Társadalomtudomány a Big Data korában*

Ságvári Bence,
az MTA Társadalomtudományi Kutatóközpont tudományos főmunkatársa
E-mail: Sagvari.Bence@tk.mta.hu

Napjainkban soha nem látott mennyiségben és részletezettségben állnak rendelkezésre adatok az emberek viselkedésével kapcsolatban; a „digitális lábnyomok" összegyűjtésére, összekapcsolására, értékesítésére és elemzésére hivatott iparág rohamtempóban fejlődik. Ezek az adatok azonban egyre inkább arra is alkalmasak, hogy segítségükkel olyan „klasszikus" társadalmi kérdéseket is vizsgáljunk, amelyek több mint egy évszázada a társadalomtudományok érdeklődésének középpontjában vannak. Egyre több lehetőség van arra, hogy olyan kérdéseket is megpróbáljunk megválaszolni, amelyek korábban elemezhető adatok hiányában nem voltak kutathatók. Ez a folyamat azonban nem feltétlenül a bevett ismeretelméleti hagyományok, illetve az egyes tudományterületek közötti megszokott szereposztások szerint fog végbe menni. A szerző a társadalomtudományok (és ezen belül elsősorban a szociológia) előtt álló fontosabb kihívásokat és lehetőségeket tekinti át.

TÁRGYSZÓ:
Big Data.
Társadalomtudományok.
Ismeretelmélet.

DOI: 10.20311/stat2017.05.hu0491

* A tanulmány az OTKA K112713 számú („Egy online közösségi hálózat életciklusa: big data elemzés") kutatási projekt keretében készült.

Napjainkban soha nem látott mennyiségben és részletezettségben állnak rendelkezésre adatok az emberek viselkedésével kapcsolatban. Százmilliók napi rutinja, a műholdak, WiFi- és mobilhálózatok által követett mozgása, kommunikációs szokásai, társas kapcsolatai, bevásárlókosarának tartalma, online keresései, hírfogyasztása – hogy csak a leggyakoribb példákat említsük – ismerhető meg közel valós időben, akár a teljes népességre vonatkozóan. Az emberek legkülönbözőbb digitális lábnyomainak összegyűjtésére, összekapcsolására, értékesítésére és elemzésére hivatott iparág rohamtempóban fejlődik, legyen szó az egyszerű célzott reklámokról, kockázatelemzésről vagy éppen kifinomult, tömeges megfigyelési technikákról (*Van Es–Schäfer* [2017]). Ezek az adatok azonban egyre inkább arra is alkalmasak, hogy segítségükkel olyan összetett társadalmi jelenségeket is vizsgáljunk, mint például a társas kapcsolatok és a kommunikáció, az egyenlőtlenségek, az oktatás, az egészségügy, a politikai részvétel vagy más, ehhez hasonló olyan „klasszikus" kérdéseket, amelyek több mint egy évszázada a társadalomtudományok érdeklődésének középpontjában vannak. Sőt, ennek az adatiparnak a fejlődése arra is lehetőséget biztosít, hogy olyan kérdéseket próbáljunk megválaszolni, amelyek korábban elemezhető adatok hiányában nem voltak kutathatók. Ehhez hozzájárulnak azok az egyre kifinomultabb technikák is, amelyek segítségével a korábban csak külön-külön, önmagukban értelmezhető adatforrások mikroadatszintű összekapcsolásával új, komplex elemzési lehetőségek válnak elérhetővé. Mindez ugyanakkor azt is jelenti, hogy ezeknek a komplex, sok esetben strukturálatlan adatoknak a „megszelídítéséhez" és „szóra bírásához" újfajta adatfeldolgozási, -elemzési módszerekre, valamint a tudományterületek, illetve a tudomány, az üzleti szféra és a kormányzatok közötti együttműködések újragondolására van szükség. Vérmérséklettől függően találkozhatunk különböző álláspontokkal azzal kapcsolatban, hogy vajon az adatforradalom – amelynek pillanatnyilag szemtanúi vagyunk – mennyiben tekinthető egy olyan paradigmaváltás részének, ahol a megszokott ismeretelméleti, kutatás-módszertani sorvezetőinket alapjaiban kell újra gondolnunk (*Mayer-Schönberger–Cukier* [2013], *Meyer–Schroeder* [2015]). Az adatvezérelt iparágak és kutatási területek jelenlegi fejlődését látva aligha kétséges, hogy ennek rendkívüli hatásai lesznek a gazdaságok és a társadalmak működésére éppúgy, mint a hivatalos statisztikai adatgyűjtésekre, a társadalomtudományok lehetőségeire, az egyes tudományágak közötti erőviszonyok átalakulására, továbbá a kormányzatok, a vállalatok és a tudományos intézetek együttműködési lehetőségeire. Napjainkban azonban még sokszor érezhetjük úgy, hogy – főleg az akadémiai szférán kívülről érkező – „pozitivista próféták" hajlamosak túlértékelni ennek pillanatnyi jelentőségét, illetve csak a jelenség pozitív társadalmi-gazdasági következményeire fókuszálnak. Ez az írás ennek az „új adatvilág-

nak" néhány olyan jellemzőit tekinti át, amelyek hatással lehetnek az empirikus társadalomtudományok (és ezen belül is elsősorban a szociológia) jövőjére.

