

NLP Text Analytics: AI Powered Sentiment Analysis

Teacher Retirement System of Texas

September 2022



Meeting Agenda

Creating Alpha Factors to inform Investment Decisions using Alternative Data



Alternative data

Text, Audio, and Video

The importance of alternative data in systematic quantitative finance investment decisions

Quant methods for text-based data

Natural Language Processing

An overview of the methods to find useful information inside text to assist investment decisions and a specific implementation of BERT.

TRS NLP Journey

Organizational considerations, MLOps, and Roadmap

Evolving organizational considerations from a POC to a vision for TRS Minimum Viable Product including MLOps.

Fundamental data vs alterative data; why alternative data is critical?

Example Corporation Income Statement Years ended December 31			
	(in thousands of dollars)		
	2021	2020	2019
Net sales	\$ 3,980	\$ 3,750	\$ 3,400
Cost of sales	3,100	2,950	2,700
Gross profit	880	800	700
Selling, general and administrative expenses	640	590	510
Operating income	240	210	190
Interest expense	20	15	15
Loss on sale of equipment	5	-	4
Income before income taxes	215	195	171
Income tax expense	50	40	36
Net income	\$ 165	\$ 155	\$ 135

See notes to the financial statements.

Verizon Communications Inc. (VZ) Q3 2018 Results - Earnings Call Transcript

Verizon Communications Inc. (NYSE:VZ) Q3 2018 Earnings Conference Call October 23, 2018 8:30 AM ET

Executives
Brady Connor - SVP, IR
Matthew Ellis - EVP and CFO

Operator

Good morning, and welcome to the Verizon Third Quarter 2018 Earnings Conference Call. At this time, all participants have been placed in a listen-only mode, and the floor will be open for questions following the presentation. [Operator Instructions] Today's conference is being recorded. If you have any objections you may disconnect at this time. It is now my pleasure to turn the call over to your host, Mr. Brady Connor, Senior Vice President, Investor Relations.

Brady Connor

Thanks, Brad. Good morning, and welcome to our third quarter earnings conference call. This is Brady Connor, and I am here with Matt Ellis, our Executive Vice President and Chief Financial Officer. As a reminder our earnings release, financial and operating information and the presentation slides are available on our investor relations website. A replay and transcript of this call will also be made available on our website.

Before I get started, I'd like to draw your attention to our Safe Harbor statement on slide 2. Information in this presentation contains statements about expected future events and financial results that are forward-looking and subject to risks and uncertainties. Discussion of factors that may affect future results is contained in Verizon's filings with the SEC which are available on our website. This presentation contains certain non-GAAP financial measures. Reconciliations of these non-GAAP measures to the most directly comparable GAAP measures are included in the financial materials on our website. The quarterly growth rates disclosed in our presentation slides and during our formal remarks are on a year-over-year basis unless otherwise noted as sequential.

Now let's take a look at consolidated earnings for the period. For the third quarter of 2018, we reported earnings of \$1.19 per share on a GAAP basis. These reported results include a few special items that I would like to walk you through. Our reported earnings include a net pre-tax loss of \$159 million primarily associated with the early debt redemption costs of \$476 million; acquisition and integration-related charges of \$137 million primarily pertaining to Oath; and a pension and benefit remeasurement credit of \$454 million.

The net impact after tax was approximately \$120 million or \$0.03 per share resulting in an adjusted earnings per share of \$1.22. Excluding the effect of these special items and the net effects of tax reform and the adoption of the revenue recognition standard, adjusted earnings per share was \$1.01 in the third quarter up 3.1% compared to \$0.98 a year ago. It has been three quarters since the adoption of the new accounting standard for revenue recognition. The effect of this change is illustrated within the table on slide 4.

As a reminder, it results in a reduction of wireless service revenue offset by an increase in wireless equipment revenue and the deferral of commission expense in both our wireless and wireline segments. The impact from this change has been fairly consistent during all three quarters of 2018 with a \$0.06 per share impact in the third quarter. We continue to expect the accretive benefit to full year earnings per share to be between \$0.27 and \$0.31.

The accretive benefit to operating income in 2018 is expected to moderate in 2019 and then become insignificant in 2020 as the timing impacts to revenues and commission costs converge. This will create year-over-year EPS pressure in both 2019 and 2020. For the remainder of this call, unless otherwise noted financial results will exclude the impact of this accounting change to provide clear comparability with prior periods.

How much data is created in last 2 years?

In the last two years alone, **90%** of the world's data has been created. 2.5 quintillion bytes of data is produced by humans every day. Most of this data is unstructured data.

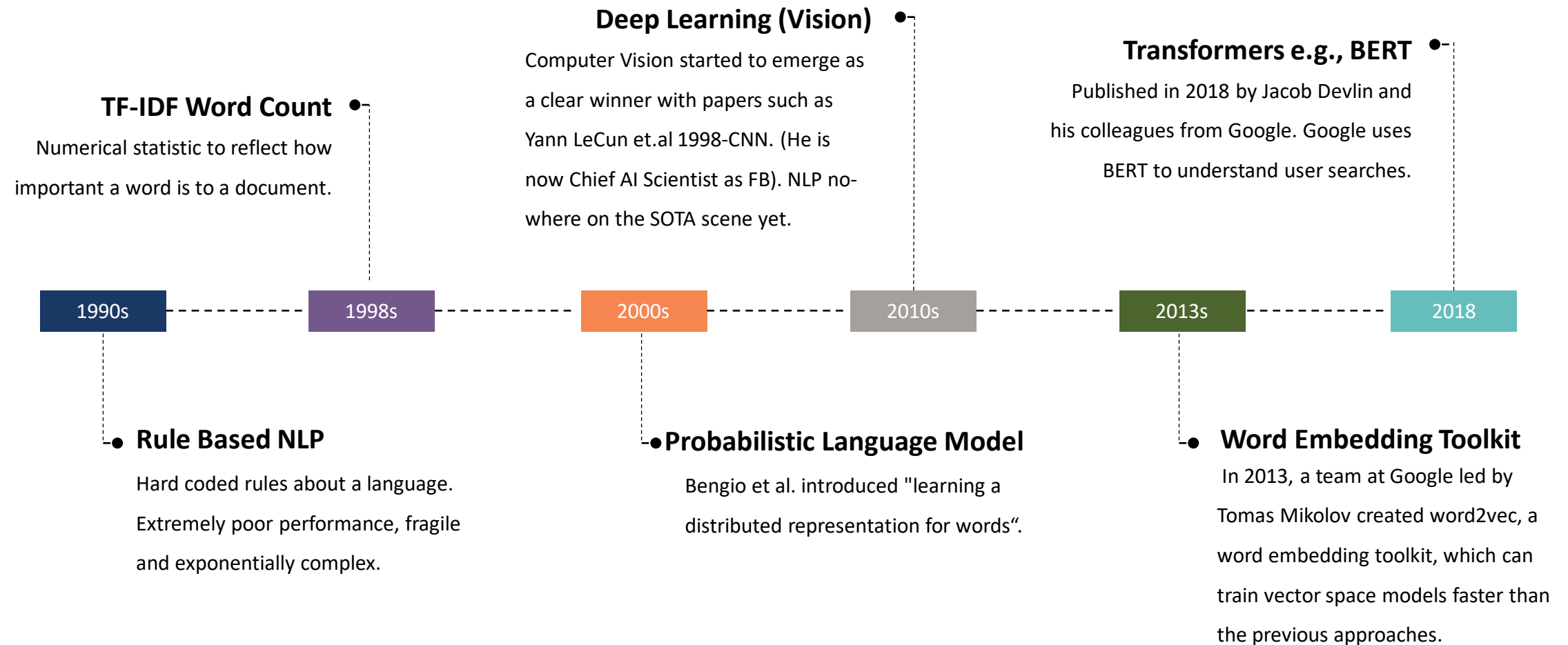
Most Primary Source data – videos, audios, text – is unstructured.

What data do investors care about?

Investors care about **Primary Source Data**. One of the sources of primary data is **Earnings Call**. It is well established that information conveyed on earnings call moves market. Material information is conveyed on these calls.

Natural Language Processing (NLP)

Major Events in the history of text processing



Google Search Example: <https://www.blog.google/products/search/search-language-understanding-bert/>

Why quants started word counting? Beginning of the NLP Journey.

Natural Language Processing: Field to quantify alternative text data.

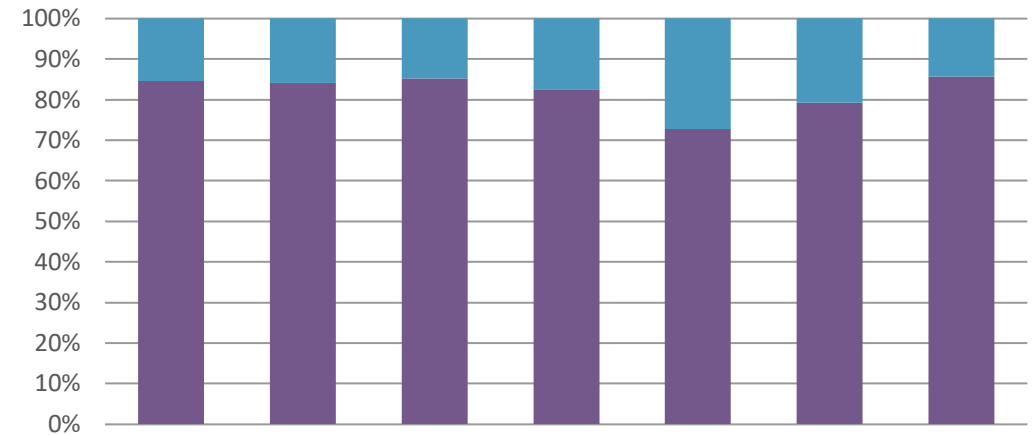
“I think from our perspective, we’ve always viewed this business as attractive in terms of its core business of ...”

- Seagate Technology Chairman & CEO Stephen J. Luczo

Loughran McDonald (S&P Global, Wolfe NLP Feed)

Every word is reduced to “root” word to be counted. Words such as “organization”, “organize”, “organized” are reduced to “organ”. Counts pre-defined positive and negative words. This approach is context blind.

Example of positive and negative word count



Time Series sentiment analysis of transcripts

+6%

Of all the words are more positive in company’s latest earnings transcripts than those in the previous period transcript.

■ **Percent positive words**

How positive is the Earnings transcript?

■ **Percent negative words**

How negative is the earnings transcript?

NLP Reduces Financial Text to Investment Factors

Calendar Q2 2017 Earnings Call. Conference Call, 26 April 2017 (96 words)

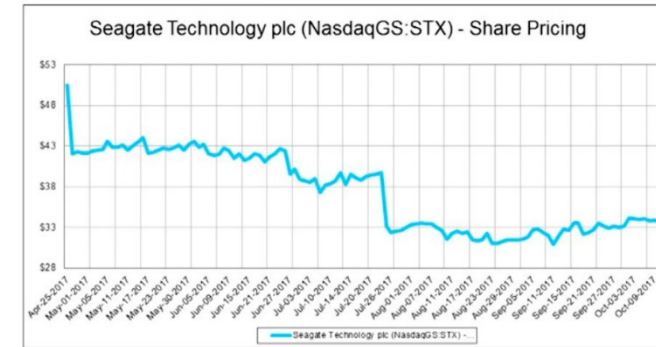
“I think from our perspective, we’ve always viewed this business as attractive in terms of its core business of selling into OEMs as well as servicing cloud service providers at one level, but really the opportunity to, I think, as architectures evolve and different customer needs evolve, to have the capability to optimize the devices either at a device level, at the subsystem level or the systems level, and if you do not have the software capability to do that, you cannot take advantage of what we think would be potentially significantly long-term trend.”

