40 years Census Decision: The census and understanding privacy protection in the transition of time

Elbsides conference
17 November 2023

KARLSTAD
UNIVERSITY
SWEDEN

Simone Fischer-Hübner

Chalmers University of Technology

Part I: 40 years Census Decision & Census Debate

Census Decision 1983 – German Constitutional Court (BVerfG)

- **Background:**
  - Planned Nation-wide census, update of citizen registry
  - Legal complaints to BverfG (Wild & Stadler-Euler / Steinmüller, Brunnstein & Podlech)
  - Declared by BVerfG as non-constitutional in December 1983

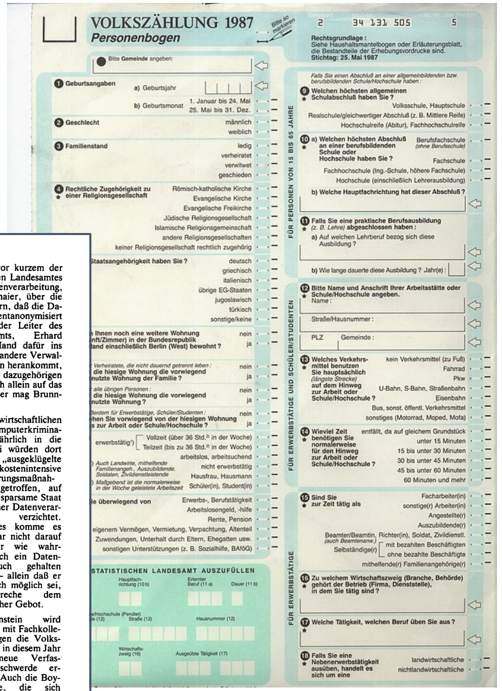
https://www.ndr.de/geschichte/chronologie/Volkszaehlung-1983-Protest-ganz-ohne-Twitter,brunnstein103.html

Main legal privacy principles declared by the German Constitutional Court (BVerfGE 65, 1):

- **Right to informational self-determination** derived from German Constitution (Art 1 I & 2 I GG)
- There are **no ”non-sensitive data”**
- Principle of **Purpose Binding** emphasized
- Privacy not only important for protecting individuals but also for **democracy & society** as a whole
- **Effective anonymisation** (”faktische Anonymisierung”) of census data demanded

Source: Michael Dick/ picture-alliance/ dpa

Census Debate 1987



Discussion:

- Does the deletion of directly identifying personal data (name, address) render the census data effectively anonymous?

-> Simple simulation model demonstrated:

Majority (>= 90%) are still identifiable

(BSc thesis - Fischer-Hübner 1986).

- Alternative: use of existing databases with (privacy-enhancing) statistical inference controls? (MSc thesis - Fischer-Hübner 1987).

VOLKSZÄHLUNG

Sicherster Ort

Computer-Experten beweisen: Die anonymen Volkszählungsdaten sind zu knacken.

An juristischen Einwänden, da waren sich Statistiker und Politiker mit Ausnahme der Grünen ganz sicher, würde die Volkszählung in diesem Jahr nicht wieder scheitern. „Peinlich genau und in vollem Umfang", verkündete das Statistische Bundesamt in Wiesbaden, habe das Bonner Parlament die Auflagen erfüllt, die ihm 1983 das Bundesverfassungsgericht in seinem Volkszählungsurteil gestellt hatte.

Der Hamburger Informatik-Professor Klaus Brunnstein hingegen, damals einer der Kläger in Karlsruhe, hat daran Zweifel. Einiges spreche dafür, so Brunnstein, daß die Strategen der Volkszählung „das Urteil entweder nur oberflächlich gelesen oder aber wesentliche Argumente daraus nicht verstanden haben".

Eine Mahnung der Richter vor allem habe der Gesetzgeber „bewußt in den Wind geschlagen". Es sei keineswegs sichergestellt, daß die Daten, die bei der großen Bürgerbefragung am 25. Mai erhoben und ohne Namen und Anschriften gespeichert werden sollen, tatsächlich anonym bleiben. Das aber, hatten die Karlsruher Richter gefordert, sei „zur Sicherung des Rechts auf informationelle Selbstbestimmung" notwendig.

