POLLS AND PRESS

STEPHEN GODFREY

PROBLEM STATEMENT

Explore and quantify the relationship between press coverage and polling performance

Data

Candidates

- Donald Trump
- John Kasich
- Ted Cruz
- Marco Rubio
- Ben Carson
- Jeb Bush
- Chris Christie
- Carly Fiorina
- Rick Santorum
- Rand Paul
- Mike Huckabee

Polls

- Source
  - FiveThirtyEight
- Summary
  - 670 polls between Jan 25, 2015 and May 3, 2016

Press

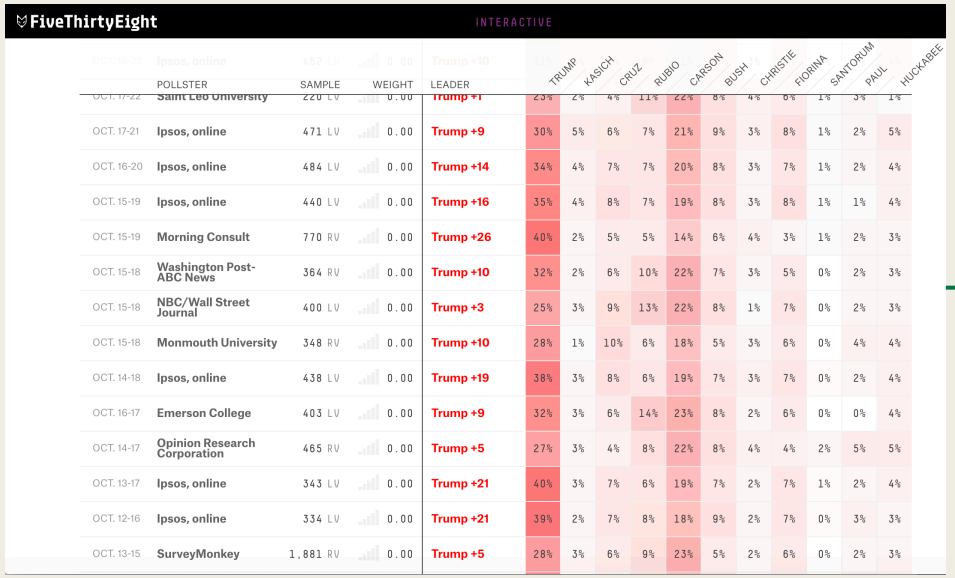
- Source
  - Global Data on Events, Language and Tone (GDELT) Global Knowledge Graph V2
- Summary
  - 5.3 MIn articles

Engineering

Data

Analysis

Polls


FiveThirtyEight

INTERACTIVE

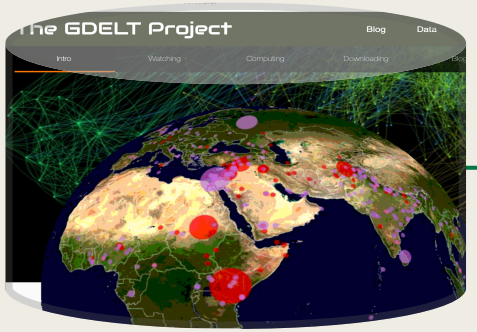
POLLSTER	SAMPLE	WEIGHT	LEADER	TRUMP	KASICH	CRUZ	RUBIO	CARSON	BUSH	CHRISTIE	FIORINA	SANTORUM	PAUL	HUCKABEE
OCT. 17-21 Ipsos, online	471 LV	0.00	Trump +9	30%	5%	6%	7%	21%	9%	3%	8%	1%	2%	5%
OCT. 16-20 Ipsos, online	484 LV	0.00	Trump +14	34%	4%	7%	7%	20%	8%	3%	7%	1%	2%	4%
OCT. 15-19 Ipsos, online	440 LV	0.00	Trump +16	35%	4%	8%	7%	19%	8%	3%	8%	1%	1%	4%
OCT. 15-19 Morning Consult	770 RV	0.00	Trump +26	40%	2%	5%	5%	14%	6%	4%	3%	1%	2%	3%
OCT. 15-18 Washington Post-ABC News	364 RV	0.00	Trump +10	32%	2%	6%	10%	22%	7%	3%	5%	0%	2%	3%
OCT. 15-18 NBC/Wall Street Journal	400 LV	0.00	Trump +3	25%	3%	9%	13%	22%	8%	1%	7%	0%	2%	3%
OCT. 15-18 Monmouth University	348 RV	0.00	Trump +10	28%	1%	10%	6%	18%	5%	3%	6%	0%	4%	4%
OCT. 14-18 Ipsos, online	438 LV	0.00	Trump +19	38%	3%	8%	6%	19%	7%	3%	7%	0%	2%	4%
OCT. 16-17 Emerson College	403 LV	0.00	Trump +9	32%	3%	6%	14%	23%	8%	2%	6%	0%	0%	4%
OCT. 14-17 Opinion Research Corporation	465 RV	0.00	Trump +5	27%	3%	4%	8%	22%	8%	4%	4%	2%	5%	5%
OCT. 13-17 Ipsos, online	343 LV	0.00	Trump +21	40%	3%	7%	6%	19%	7%	2%	7%	1%	2%	4%
OCT. 12-16 Ipsos, online	334 LV	0.00	Trump +21	39%	2%	7%	8%	18%	9%	2%	7%	0%	3%	3%
OCT. 13-15 SurveyMonkey	1,881 RV	0.00	Trump +5	28%	3%	6%	9%	23%	5%	2%	6%	0%	2%	3%

