

# Új típusú adatforrások és technológiák a kisvállalati versenyképességi vizsgálatokban



**RIERC** Regionális  
Innováció- és  
Vállalkozáskutatási Központ



**Big Data Research Group**

PÉCSI TUDOMÁNYEGYETEM  
SZENTÁGOTHAJ JÁNOS KUTATÓKÖZPONT

Hornyák Miklós  
adjunktus  
PTE KTK KMI

VÁLTOZÁS, ÚJRATERVEZÉS ÉS FEJLŐDÉS, tudományos konferencia PTE KTK, Pécs 2020. november 5-6.



Pécsi Tudományegyetem Közgazdaságtudományi Kar

# Az előadás tartalma

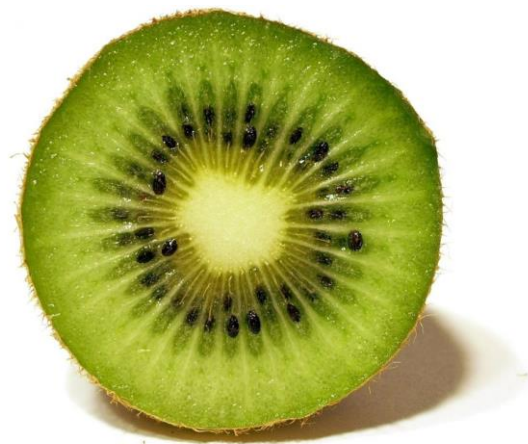
- Kutatási keret bemutatása - KIVI
- Társadalomtudományi probléma ismertetése
- Szövegbányászati szentiment elemzés - HíriX
- Online jelenlét automatizált vizsgálata - WebiX
- Csődesemény adatbányászati előrejelzése - CsődiX
- Dolgozói attitűd internet-adatvezérelt elemzése – FOI
- Városi „pezsgés” vizsgálata – 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
- 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 mérése e pillérstruktúra mentén képzett indikátorokkal – új típusúak is!
  - 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
- KIVI segítségével SWOT-típusú jellemzésre van módunk
  - Belső: Erősség, Gyengeség -> KKV kompetenciák alrendszer
  - Külső: Veszély, Lehetőség -> Intézményi alrendszer

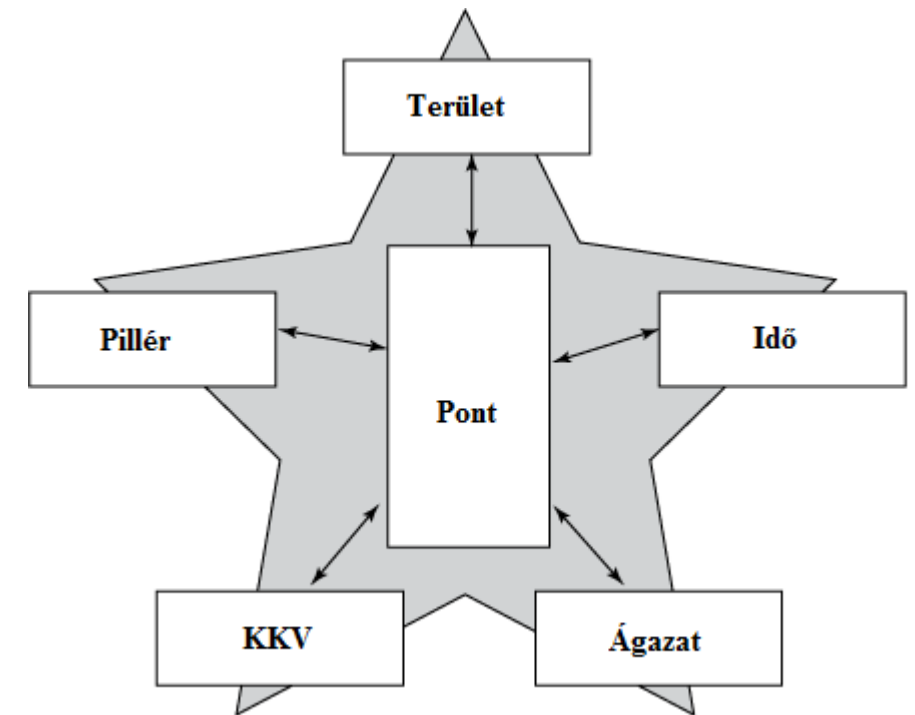


# Miért kell egy új index? - a „Big Data” jelenség

- KKV működésének környezete nem vizsgált az MKKVI-ben
- Rendszerszemléletű megközelítés alapján a rendszer eleme ÉS környezete is fontos
- Kis és középvállalati (KKV) versenyképességnél a cég a rendszer eleme
- Nem fekete doboznak tekintem, hanem erőforrás alapú (RBW) megközelítést alkalmazom
  - Belső, kompetencia alapú folyamatok vizsgálására az MKKVI alkalmas
- Külső, intézményi környezet vizsgálatát alakítom ki
- E kettő alrendszer között interfész kiépítése szükséges
- Ennek keretét az MKKVI pillérstruktúrája biztosít
- **Új típusú adatforrások és technológiák beépítése**

# Koncepcionális keretrendszer kialakítása

- Erős IT támogatás szükséges
- Összetett adatbázisszerkezet készítése
  - Változó szintű adattárolás
  - 110 egyedi változó -> 101 területi, 9 iparági szemléletű
  - 52% LAU1, 15% NUTS3, 26% NUTS2
  - Saját fejlesztésű változó a pillérek 30%-ban
- Térinformatikai keretrendszerbe illesztés
  - Alap térinformatikai funkciók
  - Térképi ábrázolhatóság
- Adatkocka alkalmazása – „Actionable KPI” képzése
  - Öt dimenzió: terület, idő, ágazat, pillér, cég
  - Metszéspontokban tényadat – komplex versenyképességi pont
  - Lefúrással pillér szintig
  - Szeleteléssel szemléletek előállítása – területi/iparági
  - Fölgörgetéssel LAU1 -> NUTS3 -> NUTS2 -> NUTS1
  - Városterzségek elméleti kezelhetősége



Egy főre jutó bruttó hazai termék vásárlóerő egységben (PPS)

- Jól elérhető adat
- Fejlődést mutató mutatóink vannak
- Növekedést jól mutatja

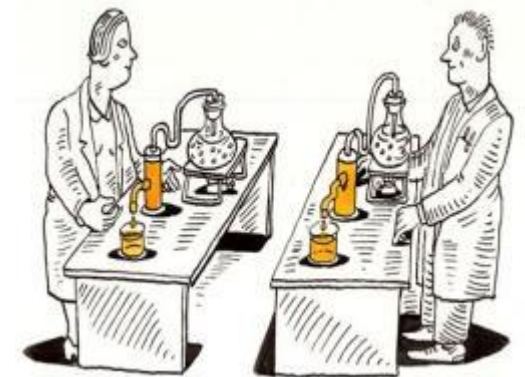


# Általános kutatói problémahalmaz

- Közösségi érzékelők alkalmazása vs. kérdőívezés
- Társadalomtudomány és új típusú adatforrások, eszközök
  - Big data jelenség kiteljesedése – 3,4 vagy 5V
  - Bottom-up vs. top-down kutatási paradigma – hipotézis vs. 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, szövegbányászat
  - Számítási kapacitás és ennek támogatása – PC vs. szolgáltatás alapú kutatás?
- Adatokhoz való legális hozzáférés nehézsége
  - Crawling vs. API vs. vásárlás
  - Platform adatok használatának korlátai
- Reprezentativitási probléma
  - Digitális lábnyom vs. digitális remete
  - Generációs különbségek – digitalizáció szintjében különbségek
- Eredmények bemutatása és értelmezése



