Synthetic Data for Disclosure Limitation: Overview and Research

Jerry Reiter

Institute of Statistics and Decision Sciences

Duke University

Standard approaches to disclosure limitation

- Suppress data
- Add random noise
- Recode variables
- Swap data

Partially synthetic data

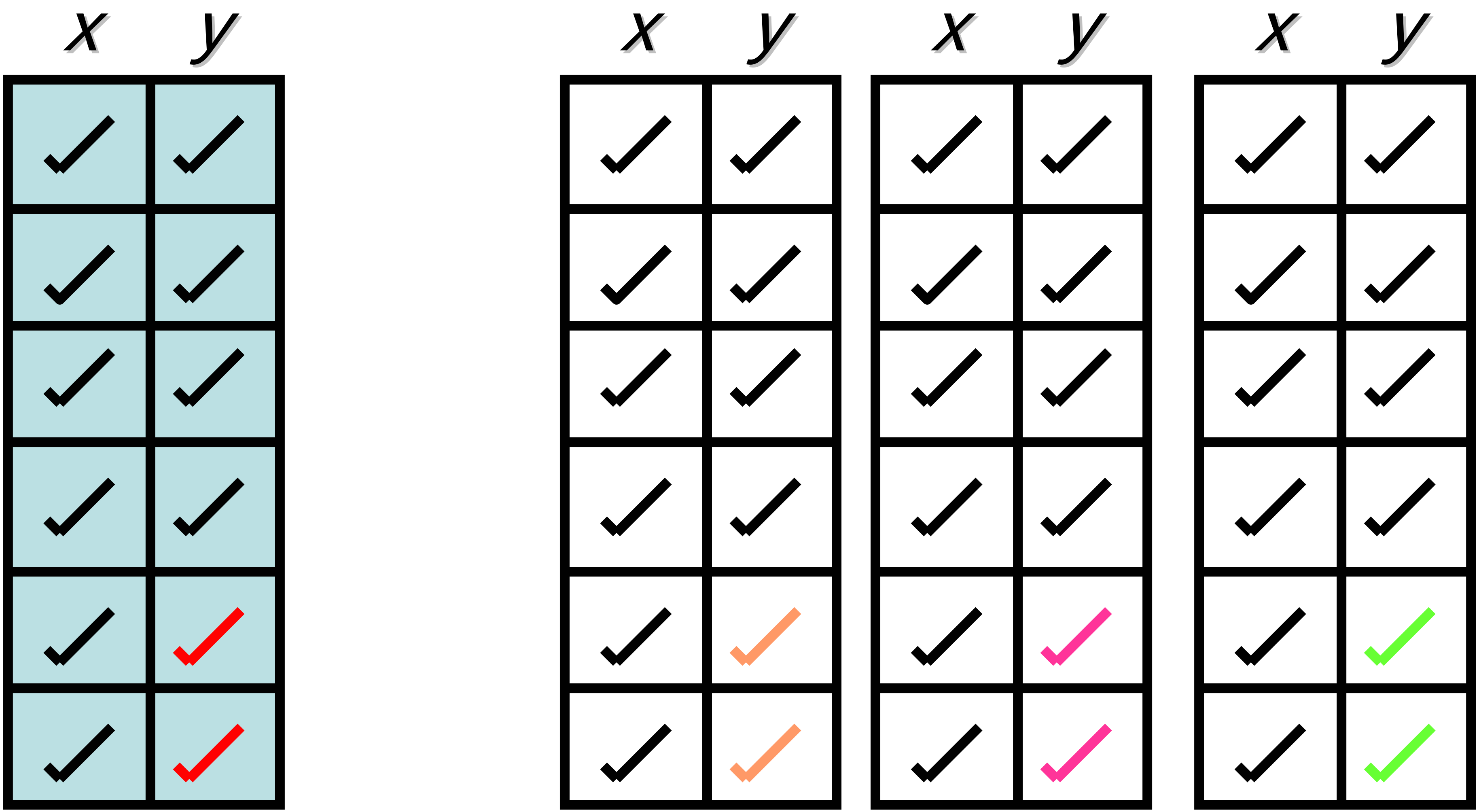
Release multiple, partially synthetic datasets so that:

- Released data comprise mix of observed and synthetic values.
- Released data look like actual data.
- Statistical procedures valid for original data are valid for released data.

