

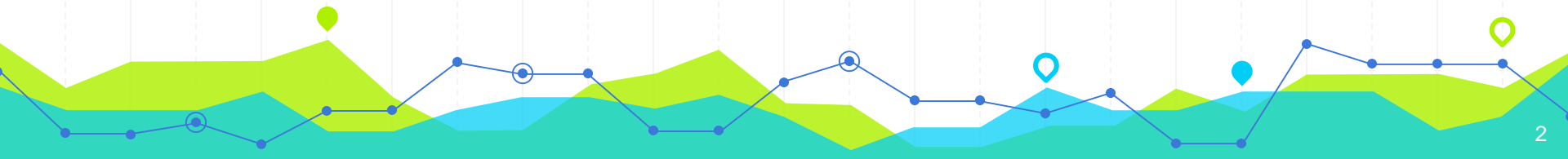
# Event-Driven Stock Prediction using NLP

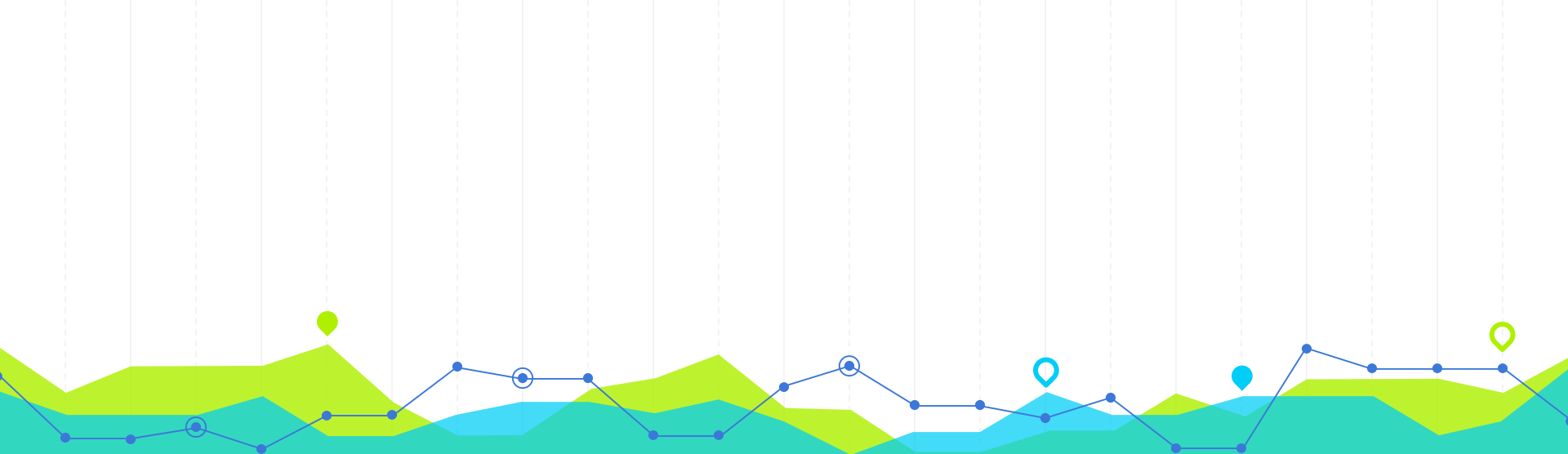
by Anant Jain, Ilya Yudkovich

CS 6120 Spring 2019 --- Prof. Lu Wang, Northeastern University

“

*The price of a security reflects all of the information available, and that everyone has a certain degree of access to the information [Efficient Market Hypothesis (EMH)]*





# What the heck?

..what does it mean?

# 1

It implies..

- Market is perfect.
- Market is predictable and it repeats itself.
- Peoples' views matter (the key source of market volatility)

In this project, we present a news surveillance system targeting financial news to predict its impact on some of the top market cap companies.



