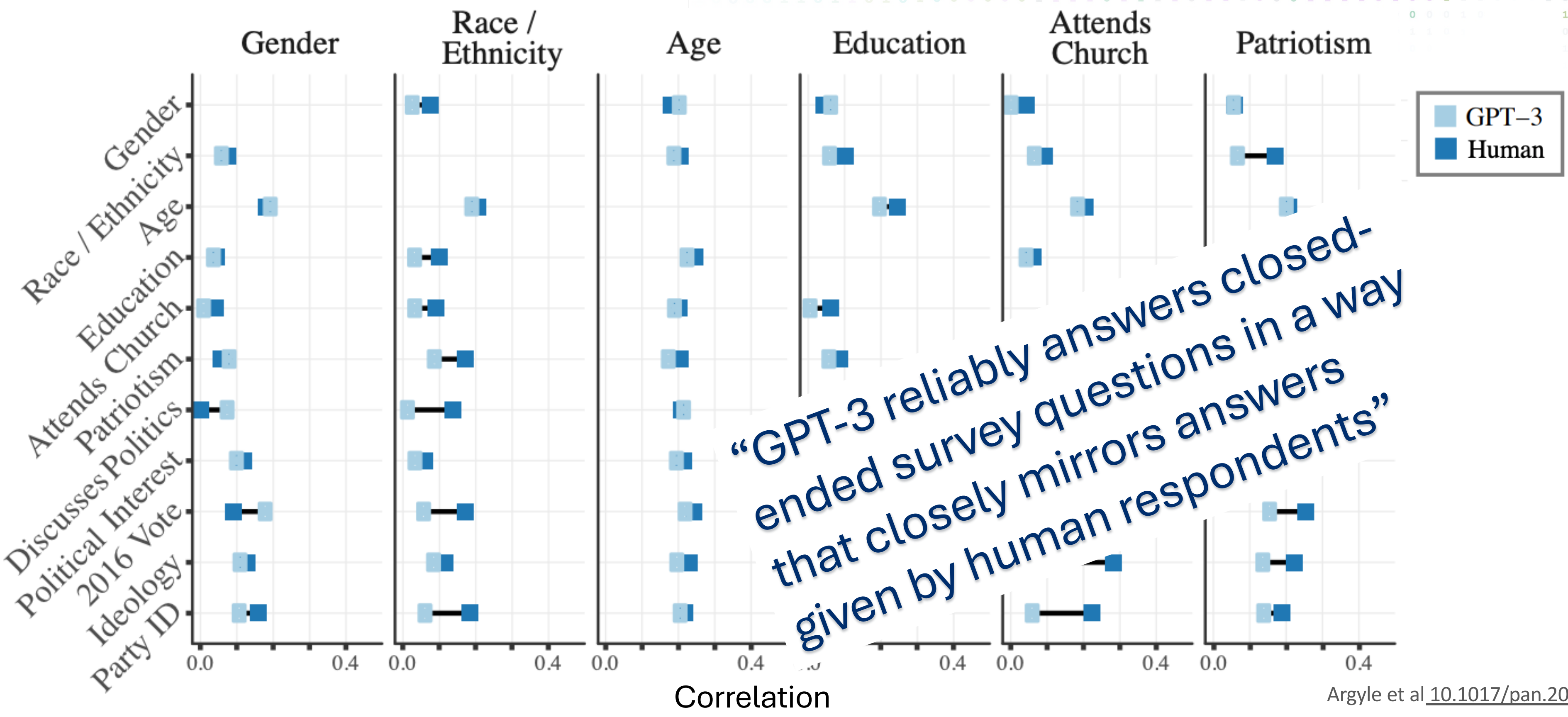
Synthetic Data for Social Science

Stephanie Eckman

Researcher & Data Scientist

University of Maryland

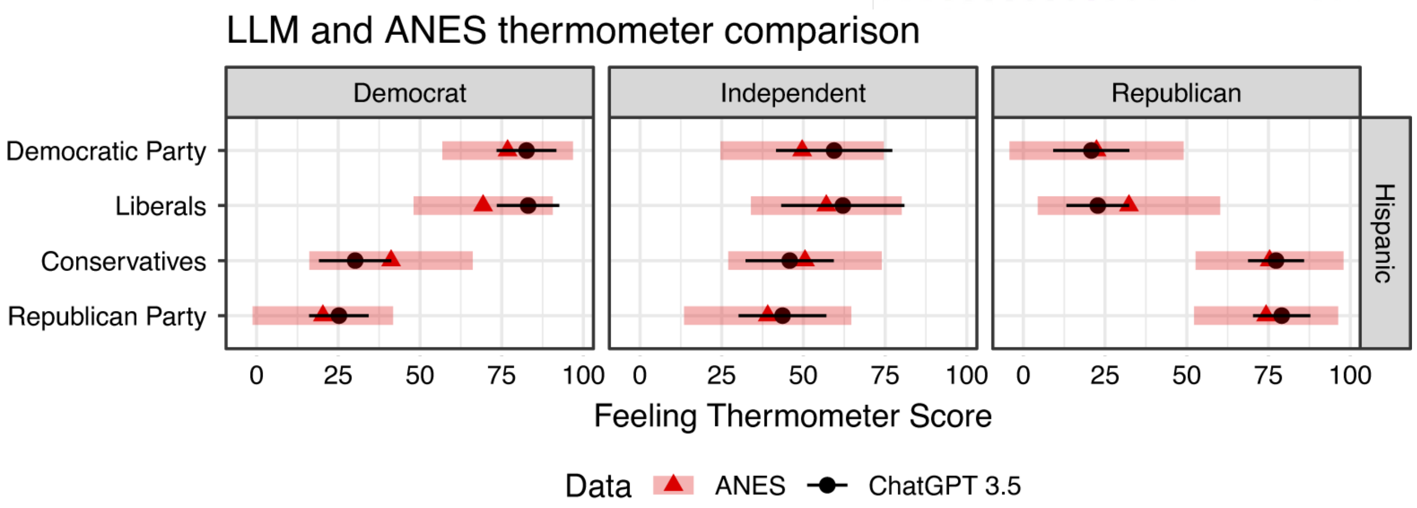
Promise of Synthetic Data



"GPT-3 reliably answers closed-ended survey questions in a way that closely mirrors answers given by human respondents"

Argyle et al 10.1017/pan.2023.2

Or Maybe Not?