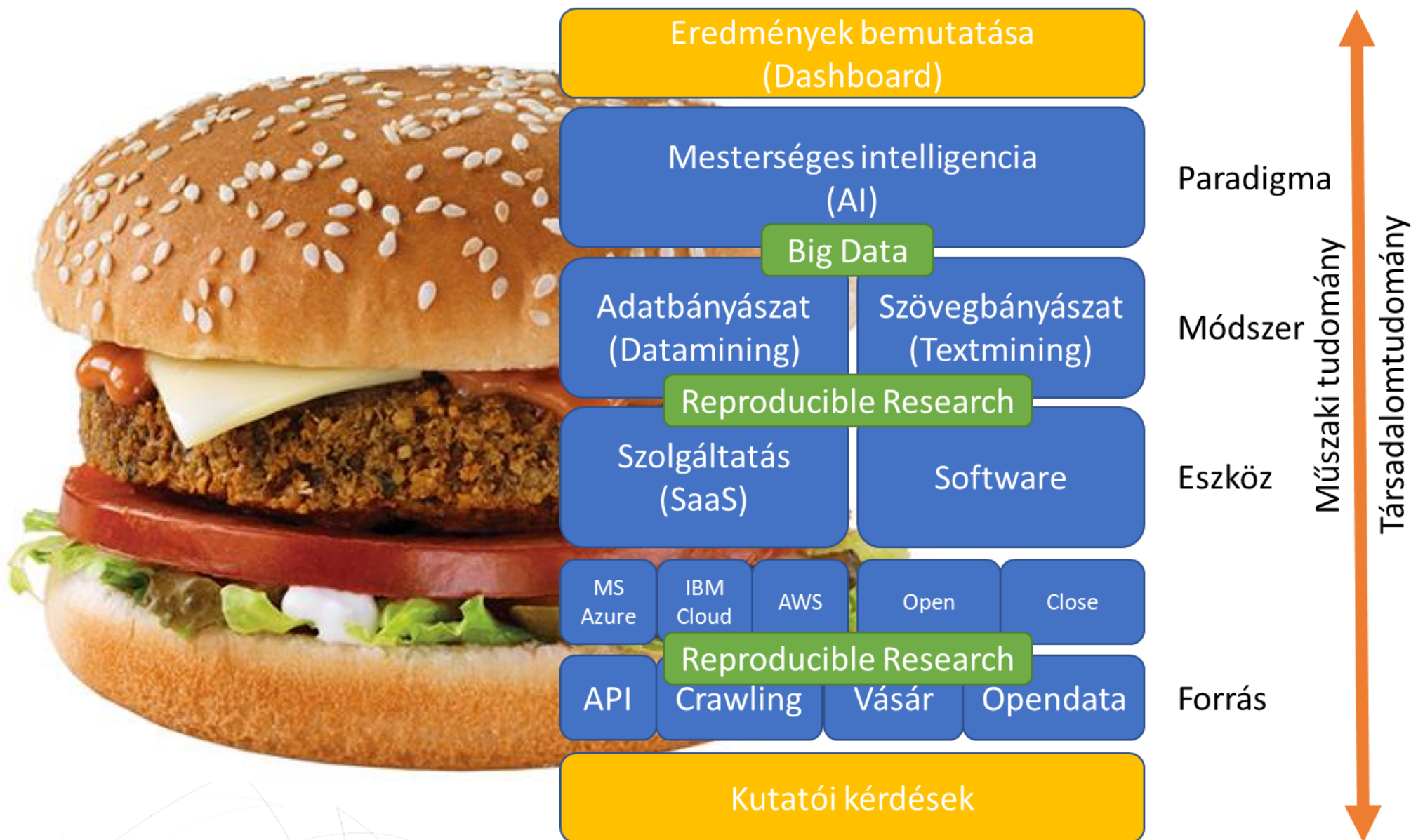
Reproducible Research



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

# „Full stack” development – a „hamburger modell”

(forrás: saját szerkesztés)



# Véleménytájolás-alapú elemzés (HírIX)

- Cél: Hírek gyűjtése, lokációhoz kapcsolása

- Pozitív, negatív, semleges szavak azonosítása

- Nehézségek

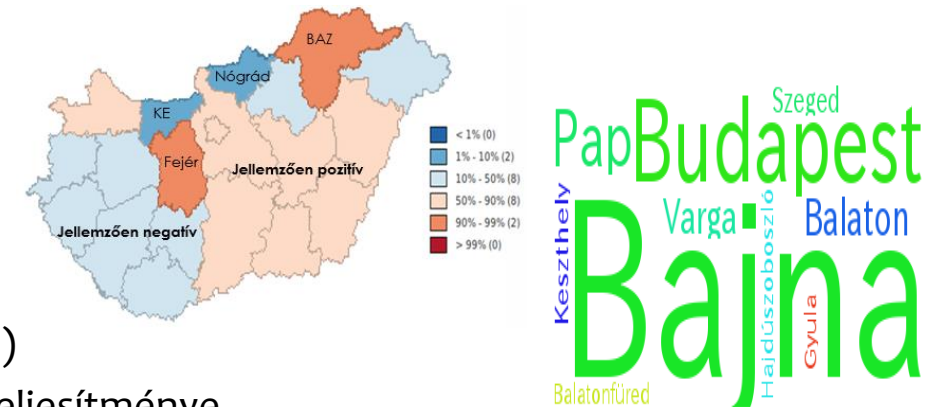
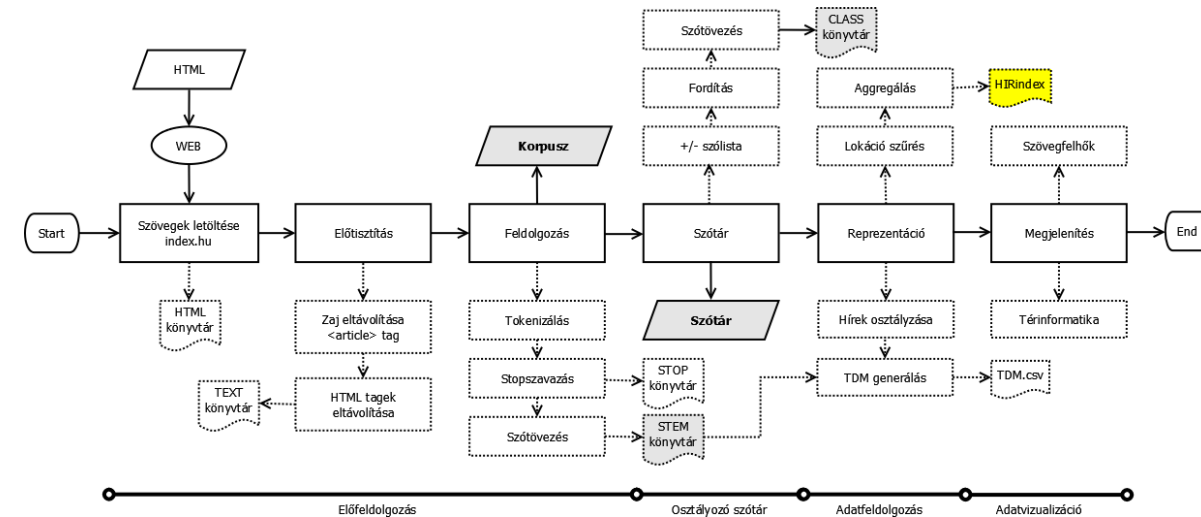
- Magyar nyelvű szövegek problémája
  - Nyelvészeti tudáselem
  - „Dobozos” szoftverek hiányosságai – Rapidminer, Python, R
  - Tudományosan validált magyar szentiment szótár hozzáférhetősége
  - Konnotációs problémák az angol fordításnál
  - Családnév versus település

