SPIA 586a: Machine Learning for Policy Analysis
Week 8: Designed Data

Brandon Stewart[1]

Princeton

March 19, 2024

[1]Huge thanks to Ted Enamerado, Erin Hartman and Matt Salganik for slides.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - spring break
  - (before that, experiments)

Where We've Been and Where We're Going...

- Last Week
  - spring break
  - (before that, experiments)
- This Week
  - designed data

Where We've Been and Where We're Going...

- Last Week
  - spring break
  - (before that, experiments)
- This Week
  - designed data
- Next Week
  - found data

Where We've Been and Where We're Going...

- Last Week
  - spring break
  - (before that, experiments)
- This Week
  - designed data
- Next Week
  - found data
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Admin Update

- Major Assignment 1 Redux
- Last Two Proposal Weeks
- Minor Assignments
- Major Assignments 2–3

Three Simplified Data Source Archetypes

Three Simplified Data Source Archetypes

- Experimental Data
  we intervene, ask, and analyze

Three Simplified Data Source Archetypes

- Experimental Data
  we intervene, ask, and analyze
- Designed Data
  we ask and analyze

Three Simplified Data Source Archetypes

- Experimental Data
  we intervene, ask, and analyze
- Designed Data
  we ask and analyze
- Found Data
  we analyze

1. Surveys
2. Total Survey Error Framework
3. Integrating Surveys with Other Data
4. Big Data Paradox
5. Writing Survey Questions is Difficult
6. Record Linkage

Surveys as Designed Data

Surveys as Designed Data

- This week we will focus in on surveys as a particular case of designed data.

Surveys as Designed Data

- This week we will focus in on surveys as a particular case of designed data.
- Other types of data share these same properties but surveys are particularly clear about the target population and the sample.

Surveys as Designed Data

- This week we will focus in on surveys as a particular case of designed data.
- Other types of data share these same properties but surveys are particularly clear about the target population and the sample.
- Key feature: a researcher controls how something is measured.

Surveys as Designed Data

- This week we will focus in on surveys as a particular case of designed data.
- Other types of data share these same properties but surveys are particularly clear about the target population and the sample.
- Key feature: a researcher controls how something is measured.
- Next week we will consider cases where we don't get to control what or how something is measured (found data).

Surveys as Designed Data

- This week we will focus in on surveys as a particular case of designed data.
- Other types of data share these same properties but surveys are particularly clear about the target population and the sample.
- Key feature: a researcher controls how something is measured.
- Next week we will consider cases where we don't get to control what or how something is measured (found data).
- The line is blurry: if someone else runs a survey but then you analyze it, is the data found or designed?

Case Study

https://commons.wikimedia.org/wiki/File:Donald_Trump_taking_his_Oath_of_Office.png

Case Study

An Evaluation of 2016 Election Polls in the U.S.

Ad Hoc Committee on 2016 Election Polling

Courtney Kennedy, Pew Research Center
Mark Blumenthal, SurveyMonkey
Scott Clement, Washington Post
JoshUA d. Clinton, Vanderbilt University
Claire Durand, University of Montreal
Charles Franklin, Marquette University
Kyley McGeeney, Pew Research Center[1]

Lee Miringoff, Marist College
Kristen Olson, University of Nebraska-Lincoln
Doug Rivers, Stanford University, YouGov
Lydia Saad, Gallup
Evans Witt, Princeton Survey Research Associates
Chris Wlezien, University of Texas at Austin

http://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

Case Study

- National polls were generally correct and accurate by historical standards.

Case Study

- National polls were generally correct and accurate by historical standards.
- State-level polls showed a competitive, uncertain contest . . .

Case Study

- National polls were generally correct and accurate by historical standards.
- State-level polls showed a competitive, uncertain contest . . .
- . . . but clearly under-estimated Trump's support in the Upper Midwest.

Case Study

"There are a number of reasons as to why polls under-estimated support for Trump. The explanations for which we found the most evidence are:"

- "Real change in vote preference during the final week or so of the campaign"

Case Study

"There are a number of reasons as to why polls under-estimated support for Trump. The explanations for which we found the most evidence are:"

- "Real change in vote preference during the final week or so of the campaign"
- "Adjusting for over-representation of college graduates was critical, but many polls did not do it"

Case Study

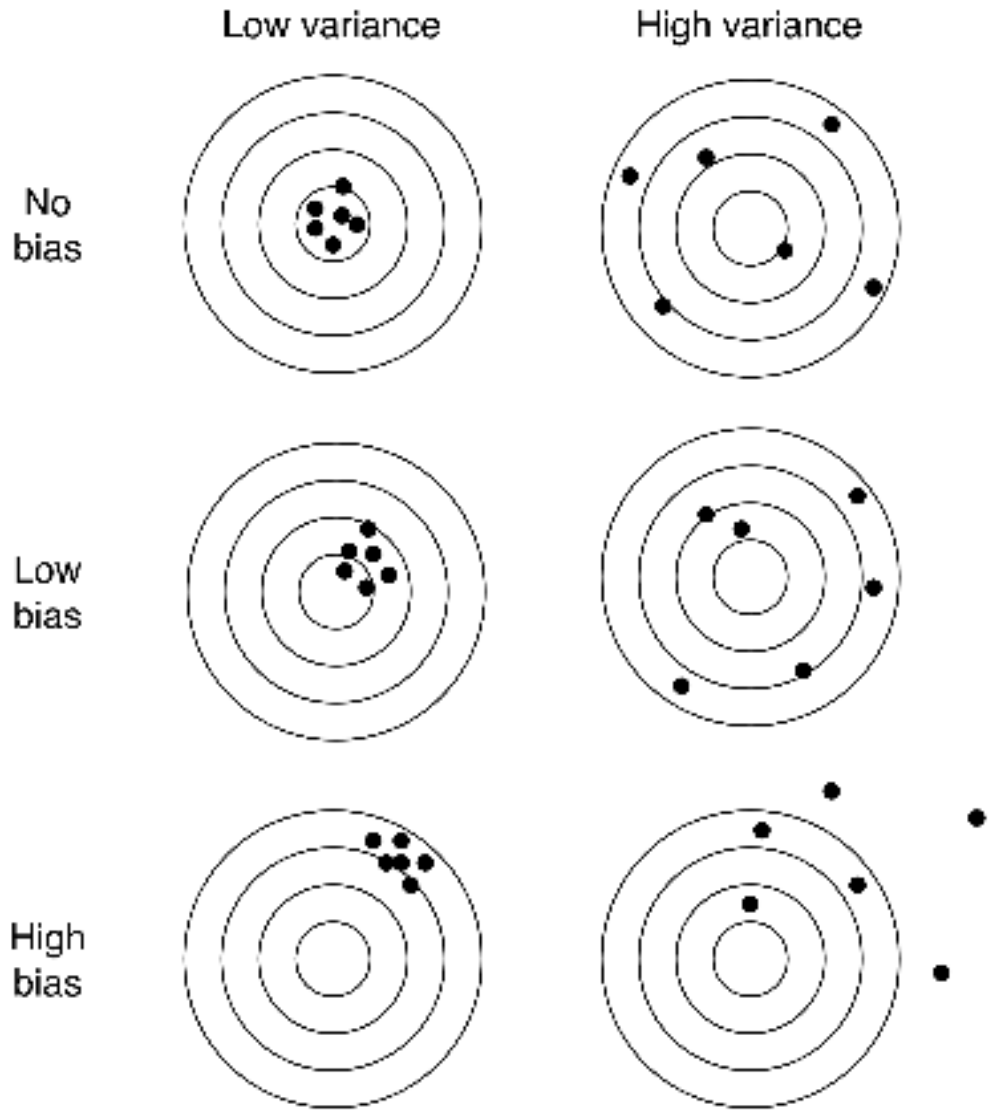
"There are a number of reasons as to why polls under-estimated support for Trump. The explanations for which we found the most evidence are:"

- "Real change in vote preference during the final week or so of the campaign"
- "Adjusting for over-representation of college graduates was critical, but many polls did not do it"
- "Some Trump voters who participated in pre-election polls did not reveal themselves as Trump voters until after the election, and they outnumbered late-revealing Clinton voters"

Full report:
`http://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx`

A Language for Things Going Wrong

A Language for Things Going Wrong



1. Surveys
2. Total Survey Error Framework
3. Integrating Surveys with Other Data
4. Big Data Paradox
5. Writing Survey Questions is Difficult
6. Record Linkage

Total Survey Error

Total Survey Error

- A framework for considering sampling, measurement, and non-sampling errors in surveys

Total Survey Error

- A framework for considering sampling, measurement, and non-sampling errors in surveys
- Has lead to acknowledgement of errors of observation (measurement) and errors of non-observation (coverage, nonresponse, sampling)

Total Survey Error

- A framework for considering sampling, measurement, and non-sampling errors in surveys
- Has lead to acknowledgement of errors of observation (measurement) and errors of non-observation (coverage, nonresponse, sampling)
- Decomposition in to bias and variance

Total Survey Error

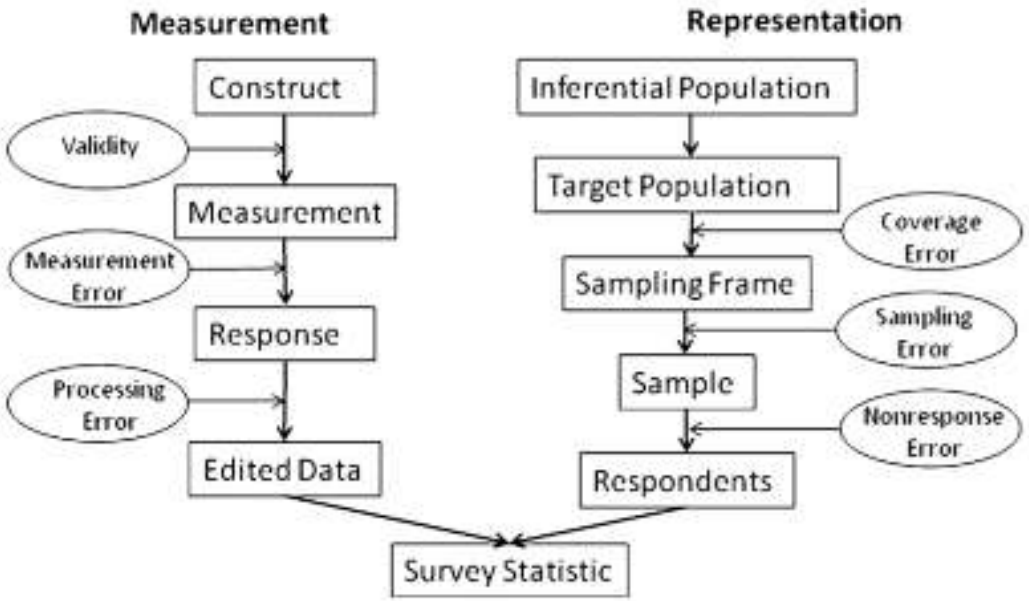
- A framework for considering sampling, measurement, and non-sampling errors in surveys
- Has lead to acknowledgement of errors of observation (measurement) and errors of non-observation (coverage, nonresponse, sampling)
- Decomposition in to bias and variance
- Learn More!
  - Robert M Groves et al. *Survey Methodology*. Vol. 561. John Wiley & Sons, 2011. Chapter 2.
  - Robert M Groves and Lars Lyberg. "Total survey error: Past, present, and future". In: *Public Opinion Quarterly* 74.5 (2010), pp. 849?879 .

Core Idea: Total Survey Error Framework

Total survey error = measurement error + representation error

Core Idea: Total Survey Error Framework

Total survey error = measurement error + representation error



What you learn depends on how you ask and who you ask.

Groves and Lyberg 2010, Fig 3

Survey Lifecycle

Design Perspective

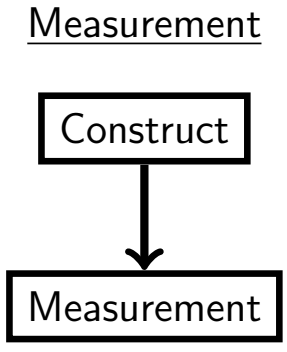
Measurement

Construct

Construct: elements of information that are sought by the researcher

Survey Lifecycle

Design Perspective

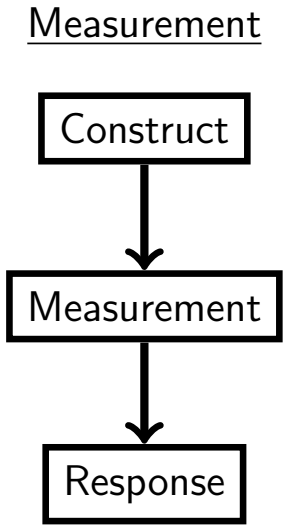


Construct: elements of information that are sought by the researcher

Measurement: ways to gather information about constructs

Survey Lifecycle

Design Perspective



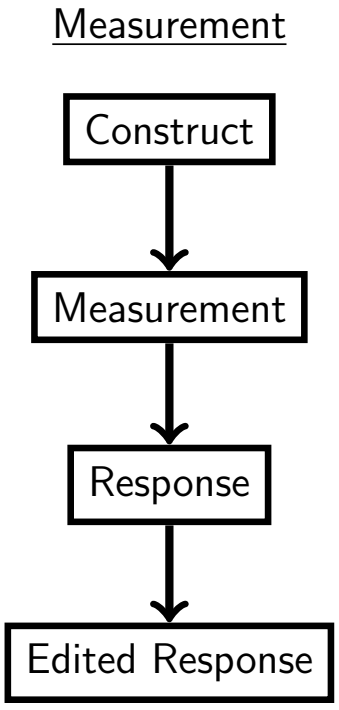
Construct: elements of information that are sought by the researcher

Measurement: ways to gather information about constructs

Response: data produced from information provided through the survey measurements

Survey Lifecycle

Design Perspective



Construct: elements of information that are sought by the researcher

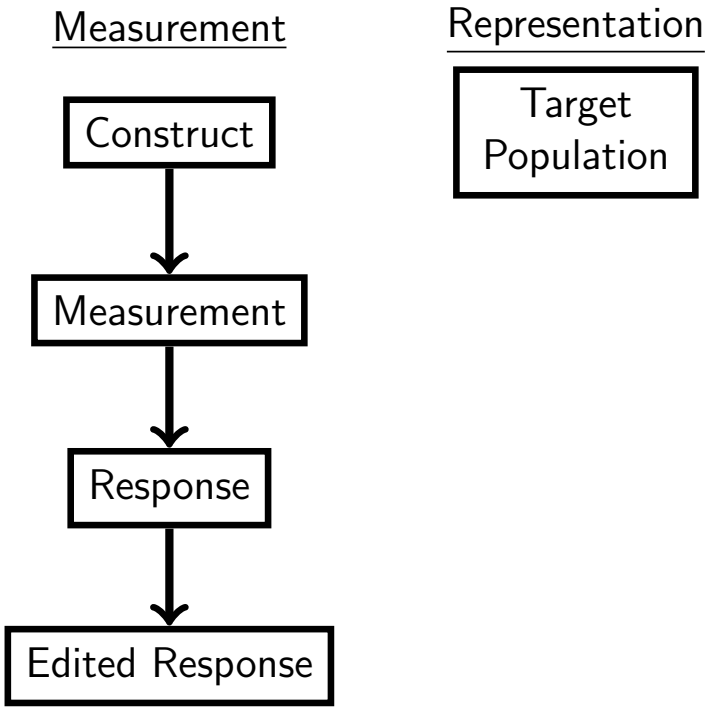
Measurement: ways to gather information about constructs

Response: data produced from information provided through the survey measurements

Edited Response: when the initial measurement provided undergoes a review prior to moving on to the next

Survey Lifecycle

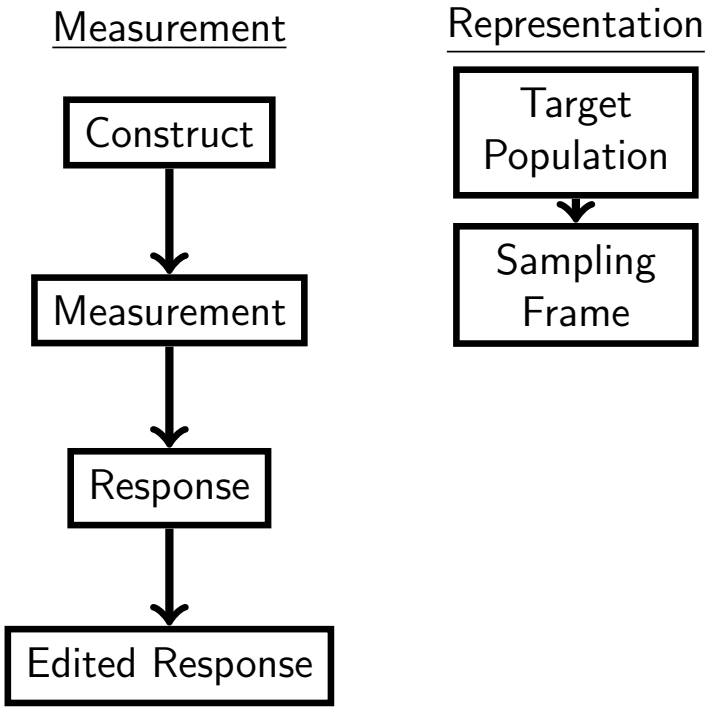
Design Perspective



Target Population: the set of units to be studied

Survey Lifecycle

Design Perspective

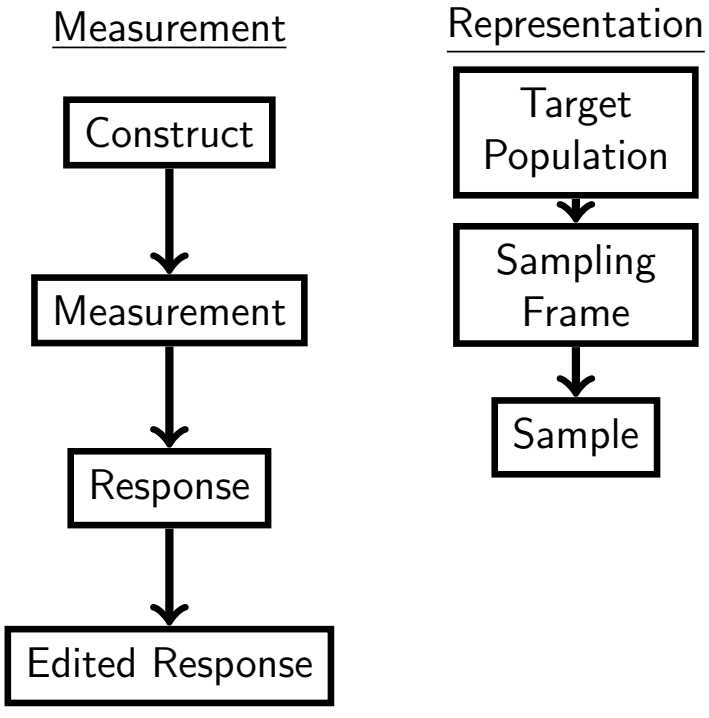


Target Population: the set of units to be studied

Sampling Frame: the set of members that has a chance to be selected into the survey sample

Survey Lifecycle

Design Perspective



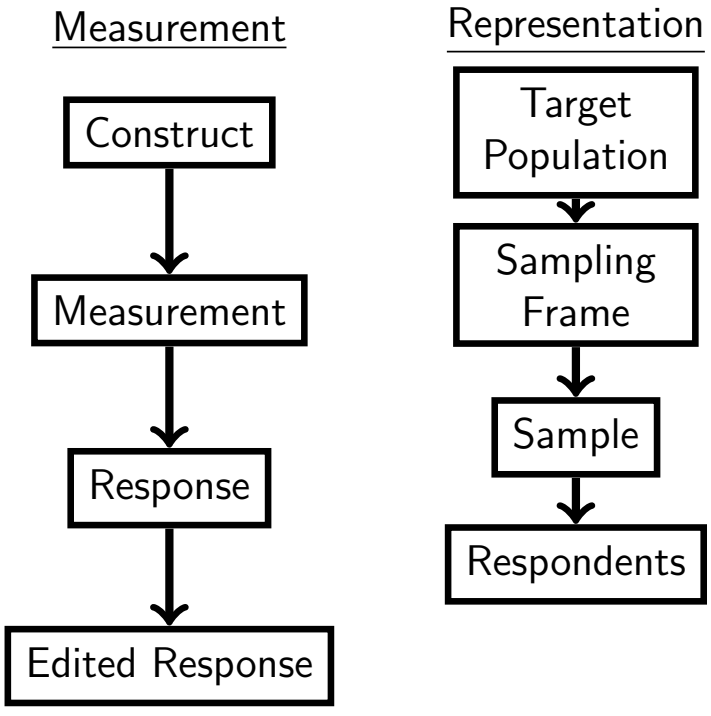
Target Population: the set of units to be studied

Sampling Frame: the set of members that has a chance to be selected into the survey sample

