Disclosure Avoidance: Protecting Student Privacy in Public Reports

METIS, July 17-19 2019

CENSORED

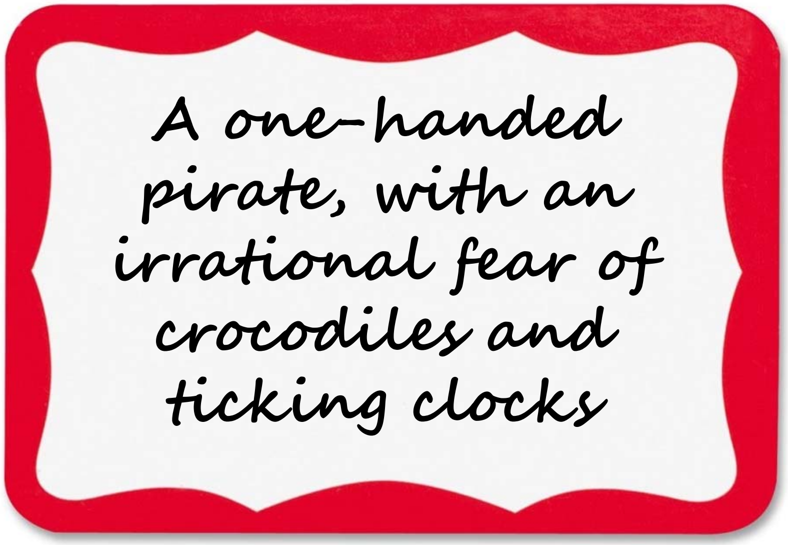
Privacy Technical Assistance Center

What is PII?

Personal Information

Captain
Hook

Personally Identifiable Information



A one-handed pirate, with an irrational fear of crocodiles and ticking clocks

PII

Name	Race /Ethnicity	Gender	Pirate Status	# of Hands	GPA
[redacted]	W	M	Y	1	2.0
[redacted]	A	F	N	2	3.5
[redacted]	B	M	N	2	3.8
[redacted]	W	F	N	2	2.8
[redacted]	H	M	N	2	3.3

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*

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 <u>linked or linkable</u> to a specific student that would allow a <u>reasonable person in the school community</u>, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty." (§ 99.3)*

Personally Identifiable Information (PII)

- **Direct Identifiers**
  - e.g., Name, SSN, Student ID Number, etc.

  (1:1 relationship to student)

- **Indirect Identifiers**
  - e.g., Birthdate, Demographic Information

  (1:Many relationship to student)

- *"**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)*

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

# of Students Proficient or Advanced on State Mathematics Assessment

Gender	Below Proficient	Above Proficient
Male	3,653	24,187
Female	2,947	23,956

# of Students Proficient or Advanced on State Mathematics Assessment

Pirate Status	Below Proficient	Above Proficient
Yes	1	0
No	6,599	48,143

# of Students Proficient or Advanced on State Mathematics Assessment

Pirate Status	Below Proficient	Above Proficient
Yes	*	0
No	6,599	48,143

Disclosure

- ***Disclosure*** means to permit access to or the release, transfer, or other communication of PII by any means. Disclosure can be authorized, such as when a parent or an eligible student gives written consent to share educational records with an authorized party, such as a researcher. Disclosure can also be unauthorized or inadvertent (accidental).

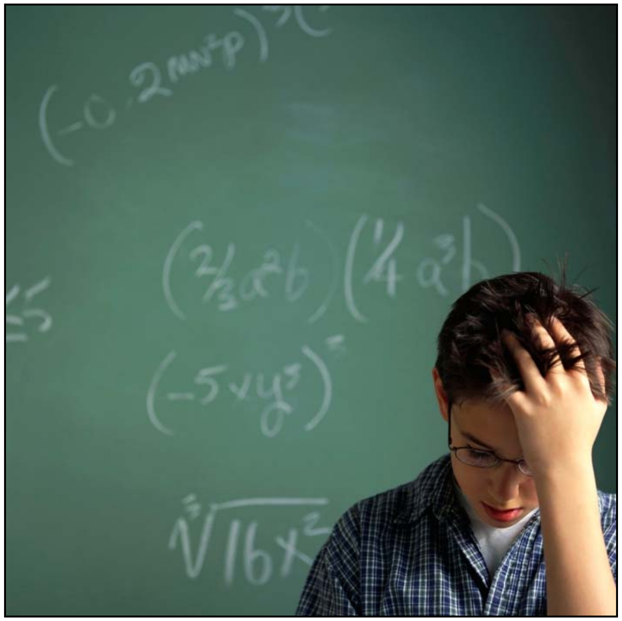
What standard is used to evaluate disclosure risk?

