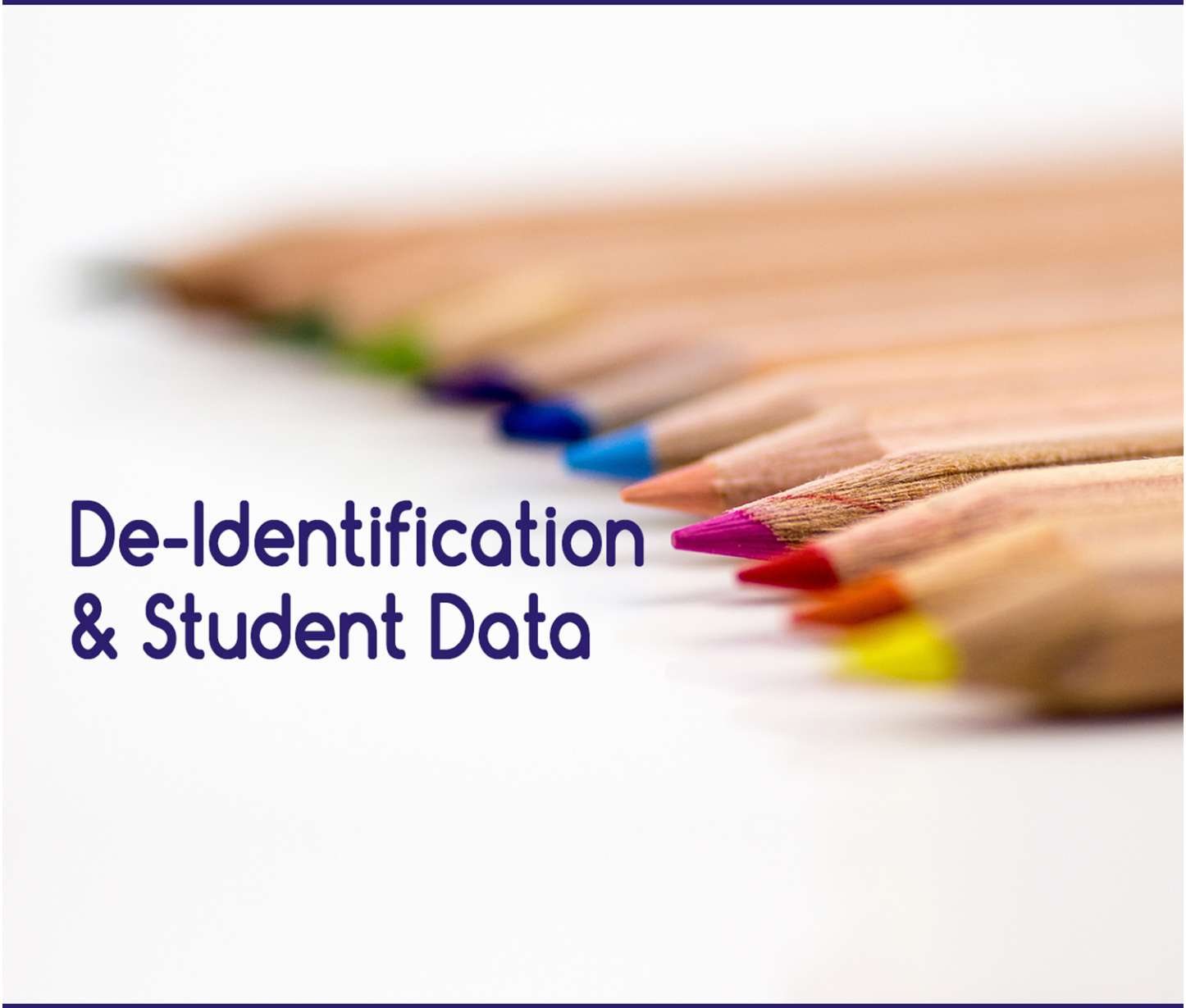
FUTURE OF PRIVACY FORUM



De-Identification & Student Data

Reg Leichty
Foresight Law + Policy
Brenda Leong
Future of Privacy Forum

August 2015

De-Identification and Student Data

Understanding De-Identification of Education Records and Related Requirements of FERPA

