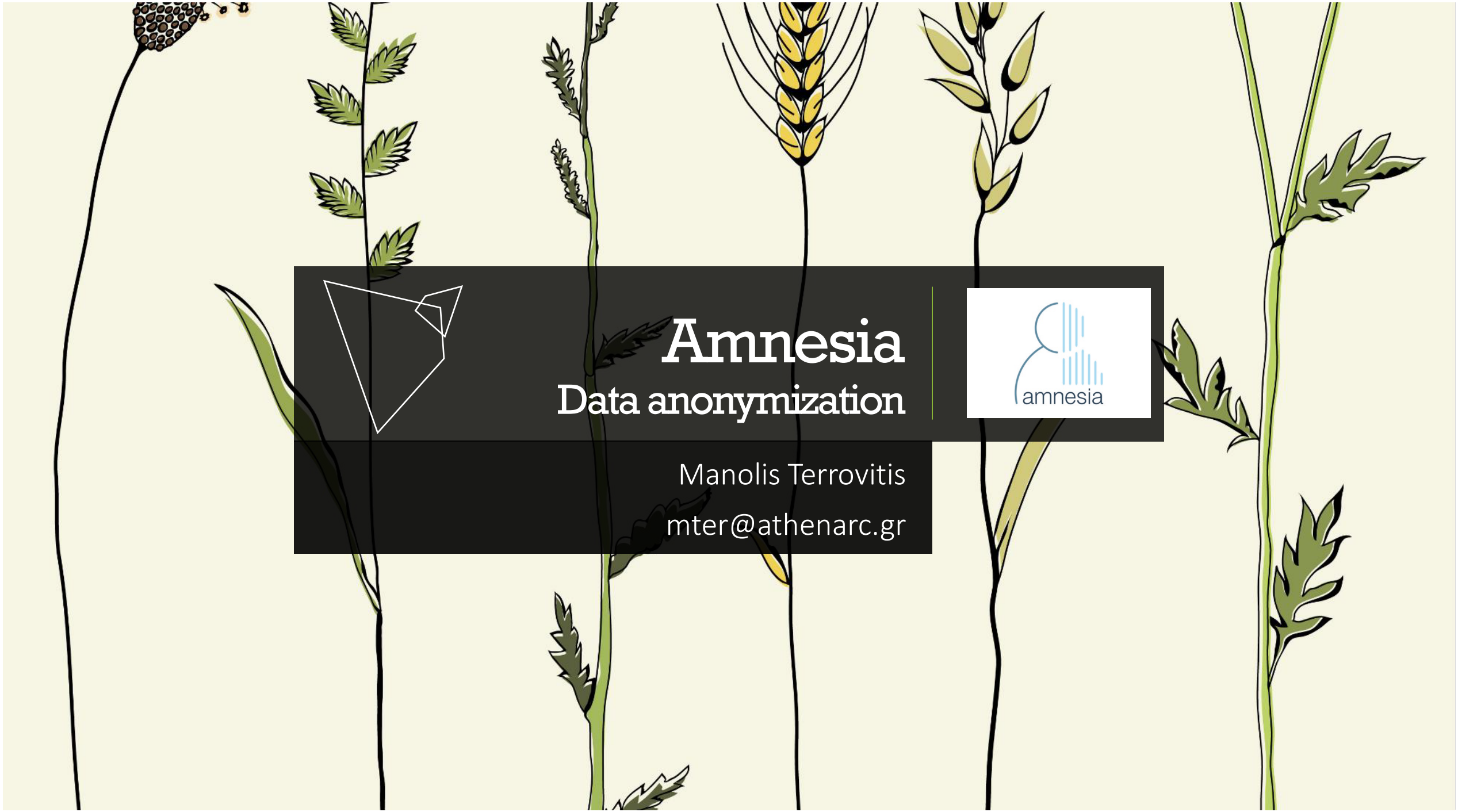


Amnesia

Data anonymization

Manolis Terrovitis

mter@athenarc.gr

- GDPR limits the usage of personal data
  - according to law and contracts
  - Consent
  - Can be used for research

- Using Personal data
  - Consent might not be given or withdrawn
  - Difficult to manage
  - Usage for research purposes requires strict internal processes
  - Cannot share with third parties

Unlock the information

- Research and studies need statistical information and properties
- Personal identification is not necessary in most fields
- Low reduction in data quality is tolerable
  - Or can be mitigated by using larger amounts of data

- Anonymization unlocks the valuable information in data
  - The anonymized data are **different from the original data**
  - Anonymization is a one-way transformation of data
  - Original data cannot be retrieved

- Pseudonymization is not Anonymization
  - In Pseudonymized data there is a way to retrieve the original data
  - Pseudonymized data are still personal data

Why Anonymize?

