Szűts Zoltán – Yoo Jinil

Big Data, az információs társadalom új paradigmája

Bevezetés

Az információ a 21. század olaja, annak elemzése pedig a robbanómotor.
(Sondergaard 2011)

A tanulmány célja, hogy az információs társadalom szempontrendszeréből kiindulva összegző képet fessen a napjainkban rendkívüli tudományos és üzleti érdeklődés fókuszába került Big Data jelenségről, bemutassa annak definíciós kísérleteit és megközelítési módjait, valamint a feltételrendszert, melynek teljesülése esetén ma erről az új informatikai, társadalmi, kereskedelmi, kormányzati paradigmáról beszélhetünk. Reményeink szerint a tanulmánynak szerepe lehet abban, hogy párbeszéd indulhasson az általunk fontosnak és aktuálisnak ítélt témák mentén úgy, mint: a lehetséges alkalmazási területek, a hozzáférés és feldolgozás (hatalmi) problémái, vagy éppen a privacy és adatbiztonság kérdésköre.[1]

A Big Data a mai információs társadalom gravitációs magja, mivel a társadalmi, politikai és gazdasági folyamatok feltérképezését segíti elő (McNeely és Hahm 2014: 304). A Big Data egyszerre adhat magyarázatot a felhasználók fogyasztói és információs viselkedésére, nyújthat segítséget a piacok felméréséhez, javíthatja a marketing- és értékesítési kampányokat, adhat irányjelzést az árképzésnél és optimalizálhatja a logisztikai folyamatokat és az árufolyamot, menthet életeket a gyógyászatban. Hasonlóképpen elengedhetetlen az okosvárosok közlekedésében: segít optimalizálni a forgalmat és az okosautókat kiszolgálva elkerülni a baleseteket. Hatékonyabbá teheti a tanítási, tanulási folyamatokat és felgyorsíthatja a tudományos felfedezéseket (Benedek és Molnár 2013).

A Big Data jelensége azonban rendkívül megosztó. Egyszerre bír utópikus és disztópikus olvasatokkal. Egyrészt hatékony eszköz a társadalmi és a jól-léti problémák feltérképezésére és a lehetséges mintázatok és válaszok megtalálására. Segítséget nyújthat az eddig gyógyíthatatlan betegségek kutatásában, a terrorizmus elleni harcban vagy a globális felmelegedés elleni küzdelemben. Ugyanakkor számos kutató – mint azt a Big Data kritikájával foglalkozó fejezetben majd részletesen is bemutatjuk – felhívja a figyelmet a személyi adatokkal való visszaélés és a privacy megsértésének lehetőségére és az államok növekvő megfigyelési és ellenőrzési erejére (Boyd és Crawford 2012: 663–664).

A Big Data egyszerre jelent szemantikai, analitikai, adattárolási és hozzáférési kihívást, hiszen napjainkban eddig nem látott nagyságrendű adatok tárolását, feldolgozását, a rejtett és váratlan összefüggések megtalálását feltételezi. Bár a fogalomnak nem létezik

[1] Amikor a Big Data jelenségével kezdtünk el foglalkozni, a szakirodalmat az EBSCO és Google Scholar rendszerében böngészve és szelektálva rá kellett jönnünk, hogy olyan nagy mennyiségű információt publikáltak a témában, hogy jó lett volna egy szelektáló gépi algoritmust használni. Ez természetesen lehetetlen, hiszen a szelekciót magának a kutatónak kell végeznie, mivel ő jelöli ki az általa vizsgálandó problémákat és veszi észre az összefüggéseket. Egy alapszintű algoritmus azonban a jövő tanulmányírói számára szemantikus elemzés alapján – a Big Datára támaszkodva – feltárhatná akár éppen a Big Data aktuális kérdésköreit vagy a legvitatottabb kérdéseit.

egységes definíciója, tanulmányunkban a Big Data alatt a strukturált és strukturálatlan információ mennyiségének exponenciális növekedését, annak elérését és felhasználását értjük. Rögtön az elején ki kell emelnünk, hogy a fogalom leginkább az adatok feldolgozásának módjára fekteti a hangsúlyt, és a big (óriási, megszámolhatóan sok) jelzővel együtt érthető meg. Big Datáról csak feldolgozható adatok esetén beszélhetünk (Bodnár 2014). A Big Data egyik alapja az a meggyőződés, miszerint az óriási mennyiségű és különböző tartalmú adatból olyan következtetéseket lehet levonni, melyek kisebb mennyiségű adat feldolgozása során nem tűnnének fel.

Az internet demokratikus környezete, a világháló szabadon írható felülete, a hálózatra kötött eszközök számának exponenciális növekedése a digitálisan rögzített adat mennyiség robbanásszerű növekedéséhez vezetett. Minden ugyanis, ami hálózat kontextusában születik, történik, megmarad, és ezzel együtt visszakereshetővé, elemezhetővé válik. Az információs társadalom korában az internet behatol a társadalom alrendszeribe is. Hálózatra költözik többek között az üzleti élet, a politika, a kormányzás, az oktatás, a gyógyítás. Ami korábban a magánélet kitüntetetten intim, zárt köre volt, az most már az interneten kinyílik a végtelenbe a minősítésekre éhes én új terepeként (Csepeli 2015: 172-173). Az internet nem felejt, erős túlzással élve környezetében nem törlődik ki semmi. „*A hálózati térben minden kapcsolat, cselekvés, érdeklődés nyomot hagy, kutathatóvá válik*" (Dessewffy és Láng 2015: 160). Ahogy életünk mind nagyobb részét online töltjük, úgy növekszik a digitális lábnyomunk is. Annak ellenére, hogy életünk számos mozzanatában digitális rendszerekkel lépünk interaktivitásba, nem gondolunk bele, milyen sok (gyakran triviális) információt hagyunk magunk után (Zadrozny és Kodali 2013).

A témát a tudomány területéről megközelítve elmondhatjuk, hogy a Big Data lehet az elmélet, kísérlet és szimuláció központú kutatás mellett a negyedik tudományos paradigmarendszer. Ezt a trendet két tendencia támogatja. Az elsőről, az adatok generálásának exponenciálisan növekvő sebességéről már értekeztünk. A másik tendencia a tárolási és számítási kapacitás növekedése, és ezzel együtt annak költségeinek csökkenése.

A Big Data számadatai is beszédesek: míg 2000-ben a világban tárolt információ csupán negyede volt digitális formában rögzítve, addig 2013-ra ez az arány már 98%-ra nőtt (Mayer-Schönberger és Cukier 2013). A világunkban létrehozott információ 90%-a az elmúlt 2 évben jött létre, mivel naponta 2.5 trillió (10^{18}) byte információ keletkezik. Jelenleg több mint 75 millió szerver generál – és tárol – adatokat. A Google keresőóriás több millió szervere folyamatosan indexel 50 milliárd weblapot. És a publikus web alatt elhelyezkedő Deep Web adatait megbecsülni is alig tudjuk (Shroff 2014).

Ezen óriási adatmennyiség elemzése összefüggések felismerését teszi lehetővé, ami segítséget nyújthat a politikai és gazdasági döntéshozóknak is. Nem meglepő tehát, ha a Big Data nem csupán a tudományos kutatások terén (Nagy Hadronütköztető – LHC), de az üzleti világban is komoly felfedezésekhez, változásokhoz vezethet (Bessis és Dobre 2014).

Ha egyes szektorokat tekintjük át, akkor elmondhatjuk, hogy a nagy- és a kiskereskedelmi, a logisztikai, a pénzügyi vállalatok és az egészségügyi intézmények mind nagyobb mennyiségű adatot generálnak és tárolnak. A Big Data adatai különféle forrásból származnak: weblapokról; az Internet of Things (IoT) szenzoraitól; közösségi média posztokból, videómegosztókból, banki vagy vásárlási tranzakciókból, kórházi adatbázisokból, okoseszközöktől, okostelefonok vagy navigációs eszközök GPS jeleiből – hogy csak egy párat említsünk.

Mint az a fentiekből kiderül, a Big Data két fő kategóriába sorolható aszerint, hogy adatai (1) konkrét emberi aktivitásból vagy (2) kizárólag gépi forrásból származnak.

Kultúra, illetve korfüggő, hogy egy átlagfelhasználó a közösségi médiában, a Facebookon, Twitteren, Instagramon, hány bejegyzést, képet oszt meg magáról, hányszor fejezi ki véleményét hozzászólások, like-ok vagy megosztások formájában, vagy éppen hány e-mailt küld, de egyszerű megfigyeléssel is megállapíthatjuk, hogy sokat. A YouTube-ra a felhasználók percenként 300 órányi videót töltenek fel[2]. Emlékeztetésképpen: ha egy átlagember 75 éven keresztül napi 8 órán át csak videókat néz, akkor 220 ezer órányi műsort tud megnézni. Ennyi új videóanyagot töltenek fel ma a videómegosztóra fél nap alatt.

Az NFC és GPS eszközök, a hálózatba kötött szenzorok folyamatosan adatokat generálnak. Az Internet of Things (IoT) forradalmának küszöbén állunk. Különböző becslések szerint 2020-ra akár 50 milliárd eszköz lehet a hálózatba kötve.

Definíciós kísérletek

„Though this be madness, yet there is method in't."
(William Shakespeare, *The Tragedy of Hamlet, Prince of Denmark*)[3]

A Big Data kifejezés olyan óriási mennyiségű, folyamatosan érkező, különböző formátumú adatot és az azokkal való munkát jelzi, amit jellemzően a különféle hálózatokon lévő gépek és az emberek közösen állítanak elő, és amelyeket a korábbi módszerekkel nem lehetett feldolgozni.

