Magyar Tudomány 181(2020)7, 869–875
DOI: 10.1556/2065.181.2020.7.2

CSILLAGÁSZATI (MENNYISÉGŰ) ADAT

ASTRONOMICAL (AMOUNT OF) DATA

Csabai István
az MTA levelező tagja, egyetemi tanár
Eötvös Loránd Tudományegyetem Természettudományi Kar Komplex Rendszerek Fizikája Tanszék, Budapest
csabai@complex.elte.hu

ÖSSZEFOGLALÁS

Ahogy haladunk előre a világ megismerésének folyamatában, a jelenségek pontosabb megértéséhez egyre több adatot kell elemezni. A közelmúltban túlléptük azt a mennyiségi küszöböt, amelyet az emberi elme még kezelni tudott. Mára nemcsak az észlelési adatok gyűjtése és tárolása, hanem azok feldolgozása, modellezése, sőt bizonyos értelemben a jelenségek megértése is gépek segítségével történik. A csillagászat évszázadokkal ezelőtt is a tudomány úttörő ágazata volt, és azoknak a területeknek egyike, ahol az adatforradalom elsőként lezajlott.

ABSTRACT

As we progress with the understanding of the world, we need to analyse more and more data to uncover the details of the phenomena. In the last decades, we have crossed a border beyond which data cannot fit into and handled by human mind alone. Today not only collection and storage of observational data are handled with machines but also modelling and in some sense the scientific understanding itself requires help from computers. Astronomy has always been at the forefront of the sciences, and it is one of the fields where the data revolution first happened.

Kulcsszavak: csillagászat, tudományos adatok, modellezés, gépi tanulás, mesterséges intelligencia

Keywords: astronomy, scientific data, modelling, machine learning, artificial intelligence

A tudományok ősidők óta fontos alapkérdésekre keresik a választ, például arra, hogy hogyan működik az élő szervezet, miként mozognak a csillagok az égen, hogyan alakítható ki igazságos társadalom. Nehéz ugyan az egyes tudományágakat összevetni egymással, de talán kijelenthetjük, hogy a csillagászat az első diszciplínák között járt a területére eső bizonyos alapjelenségek megértésében. Köszönhető ez annak, hogy a kitűzött kérdések egy része nagyon egyszerű, fun-

damentális fizikai jelenségeken alapul. Első közelítésben a Naprendszer kisszámú, nagyrészt szabad szemmel is jól látható égitestből áll, amelyek a köztük lévő távolságokhoz képest nagyon kis kiterjedésűek, vagyis pontszerűnek tekinthetők. A testek közt praktikusan „űr" van, azaz nincs súrlódás, nincs semmilyen zavaró tényező, egymás mozgását csak a néhány betűs képlettel leírható gravitációs kölcsönhatáson keresztül befolyásolják. A teljes igazság ennél persze jóval bonyolultabb, de az élet leírásának ugyanilyen szintű egyszerűsítése nemigen tehető meg. Ha például az öregedés jelenségét szeretnénk leírni, nem hanyagolhatjuk el a makromolekulák, sejtek és szövetek hierarchiáját, a gének ezreinek kölcsönhatásait, és nem valószínű, hogy bármikor is rábukkanunk egy olyan elegáns képletre, mint a newtoni gravitációé.

