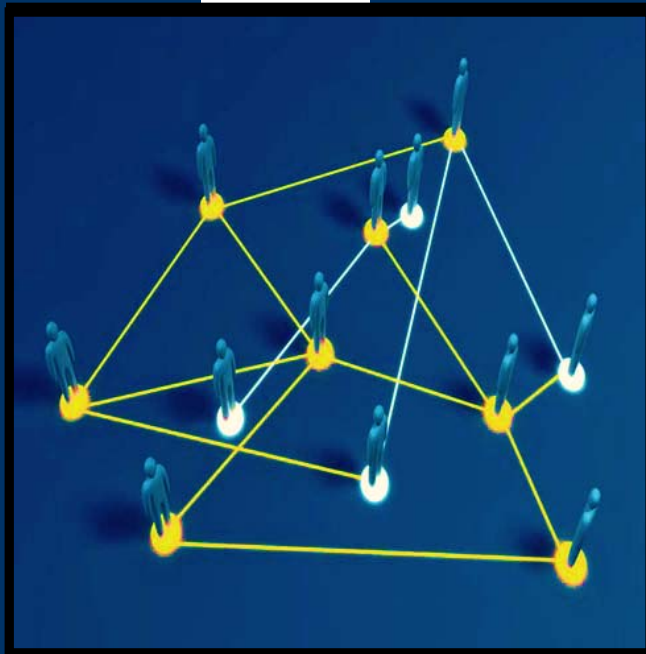


Jakobi Ákos
(ELTE)

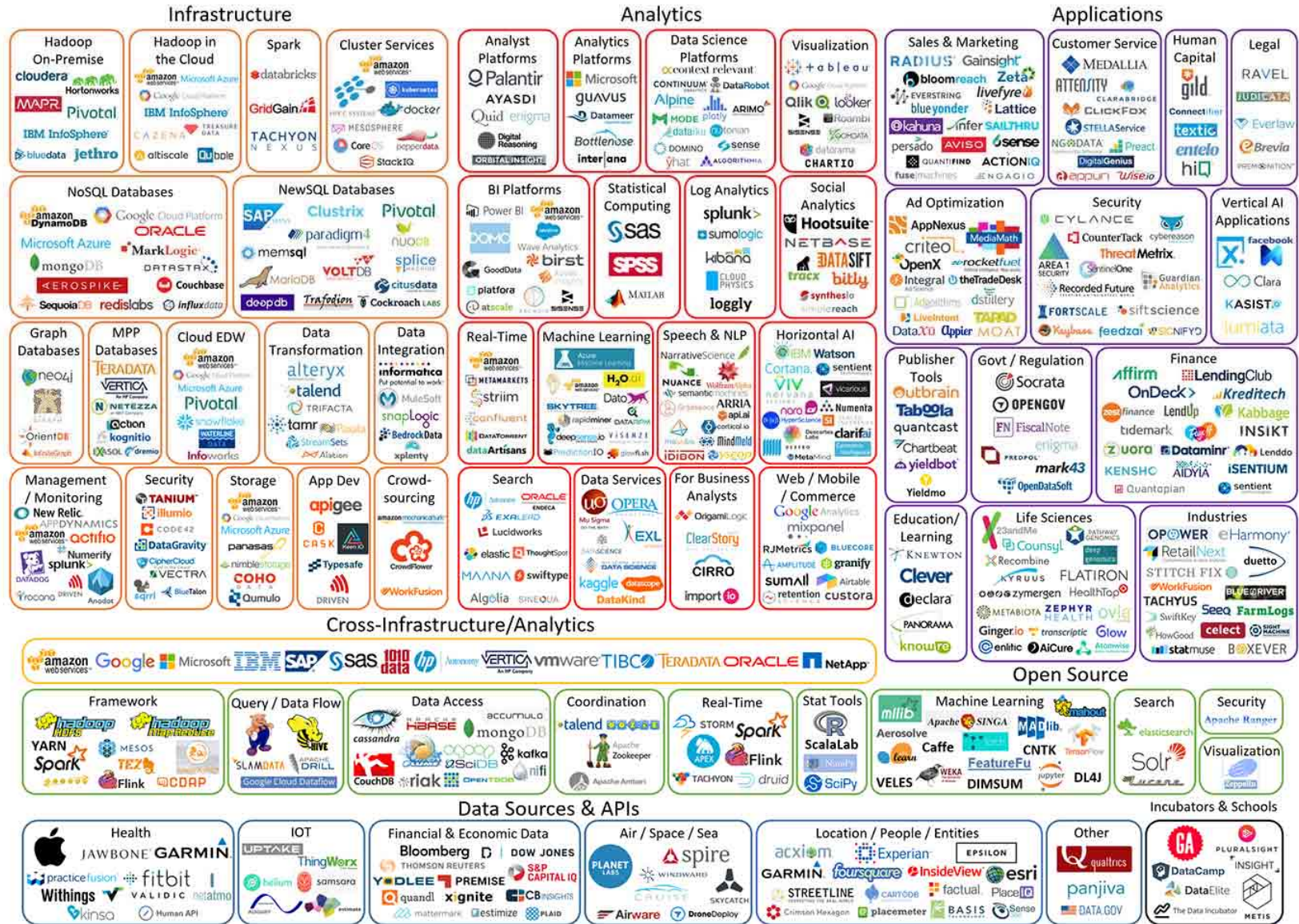


Big Spatial Data: lehetőségek és tapasztalatok

2018. február 27.

„Big Data a területi kutatásokban” konferencia, Budapest

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/6/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

Néhány gondolat a big data jelenségről

- A big data eleve digitálisan keletkezik, passzív módon termelődik a mindennapi digitális interakciók melléktermékeként
- A sokáig csak virtuális melléktermékként számon tartott napi információhalom épp akkor válik értékessé, amikor a különböző adatokat sikerül összekötni, köztük összefüggéseket, felismerhető mintázatokat találni, s mindebből értékelhető következtetéseket levonni.
- A társadalom kutatói számára valóságos aranybánya (vásárlási, munkabajárási, közlekedési és egyéb szokásaink adatai)
- Betekintést nyújt az emberi viselkedés egyedi és társadalmi szintjeibe

Big data kihívások



- Magánélet és a „surveillance society” (a Facebook nincs ingyen!)
- Adathozzáférés (piaci szereplők)
- Adatgyűjtési technológia (webscraping, webharvesting, crawler robotok)
- Mit mond nekünk valójában az adott adatforrás? (representativitás)
- Nincs előre megfogalmazott statisztikus elemzői cél, csak utólag az adatállományból találjuk azt ki
- „big data = big errors” (költséges adattisztítás)

A területi kutatások lehetőségei



- „Social (human) sensor” térbeli analitika
- Térhasználati információk (hägerstrandi térpályák)
- A közzétett tartalom területi statisztikái (lokalizált webes tartalmak)
- Az online jelenlét területi statisztikái
- Területi tranzakció-statisztikák
- ...

Big data források

Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja szerint

- Automatikus rendszerek adatai (gépek által közvetített adatok, IoT), avagy „machine to machine” típusú adatok
- Alapvetően a fizikai világ megfigyeléséből származnak és legtöbb esetben jól strukturáltak, ám méretük és keletkezési sebességük a tradicionális megközelítéseken messze túlmutat.

Big data források

Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja szerint

1. Szenzor adatok

11. Rögzített szenzorok

- 111. Otthon-automatizálás
- 112. Időjárási, szennyezési szenzorok
- 113. Közlekezési szenzorok, webkamerák
- 114. Tudományos célú szenzorok
- 115. Biztonsági videók, képek

12. Mobil szenzorok (nyomkövetés)

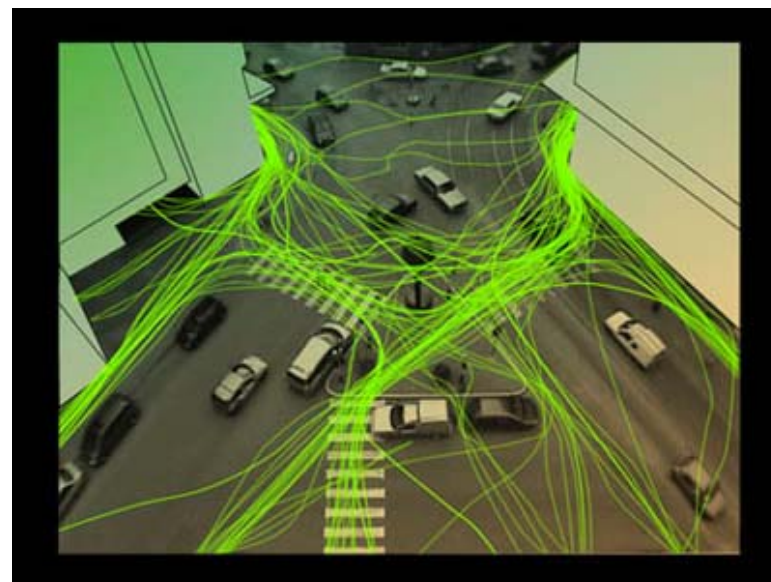
- 121. Személyes (mobil helymeghatározás)
- 122. Közúti (autók, teherszállítás)
- 123. Vasúti (vonatok)
- 124. Légi (repülőgépek)
- 125. Vízi (hajók)

13. Műholdas adatok

- 131. Topográfiai
- 132. Hőmérsékleti
- 133. Megfigyelési
- 134. Meteorológiai
- 135. Egyéb

2. Számítógépes rendszerekből származó adatok

- 210. Naplók (log-ok)
- 220. Webes naplók (web log-ok)



**Városi
térpályák**

Big data források

Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja szerint

- Hagyományos üzleti és hivatali rendszerek adatai (folyamat által közvetített adatok), avagy „people to machine” típusú adatok
- Az adatokat határozottabb strukturáltság jellemzi, kapcsolati táblákkal, metaadatokkal
 - 1. Közhivatalok által szolgáltatott adatok
 - 11. Orvosi/egészségügyi nyilvántartások
 - 2. Kereskedelem által létrehozott adatok
 - 21. Kereskedelmi tranzakciók
 - 22. Bank-/készletnyilvántartás
 - 23. E-kereskedelem
 - 24. Bankkártya/hitelkártya

Banki ügyfelek tranzakciós aktivitása



Hol
tranzaktálnak
adott bank
ügyfelei?

Big data források

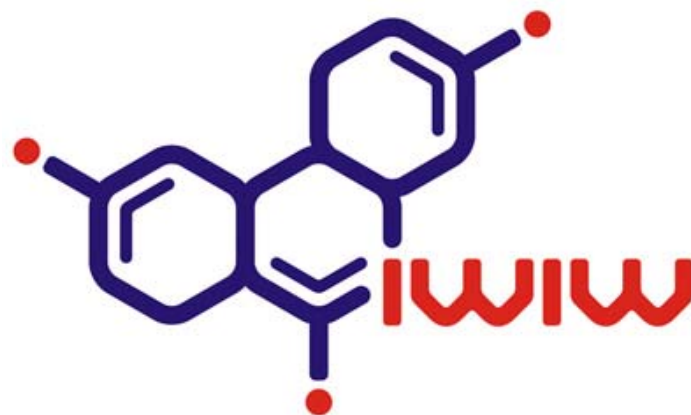
Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja szerint

- Közösségi adatok (humán-eredetű információk), avagy „people to people” típusú adatok.
- Ezek az adatok csak lazán strukturáltak és gyakran irányítási kontrol nélküliek.

- 1. Közösségi hálók: Facebook, Twitter, Tumblr stb.
- 2. Blogok, hozzászólások
- 3. Személyes dokumentumok
- 4. Képek: Instagram, Flickr, Picasa stb.
- 5. Videók: Youtube stb.
- 6. Internetes keresések
- 7. Mobil adattartalom: szöveges üzenetek
- 8. Felhasználó által generált térképek
- 9. E-Mail



WIKIPEDIA
The Free Encyclopedia





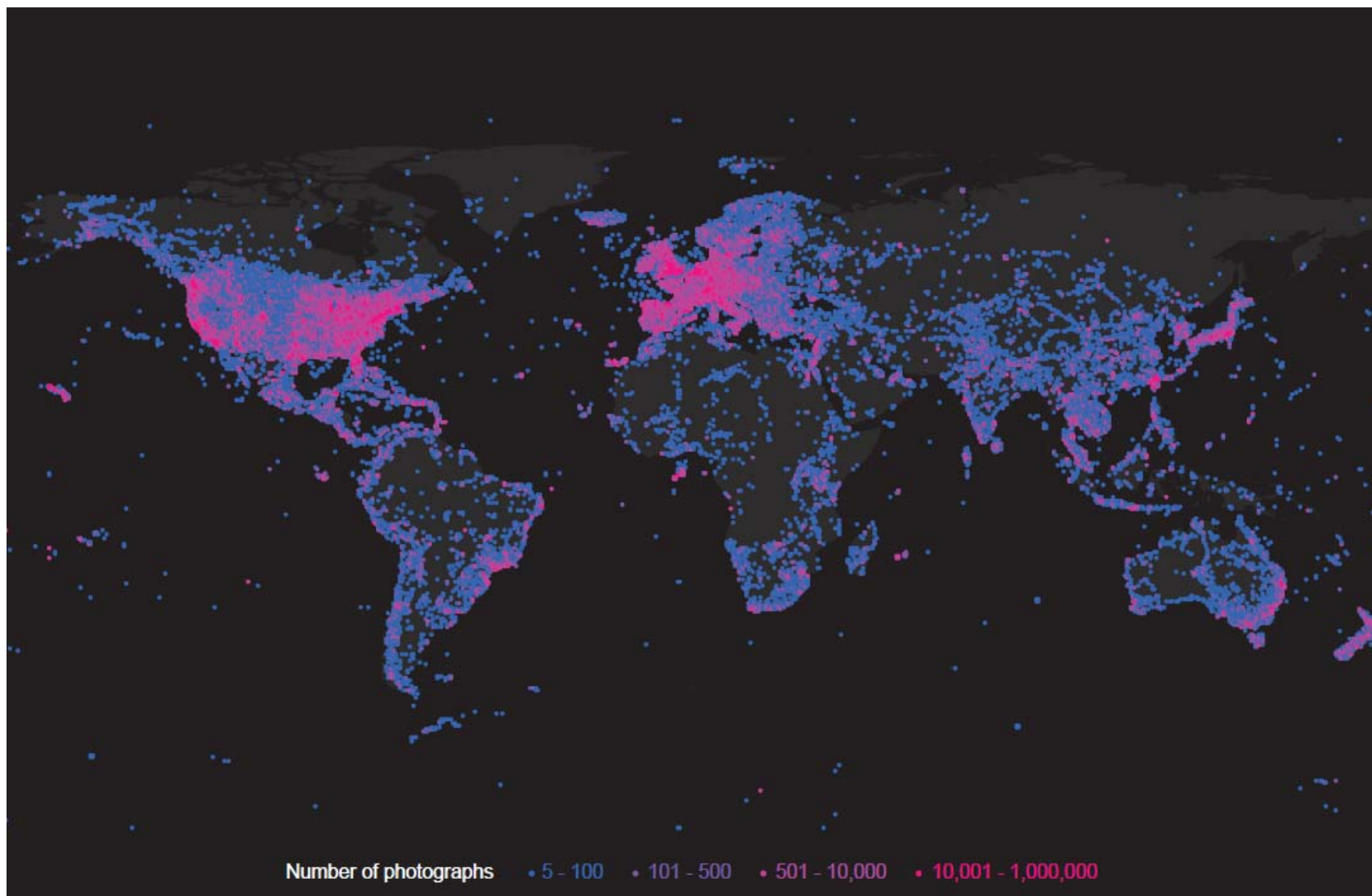
Paul Butler (2010): Visualizing friendships





