Differential Privacy

Privacy Integrated Queries:

Frank McSherry

Atlas Yu, Casey Nelson

Presentation Outline

- PINQ Overview
    - Adversarial model
    - Contributions
- Differential Privacy
    - What is it?
    - What problems does it solve?
    - How is it defined mathematically?
        - High level
- PINQ

Contributions

- PINQ: language for making transformation and aggregate computations differentially private
- Features:
  - Implemented to work in C # programs
  - Based off of LINQ
- Benefits:
  - Provide a platform which is easy to use for non-experts in privacy
  - Provide a way to safely perform calculations over data sets which were previously believed to threaten privacy
  - Easily integrated into existing applications
  - Adjustable Privacy

Adversarial Model:

- Does not aim to prevent data leaks entirely
- Does not protect from any targeted attacks on a database directly aiming to disclose information
- **The goal of PINQ is to assure that an individuals participation in a dataset is not what causes a leak**

Intuitive Definitions:

Privacy:

- Given aggregate statistics on some data set, you should not be able define the attributes of any particular subject within that set

Privacy Loss:

- Privacy is lost when one can successfully narrow down the possible participant makeups to a small pool of possibilities
  - 1 possibility = no privacy

How privacy is lost:

- Brute forcing all possible combinations of subject makeups that could lead to the published statistics
- Example:
  - “Out of a sample of 5 people, 2 liked chocolate, 2 were women, and 2 were under the age of 16”
  - **If you could make any arbitrary query to this data set, how could you de-anonymize individuals in this group building off of these statistics?**

Solutions:

- Add random noise to every computation performed on a data set
    - **What could go wrong with this?**

Solutions:

- Add random noise to every computation performed on a data set
  - What could go wrong with this?
    - Statistic can still be averaged across multiple instances in which it is used to get the original value
- Differential Privacy
  - Determine the amount of noise that can provide the most accuracy while preserving privacy
  - Multiple noises can’t be combined to recover the true statistics

Mathematical Definition:

- Computations with and without a given data point are indistinguishable
- Specifically: *We say a randomized computation M provides* ***ε****-differential privacy if for any two data sets A and B, and any set of possible outputs S ⊆ Range(M):*

$$\mathbf{Pr}[M(A) \in S] \quad \leq \quad \mathbf{Pr}[M(B) \in S] \times \exp(\epsilon \times |A \oplus B|)$$

Mathematical Definition:

$$\mathbf{Pr}[M(A) \in S] \quad \leq \quad \mathbf{Pr}[M(B) \in S] \times \exp(\epsilon \times |A \oplus B|)$$

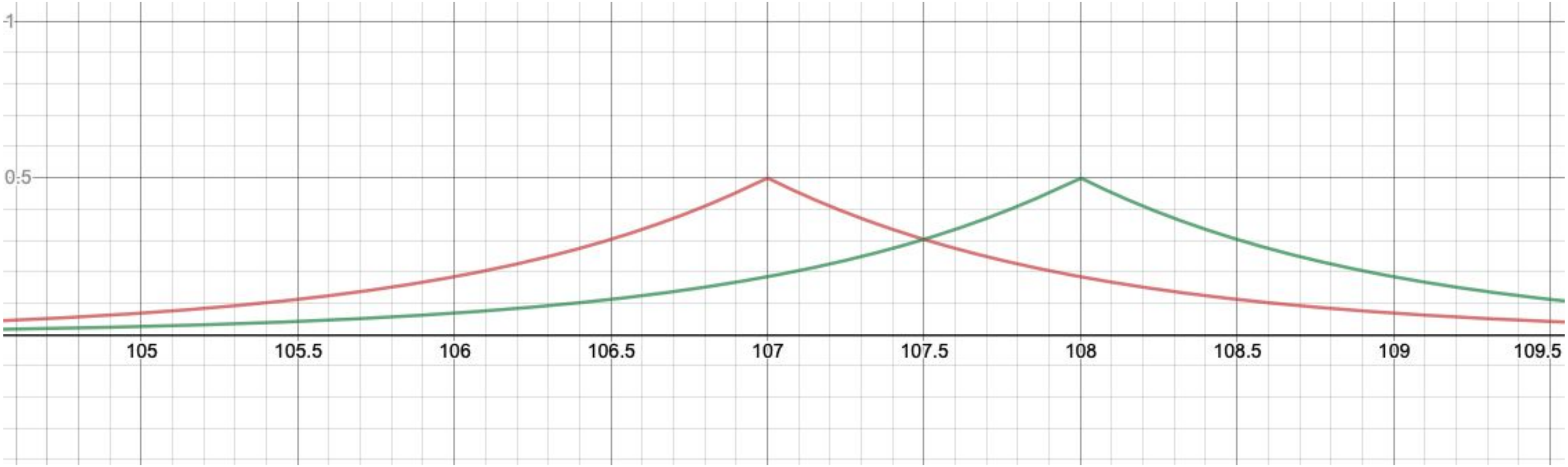
1. When A and B are identical (their difference is 0) the distributions of the computation M on these sets will be identical
2. When A and B differ by only 1 entry, the difference in their distributions under M is in terms of **ε**
    - Intuitively: **ε** represents the difference in output per differing input element

