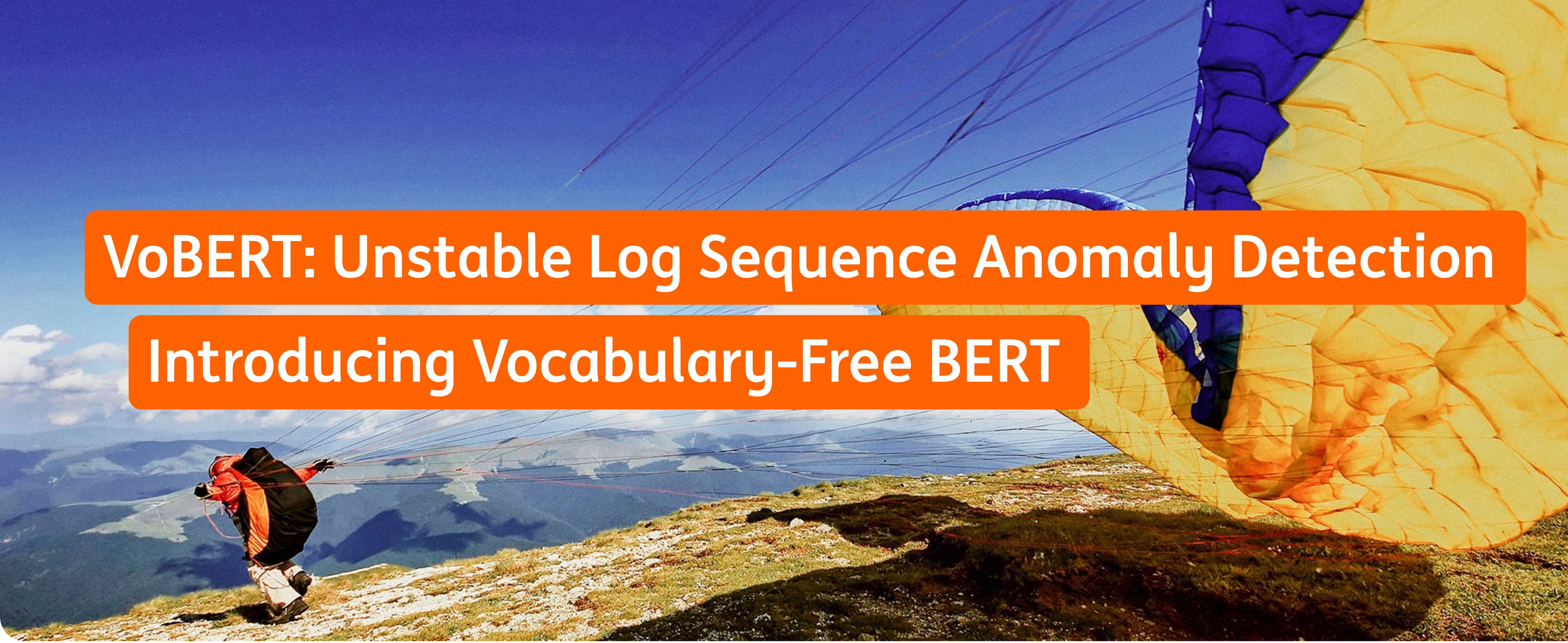


VoBERT: Unstable Log Sequence Anomaly Detection
Introducing Vocabulary-Free BERT

Dr Eduardo Barbaro
Head of Security Analytics at ING CISO
Visiting Researcher Cybersecurity Lab TUDelft

List of contributors:
Daan Hofman, ***ING*** & ***TUDelft***
Eduardo Barbaro, ***ING*** & ***TUDelft***
Yury Zhauriarovich, *Assistant professor* ***TUDelft***
Anna Lukina, *Assistant professor* ***TUDelft***

TUDelft

black hat

do your thing

Key Take Aways

Enables learning from **sequential data**

Increases robustness: Also works for **unstable** log data

Increases **explainability**: provides an element-level score

Shows the importance of **evaluating using real-world data**

Anomaly Detection?

- The identification of **rare events** or observations which deviate significantly from **most of the data**

- Finding **weird** things among **normal** things

Log Sequence?

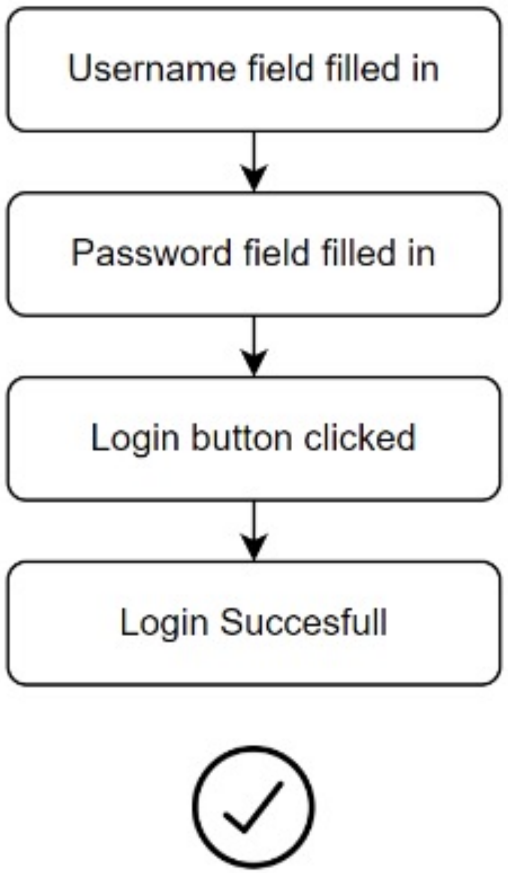
- Any software produces log files

- Log analysis is useful for:
  - Finding errors in software behaviour
  - Detect potential cybersecurity threats

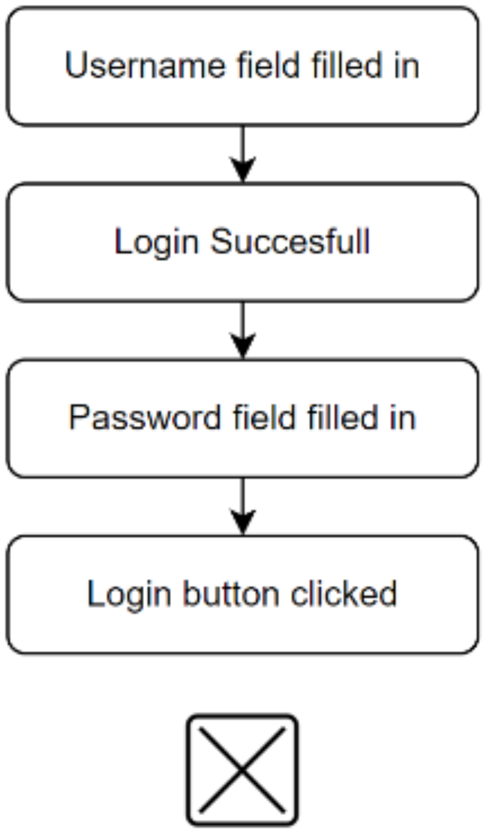
- Analysts must “manually” identify attacks

```
ERROR: Opening file "TestFile1.txt" from server MYSERVER
ERROR: Opening file "TestFile2.txt" from server MYSERVER
ERROR: Opening file "TestFile3.txt" from server MYSERVER
ERROR: Opening file "TestFile4.txt" from server MYSERVER
ERROR: Opening file "TestFile5.txt" from server MYSERVER
ERROR: Opening file "TestFile6.txt" from server MYSERVER
SUCCESS: Opening file "TestFile7.txt" from server MYSERVER
SUCCESS: Opening file "TestFile8.txt" from server MYSERVER
SUCCESS: Opening file "TestFile9.txt" from server MYSERVER
SUCCESS: Opening file "TestFile10.txt" from server MYSERVER
SUCCESS: Opening file "TestFile11.txt" from server MYSERVER
SUCCESS: Opening file "TestFile12.txt" from server MYSERVER
SUCCESS: Opening file "TestFile13.txt" from server MYSERVER
SUCCESS: Opening file "TestFile14.txt" from server MYSERVER
SUCCESS: Opening file "TestFile15.txt" from server MYSERVER
SUCCESS: Opening file "TestFile16.txt" from server MYSERVER
SUCCESS: Opening file "TestFile17.txt" from server MYSERVER
WARNING: File "TestFile18.txt" already exists.
WARNING: File "TestFile19.txt" already exists.
WARNING: File "TestFile20.txt" already exists.
SUCCESS: Opening file "TestFile21.txt" from server MYSERVER
SUCCESS: Opening file "TestFile22.txt" from server MYSERVER
SUCCESS: Opening file "TestFile23.txt" from server MYSERVER
SUCCESS: Opening file "TestFile24.txt" from server MYSERVER
SUCCESS: Opening file "TestFile25.txt" from server MYSERVER
```

