BIG DATA

A HIVATALOS STATISZTIKÁBAN

ADATÁRADAT

- 2,5 exabyte (10^{18}) adatot állítunk elő minden áldott nap (ennek a kétszerese arra elég, hogy az emberiség által valaha kimondott összes szót tartalmazza)
- 2012-ben: az emberiség által valaha termelt adatok 90%-a az utóbbi két évben keletkezett.
- A Wallmart óránként 1 millió tranzakciót rögzít.

ADATÁRADAT

De nem csak az adatok mennyisége növekszik, hanem – főleg a social media és a mobiltelefonok szolgáltatásainak széleskörű terjedése miatt – az adatok természete is változik.

Ennek az információnak a nagyja olyan adatforma, amely digitálisan követhető vagy tárolható. Többniyre cselekedetek, választások vagy preferenciák, amelyet az emberek az életük során termelnek.

BIG DATA DEFINÍCIÓJA

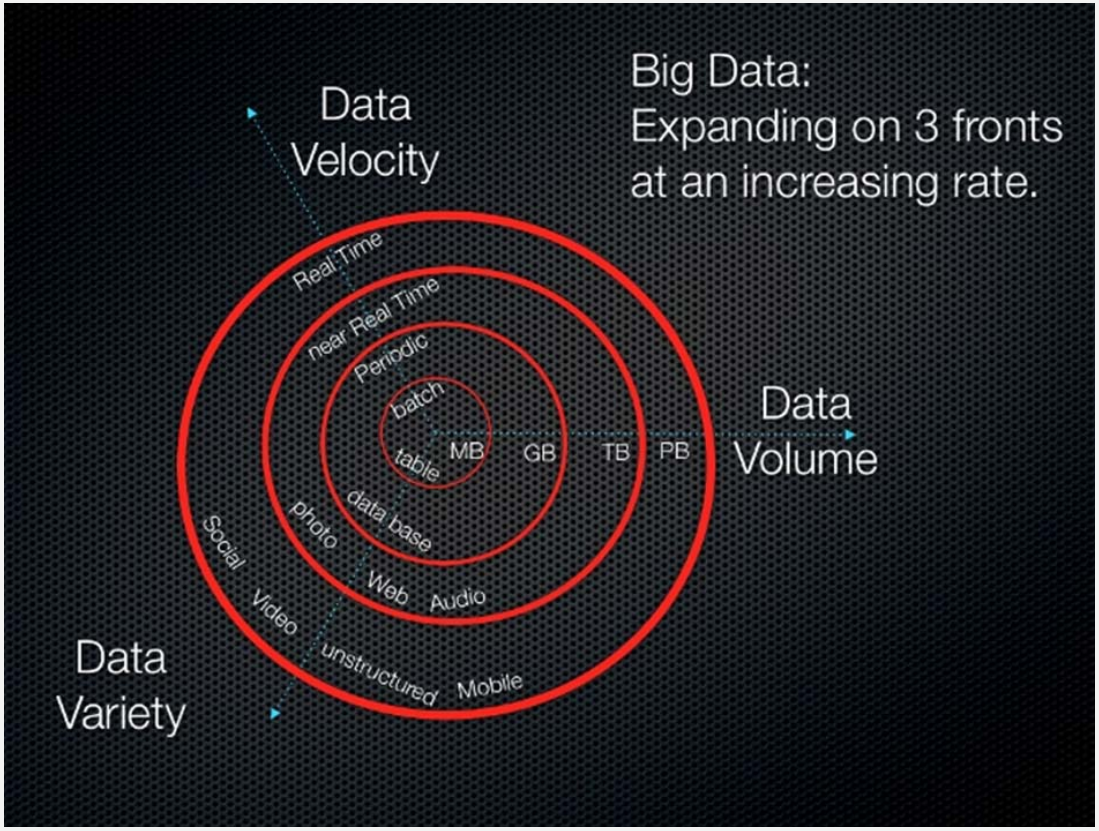
Wiki: a BIG DATA olyan nagy és komplex adathalmazok összessége, amelyeknek a kezelése hagyományos adatbáziskezelő eszközökkel nem lehetséges.