Sample: units selected from a sampling frame

Survey Lifecycle

Design Perspective



Target Population: the set of units to be studied

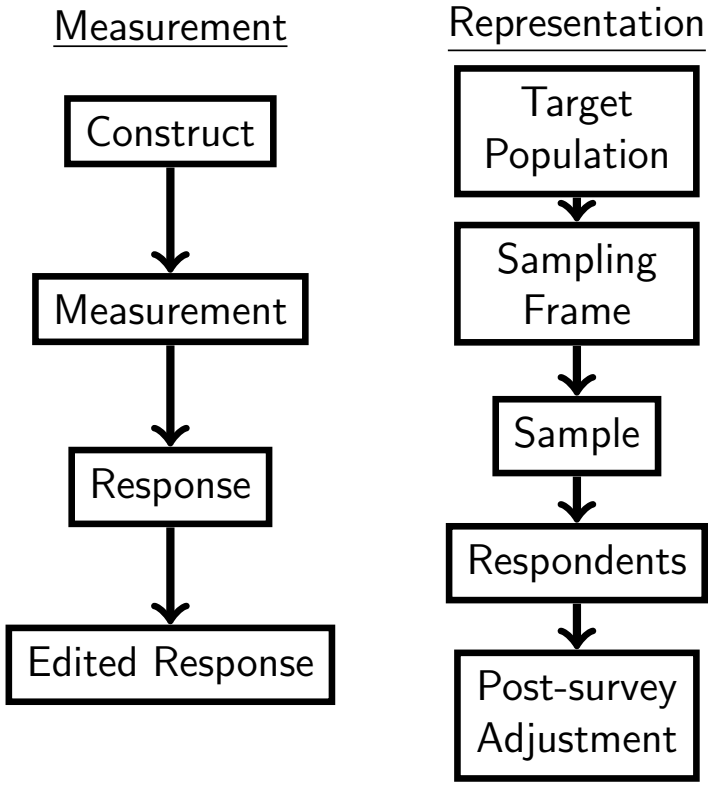
Sampling Frame: the set of members that has a chance to be selected into the survey sample

Sample: units selected from a sampling frame

Respondents: units successfully measured

Survey Lifecycle

Design Perspective



Target Population: the set of units to be studied

Sampling Frame: the set of members that has a chance to be selected into the survey sample

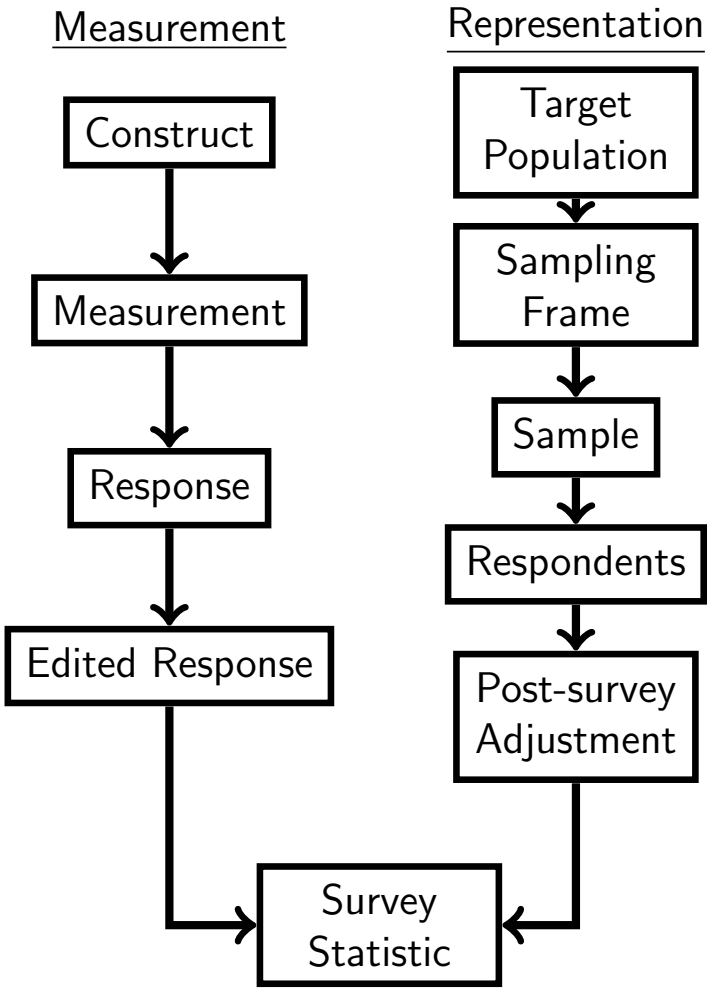
Sample: units selected from a sampling frame

Respondents: units successfully measured

Post-survey Adjustment: improve quality of estimate (e.g. adjust for non-response)

Survey Lifecycle

Design Perspective

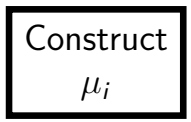


Survey Statistic: The quantity estimated in the resulting survey data

Survey Lifecycle

Quality Perspective

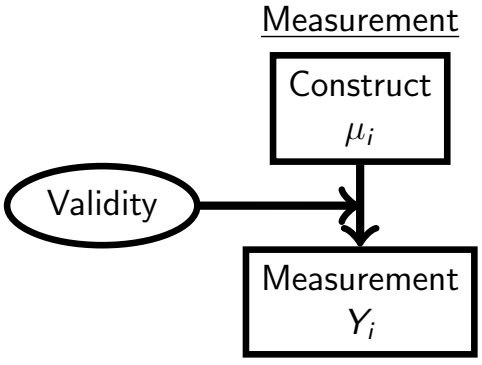
Measurement



Construct: The estimand of interest

Survey Lifecycle

Quality Perspective

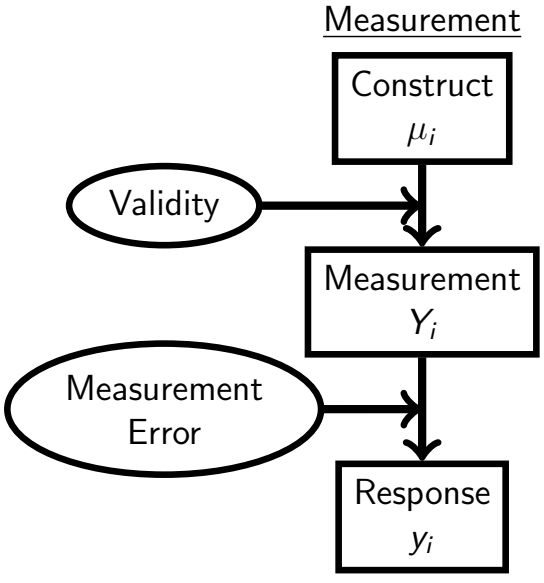


Measurement: Y_i – the value of the measurement

Construct Validity: The extent to which the measure is related to the underlying construct.

Survey Lifecycle

Quality Perspective



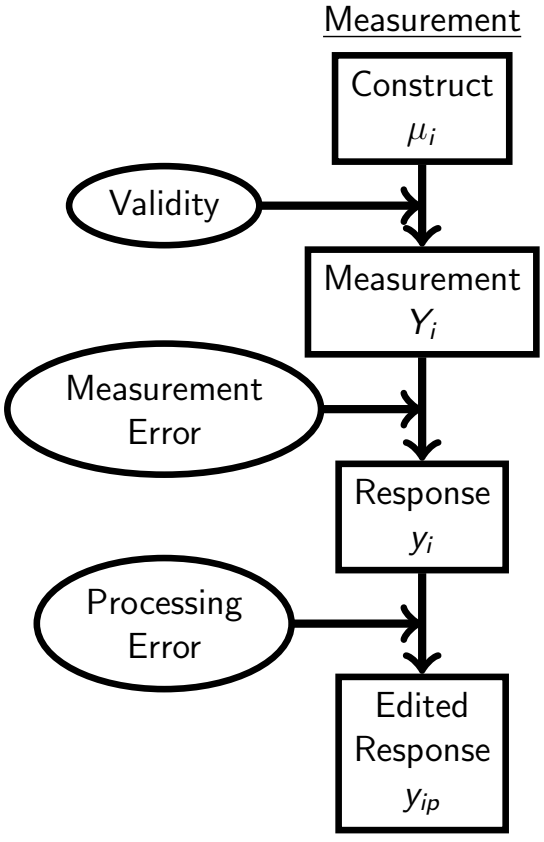
Response: y_i – value of response at application of measurement

Measurement error:
$Y_i \neq y_i$

Response Bias
$E_t[Y_{it}] \neq Y_i$

Survey Lifecycle

Quality Perspective

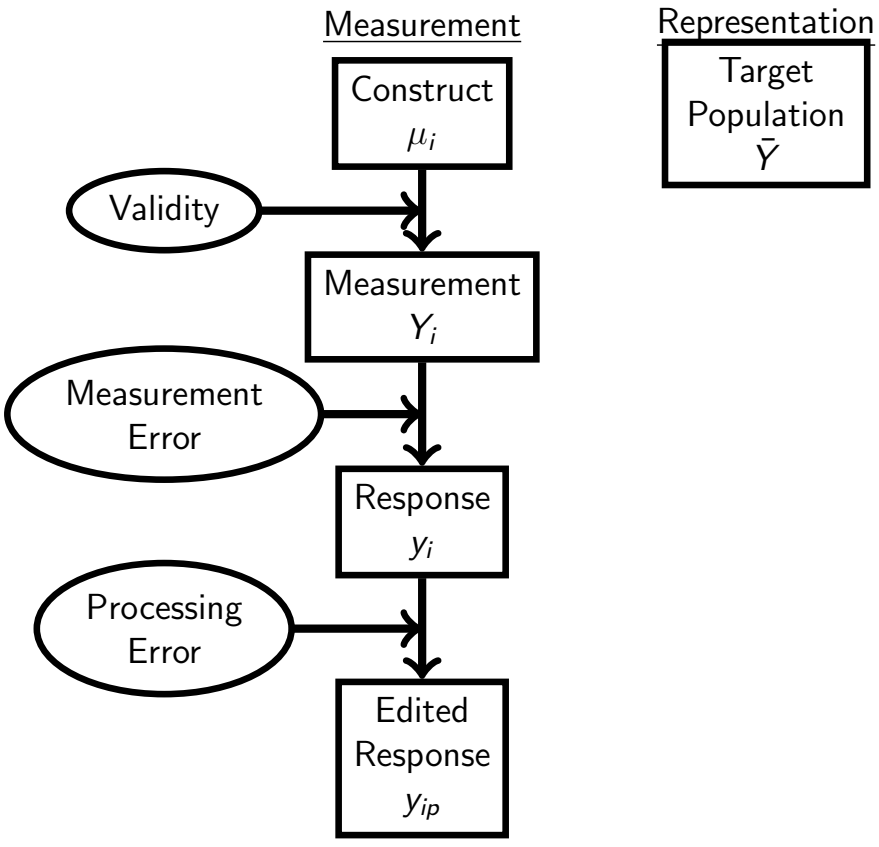


Edited Response: y_{ip} – value of response after editing and processing

Processing error: $y_i \neq y_{ip}$

Survey Lifecycle

Quality Perspective

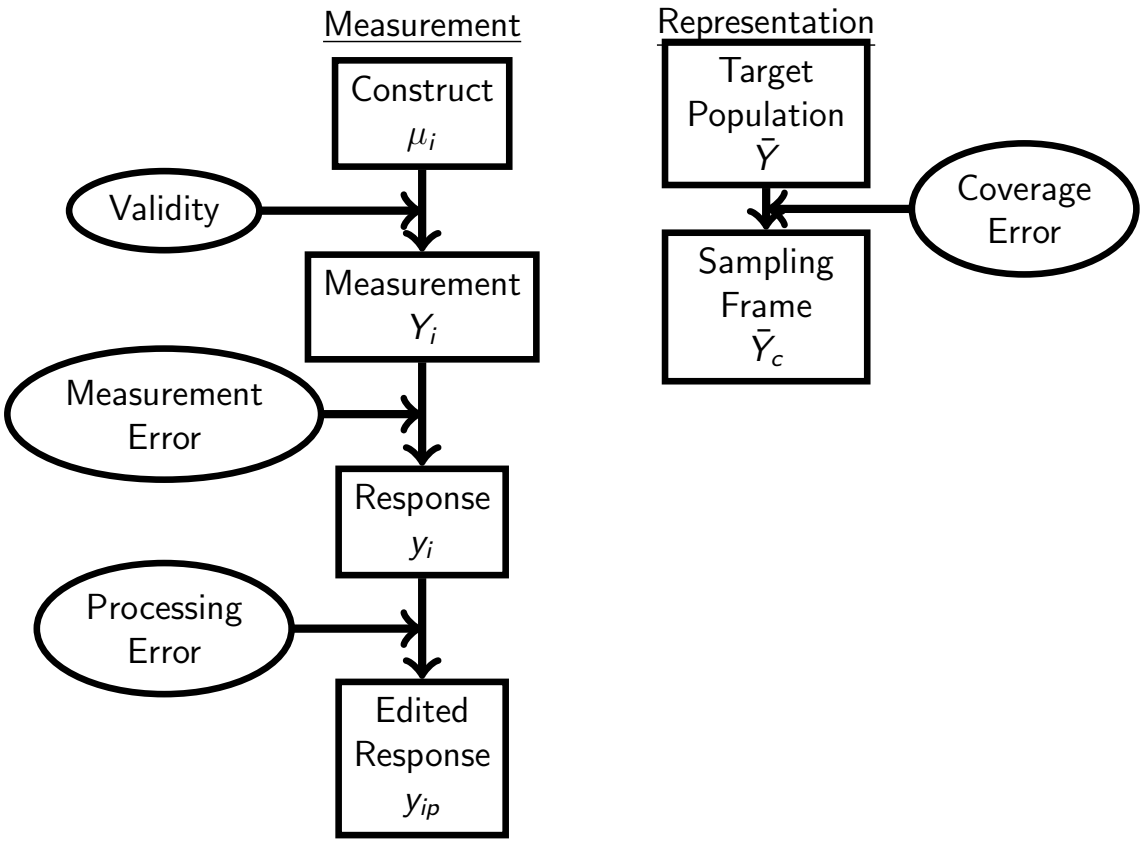


Target Population:

$\bar{Y}$ – mean of measurement in the target population

Survey Lifecycle

Quality Perspective

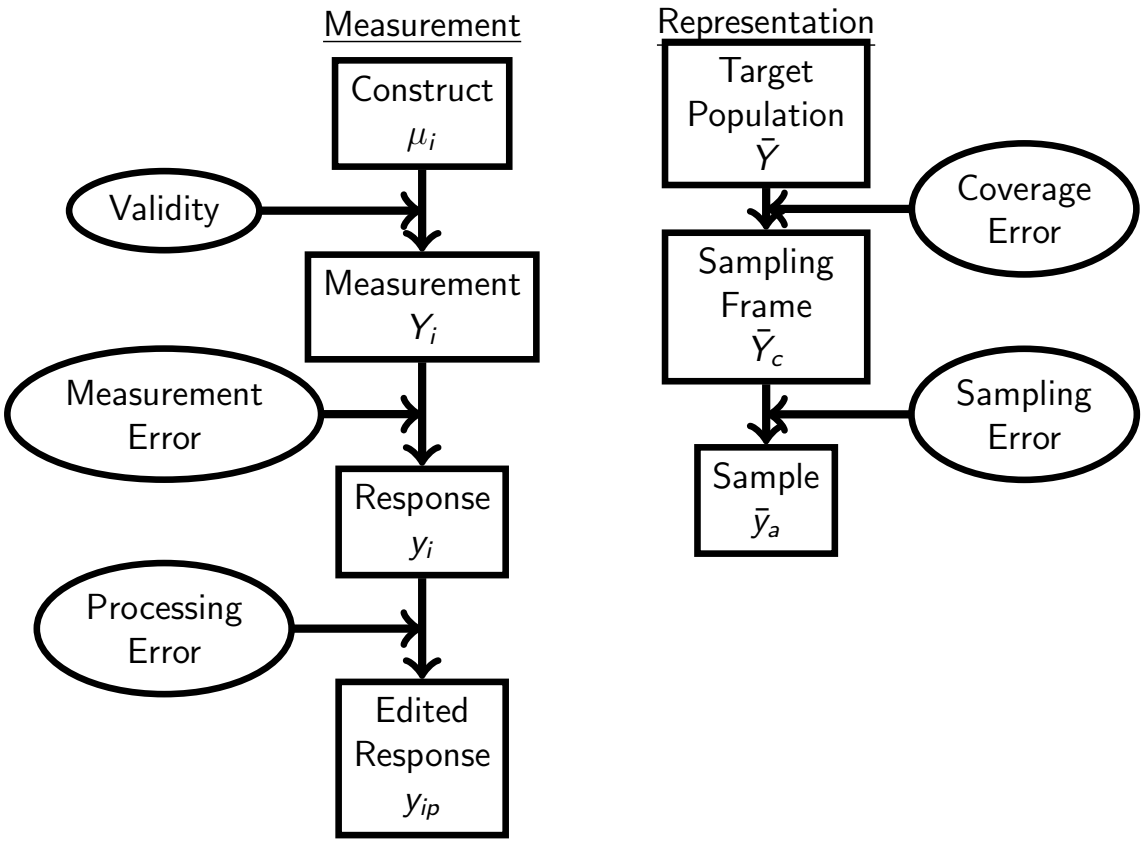


Sampling Frame: $\bar{Y}_c$ – mean of measurement in the sampling frame

Coverage error: C are covered, and U are not $\bar{Y}_C - \bar{Y} = \frac{U}{N}(\bar{Y}_C - \bar{Y}_U)$

Survey Lifecycle

Quality Perspective

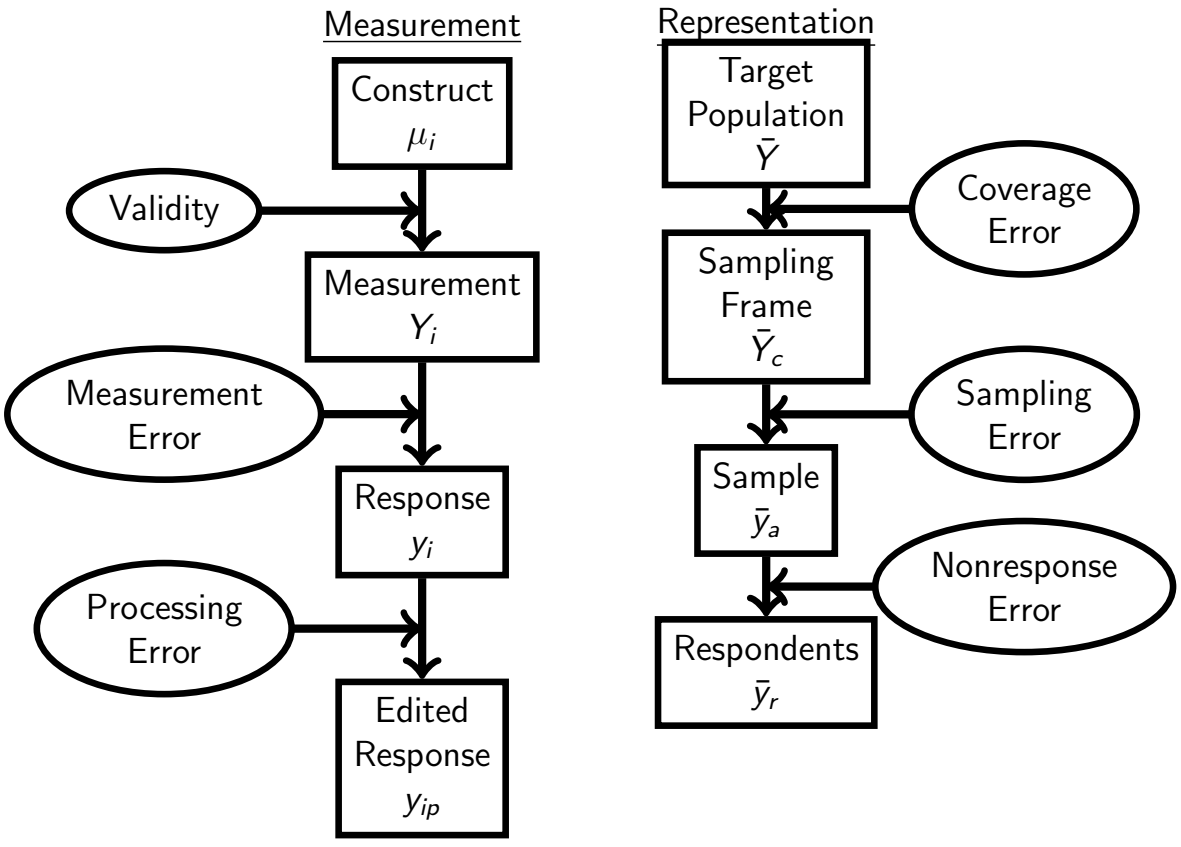


Sample: $\bar{y}_a$ – mean of response in the sample

Sampling error: approximated using known (non-zero) sampling weights

Survey Lifecycle

Quality Perspective



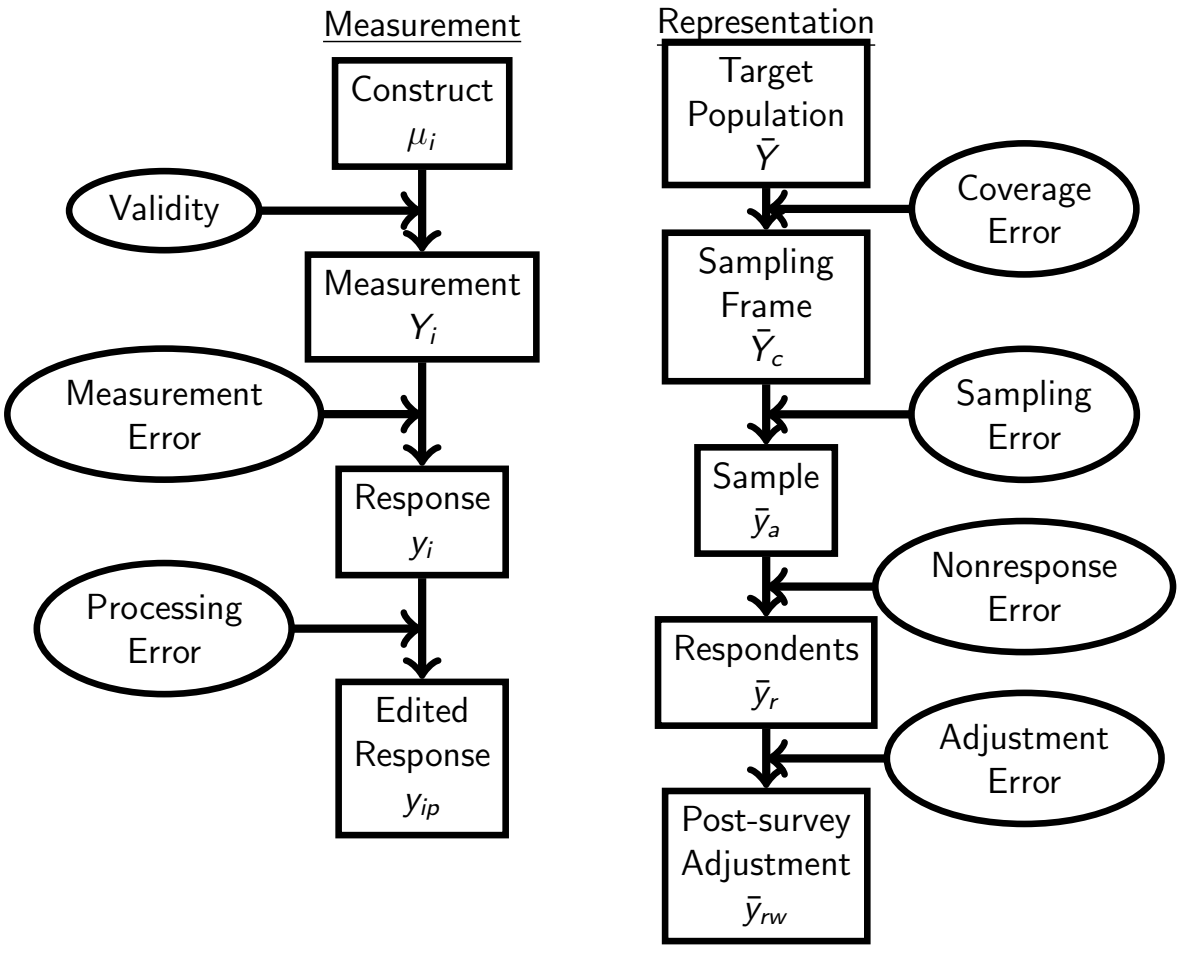
Respondents: $\bar{y}_r$ – mean of measurement among responders

