

Kardkovács Zsolt Tivadar

Adatbázisok

vendégelőadás

Adatvédelem és anonimizálás adatbázisokban

- **Jogi környezet**

- Miért kell? Miért, kell?
- Mikor és mire kötelezettek a szereplők
- Korlátok és lehetőségek

- **Matematikai probléma**

- **Technológiai megfelelési lehetőségek**

- Stratégiák
- Problémák és nyitott kérdések
- Adatelemzés-orientált adatvédelem

*„In God we trust.
All others must bring data..”
(W. Edwards Deming)*





Rövid bemutatkozó

- Kardkovács Zsolt (aka Iodoktor) – informatikus (BME), Ph.D.
- Egyetemi (ex)oktató és ipari K+F-vezető – adatelemző / gépi látás szakértő
- kardkovacs (at) u1research.org

Tapasztalatok

- Zala Megyei Kórház – orvosi kutatás nagy, érzékeny adathalmazon
- GYEMSZI – országos adattárház beszerzésének előkészítése
- Etikus adatbányászati rendszer és elv kidolgozása
- Sportadat-elemzői adatbázis és kiválasztási rendszer
- Viselkedés-elemzés a gyakorlatban

Kardkovács Zsolt Tivadar

Adatbázisok

vendégelőadás

Adatvédelem és anonimizálás adatbázisokban

- Jogi környezet
 - Miért kell? Miért, kell?
 - Mikor és mire kötelezettek a szereplők
 - Korlátok és lehetőségek
- Matematikai probléma
- Technológiai megfelelési lehetőségek
 - Stratégiák
 - Problémák és nyitott kérdések
 - Adatelemzés-orientált adatvédelem

*„In God we trust.
All others must bring data..”
(W. Edwards Deming)*





Miért foglalkozunk GDPR-ral?

Mindenki erről beszél, akkor foglalkozzunk vele...

- **GDPR (General Data Protection Regulation)**
 - Európai Parlament és a Tanács 2016/679 **rendelete** (2018. május 25.)
 - Jelenleg érvényes: 95/46/EK **irányelv**(!)
 - 2011. évi CXII. **törvény**

Mi változott? (csak a legfontosabbak...)

- Fogalmak pontosításra kerültek
- Pénzbüntetés mértéke
- Adatátadásra vonatkozó szabályok
- Szervezeti és működési követelmények kerültek előírásra
- Adatvédelmi biztos pozíciója
- Bejelentési kötelezettség



Üzleti igények

- **Vásárló (igényeinek) megismerése**
- **Vásárló (legális) manipulációja vásárlás ösztönzésére**
- **Szokásos igények**
 - Üzlet profitmaximalizálása
 - Üzleti kockázatok csökkentése

Egyéni, személyes igények

- **Lojalitás, „helyes” viselkedés elismerése**
- **Magánélethez, a közösség kizárásához való jog**
- **Az egyén tiszteletének joga**
- **Az „újrakezdés” lehetősége**

Jogalkotói igények

- **A pénzforrásnak, finanszírozónak megfelelni**
- **A választónak megfelelni**



Miért kell az adatvédelem? Miért, kell?

Az ügyfél egy befolyásolási kísérlet tárgya: feltételezzük, hogy ...

- **...az egyén védtelen**
 - Képzettség és tapasztalat nélkül nem veszi észre a manipulációt
 - Ha észre is veszi, nem feltétlenül van tisztában a korlátaival
 - És ha ismeri is ezeket, nem feltétlenül ismeri a választási lehetőségeit
- **...az egyéni haszonszerzés érdeke mindig felülírja a mások lényeges érdekeit**
 - Az üzleti környezet mindig elmegy a falig a profitmaximalizálásért
 - De legtöbbször nem hágja át az etikai-jogi akadályokat
- **...az egyén naiv és felelőtlen**
 - Magatartása jellemzően naiv (aka. mohó) algoritmust követ
 - Az ideális ember nyitott a környezetére, ezért maga nyitott a világ felé
 - ...de a világ már nem a szemhatárig terjed

A vezetők és az „értelmiség” etikai és sok esetben jogi felelőssége is, hogy védje az egyéneket a „túlzott” befolyásolástól



Mi szeretnénk elérni az adatvédelemmel?

