Classification of System Call Sequences using Anomalous Detection

William Doyle

Advisor: Aaron Cass

Imagine this scenario...

- You’re on an open network at a cafe and become the victim of a hacker

Imagine this scenario...

- You’re on an open network at a cafe and become the victim of a hacker

- The hacker attempts to read your file system and find your admin user information

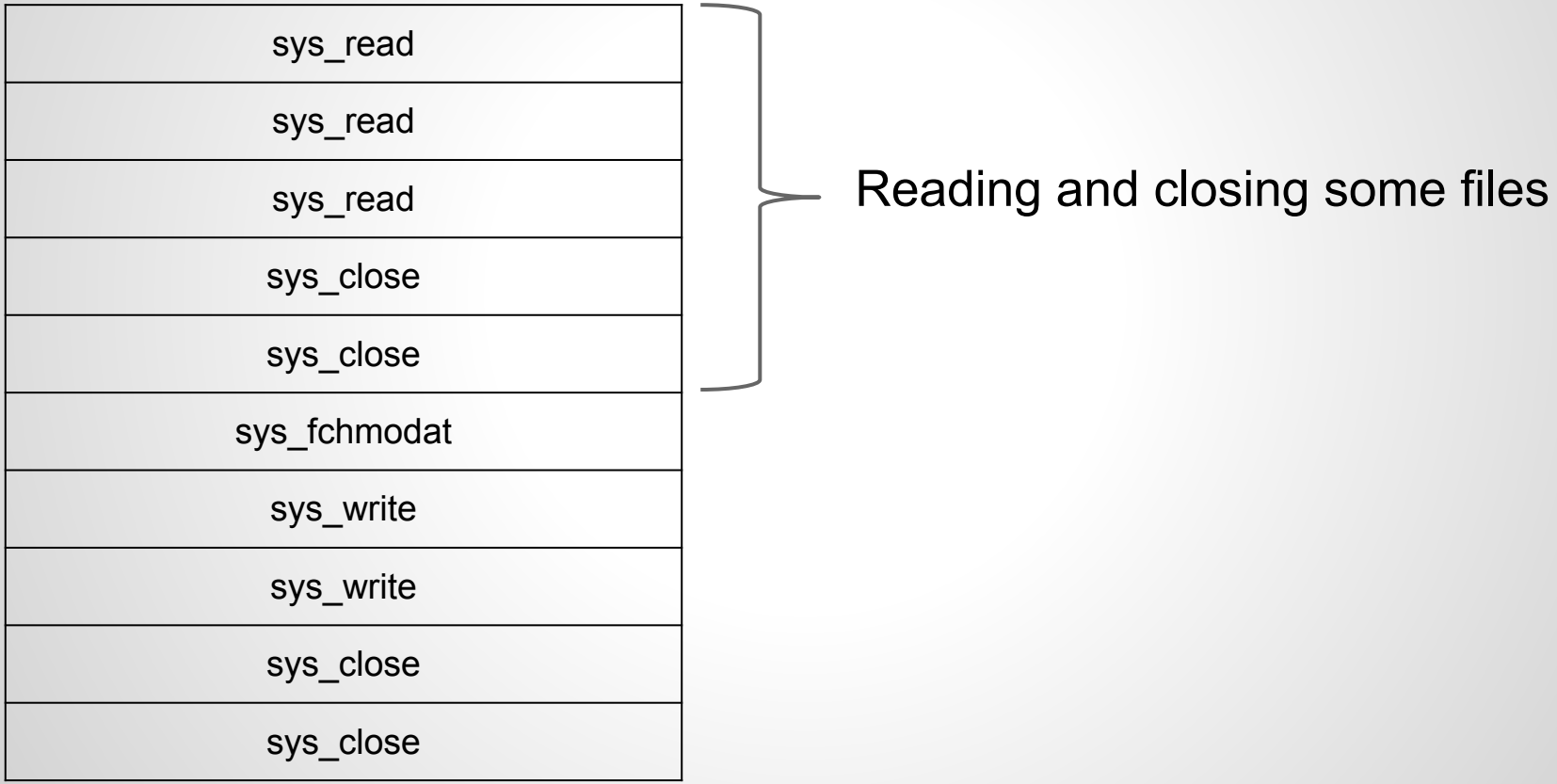
Imagine this scenario...

- You’re on an open network at a cafe and become the victim of a hacker
- The hacker attempts to read your file system and find your admin user information
- Once the hacker finds it they write a keylogger onto your computer for their benefit

