Tipikus időbeli internetezői profilok nagyméretű webes naplóállományok alapján

Schrádi Tamás

schraditamas@aut.bme.hu

Automatizálási és Alkalmazott Informatikai Tanszék

BME

A feladat

- A webszerverek naplóállományainak hasznosítása
- Felhasználói profilok kialakítása
- Tipikus felhasználói csoportok → hasznosítható tudás a szolgáltatók számára

Cookie alapú felhasználó követés

- A weboldalak third party cookie-k (C3) raknak le a kliensek böngészőjébe (1x1pixel méretű háttérszínű kép, hivatkozás a vizsgálatot végző szerverre).
- A hivatkozás az oldal első letöltésekor egy, a központi szerver által kiadott cookie-t tesz le a kliens böngészőjében.
- A C3 azonosító egyedi és azonos a webhelyek meglátogatása során.
- Ha egy tetszőlegesen figyelt oldalra látogat a felhasználó, akkor a böngésző felküldi a C3 sütijét a központi szerverre, így megoldható a felhasználó követése

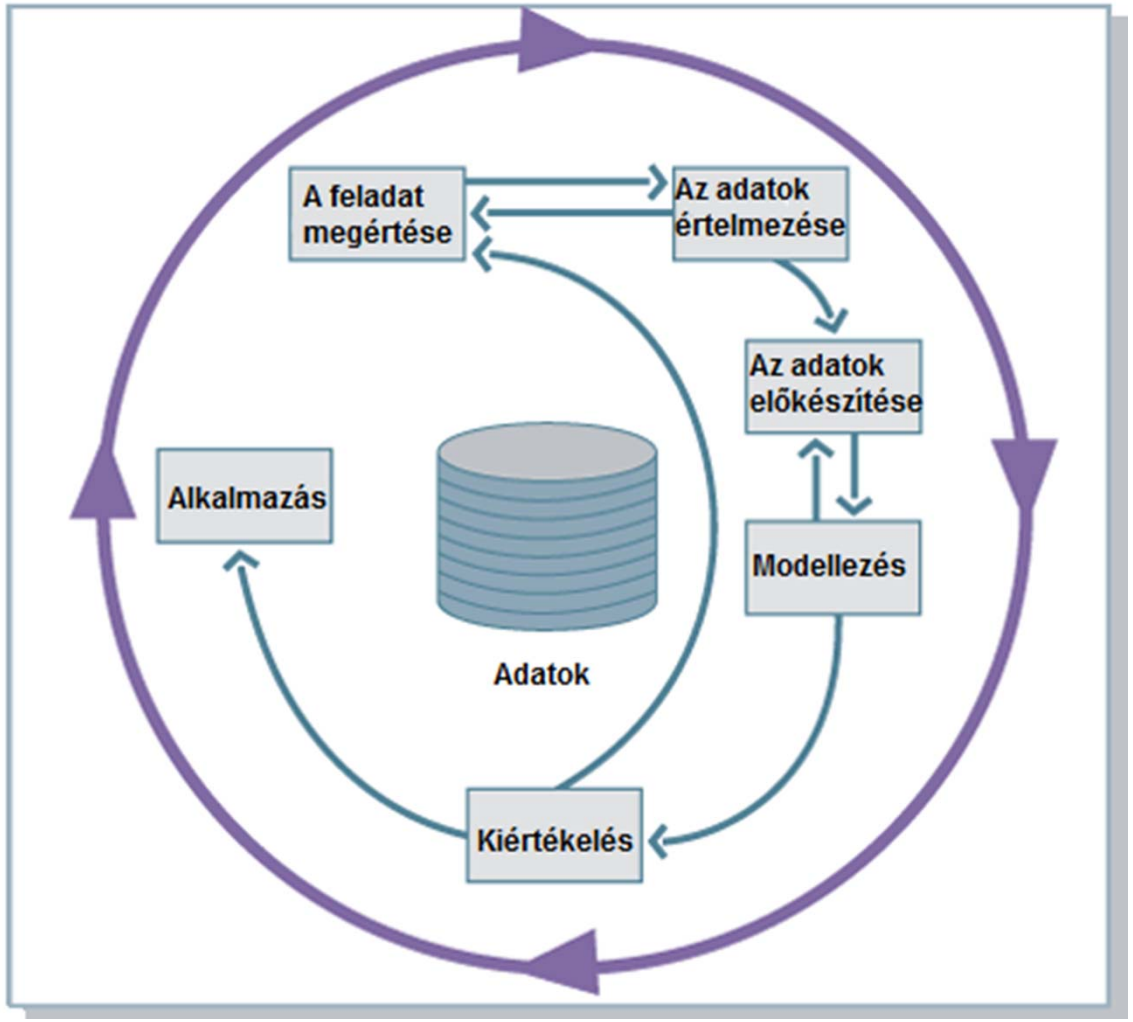
Cookie alapú felhasználó követés

- A C3 cookie-kat egyre gyakrabban törlik
- First-party cookie-k használata (C1)
- Összekapcsolható a két technika, a pontosabb azonosítás érdekében