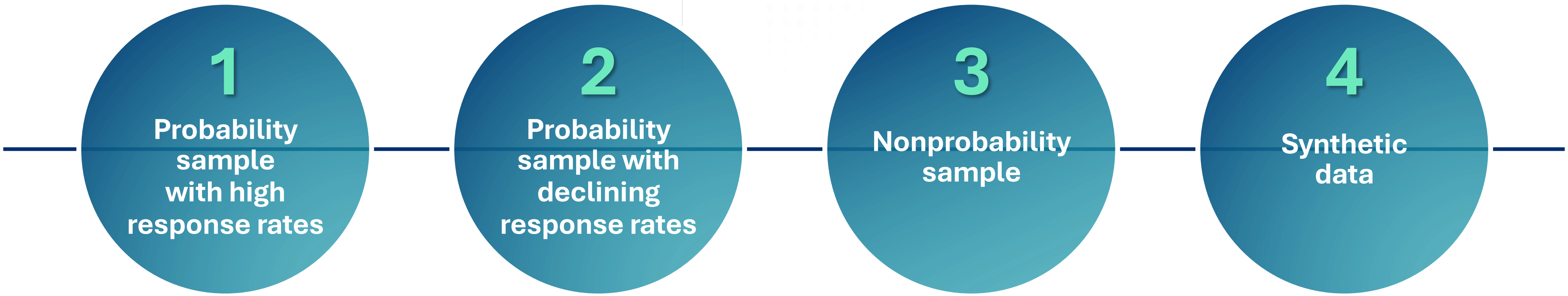
"Every synthetic mean falls within one standard deviation of the ANES average.

…

The distribution of synthetic responses for some questions exhibits far less variation than human responses"

Bisbee et al 10.1017/pan.2024.5

Survey Timeline



1 Probability sample with high response rates

2 Probability sample with declining response rates

3 Nonprobability sample

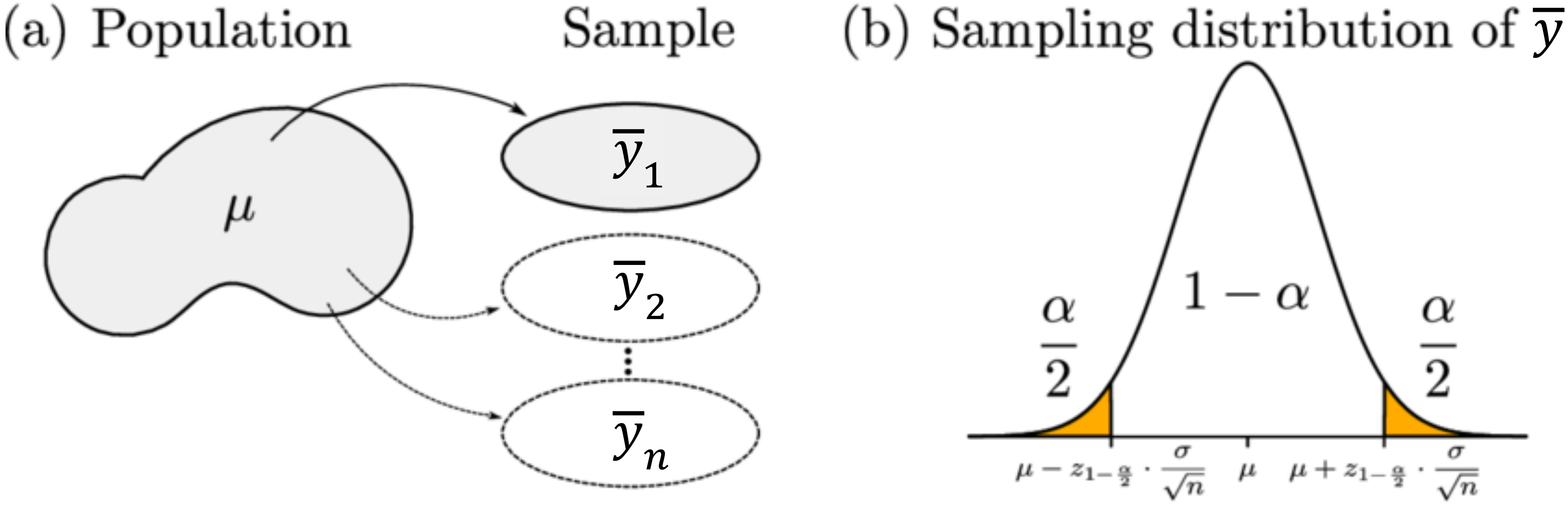
4 Synthetic data

Survey Timeline

1

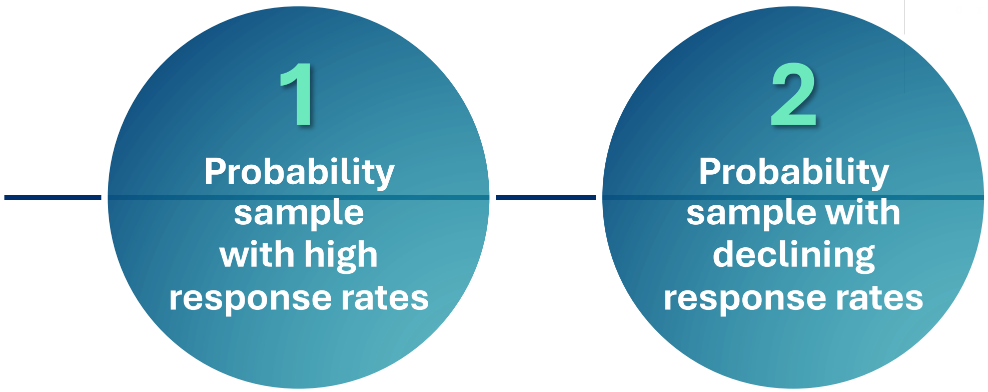
Probability sample with high response rates

Probability Sample with High RRs


(a) Population
Sample
(b) Sampling distribution of $\overline{y}$
μ
$\overline{y}_1$
$\overline{y}_2$
$\vdots$
$\overline{y}_n$
$\frac{\alpha}{2}$
$1-\alpha$
$\frac{\alpha}{2}$
$\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
μ
$\mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Adapted from Berrar 10.1016/B978-0-12-809633-8.20350-6

