MICHIGAN STATE
UNIVERSITY

Project Plan
Defeating Malware Payload Obfuscation

The Capstone Experience

Team Proofpoint

Nick Lojewski
Adam Johanknecht
Dan Somary
Vivian Qian
Derek Renusch

Department of Computer Science and Engineering
Michigan State University

Spring 2019

From Students…
…to Professionals

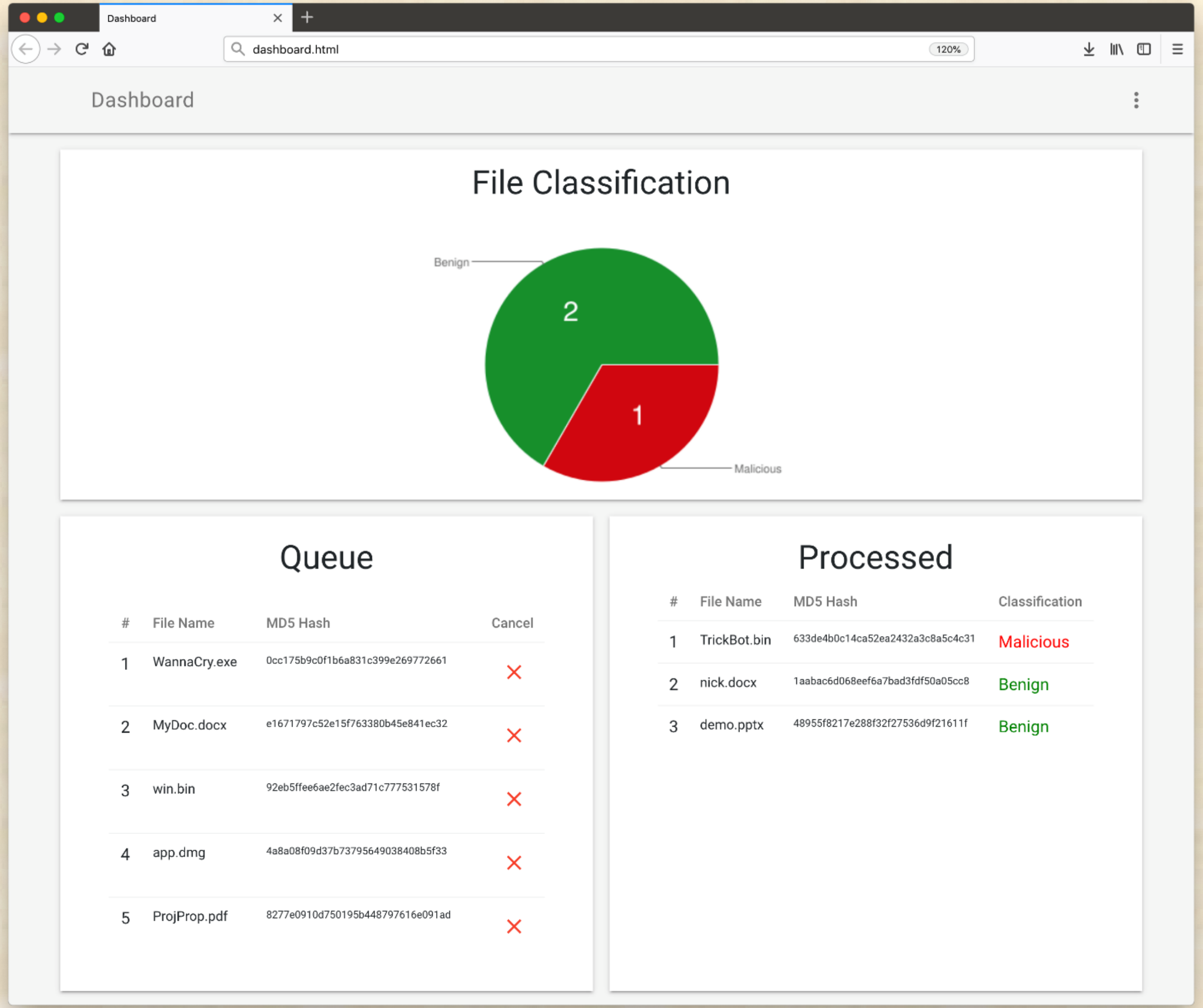
Functional Specifications

- Create a machine learning system to classify files as malicious or benign
  - Accuracy goal: have at least the same accuracy as sandbox detonation
  - Performance goal: be at least 50% faster than detonation in Cuckoo
- Display information in web dashboard
  - High level system information
  - Ability to look at details for individual files