Az „internetező” definíciója

- Internetezőt az alábbi hármas definiálja:
  - az operációs rendszer
  - az operációs rendszerbe belépett felhasználó azonosítója
  - a használt böngésző
- A cookie alapú követés indokolja
- Kifinomultabb definíció, mint az IP cím alapú azonosítás

Adatbányászati lépések



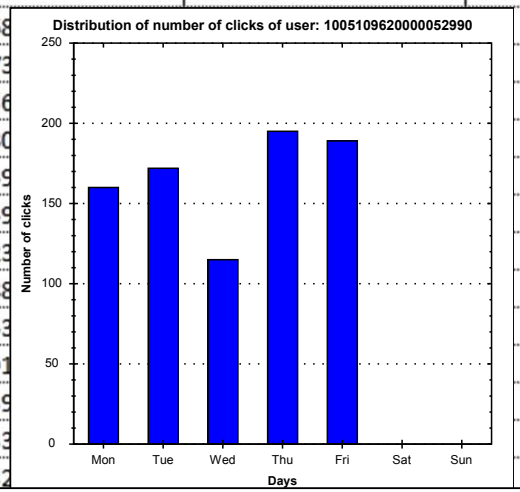
A CRISP-DM szerint

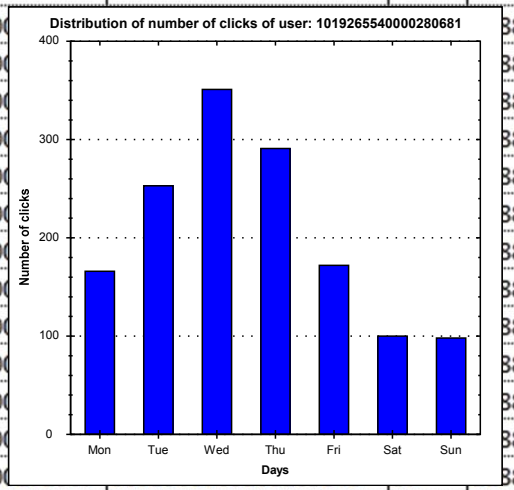
Kihívás

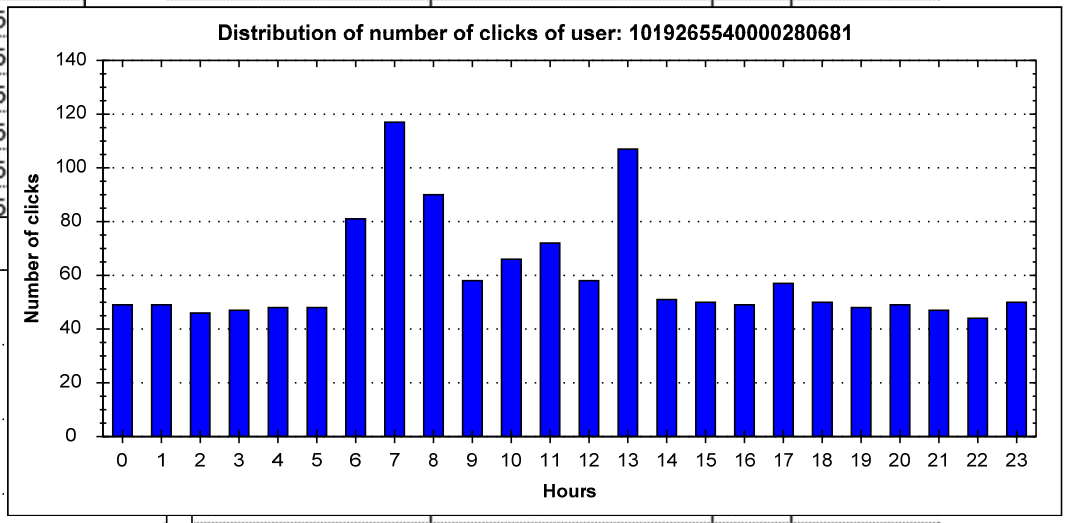
- Nagy adatmennyiség megnövekedett futási idő
- Egyetlen hónaphoz tartozó adat méretei:
  - Kb. 160 GB
  - Több mint 6 milliárd elemi rekord (32 és 64 bites egész számok) (1500 állományban)
- Nyers formájában nem dolgozható fel, nehezen értelmezhető

