MLDS Center

Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

The MLDS Synthetic Data Project: Evaluation of Research Utility & Disclosure Risk

Mark Lachowicz, UMCP
Terry V. Shaw, UMB
(and a host of many others!)

MLDS Center
Maryland Longitudinal Data System

Background

- State education and longitudinal data systems are advancing and growing in number, and the use of these data systems for education and workforce research holds great promise (Figlio, Karbownik, & Salvanes, 2017).

- Since 2005, the USDOE has supported 47 states, as well as the District of Columbia, Puerto Rico, the Virgin Islands, and American Samoa in their development of statewide education data systems (SLDS Grant Program, 2018b), representing an overall investment of $721 million in federal funding as of May 2018 (SLDS Grant Program, 2018a).

MLDS Center
Maryland Longitudinal Data System

Benefits of State Longitudinal Data Systems

- provide a number of advantages to researchers as compared to traditional survey measures, including
  - larger data sets,
  - fewer problems with attrition,
  - low rates of non-response bias, and
  - more data for rare populations
  - relatively cost-effective approach to answering policy questions as no need for costly and time-consuming primary data collection

MLDS Center
Maryland Longitudinal Data System

Overview of 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.
  - Testing the feasibility of a synthetic educational data system was one aspect of that grant.

- The concept behind the Synthetic Project:
  - data is generated based on models of the data to mimic the relational patterns among variables
  - statistical analyses with synthetic ("fake") data should yield population findings similar to the real data
  - Simultaneously, reduces the risk of privacy breach

Definitions

- **Gold Standard Data Set –** (GSDS) are simplified versions of the data housed in the MLDSC.
- **Synthetic Data Set –** (SDS) is created from a computational model such that when statistically analyzed will act like the original data.
- **Fully Synthetic Data Sets –** are comprised completely of synthesized variables, all variables and all values are synthesized.
- **Partially Synthetic Data Sets –** are a combination of synthesized and non-synthesized variables (which have the original "real" values).

MLDS Center
Maryland Longitudinal Data System

The Process…

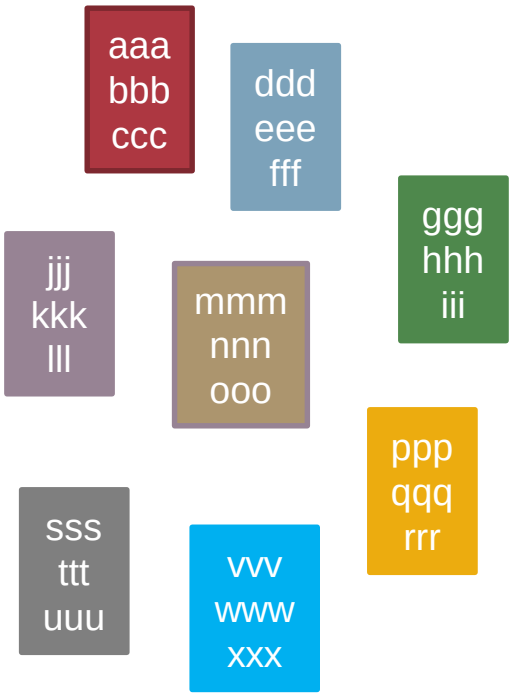
After much discussion and consultation with experts in the field it was determined that the synthetic data project needed to progress in three broad steps:

1) creation of gold standard datasets (GSDS) that integrated the complex structure of the MLDS data into a "simpler" warehouse structure.
2) synthesization of the GSDS using complex statistical processes that convert the GSDS into a Synthetic Data sets (SDS)
3) evaluation of the utility and safety of the synthetic data sets (SDS)

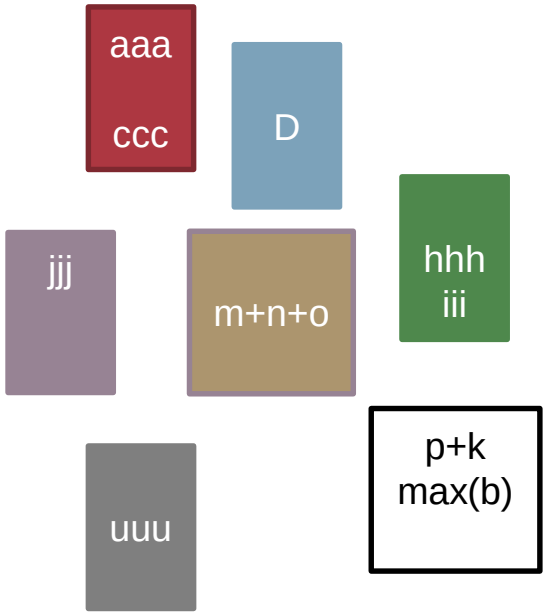
MLDS Center
Maryland Longitudinal Data System

The Process: Step 1 creating GSD

Operational Data Store (ODS)



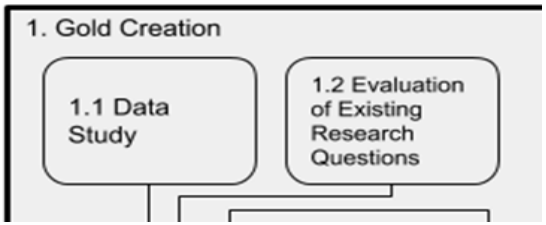
Gold Standard Data Set (GSDS)



Data Study (Step 1.1)

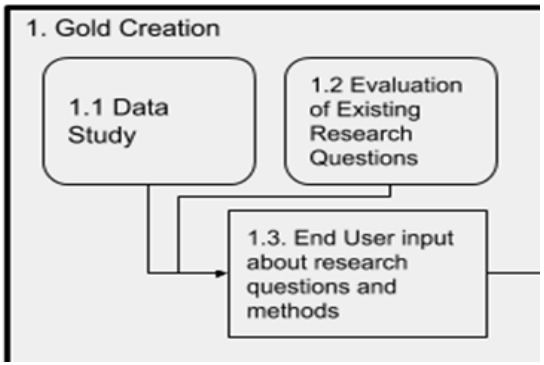
For each variable, we:

- studied the data coding by checking the consistency between the data dictionary and the values stored in the system
- examined the descriptive statistics—especially regarding outliers, missing data patterns, and in some cases the pattern of "not applicable" for some variables
- investigated the presence of redundant or overlapping information as we have multiple data sources



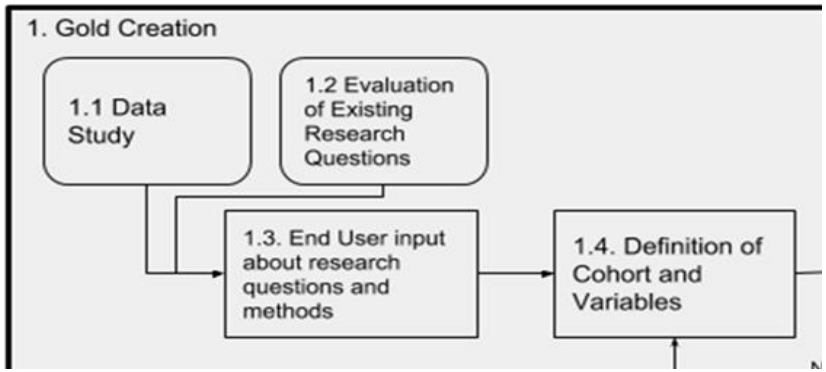
Evaluation of Existing Research Questions (Step 1.2)

- While investigating data elements in the larger data system, we evaluated the research analyses that have used the data housed in the MLDS data system along with the current research agenda of the Center. https://mldscenter.maryland.gov/ResearchAgenda.html
- To be of the greatest use, the GSDS should contain the data needed for these reports/studies



End User Input About Research Questions And Methods (Step 1.3)

- Convened a group of institutional researchers
- We asked about their research interests, the analytic methods they would use if given access to synthesized datasets of a similar type, and the desired format.
- They encouraged us to focus on a single cohort.



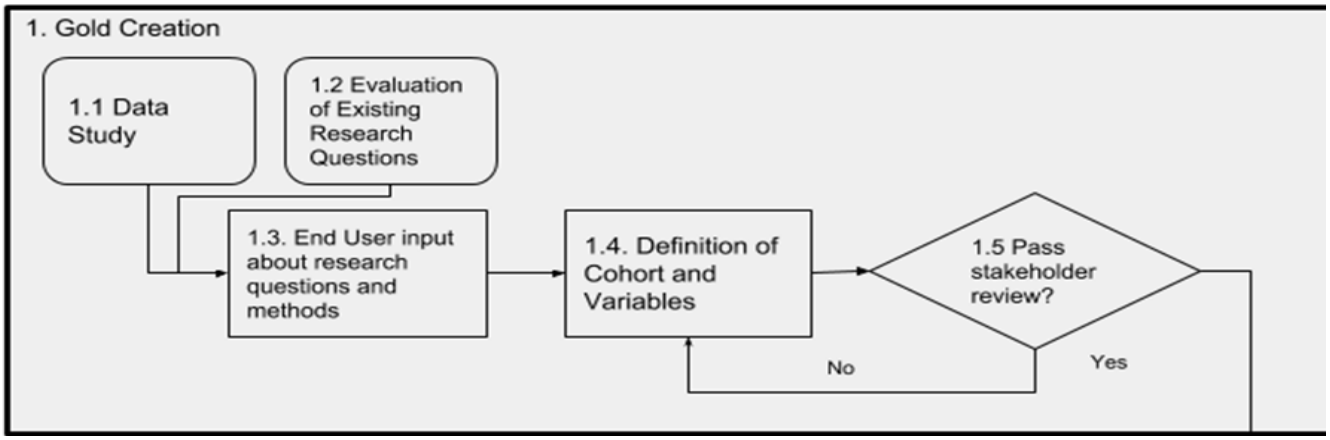
Definition of Cohort and Variables (Step 1.4)

