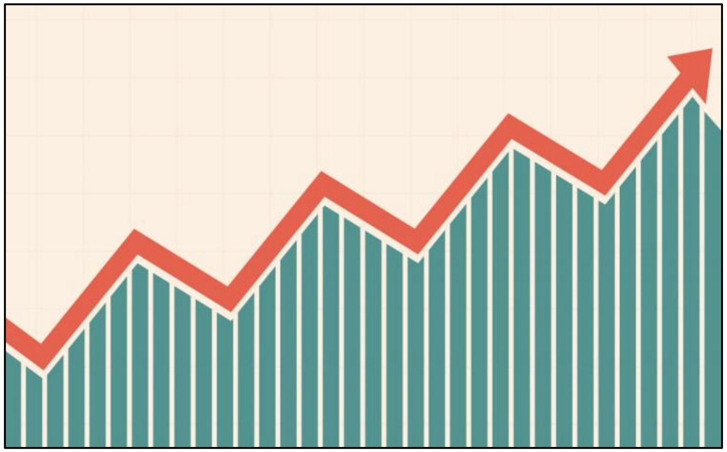
Small Area Estimation and Big Data

International Workshop on SDG Data Disaggregation
28-30 January 2019

Arturo M. Martinez Jr.
Statistician
Statistics and Data Innovation Unit
Economic Research and Regional Cooperation Department
Asian Development Bank

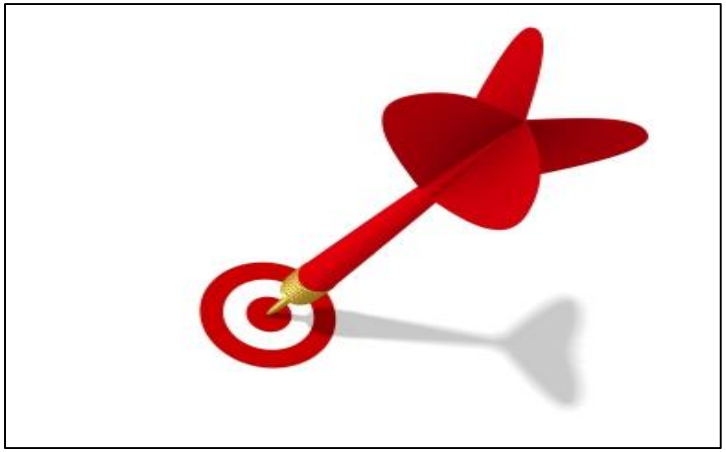
LEAVE NO ONE BEHIND

We measure (development) statistics for various reasons…

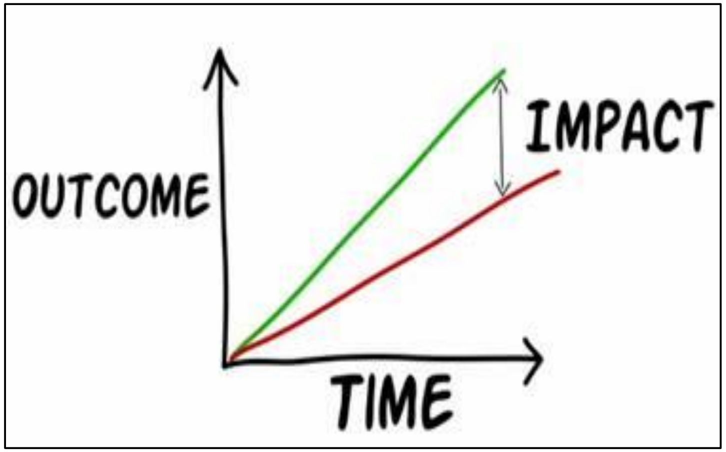


Broad Monitoring

Resource Allocation

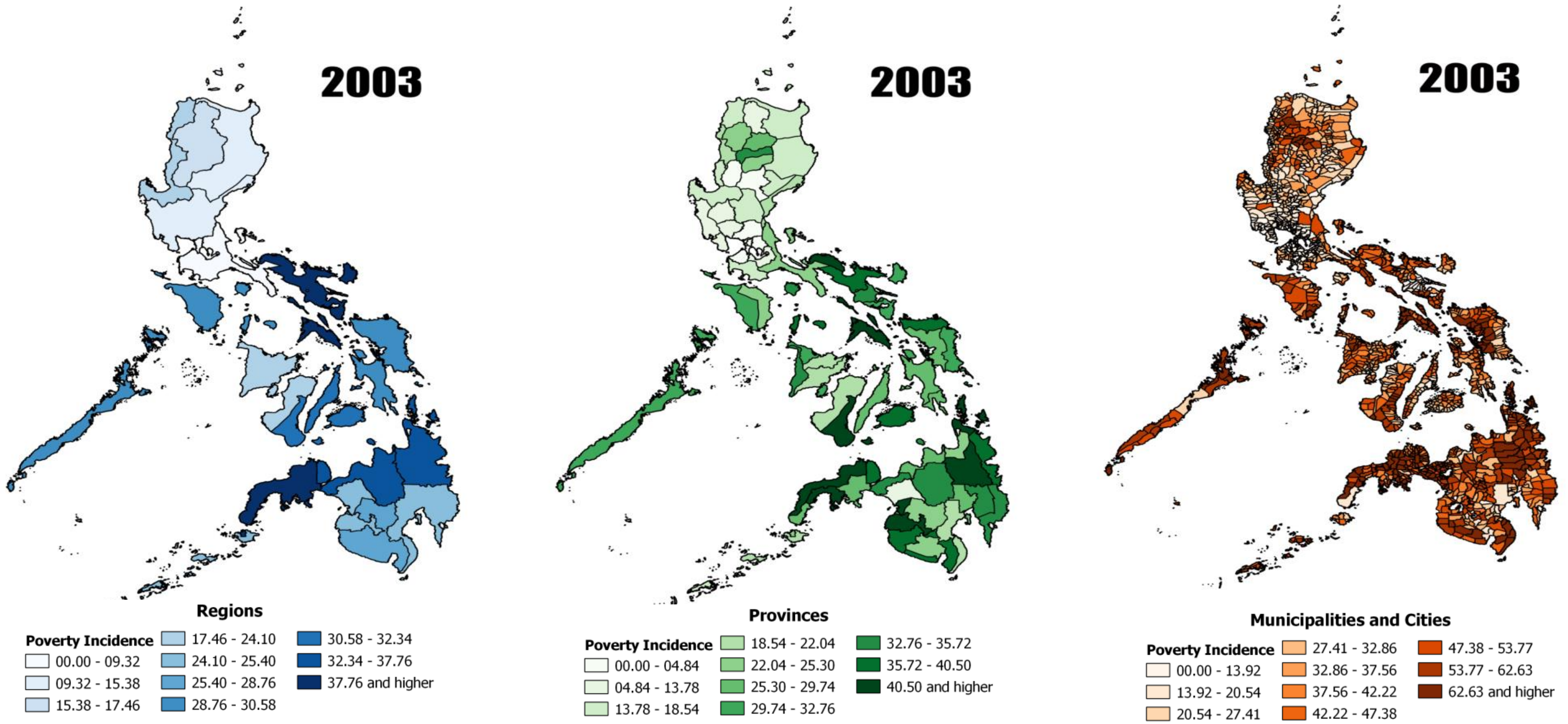


Program Targeting



Impact Evaluation

We need different levels of granularity for different objectives…



Granularity need not be spatial (could be based on population groups).

Increasing demand for granular data

Decentralization of management and governance need granular data

Economic development could be accompanied by more accentuated inequalities and disparities

The same magnitude of poverty reduction can be achieved with a fraction of cost if targeting is done efficiently

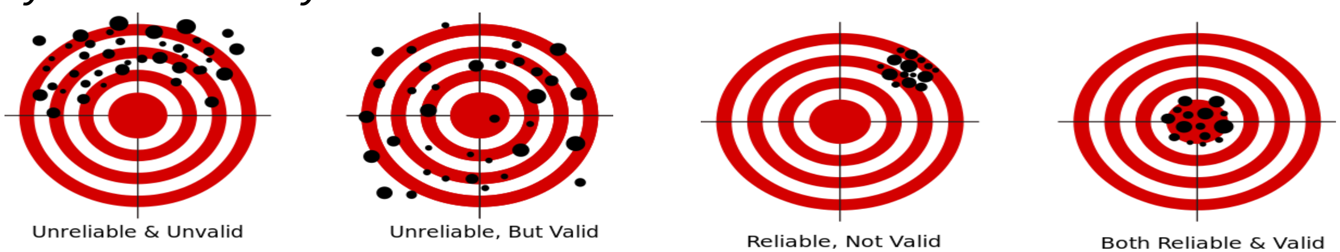
Identification of leading and lagging segments of the population

The desired level of reliability and number of analysis groups have implications on the amount of data needed…

<u>Reviewing survey sampling concepts</u>:

Survey Domain – analytical (population) subgroups for which equally reliable estimates are desired (e.g., region, province, regionxage, etc.)

