# Adatbányászat a felhőben

Kovács Gyula – Andego Tanácsadó Kft.

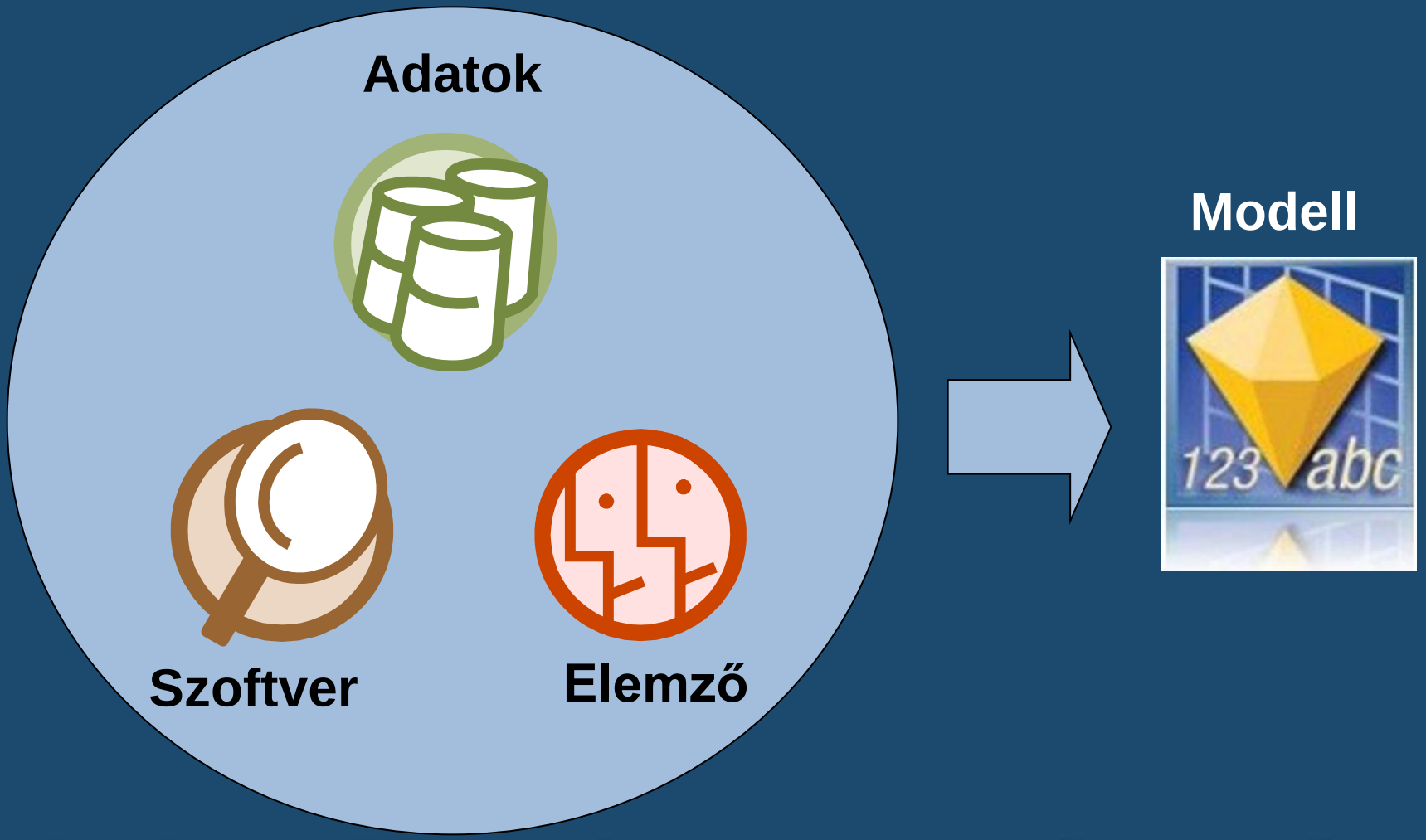
**2012 – II. Innovatív BI Konferencia**

# Magunkról

- 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 Takarék
  - **Nem pénzügy**: : Vodafone, Sanofi-Avensis, Audatex

