Statisztikai Szemle

Közzététel: 2019. november 7.

A tanulmány címe:

Könyvek a Big Data jelenségről

Szerző:

DUSEK TAMÁS, a Széchenyi István Egyetem tanszékvezető egyetemi tanára
E-mail: dusekt@sze.hu
a *Statisztikai Szemle* főszerkesztője
E-mail: Tamas.Dusek@ksh.hu

DOI: https://doi.org/10.20311/stat2019.11.hu1071

Dusek Tamás

Könyvek a Big Data jelenségről

Books on the Big Data phenomenon

DUSEK TAMÁS, a Széchenyi István Egyetem tanszékvezető egyetemi tanára
E-mail: dusekt@sze.hu
a *Statisztikai Szemle* főszerkesztője
E-mail: Tamas.Dusek@ksh.hu

A digitalizáció különféle formái révén létrejövő adatoknak és azok használatának egyre általánosabbá, a hétköznapi életben is egyre érzékelhetőbbé válásával párhuzamosan a jelenségről szóló könyvek száma is gyorsan növekedett az elmúlt években. A Big Data szóösszetétel többnyire már a könyvcímben is megjelenik, amely kifejezés nagy karriert befutva mára már elvesztette újdonságértékét, azt fokozatosan átadva új divatkifejezéseknek. A Google találatok száma a „Big Data" szavakra 188 millió volt 2019. szeptember végén; ha mindegyik lapot meg szeretné valaki nyitni, akkor másodpercenként egy lap megnyitása mellett erre 6 évre lenne szüksége. Általánosságban ritkább, de a statisztikus körökben előfordul az adatforradalom kifejezés (561 ezer Google találat a „data revolution"-ra, 847 az adatforradalomra) is. Az adattömeghez és a digitalizációhoz kapcsolódik még az adattudós, adatbányász, mesterséges intelligencia, gépi tanulás, neurális hálózatok, adatvezérelt döntéshozás, üzleti analitika, prediktív modellek kifejezések népszerűbbé válása, olykor a statisztika szó háttérbe szorításával, eltüntetésével, például egyetemi kurzusok megnevezéséből, tananyagából. A Big Data magyar megnevezése még nem honosodott meg, ezért én is maradok az ismertetésben az angol kifejezésnél, amelynek tartalmára vonatkozó sokszínű értelmezésekre az ismertetések kapcsán visszatérek majd.

A Big Datáról szóló egyes könyvek rendkívül szerteágazók tematikailag, megközelítésmódban, színvonalban és a szerzők szakmai hátterében. Egészen más egy statisztikus, egy informatikus vagy egy szociológus megközelítése, vagy alkalmazott kutatásai során a digitális adatokkal találkozó közlekedésmérnöké, klinikai kutatóé, meteorológusé, vagy a nem szakemberek, hanem a népszerű téma iránt érdeklődő újságíróké, más laikusoké. Vannak vándortémák, mint például a 3V a Big Data defi-

Statisztikai Szemle, 97. évfolyam 11. szám 1071–1086. oldal DOI: 10.20311/stat2019.11.hu1071

nációjára, de még kevés a standard, tankönyvszerűen kikristályosodott elem. Ezt valamennyire érzékeltetik majd a bemutatott könyvek is, de bizonyítani egy ennél részletesebb és szisztematikusabb áttekintéssel lehetne. Az egyes könyvekben elkerülhetetlenül megjelennek hasonló témák is, az átfedések olykor jelentősek, így egy csoportos könyvismertetéssel megtakaríthatók a könyvenkénti általános leírások és ismétlések. Mindenekelőtt a jelenségről szóló irodalom különböző műfaji és tematikus megközelítésmódjait tipizálom, elfogadva, hogy más kategorizálások is létezhetnek, valamint egyes könyvek egynél több megközelítést is alkalmazhatnak:

1. Statisztikai megközelítés. Az adatok digitális eredete a statisztika minden elemét érintő változásokkal jár együtt: adatgyűjtés, adatminőség, adattisztítás, adatszerkezet, adatvédelem, adatetika, megfigyelési egységek és sokaságok, általánosíthatóság, adatelemzés, értelmezés, transzparencia, hozzáférhetőség, ellenőrizhetőség, megismételhetőség, adattárolás, adatarchiválás, megbízhatóság, vizualizáció. Ezeknek a kérdéseknek a teljes spektrumáról bőséges és szerteágazó voltuk miatt nehéz általános könyvet írni, egy-egy szerző nehezen tudja közvetlen tapasztalatból megismerni a digitális adatgyűjtés összes forrását. Így inkább egyes részterületeket lehet jól és kimerítően tárgyalni. Jellemzően könyvek rövidebb részei és tanulmányok foglalkoznak ezekkel az elméleti, koncepcionális kérdésekkel.

2. Ismeretelméleti megközelítés. Az adatok digitális eredetének és tömegességének hatása a tudományos kutatásra.

3. Technológiai megközelítés. A digitális eredetű adatok hardverei, szoftverei, adatbáziskezelői, információmenedzsment.

4. Üzleti alkalmazások és hatások, menedzsment, döntéstudomány. A Big Data gyakorlati alkalmazásai az élet bármely területén. Elsősorban a valamilyen szempontból sikeres alkalmazásokat mutatják be, amelyek lehetnek üzletiek, közigazgatásiak, települési szintűek, közlekedésszervezésiek, orvosiak és így tovább, de foglalkozhatnak a korlátokkal is, és lehetnek technológiaellenes alapállásúak is.

5. A digitalizáció társadalmi, politikai, pszichológiai sajátosságai. A magánélet megfigyelése és ellenőrzése az állam, valamint a nagy technológiai cégek részéről, digitális kormányzás és döntéshozás, politikai kampányok, szólásszabadság az interneten, cyberbiztonság és cyberbűnözés, az oktatás, a szórakoztatóipar és a sport átalakulása, illetve hasonló témakörök.

