Synthetic Data as a Strategy for Expanding Access to Linked Administrative Data in Prevention Science

Angela K. Henneberger, Bess A. Rose, Laura M. Stapleton, & Michael E. Woolley

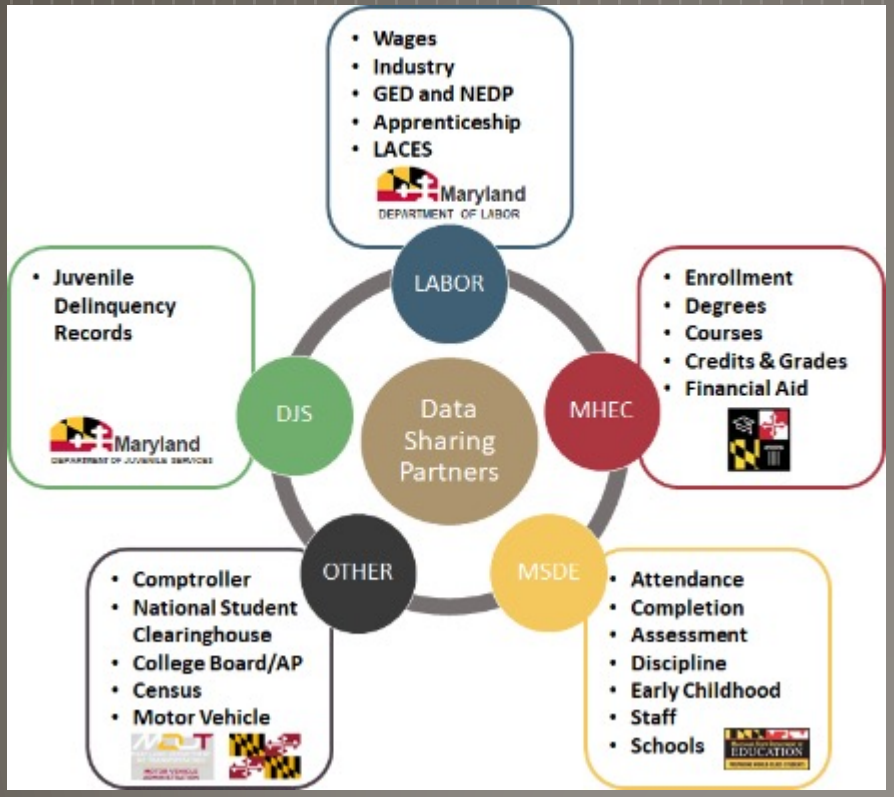
Society for Prevention Research
June 3, 2021