A Gartner kutatóintézet 2001-ben közölt definíciója szerint a Big Data nagy mennyiségű, sebességű és eltérő formátumú információt jelöl, melynek feldolgozása új típusú megközelítést kíván annak érdekében, hogy az így született eredmények segítsenek a hatékony döntéshozatalban, összefüggések felfedezésében és folyamatok optimalizálásában[4].

Fontos azonban megjegyezni, hogy maga a Big Data kifejezés már korábban is felbukkant az akadémiai környezetben, és máig folyik a vita, hogy ki használta először a kifejezést. John R. Mashey *Big Data ... and the Next Wave of Infra Stress* című, több helyen is elmondott, de alapvetően egyetemi előadásában a Stanfordon már 1998-ban említi a Big Datát[5] az általunk leírt jelenség megnevezésére, egyik korai prezentációja pedig még mindig elérhető.[6] Ugyanebben az évben, Weiss és Indurkhya (1988) is használják már a Big Data kifejezést az informatikában, illetve Diebold a statisztikában (2000).

A Big Data együttese magába foglalja a korábban soha nem látott mértékű és változatos forrásból érkező adatok rögzítését, feldolgozását, elemzését, megosztását, illetve az eredmények vizualizálását. A gravitációs mezőjébe tartozó adatok mennyisége meghaladja az általában használt adatrögzítő és feldolgozó szoftverek képességeit.

A legelterjedtebb definíció szerint (Laney 2012) a Big Datát három V jellemzi: mennyiség (*Volume*), sebesség (*Velocity*) és változatosság (*Variety*).

[2] https://www.youtube.com/yt/press/statistics.html
[3] *Őrült beszéd, őrült beszéd: de van benne rendszer.* (William Shakespeare, *Hamlet, Dán Királyfi*)
[4] http://www.gartner.com/it-glossary/big-data/
[5] http://web.stanford.edu/class/ee380/9798sum/lect06.html
[6] http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Volume: A mennyiség a másodpercenként előállított hatalmas adatözönre vonatkozik. A múltban a nagymennyiségű adat tárolási problémákat okozott. A jelenben a tárhely mérete és a tárolási sebesség növekedett, illetve ezzel együtt azok költsége csökkent. A Big Data esetében azonban új kihívásokkal kell szembenéznünk: hogyan lehet osztályozni, fontossági sorrendbe állítani az adatokat, hogyan lehet összefüggéseket észrevenni – értéket létrehozni.

Velocity: A sebesség azért fontos kérdés, mert az adatok nem nagy blokkokban jönnek, hanem folyamatosan áramolnak. Mind gyorsabban és gyorsabban kell őket feldolgozni, és lehetőleg valós időben, hogy releváns tudáshoz juthassunk.

Variety: Mégis talán a legnagyobb kihívást a változatosság jelenti, mert az egyes adatokat strukturálni kell és egymással összefüggésbe hozni, a forrásra való tekintet nélkül. A cél a kontrollálatlan adatfolyamok formázása az értékes információk kinyeréséhez[7]. A változatosságra jellemző, hogy egyszerre érkeznek adatok hagyományos adatbázisokból, szöveges dokumentumokból, videómegfigyelő rendszerekből, e-mailekből, tömegközlekedési járművekből, repülőgépek motorjaiból, telefonhívásokból, és az elemző rendszereknek összefüggéseket kell felismerniük (Majkić 2014).

Az alliteráló 3 *V* mellé számos szerző további *V*-ket helyez. Ezek lehetnek, a *Variability* (az adatok variálhatóságát jelöli), a *Virtual* (az adatok virtuális voltát jelzi), a *Veracity* (az adatok integritását jelöli) vagy a *Value* (az adatokban rejlő hasznosságot jelöli) (Zikopoulos et al. 2011) ahol az utóbbi kettő gyakran felcserélhető.

A több szempontra való figyelemfelhívás miatt fontosnak találunk két további definíciót is röviden ismertetni:

A Big Data olyan technikákat és technológiákat jelöl, melyek az extrém skálán mozgó adatok kezelését gazdaságossá teszik (Hopkins és Evelson 2011). Egyszerűbb definíciót kínálnak a McKinsey Global Institute kutatói: A Big Data olyan adattömeget jelöl, mely mérete túlmutat a hagyományos adatbázis szofverek tároló, kezelő és elemző képességein.[8]

Történeti előzmények

A személyi számítógép megjelenése végérvényesen megváltoztatta az adatokkal való bánásmódot (többek között a statisztikát is – hogy példát említsünk). Minden nehézség nélkül térképezhetők fel segítségével társadalmi folyamatok – csupán a megfelelő kérdéseket kell feltenni (Ratner 2004: 1). A Big Data előtti digitális paradigmára jellemző adatbányászat kifejezés az 1970-es évek végén, az 1980-as évek elején született. Talán ez lehet az oka, hogy az adatbázisok elemzésével foglalkozó marketingesek számára például az általunk tárgyalt Big Datra jellemző minták és összefüggések felfedezése nem hat teljesen az újdonság erejével (és varázsával) (Ratner 2004: 9–10).

De mi változott? A magányos gép helyébe a hálózatba kötött lépett, és ma már egy rendszerbe tudjuk integrálni, használható formátumba konvertálni és a számítási teljesítmény segítségével kielemezni a digitálisan rögzített az adatokat (számadatot, szöveget,

[7] https://www.it-services.hu/hirek/mi-az-a-big-data/

[8] http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Business%20Technology/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

képet, hangot, videót), és végeredményben eddig ismeretlen összefüggések tucatjaira is fel tudjuk hívni a figyelmet, mint azt az egyes ágazatok tárgyalása során majd látni fogjuk. Új szintre lépett a kereshetőség. „A digitális jel korlátlanul időtálló – évekkel ezelőtti olvasási szokásaink éppúgy kereshetőek, mint a jelenlegeik. A digitális adatbázisok összekapcsolhatóak és kereshetőek, az újságolvasási szokásokat kombinálhatjuk vásárlási adatokkal" (Dessewffy és Láng 2015: 158).

A Big Data megjelenéséhez vezető út és feltételek

A Big Data nem előzmények nélküli. Annak ellenére, hogy a jelenben az IT az üzleti és az tudományos diskurzusok egyik kulcsszava, látens formában már egy ideje jelen van. A különbség csupán annyi, hogy míg korábban egy-egy szuperszámítógép volt képes nagymennyiségű adatok elemzésére, ma ugyanez a lehetőség biztosított számos felhasználónak és vállalatnak is (Joshi 2015).

Ha történeti előzményeit keressük, akkor elmondhatjuk, hogy Vannevar Bush 1945-ben publikálta az Atlantic Monthlyban a mai hálózatba kötött számítógép elődjével, a Memex elvével foglalkozó értekezését (Bush 1945). A Memexnek az lett volna a szerepe, hogy segítse az addig felhalmozódó (nyomtatott) információmennyiségben történő eligazodást és egyes fogalmak közti kapcsolatok feltérképezését. Az igazi újdonság az adatok tárolása volt, mely a hierarchikus rendszerekkel szemben az emberi asszociációkhoz hasonlóan történt volna – címkézéssel. Bush elképzelése a kornak megfelelően még mechanikus volt, a Memexet egyfajta gépiesített, analóg magánkönyvtárként képzelte el (Szűts 2013: 49-50), elmélete azonban a Big Data eljövetelét vetítette elő, egy rendszerét, mely hatalmas mennyiségű adatot nagy sebességgel képes kezelni és – emberi segítséggel – összefüggéseket észrevenni. Ezt követően az internet megjelenése, Tim-Berners Lee hiperlinkekkel átszőtt világhálója, majd a rendszert indexelő keresőprogram, a Google mind újabb lépést jelentett a Big Data felé. Míg az internet és világháló esetében a tudományos megismerés utáni vágy volt a fejlesztés motorja, addig a keresőprogramok fejlődésében már döntő szerepet játszottak az üzleti célok is, a hatékony hirdetési rendszer kidolgozása (Shroff 2014).

A Big Data építménye

A Big Data összeségében több elemből épül fel, áll össze. Egyaránt hívta életre a felhasználók Web 2.0-ás környezetben kifejtett online aktivitása, mely digitális lábnyomot hagy, de folyamatosan szolgáltatnak adatokat a környezetünkbe mind nagyobb számba beépülő szenzorok is. Ezen adotokat hálózatok továbbítják és adatbázisok rögzítik. A digitális rögzítés azonban nem a jelen folyamataira érvényes, a múltat is folyamatosan digitalizáljuk. Ahhoz azonban, hogy mindezen adatmennyiség sikeres feldolgozására és mintázatok felismerésére vállalkozhassunk, szükség van olyan speciális szoftverekre, melyek a fejlett gépi tanulás rendszerére támaszkodnak. Végezetül, a Big Data már nem csupán a professzionális felhasználók kiváltsága, a tárhely és számítási teljesítmény növekedésével és ezek árának csökkenésével a mindennapi felhasználók is részeseivé válhatnak a Big Data paradigmájának (1. ábra).

Digitális lábnyom
Az információs társadalom jelenlegi fejlettségi szintjén a felhasználók életének számos mozzanata (keresés, böngészés, levelezés, csevegés, megosztás, értékelés, hozzászólás stb.) már online, a hálózat figyelő és mindent rögzítő szemei előtt zajlik.

