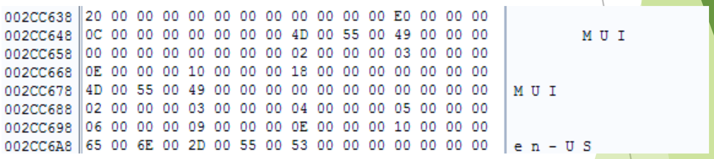
Scalable Malware Classification

Ankita J., Layton H., Omid S.

What Do People Do ?

- System (Running) level behavior (Dynamic Analysis)
- File (Binary) level information (Static Analysis)
  - Sizes
  - Code Changes
  - String Resource
  - Segment Sizes
  - Function Uses
  - Library Includes



```
002CC638 ||20 00 00 00 00 00 00 00 00 00 00 E0 00 00 00 |
002CC648 ||0C 00 00 00 00 00 00 00 4D 00 55 00 49 00 00 00 |        M U I
002CC658 ||00 00 00 00 00 00 00 00 02 00 00 00 03 00 00 00 |
002CC668 ||0E 00 00 00 10 00 00 00 18 00 00 00 00 00 00 00 |
002CC678 ||4D 00 55 00 49 00 00 00 00 00 00 00 00 00 00 00 |M U I
002CC688 ||02 00 00 00 03 00 00 00 04 00 00 00 05 00 00 00 |
002CC698 ||06 00 00 00 09 00 00 00 0E 00 00 00 10 00 00 00 |
002CC6A8 ||65 00 6E 00 2D 00 55 00 53 00 00 00 00 00 00 00 |e n - U S
```

Our Feature Selection

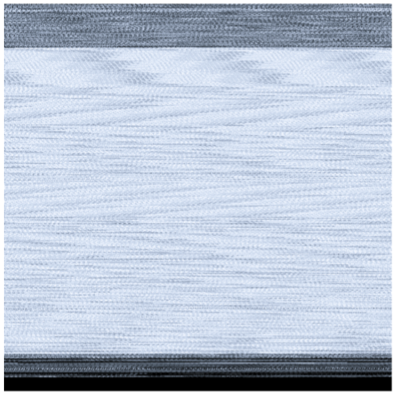
Encode malware binary into images

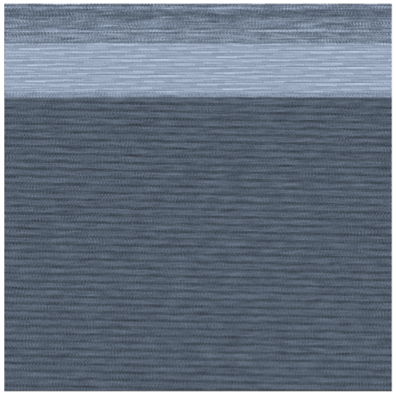
- Read bytes into pixels
- Pad with zero
- Make Square

Enrich the images with features of the decompiled binary

- Segment Sizes
- Function Count
- Library Imports
- Opcode N-grams (1-4)

Sample Images





The Network

- We trained a Convolutional Network :
  - Convolution 1 ->Convolution 2 ->Convolution 3 ->Convolution 4 ->Max Pooling 1 ->Convolution 5 ->Dense
- We Used only images in final results
- We achieved AutoLab accuracy of 93.3