MLDS Center
Maryland Longitudinal
Data System
Better Data • Informed Choices • Improved Results

Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets

Laura M. Stapleton & Michael Woolley
Maryland Longitudinal Data System Center

Outline

- Context: the data to be housed in the MLDS Center and concerns about Confidentiality & Data Disclosure
- Data Disclosure Prevention Methods
- Synthetic Data
- MLDS Center Project on Synthetic Data

Context: the data to be housed in the MLDS Center

Person Info

- Race/ Ethnicity
- Gender
- Citizenship

New ID assigned and identifiable information behind firewall

Grades	Attendance	Course	Assessments	Status
K, 1, 2	School, days absent	Pass/fail		FARM,ELL,SE, Title1, Foreign Exch, Migrant, Homeless
3 to 8	School, days absent	Pass/fail	MSA/PARCC	FARM,ELL,SE, Title1, Foreign Exch, Migrant, Homeless
9 to 12	School, days absent	Classes, Grades	HSA/PARCC, Bio/Govt, AP/PSAT/IB	FARM,ELL,SE, Title1, Foreign Exch, Migrant, Homeless

Postsecondary (MHEC)

Year	Enrollment (MHEC&NSLC)	Course	Financial Aid
1	Institution, remediation status, program	Courses, Grades	gross income, aid type, award amount
2	Institution, program	Courses, Grades	gross income, aid type, award amount
3 +	Institution, program	Courses, Grades	gross income, aid type, award amount

Workforce (DLLR)

Organization where employed	Quarterly Wages	Sector of organization

Concerns about Confidentiality & Data Disclosure

- MLDS Center, by law, cannot share individually identifiable information

  > *"Direct access to data in the Maryland Longitudinal Data System shall be restricted to authorized staff of the Center"*

- Need staff appointment to access; only a few staff have access to the identifiable information behind the firewall

- Center staff will not have time to address all possible research/policy questions; therefore providing some access to data to others would be advantageous

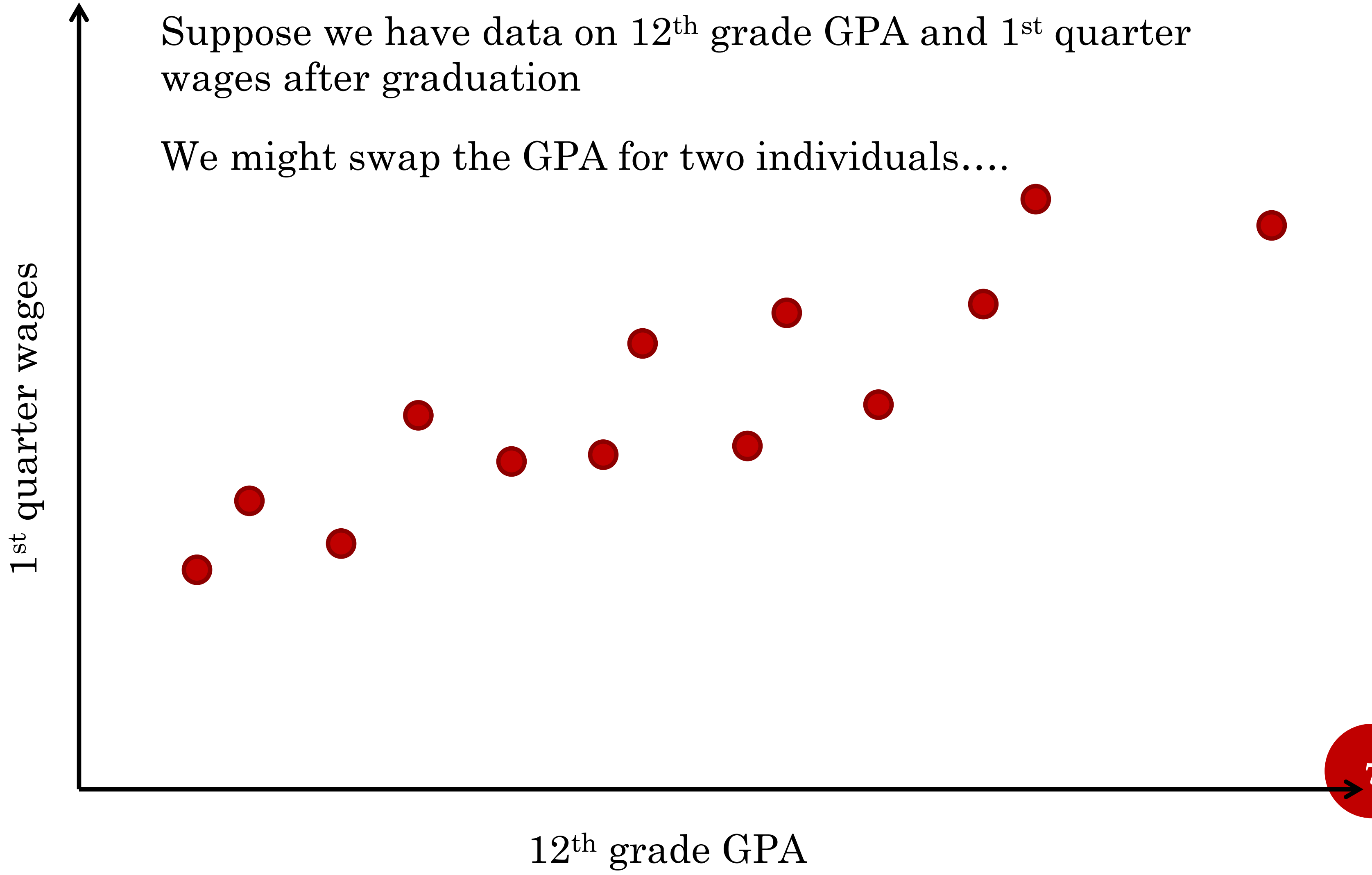
DATA DISCLOSURE PREVENTION METHODS

- Data Swapping
- Data Perturbation
- Providing Only Sample of Data from Census
- Partially Synthetic Data
- Fully Synthetic Data

Data Disclosure Prevention Methods

- Data Swapping
  - Move data from one person to another (and vice versa)
  - Not all variables are typically swapped
  - Not all observations (people) have their data swapped (referred to as the *swap rate*)
  - Some people are targeted for swapping (have unique characteristics)
  - Depending on the amount of people with swapped data, multivariate relations among variables may be affected, harming utility

