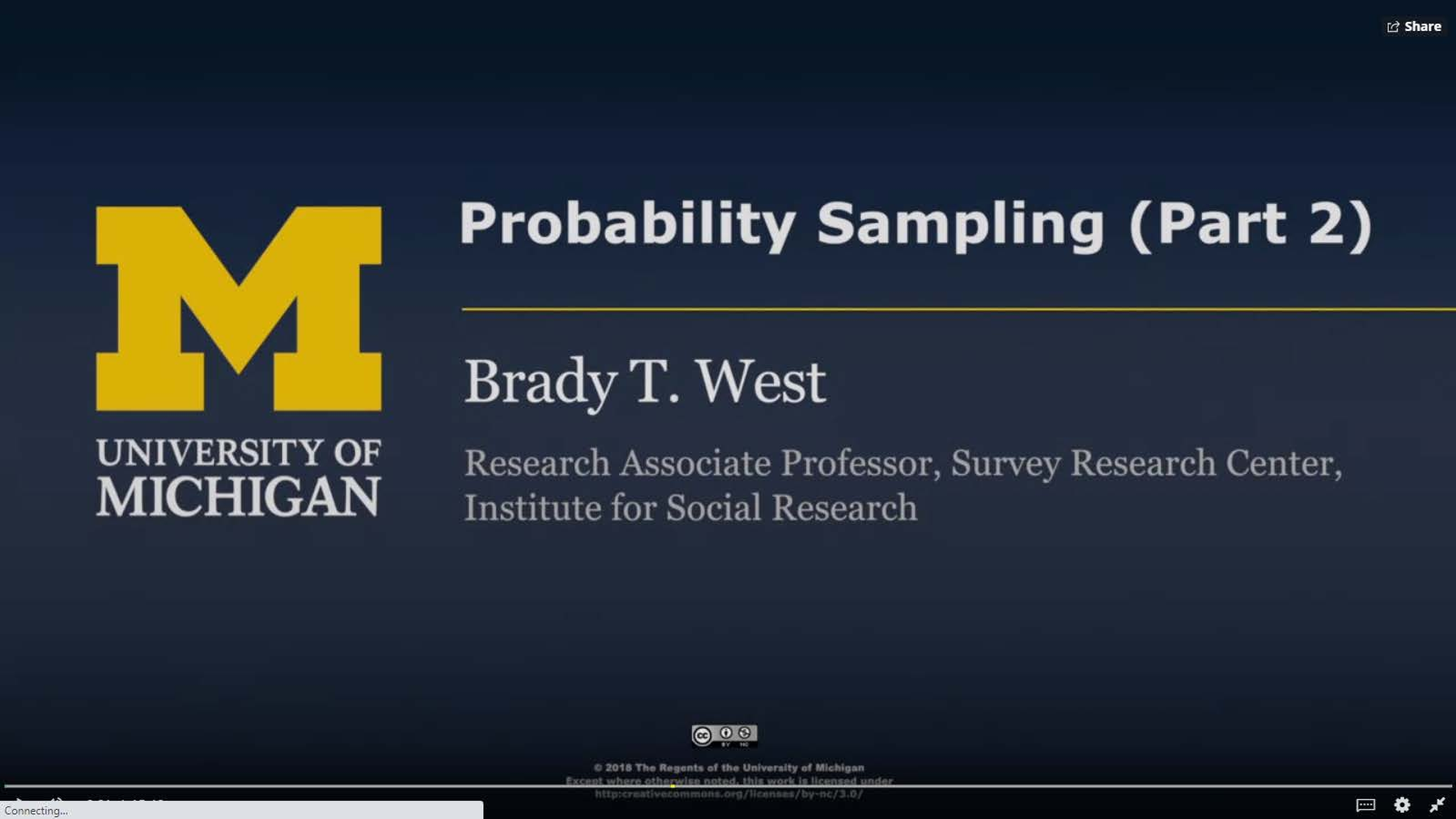
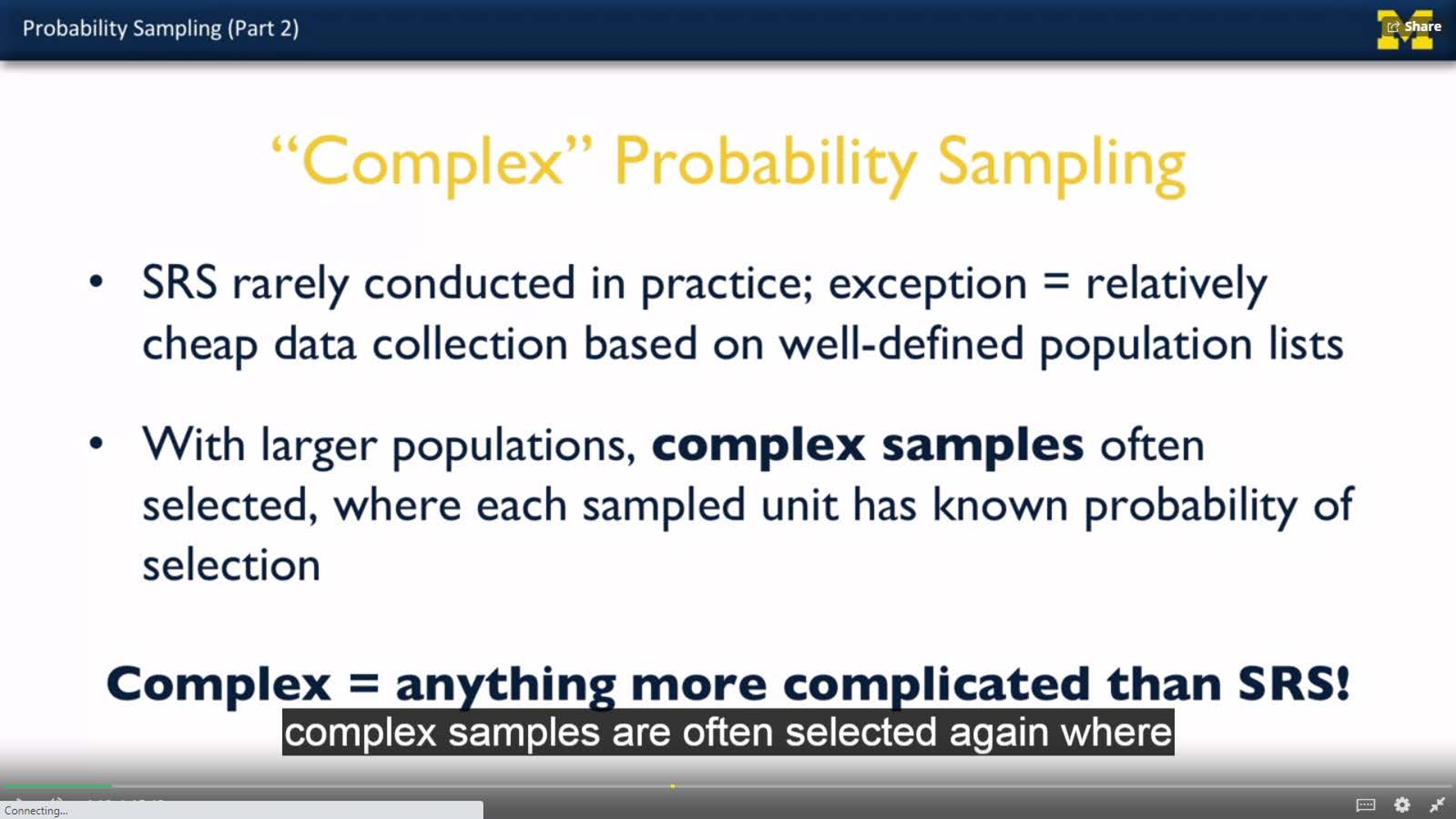