- Jövő

- Szöveg-perszonifikáció a Big5 modell alapján (LIWC szótár alkalmazása)
  - Algoritmusok fejlődésével és a corpus növelésével javítható a modell teljesítménye
  - Szövegbázis méretének növelése nagy munkát igényel, minden szempontból heterogén IT környezet
  - Szövegbázis adatforrásainak tágítása nehézkes képek
    - Twitter részadatok, hazánkban nem elérhető; Facebook bezárt; Instagram képek

- Újdonságtartalom

- Szövegbányászat alkalmazása
  - Folyamat kidolgozása és az eredmények vizualizálási módjai – szófelhő, Conway szófelhő

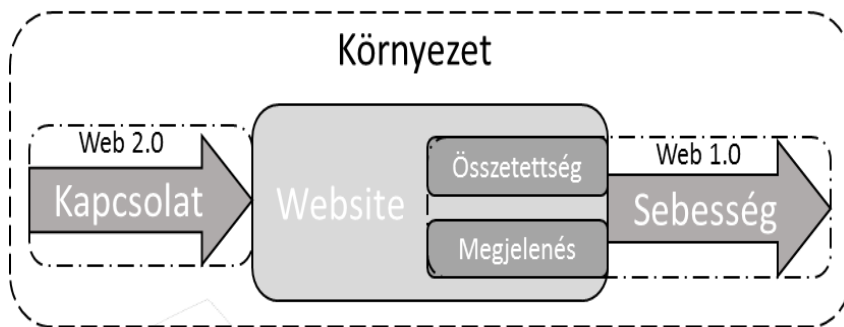


- NUTS3-as HírIX térkép
- Jelen technika nem elégséges
- $r=0,416$ , nem szignifikáns

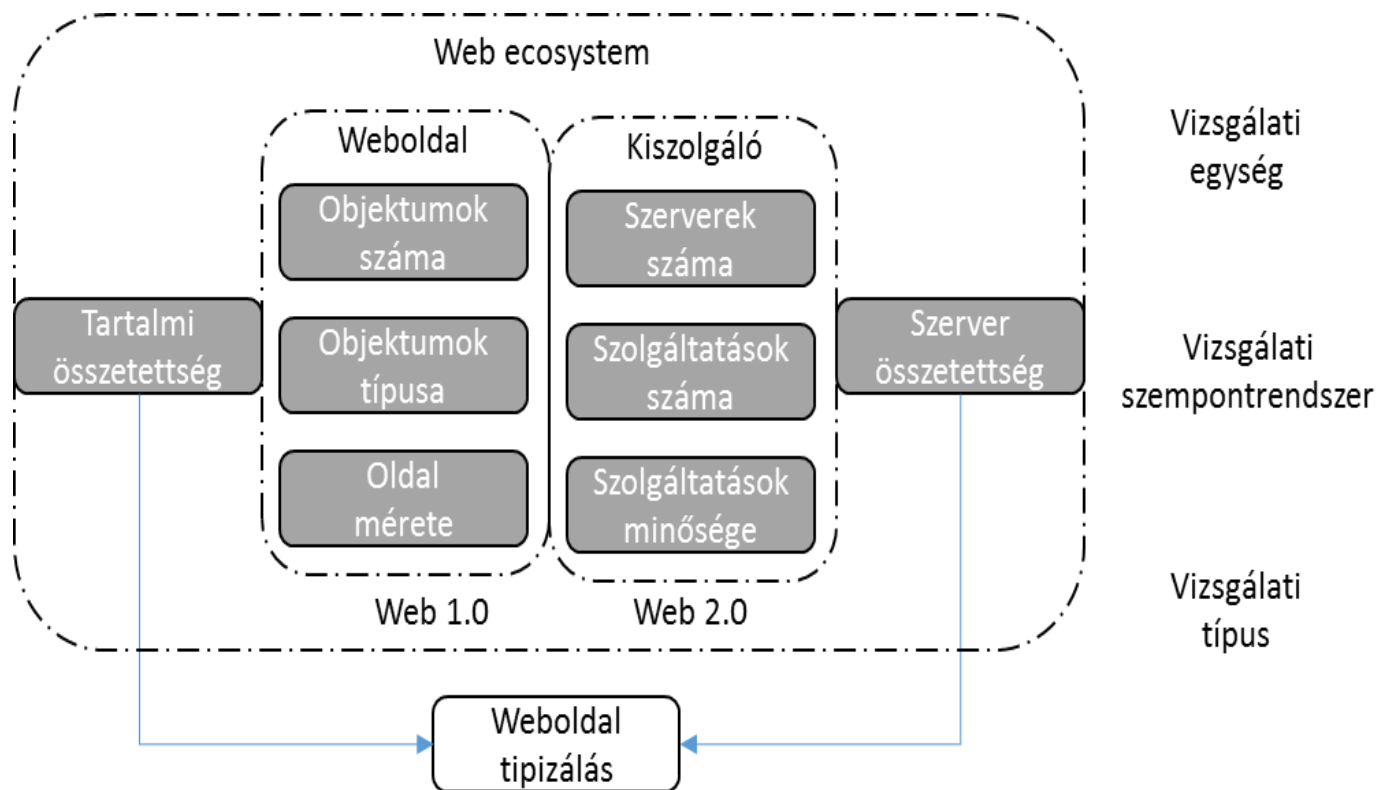


# Online jelenlét automatizált elemzése (WebIX)

- 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



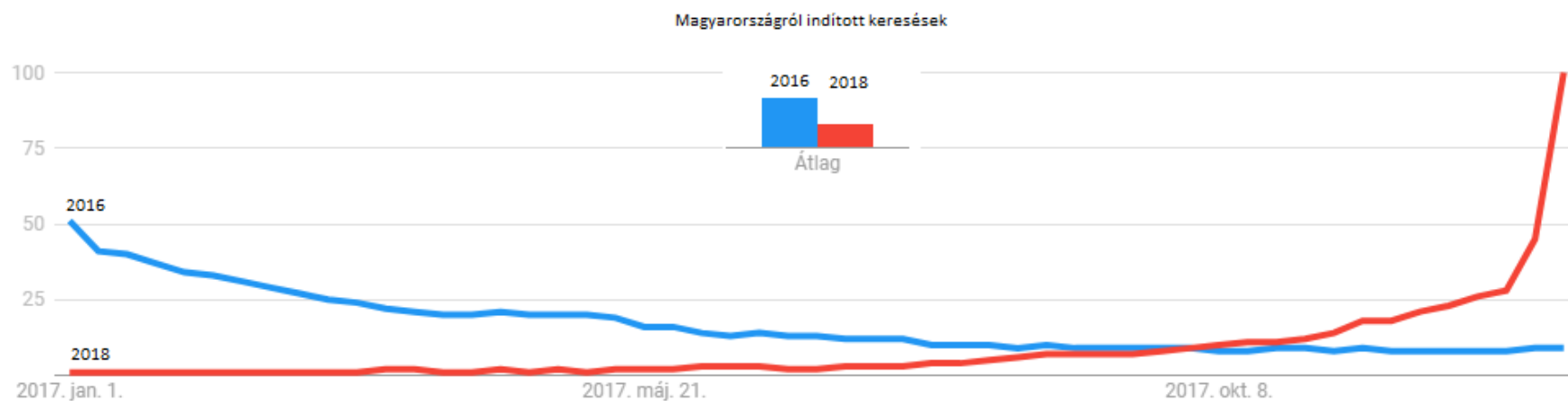
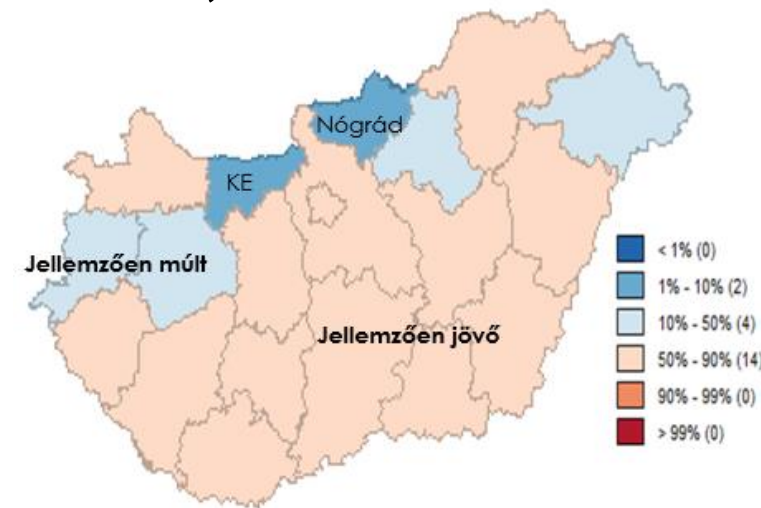
Butkiewicz modelljének sémája, forrás: Butkiewicz et al. 2011 alapján, saját szerkesztés