## Examples:

**Ticker:** AAPL.O

**News:** "'A U.S. federal judge has issued a preliminary ruling that Qualcomm Inc owes Apple Inc nearly \$1 billion in patent royalty rebate payments, though the decision is unlikely to result in Qualcomm writing a check to Apple because of other developments in the dispute.'"

**Prediction:** "Buy"

**Ticker:** AMZN.O

**News:** "'PARIS Casino's upmarket Monoprix supermarket chain is working to expand its partnership with E-commerce giant Amazon in France, following a successful launch in Paris, Monoprix's Chief Executive said on Thursday.'"

**Prediction:** "Strong Buy"

**Ticker:** MSFT.O

**News:** SAN FRANCISCO Some Microsoft Corp employees on Friday demanded that the company cancel a \$480 million hardware contract to supply the U.S. Army, with 94 workers signing a petition calling on the company to stop developing "any and all weapons technologies.'"

**Prediction:** "Sell"



# What has been done before?

..and what we are doing?!

# 2

## Related Work



**[Xie et al., 2013]**

Tree representations of information.



**[Bar-Haim et al., 2011]**

Identification of expert investors.



**[Kogan et al. 2009]**

Risk based on financial reports.



**[Siet et al., 2013]**

Sentiment analysis of news articles.



**[Luss et al., 2012]**

Word embeddings input and standard neural network prediction model.



**[Ding et al., 2014]**

Event embeddings input and convolutional neural network prediction model.

# What are we doing? What's new?

## [Basics of NLP]

- TF-IDF
- N-Grams
- Sentiment Analysis

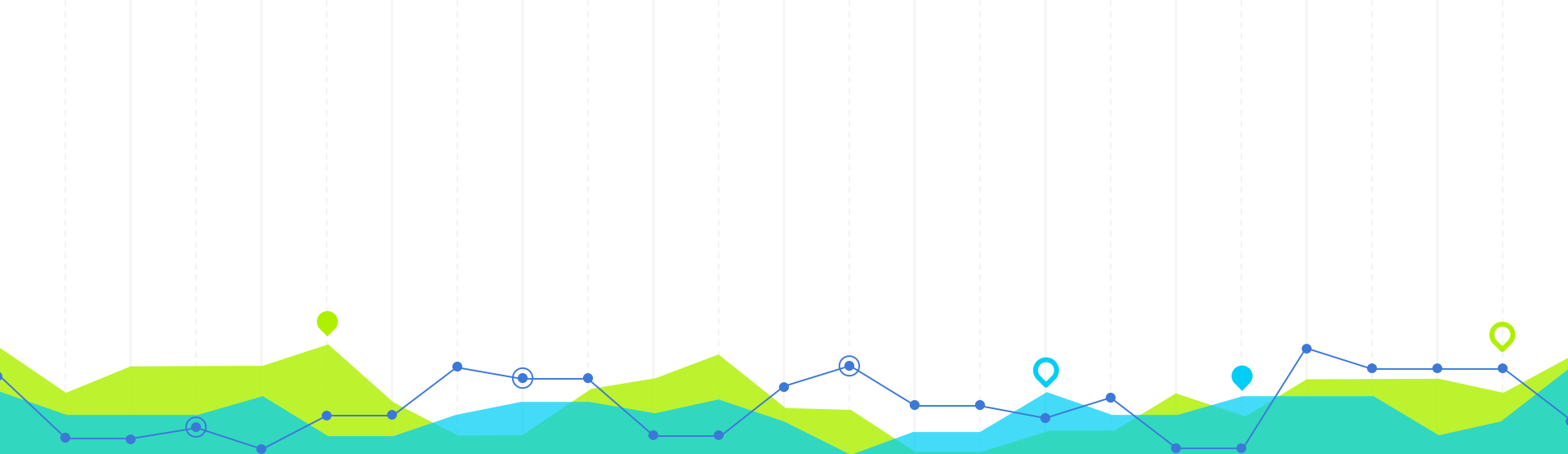
## [Traditional Deep NLP]

