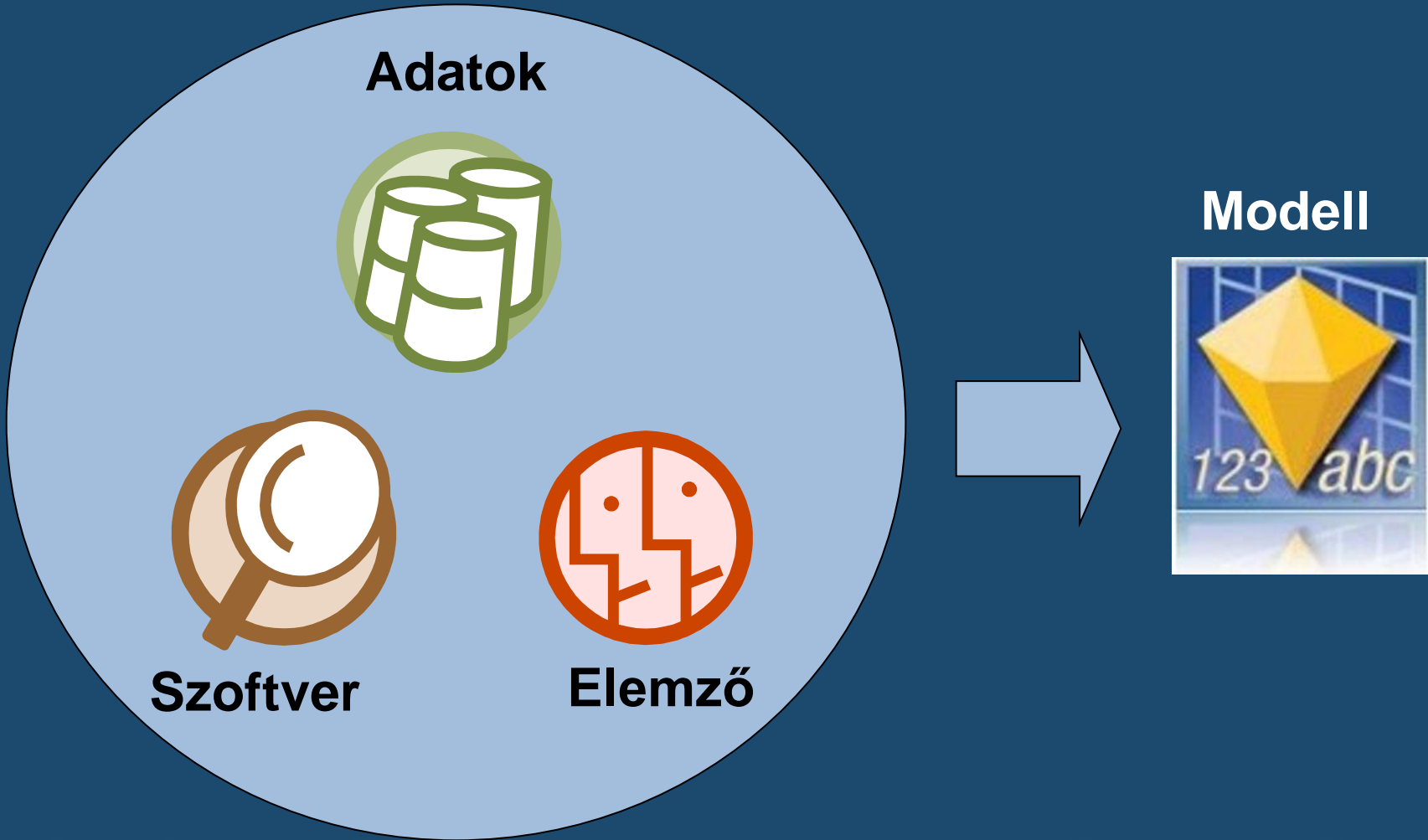


Adatbányászat a felhőben

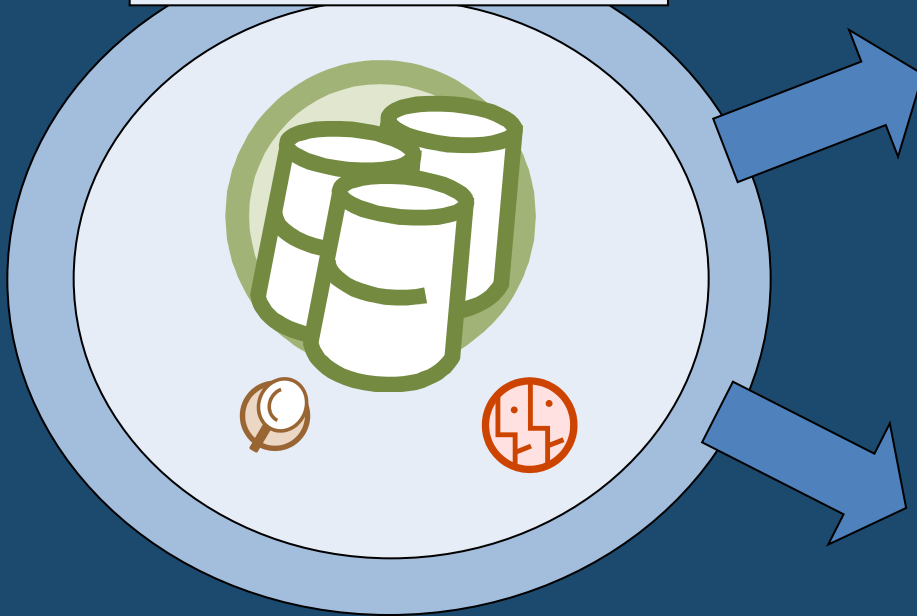
Kovács Gyula – Andego Tanácsadó Kft.

- 2010-ben alapították magánszemélyek (az alapítók több mint egy évtizedes BI tapasztalatokkal rendelkeznek)
- Andego Tanácsadó Kft. fő profilja:
 - *Intelligens applikációk kifejlesztése, és hozzá konzultáció eladása (CARculator, Sixtep hálózati szoftver, KOKAiN – Intelligens céginformációs rendszer)*
 - *Adatbányászati Akadémia (kiscsoportos tanfolyam sorozat)*
- Partnerek
 - *MentorPartner Kft., BI Consulting Kft. és SIXTEP Kft.*
- Referenciák
 - **Pénzügyi:** Lombard Lízing, OTP Bank, Raiffeisen Bank, Uniqa Biztosító, Signal Biztosító, MKB Euroleasing, PBA, Fókusz Takarékn
 - **Nem pénzügyi:** : Vodafone, Sanofi-Avensis, Audatex

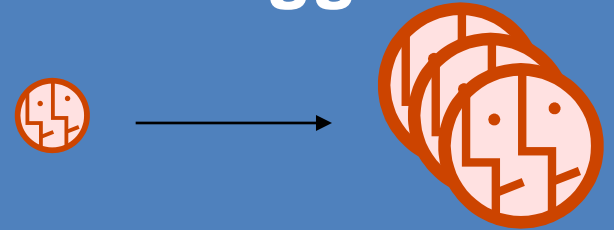


Az adatbányászati infrastruktúra általában nem szimmetrikus → felhő alapú megoldások

Adatdominancia



Kaggle



Dolgozzanak nekünk a legjobb adatbányászok!

