

Jakobi Ákos
(ELTE)



A digitális nyomok adatszerű használatának kihívásai a területi és települési vizsgálatokban

2018. október 18-19.
MRTT XVI. Vándorgyűlés, Kecskemét

Digitális nyomok

- Bankkártya-ujjlenyomat (vásárlói profil)
- Password-ök, beléptető rendszerek, ujjlenyomat-leolvasók





Utazási kártyák

- az anonim utazás vége
- fizess többet, ha azt akarod, hogy ne kövessenek

Néhány további példa ízelítőül

- Az áramszolgáltató tudja, hogy a fogyasztó reggel mikor kel fel és este mikor fekszik le.
- Digitális TV szolgáltató: tudjuk mit nézel
- GPS, adatrögzítők az autóban: tudjuk merre jártál
- Fizetőkapuk: tudjuk mikor voltál ott
- Weboldal statisztikák: preferenciák modellezése

Digitális nyomok: a big data korszak alapkövei

- 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 értékké, 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

A digitális nyomok és a „surveillance”

- information society = transparent society
- Eltolódás az eseti megfigyelésektől a térben és időben folytonos megfigyelésig
- Majdnem minden digitális interakció és tranzakció nyomokat hagy (növekvő számban)
- Létezik egy, a fizikai test árnyékában meghúzódó „adattest” (digital body), ami nem csak követ, de esetenként meg is előz bennünket. Mielőtt megérkezünk valahova már gyakran fel vagyunk mérve és osztályozva vagyunk. (Felix Stalder, Privacy is not the antidote to surveillance, 2002)

Kihívások az adatkezelésben

- Magánélet és a „surveillance society” (a Facebook nincs ingyen!)
 - „a magánjelleg nem más, mint az egyének joga ahhoz, hogy kontrolálhassák, milyen információk kerülhetnek ki róluk” (UN Global Pulse 2012)
 - A Big Data adathalmazokban az egyéneknek vajmi kevés esélye van saját adataik felülvizsgálatára
 - GDPR (General Data Protection Regulation, általános adatvédelmi rendelet)

Kihívások az adatkezelésben

- Adathozzáférés (piaci szereplők)
- „open web”: nyilvánosan hozzáférhető online adatforrások
- Adatgyűjtési technológia (webscraping, webharvesting, crawler robotok)
- Magas fokú informatikai tudás, komoly háttérkapacitás, sőt idő szükséges
- Esély: relatíve költség-hatékony eljárások
 - közösségi adatgyűjtés (a participatory sensing)
 - alacsony költségű érzékelők (low cost sensors)
 - másodlagos adatgyűjtési módszerek (pl. web-tartalom elemzés) stb.

Kihívások az adatkezelésben

- a Big Data nem tervezett, hanem organikus eredetű
- spontán módon keletkezik, „csak úgy nő” (Schiller – Burghardt 2015)
- folyamatos az adatgenerálódás
- mindig csak minta marad, hiszen a sokaság aránya nem látható
- Javasolt inkább csak a tradicionális adatforrásokkal, ha úgy tetszik „small data” eszközökkel kiegészítve és párba állítva használni (Lazer et al. 2014)

Kihívások az adatkezelésben

- Mit mond nekünk valójában az adott adatforrás? (reprezentativitás, azaz mennyiben tükrözi a vizsgált jelenség valódi sajátosságait)
 - A közösségi médiában generálódó adatok biztosan nem lehetnek 100 százalékosan valóságosak
- Nincs előre megfogalmazott statisztikus elemzői cél, csak utólag az adatállományból találjuk azt ki
- „big data = big errors” (költséges adattisztítás)
- „hype”, azaz túlzottan is felkapott és túlértékelt jelleg
 - Bryan Walsh (2014) „automatizált arrogancia avagy big data önhittség”

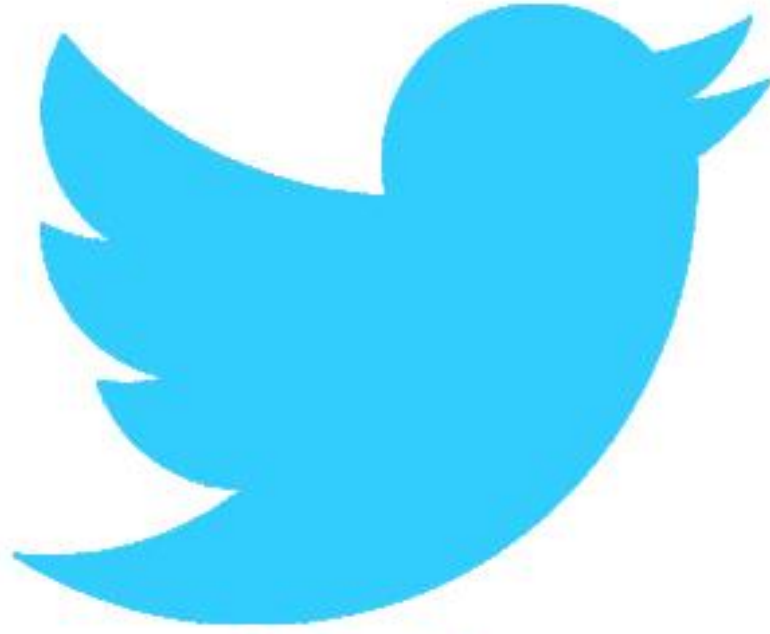
A területi kutatások lehetőségei

- Úton-útfélen hagyott *térbeli tartalommal is rendelkező* direkt vagy indirekt digitális nyomok
- Direkt digitális nyomok:
 - okostelefonok helyalkalmazásai
 - egyes honlapok forrása kiegészítve célzott geotag-ekkel
 - közvetlen fizikai szenzoros eredetűek ...
- Indirekt digitális nyomok
 - nem szándékoltnak, de mégis nagy számban keletkező spontán téradat
 - az emberi viselkedés közvetett szenzorai

