



VIKTOR MAYER-SCHÖNBERGER AND KENNETH NEIL CUKIER

BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK AND THINK

Ljudje predvsem razvijamo kulturo, ki vključuje tudi informacijsko tehnologijo in vsebino, nov izsledek tega razvoja pa so velikanske količine podatkov, t. i. masovni podatki (angl. *big data*).

Že pred petimi leti je bilo ugotovljeno, da lahko Googlova storitev Flu Trends zazna regionalne izbruhe gripe 7–10 dni prej kot uradni centri za nadzor in preprečevanje nalezljivih bolezni! Googlova storitev Flu Trends kaže velik potencial kot pravočasni, robustni in občutljivi sistem nadzora. Skrivnost uspeha Googlovega sistema je velika populacija uporabnikov, ki iščejo po svetovnem spletu, v ozadju pa so velikanske količine podatkov.

Masovni podatki pomagajo strankam pri odločanju *ad hoc* na podlagi zanesljivih napovedi, s tem pa tudi pri modernizaciji poslovnih procesov in modelov, ki ustvarjajo dodano vrednost za vlagatelje, stranke, partnerje, zaposlene in družbo kot celoto.

Masovni podatki imajo potencial, da postane zaradi njih življenje ljudi veliko bolj prijetno, še zlasti na področju energije, prometa in zdravstva, kjer lahko prispevajo pomembne izboljšave.

*Datability*¹ je inteligentna analiza obstoječih podatkov v obliki optimizacije procesov in učinkovitosti virov za spodbuditev novih priložnosti. Gre tudi za zagotavljanje koristi za navadne ljudi, kot so stranke v zdravstvenem sektorju, potniki v sektorjih zračnega in cestnega prometa, stranke v spletnem nakupovanju (Ali so med njimi tudi uporabniki velikih knjižničnih informacijskih sistemov?).

Masovni podatki so velika priložnost! Znatni porast obsega razpoložljivih podatkov iz več virov, kot so senzorji, mobilne naprave, spletne transakcije in socialna omrežja, predstavlja veliko priložnost za celotno poslovno rast z izboljšanjem proizvodov, storitev, izkušenj strank in produktivnosti podjetij vseh velikosti.

Ključ za uspeh sta dostopnost in relevantnost. Da bi izveleki bistveno vrednost iz masovnih podatkov,

potrebujejo podjetja samodejni dostop v realnem času predvsem in izključno do relevantnih podatkov, ki lahko bistveno vplivajo na njihovo poslovanje. Podjetja morajo imeti prožnost pri določanju in dinamični posodobitvi svoje strategije za dostop in zbiranje podatkov ne glede na vir podatkov, vrsto ali obliko. Relevantni in razpoložljivi podatki omogočajo inteligentno in učinkovito odločanje, ki lahko preoblikuje podjetja v agilne, na podatke usmerjene in bolj uspešne organizacije.

Po Mayerju-Schoenbergerju in Cukierju je treba razlikovati med internetom in masovnimi podatki, kljub temu da svetovni splet olajšuje zbiranje in izmenjavo podatkov. Internet je preoblikoval način komuniciranja. Masovni podatki so več kot le komunikacija in izražajo idejo, da se iz velike količine informacij lahko učimo o zadevah, ki jih sicer ne moremo razumeti z uporabo manjše količine podatkov. Masovni podatki bodo počasi spremenili naše razmišljanje o svetu, razumevanje dogajanja in sprejemanje odločitev in najbrž bomo odkrili, da so številni vidiki življenja bolj verjetnostni kot strogo določeni.

V 3. st. pr. n. št. so verjeli, da knjižnica v Aleksandriji hrani vse človeško znanje. Danes je na svetu toliko podatkov, da ima vsak človek 320-krat več informacij, kot jih je bilo shranjenih v celotni aleksandrijski zbirki, to je kar 1200 EB (eksabajtov). Če bi vse te informacije shranili na cedeje, bi pet ločenih stolpcev teh cedejev segalo do meseca!

