FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

miracum
Medical Informatics in Research and Care in University Medicine

# Datenqualität und k-Anonymität in einem MIRACUM Repository

**Marvin O. Kampf**

Freitag, 29.03.2019

MIRACUM Symposium 2019, Mainz

MEDICAL INFORMATICS INITIATIVE GERMANY

SPONSORED BY THE
Federal Ministry of Education and Research

# Gliederung

1. Was ist k-Anonymität? Wieso Datenqualität?

2. Analyse der k-Anonymität am Standort Erlangen

3. Zusammenfassung und Ausblick

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

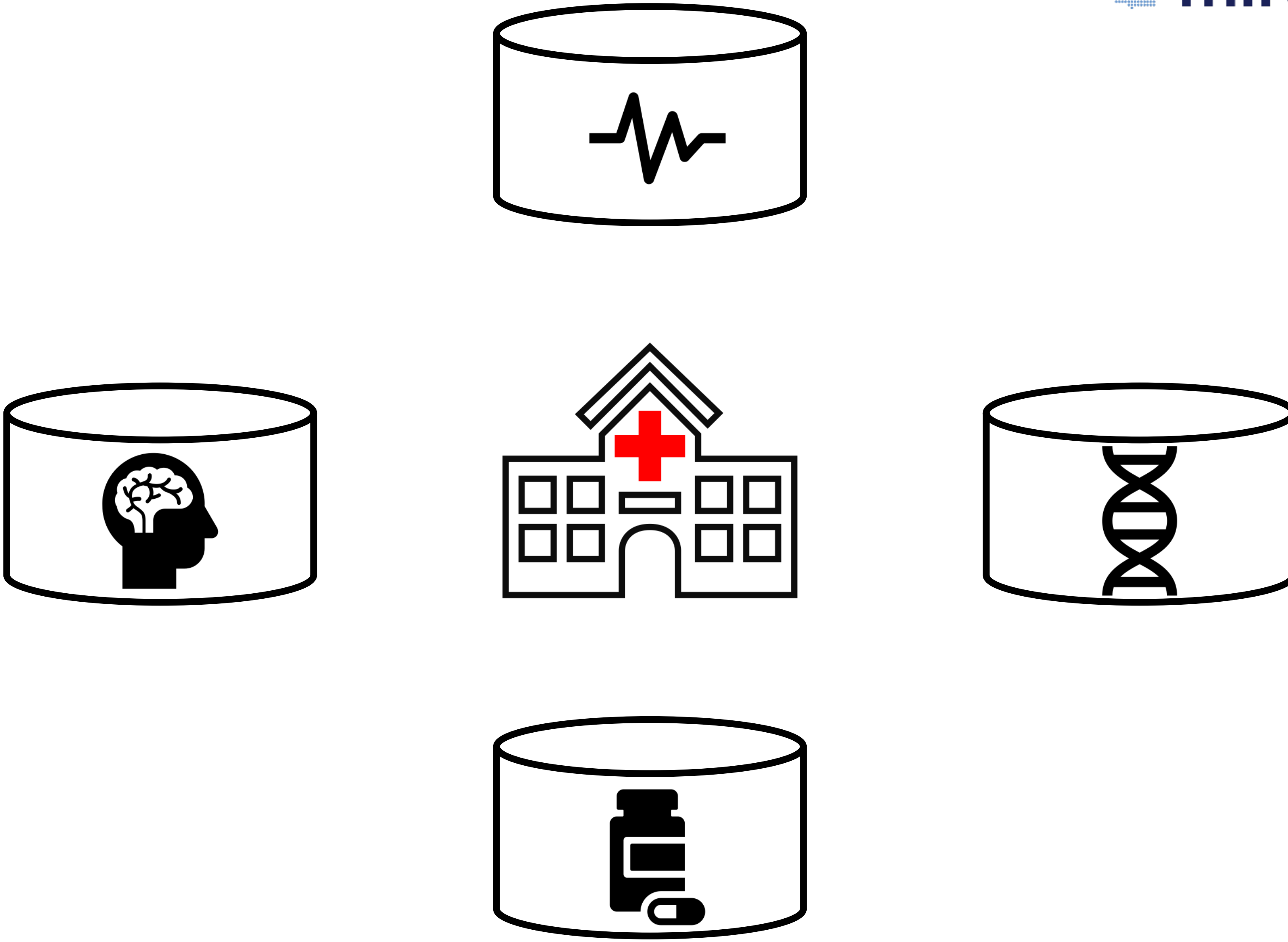
miracum
Medical Informatics in Research and Care in University Medicine

