О вопросах безопасности применения современных математических методов обезличивания при интеллектуальной обработке данных

Что такое обезличивание?

- Персональные данные - любая информация, относящаяся к прямо или косвенно определенному или определяемому физическому лицу

- Обезличивание персональных данных - действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных

Формализация

Атрибут

Человек

Имя	Рост	Вес	Возраст
Кирилл	176	80	35
Анна	160	60	20
Антон	190	100	45
Вера	163	65	35

Угрозы безопасности

- *Нарушение целостности*
- *Нарушение конфиденциальности*
- *Нарушение доступности*

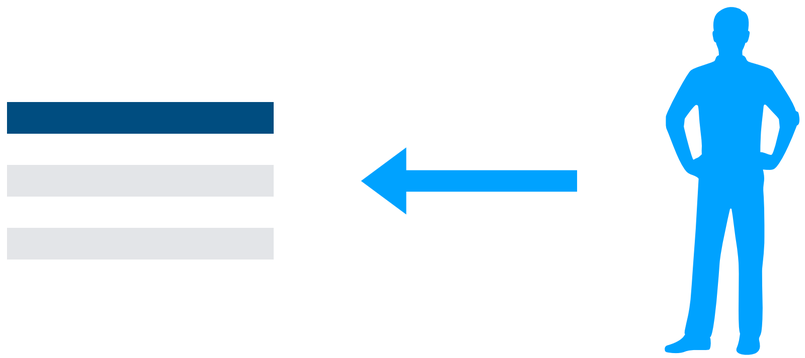
Угрозы безопасности

- *Нарушение целостности*
- *Нарушение конфиденциальности*
- *Нарушение доступности*

- *Выделение субъекта*
- *Связывание данных*
- *Определение значений*

Угрозы безопасности

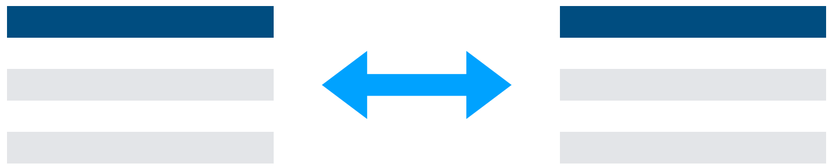
- Из набора данных выделить те записи, которые соответствуют конкретному субъекту



- *Выделение субъекта*
- *Связывание данных*
- *Определение значений*

Угрозы безопасности

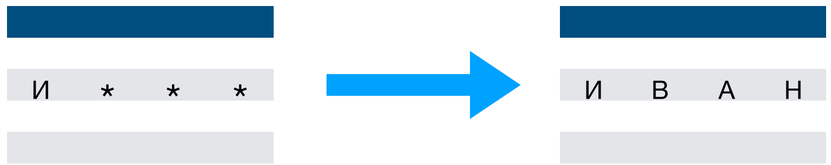
- Выделить из набора данных (например, одной таблицы или двух различных) по крайней мере две записи, соответствующих конкретному субъекту



- *Выделение субъекта*
- *Связывание данных*
- *Определение значений*

Угрозы безопасности

- Определить (возможно с некоторой вероятностью) истинные значения некоторых полей в записях данных



