VAN ÚJ A BIG DATA ALATT?
NAPJAINK LEGNÉPSZERŰBB IT-S HÍVÓSZAVA CSAK EGY LEHETŐSÉG ARRA, HOGY AZ INFORMATIKUSOK ELADJÁK MAGUKAT, VAGY VALÓDI ÚJDONSÁG?

Csicsman József (matematikus, cégvezető, Új Calculus Számítógép-alkalmazási Bt.)

ÖSSZEFOGLALÓ

A digitális technológia fejlődése előre nem látható méretű adatok tárolását és elemzését teszi lehetővé. Rohamosan nő a keletkező adatok mennyisége, és egyre inkább csökken azok feldolgozottsági szintje. Míg a hagyományos összeírás (survey) típusú adatállományok elemzési lehetőségei is folyamatosan nőnek, mindenki csak az automatikusan keletkező adatállományok, a Big Data feldolgozásáról beszél.

Igazi projektet indítani, pályázatot nyerni csak ennek a szónak, illetve a hozzá szorosan kapcsolódó Data Science fogalomnak a használatával lehet.

Nagyméretű adatállományok feldolgozásában szerzett, sokéves gyakorlat után írásomban kísérletet teszek az új ismeretek összefoglalására és a fogalmak közérthetővé tételére. Már megvalósult gyakorlati példák bemutatásával ismertetem a lehetőségek sokaságát. Olvasható lesz néhány gondolat arról, hogy a hagyományos többváltozós elemzési módszerek hogyan értelmezhetőek Big Data környezetben, illetve egy valós biztosítási probléma felvázolásával mutatom be, hogy a probléma már jelen van a biztosítási szakmában is.

SUMMARY

The revolution of digital technology provides the opportunity to store and analyse huge datasets, and we have to face the problem, that the amount of data grows rapidly while the level of data processing decrease. Since the analysis of ordinary Survey datasets became more popular, everyone focuses on handling large amount of automatically generated "Big Data".

The fact is, that the chances are minimised to vin a tender, or even launching a project in this specific area, unless Big Data being mentioned.

Throughout the years of gaining experience on analysing large datasets, I will attempt to summarize the knowledge and new ideas about the relevant issues, which I will try to make more understandable through actual practical examples. You may read thoughts and ideas about how Multivariate Analysis methods can be applied in Big Data environment, and I also would like to present you a real problem about how does it affect the Insurance sector.

Kulcsszavak: összeírás típusú adatok, adatelemzési technológiák, automatikusan keletkező adatok, közösségi médiák, szenzoros adatgyűjtések, Big Data, Data Science
Keywords: Survey Datasets, Data Analysis Technologies, Automatically Generated Data, Social Media, Sensor Data Collection, Big Data, Data Science

JEL: G22, O33

DOI: 10.18530/BK.2018.2.64
http://dx.doi.org/1018530/BK.2018.2.64

Bevezetés

Természetesen a feltett kérdésre nincs kompetenciám válaszolni, de a több mint 4 évtizedes, nagy adatállományok feldolgozásával és elemzésével eltöltött szakmai múltam tapasztalatai az alábbi cikk gondolataihoz vezettek.

Az elektronikus számítógépek megkonstruálása idején is azért volt szükség az új technológiára, mert nagyméretű adatokon igen nagy számításigényű feladatokat kellett végrehajtani, amit emberi erővel már nagyon költséges volt megoldani. Mondhatjuk tehát, hogy az első lőelem táblák (a tüzéreknek kiadott táblázatok, az ágyúk beállításáról a célzáshoz) elkészítésekor is Big Data feladatot oldottak meg a fejlesztők.

A nagyméretű adatállományok és azok speciális adatfeldolgozási technológiája az elmúlt ötven év folyamatosan fejlődő feladatai a pénzintézeti szektorban. A biztosítók sem működhetnének, ha nem lenne elegendően nagy az ügyfelek száma, és a megnövekedett ügyfélszámhoz nem tartozna megfelelő adminisztratív adatkezelő, az operatív munkát és elemzési feladatokat támogató szoftverkörnyezet. Mégis mi történt a 2000-es évek elején, hogy a Big Data fogalom bevezetésével sikerült felhívni a felhasználók figyelmét az adatelemzési feladatok fontosságára?

A technológiák fejlődése egyre nagyobb méretű adatok feldolgozását teszi lehetővé, napjainkra már nem az adat mérete, hanem a nagyméretű adatok értelmes feldolgozása jelent problémát. Tehát amikor a Big Data feladatról beszélünk, akkor elsősorban technológiai kihívásra kell gondolnunk.

Van-e, lesz-e és alkalmazható-e megfelelő algoritmus nagyméretű adatállományaink feldolgozásához akkor, amikor az adatok növekedése exponenciális, és egyre kevesebb részét dolgozzuk fel meglevő adatainknak.

Számomra az a legnagyobb újdonság a Big Datában – szemben az összeírás típusú hagyományos adatok feldolgozásával –, hogy ebben az esetben tőlünk függetlenül, előzetes specifikációk nélkül, sokszor automatikusan keletkeznek adatok, és új technológiákat kell kidolgozni az adatok kezelésére, értelmezésére és az adatból szerzett információk kigyűjtésére.

Az IBM szakértői becslés alapján arra jutottak, hogy napjainkban kétévente megduplázódik az összes adatmennyiség, vagyis huszonnégy hónap alatt annyi adat termelődik, mint a történelemben előtte összesen.

Egyértelmű tehát, hogy az információs és kommunikációs technológia fejlődésének következtében hatalmas mennyiségű adat jön létre, amelyek kihasználatlansága az adattudósok számára pazarlásnak tűnik.

A Big Data nóvuma azonban nemcsak az adatok számosságában rejlik, hanem elsősorban a közösségi média és a mobiltelefonok szolgáltatásainak széles körű terjedése miatt azok változó természetében is. Igen sok technológiai lehetőség is adódik a különféle szenzorok (automatikus érzékelők, melyek digitalizált mérési eredményeket rögzítenek) révén keletkezett adatokkal is. Az így keletkező adatok kezelése és értelmezése is a hagyományos adatkezeléstől merőben eltérő, új feladat.

Noha napjaink technikai fejlettsége egyre inkább lehetővé teszi e hatalmas adatmennyiség összegyűjtését, feldolgozását, tárolását és rendszerezését, egy hagyományos adatelemző számára mégis nehézséget jelent a módszertanok alkalmazása nagyméretű adathalmazon. A témakör bemutatása után megkísérlem összefoglalni, hogy az adatelemzési módszertanoknál milyen változásokat hozhat a Big Data.