# Was ist k-Anonymität?

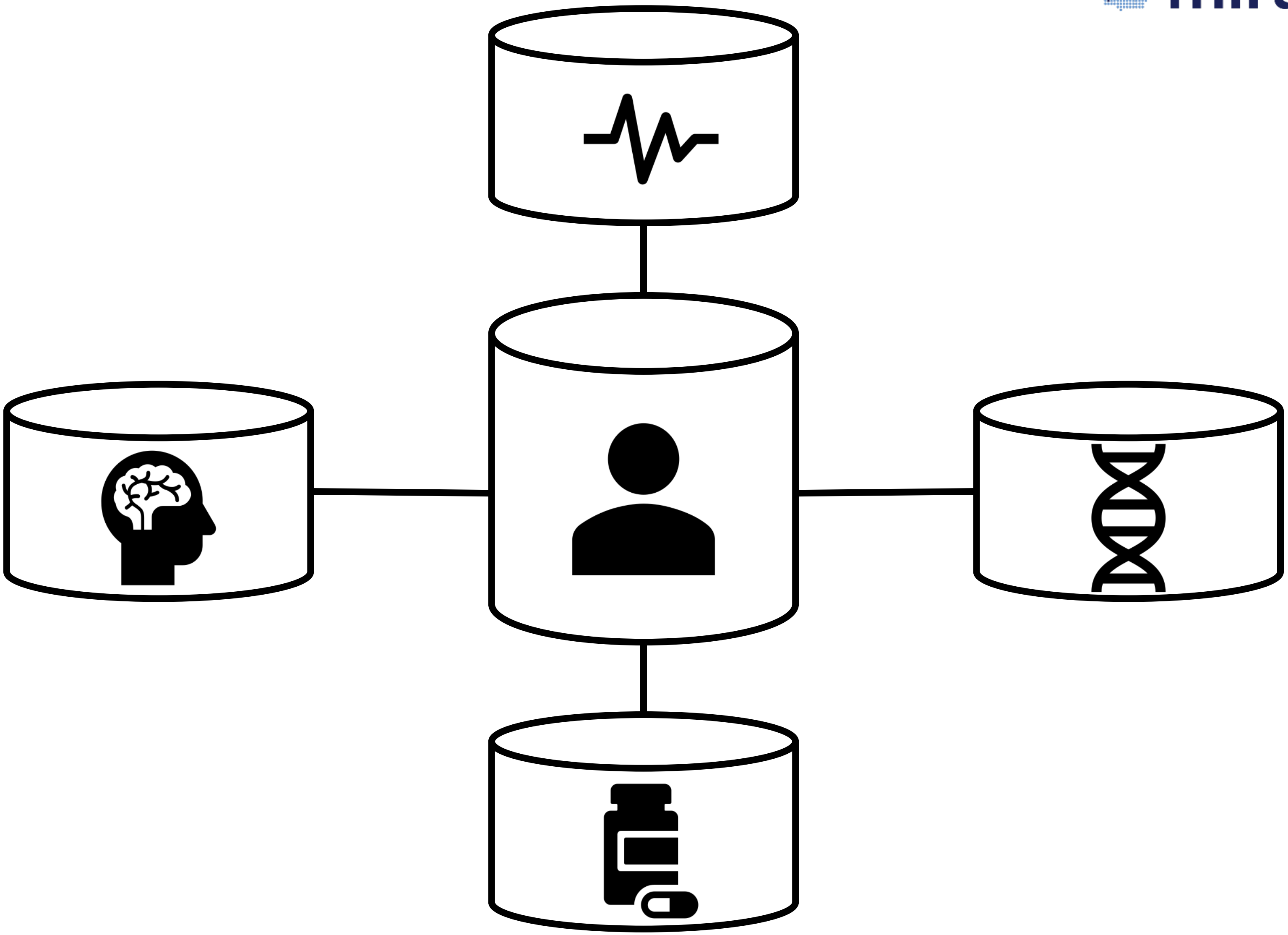
MEDICAL INFORMATICS INITIATIVE GERMANY

SPONSORED BY THE
Federal Ministry of Education and Research

miracum

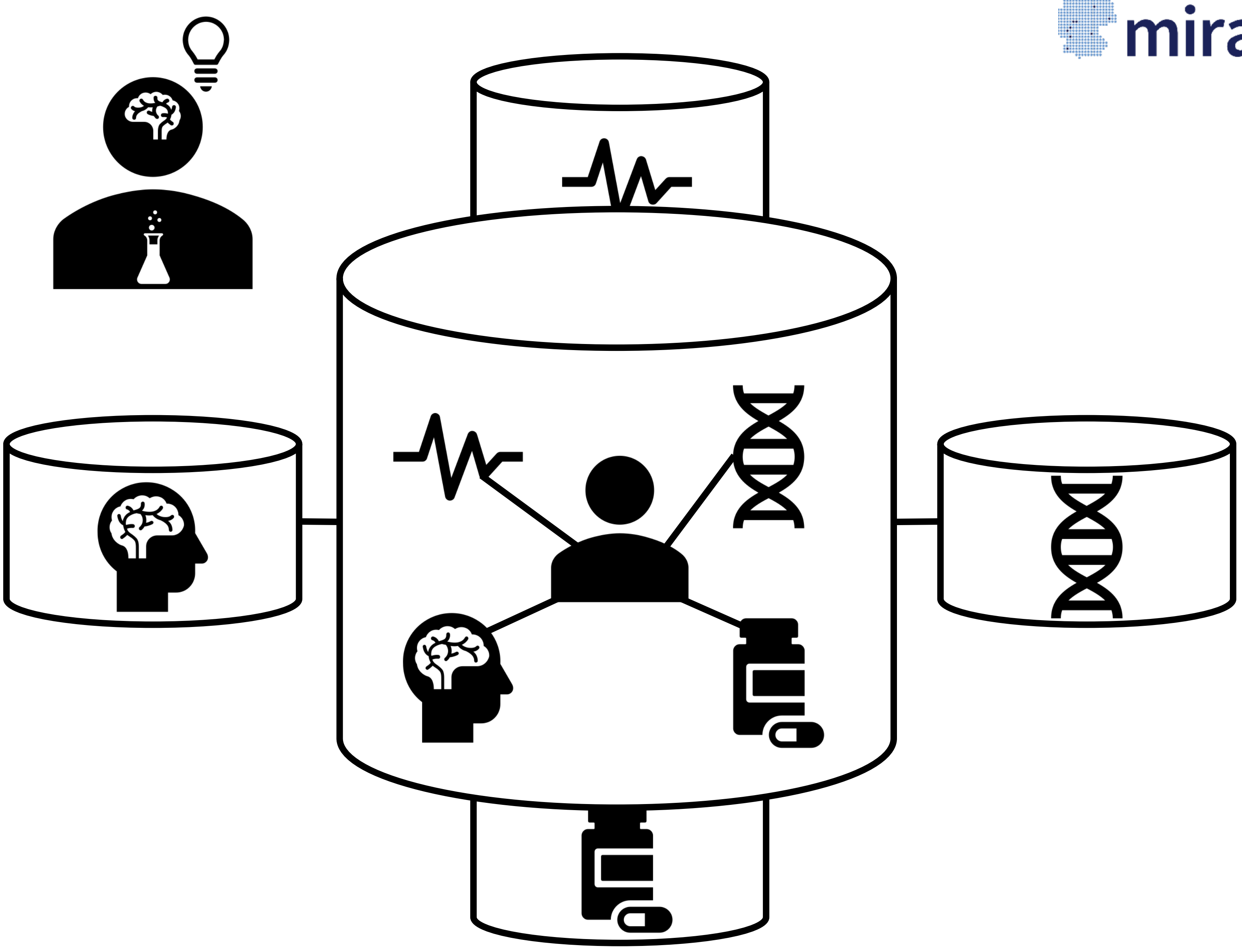


miracum



FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

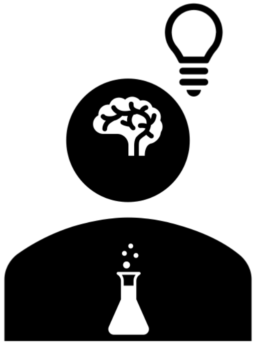


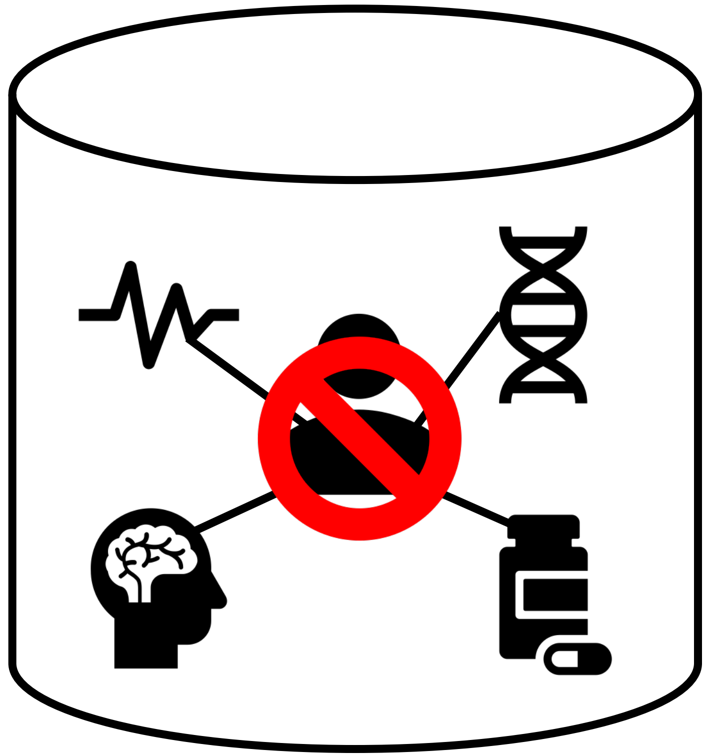
miracum

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen

miracum



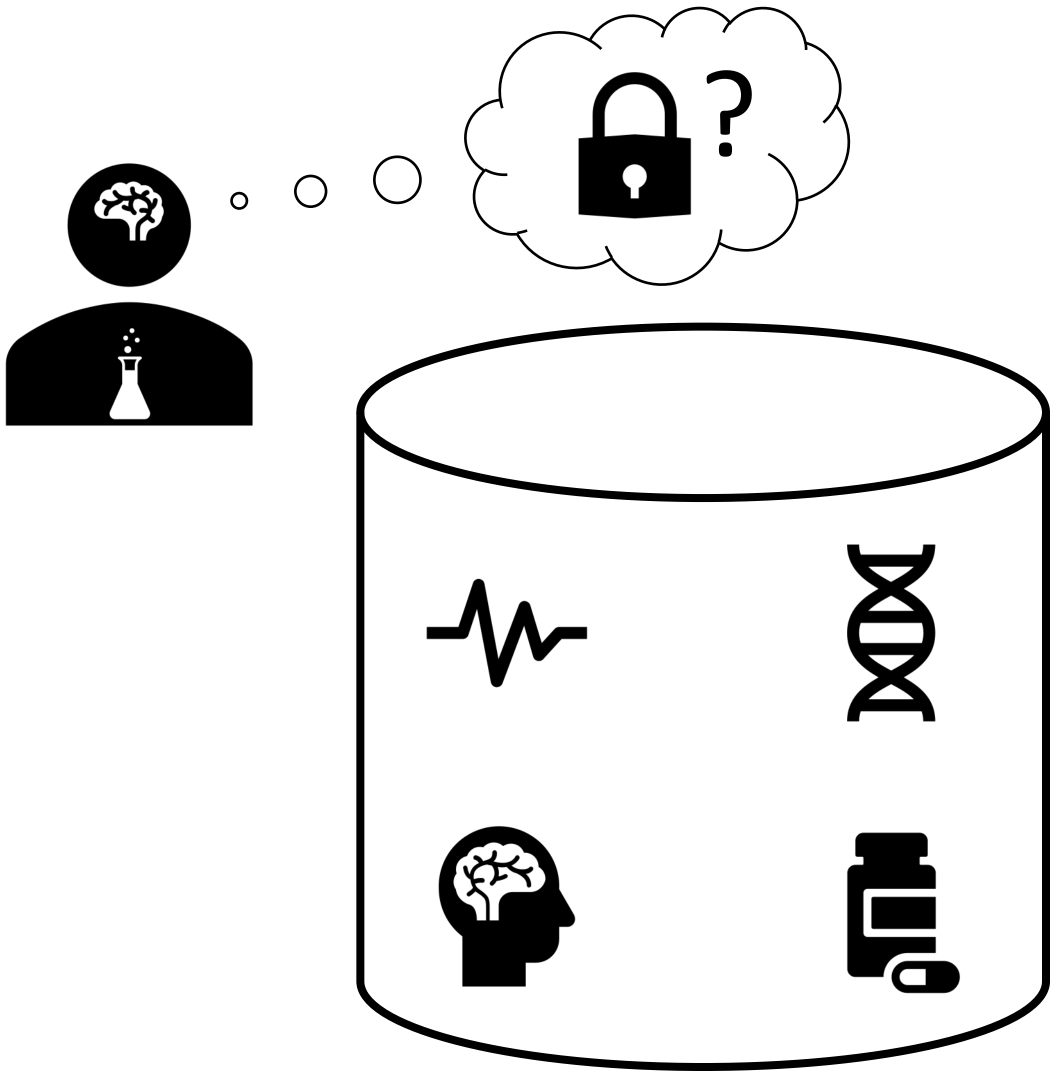




FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

miracum



# Sind die Daten jetzt anonym?

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