Data Disclosure Prevention Methods – DATA SWAPPING



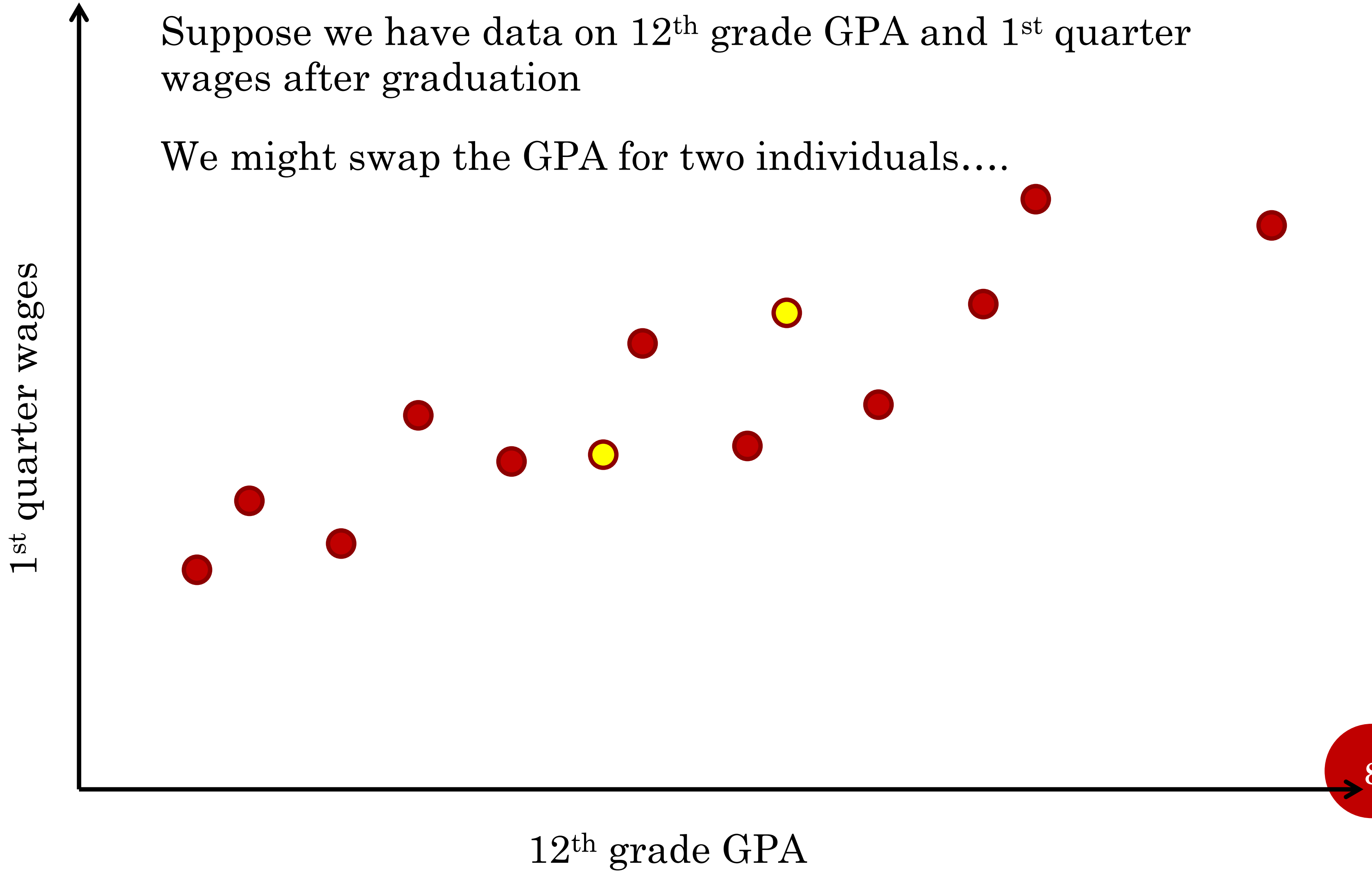
Suppose we have data on 12th grade GPA and 1st quarter wages after graduation

We might swap the GPA for two individuals….

Data Disclosure Prevention Methods – DATA SWAPPING

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation

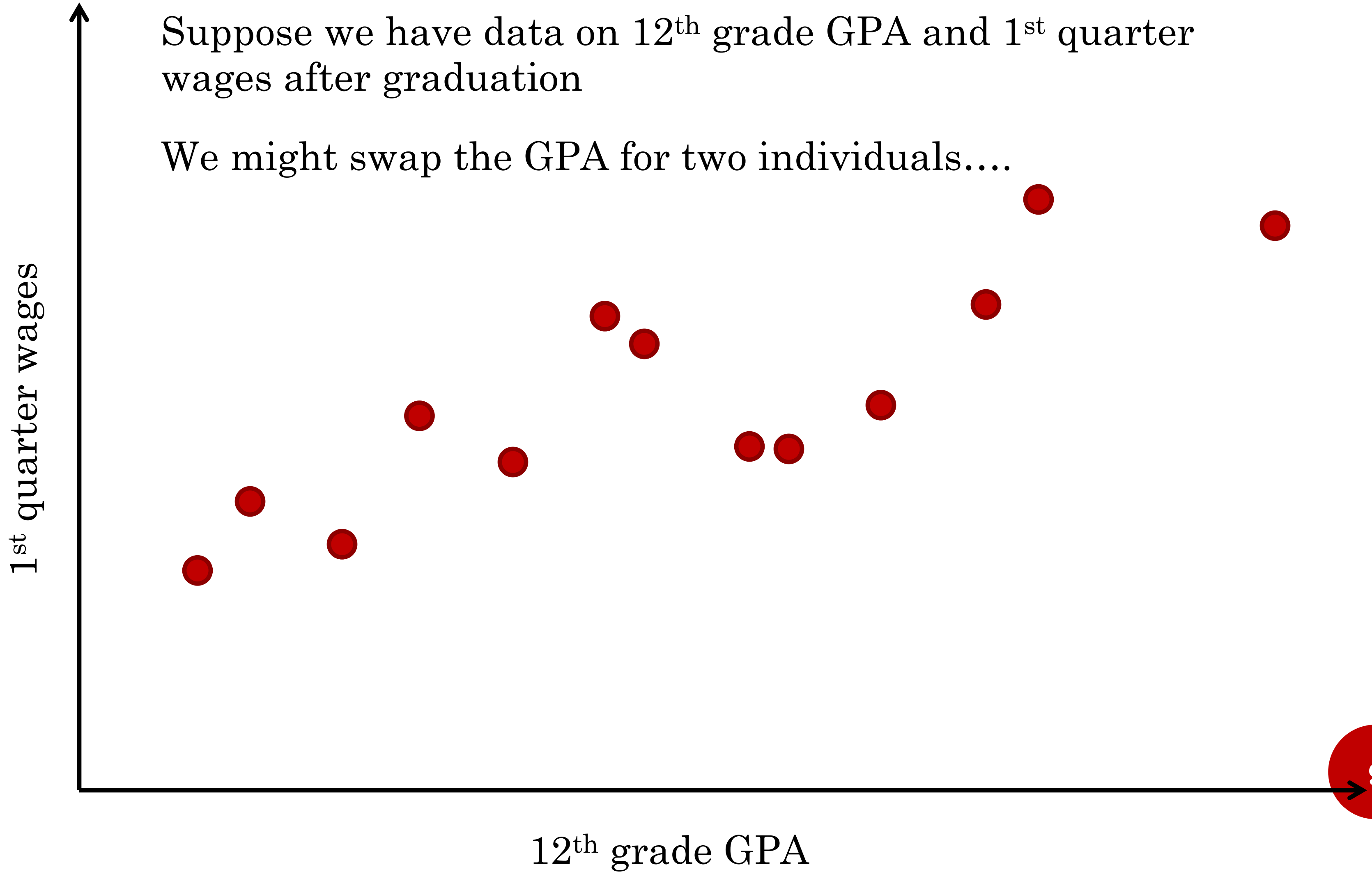
We might swap the GPA for two individuals….



Data Disclosure Prevention Methods – DATA SWAPPING

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation

We might swap the GPA for two individuals….

