DANAH BOYD, KATE CRAWFORD

Az adatrengeteg kínos kérdései – Vitaindító egy kulturális, műszaki és tudományos jelenségről

Elközelgett a Big Data, az adatrengeteg kora. A számítástudósok, fizikusok, közgazdászok, matematikusok, politológusok, bioinformatikusok, szociológusok és más tudósok fennszóval követelik, hogy az embereket, dolgokat és ezek kapcsolatait leíró, folyvást termelődő, hatalmas adatmennyiséggel dolgozhassanak. Fontos kérdések formálódnak meg. Hozzásegít-e az adatrengeteg kereshetősége az eszközök, szolgáltatások és közjavak javításához, vagy inkább a magántitoksértés és a tolakodó marketing új hullámát vezeti be? Megkönnyíti-e az adatok elemzése az online közösségek és politikai mozgalmak megértését, vagy a tiltakozók lenyomozását és a szólásjog elnyomását szolgálja majd? Átalakítja-e az emberi kommunikáció és kultúra kutatásának mai módszereit, vagy inkább beszűkíti a vizsgálható témák skáláját, magát a „kutatás" fogalmát definiálva újra? Véleményünk szerint az adatrengeteg szociotechnológiai jelenségének előretörése kapcsán kritikus vizsgálatnak kell alávetni e jelenség előfeltevéseit és előítéleteit. Cikkünk hat provokatív tézist tartalmaz, amelyekkel diszkussziót kívánunk indítani az adatrengeteg különböző aspektusairól: e technológia, elemzés és mitológia összjátékából kisarjadó kulturális, műszaki és tudományos jelenségről, amely terjedelmes retorikát gerjeszt mind utópikus, mind disztópikus hangvételben.

Kulcsszavak: *Big Data; adatelemzés; közösségi média; kommunikációtudomány; közösségi hálózati webhelyek; tudományfilozófia; episztemológia; etika; Twitter*

Szerzői információ:

danah boyd a Microsoft Research kutatója, a New York University adjunktusa, több más egyetem vendégkutatója, óraadója. Fő kutatási területe a közösségi média, a fiatalok médiahasználata, a köz- és a privátszféra közötti feszültségek és a technológia társadalmi hatásai. Doktori fokozatát 2008-ban szerezte a University of California-Berkeley-n. Kutatásain túl számos civil projektben is részt vesz önkéntesként.

Kate Crawford a University of New South Wales Journalism and Media Research Centre docense. Fő kutatási területe a társadalmi változások és a médiatechnológiák kapcsolata, különös tekintettel a mobil eszközökre és a közösségi hálózatokra. Több nagyszabású kutatás vezetője, 2008-ban az Ausztrál Akadémia díjazottja.

Így hivatkozzon erre a cikkre:

boyd, danah, Kate Crawford. „Az adatrengeteg kínos kérdései – Vitaindító egy kulturális, műszaki és tudományos jelenségről". *Információs Társadalom* XII, 2. szám (2012): 7–23.

https://dx.doi.org/10.22503/inftars.XII.2012.2.1

danah boyd – Kate Crawford

Az adatrengeteg kínos kérdései[1]

Vitaindító egy kulturális, műszaki és tudományos jelenségről

> *A technológia se nem jó, se nem rossz; nem is semleges… a technológiának a társadalom ökológiájával való interakciója olyan, hogy a műszaki újítások környezeti, szociális és emberi következményei gyakorta messze túlhaladják a szóban forgó műszaki eszközök vagy eljárások eredeti célját.* (Kranzberg 1986, 545)

> *Diskurzust kell indítanunk valamiről, amiről még nem szól számottevő diskurzus: a sokféle téri, idői és anyagi jellegről, amelyeket adatbázisainkban leképezhetünk, méghozzá úgy, hogy a tervezés a legnagyobb rugalmasságot engedje meg, és amennyire csak lehet, vegye tekintetbe a polifónia és polikrónia emergens jelenségeit. A „nyers adat" eleve oximoron, és ráadásul rossz ötlet. Az adatokat igenis elő kell főzni, méghozzá gondosan* (Bowker 2005, 183–184).

Elközelgett a Big Data, az adatrengeteg kora. A számítástudósok, fizikusok, közgazdászok, matematikusok, politológusok, bioinformatikusok, szociológusok és más tudósok fennszóval követelik, hogy az embereket, dolgokat és ezek kapcsolatait leíró, folyvást termelődő, hatalmas adatmennyiséggel dolgozhassanak. Különféle csoportok vitáznak arról, milyen előnye és mekkora költsége lehet a génszekvenciák, közösségi médiabeli interakciók, egészségügyi feljegyzések, telefonnaplók, közigazgatási iratok és más, emberek hagyta digitális nyomok elemzésének. Fontos kérdések formálódnak meg. Hozzásegít-e a masszív adatok kereshetősége az eszközök, szolgáltatások és közjavak javításához, vagy inkább a magántitoksértés és a tolakodó marketing új hullámát vezeti be? Megkönnyíti-e az adatok elemzése az online közösségek és politikai mozgalmak megértését, vagy a tiltakozók lenyomozását és a szólásjog elnyomását szolgálja majd? Átalakítja-e az adatok nagy mennyisége az emberi kommunikáció és kultúra kutatásának mai módszereit, vagy inkább beszűkíti a vizsgálható témák skáláját, magát a „kutatás" fogalmát definiálva újra?

A „Big Data" („nagy adatok"[2]) terminus sok szempontból nem túl szerencsés. Manovich (2011) rámutatott, hogy a természettudományokban korábban olyan méretű adathalmazokra használták, amelyek már szuperszámítógépet követelnek – azonban ami egykor ilyen masinát igényelt, az ma már asztali számítógépen is elemezhető különleges szoftver nélkül. Aligha kétséges, hogy ma gyakran igen nagy tömegben állnak rendelkezésre az adatok, csakhogy ez még nem az újfajta adat-ökoszisztéma meghatározásának kritériuma.

[1] Eredetiben: danah boyd and Kate Crawford: „Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon". In: *Information, Communication & Society* 15:5, p. 662–679. Fordításunk a kiadó, a Taylor & Francis Ltd. (http://www.tandfonline.com) engedélyével jelenik meg.

[2] A cikk szerzői angolul a „Big Data" kifejezés nagybetűs írásmódjával jelzik, amikor az itt megfogalmazott jelenségre, nem pedig egyszerűen „nagy adatokra" gondolnak. A magyarul másra nem használatos *adatrengeteg* szó megalkotásával e jelenség összetett voltát kívántam érzékeltetni. *(A ford.)*

Az adatrengeteg körébe tartozó egyes adathalmazok (pl. a valamely témáról szóló Twitter-üzenetek összessége) olykor meg sem közelítik más, korábbi és eredetileg nem adatrengetegnek tekintett adatok (pl. népszámlálási adatok) tömegét. Az adatrengeteg tehát nem elsősorban az adatok nagyságáról szól, hanem inkább a nagy adathalmazok keresésének, felhalmozásának és kereszthivatkozásokkal való ellátásának lehetőségéről.

A mi definíciónk szerint az adatrengeteg olyan kulturális, műszaki és tudományos jelenség, mely a következő tényezők összjátékára támaszkodik:

(1) *Technológia:* a számítókapacitás és algoritmikus pontosság maximalizálása a nagy adathalmazok gyűjtése, elemzése, összekapcsolása és összevetése végett.
(2) *Elemzés:* nagy adathalmazok vizsgálata olyan mintázatok azonosítása végett, amelyek alapján gazdasági, szociális, műszaki és jogi állítások tehetőek.
(3) *Mitológia:* az a széles körben elterjedt hit, hogy a nagy adathalmazok magasabb rendű intelligenciát jelentenek, és belőlük korábban elképzelhetetlen, az igazság, objektivitás és pontosság erényeivel vértezett tudás nyerhető.