Survey Timeline



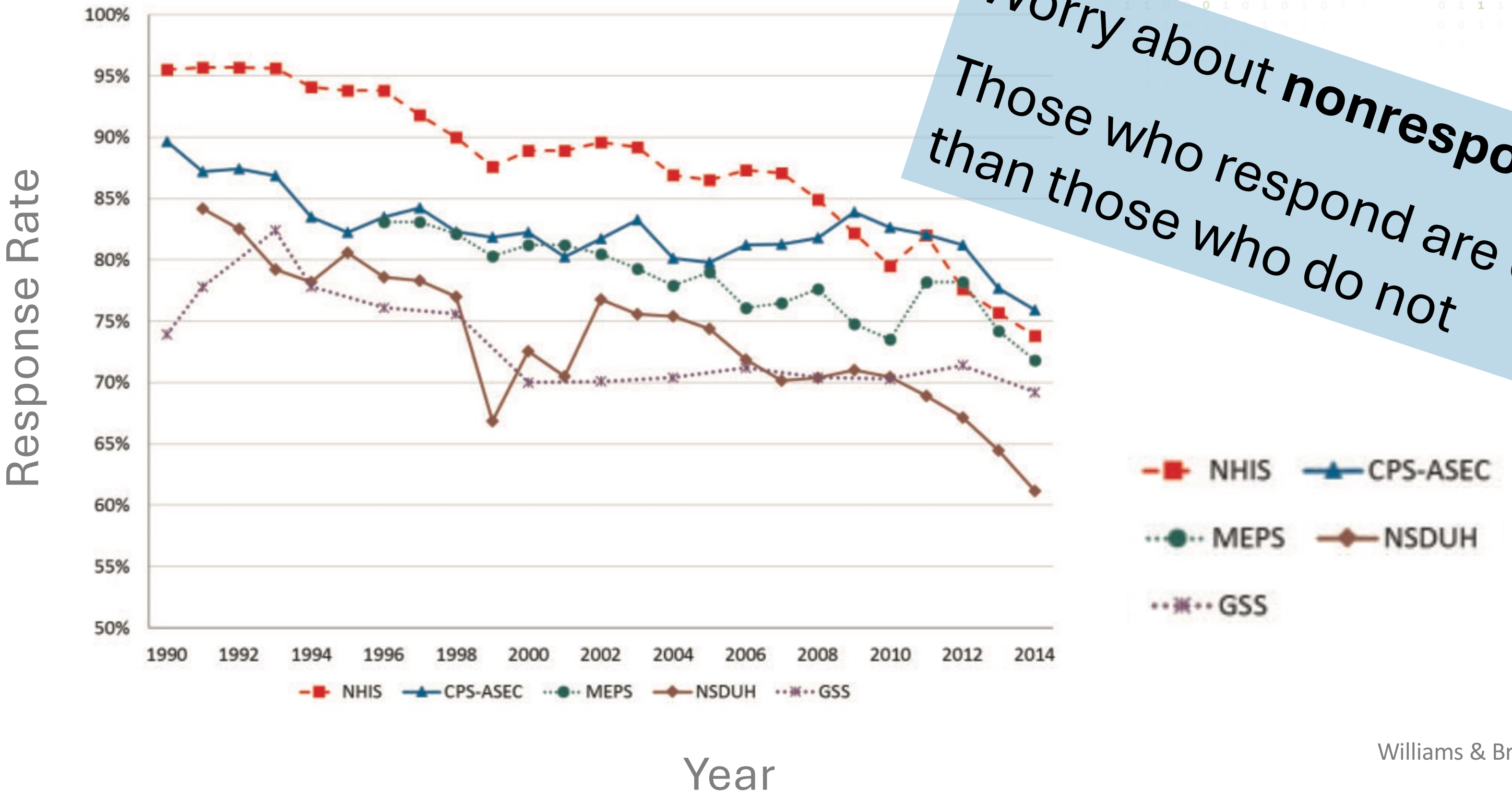
1

Probability sample with high response rates

2

Probability sample with declining response rates

Probability Sample with Declining RRs



Worry about **nonresponse bias:**
Those who respond are different than those who do not

Williams & Brick 10.1093/jssam/smx019

Weights to Reduce Nonresponse Bias

- Adjust by RR^{-1} within cells

- Propensity score models
  - Fit: $\Pr(R = 1) = logit^{-1}(X)$
  - Predict probabilities $\hat{r}$
  - Adjust by $\hat{r}^{-1}$

- **Characteristics of nonrespondents** required

	18-44	45+

Hispanic	RR = 35%
	wt = 2.9
		RR = 43%
		wt = 2.3

Non-Hisp. Black	RR = 28%
	wt = 3.6
		RR = 50%
		wt = 2.0

Non-Hisp. White	RR = 40%
	wt = 2.5
		RR = 44%
		wt = 2.3

Weights to Reduce Nonresponse Bias

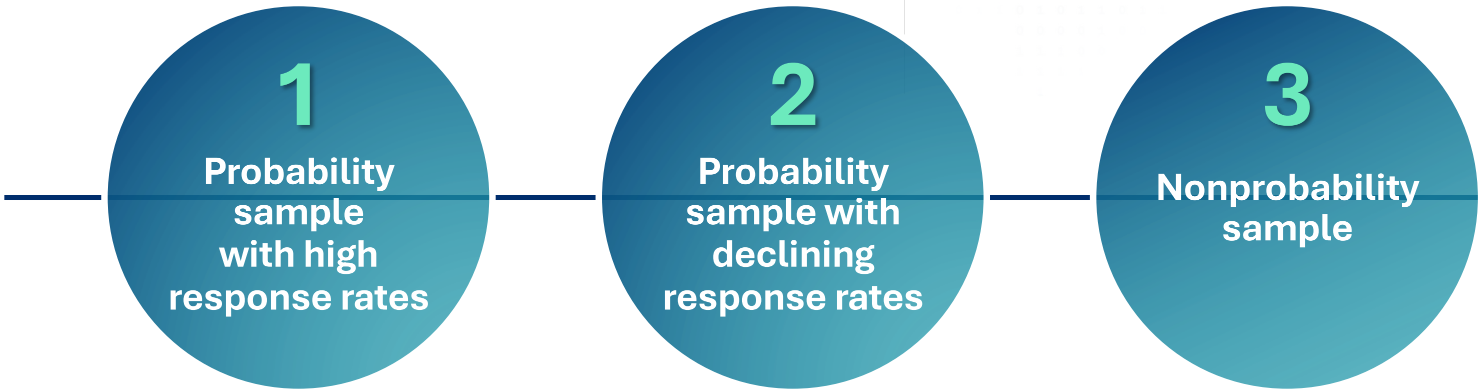
- Solve for weights that make respondents look like pop.
- **Population proportions** required

