Big Data: kako smo došli do *Velikih podataka* i kamo nas oni vode

Sažetak: *Količina informacija nastala u razmaku od otprilike 1200 godina, od osnivanja Carigrada pa do otkrića Gutenbergova tiskarskoga stroja, udvostručila se tek nakon 50 godina. Danas postojeću količinu informacija udvostručimo svake 3 godine pa je već mjerimo u eksabajtima. Tako velike količine podataka promijenile su i način na koji koristimo, ali i obrađujemo podatke. Sa sigurnošću možemo reći da smo u tijeku jedne nove velike revolucije koja ima i svoje prigodno ime Big Data – Veliki podatci. Iako su termin osmislili znanstvenici iz područja poput astronomije i genomije, Veliki podatci su posvuda. Oni su istovremeno i resurs i alat čiji je glavni zadatak informiranje. Ali, koliko god nam mogu pomoći bolje razumjeti svijet oko nas, ovisno o tome kako se njima upravlja i tko njima upravlja, mogu nas odvesti i u nekome drugome smjeru. Iako nam se brojke koje se vežu uz Velike podatke mogu u ovom trenutku činiti enormnima, moramo biti svjesni činjenice da će količina onoga što možemo prikupiti i obraditi uvijek biti samo djelić informacija koje zaista postoje na svijetu (i oko njega). No, od nečega moramo početi!*

Ključne riječi: *Big Data*, Veliki podatci, prikupljanje podataka, datafikacija, izvori podataka, grobnice podataka, podatkovni ispušni plinovi, podatkovni znanstvenici.

1. Što krije pojam 'Veliki podatci'

Veliki podatci označavaju eru u kojoj ćemo moći kvantificirati svijet i razumjeti ga. Na početku ove ere možemo reći da smo dobili i novu disciplinu. Iako se na prvi pogled nova disciplina može činiti redundantnom, moramo tu ideju ponovo razmotriti uzevši u obzir i Dieboldovo (2013) upozorenje da je cjelovitost Velikih podataka kao discipline puno veća nego zbroj njegovih dijelova. Ova disciplina otvara neke nove putanje, neka nova otkrića, koja do sada nismo mogli niti zamisliti.

Na samom početku ovoga rada, htjela bih se osvrnuti na upotrebu velikoga početnog slova u terminu Veliki podatci (*Big Data*). Iako je takvo pisanje protivno pravopisnim pravilima standardnoga hrvatskog jezika, vjerujem da se ovdje radi o iznimci. Naime, kao što ni Aleksandar Veliki, ni Antun Veliki nisu nadimak 'Veliki' dobili pri rođenju, a niti zbog svoje visine, već zbog svojih velikih djela, tako se i u ovom slučaju pridjev 'veliki' odnosi na velike mogućnosti koje ovi podatci otvaraju.

Svoj strmi uspon Veliki podatci doživjeli su tek 2008. godine, no, termin 'Big Data' pojavljuje se u akademskim i ne-akademskim radovima od sredine 1990-ih godina. Teško je utvrditi tko ga je prvi osmislio i počeo upotrebljavati. Zasad zasluge idu Johnu Masheyu, kao najvjerojatnijemu autoru termina, koji je krajem 1990-ih *„samo tražio jednostavan naziv za cijeli niz tema koji bi najbolje prenio sliku da se granice računala mijenjaju i napreduju*" (Lohr, 2013). On je u to vrijeme radio za Silicon Graphics, veliku kompaniju koja se bavila računalnom grafikom, a njezinim uslugama se koristio Hollywood za izradu specijalnih efekata, ali i agencije za video nadzore. Bila je to upotreba novih tipova podataka u velikim razmjerima.

Ova bi priča mogla imati svoje čvrsto uporište u potrazi za autorom termina, uzmemo li u obzir i da Mayer-Schönberger i Cukier (2013) također vjeruju kako su termin mogli osmisliti samo oni znanstvenici koji su imali pristup velikim podatcima kao npr. znanstvenici iz područja astronomije i genomije, i to tek u onome trenutku kad je količina podataka *prerasla* memoriju kojom računala računaju. Trebalo je naći nove načine (alate) za obradu podataka koji analiziraju velike količine podataka koji ne moraju nužno biti *pohranjeni* u klasične tablice baze podataka, poput Googleova MapReducea i Yahoova Hadoopa.

Prvu definiciju Velikih podataka kao fenomena nalazimo pak kod Diebolda (2000) koji navodi: „*U zadnje je vrijeme dosta dobre znanosti, bez obzira je li u pitanju fizika, biologija ili sociologija, bilo prisiljeno suočiti se – od čega je često i profitirala – s fenomenom Velikih podataka. Veliki podatci odnose se na eksploziju u količini (a katkad i kvaliteti) dostupnih i potencijalno relevantnih podataka, uglavnom kao posljedica skorih i besprimjerenih napredaka u tehnologiji zapisivanja i pohranjivanja podataka.“*

Američka nezavisna državna agencija, **National Science Foundation**[1], koju je osnovao Kongres SAD-a 1950. godine u svrhu promoviranja napretka u znanosti, unapređivanja nacionalnog zdravlja, prosperiteta i blagostanja, definira pojam Velikih podataka na sljedeći način: „Izraz 'Veliki podatci' u ovom se

[1] Financijski podupiru 24% svih istraživanja koja podupire Vlada SAD-a, a koja se izvode na američkim sveučilištima.

izvještaju odnosi na velike, raznolike, kompleksne, longitudinalne i/ili distribuirane podatkovne skupove koje su generirali strojevi, senzori, mrežne transakcije, elektronička pošta, video, zapisi klikova na mrežnim stranicama i/ili svi ostali digitalni izvori dostupni danas i u budućnosti".

Tek krajem 2008. termin su prihvatili i počeli koristiti vodeći računalni znanstvenici iz *Computing Community Consortiuma* objavivši izvještaj *Big Data Computing: Creating revolutionary breakthroughs in commerce, science, and society* dostupan na njihovim stranicama.

Veliki se podatci obično opisuju služeći se pojmom 3V-a koji se odnosi na **v**olumen (veeeelika količina podataka), **v**arijantnost (raznolikost tipova podataka: tradicionalne baze podataka, fotografije, dokumenti) i **v**elocitet tj. brzinu kojom se akumuliraju novi podatci (iz sličnih izvora podataka, iz prethodno arhiviranih podataka, iz podataka koji stalno pridolaze iz različitih izvora – engl. *streamed data*), ali i brzina kojom se očekuje da pristigli podatci budu dostupni za analizu (Reeve, 2013; Simon, 2013; Berman, 2013). Upravo je postojanje svih triju vrijednosti ono što razlikuje „Velike podatke" od „puno podataka", ali i ono zbog čega ova vrsta podataka zahtijeva nove metode za oblikovanje, rukovanje i analiziranje.

Iako je i količina podataka sama po sebi bitna, ono što je puno bitnije je biti u stanju prikupiti još podataka, jednostavnim putem. No, čak i u eri Velikih podataka, pojam količine ima različitu vrijednost ovisno o znanosti koja koristi podatke. Tako npr. 200 GB podataka u ekonometriji može se činiti zaista jako velikim dok istu količinu podataka fizičari mogu promatrati kao mali skup podataka jer njima „veliko" sad već znači nešto veće od 10^{15} bajta. Ali i jedni i drugi složili bi se sa Siegelom (Siegel, 2013:78): „*Size doesn't matter. It's the rate of expansion*".

S vremenom su se u jednadžbu pokušali uključiti i neki novi V-ovi (Berman, 2013; Simon, 2013) poput **v**izije (nove ideje sa starim podatcima), **v**erifikacije (mogućnost provjeravanja zadovoljavaju li podatci određeni skup specifikacija - ovaj se proces odvija prije nego što se podatci podvrgnu bilo kakvoj analizi), **v**alidacije (provjera je li svrha podataka zadovoljena i konzistentna tj. mogu li se isti točni i prikladni zaključci dobiti iz istoga skupa podataka bez obzira na broj ponavljanja analiza – ovaj se proces odvija nakon što su podatci bili podvrgnuti analizi) ili pak **v**arijabilnosti i **v**jerodostojnosti, koji se, barem za sad, nisu proširili u upotrebi.

