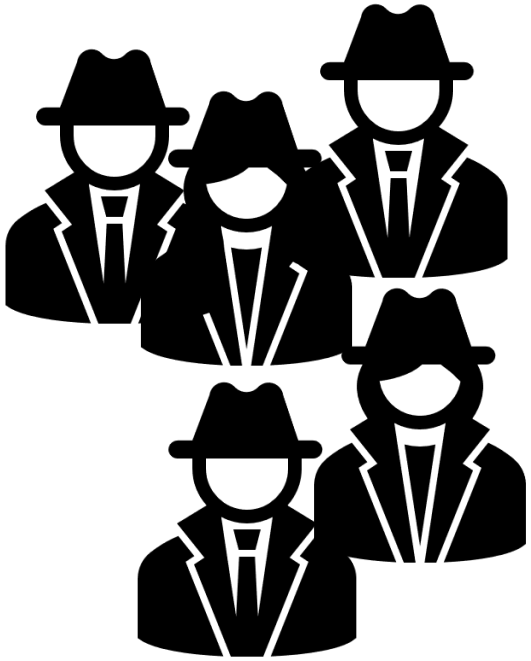
# Предотвращение разглашения информации

**Что такое раскрытие информации?**

Ненадлежащее присвоение конфиденциальной информации субъекту данных, будь то частное лицо или организация.

**Что такое предотвращение разглашения информации?**

Методы, используемые для предотвращения ненадлежащего присвоения конфиденциальной информации субъекту данных.



United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Виды разлашения

## Разглашение информации об установлении личности

- Субъект данных идентифицируется по опубликованному файлу

## Разглашение характерных признаков

- Конфиденциальная информация о субъекте данных раскрывается через опубликованный файл

## Разглашение логически обусловленной (инференциальной) информации

- Опубликованные данные позволяют определить значение какой-либо характеристики лица более точно, чем это было бы возможно в противном случае

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных

**Цели:**

- Предотвращать идентификацию отдельных респондентов
- Предотвращать преднамеренное и непреднамеренное разглашение личной информации физических лиц

**Выбор метода основывается на том, является ли:**

- Агрегированные оценки (в формате подсчета частоты или агрегированных данных о величине)
- Микроданные (отдельные единицы измерения)


United States Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Скрытие ячеек

Используя пороговое значение, отдельные ячейки определяются как чувствительные, а значения подавляются.

**Table 4: Example -- With Disclosure**

**Number of Delinquent Children by County and Education Level of Household Head**

Education Level of Household Head

| County | Low | Medium | High | Very High | Total |
|---|---|---|---|---|---|
| Alpha | 15 | **1*** | **3*** | **1*** | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | **3*** | 10 | 10 | **2*** | 25 |
| Delta | 12 | 14 | 7 | **2*** | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

**Table 6: Example -- Without Disclosure, Protected by Suppression**

**Number of Delinquent Children by County and Education Level of Household Head**

Education Level of Household Head

| County | Low | Medium | High | Very High | Total |
|---|---|---|---|---|---|
| Alpha | 15 | D | D | D | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | D | D | 10 | D | 25 |
| Delta | D | 14 | D | D | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

NOTE: D indicates data withheld to limit disclosure.

При создании таблиц обращайте внимание на пустые ячейки. Следует избегать выпуска таблиц с большим количеством пустых ячеек. Таблицы с большим количеством пустых ячеек при необходимости могут быть объединены в более широкую географию.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Произвольное округление

Значения ячеек округляются на основе произвольного решения о том, следует ли округлять в большую или меньшую сторону.

**Table 4: Example -- With Disclosure**

**Number of Delinquent Children by County and Education Level of Household Head**

Education Level of Household Head

| County | Low | Medium | High | Very High | Total |
|---|---|---|---|---|---|
| Alpha | 15 | **1*** | **3*** | **1*** | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | **3*** | 10 | 10 | **2*** | 25 |
| Delta | 12 | 14 | 7 | **2*** | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

**Table 7: Example -- Without Disclosure, Protected by Random Rounding**

**Number of Delinquent Children by County and Education Level of Household Head**

Education Level of Household Head

| County | Low | Medium | High | Very High | Total |
|---|---|---|---|---|---|
| Alpha | 15 | 0 | 0 | 0 | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | 5 | 10 | 10 | 0 | 25 |
| Delta | 15 | 15 | 10 | 0 | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Контролируемое округление

Значения ячеек округляются на основе произвольного решения о том, следует ли округлять в большую или меньшую сторону, но в некоторых случаях округление контролируется, чтобы гарантировать, что суммы строк и столбцов теперь работают.

