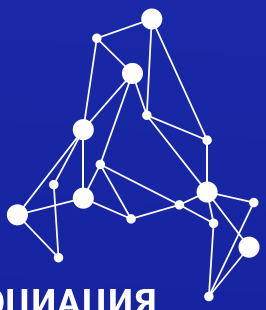

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

ДИФФЕРЕНЦИАЛЬНАЯ ПРИВАТНОСТЬ:
ТЕОРИЯ, МЕХАНИЗМЫ, ПРАКТИКА

Формальные гарантии приватности
данных в эпоху аналитики
и искусственного интеллекта

Москва
2026

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ

АННОТАЦИЯ

Настоящий документ представляет собой обзорный технический доклад (White Paper) о дифференциальной приватности — математически строгой парадигме защиты персональных данных. Документ адресован специалистам в области информационных технологий, анализа данных, информационной безопасности и защиты персональных данных, а также представителям регуляторов и руководителям, принимающим решения о внедрении технологий обработки данных.

Дифференциальная приватность (Differential Privacy, DP) — единственный на сегодняшний день подход, обеспечивающий формальные, математически доказуемые гарантии защиты приватности при публикации статистики, обучении моделей машинного обучения и синтезе данных.

В отличие от традиционных методов обезличивания, которые могут быть преодолены в ходе атак, проводимых на основе вспомогательной информации, DP гарантирует, что участие или отсутствие любого конкретного субъекта данных практически не влияет на результаты работы системы.

В документе последовательно изложены следующие аспекты: концептуальное отличие приватности от «классической» информационной безопасности; история и формальное определение DP; классификация DP (глобальная, локальная, с перемешиванием и др.); ключевые механизмы добавления шума; свойства и теоремы DP; управление бюджетом приватности; интерпретация рисков через параметры epsilon и delta; практические примеры промышленного применения. Также приведены ссылки на международные стандарты, включая NIST и ISO/IEC, и российские нормативные документы.

Ключевые параметры

Epsilon (ε) — параметр приватности (бюджет): чем он меньше, тем выше уровень защиты.
Типичные значения: 0,1 (строгая защита) ... 10 (умеренная защита).

Delta (δ) — вероятность катастрофического нарушения гарантии; обычно в диапазоне 10^{-5}... 10^{-8}.
Чувствительность (Sensitivity) — максимальное влияние одной записи на результат запроса.

ЧТО ТАКОЕ ПРИВАТНОСТЬ

Информационная безопасность

«Классическая» информационная безопасность (ИБ) включает три базовые составляющие: конфиденциальность, целостность и доступность (Confidentiality, Integrity, and Availability, CIA).

1. Конфиденциальность — требование об ограничении передачи информации третьим лицам без согласия ее обладателя.

2. Целостность — обеспечение защиты информации от неправомерного или случайного изменения и уничтожения (информация, которую «видит» тот, кому позволено ее «видеть», не меняется без ведома того, кто за эту информацию отвечает).

3. Доступность — возможность пользоваться информацией (тот, у кого есть доступ к информации, должен по возможности беспрепятственно ее получать).

Злоумышленник может атаковать любую из трех составляющих. Цель ИБ — эти атаки предотвращать.

Возможность атаковать появляется в трех основных ситуациях:

1) когда информация покоится на устройствах хранения (Data-at-Rest);

2) когда она передается по каналам связи (Data-in-Transit);

3) и даже когда она используется (Data-in-Use) по прямому назначению.

Последний пункт — самый интересный, потому что результат использования информации — это почти всегда новая информация. Она может наследовать все свойства и ограничения той информации, на основе которой создается, а может приобретать некоторые дополнительные свойства или, напротив, утрачивать какие-то из имеющихся. Например, новая информация может стать менее конфиденциальной, причем это снижение уровня конфиденциальности может производиться намеренно.

Рассмотрим пример: есть три компании одной отрасли, в которых суммарно работают 5 тыс. человек (считаем, что общая численность сотрудников — это публичная информация), и зарплата каждого из них — безусловно, конфиденциальные данные.

Для защиты конфиденциальности приняты все необходимые меры:

1. Просматривать зарплату может только генеральный директор и бухгалтер соответствующей организации (конфиденциальность).

2. Когда бухгалтер видит, сколько зарабатывает условный Иннокентий, он уверен, что Иннокентий не «сломал» систему и не удвоил себе цифру в базе данных (целостность).

3. Если бухгалтеру нужно вспомнить размер зарплаты условного Иннокентия, он может пройти двухфакторную аутентификацию (например, ввести пароль и отпечаток пальца) и получить доступ к этой информации (доступность).

Все эти требования обеспечиваются криптографическими и/или административно-техническими мерами, принятыми каждой отдельной компанией. В числе таких мер — шифрование, регламенты, защита каналов связи, политики доступа. Все они, как правило, защищают данные на стадиях хранения и передачи.

Теперь предположим, что компании решили регулярно публиковать среднюю заработную плату по отрасли. Чтобы это сделать, нужно, очевидно, сложить зарплаты всех сотрудников и поделить на их количество.

Определенная проблема состоит в том, что компаний три, они конкурируют между собой и не хотят обмениваться информацией о своих сотрудниках напрямую. Чтобы рассчитать среднюю зарплату, защитив при этом свои данные на этапе их использования, компании могут применять один из методов конфиденциальных вычислений.

В самом общем случае любой метод конфиденциальных вычислений (sMPC, HE, TEE) защищает аргументы вычисляемой функции, но не ее результат. Обычно предполагается, что свойства

результата функции (новая информация) отличаются от свойств ее аргументов (исходная информация). Например, результат может быть «менее» конфиденциален.

Так происходит и в нашем примере: средняя зарплата всех сотрудников трех компаний — гораздо менее чувствительная информация, нежели вознаграждение конкретного сотрудника конкретной компании. Более того, компании решили среднюю зарплату публиковать, то есть делать ее доступной определенному кругу лиц (назовем их исследователями). Кстати, само по себе это не делает среднюю зарплату неконфиденциальной информацией, но предполагает, что у исследователей есть право ее «видеть».

На первый взгляд, все в этой схеме хорошо: хранение и передача данных внутри компаний защищены, данные во время совместных вычислений — тоже, и исследователь, как и задумывалось, «видит» только среднюю заработную плату.

Пусть периодичность публикации средних зарплат составляет один месяц. Теперь предположим, что исследователь каким-то образом узнает, что за прошедший месяц общая численность сотрудников изменилась ровно на одного человека: условный Иннокентий уволился. Тогда, зная средние зарплаты и штатную численность сотрудников, исследователь легко может вычислить точную зарплату нашего Иннокентия. Не будем рассматривать вероятность этих событий (зарплаты сотрудников не изменились, численность сотрудников изменилась ровно на единицу, и исследователь узнал об этом), но просто зафиксируем: если все эти события произошли, то исследователь может получить доступ к конфиденциальной информации, который ему никто не предоставлял.

При этом формально все требования «классической» ИБ (конфиденциальность, целостность, доступность) соблюдались: генеральные директора и бухгалтеры по-прежнему единственные, у кого есть доступ к зарплатам конкретных людей, никакого взлома не было, и исследователь получил ровно ту информацию, которая ему предназначалась. Тем не менее условный Иннокентий пострадал: его персональная, следовательно конфиденциальная, информация была скомпрометирована.

Все дело в том, что триада CIA (конфиденциальность, целостность, доступность) — это свойства исходной информации: конкретного файла, конкретной таблицы в базе данных, конкретной записи и значения. Компрометация произошла не на уровне записей, а на уровне результата — информации, которую все рассматривали как «менее» конфиденциальную. Нельзя сказать, что это предположение совсем уж безосновательно, но оно во многом интуитивно: действительно, это же просто «среднее по больнице». Однако любое среднее складывается из отдельных значений, и если, условно говоря, не измерять степень «растворения» частного в общем, то может случиться беда.

Триада CIA остаётся необходимой основой информационной безопасности, но сама по себе не дает количественной гарантии того, насколько участие отдельного человека влияет на раскрываемый результат и что можно узнать из последовательности разрешенных публикаций. Поэтому наряду с защитой систем и данных необходимо управлять рисками приватности.

Дифференциальная приватность (Differential Privacy, DP) — один из математически строгих методов такого управления. DP ограничивает влияние данных отдельного человека на распределение публикуемого результата: результат анализа должен оставаться статистически близким независимо от того, присутствуют ли данные этого человека в анализируемом наборе. Тем самым эта модель позволяет количественно ограничивать дополнительный риск раскрытия, возникающий вследствие участия человека в анализе, но не заменяет конфиденциальность, целостность, доступность или конфиденциальные вычисления, а дополняет их защитой от выводов, которые могут быть сделаны из разрешенных статистических результатов.

Приватность: иное измерение защиты

Приватность (Privacy) — это принципиально иная плоскость. Это понятие отвечает на вопрос: что можно узнать о конкретном человеке, даже не имея права на доступ к данным о нем?

Приватность данных — это состояние, при котором обработка информации о человеке осуществляется в допустимых целях, в необходимом и соразмерном объеме и не создает для него неприемлемого риска раскрытия, идентификации, профилирования, нежелательных выводов или иного неблагоприятного воздействия.

