Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- 2016 Election!
- Homework (projects!)

Analysis of pollsters

As before, the ratings are based both on a pollster's past accuracy and on two easily measurable methodological standards:

The first standard is whether the firm participates in the American Association for Public Opinion Research Transparency Initiative, is a member of the National Council on Public Polls or contributes its data to the Roper Center for Public Opinion Research archive. Polling firms that do one or more of these things generally abide by industry-standard practices for disclosure, transparency and methodology and have historically had more accurate results.

The second standard is whether the firm usually conducts its polls by placing telephone calls with live interviewers and calls cellphones as well as landlines. Automated polls ("robopolls"), which are legally prohibited from calling cellphones, do not meet this standard even if they use hybrid or mixed-mode methodologies (for example, robocalling landlines and then supplementing with cellphone calls placed by live interviewers).[2] It's increasingly essential to call cellphones given that about half of American households no longer have a home landline. Although internet polls show promise as a potential alternative, they do not yet have a long enough or consistent enough track record to be placed on the same pedestal as high-quality, live-interview telephone polls, based on our view of the evidence.

http://fivethirtyeight.com/features/the-state-of-the-polls-2016/

How the most prolific pollsters have performed since 2014

POLLSTER	NO. OF POLLS	NCPP/ AAPOR/ ROPER	LIVE CALLER W/CELL	ONLINE	ADVANCED PLUS-MINUS
YouGov	476			✓	+0.2
Public Policy Polling	87			*	-0.4
Marist College	46	✓	✓		-1.5
Monmouth University	40	✓	✓		-1.7
SurveyUSA	40	✓		*	+0.8
Quinnipiac University	36	✓	✓		-0.8
American Research Group	36		✓		+0.7
Gravis Marketing	35				-0.2
Rasmussen Reports	34			*	+0.5
Emerson College	32	✓			+0.7
Fox News	20	✓	✓		-1.0

**Public Policy Polling, SurveyUSA and Rasmussen Reports supplement automated calls placed to landline phones with an online panel*

Simple Average Error for polls in the 21 days before an election

	PRESIDENTIAL		STATE-LEVEL		
YEAR	PRIMARY	GENERAL	GOVERNOR	U.S. SENATE	U.S. HOUSE
1998	–	–	8.2	6.8	6.8
2000	7.8	4.5	3.7	5.5	4.5
2002	–	–	5.4	4.5	5.6
2004	7.2	3.2	4.3	5.1	5.0
2006	–	–	4.6	4.3	5.8
2008	7.4	3.4	4.6	5.0	5.8
2010	–	–	4.9	5.4	6.5
2012	8.7	3.6	4.6	4.9	4.8
2014	–	–	4.5	5.4	7.9
2016	9.4	–	–	–	–
All years	**8.1**	**3.6**	**5.1**	**5.1**	**6.4**

FIVETHIRTYEIGHT

Share of polls that correctly picked the winner

	PRESIDENTIAL		STATE-LEVEL		
YEAR	PRIMARY	GENERAL	GOVERNOR	U.S. SENATE	U.S. HOUSE
1998	–	–	84%	87%	62%
2000	94%	70%	84	84	59
2002	–	–	85	82	82
2004	94	80	80	85	71
2006	–	–	91	92	74
2008	80	92	95	96	84
2010	–	–	85	79	80
2012	61	78	91	87	75
2014	–	–	76	75	91
2016	85	–	–	–	–
All years	**81**	**81**	**85**	**85**	**81**