Probability Sampling (Part 2)

Brady T. West

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

© 2018 The Regents of the University of Michigan
Except where otherwise noted, this work is licensed under
http:creativecommons.org/licenses/by-nc/3.0/

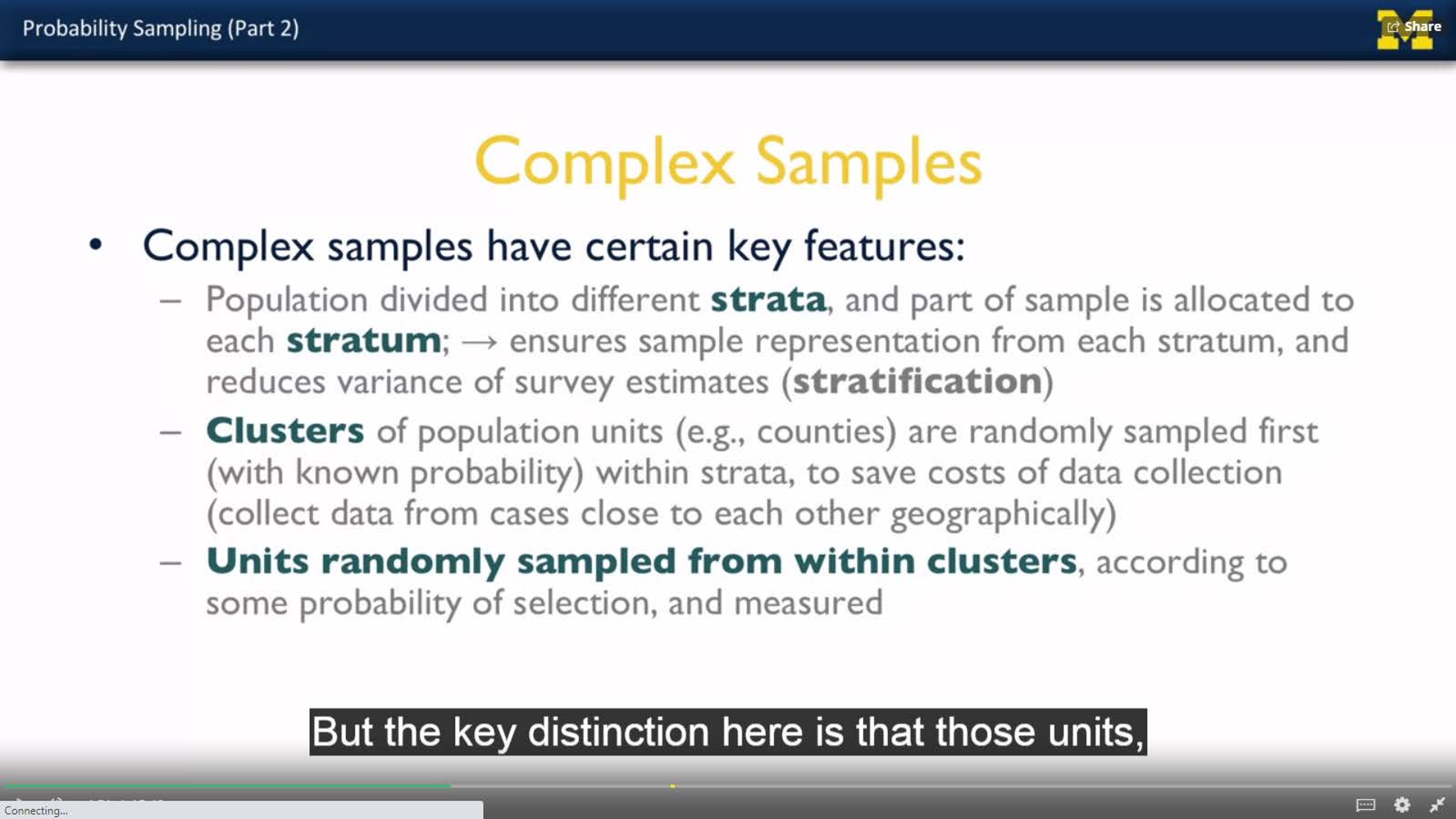


"Complex" Probability Sampling

- SRS rarely conducted in practice; exception = relatively cheap data collection based on well-defined population lists
- With larger populations, **complex samples** often selected, where each sampled unit has known probability of selection