Appropriate and well-designed student data use by schools, families, researchers, and service providers, greatly enhances teaching and learning. New technologies linked to high capacity broadband networks offer educators and other stakeholders access to powerful analytical tools, rich data, and dynamic digital resources, which can improve student outcomes and inform important education policy reforms. These technology advancements, however, also invite new risks for exposing personally identifiable student data to unauthorized disclosures, misuse, and abuse. In order to reap technology's benefits without encountering these pitfalls, educational agencies and institutions, and their outside partners, must develop and implement more effective strategies and tools for promoting students' privacy and confidentiality.

Data de-identification represents one privacy protection strategy that should be in every student data holder's playbook. Integrated with other robust privacy and security protections, appropriate de-identification – choosing the best de-identification technique based on a given data disclosure purpose and risk level – provides a pathway for protecting student privacy without compromising data's value. This paper provides a high level introduction to: (1) education records de-identification techniques; and (2) explores the Family Educational Rights and Privacy Act's (FERPA) application to de-identified education records.[1] The paper also explores how advances in mathematical and statistical techniques, computational power, and Internet connectivity may be making de-identification of student data more challenging and thus raising potential questions about FERPA's long-standing permissive structure for sharing non-personally identifiable information.

The Three-Legged Stool of De-Identification: Personally Identifiable Information, De-identification Strategies, and Data Sharing Purposes & Disclosure Risk Assessment

Data de-identification is a technically and legally complex issue with special nuances across industries and areas of law. This paper narrowly examines the issue from the perspective of education records and FERPA. The U.S. Department of Education's Privacy and Technical Assistance Center (PTAC) defines de-identification as the "process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them."[2] Understanding PTAC's definition is critical to complying with FERPA and ensuring adherence to de-identification best practice. With that goal in mind, this section introduces three core student data de-identification concepts drawn from PTAC's definition and FERPA (law and regulations): personally identifiable information (PII); de-identification processes; disclosure purpose and risk assessment.

[1] Family Educational Rights and Privacy Act, 20 U.S.C. 1232g.
[2] *Data De-identification: An Overview of Basic Terms.* U.S. Department of Education Privacy Technical Assistance Center, PTAC-GL, Oct 2012 (updated May 2013).

Personally Identifiable Information

Educational agencies and institutions, and their partners, use de-identification to sever or obscure connections between useful education data and "personally identifiable data." FERPA's sharing prohibitions and requirements (explored later in the paper) only apply to PII. In other words, non-personally identifiable information may be shared and retained without restriction (with a narrow exception related to de-identified data connected to a record locator). As a result, understanding the law's definition of PII is critical to making determinations about how student data may be used, when, and by whom. Under FERPA, PII includes, but is not limited to:

a) The student's name
b) The name of the student's parent or other family members;
c) The address of the student or student's family;
d) A personal identifier, such as the student's social security number, student number, or biometric record;
e) Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
f) 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 knowledge of the relevant circumstances, to identify the student with reasonable certainty; or
g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.[3]

Educational agencies or institutions, and partner entities, such as technology vendors, community based organizations, or researchers, interested in using de-identification as a privacy protection strategy, must pay particular attention to the definition's inclusion of "indirect identifiers" and "other information." Data de-identification techniques are used to remove the direct identifiers described above, as well as indirect identifiers and other information, which if left unaddressed, could be used to identify individual students. Other examples of indirect identifiers include race, religion, weight, activities, employment information, medical information, education information, and financial information.[4]

Data De-Identification Techniques

Data de-identification – removing or obscuring PII - begins with eliminating all direct student identifiers from an education record, but education agencies and institutions, and other data holders, must take further steps to ensure that indirect identifiers or other information do not enable an unauthorized actor from determining a student's identity. These further steps involve using sophisticated mathematical and statistical de-identification techniques, including

[3] FERPA, 10 U.S.C. 1232g; 34 CFR § 99.3.
[4] See Privacy and Technical Assistance Online Glossary: http://ptac.ed.gov/glossary. Last visited, April 12, 2015.

leveraging technology to ensure the methods are accurately and comprehensively applied across large and complex data sets. Selection of an appropriate de-identification strategy will vary based on specific context, including whether it will be applied to individual level data (information collected and recorded separately for each student) or aggregate data (data combined from several measurements). The former requires much more robust protections.

