SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2014./ 2015.

Petra Čačić

BIG DATA
Novi pristup epistemologiji i metodologiji društvenih znanosti – mogućnosti i etički problemi

Završni rad

Mentor: dr.sc. Kristina Kocijan, doc.

Zagreb, 2015.

Sadržaj

1. Uvod

Ovaj rad istražit će relativno novi fenomen u svijetu koji se označava sintagmom Big Data. Za potrebe ovog rada koristit će se hrvatski prijevod Veliki podaci koji se tek treba uvriježiti u standardnom govoru, s obzirom na njegovo nedavno pojavljivanje u domaćoj literaturi (Europska komisija, 2014; Kocijan, 2014).

U prvom poglavlju definirat će se pojam Velikih podataka i prikazati dosadašnje teorijske spoznaje, a osvrnut će se i na društveni kontekst u kojem koncept pronalazi svoju primjenu. Opisat će se svojstva koja karakteriziraju Velike podatke razlikujući ih od velike količine podataka, zbog čega dolazi do potrebe za novim metodama i alatima prikupljanja, obrade i analize podataka, kao i za novim zanimanjima.

U poglavlju Datafikacija otkrivaju se načini na koje riječi, lokacije i interakcije postaju podaci. Sam pojam 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“ (Kocijan, 2014:6). Kako datafikacija dovodi do stvaranja novih oblika vrijednosti i koji je njezin potencijal za društvene znanosti u smislu otkrivanja dosada neistraženog društvenog života grupa i pojedinaca, neka su od pitanja na koja će ovo poglavlje pokušati odgovoriti uz pomoć niza primjera iz svakodnevnog života.

Četvrto poglavlje odnosi se na Velike podatke i promjene koje oni donose u razmišljanju o podacima. Konkretno, na koji način sve veća količina podataka koju je moguće obraditi utječe na potencijalni zaokret u području metodologije znanstvenoistraživačkog rada. Hoće li tehnološka revolucija dovesti do mogućnosti da u budućnosti analiziramo sve podatke, te da uzorkovanje postane stvar prošlosti? Druga promjena odnosi se na prihvaćanje nereda u različitim područjima istraživanja i u strukturi podataka u zamjenu za širu sliku. Iako se sukobljava s intuicijom, teoretičari

Velikih podataka tvrde kako tretiranje veće količine netočnih podataka omogućuje napredne prognoze (Kitchin, 2014). Posljednja promjena bavi se odnosom korelacije i uzročnosti. Veliki podaci predlažu da se odmaknemo od fiksacije na uzročnost te zauzvrat otkrijemo korelacije koje nam ne moraju nužno reći „zašto“ se nešto događa nego nas samo obavijestiti da se to događa, te kako je to često dovoljno.

Peto poglavlje ispituje na koji način podatkovna revolucija (eng. *data revolution*) utječe na preoblikovanje konceptualizacije i prakse istraživanja u društvenim znanostima. Definiraju se pojmovi epistemologije, osnovnih metoda zaključivanja i paradigme, te preispituje mogućnost stvaranja paradigme Velikih podataka i nove znanosti koja se temelji na podacima (eng. *data-driven science*).

Šesto poglavlje bavi se etičkim implikacijama koje donose promjene u korištenju podataka. Najčešći problemi vezani su uz nadzor, privatnost, sigurnost podataka, profiliranje i tehnološko upravljanje. Ovo poglavlje nastoji opisati navedene probleme Velikih podataka te sagledati njihov utjecaj na društvo i pojedinca.

Kako bi se ova apstraktna tema pobliže objasnila, potrebno ju je primijeniti na konkretne slučajeve. Sedmo poglavlje obrađuje dvije studije slučaja. Prva se odnosi na slučaj tvrtke koja je uz pomoć Velikih podataka napravila softver koji predviđa povećanje ili smanjenje cijene avionskih karata. Drugi slučaj opisuje kako je Google pomoću velike količine podataka, procesorske moći i statističkog znanja predvidio širenje gripe.

Posljednje poglavlje pobliže opisuje prediktivnu analizu kao naprednu tehnologiju koja dovodi do značajnih promjena u različitim sferama ljudskog djelovanja. Ukratko se opisuje programski paket SPSS, namijenjen statističkoj analizi podataka te njegova suvremena inačica razvijena od strane IBM-a, SPSS Modeler, prediktivna analitička platforma približena krajnjim korisnicima i suvremenim Web tehnologijama. Također, daje se i osvrt na budućnost prediktivne analize koja ovisi o brzini kojom će tehnološki napredak omogućiti provedbu idejnih rješenja.