Big data források

Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja szerint

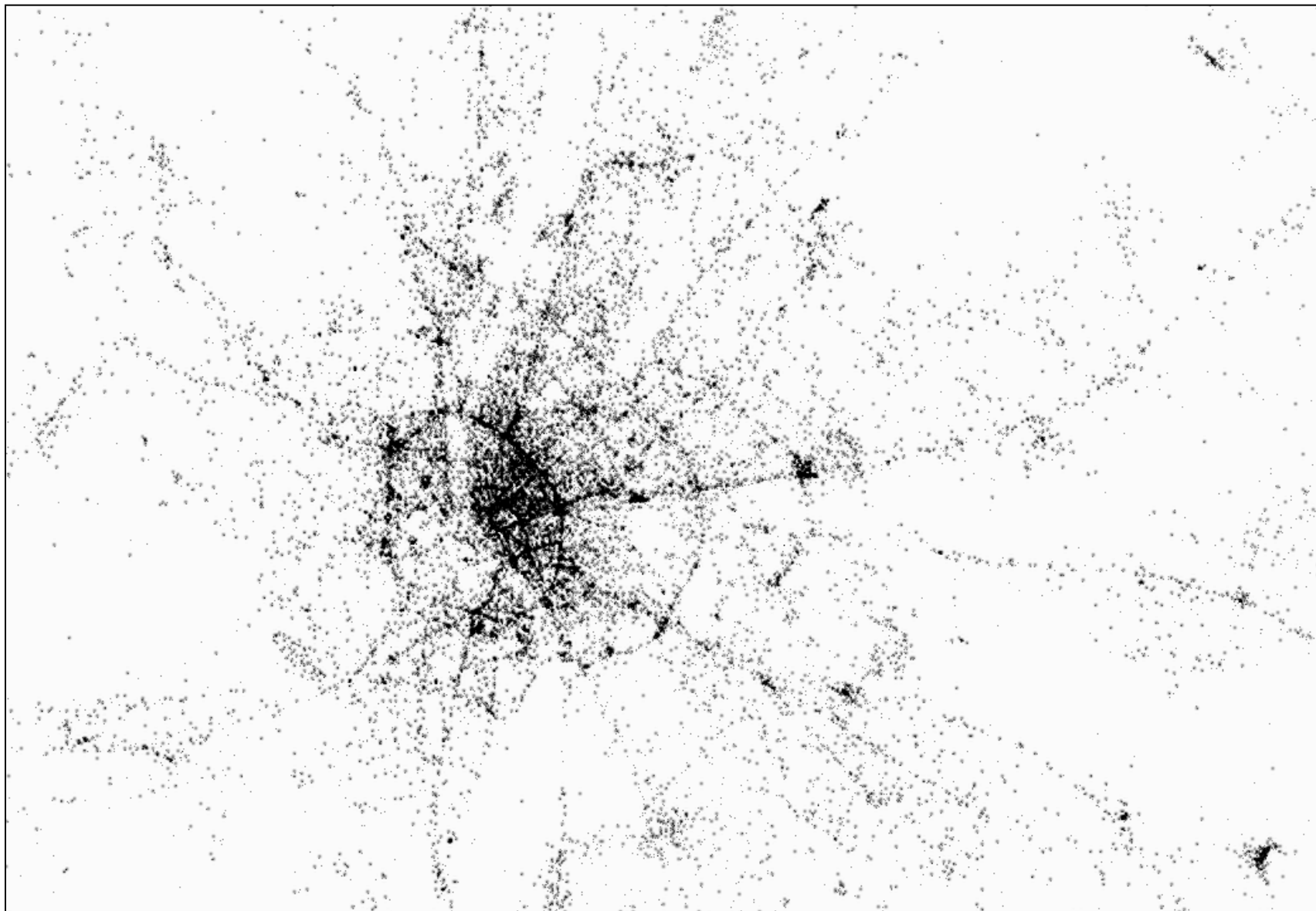
- Közösségi adatok (humán-eredetű információk), avagy „people to people” típusú adatok.
- Ezek az adatok csak lazán strukturáltak és gyakran irányítási kontrol nélküliek.
 - 1. Közösségi hálók: Facebook, Twitter, Tumblr stb.
 - 2. Blogok, hozzászólások
 - 3. Személyes dokumentumok
 - 4. Képek: Instagram, Flickr, Picasa stb.
 - 5. Videók: Youtube stb.
 - 6. Internetes keresések
 - 7. Mobil adattartalom: szöveges üzenetek
 - 8. Felhasználó által generált térképek
 - 9. E-Mail



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\u00e9 (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

Adatgyűjtés a 2018-as labdarúgó világbajnokság negyeddöntőbeli mérkőzéseinek időpontjai környékén

június 30. – Szocsi

A1	 Uruguay	2
B2	 Portugália	1

június 30. – Kazán

C1	 Franciaország	4
D2	 Argentína	3

július 2. – Szamara

E1	 Brazília	2
F2	 Mexikó	0

július 2. – Rosztov-na-Donu

G1	 Belgium	3
H2	 Japán	2

július 1. – Moszkva (Luzsnyiki)

B1	 Spanyolország	1 (3)
A2	 Oroszország	1 (4)

július 1. – Nyizsnyij Novgorod

D1	 Horvátország	1 (3)
C2	 Dánia	1 (2)

július 3. – Szentpétervár

F1	 Svédország	1
E2	 Svájc	0

július 3. – Moszkva (Otkrityije)