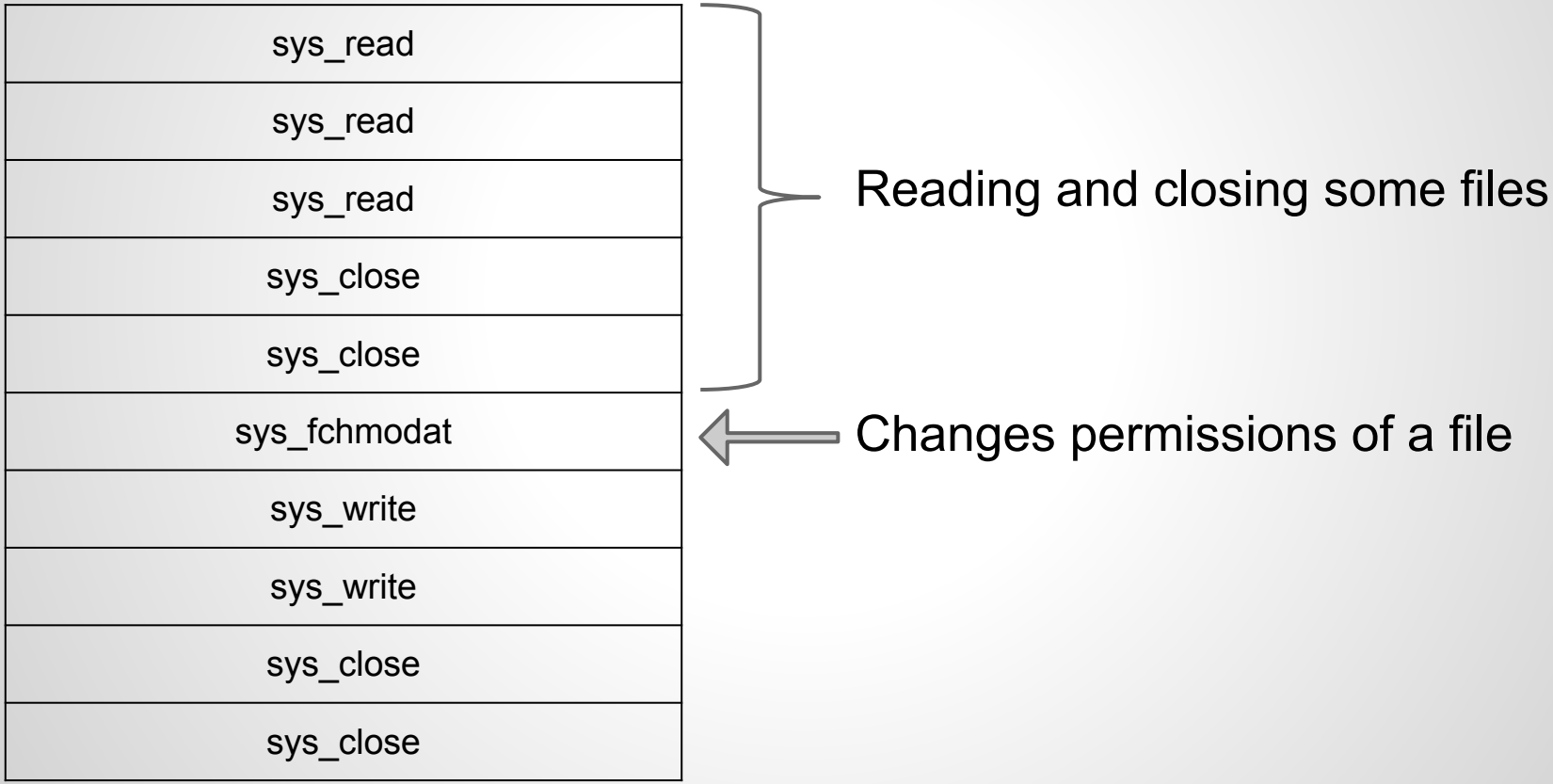
Understanding the Sequences

sys_read
sys_read
sys_read
sys_close
sys_close
sys_fchmodat
sys_write
sys_write
sys_close
sys_close