Nonresponse error: n is sample size, m is number of non-responders

$$\bar{y}_r - \bar{y}_s = \frac{m}{n}(\bar{y}_r - \bar{y}_m)$$

Survey Lifecycle

Quality Perspective

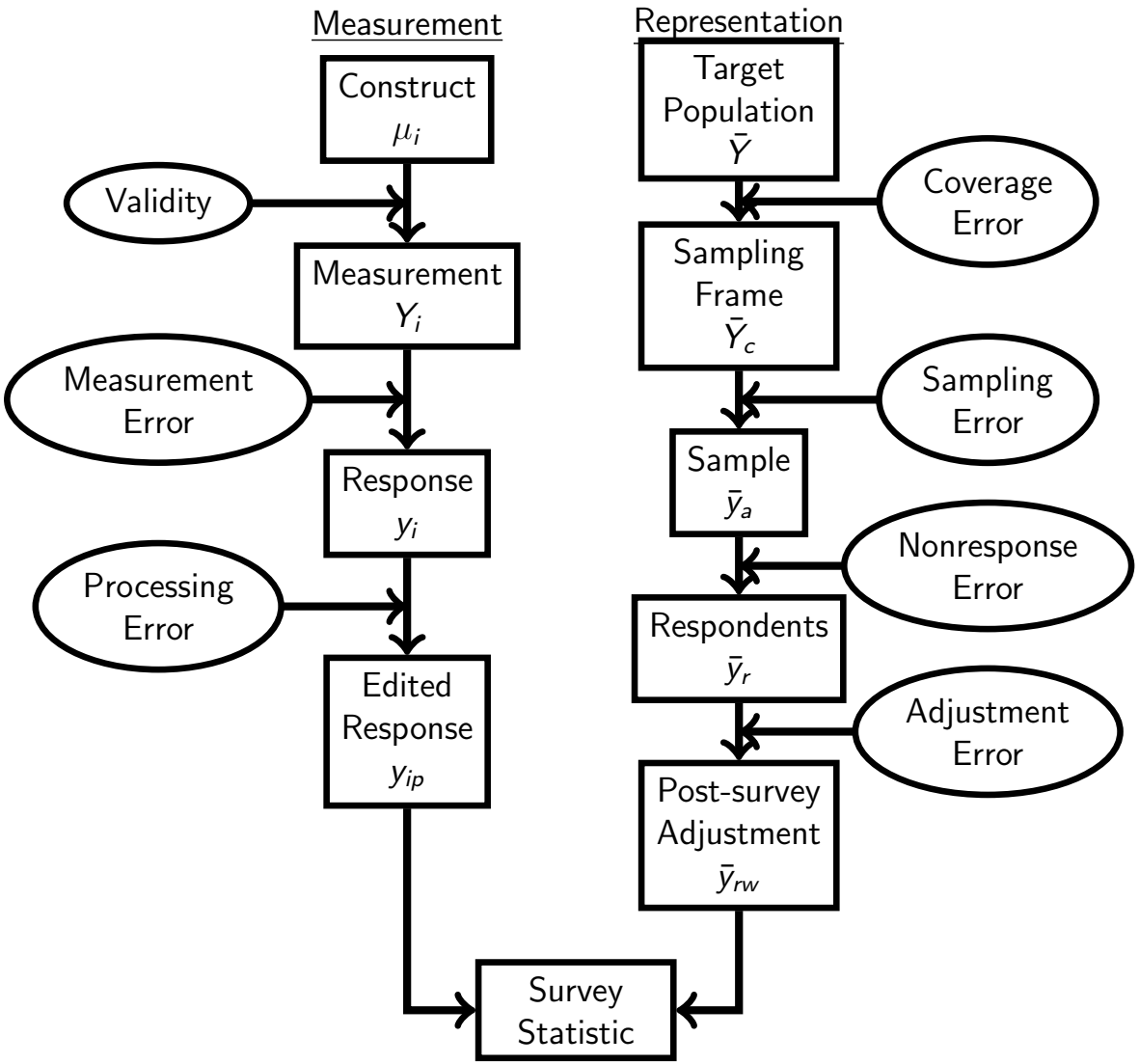


Post-survey Adjustment: $\bar{y}_{rw}$ – mean of measurement among responders, with adjustment

Adjustment error: $\bar{y}_{rw} - \bar{Y}$

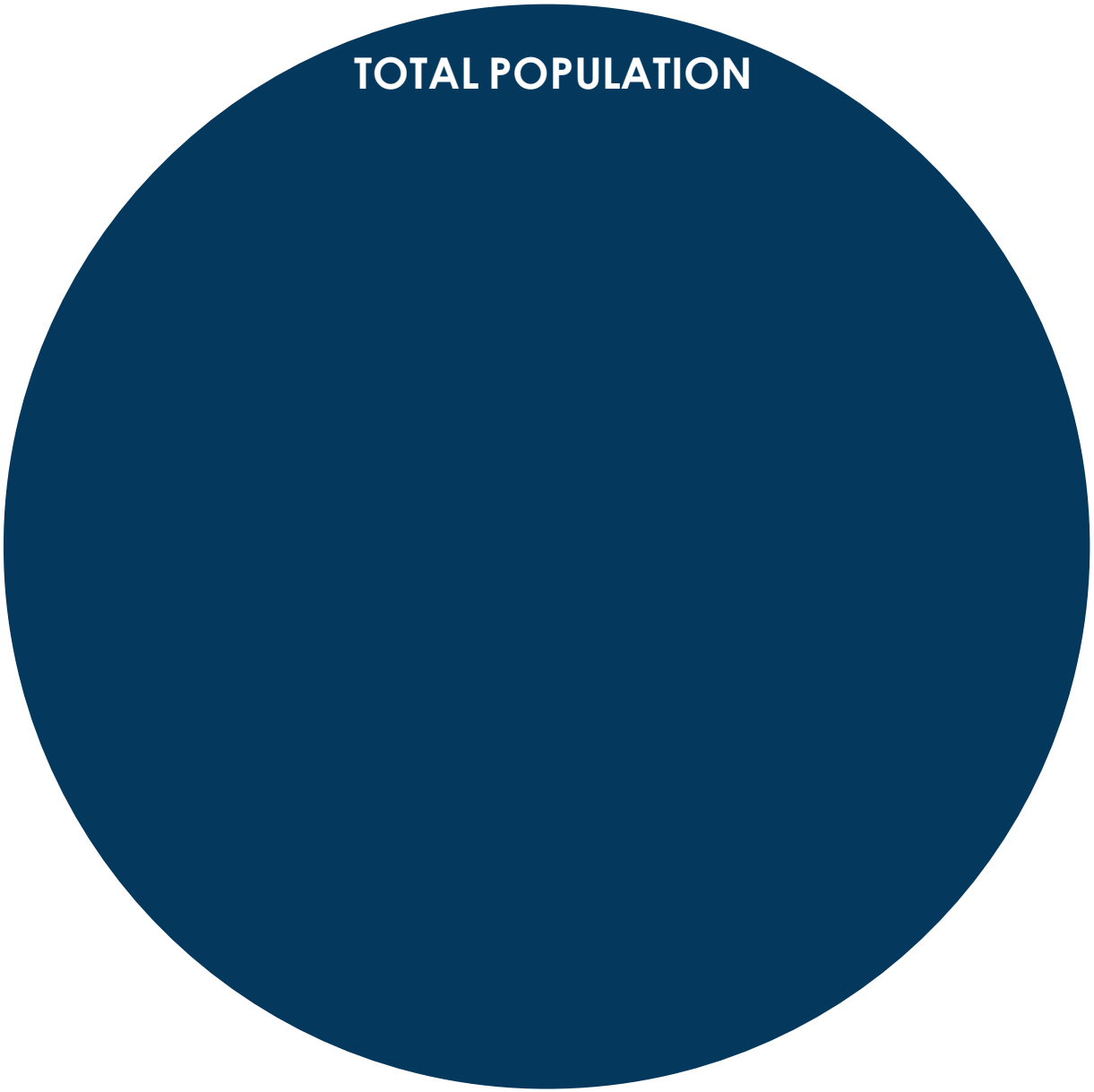
Survey Lifecycle

Quality Perspective

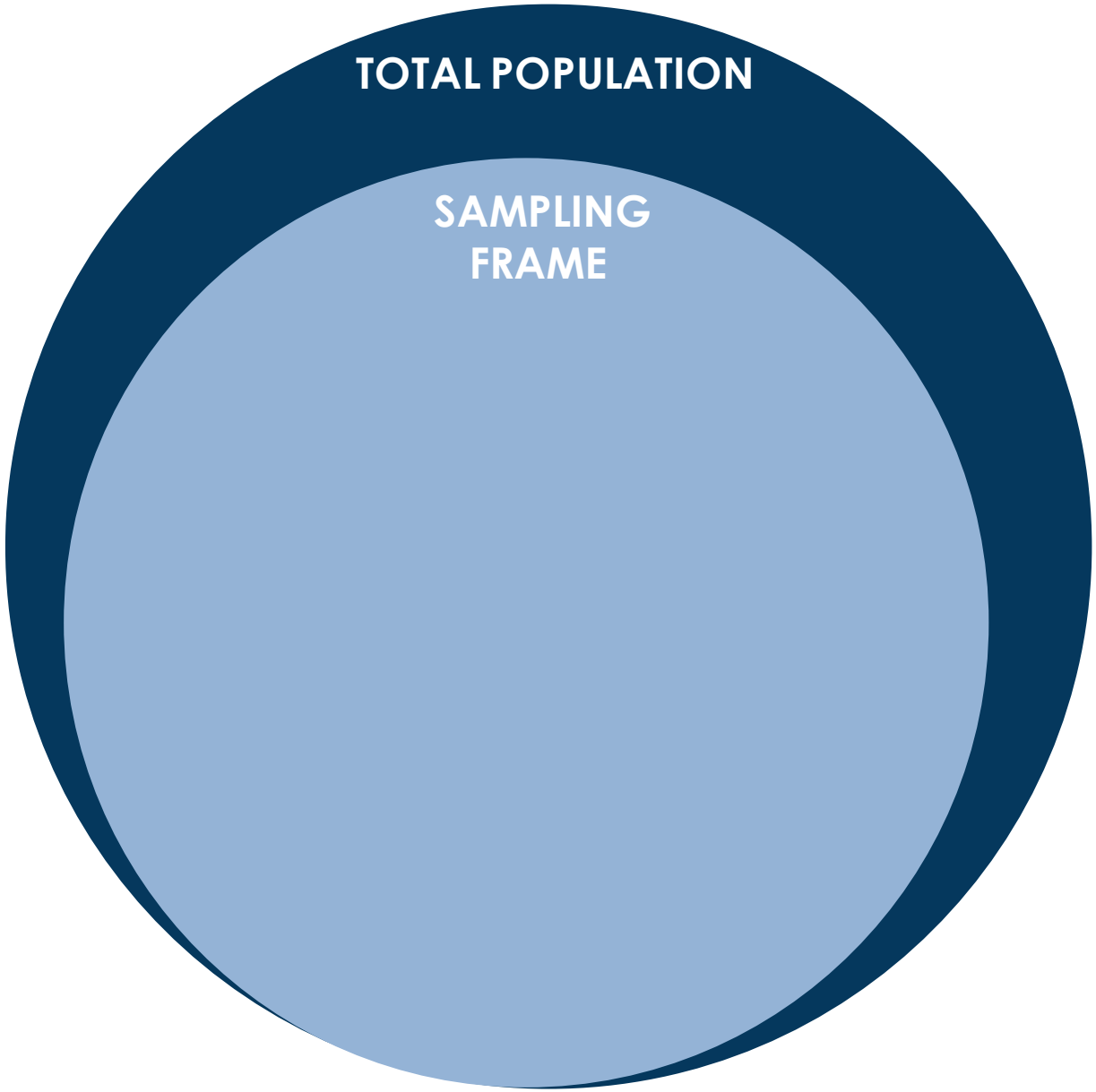


Total Survey Error: difference between the survey statistic and the construct value in the target population.