Cohort definition

- *High-school to Post-secondary* *9th graders in 2010-11 until 2015-16*
- *High-school to Workforce* *9th graders in 2010-11 until 2015-16*
- *Post-Secondary to Workforce* *first-time freshmen in 2010-11 until 2015-16*

Variable selection

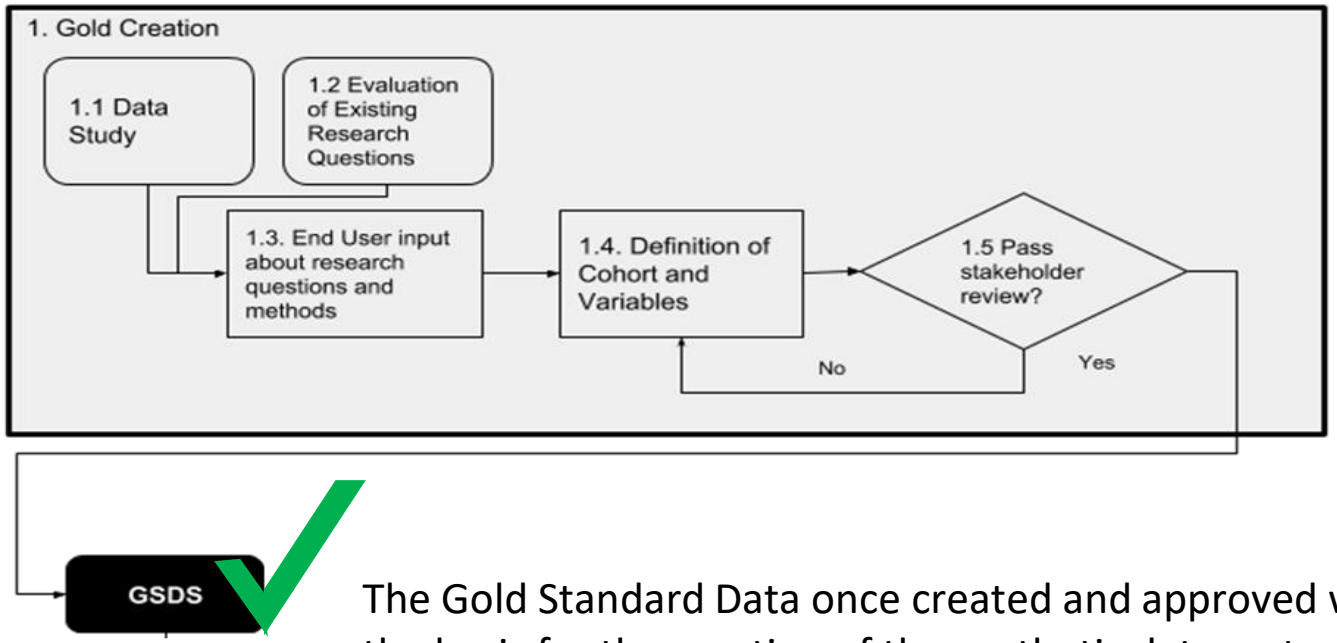
- GSDS variables selected were completed under two anticipated constraints: 1) practicality and 2) legal constraints



Decision point: Stakeholder Review (Step 1.5)

- Cohort definitions, list of variables, and simplified data structure presented to the major stakeholders within the MLDS Center.
- The creation of the GSDS was an iterative process
  - stakeholder feedback informed structure and variable selection.
  - Stakeholder review was instrumental to creating a quality functional GSDS.

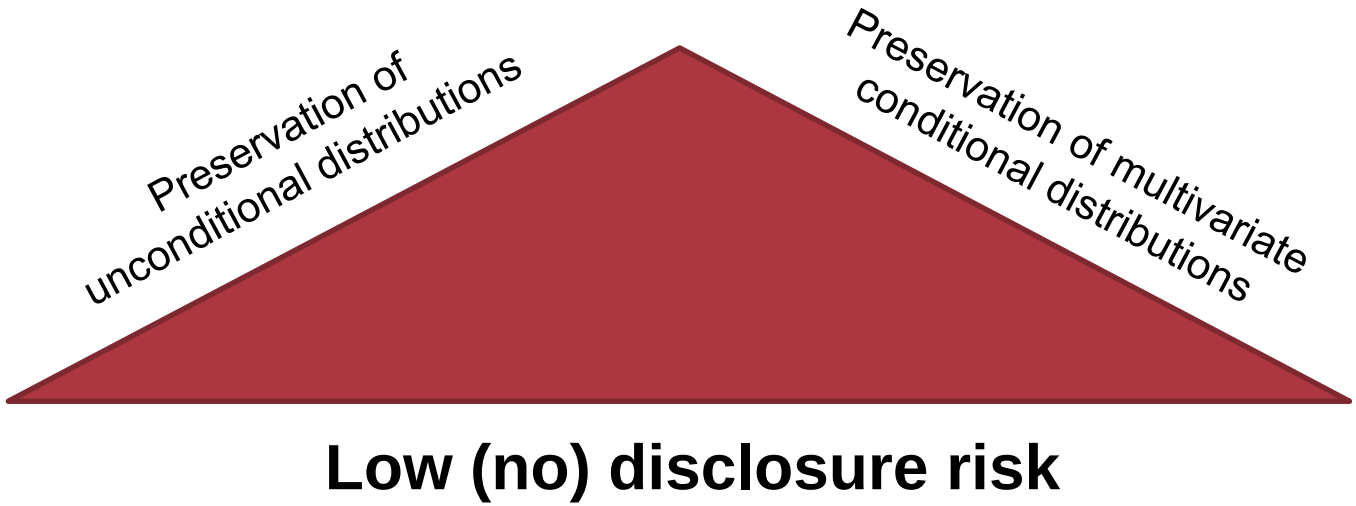
The Process, continued…



The Gold Standard Data once created and approved was the basis for the creation of the synthetic data system.

Synthesization (Step 2)

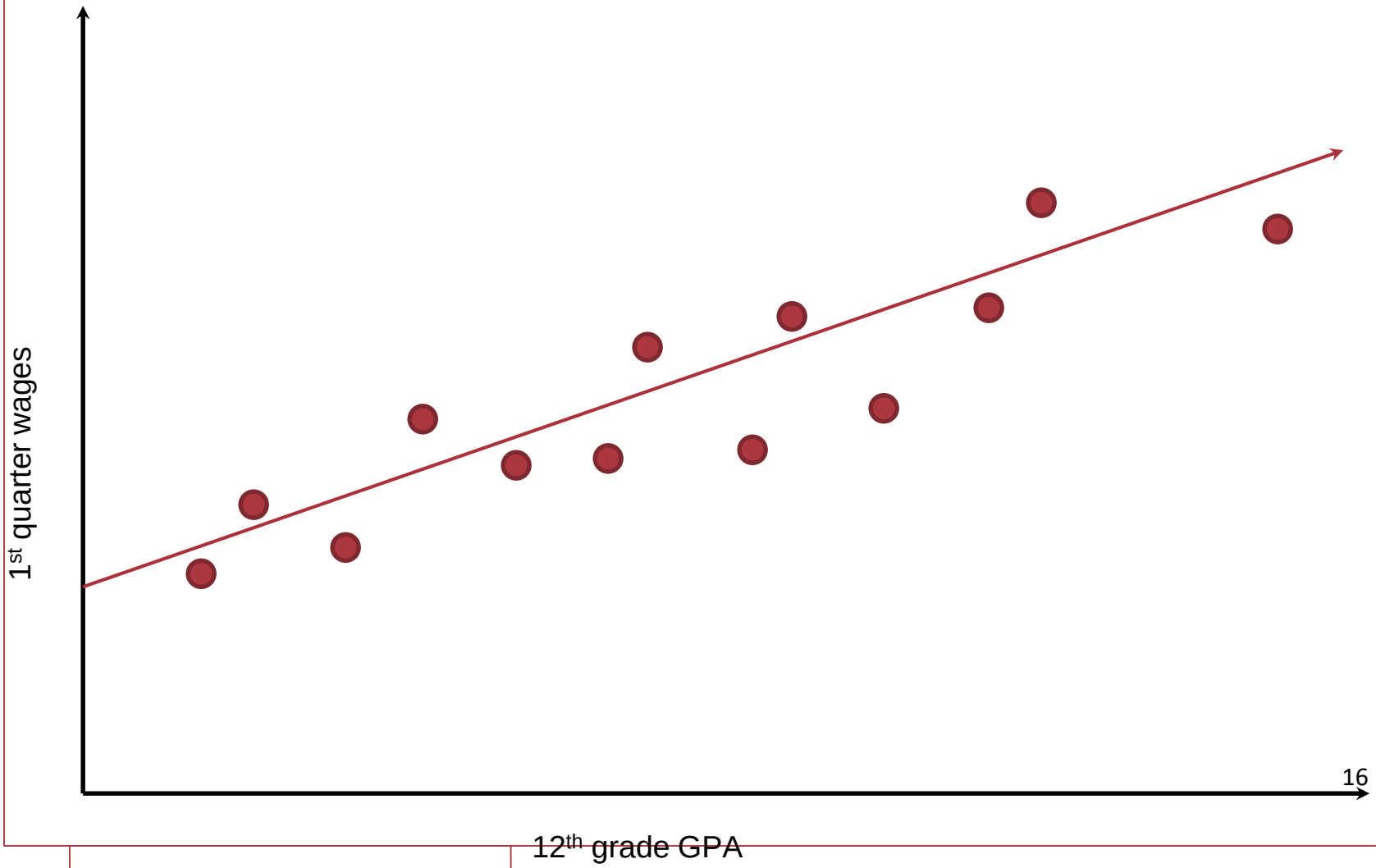
- We need to satisfy a triangular trade-off:



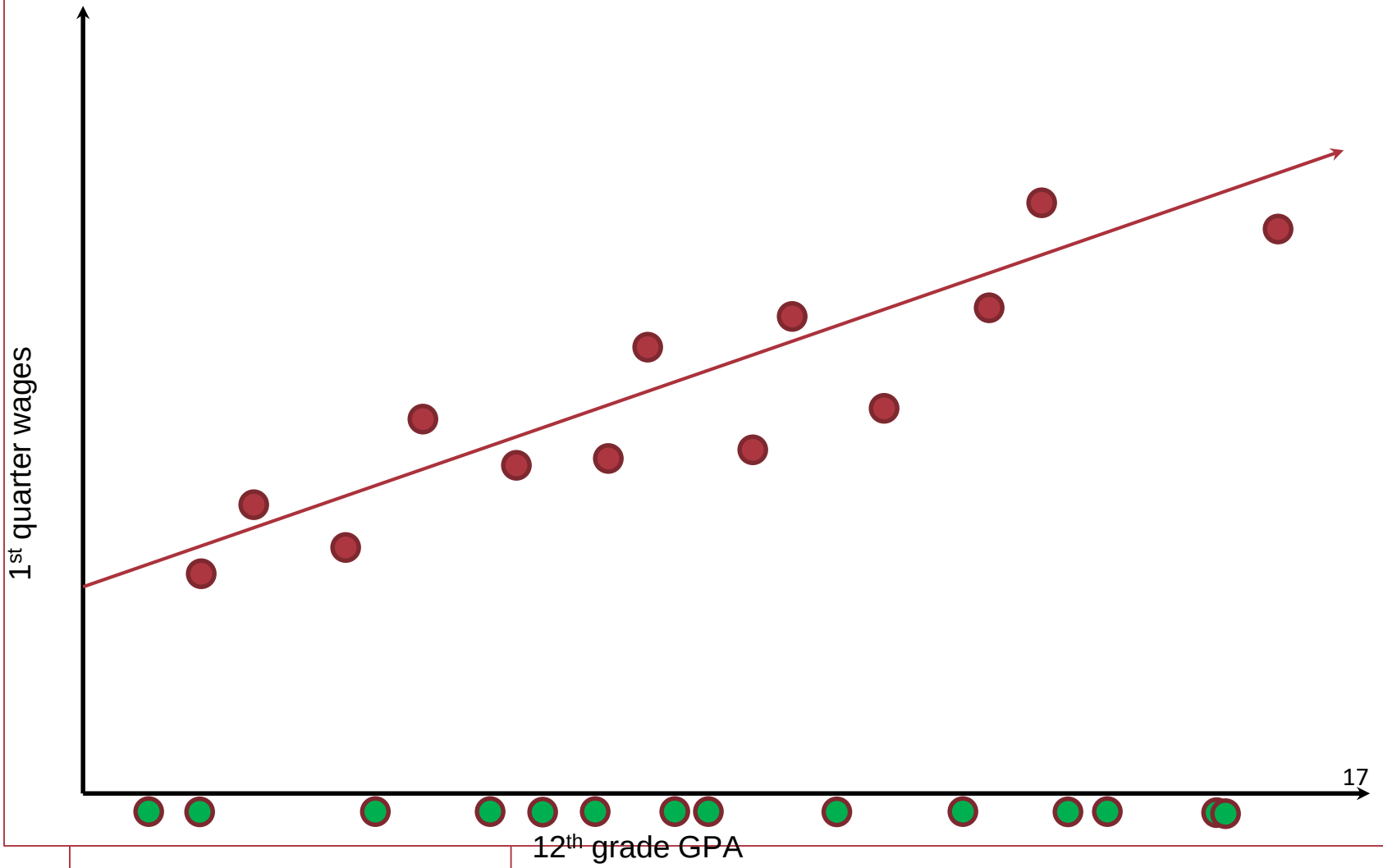
Creation of Synthetic Data

- There are various methods that can be used to generate synthetic data, all of which require some kind of strategy for modeling relations among variables in the raw data
- Synthetic data generation is traditionally accomplished with sequential regression models. Variables are arranged, and therefore synthesized, in a certain order
- For each variable, a regression model is developed against a selection of predictors among the preceding variables. The models are developed in a sequential manner until a model is developed for each variable in the data. Synthetic data are thus generated sequentially from the posterior predictive distribution for each variable

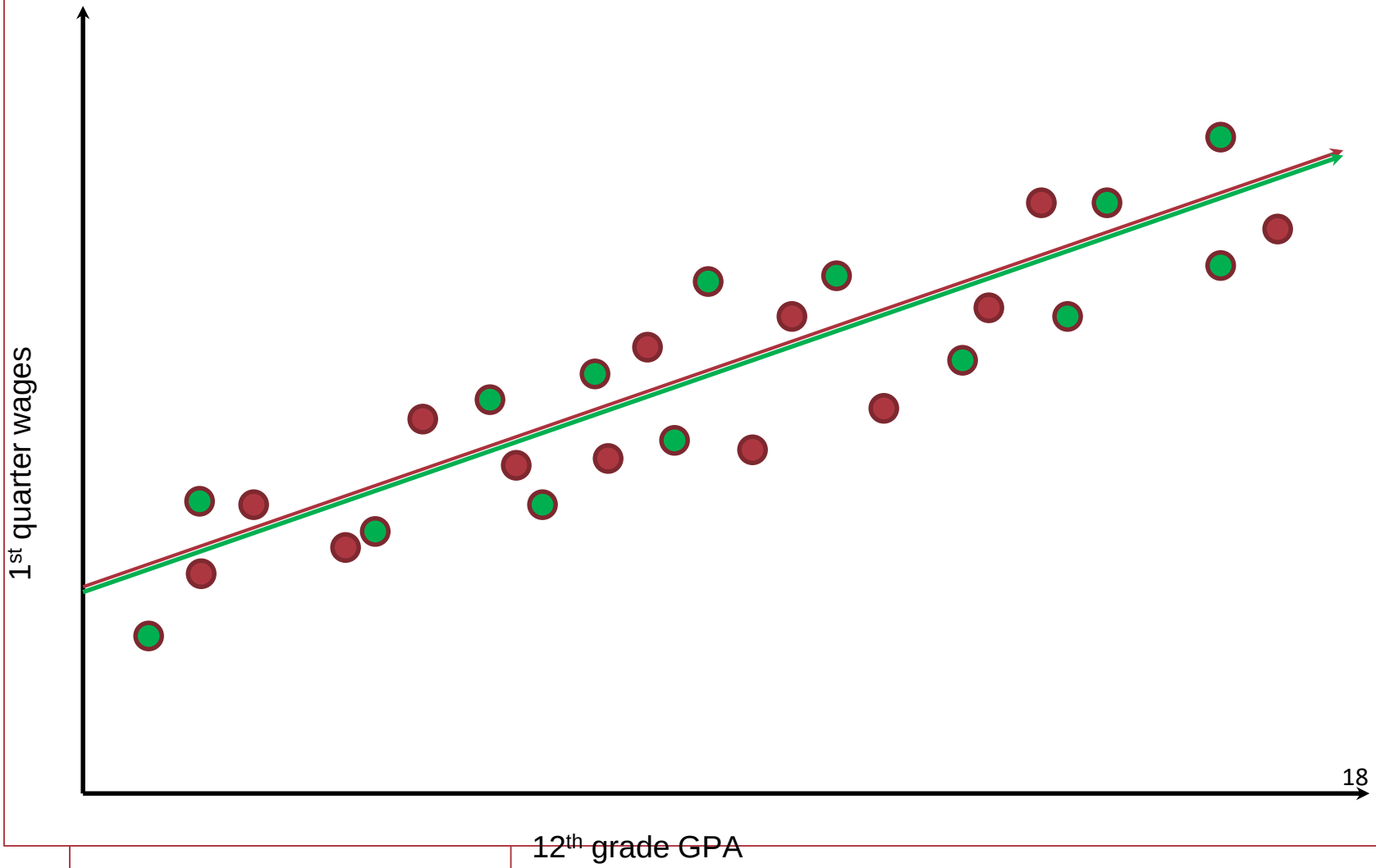
MLDS Center
Maryland Longitudinal Data System



