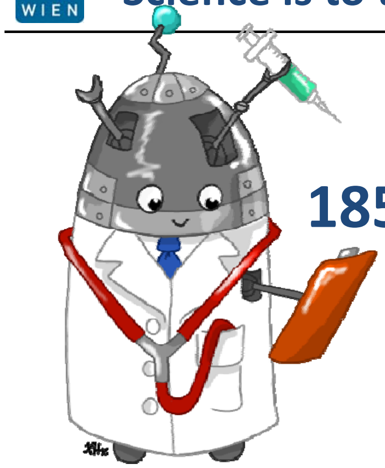


Andreas Holzinger

185.A83 Machine Learning for Health Informatics

2016S, VU, 2.0 h, 3.0 ECTS

Week 24 - 15.06.2016 17:00-20:00

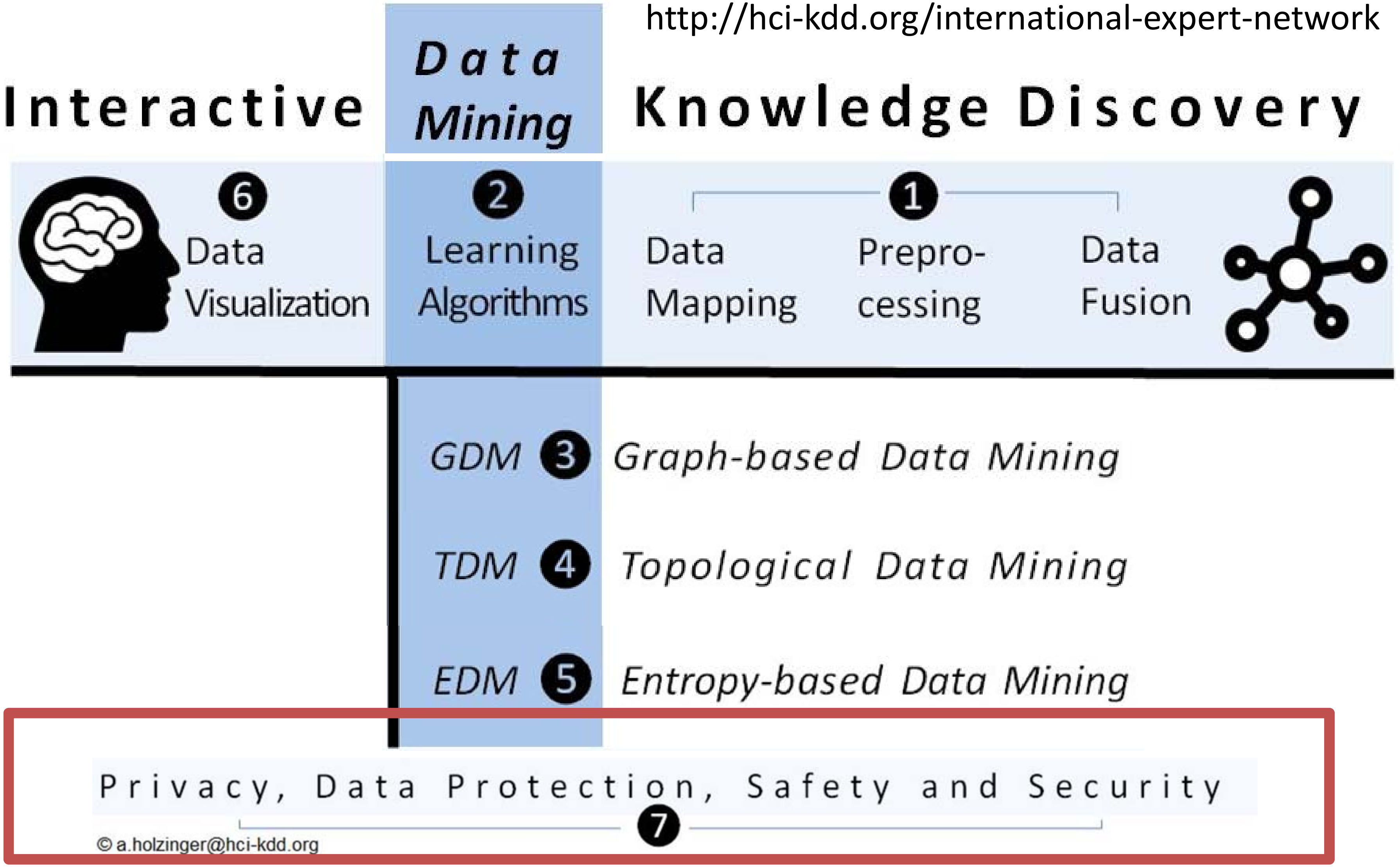
Towards Open Data Sets (k)-Anonymization of Patient Data

b.malle@hci-kdd.org

http://hci-kdd.org/machine-learning-for-health-informatics-course

ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network



Interactive
Data Mining
Knowledge Discovery
6 Data Visualization
2 Learning Algorithms
1
Data Mapping
Prepro-cessing
Data Fusion
GDM 3 Graph-based Data Mining
TDM 4 Topological Data Mining
EDM 5 Entropy-based Data Mining
Privacy, Data Protection, Safety and Security
7
© a.holzinger@hci-kdd.org

Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Overview

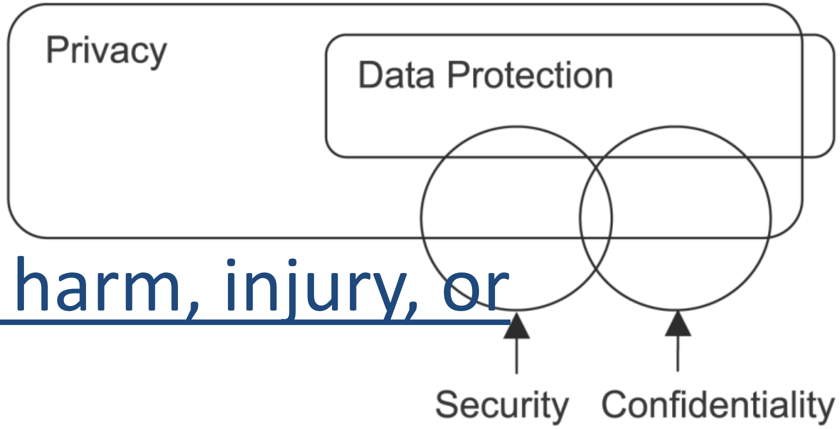
- 1. Introduction & Motivation
- 2. Properties of data & General approach
- 3. Anonymization criteria
- 4. Limits of anonymization
- 5. (Some) Algorithmic Approaches
- 6. SaNGreeA Walkthrough
- 7. Can iML help in anonymization?

Key Challenges

- Sensitive, Personal Health Data
- Mobile solutions, Cloud solutions
- Primary use of Data
- Secondary use of Data for Research
- In the medical area ALL aspects require strict
- **Privacy, Safety, Security and Data Protection!**

Horvitz, E. & Mulligan, D. 2015. Data, privacy, and the greater good. Science, 349, (6245), 253-255.

Definitions (1/2)



- **Safety** = any protection from harm, injury, or damage;
- Data Protection = all measures to ensure availability and integrity of data
- **Privacy** = (US pron. “prai …”; UK pron. “pri …”; from Latin: privatus "separated from the rest”, are the individual rights of people to protect their personal life and matters Confidentiality = secrecy (“ärztliche Schweigepflicht”)

Mills, K. S., Yao, R. S. & Chan, Y. E. (2003) Privacy in Canadian Health Networks: challenges and opportunities. *Leadership in Health Services, 16, 1, 1-10.*

Definitions (2/2)

- **Availability =** p(x) that a system is operational at a given time, i.e. the amount of time a device is actually operating as the percentage of total time it should be operating;
- **Reliability =** the probability that a system will produce correct outputs up to some given time;
- **Security =** (in terms of computer, data, information security) means protecting from unauthorized access, use, modification, disruption or destruction etc.;
- **Dependability** = the system property that integrates such attributes as reliability, availability, safety, security, survivability, maintainability (see slide 11-22);

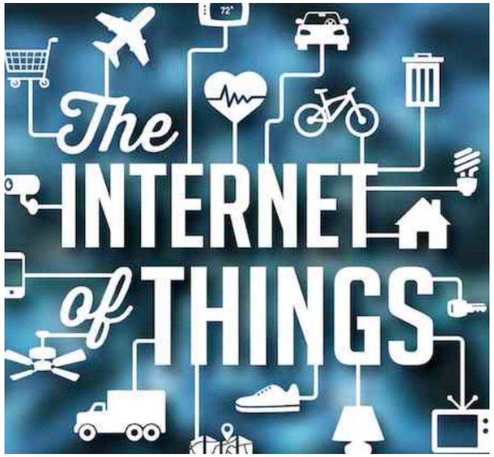
ARES Conference

International Conference on Availability, Reliability and Security

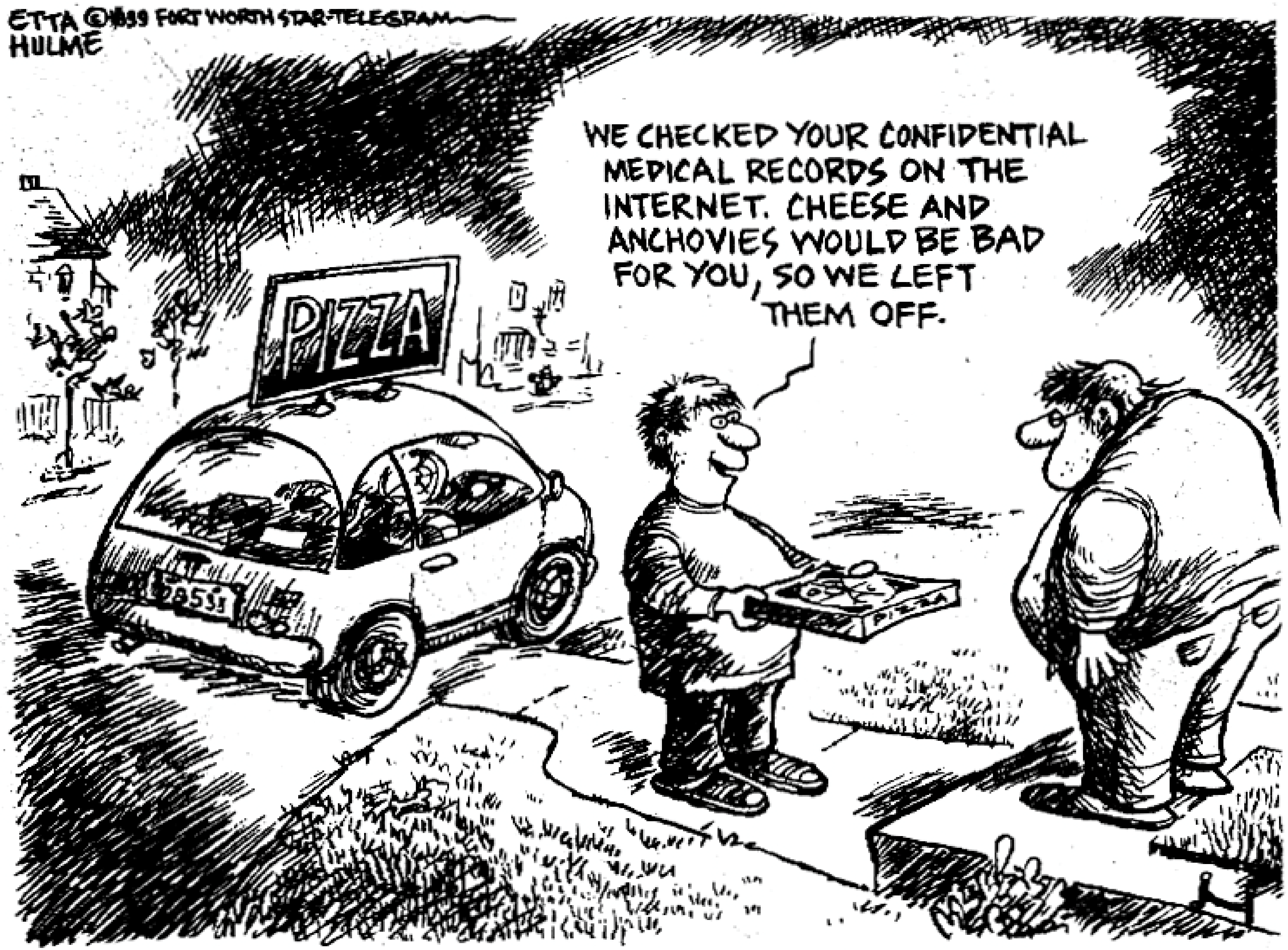
SBA Research

http://www.ares-conference.eu

http://hci-kdd.org/privacy-aware-machine-learning-for-data-science



Machine Learning and Data Privacy …



ETTA ©1999 FORT WORTH STAR-TELEGRAM
HULME

WE CHECKED YOUR CONFIDENTIAL MEDICAL RECORDS ON THE INTERNET. CHEESE AND ANCHOVIES WOULD BE BAD FOR YOU, SO WE LEFT THEM OFF.
PIZZA
PIZZA

