# Predgovor



Knjige o velikem podatkovju se uvrščajo v eno od dveh kategorij: ali ne ponujajo nikakršne razlage, kako stvari dejansko delujejo, ali pa so visoko strokovni učbeniki matematike, primerni le za podiplomske študente. Namen te, ki jo držite v rokah, je ponuditi drugačno knjigo, uvod v to, kako veliko podatkovje deluje in spreminja svet okoli nas, kakšen vpliv ima na naš vsakdan in kakšen vpliv na poslovni svet.

Pod besedo podatki smo včasih razumeli dokumente in papirje, morda opremljene z nekaj slikami. Zdaj podatki pomenijo veliko več. Spletne strani družabnih omrežij vsako minuto proizvedejo ogromne količine podatkov v obliki slik, videoposnetkov in filmov. S spletnim nakupovanjem ustvarjamo podatke ob vsakem vnosu naslova in podrobnosti o kreditni kartici. Zdaj smo na točki, ko zbiranje in shranjevanje podatkov rasteta s hitrostjo, ki je bila le nekaj desetletij nazaj nepredstavljiva, a kot bomo videli v nadaljevanju, nove tehnike analize podatkov slednje pretvarjajo v uporabne informacije. Med pisanjem knjige sem prišla do spoznanja, da se o velikem podatkovju ne moremo smiselno pogovarjati, ne da bi se pogosto vračali k temu, kako veliki komercialni akterji te podatke zbirajo, shranjujejo in analizirajo. Raziskovalni oddelki v podjetjih, kot sta Google in Amazon, so odgovorni za mnoge razvojne dosežke na področju velikega podatkovja, zato jih bom večkrat omenila.

Prvo poglavje bralca seznani z raznolikostjo podatkov na splošno, preden razloži, kako je zaradi digitalne dobe prišlo do spremembe načina definiranja podatkov. Veliko podatkovje predstavljam preprosto, s pomočjo ideje o eksploziji podatkov, v katero so vključeni računalniška znanost, statistika in stičišče med njima. V poglavjih 2–4 sem za pomoč pri razlagi nekaterih

novih metod, ki jih zahteva veliko podatkovje, zelo pogosto uporabljala grafične prikaze. Drugo poglavje razišče, zakaj je veliko podatkovje posebno, in nas tako privede do podrobnejše definicije. V 3. poglavju razpravljamo o problemih, povezanih s shranjevanjem in z upravljanjem velikega podatkovja. Večina nas ve, da je treba na osebnem računalniku ustvarjati rezervne kopije podatkov. A kako to storiti ob kolosalnih količinah podatkov, ki se danes ustvarjajo? Da bi odgovorili na to vprašanje, si bomo ogledali shranjevanje podatkovnih baz in idejo o porazdelitvi nalog med skupine računalnikov. Četrto poglavje trdi, da je veliko podatkovje uporabno le, če lahko iz njega izluščimo uporabne informacije. Za občutek, kako se podatki pretvarjajo v informacije, bo dovolj poenostavljena razlaga različnih dobro uveljavljenih tehnik.

Nato se lotimo podrobnejše razprave o praktični uporabi velikega podatkovja, ki se v 5. poglavju začne z vlogo velikega podatkovja v medicini. Šesto poglavje analizira poslovne prakse s pomočjo študij primerov Amazon in Netflix, pri čemer vsaka študija osvetljuje različne lastnosti trženja z uporabo velikega podatkovja. Sedmo poglavje se loti nekaterih varnostnih težav v zvezi z velikim podatkovjem in pomembnosti šifriranja. Kraja podatkov je postala velik problem, zato si ogledamo nekaj primerov, ki so se pojavili v poročilih, vključno s primeroma Snowden in WikiLeaks. Poglavje se zaključi s prikazom, kako se mora veliko podatkovje vse bolj varovati pred kibernetskim kriminalom. V zadnjem, osmem poglavju obravnavamo, kako veliko podatkovje z razvojem sofisticiranih robotov in vlogo le-teh na delovnih mestih spreminja družbo, v kateri živimo. Knjiga se zaključi z razmislekom o pametnih domovih in pametnih mestih prihodnosti.

V kratkem uvodu ni mogoče omeniti vsega, zato upam, da si bo bralec pomagal s priporočili v razdelku Nadaljnje branje in nadalje raziskal, kar mu je med branjem vzbudilo zanimanje.

