Szigeti Ákos[1]

A társadalomtudomány digitális átállása

https://doi.org/10.18030/socio.hu.2021.1.172

Matthew J. Salganik:
Bitről bitre. Társadalomkutatás a digitális korban

Salganik, M. J. (2018)
Bit by Bit. Social Research in the Digital Age
Princeton & Oxford: Princeton University Press

A társadalom folyamatos digitalizálódása egy olyan interdiszciplináris alapokra építkező kutatói megközelítésmódot igényel, mely képes integrálni a társadalomtudományok és a nagy adatmennyiség algoritmusok segítségével való feldolgozására szakosodott adattudomány *(data science)* ismereteit. *„A társadalomkutatók egy olyan átmenet részesei, ami a fotográfia és a mozgóképkészítés közötti váltáshoz hasonlítható”* – írja a társadalomtudomány napjainkban zajló digitalizálódásáról Matthew J. Salganik, a Princeton Egyetem szociológiaprofesszora, a számítógépes társadalomtudomány kiemelkedő alakja, számos nemzetközi élvonalba tartozó folyóiratban megjelent cikk szerzője. Publikációiban jellemzően a legújabb módszertani kérdéseket járja körbe: több ízben írt például a szociológiai kutatásban kritikus mintavételi eljárásokról, illetve a hálózatkutatás legújabb eszközeiről, így például a rejtett populációk becslésére használt ún. hálózat-felszorzó módszerről. *Bit by bit* című könyve is ebbe a keretbe illeszkedik, melyben a nagy adatmennyiséggel dolgozó, ún. *big data* módszerek alkalmazási kérdéseinek feldolgozásával kívánja elősegíteni a társadalomtudomány digitális átállását. Azt a folyamatot, melyet napjainkban a COVID-19 járvány okozta kijárási korlátozások még inkább felgyorsítottak: a társadalomkutatók is egyre inkább rákényszerülnek a digitális tér használatára, mind a kutatásaikhoz szükséges empirikus adatok összegyűjtésekor, mind eredményeik disszeminációjakor.

A könyv kifejezetten az adattudomány iránt érdeklődő társadalomkutatók számára íródott, olyan formában, mely lehetővé teszi az egyetemi kurzusokon való felhasználását is. Az oktatásban való alkalmazást segítendő, a szerző egyéni és csoportos feladatokat mellékelt a fejezetekhez, megadva azok nehézségi szintjét, és hogy szükség van-e hozzájuk adatgyűjtésre, illetve matematikai vagy éppen programozási ismeretekre. A kutatásokból vett példákkal is gazdagon illusztrált könyv nem mélyed el a technikai részletekben a feltétlenül szükséges mértéknél jobban, de további olvasnivalót nyújt azok számára, akik a matematikai-programozási háttérre is kíváncsiak. A *big data* alapjairól szóló bevezető és a kutatás jövőjéről szóló záró fejezet közötti öt tartalmi fejezet a különböző, nagy adatmennyiség gyűjtésére és elemzésére specializálódott megoldások, leginkább kurrensnek tekinthető módszerek szerint csoportosul. Külön fejezetet kapott az online viselkedés megfigyelése,

1 Szigeti Ákos doktorandusz, Nemzeti Közszolgálati Egyetem Rendészettudományi Doktori Iskola

a kérdőíves adatok és a *big data* források összekapcsolása, az online kísérletezés, a tömeges együttműködés és a *big data* etikai vonatkozásai. Salganik maga is kihasználta a tömeges együttműködésben rejlő lehetőségeket könyve írása során, amikor nyílt bírálati folyamat *(open review)* keretében adott lehetőséget a könyv írásába való becsatlakozásra, a kéziratra vonatkozó visszajelzések küldésére.

A bevezetőben a *big data* 10 jellemzőjét gyűjtötte össze a szerző, valóban sorra véve a legfontosabb tényezőket:

1. Nagy (mennyiségű), így lehetővé teszi a ritka események megfigyelését, például nyelvhelyességi szabályok változását a Google Books segítségével. Hasonlóképpen lehetőséget ad a heterogenitás vizsgálatára, például területi különbségek megfigyelésére adóbevallási adatok felhasználásával, ahogy segíti az apró különbségek detektálását is.
2. Az adatok termelődése és gyűjtése jellemzően folyamatos, ami lehetővé teszi váratlan események megfigyelését és valós idejű becslések létrehozását.
3. Nem reaktív, a vizsgált alanyok nem a kutatók érdeklődésére (kérdéseire) reagálnak, így az adatok összegyűjtése és elemzése nem befolyásolja az emberi viselkedést (legfeljebb a platform, lásd e felsorolás 8. pontját).
4. Sok esetben hiányos, nem tartalmazza az összes elemzéshez szükséges információt, például demográfiai adatokat, más internetes platformokon vagy digitális eszközökön zajló viselkedés jellemzőit stb.
5. Jogi, üzleti és etikai korlátok miatt sokszor elérhetetlen, főleg a vállalati és kormányzati adatokhoz való hozzájutás lehet nehézkes az akadémiai szférából érkező kutatók számára.
6. Sok *big data* forrás olyan nem reprezentatív mintáról tartalmaz adatokat, mely jól definiálható (például twitter felhasználók, akik a választásokkor politikai tartalmat tweetelnek), így az adott mintára vonatkozóan tudunk megállapításokat tenni, azonban ezeket ritkán tudjuk nagyobb populációra általánosítani (az említett esetben például az összes választópolgárra).
7. Bár alapvetően a *big data* longitudinális adatgyűjtés, hiszen – ahogy a felsorolás 2. pontjában láthattuk – jellemzően folyamatosan gyűjti az adatokat, ugyanakkor a populáció, az egyes platformok használata és a rendszerek folyamatosan változnak, ezáltal nehézzé téve a hosszú távú trendek megfigyelését.
8. Algoritmikailag befolyásolt, hiszen az online platformok algoritmusai a céljaik elérése érdekében befolyásolják a felhasználók viselkedését és így a kinyerhető adatokat is. A szerző által felvetett példa szerint a Facebook addig bátorítja a felhasználókat új ismerősök jelölésére, míg 20 ismerősük nem lesz.
9. Sokszor szemetet, spamet tartalmaz, melyeket akár ún. botok (bizonyos feladatokat algoritmus alapján elvégezni képest szoftverek) is létrehozhatnak, automata módon, ezáltal jelentősen befolyásolva az eredményeket. Az ilyen, botok által létrehozott adatok tisztítása nehéz és hosszadalmas lehet, kiszűrésük kulcsa inkább az adatok létrejöttének teljeskörű és mély megértésében rejlik.
10. Az összegyűjthető adatok sokszor érzékeny személyes adatokat tartalmaznak, főleg a vállalati és kormányzati platformok esetében (például egészségügyi adatokat), illetve azokat időnként a kutatási alanyok tudta nélkül használják fel. Ezekre az aggodalmakra a szerző a könyv dedikált kutatásetikai fejezetében reflektál.

Az analóg és a digitális kor közötti váltás Salganik szerint új lehetőségeket hoz a kérdőíves kutatásokban is. Szerinte a *big data* nemhogy csökkentené a kérdőíves kutatások jelentőségét, azokat inkább növelni fogja, például a *big data* források és kérdőívek adatainak összekapcsolásával. Új lehetőségeket hordoz magában továbbá a nem véletlenszerű mintavétel, egyre több kutatás támaszkodik például online panelekre, amik akár olyan új elemeket is tartalmazhatnak, mint a gamifikáció (játékossá tétel), interaktív elemekkel segítve a kutatási alanyok bevonásának sikerességét. Emellett az a tény, hogy a *big data* módszerek a viselkedés megfigyelését rendszerint a potenciálisan befolyásoló humán közreműködőt nélkülöző módon tudják megvalósítani, többletet adhat a kérdezésen alapuló módszerekkel szerzett adatok mellé, melyek esetében fennáll a kérdezőbiztos, illetve az interjúkészítő befolyásoló szerepe.

A nagy adatmennyiség összegyűjtése során lehetőség van a nagyvállalatokkal való együttműködésre, de azt a kutatók bizonyos esetekben, megfelelő engedélyekkel és technikai háttérrel maguk is elvégezhetik. Az internetes platformokon, illetve digitális eszközök bevonásával zajló kutatásokhoz nem feltétlenül szükséges olyan mennyiségű erőforrás, mint például egy országos mintán történő személyes lekérdezéshez. Salganik az ilyen környezetben zajló kutatásokkal kapcsolatban a 3 R szabályát mutatja be, melyek betartásával a későbbiekben tárgyalt kutatásetikai elveket is könnyebb érvényesíteni:

1. Pótlás *(replace):* a kutatóknak olyan lehetőségeket, természetesen jelen lévő helyzeteket érdemes keresniük, melyek pótolni tudják a nehezen kivitelezhető és rizikós kísérleteket. Például egy a szerző által hivatkozott kutatás szerint az emberek negatívabb tartalmakat posztolnak a közösségi médiában azokon a napokon, amikor esik az eső, így csupán az időjárás véletlenszerűségére támaszkodva lehetséges az üzenőfalakon megjelenő posztok variabilitásának bármilyen további befolyásolás nélküli vizsgálata.
2. Finomítás *(refine):* a kutatóknak úgy érdemes finomítani a beavatkozásaikat, hogy a lehető legkevesebb sérülést, kárt okozzák a kutatás alanyainak. Például a pozitív és negatív tartalmak blokkolása helyett a kutatók kiemelhetnek pozitív vagy negatív tartalmakat, ezzel bár befolyásolják a közösségi média üzenőfalain megjelenő tartalmakat, nem lehetetlenítik el egy adott tartalom elérését sem.
3. Csökkentés *(reduce):* a kutatóknak érdemes annyira csökkenteni a kísérletükbe bevont résztvevők számát, ahány résztvevőre mindenképpen szükségük van, ezzel minimalizálva a résztvevők sérülésének lehetőségét. Az analóg kutatásokban ez az erőforrásokkal való takarékoskodás miatt automatikusan megtörténik, a digitális kutatásokban azonban erre külön hangsúlyt kell helyezni.

Ahelyett, hogy csupán alacsony számú kutatóval, kutatási asszisztenssel együttműködve végeznénk kutatásainkat, a digitális korban már lehetőségünk van tömeges együttműködésben végezni a kutatások egyes fázisait. Salganik három csoportra osztja a tömeges együttműködés *(mass collaboration)* keretében létrejött kutatásokat:

1. Emberi számítás *(human computation):* ezekben a projektekben sok ember dolgozik egyszerű mikrofeladatokon, melyek összességével képtelenség lenne egyedül megbirkózni. Ilyen projekt volt például a Galaxy Zoo, melyben több mint százezer önkéntes kategorizált kilencszázezernél is több, galaxisokról készült fotót.
2. Nyílt felhívás *(open call):* az ilyen projektekben a kutatók egyszerűen ellenőrizhető megoldásokat kérnek a nyilvánosságtól adott problémákra, melyek közül végül kiválasztják a legmegfelelőbbet. Erre példa lehet a Netflix Prize, melyben a Netflix egymillió dollárt ajánlott fel annak, aki képes a korábbi osztályozó rendszerüknél, a Cinematchnél tíz százalékkal jobban megjósolni hárommillió olyan filmértékelést,

melyekkel a Netflix ugyan már rendelkezett, de azokat a fejlesztés érdekében szándékosan nem tette nyilvánossá.

3. Megosztott adatgyűjtés *(distributed data collection):* ezek olyan projektek, melyekben a bevont sokaság gyűjti az adatokat, jellemzően olyan mennyiségben, melyet képtelenség lenne önálló vagy akár csoportos munkában összegyűjteni. Megosztott adatgyűjtésen alapul például az eBird projekt, melyben önkéntesek fotózzák a madarakat és csatolják a megjelenésük koordinátáit, ezzel segítve az ornitológusok munkáját.

A tömeges együttműködésre épülő projektek azon túl, hogy választ adhatnak a szociológia empirikus válságára és ezzel a társadalomkutatás fejlődését szolgálják, a demokráciát is erősíthetik, hiszen az állampolgárok bevonásán alapulnak. Ahogy a Wikipedia egy nyílt, bárki által szerkeszthető enciklopédiaként megváltoztatta a tudás rendszerezésével kapcsolatos elképzeléseinket, úgy a kutatásokban alkalmazott tömeges együttműködés is teljesen megváltoztathatja a tudományos kutatás jellegét, megfelelően kiegészítve a klasszikus kutatási módszerek segítségével kapott eredményeket.

A digitális korban új etikai problémák merülnek fel a társadalomkutatásban. Salganik kiemeli a kutatók gyors tempóban növekvő hatalmát: az internetes platformokon, illetve a digitális eszközök segítségével lehetőségük lett arra, hogy a kutatási alanyok tudta és beleegyezése nélkül figyeljenek meg embereket és kísérletezzenek velük. Mindezt természetesen intézeti és a GDPR megjelenésével már európai szintű szabályozók is korlátozzák, megoldást keresve a szerző által is felvetett új problémákra. Salganik az etikus kutatások érdekében a korábban említett 3 R szabálya mellett a hagyományos kutatásokból ismert egyetemes alapelveket sorolja fel: a személyek tisztelete, az előnyösség, az igazságosság, valamint a törvények és a közérdek tiszteletben tartása. Három olyan területet emel ki, melyek különösen fontosak a *big data* források esetében:

1. A beleegyező nyilatkozat, amivel kapcsolatban a „mindenhez beleegyező nyilatkozat” *(informed consent for everything)* elvnél egy összetettebb elvet támogat, amit úgy hív, hogy „valamilyen beleegyezés a lehető legtöbb kutatáshoz” *(some form of consent for most research),* ezzel lehetővé téve az adott kutatási szituációhoz való alkalmazkodást, de figyelembe véve a fent már említett, személyek tisztelete melletti elveket is.