- GloVe Embeddings
- Bi-GRU & 1D CNN GM
- Initialization & Training

## [New Hotness]

- Transformers
- Fine-tuning
- BERT

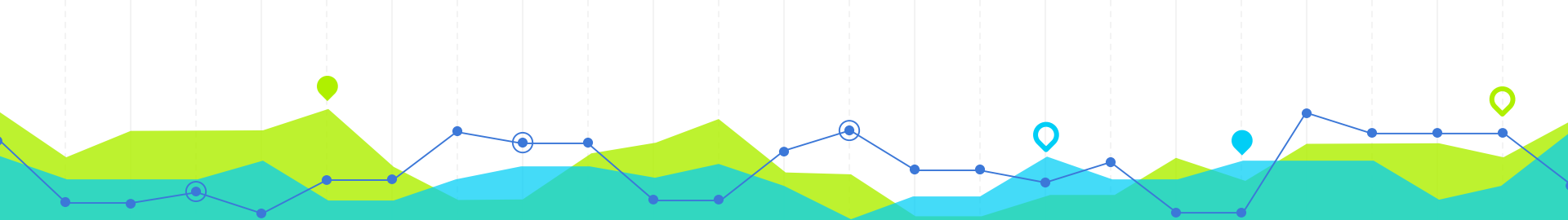




# Methodology

..how are we doing stuff?!

3



# 37%

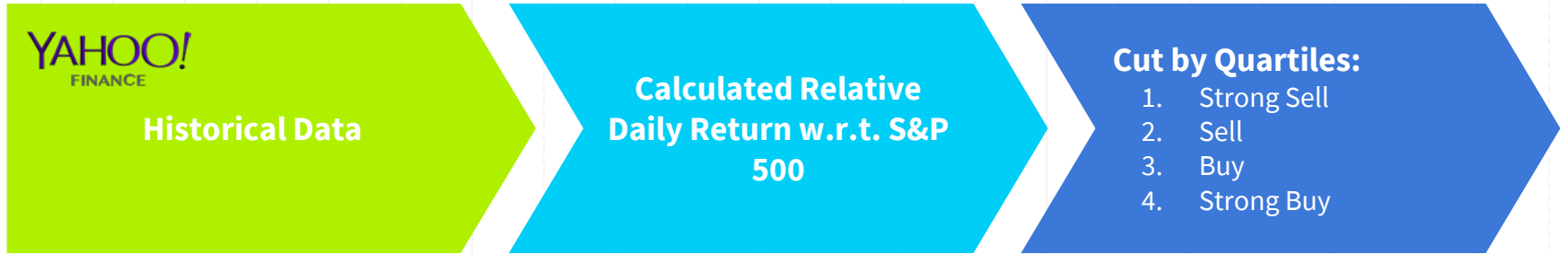
**The top market-cap companies.**

## Dataset Construction



2,590,000 pages      2,590 tickers

## Label Building Task



# Preprocessing

## Pruning

Extract top financial words which are abundantly used in financial articles.

Vocabulary ~ 20,000



Maximum Sequence Length ~ 512 - 1000.

## Padding & Trimming

Pad sequences to make the input consistent.



Dimensions ~ 300

## Event Embeddings

Use pre trained GloVe & BERT embeddings.

