# Sneaky Links: Evaluating the Usability of Private Data Sharing Guidance

Sydney Earp
*Carnegie Mellon University*

Lynn Zhang
*Carnegie Mellon University*

Abhinaya Nagalla
*Carnegie Mellon University*

Raghav Trivedi
*Carnegie Mellon University*

Sarah Scheffler
*Carnegie Mellon University*

Christine Task
*Knexus Research Group*

## 1 Introduction

When organizations need to answer questions using data distributed across multiple datasets, directly sharing or combining those records can raise significant privacy concerns. Privacy-Preserving Record Linkage (PPRL) addresses these challenges by enabling organizations to combine data without revealing individual-level information [1, 6]. However, there is currently little publicly available guidance to help practitioners determine appropriate PPRL approaches. To address this gap, initial PPRL guidance was developed to help practitioners identify and understand different forms of data-sharing problems that may require different PPRL techniques or approaches. This work contributes to that effort through a usability study of three variations of this prototype guidance. Prior studies on the usability of existing privacy frameworks suggest that they are ineffective at communicating their guidelines to expert users [2–5] emphasizing the need to investigate how best to make these documents useful. Klymenko et al. suggest there is a gap in developers' current understanding of PETs and their use in achieving privacy goals. This highlights the need to communicate PETs as technical measures to satisfy privacy frameworks [2].

**PPRL Guidance** These prototype guidance documents are intended as a first step for practitioners working with data-sharing problems that require privacy-preserving analysis or linkage. The guidance aims to help users identify what type of PPRL problem they are working with using a set of six archetypes, each representing common data-sharing scenarios with distinct characteristics. The archetypes are intended to later support the selection of appropriate privacy-preserving techniques, but that is beyond the scope of this work.

The six archetypes are the following: (A1) Private Joined Records, (A2) Private Joined Categories, (A3) Private Joined Analytics, (A4) Private Cross-Partner Categorical, (A5) Private Cross-Partner Analytics, and (A6) Private Cross-Partner Subject Tracing.

To understand how participants interact with and apply these guidance prototypes, we investigate the following questions:

**RQ1:** How effectively can participants understand and apply the PPRL guidance to real-world data-sharing scenarios?

**RQ2:** How do participants use and interpret different components of the PPRL guidance when matching scenarios to archetypes?

**RQ3:** What challenges do participants encounter when distinguishing between PPRL archetypes?

## 2 Methods

To evaluate the usability of the PPRL guidance prototypes we conducted an IRB-approved, semi-structured task-based interview study with 13 participants recruited from the Carnegie Mellon University Campus. This let us observe guidance use while collecting feedback on participant reasoning. Participants were recruited from relevant technical backgrounds, with most participants identifying as either data science/statistics-leaning or computer science-leaning.

Each 45-minute in-person session was organized into four phases. In the introduction phase, participants answered questions about their academic or professional background and received a brief verbal introduction to PPRL. In the review phase, participants were given all three guidance prototypes: a *diagram* representation where each archetype is paired with a visual example and brief description (Appendix A.1), a *flowchart* guiding users through a series of yes/no questions (Appendix A.2), and a comparative *matrix* presenting


*USENIX Symposium on Usable Privacy and Security (SOUPS)* 2026.
August 23–26, 2026, Hannover, Germany.

archetypes side-by-side (Appendix A.3). Participants were given approximately 10 minutes to review the materials and make annotations if desired. In the task phase, participants completed an archetype matching activity using six real-world data-sharing scenarios, shown in Appendix A.4. For each scenario, participants were asked to determine which PPRL archetype best fit the situation using any combination of the provided guidance prototypes while thinking aloud. In the feedback phase, researchers conducted a semi-structured interview about participants' experience with the guidance, including the perceived advantages and disadvantages of each prototype, what was confusing, what was most useful, and rationale for their responses.

Quantitative data included participant archetype assignment for each scenario and time taken, but this data was analyzed in a qualitative manner. Qualitative data included prototype preference, problem solving process, and advantages and disadvantages of each prototype. Qualitative analysis was performed via 100% double coding for all participant transcripts. The two researchers assigned to code utilized emergent coding to create a codebook (see Appendix D).

