Entropy based approach to personal data

Dr. Zoltán Alexin, PhD.
University of Szeged,
Department of Software Engineering
Árpád tér 2., Szeged, H-6720
e-mail: alexin@inf.u-szeged.hu
http://www.inf.u-szeged.hu/~alexin

Who am I

- Mathematician, PhD., information retrieval, text processing
- Since 2004, I began studying privacy issues
- Member of a Regional medical research ethics committee, 2009
- Member of the Association on Fair Data Processing, 2009
- Blogger (Facebook, www.tisztessegesadatkezeles.org)
- Has cases before Civil Courts, Hungarian Constitutional Court, European Commission on fundamental questions of medical data processing
- Achievements: excluding unsubsidized care events from the National Health Insurance Fund database, obligation of ethics approval of medical research projects without intervention, restricting the retention time of medical data at the National Health Insurance Fund
- Data protection officer at the Clinical Center of University of Szeged since 2015.
- Research on Hungarian demographic data (dataset was obtained from the Population Registry of citizens)

Content

- Motivation
- Threats of careless anonymization
- The Hungarian Population Registry data
- Conclusion

The story behind

- Itemized Medical Database (TEA, Tételes Egészségügyi Adattár), that stores accumulated health insurance accounting data from all Hungarian citizens endlessly beginning from 1998
- National Health Insurance Fund (NEAK) replaces Social Security Identifiers by a pseudonym. The interchange table is maintained by and kept endlessly by the insurance fund.
- All data items contain the date of birth, ZIP code and gender, dates, physicians, institutes, medicines
- I filed the controller of the IMD before the Constitutional Court in 2006 without success
- The law declares that the dataset is anonymous (Decree 76 of 2004 on collection and processing of medical sector data not suitable for personal identification)
- I turned to the civil court later and asked them to declare that the data is personal (not anonymous)

A data broker reckoning - Protocol

https://www.protocol.com/newsletters/policy/data-brokers-clinic-privacy?rebelltitem=10#rebelltitem10

protocol

SUBSCRIBE

NEWSLETTERS ENTERPRISE FINTECH ENTERTAINMENT WORKPLACE POLICY CHINA CLIMATE

Data brokers have always traded in our information. But following the explosive Supreme Court leak, the industry has come under fire for harvesting a specific subset of data in abortion clinic visits.

