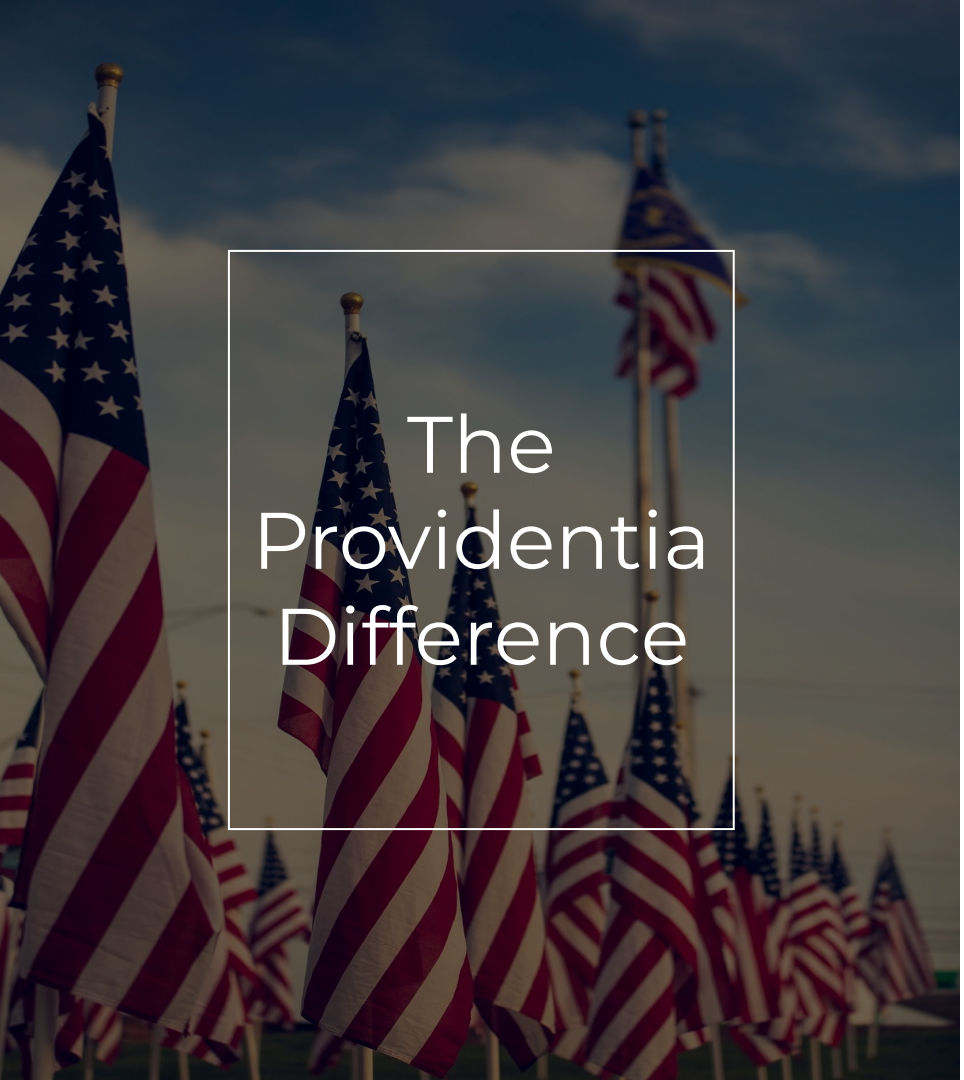

PROVIDENTIA
A new approach to
election forecasting

Providentia enables campaign strategists and political analysts to uncover new insights by leveraging machine learning to forecast elections

