Como garantir proteção de dados pessoais no seu projeto

VINTA

Apresentação

- Rebeca Sarai 👩‍💻
- Recife, Brazil ☀️ 🌊 🏖️ 🏝️ 🇧🇷
- Formada em Engenharia da Computação
- Aluna de mestrado da Universidade de Pernambuco 👩‍🎓

UPE POLI
ESCOLA POLITÉCNICA
DE PERNAMBUCO

IN PRIMIS SAPIENTIA

VINTA

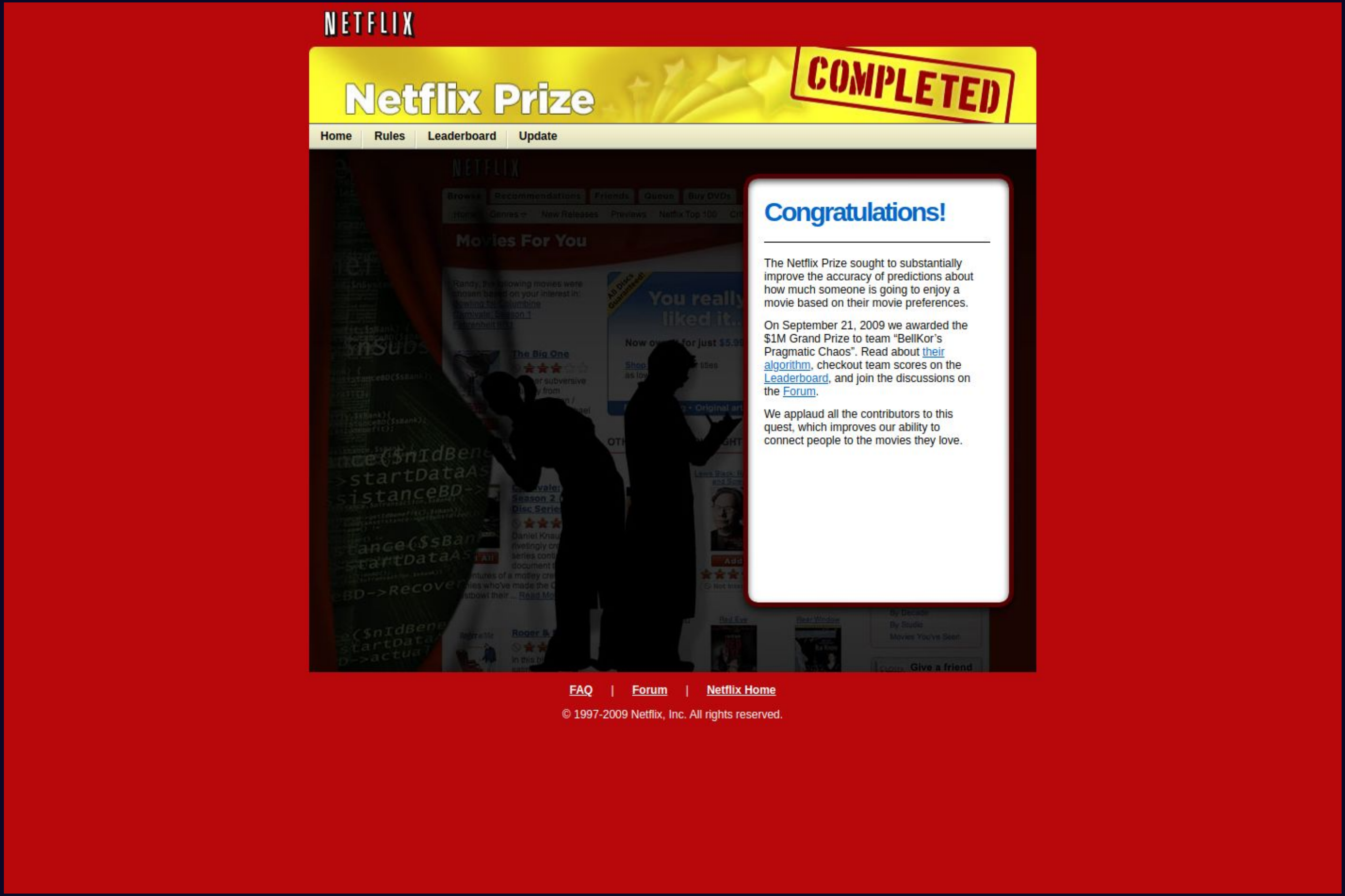
We're a **team of experts** from Brazil.
We help our clients **evolve their products** with **top notch** development and UX techniques.

Get to know us: vintasoftware.com

Como garantir proteção de dados pessoais no seu projeto

VINTA

Não é só substituir nomes...



[Netflix Prize]

@_rebecasarai

BRUCE SCHNEIER SECURITY 12.12.07 09:00 PM

WHY 'ANONYMOUS' DATA SOMETIMES ISN'T

LAST YEAR, NETFLIX published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some ofthe Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

Their research (.pdf) illustrates some inherent security problems with anonymous data, but first it's important to explain what they did and did not do.

They did *not* reverse the anonymity of the entire Netflix dataset. What they did was reverse the anonymity of the Netflix dataset for those sampled users who also entered some movie rankings, under their own names, in the IMDb. (While IMDb's records are public, crawling the site to get them is against the IMDb's terms of service, so the

WIRED

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- **Latest: Strava suggests military users 'opt out' of heatmap as row deepens**

▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

Sensitive information about the location and staffing of military bases and spy outposts around the world has been revealed by a fitness tracking company.

The Guardian

@_rebecasarai

Olhar Digital

@_rebecasarai

Anonimização é possível?

Your Data Were 'Anonymized'? These Scientists Can Still Identify You

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.

The Times

Researchers spotlight the lie of 'anonymous' data

Natasha Lomas @riptari / 7:30 am -03 • July 24, 2019 Comment

Researchers from two universities in Europe have published a method they say is able to correctly re-identify 99.98% of individuals in anonymized data sets with just 15 demographic attributes.

[TechCrunch]

Sim

Outline

- Regulamentações
- Pseudonimização (bem rápido)
- Anonimização
  - k-Anonimato (k-Anonymity)
  - Privacidade Diferencial (Differential Privacy)

Photo by Gia Oris on Unsplash

EU General Data Protection Regulation

Privacy by Design

GRPD Article 25

Pseudonimização

Pseudonimização

Dados pseudonimizados ainda **são considerados dados pessoais**, pois podem ser usados para **re-identificação** se combinados com **informações adicionais**.

Pseudonimização

- Data Masking
- Approximation
- Encryption
- Tokenization

```python
class AbstractUser(AbstractBaseUser, PermissionsMixin):
    username_validator = UnicodeUsernameValidator()
    username = models.CharField(
        _('username'),
        max_length=150,
        unique=True,
        help_text=_('Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.'),
        validators=[username_validator],
        error_messages={
            'unique': _("A user with that username already exists."),
        },
    )
    first_name = models.CharField(_('first name'), max_length=30, blank=True)
    last_name = models.CharField(_('last name'), max_length=150, blank=True)
    email = models.EmailField(_('email address'), blank=True)
```

```
class AbstractUser(Abstra
    username_validator =
    username = models.Ch
        _('username'),
        max_length=150,
        unique=True,
        help_text=_('Requ                                    and @/./+/-/_ only.'),
        validators=[user
        error_messages={
            'unique': _('
        },
    )
    first_name = models.                                 True)
    last_name = models.C                                 rue)
    email = models.Email
```

THE TOKEN OBJECT


```
{
  "id": "tok_1F7Xn02eZvKYlo2C80cAActP",
  "object": "token",
  "card": {
    "id": "card_1F7Xn02eZvKYlo2CfJ7Z3z0x",
    "object": "card",
    "address_city": null,
    "address_country": null,
    "address_line1": null,
    "address_line1_check": null,
    "address_line2": null,
    "address_state": null,
    "address_zip": null,
    "address_zip_check": null,
    "brand": "Visa",
    "country": "US",
    "cvc_check": null,
    "dynamic_last4": null,
    "exp_month": 8,
    "exp_year": 2020,
    "fingerprint": "Xt5EWLLDS7FJjR1c",
    "funding": "credit",
    "last4": "4242",
    "metadata": {},
    "name": null,
    "tokenization_method": null
  },
```


