АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

Риск-ориентированный подход к конфиденциальной обработке данных

Марат Тахавиев, руководитель GR-проектов,
Ассоциация больших данных

Предпосылки

Потребность бизнеса и государства в данных для разработки и развития продуктов и сервисов

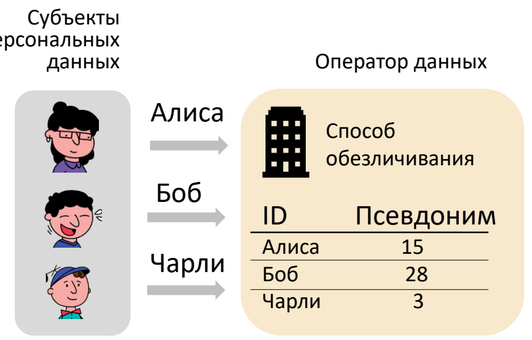
VS

Необходимость гарантий прав граждан на неприкосновенность их частной жизни при обработке данных

Отсутствие универсальных подходов к конфиденциальной обработке, подходящих для всех типов кейсов

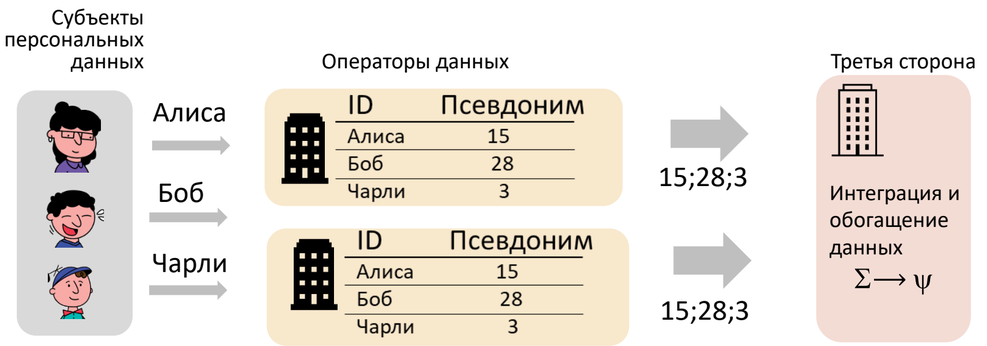
БАЗОВЫЕ СЦЕНАРИИ ОБМЕНА ДАННЫМИ

Сценарий 1. Сценарий передачи данных от пользователей в организацию

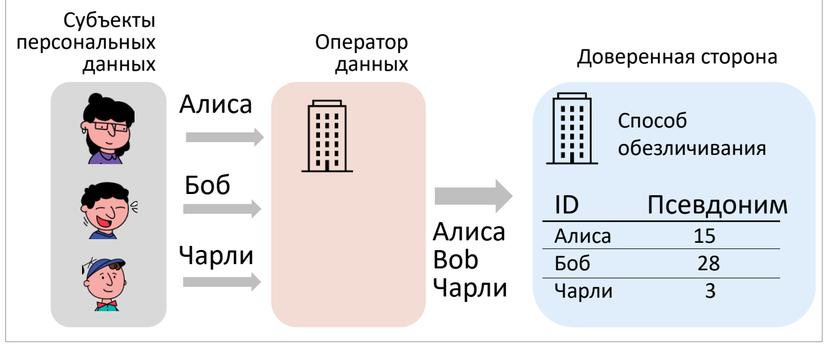


Сценарий 2. Сценарий публикации обезличенных данных третьей стороне

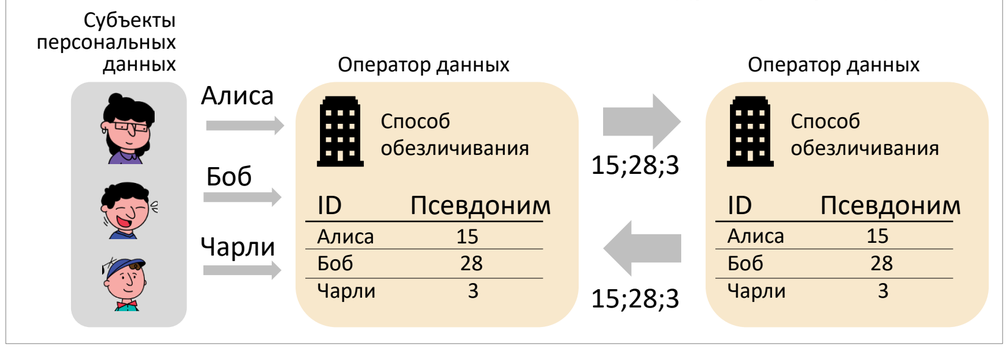
Сценарий 3. Публикация витрин



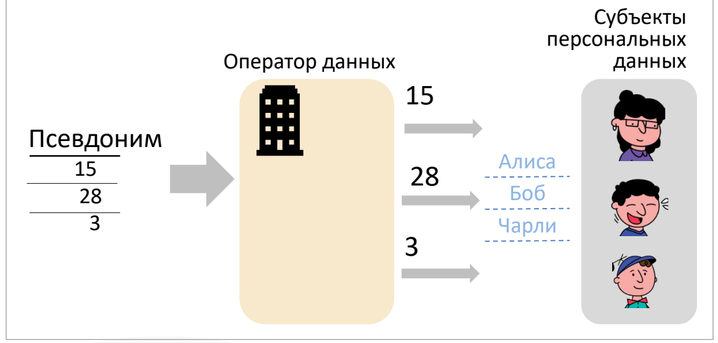
Сценарий 4. Сценарий использования агента обезличивания (доверенной стороны)



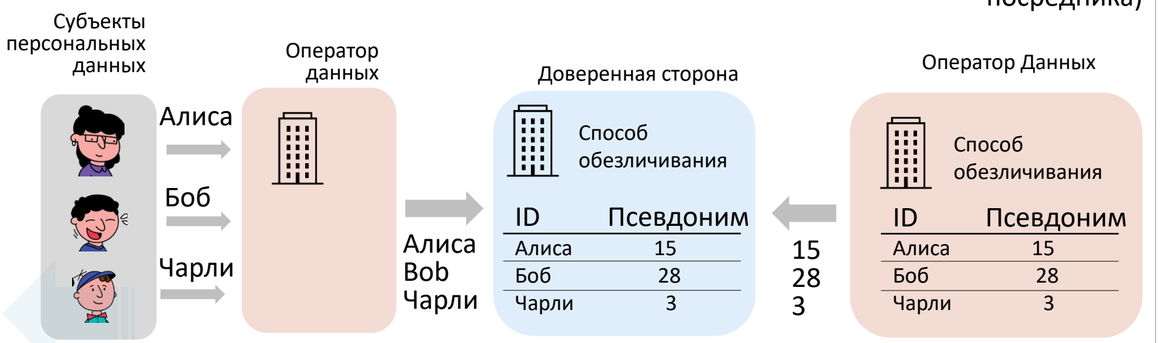
Сценарий 5. Сценарий прямого обмена информацией между операторами данных



Сценарий 6. Сценарий восстановления публичных данных



Сценарий 7. Сценарий обмена информацией через посредника)



ПУБЛИКАЦИЯ / ОБМЕН ДАННЫМИ

1 Количество участников

- 1:0 Только прием данных
- 1:1 Обмен с контрагентом
- 1:M Обмен с несколькими участниками
- 1:∞ Обмен с неопределенным количеством участников