Más szociotechnológiai jelenségekhez hasonlóan az adatrengeteg is egyaránt gerjeszt utópikus és disztópikus hangvételű retorikát. Egyfelől hathatós eszközt látunk benne, amely gyógyírt adhat a társadalom különféle problémáira és olyan változatos területeken kecsegtet új eredményekkel, mint a rákkutatás, a terrorizmus és a klímaváltozás. Másfelől a „nagy adatok" a „nagy testvér" nyomasztó megnyilvánulásának tűnnek, amely szabad utat kínál a magánszféra megsértéséhez, csökkenti a polgári szabadságot, és erősíti az államok és nagyvállalatok hatalmát. Mint minden más szociotechnológiai jelenségnél, a remény és rettegés áramlatai itt is gyakorta felismerhetetlenné teszik a folyamatban lévő finomabb, árnyaltabb változásokat.

A számítógépes adatbázis nem újdonság. A világ első automatizált számítóberendezését – a lyukkártyagépet – már 1890-ben bevetette az USA Népszámlálási Hivatala (Anderson 1988). A relációs adatbázisok az 1960-as években jelentek meg (Fry–Sibley 1974). A személyi számítógépek és az internet azóta egyre több embernek – tudósoknak, marketing-szakembereknek, kormányhivataloknak, oktatási intézményeknek és érdeklődő magánszemélyeknek egyaránt – teszi lehetővé, hogy adatokat generáljon, osszon meg, kezeljen és rendszerezzen. Ez a Savage és Burrows (2007) által az empirikus szociológia válságának nevezett jelenséghez vezetett. Olyan adathalmazokat, amelyek valaha áttekinthetetlennek és nehezen kezelhetőnek számítottak, és ekképp csak a társadalomtudósok érdeklődésére tartottak számot, ma már ömlesztett formában mindenki elérhet, akit csak érdekel – előképzettségtől függetlenül.

Kritikus kérdés, hogy mihez kezdjünk az adatrengeteg korának érkezésével. Bár a jelenség környezetét a bizonytalanság és a gyors változás jellemzi, mégis a ma hozott döntések fogják kialakítani a jövőt. Mivel egyre inkább automatizálható az adatok gyűjtése és elemzése – sőt az emberi viselkedés nagy léptékű mintázatainak kimutatására és szemléltetésére alkalmas algoritmusok is –, fel kell tennünk a kérdést, hogy mely rendszerek hajtják és melyek szabályozzák e gyakorlatokat. Lessig (1999) érvelése szerint a szociális rendszerekben négy szabályozó erő működik: a piac, a törvény, a társadalmi normák és az architektúra, vagyis – a számítástechnika esetében – a kód. Az adatrengeteg vonatkozásában e négy erő gyakran egymás elle-

nében hat. A piac csupa lehetőséget lát a nagy adattömegben: a marketingesek célzott reklámra használják, a biztosítók ajánlataik finomítására, a Wall Street bankárai pedig a piac kiismerésére. Az adatok gyűjtésének és tárolásának korlátozására már törvényjavaslatok is készültek (pl. a 2011-es amerikai Online nyomon követés elleni törvény). A személyre szabhatóság és hasonló szolgáltatások lehetővé teszik a kívánt információ gyors elérését, azonban nehéz etikai kérdéseket vetnek fel, és aggasztóan megosztják a közönséget (Pariser 2011).

Jelenleg is folynak jelentős és tartalmas kutatások az adatrengeteg kapcsán, de továbbra is fontos, hogy belegondoljunk a kínos kérdésekbe: mit jelent ez az adattömeg; ki milyen adatokhoz férhet hozzá; hogyan és milyen célokkal végezhető el az elemzés. Cikkünk hat provokatív tézist tartalmaz, amelyekkel diszkussziót kívánunk indítani az adatrengeteg különböző aspektusairól. Társadalom- és médiatudósok vagyunk, akik rendszeres eszmecserét folytatnak a számítástudomány és informatika szakértőivel. Kérdéseink nehezek: könnyű válasz nincs rájuk. Bemutatunk több különböző csapdát is, amelyek a társadalomtudósok előtt talán nyilvánvalóak, de más diszciplínák képviselői számára meglepőek lehetnek. Mivel érdekel minket a szociális média, és jártasak vagyunk e területen, itt is elsősorban a szociális média kontextusában foglalkozunk az adatrengeteggel. Ugyanakkor hisszük, hogy kérdéseink más területeken dolgozók számára is fontosak. Tisztában vagyunk azzal, hogy e kérdések csak a kezdetet jelentik, és reméljük, hogy cikkünk hatására mások is megkérdőjelezik majd az adatrengeteg fogalmához rögzült előfeltételezéseket. Az adatrengeteg komputációs kultúrája éppen azért érinti minden terület – többek között a számítástudomány, az üzlet és az orvostudomány – kutatóit, mert potenciális hatása széles körben, több diszciplínára terjed ki. Úgy véljük, ideje kritikus vizsgálatnak alávetni e jelenséget előfeltevéseivel és előítéleteivel együtt.

1. Az adatrengeteg átalakítja a „tudás" definícióját

A huszadik század első évtizedeiben Henry Ford létrehozta a tömegtermelés rendszerét, amely specializált gépeket alkalmazott szabványosított termékek gyártásához. Rövidesen ez lett a műszaki haladás elsőrendű víziója. A „fordizmus" automatizálást és gyártósorokat jelentett. Évtizedekre ez határozta meg a gyártás ortodox elveit: félre a képzett mesteremberekkel és a lassú munkával, elő a gépek új korszakával (Baca 2004). Nem egyszerűen az eszközök újultak meg: a huszadik századot a „sejtszintű" fordizmus jellemezte, amely újraértelmezte a „dolgozás" fogalmát, az ember és a munka viszonyát, sőt általában a társadalmat is.

Az adatrengeteg sem egyszerűen óriási adattömböket és az ezek kezelésére és elemzésére használatos eszközöket és eljárásokat jelenti, hanem a gondolkodás és kutatás komputációs fordulatát is (Burkholder 1992). Csakúgy, ahogyan Ford átalakította az autógyártást, s ezzel megváltoztatta a munka fogalmát, az adatrengeteg is a tudás olyan rendszerének létrejöttéhez vezetett, amely máris elkezdte megváltoztatni a tudás tárgyait, miközben arra is képes, hogy az emberi hálózatokról és közösségekről alkotott képünket befolyásolja. „Változtasd meg az eszközöket – figyelmeztet Latour (2009, p. 9) –, és megváltozik a hozzájuk tartozó társadalomelmélet egésze."

Az adatrengeteg gyökeresen átalakítja a kutatásról való gondolkodásunkat. A komputációs társadalomtudományról szólva Lazer és társai (2009) azt állítják, hogy az adatrengeteg adja meg „a képességet korábban példátlan szélességű, mélységű és terjedelmű adatok gyűjtésére és elemzésére" (722). Csakhogy nem egyszerűen nagyságrendről van itt szó, és az sem elegendő, ha a közelségből indulunk ki, vagyis abból, amit Moretti (2007) a szövegek távolról vagy közelről való elemzésének nevez. Alapvető változás ez az episztemológia és etika szintjén. Az adatrengeteg új értelmezési keretbe helyezi a tudás szerveződésének alapkérdéseit: hogyan folyhat a kutatás, hogyan kerülhetünk kapcsolatba az információval, milyen a valóság természete és hogyan szerveződik kategóriákba. Éppúgy, ahogyan Du Gay és Pryke (2002) megjegyzik, hogy „a könyvelés eszközei… nem egyszerűen a gazdasági tevékenység mérését segítik, hanem alakítják is az általuk mért valóságot" (12–13), az adatrengeteg is újrarajzolja az objektumok, a megismerési módszerek és a szociális életre vonatkozó definíciók térképét.