6. Ismeretterjesztő vagy bulvártudományos megközelítés, nem szakembereket megcélozva. Sokban hasonlít az előző megközelítéshez. Az ilyen könyvek nagyrészt információtechnológiai alkalmazá-

> sokra vonatkozó sztorik, anekdoták láncolatából állnak. Minőségükben eléggé vegyesek, de a gyengébb könyvekre vagy átlagos könyvek gyengébb részeire az jellemző, hogy hiányzik belőlük az egyes megoldások fenntarthatósága szempontjából érdekes költség-haszon elemzés, inkább a túlzó, bombasztikus megállapítások, a futurisztikus előrejelzések, a fantasztikus jövőkép, az „aki kimarad a folyamatokból, az elbukik" hozzáállás az uralkodó. Az ilyen könyvek hátsó vagy belső borítóján gyakoriak a következő jelzők: szenzációs, nélkülözhetetlen, forradalmi, egyedülálló, mély, inspiráló, elbűvölő, lenyűgöző, csodálatos, informatív, briliáns, világos, mély, átfogó, kötelező olvasmány. Az érvelés többnyire anekdotikus, vagyis bizonyítékként felsorolnak egyetlen vagy néhány pozitív példát, idéznek egy cégvezetőt, aki bevezette a rendszert, egy céget, amely sikeres volt az eljárás alkalmazásában. Mindezek általánosíthatóságával, az ellenvéleményekkel, az elbukó projektekkel pedig többnyire nem foglalkoznak.

Önmagában sem a cím, sem a kiadó nem elég annak eldöntésére, hogy milyen jellegű munkáról van szó. A Big Data sokszor marketing okokból, eladhatóságot növelő célzattal kerül be a címbe. Például *Bart Baesens* könyve a Wiley tudományos könyvkiadónál jelent meg, címét magyarra nehéz lefordítani: Analitika a Big Data világban: létfontosságú útmutató az adattudományhoz és annak alkalmazásaihoz (Analytics in a Big Data World: The Essential Guide to Data Science and its Applications). A kötet az első pár oldalban foglalkozik a Big Datával, rendkívül elemi szinten és hiányosan. Ezt követően néhány statisztikai témakört tárgyal, például mások mellett az adatstandardizálást, az adatvizualizációt (a kördiagramot ismerteti), a kategorikus változókat, a neurális hálózatokat; mindegyiket röviden. A könyv hasznos bevezető tankönyv lehet egy diáknak, de egyébként, címe ellenére, nem a Big Datával foglalkozik.

Hasonlóan a témát nem teljesen jól lefedő a Wiley egy másik tankönyvének a címe: Adattudomány és Big Data analitika: az adatok felfedezése, elemzése, vizualizációja és prezentálása (Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data). A munka egy sokszerzős kollektíváé. A bevezető fejezet jó általános áttekintést ad a Big Data jelenségről, a második fejezet az adatelemző projektek egyes részeiről. Ezt követően 7 fejezet foglalkozik az R programnyelvvel, 2 fejezet speciálisabb szoftverekkel, majd az utolsó az eredmények prezentálásával. A könyv túlnyomó része tehát adatelemző szoftvereket tárgyal. Ezek közül a szövegelemzés kapcsolódik legjobban a Big Datához.

Rátérve a részletesebb ismertetésekre, *Rob Kitchin* munkája az egyes, kettes és ötös megközelítés szintézise, és önmagában a statisztikai része rendkívül alapos, precíz, korrekt, eredeti. A mű címe: Az adatforradalom. Big Data, nyílt adatok, adat-

infrastruktúrák és azok hatásai (The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences). Az első fejezet valóban briliánsan tárgyalja az adat természetét. Érdekes, hogy angolul és magyarul ugyanúgy az ad ige a szó gyöke, csak az angolban latin eredetű, az ad jelentésű latin dare igéből származik. Ebben az értelemben az adat (data) a jelenségek bármely megfigyelhető, mérhető, feljegyezhető eleme lehetne. Azonban az adat (data) ennél szűkebb körét jelenti a jelenségeknek, nevezetesen a ténylegesen megfigyelt, megszámolt, kiszámolt, megtapasztalt és feljegyzett részét. Így technikailag, amit adatnak (data) neveznek, azt inkább kellene capta-nak (az elvesz jelentésű latin capere szóból) nevezni, amely a lehetséges összes potenciális adat (data) közül a ténylegesen megfigyelt adatot jelentené. 1950-es évekbeli szerzőkre hivatkozva Kitchin azt írja, hogy szerencsétlen történelmi véletlen, hogy az adat (datum) lett az elnevezése a captum helyett a tudomány egységnyi jelenségének, mégpedig azért, mert a tudomány nem a jelenségekkel foglalkozik a maguk adott természetében, hanem a kutatók által célzottan kiválogatott és megfigyelt jelenségekkel. Ennek a szótörténeti és jelentéstörténeti kitérőnek az volt az értelme, hogy megvilágítsa és kihangsúlyozza, hogy a begyűjtött adatok az összes lehetséges adat egy részét képezik, azt a részét, amit kiválogattak belőlük. Az adatok így lényegüknél fogva részlegesek, válogatottak és képviseletiek, kiválasztási szempontjaiknak pedig következményeik vannak.

Egy további nyelvi kitérőt érdemes tenni. Angolul történetileg a data a datum többes száma a latin eredet miatt, napjaink angoljában viszont a data a többes szám mellett egyszerre használatos egyes számban is mint tömeges főnév, hasonlóan az information szóhoz. Magyarul nincs ehhez hasonló nyelvtani bonyodalom, az adat az egyes számú, az adatok a többes számú forma. Más szemantikai problémák ugyanakkor a magyarban és az angolban egyaránt vannak akkor, amikor valaki az adat, az információ, a tény, a bizonyíték, a tudás és hasonló szavak jelentéseit, azok egymáshoz való viszonyát, illetve különbségeit próbálja meghatározni.

Az adat megelőzi az érvelést és értelmezést, amely átalakítja az adatot ténnyé, bizonyítékká és információvá. Kitchin az adatok sokféle tipizálásával foglalkozik:

– formája szerint: kvantitatív és kvalitatív adatok,
– szerkezetük szerint: strukturált, félig strukturált és nem strukturált adatok,
– forrásuk szerint: direkt, indirekt, tranziens, származtatott adatok,
– előállítójuk szerint: primer, szekunder és tercier adatok,
– típusuk szerint: kódoló, attribútum és metaadatok.

Mindezzel nem nagy terjedelemben, de mégis sokkal bőségesebb és alaposabb jellemzést ad az adatok természetéről, mint amit a tipikus statisztika tankönyvek vagy Big Datával foglalkozó kötetek tárgyalnak. Az adatok sokfélék, de közös ben-

nük, hogy az alapját képezik annak a tudáspiramisnak, amelynek a második szintje az információ (összekötött elemek), harmadik szintje a tudás (szervezett információ), negyedik szintje a bölcsesség (alkalmazott tudás). A piramis legalja, amelyen az adatok állnak, a világ. Ezzel a tudáspiramisnak nevezett koncepcióval sok helyen lehet találkozni. Nagyon sokféle értelmezése létezik mind az egyes szintek jelentésének, mind a szintek közötti kapcsolódások jellegének (ezt Kitchin tárgyalja), illetve a négy alapszintet többféle módon egészítik ki, így Kitchin az adatok alá a világ szintjét helyezte el. Nagyon sok vitapontot látok ebben a megközelítésben, amelynek kifejtése, érvekkel alátámasztása hosszadalmas lenne, de nagyjából úgy írható le a fő probléma, hogy mindegyik szintet (adat, információ, tudás, bölcsesség) homogenizálja, miközben ezen koncepciók részletes tárgyalásakor (mint itt az adatok kapcsán) mindig előkerül, hogy ezek mennyire heterogének.