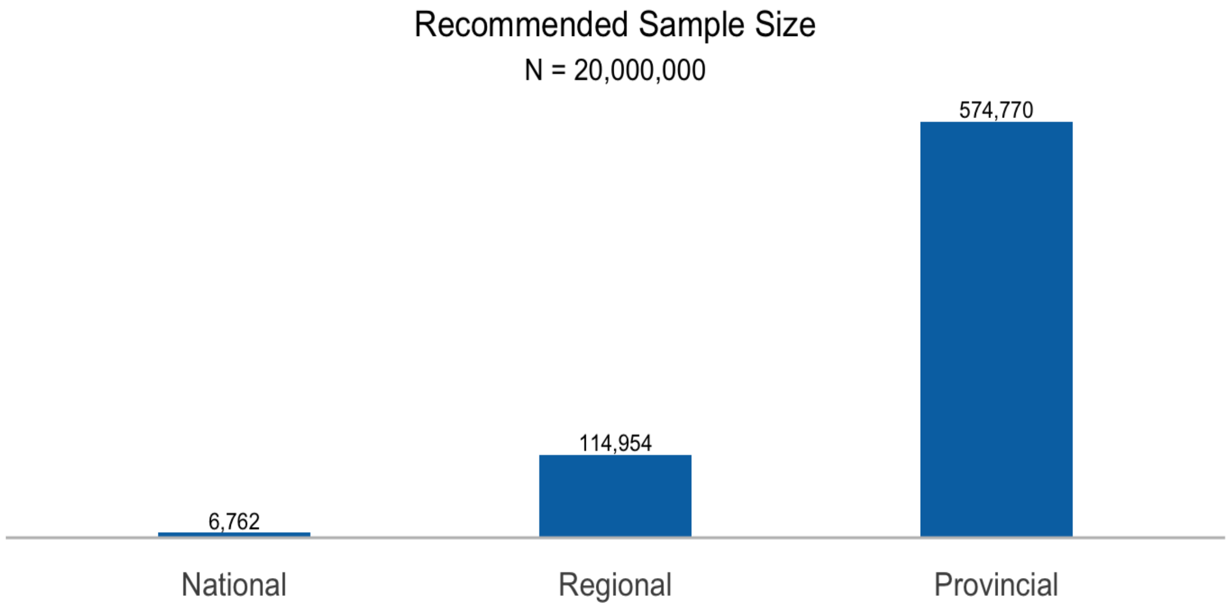
Accuracy and Reliability



Formula for calculating sample size (SRS)

$$n_{srs} = \frac{\frac{t^2_{(\alpha,N-1)}P(1-P)}{d^2}}{1+\frac{1}{N}\left(\frac{t^2_{(\alpha,N-1)}P(1-P)}{d^2}-1\right)}$$

Although having granular data is ideal, particularly in the context of meeting SDG's data requirements, it is not always practical.



The sample size increases (approximately) by a factor equal to the desired number of analysis groups.

The SDGs calls for disaggregated data based on income class, gender, ethnicity, geographic location, migration status, disability status, etc.

How can we reconcile the need for granular data and need to contain costs at manageable levels?

Conduct a survey with comprehensive coverage

- Very costly as it will require large sample size -> need to have sufficient sample size for each small area to be estimated
- Estimates are unbiased by design

Various methodologies

- Statistical techniques using survey data (e.g., rolling approach survey)
- Statistical techniques complementing survey data with auxiliary data (e.g. census, administrative records, etc.)

Small area estimation techniques combine multiple data sources to capitalize on each data source's strengths.

A typical income / consumption or living standards survey collects detailed information that can be used for estimation of our statistic of interest. It also collects other information (e.g., sociodemographic data) which may be considered as correlates of our statistic of interest.

PROS: Collects rich information on various topics
CONS: Reliable at survey domain level

A typical census collects basic information (e.g., sociodemographic data) for all units of the population.

PROS: Collects very granular data
CONS: Limited topics covered

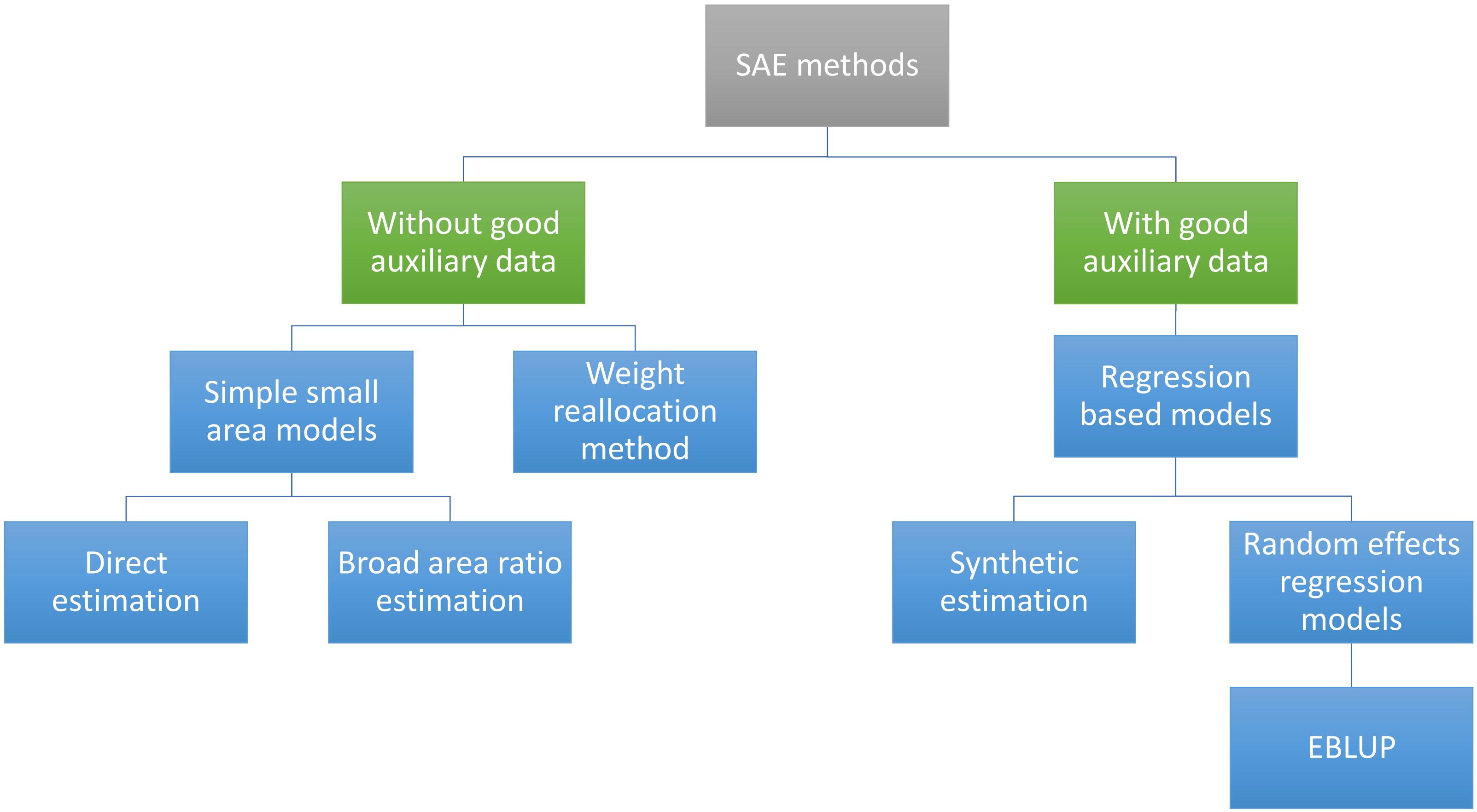
Data Requirements

- Survey data: available for the target variable y and for the independent variable x, related to y
- Auxiliary data (e.g. census, administrative records): available for x but not for y

