Exceptional service in the national interest

LESSONS LEARNED AUTOMATING GENERATION OF MALWARE METADATA SIGNATURES

Joel Schott, Charles Smutz

ML/DL 2024, Sandia National Laboratories

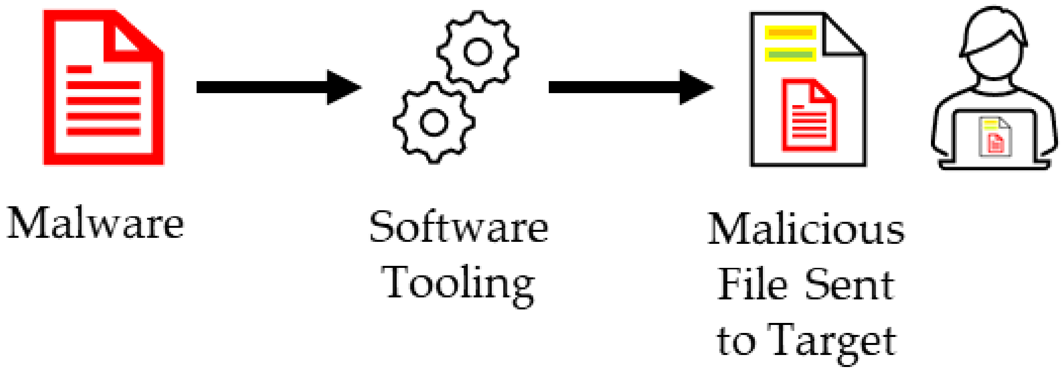
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-10421PE, SAND2024-10422V