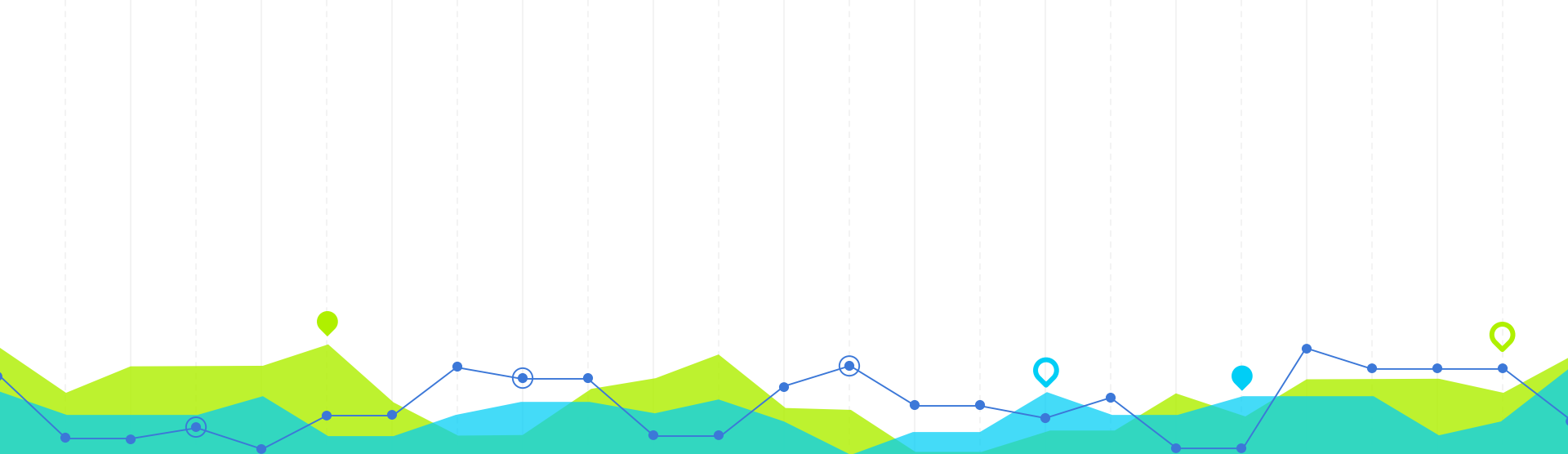


1D CNN GM, Bi-GRU, BERT.

## Model Selection

Feed the event tensors to deep learning models.





# Model Selection and Evaluation

..what worked and what did not?!

# 4

38,386 total news events

80% 30,637 training samples

20% 7,749 validation samples



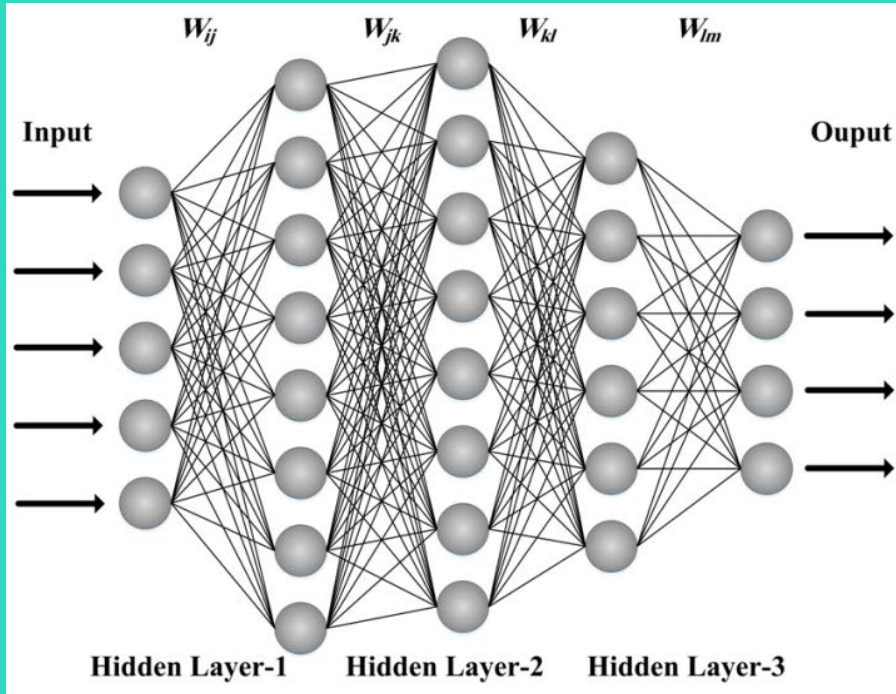
## Evaluation Metrics

- Accuracy: No problem as balanced classes!
- Matthews Correlation Coefficient (MCC):  
measures quality of the classification irrespective  
of class balance





