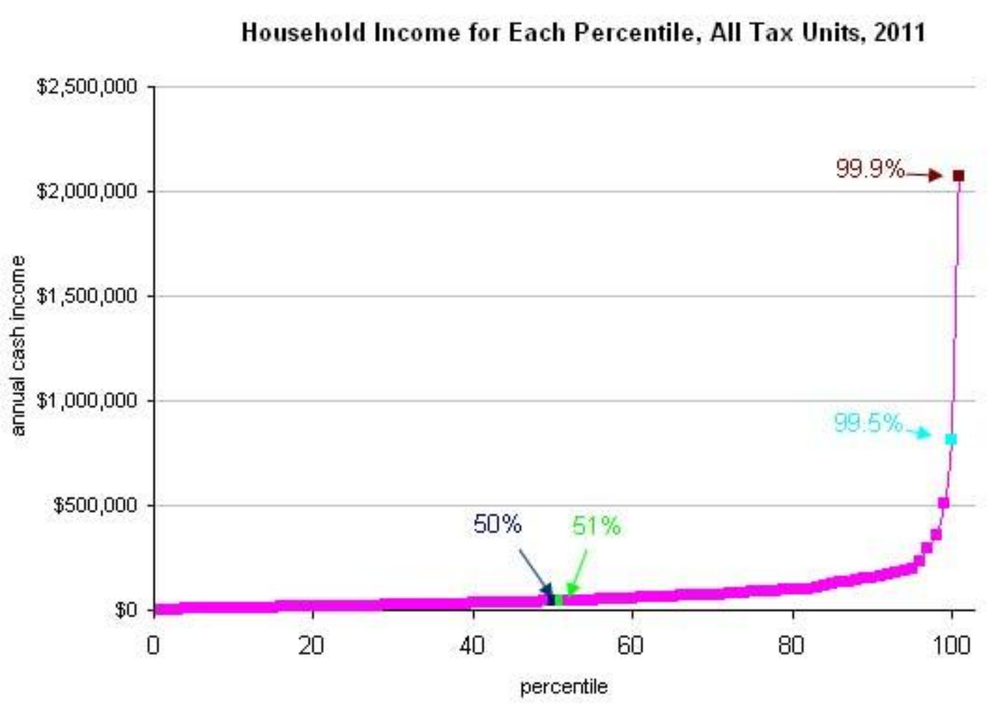
What is data?