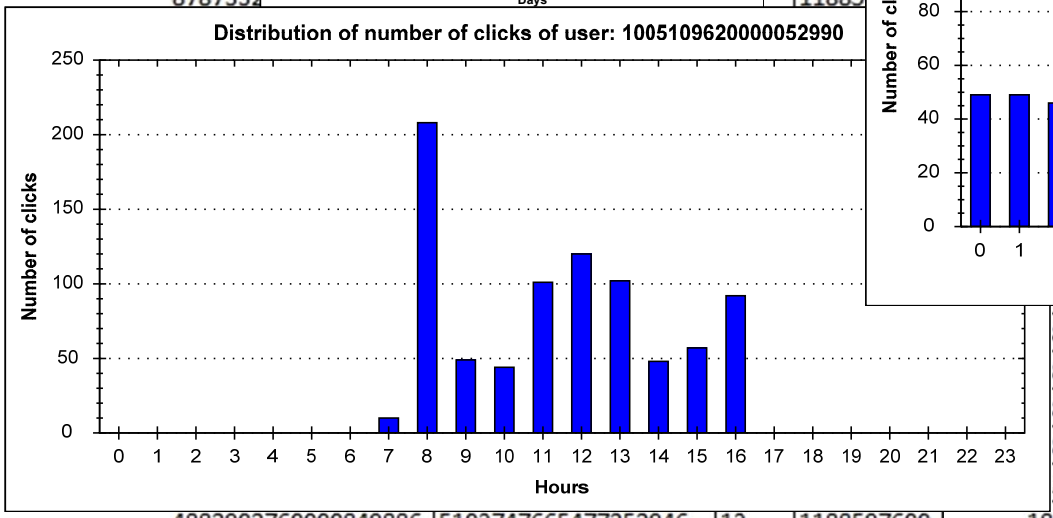
C3	C1	MID	Timestamp
1186084342000095335	5104980282598564234	0	1188597600
9783890830000696798	5061328953554657071	1	1188597600
2872913990000898253	5099371772408265811	0	1188597600
4777464930000064860	5071570822763023141	0	1188597600
2813285740000951464	5086777476597464245	0	1188597600
1657631970000702338	5091431718878549374	2	1188597600
5449354160000730116	5048440516908183211	0	1188597600
7793368			1188597600
6885973			1188597600
9689836			1188597600
5879880			1188597600
4501159			1188597600
4700859			1188597600
7623823			1188597600
6832088			11885
6556253			11885
7701791			11885
5861639			11885
8884133			11885
8787332			11885
4882902760000849886	5103747665477252946	13	1188597600
8789514750000649043	5104987875938757537	0	1188597600
8885908490000429885	5104958868084301208	14	1188597600
3830305360000821521	5100470897639457431	0	1188597600

1875158030000617913	5100362741621838793	47	1188597600
9881326130000569764	5103104541370239462	48	1188597600
1172991711000			38597600
6852833230000			38597600
7737114570000			38597600
3267183500000			38597600
8626482470000			38597600
6841867800000			38597600
4803644810000			38597600
4885692460000			38597600
1885945580000			38597600
3773501580000			38597600
3797853690000			38597600
4880757450000			38597600
5864965290000			38597600
42360390000046296	5086637044051584839	0	1188597600
07823650000925663	5071481049356831135	68	1188597600
83418160000799489	5104860333401308093	0	1188597600
65121260000619202	5096030773183422259	69	1188597600
83642098000073987	5103689850922696305	70	1188597600
1871876880000638878	5102294635091539248	71	1188597600
6802147080000853598	5091568994609706971	72	1188597600
3885949010000068416	5104987820979586474	0	1188597600
6841556980000399058	5104987829217939162	0	1188597600









Nyers erő

- 0. megközelítés: csak operatív memóriát használ
- A neve ellenére néhány megfontolás
- Hash alapú tároló a hatékony feldolgozás biztosításáért
- Ennek out-of-core kiterjesztését alkalmazzuk

