# Adatbányászat: Bevezetés

1. fejezet

Tan, Steinbach, Kumar

Bevezetés az adatbányászatba

előadás-fóliák

fordította

Ispány Márton

# Mi az adatbányászat?

- Bár hosszú évek óta alkalmazzuk még mindig nincs egyértelmű válasz erre a kérdésre.

- **Definíciók**:

Adatbányászat alatt hatékony módszerek használatát értjük adatok nagyon nagy összességének az elemzésére és hasznos, lehetőleg nem várt mintázatok kinyerésére.

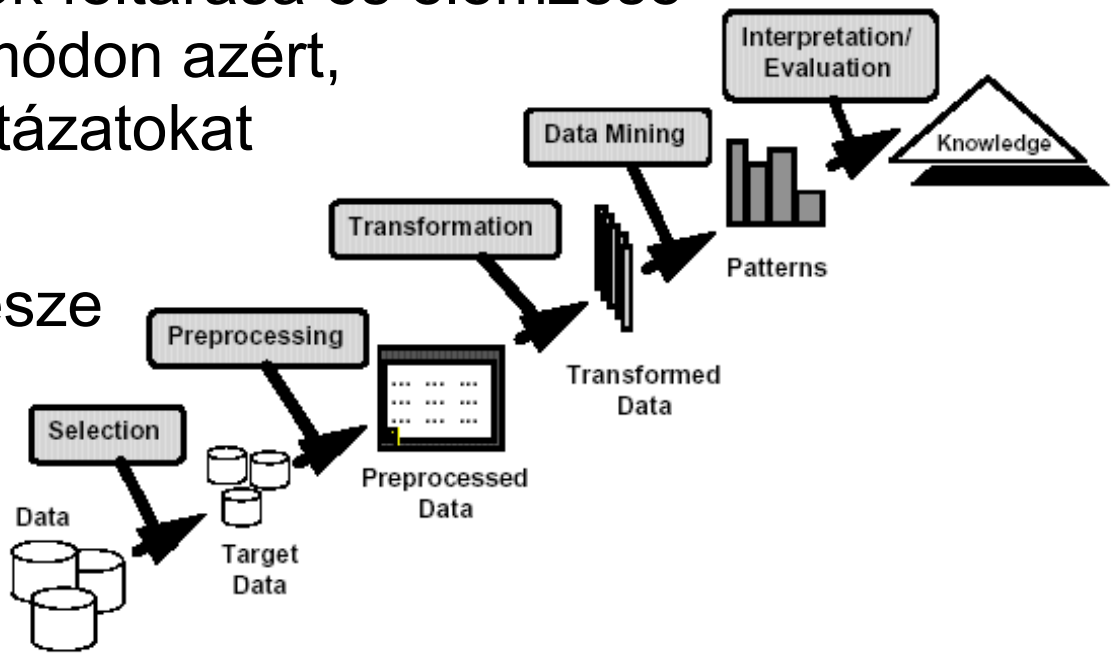
# Mi az adatbányászat?

- Definíciók:
  - Implicit (rejtett), korábban nem ismert és potenciálisan hasznos információ nem-triviális eszközökkel való feltárása.
  - Nagytömegű adatok feltárása és elemzése félig automatikus módon azért, hogy értelmes mintázatokat fedezzünk fel.
  - A KDD-folyamat része

    Knowledge

    Discovery from

    Databases



# Miért bányásszunk? Üzleti szempontok

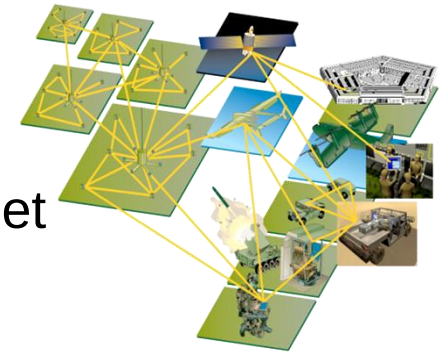
- Rengeteg adat gyűlik össze és raktározódik el adattárházakban:
  - web adatok, e-kereskedelem,
  - vásárlások áruházakban és élelmiszerboltokban,
  - bank- és hitelkártya tranzakciók.

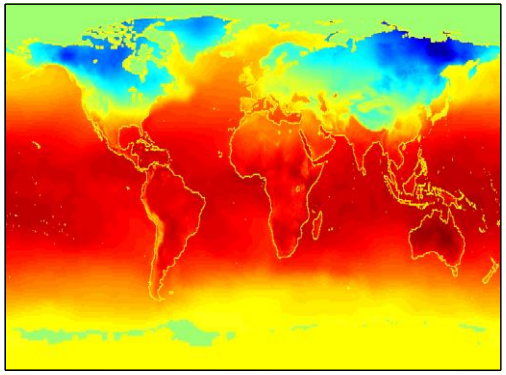


- A számítógépek egyre olcsóbbak, nagyobb teljesítményűek.
- A verseny erősödik
  - Nyújtsunk jobb, testreszabottabb szolgáltatást a versenyelőnyért (pl. CRM-ben).

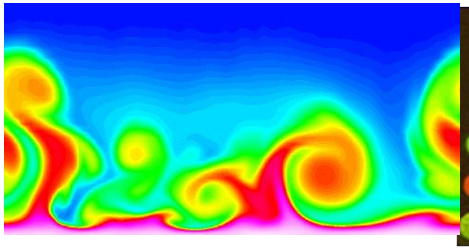
# Miért bányásszunk? Tudományos szempontok

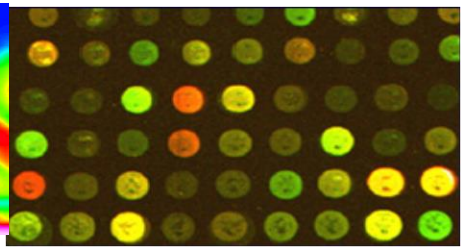
- Óriási sebességgel gyűlnek és tárolódnak az adatok (GB/óra)
  - távérzékelők műholdakon
  - távcsövek pásztázzák az eget
  - microarray mérések a génkifejeződésekre
  - szimulációk TB-nyi adatot generálnak





- Hagyományos módszerek alkalmatlansága
- Az adatbányászat segíthet a tudósoknak
  - adatok osztályozásában és szegmentálásában,
  - hipotézisek megfogalmazásában.