A szerző az első fejezetben még foglalkozik az adatok pontosságával, érvényességével, hibáival, etikájával, politikai és gazdasági aspektusaival, időbeliségével és térbeliségével, infrastruktúrájával, az adatgyűjtés környezetével. Valóban olvasásra érdemes, olyan sűrű szöveg, amelyet nehéz lenne lerövidítve átadni. A filozófiai megközelítéseiről többek közt azt írja, hogy vannak, akik úgy határozzák meg az adatot, hogy az önmagában ártalmatlan, semleges, objektív, nem tartalmaz szükségszerűen értelmezést vagy véleményt. Mások számára ez nem elfogadható közelítés, szerintük az adat mérése és használata határozza meg az adatok keretrendszerét. Az adatok nem léteznek előállításukat megelőzően, nem a semmiből keletkeznek, hanem mérési eljárások révén termelődnek. Ugyanaz a jelenség sokféle módon mérhető, mindegyik más adatbázist szolgáltat, amelyek eltérő módon elemezhetők és értelmezhetők. Úgy gondolom, ezek a kérdések egyes területeken (élettelen természet, élő természet, társadalom, pszichológia) máshogyan jelentkeznek.

A második fejezet a Small Datával foglalkozik, amely kifejezést a Big Data megjelenése előtt nem használták, hanem annak ellentétpárjaként jelent meg (angolul a Small Data mellett a Little Data is használatos). A Big Data előtti legnagyobb felmérések az országos népszámlálások voltak, amelyek azonban ritkán (néhány évente, többnyire tízévente) végrehajtottak, feldolgozásuk lassú (minimuma hónapokban mérhető), korlátozott számú és mélységű kérdést vizsgálnak, eredményeik térben aggregálva (bár településszinten vagy számlálási körzet szinten) közöltek. A Big Data ezzel szemben folyamatosan keletkezik, azonnal feldolgozható (de valójában lehetnek itt is technikai hátráltató tényezők), és térben nagyon finom felbontású. A Small Data adatoázis egy egyébként adatnélküli sivatagban, a Big Data adatözönvíz. Ez oda vezetett, hogy egyes szerzők megkérdőjelezik a Small Data jövőbeli létjogosultságát. Az akadémiai kutatási támogatások is úgy csoportosulnak át, hogy az adatszegény területeknek egyre kevesebb, az adatgazdag területeknek egyre nagyobb támogatásokat ítélnek, az alapkutatások helyett pedig az alkalmazott, iparral partnerségben végzett kutatások az inkább támogatottak, miközben azok a kutatási

kérdések, amelyeken nehéz Big Datát létrehozni, a perifériára szorulnak. A prioritások ilyen változása Kitchin szerint a Big Data természetének és a Small Data értékének a félreértésén alapul. Kitchin bemutatja azokat a tényezőket, amelyek miatt a Small Data továbbra is nélkülözhetetlen a kutatás számára, a Big Data pedig korlátozott hasznosíthatóságú. A Small Data lehet kicsi mennyiségileg és lassú a feldolgozásában, de hosszú, megalapozott módszertannal és elemzési eljárásokkal rendelkezik nagyon összetett kutatási területeken. A Small Data célzottan tud összpontosítani nagyon speciális kérdésekre, aranyat kinyerve egy vékony telérből, míg a Big Data egész hegyeket tarol le egy kis arany kibányászása érdekében. A Small Datából azáltal is több érték származhat, hogy digitalizációja és archiválása után bárki számára bármikor hozzáférhetővé és újraelemezhetővé válhat. A Big Data lehet mennyiségileg nagy, de a könnyen elérhető, könnyen megfogható felületi jelenségekről ad számot, amely az illető digitális eszközt használókra korlátozódik. A fejezetben tárgyalt az adatinfrastruktúra, adatbrókerek, adatpiacok szerepe is.

A harmadik fejezet a nyíltadat- (open data) mozgalommal foglalkozik. Röviden megfogalmazva, a nyílt adat a bárki számára elérhető és terjeszthető adatokat jelenti, de elmélyülve a kérdéskörben ez számos problémát nyitva hagy, amelyekről (így a szerzői jogokat, költségeket, fenntarthatóságot) alapos elemzést kapunk.

A negyedik fejezet elején érkezik el a szerző a Big Data kifejezés eredetének tárgyalásához. 2008-ig a kifejezés ritkán volt használatos, aztán hirtelen berobbant a köztudatba, üzleti életbe, tömegmédiába, tudományba. 2013-ra egyesek már kiábrándultságukról írtak vele kapcsolatban, deklarálva, hogy a kifejezés jelentés nélküli és halott, de a többség (üzleti élet, kormányzat és tudománytámogatás) változatlanul meg van győződve arról, hogy a Big Data a tudomány és az üzleti élet gyökeres változását eredményezi.

Ezután a Big Data legelterjedtebb meghatározását (3V) ismerteti, miszerint az nagy mennyiségű (terabájtnyi vagy petabájtnyi), nagy sebességű (egyidejű vagy majdnem egyidejű) és változatos típusú. Kitchin a három közül a sebességet emeli ki, mint igazi megkülönböztető jegyet, mert a hagyományos adatok is digitalizálhatók, nagyméretűek és változatosak, de nem állnak azonnal rendelkezésre. A Small Datához képest négy további megkülönböztető tulajdonságot részletez:

– teljesség (a megfigyelések száma megegyezik a teljes sokaság elemszámával, ami a technológia következménye);
– felbontás (a megfigyelési egységek egyedileg is nyomon követhetők, illetve részletesebbek a térbeli és időbeli adatok);
– összekapcsolhatóság (különböző adatforrásoké);
– rugalmasság (a felmérések kérdései menet közben változtathatók, a megfigyelési egységek száma tetszés szerint és folytonosan növelhető).