1. A Big Data ígérete és valósága

A bevezetőben említett komplex jelenség körülírására ma már rutinszerűen a Big Data kifejezés használatos, ez azonban sok szempontból félrevezető lehet, elsősorban azért, mert szükségképpen túlságosan is mechanikus és adatközpontú jelenséget sugall. Éppen ezért vannak olyanok, akik a – magyarul talán kissé furcsán hangzó – számítástechnikai fordulat (computational turn) kifejezés mellett érvelnek (*Van Es–Schäfer* [2017]). Ez a felismerés persze nem ma, hanem az 1950-es években történt meg, és a lényege, hogy nemcsak a tudományokban, az élet szinte minden területén egyre több feldolgozható és elemezhető adat keletkezik. Ez a több mint fél évszázaddal ezelőtt elkezdődött folyamat a globális internet, az emberek által használt sokmilliárd digitális eszköz, illetve a robbanásszerű fejlődés előtt álló „dolgok (tárgyak) internetje" (internet of things) korában új szakaszához érkezett, és ma már egyre inkább az emberi viselkedés adatalapú megértése és előrejelzése, illetve a legkülönbözőbb döntéshozatali folyamatok, társadalmi és ipari rendszerek algoritmusok segítségével történő automatizálása a fő cél.

Ettől függetlenül a szövegben a Big Data[1] kifejezést használom, részben a közel 20 éves múltja, illetve a széleskörű elterjedtsége és elfogadottsága miatt. A Big Data csupán néhány éve került a köztudatba, elsősorban az analitikai megoldásokat szállító nagy IT (információtechnológiai) vállalatok marketingtevékenységének köszönhetően (*Gandomi–Haider* [2015]). A Big Data társadalomtudományi szempontú értelmezésének is napról napra növekvő irodalma van (*Borgman* [2015], *Csepeli* [2015], *Dessewffy–Láng* [2015], *McFarland–Lewis–Goldberg* [2015], *Mutzel* [2015], *Székely* [2015]), azonban továbbra sem igazán létezik a fogalomnak egységes meghatározása. Ez persze nem véletlen, hiszen a háttérben meghúzódó technikai fejlődés olyan gyors és „rendezetlen" volt, amely aligha adott lehetőséget az ilyen jellegű filológiai munkálkodásnak. Többen is megpróbálkoztak olyan összetett definíciók megalkotásával, amelyek az üzleti és az akadémiai szféra szempontjait egyaránt figyelembe veszik, de a végeredmény szinte használhatatlanul szerteágazó lett. *De Mauro–Greco–Grimaldi* [2015] például 2014-ben több mint 1500 tanulmány és konferenciaelőadás alapján négy olyan kulcsterületet azonosítottak, amelyekben a

[1] Az angol nyelvű irodalomban a Big Data kifejezés írásmódja nagy kezdőbetűkkel terjedt el, elsősorban azért, mert így lehetett fogalomként megkülönböztetni az egyszerű „nagy adat" szókapcsolattól. Bár magyar nyelvű szövegekben ez a megkülönböztetés nem lenne szükséges, az egységesség jegyében mégis így terjedt el.

különböző Big Data-meghatározások döntő többségében megtalálhatók voltak. Ezek: *1*. az információ jellege; *2*. az összegyűjtéshez és tároláshoz használt technológia; *3*. az elemzéshez alkalmazott módszerek; *4*. továbbá a mindezek segítségével elérhető társadalmi, gazdasági hatások. Ezek alapján a következő (elég körülményes) meghatározást javasolták: „Big Data represents the Information assets characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." (A Big Data olyan információforrásokat foglal magába, amelyre a nagy mennyiség, gyors keletkezés és sokféleség jellemző, illetve sajátos technológia és elemzési módszerek szükségesek ahhoz, hogy belőle értéket nyerhessünk ki.) (*De Mauro–Greco–Grimaldi* [2015] 103. old.). Ezt kiegészítve *Kitchin* [2014] a ma már közhelyszerűen használt „V betűk", illetve mások gondolatai alapján a Big Data hét általános jellegzetességét határozta meg: *1*. nagy méret (volume); *2*. gyors elérhetőség, sebesség (velocity); *3*. nagyfokú változatosság (variety); *4*. teljesség (ezt gyakran az n = all formában jelenik meg); *5*. nagy felbontás (resolution); *6*. relációs jelleg (relational); *7*. rugalmasság és skálázhatóság.

Fontos megjegyeznünk, hogy a Big Data egyáltalán nem csak magáról az adatról szól. Nyilvánvaló, hogy az adatra alapvető szükség van, de ennek még csak nem is kell különösebben „nagynak" lennie. A keletkezés és a felhasználás jellegétől függően nagyon sokféle adattípus tekinthető Big Data-nak: például tera- és petabyte-okban mérhető éghajlati adatok, gigabyte-okban mérhető társadalmi hálózatok adatai, vagy éppen csupán megabyte-okban kifejezhető, már valamilyen formában tisztított és strukturált adat a minket körülvevő világ tetszőleges jelenségeivel kapcsolatban. A fő kérdés inkább az, hogy miként férhetünk hozzá, gyűjthetjük össze, tárolhatjuk, elemezhetjük, értelmezhetjük, vagy oszthatjuk meg másokkal ezeket az adatokat. A Big Data újszerűsége ugyanis éppen ezekben rejlik. Ha hiszünk a fejlődésre vonatkozó előrejelzéseknek, akkor azt is könnyű belátni, hogy annak az átalakulásnak, amely végső soron az adatalapú világról szól, még csak nagyon az elején járunk. A lehetőségek azonban már most is adottak, bár a kutatások ma még inkább csak a „naiv", felfedező, útkereső szakaszukban vannak. A társadalomtudományos kutatások szempontjából ez azt jelenti, hogy egyrészt régi, „klasszikus" kutatási kérdések vizsgálhatók új szempontok és megközelítések szerint, másrészt pedig korábban nem észlelt vagy az adatok hiányában kutathatatlan témák, kérdések válnak egyik pillanatról a másikra vizsgálhatóvá.