2 Направленность обмена

- Односторонняя публикация
- Двусторонний обмен
- Доверенный посредник

3 Интенсивность обмена

- Разовый (выгрузка)
- Регулярная публикация
- Непрерывный (потоковый)

4 Модель безопасности

- Внутренний контур
- Доверенная среда
- Недоверенная среда
- Публичное раскрытие информации

5 Динамика защиты

- Детерминированные схемы (одинаковы для всех сеансов)
- Частично рандомизированные схемы (изменение на уровне параметров)
- Полностью рандомизированные (включая способы)

ДЕРЕВО РЕШЕНИЙ ПРИМЕНЕНИЯ КОНФИДЕНЦИАЛЬНЫХ ТЕХНОЛОГИЙ

- DLT — Distributed Ledger Technology, Распределенные решения на базе консенсуса
- SMPC — Secure Multiparty Computation – многосторонние конфиденциальные вычисления
- ZKP — Zero-Knowledge Proof– доказательства с нулевым разглашением
- k-A — k-Anonymity – классическая модель обезличивания
- NOTARY — Доверенный централизованный посредник
- FHE — Full Homomorphic Encryption – полное гомоморфное шифрование
- PSD — Pseudonymization - псевдонимизация
- DP — Differential Privacy – дифференциальная приватность
- L-DP — Local Differential Privacy – локальная дифференциальная приватность
- G-DP — Global Differential Privacy – глобальная дифференциальная приватность
- SYNT — Synthetic Data – синтетические данные
- FL — Federated Learning – федеративное обучение

Имеется согласие пользователей на обработку их данных?