# Miért van szükség adatbányászatra?

- Igazán nagy számú nyers adat!!
  - A digitális korszakbanTB-nyi adat generálódik pillanatok alatt
    - Mobil eszközök, digitális fényképezőgépek, web-es dokumentumok.
    - Facebook adatok, tweet-ek, blog-ok, felhasználók által generált tartalmak
    - Tranzakciók, szenzor adatok, felügyeleti adatok
    - Lekérdezések, klikkek, böngészések
  - Az olcsó tárolás tette lehetővé ezeknek az adatoknak a kezelését
- Szükséges a nyers adatokat elemezni hogy tudást nyerjünk ki.

# Miért van szükség adatbányászatra?

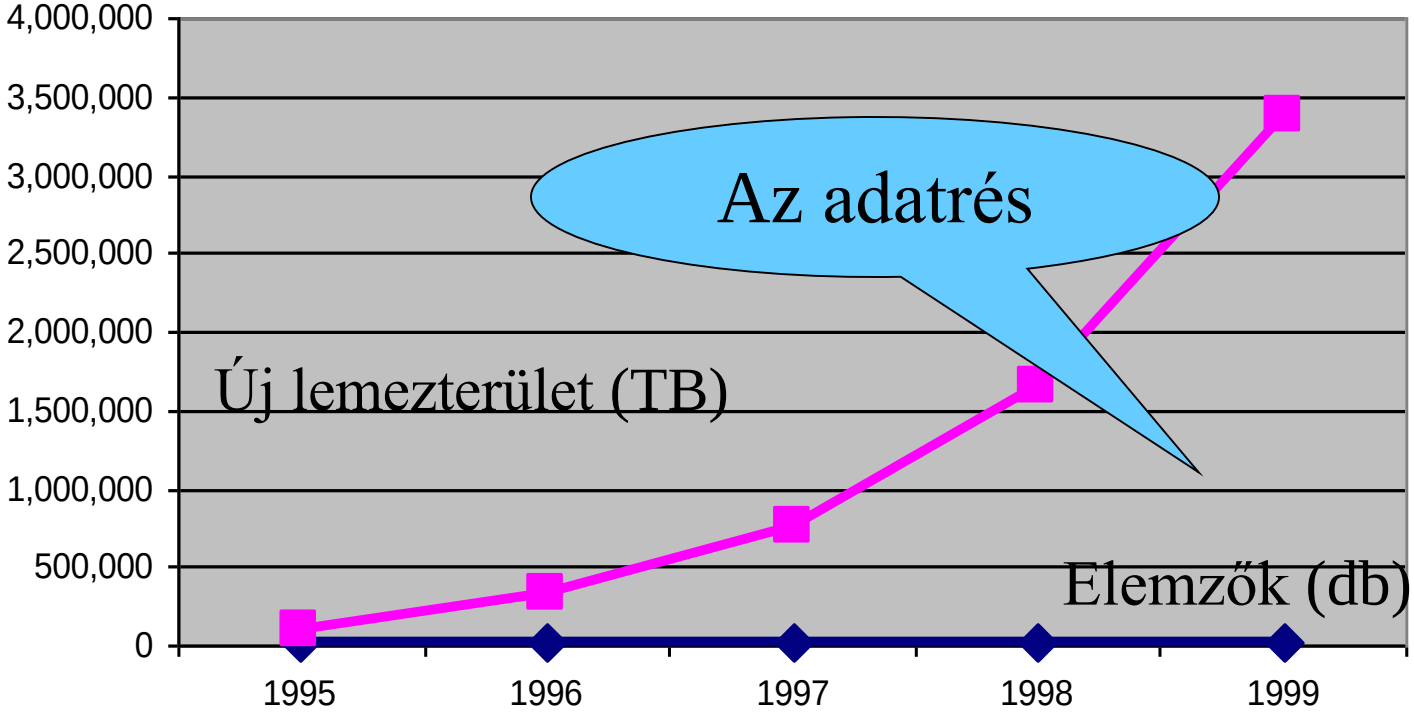
- "Az adat maga a számítógép"
  - Nagy mennyiségű adat hatásosabb lehet mint egy összetett algoritmus vagy modell
    - Google számos NLP feladatot oldott meg sikeresen, szimplán csak az adatokra támaszkodva
    - Példák: félregépelés, szinonímák
  - Az adat erő (hatalom)!
    - Napjainkban az összegyűjtött adatok az internetes nagyvállalatok legfőbb vagyonát adják
      - Google lekérdezések logjai
      - Facebook kapcsolati hálózatok
      - Twitter tweetek és követők
      - Amazon tranzakciók
  - Tudnunk kell hogyan hasznosítsuk a kollektív intelligenciát

# Az adat egyben nagyon összetett

- Adatok többféle típusa: táblázatok, idősorok, képek, grafikonok, stb.
- Tér és időbeli szempontok
- Összekapcsolt adatok különféle fajtái:
  - A mobil telefonból összegyűjthetjük a felhasználók helyét, baráti körét, megérkezését adott helyekre, twitteren közölt véleményét, kamera képeket, kereső szoftverekben elindított lekérdezéseket

# Nagy adatállományok bányászata - Motivációk

- A nem-nyilvánvaló információ gyakran ,,rejtve” van az adatokban.
- Az emberi elemzőknek hetekbe kerül míg hasznos információt találnak.
- Az adatok nagy többségét soha nem elemzik.




Forrás. R. Grossman, C. Kamath, V. Kumar: Data Mining for Scientific and Engineering Applications

# A KDD-folyamat

- Adatrögzítés
- Adattisztítás
- Adatintegráció
- Adatszelekció
- Adattranszformáció
- Adatbányászat
- Kiértékelés
- Tudásreprezentáció

A 2.-5. lépéseket az ún. *adattárház* kialakításának is nevezik az IT-n belül.