	18-44	45+	Pop.
Hispanic	$w_1 \times p_1$	$w_2 \times p_2$	30%
Non-Hisp. Black	$w_3 \times p_3$	$w_4 \times p_4$	45%
Non-Hisp. White	$w_5 \times p_5$	$w_6 \times p_6$	25%
Pop.	48%	52%	

Weights to Reduce Nonresponse Bias

- Method less important than variables (X)
- To reduce bias, we want
  - High correlation of X and Y – outcome
  - High correlation of X and R – response
- Assumptions
  - High quality population data available, on X variables
  - Given X, response is random: E(Y|X, R) = E(Y|X)

Survey Timeline



1

Probability sample with high response rates

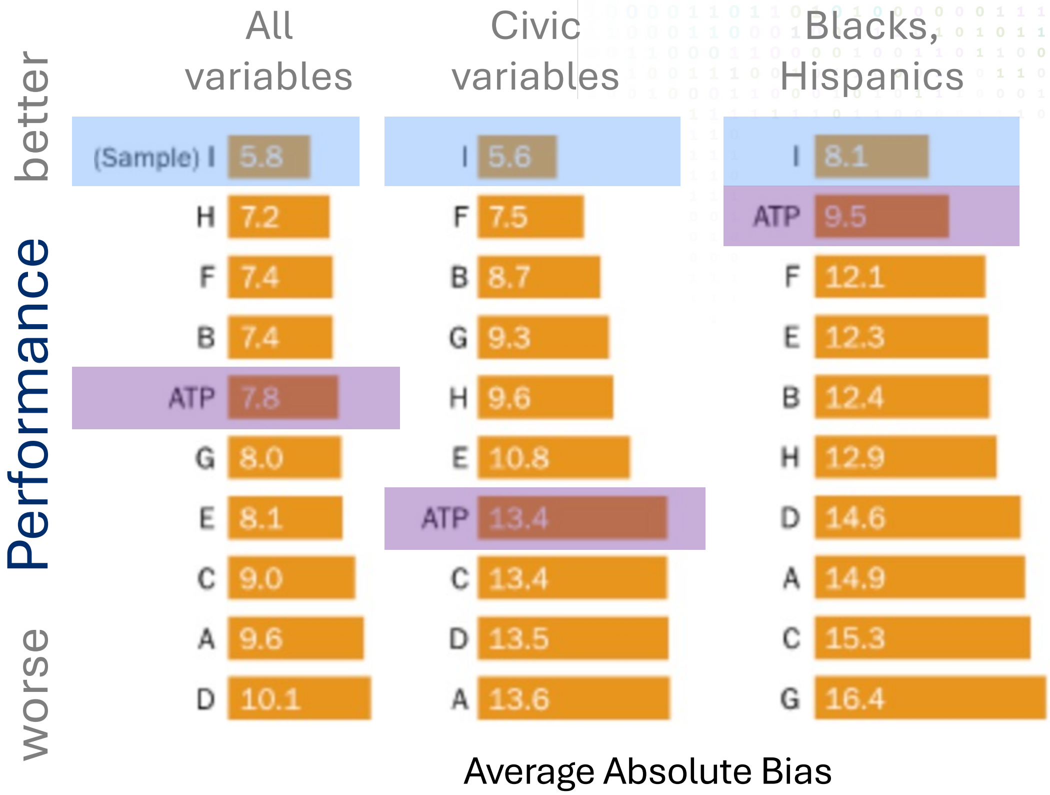
2

Probability sample with declining response rates

3

Nonprobability sample

Bias in Non-Probability Surveys



https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/

Weighting: MRP

- Step 1: model response (Y) with Xs

$$Pr(y_i = 1) = logit^{-1}(\alpha^{\mathrm{state}}_{\mathrm{s[i]}} + \alpha^{\mathrm{age}}_{\mathrm{a[i]}} + \alpha^{\mathrm{eth}}_{\mathrm{r[i]}} + \alpha^{\mathrm{educ}}_{\mathrm{e[i]}} + \beta^{\mathrm{male}} \cdot \mathrm{Male_i} + \alpha^{\mathrm{male.eth}}_{\mathrm{g[i],r[i]}} + \alpha^{\mathrm{educ.age}}_{\mathrm{e[i],a[i]}} + \alpha^{\mathrm{educ.eth}}_{\mathrm{e[i],r[i]}})$$

Predict proportions in each cell θ_j

- Step 2: Weight modeled proportions by size of cell in population

$$\theta^{MRP} = \frac{\sum N_j \theta_j}{\sum N_j}$$

Predicted 2012 election outcome from skewed sample of Xbox users: Wang et al 10.1016/j.ijforecast.2014.06.001

Multilevel Regression and Poststratification Case Studies, Juan Lopez-Martin, Justin H. Phillips, and Andrew Gelman

Weighting: Entropy Balancing

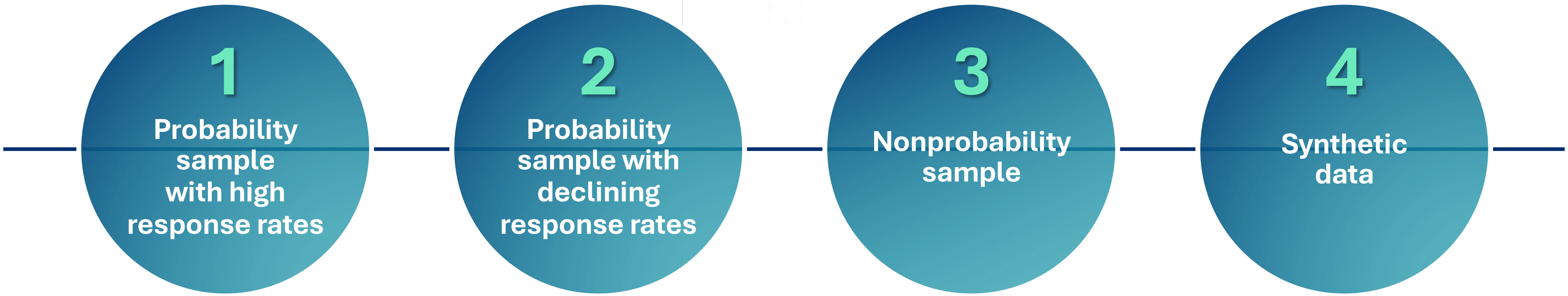
- Solve for weights w_i that:
  - Minimize entropy distance of weights from constant: $\sum w_i \times log(\frac{w_i}{k})$
  - Subject to constraint: $w_i \times x_i = \bar{X}_{pop} \pm \epsilon$
- Can also include constraints on:
  - $Var(X)$
  - $Cov(X_j, X_{j'})$