Az egyén nézőpontjából

- **Minden ember (nagyobb)egyenlő**
 - Minden vásárló egyforma
 - ...bár pozitívan diszkriminálható, ha ezek köre érdemfüggő(!)
- **Megbélyegzés tiltása**
 - Előítélet (terjedésének) tiltása
 - Az egyén mérlegelése (és nem mint csoport része)
 - De az újrakezdés lehetőségével – a múlt „elfelejthető”
- **A kihasználásának tilalma**
 - Pillanatnyi helyzetéből adódóan pl. visszaéléssel
 - Ismert gyengeségeinek kijátszásával
 - Tiltott befolyásolással
- **...de a jog szerinti bűnös bűnhődjék,**
- **...és ne korlátozzon minden más alapvető jogot**



Fogalmak (2011. évi CXII. törvény és ET 2016/679 – a GDPR – alapján)

▪ Érintett

- „**bármely** meghatározott, személyes **adat alapján azonosított vagy** - közvetlenül vagy közvetve - **azonosítható természetes személy**”

▪ Személyes adat

- „**az érintettel kapcsolatba hozható adat** [...], valamint az adatból levonható, az érintettre vonatkozó **következtetés**.”

▪ Különleges adat

- „a faji eredetre, a nemzetiséghez tartozásra, a politikai véleményre vagy pártállásra, a vallásos vagy más világnézeti meggyőződésre, az érdek-képviselői szervezeti tagságra, a szexuális életre vonatkozó személyes adat”
- „az egészségi állapotra, a kóros szenvedélyre vonatkozó személyes adat, valamint a bűnügyi személyes adat”

▪ Adatkezelés

- „az alkalmazott eljárástól függetlenül **az adaton végzett bármely művelet** vagy a műveletek összessége [...]”



Esetek a nem túl régi múltból (10+ millió felhasználó)

- **„Anonim” census adatok**
 - US választási és egészségügyi adatbázis reform (1995)
 - Hollandiai nyilvántartás
- **Publikált adatok, nem ismert összefüggések**
 - Netflix (2005), AOL (2006)
- **Biztonsági rés**
 - Heartland Payment (2008), Apple (2015), JP Morgan (2016), Equifax (2017)
- **Nem felügyelt adatbázis-elemek (kontra-audit!)**
 - CardSystem (2006)
- **Nem rendeltetésszerű felhasználása az eszközöknek**
 - FriendFinder Network (2016), Cambridge Analytica (2017)
- **Másodlagos ellenőrzés hiánya**
 - Anthem (2016)

Magyar esetek?



A jogi kérdés éles

Ha az fontos közérdek alapján
indokolt... eltérhetnek a különleges
adatok feldolgozásának tilalmától

Közérdek
(államháztartás)

Magyarország
biztosítja a ... kutatás ...
szabadságát

Kutatási érdek

„Érintettel
kapcsolatba
hozható adat”

Privát szféra
védelmének joga

Egészséghez való jog

Mindenkinek joga van
személyes adatai védelméhez

Emberi lény érdekei
megelőzik a
társadalom... érdekeit

Jogi-technikai ökölszabályok

Beleegyezés
(explicit, konkrét)

Tájékoztatási
kötelezettség

Archiválás

Ellenőrizhetőség

Tárolás,
megsemmisítés

Korlátozott
hozzáférés

Minimalitás

(szükségesség, arányosság)

Anonimizálás

Felelősség



Kardkovács Zsolt Tivadar

Adatbázisok

vendégelőadás

Adatvédelem és anonimizálás adatbázisokban

- Jogi környezet
 - Miért kell? Miért, kell?
 - Mikor és mire kötelezettek a szereplők
 - Korlátok és lehetőségek
- Matematikai probléma
- Technológiai megfelelési lehetőségek
 - Stratégiák
 - Problémák és nyitott kérdések
 - Adatelemzés-orientált adatvédelem

*„In God we trust.
All others must bring data..”
(W. Edwards Deming)*





Az elvi probléma

Kulcsfogalom: beazonosítás (re-identification)

Beazonosítás

- Beazonosításnak nevezzük azt a leképezést, folyamatot, eljárást, amelynek segítségével a rendelkezésre álló adatokból az érintettek egy nem üres halmaza egyértelmű összefüggésbe hozható a tárolt, védendő adatokkal.

