Preserving Social Science Data: How Much Replication do we Need?

Myron P. Gutmann
Nancy Y. McGovern
Bryan Beecher
T.E. Raghunathan

University of Michigan

Main Points

- Tension between resources needed for preservation and the scarcity of resources
- Replication works, but can be costly and complicated
- How do we right-size our replication strategy?
- Specific attention to Social Science data (but mostly quantitative microdata)

Threat Models: What Might we Lose & Why?

- Total File or Collection loss
  - Technical failure
  - Operator failure
- Partial data loss
  - Technical failure (bit rot, other failure)
  - Operator failure (deletes record, variable, other)
- Intentional Destruction Possible (collection, file, case, variable)

Most Data are Samples

- Long tradition of using population samples to represent the population universe
- True even of some administrative data
  - Census samples for the “long form” & ACS
  - Public use census data are samples
- Not all cases necessarily have the same sample “weight”

Statistical Analyses

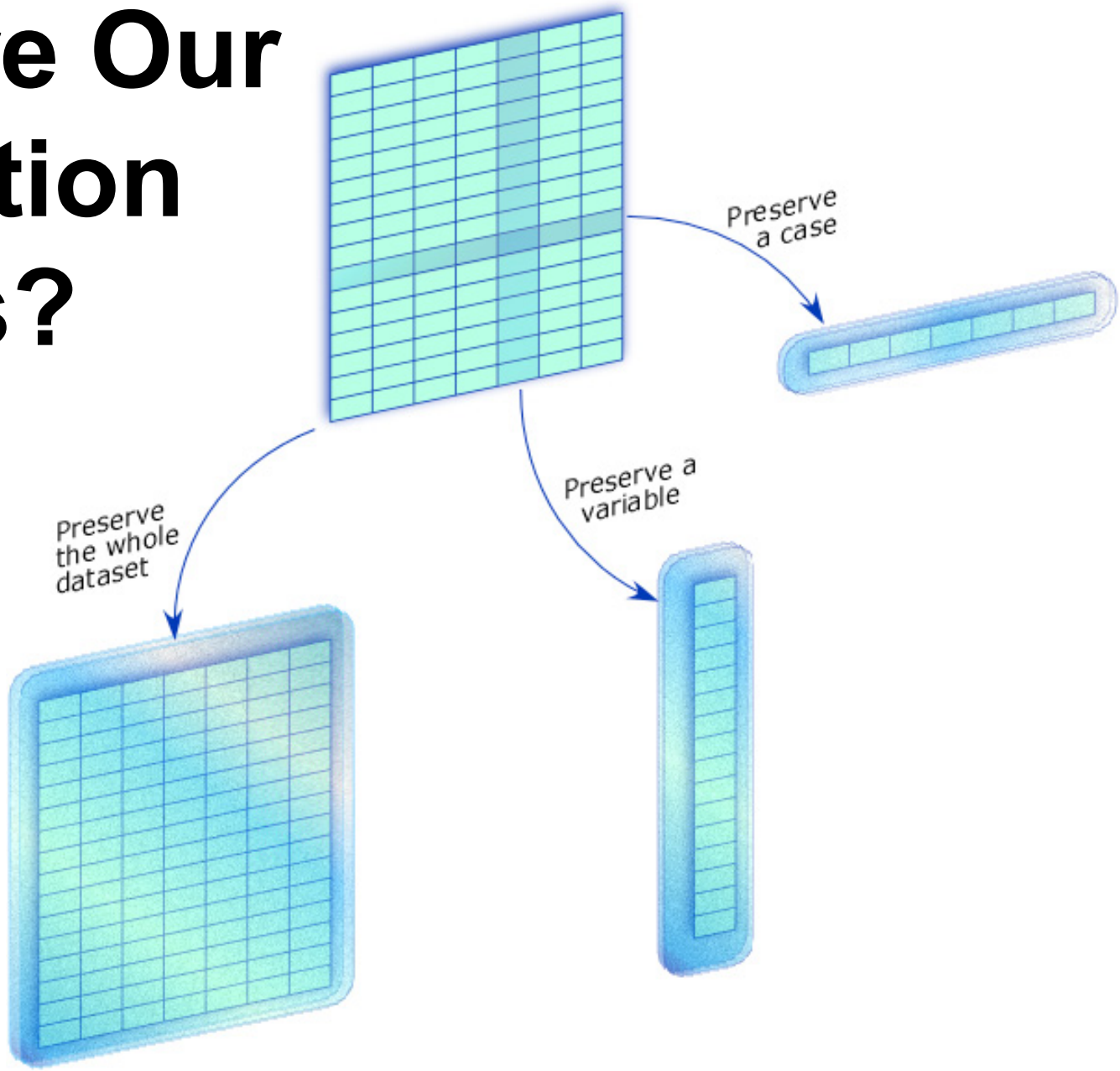
- Most sample data are analyzed using inferential statistics
- Question: is the relationship in the sample population strong enough to represent the whole population? (*few samples are really random*)
- Good methods for handling missing data
- Key finding: **possible to draw valuable conclusions from data that are sampled or otherwise incomplete**

