Imperial College
London

COMPUTATIONAL
PRIVACY
GROUP

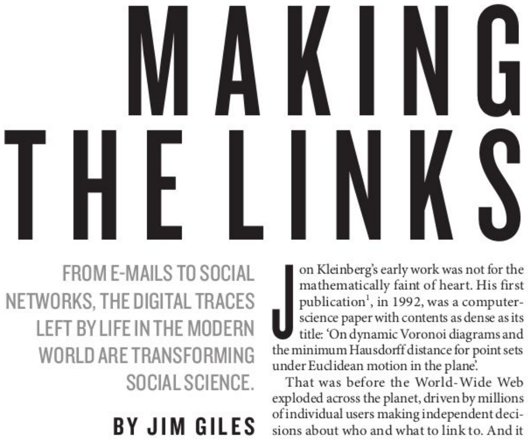
The search for anonymous data: from de-identification to systems

Yves-Alexandre de Montjoye

Imperial College London

COMPUTATIONAL PRIVACY GROUP

Disclaimer: The opinions expressed in this presentation are strict mine and not the one of the institutions I work for.



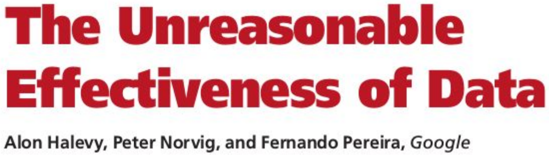
MAKING THE LINKS

FROM E-MAILS TO SOCIAL NETWORKS, THE DIGITAL TRACES LEFT BY LIFE IN THE MODERN WORLD ARE TRANSFORMING SOCIAL SCIENCE.

BY JIM GILES

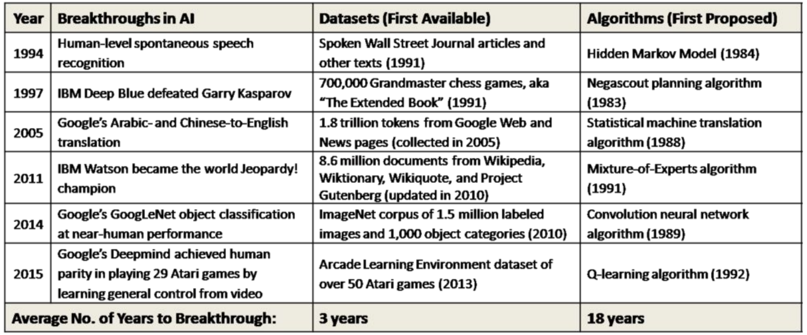
Jon Kleinberg's early work was not for the mathematically faint of heart. His first publication[1], in 1992, was a computer-science paper with contents as dense as its title: 'On dynamic Voronoi diagrams and the minimum Hausdorff distance for point sets under Euclidean motion in the plane'.

That was before the World-Wide Web exploded across the planet, driven by millions of individual users making independent decisions about who and what to link to. And it

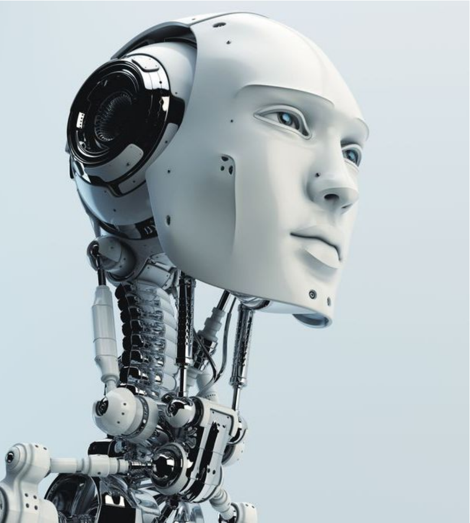


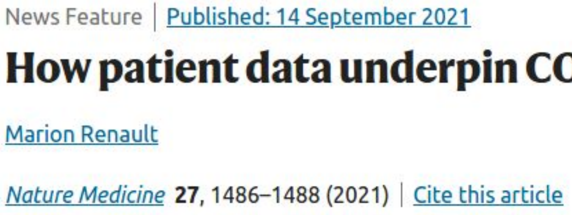
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*



Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		3 years	18 years



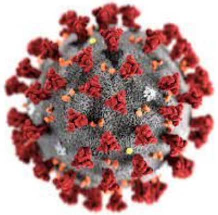


News Feature | Published: 14 September 2021

How patient data underpin COVID-19 research

Marion Renault

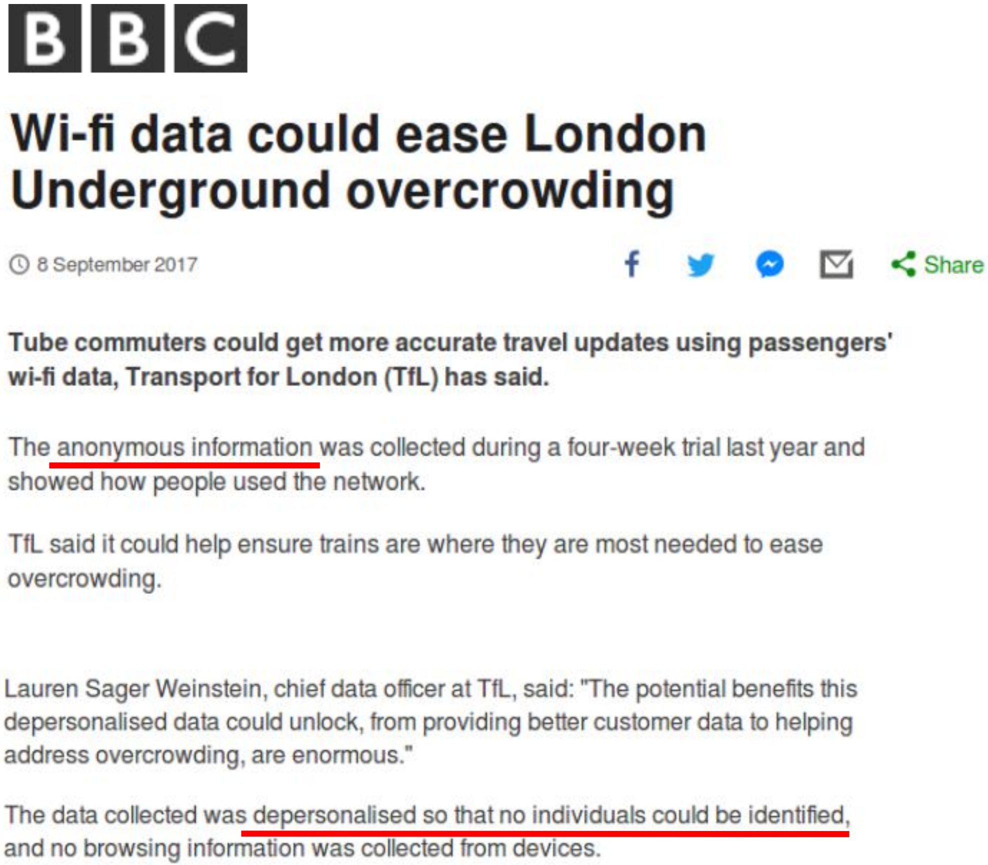
Nature Medicine **27**, 1486–1488 (2021) | Cite this article



```
H6ycJQIv.csv:
call,in,sW4aFX,2014-03-02 07:13:30,210,42.366944,-71.083611
call,out,5f0jX5G,2014-03-02 07:53:30,34,42.366944,-71.083611
text,in,5f0jX5G,2014-03-02 08:22:30,,42.386722,-71.138778
```

```
AnonID, Query, QueryTime, ItemRank, ClickURL, URL
142, www.newyorklawyersite.com, 2006-03-18 08:03:09, ,
142, westchester.gov, 2006-03-20 03:55:57, 1, http://www.westchestergov.com
1326, budget truck rental, 2006-03-24 18:27:07, ,
1326, holiday mansion houseboat, 2006-03-29 17:14:01, 5, http://everyboat.com
1326, back to the future, 2006-04-01 17:59:28, ,
```

```
Urban Outfitters   7abc1a23 09/23 $97.30
Whole Food         3092fc10 09/23 $43.78
Central Bakkery    7abc1a23 09/23 $4.33
MIT RecSport       4c7af72a 09/23 $12.29
Flour Cafe         89c0829c 09/24 $3.66
Border Cafe        7abc1a23 09/24 $35.81
```



Wi-fi data could ease London Underground overcrowding

8 September 2017

Tube commuters could get more accurate travel updates using passengers' wi-fi data, Transport for London (TfL) has said.

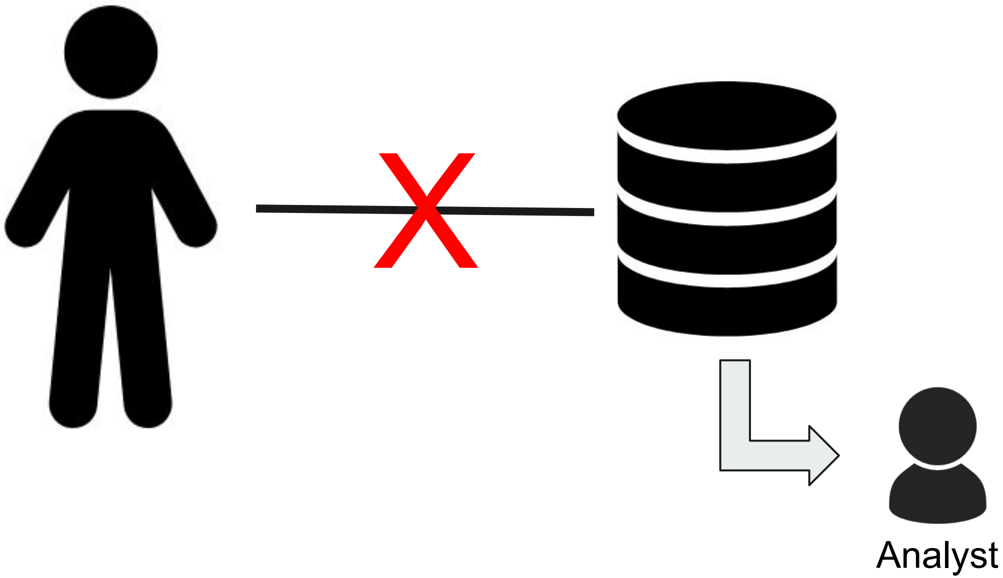
The anonymous information was collected during a four-week trial last year and showed how people used the network.

TfL said it could help ensure trains are where they are most needed to ease overcrowding.

Lauren Sager Weinstein, chief data officer at TfL, said: "The potential benefits this depersonalised data could unlock, from providing better customer data to helping address overcrowding, are enormous."

The data collected was depersonalised so that no individuals could be identified, and no browsing information was collected from devices.

Anonymity: breaking the link between a person and their data



De-identification: “40 years of litterature”

Name	DOB	G	Income [$/an]
Katerine Enter	01/1936	F	100.000
Luella Perret	04/1960	F	35.678
Dong Rice	12/1982	M	45.000
Carroll Stiner	03/1970	M	325.000
Ken Alamo	05/1969	M	125.000
Yulanda Parikh	11/1997	F	23.459
Janee Lundell	09/1995	F	75.008

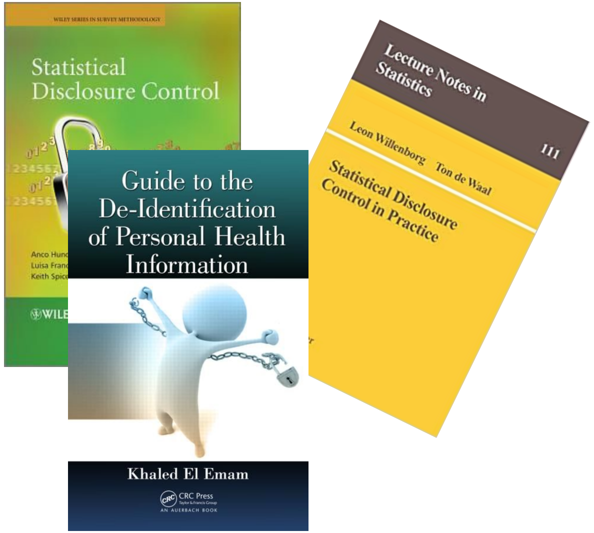
Pseudonymization

Name	DOB	G	Income [$/an]
vF0m6JGQ	01/1936	F	100.000
p0nYRG91	04/1960	F	35.678
LgRLdjaA	12/1982	M	45.000
uH4sUWLU	03/1970	M	325.000
zfyv9PRY	05/1969	M	125.000
qbu8Us1P	11/1997	F	23.459
SrQ4sonIn	09/1995	F	75.008

De-identification

- Noise addition
- Generalization
- Record swapping
- Suppression
- etc

Name	DOB	G	Income [$/an]
vF0m6JGQ	80	F	100.000
p0nYRG91	60	F	35.678
LgRLdjaA	30	M	45.000
uH4sUWLU	50	M	325.000
zfyv9PRY	50	M	125.000
qbu8Us1P	20	F	23.459
SrQ4sonIn	20	F	75.008