FIVETHIRTYEIGHT

Statistical bias in polls

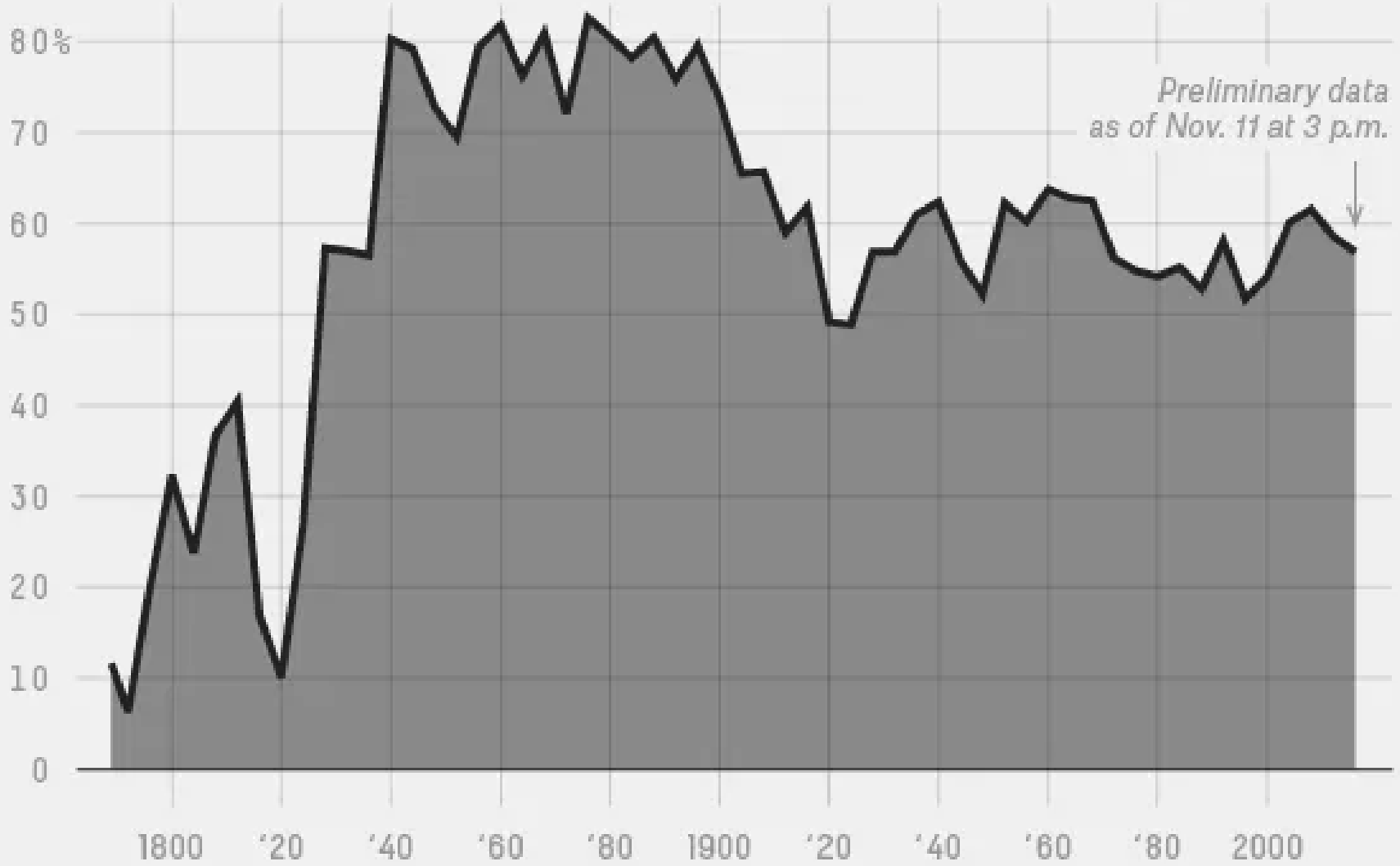
YEAR	PRESIDENTIAL GENERAL	GOVERNOR	U.S. SENATE	U.S. HOUSE
1998		R+6.1	R+4.3	R+1.9
2000	R+2.2	R+0.5	R+2.6	D+1.2
2002		D+3.4	D+1.5	D+2.0
2004	D+1.0	R+1.2	D+0.5	D+2.2
2006		R+0.6	R+2.0	D+0.3
2008	D+0.1	R+1.5	D+0.7	D+1.4
2010		R+1.3	R+2.4	D+1.0
2012	R+2.5	R+2.4	R+3.4	R+2.6
2014		D+2.7	D+3.0	D+3.3
All years	**R+0.9**	**R+0.0**	**R+1.0**	**D+1.5**