@_rebecasarai

@_rebecasarai

Anonimização

Dados anônimos **não contêm informações** que possam **potencialmente identificar** um indivíduo e **não são considerados dados pessoais** pela GDPR

Abordagens

- Anonimização Estática
- Anonimização Dinâmica
- Dados sintéticos

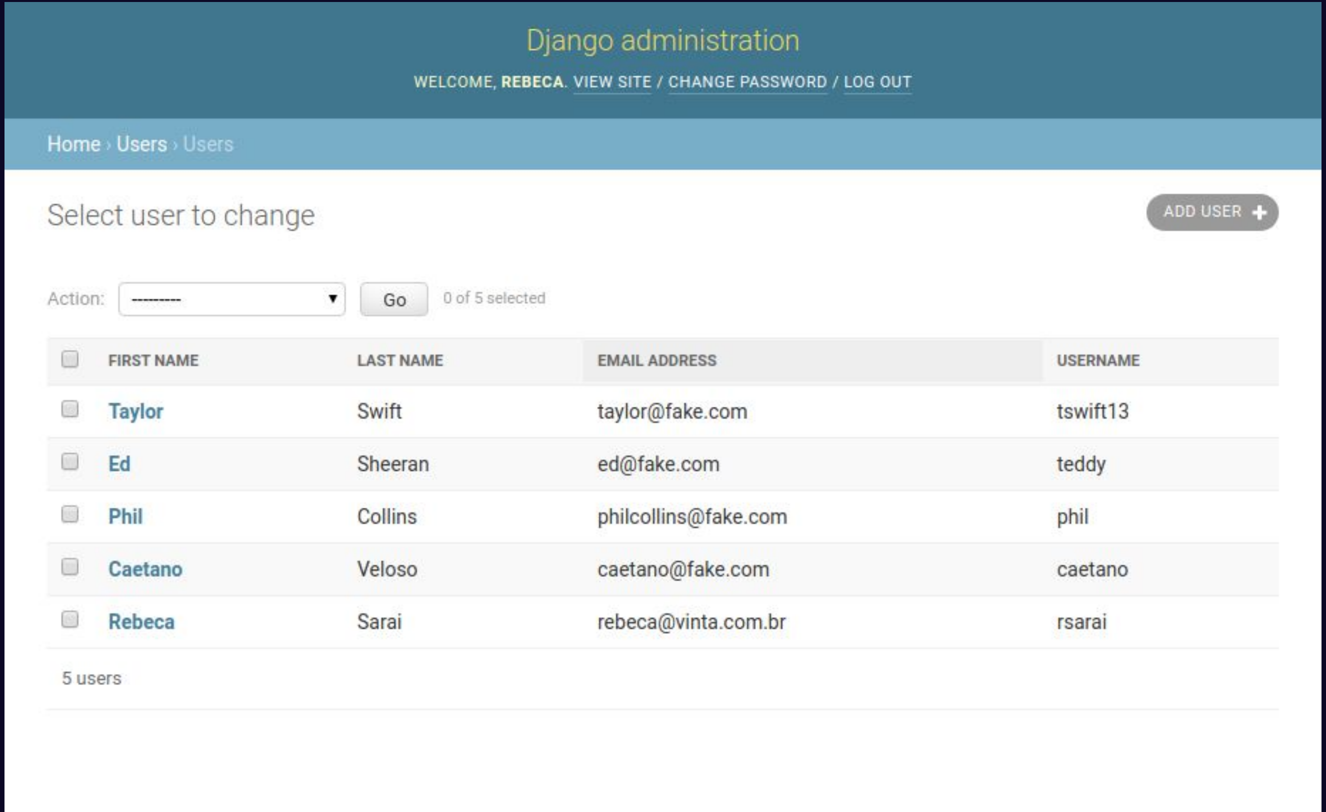
Abordagens

- Anonimização Estática
- Anonimização Dinâmica
- ~~Dados sintéticos~~

Anonimização de uma base de dados

- Anonimização **estática**
- Alterar dados **destrutivamente diretamente** no banco de dados
- Compartilhar com **third parties**
- Criar um ambiente de **teste seguro**

Anonimização de uma base de dados



@_rebecasarai

Anonimização de uma base de dados

```python
class AbstractUser(AbstractBaseUser, PermissionsMixin):
    username_validator = UnicodeUsernameValidator()
    username = models.CharField(
        _('username'),
        max_length=150,
        unique=True,
        help_text=_('Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.'),
        validators=[username_validator],
        error_messages={
            'unique': _("A user with that username already exists."),
        },
    )
    first_name = models.CharField(_('first name'), max_length=30, blank=True)
    last_name = models.CharField(_('last name'), max_length=150, blank=True)
    email = models.EmailField(_('email address'), blank=True)
    is_staff = models.BooleanField(
        _('staff status'),
        default=False,
        help_text=_('Designates whether the user can log into this admin site.'),
    )
    is_active = models.BooleanField(
        _('active'),
        default=True,
        help_text=_(
            'Designates whether this user should be treated as active. '
            'Unselect this instead of deleting accounts.'
        ),
    )
    date_joined = models.DateTimeField(_('date joined'), default=timezone.now)
```

Anonimização de uma base de dados

```python
from dj_anonymizer import anonym_field
from dj_anonymizer.register_models import AnonymBase, register_anonym, register_skip
from faker import Factory
fake = Factory.create()


class UserAnonym(AnonymBase):
    email = anonym_field.string('{seq}@fake.com', seq_callback=datetime.datetime.now)
    username = anonym_field.string('username_{seq}@fake.com', seq_callback=datetime.datetime.now)
    first_name = anonym_field.function(fake.first_name)
    last_name = anonym_field.function((fake.last_name))
    password = anonym_field.password('password')
    is_staff = anonym_field.function(lambda: False)
    ssn = anonym_field.function(fake.ssn)

    class Meta:
        queryset = User.objects.exclude(id=1)
        exclude_fields = ['is_active', 'is_superuser', 'last_login', 'date_joined',
                          'avatar', 'phone_number', 'birth_date', 'bio']
register_anonym([
    (User, UserAnonym),
])
register_skip([
    ContentType, Group, Permission, LogEntry, Session,
])
```