**Table 4: Example -- With Disclosure**

**Number of Delinquent Children by County and Education Level of Household Head**

| | Education Level of Household Head | | | | |
|---|---|---|---|---|---|
| County | Low | Medium | High | Very High | Total |
| Alpha | 15 | **1*** | **3*** | **1*** | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | **3*** | 10 | 10 | **2*** | 25 |
| Delta | 12 | 14 | 7 | **2*** | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

**Table 8: Example -- Without Disclosure, Protected by Controlled Rounding**

**Number of Delinquent Children by County and Education Level of Household Head**

| | Education Level of Household Head | | | | |
|---|---|---|---|---|---|
| County | Low | Medium | High | Very High | Total |
| Alpha | 15 | 0 | 5 | 0 | 20 |
| Beta | 20 | 10 | 10 | 15 | 55 |
| Gamma | 5 | 10 | 10 | 0 | 25 |
| Delta | 10 | 15 | 5 | 5 | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Сжатие ячеек

Группы ячеек (строк или столбцов) объединяются в одну строку.

**Table 2: Example -- With Disclosure**

**Number of Beneficiaries by Monthly Benefit Amount and County**

| | Monthly Benefit Amount | | | | | | |
|---|---|---|---|---|---|---|---|
| County | $0-19 | $20-39 | $40-59 | $60-79 | $80-99 | $100+ | Total |
| A | 2 | 4 | 18 | 20 | 7 | 1 | 52 |
| B | -- | - | 7 | 9 | - | - | 16 |
| C | -- | 6 | 30 | 15 | 4 | - | 55 |
| D | - | - | 2 | -- | - | - | 2 |

**Table 3: Example -- Without Disclosure**

**Number of Beneficiaries by Monthly Benefit Amount and County**

| | Monthly Benefit Amount | | | | | | |
|---|---|---|---|---|---|---|---|
| County | $0-19 | $20-39 | $40-59 | $60-79 | $80-99 | $100+ | Total |
| A and B | 2 | 4 | 25 | 29 | 7 | 1 | 68 |
| C and D | -- | 6 | 32 | 15 | 4 | - | 57 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Управляемая табличная настройка

Требующие защиты конфиденциальности ячейки заменяются значением, которое находится на “достаточном расстоянии” от истинного значения.

Не требующие защиты конфиденциальности ячейки настраиваются минимально, чтобы обеспечить совпадение итоговых значений.

Может быть выполнено в сочетании с округлением в качестве указания пользователю данных на то, что ячейка была скорректирована.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Верхнее кодирование и нижнее кодирование

Верхнее кодирование и нижнее кодирование подвергают цензуре данные, превышающие или опускающиеся ниже определенного значения.

Эти данные были закодированы сверху выше 250 тысяч долларов в столбце 4 и снизу ниже 10 тысяч долларов в столбце 5.

| Ид.№ | Возраст | Фактический достаток | Верхнее кодирование | Нижнее кодирование |
|---|---|---|---|---|
| 1 | 26 | $25988 | $25988 | $25988 |
| 2 | 34 | $32458 | $32458 | $32458 |
| 3 | 22 | $75483 | $75483 | $75483 |
| 4 | 45 | $18574 | $18574 | $18574 |
| 5 | 64 | $15302956 | $250000 | $15302956 |
| 6 | 41 | $192834 | $192834 | $192834 |
| 7 | 19 | $33859 | $33859 | $33859 |
| 8 | 42 | $196 | $196 | $10000 |
| 9 | 37 | $274858 | $250000 | $274858 |
| 10 | 42 | $6492 | $6492 | $10000 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Перекодирование

Перекодирование изменяет каждое значение данных в наборе микроданных. Обычно это означает, что данные перекодируются в интервал либо напрямую, либо путем округления.

