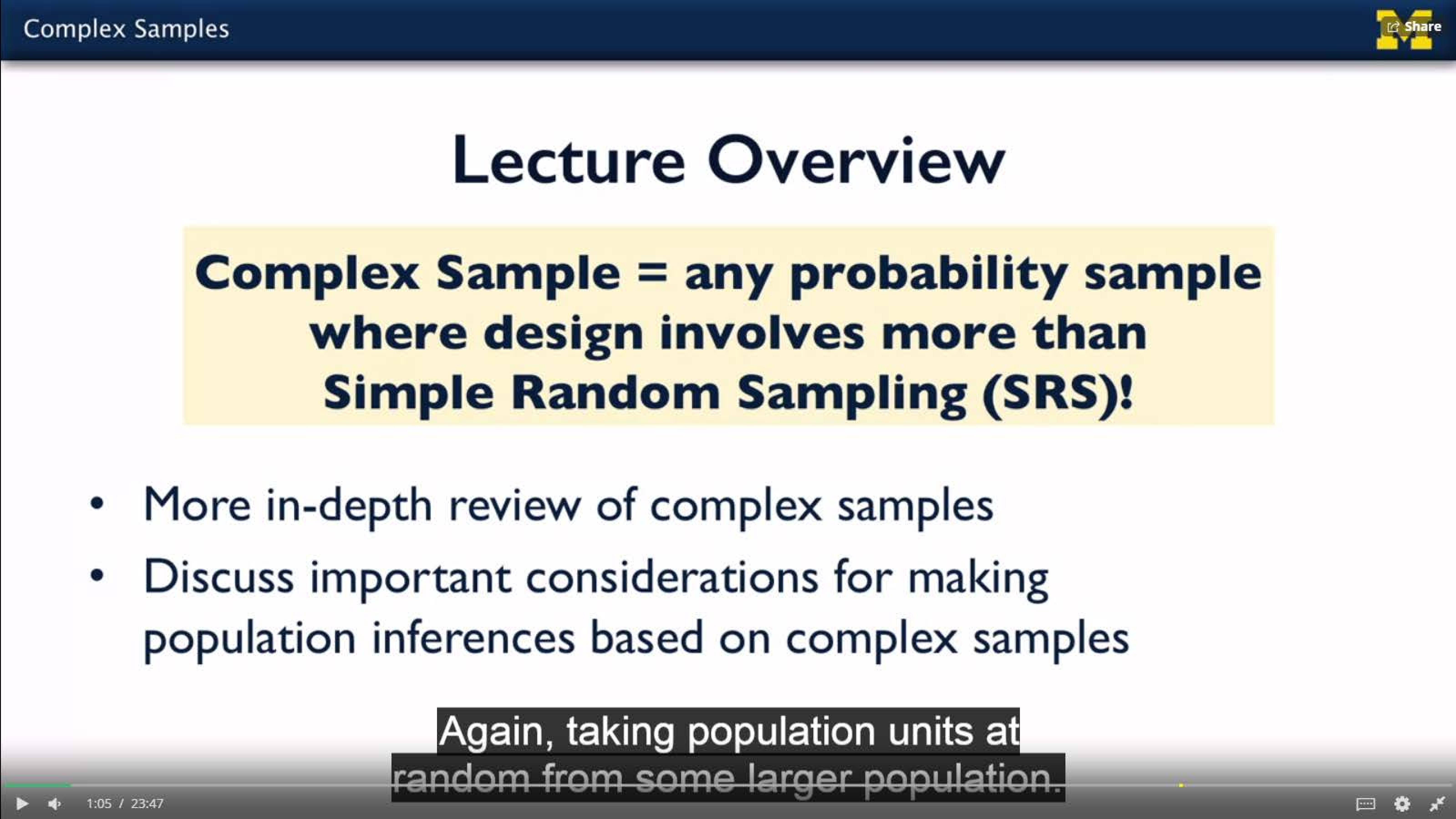
Complex Samples

Brady T. West

Research Associate Professor, Survey Research Center,
Institute for Social Research



Complex Samples

Lecture Overview

Complex Sample = any probability sample where design involves more than Simple Random Sampling (SRS)!

- More in-depth review of complex samples
- Discuss important considerations for making population inferences based on complex samples

Again, taking population units at random from some larger population.

Share

Features of Complex Samples: Stratification

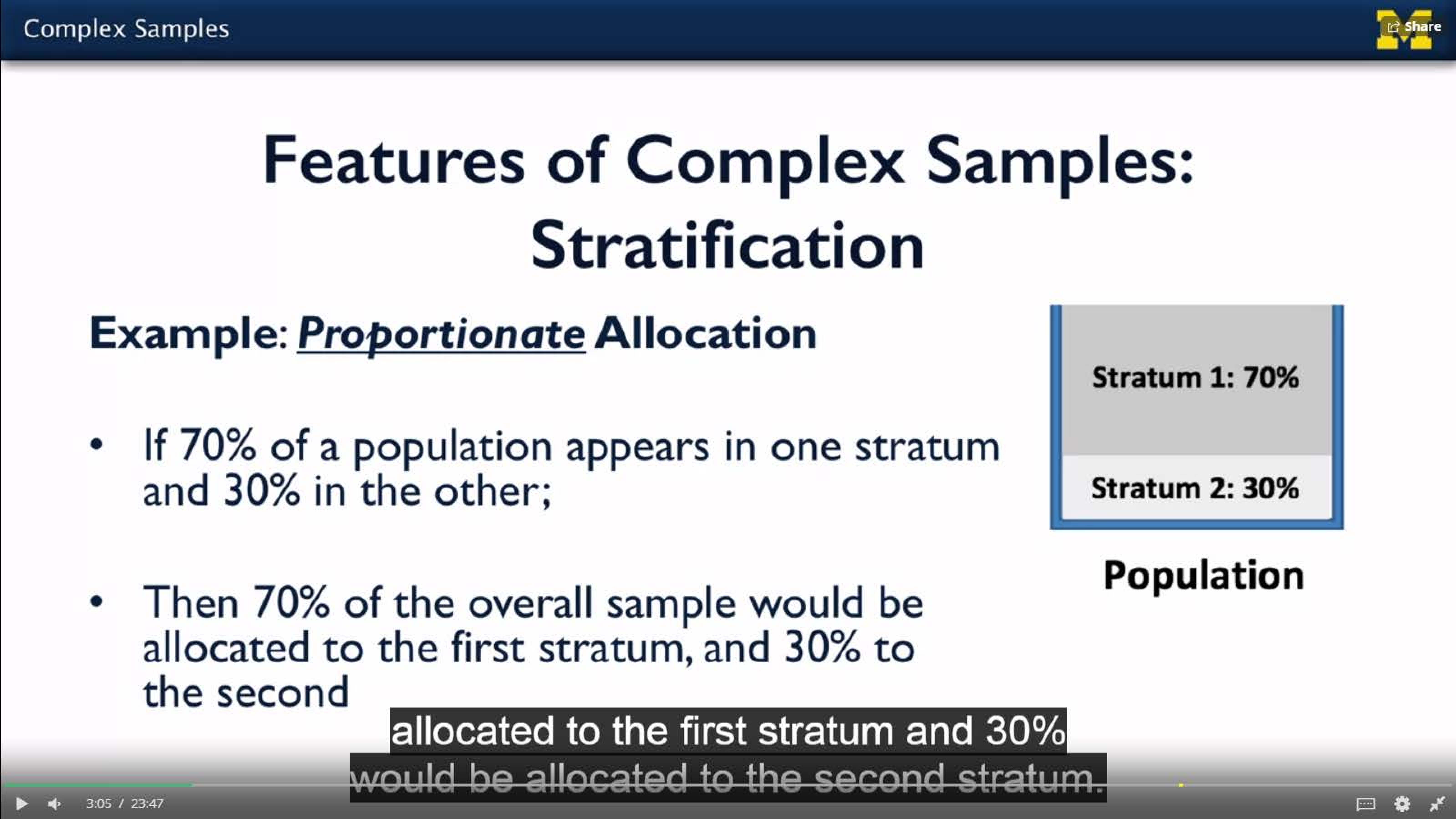
- **Stratification**: Allocation of overall sample to different "strata", or mutually exclusive divisions of the population (e.g., regions of the United States)

