survey statistician's perspective

Frauke Kreuter

JPSM – Uni Mannheim – IAB @fraukolos

Adel Javanmard (?)


Berkeley University 2018

Data Collection Design

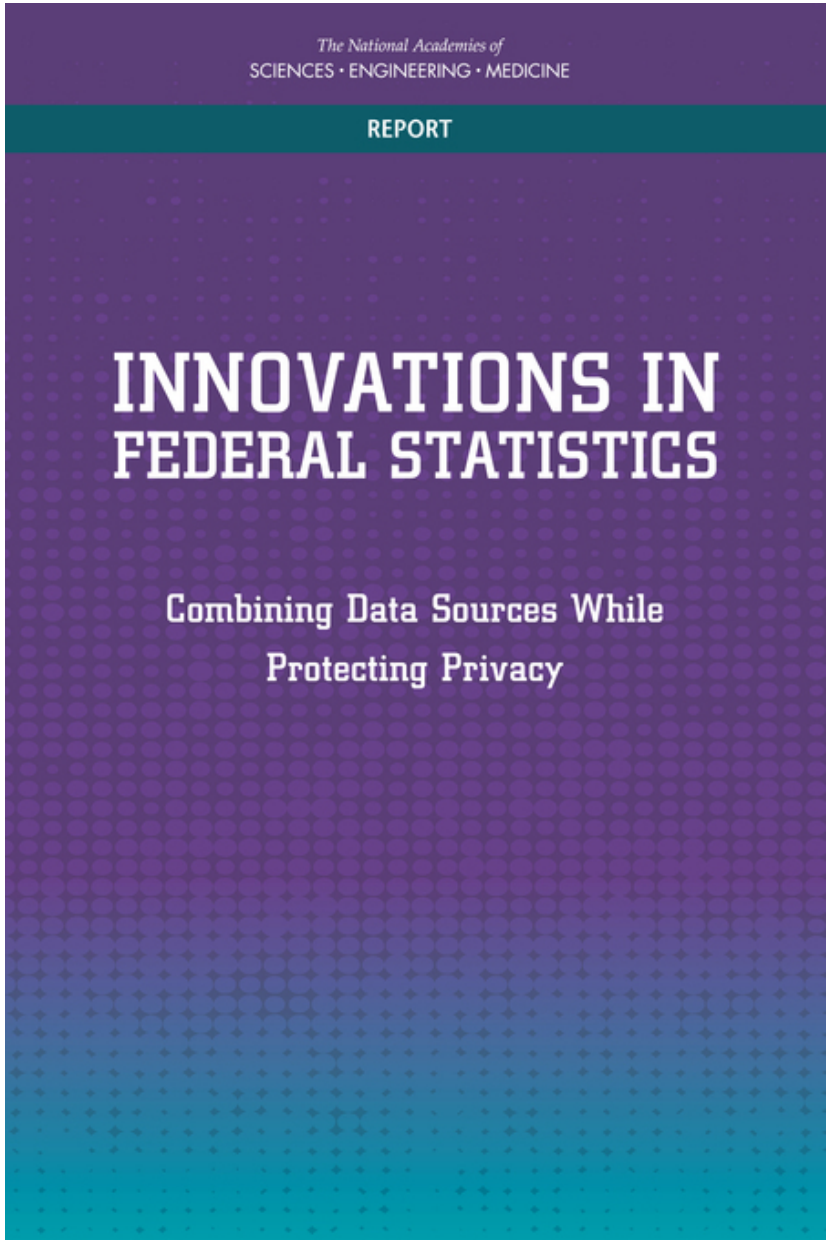
- What problem does DP solve at the recruitment stage?
- How do we deal with error prone survey answers in DP?
- Can we afford our data collection if we design for DP?

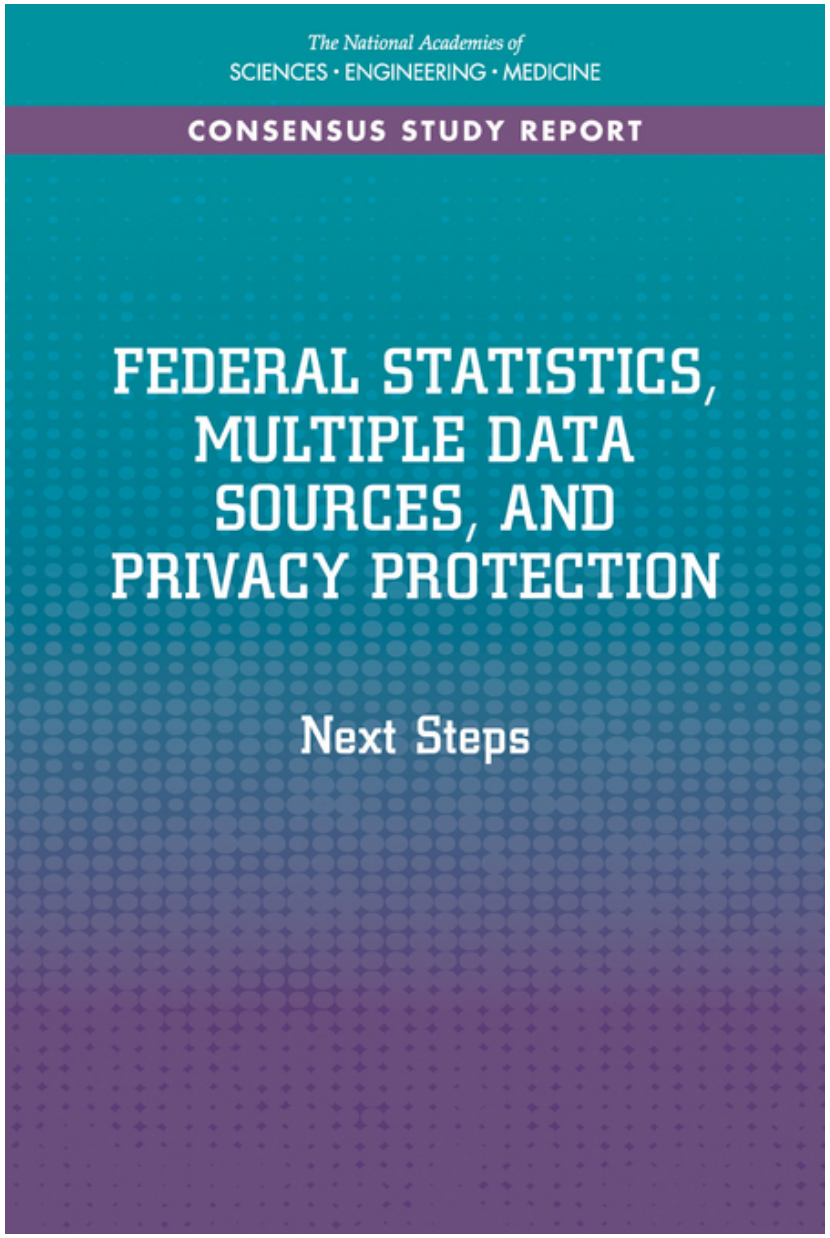
Data Analysis

- Can we still work the way we are used to with DP data?
- Do we risk distorting the benchmark?

Research Community and Replication

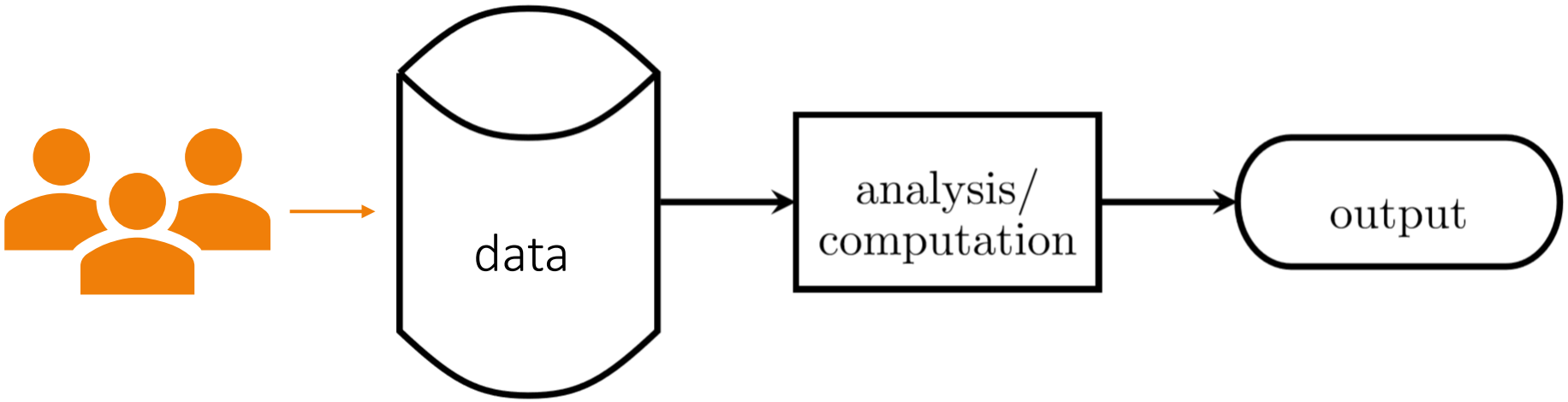
DP: Differential Privacy


The National Academies of
SCIENCES · ENGINEERING · MEDICINE
REPORT
INNOVATIONS IN
FEDERAL STATISTICS
Combining Data Sources While
Protecting Privacy


The National Academies of
SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
FEDERAL STATISTICS,
MULTIPLE DATA
SOURCES, AND
PRIVACY PROTECTION
Next Steps

Data Collection I

Recruitment and answering sensitive questions


data
analysis/
computation
output

Coutts & Jann [2011, SMR]

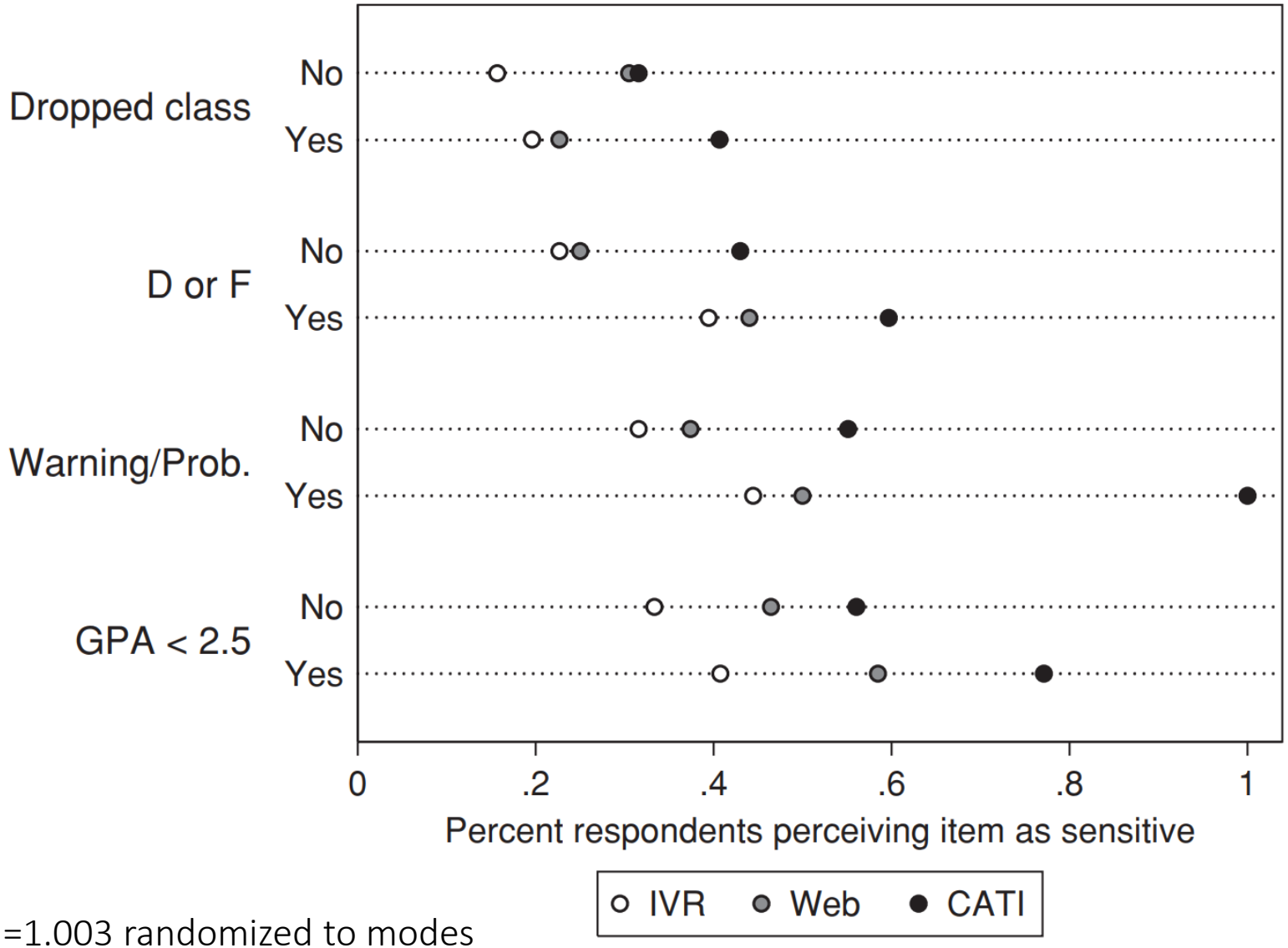
Randomized Response Techniques are problematic because of
- limited trust
- high variance due to false negative tendency (especially for more sensitive questions)

