is a ***Weight Matrix WM*** with n_v columns (number of vectors) and n_w rows (number of words in stocks' language) and is defined as follows

$$WM_{(n_v \times n_w)} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n_v} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n_v} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n_w,1} & v_{n_w,2} & \cdots & v_{n_w,n_v} \end{bmatrix}$$



Classification of training data (6th step)

- our aim is that the predictive model would, based on input OHLC sequence, trigger one of the following actions: **BUY**, **SELL**, **HOLD** **(do nothing)**,
- we label trading days from matrix *X* in training set with **trading actions** $y_{(1\times n_d)} = \left[A_1\ A_2\ \ldots\ A_{n_d}\right]^T$



- take into account *number of look ahead days* n_{la}, *trading fee* v_{fee}, *initial equity* e, *maximum number of stocks to trade* n_{max} and *stock's close price* C.

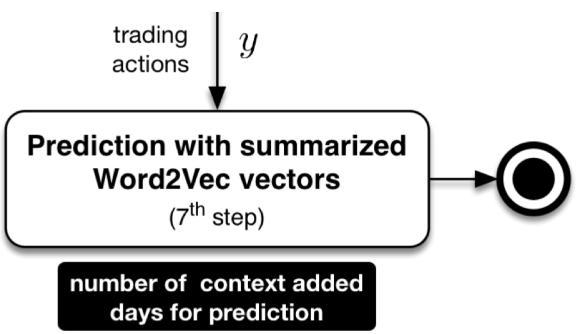
$$y_i = \begin{cases} 0 : \text{BUY} & n_{max}\cdot C_j > n_{max}\cdot C_i + 2\cdot v_{fee},\ j\in\left[i, i+n_{la}\right] \\ 1 : \text{SELL} & n_{max}\cdot C_j < n_{max}\cdot C_i - 2\cdot v_{fee},\ j\in\left[i, i+n_{la}\right] \\ 2 : \text{HOLD} & \text{otherwise} \end{cases}$$

Prediction (7th step) (1)

- **Basic prediction**
  - **normalized OHLC data** and vector of **trading actions** from "Future Teller" classification, where SoftMax classifier defines the following transformation

$$[\overline{X}_{(3\ \times\ \overline{n_d})}\ y_{(1\ \times\ \overline{n_d})}] = \left[\begin{array}{ccc|c} \frac{H_1}{O_1} & \frac{L_1}{O_1} & \frac{C_1}{O_1} & A_1 \\ \frac{H_2}{O_2} & \frac{L_2}{O_2} & \frac{C_2}{O_2} & A_2 \\ \ldots & \ldots & \ldots & \ldots \\ \frac{H_{\overline{n_d}}}{O_{\overline{n_d}}} & \frac{L_{\overline{n_d}}}{O_{\overline{n_d}}} & \frac{C_{\overline{n_d}}}{O_{\overline{n_d}}} & A_{\overline{n_d}} \end{array}\right] \rightarrow y = f\left(\frac{H}{O}, \frac{L}{O}, \frac{C}{O}\right)$$

  - does not perform well as it **does not include the context** in which OHLC candlesticks appear and influence price



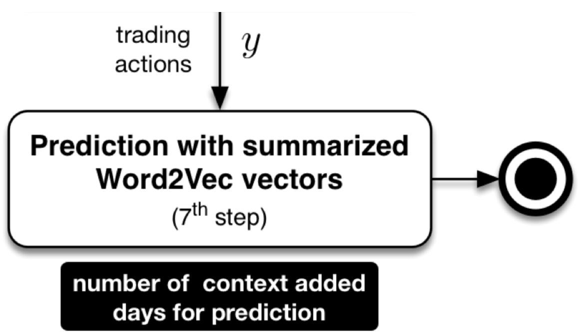
University of Ljubljana
Faculty of Computer and Information Science

Prediction (7th step) (2)