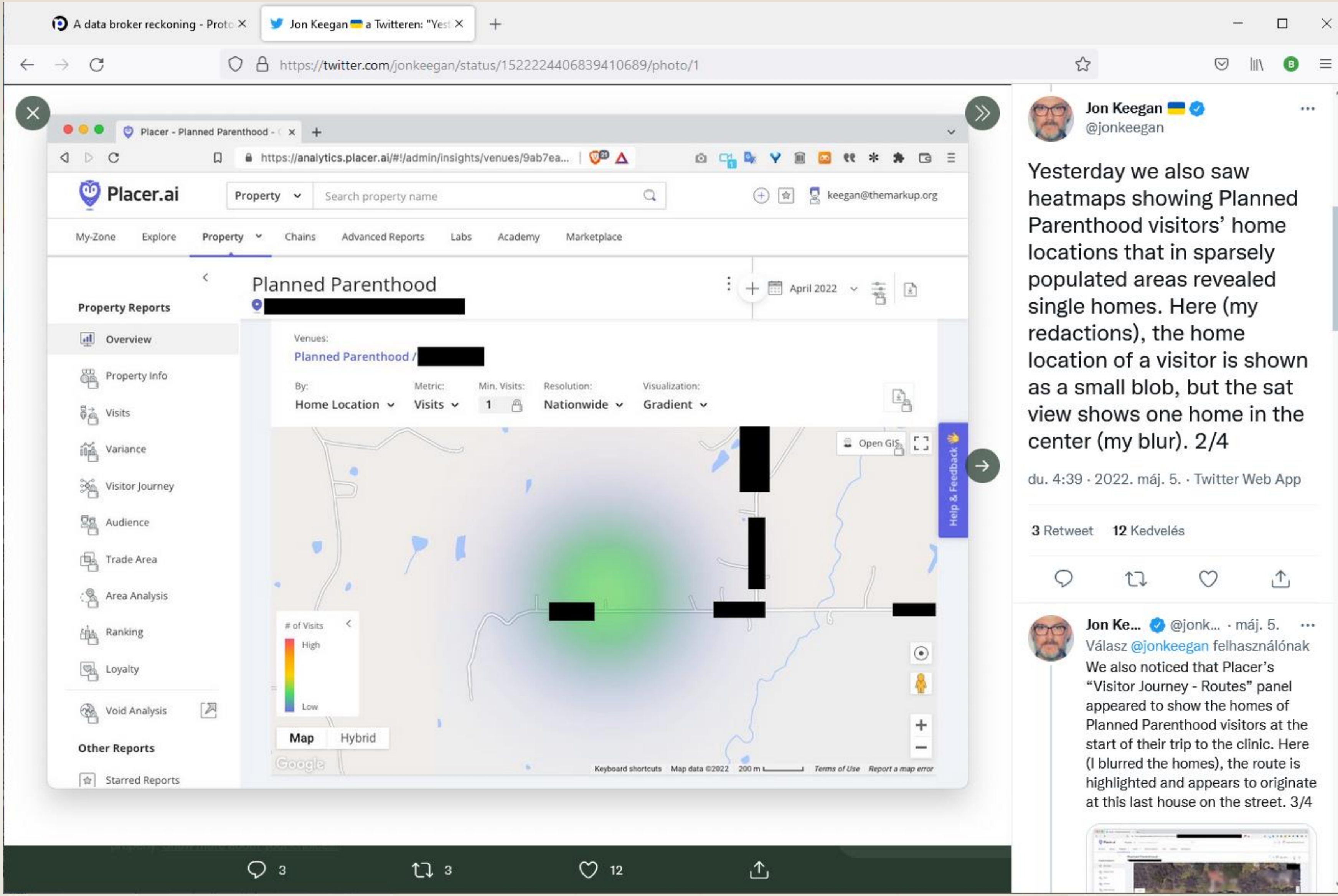
- On Tuesday, Motherboard reported that SafeGraph sold data disclosing abortion clinic visit location information. The anonymized data included timestamps and the duration of the visit.
- SafeGraph then said it would remove any data related to family-planning facilities, and that it had no evidence the data had been used maliciously.
- On Wednesday, SafeGraph CEO Auren Hoffman told Protocol, "I think it's good that we were called out" in regards to the clinic data.
- The very next day, Motherboard found another data broker, Placer.ai, offering up data that could be used to identify those visiting Planned Parenthood clinics. The company removed the sensitive data from its service after being contacted. Follow-up reporting by The Markup showed the heatmaps could be used to trace clinic visitors to specific homes, in some cases.

The companies say they weren't going out of their way to sell clinic data — but that's disturbing in its own right. Data brokers constantly collect sensitive personal information on just about everything we do.