Kirchner [2015]

No improvement of reporting accuracy with RRT compared to direct questioning (using administrative data for benchmark validation)

Caution shared by others [e.g. Holbrook and Krosnik 2010; Coutts et al. 2011; Wolter and Preisendoerfer 2013; Hoeglinger, Jann, and Diekmann 2014] though not all [e.g. Blair, Imai, Zhou 2015]

Reported Sensitivity of Survey Questions by True State and Mode of Data Collection [Kreuter, Presser, Tourangeau 2008, POQ]



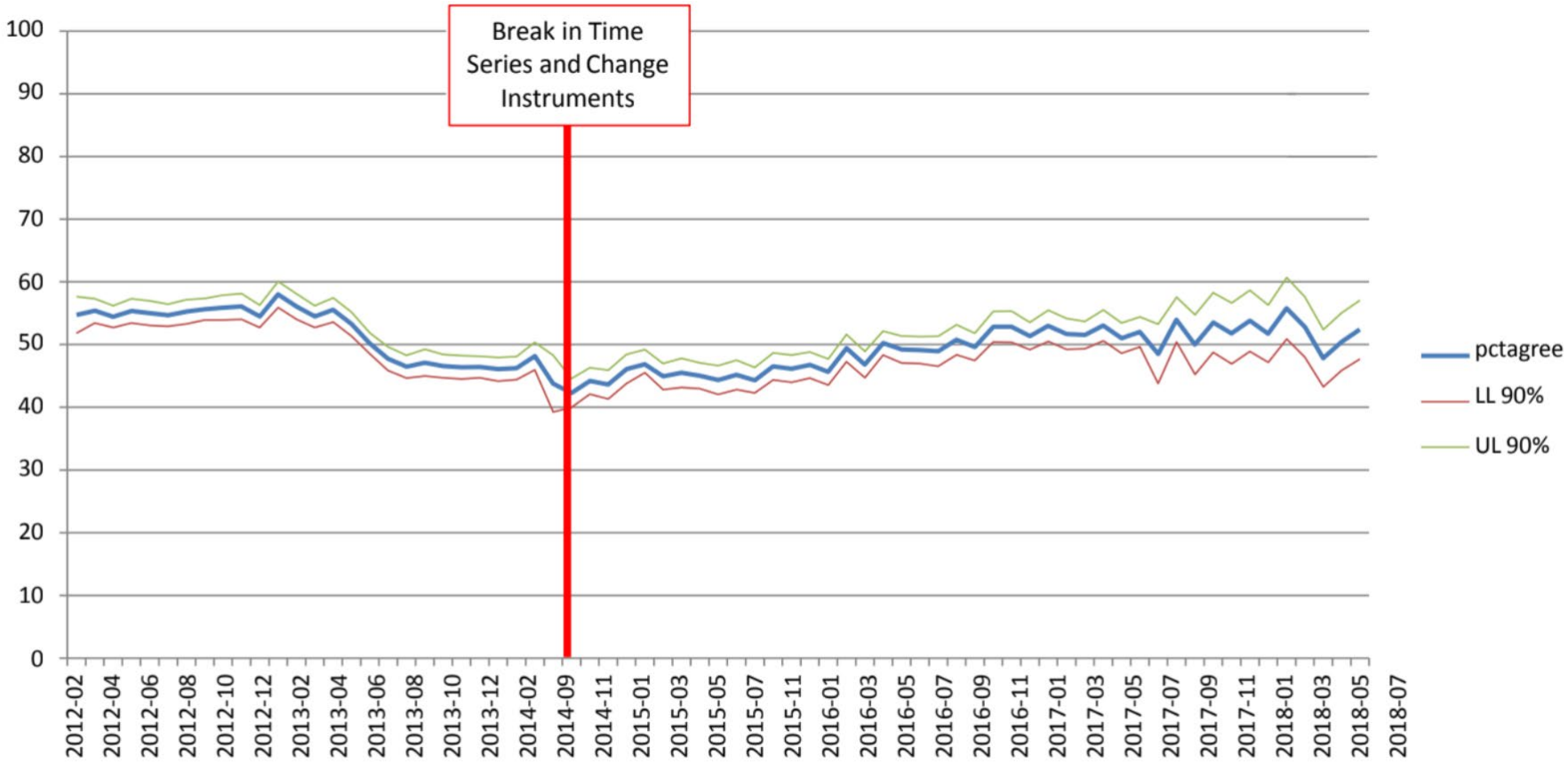
Survey of UMD alumni n=1.003 randomized to modes

Sexual Assault and Harassment in the U.S. Military V. 2 [Morral, Gore, Schell 2015, RAND]

Table 3.9
Types of Offender Behaviors Indicating Coercion/Lack of Consent for Past-Year Non-Penetrative Sexual Assaults, by Gender

Question	Men	Women
They continued even when you told them or showed them that you were unwilling	60.75% (50.44–70.39)	54.15% (50.26–58.01)
They used physical force to make you comply	13.96% (8.08–21.88)	24.04%[a] (20.63–27.72)
They physically injured you	5.02% (1.92–10.44)	4.59% (3.02–6.67)
They threatened to physically hurt you (or someone else)	7.94% (3.29–15.58)	4.69% (3.17–6.65)
They threatened you (or someone else) in some other way	15.52% (9.10–24.04)	20.36% (17.10–23.94)
They did it when you were passed out, asleep, or unconscious	7.12% (1.05–22.09)	11.64% (9.02–14.70)
They did it when you were so drunk, high, or drugged that you could not understand what was happening or could not show them that you were unwilling	10.12% (3.02–23.23)	15.61% (12.66–18.94)

Reported Believe – Data are Kept Confidential in the Federal Statistical System [Childs, Eggelston, Fobia 2018, BigSurv]



* Change in instruments coincided with a 4.8% decrease in reported belief.

Census data collection via Gallup Tracking Poll

Your participation is vital to our effort. Domestic terrorism preparedness transcends any single level of government, including the Federal government. It is a national issue that can only be effectively addressed through close cooperation at all levels—Federal, state, and local. The work of this Panel concerns nothing less than the security of our nation, the protection of our citizens' civil liberties, and the ideals of our democratic society.

Your organization has been randomly selected to represent «ORG_TYPE_TEXT» throughout the United States. The survey is being

Ellen Embrey

* U.S. Department of Defe
Representative

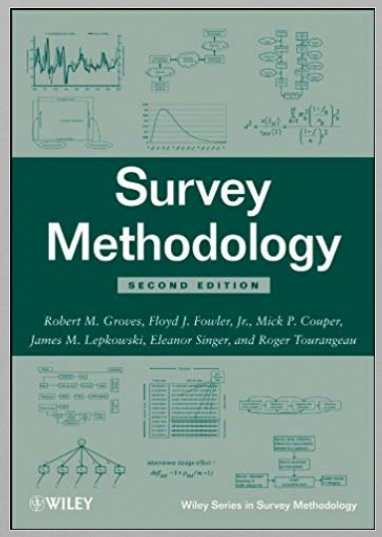
"The estimates will be the same with or without you in the data".