Miért van erre szükség?



# Dolgozói attitűd internet-adatvezérelt elemzése (FOI)

- Internetes keresések jövő orientáltsága és a GDP közötti kapcsolat (Preis et al. 2012)
  - Magyarországi IP-címekről indított szöveges keresések vizsgálatát végeztem
    - Innovációs készség, 2015-ös bázisév
- 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, o 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



- Munkavállalói attitűd jellemzése megyei szinten
- Pearson korreláció vizsgálat ( $r=0,662$ ,  $0,01$ )

# Városi „pezsgés” vizsgálata platform adatok alapján

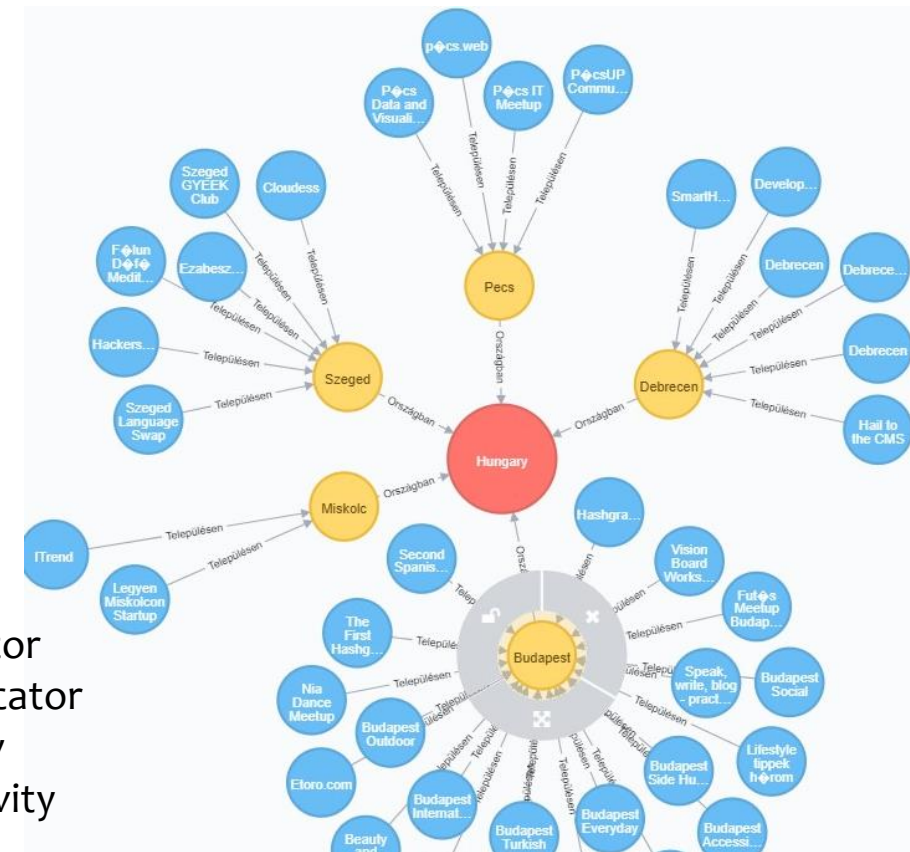
- Meetup.com platform adatai alapján várostérségek jellemzése
  - Gyökeresen más szemlélet – lokáció központú, pufferzóna alkalmazása
  - API hívások segítségével, többszörös hierarchikus felépítés – Python programozás
- Adatok gráf alapú adatbázisba leképezve – relációs adatszerkezet helyett
  - Alkalmazása nem elterjedt, új tudáselem
  - Tech Groups - Tech Group Members - Tech Events
- 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
- Meetup API változása
  - Megismételhető kutatás nehézsége

Ország mértékek:

- Meetup Tech Group Indicator
- Meetup Tech Member Indicator
- Meetup Tech Event Activity
- Meetup Tech Member Activity

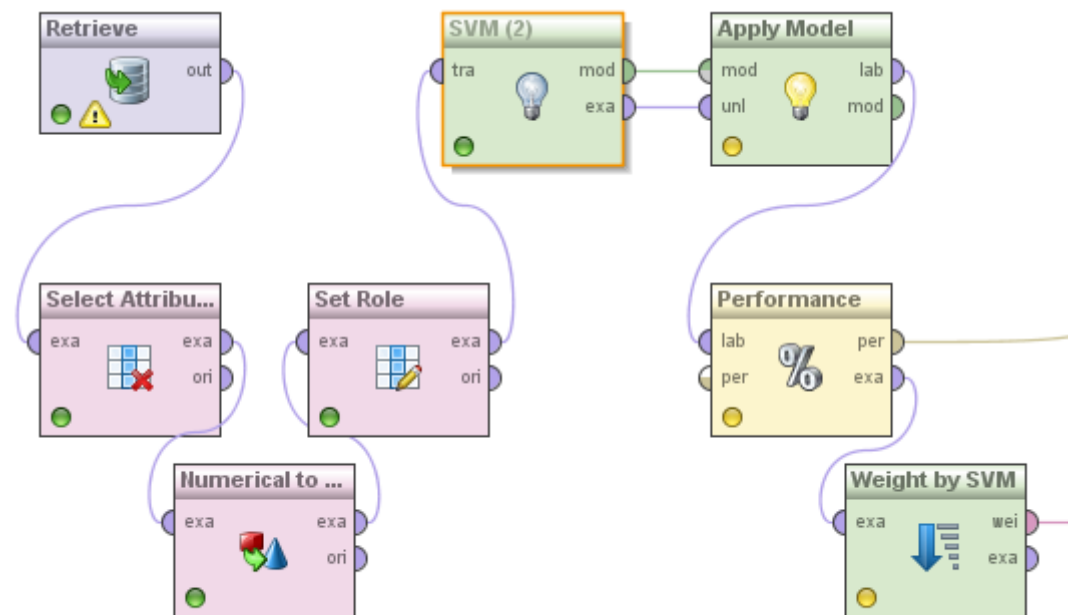


```
MATCH (a:Ország)
WHERE a.country = "hu"
MATCH (b:HUTEpülés)
MATCH (c:HUCsoport)
WHERE c.members > "100" AND c.cityID = "1016532" OR c.cityID = "1016533" OR
c.cityID = "1016534" OR c.cityID = "1016535" OR c.cityID = "1016536"
RETURN a, b, c LIMIT 50000
```