- Several different allocation schemes are possible; Aim → minimize sampling variance for particular variables given fixed costs

sampling variance for
particular variables given fixed costs.

"US 4 regions" CC-SA 3.0

2:15 / 23:47



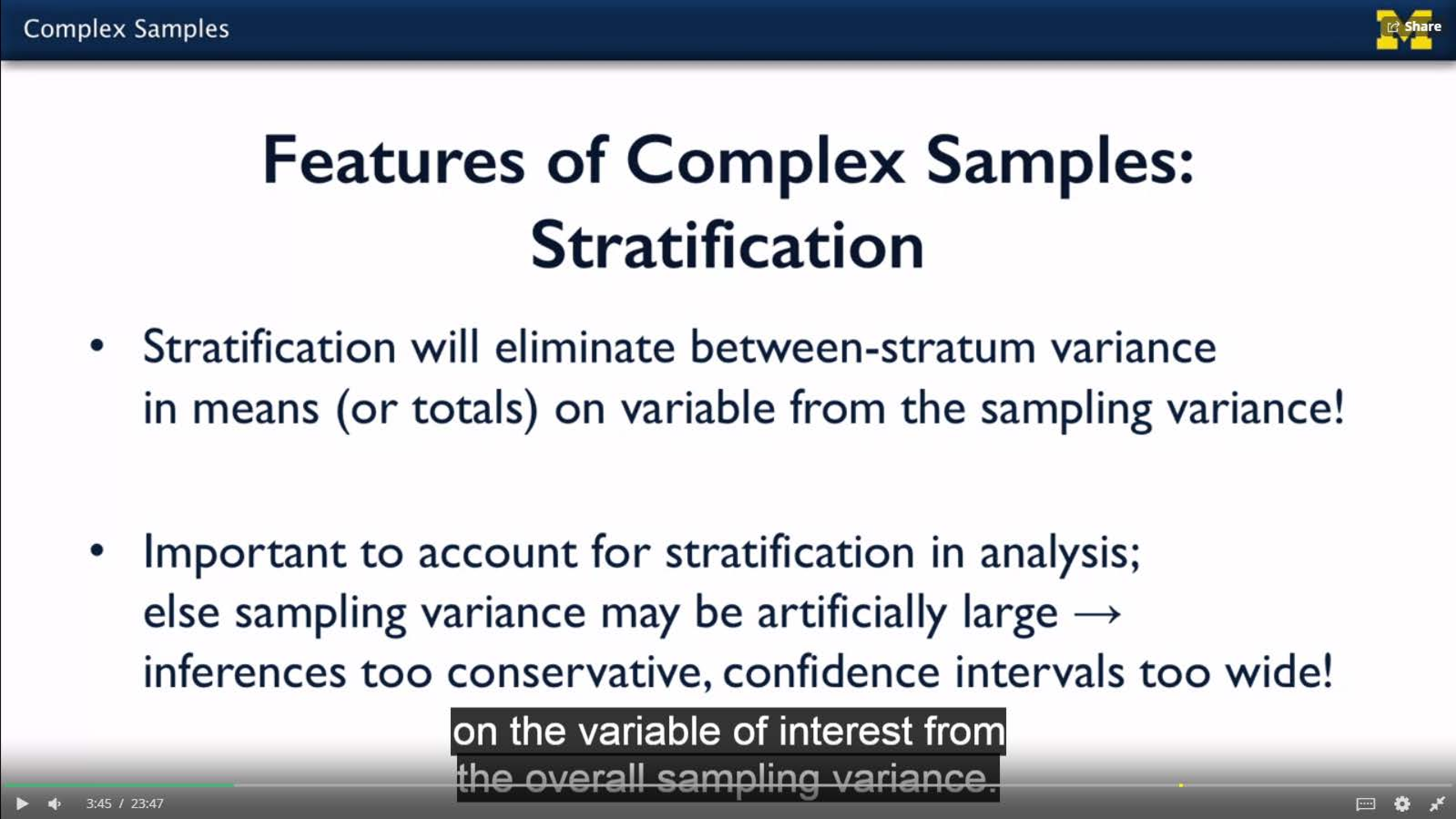
Complex Samples

Features of Complex Samples: Stratification

Example: ***Proportionate*** **Allocation**

- If 70% of a population appears in one stratum and 30% in the other;
- Then 70% of the overall sample would be allocated to the first stratum, and 30% to the second

allocated to the first stratum and 30%
would be allocated to the second stratum.



Features of Complex Samples: Stratification

- Stratification will eliminate between-stratum variance in means (or totals) on variable from the sampling variance!

- Important to account for stratification in analysis; else sampling variance may be artificially large → inferences too conservative, confidence intervals too wide!

on the variable of interest from
the overall sampling variance.

Features of Complex Samples: Clustering

- **Clustering**: Random sampling of larger clusters of population elements, possibly across multiple stages (e.g., counties, then segments, then households)