Írásom végén biztosítós mintafeladatokon keresztül vázolom, hogy az új technológia milyen új technológiai feladatokat és lehetőségeket teremt.

Ki tud több V-t kitalálni?

Amikor Big Datáról hallunk vagy olvasunk, a téma marketingesei a V betűket használják.

A Big Data definíciója az Oxford-szótárak szerint: „Extrém nagy adathalmazok, amelyek számításigényes analizálása során mintázatokat, trendeket és összefüggéseket lehet feltárni különösen az emberi viselkedés és interakciók terén."

A legtöbb számítógép-alkalmazó számára ismert MS EXCEL-ben már a magyarországi népesség adatainak kezelése is Big Data probléma.

A Wikipédián a következő olvasható: „A Big Data olyan nagy és komplex adathalmazok összessége, amelyek kezelése hagyományos adatbázis-kezelő eszközökkel nem lehetséges."

A legtöbb számítógép-alkalmazó számára ismert MS EXCEL-ben már a magyarországi népesség adatainak kezelése is Big Data probléma.

Klasszikusan tehát a következő három fogalommal jellemezhető a Big Data (ezt az angol elnevezések kezdőbetűit alkalmazva 3V-definíciónak is szokták nevezni):

1. ***Mennyiség*** *(volume).* Nehéz meghatározni, hogy mennyire nagy ez az adatmennyiség, abban azonban mindenki egyetért, hogy amit ma soknak tartunk, az holnapra még több lesz.
2. ***Változatosság*** *(variety).* A Big Data-állományok típusukat, strukturáltságukat tekintve nagyon különbözőek, és számos forrásból származnak.
   a. Ebbe az adatkörbe tartoznak a szenzorok által érzékelt és az okoseszközök adatai, illetve a közösségi hálózatok által generált „lenyomatok", vagyis minden olyan információ, amely valamilyen emberi tevékenység vagy eszköz által nyomot hagy az interneten (számítógépeken).
   b. Ilyenek például az SMS-ek, a tweetek, a hipertextek, a geolokalizációs információk, az audio- és videofájlok, a klikkek, a log fájlok, a tranzakciók és az érzékelők adatai stb.
3. ***Sebesség*** *(velocity). Groves* [2013][1] megfogalmazásával élve a Big Data élő adat, szemben a survey-típusú felvételek tervezett adataival.[2] A hatalmas adatállományok létrejöttének sebessége elsősorban az adatok „élő" jellege miatt növekszik, hiszen a folyamatosan keletkező adatok szüntelenül áramlanak. Ezzel párhuzamosan gyorsul feldolgozásuk és értelmezésük sebessége is.

A 3V-definíción túl a szakirodalom említést tesz más (ugyancsak V betűvel kezdődő) jellemzőkről is, amelyek közül a hivatalos statisztika szempontjából az adatok **valóságtartalma** (veracity) kiemelt fontosságú. E kifejezés arra utal, hogy az adatok mennyire jó minőségűek, milyen mértékben tükrözik a valóságot.

A Big Data definiálásának szempontjából további fontos jellemzők még: a ***változékonyság*** (variability), a ***megjelenítés, vizualizáció*** (visualization), az *értékes,* ***felhasználható eredmény*** (value), az ***érvényesség*** (validity), valamint az *illékonyság,* azaz az ***érvényesség hossza*** (volatility).

A jó marketinges ma már akár 27 V-t is fel tud sorolni, igaz, nem minden esetben sikerül tartalmat is rendelni a V betűhöz.

A Big Data-taxonómia

A Big Data-forrásokat több rendszerező elv szerint csoportosíthatjuk. Az adatok keletkezésük szerint három nagy csoportba sorolhatók.[3]

– Az *emberi eredetű adatok* kategóriája az emberi tapasztalatok szubjektív rekordjait takarja, amelyeket korábban könyvek, művészeti alkotások, majd fotók, videók és audioeszközök tároltak, és amelyek napjainkban csaknem mindig digitálisan (személyi számítógépeken, a közösségi hálón) keletkeznek. E típusba tartoznak a Facebook-kommentek, a lájkok és a posztok, a tweetek, a blogok, a vlogok, a személyes dokumentumok, a közösségi képmegosztókra (Pinterestre, Instagramra, Youtube-ra) feltett képek, videók, az interneten lefuttatott keresések, a mobiltelefonon küldött üzenetek és az e-mailek is.

– A *folyamateredetű adatok* közé a különböző (elsősorban az üzleti) folyamatok során keletkező adatokat soroljuk. Ezek jól strukturált, jellemzően RDBMS (relational database management system – relációs adatbázis-kezelő rendszer) adatok vagy metaadatok. Egy típusukat a nyilvántartások adatai alkotják, melyek tipikusan állami intézmények (például a közhivatalok) által fenntartott források adatai, de idetartoznak az elektronikus

egészségügyi nyilvántartások, az orvosi rekordok, a kórházi látogatások nyilvántartása, a biztosítási nyilvántartások, a banki vagy részvényadatok, a vállalkozások üzleti adatai is (ha az utóbbiakról nyilvántartás vezetését jogszabály írja elő). A folyamateredetű adatok másik csoportját a tranzakciós adatok adják; ezek közös jellemzője, hogy két entitás közötti tranzakcióból származnak. Ilyenek például a kereskedelmi tranzakciók (például az internetes vásárlások), a bank- és hitelkártya-tranzakciók, valamint az e-kereskedelem adatai (ideértve a mobilkészülékről indított tranzakciókat is) stb.

– A *gépek által előállított adatokat* klasszikusan a hangzatos Internet of Things (a dolgok internete) néven emlegetik. Idetartoznak a fix és mozgó szenzoros adatok, valamint a log fájlok. Definícióját tekintve a szenzoros adatokból származó információköteg nem más, mint a fizikai világ eseményeit rögzítő és mérő érzékelők milliárdjainak adatai. Ahogy egyre több érzékelő kerül a világban bevezetésre és aktiválásra, úgy nő az ilyen jellegű adatok volumene is. Mindent összevetve, ennek az adattípusnak a mennyisége növekszik a leggyorsabban. Szenzoros adatoknak tekinthetjük például a háztartási eszközök érzékelőinek, az időjárás- vagy a légszennyezettség-érzékelőknek, a műholdképeknek, a forgalomfigyelőknek/webkameráknak az adatait; a nyomkövető eszközös adatok közé pedig például a mobiltelefonok útvonal-/követési és a földrajzi helyzetre vonatkozó (például GPS) adatok sorolhatók. A log fájlok a számítógépek működése során, szöveges (text) formában létrehozott, rendszereseményekről szóló ún. naplóbejegyzések.

