

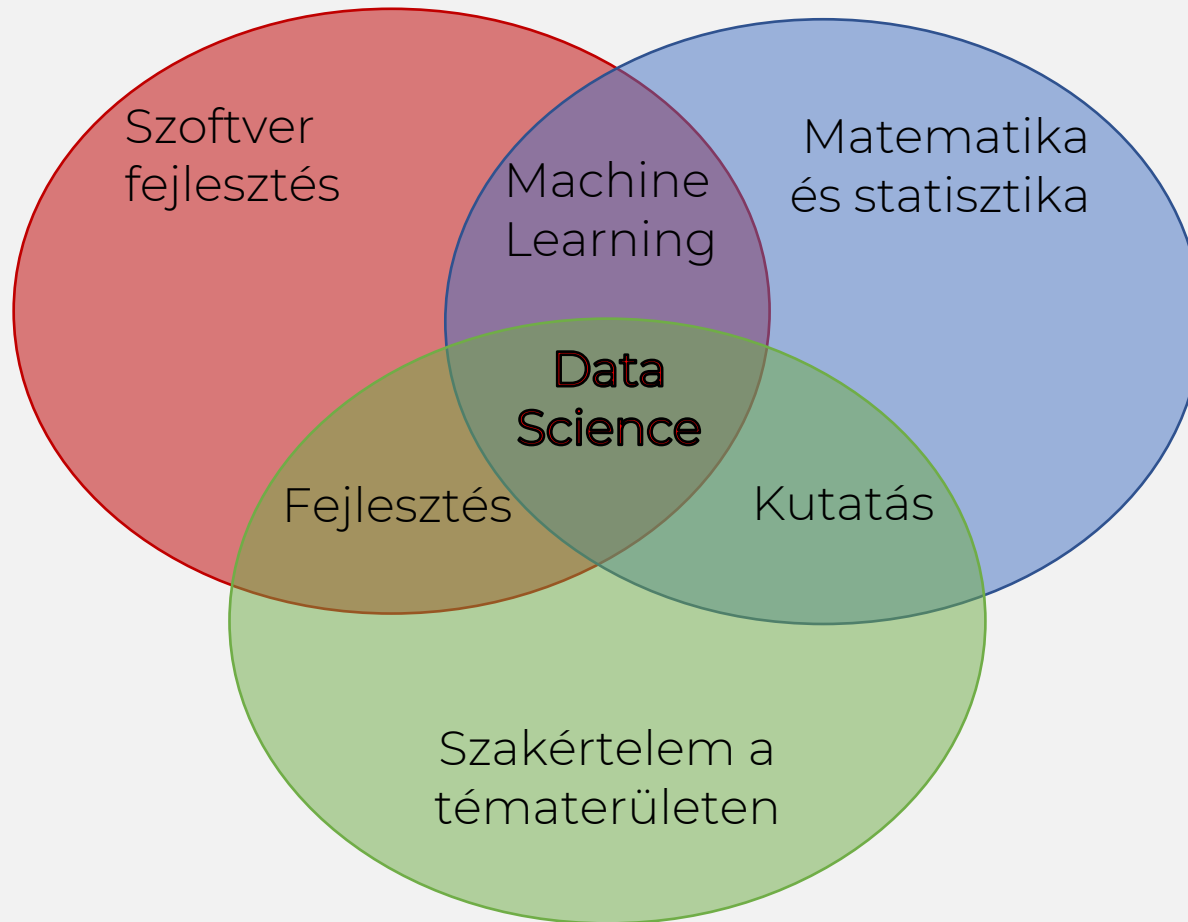
BEVEZETÉS

Lukovszki Csaba

ADATELEMZÉS FOGALMA, CÉLJA

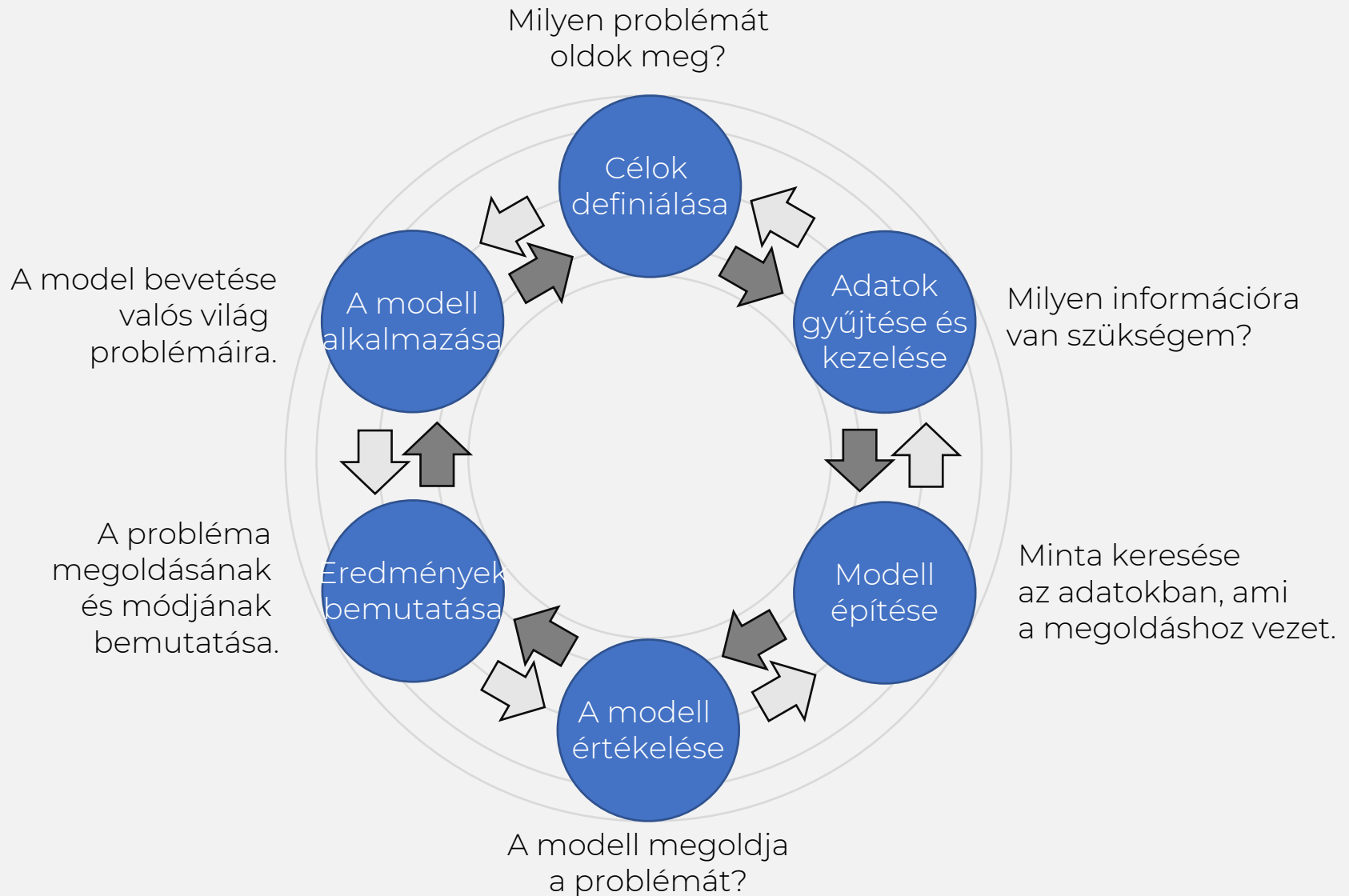
- Rengeteg adat érhető el
- Az elérhető adatok megismerését, elemzését hívjuk adatelemzésnek
- Célja:
 - Hipotézisek alátámasztása, vagy elvetése
 - Döntéstámogatás
 - Jövőbeni események becslése, jóslása
 - Általánosságban az adatok megértése, strukturált keresése
- Az adatok elemzésére – az adatok mennyisége miatt – **adatelemző szoftver**eket használunk

TÉMATERÜLETEK (DATA SCIENCE)



... és persze egy kis művészet!