FIVETHIRTYEIGHT

Election results

Turnout in U.S. presidential elections

As a share of eligible voters

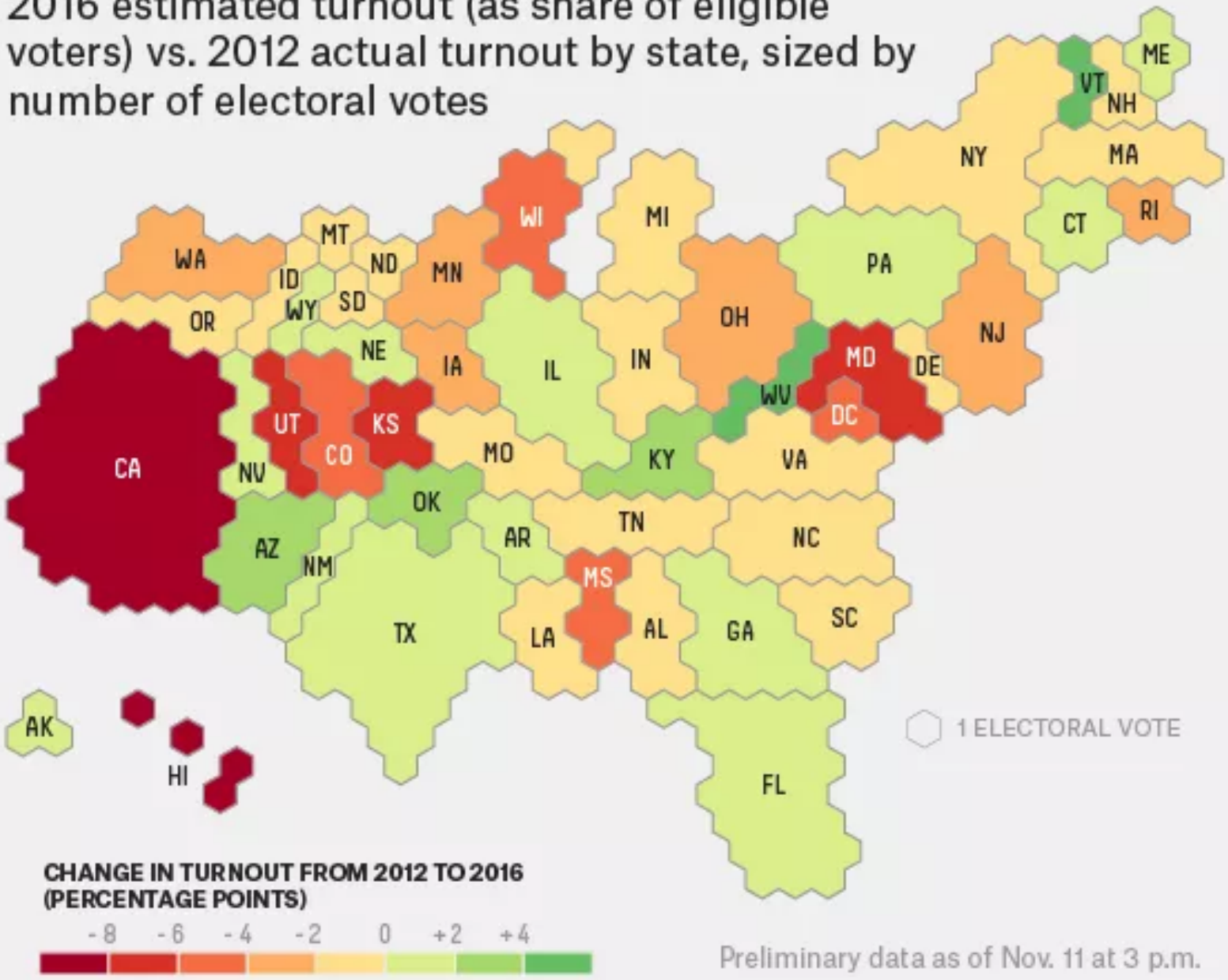


FIVETHIRTYEIGHT

SOURCE: U.S. ELECTIONS PROJECT

How turnout changed from 2012

2016 estimated turnout (as share of eligible voters) vs. 2012 actual turnout by state, sized by number of electoral votes