WILEY SERIES IN SURVEY METHODOLOGY
Statistical Disclosure Control
WILEY
Guide to the De-Identification of Personal Health Information
Khaled El Emam
CRC Press
Lecture Notes in Statistics
111
Leon Willenborg Ton de Waal
Statistical Disclosure Control in Practice

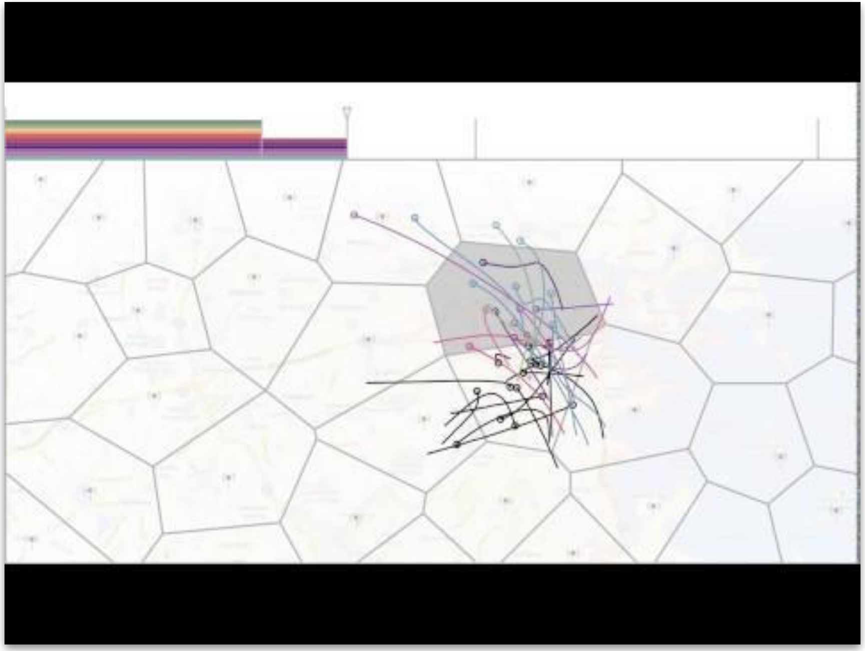
First: Pseudonymization

Pseudonymization

A person *x* who is in the dataset D

Points: place and time where we know *x* was.

Here: area of roughly 1km^2 within an hour



de Montjoye, Y.A. et al. 2013. Unique in the Crowd: The privacy bounds of human mobility, *Nature Srep, 3,* p.1376

Pseudonymization

A person *x* who is in the dataset D

Points: place and time where we know *x* was.

Here: area of roughly 1km^2 within an hour

Mobile phone dataset of 1.5M people over 15 months, **4 points (hour, location)** are enough to **uniquely identify 95% of the population**

$$\mathcal{E}_4 = .95$$

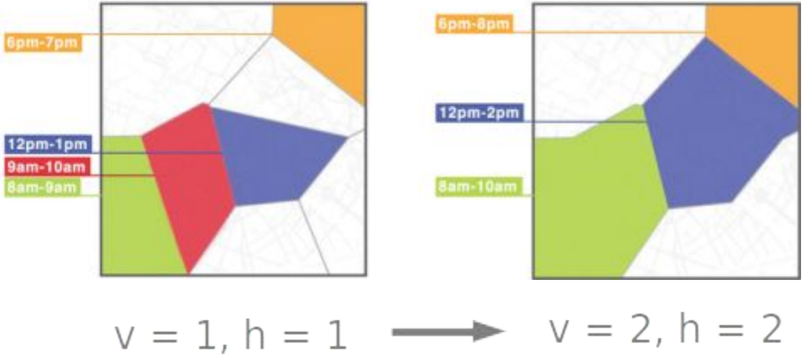
Points can be **anything** that allows me to know that the person I'm searching for what at a given place at a given time.

In a credit card dataset of 1.1M people over 3 months, **4 points (shops, day)** are enough to uniquely identify **90% of the population**

de Montjoye, Y.A. et al. 2013. Unique in the Crowd: The privacy bounds of human mobility, *Nature Srep, 3,* p.1376
de Montjoye, Y.A., et al. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221)

Second: De-identification (aka noise addition)

Noise addition (or generalization)



Using real data, we estimated that the decrease of *unicity* (risk of re-identification) follows:

$$\mathcal{E} \sim (v * h)^{-p/100}$$

with *h* the spatial resolution of the data, *v* the temporal resolution of the data, and *p* the number of points known to the attacker.

We obtain similar results on credit card data

de Montjoye, Y.A. et al. 2013. Unique in the Crowd: The privacy bounds of human mobility, *Nature Srep, 3,* p.1376
de Montjoye, Y.A., et al. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221)

Third: Scaling

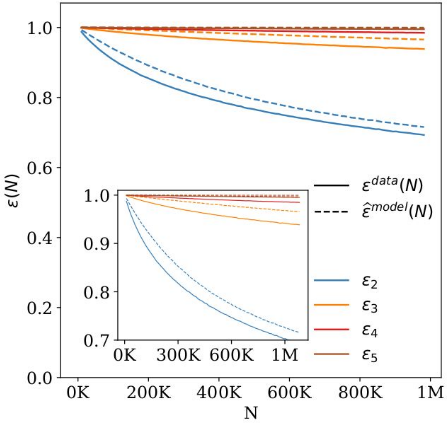
Scaling

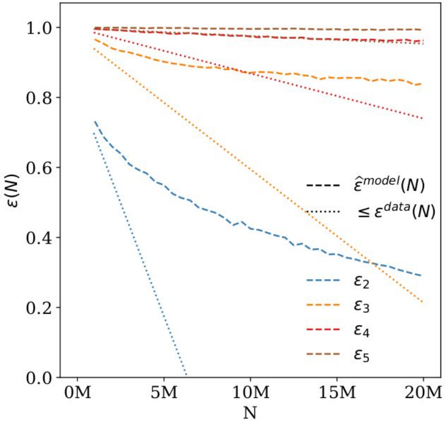
“I created a population of 22 million people with credit card transactions and a five percent sample of 1.1 million people. [...] In fact, my population needs less than one percent uniqueness to get 90 percent uniqueness in my sample.” -- Industry leader

We show through **3 different methods** that unicity is likely to be high even in large populations.

Our estimates for unicity at 20M people range from 0.99 (model) to 0.73 (tangent)

Unicity (=risk of re-identification) decreases slowly with population size





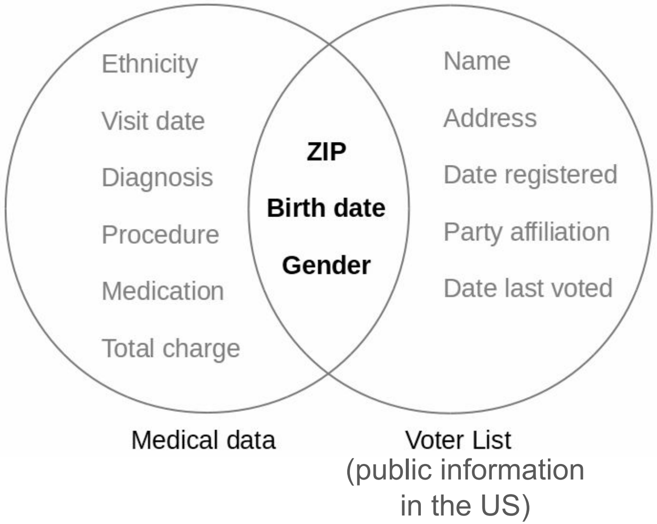
Farzanehfar, A., Houssiau, F. and de Montjoye, Y.A., 2021. The risk of re-identification remains high even in country-scale location datasets. Patterns, 2(3), p.100204.

Fourth: Uncertainty (and sampling)

The original re-identification

In 1997, William Weld, Governor of Massachusetts released properly pseudonymized medical records of state employees for research, assuming this protected privacy of individuals.

A few days later, Gov. Weld's **medical records were delivered to his office**.



Latanya Sweeney

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

Uncertainty and sampling

“The **Cambridge voter-roll** used in the attack actually included only just more than half of the true population of Cambridge at the time of Weld's collapse. This introduces a "**fatal flaw**" into the logic used to purport Weld's re-identification using voter registration data.”

Barth-Jones, D. The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections (2012).

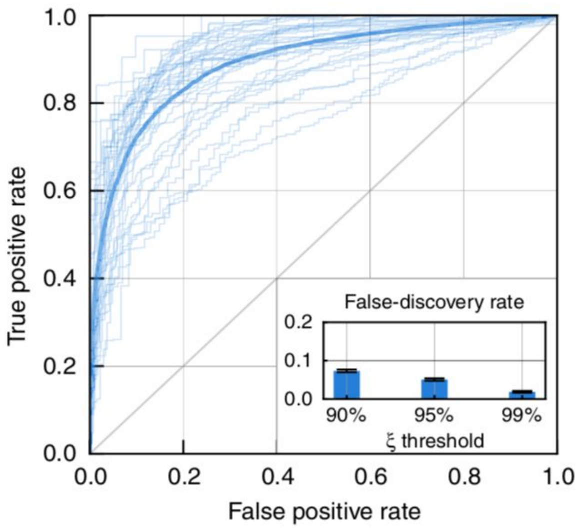
“Sampling — providing access to only a fraction of the total existing records or data, **thereby creating uncertainty** that any particular person is **even included in the dataset**”

Office of the Australian Information Commissioner in De-identification and the Privacy Act (2018)

Estimating the likelihood of a re-id to be correct

$$\xi_{\boldsymbol{x}} \equiv \mathbb{P}\Big(\underbrace{\boldsymbol{x} \text{ unique in } (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)})}_{\text{Always correctly re-identified amongst } n \text{ people}} \mid \underbrace{\exists i, \boldsymbol{x}^{(i)} = \boldsymbol{x}}_{\text{Match}}\Big)$$

$$= \big(1 - \underbrace{p(\boldsymbol{x})}_{\text{Modeled probability to draw the record } \boldsymbol{x}}\big)^{n-1}$$



We model the joint distribution with gaussian Copulas. Validated on more than **210 data collections** from census and survey data with sampling rates ranging from **0.001% to 1%**, **<AUC> = 0.87** and low FDR.

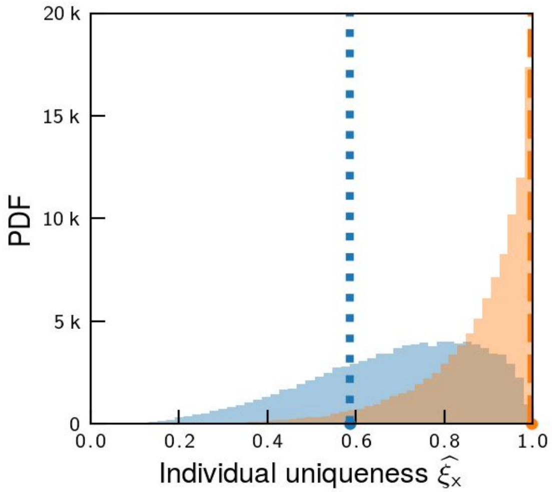
Rocher, L., Hendrickx, J.M. and de Montjoye, Y.A., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), p.3069.