Anderson, a *Wired* magazin főszerkesztője így ír az általa „a petabájtok kora" névvel illetett jelenséget méltatva:

Olyan világ ez, amelyben minden más bevethető eszköz helyére a masszív adattömeg és az alkalmazott matematika lép. Sutba az emberi viselkedés minden elméletével, a nyelvészettől a szociológiáig! Félre a rendszertannal, az ontológiával és a pszichológiával! Ki tudja, miért teszik az emberek azt, amit tesznek? A lényeg, hogy teszik – ezt pedig eddig sosem látott pontossággal tudjuk nyomon követni és mérni. Ha elegendő az adat, a számok magukért beszélnek. (2008)

A számok magukért beszélnek? Véleményünk szerint a válasz: nem. Sokatmondó részlet az is, ahogyan Anderson nagyvonalúan félresöpör minden más elméletet és diszciplínát: azt az adatrengetegről folytatott vitákban gyakran megjelenő arrogáns háttérelvet jelzi ez, amely szerint az elemzés minden más formája mellékesnek minősíthető. A számok roppant tömegében elsikkad minden más módszer, amely megállapíthatná, hogy az emberek miért tesznek valamit, írnak valamit vagy alkotnak valamit. Ez a világ sosem fogadta szívesen a szellemi munka korábbi szakágazatait. Amint Berry (2011, 8) írja, az adatrengeteg hozadéka „egyensúlyt megbontó tömegű tudás és információ, a filozófia rendszerező ereje nélkül". Filozófia helyett – amelyet pedig Kant minden intézmény racionális alapjának tartott – „ekkor a komputálhatóság tekinthető egyfajta ontoteológiának, ez pedig az ontológia új 'korszakát' vezeti fel, melyet az érthetőség történelmileg új erőviszonyai határoznak meg" (Berry 2011, 12).

Az adatrengeteg érthetőségmodelljeivel kapcsolatban még azelőtt kell feltennünk a kínos kérdéseket, mielőtt új ortodoxiákká kristályosodnának. Fordhoz visszatérve: az ő újítása abban állt, hogy a korábban összefüggő, holisztikus feladatokat a gyártósor segítségével egyszerű, atomisztikus, mechanikus feladatokra bontotta. Ezt specializált eszközök tervezésével érte el, amelyek erősen meghatározták és behatárolták a munkás cselekvését. Az adatrengeteg specializált eszközeinek ugyanígy megvannak a maguk beépített határai és korlátai. A Twitter és a Facebook például olyan forrásai az adatrengetegnek, amelyekben az archiválás és visszakereshetőség funkcionalitása igen gyenge. Ennek következtében sokkal valószínűbb, hogy a kutatók a jelen vagy a legfrissebb múlt eseményeivel foglalkozzanak – például valamely választásra, tévésorozat évadzárójára vagy természeti katasztrófára adott reakciók követésével –, mivel a régebbi adatokhoz végletesen nehéz, vagy egyenesen lehetetlen hozzáférni.

Ha azt látjuk, hogy bizonyos fajta kutatói funkciók automatizálódnak, akkor figyelembe kell vennünk a gépi eszközök beépített hiányosságait is. Nem elég azt kérdeznünk – mint Anderson javasolta –, hogy „mit tanulhat a tudomány a Google-től?", hanem azt is firtatnunk kell, hogy az adatrengeteg leszüretelői hogyan alakíthatják át a „tudásszerzés" jelentését, és milyen új lehetőségeket és új korlátokat hozhatnak a megismerés ezen új rendszerei.

2. Az objektivitás és pontosság ígérete félrevezető

„Számok, számok, számok" – írja Latour (2009). „A szociológiának mániája lett az a cél, hogy kvantitatív tudománnyá váljon." Hogy a szociológia nem érte el e célját, az Latour szerint annak tudható be, hogy hol húzza meg a határt a szociális szférában a kvantifikálható és a nem kvantifikálható tudás között.

Az adatrengeteg új lehetőséget kínál a humán diszciplínáknak, hogy kvantitatív tudomány és objektív módszertan ígéretével lépjenek fel. A korábbinál sokkal több szociális tér kvantifikálását teszi lehetővé. Ám valójában az adatrengeteggel végzett munka még mindig szubjektív, és a benne számszerűsített anyag nem feltétlenül tart jogosabb igényt az objektív igazság státusára – különösen, ha közösségi webhelyek üzeneteiről van szó. Mégis fennmarad az a téveszme, hogy míg a kvalitatív kutatást végzők történeteket értelmeznek, addig a kvantitatív kutatással foglalkozók tényeket termelnek. Fennáll tehát a kockázat, hogy az adatrengeteg újrarója a tudományos módszerről és a társadalomtudomány és humán kutatás létjogosultságáról folytatott régesrégi vita kialakult határvonalait.

Az objektivitás mindig is a tudományfilozófia és a tudományos módszerről folytatott korai viták egyik kulcskérdése volt (Durkheim 1895). Az objektivitásra való hivatkozás azt sugallja, hogy az *objektumok*, az önmagukban és önmagukért létező dolgok szférájában horgonyzunk. A szubjektivitás ellenben gyanús dolog, lévén hogy az egyéni és szociális kondicionálás különböző formái adnak neki színezetet. A tudományos módszer arra törekszik, hogy eltávolítsa magát a szubjektív szférától. Ehhez a hipotézisek felvetésének és tesztelésének szenvtelen folyamatát használja, amely végül a tudásanyag javítását eredményezi. Azonban továbbra is óhatatlanul szubjektum marad az, aki az objektivitását hangoztatja, és ezen objektivitás szubjektív megfigyeléseken és választásokon alapul.

Minden kutató értelmezője is az adatoknak. Mint azt Gitelman (2011) megjegyzi, az adatokat először is el kell képzelni adatként, és az adatok ilyetén elképzelése interpretáción alapul: „minden diszciplínának és a hozzájuk tartozó minden intézménynek megvannak a maga normái és szabványai az adatok elképzeléséhez". Amióta számítástudósok is foglalkoznak társadalomtudományi tevékenységgel, megjelent az a hajlam, hogy munkájukat a tények – nem pedig az interpretáció – munkájának állítsák be. Lehet matematikailag megalapozott a modell, érvényesnek tűnhet a kísérlet, de amint a kutató értelmet keres az eredményeiben, megkezdődik az interpretáció folyamata. Ezzel nem azt akarjuk mondani, hogy minden interpretáció egyenlőnek teremtetett, inkább csak annyit, hogy nem minden szám semleges.

Szintén az interpretáció tövéről fakadnak a kísérlettervezési döntések arról, hogy mit kell mérni. Így például a szociális média adatainak feldolgozásakor sor kerül az „adattisztogatás" folyamatára is: ennek során döntjük el, mely tulajdonságokat és változókat vesszük számításba, és melyeket nem. Ez a folyamat természeténél fogva szubjektív. Boller fejtegetése szerint:

Az adatrengeteg – mint a nyers információ nagy tömege – nem magától értetődő. Az adatok értelmezésének konkrét módszertani megközelítései viszont sok filozófiai szempontból vitathatóak. Képviselhetnek-e „objektív igazságot" az adatok, avagy minden információba szükségszerűen beleviszi az elfogultságot valamilyen szubjektív szűrő vagy az adatok „tisztogatásának" módja? (2010, 13)

E kérdésen túlmenően problémát jelentenek az adathibák is. Az internetes forrásokból származó nagy adathalmazok sokszor bizonytalanok: gyakori bennük az adatkiesés és adatvesztés. Hibáikat és hiányaikat még jobban felnagyítja, ha több adathalmazt elemeznek együtt. A társadalomtudósok körében távolról sem újkeletű az adatgyűjtéssel kapcsolatos kritikus kérdések feszegetése és az adatok esetleges részrehajló torzulásának kiküszöbölésére való igyekezet (Cain–Finch 1981; Clifford–Marcus 1986). Ehhez az adott adathalmaz méretétől függetlenül annak sajátságait és határait kell megérteni. Lehet egy halmazban sokmilliónyi adat, de ez sem véletlenszerűségét, sem reprezentatív voltát nem szavatolja. Mielőtt statisztikai alapon állítanánk valamit egy adathalmazról, tudnunk kell, honnan származnak az adatok; ugyanilyen fontos, hogy tisztában legyünk a kérdéses adatok gyengéivel, és kezeljük ezeket. Mindezen túl az adatokról adott kutatói értelmezésünk elfogultságát is tudnunk kell kezelni. Ehhez pedig fel kell ismernünk, hogy egyéniségünk és látásmódunk képes befolyásolni az elemzésünket (Behar–Gordon 1996).