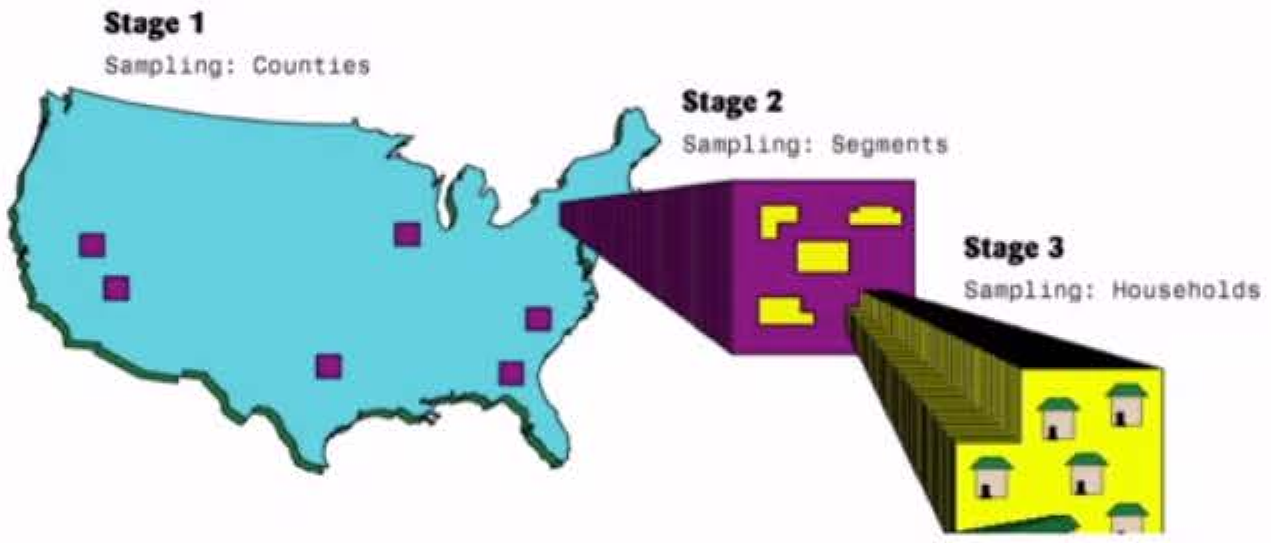
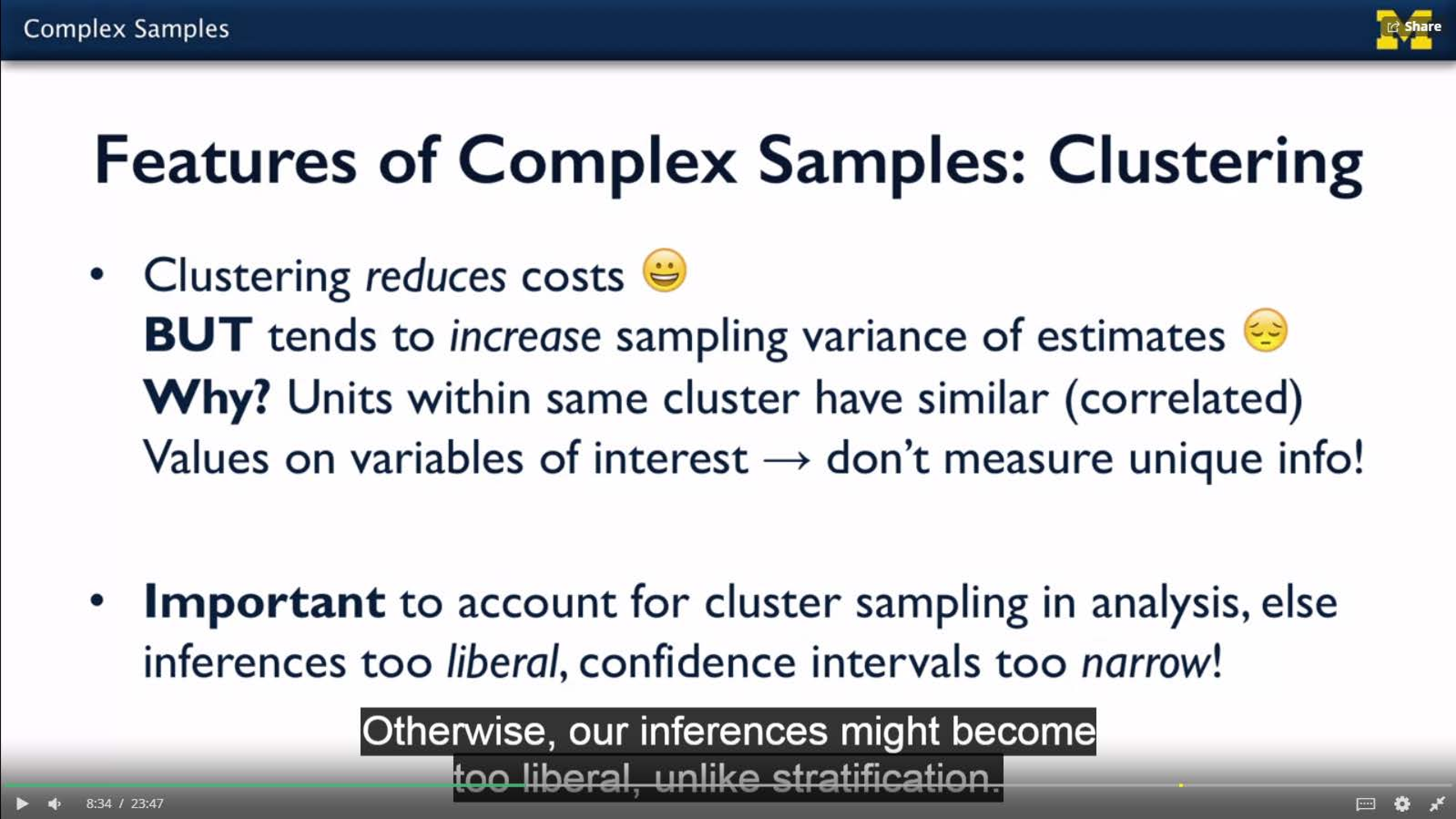


Image Credit: L. Mahadjer, Westat

- Reduces cost of data collection: expensive $$$ to visit *n* randomly sampled units from large and widespread population

but minimizing the cost
of data collection.

