MLDS Center

Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

The MLDS Synthetic Data Project

Laura Stapleton, UMCP
Angela Henneberger, UMB
(and a host of many others!)

What is the Synthetic Data Project?

- In 2015, the State of Maryland received a grant from the U.S. Department of Education's State Longitudinal Data Systems program; one of the funded projects was to create a synthetic data system of the data in the MLDS.
- Synthetic data are generated based on models to mimic the relational patterns among variables, so statistical analyses with such "fake" synthetic data should yield findings substantially similar to the real data
- Simultaneously, reduces the risk of privacy breach

MLDS Center
Maryland Longitudinal Data System

Questions in Development of the Maryland SDP

- What challenges arise in the process of creating synthetic data from a statewide longitudinal data system?
- What are the best methods for assessing the quality of the synthesized data?
- How successfully do the synthesized data fulfill the needs of the MLDS Center to provide accessible data that can inform policy while protecting individual privacy?
- To what extent do end users (applied researchers) find the synthetic data useful, and to what extent are the data actually used in analyses that inform policy?

The Process…

We needed to split the project into three broad steps:

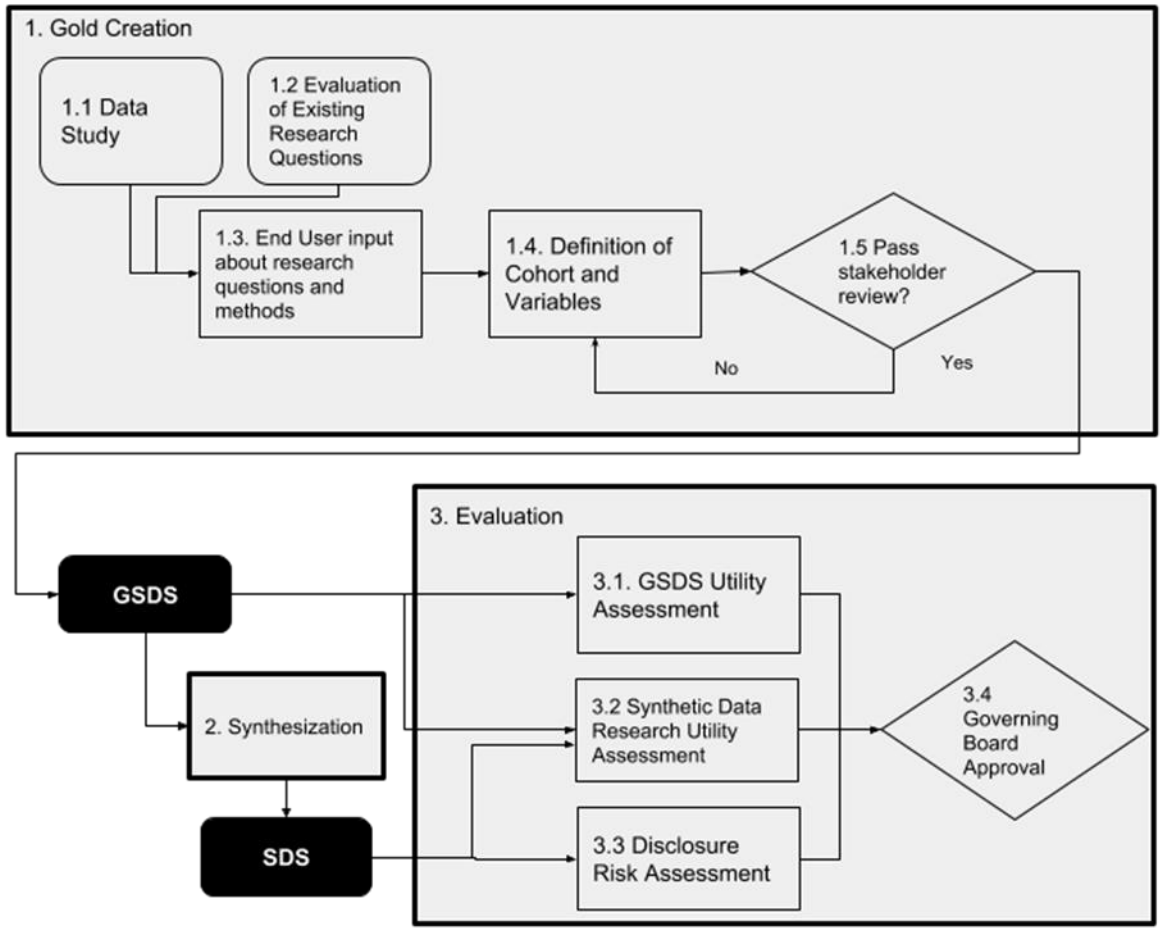
1) creation of gold standard datasets (GSDS),

2) synthesis of the GSDS, and

3) evaluation of the utility and safety of the synthetic data sets (SDS)

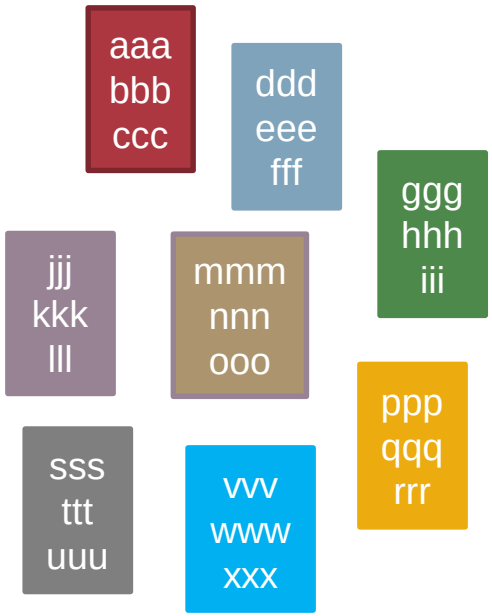
Bonnéry, D., et al. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*. doi.org/10.1080/19345747.2019.1631421