Design Specifications

- System Overview
  - Files will be placed in the queue by Proofpoint’s process
  - Extract file metadata and feed that into Machine Learning algorithm
  - Machine Learning algorithm will classify the file as benign or malicious
- Web Application
  - System Overview
  - Display detailed file information
  - System Health
- Machine Learning Framework
  - Train a Machine Learning algorithm to accurately detect malicious files
  - Determine characteristics of files that point to malicious behavior
  - Malware that can’t be classified will be detonated in Cuckoo

Screen Mockup: Job Pipeline



Dashboard

File Classification

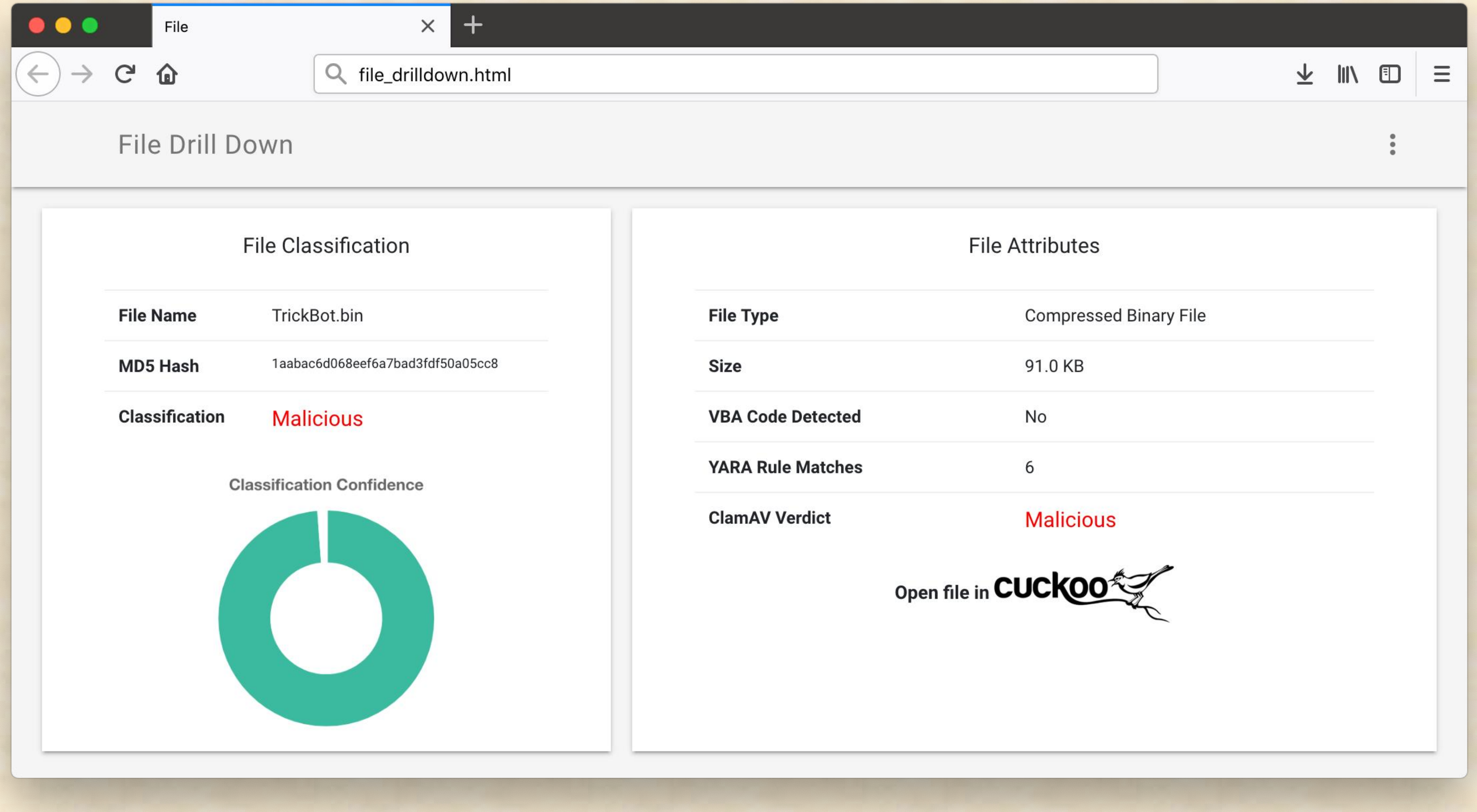