Hainmueller 10.1093/pan/mpr025

Weighting for Opt-in Samples

- To reduce bias, we want
  - High correlation of X and Y – outcome
  - High correlation of X and R – opting in to nonprob. sample
- Assumptions
  - High quality population data available for X variables
  - Given X, response is random: E(Y|X, R) = E(Y|X)

Survey Timeline



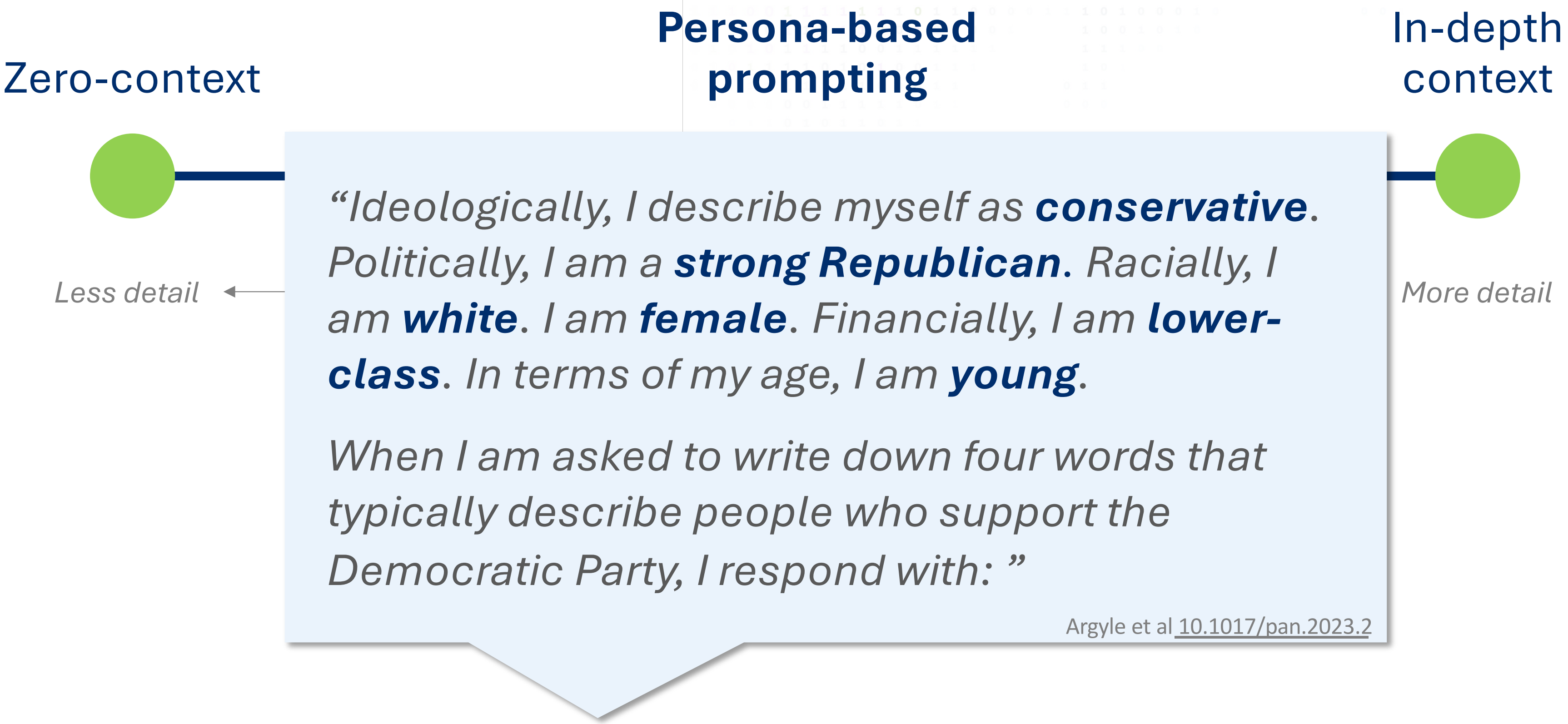
1 Probability sample with high response rates