Confidentiality Matters

- More and more social science data are not useful without confidential attributes
  - Geographic location
  - Information that might be harmful if revealed
  - Biomedical/Genetic information
- Distributing copies in multiple locations might make those data more vulnerable

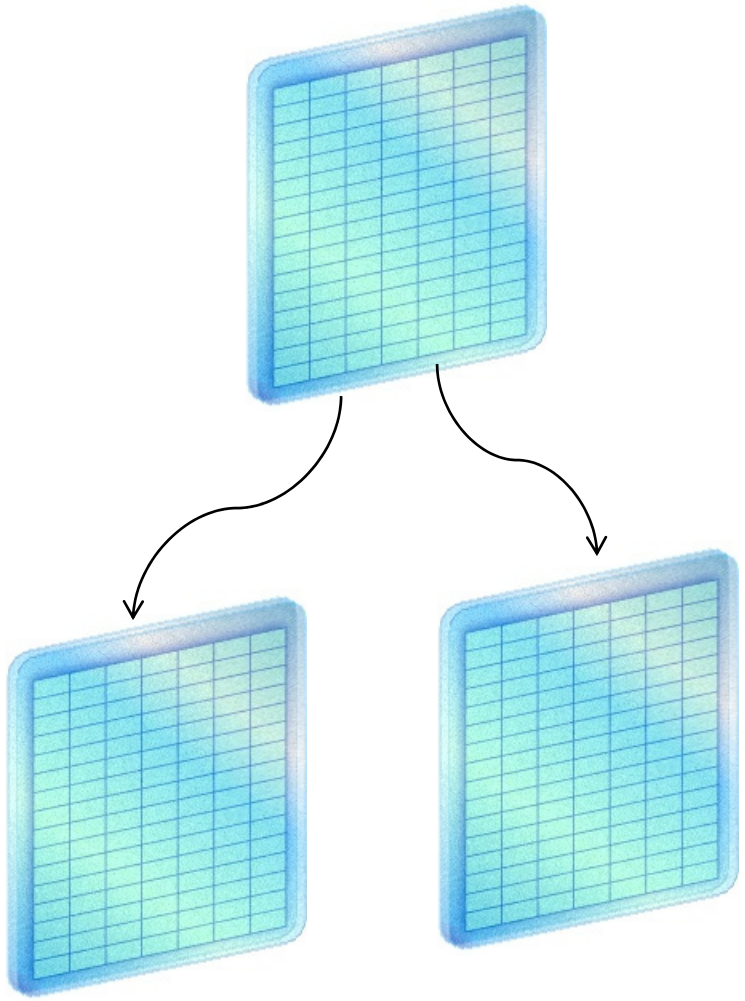
Preservation Standards for Social Science Data

What are Our Replication Options?