- Can a "**reasonable person**" in the school community who does not have personal knowledge of the relevant circumstances identify an individual in the publicly released data with reasonable certainty?

- The "reasonable person" standard
  - Hypothetical, rational, prudent, average individual in the school community
  - Does not have personal knowledge of the relevant circumstances
  - School officials, including teachers, administrators, coaches, and volunteers, are **not** included

Disclosure Avoidance Primer

- (Should we stop so you can get some coffee?)



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.

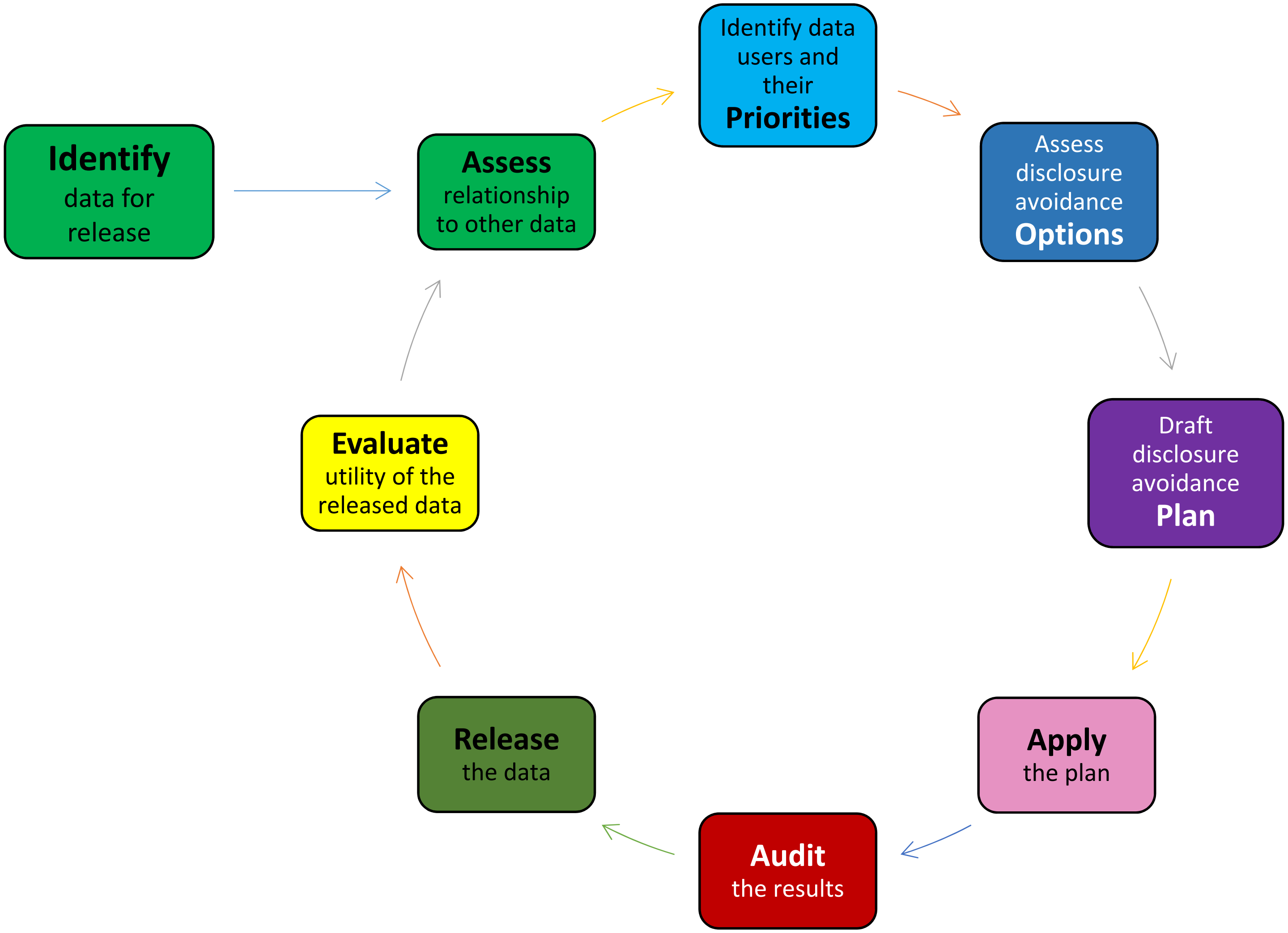
Blurring

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

Disclosure Avoidance Lifecycle



Identify data for release

Assess relationship to other data