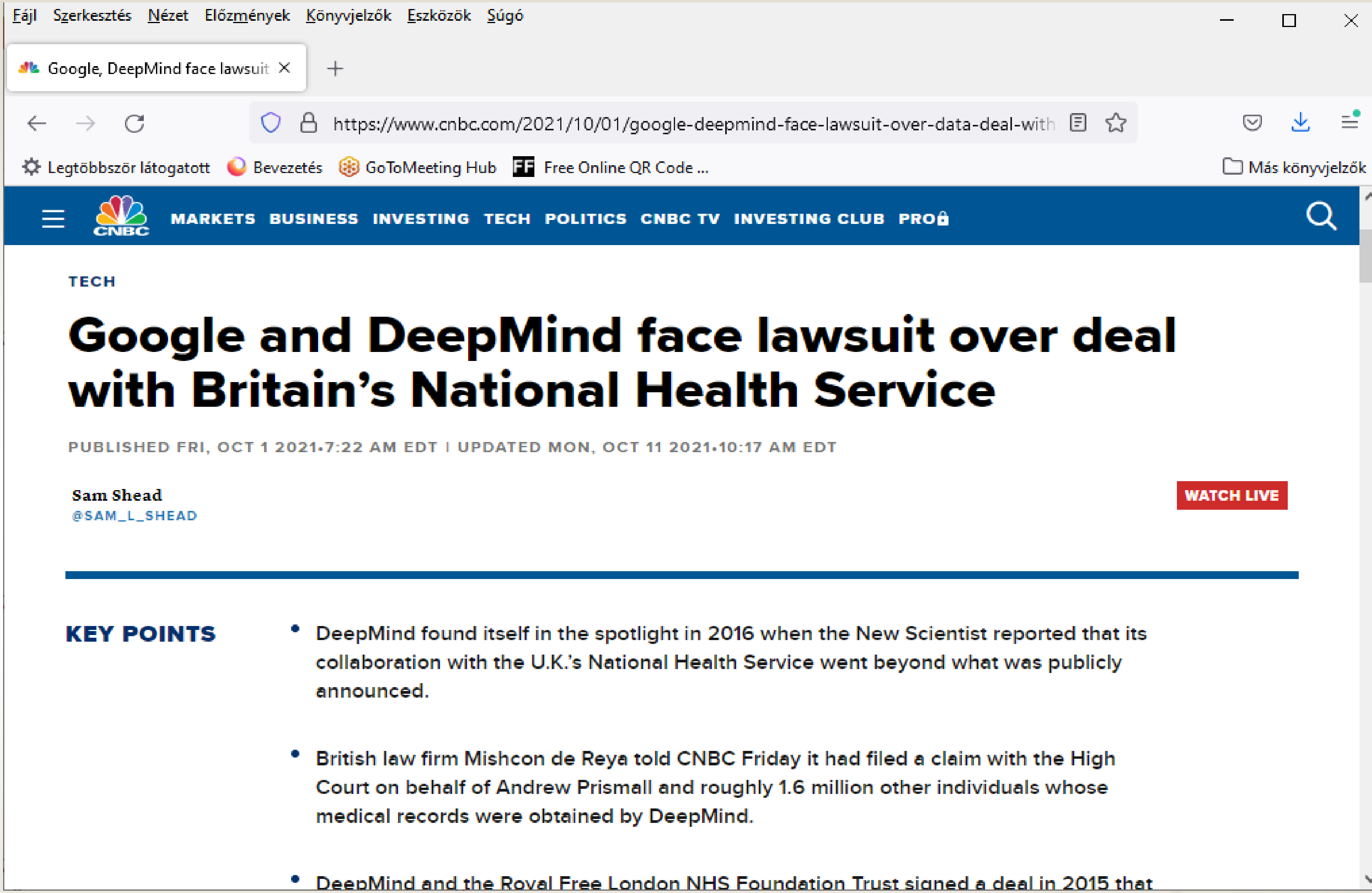
To give you the best possible experience, this site uses cookies. If you continue browsing, you accept our use of cookies. You can review our privacy policy to find out more about the cookies we use.

Accept

https://www.protocol.com/newsletters/policy/data-brokers-clinic-privacy?rebelltitem=10#rebelltitem10



https://twitter.com/jonkeegan/status/1522224406839410689/photo/1



TECH

Google and DeepMind face lawsuit over deal with Britain's National Health Service

PUBLISHED FRI, OCT 1 2021•7:22 AM EDT | UPDATED MON, OCT 11 2021•10:17 AM EDT

Sam Shead
@SAM_L_SHEAD

WATCH LIVE

KEY POINTS

- DeepMind found itself in the spotlight in 2016 when the New Scientist reported that its collaboration with the U.K.'s National Health Service went beyond what was publicly announced.
- British law firm Mishcon de Reya told CNBC Friday it had filed a claim with the High Court on behalf of Andrew Prismall and roughly 1.6 million other individuals whose medical records were obtained by DeepMind.
- DeepMind and the Royal Free London NHS Foundation Trust signed a deal in 2015 that

https://www.cnbc.com/2021/10/01/google-deepmind-face-lawsuit-over-data-deal-with-britains-nhs.html

Quasi-identifiers

- Quasi identifiers: data values that doesn’t identify an individual on its own but can become identifying in combination with other quasi identifiers.
- Quasi identifiers are not direct identifiers. Instead, they are identifiers such as an area code or zip code or date of birth. There are many people who share a zip code, and many people who share a date of birth but only few share both.
- Other words: such type of data, that an adversary can acquire together with formal identifiers like name, mother’s name etc. and can use this information to re-identify the de-identified dataset.
- A record then could be $r(q_1, q_2, q_3, q_4, q_5, \ldots, q_n, d_1, d_2, \ldots, d_m)$.
- An adversary can have $a(q_2, q_3, q_4, q_5, \text{name})$.

- The question is: what could be a quasi-identifiers? Date of birth, zip, job, gender, qualifications, schools, workplace, illness, medical operation

k-anonymity

- A dataset is called k-anonymous if for each individual there exist at least another *k* -1 distinct individuals sharing the same quasi-identifiers. This can be checked automatically by computers.
- It means, that an adversary cannot identify one single individual but at least *k* individuals (potential targets) by an attack.
- What are the acceptable values for k? Doctors say: 3, mathematicians say: 100 or 1000.
- Former Canadian data protection commissioner (Ann Cavoukian) advise 5. If an actual dataset is not at least 5-distinct then she advised additional control.