Anonimitás

- Egy reláció(!) anonim, ha abból kiindulva nem feljogosított adatkezelő számára egyetlen érintett sem beazonosítható. Matematikailag:

$r(R)$, ahol $R(A_1, K, A_N)$

I az érintettek egy halmaza,

A_i az érintettel összefüggésbe hozható jellemző

$Q \subseteq \{A_1, K, A_N\}$

$f_m : I \rightarrow r(Q)$

$f_d : r(Q) \rightarrow I$

$\neg \exists p \in I : f_d(f_m(p)) = p$

Q-t kvázi azonosítónak nevezzük

Gyakorlati probléma

avagy „a ránk úgysem vonatkozik a jogszabály...”

személyjellemzők

védett adat

ROWID	Név	Született	Nem	Irányítószám	Betegség
1	Anita	76-01-21	Nő	1107	Arthritis
2	Béla	86-03-24	Férfi	1107	Appendicitis
3	Cili	76-02-27	Nő	1117	Osterosis
4	Dénes	76-01-21	Férfi	1117	Cancer
5	Emese	86-03-24	Nő	1127	Malaria
6	Ferenc	76-02-27	Férfi	1127	Hepatitis

SQL lekérdezés eredményeképpen létrejövő
reláció is vizsgálandó!

Gyakorlati probléma

avagy „a ránk ügysem vonatkozik a jogszabály...”

átkeresztelünk

személyjellemzők

védett adat

ROWID	Név	Született	Nem	Irányítószám	Betegség
1	A	76-01-21	Nő	1107	Arthritis
2	B	86-03-24	Férfi	1107	Appendicitis
3	C	76-02-27	Nő	1117	Osterosis
4	D	76-01-21	Férfi	1117	Cancer
5	E	86-03-24	Nő	1127	Malaria
6	F	76-02-27	Férfi	1127	Hepatitis

OK, átkeresztelünk, nem látjuk közvetlenül, hogy
kiről van szó, de akkor most kit is kezelünk?

Gyakorlati probléma

avagy „a ránk úgysem vonatkozik a jogszabály...”

Most nincs név, akkor a probléma meg van oldva?

személyjellemzők

védett adat

ROWID	Született	Nem	Irányítószám	Betegség
1	76-01-21	Nő	1107	Arthritis
2	86-03-24	Férfi	1107	Appendicitis
3	76-02-27	Nő	1117	Osterosis
4	76-01-21	Férfi	1117	Cancer
5	86-03-24	Nő	1127	Malaria
6	76-02-27	Férfi	1127	Hepatitis

Eseti funkcionális függésről mikor tudjuk, hogy érdemi-e?

Ebben a relációban kulcs tulajdonságú, a valóságban pedig kvázi azonosító

Gyakorlati probléma

avagy „a ránk úgysem vonatkozik a jogszabály...”

OK, nem kell minden ebből... most jó?

töröltünk

személyjellemzők

védett adat

ROWID	Született	Nem	Irányítószám	Betegség
1	76-*	Nő	11*	Arthritis
2	86-*	*	11*	Appendicitis
3	76-*	Nő	11*	Osterosis
4	76-*	Férfi	11*	Cancer
5	86-*	Nő	11*	Malaria
6	76-*	Férfi	11*	Hepatitis

A védett adat is azonosíthat(!) ☹
Sőt, t-próba alapján is azonosíthatunk!

Gyakorlati probléma

avagy „a ránk úgysem vonatkozik a jogszabály...”



Adj egy kis időt...!

személyjellemzők

védett adat

ROWID	Született	Nem	Irányítószám	Betegség	Kezelés napja
1	76-*	Nő	11*	Arthritis	2010-03-10
2	86-*	*	11*	Appendicitis	2010-03-10
3	76-*	Nő	11*	Osterosis	2010-03-20
4	76-*	Férfi	11*	Cancer	2010-03-20
5	86-*	Nő	11*	Malaria	2010-03-30
6	76-*	Férfi	11*	Hepatitis	2010-03-30



Vigyázz!!
Idősor is azonosít!