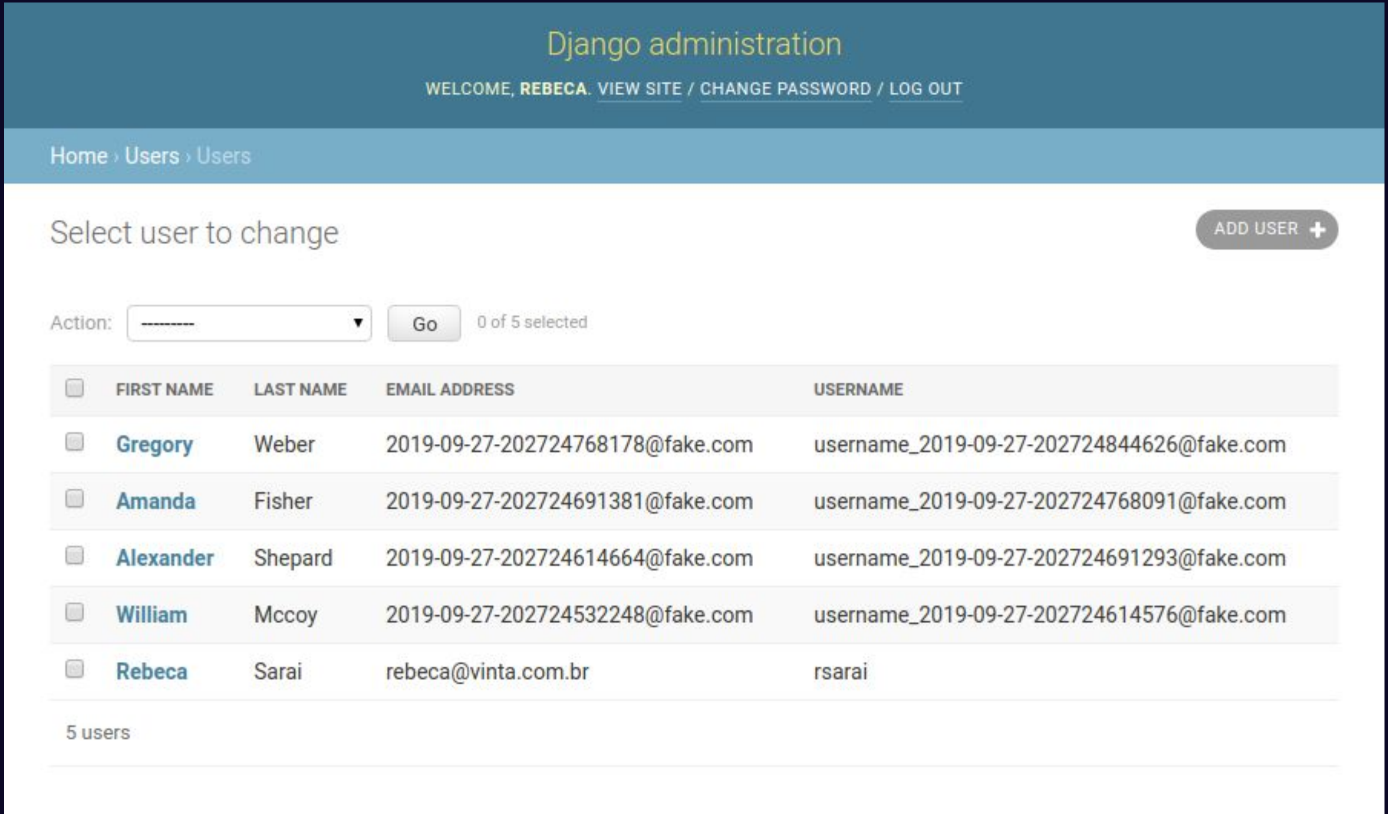
Anonimização de uma base de dados

```
✔ 21:02:57 rsarai:~/github-projects/anonymization/data-privacy/jane_doe_project (master) [2 days ago] ⤾
$ ls
anonymizer/  anonymizingdb  db.sqlite3  jane_doe_project/  janedoe  janedoeproject  manage.py*  users/
✔ 21:02:59 rsarai:~/github-projects/anonymization/data-privacy/jane_doe_project (master) [2 days ago] ⤾
$ python manage.py anonymize_db --action anonymize
Updating started

Generating fake values for model "User"


Updating finished
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
Total time (sec.): 0.3125114440917969
✔ 21:03:12 rsarai:~/github-projects/anonymization/data-privacy/jane_doe_project (master) [2 days ago] ⤾
$ exit
exit
```

Anonimização de uma base de dados



Django administration

WELCOME, REBECA. VIEW SITE / CHANGE PASSWORD / LOG OUT

Home › Users › Users

Select user to change

ADD USER +

Action: --------- Go 0 of 5 selected

FIRST NAME	LAST NAME	EMAIL ADDRESS	USERNAME
Gregory	Weber	2019-09-27-202724768178@fake.com	username_2019-09-27-202724844626@fake.com
Amanda	Fisher	2019-09-27-202724691381@fake.com	username_2019-09-27-202724768091@fake.com
Alexander	Shepard	2019-09-27-202724614664@fake.com	username_2019-09-27-202724691293@fake.com
William	Mccoy	2019-09-27-202724532248@fake.com	username_2019-09-27-202724614576@fake.com
Rebeca	Sarai	rebeca@vinta.com.br	rsarai

5 users

Anonimização de uma base de dados

- dj_anonymizer
- Mantém a **estrutura dos dados**
- **Performance**
- Anonimização deve ser **definida precisamente**
- Sujeito a ataques de **background knowledge**
- Os dados geralmente são apenas pseudonimizados

k-Anonymity

- 1° método proposto para anonimizar microdata
- Tem como objetivo criar **grupos** com pelo menos k registros compartilhando os mesmos valores de **quase-identificadores**
- Generalização e Supressão

k-Anonymity: Comportamento

Single	20

	ID	QIDs			SA
Tuple#	Name	Marital Stat	Age	ZIP Code	Crime
1	Joe	Separated	29	32042	Murder
2	Jill	Single	20	32021	Theft
3	Sue	Widowed	24	32024	Traffic
4	Abe	Separated	28	32046	Assault
5	Bob	Widowed	25	32045	Piracy
6	Amy	Single	23	32027	Indecency

k-Anonymity: Comportamento

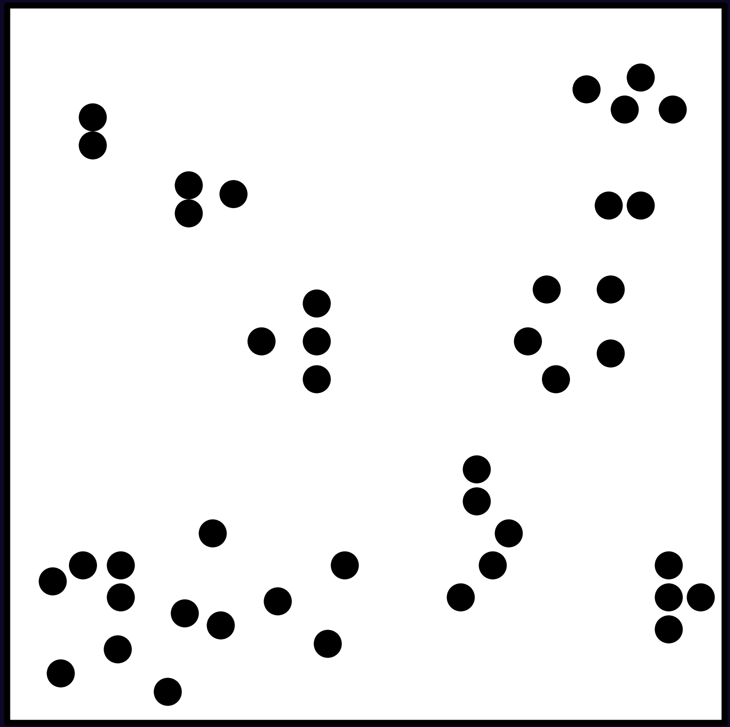
Single	20

Jill roubou alguém

	ID	QIDs			SA
Tuple#	Name	Marital Stat	Age	ZIP Code	Crime
1	Joe	Separated	29	32042	Murder
2	Jill	Single	20	32021	Theft
3	Sue	Widowed	24	32024	Traffic
4	Abe	Separated	28	32046	Assault
5	Bob	Widowed	25	32045	Piracy
6	Amy	Single	23	32027	Indecency

k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões



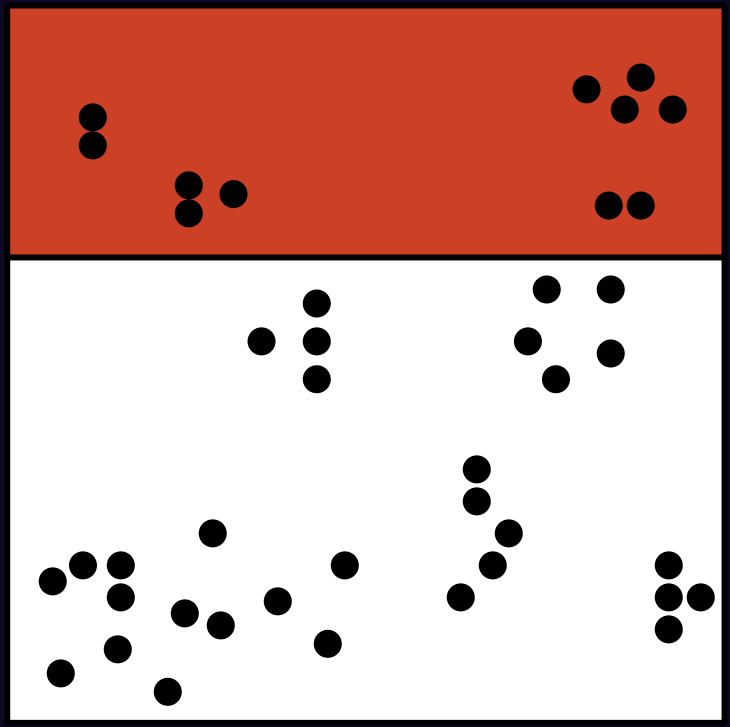
k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões

	ID	QIDs			SA
Tuple#	Name	Marital Stat	Age	ZIP Code	Crime
1	Joe	Separated	29	32042	Murder
2	Jill	Single	20	32021	Theft
3	Sue	Widowed	24	32024	Traffic
4	Abe	Separated	28	32046	Assault
5	Bob	Widowed	25	32045	Piracy
6	Amy	Single	23	32027	Indecency

k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões



@_rebecasarai

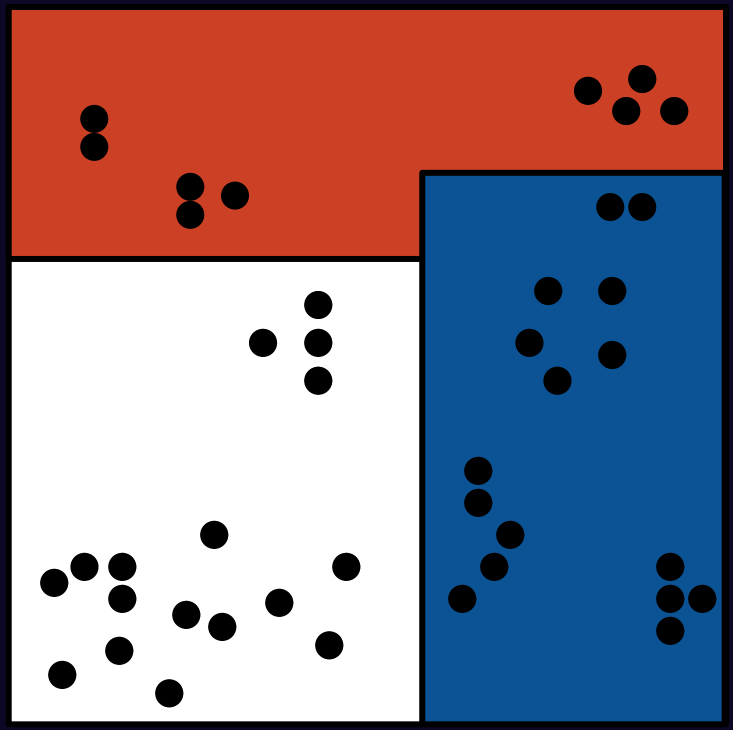
k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões

	ID	QIDs			SA
Tuple#	Name	Marital Stat	Age	ZIP Code	Crime
1	Joe	Separated	29	32042	Murder
2	Jill	Single	20	32021	Theft
3	Sue	Widowed	24	32024	Traffic
4	Abe	Separated	28	32046	Assault
5	Bob	Widowed	25	32045	Piracy
6	Amy	Single	23	32027	Indecency

k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões



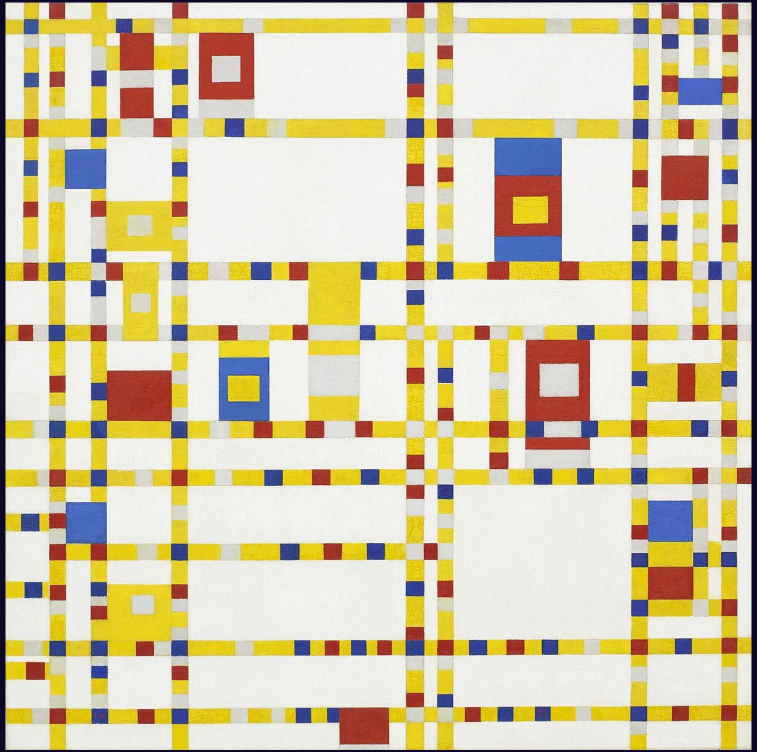
k-Anonymity: Comportamento

- Usando o algoritmo de Mondrian
- **Particiona** o espaço do domínio em várias regiões

		QIDs		Non-SA	SA
Tuple#	EQ	Marital Stat	Age	ZIP Code	Crime
1	1	Single,Separated	(23-30)	32042	Murder
4		Single,Separated	(23-30)	32046	Assault
2	2	Single,Separated	[20-23]	32021	Theft
6		Single,Separated	[20-23]	32027	Indecency
3	3	Div.,Wid.,Married,Remarried	[20-30)	32024	Traffic
5		Div.,Wid.,Married,Remarried	[20-30)	32045	Piracy

k-Anonymity

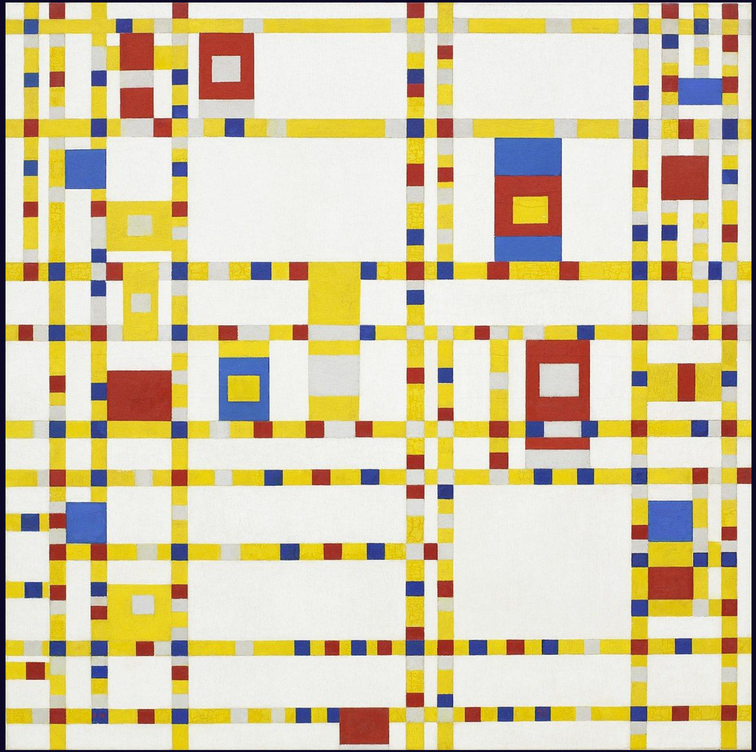
- Oferece proteção contra **divulgação de identidade**
- Não impede a divulgação dos **atributos**
- Se **múltiplas versões** dos dados são divulgadas, coordenação é necessária
- **Background knowledge**



Broadway Boogie-Woogie

Refinamentos do k-Anonymity

- l-diversity
- t-closeness
- ß-likeness
- Exige **variabilidade** nos atributos sensíveis