# k-Anonymität: Beispiel

miracum

| Identifier | Quasi-identifiers | | | Sensitive attributes |
|---|---|---|---|---|
| **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| Andrew | 37 | male | 86368 | Influenza |
| Berta | 51 | female | 87459 | Multiple Sclerosis |
| Claudia | 55 | female | 92360 | Asthma |
| Doris | 39 | female | 92660 | Hypertension |
| Eddy | 41 | male | 86399 | Influenza |
| Frances | 47 | female | 87509 | Diabetes |
| Gloria | 55 | female | 87730 | Arteriosclerosis |
| Henry | 35 | male | 73079 | Obesity |
| Ian | 43 | male | 79350 | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*



FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

# k-Anonymität: Population uniqueness

| Identifier | Quasi-identifiers | | | Sensitive attributes |
|---|---|---|---|---|
| **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| * | 37 | male | 86368 | Influenza |
| * | 51 | female | 87459 | Multiple Sclerosis |
| * | 55 | female | 92360 | Asthma |
| * | 39 | female | 92660 | Hypertension |
| * | 41 | male | 86399 | Influenza |
| * | 47 | female | 87509 | Diabetes |
| * | 55 | female | 87730 | Arteriosclerosis |
| * | 35 | male | 73079 | Obesity |
| * | 43 | male | 79350 | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*

# Anonymisierung in der EU-DSGVO

Erwägungsgrund 26:

*“[…] personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann.“*

# k-Anonymität: Generalisation & Suppression

| Identifier | Quasi-identifiers | | | Sensitive attributes |
|---|---|---|---|---|
| **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| * | 37 | male | 86368 | Influenza |
| * | 51 | female | 87459 | Multiple Sclerosis |
| * | 55 | female | 92360 | Asthma |
| * | 39 | female | 92660 | Hypertension |
| * | 41 | male | 86399 | Influenza |
| * | 47 | female | 87509 | Diabetes |
| * | 55 | female | 87730 | Arteriosclerosis |
| * | 35 | male | 73079 | Obesity |
| * | 43 | male | 79350 | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*