Számítógépes, hálózatba kötött adatbázisok
A magányos gép mítosza már a múlté. A mai rendszerek különböző formátumban rögzített adatokat tárolnak és dolgoznak fel, és ami a legfontosabb, hálózatba vannak kötve, tudásuk egységes tárházat alkot.

Múltat átmentő digitalizálás
Talán meglepő, hogy a múltba tekintés és digitalizálás a jövő problémáit oldhatja meg, például az Old Weather[9] crowdsourcing projekt keretében az Egyesült Államok haditengerészetének fedélzeti naplóit rögzítik digitális formában, így a meteorológusok Big Data környezetben vizsgálhatják a múlt időjárását és következtetéseket vonhatnak le a jövővel kapcsolatban.

A mindenhol jelenlévő számítástechnika és szenzorok
Mára elterjedt az ubiquitous computing, vagyis a mindenütt jelenlévő számítástechnika jelensége. Ezt az új paradigmát Weiser (1991) szerint az jellemzi, hogy a számítástechnika és a digitális eszközök oly módon beépültek a hétköznapi folyamatainkba, hogy már észrevétlenek maradnak, és úgy használjuk őket, hogy nem tanúsítunk ennek a ténynek jelentőségét, mivel egy automatizált folyamat részévé váltak. Hasonlóképpen, a jelen okosvárosaiban szenzorok milliói gyűjtenek adatokat az energia hálózatokkal, közlekedéssel kapcsolatban (Yoo 2014).

Gépi felismerés és machine learning
A gépi tanulás (machine learning – ML) (Samuel 1959) fontos feltétele a Big Datának, hiszen lehetővé teszi, hogy a számítógépek tanuljanak – mintákat vegyenek észre – anélkül, hogy konkrétan erre programozták volna őket. A gépi tanulásnak köszönve az elemzés során a számítógépek közvetlenül az adatokból jutnak ismeretekhez és oldanak meg problémákat. Ezen esetek többségében természetesen a számítógépeket embereknek kell tanítaniuk, az adatokat kezdetben nekünk kell megcímkéznünk és osztályoznunk, hogy később, e minta alapján, a gépek önállóan is képesek legyenek tanulni és elemezni az információkat. Abban az esetben például, amikor a gépek a közösségi médiában éppen a legvitatottabb témákat ismerik fel, ilyen emberi felügyeletet igénylő tanulásra már nincs szükség. A rendszerek ebben az esetben már maguktól tanulnak, képi elemek felismerésére azonban alkalmatlanok (Condliffe, AI Is Learning… 2016).

A számítási teljesítmény növekedése, és a tárolási kapacitás árának csökkenése

Moore még 1965-ben gyakorlati tapasztalataira alapozva mondta ki a törvényt, mely szerint az integrált áramkörökben lévő tranzisztorok száma (és ezzel együtt a számítási teljesít-

[9] www.oldweather.org

mény) másfél évente megduplázódik (Moore 1965). A kevésbé ismert, de hasonlóan fontos Kryder-törvény szerint (idézi Walter 2005) a tárolási költségek árának csökkenése még a számítási teljesítmény növekedésénél is nagyobb mértékű. A történelem során először vált elérhetővé megfizethető formában a mindennapi felhasználók számára a nagy számítási teljesítmény és az olcsó, alapvetően felhő alapú tárhely.

A Big Data céljainak megfelelően kifejlesztett rendszerek

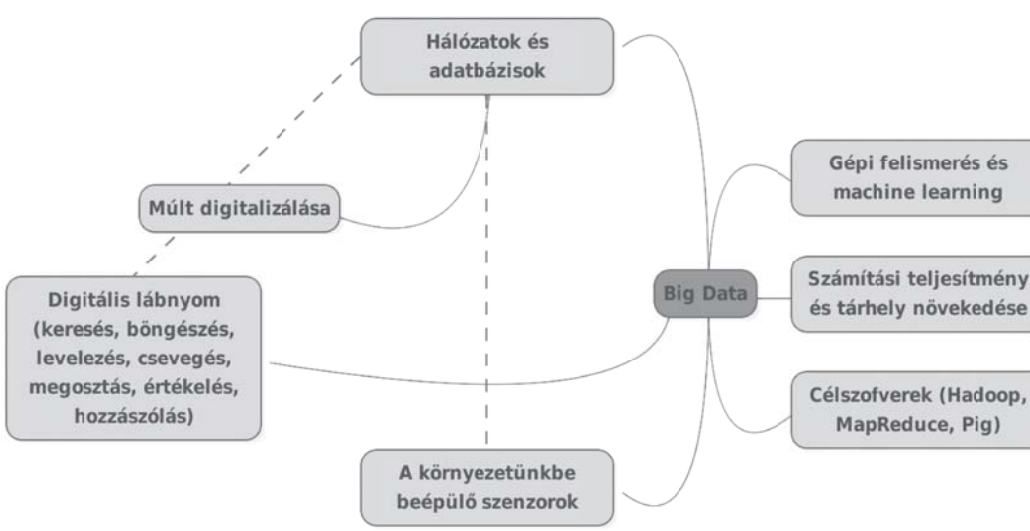
A Big Data a korábbiaktól eltérő szoftveres megközelítést kívánt, így születtek meg az ilyen nagymennyiségű adat tárolását, feldolgozását, elosztását biztosító programok, programnyelvek és futtatókörnyezetek. A teljesség igénye nélkül kiemeljük a legismertebbet:

A Big Data környezetében használt Hadoop egy nyílt forráskódú keretrendszer, amely adat-intenzív elosztott alkalmazásokat támogat. Legfőbb jellemzője, hogy nagy mennyiségű és ezzel együtt alacsony költségű, a mindennapi életben használt hardverből épített szerverfürtök létrehozását teszi lehetővé.

A Hadoophoz hasonlóan a MapReduce szintén nagy adathalmazok feldolgozására képes párhuzamosan és szerverfürtön elosztottan. A MapReduce egyszerre végez szűrést és rendezést, és végül összegzi az eredményt. A Google ezt az algoritmust használta a világháló indexelésére.

A Pig programnyelvet és futtatókörnyezetet eredetileg a Yahoo! fejlesztette ki, hogy könnyebbé tegye a Hadoopot használóknak a nagy adatmennyiség elemzését azzal, hogy kevesebb időt kellett tölteniük programozással. A Pig elnevezés metonimikus, hiszen, mint a valódi disznó, mely gyakorlatilag mindenevő, a Pig programnyelv is szinte minden típusú adattal megbirkózik (vagyis ez a 3. V).

És végül a Zookeeper a Hadoop klaszterek koordinációját végzi, dokumentálja, megnevezi és szinkronizálja szolgáltatásait.



1. ábra: A Big Data építménye (saját szerkesztés)

Az információs társadalomra jellemző szektorok, melyek profitálhatnak a Big Datából

Már a Big Data megjelenése előtt is a meteorológia, biológia, fizika, kémia és csillagászat kutatási módszerei közé tartozott a nagy mennyiségű adatok elemzése. Ez a megközelítés azonban a felsorolt szektorokon kívül nem volt jellemző. A Big Data megjelenésével a digitálisan behálózott világ lakói mindennapi életének számos területén paradigmaváltás indult meg.

Kereskedelem, média

Az Amazon online áruház vagy a Netflix videotéka ajánló rendszere Big Data analitikán alapul, hasonlóan, ahogy a Walmart is ezzel a módszerrel azonosítja egyes felhasználók kedvenc termékeit,és ennek megfelelően tölti fel raktárait. A Social Genome Big Data program lehetővé teszi az áruház számára, hogy elérje vevőit, vagy azok ismerőseit, akik érdeklődtek bizonyos termékek után online. A Walmart ugyanis ilyenkor a rendszere által relevánsnak ítélt információval, és személyre szabott kedvezménnyel keresi meg őket. Hogy képes legyen erre, a Social Genome összekapcsolja a világhálón található publikus és a közösségi médiában megjelentett információkat a vevői vásárlási adataival, illetve kontakt információval. Emellett azt is felismeri a szövegkörnyezetből, ha valaki csupa kisbetűt használva Ice Cube-ot, az énekest, vagy a jégkockát (ice cube) említi. Az eredmény végül egy folyamatosan frissülő tudásbázis, mely több millió kapcsolatot és tételt tartalmaz. Hasonlóképpen, az ugyancsak Walmart által kifejlesztett Shoppycat képes a Facebook felhasználók számára termékeket ajánlani az ismerőseik érdeklődési köre és hobbija alapján, hogy megtalálja a számukra legjobb ajándékot. Majd ezután a felhasználók geolokációja segítségével a legközelebbi áruházba irányítja őket, ahol a termék éppen kapható, ugyanis a raktárkészlet is része a vállalat Big Data rendszerének (Walmart Is Making Big Data Part Of Its DNA).[10]

Dél-Koreában a kereskedelmi láncok Big Data rendszerei a női vásárlókra fókuszálnak, ugyanis a Shinsegae áruház, a Shinhan Card és a Lotte Tour utazási iroda az adatok gyűjtése és elemzése során növelni a vásárlási hajlandóságot (Kim 2015).