The Process, continued…



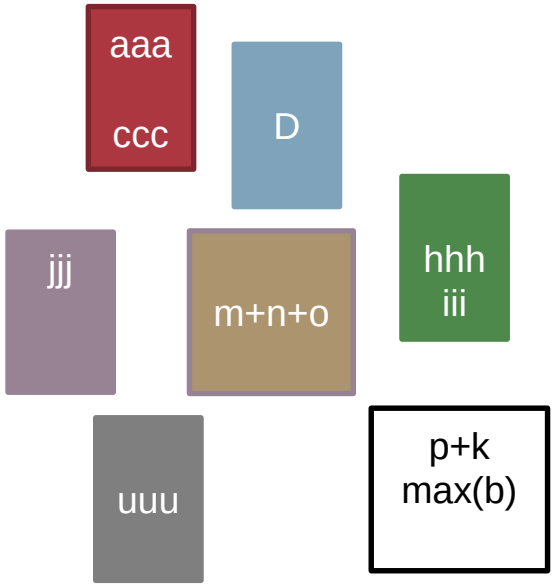
Step 1: Gold Dataset Creation…

Operational Data Store (ODS) (v=460)



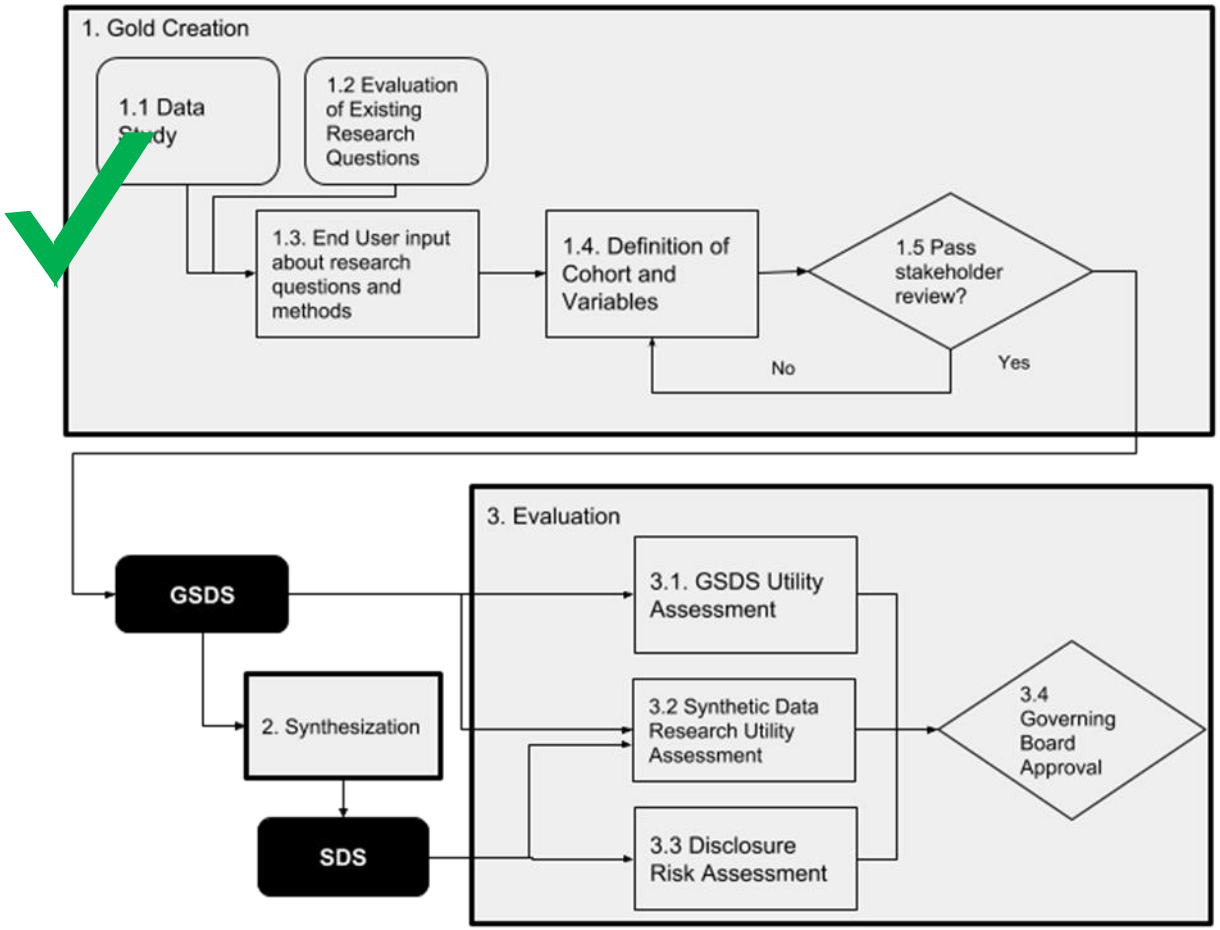
Gold Standard Data Set (GSDS) (v=65, 50, 55)

(But there are many rows of data per person!)