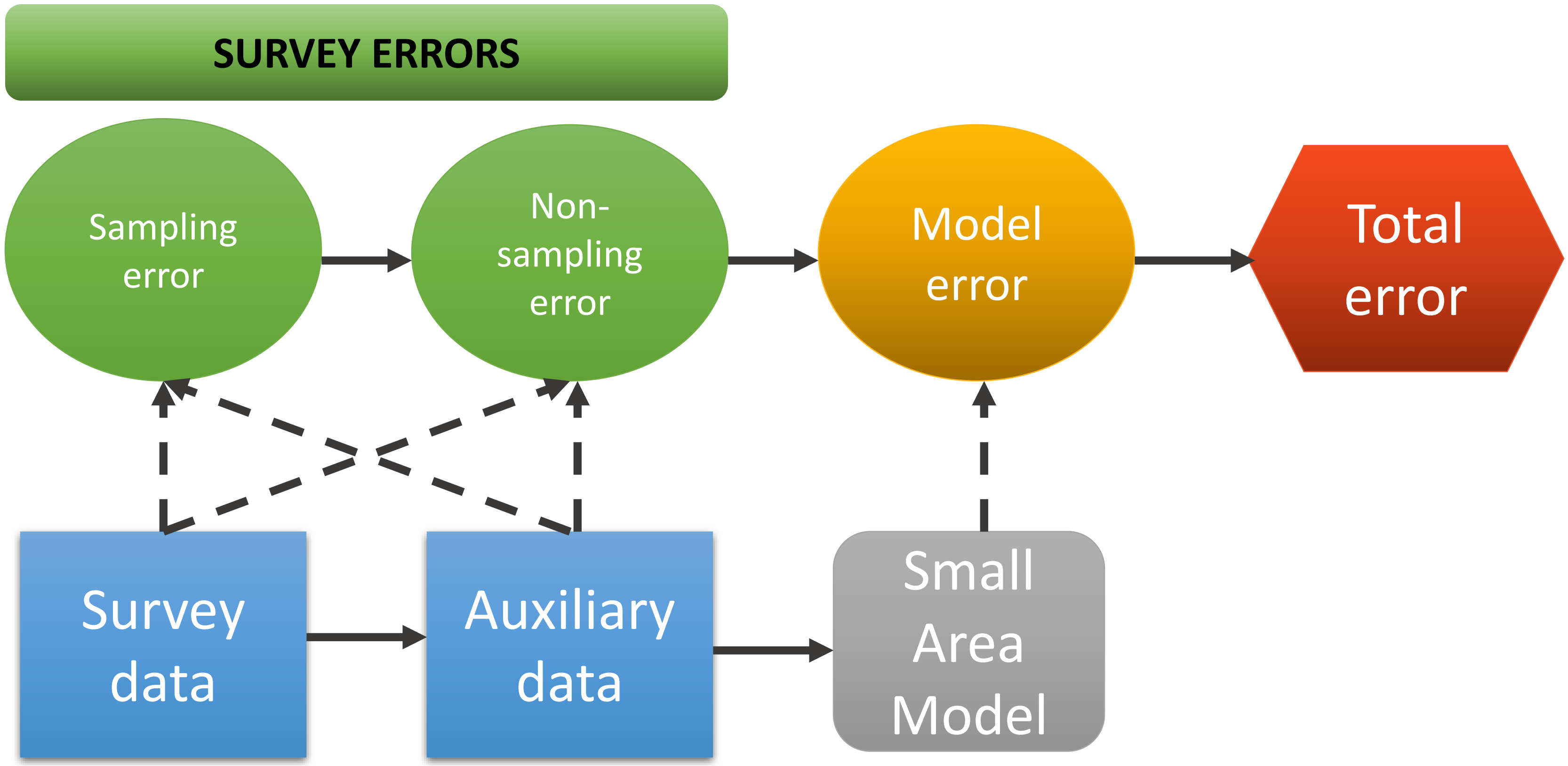
SAE borrows “strength” from auxiliary data through “x”

- Choice of SAE methods depends on availability of auxiliary data (how much auxiliary data is available? at which level is auxiliary data available?)

SAE methods based on data availability



Sources of error



Calibration Method

- Ensures that aggregate statistics are consistent with known population data from census

Weight Reallocation

- Sampled units from other neighboring sub-domains can be used to estimate characteristics for a particular sub-domain, thus "synthetically increasing" the sample size

Illustration of Weight Reallocation

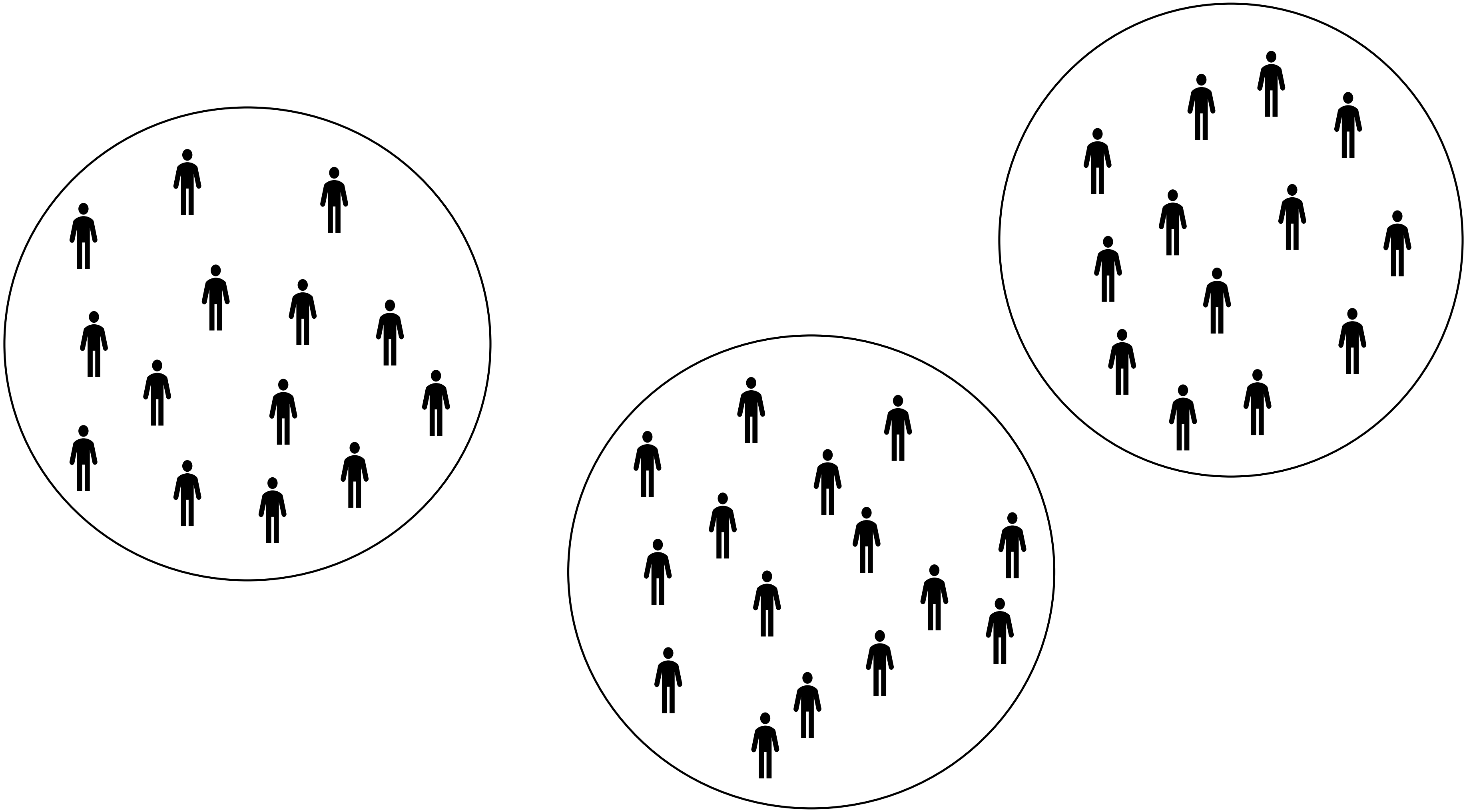
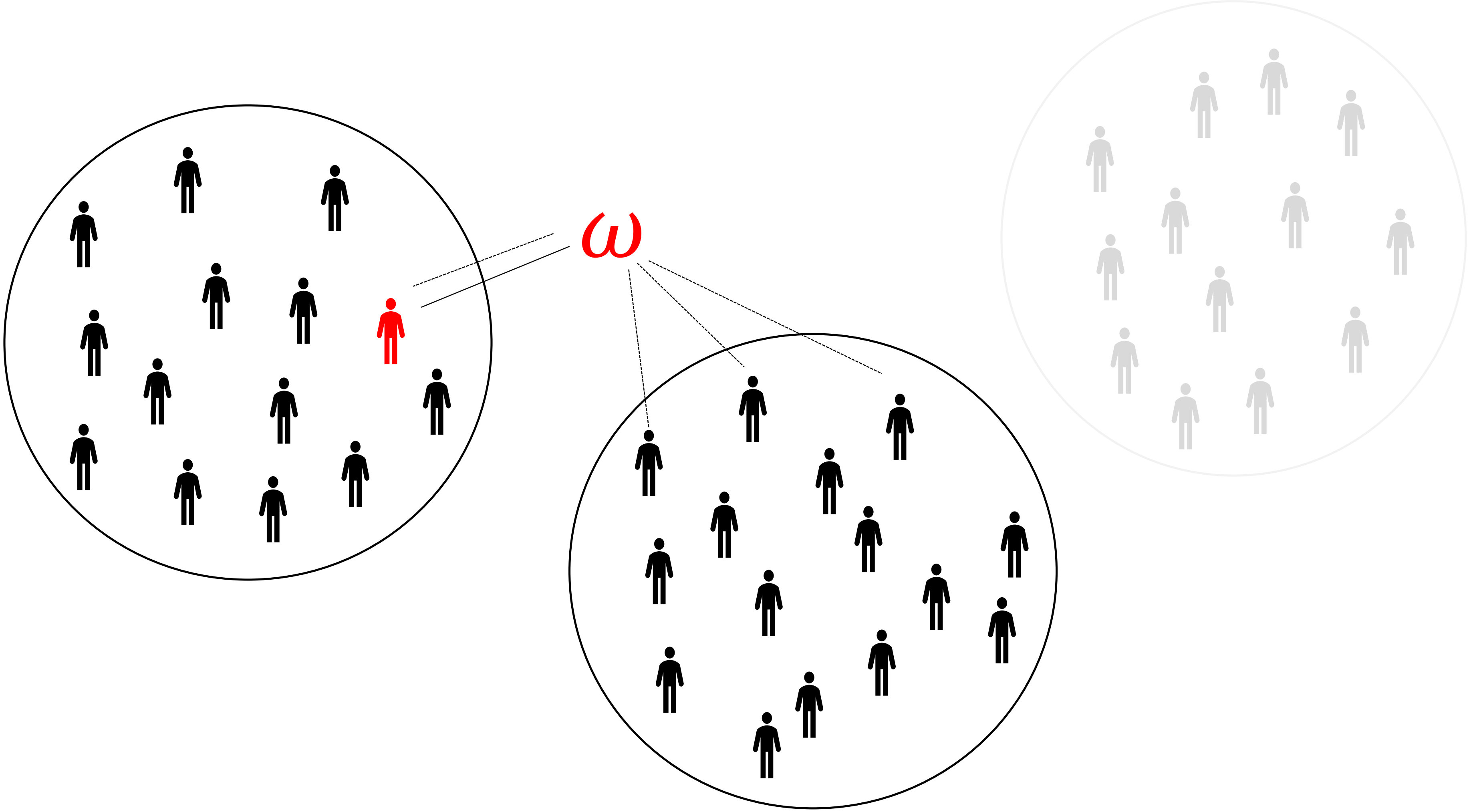