Vor der Gefahr, daß aus den Zahlenkolonnen auf den elektronischen Datenträgern wieder Erkenntnisse über Einzelpersonen gewonnen werden können, hatte Brunnstein schon früher gewarnt. Der Innenausschuß des Bundestags folgte daher dem Vorschlag des Computer-Experten, ins Volkszählungsgesetz 1987 ausdrücklich ein „Verbot der Reidentifizierung" aufzunehmen. Danach dürfen die namenlosen Daten nicht „zum Zweck der Herstellung eines Personenbezugs" zusammengeführt werden.

Allerdings: Verhindert werden kann solcher Datenmißbrauch durch das bloße Verbot nicht. Die Vorschrift hat, wie Brunnstein mittlerweile einsieht, „eher akademischen Charakter". Die Strafandrohung (Gefängnis bis zu einem Jahr oder Geldstrafe) laufe „völlig ins Leere", da niemand Verstöße gegen das Gesetz erkennen und deshalb auch nicht anzeigen könne.

Wie einfach es gelingt, mit wenigen Erhebungsmerkmalen aus der Volkszählung eine bestimmte Person aus einer großen Bevölkerungsgruppe herauszufiltern, wiesen der Professor und die Informatik-Studentin Simone Fischer-Hübner in einem Simulationsmodell auf dem Computer nach.

Aus Unterlagen des Wiesbadener Statistischen Bundesamtes und des Statistischen Landesamtes Hamburg erzeugte die Studentin in einem IBM-Personalcomputer eine Modellbevölkerung von 100 000 Personen, die gleichsam ein auf Kleinstadtformat geschrumpftes Hamburg repräsentieren.

Die Informatikerin erfaßte manche der von den Volkszählern gewünschten Personenmerkmale sogar weniger genau, als es die amtlichen Statistiker beabsichtigen. Dennoch gelang es ihr, fast alle Personen ihrer Modellbevölkerung der Datenanonymität wieder zu entreißen.

Der Werbeslogan der Volkszähler „Ihr Privatleben ist vollkommen Ihr Bier" bröselte schon unter leichter elektronischer Attacke: Für Angehörige von Minderheiten, Personen jüdischen Glaubens oder Männer und Frauen mit seltenen Berufen etwa reichen bereits drei Angaben – beispielsweise über Alter, Geschlecht und Schulabschluß – aus, sie unter 100 000 Menschen genau zu identifizieren.

„Maximal zehn Daten", so die Informatikerin, seien nötig, um nahezu jeden im Datenwust zu finden – insgesamt sind es 33 Fragen, die den Volkszählern beantwortet werden müssen.

Lediglich verwitwete Rentnerinnen über 60 sind gegen Reidentifizierung ziemlich gefeit: Da diese Frauen keinen Beruf ausüben und oft allein oder im Altersheim leben, greifen die erfolgreichsten elektronischen Suchangaben – die Daten über Beruf und im selben Haushalt lebende Verwandte – nicht. Simone Fischer-Hübner: „So gesehen sind vielleicht die Altersheime der sicherste Ort."

Zwar spottete noch vor kurzem der Präsident des Bayerischen Landesamtes für Statistik und Datenverarbeitung, Hans Helmut Schiedermaier, über die Erkenntnis von Skeptikern, daß die Daten „angeblich leicht" entanonymisiert werden können. Und der Leiter des Hamburger Landesamts, Erhard Hruschka, will seine Hand dafür ins Feuer legen, daß „keine andere Verwaltungsdienststelle an Daten herankommt, die Rückschlüsse auf die dazugehörigen Personen zulassen". Doch allein auf das Berufsethos der Statistiker mag Brunnstein nicht vertrauen.