# k-Anonymität: Generalisation & Suppression

| Identifier | Quasi-identifiers | | | Sensitive attributes |
| --- | --- | --- | --- | --- |
| **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| * | 36 < x < 42 *(37)* | male | 86* | Influenza |
| * | 46 < x < 56 *(51)* | female | 87* | Multiple Sclerosis |
| * | 46 < x < 56 *(55)* | female | 92* | Asthma |
| * | 46 < x < 56 *(39)* | female | 92* | Hypertension |
| * | 36 < x < 42 *(41)* | male | 86* | Influenza |
| * | 46 < x < 56 *(47)* | female | 87* | Diabetes |
| * | 46 < x < 56 *(55)* | female | 87* | Arteriosclerosis |
| * | 34 < x < 44 *(35)* | male | 7* | Obesity |
| * | 34 < x < 44 *(43)* | male | 7* | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*

→ Datenqualität

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

# k-Anonymität: Generalisation & Suppression

miracum

| Identifier | Quasi-identifiers | | | Sensitive attributes |
|---|---|---|---|---|
| **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| * | 36 < x < 42 *(37)* | male | 86* | Influenza |
| * | 46 < x < 56 *(51)* | female | 87* | Multiple Sclerosis |
| * | 46 < x < 56 *(55)* | female | 92* | Asthma |
| * | 46 < x < 56 *(39)* | female | 92* | Hypertension |
| * | 36 < x < 42 *(41)* | male | 86* | Influenza |
| * | 46 < x < 56 *(47)* | female | 87* | Diabetes |
| * | 46 < x < 56 *(55)* | female | 87* | Arteriosclerosis |
| * | 34 < x < 44 *(35)* | male | 7* | Obesity |
| * | 34 < x < 44 *(43)* | male | 7* | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*



FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT
Universitätsklinikum Erlangen

# k-Anonymität: Äquivalenzklassen (k=2)

| | Ident. | Quasi-identifiers | | | Sensitive attributes |
|---|---|---|---|---|---|
| **Eq. Class** | **Name** | **Age** | **Sex** | **ZIP** | **Disease** |
| A | * | 36 < x < 42 | male | 86* | Influenza |
| | * | 36 < x < 42 | male | 86* | Influenza |
| B | * | 46 < x < 56 | female | 87* | Multiple Sclerosis |
| | * | 46 < x < 56 | female | 87* | Diabetes |
| | * | 46 < x < 56 | female | 87* | Arteriosclerosis |
| C | * | 46 < x < 56 | female | 92* | Asthma |
| | * | 46 < x < 56 | female | 92* | Hypertension |
| D | * | 34 < x < 44 | male | 7* | Obesity |
| | * | 34 < x < 44 | male | 7* | Lung Cancer |

*(Any resemblance to living persons is purely coincidental.)*

# Analyse der k-Anonymität am Standort Erlangen

# Abrechnungsdaten nach § 21 KHEntgG

- Standortübergreifend standardisiertes Format
- Erste Ausbaustufe: MIRACUM Forschungsdaten-Repo
- Enthält:
  - Interne KH-Nummer, Versicherungsnummer, …
  - Demografische Daten (Name, Alter, Geschlecht, PLZ, …)
  - Fallinformationen (Datum, Aufenthaltsdauer, …)
  - Diagnosen (ICD10-GM)
  - Prozeduren (OPS-301)
- Wahl der Quasi-Ident. für Analyse:
  - **Geschlecht**, **Alter, Diagnose**, **Diagnoseart**, **Prozedur**

# Klassifikationen

- **Diagnosen: ICD10-GM**

  Beispiel:

  **E11.81** (*Diabetes mellitus, Typ 2 : Mit […]*)

  - Kapitel-Ebene: **E00-E90** (*Endokrine, Ernährungs- und […]*)
  - Gruppen-Ebene: **E10-E14** (*Diabetes mellitus*)

- **Prozeduren: OPS-301**

  Beispiel:

  **8-010.3** (*Intravenöse, kontinuierliche Applikation von […]*)

  - Kategorie-Ebene: **8-01** (*Applikation von Medikamenten […]*)

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

miracum

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | | | |
| | 3 | 1 | | | |



FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT
Universitätsklinikum Erlangen