Coming back to Gov. William Weld and the “fatal flaw”

Male in Cambridge (02132), born 07/31/1945:
58% chance of being unique

these three attributes + five children:
99.97% chance of being unique
(for any attempt, almost always correct)

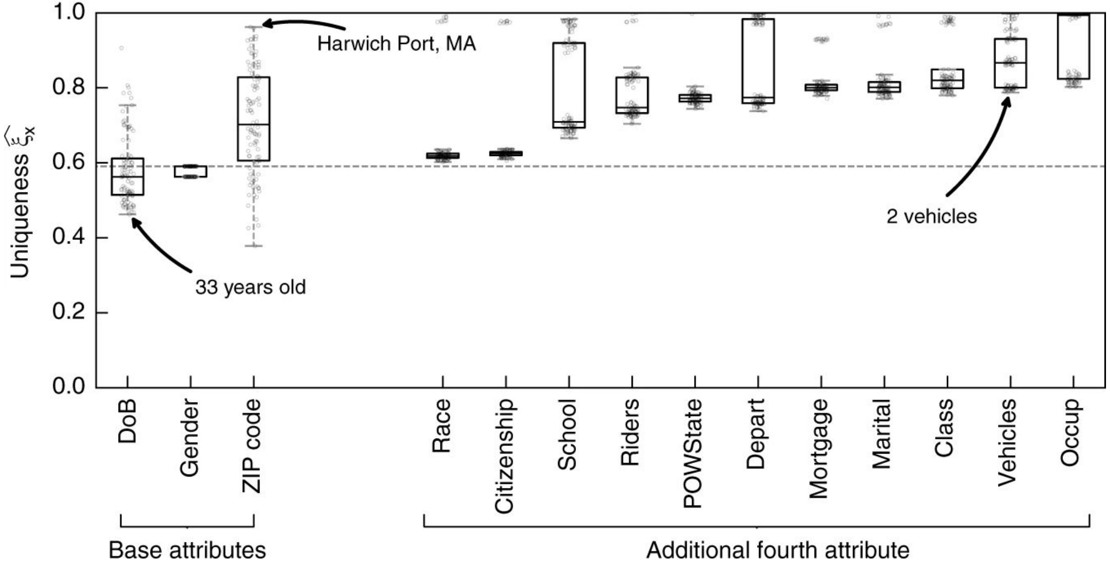
Predicted individual uniqueness for
ZIP code + DoB + gender
ZIP code + DoB + gender + nb of children



Model trained on a 5% PUMS Massachusetts census from 1990

With **15 attributes**, one can re-identify a [US] person in any dataset with **99.98% likelihood**

Using the Massachusetts census from 1990, we showed that an attacker would have had **99.97%** chance of correctly identifying Weld's medical records with ZIP code, DOB, gender, and number of children (5)



The observatory of Anonymity: 89 countries
cpg.doc.ic.ac.uk/observatory

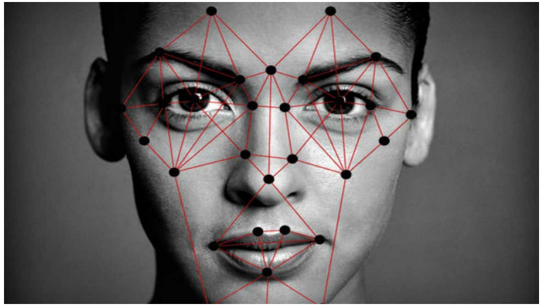
Fifth: Availability of auxiliary information

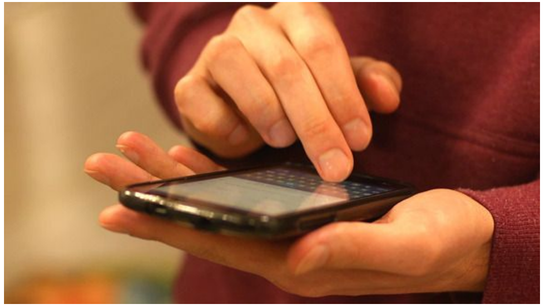
New class of attack: profiling

Demographic attributes such as gender, zip code or marital status are **fairly stable over time.** This makes them robust auxiliary information to re-identify an individual in a range of datasets.

For **time-dependent**, often large-scale data, auxiliary information that can be used to match against anonymous data **might not always be widely available*.** This has been a major argument to dismiss the risk posed by attacks in practice.

Drawing from the literature on facial recognition, **profiling attacks** are now proposed against time-dependent behavioral data.



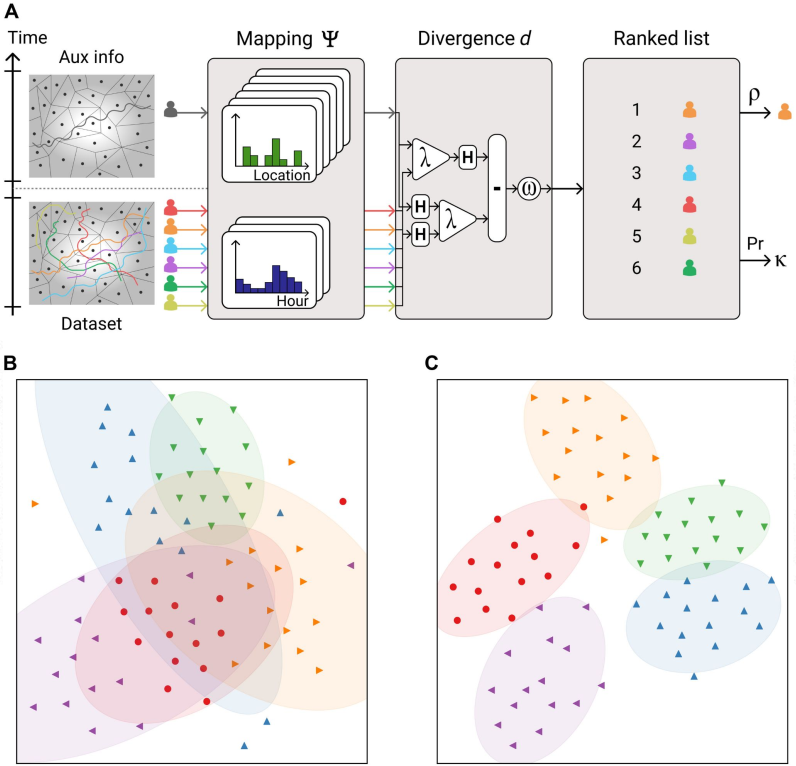


Profiling attack model for location data

We propose a profiling attack to "learn" the behavior so as to be able to **identify individual in "non-matching" data.**

Here "non-matching" can be data collected over non-overlapping time periods (what we use in our empirical setup) but also data collected at different "frequencies" (e.g. two credit cards)

Our entropy-based model learns **time-persistent profiles of individuals from their location data (space and time).** The profiles are optimized using a contrastive loss function.



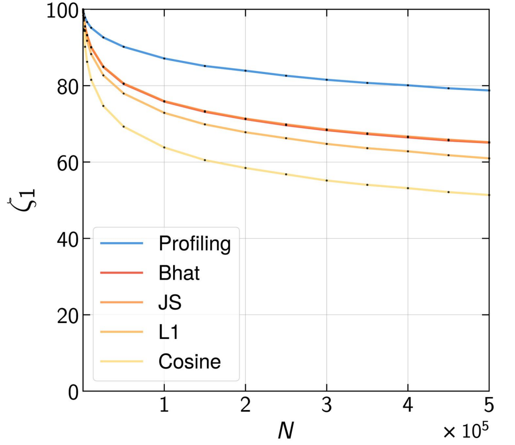
Tournier, A.J. and de Montjoye, Y.A., 2022. Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. Science Advances, 8(33), p.eabl6464.

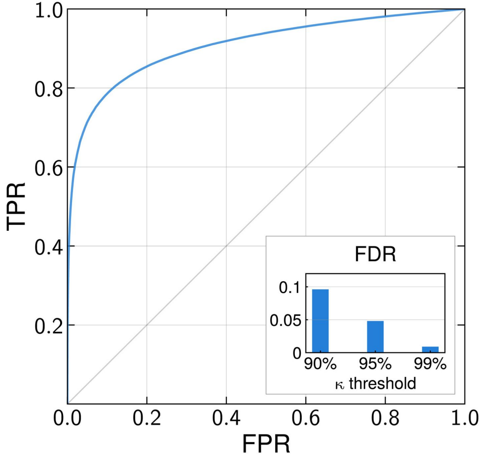
Re-identifying location data

In a location dataset of .5M people, the model correctly identifies individuals **79% of the time**, strongly outperforming all previous baselines. Rank 10 performance is 93%.

The model is **well calibrated**, allowing an attacker to evaluate the likelihood of the identification to be correct. For example, for κ > 0.95, 4.81% are false positives.

We also validate the model on a **shopping dataset** where it identifies individuals correctly 65.2% of the time (rank 10: 74%).





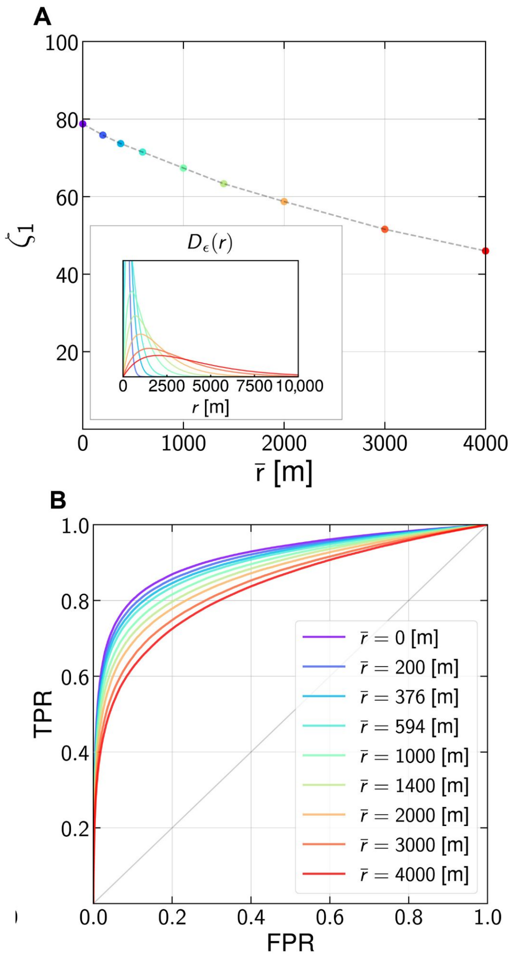
Tournier, A.J. and de Montjoye, Y.A., 2022. Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. Science Advances, 8(33), p.eabl6464.

Importantly, we also show the model to be stable over time, the membership assumption to be removed and the model to be robust to state-of-the-art noise addition

Stability of over time: the accuracy of the model only decrease by 1% per week between the dataset and the auxiliary data.

Membership assumption: using meta-classifier, we can remove the membership assumption without impacting the accuracy of the model. The AUC, e.g., only decreases from 0.91 to 0.89 for P=0.9

Robust to noise addition: we show the model to be (very!) robust to local noise addition using Geo-indistinguishability. Accuracy decreases to 78% when small amounts of noise are added (r = 100m, r_{95} = 237 m) and 71% for large amounts of noises (r = 600m, r_{95} = 1432 m)

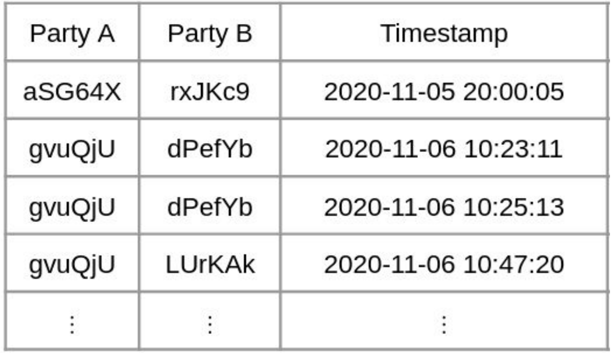


Tournier, A.J. and de Montjoye, Y.A., 2022. Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. Science Advances, 8(33), p.eabl6464.

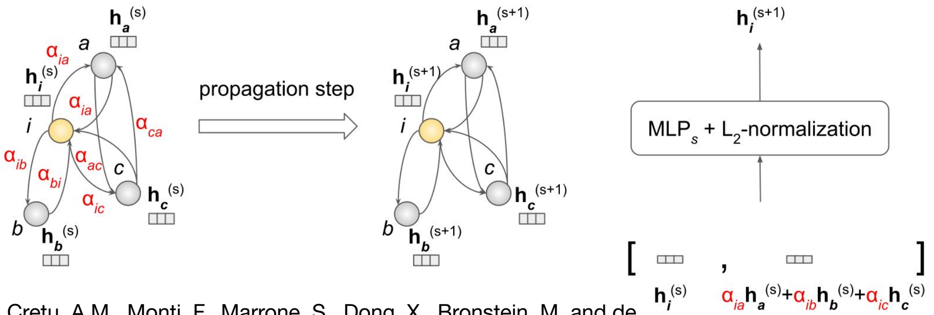
Not limited to location data: profiling using interaction data

We similarly propose a profiling attack model for **interaction data.**

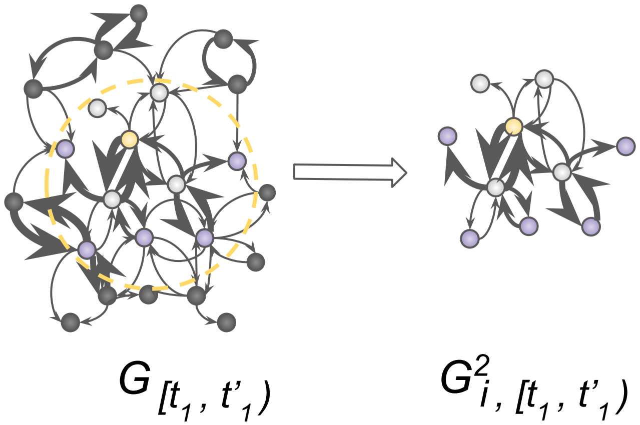
Our model computes a time-dependent profile (embedding) of an individual from their k-hop interactions using a **multi-layer graph attention network** using a set of 23 features using the bandicoot library. The embeddings are optimized using the triplet loss.