Week One - 9 January 2019

C-SoDA:
The Center for
Social Data Analytics
Lecture Hall 10 is to the left

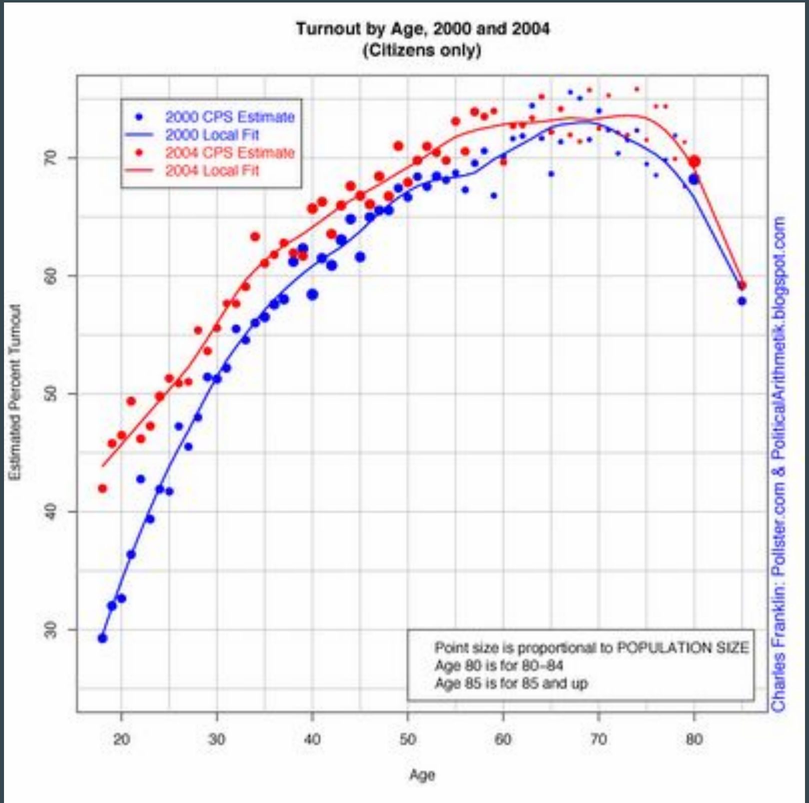
Types of data

- Continuous - real numbers



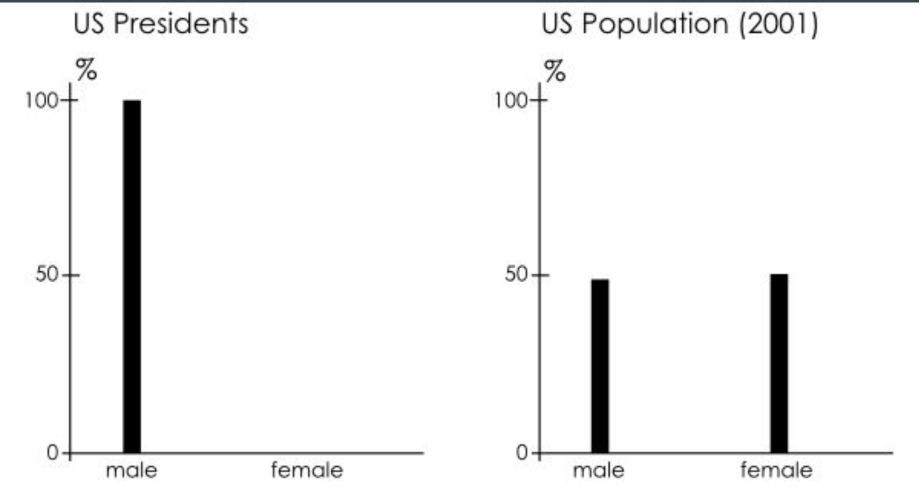
Types of data

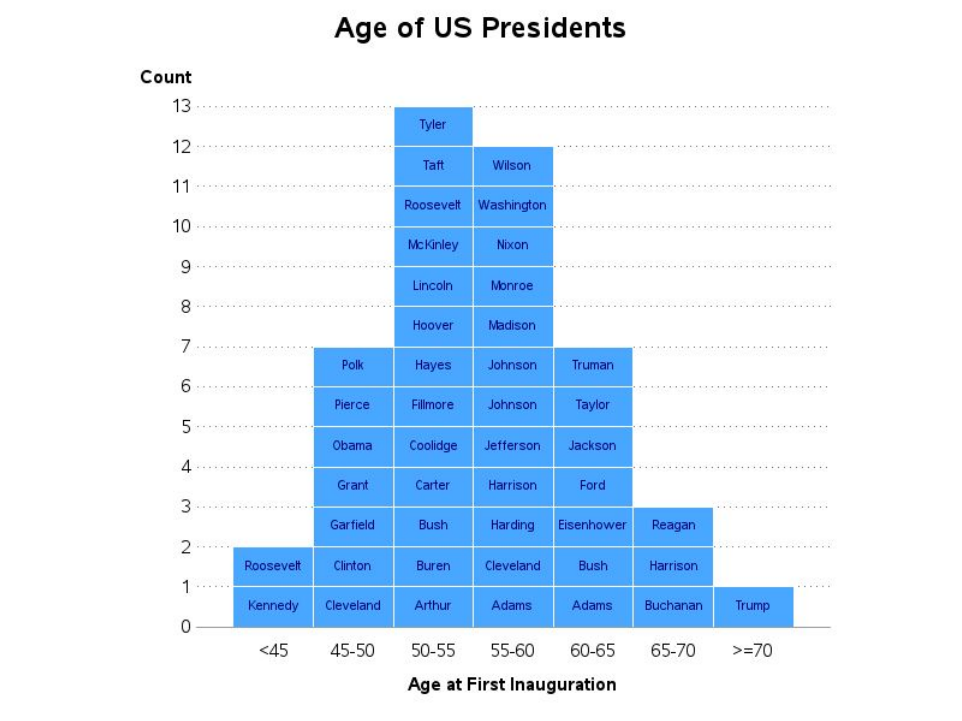
- Discrete - integers



Types of data

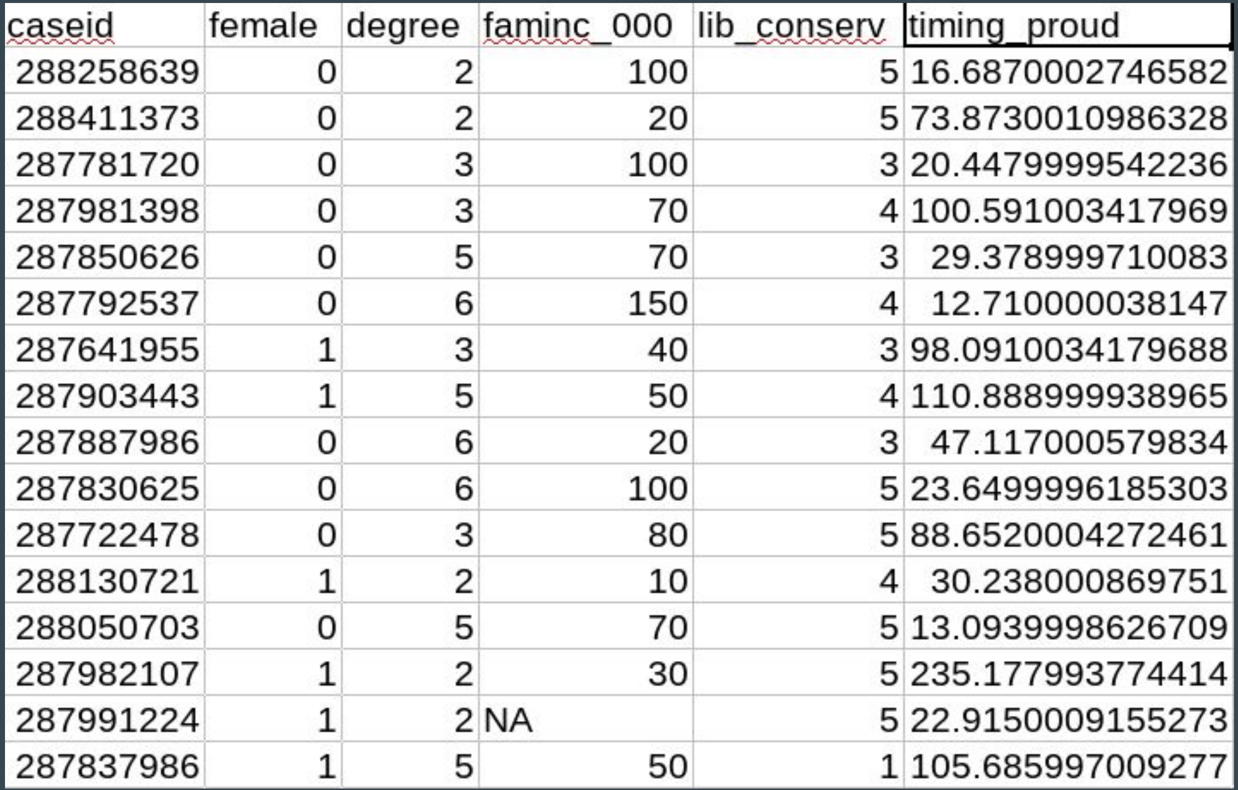
- Categorical - integers representing *types*
  - Ordinal (order matters)
  - Nominal (order does not matter)





What types of data are these?

- Ordinal
- Nominal
- Continuous

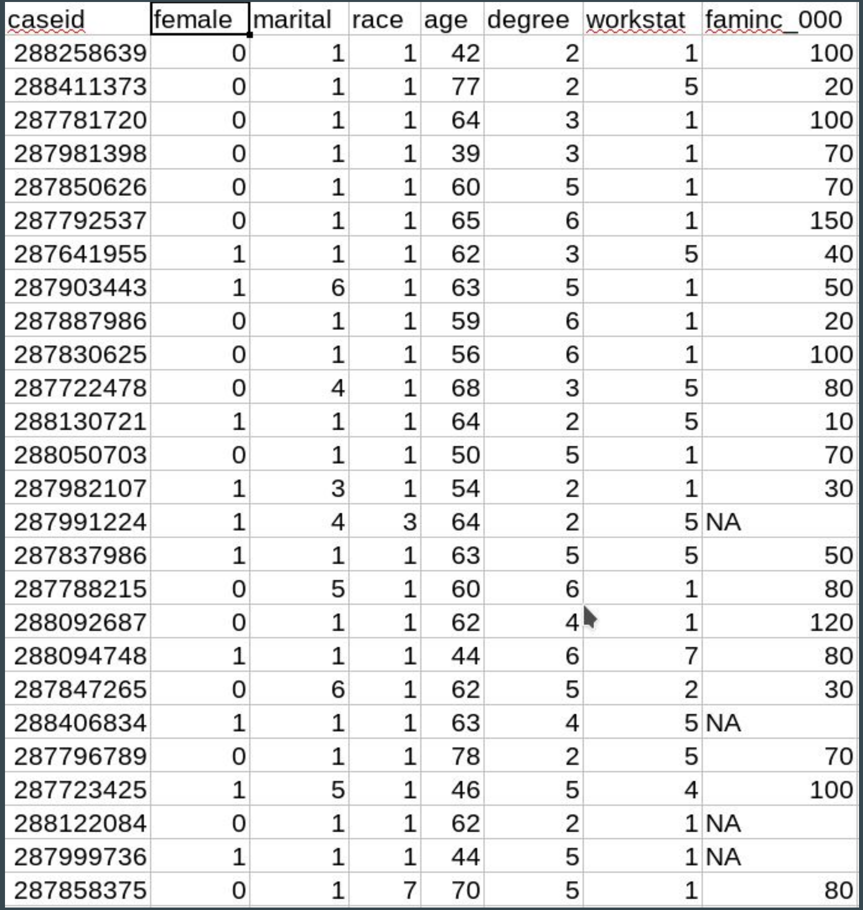


caseid	female	degree	faminc_000	lib_conserv	timing_proud
288258639	0	2	100	5	16.6870002746582
288411373	0	2	20	5	73.8730010986328
287781720	0	3	100	3	20.4479999542236
287981398	0	3	70	4	100.591003417969
287850626	0	5	70	3	29.378999710083
287792537	0	6	150	4	12.710000038147
287641955	1	3	40	3	98.0910034179688
287903443	1	5	50	4	110.888999938965
287887986	0	6	20	3	47.117000579834
287830625	0	6	100	5	23.6499996185303
287722478	0	3	80	5	88.6520004272461
288130721	1	2	10	4	30.238000869751
288050703	0	5	70	5	13.0939998626709
287982107	1	2	30	5	235.177993774414
287991224	1	2	NA	5	22.9150009155273
287837986	1	5	50	1	105.685997009277

Tabular representations

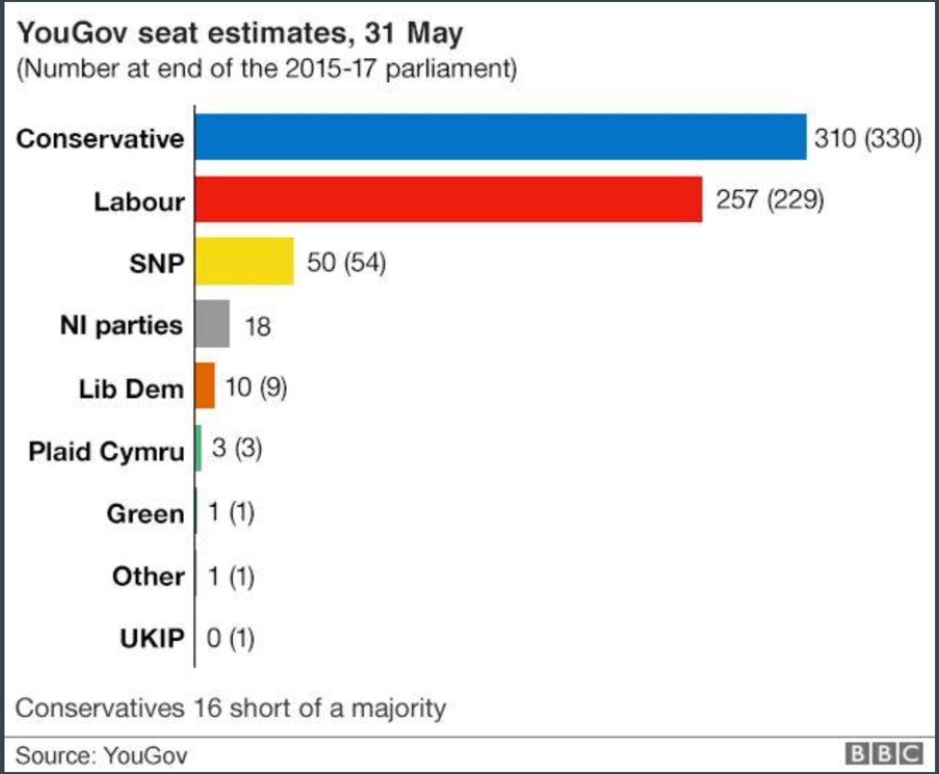
Variables (columns)

Observations (rows)



caseid	female	marital	race	age	degree	workstat	faminc_000
288258639	0	1	1	42	2	1	100
288411373	0	1	1	77	2	5	20
287781720	0	1	1	64	3	1	100
287981398	0	1	1	39	3	1	70
287850626	0	1	1	60	5	1	70
287792537	0	1	1	65	6	1	150
287641955	1	1	1	62	3	5	40
287903443	1	6	1	63	5	1	50
287887986	0	1	1	59	6	1	20
287830625	0	1	1	56	6	1	100
287722478	0	4	1	68	3	5	80
288130721	1	1	1	64	2	5	10
288050703	0	1	1	50	5	1	70
287982107	1	3	1	54	2	1	30
287991224	1	4	3	64	2	5	NA
287837986	1	1	1	63	5	5	50
287788215	0	5	1	60	6	1	80
288092687	0	1	1	62	4	1	120
288094748	1	1	1	44	6	7	80
287847265	0	6	1	62	5	2	30
288406834	1	1	1	63	4	5	NA
287796789	0	1	1	78	2	5	70
287723425	1	5	1	46	5	4	100
288122084	0	1	1	62	2	1	NA
287999736	1	1	1	44	5	1	NA
287858375	0	1	7	70	5	1	80

What are the observations and variables?



