Galántai Zoltán

Big data, tudomány, kauzalitás

Bevezetés

A Gartner 2015-ben törölte a túlértékelt technológiák listájáról a big datát (Woodie 2015), és ez minden bizonnyal azt jelzi, hogy a big data immár nem csupán ismert, de egyre szélesebb körben elterjedt és használt is. Azt azonban nem tudjuk, milyen hatásai lesznek a jövőben, mondjuk egy évtizedes távlatban. Pedig ez már csak azért is fontos, mert a big data egyes elképzelések szerint nem csupán az adatfeldolgozást, illetve a mindennapi élet különböző területeit változtathatja meg alapvetően – a vásárlói szokások és a potenciális bűnelkövetők viselkedésének feltérképezésétől a járványok előrejelzéséig –, de a tudomány fogalmát is jelentősebben megváltoztathatja, mint bármi a 17. század elejének tudományos forradalma óta.

Az alábbiakban az óvatos kételkedő pozíciójából egyfelől azt vizsgáljuk meg, hogy miként lehet újraértelmezni a tudománytörténetet a big data, illetve általában véve az adatok gyűjtése, tárolása és feldolgozása szemszögéből, másfelől, hogy ennek az átértelmezésnek milyen hosszabb távú következményei lehetnek a tudomány értelmezésére nézve. Tehát elengedhetetlen, hogy érintőlegesen a tudomány fogalmával, illetve annak egyik központi elemével, az ok – okozatisággal is foglalkozzunk, ugyanis egyes álláspontok szerint a big data szempontú értelmezés ezt legalább részben fölülírhatja. Bár nem tudjuk, hogy ez utóbbi bekövetkezik-e, felvázoljuk a jelenleg elképzelhetőnek tűnő értelmezéseket.

Big data – nagyon röviden

„A big data nagy mennyiségű, nagy sebességű és/vagy nagy változatosságú információ, amely költséghatékony és innovatív megoldásokat kínál az információfeldolgozásban, és amely lehetővé teszi a megnövelt hatékonyságú értelmezéseket, döntéshozatalt és folyamat-automatizálást”, olvasható a Gartner fogalommagyarázatában (Gartner é. n.). Eközben az sem közömbös, hogy „a gyorsan keletkező és szaporodó adattömeg… hasznosítására kevés idő áll rendelkezésre” (Bőgel 2015: 33). Mások pedig azt húzzák alá, hogy „a big data kifejezés olyan dolgokra utal, amelyeket csak nagy léptékben tekinthetünk meg… hogy segítségükkel… új felismerésekre jussunk”, és ezek a piacokat, szervezeteket, kormányokat stb. át fogják rendezni. Meg „sok minden mást” is (Mayer-Schönberberger és Cukier 2014: 14-15).

Itt – némiképp leegyszerűsítve – két dologról van szó. Egyrészt természetesen az adatmennyiségről: 2007-ben a Sloan Digital Sky Survey keretében néhány hét alatt több adatot gyűjtöttek össze, mint addig a csillagászat a kezdetek óta (Mayer-Schönberberger és Cukier 2014: 15). A Nagy Hadronütköztető pedig jelenleg évente 30 petabyte adatot gyűjt (CERN é. n.). Másrészt: a sok adat nem egyszerűen több, hanem más is. Egy hason-

lattal élve: ugyanúgy, mint ahogy a gravitáció hatása a kis rovarok számára – a víz felületi feszültségével szemben – elhanyagolható (és ennek megfelelően nem is érzékelhető), megfelelően nagy adatmennyiségek esetén is olyan jelenségeket figyelhetünk/tapasztalhatunk meg, melyeket kisebb méreteknél nem (Mayer-Schönberberger és Cukier 2014: 15).

A fizikában jól ismert probléma „a skála zsarnoksága": a tudományfilozófus Robert Batterman úgy fogalmaz, hogy csak egy redukcionista gondolhatja, hogy a jelenségeket kicsiben megfigyelve képesek leszünk a nagy rendszerek működésére, illetve az ezeket meghatározó törvényekre következtetni (Batterman 2013: 2). Azaz a big data alapú megközelítés tudományfilozófiai értelemben annak az episztemikus redukciónak az elutasítását is jelenti, mely szerint a különböző szinteken szükségképpen ugyanúgy működnek a dolgok, és ennek megfelelően azt is elutasítja, hogy létezik egyetlen, minden nagyságrendre érvényes, skálafüggetlen leírás.

A modern fizikában az 1960-as évek közepe óta viszont elfogadottnak számít valamiféle Nagy Egyesített Elmélet keresése, amely egyetlen rendszeren belül írná le a fizikai valóságot, és így egyesítené, példának okáért, az általános relativitáselméletet, valamint a kvantumfizikát. Általában véve pedig a természettudományok a lokális megfigyelésekből próbálnak következtetni a globálisra, vagyis az általános természettörvényekre (Smeenk 2013: 1).

Kísérletezés? Okozatiság?

Tehát nem lenne meglepő, ha a big data, amely per definitionem a kis és a nagy adatmennyiségek közötti különbségek eltérő voltából indul ki, legalábbis bizonyos pontokon másmilyen leírást tenne lehetővé a valósággal kapcsolatban, mint a „hagyományos" fizika. Ez ugyanis Galilei óta a modern természet-, sőt olykor egyes társadalomtudományok mintájául is szolgált abból a feltételezésből kiindulva, hogy ha bevált bizonyos, nagyon speciális területeken, például az égi mechanikában, akkor miért ne válna be mindenütt. Ezt a természettudományt pedig a matematika alapozta meg (Orrell 2007: 95) egyfajta szigorú és axiomatikus, „ha – akkor", ok-okozati logikával. A big data viszont a skálafüggetlenség elutasítása mellett éppen abban különbözik az eddigi megközelítéstől, hogy másként viszonyul a kauzalitáshoz.