JS
HTML

Selenium
Beautiful Soup

Press

Google BigQuery



SQL

Global Knowledge Graph
(public)

- 935.3 Mln rows
- 9.85 TB

Staging table
(private)

- 5.4 Mln rows
- 0.02 TB

SQL

BigQuery

Modeling dataframe

Tools

Visualization
Gaussian Naïve Bayes
K Neighbors classifier
Random Forest classifier

Features

count
tone

polarity
...
themes

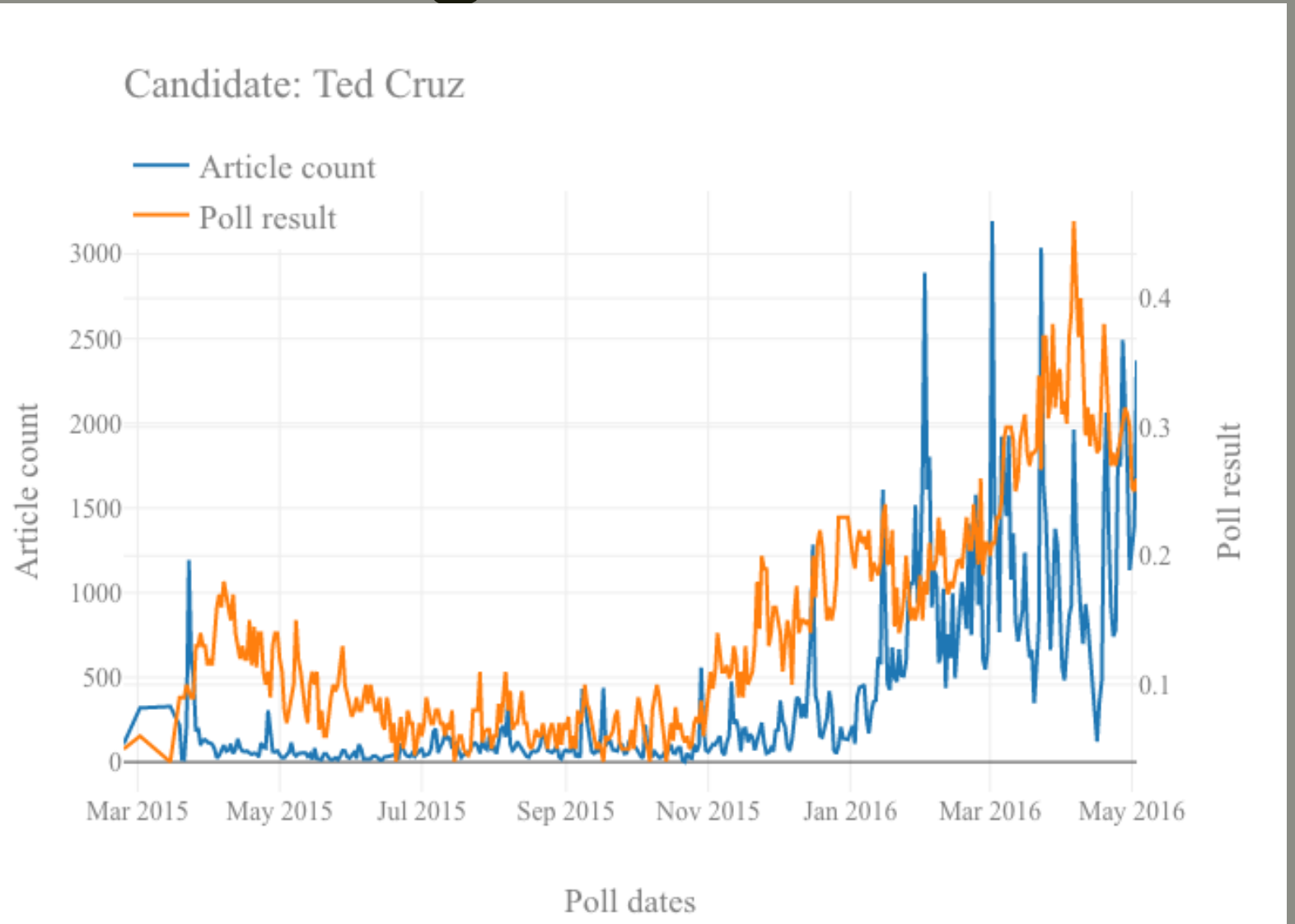
Framework

Multi-class classification

- down
- flat
- up

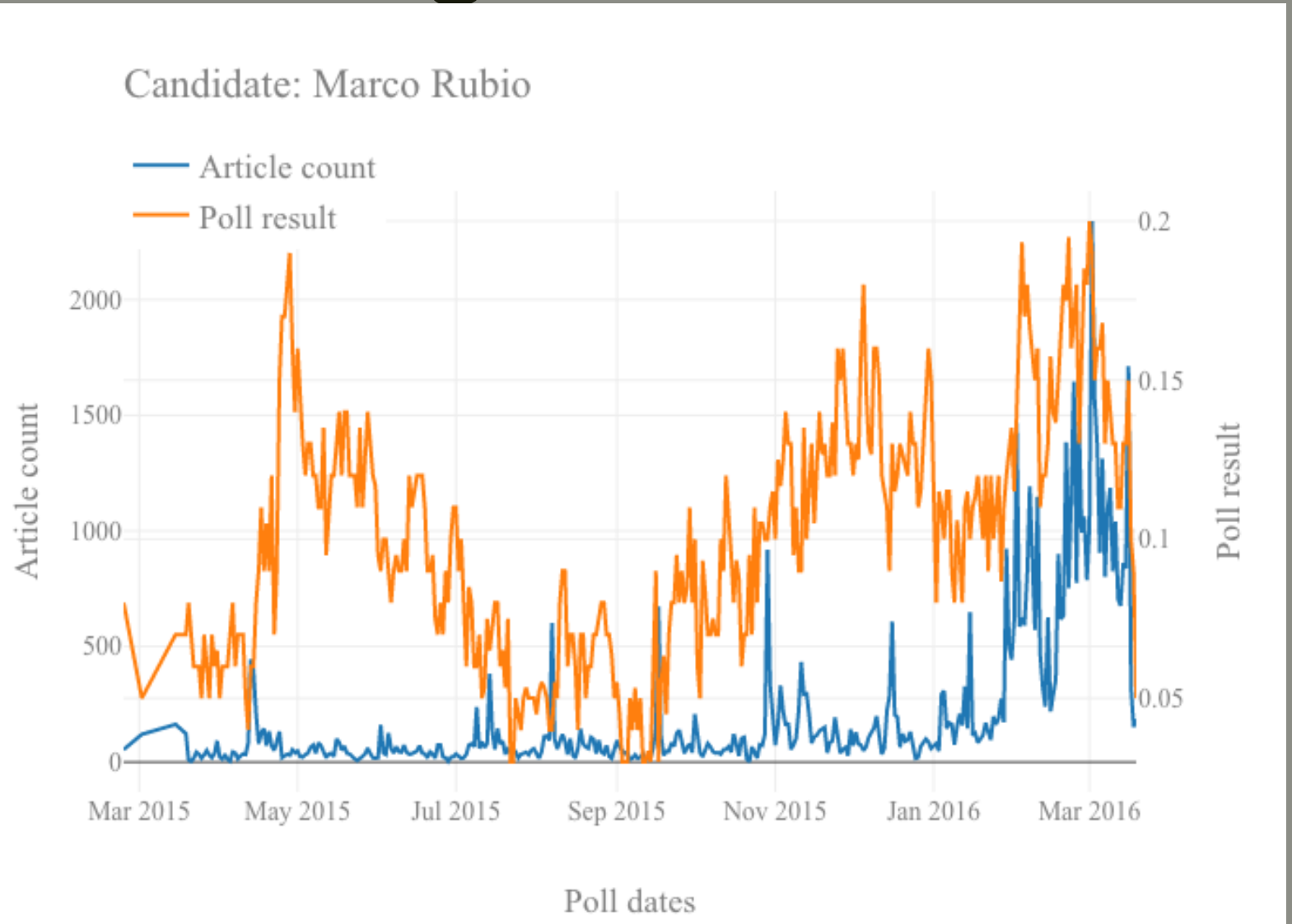
Python