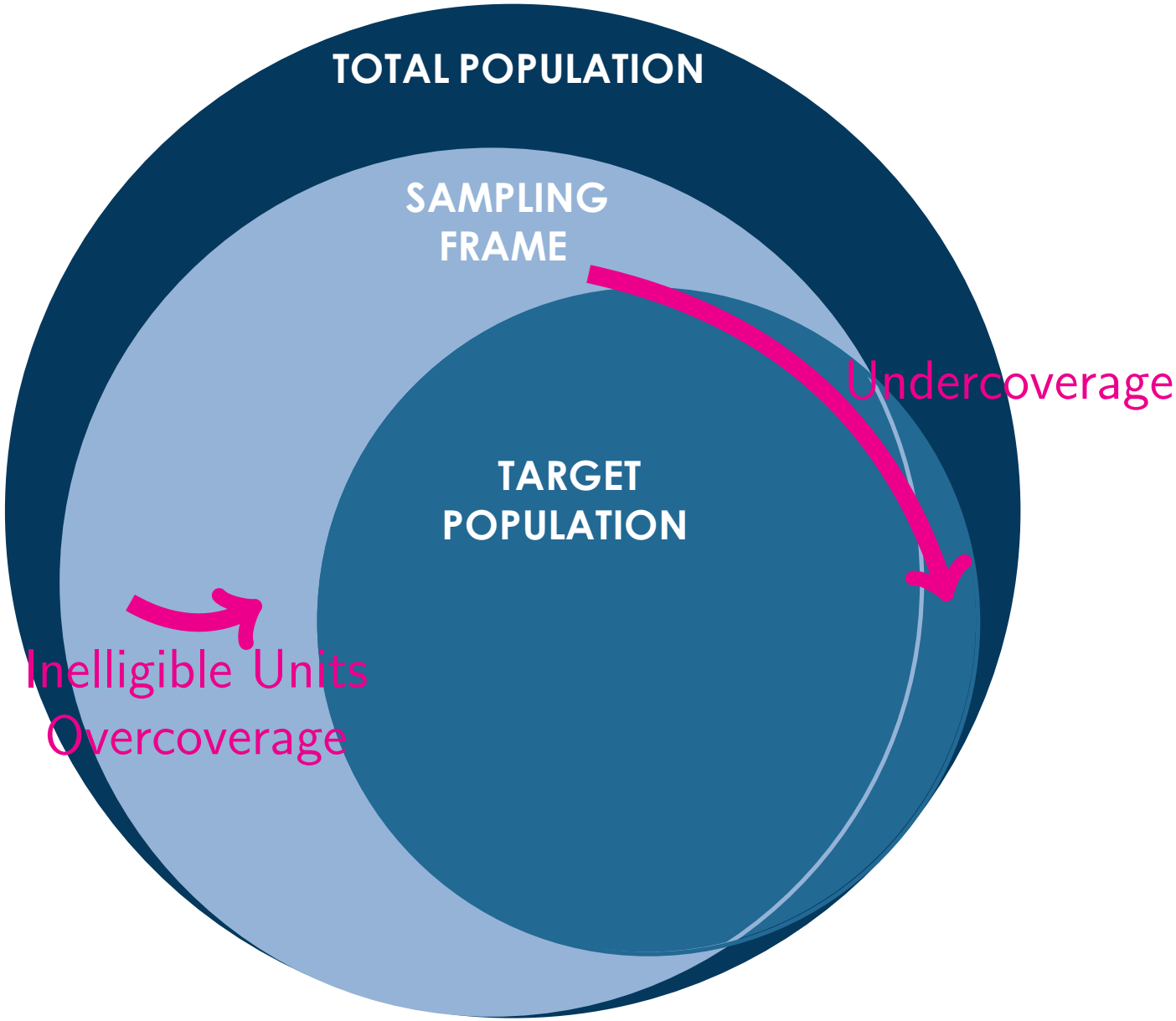
Representation



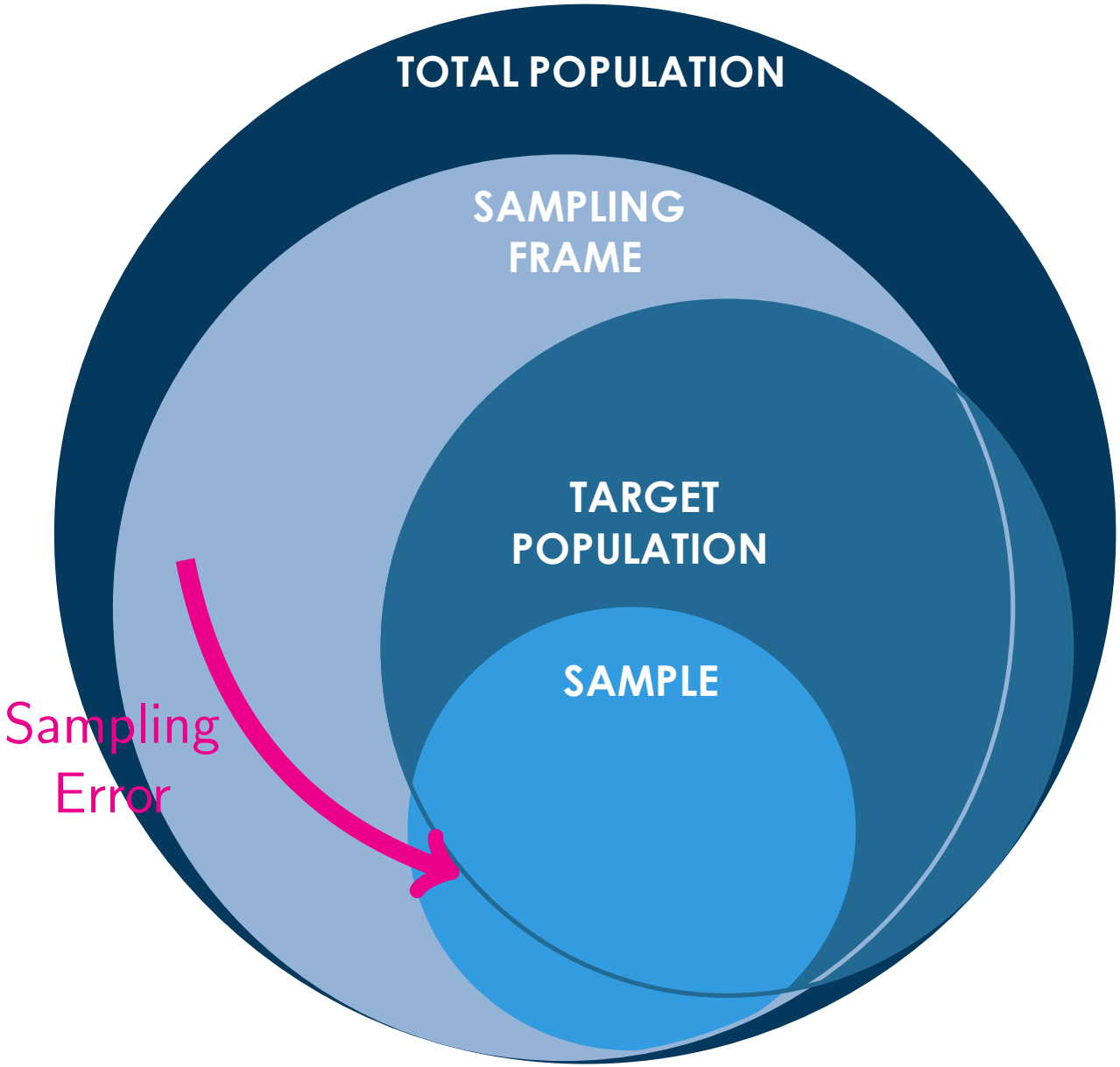
Representation



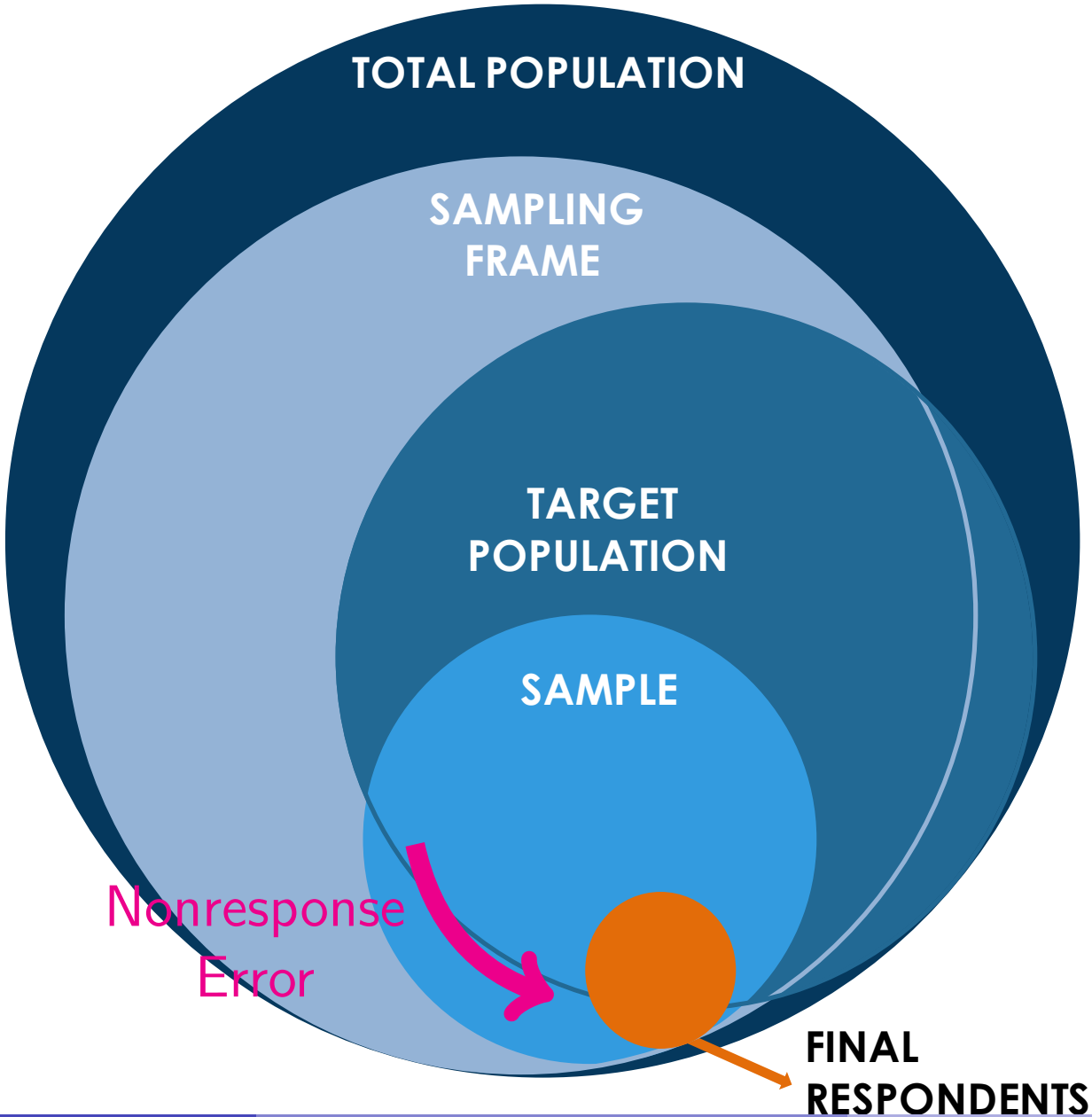
Representation



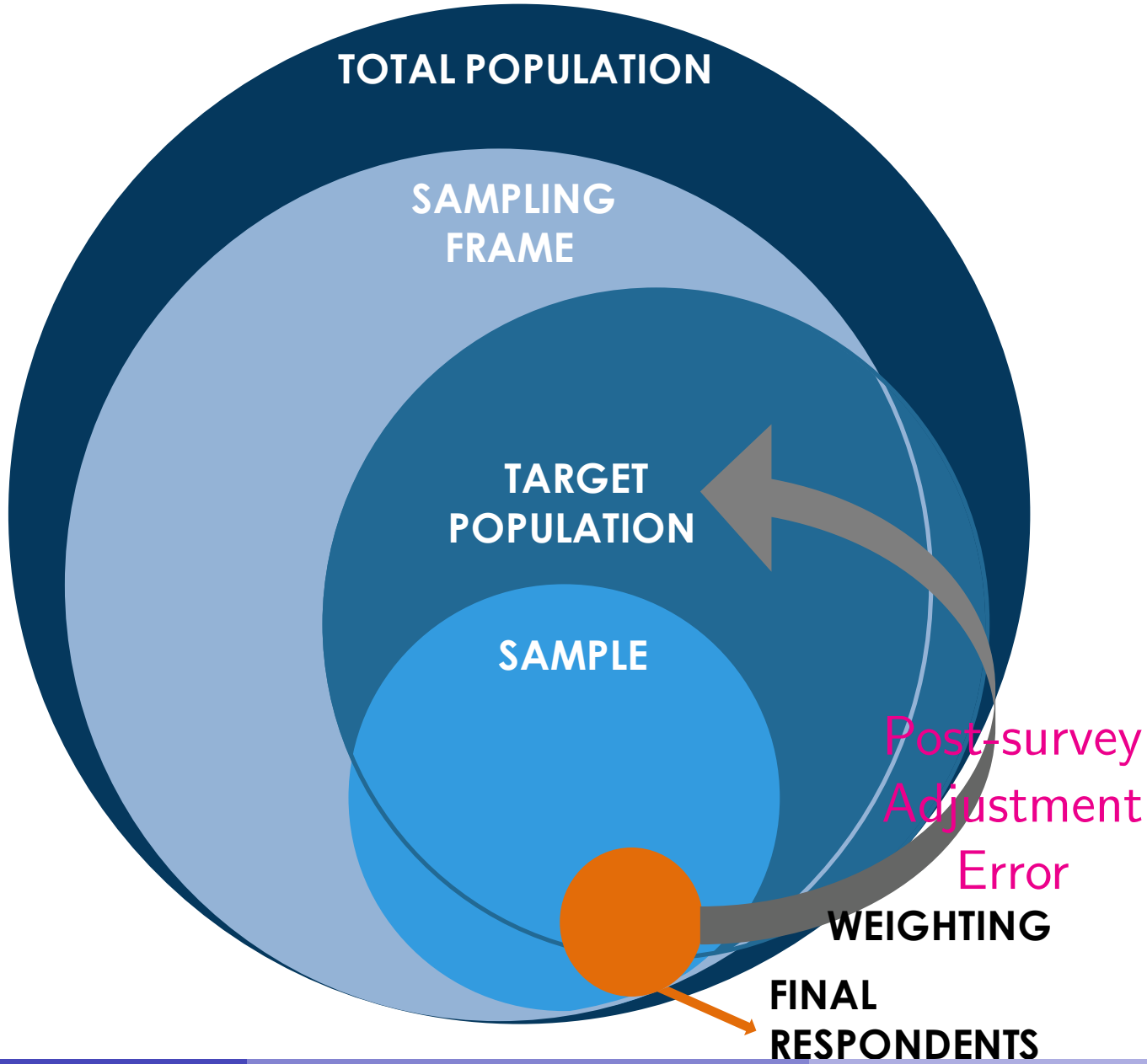
Representation



Representation



Representation



Shadow Component: Cost

Shadow Component: Cost

- Total Survey Error breaks into measurement and representation but doesn't really consider cost.

Shadow Component: Cost

- Total Survey Error breaks into measurement and representation but doesn't really consider cost.
- Cost—time and money—is a big factor in decision making for surveys.

Shadow Component: Cost

- Total Survey Error breaks into measurement and representation but doesn't really consider cost.
- Cost—time and money—is a big factor in decision making for surveys.
- We want to be right but key question is how do we invest resources in making that happen.

Shadow Component: Cost

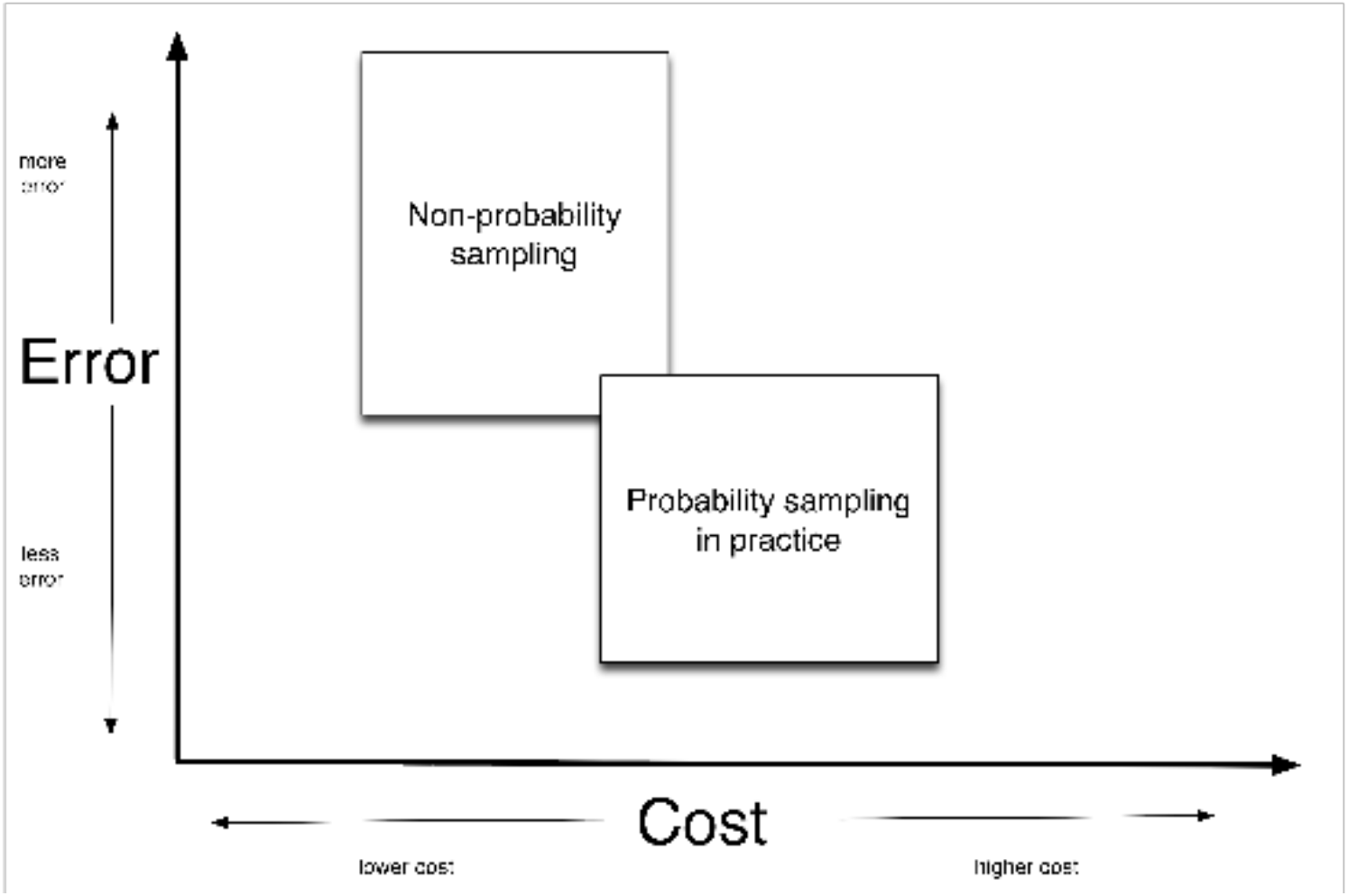
- Total Survey Error breaks into measurement and representation but doesn't really consider cost.
- Cost—time and money—is a big factor in decision making for surveys.
- We want to be right but key question is how do we invest resources in making that happen.
- It is sometimes easier to correct some parts of the process than others!

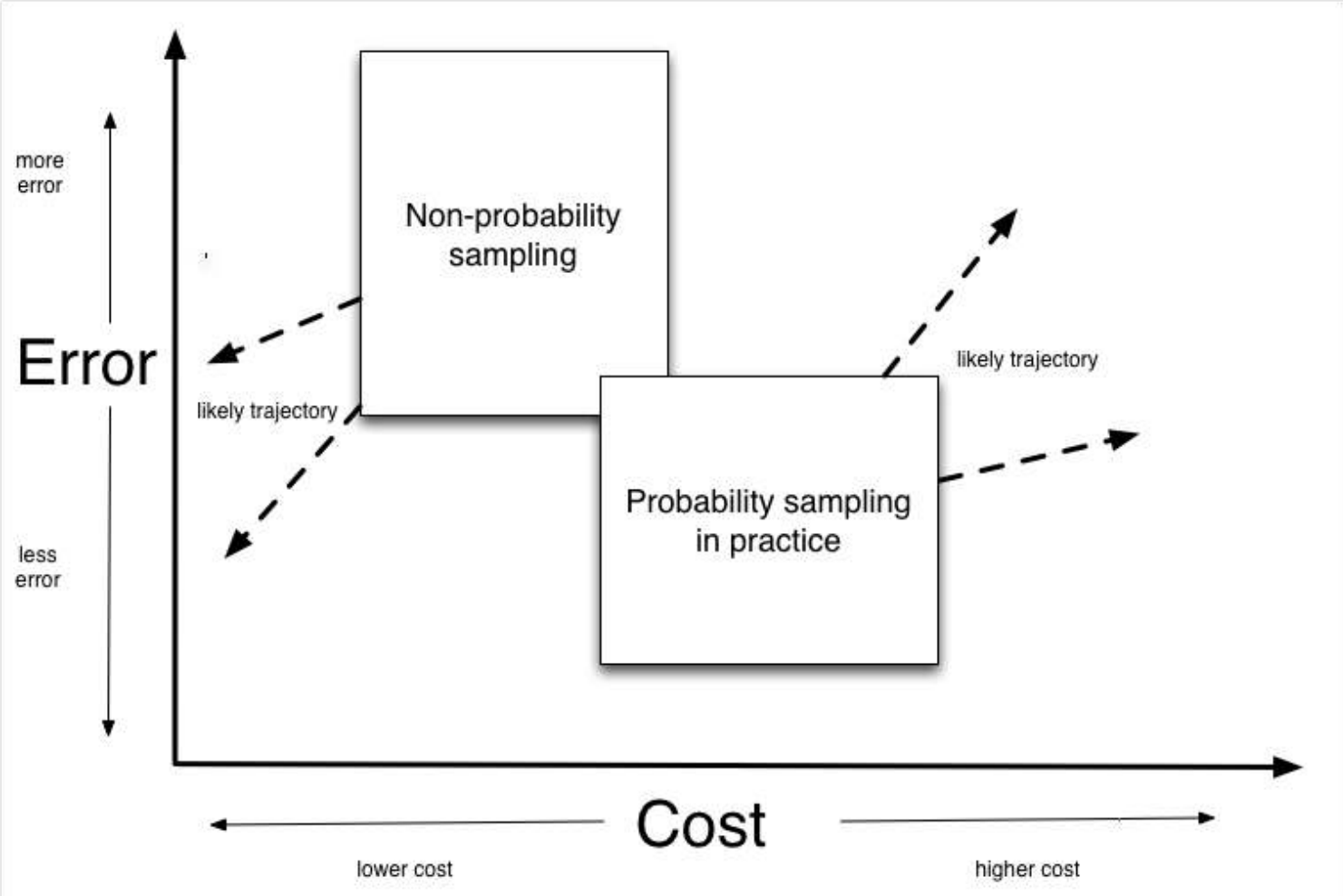
1. Surveys
2. Total Survey Error Framework
3. Integrating Surveys with Other Data
4. Big Data Paradox
5. Writing Survey Questions is Difficult
6. Record Linkage

Probability vs. Non-probability Samples

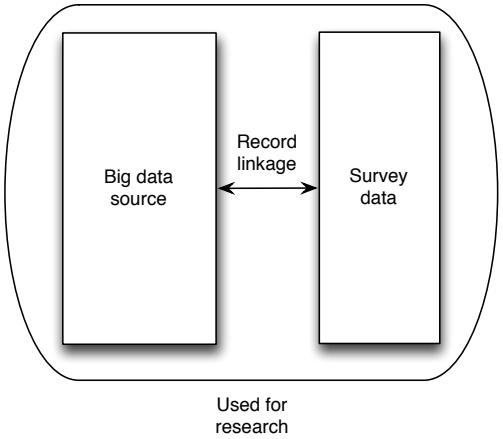
Probability vs. Non-probability Samples

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

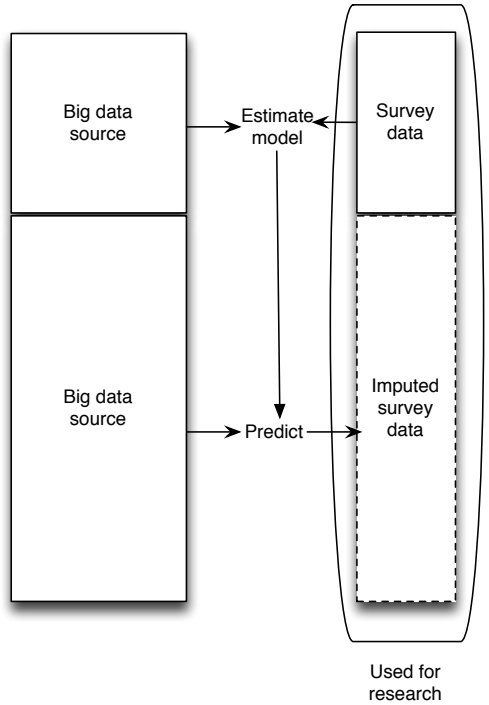

more
error
Error
less
error
Non-probability
sampling
Probability sampling
in practice
Cost
lower cost
higher cost


more
error
Error
less
error
Non-probability
sampling
likely trajectory
likely trajectory
Probability sampling
in practice
Cost
lower cost
higher cost

Enriched asking



Amplified asking

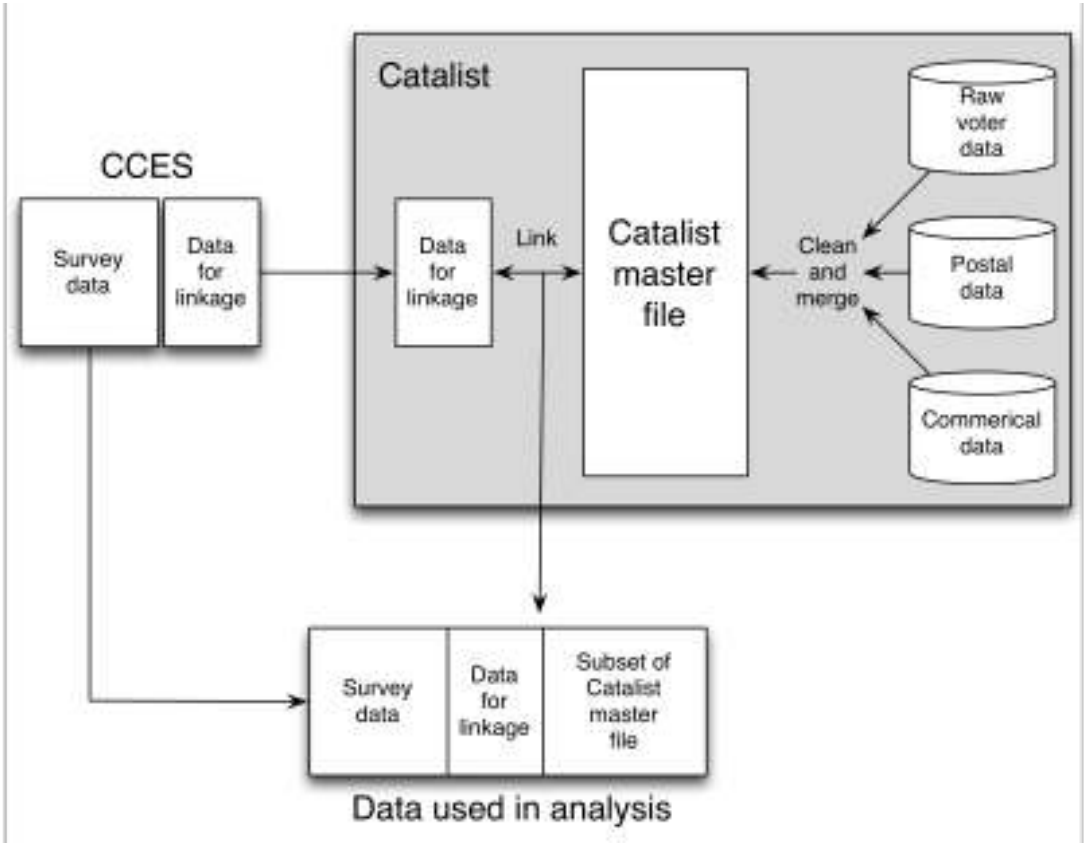


Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate

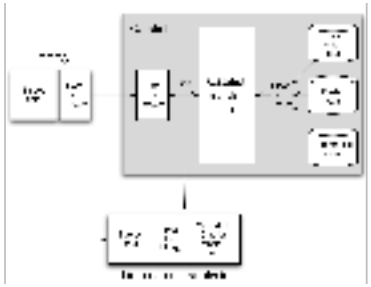
Stephen Ansolabehere
Harvard University

Eitan Hersh
Institution for Social and Policy Studies, Yale University, New Haven, CT 06520-8209
e-mail: eitan.hersh@yale.edu (corresponding author)