- Seagate Technology Chairman & CEO Stephen J. Luczo

Are things going well?

Yes, and we don’t mind talking about it in a concise manner.

Seagate Performance
-33% April 26 – October 10, 2017



Source: S&P Capital IQ platform as of 10/10/17 (shares are for illustrative purposes only)

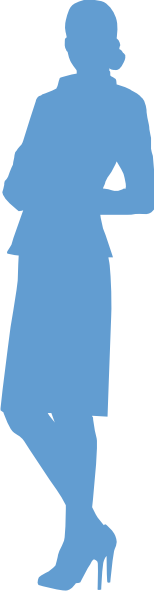
Are things NOT going well?

Yes, but things are not going well because X happened and caused Y then after. It will get well again after A, B, C is in place. Our strategic reports indicate the change is imminent.

TRS NLP Journey

Organizational Considerations

Driven by the “new” business needs



Legal & Compliance

- Goes through the dataset licenses to ensure L&C approval.
- **Licenses:** CCA 4.0, various BSD and MIT license.

Data Engineer

- Gathers, cleans, process, and prepares raw data.
- **Technology:** Python or ETL pipeline data engineer.

Data Scientist

- Manages predictive model algorithms on datasets in order to make business decisions.
- **Technology:** Spark, Sklearn, Tensorflow, Pytorch, MLFlow

MLOps

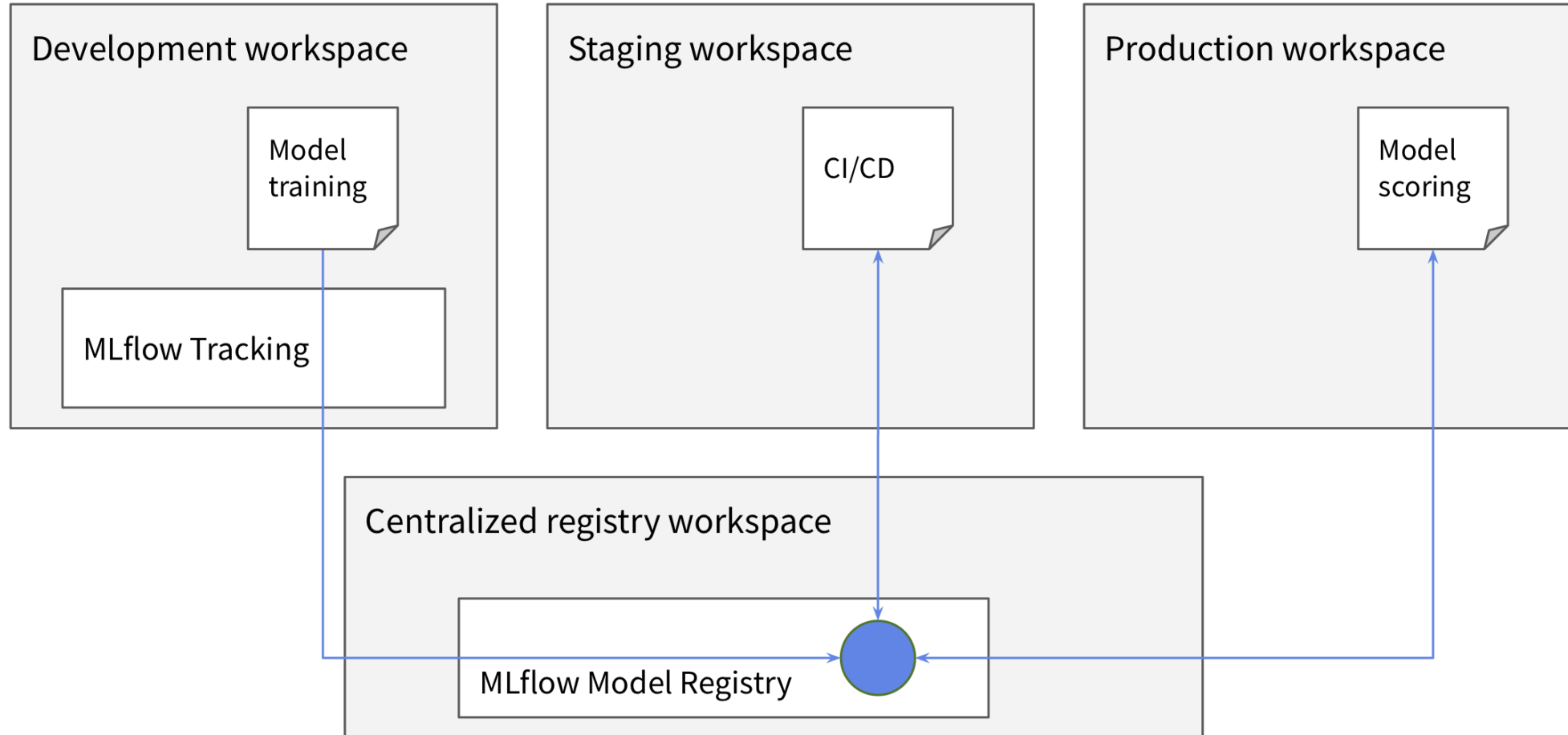
- Responsible for deploying machine learning models for Inference.
- **Technology:** Python, Pipeline ETL DataOps, MLFlow, Docker, Azure ML

TRS NLP Journey

TRS Roadmap:

- completed PoC
- Results brought to investment management committee.
- Industry moving in this direction. So, the agency is making modest investments in this direction.
- Additionally, agency is making modest investments in skill and technology.

Vision for scalable MLOps across Workspaces



POC FinBERT Results

Transcript NLP Project

Earning transcripts data from Capital IQ (CIQ)

Pilot Project

- Data Ingestion Complete
- Compute Requested
- Alpha Testing, next.

Transcripts NLP

- Topic Classification Model next.
- Focused on Fundamental Analysts

Filings – 10K, 8K

- Ensemble of NLP and Merton Model

PD – Default Prob.

- Chinese BERT
- Other EM, EAFEC Languages

Other Languages

Transcripts Project

Transcripts Project Roadmap

BERT Transcripts Model

- Investment Universe – **MSCI USA 650**
- Input Factors –
 - Sentiments
 - Number of Filings
 - Readability
 - Event Abnormal Returns
- Model Building –
 - 5-year rolling window to predict 1 months ahead stock returns.

BERT Transcripts Model

- Investment Universe – **Russell 3000 index**
- Input Factors –
 - Sentiments
 - Number of Filings
 - Readability
 - Event Abnormal Returns
- Model Building –
 - 5-year rolling window to predict 1 months ahead stock returns.

Dictionary based NLP vs Word Embedding NLP

Loughran McDonald dictionary is used by S&P Global, Wolfe, other financial NLP data vendors.

Dictionary Based

L&M Dictionary is easy to communicate but is not aware of context.

Dictionary-based models are easy to communicate with word counts.

However, count based models are not accurate as “apple” and “Apple” are one word.

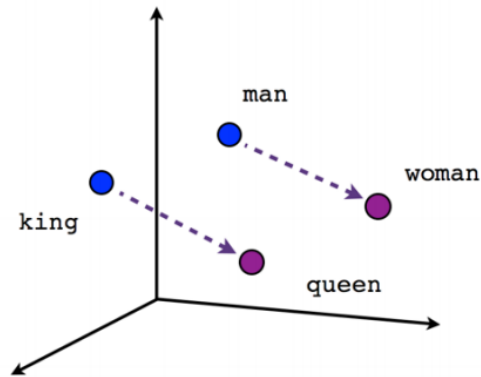
Word Embeddings

(Financial) words are mapped to numeric vectors in high dimension space.

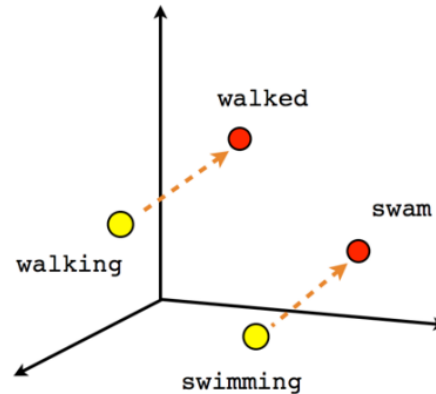
Context based model attends to prefix, suffix, noun-pronouns, and financial keywords in financial domain.

Why word embeddings are more accurate than count of words?

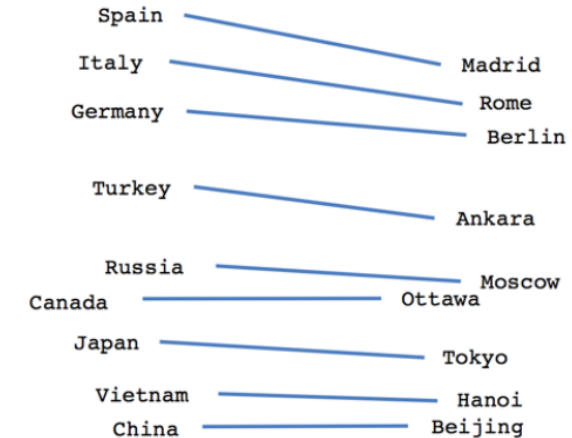
Word Embeddings



Male-Female



Verb tense



Country-Capital

Loughran McDonald (S&P Global, Wolfe NLP Feed)

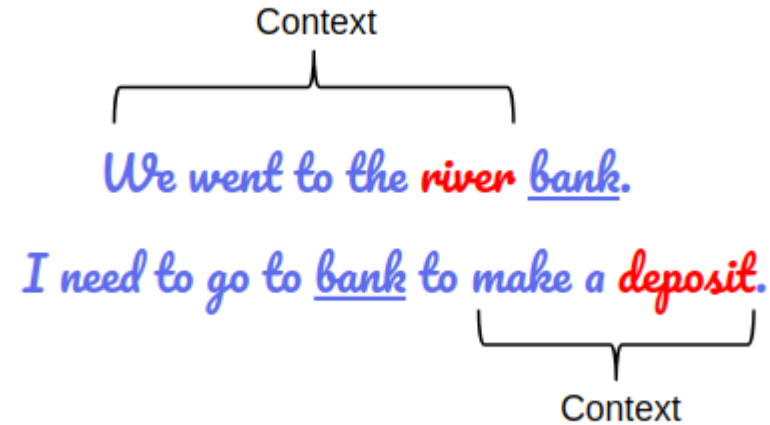
Every word is reduced to “root” word to be counted. Words such as “organization”, “organize”, “organized” are reduced to “organ”. Counts pre-defined positive and negative words. This approach is context blind.