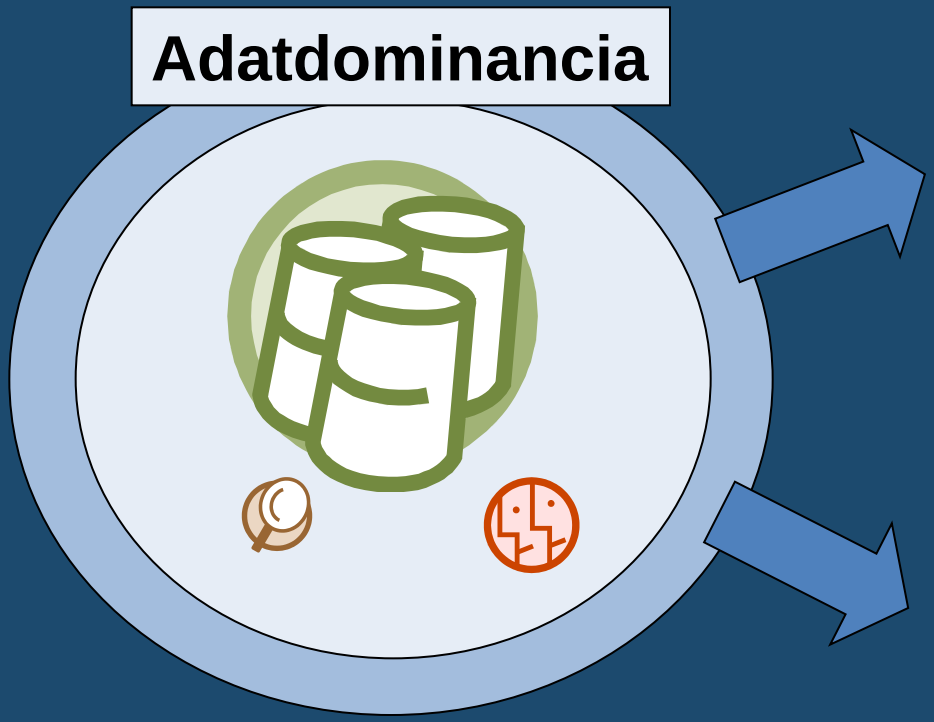
# Adatbányászat – Trinity

andego

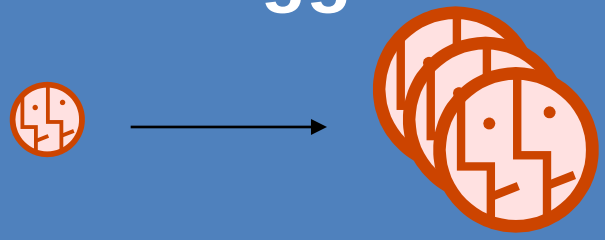


# Az adatbányászati infrastruktúra általában nem szimmetrikus –> felhő alapú megoldások

andego




## Kaggle



**Dolgozzanak nekünk a legjobb adatbányászok!**

## BigML

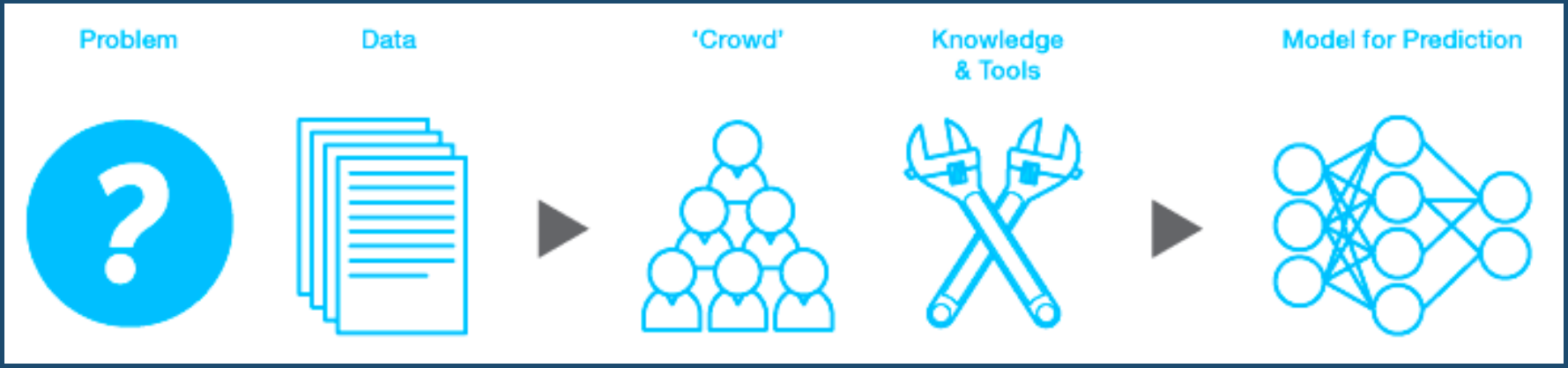

**Adatbányászat a „felhőben”!**

kaggle™

# Kaggle - We're making data science a sport.™

andego

- 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 szervezetejjek dolgoznak együtt, mint NASA, Wikipedia, Delloite, Allstate
- Mintegy 60.000 tudóst, adatbányászt, elemzőt gyűjtött össze az oldal