Adatgyűjtés és a Big Data[4]

Hagyományos megközelítés, avagy a top-down paradigma. A hivatalos statisztika általános gyakorlata szerint egy adatfelvétel előtt elsőként azt kell meghatározni, hogy milyen információkra van szükségünk, és ehhez előre kigondolt lekérdezéseket fogalmazunk meg.

Majd a következő lépéseket hajtjuk végre:

1. adatgyűjtés-tervezés,
2. adatgyűjtés,
3. adat-előkészítés,
4. adatelemzés,
5. információkinyerés az adatbázisból/a felállított hipotézis igazolása vagy cáfolata.

A top-down paradigma lényege, hogy az adatgyűjtés megtervezése során az elemzési cél(ok) meghatározásán van a hangsúly.

Big Data-megközelítés, avagy a bottom-up paradigma. A Big Data-paradigma esetében az előzőhöz képest egészen más logikát kell követnünk. Mivel itt nincs szükség az adatgyűjtés tradicionális értelemben vett megtervezésére (hiszen az adatok már megvannak, pontosabban mindenütt ott vannak), felborul a klasszikus sorrend.

A tervezés helyett ilyenkor magával az

1. adat(be)gyűjtéssel indítunk, ezt követi az
2. az adat-előkészítés,
3. az adatfeltárás (ami többnyire korrelációk keresését jelenti),
4. az algoritmusok testre szabása (elsősorban skálázható algoritmusok választása aggregálás kerülésével), végül
5. új tudás felfedezése/és az eredmények validálása (heurisztikus [mintakereső] technológiák használata az előrejelzésekhez/becslésekhez).

E megközelítés esetében a hangsúly a hozzáférhető adatok felfedezésén, vagyis olyan információértékek keresésén van, amelyeket ezekből mások még nem nyertek ki. Nyilvánvalóan ez a logika inkább az adattudósok (Data Scientists) által vizsgált problémákra kínál megoldást, akiket sokkal inkább a „Mi történik?" kérdés érdekel, mint a „Miért?" és a „Hogyan?". E speciális jellemzők miatt a Big Data integrálása a hivatalos statisztikába egyáltalán nem megy gördülékenyen.

A Big Data alkalmazási területei

Bankok

A számtalan forrásból szerzett hatalmas adatmennyiséggel szembe kell nézniük a bankoknak. A nagy adathalmazok kezelésére új módszerek szükségesek. Fontos, hogy az ügyfelek igényeit megértsük, és növeljük az elégedettségi szintjüket, továbbá minimalizálni kell a kockázatot. Nagy adatmennyiség által pontos következtetéseket vonhatunk le, ehhez mindig naprakész analízis szükséges.

Különösen a netes bankolás, a mobilok használata generál olyan információkat, melyek szinte korlátlan lehetőséget kínálnak az ügyfelek szegmentációjában, a gazdaságosság szempontjából vizsgált jó és rossz termékek elemzésében. A bankok a piaci verseny diktálta követelmények hatására egyre inkább kiépítik az új típusú bankolások feltételeit, melyek használatával igen sok információ összegyűlik a felhasználóról.

A csalások felderítésének feladatainál is segítség a szociális hálókban keletkezett adatok (pl. Facebook) elemzése. Azzal, hogy a szociális hálókban tárolt adatok nyilvánosak, mód van arra, hogy a „csaló" partner kapcsolatait kizárjuk, vagy nagyobb figyelemmel kísérjük, elkerülve az ismételt károkozás esélyét.

Oktatás

Napjainkban már a közoktatásban és különösen a felsőoktatásban digitális eszközrendszerek állnak rendelkezésre. Az elektronikus naplók segítségével követhetik a szülők gyermekeik iskolai előmenetelét. Pedagógusok, akiknek betekintésük van az adatokba, nagy behatással bírnak az iskolarendszerekre, diákokra és a tantervekre is. Big Data segítségével azonosítani tudják azokat a diákokat, akiknek másfajta oktatási rendszerre, külön odafigyelésre van szükségük, hogy követni lehessen kellő intenzitású fejlődésüket. Lehetőség nyílik jobb oktatási rendszer kialakítására, továbbá a tanárok és igazgatók támogatására.

A felsőoktatás rendszerei, a Neptun, a Coospace, a Moodle stb. igen sok adat elemzését teszik lehetővé. A személyes adatok védelmére való hivatkozás miatt nem igazán történik érdemi elemzési munka. Diplomamunkákban és doktori értekezésekben látunk igen hasznos példákat arra, hogy miképpen segíthetné az oktatási munkát, illetve az oktatási munka hatékonyságának mérését a tárolt adatok elemzése. Természetesen új probléma is felmerül a GPRS hatályba lépésével. Át kell gondolni, hogy milyen anonimizálási eljárások bevezetése után lehet folytatni a sokszor el sem kezdett adatelemzést.

Közigazgatás

Az államigazgatás hagyományosan használja az elsősorban adatgyűjtés-típusú nagyméretű adatállományokat. A definíciók szerint például egy népszámlálás nem Big Data feladat, de az ott kidolgozott technológiák sokszor egy az egyben áttehetők az új típusú feladatok megoldásához.

A társadalomkutatási feladatokban is óriási lehetőséget nyújtanak az automatikusan keletkezett adatok. Még a hagyományos adataink elemzésére sem volt elegendő kapacitás, mikor megjelentek a „végtelen” méretű új adatforrások.[5]

A közigazgatási egységek a megfelelő Big Data-analízissel jelentős előnyre tehetnek szert akár a hasznosságszámításokban, ügynökségek szervezésében, forgalomirányításban vagy a bűncselekmények megakadályozásában. Jelentős haszonnal párosul a nagy adatállományok elemzése, de az átláthatóság és a magánélet kérdéseinek határait nem lépheti túl.

Élő gyakorlat a NAV területén a gazdálkodó szervek kiválasztása az adóellenőrzésre.