ℓ-diversity

- If we have k records with the same quasi-identifiers, but several data items are the same:
- $r_1(q_1, q_2, q_3, q_4, q_5, \ldots, q_n, d_1, d_2,$ lung cancer, $\ldots, d_m)$.
- $r_2(q_1, q_2, q_3, q_4, q_5, \ldots, q_n, d_1, d_2,$ lung cancer, $\ldots, d_m)$.
- …
- $r_k(q_1, q_2, q_3, q_4, q_5, \ldots, q_n, d_1, d_2,$ lung cancer, $\ldots, d_m)$.

- Then we need not identify anybody, still be able to derive a conclusion
- A data set is said to satisfy ℓ-diversity if, for each group of records sharing a combination of quasi-identifiers, there must be at least ℓ distinct values of the sensitive attributes.

What is entropy?

$$\underbrace{011001110110011001 1}_{\substack{\text{We know the group where}\\\text{the target people is.}\\\log_2(N/k)}}\ \underbrace{XXXX}_{\substack{\text{k indistinguishable people}\\\log_2(k)}}$$

- Entropy is a weighed sum (expected value, average) amount of bits we know about a random individual in the database.

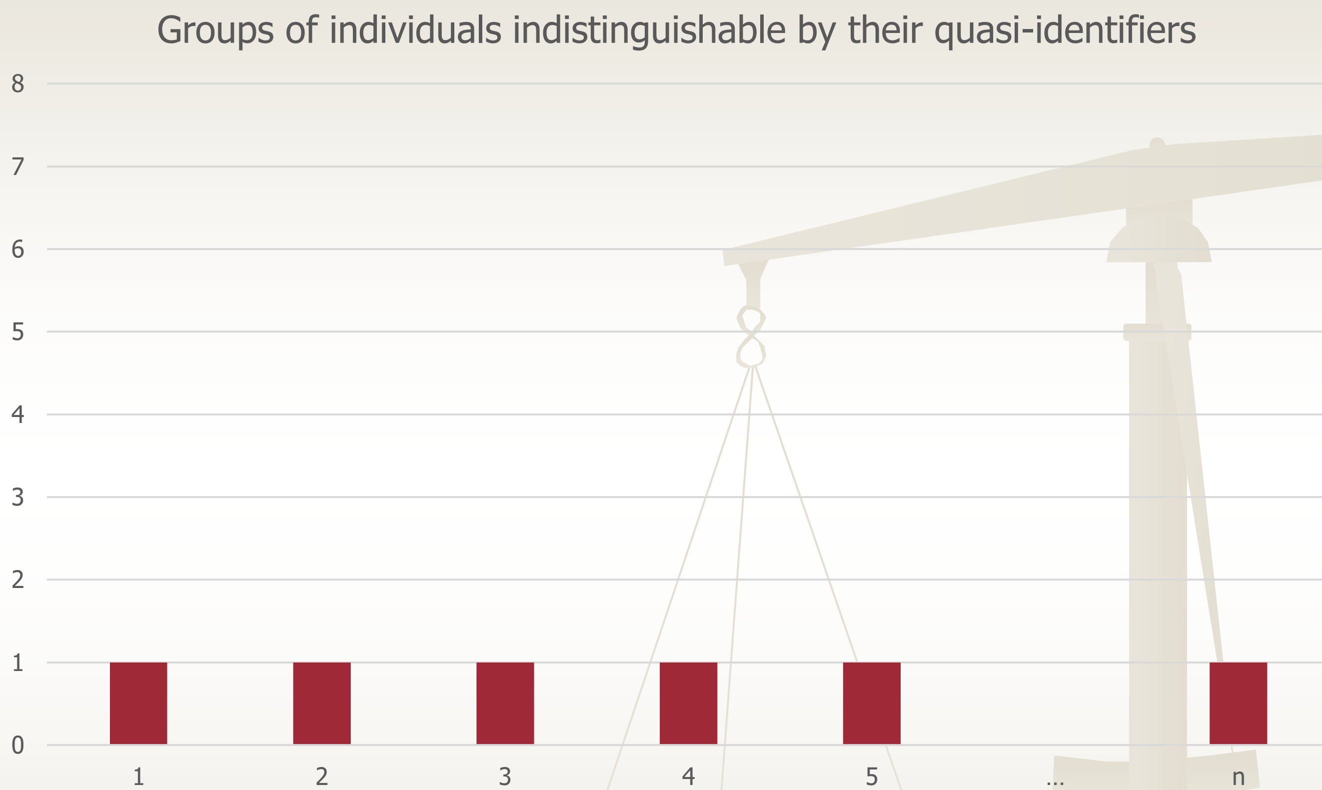
$$E(D) = \frac{1}{N}\sum_{people}\log_2\frac{N}{\#group} = \sum_{people} -\frac{1}{N}\log_2\frac{\#group}{N}$$