Longitudinal view



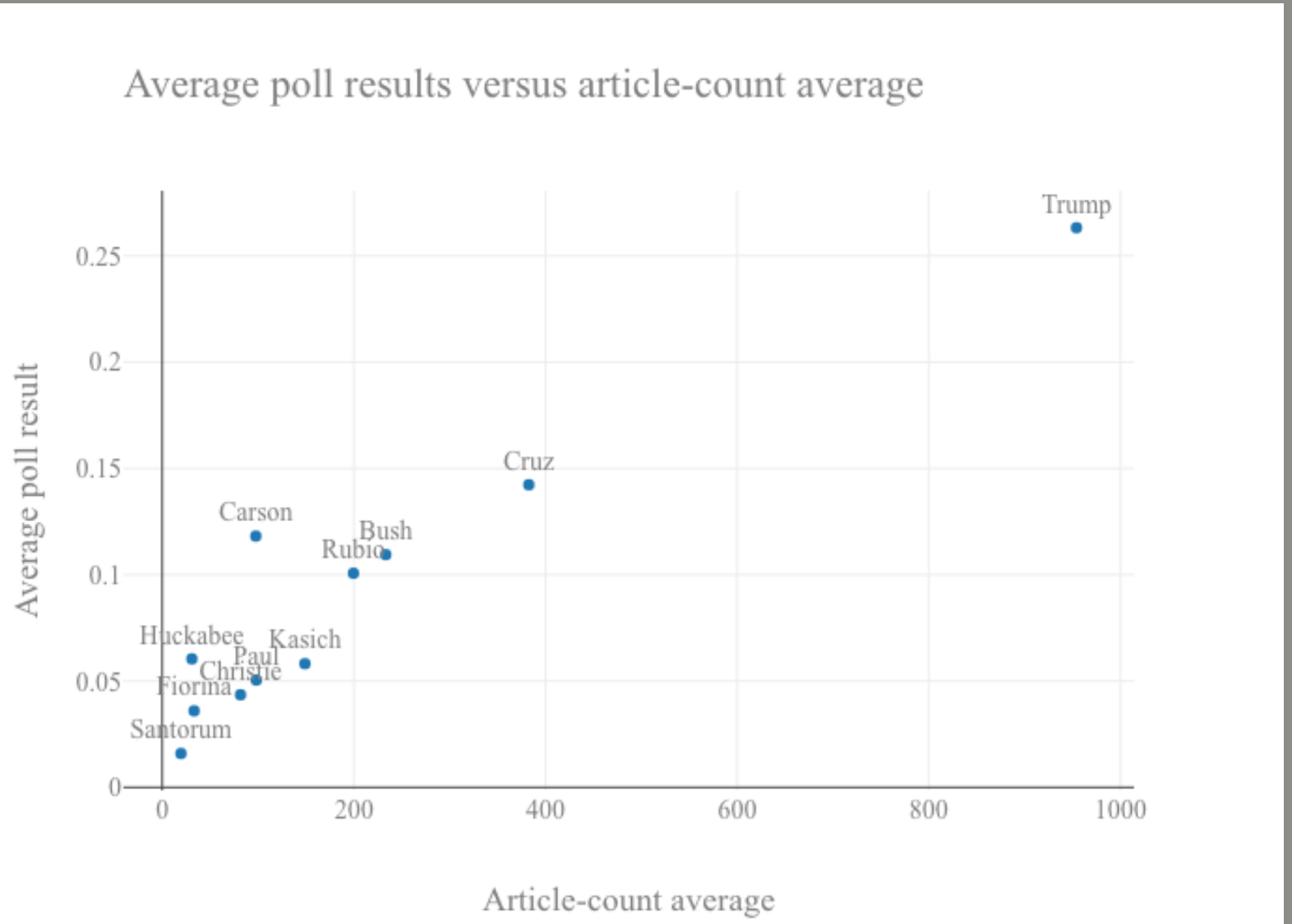
- Observations
  - *Apparent positive relationship*
  - *Coverage increases in the later stages of the race*
  - *Some variation in poll-to-poll results exist but a trend factor also seems present*