Эти данные были перекодированы с интервалами в 25 тысяч долларов.

| Ид.№ | Возраст | Фактический достаток | Перекодированный |
|---|---|---|---|
| 1 | 26 | $25988 | $25 тыс.-$50 тыс. |
| 2 | 34 | $32458 | $25 тыс.-$50 тыс. |
| 3 | 22 | $75483 | $50 тыс.-$75 тыс. |
| 4 | 45 | $18574 | $0 тыс.-$25 тыс. |
| 5 | 64 | $15302956 | $250 тыс.+ |
| 6 | 41 | $192834 | $150 тыс.-$200 тыс. |
| 7 | 19 | $33859 | $25 тыс.-$50 тыс. |
| 8 | 42 | $196 | $0 тыс.-$25 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Обмен (своппинг) данными

| Ид.№ | Возраст | Фактический достаток | Почтовый индекс |
|---|---|---|---|
| 1 | 26 | $25988 | 20942 |
| 2 | 34 | $32458 | 47892 |
| 3 | 22 | $75483 | 91003 |
| 4 | 45 | $18574 | 47743 |
| 5 | 64 | $15302956 | 10293 |
| 6 | 41 | $192834 | 88391 |
| 7 | 19 | $33859 | 20341 |
| 8 | 42 | $196 | 33061 |
| 9 | 37 | $274858 | 09281 |
| 10 | 42 | $6492 | 77801 |

| Ид.№ | Возраст | Фактический достаток | Почтовый индекс |
|---|---|---|---|
| 1 | 26 | $25,988 | 47892 |
| 2 | 34 | $32,458 | 20942 |
| 3 | 22 | $33,859 | 91003 |
| 4 | 45 | $15,302,956 | 47743 |
| 5 | 64 | $18,574 | 10293 |
| 6 | 41 | $274,858 | 88391 |
| 7 | 19 | $75,483 | 20341 |
| 8 | 42 | $196 | 77801 |
| 9 | 37 | $192,834 | 9281 |
| 10 | 42 | $6,492 | 33061 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Синтетические данные

Синтетические данные, - это смоделированные статистические данные, публикуемые в формате, который очень напоминает конфиденциальные данные.

Синтетические данные полезны, когда другие методы дают результаты, которые не имеют смысла.

Для создания синтетических данных на основе существующих данных строится модель, идентифицируются уникальные записи в данных и заменяются значением, сгенерированным на основе модели.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Шумовое воздействие

Данные, полученные от каждого респондента, немного искажены в ту или иную сторону.

Шум может быть добавлен главным образом к требующим защиты конфиденциальности ячейкам, оставляя другие ячейки в основном неизменными.

| Ид.№ | Фирма | Материальные затраты | Затраты на шум |
|---|---|---|---|
| 1 | 26 | $25988 | $26994 |
| 2 | 34 | $32458 | $32078 |
| 3 | 22 | $75483 | $76464 |
| 4 | 45 | $18574 | $18368 |
| 5 | 64 | $15302956 | $13798675 |
| 6 | 41 | $192834 | $173570 |
| 7 | 19 | $33859 | $35572 |
| 8 | 42 | $196 | $215 |
| 9 | 37 | $274858 | $250616 |
| 10 | 42 | $6492 | $6851 |

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Методы защиты данных: Микроданные

**При выпуске микроданных:**

1. Исключите информацию, которая непосредственно идентифицирует отдельных лиц, в частности данные о геолокации.
2. Скрывайте данные, которые могут косвенно идентифицировать отдельных лиц.
3. Вносите неопределенность в отчетные данные. Производите обмен данными.

**Чтобы снизить вероятность разглашения, большинство общедоступно используемых файлов микроданных:**

- Включают данные только по выборке населения.
- Не включают очевидные идентификаторы.
- Ограничивают географическую детализацию.
- Ограничивают количество и подробную разбивку категорий внутри переменных в файле.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Обсуждение вопроса об избежании разглашения информации

Каковы требования вашей организации к ограничению разглашения персональных данных?

Какой из этих приемов вы используете? И в каком контексте?

# Дифференциальная приватность

“Чрезмерно точные ответы на слишком много вопросов неизбежно разрушит частную жизнь”.