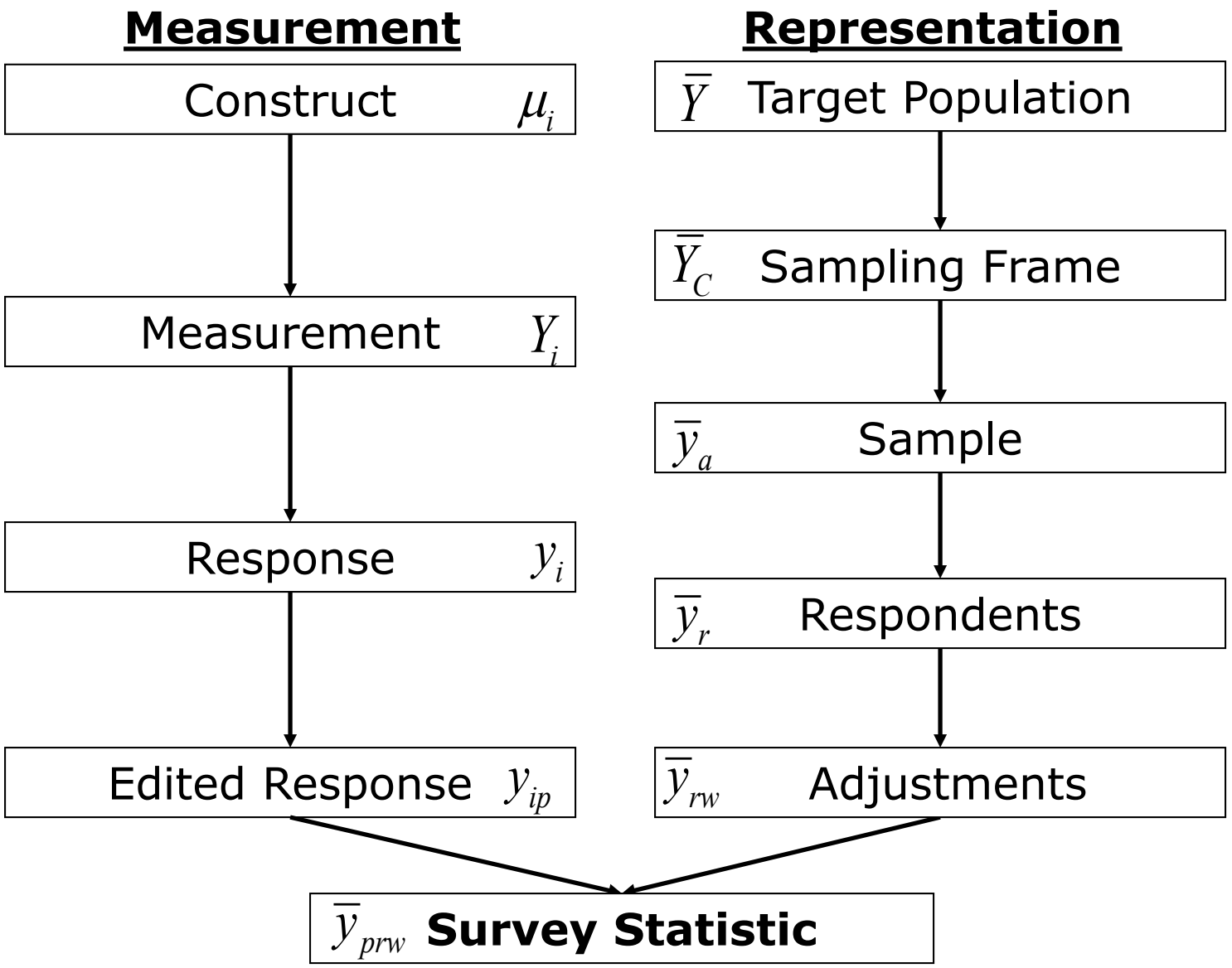
Examples of open research questions …

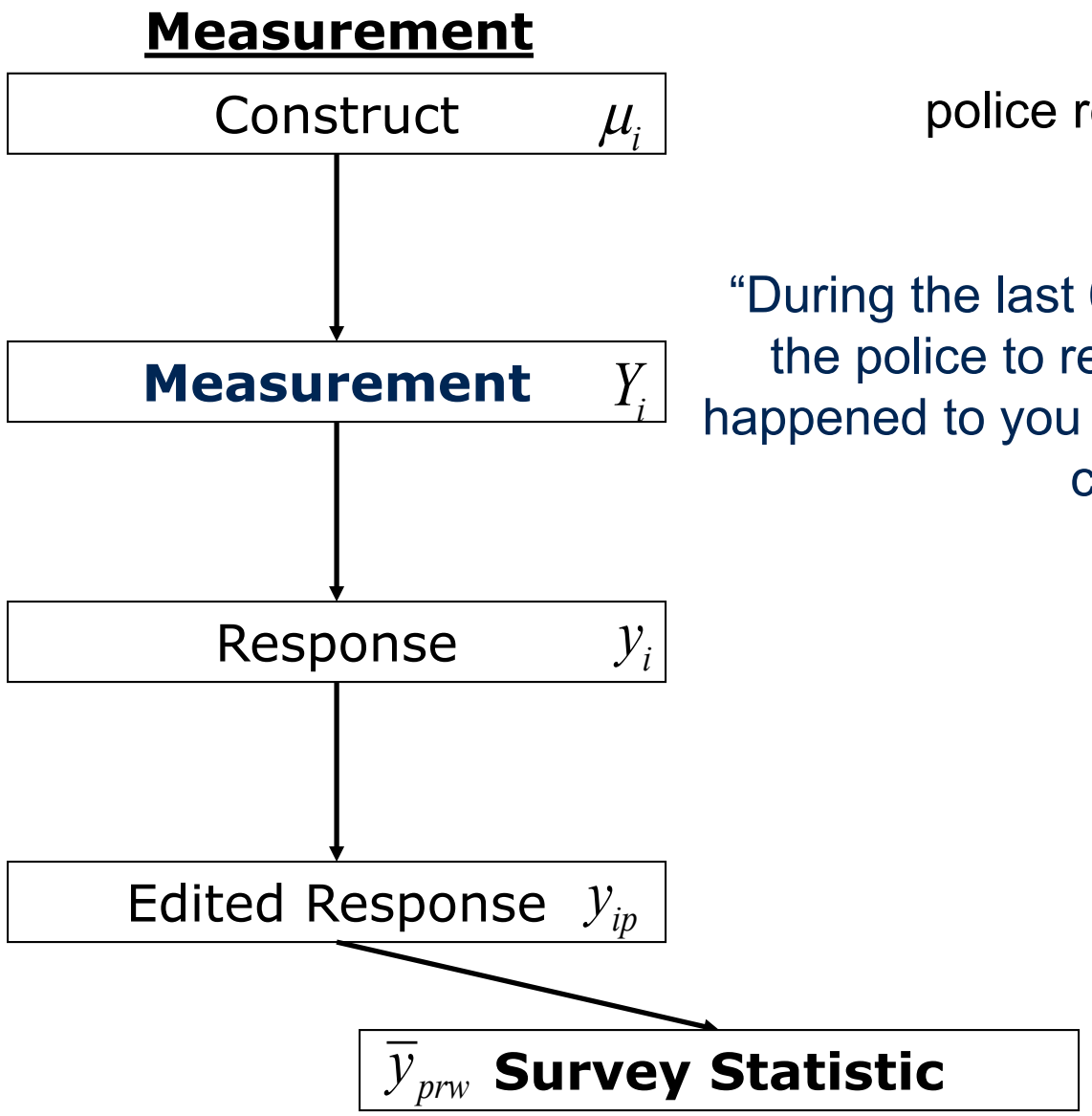
- How do we communicate the method?
- How do we establish sufficient trust?

Data Collection II

Random Noise and Missing Answers

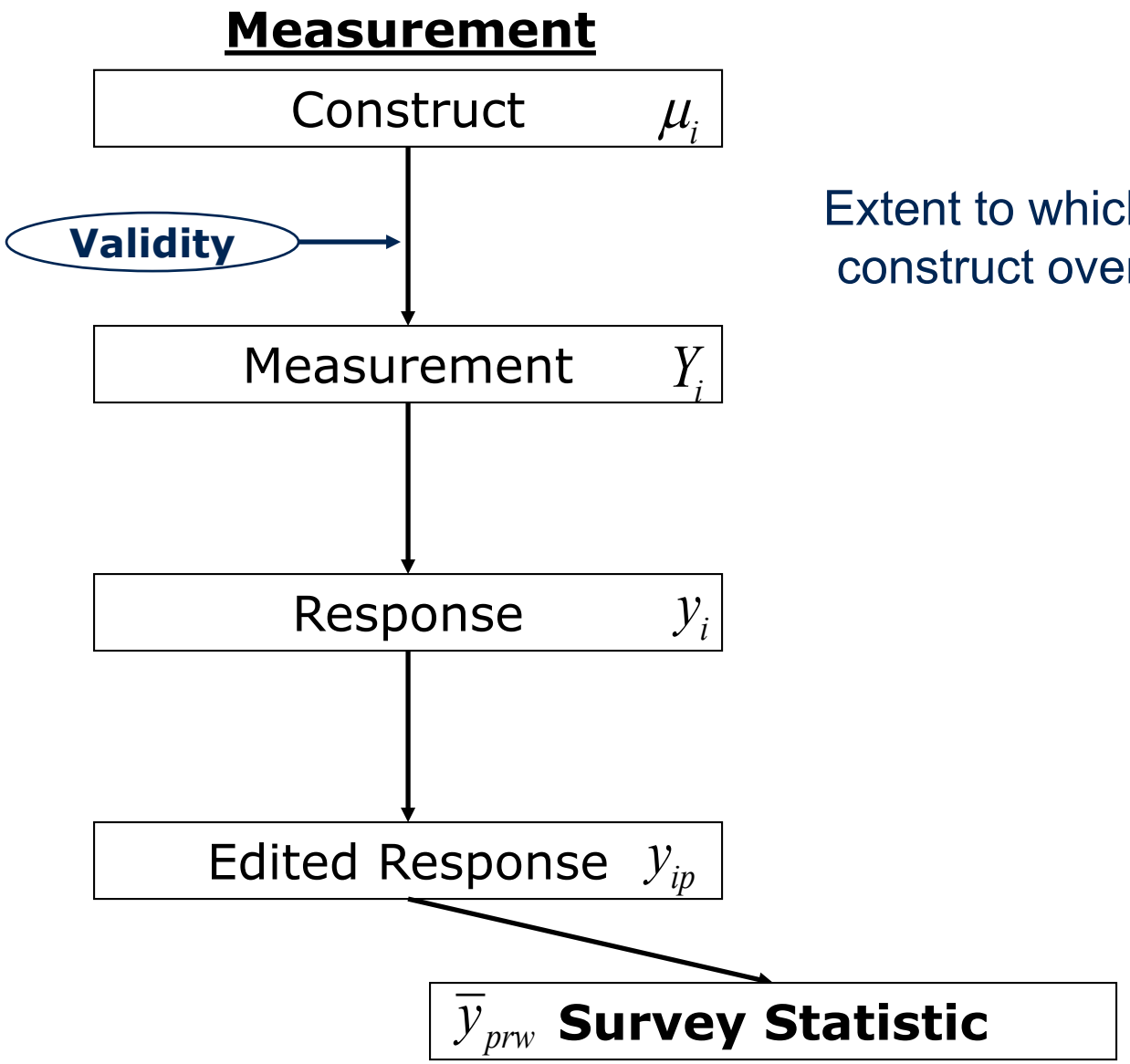



Measurement
Construct μ_i
Measurement Y_i
Response y_i
Edited Response y_{ip}
Representation
$\bar{Y}$ Target Population
$\bar{Y}_C$ Sampling Frame
$\bar{y}_a$ Sample
$\bar{y}_r$ Respondents
$\bar{y}_{rw}$ Adjustments
$\bar{y}_{prw}$ Survey Statistic

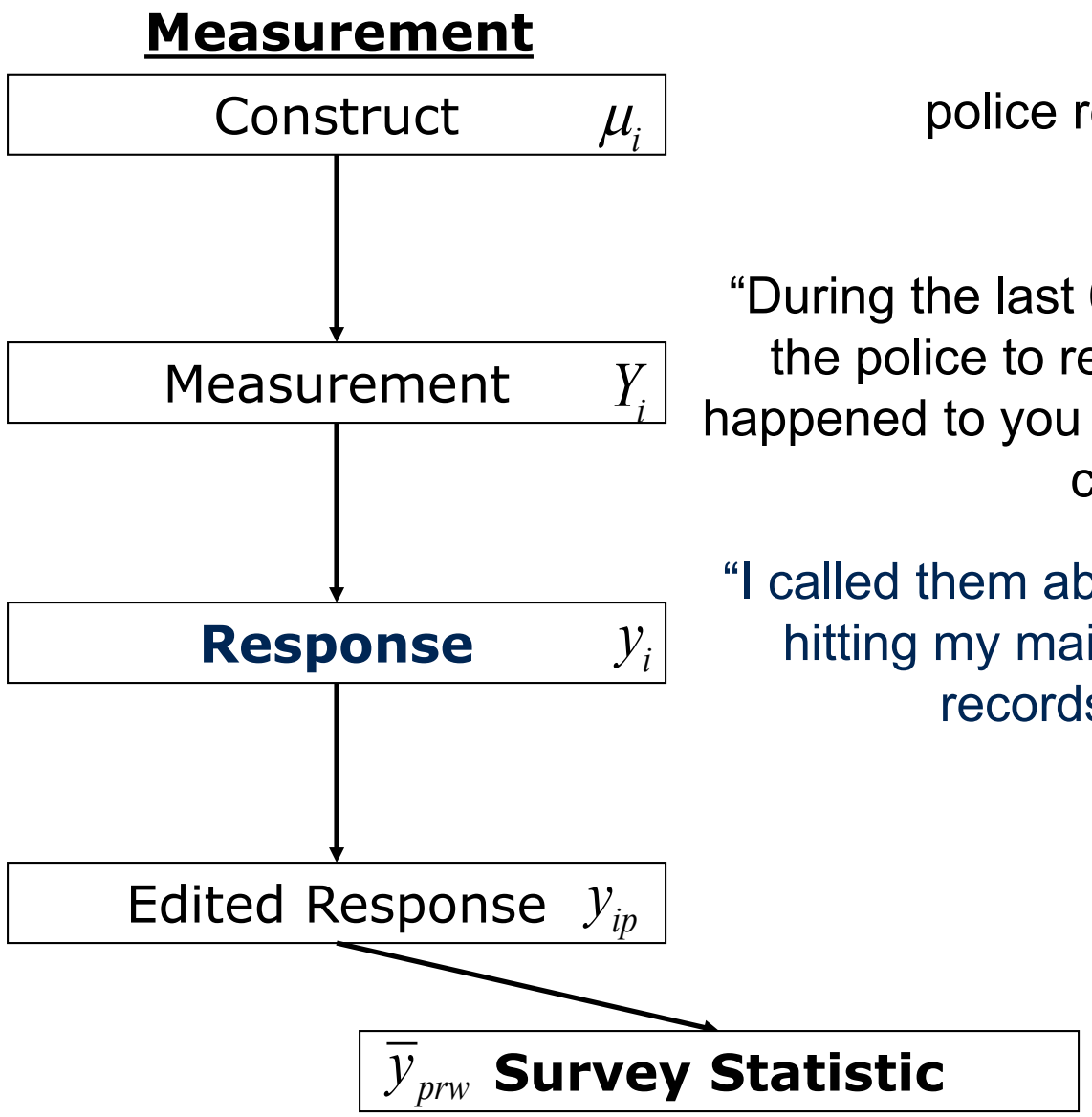


police reported crime

"During the last 6 months, did you call the police to report something that happened to you that you thought was a crime?"



