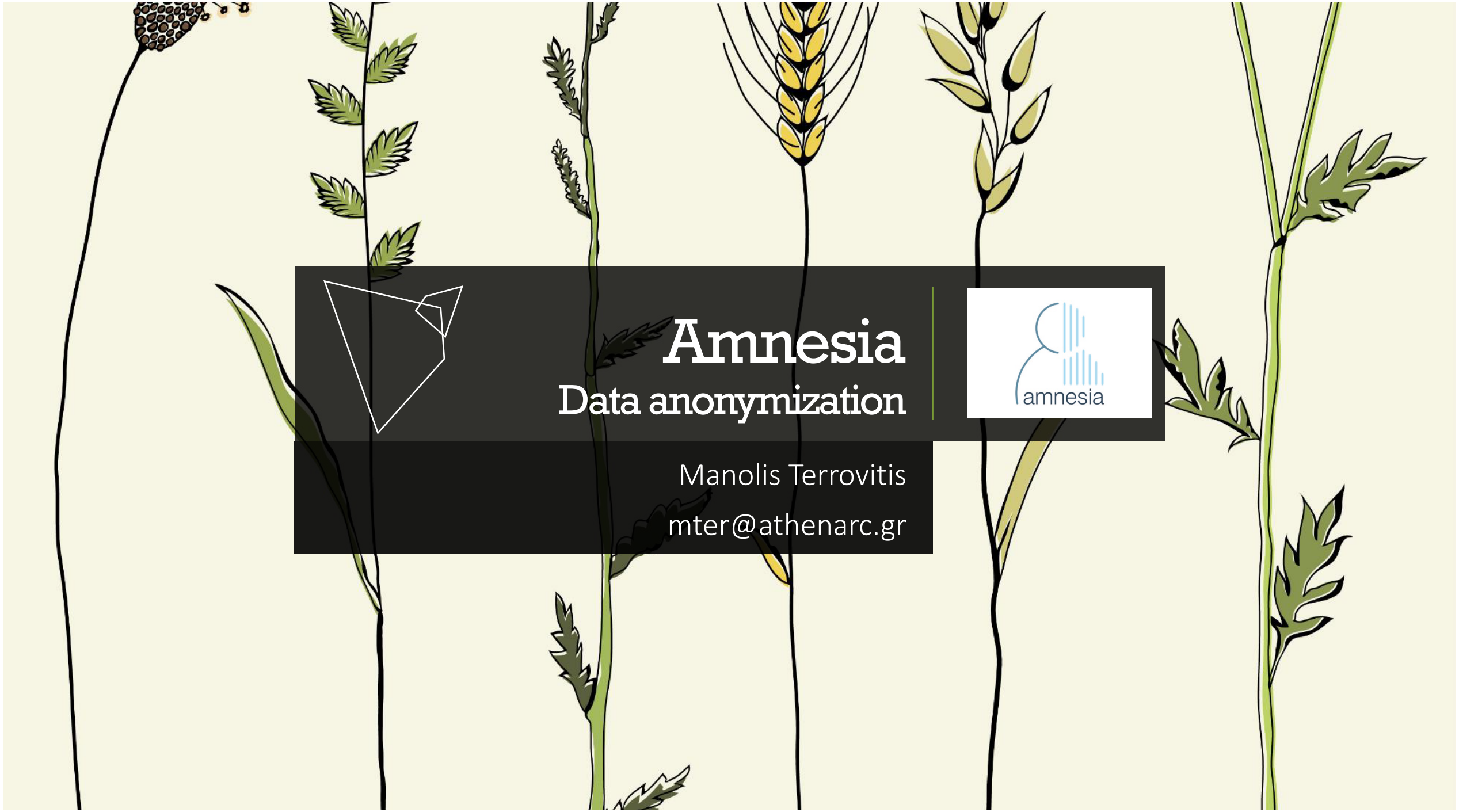


Amnesia

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  - Consent might refused or withdrawn
  - Difficult to manage
  - Usage for research purposes requires strict internal processes
  - Cannot share with third parties