A *Doug Laney* 2001-es cikkéből eredő 3V meghatározást nem tartom jónak, mert egy jó meghatározás a megkülönböztető sajátosságokra fókuszál és egyértelmű. Az adatmennyiség időben változó, nehezen mérhető, viszonyítási alapja is változik, ezért nem egyértelmű. Azért sem megkülönböztető sajátosság, mert nemcsak nagyméretű hagyományos adatbázisok léteznek (népszámlálások és hasonlók), hanem nem nagyméretű digitális eredetű adatbázisok is, amelyek keletkezési körülményeik miatt tulajdonságaikban eltérnek a nem digitális eredetű adatbázisoktól, de egyébként kényelmesen áttekinthetők és elemezhetők hagyományos módon. A sebesség a digitális eredet következménye, nem pedig egy alapvető ok. A változatosság (amely szintén nem mérhető egzaktul) a Big Data típusú adatok egészére jellemző, egy-egy Big Data típusú adatbázis többnyire egyáltalán nem változatos. Emellett a hagyományos adatbázisok összessége is rendkívül változatos.

Az adatmennyiség mérésének problémájával Kitchin is foglalkozik, különböző szerzők más-más eredményre jutnak, ami nem meglepő. Ami közös a becslésekben, az az adatok mennyiségének gyors időbeli növekedése. A tárolókapacitások növekedése révén eljutottunk abba a korba, amikor egyszerűbb mindent felvenni (rögzíteni) és tárolni, mint válogatni, szűrni és mintát meghagyni.

Az ötödik fejezet a Big Data technológiai hátterével foglalkozik, azokkal a sokszínű információtechnológiai eszközökkel, amelyek révén az adatok keletkeznek. Mindez jó történeti áttekintés és jelenkori leírás, elemzés is. Az adatforrásokat három nagy csoportra osztva tárgyalja:

– irányított adatok (megfigyelés, ellenőrzés, felderítés révén keletkeznek),
– automatikus adatok (egy eszköz normál működésének automatikus funkciójaként jönnek létre),
– önkéntes adatok (emberek által önként átadott adatok).

Ez a felosztás nagyon elterjedt lett a könyv megjelenése óta, ezért nem részletezem, de az elemzés itt is alapos, akárcsak a hatodik fejezetnél, amely az adatelemzés céljaival és eszközeivel foglalkozva szintén ismertebb lehet, és eléggé általános szinten marad, nem merülve el konkrét elemzési módszerekben. A hetedik fejezet a Big Data kormányzati, üzleti, értékteremtési és környezeti, települési szerepét elemzi. Ennél a résznél szokásosnak mondható módon hiányzik a különféle alkalmazási lehetőségek költségének tárgyalása, ezáltal úgy tűnik, mintha az információtechnológiai rendszerek létrehozatala, üzemeltetése, felügyelete és az adatok állandó elemzése során a költségektől el lehetne tekinteni.

A nyolcadik fejezet mesteri módon mutatja be az új adatvezérelt tudományos paradigma vagy máshogyan fogalmazva, az ismeretelméleti megközelítés hirdetőinek a gyengeségeit, tévedéseit. Ezen paradigma szerint a tudomány új, tiszta empiricista

fázisba érkezett. Az adatok teljesen átfogó módon mindenre kiterjednek, emberi elfogultságoktól és előítéletektől mentesen mutatják a valóságot, nincs szükség elméletre, modellre, hipotézisre, csak annak megfigyelésére, amit az adatok mutatnak. Az adatelemzéshez nem lesz szükség az adatok kontextuális sajátosságait ismerő szakterületi specialistákra sem, csak adattudósokra. Ezekben az igényekben annyi igazságmagvacska található, hogy bizonyos területeken lehetőség nyílik új vizsgálatokra, pontosabb megfigyelésekre. De egyrészt a korábbi adatgyűjtési technikák változatlanul pótolhatatlanok lesznek, másrészt számos olyan terület létezik (a társadalmi és a humán tudományok, művészetek majdnem teljesen ilyenek), amelyeknél a megértés, értelmezés, tudás nem lehetséges pusztán a digitalizált adatok ismeretében.

A kilencedik fejezet témája visszatérés az első fejezet témájához, az adat természetéhez. Az adatokhoz való hozzáférés kapcsán figyelemreméltó, hogy miközben mindenki az adatbőségről beszél, mennyire nehéz hozzájutni egy jó adatbázishoz, különösen olyanokhoz, amelyek jó időbeli és területi felbontásúak. A nagy adattermelő magáncégek, mint a telefontársaságok, pénzintézetek, kiskereskedelmi hálózatok, vagyonvédelmi cégek, közösségimédia-szolgáltatók természetesen nem kötelesek üzleti titkot képező adataikat szabadon bárkivel megosztani. Ha átadják valakinek kutatásra vagy üzleti célokra az adatbázisukat, akkor azt egyedi szerződés alapján, titoktartási és harmadik félnek való továbbadást megtiltó feltételekkel, az adatok felhasználását is szabályozóan teszik; ráadásul az adatokat sokszor aggregálva, az adatokból mintát véve, metaadatok nélkül szolgáltatják. Mindegyik adatbázis korlátozott időben, térben és a változók számában is. Az adatok minőségével, valósággnak megfelelésével, keletkezési körülményeinek ismeretével, az adatbázisok összekapcsolhatóságával kapcsolatban is számos, kevésbé köztudatban levő részletet tárgyal a szerző, sok téveszmét eloszlatva, például azt az igényt, hogy önmagában az adatbázis nagysága elegendő indok arra, hogy az adatokban található minőségi hibáktól, zajoktól el lehetne tekinteni.

A tizedik fejezet tér rá az etikai, politikai, társadalmi és jogi, adatvédelmi vonatkozásokra. Ezeknek az irodalma óriásira duzzadt mára, ehhez a szerző egy jó összefoglalással tudott hozzájárulni. Végül az utolsó fejezet összegzi az egész könyvet, amely valóban nagyszerűen, elmélyülten, építő kritikával közelíti meg az adatforradalom valamennyi aspektusát. Kisebb részterületeket lehet tovább részletezni, ahogyan számos más könyv teszi például az adatbiztonság, társadalmi és üzleti hatások kapcsán, de az összképet ennél jobban nehéz lenne bemutatni.