Word Embeddings such as Word2Vec

One-to-one vector representation for every word. This is a very powerful concept as it allows vector math on word representations. However, it has a big shortcoming. A financial institution (“Bank”) has same word vector as a river “bank” as each word has single lookup.

How can we make better contextual predictions?

BERT adds context to word embeddings



BERT Embeddings

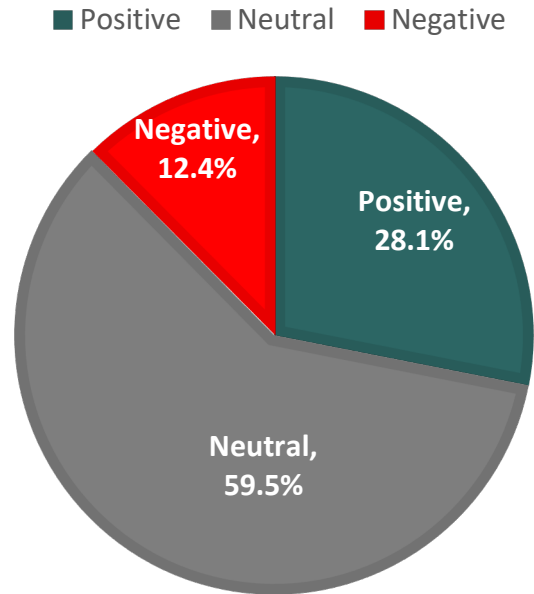
BERT solves the problem of earlier neural networks where each word got one vector representation. With BERT, each word will have a different vector representation based on its spellings and its context.

River “Bank” vs Financial institution “Bank” vs Power “Bank”

When we turn the word “bank: into a vector, the resulting vector will be based on the word and its context. This gives us a context aware model for any NLP task such as sentiment analysis.

Benchmark Dataset – Evaluation Metrics

Financial PhraseBank from Malo et al. 2014



Financial Phrasebank consists of 4,845 English sentences selected randomly from financial news found on LexisNexis database. These sentences then were annotated by 16 people with background in finance and business. The annotators were asked to give labels according to how they think the information in the sentence might affect the mentioned company stock price.

Finetuning examples

Operating profit totalled EUR 21.1 mn , up from EUR 18.6 mn in 2007 , representing 9.7 % of net sales .@positive

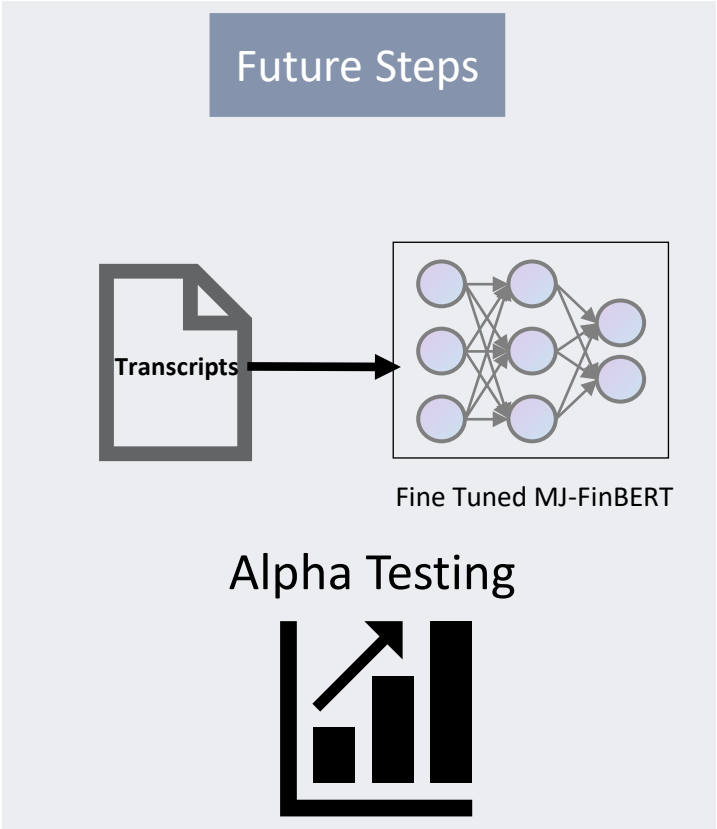
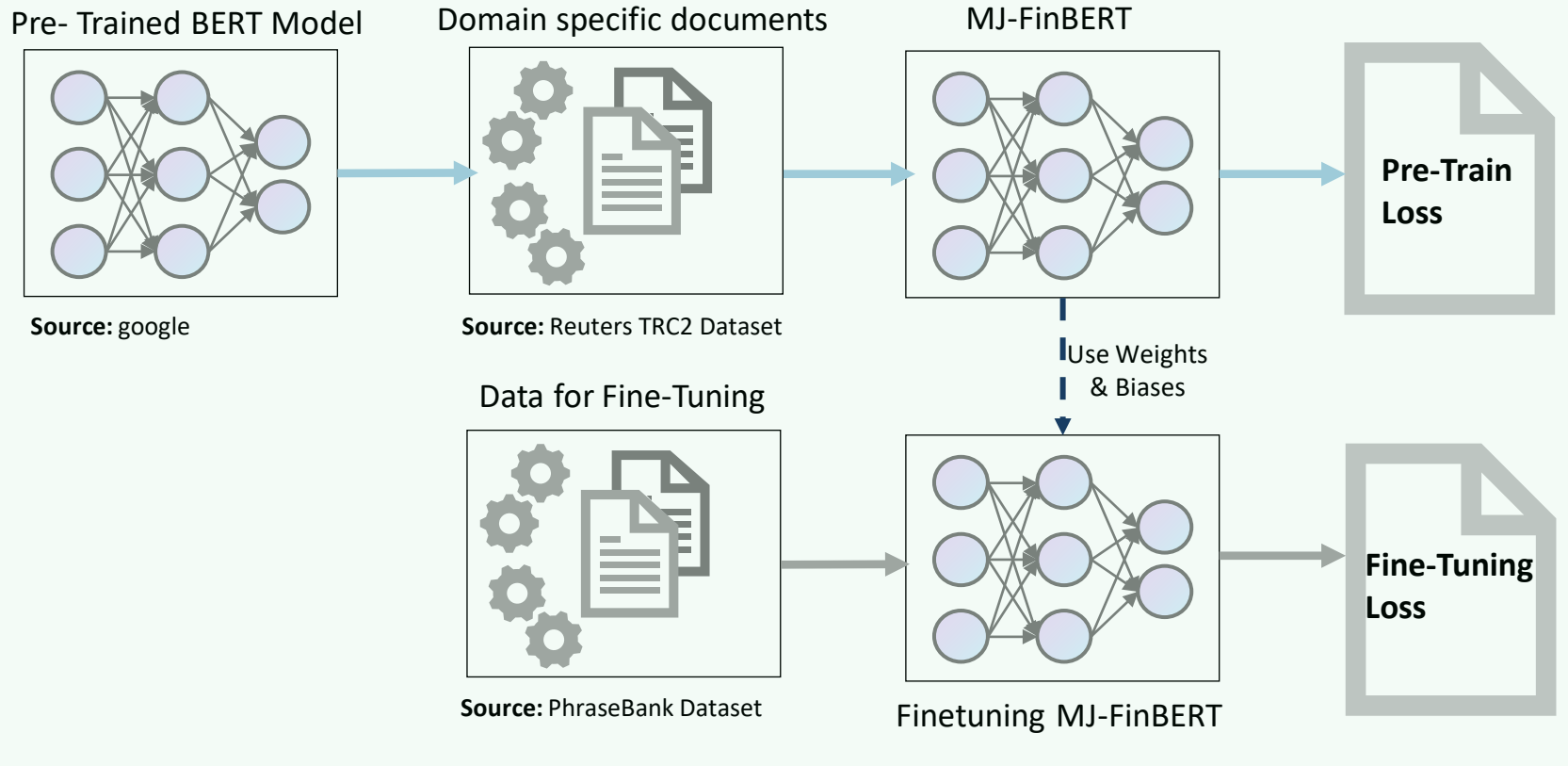
Jan. 6 -- Ford is struggling in the face of slowing truck and SUV sales and a surfeit of up-to-date , gotta-have cars .@negative

The terms and conditions of the year 2003 stock option scheme were published in a stock exchange release on 31 March 2003 .@neutral

PhraseBank Dataset source: <https://arxiv.org/pdf/1307.5336.pdf>

Model Setup - BERT Pre-training and Fine-tuning

MJ-FinBERT Model



Sources: google BERT model - <https://github.com/google-research/bert> ; Reuters TRC2 Dataset - <https://trec.nist.gov/data/reuters/reuters.html> ; PhraseBank Dataset - <https://arxiv.org/pdf/1307.5336.pdf> ; Earnings Transcripts – Capital IQ; Code – <https://github.com/huggingface/transformers> ;

Comparison of Count word model vs word Embedding model

Pos, Neg, Neu Sentiments Confusion Matrix

True label \ Predicted label	Positive	Neutral	Negative
Positive	True Positive 341 7.04%	Positives as Neut 900 18.57%	Positives as Negs 122 2.52%
Neutral	Neutrals as Pos 212 4.37%	True Neutrals 2428 50.10%	Neutrals as Neg 239 4.93%
Negative	Negatives as pos 14 0.29%	Negatives as Neuts 335 6.91%	True Negatives 255 5.26%

Accuracy=0.624

Pos, Neg, Neu Sentiments Confusion Matrix

True label \ Predicted label	Positive	Neutral	Negative
Positive	True Positive 1183 24.41%	Positives as Neut 169 3.49%	Positives as Negs 11 0.23%
Neutral	Neutrals as Pos 71 1.47%	True Neutrals 2780 57.37%	Neutrals as Neg 28 0.58%
Negative	Negatives as pos 17 0.35%	Negatives as Neuts 61 1.26%	True Negatives 526 10.85%

Accuracy=0.926

Loughran McDonald dictionary: word count model

MJ-FinBERT Results: Financial context aware model

APPENDIX

Trade off between accuracy and complexity

Unstructured data brings added complexity. Nevertheless, resources such as GPUs and deep learning researchers help.

Accuracy

Contextual Embedding Model

BERT and its successors improved scores on number of NLP tasks such as finding topics, sentiments, text search.

Each layer represents a high order representation of the text. For example, the first layer is a simple representation of sentence. But later layers are more complex and perform the task of missing word (MASK word task)

Complexity

GPU Compute Requirements

BERT inference needs neural network to compute context word vectors. So, it ideally needs GPU resources to run. However, GPUs are cheap these days.

It also adds complexity during training time, because it is learning the financial domain context.