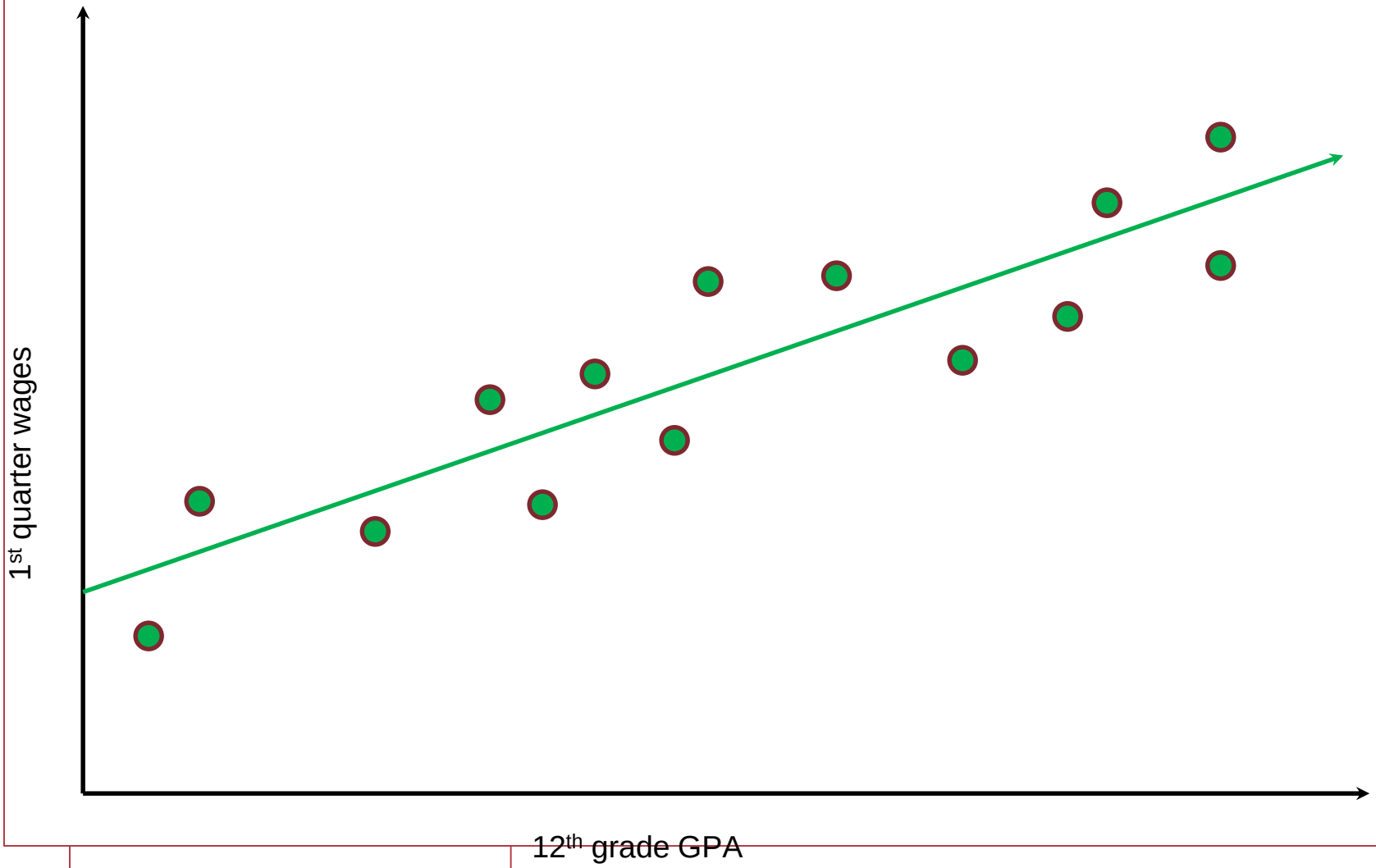
MLDS Center
Maryland Longitudinal Data System



MLDS Center
Maryland Longitudinal Data System



MLDS Center
Maryland Longitudinal Data System

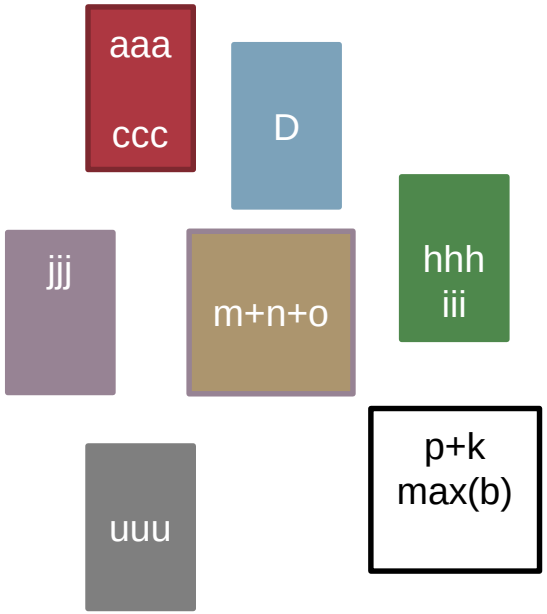


MLDS Center
Maryland Longitudinal Data System

Synthesization (Step 2), continued…

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)

Synthesization (Step 2), continued…

- 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.

Synthesization (Step 2), continued…

- Each binary partition consists of finding the best split, e.g. identifying the predictor variable and threshold that will split the dataset in two sub-datasets (nodes) for which the within-node dispersion of the dependent variable is minimal.
- The process is repeated in the resulting two sub-datasets until no potential split results in a significant between-node dispersion (or we reach an alternate stopping rule, such as N=30).

MLDS Center
Maryland Longitudinal Data System

Example of A Classification Tree for Term Grade Point Average

- Suppose that we have already synthesized several variables for 60,000 "fake" records, including:
  - 2015 2nd Term credit hours earned
  - SAT-Math, SAT-Writing
  - Gender
- We are now looking to synthesize the variable "2015 2nd Term GPA"

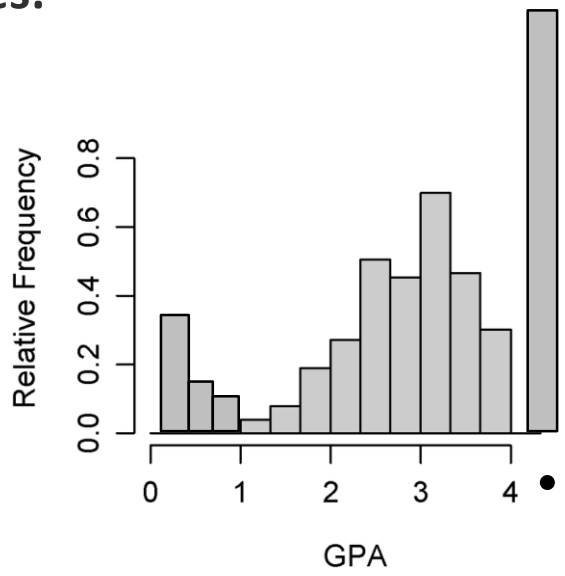
Credits	SAT-M	SAT-W	Gender	GPA
12	490	510	M	?
8	380	450	F	?
14	750	690	F	?

Example of a classification tree for term grade point average

- We will use the REAL data from GSDS to build a set of possible values for each these "fake" student

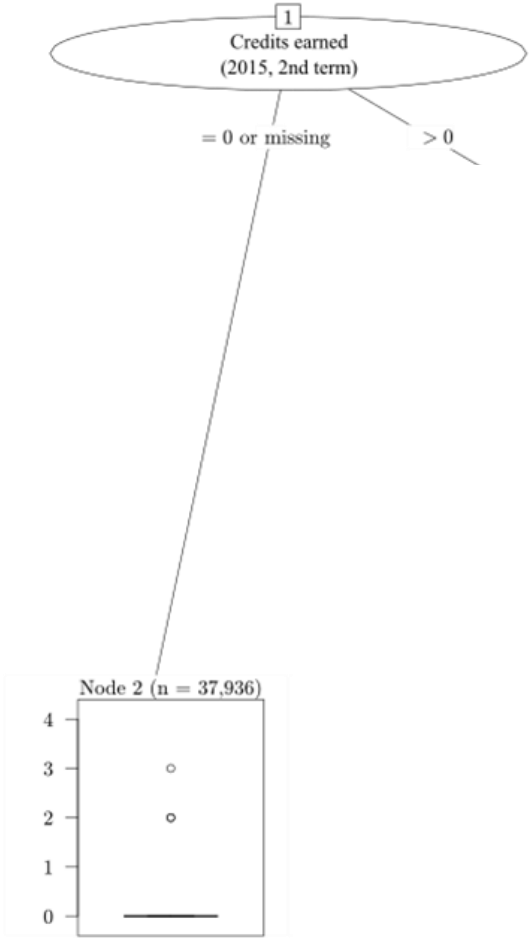
Suppose the GSDS data contained these values:

Credits	SAT-M	SAT-W	Gender	GPA
11	490	510	M	2.1
15	380	450	F	3.2
9	750	690	F	3.6
0	380	410	F	.
12	710	750	M	3.8
16	450	590	M	2.9



