Institut für Wirtschaftsinformatik
Dr. Günter Karjoth
23. Juni 2015

Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN
Wirtschaft
Institut für Wirtschaftsinformatik IWI

Information Security in Health Conference, Rotkreuz

Schutz der Patientendaten durch Anonymisierung

Wann ist gut genug?

Dr. Günter Karjoth

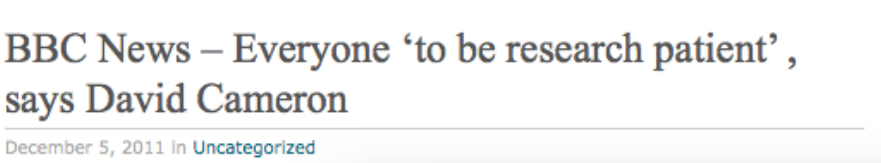
Auf dem Weg zum Dr. Algorithmus?

Potenziale von Big Data in der Medizin

Forscher greifen nach Patientendaten

Im Gesundheitswesen sei die Datenlage desolat, sagen Forscher von Public Health Schweiz. Jetzt fordern sie die Einführung eines Ausweises für Datenspender.

von **Katharina Bracher** | 15.9.2013 | **2 Kommentare**

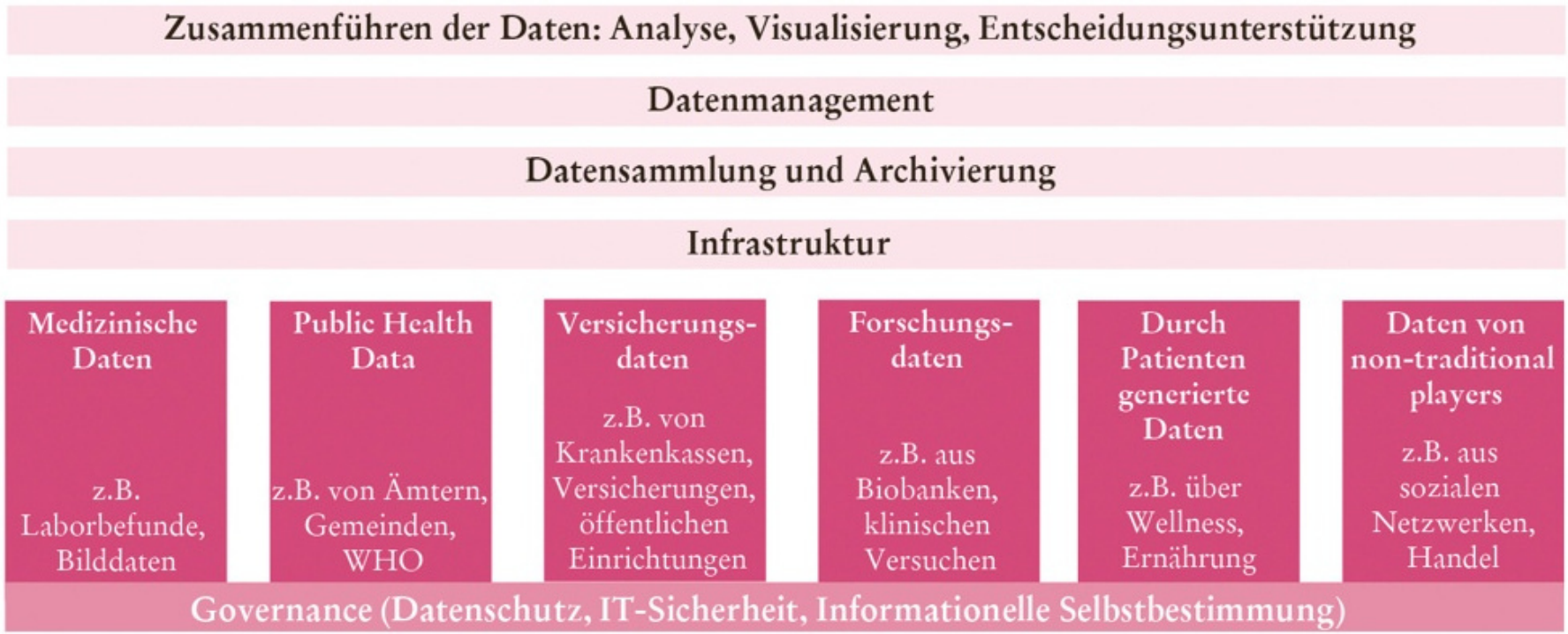


BBC News – Everyone 'to be research patient', says David Cameron

December 5, 2011 in Uncategorized

- Der Digitalisierungsgrad und damit die Menge an digitalen Daten in der Medizin wächst enorm.
- Immer mehr Patienten bringen (heute) ihre Daten selbst mit.
- Digitalisierung als Treiber für Innovationen.