Информационная безопасность и приватность связаны, но не тождественны. Система может надежно обеспечивать контроль доступа, целостность и доступность данных, однако при этом собирать избыточные сведения, использовать их не для заявленной цели, непрозрачно профилировать об отдельных участниках.

Пример нарушения приватности при соблюдении требований CIA
Медицинская база данных опубликована как агрегированная статистика заболеваний. Доступ имеют только авторизованные исследователи (требования CIA соблюдены). Однако зная, что в группе из трех человек двое здоровы, и имея данные об этих двоих из открытых источников, аналитик с высокой вероятностью устанавливает диагноз третьего.

Таким образом, приватность нарушена, причем без какого-либо взлома системы.

Типовые угрозы приватности

Угрозы приватности существуют даже при строгом соблюдении требований CIA:

- **Атака реидентификации (Re-identification Attack):** профессор Гарварда Латанья Суини (Latanya Sweeney) в своей публикации от 2000 года [1] показала, что 87% американцев однозначно идентифицируются по комбинации почтового индекса, даты рождения и пола — всего трех атрибутов, которые сами по себе не являются персональными данными.

- **Атака на основе вспомогательной информации (Auxiliary Information Attack):** объединение анонимизированных данных с данными из открытых источников. В 2008 году Арвинд Нараянан (Arvind Narayanan) и Виталий Шматиков (Vitaly Shmatikov) опубликовали данные исследования «Как нарушить анонимность набора данных Netflix Prize» («How To Break Anonymity of the Netflix Prize Dataset») [2] и показали, что 99% записей набора Netflix Prize однозначно идентифицируются при знании всего ~8 оценок с приблизительными датами, и продемонстрировали практическую деанонимизацию части записей путем сопоставления с публичными рецензиями на IMDb.

- **Атака на принадлежность (Membership Inference Attack, MIA):** определение присутствия данных о конкретном человек в обучающей выборке модели машинного обучения (ML).

- **Вывод атрибутов (Attribute Inference):** восстановление скрытых чувствительных характеристик субъекта на основе других, якобы нечувствительных полей.

- **Атака выделения (Singling Out):** идентификация уникального субъекта по редкой комбинации атрибутов в синтетических или агрегированных данных.

Почему «классических» методов недостаточно

Традиционные методы защиты приватности не обеспечивают формальных гарантий.

Метод	Суть	Уязвимость
Обезличивание	Удаление прямых идентификаторов	Реидентификация через квази идентификаторы
k-анонимность	Каждая запись неотличима от k – 1 других	l-разнообразие и t-близость не гарантированы
l-разнообразие	Разнообразие чувствительных атрибутов	Атаки на сходство, фоновые знания
Агрегация	Публикация только сводных данных	Атаки пересечения, дифференциальные запросы
Шифрование	Защита хранимых данных	Не защищает от вывода по результатам запросов

Фундаментальная проблема всех этих методов в том, что они не дают количественной оценки остаточного риска, и установить, что данные были деанонимизированы, можно лишь постфактум. Дифференциальная приватность решает эту проблему, устанавливая верхнюю границу риска еще того, как данные будут опубликованы.

ПОНЯТИЕ ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ

Исторический контекст

Понятие дифференциальной приватности было сформулировано в середине 2000-х годов в работах Синтии Дворк [3], Фрэнка Мак-Шерри, Кобби Ниссима и Адама Смита. Одна из их ключевых работ — «Калибровка шума по чувствительности при анализе приватных данных» (Calibrating Noise to Sensitivity in Private Data Analysis) — была опубликована в 2006 году [4].

Позднее теория была подробно изложена в монографии Синтии Дворк и Аарона Рота «Алгоритмические основы дифференциальной приватности» (The Algorithmic Foundations of Differential Privacy), вышедшей в 2014 году [5]. Эта работа считается одним из основных справочников по дифференциальной приватности.

В 2017 году Синтия Дворк, Фрэнк Мак-Шерри, Кобби Ниссим и Адам Смит получили премию Гёделя за вклад в развитие теории дифференциальной приватности.

Интуитивное понимание DP

Представим базу медицинских записей. Прямой запрос «Есть ли у Ивана Иванова ген X?» очевидно раскрывает сведения о конкретном человеке.

Но риск может возникнуть и при безопасном, на первый взгляд, запросе типа «Какова доля носителей гена X в базе?». Если ответ на такой запрос заметно меняется после добавления или удаления всего одной записи, то по этому изменению можно сделать определенный вывод о конкретном человеке.

Дифференциальная приватность решает именно эту проблему: она требует, чтобы результат обработки почти не зависел от того, есть ли в наборе данных запись об одном конкретном человеке. Иными словами, наблюдатель, «видящий» только результат обработки, не должен иметь возможность уверенно определить, была ли в исходных данных информация о конкретном человеке.

Формальные определения

Два набора данных называются соседними, если они отличаются только одной записью: запись добавлена, удалена или заменена (см. больше о метриках соседства в приложении А).

Случайный алгоритм обработки данных M является (ε, δ)-дифференциально приватным, если для любых соседних наборов данных D и D', а также для любого множества возможных результатов S выполняется условие:

$$Pr[\,M(D) \in S\,] \leq e^{\varepsilon} * Pr[\,M(D') \in S\,] + \delta$$

Здесь:

- ε — основной параметр приватности: чем меньше ε, тем сильнее защита, но тем больше искажение результата.
- δ — малая допустимая вероятность исключения, при котором строгая гарантия может не выполниться. Обычно δ выбирают существенно меньше 1/n, где n — число записей в наборе данных.

Основные варианты

Чистая дифференциальная приватность рассматривается, когда δ = 0. Это более строгая форма защиты, она часто используется с добавлением шума Лапласа.

Приближенная дифференциальная приватность применяется, когда δ > 0. Это более гибкая и часто более практичная форма, она обычно используется с гауссовым шумом.

Интерпретация параметров

Значение ε не следует воспринимать как универсальную шкалу «хорошо / плохо». Допустимый уровень зависит от задачи, размера набора данных, чувствительности данных и числа повторных запросов.

Ориентировочная интерпретация приведена в таблице.

Параметр	Смысл
ε = 0	Идеальная приватность, результат вообще не зависит от отдельно взятой записи. На практике почти не используется, потому что такие данные становятся бесполезными
ε ≈ 0,1	Очень строгая защита, результаты сильно зашумляются
ε ≈ 1	Сильная защита, часто рассматривается как разумный ориентир для чувствительных данных
ε ≈ 5	Более мягкая защита, может оказаться приемлемой для менее чувствительной аналитики
ε ≥ 10	Слабая защита, требует отдельного обоснования и оценки остаточного риска
Δ	Вероятность исключения из основной гарантии. Должна быть очень малой — например, в зависимости от размера набора данных и сценария применения

Важно: дифференциальная приватность не делает данные «абсолютно безопасными», она определяет измеримое ограничение в отношении того, насколько сильно результат обработки может раскрывать сведения об отдельной записи

ВИДЫ ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ

Дифференциальная приватность может применяться по-разному. Главное отличие между вариантами состоит в том, где именно добавляется шум и кому следует доверять.

Глобальная дифференциальная приватность

В модели глобальной дифференциальной приватности имеется доверенный оператор данных: он получает исходные данные, выполняет обработку и затем добавляет шум к результатам — например, к статистике, таблице, гистограмме или ответу на запрос.

Главное преимущество этого подхода — высокая точность: шум добавляется к агрегированному результату, а не к каждой отдельной записи.

Главный недостаток — необходимость доверять оператору данных. Если оператор нарушает правила или система с исходными данными скомпрометирована, то дифференциальная приватность сама по себе уже не защищает исходные записи внутри этой системы.

Типичные области применения: государственная статистика, публикация агрегированных показателей, аналитические отчеты.

Локальная дифференциальная приватность

В этой модели шум добавляется в данные до их передачи оператору. Это добавление может производиться, например, на устройстве пользователя — в этом случае сервер получает уже измененные данные и не видит исходные значения. Упрощенно говоря, пользователь сообщает неточный ответ, а случайно измененный по заранее заданному правилу. Если таких пользователей много, то по совокупности ответов можно восстановить общую статистику, но не точное значение, переданное отдельным пользователем.

Главное преимущество локальной модели в том, что при ее использовании не требуется доверять серверу. Даже если развернутая на сервере база будет раскрыта, в ней не найдется исходных ответов пользователей.

Главный недостаток этой модели — более низкая точность: поскольку шум добавляется к каждой записи отдельно, для получения надежной статистики обычно требуется существенно больше участников.

Дифференциальная приватность с перемешиванием

Эта модель занимает промежуточное положение между глобальной и локальной моделями: сначала пользователи изменяют свои данные локально, затем отдельный компонент системы перемешивает сообщения так, чтобы анализатор не смог связать конкретное сообщение с конкретным пользователем. Только после этого анализатор строит статистику.

Такой подход снижает зависимость от центрального сервера и при этом может давать более высокую точность, чем чистая локальная модель.

Главное условие применимости этой модели заключается в том, что компонент перемешивания не должен раскрывать связь между пользователями и их сообщениями. Если эта связь раскрыта, защита становится значительно слабее.

Типичные области применения модели: сбор статистики от большого числа пользователей, когда нужна более высокая точность, чем та, которую может обеспечить локальная модель, но при этом полностью доверять центральному анализатору по каким-то причинам нельзя.