7:14 / 23:47



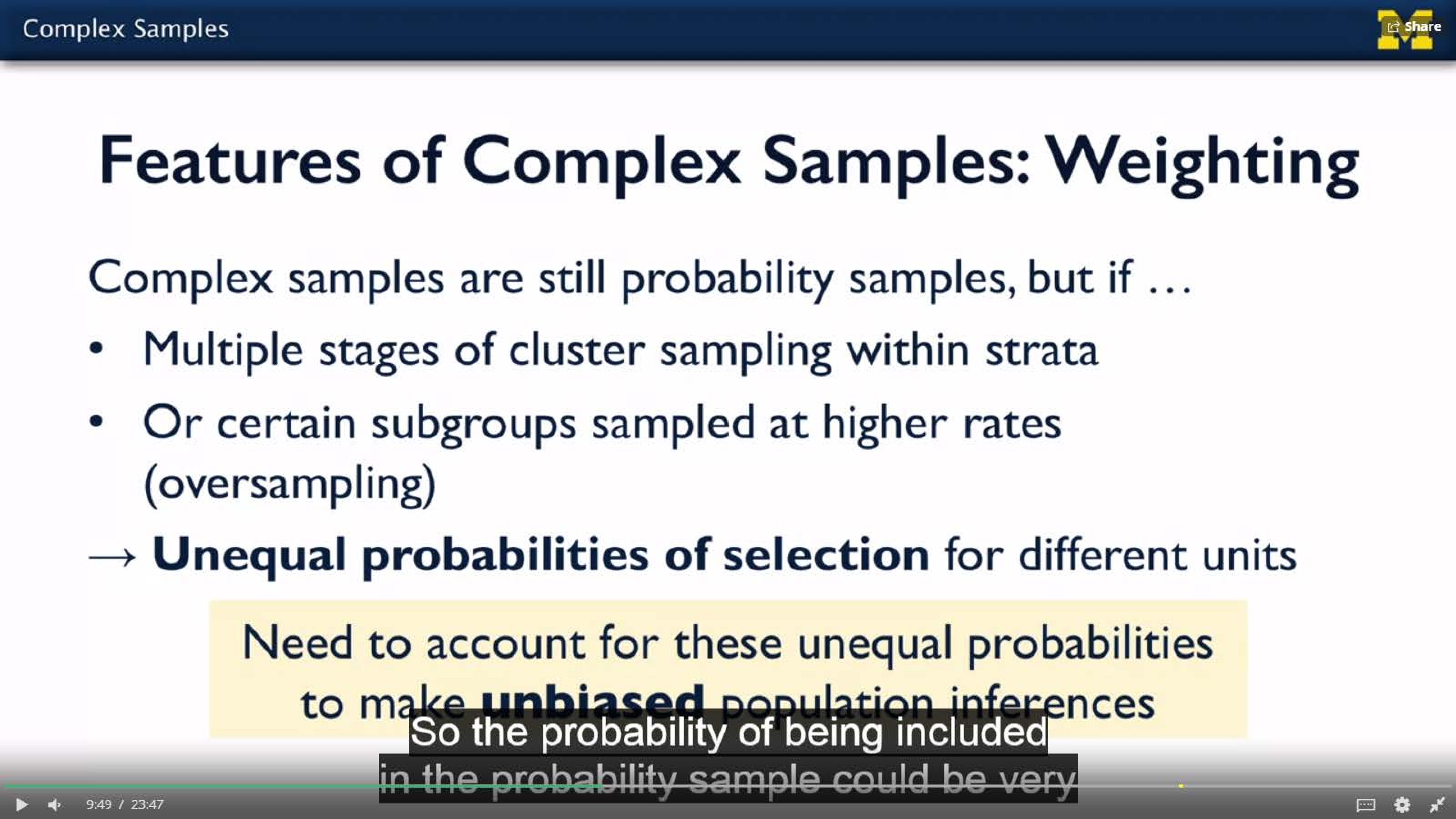
Complex Samples

Features of Complex Samples: Clustering

- Clustering *reduces* costs 😀
  BUT tends to *increase* sampling variance of estimates 😔
  Why? Units within same cluster have similar (correlated) Values on variables of interest → don't measure unique info!

- **Important** to account for cluster sampling in analysis, else inferences too *liberal*, confidence intervals too *narrow*!

Otherwise, our inferences might become
too liberal, unlike stratification.



Features of Complex Samples: Weighting

Complex samples are still probability samples, but if …

- Multiple stages of cluster sampling within strata
- Or certain subgroups sampled at higher rates (oversampling)

→ **Unequal probabilities of selection** for different units

Need to account for these unequal probabilities to make **unbiased** population inferences

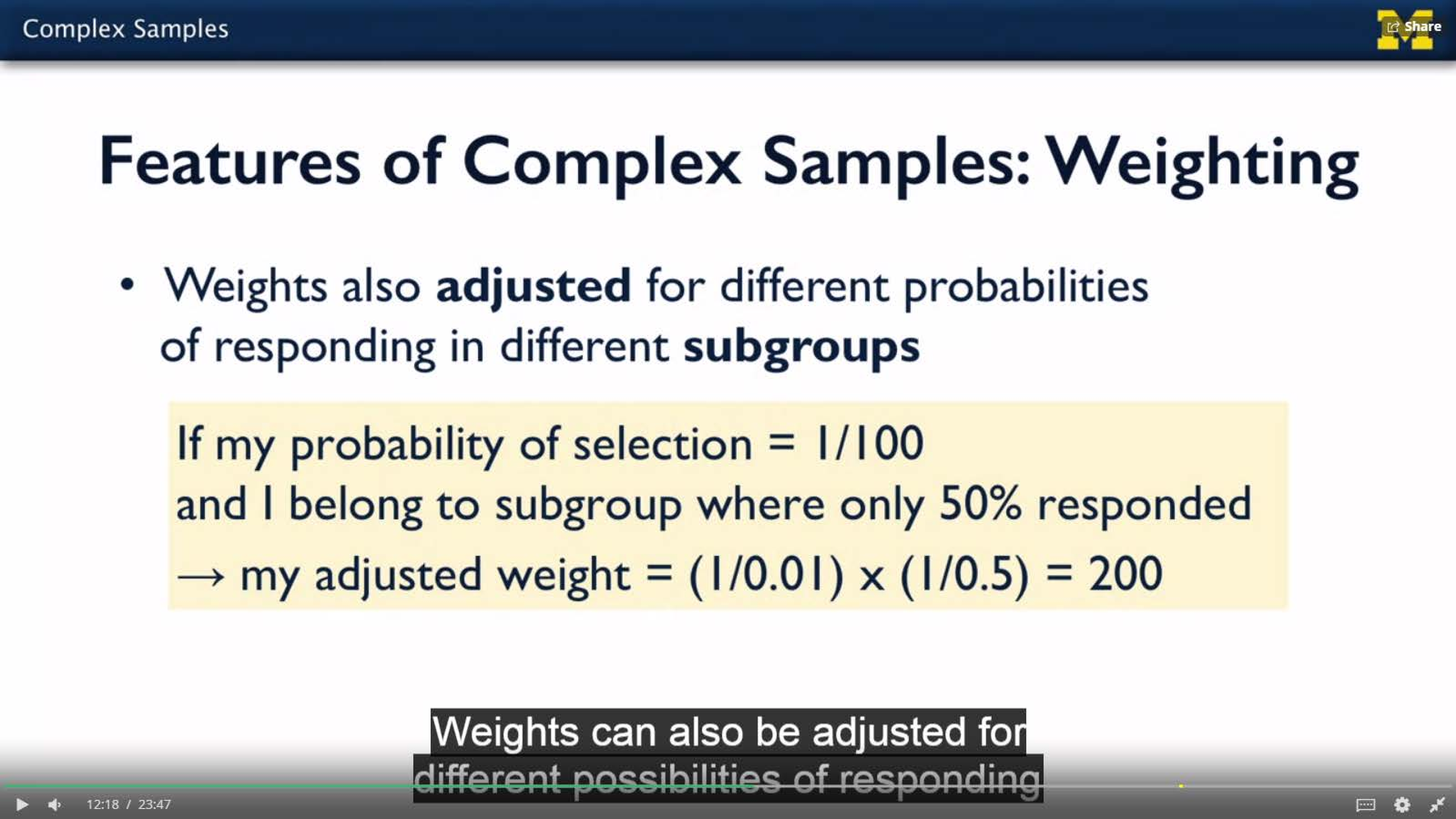
So the probability of being included
in the probability sample could be very

Features of Complex Samples: Weighting

- **How**? Use of **weights** in analysis …
  (partly) defined by **inverse of probability of selection**

If my probability is 1/100 → my weight is 100,
I represent ***myself*** and **99 others** in the population!