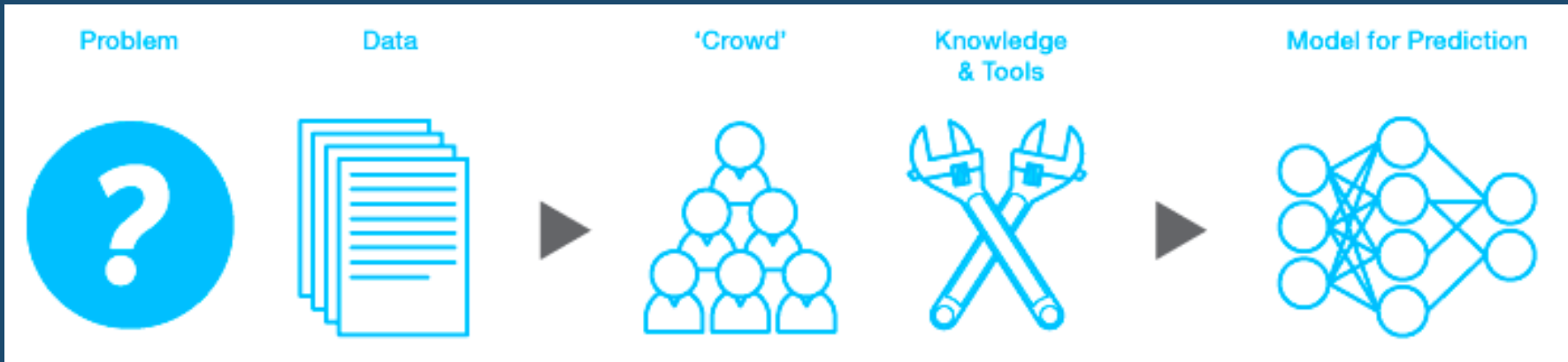
BigML



Adatbányászat a „felhőben”!

kaggle™

- 2010-ben indult az oldal, melynek célja egy adatbányász közösség kialakítása adatbányászati versenyeken keresztül (Melbourne-ből indult, de 2011-ben átköltözött San Francisco-ba a központ)
- 2011-ben 11M\$ kockázati tőkét kapott
- Olyan szervezetejkek dolgoznak együtt, mint NASA, Wikipedia, Delloite, Allstate
- Mintegy 60.000 tudóst, adatbányászt, elemzőt gyűjtött össze az oldal



Analytics

Adatbányászati verseny, ahol a cél minél jobb modellek elkészítése előre meghatározott kritériumok alapján

Recruitment

Gyakorlatilag álláshirdetés versennyel összekötve. A versenyt kiíró cég a legjobbaknak állást ajánl a verseny végén.

Host a competition for...



Analytics

Get the world's best predictive model.



Data Exploration

Find the diamonds in your data.



Recruitment

Uncover objectively brilliant candidates.



Education

Free, powerful classroom competitions.

Data Exploration

Egy cég felkérheti a Kaggle közösséget arra, hogy tegyen ajánlatot milyen elemzések végezhetők el az adatain. Egymás ajánlatait a versenyzők szavazzák meg!

Educations

Az akadémiai oldalnak nyitott versenyek. Számos tudományos verseny probléma megoldásán túl itt megtalálhatók versenyek kezdőknek is.



Predict Closed Questions on Stack Overflow

Tuesday, August 21, 2012

\$20,000 • 167 teams Saturday, November 03

Dashboard

Home/Info

Data

Make a submission

Forum

Leaderboard

Visualization

Competition Details » [Get the Data](#) » [Make a submission](#)