Party A	Party B	Timestamp
aSG64X	rxJKc9	2020-11-05 20:00:05
gvuQjU	dPefYb	2020-11-06 10:23:11
gvuQjU	dPefYb	2020-11-06 10:25:13
gvuQjU	LUrKAk	2020-11-06 10:47:20
⋮	⋮	⋮



Crețu, A.M., Monti, F., Marrone, S., Dong, X., Bronstein, M. and de Montjoye, Y.A., 2022. Interaction data are identifiable even across long periods of time. Nature Communications, 13(1), pp.1-11.

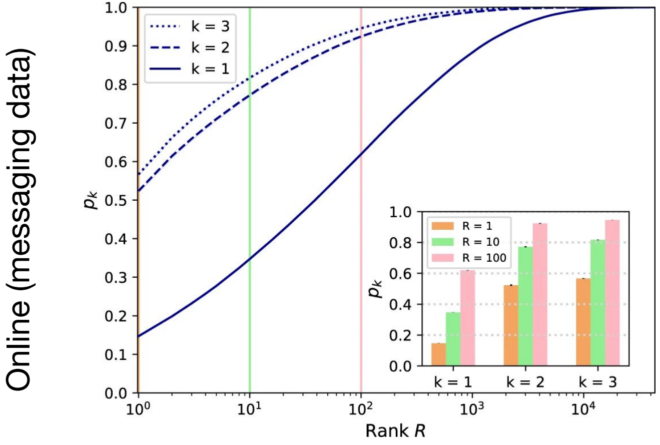


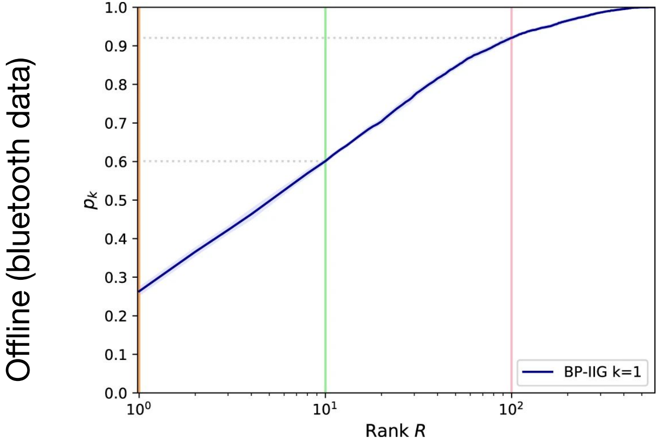
Re-identifying interaction data

In **messaging data** (e.g. WhatsApp), our approach correctly identifies individuals **52.4% of the time** out of 40k people using 2-hops graphs ($p_{k=1}$ = 14.7% and $p_{k=3}$ = 56.7%). Rank 10 and 100 results are even higher.

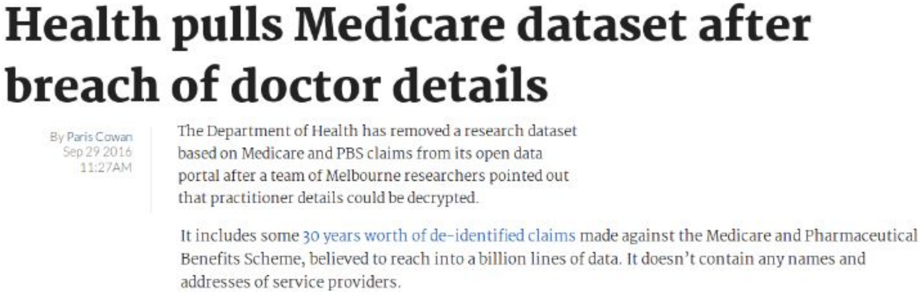
In **bluetooth data** (parties, timestamp, and RSSI), our approach correctly identifies individuals **26% of the time** out of 587 people using 1-hop graphs.

Important note: we do not believe that our results currently apply to robust privacy-preserving contact tracing protocols such as Google and Apple's Exposure Notification





Crețu, A.M., Monti, F., Marrone, S., Dong, X., Bronstein, M. and de Montjoye, Y.A., 2022. Interaction data are identifiable even across long periods of time. Nature Communications, 13(1), pp.1-11.

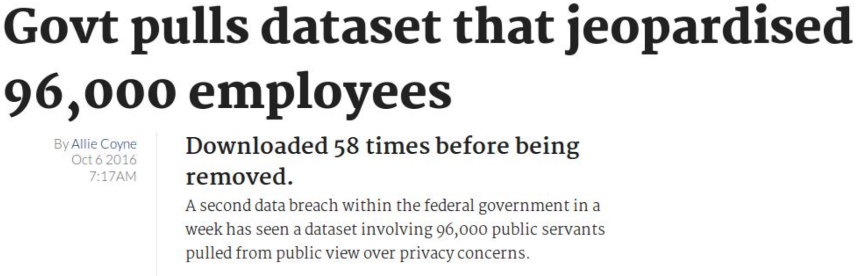


Health pulls Medicare dataset after breach of doctor details

By Paris Cowan
Sep 29 2016
11:27AM

The Department of Health has removed a research dataset based on Medicare and PBS claims from its open data portal after a team of Melbourne researchers pointed out that practitioner details could be decrypted.

It includes some 30 years worth of de-identified claims made against the Medicare and Pharmaceutical Benefits Scheme, believed to reach into a billion lines of data. It doesn't contain any names and addresses of service providers.

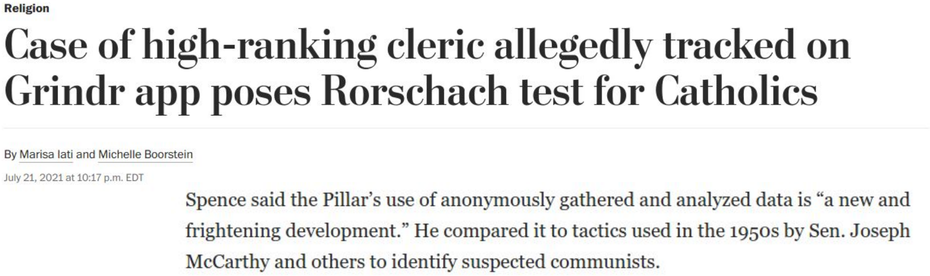


Govt pulls dataset that jeopardised 96,000 employees

By Allie Coyne
Oct 6 2016
7:17AM

Downloaded 58 times before being removed.

A second data breach within the federal government in a week has seen a dataset involving 96,000 public servants pulled from public view over privacy concerns.

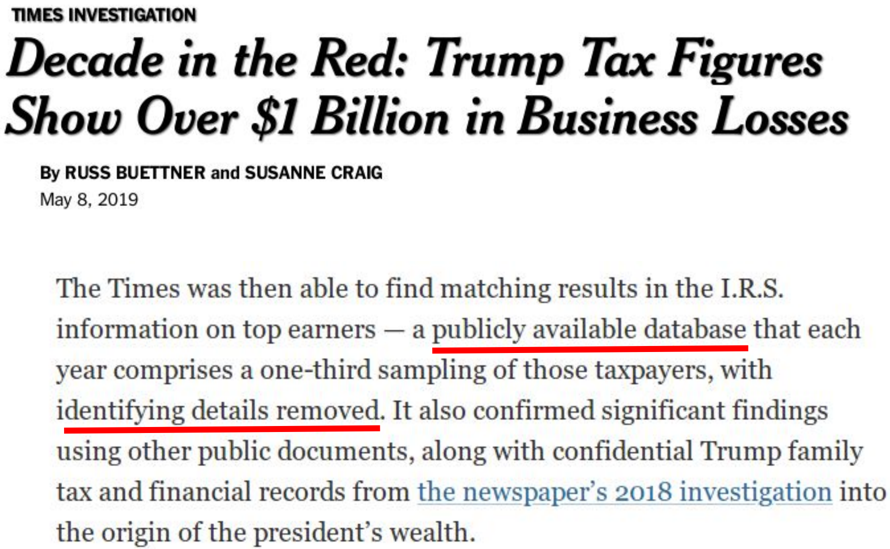


Religion

Case of high-ranking cleric allegedly tracked on Grindr app poses Rorschach test for Catholics

By Marisa Iati and Michelle Boorstein
July 21, 2021 at 10:17 p.m. EDT

Spence said the Pillar's use of anonymously gathered and analyzed data is "a new and frightening development." He compared it to tactics used in the 1950s by Sen. Joseph McCarthy and others to identify suspected communists.



TIMES INVESTIGATION

Decade in the Red: Trump Tax Figures Show Over $1 Billion in Business Losses

By RUSS BUETTNER and SUSANNE CRAIG
May 8, 2019

The Times was then able to find matching results in the I.R.S. information on top earners — a publicly available database that each year comprises a one-third sampling of those taxpayers, with identifying details removed. It also confirmed significant findings using other public documents, along with confidential Trump family tax and financial records from the newspaper's 2018 investigation into the origin of the president's wealth.

Anonymization (in the traditional, de-identification sense) doesn’t work*

* “We have currently **no reason to believe that an efficient enough, yet general, anonymization method will ever exist for high-dimensional data**, as all the evidence so far points to the contrary. The current deidentification model, where the data are anonymized and released, is obsolete and should not be used for policy.”

de Montjoye Y.-A., Pentland A.S. Response to Comment on Unique in the shopping mall: On the reidentifiability of credit card metadata, Science 351 (6279), 1274--1274, (2016)

“Anonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. [...] Anonymization remains somewhat **useful as an added safeguard**, but it is not robust against near-term future re-identification methods. **PCAST does not see it as being a useful basis for policy.**”

[US] President's Council of Advisors on Science and Technology, PCAST Report on Big Data and Privacy: A Technological Perspective

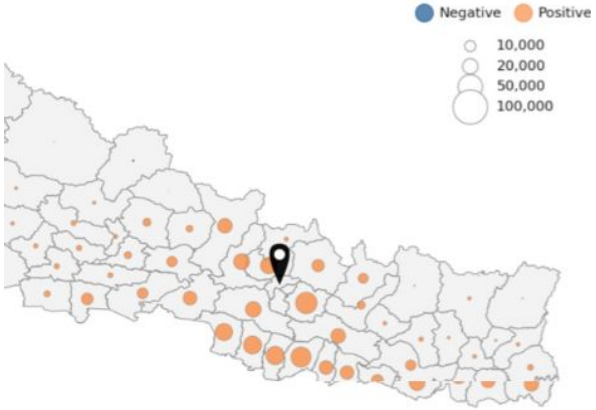
What is anonymous?

Negative definition: **Anonymous = cannot be re-identified**

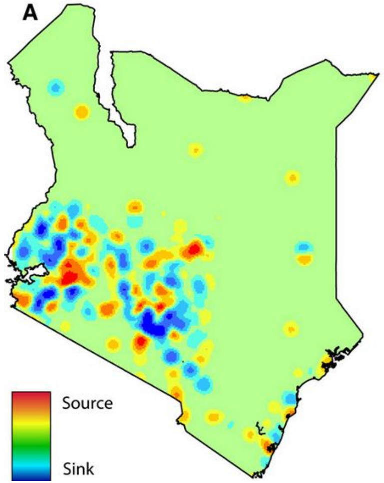
‘The principles of data protection should therefore not apply to anonymous information, [...] data rendered anonymous in such a manner that the data subject is not or no longer identifiable. [...] account should be taken of ***all the means reasonably likely to be used****, such as singling out, either* ***by the controller or by another person*** *to identify the natural person directly or indirectly. [...] taking into consideration the* ***available technology*** *at the time of the processing and technological developments.’*

Every new attack, if successful and credible, helps **defines what constitute anonymous data.**

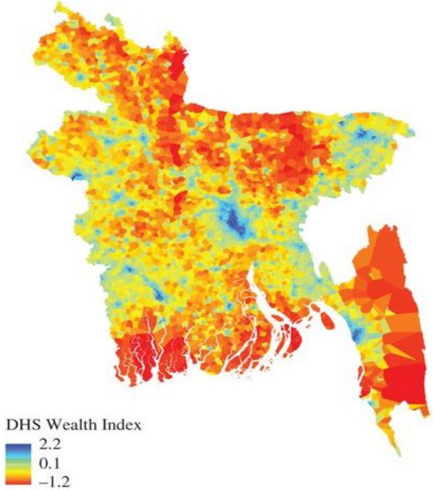
Data has a lot of potential and anonymous use of data is a strong simple promise



Wilson, R. et al. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake. PLoS currents, 8.

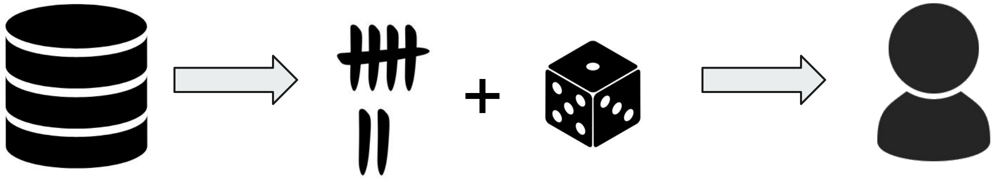


Wesolowski, A., et al. (2012). Quantifying the impact of human mobility on malaria. Science, 338(6104), 267-270.