Ansolabehere and Hersh (2012)

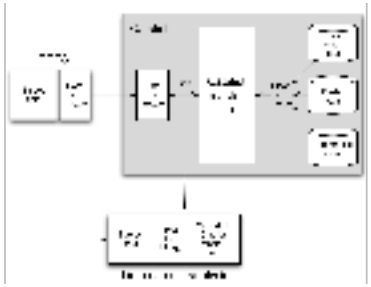


Ansolabehere and Hersh (2012)



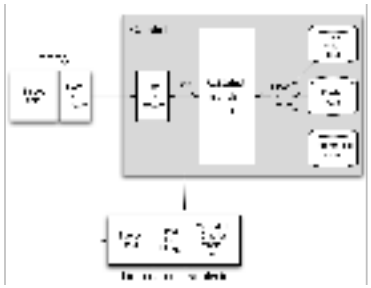
Findings:

- Over-reporting of voting is rampant: almost half of non-voters reported voting, someone who reported voting had 80% chance of actually voting



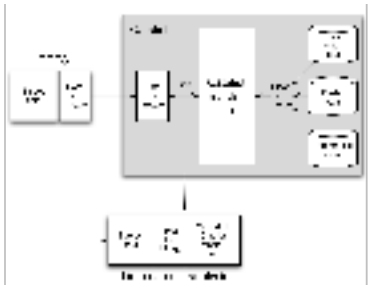
Findings:

- Over-reporting of voting is rampant: almost half of non-voters reported voting, someone who reported voting had 80% chance of actually voting
- Over-reporting is not random; it is more common among high-income, well-educated, partisans engaged in civic affairs



Findings:

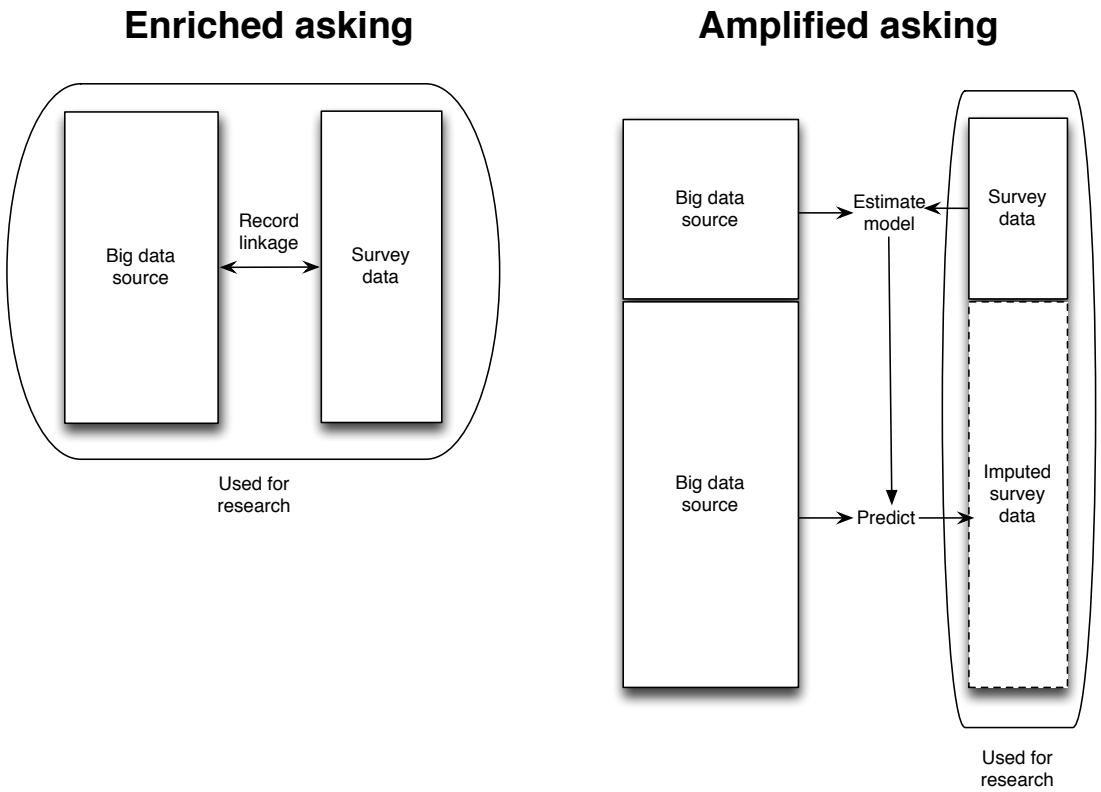
- Over-reporting of voting is rampant: almost half of non-voters reported voting, someone who reported voting had 80% chance of actually voting
- Over-reporting is not random; it is more common among high-income, well-educated, partisans engaged in civic affairs
- Because of systematic over-reporting, differences between voters and nonvoters are smaller than they appear from surveys



Findings:

- Over-reporting of voting is rampant: almost half of non-voters reported voting, someone who reported voting had 80% chance of actually voting
- Over-reporting is not random; it is more common among high-income, well-educated, partisans engaged in civic affairs
- Because of systematic over-reporting, differences between voters and nonvoters are smaller than they appear from surveys
- Existing theories are better at predicting who will reporting voting than who will actually vote

Ansolabehere and Hersh (2012)

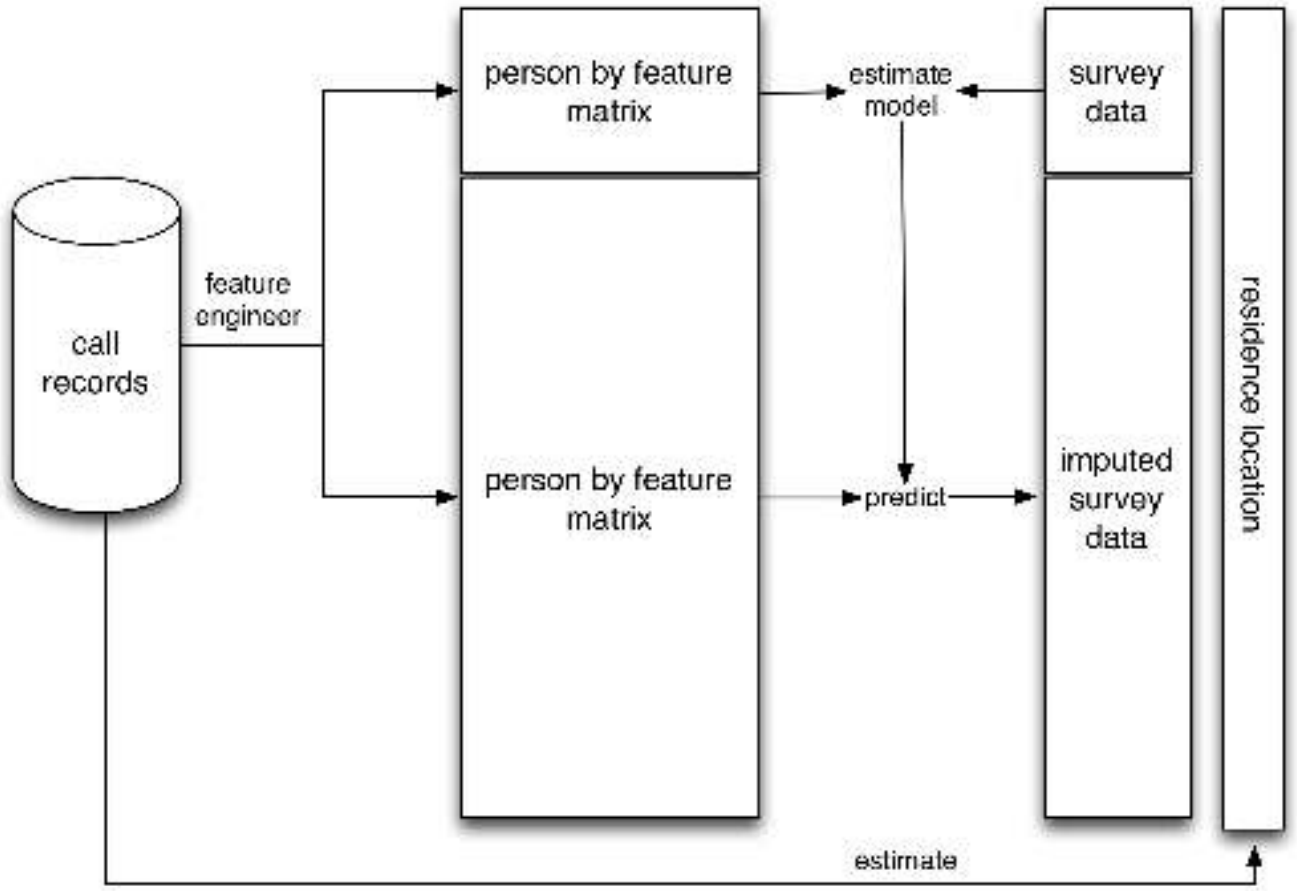


Note the different role of the big data in each case

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1]* Gabriel Cadamuro,[2] Robert On[3]

Blumenstock, Cadamuro, and On (2015)


person by feature matrix
estimate model
survey data
call records
feature engineer
person by feature matrix
predict
imputed survey data
residence location
estimate

1. Surveys
2. Total Survey Error Framework
3. Integrating Surveys with Other Data
4. Big Data Paradox
5. Writing Survey Questions is Difficult
6. Record Linkage

Bradley et. al 2021

Bradley et. al 2021

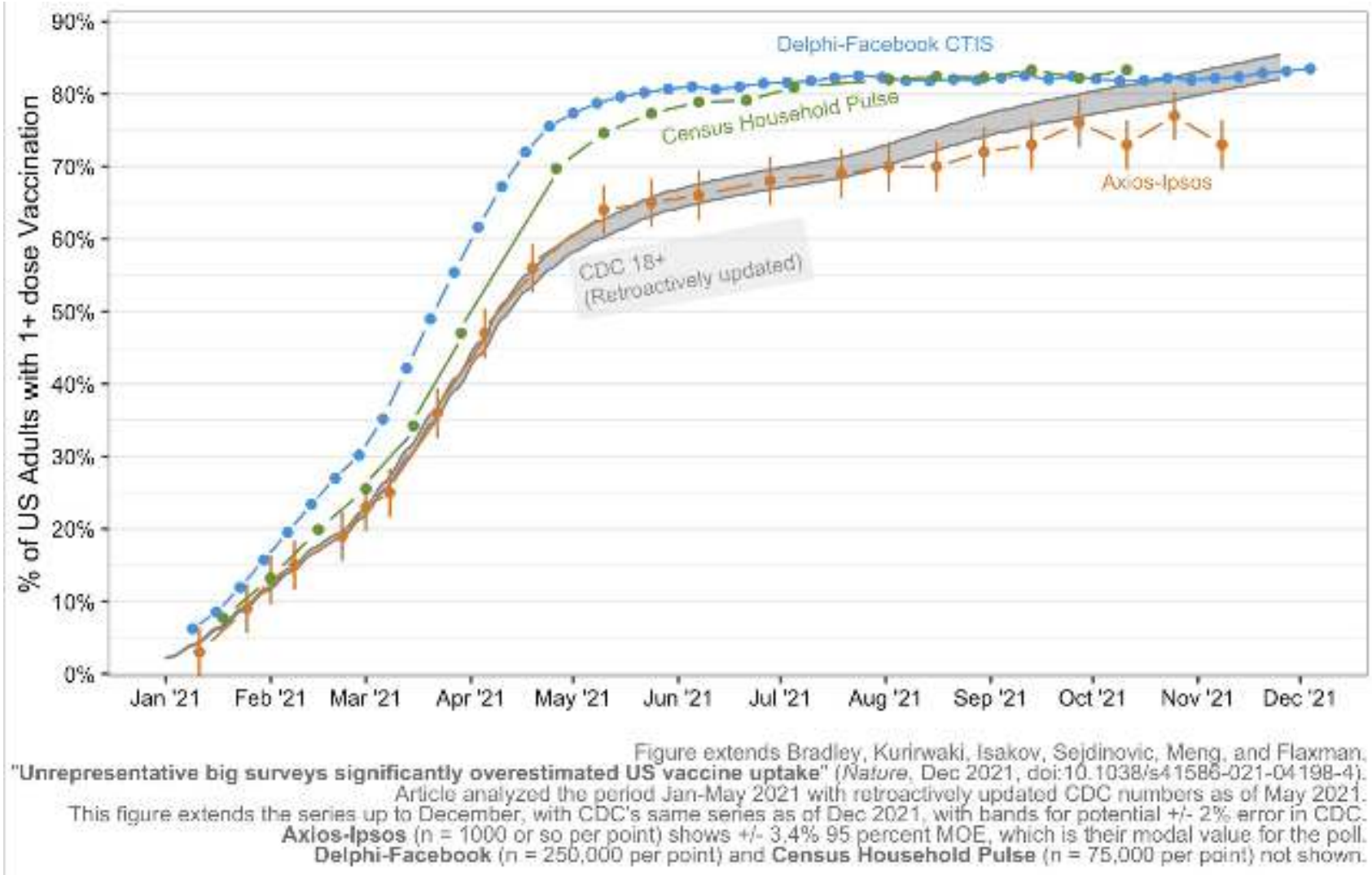


Figure extends Bradley, Kurinwaki, Isakov, Sejdinovic, Meng, and Flaxman.
"**Unrepresentative big surveys significantly overestimated US vaccine uptake**" (*Nature*, Dec 2021, doi:10.1038/s41586-021-04198-4).
Article analyzed the period Jan-May 2021 with retroactively updated CDC numbers as of May 2021.
This figure extends the series up to December, with CDC's same series as of Dec 2021, with bands for potential +/- 2% error in CDC.
Axios-Ipsos (n = 1000 or so per point) shows +/- 3.4% 95 percent MOE, which is their modal value for the poll.
Delphi-Facebook (n = 250,000 per point) and **Census Household Pulse** (n = 75,000 per point) not shown.

Bradley et. al 2021

Table 1 | Comparison of survey designs

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Recruitment mode	Address-based mail sample to Ipsos KnowledgePanel	SMS and email	Facebook Newsfeed
Interview mode	Online	Online	Online
Average size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; internet/ tablets provided to ~5% of panelists who lack home internet	Census Bureau's Master Address File (individuals for whom email / phone contact information is available)	Facebook active users
Vaccine uptake question	"Do you personally know anyone who has already received the COVID-19 vaccine?"	"Have you received a COVID-19 vaccine?"	"Have you had a COVID-19 vaccination?"
Vaccine uptake definition	"Yes, I have received the vaccine"	"Yes"	"Yes"
Other vaccine uptake response options	"Yes, a member of my immediate family", "Yes, someone else", "No"	"No"	"No", "I don't know"
Weighting variables	Gender by age, race, education, Census region, metropolitan status, household income, partisanship.	Education by age by sex by state, race/ethnicity by age by sex by state, household size	Stage 1: age, gender "other attributes which we have found in the past to correlate with survey outcomes" to FAUB; Stage 2: state by age by gender

Comparison of key design choices across the Axios–Ipsos, Census Household Pulse and Delphi–Facebook studies. All surveys target the US adult population. See Extended Data Table 1 for additional comparisons and Methods for additional implementation details.

Bradley et. al 2021

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Purpose	Measure national attitudes toward COVID-19	Sub-national social and economic impact of COVID-19	Fine-grained COVID-19 symptom surveillance
Target Pop.	18+ US general pop	18+ US general pop	18+ US general pop
Length of wave	4 days, conducted weekly	2 weeks	Daily cross-section samples, reported weekly
Average participation rate among invitees	50%	6-8%	1%
Sampling design	Inverse response propensity sampling	Systematic sample of households, adjusted for a projected response rates	Unequal-probability stratified random samples
Hesitancy / Willingness question	"How likely, if at all, are you to get the first generation COVID-19 vaccine, as soon as it's available"	"Once a vaccine preventing COVID-19 is available to you, would you..."	"If a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated?"
Vaccine hesitancy responses	"Not very / at all likely"	"Definitely/Probably NOT get a vaccine" or "Unsure"	"No, definitely/probably not"
Languages	English and Spanish	English and Spanish	English, Spanish, Brazilian Portuguese, Vietnamese, French, and Chinese
Report MoE or design effect	Both	Report standard errors for estimates from replicate weights	Report standard errors for estimates (does not include variance from weighting)
Sources for demographic benchmarks	2019 CPS March Supplement, party ID from recent ABC/WaPo polls	2018 ACS, 1-year estimates	2018 CPS March Supplement

Bradley et. al 2021

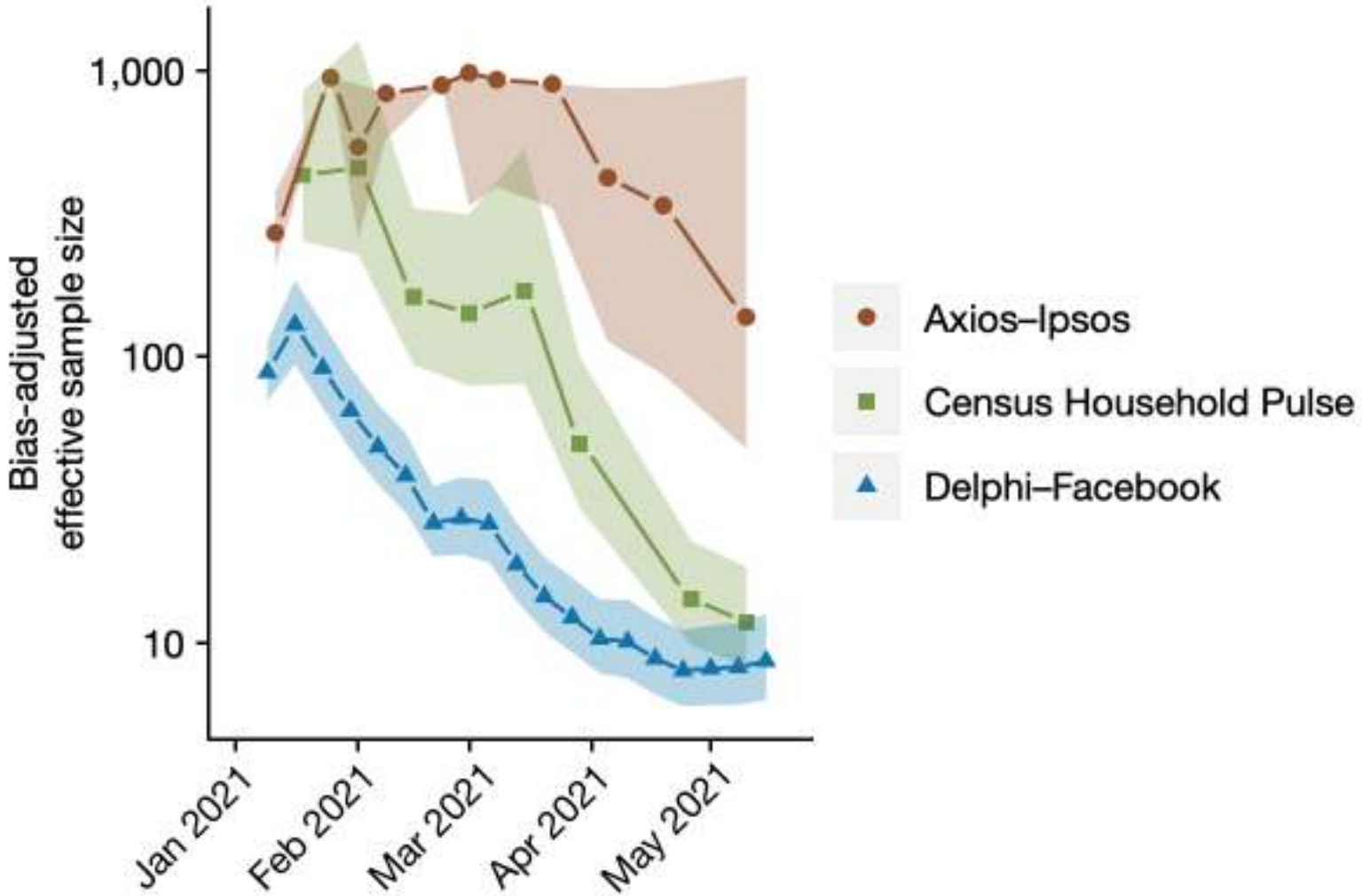


Fig 2 | Bias-adjusted effective sample size. An estimate's bias-adjusted effective sample size (different from the classic Kish effective sample size) is the size of a simple random sample that would have the same MSE as the observed estimate. Effective sample sizes are shown here on the $\log_{10}$ scale.