Illustration of Weight Reallocation



SAE without auxiliary data: weight reallocation (Schirm & Zaslavsky, 1997)

Weaknesses

- Subjective selection of what characteristics to preserve
- Computationally intensive
- Assumes that the neighborhood has same level of outcome
- Subject to **both survey and model errors**

Empirical Best Linear Unbiased Prediction

Empirical best linear unbiased predictor (EBLUP)

- combines the direct or design-based unbiased estimator with the regression-synthetic estimator
- generates efficient indirect estimators under the assumed models
- allows validation of the model from the sample data
- produces stable area specific measures of variability associated with the estimates
- allows both unit-level and area-level estimates
- subject to **model error** only

Illustration of Empirical Best Linear Unbiased Prediction

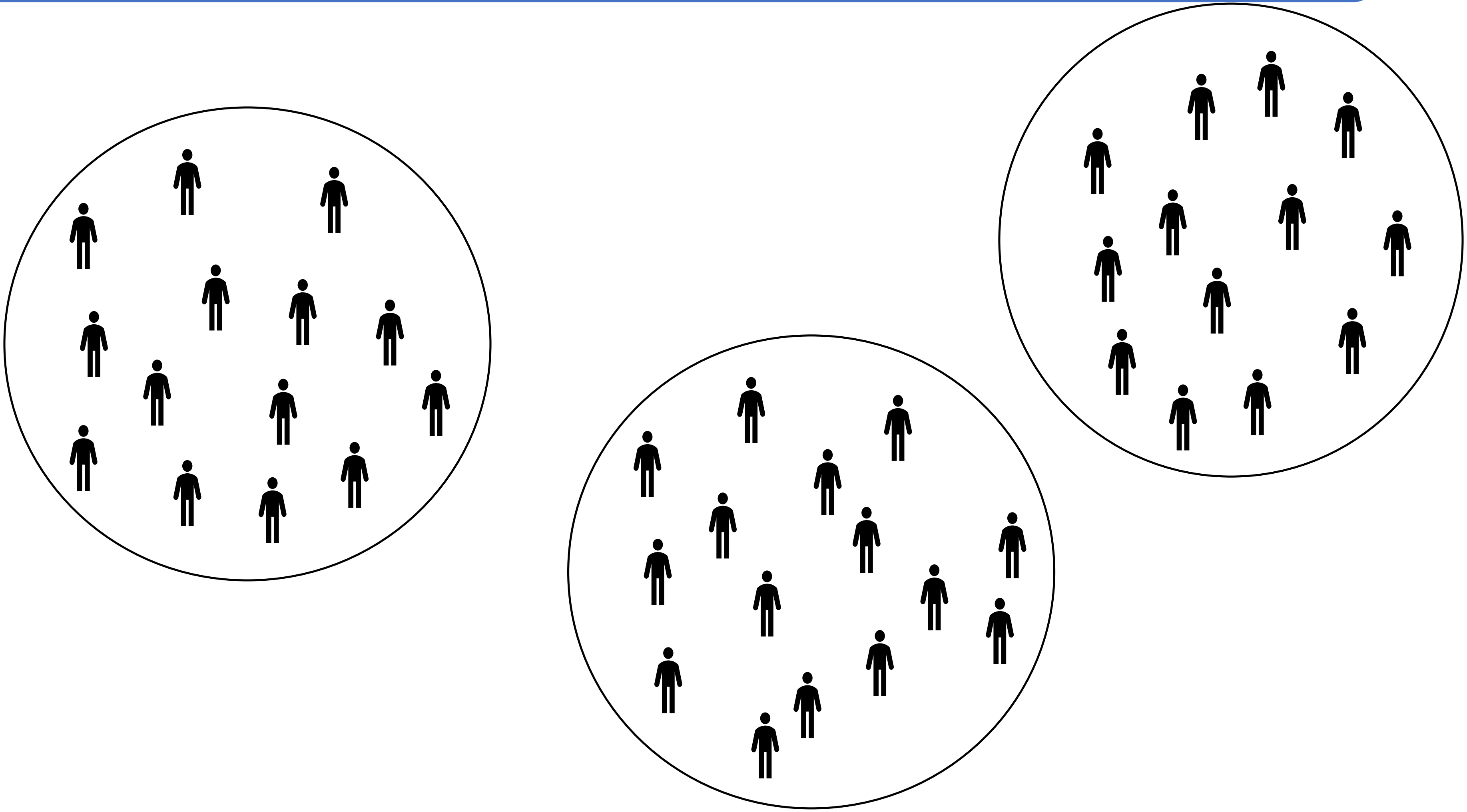


Illustration of EBLUP

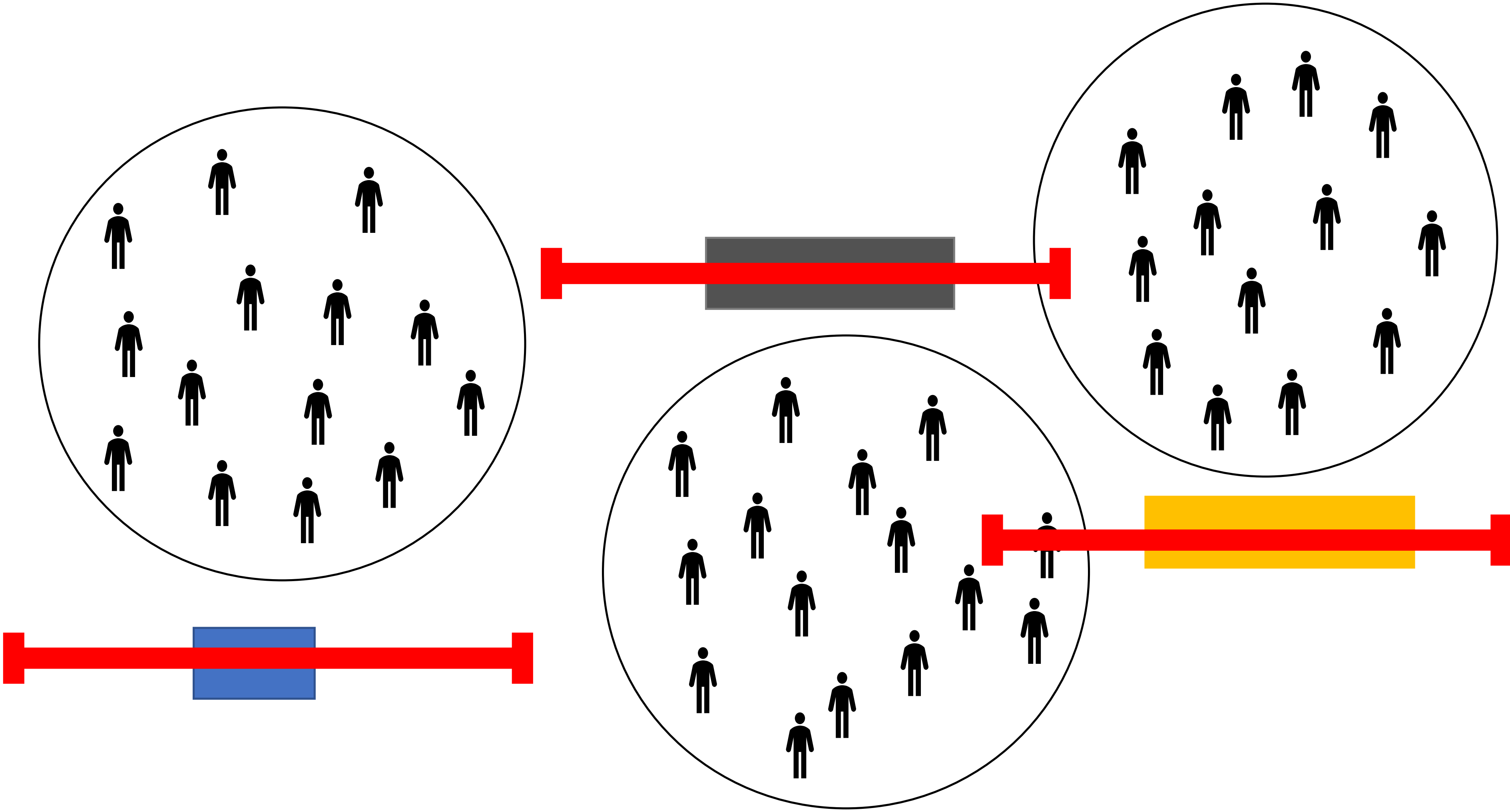
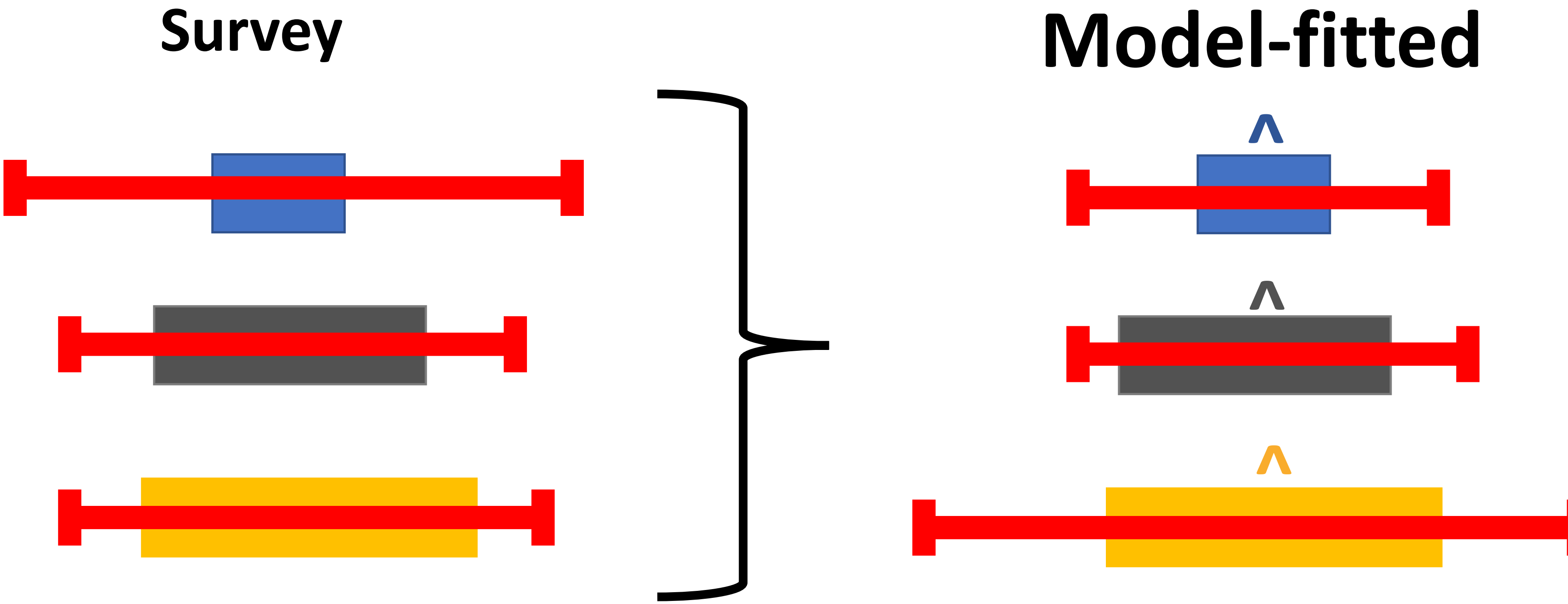