BACKGROUND

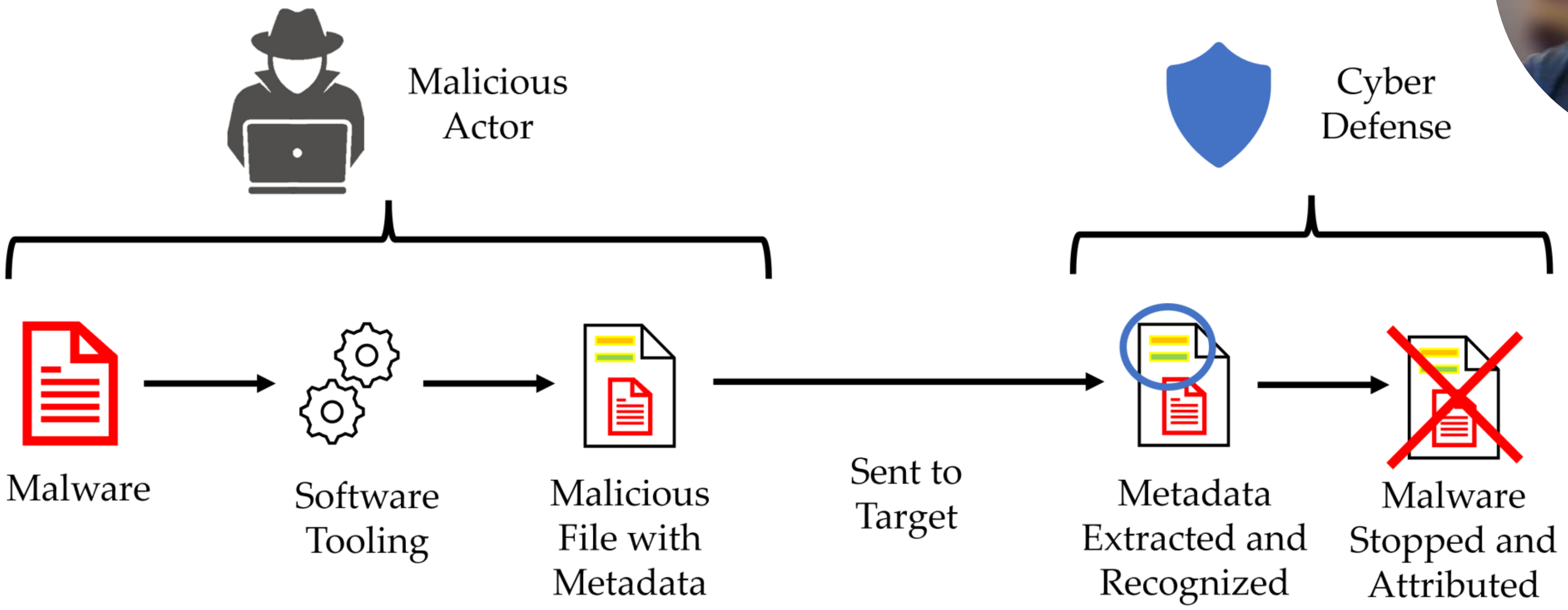
BACKGROUND

- Malware attacks from cybercrime organizations are pervasive
- These organizations have been the subject of law enforcement action
- Studying malware from these cybercrime organizations provides insights into detecting malware used in rare, targeted attacks
  - Both cybercrime organizations and perpetrators of targeted attacks use custom tools to package and send malware
  - Targeted attacks are poorly detected by commercial antivirus engines