2. Az elmélet vége az adatvezérelt tudomány korában?

Az „új empirizmust", amelynek egyik legfontosabb gondolata a címben szereplő kérdés, és amelyet gyakran együtt emlegetnek a Big Data jelenségével, leggyakrab-

ban *Chris Anderson*hoz, a *Wired* magazin korábbi főszerkesztőjéhez, a digitális világgal kapcsolatos számos nagy hatású gondolat szerzőjéhez, pontosabban az ő kiáltványszerű, 2008-as írásához szokás kötni (*Anderson* [2008]). Kiragadott és sokszor idézett fő állításai, hogy „az adatáradat a tudományos módszertant meghaladottá teszi", „ha elegendő adatunk van, akkor a számok magukért beszélnek", „[…] az oksági kapcsolat helyét átveszi a korreláció, és így a tudományos megismerés koherens modellek, egységes elméletek vagy mechanisztikus magyarázatok nélkül is lehetségessé válik. Nem kell többé a régi módszereinkhez ragaszkodnunk." Másképpen fogalmazva, a kutatás és a megismerés olyan világát vetítette előre, ahol nincs szükség előzetes elméletekre, modellekre és hipotézisekre, ami valójában éles szakítás a deduktív kutatás paradigmájával. A Big Data induktív jellegzetessége mellett a másik leggyakrabban hangoztatott érv a teljes sokaság vizsgálatának lehetősége. Ennek lényege, az olyan teljes körű (n = all) és nagy felbontású adat, amelynek segítségével azok a társadalmi csoportok, egyéni, valamint csoportszintű viselkedések, társas interakciók is megfigyelhetők és megérthetők, amelyek a hagyományos survey- és egyéb mintavétel-alapú kutatásokban korábban egyáltalán nem, vagy csak nagyon elnagyoltan és nagyfokú bizonytalansággal voltak vizsgálhatók. Érdemes azonban azt is hozzátenni, hogy az ilyen tökéletes, a teljes populációt hibátlanul tartalmazó adatforrások a valóságban szinte sohasem léteznek. A Big Data sem tartalmazza az esetek többségében a teljes sokaságot, és számtalan olyan torzítás, banális emberi vagy technológiai hiba fordulhat elő, ami miatt az elemzést végzők a „hagyományos" adatok hibáinál megismert kihívásokkal találják szembe magukat. Mindezekhez kapcsolódóan egy további sarkos vélemény, hogy ebben az „elméletlenített" és n = all világban a Big Data-forrásokból származó információkat lényegében bárki, aki a statisztikához és az adatvizualizációk olvasásához ért, valódi tudássá tudja alakítani. Ezek az állítások kétségtelenül helytállók lehetnek az üzleti analitika néhány területén, ahol kevésbé korlátozott a hozzáférés az adatokhoz, továbbá a fókusz az autonóm vagy félautonóm algoritmusalapú döntéseken van. Azonban a társadalomtudományos kutatás esetében, ahol a Big Data-t nyersanyagként kívánjuk felhasználni, néhány kritikai megjegyzést mindenképpen érdemes tenni.

Az (akadémiai) kutatások szempontjából a Big Data egyik legfontosabb jellemzője, hogy az elemzéshez használt adatok többnyire valamilyen más, általában üzleti célú tevékenység fő- vagy melléktermékei, vagy pedig kormányzati szervezetek tulajdonában vannak. Egy Big Data-alapú tudományos kutatás esetében ez szinte minden esetben azt is jelenti, hogy az adat még azelőtt keletkezett, hogy bármilyen kutatási kérdés vagy hipotézis megfogalmazódott volna. (Erre jó példa a Twitter, Facebook, Google Trends és más online szolgáltatások API-n [application programming interface – alkalmazásprogramozási felület] vagy egyéb technikával összegyűjtött adatai, továbbá valamilyen más egyedi megállapodás egy-egy üzleti vagy kormányzati adatgazdával.) Másképpen fogalmazva, erre a kutatási folyamatra

úgy is tekinthetünk, mint egy olyan új ismeretelméleti megközelítésre, amely különbözik a hagyományos deduktív logikától, ahol a hipotézisek és a megfigyelések nem valamilyen előzetes elméleti meggyőződésből, hanem a létező és a kutatók által fizikailag hozzáférhető adatok által kínált lehetőségekből (és persze korlátokból) fakadnak. Fontos továbbá, hogy bármilyen adatról is van szó, az sohasem tudományos vagy kognitív vákuumban jön létre, hanem valamilyen emberi tevékenység közvetett vagy közvetlen fő-, illetve melléktermékeként. Ezt pedig szubjektív, érdek- és értékvezérelt szempontok szükségképpen éppúgy meghatározzák, mint maga az alkalmazott technológia. Az tehát, hogy valamilyen társadalmi jelenséggel kapcsolatban rendelkezésre állnak-e kutatáshoz használható adatok, számos tényezőtől függ. Kritikai nézőpontból szemlélve pedig sok esetben talán még érdekesebb, hogy mely területeken és miért nem érhetők el adatok.