# Példa: tranzakciós adatok

- Valódi ügyfelek milliói:
  - WALMART: 40 millió tranzakció naponta, 40 petabyte adat naponta
  - AT&T 300 millió hívás naponta
  - Bankkártya társaságok: tranzakciók milliárdjai naponta.

- A pontgyűjtő kártyák lehetővé teszik a vállalatoknak, hogy információkat gyűjtsenek az ügyfeleikről.

# Példa: dokumentum adatok

- Web mint dokumentum tár: 1335 milliárd weblap

  http://www.internetlivestats.com/total-number-of-websites/

- Wikipedia: 5.5 millió cikkely

  https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

- Online hírportálok: új cikkek 100-ainak folyama nap mint nap

- Twitter:  500 millió tweet naponta

  http://www.internetlivestats.com/twitter-statistics/#trend

# Példa: hálózati adatok

- Web: 1335 milliárd lap hiperlinkekkel összekapcsolva
- Facebook: 2.2 milliárd felhasználó
- Twitter: 330 millió felhasználó
- Instant messenger: 203 millió felhasználó
- Blogok: 250 millió blog világszerte, híres emberek, politikai vezetők stb., fontos információs források

# Példa: genom-szekvenciák

- http://www.1000genomes.org/page.php
- 1000 egyed teljes szekvenciája
- $3*10^9$ nukleotida személyenként → $3*10^{12}$ nukleotida
- Még több adat: a személyek kórházi története, gén kifejeződési adatok

# Magatartási adatok

- Napjainkban a mobiltelefonok jelentős számú információt rögzítenek a felhasználók viselkedéséről:
  - GPS helyzet rekordok
  - Kamera képek
  - Telefon és SMS kommunikáció
  - Szövegek a facebook frissítésekben
  - Különféle egyedtípusok közötti kapcsolatok bejelentkezéseken keresztül
- Amazon mindent összegyűjt arról amit böngésztünk, a kosarunkba raktunk és megvásároltunk.
- Google és Bing minden böngészési aktivitásunkat képes összegyűjteni toolbar plugineken keresztül. Szintén rögzítik a lekérdezéseket, a lapokat melyet meglátogattunk és a klikket melyeket tettünk.
- Felhasználók millióinak adatait gyűjtik napi szinten

# Mi (nem) adatbányászat?

- **Mi nem adatbányászat?**
  - Egy telefonszám kikeresése a telefonkönyvből.
  - Az “Amazon” szóval kapcsolatos informá-ciók lekérdezése egy Webes keresővel. (Google)

- **Mi adatbányászat?**
  - Bizonyos nevek elterjedtebbek egyes területeken az USA-ban (O’Brien, O’Rurke, O’Reilly ír nevek Bostonban).
  - Csoportosítsuk tartalmuk alap-ján azokat a dokumentumokat, amelyeket egy keresővel kaptunk. (Pl. Amazonas esőerdő, Amazon kiadó)

# Mi az adatbányászat, ismét?

- Üzleti szempont
  - Az adatok váltak a legfontosabb versenyelőnnyé a vállalatok számára
    - Példák: Facebook, Google, Amazon
  - Annak képessége, hogy ki tudja nyerni a hasznos információkat az adatokból, kulcsfontosságúvá vált az üzleti sikerben.
- Tudományos szempont
  - A tudósok olyan, korábban elérhetetlennek tűnő, helyzetbe kerültek, hogyTB-nyi információt tudnak összegyűjteni
    - Példák: szenzor adatok, csillagászati adatok, kapcsolati hálók, gén adatok
  - Eszközök és módszerek kellenek arra, hogy elemezzük ezeket az adatokat azért, hogy jobban megértsük a világot és a tudomány fejlődjön.
- Skála (adat méret és jellemző dimenzió)
  - Miért nem használunk hagyományos analitikus módszereket?
  - Írdatlan mennyiségű adat, dimenzió probléma
  - Az adatok tömege és összetettsége nem engedi meg, hogy kézi feldolgozást alkalmazzunk, automatikus módszerek szükségesek.

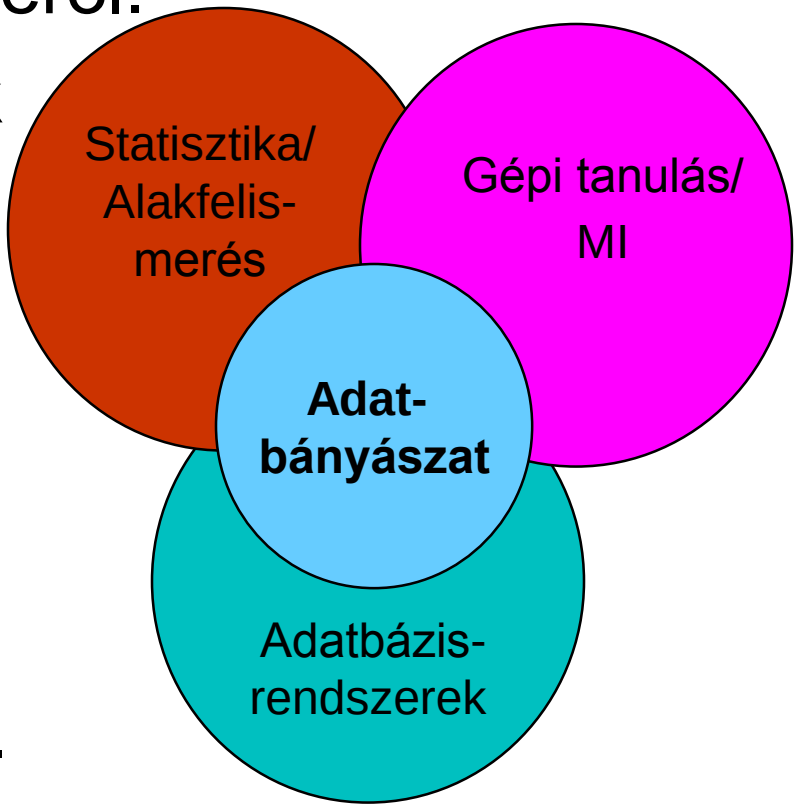
# Mi az adatbányászat, ismét?