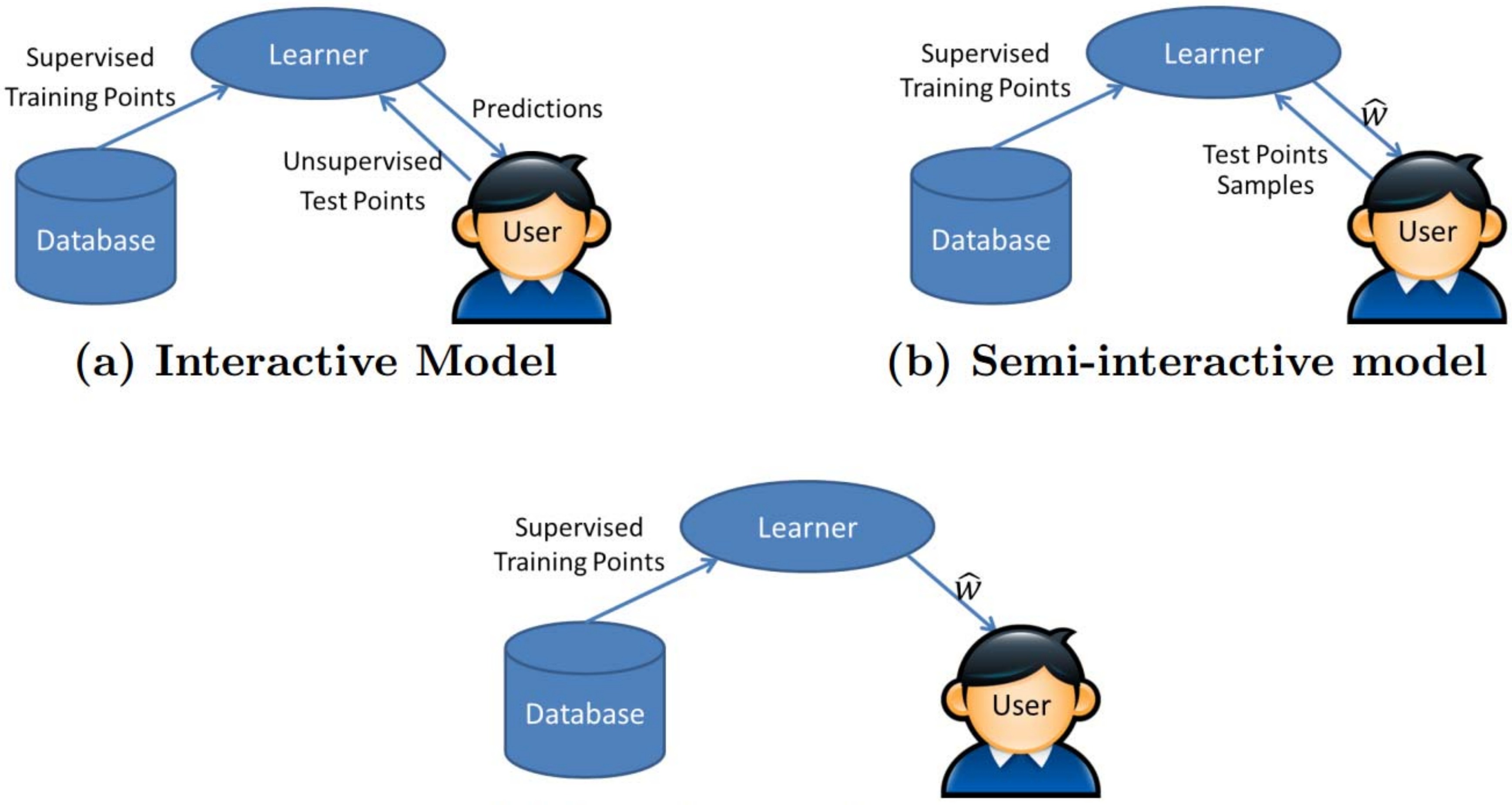
A few Privacy Principles

- Lawfulness and fairness
- Necessity of data collection and processing
- Purpose specification and purpose binding
- There are no "non-sensitive" data
- Transparency
- Data subject´s right to information correction, erasure or blocking of incorrect/ illegally stored data
- Supervision (= control by independent data protection authority) & sanctions
- Adequate organizational and technical safeguards

- **Privacy protection can be undertaken by:**
- Privacy and data protection laws promoted by government
- Self-regulation for fair information practices by codes of conducts promoted by businesses
- Privacy-enhancing technologies (PETs) adopted by individuals
- Privacy education of consumers and IT professionals

Fischer-Hübner, S. 2001. IT-security and privacy: design and use of privacy-enhancing security mechanisms, Springer-Verlag.

Example: Differentially Private Kernel Learning



Supervised Training Points

Learner

Predictions

Unsupervised Test Points

Database

User

(a) Interactive Model

Supervised Training Points

Learner

$\widehat{w}$

Test Points Samples

Database

User

(b) Semi-interactive model

Supervised Training Points

Learner

$\widehat{w}$

Database

User

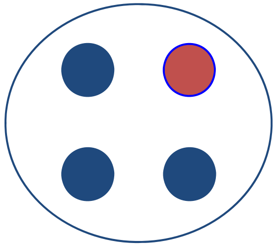
(c) Non-interactive Model

Jain, P. & Thakurta, A. 2013. Differentially Private Learning with Kernels. ICML (3), 28, 118-126.

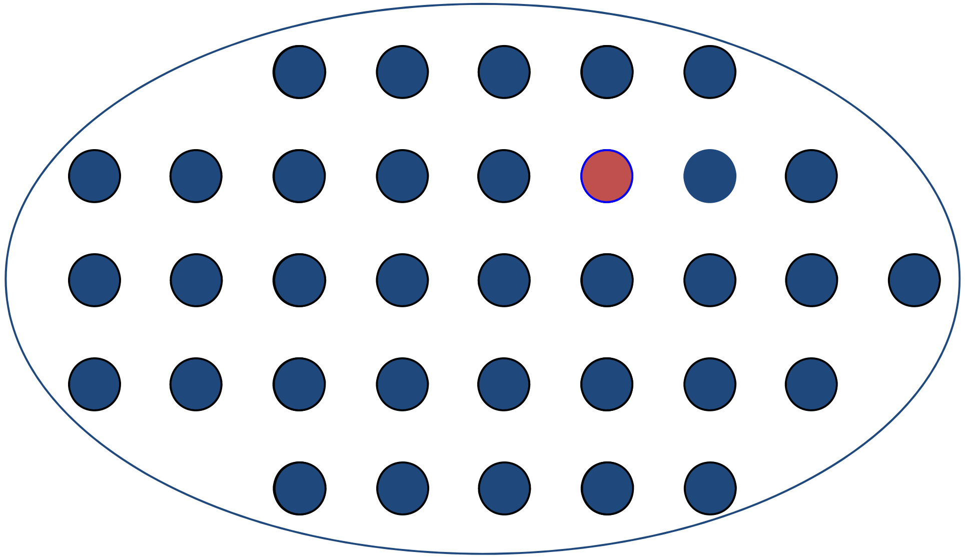
Simplest Privacy Metric

- The larger the set of indistinguishable entities, the lower probability of identifying any one of them

“Hiding in a crowd”



Less anonymous (1/4)



More anonymous (1/*n*)

Anonymity set A

$$A = \{(s_1, p_1), (s_2, p_2), \ldots, (s_n, p_n)\}$$

s_i: subject *i* who might access private data

or: *i*-th possible value for a private data attribute

p_i: probability that s_i accessed private data

or: probability that the attribute assumes the *i*-th possible value

More details see: Bharat K. Bharava (2003), Purdue University

Effective Anonymity Set Size

- Effective anonymity set size is calculated by

$$L = |A| \sum_{i=1}^{|A|} \min p_i \frac{1}{|A|}$$

Maximum value of L is |A| iff all $p_i = 1/|A|$

L below maximum when distribution is skewed

skewed when p_i have different values

Deficiency:

L does not consider violator’s *learning* behavior

Example: Entropy

- Remember: Entropy measures the randomness (uncertainty) – here private data
- Violator gains more information -> entropy decreases!
- Metric: Compare the current entropy value with its maximum value and the difference shows how much information has been leaked
- Privacy loss *D(A,t)* at time *t*, when a subset of attribute values *A* might have been disclosed:

$$D(A,t) = H^*(A) - H(A,t) \quad H(A,t) = \sum_{j=1}^{|A|} w_j \left(\sum_{\forall i} \left(-p_i \ \log_2 (p_i) \right) \right)$$

H(A)* – the maximum entropy

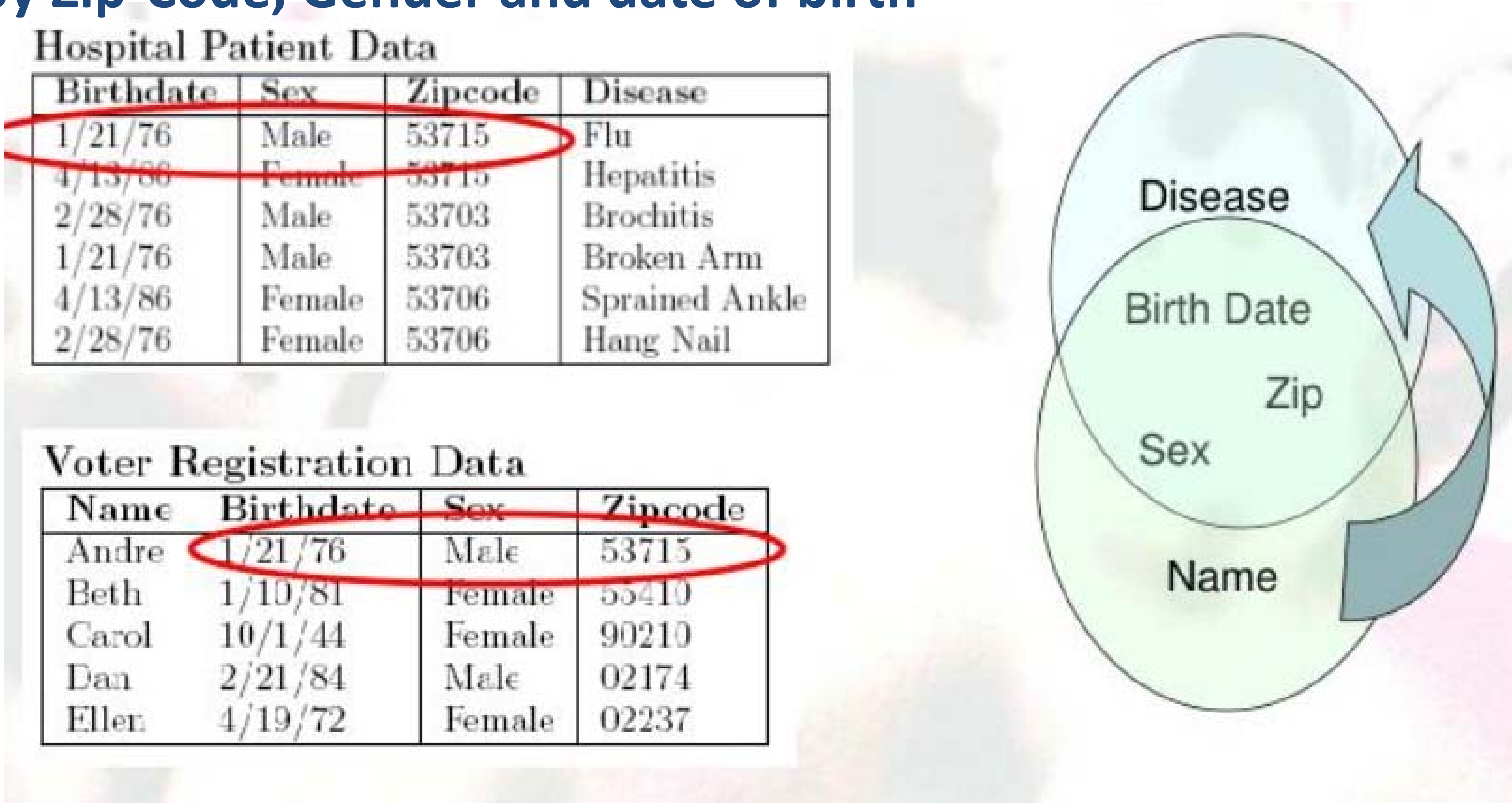
Computed when probability distribution of p_i's is uniform

H(A,t) is entropy at time *t*

w_j – weights capturing relative privacy "value" of attributes

Example : k-Anonymization of Medical Data

87 % of the population in the USA can be uniquely re-identified by Zip-Code, Gender and date of birth



Hospital Patient Data

Birthdate	Sex	Zipcode	Disease
1/21/76	Male	53715	Flu
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Sprained Ankle
2/28/76	Female	53706	Hang Nail

Voter Registration Data

Name	Birthdate	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Samarati, P. 2001. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13, (6), 1010-1027, doi:10.1109/69.971193.

Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10, (05), 571-588.

Anonymization of Patient Data

- **K-Anonymity** … not fully protected against attribute disclosure
- **L-Diversity** … extension requiring that the values of all confidential attributes within a group of *k* sets contain at least *l* clearly distinct values
- **t-Closeness** … extension requiring that the distribution of the confidential attribute within a group of *k* records is similar to the confidential attribute in the whole data set

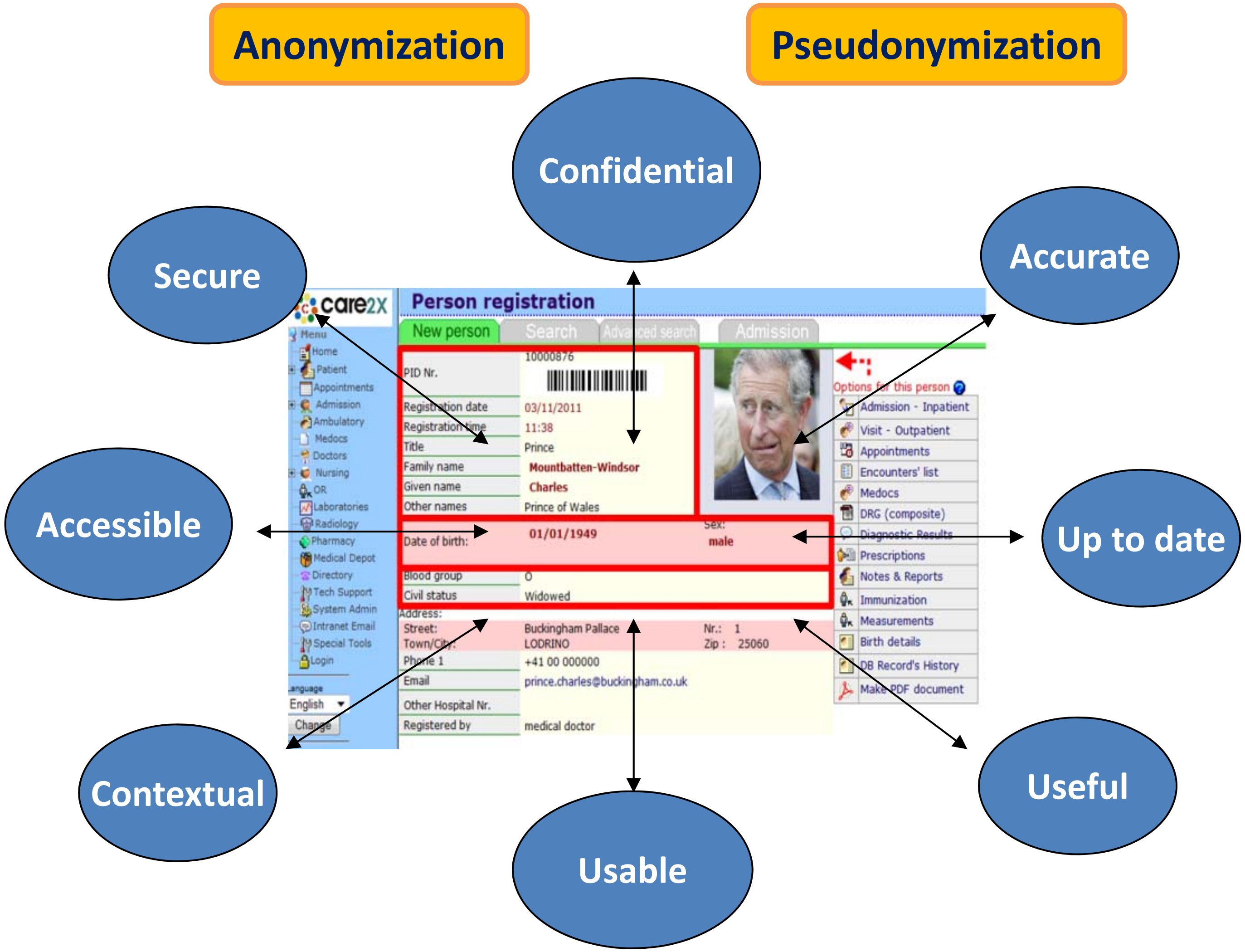
Three Examples of Freeware

- Argus: http://neon.vb.cbs.nl/casc
- ARX: http://arx.deidentifier.org
- sdcTable: http://cran.r-project.org/web/packages/sdcTable/

Future Outlook

- Privacy, Security, Safety and Data Protection are of enormous **increasing importance** in the future.
- Trend to **mobile and cloud** computing approaches.
- EHR are the fastest growing application which concern data privacy and **informed patient consent.**
- Personal health data are being stored for the purpose of maintaining a **life-long health record.**
- **Secondary use** of data, providing patient data for research.
- Production of **Open Data** to support international research efforts (e.g. cancer) without boundaries.
- **Data citation** approaches are needed for full transparency and replicability of research …

Requirements of an electronic health record (EHR)



Person registration

PID Nr.	10000876		
Registration date	03/11/2011		
Registration time	11:38		
Title	Prince		
Family name	Mountbatten-Windsor		
Given name	Charles		
Other names	Prince of Wales		
Date of birth:	01/01/1949	Sex:	male
Blood group	O		
Civil status	Widowed		
Address:			
Street:	Buckingham Pallace	Nr.:	1
Town/City:	LODRINO	Zip :	25060
Phone 1	+41 00 000000		
Email	prince.charles@buckingham.co.uk		
Other Hospital Nr.			
Registered by	medical doctor		

Anonymization: Personal data cannot be re-identified (e.g. k-Anonymization)
Pseudonymization: The personal data is replaced by a "pseudonym“, which allows later tracking back to the source data (re-identification)

What is the problem?

- 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. 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

K Anonymization criterion 1/4

k-anonymity: for every entry in the DS, there must be at least k-1 identical entries (w.r.t. QI's) => this is 3-anon:

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	25-27	4107*	Male	Allergies
X2	25-27	4107*	Male	Allergies
X3	25-27	4107*	Male	Allergies
X4	30-36	41099	*	Diabetes
X5	27-33	410**	*	Flu
X6	30-36	41099	*	Gastritis
X7	30-36	41099	*	Brain Tumor
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Alzheimer

There are 2 main possible attacks on k-anonymity…

K Anonymization criterion 2/4

1. Homogeneity attack:
   - all entries contain the same piece of sensitive information (Allergies)

Node	Age	Zip	Gender	Disease
X1	25-27	4107*	Male	**Allergies**
X2	25-27	4107*	Male	**Allergies**
X3	25-27	4107*	Male	**Allergies**

2. Background knowledge attack:

- Given two entries with identical QI sets: One has lung cancer, the other diabetes…

Node	Age	Zip	Gender	Disease
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Diabetes

Anonymization criteria 3/4

l-diversity: for every "equivalence class" of (at least k) QI-duplicates, there must be at least ***l*** different "well represented" values for the sensitive attribute

2 possible attacks:

1. Skewness attack:
   - positive 1% / negative 99%
   - If your're negative,

   other sensitive data might be revealed

2. Semantic closeness attack:
   - gastritis / gastric ulcer

Node	QI	Cancer	Drugs
X1	*	N	xyz…
X2	*	N	xyz…
X3	*	N	xyz…
X4	*	N	xyz…
X5	*	N	xyz…
X6	*	Y	xyz…
X7	*	N	xyz…
X8	*	N	xyz…
X9	*	N	xyz…

Anonymization criteria 4/4

t-closeness: an equivalence class has t-closeness if the intra-class distribution of a sensitive attribute differs no more than a threshold t from it's global distribution (whole dataset).

delta-presence:

- model the disclosed dataset (DDS) as subset of a larger DS representing a potential attacker's background knowledge
- A DDS is [d-min, d-max]-present if Pr.(individual from larger DS element DDS) is between d-min, d-max

Limits of anonymization

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

Algorithmic approaches 1 / 3

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.

Algorithmic approaches 2/3

Non-perturbative

- Generalization (hierarchies)
  - fixed ruleset
  - range partitioning (numerical values...)

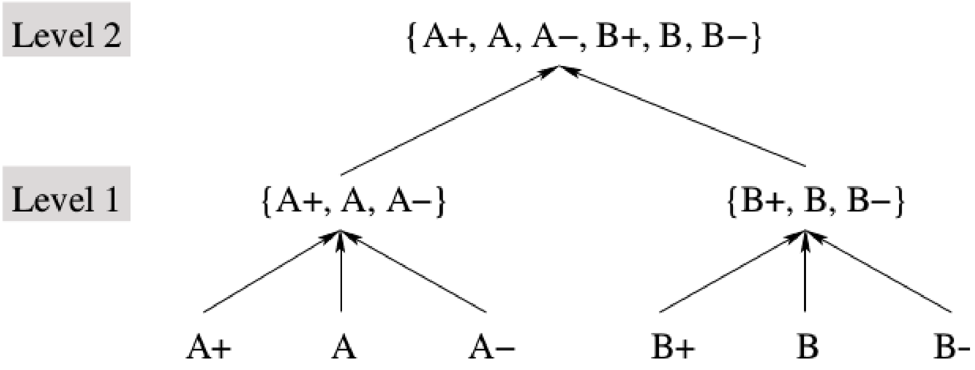


Figure 1: A possible generalization hierarchy for the attribute "Quality".

- Suppression
  - Special case of generalization (with one level)

Bayardo, R. J. & Agrawal, R. Data privacy through optimal k-anonymization. 21st International Conference on Data Engineering (ICDE'05), 2005. IEEE, 217-228.

Algorithmic approaches 3/3

Perturbative

- Adding noise (only distribution counts)
  - Value perturbation => numerical attributes
    - Idea: alter individual data points, keep distribution
  - Graph perturbation
    - (randomly) adding / deleting nodes / edges
    - very efficient / hard to reconstruct

- Microaggregation / Clustering
  - Replace node data by centroid data
  - good for numerical data, but possible also for others given rules
  - Ensures k-anonymity only when computed over all attributes at the same time
  - Exact optimal only in P when computed over just 1 attribute (else heuristic)

Practical walkthrough - SaNGreeA 1/4

“Social Network Greedy Anonymization”

- Anonymizes a dataset w.r.t 2 information categories:
  - Feature vector values => traditional, tabular
  - Graph structure => edge configuration

- Based on the concept of ‘greedy’ clustering

- Which poses the question:
  - How do we choose the next node to add to a cluster w.r.t the above two criteria?

! We need some good cost functions !

