Приватность и машинное обучение

Илья Миронов
Facebook AI

План

- Атаки на приватность, или что-то пошло не так…
- Определение приватности
- Дифференциальная приватность
  - Определение
  - Свойства
- Приватность и машинное обучение

Cynthia Dwork

Knuth Prize (2020)

Hamming Medal (2020)

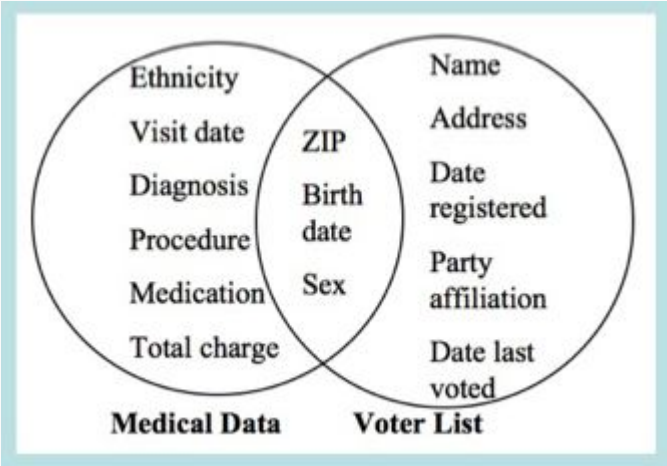
Gödel Prize (2017)

Dijkstra Prize (2007)

Атаки на приватность, или что-то пошло не так…

Вильям Велд против Латаньи Суини

Massachusetts Group Insurance Commission (1997):
Анонимизированная медицинская история всех работников штата (больничные, диагнозы, лекарства)



Латанья Суини (аспирант MIT):
$20 – список избирателей Кембриджа

Д.р.: 31 июля, 1945
почтовый индекс 02138

64%

американцев однозначно определяются почтовый индекс + д.р. + пол

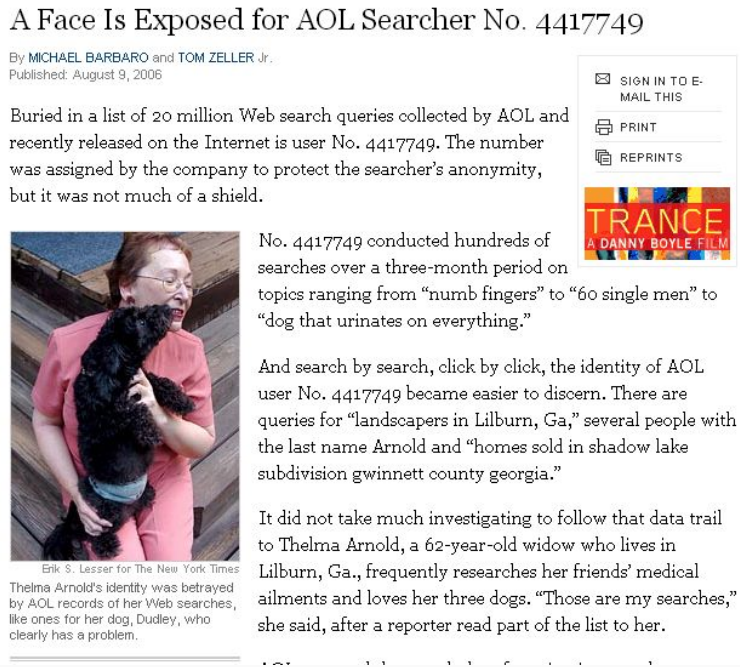
Преимущество атакующей стороны

- Дополнительная информация

Запросы AOL

4 августа, 2006: AOL Research публикует анонимные запросы 650,000 пользователей

9 августа: New York Times



A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS

TRANCE
A DANNY BOYLE FILM

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Преимущество атакующей стороны

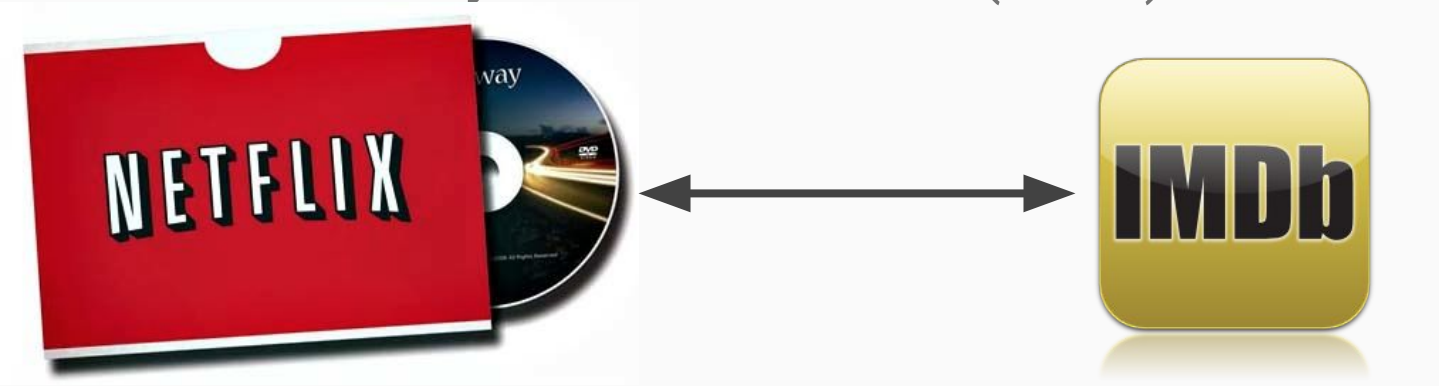
- Дополнительная информация
- Достаточно преуспеть на малой доле данных

Приз Netflix

октябрь 2006: Netflix объявляет Netflix Prize

- 10% пользователей
- в среднем 200 оценок от одного пользователя

Narayanan, Shmatikov (2006):



Приз Netflix

- Noam Chomsky in Our Times

- Fahrenheit 9/11

- Jesus of Nazareth

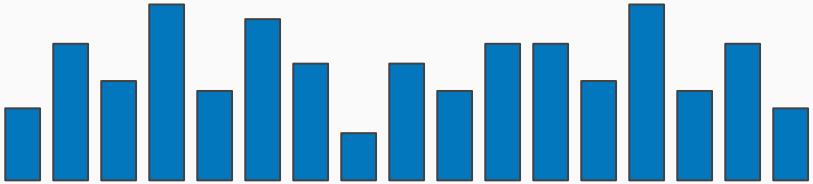
- Queer as Folk

Преимущество атакующей стороны

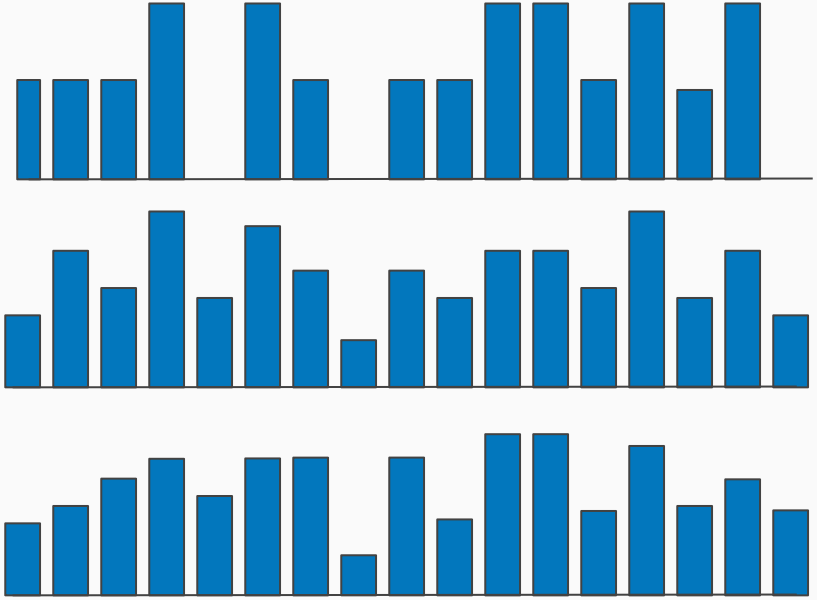
- Дополнительная информация
- Достаточно преуспеть на малой доле данных
- Большая размерность данных