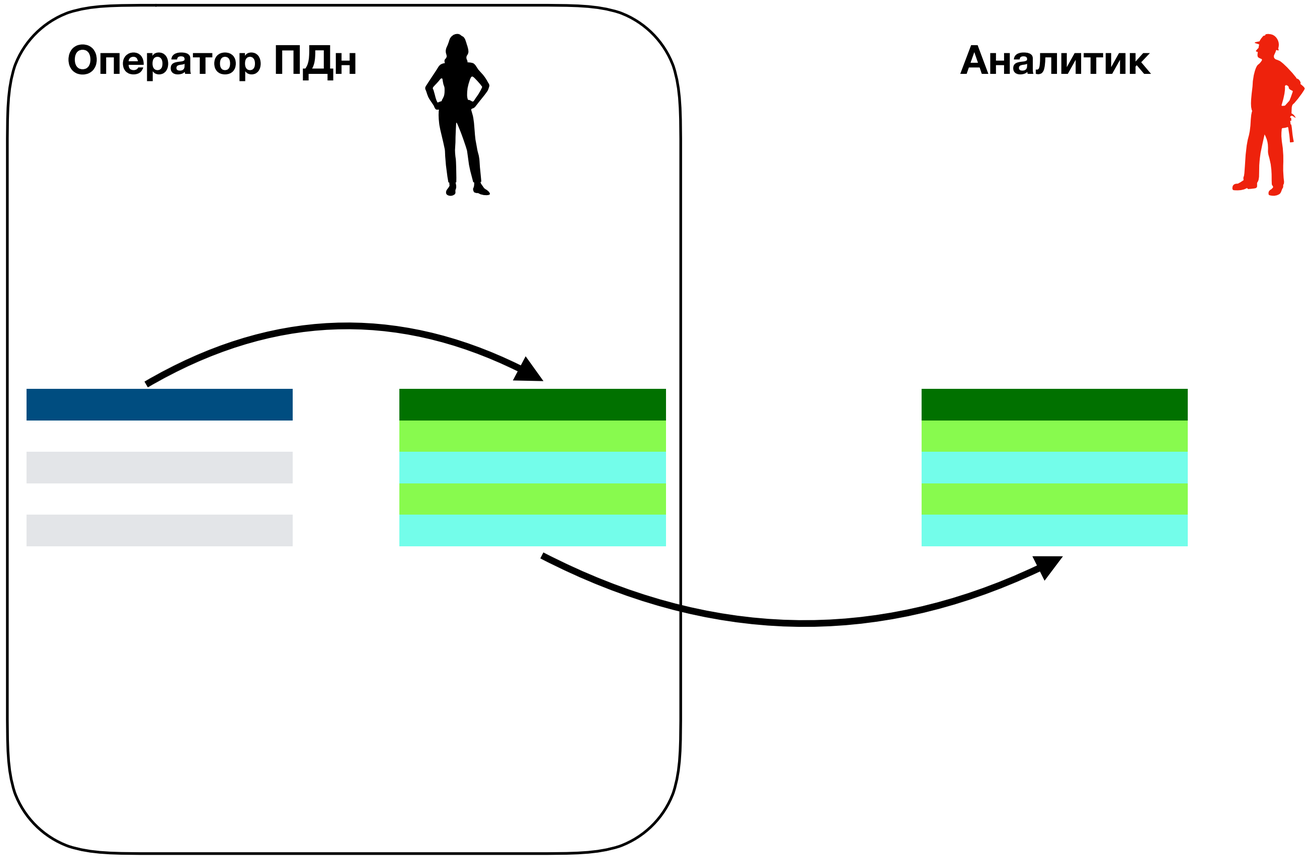
- *Выделение субъекта*
- *Связывание данных*
- *Определение значений*