Volume

Variety

Velocity

Veracity



Kép forrása: ISTAT

TAXONÓMIA

Human sourced information (eg. social networks): az emberi tapasztalatok szubjektív recordjai, amelyeket korábba könyvekben, művészi akotásokban azt követően foto/audio/video-bann tároltak

Process mediated data: üzleti folyamatok adatai, magasan strukturált, hogyományos üzleti rendszerek termékei

Machine made data: számítógépes log file-ok, szenzorok és gépek digitális adatai. Ez a típus adja magát a számítógépes feldolgozáshoz, mert jól strukturált, de a mennyisége és gyorsasága gyakran a hagyományos eszközökön túl mutat (Internet of Things).

BIG DATA KEZELÉSE

- Könnyen belátható, hogy a Big Data-dömping kezeléséhez, a folyamatosan áramló adatok gyűjtéséhez, tárolásához, előkészítéséhez és feldolgozásához meg kell felelni bizonyos előfeltételeknek.

    - Egyre növekvő számítástechnikai teljesítményre van szükség, amelyet MPP- (massive parallel processing – masszív párhuzamos fel- dolgozás) megoldásokkal lehet kezelni.
    - Elengedhetetlen az adatredisztribúció és a párhuzamos feldolgo- zás lehetőségének megteremtése (a MapReduce, a Hadoop, a Hortonworks Data Platform, az R-Rstudio stb. ismerete és alkalmazhatósága).
    - Nélkülözhetetlen a nem csak SQL-re épülő, adatmennyiség- redukáló szoftvertechnológia használata és az abban való jártasság. A statisztikusok szempontjából ugyanakkor kérdéses, hogy a Big Data alkalmazása milyen IT-ismereteket kíván meg.

BIG DATA A STATISZTIKÁBAN

A hagyományos statisztika minőségének javítása:

Hol tud javítani:

-mintavételi keret létrehozása

-jobb minták tervezése

-jobb imputálás/kalibrálás

-nem válaszolási arány csökkentése

HAGYOMÁNYOS MEGKÖZELÍTÉS

-milyen információra van szükségünk / hipotézis kell!

-az adatgyűjtés megtervezése

-adatgyűjtés

-adat előkészítése

-adatelemzés

-információ kinyerése / hipotézis igazolása vagy cáfolata

TOP-DOWN PARADIGMA
A HAGYOMÁNYOS MÓDSZER

- A hivatalos statisztika általános gyakorlata szerint egy adatfelvétel előtt elsőként azt kell meghatározni, hogy milyen információkra van szükségünk, és ehhez hipotéziseket fogalmazunk meg. Majd a következő lépéseket hajtjuk végre: ***1*. adatgyűjtés-tervezés, *2*. adatgyűjtés, *3*. adat-előkészítés, *4*. adatelemzés, *5*. információkinyerés az adatbázisból/a felállított hipotézis igazolása vagy cáfolata**
- A top-down paradigma lényege, hogy az adatgyűjtés megtervezése során az elemzési cél(ok) meghatározásán van a hangsúly. A hagyományos adatfelvételeknek tehát kulcsfontosságú eleme a tervezés, melynek részei a következők: *1*. változók, definíciók kialakítása, konceptualizálás, majd operacionalizálás, *2*. a vizsgálni kívánt sokaság kiválasztása (ez lehet teljes körű, vagy alapulhat mintavételen), *3*. az alapsokaság elérésére listák, regiszterek alkalmazása, *4*. osztályozások, kérdőívek készítése.
- Az elemzési célok eléréséhez specifikus információ(k)ra/hipotézis(ek)re van szükség, amely(ek) megszerzése/megfogalmazása után modellépítés következik. A folyamat zárása lehet valamilyen leíró statisztika, becslés vagy előrejelzés megadása.

BIG DATA PARADIGMA

-Az adat már ott van (mindenütt ott van)

-Adatgyűjtés

-Adat előkészítés

-Adat feltárás (korrelációk keresése)

