



AnoMed – Anonymisierung für medizinische Anwendungen

Projektvorstellung

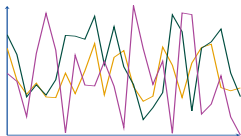
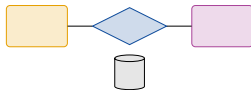
Malte Luttermann

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

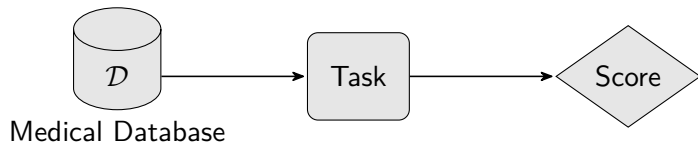
12. März 2024

Ziel des Projektes AnoMed

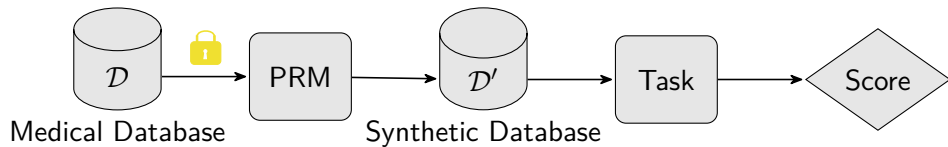
- ▶ Nutzung sensibler (personenbezogener) medizinischer Daten
 - ▶ Relationale Daten
 - ▶ Zeitreihen
 - ▶ Bilder
 - ▶ Videos



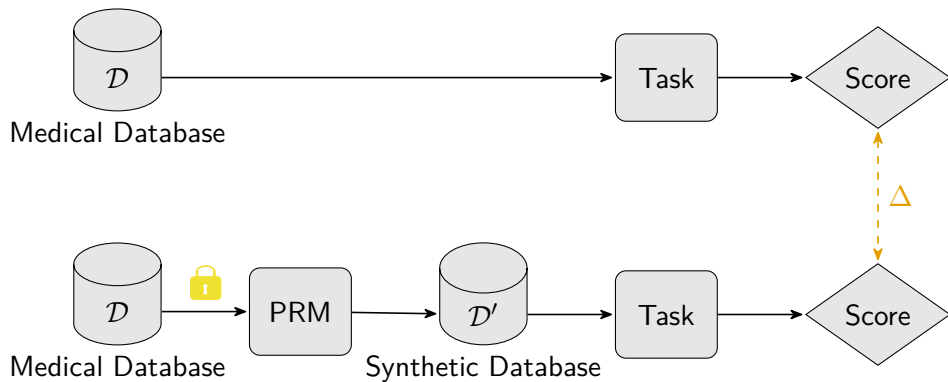
Privacy Challenge



Privacy Challenge

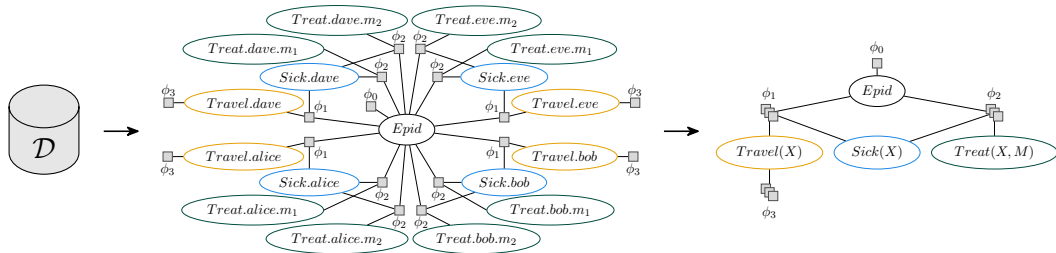


Privacy Challenge



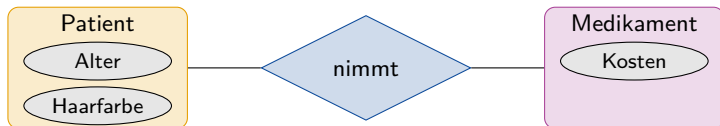
Anonymisierende Generierung von synthetischen Datensätzen

1. Lernen eines probabilistisch relationalen Modells (PRM)
2. Sampling neuer synthetischer Daten aus dem PRM



Lernen eines PRMs aus relationalen Daten

Beispieldatenbank



p_id	Alter	Haarfarbe	...	p_id	m_id	...	m_id	Kosten	...
<i>alice</i>	≥ 18	<i>hell</i>	...	<i>alice</i>	<i>m</i> ₁	...	<i>m</i> ₁	<i>high</i>	...
<i>bob</i>	≥ 18	<i>hell</i>	...	<i>alice</i>	<i>m</i> ₂	...	<i>m</i> ₂	<i>high</i>	...
<i>charlie</i>	≥ 18	<i>dunkel</i>	...	<i>bob</i>	<i>m</i> ₃	...	<i>m</i> ₃	<i>low</i>	...
<i>dave</i>	< 18	<i>dunkel</i>	...	<i>charlie</i>	<i>m</i> ₄	...	<i>m</i> ₄	<i>low</i>	...
<i>eve</i>	< 18	<i>hell</i>	...	<i>eve</i>	<i>m</i> ₅	...	<i>m</i> ₅	<i>high</i>	...
...	...	...	...	...	...	...	...	...	...