Mindez nem jelenti azt, hogy nagyon könnyű volt rájönni, hogyan is működik a Naprendszer. Évszázadokon keresztül figyelték ugyan a csillagászok a Nap, a Hold, a csillagok és a szabad szemmel látható bolygók égi pályáját, de a mérések nagyon sokáig nem voltak elég pontosak ahhoz, hogy kikényszerítsék a viszonylag egyszerű, de a fizikai okokat figyelmen kívül hagyó ptolemaioszi modell felülbírálatát. Tycho Brahe (1546–1601) dán csillagász épített végül egy akkora és olyan stabil szögmérőt (távcsövet ekkor még nem használtak), amely elég pontos adatokkal szolgált egy jobb modell kialakításához. Ezt a munkát a tudományos „big data" egyik korai példájának tekinthetjük. Tycho Brahe életének jelentős részét, mintegy harminc évet áldozott a nagy műszer megépítésére, a bolygók pozícióinak észlelésére és azok táblázatokba jegyzésére. Kevésbé közismert, hogy a „nyers" adatokból a végső, használható táblázatok elkészítése, az adatok „feldolgozása", rendszerezése Johannes Keplernek (1571–1630) ugyanennyi idejét vette igénybe. Kepler maga így írt a munka nehézségeiről: „Bízom bennetek, barátaim, hogy nem ítéltek engem teljes egészében a matematikai számítások taposómalmára, és hagytok időt filozófiai spekulációkra, amelyek az egyetlen örömömet jelentik az életben." A nyers adatok katalógusokba rendezése ma sem tartozik a kutatók kedvelt és elismert tevékenységei közé. A végül 1627-ben, a szponzoráló uralkodó tiszteletére Rudolf-táblázatok néven megjelent adathalmaz 1405 csillag és az akkor ismert bolygók légköri hatásokra korrigált pozícióját tartalmazta szögperc pontossággal. Az adatok mellett kisegítő logaritmustáblák és szemléletes példák is szerepeltek a kiadványban, megkönnyítve a horoszkópokat és a Vénusz vagy Merkúr átvonulását számoló „felhasználók" dolgát. Sajnos napjainkban nem minden közzétett adathalmazra jellemző ez az átgondolt szemlélet. A kiadás után Tycho Brahe rokonai többször próbálták megszerezni a táblázatok publikálásának jogait és hasznát. Azt állították, hogy Tycho Brahe munkájának gyümölcseit a saját családjának kellene élveznie, és nem Tycho Brahe egyik versenytársának. Kepler vitatta ezt, mivel Tycho Brahe halála előtt is évekig együttműködtek az adatok gyűjtésében, a munka jelentős részét, a számításokat, az adatok rendszerezését pedig ő végezte el. Ilyen jellegű viták a szerzőségről, a „rutin technikai" munkák elismeréséről ma is fellépnek.

Ennek a talán első tudományos „big data” projektnek az eredményei messzemenő következményekkel jártak. Kepler maga is kereste az általa felállított, azóta Kepler-törvényekként tanított tapasztalati összefüggések mögött rejlő mélyebb összefüggéseket, de végül mintegy fél évszázaddal később Isaac Newton (1643–1727) bukkant rá a földi és mennyei szférákat összekötő mechanikai törvényekre. Az így kialakult új tudományos paradigma mentén már szinte egyenes út vezetett újabb és újabb jelenségek, így a hőtan, az elektromosság és mágnesesség megértéséhez, melyekben ugyanúgy a technológiai fejlődés, az észlelések, kísérletek és matematikai modellek ciklusai segítik egymást.

Kevesebb mint száz éve, az első 100 hüvelyk átmérőjű teleszkóp tette lehetővé, hogy felismerjék: a Tejúton kívül is van világ, melyben csillagrendszerünkhöz hasonló galaxisok milliárdjai helyezkednek el egy, az addig ismertnél elképzelhetetlenül nagyobb univerzumban. A 20. század végéig mintegy ezer galaxis térbeli pozíciója vált ismertté. Ekkorra érett meg a mikroelektronika arra, hogy leváltsa az addig használt fotólemezeket. A körülbelül kétévente duplázódó kapacitású technológiára alapozva az 1990-es évek legvégén a Sloan Digitális Égboltfelmérés (SDSS) kamerája már 120 millió pixeles CCD-kamerát tartalmazott, amelynek segítségével egy szűk évtized leforgása alatt 300 millió galaxist fényképezett le, egymilliónak a színképét és abból a távolságát is meghatározta, lehetővé téve az univerzum első valamirevaló háromdimenziós térképének megalkotását. A technológia fejlődésének sebességét és gyors társadalmi beágyazódását mi sem jelzi jobban, mint hogy az egyik legnagyobb mikroelektronikai cég a napokban jelentette be egy 108 megapixeles, mobiltelefonokba szerelhető kamera elkészültét. Ahogy Keplernek is kihívást jelentett az ezer néhány száz csillag adatainak rendezése a kor technológiájával, ugyanúgy az SDSS nyers felvételeinek feldolgozása, katalógusba rendezése, közzététele a projektre szánt emberévekben számolva több munkába került, mint maga az észlelés. Mindezt nem lehetett volna megtenni a szenzorokkal párhuzamosan fejlődő számítógépek nélkül. Érdekesség, hogy a projekt indulásakor nem állt még rendelkezésre több terabájtos adathalmazok hatékony tárolására és elérésére alkalmas hardver, de az exponenciális technológiai fejlődést leíró Moore-törvény jóslata teljesült, és amikorra elkészült a felmérés, elérhetővé váltak a megfelelő számítógépek.