2. Što nisu Veliki podatci

Kao što smo vidjeli u prethodnom poglavlju, pojam usko vezan uz Velike podatke je količina podataka. No, pri tome se ne misli na puno podataka tj. na klasičnu bazu podataka koja je s vremenom narasla na veliki broj zapisa. Pojam „puno podataka" odnosi se na velike zbirke zapisa jednostavnoga formata kakav bi recimo bio popis svih studenata Filozofskoga fakulteta s podatcima o njihovom prebivalištu i prethodnom obrazovanju.

Kad govorimo o količini Velikih podataka govorimo o jednoj novoj informacijskoj paradigmi (Birnhack, 2013). Razliku (njih čak 10) između *Velikih podataka* (VP) i *standardnih podataka* (SP) najbolje je možda opisao Berman (2013):

1. ***ciljevi*** – SP daju odgovor na specifično pitanje s unaprijed određenim ciljem; VP imaju odgovore na raznovrsna pitanja s prilagodljivim ciljem
2. ***lokacija*** – SP se uglavnom nalaze unutar jedne organizacije; VP se mogu nalaziti rascjepkani na različitim lokacijama
3. ***struktura i sadržaj podataka*** – SP su strukturirani podatci s domenom iz jednoga područja, ujednačene forme; VP su nestrukturirani podatci (tekstni dokumenti, slike, filmovi, zvučni zapisi itd.) koji mogu dolaziti iz različitih domena s dodatnim vezama na podatke iz drugih izvora
4. ***priprema podataka*** – SP priprema uglavnom korisnik tih podataka; VP priprema mnogo ljudi jer su i podatci iz različitih izvora dok su korisnici podataka rijetko ljudi koji su podatke pripremili
5. ***životni vijek*** – SP imaju ograničen vijek postojanja (u prosjeku 7 godina po završetku projekta); VP sadrže podatke bez ograničenog životnog vijeka jer se većina integrira u nove projekte koji koriste VP
6. ***mjerenja*** – SP se uglavnom mjere s pomoću jednoga protokola dok se VP mjere različitim protokolima (upravo je utvrđivanje kvalitete[2] podataka kod VP najzahtjevniji posao)
7. ***reproduciranje*** – projekti koji koriste SP daju se lako reproducirati; projekti koji upotrebljavaju VP mogu se rijetko kad reproducirati
8. ***financijsko ulaganje*** – financije uložene u projekte sa SP-om relativno su male, za razliku od financija uloženih u projekte s VP-om koje mogu dovesti i do bankrota (vidi poglavlje 5.5)
9. ***introspekcija*** – pojedinačni standardni podatci mogu se identificirati s pomoću njihove lokacije određene retkom i stupcem unutar tablice, međutim, kod VP-a procedura za identifikaciju puno je složenija i ona

[2] Pri unosu podataka u sustav, između 2% i 30% podataka bude pogrešno upisano, ovisno o čimbenicima poput tipa podatka (tekstni ili brojevni zapis) ili duljini zapisa.

se (barem kod dobro oblikovanih VP resursa) može ostvariti s pomoću tehnike introspekcije

10. ***analiza*** – kod SP-a analiza se može vršiti nad svim podatcima istovremeno, kod VP-a analiza se odvija u koracima (osim u paralelnoj analizi koja se istodobno odvija na više računala) na način da se podatci izvlače, pregledavaju, smanjuju, normaliziraju, transformiraju, vizualiziraju, interpretiraju te ponovo analiziraju različitim metodama.

3. Prikupljanje podataka

Počela je nova potraga za blagom: svaki skup podataka ima neku unutarnju, skrivenu, još neotkrivenu tajnu – BLAGO, a zadatak je da ih se sve pronađe!

Veliki podatci prikupljaju se i na temelju podatkovnih tragova web pretraživanja, komunikacija na društvenim mrežama, podataka koje prikupljaju senzori, ali i podataka koje prikupljaju nadgledni sustavi. Bitno je napomenuti da samo prikupljanje podataka ne jamči poslovni uspjeh, ali zato otvara prozor u ono što je moguće.

3.1. Početci prikupljanja podataka

Veliki podatci omogućavaju uvid i razumijevanje relacija između dijelova informacija koje smo sve donedavno samo pokušavali dokučiti. Prije smo radili s malom količinom podataka jer su nam alati za prikupljanje, organiziranje, pohranjivanje i analiziranje toliko dopuštali.

Bila je to umjetno stvorena zapreka za analizom SVEGA koju nam je nametnuo tehnološki razvoj vremena kako u doba starih Egipćana i Kineza koji su među prvima počeli prikupljati *velike* podatke o svojim podanicima, preko britanskog popisa svog stanovništva objavljenog u Knjizi sudnjeg dana (engl. *Doomsday Book*) pa sve do nešto bližih nam Ureda za popis stanovništva. Ono što je karakteristično za ta prikupljanja je da su dugo trajala, podatci su bili djelomični, a do samog kraja prebrojavanja, većina ih je bila i nevažeća.

3.2.Nova era prikupljanja podataka

Danas se o svakome od nas prikuplja više informacija nego ikad prije. Ulaskom u e-eru, prikupljanje podataka prestala je biti privilegija vladinih ustanova. Podatke sada prikupljaju pa... skoro svi, počevši od državnih agencija, tajnih službi, osiguravajućih društava pa do Amazona, Googlea, Twittera, Facebooka[3], mobilnih operatera.

[3] Procjenjuje se da svaki četvrti korisnik Facebooka daje lažne informacije o sebi.

Podatke prikupljaju i čitači e-bookova koji na taj način dobivaju uvid u navike ljudi koji ih čitaju (koliko dugo čitaju stranicu, gdje čitaju, kako brzo okreću stranicu, jesu li odustali od čitanja, jesu li napravili bilješku na margini ili nešto podcrtali...). Ovi su podatci od velike važnosti i izdavačima i autorima knjiga. Tako su npr. Barnes&Noble prikupljali podatke preko svojih Nook čitača e-knjiga i otkrili kako ljudi odustaju na pola čitanja dugih knjiga koje nisu fikcija. Zato su ponudili serijal *Nook snaps* s kratkim osvrtima na teme poput zdravlja i trenutnih događaja (Mayer-Schönberger, Cukier, 2013).

Čak i stranice na kojima se nudi on-line obrazovanje (Udacity, Coursera, edX) prate mrežnu interakciju studenata kako bi otkrili bolje pedagoške metode poučavanja.

No, ne sadrže svi Veliki podatci osobne informacije. Tu spadaju senzorni podatci iz rafinerija, strojevni podatci iz tvornica (praćenje rada strojeva), podatci o inventaru skladišta ili knjižnica, vremenske baze, astronomski podatci koje prikupljaju teleskopi (*Sloan Digital Sky Survey* teleskop u New Mexicu), podatci vezani uz sljedivost u farmaceutici i prehrani, dekodiranje ljudskog genoma, GPS podatci o kretanju kamiona ili taxi službi i sl.

3.3. Datafikacija

Pojam datafikacije, kojeg uvode Mayer-Schönberger i Cukier (2013), odnosi se na proces kojim se prikupljaju informacije o svemu što nas okružuje (GPS lokacija inhalatora čijom se aktivacijom prikupljaju podatci o okolišu koji je neadekvatan za astmatičare, praćenje podrhtavanja tijela kod neuroloških pacijenata kojom liječnici prate stanje pacijenta), a potom transformiraju u format podatka kako bi se mogle prebrojati i dalje analizirati.

Tako se u samom središtu revolucije informacijske tehnologije, naglasak s tehnologije ponovo vraća na samu informaciju, odnosno, nastavlja se čovjekovo nastojanje da izmjeri, zapiše i analizira svijet u kojem živi. I dok su telefonske i internetske mreže poboljšale i ubrzale protok informacija, cilj datafikacije je obogatiti naše poimanje svijeta.

Nakon datafikacije, mogućnosti upotrebe prikupljenih informacija ograničena je samo našom dovitljivošću. Mogli bismo čak reći: *Veliki podatci i mašta, mogu svašta!*

4. Podatci, podatci, posvuda podatci

Izloženi stalnom prikupljanju, davanju i pregledavanju podataka, vjerojatno se sad već mnogi pitaju koliko li je sad podataka na svijetu. Isto pitanje mučilo je i Martina Hilberta koji je 2007. odlučio izmjeriti količinu informacija koje nas

okružuju (knjige, slike, e-mail, redovna pošta, fotografije, glazba, video, igre, tel. razgovori, auto-navigacijski sustavi, TV i radio emisije) i izračunao da je te godine postojalo 309 eksabajta (10^{18}) pohranjenih podataka (Hilbert, Lopez, 2012).

