CSCI 8360 - PROJECT 1 MALWARE CLASSIFICATION

TEAM ALPINE

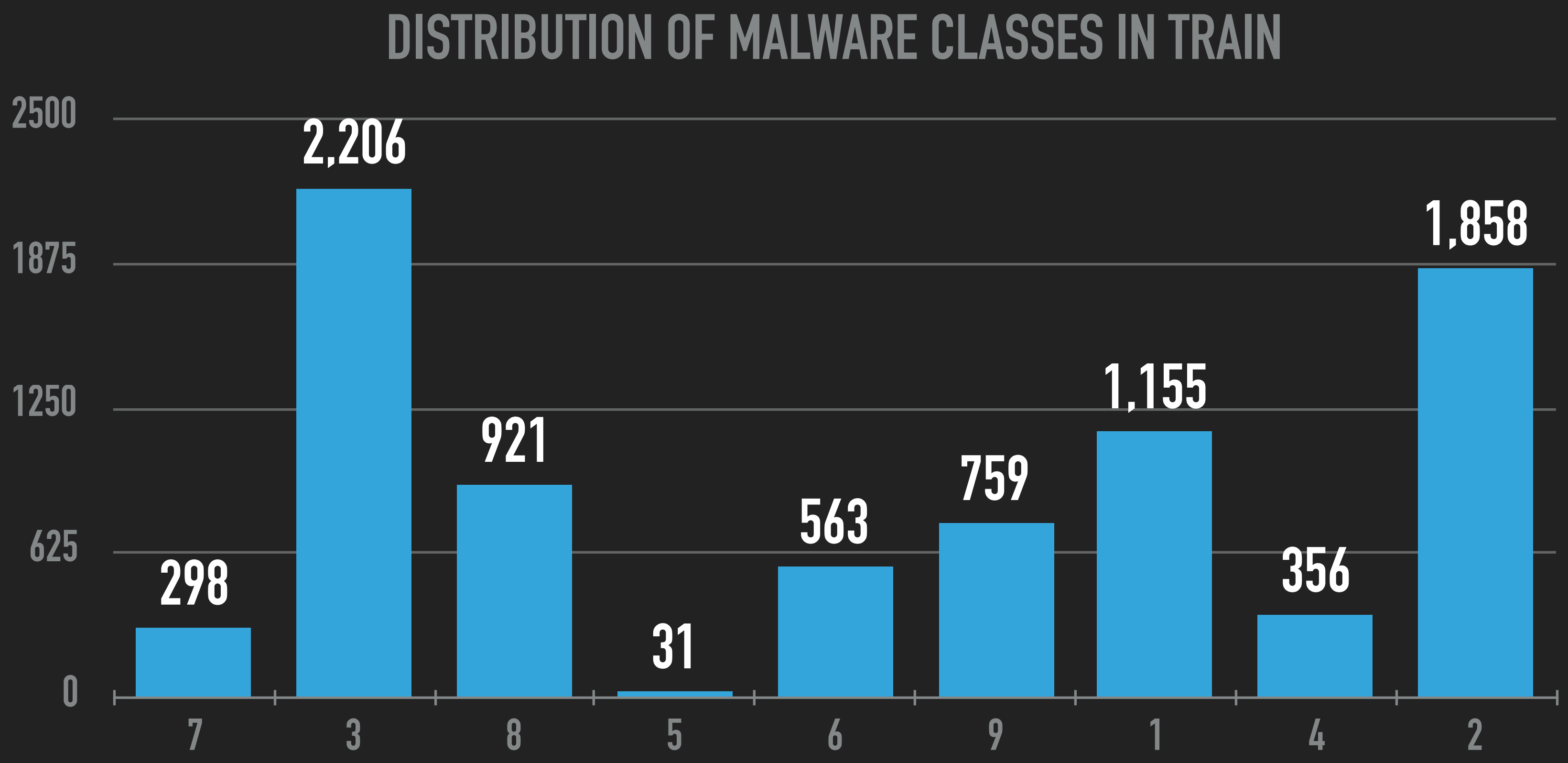
Aashish Yadavally, Hemanth Dandu, Jonathan Myers, and Sushanth Kathirvelu