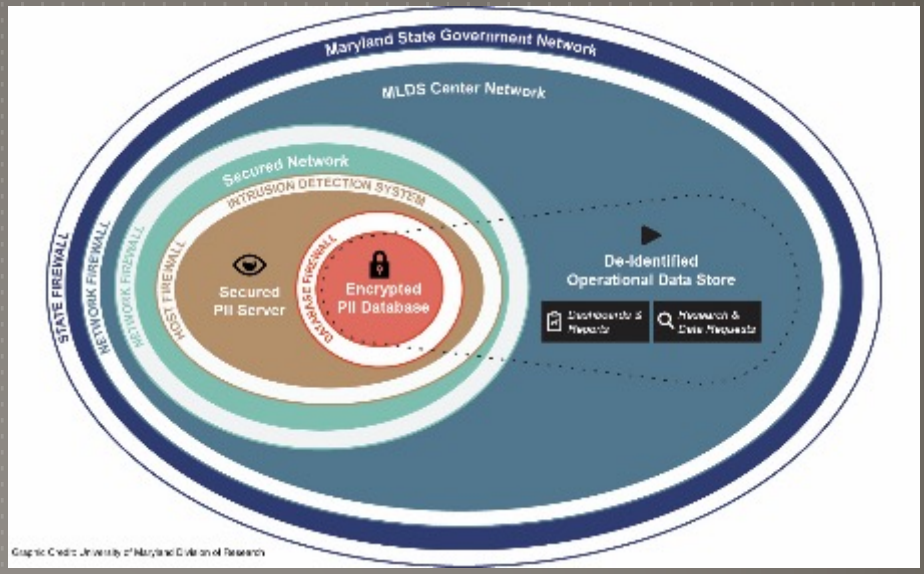
STRENGTHENING SOCIETY

University of Maryland
School of Social Work

Balancing Data Access and Privacy

The MLDS Center

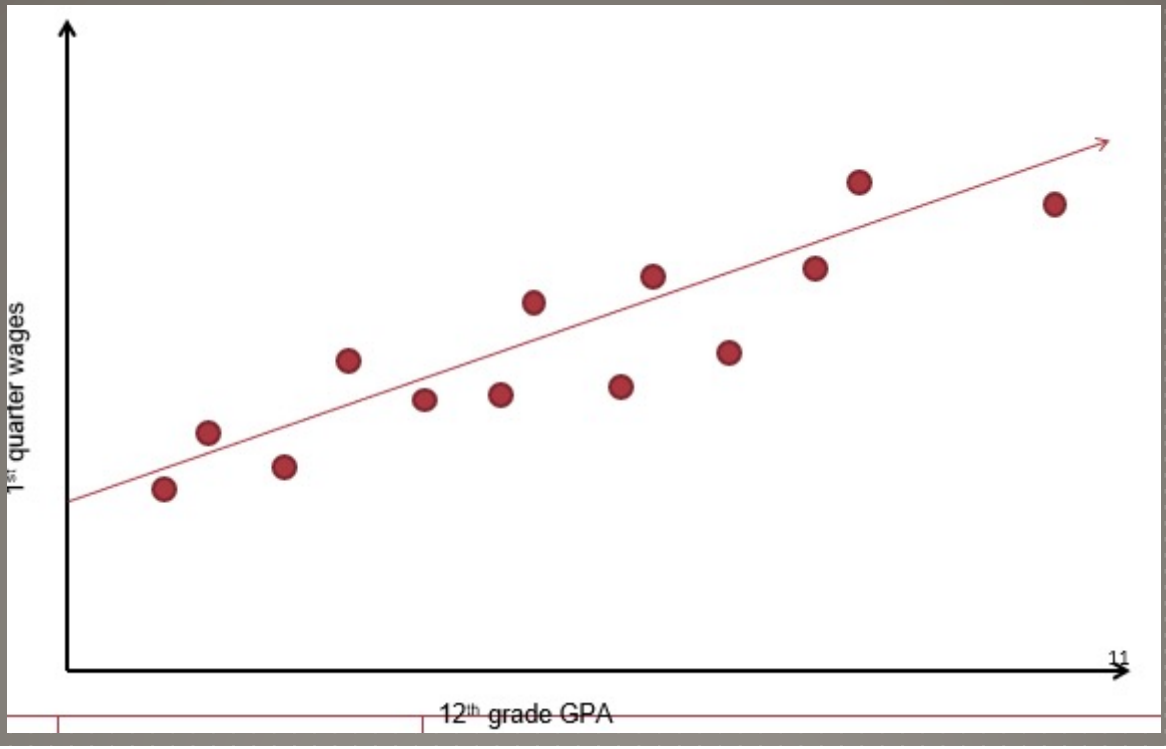




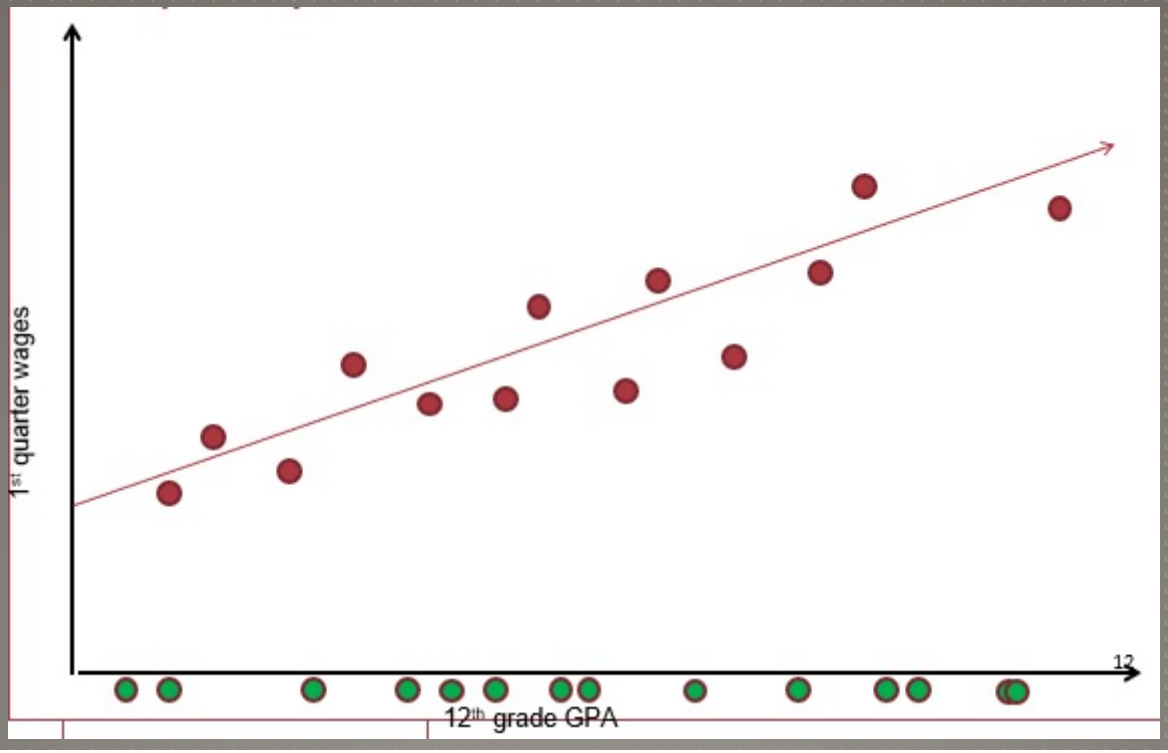
Maryland's Synthetic Data Project

- 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, known as *research utility*, so statistical analyses should yield similar findings to the real data
- Synthetic data also minimizes risks of privacy breach due to *attribute* or *identification disclosure risk*