Little (1993, *JOS*), Reiter (2003, 2004 *Surv. Meth*)

Observed Data

Synthetic Datasets



x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

Observed Data

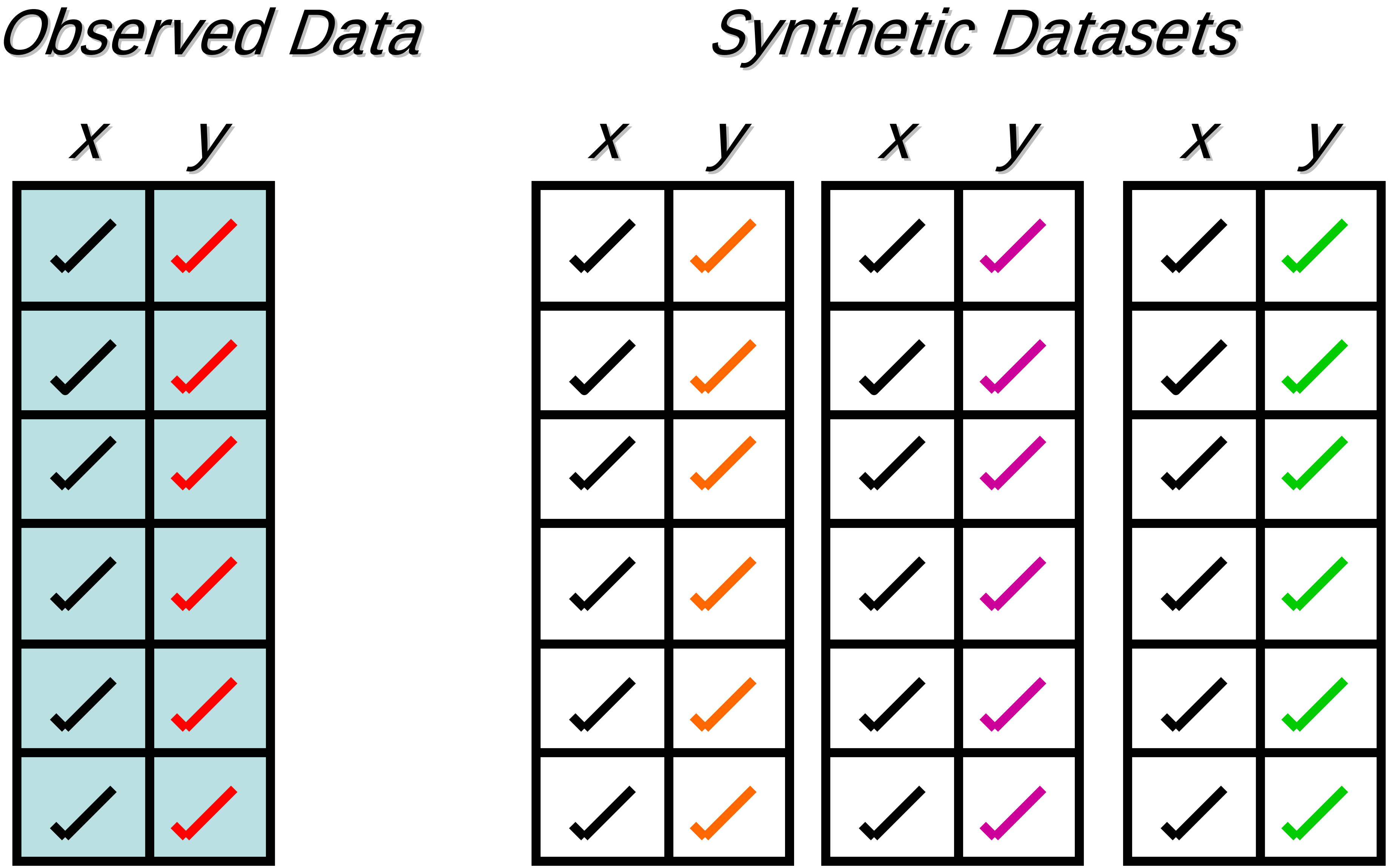
x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