A következő ismertetendő könyv egy olyan részterülettel foglalkozik, amelyet Kitchin is érintett néhányszor, de nem központi kérdésként, nevezetesen a korántsem tökéletes adatokon és algoritmusokon alapuló döntéshozás veszélyeivel. *Cathy O'Neill* könyvének címe magyarul úgy szól, hogy „A matematikai pusztítás fegyverei. Hogyan növeli az egyenlőtlenséget és fenyegeti a demokráciát a Big Data". A cím angolul egy, a math (matematika) és a mass (tömeg) hasonlóságán alapuló

szójátékot tartalmaz. A könyv műfaja igényes ismeretterjesztés, példái és hivatkozásai aktuálisak és elsősorban az elmúlt néhány év napi sajtójából származnak. A szerző matematikus végzettségű, magát matematikát szerető embernek tartja, aki gyerekkorában matematikatáborba járt, és Rubik-kockával játszott, majd matematikából, algebrai számelméletből szerezte a doktori fokozatát. Ezt azért írja le, mert könyve a matematika sötét oldalával, káros alkalmazásával foglalkozik, de nem azért, mert általában lenne matematikaellenes. Dolgozott egyetemen matematikaprofesszorként, majd a 2008-as pénzügyi válságot megelőző évben az egyik vezető befektetési alapnál kamatoztatta ismereteit. A 2008-as őszi pénzügyi válság döbbentette rá, hogy a matematika helytelen pénzügyi alkalmazásai, ahelyett, hogy segítették volna a világ problémáinak megoldását, jelentősen hozzájárultak a válság kitöréséhez és elmélyítették annak nagyságát. A számítógép pillanatok alatt képes dönteni a különböző pénzügyi kérdésekben, de a döntés alapja az emberek által betáplált döntési szabály. A bevezetőben mindezt részletezve érzékelteti kiábrándultságát a szerző, mert úgy véli, hogy a matematikai modellekbe bekódolva nemcsak tovább élnek az emberi előítéletek és félreértések, de immár a matematika magas presztízse által megszentelt és megerősített állapotban, és így a felhasználók által megtévesztő módon objektívnak elfogadva a modellek eredményeit. 2011-ben a szerző elhagyta a pénzügyi szférát, és egy elektronikus kereskedelemmel foglalkozó céghez ment dolgozni, majd onnan is távozott.

Munkahelyi és egyéb általános tapasztalatai késztették végül a könyv megírására, amelyben egy-egy fejezetet szentelt különféle témáknak, ahol az általa a matematikai pusztítás fegyvereinek (angolul rövidítve WMDs, weapons of mathematical destruction) nevezett rossz, kártékony modellek működését mutatja be. Az ilyen modellek valamilyen input adatokat használnak fel ahhoz, hogy előre jelezzenek egy outputot, amely alapján szervezetek vagy emberek életére jelentős befolyást gyakorló döntés születik (például kap-e valaki hitelt, munkát), vagy ha az eredmény alapján döntés nem is születik, maga az eredmény egy olyan nyilvános rangsor, pontszám, amely nagyban befolyásolja a szervezetek és emberek presztízsét, viselkedését (például, amikor iskolákat, kórházakat rangsorolnak). A döntést hozó algoritmusok (amelyeket O'Neill egy helyen kódba formalizált véleménynek nevez) összetett műveleteit gyakorlatilag csak számítógéppel lehet elvégezni.

A modelleket jellemzően digitális eredetű adatok táplálják, de nem mindig. Így a Big Data jelenség valójában nem jelenik meg mindenhol. A könyv gondolatmenetét összekötő elem elsődlegesen nem a Big Data, hanem a numerikus információvá átalakított minőségi jellemzőket felhasználó modellek problémái, amely jelenség terjedése ugyanakkor összefügg a digitalizációval is.

Mielőtt a káros modelleket esettanulmányszerűen ismertetné, jó modelleket is bemutat. Így a baseballra vonatkozó sokféle statisztika alapján előre lehet jelezni egyes játékosok teljesítményét. Az ilyen modellek transzparensek (vagyis bárki szá-

mára elérhetők a felhasznált adatok), folyamatosan karbantarthatók, frissíthetők az újabb és újabb adatok ismeretében, előrejelzéseik összehasonlíthatók a tényleges eredményekkel. A mezőgazdasági vizsgálatokban is elég jól mérhető inputok és outputok vannak: napos órák száma, talaj minősége, csapadék és tápanyagok mennyisége, terméseredmény, amelyek ismeretében jó modelleket lehet készíteni.

A matematikai pusztítás fegyverei típusú modellek tulajdonságaira a szerző többször visszatér, ahogyan különböző alkalmazásait bemutatja. Összességében, elszórtan az egyes példáknál, a következő sajátosságokat említi, amelyek a különböző esetekben eltérő súllyal jelentkezhetnek. Egyrészt, az adatokat felhasználó döntéstámogató, értékelő eljárások több szempontból is homályosak. A kívülállók számára nem mindig tudható, mely változókat (jellemzőket, információkat) veszik figyelembe; vagy ha a változók ismertek, az azokra vonatkozó adatok nem hozzáférhetők, nem ismertek bárki számára; vagy ha a változók és az adatok is ismertek, akkor a döntési algoritmus (algoritmus alatt az elvégzendő matematikai műveleteket értve, de ezek O'Neill szerint matematizált vélemények) nem ismerhető a kívülállók számára. Bármelyik hiányosság előfordulhat akár egyszerre is. Így nem átlátható módon születnek a döntések, az eredmények, csak annyit tudunk, hogy ezt dobta a gép, ez van, ezt kell elfogadni. Másrészt, a modellek jellemzően nagyméretűek, sok embert érintenek, sok adatot használnak fel, és növekedésre hajlamosak. Ennek a tulajdonságnak ellentmond az, hogy a szerző szervezeten belüli, így lokálisnak nevezhető példákat is bemutat. Harmadrészt, az eljárások logikája olyan, hogy a kedvezőtlen kiinduló helyzetű csoportokat (így a szegényeket, a szegény környéken élőket, az alacsony iskolázottságúakat) a korábbinál még kedvezőtlenebb helyzetbe hozza, tartósítja és elmélyíti hátrányukat; a kedvező helyzetű csoportok tagjainak helyzete pedig még kedvezőbb lesz a pozitív és negatív visszacsatolások eredményeként. Ez utóbbi tulajdonság indokolja a könyv alcímében megjelenő növekvő egyenlőtlenségeket. Van még egy negyedik jellemző is, amely rendre előkerül, bár ez az első jellemző egy alkategóriája is lehet, nevezetesen az, hogy a modellek közvetlenül nem megfigyelhető jellemzőket írnak le (mint az oktatás minősége, bűnözésre való hajlam, munkavállalói alkalmasság), így kénytelenek helyettesítő változókat használni. A helyettesítő változók pedig sokszor a könnyebben manipulálhatók közé tartoznak, teljesen alkalmatlanok arra, aminek a leírására használják.