Lernen eines PRMs aus relationalen Daten

Clustering von Entitäten

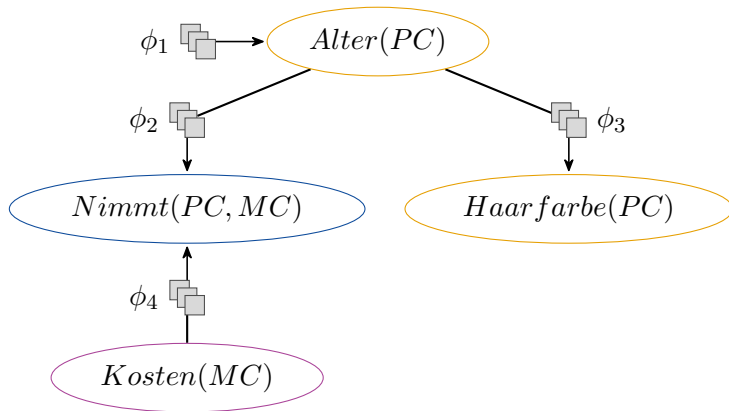
- ▶ Cluster bilden für jede Entität (hier: Patient, Medikament)
 - ▶ Logische Variable PC mit Domain $\mathcal{D}(PC) = \{pc_1, pc_2, pc_3, pc_4\}$
 - ▶ Logische Variable MC mit Domain $\mathcal{D}(MC) = \{mc_1, mc_2\}$

p_id	Alter	Haarfarbe	...	C	m_id	Kosten	...	C
alice	≥ 18	hell	...	pc_1	m_1	high	...	mc_1
bob	≥ 18	hell	...	pc_1	m_2	high	...	mc_1
charlie	≥ 18	dunkel	...	pc_2	m_3	low	...	mc_2
dave	< 18	dunkel	...	pc_3	m_4	low	...	mc_2
eve	< 18	hell	...	pc_4	m_5	high	...	mc_1
...	...	...	...	...	...	...	...	...

Lernen eines PRMs aus relationalen Daten

Lernen der Graphstruktur

- ▶ Knoten (parametrisierte Zufallsvariablen) für jedes Attribut und jede Relation
- ▶ Kanten über Unabhängigkeitstests auf Full Join der Tabellen ermitteln



Lernen eines PRMs aus relationalen Daten

Ermittlung der Potentiale

- ▶ Für jeden Parfaktor Vorkommnisse pro Clusterkombination zählen
 - ▶ z.B. in ϕ_1 Anzahl $Alter \geq 18$ und Anzahl $Alter < 18$ für $pc_1, \dots, pc_4$ zählen
- ▶ Cluster mit gleichen bzw. ähnlichen Vorkommnissen zusammenfassen

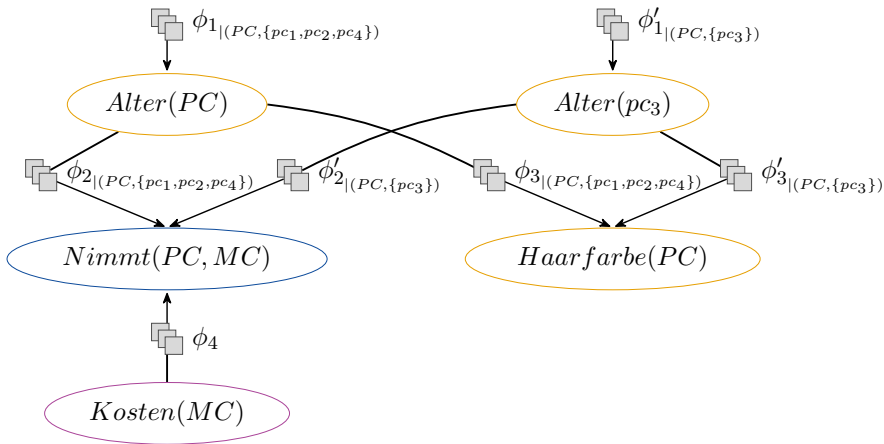
$Alter(pc_1)$	ϕ_1	$Alter(pc_2)$	ϕ_1	$Alter(pc_3)$	ϕ_1	$Alter(pc_4)$	ϕ_1
≥ 18	φ_1	≥ 18	φ_3	≥ 18	φ_5	≥ 18	φ_7
< 18	φ_2	< 18	φ_4	< 18	φ_6	< 18	φ_8

- ▶ z.B. $\phi_1(Alter(pc_1)) = \phi_1(Alter(pc_2)) = \phi_1(Alter(pc_4)) \neq \phi_1(Alter(pc_3))$

Lernen eines PRMs aus relationalen Daten

Ermittlung der Potentiale

- Cluster mit gleichen bzw. ähnlichen Vorkommnissen zusammenfassen
 - z.B. $\phi_1(\text{Alter}(pc_1)) = \phi_1(\text{Alter}(pc_2)) = \phi_1(\text{Alter}(pc_4)) \neq \phi_1(\text{Alter}(pc_3))$



Sampling neuer synthetischer Daten aus dem PRM

- ▶ Ein PRM definiert eine Wahrscheinlichkeitsverteilung
- ▶ Sampling aus der Wahrscheinlichkeitsverteilung möglich