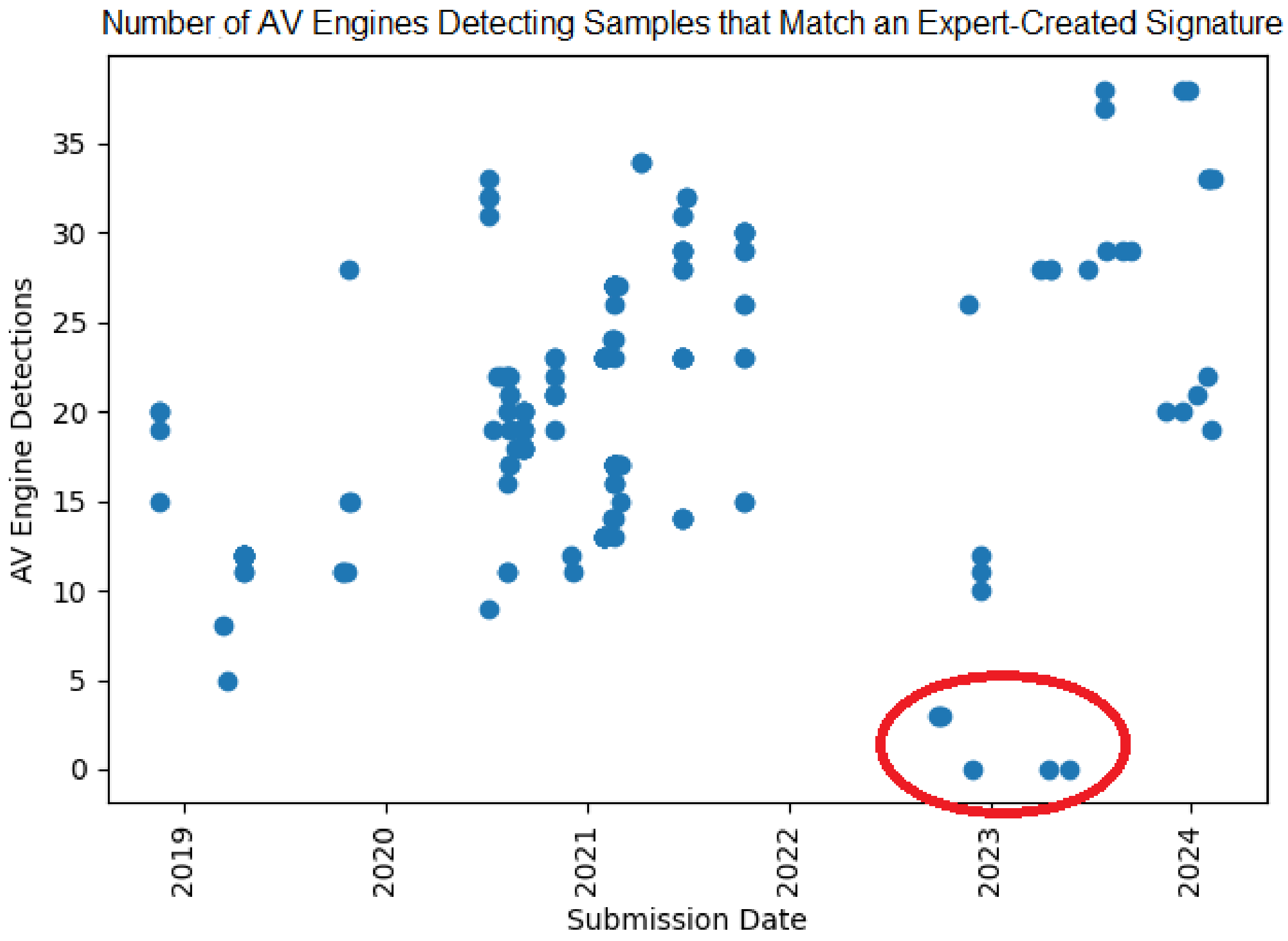
BACKGROUND

- Files or messages created and sent with custom software tools can have distinctive metadata values
- These metadata values can be used to detect malware
    - Metadata values remain distinctive even as malware evolves and changes



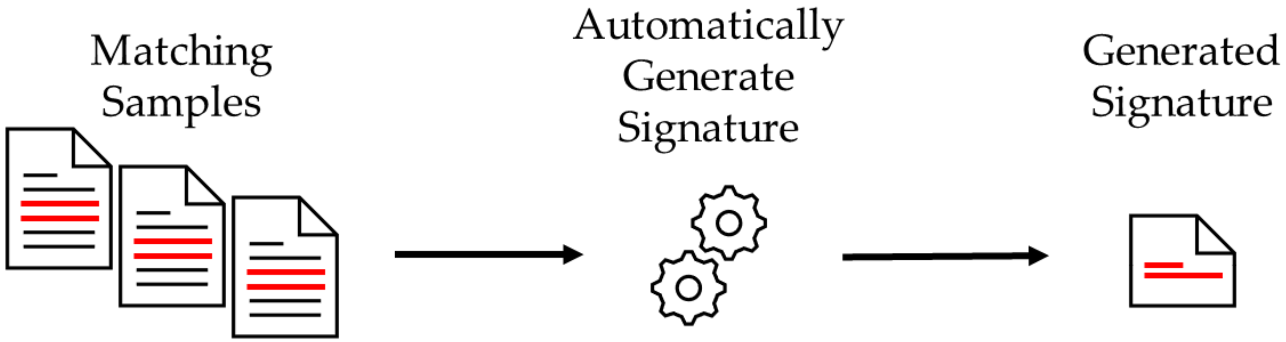
BACKGROUND

- A cybersecurity expert can create malware metadata signatures
  - More effective than commercial antivirus engines



OBJECTIVE

- Develop methods to automatically generate malware metadata signatures from malware samples



- Does not require a human expert
- Signatures can be generated more quickly
- Generated signatures are more objective
- Methods more accurately identify metadata attributes that are artifacts of tooling