Practical walkthrough – SaNGreeA 2/4

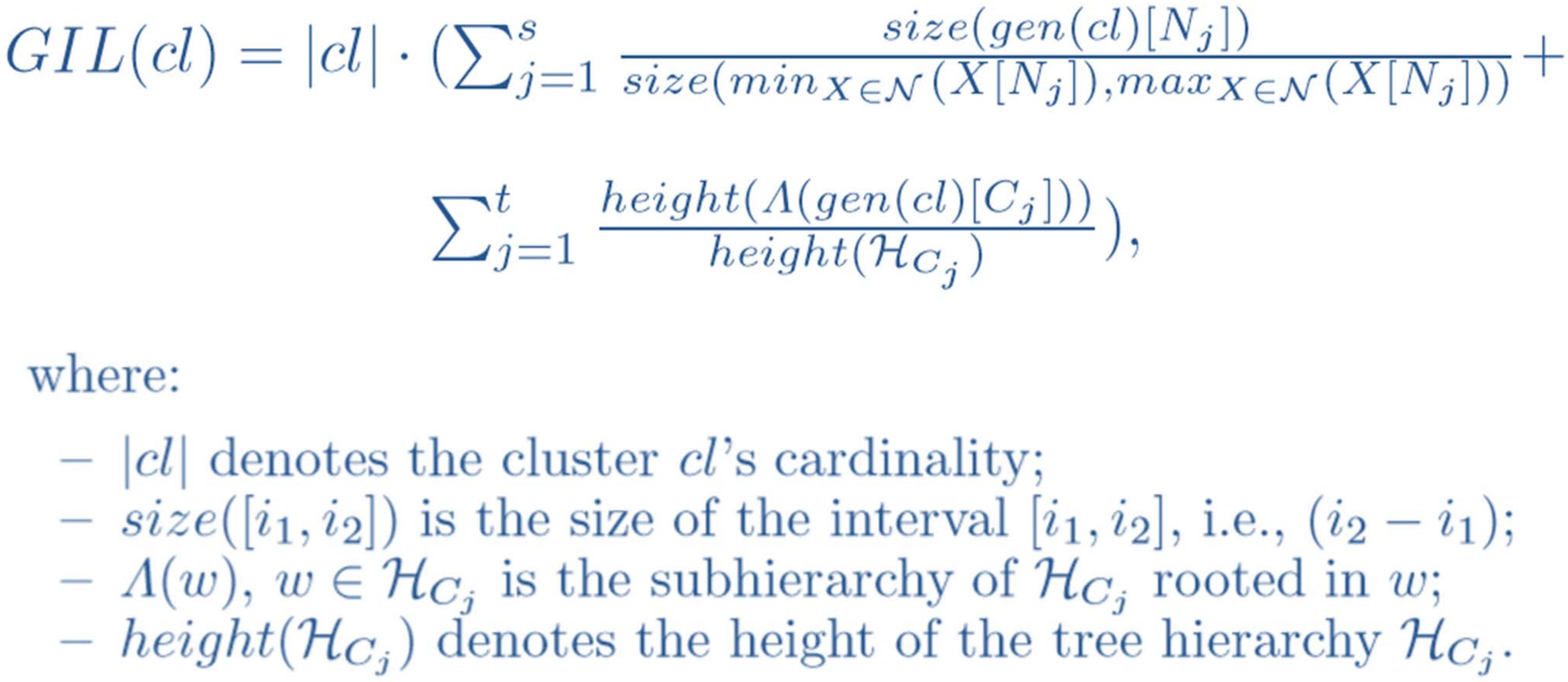
- Generalization Information loss (GIL)
  - Based on content of nodes

- We assume
  - Continuous properties (age, body height, …)
    - Candidate Nodes hold a particular value
    - Clusters have either particular value (at the start) or a generalized range
    - In order to incorporate the node into the cluster, we may have to generalize this range further, increasing the cost.

  - Categorical properties (work class, native-country, …)
    - Same preconditions as above
    - We use generalization hierarchies to determine the cost of clustering

Practical walkthrough – SaNGreeA 2/4

- Generalization information loss function:



$$GIL(cl) = |cl| \cdot (\textstyle\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{X\in\mathcal{N}}(X[N_j]),max_{X\in\mathcal{N}}(X[N_j]))} +$$

$$\textstyle\sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(\mathcal{H}_{C_j})}),$$

where:

- $|cl|$ denotes the cluster cl's cardinality;
- $size([i_1, i_2])$ is the size of the interval $[i_1, i_2]$, i.e., $(i_2 - i_1)$;
- $\Lambda(w)$, $w \in \mathcal{H}_{C_j}$ is the subhierarchy of $\mathcal{H}_{C_j}$ rooted in w;
- $height(\mathcal{H}_{C_j})$ denotes the height of the tree hierarchy $\mathcal{H}_{C_j}$.

Practical walkthrough - SaNGreeA 3/4

- Example GIL:
  - age_range overall = [11 – 91]
  - In order to cluster some nodes, we need to generalize 27 to [20 - 30]
  - Cost = (30-20)/(91-11) = 1/8

  - Given a generalization hierarchy ‘native-country’ with 4 levels
  - In order to cluster, we need to generalize ‘Austria’, ‘France’, or ‘Portugal’ to ‘Western Europe’, which is 1 level higher
  - Cost = 1/4

Practical walkthrough - SaNGreeA 4/4

- Structural Information loss (SIL)

  - Based on neighborhood information
  - Intra-SIL: Measure within a formed cluster
  - Inter-SIL: Measure between formed clusters

    - In short:

    “The probability of wrongly reconstructing a published cluster by mislabeling edges as non-edges and non-edges as edges”

Practical walkthrough – Intra-SIL – SaNGreeA 4/4

- How many potential edges within a cluster? $\binom{|cl|}{2}$
- Probability of any edge to exist? $|\mathcal{E}_{cl}| / \binom{|cl|}{2}$
- Likewise, P(edge not exists)? $1 - |\mathcal{E}_{cl}| / \binom{|cl|}{2}$
- Probability of wrongly labeling an edge as non-edge?

$$|\mathcal{E}_{cl}| \cdot \left(1 - |\mathcal{E}_{cl}| / \binom{|cl|}{2}\right)$$

- Probability of wrongly labeling a non-edge as edge?

$$\left(\binom{|cl|}{2} - |\mathcal{E}_{cl}|\right) \cdot |\mathcal{E}_{cl}| / \binom{|cl|}{2}$$

Campan, A. & Truta, T. M. 2009. Data and structural k-anonymity in social networks. Privacy, Security, and Trust in KDD. Springer, pp. 33-54.

Practical walkthrough - SaNGreeA 4/4

- All-together now:

$$intraSIL(cl) = \left(\left(\binom{|cl|}{2} - |\mathcal{E}_{cl}| \right) \cdot |\mathcal{E}_{cl}| / \binom{|cl|}{2} + |\mathcal{E}_{cl}| \cdot \left(1 - |\mathcal{E}_{cl}| / \binom{|cl|}{2} \right) \right) =$$

$$2 \cdot |\mathcal{E}_{cl}| \cdot \left(1 - |\mathcal{E}_{cl}| / \binom{|cl|}{2} \right).$$

- intraSIL is the # of non-edges times the probability of wrongly labeling a non-edge an edge plus # of edges times the probability of wrongly labeling an edge a non-edge

Campan, A. & Truta, T. M. 2009. Data and structural k-anonymity in social networks. Privacy, Security, and Trust in KDD. Springer, pp. 33-54.

Practical walkthrough – Inter-SIL – SaNGreeA 4/4

- How many potential edges betw. clusters? $|cl_1| \cdot |cl_2|$
- Probability of any edge to exist? $\frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|}$
- Same game as before

$$interSIL(cl_1, cl_2) = \underbrace{(|cl_1| \cdot |cl_2| - |\mathcal{E}_{cl_1,cl_2}|)}_{\#\text{ non-edges}} \cdot \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|} + \underbrace{|\mathcal{E}_{cl_1,cl_2}|}_{\#\text{ edges}} \cdot \left(1 - \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|}\right)$$

$$= 2 \cdot |\mathcal{E}_{cl_1,cl_2}| \cdot \left(1 - \frac{|\mathcal{E}_{cl_1,cl_2}|}{|cl_1| \cdot |cl_2|}\right).$$

- Total SIL function

$$SIL(\mathcal{G}, \mathcal{S}) = \sum_{j=1}^{v} (intraSIL(cl_j)) + \sum_{i=1}^{v} \sum_{j=i+1}^{v} (interSIL(cl_i, cl_j)).$$

Campan, A. & Truta, T. M. 2009. Data and structural k-anonymity in social networks. Privacy, Security, and Trust in KDD. Springer, pp. 33-54.

Practical walkthrough – Actual SIL – SaNGreeA 4/4

- Problem with SIL computations in greedy clustering:

  Real values can only be determined ***after*** all clusters have been built – so this doesn't help ***while*** constructing them !!!

- Solution:
  - compute neighborhood similarity between two nodes

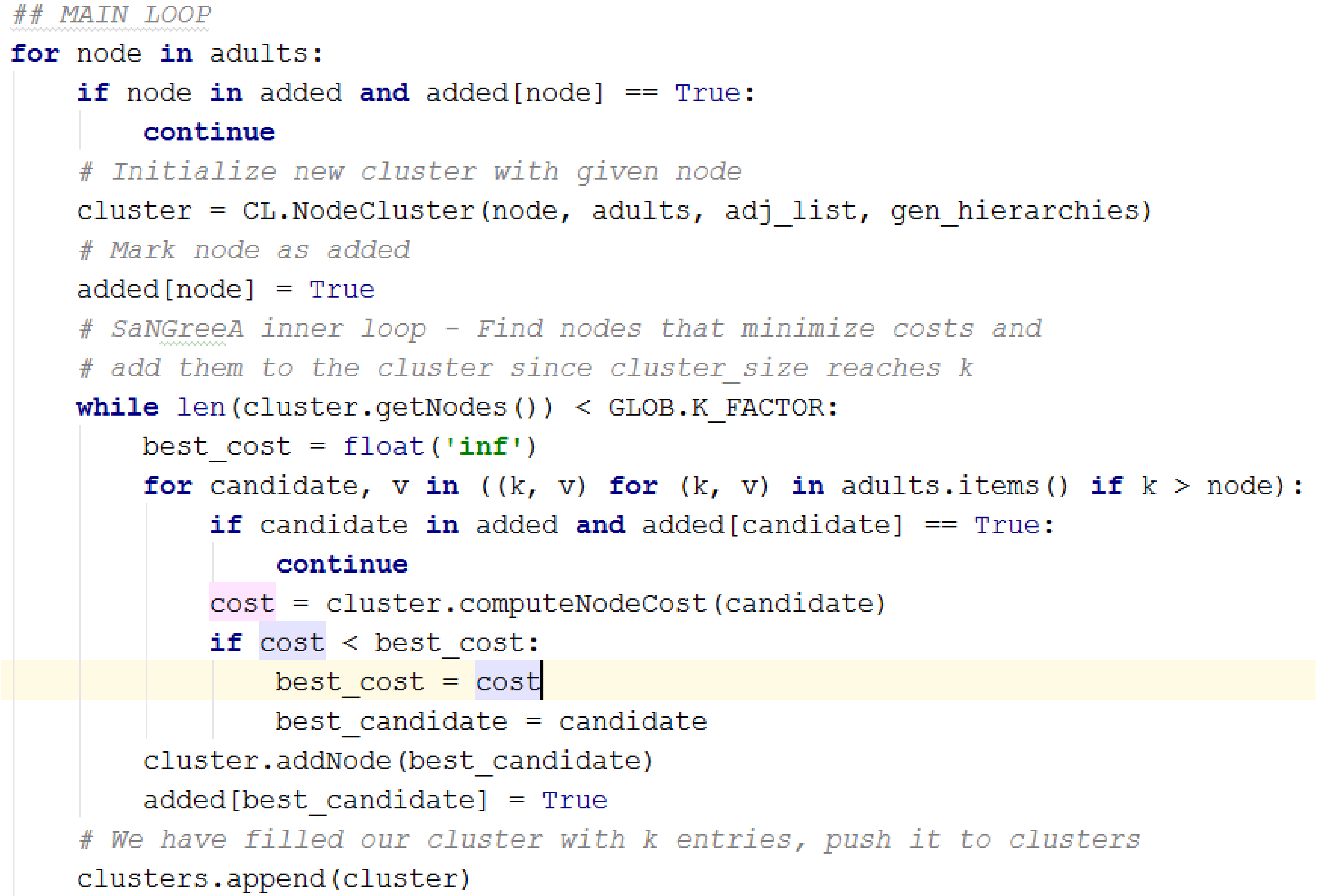
$$dist(X^i, X^j) = \frac{|\{\ell | \ell = 1..n \wedge \ell \neq i,j; b^i_\ell \neq b^j_\ell\}|}{n-2}.$$

  - as well as between a node and a cluster

$$dist(X, cl) = \frac{\sum_{X^j \in cl} dist(X, X^j)}{|cl|}.$$

Campan, A. & Truta, T. M. 2009. Data and structural k-anonymity in social networks. Privacy, Security, and Trust in KDD. Springer, pp. 33-54.

