Unclear Politics

What's going on with Real Clear Politics' poll aggregator?

Background information

- Donald Trump's surprise victory - defying expected results in key battleground states - placed the spotlight on polls & polling methods

- Can polls be trusted?

- Polling methodology

Key errors in 2016

- National polls actually weren't off by a huge amount and accurately predicted Clinton winning the national popular vote

- Crucial shift in voting patterns re: education level

Improving forecasting... but how?

- Many reputable pollsters have adjusted for this by weighting their polling sample by education now

- However, the uncertainty caused by 2016 opened the door to alternative and sometimes... questionable "innovations

- E.g. Trafalgar's 'social desirability' approach

But what about the meta?

- There are also different approaches to assessing the overall field of polls

- Highly sophisticated, highly transparent, but also still somewhat subjective: weighting the **pollsters** (FiveThirtyEight)

- Low sophistication, but also low transparency - poll **aggregators** (Real Clear Politics)

Problems with RCP

- Polling average is relatively unsophisticated
    - Taking a pure average would be more accurate if all pollsters used the same methodology

- Unclear who, when, why meets RCP's criteria for inclusion
    - Frequent changes and exclusive focus on "current" average vs trends make it hard to draw conclusions about shifts over time, and/or judge a pollster's *predictive* power

- Gives the false impression that all polls are created equal -- and prioritizes providing massive amounts of data over providing insights & context

My goals

- What - if any - insights about the **presidential race** can we derive from RCP's current selection of polls? Does RCP's approach to poll aggregation provide us with the means to make informed conclusions about **cumulative trends**?

- What can we say about how this might impact the reliability of the RCP Average as an analytical and/or predictive tool?

- Finally, what kind of basic insights can we draw about individual pollsters from the available list?

Why this matters

- As important as accurate polling + sampling methods are, even more important is how that information is presented

- Presenting casual viewers and non-experts with an overwhelming amount of information, without context and/or with an inaccurate representation of key takeaways, **does more harm than good**

- Places additional doubt on the usefulness of polling when actual outcomes are different than the above information

- At its most extreme, it also casts doubt on the integrity of the system when things go awry
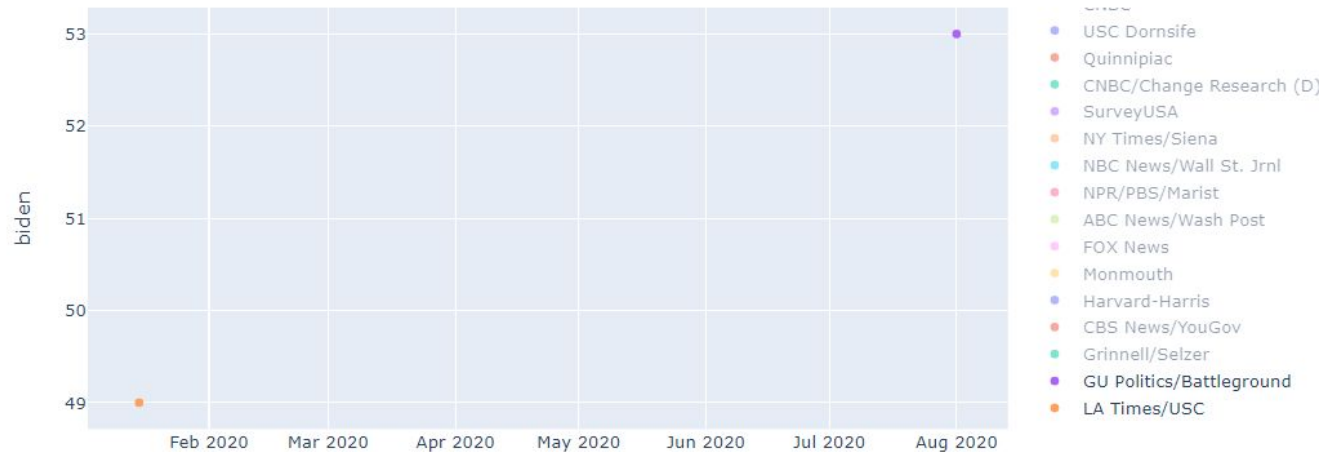
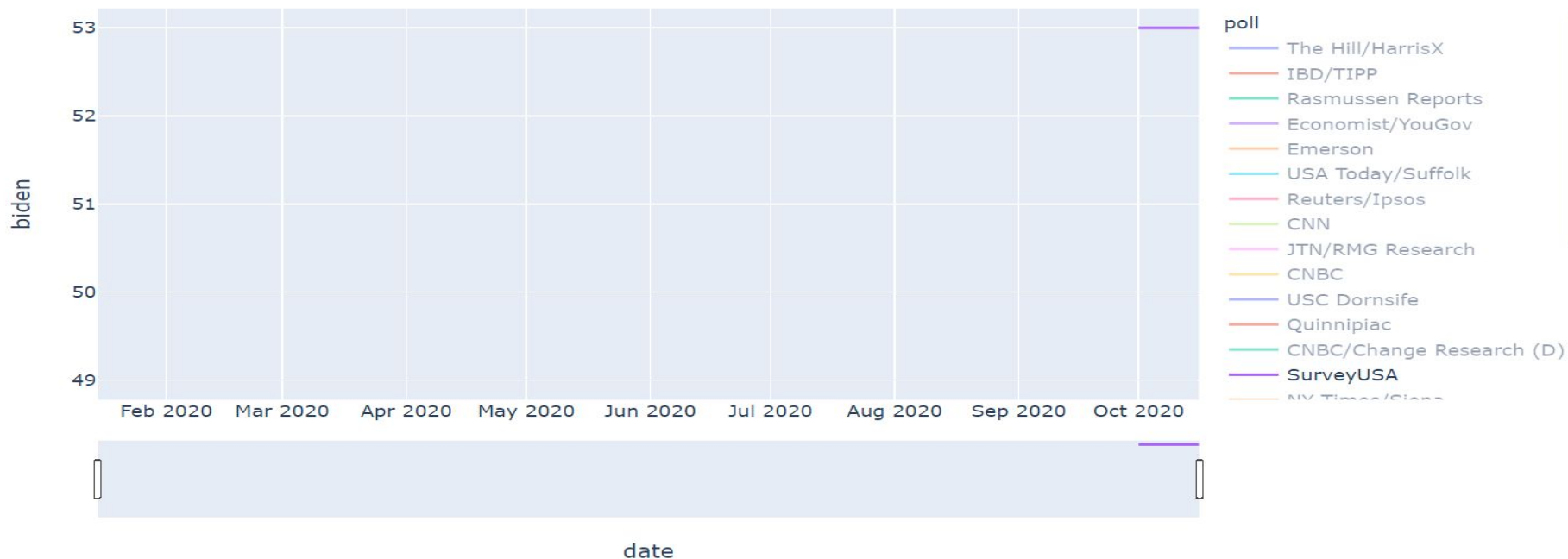# The polls