Gre za največjo eksplozijo informacij v zgodovini. Pred letom 2000 je bila le četrtnina vseh shranjenih informacij v digitalni obliki. Preostanek je bil na papirju, filmu in drugih analognih medijih. Ker pa je količina digitalnih podatkov naraščala hitro – približno vsaka tri leta se je podvojila – se je stanje popolnoma spremenilo. Danes je shranjenih informacij, ki niso v digitalni obliki, manj kot dva odstotka!

*Datafication*² je značilna za masovne podatke. Gre za možnost podatkovnega predstavljanja mnogih vidikov sveta, ki prej nikoli niso bili količinsko opredeljeni. Lokacija je na primer najprej bila določena z izumom

zemljepisne dolžine in širine, danes pa s satelitskimi sistemi GPS. Ko so računalniki zamenjali knjigo, besede obravnavamo kot podatke. Na Facebooku so preobrnjena v podatke tudi prijateljstva in stališča, kot je "všeč mi je"!

Po novem naj bi vnašali v računalnik dovolj podatkov, da bi lahko sklepali po verjetnosti, da je, recimo, semafor zelen in ne rdeč, namesto da "učimo" računalnik, kot so strokovnjaki za umetno inteligenco neuspešno poskušali storiti že desetletja, kako upravljati avto ali prevajati med jeziki!

Uporaba velike količine podatkov zahteva tri temeljite spremembe v tem, kako dostopamo in uporabljamo podatke:

- zbiramo in uporabimo veliko podatkov namesto majhne količine ali statističnih vzorcev;
- prednost damo nenatančnim podatkom namesto natančnim: v vedno večjem številu situacij se lahko tolerira nenatančnost, ker koristi od uporabe velike količine podatkov različne kakovosti odtehtajo stroške uporabe manjše količine zelo natančnih podatkov;
- v mnogih primerih se odrečemo prizadevanjem, da bi odkrili vzroke, v prid izračunu korelacij; namesto da bi poskušali razumeti natančno, zakaj se je motor pokvaril ali zakaj je izginil neželen učinek zdravila, lahko raziskovalci namesto zbiranja in analize ogromnih količin informacij iščejo vzorce, ki lahko pomagajo pri napovedovanju dogodkov v prihodnosti; masovni podatki pomagajo odgovoriti na vprašanje "kaj" in ne "zakaj", kar je pogosto "dovolj dobro".

Prej smo delali z relativno majhnimi količinami podatkov, saj so bila orodja za zbiranje, organiziranje, shranjevanje in analizo podatkov slaba. Ljudje so prebirali informacije in jih reducirali na minimum, da bi jih lažje preučevali. Šlo je za razvoj moderne statistike, ki se je začel v poznem 19. stoletju in omogočil razumevanje zahtevnejših realnosti, tudi če je obstajalo le malo podatkov. Danes pa se je tehnično okolje obrnilo za 179 stopinj. Še vedno obstajajo omejitve, koliko podatkov lahko upravljamo, vendar je teh omejitev veliko manj kot nekoč in jih bo s časom še manj!

Podatke smo obdelovali z vzorčenjem, ki zmanjša stroške in potreben čas za obdelavo. Statistično vzorčenje temelji na ideji, da z določeno toleranco lahko sklepamo o celotni populaciji na podlagi majhne podmnožice naključno izbranih podatkov. Na osnovi naključno izbrane skupine več sto ljudi lahko napovemo glasovalno obnašanje v celotni državi. Pri preprostih vprašanjih ta postopek dobro deluje, ampak odpove, ko želimo vrtati navzdol v podskupine v vzorcu. Kaj pa, če želimo vedeti, kdo je za kandidata volil samsko žensko mlajšo od 30 let?