Unlock the information

- Research requires 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

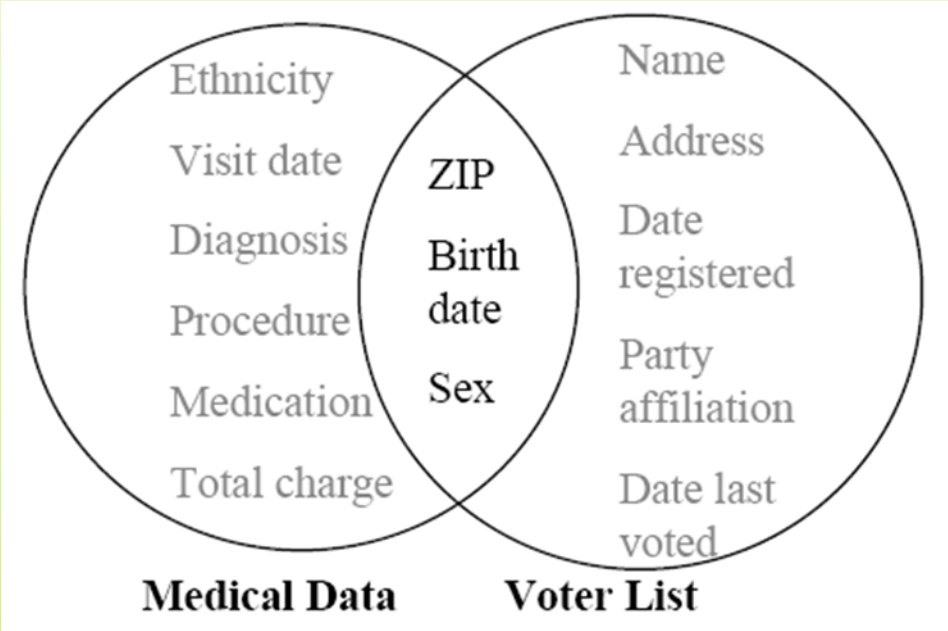
Why Anonymize?