Bradley et. al 2021

Table 2 | Composition of survey respondents by educational attainment and race/ethnicity

	Composition of US adults							Survey estimates		
	Axios–Ipsos		**Household Pulse**		**Delphi–Facebook**		**ACS**	**Household Pulse**		
	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
Education										
High school	35%	39%	14%	39%	19%	21%	39%	39%	40%	21%
Some college	29	30	32	30	36	36	30	44	38	18
Four-year college	19	17	29	17	25	25	19	54	36	10
Post-graduate	17	14	25	13	20	18	11	67	25	7
Race/ethnicity										
White	71%	63%	75%	62%	74%	68%	60%	50%	33%	17%
Black	10	12	7	11	5	6	12	42	39	19
Hispanic	11	16	10	17	11	16	16	38	48	14
Asian			5	5	2	3	6	51	43	5

Axios–Ipsos: wave ending 22 March 2021, $n = 995$. Census Household Pulse: wave ending 29 March 2021, $n = 75{,}058$. Delphi–Facebook: wave ending 27 March 2021, $n = 187{,}949$. Benchmark uses the 2019 US Census American Community Survey (ACS), composed of roughly 3 million responses. The rightmost column shows estimates of vaccine uptake (Vax), willingness (Will) and hesitancy (Hes) from the Census Household Pulse of the same wave.

1. Surveys
2. Total Survey Error Framework
3. Integrating Surveys with Other Data
4. Big Data Paradox
5. Writing Survey Questions is Difficult
6. Record Linkage

PNAS RESEARCH ARTICLE | POLITICAL SCIENCES OPEN ACCESS

Current research overstates American support for political violence

Sean J. Westwood[a] Justin Grimmer[b,1] Matthew Tyler[a] and Clayton Nall[c]

Edited by Anthony Chen, University of Chicago, Chicago, IL; received September 15, 2021; accepted February 7, 2022 by Editorial Board Member Margaret Levi

Political scientists, pundits, and citizens worry that America is entering a new period of violent partisan conflict. Provocative survey data show that a large share of Americans (between 8% and 40%) support politically motivated violence. Yet, despite media attention, political violence is rare, amounting to a little more than 1% of violent hate crimes in the United States. We reconcile these seemingly conflicting facts with four large survey experiments (n = 4,904), demonstrating that self-reported attitudes on political violence are biased upward because of respondent disengagement and survey questions that allow multiple interpretations of political violence. Addressing question wording and respondent disengagement, we find that the median of existing estimates of support for partisan violence is nearly 6 times larger than the median of our estimates (18.5% versus 2.9%). Critically, we show the prior estimates overstate support for political violence because of random responding by disengaged respondents. Respondent disengagement also inflates the relationship between support for violence and previously identified correlates by a factor of 4. Partial identification bounds imply that, under generous assumptions, support for violence among engaged and disengaged respondents is, at most, 6.86%. Finally, nearly all respondents support criminally charging suspects who commit acts of political violence. These findings suggest that, although recent acts of political violence dominate the news, they do not portend a new era of violent conflict.

political violence | affective polarization | democratic norms

Significance

Recent political events show that members of extreme political groups support partisan violence, and survey evidence supposedly shows widespread public support. We show, however, that, after accounting for survey-based measurement error, support for partisan violence is far more limited. Prior estimates overstate support for political violence because of random responding by disengaged respondents and because of a reliance on hypothetical questions about violence in general instead of questions on specific acts of political violence. These same issues also cause the magnitude of the relationship between previously identified correlates and partisan violence to be overstated. As policy makers consider interventions designed to dampen support for violence, our results provide critical information about the magnitude of the problem.

Provocative recent work (1–3)—cited in PNAS (4, 5), *The American Journal of Political Science* (6), 60 other articles and books, and 40 news articles that, together, have garnered over 2,281,136 Twitter engagements—asserts that large segments of the American population now support politically motivated violence. These studies report that up to 44% of Americans would endorse hypothetical violence in some undetermined future event (1–3, 7). This survey work fits within a media landscape that regularly raises the specter of political violence. Since 2016, we counted 2,863 mentions of political violence on news television, more than 650 news stories about political violence, and over 10 million tweets on the topic of the January 6 riot alone (see *SI Appendix*, section 1 for details for all counts in this paragraph). Political violence, however, remains exceedingly rare in the United States, amounting to 48 incidents (8) in 2019 (the most recent year for which data are available) compared to 4,526 incidents of nonpolitical violent hate crimes (9) and 1,203,808 total violent crimes (10) documented by the Department of Justice.

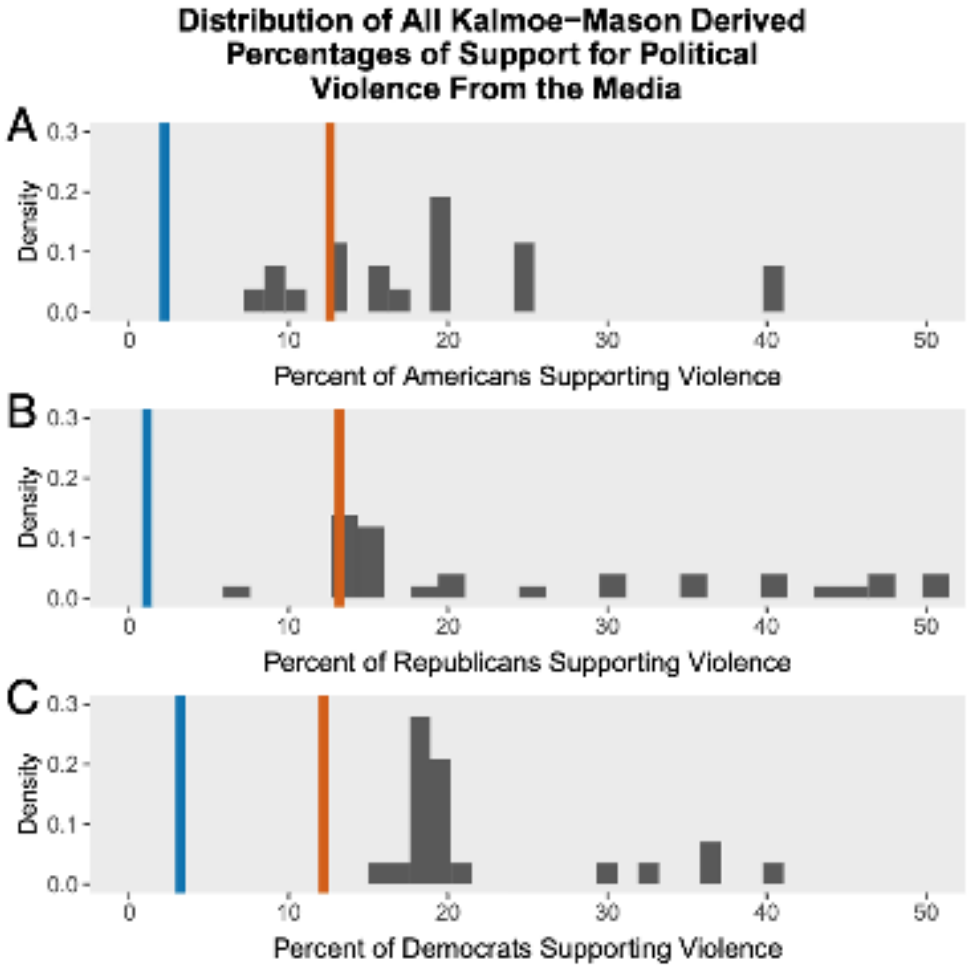


Fig. 1. We created a census of all reported estimates of support for violence using the Kalmoe–Mason measure in the media (this includes work done by authors other than Kalmoe and Mason). This figure shows their distribution. We report this in the full sample (*A*), for Republicans (*B*), and for Democrats (*C*). To contextualize the problems, in these estimates, we overlay the largest estimates (orange line) and smallest estimates (blue line) from the studies that follow. There is large variation in the reported values, but all are significantly larger than our preferred estimate. See *SI Appendix* for additional details.

Ambiguous Questions Create Upward Bias in Estimates of Support for Violence. Even if respondents truthfully report their views on political violence, vague questions make it impossible to compare responses across individuals, and render sample averages uninterpretable. For example, a measure from Kalmoe and Mason (hereafter, Kalmoe–Mason) (2, 3) asks about perceived justification for partisan violence generally: "How much do you feel it is justified for [respondent's own party] to use violence in advancing their political goals these days?" But the estimand measured by this survey item is unclear because it leaves ambiguous what "violence" refers to. Another question from Robert Pape (23), "The use of force is justified to restore Donald Trump to the presidency," offers a specific motivation, but, like the Kalmoe–Mason measures, leaves definition of "violence" to the respondent to fill in. As a simplistic example, suppose that respondents interpret the question as asking about either partisan-motivated assault or partisan-motivated murder (both acts of violence). If one individual interprets violence as "assault" while another interprets violence as "murder," then these responses are not comparable, and therefore we cannot make an inference about which respondent expresses more support for political violence (24). This also affects mean expressed support for violence. The quantity $P(\text{support partisan violence})$ is an average of respondents who interpret the question as asking about assault and others interpreting the question as asking about murder. The conditional average support for partisan violence and the relative prevalence of the components of the mixture are unknown, $P(\text{support partisan violence}) = P(\text{support partisan violence}|\text{assault})P(\text{assault}) + P(\text{support partisan violence}|\text{murder})P(\text{murder})$.

"violence" to the respondent to fill in. As a simplistic example, suppose that respondents interpret the question as asking about either partisan-motivated assault or partisan-motivated murder (both acts of violence). If one individual interprets violence as "assault" while another interprets violence as "murder," then these responses are not comparable, and therefore we cannot make an inference about which respondent expresses more support for political violence (24). This also affects mean expressed support for violence. The quantity $P(\text{support partisan violence})$ is an average of respondents who interpret the question as asking about assault and others interpreting the question as asking about murder. The conditional average support for partisan violence and the relative prevalence of the components of the mixture are unknown, $P(\text{support partisan violence}) = P(\text{support partisan violence}|\text{assault})P(\text{assault}) + P(\text{support partisan violence}|\text{murder})P(\text{murder})$.

Measuring Engagement. To assess how satisficing affects responses, we group participants based on their cognitive engagement with our survey, measured as the retention of information from vignettes. Respondents who can identify the state where the event in the vignette they just read occurred (information that was repeated in the headline and up to two times in the text) are coded as engaged, and those who cannot are coded as disengaged. Reading short passages is not a cognitively intense task, but committing content to memory is (29), and reveals engagement with cognitive work. Across our three studies, we group our sample into "engaged" respondents–those who are sufficiently motivated to carefully read and retain information–and "disengaged" respondents–those who can follow instructions but are less likely to retain facts or carefully evaluate questions.‡

Outcome Questions. We use three core questions for study 1, study 2, and study 3 and deploy two strategies to identify satisficing's upward bias on estimates of support for violence. First, in our justification outcome question, we use a dichotomous scale which should prompt disengaged respondents to randomly respond, which we can detect by observing the response distribution. Second, we offer a balanced scale with a neutral midpoint when measuring support, building on prior work that suggests disengaged respondents will respond with a "don't know" or neutral midpoint response (11). Our three questions are as follows:

1) "Was the suspect justified or unjustified?" using a dichotomous outcome scale. This differs from refs. 1–3, where the authors ask a hypothetical question with an unbalanced five-point Likert scale that is then recoded in their analysis to a dichotomous measure.
2) "Do you support or oppose the actions of [suspect]?" This is measured with a balanced Likert with a neutral midpoint and is intended to separately capture support for a suspect when a violent act is committed.
3) "Should the suspect face criminal charges?" We capture responses with a dichotomous yes/no scale, and the question is intended to determine whether the respondent thinks that someone who commits an act of violence should or should not face charges.

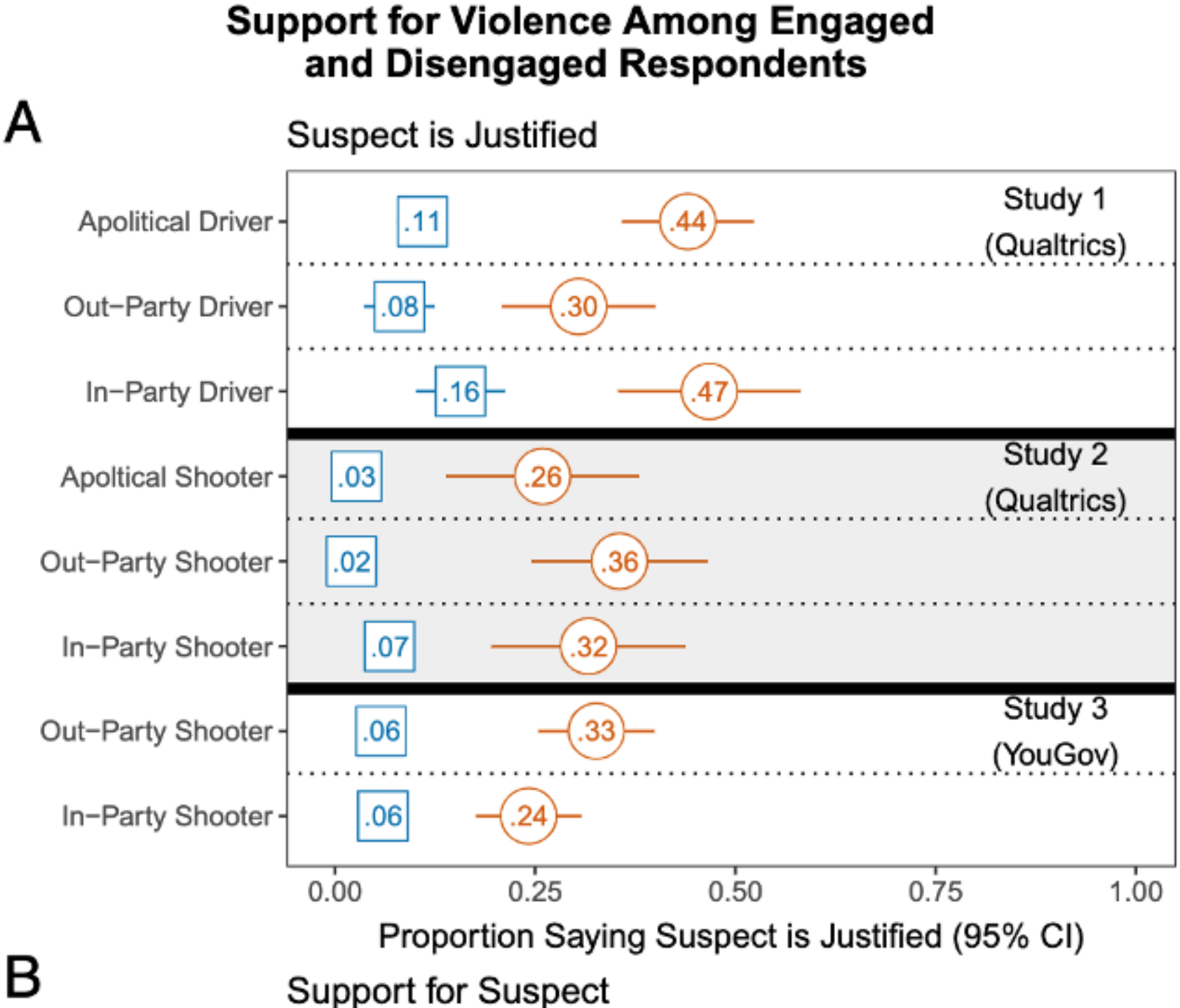
In study 1 (n = 1,002), we randomly assigned participants to read one of two stories based on real acts of political violence. In the first story, a Democratic driver was charged with hitting a group of Republicans in Florida who were registering citizens to vote. In the second story, a Republican driver was charged with assault for driving his car through Democratic protesters in Oregon. Respondents were also randomized to see the original version of the story that included partisan details or a version of the story that was altered to remove any reference to partisan motivation.

In this study, we focused on reporting details from real events. This means that, while comparable, the Democratic and Republican stories varied in several ways. To ensure that any effects we identify are not the result of those differences, we conducted a second version of this experiment. Study 2 (n = 1,023) used a single contrived story of violence in Iowa. To test the bounds of support for political violence, this story reported an extreme form of violence: murder. Similar to study 1, participants were randomly assigned to see a story with a Republican or Democratic shooter engaging in politically motivated violence or an apolitical act of murder. This story was necessarily fabricated to limit the differences across treatment conditions.

Study 3 (n = 1,863) is a replication of study 2 using the YouGov panel with the following alterations: 1) We removed the apolitical condition to focus on attitudes toward partisan violence, and 2) we removed covariate questions to reduce survey time.

Disengaged Responses Lead to Higher Estimates of Support for Political Violence. At first glance, the results of this experiment appear to align with prior surveys. Across conditions where the driver's actions are presented as political violence, we find that 21.07% of respondents in study 1 say the attack was justified. We find a similarly high level of support for the apolitical versions, where 20.56% of respondents in study 1 say the driver's action is justified. The overall support for violence is lower in study 2 and study 3, reflecting the greater severity of the violence, with 10.02% of respondents in study 2 describing the political homicide as justified and 6.70% of the respondents describing the apolitical homicide as justified. In study 3, 10.11% described the homicide as justified.

But this is biased upward by respondents who fail the engagement test (~31% of respondents in study 1, 19% of respondents in study 2, and 19% of weighted respondents in study 3). For the political treatments, 37.87% of respondents who fail the engagement test say the driver's actions were justified, while only 12.06% of respondents who passed the engagement test agree that the driver's actions are justified. For the nonpolitical treatment, we find that 44.06% of respondents who failed the engagement test say the driver's actions were justified, but only 10.93% of respondents who passed the engagement test say the driver's actions are justified. Similarly, for study 2, in the political treatments, we find that 33.82% of the respondents who fail the engagement test say the shooter's actions were justified, but only 4.37% of individuals who passed the engagement test say the action was justified. In the nonpolitical treatments, we find a similar large gap: 25.93% of respondents who fail the engagement test say the action was justified, but 2.68% of those who passed say the action was justified. The same pattern is found in study 3 (YouGov data), with 28.25% of disengaged respondents saying the shooting


Support for Violence Among Engaged and Disengaged Respondents
A
Suspect is Justified
Apolitical Driver
Out−Party Driver
In−Party Driver
Apoltical Shooter
Out−Party Shooter
In−Party Shooter
Out−Party Shooter
In−Party Shooter
.11
.44
.08
.30
.16
.47
.03
.26
.02
.36
.07
.32
.06
.33
.06
.24
Study 1 (Qualtrics)
Study 2 (Qualtrics)
Study 3 (YouGov)
0.00
0.25
0.50
0.75
1.00
Proportion Saying Suspect is Justified (95% CI)
B
Support for Suspect