ADATELEMZÉS FOLYAMATA



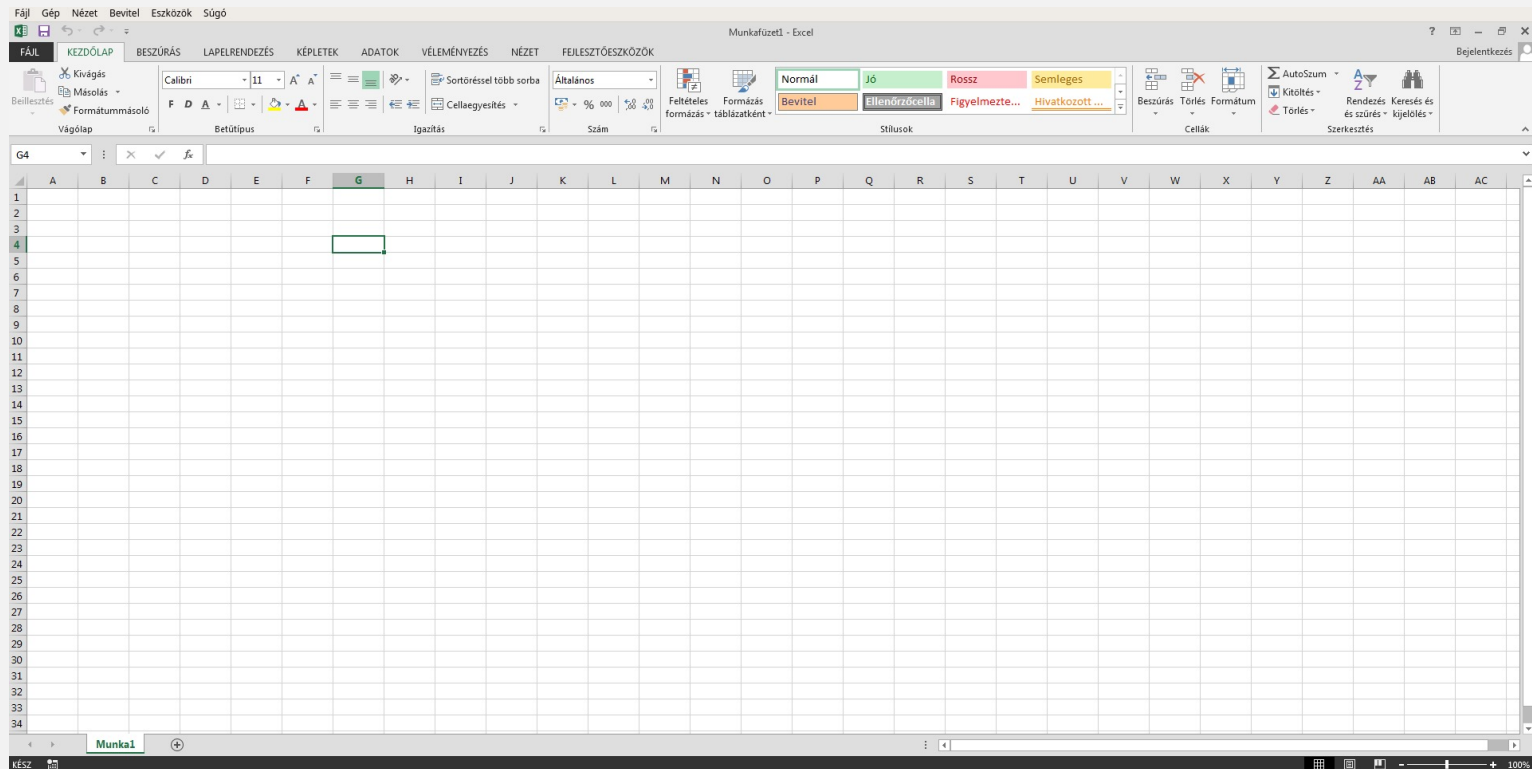
ADATELEMZŐ SZOFTVEREK

- Többféle szoftvert fejlesztettek adatelemzési célokra
- Felhasználói felület szerint
 - Grafikus felületű (Point'n'Click)
 - Programozható
- Több esetben a szoftverek mind a két módon használhatók
- Ábrák, diagramok generálásához elvárt a grafikus felület
- A tárgy keretein belül egy tipikus adatelemző szoftveren keresztül mutatjuk be az adatelemző szoftverek működését