Ezen a ponton két dologra érdemes legalább röviden kitérni. Egyfelől arra, hogy valójában nem „magától értetődő", hogy egyáltalán van értelme mai értelemben vett tudományos kísérletet végezni, és – ceteris paribus – egyetlen tényező megváltoztatása mellett azt vizsgálni, hogy miként módosul az eredmény. Arisztotelész még abból indult ki, hogy ha a kísérlet mesterséges feltételei révén beavatkoznánk a „természet rendjébe", akkor ennek két eredménye lehetne. Vagy az, hogy megváltoztatjuk a kimenetelt – és ezáltal meg is hamisítjuk a végeredményt. Vagy pedig az, hogy az eredmény ugyanaz lesz, mintha megfigyelést végeznénk – ebben az esetben pedig felesleges (Grant 2007: 49-50). Az a gondolat, hogy a megfigyelés (kísérletezés) befolyásolhatja az eredményt, bizonyos értelemben mintha a kvantummechanikáról folytatott, korai vitákban köszönne vissza. Számunkra azonban most érdekesebb, hogy az újkori „tudományos forradalom" (ahol azért indokolt az idézőjel, mert ekkoriban még nem létezett mai értelemben vett tudomány) nagyjából az 1600-as évektől az arisztotelészi, a kísérletet elutasító szemlélethez képest

egyfajta analitikus megközelítést képvisel, és azzal a feltételezéssel él, hogy az egyes elemek izoláltan is vizsgálhatóak. Az új tudományfelfogás másik, az arisztotelészi felfogástól eltérő eleme a matematikai leírás fontosságának (és érvényességének) hangsúlyozása (Henry 2002: 14). Ami viszont átvezet a kauzalitás problémájához, ugyanis ha van egy – mondjuk a gravitáció hatását leíró – egyenletünk, akkor két dolgot tehetünk, amennyiben úgy véljük, hogy a tudomány elsődleges célja a válaszadás: annak a megmagyarázása, hogy „mi történik a világban körülöttünk" (Okasha 2002: 38).Vagy feltételezzük, hogy a képlet (egyenlet, összefüggés) oksági kapcsolatot fejez ki, és a két jelenség között oksági kapcsolat van abban az értelemben, hogy az egyik szükségképpen elvezet a másikhoz. Közbevetőleg: a „kísérleti módszer" annak ellenére, hogy látszólag közel áll a modern felfogáshoz, részben az úgynevezett természeti mágián alapult, amely azzal a feltételezéssel élt, hogy létezik egyfajta, a dolgok között lévő, „rejtett kapcsolatokra" vonatkozó tudás; illetve azzal, hogy ennek a tudásnak van (vagy lehet) gyakorlati haszna (Henry 2012: 77–78). Vagy az okozatiság elfogadása helyett Hume megoldását választjuk, aki viszont abból indult ki, hogy lehetetlen meggyőződni az okozati összefüggések létezéséről (elvégre a klasszikus példa szerint mindegy, mennyi fehér hattyút figyelünk meg, nem következtethetünk teljes bizonyossággal arra, hogy a jövőben nem fogunk egy feketét találni), majd pedig ebből arra következtetett, hogy akkor valójában „nem is léteznek" oksági kapcsolatok (Okasha 2002: 51). És bár ez a tudományfilozófia számára érdekes felvetés, a modern természettudomány per definitionem az oksági magyarázaton alapuló értelmezésre épül, és ezen még a statisztikai vagy éppen sztochasztikus jelenségek vizsgálata sem változtat. Ugyanis még ezeknél is élhetünk azzal a feltételezéssel, hogy léteznek bizonyos törvények, amelyek mintegy „okozzák" a jelenségeket (még ha a magyarázatot nem is tudjuk az egyes események szintjére lebontani), és ebben az értelemben a társadalomtudományok „matematizálható" részei sem különböznek olyan nagyon. Más kérdés, hogy mi lenne a helyzet egy, a big data alapú megközelítés elterjedése esetén.

Tudománytörténetek – nagyon röviden

Mindenesetre a big data alapú értelmezés perspektívába helyezéséhez észre kell vennünk, hogy a fizika „elsődlegessége" jelentős mértékben befolyásolta az utóbbi idők tudománytörténettel kapcsolatos gondolkodását is. Ami a területtel nem hivatásszerűen foglalkozókat illeti, közöttük valószínűleg mindmáig az a megközelítés a legnépszerűbb, mely szerint a tudomány története mintegy azonos a tudományos gondolatok történetével. Mondhatni: Nagy Emberek + Nagy Eszmék = Tudomány (vagyis: egy, a feladattal magányosan megküzdő Galilei vagy Einstein tevékenysége eredményez tudományos előrelépést/áttörést); illetve, mivel a tudomány története a tudományos eszmék története, ezért a tudományos eszmék változásai a fontosak.