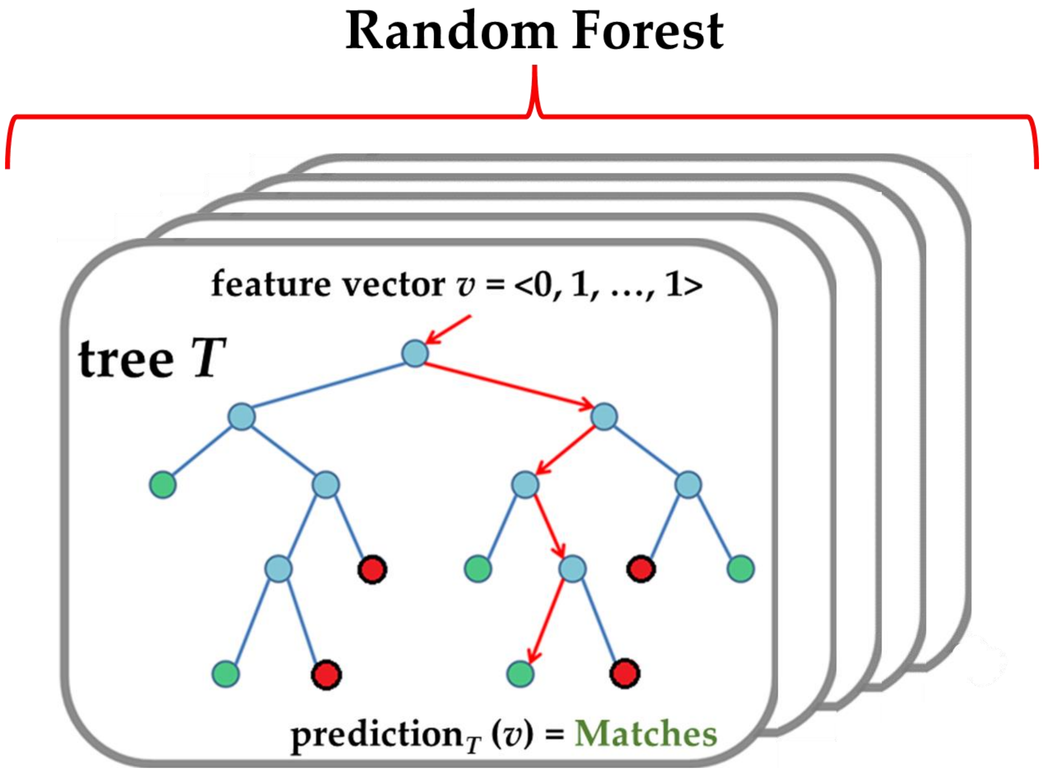
ZIP FILE SIGNATURES

ZIP FILE METADATA

- Set of features with discrete values
  - ZIP Version = 6.3
  - Method = Deflate
  - Create System = UNIX
- Signature defines possible values for a subset of the features
  - ZIP Version = 6.3 or 2.0
  - Method = Deflate