Slikovito prikazano, da su svi ti podatci u knjizi, prekrili bi cijeli SAD – 52 puta, a da su na CD-u, s njima bi se moglo podići 5 stupova od Zemlje do Mjeseca. Gledano povijesno, za količinu *proizvedenih* informacija od osnivanja Konstantinopola do otkrića Gutenbergova stroja (oko 1200 godina) trebalo je 50 godina tiskanja da se ta količina informacija udvostruči, a sad je udvostručimo svake 3 godine. To nas je dovelo do toga da je 2000. g. ¼ pohranjenih informacija u svijetu bila digitalna (¾ na papiru, filmu, LP pločama, magnetnim vrpcama...). Samo 7 godina poslije svega 7% pohranjenih informacija bilo je analogno – sve ostalo je bilo digitalno.

U svijetu astronomije poznato je da je Sloan Digital Sky Survey teleskop, izgrađen u New Mexicu, samo u prvih nekoliko tjedana prikupio više podataka nego što je prikupljeno u cijeloj povijesti astronomije. U periodu od 2000. do 2010., arhivirano je 140 terabajta informacija prikupljenih ovim teleskopom (Mayer-Schönberger, Cukier, 2013). Čini li vam se to kao jako puno podataka? Spremite se za još više podataka jer se 2016. planira puštanje u pogon *Large Synoptic Survey* teleskopa u Čileu koji bi 140 terabajta informacija trebao prikupiti svakih 5 dana.

Slična priča vezana je i uz dekodiranje ljudskoga genoma. Od 2003. do 2013. obrađeno je 3 milijarde osnovnih parova, no od 2013., ti se podatci mogu obraditi unutar jednoga dana. To nas možda i ne iznenađuje, uzmemo li u obzir da samo Google dnevno obradi 24 petabajta (10^{15}) podataka što je i tisuće puta više nego što je tiskanoga materijala u Američkoj kongresnoj knjižnici, da Facebook dnevno prikupi 3 milijarde (10^{9}) *like-ova* i/li komentara, a svaki sat preuzme oko 10 milijuna slika (10^{6}) dok na YouTube svake sekunde 800 milijuna korisnika pohrani video u trajanju više od jednoga sata.

Dok čitate ove podatke, može vam se učiniti kao da je sve što nas okružuje pretvoreno u podatak. Pa, *gotovo* sve bilo bi točnije jer, kao što i Berman (Berman, 2013:xv) skreće pozornost, posve je jasno da kad bismo išli opisati baš sve u našemu svemiru, trebao bi nam još jedan dodatni svemir u koji bismo mogli pohraniti te podatke, s tim da bi veličina toga novoga svemira trebala biti mnogo, mnogo veća od ovoga koji opisujemo.

4.1 Pristup podatcima

Vlade su prve prikupljale Velike podatke (i prije e-ere), ali su ih čuvale od pogleda ostalih. Vremena se mijenjaju pa tako i vlade koje polako otvaraju prozore u svijet podataka koje su prikupile tijekom godina. O točnosti tih podataka, kao i o odgovornosti zbog objavljivanja netočnih podataka koje je prikupila vlada, trebalo bi se posebno prodiskutirati (Washington, 2014).

Među prvim državama koje su dopustile pristup svojim podatcima bile su SAD. Predsjednik Obama izdao je 2009. direktivu da se što je moguće više podataka otvori javnosti preko stranice data.gov koja predstavlja otvoreni repozitorij s podatcima federalne vlade. Britanska je vlada otvorila *Open Data* Institut koji vodi Tim Berners-Lee (izumitelj web-a) s ciljem promoviranja novih načina upotrebe podataka i pronalaženja načina na koje će vladini dokumenti postati slobodno dostupni. EU je najavila inicijativu otvorenih podataka koja bi se mogla proširiti starim kontinentom. Australija, Brazil, Čile i Kenija počeli su se služiti strategijom otvorenih podataka.

Hrvatski sabor donio je također 2013. Zakon o pravu na pristup informacijama kojim se omogućava i osigurava pravo na pristup informacijama i ponovna uporaba informacija koje posjeduju tijela javne vlasti (tijela državne uprave, druga državna tijela, tijela jedinica lokalne i područne samouprave...).

Osim država, i mnogi su gradovi učinili slično kao i neke međunarodne organizacije poput Svjetske banke.

4.2 Trgovanje podatcima

Zadnjih godina osnovale su se razne tvrtke koje omogućavaju trgovanje podatcima, kao npr.:

- *DataMarket* (Island) – od 2008. omogućava pristup besplatnim bazama podataka iz različitih izvora (Ujedinjeni narodi, Svjetska banka, Eurostat) i zarađuje na postocima preprodaje podataka marketinškim tvrtkama
- *Factual* – omogućava pristup velikim bazama za čije je kompajliranje potrebno više vremena
- *Windows Azure Marketplace* – prodaje podatke (dajući pritom prednost visoko-kvalitetnim podatcima).

Tvrka *Import.io* savjetuje licenciranje podataka kako ih drugi ne bi mogli samo besplatno prikupiti s mreže. No, prema Birnhacku (2013) pravno gledajući,

kako broj podataka u bazi raste, sve je teže definirati tko je vlasnik podataka, tko baze, i koji su točno zakoni o privatnosti, a koji o autorskim pravima[4] na snazi.

FutureICT projekt temelji svoje istraživanje na interdisciplinarnosti Velikih podataka uključujući pristupe, metode i istraživače iz svih područja. Jedna je od glavnih uloga projekta *FuturICT* unaprjeđenje i primjena metoda ***Privacy by Design*** – možda baš oni uspiju pronaći prave odgovore na pitanja privatnosti u društvu Velikih podataka.

4.3. Smanjivanje podatkovnih vrijednosti

Iako je cijena pohranjivanja podataka pala, pa se čini kao da se stari podatci trebaju koristiti do beskonačnosti, to baš i nije tako. S vremenom, većina podataka (na različite načine i različitim tempom) ipak gubi neke od svojih vrijednosti pa njihova upotreba, umjesto da dodaje vrijednost, ustvari smanjuje vrijednost novih podataka.

Problem je odlučiti koji podatci su za upotrebu, a koji ne. Ne preporuča se odluku donositi samo na temelju vremenskoga čimbenika. Mnogi stoga izrađuju sofisticirane modele za određivanje (ne)korisnih podataka, poput Amazona koji se koristi posebno izrađenim modelom za predlaganje novih sadržaja svojim korisnicima (Mayer-Schönberger, Cukier, 2013; Simon, 2013).

5. Izvori Velikih podataka

Podatci su gorivo za informacijsko društvo bez kojega ne bi bile moguće inovacije o kojima današnji čovjek ovisi. Veliki su podatci u središtu moderne znanosti i poslovanja. Kolika je važnost Velikih podataka možda dobro ilustrira činjenica da je Engleska vlada 2013. g. dodijelila 189 milijuna funti za istraživanja u području Velikih podataka. Sintetička biologija, kao prva sljedeća disciplina po količini dodijeljenih novaca, dobila je 88 milijuna funti.

Među izvore Velikih podataka spadaju i baze s otiscima prstiju, DNA baze, zapisi aviokompanija, zapisi obrazovnih ustanova, transakcije kreditnih kartica, Facebook stranice, e-mailovi, zapisi ustanova javnog zdravstva...

U eri Velikih podataka vrijednost podatka leži u zbroju svih njegovih mogućih primjena, a te je vrijednosti moguće osloboditi na više načina: novom upotrebom „starih" podataka, spajanjem različitih skupova podataka, višenamjenskom upotrebom podataka, upotrebom podatkovnih ispušnih plinova i otvaranjem grobnica podataka. Slijede primjeri izvora pojedinačnih primjena.

[4] Nestrukturirane baze podataka (karakteristične za Velike podatke) ne podliježu zakonu o autorskim pravima (Birnhack, 2013).

5.1. Nova upotreba „starih" podataka

Među najranijim je upotrebama starih podataka (engl. *data reuse*) možda Matthew Fontaine Maury koji je još sredinom 19. stoljeća iz zapisa starih kapetanskih dnevnika otkrio kretanja oceanskih strujanja.