Complex = anything more complicated than SRS!

complex samples are often selected again where

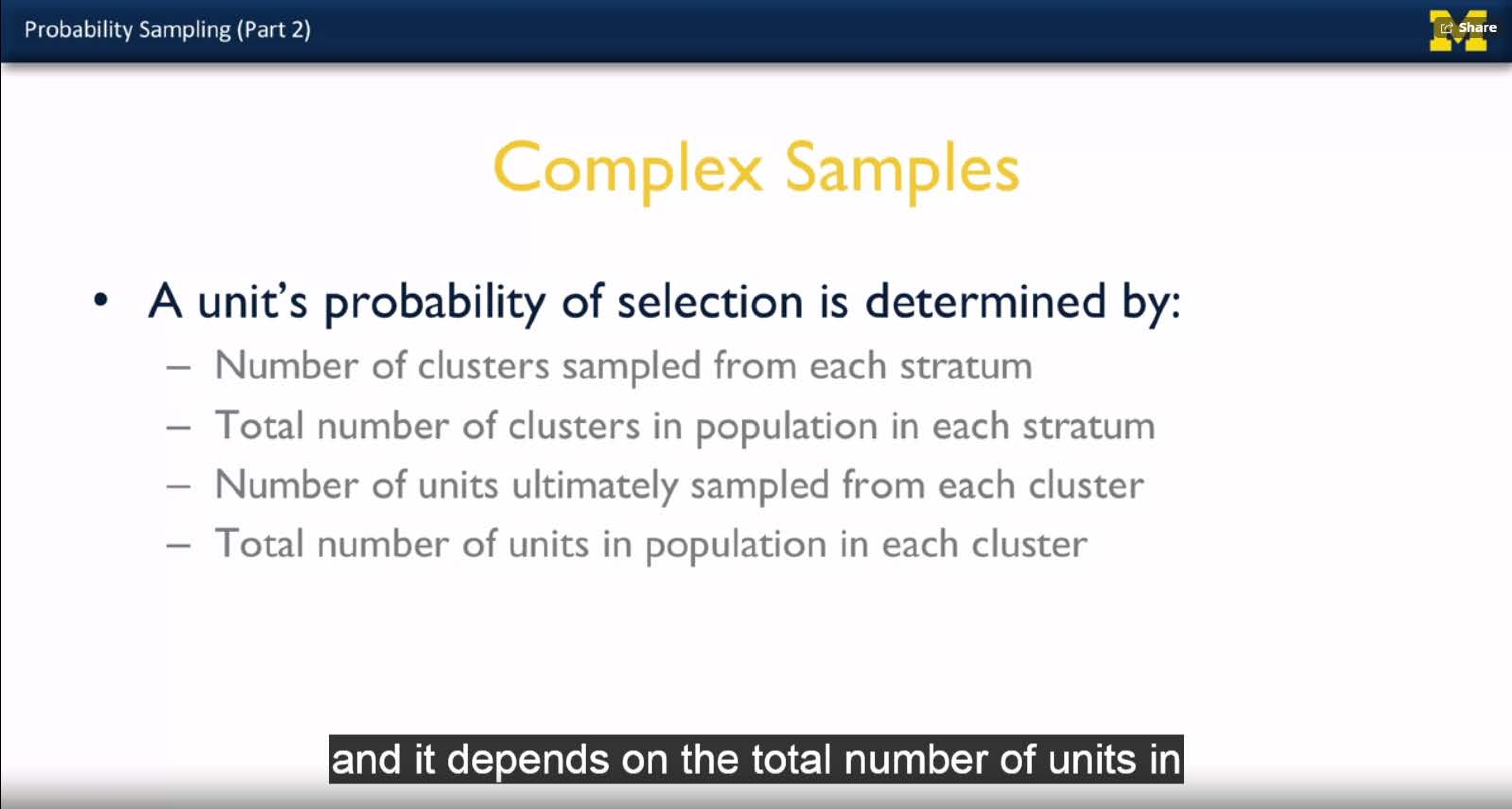


Probability Sampling (Part 2)

Complex Samples

- Complex samples have certain key features:
  - Population divided into different **strata**, and part of sample is allocated to each **stratum**; → ensures sample representation from each stratum, and reduces variance of survey estimates (**stratification**)
  - **Clusters** of population units (e.g., counties) are randomly sampled first (with known probability) within strata, to save costs of data collection (collect data from cases close to each other geographically)
  - **Units randomly sampled from within clusters**, according to some probability of selection, and measured