## 3 Results

*RQ1:* Participants were sometimes able to apply the guidance, but their reasoning often showed partial rather than comprehensive understanding of the archetypes. On average, participants correctly matched 3.15 out of 6 scenarios, with most scoring between 2 and 4 (Figure 3). Participants often engaged with the guidance by matching scenario terms to similar terms in the guidance documents; for example, using words such as "average," "correlation," "count," or "income" to determine whether the output required simple counts or more complex analytics. This worked best for scenarios with clear output cues, such as Scenario 3, where 12 out of 13 participants selected the correct analytic archetype.

While assigning archetypes to the given scenarios, participants most often considered data type, parties involved, and individual vs. aggregate output. They rarely considered shared individuals, who has what, or who learns what. This further suggests that participants could apply visible parts of the guidance, but did not necessarily understand the structural differences that distinguish the archetypes.

*RQ2:* Participants combined the prototype variations, using each format for a different purpose. The flowchart helped participants identify what to consider before selecting an archetype. The matrix helped them compare archetypes and reference more detailed information about each archetype. Although several participants initially found the matrix overwhelming, it became more useful once they knew which details to look for. Diagrams were used least, though some participants valued the visual format (Figure 2). The one participant who relied wholly on the diagrams was the only participant to match all six scenarios to their appropriate archetype, however, this may reflect their strong background in data science and policy rather than diagram format alone. The most common combination was the flowchart and matrix, with 5 participants using those documents together.

Regarding the individual components within each of the formats, participants most often relied on the flowchart visualization (7 participants), matrix output column (6 participants), and the matrix goals/definitions (5 participants). Some participants read the matrix examples to help make sense of the archetypes, but only 3 clearly used them to decide which archetype best fit a scenario.

*RQ3:* Participants often struggled to determine whether a scenario had a primary data owner. This made it harder to distinguish between asymmetric linkage archetypes (A1–A3) and symmetric linkage archetypes (A4–A6). This issue was especially visible when participants used the flowchart, because the first decision point required them to correctly determine whether a primary data owner existed. However, primary data ownership is not the only difference between asymmetric and symmetric linkage archetypes. These archetype groups also differ in how data is split across parties and who learns the final output. Only one participant explicitly considered who learns the output, suggesting that future guidance should make these additional distinctions more visible.

As shown in Figure 1, the clearest misclassification occurred in Scenario 5: 7 participants selected A1, while only 1 selected the intended answer, A5. The wording may have contributed to this error: "my school" may have implied a primary data owner, while "individuals" may have suggested individual-level output. Scenario 1 also produced A1/A6 confusion, with 4 participants selecting A1 instead of A6; however, this mix-up is less surprising because both archetypes involve shared individuals across datasets.

*Participant backgrounds* showed only small performance differences. Computer science-leaning participants scored slightly higher than data science-leaning participants, but the small sample size limits what can be concluded. Background appeared more related to guidance-use strategy than overall correctness.

Our small CMU sample may not fully represent real PPRL practitioners, though participants had relevant data, computing, or privacy backgrounds. The task also used short written scenarios, while real-world use would involve more complex data-sharing problems. Additionally, participants were not explicitly instructed to choose only one archetype, so responses with multiple selections were counted as incorrect if any selected archetype was wrong; one participant did not answer all six scenarios, so incomplete responses were excluded where applicable. Finally, the diagram prototype was modified after P1 to ensure consistent readability and condense it from two pages to one page, meaning P1 viewed a slightly different version than the remaining participants.

## References

[1] A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, and M. Kantarcioglu. Modern privacy-preserving record linkage techniques: An overview. *IEEE Transactions on Information Forensics and Security*, 16:4966–4987, 2021.

[2] O. Klymenko, O. Kosenkov, S. Meisenbacher, P. Elahidoost, D. Mendez, and F. Matthes. Understanding the implementation of technical measures in the process of data privacy compliance: A qualitative study. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2022.

[3] P. Kühtreiber, V. Pak, and D. Reinhardt. A survey on solutions to support developers in privacy-preserving iot development. *Pervasive and Mobile Computing*, 85:101656, 2022.

