Gépi tanulás a gyakorlatban

Bevezetés

Motiváció

- Nagyon gyakran találkozunk gépi tanuló alkalmazásokkal
  - Spam detekció
  - Karakter felismerés
  - Fotó címkézés
  - Szociális háló elemzés
  - Piaci szegmentáció analízis
  - Hírek téma szerinti automatikus csoportosítása
- Jelentős kereslet a gépi tanulásban jártas programozókra

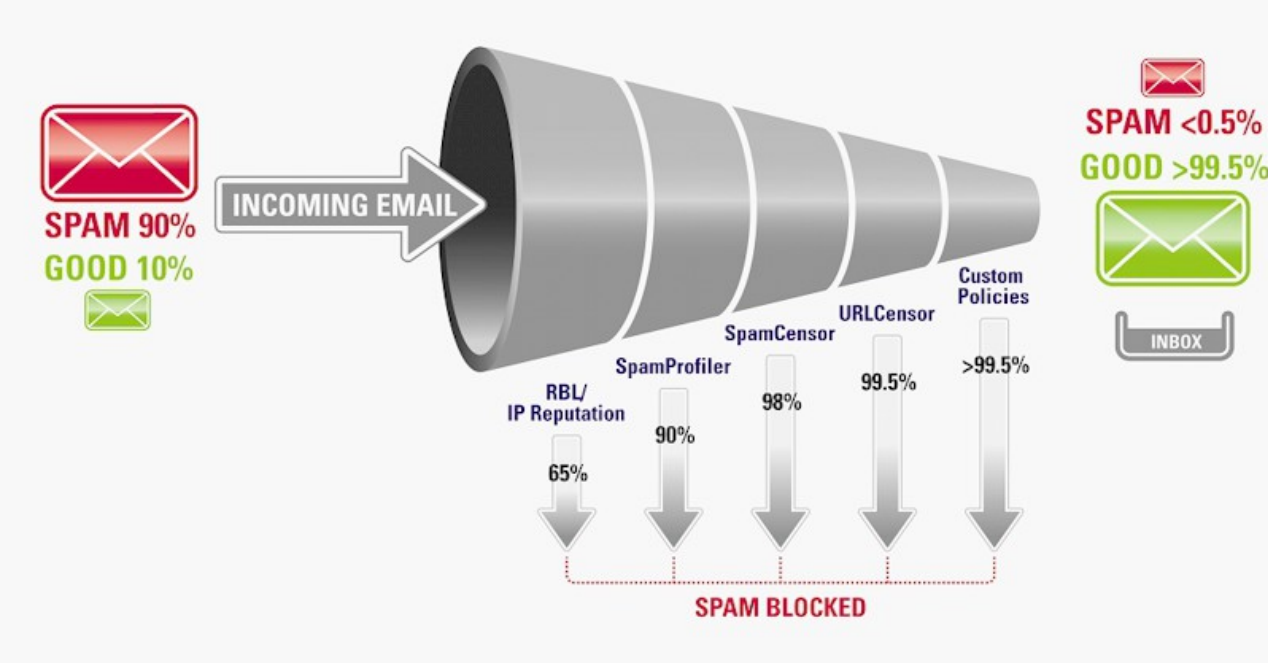
Spam detekció

Publikus adatbázisok:

UCI Spambase dataset

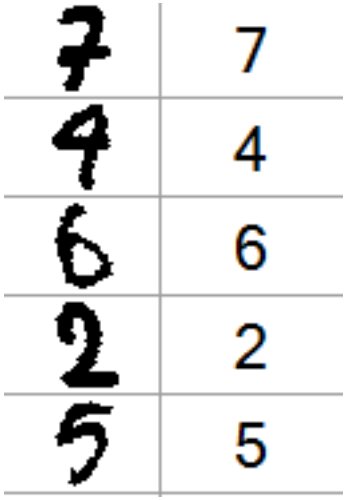
Spamassassin Public Corpus

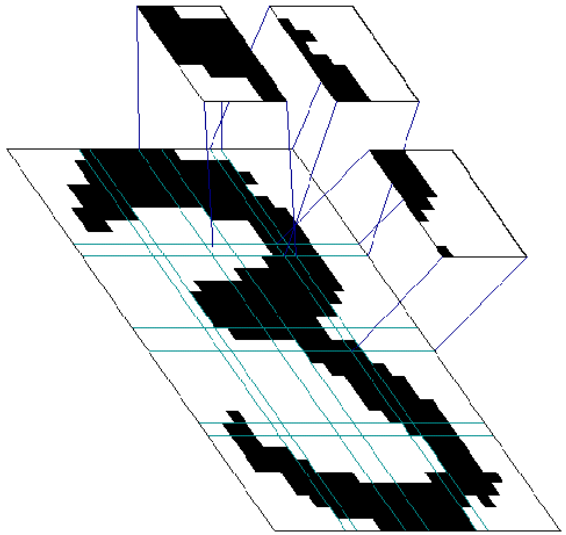
- Probléma: rengeteg kéretlen spam e-mail
- Megoldás: *szűrjük* ki a spam-et
- Spam levelek gépi úton történő felismerése nem triviális:
  - Szabály alapú megközelítés → gyorsan kijátszható, rugalmatlan
  - Jó megoldás: tanuljuk meg, hogy mi a spam
- Tanulás feltétele:
  - Gyűjtsünk *példákat*, amelyekben kézzel be van jelölve, hogy spam vagy nem.
  - Tanuló algoritmusok használatával készítsünk szűrőt.
  - Alkalmazzuk a szűrőt minden bejövő levélre