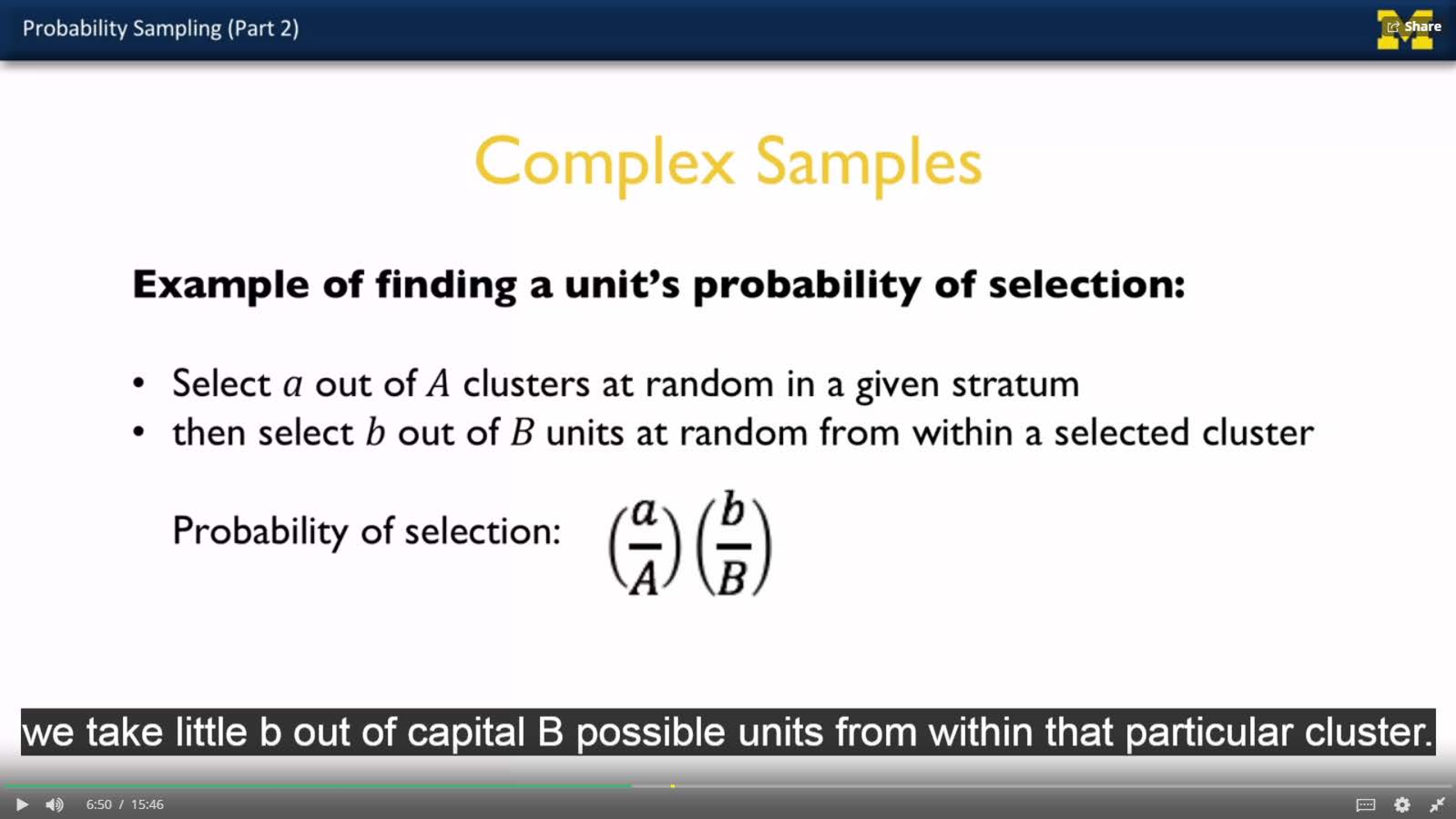
But the key distinction here is that those units,



Complex Samples

- A unit's probability of selection is determined by:
  - Number of clusters sampled from each stratum
  - Total number of clusters in population in each stratum
  - Number of units ultimately sampled from each cluster
  - Total number of units in population in each cluster

and it depends on the total number of units in



Probability Sampling (Part 2)

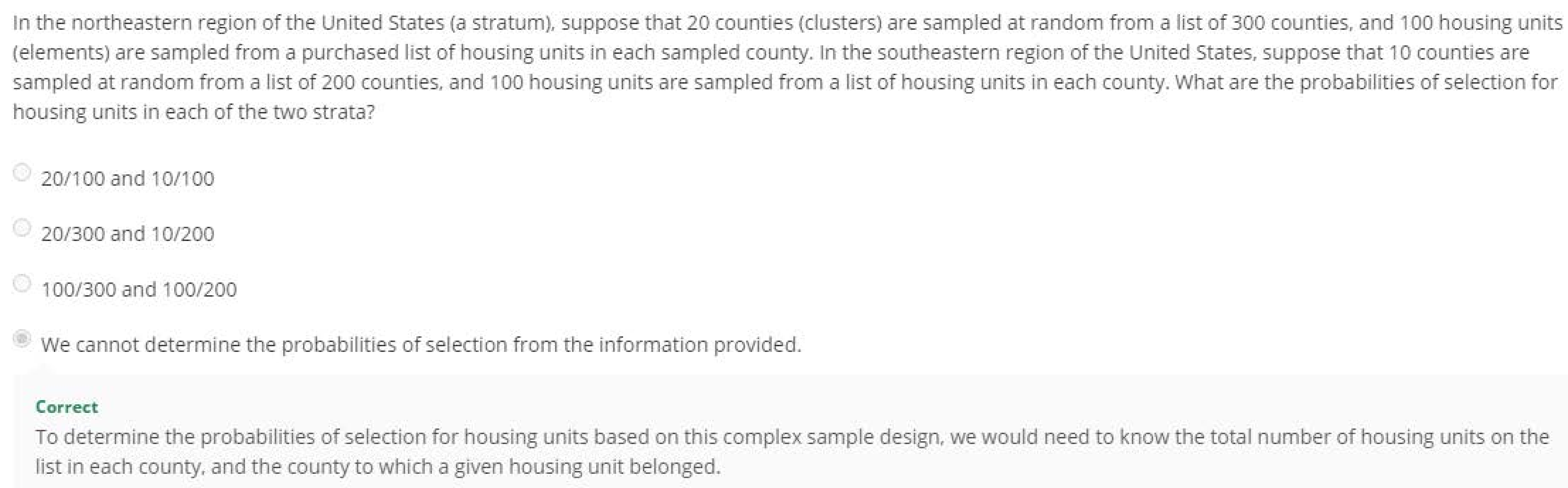
Complex Samples

Example of finding a unit's probability of selection:

- Select a out of A clusters at random in a given stratum
- then select b out of B units at random from within a selected cluster

Probability of selection: $\left(\frac{a}{A}\right)\left(\frac{b}{B}\right)$

we take little b out of capital B possible units from within that particular cluster.