Noisy Counts

- **Main Idea:** add Laplace noise to jitter statistics
- **Theorem 1:**
  - *The mechanism M(X)=|X|+Laplace(1/ε) provides ε-differential privacy.*
  - A differential privacy count mechanism can be obtained by adding a value from the 1/ε Laplace distribution
- **Generally:**
  - *The mechanism M(f(x))=f(x)+Laplace(s/ε) provides ε-differential privacy*
    - s = “sensitivity” of f(x)

Noisy Counts

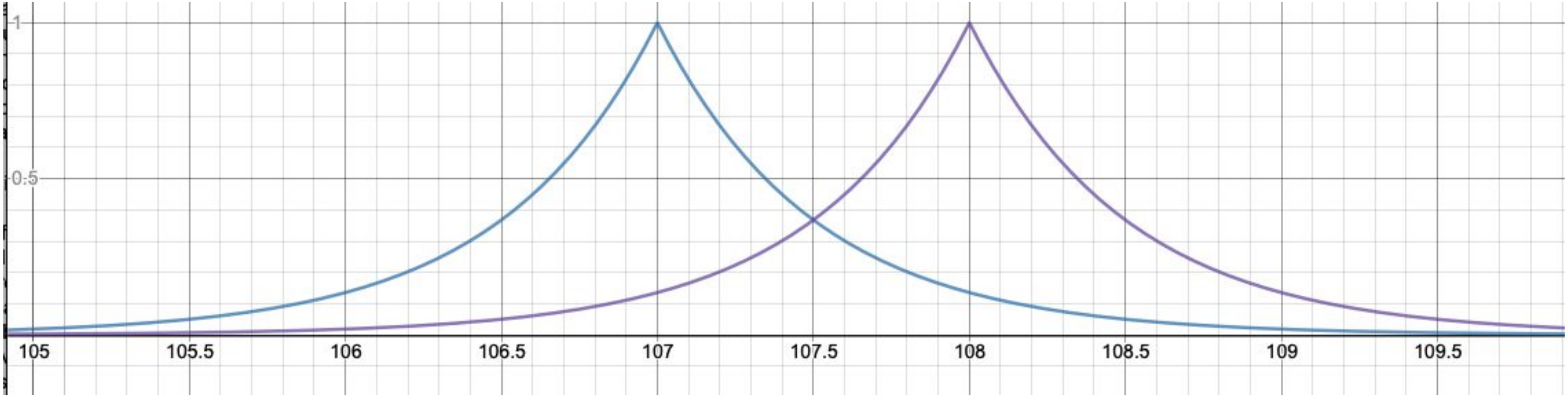
ε = 1



How would you expect this graph to change if we increase **ε?**

Noisy Counts

ε = 2



Stable Transformations

- Transformations that can be applied to a data set while still allowing differentially private computation
- C-Stable Transformation:
  - *We say a transformation T is c-stable if for any two input data sets A and B*

$$|T(A) \oplus T(B)| \quad \leq \quad c \times |A \oplus B|$$

- Why does this matter?
  - Allows the level of differential privacy of a computation over a transformation to be defined
  - Specifically: c*ε-differential privacy (theorem 2)