Log Sequence Anomaly Detection

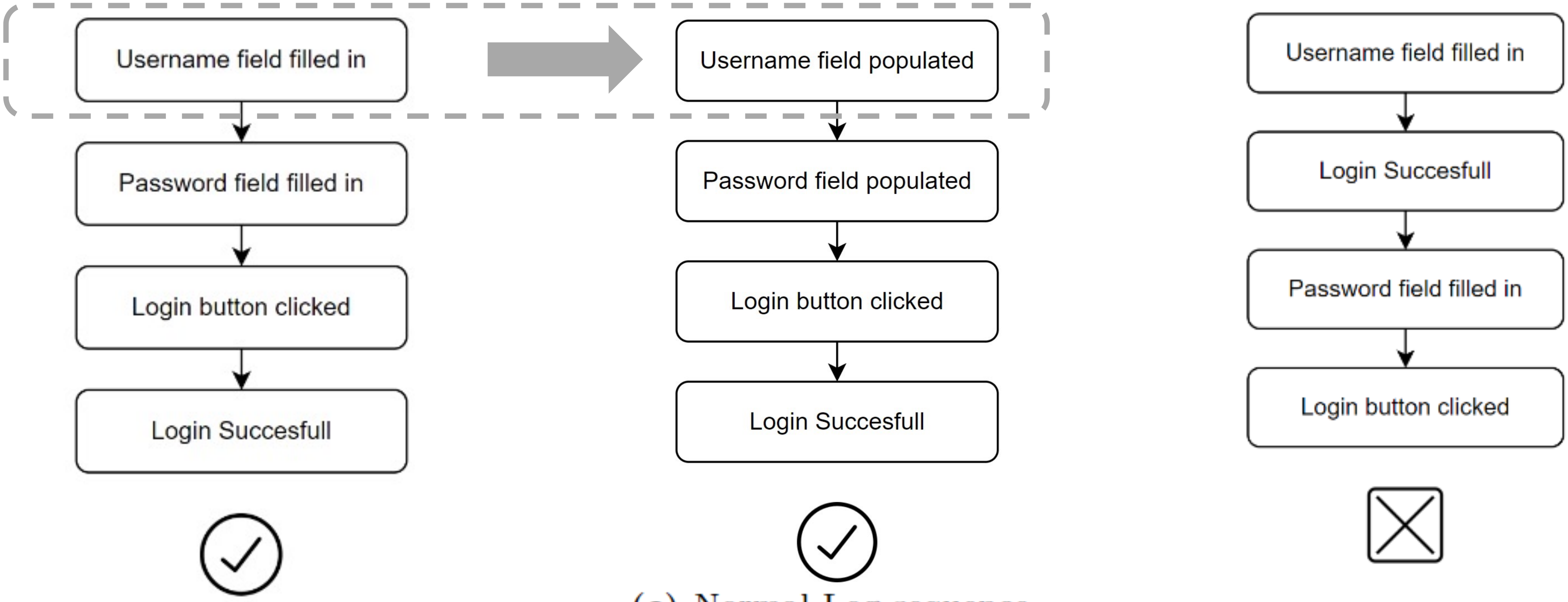


(a) Normal Log sequence



(b) Anomalous Log sequence

Unstable Log Sequence Anomaly Detection



(a) Normal Log sequence

(c) Normal Log sequence with slightly changed log messages

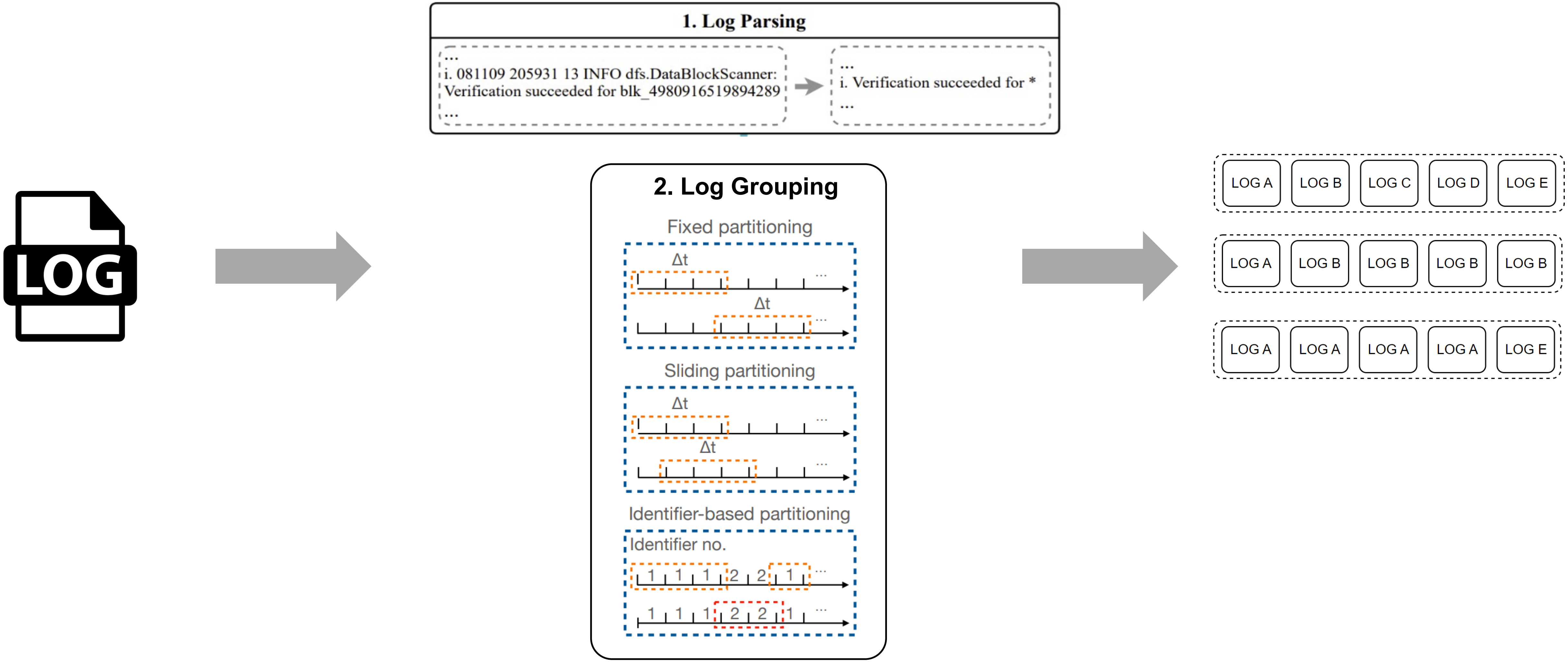
(b) Anomalous Log sequence

That begs the following question:
(and some others too)

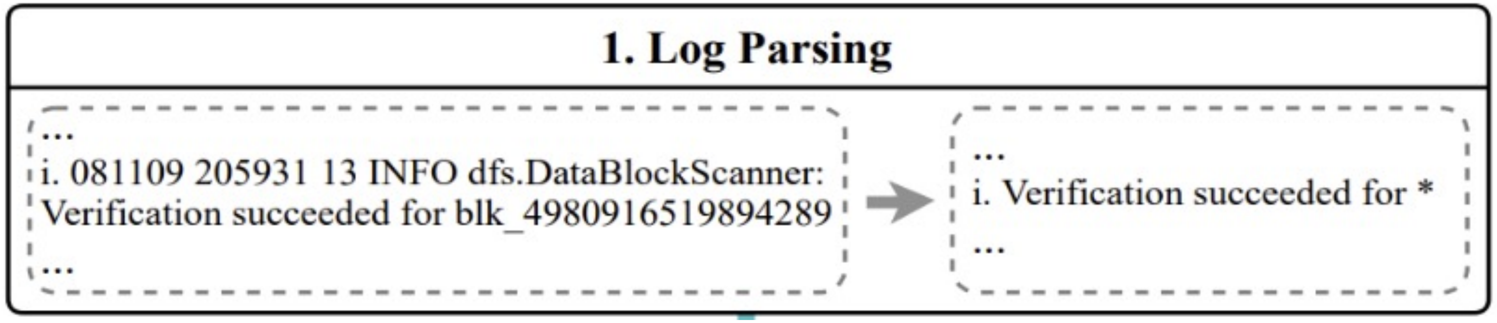
How can we identify anomalies in unstable sequential log data?

1. **Explainability:** what is the influence of each individual log/alert?
2. **Unstable Logs:** how can we deal with log instability?
3. **Real-world vs synthetic data:** how do models perform on real-world security events?

Introduction



Step 1: Log Parsing (Cleaning the house)

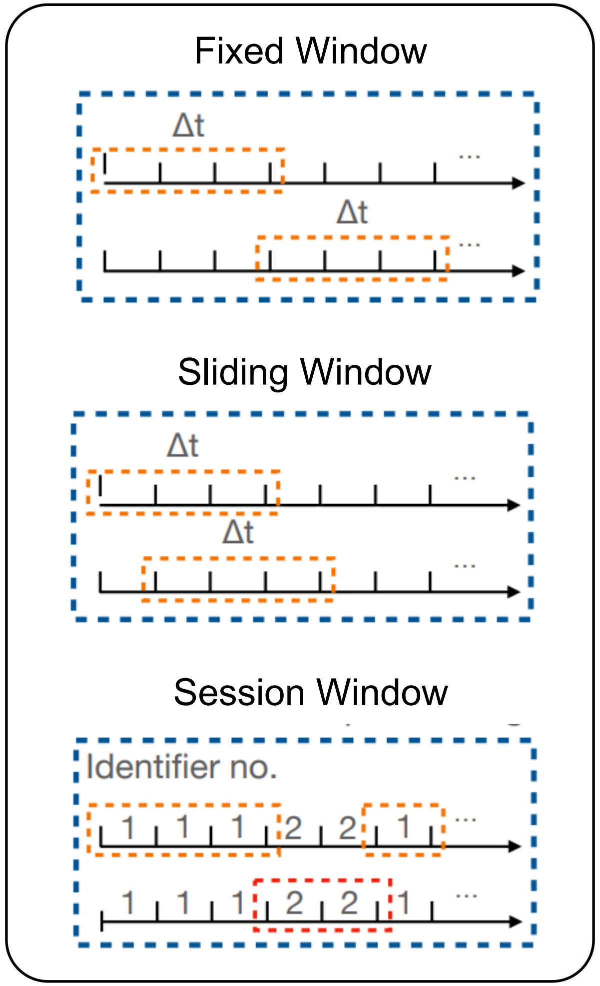


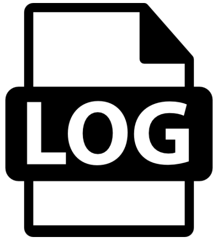
Raw log message

LogKey

Step 2: Log Grouping (too much data....)

2. Log Grouping





Pre-processing Output

Sequence 1	LOG A	LOG B	LOG C	LOG D	LOG E
Sequence 2	LOG A	LOG B	LOG B	LOG B	LOG B
Sequence 3	LOG A	LOG A	LOG A	LOG A	LOG E

How to detect anomalies?

- Rule based?
  - Too noisy (loads of false positives)
- Shallow Machine Learning?
  - We lose the temporal (order) information
- Deep Learning?
  - Potentially. But where do we start?

Background

So many words, so little numbers

Where do we start?

How do we go from words to numbers?

‘The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.’

a	and	been	earth	fascination	for	has	moon
1	1	1	1	1	1	1	1
natural	**of**	**only**	**subject**	**thousands**	**the**	**wonder**	**years**
1	2	1	1	1	1	1	1

The English Wikionary has over 700k entries

“Raise for everyone, no termination!”

or

“No raise, termination for everyone!”

The above can work, but word order has some meaning...

