Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

Anonymization

Facilitating GDPR compliance for SMEs and promoting Data Protection by Design in ICT products and services *(www.bydesign-project.eu)*

ΑΡΧΗ ΠΡΟΣΤΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

ΚΕΝΤΡΟ ΕΡΕΥΝΩΝ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ

ICT abovo Information & Communication Technologies

This presentation has been based on material provided by Dr. K. Limniotis (HDPA)

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Personal and anonymous data Definitions (GDPR)

- The term "personal data" refers to any information relating to an identified or identifiable natural person
- The data protection principles do not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person
  - However, to determine whether a natural person is identifiable, account should be taken **of all the means reasonably likely to be used**, such as singling out, by any person to identify – directly or indirectly – the natural person
    - Objective factors, such as the costs of and the amount of time required for identification, should be taken into account
  - *In simple words, we should be very careful when characterizing data as anonymous data*
  - *Have we thoroughly examined whether identification is practically fully impossible?*
    - *Identification in which context?*

The famous AOL incident (2006)

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

August 2006: research.aol.com

AOL is embarking on a new direction for its business making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community. To get started, we invite you to visit us at http://research.aol.com, where you will find:

- *…*
- ***Query streams for 500,000 users over 3 months (20 million queries)***
- *….*
- A random ID was associated to each user
  - The same (meaningless) ID, for the same user
- However, a combination of the published information with other available data could allow identification!

The user #4417749

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times — Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERALS

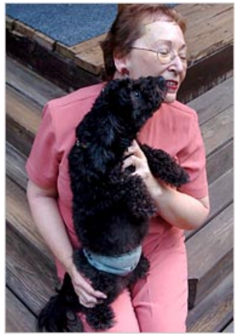
A Face Is Exposed for AOL Searcher No. 4417749

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

SIGN IN TO E-MAIL THIS | PRINT | SINGLE PAGE | REPRINTS | SAVE

ARTICLE TOOLS SPONSORED BY THE HISTORY BOYS

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



Erik S. Lesser for The New York Times

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

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

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

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

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

Multimedia

Graphic: What Revealing Search Data Reveals

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it

- The characterization of anonymous data is not an easy task
- Simply removing "obvious identifiers" is not adequate
- In other words, the notions of identifiers or "identifying data" is wide
  - Identifier in which context?

The notion of pseudonymisation

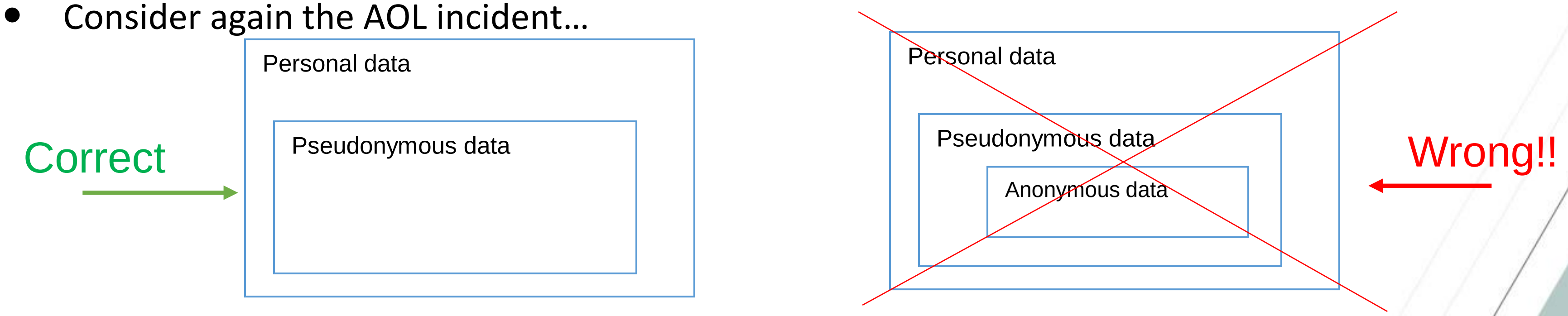
Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

- According to ISO/TS 25237:2017 standard:
- *"Pseudonymisation is a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms"*
- *De-identification is a general term for any process of reducing the association between a set of identifying data and the data subject.*
- *A pseudonym a personal identifier that is different from the normally used personal identifier and is used with pseudonymized data to provide dataset coherence linking all the information about a data subject, without disclosing the real world person identity'.*
    - As a note to the latter definition, it is stated in ISO/TS 25237:2017 that pseudonyms are usually restricted to mean an identifier that does not allow the direct derivation of the normal personal identifier. They can either be derived from the normally used personal identifier in a reversible or irreversible way or be totally unrelated.

