Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods

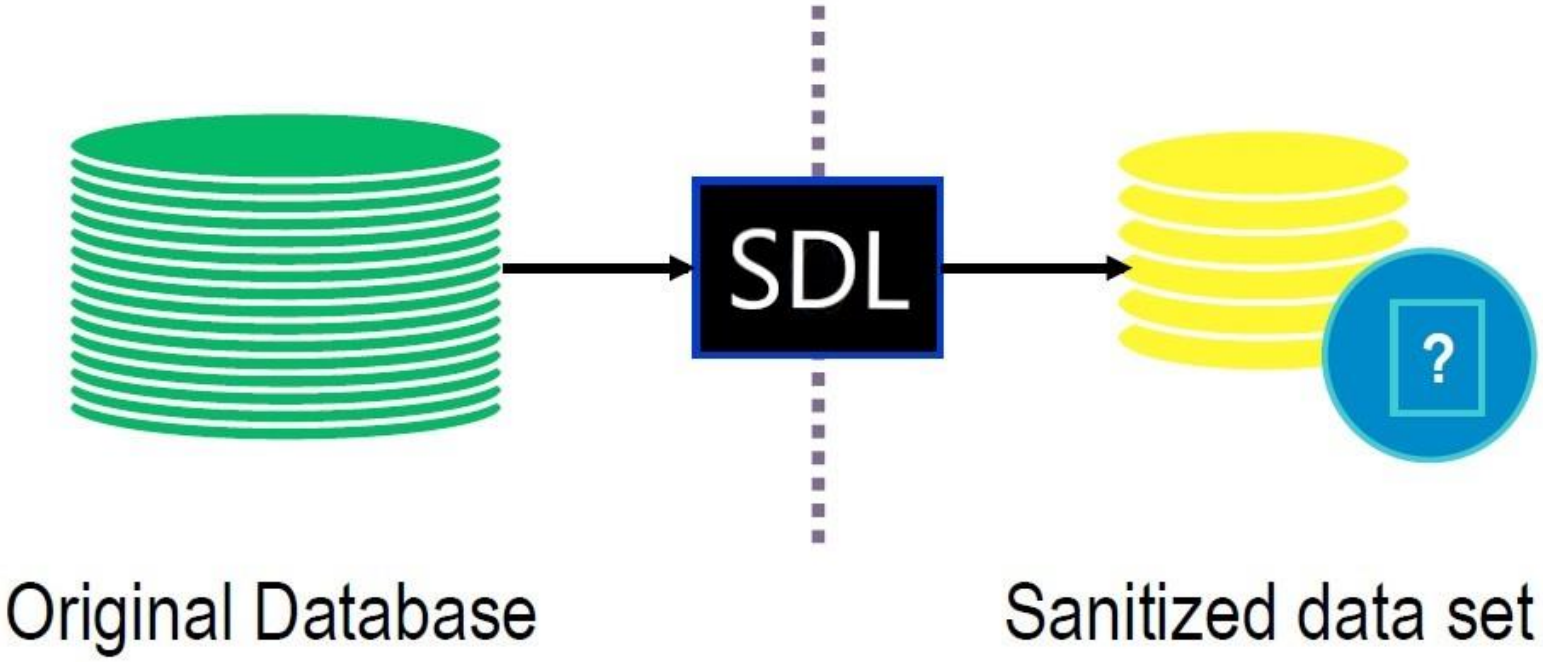
Ian M. Schmutte and Lars Vilhuber

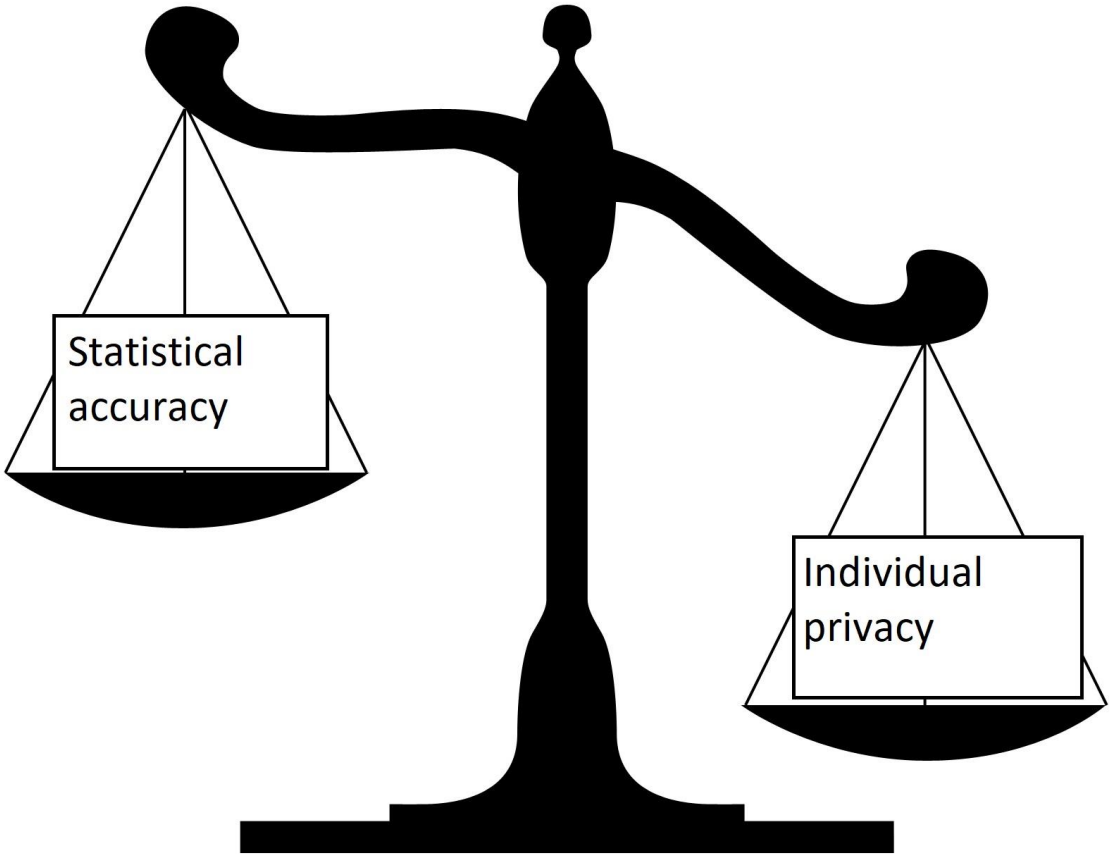
The demands on data

Knowledge about the world

Privacy for individuals

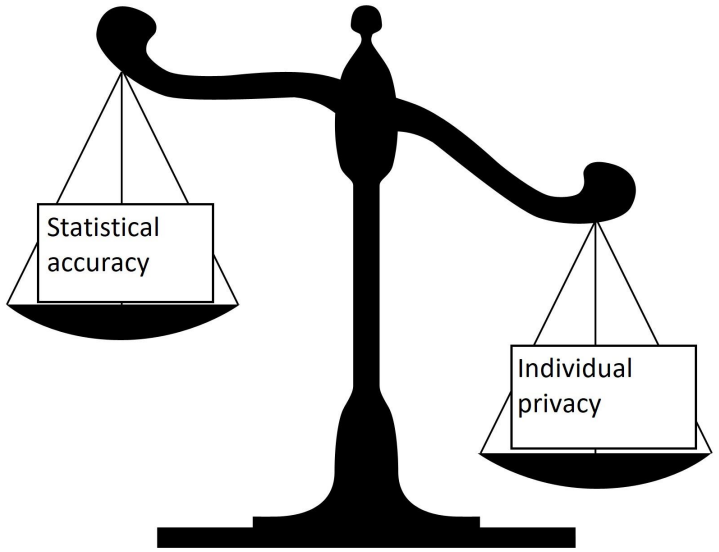
Statistical Disclosure Limitation




Statistical accuracy
Individual privacy

Factors in SDL design

1. What are the privacy requirements?
2. What analyses need to be supported?
3. Is SDL part of a broader system?



Privacy concepts

Anne: A survey participant who responds that she is a smoker

Privacy:

the right to not answer questions about smoking

Confidentiality:

the right to not have answers used against her

Identity Disclosure

Data include

- Zip code
- Gender
- Smoking status

Identity Disclosure

- Attacker knows Anne was in the study
- Only one woman in her zip code in the data.
- Now knows Anne's smoking status