Steele JE et al. (2017) Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface

Differentially private statistics

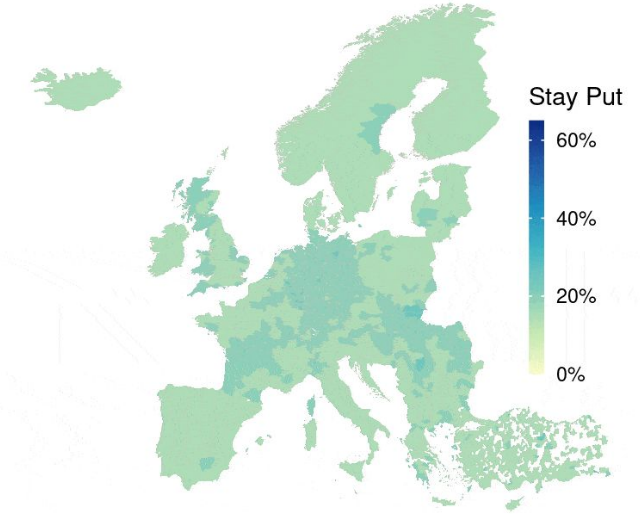


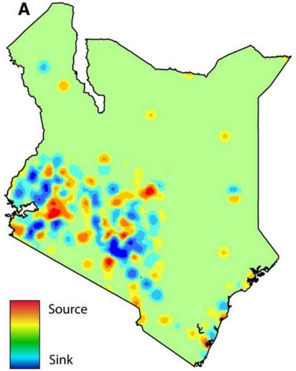
Data

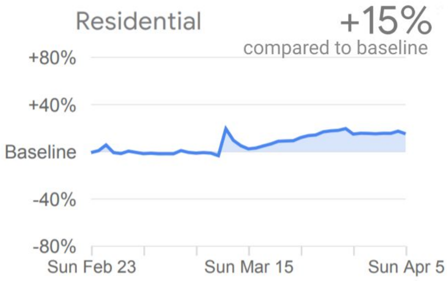
Counts + noise

Analyst

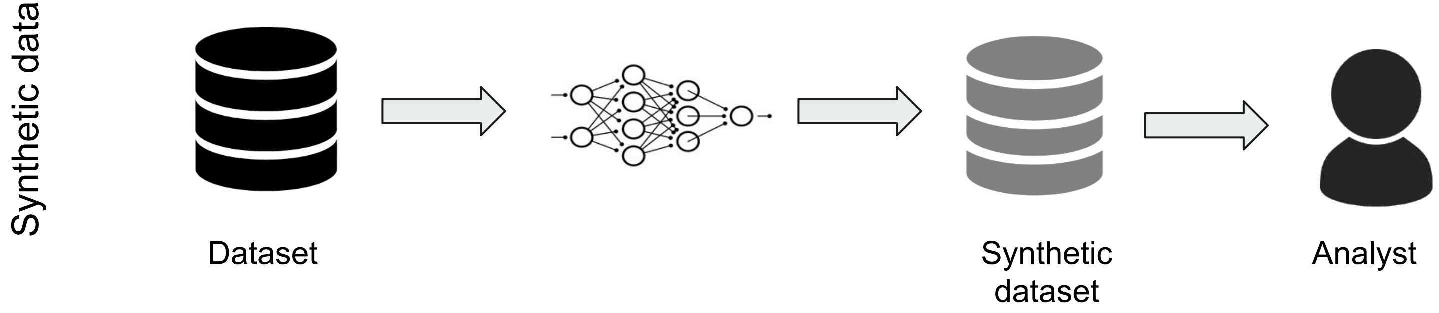
Date: 2020-03-04

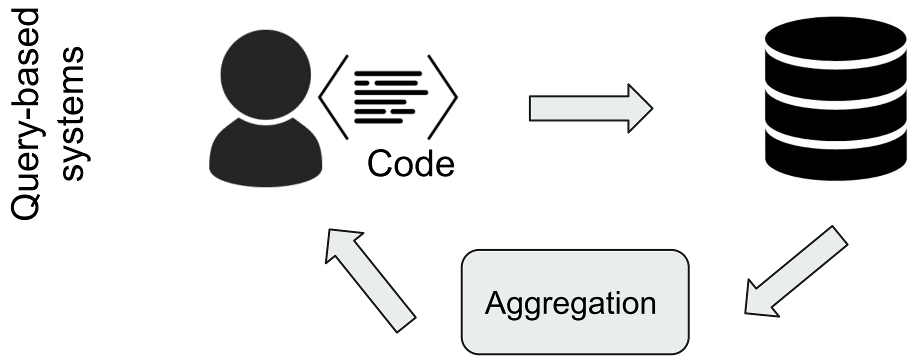






Query-based systems + synthetic data

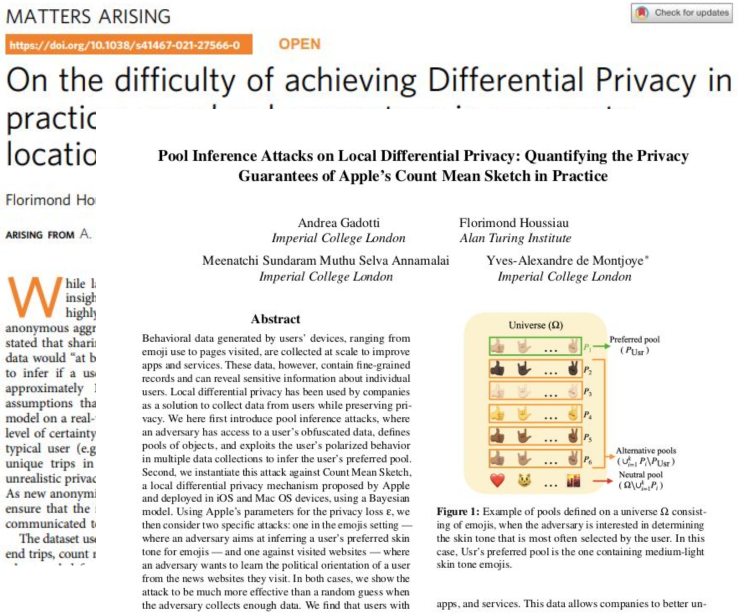




OPAL

open algorithms for better decisions

But… you **need** to (adversarially) test your solution



MATTERS ARISING

Check for updates

https://doi.org/10.1038/s41467-021-27566-0 OPEN

On the difficulty of achieving Differential Privacy in
practic
locatio

Florimond Ho

ARISING FROM A.

While l
insigh
highly
anonymous aggr
stated that shari
data would "at b
to infer if a us
approximately
assumptions tha
model on a real-
level of certainty
typical user (e.g
unique trips in
unrealistic privac
As new anonymi
ensure that the
communicated t
The dataset us
end trips, count

Pool Inference Attacks on Local Differential Privacy: Quantifying the Privacy Guarantees of Apple's Count Mean Sketch in Practice

Andrea Gadotti
Imperial College London

Florimond Houssiau
Alan Turing Institute

Meenatchi Sundaram Muthu Selva Annamalai
Imperial College London

Yves-Alexandre de Montjoye*
Imperial College London

Abstract

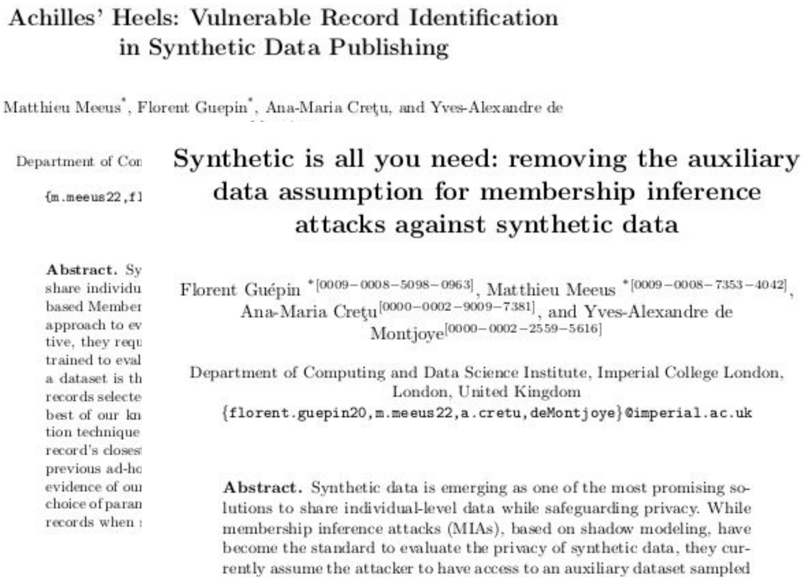
Behavioral data generated by users' devices, ranging from emoji use to pages visited, are collected at scale to improve apps and services. These data, however, contain fine-grained records and can reveal sensitive information about individual users. Local differential privacy has been used by companies as a solution to collect data from users while preserving privacy. We here first introduce pool inference attacks, where an adversary has access to a user's obfuscated data, defines pools of objects, and exploits the user's polarized behavior in multiple data collections to infer the user's preferred pool. Second, we instantiate this attack against Count Mean Sketch, a local differential privacy mechanism proposed by Apple and deployed in iOS and Mac OS devices, using a Bayesian model. Using Apple's parameters for the privacy loss ε, we then consider two specific attacks: one in the emojis setting — where an adversary aims at inferring a user's preferred skin tone for emojis — and one against visited websites — where an adversary wants to learn the political orientation of a user from the news websites they visit. In both cases, we show the attack to be much more effective than a random guess when the adversary collects enough data. We find that users with

Universe (Ω)
…
P_1
P_2
P_3
P_4
P_5
P_6
Preferred pool (P_{Usr})
Alternative pools ($\cup_{i=1}^{k} P_i \setminus P_{\text{Usr}}$)
Neutral pool ($\Omega \setminus \cup_{i=1}^{k} P_i$)

Figure 1: Example of pools defined on a universe Ω consisting of emojis, when the adversary is interested in determining the skin tone that is most often selected by the user. In this case, Usr's preferred pool is the one containing medium-light skin tone emojis.

apps, and services. This data allows companies to better un-

"Pure" DP can be hard to apply to continous data



Achilles' Heels: Vulnerable Record Identification in Synthetic Data Publishing

Matthieu Meeus*, Florent Guepin*, Ana-Maria Creţu, and Yves-Alexandre de

Department of Con

{m.meeus22,f1

Abstract. Sy
share individu
based Member
approach to ev
tive, they requ
trained to eval
a dataset is th
records selecte
best of our kn
tion technique
record's closes
previous ad-ho
evidence of ou
choice of paran
records when

Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data

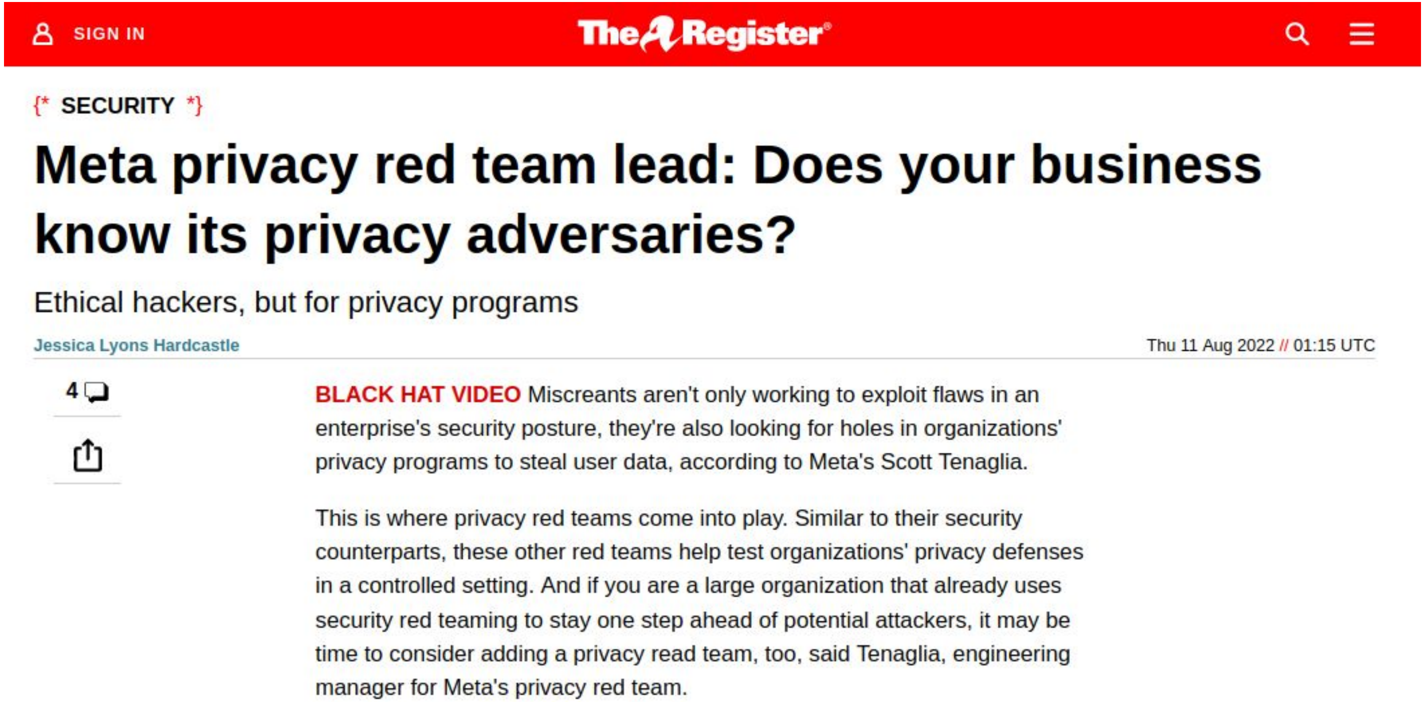
Florent Guépin *[0009-0008-5098-0963], Matthieu Meeus *[0009-0008-7353-4042], Ana-Maria Creţu[0000-0002-9009-7381], and Yves-Alexandre de Montjoye[0000-0002-2559-5616]

Department of Computing and Data Science Institute, Imperial College London, London, United Kingdom
{florent.guepin20,m.meeus22,a.cretu,deMontjoye}@imperial.ac.uk

Abstract. Synthetic data is emerging as one of the most promising solutions to share individual-level data while safeguarding privacy. While membership inference attacks (MIAs), based on shadow modeling, have become the standard to evaluate the privacy of synthetic data, they currently assume the attacker to have access to an auxiliary dataset sampled

Synthetic is not free from privacy concerns (outliers)

Adversarial attacks



SIGN IN

The Register®

{* SECURITY *}

Meta privacy red team lead: Does your business know its privacy adversaries?