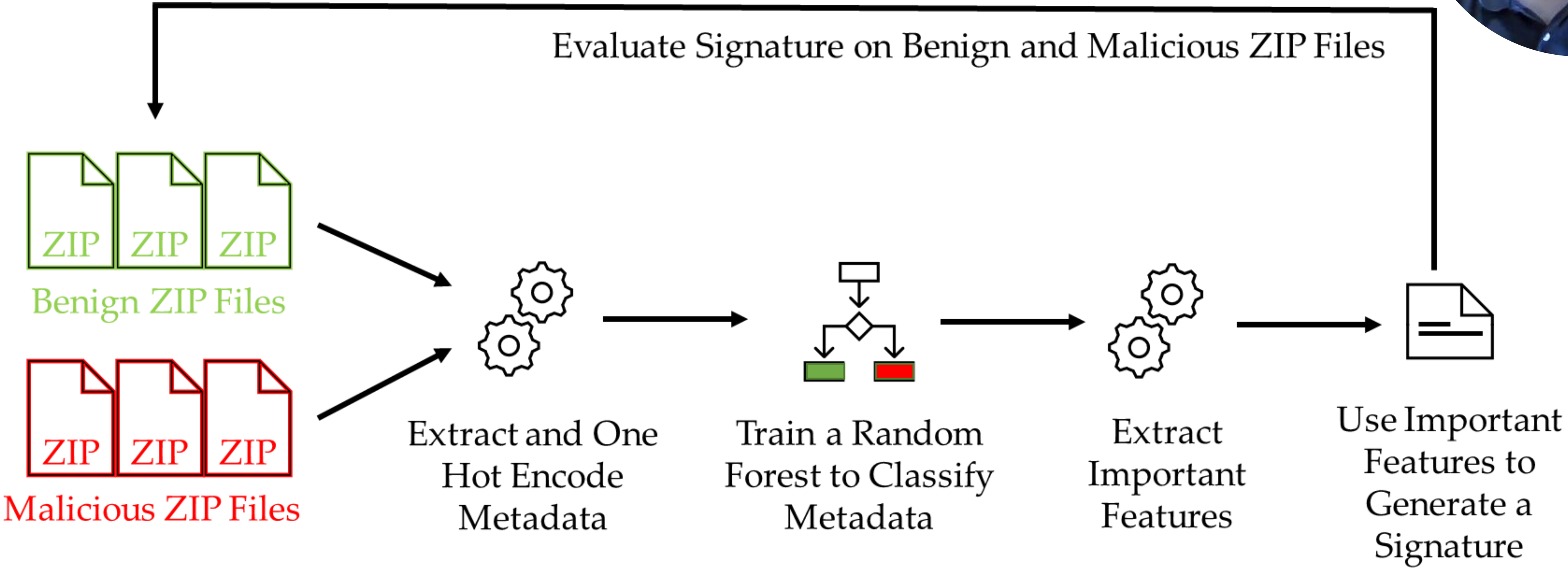
RANDOM FOREST CLASSIFIER

- Collection of decision trees
- Does not require intensive feature engineering
- Can determine feature importance



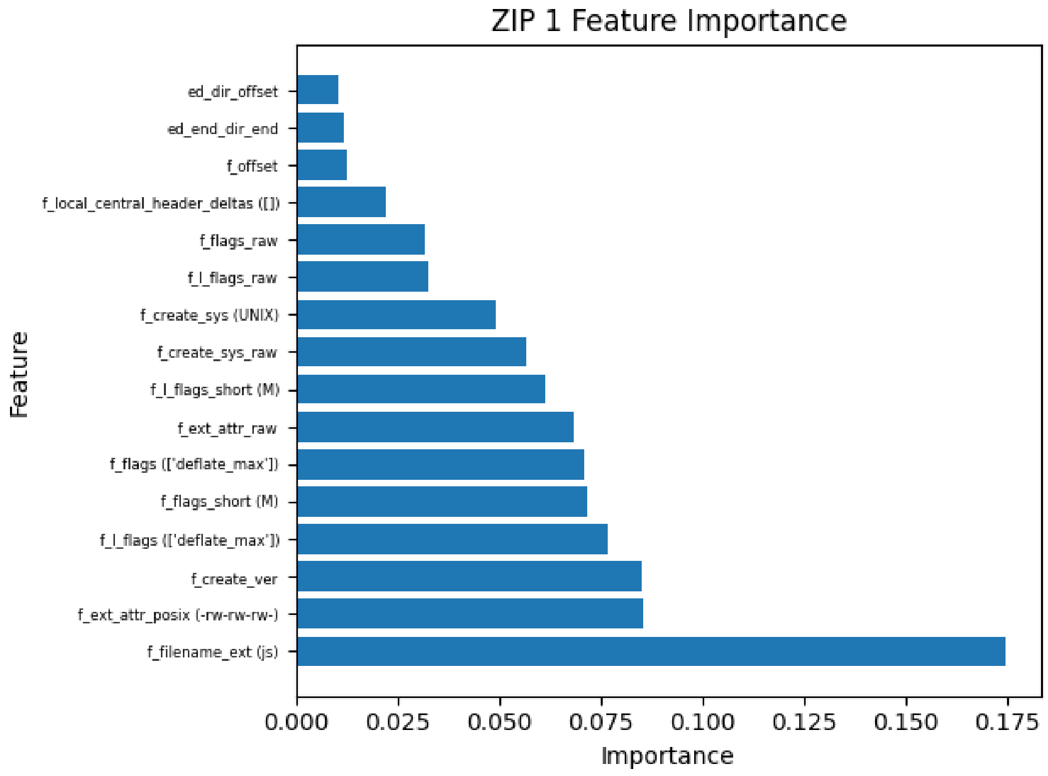
RANDOM FOREST CLASSIFIER

1. One-hot encode metadata features
2. Train a random forest to classify whether metadata came from a malicious file
3. Determine the most important metadata features
4. Use these important features for a signature