Обычная и вычислительная дифференциальная приватность

«Классическое» определение дифференциальной приватности не предполагает, что нарушителю не хватит вычислительных ресурсов. Гарантия формулируется статистически: результат обработки должен почти не меняться при добавлении или удалении одной записи.

В некоторых схемах используются криптографические методы. В этом случае часть гарантий может зависеть от способности нарушителя взломать применяемые криптографические примитивы за разумное время. Такой подход называют вычислительной дифференциальной приватностью.

Для базовых требований к защите персональных данных предпочтительно использовать

«классическую» дифференциальную приватность как более строгую и не зависящую от криптографических предположений. Криптографические методы могут применяться дополнительно, но они не должны подменять само обоснование приватности.

Варианты для машинного обучения: RDP и zCDP

При обучении моделей машинного обучения дифференциальная приватность применяется многократно — на многих шагах обучения и ко многим пакетам данных. Каждое применение механизма расходует часть общего бюджета приватности. Если оценивать суммарный расход бюджета приватности (объем внесенного шума) Грубо — простым суммированием ε по базовой теореме композиции — итоговая оценка получается сильно завышенной и плохо отражает реальный уровень защиты.

Для получения реалистичной оценки используются более точные способы учета бюджета приватности, в том числе дифференциальная приватность Реньи (Rényi Differential Privacy, RDP) и концентрированная дифференциальная приватность в варианте zCDP (Zero-Concentrated Differential Privacy) — «концентрированная в нуле», т.е. с распределением потерь приватности, сосредоточенным вокруг нулевого значения.

Основное предназначение этих видов дифференциальной приватности — получить более точную оценку суммарного расхода приватности при многократном применении механизма. Это особенно важно для алгоритма дифференциально приватного стохастического градиентного спуска (Differentially Private Stochastic Gradient Descent, DP-SGD) и других методов обучения моделей с дифференциальной приватностью.

Предлагаем такое обобщение сказанного:

- **обычная (ε, δ)-DP удобна для общего описания гарантии;**
- **RDP и zCDP удобны для расчета бюджета при обучении моделей;**
- **итоговый результат обычно все равно переводят обратно в (ε, δ)-DP для включения в отчеты и интерпретации.**

Сравнение основных вариантов

Вид дифференциальной приватности	Где добавляется шум	Кому нужно доверять	Точность	Основное применение
Глобальная	После сбора данных, к агрегированному результату	Оператору данных	Высокая	Статистика, аналитика, отчеты
Локальная	До передачи данных, на стороне пользователя	Серверу доверять не требуется	Ниже	Телеметрия, опросы, данные с мобильных устройств
С перемешиванием	Сначала у пользователя, затем сообщения перемешиваются	Компоненту перемешивания	Средняя или высокая	Массовый сбор статистики
Чистая ε-DP	Зависит от механизма	Зависит от архитектуры	Обычно ниже	Простые запросы, гистограммы
(ε,δ)-DP	Зависит от механизма	Зависит от архитектуры	Обычно выше	Аналитика, ML, синтез данных
RDP / zCDP	Зависит от механизма	Зависит от архитектуры	Высокая при точном учете	Обучение моделей, DP-SGD, PATE

Дифференциальная приватность представляет собой не один конкретный алгоритм, а семейство подходов, поэтому при выборе варианта нужно определиться с тремя главными аспектами: моделью доверия, способом добавления шума и порядком учета бюджета приватности.

ОПРЕДЕЛЕНИЕ И КОЛИЧЕСТВЕННАЯ ОЦЕНКА РИСКОВ

ε как мера информационного риска

Параметр ε показывает, насколько сильно результат обработки может зависеть от одной отдельной записи.

Если механизм удовлетворяет требованиям дифференциальной приватности, то добавление или удаление одной записи не должно существенно менять вероятность любого результата. Чем меньше ε, тем слабее влияние отдельной записи на результат и тем выше уровень приватности.

Упрощенно это можно понимать так: наблюдатель, который «видит» только результат обработки, не должен значительно повысить свою уверенность в том, что данные о конкретном человеке были в исходном наборе данных.

Более точно ε ограничивает изменение отношения шансов:
(шансы после наблюдения результата) ≤ e^{ε} × (шансы после наблюдения результата)

Здесь e^{ε} — максимальный множитель изменения.

ε	e^{ε}	Упрощенная интерпретация
0	1,00	Результат не раскрывает влияния отдельной записи
0,1	1,11	Очень сильная гарантия
1	2,72	Существенная, достаточно строгая гарантия
5	148	Слабая теоретическая гарантия
10	22026	Очень слабая теоретическая гарантия

Важно: большое значение ε не означает автоматическую утечку данных. Оно показывает, что формальная верхняя граница риска становится слабой, поэтому требуется дополнительная практическая проверка путем проведения тестовых атак.

Связь с атаками на принадлежность

Теорема (Yeom et al., 2018) [18]: вероятность успеха оптимальной атаки на принадлежность (MIA) при (ε, δ)-DP ограничена сверху:

$$\text{Advantage\_MIA} \leq (e^{\varepsilon} - 1) / (e^{\varepsilon} + 1) + \delta$$

Это означает, что DP напрямую ограничивает способность атакующего определить, была ли конкретная запись в обучающей выборке.

ε	Максимальное преимущество MIA	Упрощенная интерпретация
0,1	≈ 0,05	Почти случайное угадывание
1,0	≈ 0,46	Умеренное преимущество
5,0	≈ 0,86	Высокое преимущество
10,0	≈ 0,99	Практически полное знание

Интерпретация δ

Параметр δ задает малую вероятность исключения из основной ε -гарантии. Если ε ограничивает величину возможного изменения вероятности результата, то δ показывает, насколько мала вероятность того, что это ограничение может быть нарушено.

Поэтому δ должен быть очень малым. Обычно его выбирают меньше величины 1/n, где n — число записей в наборе данных. В более строгих сценариях используют значения порядка $1/n^2$ или ниже.

Приведем ориентировочные значения n и δ.

Размер набора данных	Верхний ориентир для δ
$n = 10^6$	Меньше 10^{-6}, лучше 10^{-8}... 10^{-12}
$n = 10^9$	Меньше 10^{-9}, лучше 10^{-10}... 10^{-12}

Важно: значение δ нельзя выбирать произвольно, оно должно соответствовать размеру набора данных, степени чувствительности данных и сценарию применения.

Потеря приватности

В анализе многократных запросов применяется понятие потери приватности.
Для конкретного результата потеря приватности определяется как логарифм отношения вероятностей получения этого результата на двух соседних наборах данных:

$$L(o) = log(Pr[M(D) = o] / Pr[M(D') = o])$$

Если механизм удовлетворяет чистой ε-DP, то потеря приватности не выходит за пределы ε.

При (ε, δ)-DP допускается малая вероятность исключения, заданная параметром δ. При выполнении многократных запросов потери приватности накапливаются, поэтому система должна вести учет бюджета приватности: сколько запросов выполнено, какие механизмы применялись и какой суммар-ный бюджет уже израсходован. Такой учет обычно называют учетом бюджета приватности.

Теоретическая гарантия и практическая проверка

Параметры (ε,δ) дают формальную верхнюю границу риска — это сильная гарантия, но она не заменяет практическую проверку реализации.

В качестве дополнительной практической меры рекомендуется проводить оценку устойчивости к кибератакам.

Проверка	Что оценивает
Атака на принадлежность	Можно ли определить, что запись участвовала в обучении модели или синтезе данных
Атака на связывание	Можно ли связать синтетическую или обезличенную запись с внешними данными
Атака на выделение	Можно ли по редкой комбинации признаков выделить уникальный субъект
Атака на вывод	Можно ли восстановить чувствительный атрибут по другим признакам
Метрики близости	Не слишком ли синтетические записи похожи на реальные

Пороговые значения для таких проверок не являются универсальными. Для правильного определения этих значений в конкретном случае нужно, чтобы в методике оценки учитывались типы данных, модели угроз и сценарии публикации.

Например, результат, близкий к случайному угадыванию в случае проверки с применением атаки на принадлежность, соответствует уровню точности около 0,5. Заметно более высокую точность атаки следует оценивать как признак повышенного риска и как повод пересмотреть параметры генерации, уровень шума или состав публикуемых данных.

ПРИМЕНЕНИЕ: КОГДА DP РАБОТАЕТ, А КОГДА НЕТ

Дифференциальная приватность не дается «бесплатно». Чтобы скрыть вклад отдельной записи, механизм добавляет в данные случайный шум. Он повышает приватность, но снижает точность результата.

Поэтому для применения DP важны четыре фактора:
- **размер выборки;**
- **число признаков;**
- **требуемый уровень приватности;**
- **сложность задачи, которую нужно решить на основе данных.**

Если имеется много данных, умеренное количество признаков и при этом задача сводится к статистике или устойчивому обучению модели, то DP может работать достаточно хорошо. Если же данных мало, признаков много, а важная информация содержится в редких комбинациях, то DP может сделать результат малополезным.