Az adatrengeteg túlságosan is könnyen enged teret az *apofréniának:* annak, hogy mintákat vegyünk észre ott, ahol valójában nincs minta – egyszerűen azért, mert az óriási adattömegben minden irányban szétsugárzó megfeleléseket lehet találni. Ennek egyik figyelemreméltó példájaként Leinweber (2007) demonstrálta, hogy az adatbányászat módszereivel erős – ámde valótlan – kapcsolatot lehet találni az S&P 500 tőzsdeindex és a bangladesi vajtermelés között.

Az interpretáció az adatelemzés kellős közepén foglal helyet. Bármennyi adattal dolgozzon is, az értelmezés korlátokkal és elfogultsággal jár. E korlátok és elfogultságok megértése és feltérképezése nélkül félreértelmezés születik. Az adatelemzés akkor a leghatékonyabb, ha a kutatók tekintettel vannak az adott adatok elemzésének hátteréül szolgáló összetett módszertani folyamatokra.

3. A több adat nem mindig jobb adat

A társadalomtudósok régóta érvelnek azzal, hogy munkájuk éppen azért kifogástalan, mert szisztematikus megközelítést alkalmaznak az adatgyűjtés és -elemzés során (McCloskey 1985). A néprajztudósok igyekeznek reflexióval ellensúlyozni értelmezéseik elfogultságát. A kísérleti tudósok kontrollt és standardizált kísérleteket használnak. A kérdőíves felmérések készítői szigorúan felülvizsgálják a mintavétel mechanizmusait és a kérdések torzítását. A kvantitatív kutatók a statisztikai szignifikanciát értékelik.

Így – és számtalan más módon – próbálják a társadalomtudósok egymás munkájának validitását felmérni. Pusztán abból, hogy az adatrengeteg rengeteg adatot kínál tálcán, még nem következik, hogy e módszertani megfontolások immár érdektelenek lennének. A minta mibenlétének megértése példának okáért fontosabb ma, mint valaha.

Például szolgálhat erre a Twitter egy statisztikai elemzés kontextusában. Mivel a Twitter adatait könnyű megszerezni – vagy „lekapirgálni"[3] –, a kutatók számos különböző mintázatot vizsgáltak már ezen a közösségi médiumon, pl. hangulatritmusokat (Golder–Macy 2011), médiaeseményekbe való bevonódást (Shamma et al. 2010), politikai felkeléseket (Lotan et al. 2011) és társalgási interakciókat (Wu et al. 2011). Bár sok humán tudós lelkiismeretesen kitér publikációiban a Twitter-adatok korlátaira, az ilyen kutatásokhoz kapcsolódó nyilvános diskurzus jobbára a rendelkezésre álló csipogások puszta számára koncentrál. Még amikor újsághír lesz az ilyen kutatásból, akkor is inkább azt hangsúlyozzák, hány millió „embert" vizsgáltak (Wang 2011).

A Twitter nem reprezentálja az „összes embert", és hiba azt feltételezni, hogy az „emberek" és a „Twitter-felhasználók" kifejezések egymás szinonimái – az utóbbi az előbbinek erősen sajátos részhalmaza. A Twittert használók populációja nem reprezentatív a globális sokaságra nézve, emellett a Twitter-fiókokat sem tekinthetjük egyenértékűnek a Twittert használókkal. Vannak több fiókkal rendelkező felhasználók, és vannak több ember által használt fiókok. Olyan emberek is vannak, akik nem készítenek saját fiókot, csak webes felületen használják a Twittert. A fiókok egy része *bot*, amely automatikusan – közvetlen emberi közreműködés nélkül – termeli az üzeneteket. Ráadásul az „aktív fiók" fogalma is problémás. Míg a felhasználók egy része gyakran tesz közzé tartalmat a Twitteren, mások inkább „hallgatózókként" vesznek részt a közösségben (Crawford 2009, p. 532). A Twitter Inc. közlése szerint az aktív felhasználók 40%-a csak azért jelentkezik be, hogy hallgatózzon (Twitter 2011). Kritikus felülvizsgálatra szorul tehát maga a „felhasználó", a „részvétel" és az „aktív" értelmezése is.

Szintúgy nem jelent az adatrengeteg teljes körű adatokat. Egy adathalmaz mérete önmagában érdektelen, ha nem vesszük figyelembe a mintavétel módját. Ha példának okáért egy kutató a csipogások téma szerinti gyakorisági eloszlását szeretné vizsgálni, de a Twitter kicenzúráz a forgalomból minden olyan csipogást, amelyben problematikus szó vagy tartalom jelenik meg (például pornográfia vagy *spam*), akkor a talált eloszlás pontatlan lesz. Akárhány csipogást dolgoz is fel kutatónk, a minta nem lesz reprezentatív, mivel az adatok kezdettől fogva torzulást tartalmaznak.

Akkor is nehéz tisztában lenni a minta sajátságaival, ha a forrás bizonytalan. A Twitter Inc. nyilvánosan elérhetővé teszi anyagának egy töredékét a Twitter API-n[4] keresztül. Ez az információs „tűzoltócső"[5] elméletileg minden, valaha elküldött nyilvános csipogást tartalmaz, miközben hangsúlyozottan hiányzik belőle minden olyan csipogás, amelyet a felhasználó privátnak vagy „védettnek" jelölt – csakhogy valójában hiányzik a tűzoltócsőből a publikusan elérhető csipogások egy része is. Emellett, bár maroknyi cég

[3] Az adatok „lekapirgálása" (scraping) itt az emberi olvasásra szánt, lazán szervezett adatok szigorúan strukturált, automatikus feldolgozásra alkalmas formában való kinyerését jelenti. *(A ford.)*

[4] Az API (Application Programming Interface) alkalmazásprogramozói felületet jelent, vagyis olyan eszközkészletet, amelynek segítségével a fejlesztők strukturált adatokhoz férhetnek hozzá.

[5] A Twitter API-ja a Firehose, vagyis Tűzoltócső nevet viseli, a rajta keresztül áramló hatalmas adatfolyamra utalva. *(A ford.)*

a teljes tűzoltócsőhöz hozzáférhet, a kutatók között nagyon kevesen élveznek ilyen szintű hozzáférést. Legtöbbjük vagy „slagot" (a nyilvános csipogások mintegy 10%-át) használhat, vagy „permetezőt" (a nyilvános csipogások mintegy 1%-át). Mások „fehérlistás" felhasználói fiókokon keresztül érhették el az API segítségével a publikus tartalom különféle részhalmazait.[6] Nem egyértelmű, hogy a különböző adatfolyamokba mely csipogások tartoznak bele, és milyen populációt reprezentál a belőlük vett minta. Lehetséges, hogy az API véletlenszerű mintát vesz a csipogásokból, vagy hogy minden órából az első néhány ezer csipogást emeli ki, vagy hogy csak a hálózati grafikon valamely szegmenséből veszi a csipogásokat. Az erre vonatkozó információ hiányában a kutatók nehezen állíthatnak bármit is az általuk elemzett adatok minőségéről. Reprezentatívak az adatok a csipogások összességére nézve? Nem, hiszen nincsenek köztük a védett fiókokhoz[7] tartozó csipogások. No de reprezentatívak-e az adatok a publikus csipogások összességére nézve? Talán, de nem feltétlenül.

A Twitter mára a tömeges adatbányászat közkedvelt forrása lett, ám a Twitter-adatokkal való munka súlyos módszertani kihívásokat von maga után, amelyekkel csak elvétve foglalkoznak az ilyen adatokat használók. Ha egy kutató adott adathalmazt vizsgál, meg kell értenie – és nyilvánosan tárgyalnia kell – nem csupán az adathalmaz korlátait, hanem annak korlátait is, hogy milyen kérdéseket intézhet ehhez a halmazhoz, és milyen értelmezéseket fogadhat el.