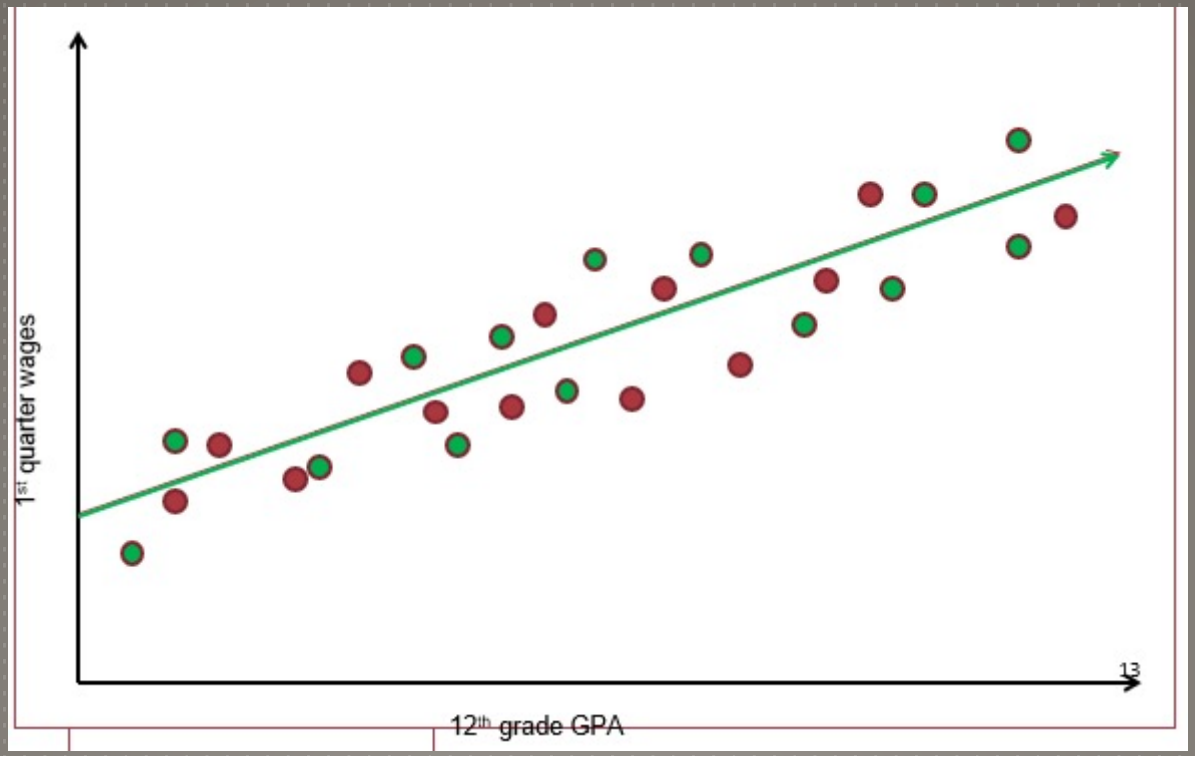
What are Synthetic Data?