Размер выборки и число признаков

Чем больше признаков в данных, тем больше информации нужно сохранить для анализа. Но при использовании DP часть информации неизбежно скрывается под воздействием шума, поэтому с ростом числа признаков требуется большее количество записей — в противном случае шум будет доминировать над полезным сигналом.
Упрощенное правило такое: чем больше d и чем меньше ε, тем более высоким должно быть n, где:
n — число записей;
d — число признаков;
ε — бюджет приватности.

Практический смысл:

- **при большом n шум усредняется и создает меньше помех для анализа;**
- **при большом d шум влияет на большее число направлений анализа;**
- **при малом ε защита сильнее, но точность ниже.**

Если число признаков велико, то перед тем, как применять DP, следует уменьшить размерность: удалить слабые признаки, объединить категории, агрегировать редкие значения или использовать методы снижения размерности.

Сложность модели

DP проще применять к простым видам статистики: средним значениям, долям, гистограммам, агрегатам. Сложнее применять DP для обучения ML-моделей: шум приходится добавлять многократно — например, на очередном шаге обучения или при расчете градиентов. Поэтому расход бюджета приватности нужно учитывать в ходе всего процесса обучения.

Чем сложнее модель, тем больше данных обычно требуется для того, чтобы сохранить их качество. Для простых моделей это ограничение может быть приемлемым. Для больших нейронных сетей и языковых моделей DP-обучение требует наличия очень больших наборов данных и аккуратного учета бюджета приватности.

Практический вывод: DP не следует рассматривать как простую кнопку «включить приватность». Для сложных моделей нужно заранее проверять, останется ли качество данных после добавления шума достаточно высоким.

Компромисс между приватностью и полезностью

В DP всегда существует компромисс:

- **ε меньше — приватность сильнее, но шум больше;**
- **ε больше — точность выше, но формальная гарантия слабее;**
- **n больше — выше шанс сохранить полезность при том же уровне приватности;**
- **d больше — риск потери качества результата выше.**

Упрощенно говоря, потеря качества растет по мере увеличения числа признаков и уменьшается при росте числа записей:

потеря качества ↑ при росте d,
потеря качества ↓ при росте n и ε.

Заметим, что это не универсальная формула, а практическое правило. Точный результат зависит от данных, алгоритма и выбранной метрики качества.

Ограничение на извлекаемую информацию

DP ограничивает не только риск раскрытия отдельной записи, но и объем информации, который можно безопасно извлечь из данных.

Если ε очень мал, а данных мало, то механизм не может одновременно:

- **хорошо скрывать вклад каждого человека;**
- **точно отвечать на сложные запросы;**
- **сохранять редкие закономерности;**
- **обучать сложную модель.**

Это фундаментальное ограничение. Его нельзя полностью обойти, применив более сложный алгоритм. Можно лишь изменить постановку задачи: увеличить выборку, уменьшить число признаков, ослабить уровень приватности или упростить требуемый результат.

Редкие события и малые группы

DP особенно плохо работает там, где полезная информация связана с редкими событиями или малыми группами.

Примеры:

- редкое заболевание;
- редкая профессия;
- редкая комбинация возраста, региона и дохода;
- редкий тип мошеннической операции;
- малая категория клиентов.

Если такая группа мала, то шум DP может полностью скрыть ее сигнал. Если же шум слишком мал, возникает риск раскрытия информации о конкретных людях.

Практический вывод: перед применением DP нужно проверять малые группы, редкие категории и выбросы. Иногда необходимо их объединять, подавлять или анализировать отдельно.

Синтетические данные с DP

При синтезе данных DP защищает не сами синтетические записи, а процесс обучения генератора или выпуска статистик, на которых строится генерация.

DP-синтез работает лучше при выполнении ряда условий:

- исходных данных много;
- признаки не слишком разрежены;
- редкие категории заранее обработаны;
- задача синтеза ограничена конкретным сценарием;
- качество проверяется не только общими метриками, но и путем проведения атак.

DP-синтез работает хуже в следующих случаях :

- набор данных мал;
- много признаков и редких комбинаций;
- требуется сохранить точные связи между многими признаками;
- нужно воспроизвести редкие, но важные события;
- механизм синтеза многократно запускается на исходных данных (переобучение, выпуск новых версий) без учета суммарного бюджета приватности.

Практическая проверка применимости DP

Перед внедрением DP рекомендуется выполнить предварительную проверку ее применимости.

Вопрос	Почему это важно
Достаточно ли записей?	При малом n шум может уничтожить полезный сигнал
Сколько признаков используется?	При большом d требуется больше данных
Есть ли редкие категории?	Малые группы могут либо исчезнуть, либо раскрыться
Какое значение ε допустимо?	Слишком маленький ε ухудшает качество результата, слишком большой ослабляет защиту
Сколько раз будет применяться механизм?	Повторные запросы расходуют общий бюджет приватности
Какая задача решается?	Реализовать DP для агрегатов проще, для сложного обучения — труднее
Проверялась ли устойчивость к атакам?	Формальная гарантия должна дополняться практической оценкой риска

Итоговое правило

Дифференциальная приватность хорошо подходит для задач, где есть достаточно большая выборка и можно принять умеренную потерю точности ради формальной гарантии приватности.

DP плохо подходит для задач, где данных мало, признаков много, а полезный результат зависит от редких индивидуальных случаев.

Решение о применении DP должно приниматься не только на основании значения ε, но и по результатам проверки качества результата, анализа редких групп и практической оценки устойчивости к атакам.

ПРАКТИЧЕСКИЕ ЗАДАЧИ И ПРИМЕНЕНИЕ

Дифференциальная приватность может применяться в разных задачах: при публикации статистики, обучении моделей, сборе телеметрии, федеративном обучении и создании синтетических данных.

Одно из наиболее востребованных направлений — синтез данных. Цель такого синтеза состоит в том, чтобы создать новый набор данных, похожий на исходный по статистическим свойствам, но не раскрывающий сведения о конкретных субъектах.

Важно различать:

- **синтетические данные как результат;**
- **механизм, с помощью которого они были получены.**

Сам факт «синтетичности» не дает гарантии приватности. Обеспечить их может только процесс создания данных, если он построен как дифференциально приватный механизм.

Синтез как DP-механизм

Генератор синтетических данных можно рассматривать как механизм M, который принимает реальные данные D и выдает синтетический набор D_syn.

Если механизм M удовлетворяет (ε, δ)-дифференциальной приватности, то выпуск D_syn наследует эту гарантию. Это означает, что наличие или отсутствие одной записи в исходных данных не должно существенно менять вероятность получения опубликованного синтетического набора.

Для DP-синтеза важно свойство постобработки: действия с уже выпущенным результатом DP-механизма не ухудшают исходную гарантию приватности.

На практике это означает, что после выпуска D_syn можно:

- **готовить агрегаты и отчеты;**
- **обучать модели на основе D_syn;**
- **строить графики и витрины данных;**
- **передавать набор данных для решения согласованных аналитических задач.**

Если последующая обработка выполняется только с уже выпущенным DP-результатом, то она не расходует дополнительный бюджет приватности.

Однако повторный запуск генератора на тех же самых исходных данных, переобучение модели или выпуск новой версии синтетического набора обычно считаются новым обращением к исходным данным. В этом случае необходимо учитывать дополнительный расход бюджета приватности.

Синтетические данные с DP могут использоваться и передаваться третьим сторонам только в рамках установленного правового, договорного и организационного режима. Дифференциальная приватность снижает риск раскрытия исходных записей, но не отменяет необходимость управления доступом, ведения системных журналов, целевого использования и проверки остаточных рисков.

Основные подходы к DP-синтезу

Существует несколько основных подходов к синтезу данных с дифференциальной приватностью.

Синтез на основе статистик: из исходных данных извлекаются статистики (частоты, распределения, корреляции, маргинальные таблицы), затем к ним добавляется шум, после чего на базе зашумленных статистик строится синтетический набор. Этот подход хорошо подходит для табличных данных, аналитики и отчетности. Его преимущество — прозрачность и управляемость, ограничение — сложность сохранения многомерных связей при большом числе признаков.

DP-Copula: синтез строится через зашумленные распределения отдельных признаков и зашумленную структуру зависимостей между ними. Подход применим для табличных данных, когда важно сохранить общие статистические связи, но не требуется точно воспроизводить редкие комбинации.

Графические модели: методы типа PrivBayes, MST и Private-PGM строят приближенную модель зависимостей между признаками и используют зашумленные маргинальные распределения. Часто они подходят для категориальных и смешанных табличных данных. Преимущество графических моделей — возможность сохранить часть зависимостей между признаками,

ограничение — снижение качества при высокой размерности и большом количестве редких категорий.

DP-GAN и DP-VAE: эти генеративные модели обучаются с применением дифференциальной приватности, обычно через DP-SGD или близкие методы. Их преимущество — гибкость, они могут применяться к сложным данным. Среди их недостатков — сложная настройка, высокий расход бюджета приватности и нестабильность качества результата, особенно на небольших наборах данных.

PATE: в этом подходе несколько моделей-«учителей» обучаются на непересекающихся частях исходных данных, затем их ответы агрегируются с добавлением шума, после чего на этих зашумленных ответах обучается модель-«студент». Подход полезен там, где исходные данные можно разделить на достаточно крупные независимые части.

