

Új típusú adatforrások és technológiák a területi vizsgálatokban



RIERC

Regionális
Innováció- és
Vállalkozáskutatási Központ

Hornyák Miklós
tanársegéd
PTE KTK KMI



PÉCSI KÖZGÁZ
ahonnan a karrier indul

Az előadás tartalma

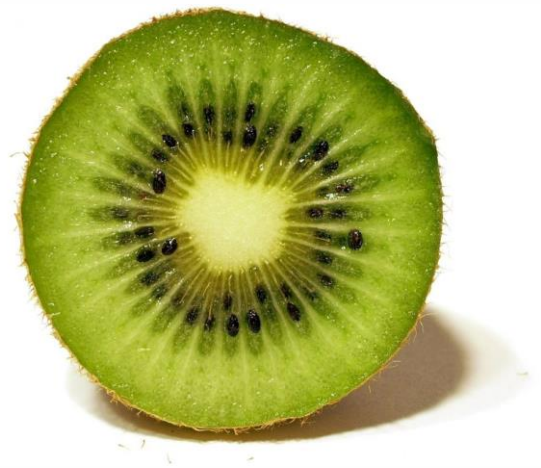
- Kutatási keret bemutatása - KIVI
- Társadalomtudományi probléma ismertetése
- Valós topológián alapuló gravitációs modell - GravityIX
- Szövegbányászati szentiment elemzés - HírIX
- Online jelenlét automatizált vizsgálata - WebIX
- Csődesemény adatbányászati előrejelzése - CsődIX
- Dolgozói attitűd internet-adatvezérelt elemzése – FOI
- Városi „pezsgés” vizsgálata – Meetup



Előadásom fókuszában a munka során tapasztalt nehézségek és kezelésük áll.

Kutatási keret ismertetése - Ökoszisztéma „puha” elemeinek mérése

- Rendszerszemléletű megközelítés a versenyképesség vizsgálatában
 - Intézményi és aktori alrendszerekre bontás
- Aktori alrendszer elemei a KKV-k, melyek vizsgálatára a Szerb-féle MKKVI modellt vettem át
 - Bővebben Szerb et al. 2014
 - WebIX és CsődIX támogatással
- MKKVI igazából a KKV belső kompetenciáit méri, működési környezettel nem foglalkozik
 - Így nem tekinthető versenyképességi indexnek, ehhez kell a környezet is
- MKKVI pillérstruktúrája az alrendszereket összekapcsoló interfész
- Intézményi alrendszer mérésére e pillérstruktúra mentén képzett indikátorokkal
 - Új típusú adatforrások és technológiák beépítése – adat és szövegbányászat, internet-adatvezéreltség
 - Iparági és területi bontás a környezet vizsgálatánál
- **Eredmény: Kompetencia és Intézményi Versenyképességi Index (KIVI)**
 - Ismereteink szerint eddig ezen alrendszereket nem kapcsolták össze egységes rendszerré
- KIVI segítségével SWOT-típusú jellemzésre van módunk
 - Belső: Erősség, Gyengeség -> KKV kompetenciák alrendszer
 - Külső: Veszély, Lehetőség -> Intézményi alrendszer

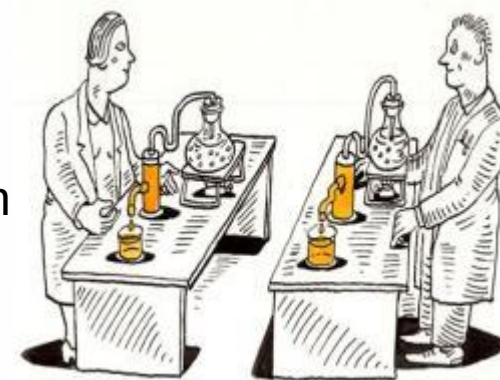


Általános kutatói problémahalmaz

- Közösségi érzékelők alkalmazása versus kérdőívezés
- Társadalomtudomány és új típusú adatforrások, eszközök
 - Big data jelenség kiterjedése – 3,4 vagy 5V
 - Bottom-up versus top-down kutatási paradigma – hipotézis versus kutatói kérdés?
 - Lehetőségek és korlátok ismerete, figyelembevétele
- Multidiszciplináris megközelítés nehézsége
 - Tudáselemek ritkán vannak meg egy személynél – domain specifikus tudás, IT ismeret, programozási ismeretek, adatok szervezése és tárolása
- Technológia gyorsan változik és fejlődik – lásd. képfeldolgozás
 - Számítási kapacitás és ennek támogatása – PC versus szolgáltatás alapú kutatás
- Adatokhoz való legális hozzáférés nehézsége
 - Crawling versus API versus vásárlás
 - Platform adatok használatának korlátai
- Reprezentativitási probléma
 - Digitális lábnyom versus digitális remete
 - Generációs különbségek – digitalizáció szintjében különbségek
- Eredmények bemutatása és értelmezése