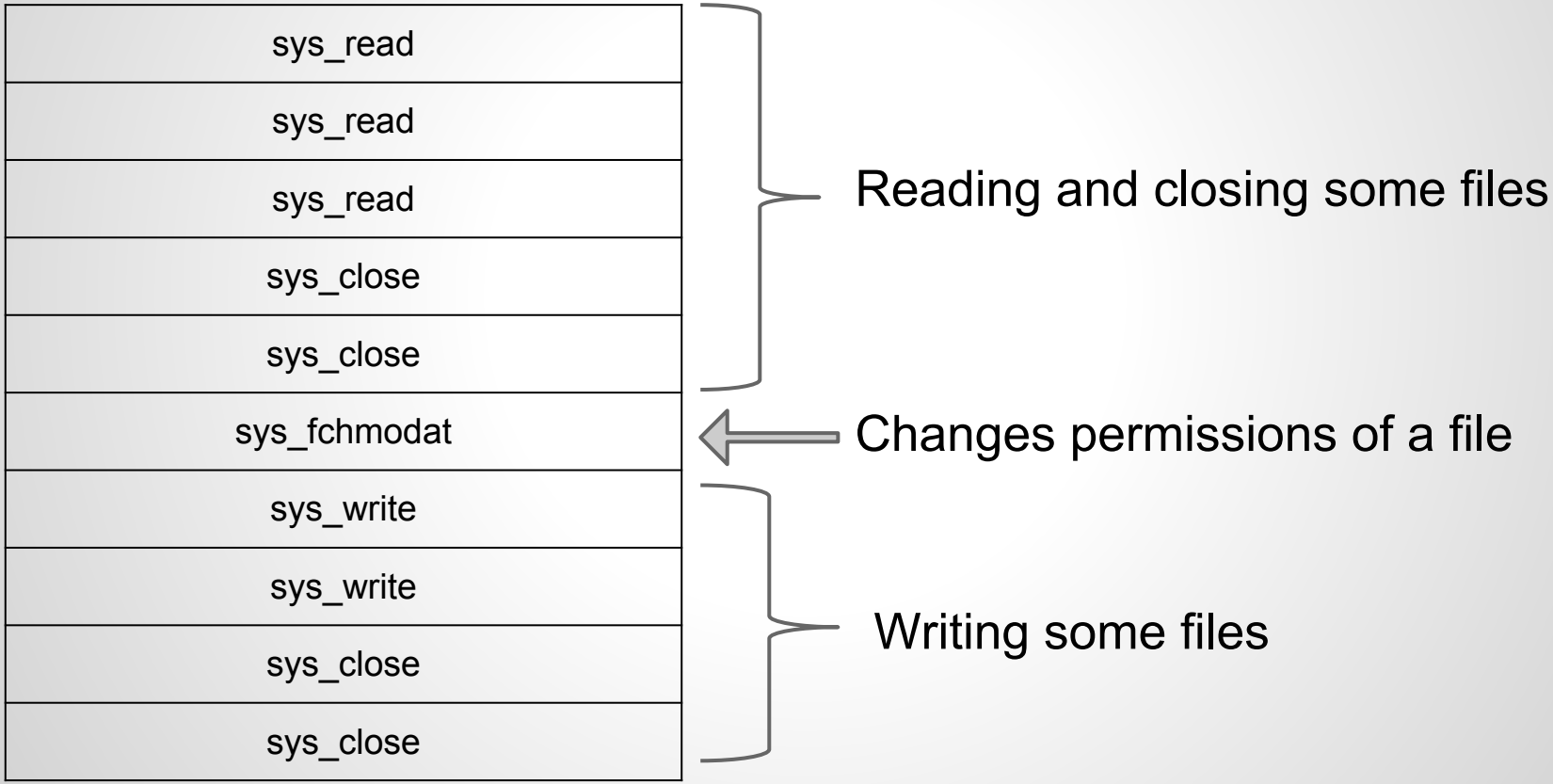
Understanding the Sequences



Understanding the Sequences



Understanding the Sequences



Understanding the Sequences

sys_read
sys_read
sys_read
sys_close
sys_close
sys_fchmodat
sys_write
sys_write
sys_close
sys_close

Understanding the Sequences

sys_read	3
sys_read	3
sys_read	3
sys_close	6
sys_close	6
sys_fchmodat	306
sys_write	4
sys_write	4
sys_close	6
sys_close	6

Understanding the Sequences

- Now that we have a sequence how can we represent it?

3
3
3
6
6
306
4
4
6
6

Understanding the Sequences

- Now that we have a sequence how can we represent it?
- **Goal**: “Learn” typical sequences used by attackers

3
3
3
6
6
306
4
4
6
6

Understanding the Sequences

- Now that we have a sequence how can we represent it?
- **Goal**: “Learn” typical sequences used by attackers
- Want to accomplish this without “learning” the exact attack sequences

3
3
3
6
6
306
4
4
6
6

Understanding the Sequences

- Now that we have a sequence how can we represent it?
- **Goal**: “Learn” typical sequences used by attackers
- Want to accomplish this without “learning” the exact attack sequences
- In other words, we want to instead find the ***anomalies*** among the sequences

3
3
3
6
6
306
4
4
6
6

Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order

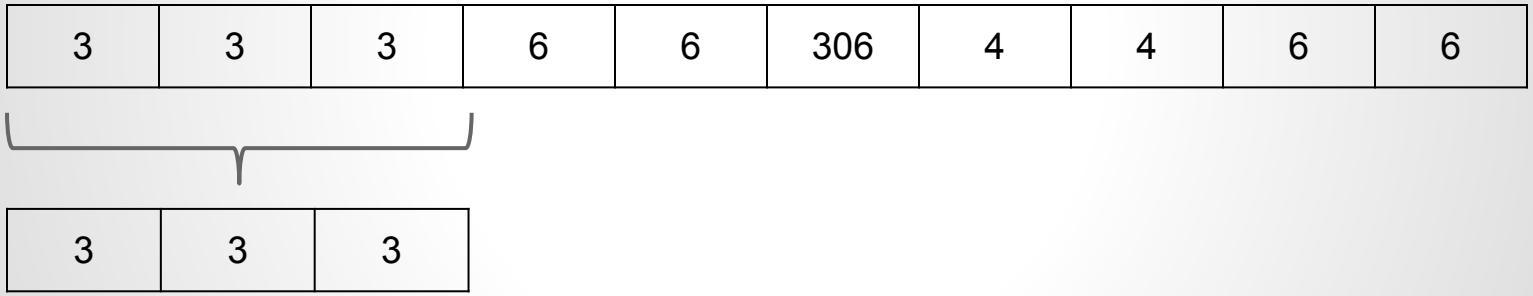
Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order.