Karakter felismerés

- Probléma: digitális képeken szereplő – akár kézzel írott – karakterek képihez *rendeljük hozzá* a tényleges karaktert
- Nehézség: eltérő írásmódok, nyomtatási hibák, stb...
  → szabály alapú megoldás nehézkes, kevéssé hibatűrő

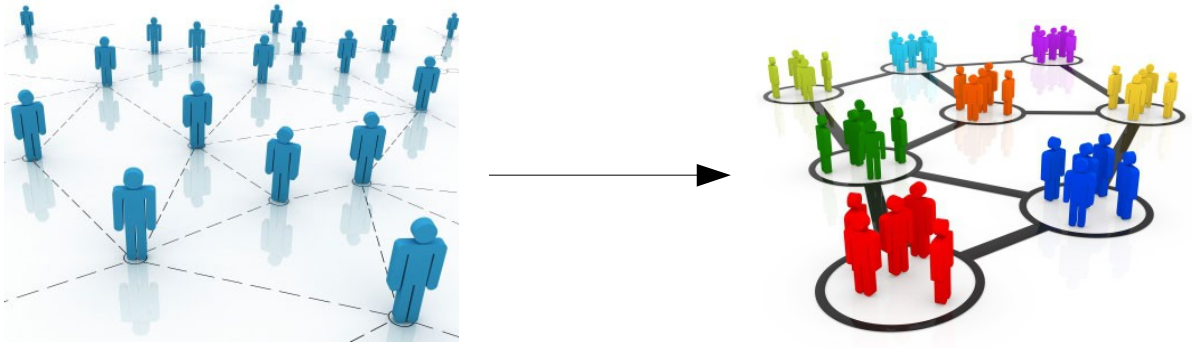




- Megoldás:
  - Gyűjtsünk adatbázisokat, ahol a képekhez, hozzá vannak rendelve a kívánt karakterek
  - Dolgozzuk fel a képeket
  - Alkalmazzunk tanuló módszert a leképezés megtanulására
  - A tanultakat használjuk újonnan írott karakterek felismerésére

