Uses and Applications of Identified vs. De-identified Data

and how to properly de-identify it!

Central Texas FERPA Conference
Austin, TX
April 12, 2013

Michael Hawes
Statistical Privacy Advisor
U.S. Department of Education

Presentation Overview

- Definitions
- Appropriate Uses of Raw, Redated, and De-Identified Data
- Disclosure Avoidance Primer
- Questions and Discussion

Confidentiality under FERPA

- Protects personally identifiable information (PII) from education records from unauthorized disclosure
- Requirement for written consent before sharing PII
- Exceptions from the consent requirement for:
  - "Studies"
  - "Audits and Evaluations"
  - Health and Safety emergencies
  - And other purposes & parties as specified in 34 CFR §99.31

Personally Identifiable Information (PII)

- Name
- Name of parents or other family members
- Address
- Personal identifier (e.g., SSN, Student ID#)
- Other indirect identifiers (e.g., date or place of birth)
- *"Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty." (§ 99.3)*

De-Identification

Can't I just de-identify the data by removing the names, SSNs, etc.?

Particularly at lower levels of geography, individuals are often easily identified by combinations of indirect identifiers and other demographic or outcome data

Aggregation

PII? But I'm only going to be using aggregate data tables…

Aggregate data tables can still contain PII if they report information on small groups, or individuals with unique or uncommon characteristics

Two Examples

Sunshine Elementary, Anywhere, USA 3rd Grade Class	FRPL Eligible	Not FRPL Eligible
White	12	16
Asian	4	3
Hispanic	6	8
African-American	2	5
Native American	1	0

Sunshine Elementary, Anywhere, USA 4th Grade Class	FRPL Eligible	Not FRPL Eligible
Male	61	0
Female	27	35

Small cells increase disclosure risk…

BUT, suppressing the small cells

may not be sufficient

Example: Suppression

Subgroup	Number of Students	Percent Proficient
American Indian	***	***
Asian	15	87.7%
Black	12	91.7%
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
Female	45	84.4%
Male	41	78.0%

Example: Suppression

Subgroup	Number of Students	Percent Proficient
American Indian	*** (1 student)	***
Asian	15	87.7%
Black	12	91.7%
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
Female	45	84.4%
Male	41	78.0%

15 + 12 + 21 + 13 + 24 = 85
45 + 41 = 86
86 – 85 = 1

Example: Suppression

Subgroup	Number of Students	Percent Proficient
American Indian	*** (1 student)	***
Asian	15 (13 proficient)	87.7%
Black	12 (11 proficient)	91.7%
Hispanic	21 (17 proficient)	81.0%
Two or More Races	13 (10 proficient)	76.9%
White	24 (19 proficient)	79.2%
Female	45 (38 proficient)	84.4%
Male	41 (32 proficient)	78.0%

87.7% = 13/15
91.7% = 11/12
etc.

Example: Suppression

Subgroup	Number of Students	Percent Proficient
American Indian	*** (1 student) (0 proficient)	0.0%
Asian	15 (13 proficient)	87.7%
Black	12 (11 proficient)	91.7%
Hispanic	21 (17 proficient)	81.0%
Two or More Races	13 (10 proficient)	76.9%
White	24 (19 proficient)	79.2%
Female	45 (38 proficient)	84.4%
Male	41 (32 proficient)	78.0%