Od 2012. godine, IBM, *Honda, Pacific Gas* i *Electric Company* u Kaliforniji zajedničkim snagama traže podatke o razmještaju električnih stanica za punjenje električnih automobila. Želja im je otkriti kad bi i gdje električni automobili mogli ostati bez struje. Istraživanje su proveli na temelju podataka poput razine napunjenosti autobaterije, pozicije automobila, doba dana, broja dostupnih punionica na najbližim električnim punionicama. Tim su podatcima još dodali i potrošnju struje iz električne mreže i povijesne zapise o modelima potrošnje struje. Na temelju prikupljenih podataka, IBM mogao je izgraditi model optimalnih mjesta za izgradnju električnih stanica. Kad se stanice izgrade, u model će se moći dodati i podatci o trenutnom vremenu, vremenska prognoza, razlika u cijenama u obližnjim električnim stanicama. Oni su iskoristili **primarne informacije iz podataka** poput indikatora razine baterije koji javlja vozaču kad je vrijeme za punjenje i podatke o korištenju energije, koje prikuplja električna mreža kako bi se održala stabilnom. Potom su iz istih podataka izračunali **sekundarne informacije** kao što su određivanje vremena i mjesta za ponovno punjenje baterije, odnosno određivanje mjesta najprikladnijih za izgradnju električnih stanica. Svojim su izračunima dodali i **dodatnu upotrebu informacija** iz GPS podataka o poziciji automobila i povijesne podatke o praćenju potrošnje energije u mreži.

Hewlett-Packard iskoristio je povijesne podatke o svojim zaposlenicima kako bi izradio sustav za predviđanje rizika od otkaza – tj. za svakoga zaposlenoga izračunat je postotak vjerojatnosti da će ta osoba dati otkaz. Izvještaju ima pristup samo mali broj menadžera koji uz broj vide još i popratnu informaciju o kontekstu na temelju kojega je lakše razumjeti izračunati postotak. Ti podatci pomažu menadžerima u osmišljavanju strategija kojima će zadržati svoje zaposlenike te na taj način smanjiti potencijalne troškove nastale zbog njihova odlaska.

Američka policija (Santa Cruz u Kaliforniji, Richmond u Virginiji, Chicago, Los Angeles, Memphis) koristi se starim podatcima (uključujući i podatke o danu u tjednu, vremenskoj prognozi, prazniku, posebnim događanjima u gradu) kako bi predvidjela moguća mjesta na kojima bi se mogla dogoditi neka kriminalna radnja. Na temelju tih podataka šalje više patrolnih automobila koja nadgledaju mjesta s većim rizikom za protuzakonite radnje.

Amazon je potpisao ugovor s AOL-om da im nude tehnologiju za njihove komercijalne stranice – usput su iskoristili (što AOL u tom trenutku nije znao da se sprema!) prikupljene podatke da vide što AOL-ovi korisnici pregledavaju, a što kupuju. Na temelju prikupljenih podataka poboljšali su svoj algoritam za davanje preporuka korisnicima koji kod njih kupuju/pregledavaju sadržaj (Mayer-Schönberger, Cukier, 2013).

Tvrtka za mjerenje web prometa Hitwise omogućuje svojim klijentima (vlasnicima web sjedišta) pretraživanje starih prikupljenih podataka kako bi iz njih mogli naučiti što vole njihovi potrošači te na taj način prilagodili/popravili svoju ponudu/uslugu.

Čak i Google omogućuje pretraživanje dijela svojih analiza pojmova koji se pretražuju preko njihova pretraživača. Upotrebom baš tih podataka, Google znanstvenici 2009. godine predvidjeli su pojavu i širenje gripe.

Drugi Google primjer je i njihova aplikacija za prepoznavanje govora, GOOG-411, koju su testirali od 2007. do 2010. Budući da nisu imali svoju tehnologiju za prepoznavanje govora, potpisali su ugovor s tvrtkom *Nuance,* no nisu naveli tko će zadržati podatke s prijevodom glasova pa ih je Google odlučio zadržati. Sada Google može statistički odrediti da neki digitalizirani zapis odgovara određenoj riječi, što je ključno za usavršavanje same tehnologije za prepoznavanje govora, ali i za neke nove usluge (Mayer-Schönberger, Cukier, 2013). Tako su stari prikupljeni podatci dobili novu upotrebu u Googleu.

Tvrtka *Forecast* prikupljala je podatke o prethodnim prodajama avionskih karata kako bi kreirali algoritam kojim se mogu predvidjeti buduća kretanja cijena aviokarata. Na taj način pomažu kupcima da u najpovoljnijem trenutku kupe kartu.

Mobilni operateri prikupljaju informacije o poziciji korisnika. Ova je informacija njima od uske – tehničke – naravi, ali je veoma vrijedna tvrtkama koje distribuiraju personalizirane reklame zasnovane na lokaciji.

Tvrtka za posuđivanje filmova *Netflix* raspisala je 2008. godine natječaj od milijun dolara za onoga tko im pomogne poboljšati njihov sustav za preporuku filmova za 10 % na temelju 100 milijuna prethodnih preporuka svojih korisnika.

Uprava bejzbolskoga tima *Yankees* zaposlila je 2002. podatkovne znanstvenike te su na temelju njihovih podataka počeli pobjeđivati (Waller, Fawcett, 2013). Tim znanstvenika koji je bio zadužen za pregled starih podataka o igrama i igračima *Yankees*a za svoje je odluke odabrao samo vitalne kriterije temeljene na statističkim vezama među određenim varijablama. Uskoro su isto

napravili i mnogi drugi bejzbolski timovi kad su uvidjeli moć koju Veliki podatci uz pomoć prediktivne analitike imaju.

Ovo su samo neki od primjera gdje se upotreba starih podataka pokazala izvrsnim izvorom informacija na temelju kojih su njihovi korisnici mogli poboljšati svoju ponudu ili uslugu, odnosno, pridonijeli su, neki manje, a neki više, poboljšanju uvjeta života.

5.2. Spajanje različitih skupova podataka

Drugi način kojim se oslobađaju nove vrijednosti prikupljenih podataka jest spajanje različitih skupova podataka. Iako su mnogi podatci vrijedni sami po sebi, pri spajanju s drugim bazama podataka njihova vrijednost može još više porasti. Spajanje ili agregacija podataka može dati posve nove uvide u područje, omogućiti neke nove upotrebe ili osigurati podatke za neke nove inovacije.

Tako primjerice, *FlyOnTIme.us* daje podatke o tome kolika je vjerojatnost da će neki let biti otkazan zbog vremenskih prilika. Oni pri tome kombiniraju podatke o letovima i vremenu iz službenih izvora koji su besplatno dostupni preko mreže: povijesni podatci o letovima koje nudi *Bureau of Transportation*, trenutni podatci o letovima iz baze Federal Aviation Administration, stari podatci o vremenu iz baze National Oceanic and Atmospheric Administration i trenutni podatci iz baze National Weather Service.

Drugi su primjer istraživači danskoga Instituta za rak koji su 2011. htjeli pronaći odgovor povećava li mobitel mogućnost dobivanja raka ili ne (Mayer-Schönberger, Cukier, 2013). U tu su svrhu iskoristili prethodno prikupljene podatke o svim mobilnim pretplatnicima od početka mobilnih mreža u Danskoj (1987. – 1995.) – ukupno 358 403 korisnika; svim pacijentima oboljelima od raka (tumor centralnog živčanog sustava) – ukupno 10 729 pacijenata (u periodu koji je slijedio upotrebu mobitela: 1990-2007); i podatke o najvišem stupnju obrazovanja i dohotku. Njihovo istraživanje, koje je uključivalo sve sudionike (N=svi), što je i cilj velikih podataka, nije pokazalo vezu između raka i vlasnika mobilnih uređaja.

Svi ovi pojedinačni skupovi podataka nemaju toliku moć pronalaženja odgovora kao što je ima njihova kombinacija. I to je ono što Velike podatke čini *velikima*. Kod velikih podataka, zbroj je puno vrjedniji od svojih dijelova. To znači da se neke veze mogu uočiti tek ako ih se koristi u velikim informacijskim korpusima.

Iako se na prvi pogled može činiti kao da više podataka nudi više informacija, to nije uvijek slučaj. Kao što se vidi iz primjera koje daju Berman (2013) te Boyd i Crawford (2011), postoji velika razlika između veličine i

potpunosti Velikih podataka. To je jedan od razloga zašto određivanje testnih uzoraka u ovome kontekstu dobiva jednu kompleksniju dimenziju koja, kako Mayer-Schönberger i Cukier (2013) vjeruju, vodi do potpunog nestanka potrebe za statističkim uzorkovanjem.