Генетические данные

Homer et al., “Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays”, PLoS Genetics, 2008



Тестирование членства



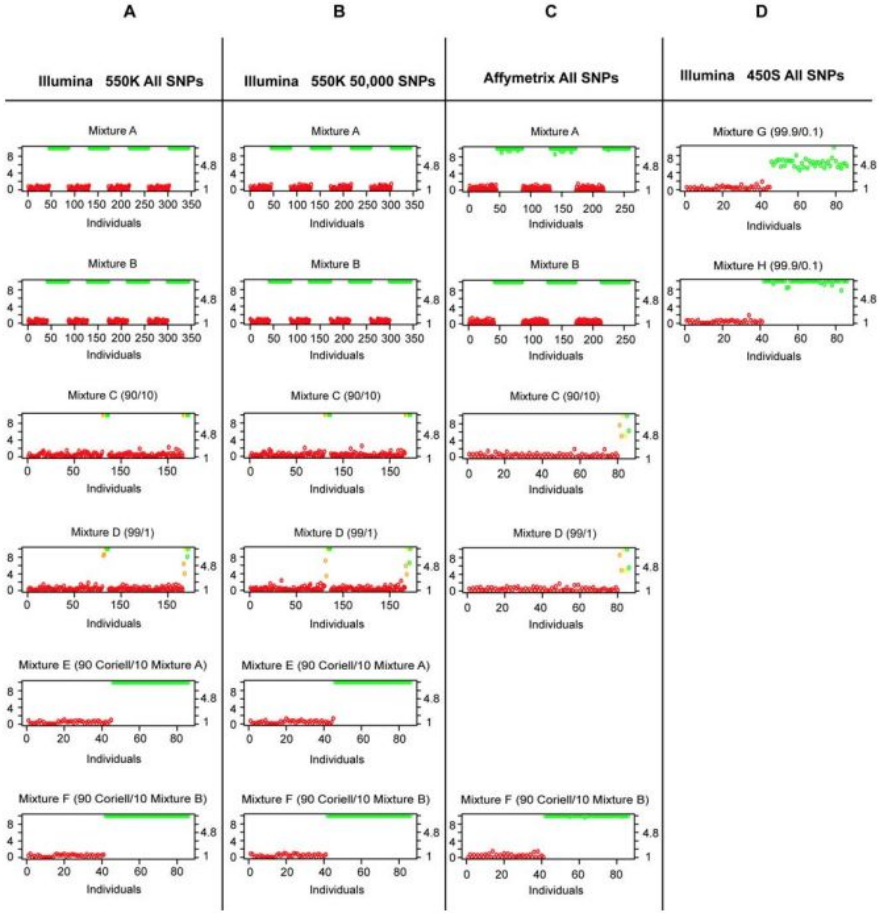


Figure 3. Experimental validation using a series of mixtures (see Methods A–F) assayed on the Affymetrix GeneChip 5.0, Illumina BeadArray 550 and the Illumina 450S Duo Human BeadChip. The x-axis shows each individual in the CEU HapMap population, the left y-axis shows the p-value (log scaled), and the right y-axis shows the value of the test statistic. For mixtures A, B, E and F those in the mixture are colored green and those not in the mixture are colored red. For mixtures C and D those individuals who are not in the mixtures are colored red, those individuals who are related to the 1% or 10% individuals in the mixtures are colored orange, those individuals who are related to the 90% or 99% are colored yellow, and those people in the mixture are colored green. In all mixtures, the identification of the presence of a person's genomic DNA was possible.

doi:10.1371/journal.pgen.1000167.g003

...неделю спустя

Zerhouni, NIH Director:

> “As a result, the NIH has removed from open-access databases the aggregate results (including P values and genotype counts) for all the GWAS that had been available on NIH sites”

Преимущество атакующей стороны

- Дополнительная информация
- Достаточно преуспеть на малой доле данных
- Большая размерность данных
- Сообразительность

Австралийский ОМС

август 2016: Правительство опубликовало медицинские данные 10% австралийцев (2.9M) за 1984–2014

- Пациент: год рождения, пол
- Госпитализации, процедуры, коды, штат, цена
- Даты изменены ±2 недели

“Чернобыль медицинской де-идентификации”

сентябрь 2016: Сотрудники Мельбурнского Университета ре-идентифицировали политиков, спортсменов, публичных фигур.

- 55 тыс женщин уникально идентифицированы благодаря рождению детей

октябрь 2016: Государство внесло закон, запрещающий ре-идентификацию правительственных данных

Отрицательные результаты

Динур-Ниссим

Данные	0	1	1	0	1	0	0	0	1	1	0	1
запрос: Σ												

Динур-Ниссим 2003:

Если ошибка $o(\sqrt{n})$, то возможно восстановление $n-o(n)$

...даже если 23.9% ошибок произвольные [DMT07]

...достаточно $O(n)$ запросов [DY08]

Дворк-Наор

Условие Тора Далениуса (“семантическая стойкость”):
“Доступ к статистической базе данных не должен позволить узнать больше об индивидууме, чем можно узнать без доступа к базе данных.” (1977)

Дворк-Наор (2006):
Всегда есть дополнительная информация, которая делает эту цель недостижимой.

MPC?

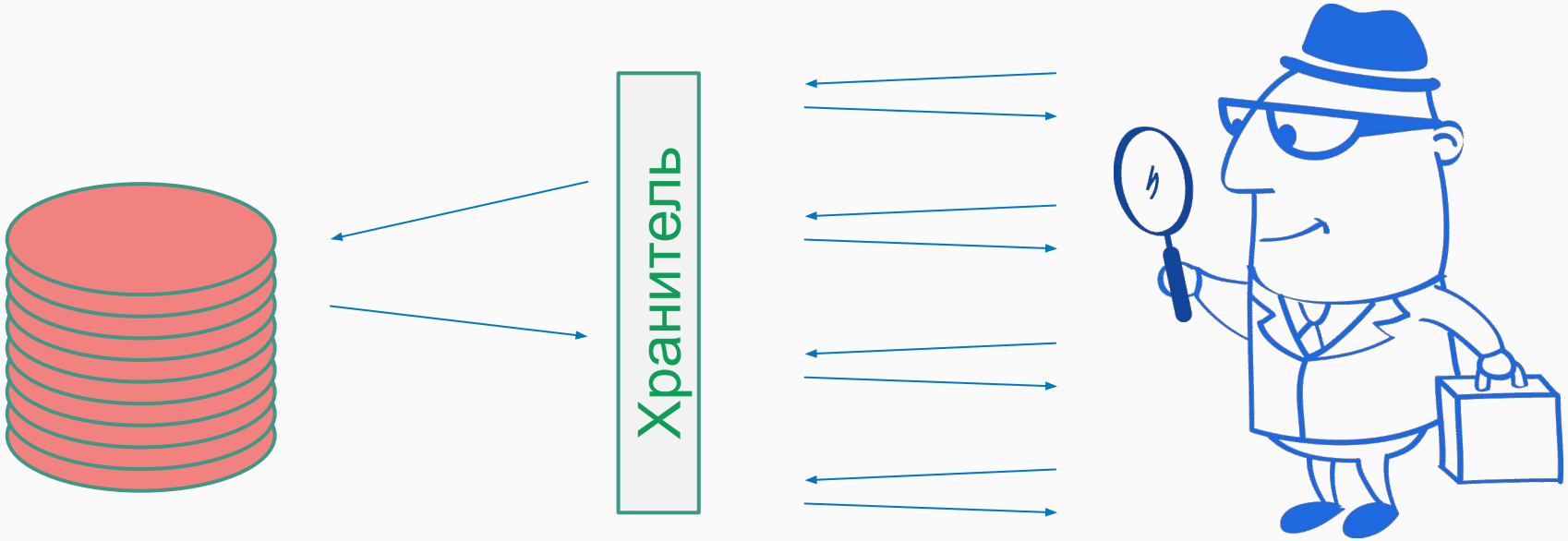
В криптографии **протокол конфиденциального вычисления** (также безопасное, защищенное или тайное многостороннее вычисление, англ. *secure multi-party computation*) — криптографический протокол, позволяющий нескольким участникам произвести вычисление, зависящее от тайных входных данных каждого из них, таким образом, чтобы ни один участник не смог получить никакой информации о чужих тайных входных данных.

