# Gyors sikerek adatbányászati módszerekkel

## Kezdő adatbányászati workshop

Petrócziné Huczman Zsuzsanna
2015.10.13.

# Bemutatkozás

- BME, műszaki informatika szak, adatbányászati szakirány – Citibank
- Data Explorer Kft.– Szerencsejáték Zrt., Magyar Telekom
- Data Solutions Kft.– MT, UPC, Kopint-Datorg, OTP, Erste, Lombard...
- Andego Kft.

# Bemutatkozás

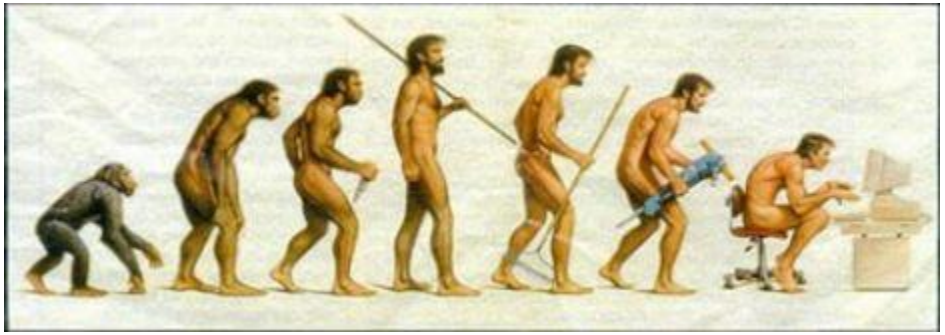
# Adatbányászat alapok – tematika

- Adatbányászati történelem
  - A halászattól a tudósokig
  - IBM Watson
- Adatbányászati projekt folyamata
  - Az igénytől az adaton át az eredményig
- Rapid Miner – ismerkedés
- 2 projekt Rapid Miner segítségével

# Adatbányászati történelem

- Mikor és miért merült fel az igény?

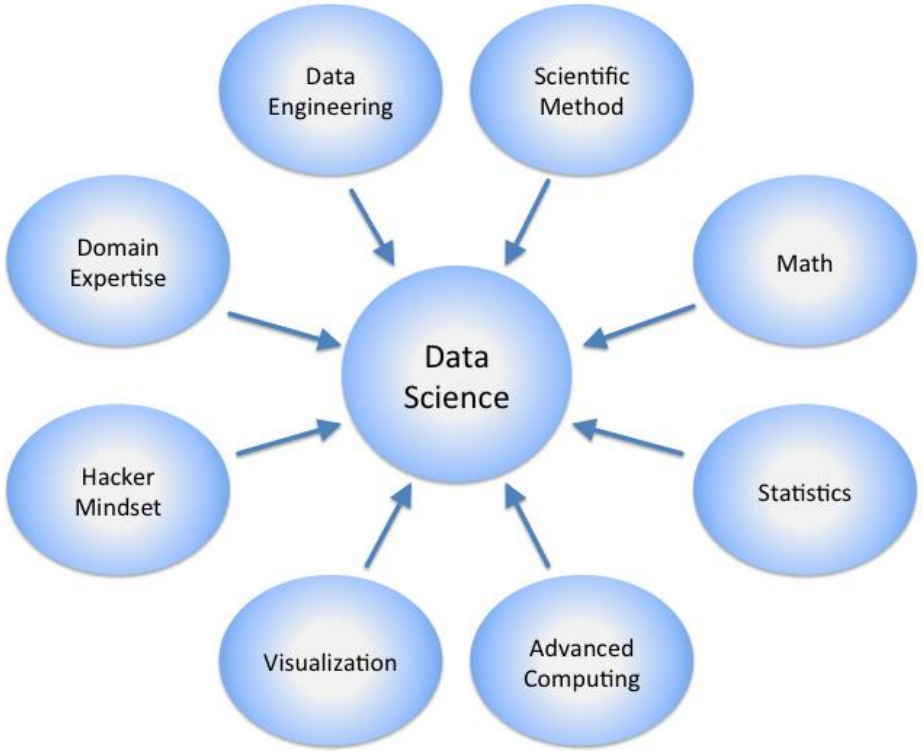
  Az adattárolás története infografika
- 1960 – Adathalászat
- 1980 – Adatbázis bányászat
- 1990 – Adatbányászat
- 2011 – Data Science



# Adatbányászati történelem

## Mi a különbség?

- Statisztikus?
- Adatbányász?
- Data Scientist?



# Adatbányászati történelem

Az adatok mennyisége, minősége

- Mennyi adat jön létre egy perc alatt?
- Az adatok 90-95%-a az utóbbi hány évben jött létre?

- Magyar „rögvalóság”
  - Céginfó adatbázis – 1.8 GB (tömörítve)
  - Vagyonbizt. utóbbi 2 évi káreseménye – 1.5 GB
  - Prezi.com logfájlok: 800 GB/nap

# Adatbányászati történelem

- 2011 – IBM Watson in Jeopardy

Demo

# Adatbányászati projekt folyamata

- 2 projektet viszünk végig ma
  - Churn projekt – Banki lemorzsolódás-elemzés
  - Marketing projekt – Banki ügyfél szegmentáció

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Adatbányászati projekt folyamata

- Milyen típusú feladatok oldhatóak meg tipikusan gépi tanuló algoritmusokkal:
  - „Kinek adhatunk hitelt?”– Osztályozás, predikció
  - „Kik a vevőink?”– Szegmentálás
  - „Akik ezt vették...” – Ajánlórendszer
  - „Holnap mennyi pénzt hoznak be?” – Idősor-előrejelzés

# Adatbányászati projekt folyamata

- Speciális feladatok:
  - „Melyik email spam?” – Szövegbányászat
  - „Írd be amit keresel” – Keresőmotor
  - „Milyen hatóanyaggal lehet gyógyítani téged” – orvosi adatbányászat
  - „Kik a csalók?” – Hálózat-elemzés

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Adatbányászati projekt folyamata

- ADAT
  - Honnan?
    - Belső forrás
    - Külső forrás (internet, piackutatás...)
  - Mennyit?
    - Hogyan lehet több adatunk?/Hogyan mintavételezzünk?
    - Mi a jó célváltozó-arány?
    - Gondoljunk a validálásra is!