Reproducible Research

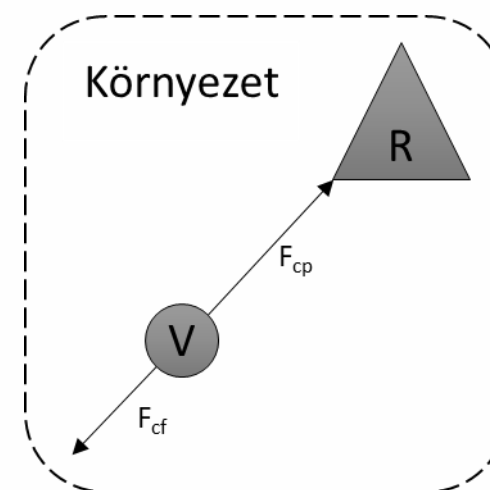
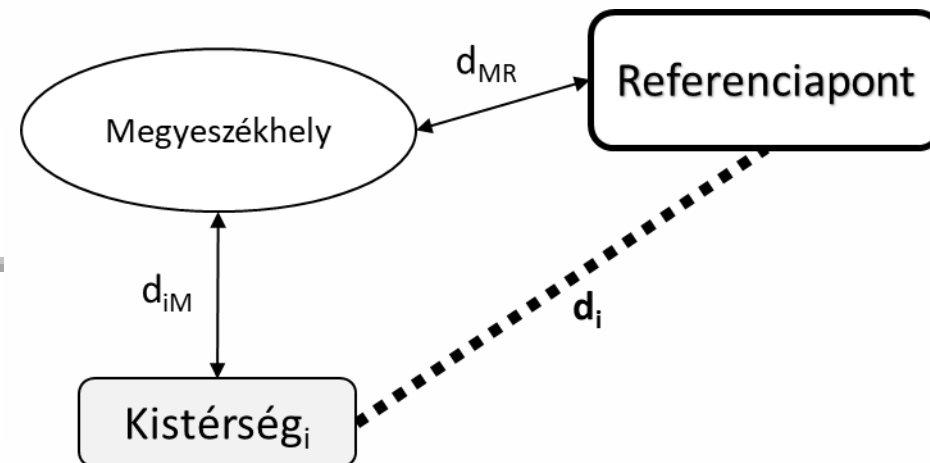


Forrás: <https://www.fosteropenscience.eu/foster-taxonomy/open-reproducible-research>

Valós topológián alapuló gravitációs modell - GravityIX

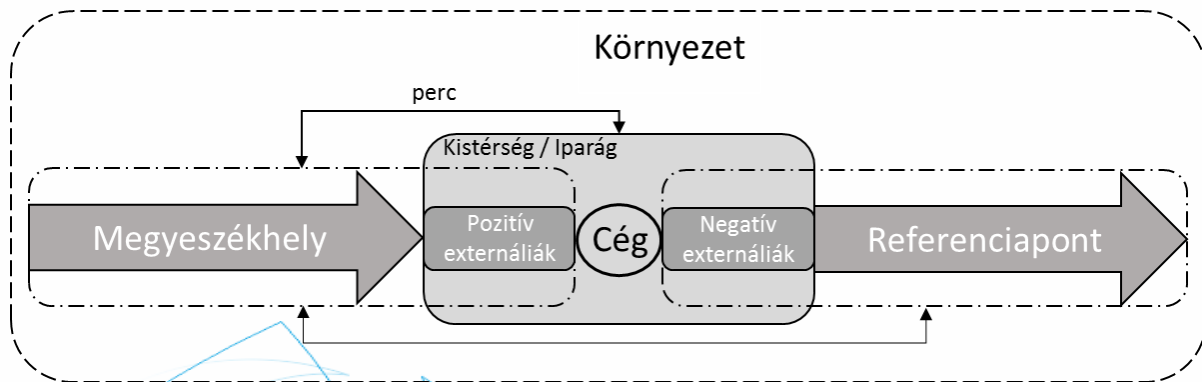
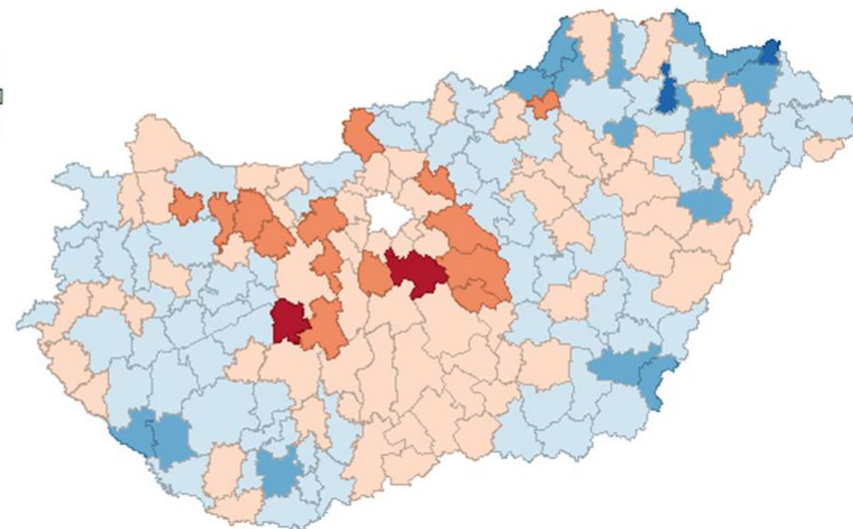
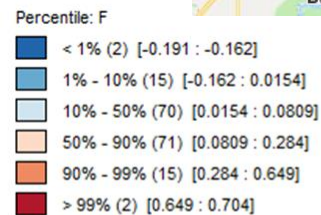
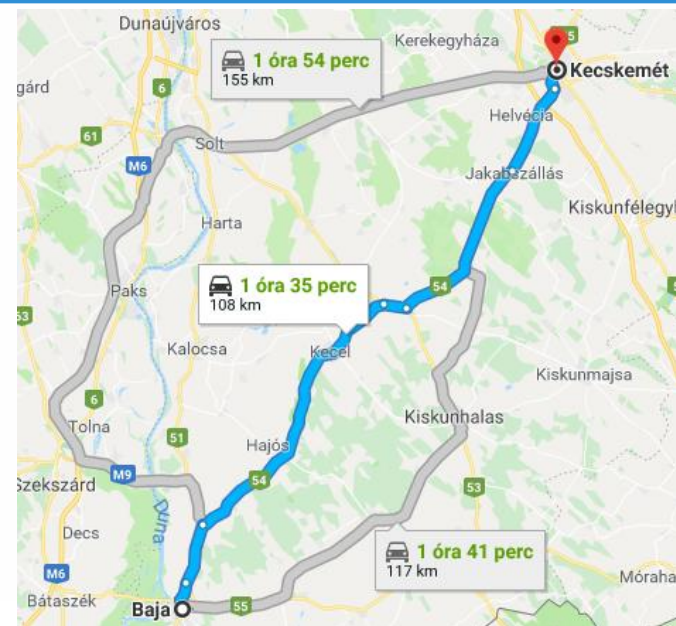
- Cél a területegységen működő KKV-re ható externáliák vizsgálata
 - Gravitációs modell mintájára
 - Egyik paraméter a távolság (d)
- Centripetális erő – pozitív externália: B2C, B2B
 - Lokális beágyazódást támogatja
- Centrifugális erő – negatív externália: iparági verseny
 - Lokális beágyazódást hátráltatja
- Modellek ritkán veszik figyelembe a valós földrajzi helyzetet
- Probléma definiálása
 - Földrajzi topológia beépítése a modellbe
 - Nem légvonalbeli mérés