Ethical hackers, but for privacy programs

Jessica Lyons Hardcastle

Thu 11 Aug 2022 // 01:15 UTC

4

BLACK HAT VIDEO Miscreants aren't only working to exploit flaws in an enterprise's security posture, they're also looking for holes in organizations' privacy programs to steal user data, according to Meta's Scott Tenaglia.

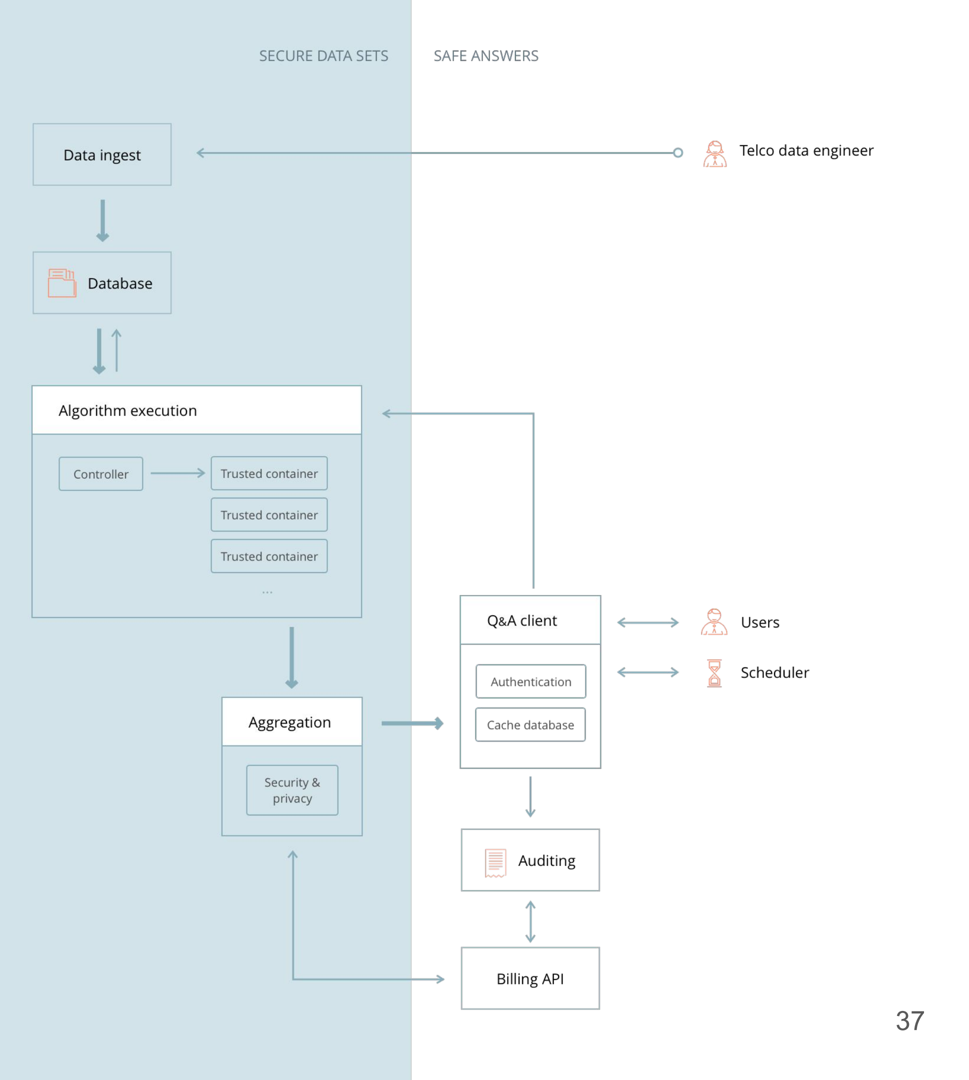
This is where privacy red teams come into play. Similar to their security counterparts, these other red teams help test organizations' privacy defenses in a controlled setting. And if you are a large organization that already uses security red teaming to stay one step ahead of potential attackers, it may be time to consider adding a privacy read team, too, said Tenaglia, engineering manager for Meta's privacy red team.

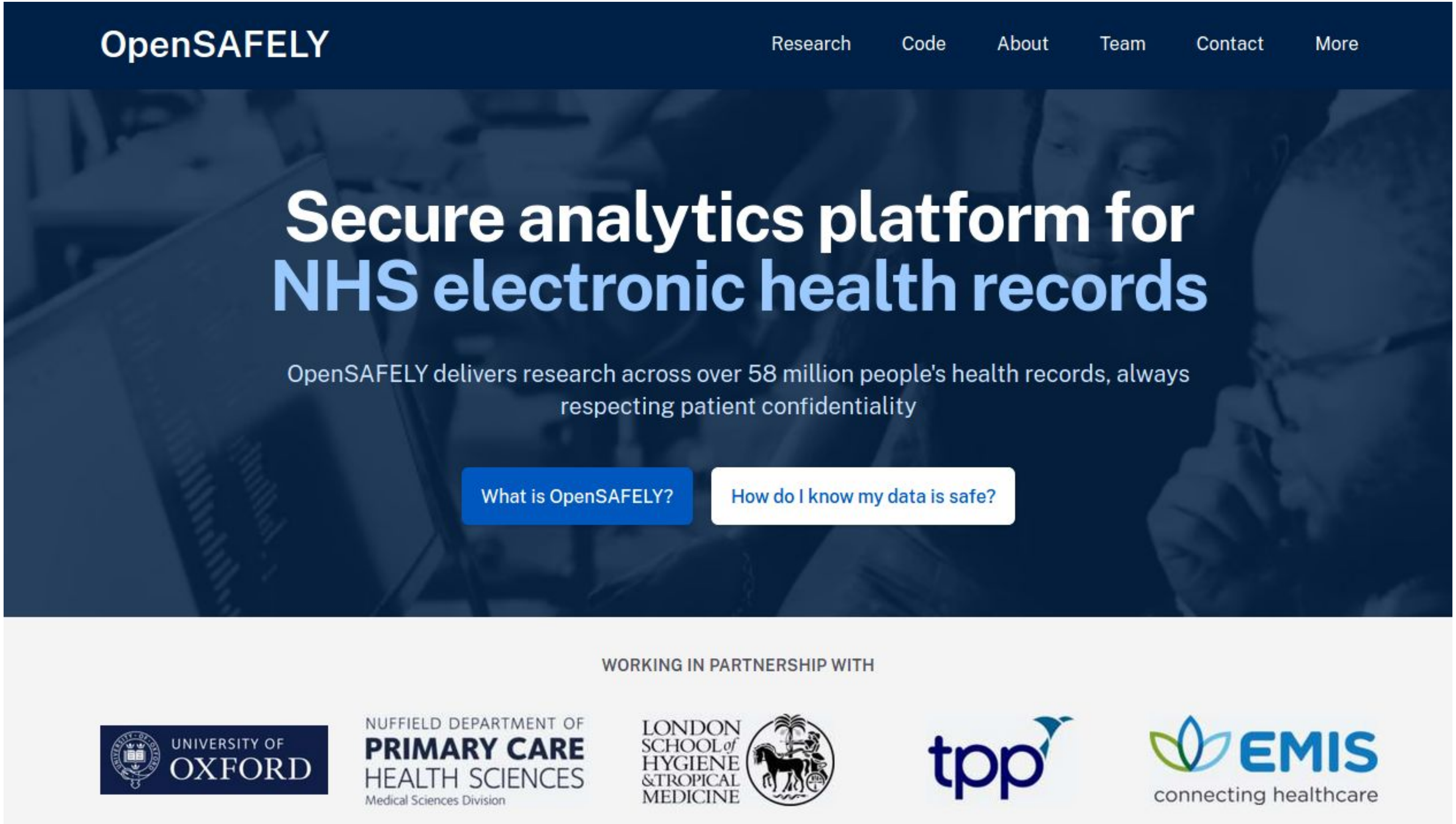
OPAL: Mobile phone data for NSOs

- Pilot projects in Colombia and Senegal
- 8M people's data in partnership with local NSOs
- All open-source software

Developed by:

With support from:




OpenSAFELY
Research
Code
About
Team
Contact
More
What is OpenSAFELY?
How do I know my data is safe?
WORKING IN PARTNERSHIP WITH
UNIVERSITY OF OXFORD
NUFFIELD DEPARTMENT OF PRIMARY CARE HEALTH SCIENCES
Medical Sciences Division
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE
tpp
EMIS
connecting healthcare

Secure analytics platform for NHS electronic health records

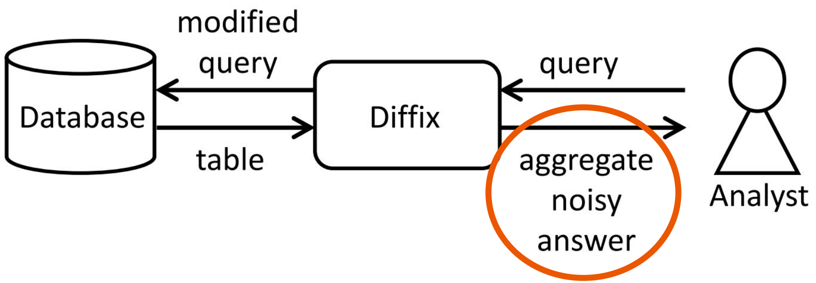
OpenSAFELY delivers research across over 58 million people's health records, always respecting patient confidentiality

Diffix is a privacy-preserving database system

Diffix is a patented commercial system developed by the company *Aircloak* and researchers at the Max Planck Institute for Software Systems.

Diffix operates as an **SQL proxy** between the analyst and the database. It aims at providing users with a **rich syntax** (great utility for general applications) and **infinite queries** (proposed as an alternative to Differential Privacy).

It was guaranteed to deliver "*GDPR-level anonymity for all use cases by CNIL (French Data Protection commission)*".



~~"How many people named Bob have a salary ≥ £100,000"~~

Q1 = "How many people have a salary ≤ £100,000"

Q2 = "How many people not named Bob have a salary ≤ £100,000"

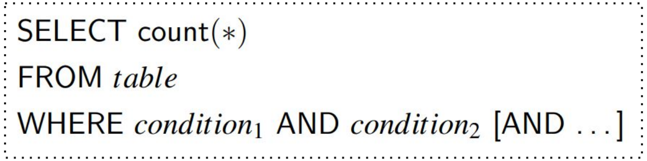
→ Q1 - Q2

This is a difference attack but there also exist averaging attacks, reconstruction attacks, membership attacks and more

A heuristic query-based system: Diffix

Diffix's aggregation: Bucket suppression

An analyst submits a (counting) SQL query to Diffix:



```
SELECT count(*)
FROM table
WHERE condition₁ AND condition₂ [AND ...]
```

(for simplicity, we write this as:)

$$Q \equiv \mathsf{count}(condition_1 \wedge condition_2 \wedge \ldots)$$

Bucket suppression

To prevent answering queries that concern only few individuals, Diffix implements **bucket suppression.** It suppresses all queries that select *too few users*.

$$T \sim \mathcal{N}(4, 1/2)$$

This means that queries directly targeting a specific user will fail.