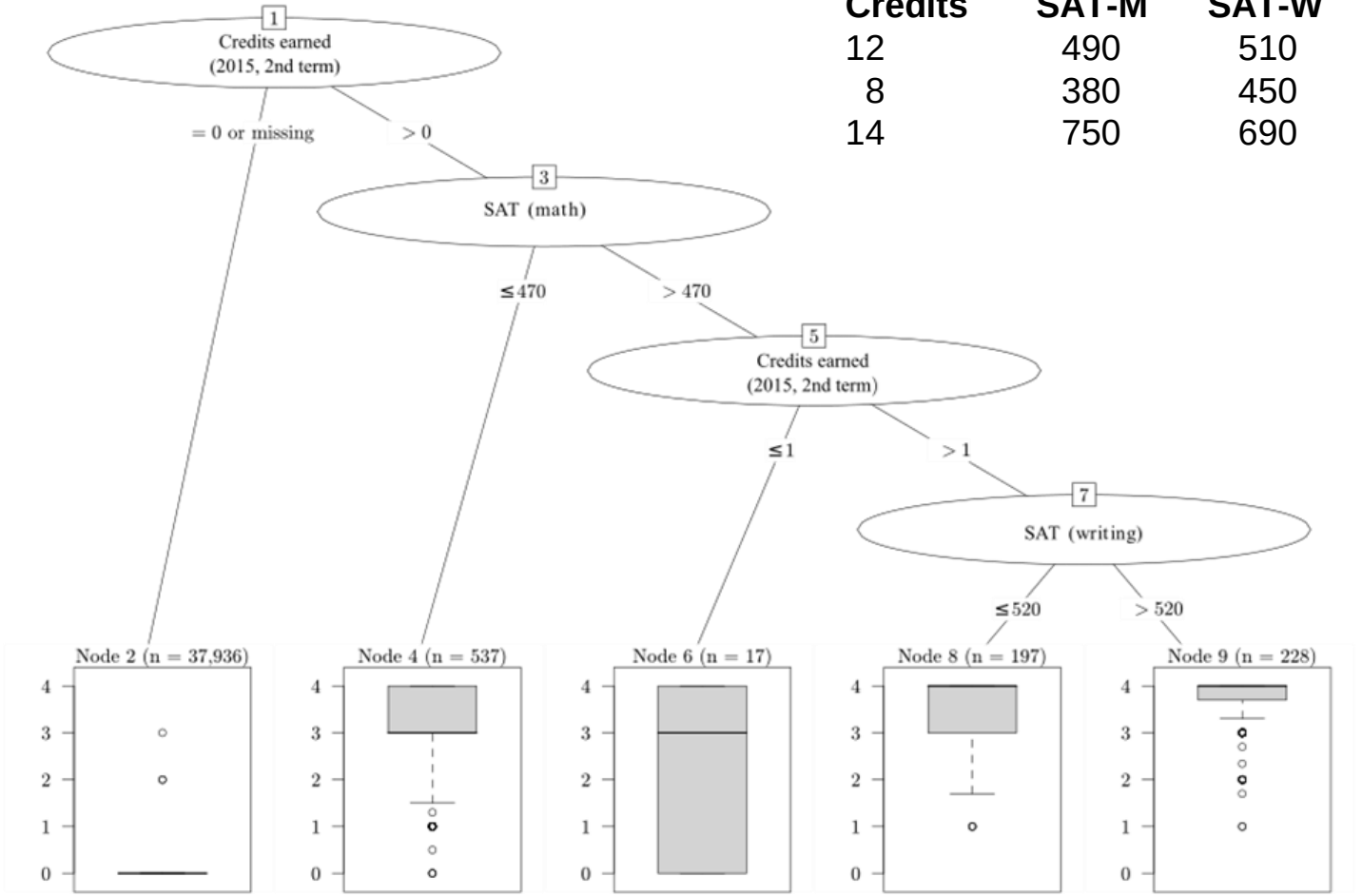
We will divide up the full distribution into homogeneous sets

Example of a classification tree for term grade point average



Example of a classification tree for term grade point average

Credits	SAT-M	SAT-W	GPA
12	490	510	3.6
8	380	450	3.1
14	750	690	4.0



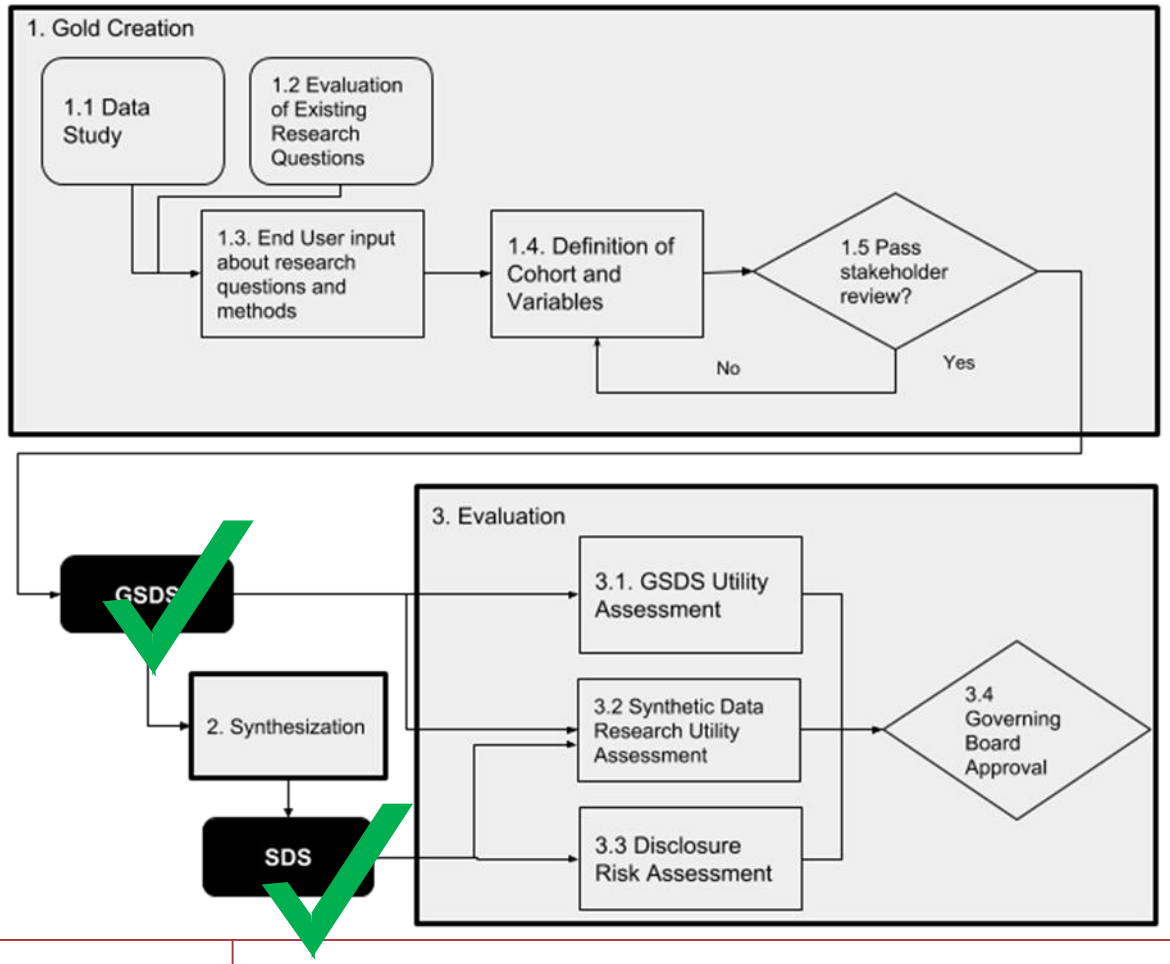
MLDS Center
Maryland Longitudinal Data System

Synthesization (Step 2), continued…

- We have fully synthesized the data for our three GSDS thirty-three times each

- In the next step we will evaluate these data but our findings will lead us to iteratively tweak our synthesis process, by including different predictor variable sets

The process, continued…



Evaluation (Step 3)

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

  Are the GSDS data useful themselves?

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

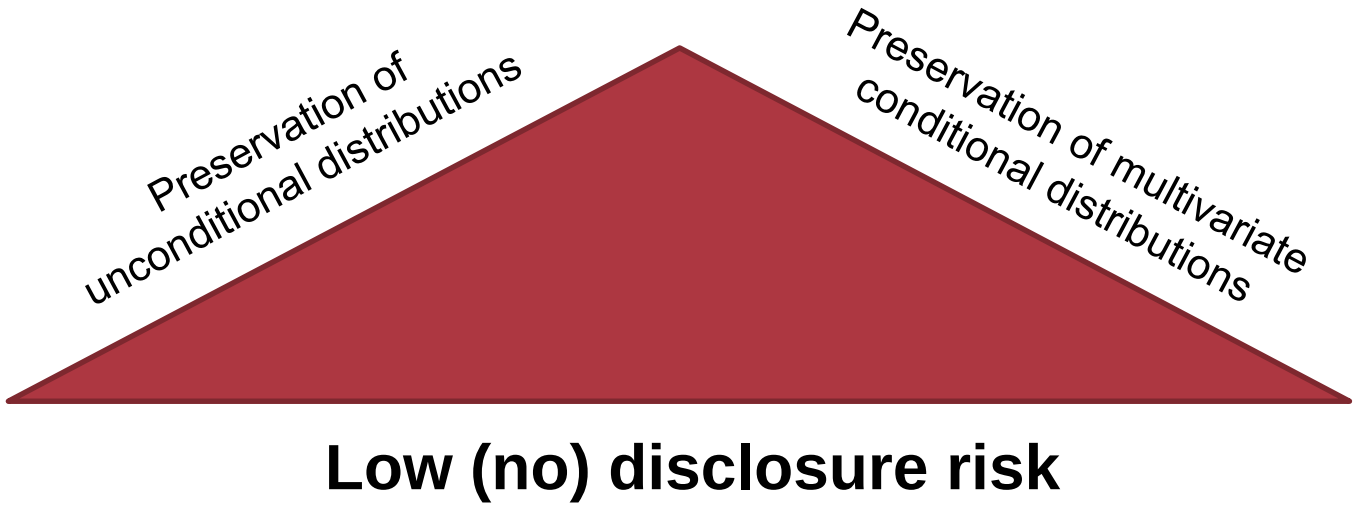
  Do you get the "right" answer from the synthetic data?

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

  Do the synthetic data pose a risk of disclosure?