APPROACHES

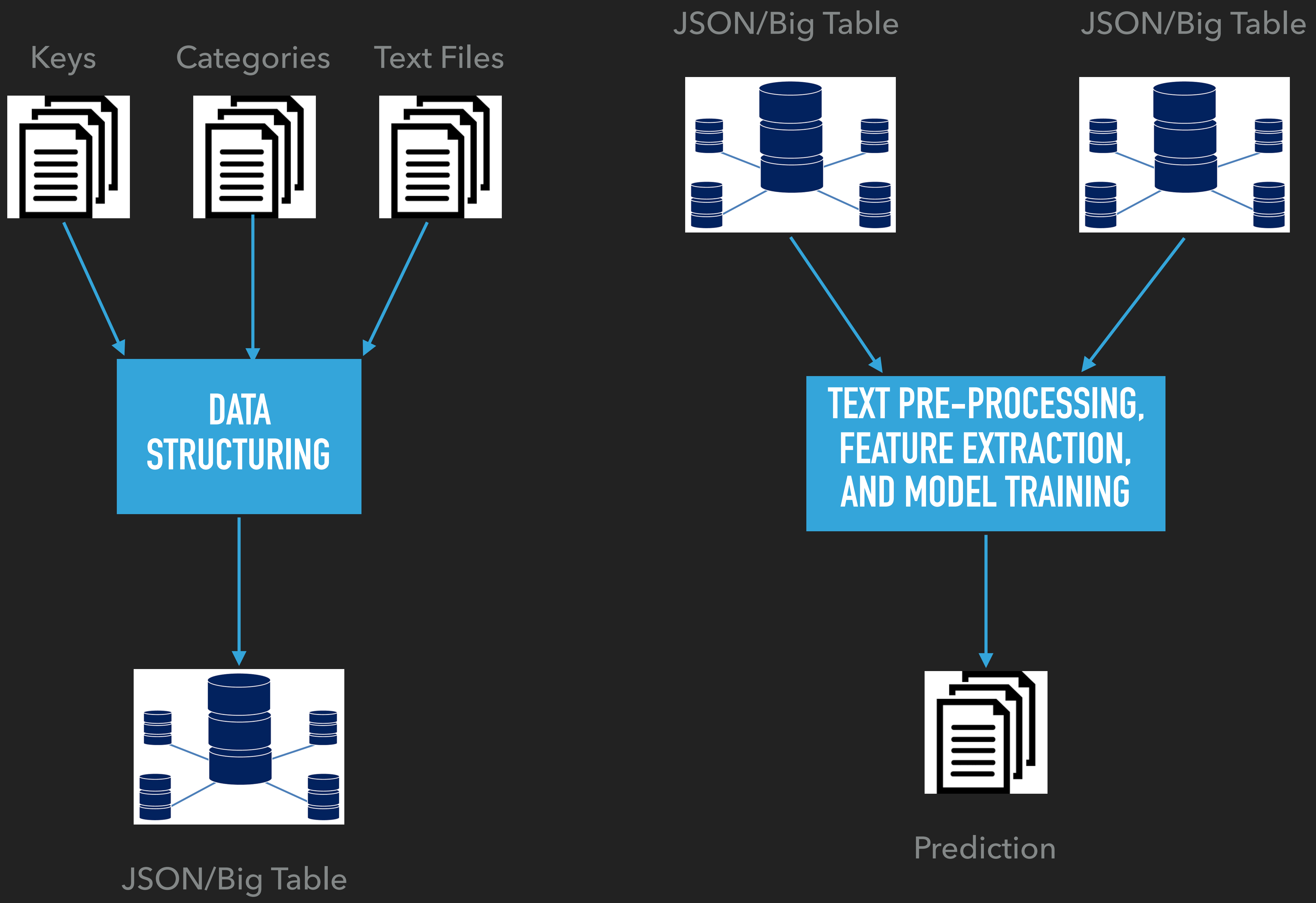
- Byte files - Extract entire text and remove line pointers
  - Text-Preprocessing: Tokenize, stopword removal(“??“,”00”), ngrams(1,2,3,4),Word count, IDF, PCA(5,20,30)
  - ML Models : Naive Bayes, Logistic Regression(with and without cross-validation), RandomForest, Support Vector Machines
  - Accuracies between 75% and 92% on test set
- Asm files - Extract only the first words from each line
  - Text-Preprocessing: Tokenize, Word count, IDF, PCA(5,20)
  - ML Models: Logistic Regression, RandomForest
  - Accuracies between 70% and 94% on test set
- Bytes + Asm - Concatenate text from bytes and asm files

TEAM ALPINE

- Word2Vec word embeddings
  - Cannot use existing W2V models
  - Too much training time on small dataset itself. Poor accuracy.
- Oversampling to handle class imbalance
  - Replicating instances with low class counts
  - Negligible impact on accuracy



SOFTWARE DESIGN



Keys

Categories

Text Files

DATA STRUCTURING

JSON/Big Table

JSON/Big Table

JSON/Big Table

TEXT PRE-PROCESSING, FEATURE EXTRACTION, AND MODEL TRAINING

Prediction

HIGHEST ACCURACY

- Bytes + Asm concatenation
    - Text Processing: Tokenization, Word Counts
    - ML algorithm : RandomForest(30 trees, 15 max depth)
    - Accuracy : 98.75%

THANK YOU ! QUESTIONS ?

Presented by Hemanth Dandu