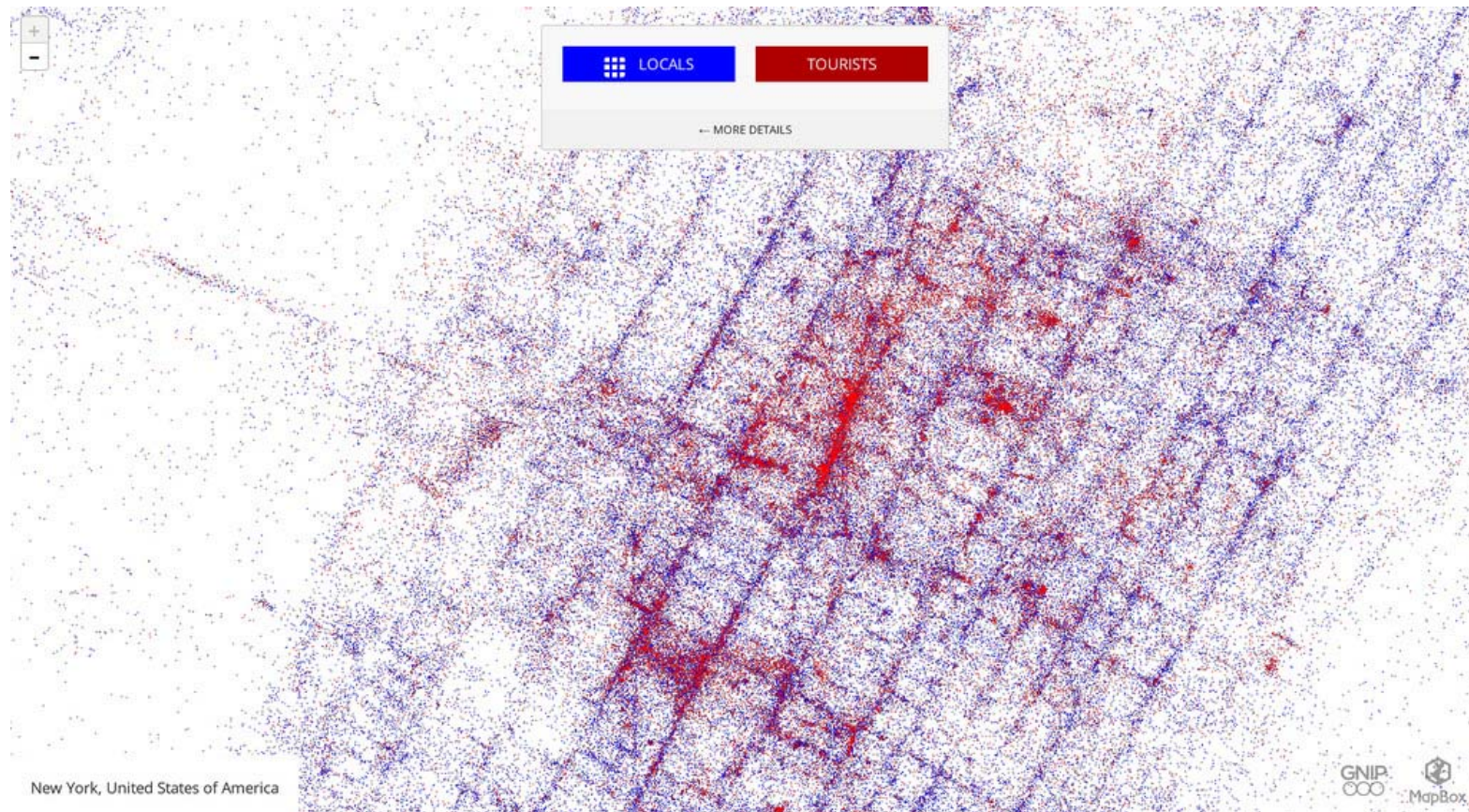


Geotag-elt Flickr fotó bejegyzések száma



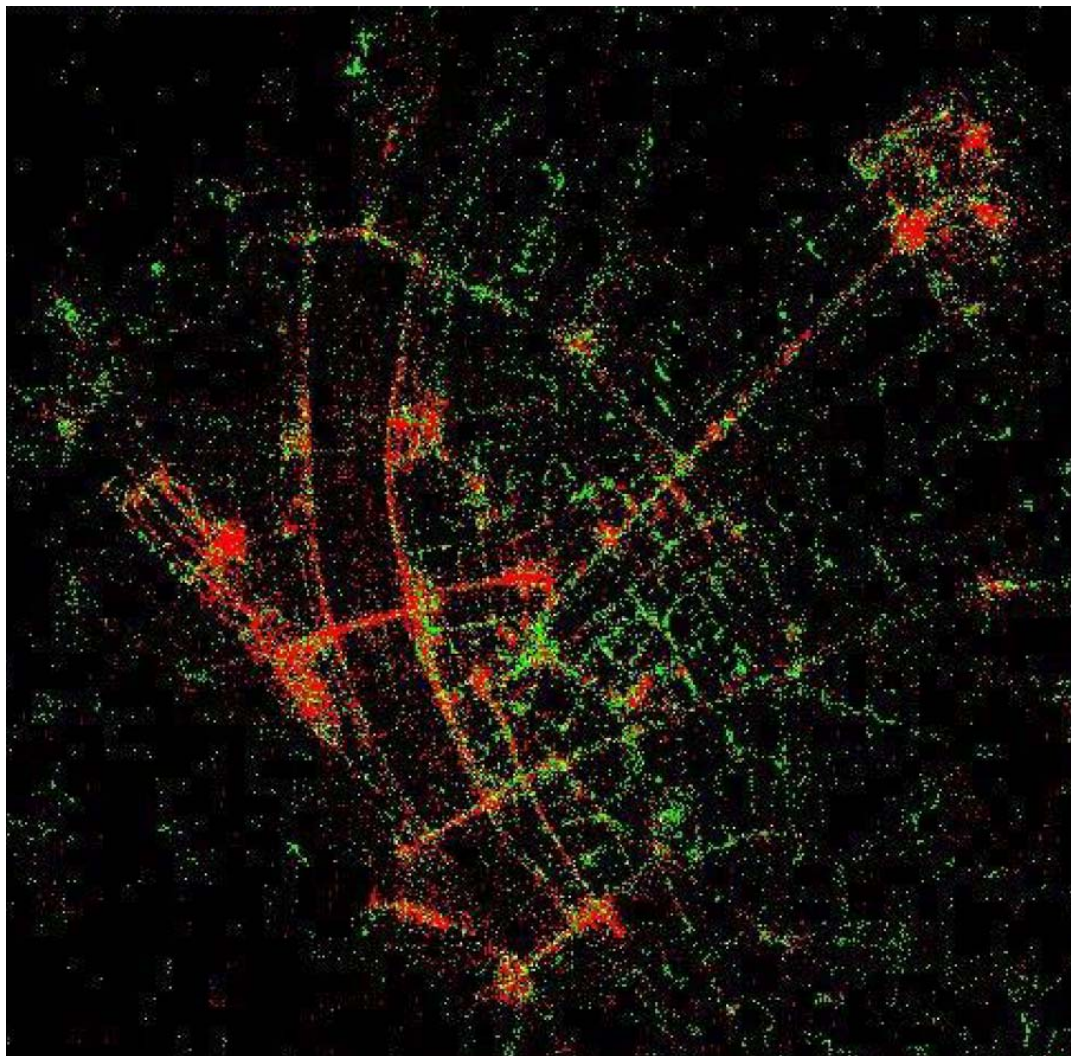
Forrás: www.convoco.co.uk

Helyi és nem helyi lakosok (turisták) Flickr fotó-bejegyzéseinek sűrűsége Manhattan-ben



Forrás: Sokol Z., 2013

Helyi és nem helyi lakosok (turisták) Flickr fotó-bejegyzéseinek sűrűsége Budapesten



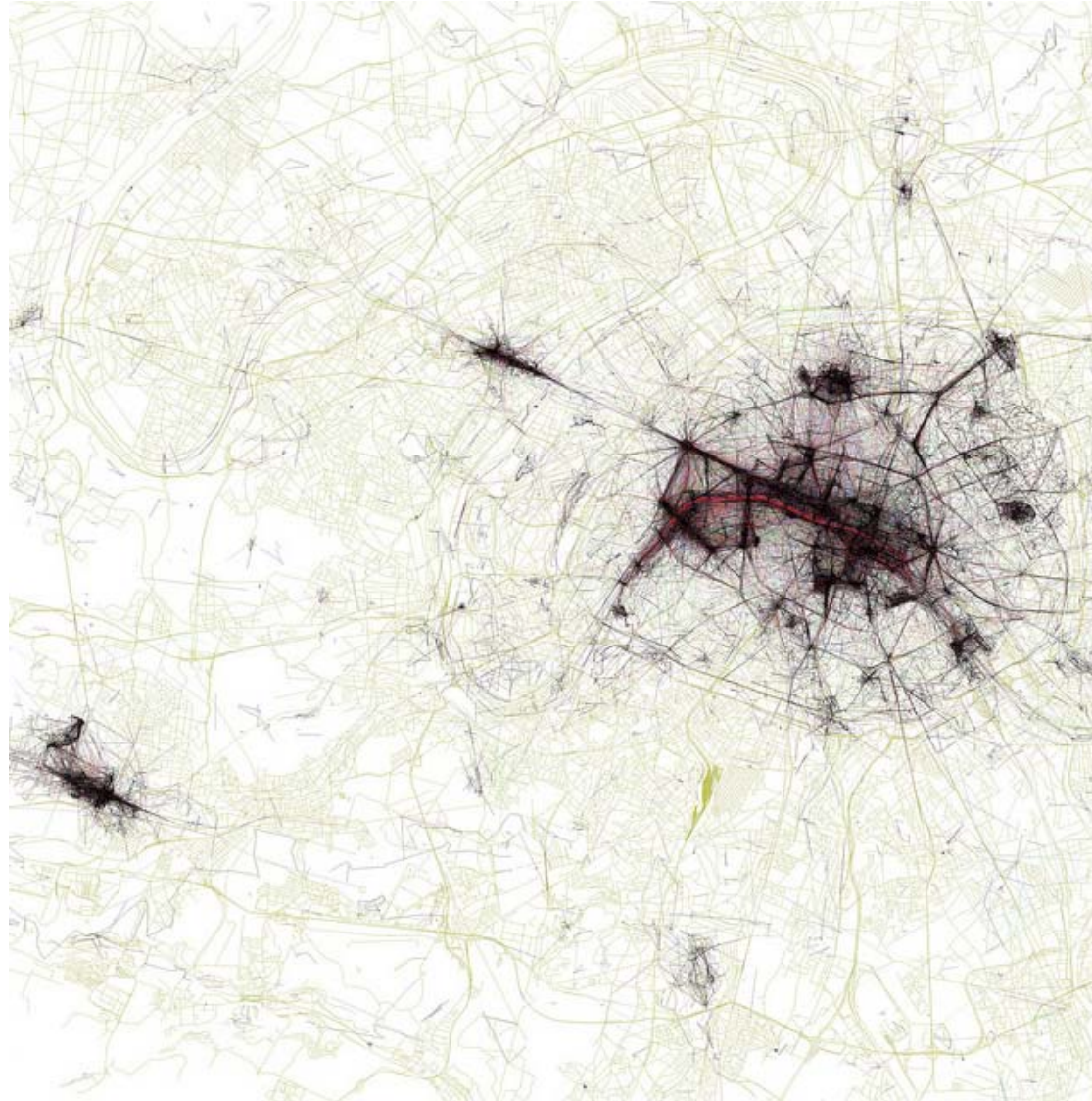
Forrás:
Kádár B. – Gede M. 2012

Csoportos térpályák Flickr fotóadatok alapján

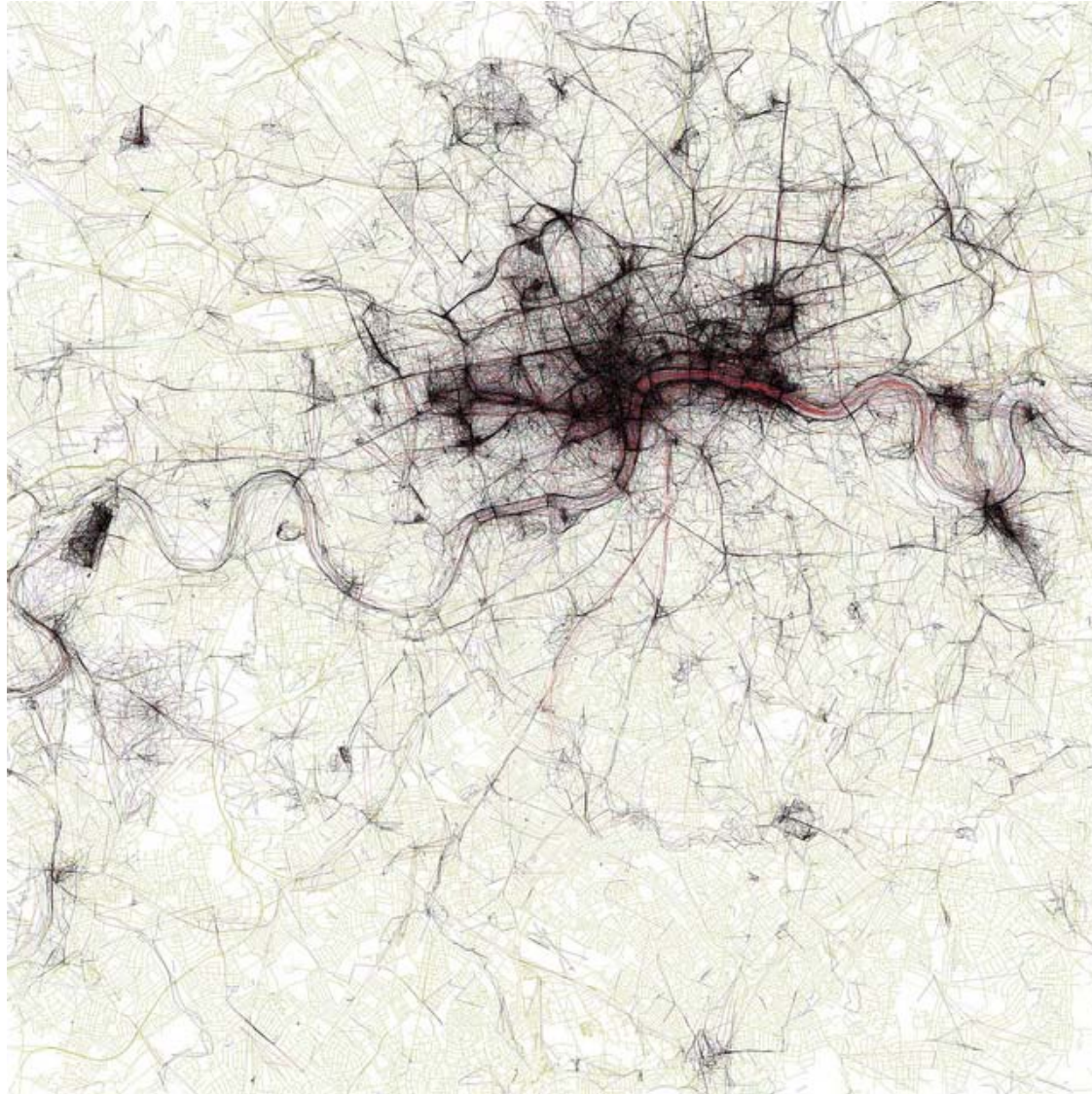


Fotókat készítő turisták
mozgási útvonalainak területi mintázata
Manhattan déli részén
(Crandall et. al., 2009)