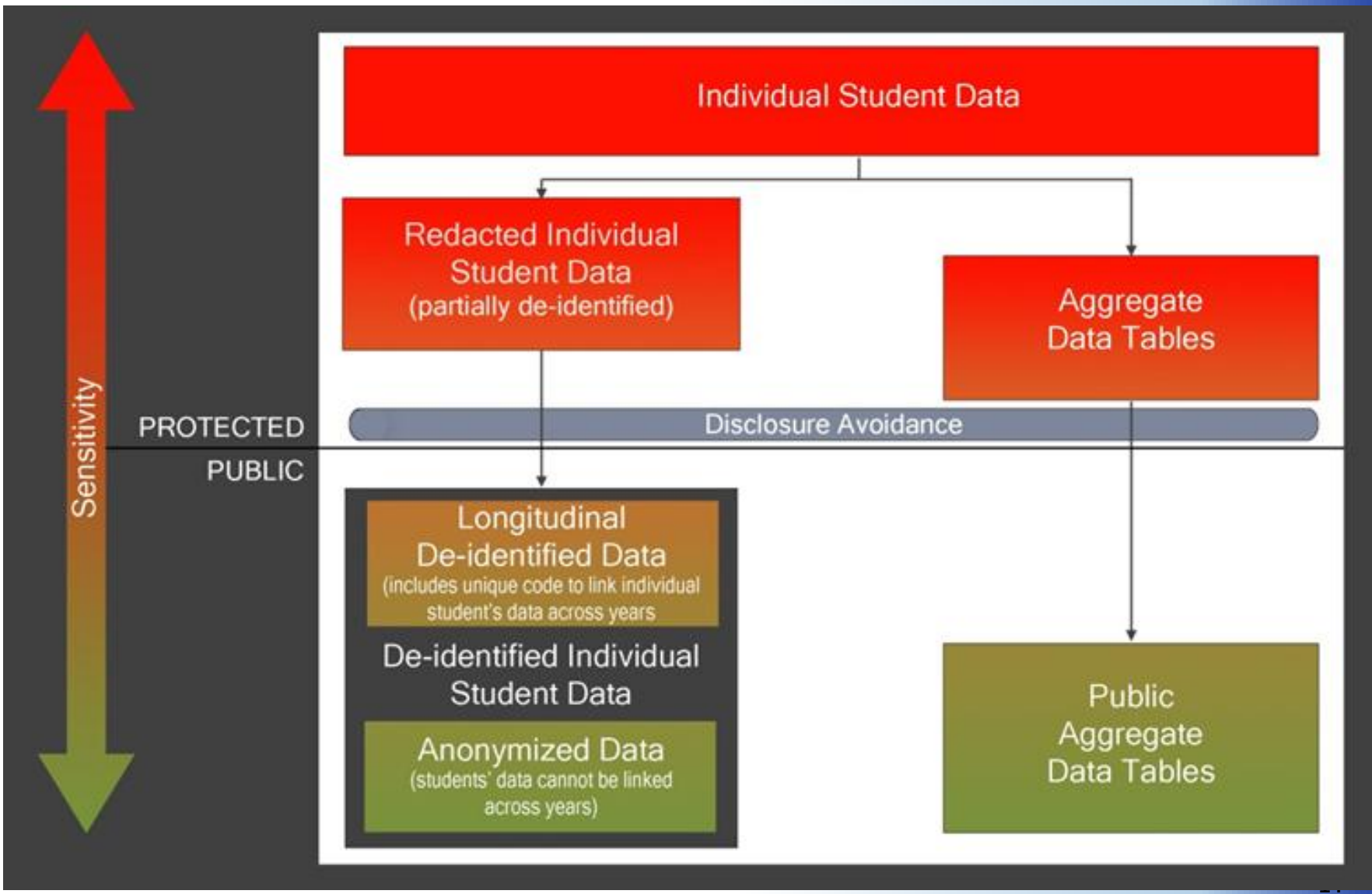
13 + 11 + 17 + 10 + 19 = 70
38 + 32 = 70
70 – 70 = 0