Нарушитель

- «Честный, но любознательный» аналитик

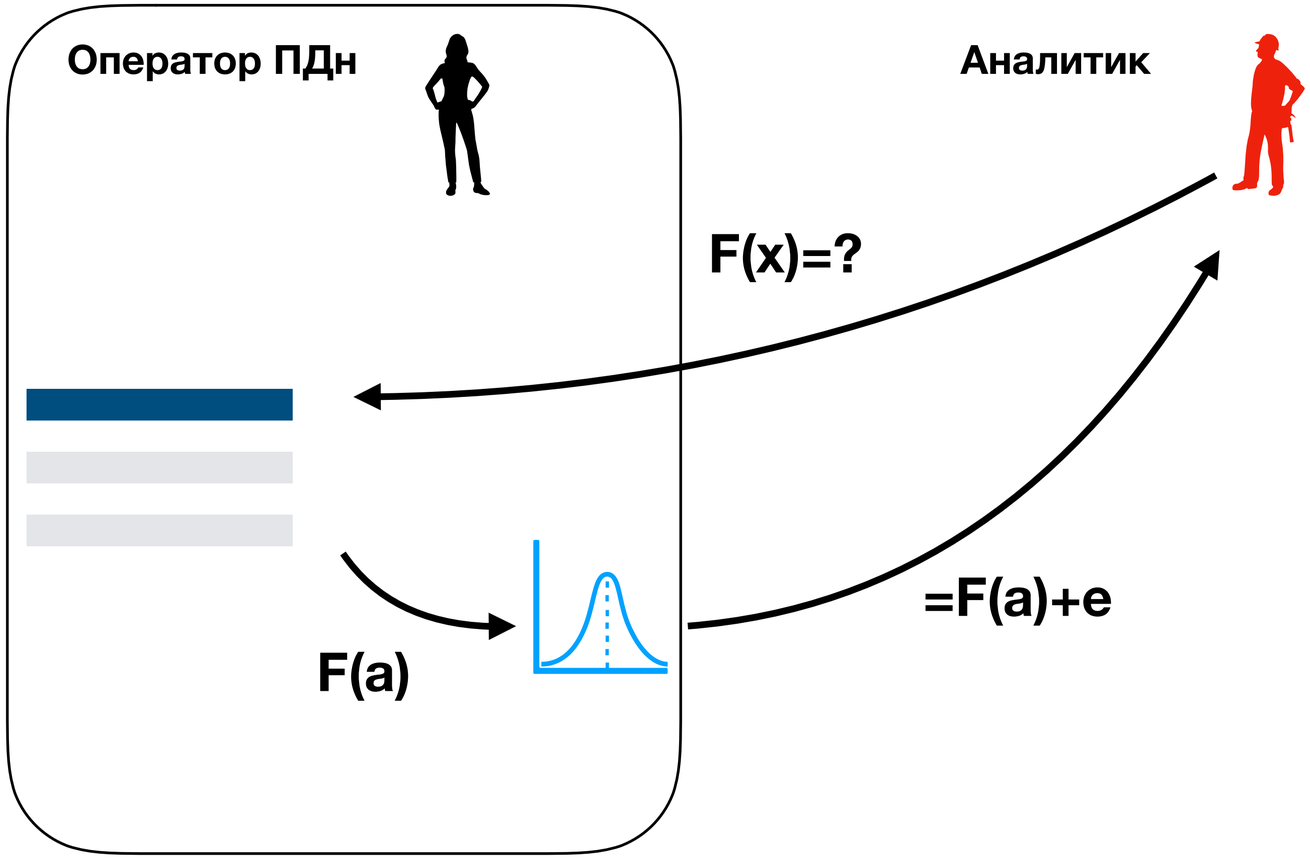
Методы

- *Классические (обработка баз данных)*
- *Сервисные (обработка запросов)*



Методы

- *Классические (обработка баз данных)*
- *Сервисные (обработка запросов)*



Классические методы

Слабый

- *Замена атрибутов на идентификаторы (псевдонимизация)*
- *Перестановки атрибутов*
- *Зашумление*
- *Обобщение/подавление*

ФИО	Паспорт	Дата Рождения
Иванов И.И.	4506 123456	17.06.86
Андреев А.А.	8105 789456	13.01.85
Федоров Ф.Ф.	3703 923675	14.12.83
Николин Н.Н.	7903 345333	01.03.83

ФИО	Паспорт	Дата Рождения
Иванов И.И.	8698472dc	17.06.86
Андреев А.А.	27ffa922f	13.01.85
Федоров Ф.Ф.	c9d53b87d	14.12.83
Николин Н.Н.	7695df95f	01.03.83

- *Малая мощность прообраза*
- *Связывание по идентификаторам*

Классические методы

- *Замена атрибутов на идентификаторы (псевдонимизация)*

Слабый

- *Перестановки атрибутов*
- *Зашумление*
- *Обобщение/подавление*

Год рождения	Должность	Доход
2000	Продавец	200000
1985	Безработный	80000
1996	Программист	12000
1980	Руководитель компании	40000

- *Перестановка коррелированных атрибутов не дает эффекта*
- *Перестановка малозначащего атрибута не дает эффекта*

Классические методы

- *Замена атрибутов на идентификаторы (псевдонимизация)*
- *Перестановки атрибутов*

Средний

- *Зашумление*
- *Обобщение/подавление*

ФИО	Паспорт	Рост
Иванов И.И.	4506 123456	160
Андреев А.А.	8105 789456	165
Федоров Ф.Ф.	3703 923675	190
Николин Н.Н.	7903 345333	180

ФИО	Паспорт	Рост
Иванов И.И.	4506 123456	160,6
Андреев А.А.	8105 789456	165,1
Федоров Ф.Ф.	3703 923675	190,9
Николин Н.Н.	7903 345333	180,3