Longitudinal view



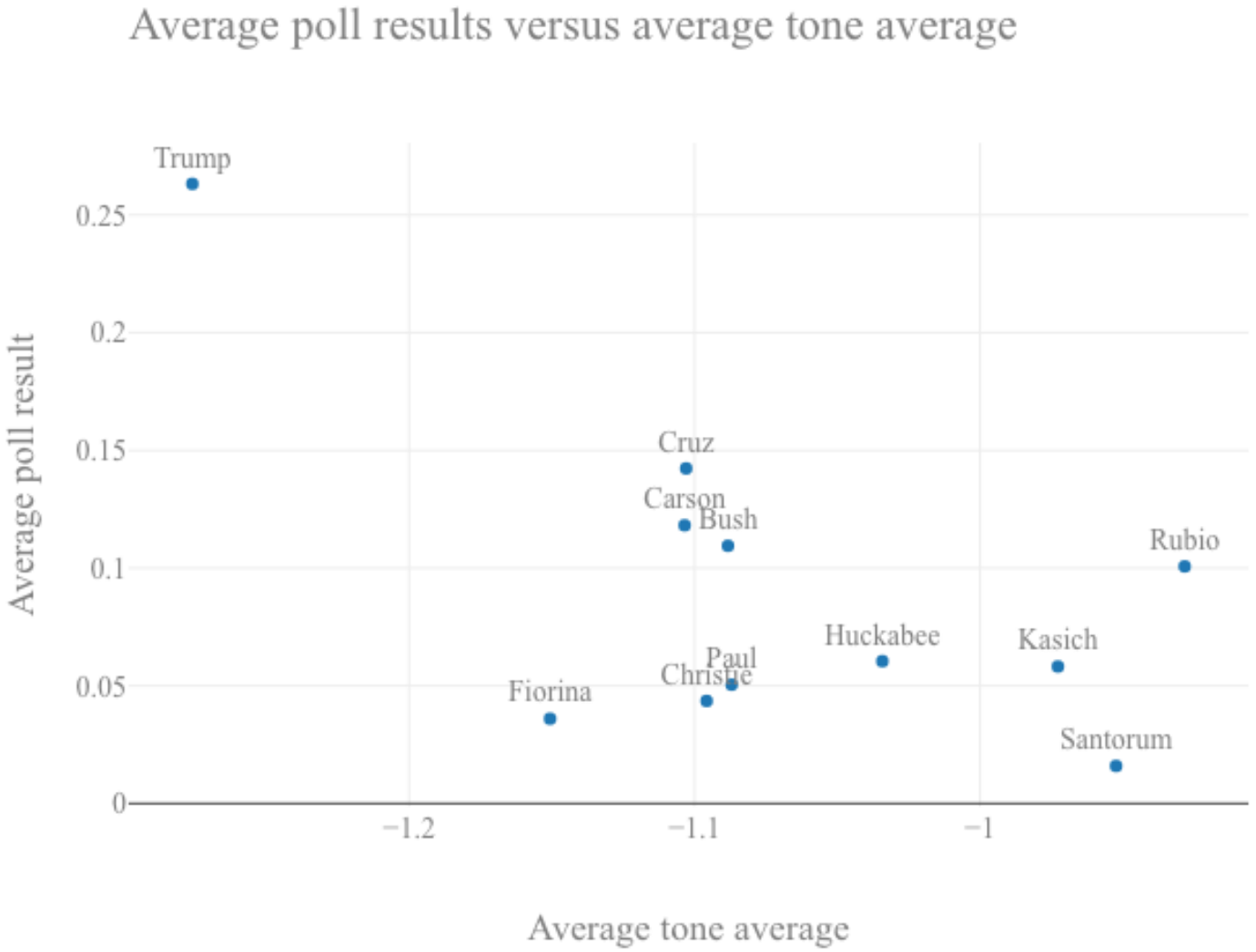
- Observations
  - *Press coverage increased as the race progresses even for a range-bound candidate*