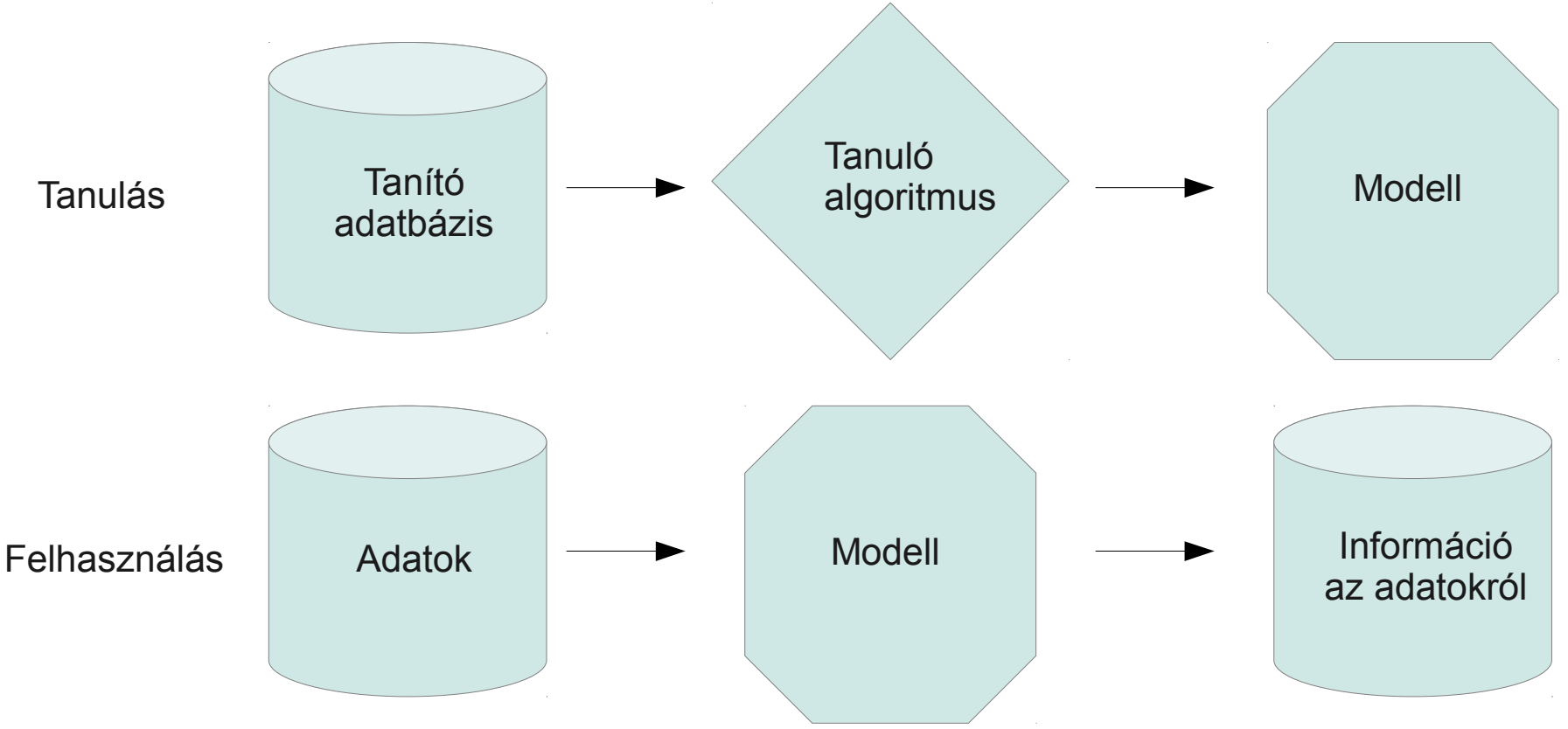
Szociális háló elemzés

- Probléma: Szeretnénk csoportokat azonosítani egy szociális hálóban
  - A „csoport” alatt bizonyos szempontból összetartozó, egyedeket értünk (pl. ugyan ott dolgoznak, hasonló ízléssel rendelkeznek)
  - A csoportokról nincs előzetes információnk ← Nagyon fontos különbség!!!
- Megoldás:
  - Adjunk hasonlósági függvényt
  - Gépi tanuló algoritmus segítségével azonosítsuk a csoportokat
  - Az új személyeket soroljuk abba a csoportba amelyhez (amely reprezentáns eleméhez) a legközelebb van



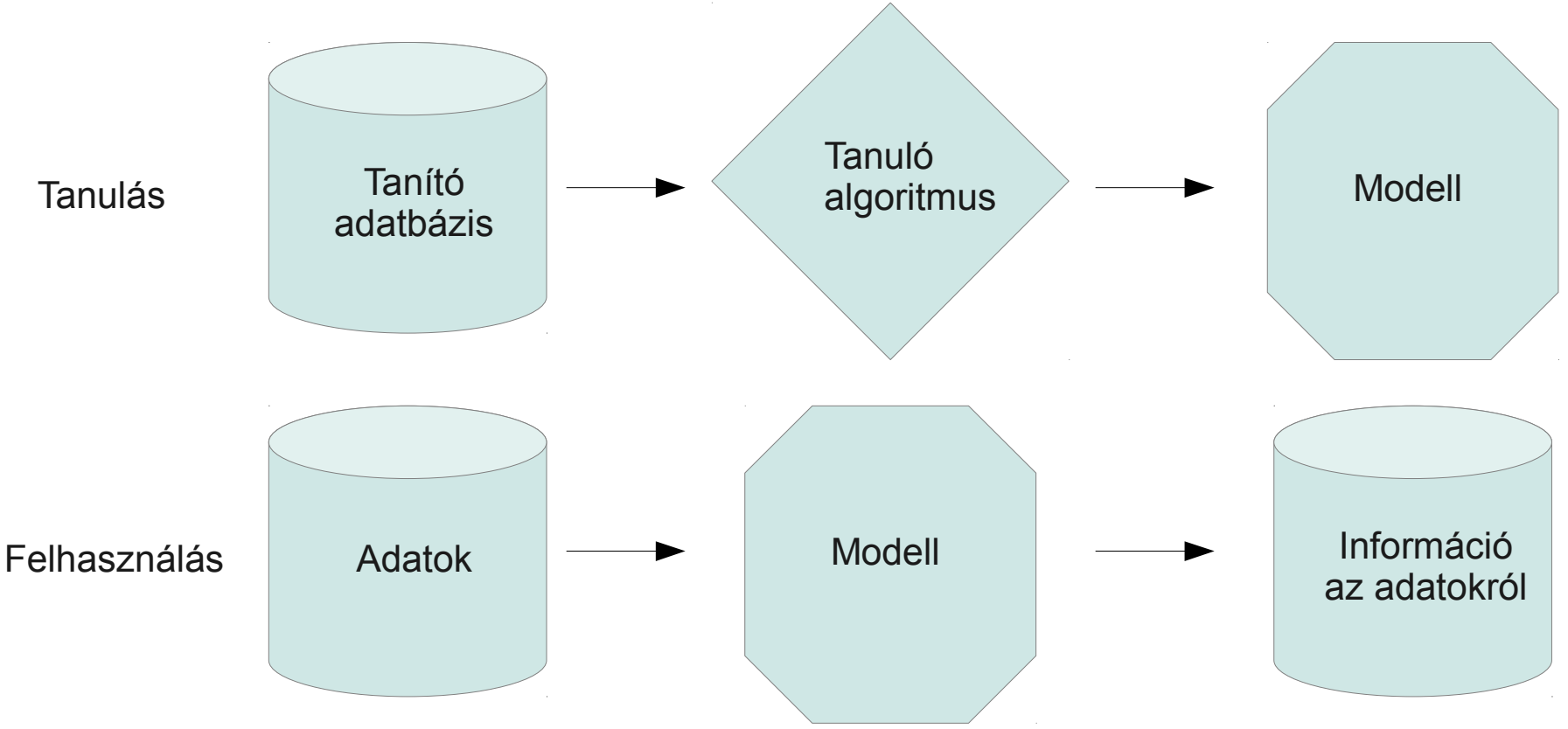
Gépi tanulás általában

- Mi volt a közös a fenti példákban? Két fázis: tanulás – felhasználás
- Spam detekció esetén:
  - *tanuló adatbázis:* (e-mail,spam/nem spam címke) párok halmaza
  - *modell:* tudás, ami alapján elvégezhető a címkézés (spam/ nem spam döntés)
  - *adatok:* e-mail-ek
  - *információ az adatokról:* spam/ nem spam címkék a megfelelő e-mail-ekre