ДА → Обмениваетесь данными о ФЛ с третьей стороной или публикуете их?

НЕТ → Получите согласие или забудьте об этом → X

ДА → Достаточно ли выводов или необходимы детальные данные?

НЕТ → Можно ли делать прогноз и вычисления на устройствах пользователей? (или необходима центральная база данных)

ДА → Возможно использование обобщенной информации или необходимы данные конкретного ФЛ?

НЕТ → Необходимы ли персональные сведения в наборе для выходных данных?

ДА → Предполагается использование AI/ML или запросы к базе данных (API)?

НЕТ → Имеется ли доступ к выходным данным или они должны оставаться в секрете?

ДА → Используйте SMPC или общение через третью доверенную сторону

NOTARY, SMPC

НЕТ → Данные структурированы? (поля, атрибуты, а не видео, звук, изображения)

ДА → Агрегация данных + дифференциальная приватность, или k-Anonymity или синтетические данные на основе рисков + Шифрование при передаче и контроль доступа

DP, k-A

НЕТ → Риск-базированная k-Anonymity или синтетические данные на основе рисков + шифрование и контроль доступа

k-A, SYNT

ДА → Предполагается одноразовое прогнозирование или обучение модели на основе данных пользователей?

ДА → Используйте оптимизацию в сложной коммуникационной среде, разделите обработку данных

DLT, NOTARY, FL, SMPC

НЕТ → Используйте федеративное обучение (потенциально с дифференциальной приватностью) или SMPC

НЕТ → Обязательно ли присутствие персональных данных непосредственно в наборе или они могут быть выделены в отдельный набор?

ДА → Шифрование при передаче и хранении, строгий контроль доступа

NOTARY, PSD

НЕТ → Данные структурированы? (или свободный формат, изображение, виде, …)

ДА → K-Anonymity (агрегация) + Дифференциальная Приватность или синтетические данные + шифрование при передаче и контроль доступа

DP, k-A, NOTARY

НЕТ → Псевдонимизация и/или синтетические данные + Шифрование данных при передаче данных и контроль доступа

PSD, SYNT

ДА → Предполагается использование AI/ML или запросы к базе данных (API)?

ДА → Используйте локальное DP или обучение на синтетических данных Рассмотрите Federated Learning

L-DP, SYNT, FL

НЕТ → Дифференциальная приватность или посредник

G-DP, NOTARY

НЕТ → Имеется ли доступ к выходным данным или они должны оставаться в секрете?

ДА → Является ли время обработки данных критичным или возможно ожидание?

ДА → Используйте Trusted Execution Environments (Intel SGX, Keybase) или посредник

TEE, NOTARY

НЕТ → Гомоморфное шифрование

HE

НЕТ → Предоставляет ли другая сторона конфиденциальные входные данные для интеграции или можно выполнить самостоятельные вычисления?

ДА → Допустимы высокие коммуникационные издержки на выполнение рекурсивных или повторяемых алгоритмов?

ДА → Конфиденциальные вычисления (Secure Multiparty Computation)

SMPC, ZKP

НЕТ → Если возможно – соглашение об обработке данных (если возможно), шифрование при передаче и хранении, контроль доступа

NOTARY

НЕТ → k-Anonymity, псевдонимизация, шифрование при передаче, контроль доступа

k-A, PSD

Адаптировано из PRIVATE AI

Разработанный подход

Основной подход к количественной оценке вероятности потери конфиденциальности заключается в разложении на множители вероятности контекстных рисков и вероятности по рискам данных

Уровень риска может быть рассчитан отдельными формулами для каждого семейства методов

$$P_{\text{повторной идентификации}} = P_{\text{контекстные риски}} \times P_{\text{риски данных}}$$