Thomas Kuhn tudományos forradalmakkal kapcsolatos elképzelései is lényegében ezen a megközelítésen alapultak az 1960-as évek elején. Legalábbis ironikus, hogy miközben Kuhn azt hangsúlyozta, hogy helytől és időtől függően változhat, mi az elfogadott tudományos paradigma, arra nem figyelt fel, hogy az általa kidolgozott paradigmafogalom is egy adott történeti kor terméke. Többek között abban az értelemben, hogy míg a természettudományok jelenleg leginkább úgy működnek, hogy egy-egy szakterület kutatói

kizárólag egyetlen, centrális paradigmát tartanak elfogadhatónak azzal kapcsolatban, mi számít tudománynak, és mik a megengedhető tudományos módszerek, korábban nem mindig volt így (és értelemszerűen a jövőben sem biztos, hogy ez lesz). De abban is magán viseli a korabeli gondolkodás lenyomatát a kuhni koncepció, hogy a tudományos elméletek elsődlegességéből indul ki.

Valójában azonban ott vannak a műszerek is, mint a lehetőségeinket befolyásoló tényezők. Freeman Dyson amerikai fizikus az 1990-es évek végén vezette be a tudománytörténész Peter Galison nyomán az „eszközvezérelt forradalom" fogalmát a kuhni tudományos forradalmakkal párhuzamba állítandó. Eközben hat nagyobb koncepcióalapú váltást különböztetett meg (mint amilyen a kopernikuszi vagy az einsteini is volt), a Galison-féléből pedig mintegy húszat Galilei távcsövétől a DNS szerkezetének felfedezéséig (Dyson 1998: 50). Ugyanekkor Galison szerint Einsteinnél az elméleti megfontolások mellett kulcsszerepet játszott az is, hogy a berni szabadalmi hivatalban dolgozva rendszeresen találkozott a korszak egyik technikai kérdésével, a pontos távolsági közlekedést lehetővé tevő elektronikus órák szinkronizálásával. Ez aztán – különböző elméleti megfontolásokkal együtt – elvezette a tér és idő kapcsolatának újraértelmezéséhez (Agar 2012: 39).

Egy olyan korban viszont, amikor a big data az érdeklődés homlokterébe került, a múltat nem csupán az eszközök vagy elméletek történeteként írhatjuk le, hanem mint az adatok gyűjtésének és kezelésének történetét is. A különböző értelmezések pedig nem feltétlenül egymást kizáróak, hanem sok esetben inkább komplementerek: az, hogy az elméleteknek hatása van a tudományra, nem zárja ki, hogy (miként Einstein esetében is történt) az eszközöknek is legyen szerepe.

Meg persze az adatoknak.

Az emberiség története adatfeldolgozás története

A történet nem a big datával és nem az internettel, de még csak nem is az írással kezdődik. David Hume már a 18. században arról beszélt, hogy két lepkegenerációt mindig elválaszt egymástól a bábállapot, tehát lehetetlen közöttük az információtovábbítás, és ez alapvetően korlátozza a lehetőségeiket (Birg 2004: 18).

Az emberiség történetében az első „információs forradalomra" akkor került sor még valamikor a neolitikum idején, amikor egy hasonló korlát szűnt meg, mivel az átlagéletkor az addigi mintegy 20 évről a másfélszeresére nőtt, és immár volt rá (valamennyi) időnk, hogy felnőttként információkat gyűjtsünk és adjunk tovább. Az írás pedig, amely e nélkül a változás nélkül nem jöhetett volna létre, a következő lépésben azt tette lehetővé, hogy az összegyűjtött információkat ne csupán a közvetlenül utánunk jövő generációnak juttathassuk el (Birg 2004: 18), illetve eljuttassuk térben távolabbi pontokra is.

Viszont az információ megbízható sokszorosítása megoldatlan maradt: a középkorban a Beowulf óangol nyelvű hőseposz másolásakor például az „elefántokból" (elpenda) gyorsan „segítők" lettek (helpenda), lévén az utóbbi jóval elterjedtebb fogalom volt az előbbinél (Fulk és Cain 2013).

Kissé leegyszerűsítve: a modern értelemben vett tudományosság legalább részben azért nem jelent meg Gutenberg előtt, mert még ha végeztek volna is többé-kevésbé pontos méréseket, nem igazán volt rá esély, hogy az adatokat hibátlanul reprodukálhassák.

A nyomtatás elterjedése viszont magával hozta a textuális stabilitást: azt, hogy az egyszer kiszedett szöveget tetszőleges mennyiségben lehetett változatlan formában sokszorosítani. Nem véletlen, hogy Tycho Brahe, az utolsó nagy csillagász, aki a távcső felfedezése és a tudományos forradalom előtt élt, már az 1500-as évek második felében nagyságrendekkel pontosabb műszereket épített, mint elődei, miközben ugyanazokat a technológiákat alkalmazta, mint ők. Ugyanis immár volt értelme pontos adatokat előállítani. Mint ahogy az sem véletlen, hogy még saját nyomdát is működtetett az eredményei közzétételére (Johns 1998: 13). És végül az sem, hogy Kepler az ő pontos adatai alapján fedezte fel a bolygók mozgását leíró törvényeit – elvégre azok immár a rendelkezésére álltak. A könyvnyomtatás elterjedését követően ugyanis, mivel már volt esély az adatok megbízható rögzítésére, volt értelme precízebb műszereket készíteni is, és Tycho műszerei a 16. század végén sokkal jobb méréseket tettek lehetővé, mint a korábbiak, miközben a megépítésükhöz – a távcsővel ellentétben – nem volt szükség új ismeretekre (Johns 1998: 9).