Broadway Boogie-Woogie

Casos de uso do k-Anonymity

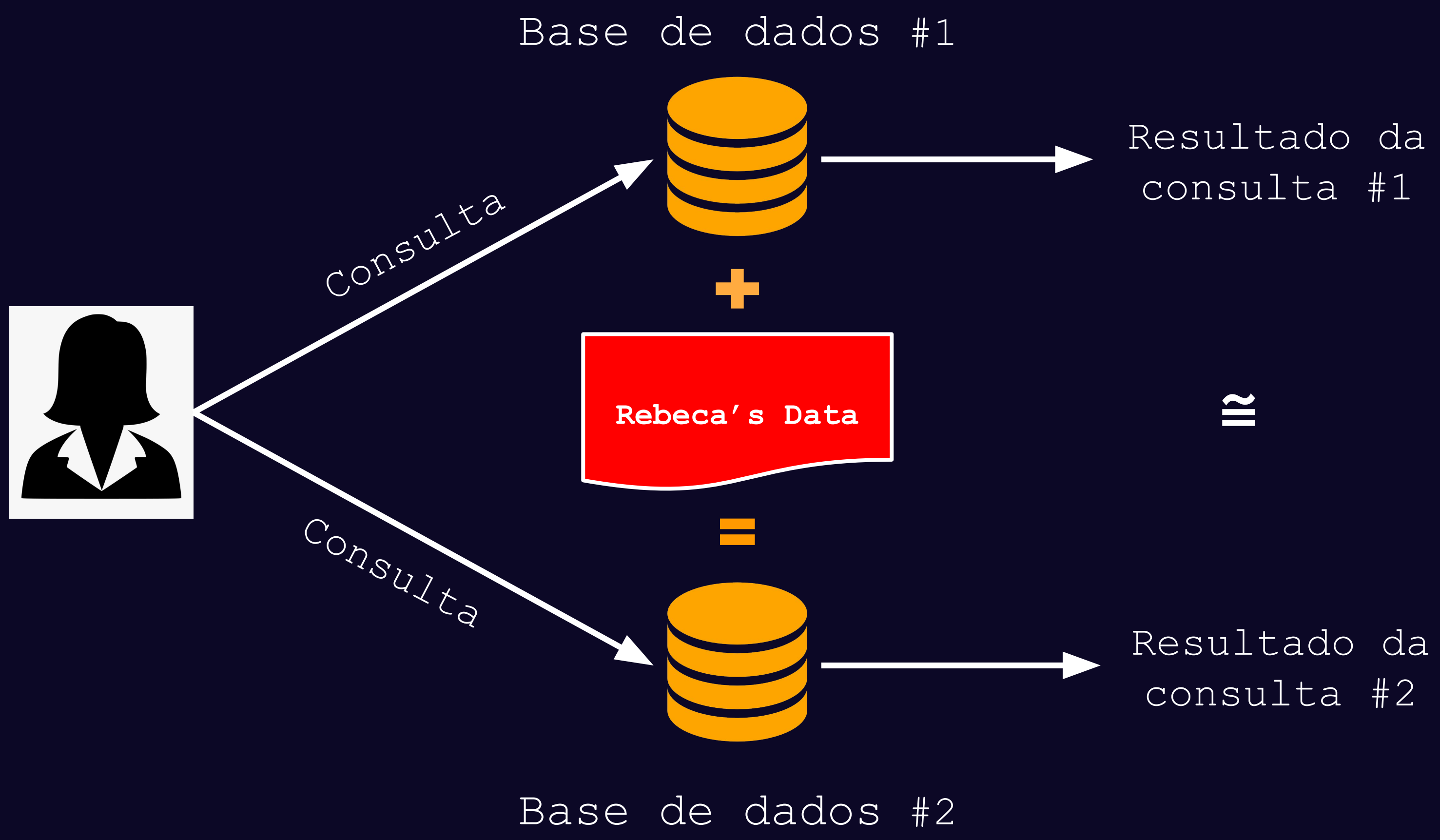
- [Haveibeenpwned]
  - [Validating Leaked Passwords with k-Anonymity]
  - [Cloudflare, Privacy and k-Anonymity]
- [Okta's PassProtect]

Casos de uso do k-Anonymity

- Haveibeenpwned
  - Validating Leaked Passwords with k-Anonymity
  - Cloudflare, Privacy and k-Anonymity
- Okta's PassProtect

Differential Privacy

- Análise de dados que preservam privacidade
- **Não é um algoritmo**
- É uma **definição formal** de privacidade
- Extrair os conhecimentos/informações da base de dados
- Sem extrair informações sobre os indivíduos na base de dados



@_rebecasarai

Differential Privacy

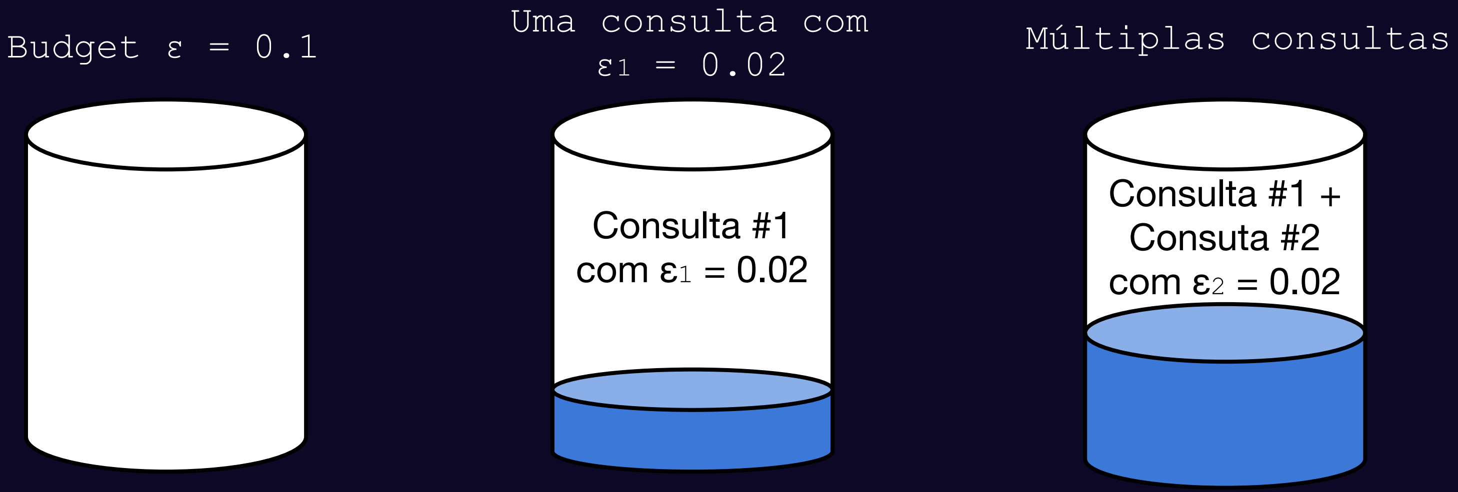
- **Negação plausível** da presença do indivíduo em uma base de dados

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

$\mathcal{M}$ é um mecanismo aleatório que fornece ε-differential privacy para todas as base de dados

Differential Privacy

- Medida de **perda de privacidade** Ɛ (**budget de privacidade**)
- Ajusta a "quantidade de privacidade"



Desafios

- Usabilidade para **não especialistas**
- **Suporte** para consultas SQL

Mecanismos

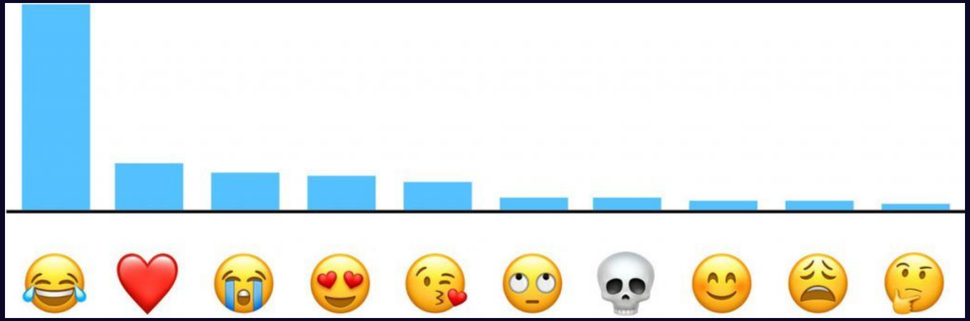
Mechanism	Strengths
Laplace Mechanism	Simple counts
PINQ	Counting and histogram queries
Elastic Sensitivity	Queries with joins
Sample & Aggregate	Statistical Estimators
Restricted Sensitivity	Graph analysis

Desafios

- Usabilidade para **não especialistas**
- **Suporte** para consultas SQL
- **Integração** com diferentes tipos de base de dados
- Lei fundamental da recuperação de informação
- Poucos exemplos do **mundo real**

Benefícios

- Proteção contra **riscos arbitrários**
- **Quantificação** da perda de privacidade
- Aplicações promissores na área de machine learning
- Usado por Microsoft, Google, Apple, Uber, etc