Párizs

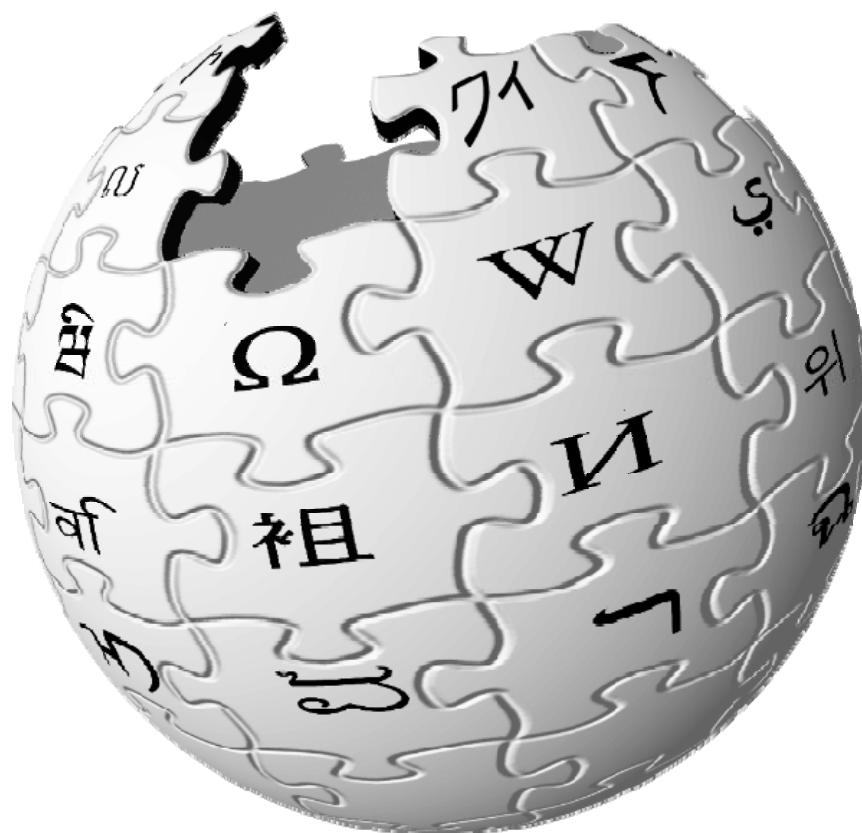


London



Budapest





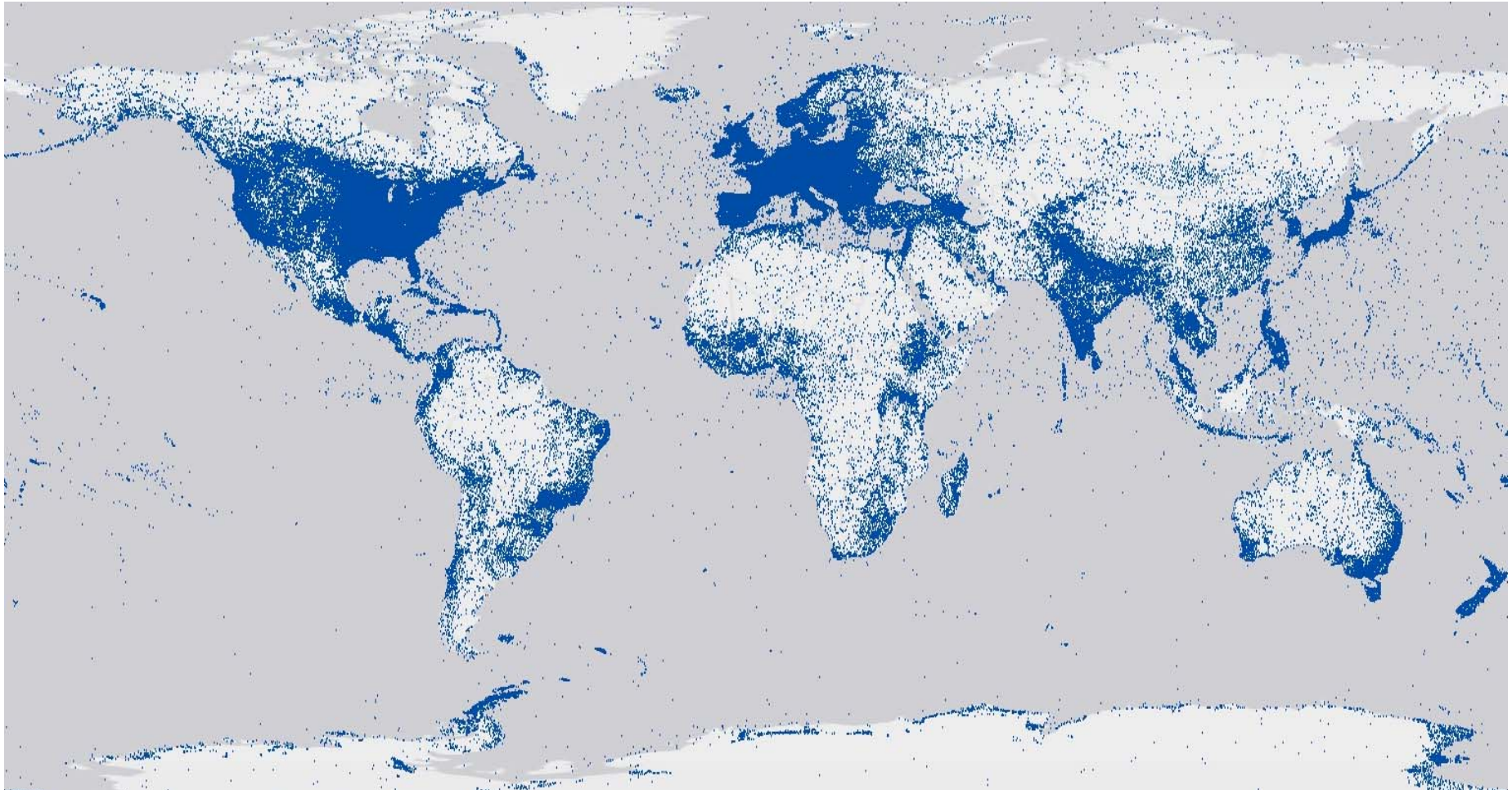
WIKIPEDIA
The Free Encyclopedia

Tartalomanalitika



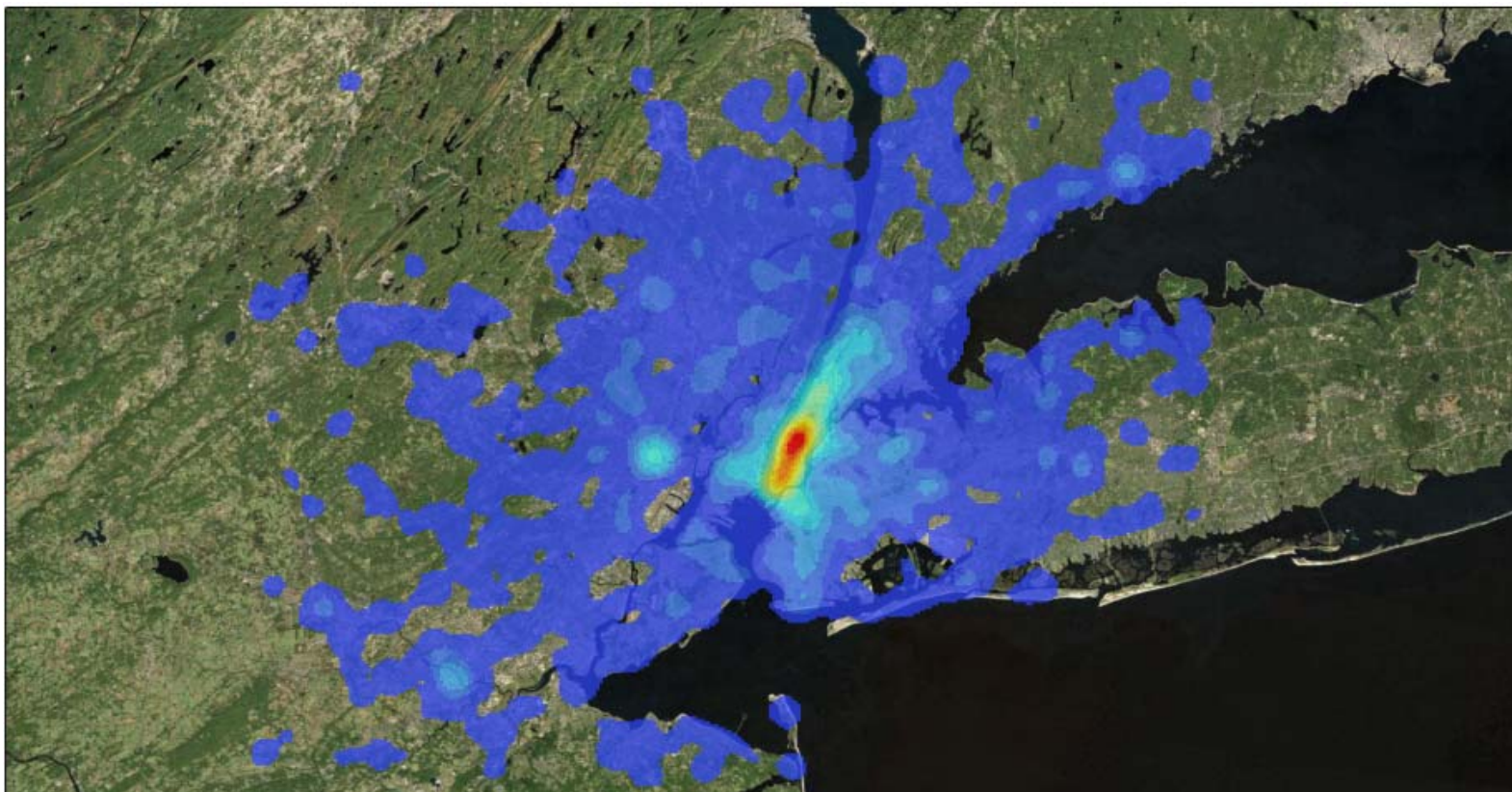
- Kik és hol tesznek közzé információkat az internetes világban
- Miről, avagy mely helyekről közölnek tipikusan információkat
- Új lehetőség az online tartalmak szövegbányászati módszerekkel történő feldolgozása
 - számszerűsített módon is meghatározható lehet egyes helyek és terek online reprezentációja
 - megbecsülhető az adott helyhez kötődő tudattartalmak minősége vagy nagysága

Geotagelt Wikipedia bejegyzések sűrűsége a világban



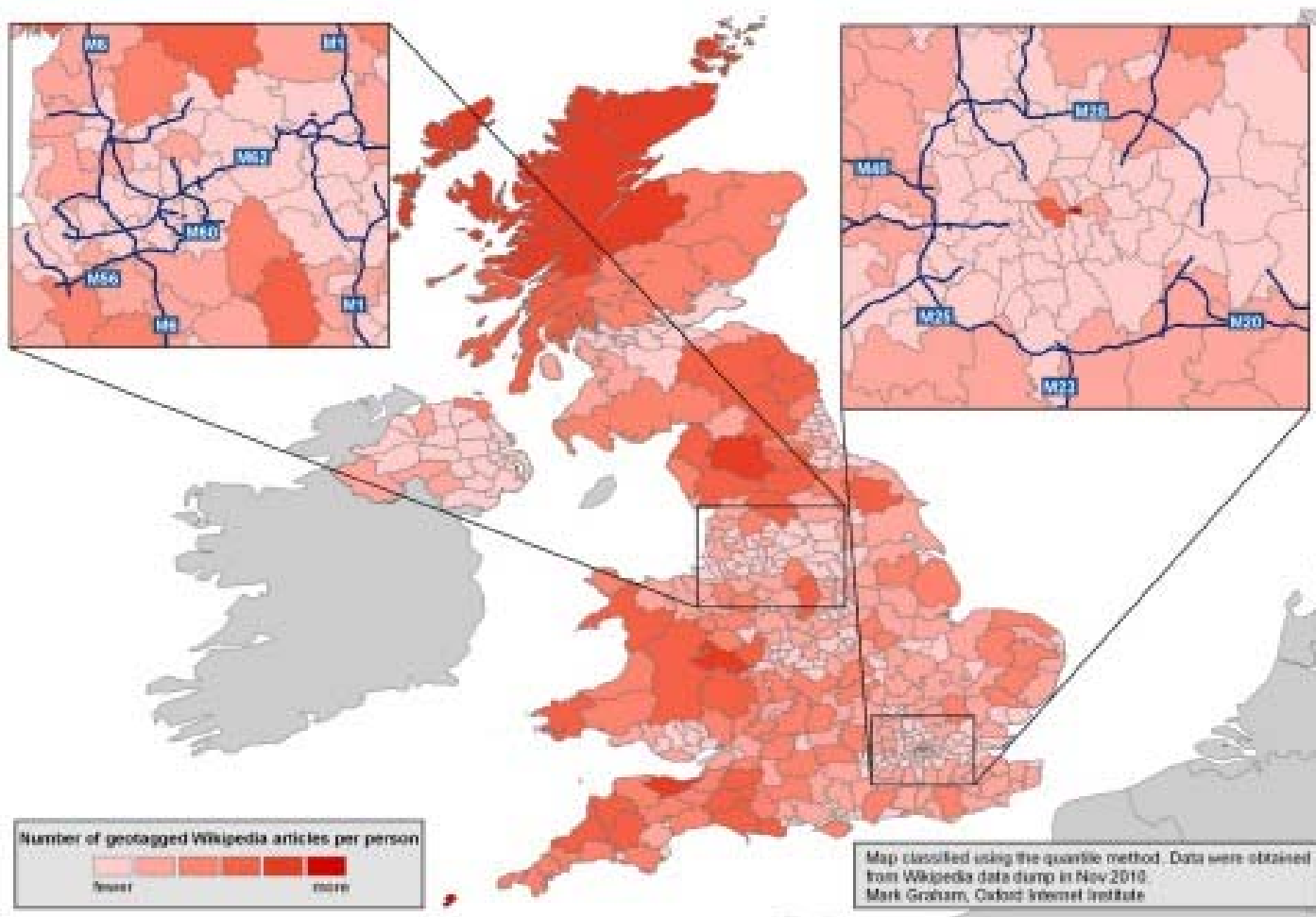
Forrás: DBpedia adatok alapján a saját szerkesztés