# Model 1: Multilayer Perceptron



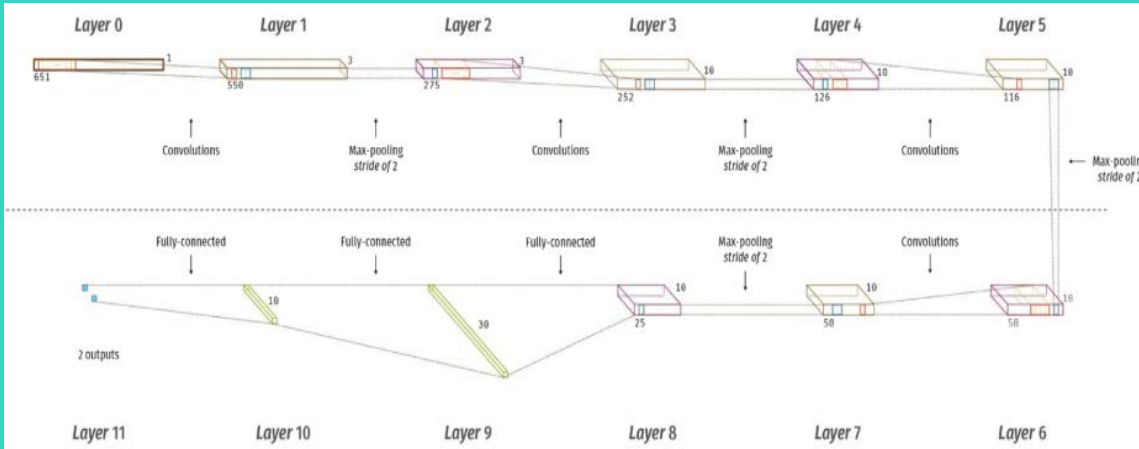
Activation: tanh  
Nodes: 256, 128, 64, 32, 16  
Solver: Adam