https://ru.wikipedia.org/wiki/Протокол_конфиденциального_вычисления

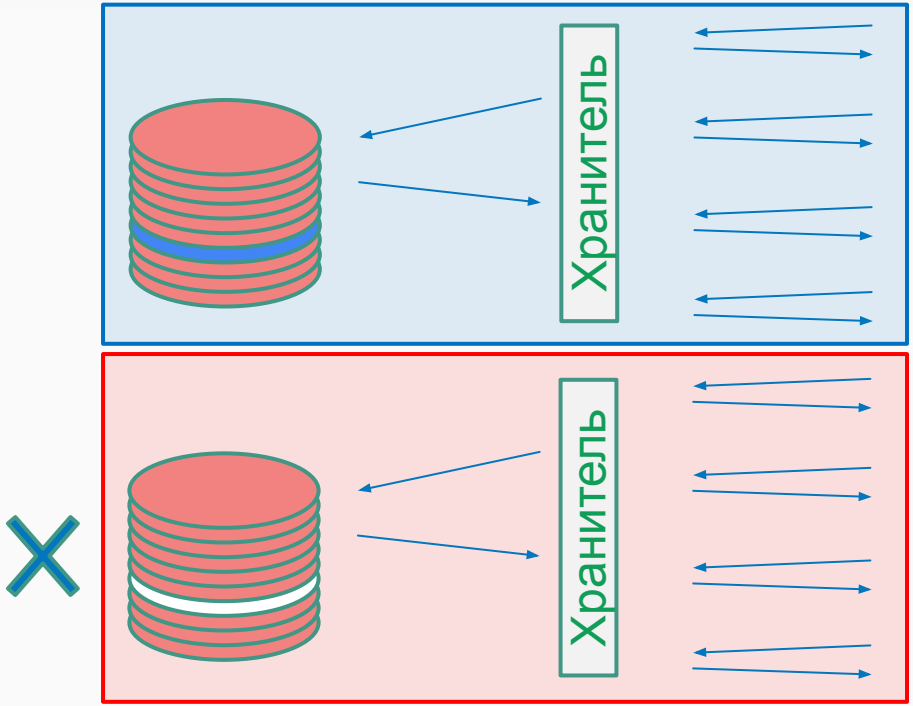
No information is leaked through the transcript except for what can be inferred from the output of the functionality about the parties' inputs

Определение дифференциальной стойкости

Определение приватности



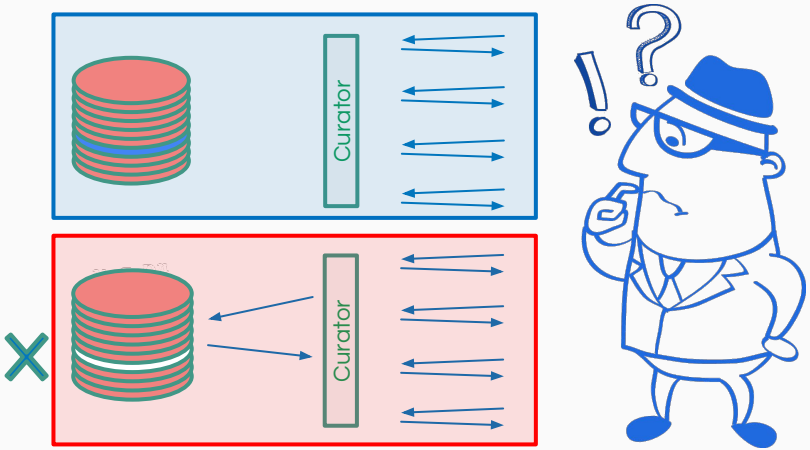
Определение приватности



Дифференциальная приватность

Базы данных D и D' *смежные*, если они отличаются в данных одного человека.

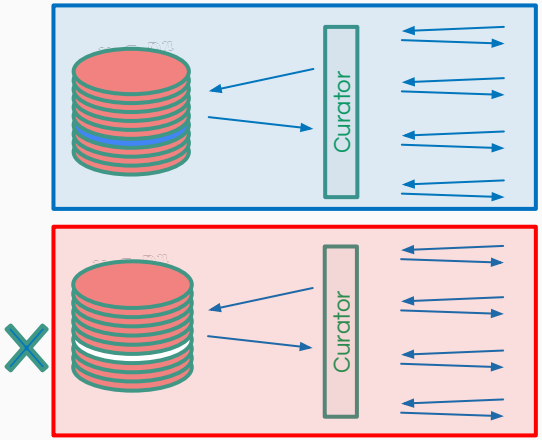
Дифференциальная приватность [DMNS06]: Для смежных баз D и D' распределение $M(D)$ почти такое же, как и $M(D')$.



Дифференциальная приватность

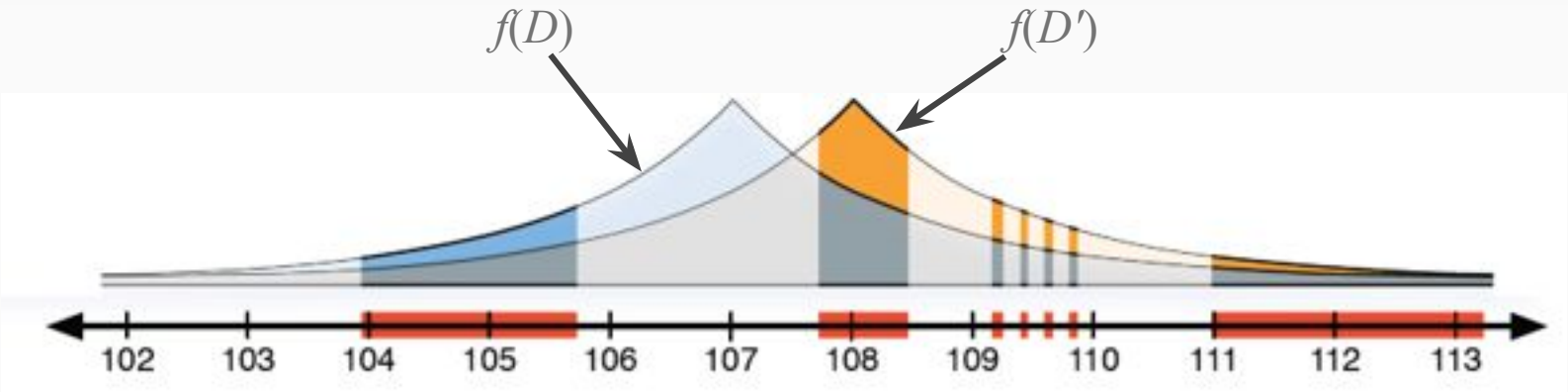
ε-ДП: Для смежных баз D и D' распределение $M(D)$ почти такое же, как и $M(D')$.

$$\forall S: \qquad \Pr[M(D) \in S] \leq \exp(\varepsilon) \cdot \Pr[M(D') \in S].$$



Параметр ε измеряет утечку информации

Интерпретация “плохой результат”



— плохие выводы

— вероятность с записью x

— вероятность без записи x

Байесовская интерпретация

- Априорная вероятность p
- Событие O
- Содержат ли данные x?

$$\frac{p(D \mid O)}{p(D' \mid O)} = \frac{p(D)}{p(D')} \cdot \frac{p(O \mid D)}{p(O \mid D')}$$

$\leq \exp(\varepsilon)$

$\exp(-\varepsilon) \leq$

Дифференциальная приватность

- Устойчива к дополнительной информации
- Пост-обработка:

  Если $M(D)$ д.п., то $f(M(D))$ тоже будет д.п.

