BOZSONYI KÁROLY

A Big Data forradalma és a jelentés ellenforradalma a társadalomtudományok módszertanában

Az ezredforduló óta kétségkívül a Big Data (BD) és a gépi tanulás (Machine Learning) a legnagyobb fordulat a társadalomtudományok módszertani eszköztárában.

Az alábbi rövid elemzésben egyrészt bemutatjuk az új paradigma nyújtotta lehetőségeket, másrészt rámutatunk a korlátaikra is. Tisztázzuk a területhez kapcsolódó fogalmakat, majd összehasonlítjuk a hagyományos és az új kutatási paradigmák sajátosságait. Megpróbálunk állást foglalni a komputációs szociológia és az adattudomány lehatárolásának kérdésében is.

(Mikor beszélhetünk Big Data adatmodellekről?) Nincs egységesen elfogadott definíció, de azonosíthatók bizonyos tulajdonságok – így a *méret,* a *strukturáltság,* az *integráció* vagy az *adatgeneráló folyamat* –, melyek valamilyen kombinációban fel-felbukkannak a Big Data különböző definícióiban.

Általában nagy méretű, masszív adatstruktúrákról beszélünk – a gigabájt, sőt terrabájt méretűek a jellemzők, több millió (akár több 100 millió) rekord, több száz (több ezer) feature (változó, mező, dimenzió) jellemzi a tipikus Big Data adatstruktúrákat.[1] Ugyanakkor nem biztos, hogy egy Big Data adatrendszer szükségképpen hatalmas méretű volna, figyelembe kell venni egyéb szempontokat is.

Jellemzők a hibrid adatstruktúrák, alacsonytól a magas strukturáltsági szintekig szóródva. Egy BD-adatstruktúrában egyszerre lehetnek jelen verbális corpusok, képek, mozgóképek, hangfájlok, pszeudo nyelvi elemek (emotikonok, ikonok stb.), koordináták, szenzoradatok és strukturált kódolású adatok. Ugyancsak jellemző a különböző adatforrásokból származó adatok integrálása – gyakran több száz vagy ezer forrás integrálódik egyetlen adatbázisba. A tapasztalat szerint a strukturálatlan adatok túlsúlyban vannak a strukturált adatokhoz képest.

Eltérően a tudományos célra, tervszerűen gyűjtött, survey adatokat tartalmazó adatrendszerektől, a humán/társadalmi viselkedéssel kapcsolatos BD-adatbázisok tényleges tranzakciós tevékenységek digitális nyomait tartalmazzák. Ezek az adatrendszerek úgy épülnek fel, hogy az alkalmazott adatmodellek és technológiák nem az adatok strukturált tárolására, hanem a felhasználói tevékenység kiszolgálására vannak optimalizálva. Mára a társadalmi cselekvés szinte elképzelhetetlen anélkül, hogy a cselekvés hatására digitális nyomok ne keletkezzenek. A telekommunikáció, a pénzügyek, a Social Media, azaz közösségi média, a webböngészés intenzív digitális naplózást generál. Ezek az adatbázisok kincsesbányái a kutatásnak azokban az országokban, ahol a kutatók számára az adatokhoz való hozzáférést nem teszik lehetetlenné a jogszabályok. Ezek az adatrendszerek elképesztő részletességgel mutatják meg azt, hogy különböző társadalmi kontextusokban mit csinálnak az emberek, de a viselkedés mögött húzódó motivációk megértéséhez általában csekély támpontot adnak.

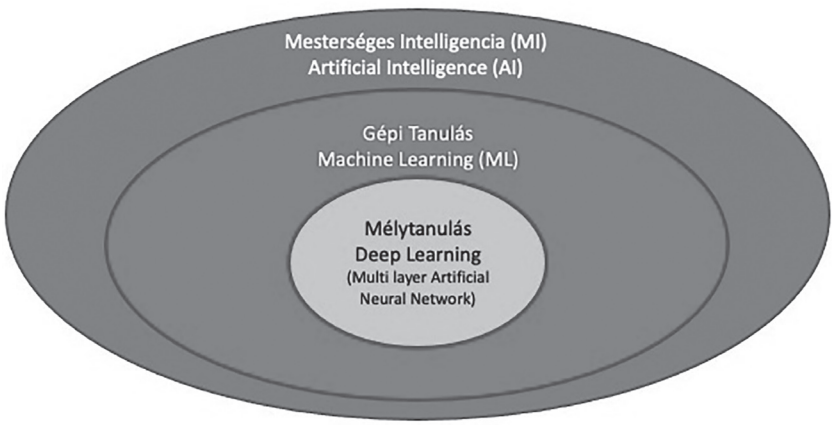
(Miért indokolt kutatási/tudományos szempontból a BD alkalmazása?) Az ipar intenzíven használja a BD-megoldásokat, mert ez üzleti előnyt jelent a számára. Ugyanakkor látni kell azt is, hogy a tudományos kutatás sem térhet ki ez elől a kihívás (lehetőség) elől.

