ADATALAPÚ MEGOLDÁSOK

Alapadatok

- Tárgyfelelős: Dr. Papp Dávid, TMIT
- Előadók: Dr. Papp Dávid, TMIT; Dr. Szűcs Gábor, TMIT
- Órák: **ea**: kedd 10:15-től, **gyak**: péntek 12:15-től
- Órai diák, feladatok:

  https://edu.vik.bme.hu/

Követelményrendszer:

- ZH + vizsga
- Megajánlott jegy: ZH + gyakorlati tesztek + kiselőadás

Megajánlott jegy

- ZH teljesítése a 7. héten (04.01)
  VAGY pót ZH teljesítése a 9. héten (04.12)

- Gyakorlati tesztek teljesítése (min. 3 db)
  - 5. héten (03.18)
  - 13. héten (05.13)

- Megajánlott jegy számítása:

$$IF\ 4{,}4 > (0{,}6 * ZH\_jegy\ +\ 0{,}4 * (\#teljesített\_gyakorlati\_tesztek))\ \geq 3{,}4:$$
$$megajánlott\ jegy = 4$$

$$ELIF\ (0{,}6 * ZH\_jegy\ +\ 0{,}4 * (\#teljesített\_gyakorlati\_tesztek))\ \geq 4{,}4:$$
$$megajánlott\ jegy = 5$$
$$ELSE:$$
$$vizsga$$

+ kiselőadás 8. héten (04.05) = 1 db teljesített gyakorlati teszt

Tantárgy tematikája

- 4 blokk
  - Numpy
  - Pandas
  - Vizualizáció
  - Gépi tanulás
- Minden blokk anyagából gyakorlati teszt
  - 30-45 perc
  - Önálló feladatmegoldás
  - GO / NO GO

Adat

- Jelentés ábrázolása (gépek számára érthető formában)
- Számokkal leírható, rögzíthető, feldolgozható, megjeleníthető
- Keletkezésük körülménye lehet:
  - kutatás
  - szenzorok
  - tranzakciók
  - digitális interakciók
  - számítások
  - mesterséges intelligencia
  - stb.

Adat

- Felhasználási területek:
  - döntéshozatal
  - probléma megoldás
  - önvezető járművek
  - szórakoztatás
  - információ közlés
  - biztonság
  - stb.

Adattípusok

- Elemi adat: nem bontható további részekre
  (pl. irányítószám, alapfizetés)
- Kvalitatív adat: szövegesen kifejezett adat
  (pl. komment, értékelés)
- Kvantitatív adat: számokkal kifejezett adat
  (pl. egyedek száma, hőmérséklet)
  kvantifikálás: kvalitatív → kvantitatív
  (pl. értékelés besorolása 1-től 5-ig terjedő skálán)
- Nyers/forrás adat: feldolgozás előtti állapot
  (pl. közvetlenül szenzorokból)
- Referencia adat: stabil, állandó értékkel bíró adat
  (pl. π)

Adattípusok

- Biztos adat: megbízható forrásból származó, elfogadott módszertan alapján ellenőrzött, tény adat
  (pl. orvosi eredmények vérvételt követően)
- Bizonytalan adat: általában kvalitatív eredetű, megfigyelésekből gyűjtött, kvantifikált adat
  (pl. páciensek elmondják tüneteiket)
  Megjegyzés: a bizonytalan adat nem azt jelenti, hogy használhatatlan, gyenge adat. Sok esetben csak ilyen jellegű adatok állnak rendelkezésünkre.
- Mérhető adat: megkérdőjelezhető technikával gyűjtött vagy statisztikailag nem szignifikáns eredetű adat
  (se nem biztos, se nem bizonytalan)

Adat típusai

- Nominális $=, \neq$
  (pl. telefonszám, színek)
- Ordinális $<, >$
  (pl. rangsorban elfoglalt hely)
- Intervallum skálájú $+, -$
  (pl. Celsius-fokban mért hőmérséklet)
- Arány skálájú $\cdot, \div$
  (pl. kilogrammban mért tömeg)

Metaadat

- **Metaadat** minden olyan adat, amely más adatokról szól
  Definíció: metaadat = adat az adatról
  (pl. film hossza, zene műfaja, stb.)
- Szükségünk van kiegészítő adatokra ahhoz, hogy az adatokat kezelni és értelmezni tudjuk. Ezeket a kiegészítő adatokat nevezzük metaadatoknak.
- A metaadatok önmaguk is adatok, így róluk is lehetnek további metaadatok.
- Megkülönböztetjük a
  - leíró és a
  - szemantikus metaadatokat.