$$D_{ij} = g\left(\frac{P_i P_j}{d_{ij}^2}\right)$$



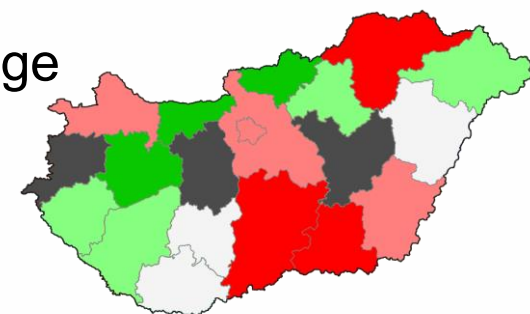
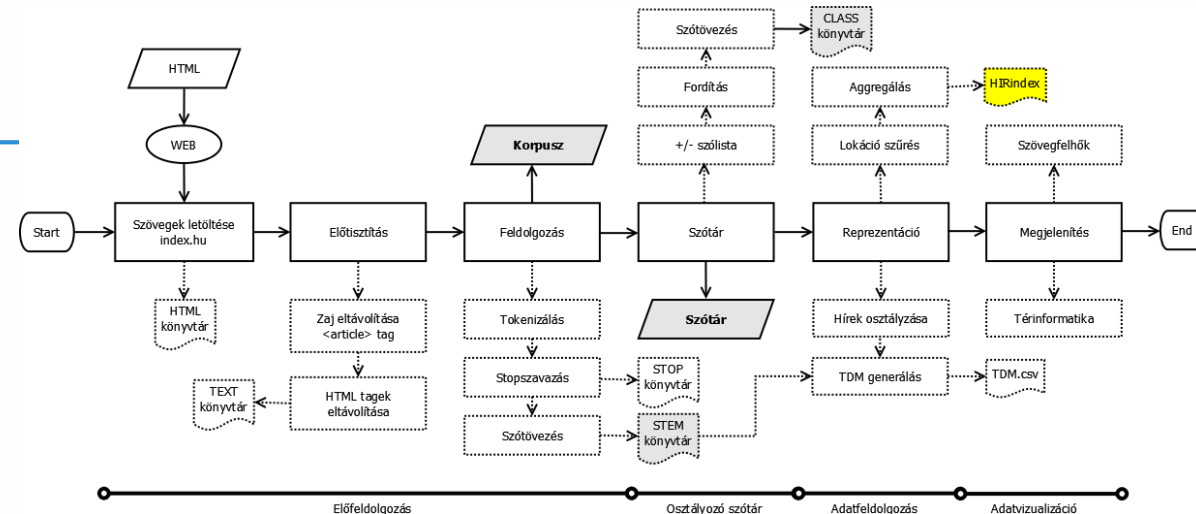
Valós topológián alapuló gravitációs modell - GravityIX

- Megoldás: térinformatika
 - Városi elérhetőség úthálózaton keresztül
- Felhasznált eszköz
 - Google Maps
 - TEIR – GeoX Kft. elérési adatok
- Alkalmazott tudáselem
 - Térinformatikai távolságmérés
- Tapasztalt nehézségek és megoldásaik
 - Google útvonalterv nehézkes
 - GeoX adatok nem voltak elérhetők, ez azóta változott!