H1	 Kolumbia	1 (3)
G2	 Anglia	1 (4)

július 6. – Nyizsnyij Novgorod

 Uruguay	0
 Franciaország	2

július 6. – Kazán

 Brazília	1
 Belgium	2

július 7. – Szocsi

 Oroszország	1 (3)
 Horvátország	1 (4)

július 7. – Szamara

 Svédország	0
 Anglia	2

július 10. – Szentpétervár

 Franciaország	1
 Belgium	0

július 15. – Moszkva (Luzsnyiki)

 Franciaország	4
 Horvátország	2

július 11. – Moszkva (Luzsnyiki)

 Horvátország	2
 Anglia	1

Bronzmérkőzés

július 14. – Szentpétervár

 Belgium	2
 Anglia	0

Twitter adatgyűjtés

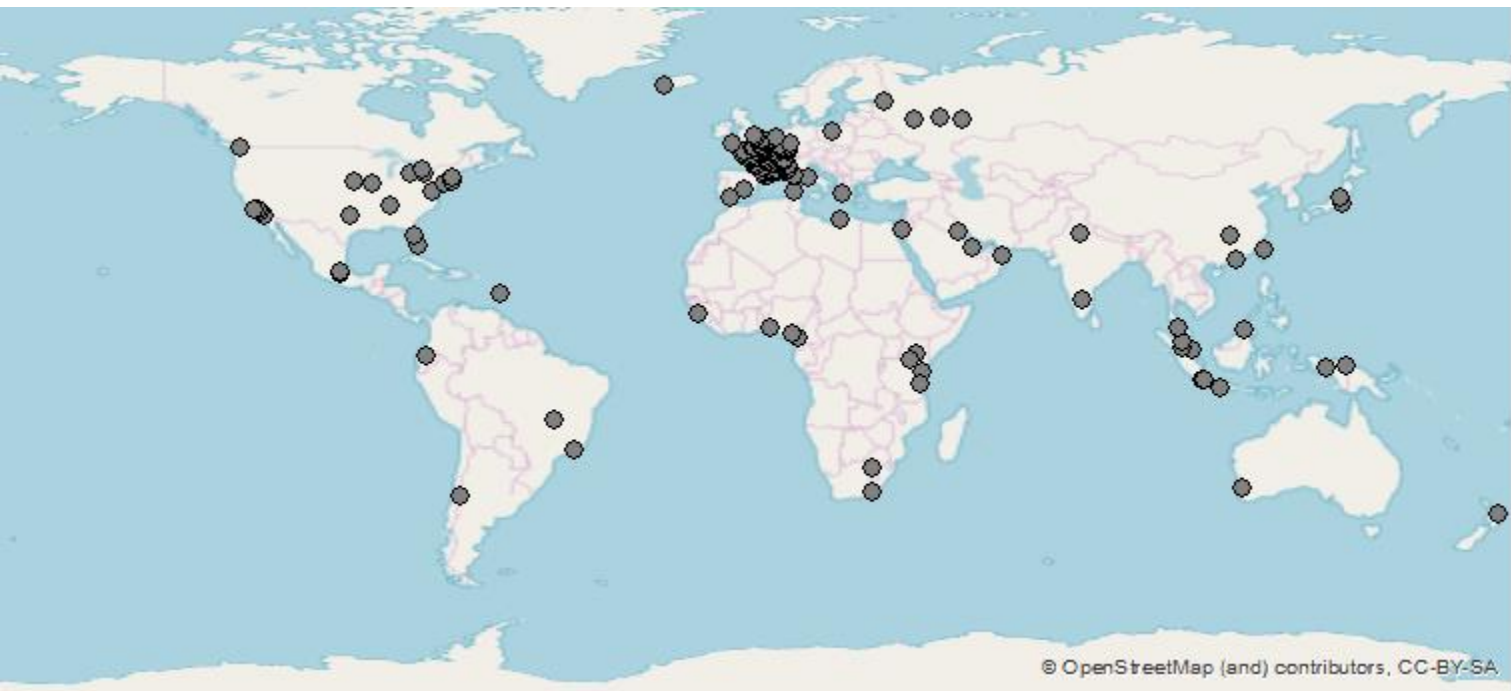
- 2018. július 6-án és július 7-én a labdarúgó mérkőzések időpontjaiban, valamint a mérkőzéseket követő másfél-két órában
- azok az esetek, amelyek az aktuálisan játszó egyik vagy másik ország nevét szövegszerűen tartalmazzák
- összesen valamivel több, mint egymillió tweet
- kb. kétezer bejegyzés geolokációs információval is rendelkezett a teljes állomány 0,2%-át adva
- GDPR miatt, úgy tűnik, egyre több felhasználó választja a geopozíció közzétételének letiltását (?)