Háttérinformációt is feltételezni kell!

Kardkovács Zsolt Tivadar

Adatbázisok

vendégelőadás

Adatvédelem és anonimizálás adatbázisokban

- Jogi környezet
 - Miért kell? Miért, kell?
 - Mikor és mire kötelezettek a szereplők
 - Korlátok és lehetőségek
- Matematikai probléma
- Technológiai megfelelési lehetőségek
 - Stratégiák
 - Problémák és nyitott kérdések
 - Adatelemzés-orientált adatvédelem

*„In God we trust.
All others must bring data..”
(W. Edwards Deming)*





Általános eljárási menetrend, stratégia

- 0. Minimalizálás**
- 1. Álnevesítés**
- 2. Kritikus pontok felderítése**
- 3. Technológiai megoldások alkalmazása**
- 4. Felügyeleti terv elkészítése**



Általános eljárási menetrend, stratégia

0. Minimalizálás

- **A tárolt adatok körét és tartalmának minimalizálása a célnak megfelelően**
 - Minden (leány)vállalati körben áttekintve
 - Időben és térben is elemezve a kérdést
 - Nézetenkénti (felhasználói csoportok szerinti) bontásban is
- **Adatátadási politika kialakítása**
 - Beszállító adatok átvételének adattartalmának ellenőrzése
 - Partnerek számára átadott adatok adattartalmának ellenőrzése
- **Napló és audit minimalizálás**
- **Adattárolás céljának és körének kommunikációja**

1. Álnevesítés
2. Kritikus pontok felderítése
3. Technológiai megoldások alkalmazása
4. Felügyeleti terv elkészítése



Általános eljárási menetrend, stratégia

0. Minimalizálás

1. Álnevesítés

- Személyes adatok azonosítása
- Védekezési terület meghatározása és az elsődleges védelem kialakítása
 - Eszköz- és gyártófüggetlen kell legyen(!)
 - Hozzáférési politikai kialakítása
 - Naplózási kérdések kialakítása
- Belső azonosítók létrehozása
 - Belső azonosítók generálása
 - Áttérés a belső azonosítók használatára