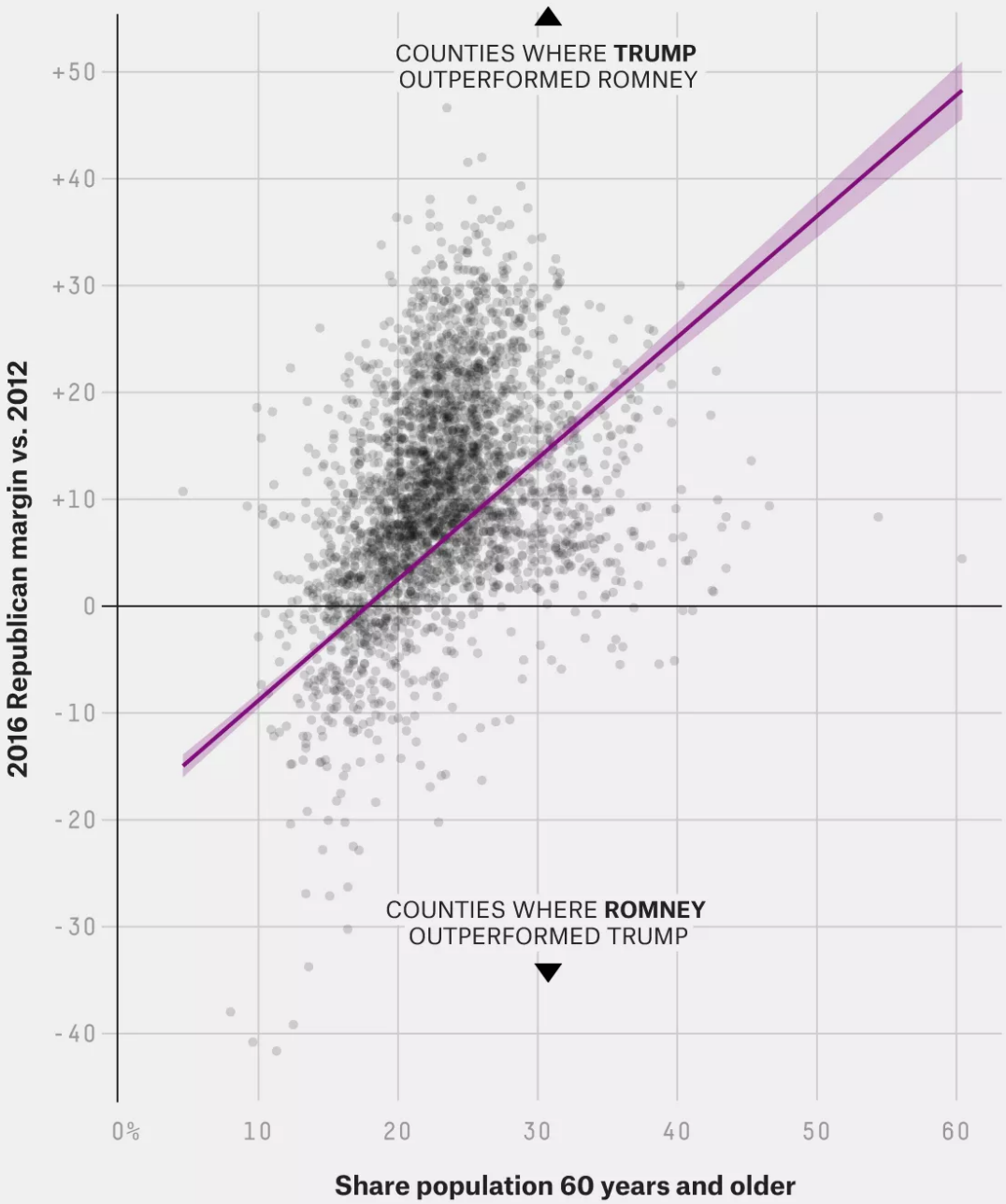
Preliminary data as of Nov. 11 at 3 p.m.

FIVETHIRTYEIGHT

SOURCE: U.S. ELECTIONS PROJECT

Trump outperformed in older counties

Trump's vote share relative to Romney's vs. share of population 60 years and older by county