# Kaggle - We’re making data science a sport.™

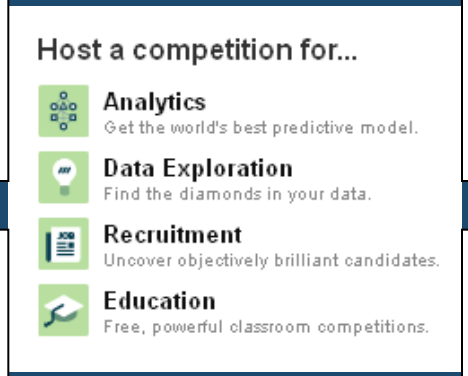
andego

## Analytics

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

## Data Exploration

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




## Recruitment

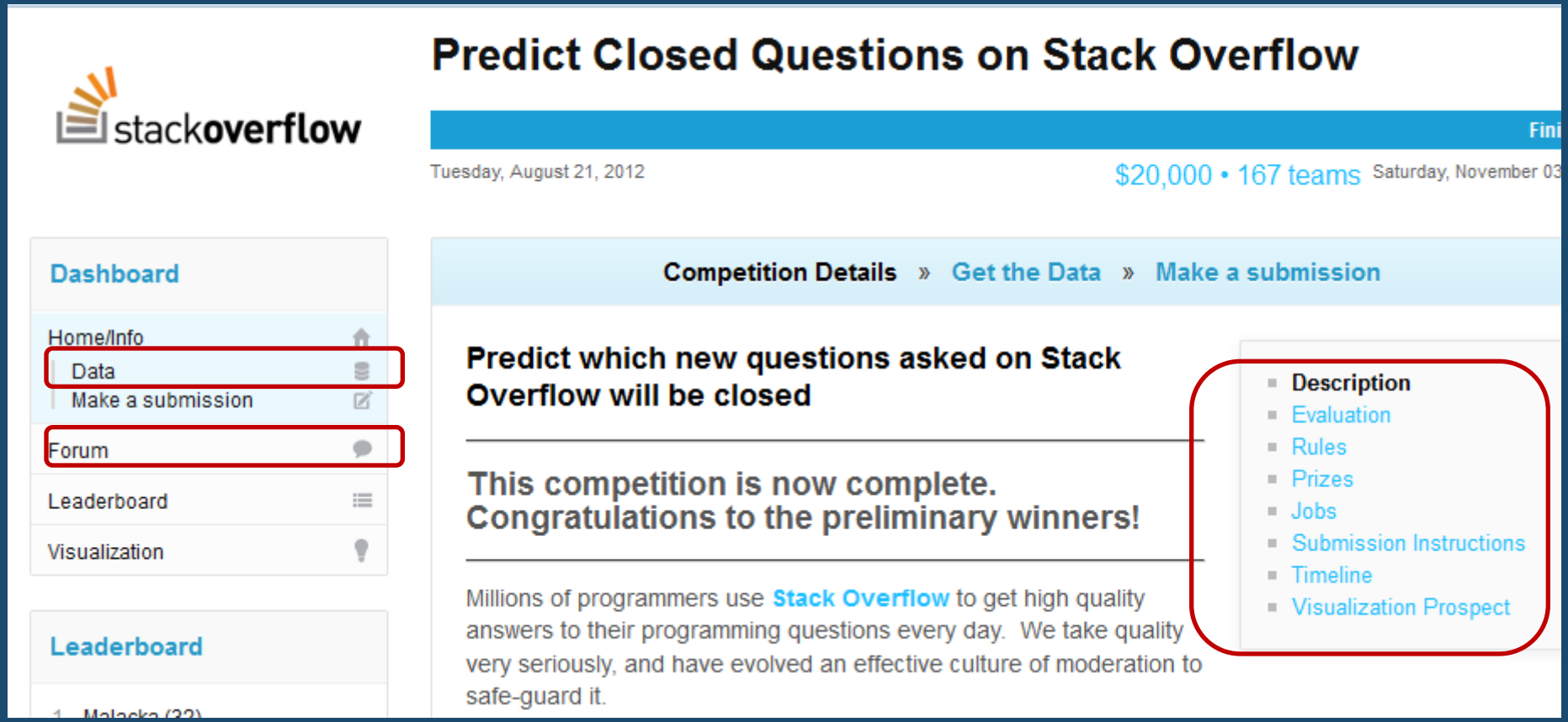
Gyakorlatilag álláshirdetés versennyel összekötve. A versenyt kiíró cég a legjobbaknak állást ajánl a verseny végén.