Koreánál maradva elmondhatjuk, hogy a rendkívül behálózott társadalmú és hatékony e-közigazgatással bíró ország élen jár a Big Datában rejlő potenciál kiaknázásában. Az államilag indított és támogatott projektek részeként például a *Bagoly* éjszakai buszrendszer keretében Szöul közlekedési vállalata az adatok elemzésével egy csupán 8 vonalból álló rendszerrel az utazni vágyok 49%-át képes szállítani a 13 milliós metropoliszban. Ezen kívül folyamatban van a taxirendszer hatékonyabbá tétele a Big Data segítségével, illetve a gyalogos biztonságának, vagy éppen a kerületi hírdetési rendszerek sikeresebbé alakítása is cél.

Végül pedig egy példa a médiából, a Global Data on Event, Location and Tone (GDELT[11]) adatbázisban 1979-től gyűjtik médiában megjelenő, a világot érintő geokódolt események adatait. A GDELT így ma már arra használható, hogy a globális társadalmak rendszereit és viselkedését feltérképezzük.

[10] https://datafloq.com/read/walmart-making-big-data-part-dna/509
[11] www.gdeltproject.org

Okosvárosok, környezetvédelem, biztonság

Az okosvárosok a Big Datát a közbiztonság növelésére, a víz és energiaellátás, a kormányzás, a közlekedés és az egészségügyi ellátás hatékonyabbá tételére használják (Yoo 2014: 28). Már több évszázada fennálló városaink is mesterséges ökoszisztémákká alakulnak, behálózott, intelligens és digitális rendszerekké válnak. E rendszerek tervezői csupán a városok digitális ütőerére kell hogy helyezzék az ujjukat annak érdekében, hogy az adatok elemzésével élhetőbbé tegyék a települést és jobbá annak lakóinak életét.

A jelenben az ipari felszerelések többségében már szenzorok vannak, melyek óriási mennyiségű adatot rögzítenek és továbbítanak. Egy gázerőmű turbinájának lapátjai egyenként napi 520 gigabyte adatot generálnak, és egy turbinában 20 lapát van. Az interkontinentális repülőjáratok több terabyte-nyi adatot továbbítanak. Mindezen információ feldolgozása segít biztonságosabbá és költséghatékonyabbá tenni a rendszereket (Zadrozny és Kodali 2013: 4).

A Global Forest Watch[12] több millió műholdképet dolgoz fel annak érdekében, hogy valós időben készítsen becsléseket az erdőirtásokról, és ez a megközelítés minden korábbinál pontosabbnak bizonyult.

A legizgalmasabb példa mégis talán New Yorkból származik. 2012-ben a város környezetvédelmi hivatala a Big Data segítségével találta meg a csatornarendszer eltömítéséért felelős éttermeket. A New York-iak már évek óta szenvedtek ugyanis a sütőolajjal bedugított lefolyók környezetkárosító hatásaitól. A hagyományos módszerek szerint az ellenőrök szúrópróba szerű ellenőrzést végeztek. A Big Data környezetében a város informatikusai összekapcsolták a szemétszállító magánvállalkozások adatbázisait az éttermek számláival és földrajzi adataival. Az így kapott eredmények alapján a felderítési ráta 95%-ra nőtt (Feuer 2013, idézi Krishnamurthya és Desouzab 2014: 166).

Orvostudomány, egészséges életmód

Az elmúlt évtizedben az egészségügyben felértékelődött az adatbázisok szerepe. A digitális képalkotás, a digitálisan tárolt egészségügyi adatok, majd az utóbbi időben a szenzorok által küldött információk forradalmának lehetünk tanúi. A hagyományos módszereket felváltja a személyre szabott, prediktív és participatív Big Data paradigma. A jövő klinikai kísérletei már nem kis mintára korlátozódnak majd, hanem a mintában gyakorlatilag mindenki, aki a rendszerben szerepel, benne lehet.

Az egyik legnagyobb forradalom a jelenben tehát az egészségügyi állapot, az életfunkciók valós idejű monitorozásáról szól. A következő generációs egészségügyi rendszerek a nap minden percében monitorozzák majd a betegek (és egészségesek) állapotát és kezeléseket, beavatkozásokat javasolnak (Timur és Son 2014: 315). Ennek előfutárai az okosórák és fitnesz karkötők. A jelenben ezen eszközök és okostelefonjaink számolják a megtett lépéseinket, mérik a pulzusunkat, figyelik az alvásciklusunkat, segítik számon tartani a folyadékbevitelünket. A jövőben arra keresik majd a választ, hogyan érezzük magunkat a kemoterápia után, vagy hogyan halad előre valamely betegségünk. A Big Data rendszerek már most is a szenzorok adatait elemezve segítenek a koraszülöttek védelmében,

[12] www.globalforestwatch.org

ugyanis algoritmusaik a csecsemők szívverése és légzésmintája alapján 24 órával előre meg tudják jósolni egy fertőzés kialakulását anélkül, hogy erre bármilyen külső tünet utalna. A jelenben már működő legegyszerűbb rendszerek szenzorai a monitorozó egységek, a szívverésben, légzésben, testhőmérséklet változatban bekövetkező változások elemzése képes felismeri az olyan anomáliákat, melyeket a tapasztalt ápolók vagy orvosok sem tudnak.

Hasonlóképpen a Big Data képes segítséget nyújtani a járványok előrejelzésében is különböző adatbázisok használatával (www.promedmail.org, www.healthmap.org, www.google.org/flutrends). Bár tanulmányunk végén részletesen is kitérünk majd a Big Data kritikájára, itt is ki kell emelnünk, hogy a járványok előrejelzésében paradigmaváltónak titulált Flu Trends projekt végül sikertelenül zárult. A Google Flu Trends Big Data algoritmusai eredetileg az influenzára jellemző tünetekkel kapcsolatos internetes felhasznólói keresések kapcsán jelezték elő az influenza (és egyes országokban a Dengue-láz) járványok kezdetét és időbeli lefolyását. A rendszer logikája egyszerűnek, mégis úttörőnek tűnt, a jelen információs és mobiltársadalmában, aki betegnek érzi magát, online keres rá a tünetekre, gyógymódokra, orvosságokra okostelefonján, táblagépén vagy számítógépén. A Flu Trends e kiindulópontra alapozó algoritmikus előrejelzései több éven át folyamatosan megelőzték az Egyesült Államok erre szakosodott hivatala, a Center for Disease Control (CDC) jelentéseit.Több éven keresztül a rendszer megbízhatóan működött, egészen addig, míg közösségi médiában megjelenő influenzával kapcsolatos hírek annyira nem kezdtél el befolyásolni a kereséseket, hogy 2013-ban a becslések és előrejelzések pontatlanná váltak (Lazer és Kennedy 2015). Az influenzával kapcsolatos félelmek ugyanis virálisan terjedtek, és a betegségről szélesebb körben értesülnek a felhasználók a Facebookon, akik akkor is rákeresnek a tünetekre, ha náluk nem jelentkeztek, tehát nem betegek.

Oktatás

Siemens (2014) szerint az oktatás sikere a nagy és komplex adatbázisok interdiszciplináris feldolgozásán múlik. Más szerzők szerint a Big Data használata a felsőoktatásban két célt szolgál: az oktatás hatékonyságát, és a költséghatékonyságot (Altbach et al. 2009).

A hallgatók visszajelzéseire is koncentráló tanítási megközelítés már egy ideje elfogadott, így az intézményekben a tanulók értékelik a tanárokat, az általuk alkalmazott módszereket, azok hatékonyságát, véleményt mondanak egyes tantárgyakról (Benedek és Molnár 2013). Gyakran azonban az ilyen jellegű kérdőíves felmérések torz vagy nem teljes, nem valós válaszokat adnak. Így a közösségi médiában megjelenő, a tanítás és tanulás témáját érintő (alapvetően őszinte) posztok Big Data alapú sentiment analysis-e arra keresi a választ, hogyan tehetnénk hatékonyabbá a tanítást.

Politikai trendek felismerése

A társadalmi és politikai elégedetlenség korában (Arab tavasz, különböző tüntetések) sokan szeretnék tudni, hogy mit is gondolnak a lakosság egyes csoportjai, hiszen a tömegkommunikációs eszközök híreinek vizsgálata erre nem alkalmas, mert ott hatalom szempontjai érvényesülnek. A média döntő hatását feltételező framingelmélet szerint a média a politikai és a gazdasági elitek ellenőrzése alatt áll, így annak elemzése nem hoz válaszokat a feltett kérdésekre (Herman és Chomsky 1998; Bajomi 2006: 82).

Ahhoz, hogy megtudjuk, egy adott közösségnek mi a véleménye, a közösségi média tartalmak olvasása lehet az egyik sikeres stratégia. Joggal merül fel azonban a kérdés, hogy egy elemző, de akár egy tucat emberből álló csoport is hány Facebook posztot, Twitter üzenetet tud elolvasni. Elolvashatja a véleményvezérek bejegyzéseit, az azokra adott válaszokat, de semmiképpen sem annyit, hogy teljes képet kapjon a közhangulatról. Ezzel szemben a Big Data elemzi és csoportosítja a véleményeket aszerint, hogy pozitívak, negatívak vagy semlegesek. Így hírtelen hallhatóvá válik számos csoport hangja (Shroff 2014: 74). A Feeltiptop (http://feeltiptop.com) például a mindennapi felhasználók számára is lehetővé teszi a közösségi médiában megjelenő posztok alapján egy adott személy elfogadottságának mérését.