Twitter adatgyűjtés

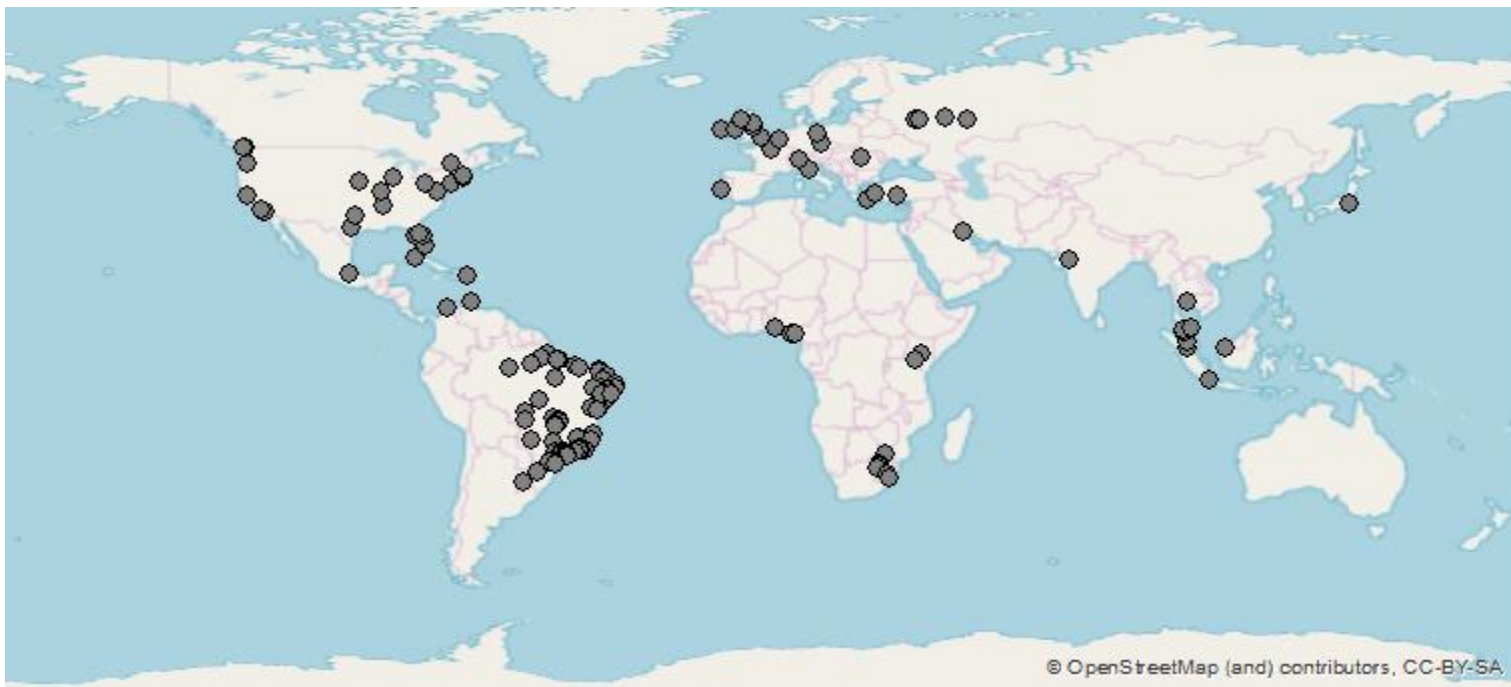
- Egy átlagos napon a bejegyzések 2,02%-a tartalmaz földrajzi metaadatokat (Leetaru et al. 2013)
- A geolokalizált tweet-ek relatíve alacsony arányát a nem lokalizált Twitter bejegyzések hirtelen (a futballesemény kapcsán) megugró száma is magyarázhatja
- Teszt: semleges időpontban a geolokalizált találatok aránya – ugyanazon kulcsszavakkal – immáron 1,95% körül alakult
- A Twitter bejegyzések számának statisztikai szempontú alakulására, úgy tűnik, hatással vannak a jelentősebb (főleg globális érdeklődést keltő) események