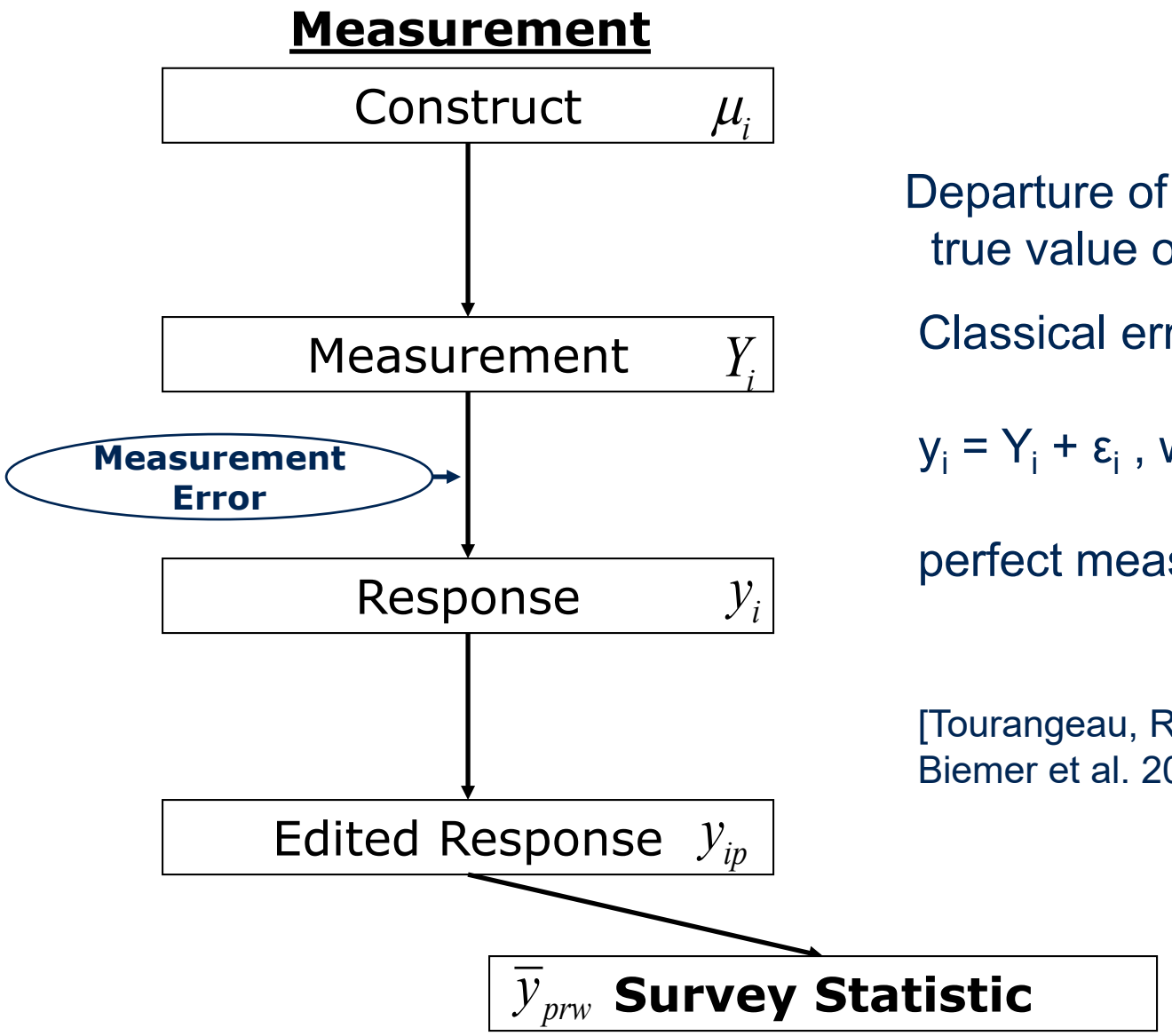
Extent to which measure reflects construct over all possible trials



police reported crime

"During the last 6 months, did you call the police to report something that happened to you that you thought was a crime?"

"I called them about my neighbor's car hitting my mailbox." – Interviewer records 1 for "Yes".



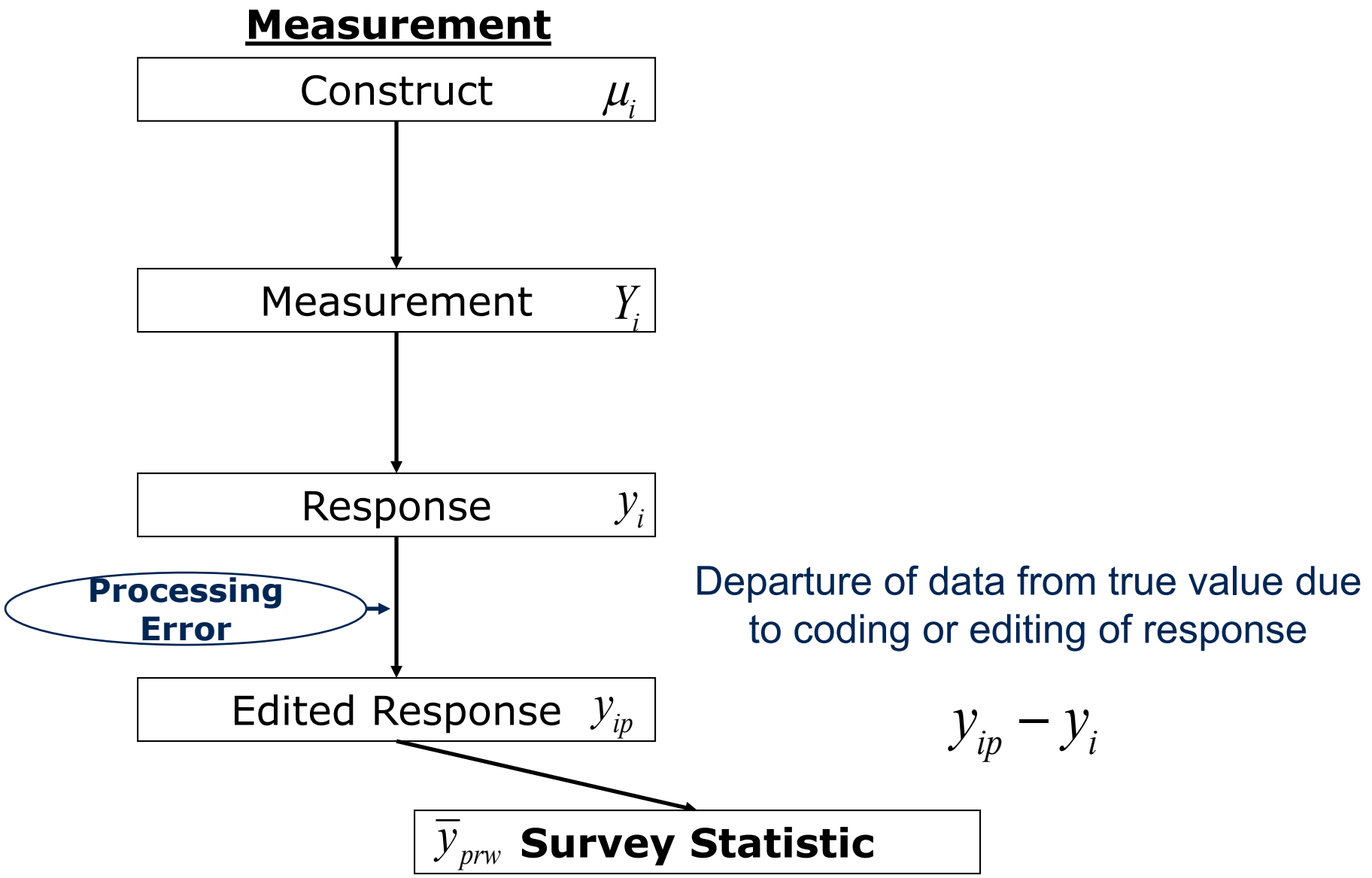
Departure of response to a measure from true value of measure for a respondent

Classical error model (Lord & Novick 1969)

$y_i = Y_i + \varepsilon_i$, with $\varepsilon \sim N(0, \sigma)$

perfect measurement means $\sigma = 0$.

[Tourangeau, Rips, Rasinski 2000; Krosnick & Presser 2010; Biemer et al. 2013; Vannette & Krosnick 2018]



Departure of data from true value due to coding or editing of response

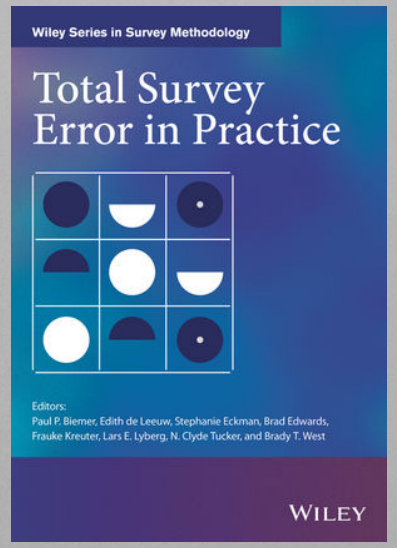
$$y_{ip} - y_i$$

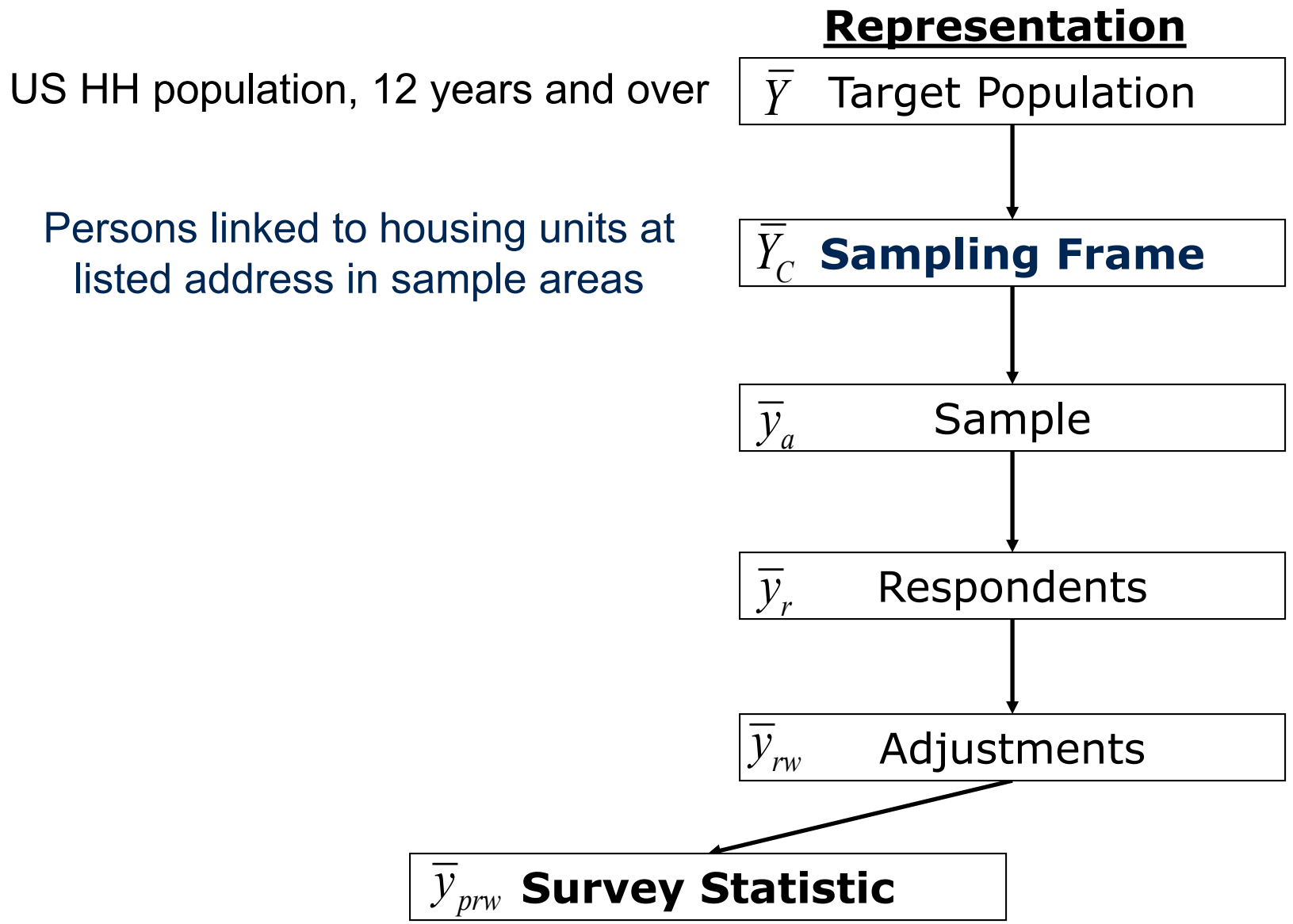
Examples of open research questions …

- What if data linkage is desired to reduce respondent burden and shorten the interview? [Sakshaug, Kreuter 2012, SRM]
- Should error-prone survey answers be considered fixed for the purpose of a DP definition? [Oberski 2019]
- Are we taking agency away from respondents
  by creating values for missing data or
  when we “improve” values for misreports?

