Short-Text Classification with Deep neural networks:

An Experimental Analysis

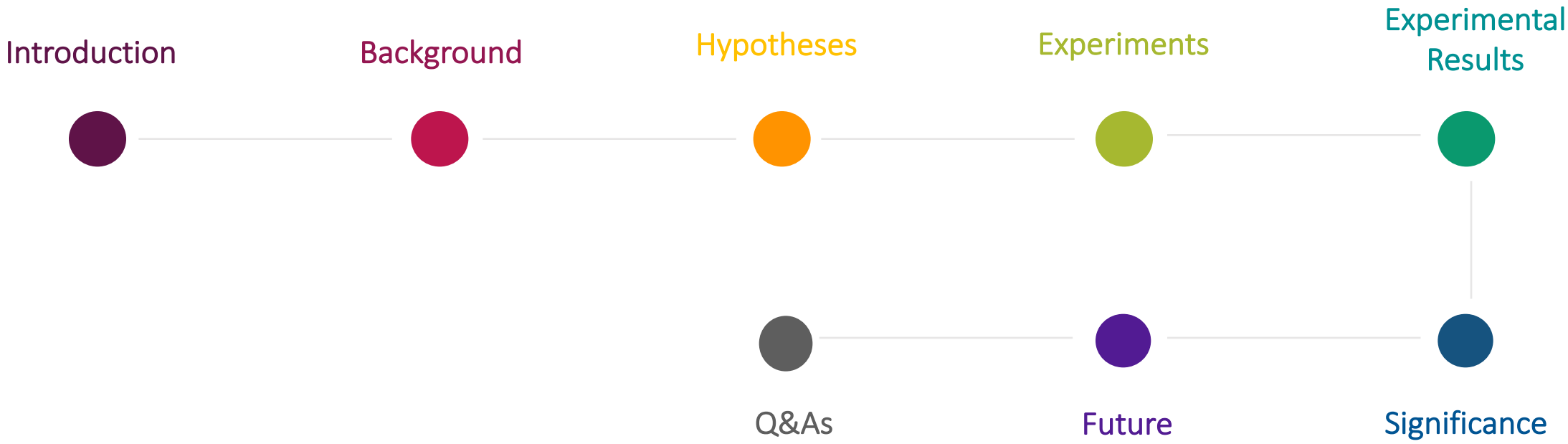
Iris Liu Chui Yi

April 2017

Final Year Project Presentation

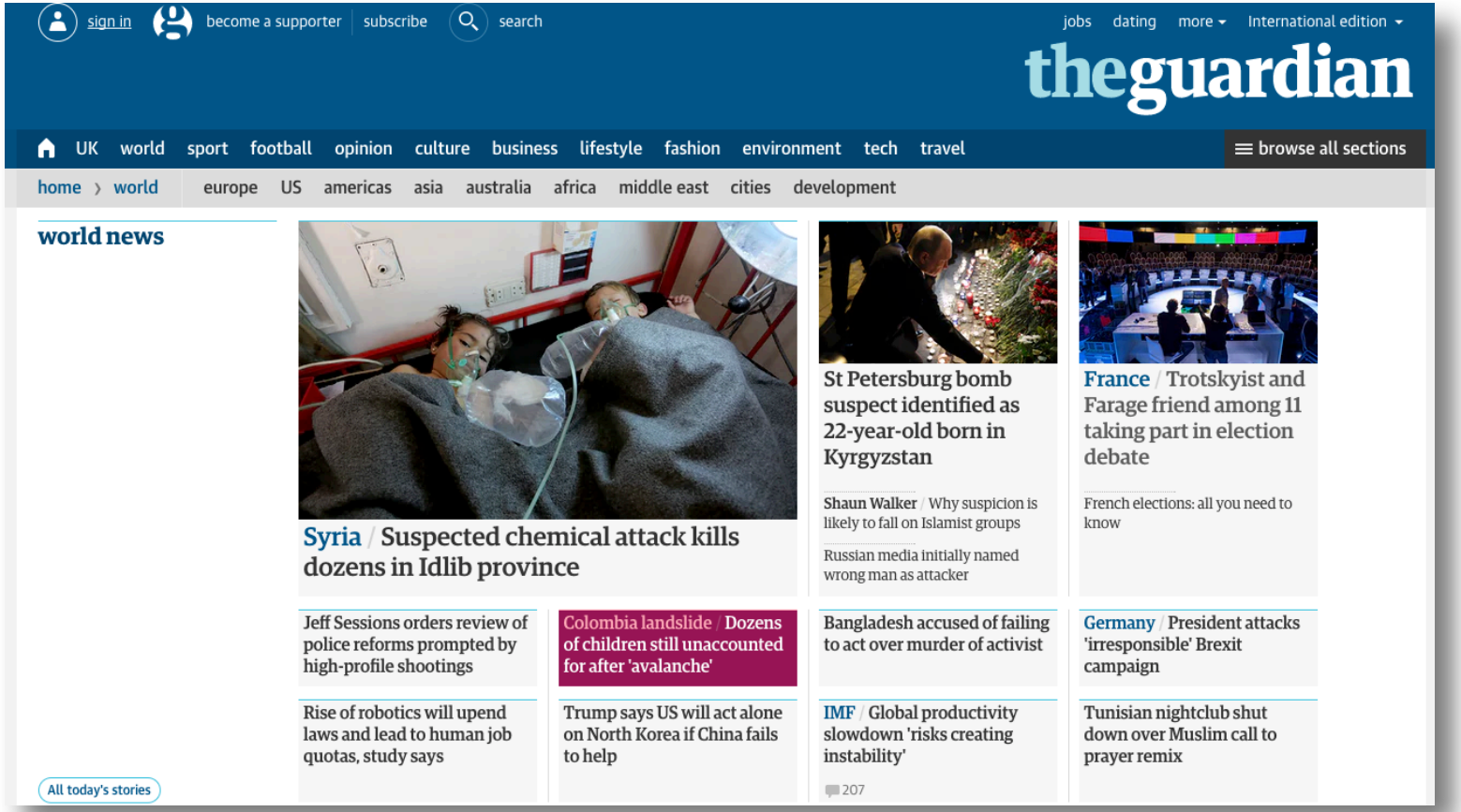
CityU
香港城市大學
City University of Hong Kong

Agenda


Introduction
Background
Hypotheses
Experiments
Experimental Results
Q&As
Future
Significance

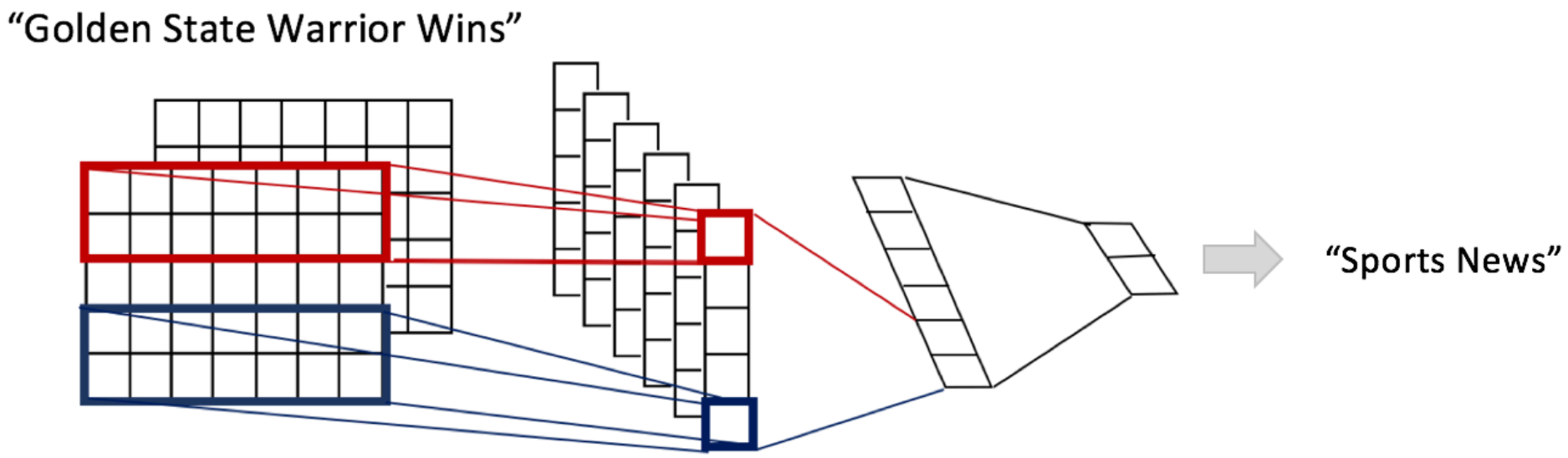
Problem Definition

- Extreme Text Classification with many class labels



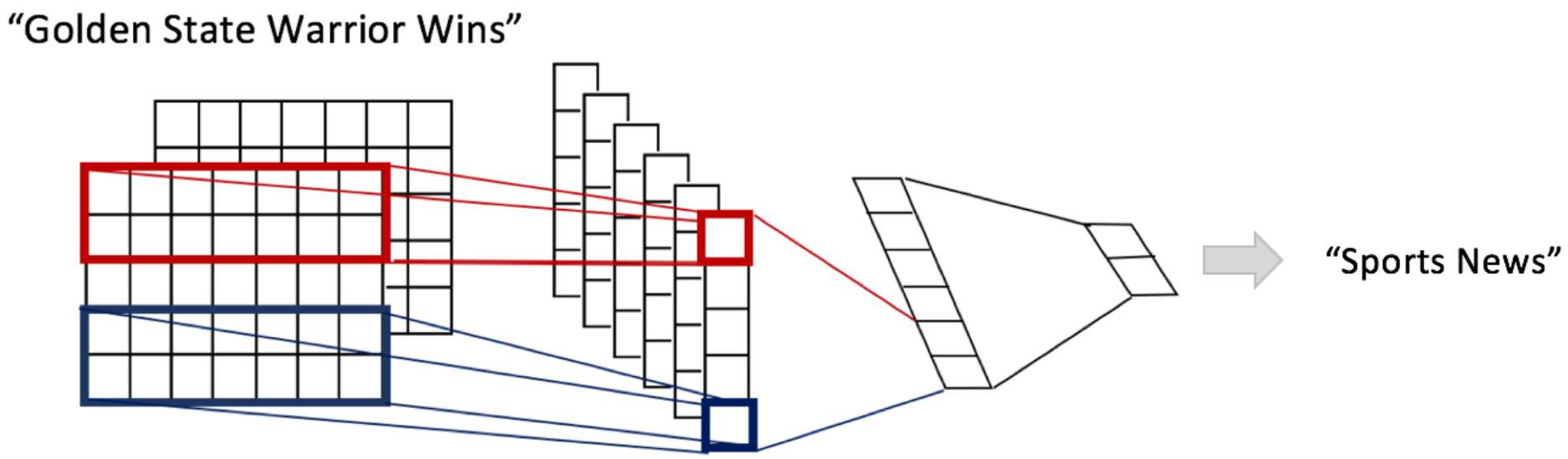
Short-Text Classification

- Input: short-text data (300 words on average)
- Output: class labels, categories



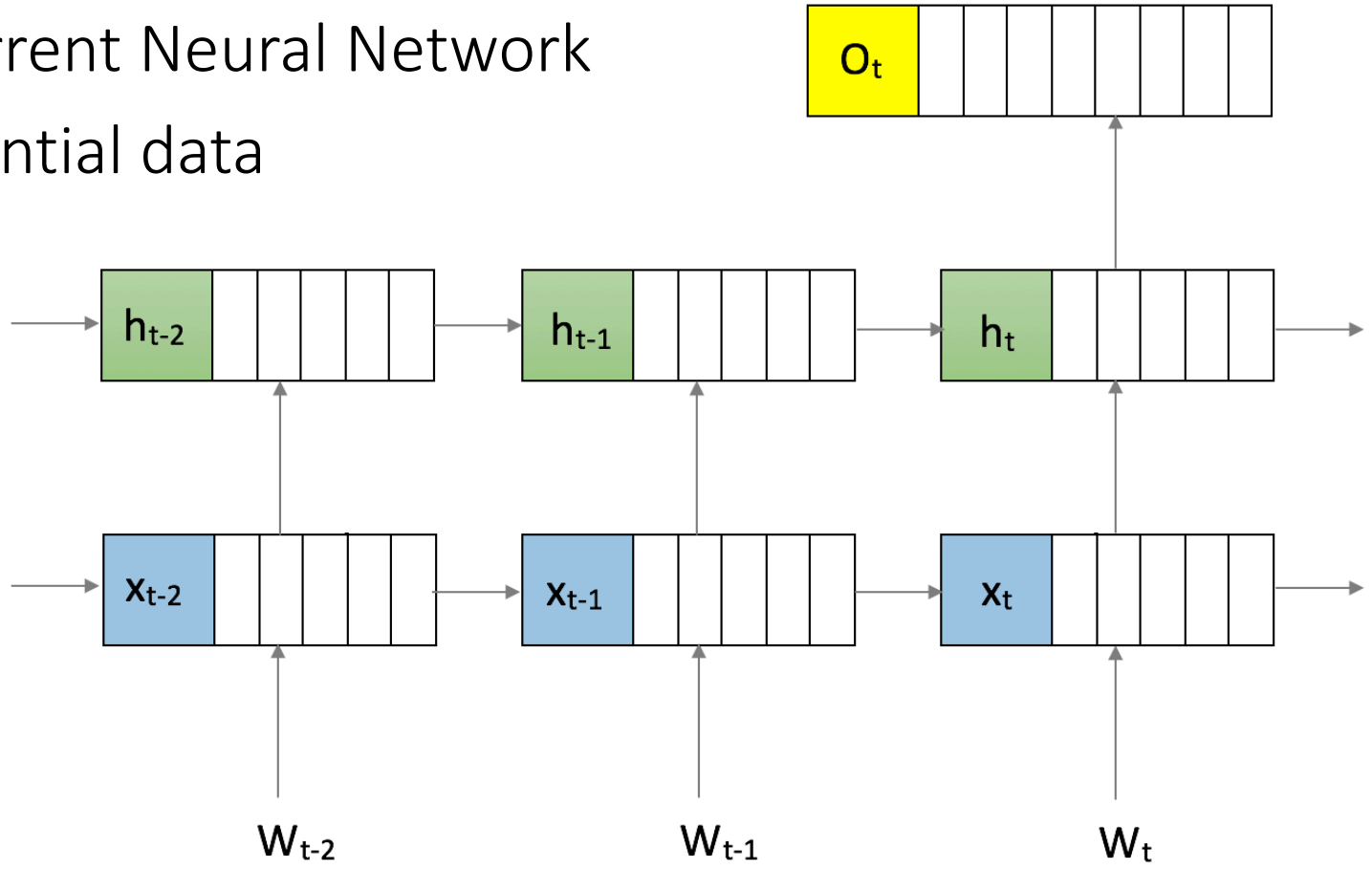
Convolutional Neural Networks (CNN)