# Csődesemény adatbányászati előrejelzése (CsődIX)

- Cél: KKV csődkockázat előrejelzése
- Adatok
  - 2011 és 2012 cégadatok – mérleg, eredmény-kimutatások, NUTS-2, iparág, csődeljárások; ANONIMITÁS
  - MKKVI cégei – 706 db
- Minta – 47 különböző attribútum
  - 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



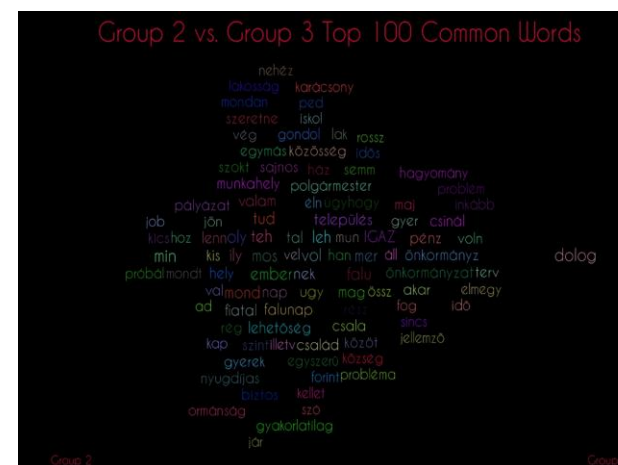


# Szövegbányászat az ormánsági kultúrakutatásban (forrás: Ragadics – Hornyák 2017)

- Szövegkorpusz
  - 25 mélyinterjú ormánsági vezetőkkel
- Term versus Document Frequencies
  - Fontos szavak vs. fontos témák
- Témaszpecifikus stopszavak pl. falu
- Kérdésenkénti vizualizálás
  - Szófelhők
- Településcsoportok képzése
  - Heterogén szövegmennyiség
  - Összehasonlító szófelhők
- Szakértői tudás mindkét területről



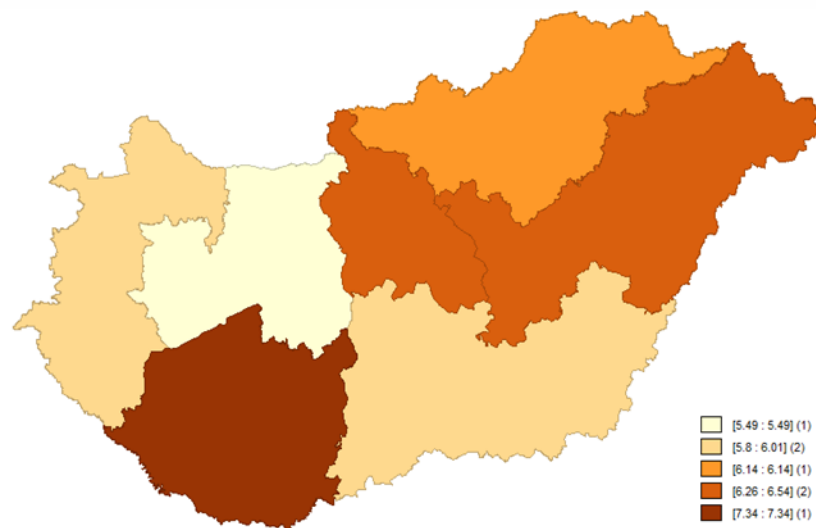
| ID | Település                                                                                                                                        | Átlagos hossz | Legrövidebb | Leghosszabb |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------|---------------|-------------|-------------|
| 1  | Drávafook, Drávasztára, Kémes, Drávacsepely                                                                                                      | 3444          | 1426        | 6119        |
| 2  | Besence, Hegyszentmárton, Ipacsfa, Sósvertike, Adorjás, Diósvizsló, Drávaszerdahely, Nagycsány, Páprád, Sámod, Sumony, Szaporca, Tengeri, Téseny | 5884          | 2350        | 13821       |
| 3  | Kákics, Marócsa, Ózdfalu, Cún, Kórós, Lúzsok, Okorág, Ózdfalu, Vejti                                                                             | 7036          | 1259        | 18023       |



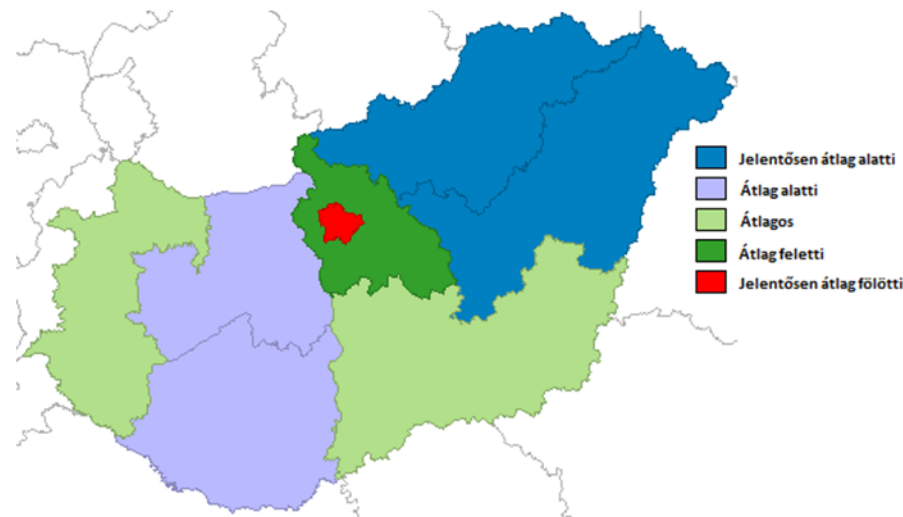
# Tapasztalatok összefoglalása és jövőbeli irányok

- Sokrétű tudáselem együttes alkalmazása
  - Kommunikációs problémák
- Adatok beszerzésének legális volta – jogi környezet ismerete
- Adatfrissítések folyamatos biztosítása – folyamatos és gyors változás
- Megfelelő algoritmusok megtalálása – folyamatos és gyors javulás
- Opensource vs. fizetős rendszerek vs. megismételhető kutatás
- Elmozdulás a vizuális elemek elemzésének irányába – folyamatos tanulás
  - Kickstarter kampányok video elemzése – ökoszisztéma szoft elemei
  - Instagram fotók elemzése – lokációk social sensor típusú vizsgálata

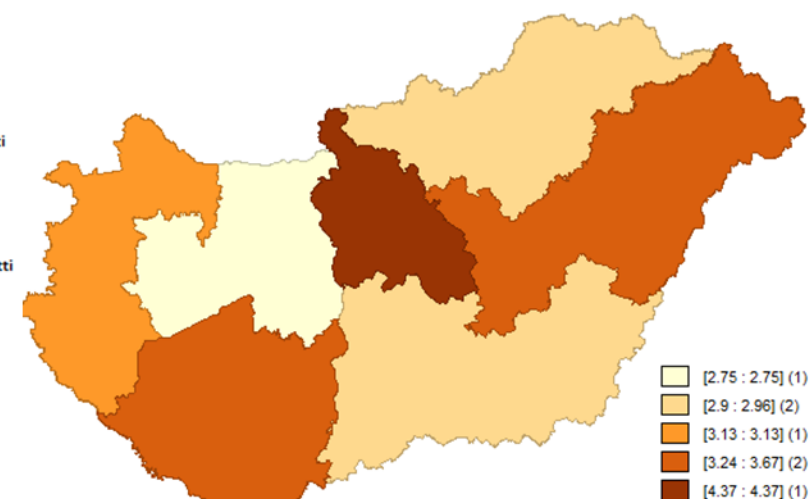
# MKKVI



# Intézményi



# KIVI



Rétegzetten reprezentatív minta belső kompetencia alrendszer versenyképességi pontjai regionális bontásban, n = 350  
(Forrás: saját szerkesztés)