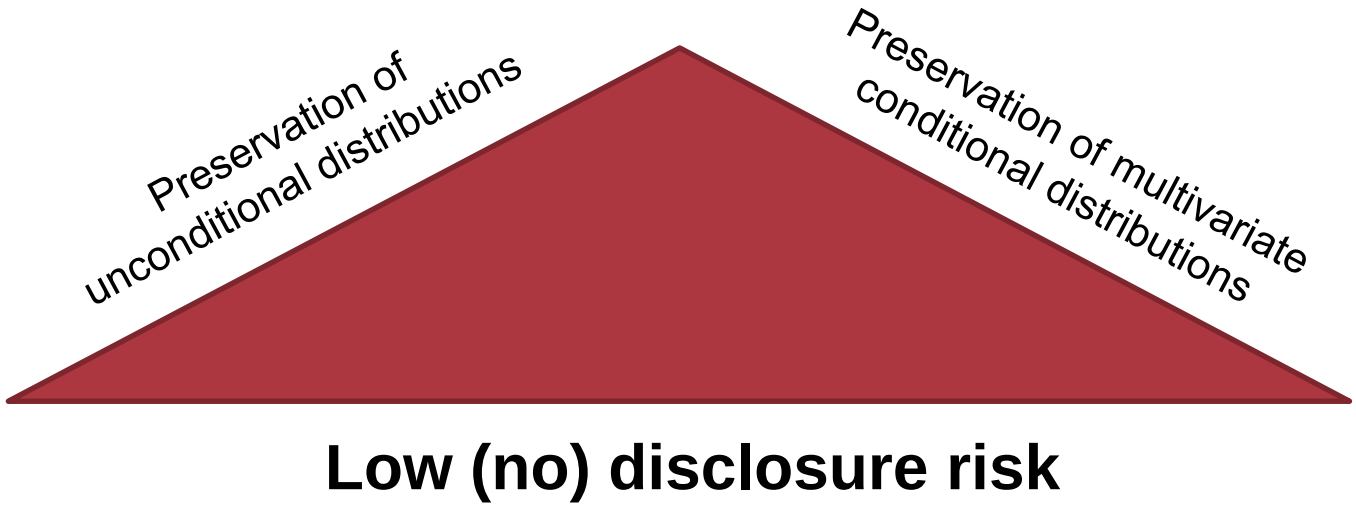
Deep dive into variable definition and possible variable use