De a textuális stabilitás önmagában még mindig kevés, mert csak az adatrögzítés megbízható – ami nem garantálja, hogy azok az adatok is megbízhatóak lesznek, amikkel például egy táblázat esetén dolgozunk. Tehát a következő lépés ezen adatok mechanikus előállítása volt azokon a területeken, ahol sok számolási feladatot kell végrehajtani és nagy az emberi hibázás esélye. A Nagy Francia Forradalom idején Gaspard de Prony vezetésével a szögfüggvények kiszámítását próbálták automatizálni úgy, hogy a „számítógép" elemeiként embereket használtak, akik mindegyike az Adam Smith-i logikát követve csak egyetlen, elemi számolási műveletet hajtott végre újra és újra (Grier 2005: 36).

A modern értelemben vett számítógép is ennek a megközelítésnek a továbbvitele – azzal a nem elhanyagolható különbséggel, hogy amikor az emberek helyét alkatrészek veszik át, akkor a rendszer sokkal megbízhatóbbá válik, és az adatokat immár nem csupán tetszőleges alkalommal tudjuk minőségromlás nélkül reprodukálni, de az előállításukba is lényegesen kevesebb hiba csúszhat.

A Turing-féle csőlátástól a végső laptopig

Eközben a 20. század második felében a számítógépek alapmodellje a Turing-gép lett, és ezzel már majdnem el is jutottunk tulajdonképpeni témánkhoz, a big datához. De csak majdnem, mert a Turing-gép absztrakt matematikai konstrukció, annak minden előnyével és hátrányával együtt. Turing ugyanis egyfajta automatizált tételbizonyításra keresve megoldást abból indult ki, hogy nem az számít, hogy hogyan, hanem kizárólag az, hogy mit csinál egy számítógép, és innentől kezdve minden, bizonyos elveket megvalósító komputer, mondhatni, ugyanazon ideális platóni komputer megvalósulásának tekinthető. Azaz a konkrét fizikai paraméterek teljesen lényegtelenek, ugyanis ha – lévén minden Turing-gép ugyanarra képes – elég időt hagyunk neki, akkor mindegyik képes lesz ugyanazokat a feladatokat megoldani, és teljesen mindegy, hogy lego-elemekből van-e összerakva vagy a legmodernebb processzorokra épül (Barrow 1992: 246).

Ezen a felfogáson alapul az úgynevezett pánkomputációs elmélet is, mely szerint, ha bármiből lehet számítógépet építeni, akkor miért ne lehetne ilyen vagy olyan formában számítógép maga az Univerzum is (Piccinini 2015)? Ez az absztrakt, Turing-féle gépen alapuló megközelítés évtizedeken keresztül uralta a számítástechnikai gondolkodást.

Ráadásul arra hajlamosította a kutatókat, hogy figyelmen kívül hagyják a konkrét, fizikai megvalósíthatóságot – mint ahogy a Turing-gép alapjául szolgáló matematika is deklaráltan a valóságban nem létező, ideális objektumokkal és az azokon végezhető műveletekkel foglalkozik. Ami azért problémás, állapította meg Rolf Landauer, az IBM kutatója már évtizedekkel ezelőtt, mert „az információ fizikai természetű", és ebből az következik, hogy minden, a valóságban elvégezhető számításnak fizikai korlátjai vannak (Landauer 1996: 188). Seth Lloyd amerikai fizikus később ki is számította, hogy legfeljebb mekkora számítási teljesítménnyel rendelkezhet az 1 kg anyagból előállított „végső laptop", ha figyelembe vesszük, hogy az adatátvitel sebességét például a fénysebesség (értsd: a természeti törvények) korlátozzák (Lloyd 2000).

Mindez azért érdemes legalább érintőlegesen megemlíteni, mert a Turing-alapú szemlélet nem teszi indokolttá olyan, a „matematikai ideák világán" kívüli szempontok figyelembe vételét, mint a processzorsebesség vagy éppen az információtárolási kapacitás növekedése, és ezért a big data mint új terület megjelenése első lépésben nem valamiféle elméleti paradigmaváltásnak, hanem a konkrét számítástechnikai megoldások fejlődésének köszönhetően következett be. Jellemző, hogy amikor a 2000-es évek elején a Microsoft által támogatott Science 2020 Group előrejelzéseket tett közzé a közeljövőben várható fejlődésről, akkor abban „mesterséges tudósok", „komputációs gondolkodás" vagy éppen „előrejelző gépek" szerepeltek. Az adatmennyiség növekedésével kapcsolatban viszont arra a következtetésre jutottak, hogy az információtovábbítás korlátai miatt „a legtöbb kutató csupán a hozzáférhető adatok kis részével fog dolgozni" (Emmott és Rison 2006: 16).

Ebben a „proto-big data" szemléletben minden bizonnyal szerepet játszik, hogy az eddigi négy, nagy információkezelési forradalom közül csupán az első (életkor-növekedés) nem kapcsolódott új technológia megjelenéséhez, a következő három: az írás, a nyomtatás és a számítástechnika (internet) azonban igen. Nota bene: a technológiavezérelt felfogásba jól beleillik a kvantumkomputer, amitől újabb, alapvető áttöréseket szokás várni – a big data viszont azzal, hogy „csak" nagyságrendekkel több és más minőségű (például strukturálatlan) adattal dolgozik, nem.

