# ATMS 517: Data Science for Geosciences
# Week 1 Assignment
# By Omkar Mehta (omehta2)

**1)(75 pts) Identify a reported violation of data ethics in recent years (not required to have occurred in your field of study), and answer the following:**

a. Provide a link to report on the violation (from a reputable (media) source) - no Buzzfeed or Reddit, please!

b. In 1-2 paragraphs, briefly summarize the data ethics violation, as well as your musings on whether or not you think it was a true violation

c. *Come to next week's synchronous session prepared to give an overview of the data ethics violation you identified to your peers in small groups!*

d. *The link for your identified data science ethics violation may be shared with your peers in a repository on our Moodle site!*

**Solution**:

A Beauty contest was judged by AI and the robots didn't like dark skin

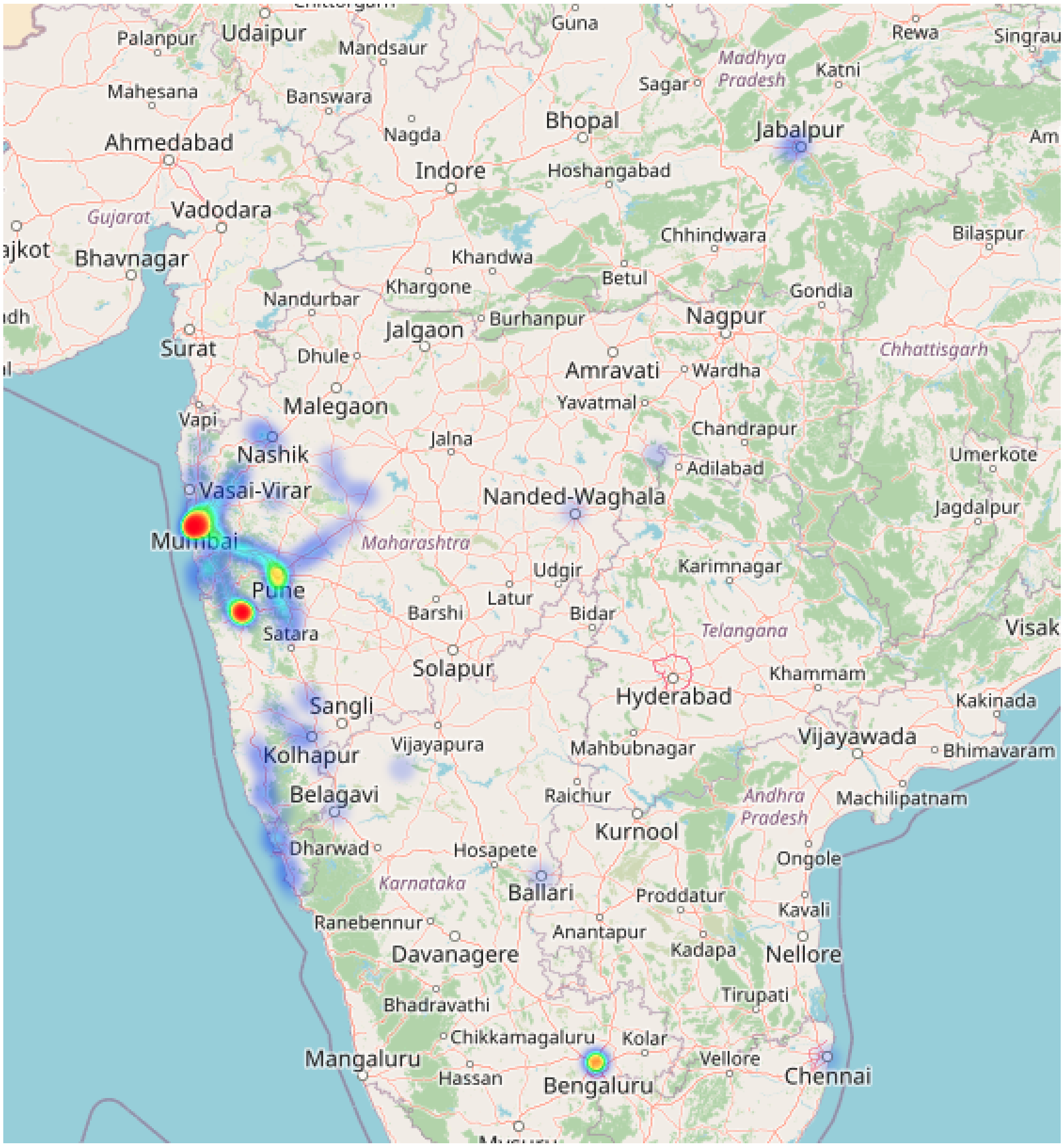
The article posted by the Guardian has highlighted a serious data ethics violation. As taught in the asynchronous lectures, the data ethics violation can occur in two ways: for the data itself and for the science done with the data. The article is more of a second type of violation, in which there was inadvertent discrimination against people of color. The algorithm that was trained by Beauty.AI's Youth Laboratory's Deep Learning team and supported by Microsoft did not include enough minorities. While it was supposed that the first international beauty contest to be judged by machines should have identified the most attractive people, based on facial symmetry, wrinkles, and other features, it did not like people with darker skin. Out of 44 winners, very few were Asian, most were white, and only one had dark skin.

