ANU Tech Launcher
Feasibility of the Web Crawler

I. Collection of images via public APIs

a. Facebook

b. Instagram

c. Twitter

d. Tumblr

e. Pinterest

II. Collection of images via application approval

III. Collection of images via automation

IV. Conclusion

I. Collection of images via public APIs

a. Facebook
Public APIs provided by Facebook allows developers with approved applications to collect images and its data. However, with the recent Cambridge Analytica scandal, we are no longer able to reach outside of "friends of friends", greatly limiting the number of collectable images. The availability of Facebook APIs does not equate to a successful use of collected data since Facebook does not allow developers to use the data for identification of individuals.

b. Instagram
Instagram was purchased by Facebook in 2012 and similar to Facebook's actions, a number of vital APIs were deprecated. For example, due to the deprecation of "following" API, we are no longer able to reach outside of the first degree of connections. This again limits the number of collectable images.

c. Twitter
While Twitter is independent of Facebook and Instagram, the effects of Cambridge Analytica led to a review of developer APIs. Their new privacy policy states that we "cannot associate the Twitter Content with any person or other individual identifiers". Similar to Facebook, while we are able to collect small amounts of images and its metadata, this is not a viable option for an application to identify persons of interest. This requires special requests from the law enforcement and must go through a review process which our team cannot pursue due to time and resource constraints of Tech Launcher.

d. Tumblr
Tumblr features an accessible library of images via its APIs, enabling approved developers to download any number of images from a specified Tumblr page. The use of this API requires an approval of the application from Tumblr. Research from Tumblr's user and developer privacy policy states that the images collected from its APIs must be used without any connection to the user.

e. Pinterest
Pinterest also features an accessible library of images ("boards" and "pins") via its API. However, their developer privacy policy strictly prohibits us from storing the downloaded images.

II. Collection of images via application approval

The approval of our application allows us to access a much bigger library of images from the respective social network sites. However, storing the collected images and identifying users from them is still an obstacle that only the law enforcement can overcome. The approval of the application implies that the application does not violate the privacy policy of the respective social network

sites, but this the nature of our project already contradicts the policies. Furthermore, the application must be completed before the review process. Our team believes that the timeframe of Tech Launcher in the first semester is not realistic to create such an application and get approved.
Our team believes that the development of the web crawler is not feasible within the timeframe of Tech Launcher in the first semester and recommend it to be reconsidered for the second semester of the project. This requires further investigation by Biometix due to the number of legal and privacy policy issues that may arise.

III. Collection of images via automation

a. Selenium and PhatomJS
   Using Selenium and PhantomJS simulates a real web browser to collect profiles and images. This allows us to collect images at a slower pace but evades the detection of using bots. Our team have used this approach on Instagram but was unsuccessful in collecting the images in a reasonable time. This method still cannot be utilized due to the privacy policy concerned mentioned in part I.

b. Beautiful Soup
   Python Beautiful Soup allows us to scrape static websites. This method was inspired from the developer consoles provided by the web browser. The team tried this method on Instagram and learnt that the user profiles were loaded dynamically, defeating the purpose of Beautiful Soup.

c. Third party social network site scraper
   Attempts at utilizing third party social network site scrapers turned out to be unsuccessful due to the deprecation of APIs in Facebook, Instagram and Twitter. After a few attempts of modifying the scrapers, one of our team member's Instagram developer account got permanently banned.

IV. Conclusion

Our team tried several methods on a number of social network sites and were successful in collecting the relevant data. However, it was not possible for us to store and utilize the collected data in any meaningful way due to several privacy and legal concerns. Facebook, Twitter, Instagram, Tumblr and Pinterest were some of the largest social network sites that we investigated and found that while some allow image collection, all of them prohibit developers to store and use the images to identify individuals. The collection and the identification of public users can only be achieved after the social network sites' approval and the team believes that it is unreasonable for a team of students to get the approval. The focus of Tech Launcher lies on the team development of a product and we believe that the web crawler involves an unreasonable amount of legal issues. Our team recommends Biometix to reconsider the web crawler throughout the first semester of Tech Launcher and introduce it in the second semester should the project proceed.