In the northeastern region of the United States (a stratum), suppose that 20 counties (clusters) are sampled at random from a list of 300 counties, and 100 housing units (elements) are sampled from a purchased list of housing units in each sampled county. In the southeastern region of the United States, suppose that 10 counties are sampled at random from a list of 200 counties, and 100 housing units are sampled from a list of housing units in each county. What are the probabilities of selection for housing units in each of the two strata?

○ 20/100 and 10/100

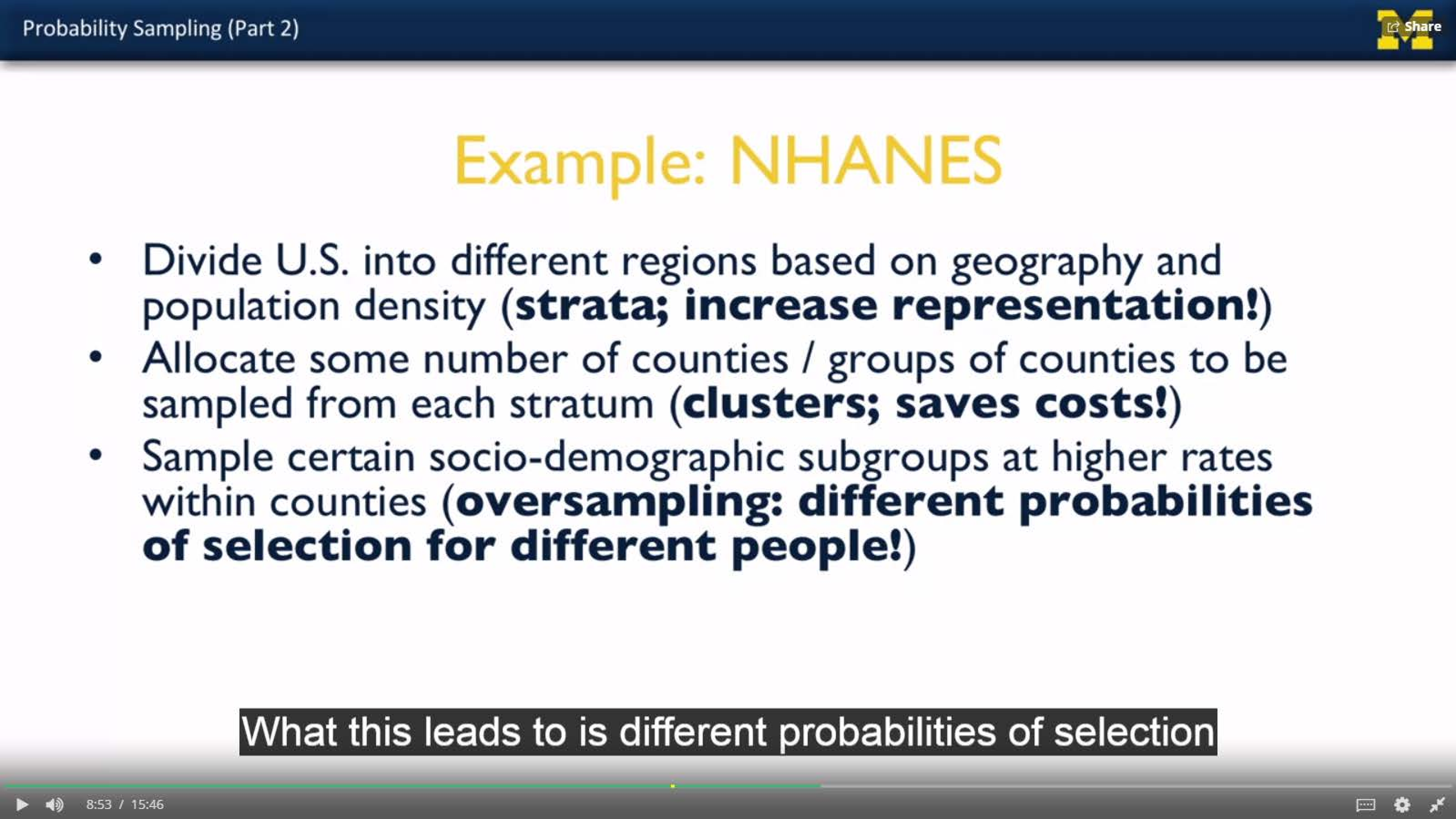
○ 20/300 and 10/200

○ 100/300 and 100/200

● We cannot determine the probabilities of selection from the information provided.

Correct

To determine the probabilities of selection for housing units based on this complex sample design, we would need to know the total number of housing units on the list in each county, and the county to which a given housing unit belonged.