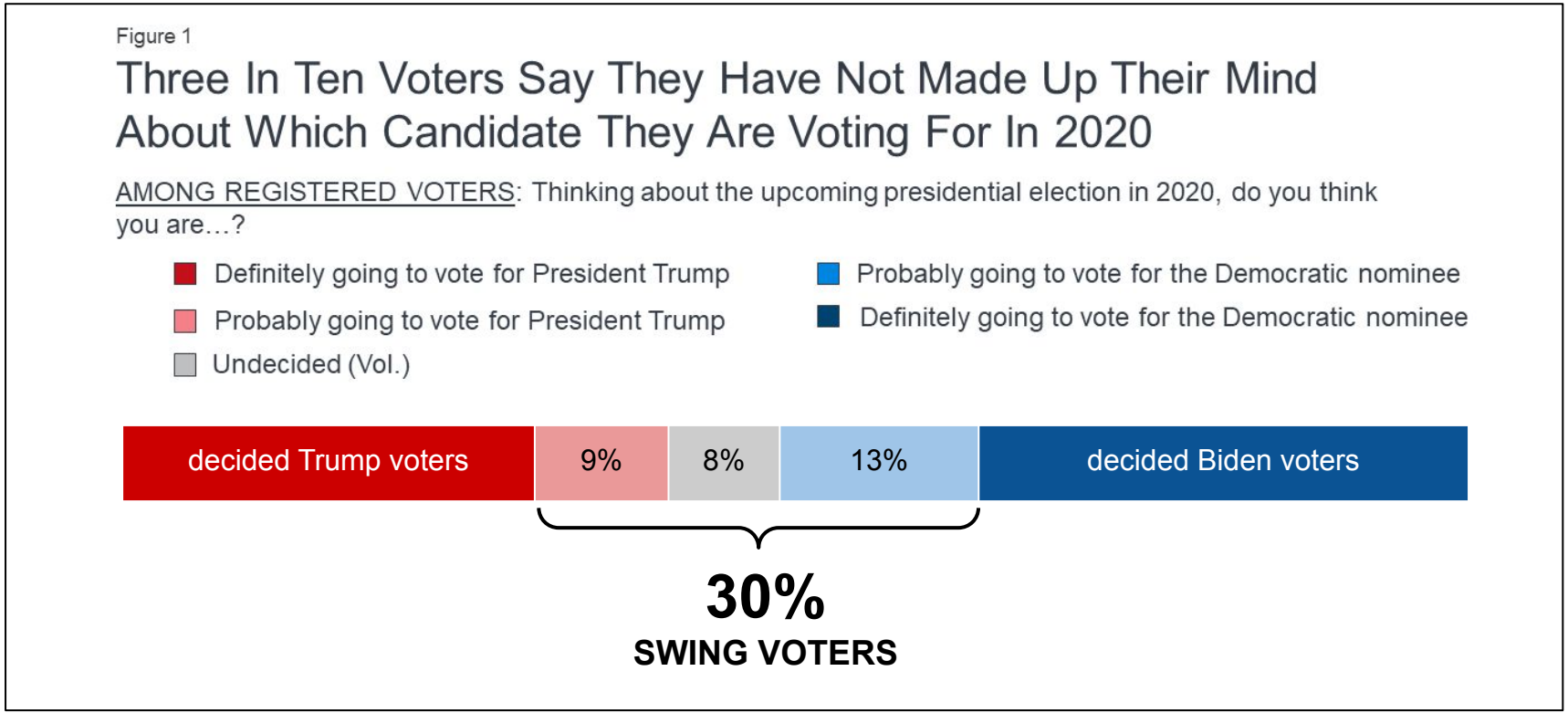


The Providentia Difference

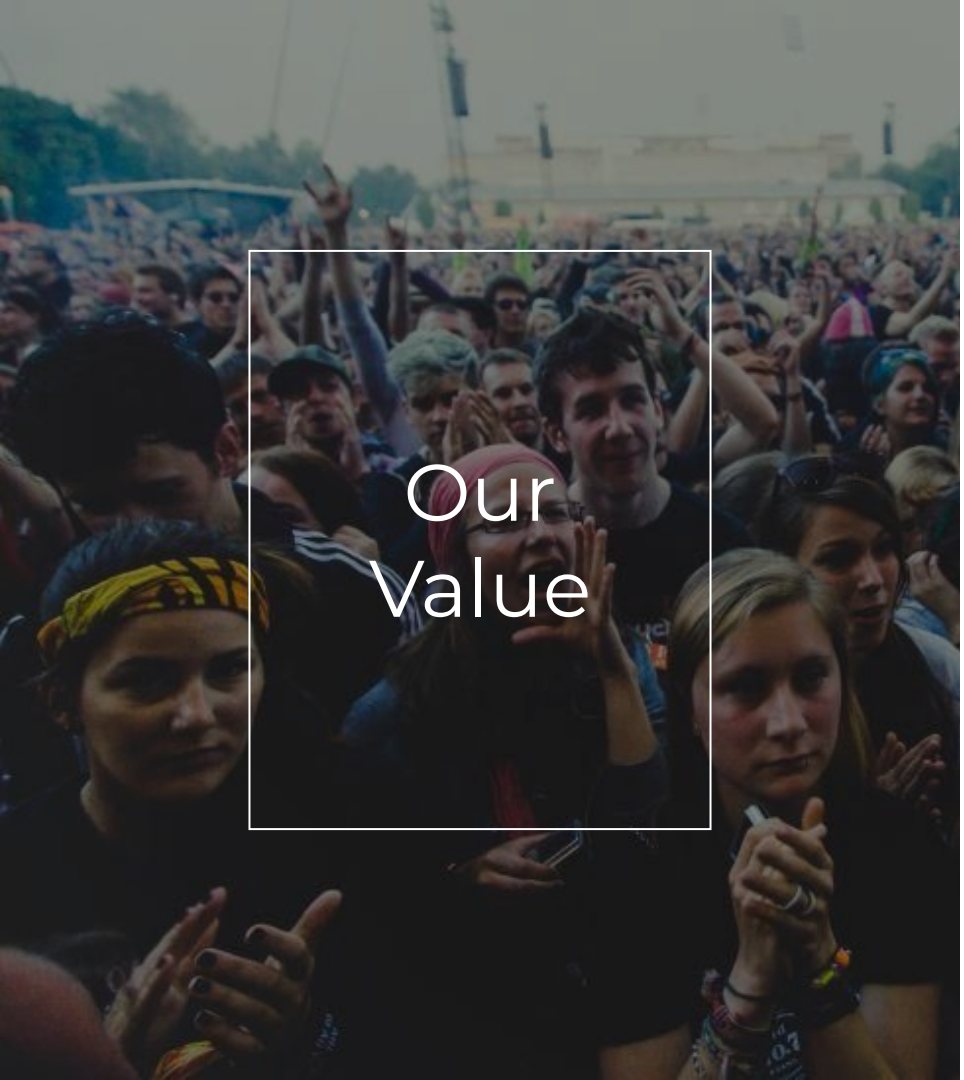
Providentia offers a non-polls based approach to election forecast that is cost-effective and generates insights with quicker turnaround time

Most election forecasts uses polls

Yet, there remains a large number of undecided voters who can swing the election



Source: KFF - Data Note: A Look At Swing voters Leading Up To The 2020 Election



Particular focus on swing states dynamics

We believe undecided voters vote by issues, not along ideological lines

Actionable insights for devising campaign plans

User friendly visualizations to show how and where to focus your resources to swing undecided voters

Overview of Modeling Approach

Datasets

1. Census Bureau demographic and economic data:

- Age groups by county
- Education levels
- Population by ethnicity
- Unemployment % by county

2. Past election results from 2000 to 2016

Undecided Voter Signals

Factors we considered:

- Twitter, headline news, and google trends sentiment for each candidate
- Voter turnout impacts
- Trending issues in each state
  - Generated from election's enter and exit polls (Healthcare, Economy, Immigration, Climate)

Random Forest Classifier & Linear Regression Model

Modeling approach:

1. Predict winner by county
2. Aggregate county results to infer winner by state
3. Aggregate electoral votes to infer winning candidate

Sentiment Analysis - RNTN & Capsule Neural Network

Modeling approach:

1. Sentiment Score on key topics and candidates
2. Mix popularity index in different parts of the country

Challenges:

1. Overfit due to small dataset
2. Imbalanced dataset
3. Cleaning inconsistent tweets

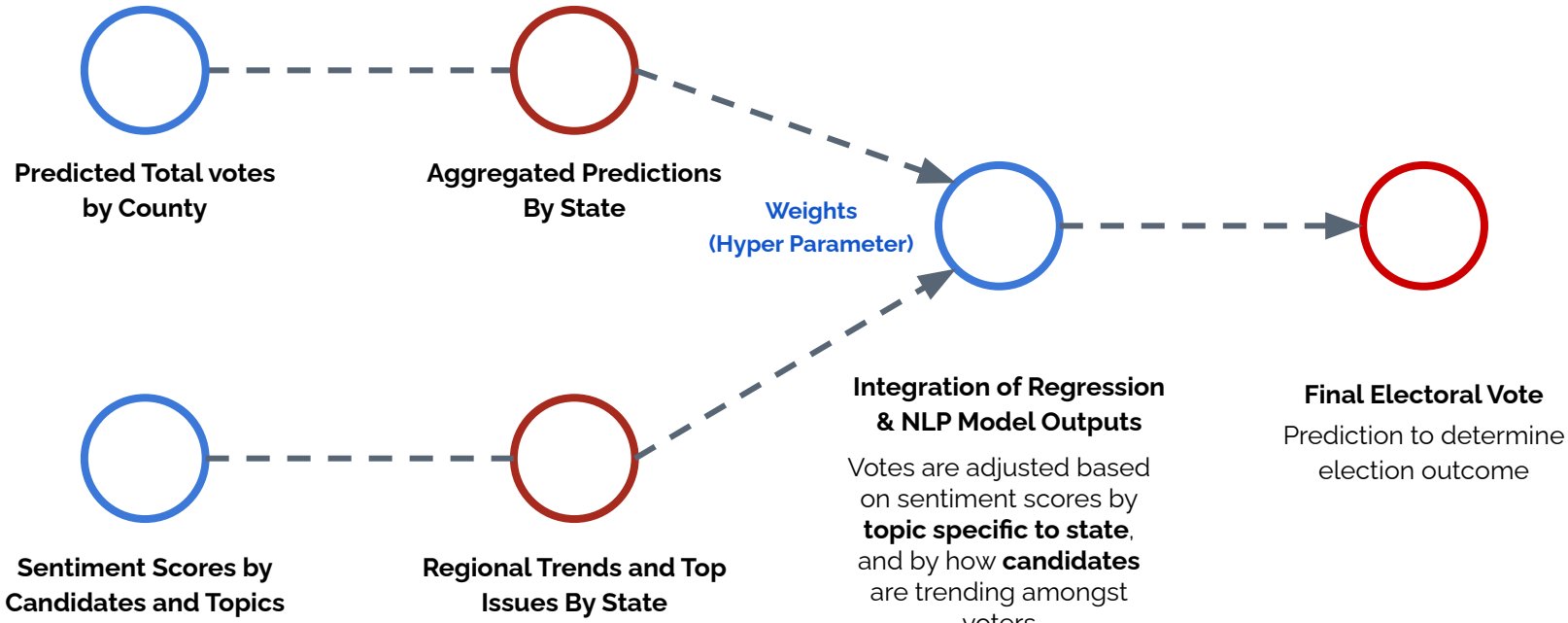
Parameter Tuning:

1. Grid search
2. Cross validation
3. Neural Network learning rate
4. Optimal tweet size

Final Feature Set:

- Total Population
- Total Votes (~Voter Turnout)
- Unemployment Rate
- Midterm elections outcome

Integration of Regression and NLP Sentiment Analysis



Model Evaluation: Accuracy and Backtesting On Past 2 Elections

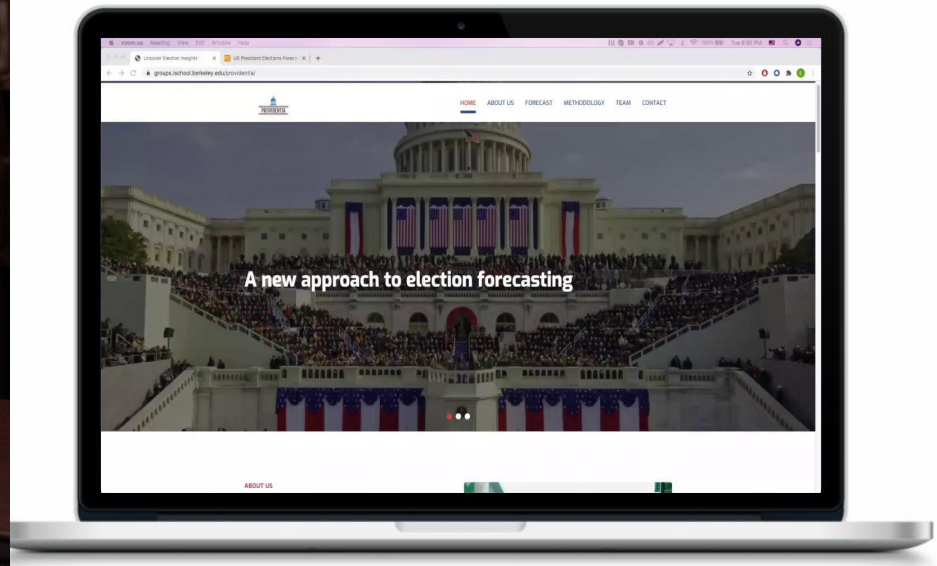
Accuracy defined as: # of correct state predictions/ all states

2012 (Romney vs. Obama)	Actual	Predicted	Accuracy
Democratic Electoral Votes	332	316	**82%**
Republican Electoral Votes	206	222	
2016 (Trump vs. Clinton)	**Actual**	**Predicted**	**Accuracy**
Democratic Electoral Votes	233	229	**88%**
Republican Electoral Votes	305	309	

PRELIMINARY 2020 ELECTION FORECAST & PRODUCT DEMO

Target Audience: campaign strategist for a political consultancy

Use Case: where does campaigns need to focus its attention and resources in order to secure victory?