- Дворк и Рот, "Алгоритмические основы дифференциальной приватности"

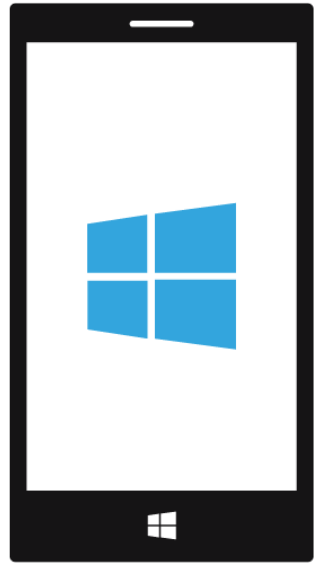
“Слишком много статистических данных, опубликованных слишком точно из конфиденциальной базы данных, почти наверняка раскрывает всю базу данных”

– Динур и Ниссим, 2003

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Работали ли вы с системой с дифференциальной приватностью?

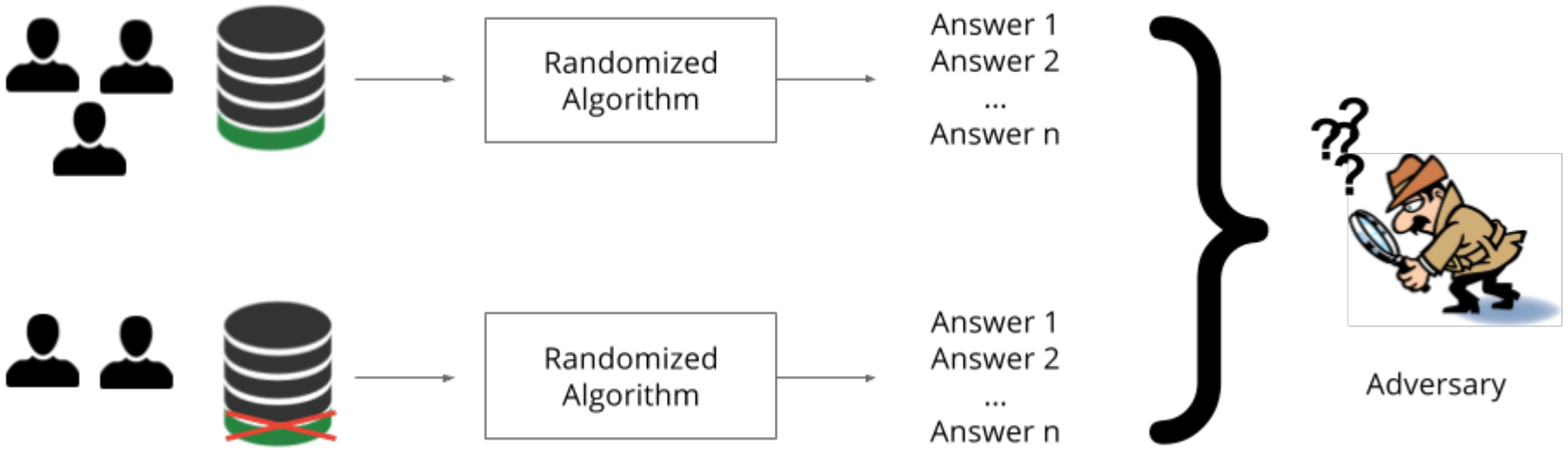


United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Что такое дифференциальная приватность?

Дифференциальная приватность, - это математическая основа для точного измерения риска разглашения, связанного с каждой публикацией конфиденциальных данных.