Identify data users and their **Priorities**

Assess disclosure avoidance **Options**

Draft disclosure avoidance **Plan**

Apply the plan

Audit the results

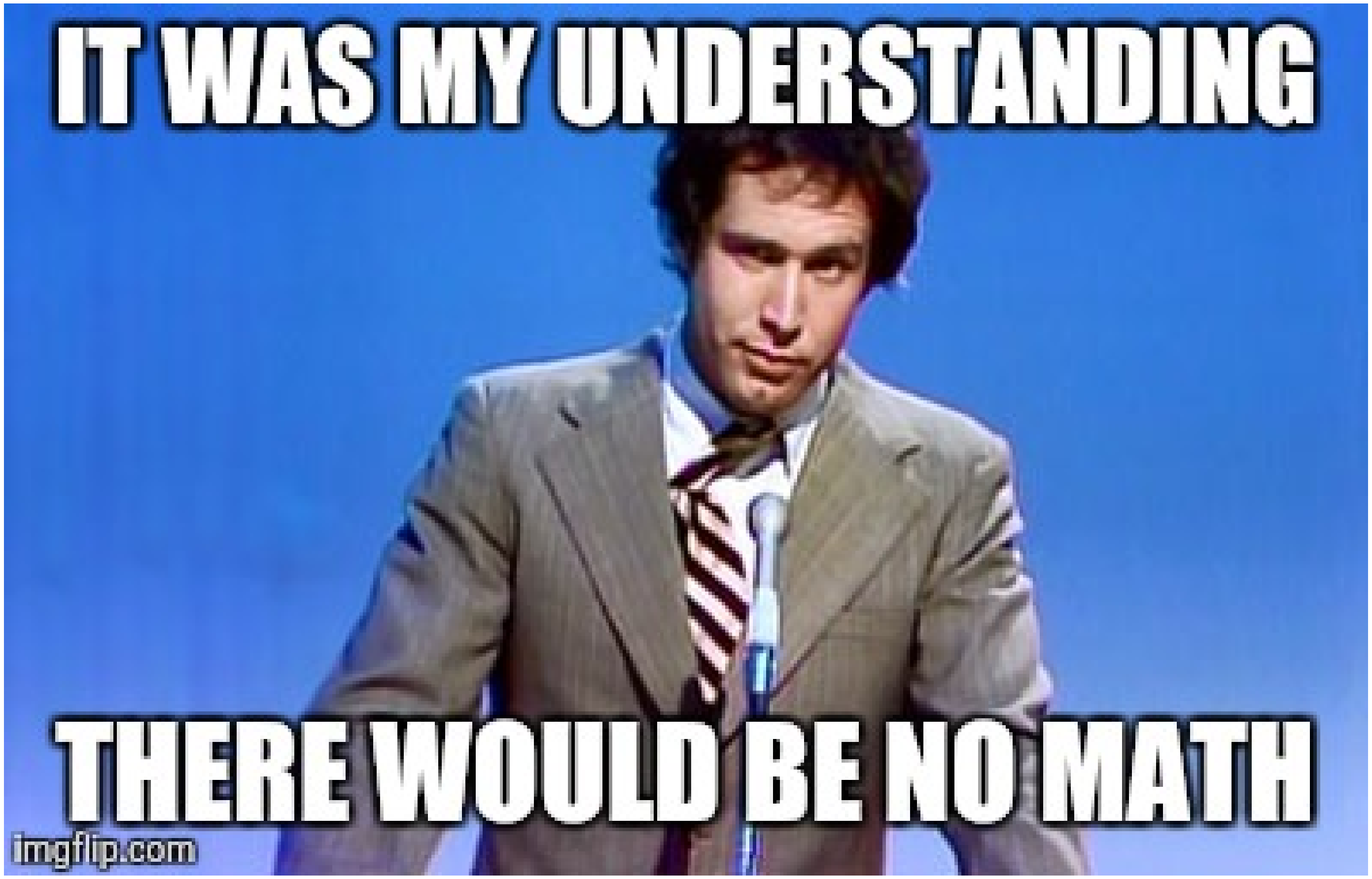
Release the data

Evaluate utility of the released 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 especially careful with individual-level data – you will probably need to use some amount of perturbation!
- Be sure to audit your results

Common Issues in Public Reporting



Population Size vs. Cell Size

Assume a minimum n-size rule of 5:

Subgroup	# Tested	# Proficient	% Proficient
Subgroup 1	6	1	16.7%

Population Size vs. Cell Size

Assume a minimum n-size rule of 5:

Subgroup	# Tested	# Proficient	% Proficient
Subgroup 1	6	1	16.7%

What if I'm that 1 student? I now know something about the other 5!

Fixed Top/Bottom Coding Thresholds

Assume a minimum n-size rule of 5:

Subgroup	# Tested	# Proficient	% Proficient
Subgroup 1	8	*	<5%

Fixed Top/Bottom Coding Thresholds

Assume a minimum n-size rule of 5:

Subgroup	# Tested	# Proficient	% Proficient
Subgroup 1	8	*	<5%

0/8 = 0%
1/8 = 12.5%

So, “<5%” of 8 students = 0 students!

A Better Approach for Handling Extreme Values

Number of Students (denominator)	Top/Bottom Coding for Percentages
1-5	Suppressed
6-15	<50%, ≥50%
16-30	≤20%, ≥80%
31-60	≤10%, ≥90%
61-300	≤5%, ≥95%
301-3,000	≤1%, ≥99%
3,001 or more	≤0.1%, ≥99.9%

What’s the missing number?

12

8

14

?

6

What’s the missing number?

$$\begin{array}{c}12\\8\\14\\?\\6\\\hline 44\end{array}$$

What’s the missing number?

$$\begin{array}{c}12\\8\\14\\4\\6\\\hline 44\end{array}$$

What’s the missing number?

