**Identification: A Teaching Moment for Privacy and Databases**

Matthew Dube and Rocko Graziano  
University of Maine at Augusta; Spring 2023

Prepared for EngageCSEdu

In relational database theory and implementation, a **primary key** represents a specified means for identifying *every* record in a table from *every other* record in the same table by using the same consistent set of attributes each time. This specifically requires that that set of attributes has values for each component attribute (NOT NULL) and that each individual record’s values for those attributes in the aggregate are distinct (UNIQUE). This same structure is mirrored in the **dictionary** concept in programming languages.

The children’s game *Guess Who?* is a great example of this concept. Each character that you try to guess has a specific name (key), but the various people have a distinct set of properties that someone could query against to find out who the opponent chose. The table that represents the *Guess Who?* game by default has two possible key structures: the *name* of the character and the *attributes* of the character (assuming that you store a value for every quality one might ask about). Crazy Games implements a free version of *Guess Who?* (<https://www.crazygames.com/game/guess-who-multiplayer>) that can be used as a demonstration of the concept. The table by which the game may judge is shown here.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Sex** | **Eyewear** | **Adornment** | **Earrings** | **Facial Hair** | **Headwear** | **Hair** | **MouthOpen** |
| Chantal | F | Glasses | Necklace | Y |  |  | Brown | Y |
| Eric | M |  | Tie | N | Goatee | Bunny Ears | Bald | N |
| Alex | M |  | Tie | N |  |  | Blond | Y |
| Bob | M |  |  | N | Beard | Bunny Ears | Blond | N |
| Paul | M |  |  | N | Mustache | Bandana | Sand | N |
| Frank | M |  |  | Y | Mustache | Headband | Black | Y |
| Zoe | F | Mask | Necklace | N |  |  | Brown | Y |
| Joe | M | Glasses | Tie | N | Mustache |  | Gray | N |
| Buba | F |  |  | Y |  | Headband | Black | N |
| Rita | F | Glasses | Tie | N |  |  | Purple | N |
| Rick | M | Glasses | Tie | N | Goatee |  | Blond | N |
| Antoine | M |  |  | N | Mustache |  | Brown | N |
| John | M | Glasses |  | N |  |  | Black | N |
| Chap | M | Glasses |  | Y | Beard |  | Bald | N |
| Evelyn | F |  | Necklace | Y |  |  | Pink | N |
| Lady | F | Glasses | Tie | N |  |  | Brown | N |
| Samantha | F |  | Necklace | N |  | Bandana | Blond | N |
| Jenny | F | Glasses | Tie | N |  |  | Pink | N |
| Javier | M |  |  | N | Mustache | Headband | Pink | Y |
| Evan | M |  | Tie | N | Mustache |  | Blond | N |
| Mathias | M | Glasses |  | N |  |  | Sand | N |
| Michael | M |  | Tie | N |  |  | Blond | N |
| Hank | M |  | Tie | N |  | Bunny Ears | Pink | Y |
| Vito | M | Mask |  | N | Beard |  | Bald | N |

*Activity: Let’s play Guess Who? The board represents a collection of characters. One player chooses a character at random from the game board. The second player will try to guess who was selected by asking a series of questions for which the only answer is YES or NO. "What is the name of the character selected?" would not be valid, but "Is the character named Eric?" would be. The goal is to identify the selected character in the fewest number of guesses.*

*Have your students pair up and play the game. Ask them to think about their strategy. Ask them to think about how what they are doing relates to identifying an individual. They should hopefully come to the conclusion that each of the questions they are asking becomes part of a filter for finding the answer they seek.*

Each character has a distinct name, satisfying the requirements for a **candidate key**. Similarly, no two characters have exactly the same properties beyond the name, therefore the entire set of properties is at worst a **superkey**.

Often, when students engage in database learning, when asked to identify a candidate key for a database, a student might do this: *look at each row and find what tells me that I am looking at this row and nothing else*. That’s obviously the name here, and intuitively, that will do the job. Whenever there is a **compound key**, however, students take the same approach and resort to finding something that is unique about that row, independent of whether or not that exact combination of attributes would identify *all* rows. This is a misconception about primary keys that students will often initially harbor. It mirrors how our brain works: we don’t necessarily care about universal categorization in the moment; we care about the insights about the situation in front of us.