A negyedik paradigma

Jim Gray számítógépes szakember utóbbi időben nagy népszerűségnek örvendő elmélete a tudományos kutatás négy paradigmáját (és ennek megfelelően négy korszakát) különíti el. Az első az „experimentális tudomány", amely empirista módon a természeti jelenségek leírását célozza meg, és nem igazán foglalkozik az okokkal. A második a „teoretikus tudomány": itt a „modellezésen és általánosításon" van a hangsúly, míg a harmadik korszak/paradigma a „számítógépes tudományoké" meg „a komplex jelenségek szimulációjáé". A negyedik pedig a „magyarázó tudomány", amelyre az jellemző, hogy „adat-intenzív, statisztikai magyarázatokkal és adatbányászattal" dolgozik ahelyett, hogy különböző törvények és szabályok létét feltételezve a valóságot próbálná modellezni (Kitchin 2015: 3).

Vegyük észre, hogy a big data esetében nem azt a tudományfilozófusok számára amúgy alapvető jelentőségű kérdést kell megválaszolnunk, hogy léteznek-e természettörvények és egyéb tudományos összefüggések a valóságban (miként a realista álláspont képviselői állítják), vagy pedig csupán pontos előrejelző képességgel rendelkező magyarázatokat

dolgozhatunk ki, miként az instrumentalisták vélik (Barrow 1988: 10–11). Hanem azt, hogy milyen módszereket használjunk a céljaink eléréséhez. Ami esetünkben átfogalmazható úgy is, hogy vajon azért az okozatisággal és matematikai képletekkel dolgoztunk-e eddig, mert nem állt a rendelkezésünkre más? De most a big data lehetővé tesz egy újfajta megközelítést, amely segítségével korábban elérhetetlen célokat is megvalósíthatunk. Például a jelenleg szükségképpen túlegyszerűsítő fogalmakat (választópolgárok, nők, társadalom stb.) a társadalomtudományokban jobb esetben az n=all (azaz a „minden adat") alapján kialakított ismeretek válthatnák fel (Dessewffy és Láng 2015: 165). Mármint ha képesek vagyunk helyesen megítélni a big data jövőbeni szerepét.

Halászok és vadászok: kauzalitás helyett komputáció?

David Edgerton brit kutató említi az úgynevezett „publikus technológiák" kérdését – ilyen volt például a II. világháború után az atomenergia, amelytől az élet alapvető megváltoztatását várták (Edgerton 2008: 2). Nem csak az atomhajtású autók, repülőgépek, vonatok stb. elterjedését, de azt is, hogy az atomenergiának köszönhetően az „elektromosság olyan olcsóvá válik, hogy mérni sem lesz érdemes" – ami persze nem következett be (Anderson 2009: 61). De ugyanígy publikus technológia volt 2000 körül az internet is. Ám míg az atomenergia leginkább nem váltotta be a hozzá fűződő, eltúlzott reményeket, addig ez utóbbi leginkább igen (valóban elterjedt, és olyan szinten a mindennapi élet részévé vált, hogy olykor már csak az „ötödik közműként" hivatkozunk rá). E példák alapján érdemes megvizsgálnunk, mennyire indokoltak vagy éppen eltúlzottak a várakozások a már-már publikus technológiává váló big datával, illetve azzal kapcsolatban, hogy az át fogja alakítani/meg fogja változtatni magának a tudománynak a természetét is.

A big data alapjául szolgáló „adatbányászat gyakran bármiféle [kiindulási] hipotézis nélkül kezdődik", állítja egy, az egyik közkeletű vélekedést visszhangzó white paper (Schmitt et al. 2015: 5). Értsd: a hagyományos és a big datán alapuló megközelítés közötti különbség a vadászok és halászok közötti különbség. Az előbbiek megtehetik, hogy egy konkrét vadra mennek, míg az utóbbiak kivetik a hálójukat, és aztán nincs más dolguk, mint várni, hogy mi akad bele.

A halászhasonlat persze nem pontos, de annyit talán érzékeltet, hogy a big datán alapuló elképzelés mintha leginkább annak a Francis Bacon-nek 17. század eleji empirista felfogására lenne visszavezethető, aki amellett érvelt, hogy az előzetes hipotézisek eltorzítják a tudományt. Ezért elutasította a matematikai összefüggések keresését is, és a hipotézisek helyett olyan „példák táblázatainak" az összeállítását szorgalmazta, melyek alapján mintegy „maguktól" kirajzolódnának az összefüggések.

Ez az álláspont tudománytörténetileg annyiból érthető, hogy a klasszikus (=tudományos forradalom előtti) „kísérlet", miként már érintettük, leginkább nem a kérdések feltevésére és megválaszolására szolgált, hanem a már ismert elméletek igazságának egyszerű demonstrálására (Henry 2002: 36-37), Bacon viszont éppen ellenkezőleg: a kísérletet tette meg a tudományosság alapjának.

Az ő nyomdokain haladó „empiristák" pedig ma is komoly fenntartásokkal viseltetnek az okozatisággal szemben, és ezt az álláspontot mintha nem egy, a big datával foglalkozó kutató is átvenné. Chris Anderson szakíró például azt hangoztatja, hogy a big data felemelke-

dése „a tudományos elméletek végét jelenti", és „az adatözönvíz elavulttá teszi a [hagyományos] tudományos módszert". Egy Jill Dyché nevű kutató pedig azt, hogy „a big data adatbányászata olyan összefüggéseket és mintázatokat fed fel, melyek létére korábban nem is gondoltunk", és „az analízist elvégzőnek immár egyáltalán nem is kell hipotézist felállítania".