A társadalomtudományok nézőpontjából alighanem a Big Data adatvezérelt, induktív, az elméleti kiindulópontokat „lefokozó" megközelítése a leginkább vitatható állítás. A Big Data a társadalmak működésével kapcsolatos új felfedezésekre ad(hat) lehetőséget, ami kellő innovativitást kíván mind a kérdésfeltevés, mind pedig a kutatási eszközök megválasztása során, valamint kétségtelen, hogy új hipotézisek és megfigyelések közvetlenül az adatokból is megszülethetnek. A megfelelő kontextus és az adott területre vonatkozó tudás nélkül az interpretáció ugyanakkor kudarcra van ítélve. Jól mutatják ezt azok az elemzések, amelyek hajlamosak figyelmen kívül hagyni a társadalomtudományok sok évtizedes elméleti és módszertani hagyományait, illetve eredményeit (*Borgman* [2015]), vagy csak egyszerűen olyan összefüggéseket bizonyítanak be Big Data-alapokon, amelyek tulajdonképpen már régóta ismertek, kis túlzással a „101-es bevezető szociológiai kurzusok" tananyagát képezik. A Big Data ugyanakkor komoly lehetőséget is kínál azon „elméletvezérelt" társadalomtudósok számára, akik képesek és hajlandók átlépni a saját tudományterületeik hagyományos határvonalait. Ezek a próbálkozások akkor lehetnek a legeredményesebbek, ha az elméleti felkészültség kiegészül azzal a gyakorlati tudással is, hogy miként lehet ezekből az új típusú adatokból valódi információt és tudást kinyerni, továbbá hogyan lehet ezt a megfelelő célközönségek számára vizualizálni, az összefüggéseket pedig tágabb társadalmi kontextusba helyezve érthetően bemutatni.

3. A különböző tudományterületek közötti szerepek átértelmezése

A Big Data és a hálózatelemzés fejlődésének következtében úgy tűnik, hogy a társadalomtudományok elvesztették korábbi privilégiumukat a társadalmak működé-

sének vizsgálatában. Egyre inkább eltűnőben vannak az egyes diszciplínák között történetileg kialakult határvonalak, amelyek többé-kevésbé kijelölték az alapvető ismeretelméleti és módszertani eszközöket, továbbá magát a kutatás tárgyát, illetve kérdéseit. A tudomány történetében talán most állt elő első alkalommal az a helyzet, hogy az internetes ipar vállalatai, valamint a mérnöki, természet- és társadalomtudományok lényegében azonos adatok felhasználásával, nagyon hasonló kutatási kérdésekre keresik a választ (*McFarland–Lewis–Goldberg* [2015]). Ez a konvergenciafolyamat hatalmas lehetőséget rejt mindazok számára, akik aktív részesei ennek az izgalmas átalakulásnak. Ugyanakkor az is nyilvánvaló, hogy az egyes tudományok közötti korábbi status quo is megváltozik. Ma még nem teljesen egyértelmű, hogy a társadalomtudományok, illetve a természet- és számítógép-tudomány közötti munkamegosztás mennyire alapul majd szimmetrikus vagy aszimmetrikus viszonyokon. Egyáltalán nem elképzelhetetlen, hogy a társadalomtudományok – és ezen belül a szociológia különösen – olyan kolonizációs folyamat elszenvedőjévé válnak, amelynek eredménye az alárendelt(ebb) szerep lesz. A „külvilág" (azaz döntően a természet- és műszaki tudományok) számára a társadalomtudományok gyakran tűnnek „könnyű prédának", elsősorban a nagyfokú fragmentációjuk miatt, ami a számos egymással versengő elméletben és az alkalmazott kutatási módszerek különbözőségében rejlik (*Balietti–Mas–Helbing* [2015], *Whitehouse et al.* [2012]). Az egyes tudományterületek közötti finom rivalizálás az elmúlt években egyre inkább a CSS (computational social science – számítógépes társadalomtudomány) területére helyeződött át (*Lazer et al.* [2009]). A definíció szerint a CSS jóval több, mint csupán Big Data, mivel ez magába foglalja többek között a hálózatelemzést és a társadalmi szimulációs modelleket, továbbá más egyéb módszereket. Az egyes tudományterületek közötti erőviszonyok megváltozását, a „ki tanul kitől" kérdését jól mutatja a *Conte* és szerzőtársai által írt, 2012-ben megjelent „Manifesto of computational social science" (A számítógépes társadalomtudomány kinyilatkoztatása) című tanulmány egy részlete (*Conte et al.* [2012] 341. old.):

> […] a szociológia különösen, a társadalomtudományok pedig általánosságban drámai paradigmaváltáson mehetnek keresztül azáltal, hogy a fizikai tudományok módszereit beemelik az eszköztárukba. A számítógépes megközelítést a kísérletek érzékeny használatával kombinálva a társadalomtudományok közelebb kerülhetnek ahhoz, hogy biztos kapcsolatot hozzanak létre az elmélet, illetve az empirikus tények és kutatás között. Ezek az összefüggések minden olyan tudomány számára kiindulópontot jelenthetnek, amelyek az emberi viselkedést kutatják; továbbá feloldhatók lesznek az olyan jellegű összeférhetetlenségek, mint a közgazdaságtanban elfogadott racionális cselekvő és a szociológia, illetve szociálpszichológia nézőpontja, amely nyíltan el-

> utasítja ezt. Másodsorban, az utóbbiak jóval inkább támaszkodnak a kísérletekből (és survey-ekből) származó tényekre, mint a hagyományos közgazdaságtan, amely tisztán absztrakt analitikus megközelítéseket alkalmaz. A számítógépes társadalomtudomány ennek a paradigmaváltásnak az egyik fő tényezője lesz. (Szerző saját fordítása.)

Részben az említettek miatt is, a társadalomtudományok területén egyre erősebb a külső nyomás és a belső késztetés arra, hogy saját tevékenységüket kifinomult kvantitatív elemzésekkel, egyre nagyobb mennyiségű adat felhasználásával és minél inkább (számító)gépesítve végezzék el. Emiatt mindjobban nő a társadalomtudományokkal foglalkozók bizonyos csoportjainak motivációja (ami részben külső nyomás, részben pedig a kimaradáshoz kapcsolódó frusztrációk [„fear of missing out"] következménye) arra, hogy újfajta adatelemzési és adatvizualizációs technikákat sajátítsanak el, és valamilyen formában megpróbáljanak a Big Data vagy a CSS tágan értelmezett kutatási területeire belépni.