Immerhin gebe es in wirtschaftlichen Rechenzentren „eine Computerkriminalität, deren Schaden jährlich in die Milliarden geht". Dabei würden dort noch „ausgeklügelte und kostenintensive Sicherungsmaßnahmen" getroffen, auf die der sparsame Staat bei seiner Datenverarbeitung verzichtet. Überdies komme es auch gar nicht darauf an, für wie wahrscheinlich ein Datenmißbrauch gehalten werde – allein daß er technisch möglich sei, widerspreche dem Karlsruher Gebot.

Brunnstein wird deshalb mit Fachkollegen gegen die Volkszählung in diesem Jahr eine neue Verfassungsbeschwerde erheben. Auch die Boykotteure, die sich grundsätzlich nicht ausforschen lassen wollen, sammeln sich wieder. Weit über 200 Gruppen („VoBo-Inis") sind, viereinhalb Monate vor dem Zähltag, bereits aktiv.

Falls weder Verweigerung noch Verfassungsbeschwerde den gewünschten Erfolg haben, wollen Volkszählungsgegner den Statistikern wenigstens die Freude an den Daten vergällen. Computer-Hacker, so wird aus der VoBo-Szene berichtet, planen, die Auswertung der Volkszählungsdaten auf den Rechnern der Statistischen Landesämter durch Computer-„Viren" lahmzulegen. Die schwer zu ortenden zerstörerischen Programmbefehle können zu nutzlosem Daten-Wirrwarr oder auch zum Zusammenbruch ganzer Systeme führen.

Die Viren brauchen wohl nicht einmal von außen in die Rechner der Statistischen Ämter geschleust zu werden. In den Rechenzentren befinden sich, wie Brunnstein glaubt, „ebenso VoBo-Sympathisanten wie überall in der Bevölkerung". ◆

Informatiker Simone Fischer-Hübner, Brunnstein
Bevölkerungsmodell auf dem Personalcomputer

DER SPIEGEL, Nr. 3/1987

89

Source: B2836 Carsten Rehder

Big boycott protests –

However, legal complaints to BverfG unsuccessful.

Part II : Lessons Learned since then…

Importance of Census Decision & Census Debate – Lessons learned

(1) No non-sensitive data & Importance of Privacy for Democrary

Private traits and attributes are predictable from digital records of human behavior

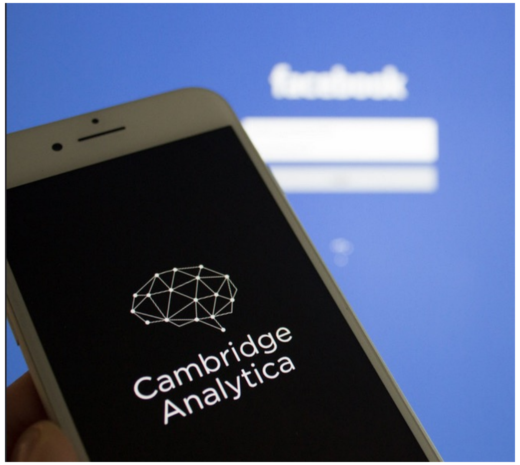
Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

Author Affiliations

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

Abstract

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest ac examples of associations between attributes and Likes and dis privacy.



Cambridge Analytica Data Breach with impact on US American Elections

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Importance of Census Decision & Census Debate – Lessons learned (II)

(1) Deletion of directly identifiable data ≠ Anonymisation – Re-identification is easy

Latanya Sweeney – experiments on 1990 US census data -

87% of the US population can be uniquely identified by gender, ZIP code and full date of birth

(L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000).

Failures of (simple) ”anonymisation” by just deleting/replacing attributes

AMERICA Online

- released a dataset of search queries from ca. 650K users, 2006
- replaced user names with numbers

New York Times reporter exemplified easy re-identification:

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

Credit: Erik S. Lesser for The New York Times

NETFLIX

- released 100M ratings from ca. 480k users, 2006
- claimed that all personal data was removed from the set

Re-identification by matching with public database: IMDb

- Netflix data: not two records are similar more than 50%.
- If the profile can be matched up to 50% similarity to a profile in IMDB, then the adversary can identity the profile with good likelihood.