Data Collection III

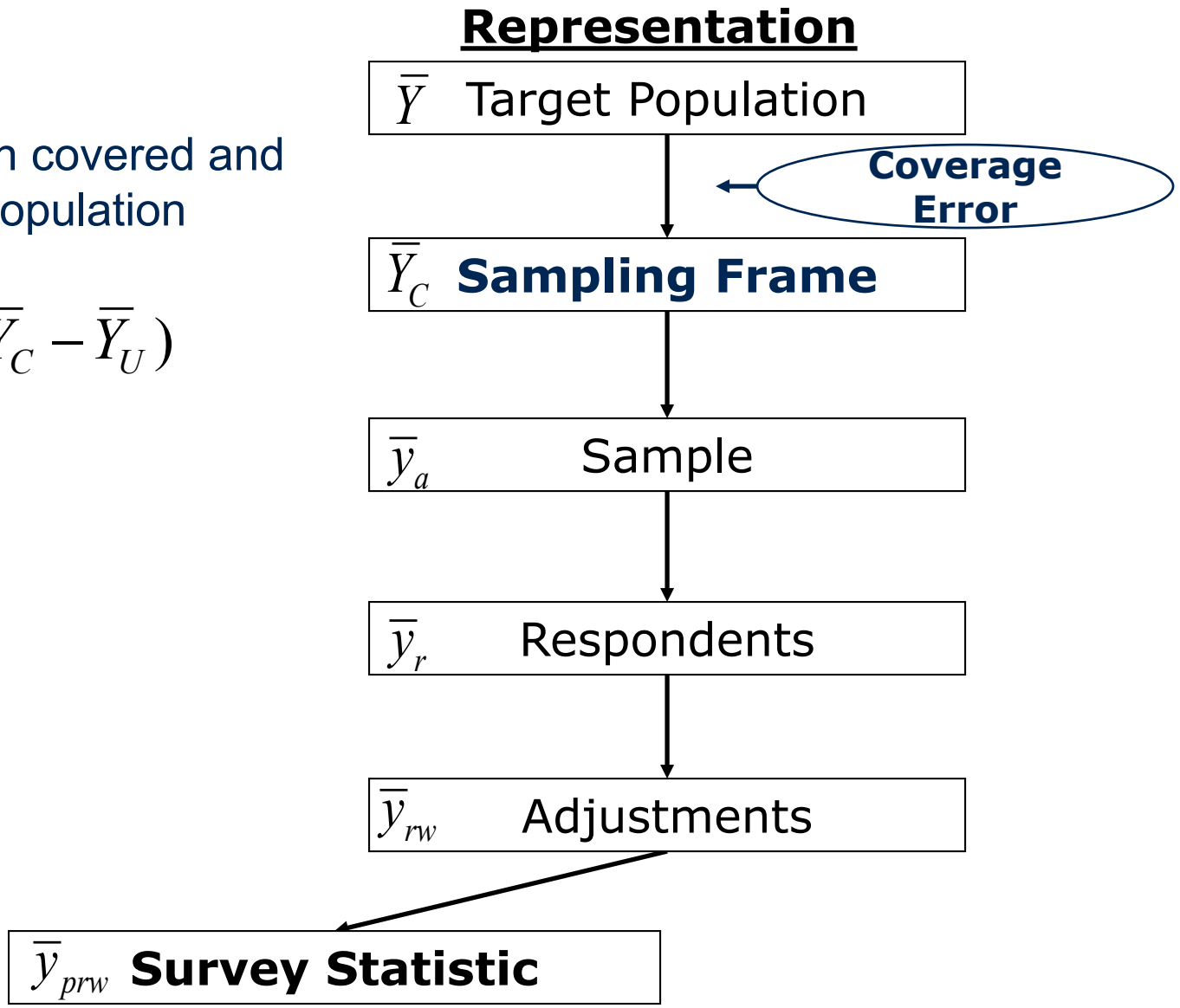
Design Decisions
Cost: $$ and time

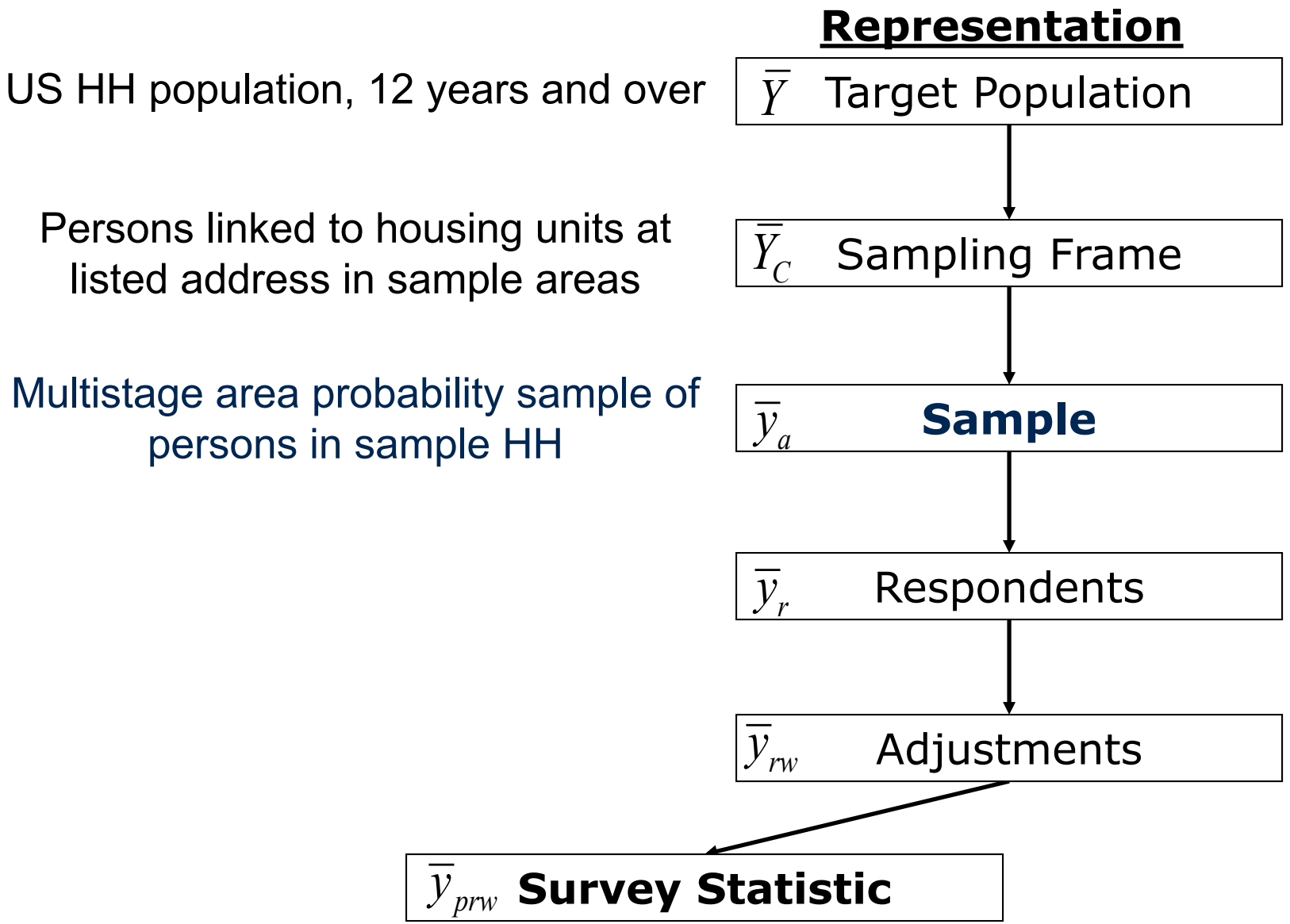



Representation
US HH population, 12 years and over
$\bar{Y}$ Target Population
Persons linked to housing units at listed address in sample areas
$\bar{Y}_C$ Sampling Frame
$\bar{y}_a$ Sample
$\bar{y}_r$ Respondents
$\bar{y}_{rw}$ Adjustments
$\bar{y}_{prw}$ Survey Statistic

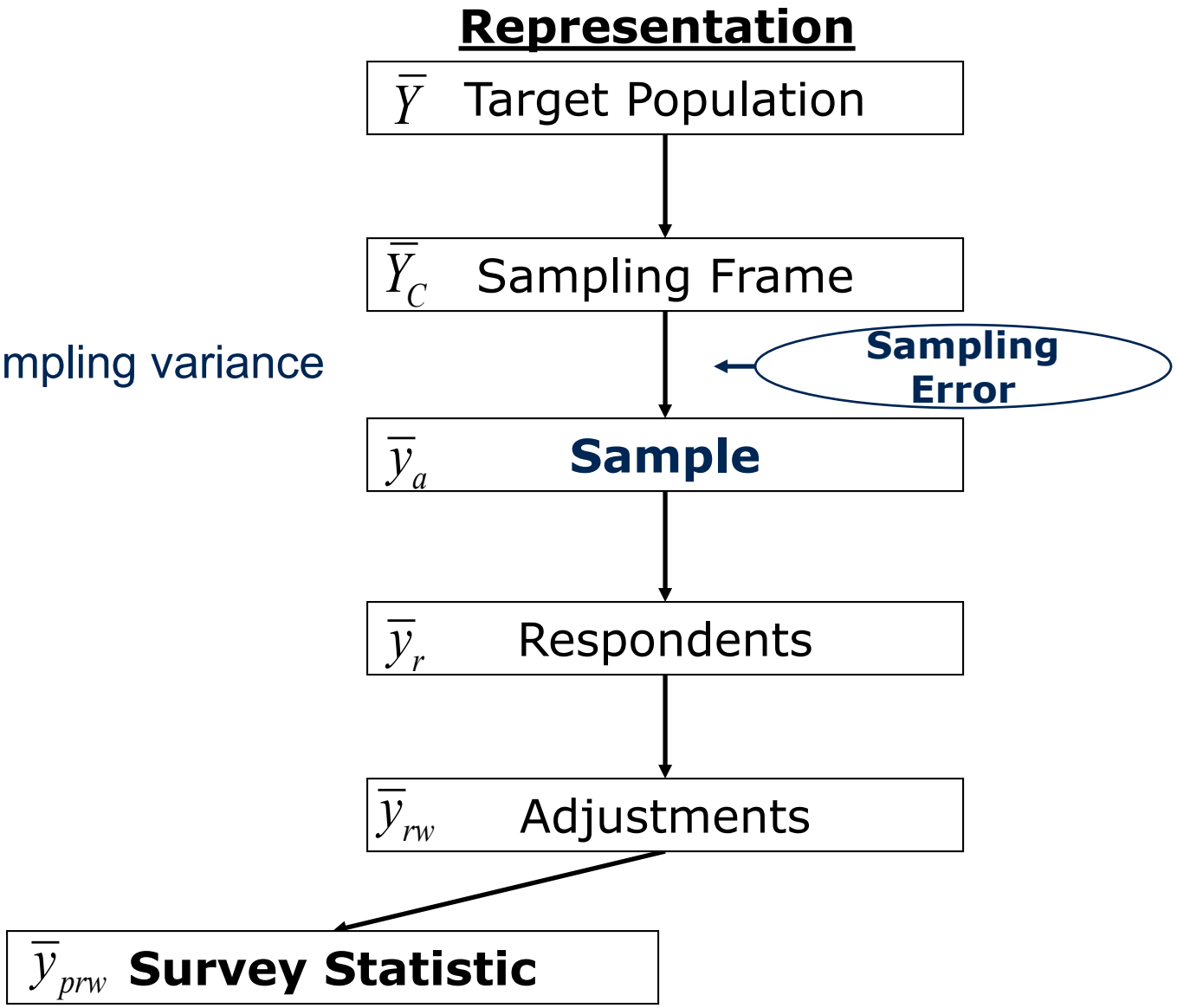
Difference between covered and noncovered population

$$\bar{Y}_c = \bar{Y}_N + \frac{U}{N}(\bar{Y}_C - \bar{Y}_U)$$




Representation
US HH population, 12 years and over
$\bar{Y}$ Target Population
Persons linked to housing units at listed address in sample areas
$\bar{Y}_C$ Sampling Frame
Multistage area probability sample of persons in sample HH
$\bar{y}_a$ Sample
$\bar{y}_r$ Respondents
$\bar{y}_{rw}$ Adjustments
$\bar{y}_{prw}$ Survey Statistic