Talán az SDSS volt az első nagy tudományos felmérés, amely annyi adatot termelt, hogy gépi segítség nélkül ember végignézni sem tudja, nemhogy alaposan megvizsgálni. Ha egy lelkes kutató vagy doktorandusz másodpercenként egy galaxis felvételét ki tudná értékelni, a nap 24 órájában lankadatlanul dolgozva is kilenc és fél év folyamatos munkájába kerülne a 300 millió objektum átnézése. Nem valószínű, hogy ezek után bárki vissza tudna emlékezni minden konkrét galaxisra, annak jellemzőire, vagy ennyi adatban összefüggéseket fedezne fel. A sarkított példa azt hivatott demonstrálni, hogy a tudomány számos területe túllépett azon a fázison, amikor az emberi érzékszervek elegendőek voltak a világ

jelenségeinek megfigyeléséhez, illetve amikor az emberi elme kapacitása és sebessége elegendő volt az adatok, összefüggések kezeléséhez. Az adatok tekintetében az egyik fontos aspektus a banálisnak tűnő adattárolás és -keresés. Amíg egy-két oldalon áttekinthető táblázatokban elfértek egy-egy kutatás eredményei, ezt a feladatot nem is igazán tekintették a tudományos munka lényegi részének. Érdekes módon az élet más területein, a bankokban, biztosító- vagy repülőtársaságoknál már korábban keletkezett annyi információ, hogy azok rendezése adatbázis-kezelő szoftvereket igényelt. Ezeket a szoftveres megoldásokat, az úgynevezett relációs adatbázis-kezelőket kellett adaptálni a tudományos adatok kezelésénél fellépő igényekhez. A feladat az SDSS esetében mind a hirtelen keletkezett nagy adatmennyiség, mind annak összetett jellege miatt számottevő kihívást jelentett. Az adatok ugyanis nem az üzleti életben megszokott nevek, elnevezések és pénzösszegek voltak, hanem térbeli koordináták, galaxisok paraméterei, színképek, képek. Ehhez új típusú adatbázisokat és sokdimenziós keresőalgoritmusokat kellett kidolgozni. A végeredményül kialakult publikus adatbázisrendszer, a SkyServer, akár Kepler Rudolf-táblázatai, számos segítő függvényt, tanító jellegű példát is tartalmaz, és azóta is alapreferenciája a kutatóközösségnek, legyen szó egy új szupernóva vagy gravitációshullám-forrás helyének meghatározásáról. Ezt az interaktív adatarchívumot néha Virtuális Obszervatóriumnak is nevezik, utalva arra, hogy ez a valódi univerzum virtuális, háromdimenziós mása, és számos jellemző újabb időigényes észlelések nélkül is gyorsan elérhető.

Az adatok rendszerezése fárasztó rutinmunka, és nem csodálkozunk, ha ilyen monoton munkában a gép segítségét vesszük igénybe. Az adatok kiértékelése, értelmezése, összefüggések feltárása sokkal inkább a kreatív emberi gondolkozás felségterülete, de ma már ez sem lehetséges gépi segítség nélkül. A puszta mennyiségen túl kihívást jelent az adatok magas dimenzionalitása, komplexitása. Az evolúció által kifejlődött elménk remekül elboldogul a háromdimenziós világban, de gondoljunk akár a csupán négydimenziós gömbökkel kapcsolatos Poincaré-tételre, máris elbizonytalanodunk, és nem sokat segít az intuíciónk. Az SDSS galaxisait, ha csak minimális paraméterekkel, színükkel – már ez is öt a megszokott RGB helyett az ultraibolya és infravörös sávok miatt –, morfológiai jellemzőikkel, égi koordinátáikkal jellemezzük, akkor is már tucatnyi dimenziónál járunk. Ha néhány pontot ábrázolunk egy szokványos grafikonon, könnyen észrevesszük a jellemző trendeket, ahogyan Kepler is felfedezte a bolygók keringési ideje és pályasugara közti összefüggést. De ki tud átlátni több millió pontot tíz dimenzióban, és azok közt szabályszerűségeket felfedezni? Részben segíthetnek azok a módszerek, amelyek tömörítik az adatokat, és az emberi elme számára kezelhető dimenziókba redukálják azokat. De mi van akkor, ha maguk az összefüggések inherensen magasabb dimenziójúak, komplexebbek az emberi elme által felfoghatónál?