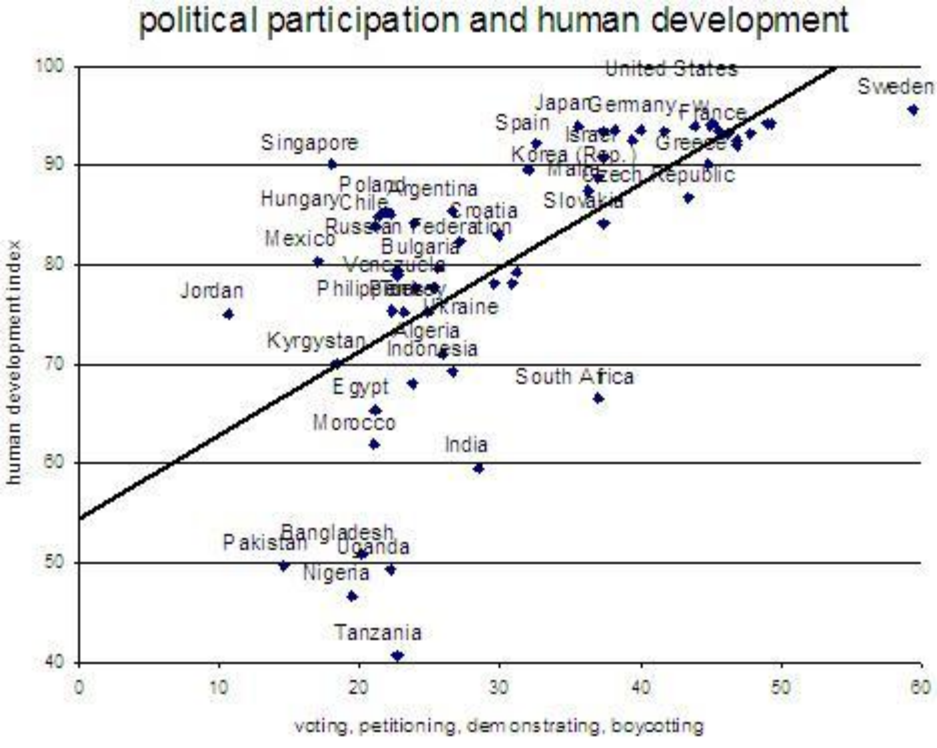
YouGov seat estimates, 31 May

(Number at end of the 2015-17 parliament)

Conservatives 16 short of a majority

Source: YouGov

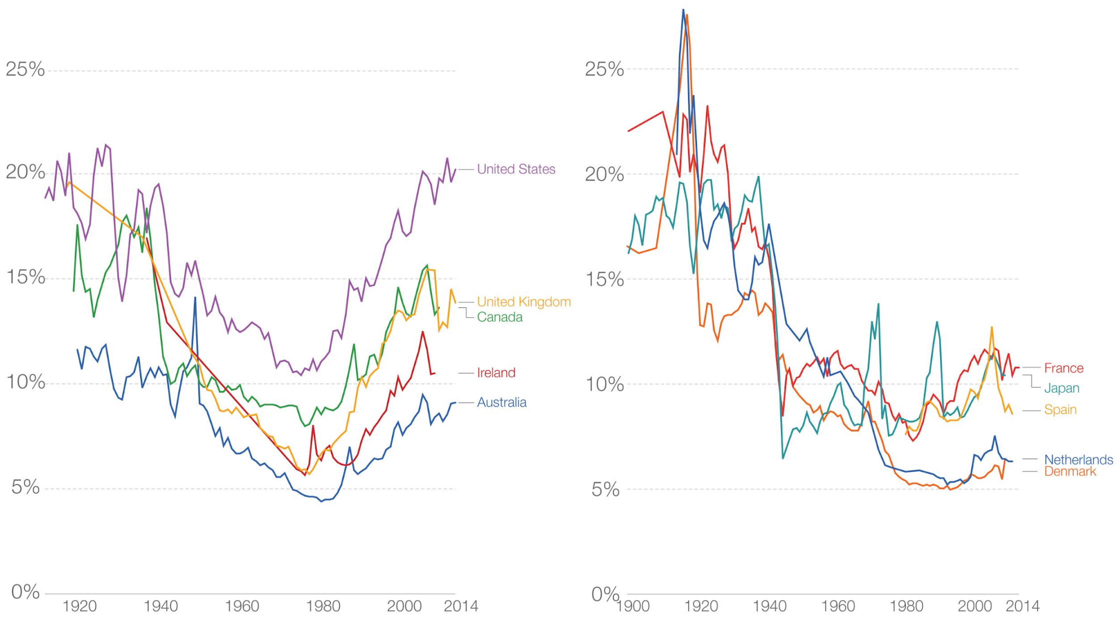
What are the observations and variables?



What are the observations and variables?

Our World in Data

Share of Total Income going to the Top 1% since 1900


The evolution of inequality in English speaking countries followed a U-shape

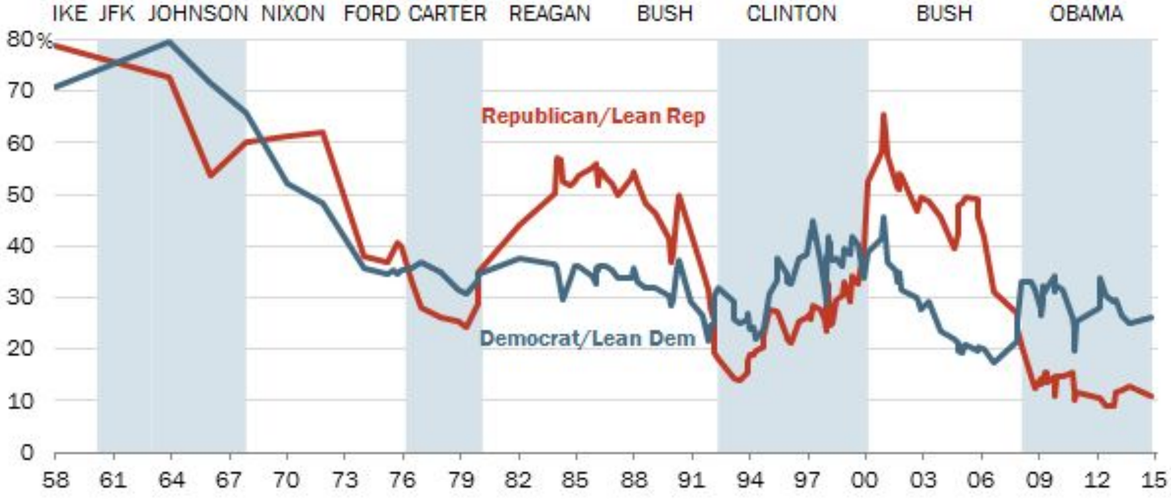
The evolution of inequality in continental Europe and Japan followed an L-shape

Data source: World Wealth and Income Database (2018). This is income before taxes and transfers.
This data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on inequality and how the world is changing. Licensed under CC-BY-SA by the author Max Roser.

What are the observations and variables?

Trust in government by party: 1958-2015

Trust federal government to do what is right just about always/most of the time ...

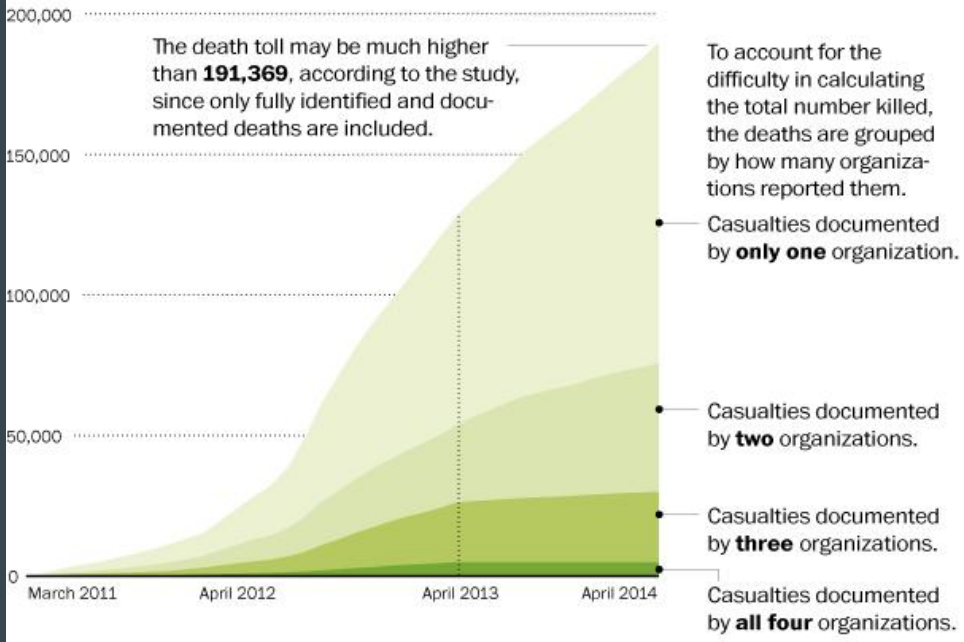


Survey conducted Aug. 27-Oct. 4, 2015. Q15. Trend sources: Pew Research Center, National Election Studies, Gallup, ABC/Washington Post, CBS/New York Times, and CNN Polls. From 1976-2014 the trend line represents a three-survey moving average.

PEW RESEARCH CENTER

What are the observations and variables?

DEATHS OVER TIME, BY NUMBER OF ORGANIZATIONS REPORTING THE DEATHS



The death toll may be much higher than **191,369**, according to the study, since only fully identified and documented deaths are included.

To account for the difficulty in calculating the total number killed, the deaths are grouped by how many organizations reported them.

Casualties documented by **only one** organization.

Casualties documented by **two** organizations.

Casualties documented by **three** organizations.

Casualties documented by **all four** organizations.

Government data, which has not been available since March 2012, has been omitted from the chart. Only three sources of data have been available since April 2013. Reports came from the Syrian Center for Statistics and Research, the Syrian Network for Human Rights, the Syrian Observatory for Human Rights (until April 2013) and the Violations Documentation Center.