Big-Data-Framework in der Medizin



Quelle: by-nc-nd/3.0/de/ Autor: Peter Langkafel für Aus Politik und Zeitgeschichte/bpb.de

Content

Die Personen «hinter» den Personendaten

- **Bestimmt** ist eine Person, wenn sich ihre Identität direkt aus der Information ergibt.
- **Bestimmbar** ist eine Person, wenn ihre identität durch die Kombination der Information mit anderen Informationen *ohne einen unverhältnismässigen Aufwand* feststellbar ist.

Beispiele:

- über Funktionsbezeichnungen («die Leiterin der Fachstelle für Gleichstellung der Stadt Zürich»)
- über einen Schlüssel (eine Fallnummer, eine Kundennummer usw.)

Besondere Personendaten

Strengere Regeln für

- religiöse, weltanschauliche, politische oder gewerkschaftliche Ansichten oder Tätigkeiten
- Gesundheit, Intimsphäre, Rassenzugehörigkeit oder ethnische Herkunft
- Sozialhilfemassnahmen
- administrative und strafrechtliche Verfolgungen oder Sanktionen.

Persönlichkeitsprofile

Zusammenstellungen von Informationen, die eine Beurteilung wesentlicher Aspekte der Persönlichkeit natürlicher Personen erlauben.

Informationen und Personendaten

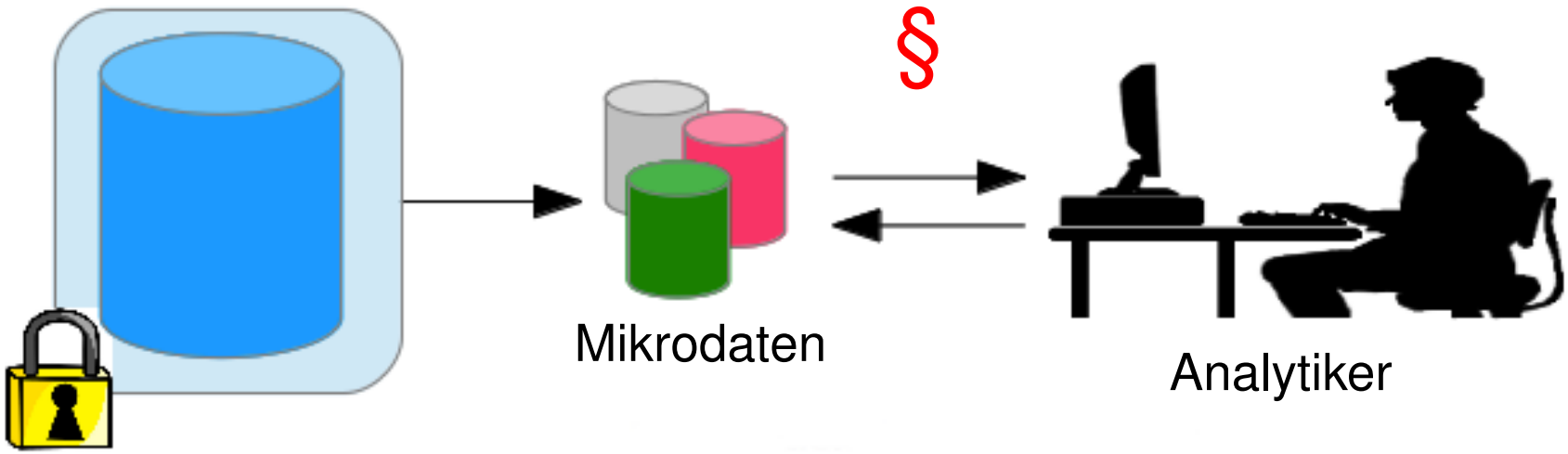
Als Informationen gelten alle Aufzeichnungen, die der Erfüllung einer öffentlichen Aufgabe dienen, und zwar unabhängig von ihrer Darstellungsform (z.B. Handnotiz, Eintrag in einer Datenbank, Bild) und ihrem Informationsträger (z.B. Papier, Festplatte oder Memorystick).