URU



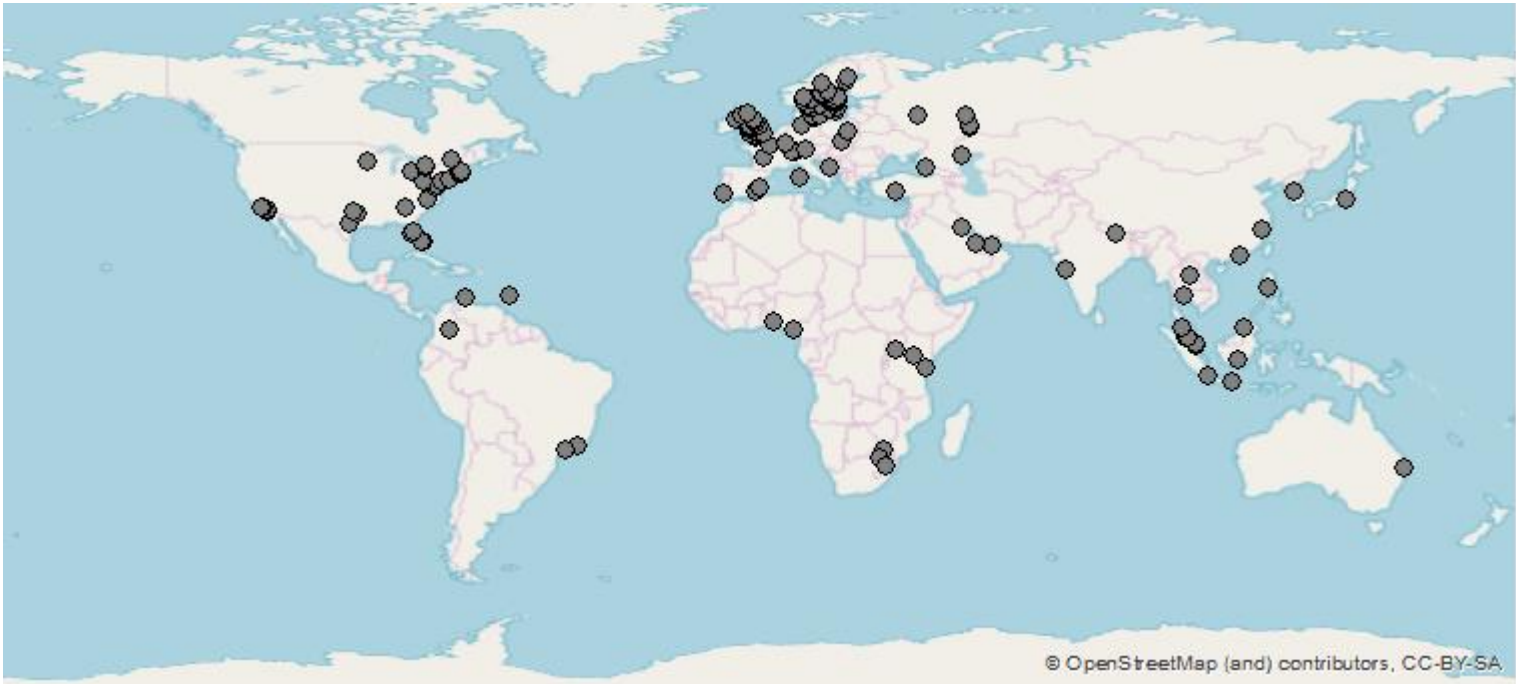
FRA



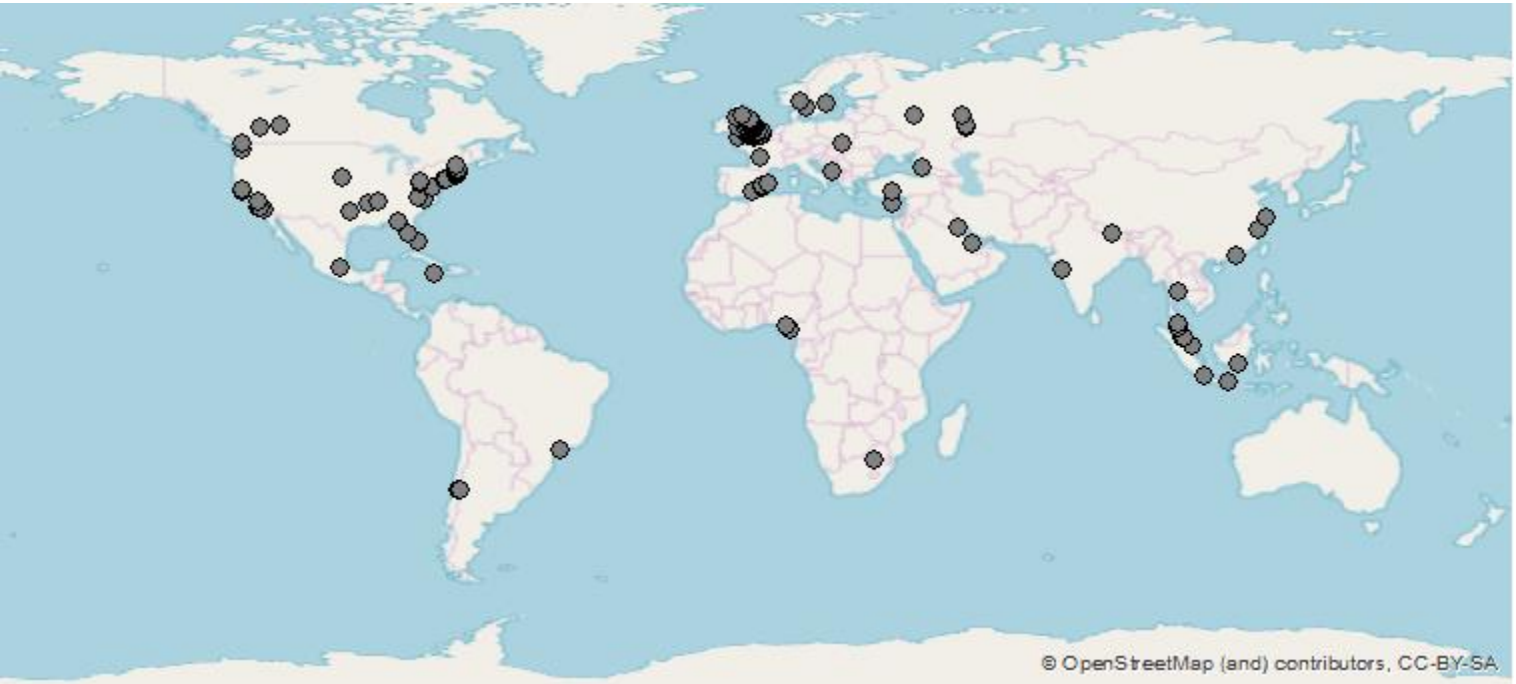
BRA



BEL



SWE



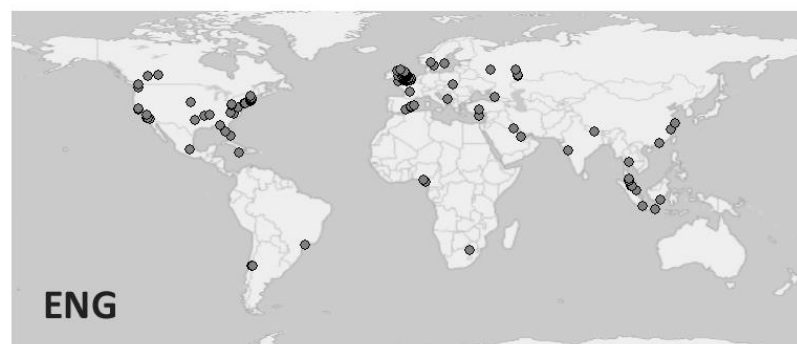
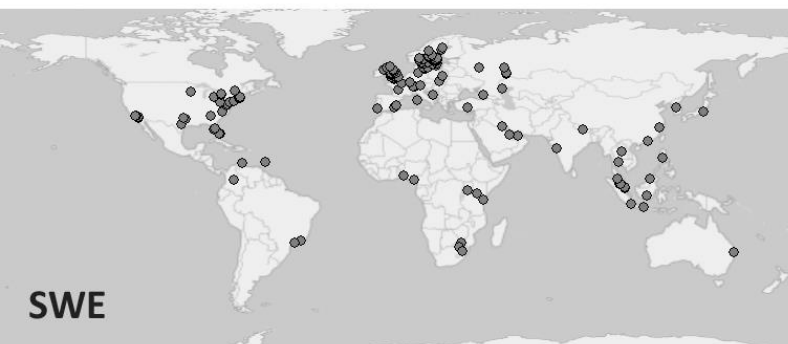
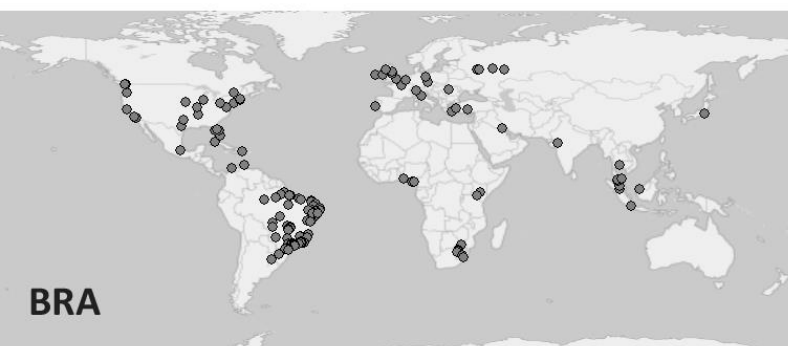
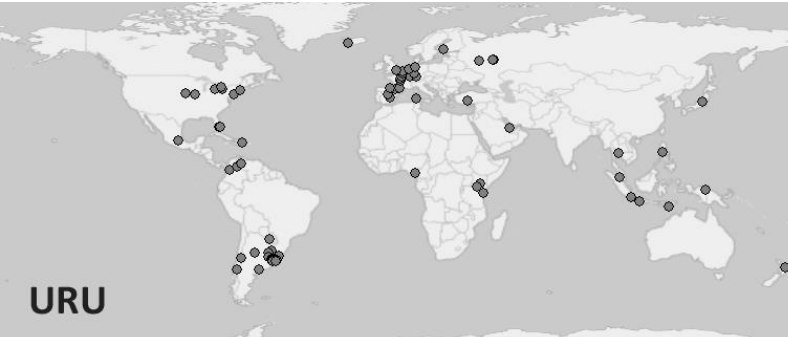
ENG



RUS



CRO



- globális Twitter aktivitás általános mintázatai
- erősen valószínű a mérkőzés kapcsán kialakult kontextuális kapcsolat
- „France” v. „Francia” (helyi nyelv v. nemzetközileg elterjedt szóhasználat)
- angolszász felülreprezentáltság
- Twitter helyi ismertsége és felhasználói penetrációja
- időzóna hatások
- a keresési kulcsszó valójában mire is vonatkozik (e rövid intervallumban valószínű, hogy tényleg a labdarúgó mérkőzéshez kötődik)
- Twitter populációs torzítás
- Geolokalizációs duplikátumok
- **Egyértelműen nem véletlenszerű mintázatok!**