- “Az adatbányászat (gyakran nagyméretű) megfigyelésen alapuló adatállomány elemzése azért, hogy nem sejtett kapcsolatokat találjunk és olyan új módon összegezzük az adatokat, amelyek az elemző számára egyaránt érthetőek és hasznosak.” (Hand, Mannila, Smyth)

- “Az adatbányászat modellek felfedezése adatok számára” (Rajaraman, Ullman)
  - Következő modellek jöhetnek szóba
    - Modellek, amelyek magyarázzák az adatokat (pl. egy egyszerű függvénykapcsolat)
    - Modellek, amelyek előrejelzik a jövőbeli adat eseteket.
    - Modellek, amelyek összegzik az adatokat
    - Modellek, amelyek kinyerik a kiemelkedő jellemzőit az adatoknak.

# Az adatbányászat eredete

- Ötleteket, módszereket merít a gépi tanulás/MI, az alakfelismerés, a statisztika és az adatbázisrendszerek területéről.
- A hagyományos módszerek alkalmatlanok lehetnek köszönhetően
  - az adattömegnek,
  - a nagy dimenziónak,
  - az adatok heterogén és elosztott természetének .



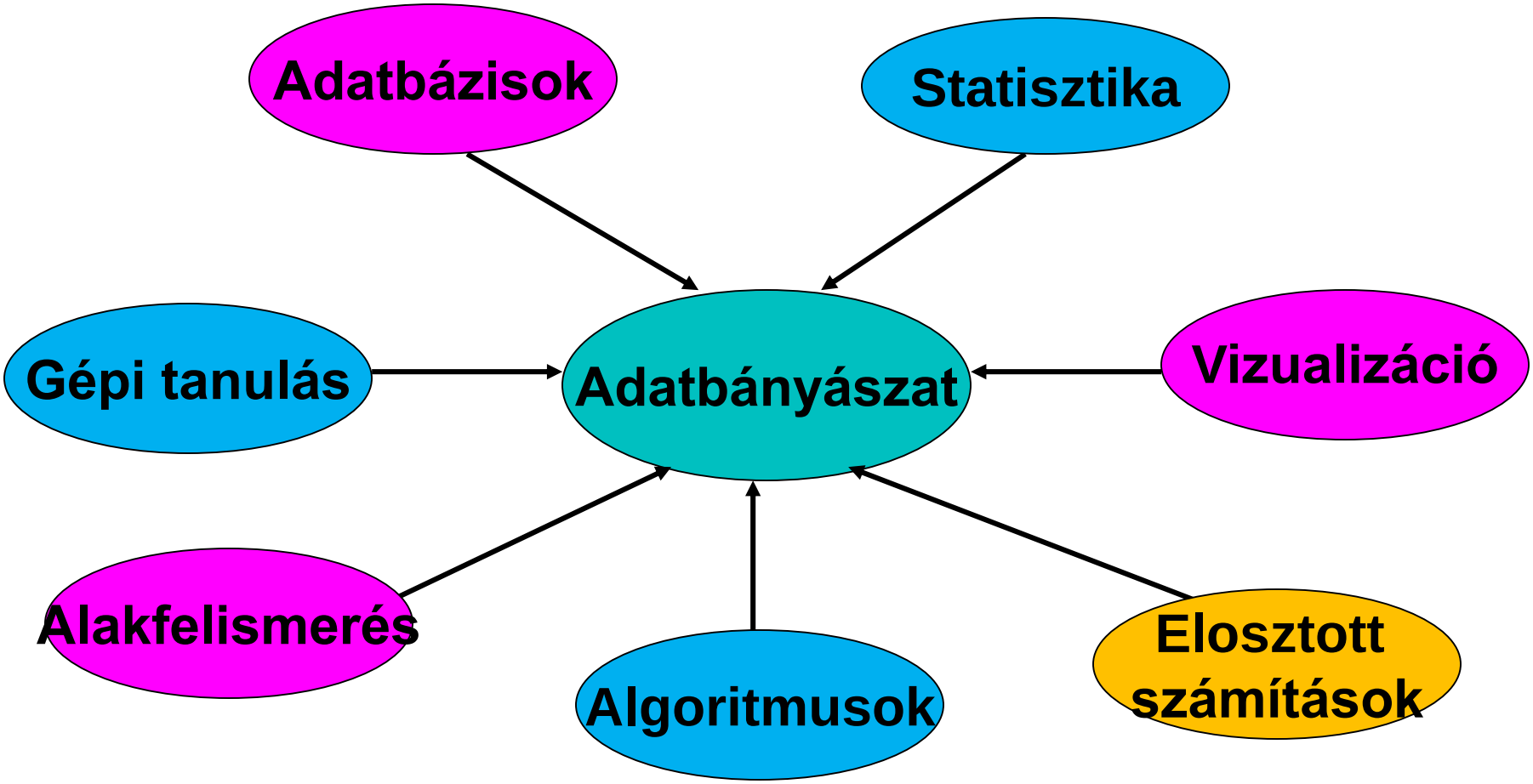
# Kultúrák

- Adatbázisok: nagy tömegű (memórián kívüli) adatra összpontosít.
- MI (gépi tanulás): komplex módszerekre és kevés adatra összpontosít.
  - Manapság az adat fontosabb mint az algoritmus. Váltás az MI-ben is.
- Statisztika: modellekre összpontosít.

# Modellek vs. Analitikus feldolgozás

- Egy adatbázis szakember számára az adatbányászat az analitikus feldolgozás egy extrém formája – olyan lekérdezések, amelyek nagy tömegű adatot vizsgálnak át.
  - Az eredmény a lekérdezésre adott válasz.
- Egy statisztikusnak az adatbányászat különféle, akár egymással versenyző, modellre való következtetés.
  - Az eredmény a modell paraméterei.

# Adatbányászat: több szakterület metszéspontja



# Adatbányászati feladatok

- Előrejelzés - predikció (Felügyelt adatbányászat)
  - Egyes változók segítségével becsüljük meg, jelezzük előre más változók ismeretlen vagy jövőbeli értékét.