Egészségügy és sportanalitika

Napjaink leggyorsabban fejlődő alkalmazási területei az élettudományok. A digitális adatgyűjtési lehetőségek műszaki fejlődése sokkal gyorsabb, mint az azokra kidolgozott elemzési technológiák.

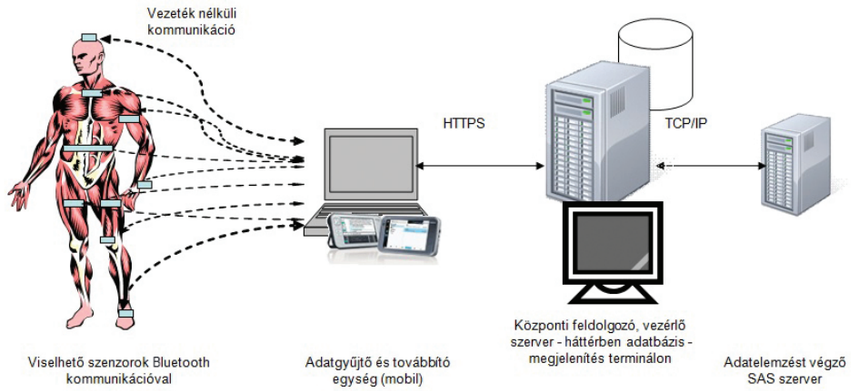
Egy orvos vagy edző sokszor fél a sok műszaki és informatikai újdonságtól. Nem pontosan érti, hogy az új és nem ritkán igen drága technológia miképpen segíti alapfeladatait. Tudják, hogy a fejlődés nem áll meg, használni kell az új eszközöket, de igen fontos az orvosok szakmai tudása is, azok nélkül a technológusok eredményei nem érnek sokat.

Egy konkrét mintaalkalmazás bemutatásával szemléltetem az új technológiák lehetőségeit. Elsősorban rehabilitációs és ortopédiai feladatok segítésére vettünk részt cégemmel egy kutatásban, ahol ún. fizioszenzorok segítségével kíséreltük meg segíteni az ortopédiai problémákkal született gyermekek gyógyítását. Elsősorban a gyógytornászok munkájának támogatására használták a módszert annak ellenőrzésére, hogy a választott terápiák segítik-e az izmok fejlesztését.

A rehabilitációs területen – például egy agyvérzésen átesett beteg esetében – a műszer olyan izomaktivitást is mér, melyet az ember már nem érzékel, így nem egy betegnél, folytatva a gyógytornász által alkalmazott terápiákat, korábban bénultnak tűnt testrészeket lehetett újra aktivizálni.

Az adatgyűjtés és feldolgozás sematikus folyamatát az 1. sz. ábra mutatja be.

1. sz. ábra: Izomaktivitásra vonatkozó adatgyűjtés és feldolgozás sematikus ábrája



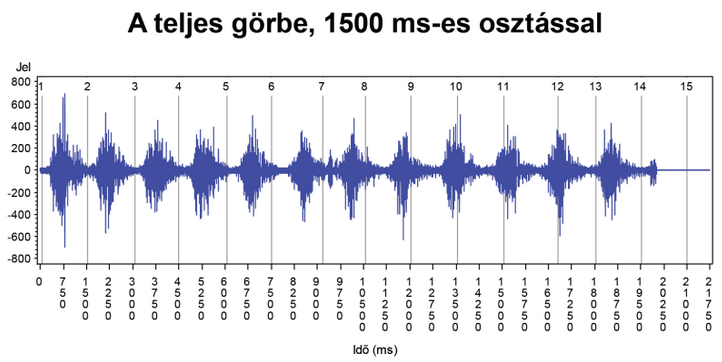
Forrás: saját szerkesztés

A páciensen elhelyezett szenzorok folyamatosan mérik az egyes izmok mozgatására gerjesztett elektromos impulzusokat. Már a szenzor vezérlője digitalizálja az adatokat, és Blutooth kommunikációval küldi akár másodpercenként egymillió mérés adatát a mobil számítógépre. Újabb speciális program küldi biztonságos adatátviteli csatornán az adatcsomagokba (blobokba) szervezett adatokat a felhőbe, a központi adattárolást biztosító szerverre. A szerveren hagyományos adatbázisba töltjük az adatokat, pontosan azonosítva a mérést végző és a mért személy azonosítóját, a mérés idejét, a mérési protokollt (a szenzorok felrakásának orvosok által diktált előírásait) és így tovább.

A tárolt adatok elemzését a SAS szoftverrel végezzük. A vizsgáló orvos végfelhasználói alkalmazással követheti a mért adatok grafikus ábráját, hangolhatja a mérés lépésközeit (milyen periódusonként kerüljön az ábrára az adat), illetve meghatározza a mélyebb adatelemzésre kiválasztott mérési intervallumot.

A mért adatokat az orvosokkal egyeztetett módon ábrázoljuk a SAS grafikai eszközeinek felhasználásával, amit a 2. sz. ábra mutat.

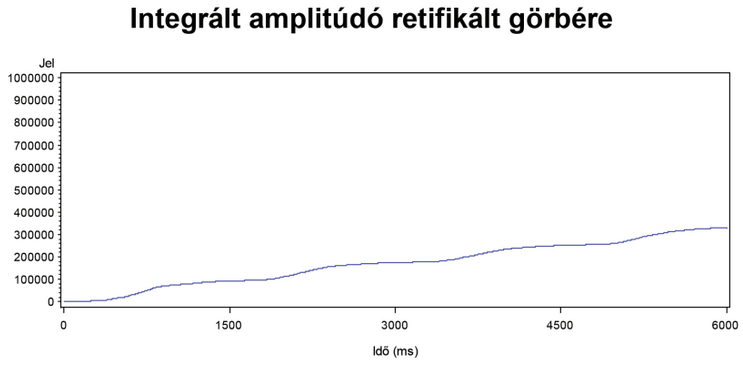
2. sz. ábra: Mért adatok ábrázolása SAS grafikai eszközzel



Forrás: saját szerkesztés

Feladat az izmok teljesítményének mérése, ezért fontos a görbék alatti területek folyamatos összegzése. A hullámgörbe jellege miatt természetesen a negatív tartományban levő értékeket is pozitív értékekkel összegezzük (3. sz. ábra).