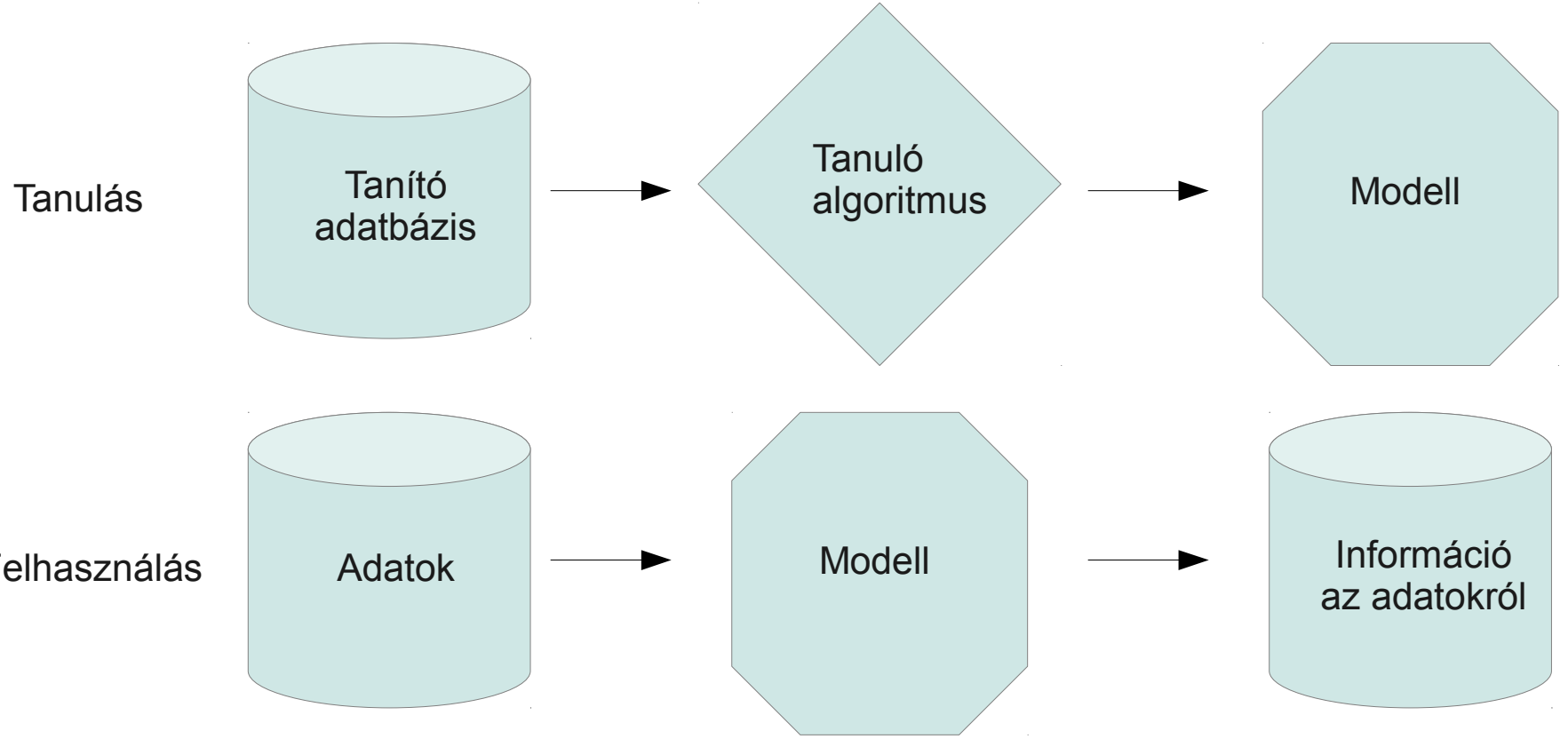
Gépi tanulás általában

- Mi volt a közös a fenti példákban? Két fázis: tanulás – felhasználás
- <u>Karakter felismerés esetén:</u>
  - *tanuló adatbázis:* (kép,karakter) párok halmaza
  - *modell:* tudás, ami alapján elvégezhető a címkézés
  - *adatok:* karaktereket ábrázoló képek
  - *információ az adatokról:* karakterek hozzárendelése a megfelelő képekhez



Gépi tanulás általában

- Mi volt a közös a fenti példákban? Két fázis: tanulás – felhasználás
- Szociális háló elemzés esetén:
  - *tanuló adatbázis:* szociális háló egyedei
  - *modell:* csoport (reprezentáns)
  - *adatok:* szociális háló egyedei
  - *információ az adatokról:* csoporthoz rendelés



Gépi tanulás általában

Szociális háló elemzés esetén:

- *tanuló adatbázis:* szociális háló egyedei
- *modell:* csoport (reprezentáns)
- *adatok:* szociális háló egyedei
- *információ az adatokról:* csoporthoz rendelés

Mi különbözik?