This begs a very important question: *what does it mean to identify*? In a computational context, the concept of a primary key or an address or a dictionary index is what we refer to as identification - if one searches a file over that value, it will have no more than one hit, independent of what value we throw at it. For example, if looking at US states, the name of the state or its postal abbreviation would satisfy this particular condition: no two states can have the same name or two-character postal code. We call this a **set** mathematically and in a data structures course. More often than not though, attributes are what are called **bags**: they can be repeated. Just because one thing is repeated, however, does not mean that all things are. This is a different approach to identification, namely all that matters is that if a search were conducted for this *particular* value, it is only found once. That type of identification acts how our brain wants to act: handle the task at hand; it is a shortcut. That’s a particular problem for privacy. Let’s explore how.

*Activity: Put students into groups and have them divvy up the rows of the table. Can you find any conditions that identify that row from all other rows (without using the name)? Now find a set of columns that identify every character in your subset from each other (again, without using the name). Share results with the group. Note anything that could be found with one piece, two pieces, etc. Then see if the subset’s key was the same for each group (it likely will not be).*

The activity promotes an understanding of the flaws of the misconception of primary keys. On the one hand, a search mechanism based on a specific trait (or trait combination) (though intuitive) is probably ineffective overall - it works very specifically. On the other hand, a unifying set of characteristics that makes everything together unique within a subset will also not necessarily carry over to a whole dataset, a flaw that we might miss if we try to identify a primary key from a sample dataset. A great example of this can be found in police lineups. If a person is suspected of a crime and they are put in a line in front of a witness or victim, if their features are immediately distinguishable from the person in the line and remotely similar to who they thought they observed, “identification” is achieved, but would not be achieved globally when other similar features might be present. A primary key does both: it identifies an individual by default and identifies every group (and subgroup) in precisely the same manner.

The misconception is, however, a valuable learning opportunity. If any record carries something distinctive within it, that record is potentially vulnerable to identification through a backdoor, even if it is not a structural data identifier in terms of indexing for a database. An example of that is found with Rita: Rita’s purple hair is unique to the set of people in the game. This likely leads participants in the game to not choose Rita as their character, because if someone ever asked the purple hair question, they would lose. Similarly, Joe’s gray hair would produce the same issue. Zoe and Vito both can be identified by their sex and eyewear: each one wears a mask, but they present as the opposite sex. Conversely, if we think about this from a hacker’s perspective, asking the purple hair question is very ineffective because it would only narrow my search for the nugget I seek by one record. This would be poor game play for *Guess Who?*.

While identification of data is important computationally, the second that we start talking about data that has anything to do with people we run into a core concept in society: **privacy**. In general, privacy is an idea that a person’s data is only available if they allow it to be; specifically, if they consent to their data being released. In this case, *data* is being used broadly: what goes on in one’s home (think about a voice assistant), health data, educational data, undisclosed attributes, etc. Though privacy touches lots of things in one’s life, these are major intersections with the data world. We have to see that in conjunction with identification: if I can identify an individual’s records in a database, I have linked them to the attributes connectable to them.

*Discussion: Is privacy important to functional society? Why or why not?*

Daniel Solove [6] published a book in 2008 (*Understanding Privacy*) about privacy. In its first paragraph, he collates different comments about privacy, including:

* Essential to democratic government
* Critical to our ability to create and maintain different sorts of social relationships with different people
* Necessary for permitting and protecting an autonomous life
* Important for emotional and psychological tranquility
* Integral to our humanity
* Heart to our liberty
* Beginning of all freedom

The comments he collated highlight that privacy is many things to different people. He goes on to quote Jonathan Franzen who calls privacy the Cheshire Cat of all values: not much substance, but a winning smile. The lack of substance is because of the lack of agreement.

There are many lenses of ethics, one of which breaks ethical issues down into four categories [4]:

* Autonomy - personal control
* Beneficence - doing the most good
* Non-Malfeasance - doing the least harm
* Justice - fairness, just desserts

Privacy is of course an ethical issue because it flags many of those four categories. Privacy as noted by the Solove collation references autonomy - we cannot have autonomy but for privacy. Making certain things public could do both good and harm to society, but that disclosure might be differentially impactful to the person concerned. Privacy also harkens to justice - is it fair for you to know something about me if I don’t provide or authorize it?