- Usually works well for spatial data such as images, speech signals
- Use convolutional window to capture features



Long-Short-Term Memory (LSTM)

- A variation of a Recurrent Neural Network
- Particularly for sequential data



x_t Word Vector

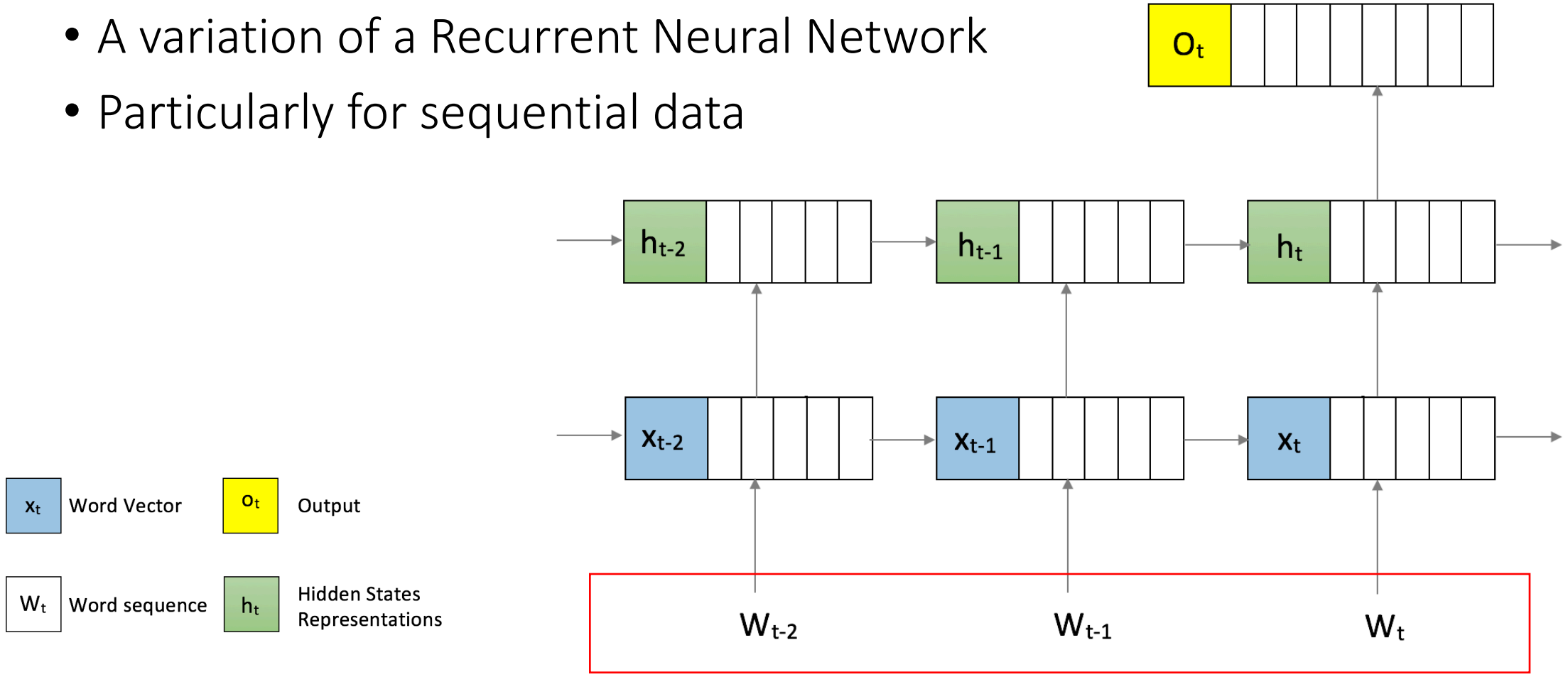
o_t Output

W_t Word sequence

h_t Hidden States Representations

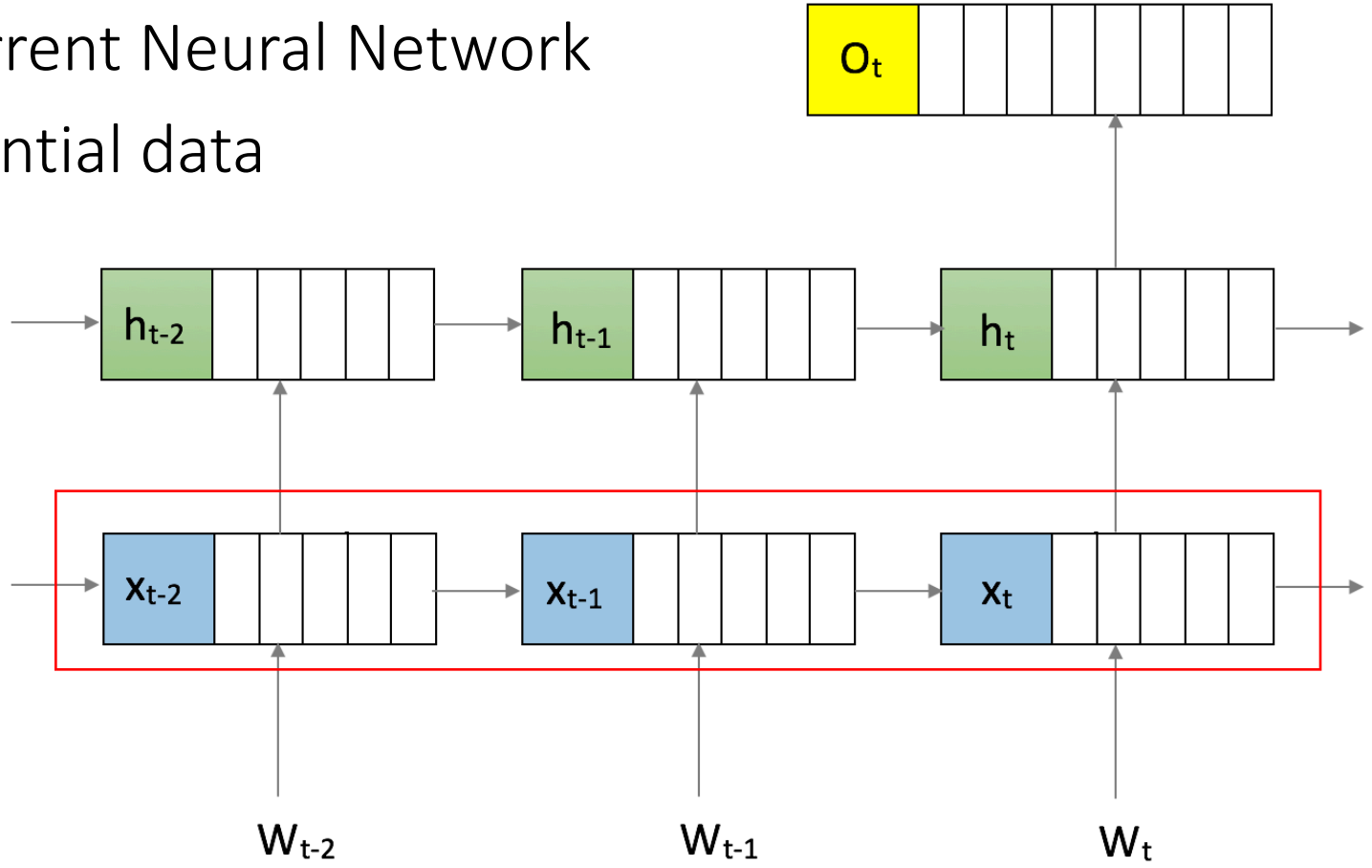
Long-Short-Term Memory (LSTM)

- A variation of a Recurrent Neural Network
- Particularly for sequential data



Long-Short-Term Memory (LSTM)

- A variation of a Recurrent Neural Network
- Particularly for sequential data



x_t Word Vector

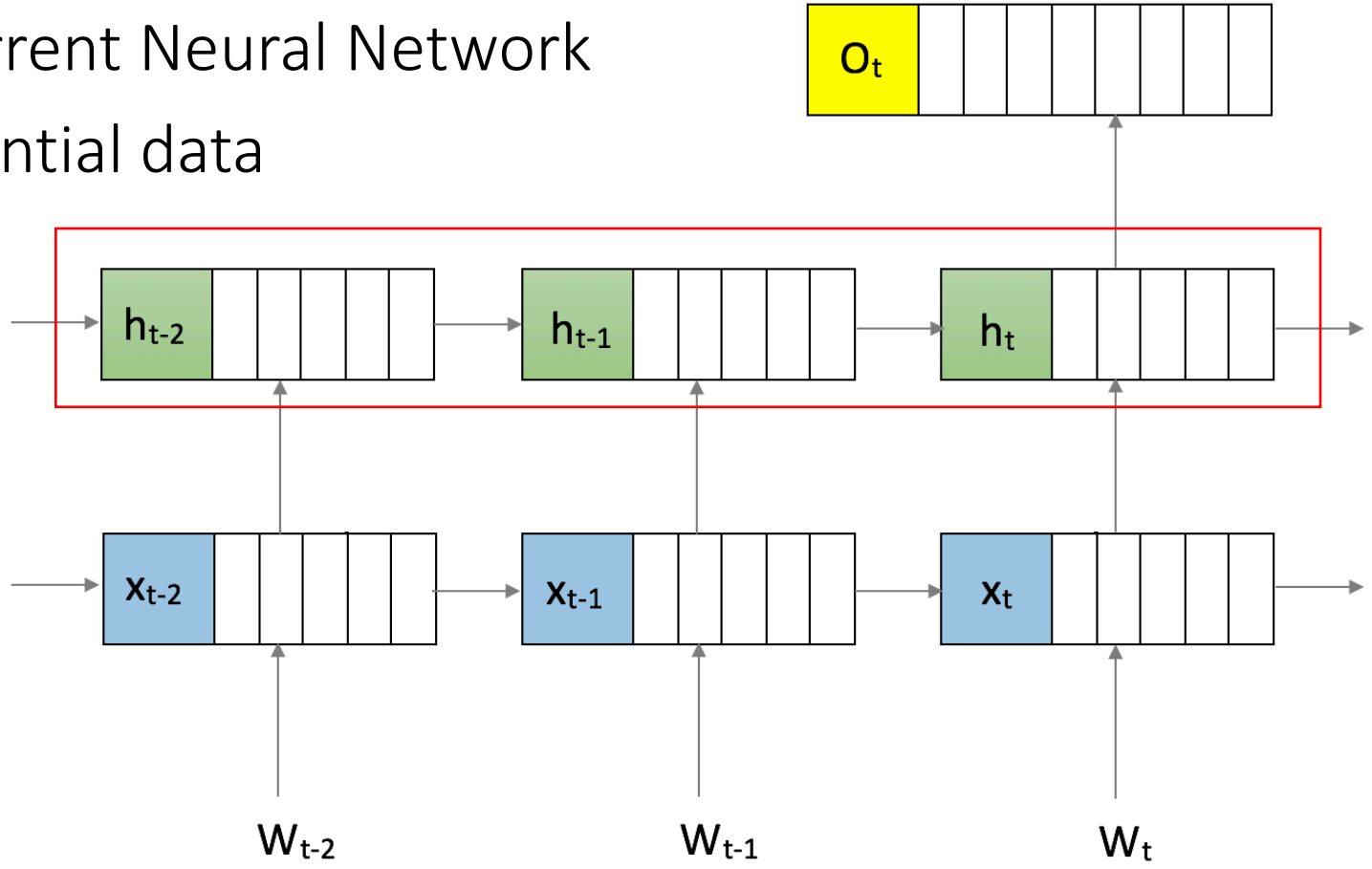
o_t Output

W_t Word sequence

h_t Hidden States Representations

Long-Short-Term Memory (LSTM)