ТЕХНОЛОГИИ ЗАЩИТЫ ПЕРСОНАЛЬНЫХ ДАННЫХ

1 ОРГАНИЗАЦИОННЫЕ И ОПЕРАТИВНО-ТЕХНИЧЕСКИЕ МЕРЫ

Организационные и оперативно-технические меры ведут к управлению контекстными рисками, которые оцениваются с помощью скоринг-модели:

$$P_{\text{контекстные риски}} = \frac{\sum_{j=1}^{n} \omega_j K_j}{\sum_{j=1}^{n} \omega_j}$$

Контекстные риски отражают способность организаций противостоять угрозам утечек данных и поддерживать процессы управления персональными данными, снижая общие угрозы

2 КРИПТОГРАФИЧЕСКИЕ МЕТОДЫ

3 МЕТОДЫ ПСЕВДОНИМИЗАЦИИ

Методы псевдонимизации – методы защиты персональных данных, при которых прямые и/или косвенные атрибуты в конкретных наборах данных заменяются на один или несколько искусственных идентификаторов

$$P_{\text{риски данных}} = \frac{k}{\langle r|R \rangle}$$

При расчете рисков данных в условиях применения методов псевдонимизации величина рисков обратна пропорциональна затраченным ресурсам (например, времени на взлом при атаках прямым перебором)

4 ОБЕЗЛИЧИВАНИЕ (АНОНИМИЗАЦИЯ)

Методы обезличивания нацелены на исключение связи отдельных атрибутов с идентификаторами. Основная идея для данного класса методов: выделении групп схожих записей (классов эквивалентности), которые могут быть отнесены более чем к одному физическому лицу.
В этом случае нарушается основное свойство определения персональных данных – соотнесение информации с прямо *или косвенно определенным или определяемом* физическом лицом.

$$P_{\text{риски данных}} = \frac{k}{\langle \tilde{E} \rangle}$$

Вероятность повторной идентификации обратна пропорциональна размеру наименьшего класса эквивалентности в наборе

5 КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ

В рамках таких методах используются комплексные криптографические схемы, а в качестве метрик риска – оценки вероятности повторения значений совокупности ключей и результатов вычисления крипто-функций на всем пространстве определения (так называемая крипто-игра).

$$P = \frac{r_{train} - r_{control}}{1 - r_{control}}$$

Доверительный интервал биноминального распределения по методу Уилсона

6 СТАТИСТИЧЕСКИЕ МЕТОДЫ И МАШИННОЕ ОБУЧЕНИЕ

Данная группа методов основана применении методов машинного обучения или статистического подхода. Методы объединяет идея замены исходных наборов данных новыми информационными массивами, сохраняющими свойства исходных наборов, но с измененными значениями.

$$P_{\text{риски данных}} = \max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s} \times I_s\right), \frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{F_s} \times I_s\right)\right)$$

Обобщенная формула Эль-Эмама

Для случаев применение дифференциальной приватности количество шума, добавленного к данным, может быть количественно определено с помощью значения ε (меньшее значение ε подразумевает более высокий уровень защиты).

Для случая "epsilon-differential privacy" (EDP)

$$P_{\text{риски данных}} = \frac{1}{n}\exp(\alpha \cdot \epsilon)$$

Гарантия EDP измеряет уровень защиты конфиденциальности, обеспечиваемый дифференциальным механизмом конфиденциальности с заданным значением ε.

Для случая "privacy-loss budget" (PLB)

$$P_{\text{риски данных}} = 1 - \frac{1}{n}\ln(\beta \cdot \epsilon)$$

Гарантия PLB измеряет величину потери конфиденциальности, разрешенную дифференциальным механизмом конфиденциальности с заданным значением ε.

Некоторые из указанных методов могут применяться последовательно, что позволяет дополнительно факторизовать (разложить на множители формулу расчета, например:

$$P_{\text{повторной идентификации}} = P_{\text{контекстные риски}} \times (P_{\text{синтетические наборы}} \times P_{\text{обезличивание}})$$

Типовые кейсы: результаты тестирования

Вероятности по рискам данных представляют из себя вероятность повторной идентификации, также могут быть дополнительно уменьшены умножением на контекстные риски, отражающие способность организаций обеспечивать общую защиту данных