Where we are now? Robo-surveillance shifts tone of CEO earnings calls

Source: <https://www.ft.com/content/ca086139-8a0f-4d36-a39d-409339227832>; Published on December 5, 2020

Artificial intelligence [+ Add to myFT](#)

Robo-surveillance shifts tone of CEO earnings calls

Trading algorithms leave a mark with deeper focus on the spoken word



Hedge funds use natural language processing to scour earnings calls, social media posts and regulatory documents for market-moving clues © FT illustration

[Twitter](#) [Facebook](#) [LinkedIn](#) [Save](#)

Robin Wigglesworth in Oslo DECEMBER 5 2020 74

Financial Times Report

By Robin Wigglesworth

“Managers of firms with higher expected machine readership exhibit more positivity and excitement in their vocal tones, justifying the anecdotal evidence that managers increasingly seek professional coaching to improve their vocal performances along the quantifiable metrics,” the paper said.

Example: ‘Au revoir to our profitability’ versus ‘we recorded a loss’ reads better in count-based NLP model.

Why vendors such as S&P Global fail to evolve? Is it compute, really?

Reason cited by David Pope, MD of Quant Research S&P Global - "state-of-the-art approach is computationally intensive".

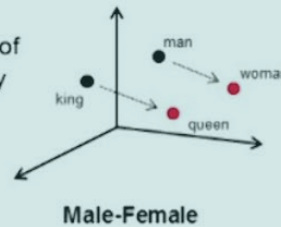
NLP Sentiment Approaches

Bag of Words – We score each word individually and tally the scores. This is the most common approach, and though primitive, works reasonably well. The 'bag of words' approach is computationally inexpensive which is important when processing a large number of transcripts. Our primer focused on the more common, computationally light approach to provide a proof statement for the transcripts data



N-Grams looks at groups of words. This is where negation can start to occur and one can score 'declining earnings' distinctly from 'declining costs'. This approach has its short comings too, as a number of words may separate 'declining' and its object.

Word Embedding approach looks at the frequency of occurrence of words in proximity to each other. This is an extremely computationally intensive approach.



S&P Global
Market Intelligence

10

#Annual2018



Our clients continue to pay big \$amt annually!

PPR: Proposed Project Request

WHO?

WHAT?

- **Background:** TRS MSG Quantitative Equity team needs a company earning transcripts natural language processing (NLP) daily feed. Our current vendors do not support this business' needs to process financial reports using deep learning NLP sentiment analysis. With Azure developer's release targeted 2021, there is an opportunity to build, deploy, and production support a solution on Azure platform.
- **Description:** This new feed will enable a new set of factors for the MSG Quantitative Equity team that extracts value from unstructured text transcripts and adds active alpha to Multi-Factor portfolios (~\$10B) managed under Mohan Balachandran's team. The daily feed will be batch-inference of machine learning NLP model (FinBERT) deployed on Azure compute services. If approved, this work would need to be balanced against existing Azure backlog to determine timing/scheduling.
- **Justification:** Natural Language Processing (NLP) via modern deep learning is promising to provide new alpha on to the MSG Multi-Factor portfolios (~\$10B). NLP deep learning models are being tested by MSG team as these models are expected to provide competitive advantages against older, limited analytical approaches. However, the delay from model research to full implementation in the Multi-Factor portfolios may be 2-3 months. In initial testing via a prior NLP POC PPR, the FinBERT model provided good results so the Quant team would like to explore the FinBERT factors on regular basis. Costs include running services and IT workhours for setup, monitoring, ML operations support and production support.

WHY NOW?

TRS NLP Journey



Completed POC

Results brought to investment management committee.

Additionally, agency is making modest investments in skill and technology.

Industry moving is this direction. So, the agency is making modest investments in this direction.



PHASE 1 - Machine Learning Feed On-Prem; subject to Infra & Info Sec Approvals

GPU to ingest data and upload Factor Feed based on a new Machine Learning Service Account with existing AD based RBAC
On-Prem solution for intermediate architecture. Next slide goes over Cloud Migration

1

DATA INGESTION

Ingest data from IMD
TRSINVSQCLAR



2

BATCH INFERENCE

GPU Batch Inference
NVIDIA V100 Compute
e.g., RHEL Box



3

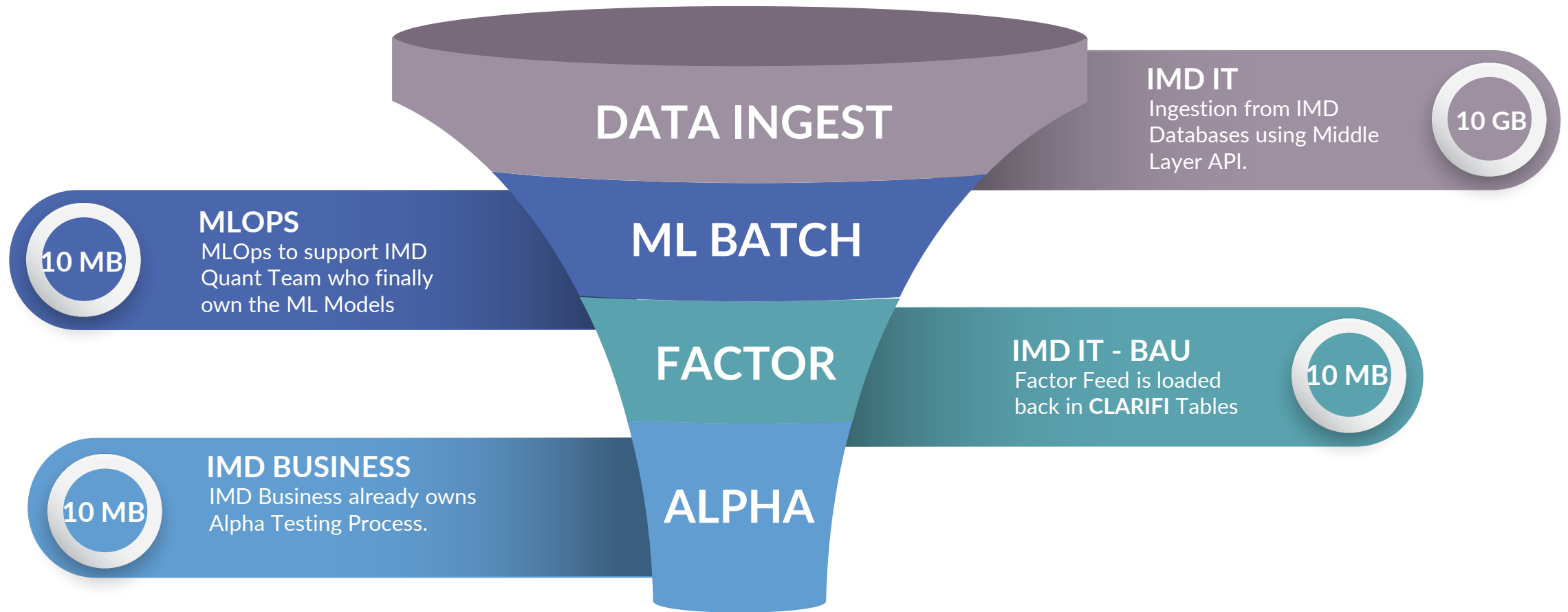
FACTOR FEED

Input for existing process
to load factors in
TRSINVSQLP01



PROCESS FLOW - OWNERSHIP

Data Size Reduces Top to Bottom



ML OPS – New*

Operationalize Machine Learning Assets such as Models, Parameters such as Batch Size etc. – Research Env to Staging to Production. Quality Assurance and Production Monitoring.

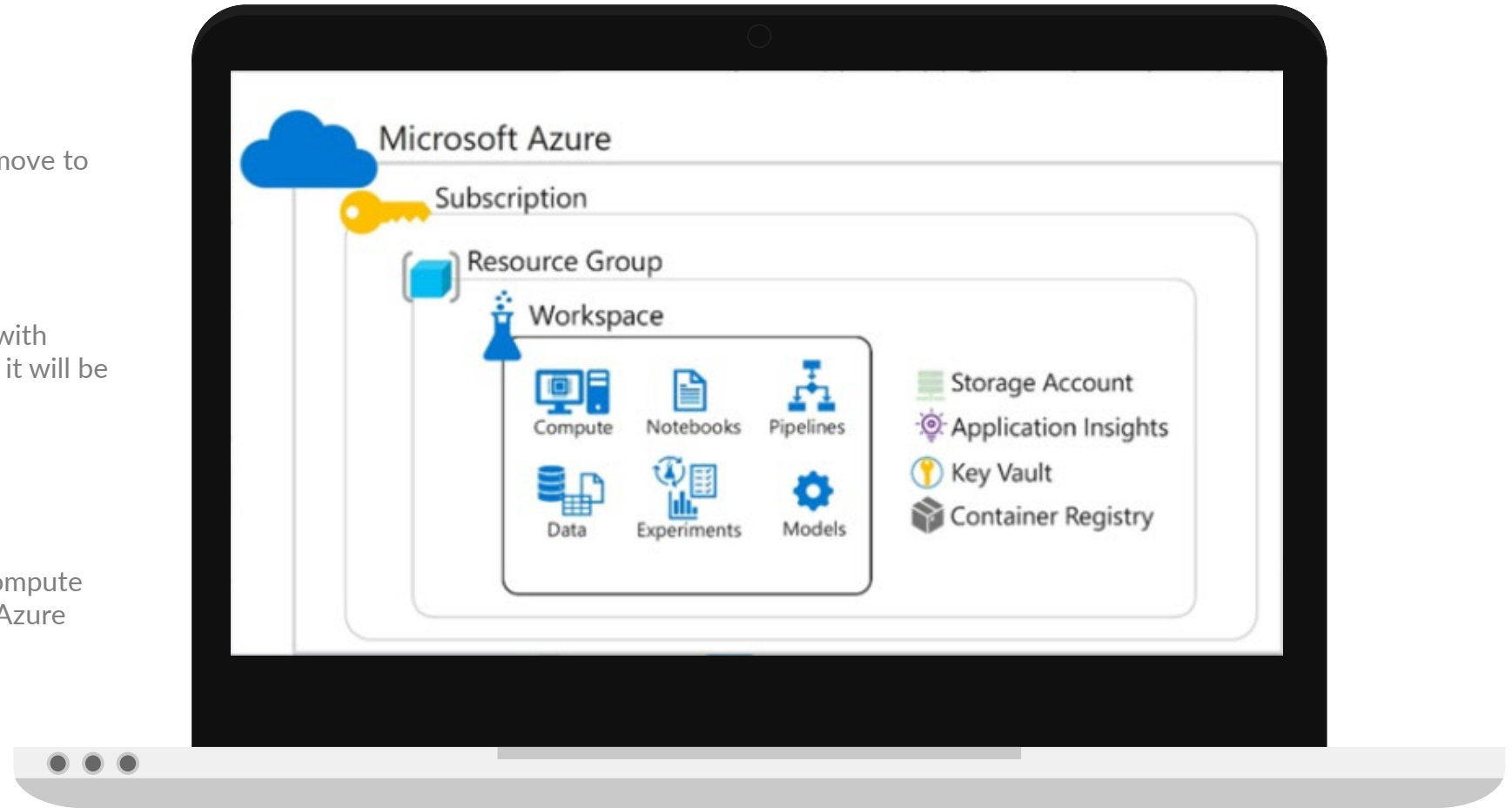
PHASE 2 – TARGET AZURE ARCH.