- A variation of a Recurrent Neural Network
- Particularly for sequential data

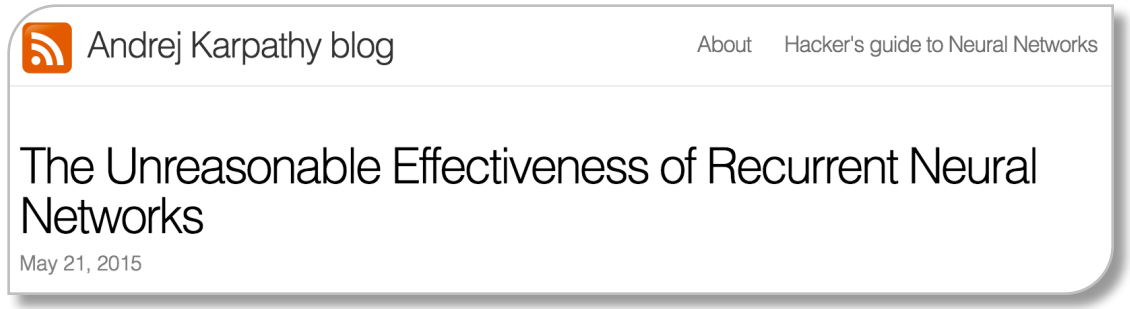


x_t Word Vector

o_t Output

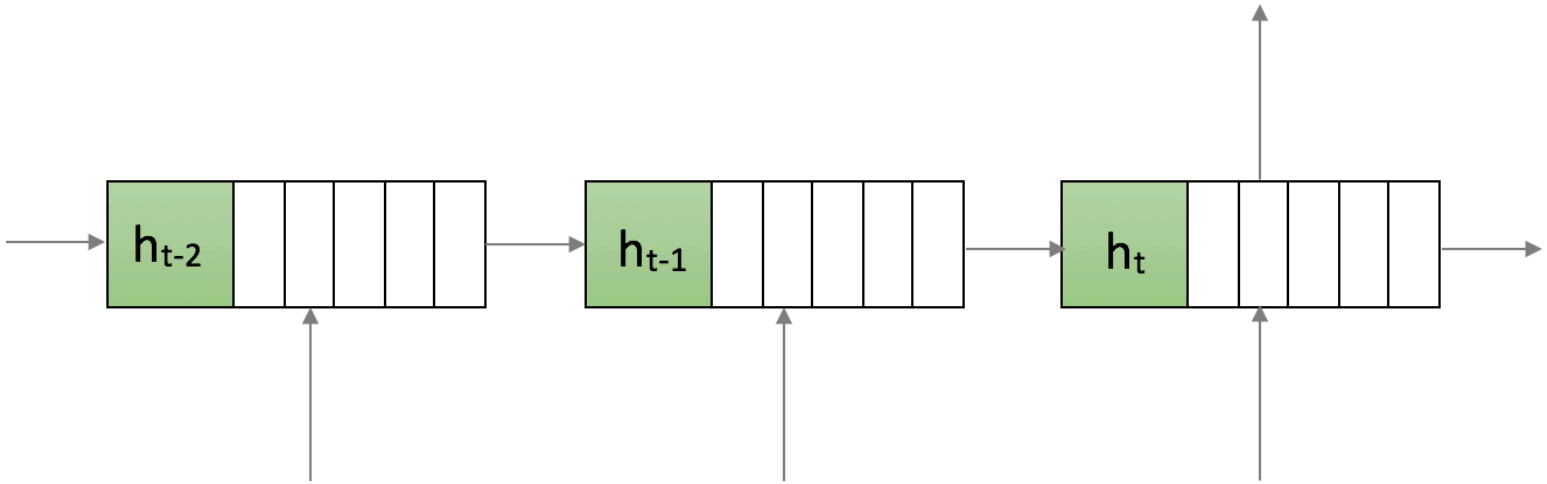
W_t Word sequence

h_t Hidden States Representations



Cell that turns on inside quotes:

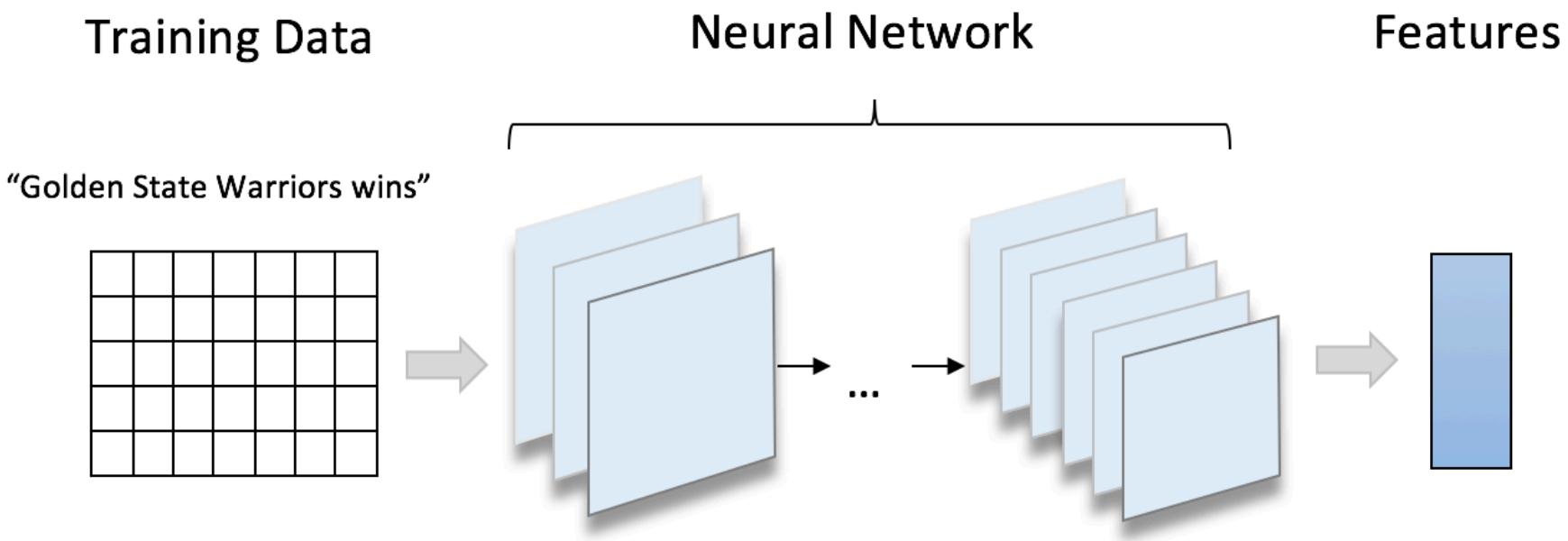
```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```



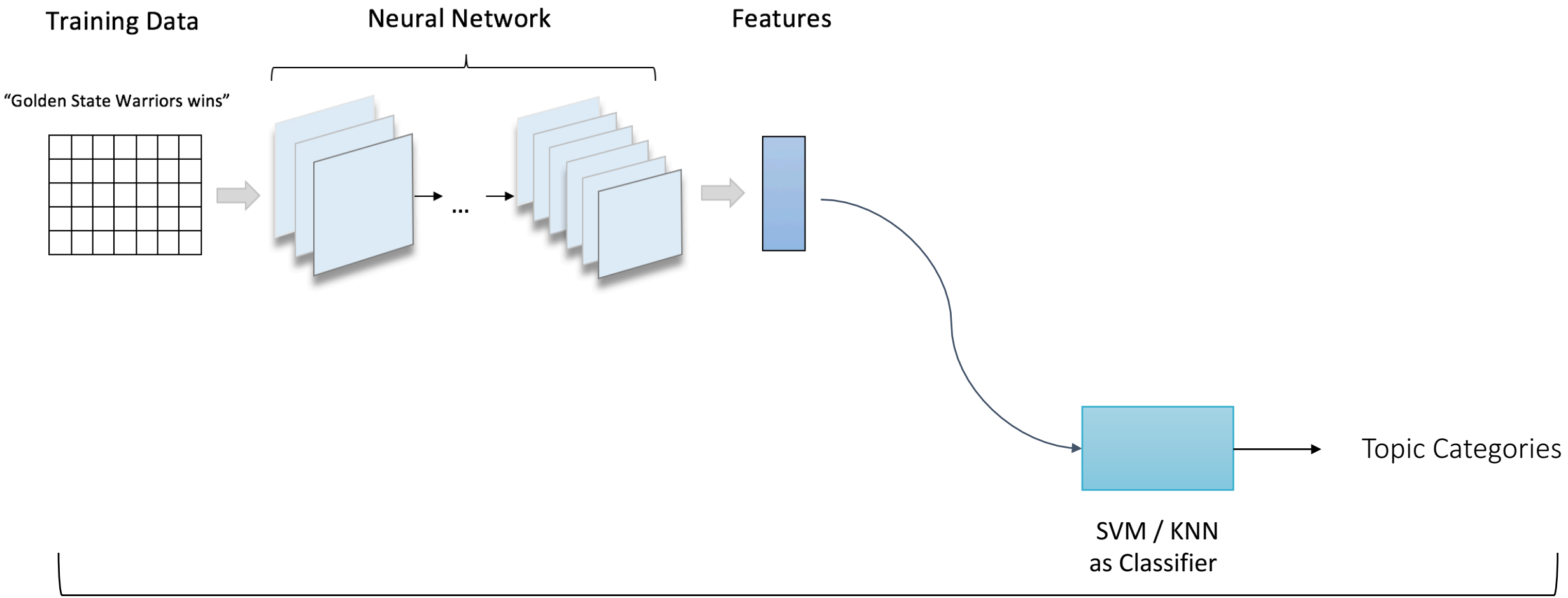
What LSTMs have captured in its hidden state can be regarded as features

What LSTMs have captured in its hidden states can be regarded as features

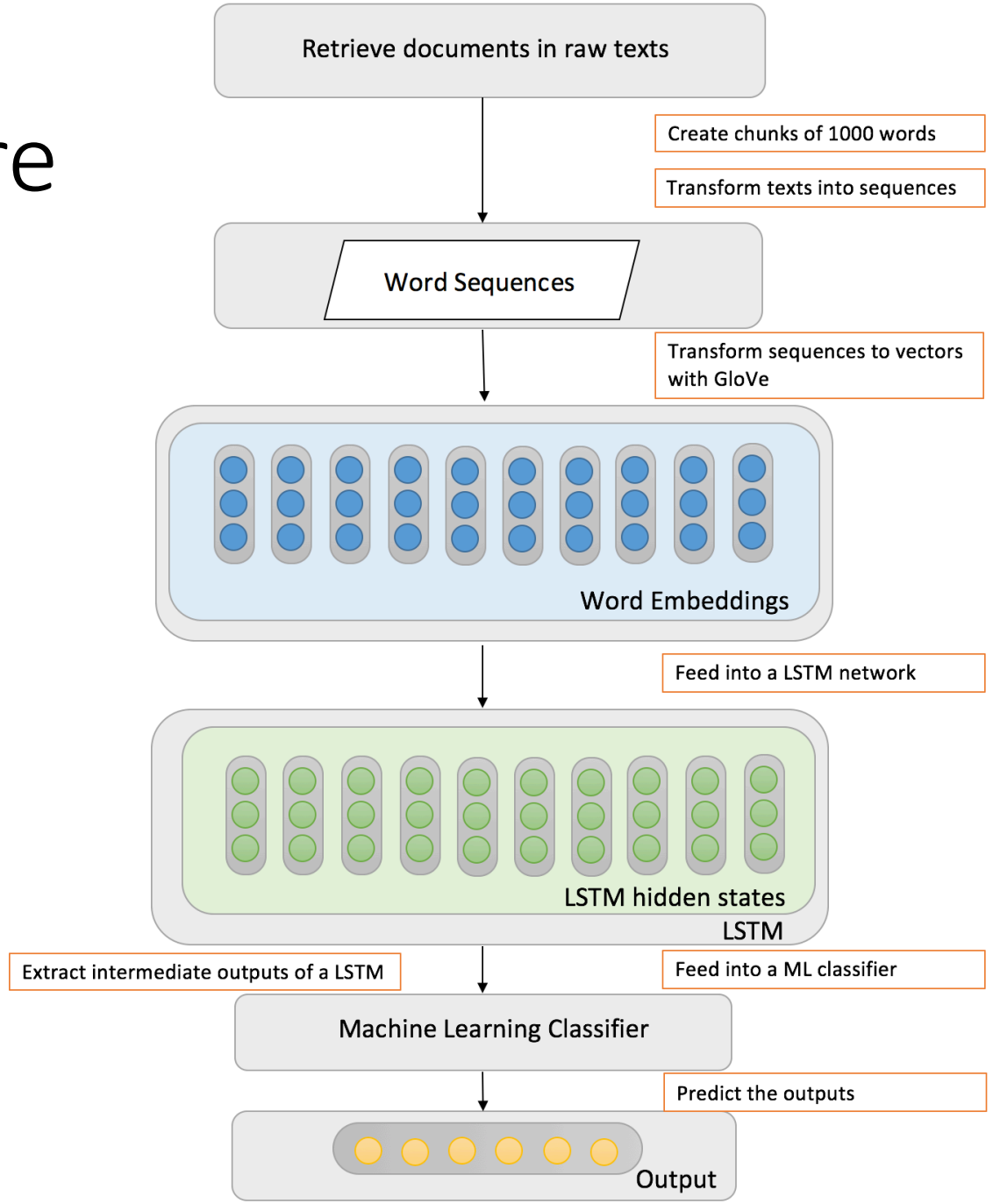
LSTM as a feature extractor



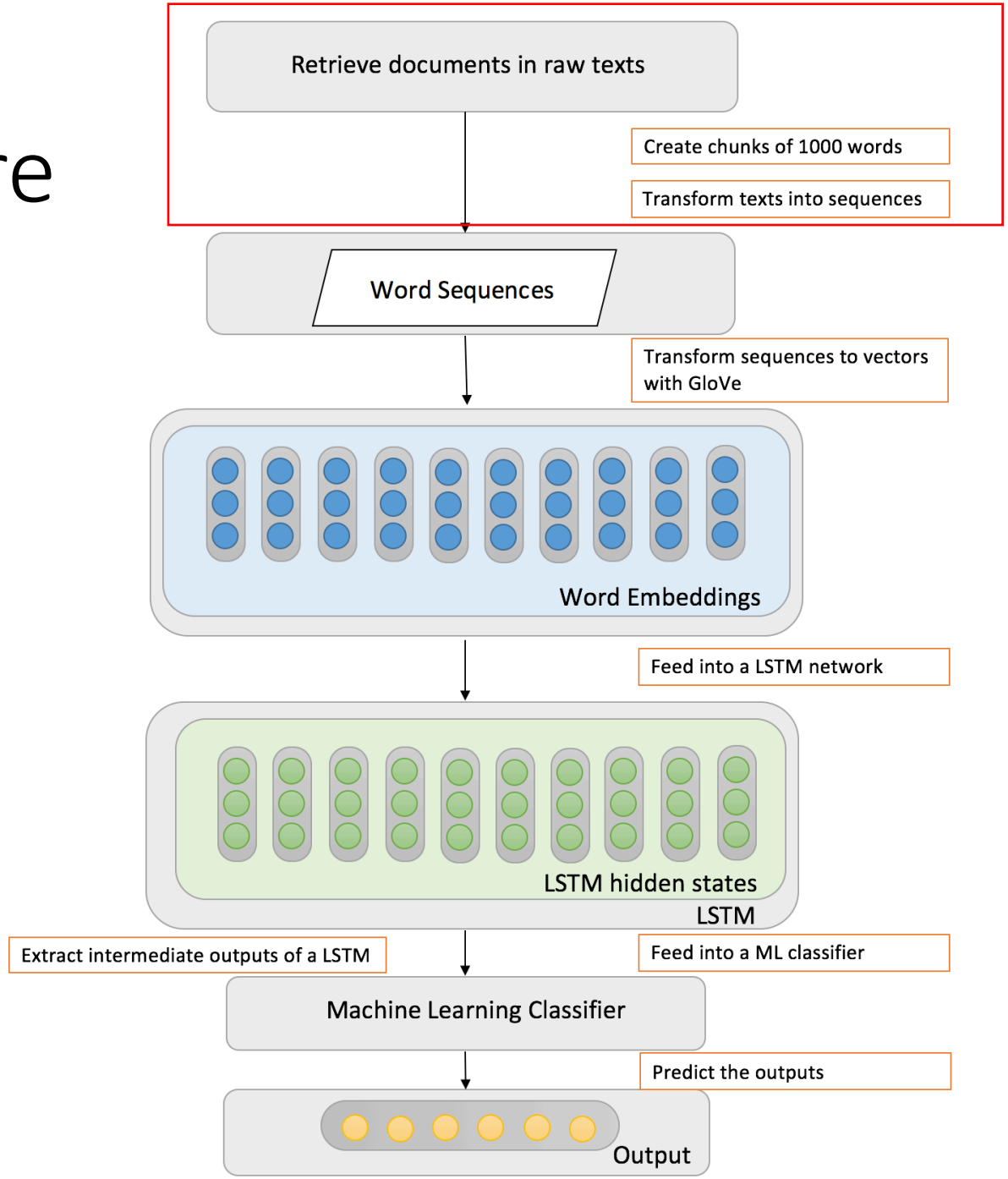
Machine Learning algorithms for classification



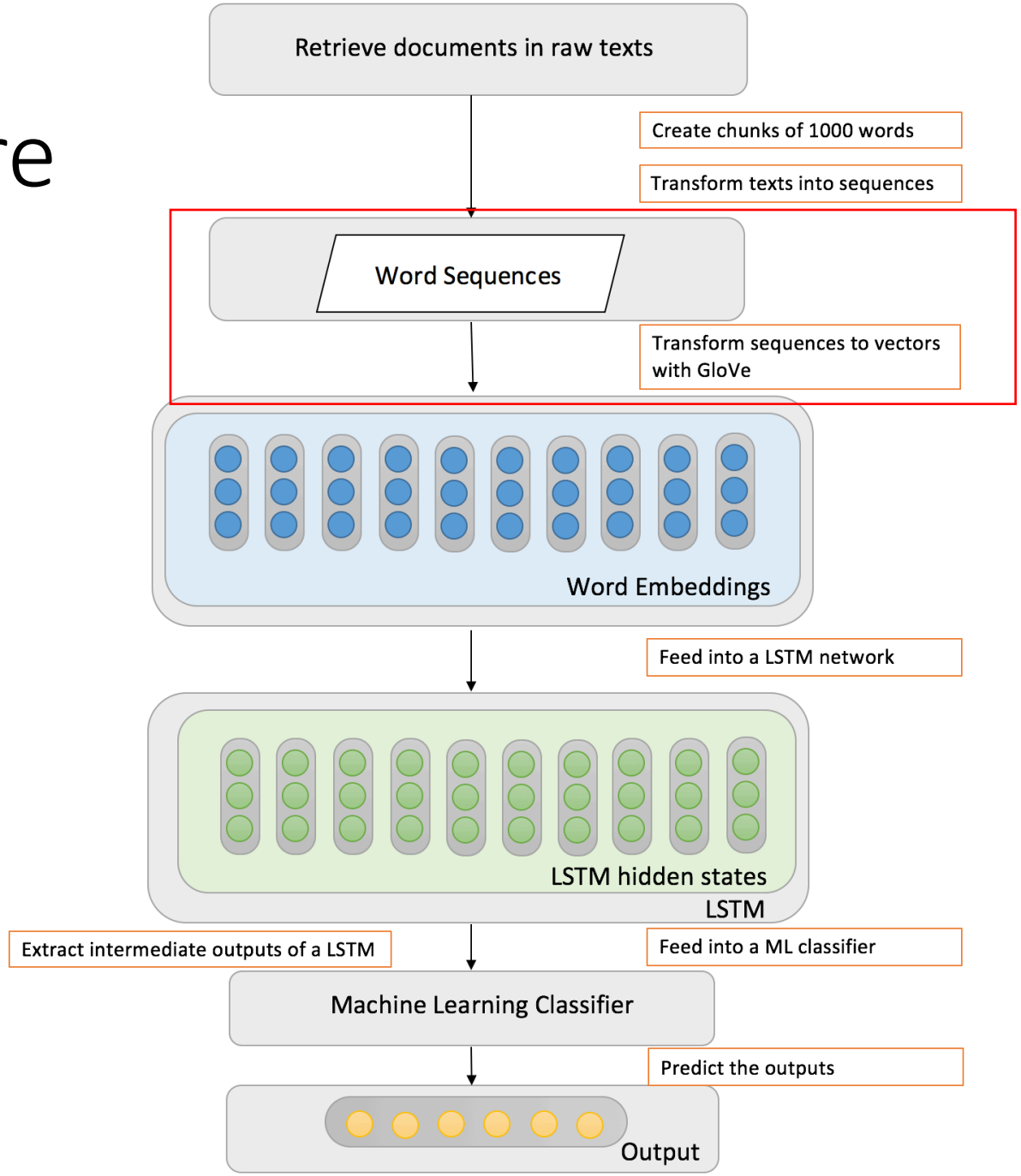
Our Network Architecture



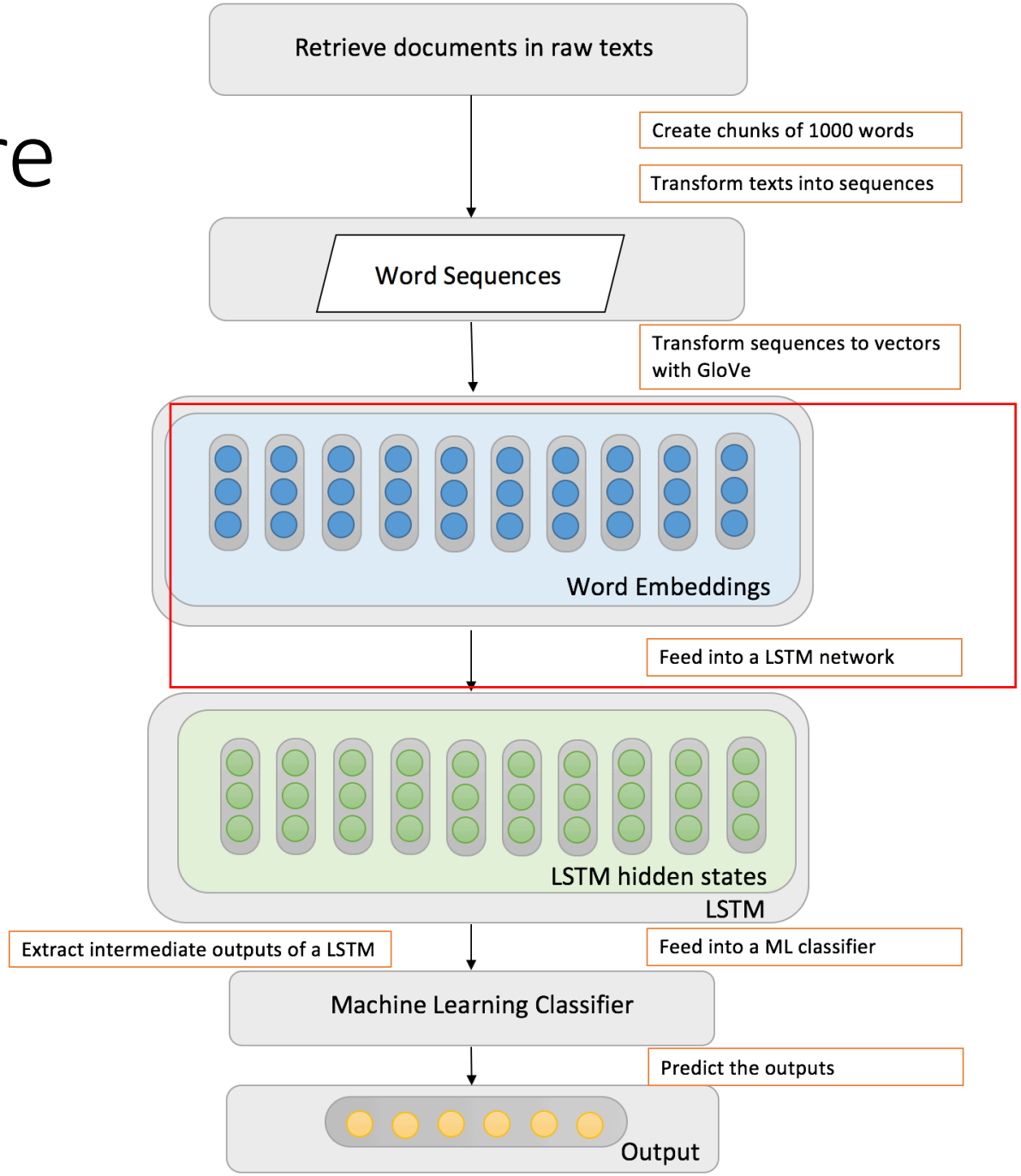
Our Network Architecture



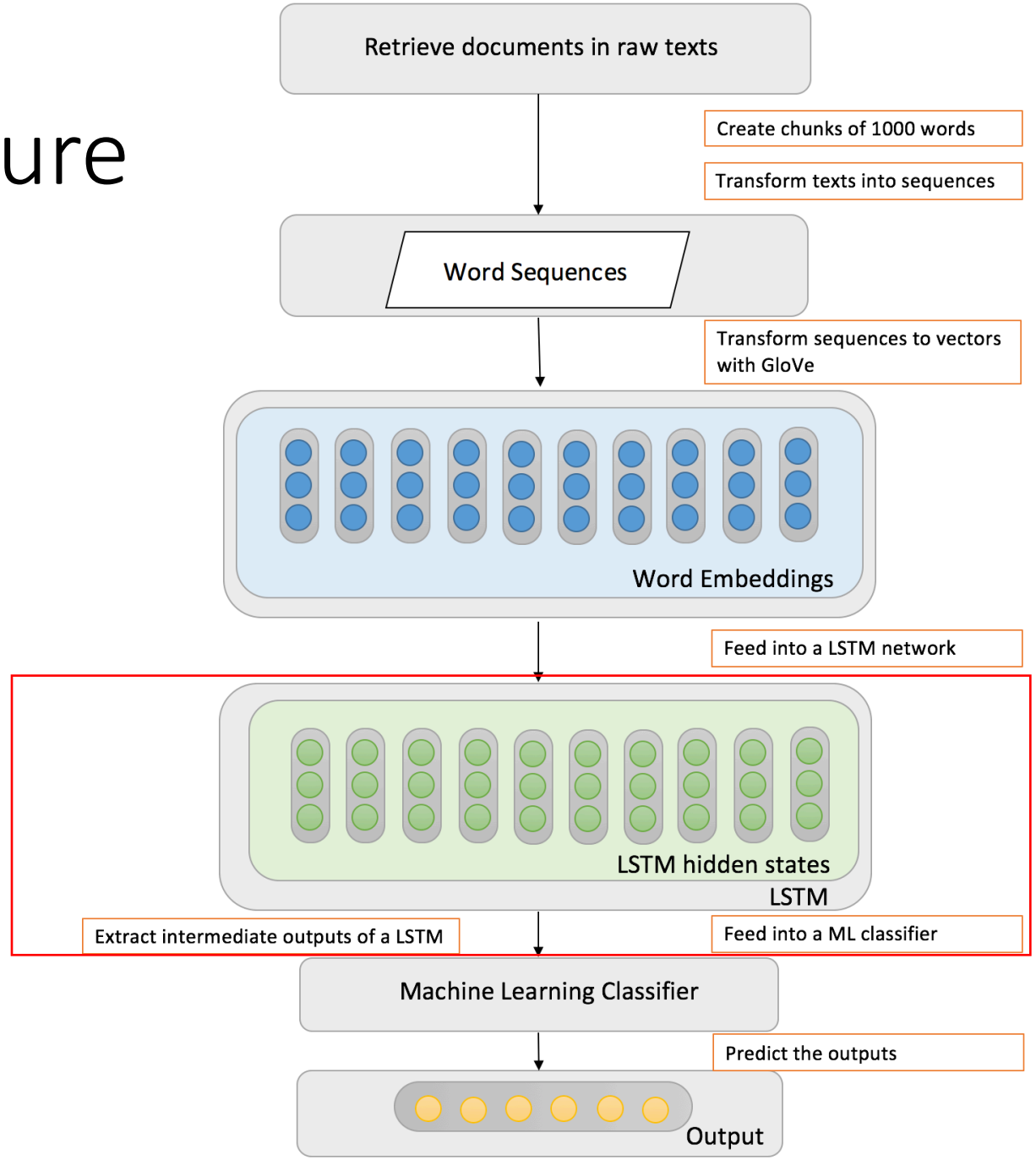
Our Network Architecture



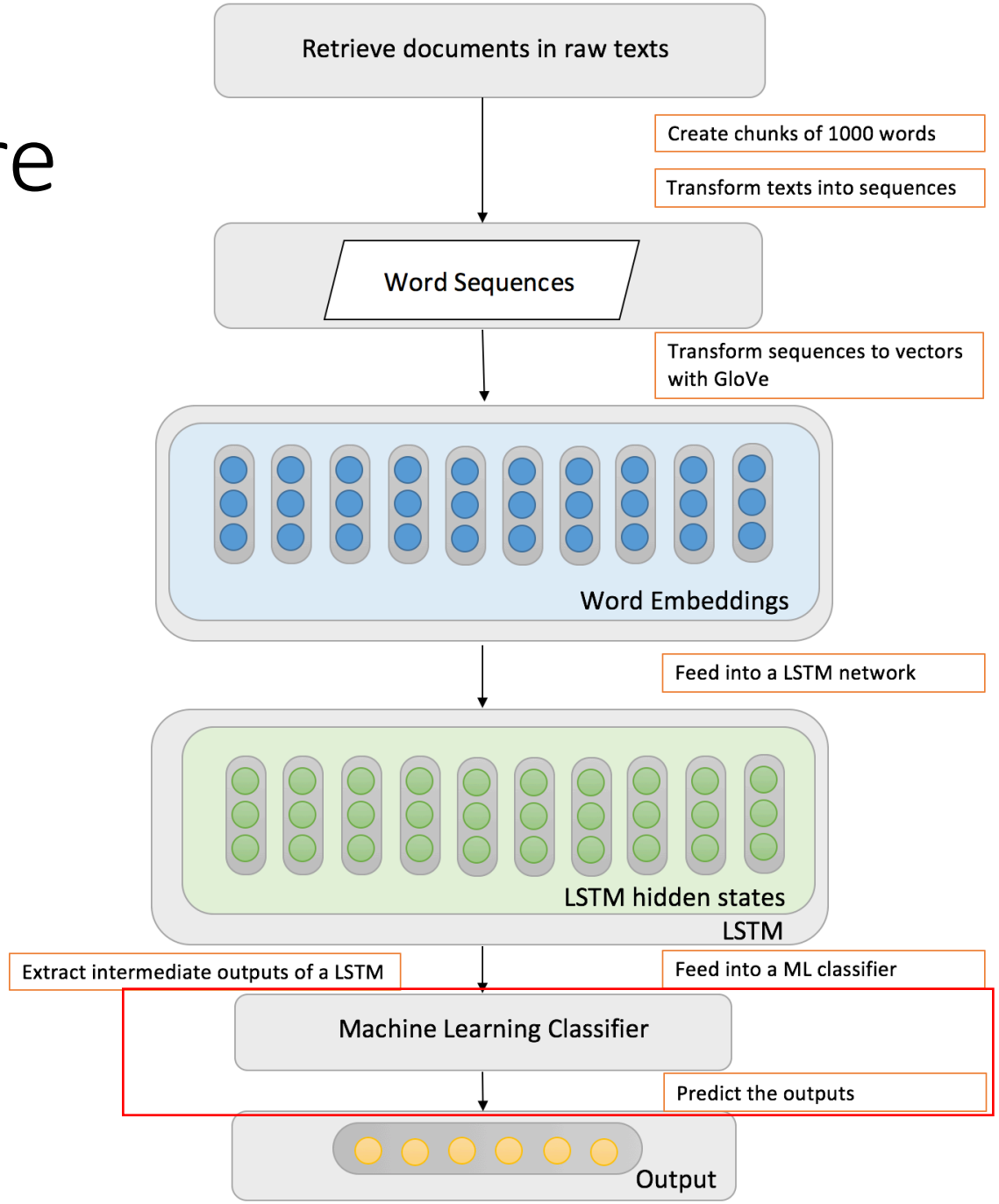
Our Network Architecture



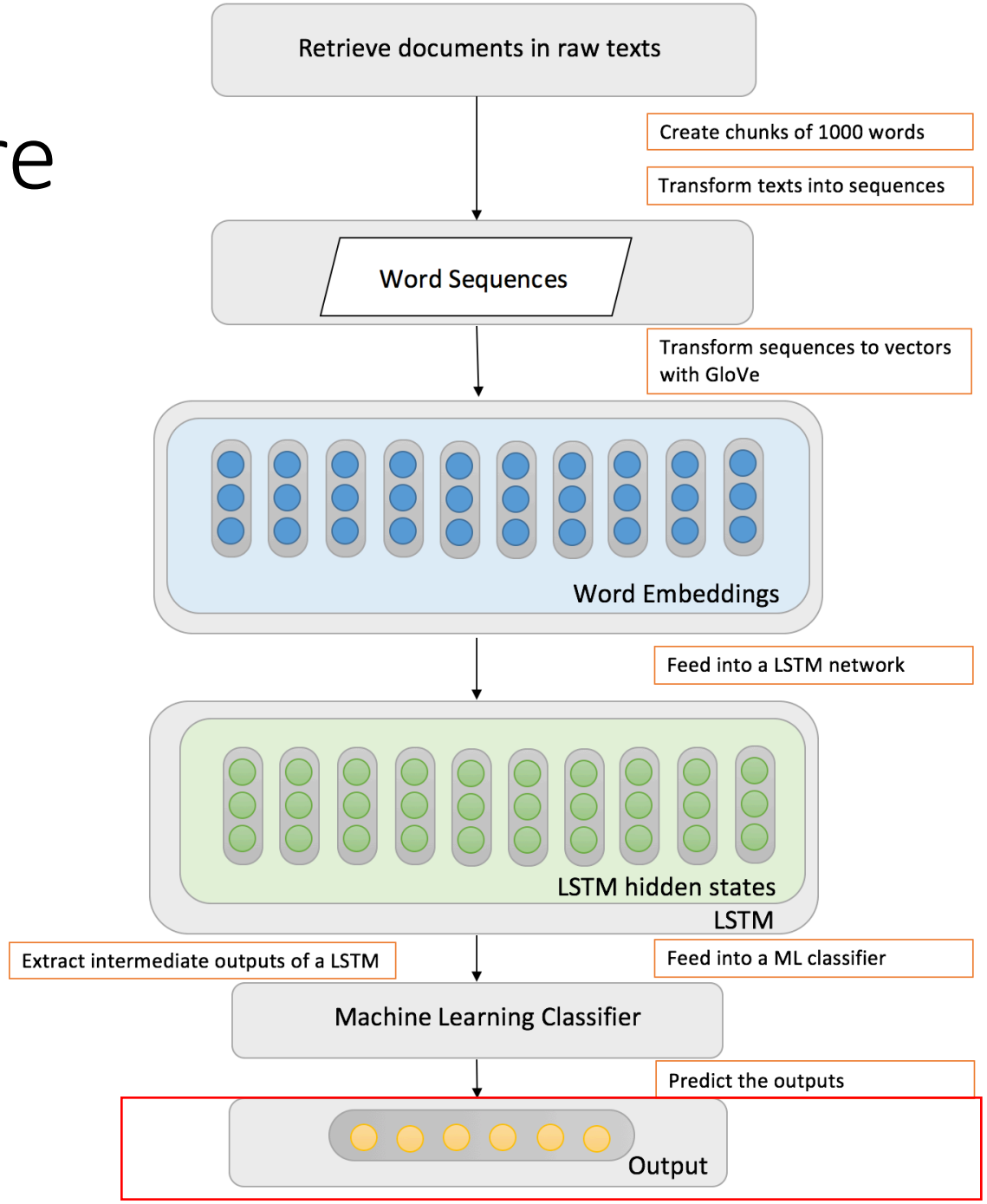
Our Network Architecture



Our Network Architecture

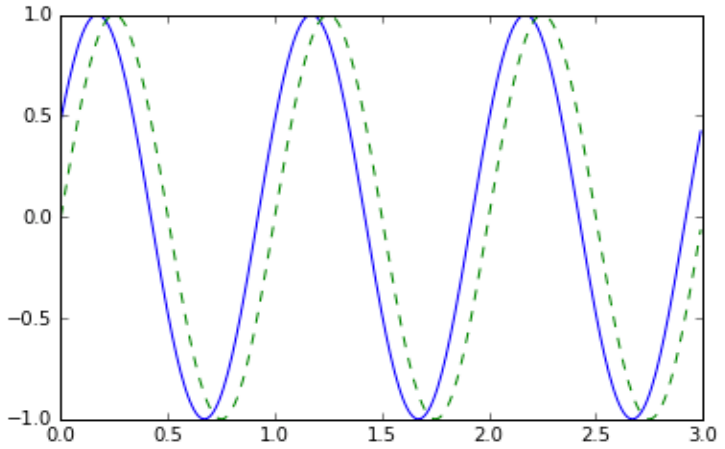


Our Network Architecture



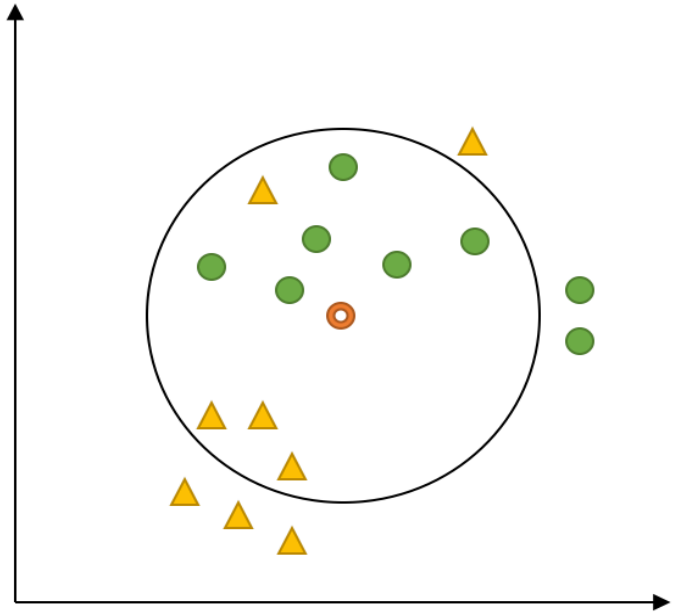
What is special about Time Series Data?

- Natural language is mostly time-series, e.g. speech signals, dialogs
- They are not only vary in sequences, but also in terms of time propagation

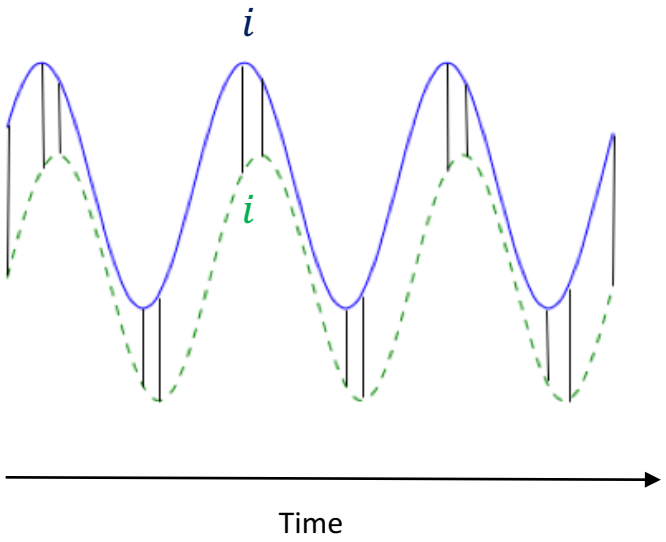


K-Nearest Neighbours