5.3. Višenamjenska upotreba podataka

Treći je izvor Velikih podatka i višenamjenska upotreba prikupljenih podataka (engl. *extensible/multiple data use*). Upotreba je podataka proširena, odnosno višenamjenska, ako se podatci ne koriste samo za primarnu, tj. onu prvu namjenu zbog koje su se počeli prikupljati, već im se nađe i neka nova primjena. Kao što se vidi i iz primjera koji slijede, katkad se nova namjena „otkrije" tek nakon što su podatci prikupljeni – no, to ne umanjuje njihovu moć.

Kao što se i očekuje, Google je jedna od najboljih tvrtki u prikupljanju podataka s višenamjenskom upotrebom na umu. Primjer su i njihova *Street View* vozila koja nisu samo slikala kuće i ulice, već su prikupljala imena Wi-Fi mreža (a vjerojatno i sadržaj koji se kretao preko njih) i GPS podatke kako bi usavršili svoje mapirajuće usluge te omogućili funkcioniranje njihova samovozećega vozila – @TED (Mayer-Schönberger, Cukier, 2013).

Novu upotrebu pronašli su i podatci koje bilježe kamere u trgovinama. Njihova osnovna upotreba trebala je biti nadgledanje kako bi se uočile krađe (osnovni razlog za postavljanje kamere – sigurnosni razlozi). No, nakon niza godina, uvidjeli su da s pomoću istih podataka mogu pratiti protočnost kroz trgovinu kako bi uočili mjesta gdje se kupci najviše zadržavaju te taj podatak iskoristili za kreiranje najboljega nacrta trgovine i položaja određenih ponuda (nova namjena podataka).

Godine 1999. programer Hank Eskin izradio je stranicu s pitanjem *Gdje je George?* (misli se na novčanicu od 1$ na kojoj je George Washington) jer je želio saznati kako se brzo i daleko „kreću" novčanice. Stranica je vrlo brzo postala iznimno posjećena i u kratkome roku dosegla je brojku od 100 milijuna zapisa koji su se sastojali od serijskoga broja novčanice i poštanskoga broja gdje ju je osoba koja je trenutno ima, dobila. Podatke koje je Eskin prikupio iskoristili su znanstvenici sa sveučilišta Northwestern i Indiana University kako bi izradili model kretanja/širenja gripe.

5.4. Podatkovni ispušni plinovi

„Ispušni plinovi" podataka pojam je koji se odnosi na one podatke koji nastaju kao sporedni proizvod (nusproizvod) korisničkih interakcija na mreži: gdje su kliknuli, koliko dugo su se zadržali na stranici, što su ukucali... (Mayer-

Schönberger, Cukier, 2013), a koje tvrtke mogu iskoristiti za poboljšavanje postojećih usluga ili za ponudu novih.

Zato Google prati koliko je puta neki termin tražen, koliko puta je tražen srodni termin, koliko smo puta kliknuli na link, ali se odmah potom vratili natrag i tražili ponovo, jesmo li kliknuli na 5. link na 1. stranici ili na 1. link na 5. stranici, jesmo li odustali od potrage i na temelju tih podataka, njihov algoritam za rangiranje odgovora prilagođava redoslijed ponuđenih odgovora.

Podatkovni „ispušni plinovi" koriste se i kod usluga poput prepoznavanja glasova (kad programu kažete da niste razumjeli - vježbate ga da poboljša svoj algoritam), filtriranja spamova i jezičnih prijevoda.

U MS **Word**u, *spell checker* uspoređuje napisane riječi s onima u rječniku točno napisanih pojmova (koji se često nadograđuje). Sustav potom „gleda" nepoznate riječi za koje nudi ispravne verzije. Održavanje ovakvoga rječnika je dosta je skupo i moguće je samo za manji broj jezika. Microsoft je za kreiranje i održavanje svojih rječnika trošio milijune dolara.

Za razliku od MS Worda, za **Google** možemo reći da ima najpotpuniji *spell checker* na svijetu koji se svakodnevno nadopunjuje, i to za SVE žive jezike, a dobio ga je skoro besplatno (kao slučajan proizvod svakodnevne upotrebe njihova web pretraživača). Naime, Google procesuira oko 3 bilijuna upita dnevno. Jedna mala (ali pametna) petlja, govori sustavu koju je riječ korisnik stvarno mislio upisati. Korisnici pomažu u tom procesu **izravno** (kad kliknu na link: *Did you mean:house?*) i **neizravno** (kad odaberu link u kojemu je ponuđena riječ dobro napisana).

Veliku grešku u procjeni pokazao je **Yahoo** koji je 2000-te imao sličnu ideju prikupljanja upita kao i Google, ali je nije uspio realizirati na ovakav način jer je, kao i Infoseek i Alta Vista, pogrešno napisane upite smatrao – SMEĆEM!

5.5. Grobnice podataka

Posljednje mjesto koje nam može poslužiti kao izvor Velikih podataka jesu grobnice podataka. To je mjesto gdje se drže prikupljeni podatci koji su se (možda) jednom iskoristili, a potom pohranili bez daljnje upotrebe. Takvi su npr. podatci prikupljeni u sklopu projekta Nacionalne biološke informacijske infrastrukture koji je tekao od 2001. do 2012. da bi tada, radi smanjivanja troškova, bio zatvoren i odbačen. Prikupljeni su podatci za sada neiskorišteni.

Kao što je vidljivo iz prethodnoga primjera, neće svi izvori Velikih podataka biti uspješni, kao ni sve njihove primjene. Istraživanja pokazuju da su čak ¾ bolničkih informacijskih sustava u Americi bile neuspješne. U Velikoj Britaniji

odbacili su investiciju od 17 milijardi dolara koja je bila uložena u informacijski sustav Velikih podataka *UK National Health Service* centra. U Americi su prekinuli 350 milijuna dolara vrijedan projekt poznat kao Cancer Biomedical Informatics Grid koji je trebao razviti standarde za označavanje i dijeljenje biomedicinskih podataka i alata za obradu podataka (Berman, 2013).

Ovo su samo neki od primjera grobnica podataka koje čekaju pravu ideju koja će im otvoriti vrata i naći im neki smisao.

6. Veliki podatci – strah i trepet novoga stoljeća ili novi oblici zaštite

Sasvim sigurno možemo reći da internet ugrožava našu privatnost – ali i da Veliki podatci to čine još više. Osim naše privatnosti (gdje smo i s kim bili, o čemu smo razgovarali, što nosili) u opasnost možemo doći zbog naših prirodnih sklonosti pa nas se može početi kažnjavati i za stvari koje još nismo napravili (a možda ni nikad i ne bismo) (Mayer-Schönberger, Cukier, 2013). Naime, profiliranje, odnosno, analiza podataka u svrhu definiranja grupe ljudi na koje se neko svojstvo odnosi, može uzrokovati diskriminaciju svih članova grupe ili ih se sve može osuditi samo zbog „pripadanja grupi", npr. ako osoba nosi muslimansko ime, bit će više sumnjiva za neki teroristički napad.

S druge strane, predviđanja o ljudima na temelju Velikih podataka nisu toliko generička jer se koristi ne-kauzalna analiza kojom se jednostavno mogu identificirati najprikladniji pojedinačni kandidati, a ne cijela grupa. Tako se grupni identitet može zamijeniti puno detaljnijim predviđanjem za svakoga pojedinačno. I ovdje se radi o profiliranju, ali puno boljem, manje diskriminirajućem, a više individualizirajućem. Bitno je naglasiti da Veliki podatci sami po sebi nisu štetni, loši niti opasni. Ono što ih čini takvima jest sposobnost čovjeka da zbog očuvanja moći nekolicine, koristi sve dostupne resurse (u ovom slučaju Velike podatke) na štetu većine.

Iako se uporno napominje da se Veliki podatci zasnivaju na korelacijama i da nisu dobar alat za uspostavljanje uzročno-posljedičnih veza (Mayer-Schönberger, Cukier, 2013), mi još uvijek živimo i razmišljamo u svijetu uzroka i posljedica. Kao rezultat toga, logično je i očekivati neki oblik zloupotrebe. Osim „napada" na našu privatnost, Veliki podatci mogu se iskoristiti kao oružje za dehumanizaciju društva čime bi nam bilo onemogućeno slobodno odlučivanje i sloboda izbora, a sve pod krinkom čuvanja toga istoga društva od mogućih počinjenih krivičnih/kriminalnih djela.