Phase 2 – Azure Components

✓ After Storage Strategy is implemented, storage will move to ADLS

✓ After **Secret Server Sync with KeyVault** is implemented, it will be incorporated.

✓ After Compute strategy is implemented, the GPU Compute will move to that such as Azure Container Instance



ML ACCESS CONTROL & PERMISSIONS

Access managed through Active Directory and RBAC

Permission	Owner	Contributor	Reader
Create workspace	X	X	
Share workspace	X		
Create compute target	X	X	
Attach compute target	X	X	
Attach data stores	X	X	
Run experiments	X	X	
View runs / metrics	X	X	X
Register model	X	X	
Create image	X	X	
Deploy web service	X	X	
View models / images	X	X	X

PHASE 2 – TARGET AZURE ARCH.

Phase 2 – Creating ML Workspace

✓ In the Microsoft Azure portal, create a new Machine Learning resource, specifying the subscription, resource group and workspace name.

✓ Use the Azure Machine Learning Python SDK to run code that creates a workspace.

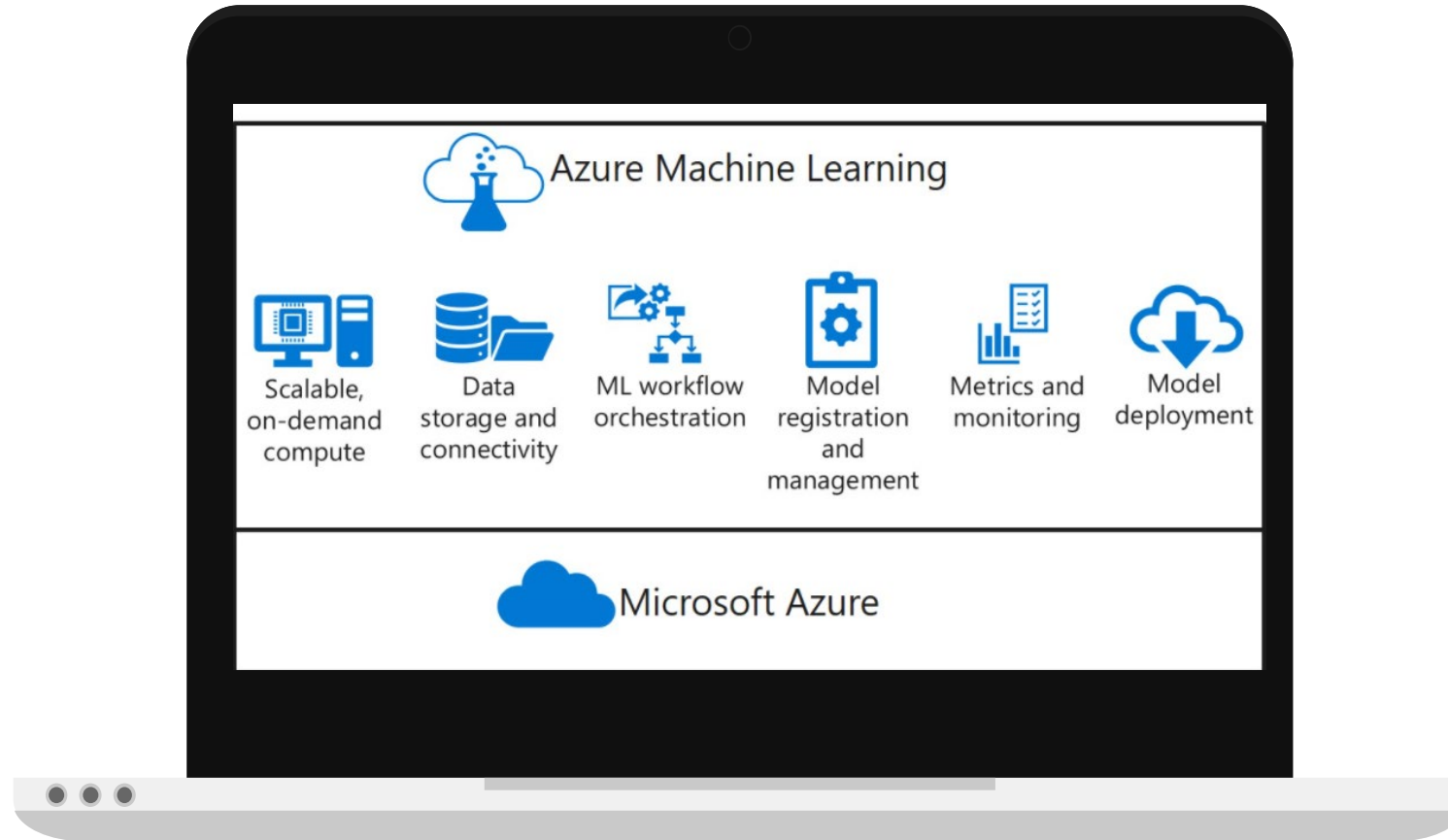
✓ Alternatively, use the Azure Command Line Interface (CLI) with the Azure Machine Learning CLI extension.

```
from azureml.core import Workspace

ws = Workspace.create(name='aml-workspace',
                    subscription_id='123456-abc-123...',
                    resource_group='aml-resources',
                    create_resource_group=True,
                    location='eastus'
                    )
```

PHASE 2 – TARGET AZURE ARCH.

Phase 2 – ML at Scale with Full MLOps



FUTURE STATE AZURE MACHINE LEARNING SERVICE

- Scalable Compute
- Data Storage connectivity to ingest data from say, snowflake.
- ML Workflow Automation to automate Training, Deployment, Management
- Model Registration, Management and Deployment

MLOps Tracking Changes b/w Phase-1 & 2

MLFlow to track experiments, audit runs etc.

```
from azureml.core import Experiment
import pandas as pd
import mlflow

# Set the MLflow tracking URI to the workspace
mlflow.set_tracking_uri(ws.get_mlflow_tracking_uri())

# Create an Azure ML experiment in your workspace
experiment = Experiment(workspace=ws, name='my-experiment')
mlflow.set_experiment(experiment.name)

# start the MLflow experiment
with mlflow.start_run():

    print("Starting experiment:", experiment.name)

    # load the data and count the rows
    data = pd.read_csv('data.csv')
    row_count = (len(data))

    # Log the row count
    mlflow.log_metric('observations', row_count)
```

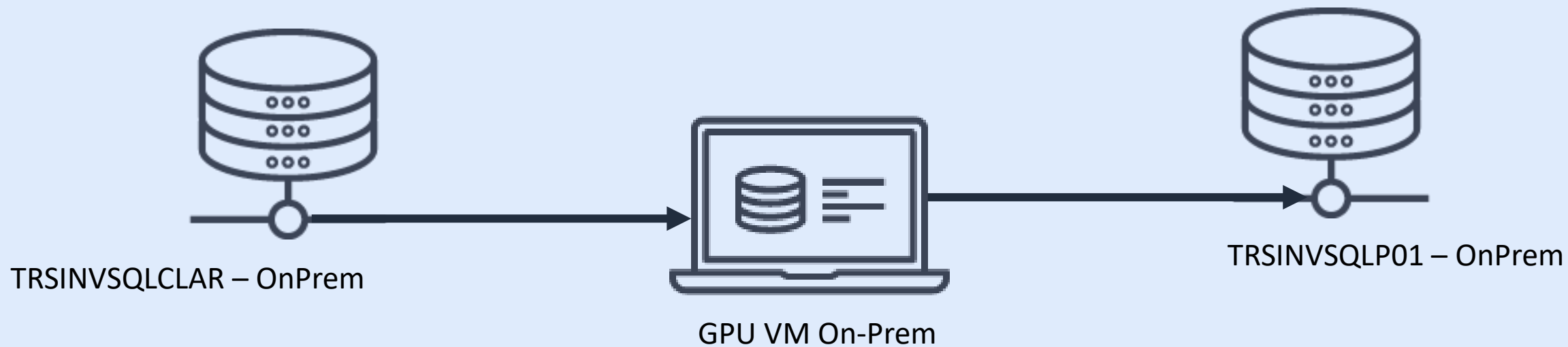
NEXT STEPS

Roadmap Document to include:

- * Training Session w/ IMD IT, if reqd.
- * MLOps Demo, if reqd.
- * Knowledge Transfer Business (e.g., Chris, Ryan, Sunny)
- * Finding and solving challenges for Infra, IT SEC and all stakeholders.

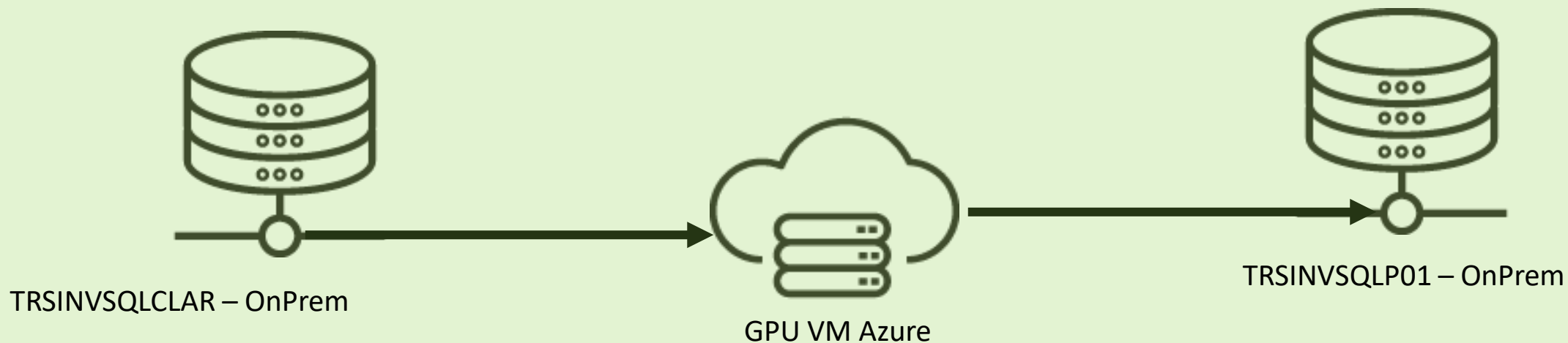
- Roadmap to include
 - Timelines - setting timelines based on prior project.
 - Leverage Technology - options for technology.. Azure or ON Prem
 - Project - monthly newsletter to all stakeholders where we are.. and Monitor Success and Communicate
 - Communication - Regular status reports to ED&A and anyone else you want me to keep in the loop.
 - Trainings - Lunch and Learn session to go over MLFlow MLOps usage for Zak's team to attend and get curious.
 - MLOps Gap - Design a MLOps system for the first steps into production.

On-Prem



Trade-Offs between the ML On-Prem vs ML on Azure solution. Storage, Connectivity, RBAC, Data Transfer, Workspaces etc.