[4] I. Ngong, B. Stenger, J. Near, and Y. Feng. Evaluating the usability of differential privacy tools with data practitioners. In *Symposium on Usable Privacy and Security (SOUPS)*, 2024.

[5] K. Renaud and L. Shepherd. How to make privacy policies both gdpr-compliant and usable. In *2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*. IEEE, 2018.

[6] D. Vatsalan, P. Christen, and V. S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

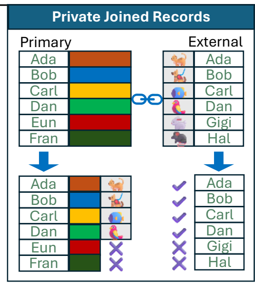
# A Study Materials

## A.1 PPRL Guidance Prototype - Diagrams

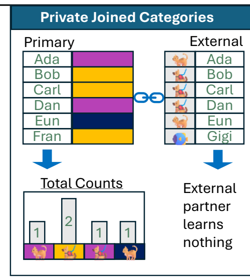
### Privacy Preserving Record Linkage (PPRL) — Diagrams and Definitions

*The diagrams use a "toy" example setup, with names representing individuals and colors representing groups of those individuals. Icons or other data represent additional information about the corresponding individual or group.*

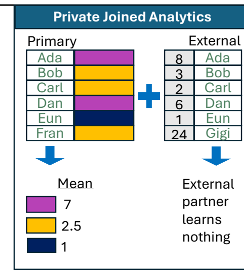
#### Asymmetric Linkage Archetypes:




**Definition:** The primary data owner wants to augment their records with information from an external partner. The primary data owner should only learn new information about people they share with the external data owner, and the external partner should learn as little as possible.

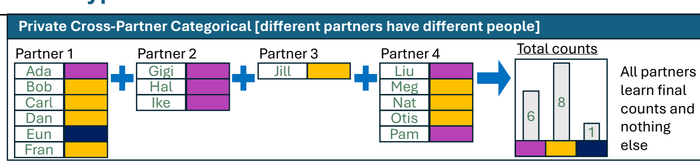



**Definition:** The primary data owner needs to make a histogram (bar chart) that combines information they have with information their external partner has. Neither side is allowed to learn the individual-level matches: the primary data owner should only learn aggregated (bar chart) information, and the external partner should learn nothing.

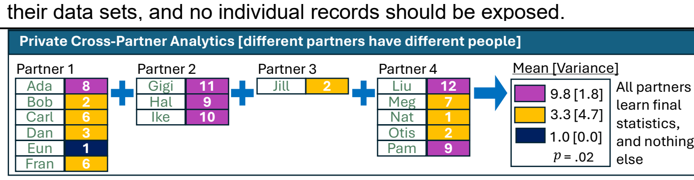



**Definition:** The primary data owner needs to collect analytics (averages, min, max, variance) about groups of people in their data. Neither side is allowed to learn the individual-level matches: the primary data owner should only learn aggregated (analytics) information, and the external partner should learn nothing.

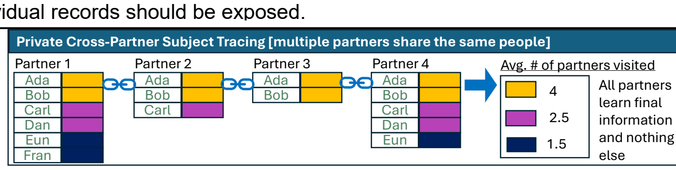
#### Symmetric Linkage Archetypes:




**Definition:** Multiple data owners each have too little data to analyze, so they want to work together as equal partners to make a histogram (bar chart) that combines information across all their data sets. We assume different partners have different individuals in their data sets, and no individual records should be exposed.