- Композиция:

  Результат применения ε_1-DP и ε_2-DP будет $(\varepsilon_1 + \varepsilon_2)$-DP

- Групповая приватность:

Сохраняется, даже если входные данные коррелированы

Каноническая задача

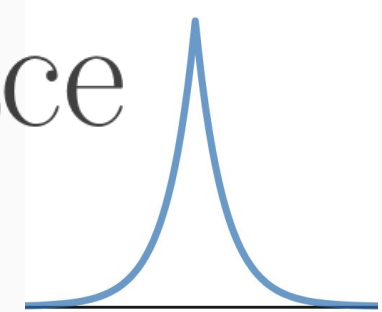
Суммирование бит

$$\sum_{i=1}^{n} x_i + noise$$

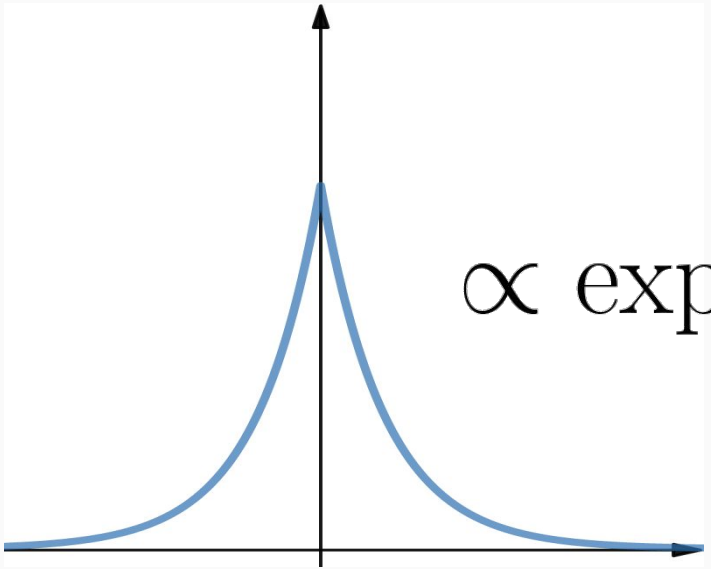
Каноническая задача

Суммирование бит

$$\sum_{i=1}^{n} x_i + \text{Laplace}$$

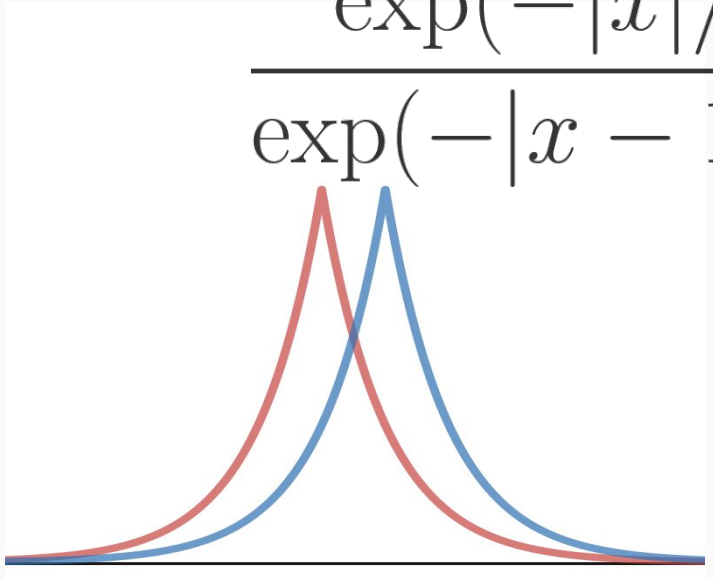


Распределение Лапласа



$$\propto \exp(-|x|/\lambda)$$

Почему ДП?



$$\frac{\exp(-|x|/\lambda)}{\exp(-|x-1|/\lambda)} = \exp\left(\frac{|x-1|-|x|}{\lambda}\right)$$

$$\leq \exp\left(\frac{1}{\lambda}\right)$$

Ослабление дифференциальной приватности

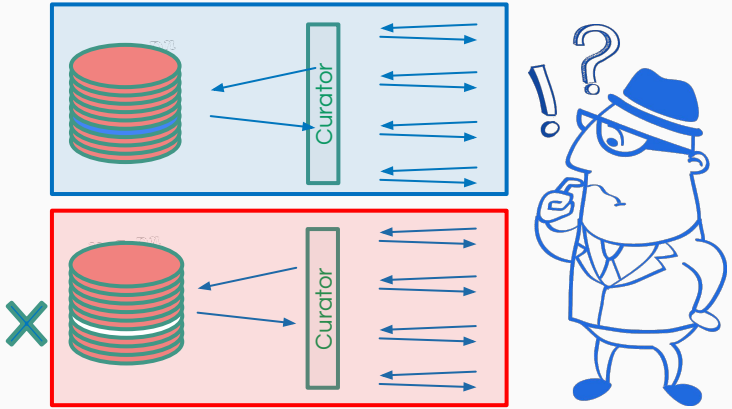
Чем плоха ε-DP

- Слишком строга
  - Псевдослучайный генератор не совместим с DP
- Слишком пессимистична
  - n применений ε-DP — $n \cdot \varepsilon$-DP

Приблизительная дифференциальная приватность

(ε, δ)-ДП: Для смежных баз D и D' распределение $M(D)$ почти такое же, как и $M(D')$.

$$\forall S: \quad \Pr[M(D) \in S] \leq \exp(\varepsilon) \cdot \Pr[M(D') \in S] + \delta.$$



Параметр ε измеряет утечку информации

Параметр δ дает небольшой люфт

Дифференциальная приватность

- Устойчива к дополнительной информации
- Пост-обработка:

  Если $M(D)$ (ε, δ)-DP, то $f(M(D))$ тоже будет (ε, δ)-DP

- Композиция:

  Результат применения $(\varepsilon_1, \delta_1)$-и $(\varepsilon_2, \delta_2)$-DP будет $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP

- Групповая приватность:

  Сохраняется, даже если входные данные коррелированы.

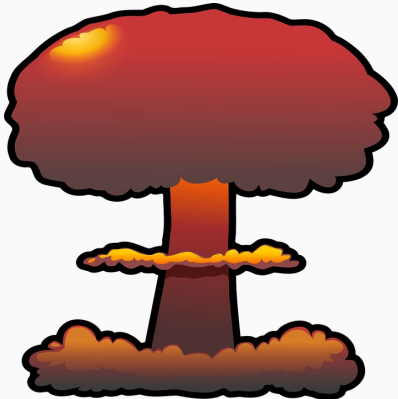
Что может скрываться за δ?