Even though we know that Deep Learning networks are black-box models, we can infer that whatever the input data is, the algorithm inherently captures the input data's features. Here, the input data had images of people of less diverse color. People with darker skin were very few. Hence, the algorithms perpetuated biases, which resulted in biased results. While this was a total case of inadvertent discrimination, the humans behind the algorithm didn't make proper assumptions. If the data processing and analysis were done in the right manner, they would have understood that the data is bad. This could have averted bad data-driven biases. This could be easily avoided by improving the input data or implementing filters to make sure that people of different races are represented with equal treatment.

**2) (25 pts) Describe 4 specific ways in which data was collected on you, your activities, and/or your housing today, in two paragraphs or less. I encourage you to think critically about non-obvious ways data on you is mined!**

**Solution:**

The first recollection and the scary nature of how big companies like Google, Facebook, etc collect data came to me when I was looking at my Google location data in 2018. Even though I used to keep my location on and I had specifically not permitted any company, even Google, to track it, I found out that my entire history is there. So, I fired up my Jupyter notebook and coded to create a heatmap of my location data.



I was shocked by how Google knows every place that I had visited, ever since I got my first smartphone. I made a big jump on the inference that Google must be collecting metadata, too. Like, what I did, whom I made contact with, etc. I am still fascinated by how these companies have made tremendous progress in NLP, CV and ML fields. I have Google Mini, subscriptions to Netflix, HBO Max, Amazon Prime, Disney Plus, YouTube music and Hulu. My life has also been simpler when these products are in my life. Like, these apps collect information on what I say, type, search, browse and buy. I get movie recommendations to my own liking from Netflix, HBO Max, Amazon Prime video, Disney Plus and Hulu, because these companies have already created a model of me. This model of me might even know better than me about my likings.

My early morning and night music tracks are perfect — thanks to YouTube music's recommendation engine. It even plays music in my mother tongue. It has diverse tastes. Google Mini does everything from scheduling my tasks on the calendar, reminding me about assignment deadlines, and buying groceries. I even know that Google Mini listens to every conversation because as soon as I browse, I get advertisements on Amazon's side page for what I might buy, based on the conversation Google Mini had heard. It is scary and useful, at the same time. This could be used for bad use, and the company must be responsible for data ethics violations. I also use Apple products like Macbook Air, Ipad Mini, iWatch, and iPhone. Somehow, I feel safe with Apple, even though it also tracks and collects data on me like how the iWatch recommends me to stand up after long hours of idle sitting (thank you for taking care of me).

So, in summary to how data is collected on me, my activities, and housing, it's through location (geographic data), conversation (natural language data), buying habits (transactional data), and smartwatch (time series data).

**3) (No points, but crucial for success in this class). Decide where you want to run Python, set it up, and successfully run the test Jupyter notebook (you can find it in Week 1 Raw Lessons as Example Jupyter Notebook).**

**Solution:**

This is actually a harder question for me because I have not installed Anaconda on my local laptop, even though I have used Anaconda, three years ago. I use Google Colab extensively for GPU. But, ever since I have understood the power of virtualization, I have moved to VS Code's Remote Access facility with Docker or Google Cloud Services. I will be using VSCode with Google Cloud Services, where I will use CPU Compute Resources — 16 GB RAM, 1 TB Storage.