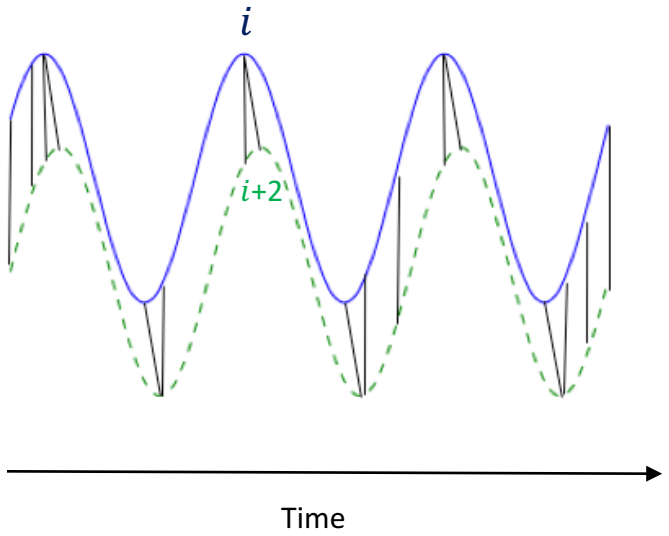
- Usually adopt Euclidean Distance as distance metrics
- Maybe not be accurate in time-series



Why Dynamic Time Warping?



Euclidean distance which align the i-th point on one sequence with the i-th point with the other will give poor similarity measure.



A non-linear alignment produces a more a intuitive similarity measure, matching sequences that are similar in shape but are out of shape.

Hypotheses

Introduction Background Hypotheses

Hypothesis 1:	Neural Networks outperform Machine Learning Algorithms

Hypothesis 2:	LSTMs should perform better than CNNs in short texts

- Compare Neural Networks with Machine Learning Models
- Compare between LSTMs & CNNs

Hypothesis 3:	Temporal LSTMs help improve the classification accuracy of machine learning algorithms.

Hypothesis 4:	Structural Models produces similar results to LSTMs

- Compare Structural Model with Machine Learning models and neural networks

Hypothesis 5:	DTW is better distance metrics than Euclidean distance for KNN

- Compare KNN-DTW with KNN-Euclidean models

Hypothesis 6: Classification accuracy drops as the number of class labels increase