- *Подбор амплитуды шума*
- *Применим только к числовым данным*

Классические методы

- *Замена атрибутов на идентификаторы (псевдонимизация)*
- *Перестановки атрибутов*
- *Зашумление*

Средний
- *Обобщение/подавление*

3-анонимная база

Имя	Возраст	Пол	Язык
Анна	30	ж	python
Андрей	24	м	java
Дарья	28	ж	c#
Николай	27	м	c++
Антон	24	м	javascript
Светлана	23	ж	c++
Лидия	19	ж	python
Вадим	29	м	javascript
Ирина	17	ж	javascript
Ольга	19	ж	java

Имя	Возраст	Пол	Язык
*	20 - 30	ж	python
*	20 - 30	м	java
*	20 - 30	ж	c#
*	20 - 30	м	c++
*	20 - 30	м	javascript
*	20 - 30	ж	c++
*	<20	ж	python
*	20 - 30	м	javascript
*	<20	ж	javascript
*	<20	ж	java

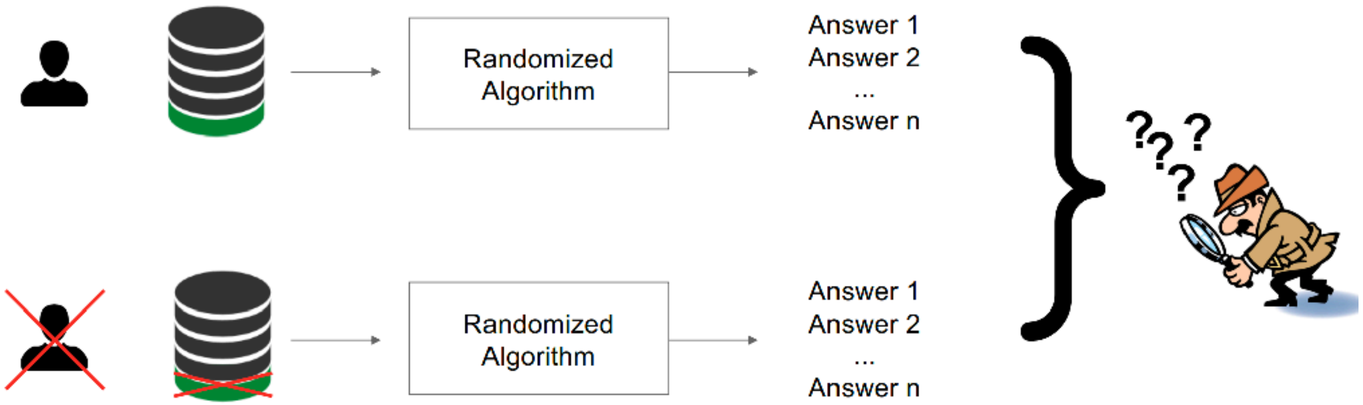
- Вычислительно сложные методы
- Уязвимы к обогащению данных
- Уязвимы к атакам на различные обобщения одной базы

Классические методы

- Классические методы хорошо изучены и, как следствие, уязвимы к различным типам атак
- Главная проблема — невозможность контроля за обезличенной базой
- Ухудшают качество аналитики
- Целесообразны к применению в защищенном контуре, где аналитик ограничен в своих возможностях

Сервисные методы — статистическое обезличивание

- *Метод зашумления*
- *Применим к числовым данным*
- *Подбор параметров шума в зависимости от …*
- *Гарантирует невозможность деобезличвания при …*



Сервисные методы — статистическое обезличивание

- *ε — параметр безопасности*
- *подбирается для:*
  - *каждой статистики*
  - *количества запросов на вычисления этой функции*
- *позволяет оценить ухудшение качества аналитики*

Для произвольной функции $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}$ механизм Лапласа определяется как

$$\mathcal{M}(x, f(), \epsilon) = f(x) + Y,$$

где Y – случайная величина, имеющая распределение $Lap(y|\frac{\Delta f}{\epsilon})$

