## Proposal: Pokemon

Team members: Ally Qi, Garet Douglass, Kyle Kinney

Team Github repository: <https://github.com/kkyle16/ETLProject.git>

* Where is the data coming from? (at least two sources)
  1. <https://www.kaggle.com/csobral/pokemon-gen-vii-pokedex?select=pokedex.json>
  2. <https://www.kaggle.com/notgibs/smogon-6v6-pokemon-tiers>
* Where is the data going to? (postgres, mongo, etc; some type of database not a flat file like a CSV.)
  1. The data will go into a Mongo database.
* What will be the structure of the data in the final database? What tables/columns/types/etc.
  1. The final data structure will be a Mongo database collection.
     + One collection will include the full dataset with Pokedex number serving as our index.
     + Another collection will group the data by Pokemon type and provide average statistics and tier counts. We will filter out unevolved and Uber tier (banned) Pokemon
     + The last collection will group the data by tier and include what proportion of the tier is made up of each generation.
* What steps are required to take the data from its current form into its final form and *who is responsible for each step?*
  1. Extraction (Ally): downloading the json and csv files into a Python Panda data frame
     + Extracting all variables from the 1st dataset (json);
     + Extracting only the Pokemon name (Name), Legendary, Generation, Mega, and Smogon 6v6 Tier (Tier) variables from the 2nd dataset (csv);
     + Will upload both datasets to github;
  2. Transformation (Kyle):
     + Data cleaning: cleaning the names of the Pokémon in both datasets and then merge them together in Python;
     + Data merging: the two datasets will be merged by name;
     + Summary statistics for variables
     + Removing uninterested variables
       - We will also add a column that combines Type1 and Type2
     + Preparing for data export to the Mongo database
  3. Loading (Garet):
     + Using the pyMongo to insert all entries to one single Mongo dataset