The U.S. Department of Education's PTAC provides helpful guidance materials, including case studies, that provide detailed information about de-identification approaches,[5] but common methods include the following strategies.[6] See Addendum A for high level examples of each technique.

Blurring	**Perturbation**	**Suppression**
Reducing the precision of disclosed data to minimize the certainty of individual identification. For example converting continuous data elements into categorical elements that subsume unique cases.	Making small changes to the data to prevent identification of individuals from unique or rare population groups. For example, swapping data among individual cells to introduce uncertainty.	Removing data, for example from a cell or row, to prevent the identification of individuals in small groups or those with unique characteristics. Usually requires suppression of non-sensitive data.

Sharing Purpose & PII Disclosure Risk assessment

Educational agencies and institutions planning to use de-identification techniques to enable unconsented data sharing – in instances when a FERPA disclosure exception does not apply - must make a "reasonable determination that the student's identity is not personally identifiable because of unique patterns of information about the student whether through single or multiple releases, and taking into account other reasonably available information."[7] The standard for making this determination is discussed later in the paper, but neither FERPA, nor the U.S. Department of Education's FERPA regulations, provide a "safe harbor" listing specific steps that lead to appropriate de-identification. Instead, federal policy provides a standard for making case-by-case judgments of PII disclosure risk at the educational agency, institution, or approved party level.[8] This case-by-case approach means that the list of indirect identifiers that must be removed or obscured to achieve appropriate de-identification will likely vary by circumstance.

[5] Privacy and Technical Assistance Center: http://ptac.ed.gov. For example, *Frequently Asked Questions on Disclosure Avoidance*, PTAC-FAQ-2, October 2012 (updated May 2013), *Data De-identification: An Overview of Basic Terms*, PTAC-GL, Oct 2012 (updated May 2013), *Case Study #5: Minimizing Access to PII: Bet Practices for Access Controls and Disclosure Avoidance Techniques*, PTAC-CS-5, October 2012.

[6] See also, Federal Committee on Statistical Methodology's Statistical Policy Working Paper 22 Report on Statistical Disclosure Limitation Methodology, (73 Fed. Reg. 74806-35, Dec 9, 2008).

[7] 73 FR 73833, December 9, 2008.

[8] 73 FR 74834, December 9, 2008.

Selecting an appropriate de-identification method depends in part on examining the planned data sharing purpose. The data sharing purpose and de-identification strategy must be compatible.[9] For example, researchers interested in examining students' performance over time might require access to detailed, accurate academic information spanning several years (limiting use of de-identification techniques that diminish a data's validity). Researchers studying a student cohort's growth toward a state's college and career ready standards using a specific pedagogy, for example, would not be able to use data de-identified using a technique that limits the data's reliability and validity. (Alternatively, this type of longitudinal research might be conducted using de-identified data linked to a record locator to enable the originating educational agency or institution to provide de-identified data for the same students over time. Use of such a locator does not render the data "personally identifiable" under FERPA, but it does trigger special requirements.) Conversely, data shared for purposes that require less data precision and accuracy, such as software training or technology research and development, could use much more aggressive de-identification strategies, such as using techniques that replace sensitive information with inauthentic or modified data.

Please note, using de-identification techniques as a privacy tool does not always involve removing all PII, but in situations when PII remains part of a given data set (i.e. where the data has not been completely de-identified), unconsented sharing may only occur with consent or consistent with an appropriate FERPA exception. For example, an educational agency or institution sharing PII under a qualified FERPA exception may wish to use de-identification techniques to minimize PII released to an outside entity, even though they may lawfully share a range of student level information. To be more specific, a researcher might conduct a study that requires a discrete list of indirect identifiers that together could lead to the student's identification, such as a student's age, race and family financial information, but not requiring other PII found in the same education records. In such an instance, these three pieces of personally identifiable student data – and other information attached them - would remain subject to FERPA's disclosure limitations and other requirements, but de-identification techniques (e.g., suppression) could provide additional protection for the student by removing data, for example from a cell or row, unnecessary to the study. Researchers lawfully using PII in this context and other cases, however, must completely de-identify any report or other information before releasing it to the public or other parties, including other researchers.[10]