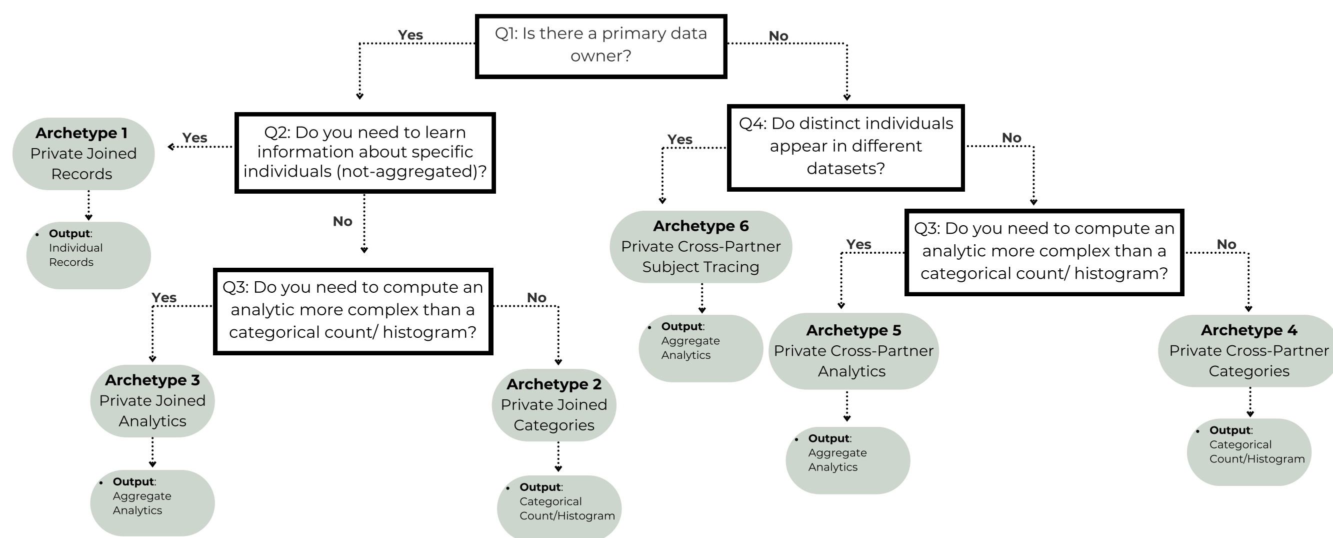

**Definition:** Multiple data owners each have too little data to analyze, so they want to work together as equal partners to compute analytical information across all their data sets. We assume different partners have different individuals in their data sets, and no individual records should be exposed.




**Definition:** Multiple data owners have overlapping data (the same individuals appear in data owned by different partners) and they want to work together to understand how individuals move between their data sets. Individuals are linked and tracked across partners, but only aggregate statistics on outcomes, events, correlations, are released. No individual data is exposed.

## A.2 PPRL Guidance Prototype - Flowchart

Privacy Preserving Record Linkage (PPRL) — Flowchart




**Q1** - Answer ‘Yes’ if **one** primary partner/data owner is collecting data from **one** external partner and only the primary partner/data owner receives the final output. Answer ‘No’ if multiple partners (**two or more**) contribute data and have equal access to the results.

**Yes**: One pet store collects data from another pet store and keeps the results.

**No**: Multiple pet stores combine their data and all receive the results.

**Q2** - Answer ‘Yes’ if individual-level data is important. Answer ‘No’ if aggregated data across unidentified individuals is sufficient.

**Yes**: A dataset is combined/joined with another dataset

**No**: A dataset is aggregated to find the average number of dog toys bought

**Q3** - A simple histogram is created by organizing data into categories and counting each category. Analytics requiring further computation include means, regression coefficients, correlations, etc). Answer ‘Yes’ if computing an analytic. Answer ‘No’ if a categorical count/histogram is adequate.

**Yes**: Finding the mean age of dogs at a shelter

**No**: Finding the count of dogs per age bracket

**Q4** - Answer ‘Yes’ if the analysis requires tracing individuals across multiple datasets (following an individual through different institutions). Answer ‘No’ if aggregate-level pooling is sufficient, and individual-level linkage is unnecessary.