Partly, weights and complex samples
are defined by the inverse of a given



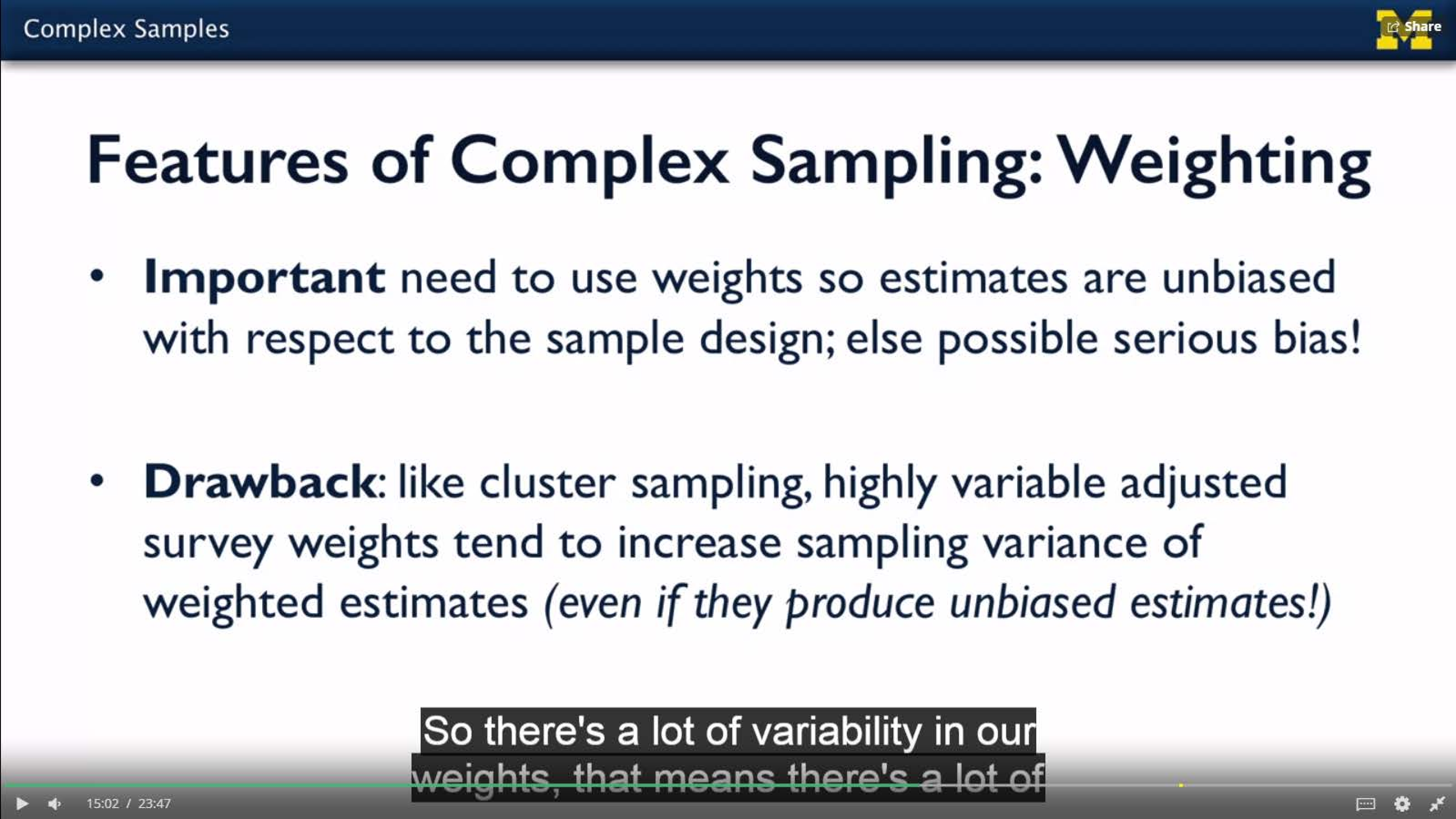
Complex Samples

Features of Complex Samples: Weighting

- Weights also **adjusted** for different probabilities of responding in different **subgroups**

If my probability of selection = 1/100
and I belong to subgroup where only 50% responded
→ my adjusted weight = (1/0.01) x (1/0.5) = 200