4. Egyenlőtlen hozzáférés az adatokhoz: „adatgazdagok" és „adatszegények"

A különböző társadalmi jelenségekkel kapcsolatos adatok összegyűjtésére és elemzésére soha nem látott lehetőségek állnak már most is rendelkezésre. Például a közösségi média adataiból pontos képet kaphatunk az emberek egymás közötti interakcióiról (*Felt* [2016]); a Twitteren használt nyelvezet elemzésével a lelkiállapot és ezen keresztül bizonyos egészségi kockázatok előrejelezhetők (*Eichstaedt et al.* [2015]), vagy éppen az online közösségi hálózatok struktúráját fizikai térben vizsgálva egy adott ország adminisztratív közigazgatási szerkezetét figyelhetjük meg (*Lengyel et al.* [2015]). Nem túl meglepő módon, ebben az adatvezérelt társadalomtudományban a siker kulcsa a mennyiségi és minőségi értelemben is hozzáférhető jó adat. Az az állítás tehát, amely úgy szól, hogy a ránk váró adatbőség korában a tudományos felfedezés korlátai eltűnni látszanak, nyilvánvalóan pontosításra szorul. Nem hagyható ugyanis figyelmen kívül, hogy újfajta választóvonalak jönnek létre „adatgazdagok" és „adatszegények" között (*boyd–Crawford* [2012]). Itt nem csupán arról az aszimmetrikus viszonyról van szó, amely az adatok tulajdonosai (akik gyűjtik, tárolják, „kibányásszák" és elemzik azt) és az adatgyűjtés tárgyai (azaz a felhasználók) között létezik (*Andrejevic* [2014]), hanem azokról a régi-új választóvonalakról, amelyek kutatók és kutatócsoportok, a vállalatok és az akadémiai világ vagy éppen országok között jönnek létre, attól függően, hogy fizikailag mennyire közel,

illetve távol vannak az adatoktól, pontosabban annak tulajdonosaitól. Csupán egy példa: a Facebook saját kutatócsoportjának tagjai[2] (amellett, hogy az egyetemi/akadémiai világnak is részei) vélhetően az egyik leginkább kiváltságos helyzetben levő kutatók a világon, hiszen közvetlen hozzáférésük van az adatokhoz, a feldolgozáshoz, és az elemzéshez pedig óriási technikai háttér áll rendelkezésükre. Az ezen a privilegizált társadalmi laboratóriumon kívül levők számára a saját adatgyűjtés API-kon, illetve webes crawlereken (keresőrobotokon) keresztül jóval korlátozottabb. Szinte lehetetlen vállalkozás kétoldalú együttműködések kialakítása a legnagyobb adatgazda-vállalatokkal (például Facebook, Google) a központtól térben és a hálózattudományban legrövidebb útként definiált mérőszám szerint is távol levő kutatók és szervezetek számára. Amire lehetőség van az olyan országokban mint Magyarország, az az egyedi megállapodások megkötése például a helyi telekommunikációs szolgáltatókkal, online médiavállalatokkal és állami/kormányzati szervezetekkel. Ehhez azonban a nagyvállalati és a bürokratikus szervezeti kultúrákban, nyelvhasználatban való jártasság éppúgy szükséges, mint a megfelelő tárgyalási készségek, illetve annak felismerése, hogy mik ezeknek a szervezeteknek az igényei, érdekei és működésük külső korlátjai (például azok a keretek, amelyeket az anyavállalat nemzetközileg kijelöl). Ezen a területen a társadalomtudományokat művelőknek biztos, hogy sokat kell még tanulniuk, fejlődniük, és képesnek kell lenniük arra, hogy meggyőzzék az adatok tulajdonosait arról, hogy a tervezett kutatásnak van értelme, az adatokat megfelelő módon használják fel, nem sérülnek üzleti érdekek, az együttműködés tehát mindkét fél számára előnyös lehet. Az adatokhoz való hozzáférés nehézsége egyébként nem csak a félperiféria és a periféria kutatói számára jelent korlátokat. A SAGE kiadó egy (nem reprezentatív) nemzetközi vizsgálatban a Big Data-alapú kutatások helyzetét és a jövőbeli lehetőségeket vizsgálta. Ennek során több ezer kutatót kérdeztek meg szerte a világon, és az eredmények azt mutatták, hogy az egyik legnagyobb problémának szinte mindenhol az adatokhoz való hozzájutás bizonyult, kiegészülve azzal, hogy a társadalomtudományt művelőkön egyre nagyobb a nyomás abba az irányba is, hogy újfajta készségekre és módszertani tudásra tegyenek szert (*Metzler et. al.* [2016]).