ADATELEMZŐ SZOFTVEREK– FUNKCIÓK

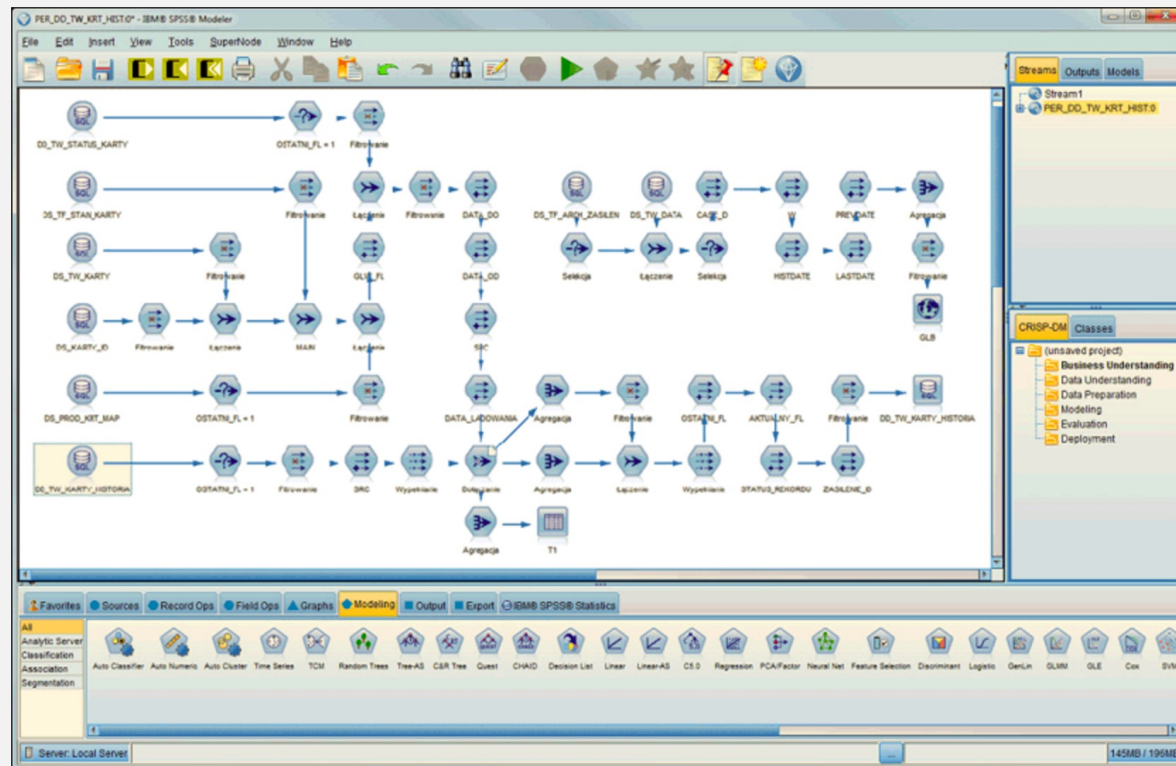
- Adatbázisok olvasása és írása
 - Adatfájlok (pl. csv, xml, yaml, json, txt)
 - Relációs adatbázisok (pl. Oracle RDBMS, MS SQL Server, ODBS)
 - NoSQL adatbázisok (pl. MongoDB)
 - Eloszott Big Data adatbázisok (pl. Hadoop)
- Adatok szűrése, tisztítása, ellenőrzése
- Adatelemzési algoritmusok, modell alkotás
 - Alapvető adat statisztikák (átlag, szórás, kvantilisek, stb...)
 - Összetett adatelemzési algoritmusok (osztályozás, regresszió, stb...)
- Grafikai megjelenítés
 - Adatok és elemzési algoritmusok eredményének ábrázolása
- Adatelemzési jelentések készítése

MICROSOFT EXCEL



- Jól ismert táblázatszerkesztő
- Számos statisztikai és adatelemzési funkcióval
- Alapvető adatelemzési feladatokra alkalmas