Berman (2013) opisuje 8 slikovito imenovanih hipoteza kako društvo doživljava Velike podatke:

1. detektivska hipoteza (engl. *gumshoe*) – prikupljanje informacija o ljudima u svrhu istražnih postupaka od strane privatnih detektiva, policije, ali i znatiželjnika koji vole njuškati i narušavati privatnost drugih
2. hipoteza Velikog Brata – prikupljanje informacija o svim članovima društva u svrhu kontroliranja
3. borgovska hipoteza – prikupljanje informacija u svrhu učenja svega o populaciji
4. *George Carlin* hipoteza – Veliki su podatci mjesto na koje možemo staviti sve naše „stvari"
5. hipoteza potrage za odbačenim stvarima (engl. *scavenger hunt*) – Veliki podatci predstavljaju zbirku svega kreiranu u svrhu pretraživanja za osobnim stvarima i činjenicama – zbirka svega o svemu što bismo ikada željeli znati
6. intelektualna hipoteza – prikupljanje informacija u svrhu izvlačenja generaliziranih znanstvenih zaključaka
7. *Facebook*[5] hipoteza – društvena arhiva koja proizvodi novac (posve suprotna intelektualnoj hipotezi)
8. nihilistička hipoteza – Veliki podatci ne postoje kao polje koje je moguće definirati – oni jednostavno predstavljaju ono što smo oduvijek radili, ali sada s malo više podataka.

Opisujuću tri paradoksa Velikih podataka (paradoks transparentnosti, identiteta i moći), Richards i King (2013) žele osvijestiti moguće opasnosti koje nosi revolucija Velikih podataka, nadajući se da će na taj način pomoći da ostanemo na pravome putu primjene Velikih podataka. S druge strane, Lerman (2013) upozorava na rizike i nepravdu koju analize Velikih podataka mogu prouzrokovati milijardama ljudi (prema podatcima *International Telecommunications Uniona* radi se o 61% populacije) koji još ne koriste internet. Ignoriranje njihovih podataka može dovesti do novoga oblika diskriminacije. Kako će oni biti zastupljeni u „globalnim" odlukama koje će se donositi na temelju Velikih podataka? Tko će se brinuti o njihovim potrebama? Tko će ih zaštiti? Ovdje se zasigurno više ne radi samo o zaštiti privatnosti, već i o ekonomskoj, političkoj i društvenoj (ne)jednakosti.

6.1. Kako se zaštiti – prijedlozi

Kako bi se zaštitili od zloupotrebe, Mayer-Schönberger i Cukier (2013) predlažu neke nove oblike zaštite osobnih podataka koji su nužni za efikasno upravljanje velikim podatcima. Za razliku od trenutno važećih zakona o zaštiti

[5] Veliki je broj korisnika Facebooka upravo onaj dio populacije koji ne bi dao pristanak za upotrebu svojih deidentificiranih podataka u svrhu znanstvenih istraživanja, ali za to isto vrijeme, sasvim besplatno i dobrovoljno ostavlja privatne/intimne zapise na Facebook stranicama gdje ih se jednostavno može prikupiti, analizirati i prodati u komercijalne svrhe.

privatnosti prema kojima osoba kontrolira želi li ili ne dati svoje podatke drugima na obradu, u kojem omjeru i komu točno, njihov prijedlog bi tu odluku prepustio kompanijama koje prikupljaju podatke uz uvjet da su oni odgovorni kako se ti podatci koriste i kakav učinak mogu imati na živote ljudi čije podatke koriste. Ovakva promjena pristupa iz „privatnosti uz odobrenje[6]" na pristup „privatnost kroz odgovornost" trebao bi omogućiti tehnološkim inovacijama da zaštite privatnost u određenim slučajevima.

Drugi je oblik zaštite očuvanje ideje da se ljudima sudi samo za ono što su uistinu učinili, a ne i za ono što bi eventualno mogli učiniti. Veliki podatci nikad i ni pod koju cijenu ne bi smjeli biti jedini izvor informacija za presudu pojedinaca jer bi to značilo uvođenje vladavine podataka do koje nikako ne smije doći. U suprotnom, naša osobna sloboda da djelujemo bila bi izgubljena. Zanimanje od posebnoga značaja postali bi tzv. algoritmičari – novonastali profesionalci iz područja računalne znanosti, matematike i statistike (Mayer-Schönberger, Cukier, 2013). Njihov bi glavni zadatak bio da nepristrano, čuvajući tajnost podataka, vrednovati izvore podataka, birati alate za analizu i interpretirati rezultate. Oni ne bi smjeli biti odani samo tvrtki za koju rade, već i ljudima na koje se rezultati analiza Velikih podataka odnose.

Siegel (2013) dio rješenja vidi u definiranju odgovora na pitanja pod kojim uvjetima i u koju svrhu: *tko?, što?, gdje?, kad?, koliko dugo?* i *zašto?* podatci mogu biti dostupni:

- **držanje** (engl. *retain*) – što se pohranjuje i koliko dugo
- **pristup** (engl. *access*) – koji zaposlenici i koje službe smiju pretraživati i imati uvid i u koje elemente
- **djeljenje** (engl. *share*) – koji se podatci mogu dijeliti dalje, kojim odjelima unutar organizacije i kojim vanjskim organizacijama
- **spajanje** (engl. *merge*) – koji se podatci mogu spajati, grupirati ili povezivati
- **djelovanje** (engl. *react*) – kako se može djelovati ovisno o podatcima.

Neka od rješenja nalazimo i kod Rubinstein (2012). Bez obzira na to koji put odabrali, čini mi se sasvim prikladnim završiti ovo poglavlje rečenicom: **Veliki podatci – molimo upravljati s iznimnim oprezom!**

6.2. Riječi utjehe

Možda kao utjeha ili olakšanje onima čiji se podatci koriste u znanstvene svrhe može poslužiti dio preuzet iz dokumenta koji je izdao američki ***National Science Foundation*** *(2012)*:

[6] U pravnim se redovima sve više preispituju zakoni o privatnosti informacija unutar okvira Velikih podataka (Birnhack, 2013).

„Zahtjev programa 'Osnovne tehnike i tehnologije za unapređenje znanosti i inženjeringa Velikih podataka (BIGDATA)' ima cilj unaprijediti temeljne znanstvene i tehnološke načine za upravljanje, analiziranje, vizualizaciju i izvlačenje korisnih informacija iz velikih, raznolikih, distribuiranih i heterogenih skupova podataka:

1. kako bi se ubrzao napredak znanstvenih otkrića i inovacija
2. kako bi se dovelo do novih područja istraživanja koja inače ne bi bila moguća
3. kako bi se potaknuo razvoj novih podatkovnih analitičkih alata i algoritama
4. kako bi se omogućila podesiva, pristupačna i održiva podatkovna infrastruktura
5. kako bi se povećalo razumijevanje ljudskih i društvenih procesa i interakcija
6. te kako bi se promovirao ekonomski rast i poboljšalo zdravlje i kvaliteta života.“

Novonastala znanja, alati, postupci i infrastrukture omogućit će revolucionarna otkrića i inovacije u znanosti, inženjeringu, medicini, trgovini, obrazovanju i državnoj sigurnosti – postavljajući tako temelj za američku konkurentnost u nadolazećim desetljećima.“

Imajući na umu gore navedeno, dijeljenje Velikih podataka od velike je važnosti za razvoj znanosti. U tom procesu dobro je iskoristiti mogućnost deidentifikacije podatkovnih objekata, ili anonimizacije podataka, kako bi se zaštitila privatnost osoba koje stoje iza tih objekata. Takvi su podatci i dalje objektivni podatci koji imaju veliku znanstvenu vrijednost, a rezultati dobiveni u istraživanjima, koja bi koristila deidentificirane podatke, uvelike bi mogla pomoći svim stanovnicima ovoga planeta.

7. Podatkovni stručnjaci

Za očekivati je da su ulazak u novu eru, pokretanje nove revolucije i stvaranje nove discipline otvorili mjesto za novoga stručnjaka – podatkovnog znanstvenika – *najseksipilnije* zanimanje 21. stoljeća (Simon, 2013). Prema izvještaju tvrtke McKinsey & Company iz 2011. godine (Manyika et al.), do 2018. godine samo na Američkom tržištu nedostajat će preko milijun i pol podatkovnih znanstvenika.

