MICHIGAN STATE
UNIVERSITY

Project Plan
Open Source Intel

The Capstone Experience

Team GM

Ben Buscarino
Will Crecelius
Igli Ndoj
Qiming Ren
Taylor Zachar

Department of Computer Science and Engineering
Michigan State University

Spring 2020

From Students…
…to Professionals

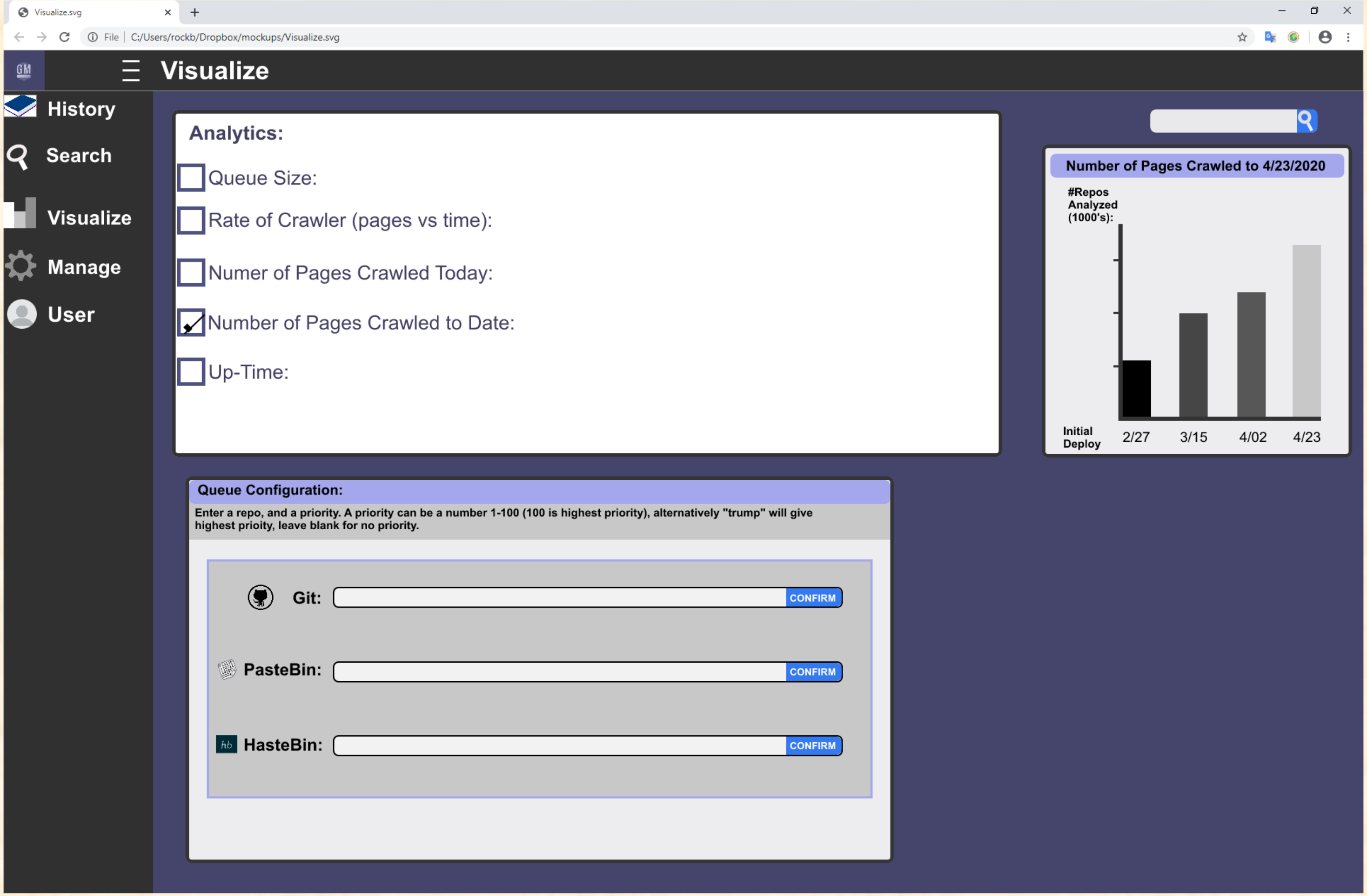
Functional Specifications

- GM is so large that it is at risk of having intellectual property leaked on public repositories.
- Open Source Intel will be able to find any GM owned IP on public repositories and then alert the GM team of findings
- A confidence rating will be given of the assurance that something is found on the site.

