# Data Cleaning

## CHAMP

December 28, 2018

### Data Cleaning

Three programming approaches were used to clean the CHAMP data: (i) direct reclassification; (ii) replacement with high frequency names; and (iii) text string extraction. For the direct reclassification, the title of speakers was removed from the speaker. For example, the name Obama, Barack President or Obama, Barack Senator, was recoded as Obama, Barack. Additionally, misspelled names were reclassified. For example, George Stephanopoullos was recoded as George Stephanopoulos. The high frequency name replacement took speakers with missing first names and replaced them with the highest frequency person with a shared last name who also has a first name. The frequency checks were done with each show. Therefore, if a show is on MSNBC frequent correspondents from MSNBC were checked before frequent correspondents from Fox News. Last, if a speaker name is missing the text string is checked for names in order of overall frequency. Once a name is identified in the text string, the program moves to the next missing speaker entry in a file. The following programs were used for cleaning:

- Reformat_CSV_Names;
- Name_Classifier_1.R;
- Name_Classifier_2.R;
- Name_Classifier_3.R;
- Name_Classifier_4.R;
- Name_Classifier_5.R;
- Name_Classifier_6.R;
- replace_names_with_high_freq.R;
- Clean_Names_High_Frequency.R and
- Extract_Names_From_Text_String_v2.R;

### Reformat_CSV_Names

The first program takes all the csv output files and then names the show by network, program, and year. That data is extracted from the first line of the parsed data and run through a classifier. For example, some shows like "NBC (" as their network. The network classifier function recodes the network as NBC. This process ensures consistency across network programs per year.

### Name Classification

The Name Classifier programs are direct reclassification of the name data. There are six programs because it took a long time to find all variations of misspellings. This was done my exporting all names and sorting them and then Caleb and I worked on finding the misspelled names. These classification were recoded into functions and then looped through all the formatted csv files.

### High Frequency Replacement

This program takes the frequency of speakers by network and then outputs the data as a list by network. It also removes titles from the speaker names, such as MAYOR, PRESIDENT, OFFICER, SPEAKER.

### High Frequency Replacement

This program loops through the list of frequency by show and then replaces people with missing first names with the first names of speakers that have the highest frequency same last name. This program also removes the titles of speakers in the data. Last, it also does some direct reclassification of data that is wrong.

### Extract Names from Text String

This program scans the first 30 characters of the text string to see if the speaker name is in the text string for missing speakers. The program searches by frequency and stops once a name is identified. For example, the most common speaker in the data is Wolf Blitzer. If the name Wolf Blitzer appears in the text string, and the speaker is missing, then speaker is coded as Blitzer, Wolf and additional names are not checked. This can introduce some error into the program. For example, the name Nicole Richie is a speaker in the data, but is less common than Nicole Rich. The program searches by frequency and stops when Nicole Rich is identified and autoreplaces the missing name with the incorrect speaker.

### NBC Cleaning

I noticed that all speakers from NBC seem to be corrupted. I set all the speaker data from NBC to missing then extracts the names from the text string. I am still working on this code.