☞ Personendaten sind auch Informationen, aber sie beziehen sich auf eine bestimmte oder auch nur bestimmbare Person.

Keine Personendaten sind

- Sachdaten, also Angaben, die keinen Bezug zu einer Person aufweisen und
- Daten, bei denen der Personenbezug dauerhaft beseitigt worden ist (anonymisierte Daten).

Offline Data Publishing



☞ Daten sind in der Granularität von Individuen!

Content

Herausforderungen im Testdatenmanagement

- Ihre Daten in Entwicklungs- und Testumgebungen sind grösstenteils 1:1 Kopien der Produktionsumgebung.
- Sie planen, Entwicklung bzw. Test ausserhalb Ihres Unternehmens durchzuführen.
- Sie haben vertrauliche Daten, welche auch ausserhalb der Produktionsumgebungen geschützt werden müssen
- Ihre Entwickler und Tester fordern realitätsnahe Daten.
- Ihre Compliance Abteilung verlangt die Einhaltung der vorgeschriebenen Datenschutzgesetze auch in Entwicklung und Test.

Schutz der Produktionsdatenbank

- realistisch & akkurat, format- und linksicher
- dynamic vs. persistent data masking

☞ Unternehmen haben sensitive Informationen über Kunden und Mitarbeiter!

- Einfaches Masking (engl. Redaction) erweitert mit Generalisierung
- Produkte
  - IBM InfoSphere Guardium Data Redaction (manual/automatic)
  - Oracle
  - ...

Die fünf Gesetze des Data Masking

Datenmaskierung ist ein Verfahren zum Erzeugen einer strukturell ähnlichen, aber unechten Version der Daten einer Organisation.

- Die Maskierung darf nicht umkehrbar sein.
- Das Ergebnis muss representativ für die Quelldaten sein.[1]
- Referentielle Integrität muss bewahrt bleiben.
- Maskiere nicht-sensitive Daten, falls sie verwendet werden können um sensitive Daten (wieder) herzustellen.
- Maskierung muss ein wiederholbarer Prozess sein.

[1] Nationale ID, AHV, Kreditkarte, Postleitzahl, Telefon, ...

Anonymisierung: Datentransformationsmethoden

Perturbative Ansätze

- Bewahren Aggregatstatistik (Mittelwert, Korrelationskoeffizient, . . .), z. B. durch
  - Hinzufügen von Rauschen,
  - Daten vertauschen,
  - Micro-aggregation,
  - Runden

  ☞ verfälschen die Daten

Nicht-perturbative Ansätze

- Verändern die Granularität der veröffentlichten Daten, z. B. durch
  - Generalisierung

    PLZ (24103 $\rightarrow$ 241**), Geschlecht (M $\rightarrow$ *), Alter (24 $\rightarrow$ [20–29])
  - Unterdrückung (“Ausreisser")

  ☞ Keine Verfälschung der Daten!

Als anonym veröffentlichte medizinische Daten

SSN	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94139	geschieden	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	verheiratet	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck
		03/18/63	M	94138	verheiratet	Kurzatmigkeit
		09/13/64	W	94141	verheiratet	Kurzatmigkeit
		09/07/64	W	94141	verheiratet	Fettsucht
		05/14/61	M	94138	ledig	Brustschmerzen
		05/08/61	M	94138	ledig	Fettsucht
		09/15/61	W	94142	Witwe	Kurzatmigkeit

Wählerliste

Name	Adresse	Stadt	PLZ	Geb.	Geschl.	Partei
Sue. J. Carlson	900 Market St.	San Francisco	94142	9/15/61	W	Demokrat

Als anonym veröffentlichte medizinische Daten

SSN	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94139	geschieden	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	verheiratet	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck
		03/18/63	M	94138	verheiratet	Kurzatmigkeit
		09/13/64	W	94141	verheiratet	Kurzatmigkeit
		09/07/64	W	94141	verheiratet	Fettsucht
		05/14/61	M	94138	ledig	Brustschmerzen
		05/08/61	M	94138	ledig	Fettsucht
		09/15/61	W	94142	Witwe	Kurzatmigkeit

Wählerliste

Name	Adresse	Stadt	PLZ	Geb.	Geschl.	Partei
Sue. J. Carlson	900 Market St.	San Francisco	94142	9/15/61	W	Demokrat

Content

Bewertung der Attribute

Quasi-Identifikator

- Eine Untermenge der Attribute, deren Wertekombination für eine Person charakteristisch sein könnte (bestimmbar macht).