5. A pozitivista Big Data-kutatások kritikája

A Big Data-alapú társadalomtudományi kutatások (pontosabban az ezekkel kapcsolatos várakozások) alapvetően a pozitivista tudományelmélet elveiből indulnak ki, ahol maga az adat egy többnyire neutrális, értéksemleges „nyersanyag”. Ez sok eset-

[2] https://research.fb.com/people/

ben így is van, azonban a társadalomtudományok a korábban már említett leértékelődésüket a kritikai adattanulmányok egyre fontosabbá váló területén tudják ellensúlyozni. Ennek érdekében úgy árnyalják a túlságosan is funkcionalista és eredményközpontú megközelítést, hogy a (nagy) adatot a hatalmi viszonyok részének tekintik (*Iliadis–Russo* [2016]). Azok a hatalmas mennyiségű felhasználói adatok, amelyeket az emberekről gyűjtenek és tárolnak, a tőke egy speciális formájaként is meghatározhatók. Ez az erőforrás – megtámogatva komplex algoritmusokkal és hatékony adatfeldolgozási képességekkel – alkalmas arra, hogy segítségével szervezetek befolyásolják az emberek érzelmeit, attitűdjeit és viselkedését. Az ezzel kapcsolatos vélt és valós aggodalmak 2016-ban leginkább a Cambridge Analytica[3] nevű vállalatnak ahhoz a tevékenységéhez köthetők, amelyet *Ted Cruz* és *Donald Trump* elnökjelölti kampányában az Egyesült Államokban, illetve a brexitet támogató kampány során az Egyesült Királyságban végzett.[4] Azokkal a talán túlságosan is sommás véleményekkel szemben, mely szerint a Big Data-ban rejlő lehetőségek ügyes kihasználása önmagában elegendő volt a választások, illetve a brexit-szavazás megnyerésére, ezek az adatok és elemzési technikák inkább csak azt tették lehetővé, hogy a korábbi módszerekhez képest jóval pontosabban modellezzék a választók személyiségét, és ezáltal még hatékonyabban, új és innovatív formában lehessen célba juttatni a személyre szabott üzeneteket. Miközben az elmúlt időszakban a közvetlen hatásokkal kapcsolatban még inkább szkeptikusak lehettünk, nyilvánvaló, hogy ezekben a módszerekben komoly lehetőségek rejlenek, ami már a nem túl távoli jövőben is komoly változásokat hozhat a politikai és üzleti kommunikáció területén. Az pedig nagyon fontos, hogy a társadalomtudományok (politológia, szociológia, szociálpszichológia, jog stb.) ezeket a változásokat kritikai nézőpontból vizsgálják. Ennek része a felelősség kérdése, amelyre itt nem csak tisztán jogi kategóriaként érdemes tekinteni (*Mann* [2017], *McDermott* [2017], *Tene–Polonetsky* [2013]. A felelős adatgazdálkodás, az etikus, személyiségi jogokat tiszteletben tartó elemzés, az adatgyűjtés és feldolgozás szükséges mértékű transzparenciája, továbbá a létrejövő új hatalmi, politikai struktúrák mind-mind alapvető vizsgálati kérdései (lehetnek) ennek a kritikai irányzatnak.

Ezen a tág kutatási területen belül van még egy kiemelt téma, amely vélhetően megkerülhetetlen lesz a jövőben, ahol a társadalomtudósok a kutatási hagyományaikból fakadóan komoly előnyben vannak. Az ebben a témában végzett kutatások és ezek gyakorlati eredményei közvetlenül befolyásolhatják a jövőbeli szakpolitikákat, a Big Data társadalmának jogi és etikai kereteit. A szóban forgó kutatási terület a Big Data- és algoritmusalapú döntéshozatal társadalmi igazságosságának/igazságtalanságának, az ezekben megjelenő diszkriminációs folyamatoknak empirikus vizsgálata.[5]

[3] https://cambridgeanalytica.org/

[4] Did Cambridge Analytica influence the Brexit vote and the US election? *Guardian*. 4 March 2017. https://www.theguardian.com/politics/2017/mar/04/nigel-oakes-cambridge-analytica-what-role-brexit-trump

[5] Lásd erről bővebben *Ságvári* [2017].

A diszkrimináció és az ennek hátterében meghúzódó előítéletek a társadalmak alapvető működési mechanizmusai közé tartoznak, és ezeket a társadalomtudományok bejáratott elméletek és módszerek segítségével évtizedek óta vizsgálják (*Sik–Simonovits* [2011]). Nagyon is emberi és szubjektív, érzelmi vagy éppen irracionális viselkedésekről van szó, amelyeket alapvetően a rendelkezésre álló ismeretek korlátozottsága vagy ennek figyelmen kívül hagyása működtet. Ezzel szemben az adatok és az algoritmusok a technológia oldaláról szemlélve, meghatározásuk szerint ennek szöges ellentétjei: gépies kiszámíthatóság, racionalitás és objektivitás. Ez azonban csak akkor igaz, ha figyelmen kívül hagyjuk a természetesen itt is meglevő emberi tényezőt. Az adatokat emberek gyűjtik, tárolják és dolgozzák fel, az algoritmusok lépéseit emberek tervezik és programozzák, még akkor is, ha egyre inkább teret kapnak a gépi tanulás és a mély tanulás elvein alapuló algoritmusok is. Az elnevezések ne tévesszenek meg senkit, hiszen a célok és a rendelkezésre álló adatnyersanyag tekintetében az emberi tényező ezeknél az eljárásoknál is döntő jelentőségű.

Az Obama-kormányzat Big Data-val foglalkozó munkacsoportja (Big Data Working Group) már 2014-ben és 2015-ben is közzétett egy-egy jelentést, amelyek az adatok kormányzati és üzleti célú felhasználásának lehetőségeit és társadalmi kockázatait vették számításba (*The White House* [2014], [2015]). A 2014-ben megjelent összegzés egyik fontos üzenete volt, hogy felhívta a figyelmet a társadalmi diszkrimináció kódolt megjelenésére az automatizált döntésekben, illetve ezeknek a rendszereknek a nem transzparens és a kívülállók számára csak nagyon korlátozottan kontrollálható működésére.