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

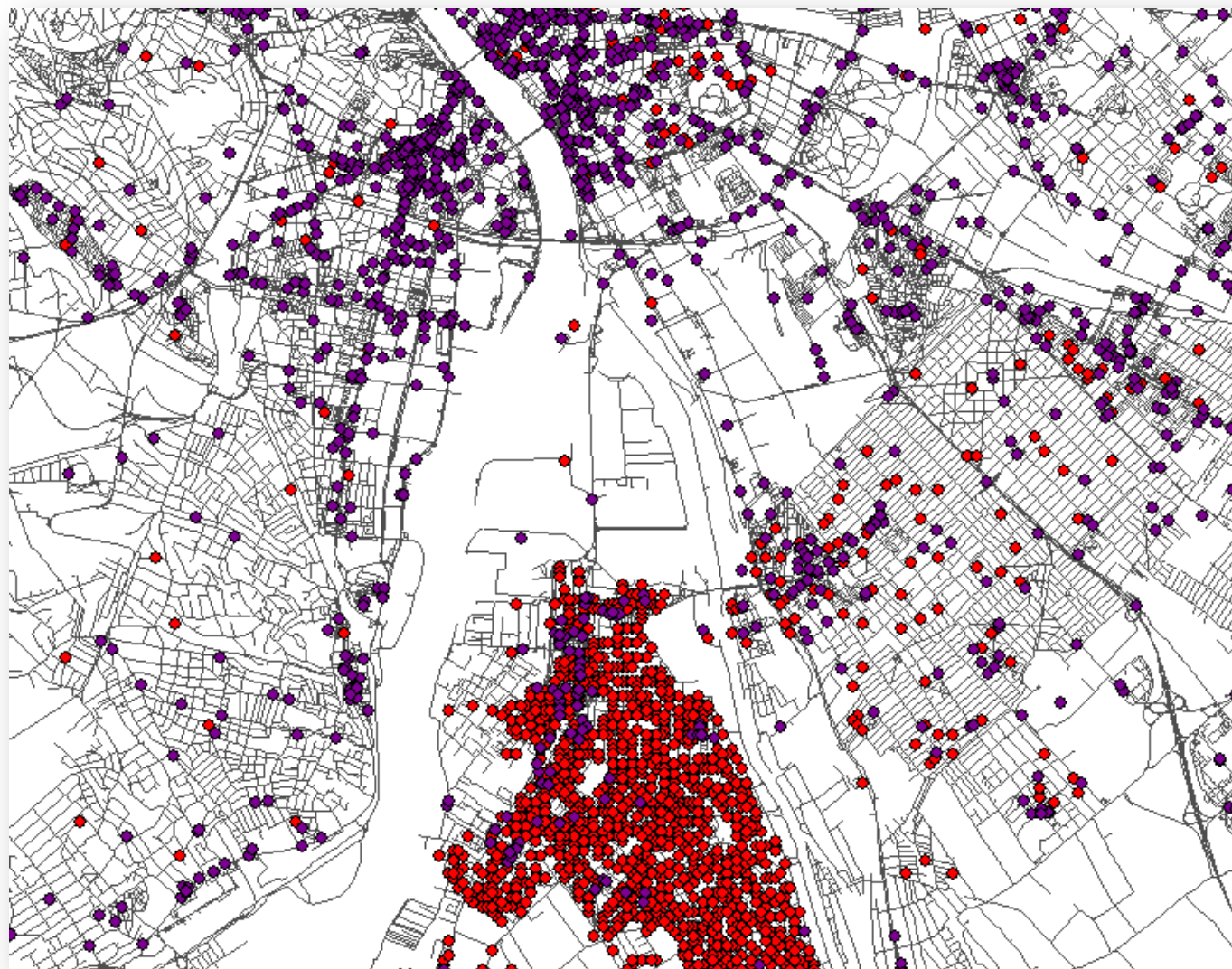


A bank
ügyfélköre
általánosságban
hol költi el a
pénzét

Banki ügyfelek tranzakciós aktivitása

- Adathozzáférés (privát vállalati adat)
- Szenzitív adatok
- Csak aggregátumok szintjén feldolgozható (külső szakembereknek)
- Csak POS tranzakciók (ATM, fióki trx, online nem)
- Csak egyetlen bank és csak azon ügyfelei, akik egyben bankkártya tulajdonosok is és valamely POS terminálon rögzített vásárlási rekorddal rendelkeznek
- Ilyen léptékű lokalizált kérdőíves vagy statisztikai felmérés amúgy nemigen elérhető

Otthon-munkahely kapcsolat azonosítása



- Adott (kiválasztott) bankfiókhoz tartozó ügyfelek lakóhelyei
- Napközbeni tranzakciós helyszínek (a kiválasztott bankfiók ügyfelei esetében)

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)

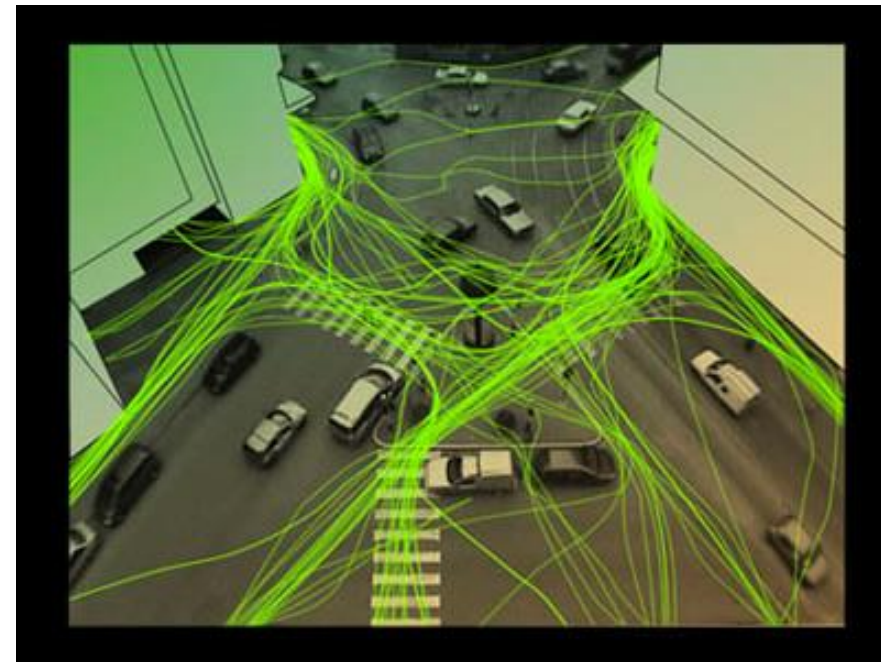
- 121. Személyes (mobil helymeghatározás)
- 122. Közúti (autók, teherszállítás)
- 123. Vasúti (vonatok)
- 124. Légi (repülőgépek)
- 125. Vízi (hajók)