Measure	Value
Accuracy	27%
MCC	0.13965

## Model 2: 1D CNN with Global Maxpooling

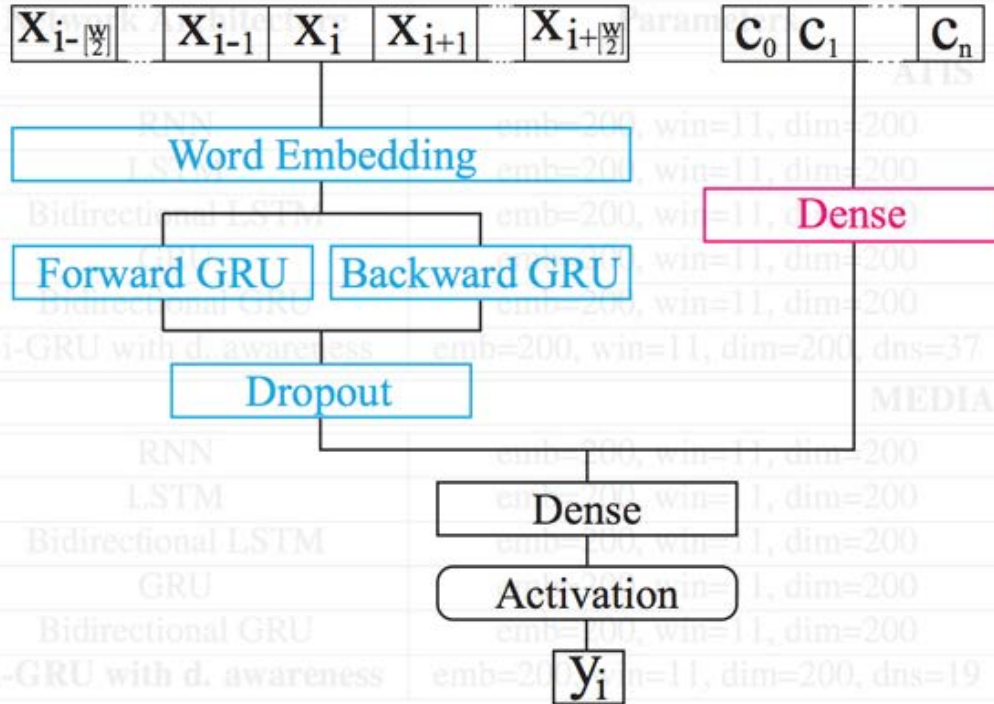
### Initialization,

- Pre-trained knowledge is in embeddings
- Everything else is 'from zero'
- Need lots of training examples



Measure	Value
Accuracy	64.43%
MCC	0.2918

## Model 3: Bidirectional-GRU



### Initialization,

- Pre-trained knowledge is in embeddings
- Everything else is 'from zero'
- Need lots of training examples

Measure	Value
Accuracy	63.34%
MCC	0.2111

# Model 4: Fine-tuning with BERT

Bidirectional Encoder Representations from *Transformers*

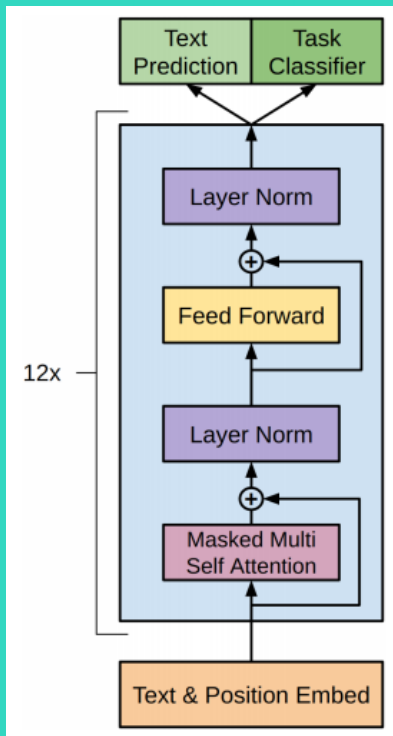
## BERT:

- Came out of Google (release 2018-10)
- It's full of transformers!

"**BERT**: Pre-training of Deep Bidirectional Transformers for Language Understanding" - Devlin *et al* (2018-10)

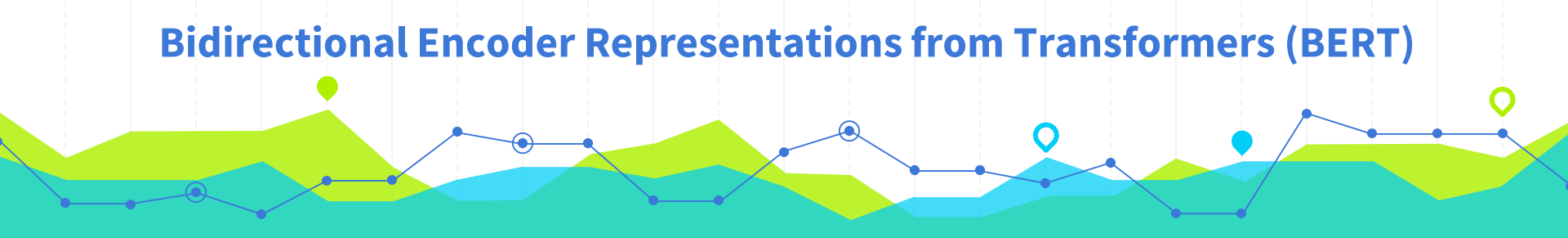
## Fine Tuning

- Take a model pre trained on huge corpus
- Do additional training on your data
- Learn actual task - using only a few examples