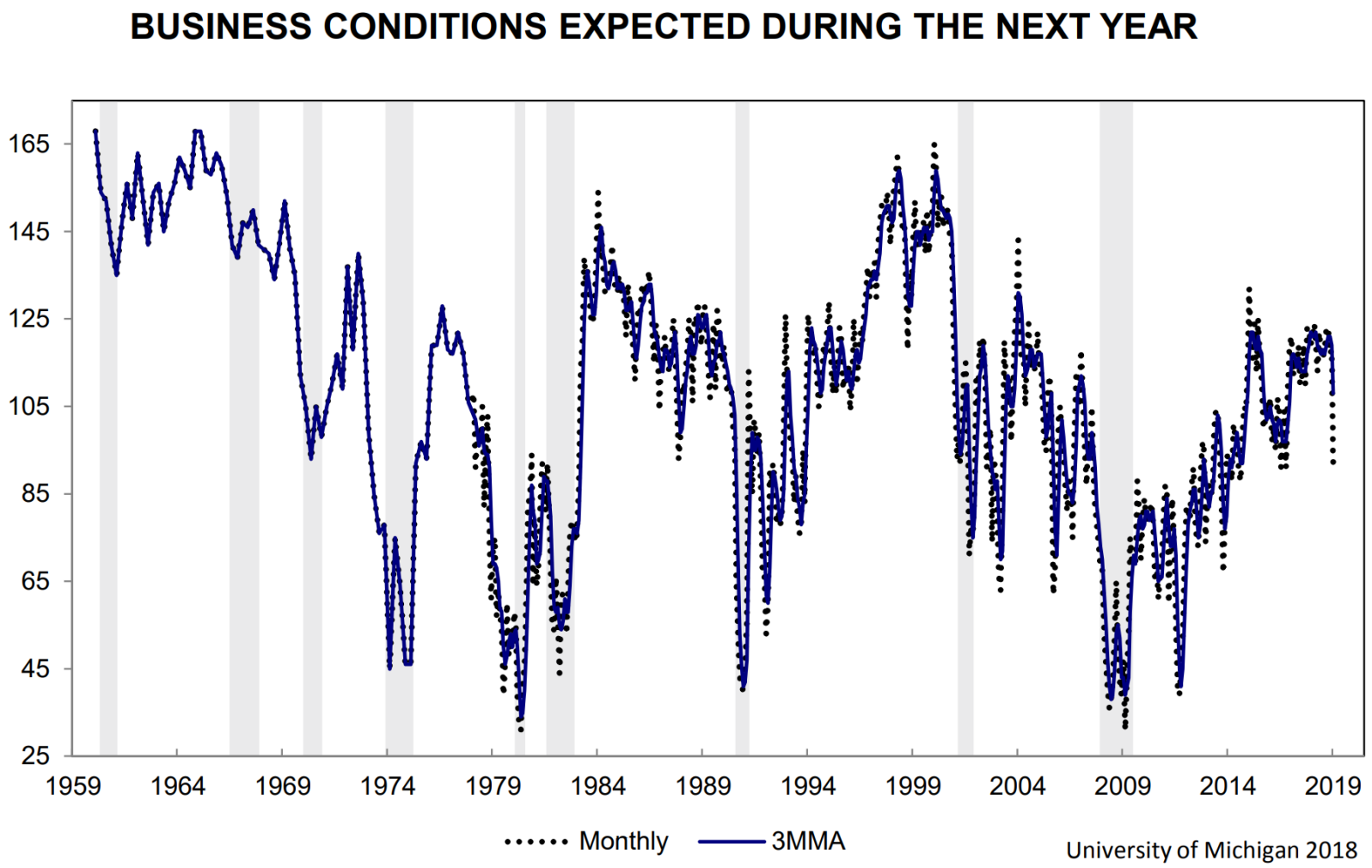
Sampling bias vs. sampling variance



Target population: Noninstitutionalized adults in contiguous U.S.

500 adults at random every month

How would you do that? What is the probability?



University of Michigan, University of Michigan: Consumer Sentiment [UMCSENT], retrieved January 29, 2019.

BJS Bureau of Justice Statistics

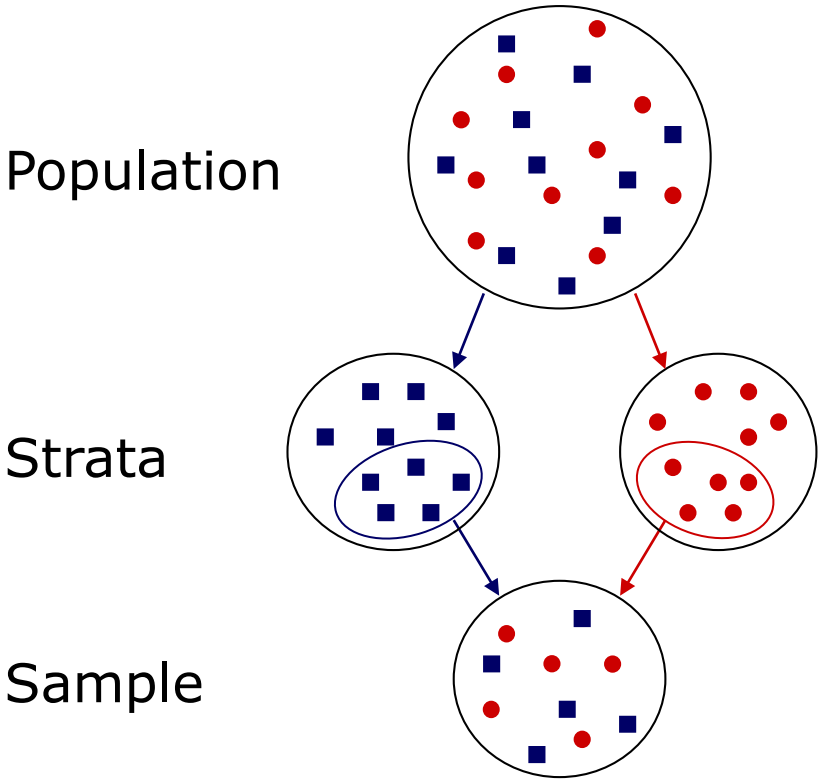
Many important surveys are not simple random samples but have one or more of these characteristics:

- Multistage
- Stratified
- Clustered
- Sampled with unequal probabilities

National Crime and Victimization Survey

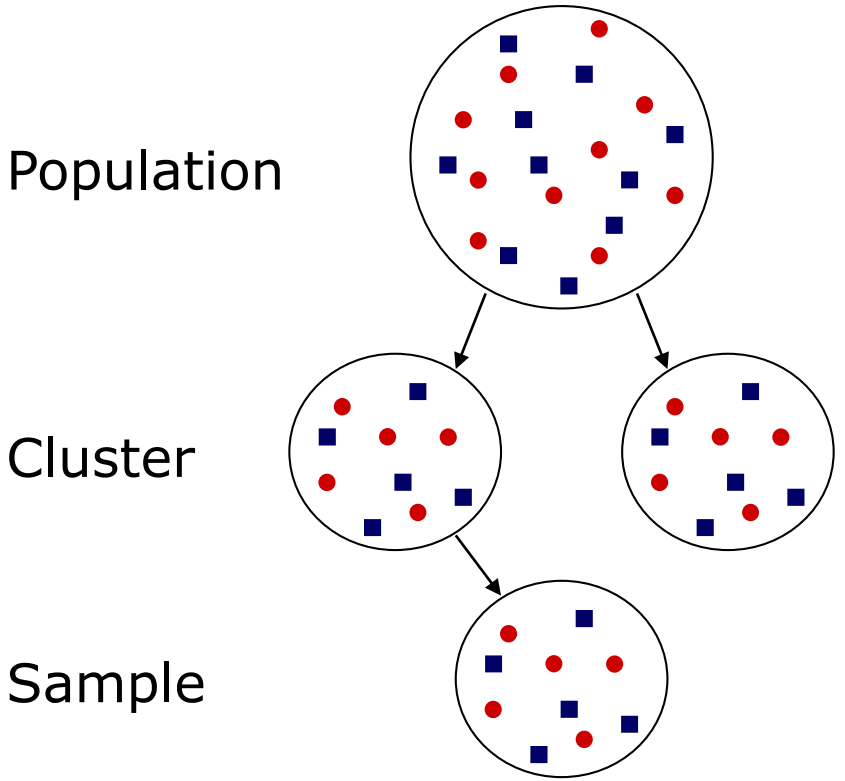
Sponsor	U.S. Bureau of Justice Statistics
Collector	U.S. Census Bureau
Purpose	Main objectives are to: • Develop detailed information about the victims and consequences of crime • Estimate the number and types of crimes not reported to the police • Provide uniform measures of selected types of crimes • Permit comparisons over time and by types of areas
Year Started	1973 (previously called the National Crime Survey, 1973–1992)
Target Population	Adults and children 12 or older, civilian and noninstitutionalized
Sampling Frame	U.S. households, enumerated through counties, blocks, listed addresses, lists of members of the household
Sample Design	Multistage, stratified, clustered area probability sample, with sample units rotating in and out of the sample over three years
Sample Size	About 41,800 households (78,600 persons)
Use of Interviewer	Interviewer administered
Mode of Administration	Face-to-face and telephone interviews
Computer Assistance	Paper questionnaire for 70% of the interviews, both face-to-face and telephone interviews; computer assistance for 30% of the interviews
Reporting Unit	Each person age 12 or older in household reports for self
Time Dimension	Ongoing rotating panel survey of addresses
Frequency	Monthly data collection
Interviews per Round of Survey	Sampled housing units are interviewed every six months over the course of three years
Levels of Observation	Victimization incident, person, household
Web Link	*http://www.ojp.usdoj.gov/bjs/cvict.htm*

STRATIFIED SAMPLING



Focus on reducing variance

CLUSTER SAMPLING



Focus on reducing costs

Valliant, Dever, & Kreuter (2015). PracTools: Tools for Designing and Weighting Survey Samples. R package version 0.3. http://CRAN.R-project.org/package=PracTools

Effect on Sample Size – Cost (rough estimates)