Шум

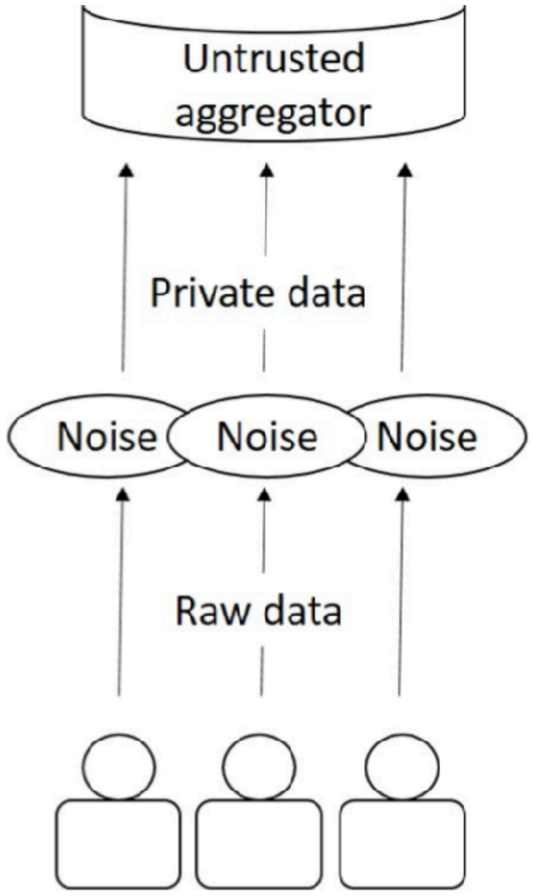
ответ

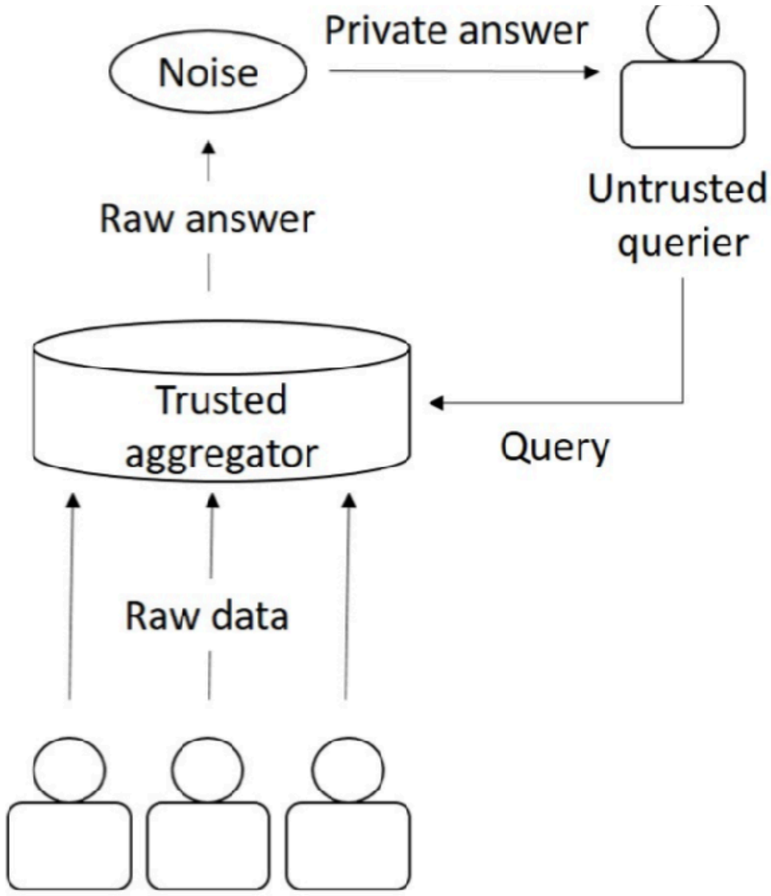
запрашиваемая функция

параметр безопасности

качество аналитики

Статистическое обезличивание — архитектуры





Статистическое обезличивание

- Позволяет получить гарантированную оценку невозможности деобезличивания при ограничениях на количество запросов и параметры шума
- Устойчиво к атакам с обогащением
- Позволяет вычислять безопасность для комбинаций преобразований (построение сложных систем)
- Требует обоснования для каждой статистики
- Применимо только к числовым данным
- Уязвимо к нарушениям условий применения (плохой шум)
- Требует контроля обращений аналитика