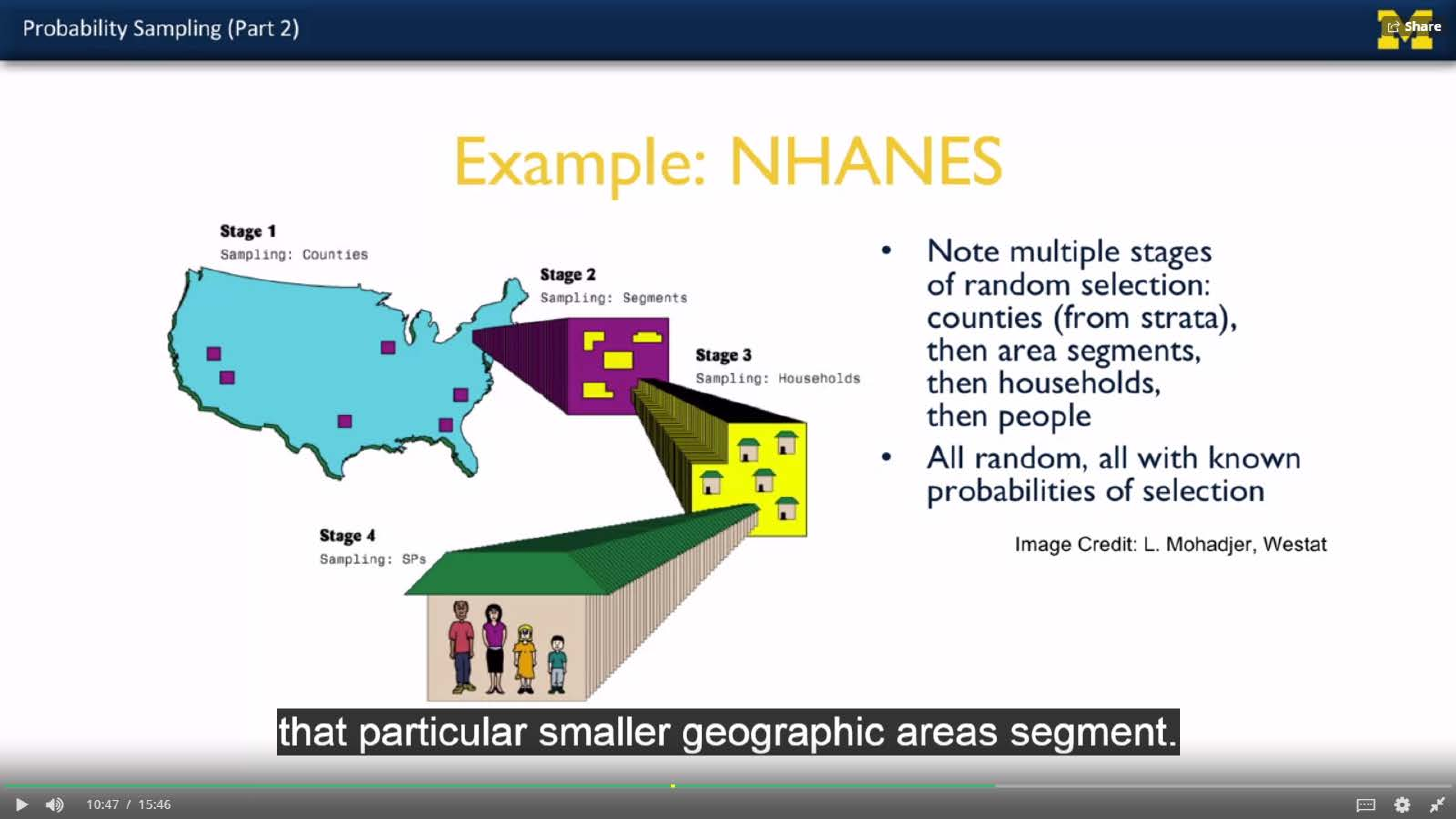


Probability Sampling (Part 2)

Example: NHANES

- Divide U.S. into different regions based on geography and population density (**strata; increase representation!**)
- Allocate some number of counties / groups of counties to be sampled from each stratum (**clusters; saves costs!**)
- Sample certain socio-demographic subgroups at higher rates within counties (**oversampling: different probabilities of selection for different people!**)

What this leads to is different probabilities of selection



Probability Sampling (Part 2)

Example: NHANES

- Note multiple stages of random selection: counties (from strata), then area segments, then households, then people
- All random, all with known probabilities of selection