Sűrűségi „hőterképek”:
Geotag-elt Wikipedia bejegyzések sűrűsödése New York körzetében



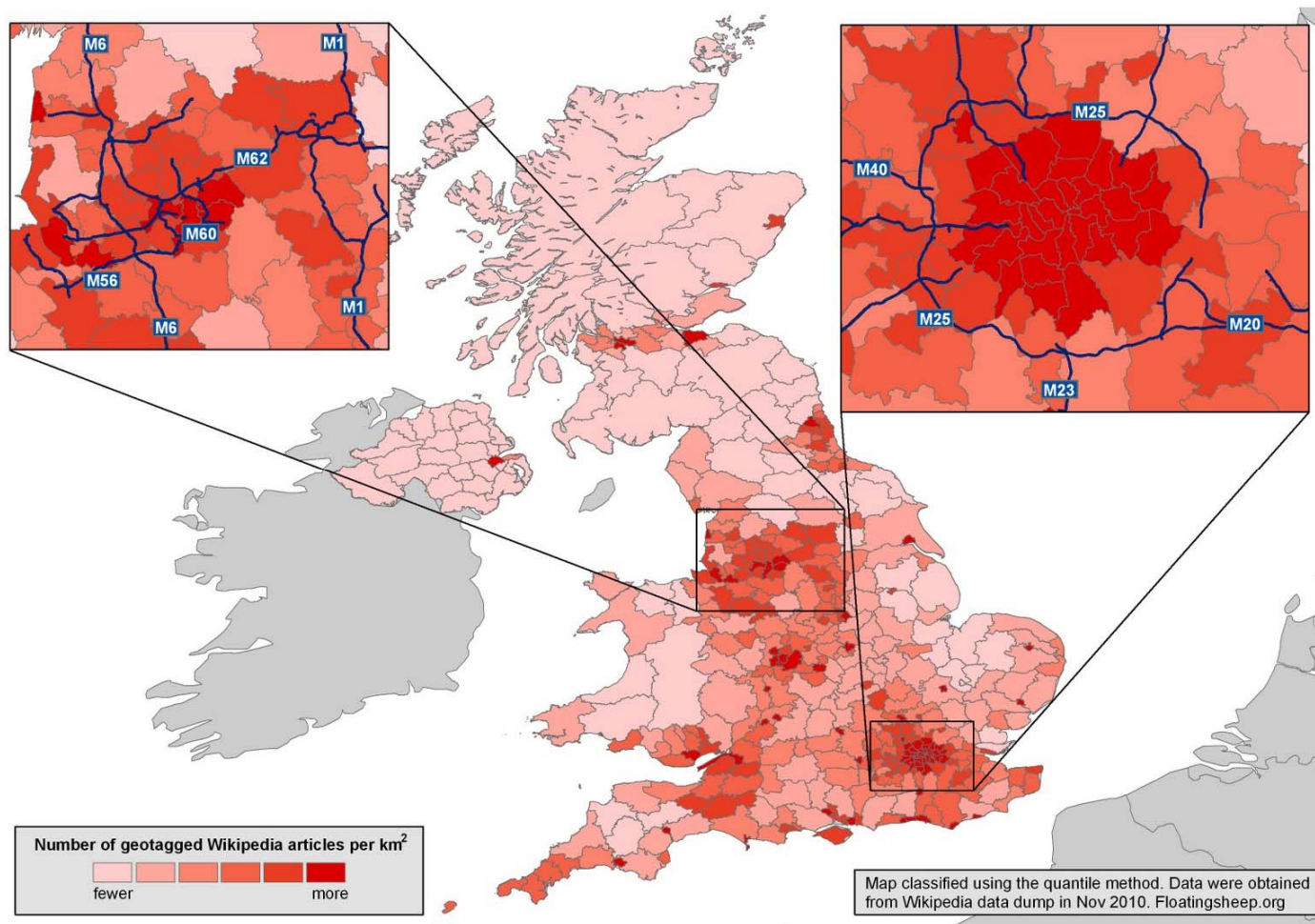
Forrás: Meeks E., 2011

Egy lakosra jutó geotag-elt Wikipedia bejegyzések Nagy-Britanniában



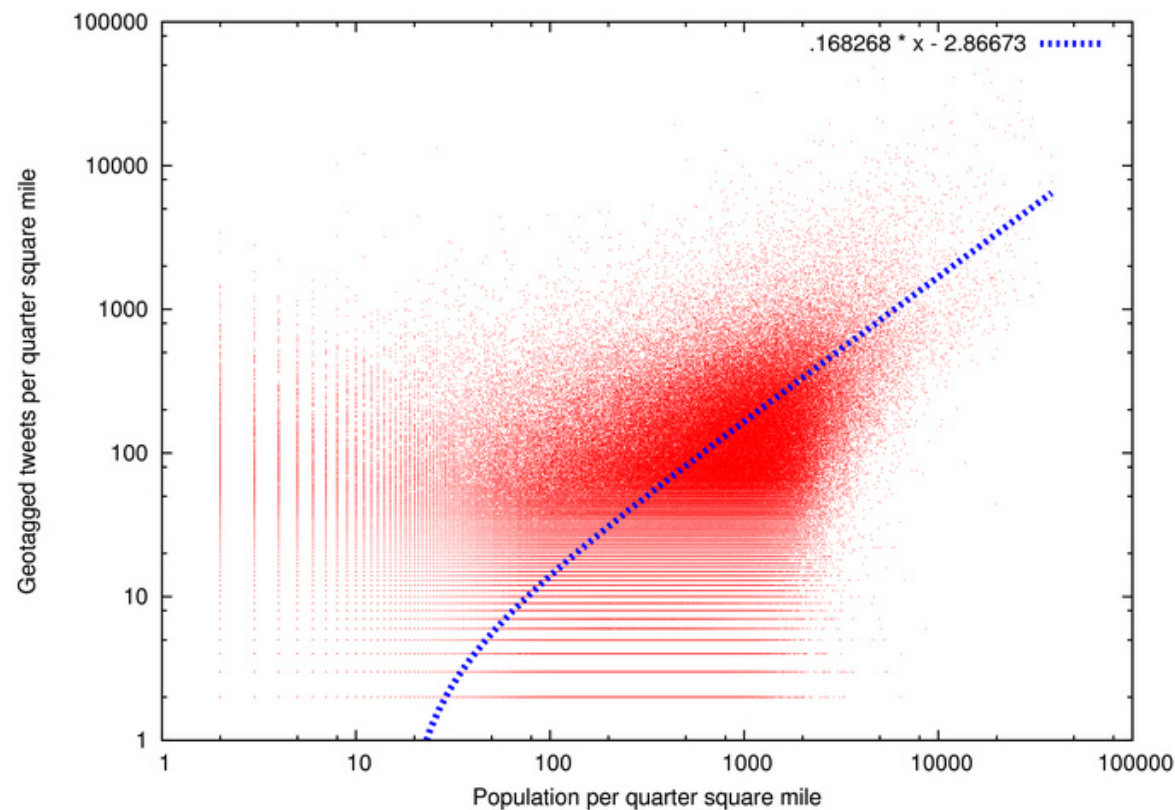
Forrás: Graham M., 2010

Egy km²-re jutó geotag-elt Wikipedia bejegyzések Nagy-Britanniában



Forrás: Graham M., 2010

A népsűrűség és a geotag-elt bejegyzések közötti kapcsolat alakulása



**A népességszám és a geotag-elt bejegyzések számának pontdiagramja
(0,5 négyzetmérföldenként, az USA-ban)
(Forrás: Fischer, E. 2012)**

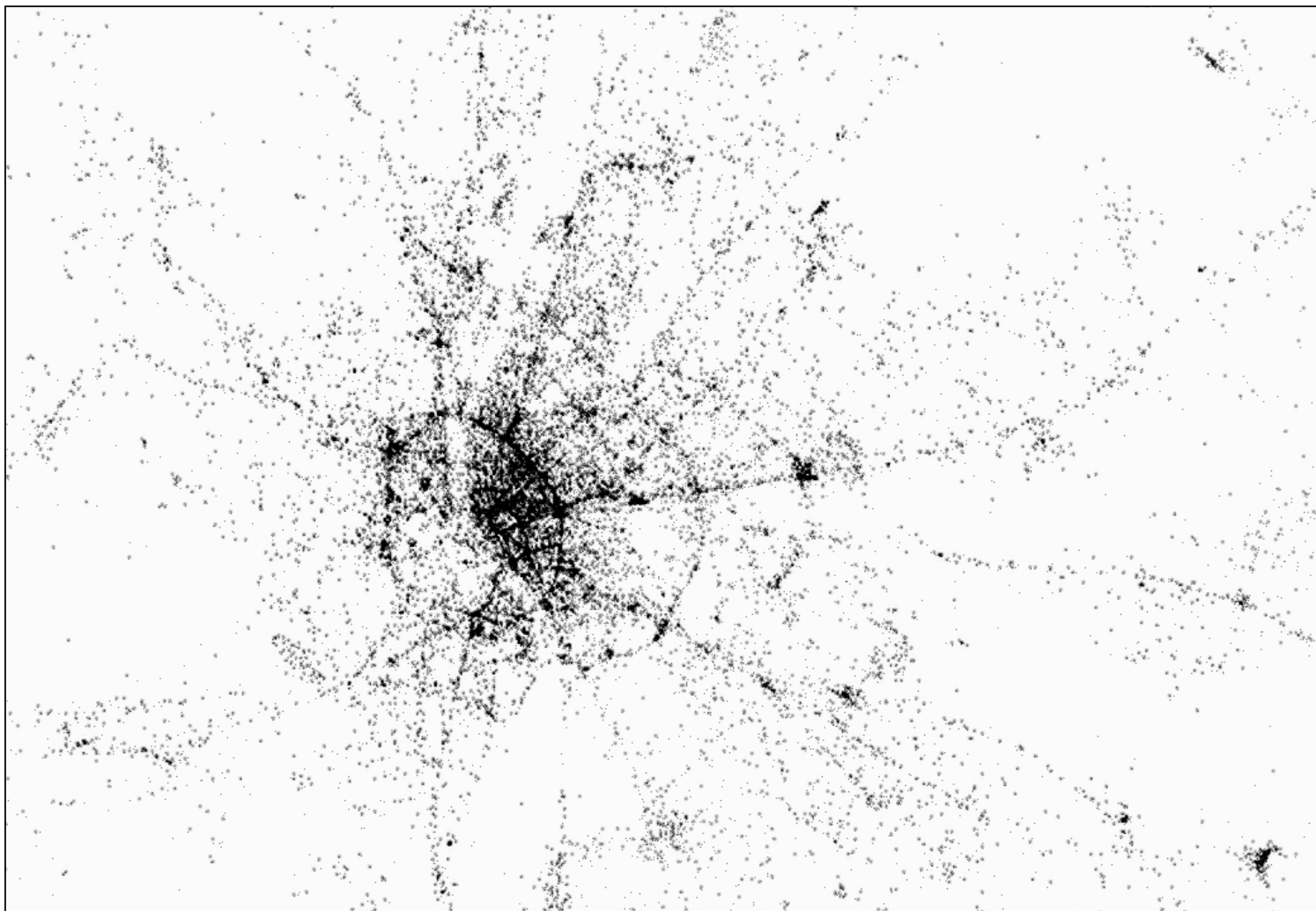
Lakatlan területeken is találhatók geotag-elt bejegyzések (pl. parkok, emlékművek stb.).



