Machine Learning for k-Anonymization

(of Patient EHR Data)

Bernd Malle
b.malle@hci-kdd.org

Holzinger Group - www.hci-kdd.org

Overview / Contents

- 1. Introduction & Motivation
- 2. Properties of data & General approach
- 3. Limits of anonymization
- 4. Input data formats
- 5. An iML approach to k-anonymization
- 6. Assignment(s) – GenHierarchies & CostFunctions

- Public release of sensitive information is useful for
  - Statistics => education, grant proposals ;-)
  - Research => prediction of disease spreading etc.

- However, personal identities need to be concealed

- In the past, simple approaches have failed to provide sufficient security:
  - data linkage of publicly available datasets
    - Netflix database, which was linked with the IMDB movie ratings database (via date of rating) => at least one user was re-identified

Re-Identifying the NYC Taxi Ride Dataset

1. Find suspicious data
2. Figure out what ONE hash represents ('0')
3. Figure out input domain for hashes
   => Medallions are 4-5 digits
   => ~20M possibilities
4. Construct inverted LUT
5. !! Whole DS hacked !!

We need robust anonymization techniques

Properties & General Approach

Data properties => Reduce granularity

Name	Age	Zip	Gender	Disease
Alex	25	41076	Male	Allergies
...	...	...	...	...

- Identifiers := immediately reveal identity
  - name, email, phone nr., SSN

  => DELETE

- Sensitive data
  - medical diagnosis, symptoms, drug intake, income

  => NECESSARY, KEEP

- Quasi-Identifiers := used in combination to retrieve identity
  - Age, zip, gender, race, profession, education

  => MAYBE USEFUL

  => MANIPULATE / GENERALIZE

Trade-off between:

- Data utility => min. information loss
- Privacy => max. information loss

Both can be easily achieved (but not together ☺)

Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer

⇔

Node	Age	Zip	Gender	Disease
X1	*	*	*	Allergies
X2	*	*	*	Allergies
X3	*	*	*	Allergies
X4	*	*	*	Diabetes
X5	*	*	*	Flu
X6	*	*	*	Gastritis
X7	*	*	*	Brain Tumor
X8	*	*	*	Lung Cancer
X9	*	*	*	Alzheimer

Two kinds of data input format

1. Microdata
   - data at the granularity of individuals (table row)

2. Graph data -> social network data, in which
   - nodes represent microdata
   - edges represent their structural context
   - graph data are harder to anonymize
     - It's harder to model the background knowledge of an attacker.
     - It is harder to quantify the information loss of modifications.

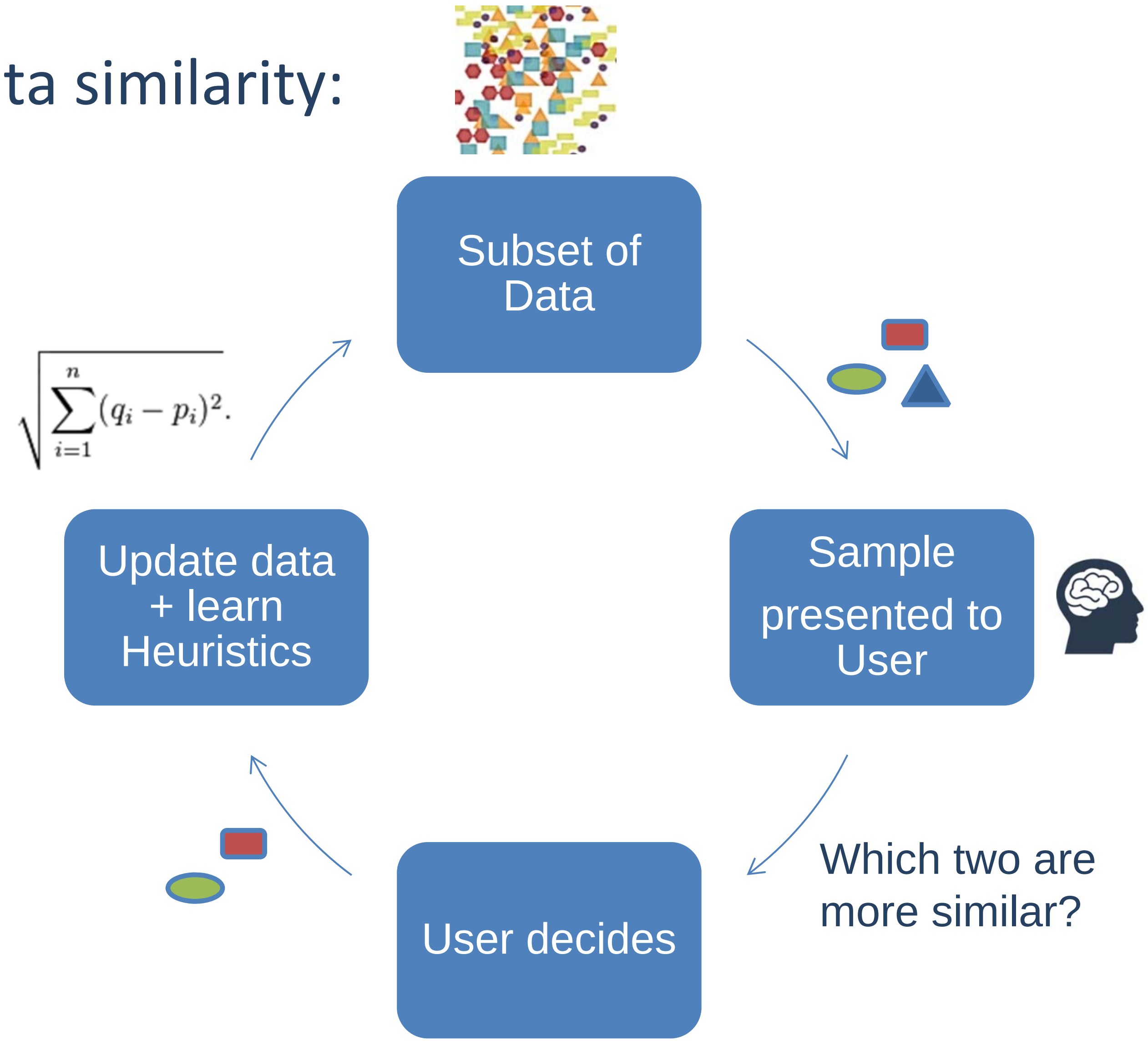
Possibilities to bring iML into anonymization?

"One cost function to rule them all ?"

- Distance functions for Clustering
  - Information loss
  - Structural loss
- Both are subjective
- "Optimality" will also depend on the specific use case (disease spreading / medication research)
- So interactive / reinforcement learning could be applied by involving a domain expert

An iML approach to anonymization 2/2

Case: data similarity:



We will walk through a k-anonymization algorithm – probably SaNGreeA (Social Network Greedy Anonymization)

1. **Lecture:** Identifying the two most important components of the greedy clustering approach
   - Tabular data generalization cost functions (GIL)
   - Network structure loss cost function (SIL)

2. **Assignment**: We will provide a Jupyter notebook with a skeleton implementation, which you will complete by filling in the necessary code for the 2 components above. Goals:
   - Correctness of sample data set
   - Playing around with different cost function parameters, thus getting a feeling for the importance of HITL.

Thank you!

Bernd Malle <b.malle@hci-kdd.org>

Wien, 2016-03-16