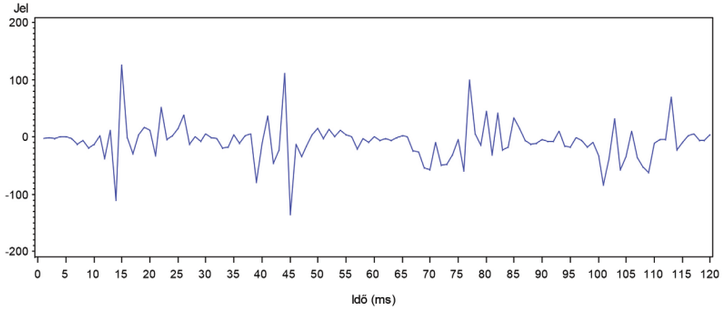
3. sz. ábra: Integrált amplitudó retifikált görbére



Forrás: saját szerkesztés

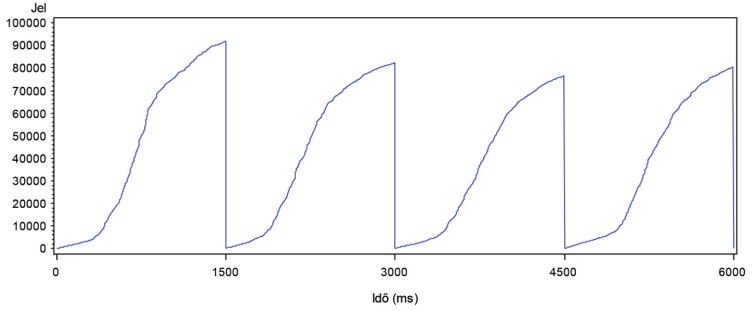
Fontos az orvosok számára a mozgás impulzusainak ábrázolása. Hasonló a szituáció ahhoz, mint amikor a kardiológus az EKG-görbénket vizsgálja (4-5. sz. ábra).

4. sz. ábra: Átlag amplitudó az eredeti görbe 50 ms-os szakaszainak átlagolásával



Forrás: saját szerkesztés

5. sz. ábra: Integrált amplitudó egységenként



Forrás: saját szerkesztés

Az orvosokkal egyeztetve a görbék vizsgálatán kívül bonyolult mérőszámokat számítunk annak érdekében, hogy az adatok hagyományos statisztikai eszközökkel elemezhetők legyenek. Ilyenkor az elemzendő adattáblában a sorok a megfigyeléseinket jelölik, oszlopaink pedig az azonosító adatokon túl egyes mérések során a görbék jellemzésére kialakított számított mutatókat is tartalmazzák.

Az élettudományok területén a sportban is nagy jelentőségű a Big Data technológiákkal kezelt adatok elemzése. Egy élvonalbeli angol, spanyol vagy német labdarúgó-bajnokságban a klubok 15-20 adatelemzőt használnak az edzések, a sportolók pszichikai állapotának, illetve a mérkőzéseknek az elemzésére. Az edzéseken viselt „hátizsákok” szenzorok segítségével gyűjtik a

játékos munkájának intenzitását, a különböző irányokban való elfordulását, a gyorsulását. Excel állományban tárolt adatok állnak rendelkezésre a különböző időpontban végzett vizsgálatokról. A mérkőzésekről is végtelen sok információ áll rendelkezésre, a jó és a rossz passzokról, a futási intenzitásról, a gól előtti akciókról. A mérkőzésen részt vett sportolók közszereplők, így az ott gyűjtött adatok nyilvánosak.

Miközben az edző arra keres választ, hogy kit állítson a lehetőleg győztes csapatba, az elemzőnek strukturális problémája van. Miképpen lehet összekapcsolni az edzések, a pszichológiai vizsgálatok és a mérkőzések adatait? Az objektumaink száma alacsony, csapatonként maximum 30 játékos van, a méréseink száma igen sok. Hogyan hat az edzés a mérkőzés idejére, pontosabban hány edzés adatát használhatjuk fel? A mérkőzés előtt mennyi idővel kell elvégezni a pszichikai vizsgálatot? Mit kezdjünk az összegyűjtött több száz adattal? Már Magyarországon is itt vannak az élvonalbeli csapatoknál a technológiák, egyetemi hallgatóimmal keresünk technikákat az adatok megfelelő elemzésére és a csapatok összeállításának segítésére, miközben tudjuk, hogy a sport mégis csak játék, és a bekövetkező események legalább 50 százalékban véletlenszerűek abban az esetben, ha azonos szintű csapatok mérik össze tudásukat.

Gyártás

Big Datával felfegyverkezve betekintést nyerhetünk a gyártás minőségébe, optimalizálhatjuk a bevételt, és minimalizálhatjuk a veszteséget. Olyan tudásra tehetünk szert, amely fontos ahhoz, hogy talpon maradjunk a mai versengő piaci környezetben. Egyre több gyártó foglalkozik analitikával, ezzel is elősegítve, hogy helyes, gyors és agilis döntéseket hozhassanak.

Az EU-s környezetvédelmi előírások és az Energetikai törvény kötelezte a nagyfogyasztókat arra, hogy kialakítsák és folyamatosan ellenőrizzék energiafelhasználásukat. Magától értetődő, hogy szenzoros adatgyűjtésekkel segítjük a folyamatok mérését. Az egészségügyhöz hasonlóan gyorsan eljutunk a mért adatok grafikus ábrázolásához, de a görbék másodlagos elemzése még nagy kihívás. Fontos lenne például a riasztás egy betört ablak esetén elszökő hőmennyiségről, a várható energia becslése a szolgáltatókkal való áralkuhoz, illetve segítség az Energiatörvény előírásainak auditálásához.

Kereskedelem

A piackutatás és a vevővel való kapcsolat elemzésének során is hatalmas adatokat kell feldolgozni és kiértékelni. A kereskedők hasznos tudást szerezhetnek a vásárlók igényeiről, a tranzakciók lebonyolításáról, és tippeket kaphatnak az üzletük fellendítésére is.

Érdekes feladat az egérmozgások elemzésével a hirdetési csalók felderítése.

Egy hotel energiafelhasználásának becslésekor rendelkezésre állnak a fogyasztási szenzorokkal mért múltbéli adatok, átvehetők az OMSZ-től a várható időjárási előrejelzések, illetve a marketingesek is tudhatják, hogy a jövő hónapban milyen lesz a szobák foglaltsága, hány rendezvény lesz a hónapban. A különböző adatforrások integrálása után becsülhető a következő hónap fogyasztása, ami valós haszonnal bír, mivel a szolgáltató akkor is büntet, ha többet, és akkor is, ha kevesebbet fogyasztunk.