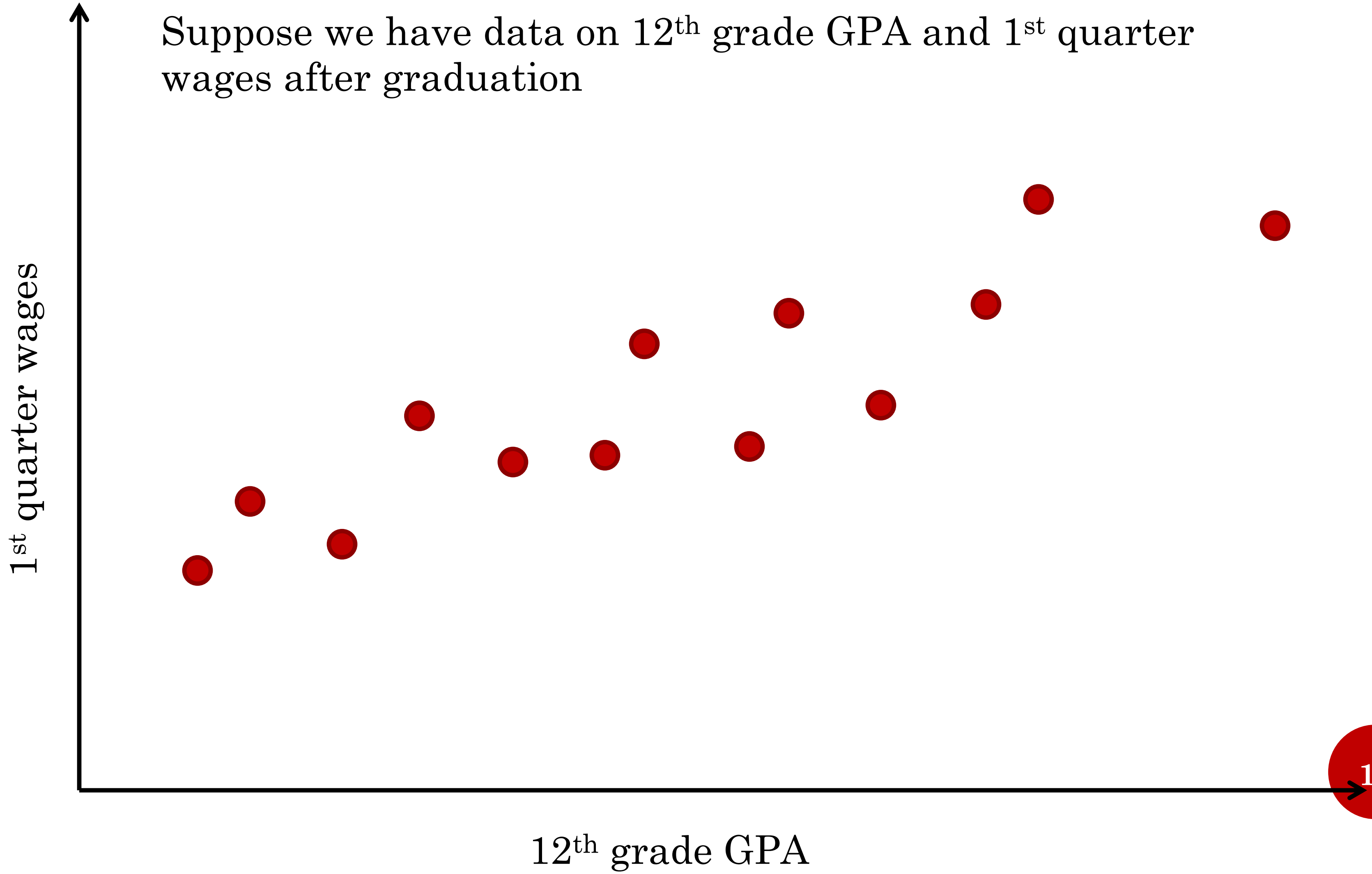


DATA DISCLOSURE PREVENTION METHODS

- Data Perturbation
  - Also referred to as "*Noise Infusion*"
  - Random error is added to each data point
  - This error may be at a specific level (e.g., 10%) so multipliers of .9 and 1.1 (with some variability) can be used
  - Complex models can be used to have differential amounts of perturbation within subgroups or across variables
  - Less likely to have adverse impacts on multivariate relations as compared to swapping

Data Disclosure Prevention Methods – DATA PERTURBATION

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation

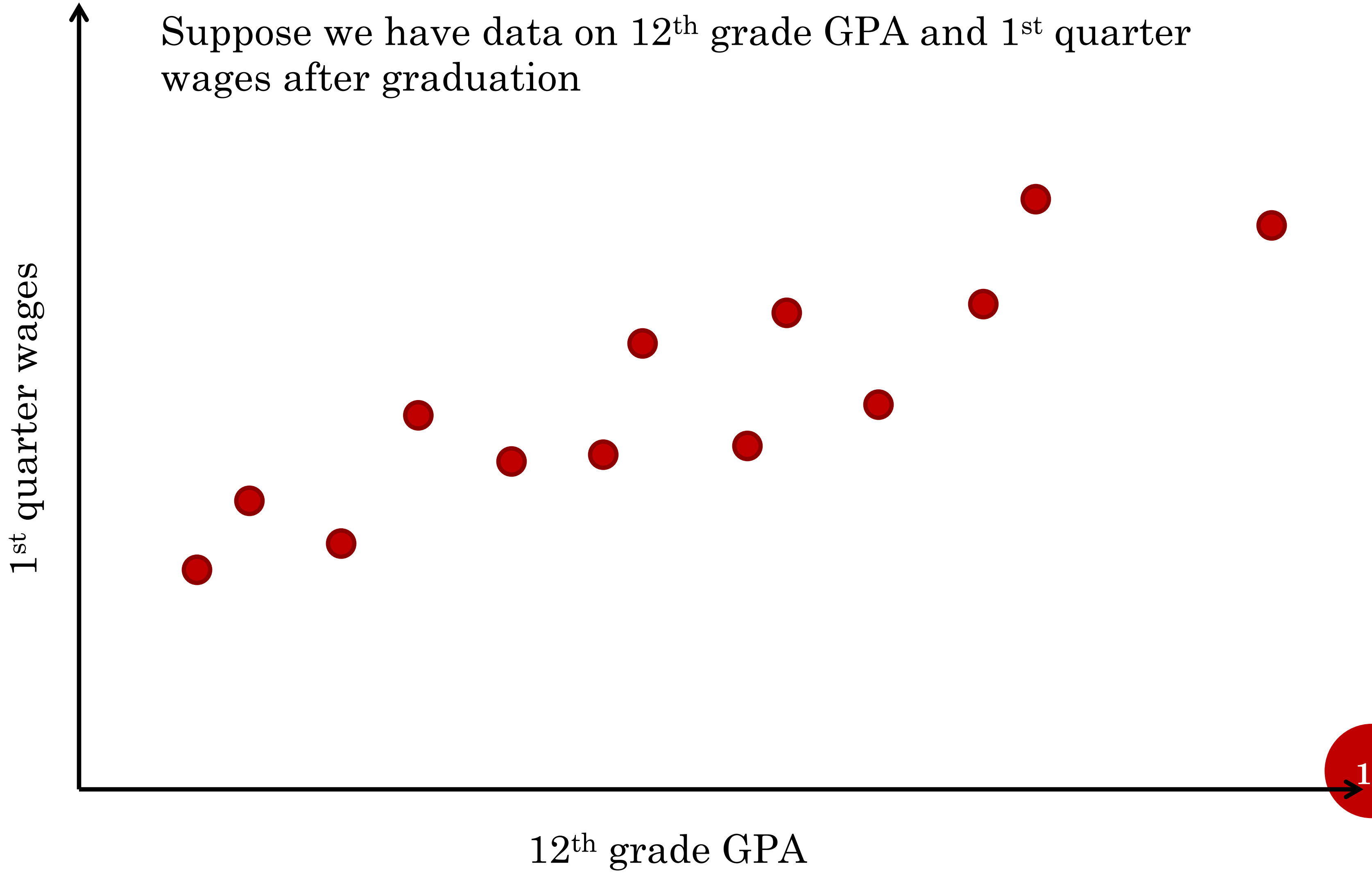


Data Disclosure Prevention Methods

- Providing Only Sample of Data
  - The MLDS Center has a census of data from the Maryland Public Schools and postsecondary institutions
  - One might release only a sample of these data (from some random selection process)
  - This process would violate the terms of the creation of the MLDS Center
  - However, this process could be used in conjunction with the synthetic data process for further identity protection