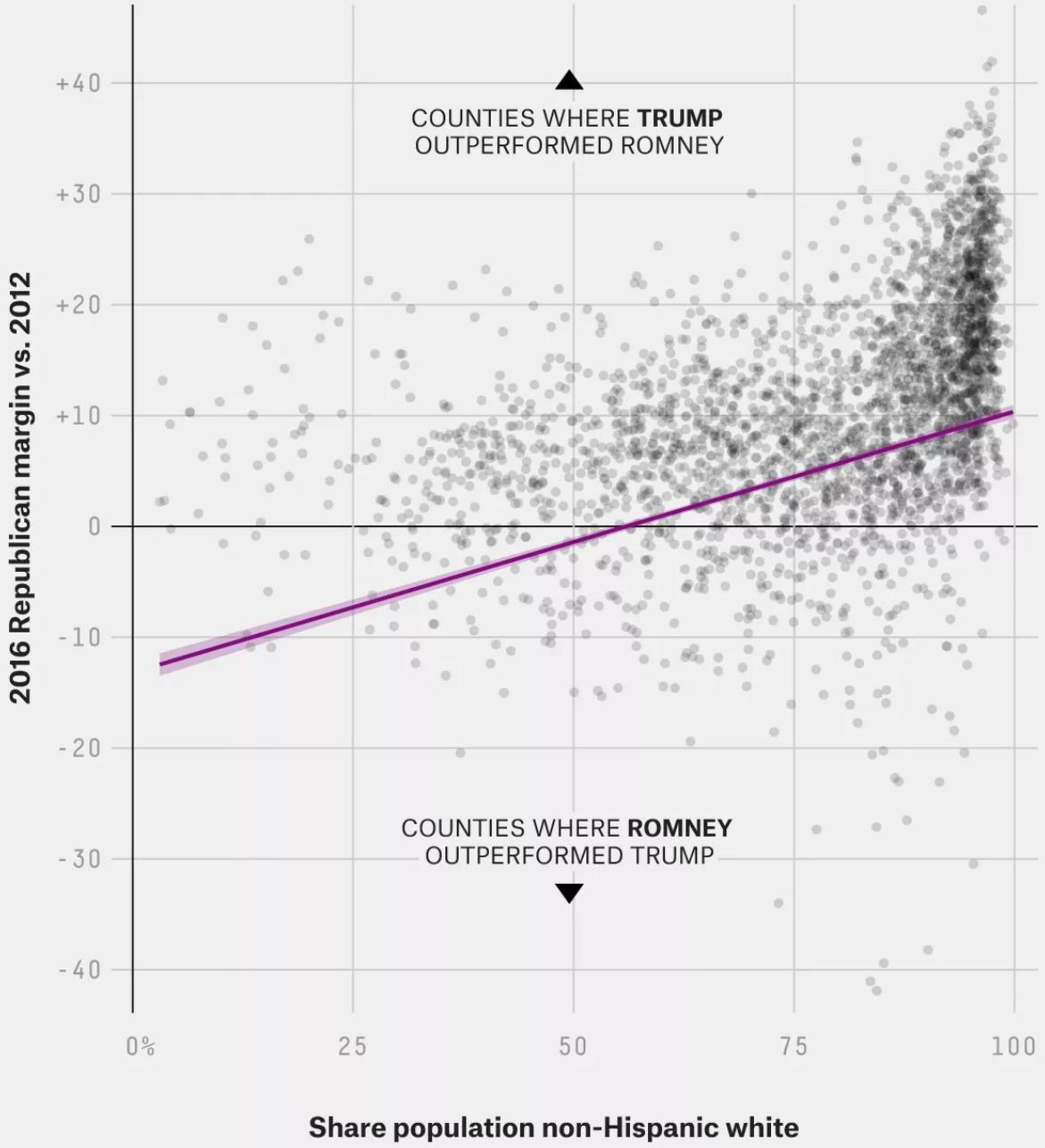


FIVETHIRTYEIGHT

SOURCES: DAVE LEIP'S PRESIDENTIAL ELECTION ATLAS, AMERICAN COMMUNITY SURVEY

Trump did best in the whitest counties

Trump's vote share relative to Romney's vs. share of population that is non-Hispanic white by county