Nenadoma je naključni vzorec v veliki meri neuporaben, saj je lahko le nekaj ljudi s temi značilnostmi v vzorcu, to pa je premalo za smiselni sklep o tem, kako bo podmnožica glasovala. Ampak če uporabimo vse podatke ($n = \text{vse}$, da uporabimo izrazoslovje statistike), problem izgine!

Ta primer odpira še eno pomanjkljivost uporabe nekaterih podatkov namesto vseh. V preteklosti, ko so ljudje zbrali le malo podatkov, so se pogosto morali že na začetku odločiti, kaj zbirati in kako uporabiti. Danes, ko lahko zberemo vse podatke, ni treba vedeti vnaprej, kaj nameravamo uporabiti in za kaj. Seveda ni mogoče vedno zbrati vseh podatkov, je pa vedno bolj izvedljivo, da zajamemo veliko več kot le vzorec, ki bi bil reprezentančen za vse. V primeru masovnih podatkov ne gre samo za ustvarjanje nekoliko večjih vzorcev, ampak za izkoriščanje po možnosti čim več obstoječih podatkov o tem, kaj proučujemo. Še vedno potrebujemo statistiko, vendar se nam ni več treba zanašati na majhne vzorce.

Včasih moramo pustiti, da govorijo podatki sami. Načina razmišljanja o podatkih – od nekaj do vseh – povzroči tretjo spremembo: od vzročnosti do korelacije. To predstavlja premik od večnega poskusa razumeti globlje razloge, kako svet deluje, k preprostemu učenju o povezanosti med pojavi. Poznavanje vzrokov je zaželeno. Problem pa je, da je vzroke pogosto izjemno težko ugotoviti, in velikokrat, ko mislimo, da smo jih ugotovili, ni nič drugega kot iluzija. Vedenjska ekonomija je pokazala, da ljudje vidijo vzroke, tudi če jih še ni. Zato moramo biti še posebej pazljivi, da preprečimo varljive kognitivne pristranskosti.

Kot smo že rekli, dober primer, zakaj so korelacije, ki jih vidimo z masovnimi podatki, zelo koristne, pa tudi če so temeljni vzroki še vedno nejasni, predstavlja Googlov model za napovedovanje gripe.

Mnogi tehnologi verjamejo, da so masovni podatki nastali v digitalni revoluciji, ki jo je označila komercialna uporaba osebnih računalnikov v začetku osemdesetih let prejšnjega stoletja, ko je zahvaljujoč napredku na področju mikroprocesorjev in računalniškega spomina bilo mogoče analizirati in shranjevati informacije kot nikoli prej. Vendar je le navidezno tako. Računalniki in internet so z znižanjem stroškov za zbiranje, shranjevanje, obdelavo in izmenjavo informacij zagotovo omogočili pojav masovnih podatkov, ki so v bistvu le zadnji korak v prizadevanju človeštva, da razume in kvantificira svet.

Ko enkrat stvari pretvorimo v podatke, lahko preoblikujemo njihov namen, podatke pa podamo v novih oblikah vrednosti. Iz teh podatkov, ki jih je mogoče shraniti in analizirati, pa lahko izvemo več o stvareh,

česar prej nismo mogli vedeti, ker stvari ni bilo mogoče izmeriti enostavno in poceni.

Masovni podatki bodo korenito spremenili način, kako vlade delujejo, in naravo politike. Ko gre za ustvarjanje gospodarske rasti, za zagotavljanje javnih storitev ali za vojne, bodo tisti, ki lahko učinkovito izkoristijo masovne podatke, vedno v veliki prednosti. Doslej se najbolj vznemirljivo delo dogaja na občinski ravni, kjer je lažji dostop do podatkov in eksperimentiranje z informacijami. Mesta naj bi si prizadevala uporabljati masovne podatke – dober zgled je New York – za izboljšanje javnih storitev in nižanje stroškov.