Example:

$\texttt{count(}\textit{name = Bob} \wedge \textit{salary < £100,000}\texttt{)}$

Diffix's noise addition: layers of consistent noise to prevent known attacks

$$\widetilde{Q}(D) = Q(D) + \sum_{i=1}^{h} \mathsf{static}[C_i] + \sum_{i=1}^{h} \mathsf{dynamic}_Q[C_i]$$

For each condition C_i, Diffix adds **static** and **dynamic** noise. Both of these are pseudorandom values drawn from a normal distribution *N*(0,1), but **seeded** in a different way (so as to give *the same noise* if some conditions are met).

The ***static*** **noise** depends only on the condition:

$$static\_seed_C = \mathsf{XOR}(\mathsf{hash}(C), salt)$$

→ Prevents *averaging attacks*

The ***dynamic*** **noise** also depends on *the user-set* of the whole query (the IDs of all the rows selected) :

$$\begin{aligned} dynamic\_seed = \mathsf{XOR}(&static\_seed_C, \\ &\mathsf{hash}(uid_1), \ldots, \mathsf{hash}(uid_m)) \end{aligned}$$

→ Prevents *difference* attacks

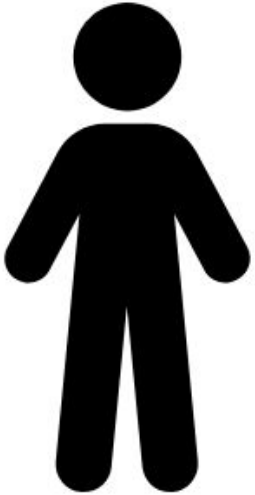
Is Bob making more than 100,000£?

Bob, a **40** years old **men** working in the department of **Computing.** Using a database of salaries at the university protected by Diffix, can I figure out if Bob is making more than 100,000£/y?

```
Q0  = count( age = 40 ∧ dept = Computing ∧ sex = men ∧
high-salary = True )
```

$Q^*_0 = |U_0|$ + static[age=40] + static[dept=Computing] + static[sex = men] + static[high-salary=True] + dynamic[age=40, U_0] + dynamic[dept=Computing, U_0] + dynamic[sex=men, U_0] + dynamic[high-salary=True, U_0]

- with $|U_0|$ = 0 or 1 with each each noise term ~ N(0,1)
- and will very almost certainly be *bucket suppressed*



Our attack(s) on Diffix: Going for the noise*

* Data-dependant noise has long been thought to be a potential risk but never lead to a vulnerability
e.g. Nissim, K., Raskhodnikova, S., & Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*

If Bob is making 100,000£

Bob :

age = 40

dept = computing

high-salary = T

(unique)

Sent: Q_1 = count(age = 40 ∧ high-salary = True) with Bob

Received: $Q^*_1 = |U_1|$ + static[age=40] + static[high-salary=True]

+ dynamic[age=40, U_1] + dynamic[high-salary=True, U_1]

Sent: Q_2 = count(age = 40 ∧ high-salary = True ∧ dept ≠ computing) without Bob

⇒ $U_1 = U_2 \cup \{\text{Bob}\}$ (because Bob is unique)

Received: $Q^*_2 = |U_1| - 1$ + static[age=40] + static[high-salary=True] + static[dept≠computing]

+ dynamic[age=40, U_2] + dynamic[high-salary=True, U_2] + dynamic[dept≠computing, U_2]

If Bob is making 100k£

Bob :

age = 40

dept = computing

high-salary = T

(unique)

Sent: Q_1 = count(age = 40 ∧ high-salary = True) with Bob

Received: Q^*_1 = ~~$|U_2|$ + static[age=40] + static[high-salary=True]~~

+ dynamic[age=40, U_1] + dynamic[high-salary=True, U_1]

Sent: Q_2 = count(age = 40 ∧ high-salary = True ∧ dept ≠ computing) without Bob

⇒ $U_1 = U_2 \cup \{\text{Bob}\}$

Received: Q^*_2 = ~~$|U_1|$~~ **- 1** ~~+ static[age=40] + static[high-salary=True] +~~ **static[dept≠computing]**

+ dynamic[age=40, U_2] + dynamic[high-salary=True, U_2] + dynamic[dept≠computing, U_2]

⇒ $\mathbf{q = Q^*_1 - Q^*_2 \sim N(1, 6)}$

If Bob is not making 100,000£

Bob:

age = 40

dept = computing

high-salary = F

(unique)

Send these queries:

Q_1 = count(age = 40 $\wedge$ high-salary = True) without Bob

Q_2 = count(age = 40 $\wedge$ high-salary = True $\wedge$ dept $\neq$ computing) without Bob

with $U_1 = U_2$ (because bob is unique)

Diffix replies:

$$Q^*_1 = |U_1| + \text{static[age=40]} + \text{static[high-salary=True]}$$
$$+ \text{dynamic[age=40, } U_1\text{]} + \text{dynamic[high-salary=True, } U_1\text{]}$$

$$Q^*_2 = |U_1| + \text{static[age=40]} + \text{static[high-salary=true]} + \text{static[dept}\neq\text{computing]}$$
$$+ \text{dynamic[age=40, } U_2\text{]} + \text{dynamic[high-salary=True, } U_2\text{]} + \text{dynamic[dept}\neq\text{computing, } U_2\text{]}$$

If Bob is not making 100k£

Bob:

```
    age = 40
    dept = computing
    high-salary = F
    (unique)
```

Send these queries:

Q_1 = count(age = 40 ∧ high-salary = True) without Bob

Q_2 = count(age = 40 ∧ high-salary = True ∧ dept ≠ computing) without Bob

⇒ $U_1 = U_2$

Diffix replies:

Q^*_1 = ~~$|U_1|$ +~~ 0 ~~+ static[age=40] + static[high-salary=True]~~

~~+ dynamic[age=40, U_1] + dynamic[high-salary=True, U_1]~~

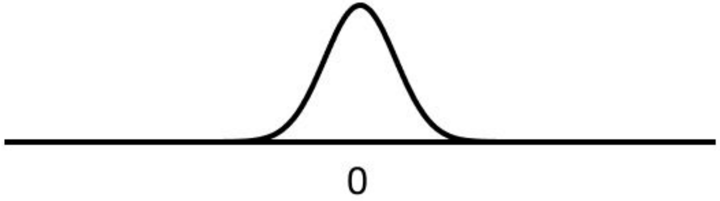
Q^*_2 = ~~$|U_1|$ + static[age=40] + static[high-salary=true] +~~ **static[dept≠computing]**

~~+ dynamic[age=40, U_2] + dynamic[high-salary=True, U_2]~~ **+ dynamic[dept≠computing, U_2]**

⇒ **$q = Q^*_1 - Q^*_2 \sim N(0, 2)$**

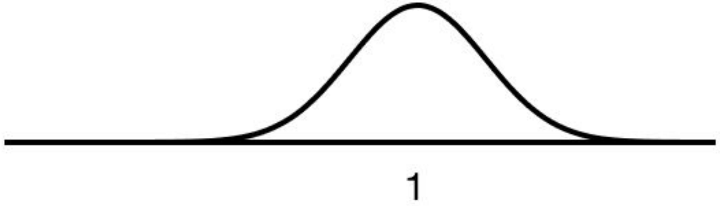
Exploiting noise in practice

if high-salary = True



Q1 - Q2 ~ ***N*(μ=0, σ=2)**

if high-salary = False



Q1 - Q2 ~ ***N*(μ=1, σ=2k+2)**

Using the attacker's background knowledge *(A,x$^{(A)}$)*:

$$Q_1 \equiv \text{count}(a_2 = x_2 \wedge \ldots \wedge a_k = x_k \wedge s = 0)$$

$$Q_1' \equiv \text{count}(a_1 \neq x_1 \wedge a_2 = x_2 \wedge \ldots$$
$$\ldots \wedge a_k = x_k \wedge s = 0)$$

Automating the attack:

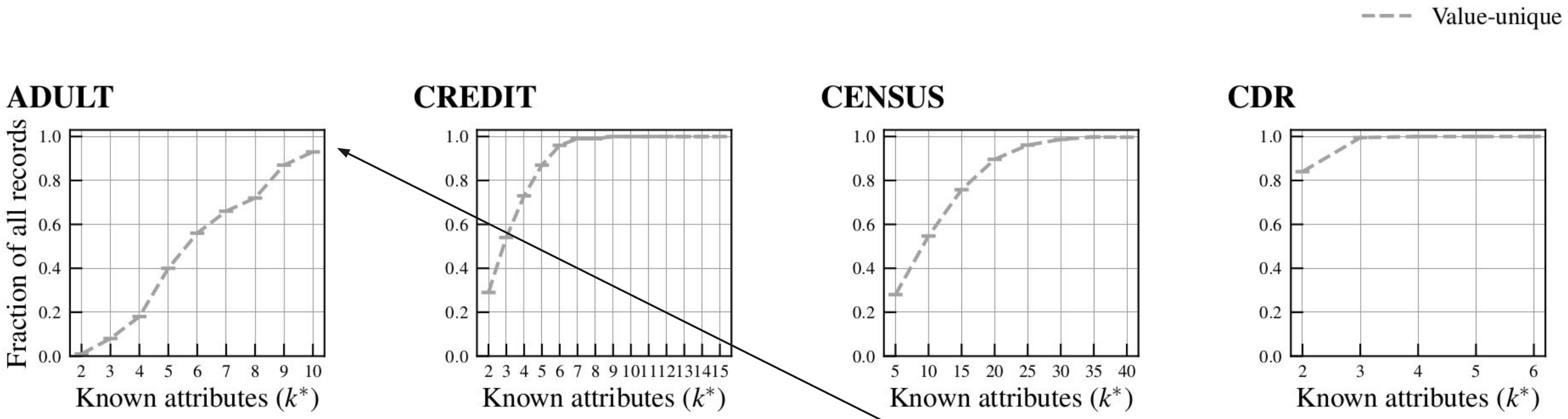
- Self-validating assumptions
- Subset exploration procedure
- Relying on dummies

Experimental validation

We implemented Diffix’s algorithm for counting queries, and applied it to four real-world datasets.

- **ADULT**: US census, 30k records, 11 attributes (secret: *salaryclass*).
- **CREDIT**: credit card applications, 690 records, 16 attributes (secret: *acceptance*).
- **CENSUS** : US census, 200k records, 42 attributes (secret: *salary >= 50k*)
- **CDR**: synthetic collection of phone metadata for a population of 2M people. Every user is treated as a record of 11.7M binary attributes (presence at a point in time+space) (secret: *presence in a random point*)

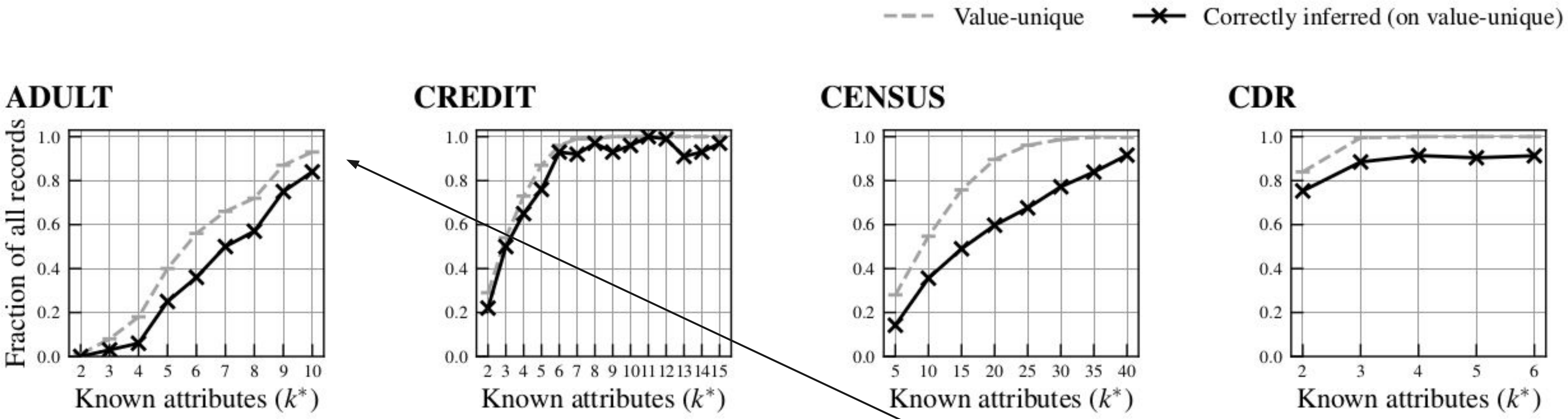
Experimental validation: without Diffix



- ADULT: US census, 30k records, 11 attributes (secret: *salaryclass*).
- CREDIT: credit card applications, 690 records, 16 attributes (secret: *acceptance*).
- CENSUS : US census, 200k records, 42 attributes (secret: *salary >= 50k*)
- CDR: synthetic collection of phone metadata for a population of 2M people. (secret: *presence in a random point*)