A nem empirista felfogást képviselő data-driven science viszont abból indul ki, hogy az adatok és az elméletek között egyfajta iteráció zajlik: az elméletek módosítják, hogy milyen adatokat keresünk, és az új adatok visszahatnak az elméletekre (Kitchin 2015: 3–6). A tisztán „adatvezérelt" kutatás pedig nem értelmezhető, hiszen valamiféle hálóra (=megfelelő eszközökre és kiindulási pontul szolgáló koncepciókra) még akkor is szükségünk van, ha nem tudjuk pontosan, mit akarunk megtalálni. Mármint ha nem akarjuk azt mondani, hogy a tudomány kizárólag bármiféle értelmezést nélkülöző leírás.

Az új adattudomány

A data-driven science a leginkább abban tér el a hagyományos tudományos felfogástól, hogy „nyitottabb a hibrid kombinációkra, ahol az abduktív, induktív és deduktív módszerek" keverednek. Azaz: megpróbálja kihasználni mind a big data, mind pedig az eddigi megközelítés előnyeit (Kitchin 2015: 9), és eközben két célja lehet. Az egyik a nagyobb merítésen alapuló új törvényszerűségek felfedezése még akkor is, ha eddig a big data nem vezetett váratlan és forradalmi tudományos áttörésekhez – és lehet, hogy nem is fog. Bár élhetünk azzal a feltételezéssel, hogy egyszer talán majd igen, abból, hogy megteremti a lehetőséget új típusú problémák vizsgálatára, nem következik szükségképpen, hogy fogunk is valamit találni. Mint ahogy az sem, hogy szükségképpen egyformán hasznos lesz a természet- és társadalomtudományok számára. Használható viszont a már meglévő (small data) eredmények tesztelésére, és például az oksági kapcsolatokra vonatkozó megérzéseink helyességét ellenőrizhetjük a segítségével (Kitchin 2015: 9).

Hogy valamivel távolabbról, a matematikából is hozzunk hasonlatot: a kis számok esetén megfigyelt összefüggések sok esetben nem alkalmazhatóak a nagy számokra. Értsd: a kis mintából származó következtetés korántsem mindig működik, és ez elengedhetetlenné teszi a minél nagyobb számokon való kísérletezést, mielőtt ha nem is egy szabályszerűséget, de legalább egy sejtést megfogalmaznánk (Guy 1988: 697).

A természettudományok annyiban jobb helyzetben vannak a matematikánál, hogy – mondhatnánk némi cinizmussal – a matematikai végtelenhez képest minden szám kicsi. Nem mintha itt nem találhatnánk magunkat szembe olyan kombinatorikai robbanással, ahol esélyünk sincs az összes releváns adatot begyűjteni és megvizsgálni.

Ami a társadalomtudományokat illeti, a big data sikersztorijai, melyek hatására egyesek egyenesen a kauzalitáson alapuló tudomány végét vizionálják, leginkább innét származnak. A Walmart adataiból például az derült ki, hogy a hurrikánok előtt az amerikaiak nem csak több zseblámpát vásárolnak, de – ki tudja, miért – több Pop-Tarts nevű cukrozott snacket is (Mayer-Schönberberger és Cukier 2014: 64). Ez persze rendkívül hasznos felismerés – mint ahogy az Amazon számára is hasznosak az arra vonatkozó adatok, hogy ha valaki megvesz egy bizonyos könyvet, akkor vajon meg fog-e venni egy másikat is, és még az okok ismeretére sincs szükség ahhoz, hogy ezt az információt egy ajánlási rendszerben fel tudjuk használni.

Ne feledjük azonban, hogy a jelenlegi felfogás szerint a természet- és a társadalomtudományok többek között abban is alapvetően különböznek egymástól, hogy a fizika például abból indul ki, hogy léteznek bizonyos törvények, melyek meghatározzák, hogy mi fog történni: hogyan esik egy elengedett tárgy a föld felé, vagy miként görbíti meg a téridőt egy nagy tömegű csillag. Azaz a fizika, illetve általában véve a természettudományok, általában előíró jellegűek. A társadalomtudományok viszont inkább azt próbálják különböző szempontok alapján bemutatni, hogy mi történt, és ezekhez próbálnak különböző, nem feltétlenül egymást kizáró magyarázatokat fűzni. Azaz: általában inkább leíró jellegűek, és a filozófiai alapjaik sem olyan egységesek, mint a természettudományoknak. Ennek megfelelően a kauzalitás sem játszik olyan központi szerepet bennük.

És a jövő? A társadalomtudományok és a kauzalitás

A természettudományok esetében az egyik szélsőséges forgatókönyvet – ahogy azt korábban már érintettük – a neobaconiánus empirizmus jelenti, a másikat pedig értelemszerűen az, mely szerint a big data múló divathóbort csupán, és a 20. században elfogadott alapok végül majd változatlanok maradnak. A talán legvalószínűbbnek tűnő középút viszont a data-driven science lenne, amely szemléletének az elfogadottá válása talán azt eredményezné, hogy a természettudományok bizonyos értelemben a számelméletre kezdenének hasonlítani, ahol viszonylag könnyen lehet új összefüggéseket, míg nehezen lehet igazolást találni hozzájuk. A 20. század egyik legismertebb matematikusa, G. H. Hardy megfogalmazásával élve: itt „a leghíresebb tételek mindegyike olyan sejtésekre épít, amelyeket olykor évszázadokkal vagy még régebben felvetettek; és amelyek [bizonyítása] a nagy mennyiségű számolásból eredő bizonyosságon alapult" (Hardy 1967: 651). Ellentétben, mondjuk, a geometriával, ahol az eukleidészi módszer szigorú, axiomatikus építkezése másfajta logikát tett lehetővé.