- Anonymized data are outside the scope of GDPR
- Anonymization provides a statistical guaranty about the risk of information leakage
- It is the most suitable way to give information to third parties, without revealing personal data

Limitations of Anonymization

Anonymized data have lost some information

- The key idea of a good anonymization algorithm is to minimize this loss and limit it in the least important information

There are gray boundaries between anonymized and pseudonymized data

Formal privacy guaranties provide a statistical guaranty for the anonymized data

- This is only an interpretation of the notion of “privacy”

It cannot easily be fully automated

- User input is needed

When to anonymize

- When you are a practitioner, and you want to share data with researchers and third parties without compromising the privacy of the user
  - After the data is anonymized, you do not need consent
- When you want to give data to recipients you do not fully trust
  - Encryption will reduce the risks of data leaks to unauthorized third parties, it will do nothing for untrusted recipients
- When you want to openly publish data and you are not fully aware of the audience
- When reduction in information quality is acceptable

Why Amnesia

- User friendly
- Works locally, no data transfer risk
- Allows users to customize the solution
- The only tool to offer anonymization for set-valued data
- The only tool to support k^m-anonymity
- Easy to incorporate to third party information systems

Statistics

https://amnesia.openaire.eu

- 110k visitors
- 415k page views
- 7.3k downloads

Status

amnesia

K-anonymity

Km-anonymity

Object relational datasets

Disk based algorithm

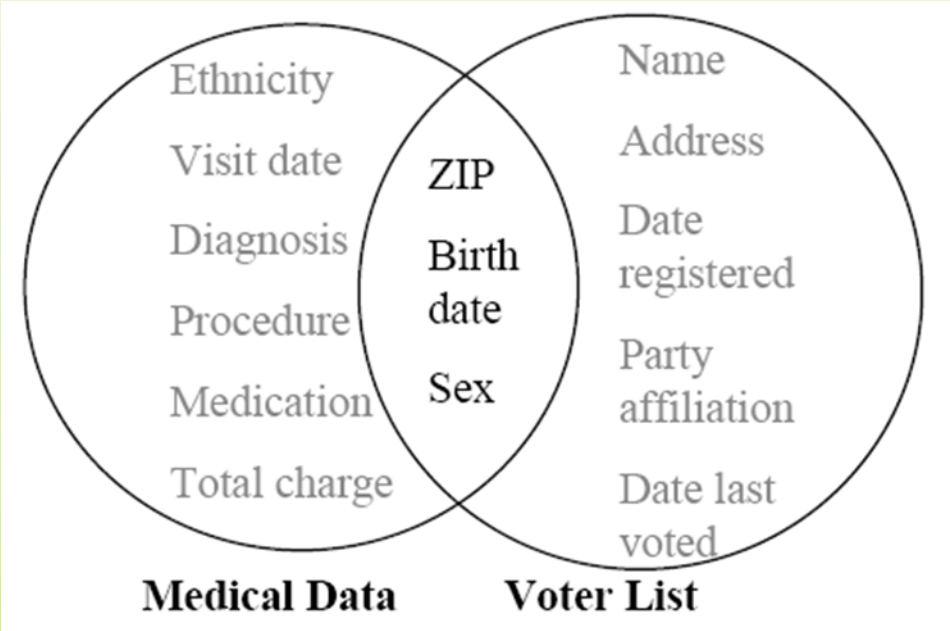
API

ReST and command line API exist to help programmers

Bugs have diminished

Queries in helpdesk are less about bugs these days

Pseudonymization - Link attacks



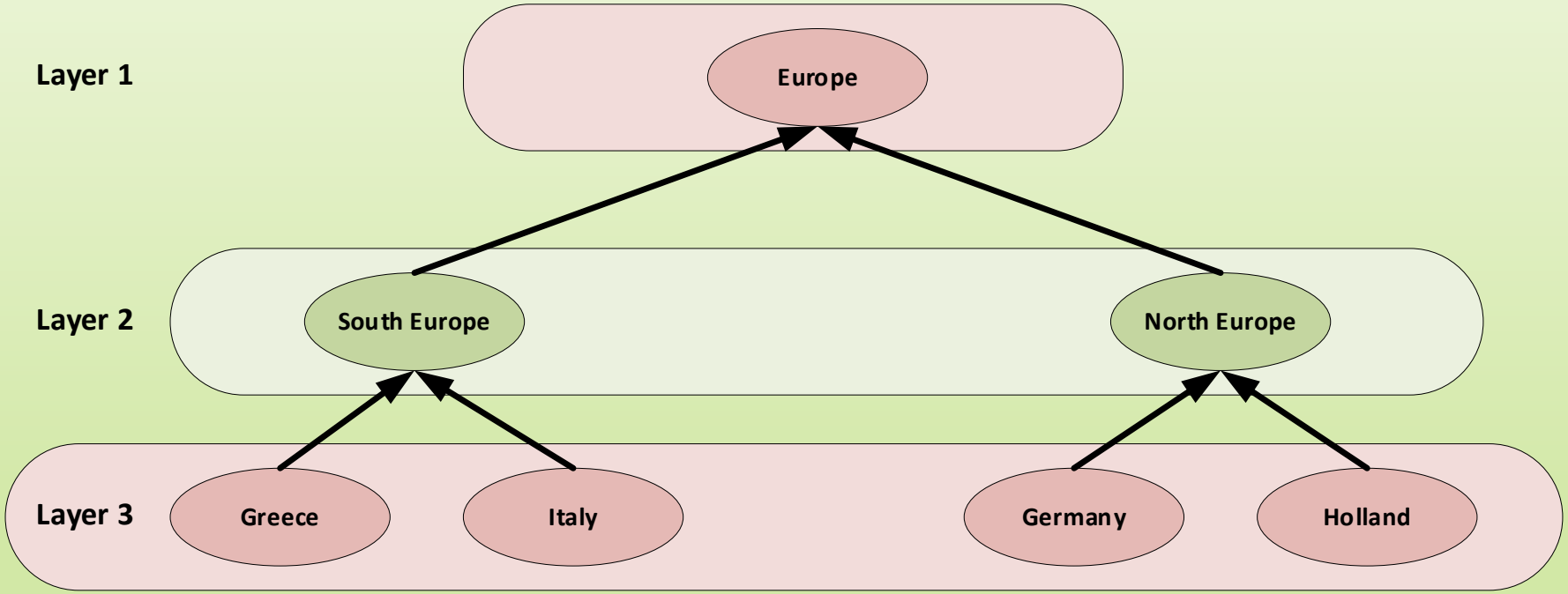
k-anonymity

- Each entry becomes indistinguishable from other *k*-1 entries

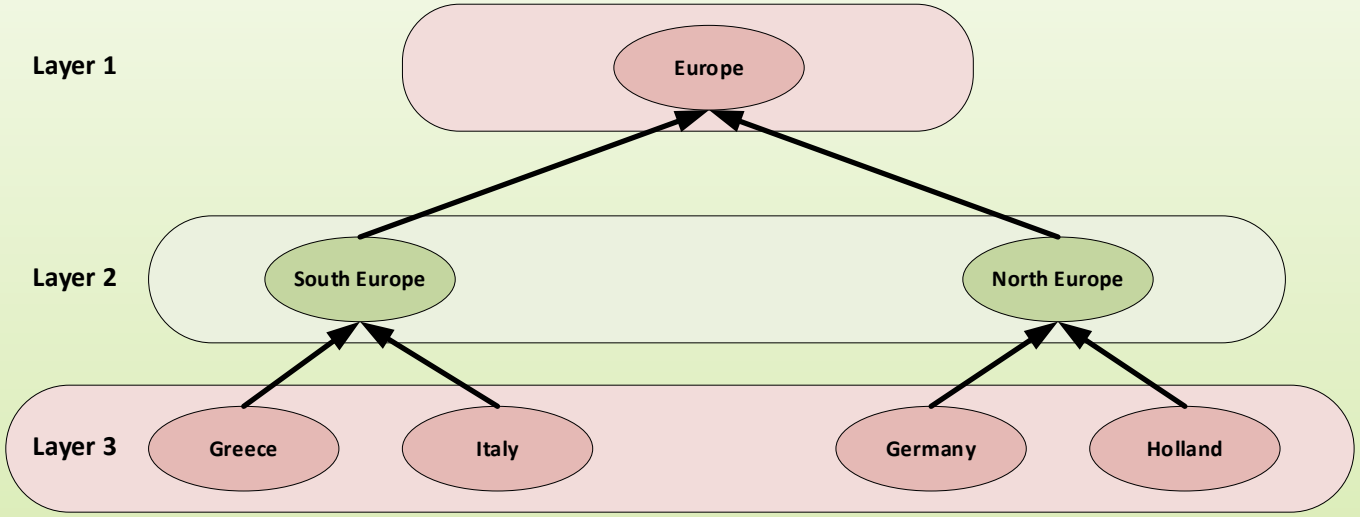
id	Zipcode	Age	National.	Disease
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