- 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

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

Pseudonymization - Link attacks



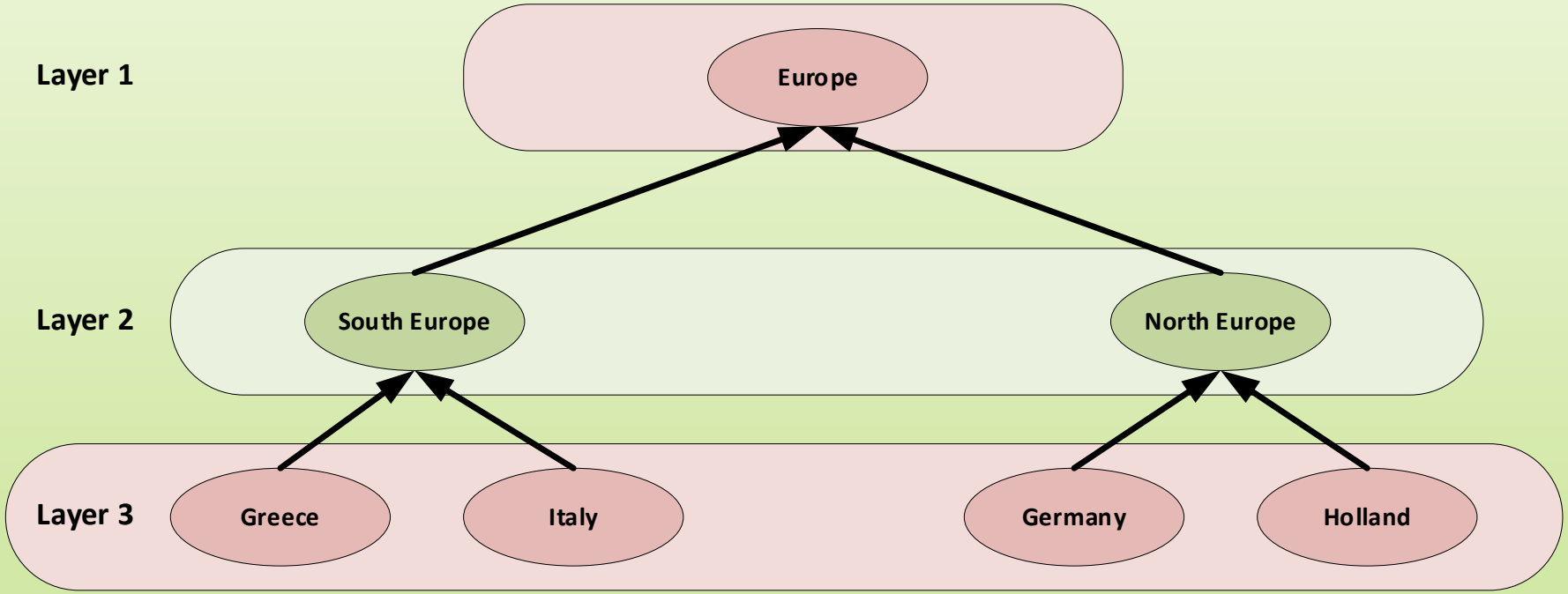
k-anonymity

- 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



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

Status

K-anonymity

Km-anonymity

Object relational datasets

Disk based algorithm

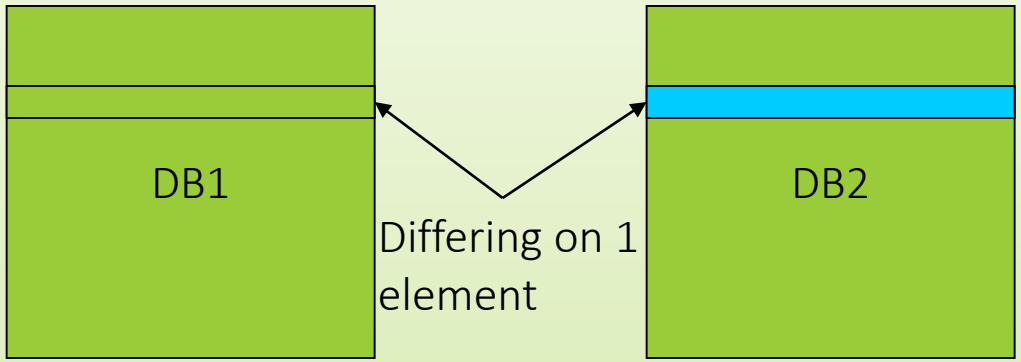
API

ReST and command line API exist to help programmers

Bugs have diminished

Queries in helpdesk are less about bugs these days

Amnesia next steps



$K: \mathcal{D} \rightarrow \mathcal{R}$ ensures ε-DP if for all adjacent datasets D_1, D_2 and for all subsets S of $\mathcal{R}$:

$$\frac{\Pr[K(D1) \in S]}{\Pr[K(D2) \in S]} \leq e^{\varepsilon}$$

Adding differential privacy

- Stricter, more generic guarantee
- Difficult for non-expert users

Differential privacy relies on noise addition to mechanisms

- It guarantees that the impact of each record in an algorithm's output is limited
- It is defined on mechanisms, i.e., algorithms, not on data

Amnesia will allow differentially private data synthesis

- Uses will define histograms and aggregates
- Amnesia will create differentially private histograms and will recreate a dataset that supports the DP histogram

Anonymized vs Synthetic data

- Synthetic data protect user privacy, but they do not provide a guarantee unless coupled with anonymization mechanisms, e.g. differential privacy
- Anonymized data provide a limit in information loss
- Synthetic data need big training sets to perform well
- Anonymized data need a large input dataset to perform well
- Amnesia is a mature tool and has been widely tested

- I expect that anonymization will perform better for low cardinality datasets and synthetic data generators for high dimensional 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 pseudo-anonymized 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

Thank You

Manolis Terrovitis

mter@athenarc.gr

https://amnesia.openaire.eu