Illustration of EBLUP



$$Y_{survey} = \beta X_{admin} + \varepsilon$$

Empirical Best Linear Unbiased Prediction

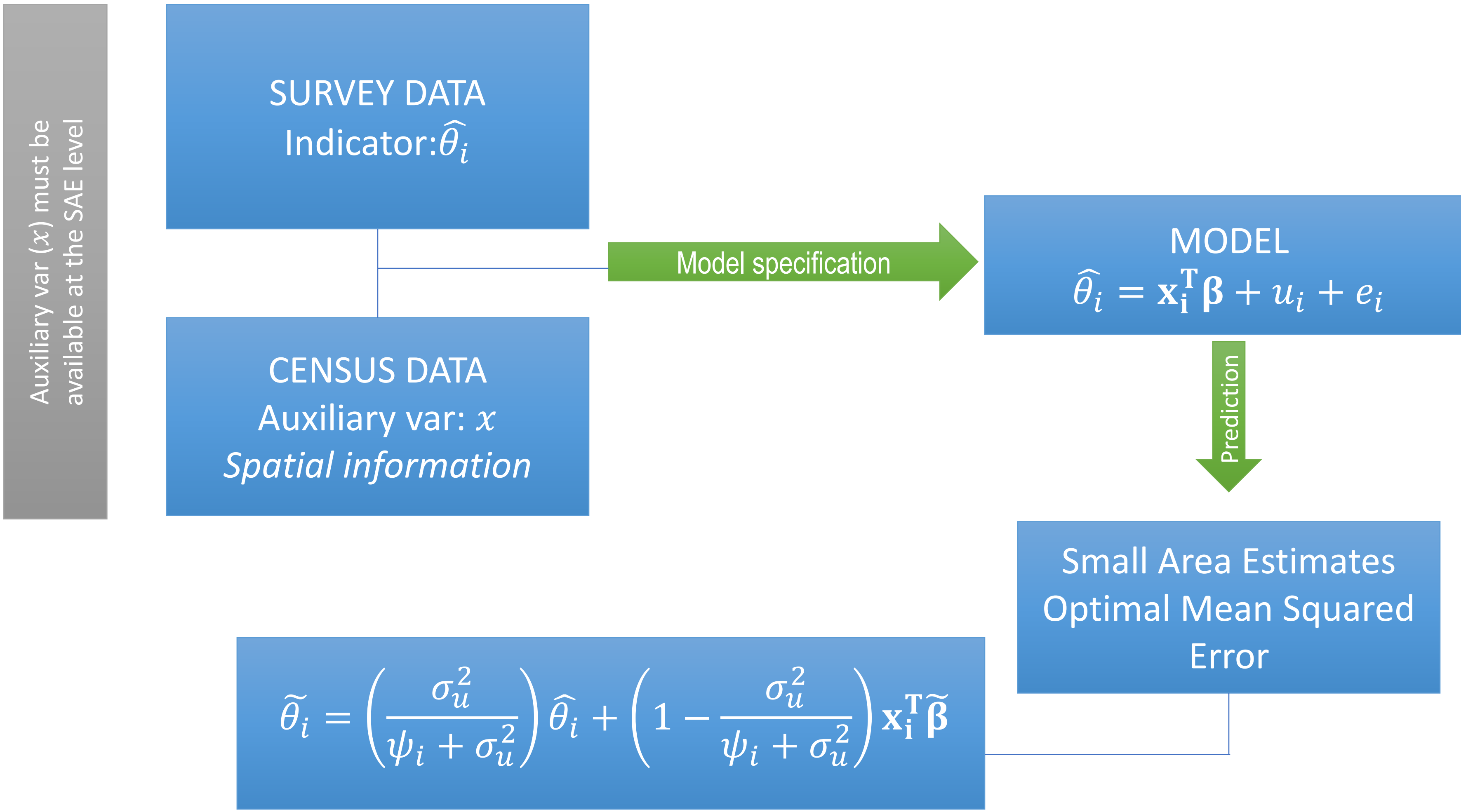
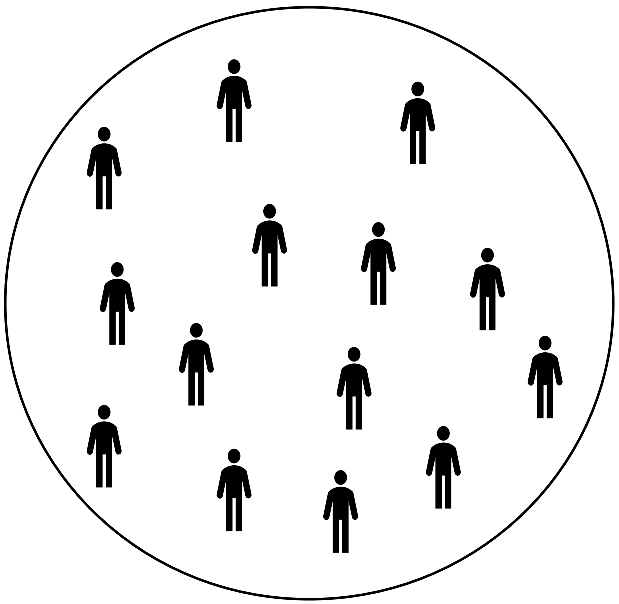
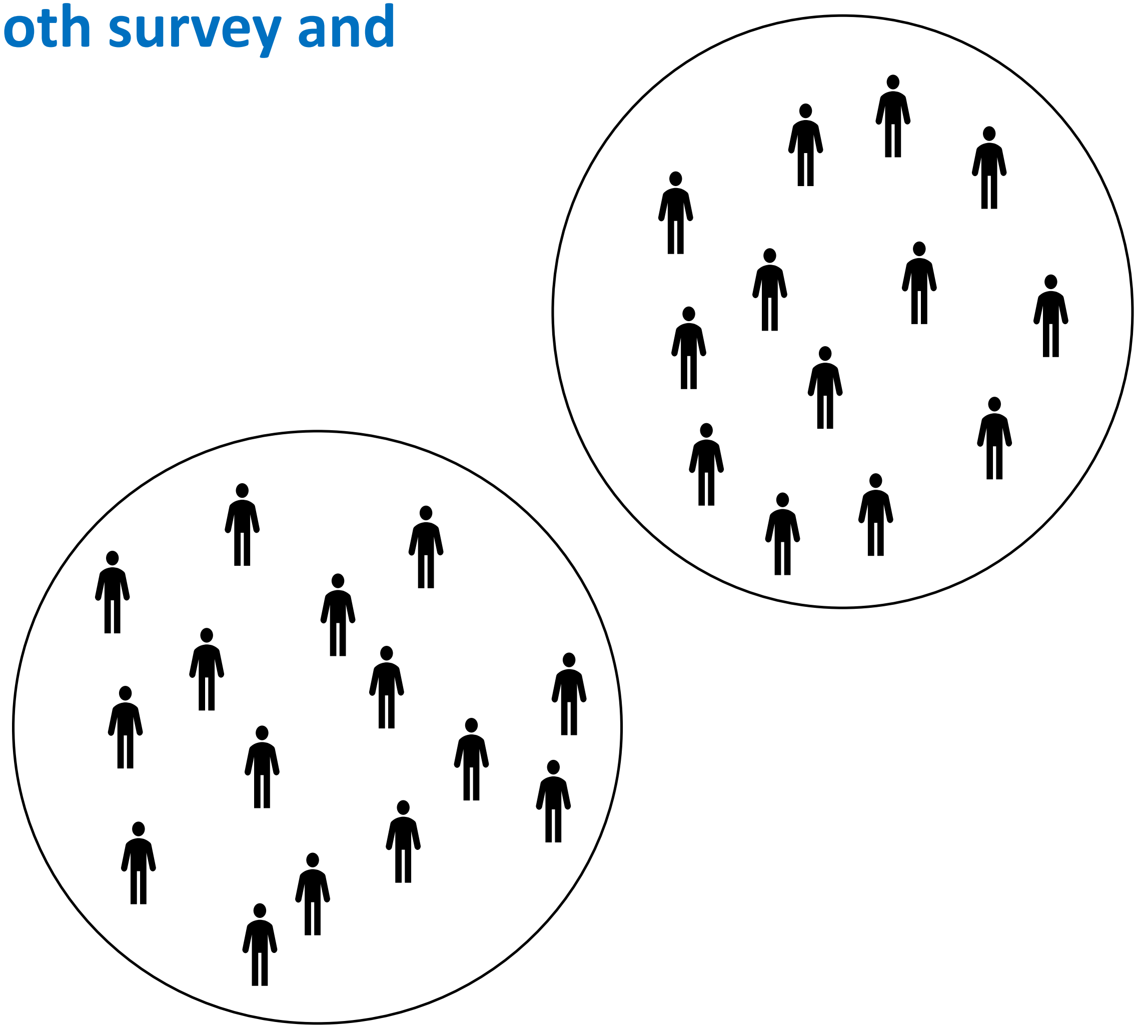