95% chances of a person to be value-unique with 10 points (attributes) if the attacker has **direct access** to the **pseudonymized dataset**

Experimental validation: with Diffix running our attack

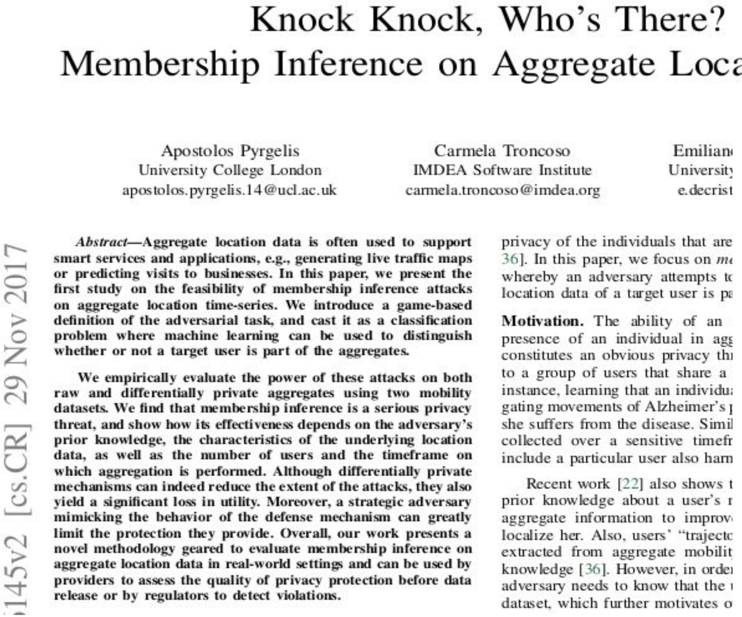


- ADULT: US census, 30k records, 11 attributes (secret: *salaryclass*).
- CREDIT: credit card applications, 690 records, 16 attributes (secret: *acceptance*).
- CENSUS : US census, 200k records, 42 attributes (secret: *salary >= 50k*)
- CDR: synthetic collection of phone metadata for a population of 2M people.

  (secret: *presence in a random point*)

Even with only access to aggregate noisy results sent by Diffix, we can **infer** the secret of a person ~90% of the time

Automating adversarial attacks



Knock Knock, Who's There? Membership Inference on Aggregate Location Data*

Apostolos Pyrgelis
University College London
apostolos.pyrgelis.14@ucl.ac.uk

Carmela Troncoso
IMDEA Software Institute
carmela.troncoso@imdea.org

Emilian
University
e.decrist

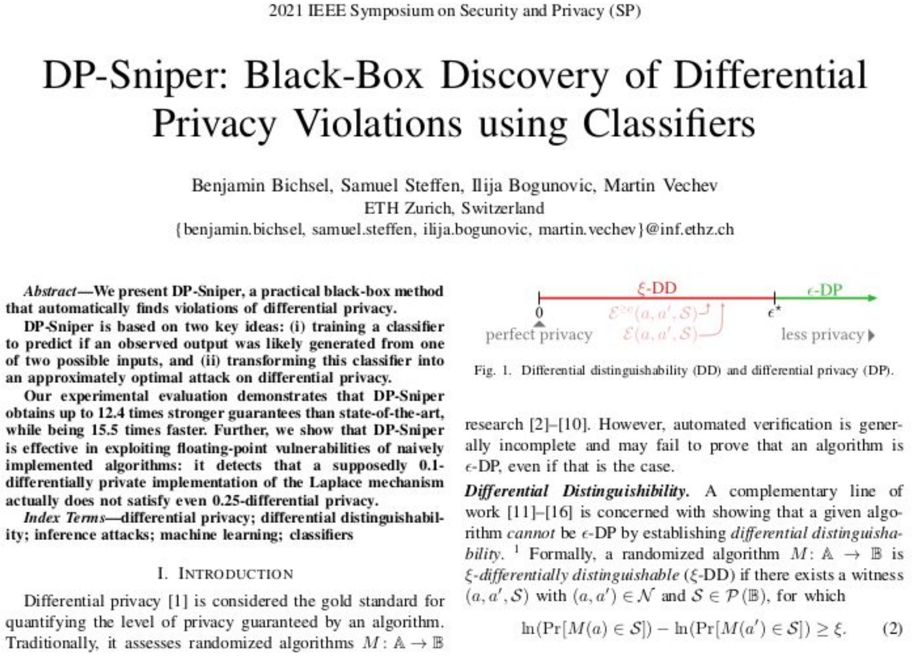
Abstract—Aggregate location data is often used to support smart services and applications, e.g., generating live traffic maps or predicting visits to businesses. In this paper, we present the first study on the feasibility of membership inference attacks on aggregate location time-series. We introduce a game-based definition of the adversarial task, and cast it as a classification problem where machine learning can be used to distinguish whether or not a target user is part of the aggregates.

We empirically evaluate the power of these attacks on both raw and differentially private aggregates using two mobility datasets. We find that membership inference is a serious privacy threat, and show how its effectiveness depends on the adversary's prior knowledge, the characteristics of the underlying location data, as well as the number of users and the timeframe on which aggregation is performed. Although differentially private mechanisms can indeed reduce the extent of the attacks, they also yield a significant loss in utility. Moreover, a strategic adversary mimicking the behavior of the defense mechanism can greatly limit the protection they provide. Overall, our work presents a novel methodology geared to evaluate membership inference on aggregate location data in real-world settings and can be used by providers to assess the quality of privacy protection before data release or by regulators to detect violations.

privacy of the individuals that are
36]. In this paper, we focus on *m*
whereby an adversary attempts t
location data of a target user is p

Motivation. The ability of an
presence of an individual in ag
constitutes an obvious privacy th
to a group of users that share a
instance, learning that an individu
gating movements of Alzheimer's
she suffers from the disease. Simil
collected over a sensitive timefr
include a particular user also harn

Recent work [22] also shows t
prior knowledge about a user's
aggregate information to improv
localize her. Also, users' "traject
extracted from aggregate mobilit
knowledge [36]. However, in orde
adversary needs to know that the
dataset, which further motivates o



2021 IEEE Symposium on Security and Privacy (SP)

DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers

Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, Martin Vechev
ETH Zurich, Switzerland
{benjamin.bichsel, samuel.steffen, ilija.bogunovic, martin.vechev}@inf.ethz.ch

Abstract—We present DP-Sniper, a practical black-box method that automatically finds violations of differential privacy.

DP-Sniper is based on two key ideas: (i) training a classifier to predict if an observed output was likely generated from one of two possible inputs, and (ii) transforming this classifier into an approximately optimal attack on differential privacy.

Our experimental evaluation demonstrates that DP-Sniper obtains up to 12.4 times stronger guarantees than state-of-the-art, while being 15.5 times faster. Further, we show that DP-Sniper is effective in exploiting floating-point vulnerabilities of naively implemented algorithms: it detects that a supposedly 0.1-differentially private implementation of the Laplace mechanism actually does not satisfy even 0.25-differential privacy.

Index Terms—differential privacy; differential distinguishability; inference attacks; machine learning; classifiers

I. Introduction

Differential privacy [1] is considered the gold standard for quantifying the level of privacy guaranteed by an algorithm. Traditionally, it assesses randomized algorithms $M: \mathbb{A} \to \mathbb{B}$

ξ-DD
ϵ-DP
0
$\epsilon^\star$
$\mathcal{E}^{2\eta}(a, a', \mathcal{S})$
$\mathcal{E}(a, a', \mathcal{S})$
perfect privacy
less privacy ▶

Fig. 1. Differential distinguishability (DD) and differential privacy (DP).

research [2]–[10]. However, automated verification is generally incomplete and may fail to prove that an algorithm is ϵ-DP, even if that is the case.

Differential Distinguishability. A complementary line of work [11]–[16] is concerned with showing that a given algorithm *cannot* be ϵ-DP by establishing *differential distinguishability.* [1] Formally, a randomized algorithm $M: \mathbb{A} \to \mathbb{B}$ is *ξ-differentially distinguishable* (ξ-DD) if there exists a witness $(a, a', \mathcal{S})$ with $(a, a') \in \mathcal{N}$ and $\mathcal{S} \in \mathcal{P}(\mathbb{B})$, for which

$$\ln(\Pr[M(a) \in \mathcal{S}]) - \ln(\Pr[M(a') \in \mathcal{S}]) \geq \xi. \qquad (2)$$

Search space of attacks

The attacks we search for consist of two components:

1. A set of queries $q_1, \ldots, q_m$

2. A mathematical (arithmetic) function *G* to combine their answers in order to infer the sensitive attribute *s*:

$$\hat{s} = G(q_1(D), \ldots, q_m(D)) \approxeq s$$

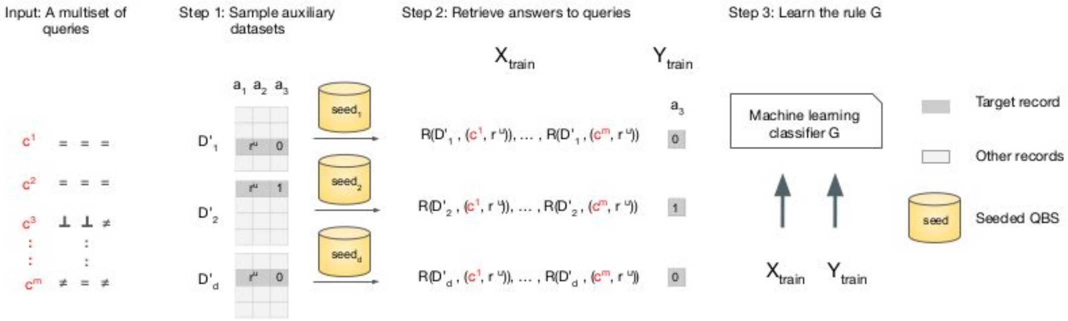
QuerySnout

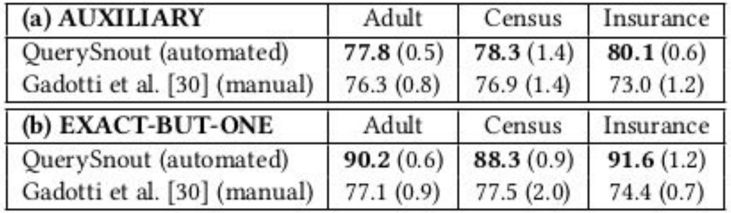
1. Given any set of queries, we use machine learning to learn a function G that infers the unknown attribute based on their answers.
2. We explore the space of queries to find the best attacks using evolutionary search techniques.

Automating the discovery of vulnerabilities

Using genetic algorithms, QuerySnout **automatically discover vulnerabilities in query-based systems.**

It takes as input a target record and the QBS as a black box, analyzes its behavior on one or more datasets, and outputs a multiset of queries together with a rule to combine answers to them in order to reveal the sensitive attribute of the target record.


Input: A multiset of queries
Step 1: Sample auxiliary datasets
Step 2: Retrieve answers to queries
Step 3: Learn the rule G
c^1 = = =
c^2 = = =
c^3 ⊥ ⊥ ≠
c^m ≠ = ≠
a_1 a_2 a_3
D'_1
D'_2
D'_d
r^u 0
r^u 1
r^u 0
$seed_1$
$seed_2$
$seed_d$
X_{train}
Y_{train}
a_3
$R(D'_1, (c^1, r^u)), \dots, R(D'_1, (c^m, r^u))$
$R(D'_2, (c^1, r^u)), \dots, R(D'_2, (c^m, r^u))$
$R(D'_d, (c^1, r^u)), \dots, R(D'_d, (c^m, r^u))$
0
1
0
Machine learning classifier G
X_{train} Y_{train}
Target record
Other records
seed
Seeded QBS



(a) AUXILIARY	Adult	Census	Insurance
QuerySnout (automated)	**77.8** (0.5)	**78.3** (1.4)	**80.1** (0.6)
Gadotti et al. [30] (manual)	76.3 (0.8)	76.9 (1.4)	73.0 (1.2)
(b) EXACT-BUT-ONE	**Adult**	**Census**	**Insurance**
QuerySnout (automated)	**90.2** (0.6)	**88.3** (0.9)	**91.6** (1.2)
Gadotti et al. [30] (manual)	77.1 (0.9)	77.5 (2.0)	74.4 (0.7)

Cretu, A.M, Houssiau F., Cully, A. and de Montjoye, Y.A., 2022. QuerySnout: Automating the Discovery of Attribute Inference Attacks against Query-Based Systems ACM CCS

Making modern anonymization work

- Help the transition of companies and government agencies to **modern anonymization techniques**
- No solution is perfect, ensure we keep **testing proposed and deployed solutions through attacks**
- Develop guidance and legislation to give **more legal certainty** and ensure an adequate level of protection

Thank you!

https://cpg.doc.ic.ac.uk/
Twitter: @yvesalexandre