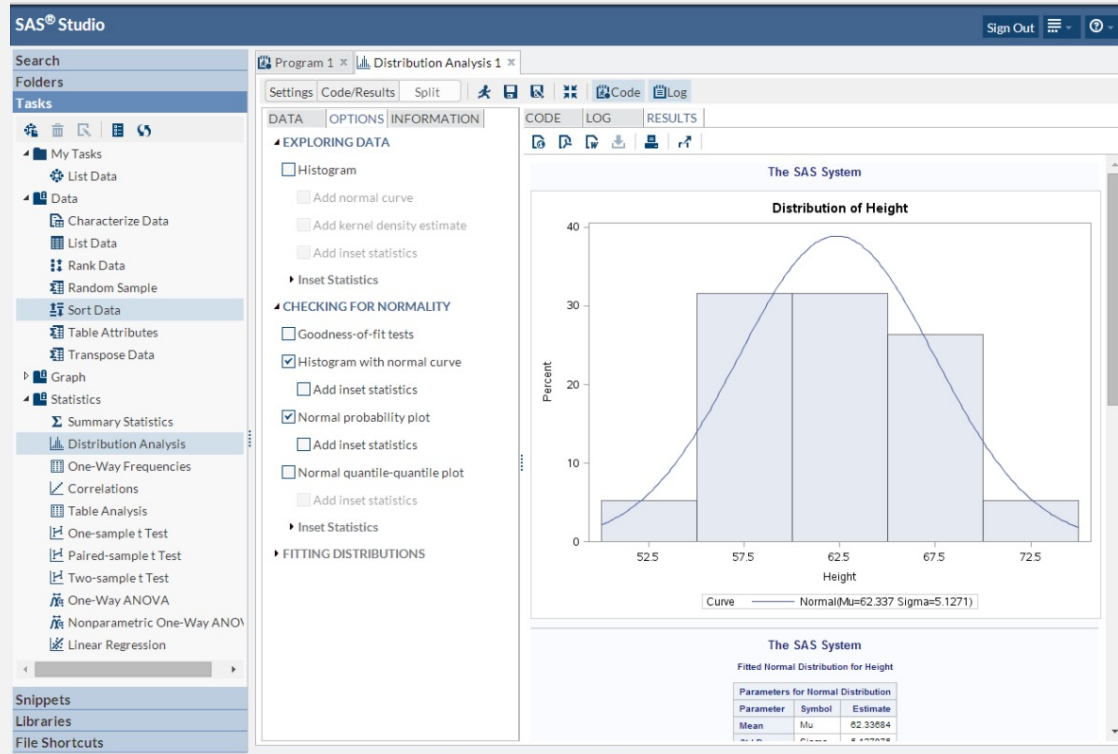
IBM SPSS STATISTICS, SPSS MODELER



- Nagy múltú statisztikai programcsomag
- Elsősorban grafikus felületű
- Command Syntax szkriptnyelven is programozható
- Számos statisztikai és adatelemzési algoritmust ismer