Mail: $50 per case

Phone: $250 per case

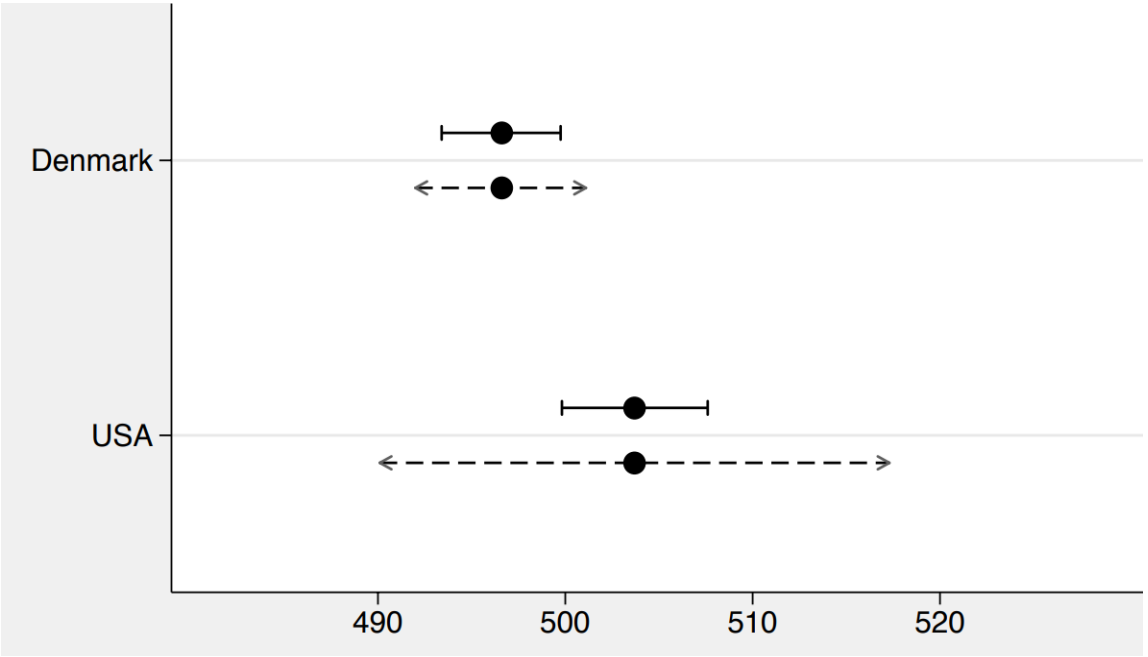
Face-to-face: $1,000 per case

Complex sample design inflates SE by a design effect of 1.4

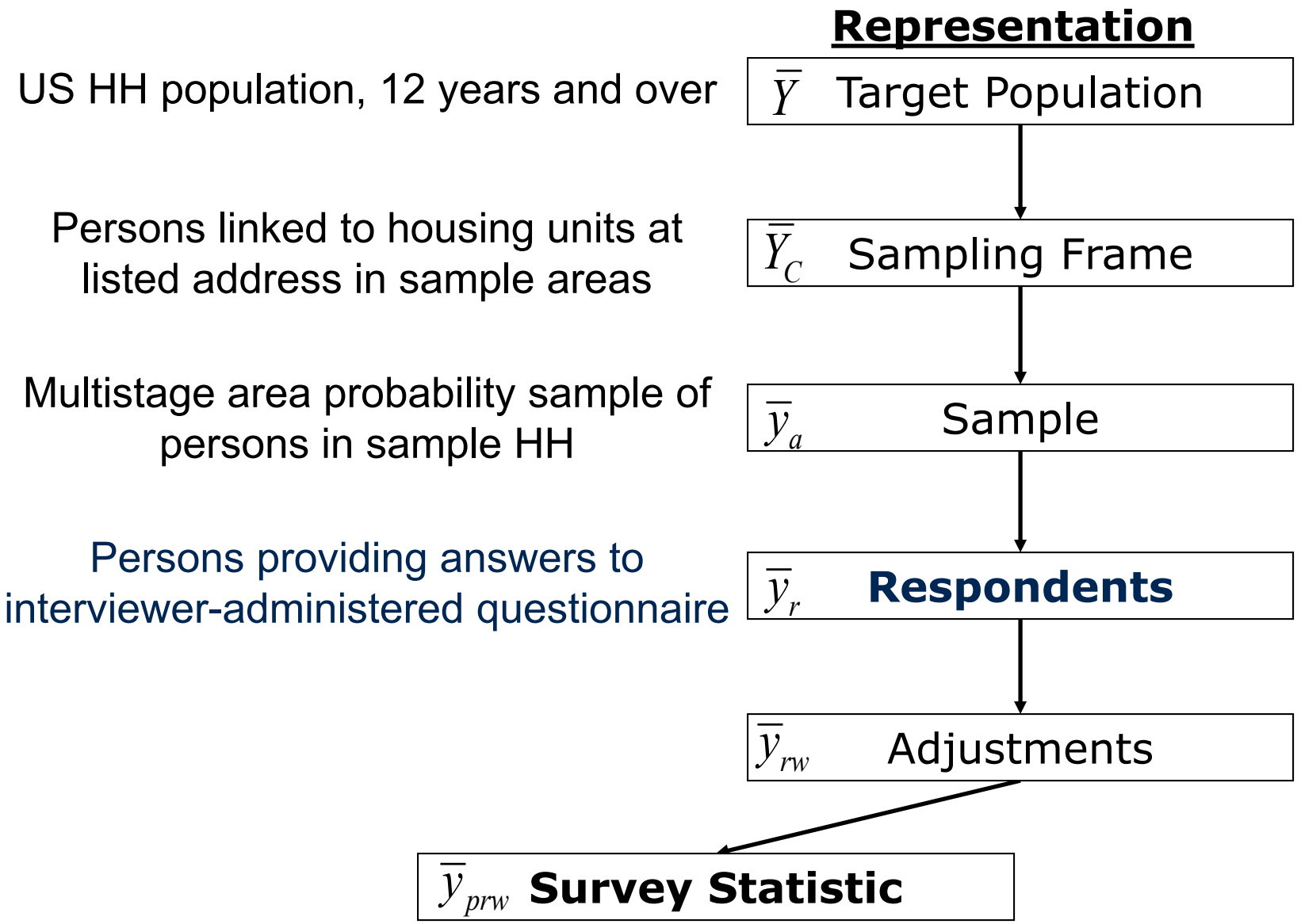
Mail: several weeks

Phone: couple of months

Face-to-face: several months

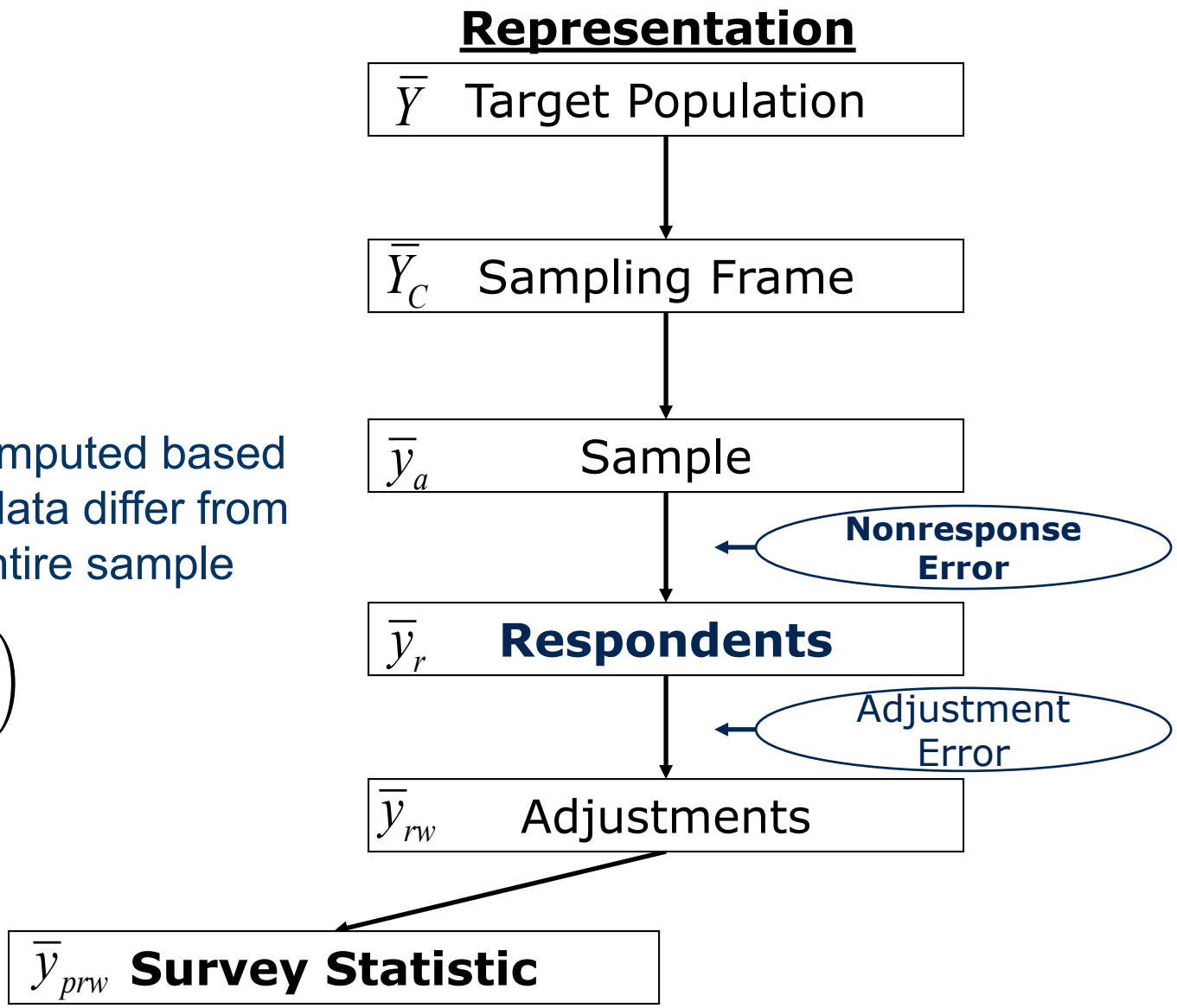


Kreuter, Valliant (2007), PISA Test scores means and confidence intervals with and without complex sample design


Representation
US HH population, 12 years and over
$\bar{Y}$ Target Population
Persons linked to housing units at listed address in sample areas
$\bar{Y}_C$ Sampling Frame
Multistage area probability sample of persons in sample HH
$\bar{y}_a$ Sample
Persons providing answers to interviewer-administered questionnaire
$\bar{y}_r$ Respondents
$\bar{y}_{rw}$ Adjustments
$\bar{y}_{prw}$ Survey Statistic

Values of statistic computed based only on respondent data differ from those based on entire sample

$$B(\bar{y}_R) \approx \left(\frac{\sigma_{y\rho}}{\bar{\rho}}\right)$$



Examples of open research questions …

- Can we still get design-unbiased estimates with DP?
- What sample size do we need to maintain desired precision with DP? Can we afford that?
- What do

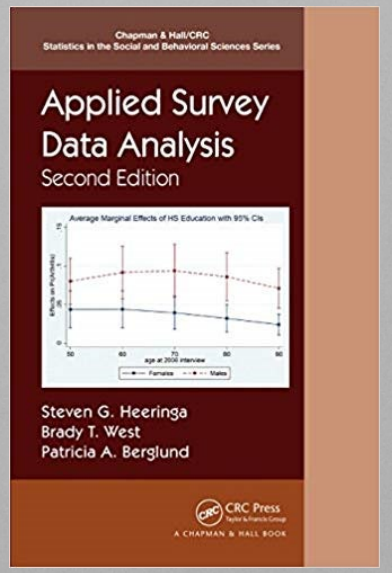
  interviewers

  longitudinal data

  households do to the DP analysis?
- Can we justify the costs if we limit access via privacy budget?

Data Analyis I

Research Questions, Workflow, Informative outliers



Typical Research Questions

"What is the relation between experiences of discrimination and the risk of PTSD among African American adults?"

Sibrava et al. 2019, American Psychologist

"How does being integrated with poor students affect the social behaviors and academic outcomes of rich students?"

Rao, 2019, American Economic Review

"Does high immigration increase inter-ethnic tension?"

Weber, 2018, European Sociological Review

Typical methods

- Psychology
  - A/B experiments
  - ANOVA
- Economics
  - Linear regression
  - Time series analysis
- Sociology
  - Multilevel (random effects) linear regression models
- Demography
  - Time to event (survival) models

....