-Az algoritmusok testreszabása

-Új tudás felfedezése / az eredmények validálása

BOTTOM UP PARADIGMA ALULRÓL FELFELÉ ÉPÍTKEZÉS

- *Big Data-megközelítés, avagy a bottom-up paradigma.* A Big Data-paradigma esetében az előzőhöz képest egészen más logikát kell követnünk. Mivel itt nincs szükség az adatgyűjtés tradicionális értelemben vett megtervezésére (hiszen az adatok már megvannak, pontosabban mindenütt ott vannak), felborul a klasszikus sorrend. A tervezés helyett ilyenkor magával az *1*. adat(be)gyűjtéssel indítunk, ezt követi az *2*. adatelőkészítés, az *3*. adatfeltárás (ami többnyire korrelációk keresését jelen- ti), *4*. az algoritmusok testreszabása (elsősorban skálázható algoritmusok választása aggregálás kerülésével), végül *5*. új tudás felfedezése/és az eredmények validálása (heurisztikus [mintakereső] technológiák használata az előrejelzésekhez/becslésekhez).
- E megközelítés esetében a hangsúly a hozzáférhető adatok felfedezésén, vagyis olyan információértékek keresésén van, amiket ezekből mások még nem nyertek ki. Nyilvánvalóan ez a logika inkább az adattudósok6 (data scientists) által vizsgált problémákra kínál megoldást, akiket sokkal inkább a „Mi történik?” kérdés érdekel, mint a „Miért?” és a „Hogyan?”. E speciális jellemzők miatt a Big Data integrálása a hivatalos statisztikába egyáltalán nem megy gördülékenyen.

GONDOK A BIG DATÁVAL

-reprezentativitás

-ismeretlen a célpopuláció

-nem tiszta a mintaegységek kiléte/miléte

-pre processing errors (mint a mérési hiba)

-social media (céltalan adatok/ robotok kiszűrése)

Nem egyértelmű korrelációk

-”az okozat halála”

-hamis korrelációk

ÉRVEK A BIG DATA MÓDSZERTANA MELLETT ÉS ELLEN

Érv	Ellenérv/Kihívás
Nincs minta	Nincs minta – reprezentativitás
Valós idejű	Lefedettség (többlet/hiány)→torzítás
Valós viselkedés, nem önbevallás	Input-/output-adatok minőségének mérése
Válaszadói tehercsökkentés	Adatforrás felhasználásának módja, potenciális validálási adatforrás elvesztése
Társíthatók más adatbázissal	Összehasonlíthatóság (jelenlegi statisztikával)
Új ismeret feltárása	IT-felszereltség, támogatás
Költségek (hosszú távon)	Költségek (rövid távon)
	Adathozzáférés
	Adatvédelem
	Stabilitás

AZ ADATFORRÁSOK JELLEMZŐI

Jellemző	Elsődleges statisztikai adatforrás	Másodlagos statisztikai adatforrás	
		Adminisztratív adatforrás	Big Data-jellegű adatforrás
Az adatok statisztikai cél(ok)ra tervezettek	igen	nem	nem
A fogalmak, a definíciók és az osztályozási rendszerek egyértelműek és ismertek	igen	gyakran	ritkán
A célsokaság jól definiált	igen	gyakran	nem
Rendelkezésre állnak metaadatok	igen	gyakran	nem
Az adatok strukturáltak	igen	igen	ritkán
Az adatok a vizsgált alapsokaságra vonatkoznak	igen	rendszerint	nem
A statisztikai adatok „kinyeréséhez” az adatok előfeldolgozása szükséges	nem	nem	igen
A lényeges/érdeklődésre számot tartó adatok közvetlenül elérhetők	igen	gyakran	nem
A segédváltozók közvetlenül elérhetők	igen	gyakran	nem
Az adatok teljes körűen lefedik a vizsgálni kívánt sokaságot	igen (cenzus) nem (survey)	gyakran	még nem
Az adatok reprezentatívak vagy adott elemzésekre reprezentatívvá tehetők	igen	gyakran	nem

Forrás: ISTAT

MIÉRT NEM MŰKÖDNEK A HAGYOMÁNYOS ELJÁRÁSOK A BIG DATA ESETÉBEN

-a számítási komplexitás határai:

Példa: inverz mátrix képzés (legkisebb négyzetek elve, maximum-likelihood via Newton-Raphson algoritmus)

-a legtöbb hagyományos algoritmust nehéz párhuzamosítani (hogy egyszerre több processzor dolgozzon a részletein) (hadoop nem tudja ezt kezelni)

-a hibás adatra / hibás szélsőrétékekre borzasztóan érzékenyek a hagyományos eljárások

-pedig a big data "zajos" és strukturálatlan (óriási adatmennyiség, nem lehet editálni, imputálni, outliert kezelni)