**Yes**: Following dogs through different veterinary institutions to study multi-year patterns

**No**: Counting how many dogs across veterinary institutions have had eye infections

## A.3 PPRL Guidance Prototype - Matrix

**Privacy Preserving Record Linkage (PPRL)** — Data Sharing Archetype Matrix

| Type of Problem (Archetype) | Goal & Definition | Participating Parties | Population Overlap | Inputs | Output | Example Scenario |
|---|---|---|---|---|---|---|
| ◆ **ASYMMETRIC LINKAGE ARCHETYPES**<br>• **Vertically split data** - *each party has a dataset with different types of information*<br>• **Two partners** - *primary and external data owner*<br>• **One party learns the output** | | | | | | |
| **Private Joined Records**<br>Archetype 1 | **Primary data owner wants to augment their records with information from an external partner.**<br>The information being learned by the primary data owner pertains to specific individuals; essentially an additional row of data is added. | **2 Partners**<br>Primary & External | **Yes**<br>Linking records on the shared individuals | **Vertically split datasets:** Each partner has a dataset with different information about some of the same individuals.<br>**Any data type** | **The primary data holder learns:**<br>• **The intersection** *(which members exist in both the primary and external datasets)*<br>• **Individual records** *(reveals complete personal information for linked individuals)*<br>**The external partner learns nothing.*** | *An insurance company and a driving app want to work together to provide a discount to an individual if they are a good driver.* |
| **Private Joined Categories**<br>Archetype 2 | **Primary data owner wants to learn aggregate categorical information across shared individuals without revealing individual information.** | **2 Partners**<br>Primary & External | **Yes**<br>Linking records on the shared individuals | **Vertically split datasets:** Each partner has a dataset with different information about some of the same individuals.<br>**Categorical data only** *(ex: sex, movie genre, color, school)* | **The primary data holder learns:**<br>• **Aggregate categorical counts** *(ex: collections of tallies, pie charts, percentages)*<br>**The external partner learns nothing.** | *A streaming service and a bookstore want to know if a user is likely to buy the book after watching the movie adaptation without revealing individual purchases or viewing histories.* |
| **Private Joined Analytics**<br>Archetype 3 | **Primary data owner wants to compute aggregate analytical information derived from shared individuals without revealing individual information.** | **2 Partners**<br>Primary & External | **Yes**<br>Linking records on the shared individuals | **Vertically split datasets:** Each partner has a dataset with different information about some of the same individuals.<br>**Includes numerical data** *(ex: age, income, test score)* | **The primary data holder learns:**<br>• **Aggregate analytics** *(ex: means, medians, regressions, significance testing)*<br>**The external partner learns nothing.** | *A bank and a retail chain calculating the average credit score of individuals who purchased a specific brand of television.* |
| ◆ **SYMMETRIC LINKAGE ARCHETYPES**<br>• **Horizontal split** - *each party has a dataset with the same types of information*<br>• **Multiple partners**<br>• **Everyone learns the output** | | | | | | |
| **Private Cross-Partner Categorical**<br>Archetype 4 | **Multiple data owners want to work together to learn aggregate categorical information without revealing individual information.** | **Multiple Partners** | **No**<br>Different partners have distinct populations | **Horizontally split datasets:** Each partner has a dataset with the same information about different individuals.<br>**Categorical data only** *(ex: sex, movie genre, color, school)* | **All parties learn:**<br>• **Aggregate categorical counts** *(ex: collections of tallies, pie charts, percentages)* | *Multiple independent bookstores pooling their sales data to determine the most popular book genre across the state, without sharing individual store metrics.* |
| **Private Cross-Partner Analytics**<br>Archetype 5 | **Multiple data owners want to work together to compute aggregate analytical information without revealing individual information.** | **Multiple Partners** | **No**<br>Different partners have distinct populations | **Horizontally split datasets:** Each partner has a dataset with no overlapping individuals.<br>**Includes numerical data** *(ex: age, income, test score)* | **All parties learn:**<br>• **Aggregate analytics** *(ex: means, medians, regressions, significance testing)* | *Several technology companies calculating the industry-wide average salary for software engineers by sex, without disclosing any single company's payroll data.* |
| **Private Cross-Partner Subject Tracing**<br>Archetype 6 | *Shared Subject Tracing*<br>**Multiple data owners want to identify aggregate patterns, correlations, or sequences of events for shared individuals across multiple datasets.** | **Multiple Partners** | **Yes**<br>Tracing the exact same individuals across different partners | **Horizontally split datasets:** Each partner has a dataset **with overlapping individuals**<br>**Includes numerical data** *(ex: age, income, test score)* | **All parties learn:**<br>• **Aggregate analytics** *(ex: means, medians, regressions, significance testing)* | *Multiple websites want to combine data to determine how many times individuals viewed an item's advertisement across various sites before purchasing the item, without disclosing individual browsing or purchasing data.* |