SaNGreeA Main Loop

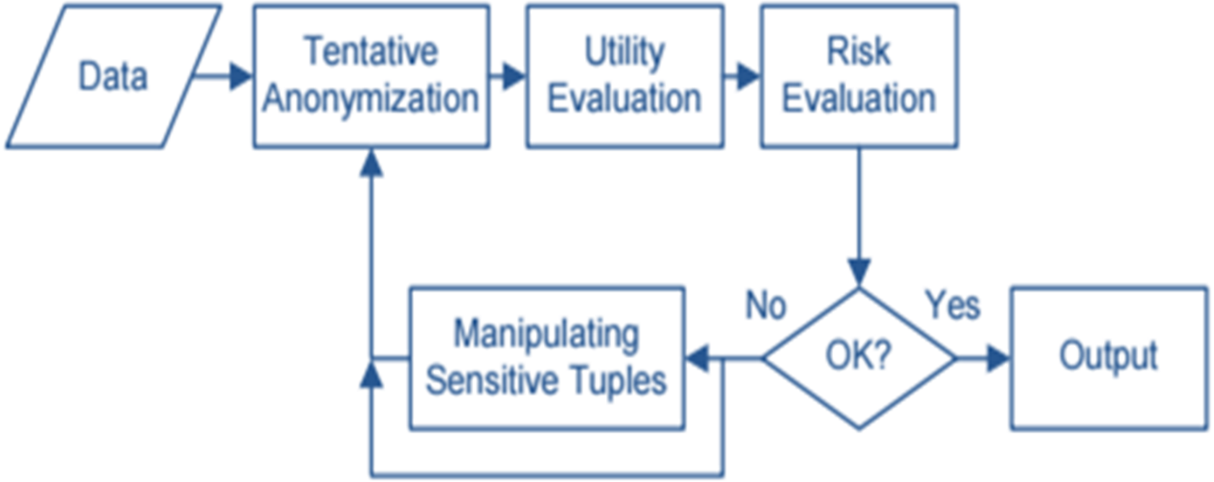


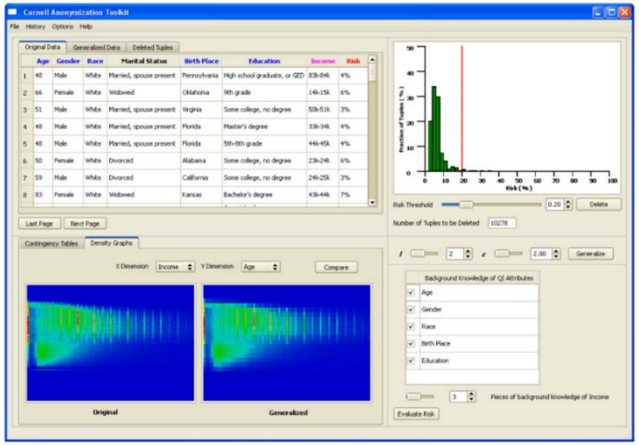
```
## MAIN LOOP
for node in adults:
    if node in added and added[node] == True:
        continue
    # Initialize new cluster with given node
    cluster = CL.NodeCluster(node, adults, adj_list, gen_hierarchies)
    # Mark node as added
    added[node] = True
    # SaNGreeA inner loop - Find nodes that minimize costs and
    # add them to the cluster since cluster_size reaches k
    while len(cluster.getNodes()) < GLOB.K_FACTOR:
        best_cost = float('inf')
        for candidate, v in ((k, v) for (k, v) in adults.items() if k > node):
            if candidate in added and added[candidate] == True:
                continue
            cost = cluster.computeNodeCost(candidate)
            if cost < best_cost:
                best_cost = cost
                best_candidate = candidate
        cluster.addNode(best_candidate)
        added[best_candidate] = True
    # We have filled our cluster with k entries, push it to clusters
    clusters.append(cluster)
```

Can iML help anonymization 1/3 ?

Examples of iML?

- The CAT (Cornell anonymization toolkit) as well as ARX (TU Munich) allow you to run utility / risk analysis
- However, they are not interactive, but only support re-running your experiment with new settings…





	Age	Gender	Race	Marital Status	Birth Place	Education	Income	Risk
1	48	Male	White	Married, spouse present	Pennsylvania	High school graduate, or GED	83k-84k	4%
2	66	Female	White	Widowed	Oklahoma	9th grade	14k-15k	6%
3	51	Male	White	Married, spouse present	Virginia	Some college, no degree	50k-51k	3%
4	48	Male	White	Married, spouse present	Florida	Master's degree	33k-34k	4%
5	48	Male	White	Married, spouse present	Florida	5th-8th grade	44k-45k	4%
6	50	Female	White	Divorced	Alabama	Some college, no degree	23k-24k	6%
7	59	Male	White	Divorced	California	Some college, no degree	24k-25k	3%
8	83	Female	White	Widowed	Kansas	Bachelor's degree	43k-44k	7%

X Dimension: Income
Y Dimension: Age

Risk Threshold: 0.20

Number of Tuples to be Deleted: 10278

l: 2 c: 2.00

Background Knowledge of QI Attributes
☑ Age
☑ Gender
☑ Race
☑ Birth Place
☑ Education

3 Pieces of background knowledge of Income

Can iML help anonymization 2/3 ?

Possibilities to bring iML into anonymization?

“One cost function for all possible application scenarios?”

- Distance functions for Clustering
  - Information loss
  - Structural loss
  - Any other, depending on algorithm

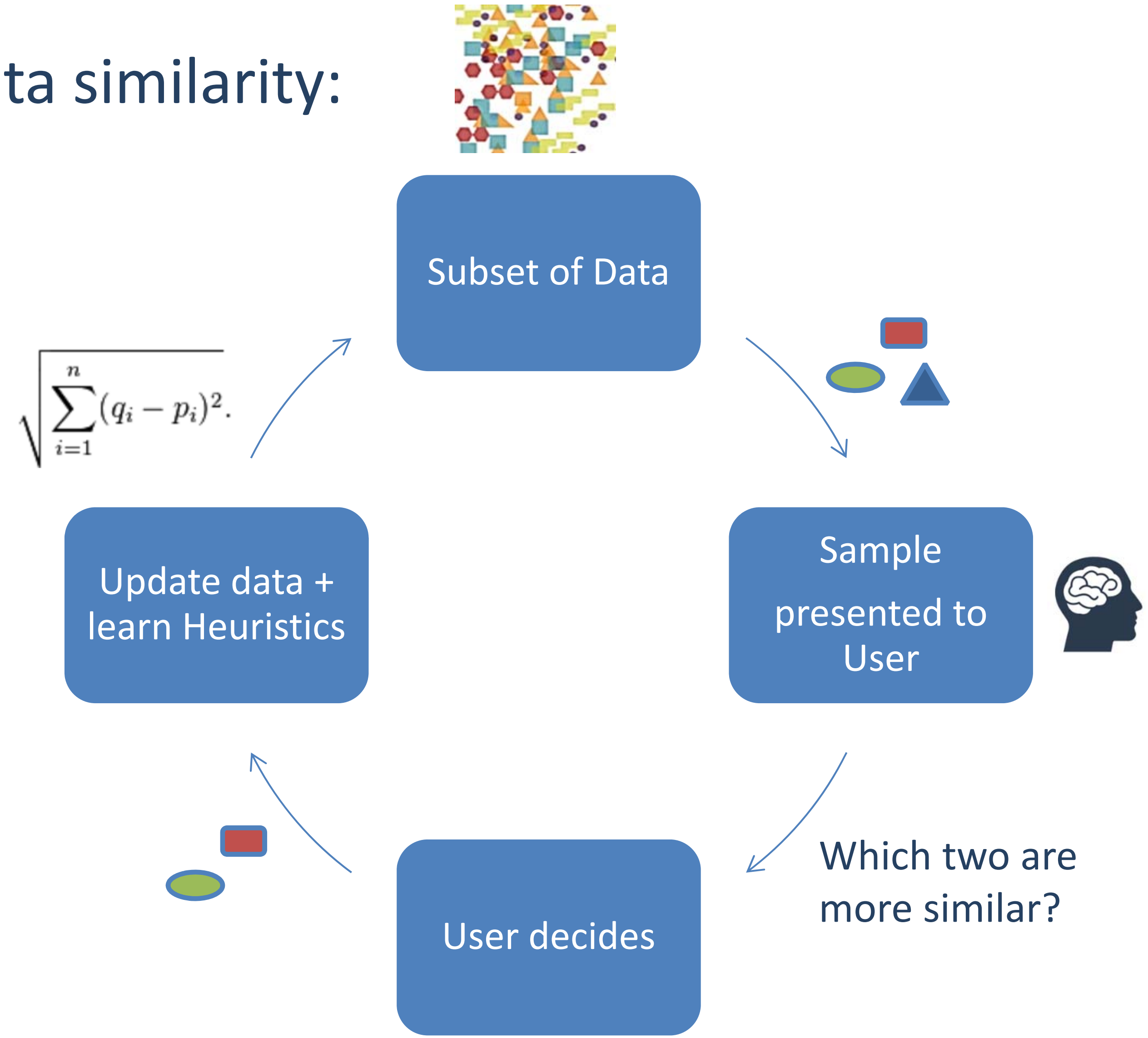
- All possible cost functions are subjective – *“what information do I want to preserve?”*

- “Optimality” will strongly depend on the specific use case (disease spreading / medication research)

- So interactive / reinforcement learning could be applied by involving a domain expert

Can iML help anonymization 3/3 ?

Case: data similarity:



Subset of Data

Sample presented to User

User decides

Update data + learn Heuristics

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

Which two are more similar?

iML Anonymization – an example

[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse

⇧

57 | Private | United-States | Male | White | Married-civ-spouse

⇩

[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*

iML Anonymization – an example

As a result, the weight vector that goes into our cost function changes:

age	workclass	native-country	sex	race	marital-status
0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

⇓

age	workclass	native-country	sex	race	marital-status
0.95	0.01	0.01	0.01	0.01	0.01

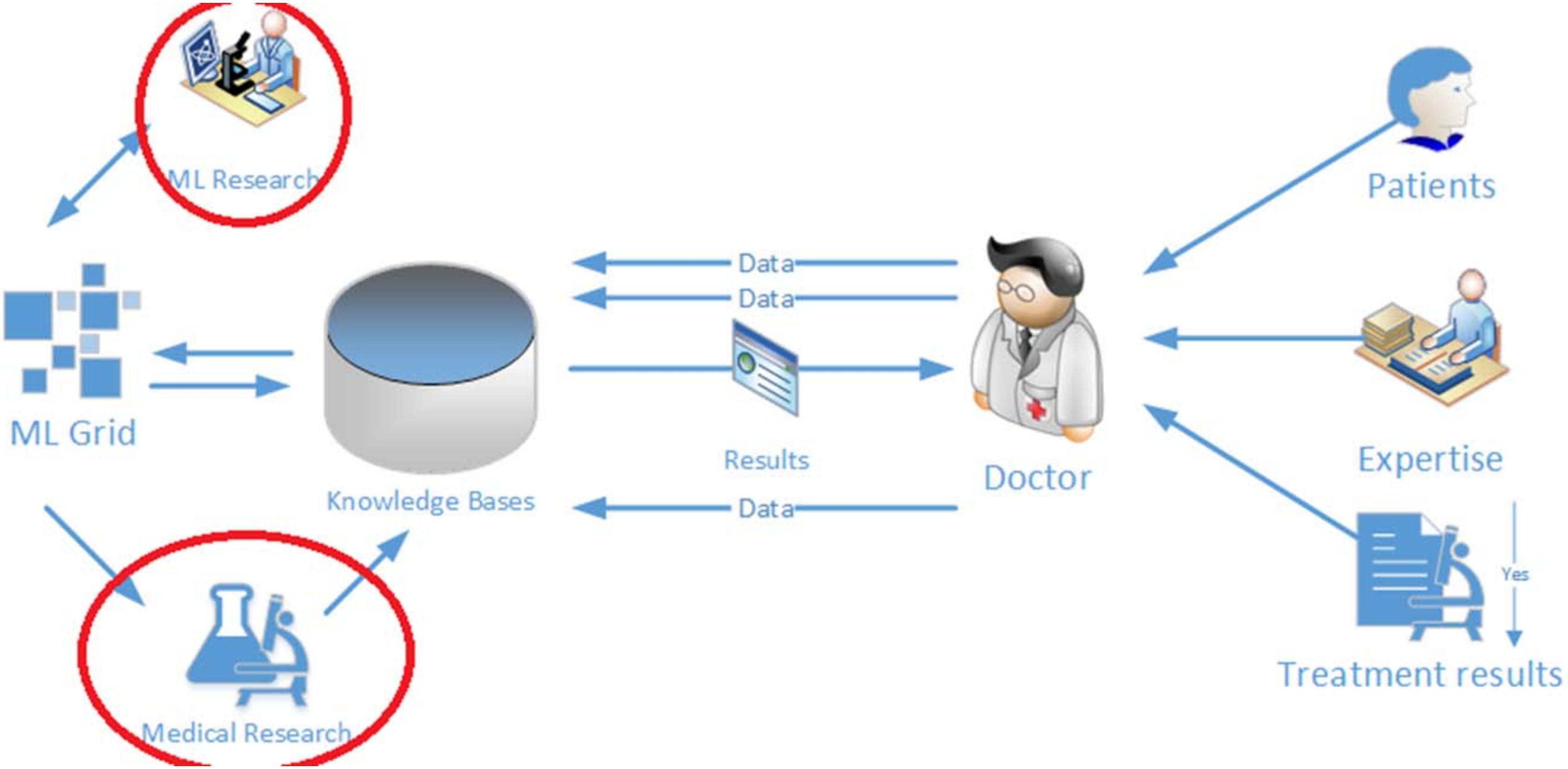
Results… (adult dataset)

Conclusion

- Conclusion: the level of privacy / security of data will always remain subjective with regard to the data set as well as potential attackers !!