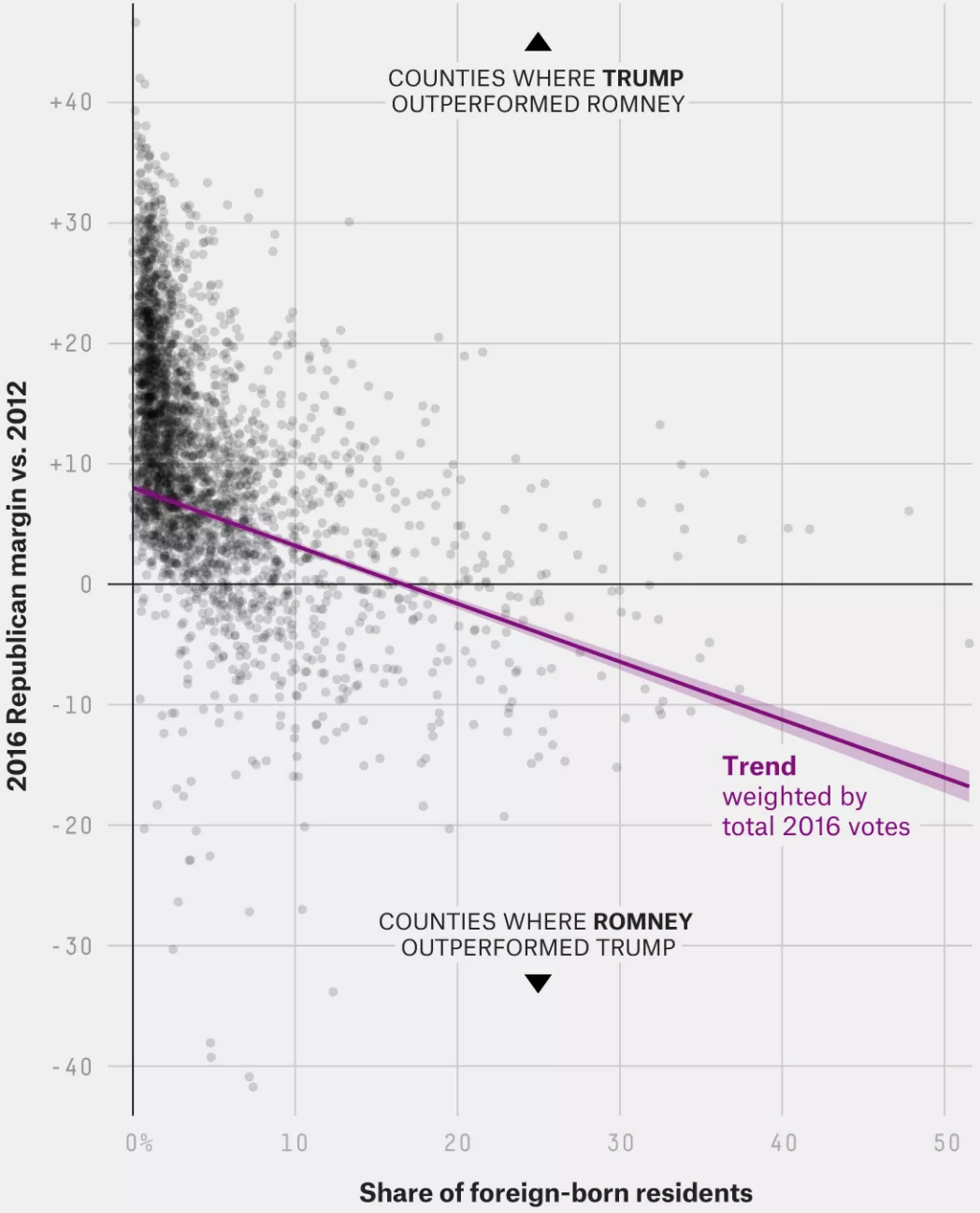


FIVETHIRTYEIGHT

SOURCES: DAVE LEIP'S PRESIDENTIAL ELECTION ATLAS, AMERICAN COMMUNITY SURVEY

Trump did best in places with fewer immigrants

Trump's vote share relative to Romney's vs. share of foreign-born residents by county