Weights can also be adjusted for
different possibilities of responding

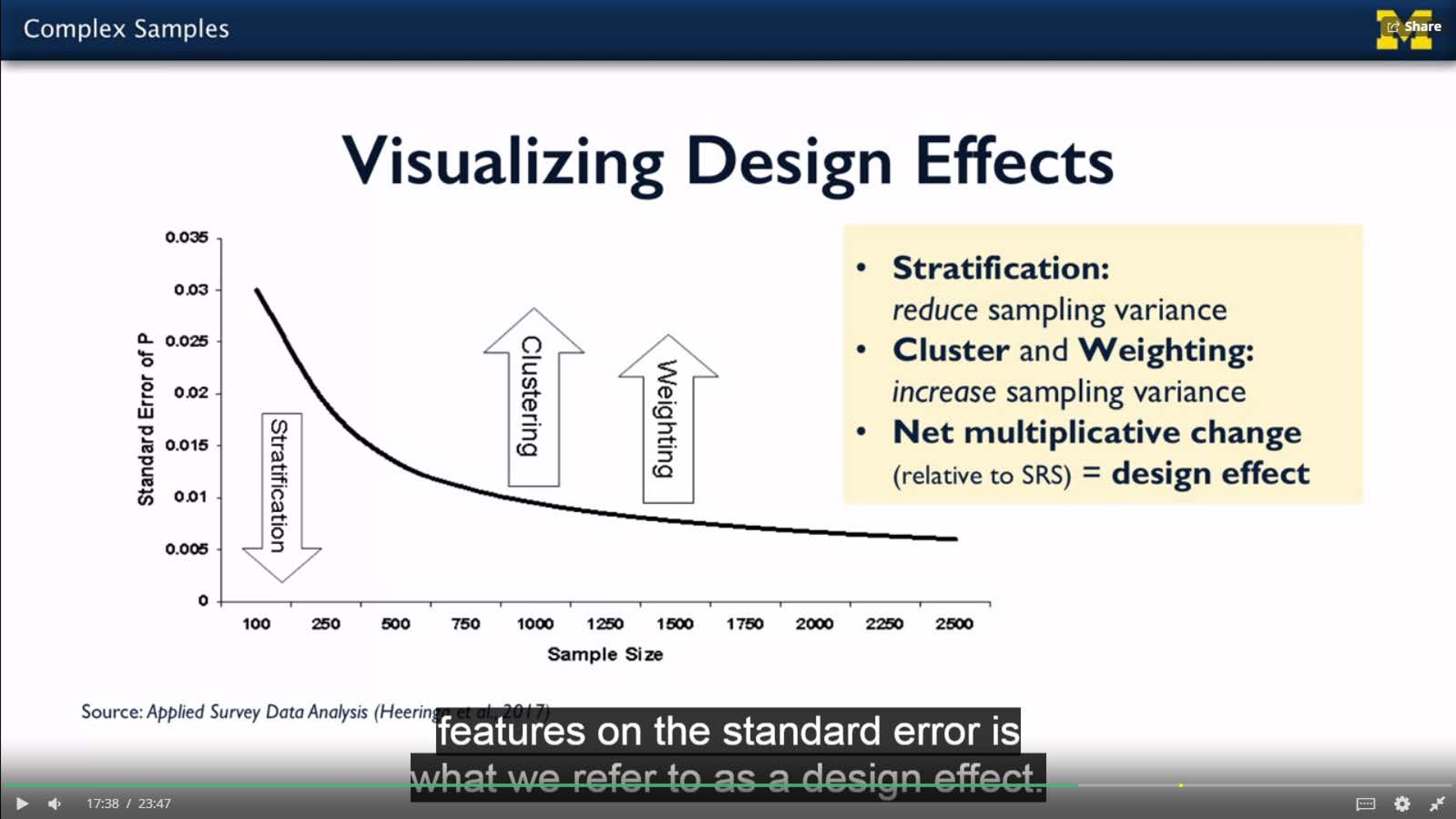


Complex Samples

Features of Complex Sampling: Weighting

- **Important** need to use weights so estimates are unbiased with respect to the sample design; else possible serious bias!
- **Drawback**: like cluster sampling, highly variable adjusted survey weights tend to increase sampling variance of weighted estimates *(even if they produce unbiased estimates!)*

So there's a lot of variability in our weights, that means there's a lot of

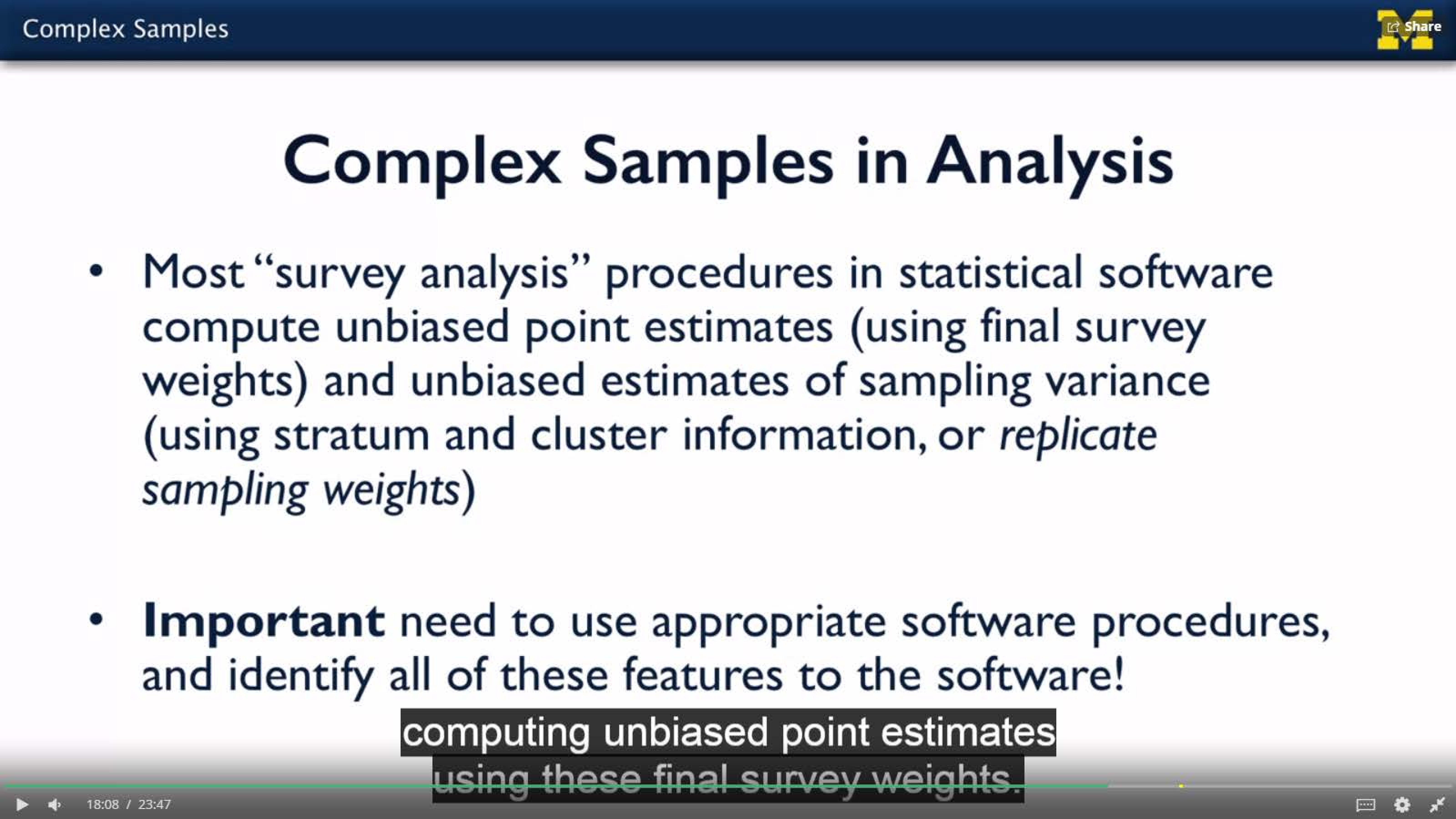


Visualizing Design Effects

- **Stratification:** *reduce* sampling variance
- **Cluster** and **Weighting:** *increase* sampling variance
- **Net multiplicative change** (relative to SRS) = **design effect**

Source: *Applied Survey Data Analysis* (Heeringa et al., 2017)

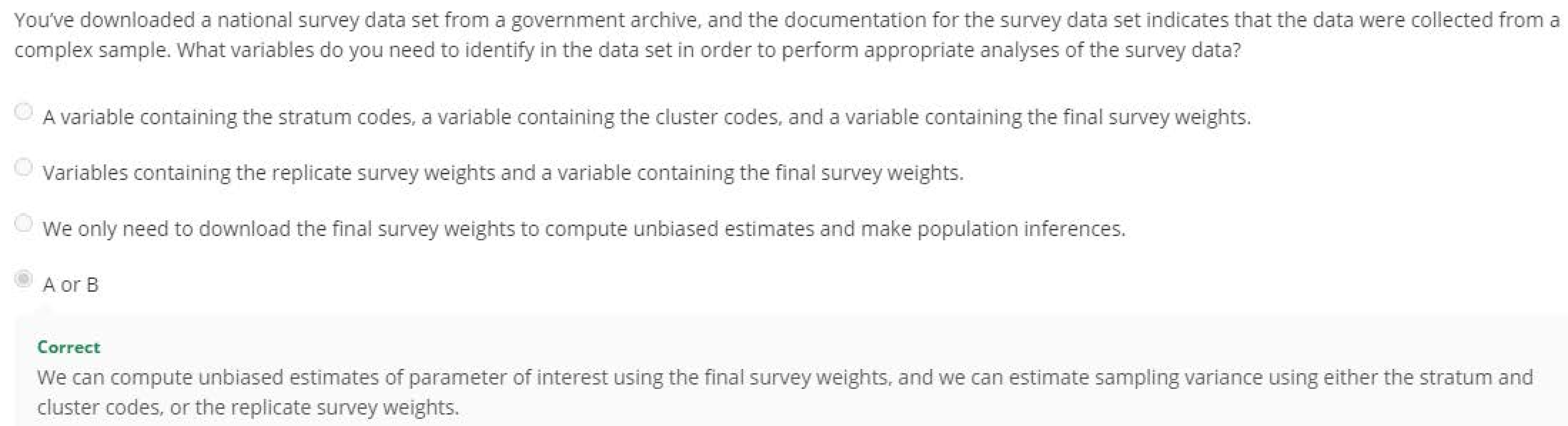
features on the standard error is
what we refer to as a design effect.