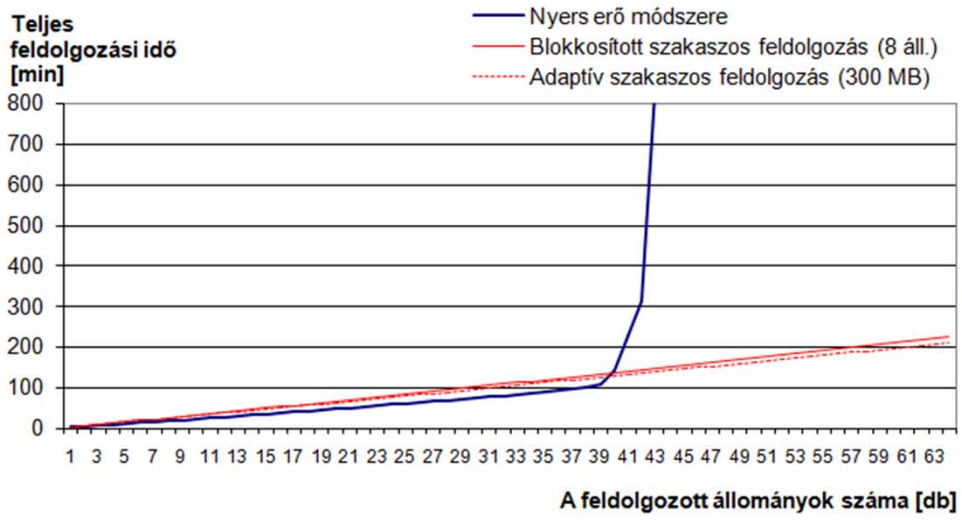
Szakaszos feldolgozás

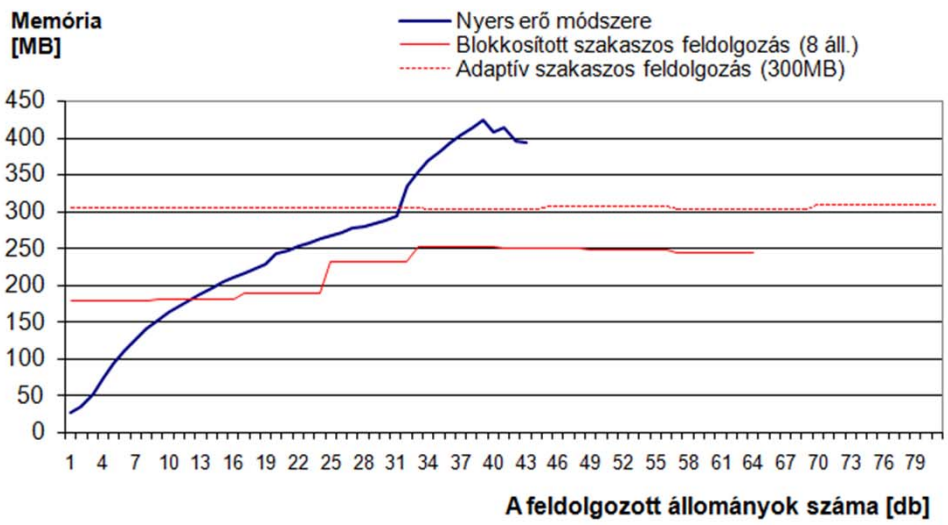
- Out-of-core módszer
- Kisebb részekben dolgozzuk fel a naplóállományokat
- Folyamatos mentés és összefésülés a merevlemezen
- Propagálja a háttértáron a megoldást
- Az utolsó összefésülés után végleges eredmény
- Rendezés
- Hibatűrés

K-utas összefésülés

- Out-of-core módszer
- A naplóállományokat külön-külön feldolgozza
- Rendezetten menti a háttértárra
- A rendezett, feldolgozott állományokon egy k-utas összefésülés
- Rendezés
- Hibatűrés

Az out-of-core módszerek szükségessége





Melyik feldolgozási módszert érdemes?

- A két bemutatott out-of-core módszer a k-utas összefésülés alapú módszernek nagyobb a háttértárigénye, míg a szakaszos feldolgozás memóriaigénye magasabb
- Mindkét módszer: hibatűrő, hatékony memóriakorlátos környezetben is (akár PC-n is)

Klaszterezés

- Elemek csoportosítása, ahol előre nem adottak a csoportdefiníciók
- Az egy csoportba kerülők hasonlítsanak jobban egymásra
- Nem egyértelmű
- Az adatbányászat egyik régi problémája