KÖVETKEZTETÉSEK

-a hivatalos statisztika jelen eljárásai (tervezett, modellre épülő mintavételi eljárások, regresszió elmélet, általános lineáris modellek, stb) hagyományos alapadatok specifikus tulajdonságain állnak vagy buknak

-nevezetesen : magas minőségű, de kevés adat

Ezek az eljárások:

-nagyon érzékenyek a hibás adatokra és a szélsőértékekre (ezért kötelező a hagyományos eljárásoknál az ellenőrzés, adattisztítás)

-tipikusan magas számítási komplexitással bírnak ezek az adatok (exponenciális viselkedés jellegzetes)

Szintézis: a jelenlegi statsiztikai eljárásoknak semmi közük a big data-hoz

Diagnózis: radikális paradigmaváltás kell a statisztikai metodológiában

KÖVETKEZTETÉSEK 2

Mit lehet tenni?

1. Robosztusabb eljárásokat használni, még akkor is, ha az némileg a pontosság rovására megy.
2. Ez a metódus közelítő és nem exact optimalizációs technikán alapuljon amelyek, képesek megbirkózni zajos objektív funkciókkal.
3. El kell fogadni a Big Data más típusú elemzéseket tesz lehetővé.

SANDBOX PROJEKT-2014

Shared Computing Enviroment

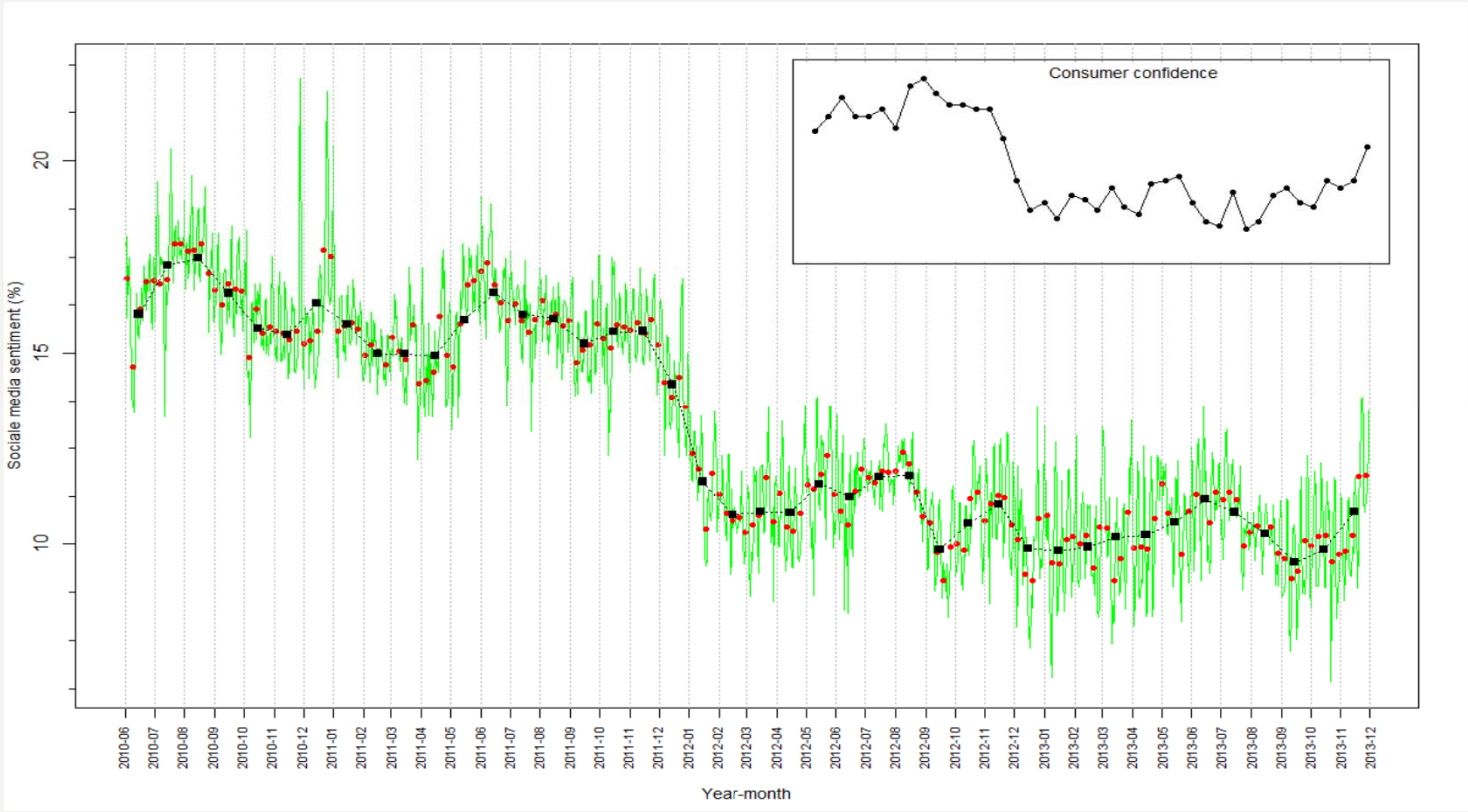
Lényege: Big Data források stabil- és ismételhető módon hozzáférhetőek, relatív könnyen és hatékonyan elérhető és ‘manipulálható’.