Privacy is also a legal issue, and famously so in the United States. Privacy as a named entity is a glaring omission in the US Constitution and its amendments. The Declaration of Independence references liberty, which one might conflate to infer privacy. Three amendments to the US Constitution from the Bill of Rights do in some sense reinforce privacy in different ways, and a fourth added in the wake of the US Civil War has been used in many instances to protect the disclosure of private acts.

* The third amendment refers to a person subject to US jurisdiction not having to quarter a soldier, an employee of the federal government who is likely armed. This assumes that one’s own home is a private space from the government
* The fourth amendment refers to a person subject to US jurisdiction not being subject to search and seizure without a warrant or probable cause. This assumes that the contents of one’s own possessions and artifacts is a private space from the government unless there is evidence of a violation of the law
* The fifth amendment refers to a person subject to US jurisdiction not having to provide self-incriminating testimony. This assumes that one’s own actions and thoughts are also private, unless someone else can corroborate them, the purpose of having witnesses in a courtroom. This of course makes trying cases quite difficult because the accused cannot be compelled to answer questions (though the act of not answering the question might harm the presumption of innocence with a jury)
* The fourteenth amendment refers to all people on US soil being afforded due process. This particular amendment has been used in many cases to uphold the rights of protected classes. One of the most famous cases of this is *Obergefell v. Hodges*, which protected the sanctity of same-sex marriage under the concept that a private attribute (sexual orientation) could not be used in a discriminatory fashion (in this case, to not recognize one’s marriage).

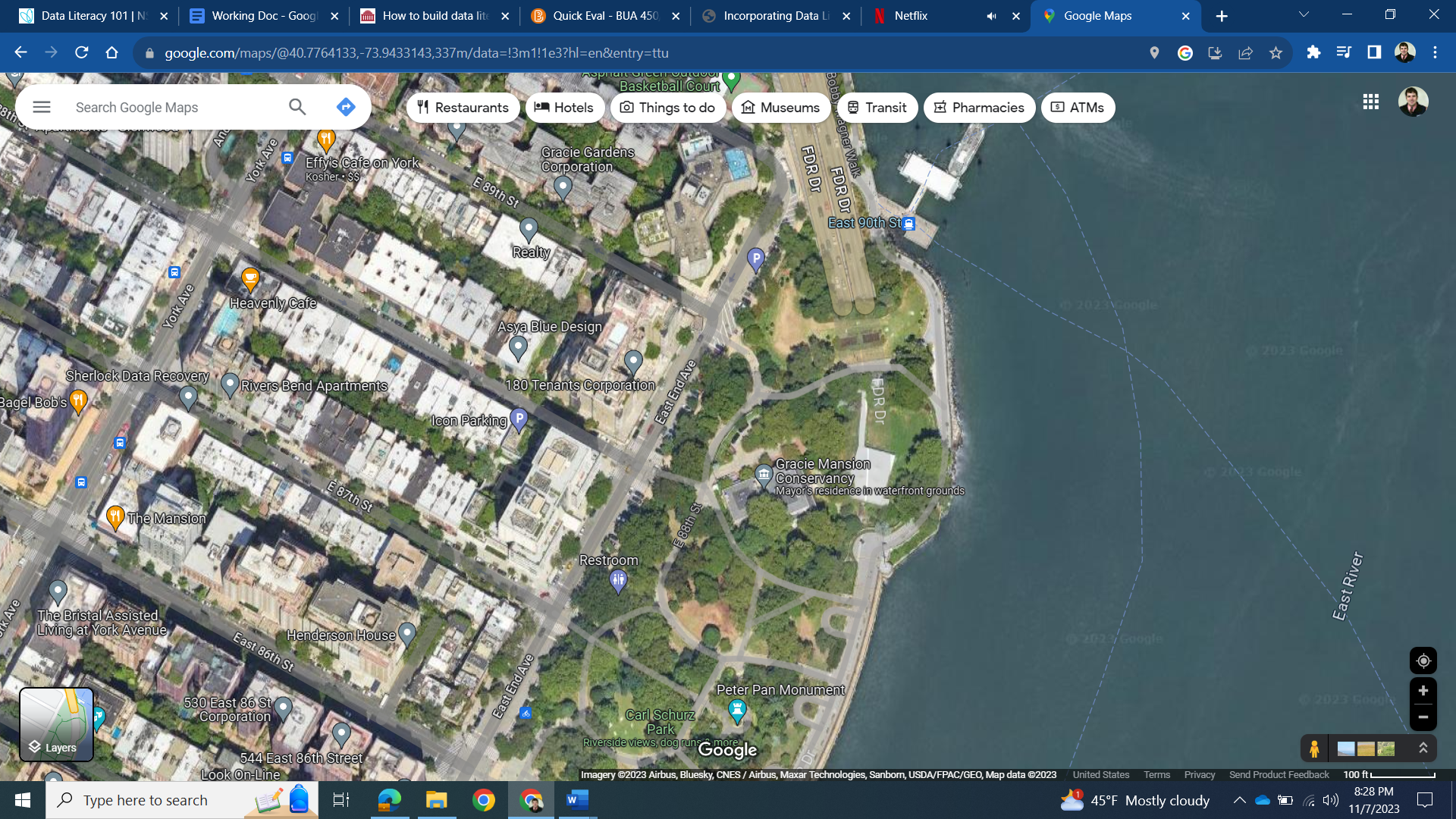
Since the way in which people view its purpose and scope is so wide, there has been great public interest in privacy, and it is much more important today due to the data-rich nature of our society. As such, people have thought about assisting or maintaining the privacy of data for a long time. **Anonymous** data refers to data that is stored without personal identifying information (PII). **Confidential** data refers to data that does not directly identify the individual, however, the researcher maintains a mechanism by which that they can identify the individual.