Synthetic Datasets

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓



Observed Data

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

Synthetic Datasets

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

x	y
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓

Existing applications

- Replace sensitive values for selected units:

  Kennickel (1997, *Record Linkage Techniques*).

- Replace values of identifiers for selected units:
  Liu and Little (2002, *JSM Proceedings*),
  American Communities Survey GQ data.

- Replace all values of sensitive variables:

  Survey of Income and Program Participation.
  Longitudinal Business Database.

Some key benefits of approach

-- Analysts use standard methods.

-- Agency can reveal nature of SDL to public.

-- Higher intensity of SDL possible (not necessary).

-- Potential to preserve associations, maintain geographies, release data in tails.

-- Handle missing and synthetic data at same time.

-- Advances in nonparametric simulation feed into SDL.

Inference with partially synthetic datasets (no missing data)

Reiter (2003, *Survey Methodology*)

- Estimand: $Q = Q(X, Y)$
- In each synthetic dataset d_i

$$q_i = Q(d_i) \quad u_i = U(d_i)$$

Quantities needed for inference (no missing data)

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2 / (m-1)$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m$$

Inference with partially synthetic data (no missing data)

- Estimate of Q : $\bar{q}_m$
- Estimate of variance is

$$T_p = \bar{u}_m + b_m / m$$

- For large *n* and *m*, use normal based inference for *Q:*

$$\bar{q}_m \pm 1.96\sqrt{T_p}$$

How do we decide on values to synthesize?

- Replace records deemed at high risk of identification or attribute disclosure.

- Requires measurements of risk at record level that should account for:

  -- multiple copies of data,
  -- intruder knowledge and behavior,
  -- information released about models.

Research: How do we decide on values to synthesize?

- Minimizing risk not adequate.
  Need to incorporate data usefulness.

- Requires measurements of utility.
  Local ones exist, global ones do not.

- Replace values to reduce disclosure risk sufficiently while minimizing impact on data usefulness.

Issues for selection

- May be sufficient from risk perspective to replace only some (not all) identifiers.

  Example: Suppose person is unique on age, race, sex, marital status, county. But not unique if either (i) county is not released exactly, (ii) age is not released exactly, or (iii) sex and race both are not released exactly.

Issues for selection

- Replacing values for some records could impact risks for other records.

  Example: Suppose intruder knows that record 1 has larger income than record 2, and could identify these records from released incomes. Synthesizing income for either record might reduce risk for both records.

Issues for selection

- Some variables have greater impact on data usefulness than others.

  Example: Suppose X is nearly uncorrelated with variables of interest to analysts. Synthesizing X may have low impact on overall data utility.

Issues for selection

- Some values have greater impact on data usefulness than others.

  Example: Suppose record has high leverage for a regression, and leverage is attributable to one variable X. Altering X has greater impact on coefficients than altering some other values.

Research: Selecting Synthetic Datasets

- Toss out synthetic datasets that give implausible results or that are too risky.
- What metrics do we use for these decisions?
- What is the effect on inferences?

Research:
Flexible imputation models

Synthetic data models should condition on as many variables as possible.

Implications?
-- accept variance to avoid bias
-- use informative prior distributions
-- sacrifice full Bayesian simulation
-- need to worry about disclosure risk
-- use semi- and non-parametric methods

Potential uses for NCES

- Restricted access file used for remote access server.

  -- no penalty for many imputed datasets.
  -- handle missing and confidential data.

- Public access file.

  -- more intense SDL possible.
  -- handle missing and confidential data.