The notion of pseudonymisation in the GDPR

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

- "Pseudonymisation" means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific person without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person
- Personal data which have undergone pseudonymisation <u>should be considered to be information on an identifiable natural person</u>.
    - That is pseudonymization does not result in anonymous data
        - Additional information to allow re-identification does exist (somewhere…)
        - Consider again the AOL incident…



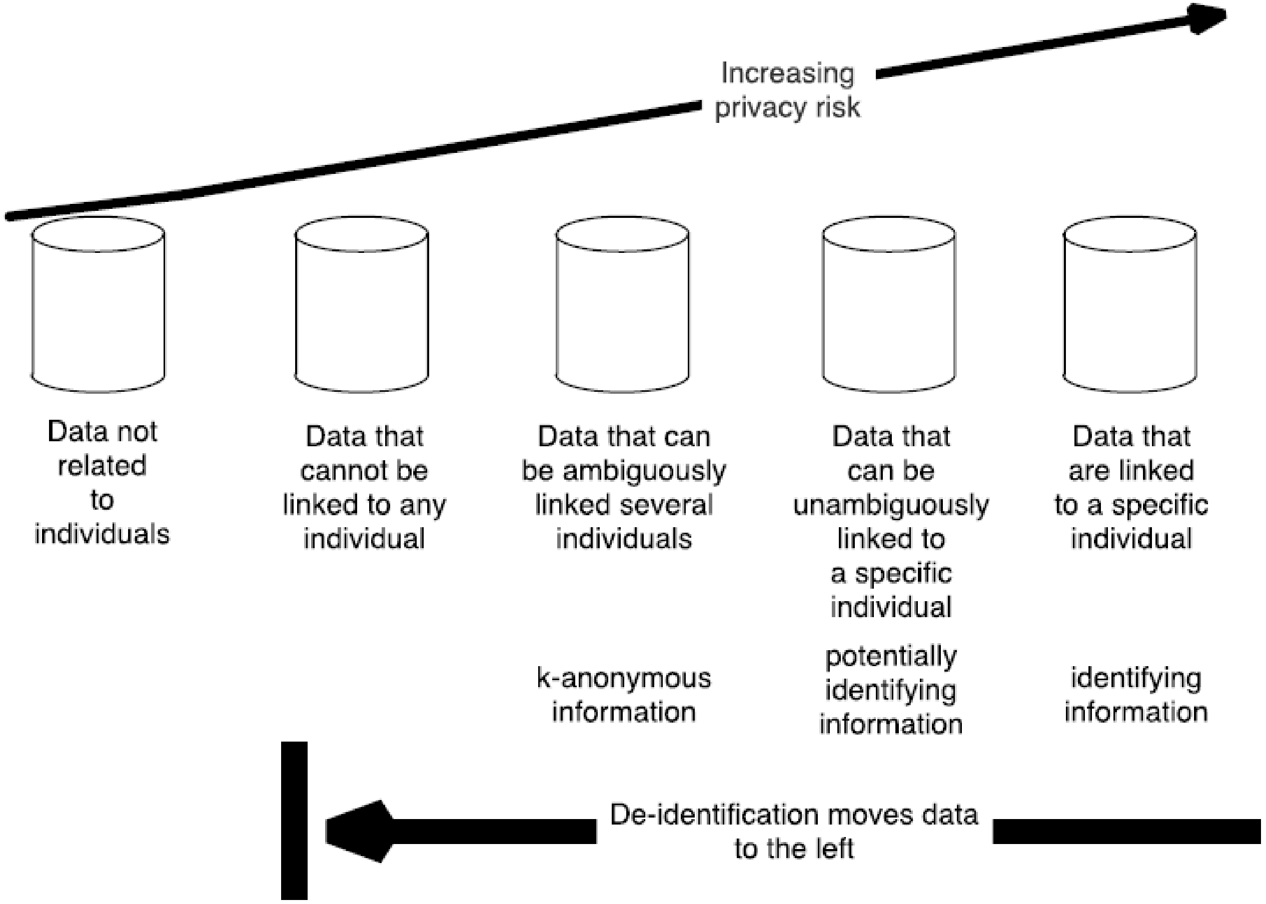
Benefits of pseudonymisation on personal data protection

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

- The GDPR makes about 15 references to pseudonymisation
  - Possible appropriate safeguard for:
    - *"purpose limitation balancing test" (art. 6, par. 4)*
    - *Data protection by design and by default (art. 25)*
    - *Security of processing (art. 32).*
    - *Processing of personal data for public interest, scientific or historical research purposes or statistical purposes (art. 89)*
- Pseudonymisation is also implied in several other places within GDPR
  - When the controller is able to demonstrate that is not in a position to identify the individual (data subject), Art. 15-20 shall not apply – i.e. right of access, right to rectification/erasure/restriction/portability *(art. 11)*
    - *Unless the data subject provides <u>additional information enabling his/her identification</u>*
  - Appropriately-implemented pseudonymisation can reduce the likelihood of individuals being identified in the event of a personal data breach

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

«Phases» of Anonymization



S. L. Garfinkel, “**De-Identification of Personal Information”,** NIST Internal Report 8053, 2015

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

When a Person can be Identified?