## Educations

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

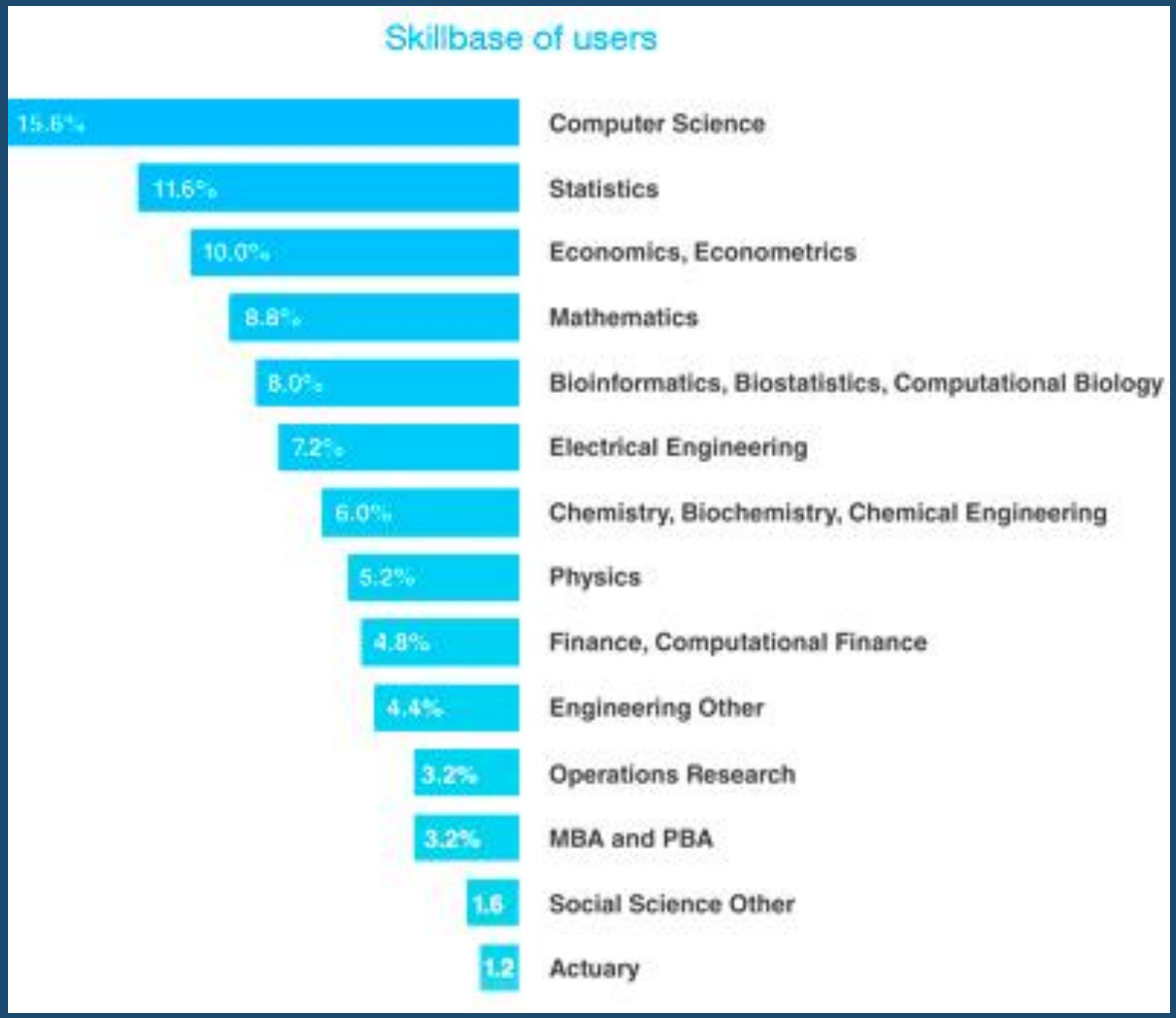
# Kaggle – Hogyan néz ki egy projekt?