2. Az információs kockázat (az információ közzététele következtében bekövetkező kár potenciálja) a digitális korban emelkedett, napjainkban mindenről adatokat gyűjtenek. Ebben a helyzetben nehéz mérni vagy megjósolni a kockázatot, így célszerű feltételezni, hogy minden adat azonosító erejű (tehát nem anonim), illetve érzékeny. Emiatt hangsúlyozza a szerző, hogy a kockázat csökkentése érdekében a kutatóknak adatvédelmi tervet kell készíteniük és követniük a kutatás folyamán.

3. A titoktartás, amivel kapcsolatban érdemes túllendülni a megszokott magánélet/nyilvánosság különbségtételen és a tartalomhoz igazítani a normákat. A tartalom három elemét, melyek 1.) a szereplők (actors, melynek tagjai: tárgy, küldő, fogadó), 2.) a jellemzők (attributes, vagyis az információ típusai) és 3.) az átadás szabályrendszerei (transmission principles, vagyis az információáramlás korlátai), mindig az adott szituációban érdemes értékelni és így döntést hozni az adatok etikus felhasználásáról.

A záró fejezet a társadalomtudományi kutatás jövőjéről szól és három fontos ajánlást fogalmaz meg a társadalomkutatók számára:

1. Érdemes kihasználni a kutatási célból létrejövő és a készen elérhető adatforrások összekapcsolásában rejlő lehetőségeket, hiszen ezen a módon is többletinformációhoz juthatunk.
2. Érdemes megfontolni a résztvevőt (kutatási alanyt) középpontba helyező adatgyűjtés alkalmazását, szemben a kutatóközpontúval, mely csupán az adatok összegyűjtésére koncentrál és nem foglalkozik eléggé azzal, hogy a digitális korban a felhasználók figyelmének megszerzéséért olyan tartalmakkal, termékekkel kell megküzdeni, melyek külön hangsúlyt fektetnek a felhasználói élmény megteremtésére.
3. Az etikai kérdéseket komolyan kell venni és a kutatásetikai szempontokat is szükséges szerepeltetni a kutatási tervekben. Ugyanakkor, a társadalomkutatóknak és az adattudósoknak közös nevezőre kell jutniuk: közelíteni szükséges a társadalomtudomány szigorú szabályokon alapuló kutatásetikai megközelítésmódját és az adattudomány ad-hoc adatvédelmi szemléletét.

A fenti ajánlások mellett a könyv erénye, hogy több ízben utal az adatkapitalizmus fogalmára, mely szerint napjainkban a pénz szerepét olyan adatok veszik át, melyek felett bizonyos tech óriások, adatoligarcha cégek (mint például a Google, a Facebook vagy a kínai Baidu) rendelkeznek. Bár e cégek működését kutatóként kevéssé befolyásolhatjuk, a Salganik által felvetett kutatásetikai kérdések és szabályok elősegíthetik, hogy a társadalomtudomány etikus módon használja fel az adatokat, illetve észrevegye és képes legyen kezelni, ha egy akadémiai vagy akár piaci szereplő nem megfelelően jár el.

Salganik könyve végén megjegyzi, hogy a jövő társadalomkutatása a társadalomtudomány és az adattudomány kombinációja lesz. A tudományok hatékony együttműködéséhez azonban (és ezt már jelen recenzió szerzője teszi hozzá) szükség van arra, hogy a tudományok értsék egymás nyelvét, megteremtsék az elméleti és módszertani interoperabilitást és ezáltal képesek legyenek interdiszciplináris kutatási projektek elvégzésére. E közös nyelv megteremtésének, illetve a tudományágak között fordítani képes szakértők képzésének fontos mérföldköve lehet a bemutatott mű megjelenése, olvasásával a *big data* módszereket alkalmazni kívánó társadalomkutatók megindulhatnak egy olyan tanulási folyamatban, melynek során képessé válnak konkrét, a könyvben bemutatott adattudományi módszerek alkalmazására.