Különösen igaz ez akkor, ha a kutatáshoz több nagy adathalmazt vonnak össze. Ez nem azt jelenti, hogy ne nyújthatna értékes meglátásokat az adatok kombinálása: Acquisti és Gross (2009) kutatásai és más, hasonló munkák már csak azért is jelentősek, mert rámutatnak, hogyan vezethet több publikus adatbázis összevonása a magánszféra súlyos megsértéséhez – például emberek társadalombiztosítási azonosítójának kiderítéséhez. Azonban, amint arra Jesper Anderson – a FreeRisk nevű nyílt hozzáférésű pénzügyi adattár egyik alapítója – is rámutatott, a több forrásból vett adatok összevonása sajátos kihívásokat teremt. „Minden egyes forrásnak vannak hibalehetőségei… Azt hiszem, ezt a problémát csak felnagyítjuk [amikor több adathalmazt kombinálunk]" (Bollier 2010, 13).

Végezetül: a jelenlegi komputációs fordulat idején különösen fontos felismernünk a „kis adatok" értékét. A kutatás bármely szinten – egészen szerény léptékben is – eredményre vezethet. Olykor akár egyetlen személy tanulmányozása is rendkívül értékes lehet. Példázhatja ezt Veinot (2007) munkája, aki egyetlen dolgozót követett nyomon (egy vízenergiával dolgozó áramszolgáltató cég szerelőakna-ellenőrét), hogy megismerje egy kékgalléros munkás információszerzési gyakorlatát. E szokatlan tanulmány során Veinot az „információszerzési gyakorlat" definícióját új keretek közé illesztette, eltávolodva az újítások iránt fogékony fehérgalléros dolgozókra való összpontosítás bevett szokásától, és az irodai és városi kontextuson kívül eső terekhez közelítve. Munkája olyan történetet mesél, amelyet több millió Facebook- vagy Twitter-fiók tömeges elemzésével sem lehetett volna felderíteni, és a vizsgálati alanyok lehető legkisebb elemszáma ellenére jelentősen hozzájárul a maga kutatási területéhez. A vizsgált adatok terjedelmének a kutatói kérdéshez kell illeszkednie – előfordul, hogy a kisebb jobb.

[6] A Twitter által nyújtott hozzáférés részletes leírása itt található: https://dev.Twitter.com/docs/streaming-api/methods. A fehérlistás fiókokat korábban széles körben használták a kutatók, de ezek már nem elérhetőek.

[7] A védett fiókok részaránya ismeretlen, bár az ilyen fiókok azonosítására tett kísérletek arra utalnak, hogy az összes felhasználói fiók kevesebb mint 10%-a védett (Meeder et al. 2010).

4. Kontextus nélkül az adatrengeteg értelmét veszti

Mivel a nagy adathalmazok modellezhetőek, ezért az adatokat gyakran úgy redukálják, hogy valamely matematikai modellhez illeszkedjenek. Azonban a kontextusától függetlenített adat értelme és értéke csökken. A közösségi hálózati webhelyek terjedése a „szociális grafikon" iránti mániákus, iparvezérelt érdeklődést váltott ki. A kutatók ezrével vetették rá magukat a Twitterre, a Facebookra és más közösségi médiumokra, hogy kielemezzék az üzenetek és fiókok közti kapcsolatokat és következtetéseket vonjanak le a szociális hálózatokról. A közösségi médiában megjelenő kapcsolatok azonban nem feltétlenül egyenértékűek a szociológusok és antropológusok kutatásaiban már az 1930-as évek óta használt szociogramokkal és rokonsági hálózatokkal (Radcliffe-Brown 1940; Freeman 2006). Az, hogy az emberek kapcsolatait meg lehet jeleníteni grafikon formájában, még nem jelenti azt, hogy e grafikon mindent elmond a kérdéses kapcsolatokról.

A szociológusok és antropológusok eleddig kérdőívek, interjúk, megfigyelések és kísérletek segítségével gyűjtöttek adatokat az emberek kapcsolatairól. Adataik alapján írták le az emberek „személyes hálózatait": az egyes emberek által létrehozott és fenntartott összeköttetések halmazát (Fischer 1982). E hálózatokból emelték ki az idők során kifejlesztett kritériumok alapján történő értékeléssel a személyes kapcsolatokat. Az adatrengeteg hozadéka két újfajta, közkedvelt szociális hálózat, amelyek adatnyomokból vezethetőek le: a „deklarált hálózat" és a „viselkedési hálózat".

A deklarált hálózat azokat a kapcsolatokat képezi le, amelyeket az emberek különféle műszaki infrastruktúrákban – például e-mailes és mobiltelefonos címtárukban, azonnali üzenetküldők barátlistáiban, közösségi webhelyek ismerőslistáiban és más szociális médianemek „követő"-listáiban – rögzítenek, vagyis deklarálnak. Számtalan oka lehet, hogy egy személy miért vesz fel valakit a különféle listákra. A végeredmény az, hogy az ilyen listákon egyaránt szerepelhetnek barátok, kollégák, felületes ismerősök, híresség, barátok barátai, közéleti személyiségek, sőt érdekes idegenek.

A viselkedési hálózatok alapjául a kommunikációs mintázatok, a mobiltelefonos cellapozíció és a közösségi médiában folytatott interakciók szolgálnak (Onnela et al. 2007; Meiss et al. 2008). Az ilyen hálózatokon szerepelhetnek olyan emberek, akik SMS-ezni szoktak egymással, olyanok, akiket a Facebookon közös fotón azonosítottak, olyanok, akik e-mailen leveleznek, és olyanok, akik egyazon térben tartózkodnak – legalábbis a mobiltelefonjuk pozíciója szerint.

Mind a viselkedési, mind a deklarált hálózatok igen értékesek a kutatók számára, ám mégsem egyenértékűek a személyes hálózatokkal. Így például a „kötődés ereje" – bár ennek értéke vitatott – az egyes kapcsolatok fontosságának jelzésére szolgál (Granovetter 1973). Ha a mobiltelefonok adatai azt mutatják, hogy egy dolgozó ember több időt tölt a kollégáival, mint a házastársával, ez nem feltétlenül jelenti azt, hogy a kollégák fontosabbak számára, mint a párja. Gyakori hiba, hogy a kötődés erejét gyakoriság alapján vagy a nyilvános deklarálások száma alapján próbálják mérni. Csakhogy a kötődés ereje – és számos, köré épült elmélet – kifinomultabb értékelést kíván annak alapján, hogy az adott ember hogyan értelmezi és értékeli a másokkal való kapcsolatait. Nem minden összeköttetés egyenértékű az összes többivel, és az érintkezés gyakori-

sága sem jelzi a kapcsolat erősségét. Sőt, még az összeköttetés hiánya sem feltétlenül jelenti azt, hogy nem létezik kapcsolat.

Az adat nem konfekcióáru. Értékes lehet az absztrahált adatok elemzése, de a kontextus megtartása – különösen bizonyos kutatási témák esetén – továbbra is kulcsfontosságú. Nagy tömegű adat esetén – és még inkább, ha az adatokat modellhez igazodva redukáljuk – a kontextus értelmezése nehéz. Az adatrengeteg vonatkozásában állandó kihívást jelent majd a kontextus kezelése.

5. Csak mert hozzáférhető, még nem biztos, hogy etikus

2006-ban egy, a Harvardon dolgozó kutatócsoport begyűjtötte 1700 felsőoktatásban tanuló Facebook-felhasználó profilját, hogy a diákok érdeklődési körének és baráti kapcsolatainak időbeni változását tanulmányozza (Lewis et al. 2008). Az elvileg anonim adatokat közzé tették a nagyvilágnak, hogy más kutatók is vizsgálhassák és elemezhessék őket. A többiek csakhamar felfedezték, hogy az adathalmaz egyes részeinek anonimitása megszüntethető, s ezzel sérül az adatgyűjtésről mit sem tudó diákok magánszférája (Zimmer 2008).