Horizontal view



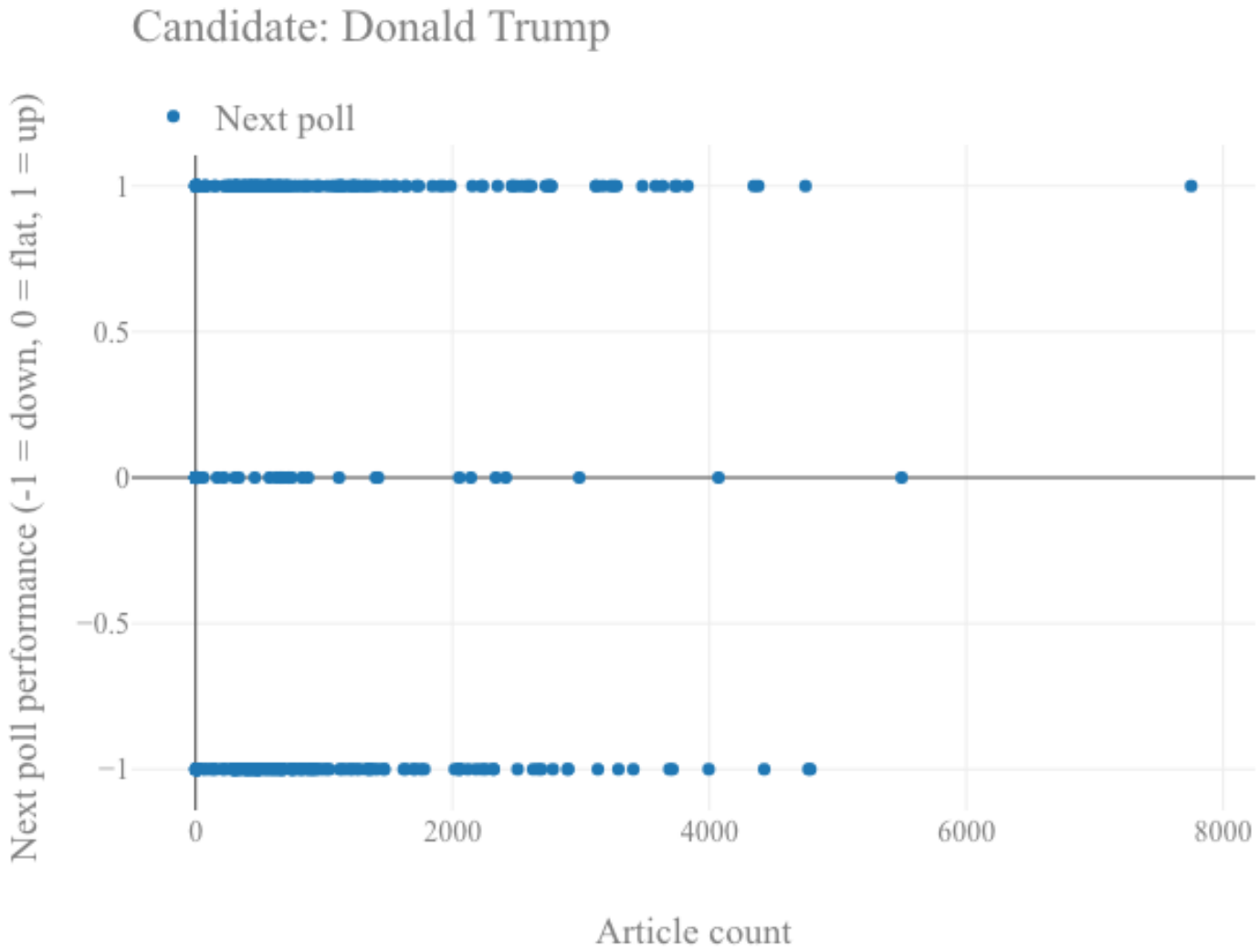
- Observations
  - *Across all candidates, an intuitive positive relationship between poll performance and press coverage seems present*
  - *Still, Trump is a noticeable outlier with substantial press coverage and the highest poll average*