Sensitive Attribute

- Attribute, welches nicht mit einer Person verknüpfbar sein sollen.

AHV	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94141	ledig	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	Witwer	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck

Ein Mass für den Schutz von Personendaten

k-Anonymität

- k Datensätze bilden eine Äquivalenzklasse (bzgl. des Quasi-Identifikators)
- schützt mit einer Konfidenz von $1/k$ vor einer 'korrekten' Verknüpfung einer Person mit ihren sensitiven Attributen

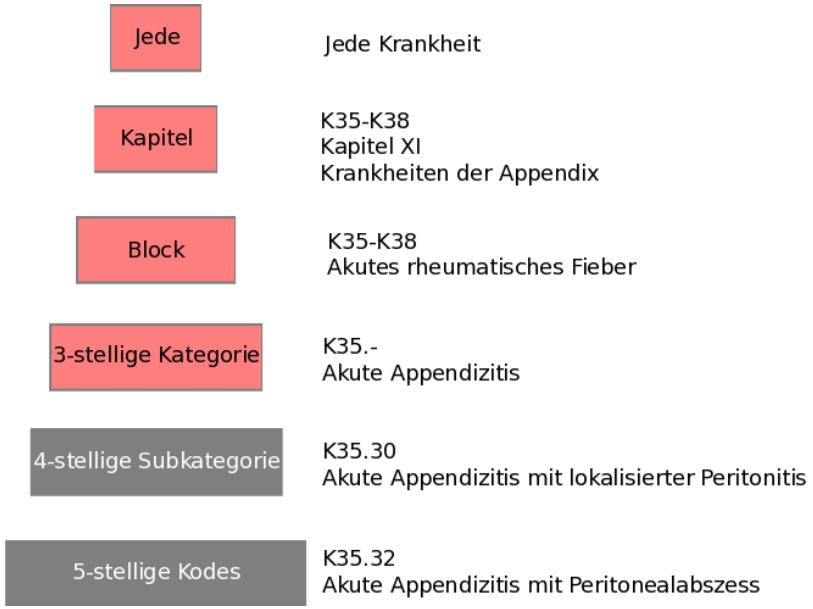
Ein Tabelle ist k-anonym, wenn jedes Tupel von mindestens $k-1$ anderen Tupeln (bis auf die sensitiven Attribute) nicht unterscheidbar ist.

Credit: Manolis Terrovitis

Techniken

Generalisierung

- Jedes Attribut hat einem Wertebereich
- Abbildung zwischen jedem Wertebereich und seiner Generalisierung



ICD-10 Klassifikation

Unterdrückung

- entfernt Daten aus der Tabelle
- meistens auf Datensatzebene (Zeile) angewendet
  - ☞ Unterstützen den Generalisierungsprozess, wenn eine beschränkte Anzahl von Tupeln mit weniger als k Exemplaren eine grosse Generalisierung erfordern würden.

Eine Tabelle

Geburtstag	Geschl.	PLZ	Ehestand	Krankheit
09/27/64	W	94139	geschieden	Bluthochdruck
09/30/64	W	94139	geschieden	Fettsucht
04/18/64	M	94139	verheiratet	Brustschmerzen
04/15/64	M	94139	verheiratet	Fettsucht
03/13/63	M	94138	verheiratet	Bluthochdruck
03/18/63	M	94138	verheiratet	Kurzatmigkeit
09/13/64	W	94141	verheiratet	Kurzatmigkeit
09/07/64	W	94141	verheiratet	Fettsucht
05/14/61	M	94138	ledig	Brustschmerzen
05/08/61	M	94138	ledig	Fettsucht
09/15/61	W	94142	Witwe	Kurzatmigkeit