Queue

#	File Name	MD5 Hash	Cancel
1	WannaCry.exe	0cc175b9c0f1b6a831c399e269772661	✕
2	MyDoc.docx	e1671797c52e15f763380b45e841ec32	✕
3	win.bin	92eb5ffee6ae2fec3ad71c777531578f	✕
4	app.dmg	4a8a08f09d37b73795649038408b5f33	✕
5	ProjProp.pdf	8277e0910d750195b448797616e091ad	✕

Processed

#	File Name	MD5 Hash	Classification
1	TrickBot.bin	633de4b0c14ca52ea2432a3c8a5c4c31	Malicious
2	nick.docx	1aabac6d068eef6a7bad3fdf50a05cc8	Benign
3	demo.pptx	48955f8217e288f32f27536d9f21611f	Benign

Screen Mockup: File Drill Down



File Drill Down

File Classification

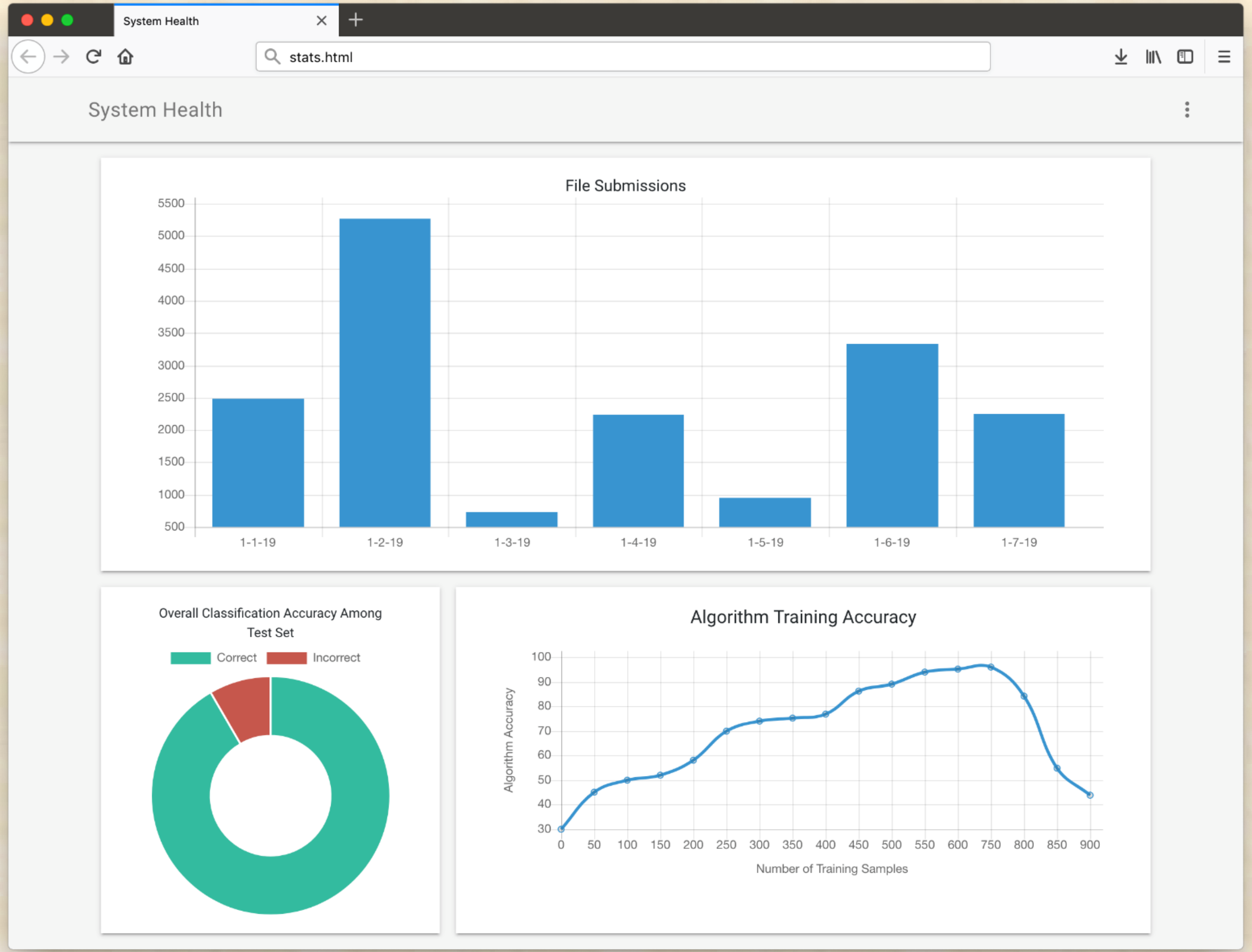
File Name	TrickBot.bin
MD5 Hash	1aabac6d068eef6a7bad3fdf50a05cc8
Classification	Malicious