The Process, continued…



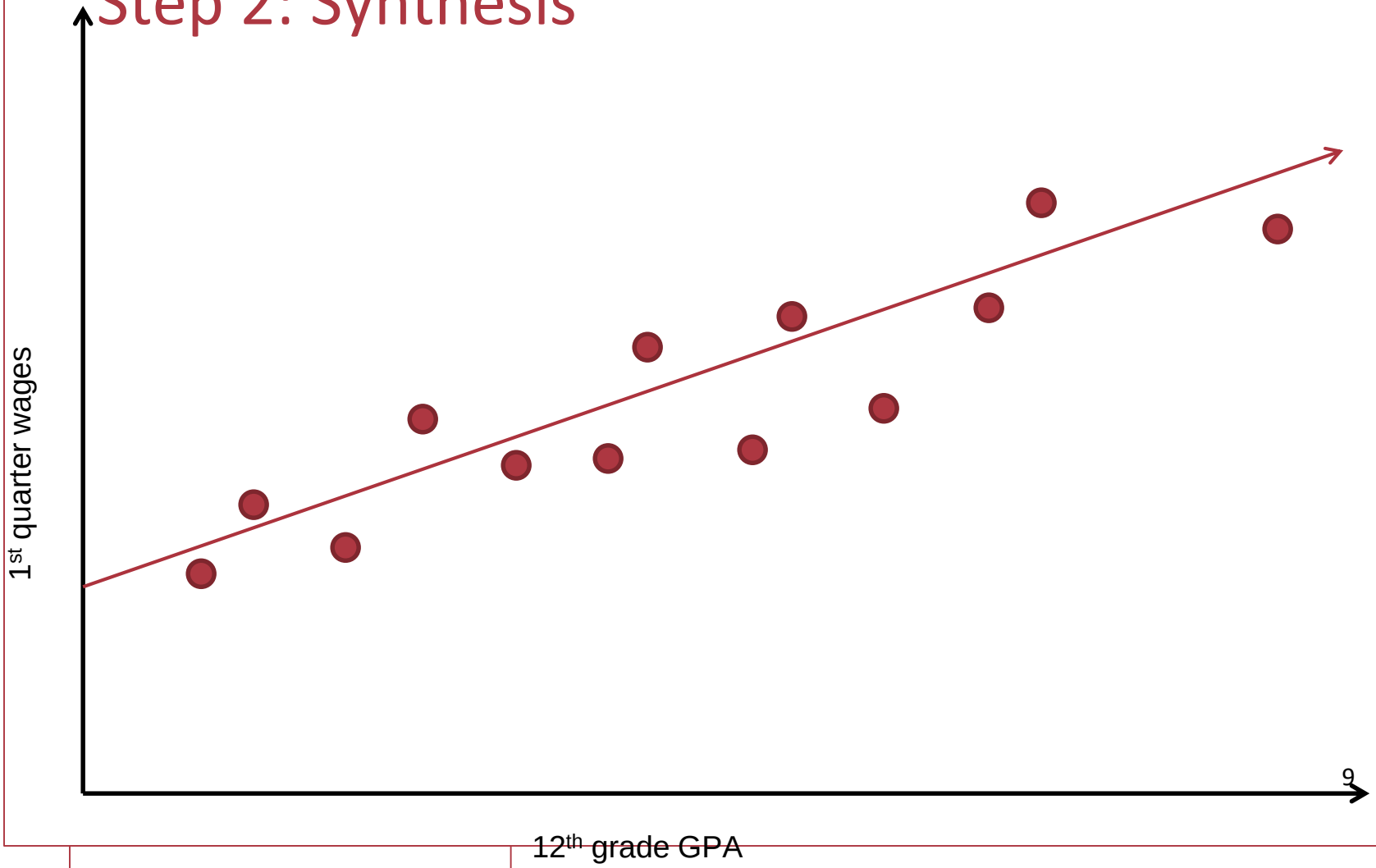
Step 2: Synthesis

- We need to satisfy a triangular trade-off:



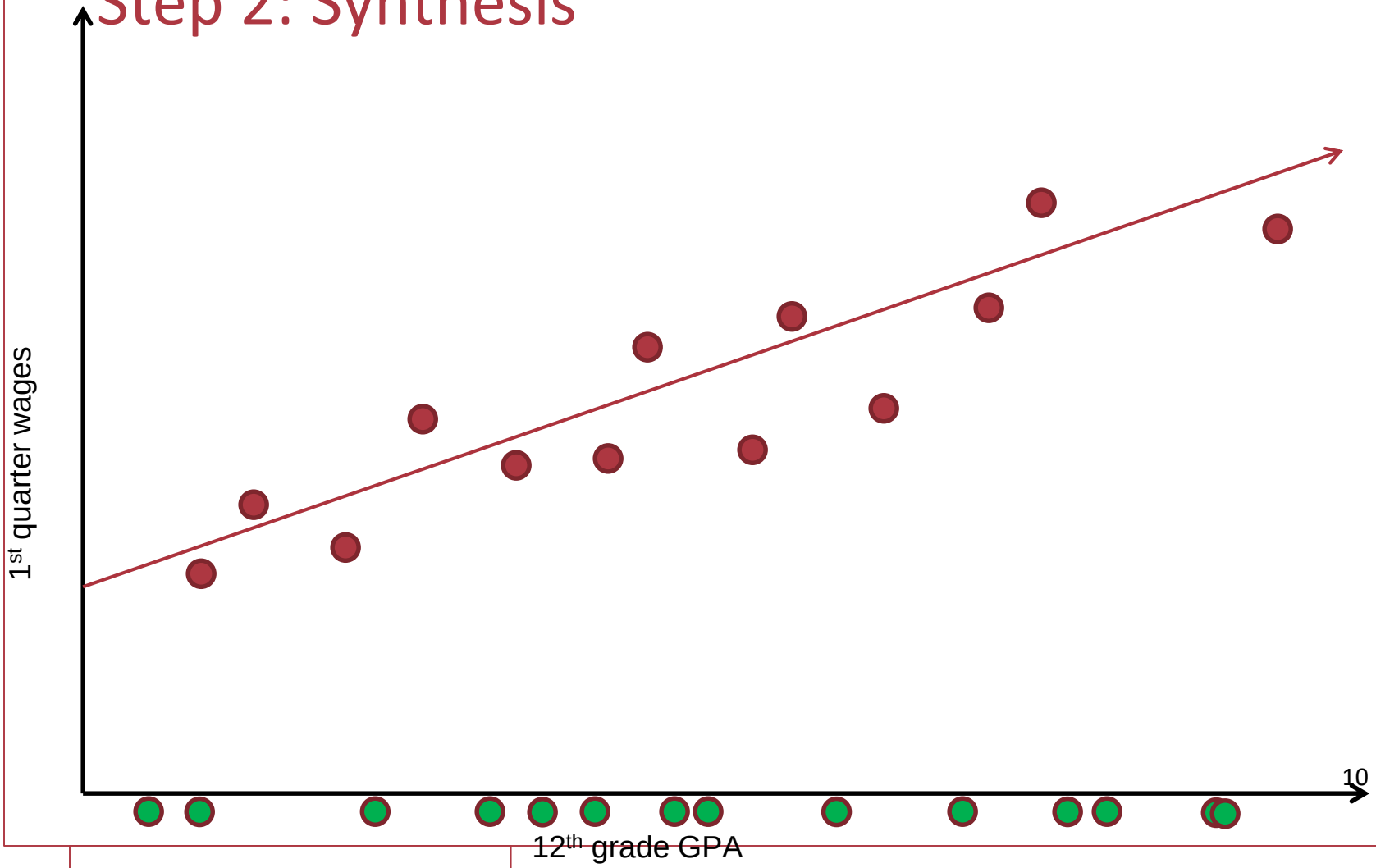
MLDS Center
Maryland Longitudinal Data System