(A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset),” in Proc. 29th IEEE Symposium on Security and Privacy, 2008.)

ELECTRONIC FRONTIER FOUNDATION

How Unique Is Your Web Browser?

Peter Eckersley*

Electronic Frontier Foundation,
pde@eff.org

Abstract. We investigate the degree to which modern web browsers are subject to "device fingerprinting" via the version and configuration information that they will transmit to websites upon request. We implemented one possible fingerprinting algorithm, and collected these fingerprints from a large sample of browsers that visited our test side, **panopticlick.eff.org**. We observe that the distribution of our fingerprint contains at least 18.1 bits of entropy, meaning that if we pick a browser at random, at best we expect that only one in 286,777 other browsers will share its fingerprint. Among browsers that support Flash or Java, the situation is worse, with the average browser carrying at least 18.8 bits of identifying information. 94.2% of browsers with Flash or Java were unique in our sample.

By observing returning visitors, we estimate how rapidly browser fingerprints might change over time. In our sample, fingerprints changed quite rapidly, but even a simple heuristic was usually able to guess when a fingerprint was an "upgraded" version of a previously observed browser's fingerprint, with 99.1% of guesses correct and a false positive rate of only 0.86%.

Eckersley, P. (2010). How unique is your web browser?. In *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Springer*

Art. 29 Data Protection Working Party – Opinion 05/2014 on Anonymisation Techniques

- Anonymisation - data must be processed in such a way that it can no longer be used to identify a natural person by using **“all the means likely reasonably to be used”** by either the **controller** or a **third party….**

Part III: Effective PETs as Enablers - Solutions & Challenges

K-Anonymity (Sweeney & Samarati)

K-anonymity: “**Each** value combination of the **quasi-identifiers (demographic data)** occurs **at least** *k* times”

(Enforced by generalisation/supression of attribute values).

Example: K-anonymisation for k=2:

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it's complicated	4	E
	24.12.98	female	1268	it's complicated	4	A

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	B
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it's complicated	4	E
	1990‘s	female	12**	it's complicated	4	A

2020 US Census & Differential Privacy

United States® Census Bureau

Partners | Researchers | Educators | Survey Respondents | News | NAICS Codes | Jobs | About Us | Contact Us

Topics | Data & Maps | Surveys & Programs | Resource Library

Search data, events, resources, and more

// Census.gov / 2020 Census Program Management / Processing the Count / Disclosure Avoidance Modernization / Understanding Differential Privacy

Within Disclosure Avoidance Modernization

Blogs

Demonstration Data & Progress Metrics

Understanding Differential Privacy

Census confidentiality protections—what we call “disclosure avoidance”—have evolved over time to keep pace with emerging threats. Since the 1990 Census we’ve added “noise”—or variations from the actual count—to the collected data. For 2020 Census data we’re applying noise using a newer protection framework based on “differential privacy.” Learn more here about why and how we’re modernizing our protections and how you can engage in the process.

Share Facebook Twitter LinkedIn

For an overview, read this brief: Why the Census Bureau Chose Differential Privacy

Harvard Data Science Review • Special Issue 2: Differential Privacy for the 2020 U.S. Census