3	3	3	6	6	306	4	4	6	6

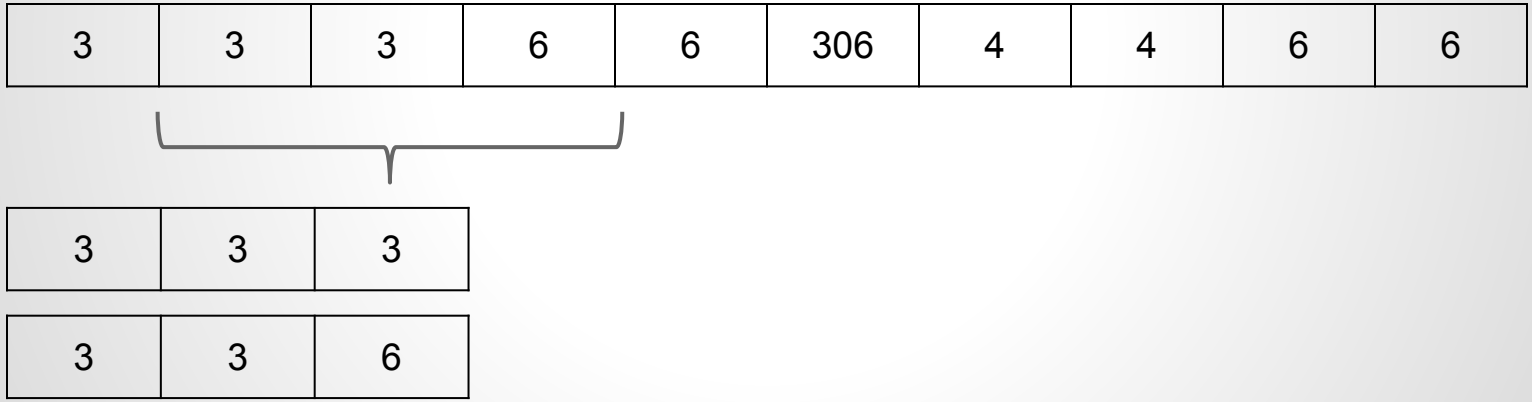
Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order.



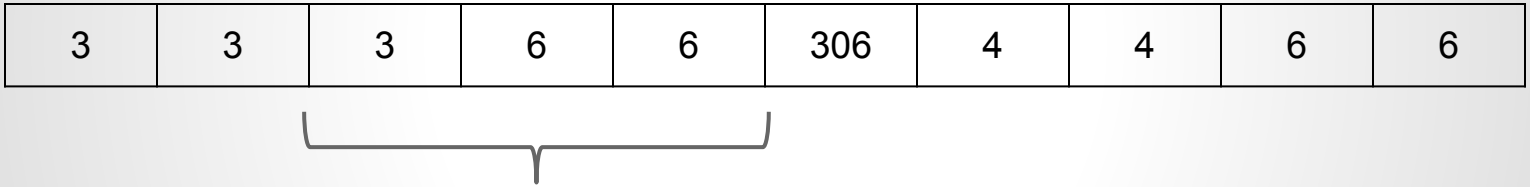
Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order.



Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order.