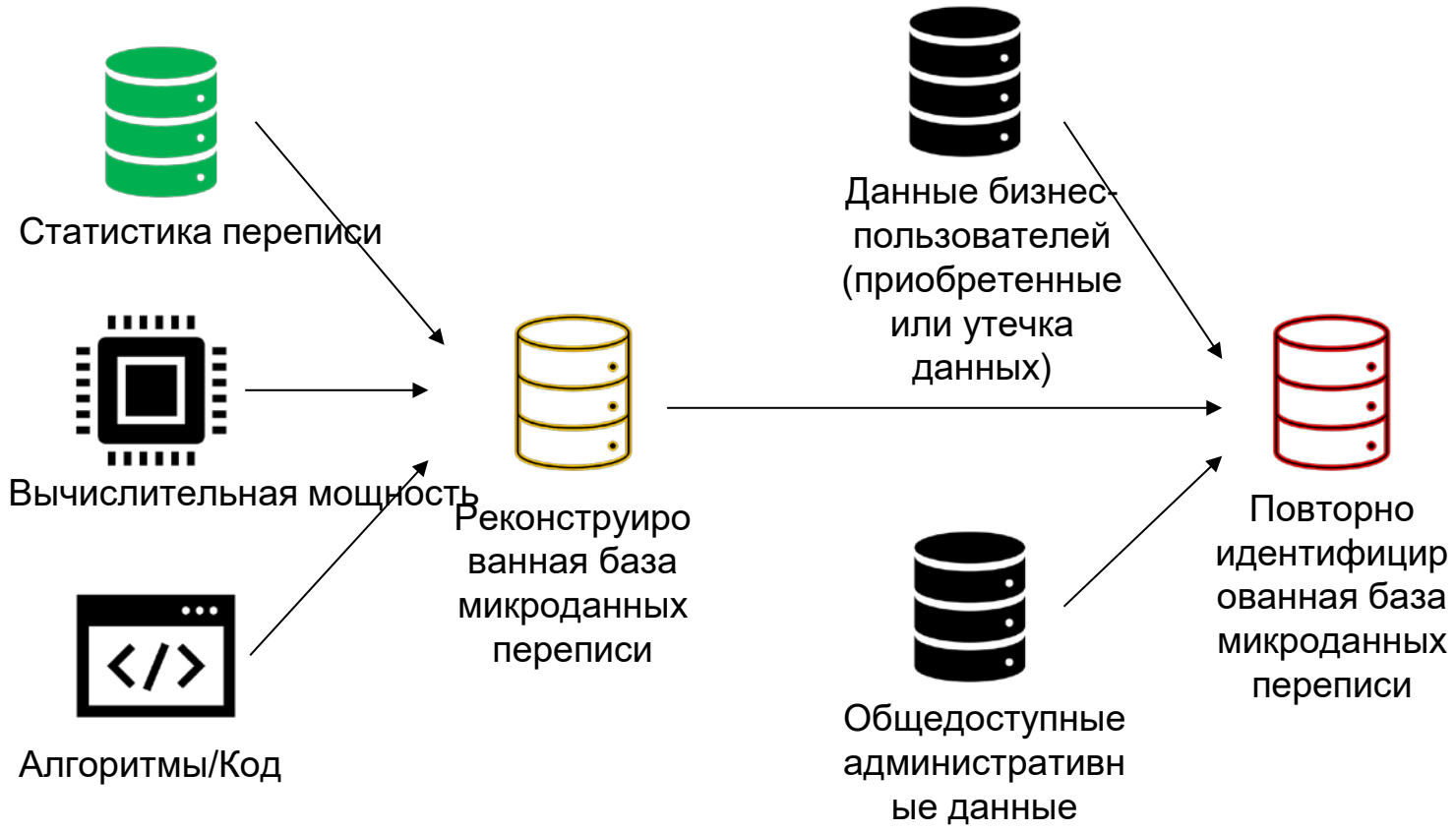

[Приватность и машинное обучение: два неожиданных союзника?]

# Атака с реконструкцией и повторной идентификацией

Для переписи населения 2010 года Бюро переписи опубликовало статистические данные о 150 миллиардах человек при населении в 310 миллионов человек. Это почти **500** статистических данных для каждого человека из анкеты из 10 вопросов!

**Атака с реконструкцией**, – использование общедоступных данных или статистики для (частичного) создания базы данных, идентичной частной базе данных. Статистика, опубликованная о базе данных, позволяет это сделать.