Image Credit: L. Mohadjer, Westat

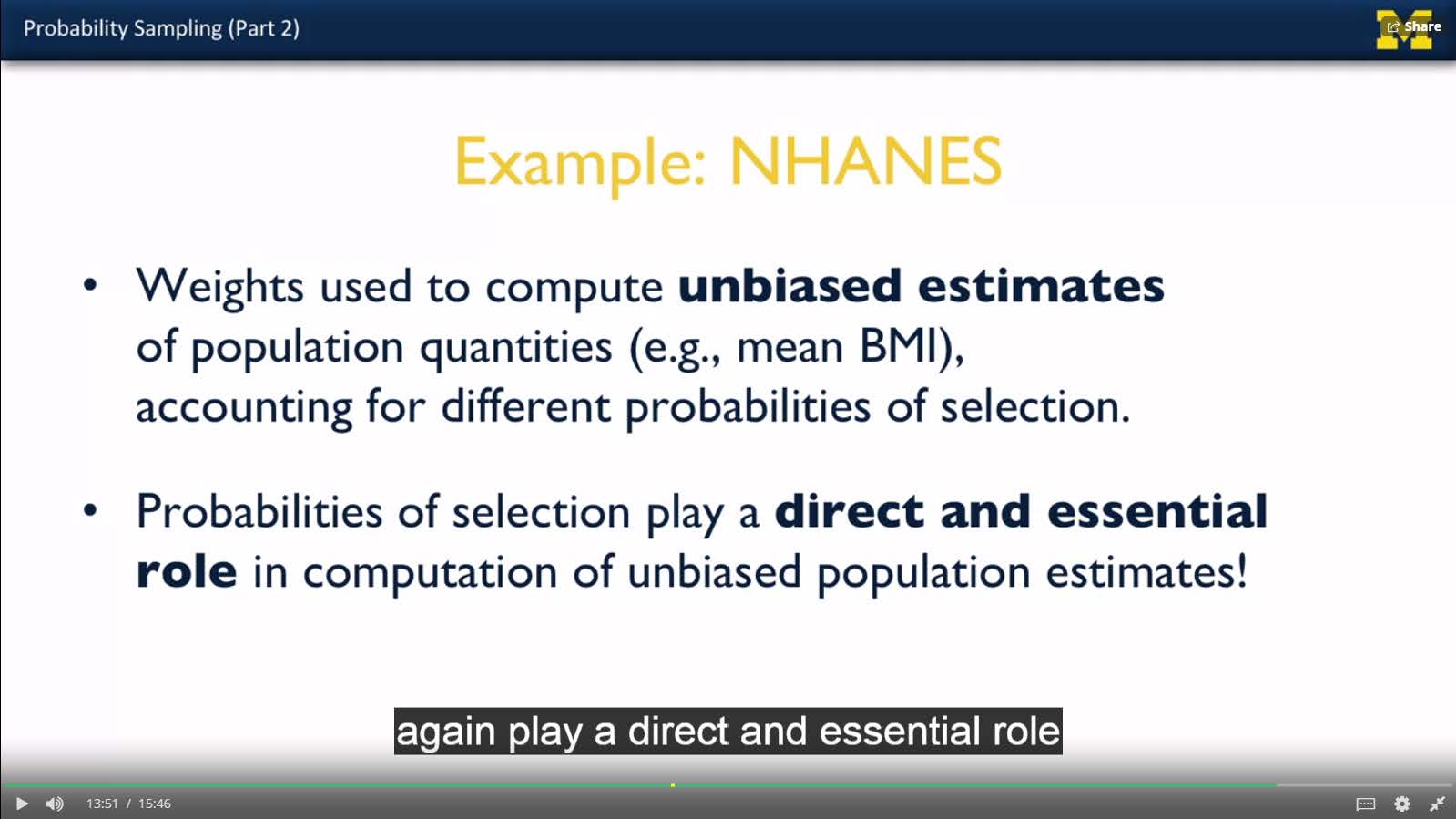
that particular smaller geographic areas segment.

Example: NHANES

- Drive a huge semi-trailer containing medical equipment and staff to each sampled county, and invite randomly selected people for an interview and a medical exam
- The inverse of a person's probability of selection is then their **sampling weight**
- If my probability is 1/100
  → my weight is 100
  I represent ***myself***
  and **99 others** in the population!

Image Credit: Steven Heeringa, Institute for Social Research, University of Michigan