Berman (2013) vjeruje da će većina tih znanstvenika biti stručnjaci koji će pripremati podatke za različite analize. Njegova podjela podatkovnih profesionalaca:

- na **profesije zadužene za izgradnju resursa** (projektanti Velikih podataka, stručnjaci za indeksiranje Velikih podataka, metadata stručnjaci, područni stručnjaci, stručnjaci za spajanje podataka iz više izvora, ontolozi i klasifikatori, programeri, konzervatori podataka i stručnjaci za naslijeđene podatke, podatkovni menadžeri i menadžeri baza podataka, mrežni stručnjaci i stručnjaci za zaštitu)[7]
- i na **profesije koje će koristiti resurse** (analitičari podataka, stručnjaci za rješavanje generaliziranih problema[8], ljudi s primjerenim programskim vještinama[9], specijalisti u kombinatorici, specijalisti za redukciju podataka, vizualizatori podataka, Big Data znanstvenici[10])

može nam značajno pomoći u izradi programa za obrazovanje tih profesija u Hrvatskoj kako se ne bi suočili s tako velikim deficitom koji se predviđa za Ameriku. Mnoga sveučilišta već nude programe za obrazovanje podatkovnih znanstvenika (Simon Fraser University u Kanadi, američka sveučilišta poput Carnegie Mellon University, Columbia University, University of California i Stanford University, Villanova University u Španjolskoj), a tu su i besplatni on-line tečajevi koje nudi Big Data University.

Što se očekuje od svih ovih novih stručnjaka? Možda to najbolje ilustrira grafički prikaz koji nude Law, Greenbacker i Eberhardt (2014) prema kojima svaki podatkovni znanstvenik treba biti upoznat s temeljnim vještinama koje su navedene u donja četiri retka tablice (mora biti kompetentan programer, imati primjereno razumijevanje matematike, statistike i analitičke metodologije, biti upoznat s okvirima distribuiranog računalstva te uz izvrsne komunikacijske vještine imati i temeljna znanja o domeni u kojoj djeluje), ali i biti ekspert u barem jednom vertikalno navedenom području (vidi tablicu 1).

Na projektima s Velikim podatcima uvijek rade timovi ljudi i svi oni rade poslove koji do prije nekoliko godina nisu postojali. Karika koja veže sva ta nova zanimanja je menadžer Velikih podataka.

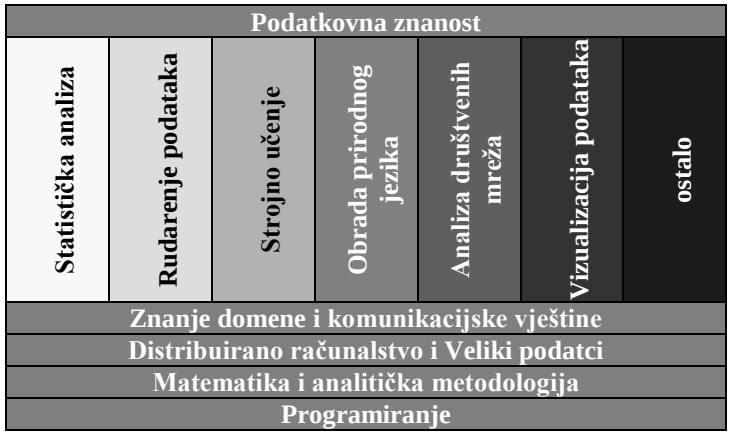
[7] Svi su ovi stručnjaci iz polja informacijskih tehnologija i najveći izazov koji mogu imati je da surađuju s ostalim članovima tima koji rade na istom resursu.

[8] Ovi su stručnjaci prema Bermanu (2013) najvažniji kadar koji sveučilišta tek trebaju obučiti kako će rješavati probleme (engl. *Generalist problem solvers*). Karakterizira ih interes za više različitih polja, po prirodi vole postavljati pitanja i talentirani su da vide odnose gdje ih drugi ne primjećuju. Oni razumiju na koji se način podatci iz različitih izvora mogu spojiti, ali i kako se problemi iz jednoga područja mogu generalizirati na ostala područja i riješiti kombiniranim podatcima i metodama iz više područja.

[9] Nije nužno da korisnici Velikih podataka znaju više od osnova skriptnih jezika poput Pythona ili Perla. Jednostavne skripte bit će sasvim dovoljne za potrebe Velikih podataka.

[10] Imat će vještine s kojima će moći otkriti sve tajne koje leže unutar izvora Velikih podataka, oni će biti savjetnici institucijama i korporacijama za načine iskoristivosti podataka kojima raspolažu, ali i oni koji će imati uvid u mogućnosti izvora velikih podataka – jesu li odgovori koje korisnik traži uopće mogući nad postojećim izvorima.

Tablica 1 – opis podatkovnog znanstvenika (prilagođeno od Law et al., 2014)



Menadžer Velikih podataka posrednik je između svih članova tima koji radi na Velikim podatcima (posebno između podatkovnog znanstvenika, analitičara i vizualizatora), ali i između tima i ostalih službi u organizaciji. Njegov je zadatak kreirati protokole kojima se opisuju procesi kojima se omogućuje ponovna identifikacija deidentificiranih objekata, ali i nadgledanje tih procesa. Oni odlučuju i hoće li podatke modelirati s pomoću klasifikacije (prema kojoj će svaki objekt imati samo jednu direktnu roditeljsku klasu) ili će izvore podataka modelirati s pomoću ontologija (prema kojoj će svaka klasa moći nasljeđivati od više roditelja). Moraju paziti i na standarde (posebno na njihove pravne aspekte) kojima se koriste jer zbog svoje veličine i različitosti, Veliki podatci jednostavno zahtijevaju različite standarde za različite tipove podataka, ali i za različite softvere za upravljanje Velikim podatcima. Njihov je zadatak i uspostaviti metode za dokumentiranje svih mogućih procedura kojima se podvrgavaju Veliki podatci (verifikacija podataka[11], validacija). Takvi protokoli moraju biti datirani i potpisani (od strane svakog člana odbora koji ima zadatak provjeriti protokol) nakon svake moguće izmjene (revizije protokola predlaže menadžer VP-a). Moraju paziti i da se protokoli poštuju kako bi izbjegli (ili barem sveli na minimum) udvajanje ili gubljenje podataka, povećan broj nepostojećih vrijednosti, zastarijevanje termina upotrebljavanih za imenovanje... Menadžeri Velikih podataka moraju znati imaju li pravo i pod kojim uvjetima prikupljati podatke i distribuirati podatke koji su im dani na upravljanje, ali isto tako moraju paziti na tajnost (čuvanje podataka tajnim) i privatnost (podatci se ne smiju upotrebljavati u svrhu uznemiravanja osobe koju podatci opisuju s dodatnim pitanjima) osoba čiji se podatci nalaze unutar Velikih podataka.

[11] Menadžeri Velikih podataka verificiraju da su podatci ispravno prikupljeni, ali ne i da su prikupljeni podatci ispravni.

Podatkovni znanstvenik (engl. *data scientist*) ima vještine statističara, programera, vizualizatora podataka i pripovjedača! Termin je osmislio Jeff Hammerbacher (Facebook). Njegov je tim otkrio da se može zaključiti hoće li netko napraviti neku radnju na Facebooku ili ne, te da to ovisi o tome hoće li isto napraviti i njihovi prijatelji (zato se toliko ističe što prijatelji rade na Facebooku!).

Podatkovni znanstvenici uglavnom imaju bakalaureat ili magisterij iz područja umjetne inteligencije, obrade prirodnog jezika ili upravljanja podatcima s jakom podlogom u matematici i/ili statistici. Od njih se očekuje da su kreativni u rješavanju problema, da jednostavno koriste ideje i koncepte iz jednog područja i uspješno ih primjenjuju na druga područja, s izvrsnim komunikacijskim vještinama prema svim organizacijskim razinama. S obzirom na to da će se alati razlikovati od jedne do druge organizacije, nužno je da podatkovni znanstvenik razumije srž alata koji se koriste za interpretiranje i analizu podataka kako bi bio u mogućnosti pomoći organizaciji da poveća svoju efikasnost, ali i profit. Za vizualni prikaz puta koji podatkovni znanstvenik treba proći može poslužiti mapa koju je izradio Swami Chandrasekaran, a koja je u obliku metro karte (dostupna na http://tinyurl.com/q55vnac).