The Mac Observer

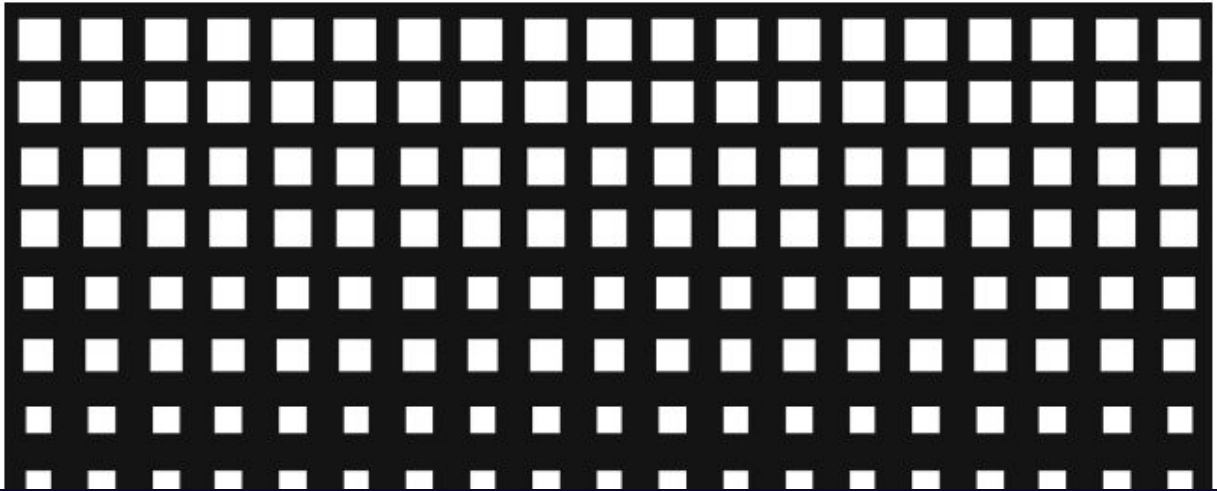
WIRED

BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION

ANDY GREENBERG SECURITY 09.15.2017 09:28 AM

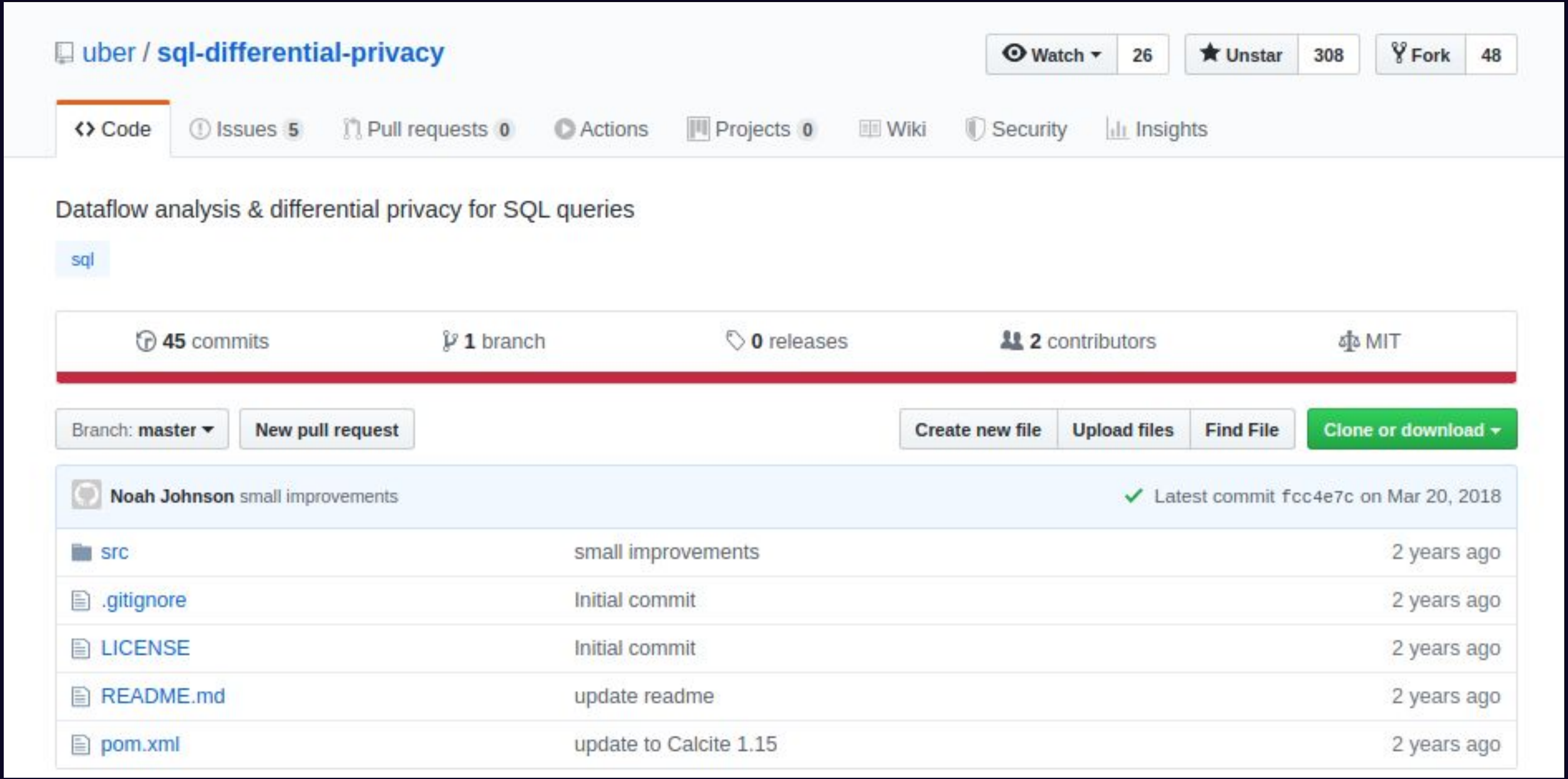
How One of Apple's Key Privacy Safeguards Falls Short

Apple has boasted of its use of a cutting-edge data science known as "differential privacy." Researchers say they're doing it wrong.