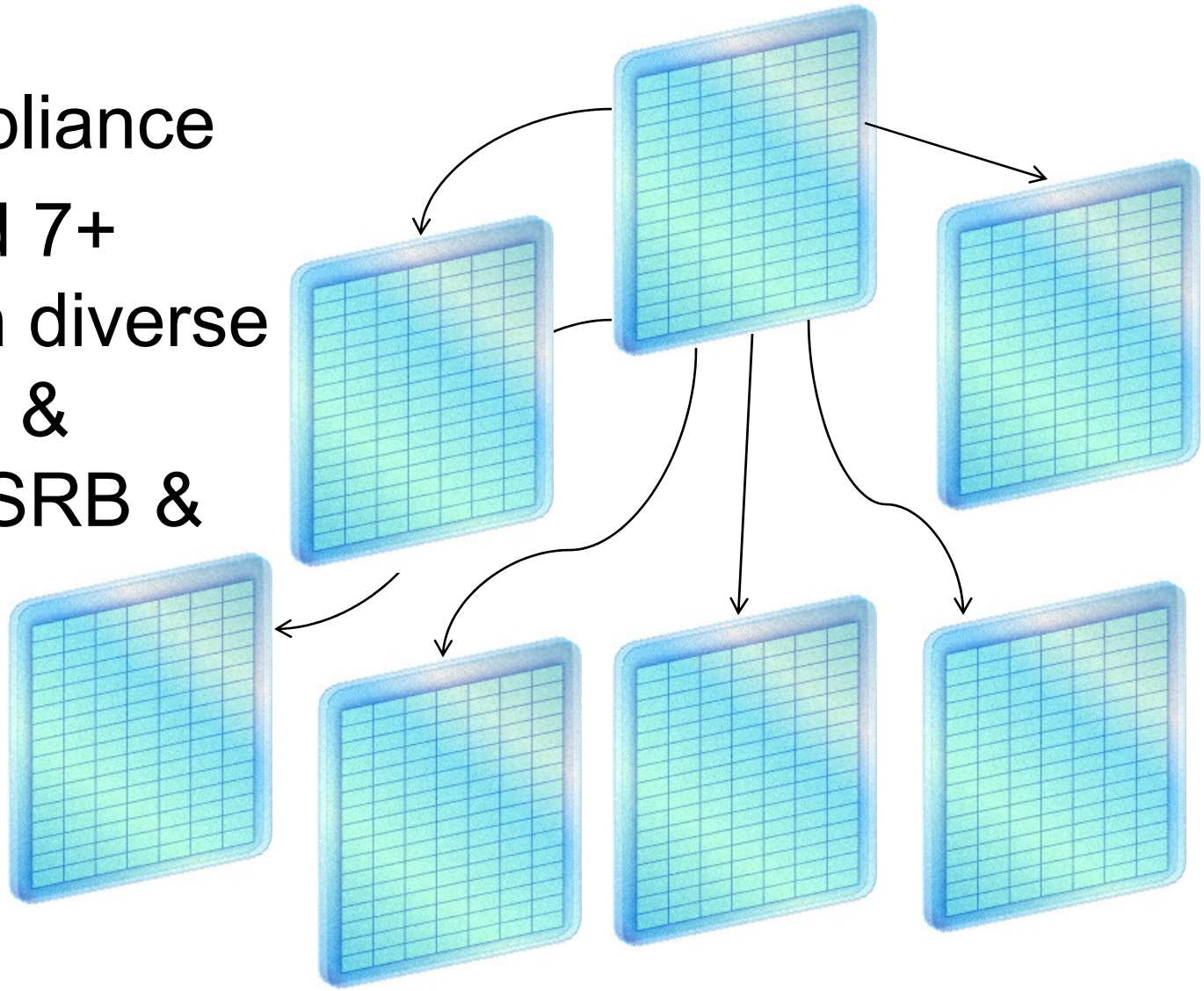
Old ways

- 60-Year history of data preservation
  - 2 to 3 copies on magnetic tape (one off-site)
  - Archive-specific preservation metadata
  - Very few losses