Classification Confidence

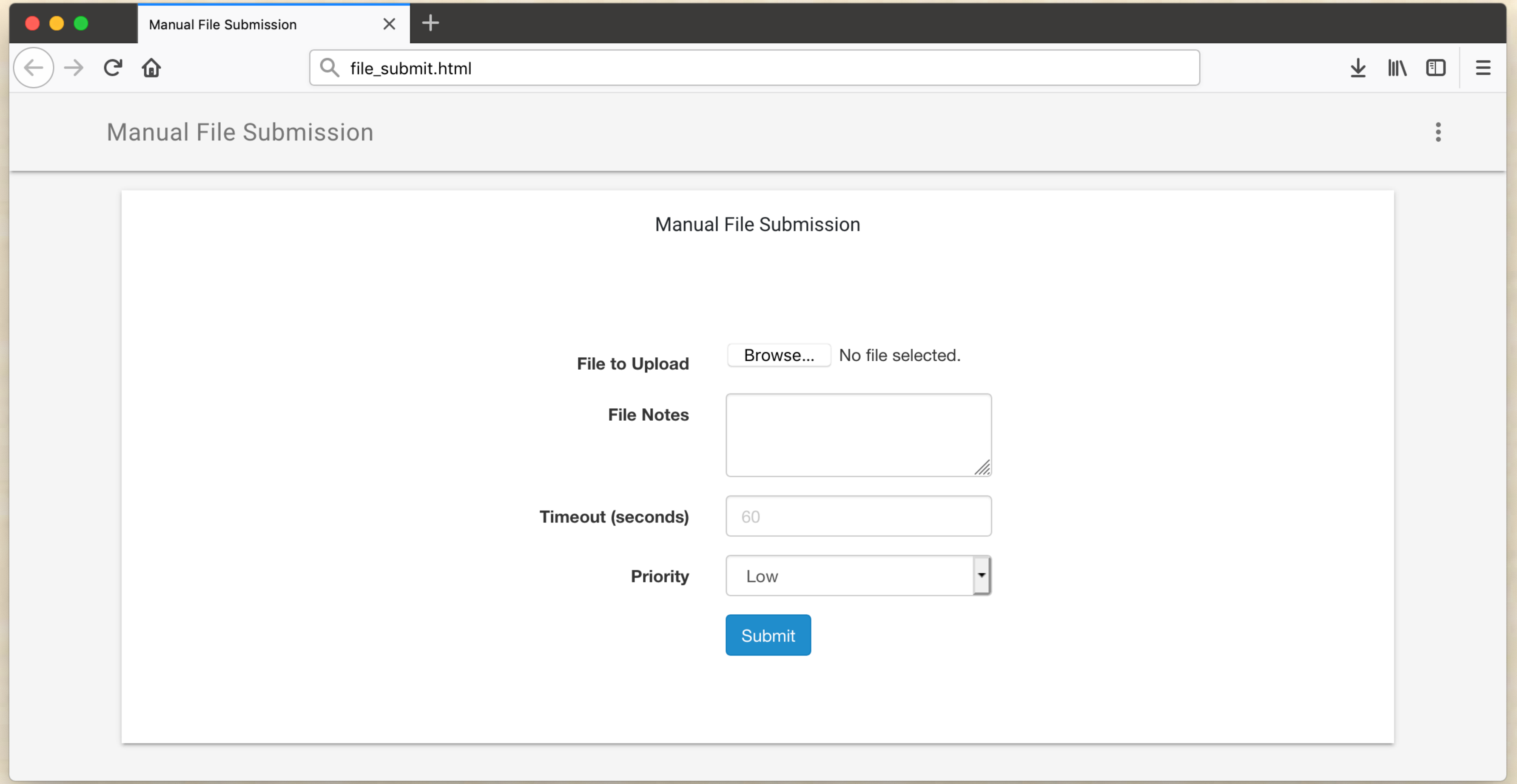
File Attributes

File Type	Compressed Binary File
Size	91.0 KB
VBA Code Detected	No
YARA Rule Matches	6
ClamAV Verdict	Malicious

Screen Mockup: System Health



Screen Mockup: Manual File Submission



Manual File Submission

File to Upload: Browse... No file selected.

File Notes:

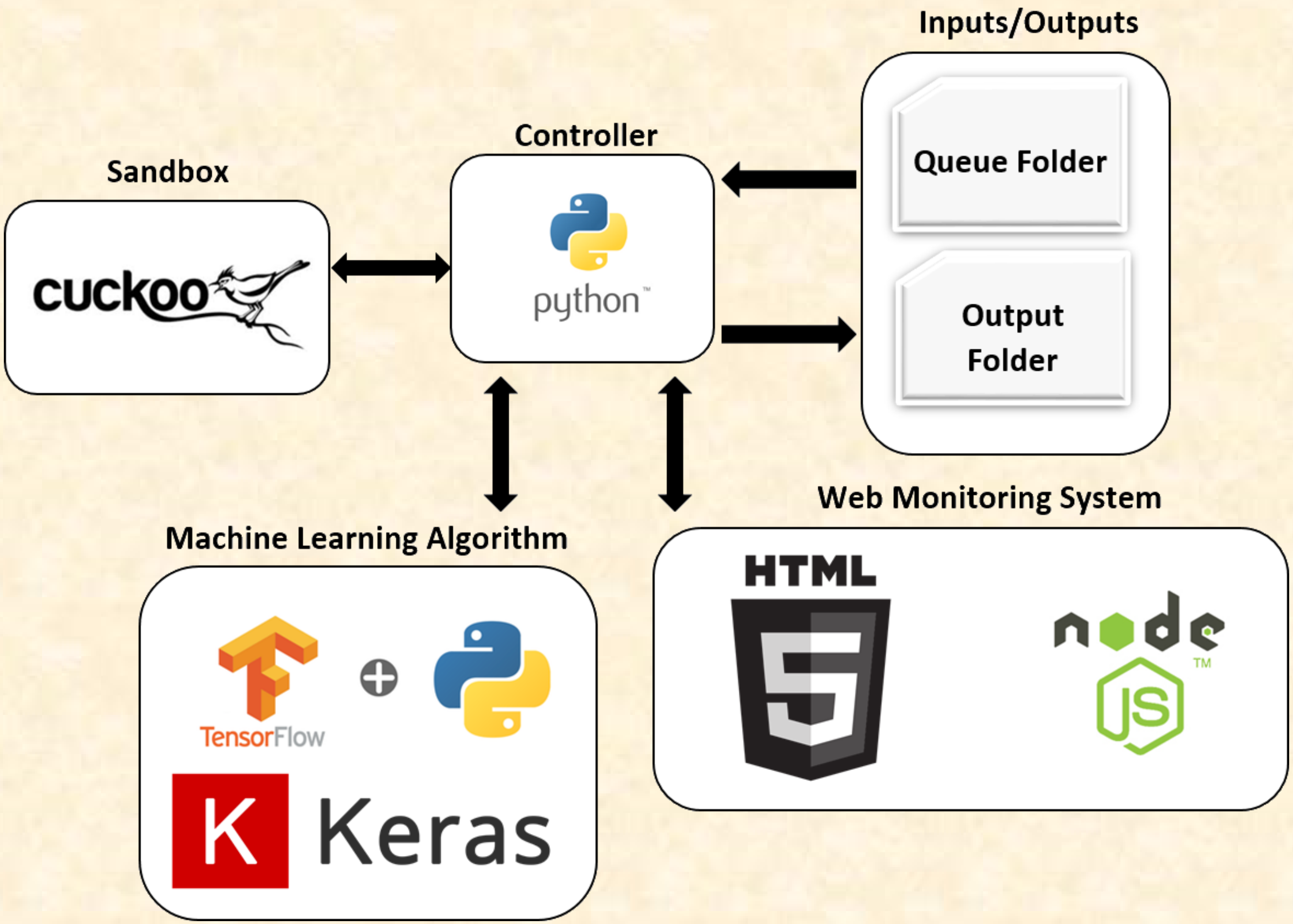
Timeout (seconds): 60

Priority: Low

Technical Specifications

- Backend Analysis
  - Python controller is used to collect and determine file types and other characters
  - Yara and pefile for initial file attribute collection on file samples
  - TensorFlow and Keras are used as Machine Learning Algorithms
  - Cuckoo is used if Machine Learning Algorithms cannot find definitive classification
- Frontend
  - HTML5 and Bootstrap CSS
  - JavaScript and Node.js

System Architecture



System Components

- Hardware Platforms
  - Proofpoint VMware ESXI server hypervisor
  - Ubuntu virtual machines
- Software Platforms / Technologies
  - VS Code, PyCharm
  - Python 3
  - Cuckoo
  - TensorFlow + Keras

Risks

- **Feasibility of using Machine Learning to analyze categories of malware**
  - It is not known if it is possible to encompass all different types of malware using a single machine learning algorithm.
  - **Mitigation:** Try to have our feature extraction be as modular as possible so that when it feeds into the machine learning algorithm, the algorithm does not need to worry about different file types.
- **Determining what file characteristics can classify malware**
  - For the ML algorithm to learn, it must be fed many files and analyze the characteristics of those files to learn what is malware. It is difficult to determine what characteristics can be used to detect malware.
  - **Mitigation:** Trial and error research into what kind of characteristics are consistent across different malware files and attempting to detect them with those characteristics before training the algorithm on it.
- **Identifying steganography**
  - The malware our project is concerned with is hidden within various payload files. It is not known if this can be detected without a full detonation.
  - **Mitigation:** Measure entropy to determine encryption. Also check for hidden values in least significant bits of RGB values in pictures.
- **Meeting target performance requirements**
  - For our project to be successful, it must be able to detect and classify malware at a faster rate than sandbox detonation but with just as much accuracy.
  - **Mitigation:** Design our algorithms with the best practices in mind and strive to have high efficiency and accuracy.

Questions?

? ? ? ?

? ?

? ? ?