Aktuális társadalmi témák

A társadalmat foglalkoztató témák is kiolvashatók a közösségi médiából. A Big Datára ebben az esetben azért van szükség, mert a témák nem azonosak a leggyakrabban megjelenő címszavakkal, hanem szemantikai elemzésre van szükség, melyet a gépi tanulásra alkalmas rendszerek végeznek. Természetesen lehetőség van ország specifikus, vagy globális témák feltérképezésére is, az utóbbi esetben további nehézséget okozhat a 3. V, mely most a különböző nyelvek formájában jelenik meg.

Kérdések és kihívások

Ha a Big Data legfontosabb kérdéseit és kihívásait szeretnénk összegezni, akkor a sort azzal kell kezdenünk, hogy kinek a birtokában van a sok adat? Ki férhet hozzá?

Mint azt bevezetőnkben már tárgyaltuk, és a mottóval is utaltunk rá, az érték, melyen a jövő világa nyugszik majd, az információ lesz. Ez az információ különböző forrásokból, különböző tulajdonosoktól származik: állami és magánszektorból; kormányoktól, vállalatoktól, intézményektől, magánszemélyektől. Több száz ország bankrendszeréből, mobilhálózatából, rendőrségi nyilvántartásából, az e-kormányzat adataiból, vásárlásokból, közlekedési eszközökből, és a sor még folytatható. A kérdés tehát: kinek a birtokában vannak az adatok, ki az „aki figyelheti", ki az, aki megtekintheti, elemezheti őket, és felhasználhatja az eredményeket. A Facebook algoritmusa nagyrészt zárt a társadalomtudósok, hálózatkutatók előtt, a teljes képet a kapcsolatrendszerekről és felhasználók közti interakcióról ők külső szemlélőként nem láthatják. De óriási a különbség a társadalom tagjai között is abban a tekintetben, hogy ki milyen mennyiségű adathoz férhet hozzá (Moorthy et al. 2015: 75).

Az egészségügyi információink talán a legérzékenyebbek, már a jelenben is szigorúan van szabályozva, ki kezelheti őket, alapvetően az állami szektor által folyik a gyűjtésük, elemzésük azonban már gyakran a magánszektorban történik (Yoon 2015: 7). És annak ellenére, hogy a Big Data egyik lényege épp a nagymennyiségű adatban rejlik, elméletben annak sincs akadálya, hogy az információk alapján konkrét személyeket azonosítsanak be (ahogy azt például a Walmart teszi célzott hirdetései során), bár ennek megakadályozását minden esetben hangsúlyozzák, például a Google mesterséges intelligencia rendszerének, a Deep Mindnak az orvostudományi célokra történő használata során (Condliffe, Deep Mind's… 2016).

Az adatok vándoroltatása fordítva is zajlik, a magánszektorból az államiba. A személygépkocsik GPS adatait kombinálják a tömegközlekedési eszközök által rögzített információkkal annak érdekében, hogy hatékonyabbá tegyék a közlekedést. Ez a migráció azonban nem mindig zökkenőmentes, így a nyilvános adatok kezeléséről szóló szabályozást is át kellett például gondolni Koreában (Lee et al. 2013).

De annak kapcsán, hogy kinek a birtokában vannak az adatok, és ki férhet hozzájuk, ki kell térnünk arra az esetre is, amikor az államok között létrejött Big Data kapcsolat segítheti a terrorizmus elleni harcot. Ez a fajta együttműködés azonban elég nehézkes, mivel számos esetben az Egyesült Államok rendvédelmi szervezetei között is akadozik az információ megosztás.

Látni kell, hogy a jelenben már nem az a modell működik, melyben a központi kormányzatok vagy a jelentős múlttal rendelkező nagyvállalatok tárolják és dolgozzák fel az adatokat, hanem sokkal inkább a hálózat korábban létrejött IT cégek: a Facebook, a Twitter, a LinkedIn, a Google és mások. A Big Data esetében is velük kell megegyezniük a kormányzatoknak, gyakran egyéni megállapodások keretében. Számos példa volt azonban a közelmúltban is, amikor az adatokat minden előzetes figyelmeztetés nélkül kiadták az állami szerveknek. A felhasználói adatokból építkező Twitter azonban az online közösség nyomására változtatott a gyakorlatán, és jelez a felhasználó felé, amikor az adataikhoz való hozzáférést kérnek a kormányzatok (Barbash 2015).

A közhiedelemben él azon elképzelés, miszerint minden adat, amit Facebookra a felhasználók feltöltenek, a közösségi oldal tulajdonává válik. A helyzet azonban ennél árnyaltabb, és érdemes magunk elé idézni a Facebook nyilatkozat, melyet minden felhasználó az oldalra való regisztrációjával automatikusan elfogad:

„Az összes tartalomnak és információnak, melyeket közzétett a Facebookon, Ön a tulajdonosa, az adatvédelmi és az alkalmazás beállítások segítségével szabályozhatja azok megosztását. A szellemi tulajdonjogok hatálya alá tartozó tartalmak vonatkozásában, mint a fényképek és videók (IP tartalom) – az adatvédelmi és alkalmazás beállítások figyelembe vétele mellett – különösen az alábbiakhoz adja hozzájárulását: nem kizárólagos, átruházható, allicencbe adható, jogdíjmentes, az egész világra érvényes licencet nyújt részünkre bármely Ön által közzétett, vagy a Facebookkal kapcsolatos, szellemi tulajdont képező tartalom felhasználásához (IP Licenc). Ez az IP Licenc megszűnik, ha szellemi tulajdont képező tartalmát vagy felhasználói fiókját törli, kivéve, ha az Ön által közzétett tartalmat másokkal megosztotta, akik azt nem törölték”[13]. Ha a „továbbá”-t „azonban”-nal helyettesítjük, akkor világossá válik, hogy ezzel egyszerre a felhasználó minden általa feltöltött infomációhoz hozzáférést ad a Facebooknak, és (Big Data) partnereinek.

Ki képes feldolgozni az adatokat?

Az állami és magánszektor Big Datája olyan vállalatok vonzáskörébe tartozik, melyek a közvélemény számára alig ismertek. A kereskedelem modern igényeit kiszolgáló Axciomot például az óriásvállalatok közé sorolják, melyről a Facebookkal vagy például a Teslával szemben a mindennapi felhasználók többsége még sohasem hallott.

[13] https://www.facebook.com/legal/terms

Hasonlóképpen a Lexis Nexis Risk Solutions adatgyűjtő és elemző vállalat az USA-ban gyakorlatilag magánhírszerzési tevékenységet folytat a kormányzat és ipari szereplők megbízásából oly módon, hogy a Big Data megoldásai a magán- és állami szektorban elérhető információkat dolgozzák fel.

Dél-Koreában az Élelmiszeri, Földművelésügyi, Erdészeti és Halászati, illetve a Közigazgatási és Biztonságügyi Minisztériumok Big Data projektet indítottak annak érdekében, hogy visszaszorítsák a patás állatok közt terjedő száj és körömfájást. Ennek érdekében kombinálták a betegséggel kapcsolatos külföldi, vám és bevándorlási, a farmokról származó adatokat az élőállomány és az állattenyésztők migrációs adataival. Ezen felül pedig a koreai Bioinformációs Központ egy nemzeti DNS menedzsment rendszert fejleszt, mely a jövő személyre szabott gyógyászati lehetőségeit biztosítja majd a lakosságnak (Kim et al. 2014: 84).

Etikai, privacy és surveillance kérdések

A Big Data egyik legvitatottabb problémaköre a privacy kérdése. A személyes adatok gyűjtése ugyanis már a kezdetek óta összefüggött a technológia társadalomra gyakorolt hatásának növekedésével (Garson 1988). Így nem meglepő, hogy számos szerző a privacy lehetséges megsértésére és visszaélésekre figyelmeztet, a legszkeptikusabbak pedig, mint az a következő fejezetben részletesebben is kifejtük, gyakran egyenlőségjelet tesznek a Big Data és a totális megfigyelés közé.

Ahogy a Big Data előzményeit is jóval a fejlett számítástechnika kora elé helyeztük, úgy az információs technológiák magánéletre gyakorolt hatását is a modern információs társadalom megjelenése előtt kell keresünk. Már 1890-ben, a modern tömegkommunikációs eszközök megjelenésekor Warren és Brandeis a következő megfigyelést tették: „amit ma otthonunk magányába elsuttogunk, azt holnap a tetőkről fogják kiáltani' (Warren és Brandeis 1890, idézi Nunan és Di Domenico 2013: 1).

A közösségi média diskurzusaiban aktívan résztvevő felhasználók okoskészülékeinek GPS jelei információkat továbbítanak azok mindennapi földrajzi helyzetéről. Ha ezen adatok hosszabb időtávon keresztül érkeznek, akkor a Big Data segítségével (földrajzi helyzet + posztolt képi és szöveges tartalom) képet kaphatunk mindennapi mozgásukról, otthonuk, munkahelyük elhelyezkedéséről, sőt a rendszer meg tudja jósolni, hol lesz az illető másnap (Shroff 2014: 190).

Jim Farley, a Ford Motor egyik vezetője szerint a cég tudja, hogy az általuk készített autók sofőrjei mikor követnek el éppen közúti vétséget, hiszen a személygépkocsik gyári navigációs rendszere és GPS jele ezt elárulja, de nyilatkozatában rögtön hozzátette, hogy ezen információkhoz való hozzáférése csak a vállalatnak van (Sedgwick 2014, idézi Doughty 2014: 7).