Stable Transformations

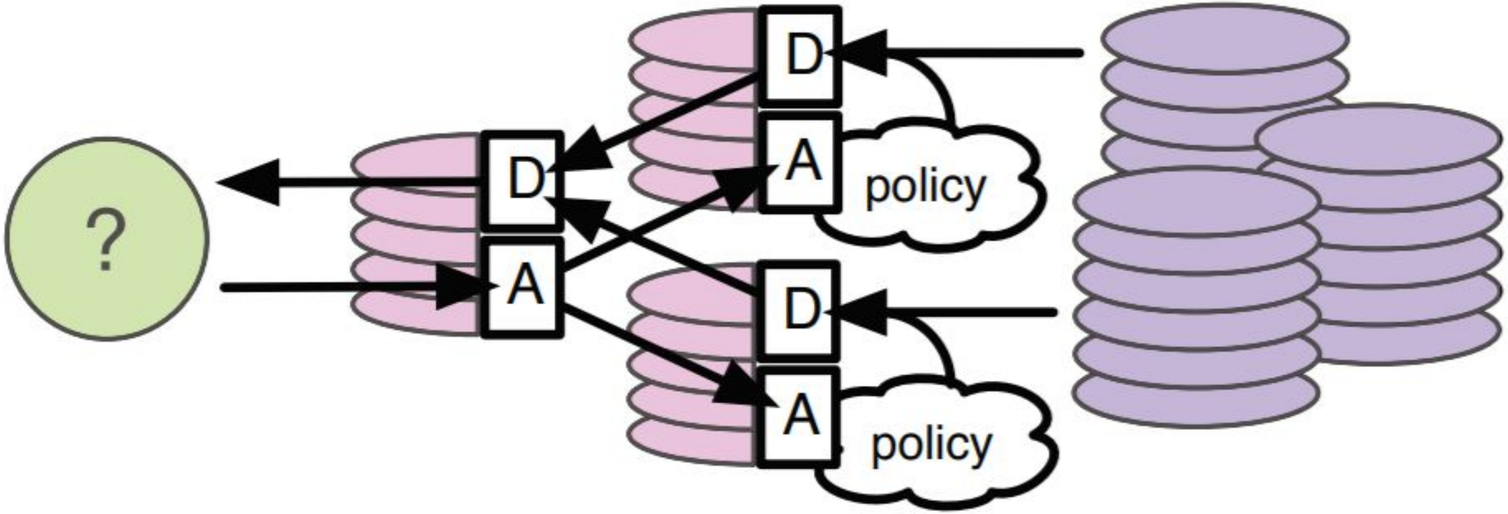
- Some transformations allowed by PINQ:

Transformation:	C-Stability:
WHERE	1
SELECT	1
GROUPBY	2
JOIN*	2*

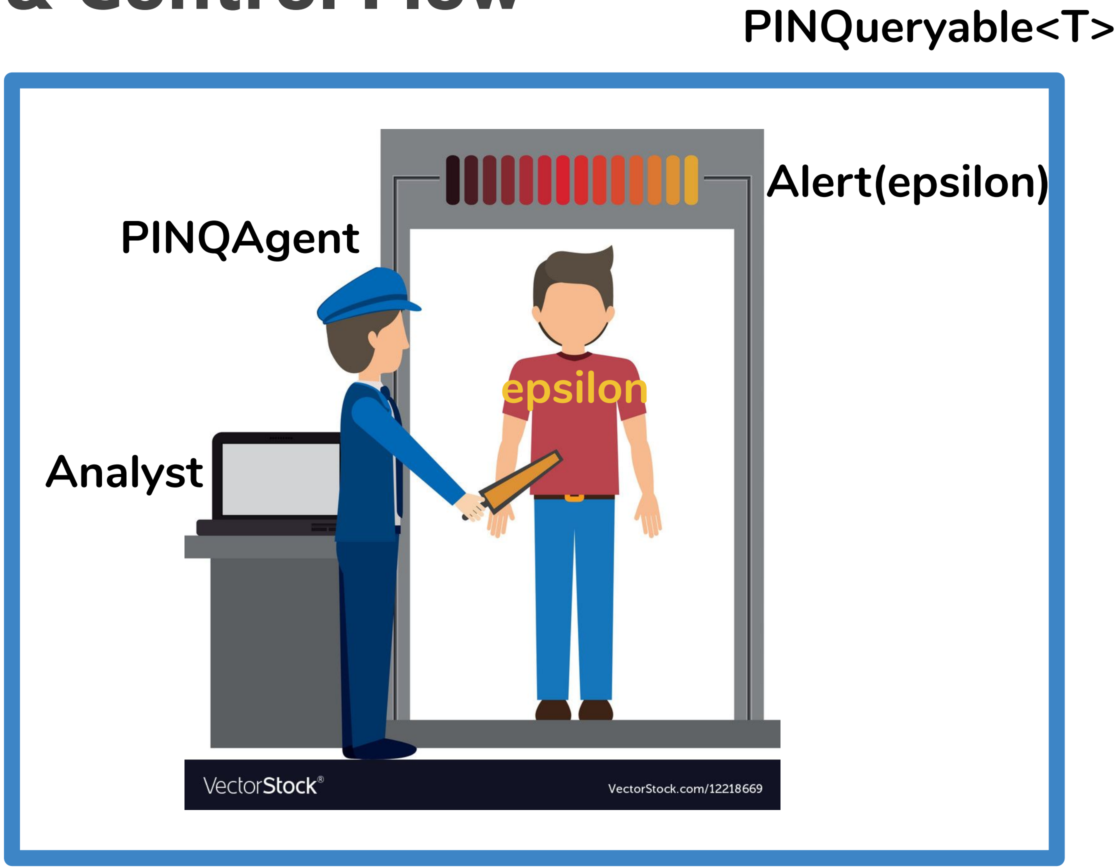
Composition

- Sequential Composition
  - Differentially private computations can be chained in sequence and still preserve differential privacy
  - Net ε defining the level of privacy of the chain of computations is the sum of the ε of each sub computation
- Parallel Composition
  - Differential privacy is also preserved in parallel computation
  - Net ε is the worst privacy guarantee of the parallel parts