Step 2: Synthesis



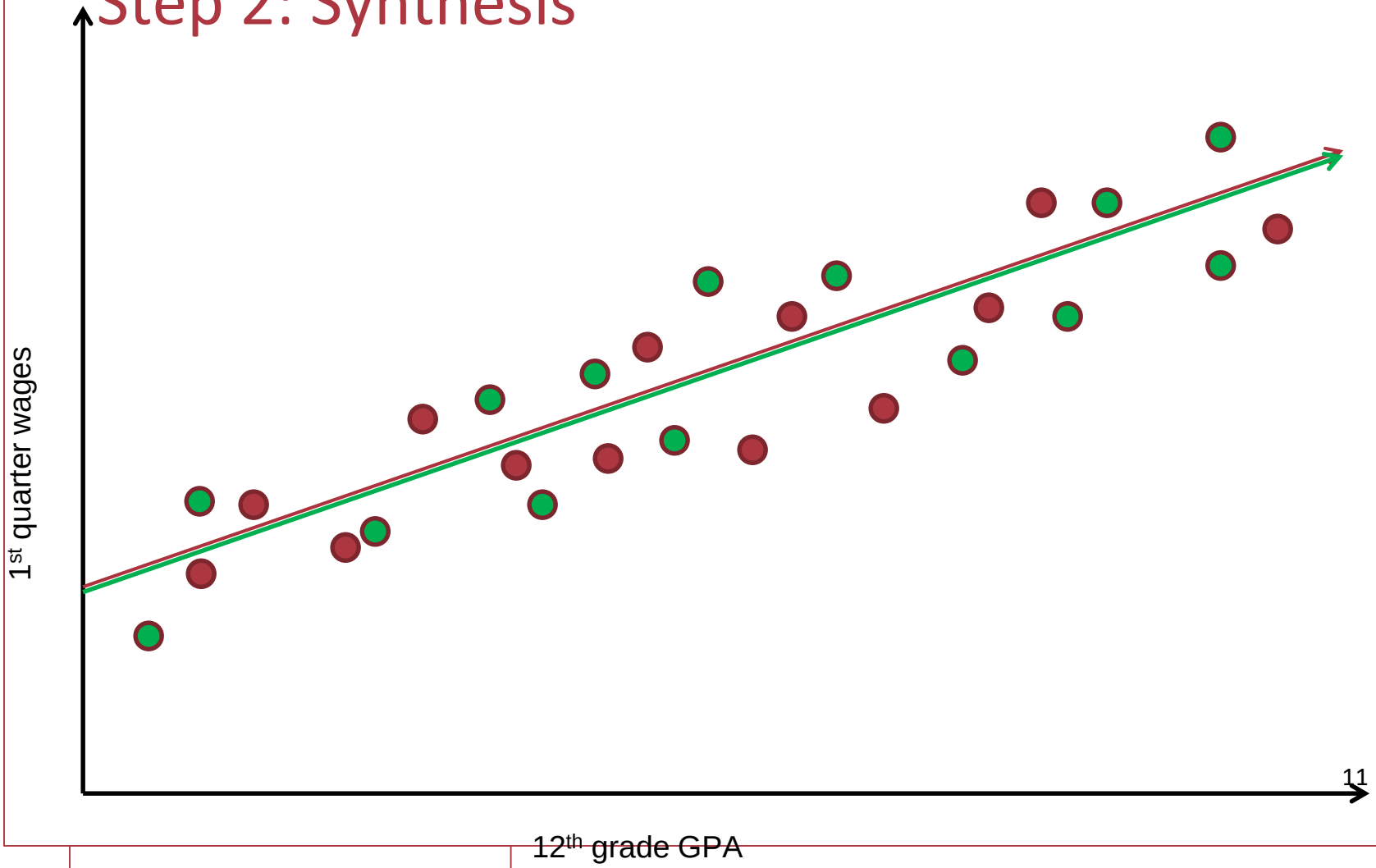
MLDS Center
Maryland Longitudinal Data System

Step 2: Synthesis



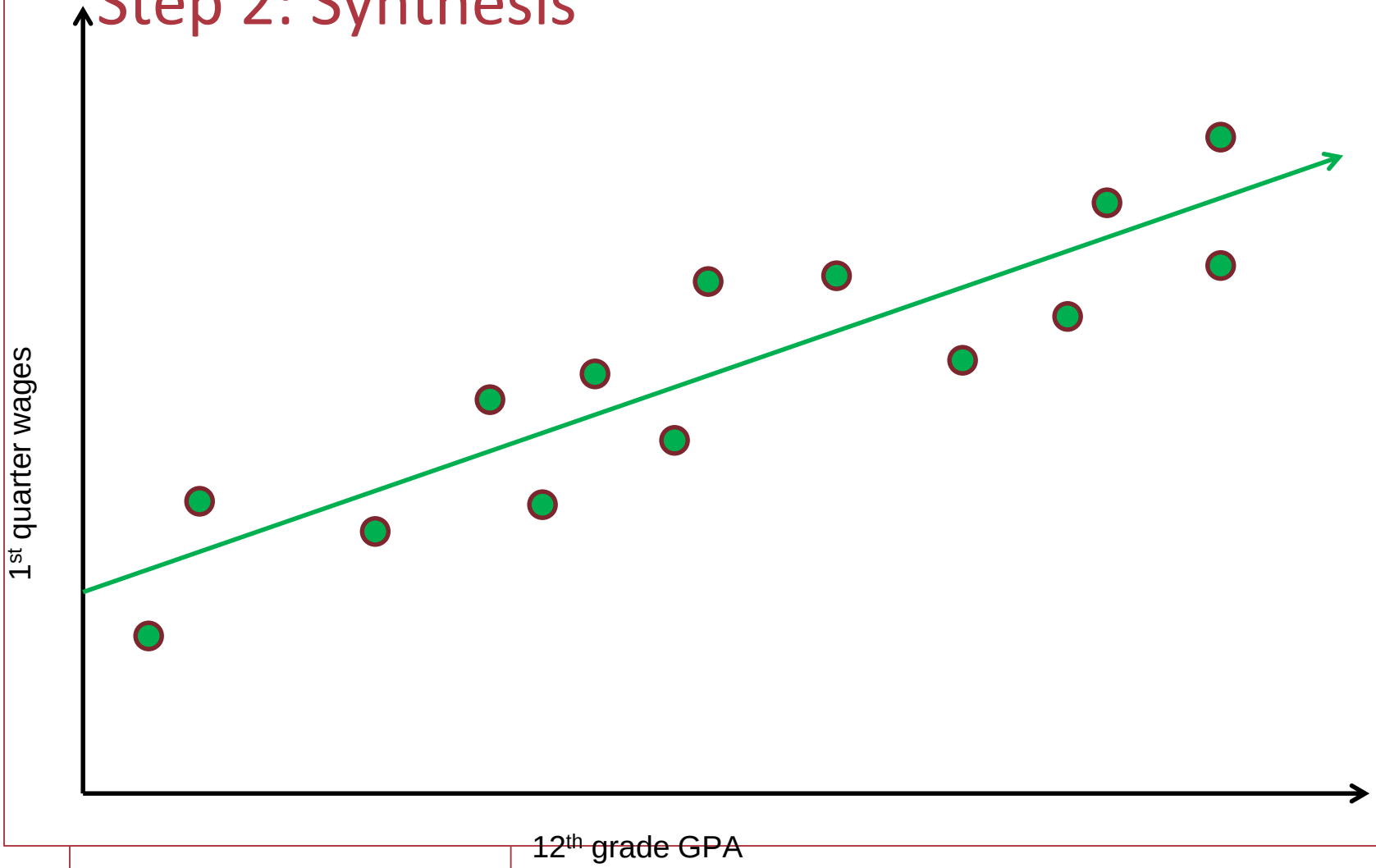
MLDS Center
Maryland Longitudinal Data System