Complex Samples in Analysis

- Most "survey analysis" procedures in statistical software compute unbiased point estimates (using final survey weights) and unbiased estimates of sampling variance (using stratum and cluster information, or *replicate sampling weights*)
- **Important** need to use appropriate software procedures, and identify all of these features to the software!

computing unbiased point estimates
using these final survey weights.



You've downloaded a national survey data set from a government archive, and the documentation for the survey data set indicates that the data were collected from a complex sample. What variables do you need to identify in the data set in order to perform appropriate analyses of the survey data?

○ A variable containing the stratum codes, a variable containing the cluster codes, and a variable containing the final survey weights.

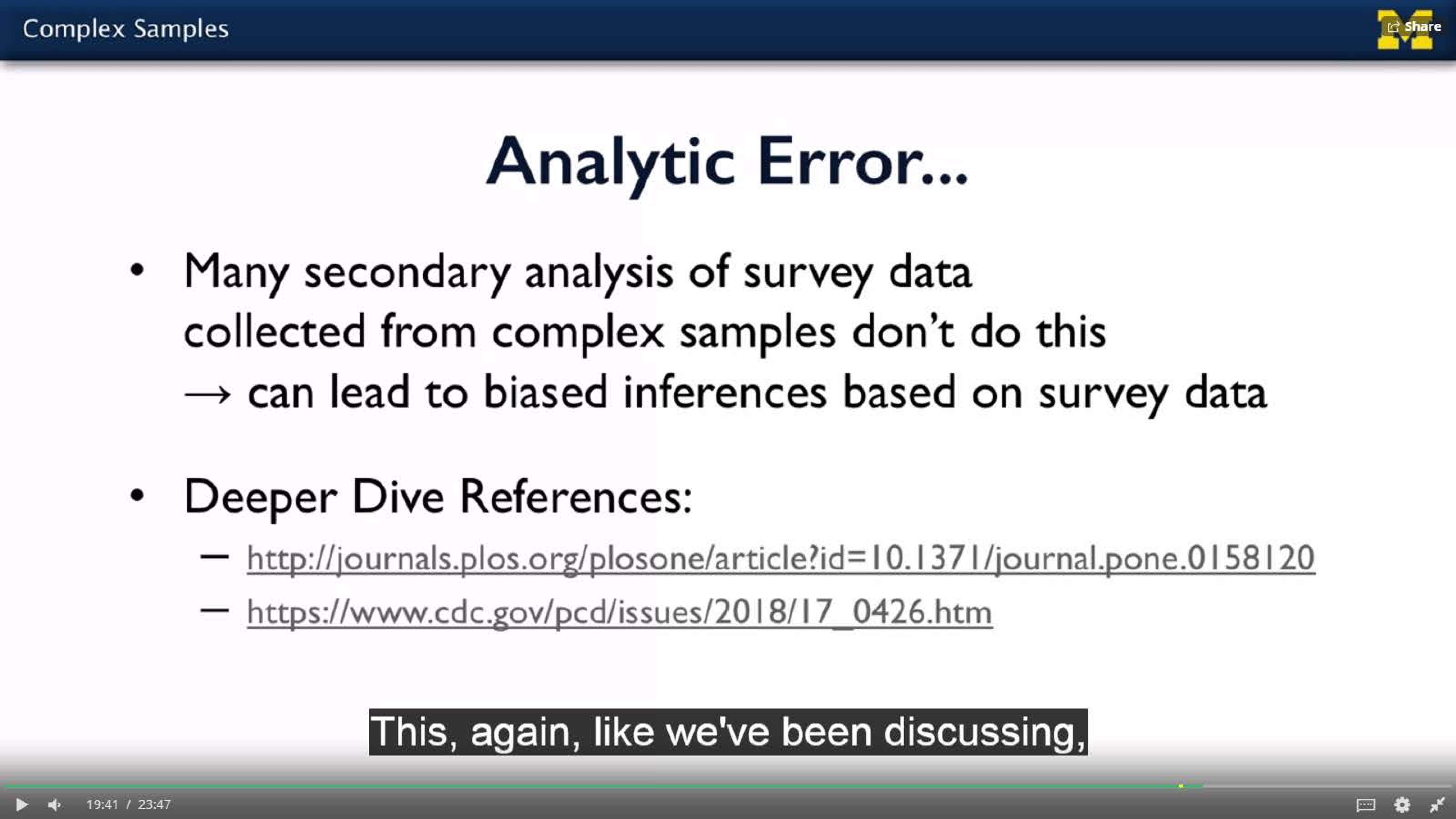
○ Variables containing the replicate survey weights and a variable containing the final survey weights.

○ We only need to download the final survey weights to compute unbiased estimates and make population inferences.

● A or B

Correct

We can compute unbiased estimates of parameter of interest using the final survey weights, and we can estimate sampling variance using either the stratum and cluster codes, or the replicate survey weights.