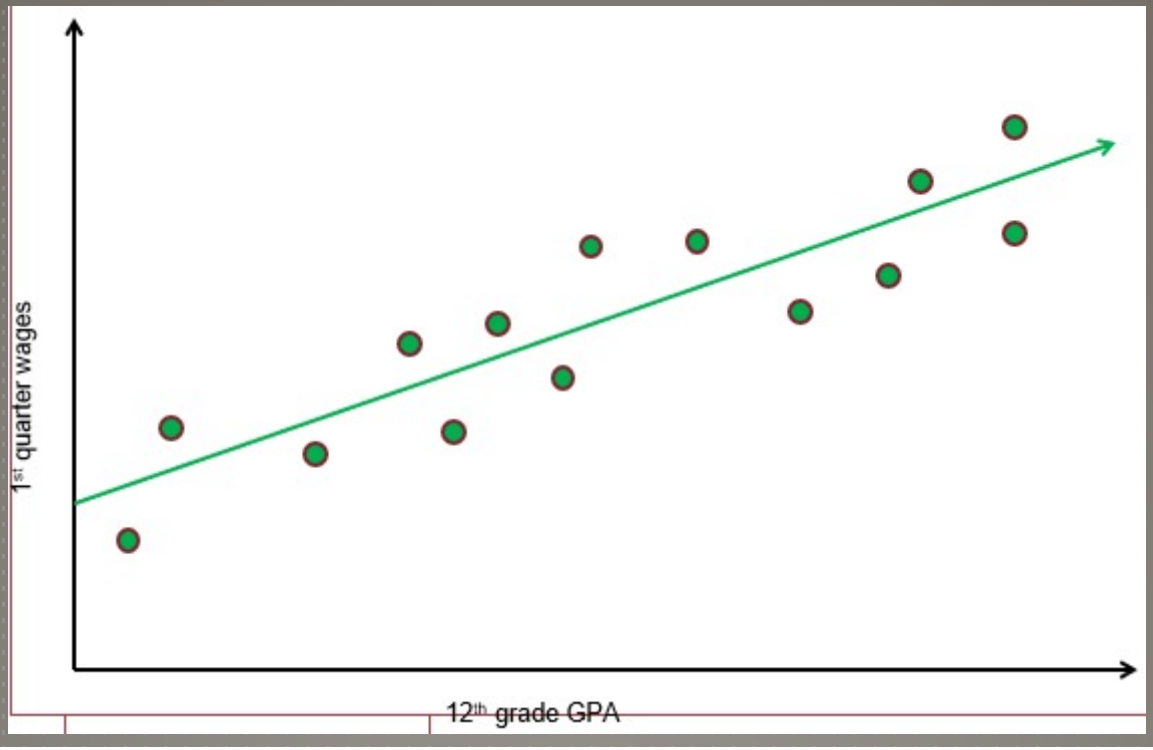
What are Synthetic Data?



What are Synthetic Data?



What are Synthetic Data?



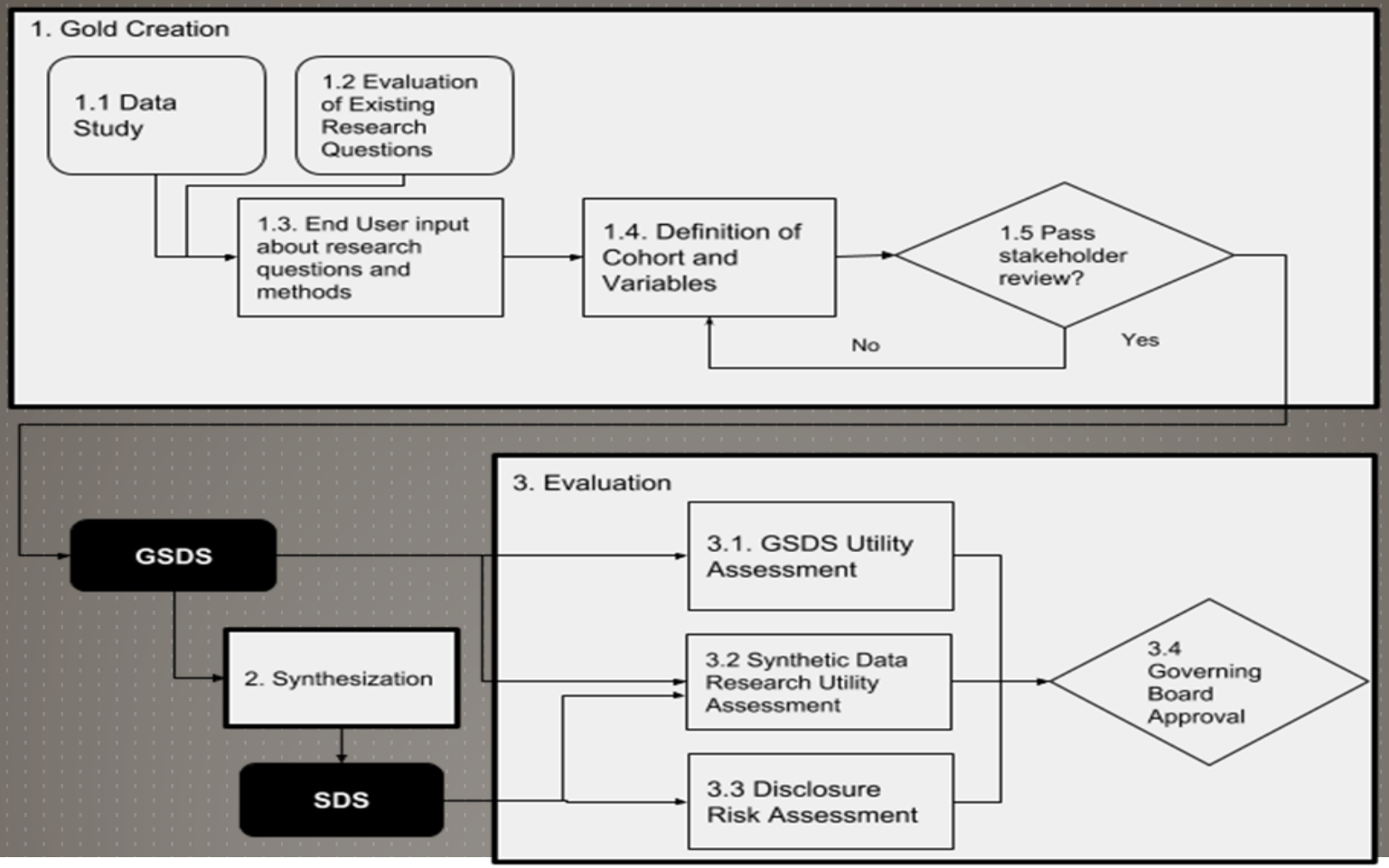
The Process

We split the project into three broad steps:

1) Creation of gold standard data sets (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

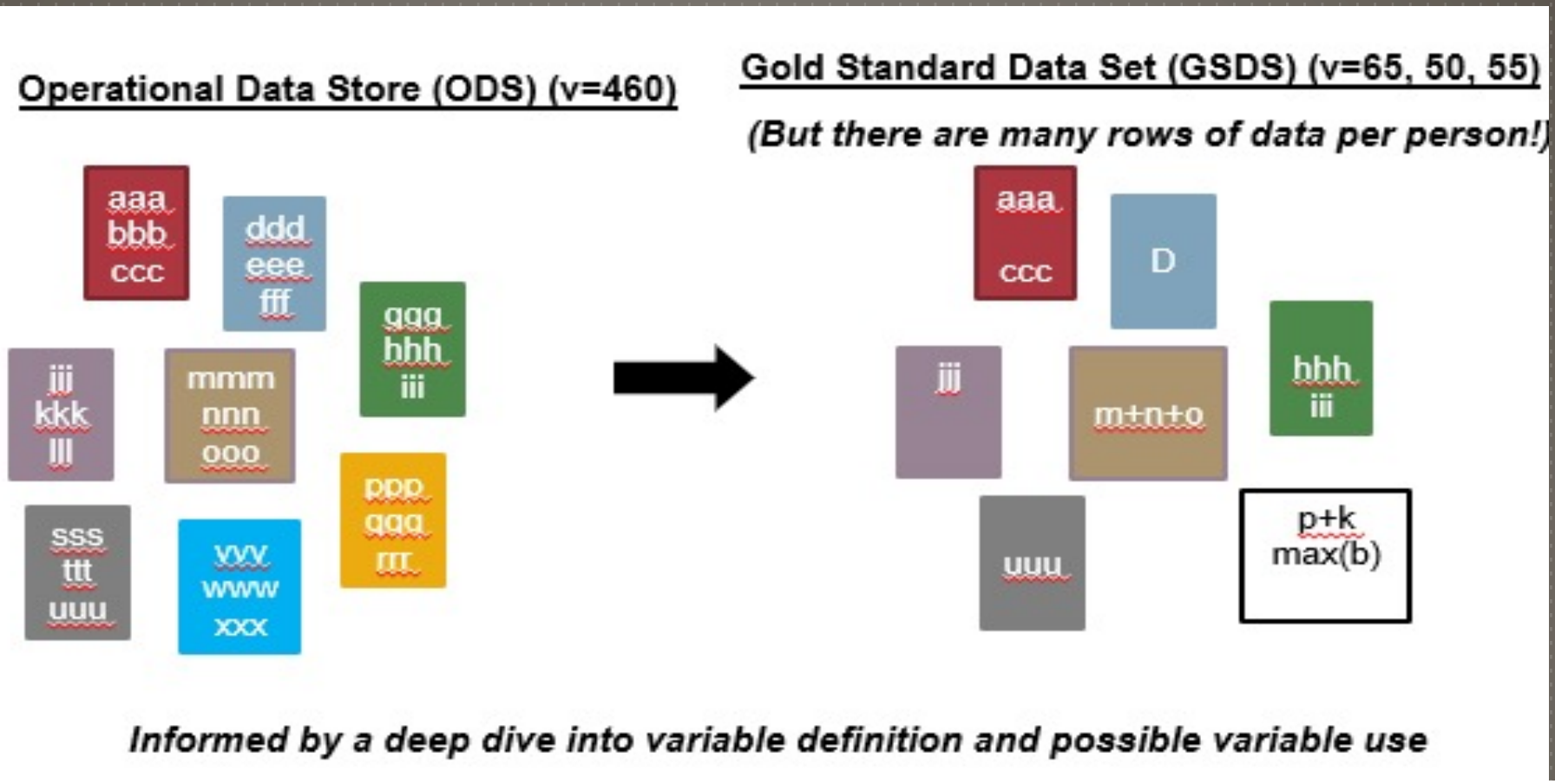
The Process, a Flowchart



Step 1: Designing the GSDS

- Identify variables in data to be included in GSDS, and therefore in synthetic data sets
- Input from MLDS Center Research Agenda, Center stakeholders, and potential users
- Balancing research utility, variable quality, and overall complexity and size of the data sets
- Three final GSDS:
  1. HS to PS - 9th Graders 2010-11 to 2015-16
  2. HS to WF - 9th Graders 2010-11 to 2015-16
  3. PS to WF - FTF 2010-11 to 2015-16