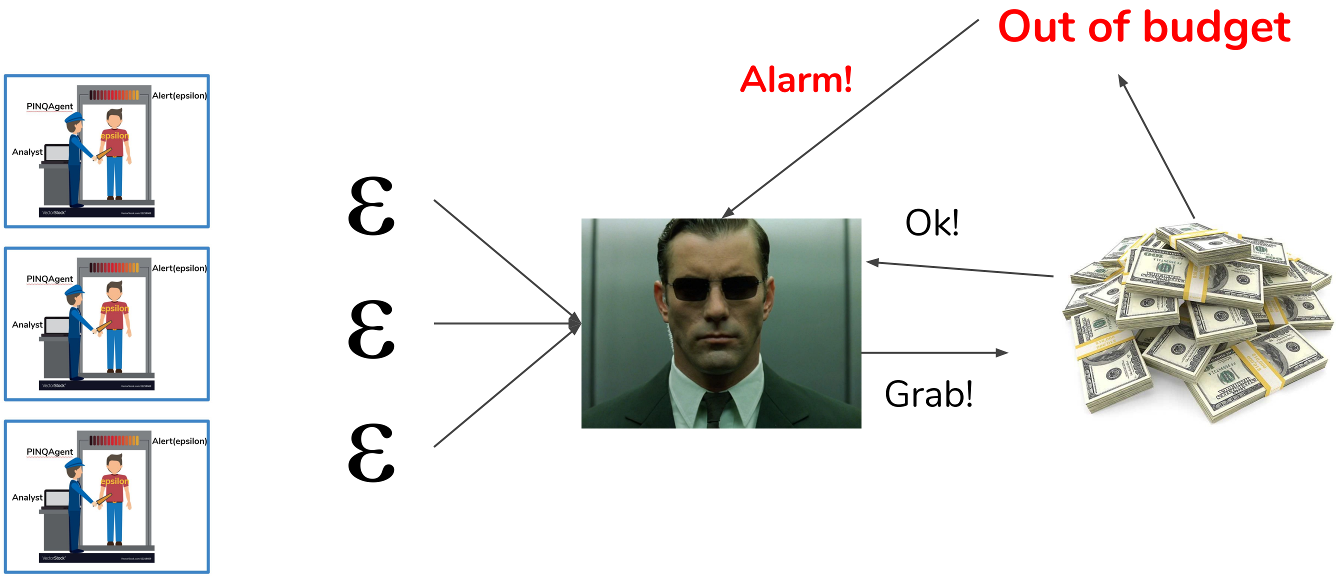
How does PINQ work



Data & Control Flow



Aggregation



ε

ε

ε

Alarm!

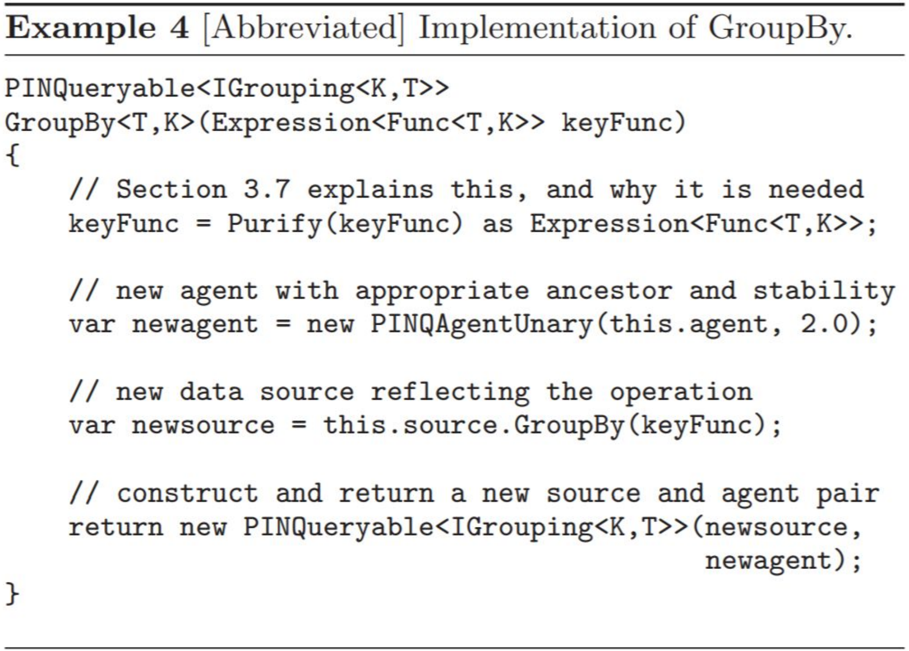
Out of budget

Ok!

Grab!

•••

Transformation



Example 4 [Abbreviated] Implementation of GroupBy.