Külső intézményi környezet alrendszer versenyképességi pontjai regionális bontásban, n = 175  
(Forrás: saját szerkesztés)

Rétegzetten reprezentatív, összevont, összetett rendszer versenyképességi pontjai regionális bontásban, n = 350  
(Forrás: saját szerkesztés)

## Köszönöm a figyelmüket!

hornyak.miklos@pte.hu

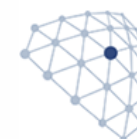


PÉCSIKÖZGÁZ  
ahonnan a karrier indul



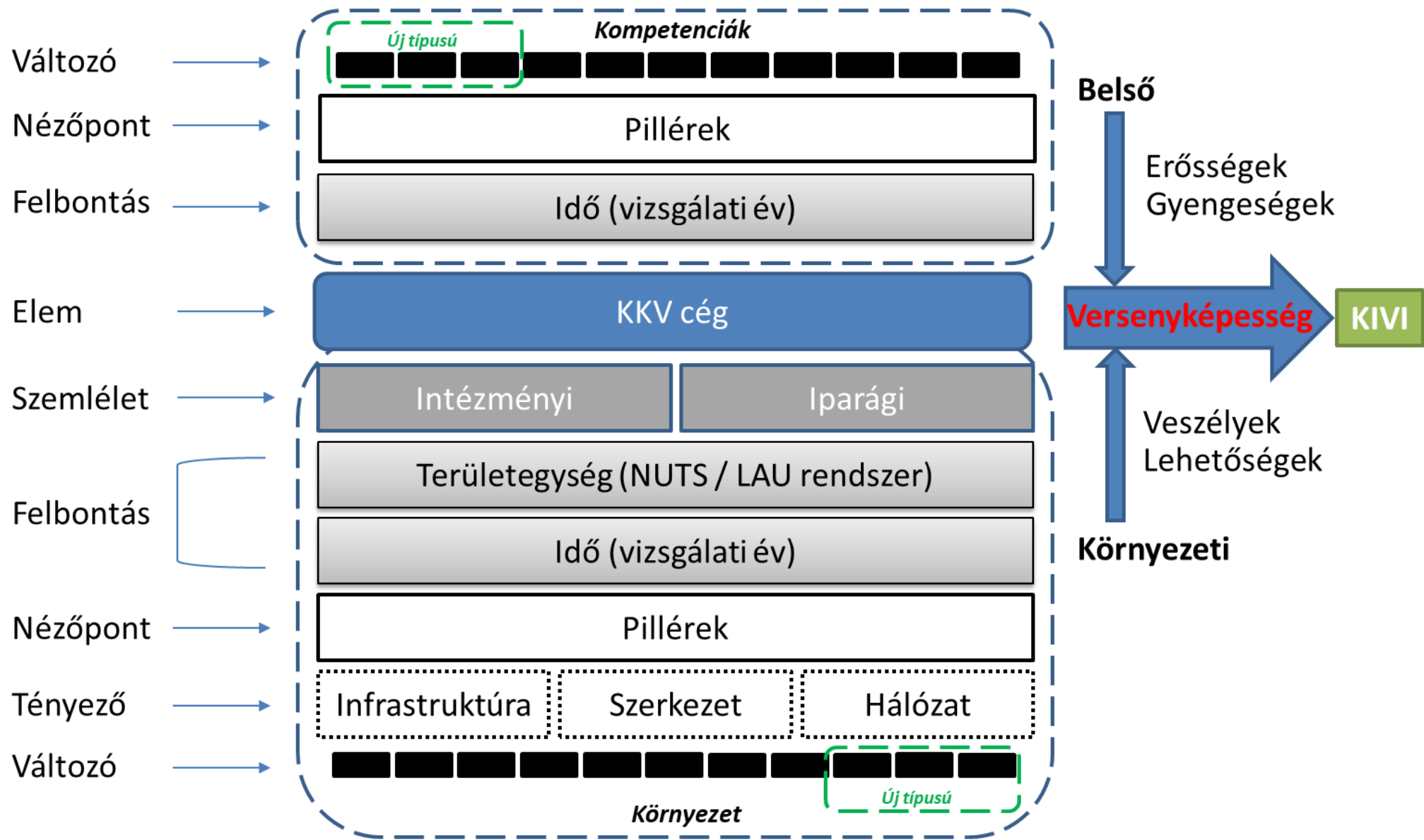
Big Data Research Group

PÉCSI TUDOMÁNYEGYETEM  
SZENTÁGOTAI JÁNOS KUTATÓKÖZPONT



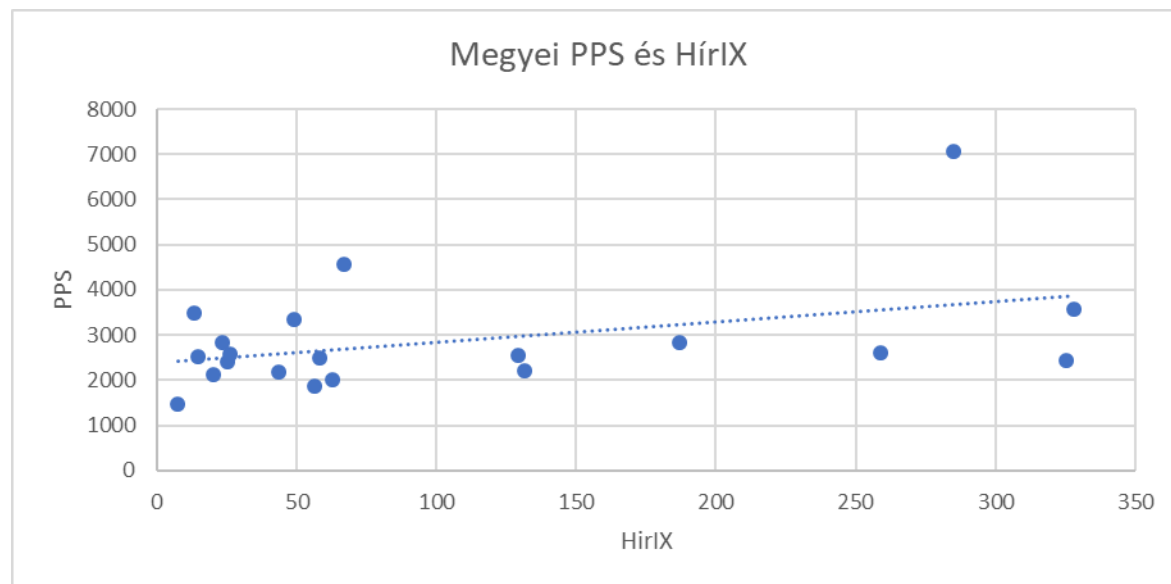
**RIERC**

Regional  
Innovation and Entrepreneurship  
Research Center



KIVI modell felépítése

# Kutatói kérdés vizsgálata – HírIX példája



- 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  $\neq 0$

- 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.