The 2020 Census Disclosure Avoidance System TopDown Algorithm

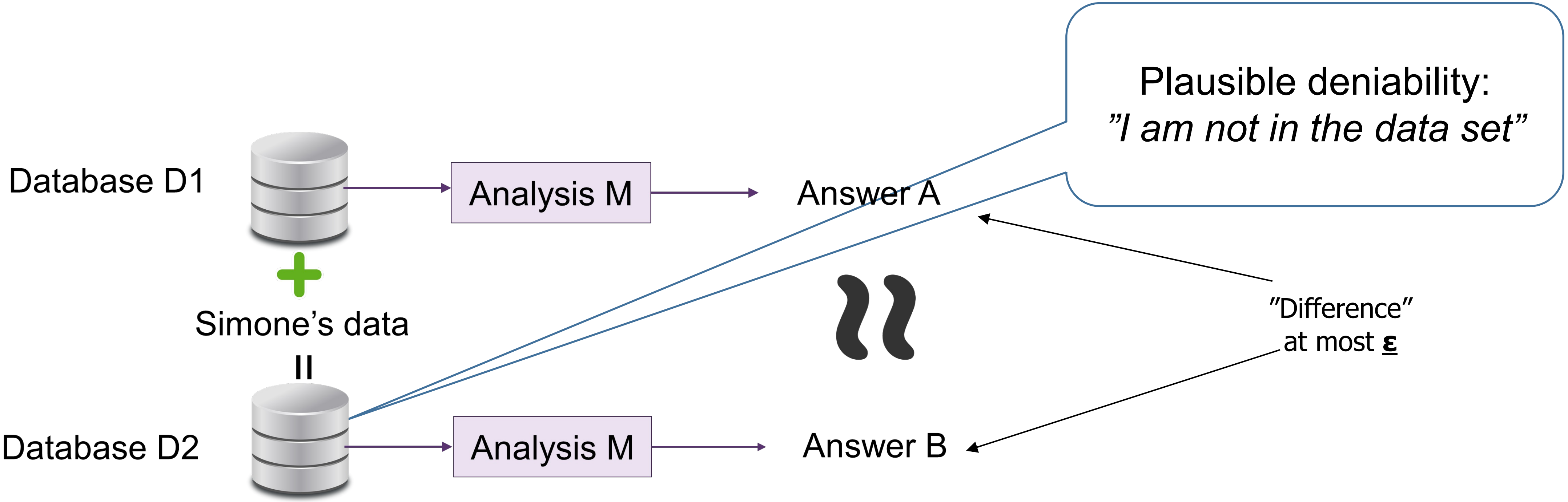
John Abowd[1] **Robert Ashmead**[1] **Ryan Cumings-Menon**[1] **Simson Garfinkel**[2] **Micah Heineck**[3] **Christine Heiss**[3] **Robert Johns**[3] **Daniel Kifer**[1,4,5] **Philip Leclerc**[1] **Ashwin Machanavajjhala**[6,7] **Brett Moran**[1] **William Sexton**[2,7] **Matthew Spence**[1] **Pavel Zhuravlev**[1]

[1]United States Census Bureau, Suitland, Maryland, United States of America,
[2]Formerly United States Census Bureau, Suitland, Maryland, United States of America,
[3]Knexus Research Corporation, Alexandria, Virginia, United States of America,
[4]Department of Computer Science and Engineering, School of Electrical Engineering and Computer Science, Pennsylvania State University, State College, Pennsylvania, United States of America,
[5]Center for Social Data Analytics, Pennsylvania State University, State College, Pennsylvania, United States of America,
[6]Department of Computer Science, Duke University, Durham, North Carolina, United States of America,
[7]Tumult Labs, Durham, North Carolina, United States of America

Published on: Jun 24, 2022

DOI: https://doi.org/10.1162/99608f92.529e3cb9

Differential Privacy (Dwork et al., 2006)