Types of variables

Depending on the question, we usually divide variables into two types:

- Outcome: a measure of the phenomena you are most interested in
- Features: numeric quantities that *explain the outcome*

Lots of terms for this:

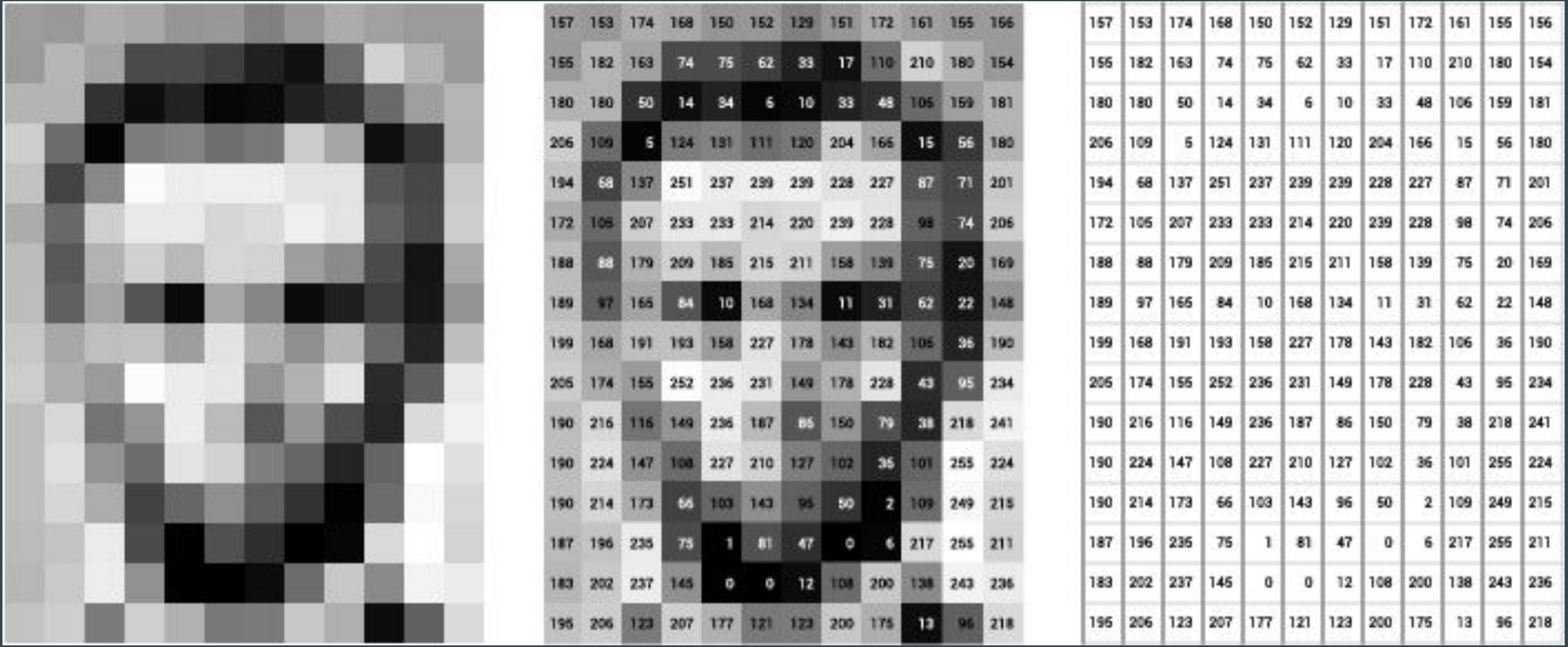
1. Independent and Dependent
2. Explanatory and Outcome
3. Features and Response

But what is data really?

But what is data really?

	But	what	is	data	really	?	observations	variables	Tabular
Slide 3	0	0	0	0	0	0	0	0	1
Slide 4	0	1	0	0	0	0	1	1	0
Slide 5	1	1	1	1	1	1	0	0	0

But what is data really?



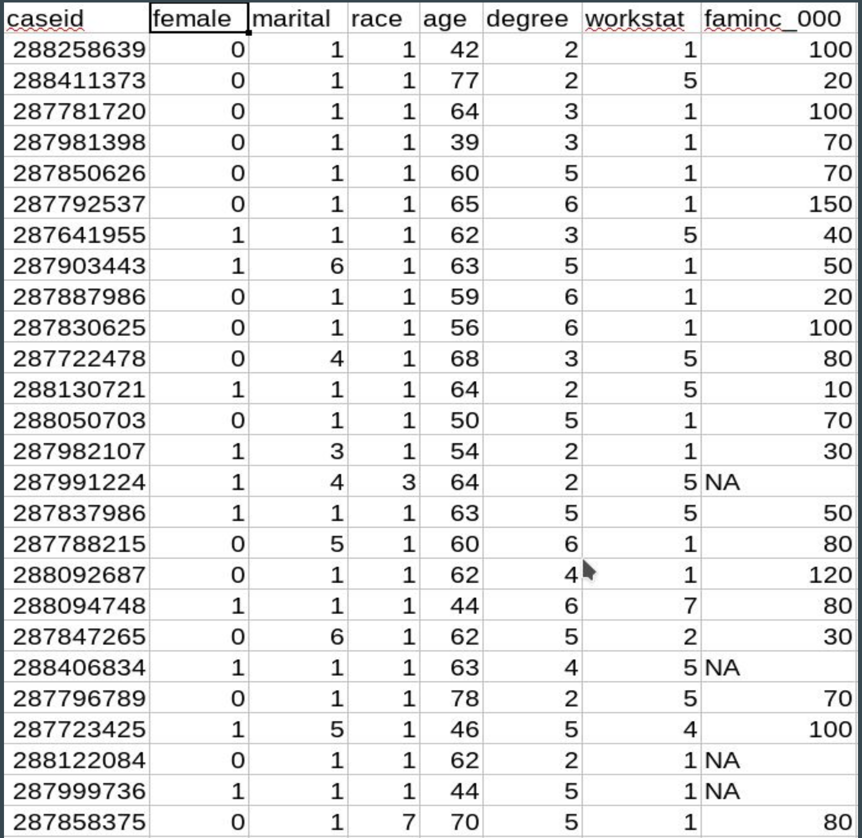
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Back to tabular representations

Variables (columns)

Observations (rows)



caseid	female	marital	race	age	degree	workstat	faminc_000
288258639	0	1	1	42	2	1	100
288411373	0	1	1	77	2	5	20
287781720	0	1	1	64	3	1	100
287981398	0	1	1	39	3	1	70
287850626	0	1	1	60	5	1	70
287792537	0	1	1	65	6	1	150
287641955	1	1	1	62	3	5	40
287903443	1	6	1	63	5	1	50
287887986	0	1	1	59	6	1	20
287830625	0	1	1	56	6	1	100
287722478	0	4	1	68	3	5	80
288130721	1	1	1	64	2	5	10
288050703	0	1	1	50	5	1	70
287982107	1	3	1	54	2	1	30
287991224	1	4	3	64	2	5	NA
287837986	1	1	1	63	5	5	50
287788215	0	5	1	60	6	1	80
288092687	0	1	1	62	4	1	120
288094748	1	1	1	44	6	7	80
287847265	0	6	1	62	5	2	30
288406834	1	1	1	63	4	5	NA
287796789	0	1	1	78	2	5	70
287723425	1	5	1	46	5	4	100
288122084	0	1	1	62	2	1	NA
287999736	1	1	1	44	5	1	NA
287858375	0	1	7	70	5	1	80

Tabular → Matrix

caseid	age	pid3	state
00001	26	1	43
00002	45	2	38
00003	62	1	14

$$\begin{pmatrix} 26 & 1 & 43 \\ 45 & 2 & 38 \\ 62 & 1 & 14 \end{pmatrix}$$

Tabular → Matrix

id	X1	X2	X3
N1	-	-	-
N2	-	-	-
N3	-	-	-

$$\begin{pmatrix} N1X1 & N1X2 & N1X3 \\ N2X1 & N2X2 & N2X3 \\ N3X1 & N3X2 & N3X3 \end{pmatrix}$$

Matrices

- A matrix is a rectangular array of numbers
  - Made up of rows (n) and columns (k)
  - Each number is an *element*, with a unique value (n, k)
- In data analysis
  - Rows = observations
  - Columns = variables
- A single observation or row (i.e. n = 3 , k=k) is a vector
  - Each vector has a *magnitude* and *direction*

Matrices

“If human beings could see in multiple dimensions, we wouldn’t need data analysis.”