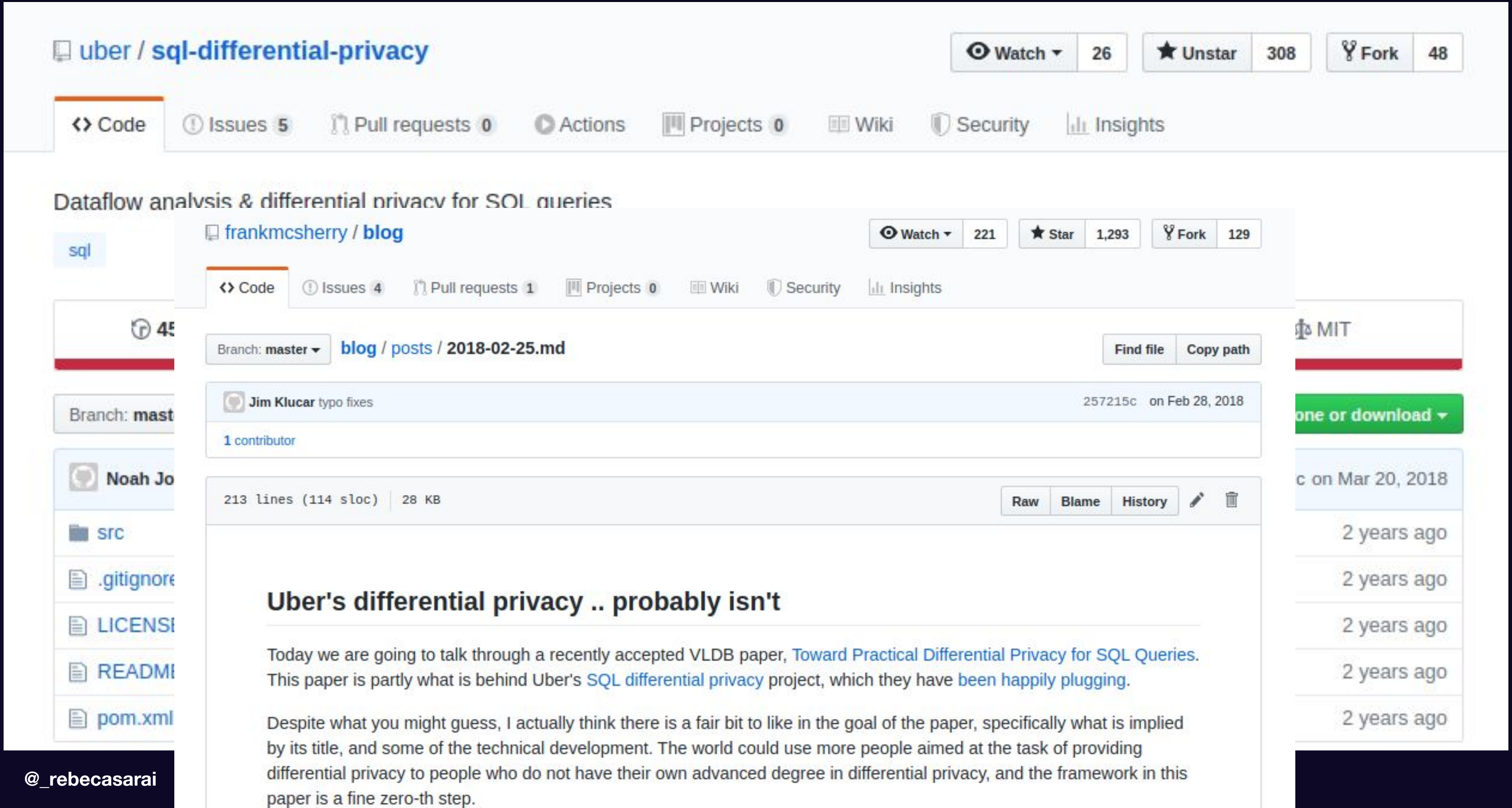


[WIRED]