Полная катастрофа:

- С вероятностью δ открыть все
- С вероятностью 1 открыть δ часть данных

Рекомендации

$\delta \ll 1/N$ or $\delta = \mathrm{negl}(1/N)$



Дифференциальная приватность Реньи [M ‘17]

(α, ε)-Rényi Differential Privacy (RDP):

$$\forall D, D' : \mathrm{D}_\alpha(D \| D') \leq \varepsilon$$

Расхождение Реньи

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log E_Q \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right]$$

Ослабление дифференциальной приватности

(∞, ε)-RDP - это ε-DP

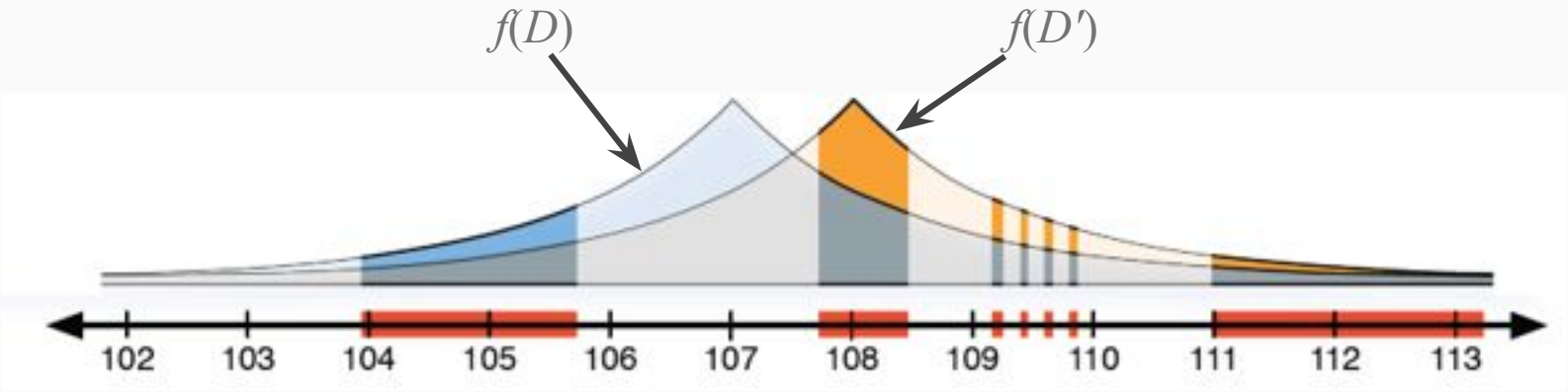
Усиление (ε, δ)-DP

(α, ε)-RDP $\Rightarrow$ $(\varepsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP для любой δ

Дифференциальная приватность Реньи

- Устойчива к дополнительной информации
- Пост-обработка::
  Если $M(D)$ удовлетворяет (α, ε)-RDP , $f(M(D))$ тоже (α, ε)-RDP
- Композиция:
  Результат применения (α, ε_1)- и (α, ε_2)-RDP является $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP
- Групповая приватность

Интерпретация “плохой результат”



— плохие выводы

— вероятность с записью x

— вероятность без записи x

Интерпретация “плохой результат”

ε-Differential Privacy: $\forall S \Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S]$

(α, ε)-Rényi Diff Privacy: $\forall S \Pr[M(D) \in S] \leq (e^{\varepsilon} \cdot \Pr[M(D') \in S])^{1-1/\alpha}$

(ε, δ)-Differential Privacy: $\forall S \Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S] + \delta$

Не бывает катастроф!

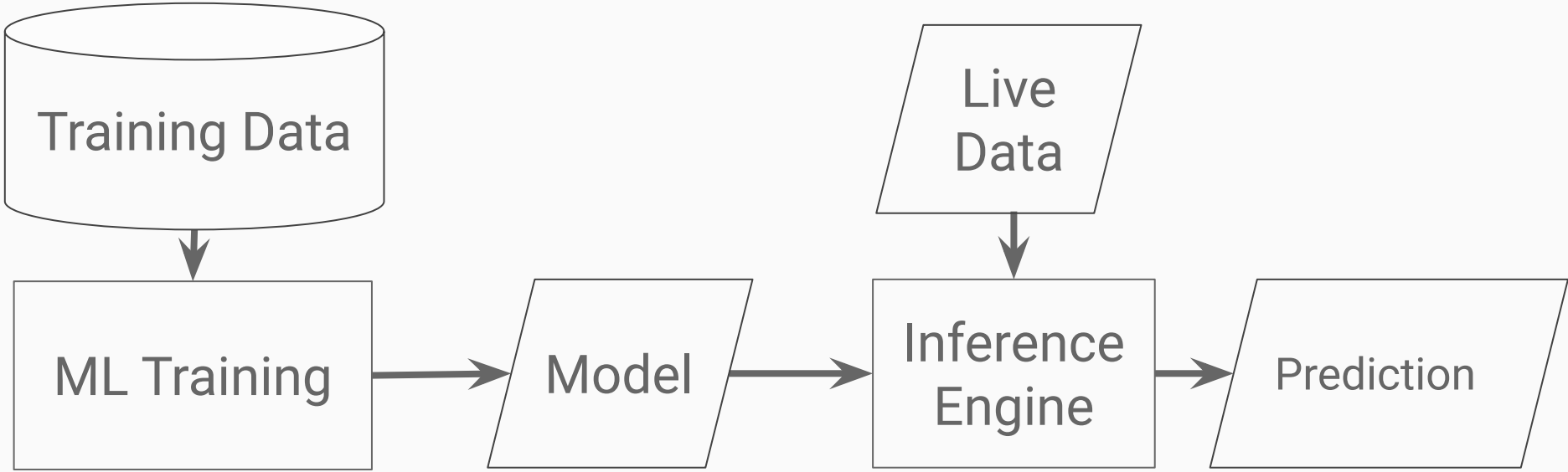
$$\Pr[M(D) \in S] \leq (e^{\varepsilon} \cdot \Pr[M(D') \in S])^{1-1/\alpha}$$