Szövegbányászati szentiment elemzés

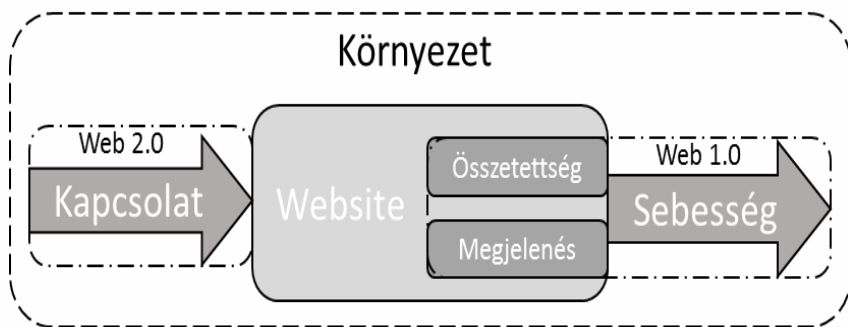
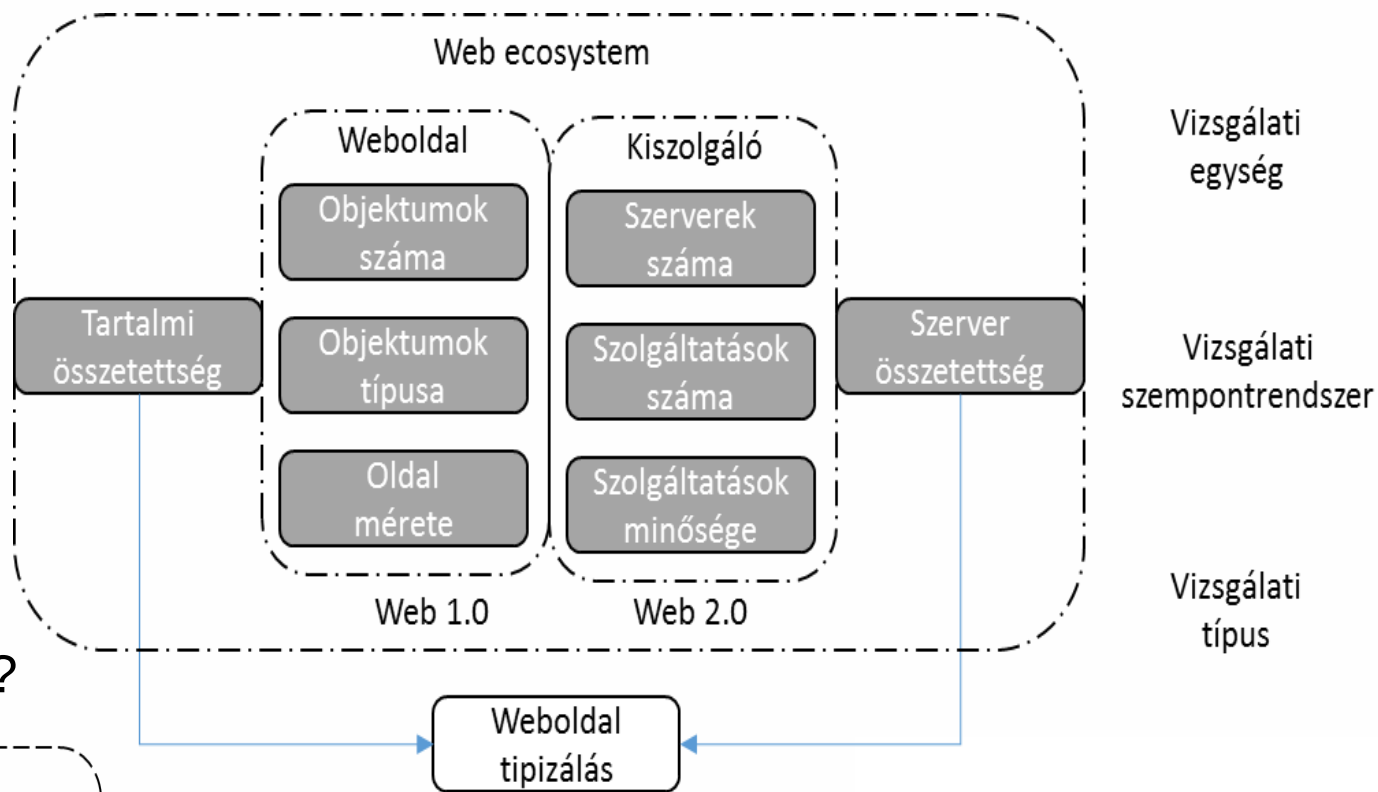
- Cél: Hírcorpus képzése, lokációhoz kapcsolása
 - Pozitív, negatív, semleges szavak azonosítása – terület inkább +/-
- Nehézség: Magyar nyelvű szövegek problémája
 - Nyelvészeti tudáselem hiánya
 - „Dobozos” szoftverek hiányosságai – Rapidminer, Python, R
 - Opensource versus fizetős rendszerek – SPSS Clementine
- Nehézség: Tudományosan validált magyar szentiment szótár hozzáférhetősége
 - Konnotációs problémák az angol fordításnál – speciális tudáselem megjelenése
- Szövegfeldolgozási anomáliák kézi kezelése
 - Családnév versus település
 - Szótövezés
- Kísérletek – szöveg-perszonifikáció a Big5 modell alapján (LIWC szótár alkalmazása)
- Algoritmusok fejlődésével és a corpus növelésével javítható a modell teljesítménye
 - Corpus méretének növelése nagy munkát igényel, minden szempontból heterogén környezet
 - Corpus adatforrásainak tágítása nehézkes – Twitter részadatok, hazánkban nem elterjedt; Facebook bezárt; Intsagram képek



Pap Budapest Szeged
Keszthely Varga Balaton
Bajna
Hajdúszoboszló Gyula
Balatonfüred

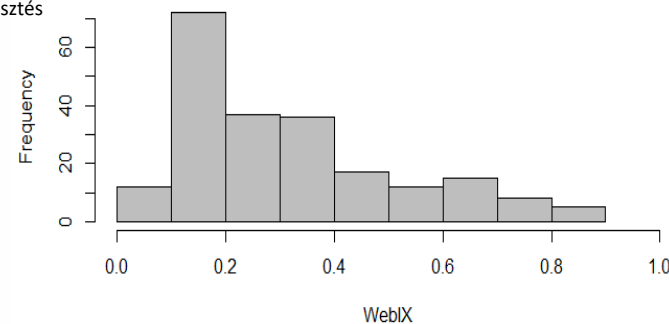
Online jelenlét automatizált elemzése - WebIX