- You can never answer the question: "Will this algorithm be good enough for our purposes?" without testing it thoroughly for your specific use cases on your own data…

More related work …



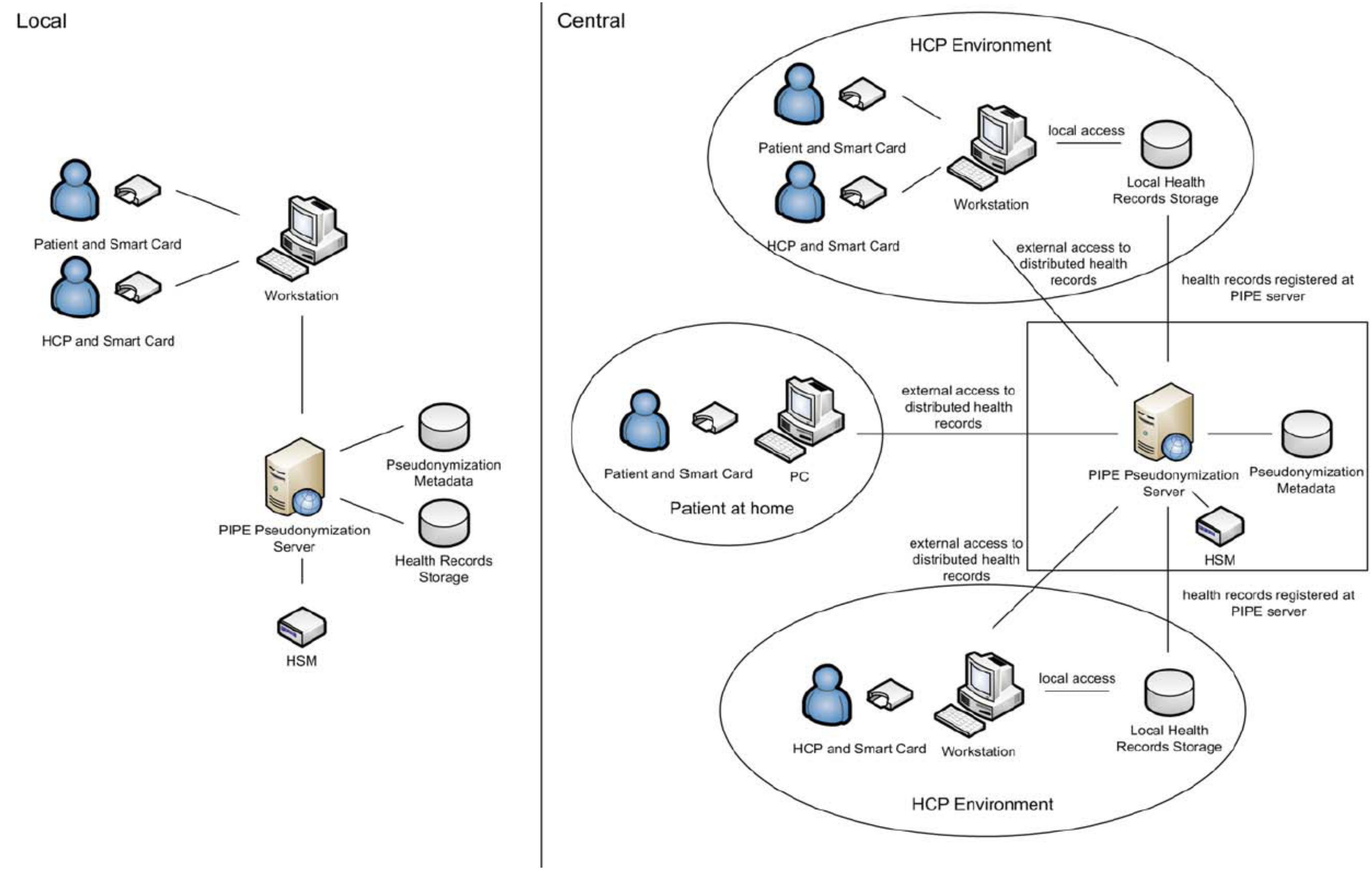
Kieseberg, P., Malle, B., Frühwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. Brain Informatics, 1-11, doi:10.1007/s40708-016-0046-2.

Thank you!

Sample Questions

- What is the difference between privacy, safety, security and data protection?
- How can iML help anonnymization?
- What is the most important issue in k-Anonymization?
- Please explain l-diversity and t-closeness!
- How does SanGReeA work?
- What are the requirements for a EHR?
- How do CAT and ARX work?
- Why is open data necessary for health informatics?
- How do you provide open data sets?

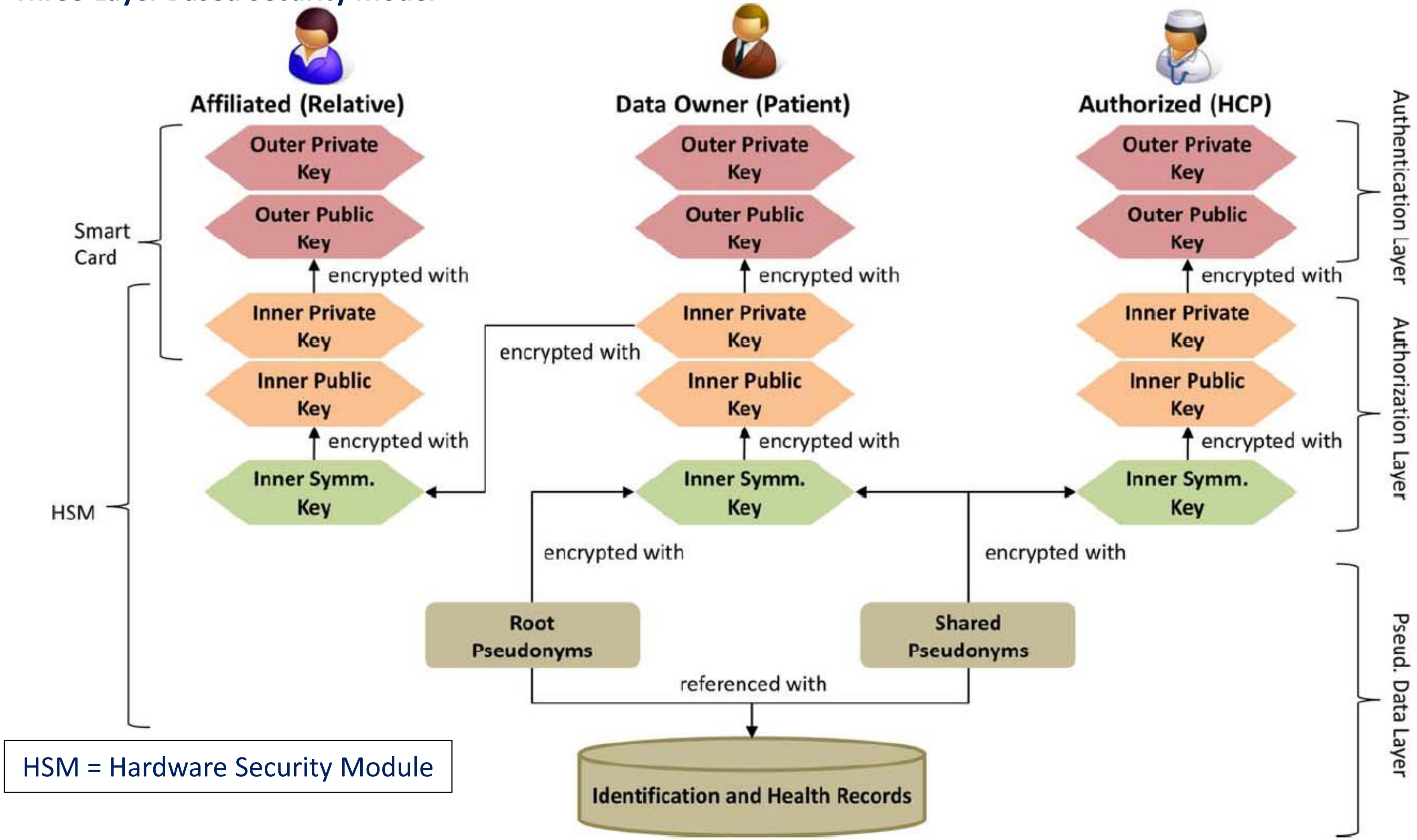
Pseudonymization of Information for Privacy 1/8


Local
Patient and Smart Card
HCP and Smart Card
Workstation
PIPE Pseudonymization Server
Pseudonymization Metadata
Health Records Storage
HSM
Central
HCP Environment
Patient and Smart Card
HCP and Smart Card
Workstation
local access
Local Health Records Storage
external access to distributed health records
health records registered at PIPE server
Patient and Smart Card
PC
Patient at home
external access to distributed health records
PIPE Pseudonymization Server
Pseudonymization Metadata
HSM
external access to distributed health records
health records registered at PIPE server
HCP and Smart Card
Workstation
local access
Local Health Records Storage
HCP Environment

Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

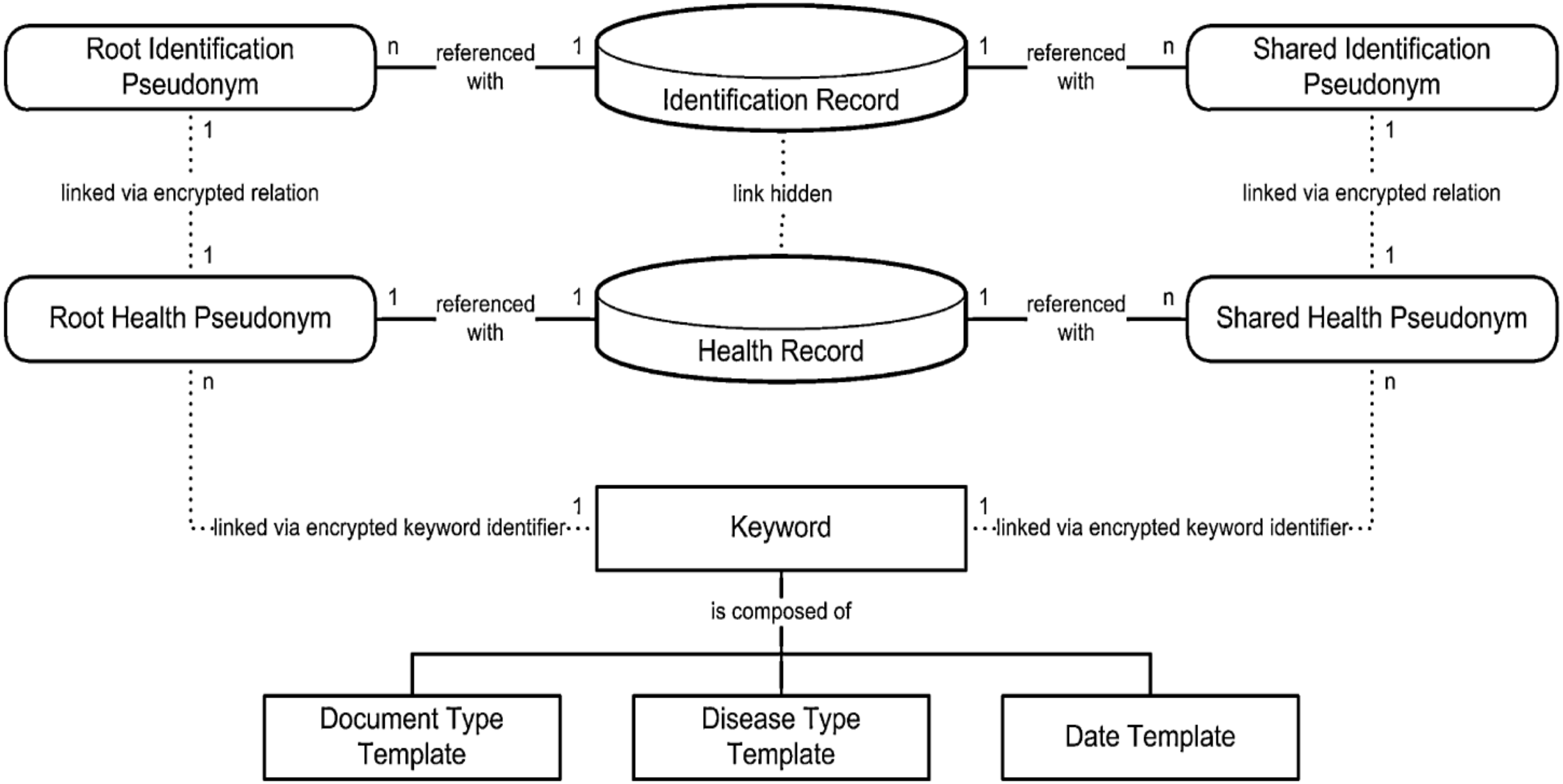
Pseudonymization of Information for Privacy 2/8