2 Probability sample with declining response rates

3 Nonprobability sample

4 Synthetic data

How to Generate Synthetic Data



Zero-context

Persona-based prompting

In-depth context

Less detail

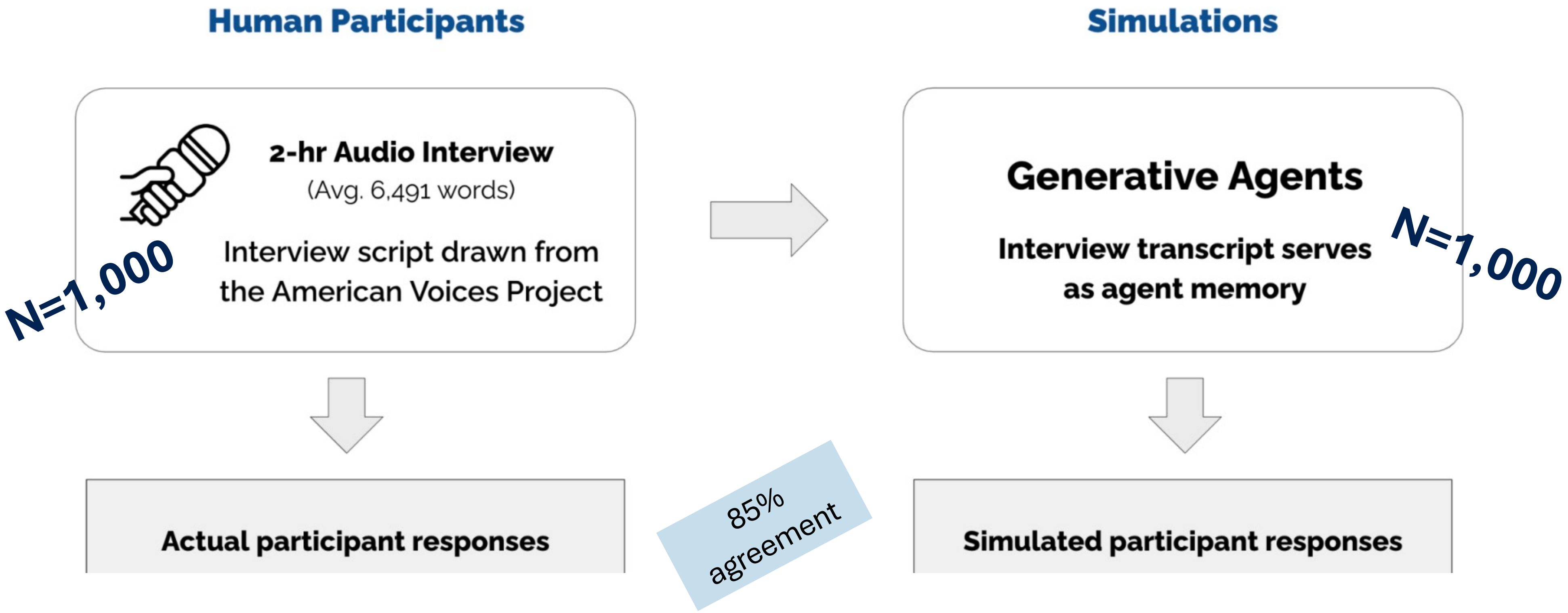
More detail

*“Ideologically, I describe myself as **conservative**. Politically, I am a **strong Republican**. Racially, I am **white**. I am **female**. Financially, I am **lower-class**. In terms of my age, I am **young**.*

When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: ”

Argyle et al 10.1017/pan.2023.2

In-depth Context



Human Participants

Simulations

2-hr Audio Interview
(Avg. 6,491 words)

Interview script drawn from the American Voices Project

N=1,000

Generative Agents

Interview transcript serves as agent memory

N=1,000

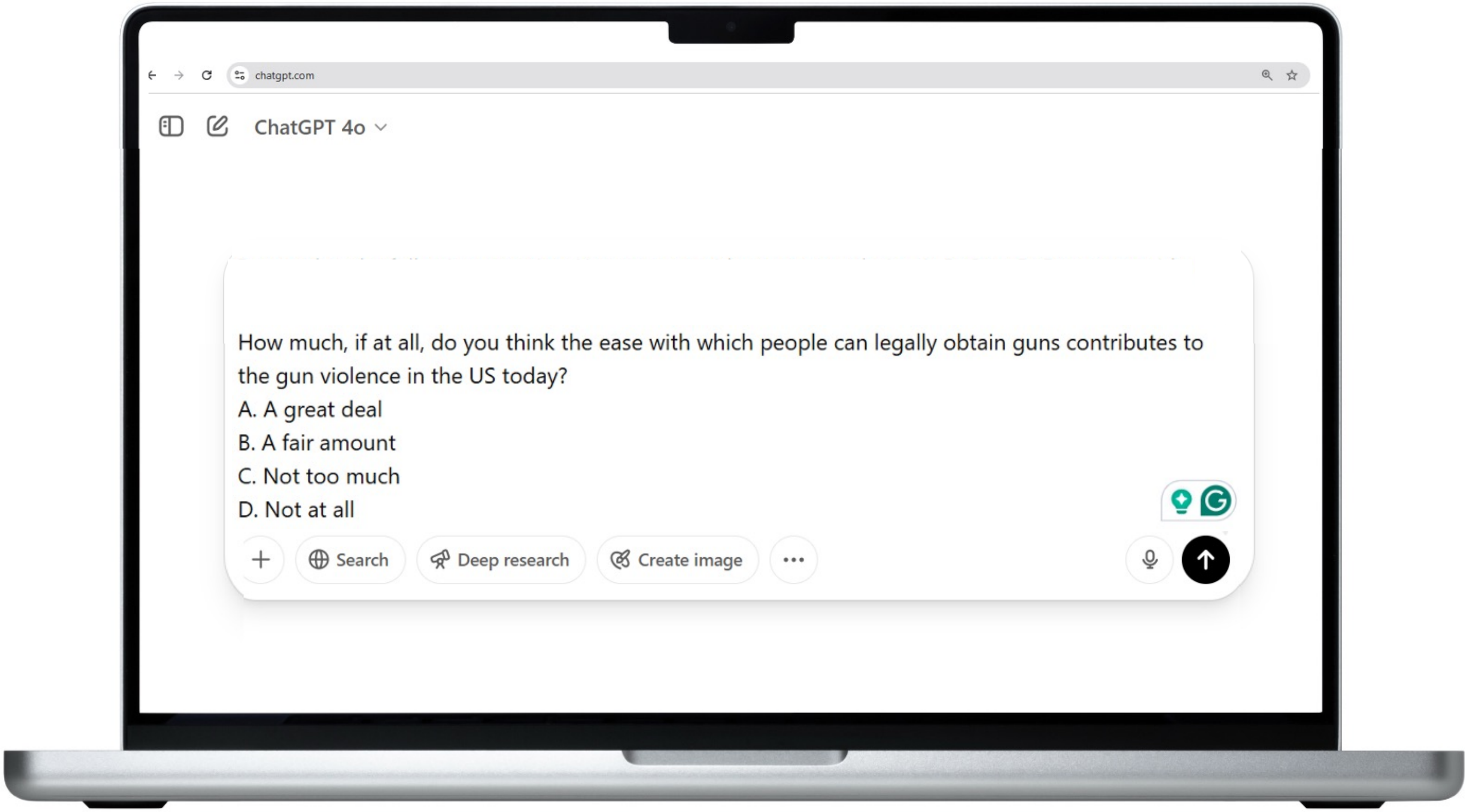
Actual participant responses

85% agreement

Simulated participant responses

Park et al 10.48550/arXiv.2411.10109

Zero-Context Prompting



How much, if at all, do you think the ease with which people can legally obtain guns contributes to the gun violence in the US today?