Predict which new questions asked on Stack Overflow will be closed

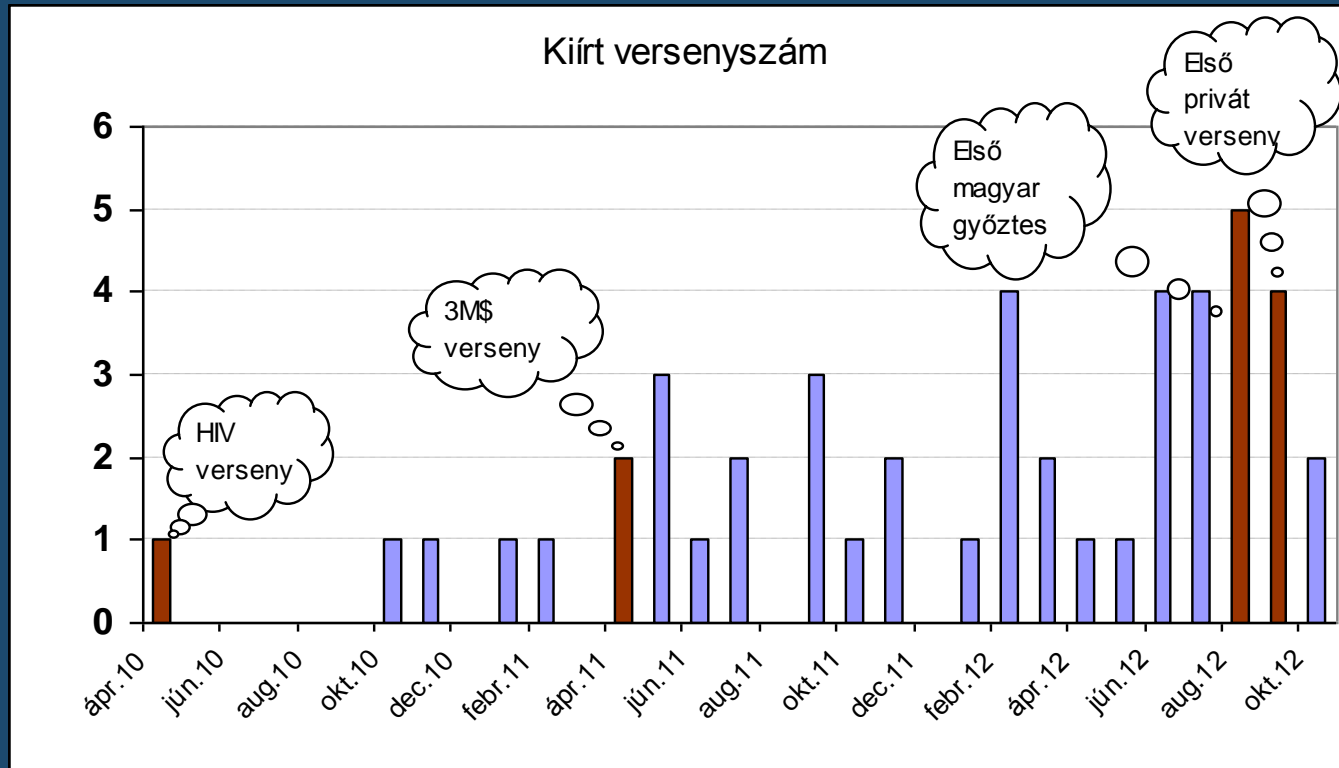
**This competition is now complete.
Congratulations to the preliminary winners!**

Millions of programmers use [Stack Overflow](#) to get high quality answers to their programming questions every day. We take quality very seriously, and have evolved an effective culture of moderation to safe-guard it.

- [Description](#)
- [Evaluation](#)
- [Rules](#)
- [Prizes](#)
- [Jobs](#)
- [Submission Instructions](#)
- [Timeline](#)
- [Visualization Prospect](#)

Az adatbányászok mihez is értenek?





- Eddig 3.5M Ft díj került kiosztásra – nyilvánvalóan a Kaggle üzleti modellje nem jutalék alapú elszámolás
- Kemény feltételek ellenére egyre több verseny

- HIV fertőzés lefolyásának előrejelzése (WHO - ez volt az első meghirdetett projekt)
- Fekete anyag keresése (magánszemélyek, csillagászok)
- Mennyire lehet előrejelezni, hogy egy új zeneszám valakinek tetszeni fog-e? (EMI)
- Egy tetszőleges fájlról mennyire pontosan dönthető el, hogy melyik Open Source projekthez tartozik (Israel Data Science Challenge)
- Twitter használat alapján lelkibetegek detektálása (Online Privacy Foundation)
- Improve Healthcare (2011.04) – Népegészségügyi adatok alapján annak előrejelzése, hogy egy beteg a következő 1 évben hány napot fog eltölteni. A projekt 2 éves, és az összdíjazás 3.000.000\$!