Tokenization

Tokenisation method	Tokens	Token count	Vocab size
Sentence	‘The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.’	1	# sentences in doc
Word	'The', 'moon,', "Earth's", 'only', 'natural', 'satellite,', 'has', 'been', 'a', 'subject', 'of', 'fascination', 'and', 'wonder', 'for', 'thousands', 'of', 'years.'	18	171K (English1)
Sub-word	'The', 'moon', ',', 'Earth', "'", 's', 'only', 'natur', 'al', 'satellite', ',', 'has', 'been', 'a', 'subject', 'of', 'fascinat', 'ion', 'and', 'wonder', 'for', 'thousand', 's', 'of', 'year’, ‘s', '.'	27	(varies)
Character	'T', 'h', 'e', ' ', 'm', 'o', 'o', 'n', ',', ' ', 'E', 'a', 'r', 't', 'h', "'", 's', ' ', 'o', 'n', 'l', 'y', ' ', 'n', 'a', 't', 'u', 'r', 'a', 'l', ' ', 's', 'a', 't', 'e', 'l', 'l', 'i', 't', 'e', ',', ' ', 'h', 'a', 's', ' ', 'b', 'e', 'e', 'n', ' ', 'a', ' ', 's', 'u', 'b', 'j', 'e', 'c', 't', ' ', 'o', 'f', ' ', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', ' ', 'a', 'n', 'd', ' ', 'w', 'o', 'n', 'd', 'e', 'r', ' ', 'f', 'o', 'r', ' ', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', 's', ' ', 'o', 'f', ' ', 'y', 'e', 'a', 'r', 's', '.'	110	52 + punctuation (English)

Tokenization

Tokenization method	Tokens	Token count	Vocab size
Sentence	‘The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.’	1	# sentences in doc
Word	'The', 'moon,', "Earth's", 'only', 'natural', 'satellite,', 'has', 'been', 'a', 'subject', 'of', 'fascination', 'and', 'wonder', 'for', 'thousands', 'of', 'years.'	18	171K (English1)
Sub-word			(varies)
Character			52 + punctuation (English)

Pros:
Intuitive.

Cons:
Big vocabularies.
Complications such as handling misspellings.
Other out-of-vocabulary words.

Tokenization

Tokenization method		unt	Vocab size
Sentence			# sentences in doc
Word			171K (English1)
Sub-word			(varies)
Character	'T', 'h', 'e', ' ', 'm', 'o', 'o', 'n', ',', ' ', 'E', 'a', 'r', 't', 'h', "'", 's', ' ', 'o', 'n', 'l', 'y', ' ', 'n', 'a', 't', 'u', 'r', 'a', 'l', ' ', 's', 'a', 't', 'e', 'l', 'l', 'i', 't', 'e', ',', ' ', 'h', 'a', 's', ' ', 'b', 'e', 'e', 'n', ' ', 'a', ' ', 's', 'u', 'b', 'j', 'e', 'c', 't', ' ', 'o', 'f', ' ', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', ' ', 'a', 'n', 'd', ' ', 'w', 'o', 'n', 'd', 'e', 'r', ' ', 'f', 'o', 'r', ' ', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', 's', ' ', 'o', 'f', ' ', 'y', 'e', 'a', 'r', 's', '.'	110	52 + punctuation (English)

Pros:
Small vocabulary.
No out-of-vocabulary words.

Cons:
Loss of context within words.
Much longer sequences for a given input.

Tokenization

Compromise

“Smart” vocabulary built from characters which co-occur frequently. More robust to novel words. Compromise

“Smart” vocabulary built from characters which co-occur frequently. More robust to novel words.

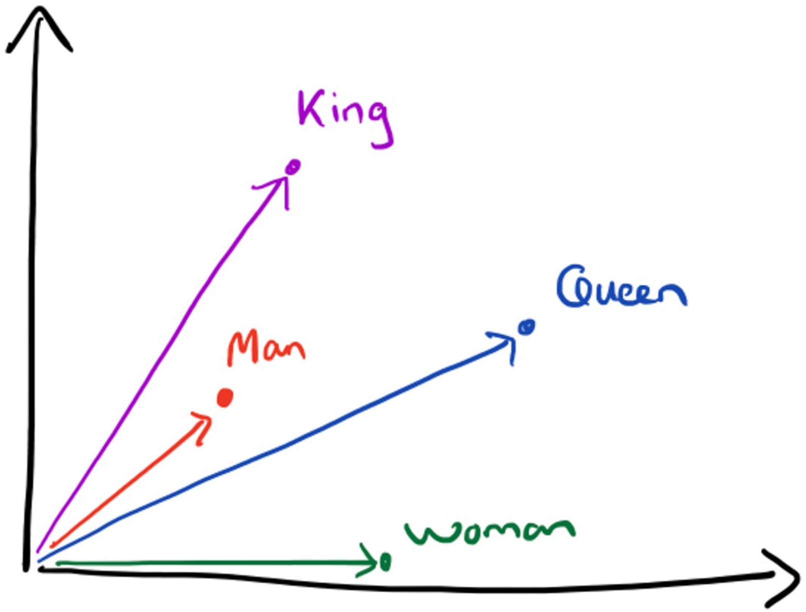
Tokenization method			size
Sentence			ces in doc
Word			glish1)
Sub-word	'The', 'moon', ',', 'Earth', "'", 's', 'only', 'natur', 'al', 'satellite', ',', 'has', 'been', 'a', 'subject', 'of', 'fascinat', 'ion', 'and', 'wonder', 'for', 'thousand', 's', 'of', 'year', 's', '.'	27	(varies)
Character	'T', 'h', 'e', ' ', 'm', 'o', 'o', 'n', ',', ' ', 'E', 'a', 'r', 't', 'h', "'", 's', ' ', 'o', 'n', 'l', 'y', ' ', 'n', 'a', 't', 'u', 'r', 'a', 'l', ' ', 's', 'a', 't', 'e', 'l', 'l', 'i', 't', 'e', ',', ' ', 'h', 'a', 's', ' ', 'b', 'e', 'e', 'n', ' ', 'a', ' ', 's', 'u', 'b', 'j', 'e', 'c', 't', ' ', 'o', 'f', ' ', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', ' ', 'a', 'n', 'd', ' ', 'w', 'o', 'n', 'd', 'e', 'r', ' ', 'f', 'o', 'r', ' ', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', 's', ' ', 'o', 'f', ' ', 'y', 'e', 'a', 'r', 's', '.'	110	52 + punctuation (English)

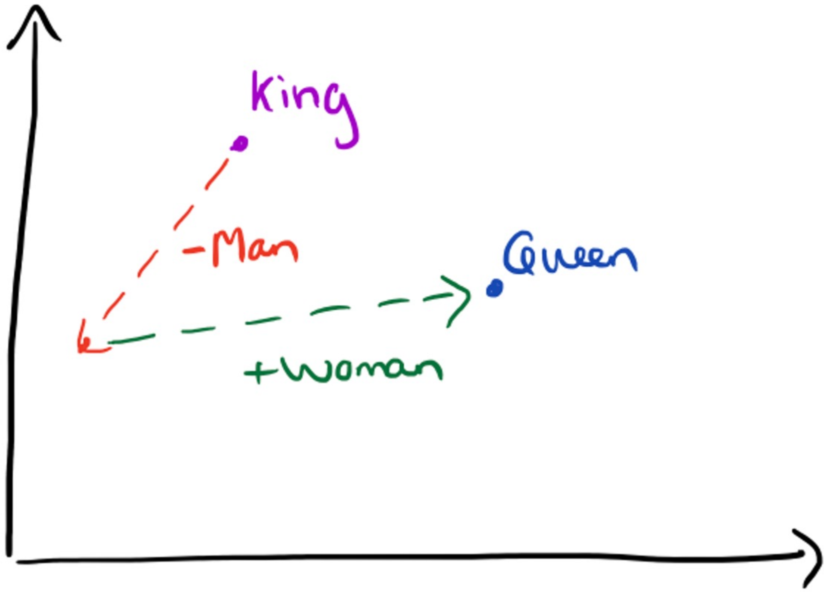
Represent words with vectors

Words with similar meaning tend to occur in similar contexts:

> The king waved to the crowd from the balcony.
> The queen waved to the subjects from the terrace.

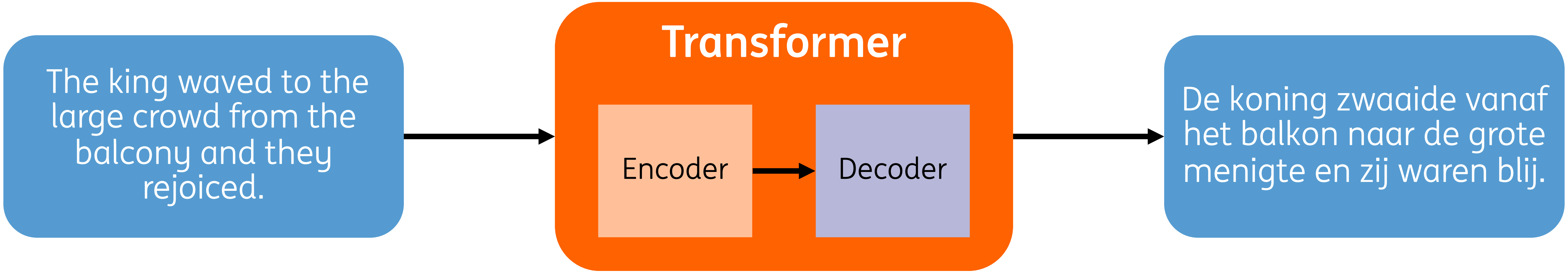
The words **king** and **queen** share context here, as do **balcony** and **terrace**.