Azure



Model Registry allows Easy Transitions

-  databricks
-  Home
-  Workspace
-  Recents
-  Data
-  Clusters
-  Jobs
-  Models
-  Search



Registered Models

Name	Latest Version	Staging	Production	Last Modified
Item_Recommender	Version 5	Version 5	Version 4	2019-10-11 15:30:02
Airline_Delay_Scikit	Version 3	–	Version 1	2019-10-11 12:41:43
Airline_Delay_SparkML	Version 5	Version 5	Version 3	2019-10-11 12:45:15
Transaction_Fraud_Classifier	Version 1	–	–	2019-10-11 15:18:05
Icon_GAN	Version 1	–	–	2019-10-12 08:20:12
Power_Forecasting_Model	Version 1	–	Version 1	2019-10-07 15:38:27
Product_Image_Classifier	Version 6	–	Version 5	2019-10-12 00:38:56
Comment_Summarizer	Version 3	Version 2	Version 3	2019-10-12 00:39:40
Movie_Recommender	Version 5	Version 5	Version 3	2019-10-10 14:07:07
Translation_Alpha	–	–	–	2019-10-11 16:45:01

RECOMMENDATIONS

On-Prem vs On Azure

Compute Options for Experiment Runs



Local Compute

- Compute where the control code for the experiment is running
- Often a development workstation or Azure Machine Learning compute instance



Compute Cluster

- Cloud-based cluster managed in an Azure Machine Learning workspace
- Starts, stops, and scales on-demand



Attached Compute

- Azure compute resource outside of a workspace
- For example:
 - Virtual Machine
 - Azure Databricks
 - Azure HDInsight

AUTO SCALING NOT A CURRENT REQMT.

Compute Clusters are not required for the initial 3-5 Machine Learning Use Cases. Auto Scaling will be required when we have a bigger ML Team and other projects.

INITIAL

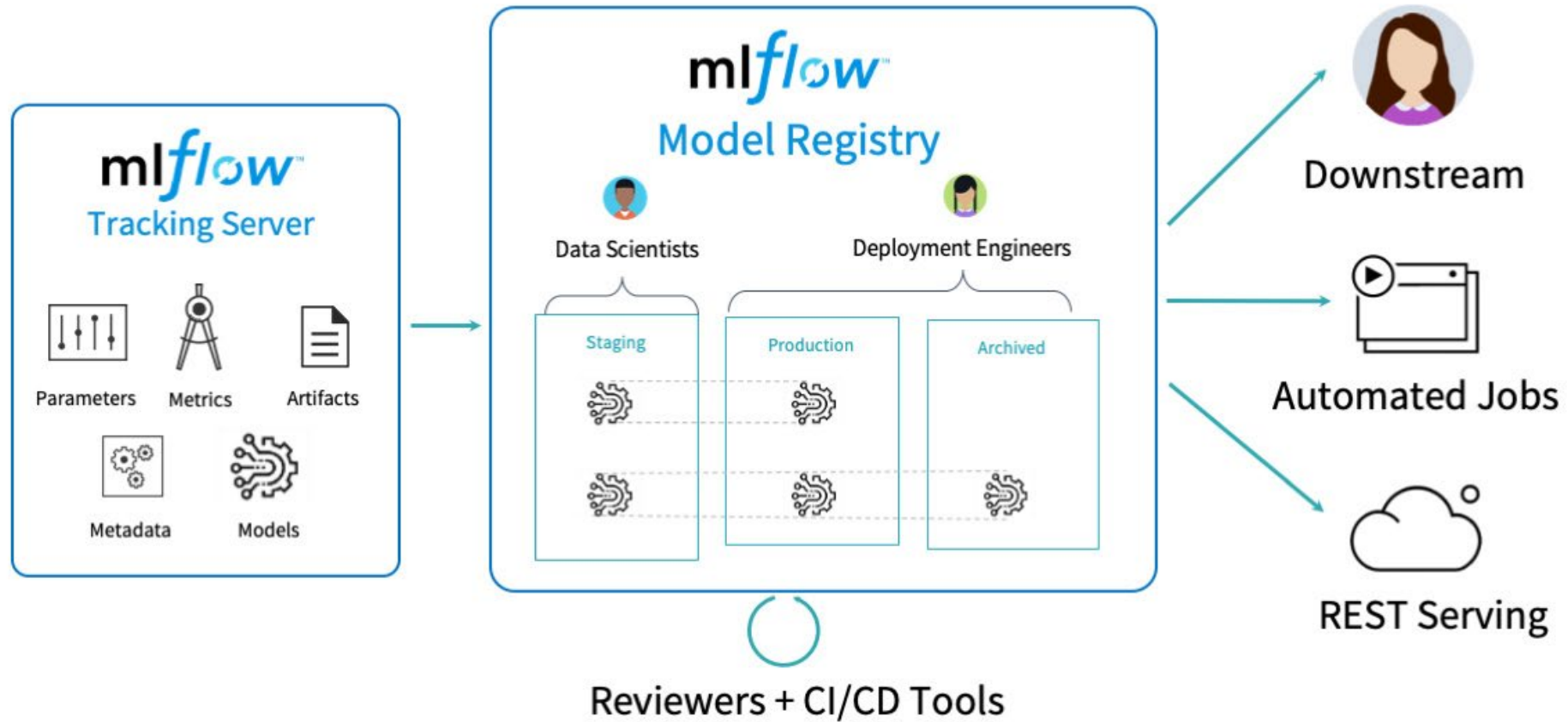
For Transcripts use case, 1300 Companies over 10 Years load is benchmarked at 48 hours of Compute time.

DAILY

For Transcripts use case, 1300 companies for daily load is benchmarked at under 1 minute Compute Time.

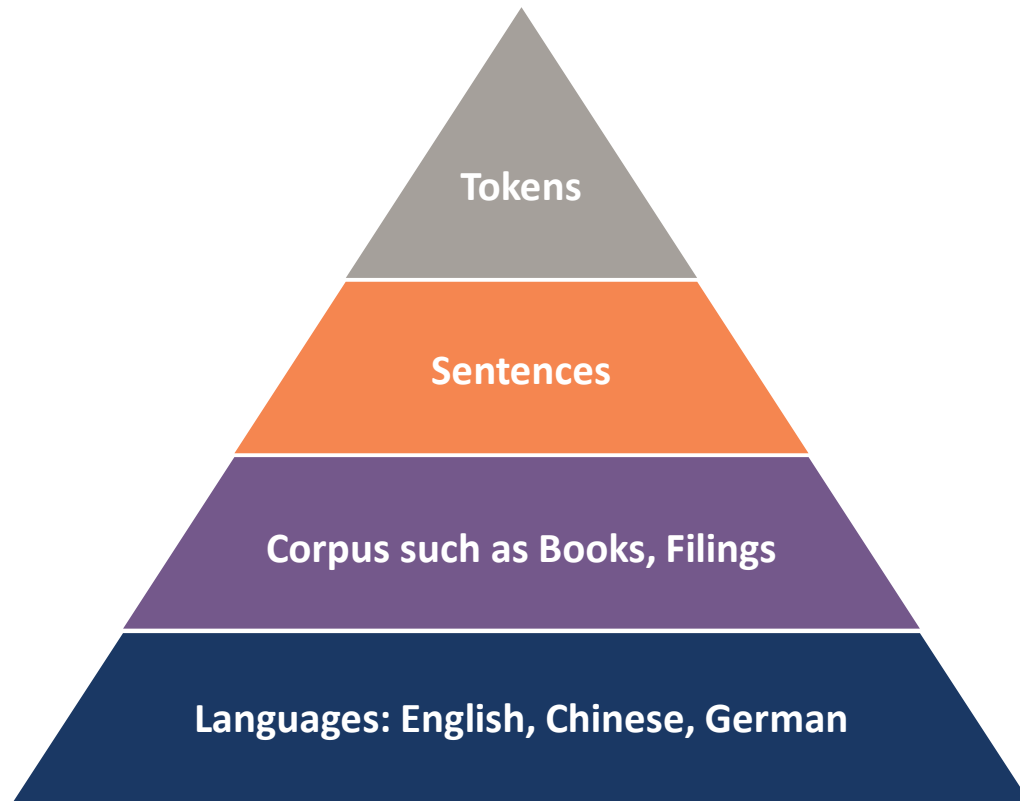
MLFlow to be used for MLOps

Experiment tracking, Model Management and Deployment



NLP: Natural Language Processing

What is NLP and what can you do with NLP?



Use Cases

NLP encompasses a variety of tasks, including:



Parts of Speech Tagging

Identifying Nouns, Verbs, Adjectives, etc.



Topic Modeling

Categorizing the documents into various topics such as Politics, Finance etc.



Sentence Classification

Classifying sentences into labels such as negative, positive, cashflow, revenue, etc.



Named Entity Recognition

Identify Person Names, Organizations, Locations, etc.

Additional benefits of Contextual Models

Systematic Quant Investment analysis sliced and diced by Geography, Sectors, Analysts, and Topic Keywords.



Aggregation by Geography – Similar context

Quantifying Language models of a specific geography such as Chinese for EM, others for EAFEC reveal trends of overall sentiment shift of a geography.



Aggregating by Sectors - Classification

Similarly, quantifying text data reveal the specific sector trend such as Airline Industry, Consumer Services, Materials, Retailers.



Aggregating by favorable analysts - NER

Playing Favorites paper by HBS Lauren Cohen and Dong Lou. Firms that call on more favorable analysts experience more negative future earnings surprises and more future earnings restatements.



Aggregating by Cash Flow, Distress - Topics

Categorizing the documents into various topics such as Free Cash Flow, Bankruptcy etc.

Natural Language Processing: Fast Evolving Field

Bidirectional Encoder Representations from Transformers



BERT was an improvement over these papers:

- Google ELMO: <https://arxiv.org/abs/1802.05365>
- OpenAI GPT: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- Google Research BERT Repo (fork) : <https://github.com/MohitJuneja/bert>



Shortcomings of Count Based Models

One can not infer much with counts of “words” without differentiating between “apple” and “Apple”.



Deep Learning & Transfer Learning

Context based language Models for multiple language tasks to avoid overfitting.



BERT and Other Transformer Models

Attention based mechanisms to attend to sub-words, co-references, and multiple meanings in that context.

How to avoid overfitting?

GLUE Scores solve for overfitting. NER models can't be good at NER. So, with transfer learning one evaluates tasks which aren't used in training.

2	CoLA	Is the sentence grammatical or ungrammatical?	This building is taller than that one. = Ungrammatical	Matthews
3	SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = .93056 (Very Positive)	Accuracy
4	MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = A Paraphrase	Accuracy / F1
5	STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = 4.6 (Very Similar)	Pearson / Spearman
6	QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = Not Similar	Accuracy / F1
7	MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = Contradiction	Accuracy
8	QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = Answerable	Accuracy
9	RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = Entailed	Accuracy

#14







GLUE Human Baseline at #14.

One might get confused how the models can outperform a human? After all, if one set of humans created the true labels then shouldn't another set of humans produce the same result? However, **performance varies widely** such as SAT/GMAT scores overtime and against peers.

What is GLUE Leaderboard? *General Language Understanding Evaluation* benchmark is a collection of datasets used for training, evaluating and analyzing NLP models relative to one another, with the goal of driving "research in development of **general and robust** natural language understanding systems". The collection consists of nine "difficult and diverse" task datasets designed to test a model's language understanding, and is crucial to understanding how transfer learning models like BERT are evaluated.