12

8

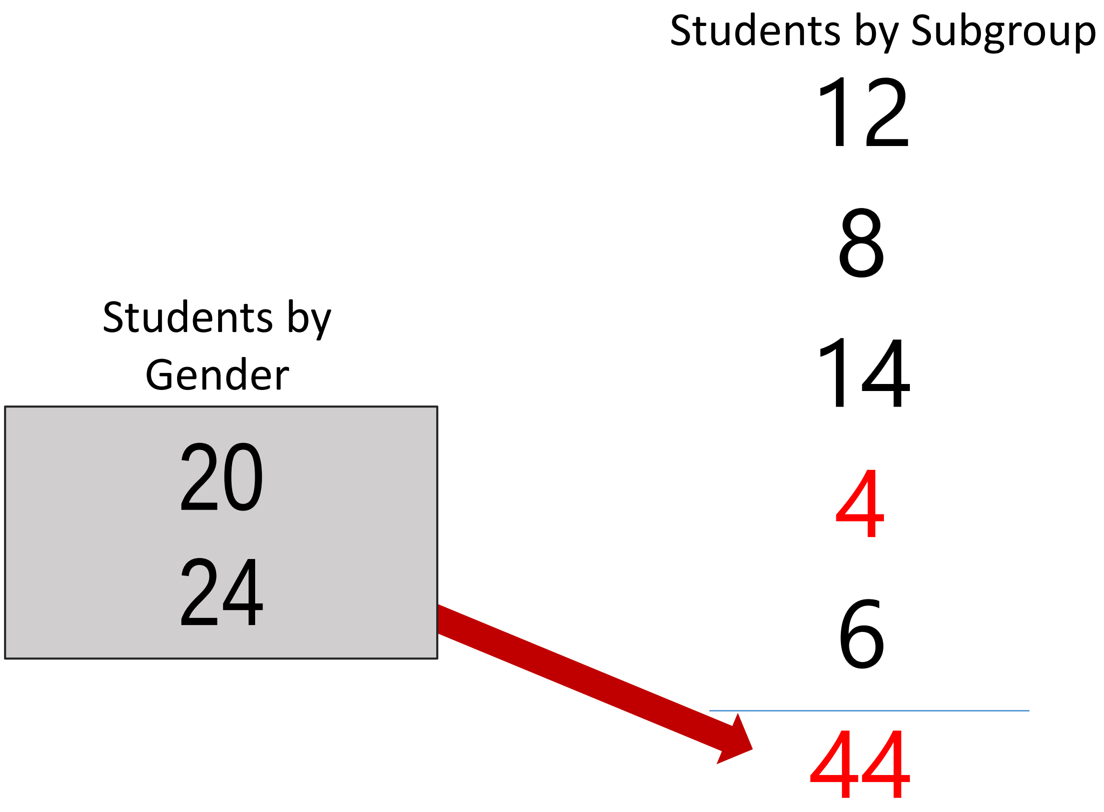
14

CENSORED

6

CENSORED

What’s the missing number?



Students by Subgroup

12

8

14

4

6

44

Students by Gender

20

24

Lack of Complementary Suppression

Subgroup	# Tested	Advanced	Proficient	Basic	Below Basic
Subgroup 1	11	0%	45%	36%	18
Subgroup 2	1	*	*	*	*
All Students	12	0%	42%	42%	17%

Lack of Complementary Suppression

Subgroup	# Tested	Advanced	Proficient	Basic	Below Basic
Subgroup 1	11	0%	45%	36%	18
Subgroup 2	1	*	*	*	*
All Students	12	0%	42%	42%	17%

Lack of Complementary Suppression

Subgroup	# Tested	Advanced	Proficient	Basic	Below Basic
Subgroup 1	11	0%	45%	36%	18
Subgroup 2	1	*	*	*100%*	*
All Students	12	0%	42%	42%	17%

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!)*

Take Home Point:
Consider All Reporting Levels

Education data are often reported in a multi-dimensional structure.

To be effective, a disclosure avoidance methodology must consider all levels of aggregation.

Take Home Point:
Data Releases by Others

When performing a disclosure risk analysis, educational agencies and institutions must consider data releases made by other organizations.

How schools, districts, states, and the Federal government release the same (or related) data, may impact the re-identifiability of the data you (or they) release!

Not All Data are Created Equal

- Disclosure avoidance is about risk assessment and risk mitigation.

- Different types of data carry different levels of reidentification risk, and thus require different approaches to disclosure avoidance.

Data Characteristics to Consider

Aggregate vs. Individual-level Data

- Individual-level Data
  - Snapshot vs. Longitudinal Data
  - Categorical vs. Continuous Measures
- Aggregate Data
  - Attribute vs. Outcome
  - Single metric vs. Composite Index
  - Student Count vs. Incident Count
  - Thresholds vs. Averages

