*What are some of the validity issues that are documented in the Hecht reading? What are possible ways something similar could arise in data either you collect, or data that is given to you?*

The research performed by Hecht focused on two study methods and areas. First, the authors sought to better understand the prevalence with which location information was readily populated by Twitter users in their profiles. Through their research methodology, the authors disproved, at the time, the common belief that this information was both regularly provided by users and contained accurate and precise geographical information. Secondly, the authors also hoped to better understand whether users unknowingly but implicitly provided geographical identifiers through their tweeting behavior on the platform. As part of this second question within the study, the authors were able to train a classifier through supervised learning that could reliably “guess” a user’s geographical location at the country and state level. As noted throughout the article, the author’s raised clear concerns around the validity of the location information provided on the Twitter platform, and this was therefore an important article that raised the important reminder to carefully vet data and ensure that its reliability is in line with expectations.

I agreed with the concerns raised by the authors around both validity and privacy when attempting to leverage this type of data in any reporting. However, I believe one “missing” piece that should be considered alongside this article is that this was published in 2011, when Twitter was still relatively new. Now, in 2022, and with users having a greater understanding of data tracking and how to enhance privacy/security, there may be a different prevalence in terms of what is provided within the location field. Based on my own experience with Twitter, I would suspect a large amount of “joke” locations to be used, perhaps even more so than what the authors were already flagging in 2011. Overall, though, the authors made clear the validity issues associated with the readily available location field and the constant need to review and clean data as much as possible before any analytic work does begin. However, the authors also made an important note in the second study about user privacy, which tied in well with the OkCupid article that was discussed last week. Even if a user’s location information is not being readily made available on the platform, this information can simply be derived through pulling tweets through the API and running a basic supervised machine learning algorithm to predict the location. This article, therefore, again puts privacy on the forefront of issues to consider when running social media analytics and is a helpful reminder that even if data can be pulled or predicted, researchers must minimize harm as much as possible and gather informed consent to ensure appropriate ethical practices are in place.

The lessons learned from this article have application today in the data that I collect in my own day to day work. I work in the healthcare space, and my company generates reporting based on medical, pharmacy, and dental claims that we receive in addition to healthcare eligibility data, as well. As part of receiving this data, there is the ability for patients to opt out entirely from reporting or omit certain demographic information, address and geography included as part of this. However, in medical claims, for example, we’d have insight into the medical providers that the patient is seeing and where their office is located, so we similarly could implicitly determine a patient’s general location through this information. With these types of problems and privacy concerns prevalent across all types of data analytic work, it is important to have very clear and comprehensive business rules in place to ensure that the data is being handled appropriately and that organizations are handling data with a high level of ethical responsibility.