# Zahvale

Ko sem mu omenila, da bi se mu rada zahvalila za njegov prispevek k tej knjigi, je Peter predlagal: »Rada bi se zahvalila Petru Harperju, brez njegove predane uporabe orodja za preverjanje črkovanja bi bila to drugačna knjiga.« Petru se prav tako zahvaljujem za strokovno kuhanje kave in smisel za humor! Ta podpora je že sama po sebi neprecenljiva, ampak Peter je naredil še veliko, veliko več, in upravičeno lahko rečem, da ta knjiga brez njegove neomajne spodbude in konstruktivnih prispevkov ne bi bila napisana.

Dawn E. Holmes, april 2017

# 1. poglavje
# Eksplozija podatkov



## Kaj so podatki?

Leta 431 pr. n. št. je Šparta napovedala vojno Atenam. Tukidid v poročilu o vojni opiše, kako je oblegovana platejska vojska, ki je bila zvesta Atenam, načrtovala pobeg s plezanjem čez obzidje okoli Plateje, ki ga je pod vodstvom Špartancev zgradila peloponeška vojska. Za uspešen pobeg so morali poznati višino obzidja, da bi lahko izdelali primerno dolge lestve. Večino peloponeškega obzidja je prekrival grob omet, vendar pa so našli predel, kjer so bile opeke še vedno jasno vidne, in tako je veliko število vojakov dobilo nalogo prešteti vrste teh nezaščitenih opek. Ker so delali s take razdalje, kjer so bili varni pred napadom sovražnika, je neizbežno prišlo do napak, a kot je razložil Tukidid, so opravili veliko štetij in rezultat, ki se je pojavil največkrat, je bil pravilen. Ker so Platejci poznali velikost uporabljenih lokalnih opek, so to največkrat ponovljeno štetje, ki bi mu danes rekli *modus*, uporabili za izračun višine obzidja in izdelali ustrezno dolge lestve za preplezanje obzidja. Vojski več sto mož je tako uspelo pobegniti in dogodek bi prav lahko imeli za najveličastnejši primer zgodovinskega zbiranja in analize podatkov. A kot bomo videli, so zbiranje, shranjevanje in analiza podatkov mnogo stoletij starejši od Tukidida.

Na palicah, kamnih in kosteh smo našli zareze, ki segajo daleč nazaj, vse do dobe mlajšega paleolitika. Te zareze naj bi predstavljale podatke, shranjene v obliki črtičnih zapisov, čeprav je to še vedno predmet akademske razprave. Morda najslavnejši primer je kost iz Ishanga, ki so jo našli v Demokratični republiki

Kongo leta 1950 in ki naj bi bila po ocenah stara 20.000 let. Kost z zarezami so si razlagali različno, kot kalkulator ali koledar, čeprav imajo nekateri rajši razlago, da so zareze tam za boljši oprijem. Kost iz Lebomba, odkrita leta 1970 v Svaziju, je še starejša in sega v čas okoli 35.000 pr. n. št. Z devetindvajsetimi povprek vrezanimi črticami je ta fragment pavijanove fibule neverjetno podoben koledarskim palicam, ki jih v daljni Namibiji uporabljajo Bušmani, kar kaže, da je bilo zarezovanje morda zares metoda, s katero so pred toliko tisočletji spremljali za svojo civilizacijo pomembne podatke.

Medtem ko o razlagi zarezanih kosti še vedno ugibamo, pa vemo, da je eden prvih primerov dobro dokumentirane uporabe podatkov popis, ki so ga opravili Babilonci leta 3800 pr. n. št. V njem so, da bi lahko dobili potrebne informacije za izračun davkov, sistematično dokumentirali število prebivalstva in dobrine, kot sta mleko in med. Tudi stari Egipčani so uporabljali podatke v obliki na les ali papirus zapisanih hieroglifov, s katerimi so beležili dostavo blaga in nadzirali davke. A zgodnji primeri uporabe podatkov nikakor niso omejeni na Evropo in Afriko. Inki in njihovi južnoameriški predhodniki, ki so vneto beležili statistiko za davčne in trgovske potrebe, so uporabljali *kipu*, sofisticiran in zamotan sistem barvnih vrvic z vozli kot desetiško zasnovan računovodski sistem. Te zavozlane vrvice, narejene iz živo obarvanega bombaža ali kamelje volne, segajo v tretje tisočletje pr. n. št., in čeprav jih je, kolikor vemo, manj kot tisoč preživelo špansko invazijo in nadaljnje poskuse uničenja, predstavljajo enega prvih znanih primerov sistemov za shranjevanje velikanskega števila podatkov. Danes razvijamo računalniške algoritme, da bi dešifrirali polni pomen *kipuja* in bolje razumeli, kako so ga uporabljali.

Čeprav za te zgodnje sisteme lahko rečemo, da uporabljajo podatke (ang. *data*), in jih tako tudi opisujemo, je »data« v