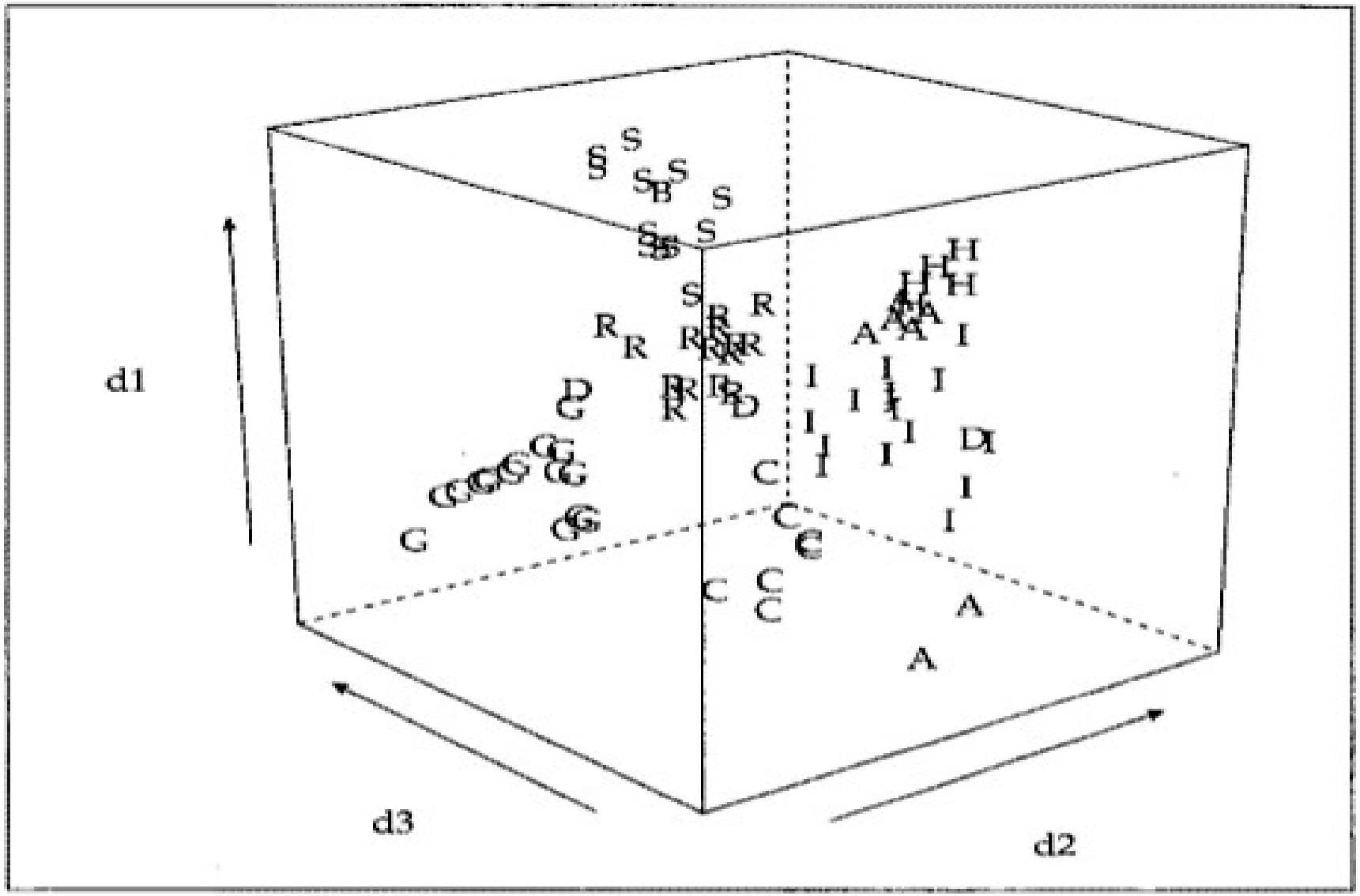
Adatbányászat az R nyelv alkalmazásával

Varjú Zoltán

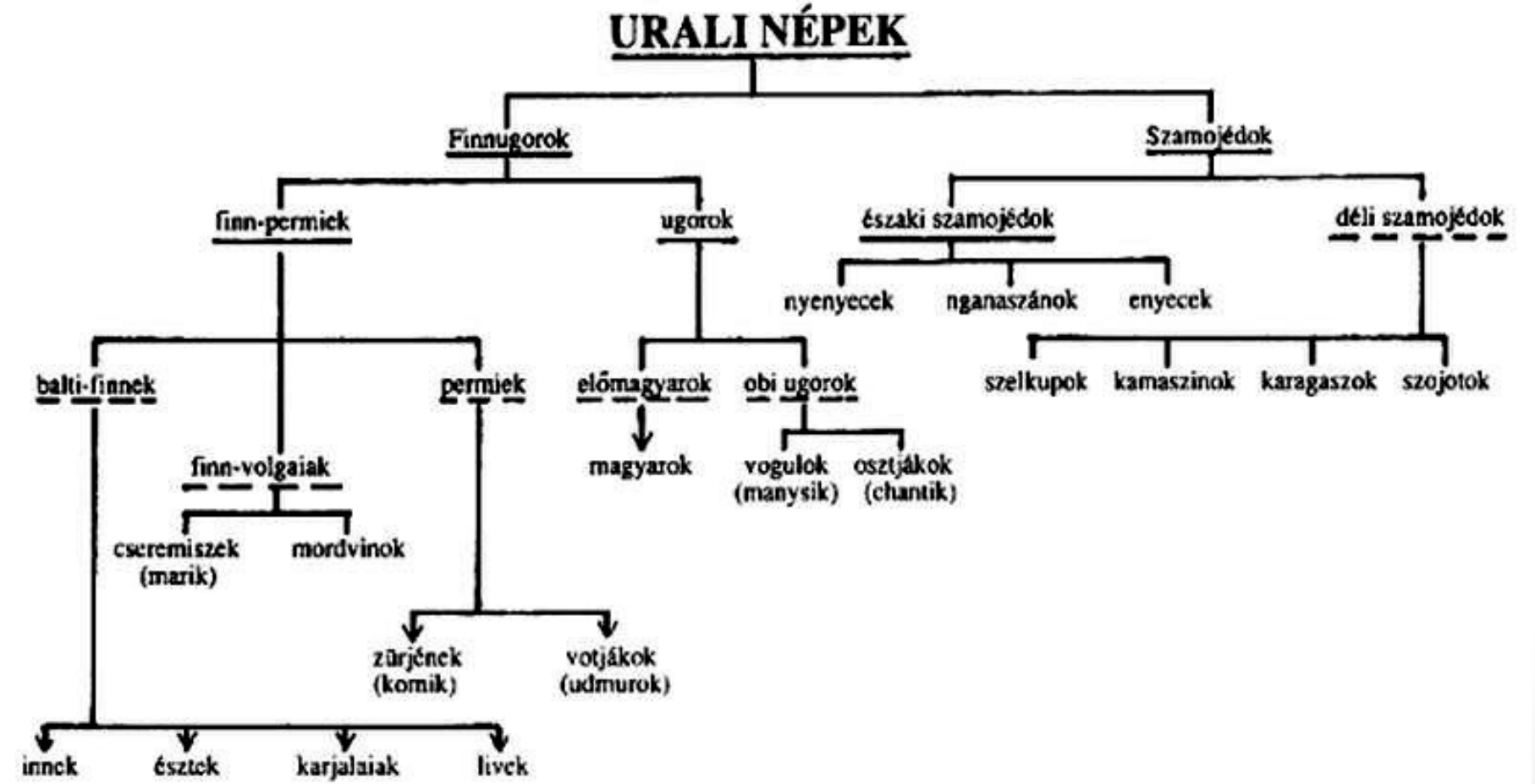
Open Source BI Fórum, 2012. okt. 3

Egy személyes példa

- Mennyire hasonlítanak egymásra a finnugor nyelvek?
- Milyen más nyelvekre hasonlítanak?
- Hogyan mérhetjük a hasonlóságot?
- Hogyan mutassuk meg a hasonlósági metrikát olyan közönségnek mely tagjai alapvetően nem a számok nyelvén beszélnek?
- Saját gyűjtésekre és online elérhető szövegekre alapozva a karaktersorok közötti hasonlóságot mértük


d1
d3
d2

A FINNUGOR NÉPEK CSALÁDFÁJA



- Az R-ben mindenre van megoldás
- Ha hozzászokik az ember a dokumentáció stílusához, akkor képes igényeire szabni a példa szkripteket
- Learning by doing
- A nyelvtudásnak több szintje van a programozási nyelvek esetében is