13. Műholdas adatok

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

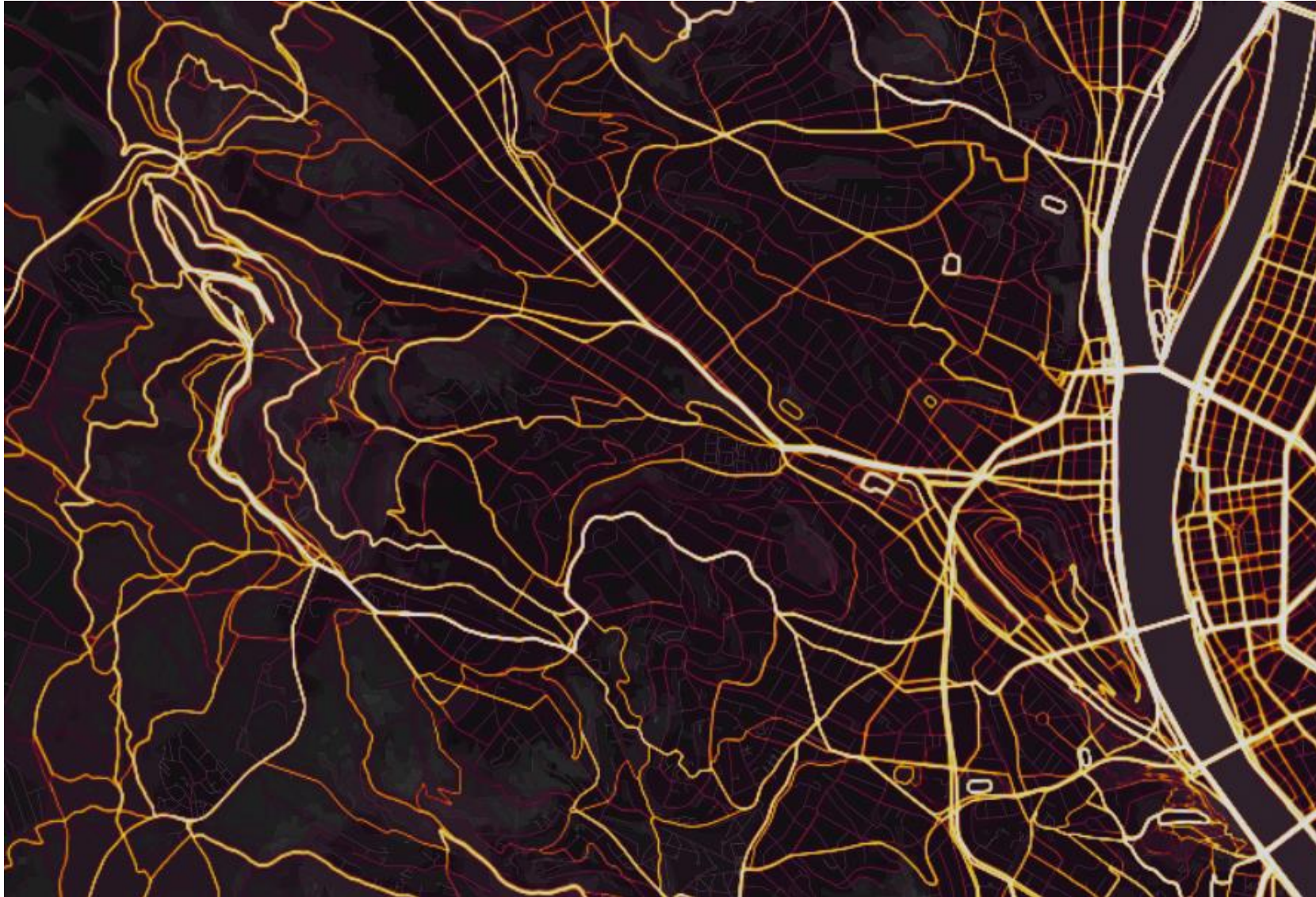
2. Számítógépes rendszerekből származó adatok

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



**Városi
térpályák**

A szabadidős sporttevékenységek (futás, kerékpározás) városi térpályái



Forrás: www.strava.com

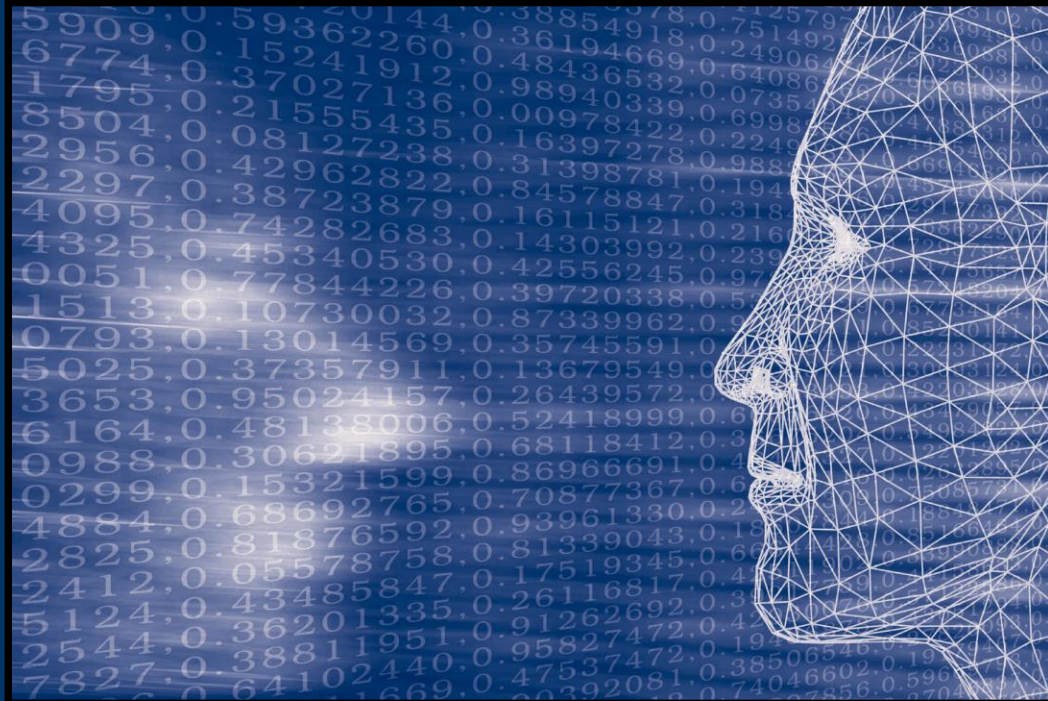
A szabadidős sporttevékenységek (futás, kerékpározás) városi térpályái



Forrás: www.strava.com

A szabadidős sporttevékenységek városi térpályái

- Láthatóvá válnak azok az utak is, amelyek csak a szabadidős sporttevékenység (futás, kerékpározás) szempontjából fontosak
- Közösségi adatgyűjtés
- Felhasználói hozzájárulás szükséges
- Itt sem lehet meghatározni a reprezentativitás mértékét
- Sporttevékenységet tekintve aktív, de az infokommunikációs eszközök használatában is jártas (valamint azzal egyáltalán rendelkező) felhasználók
- Egyedi torzítások (túl aktív felhasználók)
- Szenzitív lokációk



**Köszönöm
a figyelmet!**

Jakobi Ákos
Eötvös Loránd Tudományegyetem
jakobi@elte.hu