Hasonlóképpen a MOOC rendszerekben résztvevő hallgatók adatai is privacy kérdéseket vetnek fel, hiszen a felhasználók egy Big Data elemzés során is azonosíthatók. Ugyanis csak akkor lehet a legpontosabb eredményeket kapni és következtetéseket levonni, ha nem anonimizálják az adatokat (Daries et al. 2014: 56-58).

A Big Data által kapott eredmények ellenőrzéséhez emberi beavatkozásra van szükség. 2014-ben a chicagói rendőrség tagjai személyesen keresték meg azokat a személyeket, melyekről a rendszer úgy gondolta, hogy a jövőben bűncselekményt követhetnek el. A megkeresés során a rendőrök a rendszer által kiemelt személyek figyelmét munka és továbbképzési lehetőségre hívták fel, de egyben ismertették velük a rendszer által előre-

jelzett egyes bűncselekmények esetében a lehetséges büntetési tételeket is. Ennek hatására csökkent a bűnözés, azonban számos kritika érte a rendőrséget, mondván, hogy az ilyen típusú adatkezeléssel megsértette az emberek személyiségi jogait[14].

A privacy kérdésköre különösen a digitálisan behálózott Dél-Koreában hangsúlyos. A koreai információs társadalmat foglalkoztató egyik legaktuálisabb kérdés az, hogy a Big Data adatai személyes adatoknak számítanak, és így védelem alatt állnak-e (Kwan 2014: 125)? Amennyiben igen, úgy például a vásárlók adatai gyakran sérülnek a kereskedelmi cégek gyakorlata miatt, melynek során személyre szabott ajánlatokat kapnak. Ezért a koreai PIPA (Protect IP Act) átdolgozásra szorul (Young 2014: 355). Be kell építeni az adatok védelmét a Big Data környezetében is, a jogot, hogy a rendszer bizonyos információkat elfelejtsen (Cha 2014: 193-195). Ezzel egyidőben a mozgóképen rögzítettek személyes adatait is védeni kell (Woo 2015).

A Big Data kritikája

„Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"
(Thomas Stearns Eliot, *The Rock*)[15]

A tanulmány egyes fejezeteit olvasva az a kép alakulhat ki, miszerint a Big Data korunk varázseszköze, melynek segítségével óriási adatmennyiségekből egy kattintással összefüggések olvashatók ki, gazdasági folyamatok tehetők hatékonyabbá, társadalmi egyenlőtlenségek javíthatók ki. Hogy árnyaljuk a képet, az előző fejezetben már kitértünk a Big Data kihívásaira.

A Big Datát azonban a konstruktív kérdezés mellett számos kritika is éri. Ezek közül legerősebb azon állítás, miszerint a Big Data környezetében a nagy adattömegek elemzése csak a korrelációkat tárja fel, az okok megértésére azonban nem igazán alkalmas. Ennek következtében pedig azt az illúziót keltheti az üzleti, politikai éstudományos élet résztvevőiben és legfőbbképpen döntéshozóiban, hogy erre már nincs is szükség. Nem kell már tehát felismernünk az okokokat, keresni a miérteket. Ezt követi a kritika, miszerint a Big Data előítéles bizonyos adatokkal szemben, így nem képes teljes képet festeni, végezetül pedig gyakran fennáll az együttállásból származó téves felismerések veszélye, illetve nincs rá garancia, hogy az adatokat anonim módon kezelik, és nem használja fel megfigyelésre.

Ami a látókörön kívül esik, nem is létezik?

Kézenfekvő kérdés, hogy ami nem létezik digitálisan, az a Big Data paradigmájának szempontjából nincs. Elvben a nagymennyiségű adatok környezetében a minta mindenkitől származik. Mindenkitől, aki interakciót folytat a hálózaton, így a digitális nyomtalanok a jövőben majd másodrendű állampolgároknak számítanak?

[14] http://www.ap-institute.com/big-data-articles/how-is-big-data-used-in-practice-10-use-cases-everyone-should-read.aspx

[15] *Hol a bölcsesség, ami a tudásban elveszett? Hol a tudás, amit elvesztettünk az információban?* (Thomas Stearns Eliot, *A szikla*)

Hasonlóképpen, a Big Datában nagy a zaj, mely a kommunikáció elméletek szerint megnehezíti a befogadást (Shannon 1948) és ezzel együtt az elemzést is, miközben elfogultságra hajlamos (Boyd és Crawford 2012). A közösségi médiából gyűjtött adatok például ugyanis többnyire az Y és Z generációtól származnak, és nem reprezentálják a digitális írástudók összességét.

Valós összefüggések vagy csupán egybeesések? Az okok figyelmen kívül hagyása

A Big Data a hyperpragmatizmust népszerűsíti. Arra fókuszál, ami működik, anélkül, hogy kíváncsi lenna arra, miért és hogyan is működik. Csak annyit szükséges tudnunk, hogy valami működik, előnyos, profitot termel. Például a hitelkártya kibocsátók ügyfeleik adatait elemezve kiderítették, hogy azok, akik krómozott koponyákat vásárolnak, hajlamasak késni a részletekkel. Azok azonban, akik csúszásgátlót vesznek bútoraikra, időben törlesztenek. Arra azonban, hogy ennek mi az oka, senki sem kíváncsi.[16]

Kritikusai szerint a Big Data, amennyiben félreértik, azt a tévhitet kelti a kutatókban, hogy madártávlatból mindent észre tudnak venni, ami korábban rejtett volt a szemük elől. Olyan mintázatokat és összefüggéseket is meglátnak, melyek a valóságban nem léteznek, csupán véletlen egybeesések. Ennek oka, hogy a nagy mennyiségű adat számtalan kapcsolatot létrehoz és mintázatot kirajzol (Boyd és Crawford 2012). Erre az állításra a közösségi médiából hozhatunk példát. Akármilyen fejlett a gépi tanulás, a jelenben az elemző programok még nem képesek a szöveg valamennyi aspektusát figyelembe venni a jelentésének elemzésekor. A földrajzi, időbeli, társadalmi és kulturális beágyazottság jelenleg még túl bonyolultnak számít a jelentés tökéletes megértéséhez (Schintler és Kulkarni 2014: 345).

Hasonlóképpen, bírálói szerint a Big Data a „mi" történik kérdésre válaszol, a „miért"-tel azonban adós marad, az okokra nem képes rávilágítani. Honavar (2014) kiemeli, hogy szakadék tátong azon képességünk között, hogy megszerezzük az adatokat, és azon képességünk között, hogy összetett és helytálló következtetéseket vonjunk le.

Torzulások

A Big Data kritikusai a Google Flu Trends-t hozzák fel kedvenc példájukként, amikor torz következtetésekre akarnak rámutatni. A projektet kezdetben világszerte sikerként könyvelték el[17], hallgatóinkkal is több esetben a Big Data mintapéldájaként vizsgáltuk. Kiderült azonban, hogy eredményei torzak voltak azért, mert bizonyos körülményeket az elemzés nem vett figyelembe és tegyük hozzá, nem is tudott figyelembe venni.[18]

Mint már ismertettük, a Google Flu Trends Big Data algoritmusai eredetileg az influenzára jellemző tünetekkel kapcsolatos internetes keresések kapcsán jelezték elő a járványok kezdetét és időbeli lefolyását. Már azonban 2011-ben a Google jelentősen túlbecsülte az

[16] http://www.mtabsurveyanalysis.com/big-data-caveats-three-big-criticisms-of-big-data/

[17] http://www.ap-institute.com/big-data-articles/how-is-big-data-used-in-practice-10-use-cases-everyone-should-read.aspx

[18] Jelenleg a projekt kimutatásai vizualizált formában már nem érhetők el, a Google azonban nem tűntette el azt, csupán lezárta és leegyszerűsítette az oldalt, nem utalva a projekt megszűnésének okaira (https://www.google.org/flutrends/about/).

influenzás megbetegedések számát, csupán azért, mert nem vette figyelembe, hogy a közösségi média hatására az influenzával kapcsolatos aggodalmak felerősödnek, és a betegségre fókuszálnak a felhasználók, akkor is rákeresnek a tünetekre, ha csak kíváncsiak (Schintler és Kulkarni 2014: 344).

Szabványosítás

A Big Data nagy adatmennyiség kezelésére jött létre, ezekből igyekszik kiolvasni mintázatokat. Éppen ezért nem veszi figyelembe az egyénit, az egyedit, a mintától eltérőt. Bőgel György a Big Data kritikájában rámutat arra, hogy oktatásunk, mely a jelenben sem differenciál a jövőben még kevésbé lesz képes figyelembe venni a tanulók közti különbségeket (Bőgel 2015: 130).

Adat és valóság

Németh Renáta megkérdőjelezi a Big Data objektivitását, és (a gépi tanulás hatékonyságával szemben is) kiemeli, hogy a legjobb algoritmus sem képes értelmes kérdésfelvetésre önmagától, nem képes egyedül interpretálni a mérési adatokat (Németh 2015: 204). Így könnyen lehet, hogy délibábot látunk, olyan valóságot, mely nem létezik, és csak az adatok keltik látszatát.

A Big Data mint kutatási módszer a jövőben azt is jelentheti, hogy a kutató és a kutatás alanya véglegesen elválik, nem találkoznak többé. Eltűnik a mélyinterjú, és csupán az adatokból próbálunk majd következtetni, mit gondolnak és mit cselekszenek az emberek. Elég lesz távolról megfigyelni őket (Székely 2015: 2011).