Images source: https://blog.acolyer.org/

Now, pay attention, this is the good stuff



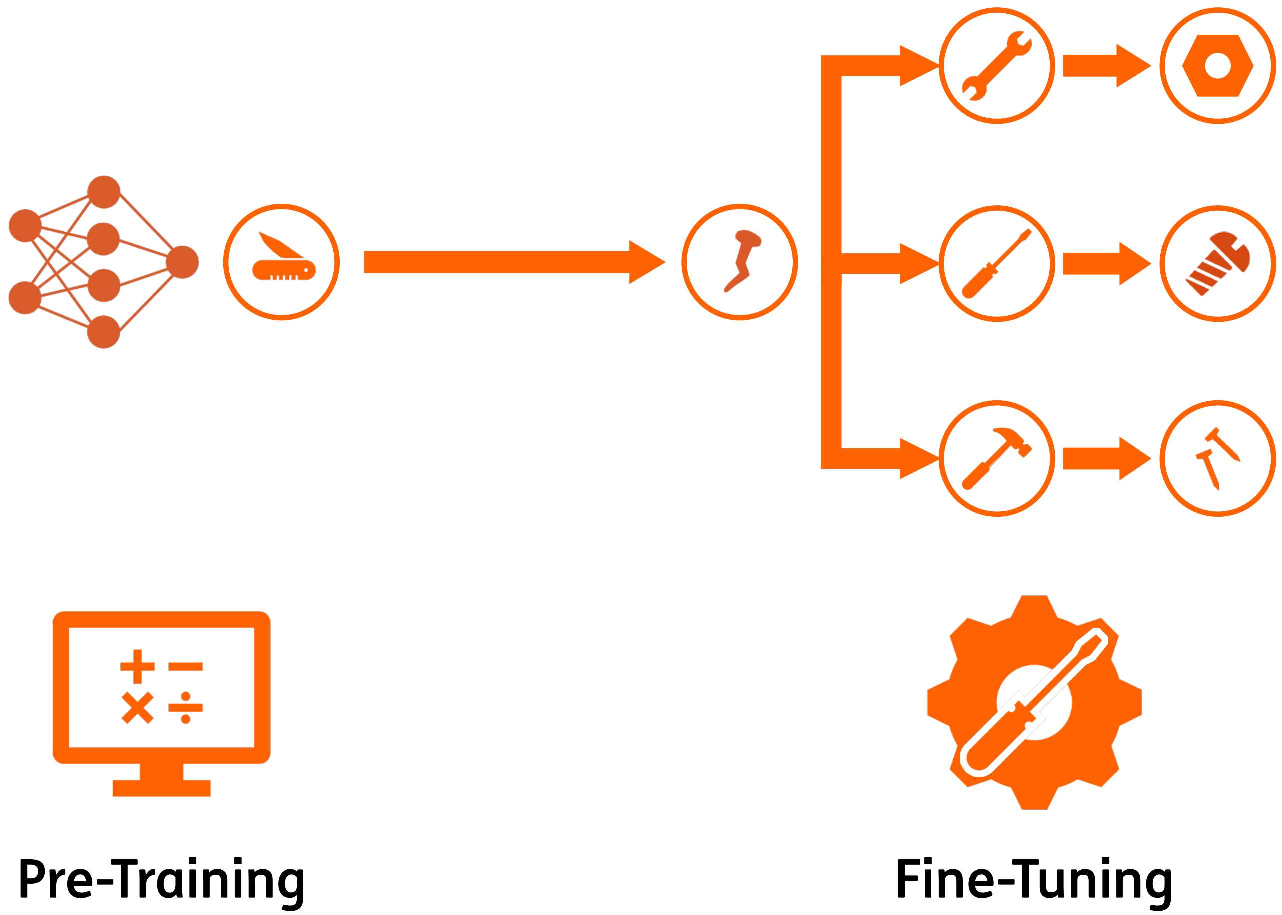
High Attention

The king waved to the large crowd from the balcony and they rejoiced.

Low Attention

Even the best models need fine-tuning

Fine-tuning enables the tailoring of LLMs to specific IT challenges, bridging the gap between generalised understanding and specialised solutions.

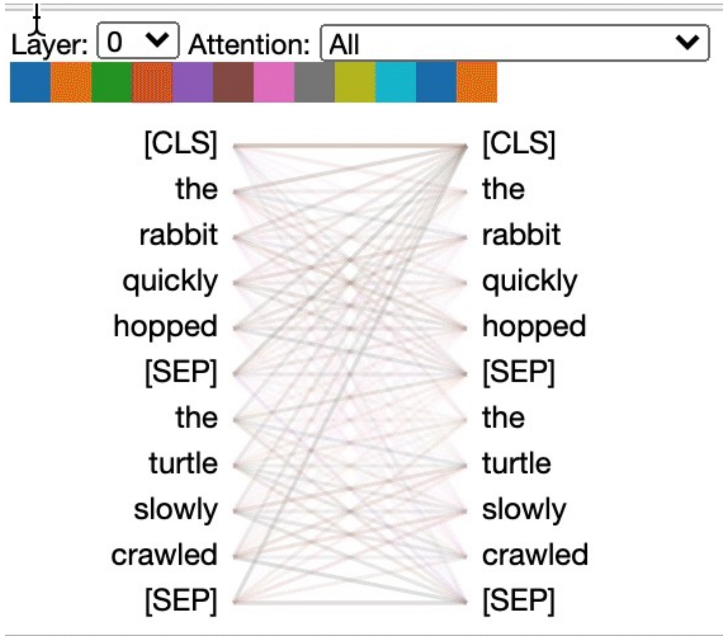


BERT models are a great starting point

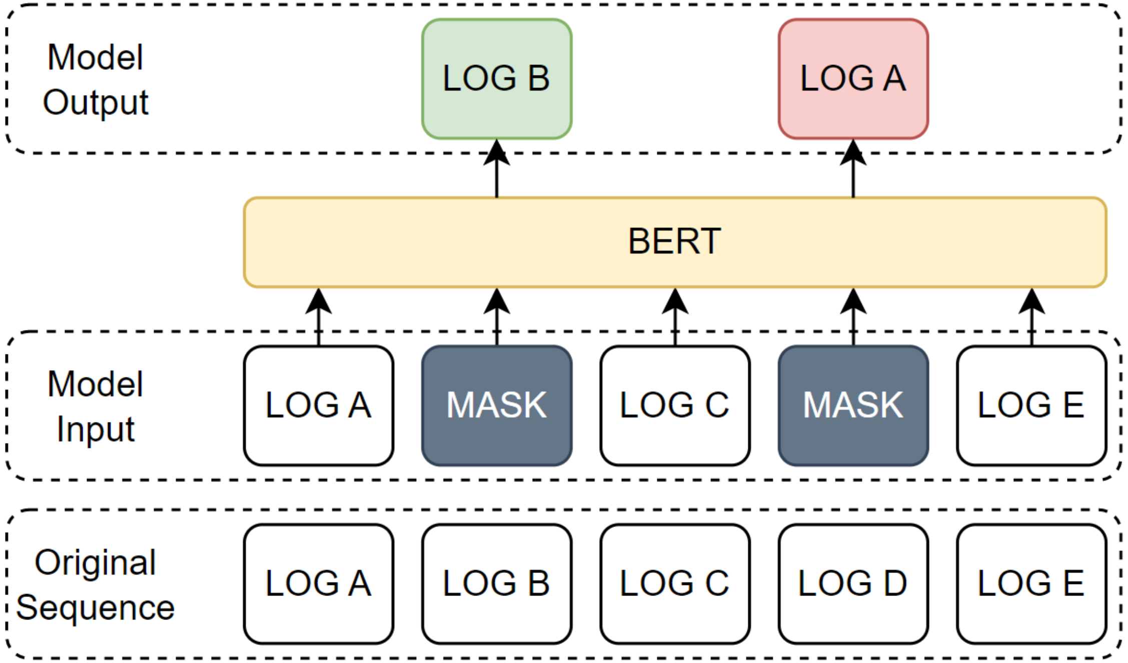
- Bidirectional Encoder Representations from Transformers

- Transformer-based architecture: Just like **GPT**
  - Transforms text based on attention-mechanism

- Word embeddings with context:
  - I go to a bar ≠ I raise the bar

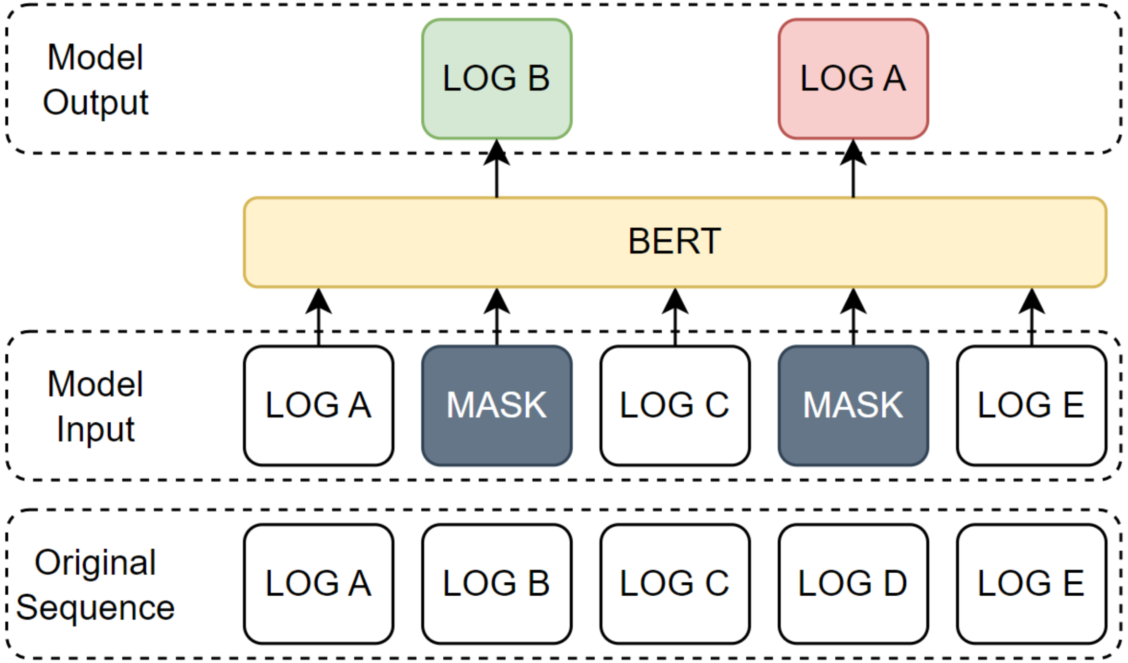


BERT training: Masked Language Modelling (MLM)