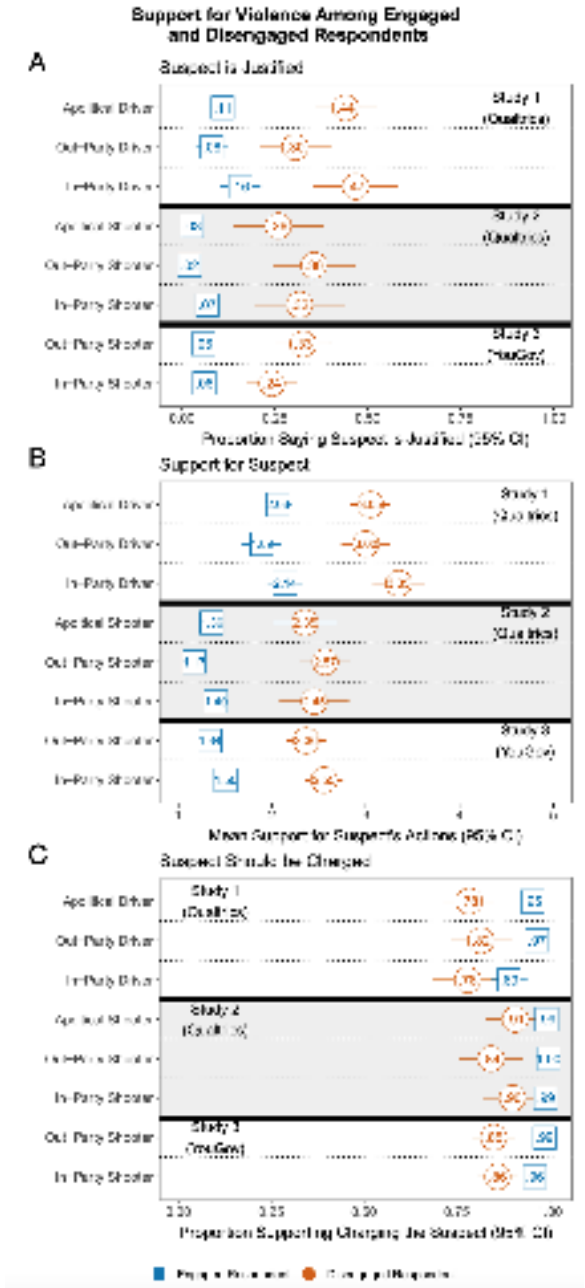
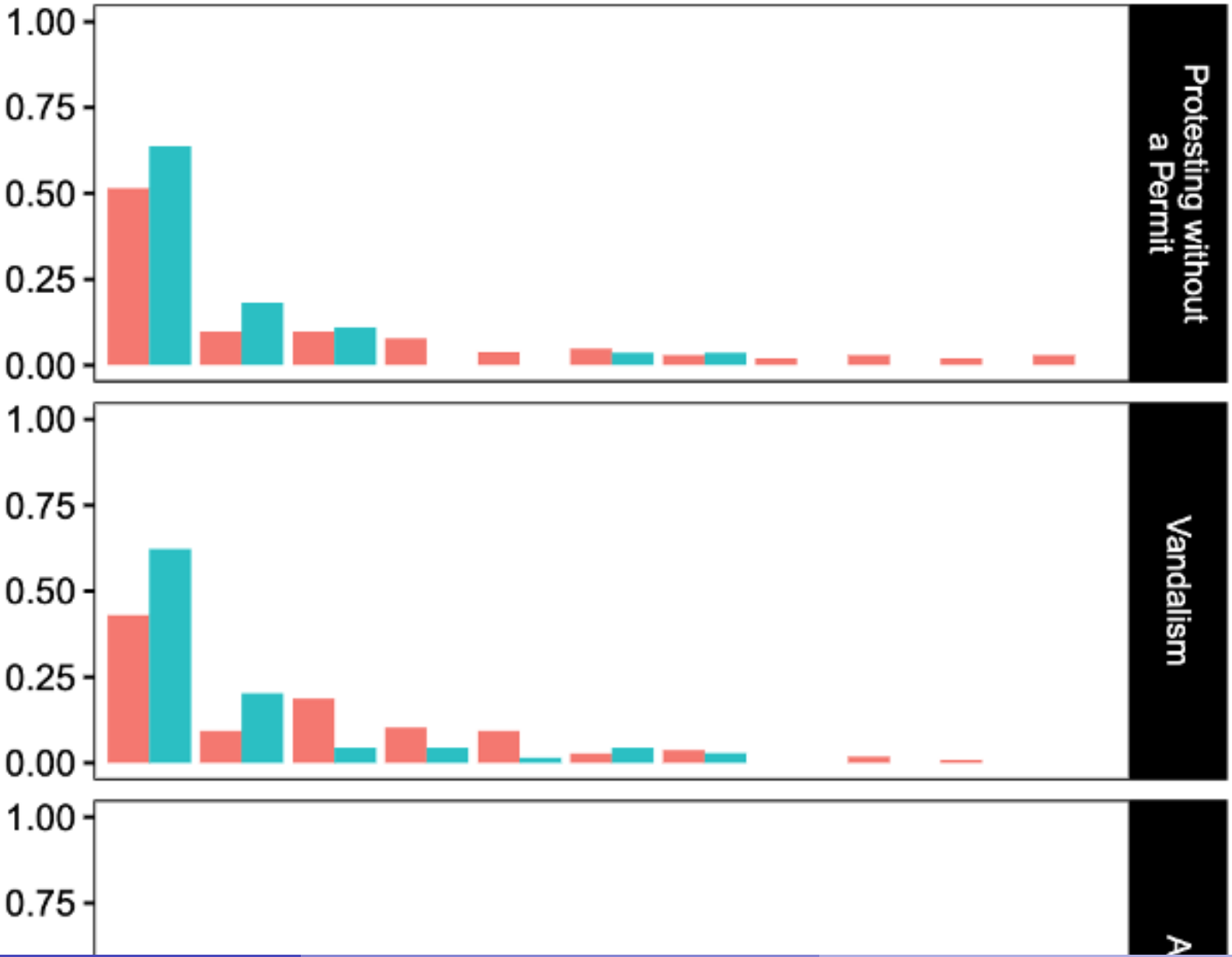
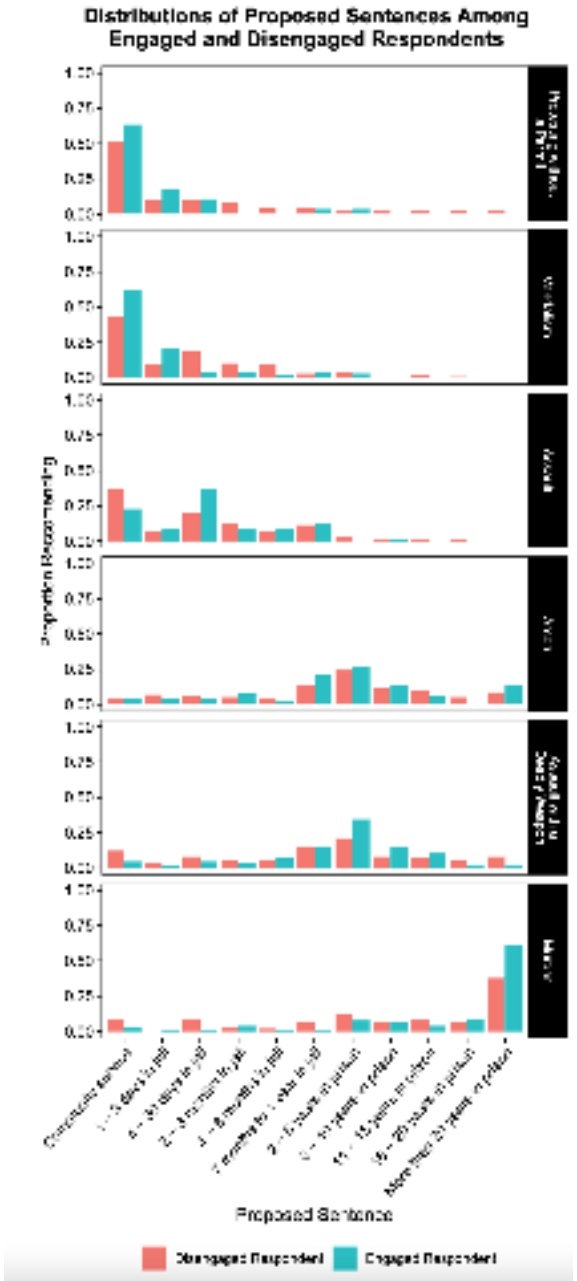

Support for Violence Among Engaged and Disengaged Respondents
A
B
C

Table 1. Kalmoe–Mason support for violence measure by engagement

	Support for violence Kalmoe–Mason measure % (N)		
	Study 1	Study 2	Study 3
Disengaged respondents	55 (312)	43 (190)	41 (354)
Engaged respondents	21 (690)	26 (833)	19 (1,509)
Combined estimate	32 (1,002)	29 (1,023)	23 (1,863)

Distributions of Proposed Sentences Among Engaged and Disengaged Respondents




Distributions of Proposed Sentences Among
Engaged and Disengaged Respondents
Proportion Responding
Proposed Sentence
Disengaged Respondent
Engaged Respondent

Linking Data

Linking Data

- Data often becomes exponentially more useful when it is combined with other data sources.

Linking Data

- Data often becomes exponentially more useful when it is combined with other data sources.
- Administrative, survey, and commercial data can all be linked when there are appropriate identifiers. We've already seen several projects that depend on this approach!

Linking Data

- Data often becomes exponentially more useful when it is combined with other data sources.
- Administrative, survey, and commercial data can all be linked when there are appropriate identifiers. We've already seen several projects that depend on this approach!
- Because the datasets come from different sources and thus information is not standardized across them!

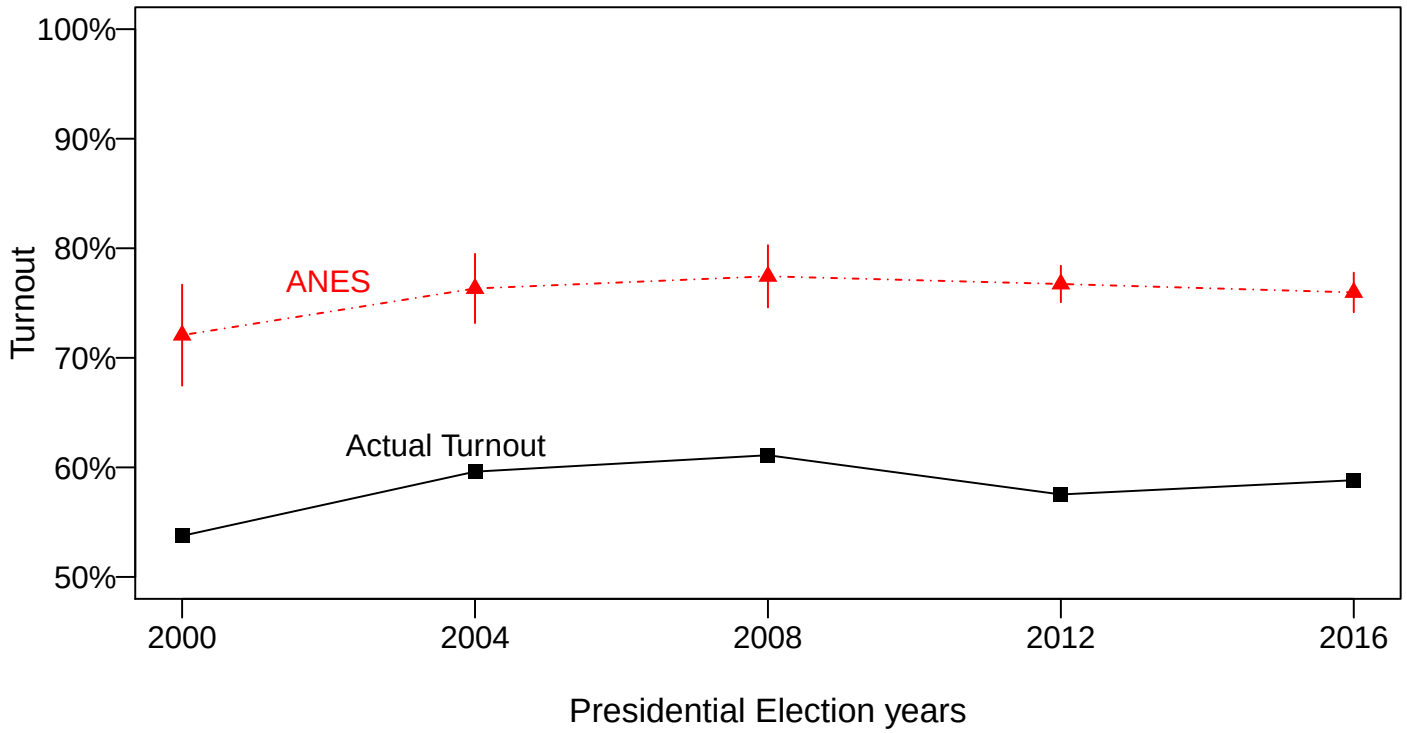
Linking Data

- Data often becomes exponentially more useful when it is combined with other data sources.
- Administrative, survey, and commercial data can all be linked when there are appropriate identifiers. We've already seen several projects that depend on this approach!
- Because the datasets come from different sources and thus information is not standardized across them!
- First-Last name combinations might not be identifying on their own and that's before even considering typos, name changes etc.

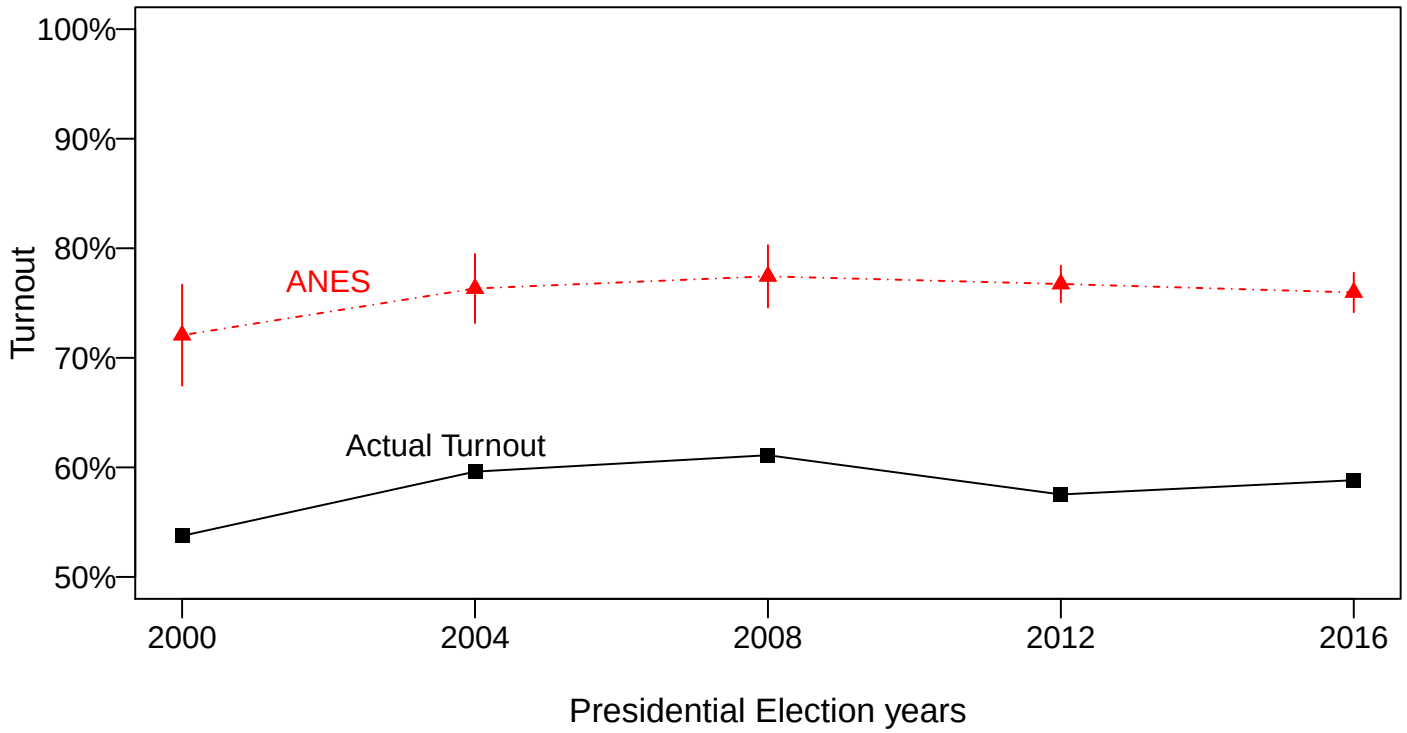
Linking Data

- Data often becomes exponentially more useful when it is combined with other data sources.
- Administrative, survey, and commercial data can all be linked when there are appropriate identifiers. We've already seen several projects that depend on this approach!
- Because the datasets come from different sources and thus information is not standardized across them!
- First-Last name combinations might not be identifying on their own and that's before even considering typos, name changes etc.
- Example: there are dozens of 'Brandon Stewart's in the FEC individual contributions data! There are even other "Brandon Stewart" professors (one in psychology and one in political science)

Linking Data Helps us Solve Problems!



Linking Data Helps us Solve Problems!



Why does this happen?

Over-reporting or Sampling issues?

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  - ★ controls false positives

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

- **Probabilistic Record Linkage (PRL)**

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

- **Probabilistic Record Linkage (PRL)**
  ★ robust to typographical errors

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

- **Probabilistic Record Linkage (PRL)**
  ★ robust to typographical errors
  ★ designed to control error rates

How We Merge Data Sets Matters

- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

- **Probabilistic Record Linkage (PRL)**
  ★ robust to typographical errors
  ★ designed to control error rates

  $\rightsquigarrow$ existing open-source implementations of PRL do not scale

How We Merge Data Sets Matters

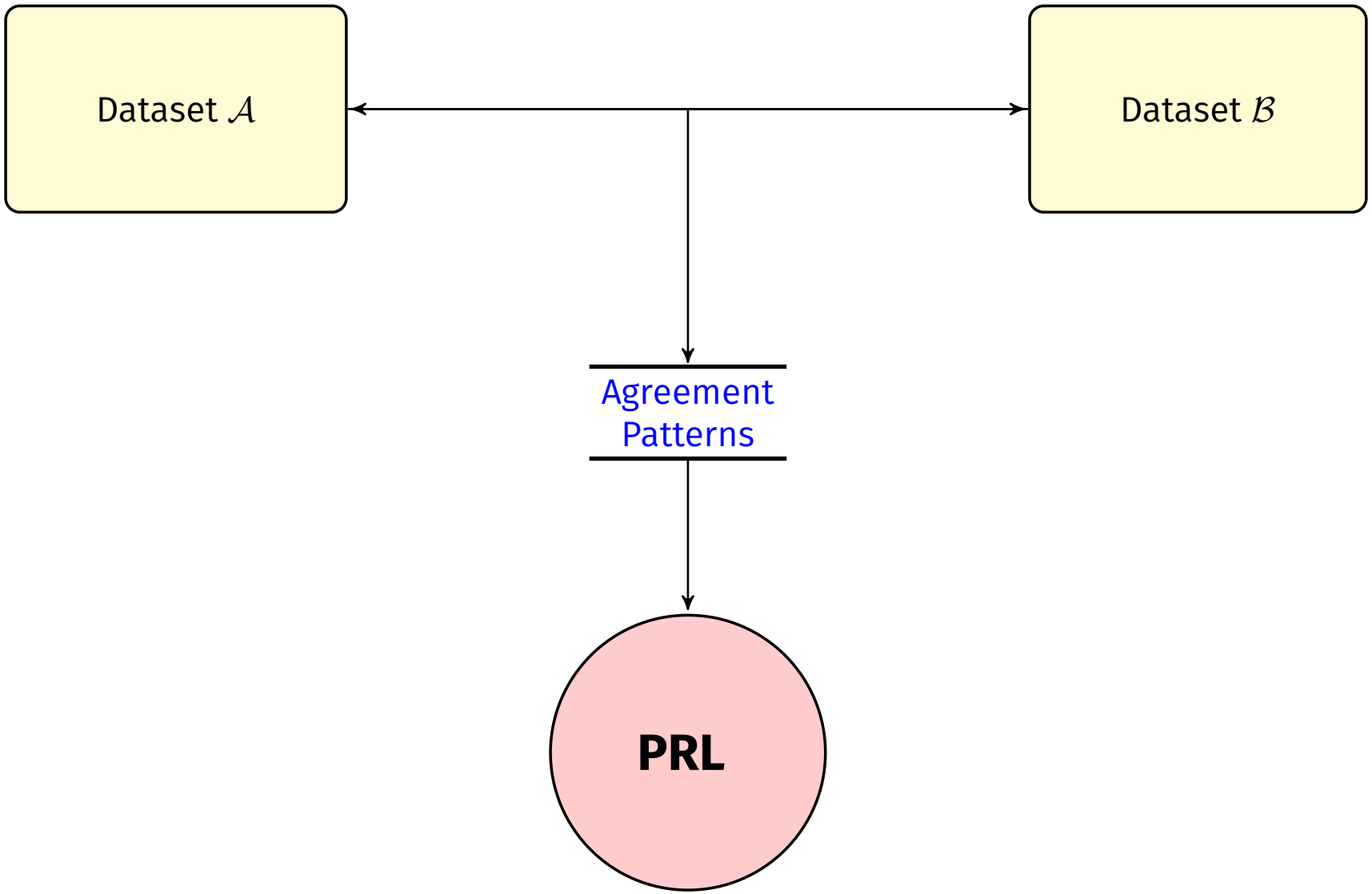
- **Deterministic methods** e.g., exact matches
  ★ controls false positives
  $\rightsquigarrow$ not robust to **typographical errors** and **missing values**

  Georgia's "Exact Match" voter identification law $\rightsquigarrow$
  In 2018, more than **53,000** records failed to match exactly and were put on hold $\rightsquigarrow$ **80%** of which are racial minorities

- **Probabilistic Record Linkage (PRL)**
  ★ robust to typographical errors
  ★ designed to control error rates

  $\rightsquigarrow$ existing open-source implementations of PRL do not scale
  Solution: `fastLink` (Enamorado, Fifield, and Imai, 2019)

The Workflow for PRL à la Fellegi-Sunter



Agreement Patterns

- Two data sets: $\mathcal{A}$ and $\mathcal{B}$
- K variables in common
- $N_{\mathcal{A}} \times N_{\mathcal{B}}$ pairs to be compared for each common variable k

Agreement Patterns

- Two data sets: $\mathcal{A}$ and $\mathcal{B}$
- K variables in common
- $N_{\mathcal{A}} \times N_{\mathcal{B}}$ pairs to be compared for each common variable k
- Agreement value in field k for a pair (i, j)

$$\gamma_k(i,j) = \begin{cases} \texttt{agree} \\ \texttt{similar} \\ \texttt{disagree} \end{cases}$$

Agreement Patterns

- Two data sets: $\mathcal{A}$ and $\mathcal{B}$
- K variables in common
- $N_{\mathcal{A}} \times N_{\mathcal{B}}$ pairs to be compared for each common variable k
- Agreement value in field k for a pair (i, j)

$$\gamma_k(i,j) = \begin{cases} \texttt{agree} \\ \texttt{similar} \\ \texttt{disagree} \end{cases}$$

Agreement pattern $\boldsymbol{\gamma}(i,j) = \{\gamma_1(i,j), \gamma_2(i,j), \ldots, \gamma_K(i,j)\}$

How to Build an Agreement Pattern

	Name			
	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	Devereux St.
2	John	Martin	56	Devereux St.
Data set $\mathcal{B}$				
1	Michael	Martinez	28	16th St.
2	James	Smitj	34	Dvereuux St.