Entities planning to use de-identification techniques must mitigate the risk of exposing the identity of individual students. Therefore, after examining the requirements of a given data sharing purpose, education data holders must also assess the risks associated with their planned disclosure, including considering past data releases (the risk of re-identification is cumulative), sample size, the nature of the data recipient,[11] whether the data will be further shared or made

[9] *Data De-identification: An Overview of Basic Terms.* U.S. Department of Education Privacy Technical Assistance Center, PTAC-GL, Oct 2012 (updated May 2013), p. 4.
[10] 73 FR 74834, December 9, 2008.
[11] The Department of Education has said "there is no statutory authority in FERPA to modify the prohibition on disclosure of personally identifiable information from education records, or the exceptions to the written consent requirement, based on the track record of the party, including journalists and researchers, in maintaining the confidentiality of information from education

public, and other contextual conditions.[12] More aggressive de-identification strategies are required in situations when the student data is potentially at greater risk of re-identification.

For example, de-identified data shared for a specific purpose with a trusted public or private entity such as a state department of education, institution of higher education, or professional vendor with strict legal and contract protections (e.g., an agreement with strict re-disclosure limitations), might be less likely to be widely available later (decreasing the re-identification threat associated with cumulative data releases), compared for example to annual school or district performance data posted directly to a public website to comply with federal and state accountability requirements. Why is greater public availability of a properly de-identified data set a potential problem? In some cases, de-identified data might be subject to nefarious comparisons with other data sets (e.g., with widely available student "directory information") or other attempts to reveal PII. When data enters the public domain, it could be exposed to cutting-edge tools and techniques designed to compare the de-identified data to other publicly available data sets and thus reveal a students' identity (the FERPA implications of such a breakthrough are discussed further below).

Although experts disagree about the extent to which new technologies and techniques can "back map" de-identified data to reveal a student's identity, a serious statistical analysis that ensures all direct and indirect identifiers have been removed can be performed to ensure any re-identification risk is remote.

In short, prudent student data holders should consider using – in light of new data mining and comparison techniques that might be more effective than is commonly accepted – the most aggressive de-identification strategies possible when data will be made public or shared widely. When data is shared with limited restricted parties under strong controls and under a FERP exception, a combination of technical, administrative and contractual controls will be appropriate for reasonable de-identification measures that may preserve greater utility of the data.

Application of FERPA to De-Identified Records

As a general rule, FERPA prohibits the disclosure of education records containing personally identifiable student data without parent or eligible student consent.[13] Therefore, the release of education records that have been appropriately de-identified – purged of direct and all necessary indirect identifiers in a given context - is not considered a "disclosure" under FERPA, since by definition such records do not contain PII.[14] Properly de-identified student data thus may be shared without limitation under FERPA (although other federal and state privacy laws may apply). Furthermore, "de-identified information from education records is not subject to any

records that they have received." (73 FR 74834). Nonetheless, the recipients' identity should likely be considered among other variables in each risk assessment.

[12] *Frequently Asked Questions – Disclosure Avoidance,* p. 4, PTAC-FAQ-2, Oct 2012 (updated May 2013). p.2-3

[13] 20 U.S.C. 1232g(b)(1)

[14] 34 CFR 99.31(b)(1)

destruction requirements because, by definition, it is not 'personally identifiable information."[15] The Department has said, however, a party releasing de-identified student data might mitigate risks associated with future data releases by independently requiring data destruction in some circumstances.[16]

There is one important exception, however, to FERPA's unconsented sharing exception for de-identified data. De-identified data coupled with a record code or locator by an educational agency or institution – allowing it to be matched later to the record source - may only be shared for education research. Although the Department's regulations and guidance do not specifically discuss the question, it appears that educational agencies or institutions may select any qualified third party to conduct research under this provision, but all secondary (non-research) uses of de-identified data with a record locator are prohibited. Furthermore, the data sharing entity may not disclose information about how it generated and assigned the record code, or other information that might allow a data recipient to identify a student based on the record code. Lastly, the record code must not be based on a student's social security number or other personal information.[17] Such a data set remains categorized as "de-identified," and may thus be shared without parent or eligible student consent, but unlike other de-identified data it may only be shared for the research purpose specified to the educational agency or institution, consistent with the other requirements described above.