- **Prediction with summarized Word2Vec vectors**
  - from vector of words and vector of trading actions
  - we replace words with a Word2Vec features vector (hyper parameter) from Weight Matrix

$$\left[w_{(1 \times \overline{n_d})}\ y_{(1 \times \overline{n_d})}\right] = \left[\begin{array}{c|c} w_1 & A_1 \\ w_2 & A_2 \\ \dots & \dots \\ w_{\overline{n_d}} & A_{\overline{n_d}} \end{array}\right]$$

$$\left[X'_{(n_v \times \overline{n_d})}\ y_{(1 \times \overline{n_d})}\right] = \left[\begin{array}{cccc|c} v_{1,1} & v_{1,2} & \dots & v_{1,n_v} & A_1 \\ v_{2,1} & v_{2,2} & \dots & v_{2,n_v} & A_2 \\ \dots & \dots & \dots & \dots & \dots \\ v_{w_{\overline{n_d}},1} & v_{w_{\overline{n_d}},2} & \dots & v_{w_{\overline{n_d}},n_v} & A_{\overline{n_d}} \end{array}\right]$$

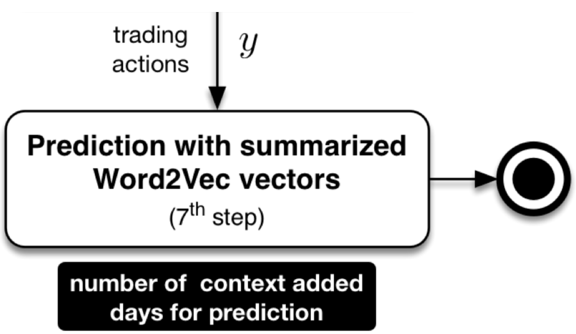


University of Ljubljana
Faculty of Computer and Information Science

Prediction (7th step) (3)

- **Prediction with summarized Word2Vec vectors**
  - we add context by **adding previous n_m trading days** to the current trading day,
  - the result is **contextualized matrix** with context vectors for given word (i.e. trading day) that is a sum of previous trading days,

$$\left[X''_{(n_v \times \overline{n_d}')}\; y_{(1 \times \overline{n_d}')}\right] = \left[\begin{array}{cccc|c} cv_{1,1} & cv_{1,2} & \dots & cv_{1,n_v} & A_1 \\ cv_{2,1} & cv_{2,2} & \dots & cv_{2,n_v} & A_2 \\ \dots & \dots & \dots & \dots & \dots \\ cv_{w'_{\overline{n_d}},1} & cv_{w'_{\overline{n_d}},2} & \dots & cv_{w'_{\overline{n_d}},n_v} & A_{\overline{n_d}'} \end{array}\right]$$