How to Build an Agreement Pattern

	Name			
	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	Devereux St.
2	John	Martin	56	Devereux St.
Data set $\mathcal{B}$				
1	Michael	Martinez	28	16th St.
2	James	Smitj	34	Dvereuux St.

How to Build an Agreement Pattern

	Name			
	First	Last	Age	Street
$1 \in \mathcal{A}$	James	Smith	35	Devereux St.
$2 \in \mathcal{B}$	James	Smitj	34	Dvereuux St.

How to Build an Agreement Pattern

Step 1: Calculate a Measure of Dissimilarity

	Name			
	First	Last	Age	Street
$1 \in \mathcal{A}$	James	Smith	35	Devereux St.
$2 \in \mathcal{B}$	James	Smitj	34	Dvereuux St.
dissimilarity	0.00	0.04		0.10

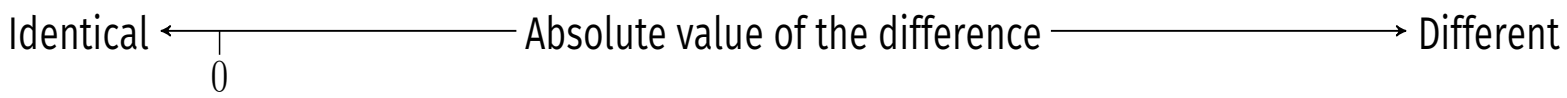
- For fields such as first name, last name, street name we use:

How to Build an Agreement Pattern

Step 1: Calculate a Measure of Dissimilarity

	Name			
	First	Last	Age	Street
$1 \in \mathcal{A}$	James	Smith	35	Devereux St.
$2 \in \mathcal{B}$	James	Smitj	34	Dvereuux St.
dissimilarity			1.00	

- For fields such as age:

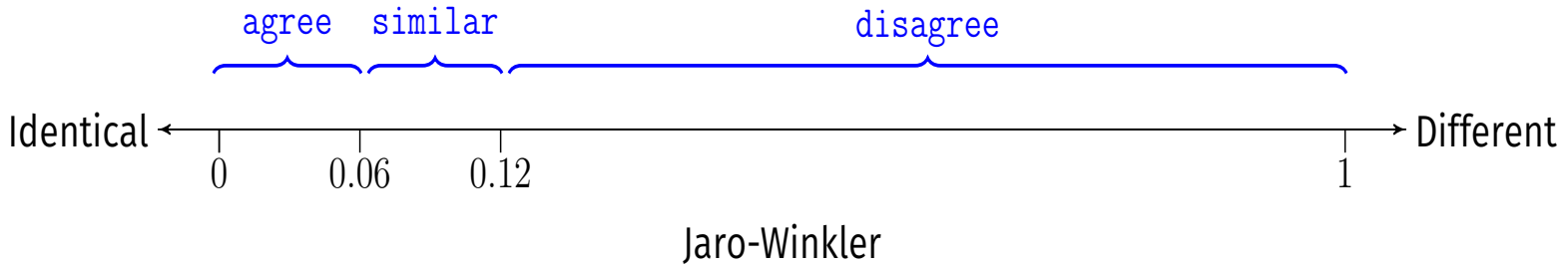


How to Build an Agreement Pattern

Step 2: Make Comparisons Discrete

	Name			
	First	Last	Age	Street
$1 \in \mathcal{A}$	James	Smith	35	Devereux St.
$2 \in \mathcal{B}$	James	Smitj	34	Dvereuux St.
dissimilarity	0.00	0.04		0.10
$\gamma(1,2)$	agree	agree		similar

- Set thresholds for each dissimilarity measure:

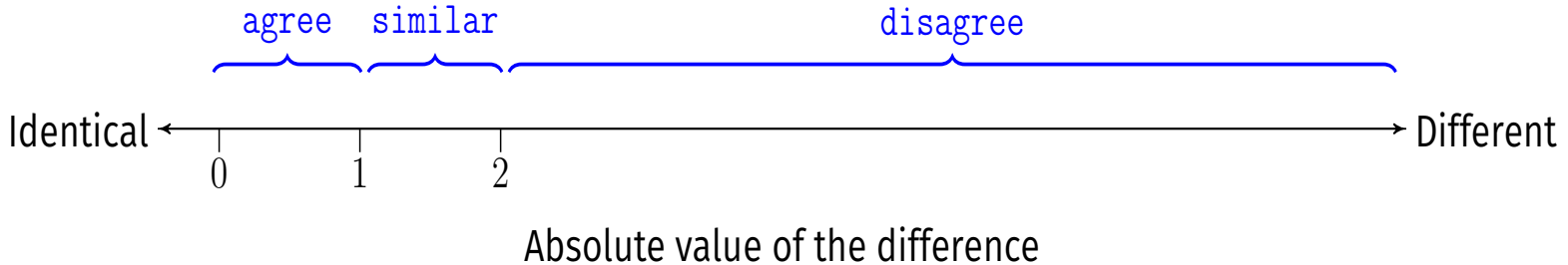


How to Build an Agreement Pattern

Step 2: Make Comparisons Discrete

	Name			
	First	Last	Age	Street
$1 \in \mathcal{A}$	James	Smith	35	Devereux St.
$2 \in \mathcal{B}$	James	Smitj	34	Dvereuux St.
dissimilarity	0.00	0.04	1.00	0.10
$\gamma(1,2)$	agree	agree	agree	similar

- Set thresholds for each dissimilarity measure:



How to Build an Agreement Pattern

	Name			
	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	Devereux St.
2	John	Martin	56	Devereux St.
Data set $\mathcal{B}$				
1	Michael	Martinez	28	16th St.
2	James	Smitj	34	Dvereuux St.
Agreement patterns $\gamma(i,j)$				
$\gamma(1,2)$	agree	agree	agree	similar

How to Build an Agreement Pattern

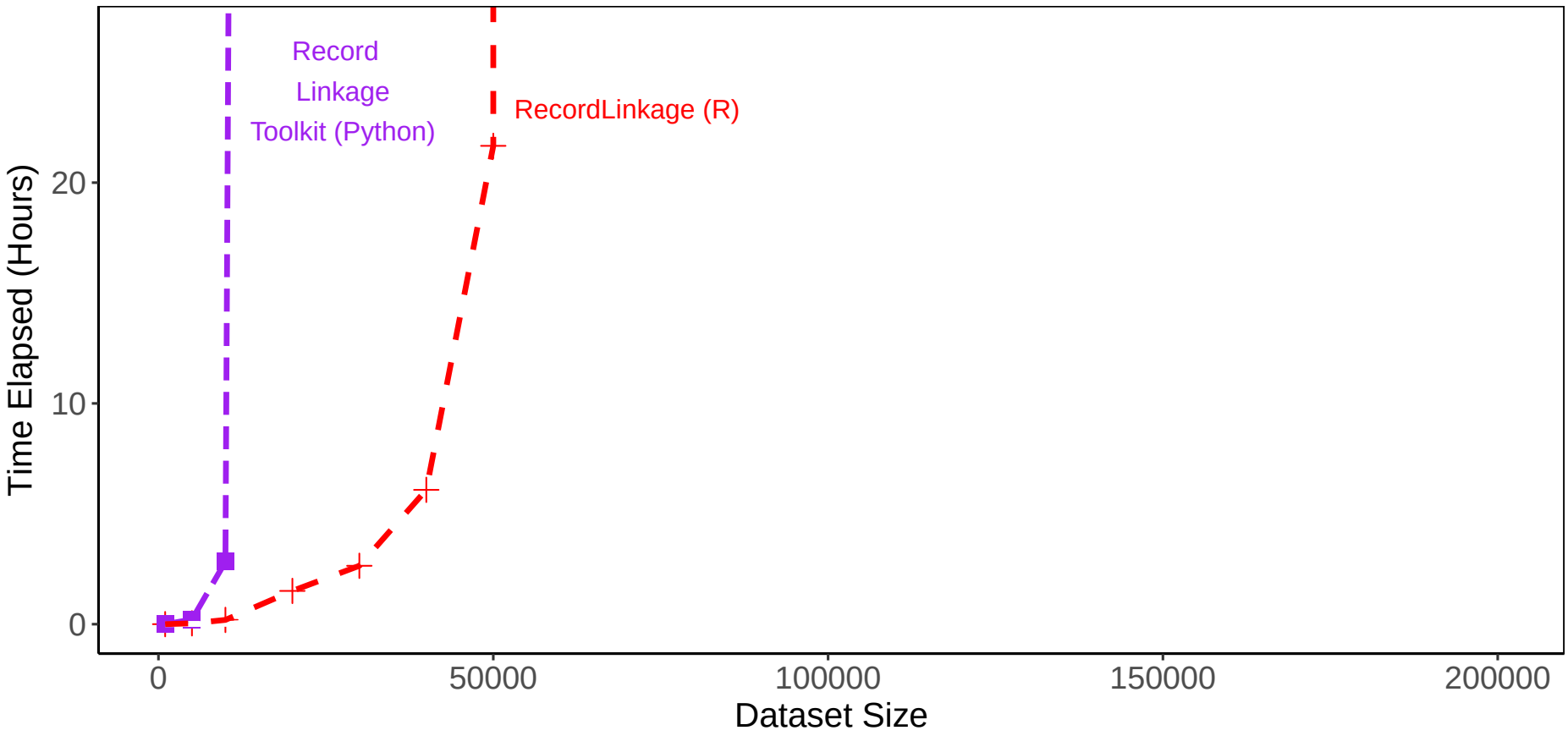
	Name			
	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	Devereux St.
2	John	Martin	56	Devereux St.
Data set $\mathcal{B}$				
1	Michael	Martinez	28	16th St.
2	James	Smitj	34	Dvereuux St.
Agreement patterns $\gamma(i,j)$				
$\gamma(1,1)$	disagree	disagree	disagree	disagree
$\gamma(1,2)$	agree	agree	agree	similar
$\gamma(2,1)$	disagree	similar	disagree	disagree
$\gamma(2,2)$	disagree	disagree	disagree	similar

How to Build an Agreement Pattern

	Name			
	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	Devereux St.
2	John	Martin	56	Devereux St.
Data set $\mathcal{B}$				
1	Michael	Martinez	28	16th St.
2	James	Smitj	34	Dvereuux St.
Agreement patterns $\gamma(i,j)$				
$\gamma(1,1)$	disagree	disagree	disagree	disagree
$\gamma(1,2)$	agree	agree	agree	similar
$\gamma(2,1)$	disagree	similar	disagree	disagree
$\gamma(2,2)$	disagree	disagree	disagree	similar

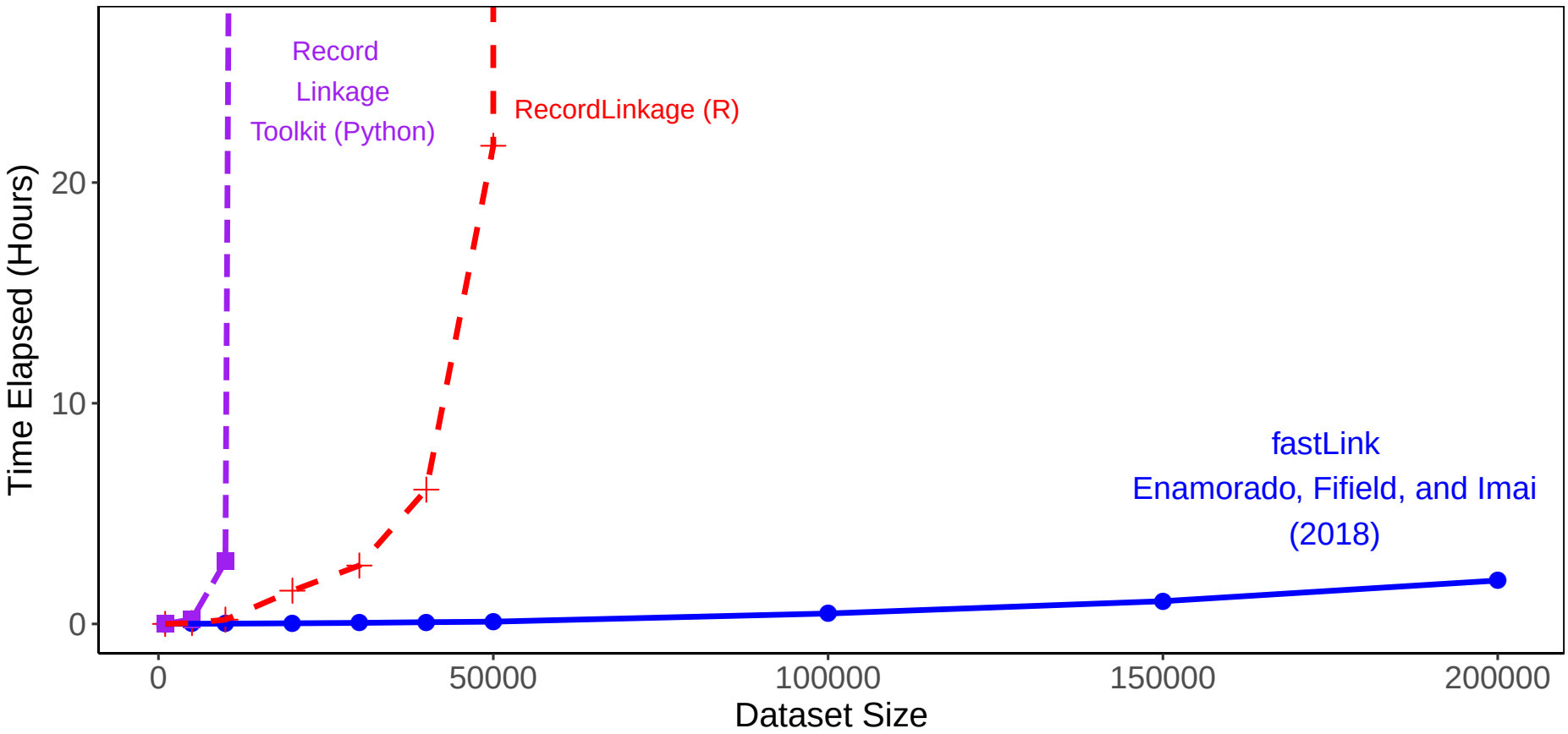
The computational bottleneck: make all these comparisons!

Runtime Comparison



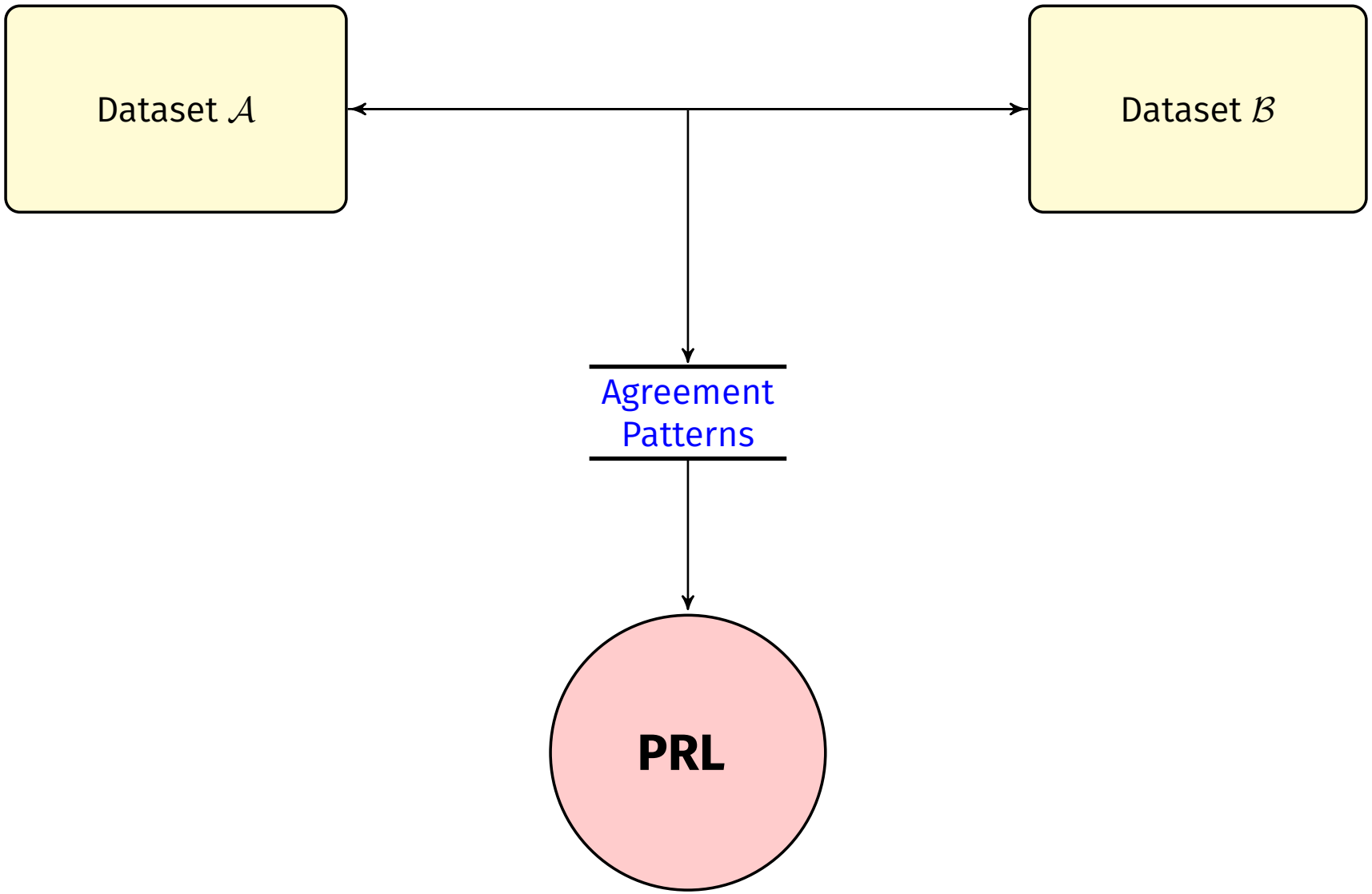
- Data sets of equal size
- Variables in common: first and last name; house number, street name and zip code; age

Runtime Comparison



- Data sets of equal size
- Variables in common: first and last name; house number, street name and zip code; age

The Workflow for PRL à la Fellegi-Sunter



The Fellegi-Sunter Model of PRL

- **We observe:** the agreement patterns $\gamma(i, j)$

The Fellegi-Sunter Model of PRL

- **We observe:** the agreement patterns $\gamma(i,j)$
- **We do not observe:** the matching status

$$M(i,j) = \begin{cases} \text{non-match} \\ \text{match} \end{cases}$$

The Fellegi-Sunter Model of PRL

- **We observe:** the agreement patterns $\gamma(i,j)$
- **We do not observe:** the matching status

$$M(i,j) = \begin{cases} \text{non-match} \\ \text{match} \end{cases}$$

Mixture Model

$$\begin{aligned} M(i,j) &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda) \\ \gamma(i,j) \mid M(i,j) = \text{non-match} &\overset{\text{i.i.d.}}{\sim} \mathcal{F}(\pi_{\text{NM}}) \\ \gamma(i,j) \mid M(i,j) = \text{match} &\overset{\text{i.i.d.}}{\sim} \mathcal{F}(\pi_{\text{M}}) \end{aligned}$$

Result: probabilities of a match between any pair
(can threshold and/or manually inspect uncertain cases)

Ways to Use Record Linkage

Ways to Use Record Linkage

1) Link datasets together (voter files, surveys, campaign contribution data, arrest data, tax records)

Ways to Use Record Linkage

1) Link datasets together (voter files, surveys, campaign contribution data, arrest data, tax records)
2) Deduplicate records (match it to itself!)

Ways to Use Record Linkage

1) Link datasets together (voter files, surveys, campaign contribution data, arrest data, tax records)
2) Deduplicate records (match it to itself!)

- Ways to speed up the package through careful use of blocking.

Ways to Use Record Linkage

1) Link datasets together (voter files, surveys, campaign contribution data, arrest data, tax records)
2) Deduplicate records (match it to itself!)

- Ways to speed up the package through careful use of blocking.
- They are always working on more improvements!

Summary

Summary

- record linkage!
- `fastLink` in `R`

Summary

- record linkage!
- `fastLink` in `R`

Going Deeper:

> Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records" *American Political Science Review*, 2019.

This Week in Review

This Week in Review

- Designed data
- Total survey error framework
- Integrating surveys with other data
- Great additional material in *Bit by Bit*:
  - New ways of asking questions
  - More on enriched and amplified asking

This Week in Review

- Designed data
- Total survey error framework
- Integrating surveys with other data
- Great additional material in *Bit by Bit*:
  - New ways of asking questions
  - More on enriched and amplified asking

Next week: Found Data

- Salganik Chapter 2: "Observing Behavior"
- Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3), 619-637.
- Voigt, Rob, et al. "Language from police body camera footage shows racial disparities in officer respect." Proceedings of the National Academy of Sciences 114.25 (2017): 6521- 6526.