Attribute Disclosure

- Attacker knows Anne was in the study
- Learns all respondents in her zip code are smokers
- Now knows Anne is a smoker

Inferential Disclosure

- Attacker knows Anne was in the study
- 99 of 100 female respondents in her zip code smoke
- Now knows Anne is *probably* a smoker

K-anonymity and l-diversity

Dataset is *k-anonymous* if, for any combination of attributes, at least *k* records have that combination

- Reduces risk of “singling out”
- Does not prevent attribute disclosure

l-diversity ensures that within each group, there is “sufficient” heterogeneity in sensitive attributes

SDL Methods

De-identification

HIPAA defines 16 identifiers to remove

- J-PAL for Stata (stata_PII_scan) and R (PII-scan)
- Innovations for Poverty Action for Python or Windows (PII_detection)

Necessary, but not sufficient

Ignorable

Coarsening

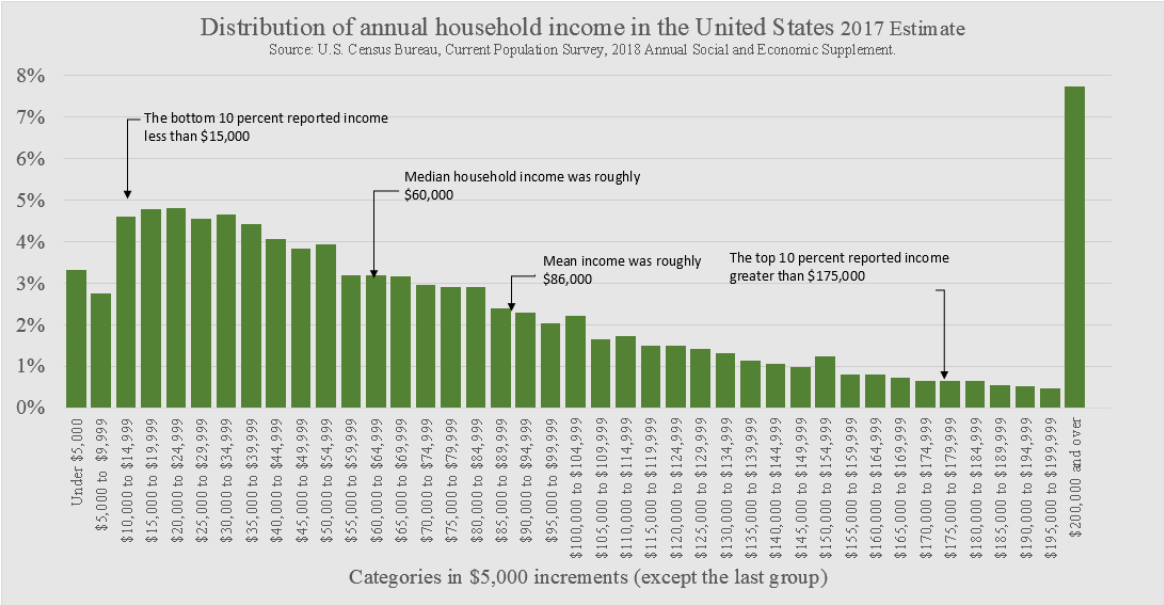
Collapse or coarsen variables that “single out” individual records

Used in combination with *k-anonymity*

Examples:

- Public use microdata areas in the American Community Survey
- Topcoding income in the Current Population Survey
- Reporting age, income in bins
- Removal of detailed geographies, like state

Topcoding



Ignorable
for inference on quantiles below topcode (e.g. 90-10 ratio in CPS)

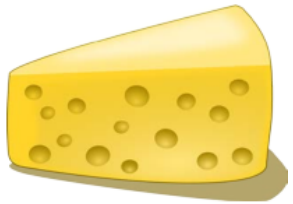
Non-ignorable
for quantiles above topcode

Cell Suppression

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

A	B	*	D
E	F	G	H
*	J	K	L
M	N	O	*

*	B	*	*
E	F	G	H
*	J	*	L
M	N	*	*



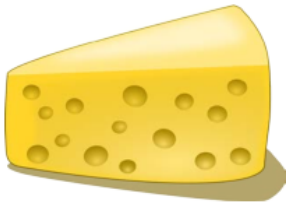
- "Blank out" cells to protect outliers
  - i.e., where one large firm dominates
- Then "blank out" more cells to prevent subtraction attack
- e.g., Economic Census, County Business Patterns

Cell Suppression

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

A	B	*	D
E	F	G	H
*	J	K	L
M	N	O	*

*	B	*	*
E	F	G	H
*	J	*	L
M	N	*	*



Not ignorable unless

…suppression was random with respect to your estimand of interest

…or you really only care about the unsuppressed data.