Ha elfogadjuk, hogy gondolkozásunk az agyunkban található idegsejtek működésének eredménye, és figyelembe vesszük azt is, hogy az evolúció milyen

feladatok megoldására optimalizálta ezt a berendezést, nem tagadhatjuk le, hogy limitált kapacitással rendelkezünk mind a befogadható adatok mennyiségét, mind pedig az információ feldolgozási sebességét tekintve. Mindennapi tapasztalataink ezt messzemenően alá is támasztják, akár olyan egyszerű feladatokra gondolva, mint tízjegyű számok szorzása vagy egy hosszabb mondat visszafelé elmondása. A tudományon belül is egyre szaporodnak azok a feladatok, ahol a nemrégiben új lendületet kapott *gépi tanulás* vagy a fellengzősebben hangzó *mesterséges intelligencia* segítségünkre lehet. A gépi intelligencia régi ábrándja a tudományos-fantasztikus regényeknek, és a tudomány is többször nekilendült megvalósításának. Neumann János, Alan Turing és a számítástechnika többi úttörői is sokat gondolkoztak az emberi elme működésén, és alapvető motivációt jelentett a számítógépek megalkotásában, noha végül azok struktúrája nem mutat sok hasonlatosságot a biológiai rendszerekéhez. Az elektronikus számítógépek megjelenésével együtt, a 60-as években alkották meg az idegsejteket utánzó első ún. perceptront, de mivel csak nagyon egyszerű feladatokat tudott megoldani, egy időre feledésbe merült ez a megközelítés. A 80-as évek végén, részben a személyi számítógépek elterjedésének, részben pedig a többrétegű perceptronok hatékony tanítási eljárásának (az ún. back propagation algoritmus) megalkotásával újabb lendületet kapott a gépi tanulás kutatása, és a cikk szerzőjének is lehetősége volt már akkor ezen a területen dolgozni. Ekkor azonban még a kor számítógépei a mesterséges neuronhálók igényeihez mérten rendkívül alacsony kapacitásúak (néhány száz kilobájt memória) és sebességűek (néhány megahertz órajel) voltak. Még ennél is nagyobb probléma volt talán, hogy akkoriban adatok se nagyon álltak rendelkezésre, így egy ún. „mesterséges intelligencia tél" köszöntött be. A tavasz napjainkban bontakozik ki a sok ezer processzormagot tartalmazó grafikus kártyáknak és az internetes adatrobbanásnak köszönhetően.

A gépek egy-két év leforgása alatt a reménytelenül gyenge teljesítményről elérték azt a szintet, amikor már a legtöbb képi felismerési feladatban az emberi megfigyelőknél jobb teljesítményre képesek. A nagy internetes cégek sokmilliós képhalmazain tanított algoritmusok nagyobb biztonsággal találják meg az összetett képeken rejtőző tárgyakat, ismerik fel a kutyák fajtáit, az emberi arcokat, mint az egyébként ilyen feladatokban otthonosan mozgó emberek. Tudományos kérdésekben gyakran kevesebb olyan adat áll rendelkezésre, amelyik fel van címkézve a megtanulandó tulajdonságokkal, kategóriákkal, így az első értékelhető eredmények az elmúlt egy-két évben születtek meg, de számuk rohamosan nő. Ha a hétköznapi felvételeken is jobban teljesít a gép, mint az ember, a szem és a látókéreg evolúciós célját figyelembe véve talán nem is annyira meglepő, hogy a gép a mindennapi tapasztalatunktól eltérő tudományos adatok elemzésében még nagyobb sikereket érhet el. Így komolyabb radiológiai tudás nélkül is fel tudtunk állítani egy olyan gépi tanulási algoritmust, amelyik a mammográfiai röntgenfelvételen a kezdődő rákos elváltozásokat megbízhatóbban ismeri fel, mint a

képzett radiológusok. Az új típusú megközelítés az „egyszerű" képfeldolgozáson túl olyan területeken is érdekes eredményeket mutat fel, ahol hagyományos matematikai módszerekkel nem kezelhető, nehéz inverz problémák lépnek fel. Egy közelmúltbeli tanulmányban azt találtuk, hogy a mesterséges neuronhálózat nemcsak pontosabban képes a kozmológiai paramétereket meghatározni gravitációs lencsék mérései alapján, hanem „feltalált" egy olyan egyszerű, de hatékony új eljárást, amely jól értelmezhető, és akár egy kozmológus is kitalálhatta volna.

