Andrew Williams & Suhail Bhat

Algorithmic Investigation - Pitch Memo

April 4, 2020

For this assignment, your group will identify and submit a brief description of 2-3 algorithmic systems that you are interested in investigating for your final reporting memo.

Your submission will consist of a 500-700 word outline for each system that includes enumerated answers to each of the questions below:

1. What system or set of decisions relies on the algorithm you want to investigate?
2. Is that system public (i.e. part of a government process) or private (i.e. part of a software or other decision-making system that is privately owned)? For the purposes of this project, if a government process relies on a private system (as in the case of COMPAS) we will still consider that public.
3. Who are the stakeholders in this system? This part should identify as specifically as possible the stakeholders who a) own and operate the system b) license and/or utilize the system and c) are targeted/affected by the system.
4. Is the algorithmic system predictive or explanatory?
5. Where can the inputs and outputs of this system be found?

**Algorithm #1 | YouTube**

Another option for our Algorithmic Investigation is to take a look at an algorithm that many people don’t think of as critical or damaging on first thought. YouTube, and Google to an extent, relies on an algorithm to give viewers suggestions on what they should watch on the platform and many content creators will tell you that the algorithm has changed countless times over the last few years. However, the tinkering of YouTube’s formula isn’t victimless. With each change in the algorithm, YouTubers can lose or gain views, subscribers and ad revenue and for those whose main source of income is from their content, this could make or break them.

Many think of YouTubers as being rich people who sit around and play video games all day, but there are many smaller content creators on the platform who live paycheck to paycheck and in the midst of COVID-19, the risk of becoming “demonetized” due to an unknown change in YouTube’s algorithm can be the straw that breaks the camel's back.

YouTube’s algorithm a private, proprietary system owned and operated by Google and it’s parent company Alphabet. YouTube is the stakeholder that owns, operates and utilizes the system while YouTubers and their content are targeted and affected by it. Viewers are also targeted and affected by it in that it changes what content they are suggested and exposed to on the platform. Google bought YouTube in 2006 and has implemented and tested various features that have been hated by users ever since. Notable examples include [forcing Google+ on users](https://www.theverge.com/2019/4/2/18290637/google-plus-shutdown-consumer-personal-account-delete) to turn the platform into a Facebook competitor and the [removal of the dislike button](https://arstechnica.com/gadgets/2019/02/youtube-doesnt-like-dislike-mobs-will-try-to-prevent-dislike-button-abuse/?comments=1). However, none have been despised as much as the reiterative changes Google has implemented on the platform’s algorithm.

YouTube’s algorithm is a predictive system. For each video, [the algorithm](https://www.youtube.com/watch?v=hPxnIix5ExI) takes in data input of a video’s title, thumbnail, descriptions and how other viewers seem to be enjoying it by looking at likes, comments and how long users are watching it. YouTube uses these criteria and weighs how likely a viewer would be to watch the content before suggesting it to them based on their viewing history.

It seems straightforward, but this system hurts smaller content creators and favors businesses and the rich, popular and often problematic YouTubers with millions of subscribers. In order for the algorithm to promote a video to users, the video needs to have high engagement but to get the engagement it needs promotion. This is no problem for larger channels since they are guaranteed engagement through their millions of subscribers while smaller channels are not. An algorithm is also used to determine whether or not a video is suitable for ads to be shown and this is how content creators make their money. How the algorithm weights each factor is a heavily guarded secret and even subtle changes to the weighting can be the beginning of the end of a YouTube channel.

Issues with YouTube’s algorithm don’t end with the demonetization of content creators. Guillaume Chaslot, a developer who worked on YouTube’s algorithm, [revealed in 2019](https://thenextweb.com/google/2019/06/14/youtube-recommendations-toxic-algorithm-google-ai/) that the algorithm intentionally recommends divisive and sensational content like fake news and conspiracy theory videos to drive up views. It favors content that rides close to being prohibited and a policy violation.

On the surface, the YouTube algorithm seems like something that really only damages companies and popular YouTubers pulling in millions of dollars, but small content creators, who are ordinary everyday people, and viewers are the ones truly being hurt by it.

**Algorithm #2 | DNA Testing**

DNA profiling has been an evolving topic in law enforcement and healthcare since the 1980s. It works on the fundamental scientific principle that each DNA sequence is unique and can be used as an identification tool in law enforcement and for health risk evaluation in medical sciences. In the United States, the Federal Bureau of Investigations maintains a national DNA database called CODIS. But, citizens can get a DNA test from a private lab such as 23andme and AncestryDNA. No doubt, DNA testing was a game-changer, but it has come up with its own problems.

In the beginning, the DNA database was expected to be a collection of the names of sex offenders, but overtime nearly everyone who goes through the system, including people who get arrested but not convicted, ended up in the database. DNA material of family members have resemblance to each other so if the law enforcement officials can’t find an exact match in their database they go after the one that falls closest to the DNA sample. This results in wrongful arrests of innocent family members.

A test sample (a saliva swab or a blood sample) goes for testing in a lab. Once the genome (structure of the DNA) is fetched, it is pushed into a system that matches it with other records in the system. Based on the match, a diagnosis is made. Of the top, the structure of DNA is complex that begs the question how a predictive model compares the two. What does it look for when comparing the results? The concept of operationalization becomes critical here. If the system is looking for, say, 1,000 points for comparison, how will the match percentage change if it looks for 2,000.

No two DNA’s match 100 percent. It is always a percentage of the genome that matches. In some cases, even if the genome matches 10-20 percent, it might be considered a match. There are two issues here. First, what percentage of matches is a confirmed match. For law enforcement agencies, it doesn’t matter. They can bring in people with 5 percent match to their samples for questioning or may arrest people as a suspect. Secondly, the genome from the new sample gets matched with a database and there is a chance that the system can make a false match. It becomes imperative to understand what roadblocks prevent the system from showing a false positive.

Some DNA tests claim they can tell the race of a person based on their tests which is absurd. They create a false link between biology and ethnicity. A collection of thousands and thousands of tests create enough data points to design a predictive system that can identify similarities and differences between DNA structures. It is plausible that calculated guesses can be made about a person’s risk of genetic diseases, but can it tell us anything about a race. That is what genome testing labs claim. But the question is what is the level of accuracy? How much can I trust it? The algorithm might get it right a few times among several predictions but that level of probability does not instill confidence to say that it can predict the race of a person with reasonable accuracy.

These two scenarios are a way to analyze the DNA testing systems and their ethical and functional challenges.

The other challenge, arguably the most talked about in the media, is the privacy concerns around DNA testing. How the data is stored and is it shared with other companies? What will happen if there is a breach? What is the potential cost of that? Can it be manipulated by bad actors by designing target-specific biological chemicals that can kill a person while making it look like a death by a natural cause?

We are in the middle of a data boom in healthcare where doctors, scientists, healthcare companies, insurance firms and big tech are looking to tap this sector. More than ever, it becomes important to analyze and understand DNA testing as it is touted as having the potential to save people from excruciating diseases by early diagnosis.