Illustration of WB Poverty Mapping Method

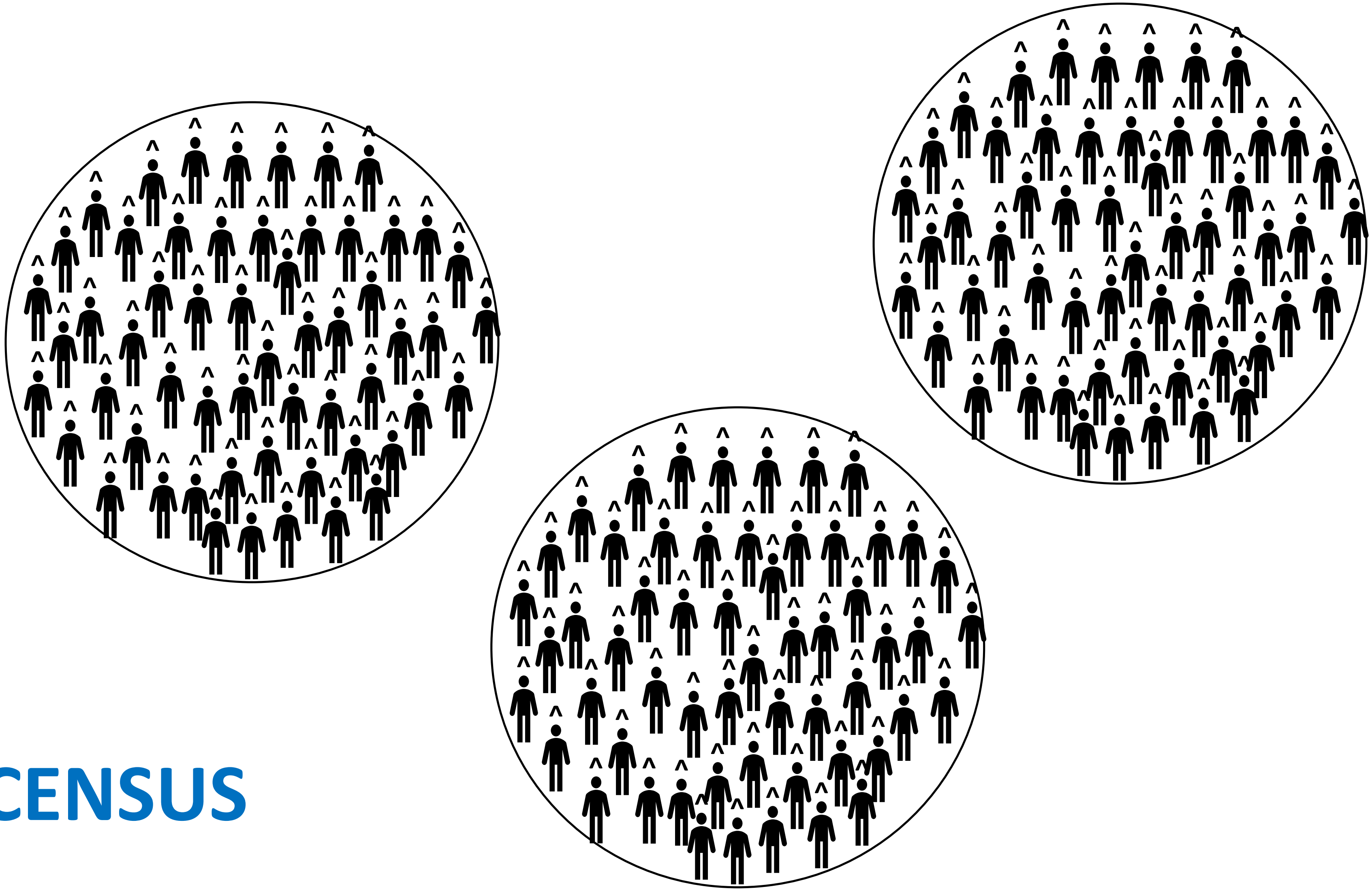
Which X's are available in both survey and census / auxiliary data?