-- Pedro Domingos, *University of Washington*

Sample vs. Population

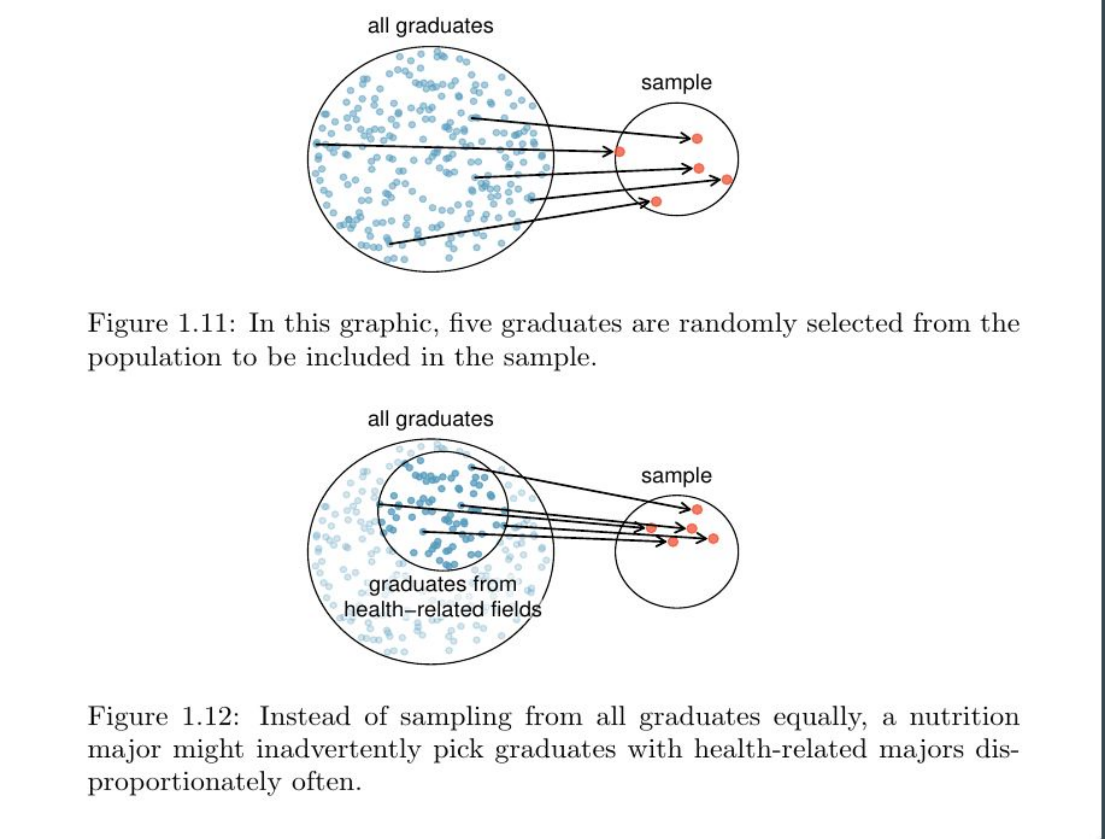

all graduates
sample

Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

all graduates
sample
graduates from health–related fields

Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

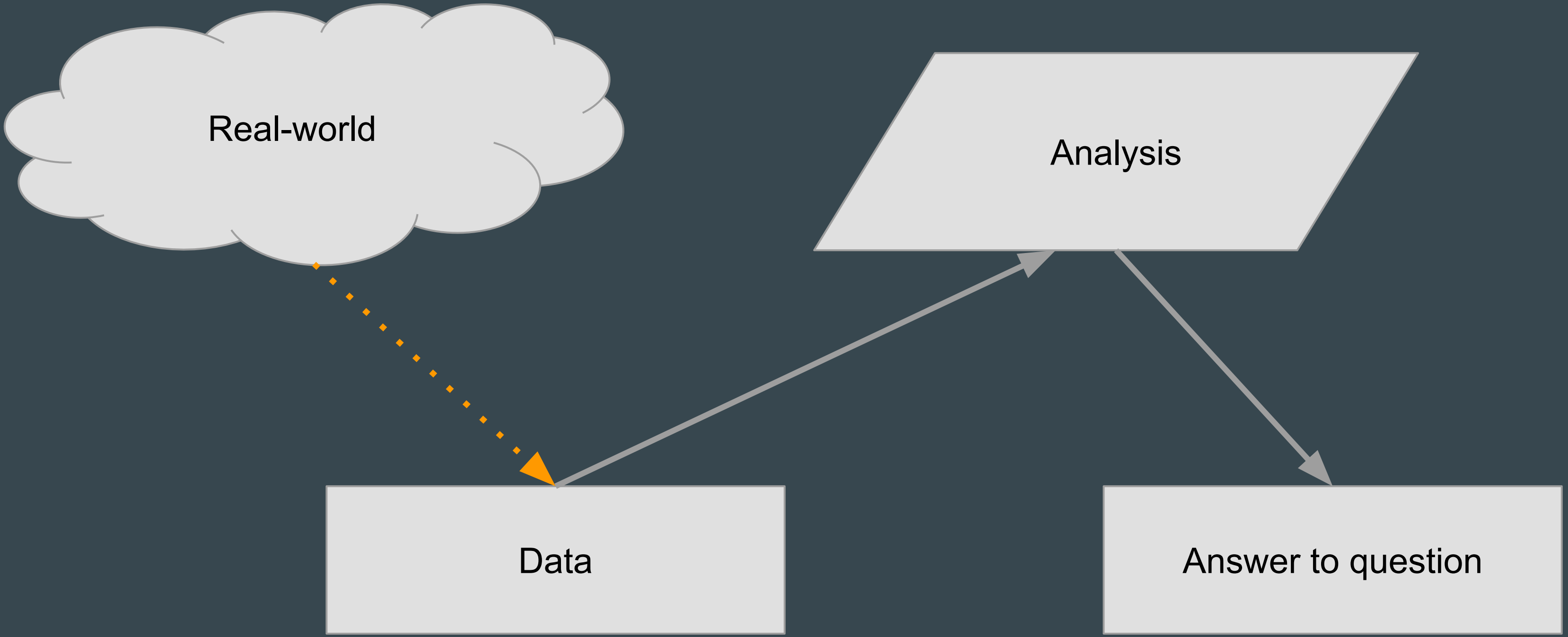
A population is the group you'd like to understand.

A sample is a segment of that group.

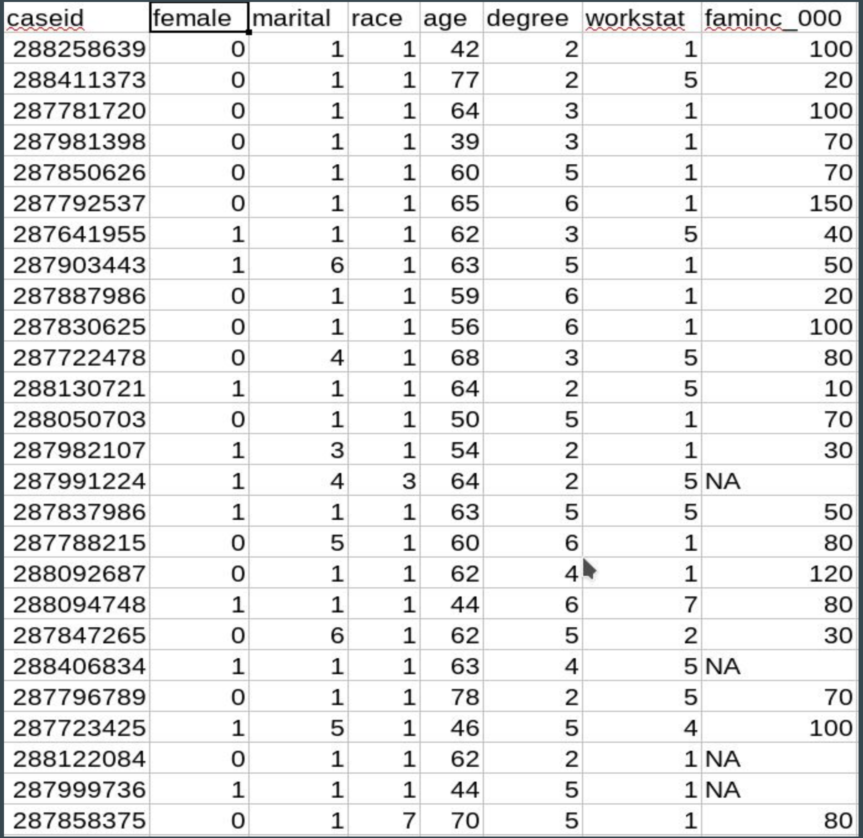
Data-generating Process (DGP)

- Two types of data collection:
    a. Experimental (artificial - controlled by the researcher)
    b. Observational (collected after the fact)
- Where does your data come from?
- How is it produced?
- A map from real-world phenomena to numbers on a spreadsheet

Data-generating Process (DGP)



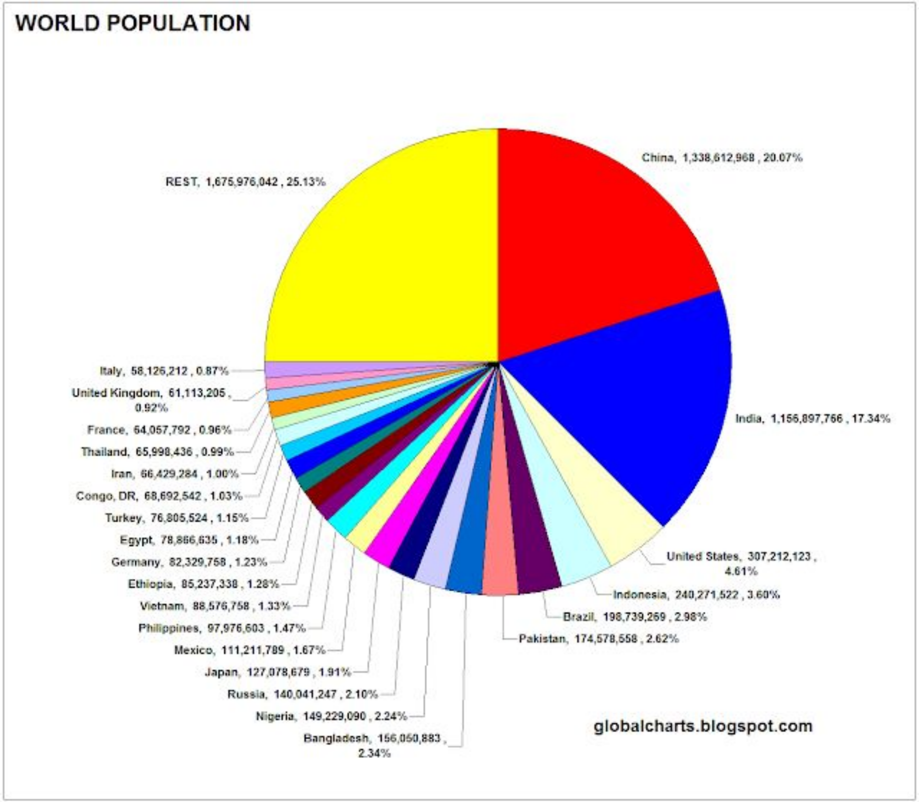
DGP Example



caseid	female	marital	race	age	degree	workstat	faminc_000
288258639	0	1	1	42	2	1	100
288411373	0	1	1	77	2	5	20
287781720	0	1	1	64	3	1	100
287981398	0	1	1	39	3	1	70
287850626	0	1	1	60	5	1	70
287792537	0	1	1	65	6	1	150
287641955	1	1	1	62	3	5	40
287903443	1	6	1	63	5	1	50
287887986	0	1	1	59	6	1	20
287830625	0	1	1	56	6	1	100
287722478	0	4	1	68	3	5	80
288130721	1	1	1	64	2	5	10
288050703	0	1	1	50	5	1	70
287982107	1	3	1	54	2	1	30
287991224	1	4	3	64	2	5	NA
287837986	1	1	1	63	5	5	50
287788215	0	5	1	60	6	1	80
288092687	0	1	1	62	4	1	120
288094748	1	1	1	44	6	7	80
287847265	0	6	1	62	5	2	30
288406834	1	1	1	63	4	5	NA
287796789	0	1	1	78	2	5	70
287723425	1	5	1	46	5	4	100
288122084	0	1	1	62	2	1	NA
287999736	1	1	1	44	5	1	NA
287858375	0	1	7	70	5	1	80