$$E(D) = \sum_{group} -\frac{\#group}{N}\log_2\frac{\#group}{N}$$

Entropy based approach

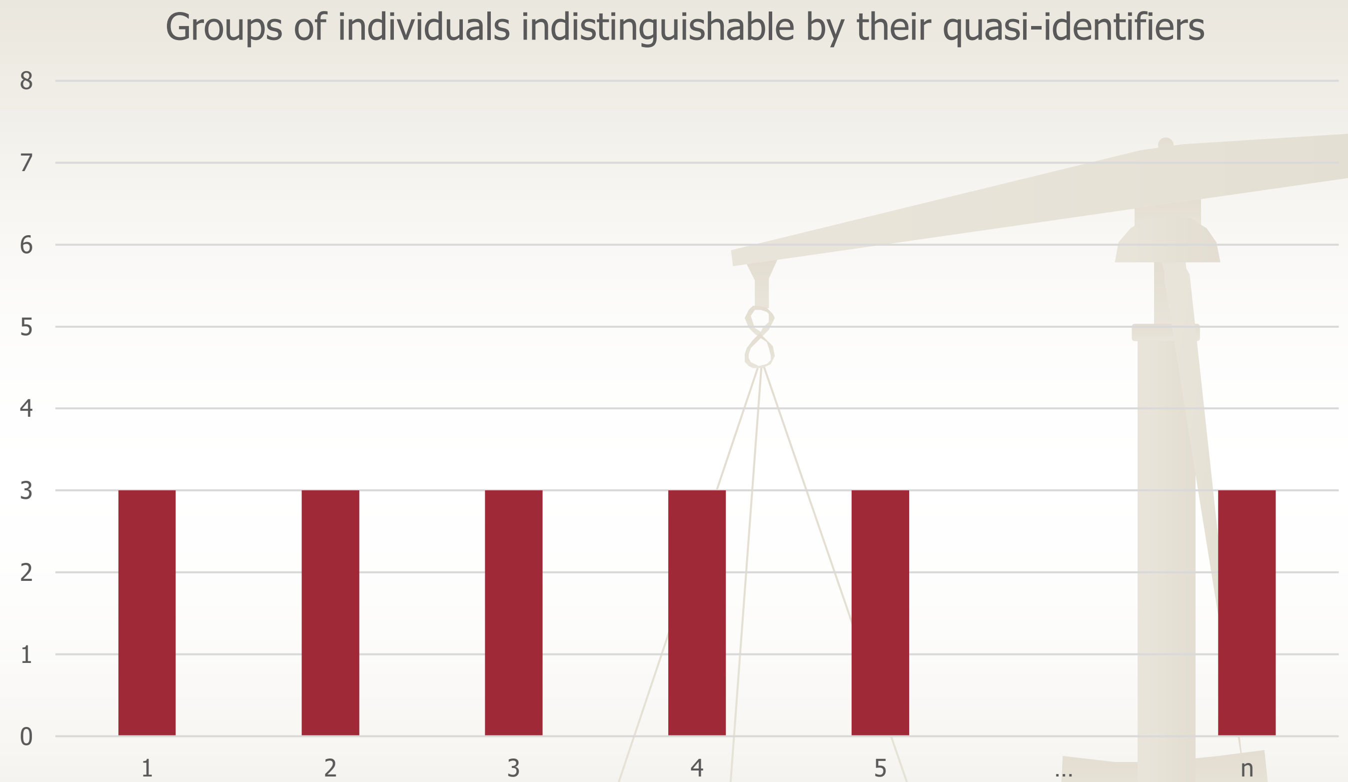
$$E(D) = \sum_{i=1,2,\dots,n}^{k_1+k_2+\dots+k_n=N} -\frac{k_i}{N}\log_2\frac{k_i}{N}$$