~~id~~	Zipcode	Age	National.	Disease
~~1~~	130**	<30	*	Heart Disease
~~2~~	130**	<30	*	Heart Disease
~~3~~	130**	<30	*	Viral Infection
~~4~~	130**	<30	*	Viral Infection
~~5~~	1485*	≥40	*	Cancer
~~6~~	1485*	≥40	*	Heart Disease
~~7~~	1485*	≥40	*	Viral Infection
~~8~~	1485*	≥40	*	Viral Infection
~~9~~	130**	3*	*	Cancer
~~10~~	130**	3*	*	Cancer
~~11~~	130**	3*	*	Cancer
~~12~~	130**	3*	*	Cancer

Data transformation – Full domain generalization



Full domain



Age	Place
28	Greece
33	Italy
21	Greece
24	Greece
37	Germany
36	Holland
35	Germany

Age	Place
[25-35]	South Europe
[25-35]	South Europe
[20-25]	South Europe
[20-25]	South Europe
[35-40]	North Europe
[35-40]	North Europe
[35-40]	North Europe

Local Recoding

- Original values are not globally generalized but only some of subsets of them

Age	Place
28	Greece
33	Italy
21	Greece
24	Greece
37	Germany
36	Holland
35	Germany

Age	Place
[25-35]	South Europe
[25-35]	South Europe
[20-25]	Greece
[20-25]	Greece
[35-40]	North Europe
[35-40]	North Europe
[35-40]	North Europe

k^m-anonymity

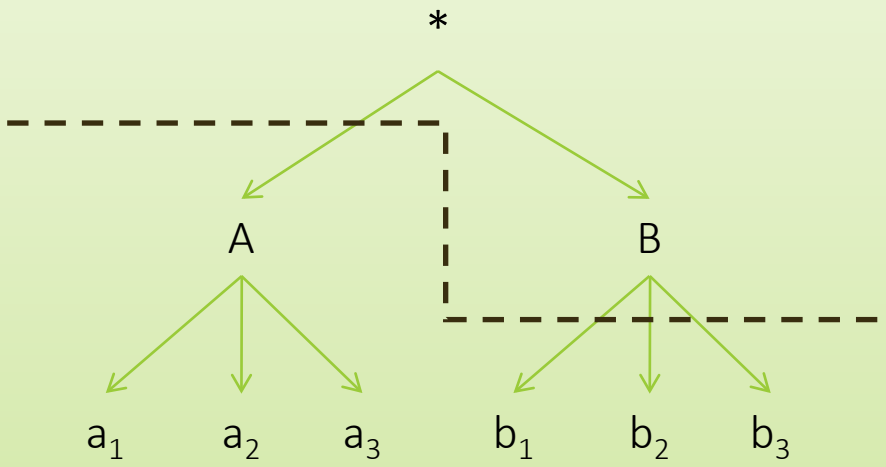
	Fruits	Meat	Vegetables	Fish
Vassilis	X	X		
Manolis	X	X	X	
Eleni			X	
Maria		X	X	
Kostas	X			X

	Fruits	Meat	Other food
Vassilis	X	X	
Manolis	X	X	X
Eleni			X
Maria		X	X
Kostas	X		X

- 2^2-anonymous
- Any combination of *m* items will not appear less than *k* times

Your Logo or Name Here

Method Description



$c_3 = \{A, b_1, b_2, b_3\}$

Amnesia Limitations

- Users are not familiar with anonymization techniques
- The process is novel and requires effort from the user’s part
- Amnesia cannot decide on privacy parameters
- K-anonymity does not protect from every type of attack

Thank You

Manolis Terrovitis

mter@athenarc.gr

https://amnesia.openaire.eu