Design Specifications

- Scrape various public repositories to identify GM property and security threats such as: usernames, API keys, and code snippets.
- User friendly web frontend to display the data collected by the backend Python scrapers.
- Once the threat has been identified and verified, GM can proceed with legal action through the website.

Screen Mockups:Visualize



Visualize.svg
C:/Users/rockb/Dropbox/mockups/Visualize.svg
GM
History
Search
Visualize
Manage
User
CONFIRM
CONFIRM
CONFIRM
Number of Pages Crawled to 4/23/2020
#Repos Analyzed (1000's):
Initial Deploy
2/27
3/15
4/02
4/23

Visualize

Analytics:

□ Queue Size:

□ Rate of Crawler (pages vs time):

□ Numer of Pages Crawled Today:

☑ Number of Pages Crawled to Date:

□ Up-Time:

Queue Configuration:

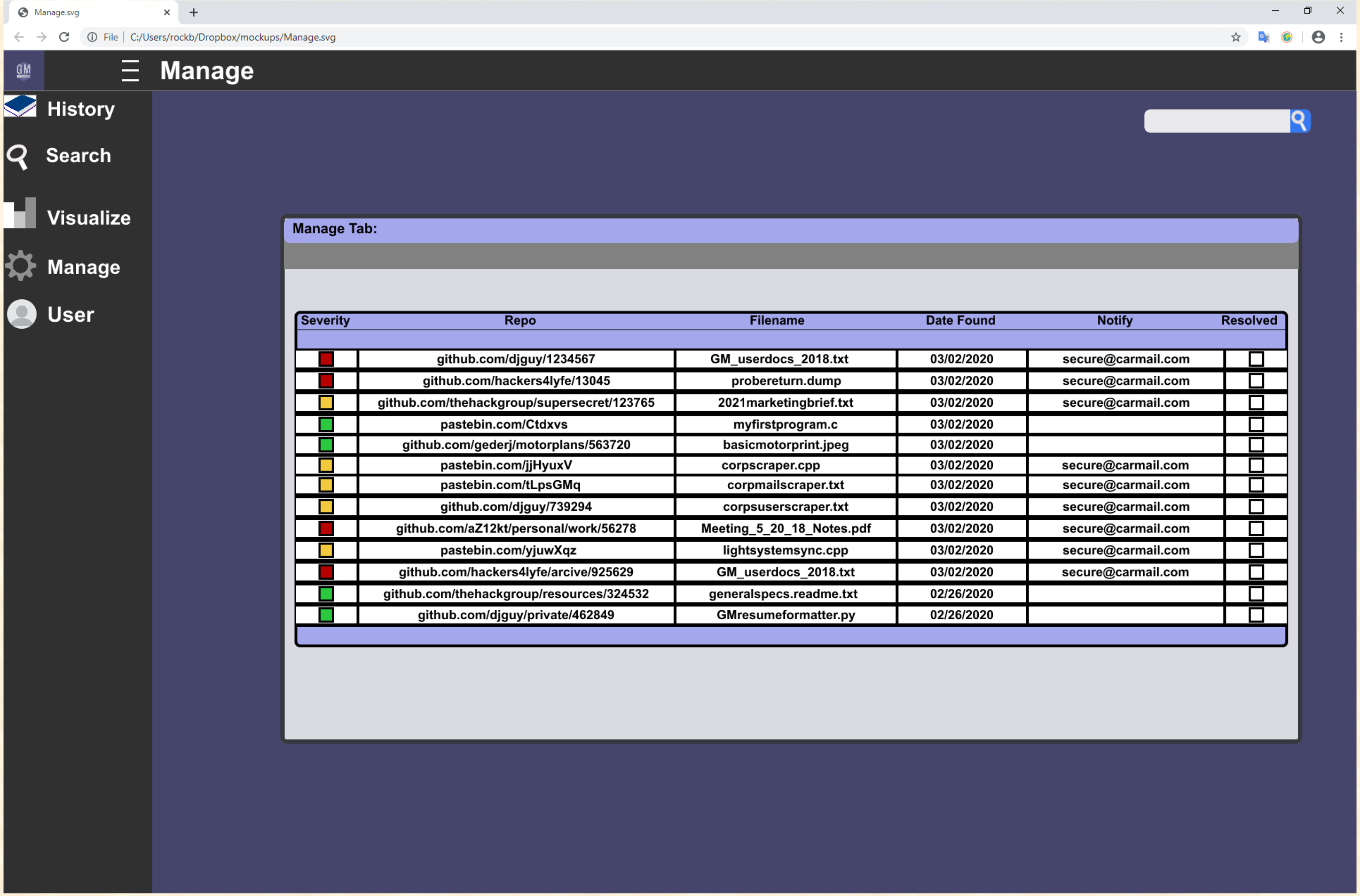
Enter a repo, and a priority. A priority can be a number 1-100 (100 is highest priority), alternatively "trump" will give highest prioity, leave blank for no priority.