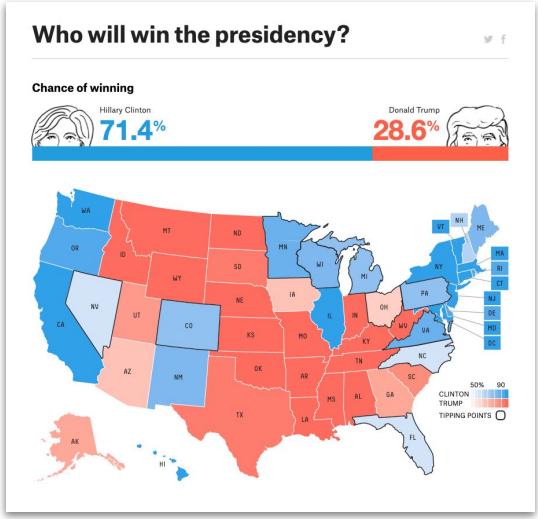


We vs. Them

Flipped winning party due to NLP/Trend adjusted sentiment score

Accurate prediction against 2016 actual results

538's 2016 Election Forecast



Providentia's Predicted 2016 Election Results

	Democrat		Republican		Comparison Against 2016 Actuals		
	Regression Prediction	NLP Adj. Prediction	Regression Prediction	NLP Adj. Prediction	538 Prediction	**Providentia** Prediction	Actual Winner
CO	1,282,494	1,168,931	988,758	1,102,320	✓ Democrat	✓ Democrat	Democrat
FL	4,474,914	4,066,054	3,702,272	✦ 4,111,131	Democrat	✓ Republican	Republican
IA	614,413	685,742	812,179	740,849	✓ Republican	✓ Republican	Republican
MI	2,591,647	2,360,214	2,036,997	2,268,429	Democrat	Democrat	Republican
MN	1,284,561	1,157,777	1,251,116	✦ 1,377,899	✓ Democrat	Republican	Democrat
NV	697,714	642,456	407,445	462,702	✓ Democrat	✓ Democrat	Democrat
NH	415,165	379,871	290,697	325,990	✓ Democrat	✓ Democrat	Democrat
NC	2,144,849	1,930,743	2,137,256	✦ 2,351,361	Democrat	✓ Republican	Republican
OH	2,741,897	2,474,911	2,597,810	✦ 2,864,795	✓ Republican	✓ Republican	Republican
PA	3,151,683	2,855,089	2,780,181	✦ 3,076,774	Democrat	✓ Republican	Republican
VA	1,806,485	1,619,038	1,942,441	2,129,887	✓ Democrat	Republican	Democrat

Key Technical Takeaways

1. **NLP to Process Twitter Data:**
   - Twitter data **lacks strict grammatical constructs**
     - Fortunately, it was correlated to the size of the tweets
     - **Removal of short tweets** improved model quality
   - **Stop words** play critical roles in training Sentiment Analysis for text with deep learning
   - **Aspect based sentiment score** is critical when a tweet or headline contains multiple candidates and/or multiple topics

2. **Geo-Mapping in Bokeh:**
   - **Limited number of tools** available for geo-mapping
   - Bokeh is a great choice if the visualization warrants limited interactivity
     - Optimizing Bokeh performance and **integration with HTML/Javascript** was challenging

3. **Data Engineering:**
   - Complexity integrating various sources of data
   - Limited experience with GCP, including integration of Bokeh; **dockerized the app environment** as a workaround

Key Learnings

Importance of Clear and Concise Communication & Storytelling

Well crafted stories can communicate complex and abstract ideas that encourage understanding and value connection

Diversity of Experience, Perspectives, and Skill Sets

Team composition and interdisciplinary interaction helped build on efficiencies and quality, while creating a unique learning experience

Future Applications

Main Topic	Category	Insights
Policy Impact	In-depth analysis of COVID impact	Analysis on voter turnout. States with higher number of cases prior to the election are likely to have lower voter turnouts. Some states may not allow vote by mail which could be a determining factor
Proactively Surface Insights (Pilot Sandbox)	Electoral vote optimization	Optimization tool on which swing states to invest resources and maximize electoral votes. Enabling scenario analysis for different combinations

Parse out undecided voters' voting trends
to enable actionable and targeted campaign tactics
in battleground states

Q&A Session