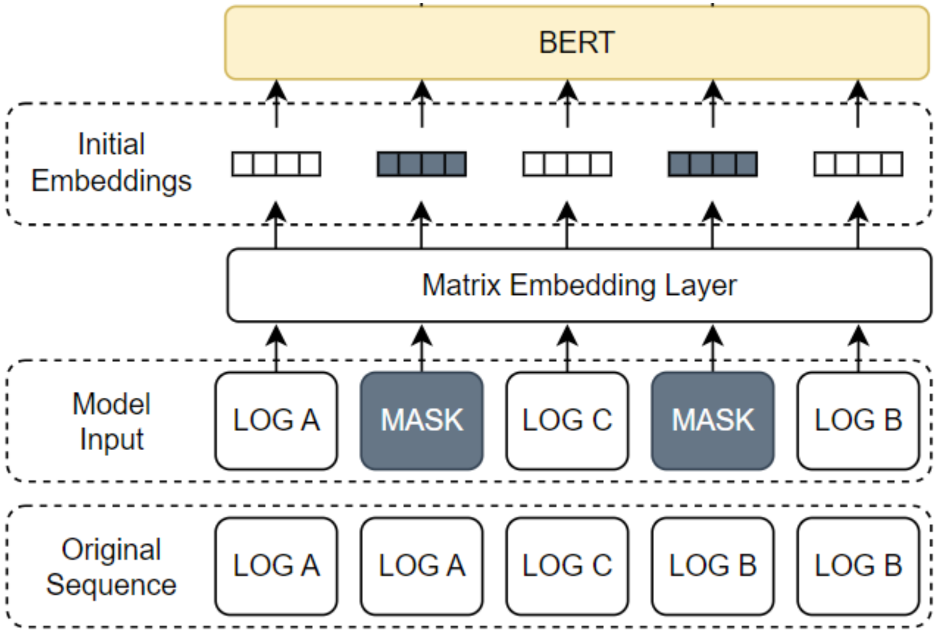
BERT for Anomaly Detection

- Trained only on normal sequences
- Predicts **poorly** on anomalous sequences



BERT for Anomaly Detection

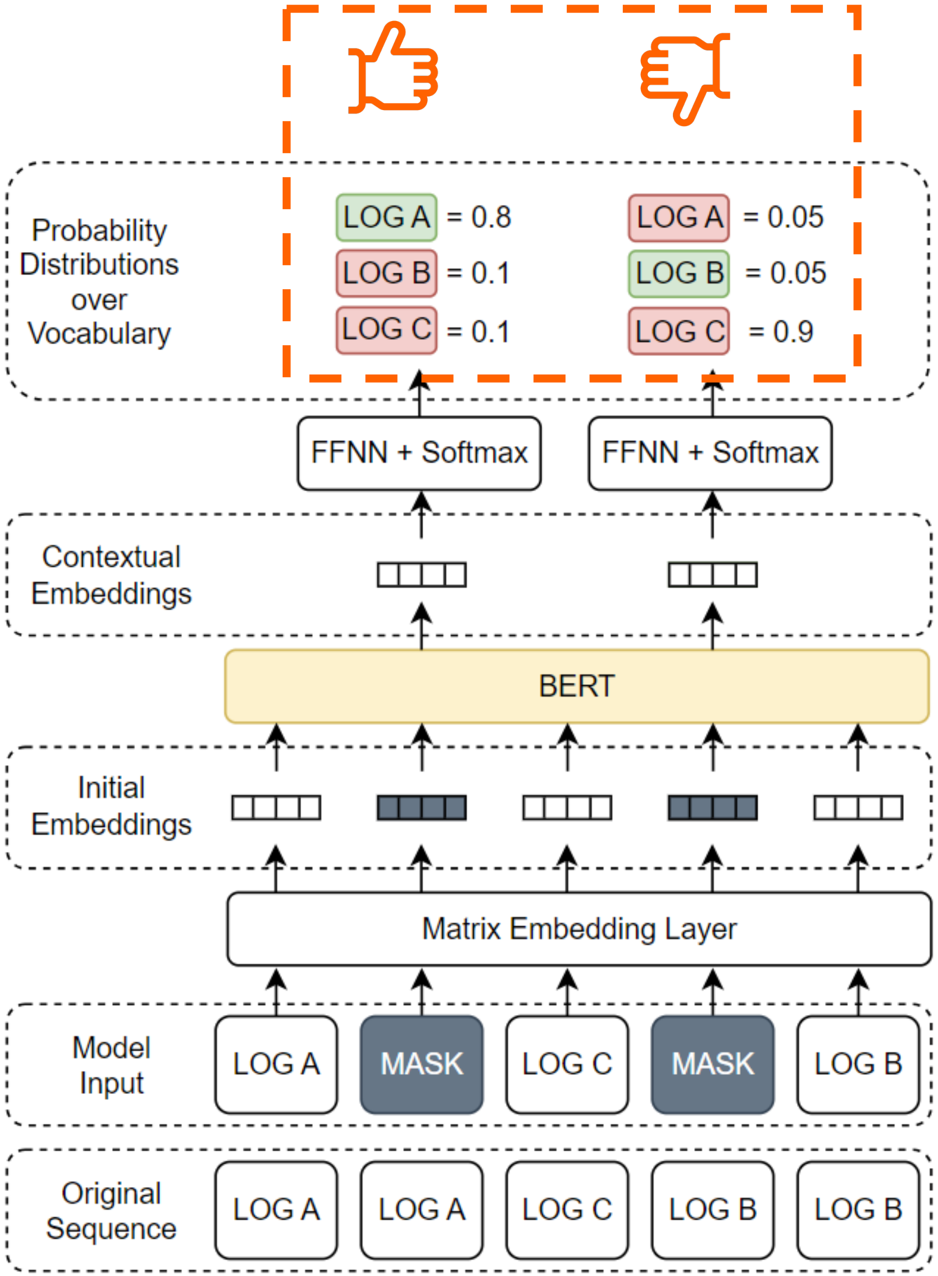
- We need Embeddings as logs need to be represented in a numerical fashion



LogBERT

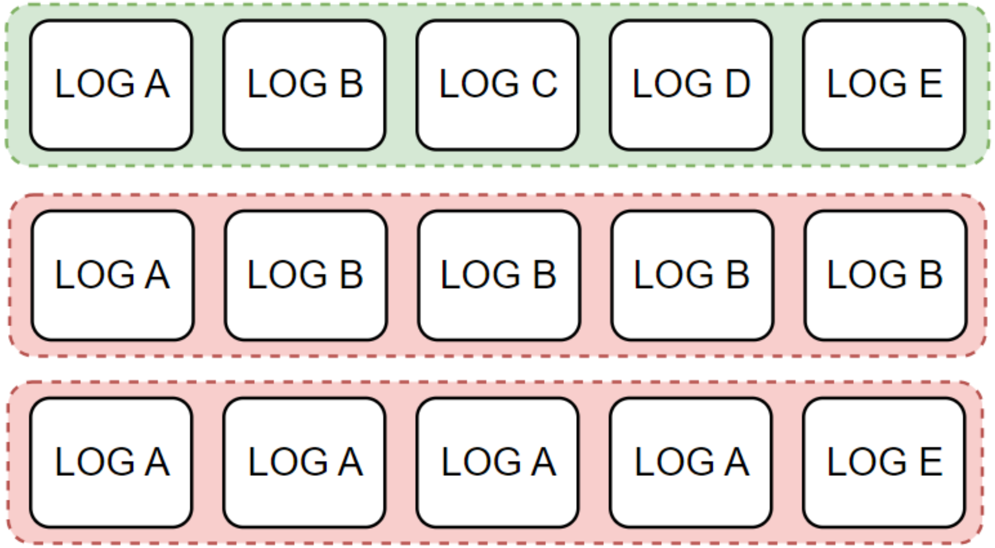
- FFNN: Single Layer

- Anomaly Criteria
  - Correct token is not in top-***g*** predicted tokens
  - More than ***t*** % of the masked log-keys are wrong

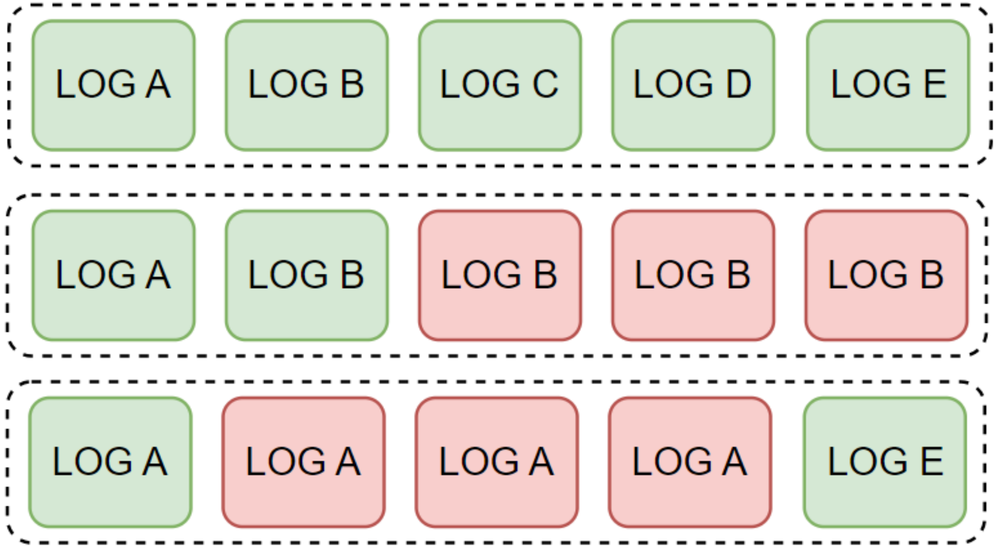



Solution
How can we solve these problems?

Explainability: Element-Level Prediction

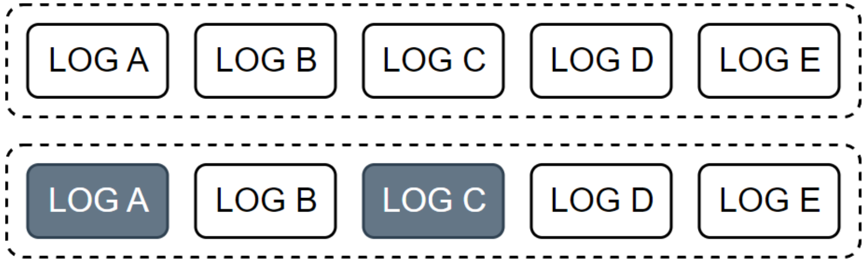


Sequence Level Predictions
(LogBERT)

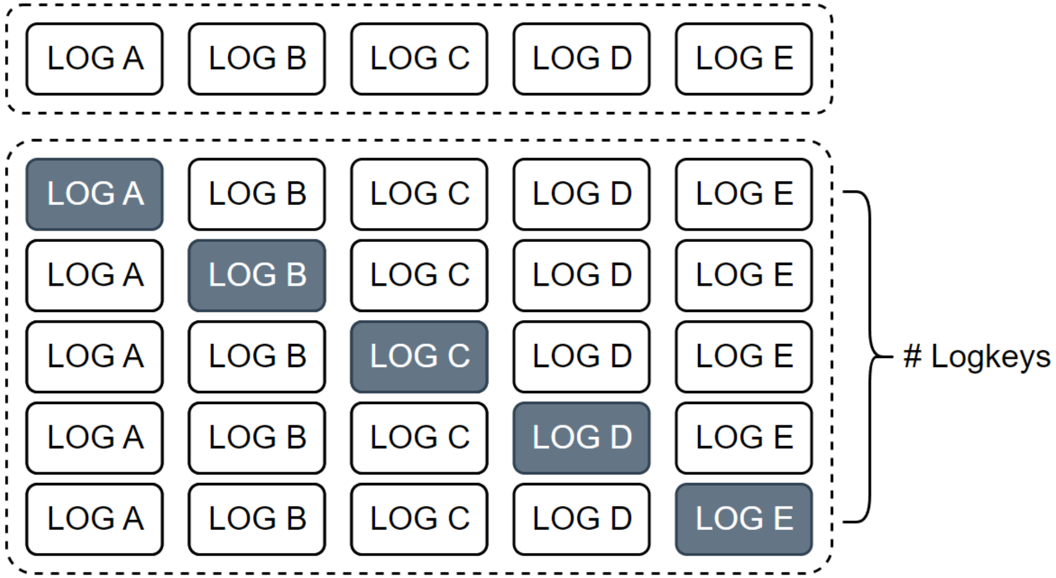


Element Level Predictions
(VoBERT)