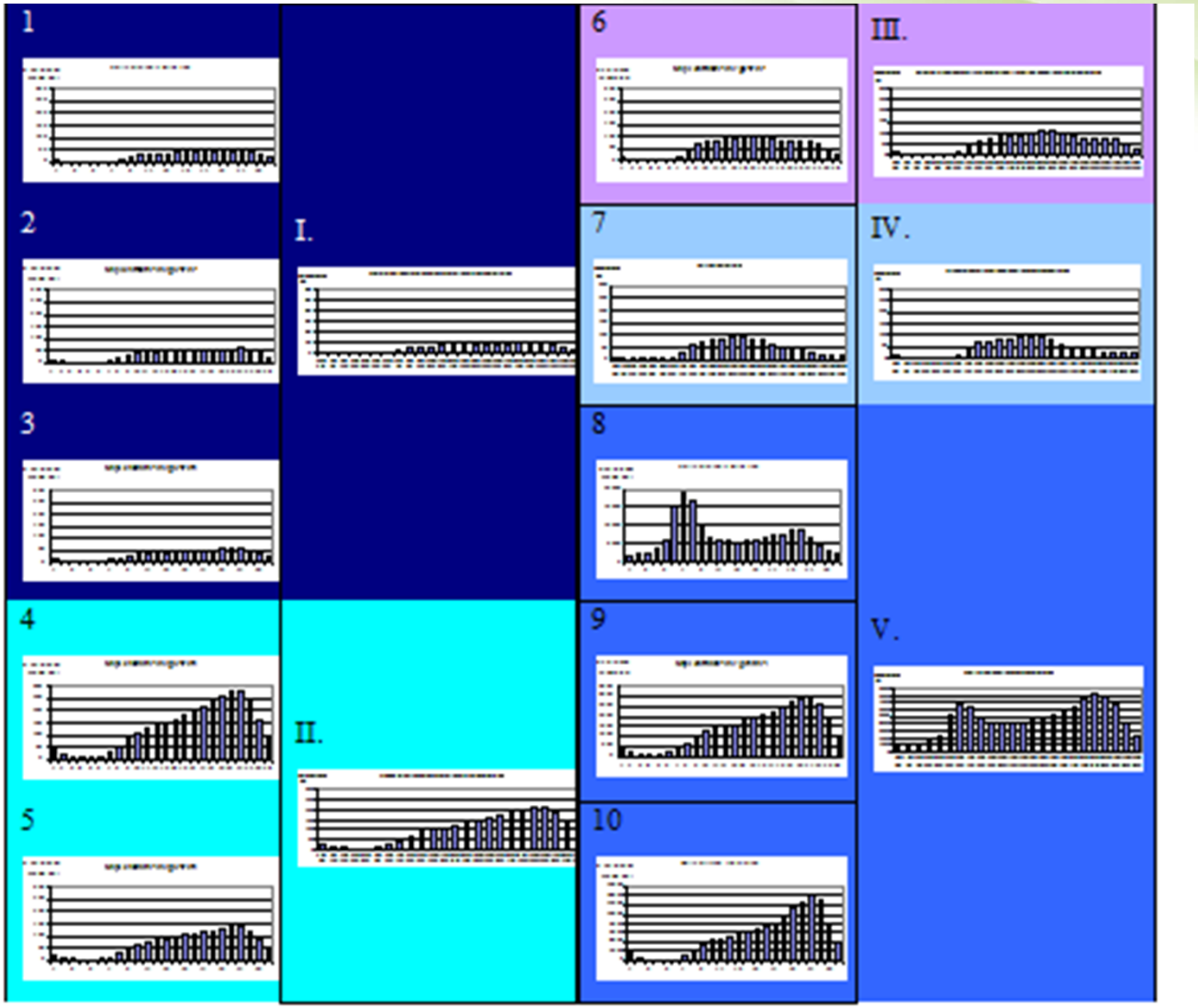
K-közép

- Egy alap algoritmus klaszterezésre
- Egy fizikai rendszer súlypontjának analógiájára épít
- Iteratív módszer
- Szegmentálásra is alkalmas
- A tapasztalat azt mutatja, hogy gyorsan lefut
- Lokális optimumba ragadhat
- A kialakítandó klaszterek számát ismerni kell

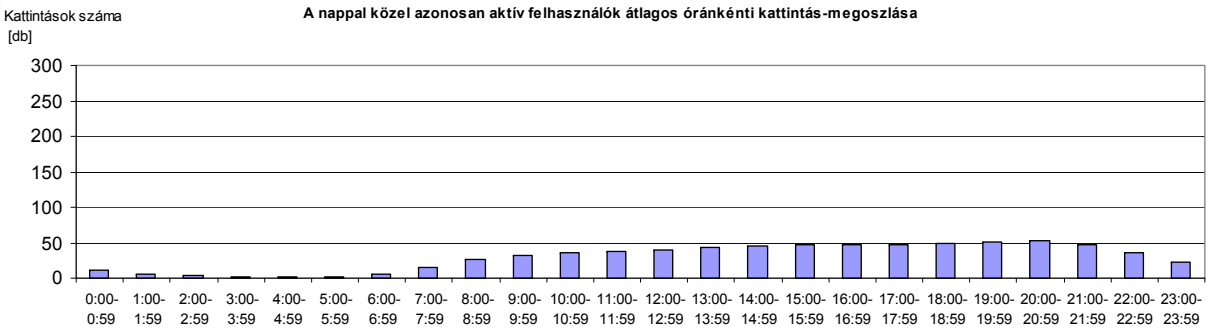
K-közép

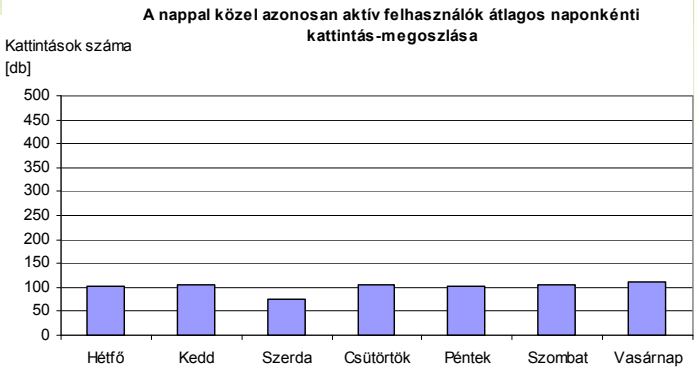
- A lokális optimumba való ragadás ellen, többszöri futtatás
- Heurisztika a klaszterek meghatározására: sok klasztert kialakítani, majd a hasonlókat összevonni (a szegmentálás miatt)