A klasszikus survey adatfelvételen alapuló elemzések generalizálhatósága és validitása egyre inkább megkérdőjeleződik. Világszerte növekszik a válaszmegtagadás, ma már nem ritka, hogy 10–15 megkeresésből jön össze egy kérdőív. Ha azonban csak minden 10. megkérdezett hajlandó válaszolni a mintában, akkor vajon mire reprezentatív a minta? A kérdőívek hosszával kapcsolatos válaszadói igények is drasztikusan változnak. A fiatal generációk nem hajlandók 10–15 percnél több idejüket rabló kérdőívet kitölteni. Vajon mennyire érvényes az a kérdőív, amelyet 60 percig kérdeztünk egy tinédzsertől, aki már a 10. perc után unta az egészet?

A tranzakciós adatbázisokon alapuló BD-rendszerek ugyanakkor teljes körűek, ezért nincs reprezentativitás kérdés, sőt a nagyon kis elemszámú, reprezentatív mintával egyébként elérhetetlen társadalmi csoportok is elemezhetővé válnak. Ugyanakkor itt nem a kutató tematizálja a kutatási kérdéseket, hanem a rendelkezésre álló adatokhoz kell illeszteni a kérdéseinket.

(Soha nem látott számítási kapacitások) Az informatikai adatokat feldolgozó rendszerek sebessége és kapacitása az elmúlt ötven évben exponenciális sebességgel növekedett. Ennek következtében ma egy átlagos notebook számítógép számítási kapacitása meghaladja a 10–15 évvel ezelőtti szuperszámítógépekét. Ráadásul a szuperkomputing a felhőalapú IT-szolgáltatások elterjedésével gyakorlatilag üzleti alapra helyeződött és jó néhány[2] nemzetközi szolgáltatónál lehet néhány dollár/perc alapon szuperszámítási kapacitást bérelni. Ez a komputációs forradalom olyan matematikai, módszertani lehetőségeket nyit meg, amelyek lehetővé teszik a BD-adatrendszerek információtartalmának maximális kiaknázását.

Az új módszertani paradigma kulcsfogalmai a *Mesterséges Intelligencia* (Artificial Intelligence), a *Gépi tanulás* (Machine Learning) és a *Mesterséges Neurális Hálók*kal (Artificial Neural Network) megvalósított *Mélytanulás* (Deep Learning). Ábránkon e fogalmak, modellosztályok egymáshoz való viszonyát mutatjuk be.



A modern adatelemzési modellek egymáshoz való viszonya

Az új modellek komputációsan nagyon hatékonyak, tehát nagy adatrendszereken is gyorsan eredményt képesek elérni. Ugyanakkor az eredményeket létrehozó modellek interpretációja gyakran nehézkes vagy egyenesen lehetetlen. Egy neurális háló működése mögött

nincsen kevés számú, de jól interpretálható regressziós paraméter, mint egy lineáris modell esetében, van viszont rengeteg nem vagy nehezen interpretálható súlyozott paraméter, ami nem teszi lehetővé a modellek működésének hagyományos értelemben vett megértését.

(Nagy számítási kapacitás hatásai a módszertani lehetőségekre) Az olcsón bárki által elérhető nagy számítási kapacitások nemcsak új, korábban nem látott modellek alkalmazását teszik lehetővé, hanem régóta, akár évtizedek vagy évszázadok óta ismert elméleti megközelítések gyakorlati alkalmazására nyílik lehetőség. Gondoljunk arra, hogy a Bayes-statisztika a frekventista (hagyományos) statisztikai megközelítésekkel szemben az elmúlt kétszáz évben nem a matematikai/elméleti alacsonyabbrendűsége, hanem a megvalósíthatatlan komputációs költségei miatt szorult háttérbe. Az utóbbi egy-két évtizedben a kvantitatív kutatók standard eszközeivé váltak a *resampling technikák* (Good 2006), mint amilyen a Bootstrapp, a Jacknife, a Splithalf stb., s ezzel lehetővé vált az aszimptotikus módszertanoktól való elszakadás azokban az esetekben, amikor az adatok nem felelnek meg az aszimptotikus elméletek alkalmazásánál megkövetelt feltételeknek.