If ultimately, my probability of selection was one divided by 100,

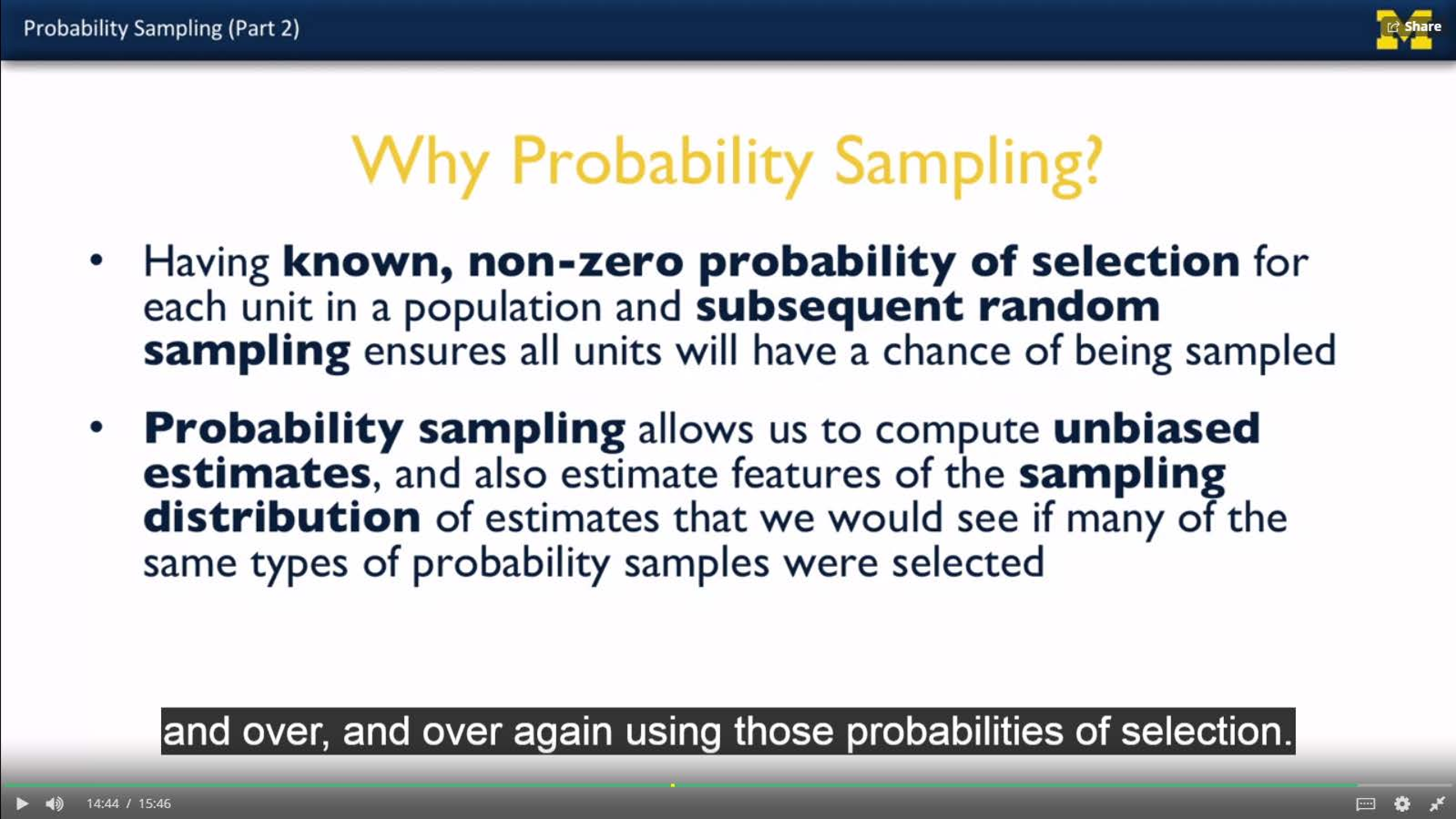


Probability Sampling (Part 2)

Example: NHANES

- Weights used to compute **unbiased estimates** of population quantities (e.g., mean BMI), accounting for different probabilities of selection.
- Probabilities of selection play a **direct and essential role** in computation of unbiased population estimates!

again play a direct and essential role

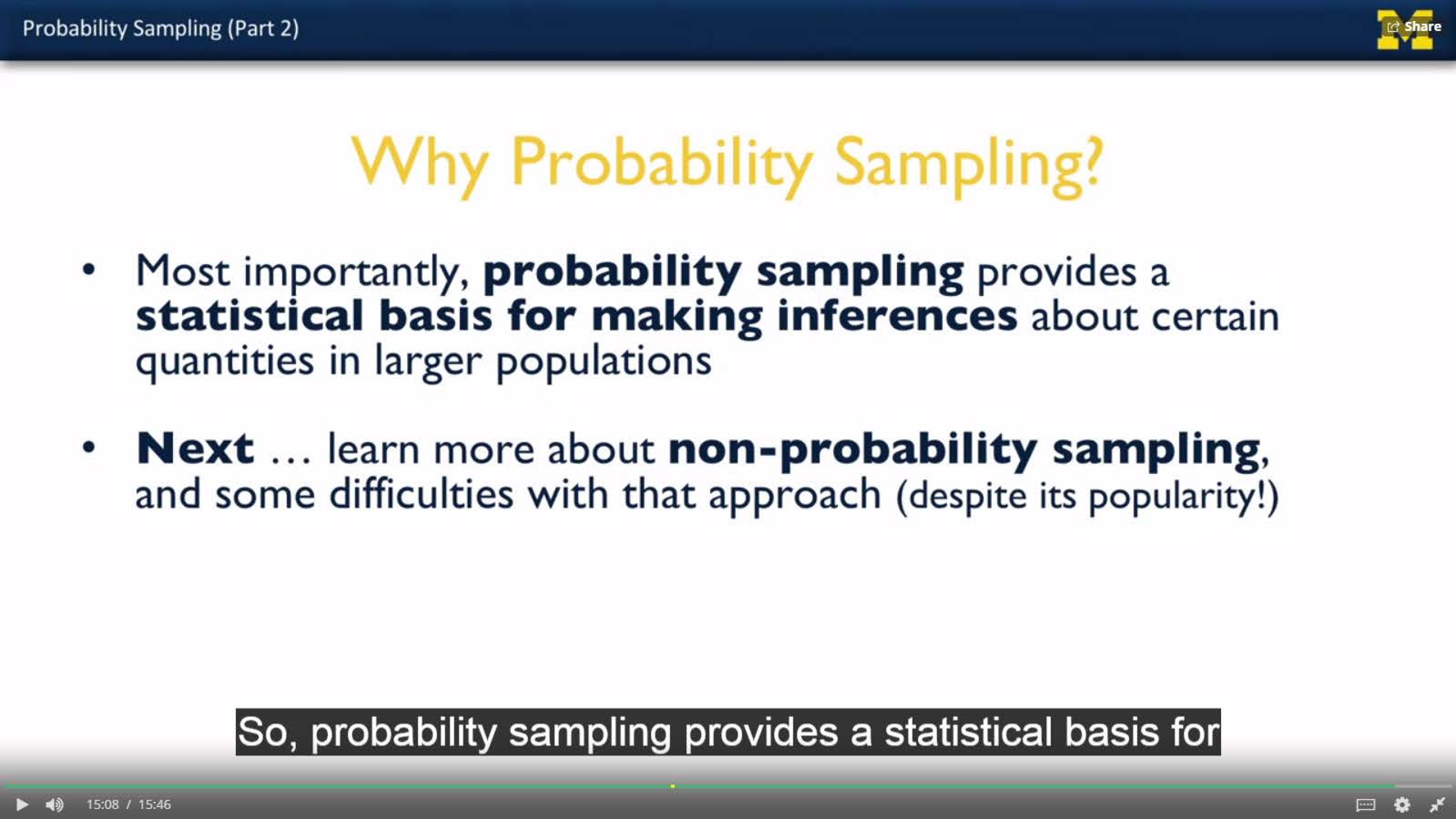


Probability Sampling (Part 2)

Why Probability Sampling?

- Having **known, non-zero probability of selection** for each unit in a population and **subsequent random sampling** ensures all units will have a chance of being sampled
- **Probability sampling** allows us to compute **unbiased estimates**, and also estimate features of the **sampling distribution** of estimates that we would see if many of the same types of probability samples were selected

and over, and over again using those probabilities of selection.



Probability Sampling (Part 2)

Why Probability Sampling?

- Most importantly, **probability sampling** provides a **statistical basis for making inferences** about certain quantities in larger populations
- **Next** … learn more about **non-probability sampling**, and some difficulties with that approach (despite its popularity!)

So, probability sampling provides a statistical basis for