Entropy based approach



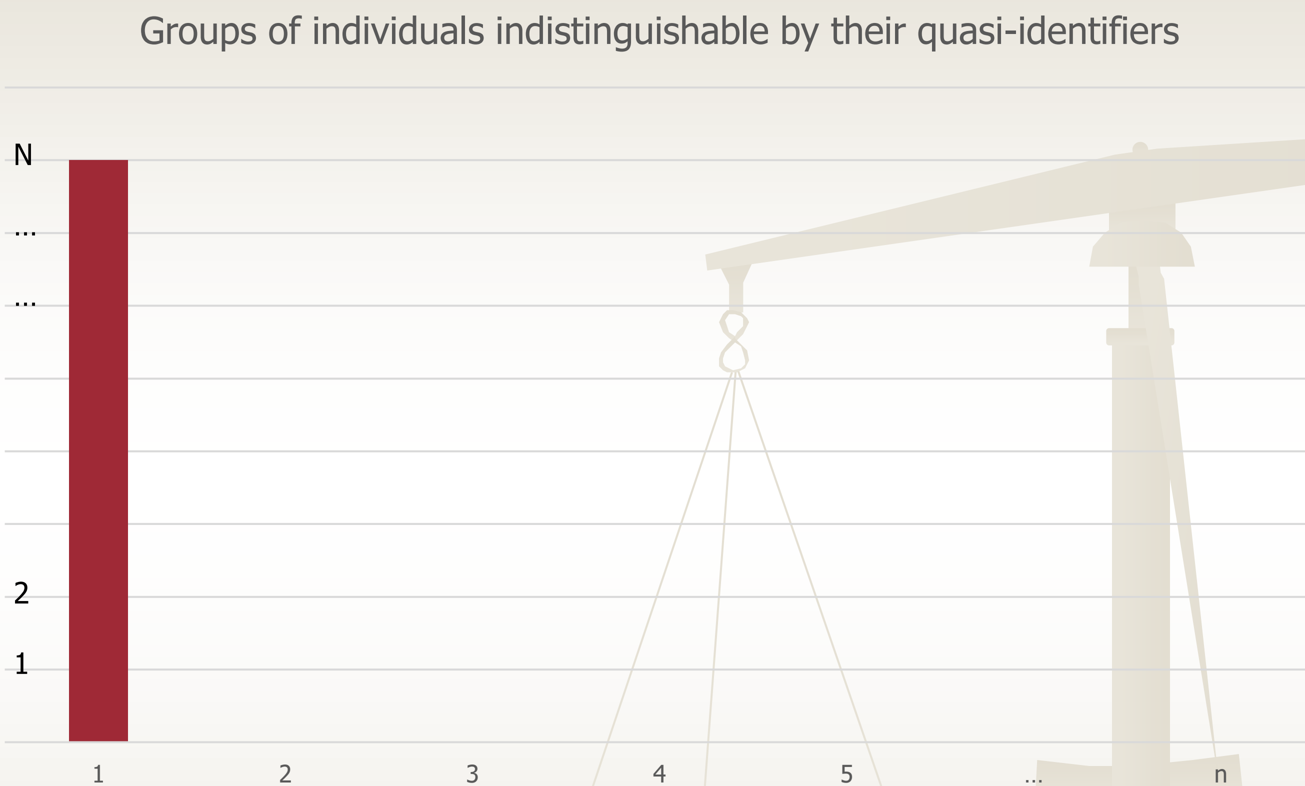
$$E(D) = \sum_{i=1,2,\ldots,n}^{1+1+\cdots+1=N} -\frac{1}{N}\log_2\frac{1}{N} = -N\frac{1}{N}\log_2\frac{1}{N} = -\log_2\frac{1}{N} = \log_2 N$$

Entropy based approach



$$E(D) = \sum_{i=1,2,\ldots,N/3}^{3+3+\cdots+3=N} -\frac{3}{N}\log_2\frac{3}{N} = -\frac{N}{3}\frac{3}{N}\log_2\frac{3}{N} = -\log_2\frac{3}{N} = \log_2 N/3$$