Step 2: Synthesis



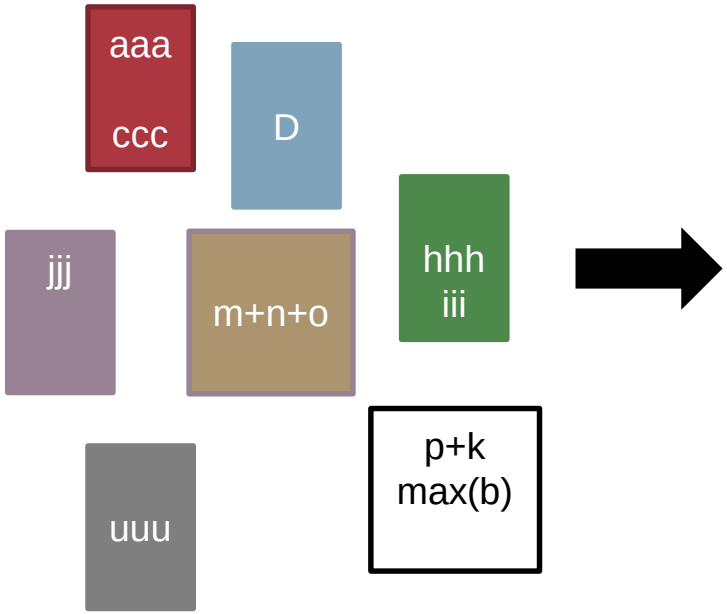
MLDS Center
Maryland Longitudinal Data System

Step 2: Synthesis



Step 2: Synthesis

Gold Standard Data Set (GSDS) (v=65, 50, 55)



Transformed (v=4000, 4700, 5900)

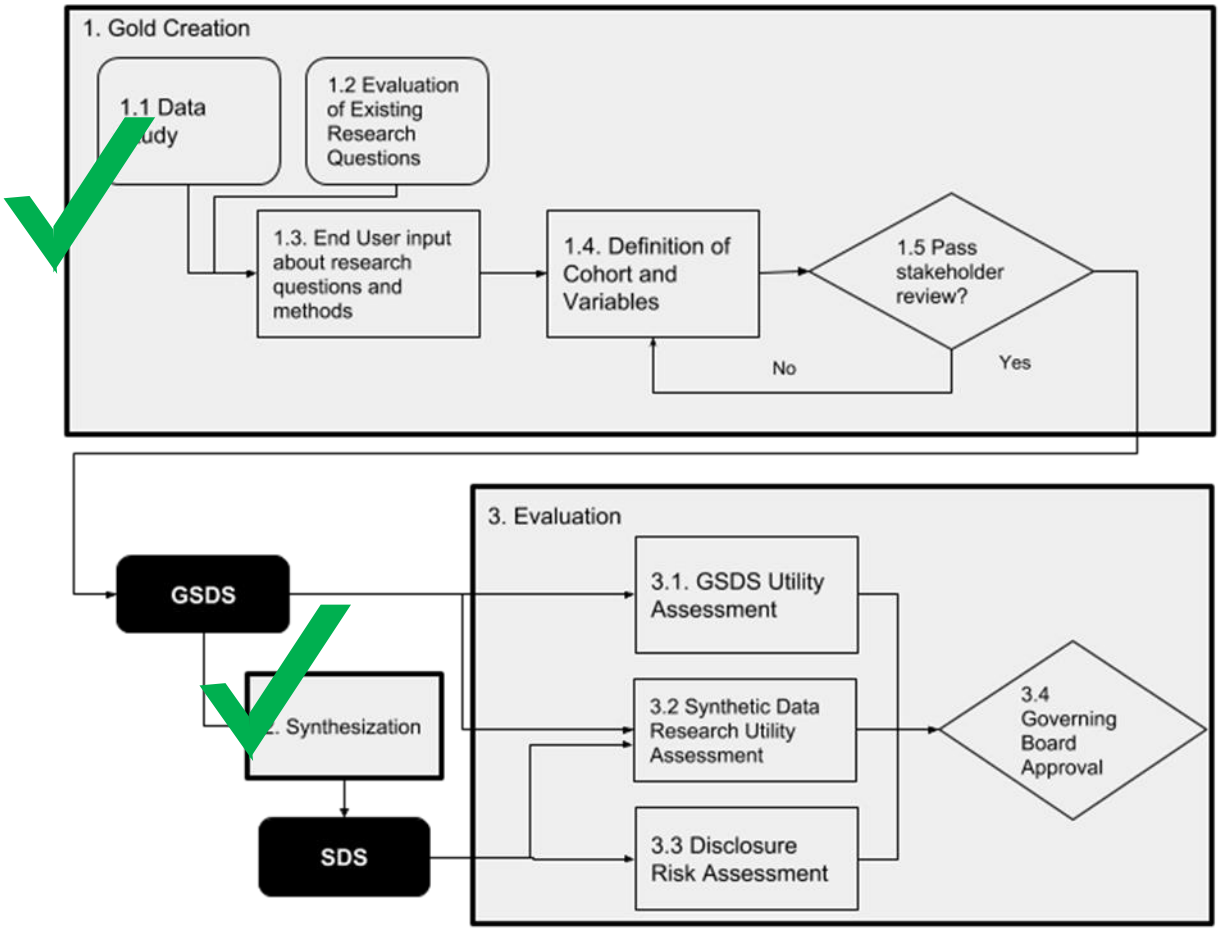
For synthesis, we need one wide record per individual