Evaluation (Step 3)

- We need to satisfy a triangular trade-off:



Research Utility Assessment

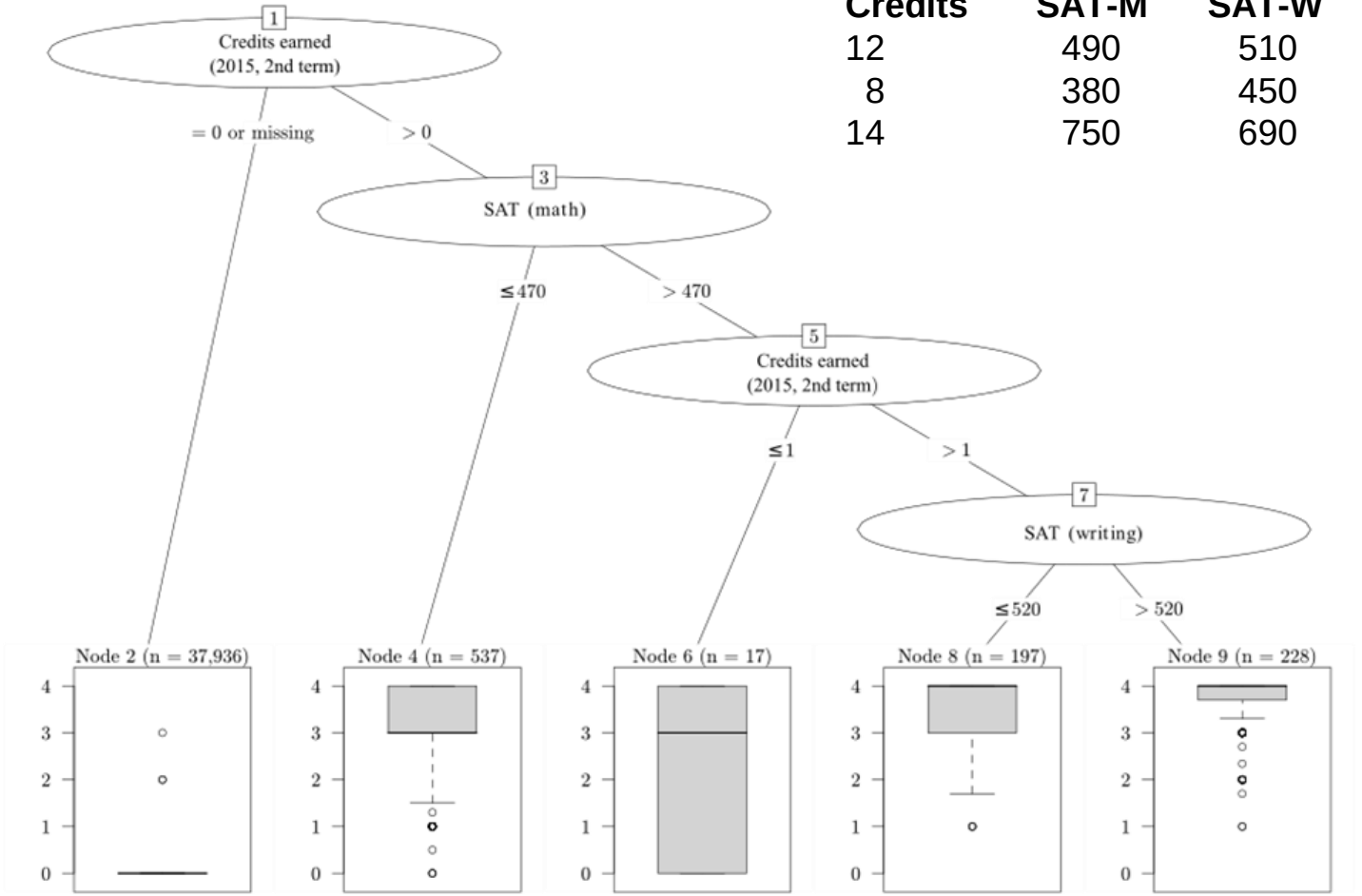
- How well does the synthetic data reflect the characteristics of the gold standard data?
- Utility of the synthetic data depends primarily on the quality of the generating model
  - Relationships not modeled will not be present in the synthetic data
- Four dimensions
  - Model evaluation
  - Exploratory comparisons
  - Model specific measures
  - Global utility

RU - Model evaluation

- CART model used to generate synthetic datasets
- Nonparametric model → no distributional assumptions
- Assessment
  - For a selection of important variables (e.g., total wages), confirm trees worked as intended
  - Multiple splits on different variables evidence of multivariate relationships modeled
  - Leafs with zero variance
  - Leafs with sample size < 30

Example of a classification tree for term grade point average

Credits	SAT-M	SAT-W	GPA
12	490	510	3.6
8	380	450	3.1
14	750	690	4.0



RU - Exploratory Comparisons

- First assessment step for the synthetic datasets
- At minimum, synthetic variable distributions should be comparable to their corresponding gold standard variables
  - Marginal distributions of variables with few missing values should be very close
  - Distributions expected to diverge for sparse variables and within smaller N subgroups
- Compare quantiles, means, histograms, etc.
- Check for reasonability
  - No negative incomes, etc.

RU - Exploratory Comparisons

- In total, ~ 100 unique variables in the GSDS
  - Measures for many aspects of education in high school and post-secondary programs
  - Repeated measures for individuals on many variables over time (e.g., GPA, wages)
- Multiple synthetic datasets
  - Same model used to generate replicates
  - Variability across synthetic datasets

RU - Exploratory example

- Categorical - marginal distribution of Race

	GSDS N (%)	AVG SDS N (%)
Asian	3082 (5.9)	3089.7 (6.0)
Black	19014 (36.6)	19005.46 (36.7)
2 or More	2169 (4.2)	2195.03 (4.2)
White	26253 (50.6)	26254.42 (50.6)
Missing	744 (1.4)	734.61 (1.4)

RU - Exploratory example

- Continuous - marginal distribution of GPA

	GSDS	AVG SDS M (SD)		GSDS	AVG SDS M (SD)
N	26864	27898.79 (184.6)	Min	0	0.16 (0.465)
Missing	2267	2217.24 (44.52)	Q10	2.56	2.56 (0.003)
Mean	3.204	3.193 (0.003)	Q25	2.87	2.862 (0.005)
SD	0.466	0.465 (0.002)	Median	3.22	3.206 (0.006)
Skew	-0.269	-0.245 (0.019)	Q75	3.57	3.555 (0.005)
Kurtosis	2.509	2.502 (0.11)	Q90	3.83	3.818 (0.005)
			Max	4.26	4.254 (0.012)

RU - Specific Utility Assessment

- How well does synthetic data reproduce the results of specific analyses?
- Gold standard analyses
  - Group means and mean differences
  - Strength of relationship - bivariate
  - Strength of relationship - multivariate
  - Longitudinal associations
- Informed by variable definitions and content knowledge

RU - Analyzing synthetic data

- Multiple synthetic datasets → pooled results
- Analyze each synthetic dataset as if gold standard
- Combine results
- Comparable to multiple imputation

RU - Analyzing synthetic data

- Parameter ***Q*** we want to estimate
- ***q*** is the sample estimate of ***Q***
- ***u*** is the variance of estimate ***q***
- $q^{(i)}$ and $u^{(i)}$ are estimates from a synthetic dataset ***i***

RU - Analyzing synthetic data

- Compute $\bar{q}$ across datasets

$$\bar{q} = \sum q^{(i)} / m$$

- Compute variance between datasets (***b***)

$$b = \sum (q^{(i)} - \bar{q})^2 / (m - 1)$$

- Compute average variance across datasets

$$\bar{u} = \sum u^{(i)} / m$$

- Compute pooled variance estimate ***T***

$$T = (1 + m^{-1})b - \bar{u}$$

RU - Analyzing synthetic data

- Inferences about ***Q*** can be made using $\overline{q}$ and ***T***

- For large samples and large ***m***, 95% CI for $\overline{q}$

$$\overline{q} \pm 1.96\sqrt{T}$$

- For small samples and/or small ***m***, use *t*-distribution with degrees of freedom (***df***)

$$df = (m-1)(1-\overline{u}/((1+m^{-1})b))^2$$

RU - Comparing results

- Comparing point estimates not sufficient, need to incorporate variability
- Standardized differences

$$SD = \frac{\beta_{SDS} - \beta_{GSDS}}{SE_{GSDS}}$$

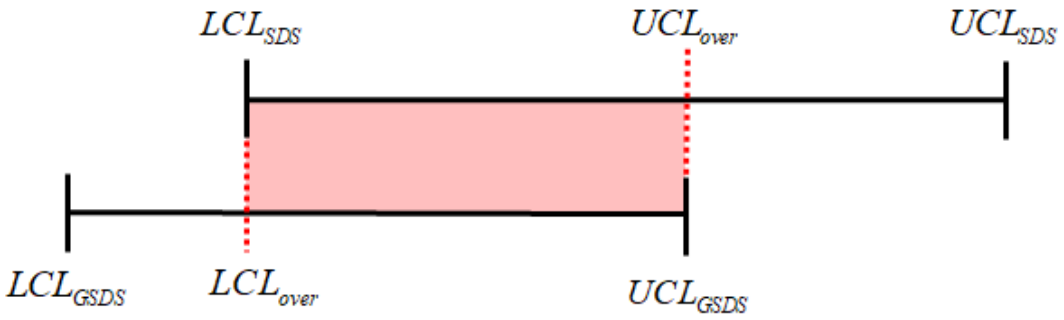
RU - Comparing results

- Point estimates may appear substantially different but inferences may be similar due to uncertainty
- Confidence interval overlap

$$IO = \frac{1}{2}\left\{ \frac{UCL_{over} - LCL_{over}}{UCL_{GSDS} - LCL_{GSDS}} + \frac{UCL_{over} - LCL_{over}}{UCL_{SDS} - LCL_{SDS}} \right\}$$