-a választott források, a hivatalos statsiztikák értékeléséhez általában használt minőségi kritériumoknak megfelelőek.

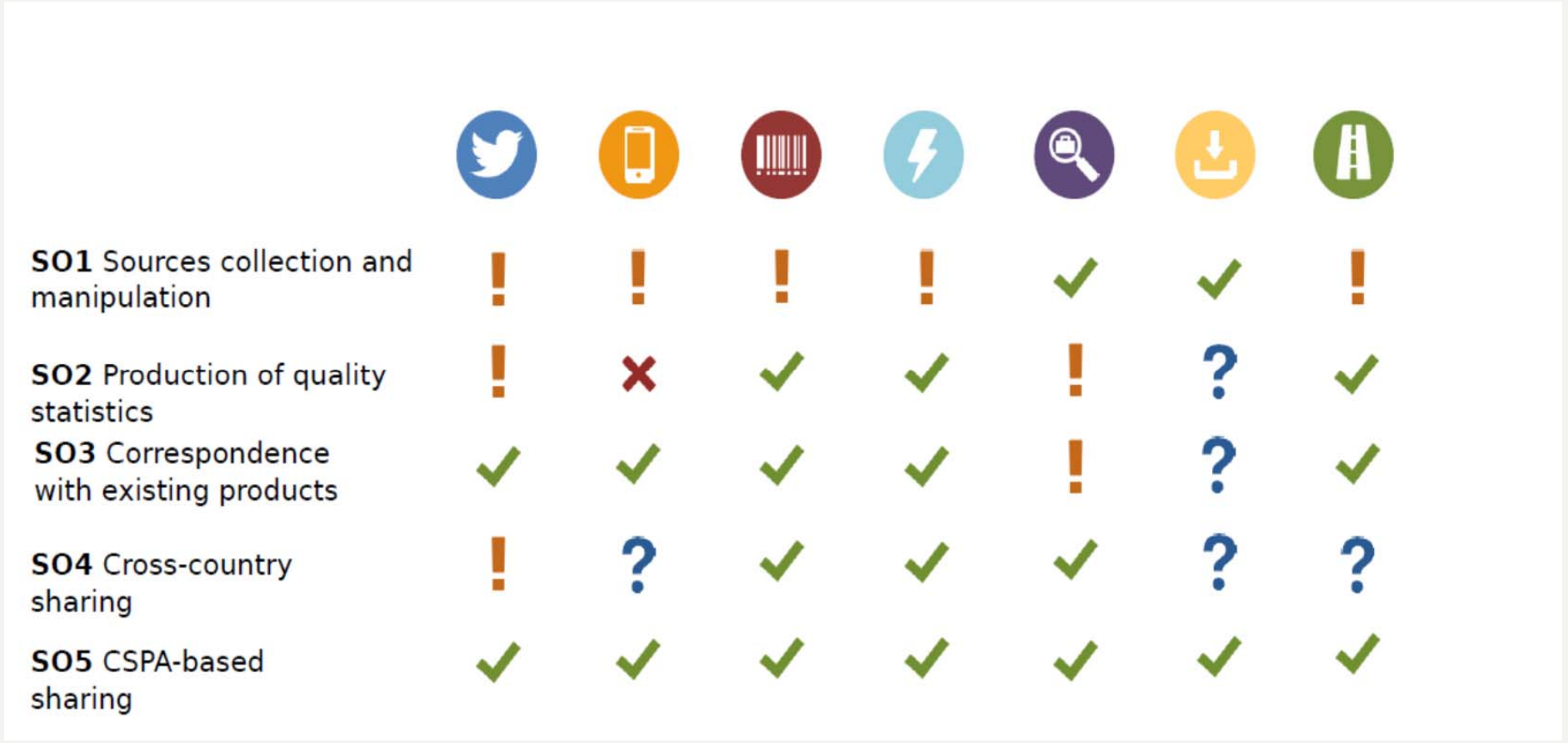
-Létező, sokak által használt adatokkal megfeleltethetőek, HBS indikátorok, árstatisztika.