$= 0$

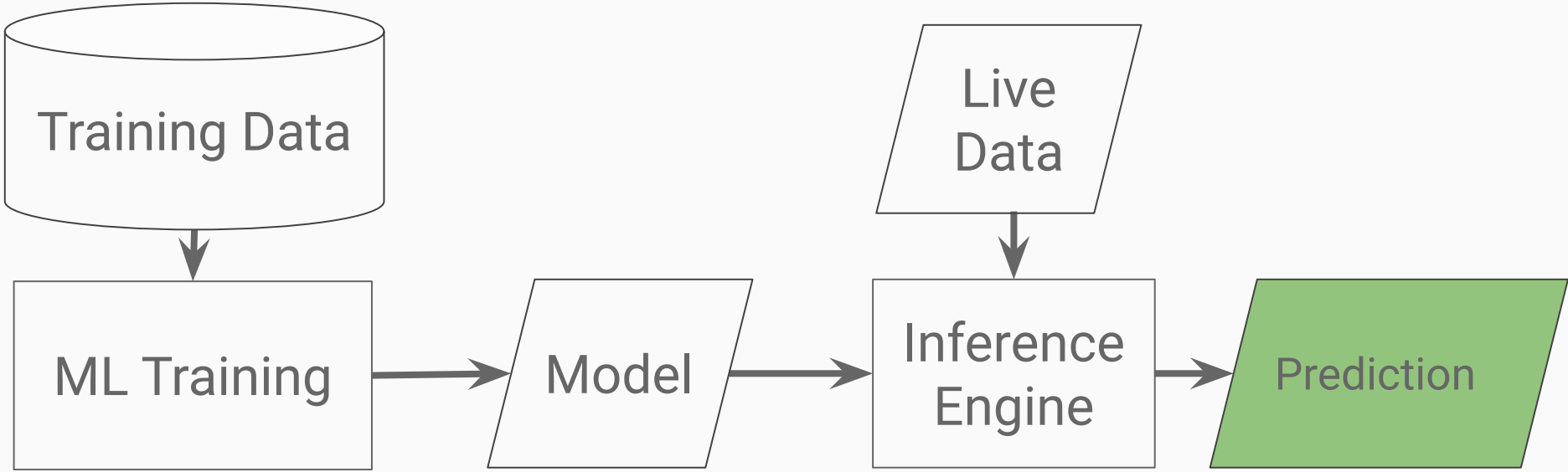
$= 0$

Дифференциальная приватность и машинное обучение

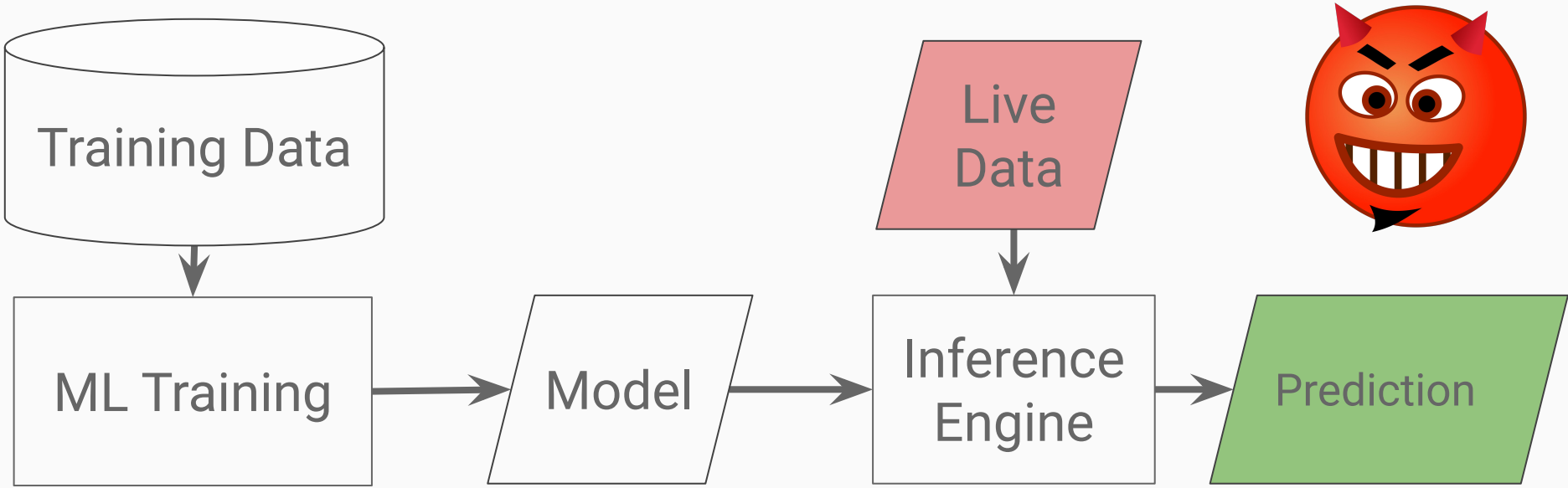
Машинное обучение



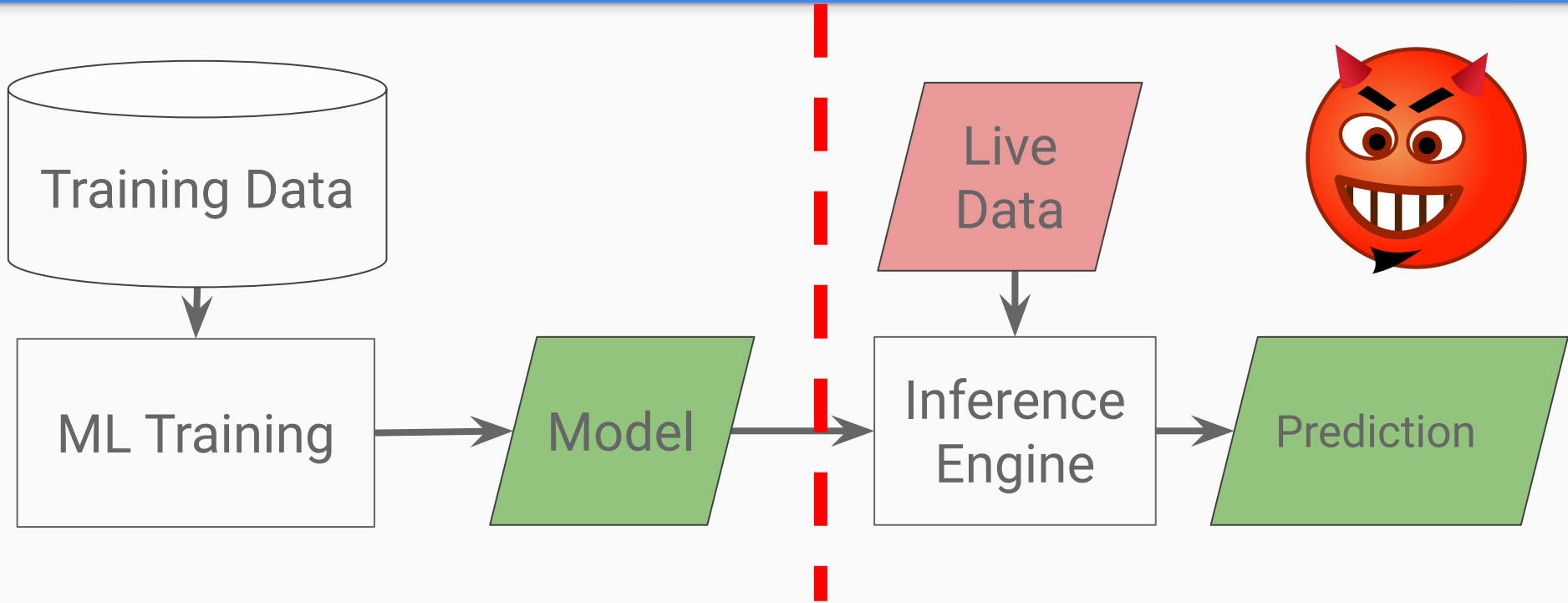
Машинное обучение и модель угроз



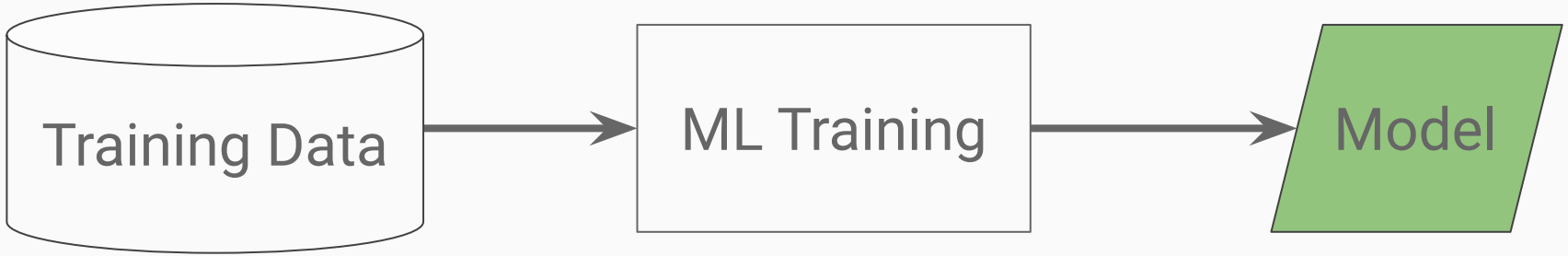
Машинное обучение и модель угроз



Машинное обучение и модель угроз



Машинное обучение и модель угроз



Заблуждение: обобщение гарантирует приватность

Наши модели гарантируют обобщение, значит они приватны

Заблуждение: обобщение гарантирует приватность

Обобщение

- в среднем
- точность модели

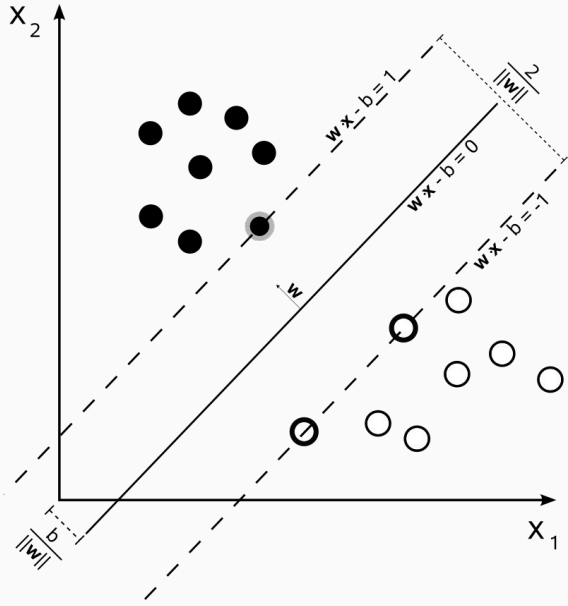
Приватность

- худший случай
- все параметры

Заблуждение: обобщение гарантирует приватность

- Примеры, когда это не так
  - Ближайшие соседи
  - Опорные векторы

- Модели могут быть очень большими
  - Миллиарды параметров



С сомалийского на английский

С сомалийского на английский

С сомалийского на английский

С сомалийского на английский

Somali

Translate from Irish

ag ag ag ag ag ag ag ag ag ag ag ag ag

English

And they came to be at the king 's gate by the valley of the tribes

С сомалийского на английский

Somali