Elméleti jelentőségüknek és gyakorlati hasznuknak megfelelő helyet kapnak a Bayes-statisztikán alapuló modellek (Congdon 2006, McElreath 2020), mert széles körű alkalmazásuknak már nem képezik akadályát a számítási korlátok.

Rengeteg térben megfigyelt, területileg aggregált adat van a gazdasági-társadalmi statisztikai rendszerekben, s e térbeli adatok matematikailag korrekt kezelését a fölöttébb számításigényes tér-ökonometriai (spatial oeconometrics) modellek jelentik (Golgher and Voss 2016). Ezek az elemzések új lehetőséget teremtenek releváns társadalomtudományi és policy elemzések elvégzésére egyaránt.

A szinte korlátlan méretű digitális nyelvi korpuszok és hatalmas számítási teljesítmények tették lehetővé a természetes nyelvi feldolgozás (Natural Language Processing; NLP) modelljeit. Ezen eszközök a közösségimédia-platformokon[3] zajló nyilvános diskurzus alapján számtalan releváns társadalomtudományi elemzést tesznek lehetővé a társadalomtudományok legkülönbözőbb ágaiban (Koltai et al. 2021). Ugyanakkor a szerkesztett tartalmak (sajtó, online média) elemzése új izgalmas alternatíváját jelentik a hagyományos tartalomelemzési (Quantitative Content Analysis) módszertanoknak. A hagyományos tartalomelemzés mindig a kutató kódrendszeréhez igazodva tárja fel a verbális korpusz tartalmi jellegzetességeit. Ellenben a természetes nyelvi feldolgozás eszköztára [Topic modellezés, Látens Szemantikai Analízis (LSA), Látens Dirichlet Allokáció (LDA)] lehetővé teszi, hogy a korpusz belső logikája alapján, előzetes kódutasítások nélkül kerüljön kinyerésre (extrachálásra) a tartalmi összefüggések rendszere.

A komputációs robbanás a hálózatelemzés paradigmájában is forradalmi változást hozott. Az évtizedeken át uralkodó egocentrikus paradigmát és a szociometriát felválthatta a teljes hálózat explorációján alapuló kvantitatív hálózatkutatás. A gyors algoritmusok akár több százezer vagy millió csúcspontból (node-ból, vertexből) álló hálózatok elemzését is támogatják. Az emberek kapcsolatait pedig elemezhetővé naplózzák a közösségimédia-platformok és a telekommunikációs adatbázisok. Felmerül a kérdés, hogy mi a különbség a néhány tucat ember kapcsolatrendszerét elemző szociometria és a néhány ezer (vagy még több) ember hálózatát elemző hálózatelemzés (Social Network Analysis) között. A válasz egyszerű. A méret. A helyzet az, hogy kis méretekben, néhány tucat csúcsig minden véletlen hálózatra jól illeszkedik az Erdős–Rényi-féle (ER) véletlengráf-modell. Ezért a kis véletlen hálózatok mind hasonlítanak egymásra, nincs érdekes topológia, az egyes emberek hálózati pozíciója az érdekes kutatási kérdés. A nagy hálózatokban már megjelennek különböző jelentős eltérések az ER-modelltől és ettől

önmagában egy hálózat nagy skálájú topológiájának felderítése is érdekes kéréssé válik. Valahogy úgy kell ezt elképzelni, hogy egy teniszpálya mindig hasonlít egy másik teniszpályára, függetlenül attól, hogy az egyik egy hatalmas gömb felületén, a másik pedig egy hatalmas kürtőskalács felületén helyezkedik el. Kis skálán nem látszanak a nagy skálán jelentkező topologikus különbségek. A hálózatok topológiája azért fontos, mert alapvetően befolyásolja a hálózatok stabilitását, valamint a hálózatokon zajló terjedési folyamatok dinamikáját[4]. A nagy hálózatokban azonban nem csak a topologikus sajátosságok feltárása izgalmas, hanem érdemes a hálózatban „közösségeket" keresni, ami a gráf particionálásának végrehajtását jelenti. A nagy hálózatok esetében a hálózatok vizuális megjelenítése is fontos kihívás.