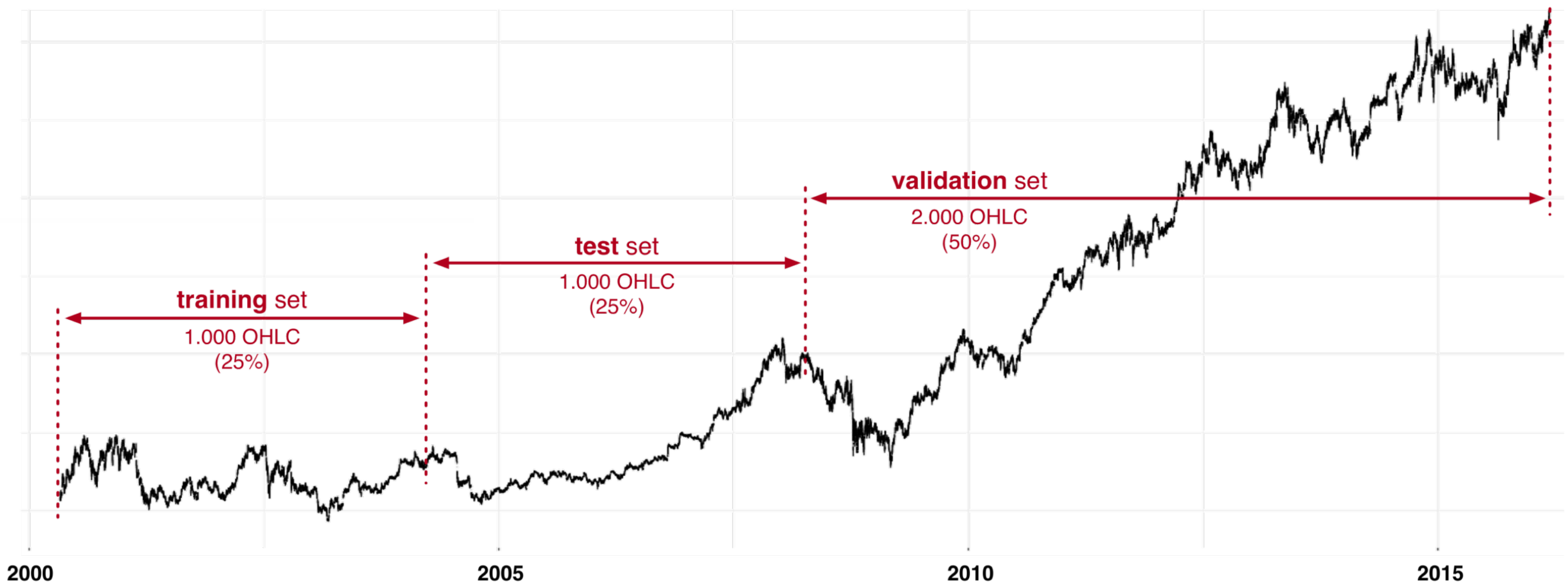


Evaluation (1)

- considered various performance metrics (e.g. total hit ratio, MSE, classification accuracy, AUC, logarithmic loss etc.),
- decided to evaluate using a **trading strategy with initial equity** ($10.000,00) and selected prediction model, **including trading fees** ($15) that penalize numerous trading actions which decrease the profitability of prediction model utilization.

Evaluation (2)

- historical data included **4.000 OHLC** trading days, starting from 1. 5. 2000.



Evaluation (3)

- the initial proposed model was evaluated on shares of **Apple (AAPL)**, **Microsoft (MSFT)** and **Coca-Cola (KO)**,
- the final evaluation was performed on **Russell Top 50 Index** (50 stocks of the largest companies in the U.S. stock market),
  - the model was trained for each individual stock,
- the results compared with existing trading strategies **Buy & Hold**, **Moving Average (MA)** and **MACD**.

Evaluation (4)

- **average yield** with initial equity of $10.000,00

	Buy & Hold	MA(50,100)	MACD	W2V
Test phase	$2,818.98	$1,073.06	-$482.04	$11,725.25
Validation phase	$16,590.83	$6,238.43	$395,10	$10,324.24

- **Wilcoxon Signed Rank Test** for forecast models

	Buy & Hold		MA(50,100)		MACD	
	W	p-value	W	p-value	W	p-value
Test phase	2	$< .0001$	1	$< .0001$	1	$< .0001$
Validation phase	-	-	427	$< .0210$	155	$< .0001$

  – with the correct selection of parameters our model achieves statistically significantly better yields than the reference popular methods.

Conclusion

- proposed model has a great **potential for practical use**,
- it is too early to conclude that the proposed model provides a financial gain,
  - selected model parameters are not equally appropriate for different time periods in terms of yield,
- the forecast model is strongly influenced by the training data set,
- if training with data that contains bear trend, the model might be very cautious despite the general growth trend,
  - the problem is due to **overfitting**, so training with **more data** would help.

Future Work

- incorporate the **stop loss function** and already known and proven technical indicators,
- the use of **OHLC data of other stocks** in the training phase as we acquire more diverse patterns that helps algorithms to detect the underlying pattern better,
- to improve classification accuracy and logarithmic loss, the SoftMax algorithm could also be replaced with **advanced machine learning classification algorithms**,
- alternative method of **forecasting as a simple linear operation of aggregating vector representations** of the last n Japanese candlesticks,
  - we could obtain a daily, weekly or monthly trend forecast.

Thank you for your attention

Marko Poženel and **Dejan Lavbič**

What Language Do Stocks Speak?

Dejan.Lavbic@fri.uni-lj.si

https://www.linkedin.com/in/**dejan/**

http://www.**lavbic.net**

@dlavbic