A metaadatok kategorizálása

- **Leíró metaadat:**
  jelentése nem kapcsolódik közvetlenül az adat jelentéséhez. Az adat keletkezésének és/vagy módosításának körülményeit írja le.
  (pl. dokumentum szerzője, videó hossza, utolsó módosítás dátuma, GPS adat, kamera lencse fókusz értéke, stb.)
- **Szemantikus metaadat:**
  az adat jelentéséről hordoz információt
  (pl. dokumentum kulcs szavai, film jelenetek címkéi, stb.)

Adatmodellek

- Az adatokat azért tároljuk és kezeljük, hogy később különböző célokra felhasználhassuk. Hatékonyabb felhasználás lehetséges ha nem „ömlesztett”, hanem szervezett adatokat kezelünk.
- Szükséges:
  - A tártolt adatok szerkezetét és felhasználási szabályait leíró modell (formális jelölésrendszerrel az adatok, adatkapcsolatok, és az azokon végrehajtható műveletek leírása.)
  - A leíró módszereknek szabványosaknak kell lenniük
- A különböző mértékben szervezett informatikai adatkezelés különböző adatmodellezést tesz lehetővé.

Strukturált adatok

- Ha a tárolás által meghatározott struktúra jól illeszkedik az adatok által hordozott információ struktúrájához, akkor **strukturált adatok**ról beszélünk.
- A meglévő és a jövőben eltárolandó adatok struktúrája leírható egy állandó sémával (adatbázis esetén egy adatbázis sémával).
- Példa: film adatbázis a filmek címéről, alkotóiról, jogtulajdonosairól, technikai jellemzőiről (időtartam, képminőség, hangrendszer, hangsávok, stb.)

Strukturálatlan adatok

- Ha az adatok által megjelenített információ értelmesen nem strukturálható, az adathalmaz egésze hordozza az információt, akkor **strukturálatlan adatok**ról beszélünk.
- Példa: pixelgrafikus kép, ahol az egész kép hordozza az információt (pl. emberi arckép), a képet alkotó biteket feldarabolni – tartalmilag – értelmetlen.

Félstrukturált adatok

- Ha az adatok tárolása által meghatározott struktúra nem jól illeszkedik az adatok információtartalma által meghatározott struktúrához (az adatok értelméhez, azaz szemantikájához), akkor **félstrukturált adatok**ról beszélünk.
- Példa: közösségi média-profil. Vannak adatmezők, ám azokban tárolási struktúrával el nem látott adatok lehetnek.

Adatok felhasználási céljai

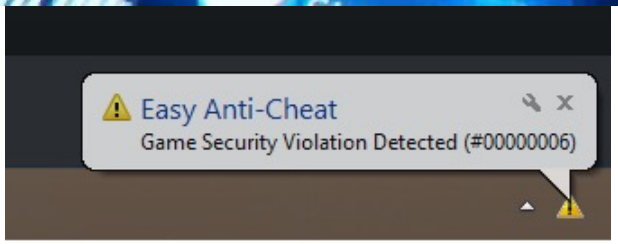
- **Ellenőrzés** - az adatokat azért gyűjtjük és elemezzük, hogy ezzel tudjuk felügyelni, ellenőrizni a megfigyelt folyamatokat.
- **Hatékonyságnövelés** - az adatok ismeretében egyes folyamatok optimalizálhatók, fejleszthetők.
- **Önreflexió** - ebben az esetben az adatokat azok a szereplők használják fel közvetlenül, akikről maguk az adatok szólnak.

Adatalapú megoldások

- **Adatalapú** azt jelenti, hogy a cselekvések véghezvitele és az irányelvek kialakítása adatok felhasználásával történik.
- Lehetőség nyílik hatékonyabb eredmények elérésére, a „találgatás”-hoz képest.
- Az adatalapú megoldások szuboptimálisak, ha
  - félreértelmezett-, ismeretlen-, hibás-, hiányzó adatok,
  - helytelenül kialakított adatmodellek,
  - félretervezett algoritmusok és/vagy
  - hibás emberi következtetések vannak.

