# Egészségügyi adatok anonimizálása

Dr. Vassányi István, PE-MIK VIRT

Források:

- Zoltán Alexin. Does fair anonymization exist?, International Review of Law, Computers & Technology (2014), DOI: 10.1080/13600869.2013.869909
- Dr. Alexin Zoltán. Kockázatokat rejt az egészségügyi adatok anonimizálása. IME XIII. 2. szám, 2014. március, 68-72 o.

# Bevezetés

- Anonimia: görög szó, értelme “név nélküli”
- Különösen egészségügyi, szociológiai, bűnügyi stb. kutatások esetén lényeges az érintettek személyiségi jogainak védelme
- Anonimizálás nélkül csak különleges esetben és feltételekkel adható ki adat (etikai engedély, adatvédelmi felügyelő, adatvédelmi szabályzat a hibák szankcióival, az adatok törlése a felhasználás után, az adatfeldolgozó kutatóhely fizikai és informatikai biztonsága, ellenőrzések…)
- **Ideális** anonimizálás esetén a kutatás elvégezhető, a statisztikai következtetések helytállóak, DE az egyes esetek gyakorlatilag nem azonosíthatóak
- Az adatok forrása lehet: kórházi információs rendszerek, adminisztratív adatbázisok, közösségi oldalak, …
- A támadó több forrást kombinálhat

## Törvényi szabályozás, ajánlások

- 1996: US Department of Health and Human Services
  HIPAA (Health Insurance Portability and Accountability Act)
  - Szakértő általi manuális anonimizálás (ad hoc) VAGY
  - ‘safe harbor’ módszer: az irányítószámok és dátumok generalizálása + bizonyos adatok törlése
  - 5 jegyű ir.szám  első 3 jegye, **minden** dátumból csak az év megtartása (vizsgálatok, események dátumára is)
  - 90 év felett nincs születési információ, csak ‘90 évnél idősebb’
  - Egyéb adatok, pl. testsúly stb. kategorizálása (amúgy is gyakran előfeltétele az adatfeldolgozásnak)
  - 5-nél kisebb elemszámú csoportot tartalmazó bontás ne legyen (módosítani kell a lekérdezést)
- Robosztus, de nem 100%-os hatékonyságú, több esetben is támadhatónak bizonyult (kiegészítő adatforrásokkal)

# Az azonosítási kockázat

- Alapelv: az anonimizálási módszer megválasztásához meg kell becsülni az azonosítás kockázatát (ami a konkrét adatok eloszlásától is függ)

$$kockázat = \frac{azonosítható\ személyek\ száma}{összes\ személy}$$

- Elfogadható kockázat: <0.1%, <1% (alkalmazás-függő)

## Az azonosítási kockázat becslése

- Ha egy n személyből álló csoportból az egyéneket véletlenszerûen b számú alcsoportba helyezzük (pl. születések dátuma éven belül), akkor az i embert tartalmazó alcsoportok száma:

$$f_n(i) = \binom{n}{i} b^{1-n} (b-1)^{n-i}$$

- Az egyenletes eloszlás sok esetben feltételezhető (pl. születési dátumok)
- g-különböző személyek: akiket nem lehet megkülönböztetni MAXIMUM g-1 személytől (1-különböző: egyke)

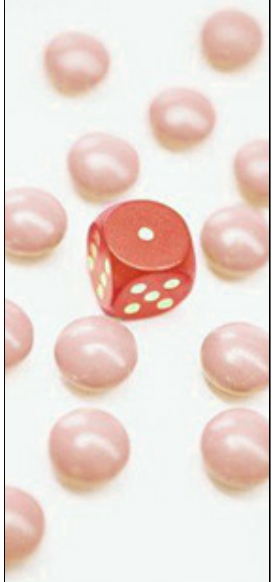
- Példa: 70 ember közül hányan vannak, akik születésnapján nincs senki másnak születésnapja (egykék)?

$$70 \times 365^{-69} \times 364^{69} \approx 57{,}93$$

- És hány napon lesz 2 embernek is születésnapja?