Three-Layer Based Security Model



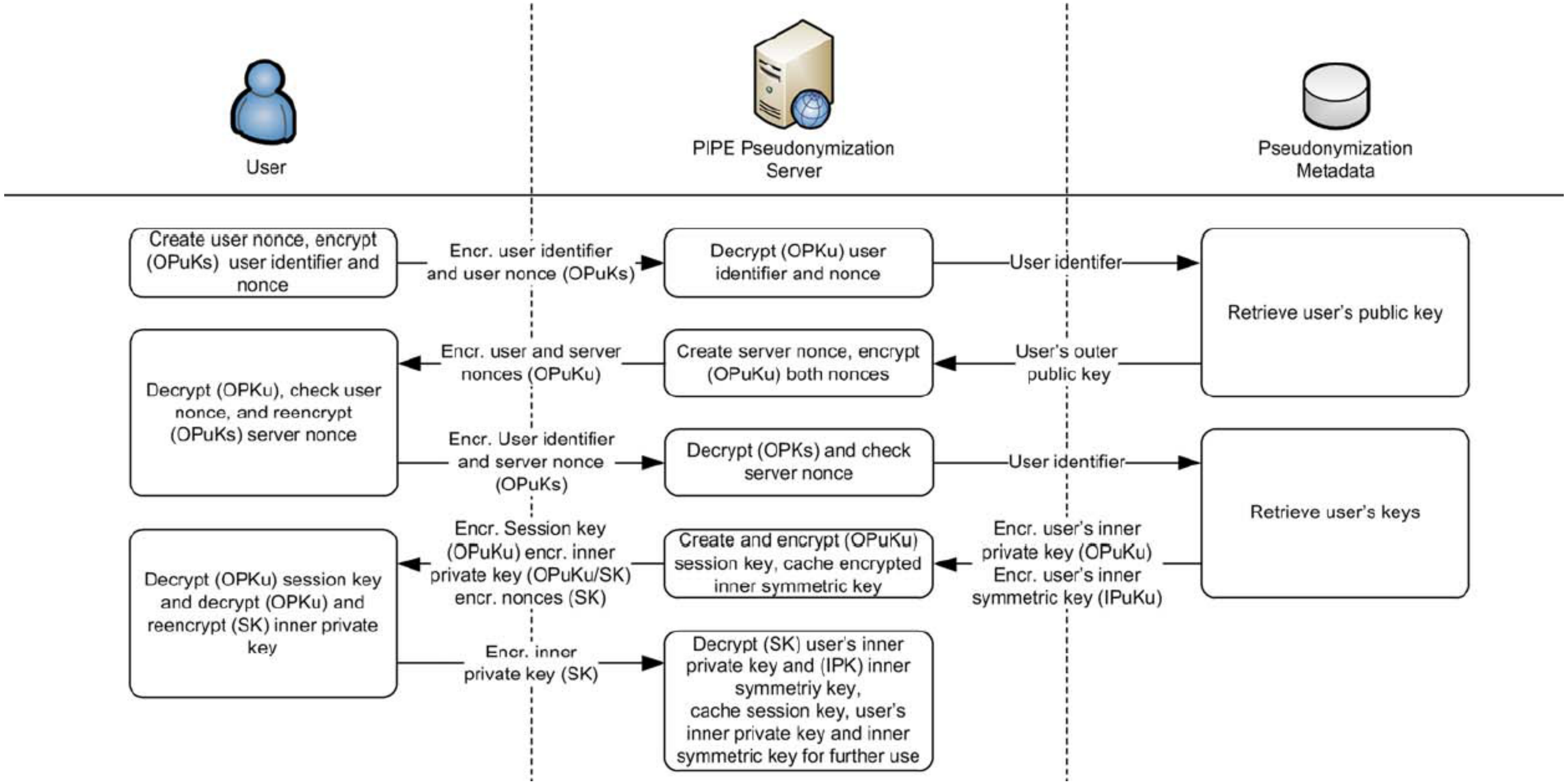
HSM = Hardware Security Module

Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*



Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

Pseudonymization of Information for Privacy 4/8


User

PIPE Pseudonymization Server

Pseudonymization Metadata

Create user nonce, encrypt (OPuKs) user identifier and nonce

Encr. user identifier and user nonce (OPuKs)

Decrypt (OPKu) user identifier and nonce

User identifer

Retrieve user's public key

User's outer public key

Create server nonce, encrypt (OPuKu) both nonces

Encr. user and server nonces (OPuKu)

Decrypt (OPKu), check user nonce, and reencrypt (OPuKs) server nonce

Encr. User identifier and server nonce (OPuKs)

Decrypt (OPKs) and check server nonce

User identifier

Retrieve user's keys

Encr. user's inner private key (OPuKu) Encr. user's inner symmetric key (IPuKu)

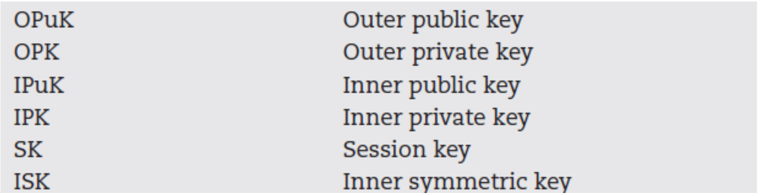
Create and encrypt (OPuKu) session key, cache encrypted inner symmetric key

Encr. Session key (OPuKu) encr. inner private key (OPuKu/SK) encr. nonces (SK)

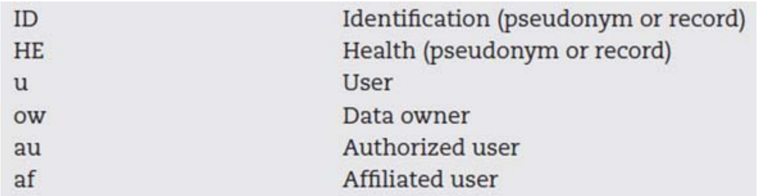
Decrypt (OPKu) session key and decrypt (OPKu) and reencrypt (SK) inner private key

Encr. inner private key (SK)

Decrypt (SK) user's inner private key and (IPK) inner symmetriy key, cache session key, user's inner private key and inner symmetric key for further use



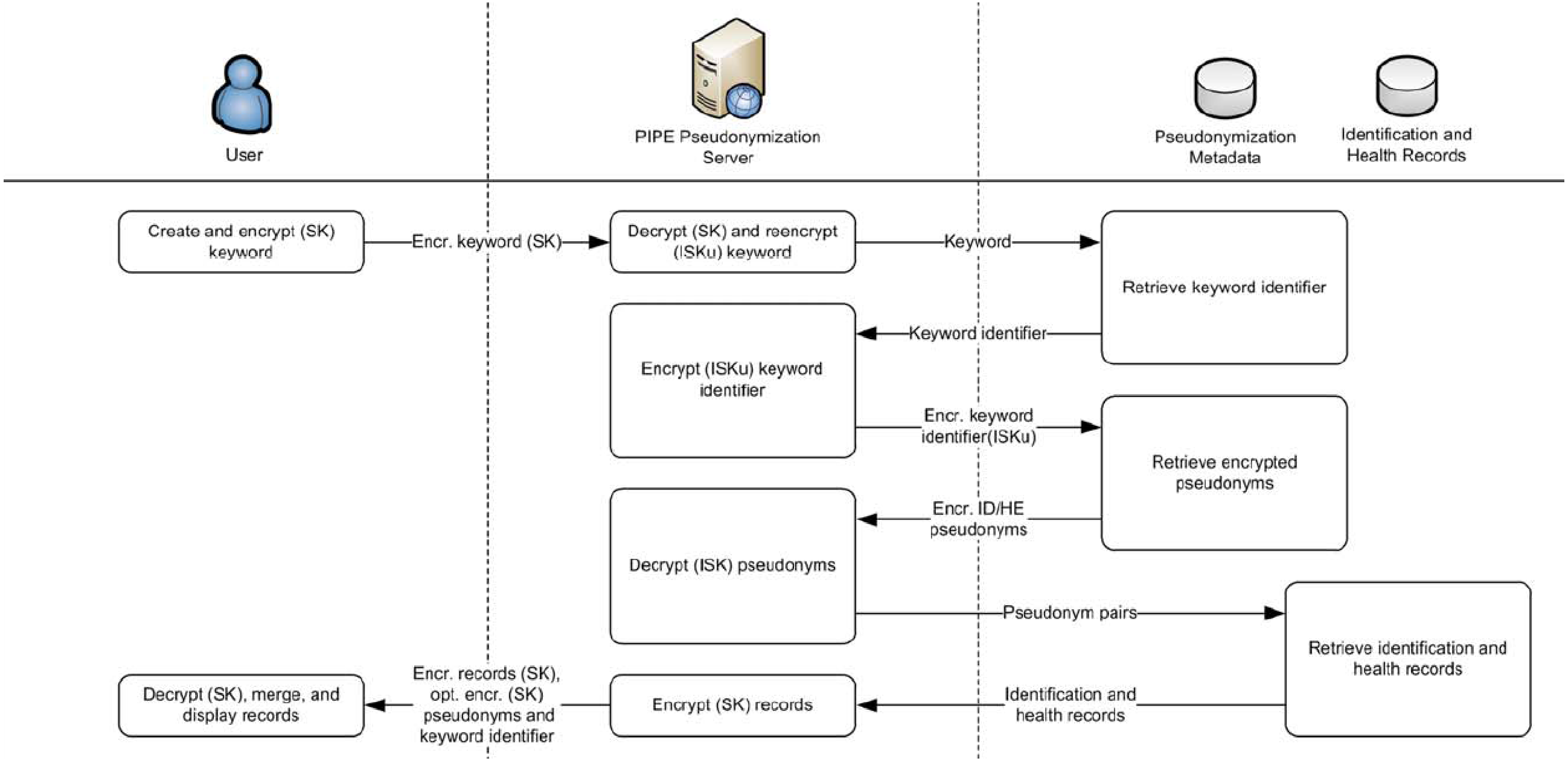
OPuK	Outer public key
OPK	Outer private key
IPuK	Inner public key
IPK	Inner private key
SK	Session key
ISK	Inner symmetric key



ID	Identification (pseudonym or record)
HE	Health (pseudonym or record)
u	User
ow	Data owner
au	Authorized user
af	Affiliated user