Explainability: Per Element Masking



Ratio Masking O(1)
(LogBERT)



Per Element Masking
O(n)
(VoBERT)

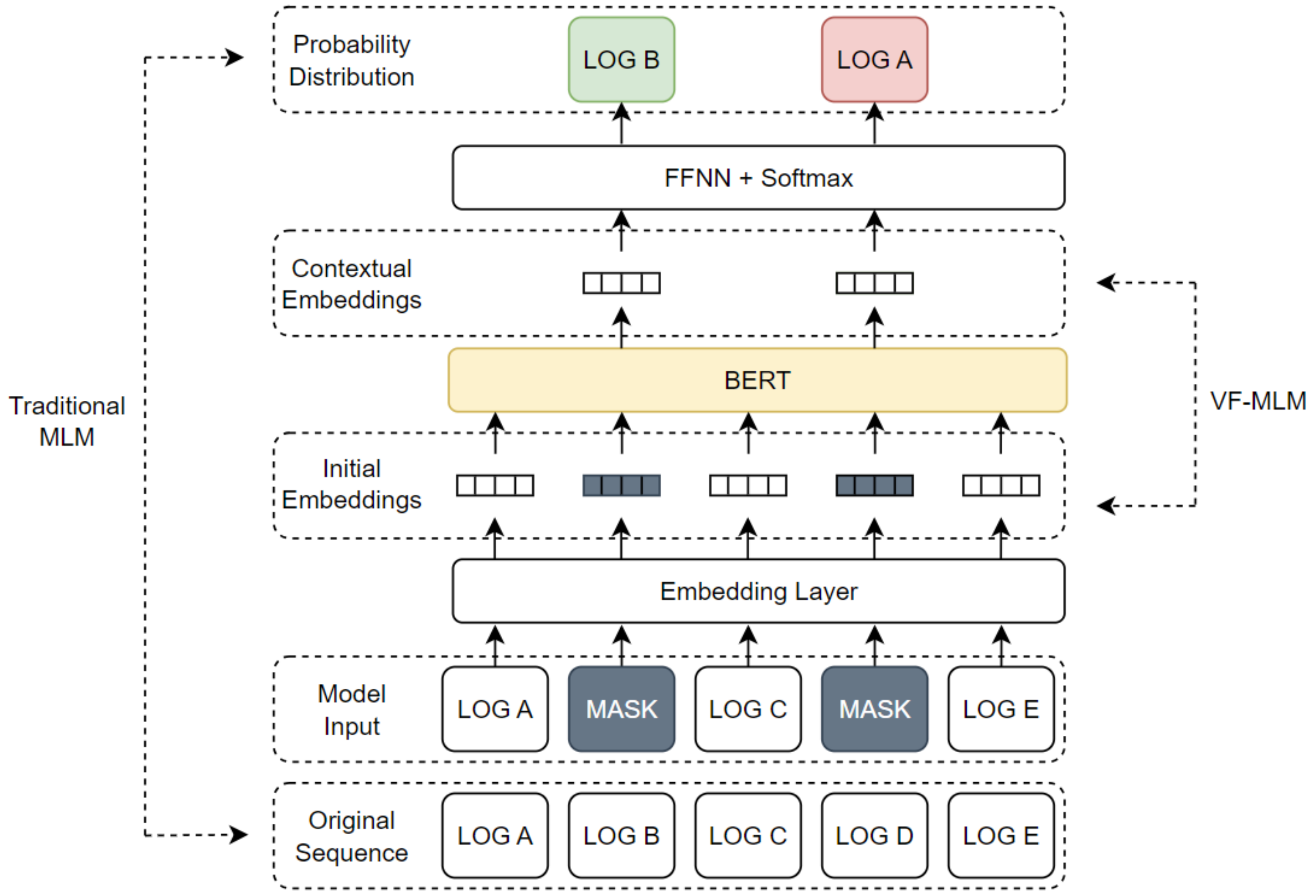
Log-stability: How can we make it more robust?

- By making it Vocabulary-Free:
  - Model architecture cannot depend on vocabulary
  - Embedding layer works with out-of-vocabulary log-keys

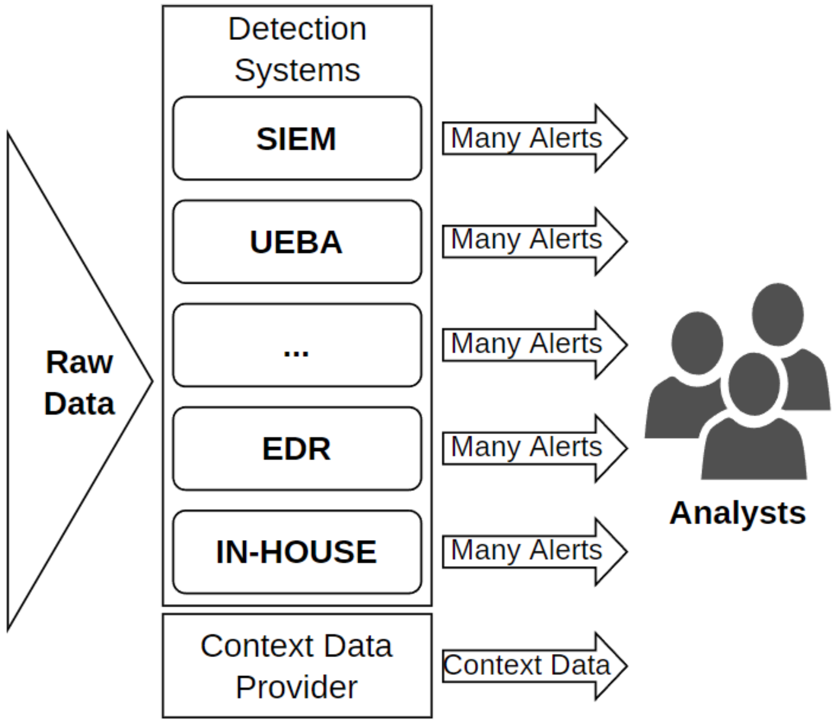
- Novel pre-training task based on this key insight:
  - No need to actually reconstruct the whole sequence, **we just need to know how close the model was**

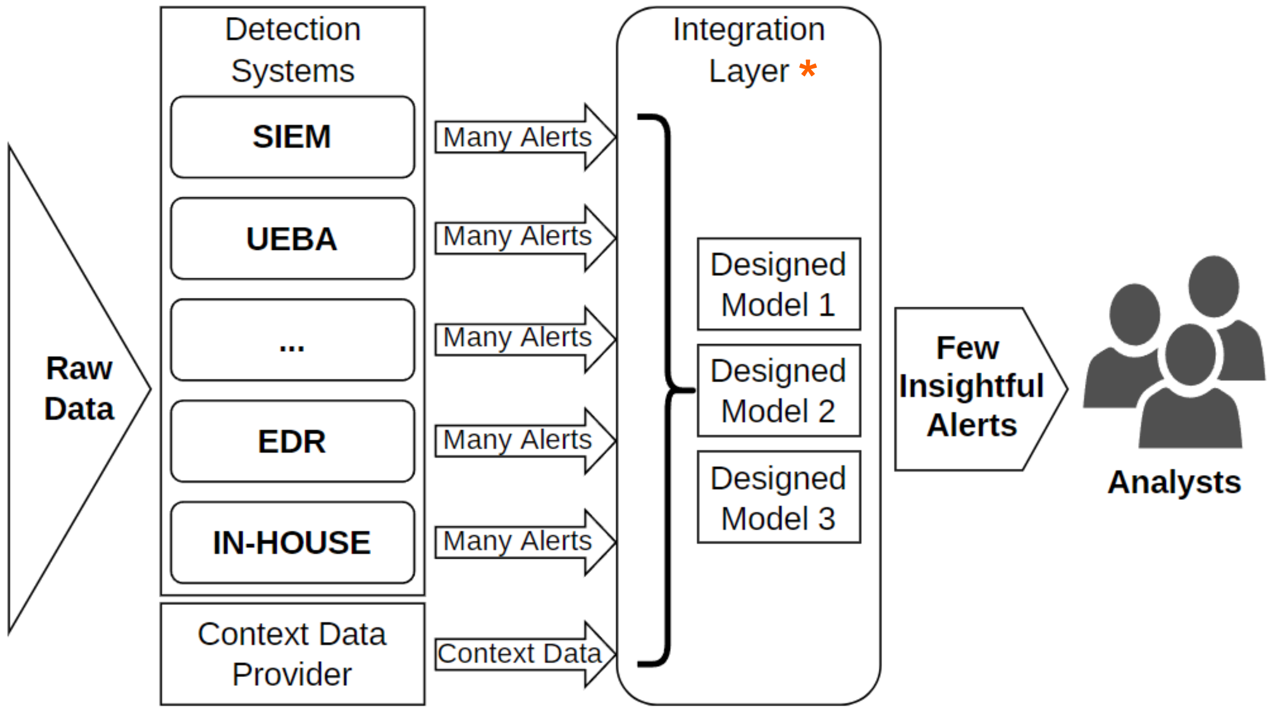
Log-stability: Vocabulary-Free MLM

- Architecture requirement:
  - Compare embeddings directly
- Embedding layer requirement:
  - Semantic embedding layer



Real-world data: meet the security detection framework





*ACCEPTED Roelofs et. al. Finding Harmony in the Noise: Blending Security Alerts for Attack Detection, **39th ACM/SIGAAP Symposium on Applied Computing 2024**

Results

How did we do?