. . . und ihre minimale Generalisierung

Geburtsjahr	Geschl.	PLZ	Ehestand
64	-	941**	-
64	-	941**	-
64	-	941**	-
64	-	941**	-
63	-	941**	-
63	-	941**	-
64	-	941**	-
64	-	941**	-
61	-	941**	-
61	-	941**	-
61	-	941**	-

Geburtsjahr	Geschl.	PLZ	Ehestand
$[61-64]$	W	9413*	nicht-ledig
$[61-64]$	W	9413*	nicht-ledig
$[61-64]$	M	9413*	nicht-ledig
$[61-64]$	M	9413*	nicht-ledig
$[61-64]$	M	9413*	nicht-ledig
$[61-64]$	M	9413*	nicht-ledig
$[61-64]$	W	9414*	nicht-ledig
$[61-64]$	W	9414*	nicht-ledig
$[61-64]$	M	9413*	ledig
$[61-64]$	M	9413*	ledig
$[61-64]$	W	9414*	nicht-ledig

☞ Die Berechnung einer optimalen k-Anonymität Tabelle ist ein NP-schweres Problem, unabhängig von der Granularitätsstufe.

Mögliche Angriffe

Homogenität

Urs

PLZ	Alter
74678	26

Hintergrundwissen

Satoshi (Japaner)

PLZ	Alter
74673	36

Eine 3-anonyme Patiententabelle

PLZ	Alter	Gehalt	Krankheit
746**	2*	20K	Herzerkrankung
746**	2*	30K	Herzerkrankung
746**	2*	40K	Herzerkrankung
7490*	≥ 40	50K	Gastritis
7490*	≥ 40	100K	Grippe
7490*	≥ 40	70K	Bronchitis
746**	3*	60K	Herzerkrankung
746**	3*	80K	Krebs
746**	3*	90K	Krebs

k-Anonymität kann versagen, falls

- es den sensitiven Werten in einer Äquivalenzklasse an Vielfalt mangelt, oder
- der Angreifer Hintergrundwissen besitzt.

Prinzip der *l*-Diversität

Eine Äquivalenzklasse (Block) ist l-divers falls es mindestens l "wohl-vertretene" Werte des sensitiven Attributes S enthält.

Eine Tabelle is l-divers falls jeder Block l-divers ist.

☞ der Angreifer benötigt $l-1$ "damaging pieces", um eine positive Offenlegung zu erreichen.

Alter	Geschlecht	Krankheit
$[26-27]$	M	Grippe
$[26-27]$	M	Grippe
$[23-25]$	*	Erkältung
$[23-25]$	*	Diabetes
22	M	Grippe
22	M	Krebs

$k = 2$

Alter	Geschlecht	Krankheit
$[25-27]$	M	Grippe
$[25-27]$	M	Grippe
$[25-27]$	M	Erkältung
$[22-24]$	*	Diabetes
$[22-24]$	*	Grippe
$[22-24]$	*	Krebs

k = 3, $E \geq log(1.9)$

Offenlegung sensitiver Attribute

Ähnlichkeitsangriff

Regula

PLZ	Alter
7468	26

Schlussfolgerung

→ Regulas Gehalt ist im Bereich [20k,40k], was relativ wenig ist.

→ Regula hat eine magen-bezogene Krankheit.

Eine 3-diverse Patiententabelle

PLZ	Alter	Gehalt	Krankheit
746*	2*	20K	Magengeschwür
746*	2*	30K	Gastritis
746*	2*	40K	Magenkrebs
804*	≥ 40	50K	Gastritis
804*	≥ 40	100K	Grippe
804*	≥ 40	70K	Bronchitis
86**	3*	60K	Bronchitis
86**	3*	80K	Lungenentzündung
86**	3*	90K	Magenkrebs

☞ *l*-Diversität erfasst nicht die Semantik von sensitiven Werten!