Data Disclosure Prevention Methods – PROVIDING A SAMPLE

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



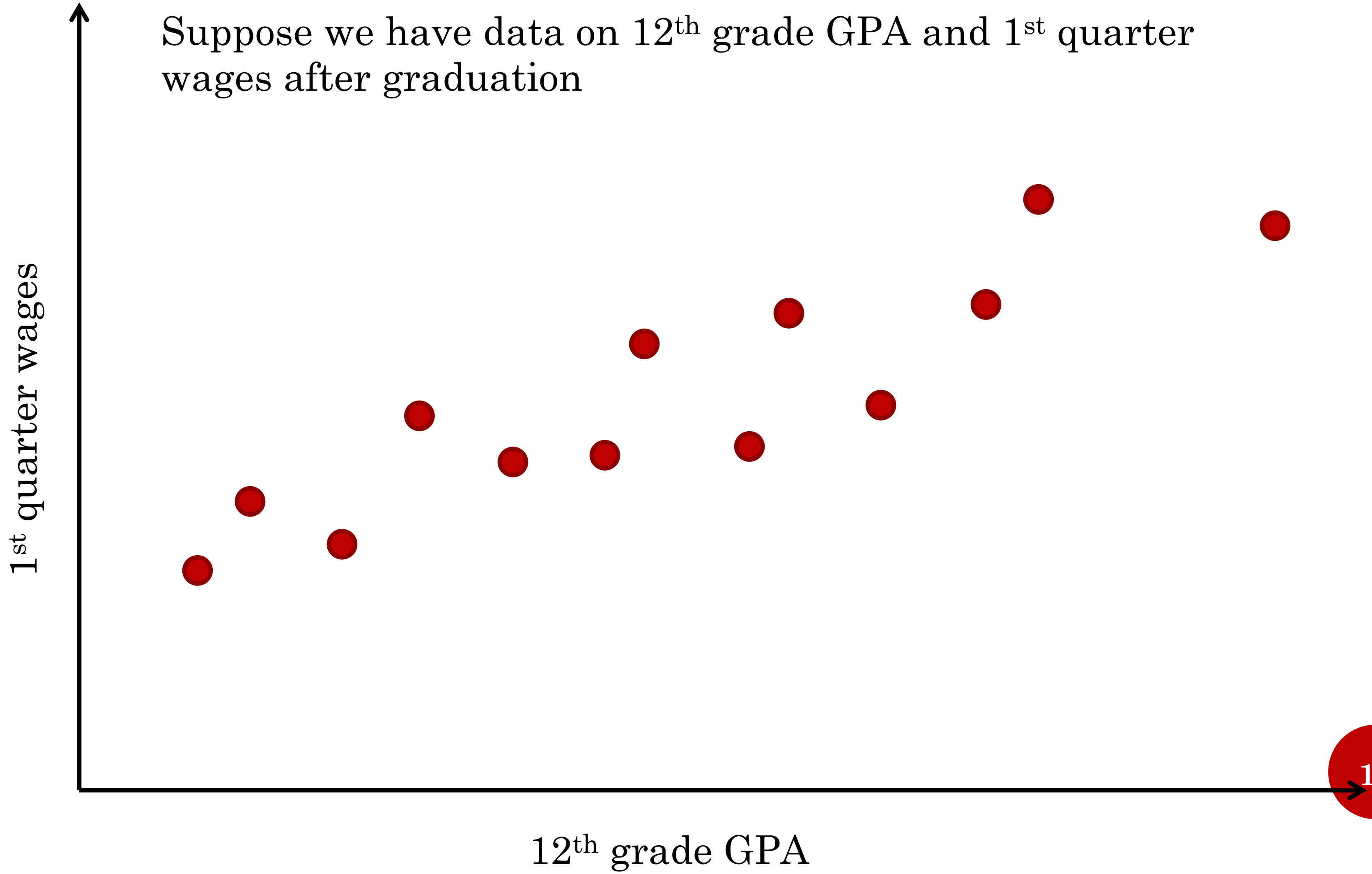
Data Disclosure Prevention Methods

- Partially Synthetic Data
  - Create a dataset that contains the source data
  - Partially fabricate some of the data (instead of perturbing a variable value or swapping it out, create a new value)
  - Data are fabricated based on known characteristics about the source data (distribution, relations with other variables)
  - If individual-level source data are retained, would violate terms of MLDS

- Fully Synthetic Data
  - Create a dataset that shares characteristics of the source data
  - Entirely fabricated data

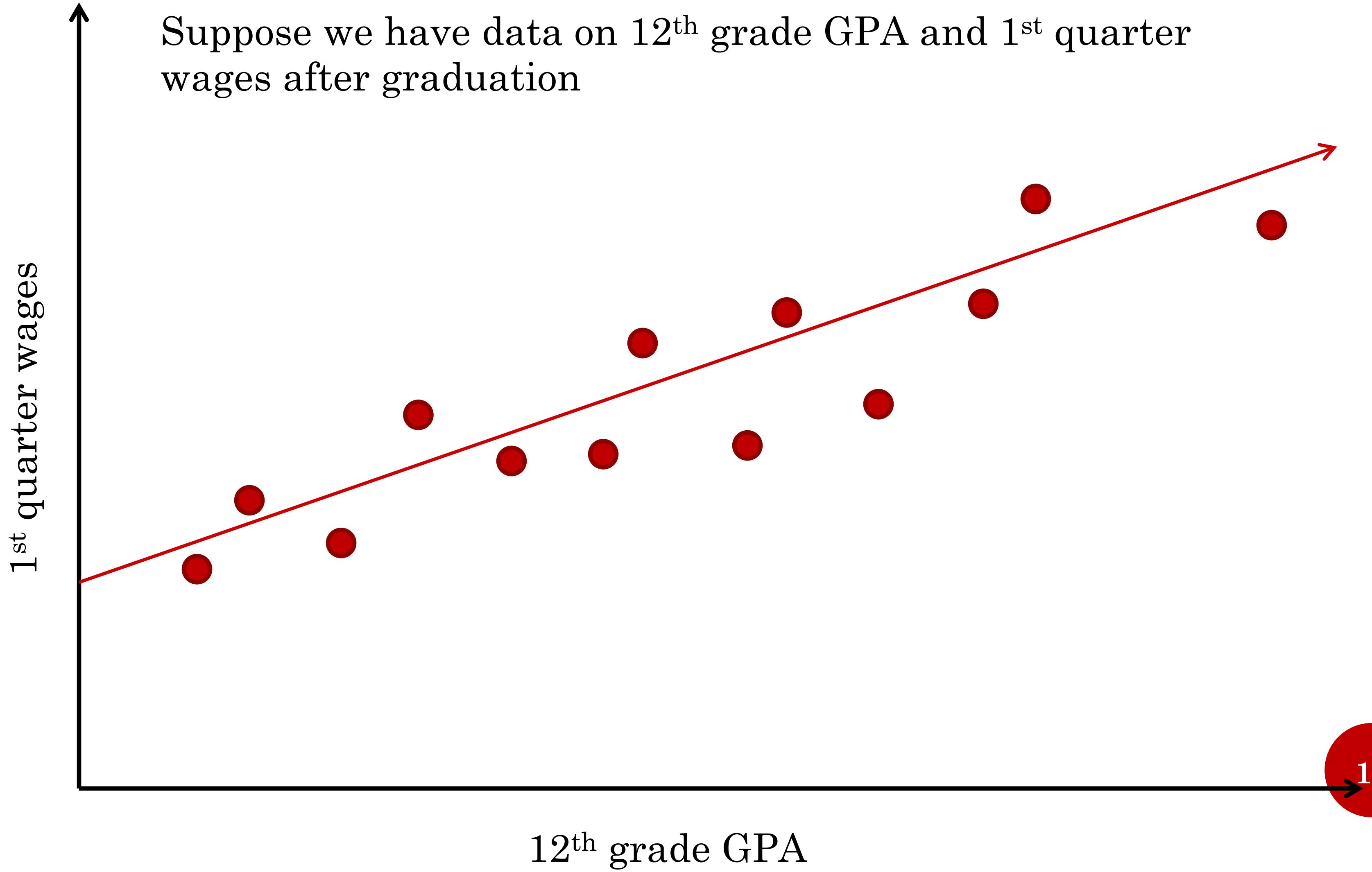
Data Disclosure Prevention Methods – SYNTHETIC DATA