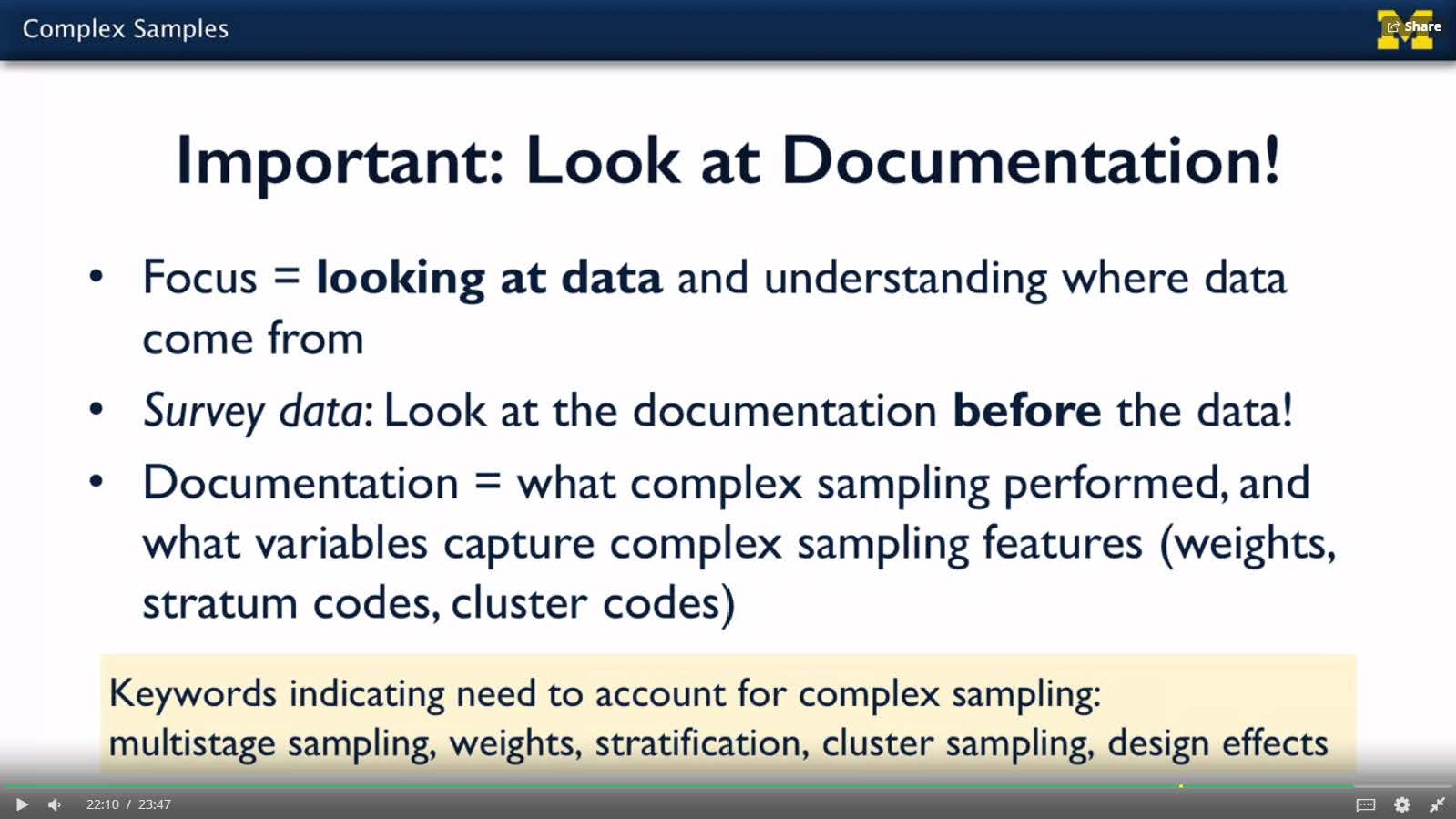


Analytic Error...

- Many secondary analysis of survey data collected from complex samples don’t do this → can lead to biased inferences based on survey data

- Deeper Dive References:
  - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0158120
  - https://www.cdc.gov/pcd/issues/2018/17_0426.htm

This, again, like we've been discussing,

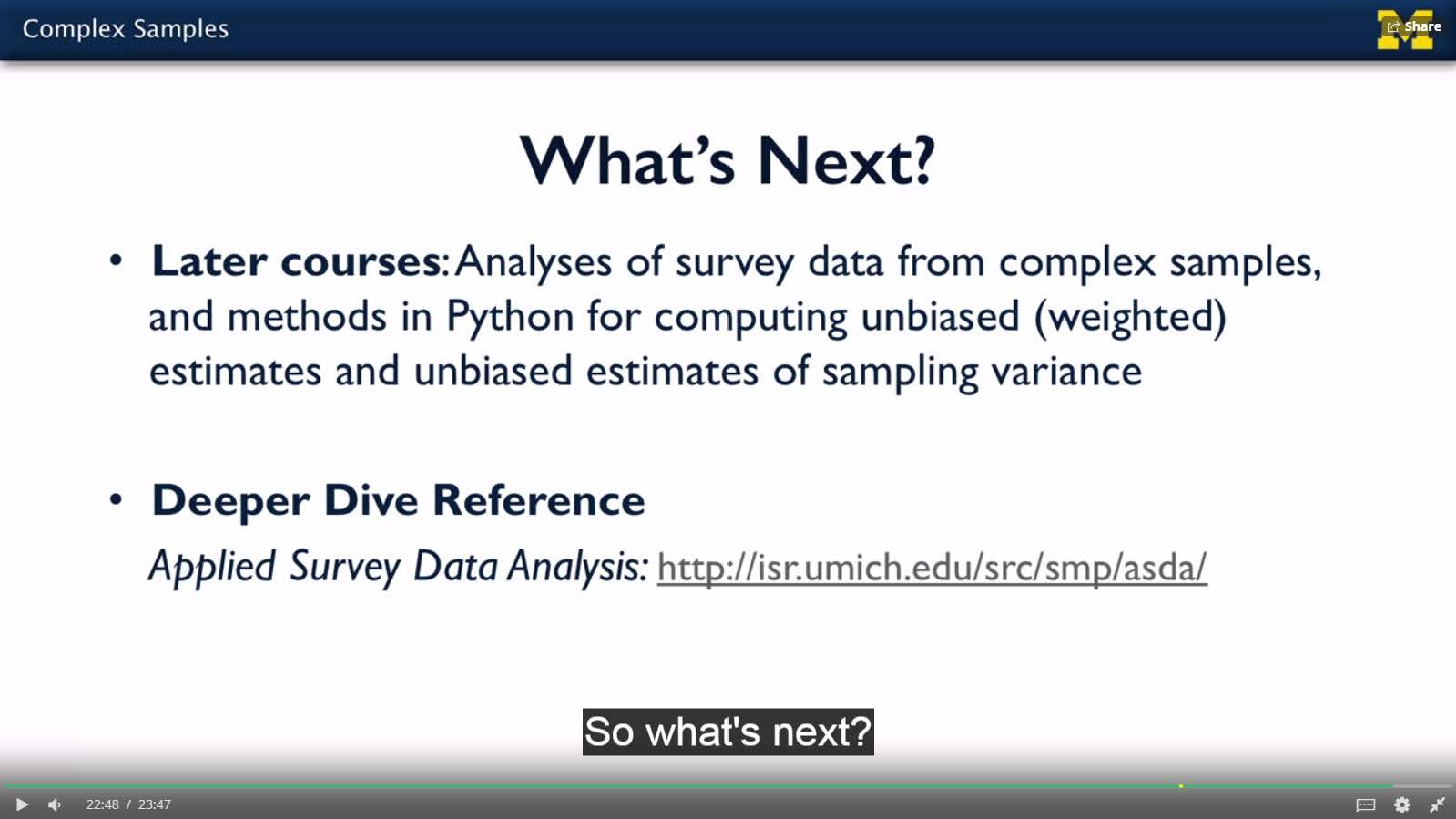


Complex Samples

Important: Look at Documentation!

- Focus = **looking at data** and understanding where data come from
- *Survey data*: Look at the documentation **before** the data!
- Documentation = what complex sampling performed, and what variables capture complex sampling features (weights, stratum codes, cluster codes)

Keywords indicating need to account for complex sampling:
multistage sampling, weights, stratification, cluster sampling, design effects



What's Next?

- **Later courses:** Analyses of survey data from complex samples, and methods in Python for computing unbiased (weighted) estimates and unbiased estimates of sampling variance

- **Deeper Dive Reference**

  Applied Survey Data Analysis: http://isr.umich.edu/src/smp/asda/

So what's next?