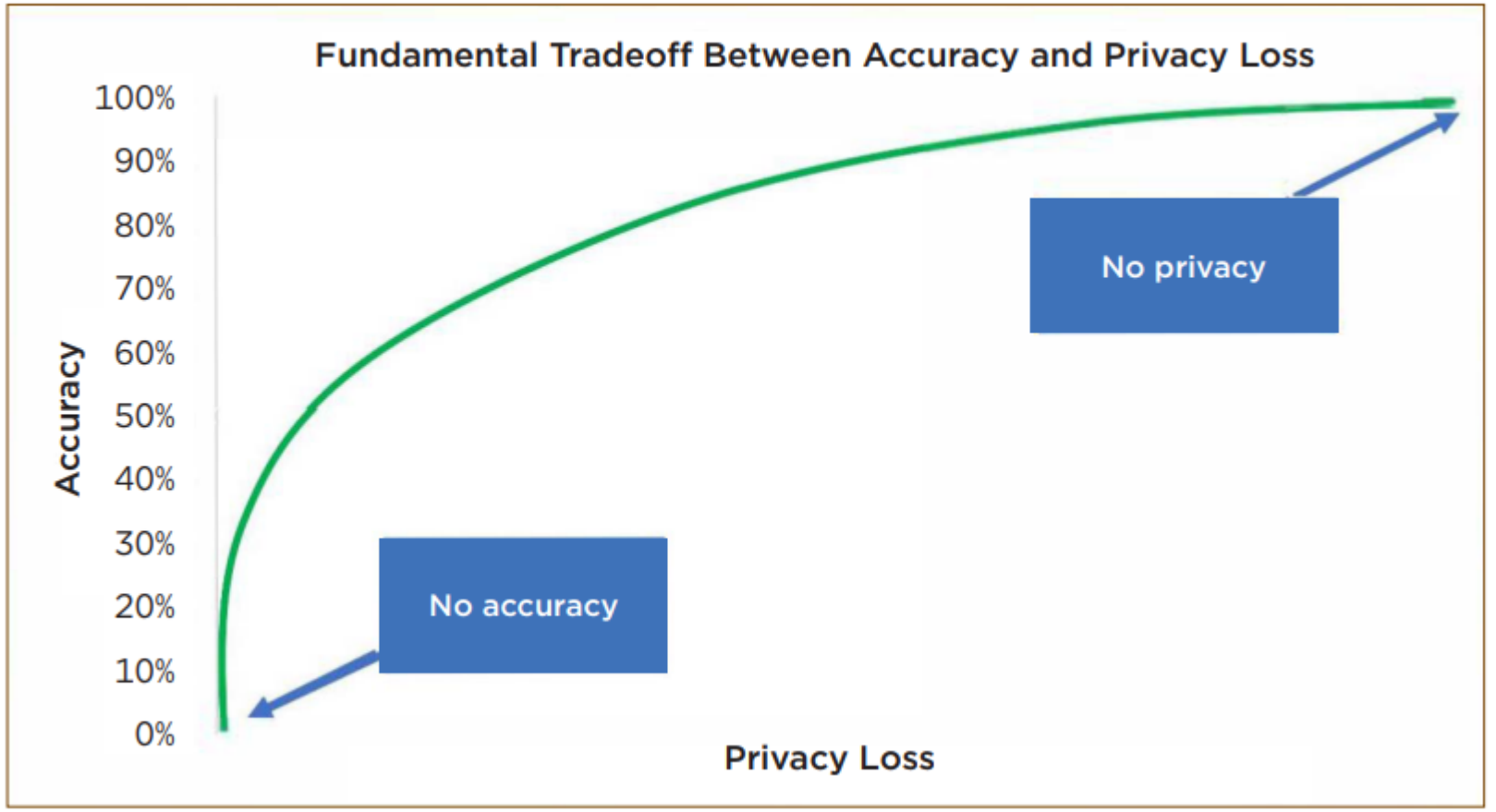
**Атака с повторной идентификацией (атака с привязкой)**, – использование общедоступных данных или вспомогательной информации для идентификации физических лиц в частной базе данных.




United States Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Приватность по сравнению с Точностью




Уклонение от раскрытия информации в ходе переписи 2020 года: Введение

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Преимущества и недостатки

## Преимущества

- Закрыто по составу
- Устойчиво к последующей обработке
- Ориентировано на будущее
- Доказуемо и настраиваемо
- Публично и объяснимо
- Защищает от атак на восстановление базы данных