☞ *t*-closeness

Vergleich der Anonymitätsmasse

Alter	Geschl.	Krankheit
$[26-27]$	M	Grippe
$[26-27]$	M	Grippe
$[23-25]$	*	Erkältung
$[23-25]$	*	Diabetes
22	M	Grippe
22	M	Krebs

$k = 2$

Alter	Geschl.	Krankheit
$[25-27]$	M	Grippe
$[25-27]$	M	Grippe
$[25-27]$	M	Erkältung
$[22-24]$	*	Diabetes
$[22-24]$	*	Grippe
$[22-24]$	*	Krebs

$k = 3$

Alter	Geschl.	Krankheit
$[22-27]$	*	Grippe
$[22-27]$	*	Grippe
$[22-27]$	*	Erkältung
$[22-27]$	*	Diabetes
$[22-27]$	*	Grippe
$[22-27]$	*	Krebs

$k = 6$

- Es ist schwer *l*-Diversität zu erzielen, wenn einer der sensitiven Werte sehr geläufig ist; z. B. 90% haben “Herzprobleme".
- Sind einige positive Offenlegungen akzeptierbar, könnte man weniger zurückhaltend sein.

  ☞ Güterabwägung zwischen Data Utility und Privacy

Content

Daten-Representation

- Relationale Daten
  - Registrierungs- und demographische Daten
- Mengenwertige Daten
  - Abrechnungen
- Sequentielle Daten
  - DNA
- Trajektorien (Bahnkurven)
  - Ortsdaten von Mobiltelefonen
- Graphen
  - Soziale Netzwerke
- Text
  - Klinische Aufzeichnungen, Tweets

Electronic Medical Records			
Name	Geburtsjahr	ICD	DNA
Beat	1955	493.00, 185	C . . . T
Vreni	1943	185, 157.3	A . . . G
Vreni	1943	493.01	C . . . G
Ursula	1965	493.02	C . . . G
Urs	1973	157.9, 493.03	G . . . C
Urs	1973	157.3	A . . . T

19 Jahre alter Mann mit Vorgeschichte Ekzem im Kleinkindalter, jetzt sporadische lokale Beschwerden im Mund nach Erdnussverzehr und Rhinokonjunktivitis während der Pollensaison.

Einzigartigkeit von Daten

Wieviele persönliche Daten sind notwendig um jemand re-identifizieren zu können:

- (Geburtsjahr, Geschlecht, 3-stellige PLZ)
  $\rightarrow$ 0.04% der amerikanischen Bevölkerung
- (Geburtsdatum, Geschlecht, 5-stellige PLZ)
  $\rightarrow$ 63–87 % der amerikanischen Bevölkerung
- 2 spatio-temporale Punkte $\rightarrow$ 50%
- 4 spatio-temporale Punkte $\rightarrow$ 95%
- 2 ICD Nummern $\rightarrow > 90\%$

Text De-identification

Klinische Vorgeschichte

77 year old female with a history of B-cell lymphona (Marginal zone, SH-02-22222, 6/22/01).

Flow cytometry and molecular diagnostics drawn.

- Finde die persönlichen Identifikatoren (z. B. Name, Record#, AHV).
- Ersetze oder entferne die gefundenen persönlichen Identifikatoren.

☞ Bewahre die Integrität der Information während die persönlichen Identifikatoren faktisch verborgen sind.

Techniken

- White lists (Worte mit grosser Häufigkeit bleiben an ihrem Originalplatz)
- Regel- und Wörterbuch-basiert (Pattern Matching)
- statistisches Lernen

☞ Benötigt eine konsistente Ersetzungsstrategie!!

Quelle: J. Gardner & L. Xiong: HIDE: An integrated system for Health Information DE-identification. 2006.

t-Plausibilität

Ein Einwohner von Aarau kaufte Marihuana gegen lumbale Schmerzen, verursacht durch Leberkrebs.

Ein Einwohner von ~~Aarau~~ kaufte ~~Marihuana~~ gegen ~~lumbale Schmerzen~~, verursacht durch ~~Leberkrebs~~.

t-Plausibilität verallgemeinert sensitive Terme zu semantisch ähnlichen Termen, z. B. “Tuberkulose” $\rightarrow$ “Infektion”.

Ist eine Wortontologie und ein Grenzwert t gegeben, kann der gesäuberte Text mindestens $t-1$ anderen Texten zugeordnet werden.

Ein Einwohner von Kantonshauptstadt kaufte Droge gegen Schmerzen, verursacht durch Karzinom.

Content

Resümee

Der Wert von Daten mit Personenbezug erschöpft sich nicht schon in ihrer ersten Verwendung. Aber wie können sie Dritten sicher zugänglich gemacht werden?