Levels of Protection

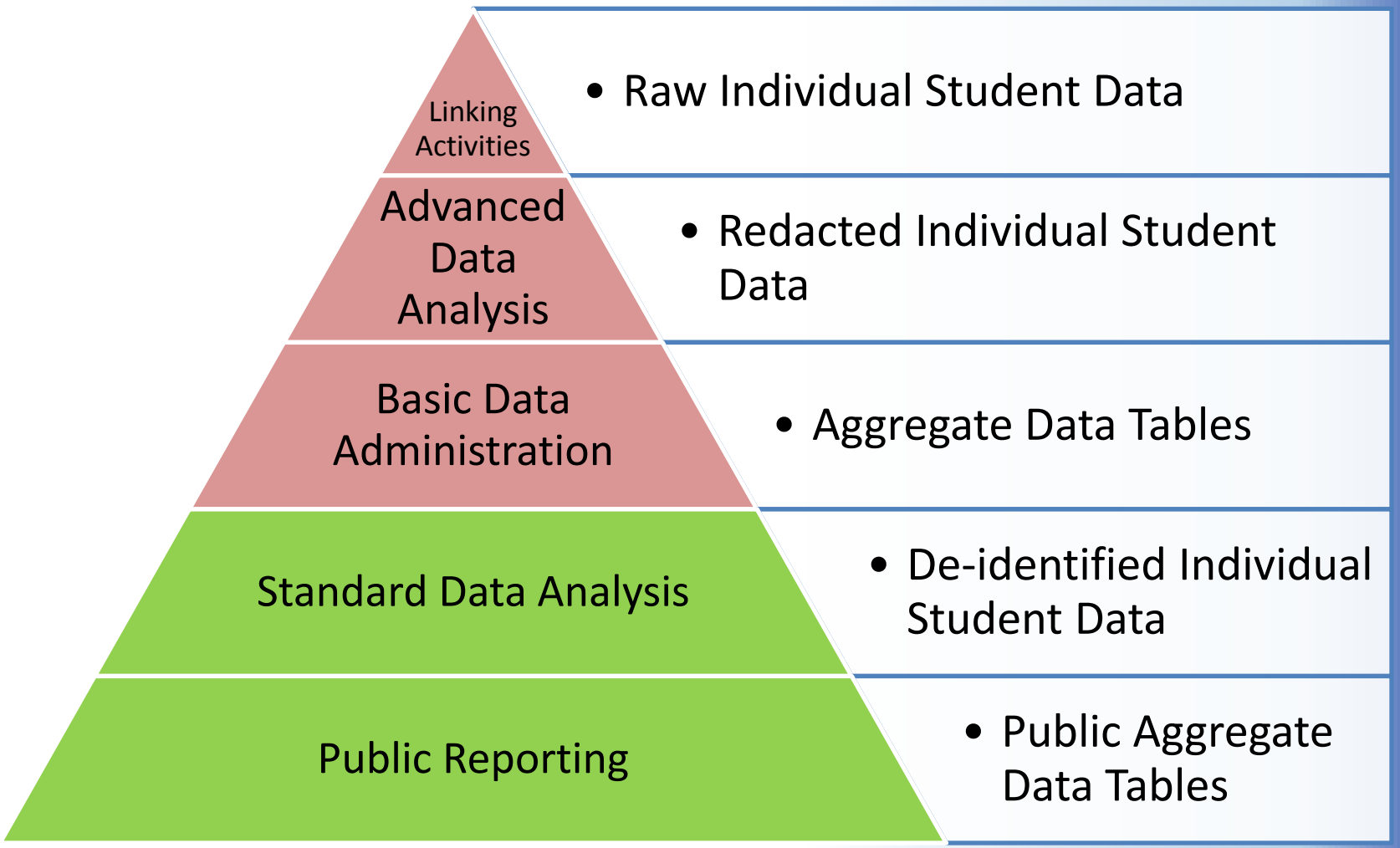
Legal protections (e.g., FERPA) are typically binary: Data is either protected or it is not.

Best practices, however, suggest using a spectrum of privacy protections with access to data of varying sensitivies controlled according to specific role-based and use-based needs.

