Combining Data from Survey and Non-Survey Sources: Challenges and Opportunities

Sharon Lohr, Westat
May 6, 2016
sharonlohr@westat.com

The views presented in this talk are those of the author and do not represent the views of any Federal Government agency/department or Westat.

Outline

- Why combine survey with non-survey data?
- Challenges for combining
- Statistical methods
- Research opportunities
- Case Study: Model health care costs
  - MEPS, MCBS, NHIS, NHANES, other surveys
  - Medicare claims, provider data, prescription prices, …

Why Combine With Other Sources?

- Probability samples
  - Cost ↗
  - Nonresponse rates ↗
- Cost for other sources ↘
- More demand for
  - Faster statistics
  - Detailed information on subpopulations
- Leverage advantages of each source
- More and cheaper information

Combining Data Sources: Old News

- Design (before survey)
- Calibration (after survey)
- **Assume**
- External source represents population of inference
- Design: frame complete, accurate
- Control totals accurate
  - From same population as survey
  - Variables represent same characteristic
- Calibration model removes bias

Challenges in Combining Data Sources

- Population correspondence among sources
  - Coverage, Respondents, Self-selection
- Variable correspondence
  - Questions / ordering
  - Mode / source / sponsor
- Access to, continued availability of sources
- Transparency
- Inference: does 95% CI have 95% coverage?
- Protecting privacy

Statistical Methods

- Link records
  - Deterministic or probabilistic
  - Accuracy?
  - Protecting privacy
- Imputation
- Multiple frame
- Small area estimation
- Hierarchical models