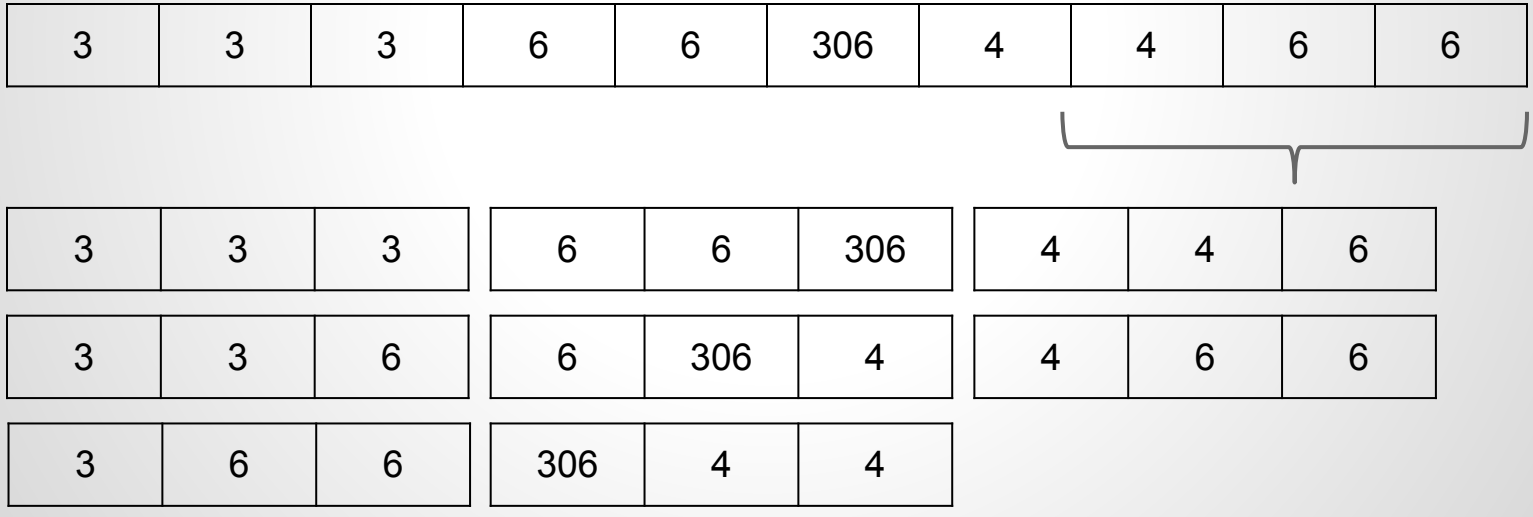
3	3	3

3	3	6

3	6	6

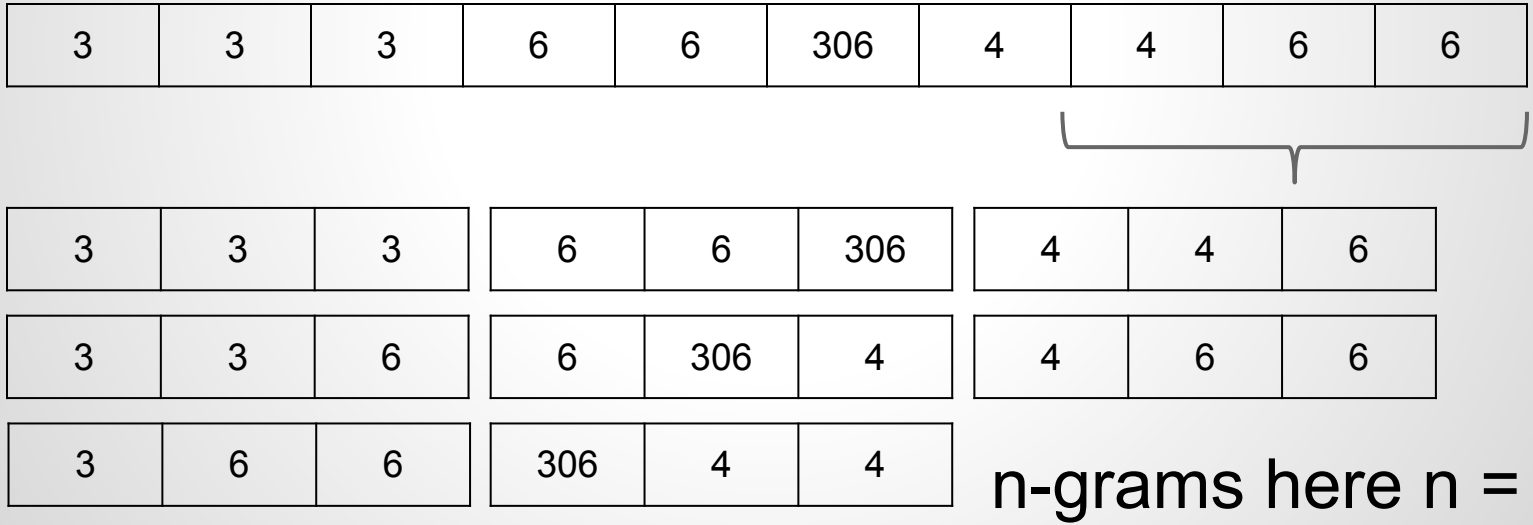
Understanding the Sequences

- **Solution:** Represent the data as smaller subsequences in the same order.