Dióhéjban, a diszkrimináció az algoritmusalapú rendszerekben kétféle módon (illetve ezek együttes kombinációjaként) jelenhet meg. Az első esetben már maguk az algoritmus számára bemenő információt (input) szolgáltató adatok is megkérdőjelezhető minőségűek. A társadalmi igazságosság, az egyenlő bánásmód szempontjából pedig az eredmény az alkalmazott algoritmus tökéletességétől és részrehajlásától függetlenül problémás, hiszen már a kezdetektől fogva jelen van a „rossz" adat a rendszerben.

Második esetben nem az adattal, hanem az azt feldolgozó algoritmus működésével van a probléma. Ennek lényege, hogy a matematikai-statisztikai alapon működő döntések, az eredeti szándéktól akár teljesen függetlenül is, a bennük megjelenő „kulturális kód" következményeként a meglevő társadalmi igazságtalanságokat erősítik fel, vagy hoznak létre újfajta diszkriminatív helyzeteket.

Tehát a lehetséges társadalmi következmények nagyságrendjét tekintve egy viszonylag új jelenségről van szó. A probléma és a mögöttes mechanizmusok ismertek, egyre több az ezzel kapcsolatos hír, és mind többet tudunk arról is, milyen szándékos vagy nem szándékos lépéseken keresztül jöhetnek létre igazságtalan és diszkriminatív döntések, illetve az ezekre épülő komplex rendszerek. Ennek a tudásnak a rendszerezése, az alkalmazható kutatási módszertanok megtalálása, és a kérdés „társa-

dalmasítása" olyan feladat, amelyet hatékonyan az IT-szakemberek, adatelemzők, jogászok, szociológusok és más társadalomtudósok közösen alkotott kutatócsoportjai tudnak elvégezni.

6. Merre tovább társadalomtudomány?

Az előzőkben bemutatott főbb trendek alapján adódik a kérdés, hogy vajon milyen szerepet tölthetnek be a társadalomtudományokkal foglalkozók ebben az új környezetben? Természetesen univerzális receptek nem léteznek, így minden egyéni stratégia más és más. Megkockáztathatjuk, hogy a jelenleg zajló folyamatok ahhoz hasonló horderejű változásokat fognak okozni, mint ami a XX. század második felében a szociológián belül következett be a többváltozós statisztikai modellek és a survey-kutatások előtérbe kerülésével, ami az akadémiai kutatások piacát éppúgy átalakította, mint az egyetemi képzések tartalmát, illetve a „versenyképes" szociológustól elvárt készségeket. Így tehát egyre nagyobb igény lesz olyan alapvetően társadalomtudományos indíttatású szakemberekre, akik az elméleti felkészültségük és a nyilvánvalóan elvárható társadalmi érzékenységük mellett az új típusú adatok által megkívánt módszertani ismeretekkel és gyakorlati (programozói) tudással is rendelkeznek. A számítógépes etnográfia, számítógépes nyelvészet, hálózattudomány, gépi tanulás, Big Data-alapú kísérletek mind olyan új kutatási irányok, ahol szükség van erre a hibrid tudásra, és ahol hatalmas lehetőségek vannak az interdiszciplináris együttműködésekre. Ezen az új tudományos „piactéren" a szerepek vélhetően nem egyenlően lesznek elosztva, és a társadalomtudományok nem feltétlenül indulnak a legjobb pozícióból. A „meccs" azonban még korántsem lefutott. Az új típusú adatok megszerzésében és elemzési célú felhasználásában való jártasság, a pozitivista szemlélet némi szkepticizmussal fűszerezve és a társadalomtudományos nézőpontból fakadó kritikai attitűd lehetnek azok a készségek és kutatói szerepfelfogások, amelyekkel a Big Data korában és világában a társadalomtudósok sikeresek lehetnek.

Irodalom

ANDERSON, C. [2008]: The end of theory: The data deluge makes the scientific method obsolete. *WIRED magazine*. 23 Juny. https://www.wired.com/2008/06/pb-theory/

ANDREJEVIC, M. [2014]: The Big Data Divide. *International Journal of Communication*. Vol. 8. No. 8. pp. 1673–1689.

BALIETTI, S. – MAS, M. – HELBING, D. [2015]: On disciplinary fragmentation and scientific progress. *PLoS One*. Vol. 10. No. 3. pp. 1–26. http://dx.doi.org/10.1371/journal.pone.0118747

BORGMAN, C. L. [2015]: *Big Data, Little Data, No Data. Scholarship in the Networked World*. The MIT Press. Cambridge.

BOYD, D. – CRAWFORD, K. [2012]: Critical questions for Big Data. *Information, Communication & Society*. Vol. 15. No. 5. pp. 662–679. http://dx.doi.org/10.1080/1369118x.2012.678878

CONTE, R. – GILBERT, N. – BONELLI, G. – CIOFFI-REVILLA, C. – DEFFUANT, G. – KERTESZ, J. – LORETO, V. – MOAT, S. – NADAL, J.-P. – SANCHEZ, A. – NOWAK, A. – FLACHE, A. – SAN MIGUEL, M. – HELBING, D. [2012]: Manifesto of computational social science. *The European Physical Journal Special Topics*. Vol. 214. Issue 1. pp. 325–346. http://dx.doi.org/10.1140/epjst/e2012-01697-8

CSEPELI G. [2015]: A szociológia és a Big Data. *Replika*. 92–93. sz. 171–176. old.

DE MAURO, A. – GRECO, M. – GRIMALDI, M. [2015]: What is Big Data? A consensual definition and a review of key research topics. *AIP Conference Proceedings.* Vol. 1644. Issue 1. pp. 97–104. http://dx.doi.org/10.1063/1.4907823

DESSEWFFY, T. – LÁNG, L. [2015]: Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*. 92–93. sz. 227–230. old.