- **Probléma**
 - Magas az erőforrás igénye
 - Szubjektivitással terhelt
- **Web ökoszisztéma vizsgálata**
- **Technicista elemzési módszer**
- **Megoldás**
 - Python program
 - 22 mutató – 4 pillér – 2 alindex
- **Online jelenlét – weboldal vs. Fb.?**



WebIX koncepcionális modellje, forrás: saját szerkesztés

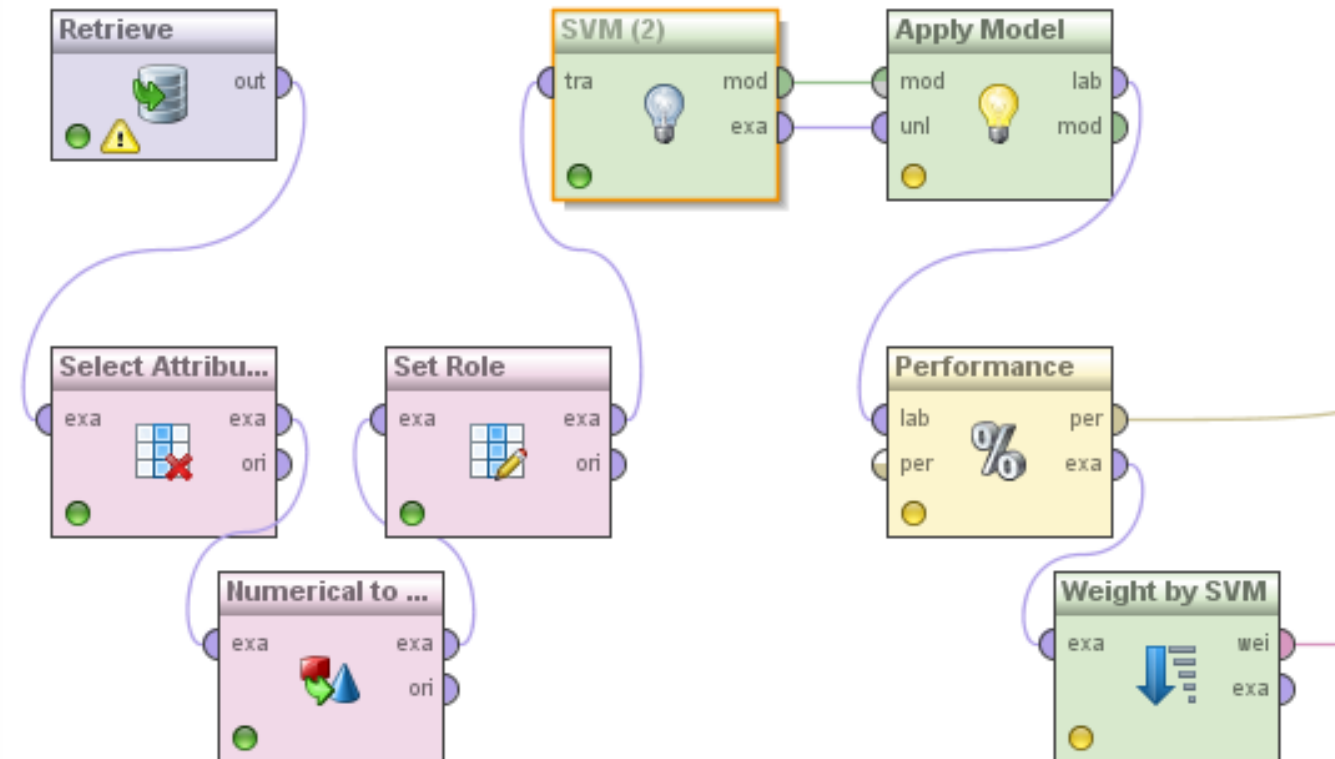
Butkiewicz modelljének sémája, forrás: Butkiewicz et al. 2011 alapján, saját szerkesztés



Csődesemény adatbányászati előrejelzése - CsődlX



- Cél: KKV csődkockázat mérése
- Adatok
 - 2011 és 2012 cégs adatok – mérleg, eredmény-kimutatások, NUTS-2, iparág, eljárások
 - MKKVI cégei – 706 db
- Minta
 - Tanulóminta – 1180 db cég – 50% ban 2013-as csődesemény
 - Teszt adatok – 506 db cég – 50%-ban 2013-s csődesemény
- Tudáselem
 - Adatbányászati gépi tanulási modellek ismerete
- Modell
 - Neurális háló segítségével predikciós modell építése
 - SVM – 82%-os pontosság
- Eredmény – optimista megközelítés
 - 144 cég esetében csődkockázatosnak minősíti
- Probléma
 - Tanuló- és tesztminta beszerzése
 - Frissítési nehézségek miatt lezártuk a projektet



Dolgozói attitűd internet-adatvezérelt elemzése (forrás: saját szerkesztés)

- Internetes keresések jövő orientáltsága és a GDP közötti kapcsolat (Preis et al. 2012)