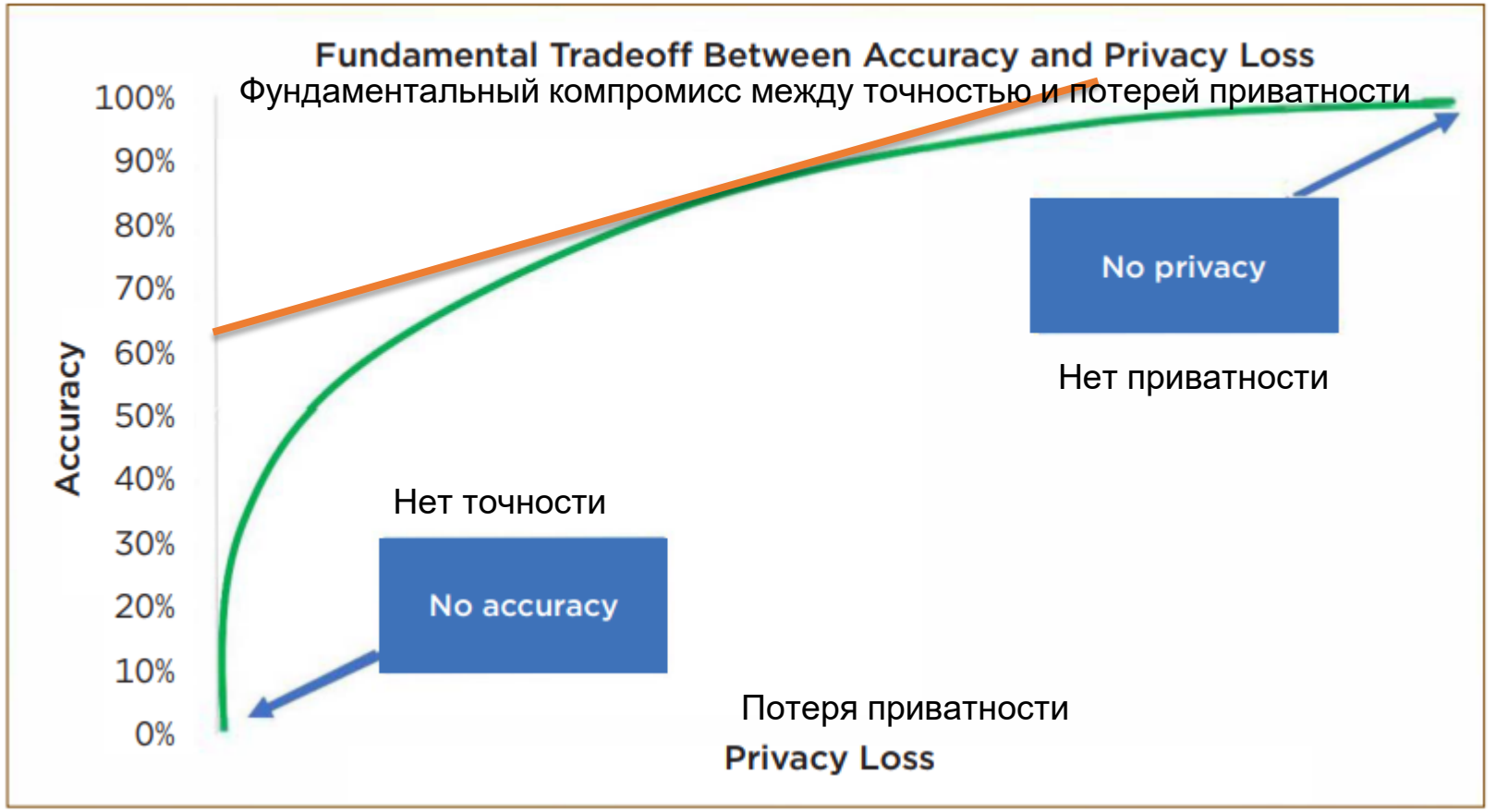
## Недостатки

- Вся страна должна быть обработана одновременно для достижения наилучшей точности.
- Каждое использование личных данных должно учитываться в бюджете потерь приватности.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

# Потеря приватности по сравнению с Точностью, как социальным выбором




[Уклонение от разглашения информации в ходе переписи 2020 года: Введение]

# Потеря приватности по сравнению с обсуждением точности

Какие факторы влияют на то, как ваши организации подходят к компромиссу между потерей приватности и точностью?

Как, по-вашему, ваши заинтересованные стороны оценивают компромисс между потерей приватности и точностью? Совпадает ли их точка зрения с вашей?

Существуют ли какие-либо ограничения или препятствия, с которыми вы можете столкнуться при попытке защитить приватность или повысить точность?