Git:

PasteBin:

HasteBin:

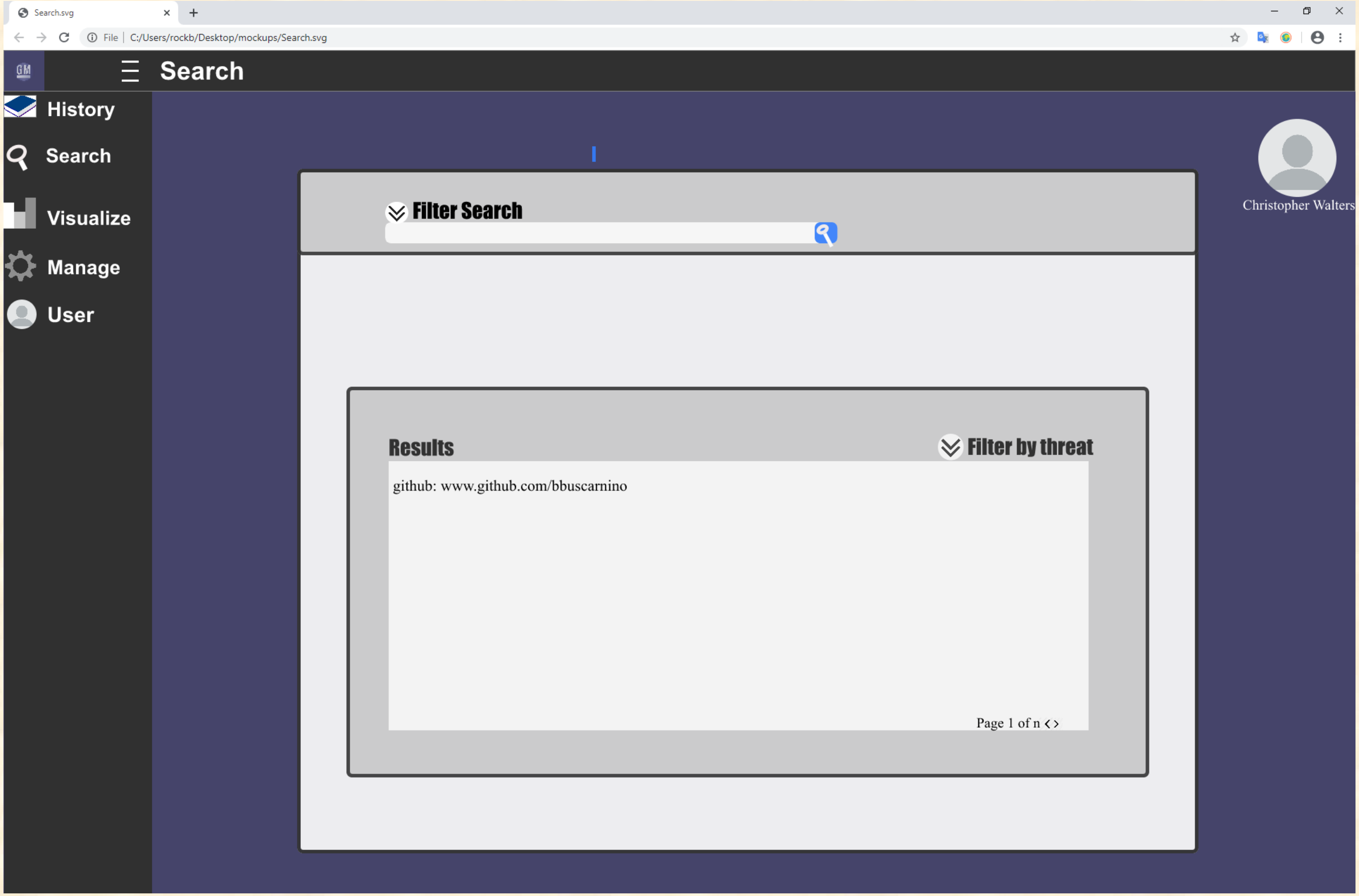
Screen Mockups: Manage



Manage Tab:

Severity	Repo	Filename	Date Found	Notify	Resolved
🟥	github.com/djguy/1234567	GM_userdocs_2018.txt	03/02/2020	secure@carmail.com	□
🟥	github.com/hackers4lyfe/13045	probereturn.dump	03/02/2020	secure@carmail.com	□
🟨	github.com/thehackgroup/supersecret/123765	2021marketingbrief.txt	03/02/2020	secure@carmail.com	□
🟩	pastebin.com/Ctdxvs	myfirstprogram.c	03/02/2020		□
🟩	github.com/gederj/motorplans/563720	basicmotorprint.jpeg	03/02/2020		□
🟨	pastebin.com/jjHyuxV	corpscraper.cpp	03/02/2020	secure@carmail.com	□
🟨	pastebin.com/tLpsGMq	corpmailscraper.txt	03/02/2020	secure@carmail.com	□
🟨	github.com/djguy/739294	corpsuserscraper.txt	03/02/2020	secure@carmail.com	□
🟥	github.com/aZ12kt/personal/work/56278	Meeting_5_20_18_Notes.pdf	03/02/2020	secure@carmail.com	□
🟨	pastebin.com/yjuwXqz	lightsystemsync.cpp	03/02/2020	secure@carmail.com	□
🟥	github.com/hackers4lyfe/arcive/925629	GM_userdocs_2018.txt	03/02/2020	secure@carmail.com	□
🟩	github.com/thehackgroup/resources/324532	generalspecs.readme.txt	02/26/2020		□
🟩	github.com/djguy/private/462849	GMresumeformatter.py	02/26/2020		□