# Jövőorientációs attitűd NUTS3-as területi szemléletben

## Algoritmus 1 FOI értékek számítása

**input:**  $q_1$ ,  $q_2$ ,  $B$ ,  $L_m$

**output:** FOI érték

Google Trends adatbevitel

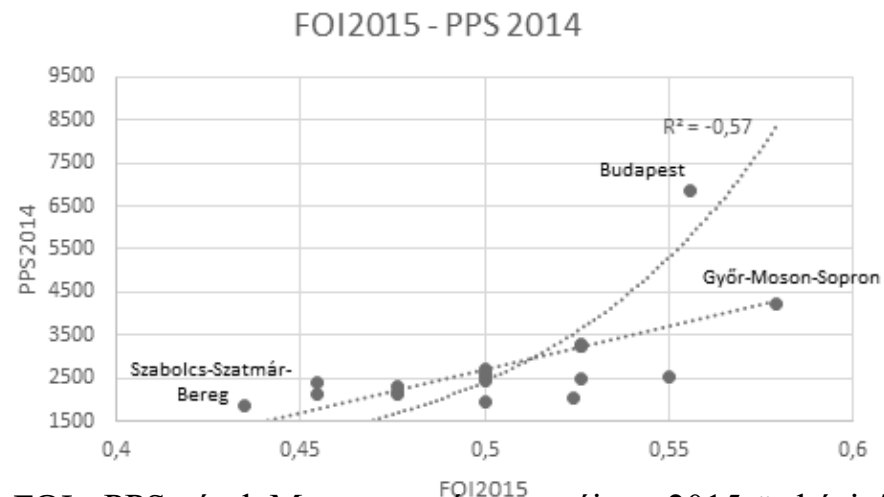
- múlt ( $q_1$ ), jövő ( $q_2$ ) irányultságot mutató keresési kulcsszó megadása  
pl. 2014, 2016
- $B$  bázisév beállítása, amelyhez időszakban indított kereséseket szűrjük  
pl. 2015
- $L_m$  keresések indítási helyének beállítása  
pl. Magyarország – Baranya megye

Google Trends keresési logok összesítése

- múlt ( $q_1^m$ ), jövő ( $q_2^m$ ) irányultságot mutató átlagos találati értékek kiolvasása  $m$  lokációnként

$FOI_B^m$  érték számítása  $B$  bázisévre  $m$  lokációnként

- $FOI_B^m = q_2^m / q_1^m$



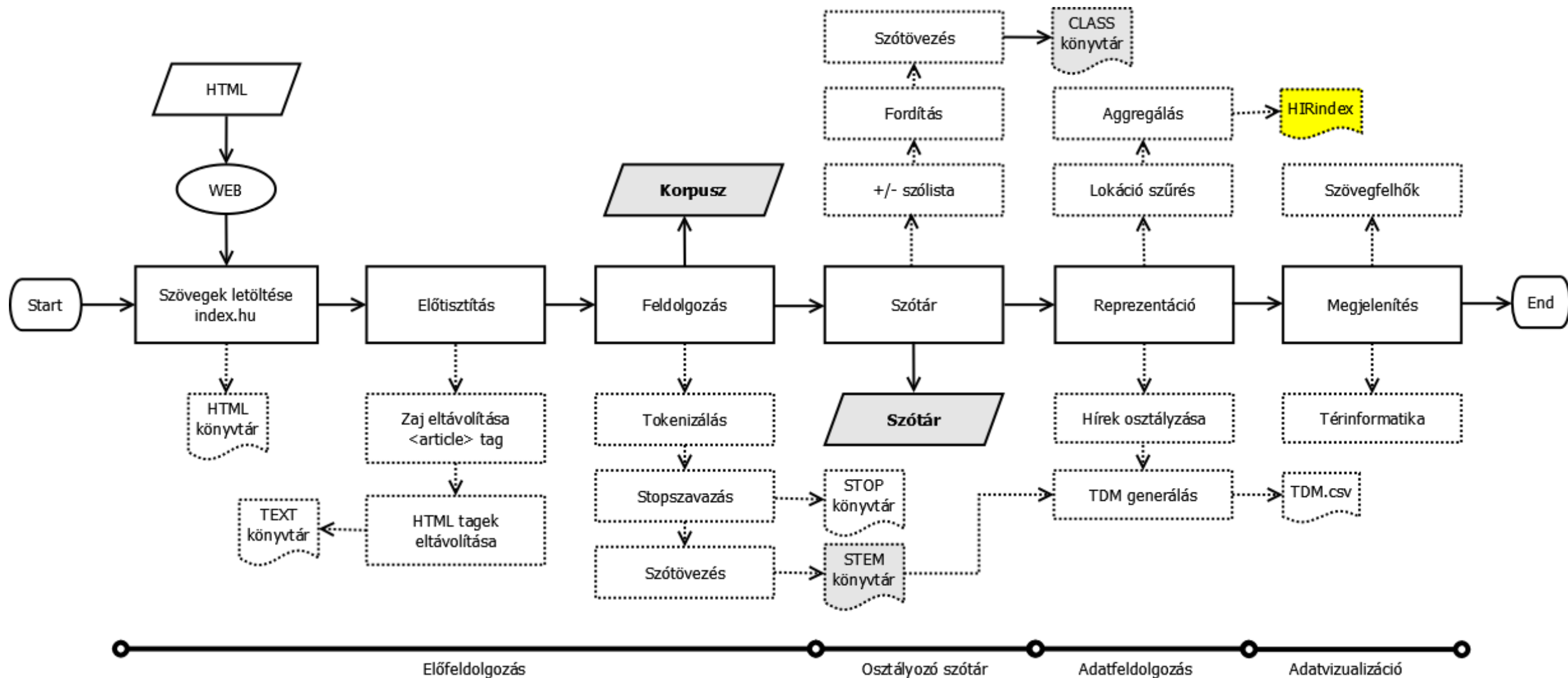
FOI - PPS párok Magyarország megyéire a 2015-ös bázisévben  
(Forrás: saját szerkesztés)

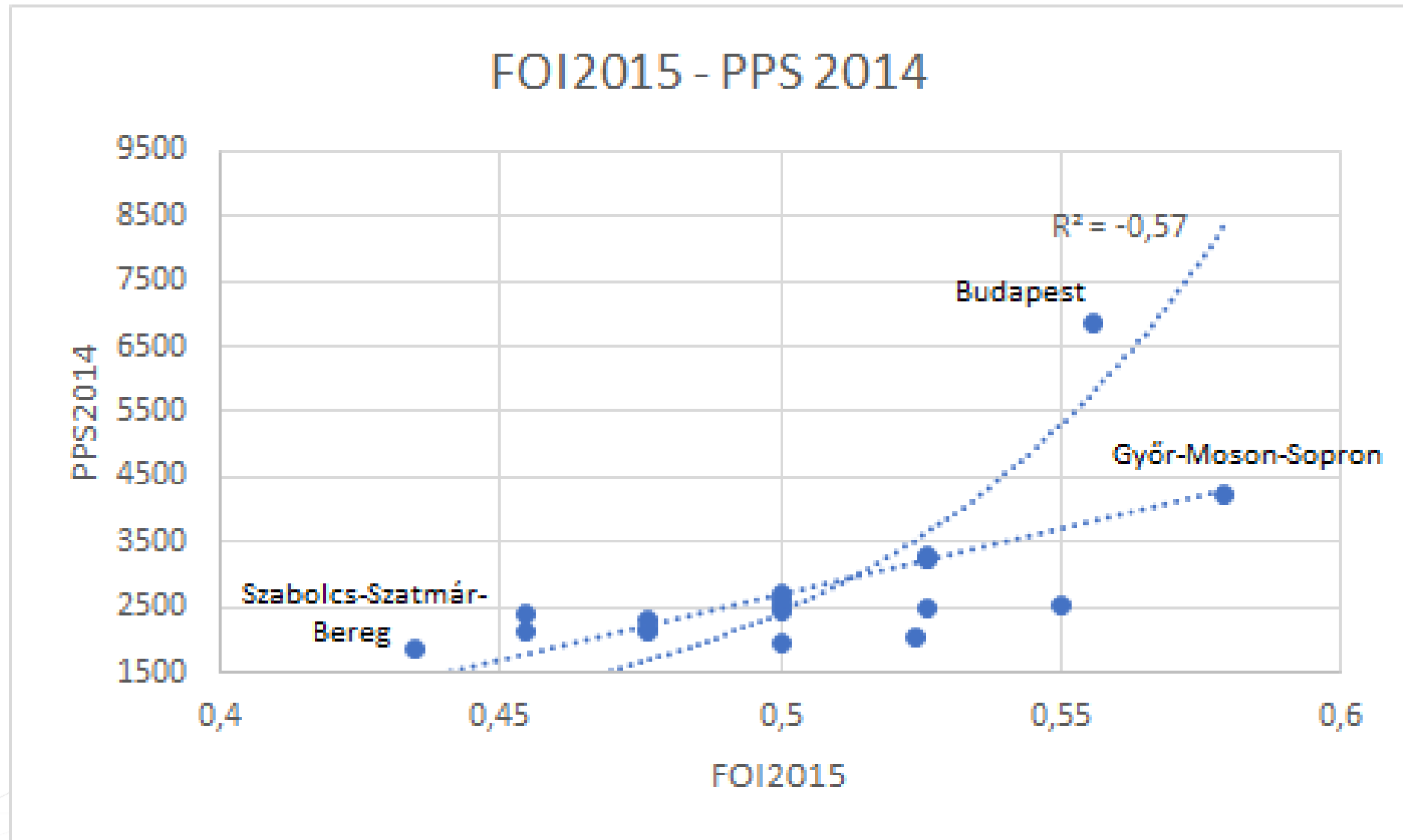
|                                                              |                     | PPS    | FOI    |
|--------------------------------------------------------------|---------------------|--------|--------|
| PPS                                                          | Pearson Correlation | 1      | ,662** |
|                                                              | Sig. (2-tailed)     |        | ,001   |
| FOI                                                          | Pearson Correlation | ,662** | 1      |
|                                                              | Sig. (2-tailed)     | ,001   |        |
| **. Correlation is significant at the 0.01 level (2-tailed). |                     |        |        |