☞ Datenschutzgefahren: Aufdeckung der Identität und/oder sensitiver Attribute

Eine Anonymisierung von Daten gibt keine (strikte) Garantie der Anonymität!

- *k*-Anonymität – Schutz gegen Verknüpfung von Identitäten
- *l*-Diversität – Schutz gegen die Offenlegung von Attributen.

☞ In unstrukturierten Daten ist es noch schwieriger “Quasi-Identifikatoren” zu benennen.

Wird die Anonymitätsgarantie verstärkt, verringert sich die Datenqualität: Es benötigt eine Güterabwägung zwischen Nutzwert und Datenschutz.

☞ Was sind geeignete Parameter für den Schutz und dem Nutzwert?

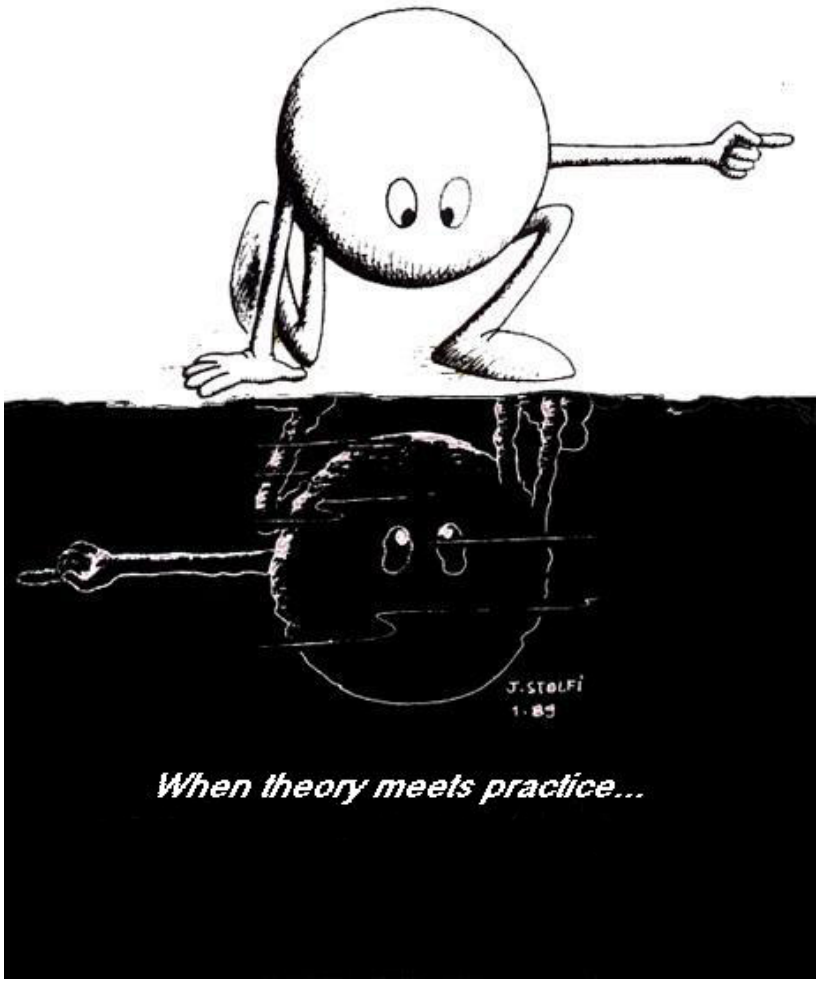
Zur Erinnerung!

Werden Daten offengelegt, spielt es keine Rolle, wie sensibel die Daten für uns sind, sondern wie charakteristisch. Es ist das Letztere, der den nötigen Aufwand bestimmt, um sie mit anderen Daten zu verknüpfen, um damit unsere Identität aufzudecken.

Abgesehen von Kombinationen von demografischen Daten, sind einige der möglichen Dinge, die sie eindeutig identifizieren können, unter anderem:

- Diagnosecodes
- Laborresultate
- sportliche Tätigkeiten
- oder sogar ihre soziales Netz.

Vielen Dank für ihre Aufmerksamkeit!



Credit: Jorge Stolfi

Hochschule Luzern

Competence Center Information Security

Dr. Günter Karjoth

Forschungsdozent

Zentralstrasse 9, CH-6002 Luzern

T: +41 41 228 99 78

guenter.karjoth@hslu.ch