andego

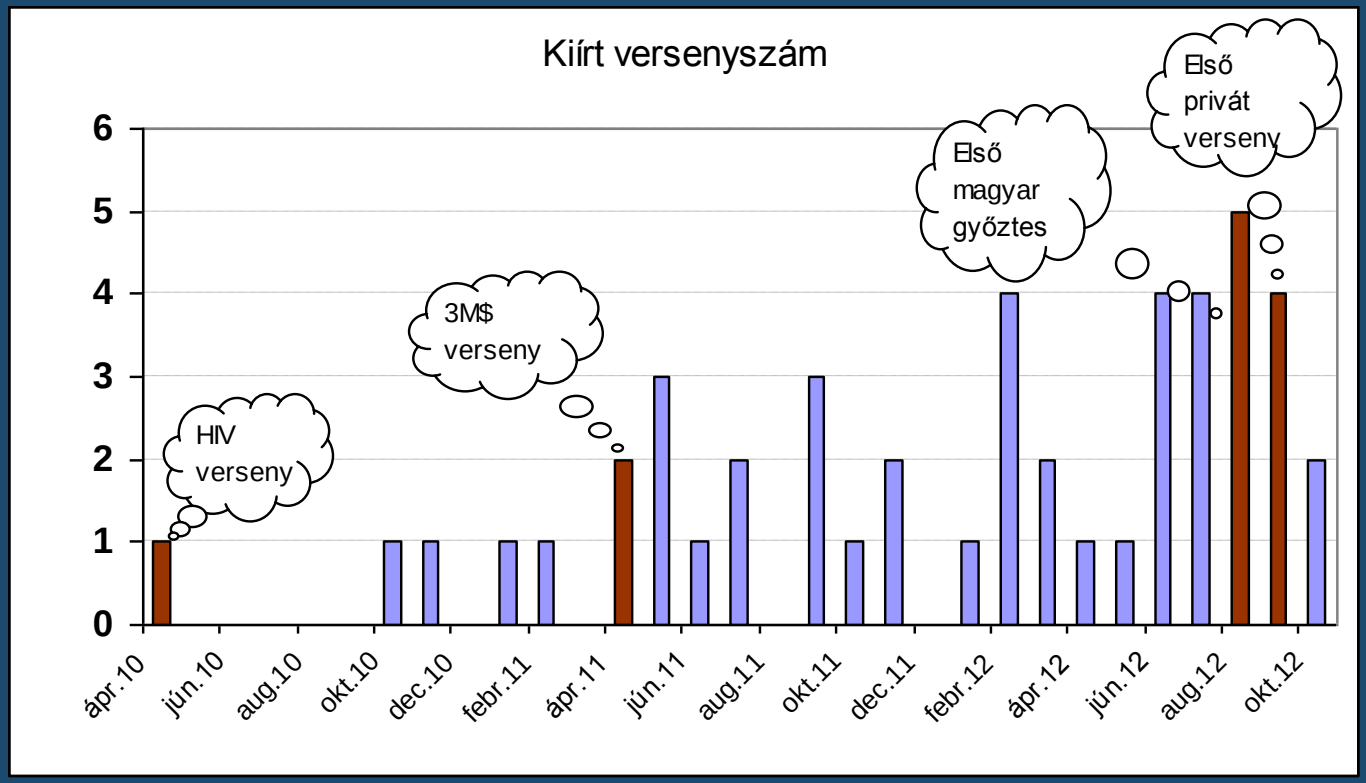


# Az adatbányászok mihez is értenek?

andego



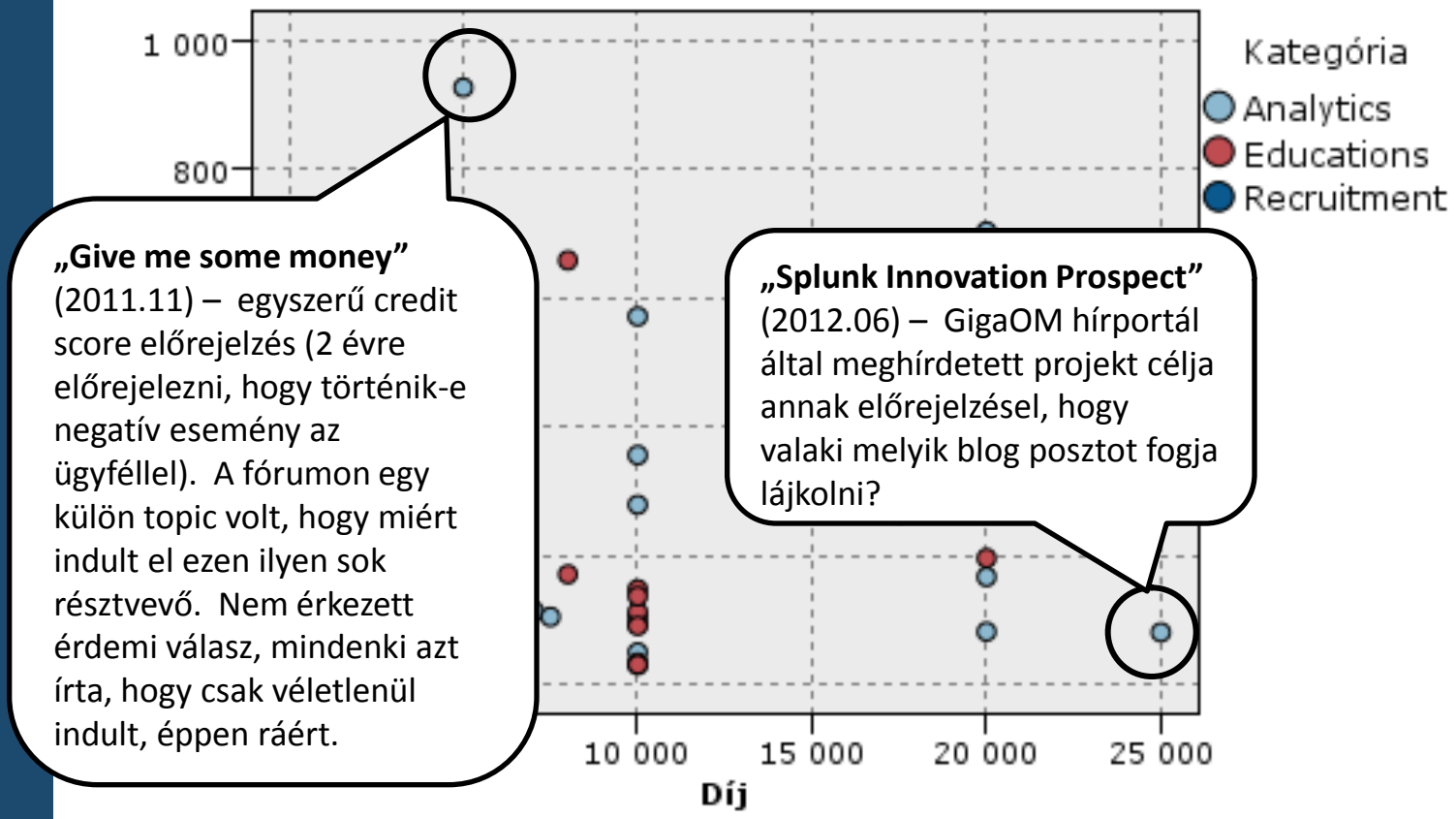
# Mennyire népszerű az oldal?