- Leírás - jellemzés (Nem-felügyelt adatbányászat)
  - Találjunk olyan, az emberek számára interpretálható mintázatot, amely jellemzi az adatot.

Forrás. Fayyad tsai: Advances in Knowledge Discovery and Data Mining, 1996

# Adatbányászati alapfeladatok

- Osztályozás [Felügyelt]
- Csoportosítás [Nem-felügyelt]
- Társítási szabályok keresése [Nem-felügyelt]
- Szekvenciális mintázatok keresése [Nem-felügyelt]
- Regresszió [Felügyelt]
- Eltérés keresés [Felügyelt]

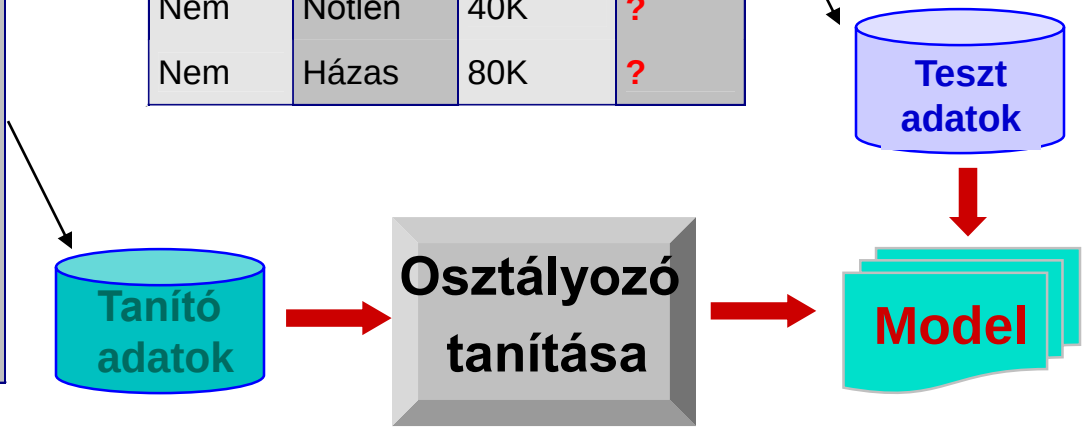
# Az osztályozás definíciója

- Adott rekordok egy halmaza (*tanító adatállomány*)
  - Minden rekord *attributumok* értékeinek egy halmazából áll, az attributumok egyike (vagy némelyike) az ún. *osztályozó* változó.
- Találjunk olyan *modellt* az osztályozó attributum-ra, amely más attributumok függvényeként állítja elő.
- Cél: korábban nem ismert rekordokat kell olyan pontosan osztályozni ahogyan csak lehetséges.
  - A *teszt adatállomány* a modell pontosságának meghatározására szolgál. Az adatállományt két részre bontjuk, a tanítón illesztjük a modellt, a tesztelőn pedig megállapítjuk a hibáját.

# Példa osztályozásra

| *Tid* | Vissza-térítés (kategórikus) | Családi állapot (kategórikus) | Jöve-delem (folytonos) | Csalás (osztályozó) |
|---|---|---|---|---|
| 1 | Igen | Nőtlen | 125K | Nem |
| 2 | Nem | Házas | 100K | Nem |
| 3 | Nem | Nőtlen | 70K | Nem |
| 4 | Igen | Házas | 120K | Nem |
| 5 | Nem | Elvált | 95K | Igen |
| 6 | Nem | Házas | 60K | Nem |
| 7 | Igen | Elvált | 220K | Nem |
| 8 | Nem | Nőtlen | 85K | Igen |
| 9 | Nem | Házas | 75K | Nem |
| 10 | Nem | Nőtlen | 90K | Igen |

| Vissza-térítés | Családi állapot | Jöve-delem | Csalás |
|---|---|---|---|
| Nem | Nőtlen | 75K | ? |
| Igen | Házas | 50K | ? |
| Nem | Házas | 150K | ? |
| Igen | Elvált | 90K | ? |
| Nem | Nőtlen | 40K | ? |
| Nem | Házas | 80K | ? |



# Osztályozás: 1. alkalmazás

- Direkt marketing
  - Cél: a levelezés költség csökkentése azon ügyfelek halmazának *megcélzásával* akik valószínűleg megvásárolják az új telefont.
  - Megközelítés:
    - Használjuk fel a korábban bevezetett hasonló termékekkel kapcsolatos adatokat.
    - Ismerjük, hogy mely ügyfél dönt úgy, hogy vásárol és melyik dönt másképp. Ez a *{vásárol, nem vásárol}* döntés képezi az *osztályozó attributumot*.
    - Gyűjtsük össze az összes ilyen ügyféllel kapcsolatos információt: demográfiai adatok, életstílus, céges előtörténet stb.
      - Foglalkozás, lakhely, mennyit keres stb.
    - Használjuk mindezen információt mint input attributumokat arra, hogy egy osztályozó modellt tanítsunk.

Forrás. Berry & Linoff: Data Mining Techniques, 1997

# Osztályozás: 2. alkalmazás

- Csalás keresés
  - Cél: a csalásnak tűnő esetek előrejelzése  hitelkártya tranzakcióknál.
  - Megközelítés:
    - Használjuk fel a hitelkártya tranzakciókat és a számlatulajdonossal kapcsolatos információkat.
      - Vásárláskor egy ügyfél mit vesz, milyen gyakran fizet
    - Címkézzük meg a múltbeli tranzakciókat: csalás ill. jó. Ez alkotja az osztályozó attributumot.
    - Tanítsunk egy modellt a tranzakciók egy halmazán.
    - Használjuk ezt a modellt arra, hogy a számlákhoz tartozó hitelkártya tranzakcióknál a csalást előre-jelezzük.