2. Kritikus pontok felderítése

3. Technológiai megoldások alkalmazása

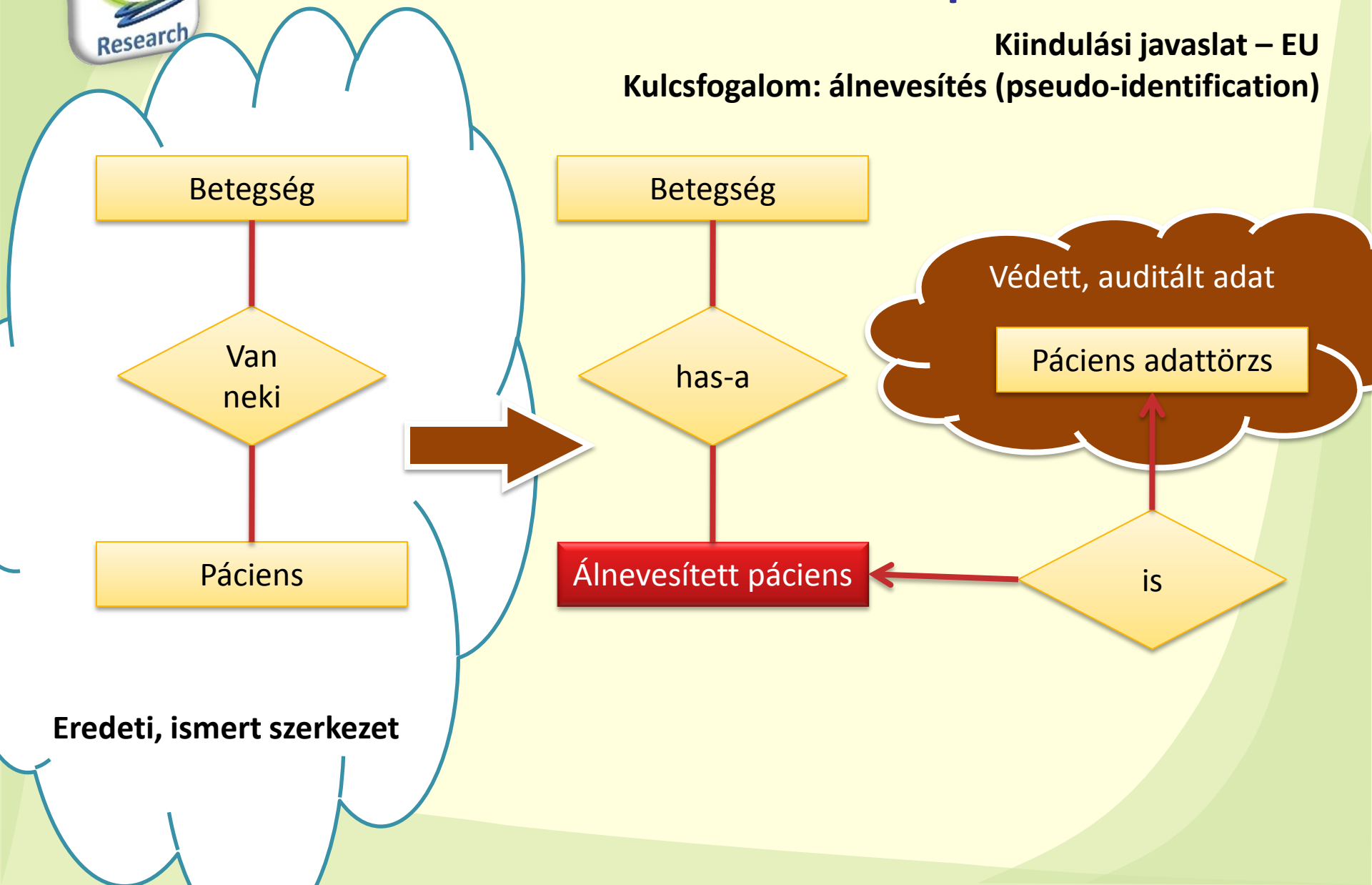
4. Felügyeleti terv elkészítése



1. lépés – álnevesítés

Kiindulási javaslat – EU

Kulcsfogalom: álnevesítés (pseudo-identification)



Gyakorlati probléma

Álnevesítés

átkeresztelünk

személyjellemzők

védett adat

ROWID	Név	Született	Nem	Irányítószám	Betegség
1	A	76-01-21	Nő	1107	Arthritis
2	B	86-03-24	Férfi	1107	Appendicitis
3	C	76-02-27	Nő	1117	Osterosis
4	D	76-01-21	Férfi	1117	Cancer
5	E	86-03-24	Nő	1127	Malaria
6	F	76-02-27	Férfi	1127	Hepatitis

Álnevesítés nem anonimizálás



Ez egy kutya...



Általános eljárási menetrend, stratégia

0. Minimalizálás
1. Álnevesítés
- 2. Kritikus pontok felderítése (stressz teszt)**
 - a. Technológiai problémák feltárása (biztonságtechnika)
 - b. Kvázi és indirekt azonosítók meghatározása
 - a. Rejtett kockázati források feltárása
 - b. Alternatívák keresése
 - c. Védekezési politika kialakítása
3. Technológiai megoldások alkalmazása
4. Felügyeleti terv elkészítése



2. lépés – kritikus pontok felderítése

Lehet, hogy van kész megoldás

- Speciális esetekre jellemző
 - USA HIPAA javaslat (Safe Harbor)
 - British Health Care és az EU 29-es munkacsoport ajánlásai
 - NAIH „ajánlások”
- Jogtárból, ítéletekből meríthetünk
- Szakirodalomból meríthetünk

...de többnyire nincs kész megoldás

- Ezért mindig törekedj a minimális attribútumhalmaz és leíróerő létrehozására
- ...magunknak kell meghatároznunk legalább a kvázi azonosítók halmazát
- ...és ügyelnünk kell a rejtett, például idősoros adatok kezelésére

Hogyan?