- **Az akadémiai köntöst nem igazán sikerült levetnie** – a projektek jelentős része valamilyen tudományos kutatáshoz köthető (vagy valamilyen konferencia tervezett előadásához kapcsolódik)

- **Komoly probléma az adatvédelem** – a feltett adatok máshol is publikálható adatok. Egy új kategória jelent meg (Private) – ez egy meghívásos verseny, ahol már cégadatok kerülnek átadásra

- **Egyenlőre mérsékelt növekedés jellemző** – nagyon szűk azon intézmények köre, amelyek projekteket indítanak.

bigml

- 2011-ben indult az oldal az amerikai Corvallisban (Oregon), egy innovációs centrumban. Az oldal célja minél több üzleti felhasználó számára elérhető tenni az adatbányászatot „Big Data”-n



Set up a Data Source

Upload, share, or stream your data securely.



Create a Dataset

Quickly format and shape your data source to get useful analytics.



Create a Model

Turn your dataset into an actionable and visual predictive model.



Generate Predictions

Input new data to get immediate predictions from your model.

Bigml – A modellezésért fizetni kell a méret függvényében



The screenshot shows the Bigml website's pricing calculator. At the top, there is a navigation bar with the Bigml logo and links for Home, How It Works, Features, Gallery, Developers, and Blog. A 'Login' button is in the top right corner. The main heading is 'How big is your data?' with three unit selection buttons: MB, GB (selected), and TB. Below this are three sliders for 'DATASETS', 'MODELS', and 'PREDICTIONS'. The DATASETS slider is set to 20GB, MODELS to 10GB, and PREDICTIONS to 26K. A green banner at the bottom displays the total cost: 61,700 × 1¢ = \$617.00.

Category	Value
DATASETS	20GB
MODELS	10GB
PREDICTIONS	26K

61,700 × 1¢ = \$617.00

- **Privát modellek** – kizárólag az láthatja, aki az adatokat feltette és a modelleket elkészítette. Az összes modell alapból privát.
- **White Box modellek** – bárki megveheti a modellt, és ugyanúgy használhatja, mintha sajátja lenne. A modelleket vizsgálhatja és készíthet predikciókat.
- **Black Box modellek** – a modelleket nem tudják mások megnézni, kizárólag predikciókat tudnak elkészíteni.

Modell galéria – üzleti modellek, kevésbé jellemzők az akadémiai modellezések

- AEROSPACE & DEFENSE
- AUTOMOTIVE, ENGINEERING & MANUFACTURING
- BANKING & FINANCE**
- CHEMICAL & PHARMACEUTICAL
- CONSUMER & RETAIL
- DEMOGRAPHICS & SURVEYS
- ENERGY, OIL & GAS
- FRAUD & CRIME
- HEALTHCARE
- HIGHER EDUCATION & SCIENTIFIC RESEARCH
- HUMAN RESOURCES & PSYCHOLOGY
- INSURANCE
- LAW & ORDER
- MEDIA, MARKETING & ADVERTISING
- MISCELLANEOUS
- PHYSICAL, EARTH & LIFE SCIENCES
- PROFESSIONAL SERVICES
- PUBLIC SECTOR & NONPROFIT
- SPORTS & GAMES
- TECHNOLOGY & COMMUNICATIONS
- TRANSPORTATION & LOGISTICS
- TRAVEL & LEISURE
- UTILITIES