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

# Mennyire népszerű az oldal?

andego




**„Give me some money”** (2011.11) – egyszerű credit score előrejelzés (2 évre előrejelezni, hogy történik-e negatív esemény az ügyféllel). A fórumon egy külön topic volt, hogy miért indult el ezen ilyen sok résztvevő. Nem érkezett érdemi válasz, mindenki azt írta, hogy csak véletlenül indult, éppen ráért.

**„Splunk Innovation Prospect”** (2012.06) – GigaOM hírportál által meghírdetett projekt célja annak előrejelzésel, hogy valaki melyik blog posztot fogja lájkolni?

# Néhány érdekesebb projekt

andego

- 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 )
- Imporve Healthcare (2011.04) – Népegészségügyi adatok alapján annak előrejelzése, hoyg egy betek a következő 1 évben hány napot fog eltölteni. A projekt 2 éves, és az összdíjazás 3.000.000$!

# Kaggle – hol tart?

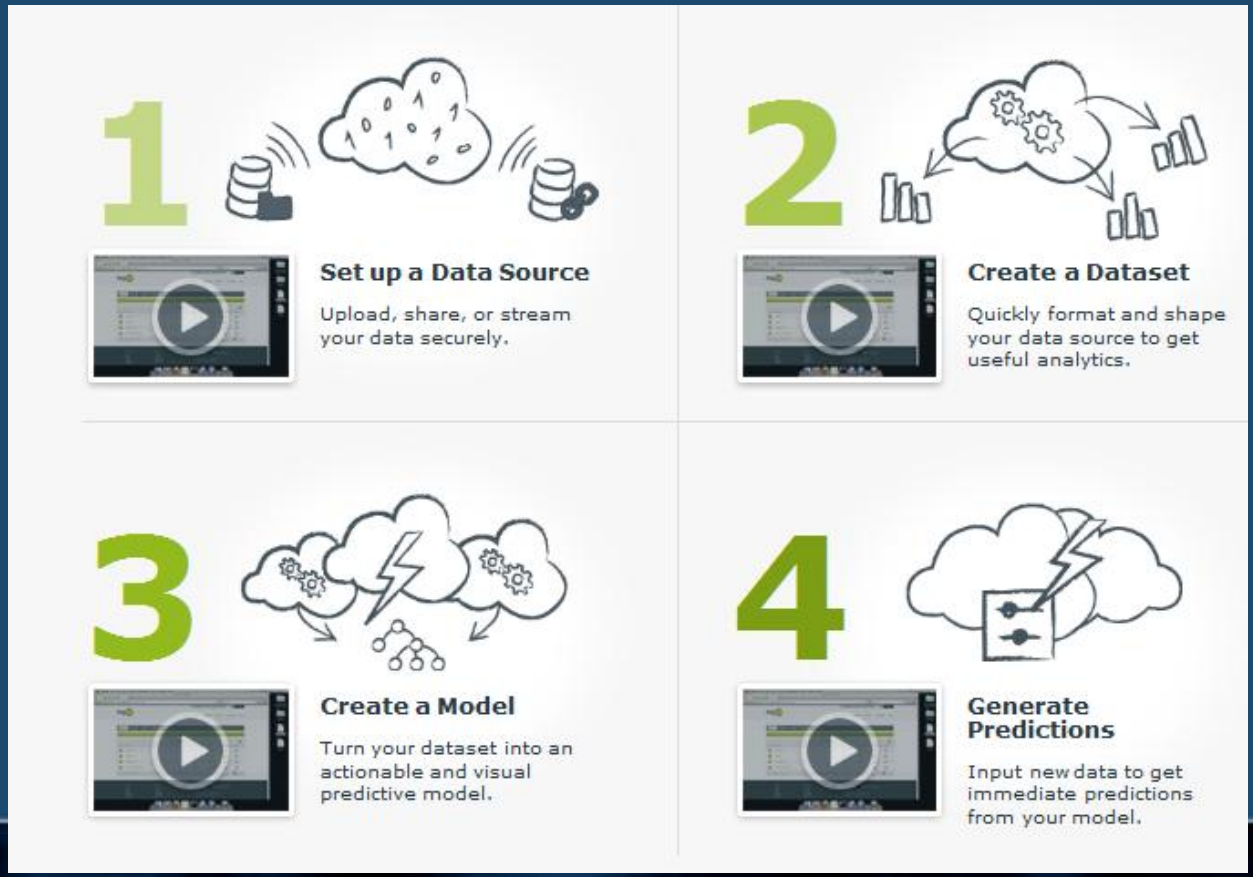
andego

- **Az akadémiai köntöst nem igazán sikerült levetnie** – a projektek jelentős része valamilyen tudományos kutatáshoz köthetők (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