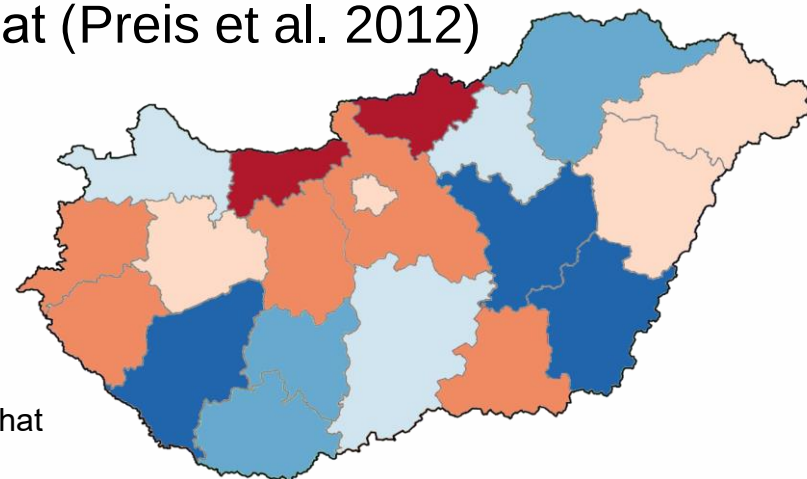
- Magyarországi IP-címekről indított szöveges keresések vizsgálatát végeztem

- Google Trends alkalmazása

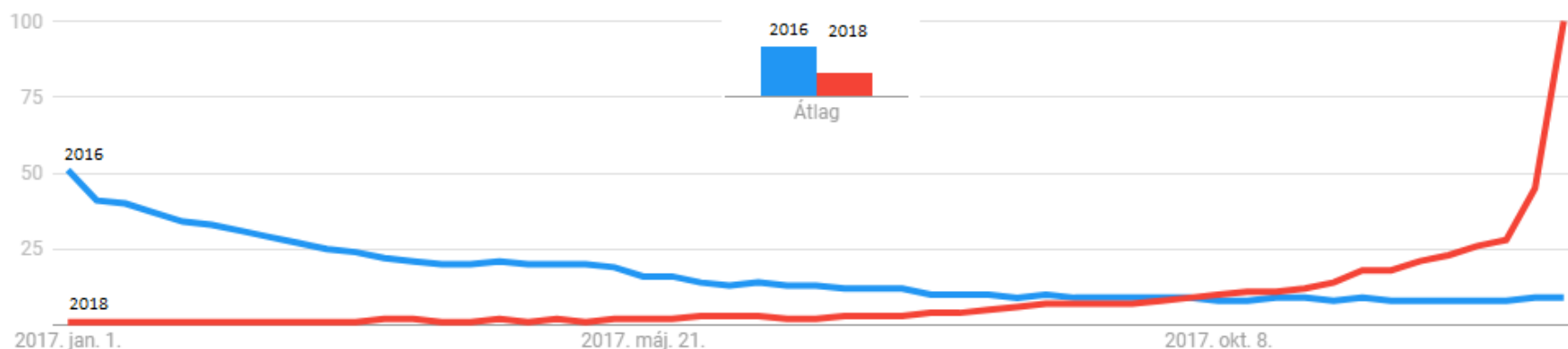
- Lokáció, időszak, témacsoport és a keresési kifejezés szűrés
- Megyékre szűrés – bázis évhez viszonyítás – egyszerű keresési kifejezés

- Korlátok

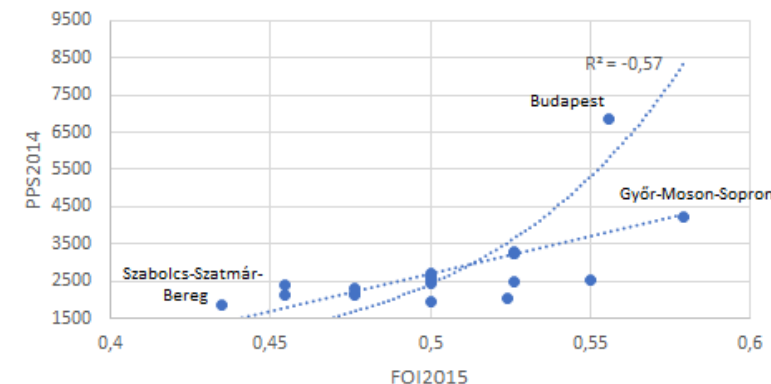
- Nem abszolút értéket szolgáltat a Google, csak arányokat pl. 100 a leggyakoribb, 0 nincs adat
- Nem ismert az alkalmazott algoritmus, így fekete dobozként tekintünk rá – bármikor és bárhogy változhat
- Internet penetráció és digitális írástudás – bizonyos rétegek kimaradnak a vizsgálatból



Magyarországról indított keresések



FOI2015 - PPS 2014



Városi „pezsgés” vizsgálata platform adatok alapján (forrás: saját szerkesztés)



- Meetup.com platform adatai alapján várostérségek jellemzése

- Gyökeresen más szemlélet – lokáció központú, pufferzóna alkalmazása
- API hívások segítségével, többszörös hierarchikus felépítés – Python programozás