Horizontal view



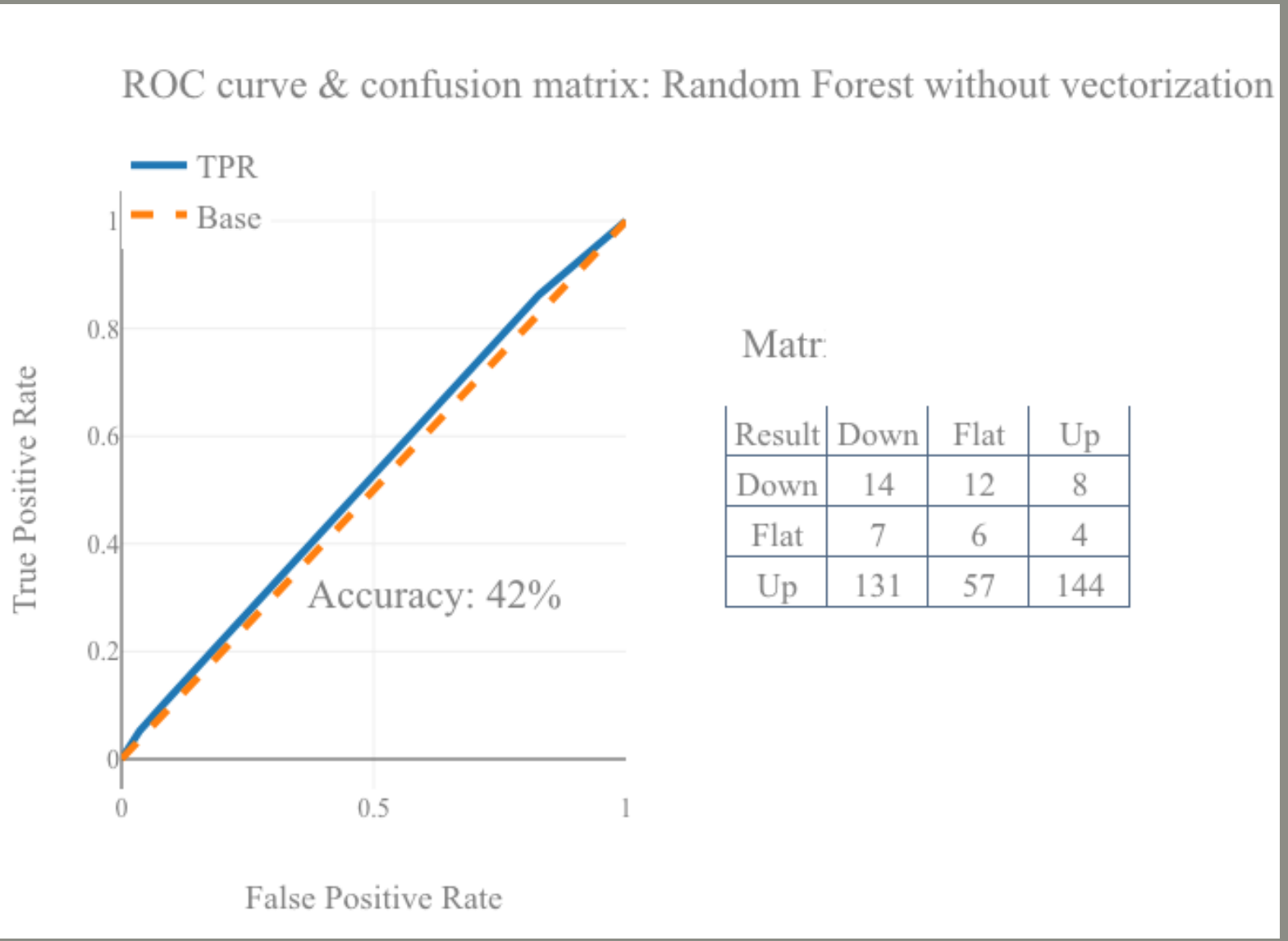
- Observations
  - *Less of pattern is present when poll performance is plotted versus average of the tone average*
  - *Still, Trump is a noticeable outlier with negative average coverage yet the highest poll average*

The target variable



- Observations
  - *Comparing the target variable - down, up or flat – to article count shows few distinguishable patterns and highlights modeling challenges*

