Disclosure Avoidance:

The Good, The Bad, and the Ugly

PTAC CT Site Visit
August 20, 2014
Cromwell, CT

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

Which of the Following are NOT considered PII?

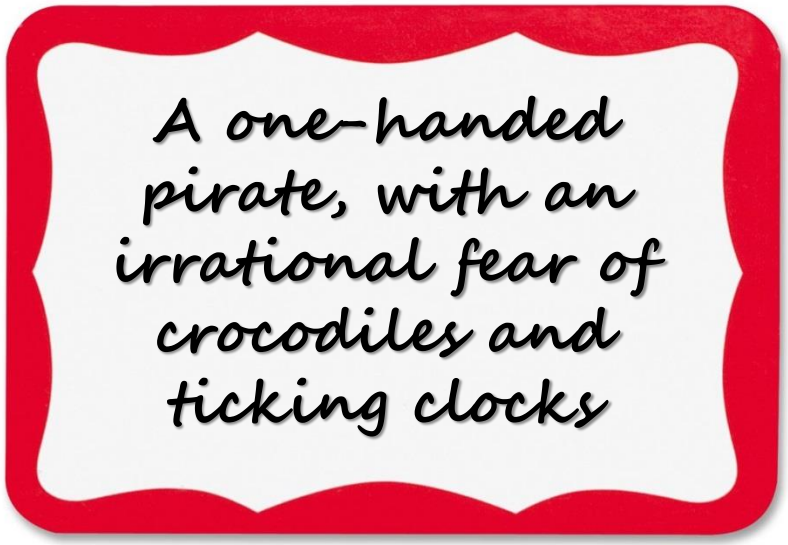
- *Name*
- *Social Security Number*
- *Address*
- *Month of Birth*
- *Telephone Number*
- *Shoe Size*
- *Job Title*
- *Email Address*
- *Office Number*
- *Racial/Ethnic Group*
- *Pet's Name*
- *Criminal Record*
- *School Attended*
- *1st Grade Teacher*
- *License Plate*
- *Mother's Maiden Name*
- *Bank Account Number*
- *Favorite Movie*
- *Performance Rating*
- *Grades*
- *Test Scores*

PII is:

Personal Information

PII is:

Personally Identifiable Information



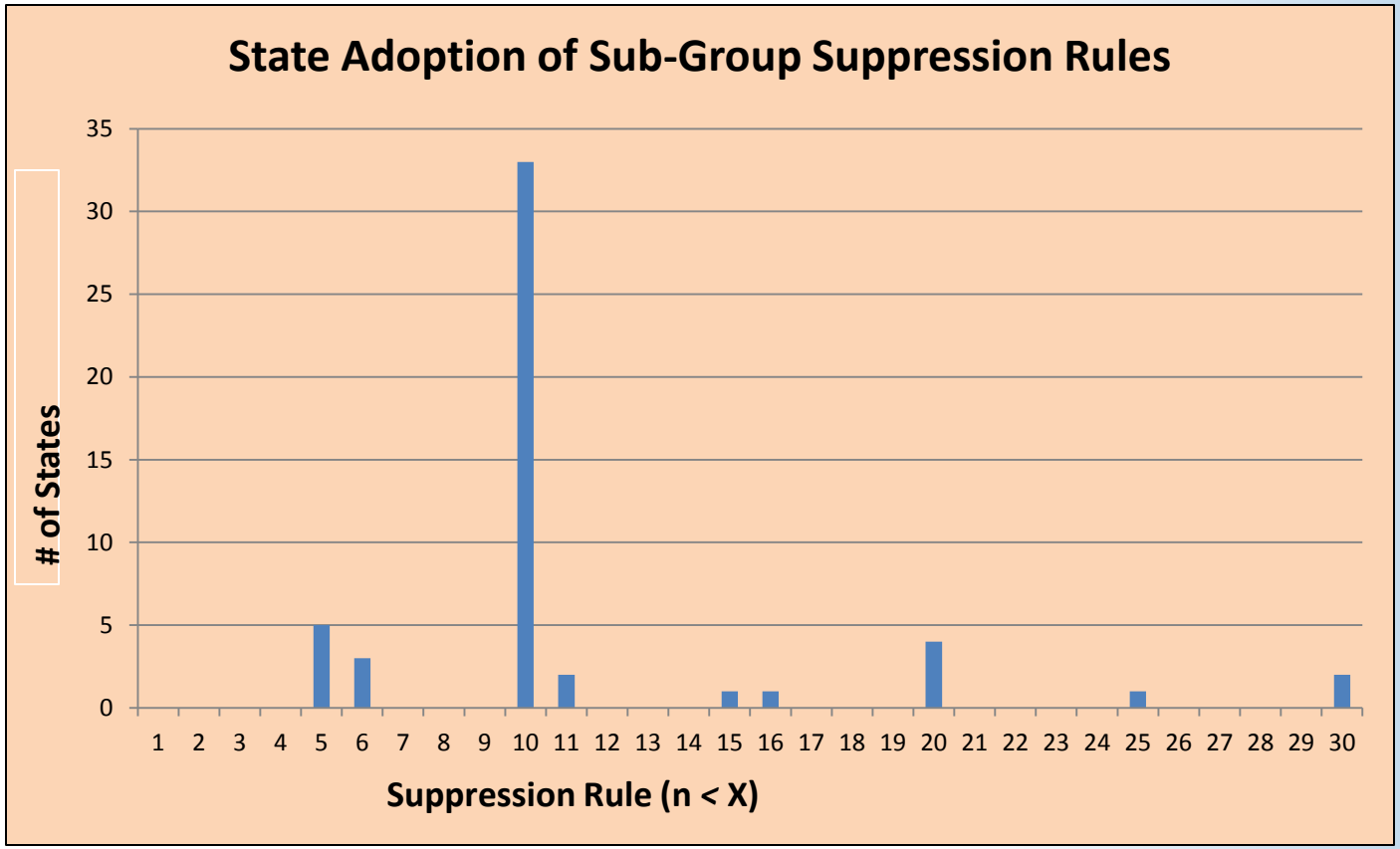
Personally Identifiable Information (PII) under FERPA

- 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)*

PII? But I'm only releasing aggregate data tables…

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

How States are Doing It



Small cells increase disclosure risk…

BUT, suppressing the small cells

may not be sufficient

Common Mistakes in Public Reporting

Population Size vs. Cell Size

End of Grade (Mathematics) Grade 7
Number and Percent of Students At or Above Achievement Level III in Mathematics
Students Taking All Tests

CENSORED

Student Subgroup	2011-2012 # At or Above Level III	2011-2012 # Valid Scores	2011-2012 Percent At or Above Level III	2012-2013 # At or Above Level III	2012-2013 # Valid Scores	2012-2013 Percent At or Above Level III
Students With Disabilities	19	41	46.3%	-	46	<5%
Non-Disabled Students	139	183	76.0%	40	195	20.5%
Academically Gifted	-	51	>95%	32	48	66.7%
Academically Gifted Math	-	45	>95%	29	42	69.0%
Academically Gifted Reading	-	43	>95%	21	31	67.7%
Autistic	*	*	*	1	6	16.7%

What's the missing number?

12

8

14

?

6

What's the missing number?

12

8

14

?

6

44

What's the missing number?

12

8

14

4

6

44

What's the missing number?