(A régi és az új empírikus kutatási paradigma összehasonlítása) Hagyományos paradigma*:* a konfirmatív és az exploratív kutatási stratégia élesen elválik. Az előzetesen megfogalmazott hipotézisek szerepe nagy. A kutatás eredményeinek értelmezése magától értetődő, hiszen a kutatási protokoll általában úgy van megalkotva, hogy az eredmények interpretálhatók legyenek. A fogalomhasználatban a kutatók által használt fogalmak reflektálnak a tudományos és társadalmi jelentésük esetleges különbségeire.

Új paradigma, azaz a Big Data/Data Science: az előre megfogalmazott kutatási hipotézisek mellett (gyakran akár helyettük is) a kutatásvezető protokollok gyakran heurisztikák vagy más exploratív kutatási stratégiák. Jellemző a Mesterséges intelligencia (Artificial Intelligence) és Gépi tanulás (Machine Learning) megoldásokkal támogatott elemző eszközök intenzív használata. Mivel a kutatás keretei a rendelkezésre álló adatmodellek által behatároltak, gyakran a szignifikáns eredmények birtokában is interpretációs problémák merülnek fel. Az interpretációs nehézségeknek tipikusan két forrása van. Egyrészt az adatok tranzakciós viselkedések alapján generálódnak és a társadalmi cselekvők viselkedése mögötti kognitív, emocionális, fogalmi struktúrákról nincsenek direkt visszacsatolások az adatokban. Másrészt a sikeres modellek gyakran nagy prediktív erővel bírnak ugyan, de az alkalmazott elemzési módszertan (pl. mélytanulás) miatt nincs mód a működésük megértésére.

		Megfigyelések (adatok) mennyisége	
		korlátozott	**nagyon nagy**
Adatok strukturáltsága	**magas**	Survey paradigma	Gépi tanulás, Mesterséges Intelligencia, Mélytanulás
	alacsony	Kvalitatív módszertanok és tartalomelemzés	Természetes nyelvi feldolgozás Mesterséges Intelligencia

A két kutatási paradigma egymáshoz való viszonya

(Kvantitatív és kvalitatív módszertanok viszonya) A hagyományos kutatási logikában a kvalitatív és kvantitatív módszerek két külön világ polgárai, és ha néha találkoznak, akkor tipikusan a kvalitatív módszereknek a kvantitatív vizsgálatok tervezésében,

előkészítésében van szerepe. Paradox módon a Big Data korszakában felértékelődik a kvalitatív vizsgálatok szerepe. És megváltozik a kutatási folyamatban elfoglalt helyük is. A korábban már említett interpretációs nehézségek miatt, a kvalitatív elemzés a kvantitatív vizsgálatok után válik fontossá és érdekessé, mert lehetővé teszi a szignifikáns kvantitatív eredmények megértését, interpretációját. Véleményem szerint egyenesen kötelező bizonyos BD-módszertanok használata során a talált összefüggéseket kvalitatív eszközökkel validálni.

(A Data Science társadalomképe) Az adattudomány módszertana felől nézve, a társadalom működése során folyamatosan előáll egy adathalmaz, ami elemzési lehetőségek tág terét nyitja meg. Technikai szempontból a társadalmi cselekvők nem mások, mint részlegesen megfigyelt adatgeneráló folyamatok (Data Generating Process), a társadalom pedig nem más, mint ezen adatgeneráló folyamatok bizonyos dimenziók mentén történő realizációjának varianca–kovariancia mátrixa, melyet bizonyos időpontokban meg tudunk figyelni. Ebben a világképben a szociológiai kutatás már nem több mint kovariancia mátrix látens struktúrákra való faktorizálása és/vagy az adatgeneráló folyamatokhoz kapcsolódó megfigyelési egységek halmazának particionálása.

Ez a szemlélet két irányból is problémákat vet fel. Egyrészt az adatbázisok determinálják, hogy mit látunk a világból és mit nem. Az adatsiló határain belül lehet, hogy működnek bizonyos modellek, de soha nem zárhatjuk ki, hogy a megtalált összefüggések csak korlátozottan érvényesek. Ez azonban a kisebb probléma. A nagyobb baj az, hogy a technikai adatelemzés általában nem tud számot vetni a társadalmi lét jelentéstelítettségével. Képzeljük el, hogy több százezer szenzor elhelyezésével beműszerezünk egy futballstadiont. Mindent mérünk és a hatalmas adatbázist elemezzük. Képesek leszünk az adatbázisból megérteni azt, hogy mit jelent a gól? Megértjük, hogy a szurkolók mikor és miért ordítanak, és mikor tapsolnak? Megértjük, hogy a Fradi szurkolók és az Újpest tábor mit gondol egymásról? Semmit sem fogunk megtudni a futball társadalmi beágyazódásáról, funkciójáról és a jelentőségéről sem.