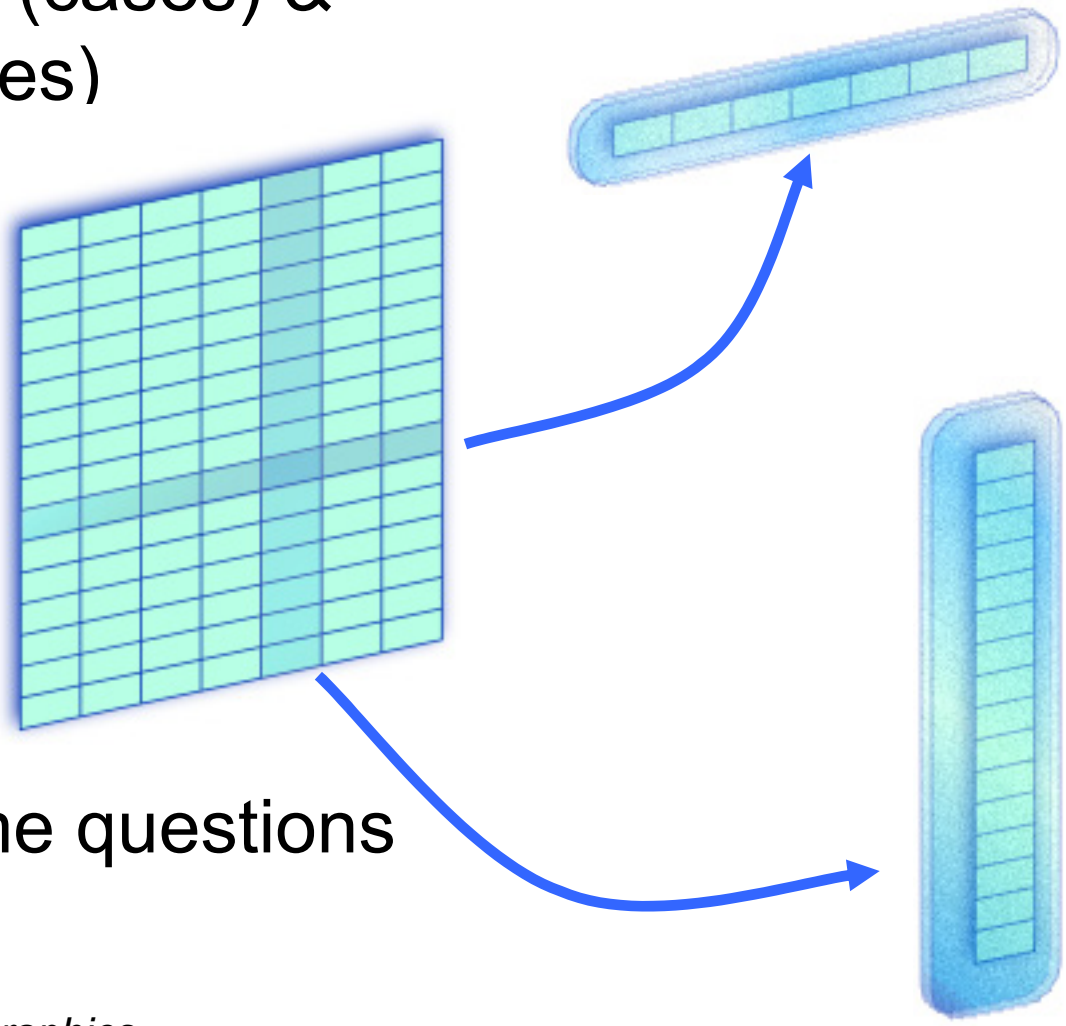
… and New

- OAIS compliance
- Disk-based 7+ copies with diverse technology & locations (SRB & LOCKSS)



My Key Concepts

- Data are in rows (cases) & columns (variables)
- Cases are generally sampled
- Potential re-spondents may decline to participate, or only answer some questions



**Thanks to Cole Whiteman for the Graphics*

Data Preservation Today

- High level of granularity allows single data element (variable within case) to be extracted from the preserved record
- Non-proprietary encoding
- No compression