Evaluation Log Data

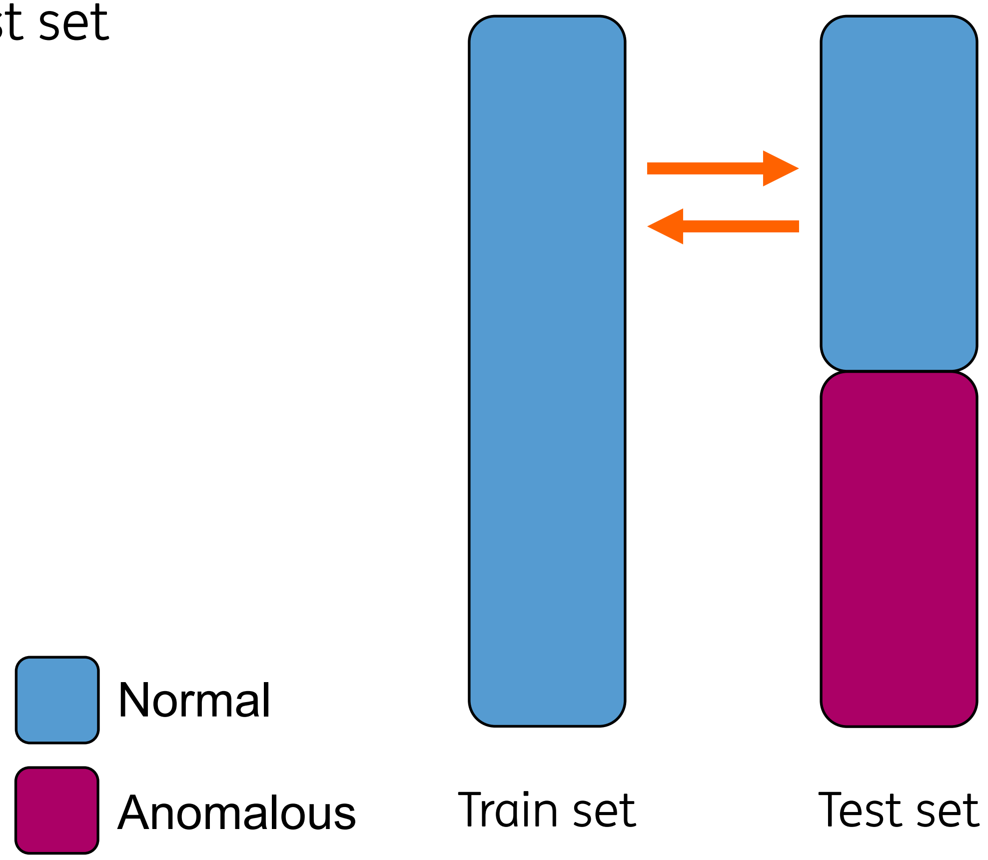
- 3 most frequently used High Performance Computing (HPC) log datasets
  - Hadoop Distributed File System (HDFS)
  - BlueGene/L Supercomputer System (BGL)
  - Thunderbird (TBird)

- ING alert dataset

Dataset	Number of logs	Of which anomalous
HDFS (Logs)	11,172,157	284,818 (3%)
BGL (Logs)	4,747,963	348,460 (1%)
TBird Small (Logs)	20,000,000	758,562 (4%)
ING (Alerts)	399,061	6,466 (2%)

Data Instability

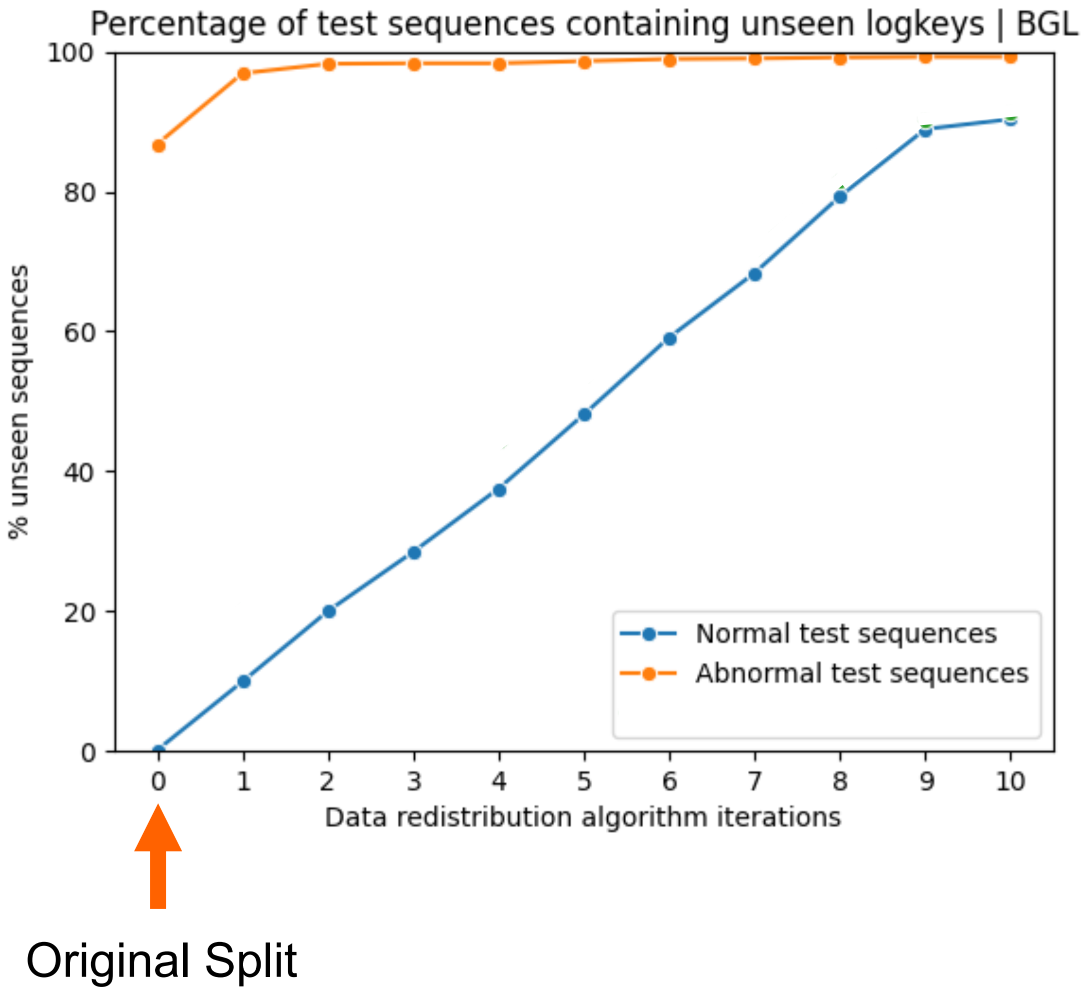
- Proxy: Percentage of unseen log-keys in the normal test set
- Data redistribution algorithm
  - Reshuffle train-test split
  - Train and test size remain fixed



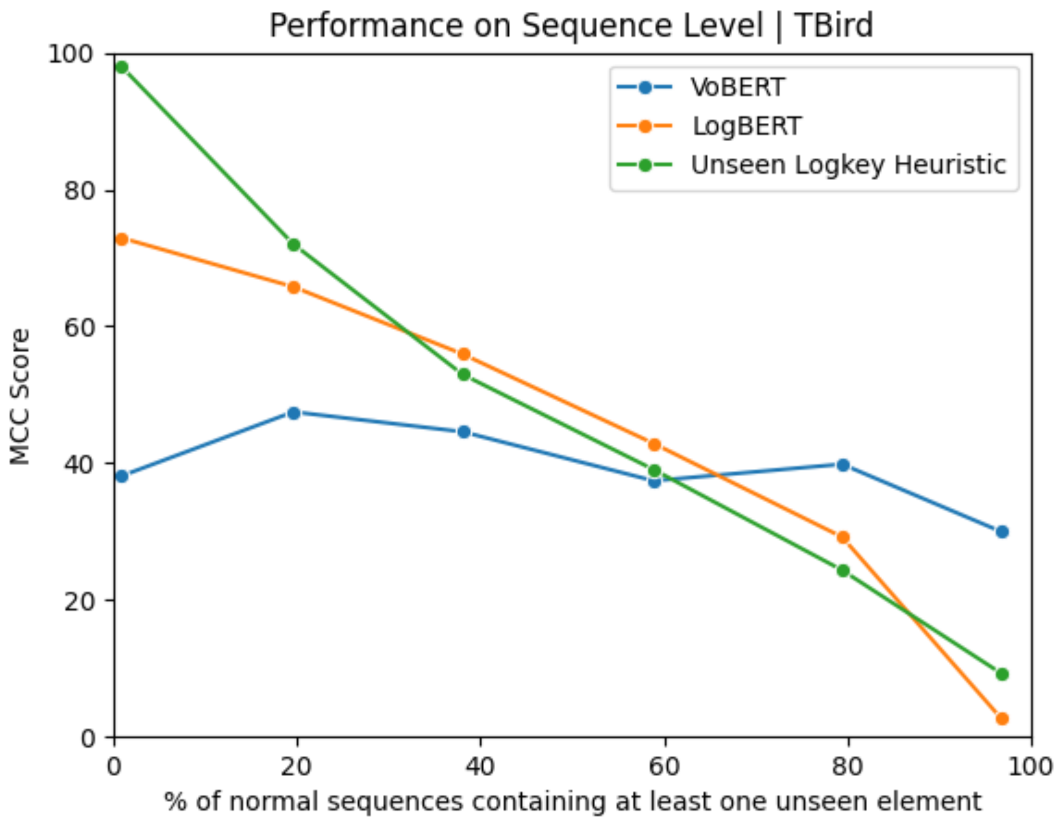
Increasing Data Instability (BGL Dataset)