Természetesen egy hotel energetikusa nem képes a sokfajta technológia integrálására. Fejlesztőink ún. Dashboard alkalmazást készítettek, mely olvassa az energetikai szenzorok adatait negyedóránként, átveszi az OMSZ adatait elektronikusan, megkapja a foglaltságot és a várható rendezvények Excel tábláit, és javaslatot tesz a következő hónap energiaigényére.

A Big Data és az adatelemzés kapcsolata: Ki a Data Scientist?

Napjaink egyik legdivatosabb munkaköre a Data Scientist (adattudós, adatelemző-kutató), aki ismeri a Big Data eszközrendszerét, és tudást hoz létre a tárolt adatok elemzésével. A Data Scientist képes a Big Data technológiákkal keletkező adatok kiaknázására. Ezek az emberek nem alkotnak homogén halmazt, a legkülönbözőbb területekről érkeznek.

> **Napjaink egyik legdivatosabb munkaköre a Data Scientist, aki ismeri a Big Data eszközrendszerét, és tudást hoz létre a tárolt adatok elemzésével.**

Milyen eszközrendszere van a Data Scientistnek, és hogyan viszonyulnak ezek a módszerek a hagyományos adatelemzési módszertanokhoz?

Az olvasó számára ismert lehet Dr. Kovács Erzsébet tankönyve.[6] A továbbiakban a könyv tematikájának megfelelően teszek kísérletet a javasolt módszertanok használatára a Big Data világában.

Leíró és feltáró adatelemzés

A többváltozós adatelemzés alapja az „adat", ami a számítógépes elemzés érdekében **mátrixba** rendezett. Szokásos elrendezése szerint soraiban találjuk a megfigyeléseket, és az oszlopok tartalmazzák a megfigyeléseken mért változókat. A Big Data feladatoknál a legnagyobb előkészítő munka a különböző adatforrásokból szerzett adatok rendezése. A gyakran hallott Hadoop technológia éppen arra ad módszertant, hogy a legkülönbözőbb forrásokból származó nagyméretű adatokat össze tudjuk rendezni.

A változók jellemzőinek feltárása mellett gyakran nem tudhatjuk a megfigyelt változók értékkészletét. A folyamatosan keletkező új adatok újabb és újabb értékeket is keletkeztethetnek. Megfontolást érdemel a hiányzó adatok pótlása, a kilógó egyedek feltárása, mivel van adat. Elég könnyű a válasz abban az értelemben, hogy a hiányzó adatokat egyszerűen kiszűrjük az elemzés előkészítő szakaszában.

Külön kutatást igényel, hogy hol van a többváltozós elemzések mérethatára. A számítógépek már igen gyorsak, így valószínű nem a számítási idő a szűk keresztmetszet a nagyméretű adatok elemzésekor.

Sokkal inkább lehet gond az, hogy a mai gépek 64 bites adatokkal tudnak dolgozni, ami általában elegendő, de ezt a méretet kellően nagy számok esetén gyorsan kinőjük, illetve a 12 decimális jegy gyakran kevés a megfelelő pontosság elérésére. Speciális esetekben léteznek algoritmusok a nagyméretű számok aritmetikai problémáinak megoldására, illetve egyre inkább fejlődik a párhuzamos programozás technikája, de ezek a megoldások extra módon megnövelik a feladat elvégzéséhez szükséges kapacitásokat.

Nagyméretű feladatainkat gyakran csak egy mintán tudjuk végrehajtani és annak eredményeként következtetni a teljes sokaságra.

A változók mérési skálája

Nominális skálán mérünk, ha csak megkülönböztetést jeleznek a számok vagy a betűk. Ilyenkor általában nem is egyértelmű, hogy egy-egy kategóriát mivel jelölünk. Különösen a szöveges adatforrásoknál kell sokat dolgozni az egyes szavak egységes kategóriákba rendezésével.

Ritkán mondhatjuk el, hogy a Big Data adatforrása **ordinális skálán** mérhető, azaz rendezhető.

Intervallum skálán mért adatok között már eltérést is számolunk és értelmezünk. Az intervallum hossza a két megfigyelés közötti eltérést tükrözi.

Különösen a szenzoros adatgyűjtésből származó adatoknál kell sokat foglalkozni azzal, hogy a mért adatból miképpen lehet érték adatot képezni. Gyakran akár másodpercenként is milliós nagyságrendű adat áll rendelkezésre. Milyen szakaszokra érdemes mintát venni az adatból? A gyűjtött adat inkább függvényekkel írható le, és sokkal inkább a függvény tulajdonságaival jellemezhető. Például adott intervallumban kiszámolt görbe alatti terület lesz a mérés alapja, többek között a kifejtett erő mértékét így határozhatjuk meg fizioszenzoros adatgyűjtésnél.

Megváltozik a **leíró statisztikák** értelmezése is. Az előző példát követve az orvosok különböző mérőszámokat várnak el egy mozgás intenzitásának és periodicitásának jellemzésére, illetve a különböző időpontokban mért adatok eltérése lesz a fontos információ.

Kategóriák és kereszttáblák elemzése

A hagyományos kereszttáblák is jól jellemezhetik a Big Data sokaságot, viszont igen gyakran be kell vezetni az időt mint új dimenziót. Táblázataink egy percre, órára vagy hónapra vonatkozóan adnak majd megfelelő információkat. Nehéz a hagyományos adattárház technológiákhoz hasonlóan megállítani az időt, mivel méréseink általában folytonosak.

Többváltozós regressziószámítás

Többváltozós lineáris regressziós modellt írunk fel akkor, ha több független magyarázó változó lineáris kombinációjával becsüljük a magyarázni kívánt y változót.

Ha a változók számához képest elegendően sok mérésünk van, akkor az objektumok számának növelése nem hoz új információt, azaz elegendő egy megfelelő minta kiválasztása egyenletesen vagy véletlenszerűen az elemzendő adatállományból. Az eredmények validálása és tesztelése igen fontos és folyamatosan elvégzendő feladat. Félő, hogy az új adatok az idő függvényében, valamilyen ok miatt egészen másképp viselkednek, mint a múlt adatai.