(A jelentés ellenforradalma) Az ember fogalomalkotó és jelentés-vezérelt lény. A világról alkotott interpretációink és a hozzájuk kapcsolódó narratívák alapvetően meghatározzák viselkedésünket. A jelentésteliség feltárása nélkül készíthetünk hasznos előrejelző modelleket a társadalmi-emberi viselkedésről, de ezek a modellek valójában veszélyes babonákká válnak, ha nem vagyunk képesek interpretálni és megérteni működésüket.

Készíthetünk üzletileg nagyon hasznosan működő ajánlórendszereket is. Az ajánlórendszerek alapvető szerepet töltenek be a bőség gazdaságtanában, hiszen korunk embere technikai támogatás nélkül összeomlana a választás lehetséges alternatíváinak tömege alatt. Üzleti-technikai alkalmazásokban tehát valóban nagyon hasznosak a modern adatelemzésen alapuló BD-rendszerek.

Tudományos kutatás céljaira alkalmazva őket azonban azt kell mondanunk, hogy a nagy adatbázisokon futtatott modern elemzési technikáink ugyan gyakran találnak szignifikáns eredményeket az adatok között, de ezeknek az összefüggéseknek a jelentés nélkülisége, és hogy nem interpretálhatók, az egyrészt magában rejti a tévedés lehetőségét, másrészt nem támogatják az emberi társadalom működésének megértését. Ezen okokból a kvalitatív módszertanok alkalmazása felértékelődik, bár helyük a kutatási folyamatban – mint azt említettük –megváltozik.

A hagyományos kutatási paradigmában általában a kvalitatív módszertanok exploratív eredményeit konfirmáltuk kvantitatív vizsgálatokkal, a Big Data korában viszont a masszív adatrendszereket exploráló elaborált matematikai módszertanok eredményeit kell utólag jelentéstelivé tenni alkalmasan megválasztott kvalitatív technikák alkalmazásával. És pontosan ez a társadalmi jelentésre való reflexió választja el egymástól a komputációs szociológiát (társadalomtudományt) és a pusztán technikai elemzésre fókuszáló adattudományt.
A komputációs szociológiára nagy lehetőségek várnak és joggal várjuk, hogy alapvető eredményekkel fog hozzájárulni a társadalom működésének lényeges megértéséhez.

IRODALOM

Good, P. I.: *Resampling Methods.* Boston, 2006, Birkhäuser MA, https://doi.org/10.1007/0-8176-4444-X
Congdon, P.: *Bayesian Statistical Modeling.* Chichester, 2006, John Wiley and Son.
McElreath, R.: *Statistical Rethinking, A Bayesian Course with Examples.* in R and STAN, Taylor and Francis Ltd, 2020, DOI https://doi.org/10.1201/9780429029608
Golgher, A. B. – Voss, P.R: How to Interpret the Coefficients of Spatial Models: Spillovers, Direct and Indirect Effects. *Spat Demogr* 4, 175–205. 2016. https://doi.org/10.1007/s40980-015-0016-y
Koltai, J. – Kmetty, Z. – Bozsonyi, K.: From Durkheim to Machine Learning: Finding the Relevant Sociological Content in Depression and Suicide-Related Social Media Discourses. In Rudas, T. – Péli, G. (eds): *Pathways Between Social Science and Computational Social Science. Computational Social Sciences.* 2021, Springer, Cham. https://doi.org/10.1007/978-3-030-54936-7_11

JEGYZETEK

1 Például a US Army által működtetett globális Argus felderítő rendszer 40 gigabájt információt gyűjt másodpercenként. A kamerák és szenzorok 6000 terabájt képet és adatot küldenek az elemző központba naponta.
2 Például az IBM Supercomputing vagy az Amazon Web Services (AWS) High Performance Computing szolgáltatása, továbbá a Microsoft Azure és Google Cloud Platform szolgáltatása, ami magában foglalja a Tensor Flow nevű DeepLearning neurális háló számítási környezetet is
3 A Twitter és a Facebook különösen alkalmas erre az elemzési módszertanra.
4 Ha az emberek közti kapcsolatok ER-gráf szerint működnének, akkor nem lennének ennyire veszélyesek a járványok, ugyanis egyetlen pontról indítva nem lehetne végigfertőzni a teljes hálózatot (legalábbis az emberi társadalmakra jellemző alacsony sűrűségű véletlen gráfokkal modellezhető hálózatokat nem).