Eltérés az adatok szerkezetében!

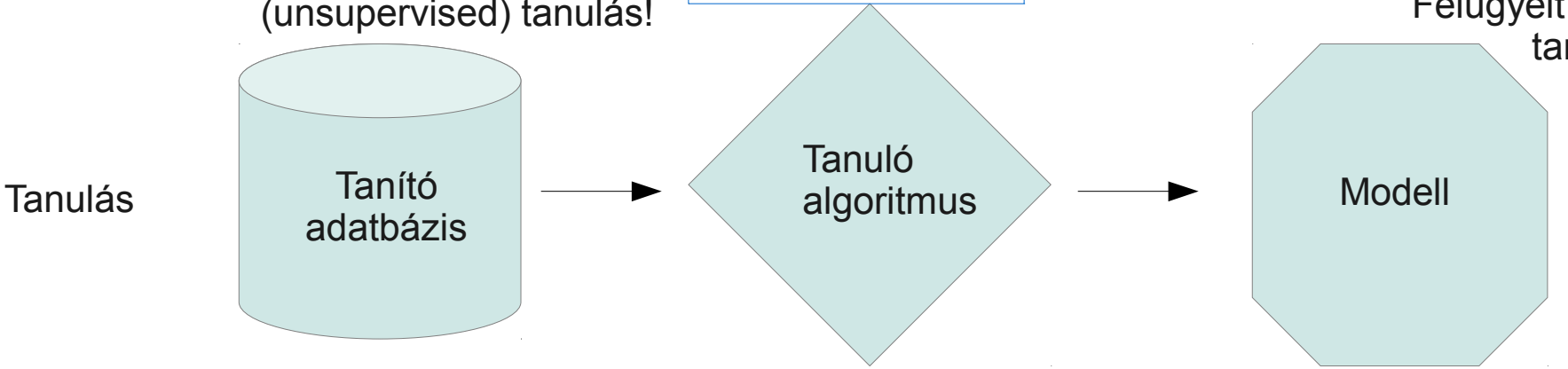
Karakter felismerés esetén többlet információ!

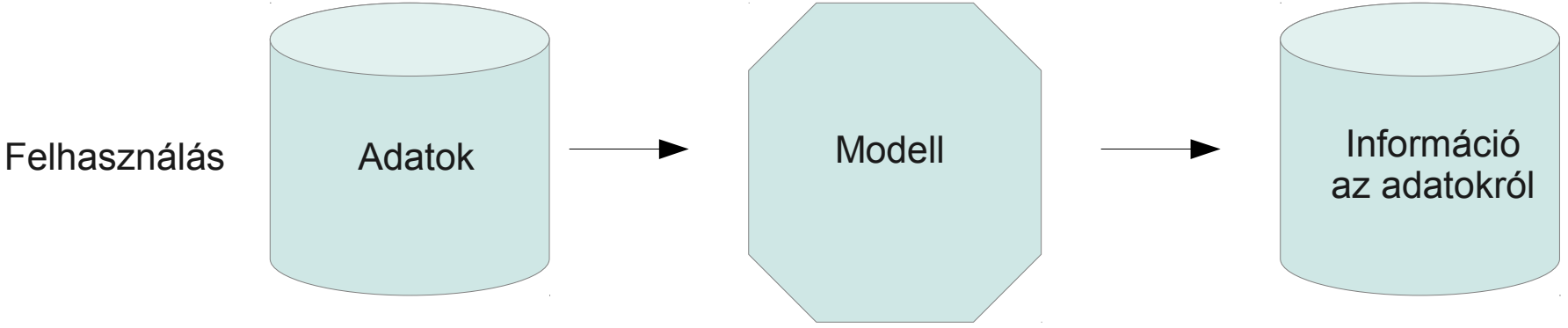
Karakter felismerés esetén:

- *tanuló adatbázis:* (kép,karakter) párok halmaza
- *modell:* tudás, ami alapján elvégezhető a címkézés
- *adatok:* karaktereket ábrázoló képek
- *információ az adatokról:* karakterek hozzárendelése a megfelelő képekhez

Felügyelet nélküli (unsupervised) tanulás!

Felügyelt (supervised) tanulás!