A szerző a következő területeken mutatja be, részletezi a problémákat egy-egy fejezetben: általános iskolák oktatóinak rangsorolása, felsőoktatási intézmények rangsorolása, online reklámok, bűnüldözés, rendőri és bírói munka, állásra jelentkezők rangsorolása, rugalmas munkaidőbeosztású helyeken dolgozók munkaidőbeosztása, hitelkérelmek elbírálása, biztosítások, politikai kampányok. Mindegyik amerikai példa, de általános voltuk miatt különösebb háttérismeretet nem követel meg a megértésük.

Az amerikai egyetemek rangsorolása kapcsán a szerző a US News médiatársaság által 1983 óta közölt, Amerikában széles körben ismert és hivatkozott rangsorral

foglalkozik. Bemutatja az inputokat, amelyek az oktatás minőségét és a diákélet jellemzőit hivatottak mérni. Ezek mind helyettesítő változók, mivel az oktatás minősége közvetlenül nem mérhető. A rangsort 25 százalékban egy szubjektív elem határozza meg, egyes megkérdezettek véleménye az egyetemekről. Ezt a rangsor homályos elemének nevezi a szerző. Bírálja, hogy kimaradt a tandíj és az oktatási költség, mert ezzel az egyetemi vezetők egy aranyozott csekkfüzetet kaptak a rangsor készítőitől azáltal, hogy 15 egyéb indikátornál kell ügyelniük a jó teljesítményre, amely között költségek nem szerepelnek semmilyen formában. A tandíjak 1985 és 2013 között 500 százalékkal emelkedtek, ami csaknem négyszerese az inflációs rátának. A tandíjemelkedés nem önmagában a rangsor következménye, de megerősített egy létező trendet és érzést, miszerint a magas presztízsű egyetemek költségesek, de megérik a tandíjukat, mert az ottani végzettség a társadalom jól fizetett felső rétegébe való belépést segíti. Nemcsak a tandíjak növekedtek jelentősen reálértékben, hanem a marketingre, reklámozásra fordított kiadások, a konzulens cégeknek fizetett tanácsadói díjak, az egyetemi kiegészítő létesítményekre (sportcsarnokokra, diákszállásokra) fordított kiadások is.

További nem szándékolt negatív következményként az adatok manipulálása, meghamisítása is megjelent, amire O'Neill több példát hoz. Végül a rangsor önbeteljesítő jóslatként kezd működni, a rangsor végén levő intézmények híre még rosszabb lesz, még kevesebb diák választja őket, akik között még nagyobb lesz az aránya a gyengébb középiskolai eredménnyel rendelkezőknek, növekszik a különbség a sor elején és végén levők között. A formalizált rangsorok előtt is volt képe az embereknek arról, hogy mekkora presztízse van egy-egy egyetemnek, de ez nem egzaktul, pontozásosan vagy rangsorral meghatározott volt. Pontokkal kifejezve a különbséget az egzaktságnak csak a látszatát kapjuk, mert a vizsgált jellemző, az egyetemi minőség természete miatt nem kvantifikálható.

Felsőoktatási példa megjelenik a következő fejezetben is, amely az internethasználati szokások megfigyelésén alapuló, személyre szabott online reklámok árnyoldalairól szól. Ezek a legsérülékenyebb embereket veszik célba, drága hiteleket, termékeket, képzéseket ajánlva nekik. Ők azok, akik hitelt is felvéve iratoznak be olyan online képzésekre, amelyek ötödakkora költséggel is elérhetők lennének. A Phoenixi Egyetem, amely hallgatónként 892 dollárt költ oktatásra, 2 225 dollárt marketingre, 50 millió dollárt költött szegényeket megcélzó Google hirdetésekre, a felfelé történő társadalmi mobilitás csalétkét használva. A Corinthian College online jogi asszisztens képzésének tandíja 68 800 dollár, miközben ilyen képzések 10 000 dollár alatt elérhetők. Az intézmény személyre szabott hirdetései az alacsony önbizalmú és társadalmilag elszigetelt internetfelhasználókat célozza. Egyes felsőoktatási intézmények egyéb internetes hirdetési trükköket is bevetnek, együttműködve marketingcégekkel. Például hamis álláshirdetéseket jelentettek meg, amelyre az érdeklődők megadták a telefonszámukat. Az álláshirdetésre jelentkezőket később visz-

szahívták, a visszahívottak 5 százaléka érdeklődést mutatott egyetemi képzések iránt. Ez arányaiban kicsi, de abszolút értelemben elég nagy lehet, és sokat érhet a túlárazott kurzusok miatt.

Ezek a módszerek azért hatékonyak, mert egy olyan részsokaságot céloznak meg, amely az átlagnál jóval nagyobb arányban meggyőzhető a reklámmal. Nem konkrétan valakit szeretnének a vásárlók között látni, hanem bárkit a megcélzott csoportból. Módszertanilag helyes, etikailag bírálható az eljárás. A szerző azonban olyan területeket is bemutat, amelyek módszertanilag helytelenek, mert statisztikai összefüggések alapján egyéni viselkedésre és nem egyének csoportjainak a viselkedésére következtetnek. Ilyen modelleket használnak a bírósági és rendőrségi gyakorlatban, nagyvállalatok munkaerő-felvételénél. Az az elítélt, aki munkanélküli, magas bűnözési rátájú körzetben lakik, szomszédjai és rokonai között vannak elítéltek, ugyanazért a bűntényért nagyobb büntetést kap, mint akire nem érvényesek az említett környezeti jellemzők. A modellek az egyénre jellemző sajátosságokat nem veszik figyelembe, ezeket helyettesítik néhány olyan csoportjellemzővel, amelyek szerepelnek a modellben.

Egyes esetekben a szerző gondolatmenetét nem látom alátámasztottnak. O'Neill a biztosítások egyénre szabását is bírálja, mert ez alapján szerinte a biztosítás az eredeti funkcióját elveszti. Ezzel nem értek egyet, ez csak abban az esetben lenne így, ha nem csoportjellemzők alapján történne a biztosítási díj egyénre szabása, hanem valóban egyedi szempontok alapján. A rendőrök utcai járőrözésének tervezését segítik a korábban ismertté vált bűncselekmények. Ahol korábban több bűncselekmény volt, ott a jövőben is nagyobb az esélye a több bűncselekménynek (ez nemcsak feltételezés, hanem tapasztalati megfigyelés is), ezért ott többet járőröznek a rendőrök. Így volt ez a Big Data előtti korszakban is, amikor még nem digitálisan tárolták az adatokat, hanem nagyméretű fali gombostűtérképeken jelölték be az eseteket. A szerző gondolatmenete szerint ez azért támadható, mert ahol több a járőr, ott több a felderített kihágás is, így ismét egy ördögi kör alakul ki szerinte, amely szintén a szegényebb, hátrányosabb helyzetű lakókörzeteket és azok lakóit érinti.