Példa kész megoldásra

Az alábbi adatok tárolását kell elkülönítetten (biztonságos módon) kezelni, a releváns orvosi (védett) adatokkal össze nem keverhető módon tárolni, tovább az orvosi szövegekből ezen utalásokat kell eltávolítani.

Safe Harbor

- Név
- Évtől különböző dátum
- Fax
- TB azonosító
- Egészségügyi szám
- Oklevél vagy licenc
- Eszköz és sorozatszám
- IP cím
- Arcképes fénykép
- Lokáció <20.000 lakosra
- Telefonszám
- Email cím
- Orvosi iratszám
- Bankszámlaszám (kötvényszám)
- Rendszám
- URL
- Biometrikus azonosító
- Egyedi azonosító, jellemző

Támadható... de nem könnyen



Kvázi azonosítók feltárása

Jó hír, van ismert módszer – avagy a KeLeT világa:

▪ **k-anonimitás**

- Egy reláció(!) k -anonim, ha a személyes jellemzőire vetített (zsákszemantikájú) eredményrelációja legalább k elemű.

▪ **ℓ -diverzitás**

- Egy reláció ℓ -diverz, ha bármely védendő attribútumára igaz, hogy a séma egyéb attribútumain a reláció legalább ℓ értéket vesz fel. (Vagyis a séma egyéb attribútumainak entrópiája nagyobb, mint az érték ℓ logaritmus.)

▪ **t -lezárt**

- (Egyszerűsített változat) Egy reláció érzékeny adatainak eloszlása és a teljes reláció eloszlásának távolsága (lásd t -próba) meghalad egy t küszöböt, akkor a reláció t -lezárt.

Kvázi azonosítók feltárása

k-anonimitás

- `SELECT Született, Nem, Irányítószám, COUNT(*) FROM t GROUP BY 1, 2, 3;`
- Egy 2-anonim reláció

ROWID	Született	Nem	Irányítószám	Betegség
1	76-*	Nő	11*	Arthritis
2	86-*	Férfi	11*	Appendicitis
3	76-*	Nő	11*	Osterosis
4	76-*	Férfi	11*	Cancer
5	86-*	Nő	11*	Malaria
6	76-*	Férfi	11*	Hepatitis

személyjellemzők

védett adat



Kvázi azonosítók feltárása

Jó hír, van ismert módszer – avagy a KeLeT világa:

- **k -anonimitás**

- Egy reláció(!) k -anonim, ha a személyes jellemzőire vetített (zsákszemantikájú) eredményrelációja legalább k elemű.

- **ℓ -diverzitás**

- Egy reláció ℓ -diverz, ha bármely védendő attribútumára igaz, hogy a séma egyéb attribútumain a reláció legalább ℓ értéket vesz fel. (Vagyis a séma egyéb attribútumainak entrópiája nagyobb, mint az érték ℓ logaritmusa.)

- **t -lezárt**

- (Egyszerűsített változat) Egy reláció érzékeny adatainak eloszlása és a teljes reláció eloszlásának távolsága (lásd t -próba) meghalad egy t küszöböt, akkor a reláció t -lezárt.

Kvázi azonosítók feltárása

A ℓ -diverzitás (k -anonimitás kiterjesztése)

- Ez általában nem megy adattorzítás nélkül...
- Egy 2-diverz (és 2-anonim) reláció:

aggregálunk

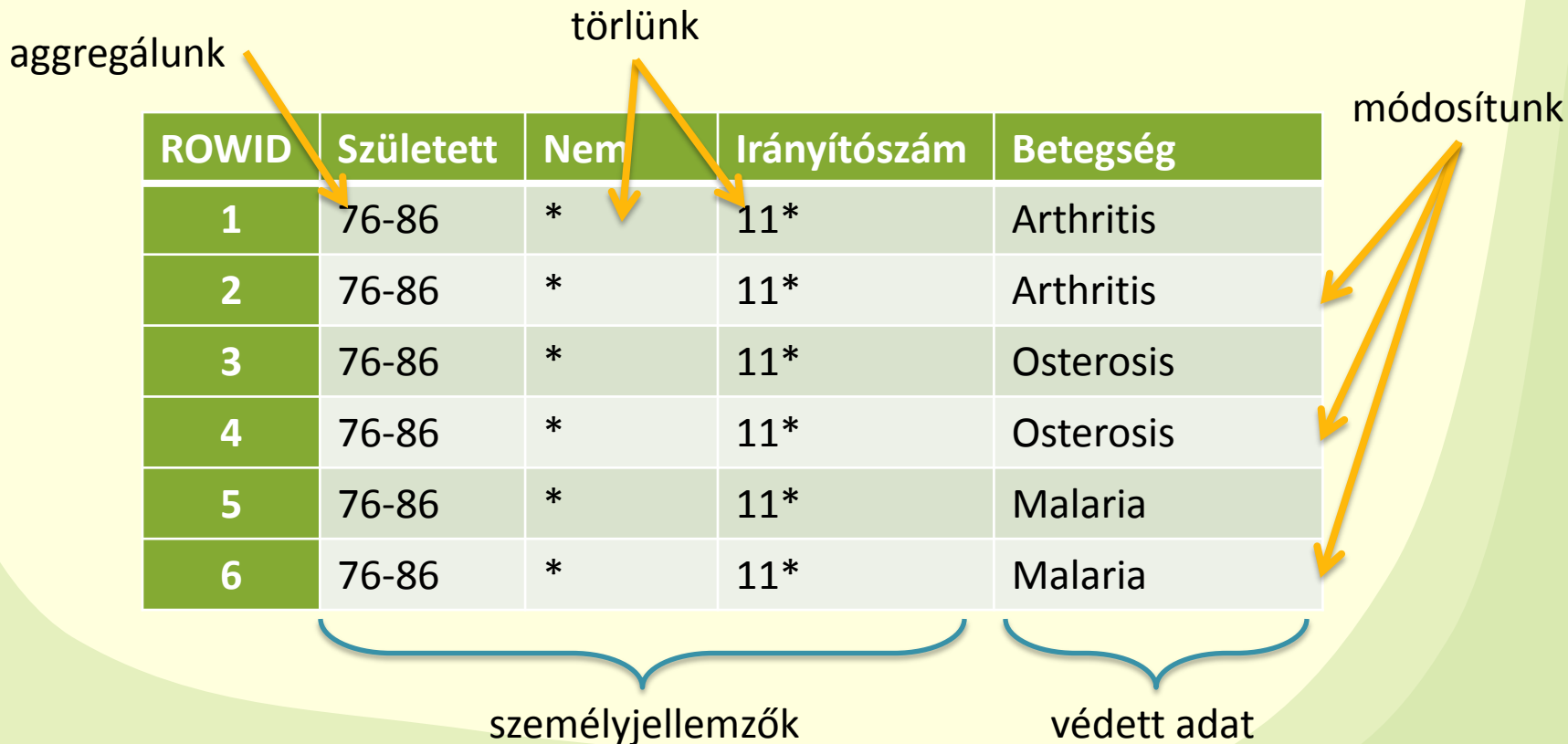
törlünk

módosítunk

ROWID	Született	Nem	Irányítószám	Betegség
1	76-86	*	11*	Arthritis
2	76-86	*	11*	Arthritis
3	76-86	*	11*	Osterosis
4	76-86	*	11*	Osterosis
5	76-86	*	11*	Malaria
6	76-86	*	11*	Malaria

személyjellemzők

védett adat





Kvázi azonosítók feltárása

Jó hír, van ismert módszer – avagy a KeLeT világa:

- **k -anonimitás**
 - Egy reláció(!) k -anonim, ha a személyes jellemzőire vetített (zsákszemantikájú) eredményrelációja legalább k elemű.
- **ℓ -diverzitás**
 - Egy reláció ℓ -diverz, ha bármely védendő attribútumára igaz, hogy a séma egyéb attribútumain a reláció legalább ℓ értéket vesz fel. (Vagyis a séma egyéb attribútumainak entrópiája nagyobb, mint az érték ℓ logaritmusa.)
- **t -lezárt**
 - (Egyszerűsített változat) Egy reláció védett adatainak eloszlása és a teljes reláció eloszlásának távolsága (lásd t -próba) meghalad egy t küszöböt, akkor a reláció t -lezárt.