- Adatok gráf alapú adatbázisba leképezve – relációs adatszerkezet helyett

- Alkalmazása nem elterjedt, új tudáselem

- Előny a kapcsolatok és csomópontok természetes használata

- Cities – Groups – Events - Topics

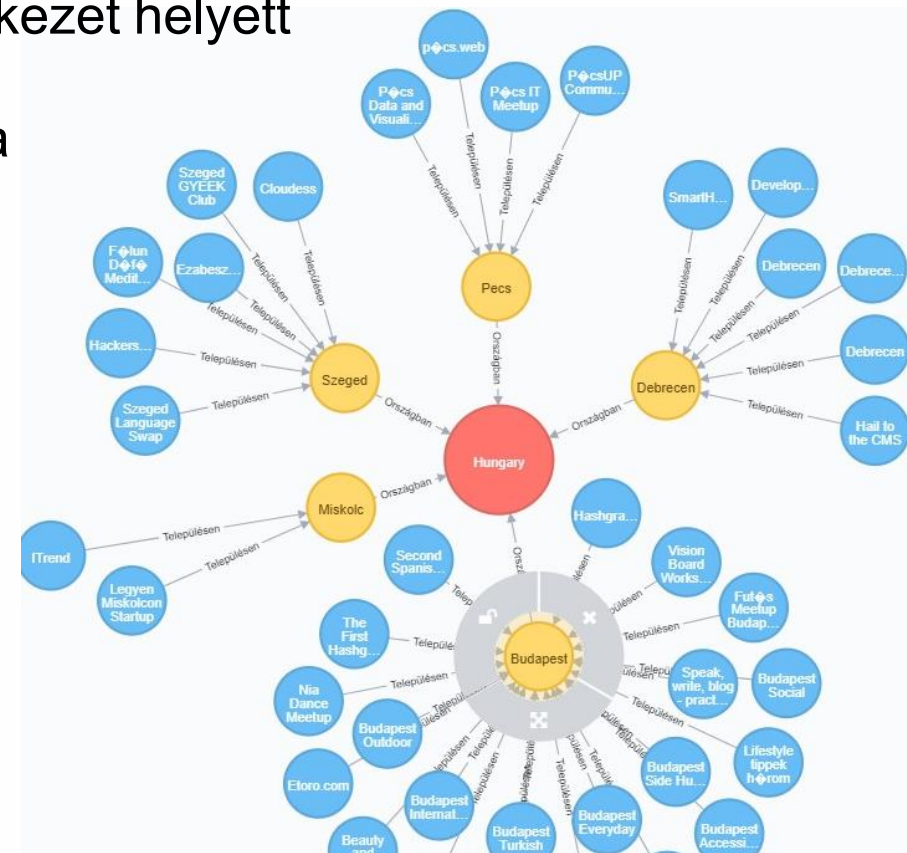
- Előny a gyors, interaktív és intuitív eredmény

- Cypher és Graph Data Visualization

- Nehézség az új szemlélet elsajátítása

- strukturált vs. strukturálatlan adatok tárolása

```
MATCH (a:Ország)
WHERE a.country = "hu"
MATCH (b:HUTelepülés)
MATCH (c:HUCsoport)
WHERE c.members > "100" AND c.cityID = "1016532" OR c.cityID = "1016533" OR
c.cityID = "1016534" OR c.cityID = "1016535" OR c.cityID = "1016536"
RETURN a, b, c LIMIT 50000
```



Tapasztalatok összefoglalása és jövőbeli irányok

Tapasztalatok összegzése

- Sokrétű tudáselem együttes alkalmazása
 - Kommunikációs problémák
- Adatok beszerzésének legális volta
- Adatfrissítések folyamatos biztosítása
- Megfelelő algoritmusok
- Opensource vs. Fizetős rendszerek
- Elmozdulás a vizuális elemek elemzésének irányába
 - Kickstarter kampányok video elemzése – ökoszisztéma szoft elemei
 - Instagram fotók elemzése – lokációk social sensor típusú vizsgálata



RIERC

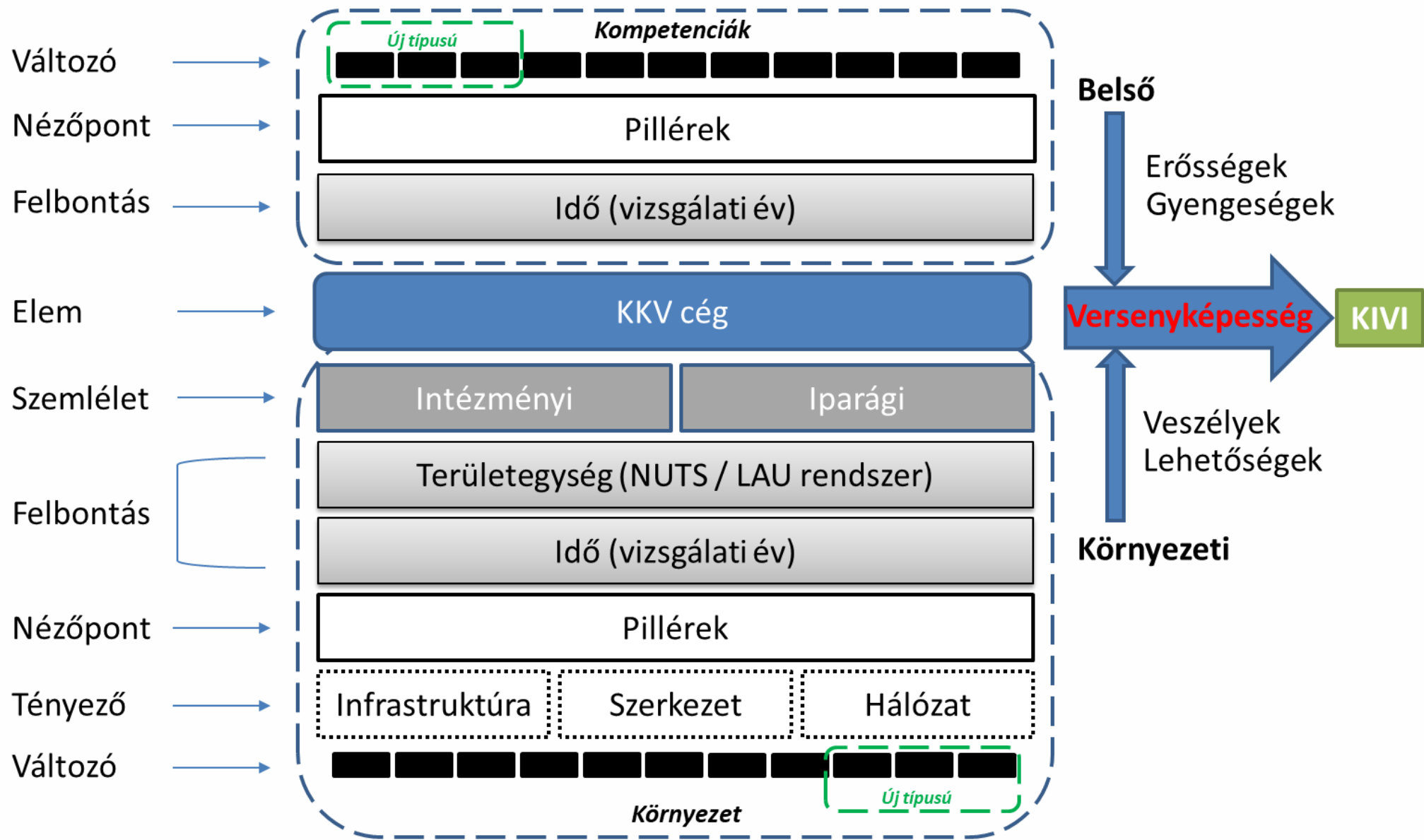
Regional
Innovation and Entrepreneurship
Research Center