-A különböző országok által használt platformok, módszerek, eszközök és adatbázisok megfeleltehetők legyenek.

-Össznépi összjáték, (módszerek megoszthatók legyenek).


Consumer confidence
Sociale media sentiment (%)
10
15
20
Year-month
2010-06
2010-07
2010-08
2010-09
2010-10
2010-11
2010-12
2011-01
2011-02
2011-03
2011-04
2011-05
2011-06
2011-07
2011-08
2011-09
2011-10
2011-11
2011-12
2012-01
2012-02
2012-03
2012-04
2012-05
2012-06
2012-07
2012-08
2012-09
2012-10
2012-11
2012-12
2013-01
2013-02
2013-03
2013-04
2013-05
2013-06
2013-07
2013-08
2013-09
2013-10
2013-11
2013-12

Forrás: ISTAT



Common Statistical Production Architecture

Forrás: ISTAT

AZ OLASZ MELÓ

2013 -2015: megállapodások kötése: NSI, Akadémia, Private Sector

Adatforrás	Domain
Online keresés	LFS
Internet-scraped Data	ICT, Árstatisztika
Mobiltelefon adatok	Turizmus statisztika
Scanner data	Árstatisztika
Social Media	Társadalomstatisztika (!)
Képek: Közlekedési webkamerák	Közlekedésstatisztika

A MIXED METHOD BIG DATA ÉRTELMEZÉSE

„ICT in enterprises”:

Kérdőív kiküldése:

1) web-scraping
2) Text mining

	Use of a Website or Home Page		
B7.	**In January 2013, did your enterprise have a Website or Home Page?** (Filter question)	Yes	No -> go to B9
B8.	**In January 2013, did the Website or Home Page have any of the following?**	Yes	No
	*[6] a) Online ordering or reservation or booking, e.g. shopping cart		
	b) A privacy policy statement, a privacy seal or certification related to website safety		
	c) Product catalogues or price lists		
	d) Order tracking available on line		
	e) Possibility for visitors to customise or design the products		
	f) Personalised content in the website for regular/repeated visitors		
	g) Advertisement of open job positions or online job application - *Optional*		

Forrás: ISTAT

EURÓPAI PÉLDÁK

Hollandia: Road sensors (Traffic loops)

–index of traffic intenstiy

-főutakon elhaladó autók száma

-230 millió record/nap

-komplex adattisztítási eljárások

Észtország: turizmus (mobiltelefon adatokkal)

-turizmus inbound/outbond

-transportation flows

-mindennapi mobilitás

-egyéni közösségi jellemzők (találkozóhelyek)

UK: Twitter adatok

-Geo located tweets

KELL EGY KOMPAKT PROJEKT: IDŐMÉRLEG

Ki?

Mikor?

Hol?

Pontosan mit?

Kivel?

Mennyi ideig?

A HAGYOMÁNYOS IDŐMÉRELGNAPLÓ ÉS AZ INNOVÁCIÓ

II. A VIZSGÁLT NAP IDŐMÉRLEGE

Online vagy nem

SWB

Párhuzamos2

A tevékenység								
Sorszáma	kezdete óra, perc	vége	pontos megnevezése	Történt a naplóba bejegyzés előtt vagy után?	Hol történt?	Mivel végezte együtt, kivel volt beszélgetésben?	Ki végezte még együtt?	Egyidejűleg végzett egyéb tevékenysége
a	b	c	d	e	f	g	h	i
1	4, 00	6.10	alvás		hálószoba			
	6.11	6.30	mosakodás		fürdőszoba			
	6.31	6.50	reggeli készítése	s. ht	konyha	férj		beszélgetés
	6.51							

Nyitott napló

BIG DATA VIZUALIZÁCIÓ

http://www.urbanmobs.fr/fr/france/