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

miracum

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | | | |
| | 2 | 1.748 | | | |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

miracum

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | | | |
| | 2 | 110.616 | | | |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | Both above | 1 | 5.835 |
| | 2 | 110.616 | | 2 | 3.453 |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | Both above | 1 | 5.835 |
| | 2 | 110.616 | | 2 | 3.453 |
| {sex,age,diagnosis} | | | Recoding to Group Level | 1 | 206 |
| | | | | 2 | 104 |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | Both above | 1 | 5.835 |
| | 2 | 110.616 | | 2 | 3.453 |
| {sex,age,diagnosis} | | | Recoding to Group Level | 1 | 206 |
| | | | | 2 | 104 |
| | | | Recoding to Chapter Lvl. | 1 | 14 |
| | | | | 2 | 10 |

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | Both above | 1 | 5.835 |
| | 2 | 110.616 | | 2 | 3.453 |
| {sex,age,diagnosis} | | | Recoding to Group Level | 1 | 206 |
| | | | | 2 | 104 |
| | | | Recoding to Chapter Lvl. | 1 | 14 |
| | | | | 2 | 10 |
| {sex,age,procedure} | 1 | 206.260 | | | |

# # Äquivalenzklassen für best. Attribute
## (2004-2017; 423.087 Patienten)

| Attribute Set | k | # Eq. Classes | Transformation | k | # Eq. Classes |
|---|---|---|---|---|---|
| {age} | < 3 | 0 | Age grouping by 10 yrs | < 105 | 0 |
| | 3 | 1 | | 105 | 1 |
| {diagnosis,type} | 1 | 3.351 | 3-digit ICD10 | 1 | 97 |
| | 2 | 1.748 | | 2 | 66 |
| {sex,age,diagnosis,type} | 1 | 287.733 | Both above | 1 | 5.835 |
| | 2 | 110.616 | | 2 | 3.453 |
| {sex,age,diagnosis} | | | Recoding to Group Level | 1 | 206 |
| | | | | 2 | 104 |
| | | | Recoding to Chapter Lvl. | 1 | 14 |
| | | | | 2 | 10 |
| {sex,age,procedure} | 1 | 206.260 | Recod. Cat. Lvl. | 1 | 50 |



FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen

# Zusammenfassung und Ausblick

# Zusammenfassung

- Allgemein hohe Population Uniqueness ohne Transformationen
- Mit steigender Anzahl an zu betrachtenden Attributen (Quasi-Identifikatoren), steigt Population Uniqueness
- Generalisierung von ICD10 und OPS zu höheren Hierarchieebenen reduziert Population Uniqueness (allerdings keine 2-Anonymität)
  - Möglicherweise machbar für Sub-Kohorten
  - Individuen ausschneiden (Volle Suppression?)

# Zusammenfassung

- Anonymisierung muss immer spezifisch für ein Projekt bzw. einen Datensatz vorgenommen werden
- Anonymisierung sorgt immer für Informationsverlust

**Anonymität vs. Datenqualität**

- Es ist bei gewisser Datenqualität nie unmöglich zu re-identifizieren, allerdings kann Aufwand gesteigert werden

**“Akademisches Risiko“ vs. Real-life Risiko**

# Ausblick

- **WICHTIG: Wahl der Quasi-Identifikatoren für Betrachtung realitätsnah gestalten!**
- Weitere Anonymisierungstechniken, z.B.:
  - **Utility constraints**
    - *G. Poulis et al., Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints*
  - **Global/local recoding**
    - *J. Xu et al., Utility-based anonymization using local recoding*
- Weitere Risiko-Modelle, z.B.:
  - **l-Diversität**
    - *A. Machanavajjhala, D. Kifer, J. Gerke et al., L-diversity: Privacy Beyond K-anonymity*
  - **t-Closeness**
    - *N. Li, T. Li, and S. Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*

# Dankeschön!

MEDICAL INFORMATICS INITIATIVE GERMANY
MEMBER OF TMF
SPONSORED BY THE Federal Ministry of Education and Research
FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG MEDIZINISCHE FAKULTÄT
Universitätsklinikum Erlangen

# Glossar

- **k-Anonymität**: Für jede Kombination von quasi-identifizierenden Attributwerten gibt es mindestens zwei Datensätze mit diesen Werten
- **l-Diversität**: Jede Äquivalenzklasse enthält sensible Attribute, deren Werte „ausreichend repräsentativ“ sind

# Paper Quellen


[1] F.K. Dankar, K. El Emam, A. Neisa, and T. Roffey, Estimating the re-identification risk of clinical data sets, BMC Med. Inform. Decis. Mak. 12 (2012). doi:10.1186/1472-6947-12-66.