$$\frac{70 \times 69}{2} \times 365^{-69} \times 364^{68} \approx 5{,}49$$

# A képlet igazolása



LEMMA A.1. *Assume that $n$ individuals are distributed uniformly independently at random into $N$ bins. Let $f_i(n)$ be the expected number of bins that contain $i$ individuals. Then*

$$f_i(n) = \binom{n}{i} N^{1-n}(N-1)^{n-i}$$

PROOF. The value $f_i(n)$ is determined for all $i \geq 0$ and all $n \geq 0$ by the following formulas

$$\begin{aligned} f_0(0) &= N \\ f_0(n+1) &= f_0(n)(1-1/N) \\ f_i(n+1) &= f_i(n)(1-1/N) + f_{i-1}(n)/N \end{aligned}$$

Let us define $g_i(n) = f_i(n)N^{i-1}(1-1/N)^{i-n}$. The equations above become:

$$\begin{aligned} g_0(0) &= 1 \\ g_0(n+1) &= g_0(n) \\ g_i(n+1) &= g_i(n) + g_{i-1}(n) \end{aligned}$$

Therefore $g_i(n) = \binom{n}{i}$ and $f_i(n) = \binom{n}{i}N^{1-i}(1-1/N)^{n-i} = \binom{n}{i}N^{1-n}(N-1)^{n-i}$. □

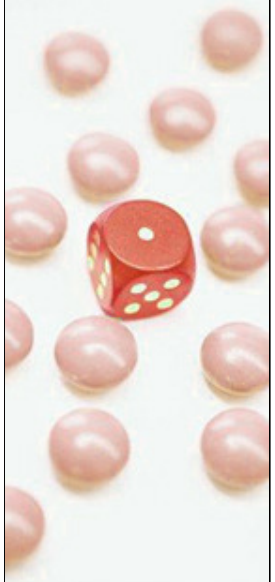
# Az azonosítási kockázat becslése

- Tehát a 70 emberből várhatóan 57.93 + 2*5.49 = 68.91 lesz 2-különböző
- k-ikrek: akik PONTOSAN k-1 személlyel megkülönböztethetelenek (adott tulajdonságok ismeretében)
- Minél kisebb a sokaság (pl. kis település), annál nagyobb hányad az 1-iker (egyértelműen azonosítható)
- Ha i db. megkülönböztethetelen személy közül 1/i valószínűséggel találjuk el, ki az emberünk, akkor:

$$\text{kockázat}(k\text{-különböző}) = \frac{\sum_{1 \le i \le k} i \times number\ of\ (i\text{-iker}) \times \frac{1}{i}}{\text{összes személy}} = \frac{\sum_{1 \le i \le k} number\ of\ (i\text{-iker})}{\text{összes személy}}$$

- Azonban sokszor nem véletlen a választás (kiegészítő információk alapján)

# Egy konkrét példa



Budai IK konvenciója tudományos kutatásra kiadott adatok esetén:

A reidentifikációs probléma kezelése, az adatvédelmi incidensek megelőzése céljából az alábbi konvenciók kerültek alkalmazásra.

A dokumentumok döntő mértékben valós adatokból állnak elő, bizonyos módosításokkal. Ezek lényege, hogy a demográfiai adatok olyan mértékben vannak megkeverve, hogy a dokumentum alanya ne legyen azonosítható (összhangban a hatályos adatvédelmi szabályozással), de minél több feldolgozási célból értékes jellemző megmaradjon:

- A beteg és anyja nevének vezetékneve és keresztneve lecserélésre került egy statisztikai alapon kiválasztott véletlenszerű értékre. Vezetéknevek esetén a forráspopulációban előforduló leggyakoribb 10%-ából, míg utónevek esetén 60%-ából történik a választás. Az új keresztnév esetén a nemnek megfelelő utónév került beállításra.
- A születéskori név nőnél a fentiek szerint történik, míg a férfiaknál kis valószínűséggel generálásra került (utóbbi esetben a forráspopulációban ritkán megadott), egyébként üres.
- A TAJ szám véletlenszerű, de algoritmikusan helyes értékre került lecserélésre.
- A születési idő a valódi év és hónap első napjára lett módosítva (így megőrződik az életkor!)
- A születési hely egy statisztikai alapon kiválasztott véletlenszerű érték (tipikusan Budapest).
- A lakcím irányítószám és település marad az eredeti (az adatok geográfiailag értékelhetőek maradnak!), az utcanév egy fix értékre kerül lecserélésre.
- A telefonszám(ok) és az email egy véletlenszerű (nem valós) értéket vesz fel, 27%-os valószínűséggel, egyébként üres marad.

Ugyanannak a betegnek a különböző eseteinél a fentiek szerint eltérésekkel (pl. más-más név alatt) jönnek létre a demográfiai adatok