a1 a2 a3 a4 a5 a6 c1 c2 c3 c4 c5 c6 D1 D2 …max(b)

Step 2: Synthesis

- Given the sheer number of variables (in wide format) and the potential for interactions and non-linearities….
- After initial testing and evaluation of the different existing methods, the decision was made to implement the CART method (described in Reiter, 2005b)
- A CART is the outcome of a general empirical method to model a dependent variable conditionally to a set of predictor variables. It partitions the joint predictor space obtained after applying a binary partition recursively.

The Process, continued…



Step 3: Evaluation

- **GSDS utility assessment (Step 3.1)**

  Are the GSDS data useful themselves?

- **Synthetic data research utility assessment (Step 3.2)**

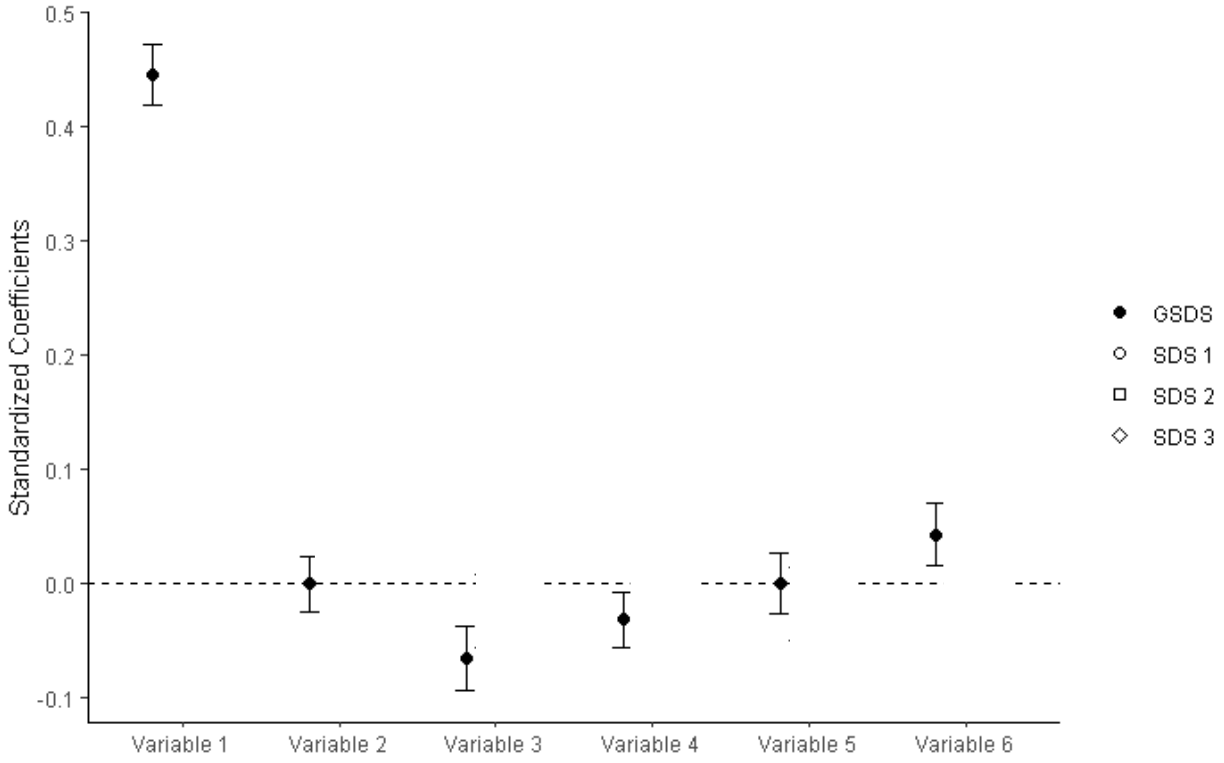
  Do you get the "right" answer from the synthetic data?