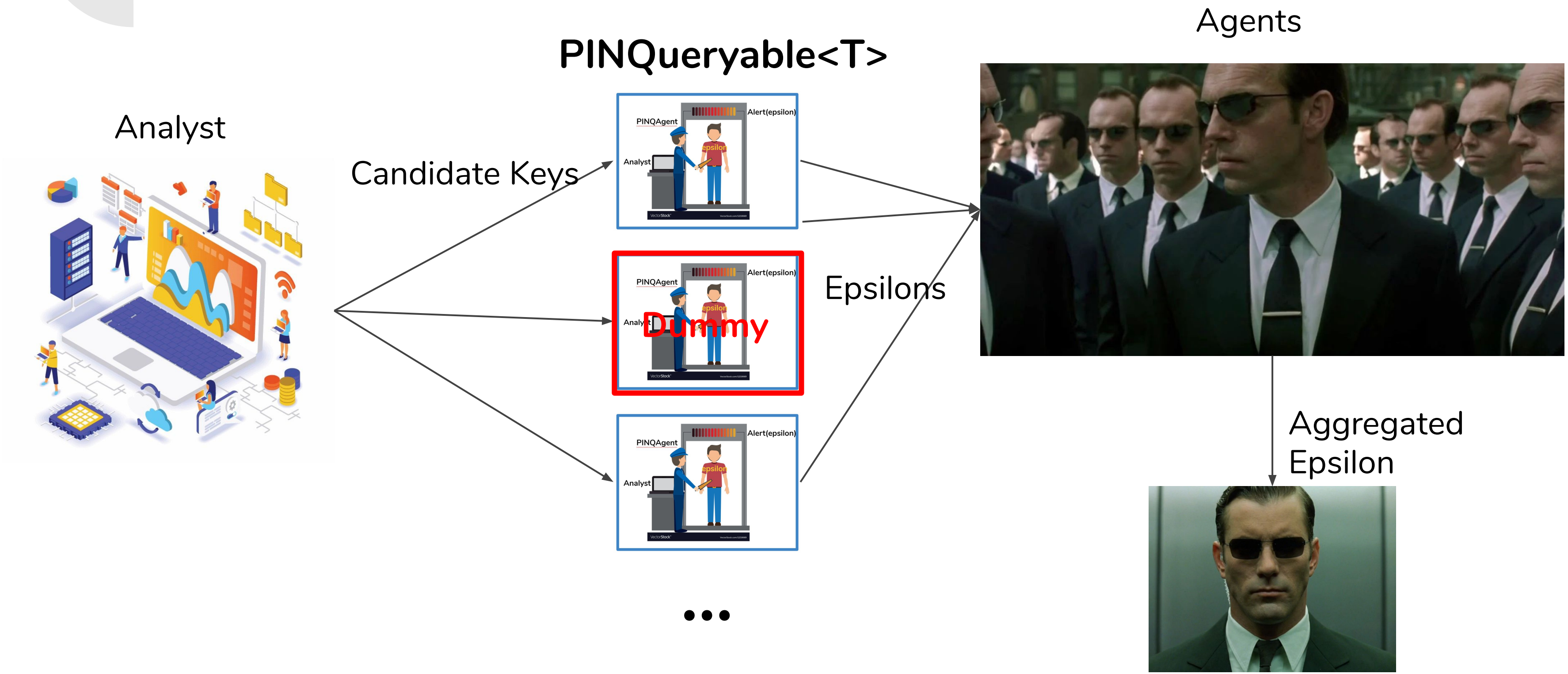
```
PINQueryable<IGrouping<K,T>>
GroupBy<T,K>(Expression<Func<T,K>> keyFunc)
{
    // Section 3.7 explains this, and why it is needed
    keyFunc = Purify(keyFunc) as Expression<Func<T,K>>;

    // new agent with appropriate ancestor and stability
    var newagent = new PINQAgentUnary(this.agent, 2.0);

    // new data source reflecting the operation
    var newsource = this.source.GroupBy(keyFunc);

    // construct and return a new source and agent pair
    return new PINQueryable<IGrouping<K,T>>(newsource,
                                            newagent);
}
```

Partition



PINQueryable<T>

Analyst

Candidate Keys

Epsilons

Agents

Aggregated Epsilon

…

Question

How can PINQ help online-ads platforms better protect user privacy and obey GDPR rules?

Additional Question

What’s the tradeoff between privacy and accuracy?

When is it acceptable to lower the accuracy of statistics to preserve an individual’s privacy?

When is it acceptable to loosen privacy constraints for more accurate statistics?

Question

Is PINQ really easy to use for those without expertise in privacy?

For example, many of you mentioned in the reading quiz that you weren’t sure how to choose epsilon for question 1. Do you see this as a real problem non-expert users would face?

Addition Questions?