Often data collected for other purposes (research or gov. statistics)

- ICPSR data archive in the U.S.
- Essex - UK data archive
- GESIS – data archive, Germany

Table 4 Risk factors for depression among immigrants and ethnic minorities in Europe: country level effects (random slopes) (weighted data)

	Model 1			Model 2			Model 3			Model 4		
	P.E.	S.E.	*p*	P.E.	S.E.	*p*	P.E.	S.E.	*p*	P.E.	S.E.	*p*
Parameter variance												
Between countries	1.60	0.50	***	1.58	0.50	***	1.56	0.49	***	1.88	0.63	**
Within countries	14.67	0.11	***	14.66	0.11	***	14.64	0.11	***	13.20	0.10	***
Variance components												
Ethnic minority	0.37	0.19	*	0.36	0.19	*	0.26	0.15		0.21	0.14	
First generation	0.11	0.09		0.02	0.08		0.01	0.08		0.00	0.08	
Gender	0.10	0.04	*	0.10	0.04	*	0.10	0.04	*	0.07	0.03	*
Partner	0.13	0.05	**	0.13	0.05	*	0.13	0.05	*	0.08	0.04	*
21–35 years	0.26	0.10	**	0.25	0.10	**	0.25	0.10	**	0.12	0.06	*
50–64 years	0.13	0.06	*	0.13	0.06	*	0.13	0.06	*	0.04	0.03	
65 years or older	0.80	0.26	**	0.79	0.26	**	0.79	0.26	**	0.28	0.12	
Outside Europe				0.28	0.18		0.26	0.18		0.15	0.14	
Ethnic discrimination							0.24	0.20		0.18	0.18	
Finding it very difficult										1.01	0.39	**
Finding it difficult										0.11	0.05	*
Student										0.05	0.14	
Unemployed										0.37	0.19	
Sick/handicapped										0.28	0.22	
Pension										0.23	0.10	*
Other										0.04	0.04	
Years of education										0.00	0.00	**

Source: European Social Survey, 3rd round, own calculations

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, Wald Z test

1. Focus on coefficients (parameters) and standard errors (statistical sampling variation)

2. Model exploration is part of the analysis

3. Mix of categorical and continuous variables

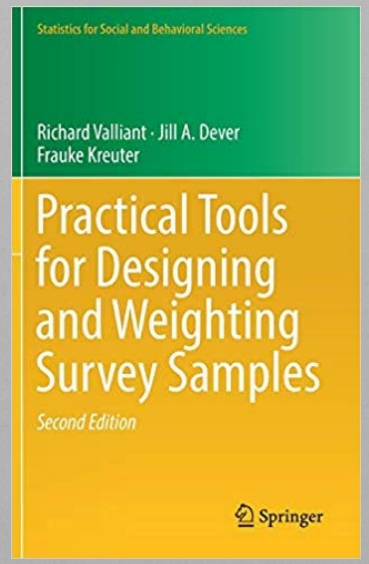
4. Complex data structures

Examples of open research questions …

- Can microdata be used for iterative analysis processes?
- Can <INSERT FAVORITE METHOD> be used?
- Can we learn enough about coefficients?
- Can outliers still lead to insights?
- What about responsive design /
  predictive (policy) intervention applications?

Data Analysis II

Variance Estimates, Weights and Benchmarks



Variance Estimation

Exact formulas

- Only possible for ‘linear’ estimators

A *linear* estimator is one that can be written as $\hat{\theta} = \sum_{i \in U} \delta_i \alpha_i y_i$ where α_i is a constant. The value for element i is the same regardless of the set of sample units that is selected. δ_i indicates whether unit i is in sample or not (0 or 1)

Examples of linear estimators: totals, means of the form $\bar{y} = \sum_{i \in s} \alpha_i y_i / N$

Examples of nonlinear estimators: ratio of means, log(mean), median, regression coefficients

Linearization (Taylor series)

- Used for ‘nonlinear’ estimators

Replication

- Applies to linear and nonlinear estimators

In practice often ignored! [West, Sakshaug, Aurelien 2016, PlusOne]

Replicate Weights

General procedure

1. Dived full sample into subsamples (replicates)
2. Repeat weight computation for each subsample
   - base weight
   - adjustment for subsampling
   - nonresponse adjustment
   - calibration
3. Each sample element has a full sample weight and a series of replicate weights
4. Uses receive (large) file with data and all weights

GREG – Nonresponse Adjustment

Categorical and continuous variables can be used

Estimator of total is

$$\hat{T}_{yGREG} = \hat{t}_y + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^T \hat{\mathbf{B}}$$

$\hat{t}_y$ is est'd total using input weights (base or NR-adjusted)

$\mathbf{t}_\mathbf{x}$ is vector of pop totals of x's

$\hat{\mathbf{t}}_x$ is vector of estimated pop totals of x's using input weights

$\hat{\mathbf{B}}$ is (input weighted) slope of y on $\mathbf{x}$

Underlying model for GREG is

$y_i = \mathbf{x}^T\beta + \varepsilon_i,\ \varepsilon_i \sim (0, v_i)$

$$\begin{aligned} \hat{T}_{yGREG} &= \sum_{i\in s}\left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^T \left(\mathbf{X}^T\mathbf{D}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{x}_i/v_i\right] d_i y_i \\ &= \sum_{i\in s} g_i d_i y_i \end{aligned}$$

Examples of open research questions …

- What about replicate weights in DP?
- What if population benchmarks for nonresponse adjustments are DP? [Dever & Valliant 2016, JSSM; Lee & Valliant 2015, JOS; Liao & Valliant 2012]
- How do get fixed privacy guarantee with need for valid variance estimation (multiple synthetic data sets)


Berkeley University 2018

Example: IAB – SMART [Kreuter et al. 2019]

Sample of households with at least one welfare benefit recipient (at reference date)

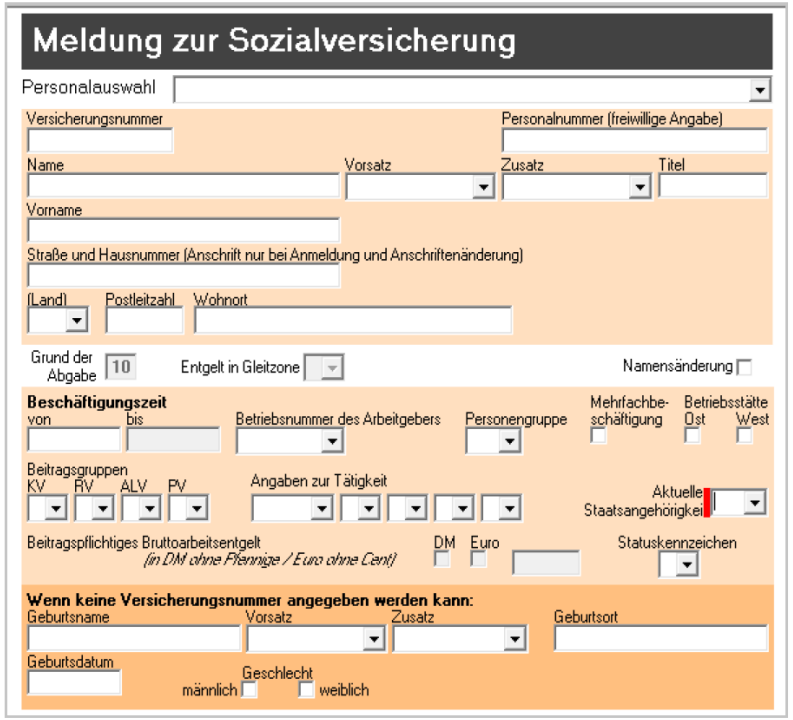
Refreshed annually

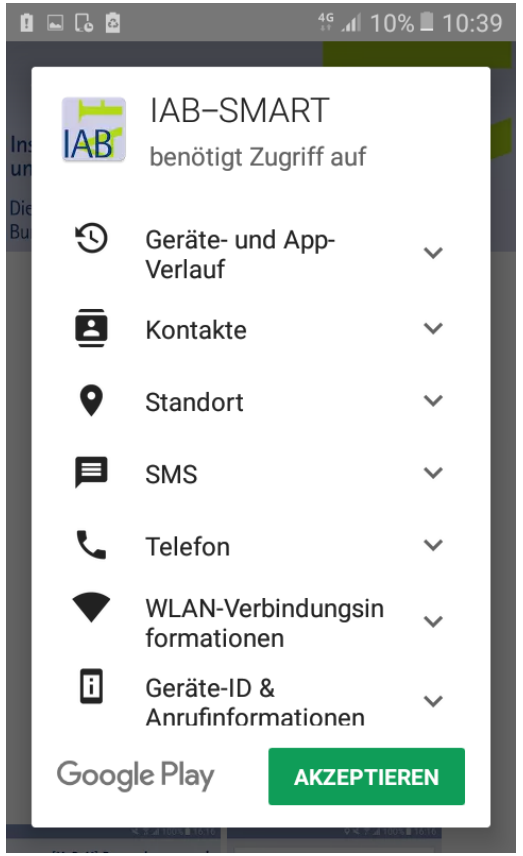
Surveyed annually

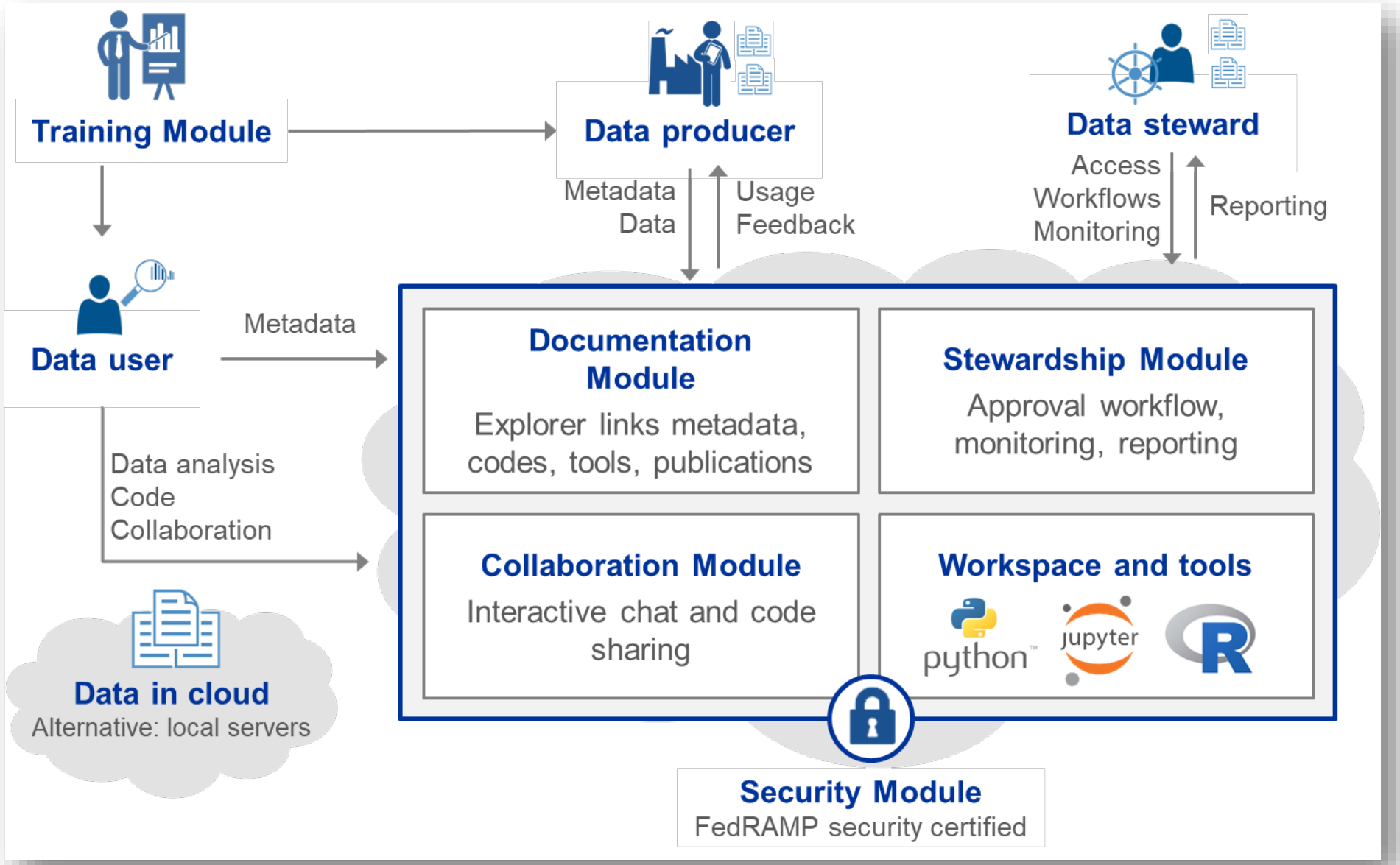
Random household sample of resident population

Refreshed annually

Surveyed annually






Training Module
Data producer
Data steward
Metadata
Data
Usage
Feedback
Access
Workflows
Monitoring
Reporting
Data user
Metadata
Documentation Module
Explorer links metadata, codes, tools, publications
Stewardship Module
Approval workflow, monitoring, reporting
Data analysis
Code
Collaboration
Collaboration Module
Interactive chat and code sharing
Workspace and tools
python
jupyter
Data in cloud
Alternative: local servers
Security Module
FedRAMP security certified

Examples of open research questions …

- Focus on input or output control?
- How to automate either or both?
- How to automate the rich context documentation?
- How to scale the use?