Köszönöm a figyelmüket!

hornyak.miklos@pte.hu

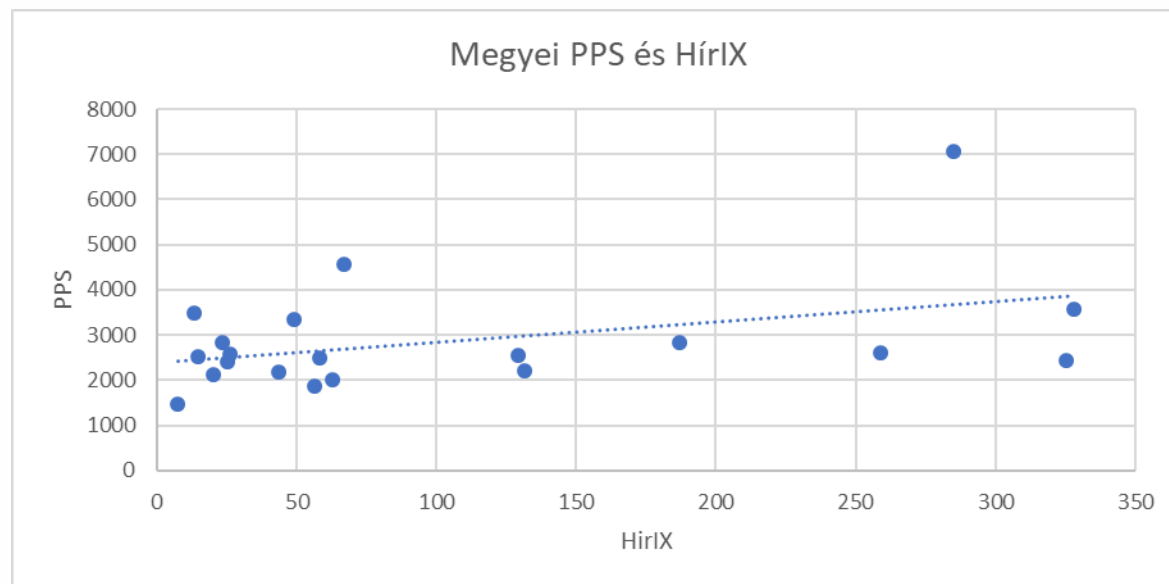


PÉCSIKÖZGÁZ
ahonnan a karrier indul



KIVI modell felépítése

Kutatói kérdés vizsgálata – HírIX példájában



- T-próba

• Szabadságfok (f)	19
• Szignifikancia szint	0,01
• Student tábla ($t_{0,01}$)	2,539
• T-próba (t_p)	1,9355275037

Mivel $t_p < t_{0,01}$ ezért a nullhipotézist elfogadjuk, vagyis a megyei PPS és HírIX értékek közti korreláció szignifikánsan nem tér el a nullától 99%-os szinten.

- Normalitás vizsgálat

A ferdeség (-0,14) és a csúcsosság (1,14) értékek alapján a minta normális eloszlásúnak tekinthető.

- Korreláció

Pearson-féle korrelációs együttható értéke ($r = 0,415056282$) alapján a megyei PPS és FOI értékek között kapcsolat azonosítható.

- Szignifikancia

- Nullhipotézis: a korrelációs együttható a megyei PPS és HírIX értékek között $= 0$
- Alternatív hipotézis: a korrelációs együttható a megyei PPS és HírIX értékek között $\neq 0$

- Szerb, L. et al. (2014): *Mennyire versenyképesek a magyar kisvállalatok? A magyar kisvállalatok (MKKV szektor) versenyképességének egyéni-vállalati szintű mérése és komplex vizsgálata*, Marketing és Menedzsment 11/2014; XLVIII.(Különszám), pp. 3-21.
- Preis T., Moat H. S., Stanley H. E., Bishop S. R., (2012): *Quantifying the advantage of looking forward*, Scientific Reports 2 350, doi:10.1038/srep00350