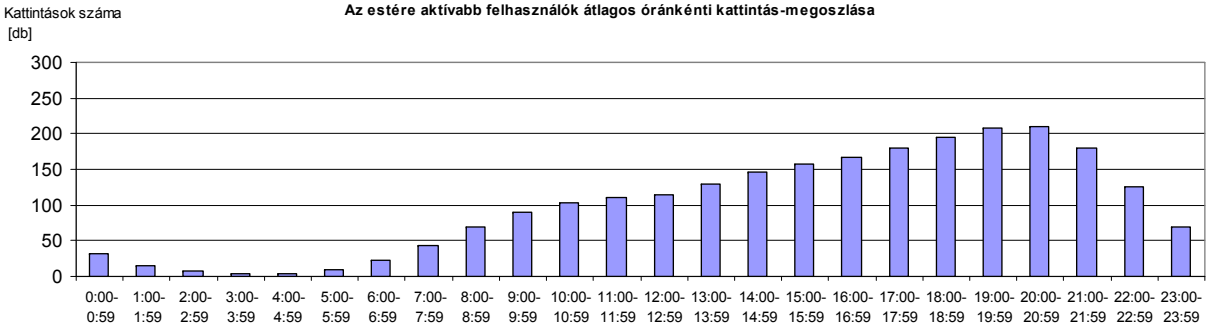
A kialakított klaszterek

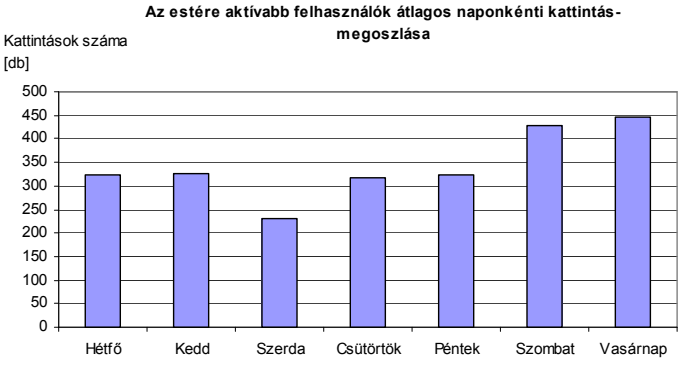


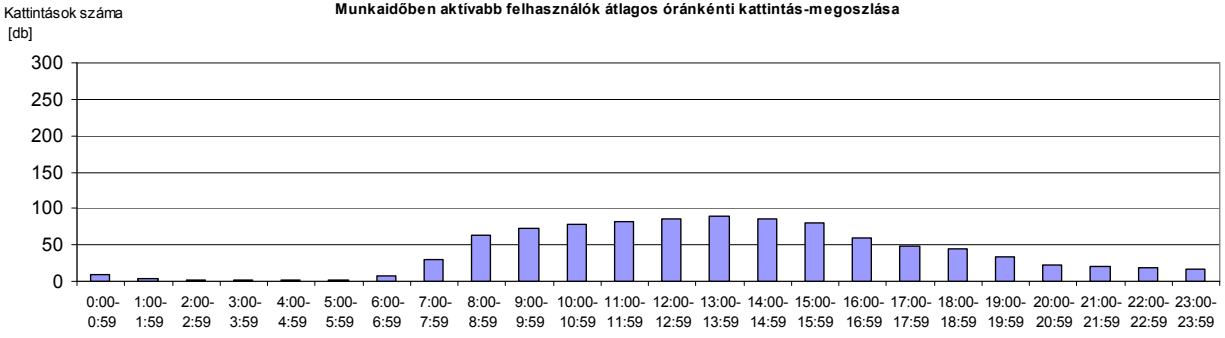
A kialakított klaszterek