Természetesen igen fontos, hogy teljesüljenek a regressziószámítás használatának feltételei, mely kritériumok teljesítése nagy objektumszámnál igen nehézkes is lehet. A túlságosan nagyméretű feladatoknál ebben az esetben is gondot jelent a futási idő és a számítási aritmetika mértéke. Félő, hogy az adatok túlcsordulása miatt hibás eredményeket kaphatunk.

Logisztikus regresszió

Bár a logisztikus regresszió a lineárishoz hasonló algoritmusokkal oldható meg, kevesebb feltétel meglétét követeli a független változóktól. Megállapításaim hasonlóak a lineáris regresszióhoz.

Pénzügyi alkalmazásoknál vigyázni kell a mintavételezés technikájára, mert előfordulhat, hogy mintánkban például már nem lesz elegendő mértékű negatív esemény, és a modellezés nem fog sikerülni. Tapasztalati képlet, hogy egy százalék alatti negatív esemény alatt nem működik az eljárás, a jó modellezéshez legalább 10 százalékra van szükség.

Faktorelemzés

Faktorelemzést célszerű alkalmazni, ha a mért/vizsgált változók között erős lineáris kapcsolatot tételezünk fel. Ilyenkor a faktorelemzéssel hatékonyan valósítható meg a változók információtartalmának sűrítése. Az alkalmazás előfeltétele, hogy elegendően sok objektumunk/megfigyelésünk legyen a változók számához képest (azaz az adattáblában legalább ötször több sor legyen, mint oszlop). Ilyenkor a megfigyelések számának további növelése már nem pontosítja érdemben az eredményeket.

Diszkriminancia elemzés

A diszkriminancia analízis jó módszertan lehetne a nagyméretű adatállományunk csoportosítására, sőt, jó lenne, ha könnyen megtalálnánk egy ismeretlen sokaságot szétválasztó változót. A gond a módszertan alkalmazásánál éppen az előzetes feltételek teljesülése. Nehezen dönthető el, hogy

- a változók többváltozós normális eloszlást követnek, és
- minden csoportnak azonos a kovariancia mátrixa.

Sokdimenziós skálázás

Ez a módszertan igen hasznos a változók számának csökkentésében úgy, hogy az objektumok közötti távolság ne változzon. Amikor azt állítom, hogy a skálázás nem lesz a Big Data-felhasználó gyakran használt módszere, akkor az igen nagy pontosságot igénylő számításokra gondolok.

A neurális háló módszertanának különleges helyzete

Elsősorban a Credit Scoring feladatoknál használtuk, konkurens algoritmusként a bináris fa és a logisztikus regresszió mellett a neurális hálókat. Jó tíz évvel ezelőtt kritizáltuk a módszert adatfüggősége és az eredmények nehéz értelmezése miatt, viszont sok esetben a neurális háló jobb modellt adott, mint a másik két eszköz.

Napjainkban, a Big Data világában éppen az adatoktól való függőségét emelik ki pozitív tulajdonságként. Ismét népszerű a mesterséges intelligencia fogalma, és ott jolly joker lett a neurális háló, amely az egyre növekvő adatok körét fel tudja dolgozni, igaz, nem ritkán eltérő eredménnyel. Az esetek 70-80 százalékát jól becsüli a módszertan, a maradékot pedig mindenképpen egyedi módszerekkel kell kezelni.

A Big Data kezelésének szoftveres háttere

Biztosítási ajánlatok a roaming területre való belépéskor

Már Magyarországon is élő alkalmazás, mikor az országot mobiltelefonnal elhagyó utas biztosítási ajánlatot kap. Az elfelejtett utasbiztosítás esetében hálásak lehetünk az ajánlattevőnek, de semmiképpen sem vesszük a figyelmeztetést zaklatásnak, ellentétben a telefonos megkeresésekkel. Az ajánlattétel mögött tipikus Big Data feladat van, melynek végrehajtásán sok partnernek kell közösen dolgozni.

> **Már Magyarországon is élő Big Data alkalmazás, mikor az országot mobiltelefonnal elhagyó utas biztosítási ajánlatot kap.**

1. A cellainformációk szerint a mobilszolgáltató érzékeli, hogy valaki elhagyja az országot. Ez nem jelent extra feladatot, megszokott üzenet, mikor az országhatár elhagyásakor SMS-t kapunk a roaming feltételekről. Már csak azt kell kimódolni, hogy ezzel az SMS-sel egy időben a biztosító is kapjon jelzést, hogy milyen telefonszám birtokában hagyta el az utas Magyarországot.
2. Ellenőrizni kell, hogy az utasnak van-e már utasbiztosítása. Természetesen, ha van ilyen, különösen saját társaságunknál, akkor nem küldhetünk ajánlatot.
3. Ha nincs megkötött biztosítás, akkor attól függően, hogy az utas partnere vagy sem a biztosítónknak, konstruálunk ajánlatot, és küldjük el azt.
4. Ha van jó Fintech/Insurtech alkalmazásunk, azaz van szerződéskötési applikációnk, akkor már csak az üzletet kell realizálni, mely az esetek döntő százalékában emberi beavatkozás nélkül megtörténhet.

A Big Data technológia hatása a biztosítókra. Az önvezető autók elterjedése [7 8]

Napjaink legintenzívebb kutatása a gépjárműiparban az önvezető autók területén folyik. Az országutakon elhelyezett jeladók és a már hagyományos GPS-navigáció alapján az okosautó elvezet a kívánt célba, mint ahogyan azt a 4-es metró teszi a budapesti mindennapokban. Ma már érthető, hogyan készülhet olyan szoftver és azt kiszolgáló hardver környezet, mely a feladatot megoldhatóvá teszi.

A számítógépek méretének csökkentése és a hálózati megoldások új generációja gyakorlati lehetőséget ad a feladatra néhány éven belül, tehát a biztosítóknak is fel kell készülniük az új kihívásokra.

A technológiai fejlődés csökkenti vagy növeli a gépjármű-biztosítások kockázatát és díját?

Az önvezető autózás biztosan csökkenti a kockázatot, viszont a gépjárművek értéke, így a kártérítési díjak emelkedni fognak. Azzal, hogy az önvezető technológia soha nem lesz kizárólagos, gondolhat-e az okosautó a partnerek nem várt viselkedésére? Lesznek-e elég fejlettek a szoftverek, hogy kiszűrjék az összes nem várt eseményt, melyet egy gyakorlott sofőr reflexe kiküszöböl? A lakosság körében mindig lesz igény hagyományos és hibrid megoldásokra.