A társadalomtudományok esetében szintén lehetséges, hogy minden marad a régiben. Ezt az álláspontot képviselők szeretik hangoztatni, hogy „az a folyamat, mely révén az irodalmat adattá változtatjuk, elveszi az egész ízét", és szerintük hasonló mondható általában véve is a társadalomtudományok „adatosításával" kapcsolatban. Az ezzel ellentétes álláspont szerint viszont a big data alapú irodalom(vagy általában véve társadalom)kutatás nem felváltja, hanem kiegészíti az eddigieket – még akkor is, ha itt is beleütközünk az okokozatiság problémájába (Kitchin 2015: 8). Aminek viszont még súlyosabb következményei lehetnek, mint a természettudományok esetében.

A big data árnyoldalairól beszélve minden bizonnyal a privacy-vel kapcsolatos problémák is az eszünkbe jutnak. Például az, hogy a small data korával ellentétben immár nem megoldás, ha hozzájáruláshoz kötjük az adatgyűjtést. Amikor a Google Street View lehetővé tette, hogy Németországban az emberek elhomályosítsák a házukról készült felvételt, ha attól tartanak, hogy az vonzó célpontnak látszana a bűnözők számára, akkor éppen az ilyen, elhomályosított képek váltak árulkodóvá (Mayer-Schönberberger és Cukier 2014: 170). És hasonlóképpen: egy hálózatosodott társadalomban már nem elég csupán azt meggátolni, hogy valaki hozzáférjen a személyes adatinkhoz, amennyiben megpróbáljuk megőrizni a hagyományos értelemben vett privacy-t. Amikor az AOL a 2000-es évek elején kutatási célokra nagy mennyiségű anonimizált adatot tett közzé (például a keresésekhez

kapcsolódó IP-címeket törölve), akkor kiderült, hogy a metaadatok is elegendőek az egyes felhasználók azonosításához (Mayer-Schönberberger és Cukier 2014: 171). De változik az adatok időbeni hozzáférhetősége is, úgyhogy ma már mintegy a big datára adott válaszul létezik a RTBF (Right To Be Forgotten). Ez nem csak arra fókuszál, hogy milyen információk érhetőek el rólunk aktuálisan, hanem arra is, hogy meddig (Székely 2015: 221).

Az, hogy a big datának ilyen alapvető következményei vannak, valójában nem meglepő. A modern értelemben vett privacy is az olyan újabb információs technikák felemelkedésével jelent meg, mint amilyen a telefon, a gyors sajtó vagy éppen a Kodak fényképezőgépe (Smith 2000: 126), és amennyiben valóban új technológia jön létre, az szükségképpen hatni fog az olyan, különböző társadalomtechnikai rendszerekre, mint amilyen például a jog is, amely ma lényegében a klasszikus fizika kauzalitását veszi alapul abból kiindulva, hogy az elkövetőt (ok) és az elkövetett tettet (okozat) összekapcsolva az „okozást" kell büntetni a klasszikus „ha – akkor" logika alapján. Csak éppen mi lenne, amennyiben nagy, sőt nagyon nagy valószínűséggel előre jelezhetnénk valaki cselekedeteit, mivel már olyan big data módszerek állnának a rendelkezésünkre, melyek a modern jog kialakulásakor még nem? Azaz: korábban nem csupán azért választottuk-e a hagyományos megoldást, mert nem volt más?

Természetesen nem amellett akarok érvelni, hogy zárjunk valakit börtönbe csak azért, mert a „viselkedési mintázata", a vele kapcsolatos adatok (és így tovább) arra utalnak, hogy potenciális bűnelkövető. Szó sincs erről. Arról viszont nagyon is szó van, hogy a társadalomtudományok logikájától, ahol a kauzalitás általában más (és kisebb) szerepet játszott, mint a természettudományokban, kevésbé áll távol, hogy inkább a mintázatokat keresse, mint az okozatiságot. A hagyományos érvelés szerint „a társadalom… túl komplex, esetleges és rendezetlen ahhoz, hogy képletekre és [nem jogi értelemben vett] törvényekre vezessük vissza" (Kitchin 2015: 8). Azonban most éppen, hogy nem szigorú értelemben vett természeti törvények, hanem csupán „általában igaz" szabályszerűségek után kutatunk, melyek a big data segítségével talán jobban megragadhatóak. Abból, hogy a természettudomány kauzalitása ezeken a területeken nem igazán vált be, nem szükségképpen következik, hogy nagyobb, de okozatisággal nem alátámasztható összefüggések sem fognak kirajzolódni.

Aztán persze ki tudja, hogy tényleg így lesz-e. Különösen, hogy amennyiben a big data tényleg alapvető változásokhoz fog vezetni, akkor joggal tételezzük fel, hogy amit jelenleg látunk, az csak az 1.0-ás verzió. Egyelőre „a szervezetek „új [big data] eszközeiket… még a régi módon használják: azt igyekeznek vele jobbá tenni, amit eddig is csináltak (Bőgel 2015: 42), és bármennyire is megírható a tudomány története a big data felől nézve, ebből legfeljebb az derül ki, hogy milyen lehetett a múlt, de az nem, hogy a jövő milyen lesz.

Irodalom

Agar, John, *Science in the Twentieth Century and Beyond*, Politiy, 2012.

Barrow, John D., *Pi in he Sky. Counting, Thinking and Being*, Little Brown and Company, Boston, 1992.

Barrow, John D., *World within World*, Oxford University Press 1988.