twitter

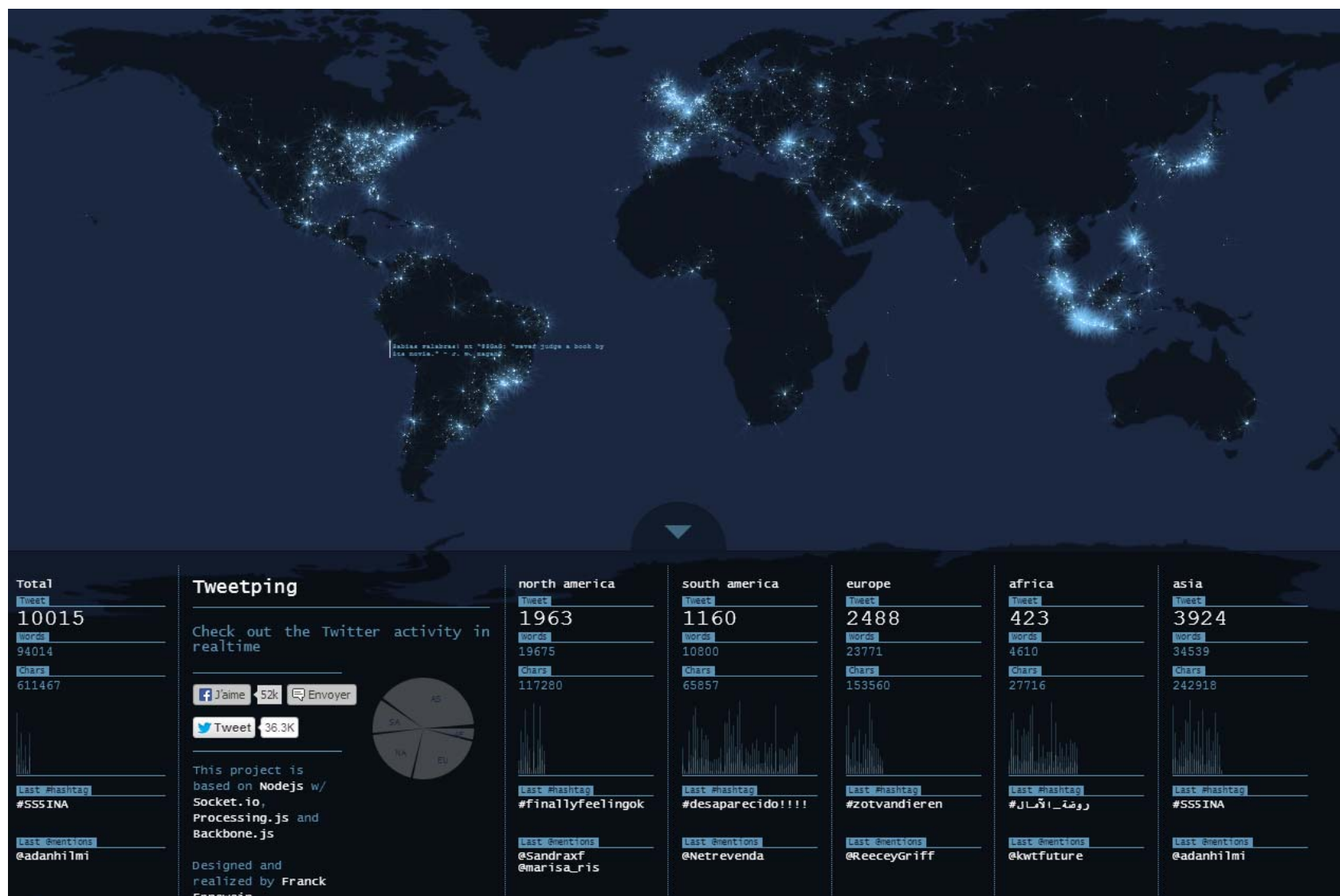
Twitter Streaming API

null	1423154659.33	DENISS PASHKEVICH QUARTET \ / ARAMAIC MAN \ / BUDAPEST OPUS JAZZ CLUB by Deniss Pashkevich http://t.co/xj5K4fVzPU
null	1423154659.95	Great Britain fall to impressive Turkey in Fed Cup: A shock 2-1 defeat to Turkey in Budapest left Great Britai... http://t.co/HddJSPlkxZ
null	1423154695.27	We take you inside St. Stephen's Basilica in #Budapest http://t.co/oWMHJK8VRi @VikingRiver #travel #boomers
null	1423154718.0	Ho comprato i biglietti per Budapest che, compresa l'iscrizione al MUN, mi costeranno 160\u20ac. Da oggi sono in regime di Fiscal Compact.
null	1423154735.16	Ya se estren\u00f3 en M\u00e9xico The Grand Budapest Hotel? tengo que verla! Por qu\u00e9 no se nada? #WesAnderson #LorenaG
[47.479262,19.067871]}	1423154735.65	I'm at Kaeng Som Tom Yum - Thai Takeaway in Budapest https://t.co/217GJd02oL
[47.499047,19.059994]}	1423154739.03	Usher, cider and budapest.
null	1423154744.65	Packing time..ready for Budapest!\nCheyenne+ @OneInkSeven @Intenzeink @HustleButter @H2Ocean @KillerInkTattoo http://t.co/qvJNLs00Un
null	1423154748.15	Budapest is a must \ud83d\ude4a
null	1423154774.52	Video: OneRepublic - Budapest (George Ezra Cover)\u00a0 http://t.co/KV8vyGpHRL
null	1423154775.27	RT @BestMovieLine: The Grand Budapest Hotel http://t.co/EO5v3zHaJp
[47.552201,19.045819]}	1423155187.3	I'm at Filatorig\u00e9t (H5) - @bkkbudapest in Budapest https://t.co/81RdAVwWxS
null	1423154792.4	Gotta leave Budapest a day early man allow



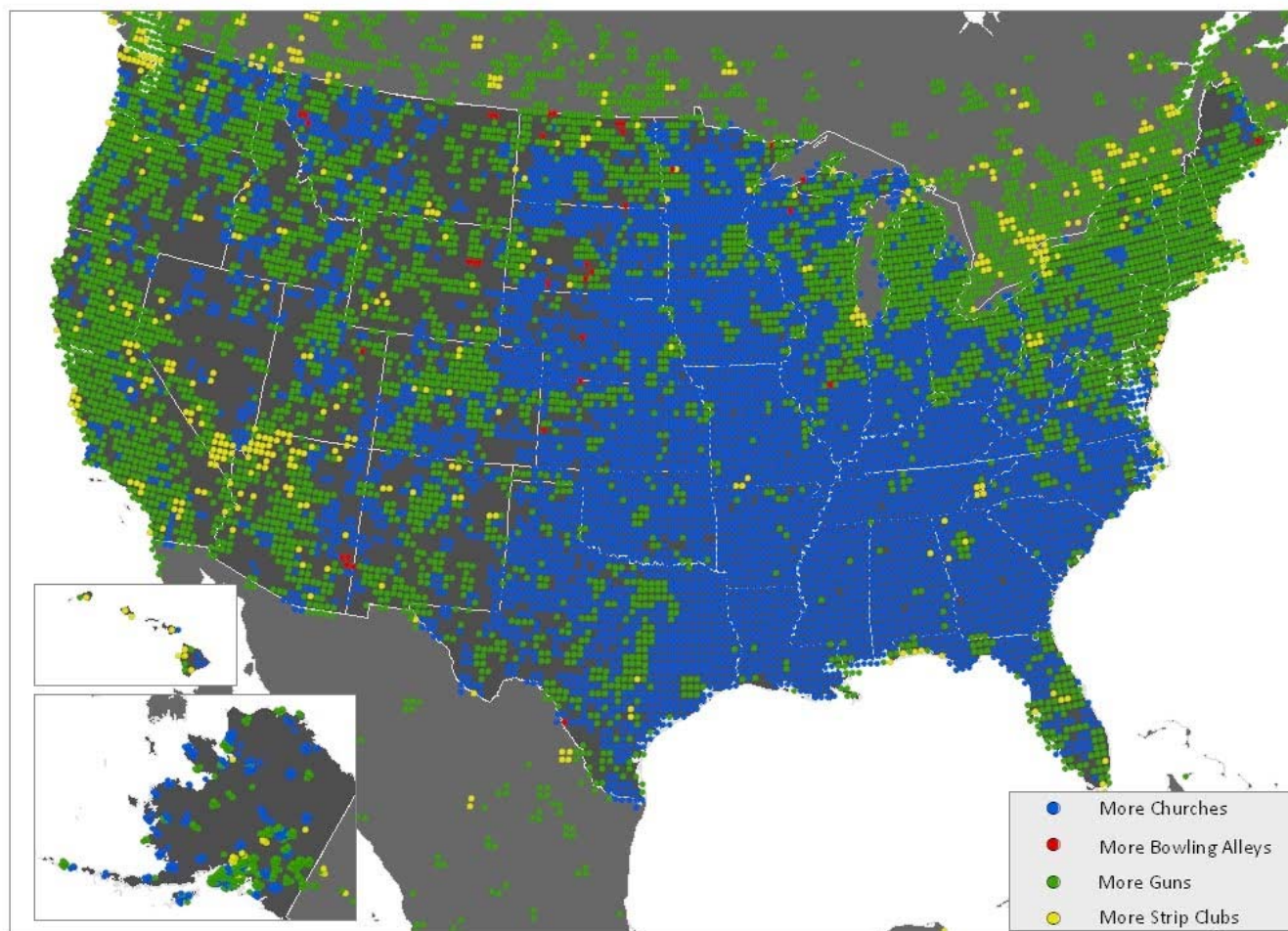
Geotag-elt Twitter bejegyzések sűrűsége Budapesten.
Forrás: Fisher (2013) alapján

A Twitter aktivitás globális területi mintázata (N=10015, 2013.05.07)



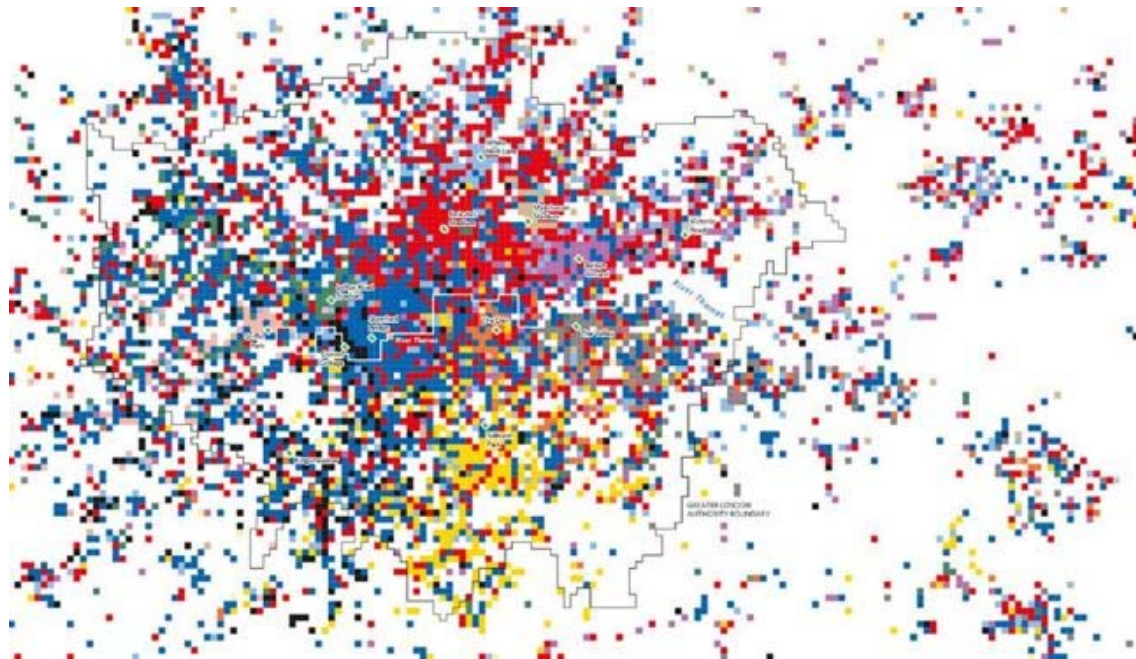
Forrás: tweetping.net

Néhány kiválasztott kulcsszó előfordulási viszonyai a geolokalizált Twitter bejegyzésekben az Egyesült Államokban



Forrás: Floatingsheep Team (2010)

A legnépszerűbb futball klubok a Twitter bejegyzések alapján Londonban



PREMIER LEAGUE

■ CHELSEA
Stamford Bridge
2013-14 record: 25-7-6 (3rd place)

WINS
LOSSES
DRAWS

■ ARSENAL
Emirates Stadium
2013-14 record: 24-7-7 (4th)

■ TOTTENHAM
White Hart Lane
2013-14 record: 21-6-11 (6th)

■ CRYSTAL PALACE
Selhurst Park
2013-14 record: 13-6-19 (11th)

■ WEST HAM
Boleyn Ground
2013-14 record: 11-7-20 (13th)

■ FULHAM
Craven Cottage
2013-14 record: 9-5-24 (19th)