* If a low-privacy technique is used the external partner may learn what individuals exist in both datasets.

## A.4 Real-world Example Scenario Task Sheet

# Activities:

**Real World Use Cases:**

Match each problem below with its archetype problem, by writing the correct archetype on each use case.

| | |
|---|---|
| 1. How many visits does it typically take for a patient to be diagnosed with disease X? I know about visits at my hospital, but I don't know about visits at other hospitals. | 2. Is a particular virus variant more prominent among homeless populations? I want to know the overall infection pattern, but I shouldn't learn which of my patients are homeless. |
| 3. How are student outcomes at my school correlated with income? I can't learn my students' incomes. | 4. There's a virus variant spreading to patients that use the town sewer system. We could alert them, but we don't know which individual patients are at risk. The city knows which households use the sewer. |
| 5. What are the average standardized test scores for students who receive supplemental instruction? My school only has two individuals who receive supplemental instruction, and we can't share student data with other schools. | 6. How effective is my job training program across different demographic groups? The classes shouldn't directly share student data, and classes may only have 1 or 2 students from some demographic groups. |

# B Results

**Archetype Assignments per Scenario**

| | S1 (Correct = A6) | S2 (Correct = A2) | S3 (Correct = A3) | S4 (Correct = A1) | S5 (Correct = A5) | S6 (Correct = A4 or A5) |
|---|---|---|---|---|---|---|
| **A1** | 4 | 0 | 0 | **7** | 7 | 1 |
| **A2** | 1 | **7** | 0 | 2 | 2 | 1 |
| **A3** | 2 | 3 | **12** | 0 | 0 | 2 |
| **A4** | 3 | 2 | 1 | 0 | 0 | 3 |
| **A5** | 1 | 1 | 0 | 1 | **1** | 5 |
| **A6** | **1** | 0 | 0 | 3 | 3 | 2 |

Figure 1: Participant archetype assignments across the six task scenarios. Green cells indicate correct assignments. Red cells highlight the most common incorrect assignments. Numbers show how many participants selected each archetype for each scenario.

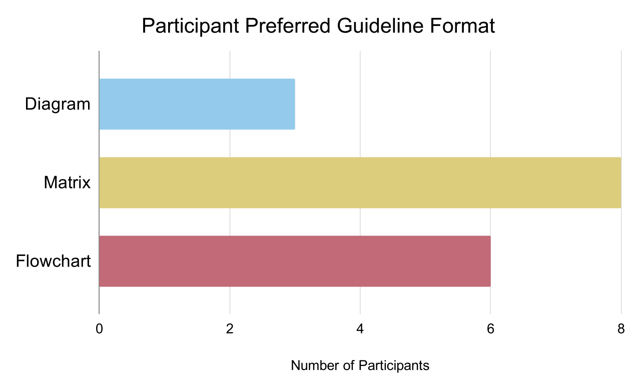



Figure 2: Preferred guidance formats. Counts include participants who had more than one preferred format.

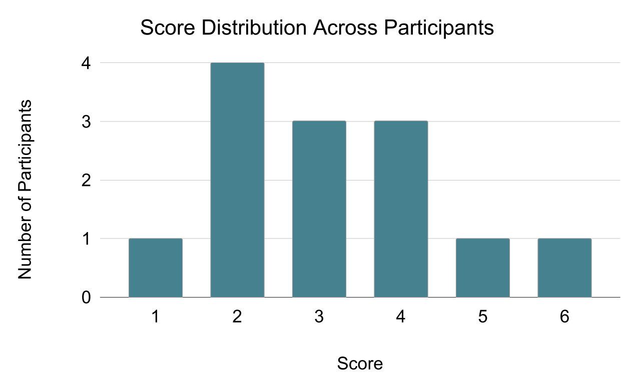



Figure 3: Participant score distribution across the six scenario-matching tasks.