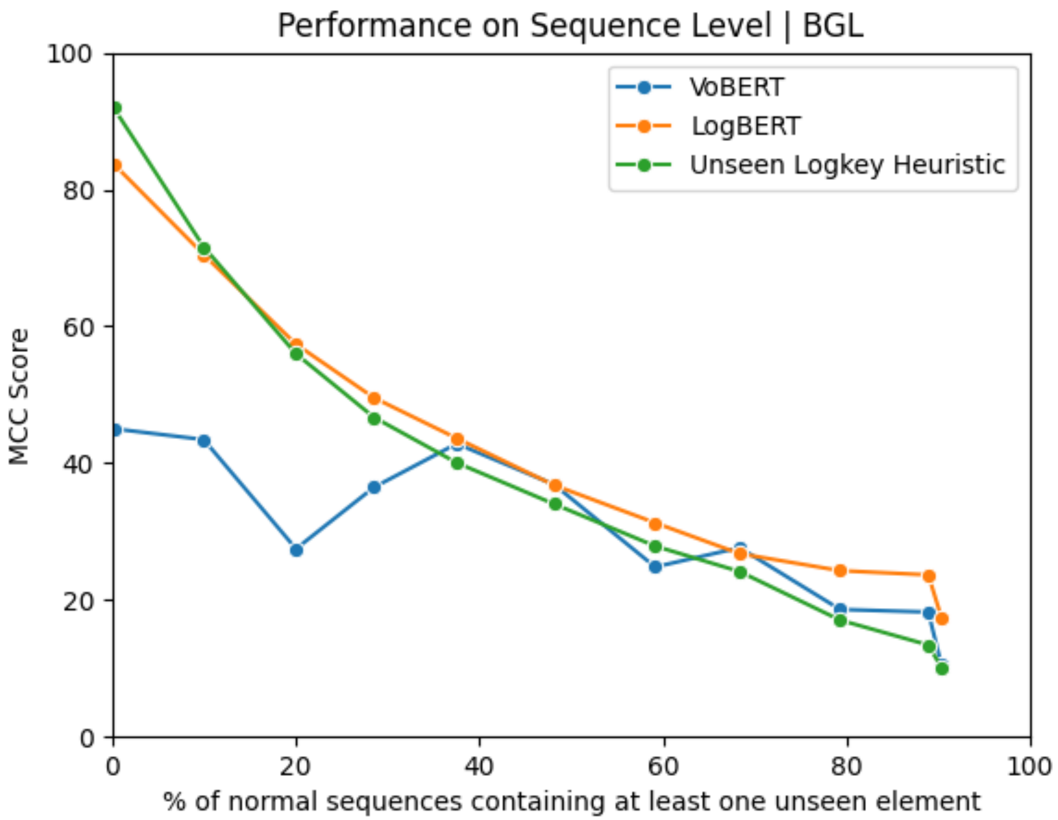
- Original split:
  - Normal sequences contain <5% unseen
  - Anomalous sequences contain >80% unseen
  - Few unseen log-keys in total

- We increase the unseen percentage to >80%



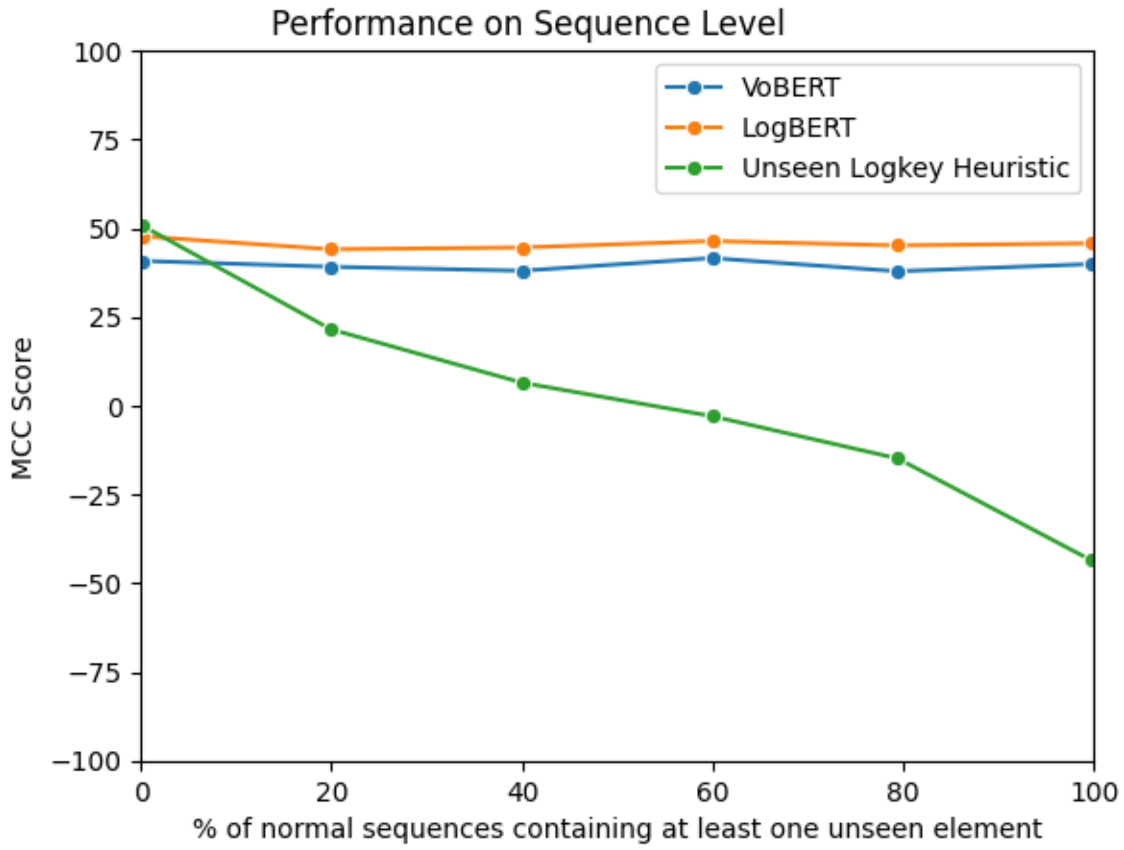
Results Public Datasets





Results using real-world data

- Simple heuristic did not work
- VoBERT had similar performance to LogBERT
- LogBERT performance was stable
  - Why? The average percentage of unseen log-keys in the sequences did not increase
  - We will probably see this effect when using this metric instead



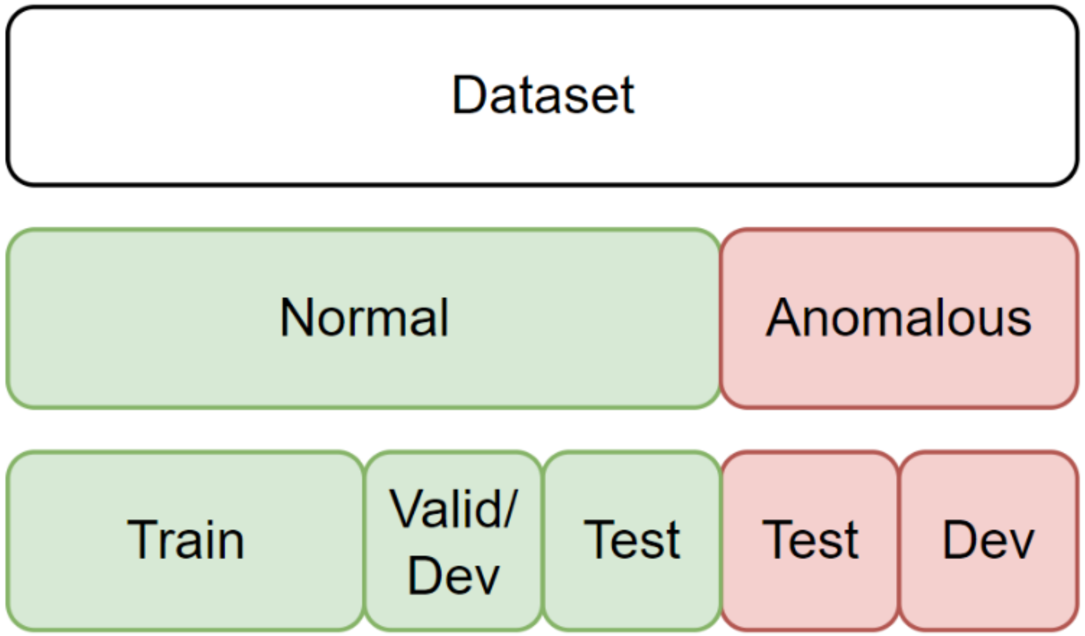
Conclusion

- Using a transformer model allows us to **leverage sequential data** (keep the alert order)
- Our solution is robust in **unstable log data** environments
- **Explainability**: Element-level evaluation performance can provide extra insights, but at a significant computational cost
  - Use it to further investigate suspicious alerts/logs
- LogBERT's performance that was not representative of a real-world situation
  - Don't blindly trust published research
  - It is important to **evaluate on real-world data**

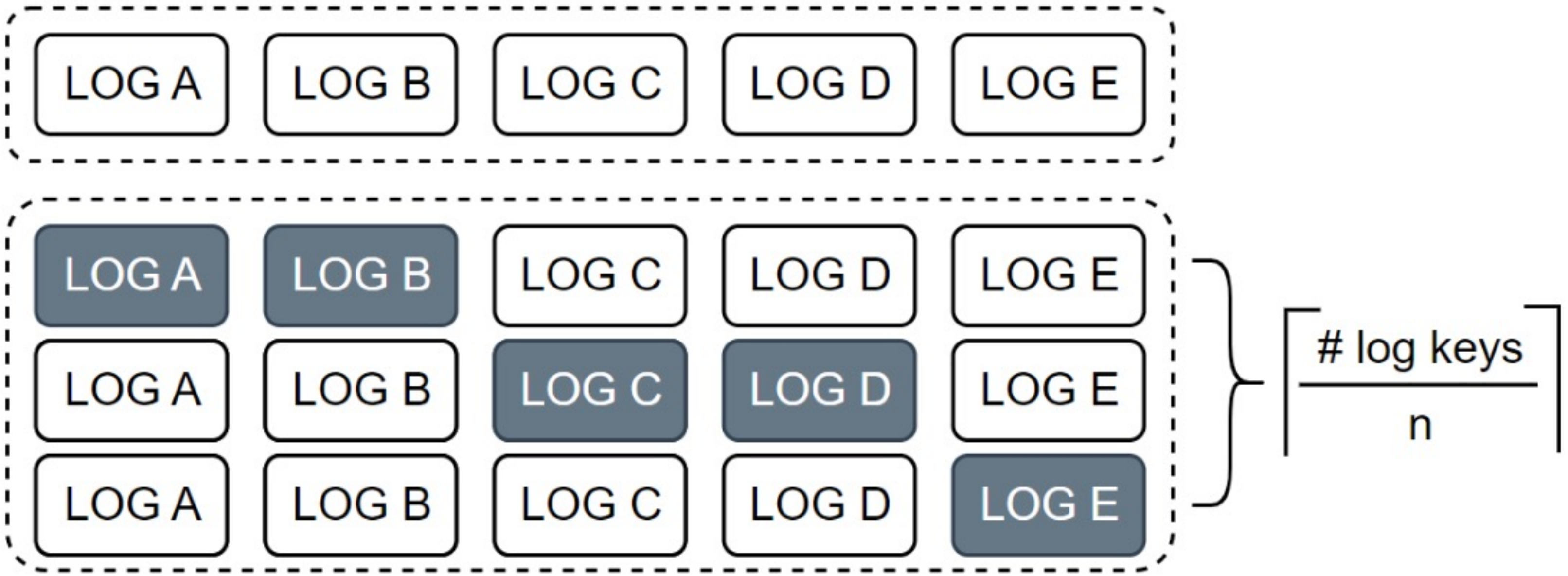
Visit us at Booth 436

do your thing

Data set split



Future Work: n-gram masking



Data Instability: Case Study