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Adatbányászati projekt folyamata

- Az adat közelebbről
  - Objektumot leíró változók (+ Célváltozó)
- Adat-átalakítás
  - Aggregálás, adatforrások összevonása
  - Új változók generálása
  - Hiányzó értékek kezelése
  - Adattípus-átalakítások

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Adatbányászati projekt folyamata

- Milyen gépi tanuló algoritmus-családok vannak?
  - Döntési fák
  - Neurális hálók
  - Klaszterező eljárások
  - Regressziós modellek (Idősor-előrejelzés, predikció)
  - ...

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Adatbányászati projekt folyamata

- Hogyan mérjünk teljesítményt?
  - Modell-építéstől független adatbázison!
    - Vagy azonos időszakon, eddig nem látott adatokon
    - Vagy különítsünk el teszt-időszakot
- Külön adathalmazok:
  - Tanítás, Tesztelés, Validálás

# Adatbányászati projekt folyamata

- Határozzuk meg az üzleti célt
- Gyűjtsünk, szerezzünk adatokat
- Készítsük elő az adatot
- Építsünk modelleket
- Nézzük meg, hogy teljesít
- Építsük be a folyamatokba

CRISP-DM

# Rapid Miner – tematika

- Alap-információk
- Technológiai háttér
- Építőkövek
- Előnyei

# Alap-információk

- Egyetemi fejlesztés
- Open source
- 3 verzió
  - Alap
  - Community
  - Professional (2000 $ / év)
    Többféle adatforrás + MarketPlace+ Support + Cloud
- Tutorial, példa-adatbázisok, wiki, fórum…

# Technológiai háttér

- XML alapú, GUI felület
- Párhuzamos futtatás
- Java library-ként használható
- Market
  TextMining, DataStream, ...
- Operátorok „egyszerűen” írhatóak

# Építőkövek

- Alapegységek: operátorok
- Az operátorok rendelkeznek:
  - Bemenet
  - Kimenet
  - Opcionálisan belső operátorok

# Előnyei

- Egyszerűen telepíthető
- Vizualizációs eszközök
- Számos algoritmus, egyszerű fejleszthetőség
- Platform-független

# Projektek

TELEPÍTÉS

# Churn projekt

- Válasszunk egy szimpatikus bankot

POSTABANK

# Churn projekt

- Üzleti cél:
  - Szeretnénk tudni, hogy az elkövetkező fél évben ki fogja nagy valószínűséggel lemondani a szolgáltatását.
- Adatbányászati célra lefordítva:
  - Milyen ügyfélkarakterisztikával rendelkezik az a csoport, aki az utóbbi időben lemondta a szolgáltatást.

Osztályozás

# Churn projekt

- Adatok – milyen adatokat kérjünk?
  - Mindent, ami rendelkezésre áll az ügyféllel kapcsolatban.
    - Viselkedési adatok
    - Demográfia
    - Termékei
    - Célváltozó

# Churn projekt

- Adatok feldolgozása
  - Források egyesítése
  - Alapstatisztikák (mit nézzünk?)
    - Extrém értékek kiszűrése
    - Új változók (ötletek?)
  - Algoritmus-függő
    - Hiányzó adatok kezelése
    - Adattípus-átalakítás

# Churn projekt

- Modellezés
  - Próbáljunk ki többféle modellt, többféle paraméterezéssel

# Churn projekt

- Validálás
- Válasszuk ki a legjobb modellt

# Churn projekt

- Beépítés az üzleti folyamatokba
  - Pl. Ajánljunk nekik kedvezményt, hogy megtartsuk őket

# Marketing projekt

- Üzleti cél
  - A bankunk célzott marketingakciókra készül, hogy így növelje az ügyfeleknek értékesített szolgáltatásokat.
- Adatbányászati cél
  - Csoportosítsuk az ügyfeleket annak alapján, hogy azok milyen gyakran használják a bank adott szolgáltatásait.

Szegmentáció

# Marketing projekt

- Adatok – milyen adatokat kérjünk?

100.000 ügyfél viselkedési történetét tartalmazza. A rögzített tranzakciók típusa:

- hagyományos banki tranzakciók (*TBM - traditional banking methods*);
- ATM tranzakciók (*ATM - automatic teller machine*);
- POS tranzakciók (*POS - point of sale*);
- ügyfélszolgálati tranzakciók (*CSC - costumer service*).

# Marketing projekt

- Adatok feldolgozása
  - Alapstatisztikák (mit figyeljünk?)
    - Hibás az adatbázis!
  - Transzformáljuk az adatokat; új változó
  - Algoritmus-függő
    - Hiányzó adatok kezelése
    - Adattípus-átalakítás

# Marketing projekt

- Modellezés
  - Próbáljunk ki több modellt, többféle paraméterezéssel

# Marketing projekt

- „Validálás”
  - A szegmentációs modelleket nehéz „validálni”, a cél, hogy minél jobban különböző, jól leírható, magyarázható szegmenseket kapjunk.
  - Válasszuk ki a legjobb modellt

# Marketing projekt

- Beépítés az üzleti folyamatokba
  - Az eredmények alapján milyen javaslatot tennénk a marketing osztálynak?

# Köszönöm a figyelmet!

huczman@andego.hu