Новые методы отбора синтетических кандидатов: в некоторых новых подходах предварительно обученная модель используется для генерации кандидатов, а DP-бюджет расходуется на их отбор или оценку. Это может быть полезно, когда прямое обучение генератора с DP оказывается слишком затратным или ведет себя нестабильно. Такие методы требуют отдельной проверки качества, смещений и остаточных рисков.

Другие области применения DP

Помимо синтеза данных, дифференциальная приватность применяется в нескольких крупных направлениях.

Машинное обучение: для обучения моделей используется DP-SGD и близкие к нему методы. В таких алгоритмах вклад отдельных записей ограничивается, а затем к обновлениям модели добавляется шум — это снижает риск того, что обученная модель запомнит сведения об отдельном человеке. Такой подход особенно важен при обучении моделей на чувствительных данных. Качество модели при этом может снижаться, поэтому требуется обеспечить отдельную проверку точности и учет общего бюджета приватности.

Государственная и публичная статистика: DP может применяться при публикации статистических таблиц, агрегатов и отчетов. Основная задача — снизить риск раскрытия сведений о малых группах и отдельных субъектах при сохранении полезности статистики. Такие системы требуют особенно аккуратного выбора бюджета приватности, так как слишком сильный шум может ухудшить качество официальной статистики.

Телеметрия и пользовательская аналитика: в локальной DP шум добавляется на стороне пользователя до передачи его данных на сервер — такой подход позволяет собирать общую статистику использования продуктов без получения точных исходных значений от каждого пользователя. Подход хорошо работает при очень большом числе участников. Для малых выборок локальная DP часто дает слишком зашумленные результаты.

Федеративное обучение с DP: в этом сценарии данные остаются у участников, а в центральную систему передаются обновления модели или агрегированные параметры. Если к этим обновлениям добавляется DP-шум, то снижается риск восстановления информации об отдельных участниках. Этот подход применяется там, где исходные данные нельзя или нежелательно передавать в единое хранилище.

Финансовый сектор и данные с повышенной чувствительностью: в финансовых задачах DP может применяться для аналитики, тестирования моделей, синтеза транзакционных данных, межорганизационного обмена и в сценариях антифрода. Однако в таких задачах часто встречаются редкие события, выбросы и малые группы, поэтому применение DP должно сочетаться с обработкой редких категорий, практической проверкой на устойчивость к атакам и правовым контролем доступа к результатам.

Практический конвейер применения

В практических проектах можно использовать следующий порядок:

- согласование цели
- подготовка данных
- обработка редких категорий
- выбор DP-механизма
- синтез или обучение
- оценка качества
- проверка на устойчивость к атакам
- отчет и аудит

В отчетах рекомендуется фиксировать:

- цель применения;
- состав исходных данных;
- единицу приватности;
- выбранный механизм;
- значения ε и δ;
- число запусков механизма;
- метрики качества;
- результаты проверок на устойчивость к атакам;
- ограничения на использование результата.

Такой подход позволяет обеспечивать управляемое и проверяемое применение DP. Кроме того, он помогает отделить техническую гарантию приватности от юридических и организационных требований.

Полезность технологии

Дифференциальная приватность наиболее полезна там, где нужно получить измеримую техническую гарантию, сохранив часть аналитической ценности данных.

Применение DP в синтезе данных позволяет выпускать наборы, пригодные для аналитики, тестирования и ограниченного обучения моделей. Качество результата будет зависеть от размера данных, числа признаков, редких категорий, выбранного бюджета приватности и сценария использования. Поэтому DP следует рассматривать не как универсальную замену обезличиванию или правовой оценке, а как один из строгих технических инструментов в составе общей системы управления приватностью.

РЕГУЛИРОВАНИЕ И СТАНДАРТЫ

Российское регулирование рассматривает дифференциальную приватность как разновидность обезличивания персональных данных, поэтому на нее можно распространить положения, описанные нами ранее в Докладе, посвященном обезличиванию (Технологии защищенной обработки данных. Обезличивание (2025)).

Использование DP само по себе не приводит к изменению статуса обрабатываемых компанией-оператором данных — они остаются персональными. При этом внедрение механизмов бережливой обработки данных, к которым, в частности, относится и DP, приносит пользу как самому оператору, так и непосредственно субъектам персональных данных и иным лицам.

Например, это позволяет снизить потенциальные негативные последствия утечки для субъектов, а также помочь оператору в случае утечки утверждать, что он принимал не только минимально необходимые по закону, но и повышенные меры защиты — такие заявления могут стать вескими аргументами в пользу снижения административного штрафа или даже освобождения от ответственности.

Кроме того, использование DP при обработке персональных данных на основании законного интереса может служить дополнительным аргументом в пользу соблюдения баланса интересов оператора и субъектов персональных данных.

Наконец, в целом внедрение технологий защищенной обработки данных (Privacy-Enhancing Technologies, PET) повышает уровень репутации оператора как в глазах своих контрагентов, так и с точки зрения регуляторов.

Ниже перечислены некоторые наиболее применимые стандарты дифференциальной приватности, действующие в России и мире, и приведены примеры инструментов, пригодных для практической реализации DP.

Международные стандарты

Документ	Организация	Релевантность для DP
NIST Privacy Framework v1.0 (2020)	NIST (США)	Управление рисками приватности; DP совместима с требованиями фреймворка
NIST SP 800-188 «De-Identifying Government Datasets: Techniques and Governance»	NIST	Рекомендует DP как метод обезличивания государственных данных
NIST IR 8062 (Privacy Risk Management)	NIST	Методология оценки рисков приватности; DP как инструмент снижения риска
ISO/IEC 20889:2018	ISO/IEC JTC 1/SC 27	Определение DP (п. 3.9); нормативная основа для терминологии
ISO/IEC 29101 (Privacy Architecture)	ISO/IEC	Архитектура систем защиты приватности
ISO/IEC 27559 (De-identification framework)	ISO/IEC	Фреймворк обезличивания, включая DP-методы

Российские нормативные акты

Документ	Орган	Связь с DP
Приказ РКН N140 от 19.06.2025	Роскомнадзор	Требования к обезличиванию; DP как математически обоснованный метод
ПНСТ 1.11.164-1.361.25 (Ч. 1–3)	Росстандарт / ТК 164	Предстандарты синтеза данных с DP-гарантиями (в стадии экспертного рассмотрения)
ГОСТ Р ISO/IEC 29100-2021	Росстандарт	Адаптация ISO/IEC 29100 — приватность в ИС

Примеры программных инструментов

Инструмент	Разработчик	Тип	Применение
TensorFlow Privacy	Google	Open Source	DP-SGD для TensorFlow
Tumult Analytics	Tumult Labs / ex-Census	Open Source	Запросы на DP-SQL
SmartNoise (OpenDP)	Microsoft / Harvard	Open Source	DP-механизмы, синтез
IBM Diffprivlib	IBM	Open Source	DP-статистика и ML
Chorus	UC Berkeley / Uber	Исследовательский, с опытным промышленным развертыванием	DP поверх SQL СУБД
Gretel.ai	Gretel	Коммерческий	DP-синтез данных
Mostly.AI	Mostly AI	Коммерческий	SaaSDP-синтез, финансы, HR

ОГРАНИЧЕНИЯ

Дифференциальная приватность — мощный инструмент, но она не является универсальным решением всех проблем, связанных с приватностью. Практические специалисты должны понимать ее ограничения.

Компромисс между приватностью и полезностью

Добавление шума неизбежно снижает точность результатов. Строгие гарантии (малый ε) приводят к значительной потере полезности данных. Для малых датасетов ($n < 1000$) с высокой размерностью ($d > 100$) DP-системы могут оказаться непрактичными.

Универсального «безопасного» значения ε не существует. Его оптимальная величина зависит от конкретного контекста, размера выборки, чувствительности данных и требований к точности.

Ограничения на уровне записи

«Классическая» DP защищает данные на уровне записей, но не гарантирует защиту от утечки информации об агрегатных паттернах и корреляциях. Если группа субъектов имеет уникальную комбинацию атрибутов, то это может быть обнаружено даже в DP-данных (при слабых ε).

Как показывает анализ атак на восстановление паттерна (Pattern Reconstruction Attack), при высокой полезности синтетических данных корреляция Спирмена между значимостью признаков в реальных и синтетических данных $rho \to 1$, что указывает на утечку структуры модели.

Сложность реализации

Корректная реализация DP требует аккуратного учета некоторых аспектов: чувствительности всех функций, полного бюджета (включая все промежуточные запросы), корректного клиппинга и добавления шума. Ошибки в любом из них могут привести к нарушению заявленных гарантий без явных ошибок кода.

DP не заменяет другие меры безопасности

Дифференциальная приватность — это компонент защиты, но она не определяет контуры полной архитектуры безопасности. Необходимо сочетать DP с рядом других мер:

- **контролем доступа и аутентификацией (CIA-модель);**
- **шифрованием данных при их хранении и передаче;**
- **процедурными мерами — политиками обработки данных и аудитом;**
- **юридическими мерами — соглашениями о неразглашении и DPA.**

DP гарантирует, что, даже если данные украдены или произошла их утечка, злоумышленник не сможет извлечь информацию сверх гарантированного предела. Но DP не предотвращает саму утечку.