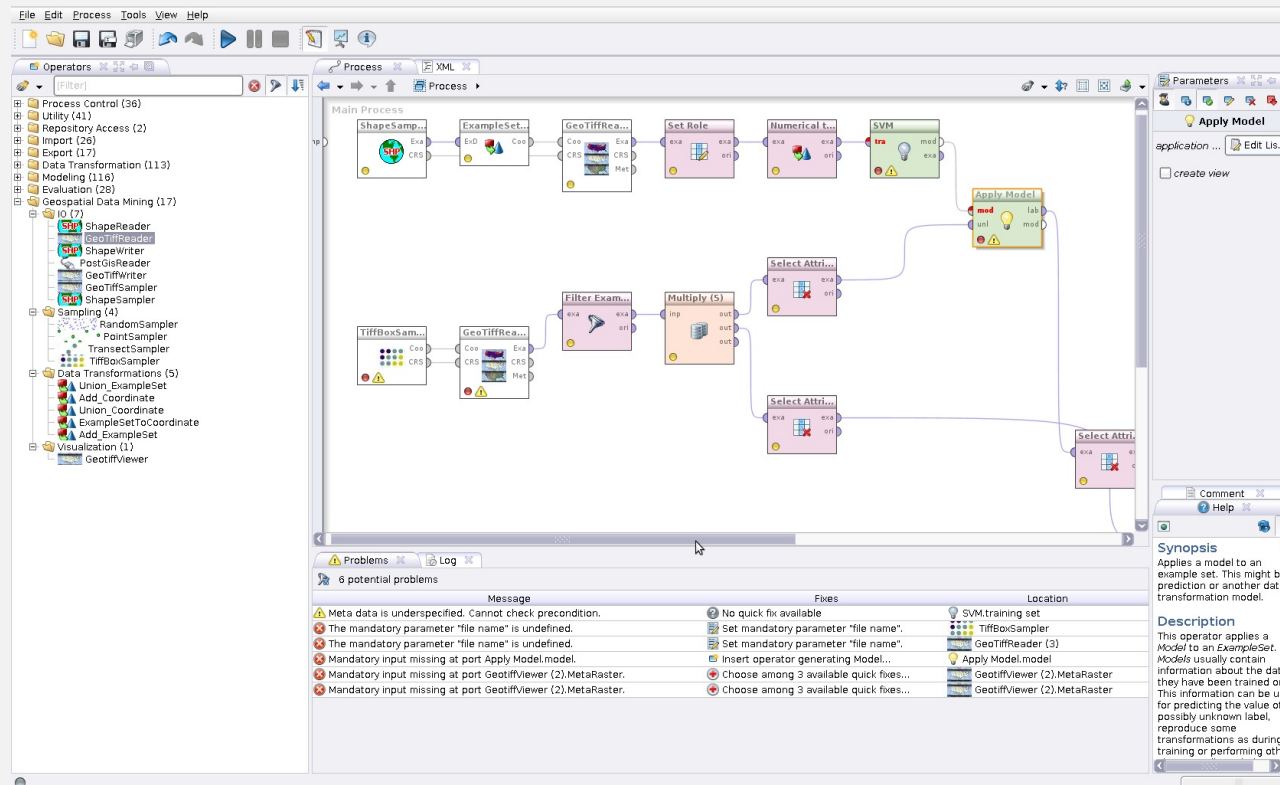
SAS STUDIO

Statistical Analysis System



- Egyik legrégebbi és leghíresebb adatelemzési program (1972)
- Point'n'Click, de a SAS nyelv segítségével programozható

RAPID MINER



- A legszélesebb körben alkalmazott adatbányászati szoftver
- Szinte minden megvalósítható benne programkód fejlesztése nélkül is
- Számos plugin használható és kompatibilis az R és Python szkriptekkel

EGYÉB MEGOLDÁSOK

- Python – NumPy, Pandas, Matplotlib, SciPy
 - A Python nyelvhez fejlesztett, tudományos számításokat támogató könyvtárak
 - Statisztikai számítások, megjelenítés, adatkezelés
- Weka
 - Grafikus felületű, gépi tanulási feladatokra használható szoftver
- Octave, Matlab
 - Mátrix alapú számításokra fejlesztett, de azóta számos egyéb témakörben használható alkalmazások
- R nyelv

AZ R NYELV

R, MINT ADATELEMZŐ NYELV

- Az R egy szabad forráskódú programozási nyelv
 - Statisztikai és adatbányászati alkalmazásokra terveztek
- Az R az 1976-ban született S nyelv egyik implementációja
 - Az S nyelv utolsó szabványa az 1998-es S4 szabvány
- Az első stabil változata 2000-ben jelent meg
- Tipikus adatelemzési feladatok, amire az R alkalmas...
 - Adatelőkészítés, adatszűrés
 - Gépi tanulás, osztályozás, klaszterezés
 - Regresszió, predikció, idősorok elemzése
 - Anomália detekció
 - Adatvizualizáció
- ... és amire kevésbé
 - Extrém méretű adatok kezelése (bár Hadoop csomagok is léteznek)

A FELHASZNÁLÓI FELÜLET

- Felhasználói felületek
 - [Rstudio](#)
 - RCommander
 - RKWard
 - Rattle
- A felületek támogatják a hatékony fejlesztést
 - Szkriptek kezelése
 - Automatikus kiegészítés
 - Stb...
- Az R nyelv funkciói csomagokkal (package) bővíthetők
 - Végtelen számú csomag érhető el (~15360, 2020. január)
 - Algoritmus implementációk
 - Adatkezelési könyvtárak (json, xml, xls, mongodb, stb...)
 - Grafikai könyvtárak
 - Stb...