*Discussion: are anonymity and confidentiality enough to protect the privacy of data?*

While both of these concepts certainly have a place and protect some data, they are not in and of themselves enough. We can see that these two concepts are not enough in two very different lights: the first is through the connectability of datasets (which we will discuss now), while the second is about inferring a piece of information not provided through machine learning (which we will discuss later).

Anonymity and confidentiality motivate how records are stored in human contexts. The results of census operations or evaluation surveys are often presented in an aggregate form to protect individuals from identification (anonymity). While a researcher might be able to identify the community, they may not be able to identify a member of that community from the data itself (note that confidentiality would in effect have a lookup table to reconnect identity, but that is locked away). This provides an essential safeguard to personal demographic or personal opinion privacy. This philosophy of course falls apart when the community does not have many members within it as the sheer answers provided may actually suggest their provider.

This is somewhat common in census geometries, particularly within cities. A mayor’s residence or governor’s residence or something like that has very few neighbors, often separated by streets. Those streets produce census blocks that have very limited population or household counts, thus introducing something that could start to assign a question’s answer to an individual at a much higher likelihood. Depending on how these blocks are used, they could create opportunities where someone’s name could get accessed. This image of Hell’s Gate in New York City shows Gracie Mansion (center), the residence of the mayor of New York City. It is separated by roads from any other building, abutted by FDR Drive, Gracie Square, and East End Avenue.



*Hell’s Gate, Manhattan, location of Gracie Mansion*.

*Discussion: why do we have to be concerned about giving away data with a personal identifier in several different contexts, even if the things you don’t want connected to one another are never given together?*

In August 2021, Jesse Clark, Lindsey Cormack, and Sam Wang [1] analyzed the cast vote record for New York City, which aggregates the choices of the population of the election district, and the New York State voter history file, which includes names, addresses, election districts, voter history, party affiliation, and registration status for each historical voter. By combining the results together, if an election district had a single voter in a particular election, these two datasets together would isolate it, because the voter history file would have only a single entry with that election district for that election. That then ties the name to the logged singular vote. In essence, the voter was able to be identified positively due to the election district because the election district had one ballot cast and another file could isolate what the vote actually was. This was true for 378 separate voters in the primaries for New York City Mayor in 2021. This is a gross violation of personal privacy: exposing how one voted in a secret ballot. This is a direct example of the first element of Solove’s collation: essential to democratic government. This opens up the door to many potential negative outcomes, including persecution if the leader would choose to pursue them. This would not have been possible but for the connectability of the two datasets.