ВЫВОДЫ

Дифференциальная приватность представляет собой математически строгую и практически применимую парадигму защиты приватности персональных данных. На основе изложенного в настоящем документе можно сформулировать следующие ключевые выводы:

1. DP — это единственная формальная гарантия, единственный подход, дающий количественно измеряемую и математически доказуемую верхнюю границу риска утечки информации о конкретном субъекте данных, не зависящую от вспомогательных знаний атакующего.

2. Параметры ε и δ имеют прямую интерпретацию через вероятностные риски: ε ограничивает максимальное «байесовское обновление» знаний атакующего о субъекте, а δ — вероятность катастрофического нарушения гарантии.

3. Теоремы о постобработке и композиции делают DP методологией, практически применимой в сложных конвейерах: любая последующая обработка DP-результата бесплатна с точки зрения бюджета, к тому же суммарный бюджет поддается точному учету.

4. На рынке имеется вполне зрелая экосистема инструментов: TensorFlow Privacy, Opacus, Tumult Analytics, SmartNoise/OpenDP, Mostly.AI, Gretel.ai — это пригодные для промышленного применения решения с активной технической поддержкой.

5. Промышленная зрелость подтверждена масштабными развертываниями: кейсы в US Census Bureau (~330 млн записей), Apple (сотни миллионов устройств), Google (RAPPOR) и LinkedIn демонстрируют применимость DP в реальных продуктивных средах.

6. В российском контексте DP совместима с 152-ФЗ. Формальные гарантии ε-DP могут быть использованы для обоснования уровня защиты персональных данных перед Роскомнадзором. Российские предстандарты ПНСТ (ТК 164) закрепляют позиции DP как метода защиты при синтезе данных.

7. Ограничения требуют осознанного управления: DP не устраняет компромисс между приватностью и полезностью; малые дата-сеты с высокой размерностью требуют слабых гарантий; DP не заменяет традиционные меры ИБ.

8. Выбор ε — операционное решение, требующее понимания контекста. Универсального «безопасного» значения ε не существует, его выбор определяется размером выборки, чувствительностью данных, требованиями к точности и регуляторными требованиями.

ПРИЛОЖЕНИЯ

Приложение А.
ЕДИНИЦА ПРИВАТНОСТИ И СОСЕДНИЕ НАБОРЫ ДАННЫХ

Общие положения

Первым шагом к применению дифференциальной приватности является определение объекта, который нужно защищать. Для этого используется понятие соседних наборов данных.

Два набора данных называются соседними, если они отличаются минимальным защищаемым элементом: одной записью, одним событием, одним пользователем или группой записей.

Выбор соседства определяет смысл гарантии. Нельзя корректно указать ε и δ, не указав единицу приватности.

Единица приватности

Единица приватности — это минимальный объект, вклад которого должен быть защищен.

Уровень	Определение соседства	Что защищается	Примеры
Событие	Отличается одним событием	Факт отдельного действия	Одна покупка, один переход, одна транзакция
Запись	Отличается одной строкой таблицы	Одна запись	Одна строка в медицинской или финансовой таблице
Пользователь	Отличается всеми записями одного пользователя	Все поведение пользователя	История операций, история действий в браузере
Группа	Отличается группой объемом до k записей или субъектов	Совместный вклад группы	Семья, домохозяйство, рабочая группа

Защита на уровне пользователя обычно сильнее защиты на уровне записи, но требует большего зашумления. Если один пользователь может иметь до k записей, то чувствительность механизма может увеличиться до k раз. На практике ее ограничивают с помощью ограничения вклада: например, не более k событий от одного пользователя.

Варианты соседства

Наиболее распространены следующие определения соседства:

Вид соседства	Описание	Комментарий
Добавление или удаление	Один набор получается из другого добавлением или удалением одной записи	Часто используется в теории DP
Замена записи	Наборы отличаются заменой одной записи	Часто используется при фиксированном размере выборки
Групповое соседство	Наборы отличаются группой до k записей	Используется для групповой приватности
Метрическое соседство	Расстояние между объектами задается метрикой d(x, x′)	Используется для координат, текстов, временных рядов

При расчете чувствительности важно явно указать, какое соседство используется. Например, значения чувствительности гистограммы при добавлении записи и при замене записи существенно различаются.

Метрическая дифференциальная приватность

В «классической» DP соседство обычно задается для наборов данных. В некоторых задачах полезно применять обобщение, в котором близость определяется метрикой между объектами.

Механизм M удовлетворяет метрической DP с параметром ε, если для любых x, x' и любого множества результатов S выполняется условие:

$$Pr[M(x) \in S] \leq exp(\varepsilon \cdot d(x, x')) \cdot Pr[M(x') \in S]$$

Здесь $d(x,\ x')$ — выбранное расстояние между объектами.

Если $d(x,\ x')$ равно 0 при $x = x'$ и равно 1 при иных x и x', то получается обычная локальная форма DP.
В таблице ниже приведены некоторые примеры применения.

Область	Метрика	Что защищается
Геолокация	Расстояние между точками с координатами	Точное местоположение
Текст	Расстояние между словами или векторами	Точное слово или фраза
Временные ряды	L2, DTW или иная метрика	Точная траектория изменения сигнала

Использование метрической DP требует аккуратного выбора метрики. Если метрика плохо отражает реальную близость объектов, то формальная гарантия может оказаться слишком слабой для конкретной практической задачи.

Приложение Б.
ЧУВСТВИТЕЛЬНОСТЬ ФУНКЦИЙ

Что такое чувствительность функций

Чувствительность функции показывает, насколько сильно может измениться результат запроса при изменении одной единицы приватности.

Пусть f — функция запроса, а D и D' — соседние наборы данных. Чувствительность определяется как максимальное изменение результата:

$$\Delta f = max \, ||f(D) - f(D')||$$

Норма выбирается в зависимости от механизма:

- $L1$ — обычно для механизма Лапласа;
- $L2$ — обычно для механизма Гаусса;
- чувствительность функции полезности — для экспоненциального механизма.

Чем выше чувствительность, тем больше шума нужно добавить, чтобы обеспечить ту же гарантию приватности.

Ограничение диапазона и вклада

Перед применением DP необходимо ограничить вклад одной записи или одного пользователя — без этого чувствительность может быть неограниченной.

Примеры ограничений:

- значение дохода ограничивается диапазоном [0, R];
- число событий от одного пользователя ограничивается k;
- градиент в DP-SGD ограничивается нормой C;
- редкие категории объединяются или подавляются.

Ограничение вклада является частью DP-механизма и должно быть документировано.

L1-чувствительность

$L1$-чувствительность определяется по формуле:

$$\Delta_1 f = max \, \left|\left|f(D) - f(D')\right|\right|_1$$

Для скалярной функции это обычный модуль разности.

Примеры при соседстве по добавлению или удалению одной записи приведены в таблице.

Запрос	Условие	*L*1-чувствительность
COUNT	Одна запись добавляется или удаляется	1
SUM	Значение ограничено диапазоном [0, 1]	1
SUM	Значение ограничено диапазоном [0, R]	R
Среднее	Значение в [0, R], фиксированный размер n	R / n
Гистограмма	Одна запись попадает в один интервал	1

При соседстве по замене записи чувствительность может быть выше. Например, для гистограммы замена одной записи может уменьшить один интервал и увеличить другой, поэтому $L1$-чувствительность равна 2.

*L*2-чувствительность

$L2$-чувствительность определяется по формуле:

$$\Delta_2 f = max \left|\left|f(D) - f(D')\right|\right|_2$$

Она используется для векторных запросов и Гауссова механизма.

Примеры приведены в таблице.

Запрос	Условие	*L*2-чувствительность
COUNT	Скалярный счетчик	1
Гистограмма	Добавление одной записи	1
Гистограмма	Замена одной записи	$\sqrt{2}$
Вектор средних	Каждый признак ограничен	Зависит от диапазонов и размерности
Средний градиент	Норма каждого градиента ограничена C, размер батча равен B	Обычно порядка C / B

В DP-SGD вклад каждой записи ограничивается клиппингом градиента:

$$g_{clipped} = g / max(1, ||g||_2 / C)$$

После этого к среднему градиенту добавляется Гауссов шум.

Сводная таблица

Объект	Чувствительность	Типичный механизм
Счетчики	Δ1 = 1	Лаплас, дискретный шум
Суммы с ограниченным диапазоном	Зависит от диапазона	Лаплас или Гаусс
Средние значения	Зависит от диапазона и величины n	Лаплас или Гаусс
Гистограммы	Зависит от определения соседства	Лаплас, Гаусс, дискретный шум
Векторные статистики	*L2*-чувствительность	Гаусс
Градиенты модели	Ограничиваются клиппингом	Гаусс, DP-SGD
Выбор категории	Чувствительность функции полезности	Экспоненциальный механизм

Приложение В.
МЕХАНИЗМЫ DP

Общие положения

DP-механизм — это алгоритм, который преобразует результат запроса так, чтобы ограничить влияние одной единицы приватности.

В простейшем случае механизм имеет вид:

$$M(D) = f(D) + \text{шум}$$

Шум должен быть откалиброван по чувствительности функции и выбранным параметрам приватности.