Surveillance, totális megfigyelés

A Big Data kritikája a surveillance és totális megfigyelés kapcsán a legerősebb. A privacy kapcsán már értekeztünk az adatokhoz való hozzáférés kérdéséről. Egyik legszkeptikusabb olvasatát adja a Big Data jelenségnek John Feffer (2014), aki Foucault (1977) Panopticonját idézi, melyben a börtönigazgató az intézmény középpontjában ülve mindent megfigyelhetett. Ezt a megfigyelési paradigmát követte a globális megfigyelés hidegháborús időszaka, melyben a megfigyelés a társadalom minden rétegébe begyűrűzött. Végül pedig információs társadalmunk és mediatizált világunk metaforájává vált a Nagy Testvér, a megfigyelés és magamutogatás kombinációja. A mozzanat, melynek során a kamerát magunk felé fordítjuk (szelfi), a közösségi médiában minden lépésünket megosztjuk, ezzel kiszolgáltatva magunkat a Big Data totális surveillance gépezetének.

Z. Karvalics a kritika kritikájával él, miközben kijelenti, a survaillance-szel kapcsolatos félelmek túlzottak, hogy az online tranzakciós lábnyom nem egyenlő magával az emberrel, annak a gazdagságával és teljességével. A Big Data segítségével így „nem totális megfigyelésre és profilalkotásra törekszenek – hiszen ez képtelenség lenne, hanem profitnövelésre, fontos azonban, hogy a döntés a vásárlással, fogyasztással, szolgáltatás igénybevételével kapcsolatban alapvetően még mindig a felhasználóé" (Z. Karvalics 2015: 195).

Összefoglalás

A jelen információs társadalmában, a mindenhol jelenlévő számítástechnika, az adatbázisok és szenzorok, a Web 2.0 szabadon írható közösségi média, illetve a Web 3.0 szemantikus környezetében a Big Data számos, eddig rejtett összefüggést képes feltárni. A nagy mennyiségű adat valós idejű feldolgozása jelentős sikereket hozhat a kereskedelemben, oktatásban, segítheti az okosvárosok és az orvostudomány fejlődését. Hasonlóképpen fontos szerepet játszhat a politika trendek és aktuális társadalmi problémák felismerésében, és így végső soron támogathatja a békés megoldások megtalálását.

Látnunk kell azonban, hogy még a Big Data forradalom kezdetén vagyunk. Folyamatosan jelennek meg az új paradigma tulajdonságait figyelembe vevő szoftverek, algoritmusok, fejlődik az eredmények vizualizációja. Most, hogy a számítási teljesítmény és tárhely alacsony árai elvben lehetővé teszik a mindennapi felhasználó számára is a Big Data univerzumába való belépést, vajon valóban elmondhatjuk, hogy a nagy adatok világa demokratikus, és mindenki hozzáfér?

Számos kérdéssel kell tehát még szembenézünk: a Big Data új kihívások elé állít minket a személyi adatok védelme, a privacy terén. Hasonlóképpen fel kell ismernünk a korlátait, a lehetséges torzulásokat és elfogultságokat, a mintán kívül maradt csoportokat és figyelmen kívül hagyott tényezőket, az eseteket, amikor az eredmények véletlen egybeesésről és nem valós összefüggésekről árulkodnak.

Irodalom

Akerkar, Rajendra, Guillermo Vega-Gorgojo, Grunde Løvoll, Stephane Grumbach, Aurelien Faravelon, Rachel Finn, Kush Wadhwa, Anna Donovan, Lorenzo Bigagli, *Understanding and mapping big data, Big data roadmap and cross-disciplinary community for addressing societal externalities*, 2015.

Altbach, Philip G.,Liz Reisberg, Laura E. Rumbley, *Trendsin Global Higher Education: Tracking an Academic Revolution, Report Prepared for the UNESCO 2009 World Conferenceon Higher Education*, United Nations Educational, Scientific and Cultural Organization, Paris, 2009.

Bajomi Lázár Péter, „Manipulál-e a média?", *Médiakutató*, (2006/nyár), 81–85. old.

Barbash, Fred, Twitter warnsusers of 'states ponsored' hack apparently in pursuit of private information, *The Washington Post*, December 14, 2015. https://www.washingtonpost.com/news/morning-mix/wp/2015/12/14/twitter-warns-users-of-state-sponsored-hack-apparently-in-pursuit-of-private-information/

Benedek András, Molnár György, „ICT Related Tasks and Challenges In The New Model of Technical Teacher Training", in John Terzakis, Constantin Paleologu, Gyires Tibor (eds.), *Eighth International Multi-Conferenceon Computinginthe Global Information Technology*, Nice, InfoWare, 2013, pp. 40-44.

Benedek András, Molnár, György, „Supporting the m-learning based knowledge transfer in university education and corporate sector", in Arnedillo Sánchez, Pedro Isaías (eds.), *Proceedings of the 10th International Conferenceon Mobile Learning*, Madrid, IADIS Press, 2014, pp. 339-343.

Bessis, Nik, Dobre Ciprian (eds.), *Big Data and Internet of Things: A Roadmap for Smart Environment*, Springer, New York, 2014.

Beyan, Timur, Yes¸im Aydın Son, „Emerging Technologies in Health Information Systems: Genomics Driven Wellness Tracking and Management System (GO-WELL)", in Nik Bessis, Dobre Ciprian (eds.), *Big Data and Internet of Things: A Roadmap for Smart Environment*, Springer, New York, 2014, pp. 315-342.

Bodnár Csaba, „Mi is az a Big Data?", *IT Café*, 2014. http://itcafe.hu/hir/mi_is_az_a_big_data.html
Boyd, Danah, Kate Crawford, „Critical questions for big data", *Information, Communication and Society*, 15 (-), 2012, pp. 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878
Bőgel György, *A Big Data ökoszisztémája*, Typotex, Budapest, 2015.
Brownstein, John S., Clark C. Freifeld, Lawrence C. Madoff, „Digital Disease Detection – Harnessing the Web for Public Health Surveillance", *New England Journal of Medicine*, no 360 (2009), pp. 2153-2157. http://dx.doi.org/10.1056/NEJMp0900702
Bush, Vannevar, „As We May Think", *The Atlantic Monthly*, 176(1945), pp. 101-108. http://www.theatlantic.com/doc/194507/bush [Magyarul: Vannevar Bush „Út az új gondolkodás felé", in Sugár János (szerk.), *Hipertext + multimédia*, Artpool, Budapest, 1996.] http://www.artpool.hu/hipermedia/bush.html
Cha, Sang-Yook (차상육) „A Study on Big Data Circumstance and Privacy Protection" (빅데이터(Big Data) 환경과 프라이버시의 보호), IT와 법 연구, 8(-), 2014, pp. 193-259.
Condliffe, Jamie, „Deep Mind's First Medical Research Gig Will Use AI to Diagnose Eye Disease", *MIT Technology Review*, 2016, https://www.technologyreview.com/s/601845/deepminds-first-medical-research-gig-will-use-ai-to-diagnose-eye-disease/
Condliffe, Jamie, „AI Is Learning to Seethe World—But Not the Way Humans Do", *MIT TechnologyReview*, 2016, https://www.technologyreview.com/s/601819/ai-is-learning-to-see-the-world-but-not-the-way-humans-do/
Curran, Kevin, Niamh Curran, „Social Networking Analysis" in Nik Bessis, Dobre Ciprian (eds.), *Big Data and Internet of Things: A Roadmap for Smart Environment*, New York, Springer, 2014, pp. 367- 378.
Csepeli György, „A szociológia és a Big Data", *Replika*, 92-93. szám (2015/3-4), 171-176 old.
Daries, Jon P., Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Andrew Dean, Ho, Daniel Thomas Seaton, Isaac, Chuang, „Privacy, Anonymity, and Big Data in the Social Sciences", *Education*, 12(7), 2014, pp. 56-63.
Dessewffy Tibor, Láng László, „Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon", *Replika*, 92-93 szám, (2015/3-4), 157-170.
Diebold, Francis, Big Data Dynamic Factor Modelsfor Macroeconomic Measurement and Forecasting, *Eighth World Congress of the Econometric Society*, (Seattle, augusztus), 2000. http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF
Doughty, Howard A., „Surveillance, Big Data Analytics and theDeath of Privacy", *College Quarterly*, 17(3), 2014, pp. 1-21.
Feffer, John, Participatory totalitarianism, 2014. http://www/commondreams.org/view/2014/06/04-10
Foucault, Michael, „Discipline and punish: Panopticism", in Alan Sheridan, (ed.), *Discipline and punish: The birth of the prison*, New York, VintageBooks, 1977.
Garson, Barbara, *The Electronic Sweatshop: How Computersare Transforming the Office of the Future in to the Factory of the Past*, Simon & Schuster, New York, 1988.
Herman, Edward S., Noam Chomsky, *Manufacturing Consent: The Political Economy of the Mass Media*, Pantheon Books, New York, 1988.
Honavar, Vasant G., „The Promise and Potential of Big Data: A Casefor Discovery Informatics", *Review of Policy Research*, 31(4), 2014, pp. 326–330. http://dx.doi.org/10.1111/ropr.12080
Hopkins, Brian, Boris Evelson, *Expand Your Digital Horizon With Big data*, Forrester Research, Cambridge, 2011. http://www.asterdata.com/newsletter-images/30-04-2012/resources/forrester_expand_your_digital_horiz.pdf
Joshi, Pramila, „Analyzing Big Data Tools and Deployment Platforms", *International Journal of Multidisciplinary Approach and Studies*, 2(2), 2015, pp. 45-56.
Kim, Eun-ju (김은주), „Big Data를 활용한 여성소비자의 특성연구", Journal of Digital Convergence (디지털융복합연구), 13(10), 2015, pp. 185-194. http://dx.doi.org/10.14400/JDC.2015.13.10.185

Kim, Gang-Hoo, Silvana Trimi, Ji-Hyong Chung, „Big-Data Applications int he Government Sector", *Communications of the ACM*, 57(3), 2014, pp. 78-85.http://dx.doi.org/10.1145/2500873

Krishnamurthya, Rashmi, Kevin C. Desouzab, „Big data analytics: The case of the social security administration", *Information Polity*, 19(3-4), 2014, pp. 165–178. http://dx.doi.org/10.3233/IP-140337

Kwak, Kwan Hoon (곽관훈), „Special Issue 2: Data Protection and User`s Rights : Big Data and Personal Information Protection", (특집 2 : 개인정보 보호와 이용자 권리 ; 기업의 빅데이터(Big Data) 활용과 개인정보의 보호의 조화), 鑑法學 (Ilkam Llaw Rreview (鑑法學)), 27(-), 2014, pp. 125-153.