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



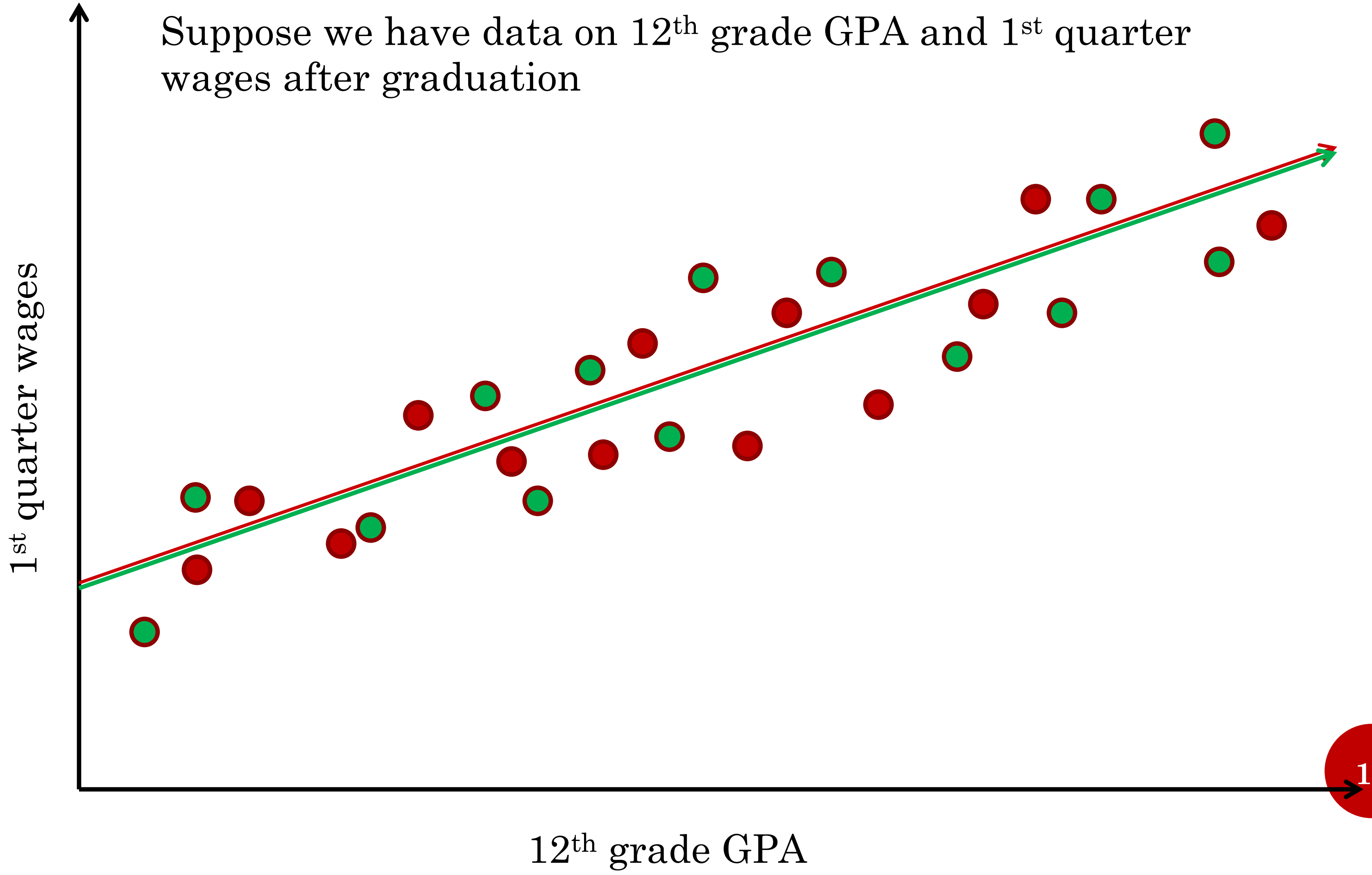
Data Disclosure Prevention Methods – SYNTHETIC DATA

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



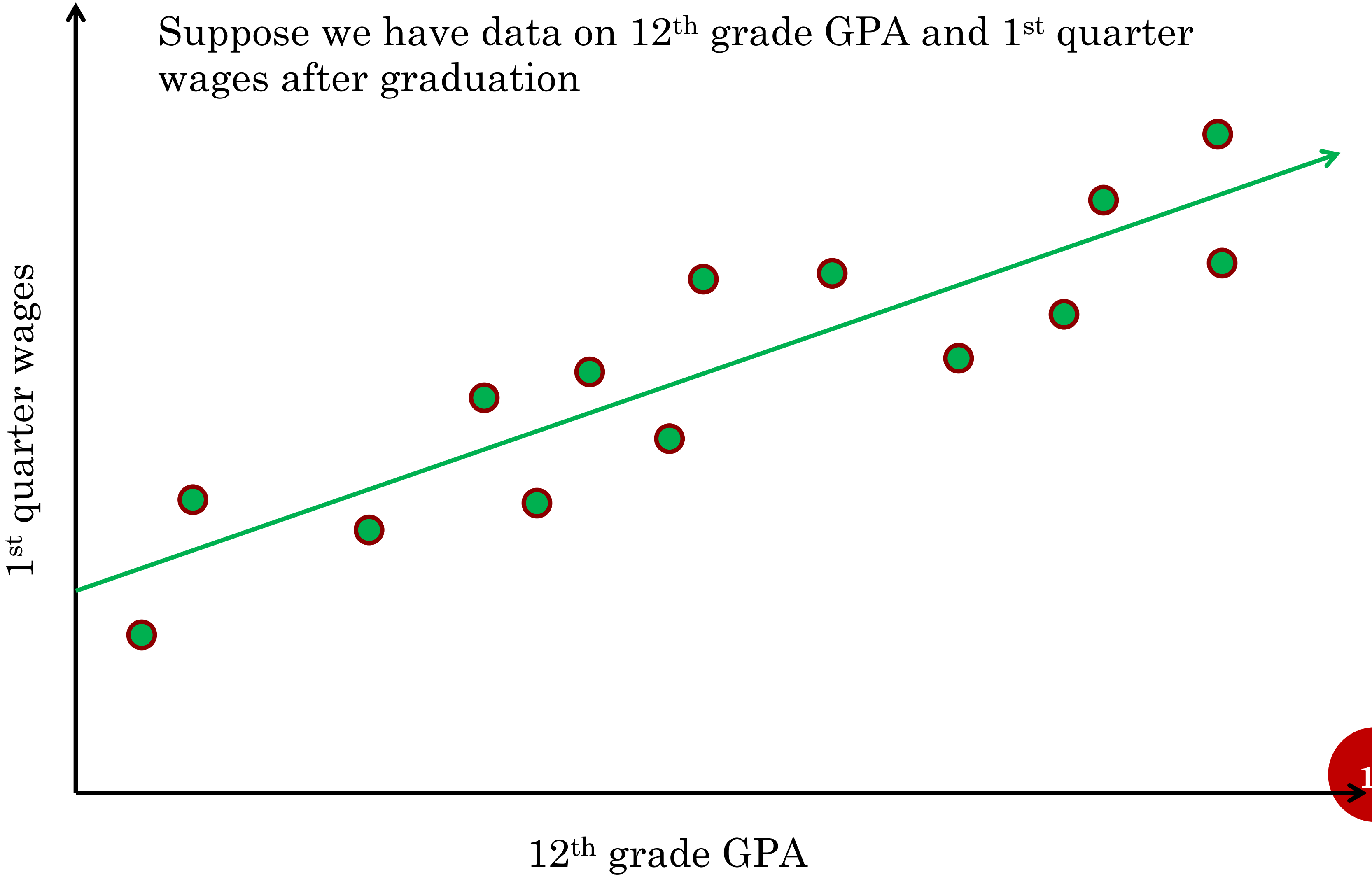
Data Disclosure Prevention Methods – SYNTHETIC DATA