Механизм Лапласа

Механизм Лапласа применяется к числовым запросам. Он использует *L*1-чувствительность.

$$M(D) = f(D) + \eta$$

Здесь:

$$\eta \sim Lap(0, b)$$

$$b = \Delta 1\, f \,/\, \varepsilon$$

Механизм Лапласа обеспечивает чистую ε-DP, то есть вариант с δ = 0.

Типичные применения:

- **счетчики;**
- **суммы;**
- **средние значения;**
- **гистограммы;**
- **простые аналитические запросы.**

Практический порядок применения механизма Лапласа:

1. определить функцию f;
2. задать единицу приватности и соседство;
3. ограничить диапазон значений или вклад пользователя;
4. вычислить Δ1 f;
5. выбрать ε ;
6. добавить шум Лапласа

Механизм Гаусса

Механизм Гаусса применяется к числовым и векторным запросам и опирается на *L*2-чувствительность. Общая формула аналогична:

$$M(D) = f(D) + \eta$$

Здесь:

$$\eta \sim N(0, \sigma^2 I)$$

Одна из простых достаточных калибровок выглядит так:

$$\sigma \geq \Delta 2\, f \cdot sqrt(2\, ln(1{,}25 \,/\, \delta)) \,/\, \varepsilon$$

Эта формула используется как базовая оценка для (ε, δ)-DP при стандартных условиях. В современных системах для точного расчета часто применяют RDP, zCDP или специальные инструменты учета бюджета приватности.

Типичные применения механизма Гаусса:

- **векторные статистики;**
- **матрицы корреляций;**
- **градиенты моделей;**
- **DP-SGD;**
- **федеративное обучение;**
- **часть методов DP-синтеза.**

Экспоненциальный механизм

Экспоненциальный механизм используется, когда результатом является не число, а выбор объекта из множества вариантов.

Пусть $u(D, r)$ — функция полезности, показывающая, насколько результат r хорош для набора данных D.

Чувствительность функции полезности можно оценить так:

$$\Delta u = max\, |u(D, r) - u(D', r)|$$

Экспоненциальный механизм выбирает результат r с вероятностью, вычисляемой по формуле:

$$exp(\varepsilon \cdot u(D, r) \,/\, (2\Delta u))$$

Механизм чаще выбирает более полезные ответы, но сохраняет случайность выбора.

Типичные применения экспоненциального механизма:

- **выбор категории;**
- **выбор top-k элементов;**
- **выбор признаков;**
- **выбор синтетических кандидатов;**
- **некоторые методы генерации текста и синтетических данных.**

Рандомизированный ответ и локальные механизмы

Рандомизированный ответ (Randomized Response) является базовым локальным механизмом для чувствительных запросов.

Для бинарного ответа механизм сообщает истинный ответ со следующей вероятностью:

$$p = e^{\varepsilon} / (e^{\varepsilon} + 1)$$

и противоположный ответ с вероятностью:

$$1 - p = 1 / (e^{\varepsilon} + 1)$$

Такой механизм обеспечивает ε-локальную дифференциальную приватность.

Локальные механизмы применяются, когда сервер не должен видеть исходные значения пользователей. Их типичные области применения:

- **опросы;**
- **пользовательская телеметрия;**
- **сбор статистики с устройств;**
- **задачи, выполняемые без доверенного оператора данных.**

Главное ограничение локальной DP — более низкая точность по сравнению с глобальной DP при том же числе участников.

Сводная таблица механизмов

Механизм	Тип выхода	Гарантия	Основное применение
Лаплас	Числовой	ε-DP	Счетчики, суммы, гистограммы
Гаусс	Числовой или векторный	(ε, δ)-DP	Векторы, градиенты, ML
Экспоненциальный	Дискретный выбор	ε-DP	Категории, top-k, выбор признаков
Randomized Response	Локальный бинарный ответ	ε-LDP	Опросы, телеметрия
Дискретный Лаплас / геометрический	Целочисленный	ε-DP	Целочисленные счетчики
DP-SGD	Обучение модели	Обычно (ε, δ)-DP	Нейронные сети и ML
PATE	Обучение через ансамбль	Обычно (ε, δ)-DP	Обучение модели-«студента» по зашумленным ответам

Приложение Г.
ОСНОВНЫЕ СВОЙСТВА ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ

Постобработка

Если механизм M удовлетворяет требованиям (ε, δ)-DP, то любая функция от его результата также удовлетворяет (ε,δ)-DP: g(M(D)) сохраняет ту же DP-гарантию. Это свойство называется устойчивостью к постобработке.

Практические следствия:

- нормировка DP-статистики не требует дополнительного бюджета;
- визуализация DP-результата не требует дополнительного бюджета;
- обучение модели на уже выпущенном DP-синтетическом наборе не требует дополнительного бюджета;
- анализ уже опубликованного DP-результата не ухудшает формальную гарантию.

Ограничение: постобработка не улучшает гарантию. Если результат соответствует (ε,δ)-DP, то последующая обработка не делает его автоматически более приватным.

Кроме того, постобработка не отменяет правовые и организационные ограничения на использование результата.

Последовательная композиция

Если к одним и тем же данным последовательно применяются несколько DP-механизмов, то бюджеты приватности накапливаются.

Если механизм M_i имеет гарантию $(\varepsilon_i, \delta_i)$ для $i = 1, \ldots, k$, то базовая оценка суммарной гарантии будет следующей:

$$\varepsilon_{total} = \sum \varepsilon_i$$

$$\delta\_total = \sum \delta_i$$

Практический пример: если выполнить 10 независимых запросов с ε = 0,2, то в рамках базовой композиции суммарный бюджет составит $\varepsilon_{total} = 2$.

Продвинутая композиция и учет бюджета

Базовая композиция часто дает завышенную оценку расхода бюджета, особенно при большом числе шагов обучения.

Для более точного учета применяются следующие инструменты:

- продвинутая композиция;
- RDP-accountant (на основе дивергенции Реньи);
- zCDP-accountant (концентрированная в нуле дифференциальная приватность);
- PRV-accountant (Privacy Random Variable — учет на основе случайной величины потерь);
- специальные инструменты в библиотеках DP-SGD.

Их практический смысл в том, что система должна не просто суммировать запросы, а вести формальный учет бюджета приватности, применяя для этого выбранный метод, и фиксировать итоговые значения пар показателей (ε, δ).

Усиление приватности подвыборкой

Если механизм применяется не ко всей базе, а к случайной подвыборке, приватность может усиливаться.

Интуитивная трактовка этого утверждения проста: если конкретная запись попадает в обработку только с вероятностью q, то ее влияние на результат уменьшается.

Это свойство особенно важно для DP-SGD, где каждый шаг обучения выполняется на фрагменте (мини-батче), а не на всей выборке.

Точная формула усиления зависит от способа подвыборки, механизма и метода учета бюджета, поэтому в практических системах нужно применять для расчетов методику privacy accountant, а не грубую формулу с вычислением приватности вручную.

Групповая приватность

Если механизм защищает одну запись с параметром ε, то защита группы из k записей обычно требует бюджета порядка $k\varepsilon$.

Иными словами, при увеличении защищаемого вклада растет и необходимый для его защиты шум.

Практические выводы таковы:

- **при использовании дифференциальной приватности на уровне записей (record-level DP) защищается одна запись;**
- **при применении дифференциальной приватности на уровне пользователей (user-level DP) нужно учитывать все записи пользователя;**
- **если пользователь может внести до k записей, вклад нужно ограничить или перейти к учету роста чувствительности.**

Без ограничения вклада user-level DP может стать практически неприменимой из-за слишком большого шума.

Параллельная композиция

Если данные разбиты на непересекающиеся части и каждая единица приватности попадает только в одну из частей, то применение DP-механизмов к разным частям не обязательно суммирует бюджеты.

В простом случае суммарная гарантия определяется наибольшим среди имеющихся частей бюджетом:

$$\varepsilon_{total} = max\ \varepsilon_i$$

$$\delta_{total} = max\ \delta_i$$

Условие применимости: один и тот же субъект или одна и та же защищаемая запись не должны участвовать в нескольких частях одновременно.

Если это условие нарушено, необходимо использовать последовательную композицию или более строгую оценку.

Сводная таблица механизмов

Свойство	Практический смысл
Постобработка	Работа с уже выпущенным DP-результатом не требует дополнительного бюджета
Последовательная композиция	Повторные запросы к одним и тем же данным расходуют общий бюджет
Продвинутая композиция	Дает более точный учет бюджета при большом числе операций
Подвыборка	Случайное включение записей может усиливать приватность
Групповая приватность	Защита группы требует большего шума
Параллельная композиция	Для непересекающихся частей берется максимум среди бюджетов, а не сумма

Приложение Д.
БЮДЖЕТ ПРИВАТНОСТИ И УПРАВЛЕНИЕ ИМ

Понятие бюджета приватности

Бюджет приватности — это заранее установленный предел совокупного раскрытия информации через DP-механизмы.

Каждый запрос к исходным данным, каждый запуск генератора, каждое обучение модели или выпуск новой версии результата могут расходовать часть бюджета.

Бюджет должен задаваться до начала обработки данных и фиксироваться в документации.

Необходимо определить как минимум следующие характеристики:

- **единицу приватности;**
- **допустимые значения ε и δ;**
- **список операций, расходующих бюджет;**
- **метод учета композиции;**
- **порядок действий при исчерпании бюджета.**

Что расходует бюджет

Бюджет расходуется при обращении к исходным данным через DP-механизм.