[2] K. El Emam, Risk-based de-identification of health data, IEEE Secur. Priv. 8 (2010) 64–67. doi:10.1109/MSP.2010.103.

[3] T. Ganslandt, M. Boeker, M. Löbe, F. Prasser, J. Schepers, S.C. Semler, S. Thun, and U. Sax, Der Kerndatensatz der Medizininformatik-Initiative: Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene, Forum Der Medizin-Dokumentation Und Medizin-Informatik. 20 (2018) 17–21.

[4] T. Ganslandt, S. Mate, K. Helbing, U. Sax, and H.-U. Prokosch, Unlocking Data for Clinical Research – The German i2b2 Experience, Appl. Clin. Inform. 2 (2011) 116–127. doi:10.4338/ACI-2010-09-CR-0051.

[5] C. Haverkamp, T. Ganslandt, P. Horki, M. Boeker, A. Dörfler, S. Schwab, J. Berkefeld, W. Pfeilschifter, W.D. Niesen, K. Egger, M. Kaps, M.A. Brockmann, E. Neumaier-Probst, K. Szabo, M. Skalej, S. Bien, C. Best, H.-U. Prokosch, and H. Urbach, Regional Differences in Thrombectomy Rates: Secondary use of Billing Codes in the MIRACUM (Medical Informatics for Research and Care in University Medicine) Consortium, Clin. Neuroradiol. 28 (2018) 225–234. doi:10.1007/s00062-017-0656-y.

[6] M. Langarizadeh, A. Orooji, and A. Sheikhtaheri, Effectiveness of Anonymization Methods in Preserving Patients ' Privacy : A Systematic Literature Review, (2018) 80–87. doi:10.3233/978-1-61499-858-7-80.

[7] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.-U. Prokosch, and M. Sedlmayr, Towards Implementation of OMOP in a German University Hospital Consortium, Appl. Clin. Inform. 9 (2018) 54–61. doi:10.1055/s-0037-1617452.

[8] Medical Informatics Initiative, MII Core Data Set, (2017). http://www.medizininformatik-initiative.de/en/core-data-set (accessed October 27, 2018).

[9] G. Poulis, G. Loukides, S. Skiadopoulos, and A. Gkoulalas-Divanis, Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints, J. Biomed. Inform. 65 (2017) 76–96. doi:10.1016/j.jbi.2016.11.001.

[10] F. Prasser, F. Kohlmayer, and K.A. Kuhn, The importance of context: Risk-based de-identification of biomedical data, Methods Inf. Med. 55 (2016) 347–355. doi:10.3414/ME16-01-0012.

[11] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, M. Neumaier, K. Marquardt, H. Renz, H.-J. Rothkötter, C. Schade-Brittinger, P. Schmücker, J. Schüttler, M. Sedlmayr, H. Serve, K. Sohrabi, and H. Storf, MIRACUM: Medical Informatics in Research and Care in University Medicine, Methods Inf. Med. 57 (2018) e82–e91. doi:10.3414/ME17-02-0025.

[12] P. Samarati, Protecting Respondents' Identities in Microdata Release, IEEE Trans. Knowl. Data Eng. 13 (2001) 1–29.

[13] P. Samarati, and L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion., Proc. IEEE Symp. Res. Secur. Priv. (1998) 384–393. doi:http://dx.doi.org/10.1145/1150402.1150499.

[14] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative., Methods Inf. Med. 57 (2018) e50–e56. doi:10.3414/ME18-03-0003.

[15] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, Int. J. Uncertainty, Fuzziness Knowledge-Based Syst. 10 (2002) 571–588. doi:10.1142/S021848850200165X.

[16] L. Sweeney, k-Anonymity: a Model for Protecting Privacy, Int. J. Uncertainty, Fuzziness Knowledge-Based Syst. 10 (2002) 557–570. doi:10.1142/S0218488502001648.

[17] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, Utility-based anonymization using local recoding, in: Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '06, 2006: p. 785. doi:10.1145/1150402.1150504.




FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum Erlangen