CI Synthetic

CI Gold



RU - Specific Utility Example

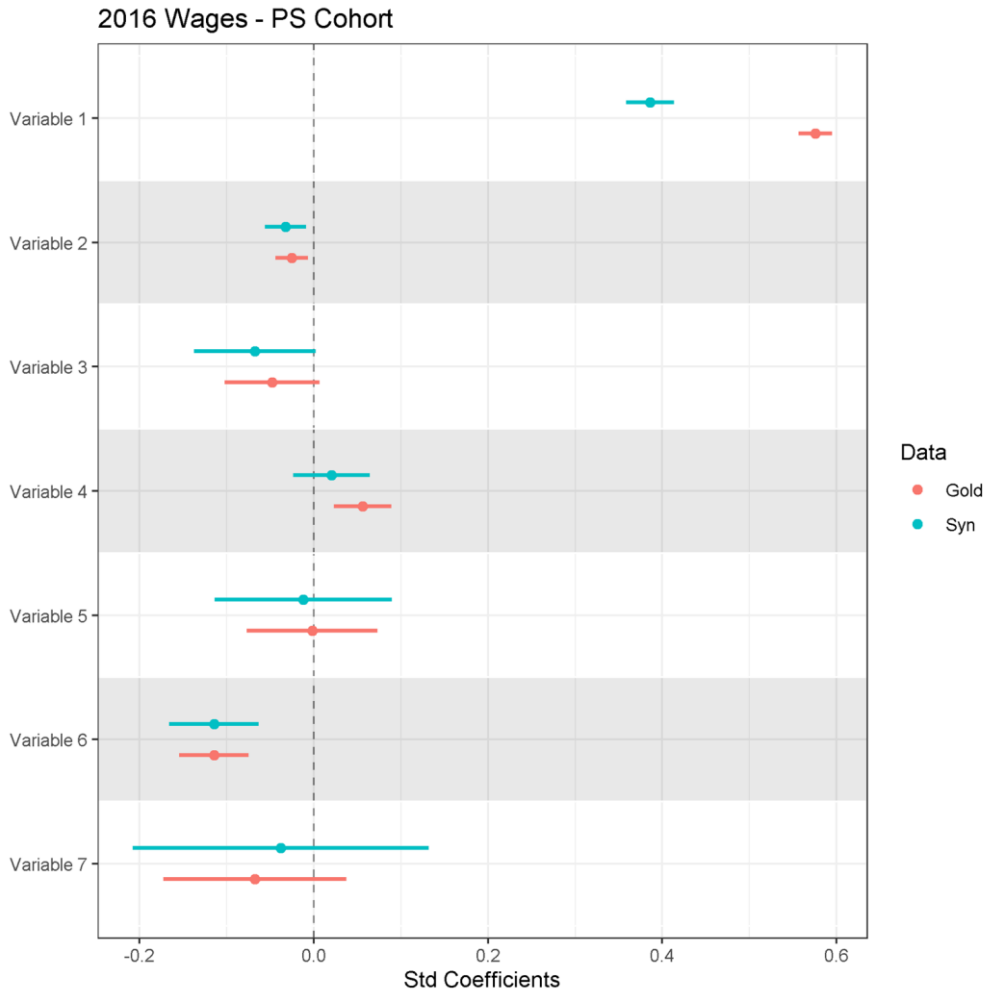
- To illustrate components of utility assessment, we use a subset of the PS → WF cohort
- Regressed (log transformed) 2016 wages on gender, cumulative standardized test (SAT or ACT), transformed 2015 wages, race/ethnicity categories
- The sample size of this cohort was 51,863 students
- Results first from early synthesis model

$$lWage\_2016_i = b_0 + b_1 lWage\_2015_i + b_2 CSA_i + b_3 Sex_i + b_4 Ethn_i + \sum b_j Race_i + e_i$$

RU - Specific Utility Example

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	CI Overlap
Variable 1	0.576 (0.01)	0.386 (0.032)	3.334	-1.821
Variable 2	-0.025 (0.01)	-0.032 (0.012)	0.128	0.84
Variable 3	-0.048 (0.028)	-0.067 (0.035)	0.122	0.85
Variable 4	0.056 (0.017)	0.02 (0.022)	0.372	0.532
Variable 5	-0.002 (0.038)	-0.012 (0.05)	0.047	0.881
Variable 6	-0.114 (0.02)	-0.114 (0.026)	< 0.001	0.898
Variable 7	-0.068 (0.054)	-0.038 (0.084)	0.097	0.819

RU - Specific Utility Example



RU - Specific Utility Example

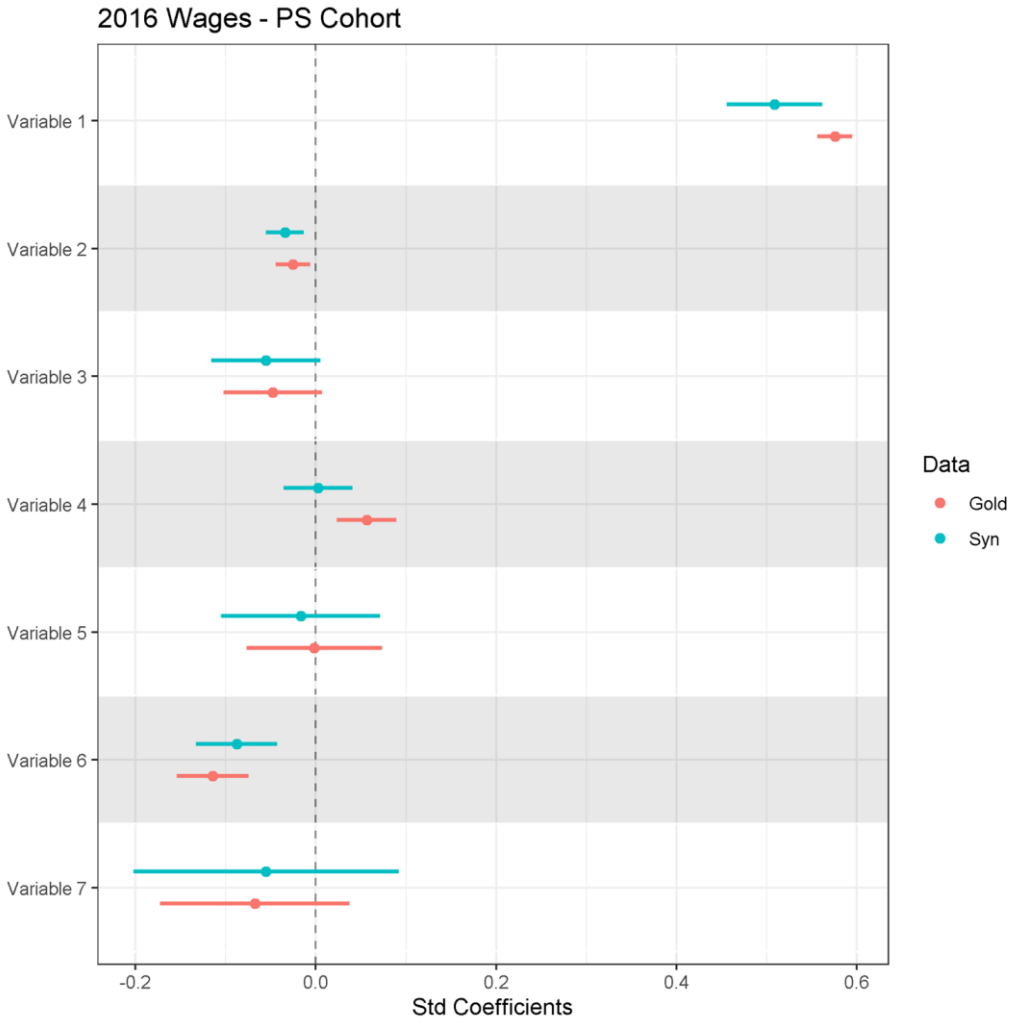
- Issue
  - Synthesis model was not well tuned for wages
  - Only one lag was used for employment in each sector
  - Quarterly wage by sector was creating sparse data

- Solution
  - More lags for wages used in the predictor set
  - Yearly global wage is synthesized first with more lags
  - Quarterly percentages with lags
  - Sector percentage within quarterly with same sector lags and all quarters

RU - Specific Utility Example

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	CI Overlap
Variable 1	0.576 (0.01)	0.509 (0.026)	1.176	0.094
Variable 2	-0.025 (0.01)	-0.034 (0.01)	0.162	0.776
Variable 3	-0.048 (0.028)	-0.055 (0.03)	0.046	0.938
Variable 4	0.129 (0.026)	0.107 (0.033)	0.148	0.827
Variable 5	0.056 (0.017)	0.003 (0.019)	0.556	0.246
Variable 6	-0.002 (0.038)	-0.017 (0.043)	0.069	0.912
Variable 7	-0.114 (0.02)	-0.088 (0.022)	-0.229	0.688

RU - Specific Utility Example



RU - General Utility

- How well does the synthetic data reproduce the variable relationships in the GSDS in general
  - Global measure
  - Not tied to a specific analysis
- Options
  - Kullback-Leibler divergence
  - Cluster analysis
  - Propensity scores ←
- How well one can discriminate between gold and synthetic data

RU - General Utility

- Propensity score method

	Dataset	Subj ID	Variable 1	Variable 2	Variable 3
Gold Standard	0	1	1	9	3
	0	2	0	12	5
	0	3	0	4	1
	0	4	1	6	1
	...	...	...	...	...
Synthetic	1	S1	1	10	0
	1	S2	0	12	0
	1	S3	0	5	0
	1	S4	1	4	0

RU - General Utility

- Overall measure of utility Snoke et al. 2018 and Woo et al. 2009
  - Mean square error of propensity scores (pMSE)
  - pMSE → 0, less discrepancy between real and synthetic datasets
  - Relative utility measure
- Variable importance
  - Variables with high importance indicate discrepancies between the GSDS and SDS
  - Useful diagnostic method