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



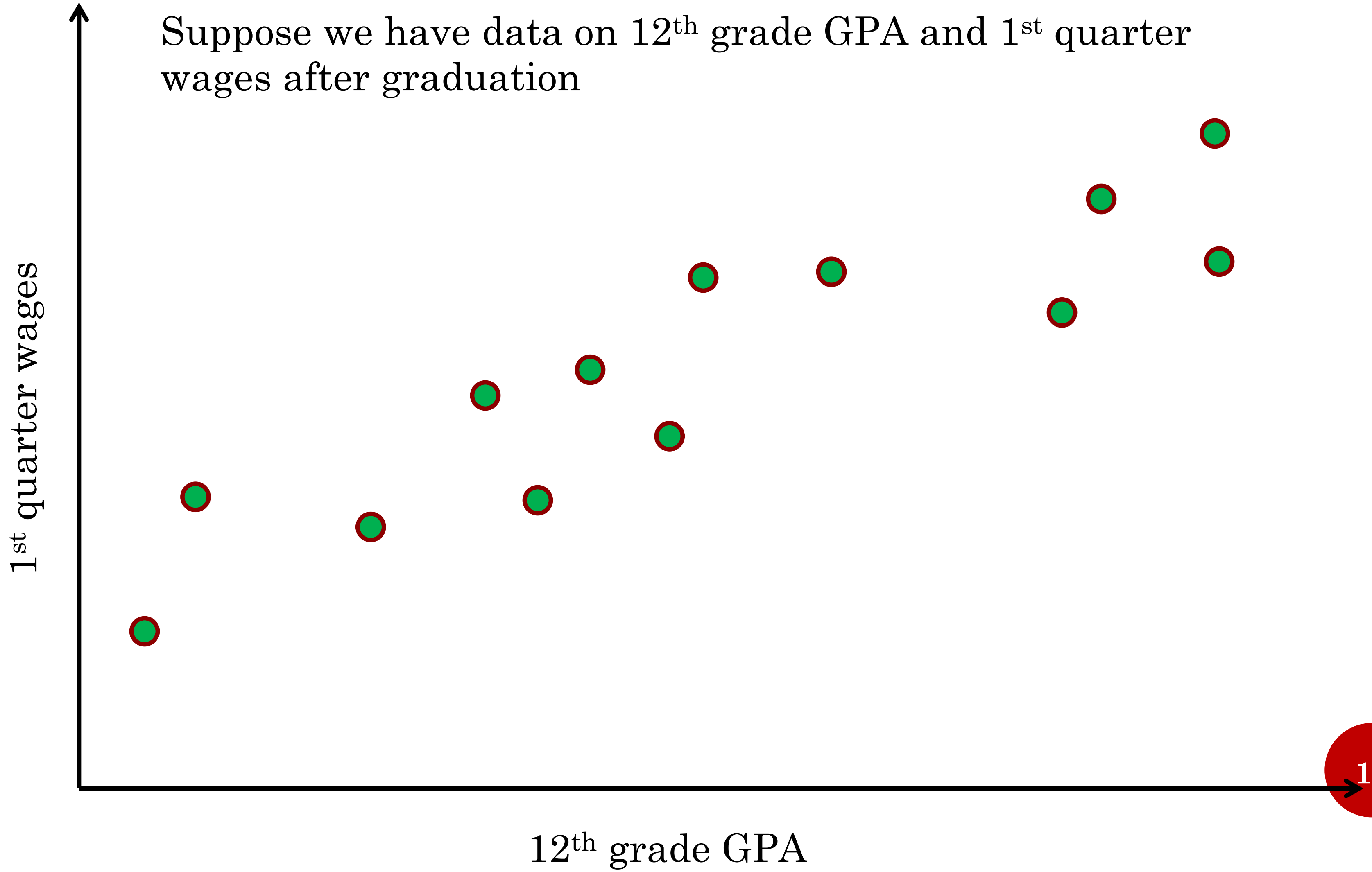
Data Disclosure Prevention Methods – SYNTHETIC DATA

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



Data Disclosure Prevention Methods – SYNTHETIC DATA

Suppose we have data on 12th grade GPA and 1st quarter wages after graduation



Data Disclosure Prevention Methods – SYNTHETIC DATA

Another Example…

First, let's talk about missing data…

X	Y
8	10
5	8
8	9
2	4
7	7
8	9
7	7
7	6
3	?
2	?

Correlation$_{X,Y}$ = .87

Using that correlation, we can impute values for the missing values

X	Y
8	10
5	8
8	9
2	4
7	7
8	9
7	7
7	6
3	**2**
2	**3**

These imputed values are a random draw from a probability distribution

Data Disclosure Prevention Methods – SYNTHETIC DATA

An entirely synthetic data set could be created, utilizing known characteristics of the data:

X	Y
8	10
5	8
8	9
2	4
7	7
8	9
7	7
3	2
7	6
2	3

$\text{Correlation}_{X,Y}$ =		.87
X:	mean =	5.7
	variance =	5.6
	skew =	-0.7
	kurtosis =	-1.4
Y:	mean =	6.5
	variance =	6.7
	skew =	-0.5
	kurtosis =	-1.0

X	Y
2	**2**
7	**6**
7	**7**
6	**7**
4	**6**
9	**8**
5	**6**
7	**10**
3	**2**
8	**9**

Data Disclosure Prevention Methods – SYNTHETIC DATA

Once synthetic data are created, evaluate the utility (or how close the synthetic data mirrors truth):

Gold Standard

$\text{Correlation}_{X,Y}$ =		.87
X:	mean =	5.7
	variance =	5.6
	skew =	-0.7
	kurtosis =	-1.4
Y:	mean =	6.5
	variance =	6.7
	skew =	-0.5
	kurtosis =	-1.0

Synthetic

$\text{Correlation}_{X,Y}$ =		.86
X:	mean =	5.8
	variance =	4.6
	skew =	-0.4
	kurtosis =	-1.8
Y:	mean =	6.3
	variance =	6.2
	skew =	-0.6
	kurtosis =	-.06