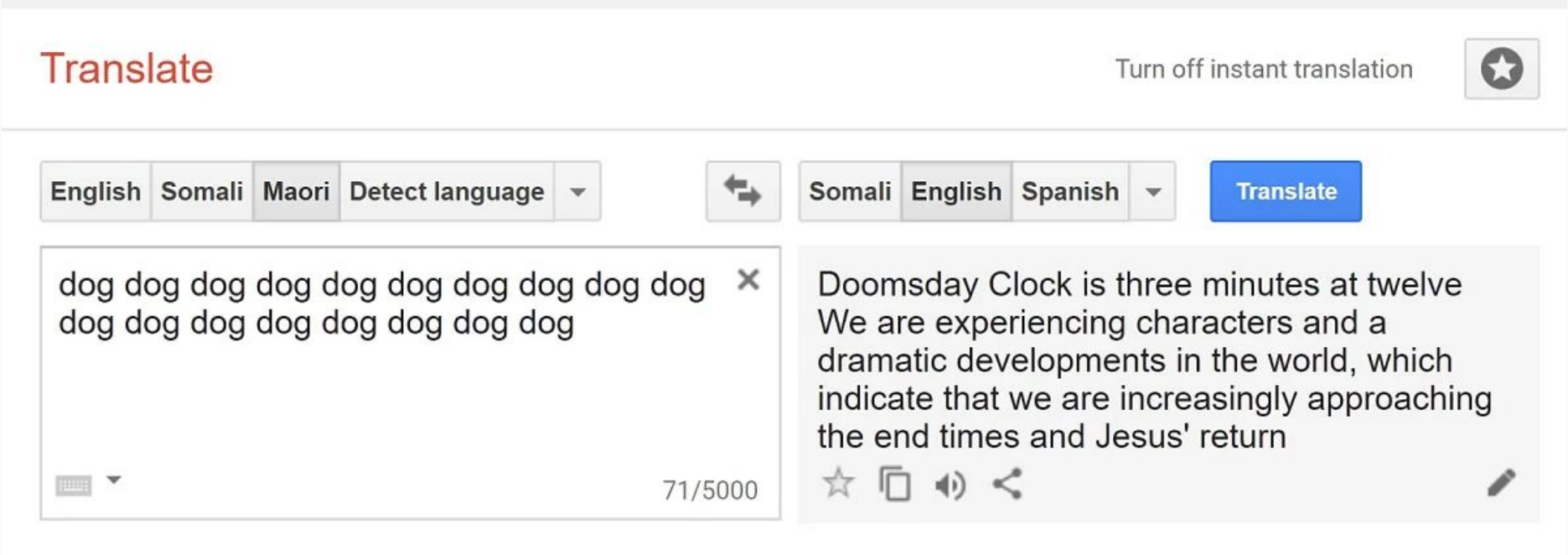
Translate from Irish

ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag

English

Numbers 520,000 agon Numbers Away From the Nation of the Lions

С маори на английский

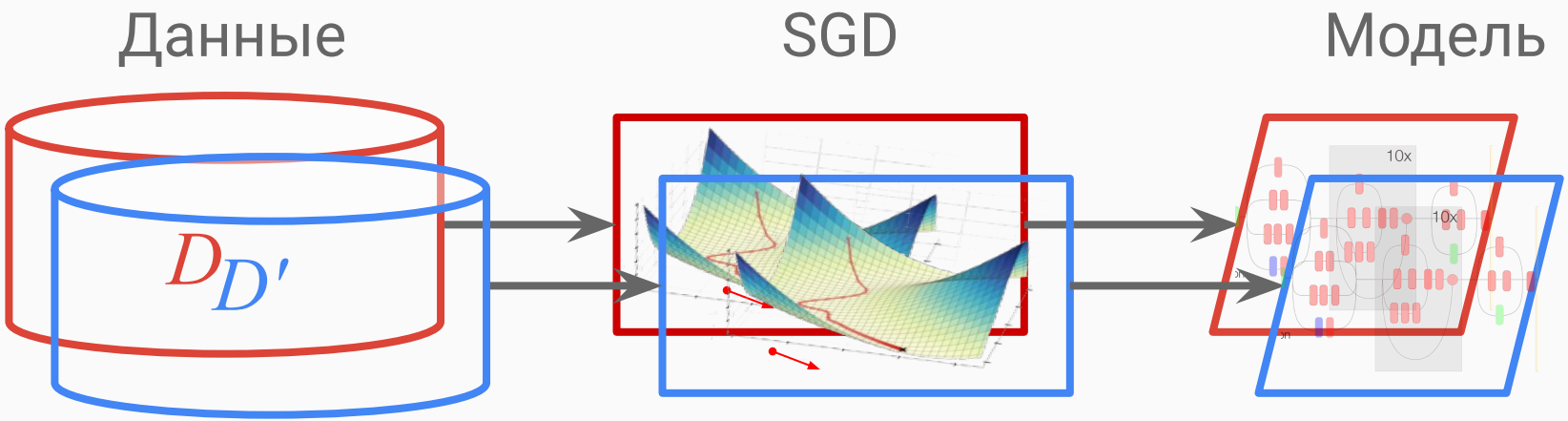


“Understanding Deep Networks Requires Rethinking Generalization”, Zhang et al.’17

data aug	dropout	weight decay	top-1 train	top-5 train	top-1 test	top-5 test
ImageNet 1000 classes with the original labels						
yes	yes	yes	92.18	99.21	77.84	93.92
yes	no	no	92.33	99.17	72.95	90.43
no	no	yes	90.60	100.0	67.18 (72.57)	86.44 (91.31)
no	no	no	99.53	100.0	59.80 (63.16)	80.38 (84.49)
Alexnet (Krizhevsky et al., 2012)			-	-	-	83.6
ImageNet 1000 classes with random labels						
no	yes	yes	91.18	97.95	0.09	0.49
no	no	yes	87.81	96.15	0.12	0.50
no	no	no	95.20	99.14	0.11	0.56

Дифференциальная приватность и машинное обучение

Дифференциально-приватный SGD



Стохастический градиентный спуск (SGD)

Найти $\nabla L(\theta_i)$ на выборке → $\theta_{i+1} := \theta_i - \eta \nabla L(\theta_i)$ → Найти $\nabla L(\theta_{i+1})$ на выборке → $\theta_{i+2} := \theta_{i+1} - \eta \nabla L(\theta_{i+1})$ →