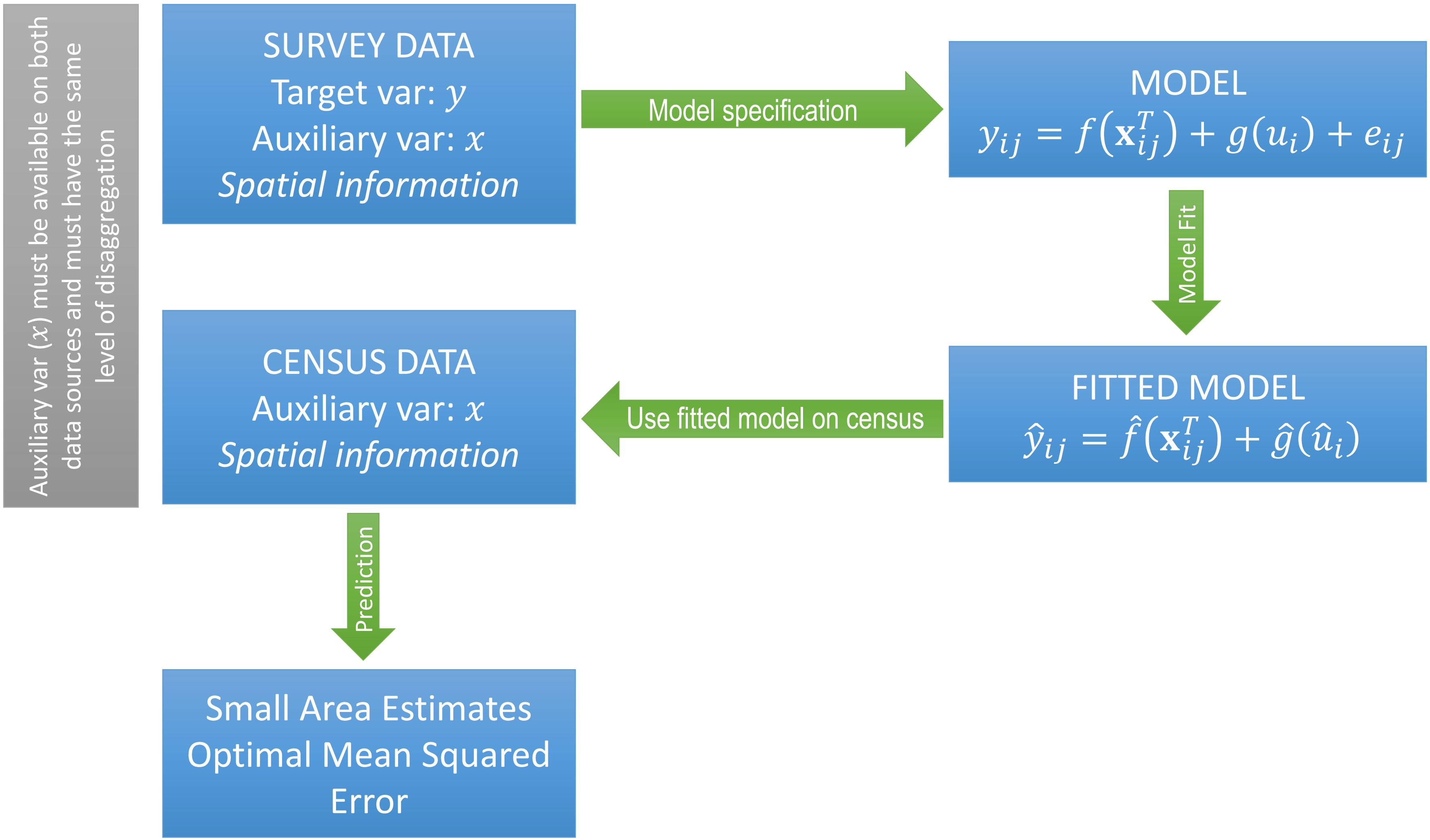
SURVEY

Illustration of WB Poverty Mapping Method



CENSUS

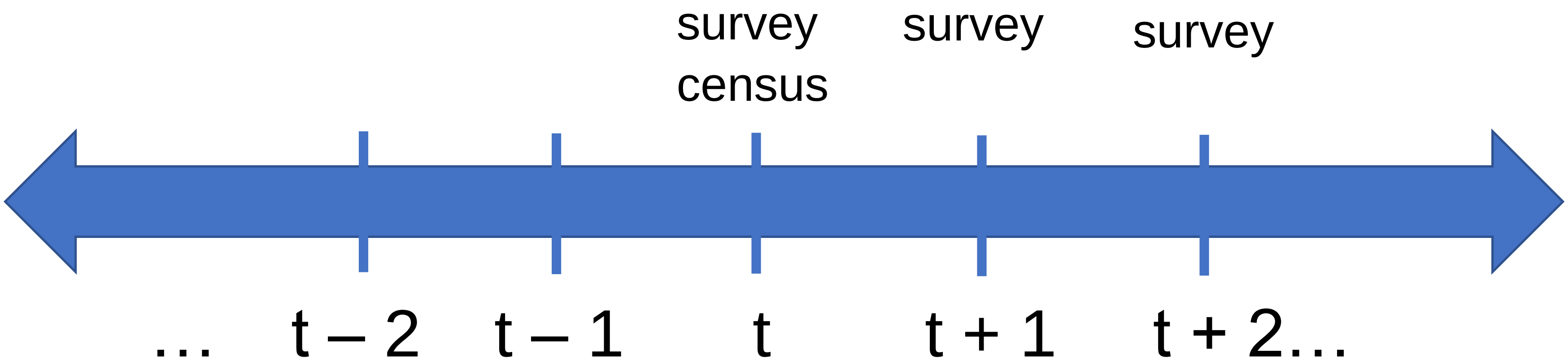
WB Poverty Mapping Method



Measurement Issues

$$Y_{(s)} = X'_{(s)}\beta_{(s)} + \varepsilon_{(s)}$$

$$\hat{Y} = X'_{(c)}\hat{\beta}_{(s)}$$



Measurement Issues

<u>X</u> should be time-invariant (e.g., sex, religion, educational attainment, parental characteristics, etc.)

→ Limited covariates that satisfy time-invariance assumption → Poor model fit
→ Improved reliability but (possibly) lower validity

Where can we get additional auxiliary data?

- 'Big data' generally refers to the type of data arising from people's **digital transactions** with computers, social media, mobile phones, photos, satellite images, sensors, and other types of digital technology.

Category	Source
Exhaust data	Mobile phone data / Financial transactions / Online search and access logs / Administrative data / citizen cards / Postal data
Sensing data	Satellite and UAV imagery / Sensors in cities, transport and homes / Sensors in nature, agriculture and water / Wearable technology / Biometric data
Digital content	Social media data / Web scraping / Participatory sensing / crowdsourcing / Health records / Radio content

Source: Key Indicators for Asia and the Pacific 2016

Finding Auxiliary Information from Innovative Data Sources

Innovative Data Sources

Satellite images, mobile phone records, social media data, and other types of big data

Survey $X_{(s)}$ ← Innovative Data $X_{(id)}$ → Census $X_{(c)}$

$$Y_{(s)} = X'_{(s)}\beta_{(s)} + X'_{(id)}\gamma_{(id)} + \varepsilon_{(s)}$$

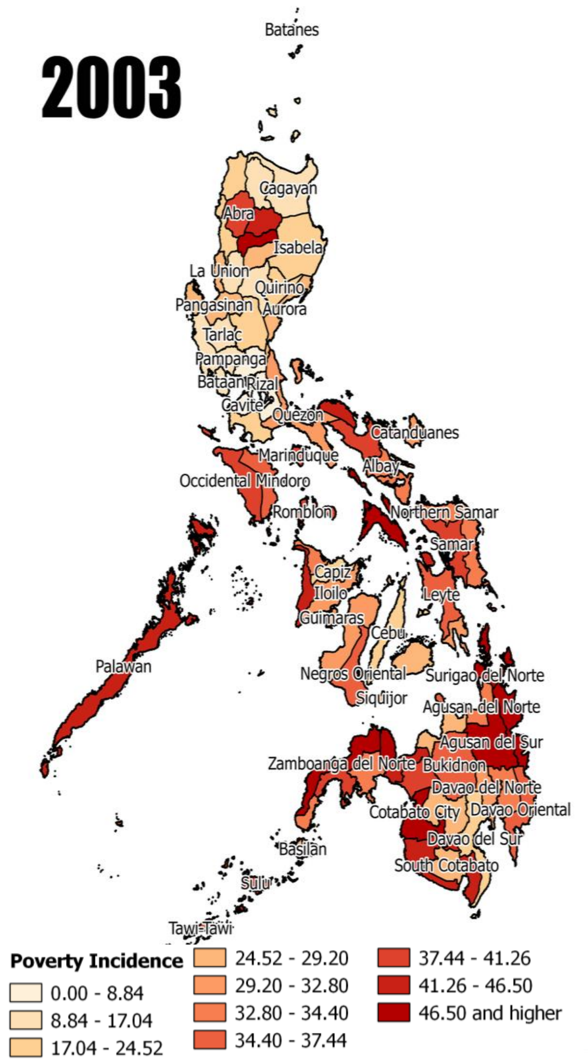
$$\hat{Y} = X'_{(c)}\hat{\beta}_{(s)} + X'_{(id)}\hat{\gamma}_{(id)}$$

What specific types of innovative data source could we include in our SAE model?

Global Distribution of Intensity of Nighttime Lights

1992

Correlation between Poverty Rates and NTL Values



Source: ADB Key Indicators for Asia and the Pacific 2016.