12

8

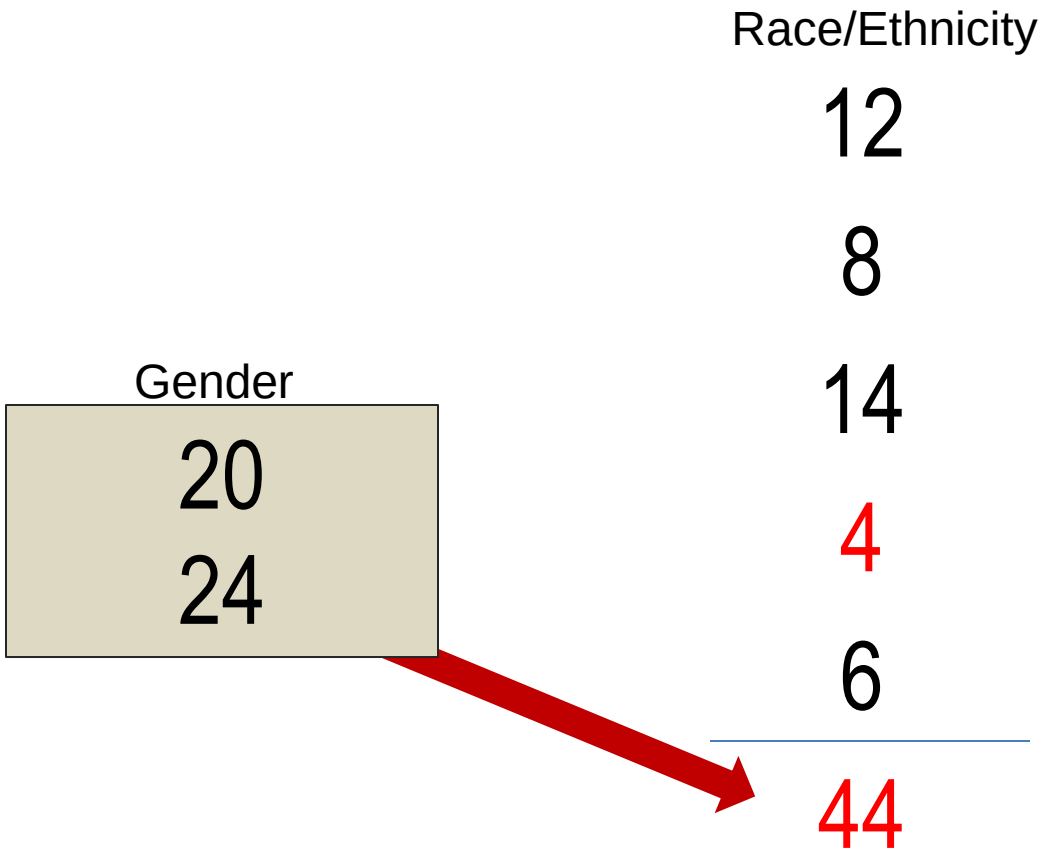
14

CENSORED

6

CENSORED

What's the missing number?



Lack of Complementary Suppression

	GR						
	School						
	Stud. Incl	Part. Rate	% at Each Level				
Student Group	#	%	A	P	NI	W	
Accountability Subgroups							
Students w/disabilities	4	-	-	-	-	-	
ELL and Former ELL		-	-	-	-	-	
Low income	3	-	-	-	-	-	
High needs	7	-	-	-	-	-	
Afr. Amer./Black		-	-	-	-	-	
Amer. Ind. or Alaska Nat.		-	-	-	-	-	
Asian		-	-	-	-	-	
Hispanic/Latino	1	-	-	-	-	-	
Multi-race, Non-Hisp./Lat.		-	-	-	-	-	
Nat. Haw. or Pacif. Isl.		-	-	-	-	-	
White	11	100	0	45	36	18	8
Other Subgroups							
Male	5	-	-	-	-	-	
Female	7	-	-	-	-	-	
Title1	4	-	-	-	-	-	
Non-Title1	8	-	-	-	-	-	
Non-Low Income	9	-	-	-	-	-	
ELL		-	-	-	-	-	
Former ELL		-	-	-	-	-	
1st Year ELL		-	-	-	-	-	
Ever ELL		-	-	-	-	-	
All Students							
2013	12	100	0	42	42	17	8
2012	15	100	7	27	53	13	7

Lack of Complementary Suppression

GR							
	School						
	Stud. Incl	Part. Rate	% at Each Level				
Student Group	#	%	A	P	NI	W	
Accountability Subgroups							
Students w/disabilities	4	-	-	-	-	-	
ELL and Former ELL		-	-	-	-	-	
Low income	3	-	-	-	-	-	
High needs	7	-	-	-	-	-	
Afr. Amer./Black		-	-	-	-	-	
Amer. Ind. or Alaska Nat.		-	-	-	-	-	
Asian		-	-	-	-	-	
Hispanic/Latino	1	-	-	-	-	-	
Multi-race, Non-Hisp./Lat.		-	-	-	-	-	
Nat. Haw. or Pacif. Isl.		-	-	-	-	-	
White	11	100	0	45	36	18	8
Other Subgroups							
Male	5	-	-	-	-	-	
Female	7	-	-	-	-	-	
Title1	4	-	-	-	-	-	
Non-Title1	8	-	-	-	-	-	
Non-Low Income	9	-	-	-	-	-	
ELL		-	-	-	-	-	
Former ELL		-	-	-	-	-	
1st Year ELL		-	-	-	-	-	
Ever ELL		-	-	-	-		
All Students							
2013	12	100	0	42	42	17	8
2012	15	100	7	27	53	13	7