№	КЕЙС	СЦЕНАРИЙ ОБМЕНА ИНФОРМАЦИЕЙ	МЕТОДЫ ЗАЩИТЫ	ОПИСАНИЕ МЕТОДОВ	СЛОЖНОСТЬ АТАКИ	УСПЕХ АТАКИ
1	**СКОРИНГ С ПРИВЛЕЧЕНИЕМ 3-ИХ СТОРОН** ОБМЕН ДАННЫМИ С ИСПОЛЬЗОВАНИЕМ ОДНОСТОРОННИХ КРИПТОГРАФИЧЕСКИХ ФУНКЦИЙ	Сторона А (финансовая организация) формирует массив своих клиентов (представленных мобильными номерами телефонов) и запрашивает сторону В (скоринговое агентство) для формирования скоринга на основании обогащенной информации. сторона В формирует массив данных с номерами телефонов и соответствующих им скоринговых коэффициентов, возвращает его стороне А. Требуется защита схемы обмена.	БАЗОВАЯ СХЕМА: Хэширование MD5 с использованием статической соли	На стороне А номера телефонов хэшируется с заранее согласованной солью – случайной строкой	НИЗКАЯ 1 мин – 10 минут	0.3-0.7 (в зависимости от соли)
			УЛУЧШЕННАЯ СХЕМА: Использование динамической соли и усиленных методов хэширования	Переход к использованию функция SHA-3. Дополнительно организуется отдельный сеанс с передачей набора динамических солей для каждого номера	ВЫСОКАЯ месяцы-годы (в пределе бесконечность)	0.00001
2	**МАРКЕТИНГОВОЕ КАСАНИЕ** НАХОЖДЕНИЕ ПЕРЕСЕЧЕНИЯ ПРИВАТНЫХ НАБОРОВ ДЛЯ ПОСТРОЕНИЯ АНАЛИТИЧЕСКИХ МОДЕЛЕЙ	Сторона А (маркетинговая платформа) формирует данные о просмотрах рекламы пользователями, представленными номерами телефонов. Сторона В (финансовая организация) представляет информацию о платежах по рекламируемому продукту, а также основной профиль ФЛ, осуществившего платеж. Сторона С совмещает наборы информации от А и В, строит модель для машинного обучения, позволяющую осуществить маркетинговое таргетирование. Требуется защита схемы обмена.	БАЗОВАЯ СХЕМА: Формирование для телефонов микросегментов и классическое обезличивание	Представление телефонов, как чисел из заданного промежутка и затем обобщение информации для выбранного диапазона	НИЗКАЯ	0.5
			УЛУЧШЕННАЯ СХЕМА: Реализация PS3I схемы многосторонних безопасных вычислений	Сложная схема обмена информации на основании нескольких ключей и гомоморфного шифрования, а также забывчатой передачи (Oblivious Transferring)	ВЫСОКАЯ (взломать невозможно с учетом защиты каналов)	0.00000001
3	**ОЗЕРО ДАННЫХ** ОБЕЗЛИЧИВАНИЕ ДЛЯ БОЛЬШИХ ИНФОРМАЦИОННЫХ КОМПЛЕКСОВ	СТОРОНА А (финансовая организация) реализует хранилище /озеро данных для управления данными своими клиентов (цифровые профили, майнинг процессов, маркетинг). К озеру данных имеют доступ аналитики данных, а также привлеченные третьи стороны (строят модели и прогнозы). Необходима модель защиты в условиях постоянно наращиваемых больших данных	БАЗОВАЯ СХЕМА: Маскеризация данных	Замена данных или их частей символами-заменителями	СРЕДНЯЯ (в зависимости от степени обобщения)	0.2-0.5
			УЛУЧШЕННАЯ СХЕМА: Многомерная схема обезличивания на основе метода Мондриана	Разбиение исходных наборов на зоны с одинаковыми классами эквивалентности и проведение итерации по “жадному” поиску	ВЫСОКАЯ	< 0.1

Предложения АБД (Правовая модель)

Внесение изменений в 152-ФЗ, допускающих обработку обезличенных данных без согласия субъекта персональных данных при условии, что уровень риска деобезличивания ниже установленного порога

Закрепление модели оценки риска деобезличивания в подзаконных нормативных актах Минцифры России

Предоставление права владельцам данных самостоятельно оценивать и управлять риском деобезличивания данных в соответствии с утвержденной моделью и принимать решение о дальнейшей обработке

Оценка риска деобезличивания включает оценку обрабатываемого датасета (риск данных) и организационно-технических мер, применяемых в процессе обработки (контекстные риски)

Опционально: страхование ответственности участников рынка, применяющих риск-ориентированный подход, в случае реализации риска деобезличивания