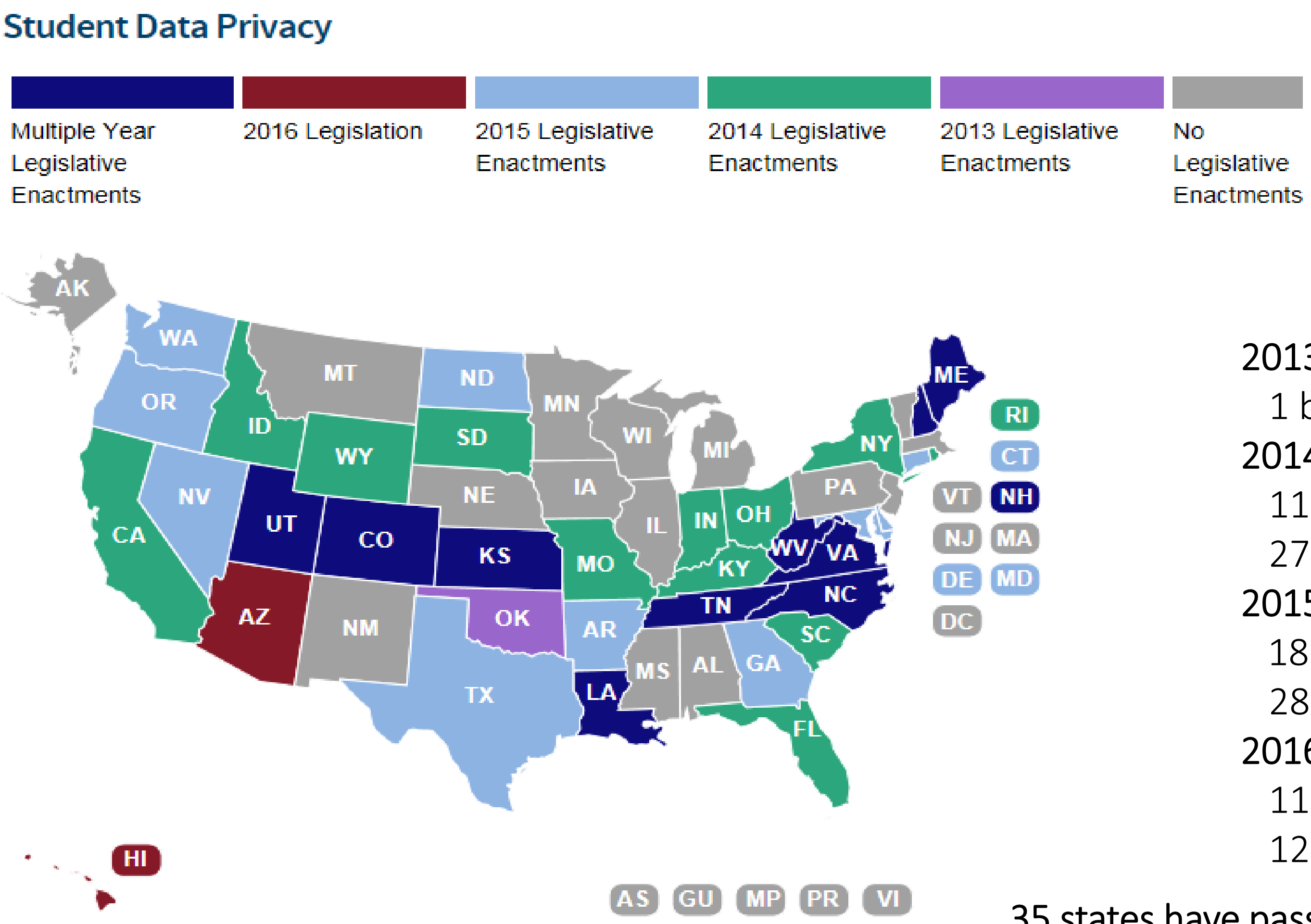
It’s all about risk

“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”