Az eset a lapok címoldalára került, a tudósokat pedig fogas kérdések elé állította: mi hát a közösségi webhelyeken lévő, úgynevezett „publikus" adatok státusa? Fel szabad ezeket használni külön engedély nélkül? Mi lenne a kutatók számára a helyes etikai gyakorlat? A személyi adatok védelméért kampányolók szemében ez a terület ma már a legfontosabb frontok egyike, ahol erősebben kell védeni a magánszférát. Külön nehézség, hogy a magántitoksértést nehéz konkrétan megfogalmazni. Történt az adott pillanatban károkozás? És történik-e 20 év múlva? „Minden emberi alanyokra vonatkozó adat kapcsán óhatatlanul felmerül a magánszféra kérdése, és a személyi adatokkal való visszaélés tényleges kockázatát nehéz számszerűsíteni" (*Nature*, idézi Berry 2011).

Az intézményi kutatásetikai bizottságok (IKEB-ek) – és hasonló etikai testületek – az 1970-es években alakultak meg a humán kutatások felügyelete céljából. Az IKEB-ek célja – bár ennek megvalósítása tagadhatatlanul problematikus (Schrag 2010) – az, hogy keretet állítson fel, amelyben egy-egy adott vizsgálat etikus volta értékelhető, valamint hogy fékek és ellensúlyok rendszerével biztosítsa a vizsgálati alanyok védelmét. Az „informált beleegyezés" gyakorlatához hasonló eszközök a válaszadók személyi adatainak védelmével együtt arra szolgálnak, hogy felvértezzék a vizsgálatok résztvevőit az orvos- és társadalomtudományban korábban előforduló visszaélések ellen (Blass 2004; Reverby 2009). Bár az IKEB-ek nem képesek minden egyes vizsgálat károkozó potenciálját megjósolni (és sajnos nem ritkán az etikától teljesen független indokokkal akadályozzák meg a kutatók munkáját), kétségtelen hasznuk az, hogy magukat a kutatókat is munkájuk etikai vonatkozásainak kritikus átgondolására késztetik.

Nagyon kevéssé vagyunk tisztában az adatrengeteg jelenségének hátterében meghúzódó etikai tényezőkkel. Szabad-e egy embert egy óriási adathalmaz részeként kezelnünk? Mi lesz, ha valakinek a „nyilvános" blogbejegyzését megfosztjuk a kontextusától, és a szerző által sohasem sejtett módon elemezzük? Mit jelent az embernek, ha rivaldafénybe kerül, vagy ha tudta nélkül elemzik ki? Ki a felelős azért, hogy a kutatási folyamat ne sértsen se egyéneket, se közösségeket? Hogy nézhet itt ki az informált beleegyezés?

Talán ésszerűtlen elvárnunk a kutatóktól, hogy beleegyező nyilatkozatot szerezzenek mindenkitől, aki csipogott valamit a Twitteren, de az is aggasztó, ha a kutatók vizsgálódásuk etikus voltát hangoztathatják pusztán azért, mert hozzáférhető adatokat használnak. Az, hogy valamilyen tartalom nyilvánosan elérhető, még nem jelenti azt, hogy bárki bármire használhatja. Az online adatok begyűjtésének és elemzésének komoly etikai vonzatai vannak (Ess 2002). Nem lehet eltekinteni a kutatás etikai bírálatától pusztán azért, mert az adatok látszólag publikusak. A tudósoknak folyamatosan kérdőre kell vonniuk önmagukat – és kollégáikat – arról, hogy adatgyűjtési, -elemzési és publikációs módszereik mennyire etikusak.

Az etikus gyakorlathoz fontos, hogy a kutatók eltöprengjenek az elszámoltathatóság jelentőségéről – mind a szakterületüket, mind a kutatás alanyait illetően. Elszámoltathatóság alatt a magánszféra védelménél szélesebb körű fogalmat értünk, ahogy azt Troshynski és társai (2008) írták le. Ez érvényes lehet akkor is, amikor a magánszféra hagyományos értelemben vett megsértése nem merül fel. Az elszámoltathatóság többirányú kapcsolat: tartozhat valaki ilyen kötelezettséggel a feletteseinek, kollégáinak, kutatása résztvevőinek, sőt a nyilvánosságnak is (Dourish–Bell 2011). Amikor egyetemi tudósok emberi alanyokkal dolgoznak, a szakterületükre vonatkozó sajátos szabványok kényszerítik őket válaszadóik jogainak és jóllétének megóvására. Csakhogy az etikai bizottságok gyakran az adatrengeteg bányászatának és anonimizálásának folyamataival sincsenek tisztában, nemhogy azzal, hogy milyen hibák eredményezhetik az adatok személy szerinti azonosíthatóságát. Az elszámoltathatóság érvényesítéséhez nem elegendő azt gondolni, hogy az etikai bizottságok majd úgyis gondoskodnak az emberek védelméről, hanem komolyan fontolóra kell venni az adatrengeteg jelenségének burjánzó következményeit.

Igen jelentős az adatrengeteg tanulmányozásában az igazság, irányítás és hatalom kérdése is: a kutatók rendelkeznek eszközökkel és hozzáféréssel, a közösségi média felhasználói pedig általában véve nem. Ők fokozottan kontextusérzékeny térben hozták létre adataikat, és teljesen elképzelhető, hogy a felhasználók egy része nem engedélyezné, hogy ezen adatokat más környezetben felhasználják. Sokan tudatában sincsenek, hogy bármely pillanatban milyen sok és sokféle közvetítő és algoritmus gyűjti és tárolja adataikat jövőbeli felhasználás céljából. A felhasználók ritkán gondolnak kutatókra, amikor közönségüket elképzelik. Annak sincsenek feltétlenül tudatában, hogy az általuk közzétett információ mennyiféle módon használható anyagi vagy egyéb nyereség szerzésére. Meglehet, hogy publikusnak (vagy részben publikusnak) nyilvánítják adataikat, de ezt nem lehet egy kalap alá venni azzal, hogy bármiféle felhasználást szabadon engedélyeznének. Az adatrengeteg kutatói ritkán ismerik el, hogy a nyilvános jelenlét (mint amikor valaki egy parkban üldögél) távolról sem ugyanazt jelenti, mint a nyilvánosság (vagyis a figyelem aktív felhívása) (boyd–Marwick 2011).

6. Az adatrengeteg hozzáférhetőségének korlátozása új digitális határokat teremt

Golder az adatrengetegről szóló esszéjében (2010) Homans szociológust (1974) idézi: „A társadalomtudomány módszerei időben és pénzben is drágák, és napról napra drágábbak." Adatokat gyűjteni mindeddig nehéz, idő- és erőforrás-igényes munka volt.

Az adatrengeteg iránti lelkesedés nem kis részben abból a vélekedésből táplálkozik, hogy most már óriási adattömegekhez lehet könnyedén hozzáférni.

No de ki fér hozzájuk? Milyen céllal? Milyen kontextusban? És milyen megszorításokkal? Elnézve, milyen robbanásszerűen szaporodtak el a közösségi médiából eredő adatokat felhasználó kutatások, azt gondolhatnánk, hogy könnyű ilyen adatokat szerezni – ám korántsem ez a helyzet. Amint Manovich (2011) megjegyzi, „csakis a közösségi médiát szolgáltató cégek férhetnek hozzá az igazán nagy tömegű szociális adatokhoz – különösen a dinamikus (más néven tranzakciós) adatokhoz. A Facebook alkalmazásában álló antropológus vagy a Google alkalmazásában álló szociológus olyan adatokhoz férhet hozzá, amelyek tudóstársai számára elérhetetlenek." Vannak cégek, amelyek teljesen korlátozzák adataik elérhetőségét; mások díjat szabnak a hozzáférésért; megint mások szűkített adathalmazokat tesznek elérhetővé az egyetemi kutatók számára. Ez erősen egyenlőtlenné teszi a rendszert: akinek van pénze (vagy aki bennfentes egy cégnél), az máshogyan kutathat, mint a többiek. Aki nem rendelkezik hozzáféréssel, az sem megismételni, sem bírálni nem tudja azokat a módszereket, melyekről privilegizált szaktársai számolnak be.