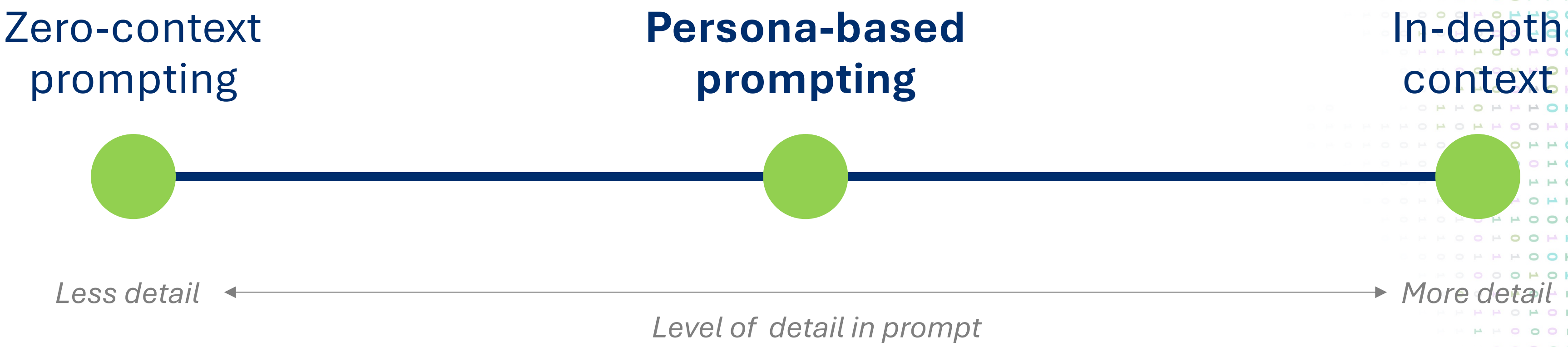
A. A great deal

B. A fair amount

C. Not too much

D. Not at all

What Xs to Give Model?



Zero-context prompting

Persona-based prompting

In-depth context

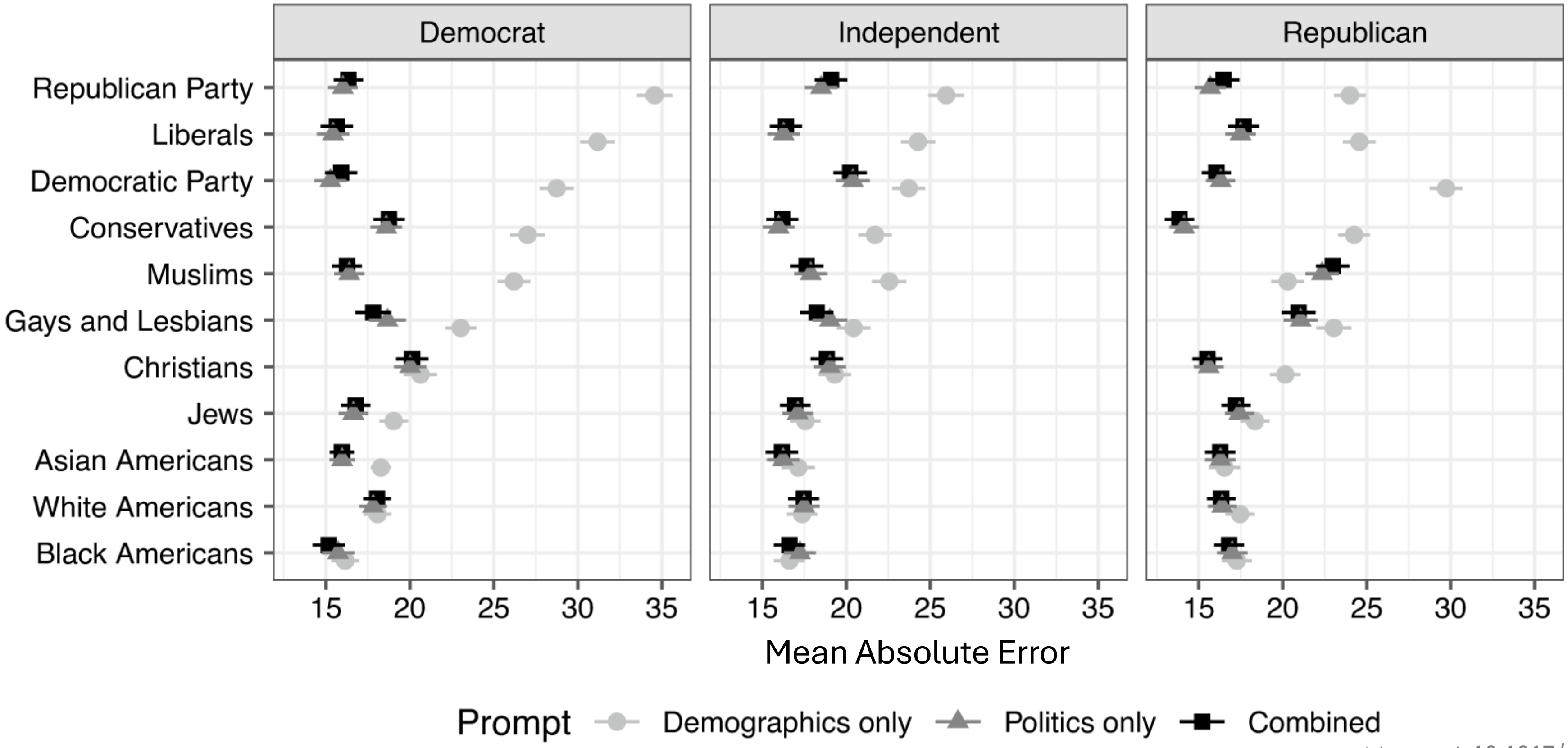
Less detail

More detail

Level of detail in prompt

- Same issue as weighting, nonprobability surveys
  - Sometimes adjustments work
  - Likely related to correlation of X & R, X & Y – with model as intermediary