Measure	Value
Accuracy	66.43%
MCC	0.3010

# Bidirectional Encoder Representations from Transformers (BERT)



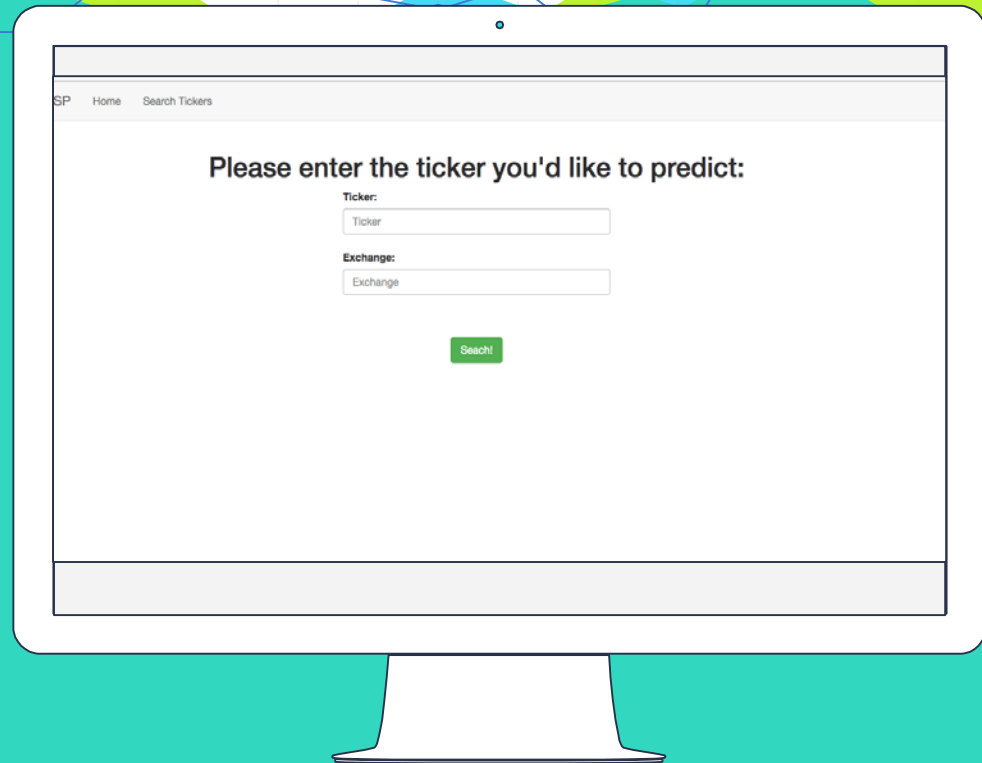
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>91.1</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>81.9</b>

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); BERT<sub>BASE</sub> = (L=12, H=768, A=12); BERT<sub>LARGE</sub> = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from <https://gluebenchmark.com/leaderboard> and <https://blog.openai.com/language-unsupervised/>.

## Comparing Results

Model	Accuracy	MCC
Luss and d'Aspremont [2012]	56.42%	0.0711
Ding et al.[2014]	65.08%	0.4357
1D-CNN GM (this project)	64.43%	0.2918
Bi-GRU (this project)	63.34%	0.2111
BERT (this project)	66.43%	0.3010

LIVE PROJECT



## Future work

- Better label: comparing the stock price of the company to the corresponding industry, instead of comparing everything with S&P 500
- Use data from other platforms (eg. Bloomberg)
- Experiment more with BERT (give it more time)



# THANKS!

## Any questions?

Anant Jain

Ilya Yudkovich

Team #: 12