- Compare different models across different number of classes
- Created Benchmark datasets with different datasets of different number of classes

Experimental Setups

Datasets

20 Newsgroups

Yahoo! Answers

Ohsumed

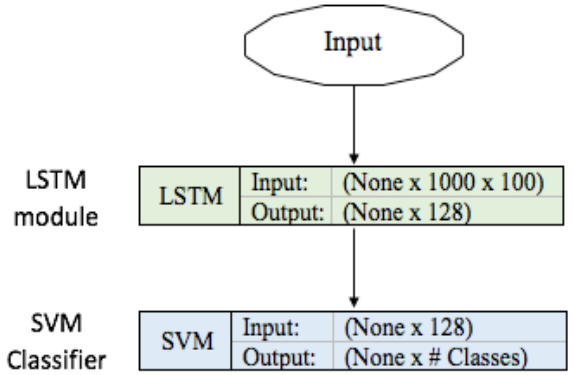
Model Set-ups

- SVM (with Tf-idf)
- KNN (with Tf-idf)
- LSTM
- CNN
- LSTM+SVM
- LSTM+KNN-Euclidean
- LSTM+KNN-DTW

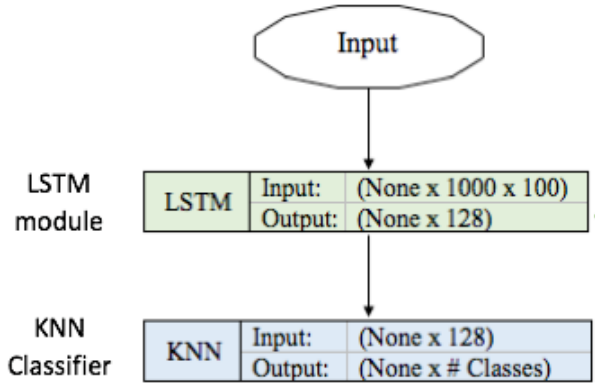
Machine Learning Models

Neural Networks

Structural Models



LSTM+SVM model

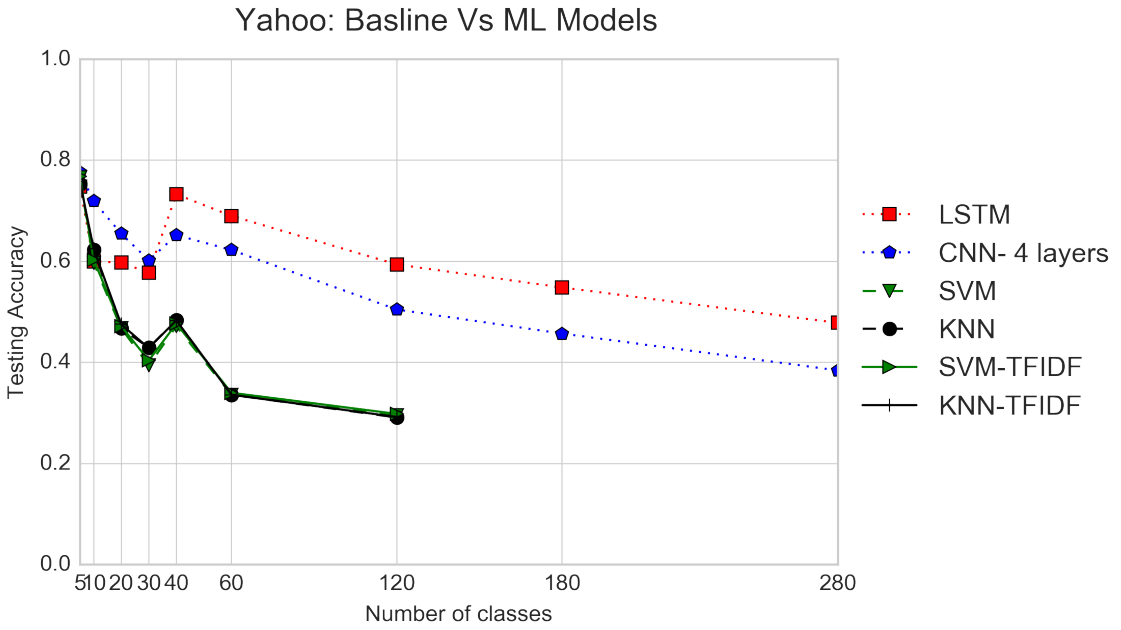


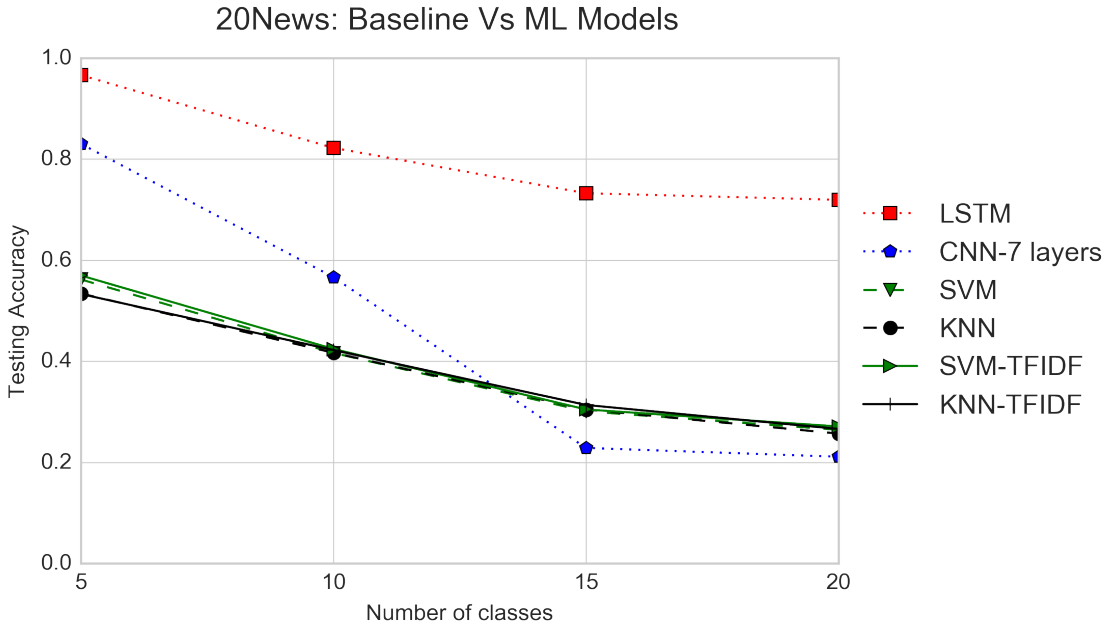
LSTM+KNN model

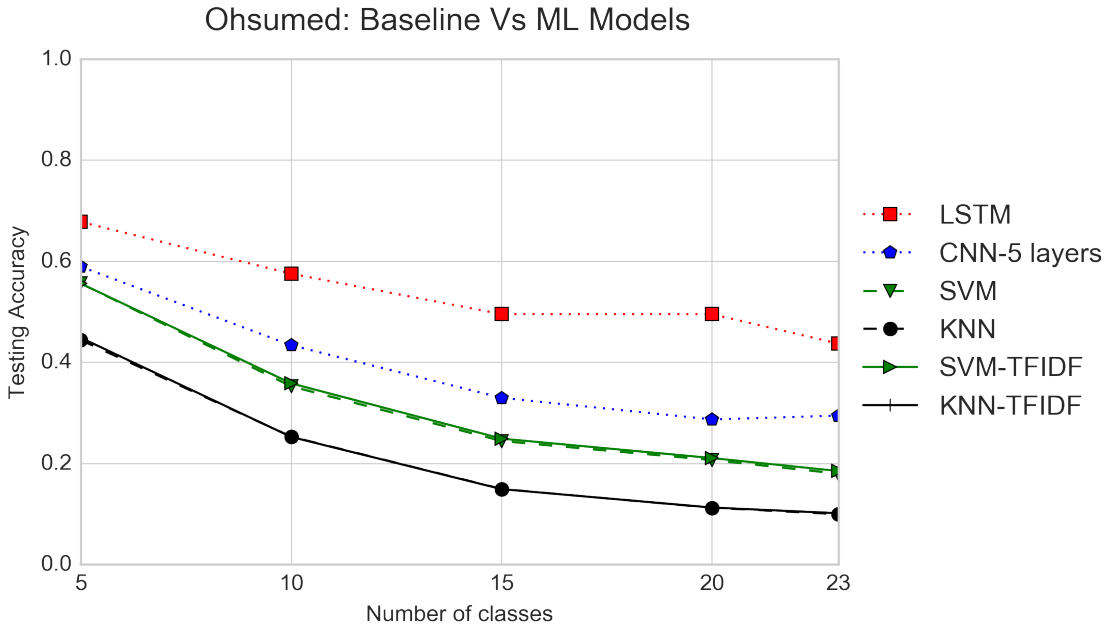
Experimental Results

Findings

1. Neural Networks outperform ML models
2. LSTMs outperform CNNs

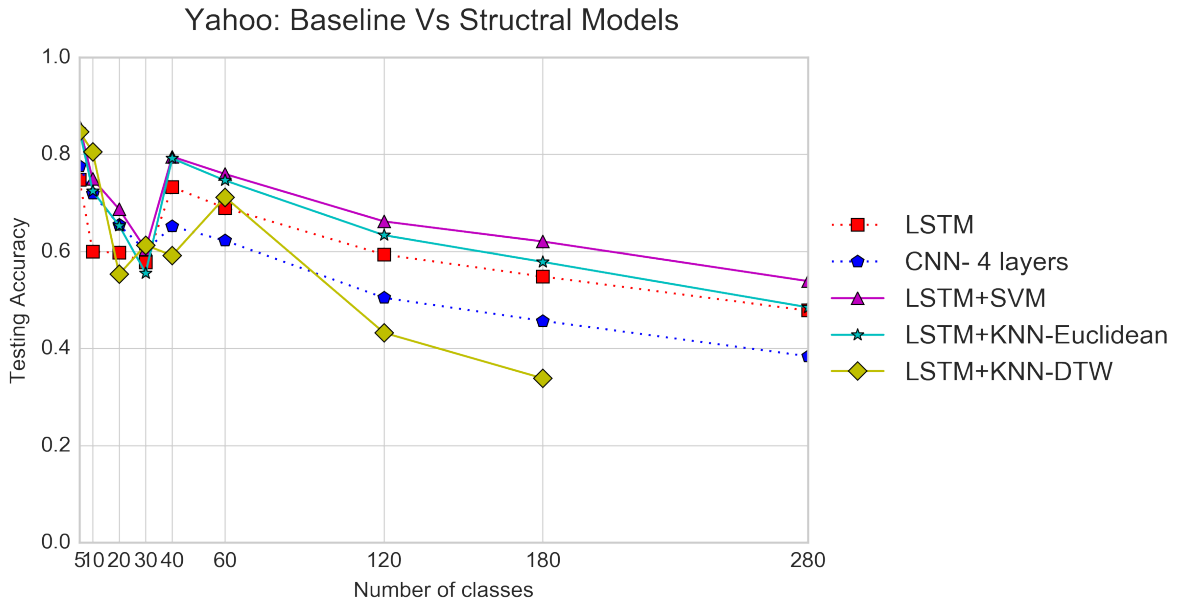


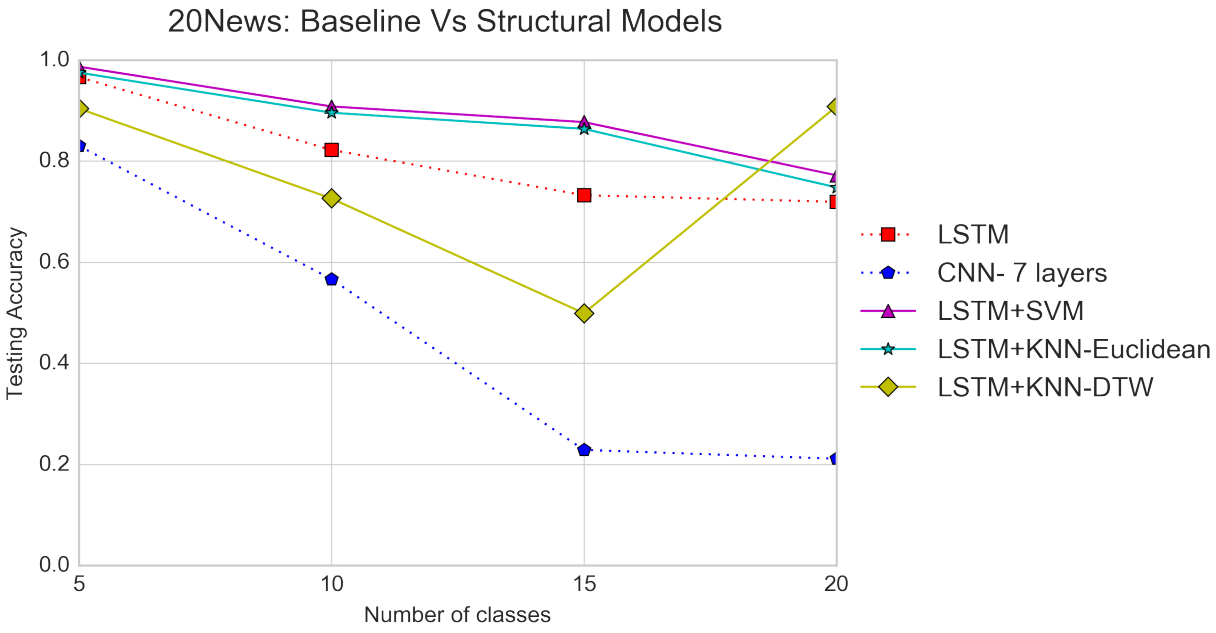


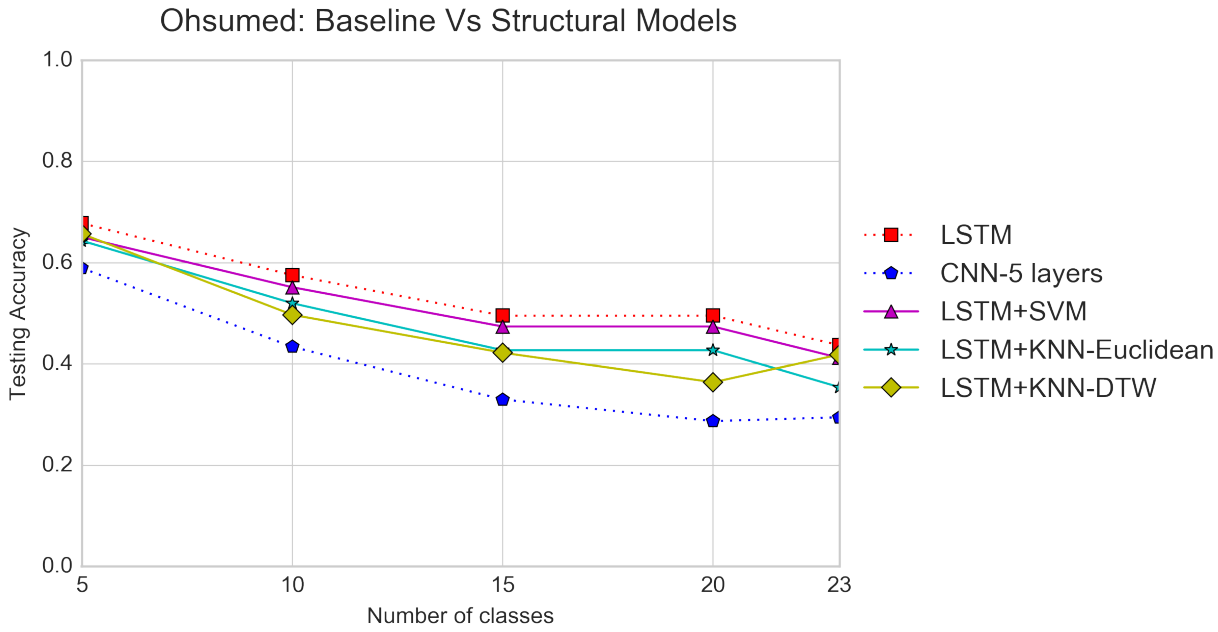


Findings

3. Structural models outperform Neural Networks on 20news and Yahoo datasets

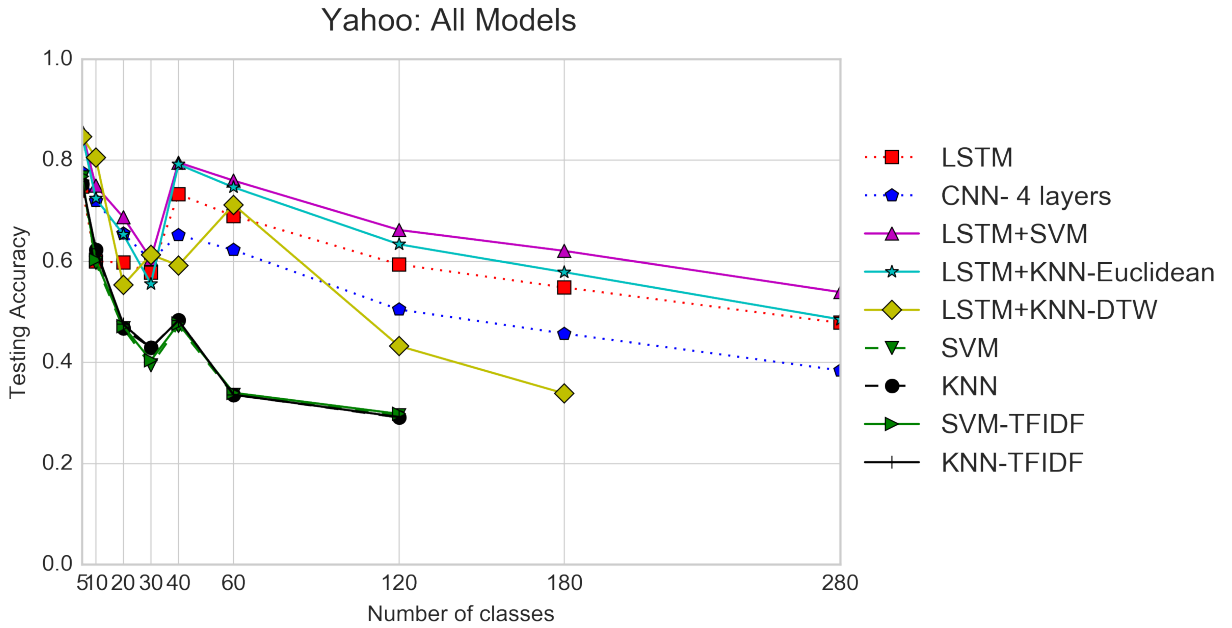


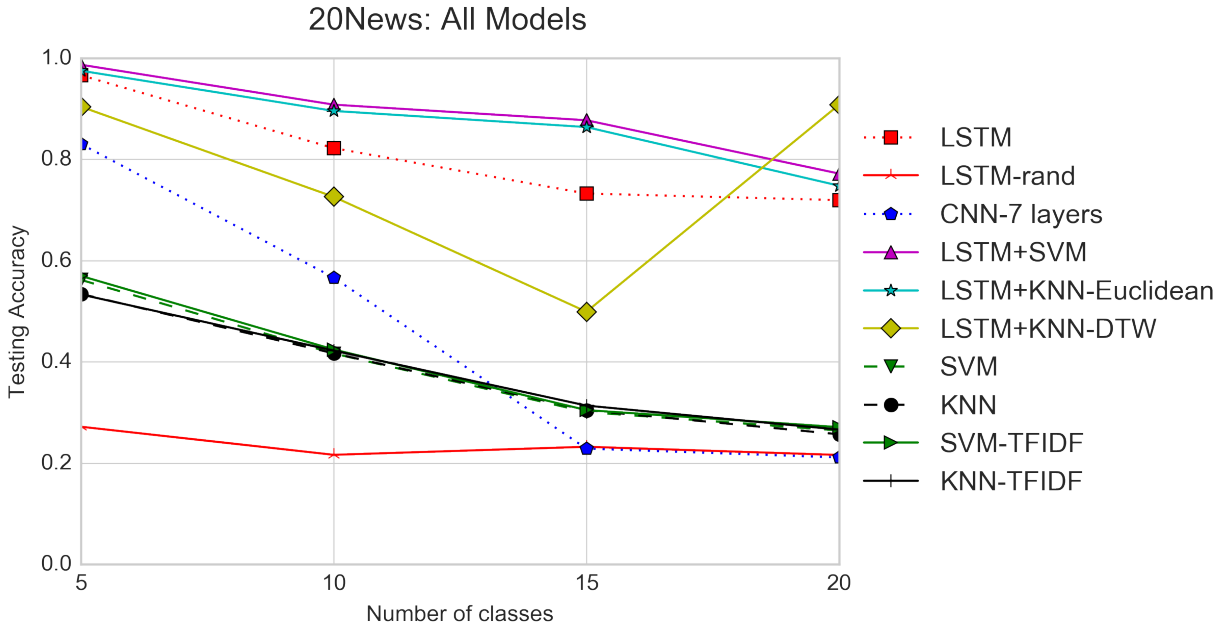


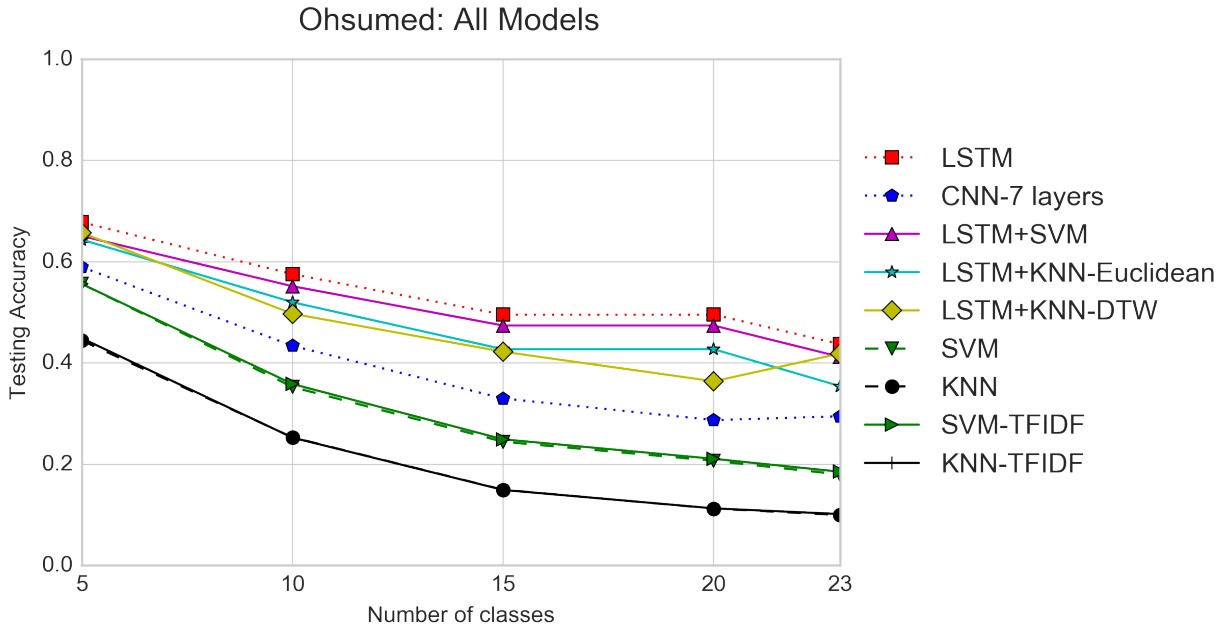


Findings

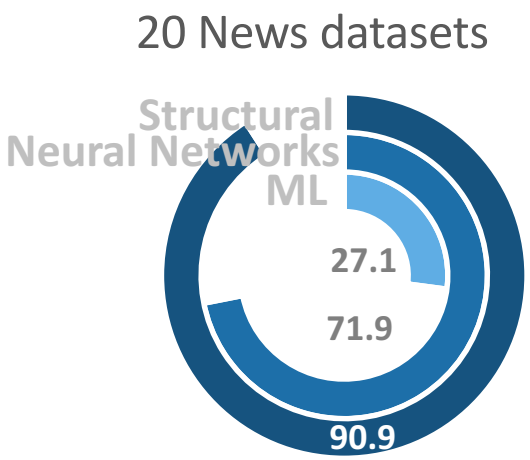
4. Structural Models outperform ML models largely
5. DTW does not shows an improvement with KNN