O'Neill javaslata a helyzet megoldására több részből áll. Egyrészt az adattudósoknak az orvosok hippokratészi esküjéhez hasonló fogadalmat kellene tenniük, amely megtiltja a modellek helytelen használatát és félreértelmezését. A törvényeknek is változniuk kellene, hogy védjék a kiszolgáltatott embereket a nem etikus internetes hirdetésektől.

A könyv felvállaltan, deklaráltan elfogult és csak a negatív következményekkel foglalkozik. Erénye, hogy nem csupán értékítéletet, ideológiát jelenít meg, hanem tapasztalatokon alapulóan érvel álláspontja mellett, amit sokszor elfogadhatónak tartok, de olyan esetekben nem, amikor már inkább egy rögeszméhez hasonlatossá válik. Ha annyi lenne az üzenete, hogy léteznek valódi veszélyek, amelyek nem csupán a képzelet szülöttei, akkor ezt a célt teljesítette volna. Mindezen korlátok mellett, elfogultságai ellenére is értékes olvasmányról van szó.

A következő kötet a hatodik megközelítést képviseli, jól érzékelhetően különbözik az előző könyvektől. Magyarul is megjelent a HVG Könyvek ismeretterjesztő sorozatában 2014-ben, alig egy évvel az angol megjelenést követően. A címe: Big Data. Forradalmi módszer, amely megváltoztatja munkánkat, gondolkodásunkat és egész életünket. Angol címében nem szerepel a zavaróan ható módszer, és jobban érződik a jövőorientáltság: Big Data: A Revolution That Will Transform How We Live, Work, and Think. A szerzők *Viktor Mayer-Schönberger* informatikus, jogász és *Kenneth Cukier* újságíró. A könyv erénye a szerteágazó történeti és aktuális példaanyag, amely a statisztikai adatok használatával megoldott sokféle különböző problémára vonatkozik. Ezekkel valóban jól lehet szemléltetni a statisztikai adatok és elemzések hasznosságát, de amely példa az egyik kutatási területen jól működhet, azt nem lehet automatikusan általánosítani mindenre úgy, ahogyan azt a szerzők rendre teszik. Emellett a példákat nem egy koherens gondolatmenet fűzi össze, hanem ad hoc módon jönnek egymás után, sokszor össze nem tartozó hirtelen váltásokkal, amelyek azt hivatottak bizonyítani, hogy a Big Data az élet minden területét meg fogja változtatni vagy már megváltoztatta.

A könyvet nem a 10 fejezet sorrendjében ismertetem, hanem tematikusan tárgyalom a főbb gyengeségeit. Ezek azért is tanulságosak, mert nemcsak ebben a könyvben lehet velük találkozni elszigetelten, de népszerűek és ismertek lehetnek a tudományos ismeretterjesztés és a felületes „tálalásra" hajlamos tömegmédia világából is. Nem a szerzők személyét bírálom velük, hanem a kifejtett nézeteket.

Az első problémakör a példák értelmezése és indokolatlan általánosítása. Egy-egy jó példa nagy segítség lehet absztrakt koncepciók szemléletessé tételéhez, a rossz és inadekvát példák viszont nem hasznosak, mert nem szabadna speciális körülményekre (információtechnológiai területen működő cégekre) vonatkozó tapasztalatokat általánosítani (az összes cégre). A könyv terjedelmének nagy részét a példák teszik ki, kiemelve a könyvből ezeket, nagyjából egy fejezetnyi szöveg maradna. Az összekötő szövegekben a szerzők egyfolytában a következőkhöz hasonló klisék köré építenek egész bekezdéseket: „útkeresés zajlik", „világunk átalakul", „a Big Data átformálja az üzleti életet", „a társadalom számára sok előnnyel jár majd mindez", „a világ a kauzalitás keresésétől a korreláció felé fordul".

Számos példa végighúzódik a könyvön, a két legnépszerűbb a Google influenza-előrejelző rendszere és a Farecast repülőjegyek árát előrejelző rendszere. Érdekes módon mára mindkét rendszer megszűnt. A Google influenza-előrejelző rendszere az internetes keresések révén prognosztizálta az influenzajárványok kitörését. Az azóta már tapasztalati okok, vagyis a rossz előrejelzések miatt megbukott projektről a szerzők lelkesen írtak: „A Google módszeréhez nincs szükség kenetek vizsgálatára, és háziorvoshoz sem kell fordulni. Ez a módszer az ún. big datára épül, vagyis a társadalom azon képességére, hogy az információt új módon munkára fogva hasznos felismerésekhez, értékes árucikkekhez vagy szolgáltatásokhoz juthassunk. Ennek

köszönhetően az emberiség már egy korábbiaknál hatékonyabb eszköz birtokában készülhet fel egy esetleges új világjárványra." (10–11. old.) A mintavétel hátrányait, a teljes sokaság előnyeit bemutató részben olvashatjuk: „A Google influenza-előrejelző rendszere, a Flu Trends éppen ezért nem kisszámú véletlenszerűen kiválasztott mintára hagyatkozik, hanem az Egyesült Államokban végzett több milliárd internetes keresésre. Az összes adat felhasználása olyan mértékben javítja az elemzést, hogy az influenza terjedése már nemcsak az egész ország vagy egy állam, hanem egy-egy konkrét város vonatkozásában is megjósolható." (36. old.) A korrelációról írva: „Ilyen erős korrelációt láttunk például a Google Flu Trends esetében: minél többen kerestek rá bizonyos kifejezésekre a Google-lal egy adott területen, annál többen voltak ott influenzások." (63. old.) A hipotézisalapú megközelítés idejétmúlt, elavult módjáról írva: „A kifinomult számítógépes elemzésekkel ma már ki lehet választani az optimális jelzőértékeket, ahogyan az csaknem félmilliárd matematikai modell átrágását követően a Goggle Flue Trends esetében is történt." (65. old.) „Korrelációs elemzésnek vethetjük alá a big datát, hogy megmondja nekünk, milyen keresési lekérdezések a legjobb jelzőértékei az influenzának." (66. old.) Arról írva is megjelenik példaként az influenza előrejelzése, hogy az oksági kapcsolatokat ritkán lehet bizonyítani, ha egyáltalán lehet bizonyítani: „Hogyan végezhetünk oksági kísérletet annak kimutatására, hogy bizonyos kifejezések internetes keresése miért jelzi előre az influenza terjedését?" (77. old.) A későbbi részekben még legalább ennyiszer ismétlődik a példa.