Other potential supplementary sources of info

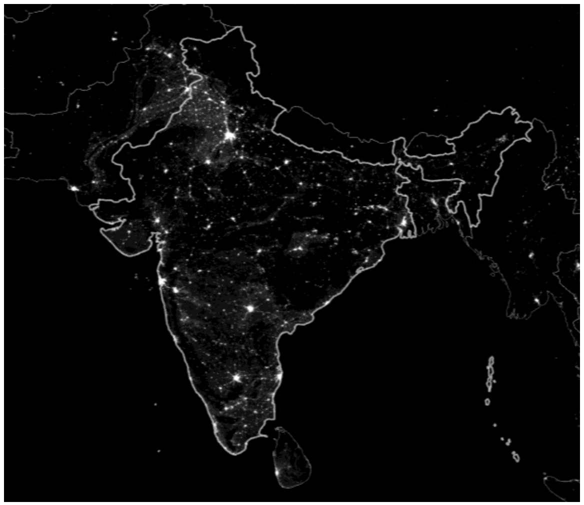


Image Source: DevelopmentSEED


Image Source: Solar Quotation


Image Source: Earth Imaging Journal

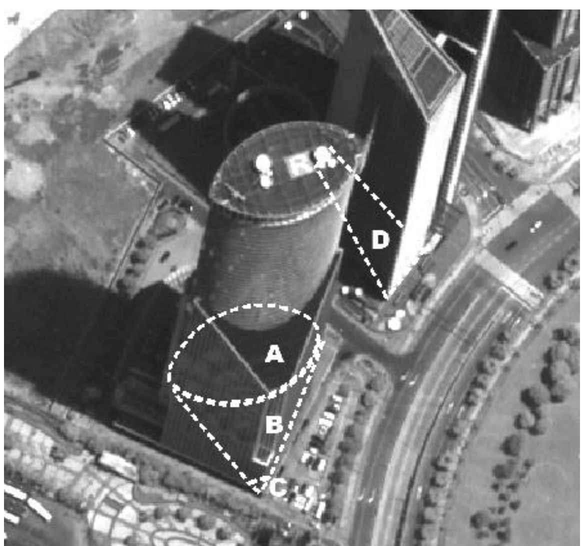


Image Source: Pan et al., 2008

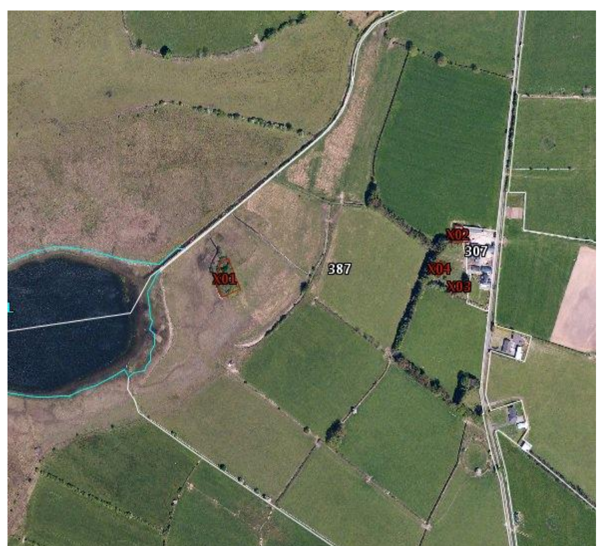


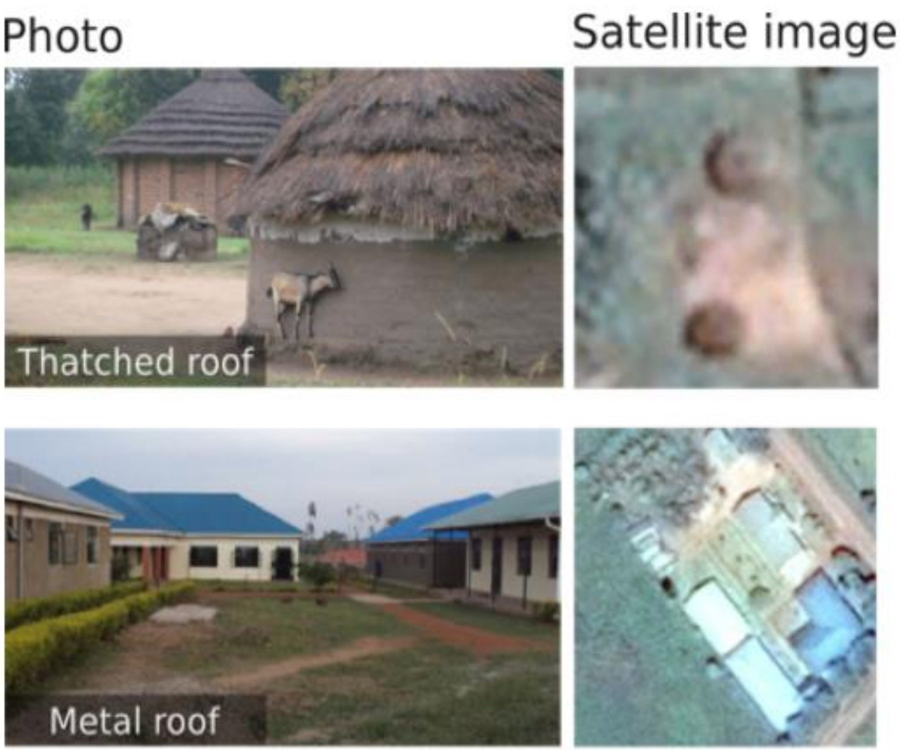
Image Source: AgriLand


Image Source: Wang et al., 2016

Measuring Poverty with Machine Roof Counting

Video source: https://www.youtube.com/watch?v=rXFVejLDGAA

Measuring Poverty with Machine Roof Counting



Source: UN Global Pulse

https://www.unglobalpulse.org/projects/measuring-poverty-machine-roof-counting

Combining Satellite Imagery and Machine Learning to Predict Poverty

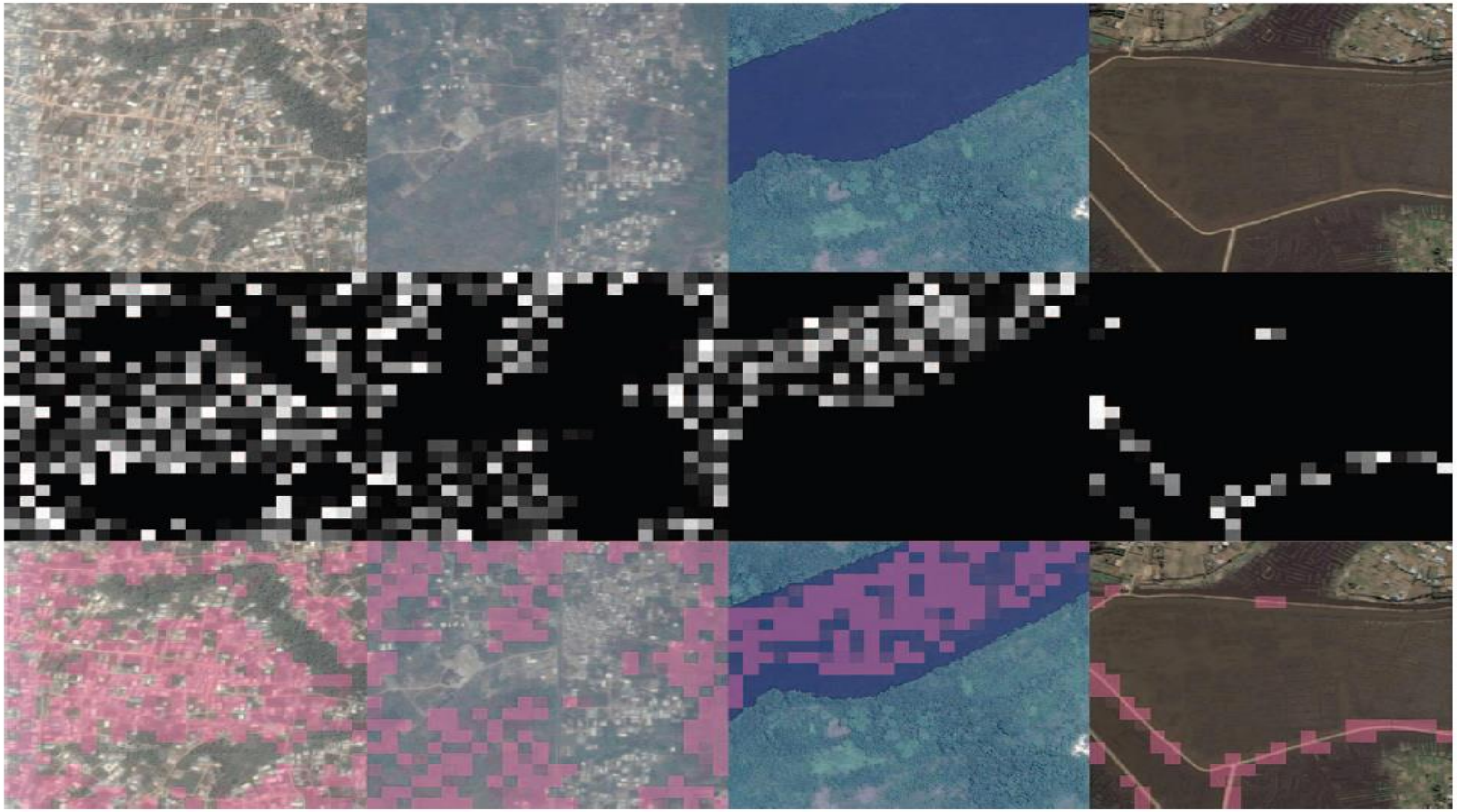


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Source: Jean et al., 2016

Moving Forward…

Do we really need survey and census data to measure poverty? Could we capitalize on AI / machine learning for such purpose?

Science

Home News Journals Topics Careers

Log in | My acc

SHARE

721

0

RESEARCH ARTICLE

Combining satellite imagery and machine learning to predict poverty

Neal Jean[1,2,*], Marshall Burke[3,4,5,*,†], Michael Xie[1], W. Matthew Davis[4], David B. Lobell[3,4], Stefano Ermon[1]
+ See all authors and affiliations

Science 19 Aug 2016:
Vol. 353, Issue 6301, pp. 790-794
DOI: 10.1126/science.aaf7894

Article | Figures & Data | Info & Metrics | eLetters | PDF

Measuring consumption and wealth remotely

Nighttime lighting is a rough proxy for economic wealth, and nighttime maps of the world show that many developing countries are sparsely illuminated. Jean *et al.* combined nighttime maps with high-resolution daytime satellite images (see the Perspective by Blumenstock). With a bit of machine-learning wizardry, the combined images can be converted into accurate estimates of household consumption and assets, both of which are hard to measure in poorer countries. Furthermore, the night- and day-time data are publicly available and nonproprietary.

Science, this issue p. 790; see also p. 753

Moving Forward…

Enhance: Granularity & Timeliness

- ✓ **Channel resources and implement poverty intervention programs more effectively**

ADB’s Data for Development Technical Assistance

Aims to build the capacity of DMCs in compiling disaggregated data for select indicators of the SDGs using combination of traditional and innovative forms of data in accordance with the SDGs’ “leave no one behind” principle’s granular data requirements.

ADB is collaborating with UNESCAP, PARIS21, World Data Lab, and other development partners

ADB's Data for Development Technical Assistance

Country-Specific Case Studies on Data Disaggregation and Big Data Analytics

Issues:

Expanding scope of surveys and other data collection systems is costly

The range of small area estimation techniques is constrained by availability of surveys and census / administrative data

Conventional data collection systems provide dated information

→ What is the benefit of complementing conventional with innovative data sources?

ADB's Data for Development Technical Assistance

Technical Manual on Disaggregation of Official Statistics and SDGs

Strategically-designed training workshops targeted to NSO staff

Online Course Modules on SAE and Big Data Analytics

Thank you very much!

email:
amartinezjr@adb.org