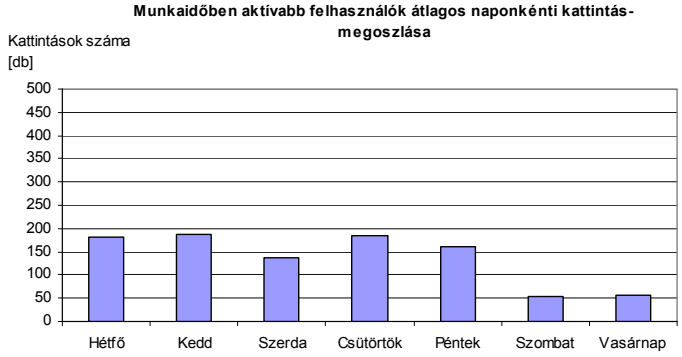




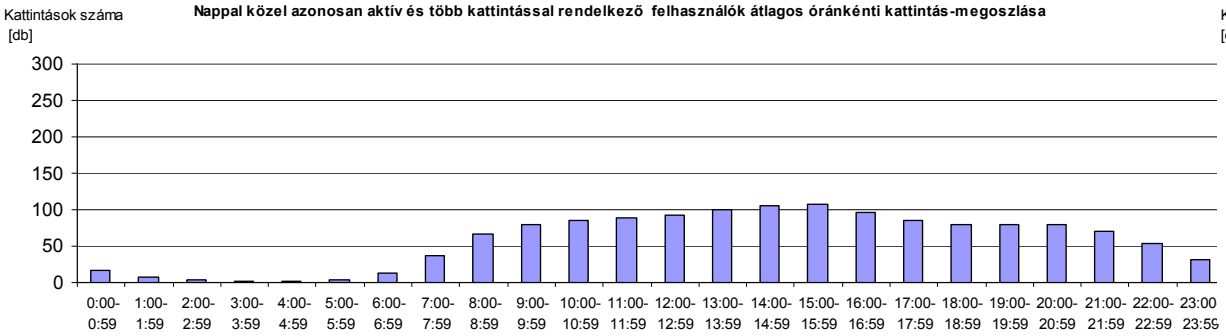


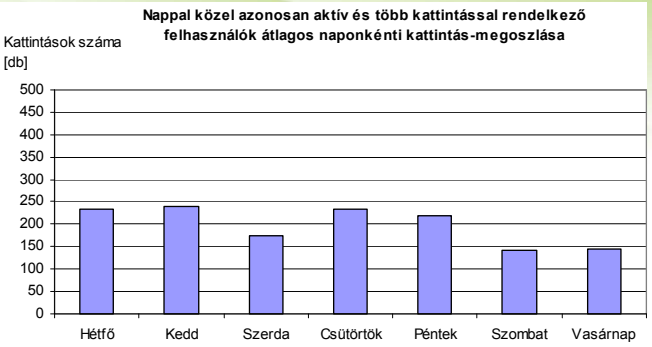


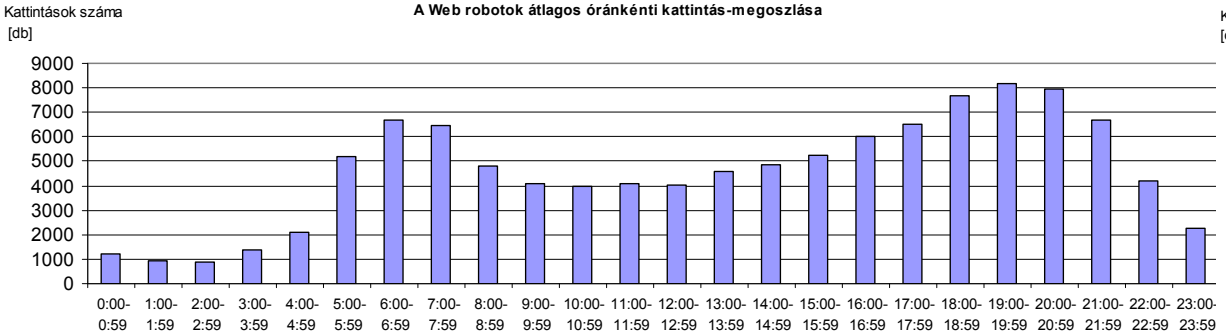


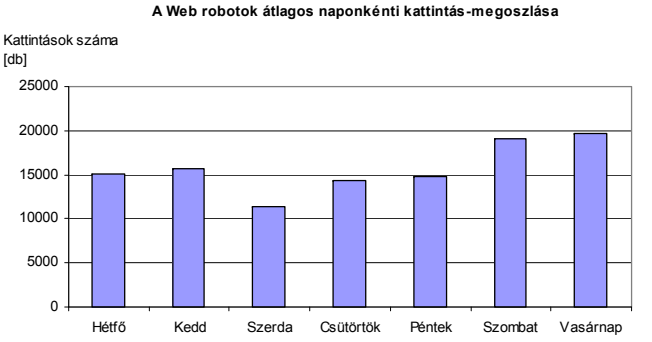