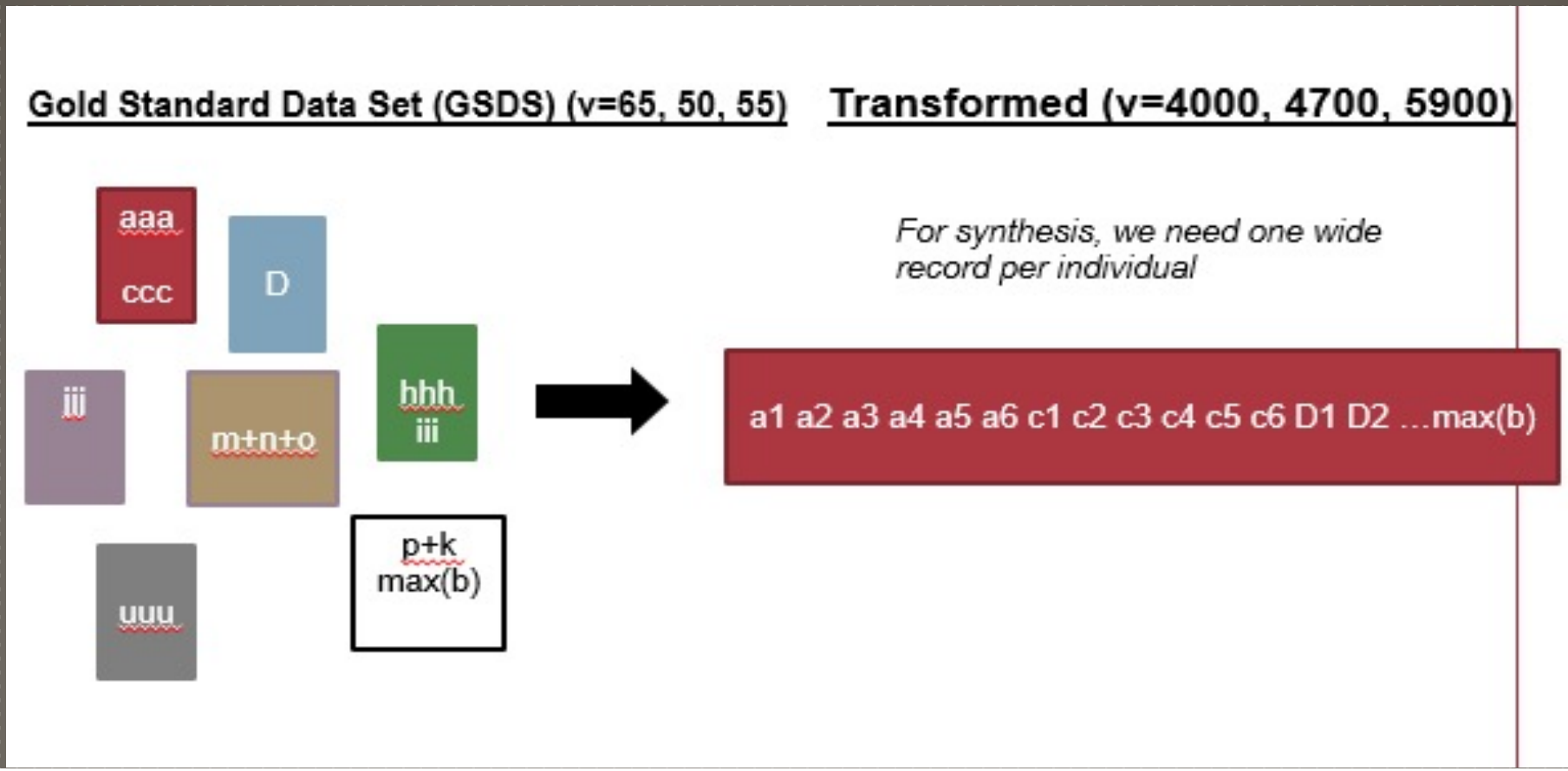
Step 1: GSDS Creation



Step 2: Synthesizing the GSDS

- We needed to satisfy a triangular trade-off:

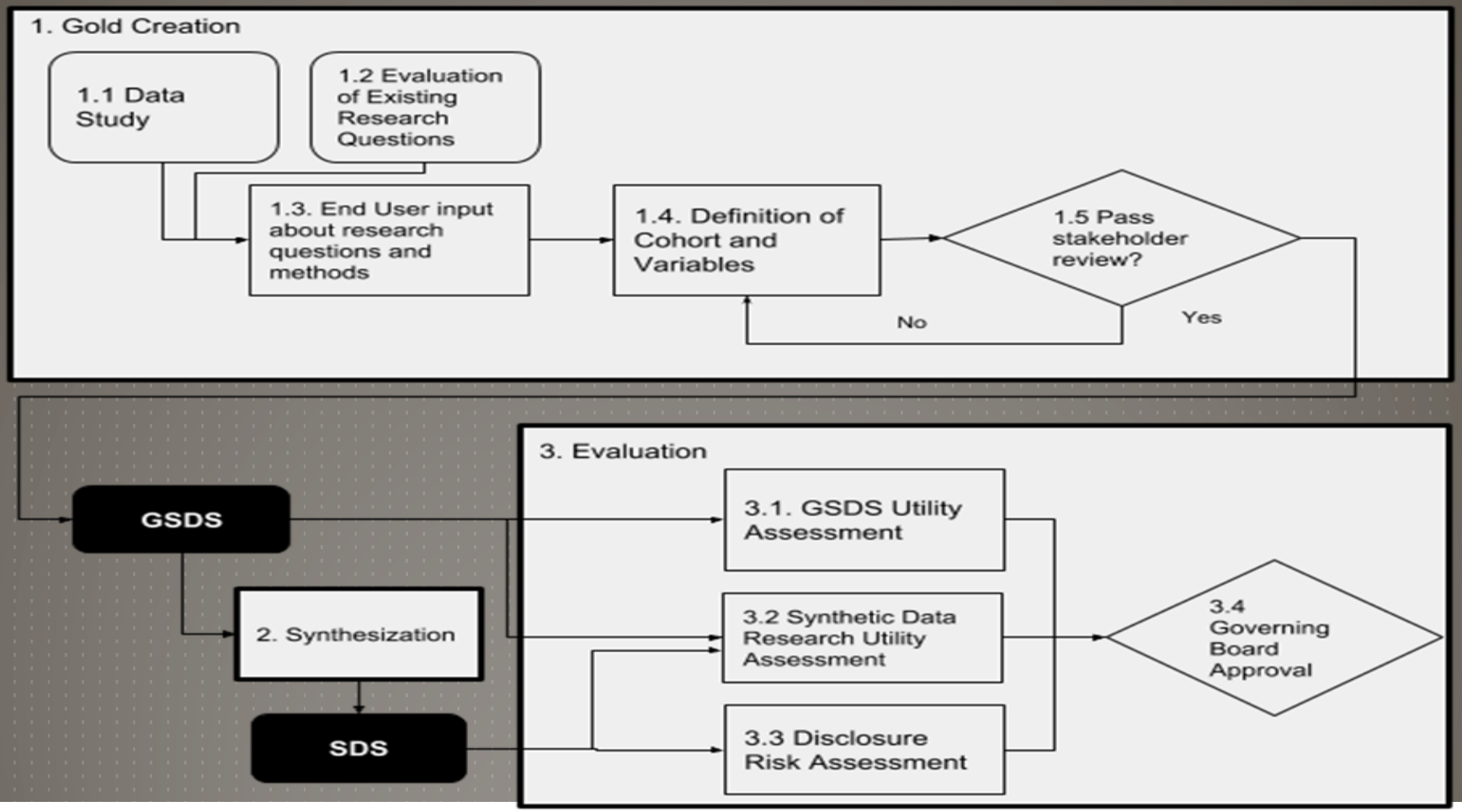
Step 2: Synthesizing the GSDS



Step 2: Synthesizing the GSDS

- Given the large number of variables (in wide format), both categorical and continuous variables, and the potential for interactions and non-linearities….
- We decided after initial testing and evaluation of the different existing methods, to implement the CART method (see Reiter, 2005)
- CART is the outcome of a general empirical method to model a dependent variable conditionally on a set of predictor variables.

The Process


1. Gold Creation
1.1 Data Study
1.2 Evaluation of Existing Research Questions
1.3. End User input about research questions and methods
1.4. Definition of Cohort and Variables
1.5 Pass stakeholder review?
No
Yes
GSDS
2. Synthesization
SDS
3. Evaluation
3.1. GSDS Utility Assessment
3.2 Synthetic Data Research Utility Assessment
3.3 Disclosure Risk Assessment
3.4 Governing Board Approval

Step 3: Evaluation of the SDS

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

  Are the GSDS data useful themselves?

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

  Do you get the "right" answer from the SDS?