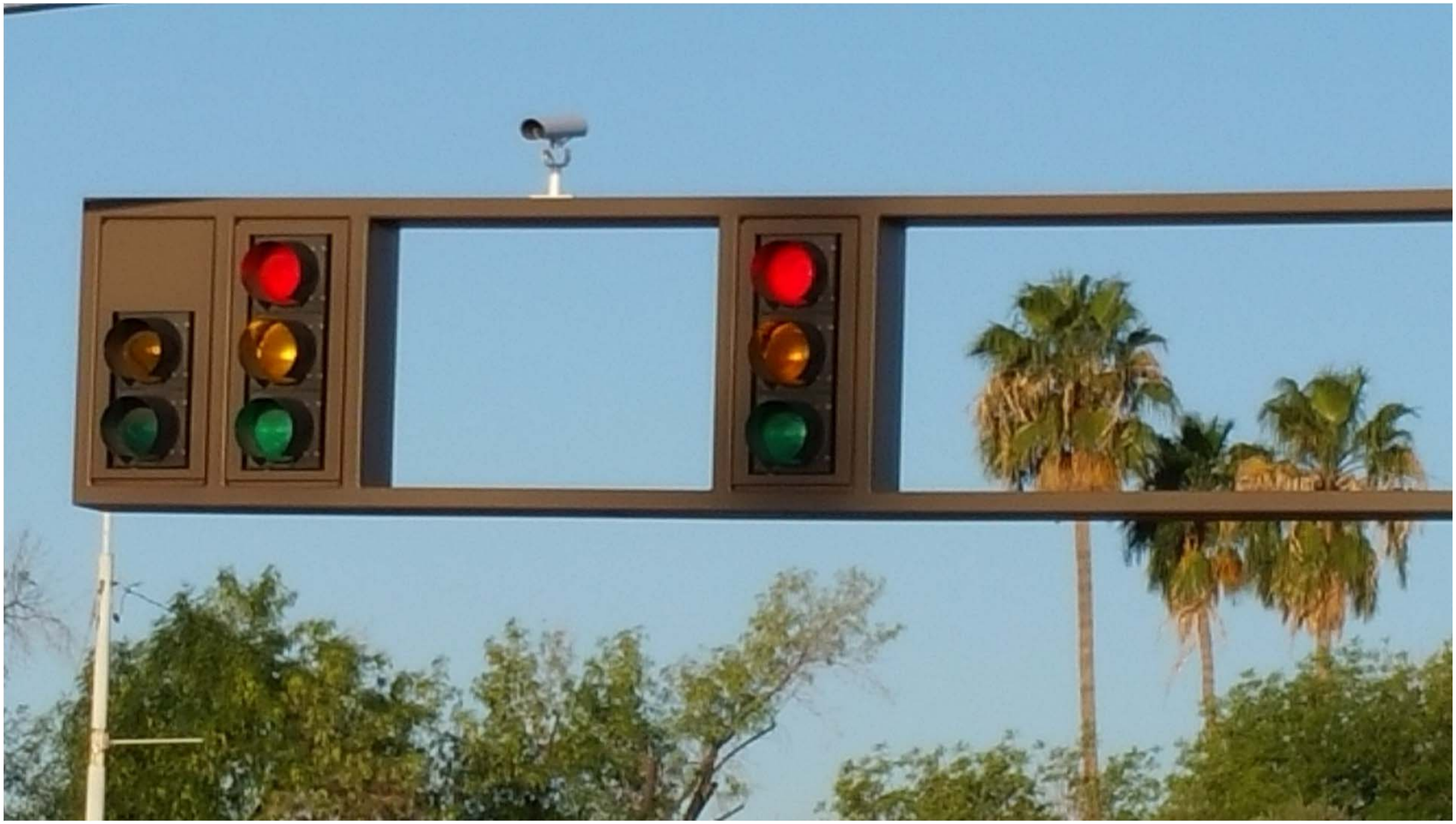
Linking records: Canadian Income Survey

- Instead of asking about all aspects of income …
- “Statistics Canada plans to combine your household’s survey information with tax data. The combined data will be used for statistical purposes only, and will be kept confidential.”
- Calibration to tax record data, demographics

UMETRICS Initiative

- Link grad students who received research funds
  - University administrative data: expenditures
  - W-2 data
  - Survey of Earned Doctorates
  - Proquest dissertation database
  - Longitudinal Household Employer Dynamics
  - Census Business Register