RESULTS

- Random forest classifiers and generated signatures were extremely effective
- However, signatures often included many redundant features
  - Limits signature generalization and explainability
- Difficult to determine the feature importance threshold



CONCLUSIONS

- Knowing which features are important is different from knowing which features are sufficient
- A greedy algorithm for building combinations of features created shorter, equally effective signatures

EMAIL SIGNATURES

EMAIL METADATA

- Sequences of characters
- Signatures are regular expressions that must match the entire metadata value

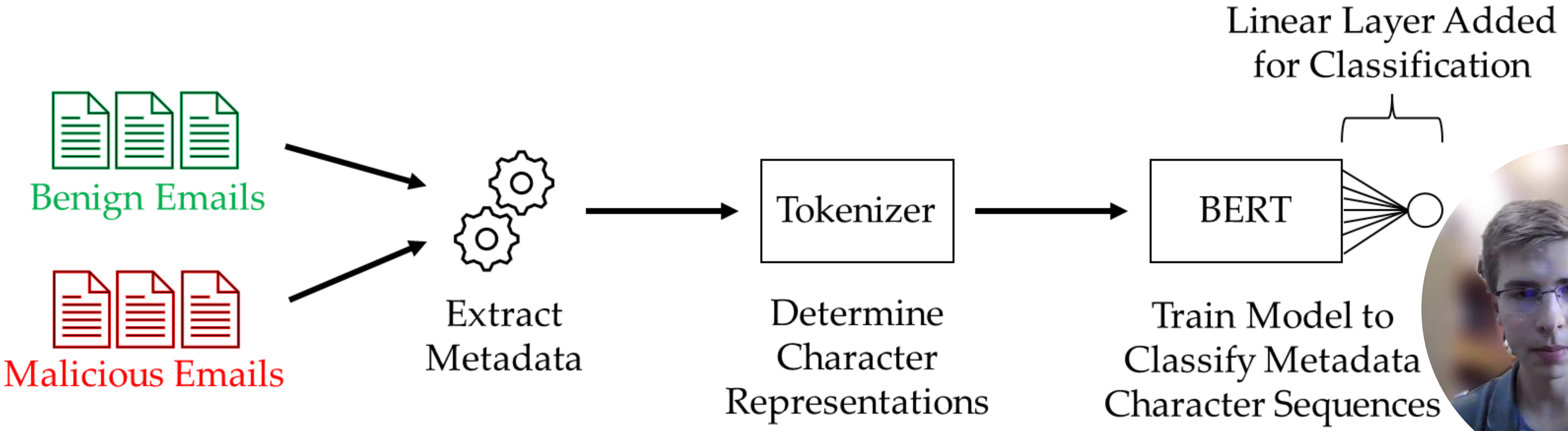
Example Metadata Value: ----=_NextPart_000_0012.08724505