R MARKDOWN, R NOTEBOOK

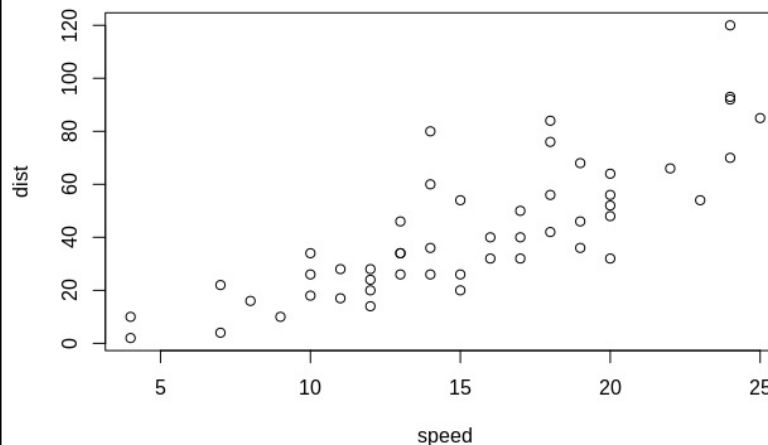
```
1 ---
2 title: "R Notebook"
3 output: html_notebook
4 ---
5
6 This is an [R Markdown](http://rmarkdown.rstudio.com)
7 notebook, the results appear beneath the code.
8 Try executing this chunk by clicking the Run button or by
9 placing your cursor inside it and pressing Ctrl+Shift+Enter.
10
11 ```{r}
12 plot(cars)
13 ```
14 Add a new chunk by clicking the Insert Chunk button or by
15 pressing Ctrl+Alt+I.
16 When you save the notebook, an HTML file containing the code
17 and output will be saved alongside it (click the Preview
18 button or press Ctrl+Shift+K to preview the HTML file).
19 The preview shows you a rendered HTML copy of the contents of
20 the editor. Consequently, unlike Knitr, Preview does not
21 run any R code chunks. Instead, the output of the chunk when
22 it was last run in the editor is displayed.
```

R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

plot(cars)



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knitr*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

R BEMUTATÁSA

- Az R interpretált nyelv
- Az R nyelvben minden, még az egyes nyelvi elemek is objektumok
- Az egyes objektumoknak típusa van, többfajta is lehet
 - mode: Az S nyelv szabályai szerinti típus
 - typeof: Az R nyelvi típus
 - storage: A tárolás szerinti típus
 - class: Az értelmezés szerinti típus
- A gyakorlatban legtöbbször a mode, typeof és a storage megegyezik
- Számos alaptípus, ezekből a három legfontosabb
 - vektorok
 - listák
 - függvények

OBJEKTUMOK

	class	mode	typeof	storage
TRUE	logical	logical	logical	logical
c(TRUE, TRUE)	logical	logical	logical	logical
5L	integer	numeric	integer	integer
c(1L,1L,1L,1L)	integer	numeric	integer	integer
5	numeric	numeric	double	double
c(1,2,3,4)	numeric	numeric	double	double
"test"	character	character	character	character
c("one", "two", "three")	character	character	character	character
raw()	raw	raw	raw	raw
factor()	factor	factor	factor	factor
matrix(logical())	matrix	logical	logical	logical
matrix(numeric())	matrix	numeric	numeric	numeric
array(numeric(8), c(2,2,2))	array	numeric	double	double
list()	list	list	list	list
data.frame()	data.frame	list	list	list
cos()	function	function	function	function

R KÖRNYEZET

- R használata
 - RStudio
 - Desktop
 - Telepítése: Free Desktop RStudio www.rstudio.com
 - Szerver web hozzáférés
<http://batman.tmit.bme.hu/rstudio>
 - Belépés
 - `st_<neptun kód>/<neptun kód>`

HA NEM ÉRTJÜK, KÉRDEZZÜNK