Source: https://docs.google.com/spreadsheets/d/1BrOdjJgky7FfeiwC_VDURZuRPUFUaz_jfczPPT35P00/edit#gid=70328292

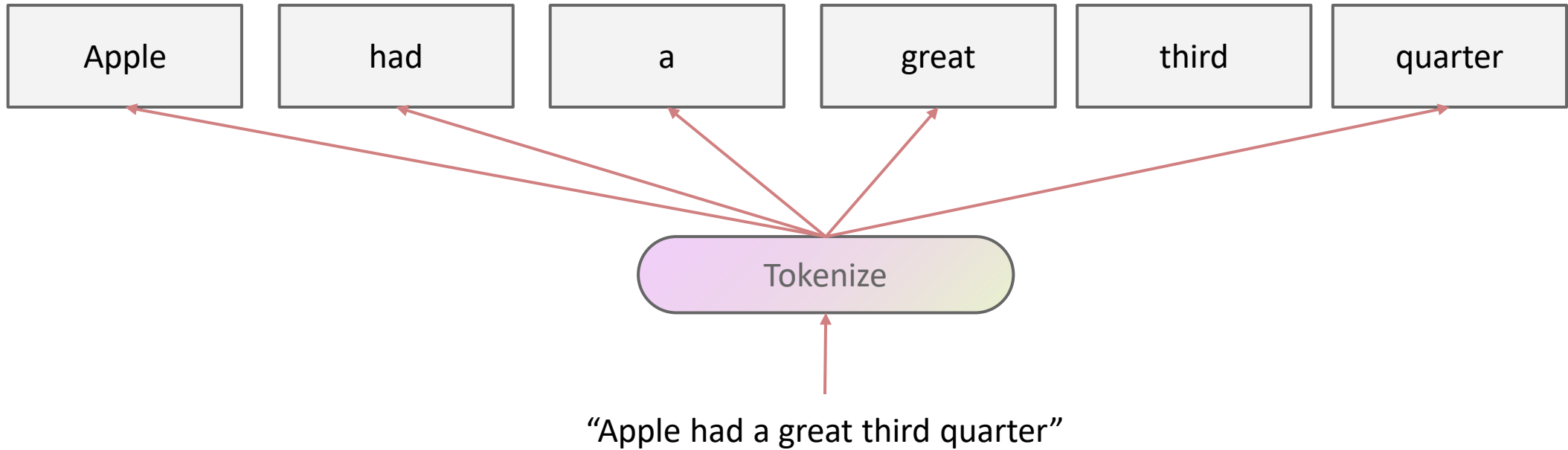
General Language GLUE Leaderboard – To avoid overfitting

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	M
1	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	
+ 2	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	
+ 3	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	
4	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	
5	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	
6	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART			89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	
+ 7	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	
+ 8	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	
+ 9	Huawei Noah's Ark Lab	NEZHA-Large		89.1	69.9	97.3	93.3/91.0	92.4/91.9	74.2/90.6	91.0	

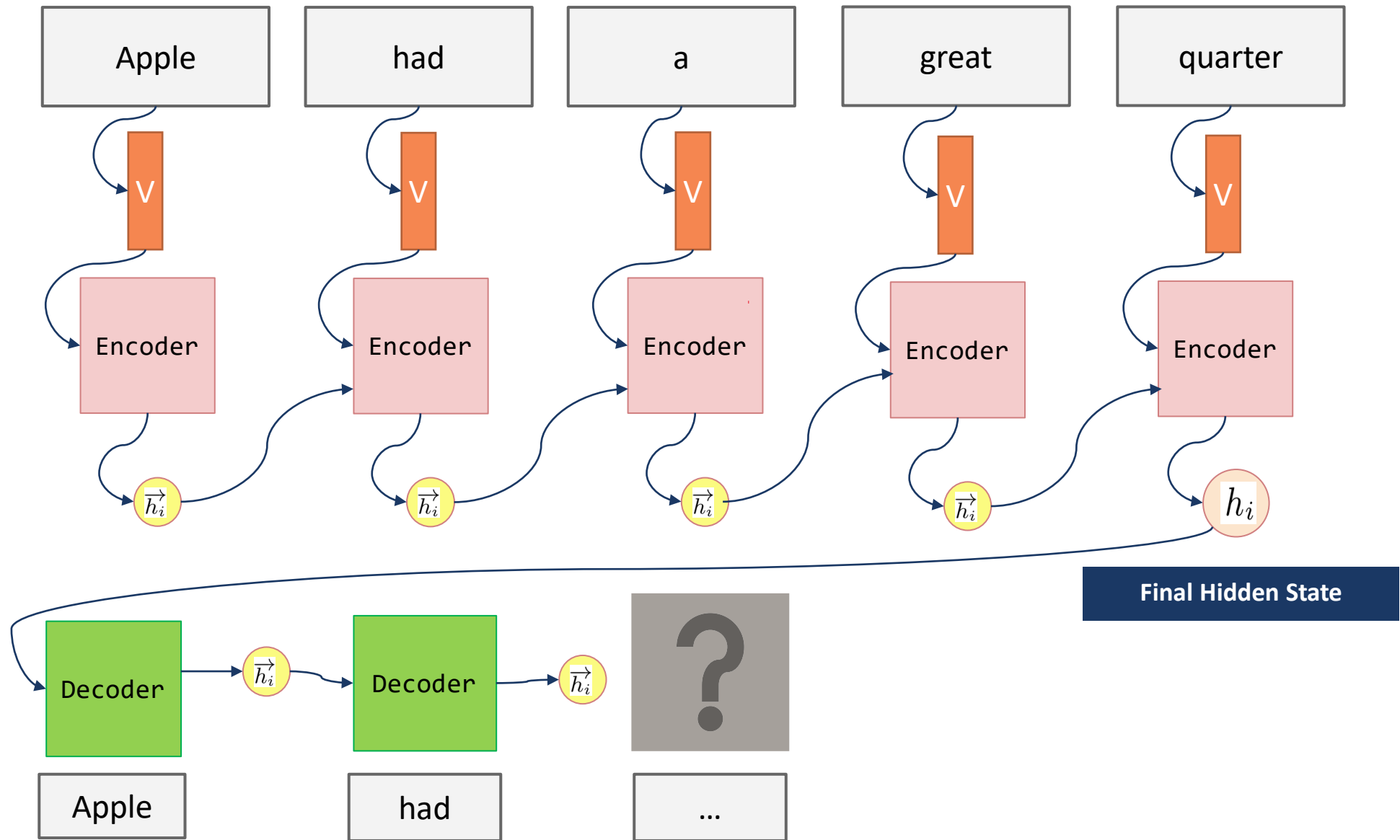
Metrics: GLUE Leaderboard [source: <https://gluebenchmark.com/leaderboard>]

Language and Neural Networks – Tokenization during pre-training

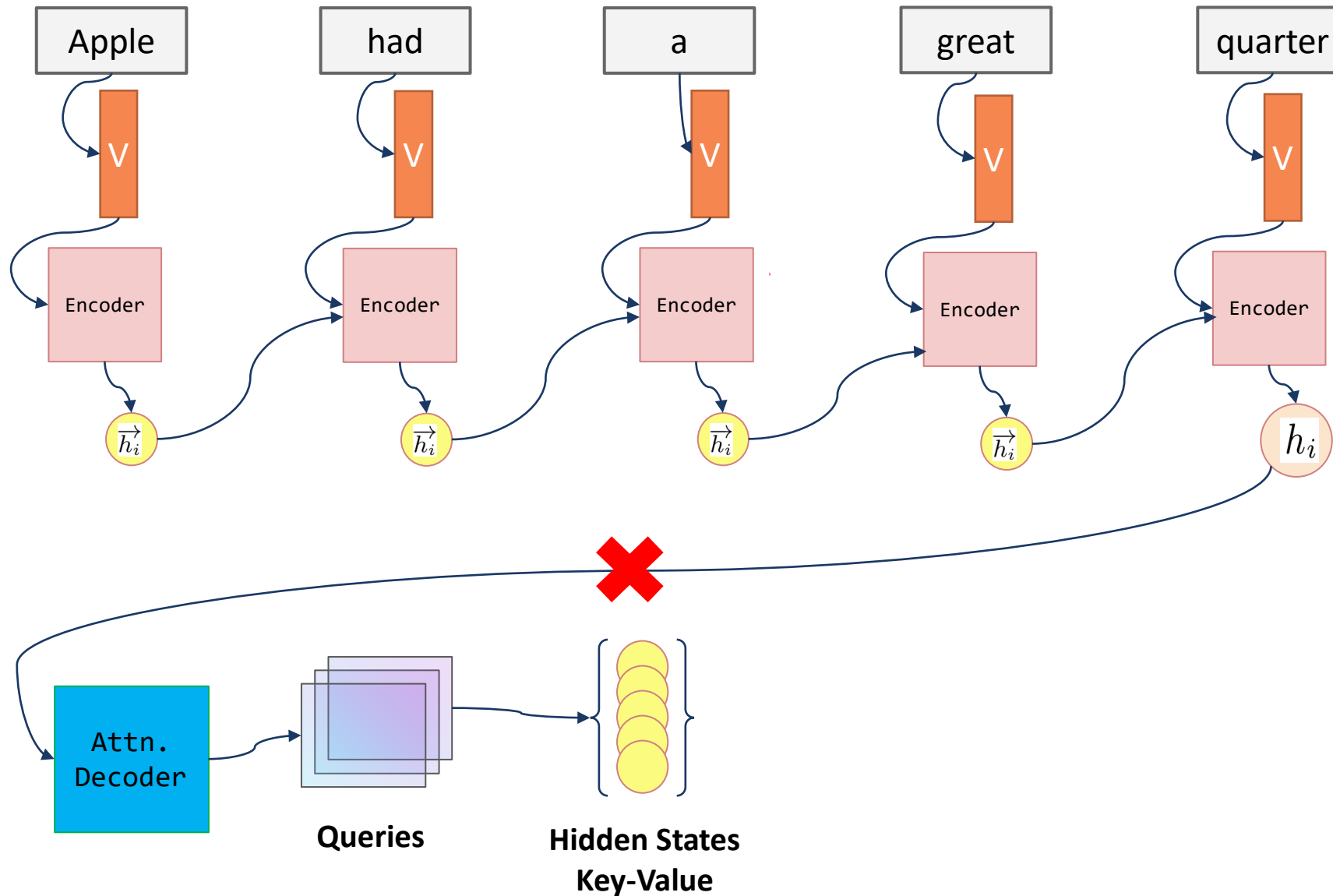
Recurrent Neural Networks (2010 to 2017)



