





# Industry Classification Using Public Data

Apoorva Dhanala

Supervised by Javier Miranda, Anne Russell, and  
Amichai Joshua Goldsman

Economy-Wide Statistics Division

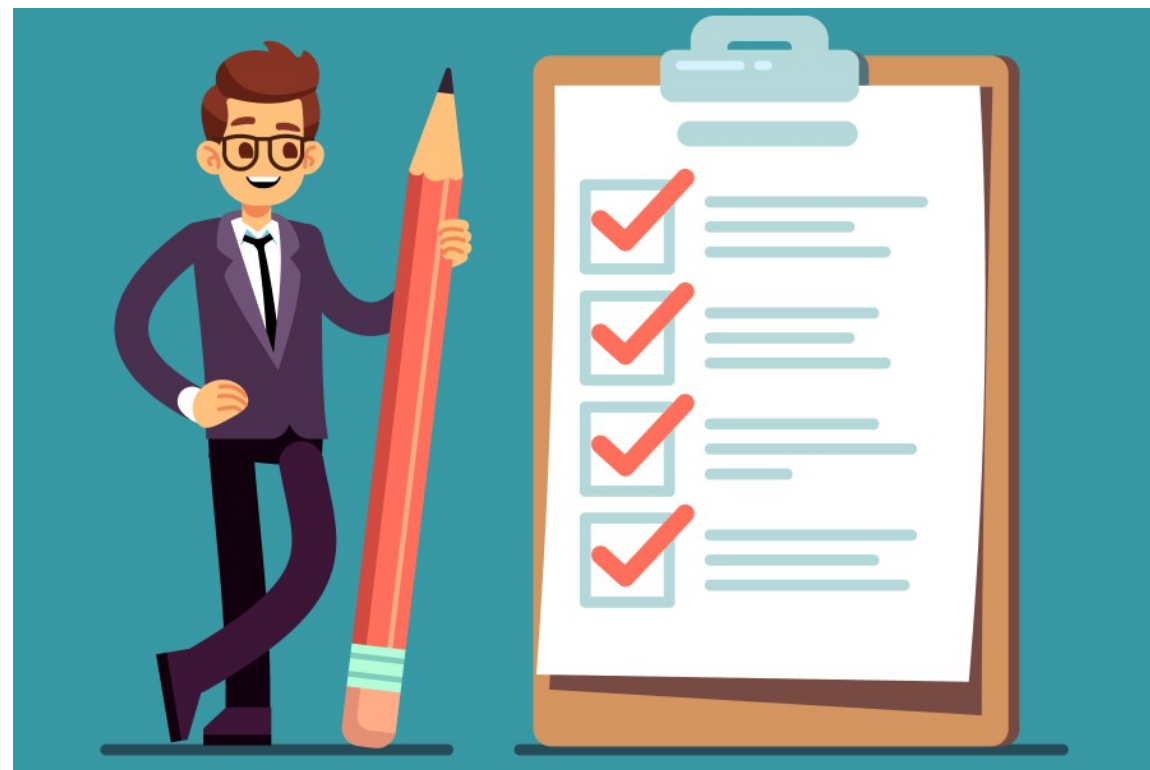
CIVIC   
DIGITAL   
FELLOWSHIP   


# PURPOSE & GOALS

- This project is a continuation and extension of a previous Civic Digital Fellowship project focused on using a Machine Learning pipeline to generate 6-digit NAICS codes using public data
  - NAICS Code Example:
    - 4853: Taxi & Limousine Service, 485320: Limousine Service
  - Using public data and the Google API to classify and improve the quality and accuracy of statistics produced and shown by the Economic census
  - Needed because about 20% of the NAICS code data sourced from administration data is incorrect at the six digit level

# BENEFITS

- Improving the quality and accuracy of statistics produced
  - Ex:
    - Employment rate/numbers
    - Revenue information
- Reducing amount of incorrect classification
  - Reduce the number of incorrectly sent forms mailed to businesses
  - Minimize write-in responses while improving businesses' response rate overall



# CHALLENGES

- The Google API doesn't allow for search customization for specific features, such as:
  - Distinguish between businesses with and without employees
  - Limits data available to developers such as a business's original industry classification
- The Google API is a costly and relatively inefficient tool for data collection



# METHODS & SOLUTIONS

- Creating algorithms and approaches that will allow us to focus on targeting businesses we care about
- Sorting through Economic Census data in order to:
  - Determine whether business density is a better metric than population density to inform search and collection criteria
  - Create an algorithm focused on identifying employer business density by: county, census tract, and NAICS code
  - Understand gaps in industry coverage by looking at county and Census tract



# FUTURE PLANS

- Modify Python script conducting classification to pass data through the pipeline as a function of business density rather than population density
- Optimize the set of keywords that we use in our grid search for data collection
- Test the new scripts and expand the Google API collection
- Update the Machine Learning models using the new data collected from the scripts