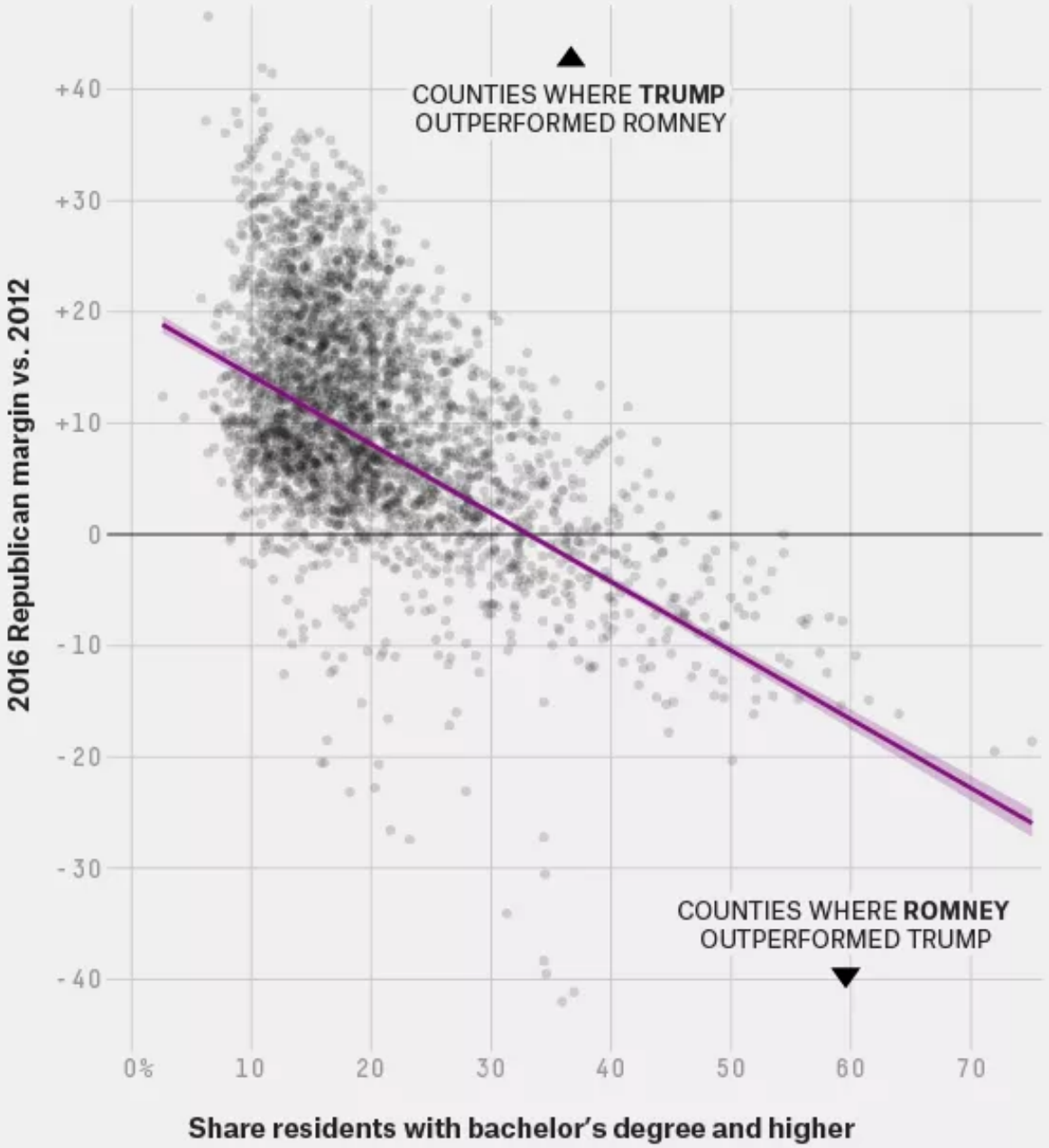


FIVETHIRTYEIGHT

SOURCES: DAVE LEIP'S PRESIDENTIAL ELECTIONS ATLAS, AMERICAN COMMUNITY SURVEY

Trump outperformed in less educated places

Trump's vote share relative to Romney's vs. share of population bachelor's degree and higher by county



FIVETHIRTYEIGHT

SOURCES: DAVE LEIP'S PRESIDENTIAL ELECTION ATLAS, AMERICAN COMMUNITY SURVEY

Why results missed the mark

http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/

One likely culprit is what pollsters refer to as nonresponse bias. This occurs when certain kinds of people systematically do not respond to surveys despite equal opportunity outreach to all parts of the electorate. We know that some groups – including the less educated voters who were a key demographic for Trump on Election Day – are consistently hard for pollsters to reach.