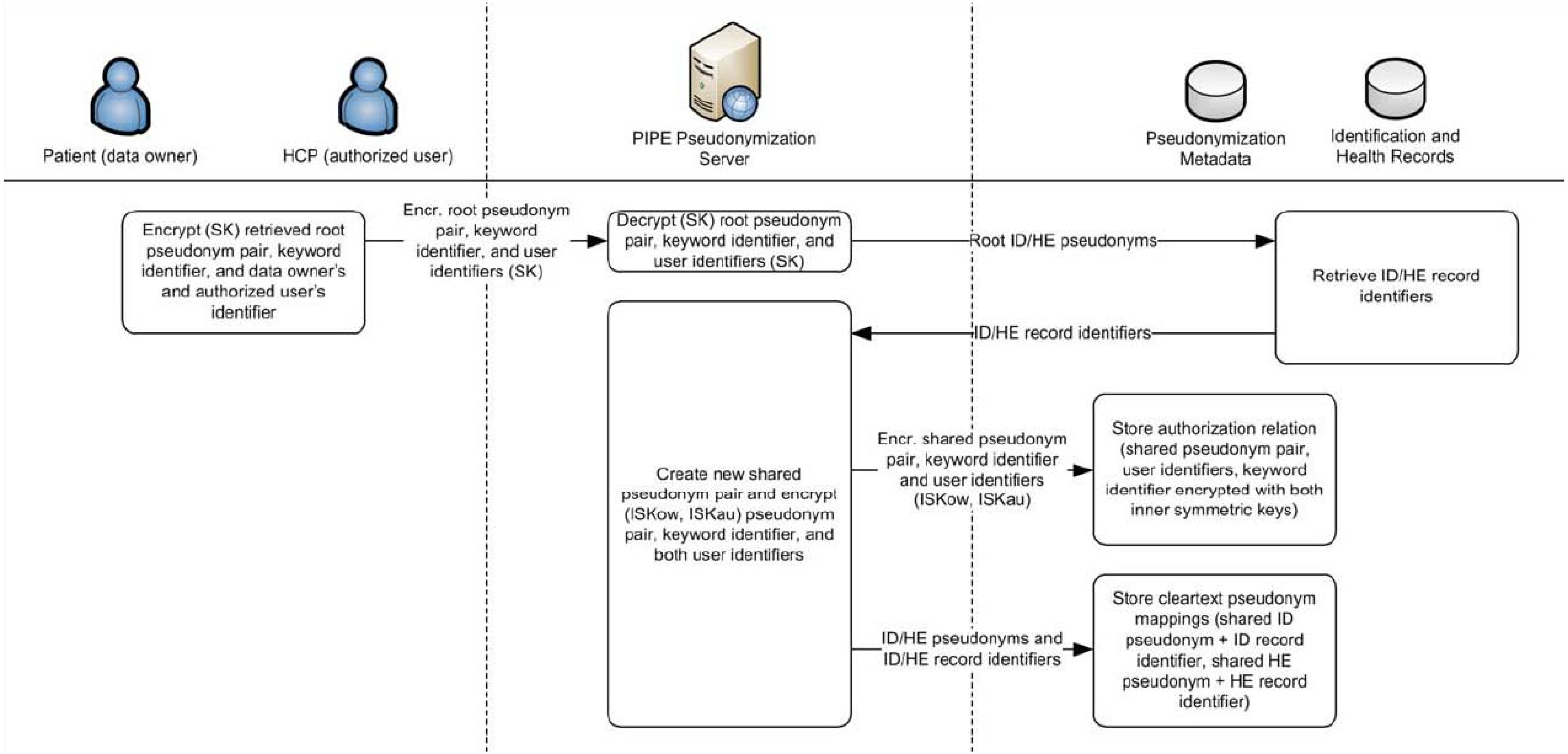
Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

Pseudonymization of Information for Privacy 5/8


User
PIPE Pseudonymization Server
Pseudonymization Metadata
Identification and Health Records
Create and encrypt (SK) keyword
Encr. keyword (SK)
Decrypt (SK) and reencrypt (ISKu) keyword
Keyword
Retrieve keyword identifier
Keyword identifier
Encrypt (ISKu) keyword identifier
Encr. keyword identifier(ISKu)
Retrieve encrypted pseudonyms
Encr. ID/HE pseudonyms
Decrypt (ISK) pseudonyms
Pseudonym pairs
Retrieve identification and health records
Identification and health records
Encrypt (SK) records
Encr. records (SK), opt. encr. (SK) pseudonyms and keyword identifier
Decrypt (SK), merge, and display records

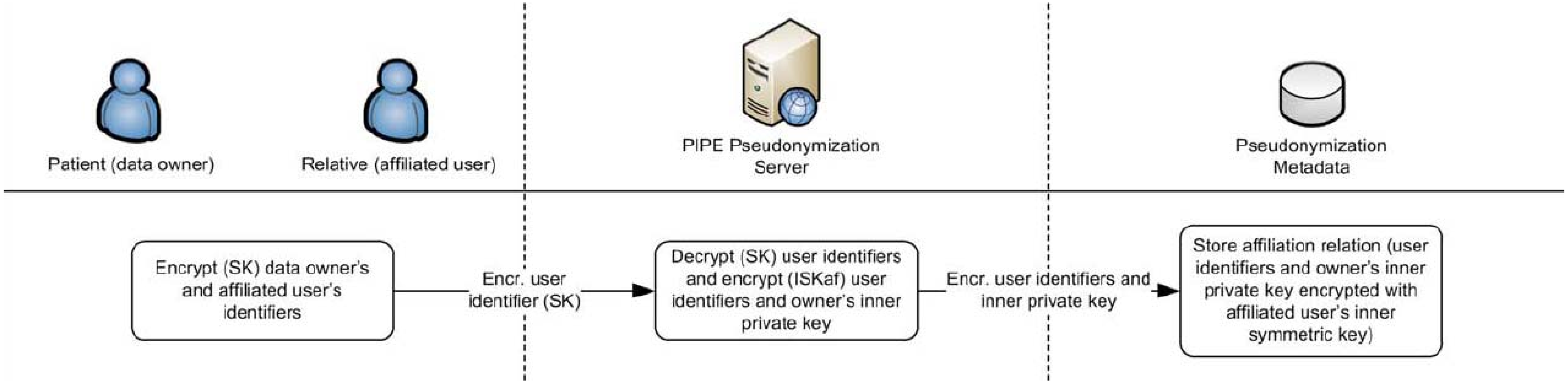
Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

Pseudonymization of Information for Privacy 6/8


Patient (data owner)
HCP (authorized user)
PIPE Pseudonymization Server
Pseudonymization Metadata
Identification and Health Records
Encrypt (SK) retrieved root pseudonym pair, keyword identifier, and data owner's and authorized user's identifier
Encr. root pseudonym pair, keyword identifier, and user identifiers (SK)
Decrypt (SK) root pseudonym pair, keyword identifier, and user identifiers (SK)
Root ID/HE pseudonyms
Retrieve ID/HE record identifiers
ID/HE record identifiers
Create new shared pseudonym pair and encrypt (ISKow, ISKau) pseudonym pair, keyword identifier, and both user identifiers
Encr. shared pseudonym pair, keyword identifier and user identifiers (ISKow, ISKau)
Store authorization relation (shared pseudonym pair, user identifiers, keyword identifier encrypted with both inner symmetric keys)
ID/HE pseudonyms and ID/HE record identifiers
Store cleartext pseudonym mappings (shared ID pseudonym + ID record identifier, shared HE pseudonym + HE record identifier)

Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

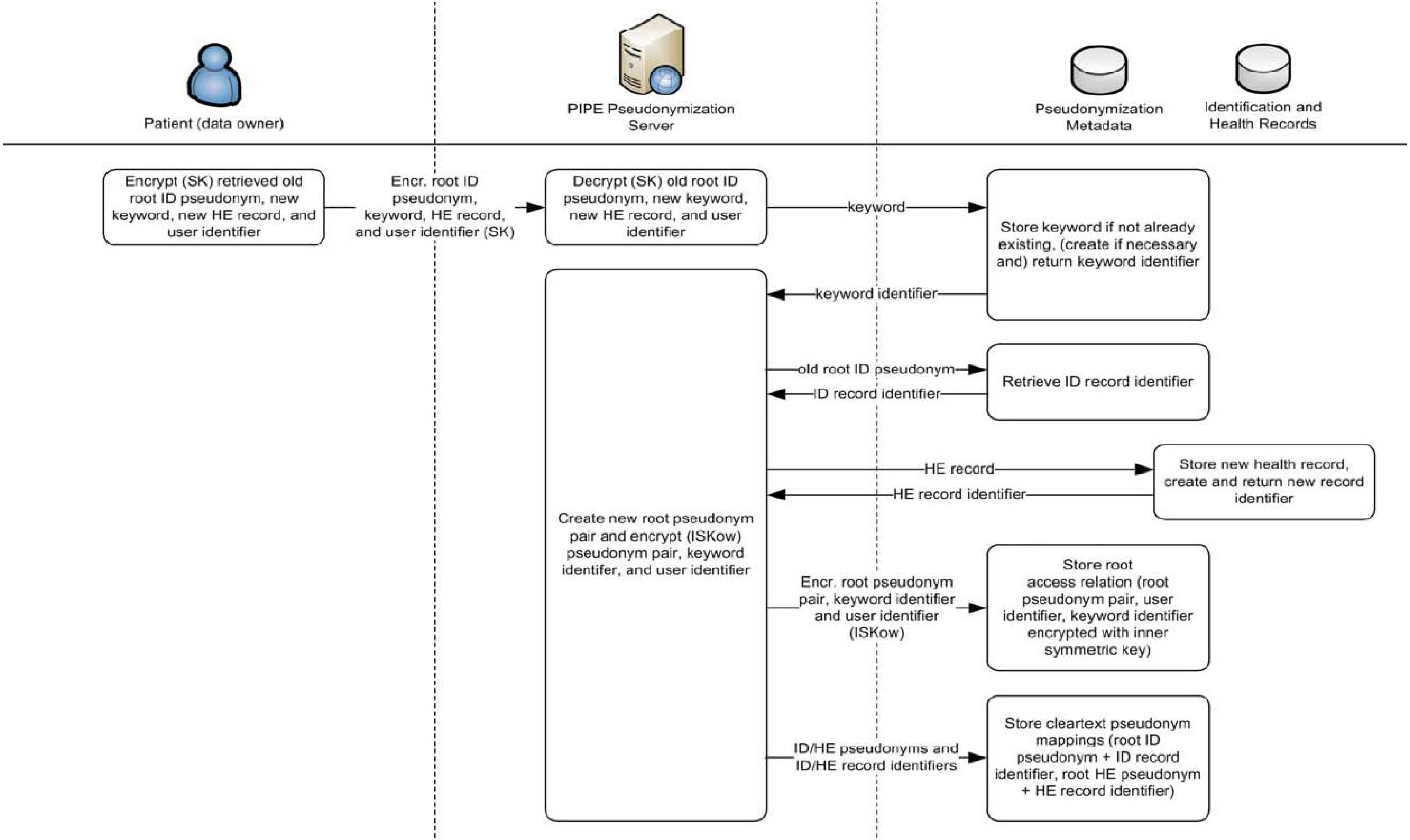
Pseudonymization of Information for Privacy (7/8)


Patient (data owner)
Relative (affiliated user)
PIPE Pseudonymization Server
Pseudonymization Metadata
Encrypt (SK) data owner's and affiliated user's identifiers
Encr. user identifier (SK)
Decrypt (SK) user identifiers and encrypt (ISKaf) user identifiers and owner's inner private key
Encr. user identifiers and inner private key
Store affiliation relation (user identifiers and owner's inner private key encrypted with affiliated user's inner symmetric key)

Note: Similar to authorization, a user affiliation requires that both the patient as data owner and the trusted relative as affiliated user are authenticated at the same workstation. Consequently, both user identifiers are transferred to the pseudonymization server where they are encrypted with both the users' inner symmetric keys. The patient's inner private key is also encrypted with the relative's inner symmetric key, and all elements are stored in the pseudonymization metadata storage as affiliation relation.

Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*

Pseudonymization of Information for Privacy (8/8)



Patient (data owner)

PIPE Pseudonymization Server

Pseudonymization Metadata

Identification and Health Records

Encrypt (SK) retrieved old root ID pseudonym, new keyword, new HE record, and user identifier

Encr. root ID pseudonym, keyword, HE record, and user identifier (SK)

Decrypt (SK) old root ID pseudonym, new keyword, new HE record, and user identifier

keyword

Store keyword if not already existing, (create if necessary and) return keyword identifier

keyword identifier

old root ID pseudonym

ID record identifier

Retrieve ID record identifier

HE record

HE record identifier

Store new health record, create and return new record identifier

Create new root pseudonym pair and encrypt (ISKow) pseudonym pair, keyword identifer, and user identifier

Encr. root pseudonym pair, keyword identifier and user identifier (ISKow)

Store root access relation (root pseudonym pair, user identifier, keyword identifier encrypted with inner symmetric key)

ID/HE pseudonyms and ID/HE record identifiers

Store cleartext pseudonym mappings (root ID pseudonym + ID record identifier, root HE pseudonym + HE record identifier)

Neubauer, T. & Heurix, J. (2011) A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics, 80, 3, 190-204.*