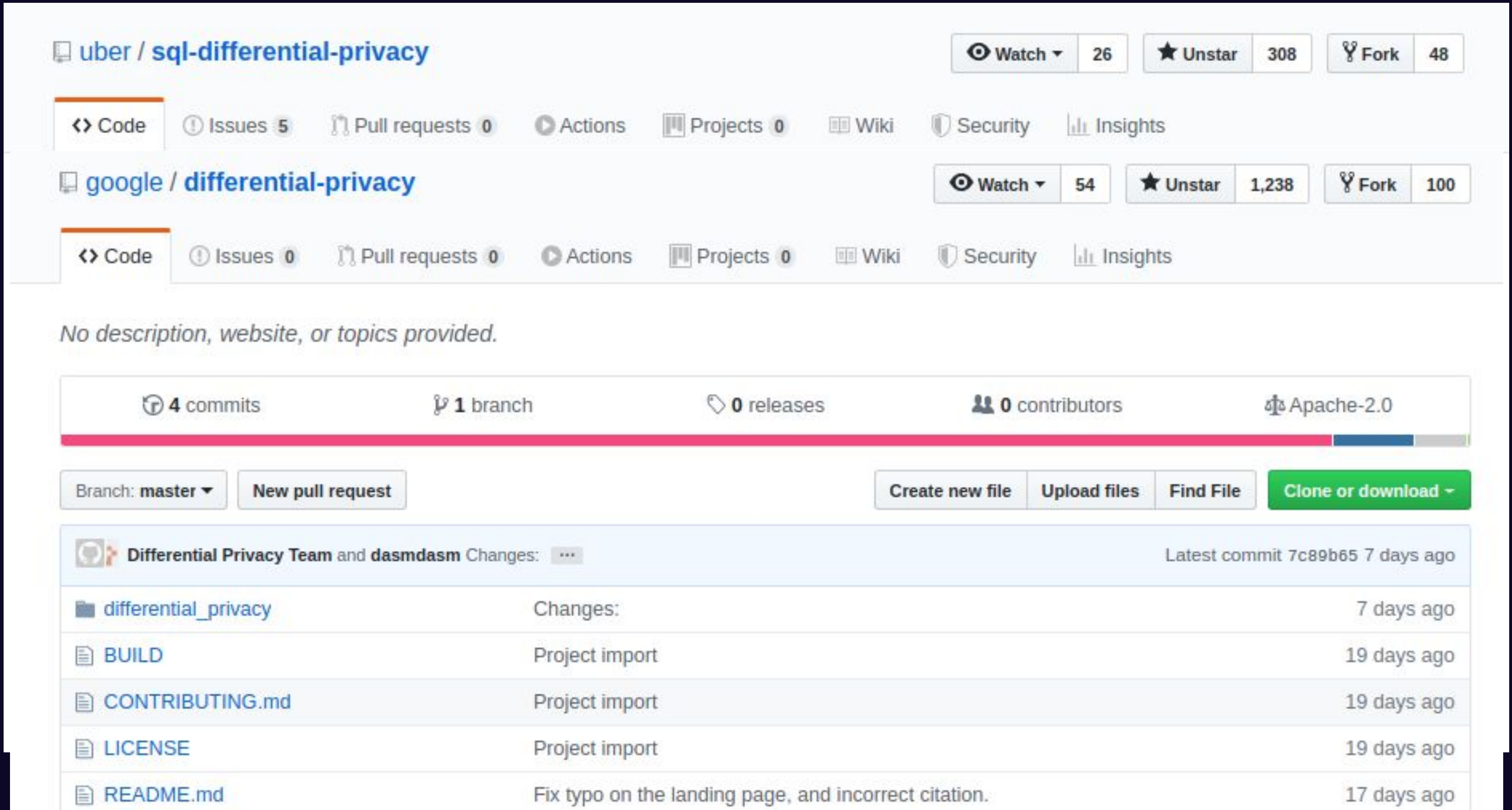
@_rebecasarai

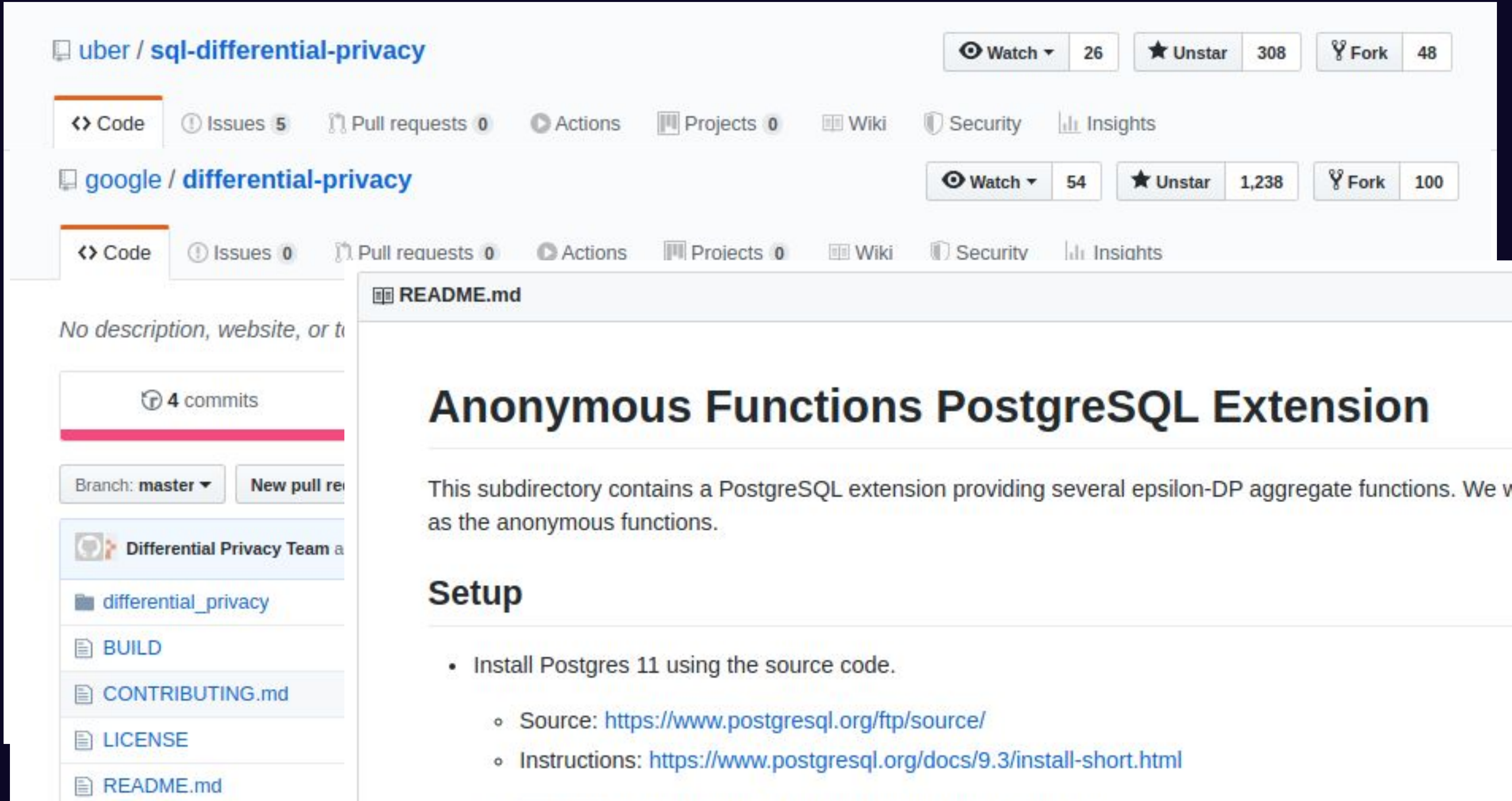


@_rebecasarai



@_rebecasarai


uber / sql-differential-privacy
Watch 26
Unstar 308
Fork 48
Code
Issues 5
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
google / differential-privacy
Watch 54
Unstar 1,238
Fork 100
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
No description, website, or topics provided.
4 commits
1 branch
0 releases
0 contributors
Apache-2.0
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download
Differential Privacy Team and dasmdasm Changes:
Latest commit 7c89b65 7 days ago
differential_privacy
Changes:
7 days ago
BUILD
Project import
19 days ago
CONTRIBUTING.md
Project import
19 days ago
LICENSE
Project import
19 days ago
README.md
Fix typo on the landing page, and incorrect citation.
17 days ago


uber / sql-differential-privacy
Watch 26
Unstar 308
Fork 48
Code
Issues 5
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
google / differential-privacy
Watch 54
Unstar 1,238
Fork 100
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
No description, website, or t
4 commits
Branch: master
New pull re
Differential Privacy Team a
differential_privacy
BUILD
CONTRIBUTING.md
LICENSE
README.md
README.md
Anonymous Functions PostgreSQL Extension
This subdirectory contains a PostgreSQL extension providing several epsilon-DP aggregate functions. We w
as the anonymous functions.
Setup
Install Postgres 11 using the source code.
Source: https://www.postgresql.org/ftp/source/
Instructions: https://www.postgresql.org/docs/9.3/install-short.html

sciendo
Proceedings on Privacy Enhancing Technologies ..; .. (..):1–20

Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson*

Differentially Private SQL with Bounded User Contribution

Abstract: Differential privacy (DP) provides formal guarantees that the output of a database query does not reveal too much information about any individual present in the database. While many differentially private algorithms have been proposed in the scientific literature, there are only a few end-to-end implementations of differentially private query engines. Crucially, existing systems assume that each individual is associated with at most one database record, which is unrealistic in practice. We propose a generic and scalable method to perform differentially private aggregations on databases, even when individuals can each be associated with arbitrarily many rows. We express this method as an operator in relational algebra, and implement it in an SQL engine. To validate this system, we test the utility of typical queries on industry benchmarks, and verify its correctness with a stochastic test framework we developed. We highlight the promises and pitfalls learned when deploying such a system in practice, and we publish its core components as open-source software.

Keywords: differential privacy, database, SQL

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

a population without revealing too much about individuals is a long-standing field of research. The standard definition used in this context is differential privacy (DP): it provides a formal guarantee on how much the output of an algorithm reveals about any individual in its input [10, 11, 14]. Differential privacy states that the distribution of results derived from private data cannot reveal "too much" about a single person's contribution, or lack thereof, to that data [12]. By using differential privacy when analyzing data, organizations can minimize the disclosure risk of sensitive information about their users.

Query engines are a major analysis tool for data scientists, and one of the most common ways for analysts to write queries is with Structured Query Language (SQL). As a result, multiple query engines have been developed to enable data analysis while enforcing DP [2, 21, 26, 33], and all of them use a SQL-like syntax.

However, as we discuss in Section 2, these differentially private query engines make some implicit assumptions, notably that each individual in the underlying database is associated with at most one database record. This does not hold in many real-world datasets, so the privacy guarantee offered by these systems is weaker than advertised for those databases. To overcome this uber / sql-differe

<> Code | Issues

google / differen

<> Code | Issues

No description, websi

4 commits

Branch: master | Ne

Differential Privacy

differential_privacy

BUILD

CONTRIBUTING.m

LICENSE

README.md

sciendo

Proceedings on Privacy Enhancing Technologies ..; .. (..):1–20

Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson*

Differentially Private SQL with Bounded User Contribution

Abstract: Differential pri… guarantees that the outp… not reveal too much infor… present in the database. V… vate algorithms have been… erature, there are only a… tions of differentially priv… existing systems assume t… ated with at most one da… realistic in practice. We p… method to perform differen… databases, even when indiv… with arbitrarily many row… an operator in relational al… SQL engine. To validate th… of typical queries on industry benchmarks, and verify its correctness with a stochastic test framework we developed. We highlight the promises and pitfalls learned when deploying such a system in practice, and we publish its core components as open-source software.

Keywords: differential privacy, database, SQL

DOI Editor to enter DOI

Received ..; revised ..; accepted ...

…too much about indi-… of research. The stan-… t is differential privacy… ntee on how much the… bout any individual in… privacy states that the… m private data cannot… person's contribution,… . By using differential… ganizations can mini-… ive information about…

…analysis tool for data… common ways for an-… tructured Query Language (SQL). As a result, multiple query engines have been developed to enable data analysis while enforcing DP [2, 21, 26, 33], and all of them use a SQL-like syntax.

However, as we discuss in Section 2, these differentially private query engines make some implicit assumptions, notably that each individual in the underlying database is associated with at most one database record. This does not hold in many real-world datasets, so the privacy guarantee offered by these systems is weaker than advertised for those databases. To overcome this

308 | Fork | 48

1,238 | Fork | 100

Extension

P aggregate functions. We w

Preocupações ao anonimizar datasets

- Dados não ser **completamente anonimizados** e se manter **úteis**
- **Re-identificação** não é o único risco
- Consultas em **grandes datasets** não garantem privacidade
- **Auditoria** de consultas nem sempre é viável

Concerns when anonymizing datasets

- Divulgar **somente estatísticas** não é seguro
- Divulgar **fatos ordinários** pode ser problemático
- **Segurança** não quer dizer privacidade

Obrigada!

Perguntas?

Give me feedback on **@_rebecasarai**

Rebeca Sarai

Software Developer

rebeca@vinta.com.br

@_rebecasarai

/rsarai

References

- CAO, Jianneng; KARRAS, Panagiotis. **Publishing microdata with a robust privacy guarantee**. Proceedings of the VLDB Endowment, v. 5, n. 11, p. 1388-1399, 2012.
- The Algorithmic Foundations of Differential Privacy (here)
- Differentially Private SQL with Bounded User Contribution (here)
- Differential Privacy at Scale: Uber and Berkeley Collaboration (video)
- Tutorial: Differential Privacy and Learning: The Tools, The Results, and The Frontier (video)
- Keeping Your Data Secure While Learning From It - Andreas Dewes and Katharine Jarmul (video)
- 9 Data Anonymization Use Cases You Need To Know Of (here)
- The Definition of Differential Privacy - Cynthia Dwork (video)
- Protecting Personal Data with Django (because it's the law) (video)
- Pseu, Pseu, Pseudio. Pseudonymization in Django. by Frank Valcarcel (video)
- DOMINGO-FERRER, Josep; SORIA-COMAS, Jordi. **Anonymization in the time of big data**. In: International Conference on Privacy in Statistical Databases. Springer, Cham, 2016. p. 57-68.
- LI, Ninghui; LI, Tiancheng; VENKATASUBRAMANIAN, Suresh. **t-closeness: Privacy beyond k-anonymity and l-diversity**. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007. p. 106-115.
- SWEENEY, Latanya. **k-anonymity: A model for protecting privacy**. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, v. 10, n. 05, p. 557-570, 2002.
- Apple Releases Details on Differential Privacy, and the Big Takeaway Is Which Emoji Is Most Popular (here)
- Differential Privacy In Action (here)

Other differential privacy projects

- https://github.com/google/rappor
- https://github.com/prashmohan/GUPT
- https://github.com/LLGemini/PINQ
- https://github.com/ektelo/ektelo