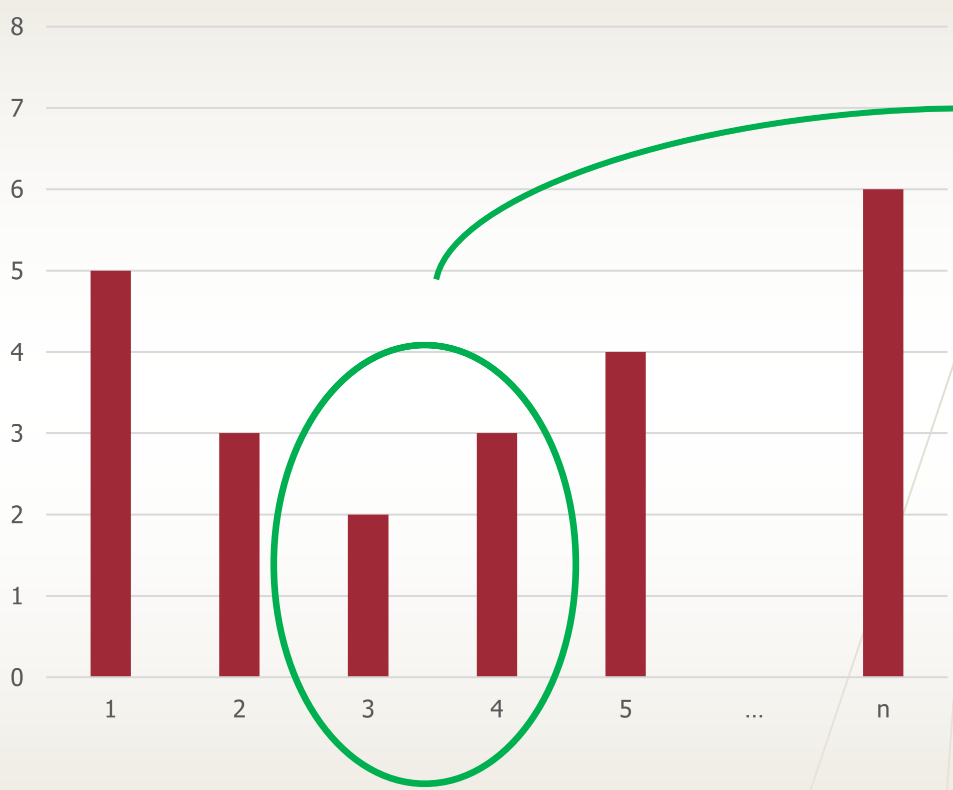
Entropy based approach



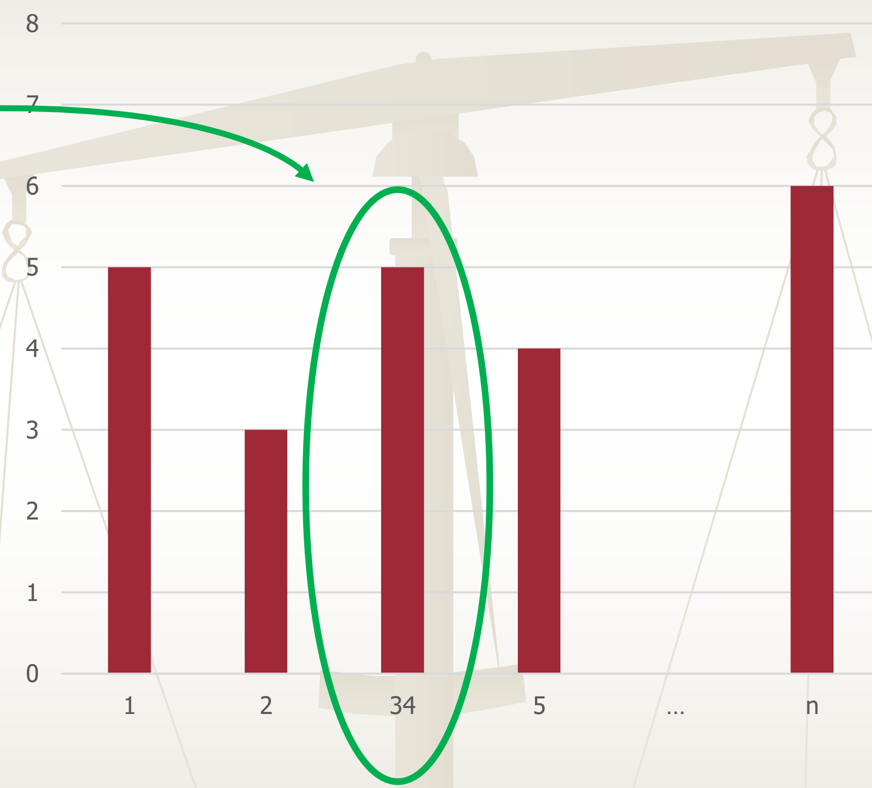
$$E(D) = \sum_{i=1,}^{\mathrm{N}=N} -\frac{N}{N}\log_2\frac{N}{N} = -\log_2 1 = 0$$