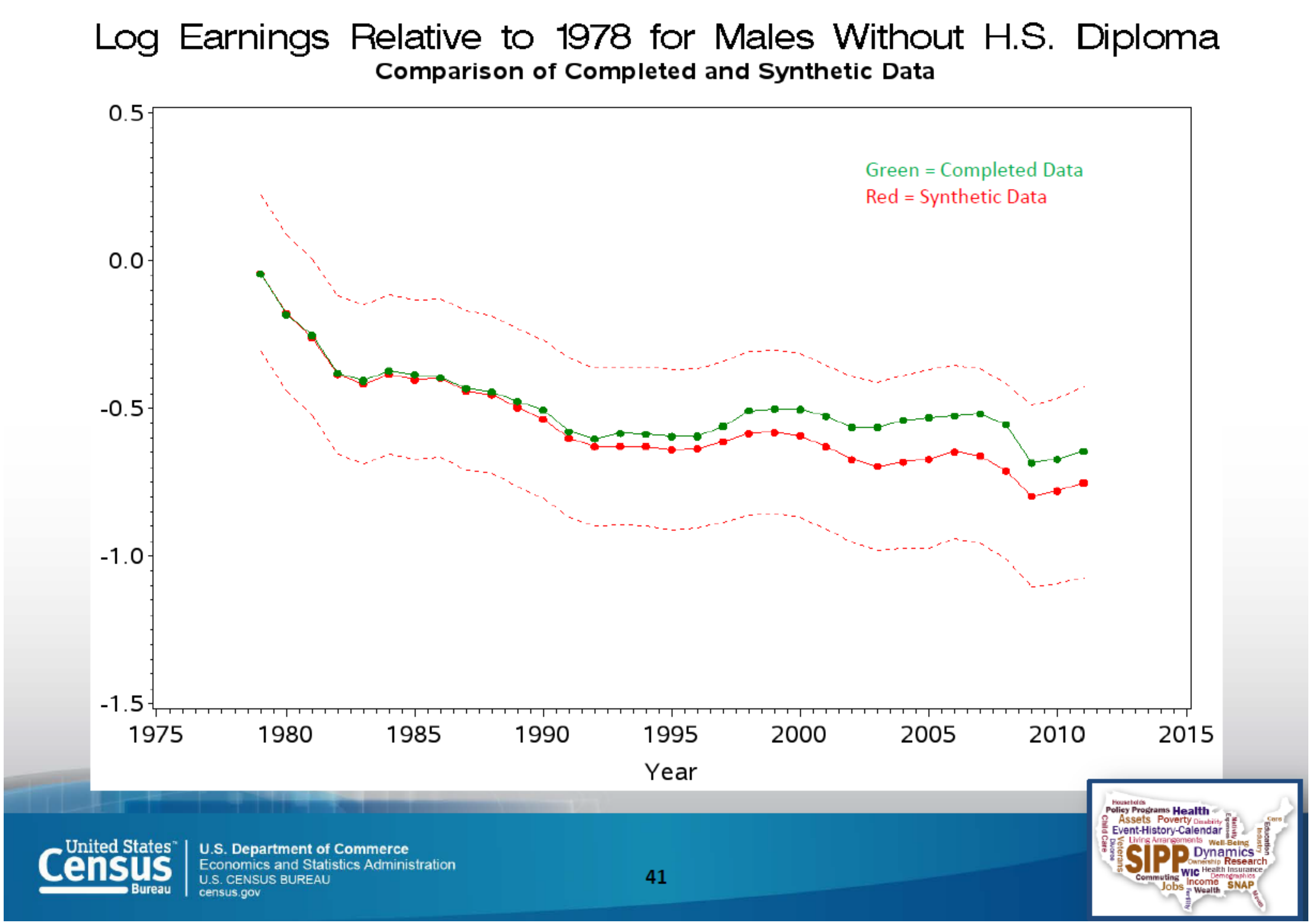
Synthetic Data

- The synthetic data process involves several steps:
  - Identifying variables to synthesize
  - Evaluating distributions of those variables in *Gold Standard* data
  - Defining models that would inform the conditional distribution of the variable
  - Identifying subgroups of individuals of interest (on which the models would be imposed)
  - Imputing (synthesizing) data values from conditional probability distributions within subgroups, typically sequentially (called *synthetic implicates*)
  - Producing multiple sets of synthesized data
  - Evaluating the data for: utility, disclosure risk

SYNTHETIC DATA

- The U.S. Census SIPP program has a public access synthetic file: SSB
  - Link survey participation in SIPP with government administrative data about individuals
  - Uses a partially-synthetic process -- the only gold standard variables are gender and link to spouse
  - Chose list of variables that was “long enough to be useful” and short enough to be protected and processed in a reasonable amount of time
  - Subgroups need at least 1,000 observations for marginal probability distribution estimation
  - Started process in 2000, now up to version 6.0. Publish new file every 3-4 years.
  - SSB users can submit code to Census to have analysis run on Gold Standard data (*2-3 week turn around time*)

Synthetic Data



Log Earnings Relative to 1978 for Males Without H.S. Diploma

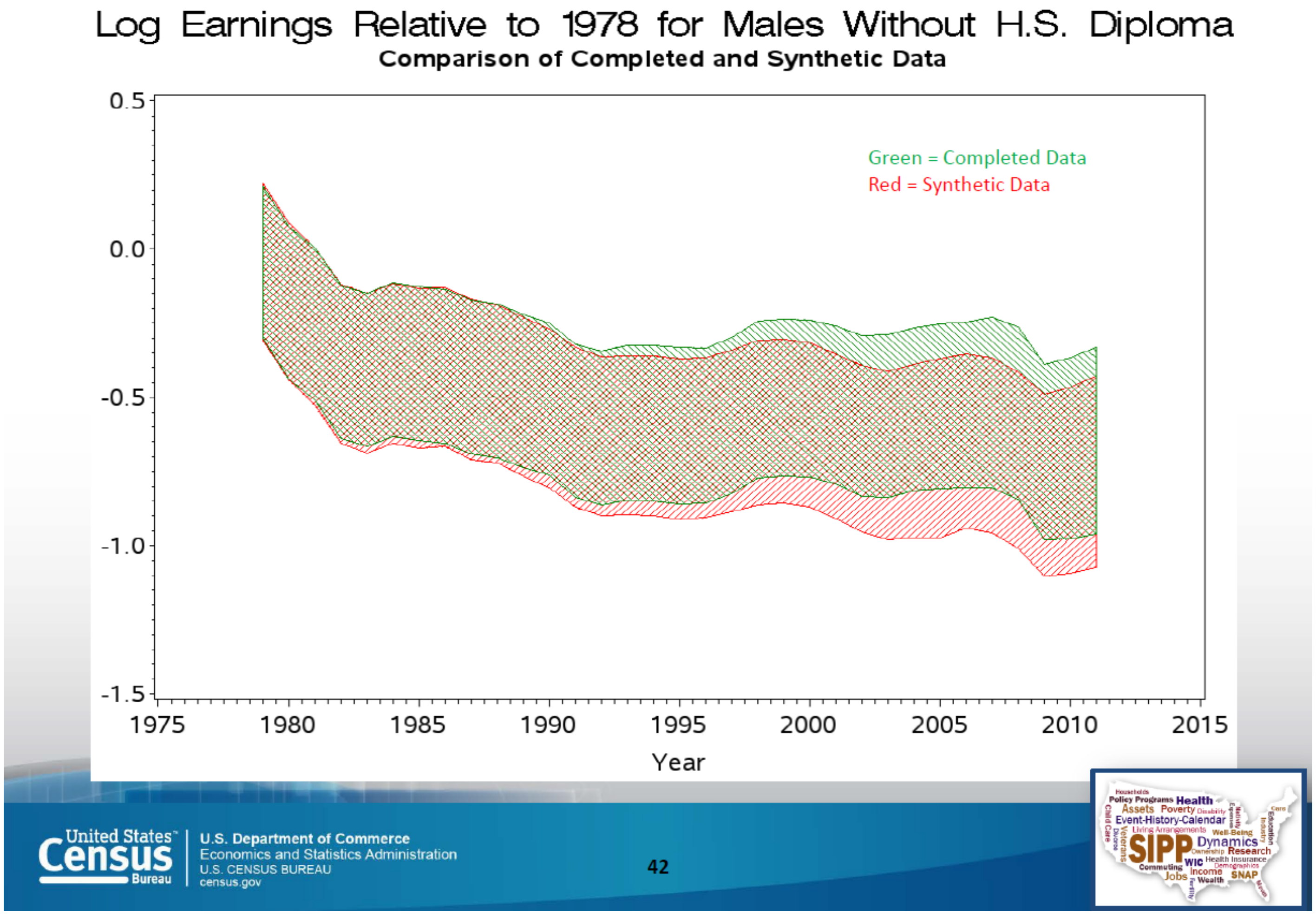
Comparison of Completed and Synthetic Data