# Irodalomjegyzék

- 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
- Hornyák Miklós - Ragadics Tamás: *Kihívások és válaszok a kiskereskedelembe – vállalkozások az Ormánságban*.  
In: Sikos T. Tamás (szerk.) *A válság hatása a kiskereskedelemre*, Szent István Egyetemi Kiadó, Gödöllő, pp. 131-145. (2013)

# HírIX folyamatábrája





# WebIX mértékek

| Mutató                | Pillér                              |
|-----------------------|-------------------------------------|
| Page Size (byte)      | Sebesség<br>(Load Time / Page Size) |
| Load Time (másodperc) |                                     |

| Mutató           | Pillér                |
|------------------|-----------------------|
| Inner Links (db) | Összetettség<br>(SUM) |
| Outer Links (db) |                       |
| Images (db)      |                       |
| Scripts (db)     |                       |
| CSS (db)         |                       |
| Forms (db)       |                       |

| Mutató                  | Pillér              |
|-------------------------|---------------------|
| SSL (0/1)               | Megjelenés<br>(SUM) |
| Page Redirect (0/1)     |                     |
| Cookie (0/1)            |                     |
| Browser Caching (0/1)   |                     |
| GAnalytics (0/1)        |                     |
| Page Title (<65)        |                     |
| Meta Description (<155) |                     |
| Mobile Viewport (0/1)   |                     |

| Mutató          | Pillér             |
|-----------------|--------------------|
| Mail (0/1)      | Kapcsolat<br>(SUM) |
| Apple (0/1)     |                    |
| GPlus (0/1)     |                    |
| Facebook (0/1)  |                    |
| Twitter (0/1)   |                    |
| Instagram (0/1) |                    |

## WebIX – 0 és 1 közé eső érték

| Pillér       | Alindex | Index |
|--------------|---------|-------|
| Sebesség     | Web 1.0 | WebIX |
| Összetettség |         |       |
| Megjelenés   |         |       |
| Kapcsolat    | Web 2.0 |       |

$$AI_{Web\ 1.0} = \frac{P_{Sebesség} + P_{Összetettség} + P_{Megjelenés}}{3}$$

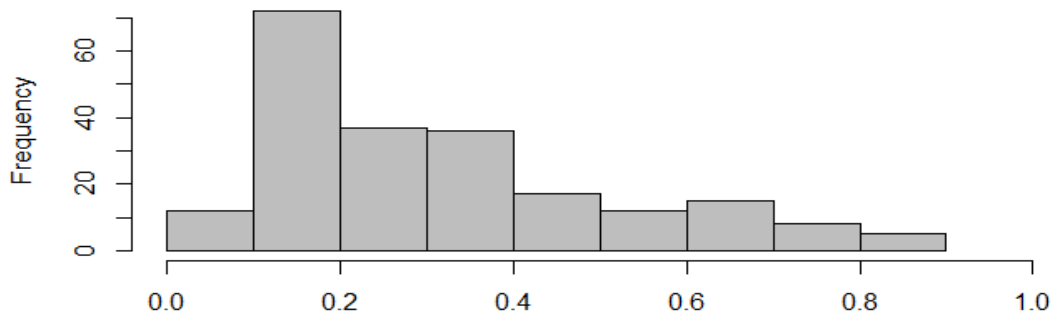
$$AI_{Web\ 2.0} = P_{Kapcsolat}$$

$$WebIX = \frac{(AI_{Web\ 1.0} + AI_{Web\ 2.0})}{2}$$



# WebIX eredmények

| Iparág | A    | B+C+D+E | F    | G+H+I | J    | L    | M+N  | R+S+T+U | Átlag |
|--------|------|---------|------|-------|------|------|------|---------|-------|
| WebIX  | 0,06 | 0,17    | 0,21 | 0,32  | 0,40 | 0,58 | 0,41 | 0,33    | 0,26  |



WebIX

WebIX hisztogramja, n = 214, forrás: saját szerkesztés

- a) 0-0,1 közti WebIX érték - cégek ~ 3 %
- b) 0,1 – 0,4 közti WebIX érték
- c) 0,4 – 0,7 közti WebIX érték
- d) 0,7 fölötti WebIX érték – cégek ~ 5%

|                    | TEÁORo8                             | Létszámkategória | Cégkor<br>2013 (év) | Megye    | Jogállás |
|--------------------|-------------------------------------|------------------|---------------------|----------|----------|
| <b>Legjobb</b>     | Műszaki vizsgálat, elemzés          | 1                | 9                   | Budapest | főváros  |
| <b>Legrosszabb</b> | Egyéb ruházati kiegészítők gyártása | 4                | 19                  | Baranya  | város    |

|              | Sebesség | Összetettség | Megjelenés | Kapcsolat |
|--------------|----------|--------------|------------|-----------|
| Sebesség     | 1        | ,103**       | ,059       | ,213**    |
| Összetettség | ,103**   | 1            | ,267**     | ,478**    |
| Megjelenés   | ,059     | ,267**       | 1          | ,213**    |
| Kapcsolat    | ,213**   | ,478**       | ,213**     | 1         |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

WebIX pillérértékek Pearson-féle korrelációs mátrixa, n=214, forrás: saját számítás

|                     | NUTS2  |        |        |        |        |        |        |
|---------------------|--------|--------|--------|--------|--------|--------|--------|
|                     | 1      | 2      | 3      | 4      | 5      | 6      | 7      |
| <b>Web aránya</b>   | 70,73% | 71,85% | 72,13% | 68,52% | 73,85% | 68,29% | 81,42% |
| <b>WebIX átlaga</b> | 0,264  | 0,222  | 0,268  | 0,261  | 0,241  | 0,224  | 0,246  |