Automated License Plate Readers



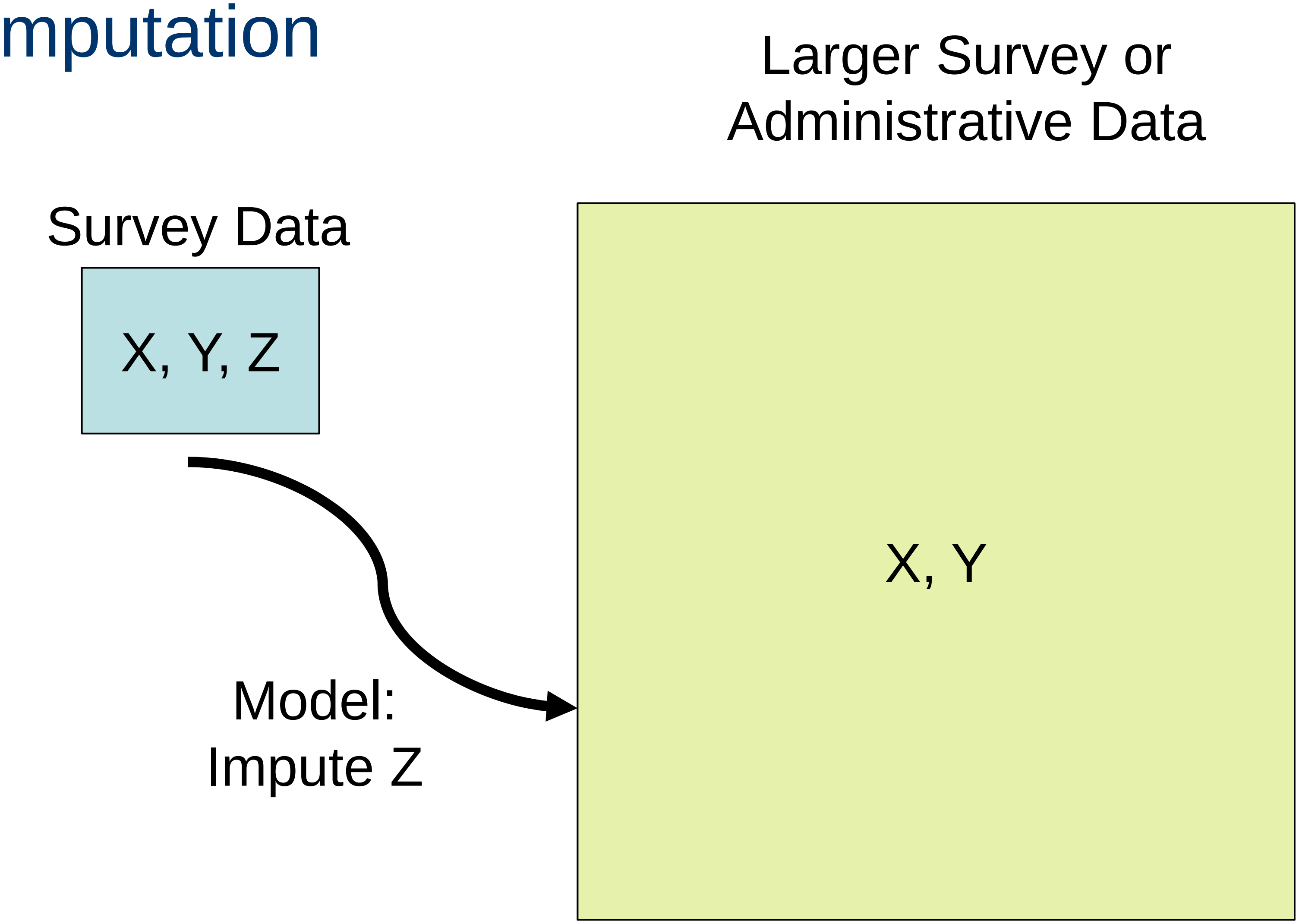
Record Linkage

- Increases number of variables
- Can be used to merge data sources containing different records, augment size of data
- Quality, inference depend on linkage
- Privacy
- Form of imputation

Imputation

- Combining data sources is missing data problem
- Each source is missing observations, variables
- Model-based imputation