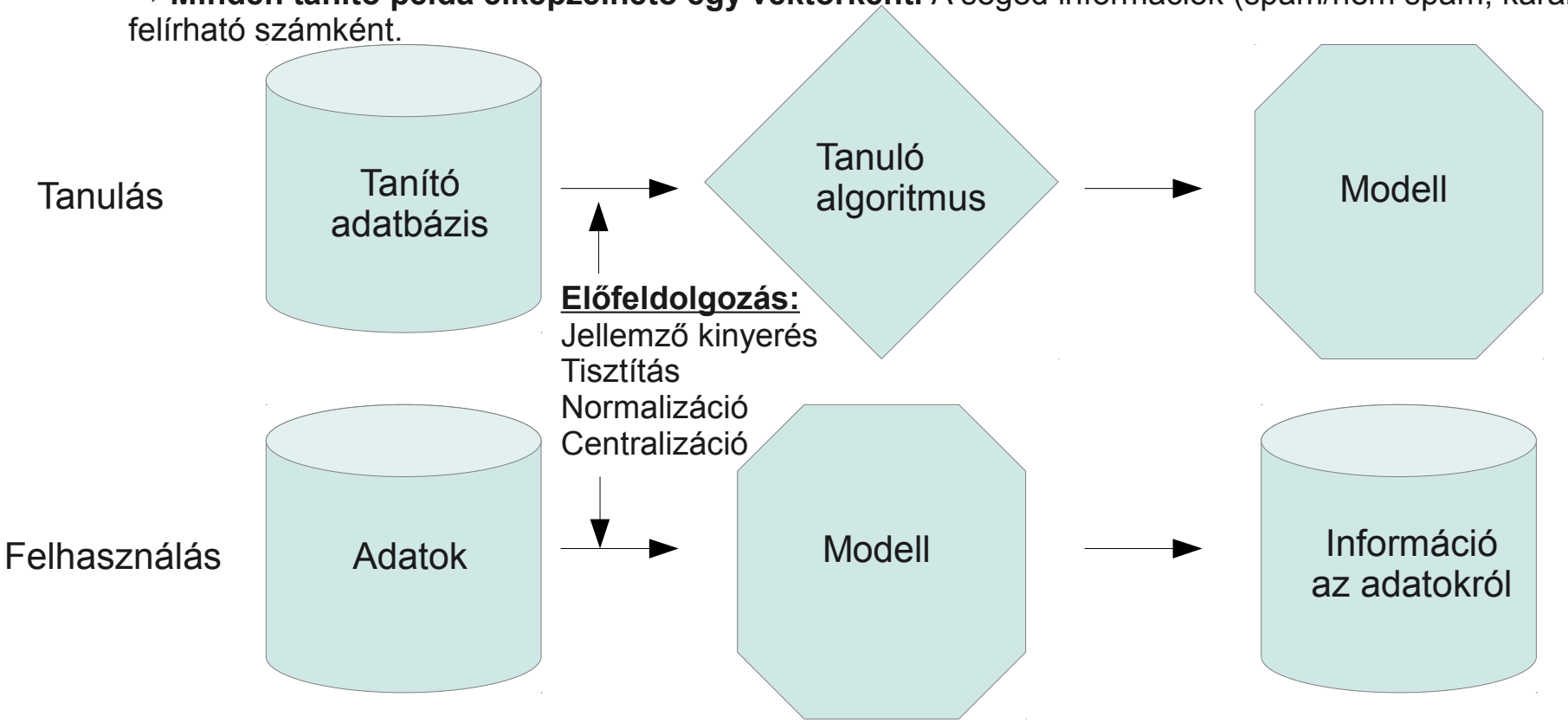




Jellemző kinyerés

- A tanulás és a felhasználás megkezdése előtt a nyers adatok feldolgozása szükséges
  - Jellemző kinyerés: nyers adatokból → R^d-beli vektor reprezentáció készítése → minden tanuló példa egy vektor. Példák:
    - *E-mail szövegek:* választott szavak (dimenzió) előfordulásaiból álló vektor
    - *Karakterek képei:* régiókhoz (dimenzió) kapcsolódó sűrűségi statisztikák
    - *Szociális háló egyedei:* választott tulajdonság, preferencia (dimenziók) meglétéből képzett vektor

→ **Minden tanító példa elképzelhető egy vektorként.** A segéd információk (spam/nem spam, karakter) felírható számként.



Felügyelt/felügyelet nélküli tanulás

- A gépi tanulási feladatok és így a felhasznált módszerek feloszthatók „informáltság” szerint:
  - Felügyelt tanulás/módszer: A tanító adatbázis olyan, hogy a tanuló példákhoz a feladat megoldására vonatkozó *közvetlen információk is rendelkezésre állnak*. Pl. *osztályozási* feladat, ahol az osztálycímkék adottak (pl. karakter felismerés, spam detekció)
  - Felügyelet nélküli tanulás/módszer: A tanító adatbázisban nincs közvetlen segéd információ a megoldásra vonatkozóan. Pl. a módszernek kell felderítenie a nyers adatok alapján az adatbázisban rejlő belső struktúrákat, csoportokat; azaz *klaszterezni*, csoportokba sorolni kell azokat (pl. szociális hálózatban azonos preferenciával rendelkező csoportok azonosítása)
  - Félig felügyelt tanulás/módszerek: A *fenti két típusú adatokat egyszerre tartalmazó adatbázissal rendelkező feladatok*. Pl. természetes nyelven íródott hírek valamilyen szempont szerinti releváns/nem releváns osztályozása során, felhasználhatunk az interneten elérhető bármilyen szövege, mint címkézetlen adat.