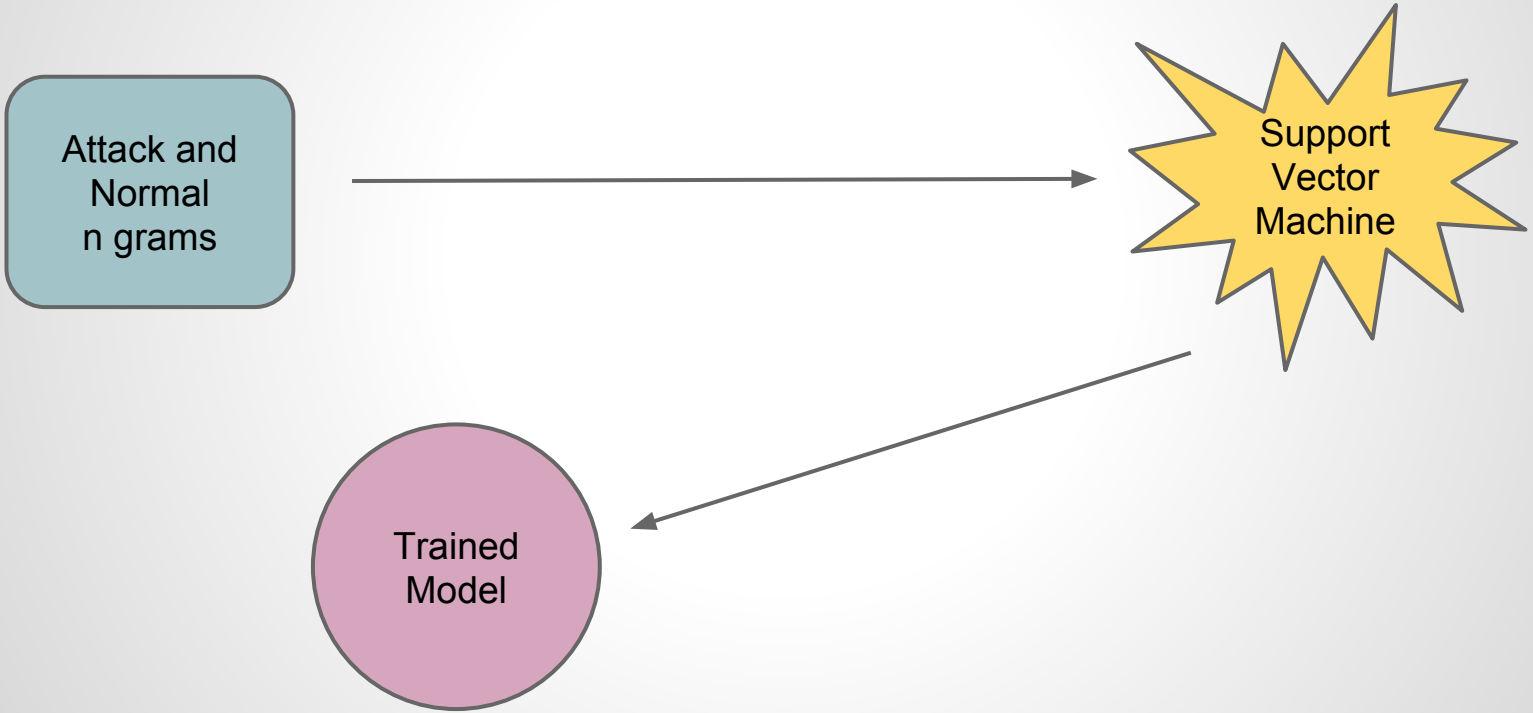


Understanding the Sequences

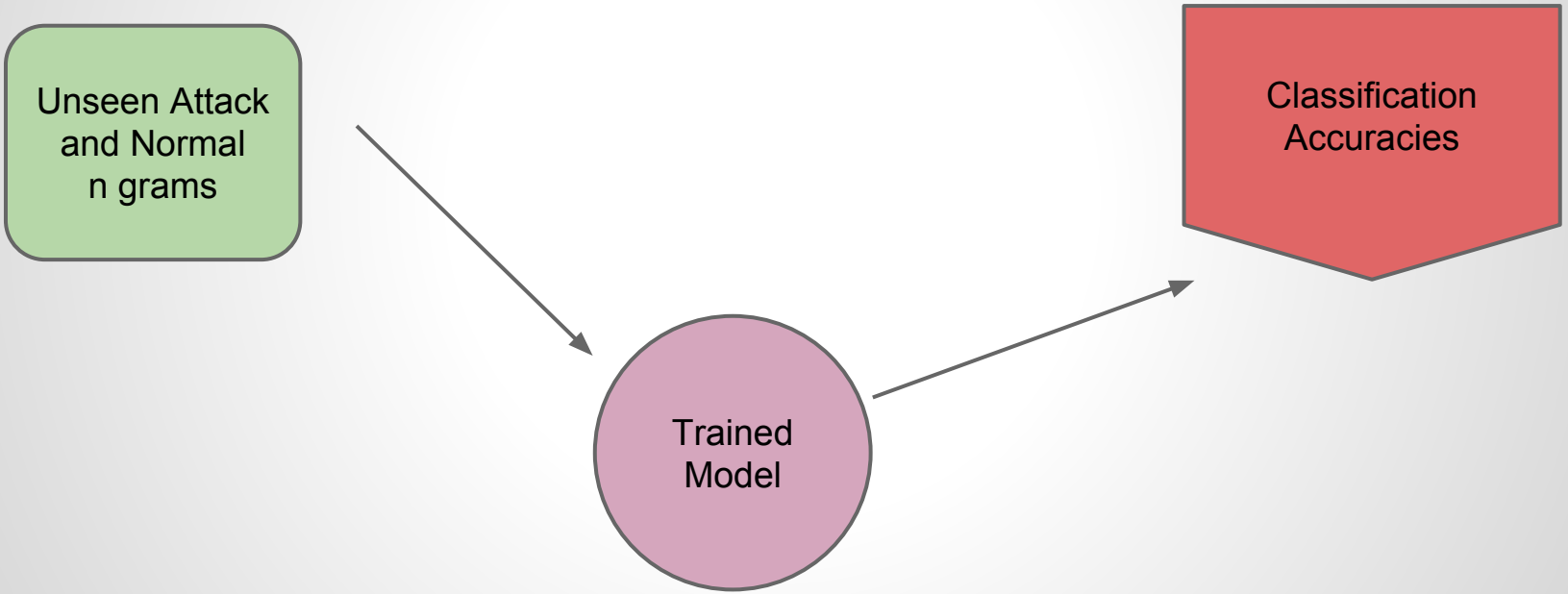
- **Solution:** Represent the data as smaller subsequences in the same order.



Training the Model



Testing the Model



Understanding the Sequences

- We have large typical attack and normal sequences

Understanding the Sequences

- We have large typical attack and normal sequences
- Instead of looking at the whole sequence we will break it down into n-grams

Understanding the Sequences

- We have large typical attack and normal sequences
- Instead of looking at the whole sequence we will break it down into n-grams
- These will be the pieces that we will learn from to ***classify*** the difference when looking at a new sequence

Understanding the Sequences

- We have large typical attack and normal sequences
- Instead of looking at the whole sequence we will break it down into n-grams
- These will be the pieces that we will learn from to ***classify*** the difference when looking at a new sequence
- **Theory:** Attack subsequences will have patterns separate from normal subsequences

Features of the Sequences

- Because there are many possible sequences, can we look to trim them down without degrading performance