Bőgel, György, *A Big Data ökoszisztémája*, Typotex Kiadó 2015.

Batterman, Robert, „The Tyranny of Scales", in Robert Batterman (ed.), *The Oxford Handbook of Philosophy of Physics*, Oxford University Press, New York, 2013, pp. 1-23.

Birg, Herwig, *A világ népessége. Dinamikus növekedés és leselkedő csapdák*, Corvina Kiadó, 2004 [1996].

CERN, „Computing", dátum nélkül, https://home.cern/about/computing

Dessewffy Tibor és Láng László, „Big Data és a társadalomtudományok találkozása a műtőasztalon", Replika, 92–93 (2015-09-01), 157-170. old.

Dyson, Freeman, *Imagined Worlds*, Harvard University Press, 1998.

Edgerton, David, *The Shock of the Old. Technology and global history since 1900*, Profile Books, London, 2008 [2006].

Emmott, Stephen and Stuart Rison, „*Towards 2020 Science*", Microsoft Research, 2006. http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf

Fulk, R. D. and Christopher Cain, *A History of Old English Literature*, Wiley-Blackwell, 2013. https://books.google.hu/books?id=luXWM2fI8fEC&pg=PT58&dq=beowulf+elephant+helper&hl=hu&sa=X&ved=0ahUKEwi93MnFyqXOAhUqMJoKHeWPA2wQ6AEIHDAA#v=onepage&q=beowulf%20elephant%20helper&f=false

Gartner, „IT Glossary. Big Data", dátum nélkül, http://www.gartner.com/it-glossary/big-data/

Grant, Edward, *A History of Natural Philosophy From the Ancient World to the Nineteenth Century*. Oxford University Press, 2007.

Grier, David Alan, *When Computers Were Human*, Princeton University Press, Princeton and Oxford, 2005.

Guy, Richard K., „The Strong Law of Small Numbers", *American Mathematical Monthly*, 95 (8): pp. 697-712, 1988. https://doi.org/10.2307/2322249

Hardy, Godfrey H, *Collected papers of G. H. Hardy*, vol. 2, Clarendon Press, Oxford 1967. Henry, John, A Short History of Scientific Thought, Palgrave MacMillan, New York 2012 [2011].

Henry, John, *The Scientific Revolution and the Origins of Modern Science*, Palgrave, New York 2002 [1997].

Johns, Adrian, *The Nature of the Book. Print and Knowledge in the Making*, The University of Chicago Press, Chicago and London, 1998.

Kitchin, Rob, „Big Data, new epistemologies and paradigm shifts", *Big Data and Society*, April – June 2014, pp. 1–12. http://dx.doi.org/10.1177/2053951714528481

Landauer, Rolf, „The physical nature of information", *Physics Letters A* 217, 1996, pp. 188-193. http://dx.doi.org/10.1016/0375-9601(96)00453-7

Lloyd, Steh, „Ultimate physical limits to computation", https://arxiv.org/abs/quant-ph/9908043v3, http://dx.doi.org/10.1038/35023282

Mayer-Schönberberger, Viktor – Kenneth Cukier, *Big data. Forradalmi módszer, amely megváltoztatja munkánkat, gondolkodásunkat és egész életünket*, HVG Könyvek, 2014 [2012].

Okasha, Samir, *Philosophy of Science. A Very Short Introduction*, Oxford University Press, 2002.

Orrell, David, *The Future of Everything. The Science of Prediction*, Thunder's Mouth Press, New York, 2007.

Piccinini, Gualtiero, „Computation in Physical Systems", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, The Methaphysics Research Lab, 2015. http://plato.stanford.edu/archives/sum2015/entries/computation-physicalsystems/

Charles P. Schmitt, Steven Cox, Karamarie Fecho, Ray Idaszak, Howard Lander, Arcot Rajasekar, Sidharth , „Scientific Discovery in the Era of the Big Data: More than the Scientific Method", Renci White Paper Series, Vol 3, No. 6. 1 2015 http://renci.org/wp-content/uploads/2015/11/SCi-Discovery-BigData-FINAL-11.23.15.pdf

Smeenk, Chris, „Philosophy of Cosmology", in Robert Batterman (ed.), *The Oxford Handbook of Philosophy of Physics*, Oxford University Press, New York, pp. 1-34, 2013.

Smith, Robert Ellis, *Ben Franklin's Web Site. Privacy and Curiosity from Plymouth Rock to the Internet*. Privacy Journal, Providence, 2000

Székely Iván, „Az adatmentes zónák szükségessége és esélye. Helytelen reflexió Dessewffy Tibor és Láng László írására", *Replika*, 92-93 (2015-09-01), 209-225. old.

Woodie, Alex, „Why Gartner Dropped Big Data Off the Hype Curve", *Datanami*, August 26, 2015, http://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/

Galántai Zoltán (1964) jelenleg a BME BTK Pénzügyek Tanszék docense. Korábban az MTA Jövőkutatási, majd az MTA Tudomány- és Technikatörténeti Komplex Bizottsága titkára és a Magyar UNESCO-bizottság tagja volt. Könyvei jelentek meg többek között a földönkívüli értelem kutatásának kultúrtörténetéről (Marscsatornák, idegen civilizációk, angyalok… Pesti Szalon 1996); a távoli jövő kutatásáról (Majdnem az örökkévalóságig, Arisztotelész 2006); az olvasás és írás jövőjéről (Könyvkettő, eClassic 2013). Érdeklődési köre: tudománytörténet és tudományfilozófia.