While more than one person lived in that election district, the key here is that the situation produced a relevant filter that would turn something that was not unique into something that was, much like *Guess Who?* We might call this a conditional functional dependency. The most recognizable person affected by this was Dante DeBlasio, the son of the outgoing mayor. While he is not the only resident of that election geography, he was, however, the only one that voted in the Democratic primary. Because there are two primaries, that adds additional opportunities for identification. Because New York City is hyper-Democratic in partisan distribution, there are numerous opportunities for lone Republican ballots to be cast in electoral districts. The scenario here is akin to trying to identify Zoe or Vito in our sample dataset: once their sex is known (not really identifying), they can be identified by having a list of mask wearers (also not highly identifiable on its own in practice, though in this dataset, quite restrictive). The same would be true for Paul and Samantha with their bandanas. The same concept is exhibited in a biased police lineup where the person you suspect did it looks very different than any other member of the lineup, a concept that exhibits the foundational problem of implicit bias.

Identifiability is not always bad, however. **Classification** is a process by which different observations might in their similarities predict a particular outcome. A canonical example of this is found in research about breast cancer cells [5]. Using logistic regression techniques, a dataset is classified into benign and malignant based on a series of measurements from microscope slides. This type of approach produced a highly accurate model which could then be used to do a quick initial screen for the condition. This is another example of inferring one thing from another. This case is particularly interesting because the doctor would want to know to whom the sample belonged, representing confidential data. Because these items *can* be tied together by at least someone, this opens the risk to medical data being connected to other items. This is a price that as a society we have shown that we might be willing to pay, but in so doing, we enact specific standards to protect that data.

Another practical implication for this thought process is to think about ways we design methods in programming languages. There are libraries that search spatial data for prepositional relations. These libraries have methods that isolate a specific relation (such as contains) or try to generate the relation (such as a 9-intersection model or a logical dictionary from measurements). The first one can be made far more efficiently than the second one. It need only have the smallest number of measures unique to it to facilitate the task. The second one requires a primary key condition. Egenhofer, Dube, and colleagues [2,3] studied these concepts for vector and raster spaces. While this is not a social risk, it is an example of an interdisciplinary application of cognition and linguistics to computing. We want to ask questions of datasets, and those questions are motivated in human terms. It is not always necessary to rely on the primary key to ask questions in the most efficient manner; sometimes a simple property identifies that which is required.

A proper identifier (in a database context) can also be problematic with respect to personal identification. The Social Security Administration created a unique account number for people who would receive benefits from that social program and would contribute in kind to it. Within ten years, many governmental agencies started to use this number as an identifier in their own data (which was actually not its expressed written purpose, which is likely surprising today). As databases grew up in the government, a consistent value could be used to connect records across agencies. The most efficient way to search these databases (their primary key) became a window into almost any personal database possessed by the government. Similarly, businesses, educational institutions, and healthcare facilities started to use the Social Security Number in their applications out of convenience. Personnel, students, and patients began to see one of the most sensitive pieces of information about a person shared out in the open. This type of primary key is quite dangerous. While it serves the purpose and is memorable for who it applies to, it unlocks the door to the same type of filtering that we saw in the DeBlasio case, but in that instance, it is far more open season and can readily impact anyone and everyone.

*Discussion: What types of things have you had to provide your Social Security Number for? What other information did you provide simultaneously? Why are we talking about this?*

Sometimes the right “identifier” or set of attributes can lead to joins that might infer a protected concept or (as in the case of DeBlasio) connect unintended items that were otherwise provided in good faith. In machine learning, we have seen that this can lead to the issue of ***proxy discrimination*** – the structured training data presented to the algorithm leads to categorizations which inadvertently discriminate based on a protected class – race, age, gender, etc – when that class can be inferred from the collected attributes. Notice the word *inferred* - the protected class was never actually collected.

A useful example of this is the collection of ZIP code information, perhaps as part of the applicant process for a mortgage loan. Now, imagine training an algorithm to make recommendations on loan applications – that is, judge the credit-worthiness of the applicant. To train the algorithm we present it the historical records of past applications and their approval status (approved or rejected), or even the results of the loan (paid back, defaulted).

What happens when that data reflects past discriminatory practices? When certain neighborhoods were ‘red-lined’ due to their majority-minority population? When applicants who had a hard time obtaining credit were subject to predatory loans with abnormally high interest rates? These past practices will become the “pattern” learned by the algorithm, which will then perpetuate those inherent biases in judging the credit-worthiness of future applicants.