Model evaluation

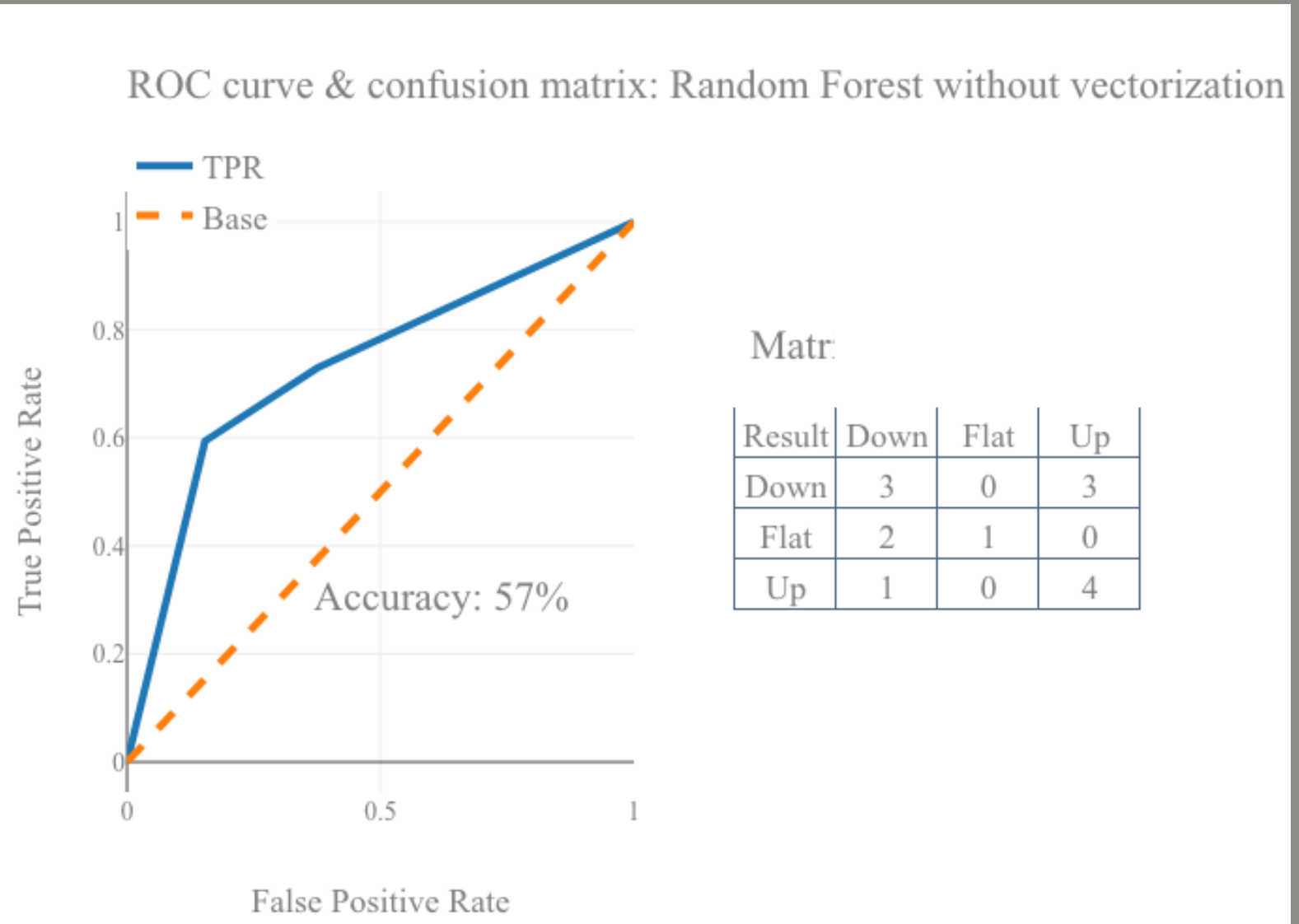


Result	Down	Flat	Up
Down	14	12	8
Flat	7	6	4
Up	131	57	144

- Observations
  - *Most classification models show little predictive power*

- Model parameters
  - *Data - All candidates*
  - *Target - Next poll*
  - *Features - Article count*
  - *Model - Random Forest*

Model evaluation



Result	Down	Flat	Up
Down	3	0	3
Flat	2	1	0
Up	1	0	4

- Observations
  - *Some evidence that predictive power improves when the model is focused on specific candidates and looks to future polls (small test sample)*
- Model parameters
  - *Data - Donald Trump*
  - *Target - Third poll*
  - *Features - Article count, Tone avg, Polarity avg*
  - *Model - Random Forest*

CONCLUSION

1. Building the dataset was a challenge, but it produces some interesting longitudinal and horizontal visualizations

2. Broad press metrics averaged across a wide candidate field provide little power in predicting performance in upcoming polls

3. More selectively modeling including focusing on specific candidates or selected publications may be more fruitful