Imputation



Larger Survey or Administrative Data

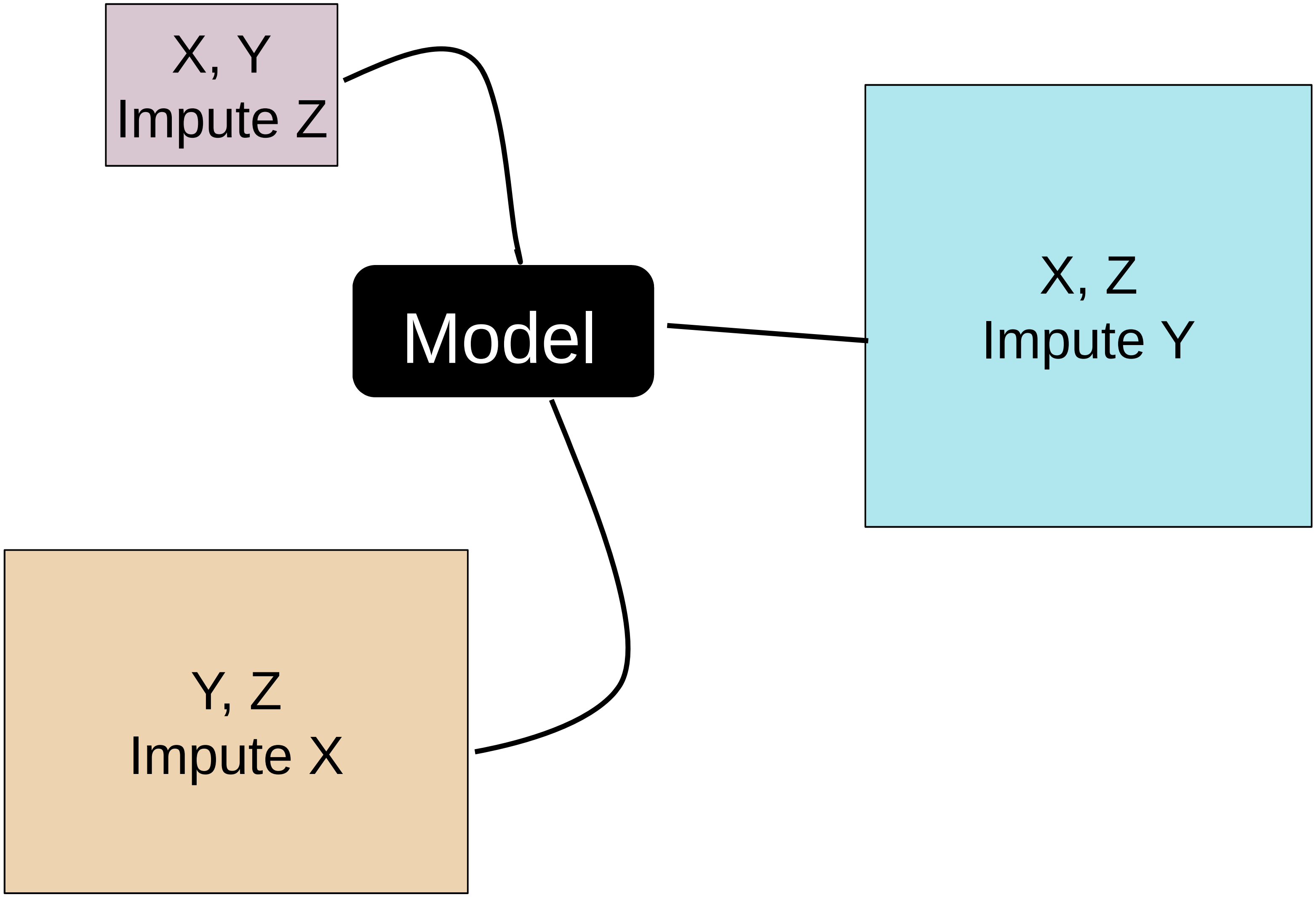
Survey Data

X, Y, Z

X, Y

Model:
Impute Z

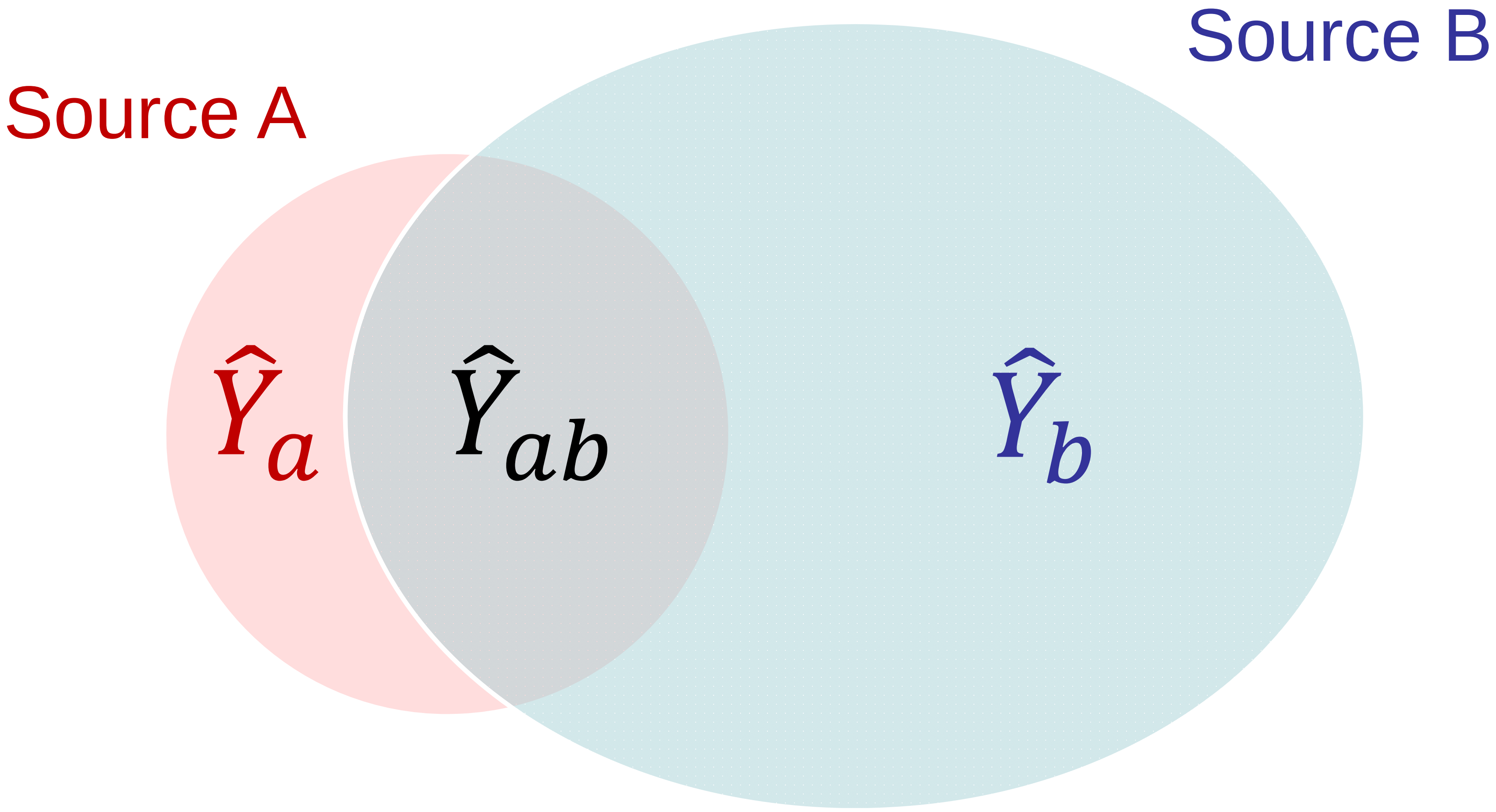
Imputation



Imputation

- Transparency: Can use explicit model
- Variable correspondence depends on model
- Assumes relationship in one source holds for other sources, nonrespondents