# Osztályozás: 3. alkalmazás

- Ügyfél lemorzsolódás
  - Cél: egy ügyfél elvesztésének előrejelzése (egy versenytárshoz való átpártolás)
  - Megközelítés:
    - Használjuk az összes múlt és jelenbeli ügyfélhez kapcsolódó tranzakciót attributumok keresésére.
      - Milyen gyakran telefonál, hol telefonál, leginkább melyik napszakban telefonál, pénzügyi helyzete, családi állapota stb.
    - Címkézzük meg az ügyfeleket aszerint, hogy hűségesek (lojálisak) vagy hűtlenek.
    - Találjunk modellt a hűségesek leírására.

Forrás. Berry & Linoff: Data Mining Techniques, 1997

# Osztályozás: 4. alkalmazás

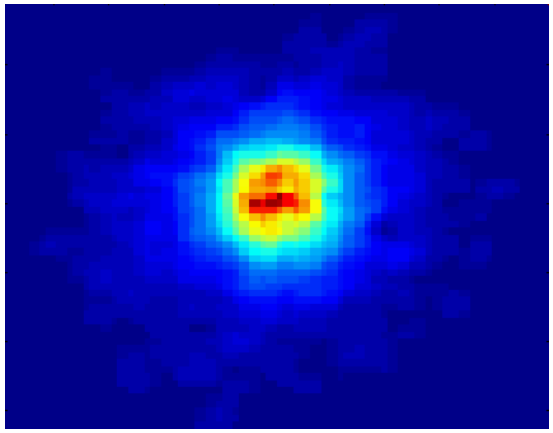
- Égboltfelmérés katalógizálása
  - Cél: égi objektumok osztályainak (csillag vagy galaxis) előrejelzése, figyelembe véve még az alig láthatóakat is. (Forrás: Palomar Obszervatórium)
    - 3000 kép, 23,040 x 23,040 pixel képenként.
  - Megközelítés:
    - Szegmentáljuk a képeket.
    - Mérjük meg a kép attributumait (features - jellemzők) - 40 db objektumonként.
    - Modellezzük az osztályokat ezen jellemzők alapján.
    - Sikertörténet: 16 új vörös-eltolódású kvazárt találtak, amely a legtávolabbi objektumok egyike és amelyet nehéz megtalálni!

Forrás. Fayyad tsai: Advances in Knowledge Discovery and Data Mining, 1996

# Galaxisok osztályozása

Forrás: http://aps.umn.edu

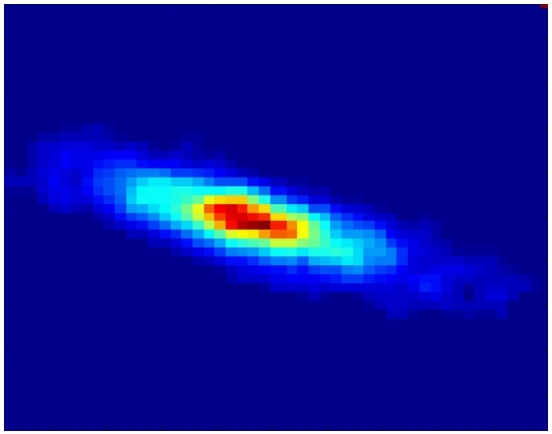
*Fiatal*



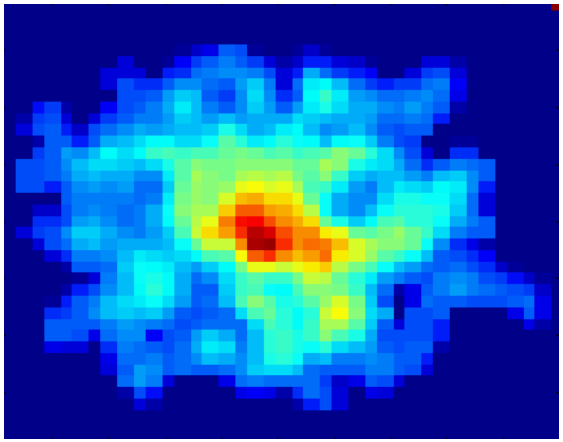
**Osztályozó változó:**
- **Az alakzat állapotai**

**Attributumok:**
- **Képi jellemzők**
- **A vett fényhullámok karakterisztikája stb.**

*Középkorú*



*Idős*



**Adatnagyság:**
- **72 millió csillag, 20 millió galaxis**
- **Objektum katalógus: 9 GB**
- **Kép adatbázis: 150 GB**

# A csoportosítás definíciója

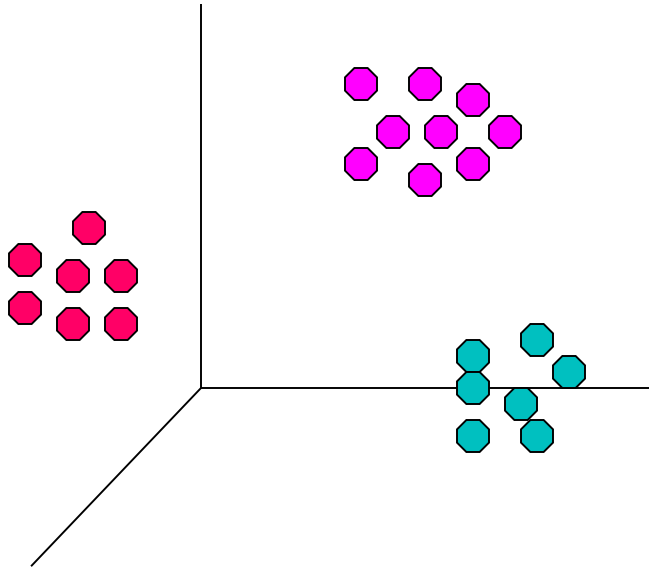
- Adott rekordok (pontok) egy halmaza, melyeket attributumok egy halmazával írunk le, továbbá adott közöttük egy hasonlósági mérték. Találjunk olyan csoportokat (klasztereket), amelyekre
  - az azonos csoportban lévő rekordok minél hasonlóbbak,
  - a különböző csoportokban lévők pedig minél kevésbé hasonlóak.
- Hasonlósági mértékek:
  - euklideszi távolság, ha az attributumok folytonosak,
  - egyéb, a feladattól függő mérőszámok.