Analitičari Velikih podataka prikupljaju i organiziraju podatke. Oni modeliraju događaje i procese koji se trebaju dogoditi u slučajevima novih poslovnih prilika odnosno neprilika. Osmišljeni se modeli potom spajaju sa stvarnim podatcima i osluškuju podatke u potrazi za naznačenim signalnim vrijednostima. Oni odlučuju na koji će način riješiti problem izostavljenih vrijednosti, nemogućih vrijednosti i vrijednosti koje odskaču od definirane domene. Oni biraju i tehnike kojima će smanjiti dimenzionalnost podataka. Sve svoje odluke moraju dokumentirati.

Podatkovni vizualizatori su vješti analitičari s izvrsnim komunikacijskim vještinama koji prevode podatke u informacije kako bi se one mogle efikasnije koristiti. Oni moraju istražiti postojeće podatke kako bi mogli identificirati što oni točno znače i koji bi utjecaj mogli imati na posao, a potom te informacije na jasan i pristupačan način prezentirati (vizualizirati) ne-tehničkom osoblju i upravi.

Podatkovni arhitekti programeri su s tradicionalnim obrazovanjem u području programiranja i poslovne inteligencije obučeni za rad s nestrukturiranim podatcima. Njih ne plaše višeznačnosti i ustrajni su u rješavanju podatkovnih problema na nove i inovativne načine.

Podatkovni inženjeri i operatori analitičari su IT ili IS sustava koji imaju zadatak izgraditi sustav koji su osmislili podatkovni arhitekti, a potom održavati, testirati i vrednovati sustave koji koriste Velike podatke

Agenti podatkovnih promjena također su analitičari s izvrsnim komunikacijskim vještinama s osnovnim ciljem uvođenja promjena unutar organizacija na temelju postojećih podatkovnih analiza.

Konzervatori (engl. *curator*) **izvora Velikih podataka** imaju zadatak prikupljati naslijeđene (engl. *legacy)* i buduće podatke u izvor, moraju paziti na postojanje adekvatnih protokola za verifikaciju podataka, moraju odabrati prikladno nazivlje za anotiranje (obilježavanje) podataka, moraju anotirati podatke, ali i napraviti potrebne prilagodbe u slučajevima kad se pojave nove verzije nazivlja ili se pak neko nazivlje zamijeni drugim.

8. Zaključak

Trebala je nova era i nova revolucija da čovjek ponovo počne obraćati pozornost na informaciju, a ne samo na tehnologiju koja tu informaciju omogućava. Čovjekovu vječnu želju da izmjeri, zapiše i analizira svijet u kojemu živi moguće je sada ostvariti na jednoj novoj razini s pomoću Velikih podataka.

Nije bitna samo količina podataka, već i brzina akumuliranja novih podataka kao i raznolikost tipova prikupljenih podataka. Primjeri upotrebe Velikih podataka pokazuju nam da su oni ušli u sve sfere našega života i da ih više ne možemo ignorirati. Potrebno je u kratkom roku početi s obrazovanjem kadra (Je li moguće da mi već kasnimo?) koji će se moći nositi s obradom, prikupljanjem, analiziranjem, upravljanjem, vizualizacijom Velikih podataka, ali prije svega na jedan etičan način imajući na umu osobu-čovjeka koju ti podatci opisuju. Jer čovjek je razlog zašto tražimo bolja rješenja i cijela znanost trebala bi biti tu upravo radi čovjeka – kako bi mu se omogućio bolji život i bolji osobni razvoj.

Poseban značaj Velikih podataka mogu imati interdisciplinarna istraživanja u kojima bi se mogli spajati podatci prikupljani pojedinačno za svaku od disciplina. Odjednom nam se otvara cijeli niz pitanja na koja je možda jedino moguće odgovoriti upotrebom Velikih podataka (veza između ekoloških katastrofa i epidemija bolesti, migracija ptica i globalnih vremenskih uvjeta i sl.), ali i uvidjeti iskoristivost rješenja jednoga područja za rješavanje ekvivalentnih problema nekog drugog područja.

Nove spoznaje i nova rješenja već su tu, u podatcima koji nas okružuju. Treba se samo znati zagledati kako bi se uočili ti uzorci koji će nam otvoriti jednu novu škrinju znanja i pomoći nam u rješavanju mnogih zagonetki, imajući

uvijek na umu etičnost, i za korištenje Velikih podataka i za primjenu rješenja koja ti podatci nude.

9. Literatura

Berman, J. J. (2013), *Principles of big data: preparing, sharing, and analyzing complex information*, Elsevier, Morgan Kaufman, Amsterdam.

Birnhack, M. (2013), „S-M-L-XL Data: Big Data as a New Informational Privacy Paradigm“, u Big Data and Privacy: Making Ends Meet 7-10 (Future of Privacy Forum & Center for Internet & Society, Stanford Law School), dostupno na: http://ssrn.com/abstract=2310700 (pristupljeno 3.4.2014.).

Boyd, D., Crawford K. (2011)„Six Provocations for Big Data, in A Decade in Internet Time“, Symposium on the Dynamics of the Internet and Society, dostupno na: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431 (pristupljeno 1.4.2014.).

Computing Community Consortium (2008), "Big Data Computing: Creating revolutionary breakthroughs in commerce, science, and society", dostupno na: *http://www.cra.org/ccc/files/docs/init/Big_Data.pdf* (pristupljeno 20.3.2014.).

Diebold, F. X. (2000), „'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting“ Discussion Read to the Eighth World Congress of the Econometric Society, Seattle, August, dostupno na: http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF (pristupljeno 1.4.2014.).

Diebold, F.X. (2013), „A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline“, Second Version, PIER Working Paper No. 13-003, dostupno na: http://ssrn.com/abstract=2202843 (pristupljeno 1.4.2014.).

Hilbert, M., Lopez, P. (2012), „How to Measure the World's Technological capacity to Communicate, Store, and Compute Information Part I: Results and Scope“ u International Journal of Communication 6, 956-979, dostupno na: http://ijoc.org/index.php/ijoc/article/view/1562/742 (pristupljeno 3.4.2014.).

Hrvatski Sabor (2013), „Zakon o pravu na pristup informacijama“, dostupno na: http://narodne-novine.nn.hr/clanci/sluzbeni/2013_02_25_403.html (pristupljeno 25.3.2014.).

Law, D., Greenbacker, C., Eberhardt, J. (2014), „Do You Know Big Data?“, poster, dostupno na: http://www.ctovision.com/download/know-big-data/ (pristupljeno 10.6.2014.).

Lerman, J. (2013), „Big Data and Its Exclusions“ u 66 Stanford Law Review Online 55, dostupno na http://ssrn.com/abstract=2293765 (pristupljeno 7.4.2014.).

Lohr, S. (2013), „The Origins of 'Big Data': An Etymological Detective Story“, dostupno na: http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/, (pristupljeno 29.3.2014.).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. (2011), „A. Hung Byers: Big Data: The next frontier for innovation, competition, and productivity“, Report by McKinsey Global Institute, dostupno na: http://tinyurl.com/cplxu6p (pristupljeno 15.3.2014.).

Mayer-Schönberger, V., Cukier, K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt, Boston.

National Science Foundation (2012), „Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)“, National Science Foundation program solicitation NSF 12-499, dostupno na: http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf (pristupljeno 1.4.2014.).

Reeve, A. (2013), *Managing Data in Motion: Data Integration Best Practice, Techniques and Technologies*, Morgan Kaufmann, Waltham.

Richrads, N.M., King, J.H. (2013), „Three Paradoxes of Big Data“ u 66 Stanford Law Review Online 41, dostupno na http://ssrn.com/abstract=2325537 (pristupljeno 7.4.2014.).

Rubinstein, I.S. (2013), „Big Data: The End of Privacy or a New Beginning?“ u International Data Privacy Law; NYU School of Law, Public Law Research Paper No. 12-56, dostupno na: http://ssrn.com/abstract=2157659, (pristupljeno 5.4.2014.).

Siegel, E. (2013), *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Simon, P. (2013), *Too Big to Ignore*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Waller, M.A., Fawcett, S.E. (2013), „Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain“ u Journal of Business Logistics, Vol. 34, Nb.4, dostupno na http://ssrn.com/abstract=2339972 (pristupljeno 2.42014.).

Washington, A.L. (2014), „Big Data and Public Sector Information, Online International Forum on Postal Big Data“, dostupno na: http://ssrn.com/sol3/abstract=2386150 (pristupljeno 3.4.2014.).