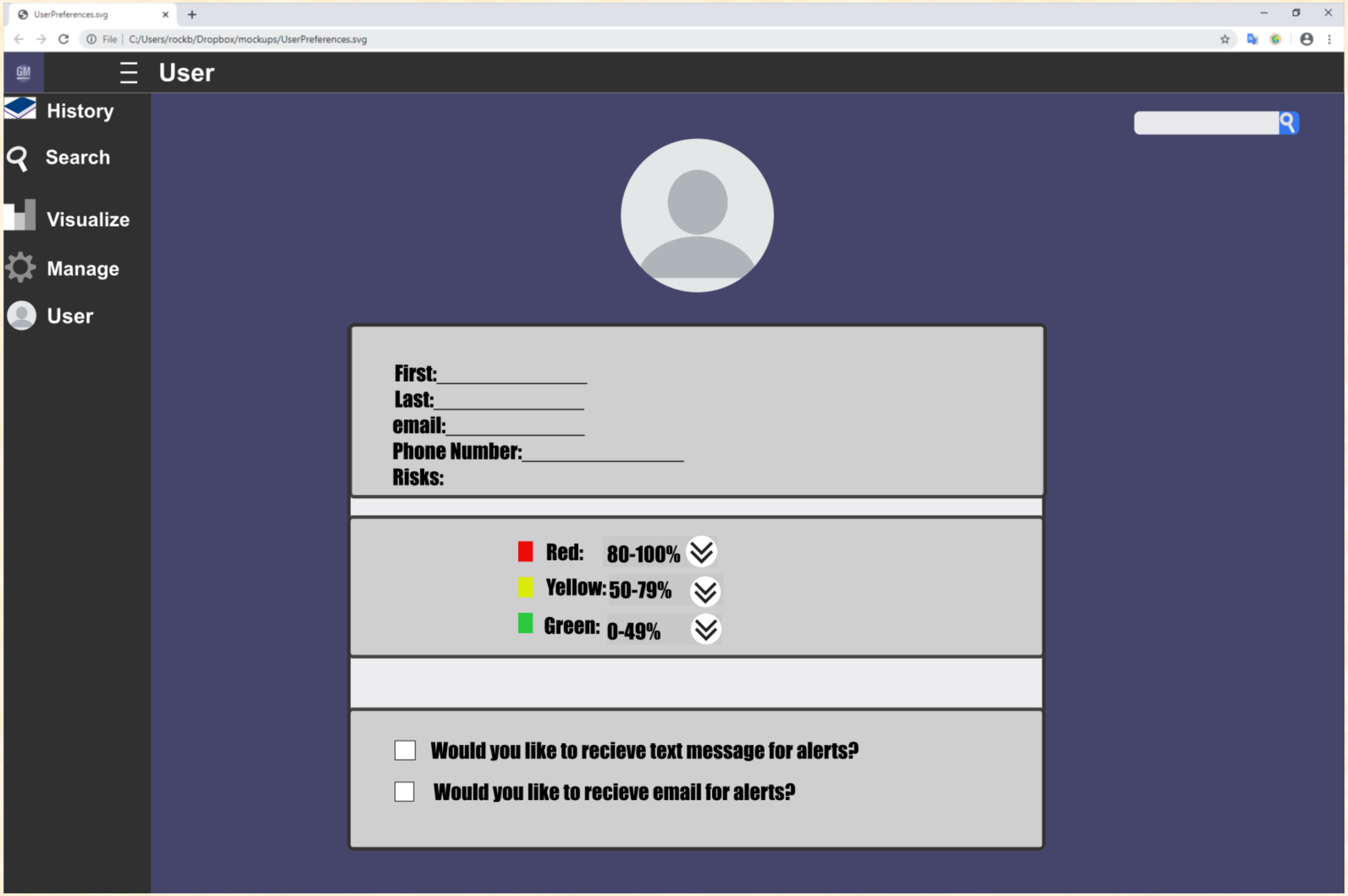
Screen Mockups: Search



github: www.github.com/bbuscarnino

Page 1 of n < >

Screen Mockups: User-Preferences



First:_______________

Last:_______________

email:_______________

Phone Number:________________

Risks:

🟥 Red: 80-100%

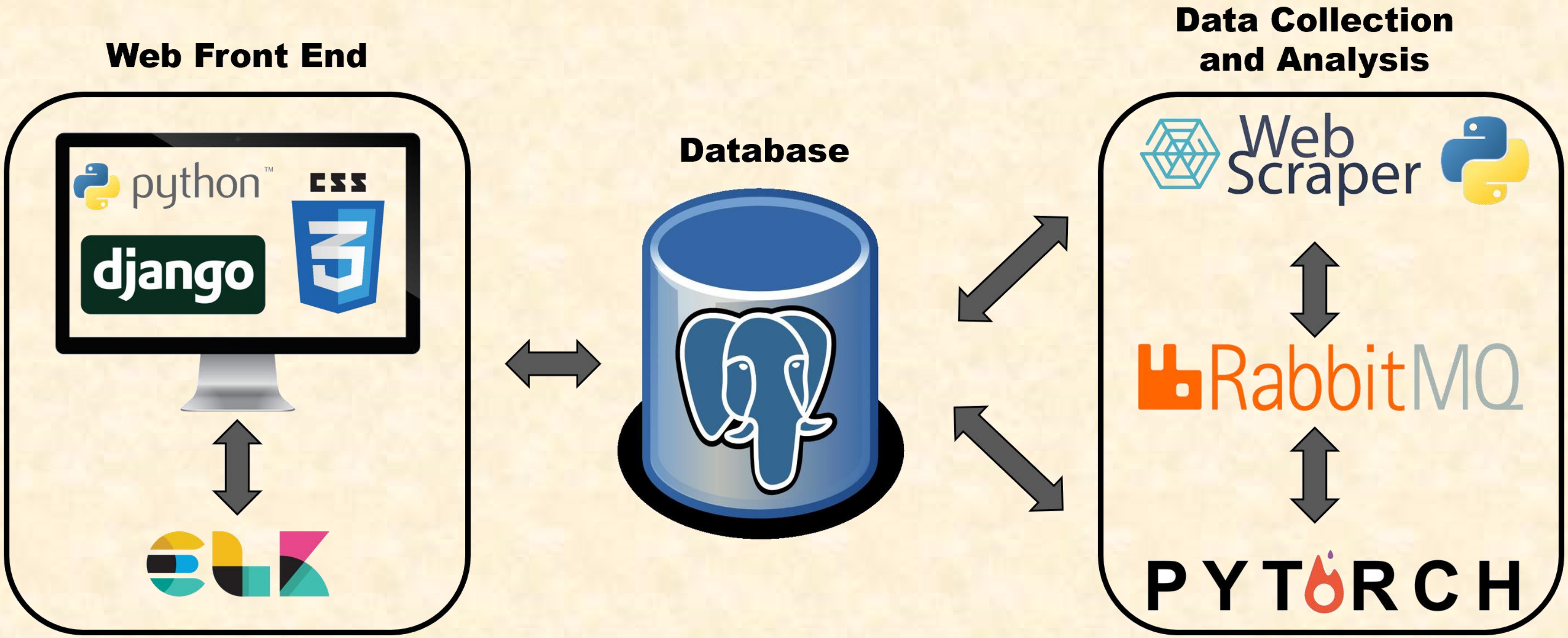
🟨 Yellow: 50-79%

🟩 Green: 0-49%

□ Would you like to recieve text message for alerts?

□ Would you like to recieve email for alerts?

System Architecture



Technical Specifications

- GM employees interact with a Python web application.
- The front end uses Django, a high-level Python web framework, styled with the Semantic UI CSS Framework.
- RabbitMQ serving as a work-queue between a suite of web crawlers and PyTorch machine learning service.
- PostgreSQL used for data warehousing of confidence rating, crawled websites, and details of identified leaks.

System Components

- Hardware Platforms
  - Rack-mounted server running NixOS.
  - MacOS Catalina.
- Software Platforms / Technologies
  - Python, Django, Semantic UI.
  - PostgreSQL and RabbitMQ.
  - Web scraper and ML services using Beautiful Soup, Requests, and PyTorch.
  - Kubernetes and Azure Kubernetes Service.

Risks

- Risk 1: Computing constraints
  - Scanning the entirety of GitHub/GitLab/Bitbucket/Pastebin plus their revision histories.
  - Mitigation: Leveraging bounding and caching much as possible to reduce the amount of computationally expensive work done.

- Risk 2: Identifying problematic content
  - GM is such a large company that they don't know all types of IP the enterprise may have exposed.
  - Mitigation: The crawlers will need to have a large breadth for their search,. Treating findings as metaphorical "breadcrumbs" and showing them to the user in a meaningful way so that the user can finish the investigation is an encouraged solution offered by our client.

- Risk 3: Bringing machine learning to production
  - Nobody on the team has experience with machine learning.
  - Mitigation: Leverage outside resources such as MSU CSE faculty and industry contacts.

Questions?

? ? ? ?

? ?

? ? ?