Before such data sharing can occur, however, the education record must be properly de-identified. As referenced above, the "releasing party is responsible for conducting its own analysis and identifying the best methods to protect the confidentiality of information from education records it chooses to release."[18] This determination depends on FERPA's disclosure risk assessment standard. This standard asks whether a "reasonable person in the school community who does not have personal knowledge of the relevant circumstances" could use the released data, and other publicly available data, to identify an individual student with "reasonable certainty."[19] This standard extends to possible data holders beyond the literal school community.

The Department of Education does not require educational agencies and institutions to use specific data disclosure avoidance techniques to achieve this standard, and stated in a recent rulemaking, "it is not possible to prescribe or identify a single method to minimize the risk of disclosing personally identifiable information that will apply in every circumstance…"[20] The Department has also said "determining whether a particular set of methods for de-identifying data and limiting disclosure risk is adequate cannot be made without examining the underlying data sets, other data that have been released, publicly available directories and other data that are linked or linkable to the information in questions.[21] In other words, the party releasing data

[15] 73 FR 15585, March 24, 2008
[16] 73 FR 74835, December 9, 2008
[17] 34 CFR 99.31(b)(2)(i)-(iii).
[18] 73 FR 74835, December 9, 2008.
[19] 34 CFR § 99.3, 34 CFR §99.31(b)(1)
[20] 73 FR 74835, December 9, 2008
[21] Ibid at 74835

must perform a context specific analysis and identify the best method for protecting student information subject to disclosures. Proper application of the accepted mathematical and statistical de-identification strategies described earlier in the paper meet this legal standard in many instances, but by law each sharing context must be independently analyzed against the Department's reasonableness standard.[22]

Some experts have argued that given recent cases where researchers have leveraged access to other publicly available data sets to identify specific individuals, absolute data de-identification may be impossible, or at a minimum, increasingly difficult.[23] In light of this uncertainty, data sharing parties should very carefully analyze each proposed disclosure of de-identified data against FERPA's reasonableness standard and also consider using contracts that specify protections – above and beyond FERPA - that could further minimize the risk of re-identification.

De-Identified Data: Retention and Destruction

FERPA permits third party data holders, including vendors, to retain and use appropriately de-identified data – so long as it is not associated with a record locator -for any secondary purpose. Furthermore, FERPA does not describe how de-identified data should be managed, including, as described above, when and how the data should be destroyed. Vendors and other third party holders must, however, ensure that a given de-identified data set is not subject to relevant contract terms, or other Federal, state, and local privacy laws and regulations, which might contain more stringent data retention or destruction requirements.[24] For example, personal data subject to the Children's Online Privacy Protection Act may only be retained so long as is necessary to fulfill the purpose for which it was collected, and COPPA covered entities must delete the information using reasonable measures to protect against its unauthorized access or use.[25]

Although FERPA does not govern the use, retention and destruction of properly de-identified data, third parties should have sound policies – guided by National Institute of Standards and Technology or PTAC best practice recommendations - addressing these issues. This internal, independent step includes ensuring that de-identified data is destroyed when it is no longer needed, in order to minimize re-identification risks associated with possible future efforts to compare and link the data with other data sets. Data holders must also ensure that they take proper actions to destroy data. Simply deleting data is not sufficient in most cases and PTAC's data destruction best practices provide helpful guidance. PTAC recommends that data holders "make risk-based decisions on which [destruction] method - [e.g. clearing, purging, or destroying data] - is most appropriate based on the data type, risk of disclosure, and the impact if that data were to be disclosed without authorization."[26] The data de-identification method used to remove

[22] 34 CFR 99.31.(b)(1). See also, PTAC *Frequently Asked Questions – Disclosure Avoidance,* p. 4, PTAC-FAQ-2, Oct 2012 (updated May 2013).
[23] *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* Paul Ohm, University of Colorado Law School, UCLA Law Review, Vol. 57, p. 1701, 2010 .
[24] Privacy and Technical Assistance Center, *Best Practices for Data Destruction,* p. 5, PTAC-IB-5, May 2014.
[25] 16 C.F.R. § 312.10.
[26] PTAC Best Practices for Data Destruction, p. 5.