Minden jel arra mutat, hogy a tudományosadat-forradalomnak még csak az elején tartunk. Hamarosan indul a Large Synoptic Survey Telescope (LSST) projekt, amely hetente több adatot gyűjt, mint az SDSS egy évtized alatt. És ez csak egyetlen földi bázisú csillagászati észlelési projekt a látható tartományban. Emellett számos más távcső készül, amelyek a földfelszínről vagy az űrből észlelik egyre részletesebben az elektromágneses spektrum széles tartományát a rádióhullámoktól a gamma-sugárzásig, sőt a közelmúltban új modalitásként a gravitációshullám-detektorok is csatlakoztak. Mindezek a „csillagászati mennyiségű" adatok azonban eltörpülnek például a modern orvosbiológia által termelt adatmennyiség mellett. A néhány éve még hárommilliárd dollárból és több mint egy évtized alatt megvalósult humán genom szekvenálás mára már csupán pár napot vesz igénybe, és néhány száz dollárba kerül, közel terabájtnyi adatot generálva mintánként. A rutinszerűen, egy-egy páciens szöveti mintáiból készült mikroszkópos felvétel nagyjából 4 gigapixeles, a tomográfok egyre nagyobb felbontású, 3 dimenziós felvételei még nagyobb adatmennyiséget hordoznak. Az *adattudományt* ma már sokan önálló diszciplínaként kezelik, ami a tudomány hagyományos területein túl az ipari alkalmazások és az üzleti élet szinte minden ágazatában egyre nagyobb szerepet kap. Ahogy Galilei és Newton egyesítette a mennyei és földi szférák leírását, az adattudomány ugyanannak az észlelés-modellezés-jóslás-tesztelés paradigmának a kiszélesítése, amelyet a tudományok évszázadok óta követnek: ma már a mindennapi élet számos területe is tudományos megközelítést igényel.

Ahhoz, hogy az univerzum történetét, az élő szervezet, a társadalom vagy a gazdaság komplex jelenségeit megértsük, szükség is van a nagy adathalmazokra: megbízható komplex modellek nem alkothatók kevés adatpontból. A sok információ feldolgozásához pedig szükség van az adatokat rendezni és elemezni képes új, számítógépes módszerekre, köztük a gépi tanulás eszköztárára is. A technológiai innováció és a tudomány mindig is kéz a kézben járt. A tudomány épít az új technológiákra, a pontosabb szögmérőre, a precízen csiszolt távcsőtükör nagy fénygyűjtő felületére, a fotólemez vagy a CCD-csip érzékenységére, a számítógépek gyors információfeldolgozó képességére. Ugyanakkor a modern fizika az embernek abból a „haszontalan" álmodozásából született, hogy megértse a mennyei szférák harmóniáját, a hirtelen a modern gazdaság motorjává váló mesterséges intelligencia pedig abból a filozofikus vágyból, hogy megértsük az emberi gon-

dolkodás mikéntjét. A mélyebb összefüggéseket firtató tudományos elmélkedés esetenként váratlanul, de időről időre megbízhatóan terem olyan gyümölcsöket, amelyekre új technológiák alapozhatóak.

Arthur C. Clark szavait idézve: „Bármely kellően fejlett technológia megkülönböztethetetlen a mágiától." És valóban, a mechanika törvényeinek megismerése lehetővé tette, hogy katedrálisokat építsünk, és egyszerű gépekkel olyan tárgyakat emeljünk fel, amelyeket emberi erővel lehetetlen. A termodinamikai ismeretekre építve képesek lettünk kontinenseket és óceánokat átszelni, és mindenki garázsában ott van a „hétmérföldes csizma". Az elektromosság és kvantummechanika törvényeinek feltárása elhozta a villamosítást és az internetet, a mobiltelefon „varázstükrével" pedig távolba láthatunk és hallhatunk. Vajon miféle új csodákat hoz az adattudomány és a mesterséges intelligencia?