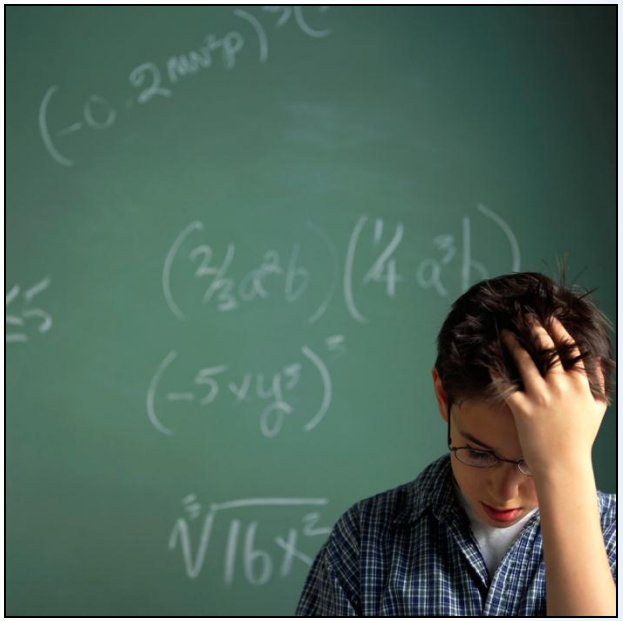
Fully Identified, Redacted, Aggregated, and De-identified Data



Samply Hierarchy of Access



Disclosure Avoidance Primer



(aren't you glad you had coffee this morning?)

It's all about risk

"The release of any data usually entails at least some element of risk. A decision to eliminate all risk of disclosure would curtail [data] releases drastically, if not completely. Thus, for any proposed release of [data] the acceptability of the level of risk of disclosure must be evaluated."

Federal Committee on Statistical Methodology, "Statistical Working Paper #2"

3 Basic Flavors of Disclosure Avoidance

- Suppression
- Blurring
- Perturbation

Suppression

Definition:	Removing data to prevent the identification of individuals in small cells or with unique characteristics
Examples:	• Cell Suppression • Row Suppression • Sampling
Effect on Data Utility:	• Results in very little data being produced for small populations • Requires suppression of additional, non-sensitive data (e.g., complimentary suppression)
Residual Risk of Disclosure:	• Suppression can be difficult to perform correctly (especially for large multi-dimensional tables) • If additional data is available elsewhere, the suppressed data may be re-calculated.

Blurring

Definition:	Reducing the precision of data that is presented to reduce the certainty of identification
Examples:	• Aggregation • Percents • Ranges • Top/Bottom-Coding • Rounding
Effect on Data Utility:	• Users cannot make inferences about small changes in the data • Reduces the ability to perform time-series or cross-case analysis
Residual Risk of Disclosure:	• Generally low risk, but if row/column totals are published (or available elsewhere) then it may be possible to calculate the actual values of sensitive cells