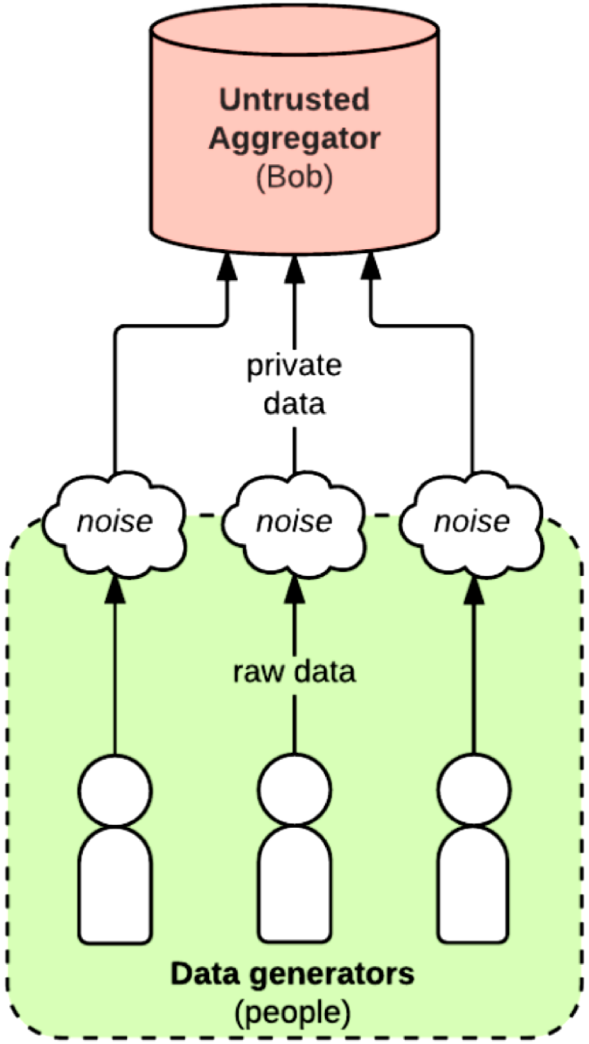
Definition: ϵ-Differential Privacy

$$\frac{\Pr(M(D) = C)}{\Pr(M(D_{\pm i}) = C)} < \mathrm{e}^{\epsilon}$$

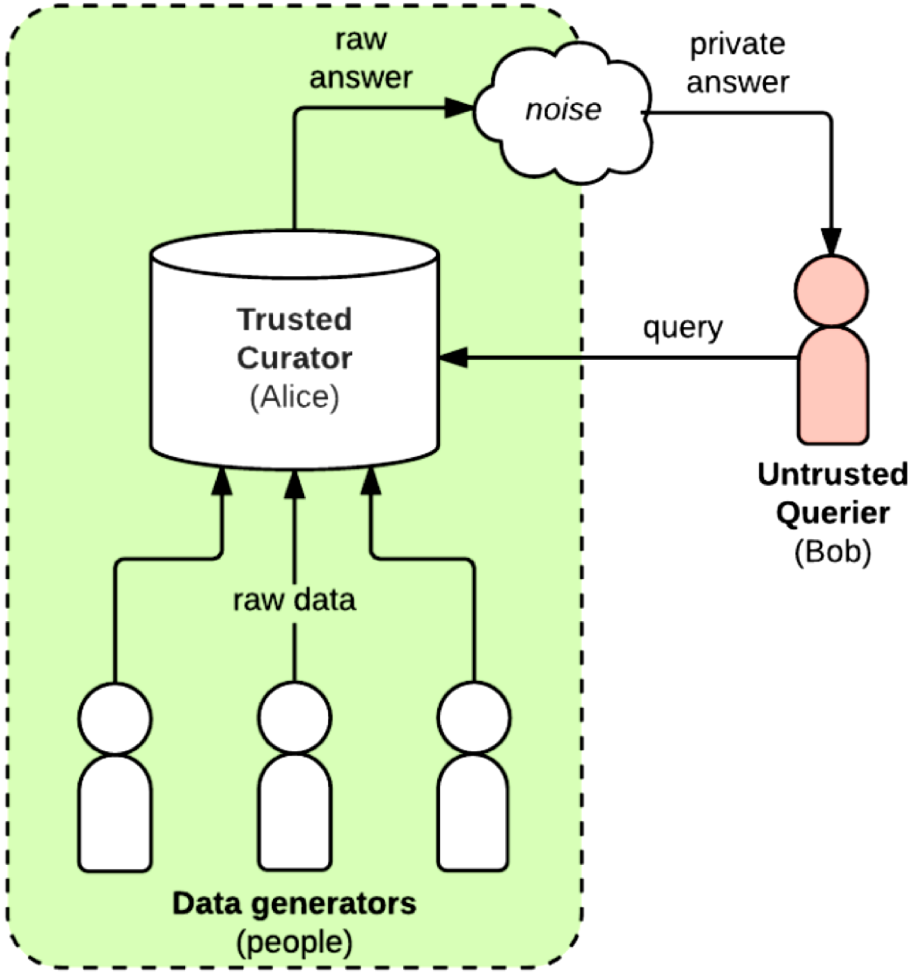
For any $|D_{\pm i} - D| \leq 1$ and any $C \in Range(M)$.