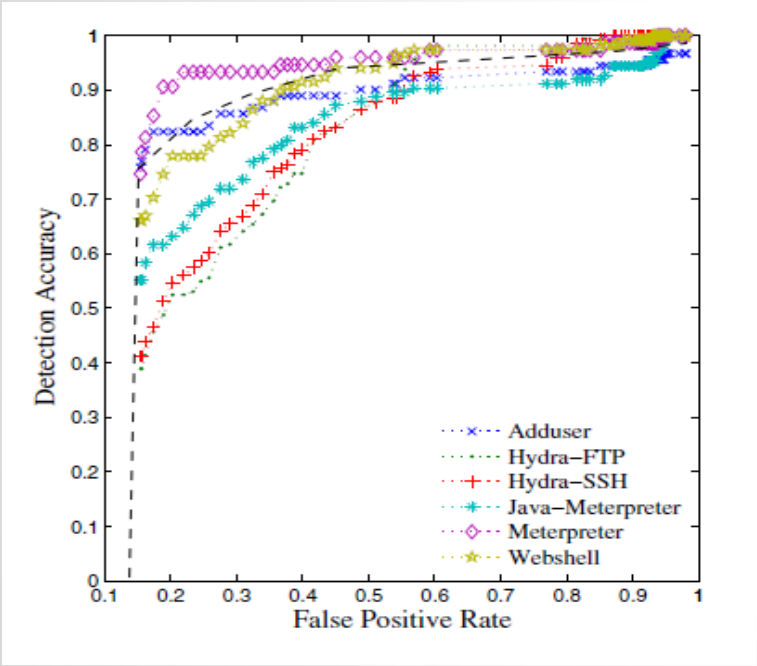
Features of the Sequences

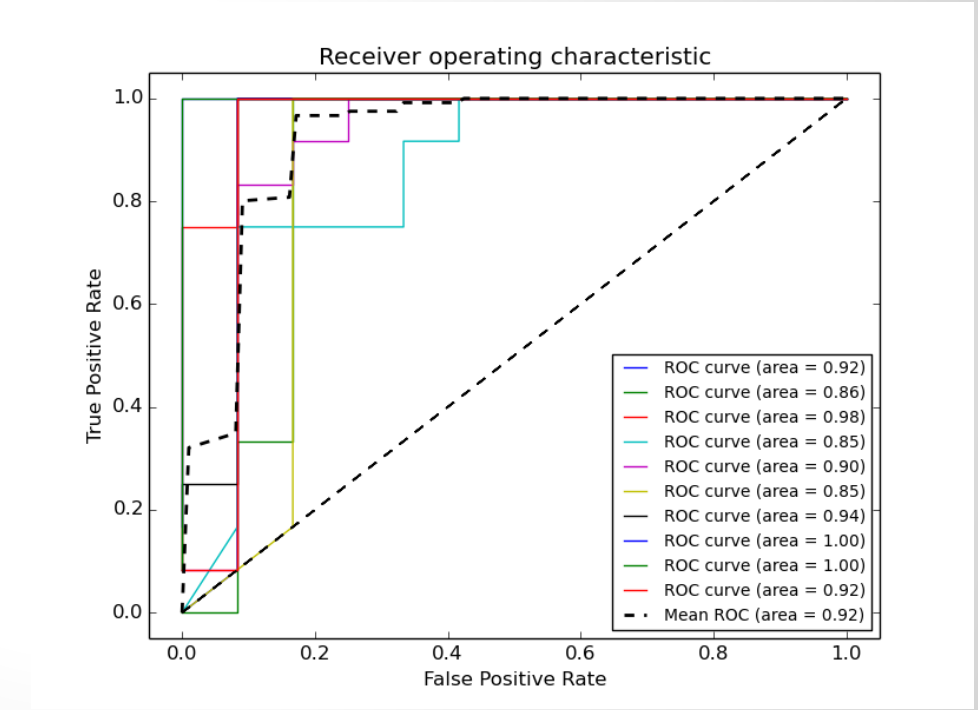
- Because there are many possible sequences, can we look to trim them down without degrading performance
- We suspect we can “combine” system calls during the preprocessing phase of the classifier

Features of the Sequences

- Because there are many possible sequences, can we look to trim them down without degrading performance
- We suspect we can “combine” system calls during the preprocessing phase of the classifier
- Aid in lowering computation as well as possibly increasing correct classification