A kialakított klaszterek

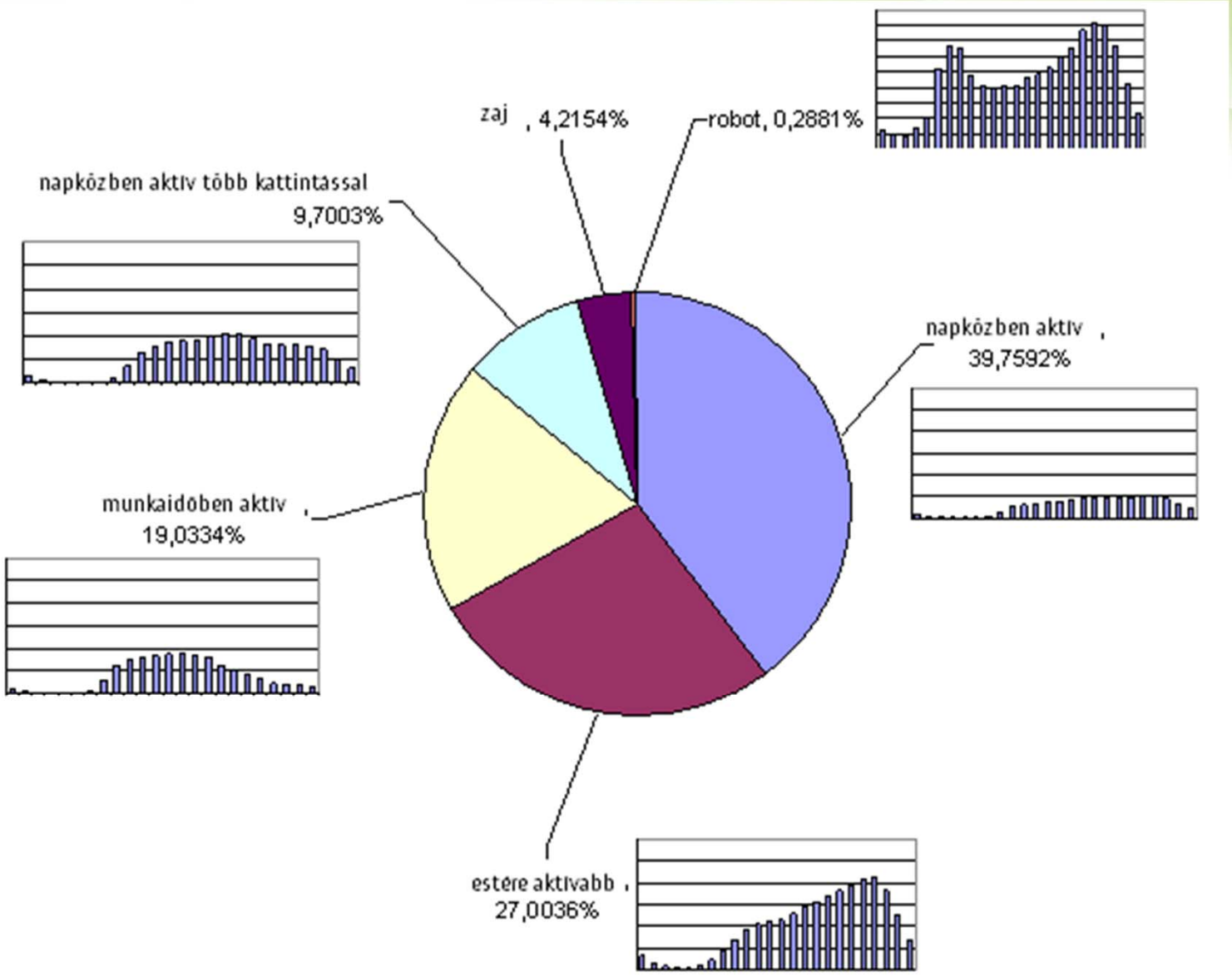




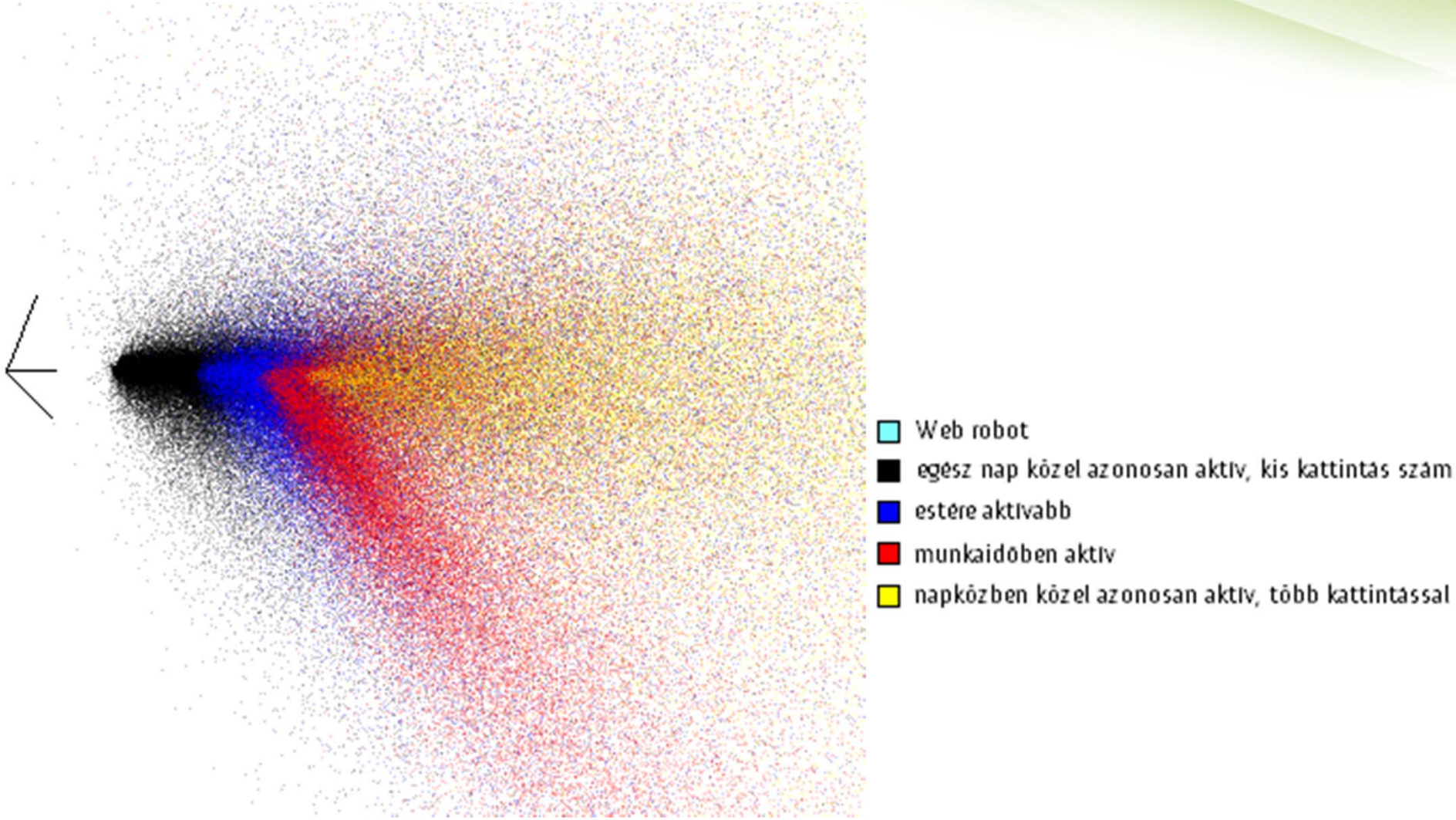




A kialakított klaszterek



A kialakított klaszterek



Köszönöm a figyelmet!