Better Ys with More Xs



Bisbee et al. 10.1017/pan.2024.5

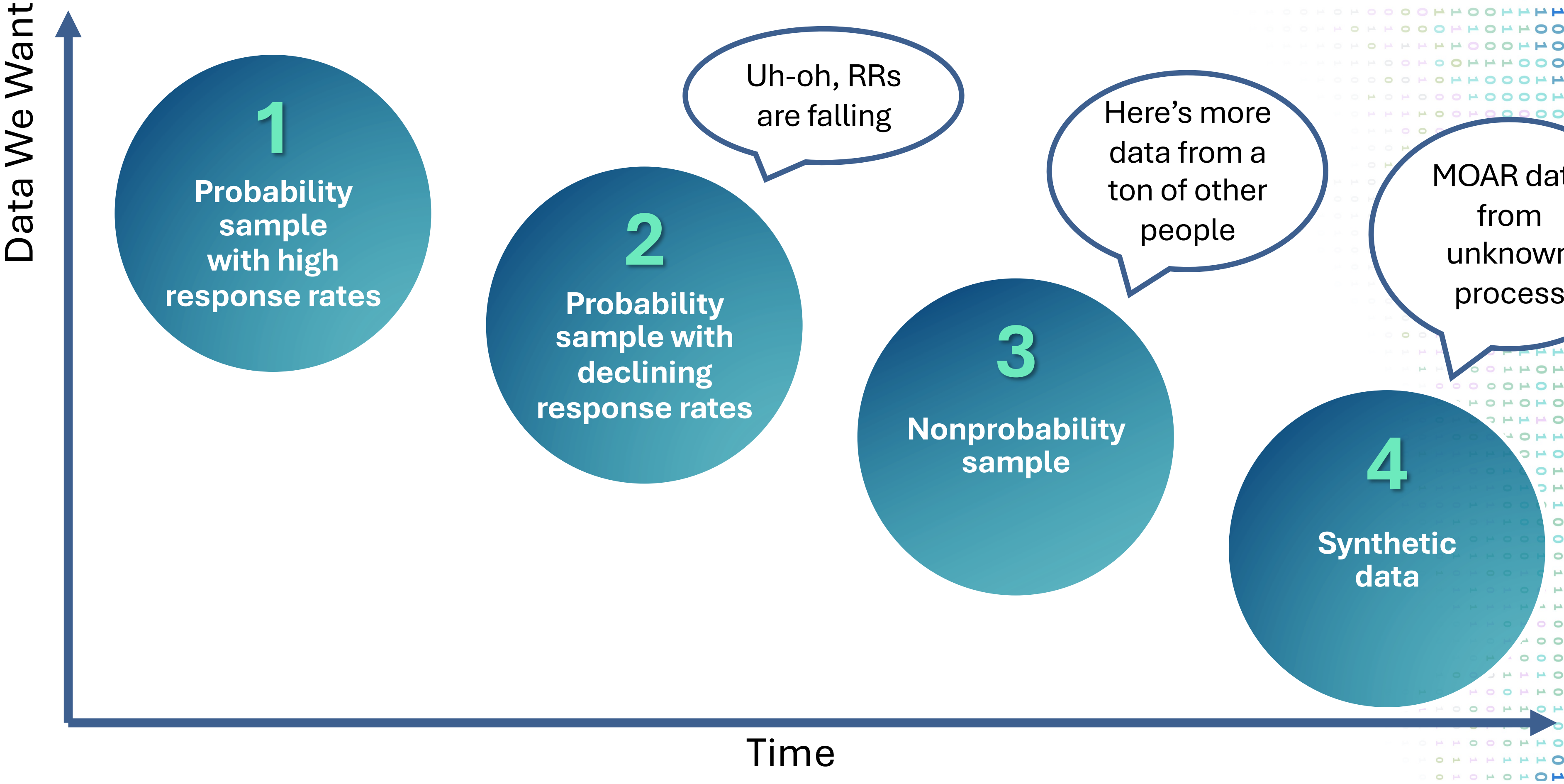
Given X, Can Model Predict Y?

- We don't understand the data generating process

$$Y = f(X)$$

- What can go wrong?
  - Hallucinations
  - Temporal issues
  - **Social bias**: consistent bias towards majority group
  - **Machine bias**: inconsistent bias across topics, groups
  - Sensitivity to prompt style, wording, order, etc

Solving the Wrong Problem



Data We Want

1 Probability sample with high response rates

2 Probability sample with declining response rates

Uh-oh, RRs are falling

3 Nonprobability sample

Here's more data from a ton of other people

4 Synthetic data

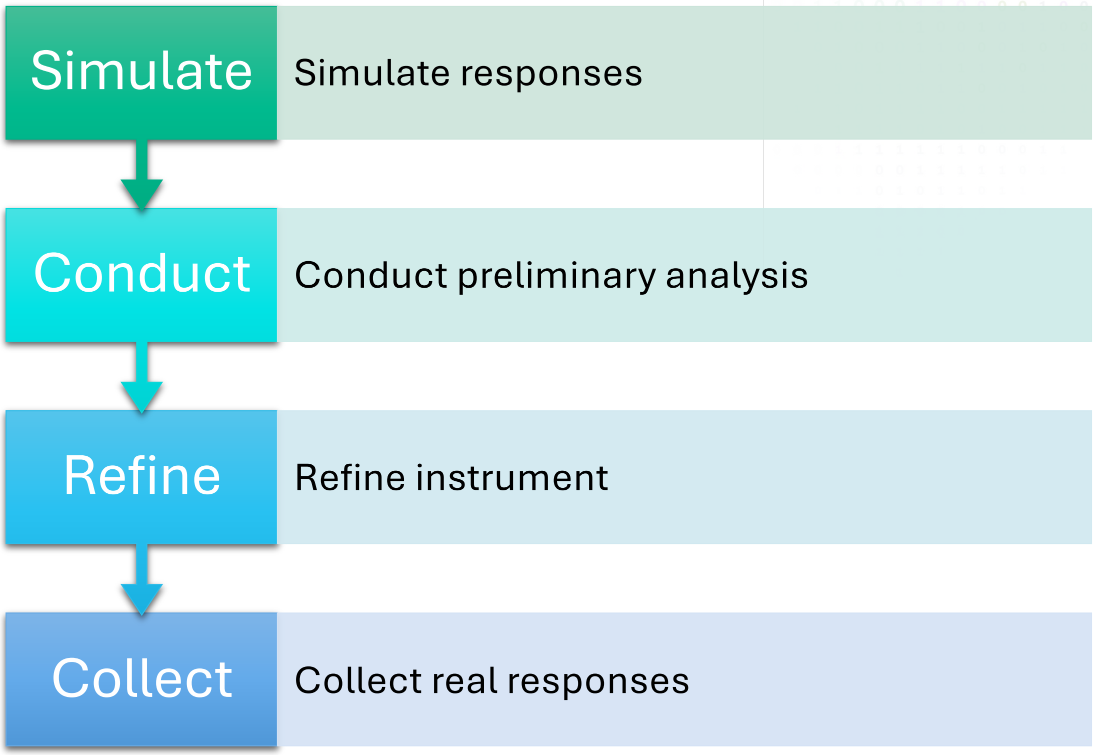
MOAR data from unknown process

Time

The Right Problems

- Improve bias reduction methods
  - High quality data on more X variables
  - Theory to know the right X variables
- Ethical issues
  - Autonomy, agency, and consent
  - False sense of inclusion, representativity
  - Lack of reproducibility

Pilot Testing with Synthetic Data



Simulate — Simulate responses

Conduct — Conduct preliminary analysis

Refine — Refine instrument

Collect — Collect real responses

Thank You

Stephanie Eckman

www.stepheckman.com