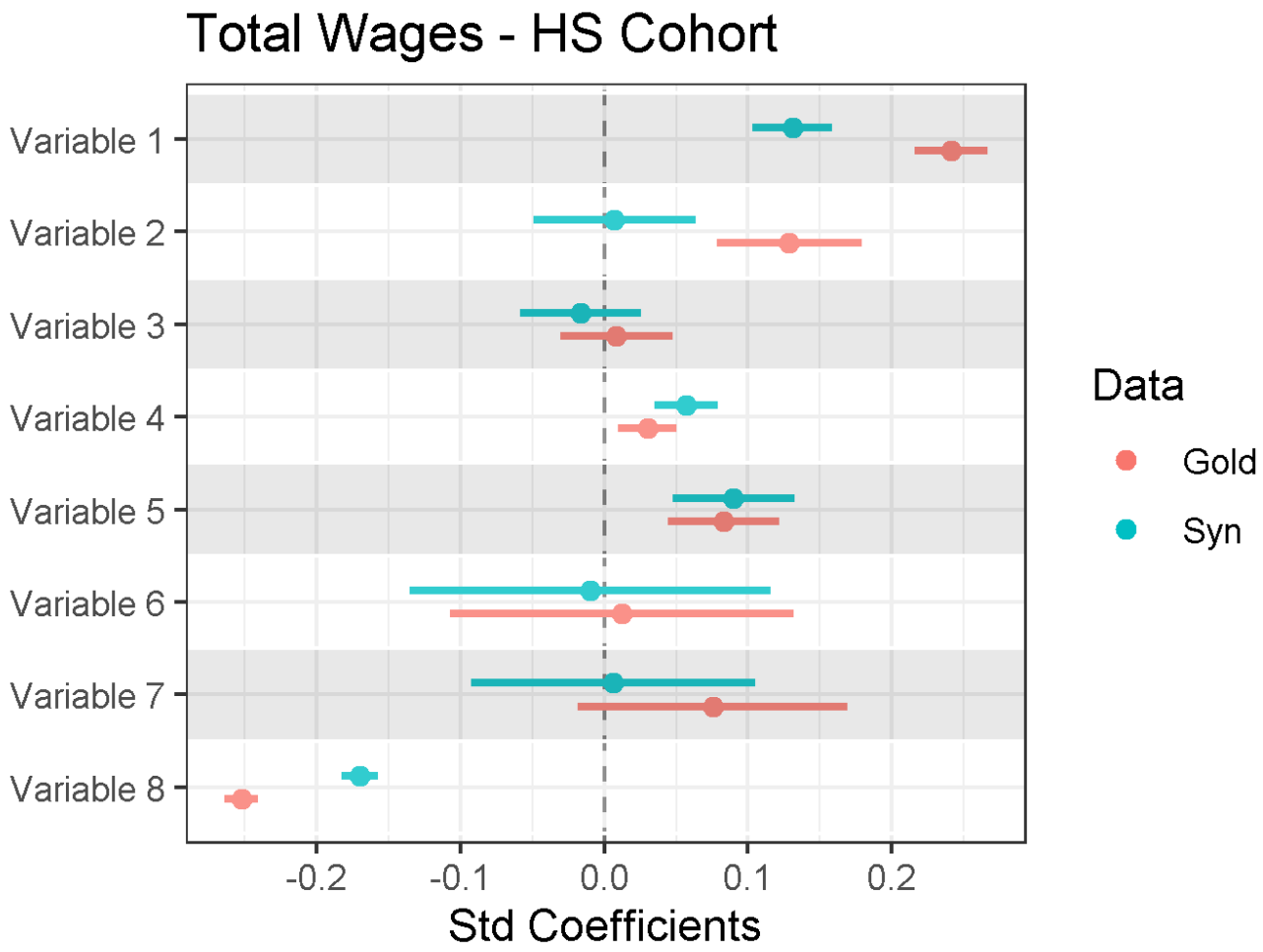
RU - Beta Testing

- Research utility beta testers each developed 4 models to test on the synthetic data in R and SAS
- Models included
  - Linear regression
  - Logistic regression
  - Multinomial regression
- Asked to attempt pooling results
- Completed code submitted to be run on the gold standard dataset
- Compare inferences

Beta Testing – RU Results

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	CI Overlap
Variable 1	0.241 (0.013)	0.131 (0.013)	1.617	-1.145
Variable 2	0.129 (0.026)	0.007 (0.028)	0.893	-0.164
Variable 3	0.008 (0.02)	-0.016 (0.02)	0.237	0.685
Variable 4	0.03 (0.01)	0.057 (0.011)	-0.489	0.354
Variable 5	0.083 (0.02)	0.09 (0.021)	-0.064	0.915
Variable 6	0.012 (0.061)	-0.01 (0.061)	0.068	0.908
Variable 7	0.075 (0.048)	0.006 (0.048)	0.275	0.631
Variable 8	-0.253 (0.006)	-0.17 (0.006)	-2.553	-2.42

Beta Testing – RU Results



MLDS Center
Maryland Longitudinal Data System

Disclosure Risk

MLDS Center
Maryland Longitudinal Data System

DR - Data in GSDS and Synthetic Data

- All variables in all synthetic data sets are synthesized
- No ID numbers or identifying information carried over from GSDS to synthetic data
- No geographic information in the synthetic data: school districts, zip codes, or census tracks or blocks
- No identifying information about schools, colleges, universities, or employers in the synthetic data

Disclosure Risk

Do the synthetic data pose a risk to the disclosure of confidential protected information?

We use the GSDS as our 'external dataset.' This assumes a worst case situation where an intruder might know almost everything about specific individuals or subgroups.

MLDS Center
Maryland Longitudinal Data System

Disclosure Risk Assessment

Two types of Disclosure Risk:

A. *Identification Disclosure:* The potential for an intruder to match a given record with a specific individual

A. *Attribute Disclosure:* The possibility that sensitive characteristics of smaller subpopulations could be determined

DR - Assessing Risk: Identification Disclosure

- Identification Disclosure rests on the assumption that the synthesized data contains identifiable information about individuals from the GSDS on which it was modeled.
- For fully synthesized data the “cases” do not exist (there are no “real” records), so theoretically, there is no identification disclosure risk (the probability would conservatively be 1/N).
- One way to examine identification disclosure in fully synthesized data is to see if it is possible to determine if a specific record from the GSDS is in the SD.

MLDS Center
Maryland Longitudinal Data System

Disclosure Risk Findings

- Disclosure Risk assessments rests on the assumption that the synthesized data contains identifiable information about individuals from the GSDS on which it was modeled.
  - The created synthetic data sets are **Fully Synthetic** datasets
  - Fully Synthetic data **anonymizes records -** no actual sets of values from the GSDS exist in the data.
  - The synthetic implicates we have created have **extremely low** disclosure risk

DR - High School Cohort

Category	Disclosure Risk
Overall Disclosure Risk	0.000002
Disclosure Risk for Average Person (records near the median across categories)	0.000029
Known NAIC codes (NAIC=22 Utilities Sector)	0.001314

DR - Post-Secondary Cohort

Category	Disclosure Risk
Overall Disclosure Risk	0.000006
Disclosure Risk for Average Person (records near the median across categories)	0.000114
Known NAIC codes (NAIC=21 Mining/Extraction)	0.002994

MLDS Center
Maryland Longitudinal Data System

DR - Assessing Risk: Attribute Disclosure

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our “outside source” of information
- The use of the original data provides a worst case scenario of external information an intruder might possess
- Disclosure risk is calculated as the odds of determining sensitive information (such as wages or test scores) using a process of probability matching between the synthetic and “outside” data

MLDS Center
Maryland Longitudinal Data System

Assessing Attribute Disclosure Risk

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our "outside source" of information

There is a 1% chance of finding a set of attributes across demographic characteristics in both GSDS and SDP

- 0% of these attributes provided accurate grade information
- <0.01% had wage info within 20% of average wage and 0% within 10%

DR - High School Cohort

- Focusing on the 824 unique cases found in the GSD as the 'external' data set.
- Match with the SD by characteristics.

- 284 (34.5%) have no records in SD (no risk)
- 651 (79.0%) have no records in at least one of the SD sets (uncertainty)
- 173 (21.0%) match on all characteristics

DR - High School Cohort

- Comparison to HS
- 0% of cases had both school and acad year
  - No acad year match
  - 31% school achievement match
- Comparison to Avg Wages
- 0% of cases had matching information on NAIC and Wages.
  - 60% of records had a match with NAIC codes
  - <2% had an average wage within 20% of the GS avg wages and none within10%.

DR - Post-Secondary Cohort

- Focusing on the 3,501 unique cases found in the GSD as the 'external' data set.
- Match with the SD by characteristics.

- 1575 (45.0%) have no records in SD (no risk)
- 1,556 (18.9%) have no records in at least one of the SD sets (uncertainty)
- 370 (10.6%) match on all characteristics

MLDS Center
Maryland Longitudinal Data System

DR - Assessing Attribute Disclosure Risk

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our "outside source" of information

There is a 1% chance of finding a set of attributes across demographic characteristics in both GSDS and SDP

- 0% of these attributes provided accurate grade information
- <0.01% had wage info within 20% of average wage and 0% within 10%

MLDS Center
Maryland Longitudinal Data System

DR - Beta Testing

- To date we have had two groups working on DR beta-testing.

- Using publicly available data can the beta tester identify sample uniques and gain information based on known information.
  - examining national education data sources for base information to use to try and gain sensitive information through the SDS?
  - using publicly available statistical software can they isolate uniques and gain sensitive information?

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: Mark Lachowicz - mark.lachowicz@Maryland.gov

    Terry Shaw - terry.shaw@maryland.gov

- Acknowledgement:
  - This presentation was prepared for the Research Branch of the Maryland Longitudinal Data System Center (MLDSC) as part of funding from the U.S. Department of Education (R372A150045)
  - We would like to thank the entire SDP team and the staff of the MLDSC for their assistance with the work and the presentation.
  - The opinions do not represent those of the MLDS Center, its agency partners, or the federal government.