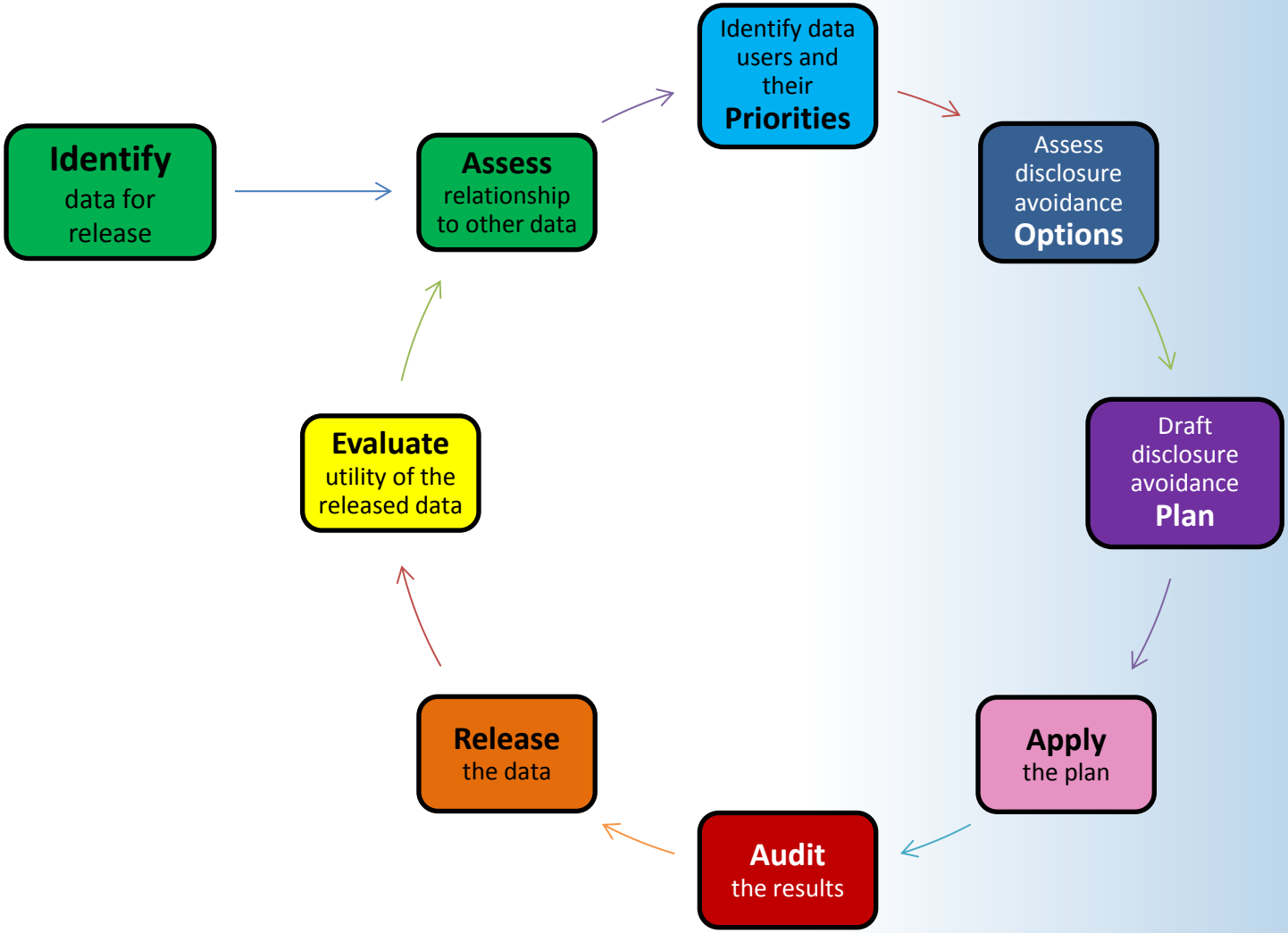
Perturbation

Definition:	Making small changes to the data to prevent identification of individuals from unique or rare characteristics
Examples:	• Data Swapping • Noise • Synthetic Data
Effect on Data Utility:	• Can minimize loss of utility compared to other methods • May affect public credibility of the data; transparency and messaging is key
Residual Risk of Disclosure:	• If someone has access to some (e.g., a single state's) original data, they may be able to reverse-engineer the perturbation rules used to alter the rest of the data

Some tips to consider:

- You don't have to limit your plan to a single method – you can adopt multiple methods that compliment each other (e.g., suppression and top/bottom coding)
- If using suppression, be especially aware of row/column totals, and related tables – complimentary suppression will most likely be necessary
- When reporting in percentages, round to whole numbers whenever possible
- Be sure to audit your results

Disclosure Avoidance Lifecycle



Take Home Points on Disclosure Avoidance

- There is no single “right” way to perform disclosure avoidance
- Different methods affect the utility of the data differently – Know your users and what they want!
- Be aware of how other available data may impact your choice of disclosure avoidance method(s)
- Be aware that the method(s) you adopt for one release may affect how you can release related data
- Disclosure avoidance doesn’t end when you release the data – periodically re-evaluate re-identification risk and data utility
- Whenever possible, educate your users about the disclosure avoidance methods that have been applied

Questions and Discussion

Michael Hawes
Statistical Privacy Advisor
U.S. Department of Education
Michael.Hawes@ed.gov
(202) 453-7017