Azt is tudatosítanunk kell, hogy az adatrengeteg arisztokratáinak osztályát az egyetemek rendszere is erősíteni fogja: az elsőrangú, bőséges forrásokkal rendelkező egyetemek megengedhetik maguknak, hogy adathozzáférést vásároljanak, és e kiemelt egyetemek diákjait hívják majd legnagyobb eséllyel dolgozni a nagy közösségimédia-szolgáltatók. A perifériára szorultak kevesebb állásajánlatot kapnak, és még kevesebb lehetőségük lesz a szakmai fejlődésre. A társadalomtudósokat megosztó hasadékok így még sokkal jobban kiszélesednek.

A hozzáférés kérdésén túl felmerül a hozzáértés kérdése is. Az API-val való birkózás, a hatalmas adattömegek lekapirgálása és elemzése olyan készségeket igényel, amelyekkel jobbára csak a számítástechnikai háttérrel rendelkező tudósok bírnak. Ha a számítástudományi szakértelmet fogadjuk el legfontosabbnak, kérdésessé válik, hogy ebben a kontextusban ki indul előnyös, és ki hátrányos helyzetből. Így voltaképpen új hierarchiák rögzülnek aszerint, hogy „ki tud olvasni a számokból" – annak elismerése helyett, hogy mind a számítástudósok, mind a társadalomtudósok nézőpontja értékes meglátásokra vezethet. Nem közömbös az sem, hogy ez az elhatárolódás egyben nemek szerinti elhatárolódás is. A számítástechnikában járatos kutatók zöme jelenleg férfi, és – mint azt a feminista történészek és tudományfilozófusok igazolták – a vizsgált kérdéseket meghatározza az, hogy ki kérdez (Harding 2010; Forsythe 2001). Igen összetett az a kérdéskör, hogy milyen kutatói készségek lesznek értékesek a jövőben és hogyan taníthatóak e készségek. Hogyan képezhetünk olyan diákokat, akik az algoritmusokban és adatelemzésben ugyanúgy eligazodnak majd, mint a szociális analízis és társadalomelmélet terén?

Végezetül: az adatrengeteghez való hozzáférés nehézsége és költsége a kutatási eredmények kultúráját is korlátozza. A nagy adatkezelő cégeket semmi sem kötelezi adataik közzétételére, ellenben teljességgel megszabhatják, ki férhessen hozzá ezekhez. A védett adathalmazokhoz hozzáféréssel bíró adatrengeteg-kutatók – attól tartva, hogy elveszíthetik e hozzáférésüket – kisebb valószínűséggel választanak olyan kérdéseket, amelyek a közösségimédia-szolgáltatók érdekeit csorbíthatják. Az adatrengeteg jövőjéről elmélkedve mindenképpen fontolóra kell vennünk ezt a nyomasztó folyamatot, amely kihat arra, hogy miféle kérdésekkel foglalkozhat – akár nyilvánosan, akár privát keretek között – a kutatás.

Az adatrengeteg körül kialakult jelenlegi ökoszisztéma új digitális határvonalat teremt: elkülöníti az adatokban gazdagokat az adatokban szegényektől. Egyes céges kutatók már egyenesen azt is felvetették, hogy az egyetemi kutatók talán ne is vesződjenek a közösségi média adatainak tanulmányozásával. Jimmy Lin professzor, miután kutatószabadságát a Twitternél folytatott ipari gyakorlattal töltötte, azt indítványozta, hogy akadémikusok ne foglalkozzanak olyan kutatással, amelyet az iparban dolgozók „jobban is el tudnak végezni" (Conover 2011). Bár nem újdonság, hogy valaki ilyen nyíltan próbál határt vonni a „bennfentes" és „kívülálló" kutatók között, az ilyesmi mindenképpen kárára válik a kutatói közösségnek. „A demokratizálás hatékonyságát – állította Derrida (1996) – mindig mérhetjük e lényegi kritériummal: ki kerül be az archívumba, ki fér hozzá, miből áll és hogyan értelmezik" (4).

Ha egy rendszerbe expliciten beleíródnak az egyenlőtlenségek, az elkerülhetetlenül osztályalapú struktúrákat eredményez. Manovich (2011) szerint az adatrengeteg birodalmában az embereknek három osztályuk van: „akik adatokat alkotnak (akár tudatosan, akár digitális lábnyomuk hátrahagyásával), akiknek módjukban áll az adatokat begyűjteni, és akik képesek kielemezni azokat". Tudjuk, hogy az utolsó csoport a legkisebb és a leginkább privilegizált: ők egyszersmind azok, akik az adatrengeteg felhasználásának szabályait írhatják, és megszabhatják, ki vehet részt a folyamatban. Az intézményi egyenlőtlenség vélhetően semmiféle meglepetést nem okoz akadémiai körökben, mégis szükség van a vizsgálatára és kritikájára, mivel az effajta egyenlőtlenség torzítja az adatokat és befolyásolja a jövőbeni kutatás típusait.

Amikor azt mondjuk, hogy az adatrengeteg jelensége ludas lehet több nagyívű történelmi és filozófiai változásban, távolról sem állítjuk, hogy e változásoknak ez az egyetlen oka; a komputációs fordulatnak nem az akadémia világa az egyetlen hajtóereje. Mind a kormányok, mind az ipar keményen igyekszik a lehető legtöbb adatot gyűjteni és belőlük maximális információt kivonni – akár célzott hirdetés, akár terméktervezés, közlekedésszervezés vagy bűnmegelőzés végett. Igenis állítjuk viszont, hogy az adatrengeteg operacionalizálása komoly és széles körű következményekkel jár, és kihat a kutatások jövőbeli témaválasztására is. Amint Suchman (2011) – Levi-Strauss nyomán – megjegyzi, „a szerszámaink mi vagyunk". Át kell gondolnunk, hogy a szerszámok hogyan alakítják velünk együtt a világot, miközben használjuk őket. Az adatrengeteg kora még alig hogy elkezdődött, de máris fontos, hogy a kutatás ezen új hullámának előfeltevéseit, értékeit és előítéleteit megkérdőjelezzük. A tudás előállítására hivatott tudósokként az effajta kérdések felvetése munkánk lényegi elemei közé tartozik.

Köszönetnyilvánítás

Köszönetet szeretnénk mondani Heather Casteelnek a cikk elkészítéséhez nyújtott segítségéért. Mélyen hálásak vagyunk továbbá Eytan Adarnak, Tarleton Gillespienek, Bernie Hogannak, Mor Naamannak, Jussi Parikkának, Christian Sandvignek, valamint a Microsoft Research Social Media Collective minden tagjának az ösztönző beszélgetésekért, tanácsaikért és visszajelzésükért. Hálával tartozunk azoknak is, akik az Oxford Internet Institute tizedik évfordulója alkalmával adtak visszajelzést. Végezetül köszönjük névtelen lektoraink hasznos megjegyzéseit.

Balogh Dániel fordítása

Irodalom

Acquisti, A. & Gross, R. (2009) 'Predicting social security numbers from public data', *Proceedings of the National Academy of Science*, vol. 106, no. 27, pp. 10975–10980.

Anderson, C. (2008) 'The end of theory, will the data deluge makes the scientific method obsolete?', *Edge*, [Online] Elérhető: http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (2011. július 25.).

Anderson, M. (1988) *The American Census: A Social History*. Yale University Press, New Haven, CT.

Baca, G. (2004) 'Legends of Fordism: between myth, history, and foregone conclusions', *Social Analysis*, vol. 48, no. 3, pp. 169–178.

Behar, R. & Gordon, D. A. (eds) (1996) *Women Writing Culture*. University of California Press, Berkeley, CA.

Berry, D. (2011) 'The computational turn: thinking about the digital humanities'. *Culture Machine*, vol. 12, [Online] Elérhető: http://www.culturemachine.net/index.php/cm/article/view/440/470 (2011. július 11.).

Blass, T. (2004) The Man Who Shocked the World: The Life and Legacy of Stanley Milgram, Basic Books, New York.

Bollier, D. (2010) 'The promise and peril of big data'. [Online] Elérhető: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (2011. július 11.)

Bowker, G. C. (2005) *Memory Practices in the Sciences*. MIT Press, Cambridge, MA.