What's going on with the variability?

| poll	biden	trump
ABC News/Wash Post	2.133910	1.414214
CBS News/YouGov	2.309401	1.272418
CNBC	2.258318	2.786874
CNBC/Change Research (D)	1.414214	1.000000
CNN	2.179449	1.900292
Economist/YouGov	2.026562	1.131308
Emerson	1.964971	3.000000
FOX News	2.828427	2.118700
GU Politics/Battleground	NaN	NaN
Grinnell/Selzer	1.414214	1.414214
Harvard-Harris	2.825269	1.187735
IBD/TIPP	2.174229	2.576114
JTN/RMG Research	1.133893	0.786796
LA Times/USC	NaN	NaN
Monmouth	1.767767	1.908627
NBC News/Wall St. Jrnl	1.490712	1.632993
NPR/PBS/Marist	1.602082	0.752773
NY Times/Siena	0.577350	2.886751
Quinnipiac	1.301708	1.787301
Rasmussen Reports	1.581139	2.073802
Reuters/Ipsos	1.581139	0.927961
SurveyUSA	0.000000	0.000000
The Hill/HarrisX	1.856310	1.648822
USA Today/Suffolk	1.500000	1.825742
USC Dornsife	1.154701	0.000000

What's going on with random inclusions?

What's going on with random inclusions?

Poll clustering?

- Results from different pollsters start clustering in on each other as we near the end of an election season
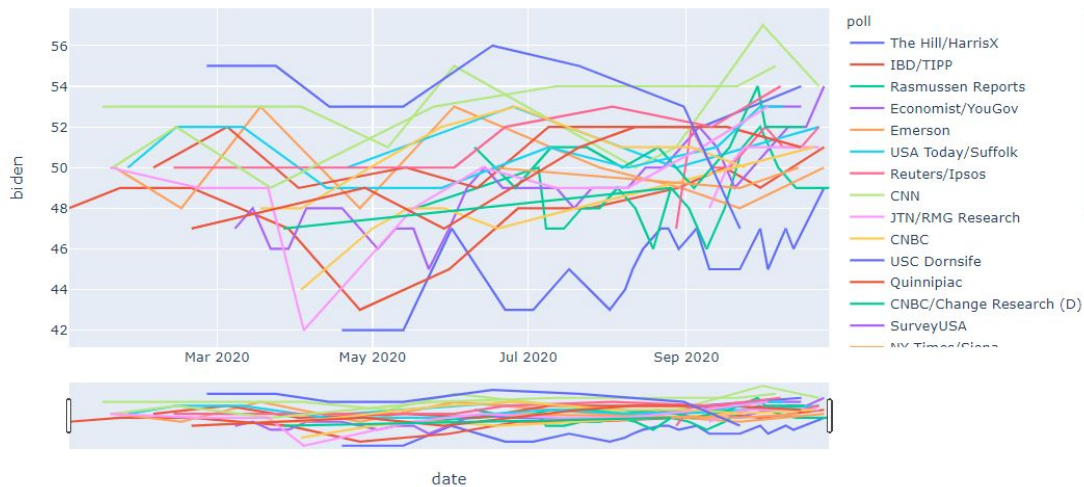
Possible poll inclusion strategy

- Not much can be gleaned from RCP data as a base level of analysis -- and that's part of the problem

- The bulk of the included polls trend toward the same results/range, but RCP also includes outlier polls on both directions that provide drastically different views of the race / reality than the rest of the data

- This may have been the exact strategy behind their selections - aim to provide a majority of polls with similar variation + results, but randomly toss in outlier polls to "cover all bases"

Next Steps

- Time series analysis
  - Plot a moving average (likely 2-week intervals) for each individual poll
  - Plot standard deviation lines to get a better grasp of variability

- Finish building out a Plotly Dash app to allow users to explore how including/excluding some of RCP's chosen polls would affect outcomes