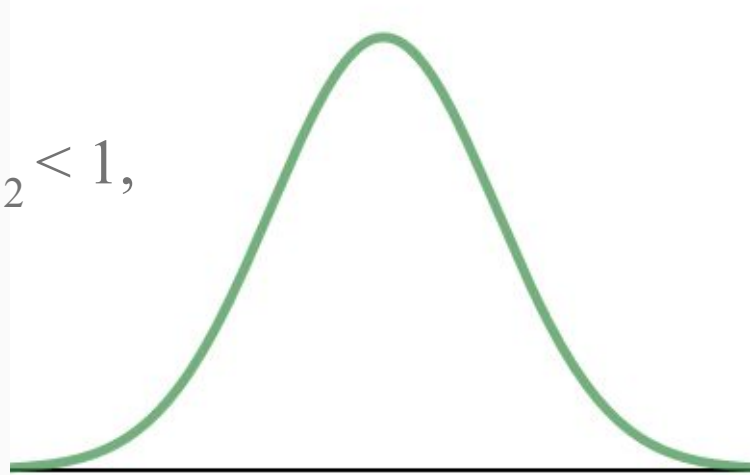
Дифференциальная приватность: гауссовый механизм

ℓ_2-чувствительность $f: \mathcal{D} \to \mathbb{R}^n$:

$$\max_{D,D'} \|f(D) - f(D')\|_2 < 1,$$

тогда гауссовый механизм

$$f(D) + N^n(0, \sigma^2)$$



гарантирует (ε, δ)-DP, где $\delta \approx \exp(-(\varepsilon\sigma)^2/2)$.

Простой рецепт

Чтобы вычислить f с дифференциальной приватностью

1. Оценить сверху чувствительность f
2. Применить гауссовый механизм

Базовая композиция

Если f удовлетворяет $(\varepsilon_1, \delta_1)$-DP, а g $(\varepsilon_2, \delta_2)$-DP, то

$f(D), g(D)$ удовлетворяет $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP

Простой рецепт для сложных функций

Чтобы вычислить сложную функцию f с дифференциальной приватностью

1. Оценить сверху чувствительность частей f
2. Применить гауссовый механизм к каждой части
3. Подсчитать суммарную приватность

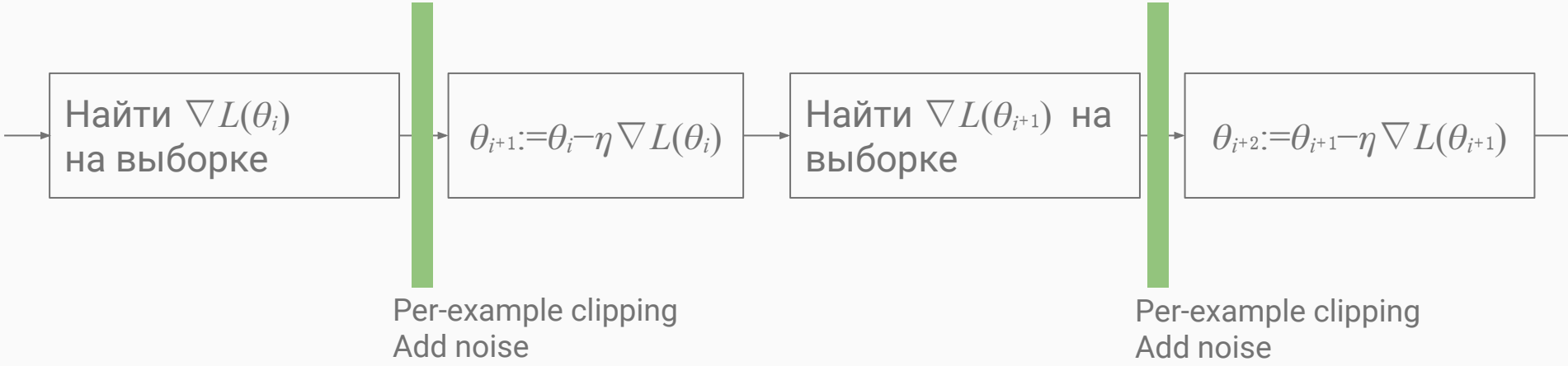
SGD с дифференциальной приватностью?

выборка $\{x_j\}$

$$\nabla L(\theta_i) = \Sigma \nabla L(x_j, \theta_i)$$

Найти $\nabla L(\theta_i)$ на выборке → $\theta_{i+1} := \theta_i - \eta \nabla L(\theta_i)$ → Найти $\nabla L(\theta_{i+1})$ на выборке → $\theta_{i+2} := \theta_{i+1} - \eta \nabla L(\theta_{i+1})$

SGD с дифференциальной приватностью



MNIST

28×28 images

60,000 обучение

10,000 тестирование

Наивный анализ

1. Выбираем $\sigma = \frac{\sqrt{2\log 1/\delta}}{\varepsilon}$ $= 4$
2. Каждый шаг (ε, δ)-DP $(1.2, 10^{-5})$-DP
3. Число шагов T $10{,}000$
4. Композиция: $(T\varepsilon, T\delta)$-DP $(12{,}000, .1)$-DP

Сильная композиция

1. Выбираем $\sigma = \frac{\sqrt{2\log 1/\delta}}{\varepsilon}$ $= 4$
2. Каждый шаг (ε, δ)-DP $(1.2, 10^{-5})$-DP
3. Число шагов T $10{,}000$
4. С. комп: $(\varepsilon\sqrt{T\log 1/\delta}, T\delta)$-DP $(360, .1)$-DP

Dwork, Rothblum, Vadhan, “Boosting and Differential Privacy”, FOCS 2010
Dwork, Rothblum, “Concentrated Differential Privacy”, https://arxiv.org/abs/1603.0188

Используем случайность выборки!

1. Выбираем $\sigma = \frac{\sqrt{2\log 1/\delta}}{\varepsilon}$ $= 4$
2. Каждая выборка - доля q 1%
3. Каждый шаг $(2q\varepsilon, q\delta)$-DP $(.024, 10^{-7})$-DP
4. Число шагов T $10{,}000$
5. С. комп: $(2q\varepsilon\sqrt{T\log 1/\delta}, qT\delta)$-DP $(10, .001)$-DP

S. Kasiviswanathan, H. Lee, K. Nissim, S. Raskhodnikova, A. Smith, “What Can We Learn Privately?”, SIAM J. Comp, 2011

Renyi Differential Privacy Accountant

1. Выбираем $\sigma = \frac{\sqrt{2\log 1/\delta}}{\varepsilon}$ $= 4$
2. Каждая выборка - доля q 1%
3. Отслеживаем приватность Реньи
4. Число шагов T $10{,}000$
5. RDP: $(2q\varepsilon\sqrt{T}, \delta)$-DP $(1.25, 10^{-5})$-DP

Дифференциальная приватность в TensorFlow

tensorflow / privacy

Unwatch 41 | Unstar 692 | Fork 88

Code | Issues 9 | Pull requests 0 | Insights | Settings

Library for training machine learning models with privacy for training data

Edit

machine-learning | privacy | Manage topics

110 commits | 1 branch | 0 releases | 14 contributors | Apache-2.0

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

tensorflower-gardener Check batch_size % microbatches = 0 and calculate privacy budget only... ...

Latest commit ab466b1 9 hours ago

Дифференциальная приватность в PyTorch

facebookresearch / pytorch-dp

Unwatch 16 | Unstar 116 | Fork 16

Code | Issues 1 | Pull requests 1 | Actions | Projects 0 | Security 0 | Insights | FB Internal

Training PyTorch models with differential privacy

38 commits | 1 branch | 0 packages | 0 releases | 10 contributors | Apache-2.0

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download