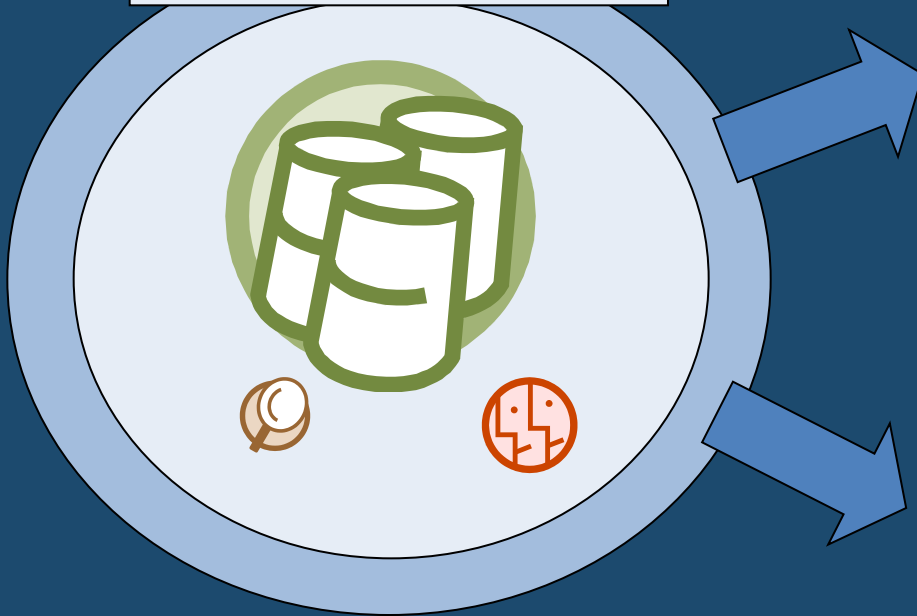
The image displays a grid of 10 decision tree models from the Andego gallery. Each model is represented by a small decision tree diagram at the top, followed by a title, a brief description, and metadata such as file size, number of fields, and number of instances. A red arrow points from the 'BANKING & FINANCE' category in the left sidebar to the first model in the grid.

- Loan Risk**
Model that will predict the quality of risk of a loan application. This dataset was initially...
Fields: France, RiskAnalysis, Loan
134.7 KB, 21 fields / 1000 instances
- APPLE Shares: 1954 - Now**
Apple shares from 1954 to 2012
Close value prediction
Fields: Close, Apple, Shares
292.4 KB, 3 fields / 7000 instances
- LendingClub Loans**
A model predicting loan delinquency rat for loans given by LendingClub.com based on about 50000 loans...
Fields: Credit, Risk
4.3 MB, 19 fields / 40599 instances
- Europe Debt: Debt per capi...**
Debt in Europe by country and years. Debt per capita prediction.
Fields: Debt PerCapita
12.7 KB, 5 fields / 411 instances
- Venture Deals**
Uses some venture deals from crunchbase to predict amount raised
Fields: AmountRaised (\$M)
294.3 KB, 13 fields / 1407 instances
- Europe Debt: Total debt pr...**
Debt in Europe by country and years. Total debt prediction.
Fields: Debt MilionEuro, Europe, Debt
12.7 KB, 5 fields / 411 instances
- Prosper Loans**
Predict loan status based on Prosper loans between November 2005 and October 2012.
Source: Prosper
Fields: Status, Repaid-per-loans, Prosper
4.3 MB, 13 fields / 80949 instances
- Europe Debt: Country pred...**
Debt in Europe by country and years. Country prediction.
Fields: Country, Europe, Debt, Economy
12.7 KB, 5 fields / 411 instances

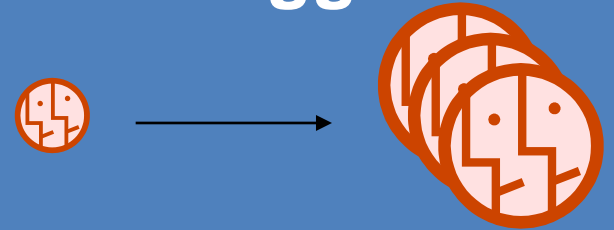
- Egy nagyon új kezdeményezés kimondottan üzleti felhasználók számára
- Az oldal legnagyobb pozitívuma, hogy nincsen adatbázis méret korlát, kimondottan nagy adatokra fejlesztették ki a rendszert („Big Data kompatibilis”).
- Maga a modellezés egyszerű, és maga a felhasználó nem nyúlhat bele a modellezés folyamatába.
- A modellek értékesíthetők – kérdés mennyire lehet univerzális modelleket elkészíteni.

Az adatbányászati infrastruktúra általában nem szimmetrikus → felhő alapú megoldások

Adatdominancia



Kaggle



Dolgozzanak nekünk a legjobb adatbányászok!

BigML



Adatbányászat a „felhőben”!