- **Disclosure risk assessment (Step 3.3)**

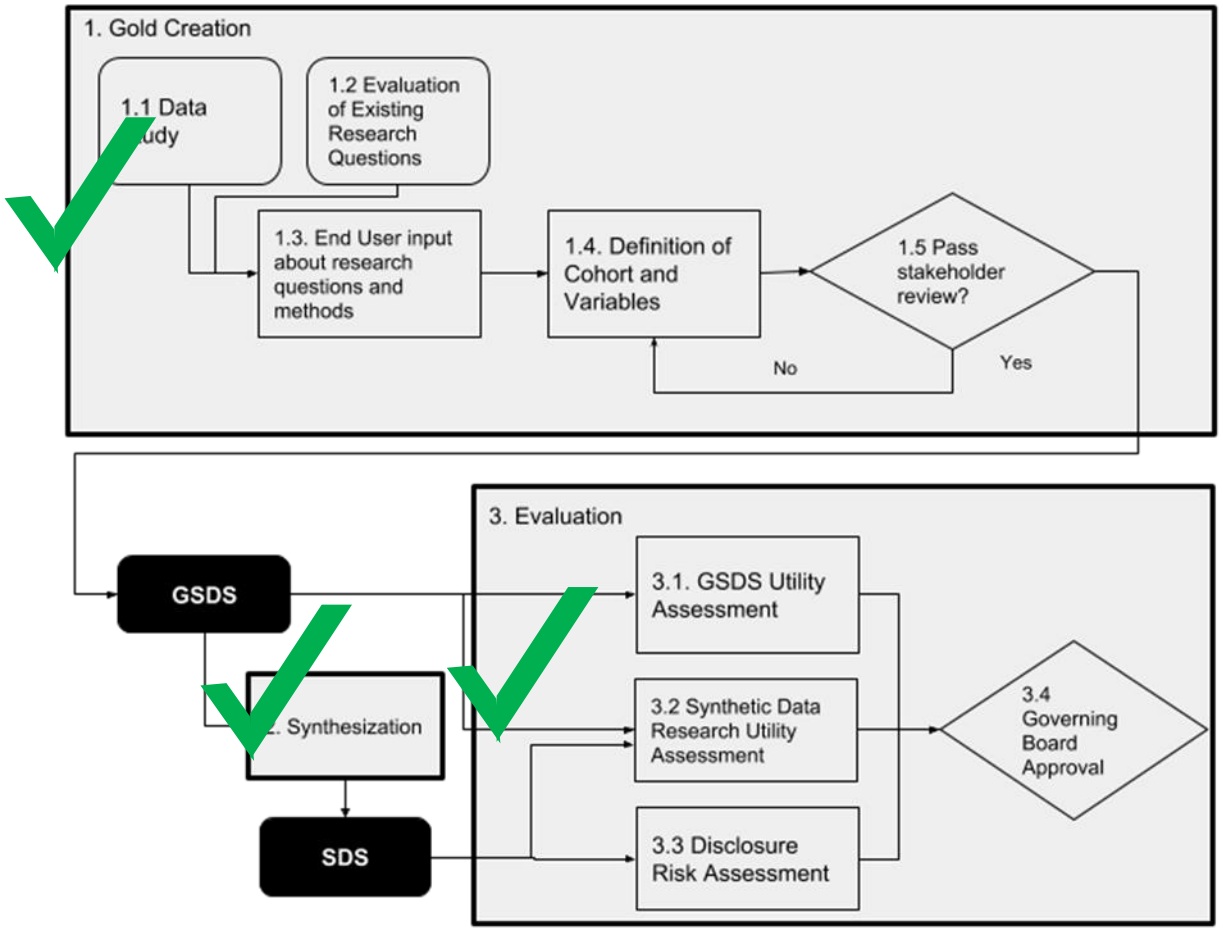
  Do the synthetic data pose a risk of disclosure?

Step 3: Evaluation Example

Comparisons of Standardized Multiple Regression Coefficient Estimates



The Process, continued…



MLDS Center
Maryland Longitudinal Data System

Key Recommendations and Lessons Learned

- Importance of buy-in throughout the process
- Legal infrastructure may create barriers
- Many decisions need to be made – careful selection of stakeholders for each set of decisions
- Clarity in goals for the synthetic data (e.g., training or research)
- Do not underestimate the time, cost, and expertise needed

MLDS Center
Maryland Longitudinal Data System

Informing a Decision about Synthetic Data as a Data Sharing Strategy

- **Safe Data Sharing Strategy:** Our findings indicate that synthetic data are a safe method for sharing sensitive protected data
- **Robust for Research:** Data can be synthesized such that, when analyzed, findings reflect the "real" data
- **Leverage Human Capital:** Synthetic data can be a way to increase the number of people using statewide data, expanding what is learned and advancing research-informed decision-making

MLDS Center
Maryland Longitudinal Data System

Thank you!

- Contributors:
  - Daniel Bonnery, Yi Feng, Angie Henneberger, Tessa Johnson, Mark Lachowicz, Bess Rose, Terry Shaw, Laura Stapleton, Mike Woolley
  - Email: Lstaplet@umd.edu

- Acknowledgement:
  - This presentation was prepared by the Research Branch of the Maryland Longitudinal Data System Center (MLDSC) as part of funding from the U.S. Department of Education (R372A150045)
  - The Research Branch would like to thank the entire staff of the MLDSC for their assistance with the work and the presentation.