EICHSTAEDT, J. C. – SCHWARTZ, H. A. – KERN, M. L. – PARK, G. – LABARTHE, D. R. – MERCHANT, R. M. – JHA, S. – AGRAWAL, M. – DZIURZYNSKI, L. A. – SAP, M. – WEEG, C. – LARSON, E. E. – UNGAR, L. H. – SELIGMAN, M. E. P. [2015]: Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science.* Vol. 26. Issue 2. pp. 159–169. http://dx.doi.org/10.1177/0956797614557867

FELT, M. [2016]: Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*. Vol. 3. No. 1. pp. 1–15. http://dx.doi.org/10.1177/2053951716645828

GANDOMI, A. – HAIDER, M. [2015]: Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management*. Vol. 35. No. 2. pp. 137–144. http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007

ILIADIS, A. – RUSSO, F. [2016]: Critical data studies: An introduction. *Big Data & Society*. Vol. 3. No. 2. pp. 1–7. http://dx.doi.org/10.1177/2053951716674238

KITCHIN, R. [2014]: Big Data, new epistemologies and paradigm shifts. *Big Data & Society*. Vol. 1. No. 1. pp. 1–12. http://dx.doi.org/10.1177/2053951714528481

LAZER, D. – PENTLAND, A. – ADAMIC, L. – ARAL, S. – BARABÁSI, A.-L. – BREWER, D. – CHRISTAKIS, N. – CONTRACTOR, N. – FOWLER, J. – GUTMANN, M. – JEBARA, T. – KING, G. – MACY, M. – ROY, D. – VAN ALSTYNE, M. [2009]: Life in the network: The coming age of computational social science. *Science*. Vol. 323. Issue 5915. pp. 721–723. http://dx.doi.org/10.1126/science.1167742

LENGYEL, B. – VARGA, A. – SÁGVÁRI, B. – JAKOBI, A. – KERTÉSZ, J. [2015]: Geographies of an Online Social Network. *PLoS One*. Vol. 10. No. 9. pp. 1–13. http://dx.doi.org/10.1371/journal.pone.0137248

MANN, S. [2017]: Big Data is a big lie without little data: Humanistic intelligence as a human right. *Big Data & Society*. Vol. 4. No. 1. pp. 1–10. http://dx.doi.org/10.1177/2053951717691550

MAYER-SCHÖNBERGER, V. – CUKIER, K. [2013]: *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Houghton Mifflin Harcourt. Boston, New York.

MCDERMOTT, Y. [2017]: Conceptualising the right to data protection in an era of Big Data. *Big Data & Society*. Vol. 4. No. 1. pp. 1–7. http://dx.doi.org/10.1177/2053951716686994

McFarland, D. A. – Lewis, K. – Goldberg, A. [2015]: Sociology in the era of Big Data: The ascent of forensic social science. *The American Sociologist*. Vol. 47. Issue 1. pp. 12–35. http://dx.doi.org/10.1007/s12108-015-9291-8

Metzler, K. – Kim, D. A. – Allum, N. – Denman, A. [2016]: *Who Is Doing Computational Social Science? Trends in Big Data Research.* Technical Report. SAGE. London.

Meyer, E. T. – Schroeder, R. [2015]: *Knowledge Machines: Digital Transformations of the Sciences and Humanities*. MIT Press. Cambridge.

Mützel, S. [2015]: Facing Big Data: Making sociology relevant. *Big Data & Society*. Vol. 2. No. 2. pp. 1–4. http://dx.doi.org/10.1177/2053951715599179

Ságvári B. [2017]: Diszkrimináció, átláthatóság és ellenőrizhetőség. Bevezetés az algoritmusetikába. *Replika*. Megjelenés alatt.

Sik E. – Simonovits B. [2011]: Adalékok a diszkriminációtesztelés kutatási problémáinak megismeréséhez. In:. *Sik E. – Simonovits B.* (szerk.): *A diszkrimináció mérése.* ELTE TáTK. Budapest. 177–207. old.

Székely I. [2015]: Az adatmentes zónák szükségessége és esélye. *Replika*. 92–93. sz. 209–226. old.

Tene, O. – Polonetsky, J. [2013]: Big Data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*. Vol. 11. Issue 5. pp. 240–273.

The White House [2014]: *Big Data: Seizing Opportunities, Preserving Values*. Washington, D.C.

The White House [2015]: *Big Data: Seizing Opportunities, Preserving Values. Interim Progress Report*. Washington, D.C.

Van Es, K. – Schäfer, M. T. [2017]: Introduction. New brave world. In: *Schäfer, M. T. – Van Es, K.* (eds.): *The Datafied Society.* Amsterdam University Press. Amsterdam. pp. 13–23.

Whitehouse, H. – Kahn, K. – Hochberg, M. E. – Bryson, J. J. [2012]: The role for simulations in theory construction for the social sciences: Case studies concerning divergent modes of religiosity. *Religion, Brain & Behavior*. Vol. 2. No. 3. pp. 182–201. http://dx.doi.org /10.1080/2153599x.2012.691033

Summary

Today large amounts of data are available to research on human behaviour, and the industry that relies on collecting, combining, selling and analysing digital footprints is developing with lightning speed. Such data can increasingly be used to address those "classic" societal issues that have been in the focus of social science for more than a century. Also, the advances in the use of such data in social sciences offer the possibility to answer questions that were beyond research in the past due to the lack of available data. It is likely that this transformation might not follow the established rules of epistemology and the traditional division between fields of sciences. The author presents some important challenges and opportunities that social sciences (particularly sociology) are confronted with.