Adatalapú megoldások (példa)

- Biztonság és csalás felderítés



Adatalapú megoldások (példa)

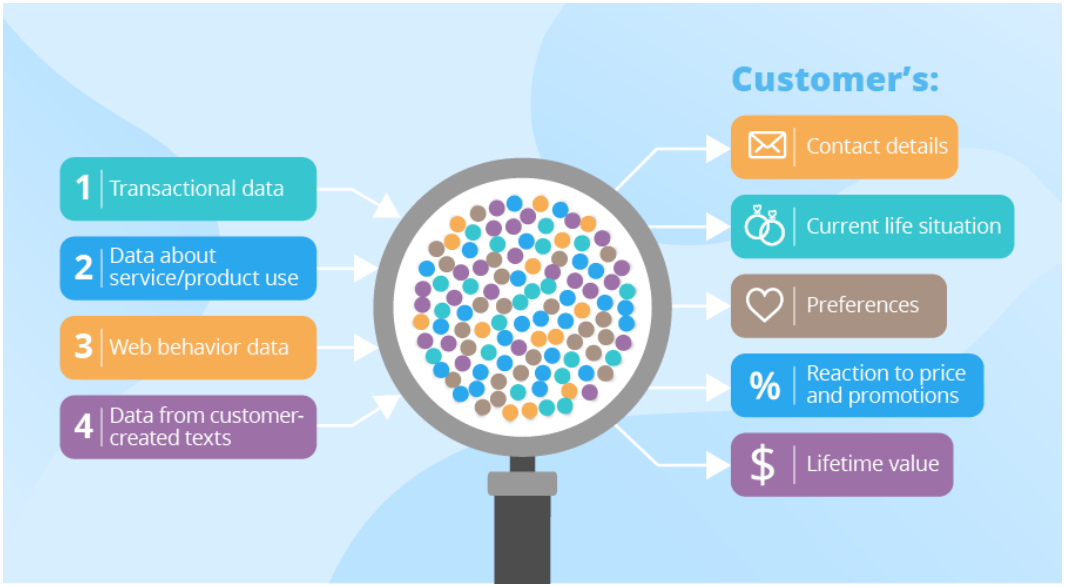
- Banki automatizálás

Adatalapú megoldások (példa)

- Időjárás előrejelzés

Adatalapú megoldások (példa)

- Ügyfélanalitika



Szoftver eszközök

Anaconda

- Ingyenes, nyílt forrás kódú
- Python és R
- Adattudomány és gépi tanulás
- https://www.anaconda.com/distribution/

Python

- Általános célú, magas szintű programozási nyelv
- Jól szervezett, könnyen áttekinthető kód
- Számos könyvtár csomag
  - NumPy: tömbösített, homogén adatok kezelése
  - Pandas: címkézett, heterogén adatok kezelése
  - SciPy: tudományos számítási feladatok
  - Matplotlib: publikáció minőségű vizualizáció
  - IPython: interaktívan végrehajtható, megosztható kód
  - Scikit-Learn: gépi tanulás
  - stb.

(Ana)conda virtual environment

- Különálló, letisztult környezet
- Minden projekthez új környezet
- Több eltérő verzió párhuzamosan
- Nincs verzióütközés
- Függőségek kezelése
- Hasonló a python virtuális környezetéhez
  DE nem csak python nyelvhez (R)

conda vs. pip

- pip: python csomag könyvtár kezelő
- venv: python virtuális környezet kezelő
- conda: csomag könyvtár ÉS virtuális környezet kezelő

```
conda create --name myenv
```

```
conda create --name myenv python=3.6
```

```
conda create --name myenv numpy
```

Windows: activate / deactivate

Linux: source activate / source deactivate

Jupyter Notebook

- Interactive Python (IPython)
- Önálló python szkriptek
- Jupyter Notebook: ezek kombinációja
- Anaconda disztribúció része
- Grafikus, böngésző alapú interfész
- Kétféle cella:
  - futtatható kód
  - formázott szöveg (markdown)
- Notebook szerver indítása
  - Anaconda prompt (vagy terminál): >> jupyter notebook
  - Anaconda Navigator: GUI