- Right-sizing replication requires paying attention to these attributes

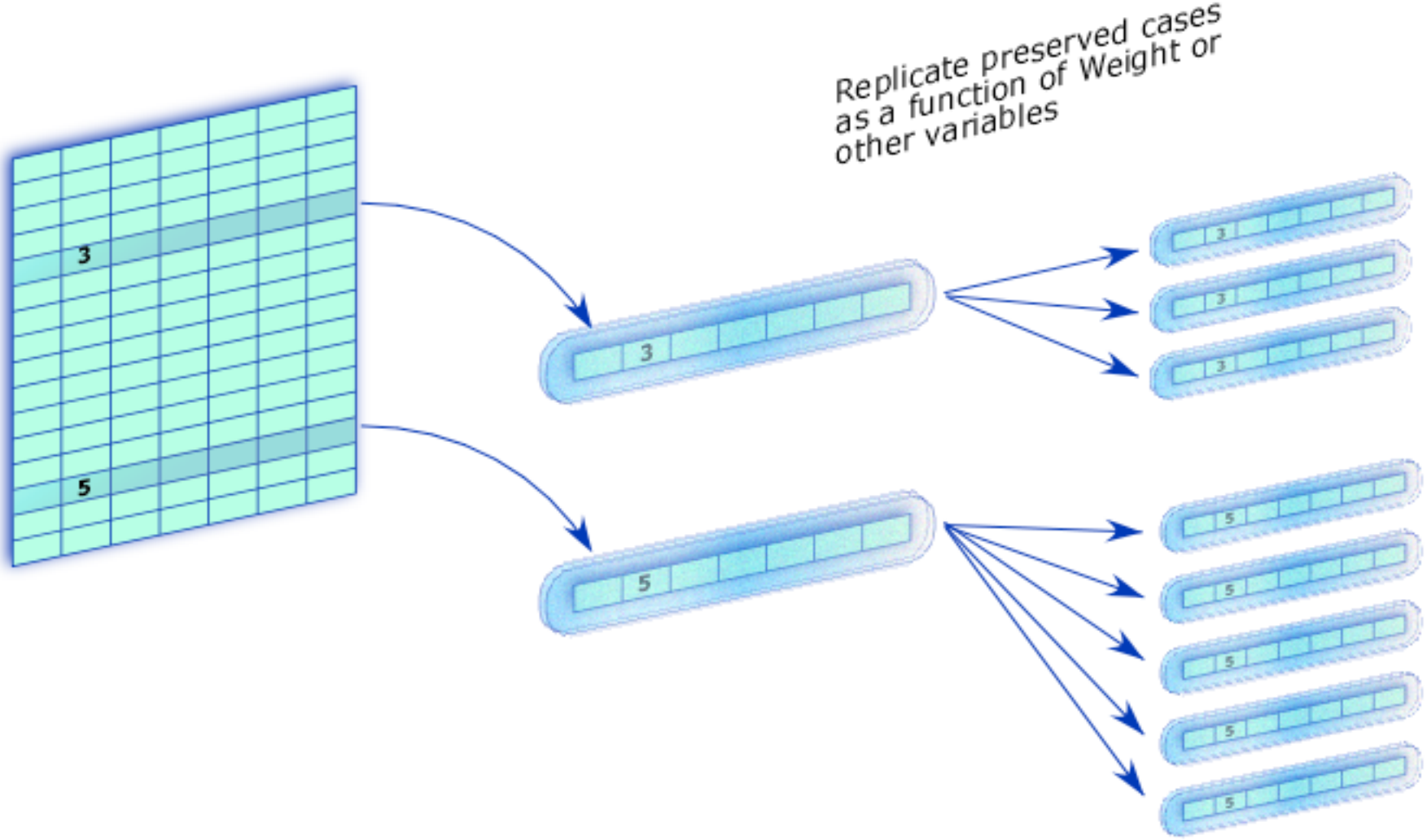
Five Key Questions

1. What is the right level of Granularity for Replication of Social Science Objects?
2. Do some objects require more replicas than others?
3. Is it possible to reduce replication for some objects?
4. Are diverse devices always required?
5. Does confidential data matter?

