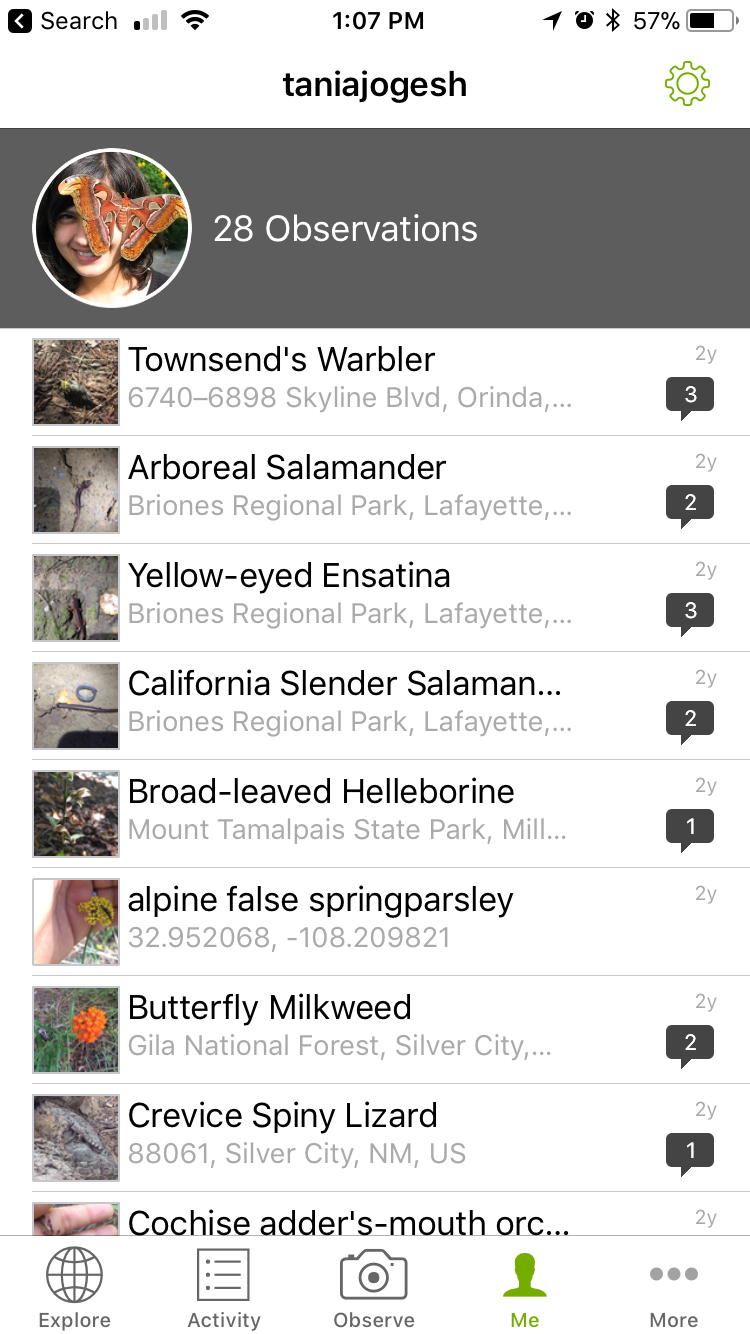
CAPSTONE 2

1. **What is the problem you want to solve?**

Develop a classifier that predicts if an identification posted via the Citizen Science Biodiversity app iNaturalist (https://www.inaturalist.org/) is likely to be correct. iNaturalist is a mobile application that allows users to upload photos and information on natural history observations (plant and animal sightings). The online community can subsequently verify observations uploaded by users. However, there are a lot of observations and not all plants/animals/fungi are easy to identify. Instead of relying solely on the community to verify an ID, my classifier will help iNaturalist decide which observations are more likely to be correct or misidentified.

1. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

My client is iNaturalist and they can use this classifier to determine which observations need additional expert identification and can tag those that are likely to be correct.

1. **What data are you going to use for this? How will you acquire this data?**

iNaturalist’s user data and observations are freely available to the public and can be queried and obtained via an API. A subset of these data include whether or not an observation was accurately classified by experts in the community so the data are already labeled for classification.

1. **In brief, outline your approach to solving this problem (knowing that this might change later).**

The iNaturalist dataset is really large so the first goal would be to acquire and clean the data with Spark. I will then select appropriate features and iteratively test and optimize multiple classifiers.

1. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

The deliverables will include the code, a report and a slide deck.