Masovni podatki prav tako pomagajo povečati preglednost demokratičnega upravljanja. Okoli ideje odprtih podatkov je zraslo gibanje, ki presega zakone o svobodi informacij, ki so vsakdanji pojav v razvitih demokracijah. Zagovorniki pozivajo vlade, naj skrbijo za velike količine neškodljivih podatkov, ki bi bili na voljo javnosti. ZDA so bile v ospredju s svojo spletno stranjo Data.gov. in mnoge druge države so jim sledile.

Živimo v dobi, v kateri se, kot rečeno, proizvaja neverjetna količina informacij. Mayer-Schönberger in Cukier obravnavata v knjigi tudi glavna etična vprašanja v zvezi z masovnimi podatki, in sicer: Kaj se dogaja z zasebnostjo? Bodo stroji, ne pa ljudje, sprejemali naše glavne odločitve?

Ob istem času, ko so vlade spodbujale uporabo množičnih podatkov, so morale tudi poskrbeti za zaščito državljanov pred škodljivo prevlado trga. Podjetja, kot so Google, Amazon in Facebook, so kopičila velikanske količine informacij o vseh in o vsem. Zakoni, ki ščitijo pred monopolizacijo trgov blaga in storitev, kot so programska oprema in mediji, saj so velikosti trgov teh proizvodov relativno enostavni za oceno. Toda kako naj vlade uporabijo protimonopolne predpise na masovnih podatkih, na trgu, ki ga je težko definirati in ki nenehno spreminja obliko? Zasebnost bo postala še večja skrb, saj bo več podatkov zagotovo ogrozilo zasebne podatke, kar predstavlja negativno stran "big data", in se zdi malo verjetno, da bi jo preprečili zakoni in sodobna tehnologija.

Predpisi, ki urejajo masovne podatke, morda celo predstavljajo bojišče med državami. Evropske vlade so že preverjale Google po scenariju, ki spominja na protimonopolne ukrepe Evropske komisije proti Microsoftu, ki so se začeli pred desetletjem. Facebook bi lahko postal tarča podobnih ukrepov po vsem svetu, saj se tam kopiči ogromna količina podatkov o posameznikih.

Avtorja vidita uporabo modela masovni podatki tudi v dejavnostih vohunjenja. Lani je Edward Snowden s svojimi razkritji sprožil veliko prisluškovalno afero, ki v tem hipu pretresa sodobni svet zaradi negativnih posledic za številne države in množice posameznikov, ki so tarča obveščevalnih služb na čelu z ameriško varnostno agencijo (National security agency – NSA)!

Ni naključje, da bodo masovni podatki in *dataability* vodilna tema srečanja CeBIT 2014, toliko sta namreč ti temi pomembni za sodobno globalno informacijsko družbo!

Po napovedih skupine Gartner 2012 več kot 85 % organizacij ne bo moglo učinkovito izkoristiti priložnosti masovnih podatkov do leta 2015, in sicer predvsem zato, ker se soočajo z:

- velikansko količino – zaposleni, stranke, partnerji in konkurenti nenehno generirajo terabajte in petabajte dragocenih podatkov;
- raznolikostjo – masovni podatki prihajajo v različnih, strukturiranih in nestrukturiranih oblikah iz mnogih (notranjih in zunanjih) virov;
- dinamičnostjo – masovni podatki rastejo in se hitro spreminjajo v realnem času in zato zahtevajo krajše cikle predelave.

Tistim, ki sem jih nagovoril, da bodo prebrali predstavljeno knjigo, želim dobrodošlico v revoluciji masovnih podatkov!

Opombe

- 1 V slovenščini še ni ustreznega izraza za tujko, ki združuje v sebi izraze za masovne podatke, zmožnost, trajnost in odgovornost (angl. big data, ability, sustainability, responsibility), zato jo v besedilu puščamo neprevedeno in v ležečem tisku (op. lekt.).
- 2 V slovenščini še ni ustreznega izraza za tujko, ki opisuje prenos vsega v numerično obliko ali podatek, zato jo v besedilu puščamo neprevedeno in v ležečem tisku (op. lekt.).

Tvrtko-Matija Šercar