## C Participant Demographics

| Participant | Occupation | Experience with Data Science/ Statistics | Experience with Computer Science | Experience with Privacy | DS or CS |
|---|---|---|---|---|---|
| P1 | PhD | Practical Experience | Basic Familiarity | Basic Familiarity | DS |
| P2 | Professional | Practical Experience | Practical Experience | Practical Experience | DS |
| P3 | Undergraduate | Basic Familiarity | Practical Experience | Basic Familiarity | CS |
| P4 | Masters | Basic Familiarity | Basic Familiarity | Basic Familiarity | DS |
| P5 | Masters | Basic Familiarity | Practical Experience | Practical Experience | CS |
| P6 | Undergraduate | Practical Experience | Basic Familiarity | No Experience | DS |
| P7 | Undergraduate | Basic Familiarity | Basic Familiarity | No Experience | DS |
| P8 | PhD | Practical Experience | Practical Experience | Practical Experience | DS |
| P9 | Masters | Practical Experience | Practical Experience | Basic Familiarity | CS |
| P10 | Undergraduate | Basic Familiarity | Basic Familiarity | No Experience | DS |
| P11 | Masters | Basic Familiarity | Practical Experience | Practical Experience | CS |
| P12 | PhD | Practical Experience | Practical Experience | Practical Experience | DS |
| P13 | Masters | Basic Familiarity | Practical Experience | Basic Familiarity | CS |

## D Codebook

### D.1 Features Considered Per Scenario

| Code | Description | Example |
|---|---|---|
| Data type | The participant considers the type of data involved in the scenario. | "test scores is an average, not categorical"; "do people have a job year out or not, in which case it would be like categorical" |
| Parties involved | The participant considers the number of or which parties are involved in the scenario. | "it sounds like there's just one primary data owner"; "I feel like there are multiple parties involved"; "different classes are different partners" |
| Shared individuals | The participant considers whether the individuals exist across multiple datasets. | "each person is at only one school"; "do people appear in multiple classes?" |
| Who learns what | The participant considers what parties learn what information. | "it's not entirely clear to me if they have equal access to the results" |
| Who has what | The participant considers what information each party has. | "the hospital knows which patients are infected [...] whereas the government knows which all households are connected" |
| Symmetric vs. asymmetric | The participant considers whether the scenario sounds asymmetric or symmetric. | "we know that this will be symmetric and not asymmetric"; "this does sound like asymmetric" |
| Individual vs. aggregate | The participant considers whether the scenario requires information on specific individuals or aggregate information. | "maybe I do need to know about specific individuals"; "I don't care about like the individual level data" |
| Method used in scenario | The participant considers what method may be used in the scenario to obtain the desired output. | "assume you're comparing a baseline"; "we're trying to get a frequency of virus among homeless versus housed" |
| No additional features considered | Participant does not indicate consideration of any particular features. | Participant only answered flowchart questions or participant did not discuss any relevant features. |

## D.2 Per-Scenario Responses

| Code | Description |
|---|---|
| A1 Private Joined Records | The participant selected *Archetype 1: Private Joined Records* for the respective scenario. |
| A2 Private Joined Categories | The participant selected *Archetype 2: Private Joined Categories* for the respective scenario. |
| A3 Private Joined Analytics | The participant selected *Archetype 3: Private Joined Analytics* for the respective scenario. |
| A4 Private Cross-Partner Categories | The participant selected *Archetype 4: Private Cross-Partner Categories* for the respective scenario. |
| A5 Private Cross-Partner Analytics | The participant selected *Archetype 5: Private Cross-Partner Analytics* for the respective scenario. |
| A6 Private Cross-Partner Subject Tracing | The participant selected *Archetype 6: Private Cross-Partner Subject Tracing* for the respective scenario. |