1. YouGov creates survey pool by collecting volunteers
2. They sample volunteers according to U.S. census
3. Respondents fill out form online

DGP Example



1. Countries distribute surveys to individual respondents
2. Respondents fill out surveys
3. Statistical offices for specific countries tabulate information and submit to UN

DGP Example

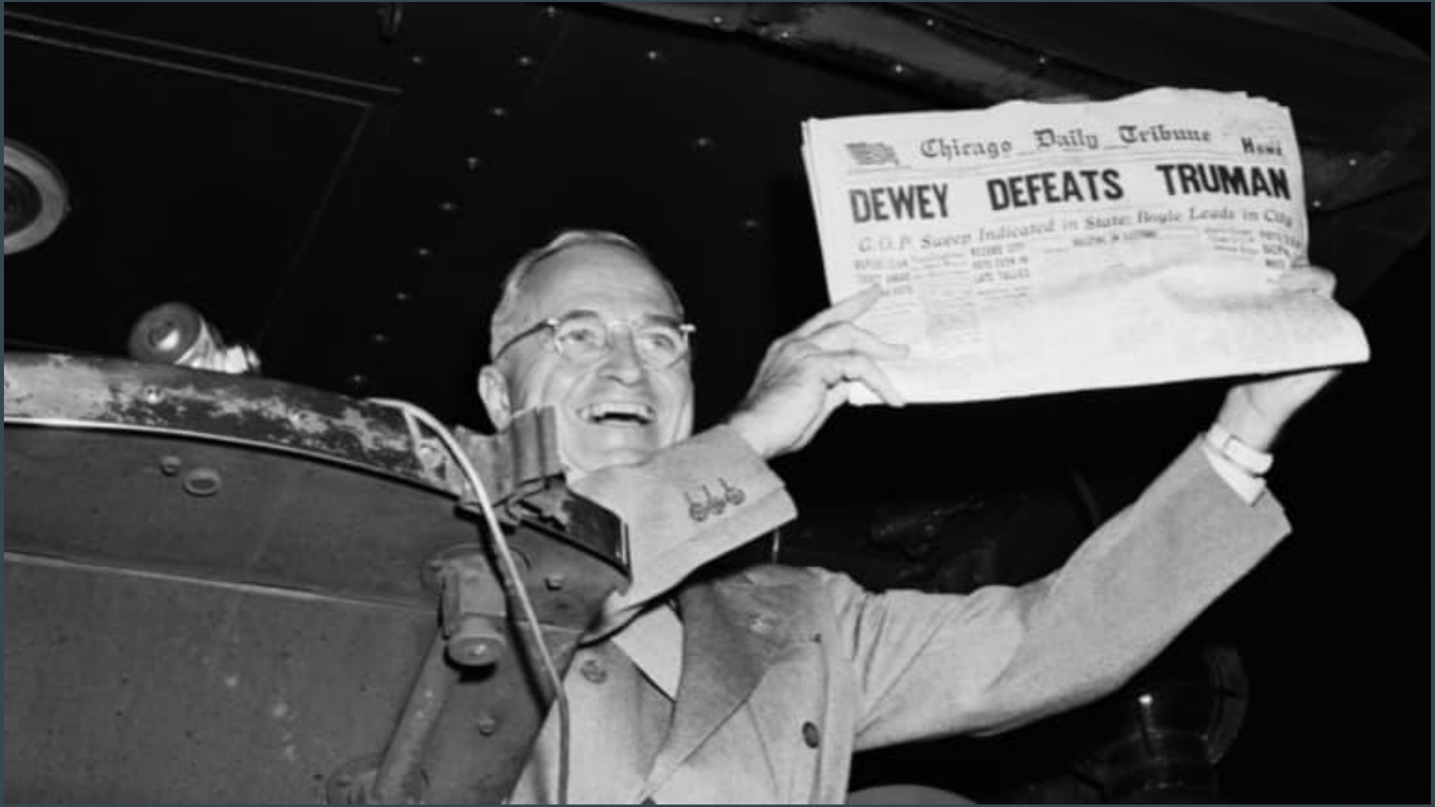
9031
901
9023
TERRITOIRE DE KAZUMBA
5027
TERRITOIRE BASOKO

When we get the DGP wrong

“On two occasions I have been asked, ‘Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?’ … I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.”

-- Charles Babbage

When we get the DGP wrong



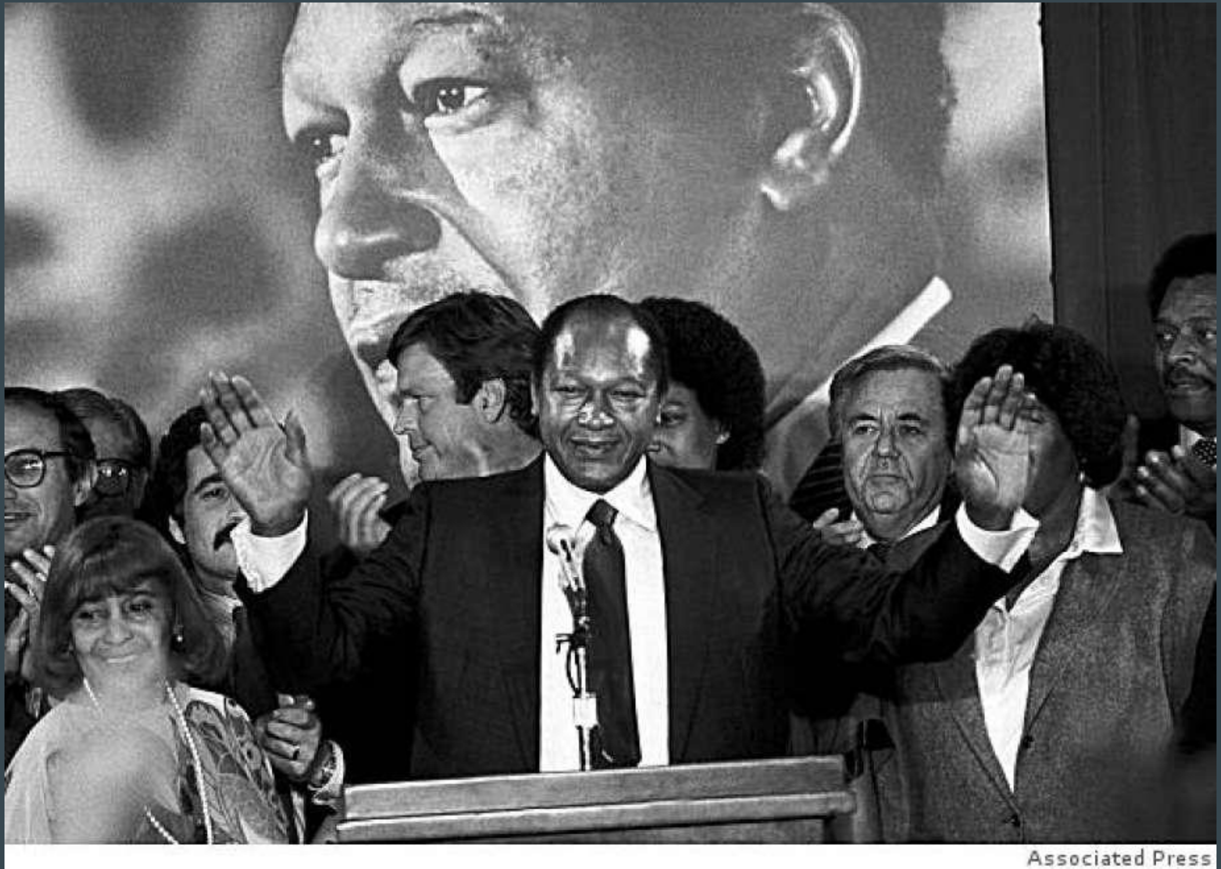
Problems with sampling

DGP used for Gallup poll predicting Dewey victory:

1. Divide US Census into discrete categories (i.e. urban white women, rural African-American men, etc.)
2. Each interviewer is assigned to collect interviews from each category
3. Size of categories are the same ratio as U.S. population