Recent Trends and Challenges

Student Privacy and State Legislation



2013
  1 bill in 1 state

2014
  110 bills in 36 states
  27 passed in 20 states

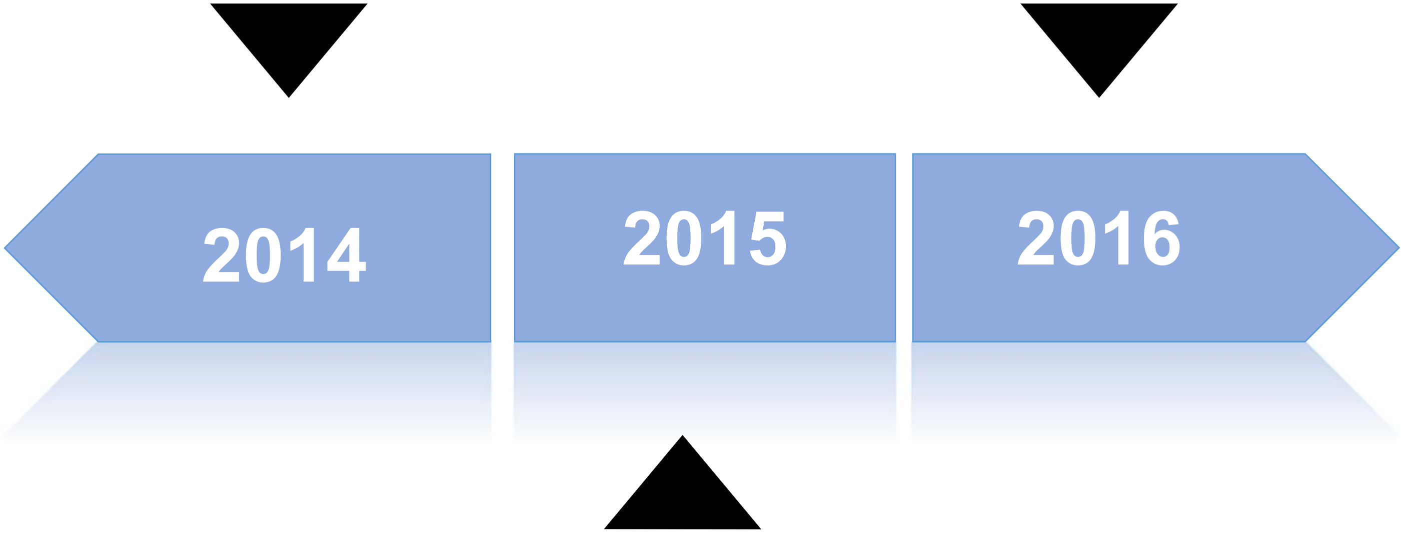
2015
  188 bills in 47 states
  28 passed in 15 states

2016
  111 bills in 34 states
  12 passed in 9 states

35 states have passed 73 laws since 2013

National Conference of State Legislatures June 28, 2016

Trends in Types of Legislation



2014

Banning Collection of Specific Data
Ensuring Data is Used Appropriately

2015

Protecting Data From Commercial Use
Transparency
Ensuring Data is Used Appropriately

2016

Ensuring Data is Used Appropriately
3rd Party Provider Restrictions
Transparency
Parental Consent and/or Notification

How ED is Using Disclosure Avoidance

School and Local Educational Agency (LEA)-level Assessment Data:

When publishing the two outcome category school and LEA-level math and language arts assessment data, the Department employs a combination of primary cell suppression for very small subgroups, and blurring of data for medium-sized groups using ranges and top/bottom-coding with varying widths, depending on the size of the reported subgroup.

How ED is Using Disclosure Avoidance

State-level IDEA and Special Education Data: For IDEA and special education data releases, the Department typically relies on aggregation to the State-level, coupled with primary cell suppression, complementary cell suppression, and/or top/bottom-coding, as necessary, to protect privacy and prevent reidentification of specific individuals.

How ED is Using Disclosure Avoidance

Civil Rights Data Collection (CRDC): The public - release version of the Civil Rights Data Collection employs a sophisticated rounding routine to protect privacy and prevent reidentification. Most CRDC data elements are blurred using rounding, while data elements relating to outcome/performance data and those pertaining to IDEA and special education are protected using a combination of bottom-coding and rounding. All rounding routines for the CRDC are applied at the lowest level of subgroup disaggregation, and all row, column, and multidimensional tabular totals are calculated using the rounded values.

Policy Update

April 2016 Letter to Louisiana

States and Districts may release **Basic Enrollment data** (student counts disaggregated by **Race/Ethnicity X Gender**) without privacy protections

Student counts disaggregated by other characteristics, and student outcome and performance data will likely still need disclosure risk analysis and the application of statistical disclosure limitation methods.

Looking Ahead

Privacy Technical Assistance Center (PTAC) Resources

Student Privacy Website:
https://studentprivacy.ed.gov

- Issue Briefs
- Checklists
- FAQs
- Case Studies
- Webinars
- Policy Letters
- Etc.

Help Desk: PrivacyTA@ed.gov

On-site Assistance
(site visits, trainings, etc.)

Selected PTAC Resources on Disclosure Avoidance:

Frequently Asked Questions—Disclosure Avoidance

Data De-identification: An Overview of Basic Terms

Case Study #5: Minimizing PII Access

CONTACT INFORMATION

United States Department of Education,
Privacy Technical Assistance Center

☎ (855) 249-3072
(202) 260-3887

✉ privacyTA@ed.gov

🖥 https://studentprivacy.ed.gov

📠 (855) 249-3073