Примеры операций, расходующих бюджет:

- **публикация DP-статистики;**
- **повторный запрос к исходным данным;**
- **обучение модели с DP;**
- **запуск DP-генератора синтетических данных;**
- **выпуск новой версии синтетического набора;**
- **подбор параметров, если он использует исходные данные.**

Не расходуют новый бюджет операции, выполняемые только над уже выпущенным DP-результатом:

- **построение графиков;**
- **нормирование таблиц;**
- **обучение модели на опубликованном DP-синтетическом наборе;**
- **расчет агрегатов по уже выпущенному DP-набору.**

Инструменты учета бюджета

Для простых запросов может быть достаточно базовой композиции. Для сложных процессов, особенно при обучении моделей, требуется специальный учет бюджета.

Инструмент	Где применяется
Базовая композиция	Небольшое число простых запросов
Продвинутая композиция	Большое число повторных запросов
RDP-accountant	DP-SGD, обучение моделей
zCDP-accountant	Гауссовы механизмы, аналитические оценки
PRV-accountant	Точный численный учет бюджета
PATE-accountant	Обучение через ансамбль учителей

В отчете следует указывать не только итоговые ε и δ, но и метод, применявшийся для их расчета.

Выбор ε

Универсального безопасного значения ε не существует. Его выбор зависит от следующих параметров:

- **чувствительность данных;**
- **размер выборки;**
- **единица приватности;**
- **число повторных операций;**
- **допустимая потеря качества;**
- **модель угроз;**
- **правовой и организационный контекст.**

Малое значение ε дает более сильную формальную гарантию, но увеличивает шум. Большое значение ε сохраняет больше полезности, но ослабляет гарантию приватности.

Предлагаем следующие ориентиры по интерпретации величины ε.

Диапазон ε	Интерпретация
$\varepsilon < 1$	Строгая гарантия, часто сопровождаемая значимой потерей качества
$1 \le \varepsilon \le 4$	Умеренный диапазон, требует проверки полезности и устойчивости к атакам
$4 < \varepsilon \le 8$	Практический диапазон для некоторых задач синтеза и ML, требует обоснования
$\varepsilon > 8$	Слабая формальная гарантия, необходима отдельная оценка остаточных рисков

Эти диапазоны не являются нормативными порогами. Они могут использоваться только как ориентиры для обсуждения и должны подтверждаться проверкой качества и актуальных рисков.

Выбор δ

Параметр δ должен быть малым по сравнению с размером набора данных.

Практическое правило такое:

$$\delta \ll 1 / n$$

Здесь n — число защищаемых единиц приватности.

Для чувствительных данных часто выбирают более строгие значения, например порядка $1/n^2$ или ниже. Конкретное значение должно сопровождаться обоснованием в отчете.

Управление бюджетом в проекте

Рекомендуемый порядок управления бюджетом приватности следующий:

- определить цель обработки;
- определить единицу приватности;
- выбрать модель соседства;
- установить предельные значения ε_max и δ_max;
- распределить бюджет между операциями;
- указать метод учета композиции;
- обеспечить ведение журнала расхода бюджета;
- при исчерпании бюджета остановить новые обращения к исходным данным;
- пересматривать бюджет только в соответствии с формальной процедурой изменения требований

Приведем пример записи в журнале расхода бюджета.

Операция	Механизм	Единица приватности	E	δ	Метод учета
Публикация гистограммы	Лаплас	Пользователь	0,5	0	Базовая композиция
Обучение модели	DP-SGD	Пользователь	2,0	10^{-8}	RDP-accountant
Синтез данных	DP-Copula	Пользователь	1,5	10^{-8}	Гауссов механизм / RDP

Бюджет приватности должен рассматриваться как управляемый ресурс проекта. Нельзя выбирать ε после просмотра результата или многократно запускать механизм до получения «хороших» данных без учета расхода бюджета.

Корректная DP-система должна фиксировать единицу приватности, механизм, параметры, число запусков и итоговый расход бюджета.

ОБ АВТОРАХ ДОКЛАДА

АЛЕКСЕЙ НЕЙМАН
Исполнительный директор
Ассоциации больших данных,
руководитель FIT Academy of Russia,
Master of Data Science, CDMP, PMP

ВАЛЕРИЙ ХВАТОВ
Специалист
по кибербезопасности
и распределенным вычислениям,
технический директор
DGT Network

ПЕТР ЕМЕЛЬЯНОВ
CEO,
ООО «Блумтех»

АЛЕКСАНДР ПАРТИН
Адвокат и партнер
Privacy Advocates,
соучредитель «РППА.Офис»,
сопредседатель
Privacy & Legal Innovation
кластера РАЭК, CIPP/E, CIPM

СПИСОК ЛИТЕРАТУРЫ

Основополагающие работы

[1] Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000
[2] Narayanan A., Shmatikov V. Robust De-anonymization of Large Sparse Datasets. IEEE S&P 2008
[3] Dwork, C. (2006). Differential Privacy. ICALP 2006, LNCS 4052, pp. 1-12.
[4] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006, LNCS 3876, pp. 265-284.
[5] Dwork, C., Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.
[6] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M. (2006). Our Data, Ourselves: Privacy Via Distributed Noise Generation. EUROCRYPT 2006.

Механизмы и алгоритмы

[7] McSherry, F., Talwar, K. (2007). Mechanism Design via Differential Privacy. FOCS 2007.
[8] Abadi, M. et al. (2016). Deep Learning with Differential Privacy. ACM CCS 2016, pp. 308-318.
[9] Mironov, I. (2017). Renyi Differential Privacy. IEEE CSF 2017, pp. 263-275.
[10] Bun, M., & Steinke, T. (2016). Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In M. Hirt & A. D. Smith (Eds.), Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, October 31–November 3, 2016, Proceedings, Part I (LNCS 9985, pp. 635–658). Springer. / https://doi.org/10.1007/978-3-662-53641-4_24
[11] Erlingsson, U. et al. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. ACM CCS 2014.
[12] Kairouz, P. et al. (2015). The Composition Theorem for Differential Privacy. ICML 2015.

Теоретические границы

[13] Bassily, R., Smith, A., Thakurta, A. (2014). Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. FOCS 2014.
[14] Duchi, J., Jordan, M., Wainwright, M. (2018). Minimax Optimal Procedures for Locally Private Estimation. JASA, 113(521), 182-201.
[15] Bun, M., Steinke, T., & Ullman, J. (2017). Make Up Your Mind: The Price of Online Queries in Differential Privacy. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2017) (pp. 1306–1325). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974782.85.
[16] Aumüller, M., Lebeda, C. J., & Pagh, R. (2021). Differentially Private Sparse Vectors with Low Error, Optimal Space, and Fast Access. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21), pp. 1223–1236. Association for Computing Machinery. https://doi.org/10.1145/3460120.3484735.
[17] Cohen-Addad, V., Fan, C., Lattanzi, S., Mitrović, S., Norouzi-Fard, A., Parotsidis, N., & Tarnawski, J. (2022). Near-Optimal Correlation Clustering with Privacy. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 33702–33715.
[18] Yeom, S. et al. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. IEEE CSF 2018.

Метрическая DP и расширения

[19] Chatzikokolakis, K. et al. (2013). Broadening the Scope of Differential Privacy Using Metrics. PETS 2013.
[20] Papernot, N. et al. (2017). Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. ICLR 2017 (PATE).
[21] Lin, Z. et al. (2023). Differentially Private Synthetic Data via Foundation Model APIs. ICML 2023 (Private Evolution / DPSDA).
[22] Gopi, S. et al. (2021). Numerical Composition of Differential Privacy. NeurIPS 2021 (PRV Accountant).

Стандарты и нормативные документы

[23] NIST Privacy Framework Version 1.0. National Institute of Standards and Technology, 2020.
[24] NIST SP 800-188: De-Identifying Government Datasets. Draft, NIST, 2016.
[25] NIST IR 8062: An Introduction to Privacy Engineering and Risk Management. NIST, 2017.
[26] ISO/IEC 20889:2018. Privacy enhancing data de-identification terminology and classification of techniques.
[27] ISO/IEC 29101:2018. Privacy architecture framework.
[28] ПНСТ 1.11.164-1.361.25. Защита приватности — методы синтеза данных. Части 1–3. Росстандарт / ТК 164, 2025.
[29] Федеральный закон от 27.07.2006 152-ФЗ «О персональных данных». Российская Федерация.

Российская практика

[30] АБД Пояснительная записка 4 к серии ПНСТ: Анализ международного опыта и российских кейсов применения синтеза данных с DP. 2025.
[31] АБД. Кейс ABD-MA-DPCOPULA-2024-001: Маркетинговая атрибуция с PSI и DP-синтезом. АБД, 2024.
[32] АБД. Кейс ABD-HH-DPSYNTH-2025-002: Синтетические резюме HeadHunter. АБД, 2025.
[33] АБД. Кейс ABD-FRAUD-DPSYNTH-2025-003: Антифрод-конвейер с Stratified DP. АБД, 2025.

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

www.rubda.ru

Адрес: Москва,
Пресненская набережная, 10с2

+7 (495) 252-72-60
info@rubda.ru