FOOTBALL LEAGUE CHAMPIONSHIP

■ Queens Park Rangers
■ Charlton Athletic
■ Millwall

LEAGUE ONE

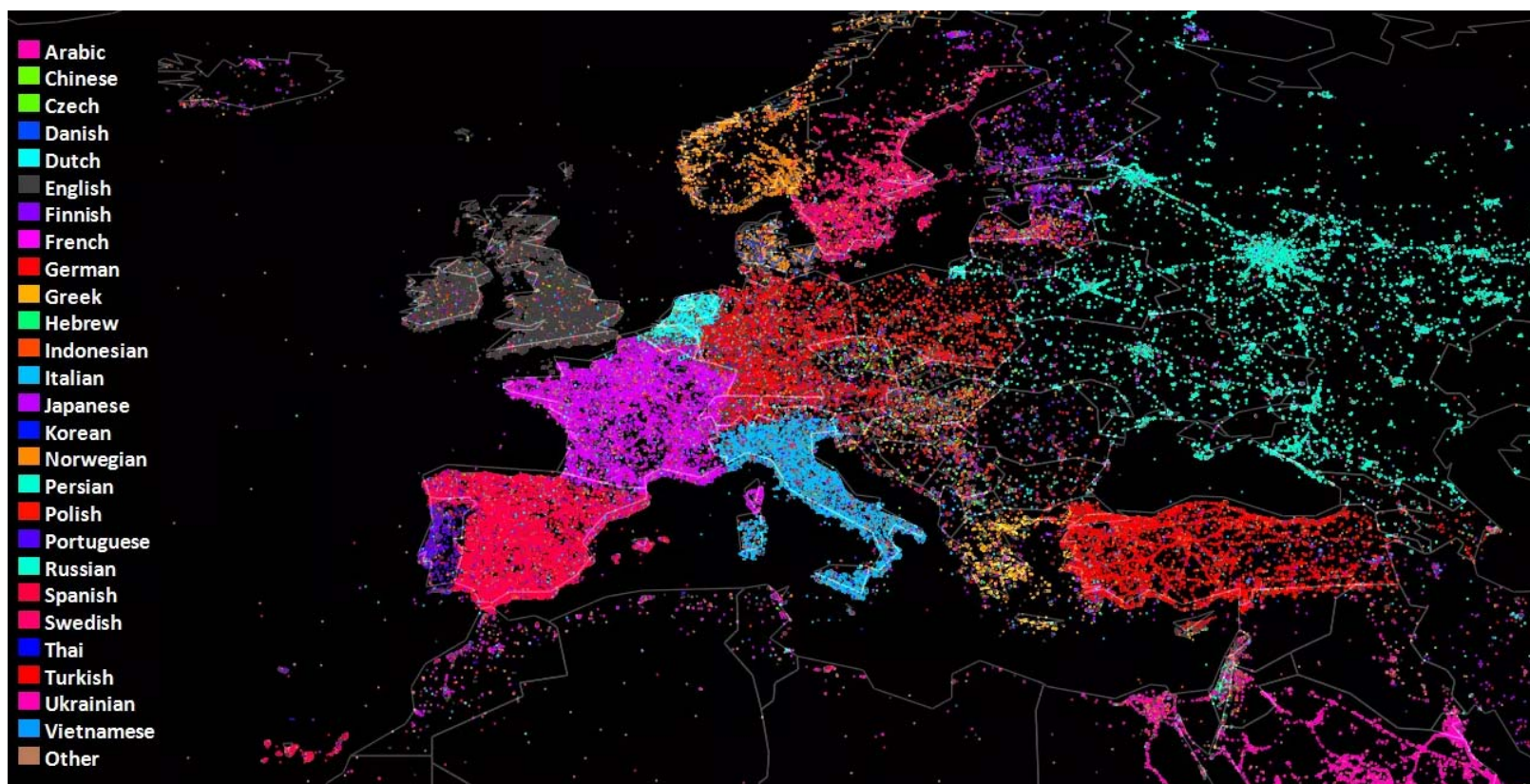
■ Brentford
■ Leyton Orient

LEAGUE TWO

■ Dagenham & Redbridge
■ AFC Wimbledon

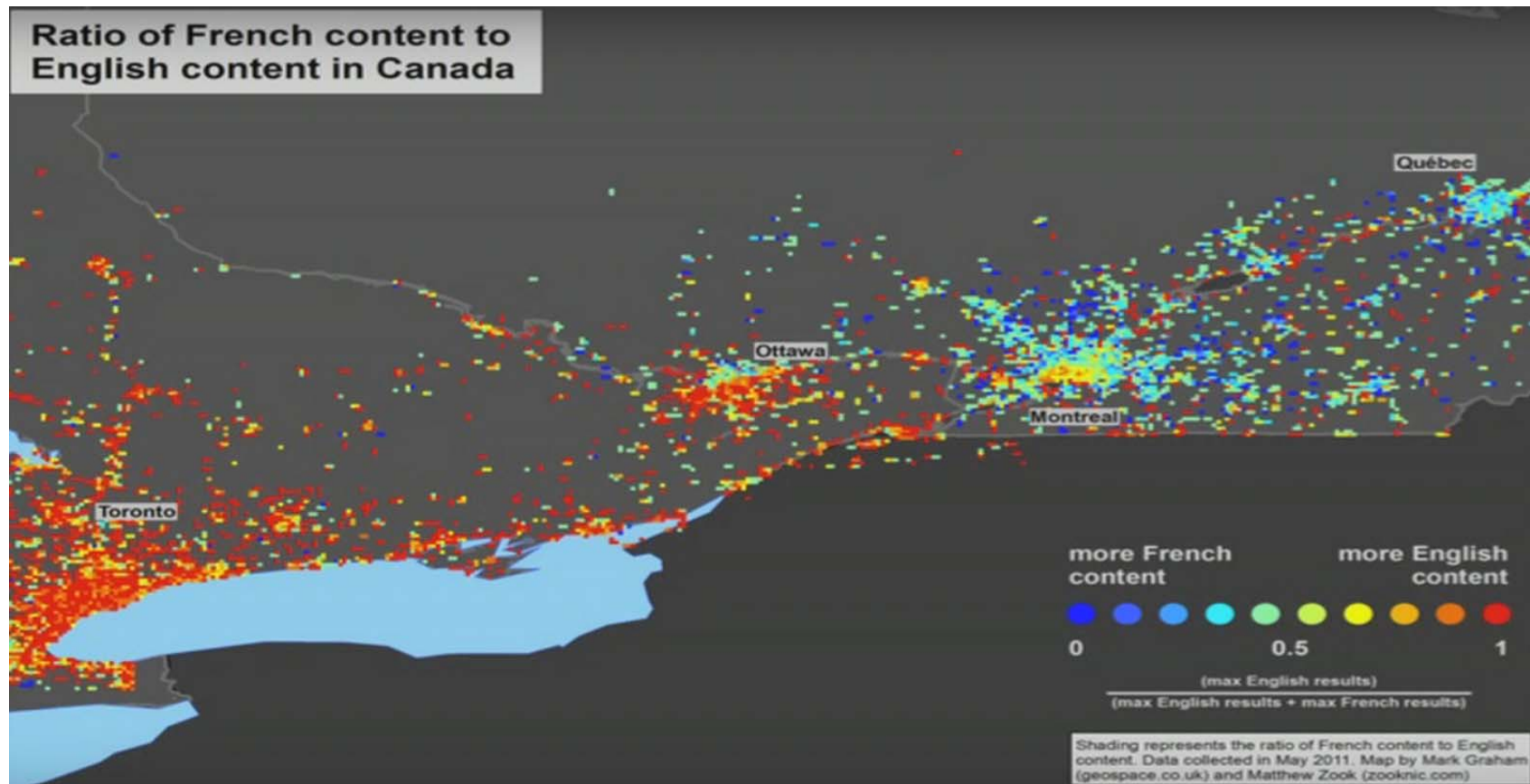
Cheshire - Uberiti (2014)

Geolokalizált Twitter bejegyzések nyelv szerinti megoszlása Európában



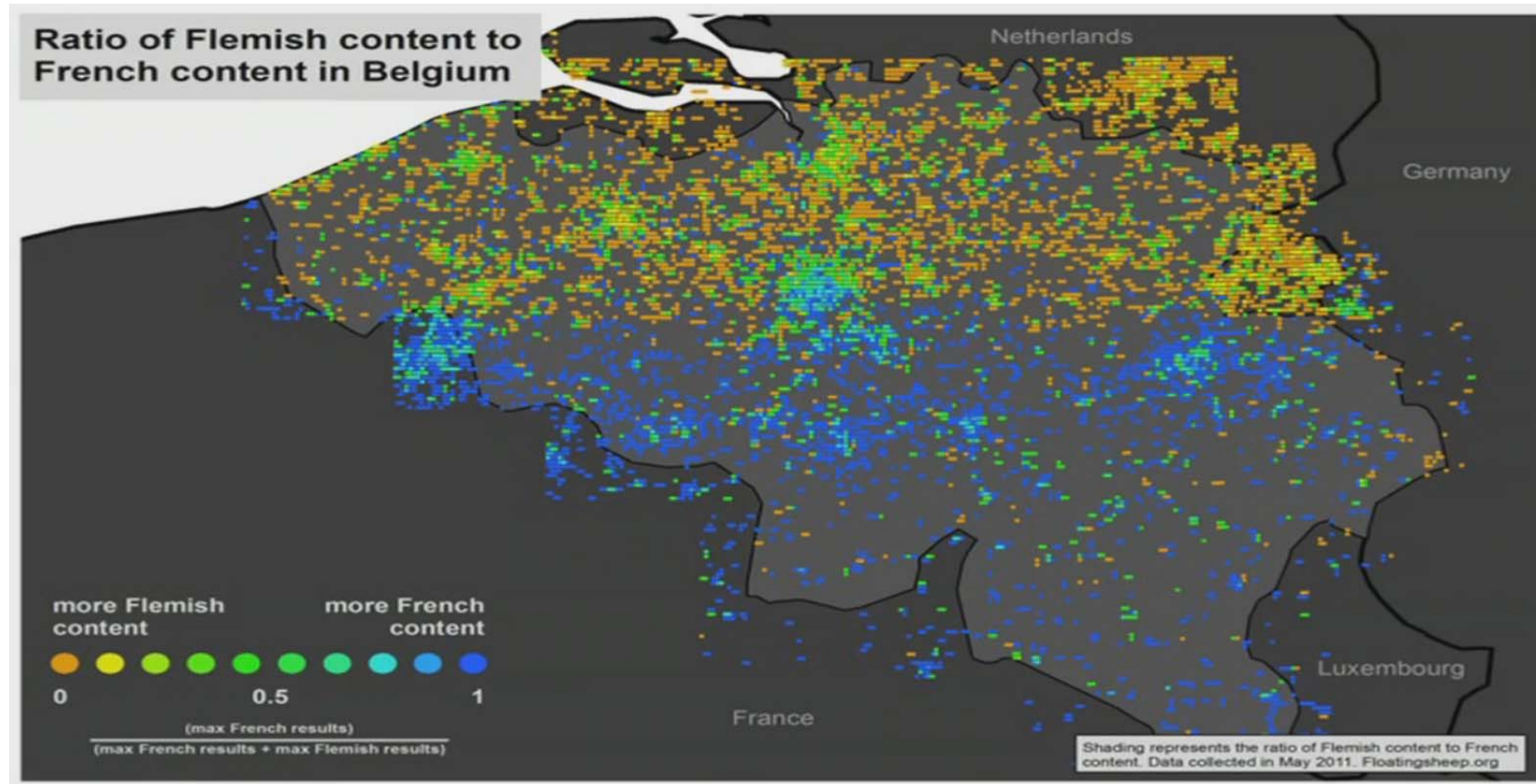
Forrás: Fischer (2011)

Geolingvisztikai térkép Twitter adatok alapján

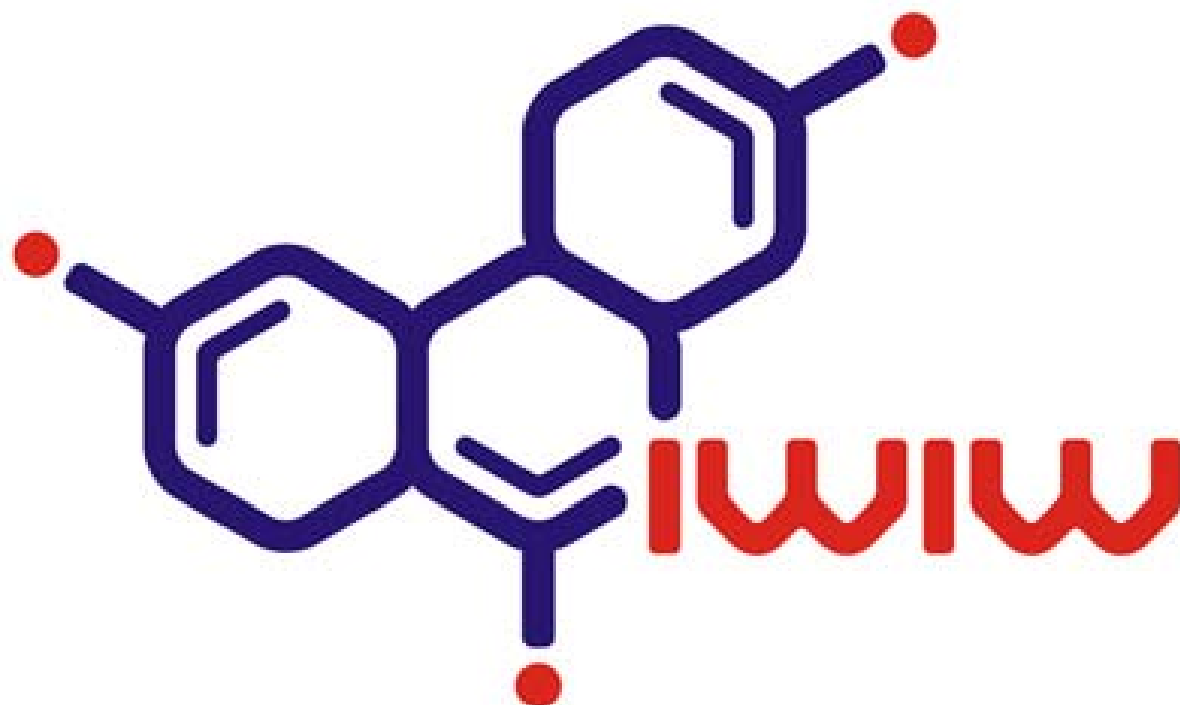


Forr s: Graham and Zook 2011

Geolingvisztikai térkép Twitter adatok alapján



Forrás: Graham and Zook 2011



Az online közösségi hálók (Online Social Networks, OSN) területi kutatásának empíriái

- Csomópontok = felhasználók
- Élek = kapcsolatok
- Kibertéri keretek: elméleti lehetőség a „tértelen” kapcsolatokra (death of distance, Cairncross 1997)
- OSN kapcsolati hálók mégis térbeli megkötöttségeket mutatnak (geography matters, de Blij 2007, Lengyel-Jakobi 2013, Jakobi-Lengyel 2014)
- Közösségi oldalak, ahol döntő többségben azokkal az emberekkel való kapcsolatainkat dokumentáljuk, akikkel az eddigi offline világ hálózataiban kapcsolatba kerültünk

iWiW esettanulmány: adatbázis jellemzők

- Adattartalom (kapcsolati adatok):
 - Felhasználói létszámok (települési aggregátumban)
 - Felhasználók közötti kapcsolatszámok (települési kapcsolat-mátrix)
- A felhasználók lokalizációja profilinformációk alapján
 - Legördülő menüből
 - Kötelező
 - Hibák nem szűrhetők
- Geolokalizáció: települési szinten

iwiw térbeli kapcsolatstruktúra-elemzés

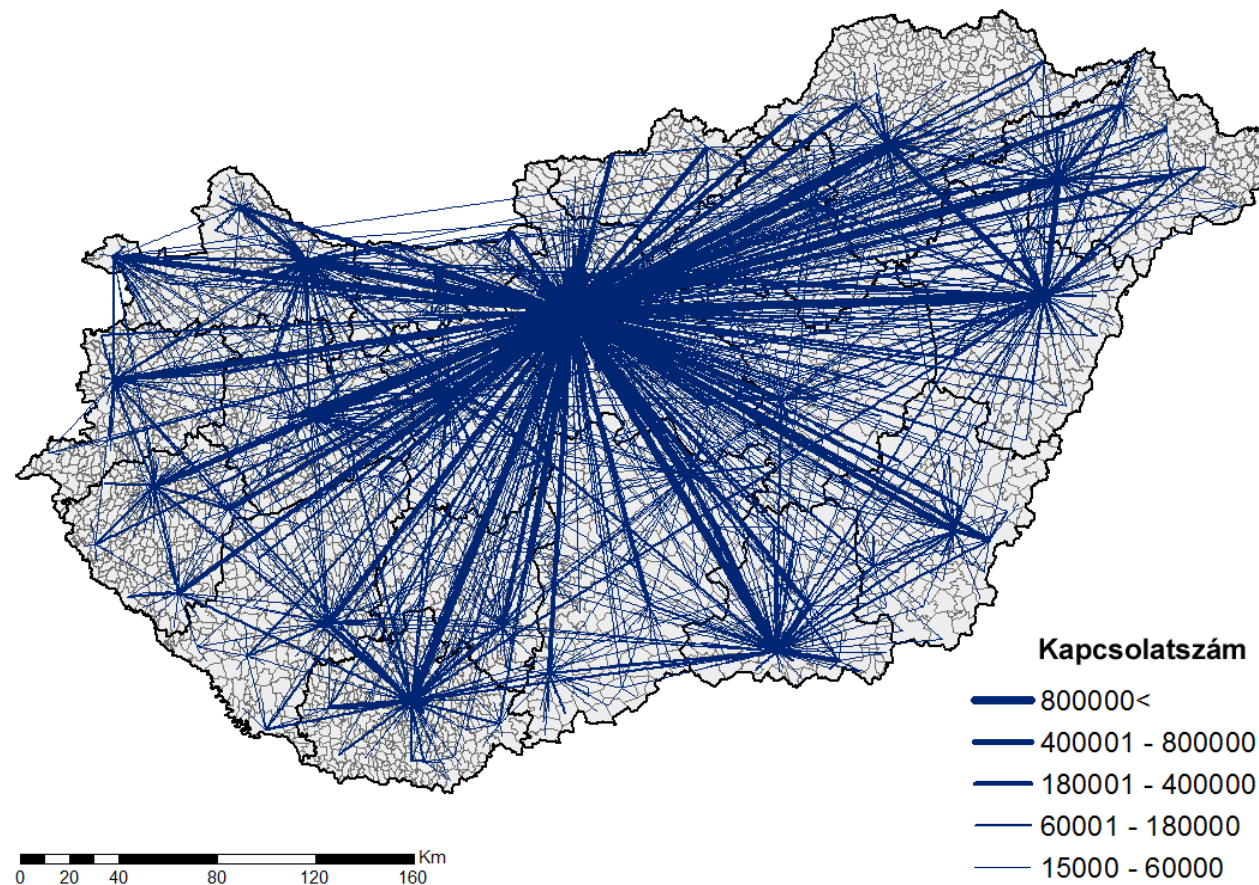
- Települési kapcsolati koordináta-párok
- Kapcsolatszám (azon felhasználók száma, akiknek a másik településen is van ismerősük)
- Hurkok az adatvizualizációból kihagyva

id1	id2	connect	latitude_1	longitude_1	latitude_2	longitude_2	weight
368	574	383	46,97134	16,50487	47,10362	16,62227	3,582
571	574	923	47,05568	16,69594	47,10362	16,62227	6,213
172	574	442	47,15947	16,63766	47,10362	16,62227	4,384
286	574	98	47,25337	16,82970	47,10362	16,62227	3,455
324	574	150	47,32122	16,48816	47,10362	16,62227	3,729
18	582	952	46,85588	18,14873	46,92726	18,23469	3,597
469	582	4 500	46,86935	18,44363	46,92726	18,23469	5,088
367	582	309	47,04457	18,18686	46,92726	18,23469	3,294
156	1 113	95	46,70373	17,39398	47,17663	17,61292	3,024
798	1 113	156	47,03306	17,53200	47,17663	17,61292	3,515

iWiW esettanulmány: adatbázis jellemzők

	Felhasználók	Települések
Csomópontok száma	4.058.505	2.562
Élek (kapcsolatok) száma	785.841.313	1.372.540
Településen belüli élek száma	369.789.373	2.562
Települések közötti élek száma	415.653.749	1.369.978