$(e^{\varepsilon} \approx 1 + \varepsilon$ *for small* $\varepsilon)$

Differential privacy - models



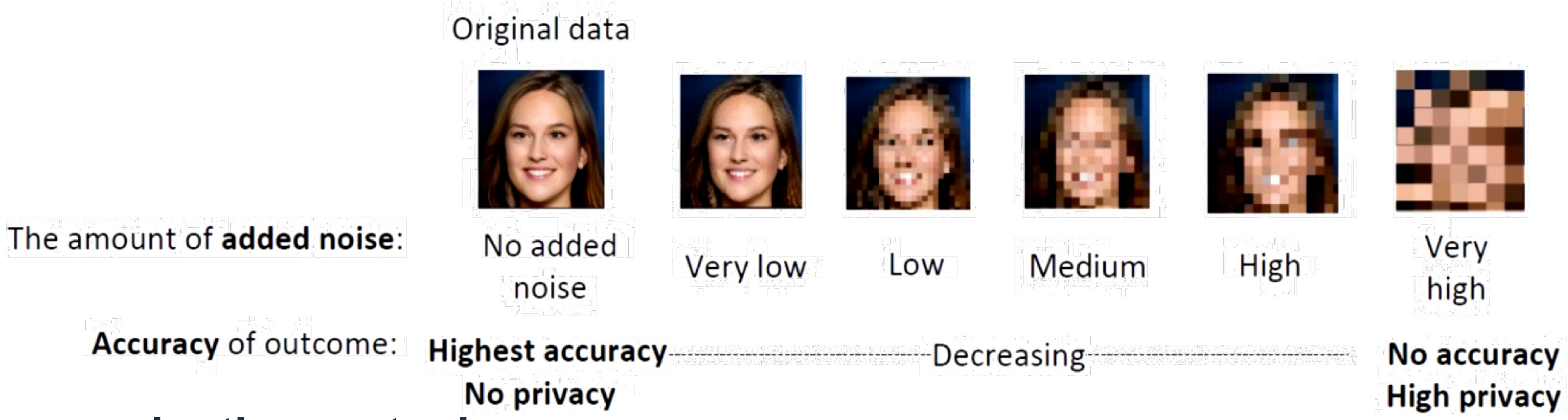
Local privacy



Global privacy

Image credit: Bennett Cyphers

Challenge: Explaining Differential Privacy (DP)



Interview-based User-Study – evaluating metaphors:

- **Misconceptions, triggered by digital-world analogies:**
  - Knowledge of encryption - > DP is reversible
  - ….
- **Further misconceptions:**
  - Knowledge of DP may allow to reverse
  - ….
- Focus on utility tradeoff / loss (rather than privacy gain)

Conclusions:

- Put emphasis on illustrating risk reduction
- Guidance on adequate risks per context and the implications

Karegar, F. Alaqra, A.S., Fischer-Hübner, S. Exploring {User-Suitable} Metaphors for Differentially Private Data Analyses, 18th Symposium on Usable Privacy and Security - SOUPS 2022.

Conclusions & Discussion

- Census debate & decision – important milestone
- Lessons learned: ”Anonymised” data can never be totally anonymous
- PETs can minimise risks – but come with utility trade-off & usability challenges
- Census and/vs. statistics on existing databases (see Zensus 2011, 2022)