Now, imagine that same problem being reflected by algorithms used to recommend sentences for those convicted of crimes, trained on data which reflects the societal bias of over-zealous policing in majority-minority urban centers. A ProPublica [7] report uncovered concerning patterns with Criminal Risk Assessment algorithms in use in Broward County, Florida, patterns which suggested that risk scores reflected the race of the defendant.

*Discussion: what can be done to prevent these outcomes? Is the algorithm biased? What responsibility do data scientists have to uncover and prevent proxy discrimination?*

As should be evident, the more data that we store (particularly about people), we are taking risks. The risk we take is that something that is sufficiently rare can create potentially dire consequences for a person’s privacy or lead to very targeted decisions motivated from past biased practices. At the same time, this inherent risk seems to be accepted because it provides the opportunity for efficient searching. This split notion of identity challenges the idea that things meant to be private can truly stay private. This should lead to asking questions about what should I put on social media, should I apply for things that I have no intention of actually utilizing, and many other questions.

**What Can We Do About It?**

While the things that we have talked about today are quite scary, there might be several ways in which we can mitigate some of these problems. We have alluded to some throughout this document, but let’s think about them a bit more:

*What We Collect*

We wish to collect data that will fulfill our intended purposes, the underlying goal of any information system at its core. An additional concern when we design information systems is to consider what might be needed in the future for potentially related tasks. In system design careers, being able to look toward future needs will allow the data we collect to have long-term value, and avoid the need to retire antiquated systems (or hold onto them beyond their useful life, providing more opportunities for accidental identification or unintended joins or unauthorized breaches.

While this at first seems to be a problem, it might not be as big of a problem as we might expect. Data mining methods routinely hold back data from model generation so that they can better assess how a model performs on non-training data. Data missing particular values become prime candidates for employing such methods as our models can help to potentially predict what is missing. This however runs the risk of not being a true representation. This is sort of a reframing of one issue into a chance for improvements vis a vis another.

*How We Collect It or Store It*

For data about people, we discussed in one way or another three separate ways in which privacy is at least attempted to be protected based on how we collect it. We can collect data free from personally identifiable information (anonymous); we can collect data identifiable only to the collector (confidential); we can aggregate individual data into area or temporal units to treat occurrences as more rate-based than event-based. We also saw that through our discussions, these techniques do not always lead to the ultimate full protection of the data. In the case of anonymous data, other things might be inferrable or connectable to something that is shared with a set that contains the identifier; confidential data maintains a vestige of identification and (if compromised), provides the key to potentially other doors (this leads to the practice of data discussions with Institutional Review Boards); aggregate data can work well, but only if there are sufficient numbers of items that together construct the aggregate. All of these can create false senses of security. Additionally, aggregate data transforms how we conceive of the data itself, which may or may not be a welcome concept.

*How We Allow It to be Viewed*

Database management systems allow for the construction of views. Views restrict what can be seen from a dataset at the user level. A good example of this is in academic records systems. Financial aid personnel or student employment personnel need access to something like a social security number. Faculty members or advising staff do not. As such, the social security number needs to be collected, but we can mitigate the issue by only making the social security number available to appropriate personnel, managed by user roles in the database system. While the data is still collected (and thus induces vulnerability), it is not as universally accessible within the organization. Organizations also enact policies outside of computerized systems that dictate what is fair and ethical usage of data. FERPA and HIPAA represent federal law approaches to this problem, through which organizations that work with this type of data have to justify their procedures and thus operate.

**References**

[1] Clark, J., Cormack, L., & Wang, S. 2021. Privacy Concerns in New York City Elections. Technical Report, Princeton University.

[2] Dube, M., Barrett, J., and Egenhofer, M. 2015. From Metric to Topology: Determining Relations in Discrete Space. Proceedings of the 2015 International Conference on Spatial Information Theory, 151-171.

[3] Egenhofer, M., and Dube, M. 2009. Topological Relations from Metric Refinements. Proceedings of the 17th ACM SIGSPATIAL 2009, 158-167.

[4] Gillon, R. 1994. Medical ethics: four principles plus attention to scope. Bmj 309.6948, 184-188.

[5] Mangasarian, O., Street W., and Wolberg, W. 1995. Breast cancer diagnosis and prognosis via linear programming. Operations research 43.4, 570-577.

[6] Solove, D. 2008. Understanding privacy.

[7] Angwin, Larson, Mattu, Kirchner (2016). Machine Bias, ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>