Multiple Frame Methods



Source A

Source B

$\hat{Y}_a$

$\hat{Y}_{ab}$

$\hat{Y}_b$

$$\hat{Y}_{ab} = \lambda\hat{Y}_{ab} + (1-\lambda)\hat{Y}_{ab}$$

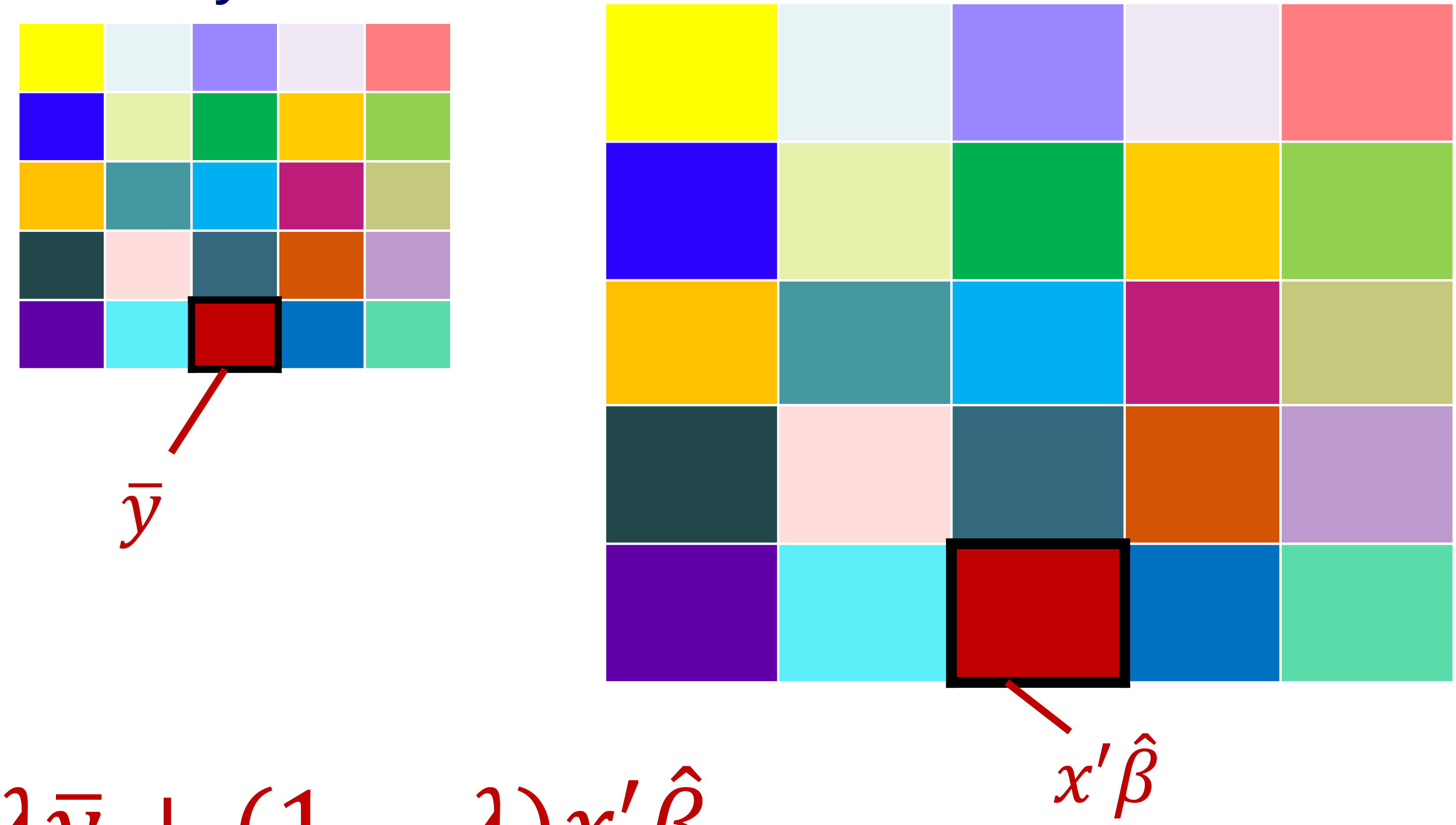
Multiple Frame Methods

- If assumptions met, get
  - Better coverage
  - More data (esp. if source B cheap)
- Most work assumes
  - We know who is in the overlap
  - Variable y is same in both sources

Small Area (Subpopulation) Estimation

Survey Data

Administrative Data



$$\lambda\bar{y} + (1 - \lambda)x'\hat{\beta}$$

Small Area Estimation

- Improves precision (under assumed model) by using administrative data
- Uses summary statistics (area-level model)
- Does model hold for areas where we have no (or very little) survey information?

Hierarchical Bayesian Methods

- Related to meta-analysis in biostatistics
- Model for mean $\bar{y}_{aj}$ in area a, source j:

$$\bar{y}_{aj} = \theta_a + \delta_{aj}$$

random effect

$$\sim N(\mu_j, \tau_j^2)$$

- Lots of variations

Hierarchical Bayesian Methods Can

- Explicitly model bias (but need to define something as unbiased)
- Use prior information on reliability of sources
- Capture between-source differences in standard error
- Use
  - Area-level statistics (use weights in each survey) or
  - Individual data records (nested in sources)

Hierarchical Bayesian Methods

- Strong assumptions on bias, model form
  - Do we have a gold standard source?
- Sensitivity to prior information
- Survey weights, nonresponse, overlap
- Standard errors do not capture model inadequacies

Evaluating Methods

Method	Fit for Use, Timely	Transparency	Accurate Inference	Protect Privacy
Linkage				
Imputation				
Multiple Frame				
Small Area				
Hierarchical Bayes				

Research opportunities

- Design: how do we make use of multiple sources at beginning rather than just for calibration?
- Who is missing in different sources?
- Self-selection issues
- Standard errors that include
  - Nonsampling error
  - Model misspecification
- Metrics for quality of data sources

Research opportunities

- Dynamic estimates
- Use different methods for different subpops
- Combine methods to capture best features
- Privacy protection
- What happens if sources disappear, change?
- Bowley & Burnett-Hurst (1915)
  Livelihood and Poverty
  - 1-in-20 sample of addresses
  - “peculiar safety in the process of averaging”