- In addition to the identifiers, there are the quasi-identifiers which when combined can lead to the identification of a person!

Identifier	Quasi-identifier			Sensitive attribute
Name	**DOB**	**Gender**	**Zipcode**	**Disease**
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Removal of Identifiers cannot guarantee anonymity

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

An example of «Bad Anonymization»

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

- Assume that a Hospital provides the above “anonymized” table (after removal of all data that could lead to the identification of a person (Name, ID number, VAT number, Social security number etc).

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example of «Bad Anonymization»

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Source: B. Fung et.al., Privacy-Preserving Data Publishing: A Survey of Recent Developments, ACM Computing Surveys, 2010

- Assume that somebody knows that the list provided by the Hospital includes some specific persons (e.g. residents of a small village)
- For these persons data can be easily found from publicly available sources (Table b)
- By combining the two Tables we can identify some persons
    - E.g. (Job, Sex, Age) = (Laywer, Male, 38) reveals that Doug suffers form HIV

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

Addressing the Problem – «Generalization»

- To avoid this type of attacks we can appropriately modify the values of quasi-identifiers, through generalization:
  - E.g. we do not release the precise age but, instead, an age range (for instance 30-40)
    - The greater the Generalization the better the anonymity, although we may miss useful information
    - The aim is to achieve the best possible anonymization with the least possible loss of information

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

«Generalizing» the previous table

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

«Generalization»

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

??

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Generalization Criteria

- k-anonymity – (Samarati-Sweeney, 1998):

  In an anonymous table the number of records with the same quasi-identifiers values is at least k
- Clearly, the bigger k is, the better the anonymity
- For the previous example: Anonymous with k = 3

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Alternative Approach

- Suppression: Some fields or entire records are deleted

#	**Zip**	**Age**	**Nationality**	**Condition**
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Flu

Generalization Suppression

- Maximum Generalization is equivalent to Suppression

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Is k-anonymity enough ?

- Let us assume the following:

	Zip	Age	National
Bob →	13053	31	American
Akira →	13068	21	Japanese

- and that someone makes public the following data:

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Data Set

	Non-Sensitive Data			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	HIV
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	HIV
10	13053	37	Indian	HIV
11	13068	36	Japanese	HIV
12	13068	35	American	HIV

k-anonymity with k=4

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*HIV*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485**	*> = 40*	*	*Viral Infection*
9	130**	3*	*	HIV
10	130**	3*	*	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Akira

Bob

k-anonymity with k=4

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

	Non-Sensitive Data			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*HIV*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485*[illegible]	[illegible]	[illegible]	*Viral Infection*
9	130*[illegible]	[illegible]	[illegible]	HIV
10	130*[illegible]	[illegible]	[illegible]	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Akira

Bob

Bob has HIV!!

byDesign

k-anonymity with k=4

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

	Non-Sensitive Data			*Sensitive Data*
#	[illegible]	[illegible]	[illegible]	**Condition**
1	130*	[illegible]	[illegible]	Heart Disease
2	130*	[illegible]	[illegible]	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*HIV*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485*	[illegible]	[illegible]	*Viral Infection*
9	130*	[illegible]	[illegible]	HIV
10	130*	[illegible]	[illegible]	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Akira

Bob

If we know that heart diseases are extremely rare in Japan, then it is highly likely that Akira has been infected by a virus

Bob has HIV!!

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Anonymity with l-diversity

- The total number of non-distinct records (have same QID values) form an equivalence class
- Distinct l-diversity (Machanavajjhala et al., 2006): Every equivalence class should include at least l distinct values of the sensitive field.

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Distinct 3-diversity

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	1305*	<= 40	*	Heart Disease
2	1305*	<= 40	*	Viral Infection
3	1305*	<= 40	*	HIV
4	1305*	<= 40	*	HIV
5	*1485**	*>= 40*	*	*HIV*
6	*1485**	*>= 40*	*	*Heart Disease*
7	*1485**	*>= 40*	*	*Viral Infection*
8	*1485**	*>= 40*	*	*Viral Infection*
9	1306*	<= 40	*	Heart Disease
10	1306*	<= 40	*	Viral Infection
11	1306*	<= 40	*	HIV
12	1306*	<= 40	*	HIV

Bob and Akira Belong here

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Is Distinct l-Diversity enough ?

- Probabilistic inference attacks are still possible

...	**Disease**
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

10 records - Equivalence class

8 out of 10 have HIV

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Anonymization Tools

- ARX (https://arx.deidentifier.org/)
- Amnesia (https://amnesia.openaire.eu/)
- UTD Anonymisation toolbox (http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home)
- Anonimatron (https://realrolfje.github.io/anonimatron/)