Non-random sampling

When we get the DGP wrong



Associated Press

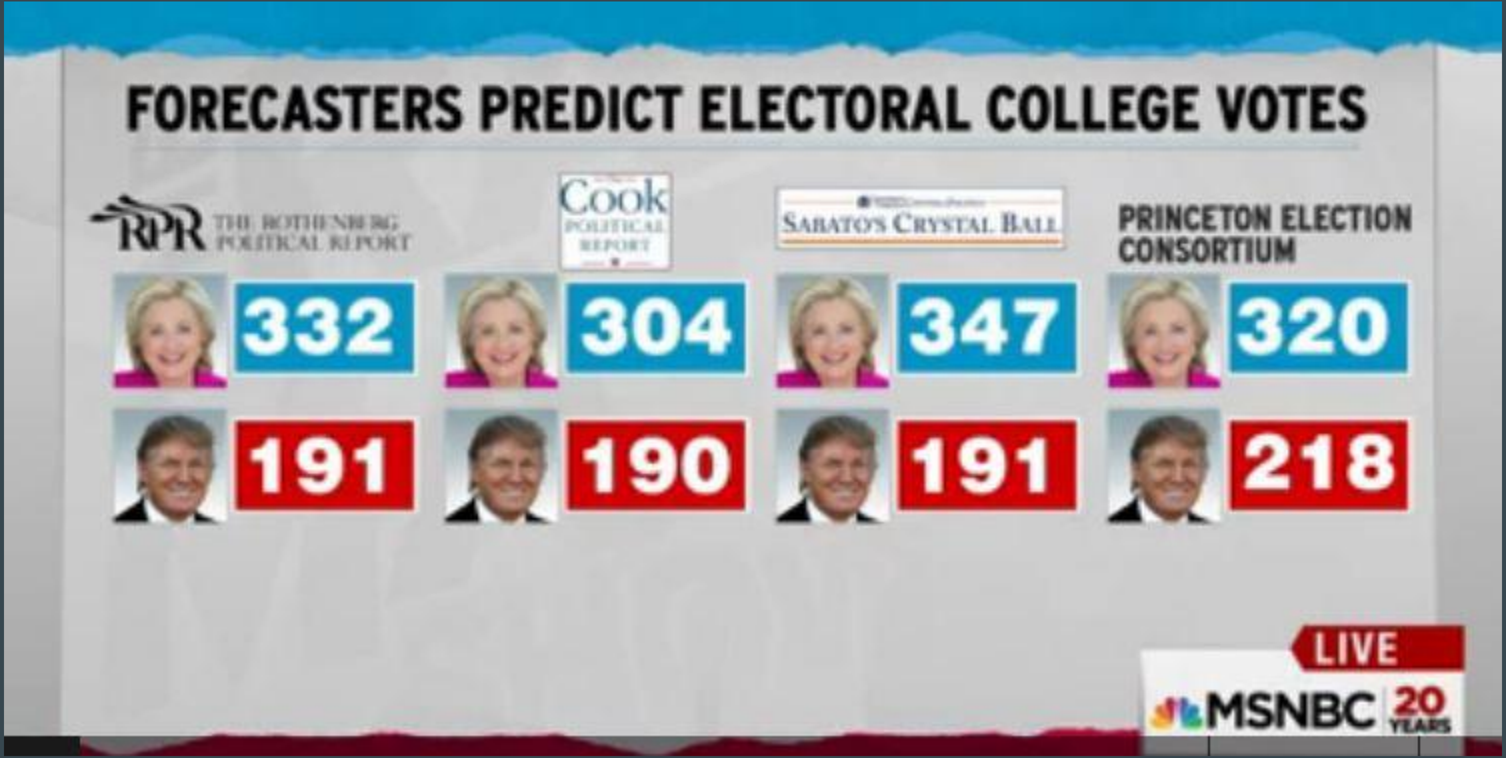
Problems with measurement

DGP for polls predicting Bradley victory:

1. Selects individuals for survey from a *random sample* of voting age population
2. Individuals respond to the in-person interviewer with their vote preference

Social Desirability Bias

When we get the DGP wrong



FORECASTERS PREDICT ELECTORAL COLLEGE VOTES

RPR THE ROTHENBERG POLITICAL REPORT	Cook POLITICAL REPORT	SABATO'S CRYSTAL BALL	PRINCETON ELECTION CONSORTIUM
332	304	347	320
191	190	191	218

DGP for 2016 Election Forecasts

1. Independent surveys are conducted:
   a. Firms determine ‘likely voters’
   b. A random sample of likely voters is drawn
   c. Interviewers attempt to contact those voters
   d. Voters who consent to interview have their preferences recorded
2. Forecasters average various independent surveys for each state
   a. Rank the quality of the source
   b. Assume that errors among pollsters are random

DGP for 2016 Election Forecasts

1. Independent surveys are conducted:
   a. Firms determine ‘likely voters’
   b. A random sample of likely voters is drawn
   c. Interviewers attempt to contact those voters
   d. Voters who consent to interview have their preferences recorded
2. Forecasters average various independent surveys for each state
   a. Rank the quality of the source
   b. Assume that errors among pollsters are random

No such thing as bad data, just bad assumptions

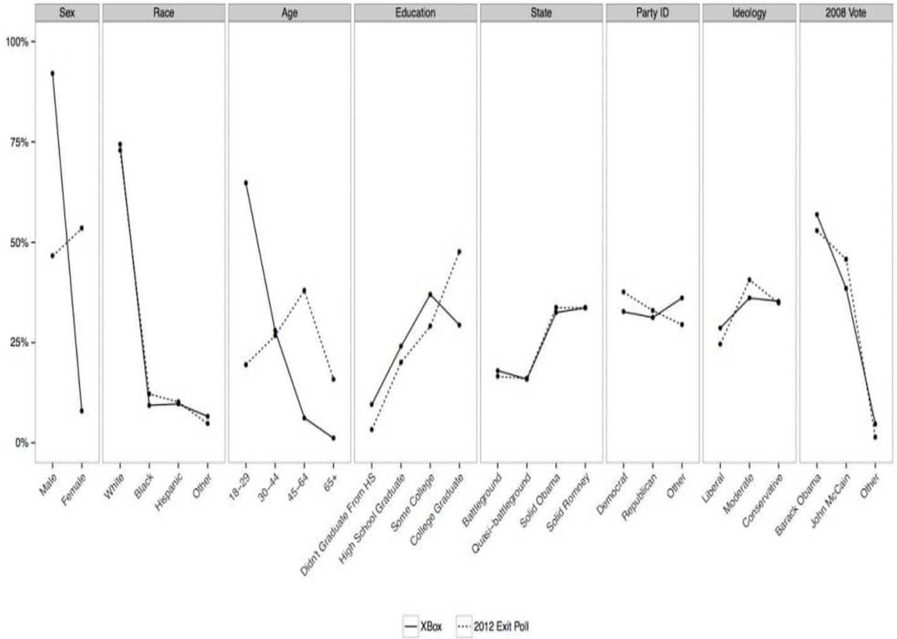


Figure 1: A comparison of the demographic, partisan, and 2008 vote distribution in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). The sex and age distributions, as one might expect, exhibit considerable differences.