ROC Curve Analysis

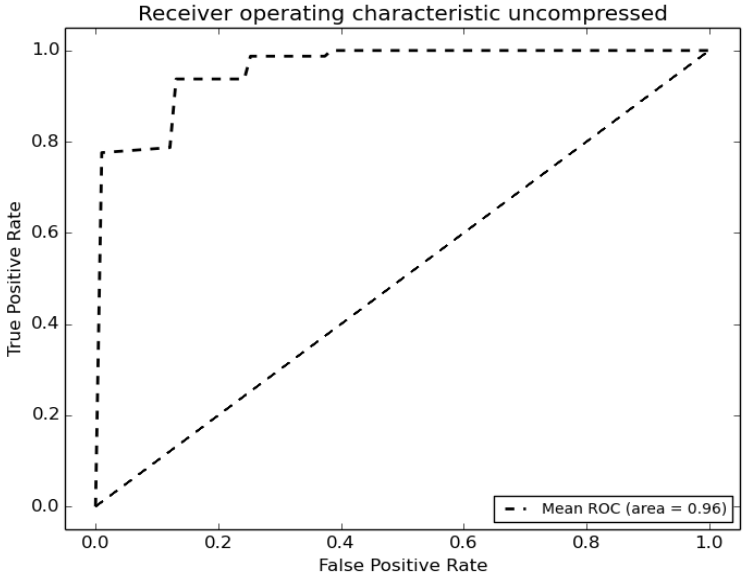


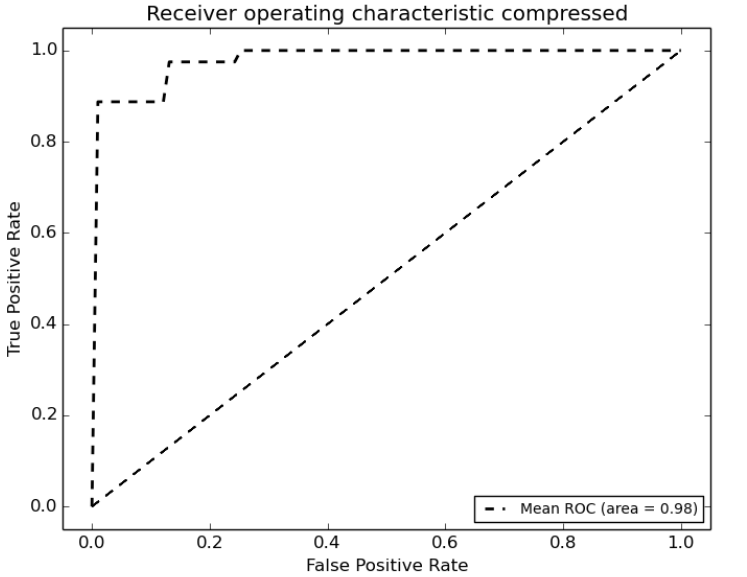


Xie et. al

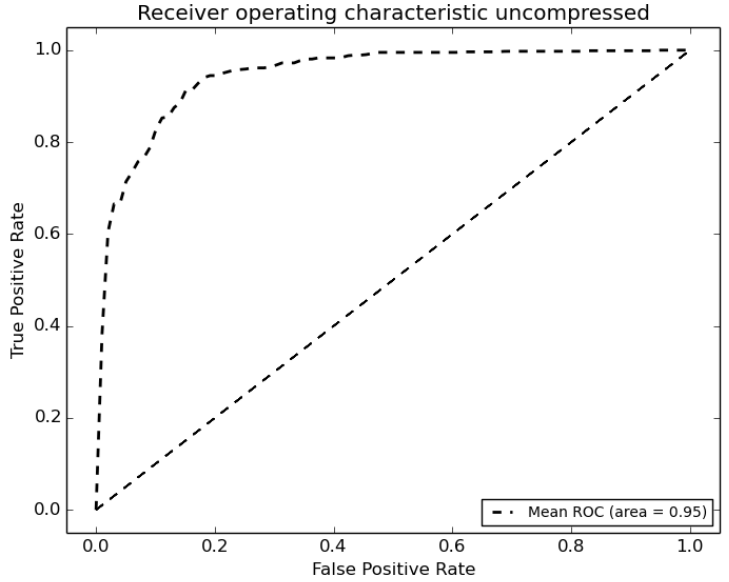
Our initial mean ROC curve

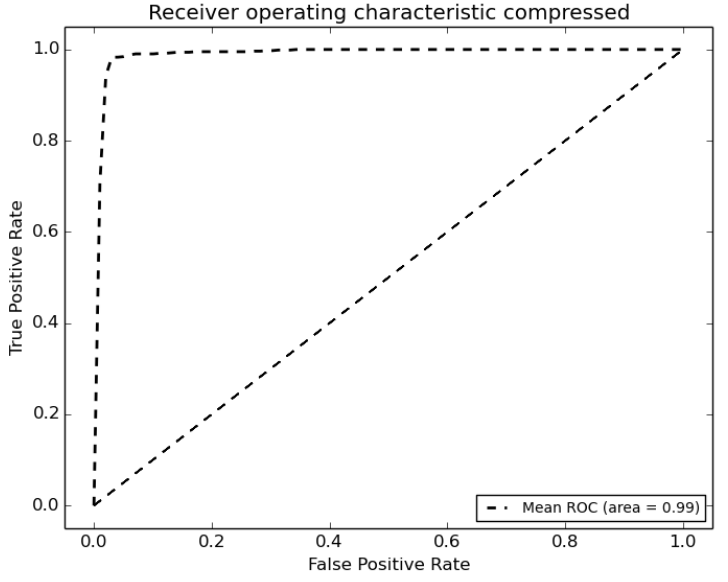
ROC Curve Analysis (subset)





ROC Curve Analysis (full dataset)





Summary

- Classified system call traces into normal and attack categories based on n-gram analysis

Summary

- Classified system call traces into normal and attack categories based on n-gram analysis
- Compressed less frequent system calls into one dummy system call

Summary

- Classified system call traces into normal and attack categories based on n-gram analysis
- Compressed less frequent system calls into one dummy system call
- Found a statistical difference between compressing versus uncompressed using subsets of the data

Future Work

- Identify when the model stops being significant when training compressed models from large datasets

- Optimize the classifier further:
  - Alter the length of our n-grams generated
  - Changing the metric for compression
  - Adjusting the classification threshold in our model

Questions...