6. Overall decreasing trends are observed

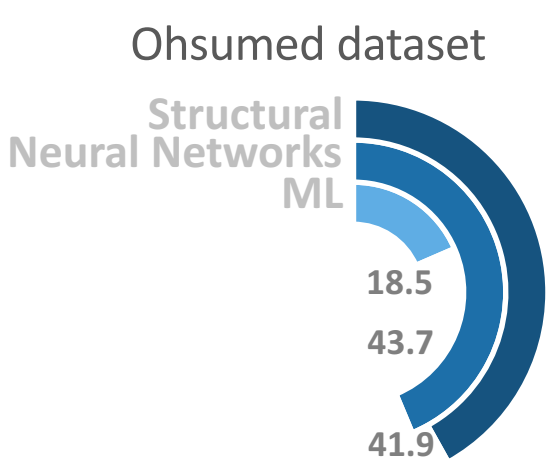


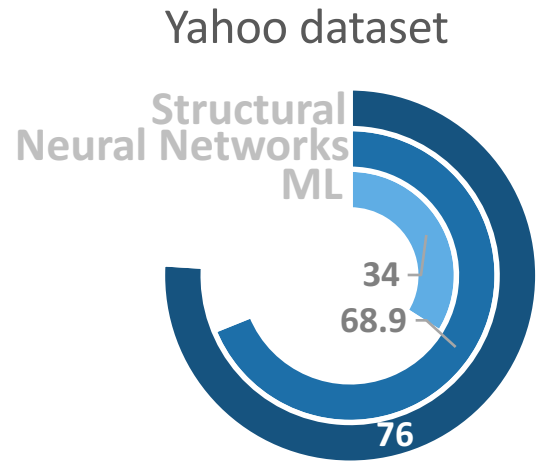


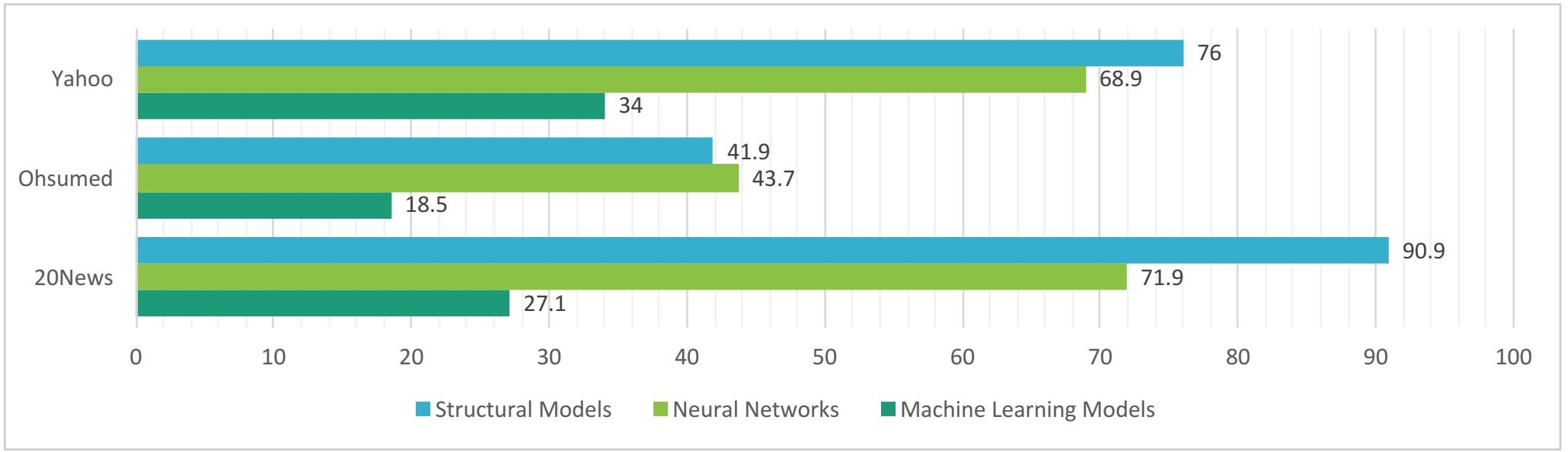


Performance dashboard of all models









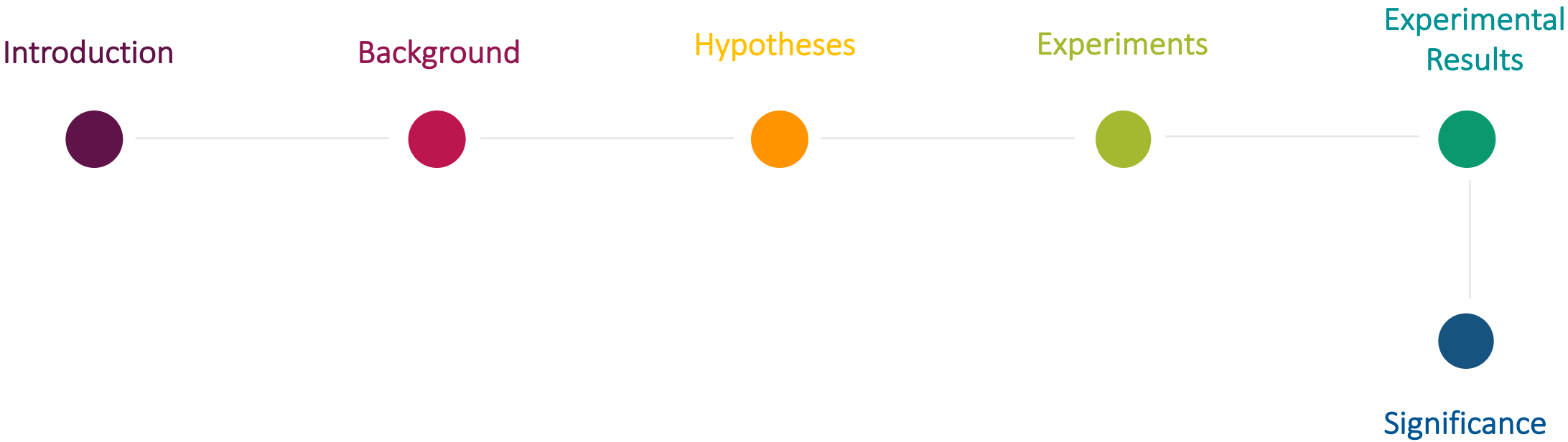
Conclusion

Hypothesis 1:	Neural Networks outperform Machine Learning Algorithms
Hypothesis 2:	LSTMs should perform better than CNNs in short texts
Hypothesis 3:	Temporal LSTMs help improve the classification accuracy of machine learning algorithms.
Hypothesis 4:	Structural Models produces similar results to LSTMs
Hypothesis 5:	DTW is a better distance metrics than Euclidean distance for KNN
Hypothesis 6:	Classification accuracy drops as the number of class labels increase

Conclusion

Hypothesis 1:	Neural Networks outperform Machine Learning Algorithms
Hypothesis 2:	LSTMs should perform better than CNNs in short texts
Hypothesis 3:	Temporal LSTMs help improve the classification accuracy of machine learning algorithms.
Hypothesis 4:	Structural Models produces similar results to LSTMs
Hypothesis 5:	DTW is not a better distance metrics than Euclidean distance for KNN