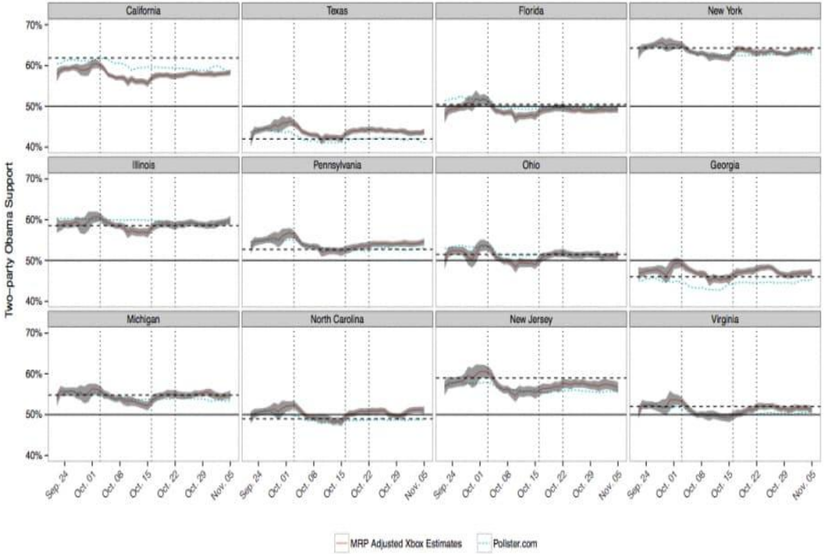
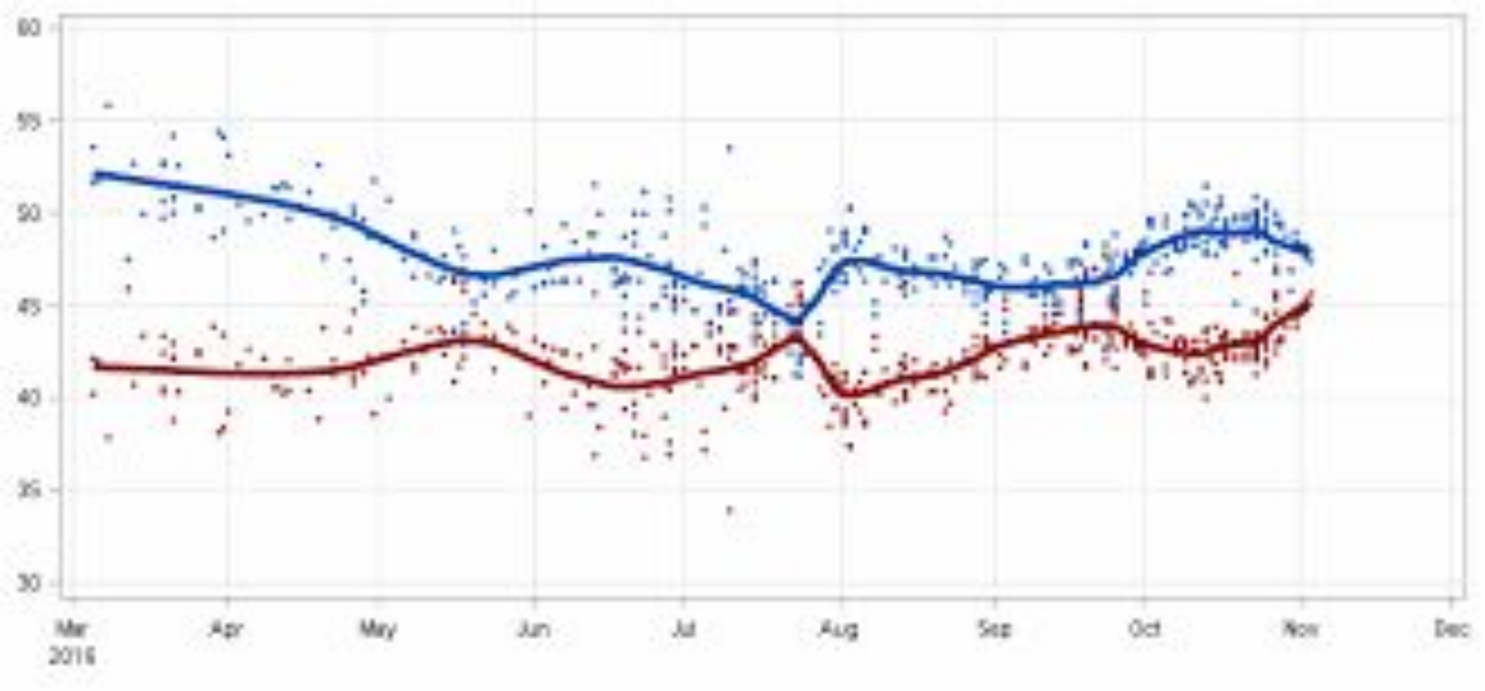
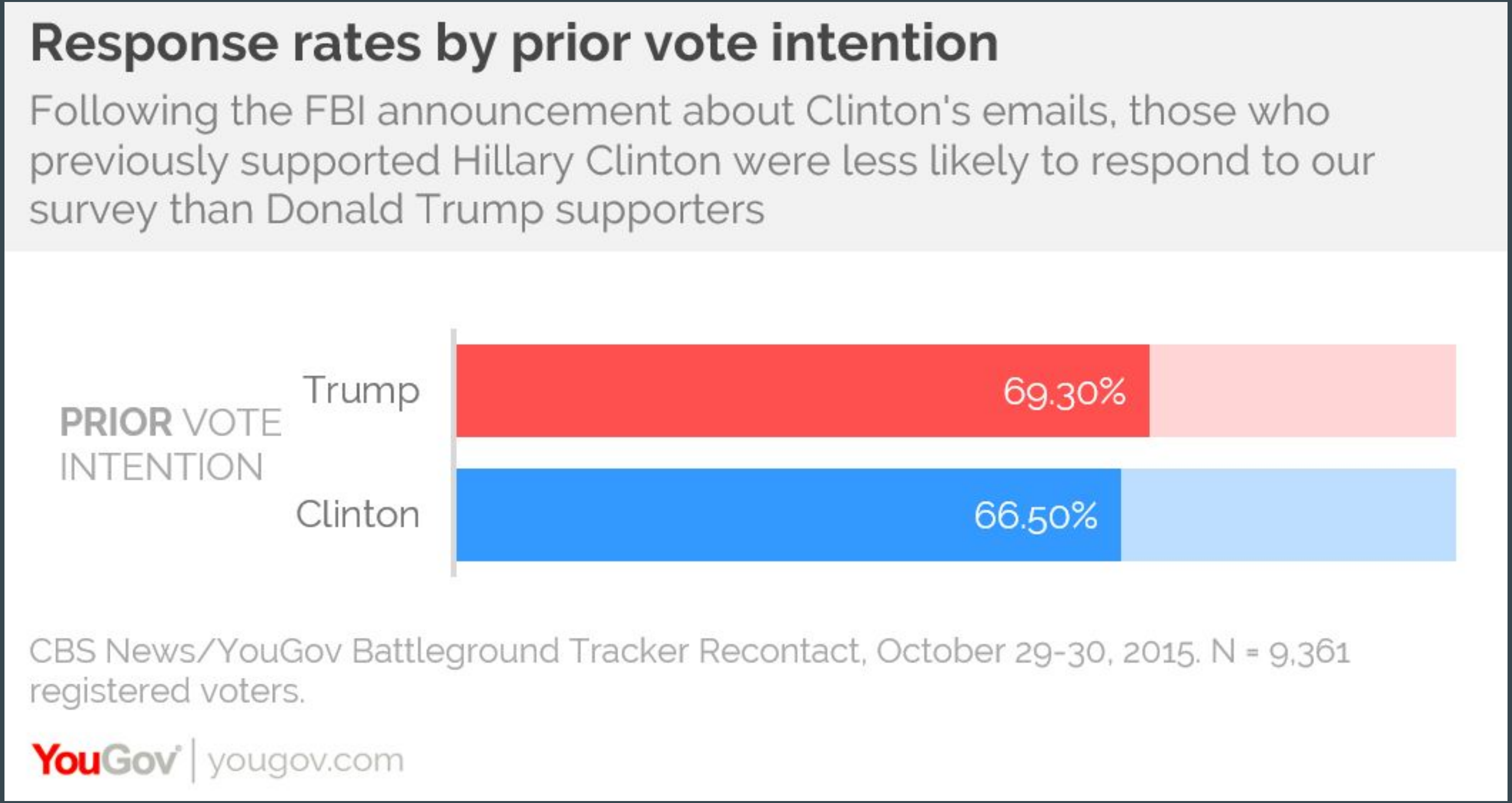


Figure 4: MRP-adjusted daily voter intent for the 12 states with the most electoral votes, and the associated 95% confidence bands. The horizontal dashed lines in each panel give the actual two-party Obama vote shares in that state. The mean and median absolute errors of the last day voter intent across the 51 Electoral College races are 2.5 and 1.8 percentage points, respectively. The state-by-state daily aggregated polling results from Pollster.com, given in the dotted blue lines, are broadly consistent with the estimates from the Xbox data.

The myth of “swing voters”



The myth of “swing voters”



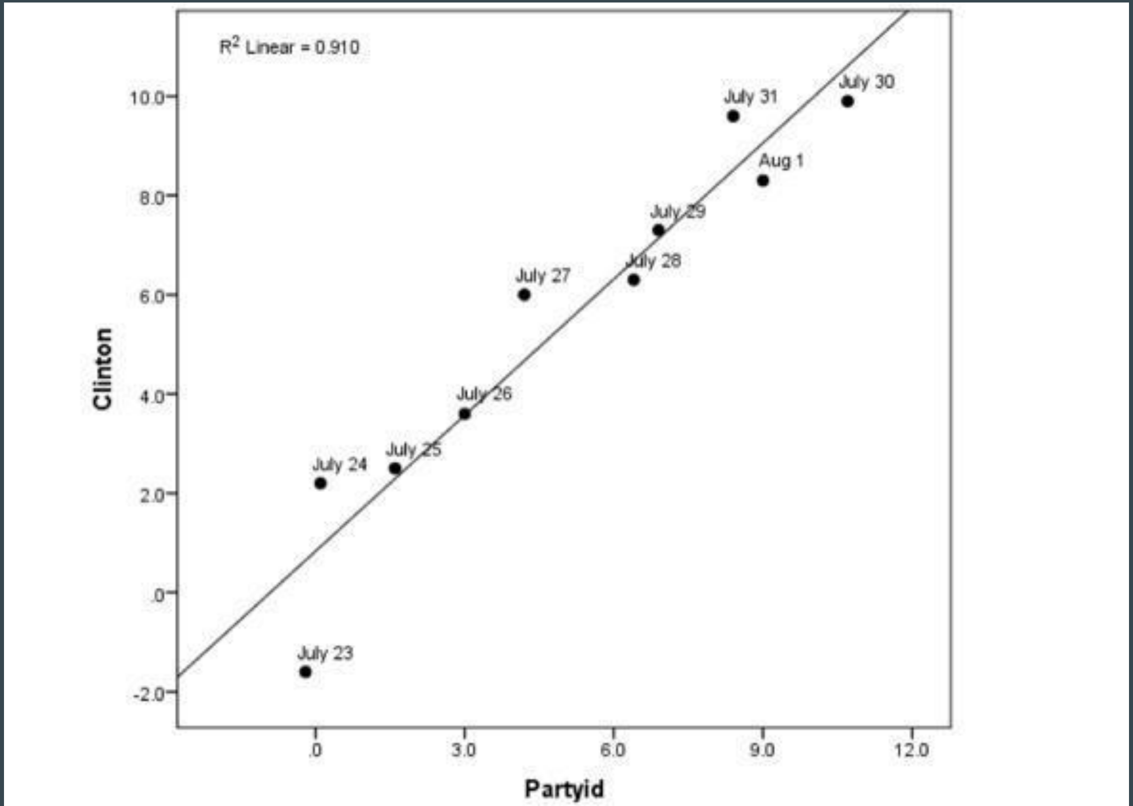
Response rates by prior vote intention

Following the FBI announcement about Clinton's emails, those who previously supported Hillary Clinton were less likely to respond to our survey than Donald Trump supporters

CBS News/YouGov Battleground Tracker Recontact, October 29-30, 2015. N = 9,361 registered voters.

YouGov | yougov.com

The myth of “swing voters”



Review

Types of data:

1. Continuous
2. Discrete
3. Categorical
   a. Nominal
   b. Ordinal
   c. Binary/Dummy

Review

How is data represented?

1. Tabular
   a. Rows = observations
   b. Columns = variables
   c. Labelled, computer display
2. Matrix
   a. Rows = observations
   b. Columns = variables
   c. Unlabelled, computer operations
   d. Defines a geometric space

Review

Data-generating Process (DGP)

1. From real-world phenomena to numbers
2. Step-by-step recipe for data collection
3. Two types of data collection
   a. Experimental - *controlled by researcher*
   b. Observation - *collected by researched after the fact*

Review

Miscellaneous vocabulary:

1. Outcome variable(s):
   - what measures the primary outcome of interest
2. Explanatory variables:
   - what explains the primary outcome of interest
3. Population:
   - the group you which to answer questions about
4. Sample:
   - a subset of the population that you use in your data analysis