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

41

Synthetic Data



Log Earnings Relative to 1978 for Males Without H.S. Diploma

Comparison of Completed and Synthetic Data

Green = Completed Data

Red = Synthetic Data

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

42

SYNTHETIC DATA

- Several government programs (U.S. and other countries) have synthetic data approaches to data disclosure prevention that we can learn from

- No State Longitudinal Data System is using a synthetic data approach

MLDS Center Project on Synthetic Data

- Approximately $2.6 million as part of 2015 SLDS grant from the U.S. Department of Education to the State of Maryland

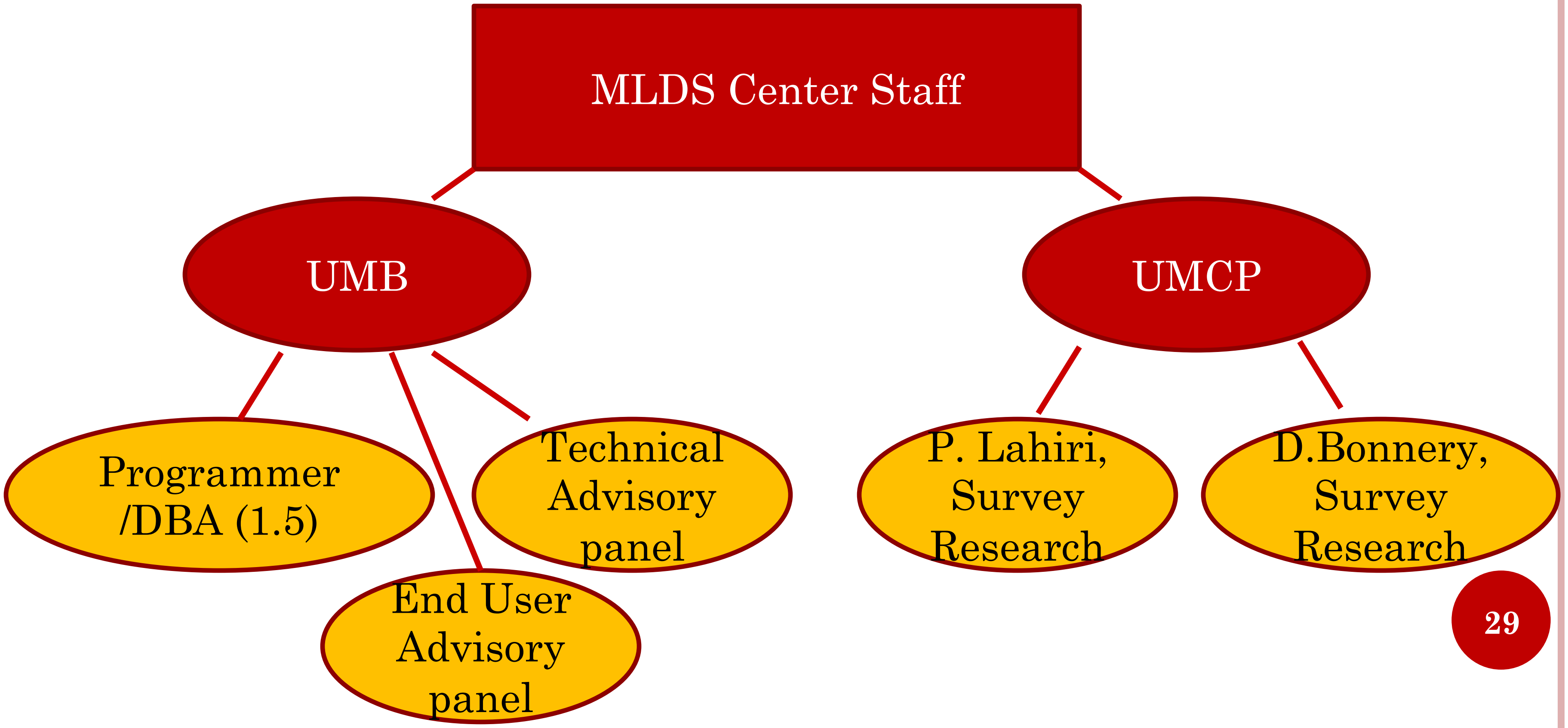
- Joint work of:

  MSDE, MLDS Center, UMB, UMCP

- Overarching goals of:
  - Creating three data files to facilitate center work
  - Creating synthetic replicas of these warehouses
  - Examining the feasibility of retaining cluster specific variance components within the synthetic data

MLDS Center Project on Synthetic Data

- PERSONNEL



MLDS Center Staff

UMB

UMCP

Programmer /DBA (1.5)

Technical Advisory panel

End User Advisory panel

P. Lahiri, Survey Research

D.Bonnery, Survey Research

MLDS Center Project on Synthetic Data

Project 1.1 – Create the three data warehouses

- Content of three files:
  - K-12 to Postsecondary focus
  - Postsecondary to Workforce focus
  - K-12 to Workforce focus
- End-user panel input to define needs in data files
  - Variables to include (and exclude)
  - Anticipated models/parameters of interest
- Hire programming staff to create the data file structure and facilitate extracts from MLDS system
- These data files will be considered the "*Gold Standard Files*"
- ***Anticipated completion – by beginning of 2017***

MLDS Center Project on Synthetic Data

Project 1.2 – Populate data files with synthetic data

- Build models for variable probability distributions
  - Input from Technical Advisory Panel and Consultant
  - Test creation models
- Fully populate the synthetic data files
- Validate the system
  - Utility rates
  - Disclosure testing
- Evaluate the use of multiple synthetic files
- Beta testing with end users
- ***Anticipated completion – by late 2018***

MLDS Center Project on Synthetic Data

Project 1.3 – Disseminate information about files

- Design web portal for access to synthetic files
- Host Education Researcher Summit
  - Training materials developed
  - Evaluate needs of the researchers

- ***Anticipated completion – by mid 2019***

MLDS Center Project on Synthetic Data

Project 1.4 – Examine feasibility of synthetic data for cluster-specific or random effects estimation

- Evaluate whether it is possible to retain some cluster-specific information in the synthetic data files
    - Partially synthetic?
    - Synthetic random effects
- Validate cluster-specific files
    - Usability rates
    - Data disclosure rates
- ***Anticipated completion – by mid 2019***

- WHAT YOU CAN DO
  - Offer to serve on the End User panel
  - Attend open forums (such as this) to discuss the issues

- QUESTIONS?