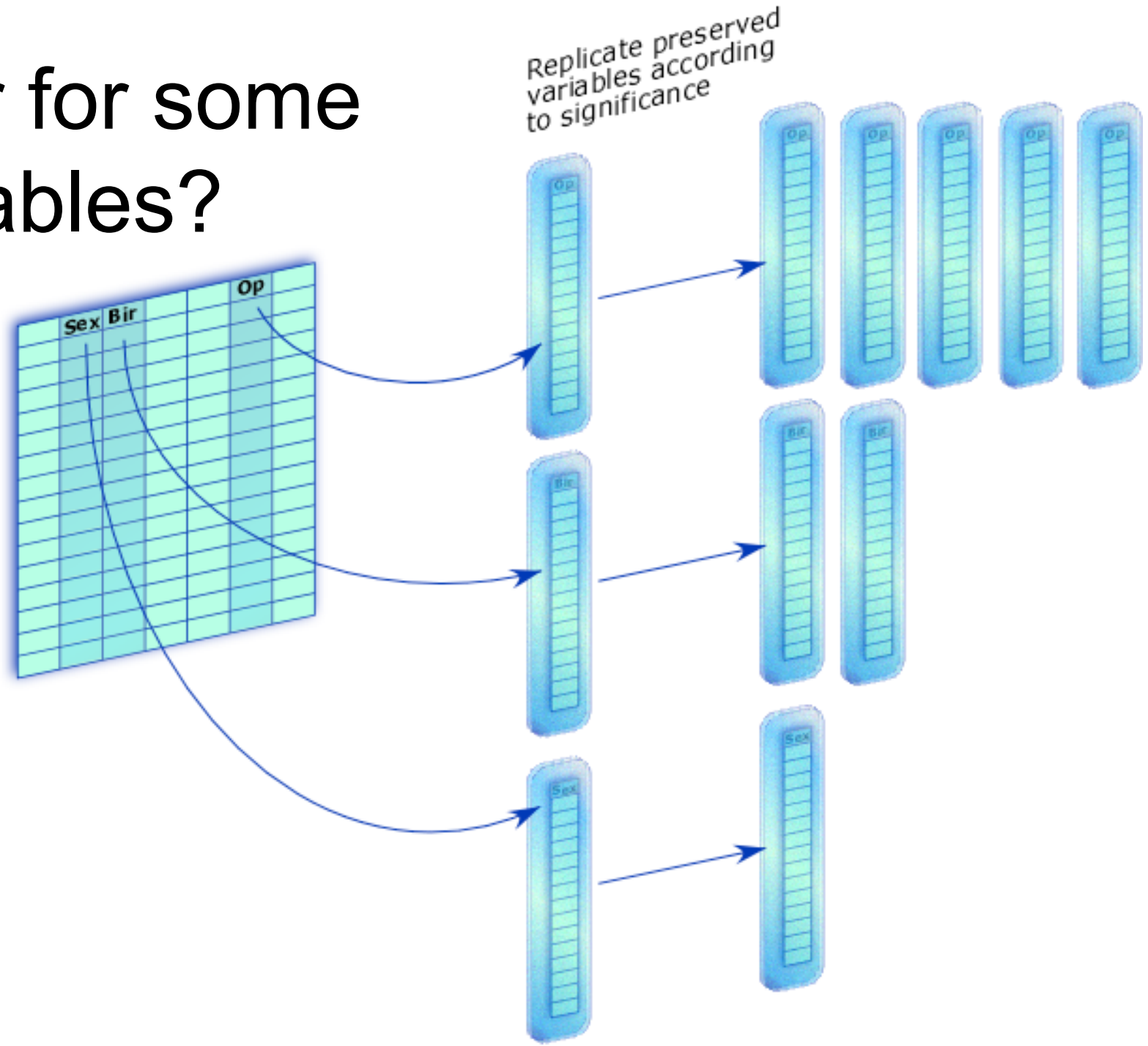
Answers: Level of Granularity

- **The case** is probably best, but should be able to identify an object as small as a variable
- Challenge: appropriate audit strategy for case/variable, rather than file

More Replicas for Some Cases?



... or for some variables?



Do Fewer Replicas Ever Work?

- Sampled data could require fewer replicas than universe data, if...
  - Potential unit of loss is the case/variable
  - Replication and audit can operate at those levels
- Effective inference can compensate for sample size, sample weights, missing data
- No argument for fewer than required for Byzantine fault tolerance

What about Diverse Devices?

- Possible to think about reducing diversity of devices, for ...
  - Within-case replication (multiple replicas of a single variable within a case)
  - Within-study replication (multiple replicas of a single case within a study)
- Ensures against failure at the bit-loss level but not necessarily at the device- or file-level

Does Confidentiality Matter?

- Confidential data requires the same replication approach as non-confidential data, but has its own issues
- Public replication schemes with multiple locations are potentially dangerous
- New approaches are required

Summary

- Social Science data preservation allows more replicas for some objects than others
- Sampling allows fewer replicas than otherwise required
- Diverse replication strategies may be warranted
- Confidential data requires the same number of replicas, but increased control

Costs vs. Benefits

- Benefits:
  - New approach takes into account the nature of social science data
  - May lower costs if very many replicas are the only option and high granularity is possible
- Costs:
  - New tools for managing & auditing replicas
  - Need to study this approach, probably with simulation studies

gutmann@umich.edu

THANK YOU!