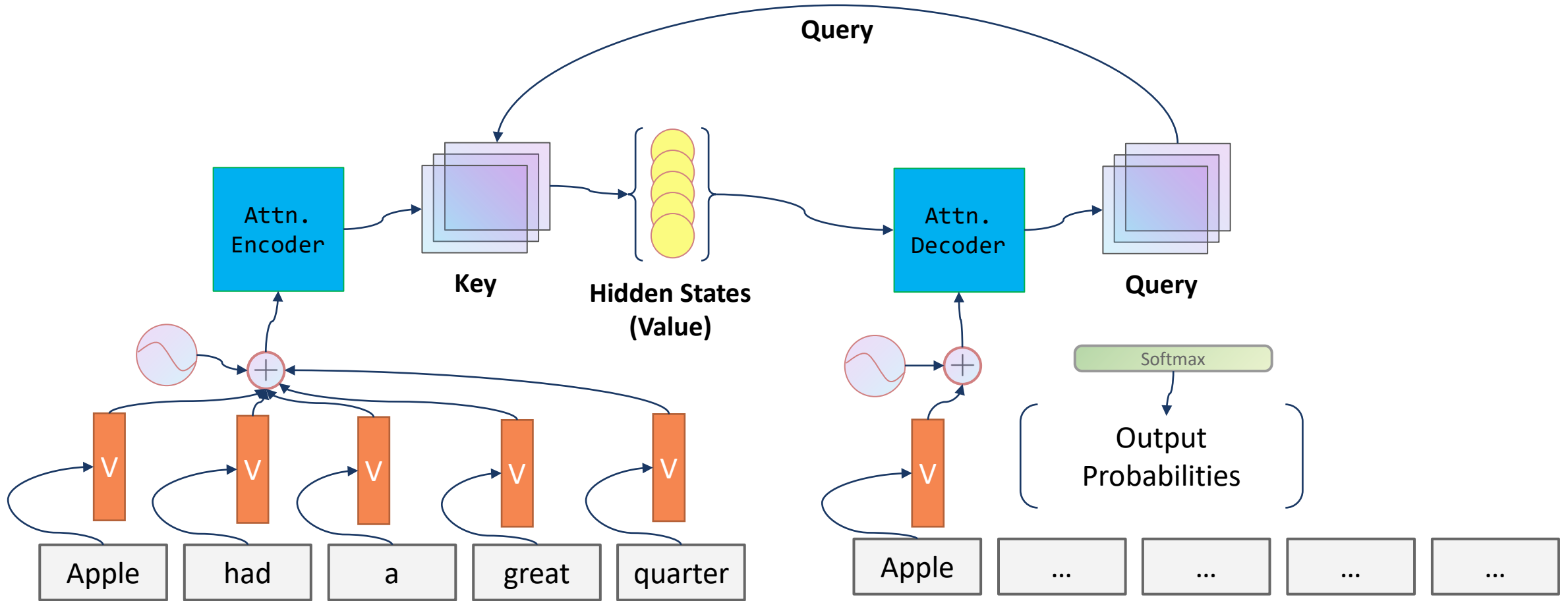
RNN Shortcomings – Slow and Small Sentences length; Not good for Long Range



Attention Decoder Idea – Attend to all hidden states.



Attention – Similar Changes to Encoder – Pre-training tasks: Masking and NSP



Language Understanding – Example for Masking and NSP

Next Sentence Prediction (NSP)

Sentence 1

[CLS] **Apple** had a great [MASK] quarter [SEP]



Sentence 2

They sold a record number of [MASK] phones [SEP]

Label

IsNext

Sentence 1

[CLS] Apple had a great [MASK] quarter [SEP]

Sentence 2

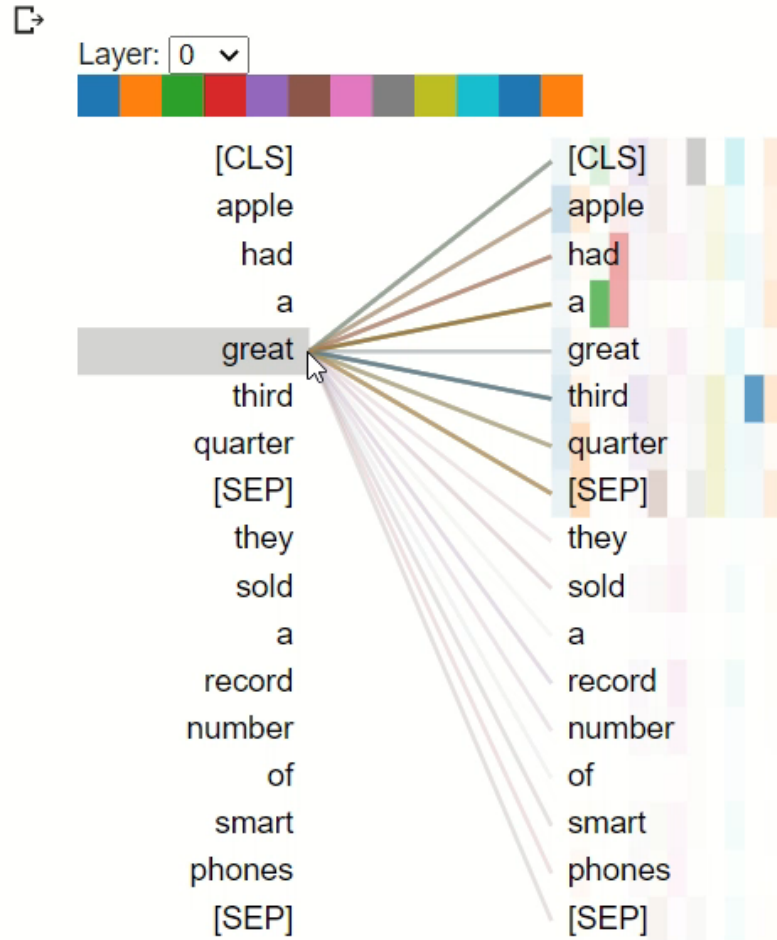
Many died during COVID since [MASK] of March [SEP]

Label

NotNext

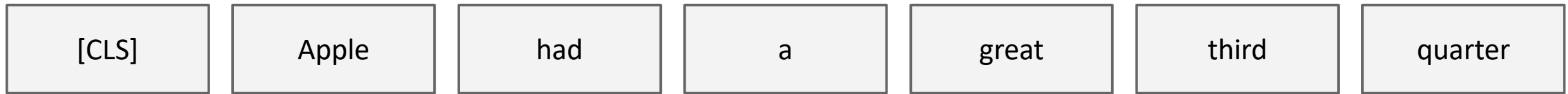
Attention based context distributions – CLS token is used during fine-tuning.

[CLS] Apple had a great third quarter [SEP]



Sentiment Classification Intuition

Classification – Supervised Task on top of Pre-trained BERT

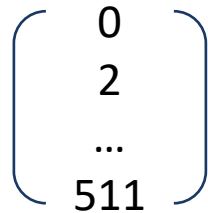


Class Label

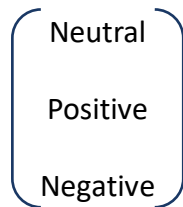
Learned during Fine Tuning

Weights
 $w_0, w_1, w_2, \dots, w_{511}$
 $w_0, w_1, w_2, \dots, w_{511}$
 $w_0, w_1, w_2, \dots, w_{511}$

CLS Token



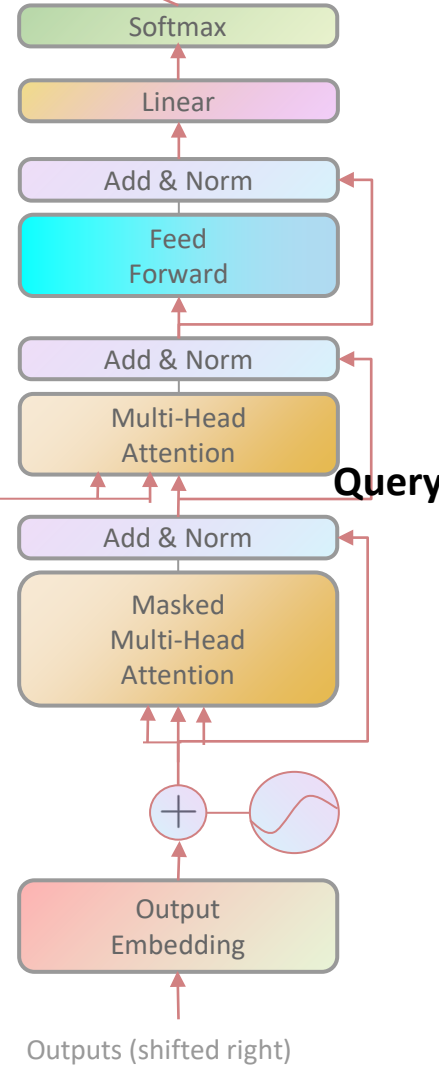
Class Vector



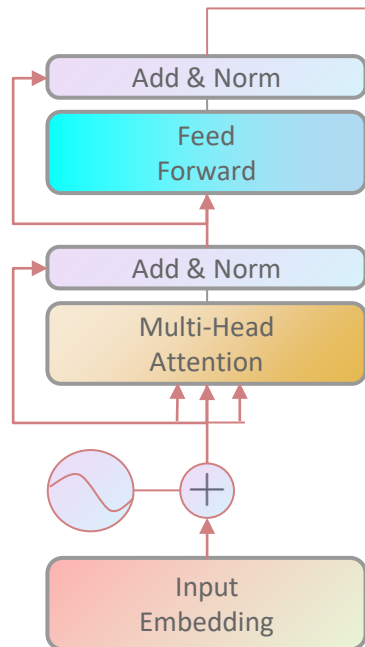
$$\begin{matrix} - \\ + \end{matrix} \begin{bmatrix} 0 & 0 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.85 & 0.7 & 0.95 & 0.8 & 0.2 & 1 \\ 0.33 & 0.6 & 0.01 & 0.1 & 0.15 & 0.3 \end{bmatrix} \times \begin{matrix} - \\ + \end{matrix} \begin{bmatrix} 0 \\ 0.2 \\ 1 \\ 0.1 \\ 0.3 \\ 0.1 \end{bmatrix}$$

Transformer Network – Language Understanding

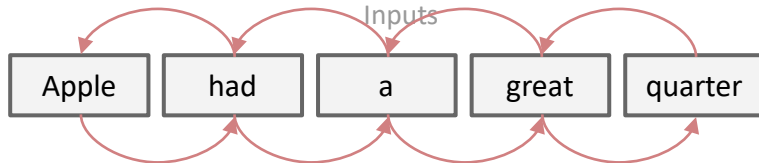
$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Qk^T}{\sqrt{d_k}}\right) v$$



Keys & Values



BERT: Bidirectional



Earnings Call transcript sections

Presentation and Q&A

Presentation

30,000 MSCI USA Transcripts

Management adapts quickly to how transcripts are taken by the market during earnings call.

Several London based executive coaches purchases earnings call transcripts to coach executives how to behave during earnings call.

Management controls what happens during prepared remarks.

Q & A

30,000 MSCI USA Transcripts

During Q&A would have fewer scripted remarks based on analyst questions.

Playing Favorites paper by HBS Lauren Cohen and Dong Lou. Firms that call on more favorable analysts experience more negative future earnings surprises and more future earnings restatements. [Metadata Tagging – Analysts vs Estimates]

Source: <https://dash.harvard.edu/bitstream/handle/1/11508220/14-021.pdf>

History of Natural Language Processing (NLP)

THANK YOU