Hypothesis 6:	Classification accuracy increases as the number of class labels increase

Significance



Traditional feature engineering of Machine Learning Model

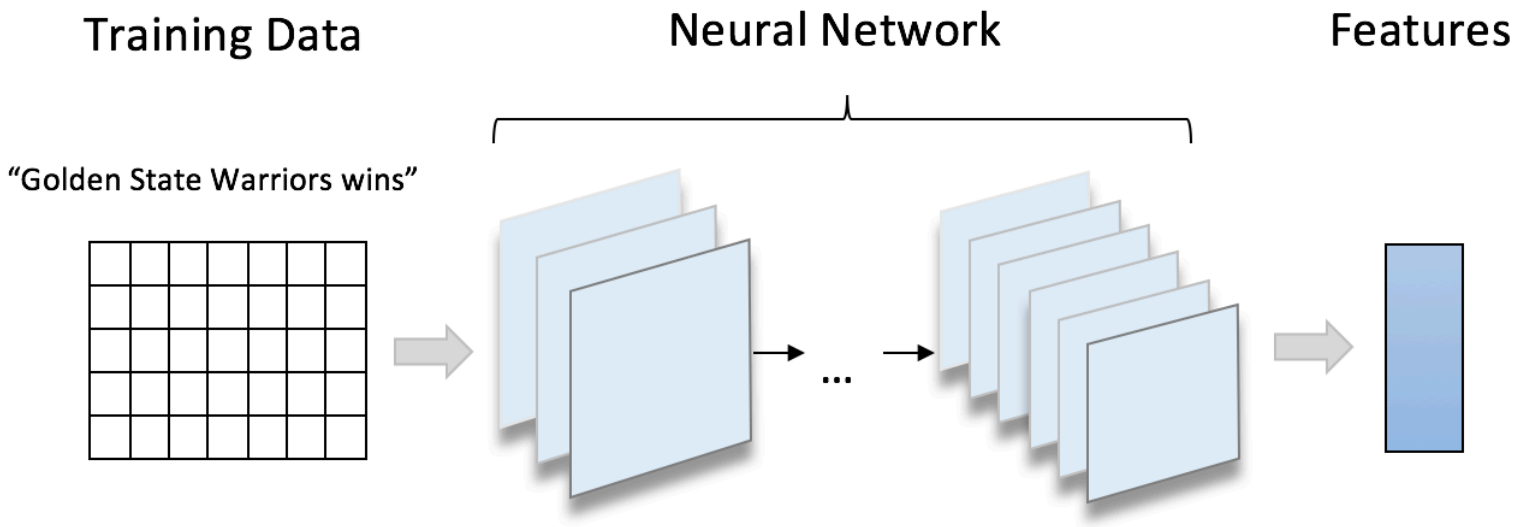
- Involves pre-defined dictionary as the first step
- Followed by many domain-specific steps, e.g. lemmatization, stemming e.tc..
- Require prior knowledge.

Traditional feature engineering of Machine Learning Model

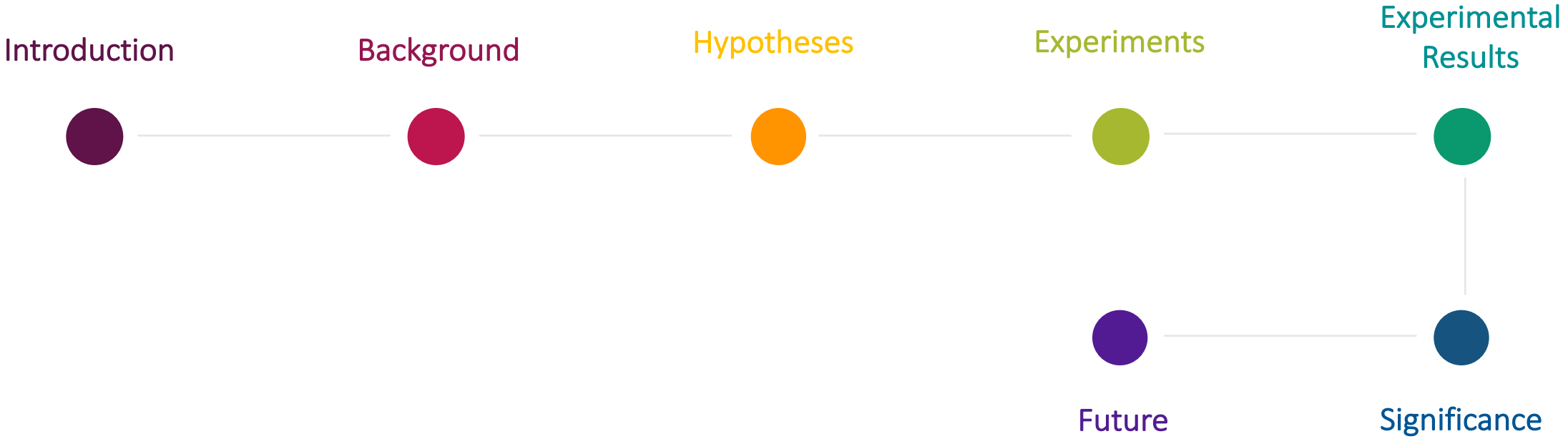
- Involves pre-defined dictionary as the first step
- Follows by many domain-specific steps, e.g. lemmatization, stemming e.tc.. Time-expensive
- Require prior knowledge. Labor-Intensive

“Zero cost” on feature engineering

- LSTMs capture feature representations in its hidden states
- Exhibits no previous knowledge
- Increase accuracy to up to 3.5X



Future


Introduction
Background
Hypotheses
Experiments
Experimental Results
Future
Significance

- Apply unsupervised learning techniques
  - How well temporal LSTMs have learnt and predict?
- Restricted Boltzman Machine as temporal networks
  - Further improve the training results with ML models
- Apply on short-sequential data e.g. speech signals, dialogs

THANK YOU!

Q&As

https://github.com/irisliu0616