Boyd, D. & Marwick, A. (2011) 'Social privacy in networked publics: teens' attitudes, practices, and strategies', az *Oxford Internet Institute*-ban elhangzott előadás. [Online] Elérhető: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1925128 (2011. szeptember 28.).

Burkholder, L. (ed.) (1992) *Philosophy and the Computer*, Westview Press, Boulder, San Francisco és Oxford.

Cain, M. & Finch, J. (1981) 'Towards a rehabilitation of data' in *Practice and Progress: British Sociology* 1950–1980, eds P. Abrams, R. Deem, J. Finch & P. Rock, George Allen and Unwin, London, pp. 105–119.

Clifford, J. & Marcus, G. E. (eds) (1986) *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA.

Conover, M. (2011) 'Jimmy Lin', *Complexity and Social Networks Blog* [Online] Elérhető: http://www.iq.harvard.edu/blog/netgov/2011/07/the_international_conference_o.html (2011. december 9.)

Crawford, K. (2009) 'Following you: disciplines of listening in social media', *Continuum: Journal of Media & Cultural Studies*, vol. 23, no. 4, pp. 532–533.

Derrida, J. (1996) *Archive Fever: A Freudian Impression*. Ford.: Eric Prenowitz, University of Chicago Press, Chicago.

Dourish, P. & Bell, G. (2011) Divining a Digital Future: Mess and Mythology in Ubiquitous Computing, MIT Press, Cambridge, MA.

Du Gay, P. & Pryke, M. (2002) Cultural Economy: Cultural Analysis and Commercial Life, Sage, London.

Durkheim, E. (1895/1982) *Rules of Sociological Method*. The Free Press, New York, NY.

Ess, C. (2002) 'Ethical decision-making and Internet research: recommendations from the aoir ethics working committee'. *Association of Internet Researchers* [Online] Elérhető: http://aoir.org/reports/ethics.pdf (2011. szeptember 12.)

Fischer, C. (1982) *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago, Chicago.

Forsythe, D. (2001) *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press, Stanford.

Freeman, L. (2006) *The Development of Social Network Analysis*. Empirical Press, Vancouver.

Fry, J. P. & Sibley, E. H. (1996) [1974] 'Evolution of database management systems'. *Computing Surveys*, vol. 8, no. 1.1, pp. 7–42. Reprint (1996): *Great Papers in Computer Science*, ed. L. Laplante, IEEE Press, New York.

Gitelman, L. (2011) *Notes for the Upcoming Collection 'Raw Data' is an Oxymoron* [Online] Elérhető: https://files.nyu.edu/lg91/public/ (2011. július 23.)

Golder, S. (2010) 'Scaling social science with hadoop', *Cloudera Blog* [Online] Elérhető: http://www.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop/ (2011. június 18.)

Golder, S. & Macy, M.W. (2011) 'Diurnal and seasonalmood vary withwork, sleep and daylength across diverse cultures'. *Science*, vol. 333, no. 6051, pp. 1878–1881, [Online] Elérhető: http://www.sciencemag.org/content/333/6051/1878.

Granovetter, M. S. (1973) 'The strength of weak ties'. *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380.

Harding, S. (2010) 'Feminism, science and the anti-Enlightenment critiques'. In: *Women, Knowledge and Reality: Explorations in Feminist Philosophy*, eds A. Garry & M. Pearsall, Unwin Hyman, Boston, MA, pp. 298–320.

Homans, G. C. (1974) *Social Behavior: Its Elementary Forms*. Harvard University Press, Cambridge, MA.

Kranzberg, M. (1986) 'Technology and history: kranzberg's laws'. *Technology and Culture*, vol. 27, no. 3, pp. 544–560.

Latour, B. (2009) 'Tarde's idea of quantification'. In *The Social after Gabriel Tarde: Debates and Assessments*, ed. M. Candea, Routledge, London, pp. 145–162 [Online] Elérhető: http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf (2011. június 19.)

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Van Alstyne, M. (2009) 'Computational social science'. *Science*, vol. 323, no. 5915, pp. 721–723.

Leinweber, D. (2007) 'Stupid data miner tricks: overfitting the S&P 500'. *The Journal of Investing*, vol. 16, no. 1, pp. 15–22.

Lessig, L. (1999) *Code: and Other Laws of Cyberspace*. Basic Books, New York, NY.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. & Christakis, N. (2008) 'Tastes, ties, and time: a new social network dataset using Facebook.com'. *Social Networks*, vol. 30, no. 4, pp. 330–342.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. & boyd, D. (2011) 'The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions'. *International Journal of Communications*, vol. 5, pp. 1375–1405 [Online] Elérhető: http://ijoc.org/ojs/index.php/ijoc/article/view/1246.

Manovich, L. (2011) 'Trending: the promises and the challenges of big social data'. In *Debates in the Digital Humanities* ed. M. K. Gold, The University of Minnesota Press, Minneapolis,

MN, [Online] Elérhető: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (2011. július 15.)

McCloskey, D. N. (ed.) (1985) 'From methodology to rhetoric'. *The Rhetoric of Economics*. University of Wisconsin Press, Madison, pp. 20–35.

Meeder, B., Tam, J., Gage Kelley, P. & Faith Cranor, L. (2010) 'RT@IWantPrivacy: widespread violation of privacy settings in the Twitter social network'. A *Web 2.0 Security and Privacy* (W2SP 2011) alkalmából elhangzott előadás, Oakland, CA.

Meiss, M. R., Menczer, F. & Vespignani, A. (2008) 'Structural analysis of behavioral networks from the Internet'. *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, pp. 220–224.

Moretti, F. (2007) *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, London.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. & Barabási, A. L. (2007) 'Structure and tie strengths in mobile communication networks'. *Proceedings from the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336.

Pariser, E. (2011) *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, New York.

Radcliffe-Brown, A. R. (1940) 'On social structure'. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, vol. 70, no. 1, pp. 1–12.

Reverby, S. M. (2009) *Examining Tuskegee: The Infamous Syphilis Study and Its Legacy*. University of North Carolina Press, Chapel Hill, NC.

Savage, M. & Burrows, R. (2007) 'The coming crisis of empirical sociology'. *Sociology*, vol. 41, no. 5, pp. 885–899.

Schrag, Z. M. (2010) *Ethical Imperialism: Institutional Review Boards and the Social Sciences*, 1965–2009. Johns Hopkins University Press, Baltimore, MD.

Shamma, D. A., Kennedy, L., & Churchill, E. F. (2010) 'Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? A *Computer-Supported Cooperative Work-2010* alkalmából elhangzott előadás, *Association for Computing Machinery*, 2010. február 6–10., Savannah, Georgia USA. Elérhető: http://research.yahoo.com/pub/3041.

Suchman, L. (2011) 'Consuming anthropology'. In *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences* eds A. Barry & G. Born, Routledge, London [Online] Elérhető: http://www.lancs.ac.uk/fass/doc_library/sociology/Suchman_consuming_anthroploogy.pdf.

Troshynski, E., Lee, C. & Dourish, P. (2008) 'Accountabilities of presence: reframing location-based systems'. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008. április 5–10., Firenze, Olaszország.

Twitter (2011) 'One hundred million voices'. *Twitter Blog* [Online] Elérhető: http://blog.Twitter.com/2011/09/one-hundred-million-voices.html (2011. szeprember 12.)

Veinot, T. (2007) 'The eyes of the power company: workplace information practices of a vault inspector'. *The Library Quarterly*, vol. 77, no. 2, pp. 157–180.

Wang, X. (2011) 'Twitter posts show workers worldwide are stressed out on the job'. *Bloomberg Businessweek* [Online] Elérhető: http://www.businessweek.com/news/2011-09-29/Twitter-posts-show-workers-worldwide-are-stressed-out-on-the-job.html (2012. március 12.)

Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J. (2011) 'Who says what to whom on Twitter', Az *International World Wide Web Conference (WWW 2011)* konferenciakiadványa, március 28–április 1., Hyderabad, India, pp. 705–714.

Zimmer, M. (2008) 'More on the "Anonymity" of the Facebook dataset – it's Harvard College'. *MichaelZimmer.org Blog* [Online] Elérhető: http://www.michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/ (2011. június 20.)