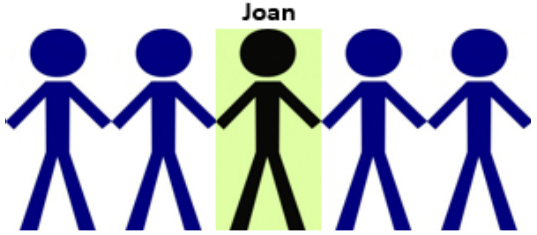
PII from a data set should be a central factor in making this determination. Data holders seeking additional guidance on proper destruction strategies should consult recommendations made by the National Institute of Standards and Technology and other expert sources.[27]

Conclusion

De-identification offers an important tool for educational agencies, institutions and their partners seeking to maximize student data's potential value to improving teaching and learning, while also carefully protecting student privacy and confidentiality. Proper data de-identification requires, however, deep technical knowledge and expertise and adherence to industry best practice. Therefore, student data holders should not attempt to de-identify student data sets without competent support. They should also consult competent legal counsel to ensure that their data management policies and practices – including de-identification strategies - comply with FERPA and all other relevant federal, state, and local laws and requirements potentially applicable to the data they manage.

[27] National Institute of Standards and Technology (NIST) Special Publication 800-88 Rev. 1: Guidelines for Media Sanitization. December 2014.

Illustration of Common De-Identification Measures in Aggregate Data Sets



Raw Individual Student Data in Aggregate Data Table

Joan's Director Identifiers	**Joan's Indirect Identifiers**
Student Name: Joan Smith Students Parents: John Smith & Jackie Smith Address: 0000 00th Street, Washington,D.C. Student Number: 4444 Social Security Number: 555-555-555	Data of Birth: 11/01/2000 Race: Alaska Native Gender: Female Place of Birth: Washington, D.C. Family Income: $85,000 GPA: 3.75

Redacted Individual Student Level Data in Aggregate Data Table

	Joan's Indirect Identifiers
All Direct Identifiers Removed	Data of Birth: 11/01/2000 Race: Alaska Native Gender: Female Place of Birth: Washington, D.C. Family Income: $85,000 GPA: 3.75

Blurring (Reducing Data Precision including Using Broader Categories)

	Joan's Indirect Identifiers
All Direct Identifiers Removed	Data of Birth: 2000 Race: Minority Gender: Female Mother's Maiden Name: Johnson Place of Birth: Mid-Atlantic Family Income: $50,000 - $100,000 GPA: 3.5 – 4.0

Perturbation (Small Data Changes, including through Swapping Data among Cells)

Mike's Indirect Identifiers	**Joan's Indirect Identifiers**
Data of Birth: 1999 Race: Unique Characteristic Removed Gender: Female ⟷ Mother's Maiden Name: Unique Characteristic Removed Place of Birth: Midwest ⟷ Family Income: $50,000 - $100,000 GPA: 3.5 – 4.0	Data of Birth: 2000 Race: Unique Characteristic Removed Gender: Male Mother's Maiden Name: Unique Characteristic Removed Place of Birth: Northeast Family Income: $50,000 - $100,000 GPA: 3.5 – 4.0

Suppression (Removing Data from a Cell or Row)

	Joan's Indirect Identifiers
All Direct Identifiers Removed	Data of Birth: 2000 ~~Race: Unique Characteristic Removed~~ Gender: Female ~~Mother's Maiden Name: Unique Characteristic~~ Removed Place of Birth: Mid-Atlantic Family Income: $50,000 - $100,000 GPA: 3.5 – 4.0