Adatbányászat az R nyelv alkalmazásával

- Rövid áttekintés
- R a gyakorlatban
- R és Big Data
- Jövő

Kis R történelem és háttér

- Az R nyelv az S statisztikai programozási nyelv “dialektusa”

- 1976 Bell Labs, John Chambers és tsai Fortranban implementálták az S nyelvet

- 1988 C-ben újraírták az S-t

- 1993-ban Insightful Corp. megkapja az S licencét, ma TIBCO birtokolja S-PLUS néven

Kis R történelem és háttér

- 1991 Ross Ihaka és Robert Gentleman elindítja az R fejlesztését
- 1995 GPL licenc alatt jelenik meg a nyelv
- 2000 R 1.0.0
- 2012.06.22 R 2.15.1 “Roasted Marshmallows”

Kis R történelem és háttér

- CRAN

  http://cran.r-project.org/

- 5300 csomag

- CRAN Task Views

- ggplot2
  http://ggplot2.org/

- Wilkson et al. ***The Grammar of Graphics***

Miért R?

- CRAN 5300 csomag
- A statisztikai lingua franca
- A legjobban dokumentált nyelv a területen
- Aktív fejlesztői közösség
- Aktív felhasználói közösség
- Trendi

R a gyakorlatban

"[W]e wanted **users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming.** Then as their needs became clearer and their sophistication increased, **they should be able to slide gradually into programming,** when the language and system aspects would become more important." Chambers

(http://www.stat.bell-labs.com/S/history.html)

R a gyakorlatban

- R, mint adatfelfedezési eszköz (EDA)

- R, mint "number cruncher"

R a gyakorlatban - workflow

- Elméletileg az R “képességei” lefedik az egész elemzési folyamatot
- Gyakorlatban az előfeldolgozás általában más nyelven történik (pl. Python, Perl)
- “interaktív” szkriptek a repl segítségével
- A szkriptek “refaktorálása” riportokhoz
- prototyping

R a gyakorlatban

- Szoftverfejlesztés vs. algoritmikus elemzés

- Kelly – Mill: *A Software Chasm*

- Az R a legelterjedtebb nyelv kutatói és ipari körökben is

- Prabhu – Jablin: *A survey of practice of computational science*

R és Big Data

- In-memory eszköz

- a rendelkezésre álló RAM 10-20 százalékát meghaladó adatmennyiség már lassuláshoz vezet, 50% felett pedig tkp. használhatatlan