Kvíz

1. Kvíz: Szeretnénk *előrejelezni ingatlanok árait*.
Ehhez a rendelkezésünkre áll egy olyan előfeldolgozott adathalmaz, aminek sorai az egyes ingatlanokat írják le. Az adathalmaznak két oszlopa van.

- Az első az *ingatlanok területét*,
- a második az *ingatlanok árait tartalmazza*.

Ez a probléma:

1. Felügyelet nélküli
2. Felügyelt
3. Félig felügyelt
4. Nem gépi tanuló feladat

Felügyelt tanulás – Regresszió

- Legyen adott az előbb említett adatbázis.

  Szeretnénk *előrejelezni ingatlanok árait*.
  Ehhez a rendelkezésünkre áll egy olyan előfeldolgozott adathalmaz, aminek sorai az egyes ingatlanokat írják le. Az adathalmaznak két oszlopa van.

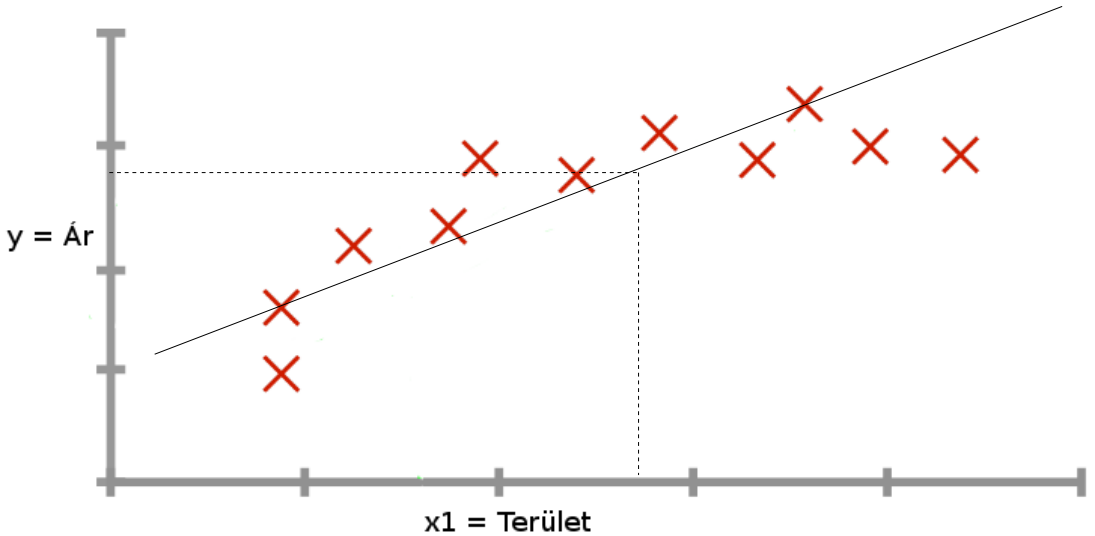
  - Az első az *ingatlanok területét*,
  - a második az *ingatlanok árait tartalmazza*.

- Ez megjeleníthető az ábrán látható formában

Felügyelt módszer, valós segéd információval → regressziós feladat

A modell egyenes (sík) → Lineáris regresszió

- Az egyenes egy lehetséges modellt ábrázol, amit a *tanítás során* előállítottunk
- A felhasználás során a bejövő példák (ár nélkül, csak terület) árát *az egyenes határozza meg (modell)*.



Felügyelt tanulás – Osztályozás

- Legyen adott egy előfeldolgozott spam detekciós adatbázis. Az előfeldolgozás során két szó meglétét vizsgáltuk:
  - rolex: (x1 értéke *n*, ha *n*-szer szerepel az e-mailben a rolex szó)
  - elad: (x2 értéke *n*, ha *n*-szer szerepel az adott e-mailben az elad szó)

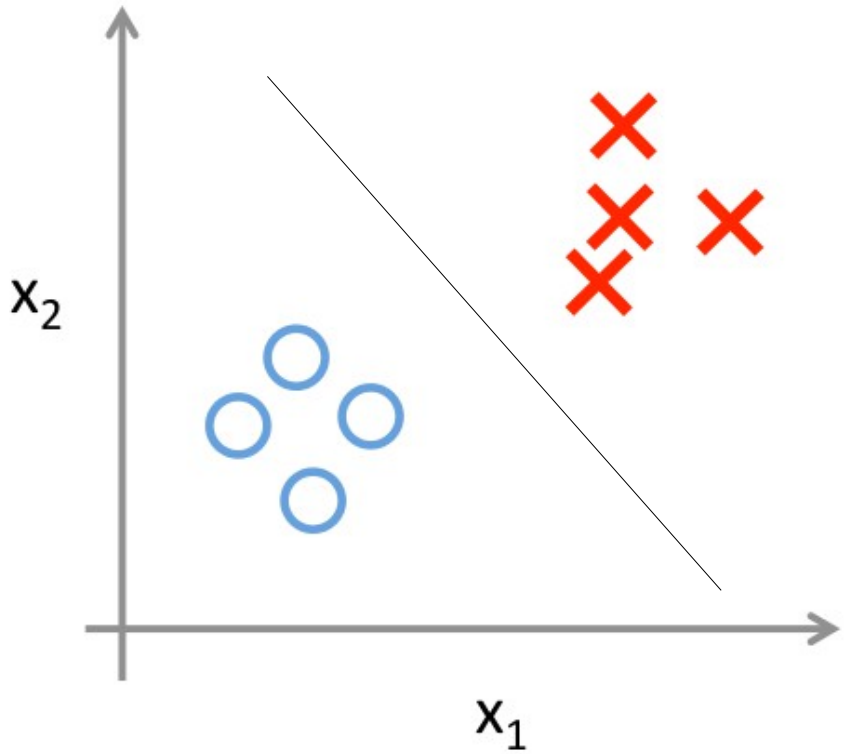
Felügyelt módszer, diszkrét segéd információval → osztályozási feladat