A második problémakör a mintákkal, adatmennyiséggel, adatminőséggel, adattartammal és néhány további kapcsolódó területtel kapcsolatos. A szerzők szerint a Big Data korszakban „a véletlen mintavételhez folyamodni olyan, mint lovaglóostort ragadni egy motorizált világban". (41. old.) Ez lényegesen más hasonlat, mint Kitchin aranyteléres (Small Data) és hegyeket elbontó külszíni fejtéses (Big Data) aranybányászata, és sejthető, hogy félreértéseken alapul. Állításuk szerint a digitális adatok tömege, nagyságrendje bőségesen kompenzálja az adatokban rejlő hibákat és torzításokat. Ezt elméletileg és gyakorlatilag is már régen megcáfolták, egy nagy és torz minta még a kicsi torz mintánál is rosszabb lehet a méretből fakadó más típusú problémák, adathibák miatt. „Az információ kissé pontatlanabb lesz, de a hatalmas adattömegért ez mégis elfogadható ár." (44. old.) A szerzők nem hasonlítják össze szisztematikusan a mintavételből származó és a Big Data típusú adatokat, alig érintik a teljes körű felméréseket, és egyáltalán nem írnak a teljes körű adminisztratív adatforrásokról. Nem foglalkoznak azzal, hogy az információ nemcsak pontatlanabb lesz, hanem tartalmában és jellegében tipikusan másra vonatkozik, mint a kérdőíves felmérésekből nyerhető információ; de tartalmában azonos információ is más lehet az adatfelvételi módszer hatása miatt. Azt sem érintik, hogy a kérdőíves felmérések változógazdagságát hogyan lehet pótolni a jellemzően egy vagy néhány változós Big Data típusú adatforrásokkal. Nem írják le, hogy hol vannak azok a Big Data típusú adatok, amelyek tartalmukban és

mélységükben hasonlók a változógazdag kérdőíves felmérésekhez. Ilyen példáik nincsenek. Egy konkrét összehasonlításuk a szumómérkőzések 11 évnyi összes (64 ezernyi) eredményének a vizsgálatára vonatkozik. Szerintük a következtetéseket (így azt, hogy a nagyobb tétért birkózók 25 százalékkal gyakrabban nyertek, mint ahogy normális lett volna) „véletlenszerű mintavétellel azonban valószínűleg nem lehetett volna kimutatni" (38. old.). Ez az állítás a véletlen mintavétellel kapcsolatos elméleti ismeretek minden alapot nélkülöző tagadása, ráadásul az elméletet tapasztalati és szimulációs úton is számtalanszor bizonyították már.

A harmadik problémakör az oksággal, korrelációval, elméletekkel, adatvezérelt tudománnyal kapcsolatos. Ennek a korlátait Kitchin is tárgyalta, egyoldalú látásmódját és a tudományos kutatásra gyakorolt hatását, amely a témaválasztáson (digitális adatban gazdag területek, részletkérdések) és a tudományfinanszírozáson keresztül is érvényesül. A szerzők számos alkalommal megismétlik, hogy új tudományos korszakba érkeztünk, ahol az adatok maguk beszélnek, nincs szükség elméletekre, oksági magyarázatokra és miértekre, a korreláció elegendő. „A korrelációk nemcsak azért hatékonyak, mert segítenek a megismerésben, hanem azért is, mert az így szerzett ismeretek viszonylag egyértelműek. Amikor azonban ismét az oksági viszonyokat helyezzük előtérbe, ez a tudás gyakran elhomályosul." (77. old.) Mindez nem azért hibás, mert tapasztalatilag sem alátámasztott (könyvtárnyi empirikus irodalma van a különböző eredetű látszólagos korrelációknak és a figyelmen kívül hagyott zavaró változók miatti látszólagos korrelálatlanságnak, az ökológiai tévkövetkeztetésnek, a Simpson-paradoxonnak, a módosítható egység problémájának), hanem mert elméletileg és módszertanilag is a tudomány előtti helyzetet hozná vissza. A tudományos kutatás lényegi sajátossága, hogy nemcsak feljegyzi a jelenségeket, hanem az azok közötti kapcsolatokra konzisztens elméleti rendszerekbe ágyazott oksági magyarázatokat kíván adni.

Az utóbbi két kérdéskör koncepcionálisan különösen fontos. A könyvnek vannak érdemei is, jó példái, részterületei. A leginkább problémás részek az első öt fejezetre koncentrálódnak, míg a további fejezetek kevésbé foglalkoznak fontosabb koncepcionális kérdéssel, ott a példák sokszor megfelelők. Itt esik szó a lehetséges veszélyekről, árnyoldalakról is, bár azt a hiányérzetet nem oldják fel, amely abból adódik, hogy nem foglalkoznak a Big Data megoldások költségeivel, elsősorban a megoldásokhoz szükséges jelentős emberi idővel. A magyar kiadás hátoldalán azt olvashatjuk a Big Data helytelen felhasználásának veszélyei között, hogy „sőt még az is előfordulhat, hogy valakit olyasmiért ítélnek el, amit el sem követett, egyszerűen azért, mert a big data képes előre jelezni a jövőbeli viselkedést". Szerencsére ilyet nem írnak a szerzők az adatvédelmi kérdéseket tárgyalva, hanem O'Neill könyvéhez hasonlóan mutatják be, hogy a bűnözésre való hajlamot előre jelző rendszerek használatakor fennáll az a lehetőség, hogy a magas kockázatú személyeket a bűnmegelőzés céljából zaklassák a hatóságok (194–196. old.). Olyat azonban nem írnak sehol, hogy ez alapján bárkit el lehetne ítélni.

Ismertetett kötetek

BAESENS, B. [2014]: *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley. Hoboken.

KITCHIN, R. [2014]: *The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage. London.

LONG, C. (exe. ed.) [2015]: *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley. Indianapolis.

O'NEIL, K. [2016]: *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy.* Penguin Books. London.

MAYER-SCHÖNBERGER, V. – CUKIER, K. [2014]: *Big data. Forradalmi módszer, amely megváltoztatja munkánkat, gondolkodásunkat és egész életünket*. HVG Könyvek. Budapest.