Az offline és online térbeliség összefüggései



Települések közötti hálózati kapcsolatok (összes iWiW kapcsolatszám, 15e db. felett)
(abszolút élsúlyokkal)

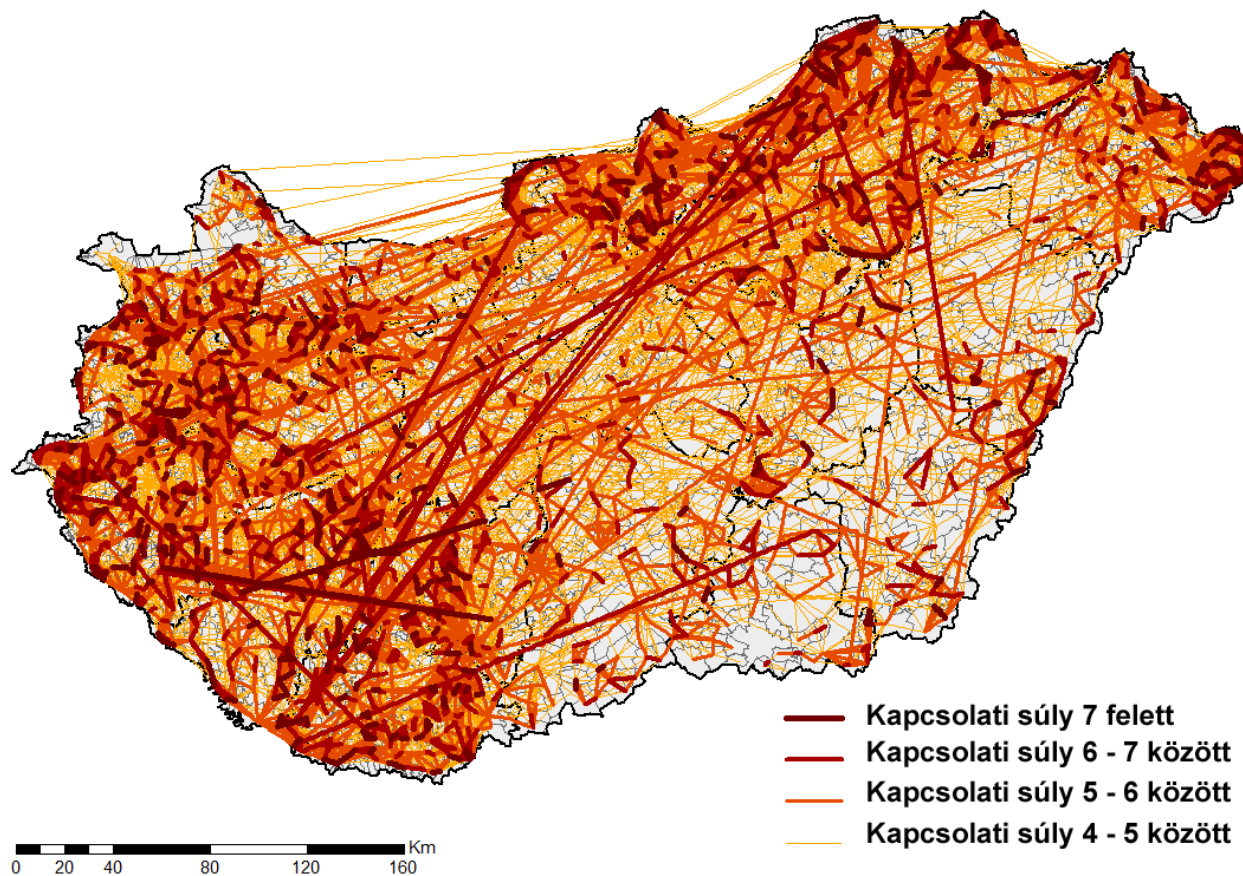
A kapcsolati adatok súlyozása

- A települések közti összeköttetéseket minden település esetében a fontosságuk szerint súlyozni lehet.
- A kialakított log-normalizált kapcsolatsúly eredmények lehetővé teszik a fontos kapcsolatok kihangsúlyozását a településmérettől függetlenül is
- A kapcsolatsúly a megfigyelt és a véletlenszerűen várható településközi kapcsolatszám egymáshoz viszonyított arányaként került meghatározásra:

$$K_{ij} = \text{Log}\left(\frac{w_{ij}}{e_{ij}}\right) = \text{Log}\left(w_{ij} / \frac{s_i s_j}{\sum_{i=1, j=1}^n w_{ij}}\right)$$

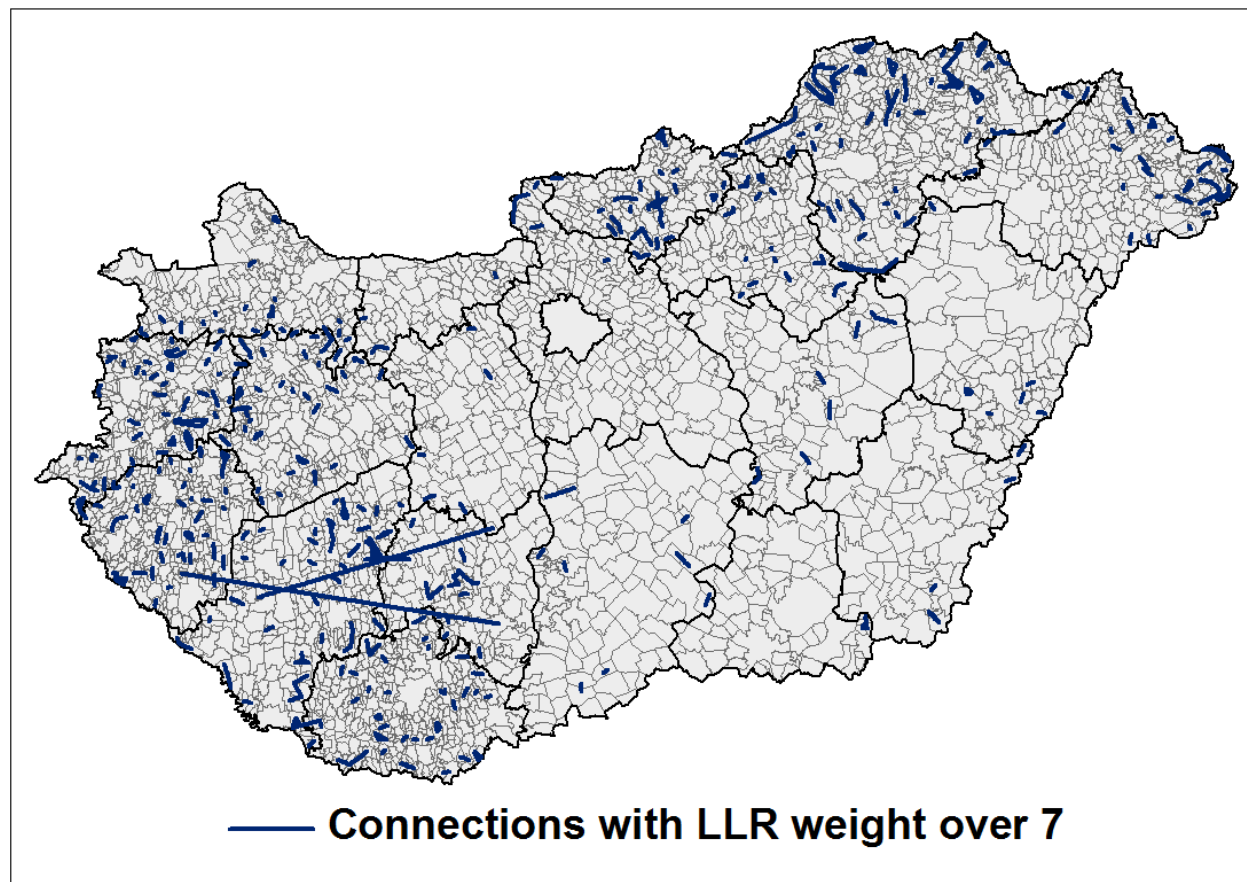
- A magas pozitív kapcsolatsúly értékek erős településközi kapcsolatokra utalnak, a nagyon alacsony negatív értékek gyenge összeköttetéseket tükröznek

Hol találhatók a szoros kapcsolatok?

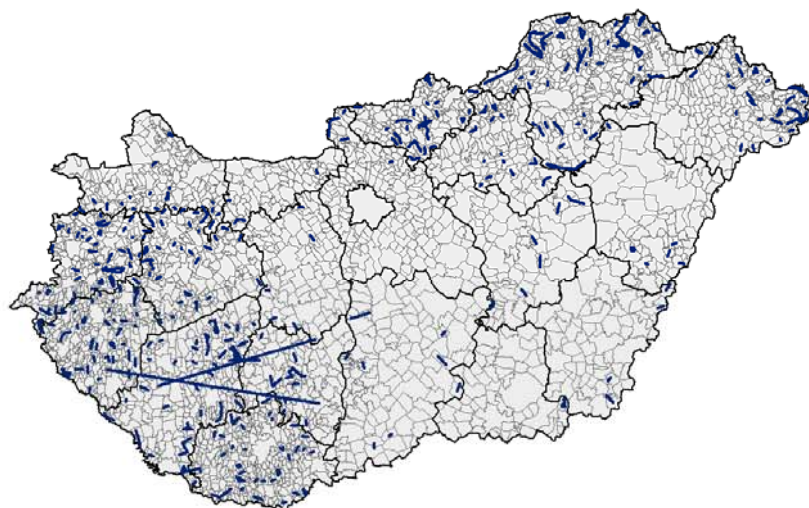


Települések közötti súlyozott hálózati kapcsolatok (log-normalizált élsúlyokkal)

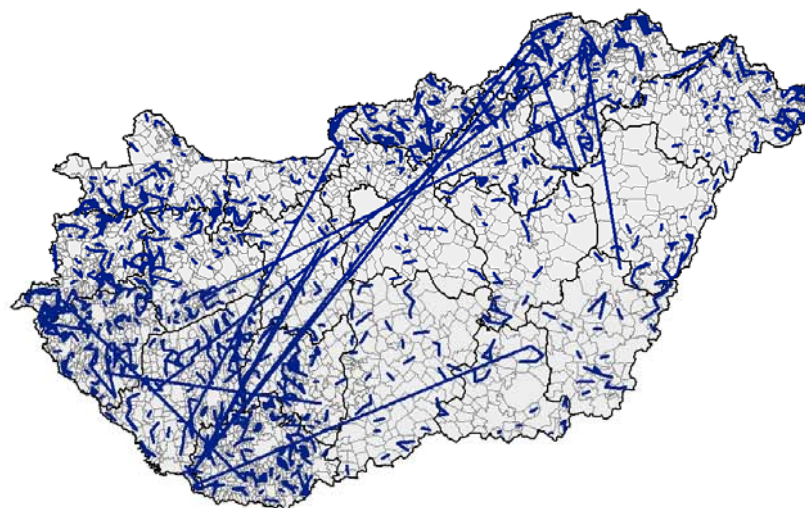
Hol találhatók a nagyon szoros kapcsolatok?