Example Metadata Signature: ----=_NextPart_00[0-9]_[0-9]{1,5}\.[0-9]{1,10}

Constant   Character Class

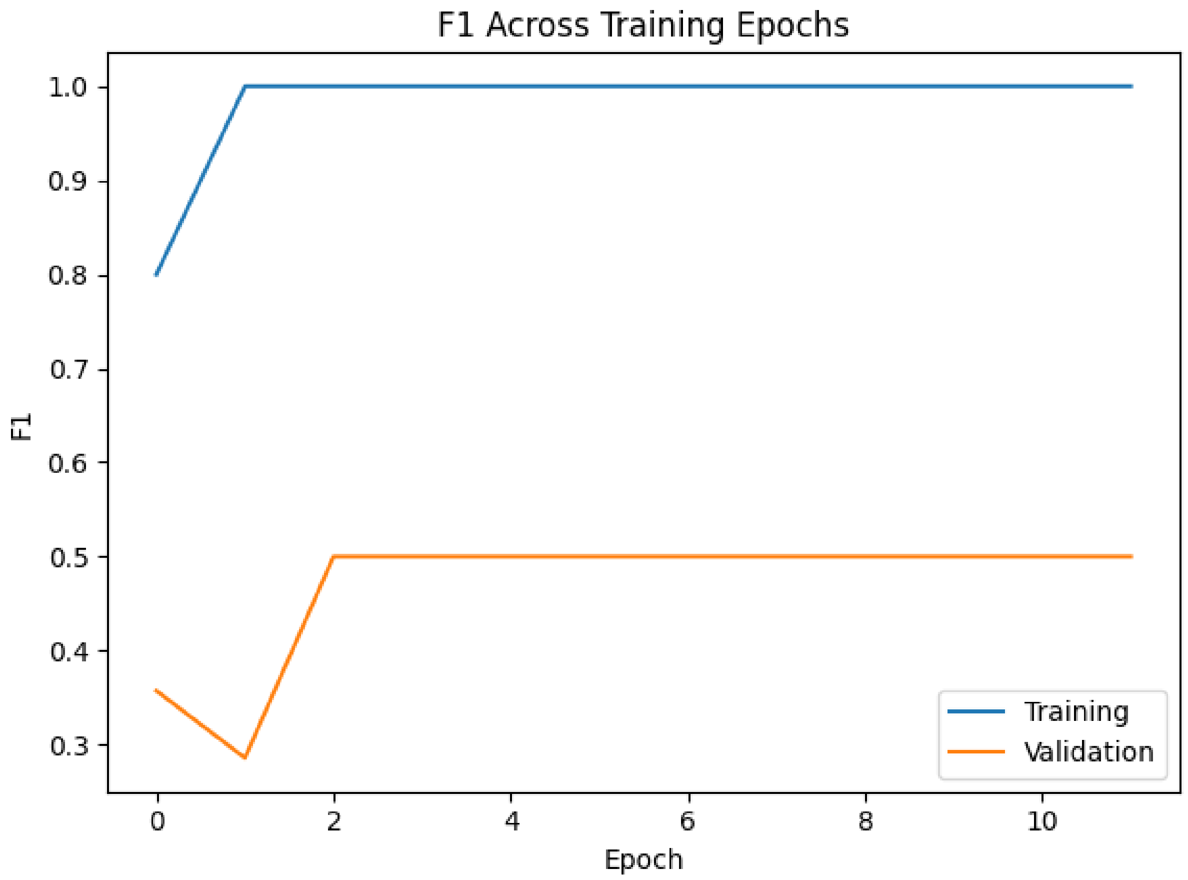
TRANSFORMER MODEL

- Train a transformer neural network to classify whether metadata came from a malicious email
  - Transformers learn context and meaning in sequential data
- Bidirectional Encoder Representations from Transformers (BERT) is a transformer architecture previously used for malware analysis
  - BERT considers both preceding and succeeding elements when learning context



RESULTS

- Performance on training data quickly neared 100% accuracy
- Poor generalization to unseen testing data



CONCLUSIONS

- Insufficient data for training a transformer model capable of generalizing
  - For some signatures, few malicious emails are available
  - Possible to generate synthetic data using the known signatures
- Signatures are based on a pattern in the entire sequence rather than the contextual meaning of the individual characters
  - A transformer model attempts to learn contextual meaning that may not exist
- An algorithm that used Multiple Sequence Alignment could create effective signatures and required few training samples
  - Required encoding knowledge specific to the domain and the signature format