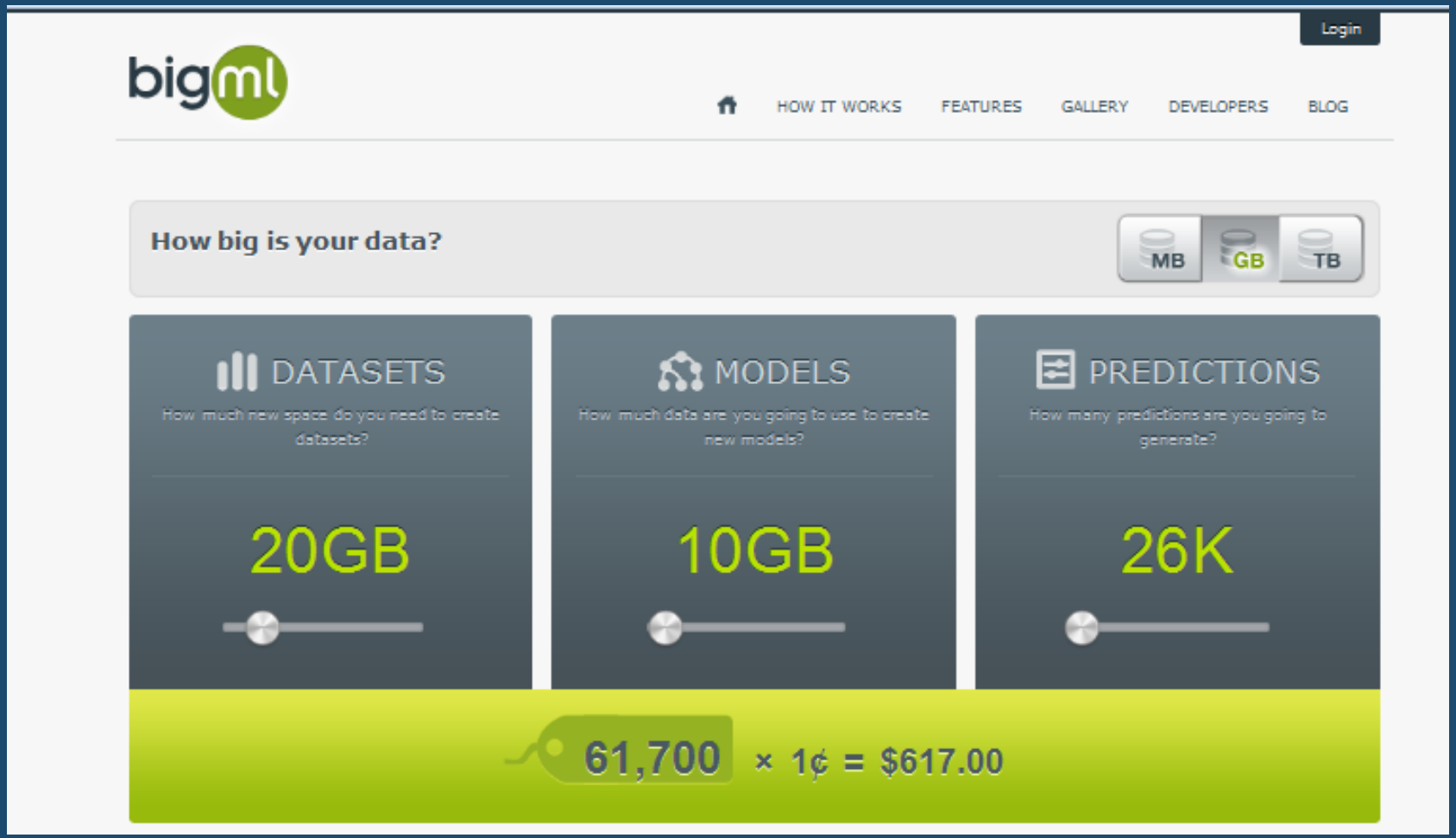
# Bigml – Machine Learning for Everyone

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



# Bigml – A modellezésért fizetni kell a méret függvényében

andego



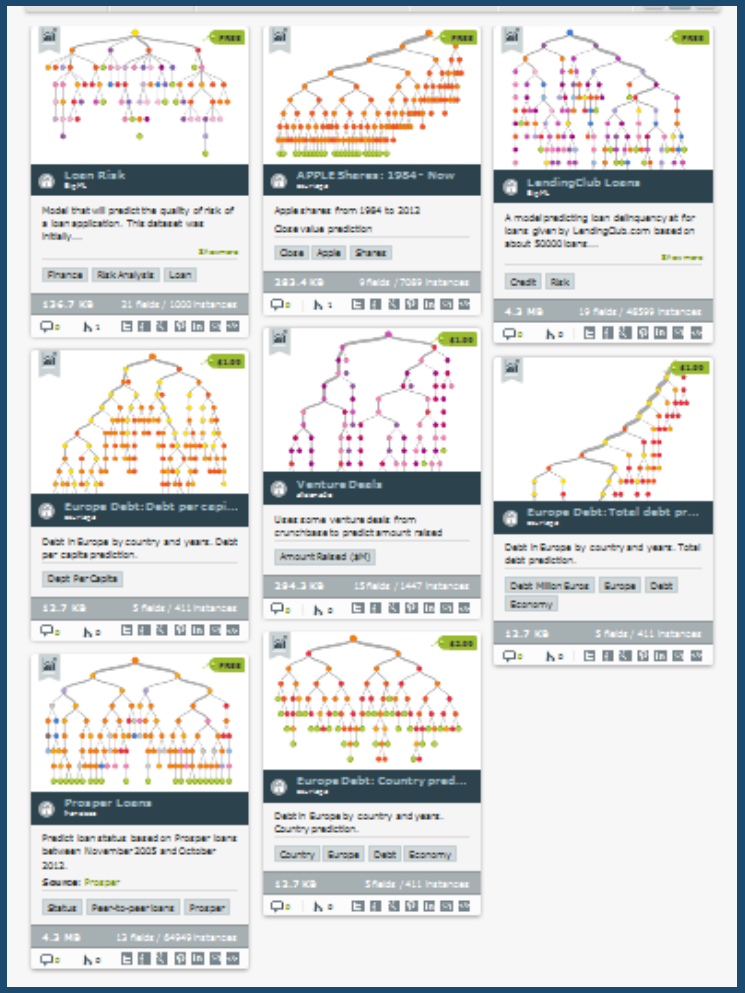
# Modellek elérhetősége

- **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

andego

- 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

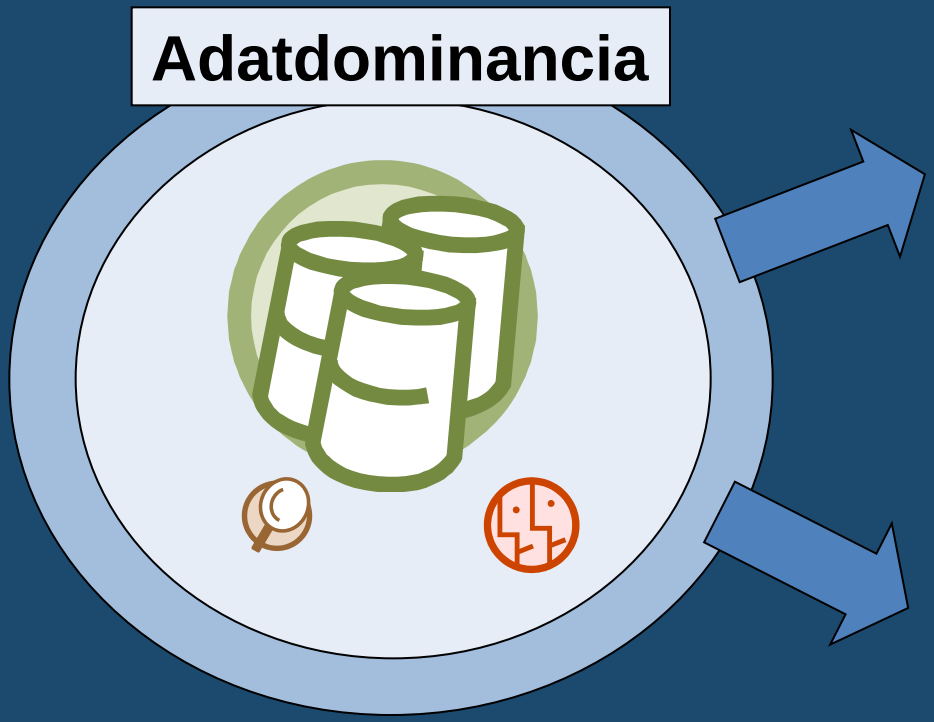


# BigML– hol tart?

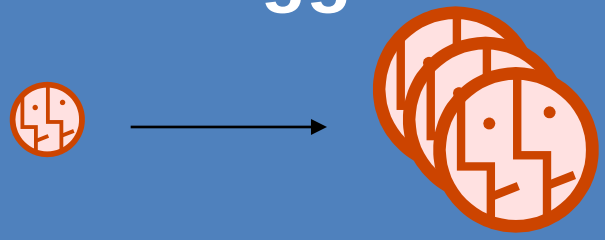
- Egy nagyon új kezdeményezés kimondottan üzelti felhasználók számára
- Az oldal legnagyobb pozitívuma, hogy nincsen adatbázis méret korlát, kimondottan nagy adatokra fejlesztették ki a rendszert („Big Data kompatibilis”).
- Maga a modellezés egyszerű, és maga a felhasználó nem nyúlhat bele a modellezés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

# Az adatbányászati infrastruktúra általában nem szimmetrikus –> felhő alapú megoldások

andego




## Kaggle



**Dolgozzanak nekünk a legjobb adatbányászok!**

## BigML


**Adatbányászat a „felhőben”!**