# A csoportosítás szemléltetése

Euklideszi távolságon alapuló csoportosítás a háromdimenziós térben.

A csoportokon belüli távolságot minimalizáljuk

A csoportok közötti távolságot maximalizáljuk



# Csoportosítás: 1. alkalmazás

- Piac szegmentáció
  - Cél: a piac felosztása az ügyfelek diszjunk halmazok-ra való bontása útján, ahol minden egyes potenciális célcsoportot, piaci szegmenst különböző marketing eszközökkel tervezünk elérni.
  - Megközelítés:
    - Gyűjtsük össze az ügyfeleket jellemző attributumokat, amelyek pl. földrajzi és életstílushoz kapcsolódó információk.
    - Keressük hasonló ügyfelek csoportjait.
    - Mérjük meg a csoportosítás (szegmentálás) jóságát az ügyfelek vásárlási mintáit vizsgálva. Az egy csoportba eső ügyfelek hasonlóan viselkednek-e szemben a más csoportokba esők különböző viselkedéséhez képest.

# Csoportosítás: 2. alkalmazás

- Dokumentumok csoportosítása
  - Cél: egymáshoz hasonló dokumentumok csoportjainak keresése a bennük megjelenő fontosabb kulcsszavak alapján.
  - Megközelítés: azonosítsuk a leggyakrabban előforduló kifejezéseket a dokumentumokban. Definiáljunk egy hasonlósági mértéket a különböző kifejezések gyakorisága alapján. Használjuk ezt a csoportosításra.
  - Haszon: információ kinyerésre használhatjuk a csoportokat  új dokumentum beillesztésével vagy kifejezések (kulcsszavak) keresésével a csoportosított dokumentumokban.

# Dokumentum csoportosítás szemléltetése

- Csoportosítandó: 3204 cikk a Los Angeles Timesból.
- Hasonlósági mérték: mennyi közös szó van a dokumentumokban (előfeldolgozás után).

| Kategória | Összes cikk | Helyes osztály |
|---|---|---|
| Gazdaság | 555 | 364 |
| Külföld | 341 | 260 |
| Belföld | 273 | 36 |
| Közlekedés | 943 | 746 |
| Sport | 738 | 573 |
| Kultúra | 354 | 278 |

# S&P 500 részvény adatok

- Minden nap megfigyeljük a részvények mozgását.
- Csoportosítandó rekordok: Részvény-{FEL/LE}
- Hasonlósági mérték: két rekord hasonló, ha az őket leíró események gyakran fordulnak elő azonos napokon.
  - Társítási szabályt használtunk a hasonlósági mérőszám meghatározására.

| | *Talált klaszterek* | *Ipari csoport* |
|---|---|---|
| **1** | Applied-Matl-LE,Bay-Network-LE,3-COM-LE, Cabletron-Sys-LE,CISCO-LE,HP-LE, DSC-Comm-LE,INTEL-LE,LSI-Logic-LE, Micron-Tech-LE,Texas-Inst-LE,Tellabs-Inc-LE, Natl-Semiconduct-LE,Oracl-LE,SGI-LE, Sun-LE | Technológia1-LE |
| **2** | Apple-Comp-LE,Autodesk-LE,DEC-LE, ADV-Micro-Device-LE,Andrew-Corp-LE, Computer-Assoc-LE,Circuit-City-LE, Compaq-LE, EMC-Corp-LE, Gen-Inst-LE, Motorola-LE,Microsoft-LE,Scientific-Atl-LE | Technológia2-LE |
| **3** | Fannie-Mae-LE,Fed-Home-Loan-LE, MBNA-Corp-LE,Morgan-Stanley-LE | Pénzügy-LE |
| **4** | Baker-Hughes-FEL,Dresser-Inds-FEL, Halliburton-HLD-FEL, Louisiana-Land-FEL, Phillips-Petro-FEL,Unocal-FEL, Schlumberger-FEL | Olaj-FEL |

# Társítási szabályok definíciója

- Adott rekordok egy halmaza, amely tételek (termékek) egy összességét tartalmazza.
  - Keressünk olyan összefüggéseket, következtetéseket, amely egyes tételek előfordulását előrejelzi más tételek előfordulása alapján.

| *TID* | *Tételek* |
|---|---|
| 1 | Kenyér, Kóla, Tej |
| 2 | Sör, Kenyér |
| 3 | Sör, Kóla, Pelenka, Tej |
| 4 | Sör, Kenyér, Pelenka, Taj |
| 5 | Kóla, Pelenka, Tej |

Feltárt szabályok:

**{Tej} --> {Kóla}**

**{Pelenka, Tej} --> {Sör}**

# Társítási szabályok: 1. alkalmazás

- Marketing és reklám
  - Legyen a feltárt szabály

    *{Édessütemény, … } --> {Burgonyaszirom}*
  - Burgonyaszirom mint következmény => Arra használható, hogy meghatározzuk mit tegyünk az eladás meggyorsításáért.
  - Édessütemény mint előzmény => Arra használható, hogy lássuk mely termékekre van hatással az, ha a bolt felhagy az édességek forgalmazásával.
  - Édessütemény mint előzmény *és* burgonyaszirom mint következmény => Arra használható, hogy lássuk mely termékeket kell az édességek mellett árulni, hogy előmozdítsuk a burgonyaszirom forgalmát!

# Társítási szabályok: 2. alkalmazás

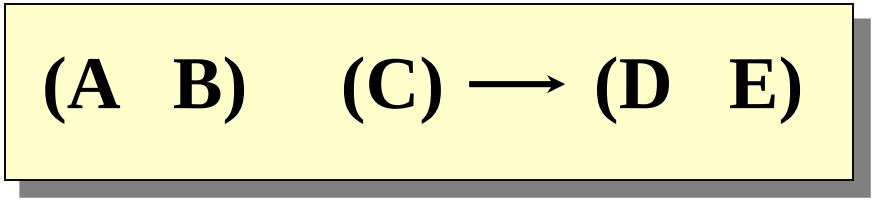
- Bevásárlóközpontok polckezelése
  - Cél: azon termékeknek a meghatározása, amelyeket elég sok vásárló vesz meg egyszerre.
  - Megközelítés: dolgozzuk fel az automatizált vásárlás során a vonalkód leolvasóval gyűjtött adatokat a termékek között kapcsolatokat keresve.
  - Egy klasszikus szabály:
    - Ha egy vásárló pelenkát és tejet vesz, akkor nagy eséllyel vesz sört is.
    - Ne lepődjünk meg ha a pelenkák után 6-os csomagban sört találunk!