  General Utility and Specific Utility

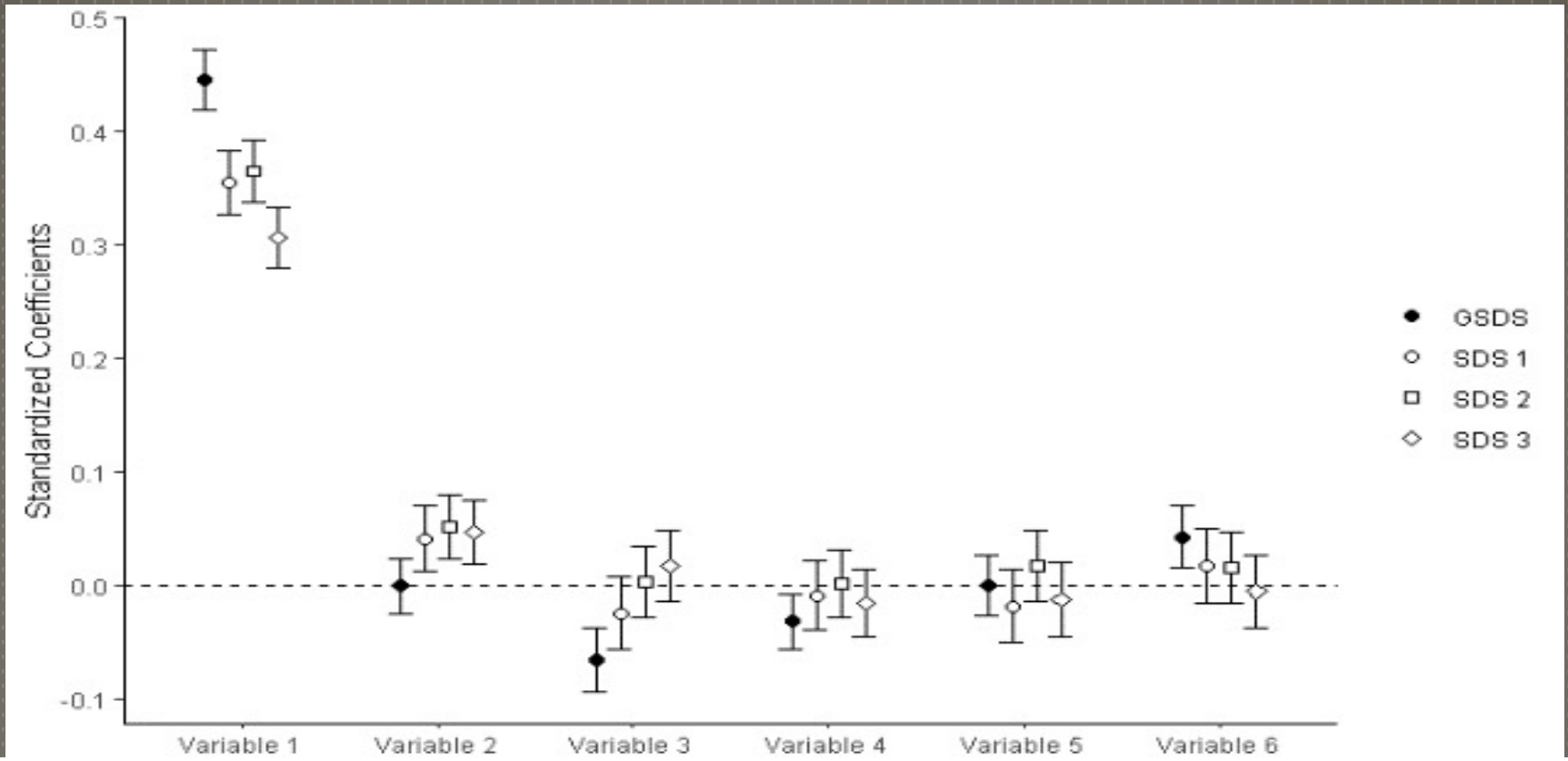
- **Disclosure risk assessment (Step 3.3)**

  Do the SDS pose a risk of disclosure?

  Identification and Attribute Disclosure Risk

Step 3: Specific Utility Evaluation Example

Comparison of Standardized Multiple Regression Coefficient Estimates



Step 3: Disclosure Risk

- *Identification Disclosure Risk* is the potential to identify an individual included in the GSDS in the SDS.
- However, we created fully synthesized data sets and therefore no actual "cases" or real individuals exist in the SDS, so there is essentially zero identification disclosure risk.
- *Attribute Disclosure Risk* is the potential to determine sensitive values for small/rare groups in the GSDS and relies on utilizing outside information (such as additional data) to create inferences as a means to identify at-risk groups (<10; 505 and 2,742).
- We examined ADR using a "worst case scenario" assumption.
- We determined that an intruder could not learn sensitive information about such small/rare groups to any more precision than could be gleaned from knowing basic education and career information about such individuals.

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

Informing a Decision About Synthetic Data for Prevention Science

- **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, policy, and practice in prevention science

Acknowledgements

- This presentation was prepared by the Research Branch of the Maryland Longitudinal Data System (MLDS) Center as part of funding from the U.S. Department of Education (R372A150045)
- We are grateful for the research support provided by the MLDS Center. The views in this presentation do not necessarily represent the views of the MLDS Center, its partner agencies, or the federal government.
- Contributors: Daniel Bonnery, Yi Feng, Tessa Johnson, Mark Lachowicz, Terry Shaw

Contact

ahenneberger@ssw.umaryland.edu

mwoolley@ssw.umaryland.edu