Laney, Douglas, *3D Data Management: Controlling Data Volume, Velocity and Variety*, Gartner, 2012. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, David, Ryan Kennedy, „We Can Learn From the Epic Failure of Google Flu Trends", Wired, 2015. http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lee, Seokjoo (이석주), Yeon, Jiyoun (연지윤), Cheon, Seunghoon (천승훈) „Big Data for Transportation Policies and Their Applications" (빅데이터를 이용한 교통정책 개발 및 활용성 증대방안),. 한국교통연구원 기본연구보고서, -(-), 2013, pp. 1-124.

Löffler, Sven, Mi az a Big Data, 2014. https://www.it-services.hu/hirek/mi-az-a-big-data/

Majkić, Zoran, *Big Data Integration Theory Theory and Methods of Database Mappings, Programming Languages, and Semantics*, Springer, New York, 2014.

Mashey, John R., *Big Data ... and the Next Wave of Infra Stress*, 1999. http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Mayer-Schönberger, Victor, Kenneth Cukier, *Big data: A revolution that will transform how we live, work, and think*, HoughtonMifflin Harcourt, New York, 2013.

McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, 2011. http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Business%20Technology/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

McNeely, Connie L., Jong-onHahm, „The Big (Data) Bang: Policy, Prospects, and Challenges", *Review of Policy Research*, 31(4), 2014, pp. 304-310.http://dx.doi.org/10.1111/ropr.12082

Moore, Gordon E., „Cramming more components onto integrated circuits", *Electronics*, 38(8), 1965. 114–117. http://dx.doi.org/10.1109/N-SSC.2006.4785860

Moorthy, Janakiraman, Rangin Lahiri, Neelanjan Biswas, Dipyaman Sanyal, Jayanthi Ranjan, Krishnadas Nanath, Pulak Ghosh, „Big Data: Prospects and Challenges", *VIKALPA The Journal for Decision Makers*, 40(1), 2015, pp. 74–96. http://dx.doi.org/10.1177/0256090915575450

n. a. *A Facebook jogi nyilatkozata*. https://www.facebook.com/legal/terms

n. a. *Big Data*. http://www.gartner.com/it-glossary/big-data/

n. a. *How is Big Data Usedin Practice? 10 Use Cases Everyone Must Read*. http://www.ap-institute.com/big-data-articles/how-is-big-data-used-in-practice-10-use-cases-everyone-should-read.aspx

n. a. https://www.youtube.com/yt/press/statistics.html

n. a. *Walmart Is Making Big Data Part Of Its DNA*. https://datafloq.com/read/walmart-making-big-data-part-dna/509

National Research Council of Italy, http://byte-project.eu/wp-content/uploads/2016/03/BYTE-D1.1-FINAL-post-Y1-review.compressed-1.pdf

Németh Renáta, „A számok tényleg magukért beszélnek?", *Replika*, 92-93 szám, (2015/3-4), 203-208.

Nunan, Daniel, MariaLaura Di Domenico, „Market research and the ethics of big data", *International Journal of Market Research*, 55(4), 2013, pp. 1-13.

Ratner, Bruce, *Statistical modeling and analysis for database marketing: effective techniques for mining big data*, Cleveland, Chapman and Hall/CRC Press, 2004.

Samuel, Arthur, „Some Studies in Machine Learning Usingt he Game of Checkers", *IBM Journal* 3 (3), 1959, pp. 210–229.

Schintler, Laurie A.,Rajendra Kulkarni, „Big Data for Policy Analysis: The Good, the Bad, and the Ugly", *Review of Policy Research*, 31(4), 2014, pp. 343-348.http://dx.doi.org/10.1111/ropr.12079

Shin, Kwan Woo (신관우), „A Study on Utilizing Video Big data of Private Security -Focusing on Protection of the Personal Video Information" (민간경비의 영상 빅데이터 활용을 위한 과제: 개인영상정보 보호를 중심으로), South Korea Private Security Science Review (한국민간경비학회보), 14(1), 2015, pp. 212-231.

Shroff, Gautam, *The Intelligent Web: Search, smart algorithms, and big data*, Oxford, Oxford University Press, 2014.

Siemens, George, „Supporting and promoting learning analytics research", *Journal of Learning Analytics*1(1), 2014, pp. 3–5.

Son, Young-Hoa (손영화), „The Protection of Personal Information in the Era of Big Data", (빅데이터 시대의 개인정보 보호방안),. Study on Enterprise Law (企業法研究), 28(3), 2014, pp. 355-393.

Sondergaard, Peter (2011) Conference speach, *Gartner Symposium/ITxpo* 2011, October 16-20, Orlando. http://www.gartner.com/newsroom/id/1824919

Székely Iván, Az adatmentes zónák szükségessége és esélye, *Replika*, 92-93 szám, (2015/3-4), 209-226. old.

Szűts Zoltán, *A világháló metaforái*, Osiris, Budapest, 2013.

Walter, Chip, „Kryder's Law", *Scientific American*, 2005. 08. 01. http://www.scientificamerican.com/article/kryders-law/

Weiser, Mark, „The computer forthe 21st century", *Scientific American* 265(3), 1991, pp. 94-104.

Weiss, Sholom, Nitin Indurkhya, *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers Inc., Burlington, 1998.

White, Patricia, R. Saylor Breckenridge, „Trade-Offs, Limitations, and Promises of Big Data in Social Science Research", *Review of Policy Research*, 31(4), 2014, pp. 331-338. http://dx.doi.org/10.1111/ropr.12078

Yoo, Jinil, „A civil kérdések esélyei és kihívásai az okos (digitálisan behálózott) városokban a dél-koreai New Songdo City példáján keresztül", *Civil Szemle*, 11(2), 2014, 25-48 old.

Yoon, Seok Jin (윤석진), „Conflicts of Personal Information Protection and advantage of Big Data, the problem and the legislative policy issues – Focus on the Health & Medical Big Data" (개인정보 보호와 빅데이터 활용의 충돌, 그 문제와 입법정책 과제 -보건의료 빅데이터를 중심으로),. *Central University (中央法學)*, 17(1), 2015, pp. 7-47.

Z. Karvalics László, „A Nagy Adat-jelenség társadalomtudományi lehorgonyozásához", *Replika*, 92-93, (2015/3-4), 189-201 old.

Zadrozny, Peter, Raghu Kodali, *Big Data Analytics Using Splunk: Deriving Operation al Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*, Apress, New York, 2013.

Zikopoulos, Paul, Chris Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill Professional, New York, 2011.

Szűts Zoltán médiakutató, az ELTE-n diplomázott, doktorált és habilitált. A Zsigmond Király Egyetem tanszékvezető főiskolai tanára és a BME Műszaki Pedagógia Tanszékének oktatója. Rendszeresen publikál az újmédia, információs társadalom, digitális pedagógia és online művészetek témájában tanulmányokat és ismeretterjesztő cikkeket a hazai tudományos lapokban. A világháló metaforái – Bevezetés az új média művészetébe és az Egyetem 2.0 kötetek szerzője. 2004 és 2007 között a szöuli Hankuk University of Foreign Studies vendégtanára volt. Kutatási területe az online kommunikáció, hipertext, az online közösségek és a világháló művészete. Legutóbbi publikációja az Információs Társadalomban: A netsemlegesség – definíciók, törvényhozói, tartalomszolgáltatói, internetszolgáltatói és felhasználói olvasatok, 2015/3.

Yoo Jinil irodalomtörténész, doktori értekezését az ELTE-n írta. A szöuli Hankuk University of Foreign Studies magyar tanszékének tanára. Korábban a Korean Association of Central&Eastern European and Balkan Studies munkatársa volt. Rendszeresen publikál tanulmányokat és tudománynépszerűsítő cikkeket koreai és magyar tudományos folyóiratokban. Kutatási területe a magyar irodalom, a közép-európai és koreai kulturális kapcsolatok, digitális kultúra. Legutóbbi publikációja az Információs Társadalomban: A netsemlegesség – definíciók, törvényhozói, tartalomszolgáltatói, internetszolgáltatói és felhasználói olvasatok, 2015/3.