# Társítási szabályok: 3. alkalmazás

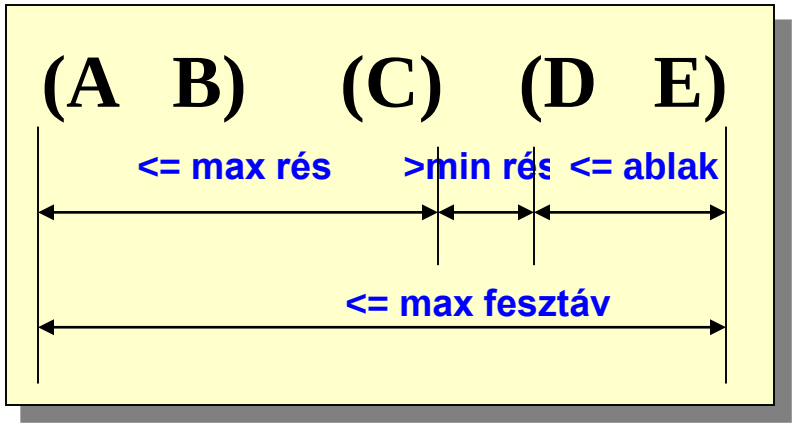
- Alkatrész gazdálkodás
  - Cél: egy háztartási berendezéseket javító vállalat szeretné előre látni a szükséges javítások fajtáit, hogy a megfelelő alkatrészekkel legyenek felszerelve a szervízautók és így a kiszállások számát csökkentsék.
  - Megközelítés: a különböző fogyasztói helyeken végzett korábbi javításokhoz szükséges eszközök és alkatrészek adatainak összegyűjtése és a közös előfordulások mintáinak feltárása.

# Szekvenciális mintázatok definíciója

- Adott *objektumok* egy halmaza úgy, hogy minden objektumhoz tartozik *eseményeknek egy sorozata*. Keressünk olyan szabályokat, amelyek a különböző események között minél erősebb szekvenciális függéseket jeleznek előre.




- A szabályokat az első felfedezett mintázatok alakítják ki. A mintázatokban előforduló eseményeknek időbeli peremfeltételeknek kell eleget tenniük.



# Példák szekvenciális mintázatokra

- Hibaüzenet a telekommunikációban:
  - (Átalakító_hiba  Túlzott_vezeték_áram)

    (Egyenirányító_riadó) --> (Tűz_riadó)
- Tranzakciók sorozata automatizált vásárlásnál:
  - Számítástechnikai könyvesbolt:

    (Bevezetés_a_Visual_C_be)  (Bevezetés_C++_ba) -->
    (Perl_kezdőknek, Tcl_Tk_nyelv)
  - Sportruházat bolt:

    (Cipő) (Teniszütő, Teniszlabda) --> (Sport_dzseki)

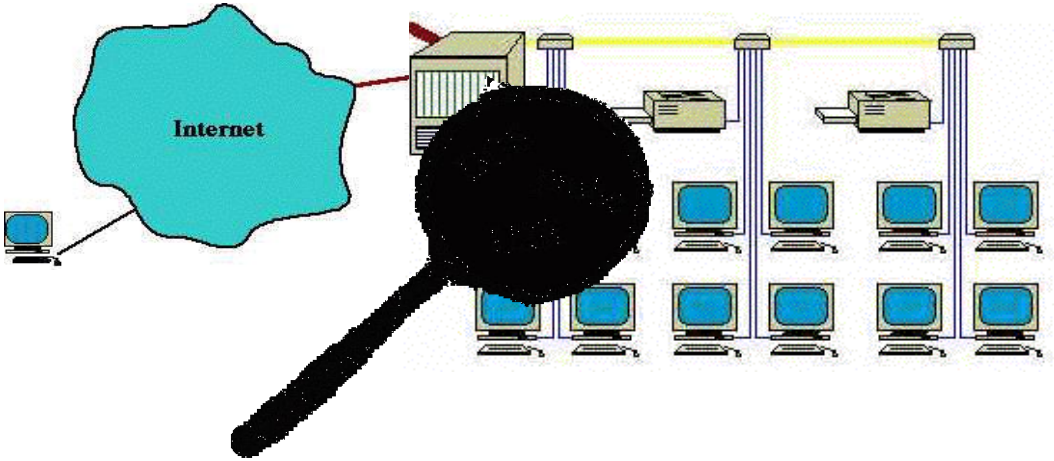
# Regresszió

- Jelezzük előre egy adott folytonos változó értékét más változók értékeit felhasználva, lineáris vagy nemlineáris függőséget feltételezve.
- Alaposan vizsgálták a statisztika és a neurális hálók területén.
- Példák:
  - Egy új termékből eladott mennyiség előrejelzése a reklámköltségek alapján.
  - A szélsebesség előrejelzése a hőmérséklet, a páratartalom, a légnyomás stb. segítségével.
  - A részvény-indexek idősorral való előrejelzése.

# Eltérés/Rendellenesség keresése

- A normális viselkedéstől szignifikáns eltérések keresése.
- Alkalmazások:
  - Hitelkártya csalások keresése


  - Hálózati behatolás érzékelése




*Egyetemi szinten átlagos hálózati forgalom esetén 100 millió kapcsolat jön létre naponta*

# Kihívások az adatbányászatban

- Skálázhatóság
- Dimenzió probléma
- Összetett és heterogén adatok
- Nem-hagyományos elemzés
- Adatminőség
- Jogosultság kezelés és elosztott adatok
- Adatvédelem
- Adatfolyamok