Entropy based approach

Groups of individuals indistinguishable by their quasi-identifiers (D_1)



Groups of individuals indistinguishable by their quasi-identifiers (D_2)



$$E(D_1) > E(D_2)$$

k-Anonymity

$$-\log_2 \frac{k}{N} \geq \mathrm{E(D)} = \sum_{i=1,2,\ldots,n}^{k_1+k_2+\cdots+k_n=N,\, k_i \geq k} -\frac{k_i}{N} \log_2 \frac{k_i}{N}$$

- The population of Hungary was (N=) 10 004 090 at the time of the snapshot have been taken, $\log_2(N)$ = 23.254 bits.
- The above inequality says that if a dataset is *k*-anonymous then its entropy is less than $\log_2(N/k)$.
- If we have computed the entropy then can have an estimated k

$$k = \frac{N}{2^{Entropy}}$$

Entropy of ZIP codes

ZIP code	Settlement	Population	Bits	Entropy
1011	Budapest I.	3286	11.5719	0.003800
1012	Budapest I.	4446	11.1357	0.004948
1013	Budapest I.	3404	11.5210	0.003920
…				
9982	Apátistvánfalva	589	14.0519	0.000827
9983	Szakonyfalu	769	13.6672	0.001050
9985	Felsőszölnök	589	14.0519	0.000827
Sum:		10,004,090		10.303428

The result of the computation shows that that the entropy of ZIP codes is 10.3 bits. It means that statistically, for a random citizen the expected amount of information in his/her ZIP code is 10.3 bits. It corresponds to 7916-anonymity.

Entropy of birthdate x ZIP codes

Birthdate x ZIP code	Population	Bits	Entropy
(1894.12.31., 3744)	1	23.254	2.324458e-6
…			
(1975.08.04., 9400)	4	21.254	8.498159e-6
(1975.08.04., 9407)	1	23.254	2.324458e-6
(1975.08.04., 9473)	1	23.254	2.324458e-6
(1975.08.04., 9523)	1	23.254	2.324458e-6
(1975.08.04., 9600)	1	23.254	2.324458e-6
(1975.08.04., 9700)	6	20.669	1.239640e-5
…			
Sum:	10,004,090		22.79385

Bits	Population	Ratio
23	6635838	66.33%
22	8629982	86.26%
21	9692881	96.89%
20	9996707	99.93%
19	10004090	100.00%

- The entropy is 22.7985 bits.
- It corresponds to 1.37-anonymity. This database poses substantial risk for re-identification.
- The ratio of singletons is greater the 54% of the population, in fact it was 6,635,838 individuals.

Birthdate x ZIP x gender

Birthdate x ZIP x gender	Population	Bits	Entropy
(1894.12.31., 3744, M)	1	23.254	2.324458e-6
...			
(1954.04.14., 6000, M)	1	23.254	2.324458e-6
(1954.04.14., 6041, M)	1	23.254	2.324458e-6
(1954.04.14., 6066, M)	2	22.254	4.448998e-6
(1954.04.14., 6070, F)	1	23.254	2.324458e-6
(1954.04.14., 6097, F)	1	23.254	2.324458e-6
(1954.04.14., 6097, M)	1	23.254	2.324458e-6
...			
Sum:	10,004,090		22.992721

Bits	Population	Ratio
23	7845850	78.43%
22	9403904	94.00%
21	9942428	99.38%
20	10003959	99.99%
19	10004090	100.00%

- The entropy is 22.992721 bits.
- It corresponds to 1.19-anonymity. This database poses substantial risk for re-identification.
- The ratio of singletons is greater the 74% of the population, in fact it was 7,845,850 individuals

Conclusion

- Pessimistic: medical research data is never anonymous?
- Paul Ohm: The broken promises of privacy
- Fiona Caldicott: The Information Governance Review, 2013
  - Large medical data can be processed only in a controlled and safe environment called „accredited safe havens”
- People put out uncontrollably, everything to the Internet, including very personal, sensitive information, like genetic findings, DNA fingerprints.
- Crucial point is trust and responsibility.

informed consent.

Thank you for your attention!