Why results missed the mark

Some have also suggested that many of those who were polled simply were not honest about whom they intended to vote for. If this were the case, we would expect to see Trump perform systematically better in online surveys, as research has found that people are less likely to report socially undesirable behavior when they are talking to a live interviewer. Politico and Morning Consult conducted an experiment (http://www.politico.com/story/2016/11/poll-shy-voters-trump-230667) to see if this was the case, and found that overall, there was little indication of an effect, though they did find some suggestion that college-educated and higher-income voters might have been more likely to support Trump online.

Why results missed the mark

A third possibility involves the way pollsters identify likely voters. Because we can't know in advance who is actually going to vote, pollsters develop models predicting who is going to vote and what the electorate will look like on Election Day. This is a notoriously difficult task, and small differences in assumptions can produce sizable differences in election predictions (http://www.pewresearch.org/2016/01/07/can-likely-voter-models-be-improved/) .

Possible explanations

https://www.yahoo.com/news/the-failure-of-election-polling-was-about-3-key-things-182053428.html?ref=gs

1. Polls did not fully account for the Shy Trump Voter

White women, in particular, proved to be a surprise: 53% of them voted for Trump overall, led by those without a college degree, who went for Trump by a 2-1 margin. White women with a college degree went for Clinton, but only barely, by six percentage points.

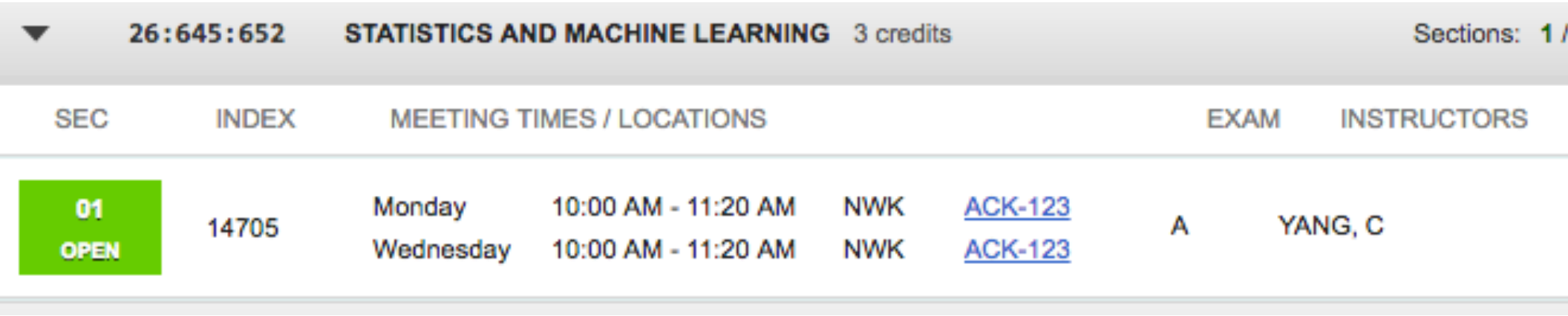
2. Polling methods need to change

In 2003, Gallup wrote a post about the falling response rates in polls. If you start with a target sample size of 1,000 households, Gallup wrote, at least 200 households fall out because they are businesses or non-working numbers. Of the 800 left, another 200 "may be unreachable in the time frame allocated by the researcher… household members at these numbers may use caller ID or other screening devices and refuse to answer." Now you're down to 600, of which 200 more people may pick up the phone but refuse to participate in the poll. Suddenly, the sample size has shrunk from 1,000 to a mere 400 households.

3. The bigger failure was interpretation of the polls

After an initial immediate backlash to the polls, a newer narrative is already emerging: the polls didn't fail as terribly as everyone is saying they did.

Spring course!



26:645:652 **STATISTICS AND MACHINE LEARNING** 3 credits Sections: 1 /

SEC	INDEX	MEETING TIMES / LOCATIONS				EXAM	INSTRUCTORS
01 OPEN	14705	Monday Wednesday	10:00 AM - 11:20 AM 10:00 AM - 11:20 AM	NWK NWK	ACK-123 ACK-123	A	YANG, C

Enroll!

Homework for Monday

- Speed talks: 2 minutes each + 1 question
- (No slides!)
- You will lose points if you go over time. Give your pitch in the time allotted.
- If you aren't here when I call your name, you will get a zero.

Projects!

First round due Nov 18th by 11:59 pm

Will then get comments

Give presentation in class

Turn in final assignment Dec 14 11:59 pm