Lack of Complementary Suppression

GR							
	School						
	Stud. Incl	Part. Rate	% at Each Level				
Student Group	#	%	A	P	NI	W	
Accountability Subgroups							
Students w/disabilities	4	-	-	-	-	-	
ELL and Former ELL		-	-	-	-	-	
Low income	3	-	-	-	-	-	
High needs	7	-	-	-	-	-	
Afr. Amer./Black		-	-	-	-	-	
Amer. Ind. or Alaska Nat.		-	-	-	-	-	
Asian		-	-	-	-	-	
Hispanic/Latino	1	-	-	-	-	-	
Multi-race, Non-Hisp./Lat.		-	-	-	-	-	
Nat. Haw. or Pacif. Isl.		-	-	-	-	-	
White	11	100	0	45	36	18	8
Other Subgroups							
Male	5	-	-	-	-	-	
Female	7	-	-	-	-	-	
Title1	4	-	-	-	-	-	
Non-Title1	8	-	-	-	-	-	
Non-Low Income	9	-	-	-	-	-	
ELL		-	-	-	-	-	
Former ELL		-	-	-	-	-	
1st Year ELL		-	-	-	-	-	
Ever ELL		-	-	-	-		
All Students							
2013	12	100	0	42	42	17	8
2012	15	100	7	27	53	13	7

Fixed Top/Bottom Coding Thresholds

End of Grade (Mathematics) Grade 3
Number and Percent of Students At or Above Achievement Level III in Mathematics
Students Taking All Tests

CENSORED

Student Subgroup	2011-2012 # At or Above Level III	2011-2012 # Valid Scores	2011-2012 Percent At or Above Level III	2012-2013 # At or Above Level III	2012-2013 # Valid Scores	2012-2013 Percent At or Above Level III
Students With Disabilities	6	15	40.0%	-	8	<5%
Non-Disabled Students	89	115	77.4%	20	76	26.3%
Academically Gifted	-	11	>95%	9	11	81.8%
Academically Gifted Math	*	*	*	-	8	>95%

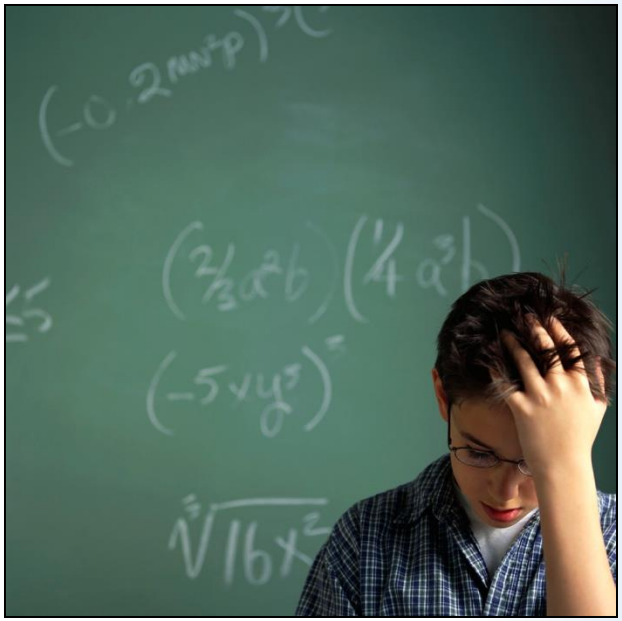
The Trouble with Cell Size Rules

Remember: It's not just the small cells that are important.