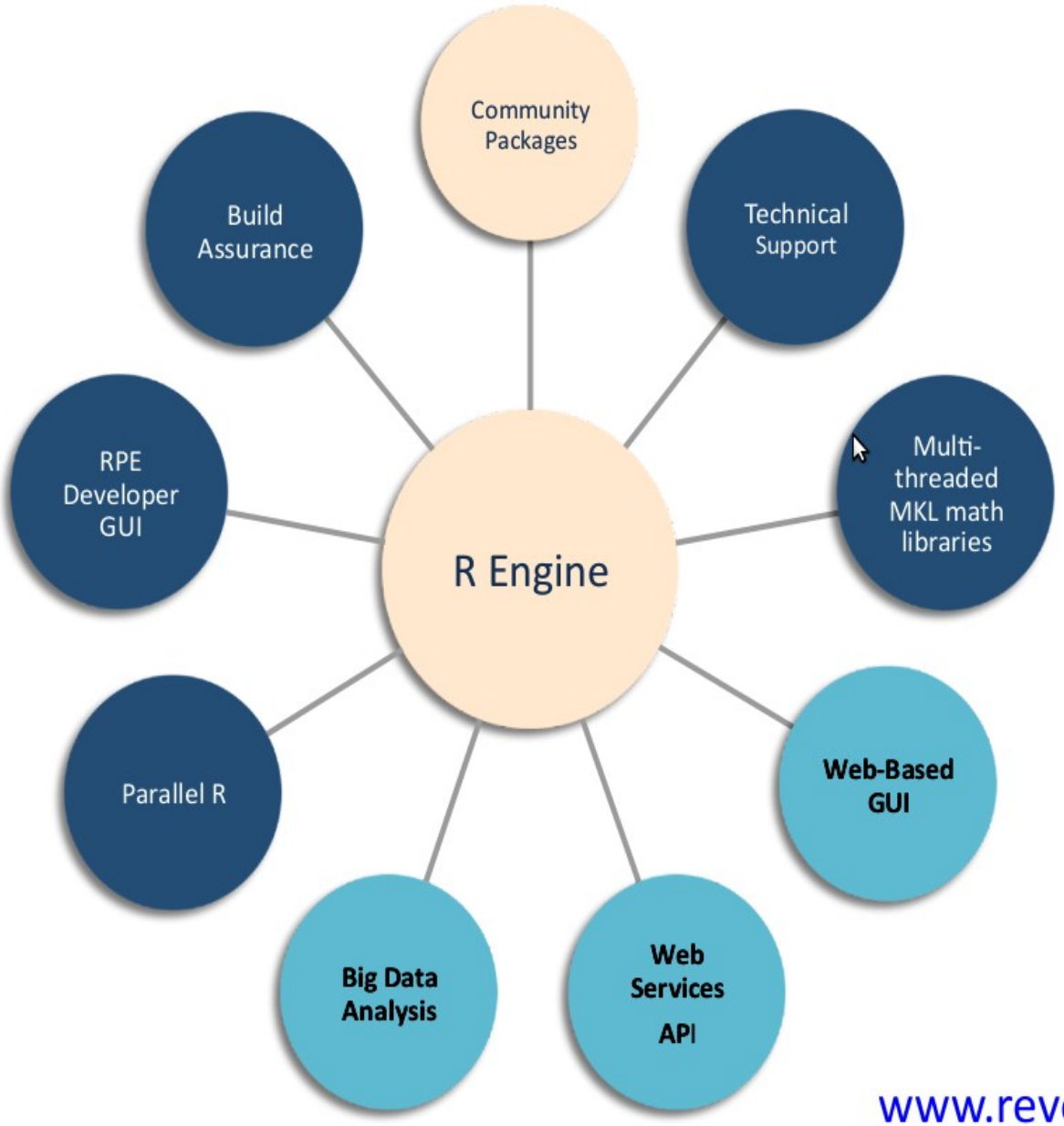
- Bigmemory - köztes megoldást kínál desktop környezetben, az ún. memory-mapping technikával egyfajta virtuális memoriát használ

- Snow, Multicore, Parallel, R+Hadoop, RHIPE, Segue – open source megoldások, amik jók, de nem triviális R elősimereteket kíván meg alkalmazásuk

R és Big Data

REVOLUTION ANALYTICS

R és Big Data



Community - Open Source

Revolution – Proprietary additions

Revolution – Forthcoming proprietary additions

www.revolutionanalytics.com/our-vision

R és Big Data

- Az átlag R felhasználó nem programozó

- Humán faktor: miért tanuljon meg valaki még egy új nyelvet, technológiát stb. ?

- $1000 / workstation Revolution Analytics licenc díj

Jövő

- R fejlesztések orvosolják a problémákat

- Egyre többen ismerik a nyelvet *"... the sexy job in the next 10 years will be statisticians"* (Hal Varian, Google)

Alternatívák

- Funkcionális JVM nyelvek
- Clojure Incanter
- http://incanter.org/

Clojure

- Scala breeze
- https://github.com/scalanlp/breeze

Köszönöm a figyelmet!

- Kereső Világ blog

  http://kereses.blog.hu/

Számítógépes nyelvészet szakmai blog
http://szamitogepesnyelveszet.blogspot.com/

- varju.zoltan@weblib.hu

- @zoltanvarju