A legszorosabb kapcsolatok térképe



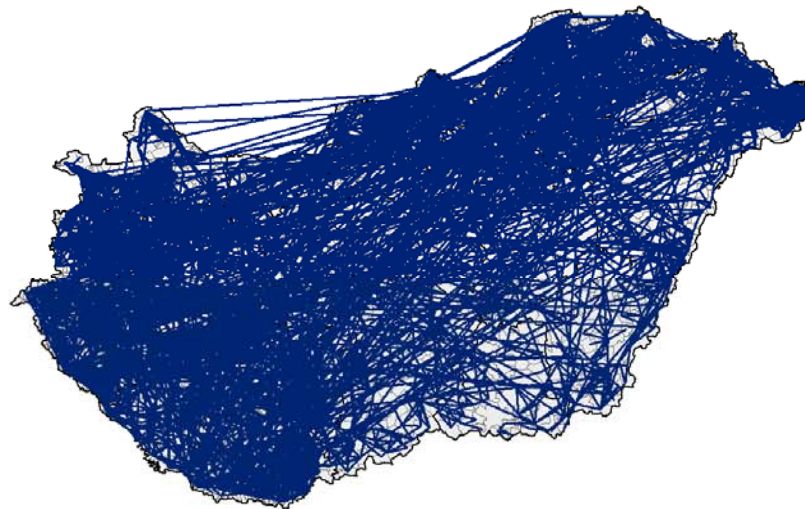
— Kapcsolatsúly 7 felett



— Kapcsolatsúly 6-7 között

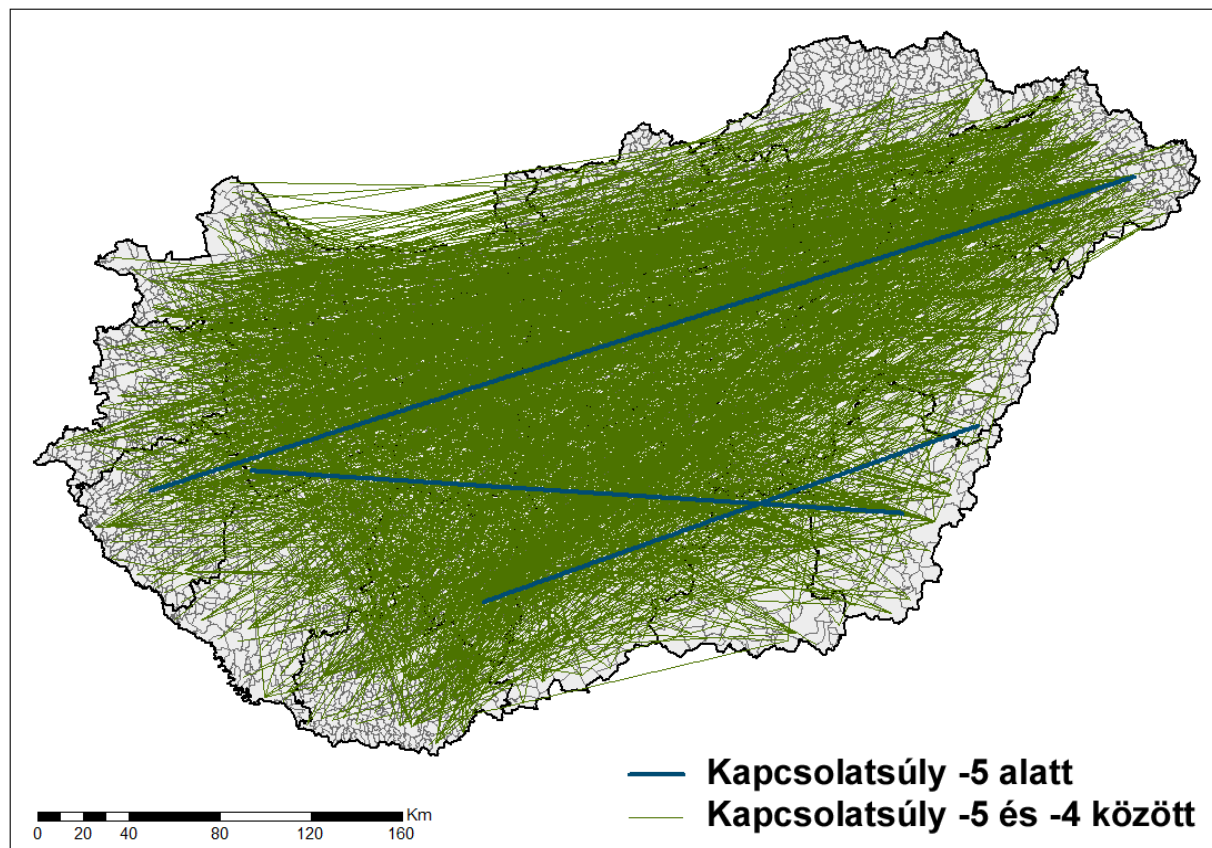


— Kapcsolatsúly 5-6 között



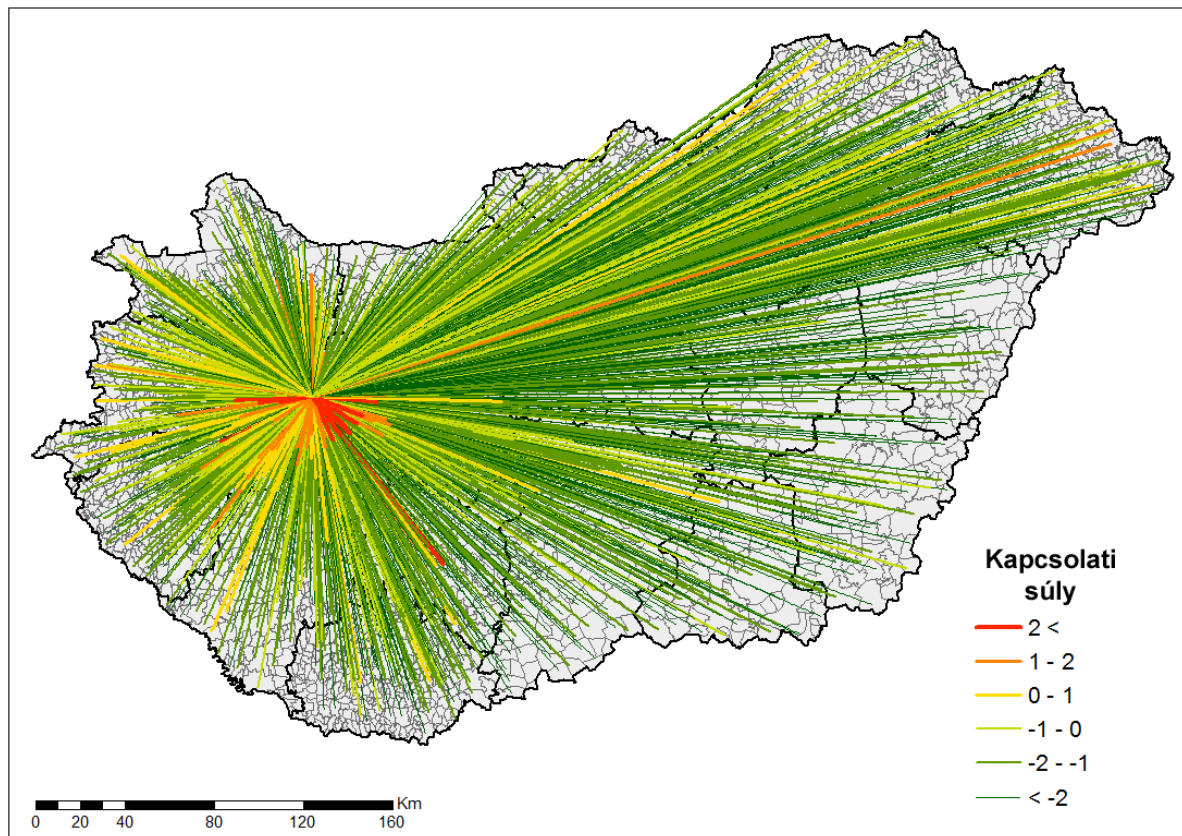
— Kapcsolatsúly 4-5 között

Hol jellemzőek a gyenge kapcsolatok?



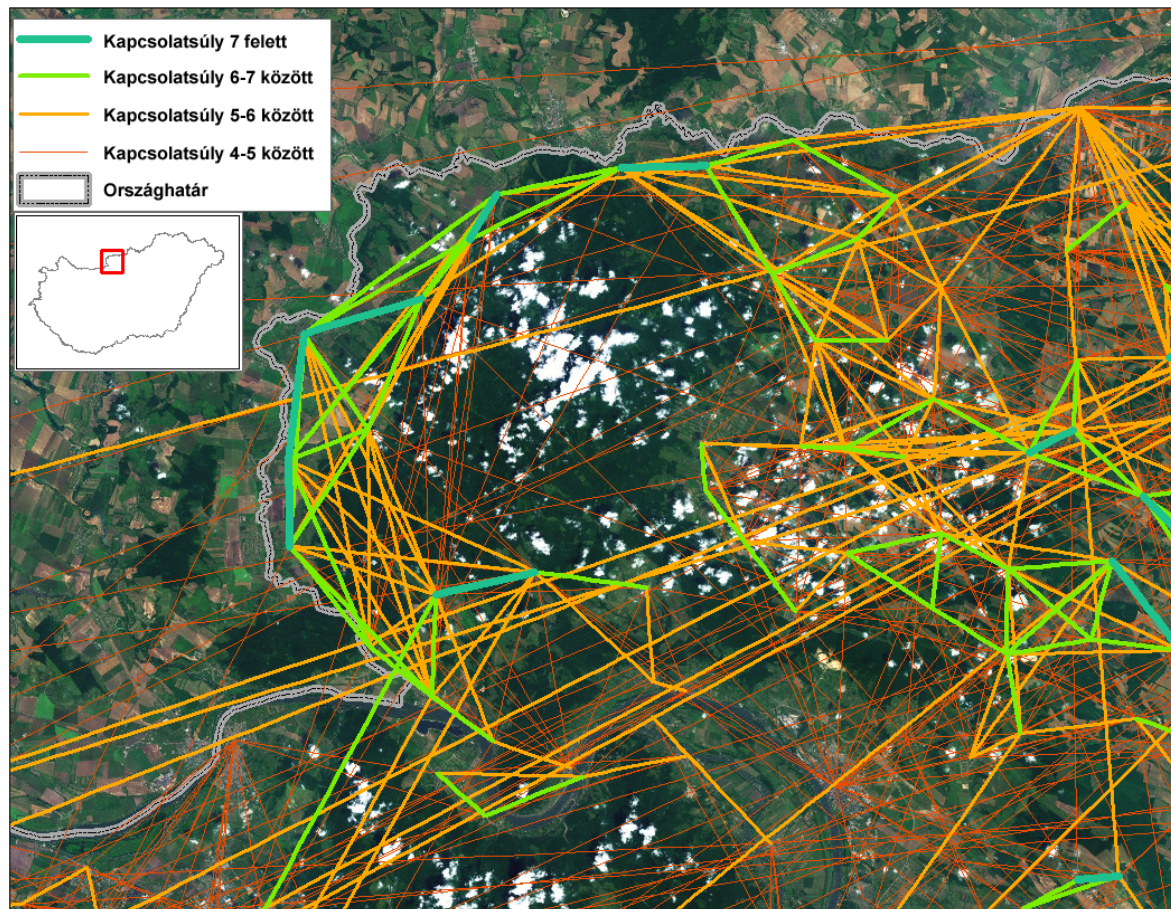
A gyenge kapcsolatok térképe

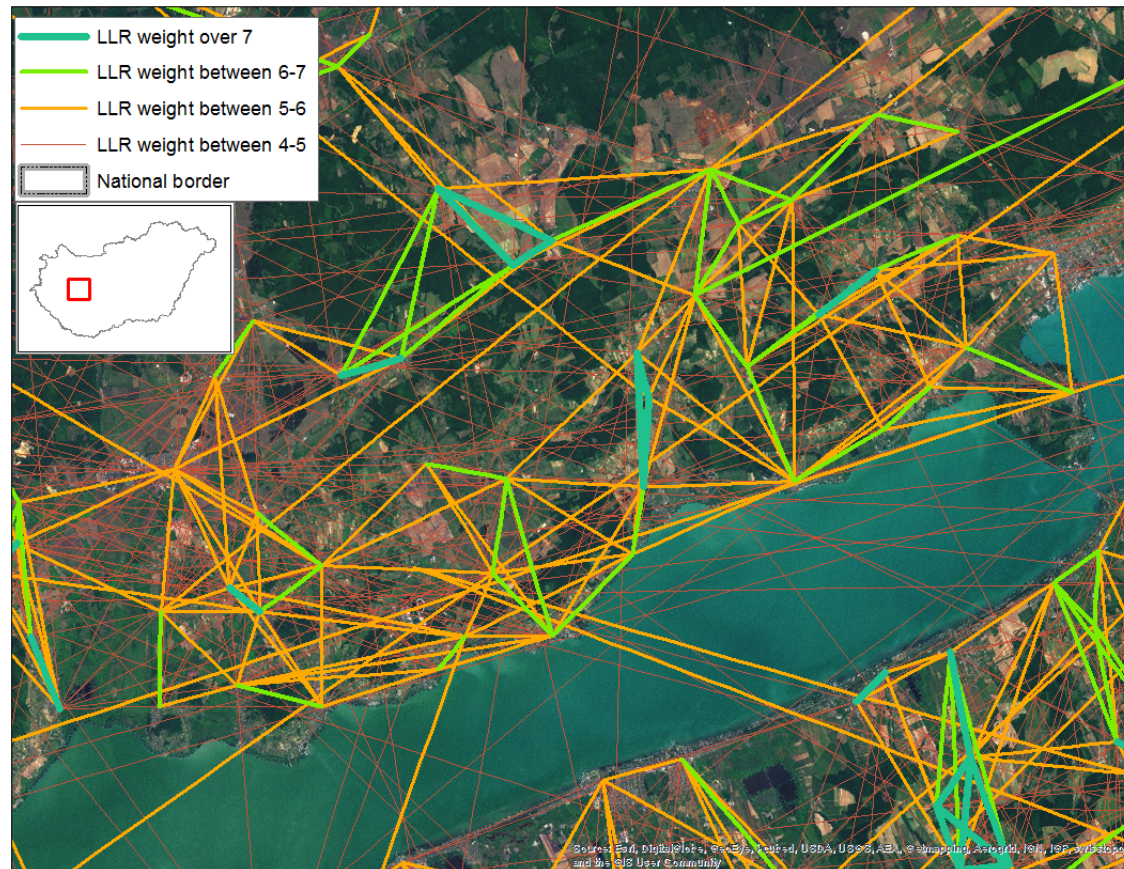
Egyedi települési kapcsolati háló



Herend hálózati kapcsolatai

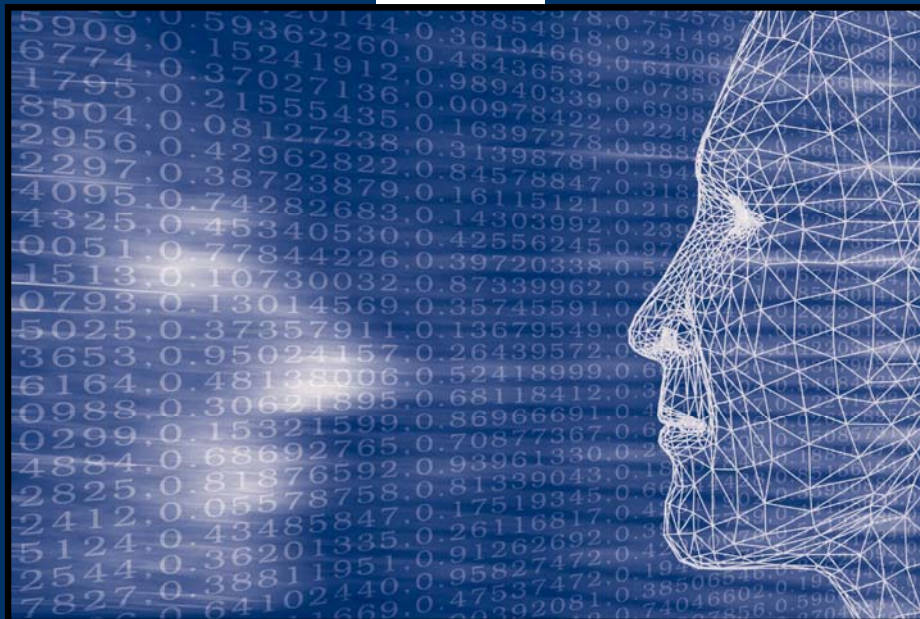
iWiW kapcsolatok és a földrajz





Összegzésként

- A legtöbb Big Data halmaz magában hordozza a területi (spatial) komponenst is
- Új lehetőségek, de új kihívások is!



**Köszönöm
a figyelmet!**

Jakobi Ákos
Eötvös Loránd Tudományegyetem
jakobi@caesar.elte.hu