Bigger cells/values can still be disclosive if:

- they are extreme values *(e.g., ~0% or ~100% of students in a group)*, or
- they can be used to calculate the values of protected cells elsewhere *(in the same table, or even in another data release!)*

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

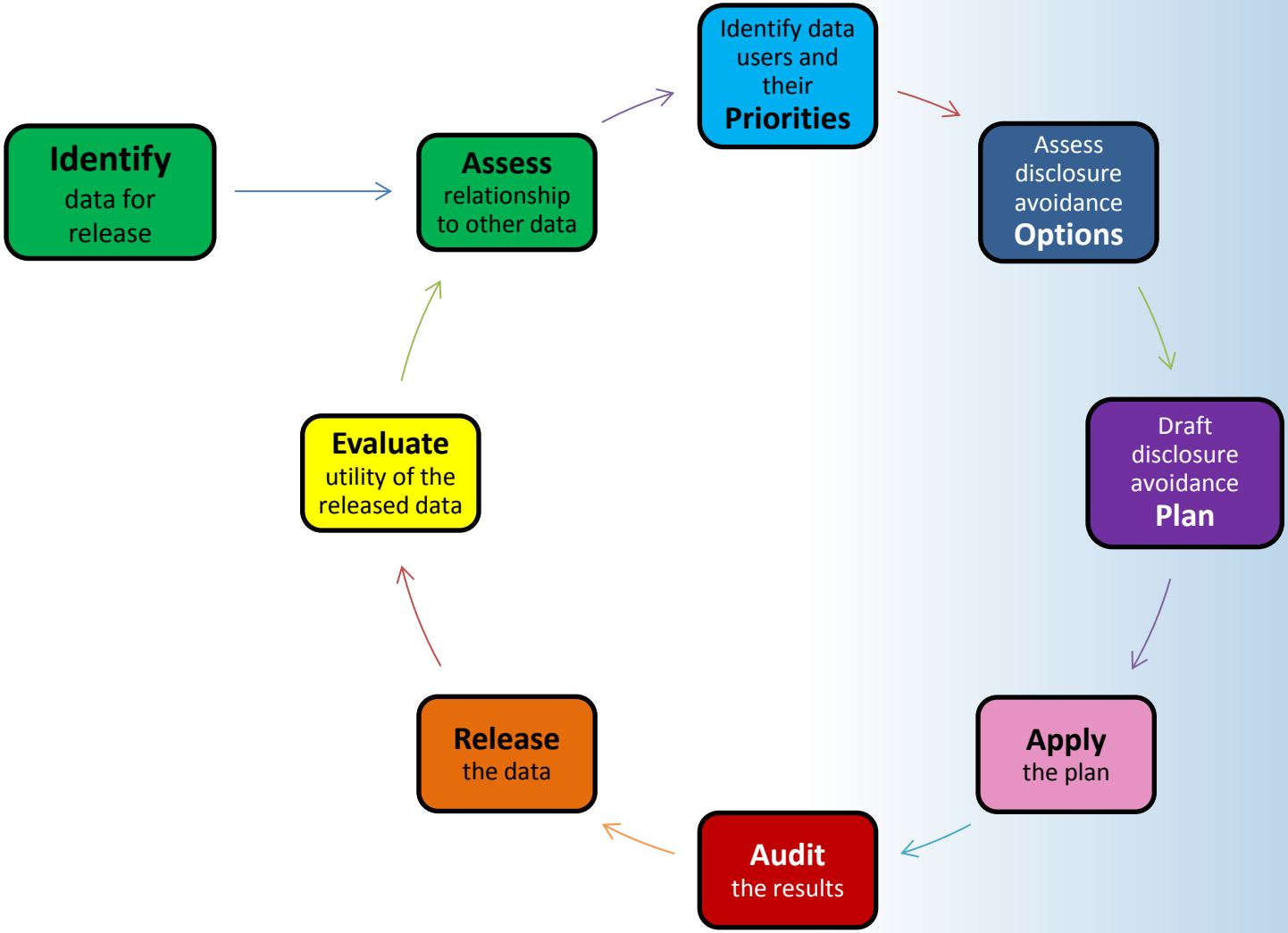
Perturbation

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 • Seen as inappropriate for program data because it reduces the transparency and credibility of the data, which can have enforcement and regulatory implications
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



Questions and Discussion

Michael Hawes

Statistical Privacy Advisor

U.S. Department of Education

Michael.Hawes@ed.gov

(202) 453-7017