A modell egyenes (sík) → Szeparáló hipersík

Az osztály címke bináris (y): 1 (piros x), ha az adott levél spam, 0 (kék o), ha nem

Szeretnénk *előrejelezni hogy egy levél spam vagy nem.*

- Ez megjeleníthető az ábrán látható formában
- Az egyenes egy lehetséges modellt ábrázol, amit a *tanítás során* előállítottunk
- A felhasználás során a bejövő példák (x1,x2) címkéjét az alapján határozzuk meg, hogy az egyenes *(modell) melyik oldalára esik.*



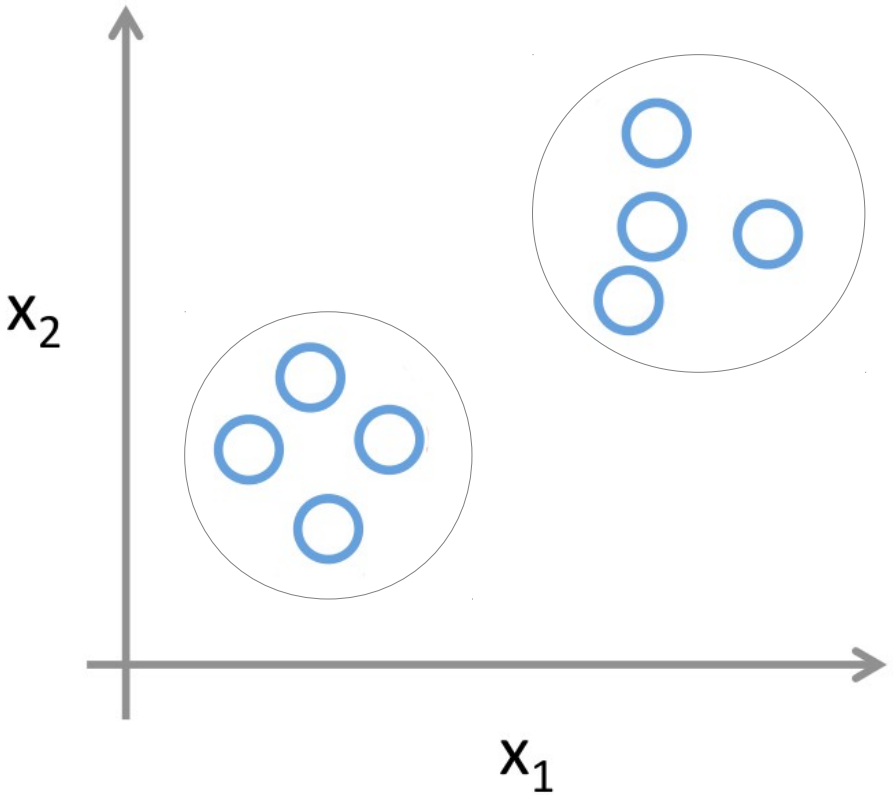
Felügyelet nélküli tanulás – Klaszterezés

- Legyen adott egy előfeldolgozott piac szegmentálási adatbázis. Az adatbázis emberek film kölcsönzési szokásairól tartalmaz információt 2 dimenzióban:
  - akció film kölcsönzése: (x1 értéke *n*, ha *n* darab akciófilmet kölcsönzött ki az adott ember 1 hét alatt)
  - vígjáték film kölcsönzés: (x2 értéke *n*, ha *n* darab vígjátékot kölcsönzött ki az adott ember 1 hét alatt)

Felügyelet nélküli módszer módszer, csoportok keresése → klaszterezés

Szeretnénk *azonosítani a hasonló szokásokkal rendelkező kölcsönzők csoportjait.*

- Ez megjeleníthető az ábrán látható formában
- Az körök egy lehetséges klaszterezést ábrázolnak (modell)
- Modelleket középpontjukkal azonosíthatjuk
- A felhasználás során a bejövő példák (x1,x2) csoportját az alapján határozzuk meg, hogy *melyik kalszter középponthoz (modell) esik közelebb.*



Weka demo

- A kurzus további részében konkrét tanuló algoritmusok megismerésére fogunk fókuszálni.
- Szerencsére, ha éles alkalmazásban akarjuk használni őket, akkor nem kell mindent leprogramoznunk, mivel számos könyvtár/alkalmazás áll a rendelkezésünkre:
  - Weka
  - Mahout
- Azt viszont nagyon fontos látni, hogy – megfelelő ismeretek hiányában – nagyon könnyen lehet hibásan alkalmazni az algoritmusokat! A félév során erre több példát is fogunk látni.