Kvázi azonosítók feltárása

Rossz hír is van...

- A legegyszerűbb k -anonimitás ellenőrzés is $O(n \log n)$ komplexitású, de annak eldöntése, hogy valamely reláció k -anonim általánosságban NP-nehéz probléma
- Ha ismétlődő azonosítók is vannak, akkor b^n legalább tárhelyigényre van szükség, ahol b az érintettek és n az ismétlődések száma
- ...és **ez még csak a jól strukturált adatok** problematikája, nem beszéltünk
 - képről,
 - hangról,
 - szövegről,
 - videókról
 - stb.



Általános eljárási menetrend, stratégia

0. Minimalizálás
1. Álnevesítés
2. Kritikus pontok felderítése
3. **Technológiai megoldások alkalmazása, implementáció**
 - **Hozzáférés-korlátozás, védelmi felügyelet**
 - Titkosítás, tiltás
 - Naplózás, audit
 - **Adatredukció avagy adatcsonkolás**
 - Aggregált nézetek kialakítása
 - Csonkolásos
 - **Megtévesztés avagy funkcionális anonimizálás**
4. Felügyeleti terv elkészítése



Biztos, hogy szükségünk van a teljes információra?

- **Levágás (pl. rendszernaplók)**
 - IP cím (személyes adat!): 152.66.x.x
 - Kor meghatározásához elég az év: 1990
- **Aggregáció (pl. Statisztikai Hivatal adatai)**
 - Meglátogatott lapok száma
 - Eltöltött idő
 - Mezőtúr cenzusadatai
- **Rotáció**
 - Sütik – cookie (nem kell személyhez kötött legyen!)
- **Nem invertálható leképezés (pl. személyfelismerésnél)**
 - Hashing – ezzel óvatosan!
 - Arcpontok közötti viszony

Gyakorlati probléma

Adatcsonkolás

töröltünk

személyjellemzők

védett adat

ROWID	Született	Nem	Irányítószám	Betegség
1	76-*	Nő	11*	Arthritis
2	86-*	*	11*	Appendicitis
3	76-*	Nő	11*	Osterosis
4	76-*	Férfi	11*	Cancer
5	86-*	Nő	11*	Malaria
6	76-*	Férfi	11*	Hepatitis



Funkcionális anonimizálás

Észrevétel

- Az adat az érintettel való összekapcsolhatóságát akkor is elveszíti, ha nem egyértelmű kire vonatkozik – ezt nevezzük funkcionális anonimizálásnak
- ...viszont az adatbázis integritása sérül(het) – mivel áladatok keletkeznek

Lehetséges eszköztár

- Zaj hozzáadása
- Random permutáció
- Mintavételezés
- Többértelműsítés
- Szakirodalom által definiált minőségi kritériumok teljesítése
- ...és ezek lineáris kombinációja



Általános eljárási menetrend, stratégia

0. Minimalizálás
1. Álnevesítés
2. Kritikus pontok felderítése
3. Technológiai megoldások alkalmazása
- 4. Felügyeleti terv elkészítése**
 - Tárolt adatok felülvizsgálatának időzítése
 - Betekintési politikai kialakítása
 - Lekérési felületek és ellenőrzési pontok
 - Riportfelületek kialakítása jogosultságokkal körülírva
 - Adatmegsemmisítési politika kialakítása
 - Érintettek igényeinek azonosítása
 - Elvek meghatározása és véglegesítése (compliance policy)
 - Ügymenet kidolgozása és implementációja (iratkezelés!)
 - Megsemmisítési politika kommunikációja

Kardkovács Zsolt Tivadar

Adatbázisok

vendégelőadás

Adatvédelem és anonimizálás adatbázisokban

- **Jogi környezet**

- Miért kell? Miért, kell?
- Mikor és mire kötelezettek a szereplők
- Korlátok és lehetőségek

- **Matematikai probléma**

- **Technológiai megfelelési lehetőségek**

- Stratégiák
- Problémák és nyitott kérdések
- Adatelemzés-orientált adatvédelem

*„In God we trust.
All others must bring data..”
(W. Edwards Deming)*