Hogyan érjük el, hogy a biztosító megtartsa szerepét, ne vegyék át azt a technológiai fejlődést kihasználó autógyártók, és a bekövetkező károkat ne a számukra megfelelő szervizbe irányítsák? A szabályozásnak és a biztosítási iparnak a hagyományos és önvezető technológia együttélésére kell felkészülnie, nem csak átmeneti időre.

Már a mai autók is rendelkeznek mindenféle okos, Front-assist környezetfigyelő rendszerekkel (városivészfék-funkció, gyalogosfelismerés, automatikus követésitávolság-tartás (radarral), sávtartó asszisztens, parkolóasszisztens 3.0, környezeti kamera (area view), gumiabroncs-ellenőrzés, másodlagos ütközéselkerülés, fáradtságfelismerés stb.). Miképpen fogja a biztosító szponzorálni az ilyen eszközök meglétét a díjakban?

Végül érdekes és még nyitott kérdések az új technológiák elterjedésekor:

1. Ki rendelkezik majd biztosítással? A gyártó? Ki írja alá a szerződést?
2. A baleset elszenvedői/okozói egyeztetnek, fotó és kárbejelentő lap készül, vagy az
3. egész úton felvett képanyagot használják?
4. A baleset helyén rendőri helyszínelés, jegyzőkönyv készül, vagy az önvezető autó megáll, és hívja a rendőrrobotot?

Befejező gondolatként: engem személyesen örömmel tölt el, hogy a nagyméretű adatállományok elemzésének reneszánsza van a Big Data elterjedésével. Legyen is így a jövőben, de ne feledjük, hogy a technológiák ismeretéhez nagy tudású programozóra és az adatok elemzéséhez megfelelő matematikai ismeretekkel és szakmai tudással is rendelkező szakértőre lesz szükség. Míg a jó programozóknak szilárdabb matematikai alapokra, statisztikai ismeretekre, az üzleti-biztosítási területen pedig az informatikai problémák megoldásához való nagyobb toleranciára van szükségük.

Akkor lesznek sikeres eredményeink a Big Data és a Data Science területén, ha a különböző szakterületek képviselői együtt tudnak dolgozni, és eltűnnek a napjaink szigorú projektmenedzsmentjében felállított falak. Ha nem értjük meg a másik fél problémáját, megoldási nehézségeit, és csak az adminisztrációs feladatok teljesítésére koncentrálunk, talán teljesülnek a projektek, de nem lesz igazi sikerélményünk.

HIVATKOZÁSOK

[1]P. Groves – B. Kayyali – D. Knott – S. Van Kuiken (McKinsey Quarterly): The 'big data'revolution in healthcare
2013 - pharmatalents.es
[2]PJH. Daas – MJ. Puts – B. Buelens: Big data as a source for official statistics
- Journal of Official …, 2015 - degruyter.com
[3]B. Buelens – P. Daas – J. Burger – M. Puts: Selectivity of Big data
URL http://www …, 2014 - pdfs.semanticscholar.org
[4]Giczi J. – Szőke K.: Hivatalos statisztika és a Big Data
STATISZTIKAI SZEMLE, 2017 - real.mtak.hu
[5]Ságvári B.: Társadalomtudomány a Big Data korában
STATISZTIKAI SZEMLE, 2017 - real.mtak.hu
[6]http://etananyag.ttk.elte.hu/FiLeS/downloads/14_KOVACS_E_Tobbvalt_adatelemzes.pdf
URL http://www …, 2014 - pdfs.semanticscholar.org
[7]Beleznai Endre: Technológiai változások hatása a gépjármű-biztosításra
Gépjármű- és Lakossági Vagyonbiztosítási Igazgatóság, Generali Biztosító
[8]Dr. Kovács Erzsébet – Trinh Anh Tuan: Önvezető autók és a biztosítás
Corvinus-Generali Akadémia III.
2018. április 19.

IRODALOMJEGYZÉK

Peter Groves – Basel Kayyali – David Knott – Steve Van Kuiken (McKinsey Quarterly): The 'big data'revolution in healthcare
https://www.mckinsey.com/~/media/mckinsey/industries/healthcare%20systems%20and%20services/our%20insights/the%20big%20data%20revolution%20in%20us%20health%20care/the_big_data_revolution_in_healthcare.ashx
Letötés ideje: 2018. 05. 14.
2013 - pharmatalents.es
PJH. Daas – MJ. Puts – B. Buelens: Big data as a source for official statistics
https://www.researchgate.net/publication/281558593_Big_Data_as_a_Source_for_Official_Statistics
Letöltés: 2018. 05. 14.
- Journal of Official, 2015 - degruyter.com
B. Buelens – P. Daas – J. Burger – M. Puts: Selectivity of Big data
https://www.researchgate.net/publication/261436243_Selectivity_of_Big_data
letöltés: 2018. 05. 14

Giczi Johanna – Szőke Katalin: Hivatalos statisztika és a Big Data
STATISZTIKAI SZEMLE, 2017 - real.mtak.hu
https://doi.org/10.20311/stat2017.05.hu0461
Ságvári Bence: Társadalomtudomány a Big Data korában
STATISZTIKAI SZEMLE, 2017 - real.mtak.hu
https://www.researchgate.net/profile/Bence_Sagvari/publication/317713050_Tarsadalomtudomany_a_Big_Data_koraban/links/597f7f77a6fdcc1a9acec111/Tarsadalomtudomany-a-Big-Data-koraban.pdf
Letöltés ideje: 2018. 05.04
https://doi.org/10.20311/stat2017.05.hu0491
Kovács Erzsébet: Többváltozós adatelemzés
http://etananyag.ttk.elte.hu/FiLeS/downloads/14_KOVACS_E_Tobbvalt_adatelemzes.pdf
Letöltés ideje: 2018.05.14.
Szabó Dániel: Adatbányászat a biztosítási szektorban
Biztosítás és Kockázat, 2015. december
http://www.mabisz.hu/images/stories/docs/biztositas-es-kockazat/2_4/biztositas-es-kockazat-2-evf-4-szam-5-cikk.pdf Letöltés: 2018.05.14.
https://doi.org/10.18530/bk.2015.4.62
Bulcsú Fajszi – László Cser – Tamás Fehér: Business Value in an Ocean of Data: Data Mining from a User Perspective, Kindle Edition
https://www.amazon.com/Business-Value-Ocean-Data-Perspective-ebook/dp/B07323J5N7
Letöltés ideje:2018. 05.04.