So then what?

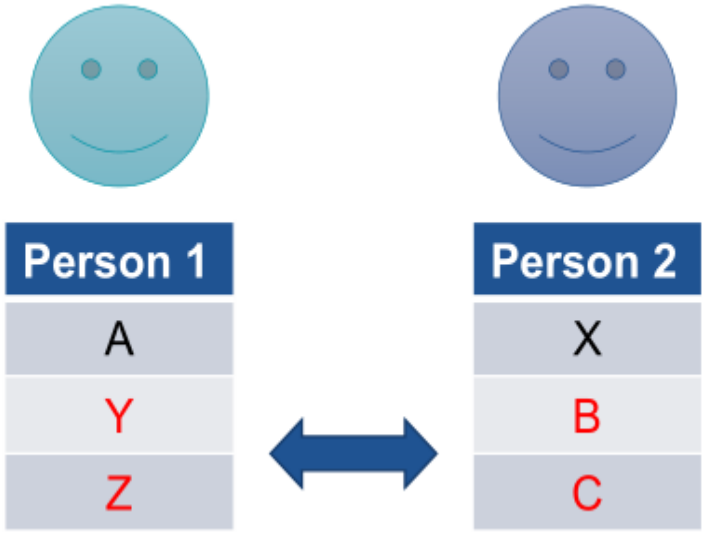
Swapping

High-risk records:

- Matched to a “nearby” record
- .. And swapped

Preserves counts on key characteristics

May prevent disclosure of sensitive attributes



Swapping

Ignorable if..

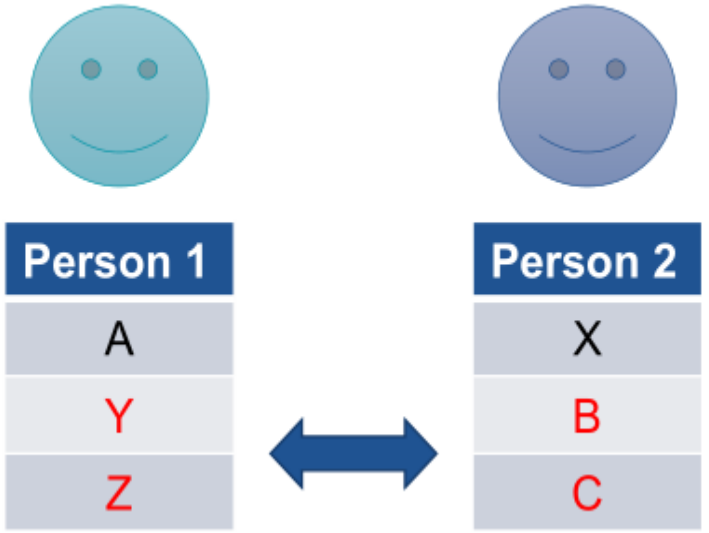
only care about matching variables

Non-ignorable for

covariance between matching and other variables

Parameters are secret

- Swap rate
- Sensitive chars
- Swap domain
- Etc.



Noise Infusion

Add randomly distributed noise to each unit

Add up the distorted units

Noise averages out in larger cells

Ignorable for means;
Non-ignorable for variances

14	41	50	58	65
15	24	26	30	25
52	53	66	47	51
68	6	44	17	32
38	26	33	42	64

$p_{Z|Y}(Z|Y,\theta_M)$

13	41	51	58	65
15	24	25	30	24
51	54	66	48	51
68	6	44	16	32
38	25	33	42	65

Why should you read this chapter?

- Further description of methods
- Links to tools and the broader literature
- Connections between SDL and formal privacy

Thank You!

Ian M. Schmutte
http://ianschmutte.org
schmutte@uga.edu