# Forecasting the 2024 U.S. Presidential Election*

## Using Aggregated Polling Data to Predict the Outcome of the 2024 Presidential Race

Andy Jiang Onon Burentuvshin Rui Hu

November 4, 2024

The 2024 U.S. presidential election is shaping up to be a highly swing race, with polling data playing a crucial role in gauging public sentiment and forecasting the potential outcome. This paper aims to build a predictive model using a poll-of-polls approach, aggregating data from multiple pollsters to forecast the winner of the upcoming election. The analysis incorporates polling data from various sources, filtered to include only high-quality pollsters with a numeric grade above 3, to ensure the reliability of the predictions. Additionally, a detailed examination of one selected pollster's methodology is conducted, highlighting strengths, weaknesses, and biases in the survey process. The paper concludes with an idealized polling methodology, designed within a $100K budget, to enhance the accuracy of election forecasts. Results are presented in the context of understanding polling trends, potential biases, and the evolving landscape of voter preferences as the election approaches.

## 1 Introduction

The 2024 U.S. presidential election is shaping up to be one of the most closely watched and potentially pivotal elections in recent history. As public opinion shifts and political dynamics evolve, polling data has emerged as a critical tool for forecasting the outcome and understanding voter preferences. Polls provide a snapshot of where candidates stand at various points in time, reflecting the impact of campaigns, debates, and major events. However, the reliability and accuracy of individual polls can vary significantly based on methodology, sample size, and other factors, making it necessary to aggregate multiple sources of data for a more robust prediction.

*Code and data are available at: [https://github.com/AndyYanxunJiang/president-polls).

This paper aims to forecast the outcome of the 2024 U.S. presidential election by leveraging a "poll-of-polls" approach, which aggregates data from various pollsters to create a more comprehensive picture of the race. By focusing on high-quality polls and conducting a detailed analysis of the polling trends for the two main candidates, Donald Trump and Kamala Harris, we seek to uncover patterns in voter support and potential factors influencing shifts in public opinion. Additionally, the paper examines the methodologies of different pollsters to assess their strengths and weaknesses, which can help identify possible biases in the aggregated data.

The analysis will further explore the geographic distribution of support across key swing states, providing insights into the electoral landscape. Finally, the paper presents an idealized polling methodology with a $100,000 budget, designed to improve the accuracy and representativeness of election forecasts. By integrating multiple data sources and critically evaluating polling practices, this study aims to contribute to the ongoing discussion on the role of polling in democratic processes and election forecasting.

# 2 Data

The data used in this paper was obtained from the FiveThirtyEight website (Best, Bycoffe, et al. 2024), which aggregates polling data and other relevant indicators to track the U.S. election landscape. We accessed this data using custom web scraping tools and APIs provided by FiveThirtyEight. The dataset includes information on pollster, state, start and end dates of polls, sample size, population type (e.g., registered voters or likely voters), and party affiliation.

## 2.1 Overview

Our analysis was conducted using R (R Core Team 2022) along with several supporting packages. We used the **tidyverse** for data manipulation (Wickham et al. 2019), **readr** for data import (Wickham, Hester, and Bryan 2023), and **here** for file management (Müller 2020).

Dynamic reporting was handled with **knitr** (Xie 2024), while **kableExtra** allowed for enhanced table formatting (Zhu 2024). We used **scales** to adjust visualization scales (Wickham, Pedersen, and Seidel 2023), and **arrow** for efficient data processing (Richardson et al. 2024).

For Bayesian modeling and visualization, we used **bayesplot** (Gabry and Mahr 2024) and **rstanarm** (Goodrich et al. 2024).