2. Veliki podaci

Termin „Veliki podaci“ odnosi se na našu novu mogućnost da obradimo ogromne količine informacija, automatski ih analiziramo te iz njih izvučemo ponekad zapanjujuće zaključke. Veliki podaci opisani su kao dio računalne znanosti koja se naziva umjetna inteligencija, preciznije, područje strojnog učenja; međutim oni ne pokušavaju naučiti računalo da razmišlja kao ljudi (Mayer-Schönberger, Cukier; 2013:11). Točnije, radi se o primjeni matematike na ogromnim količinama podataka s ciljem povećanja vjerojatnosti. Na primjer, koliko je vjerojatno da je email poruka spam ili da su slova „teh“ trebala biti „the“. U svojoj srži Veliki podaci temelje se na predviđanjima, a ključni faktor je da sustav dobro radi zato što koristi velike količine podataka te da s vremenom i novim podacima postaje sve bolji. Međutim, osim što je za funkcioniranje sustava potrebno mnogo podataka, potrebna je i velika procesorska snaga i prostor na kojem će se svi ti podaci skladištiti. S obzirom da su navedeni faktori potrebni za njegovu realizaciju bili preskupi, fenomen Velikih podataka uzima maha upravo u ovom vremenu, a ne nekoliko godina ranije.

Prema jednoj od definicija, Veliki podaci su svi podaci kojima nije moguće upravljati, niti ih analizirati pomoću standardnih alata i tehnika analize podataka (Manyika et al., 2011). Kako bi pokušali predočiti ljudima što znači tolika količina podataka Mayer-Schönberger i Cukier (2012) opisali su kako bi 1,200 exabajta podataka procijenjenih u 2013. godini, u slučaju da su u obliku tiskanih knjiga – pokrili cijelu površinu SAD-a u debljini od 52 sloja. Podaci proizvedeni od znanstvenika, korisnika Interneta i fizičkih sistema zajedno se nazivaju Veliki podaci koje treba razlikovati od velike količine podataka. Kako bi se određeni skup podataka mogao smatrati Velikim podacima potrebno je da posjeduje tri karakteristike popularno nazvane „3V“ prema početnim slovima značajki na engleskom jeziku *volume, variety, velocity*. Volumen (eng. *volume*) se odnosi na „veliku količinu podataka“, varijantnost (eng. *variety*) na „raznolikost tipova podataka, uključujući tradicionalne baze podataka, fotografije,

dokumente i složene zapise“, a velocitet (eng. *velocity*) na „brzinu kojom se akumuliraju novi podatci iz sličnih izvora podataka, iz prethodno arhiviranih podataka ili naslijeđenih zbirki, i iz prenesenih podataka koji pridolaze iz različitih izvora“ (Berman, 2013:xx). Postojanje ovih triju vrijednosti zahtijeva nove metode za oblikovanje, rukovanje i analiziranje Velikih podataka. Ponekad se spominje i četvrti „V“ koji pretpostavlja kako podaci sami po sebi posjeduju određenu vrijednost (eng. *value*) (Zhang, 2013).

Još neke od karakteristika koje se odnose na Velike podatke su istinitost (eng. *veracity*) u smislu kvalitete prikupljenih podataka koja može uvelike varirati (i tako utjecati na točnost analize), složenost (eng. *complexity*) upravljanja podacima s obzirom da dolaze iz različitih izvora, neodređenost ili neizvjesnost (eng. *vagueness*) vezana uz rezultate analize. Učestalo je i da se prije bilo kakve analize podaci podvrgnu postupku verifikacije, to jest, da se provjeri zadovoljavaju li određeni skup podataka, a nakon analize postupku validacije - provjere je li svrha podataka zadovoljena i konzistentna, to jest, mogu li se isti zaključci dobiti iz istoga skupa podataka i u ponovljenoj analizi (Kocijan, 2014).

Već sada postoje različita mišljenja o tome koja je svrha prikupljanja i obrade Velikih podataka, a Berman (2013) navodi neka od njih kao što su prikupljanje informacija u svrhu kontroliranja društva (Veliki brat) i istražnih postupaka, učenja i izvlačenja generaliziranih znanstvenih zaključaka ili u svrhu povećanja profita.

Proučavajući etimologiju termina Velikih podataka, Francis Diebold došao je do sredine 90-ih godina 20. stoljeća i njegovog najvjerojatnijeg tvorca Johna Masheya, vodećeg znanstvenika u tvrtki Silicon Graphics koja je upotrebljavala nove tipove podataka u velikim razmjerima. U 2008. godini još uvijek tek nekolicina ljudi koristi ovaj termin, dok 2013. godine on postaje izrazito prisutan (eng. *buzzword*) u poslovanju, popularnim medijima i znanstvenim časopisima (Kitchin, 3014:67).

S obzirom na brzinu kojom se stvari odvijaju, može se reći da se radi o svojevrsnoj podatkovnoj revoluciji (eng. *data revolution*) koja ima potencijal da promijeni način na koji živimo, radimo i razmišljamo. Stvara se nova disciplina i cijeli niz novih zanimanja kao što su podatkovni znanstvenik (eng. *data scientist*), menadžer Velikih podataka, analitičar Velikih podataka, ili podatkovni arhitekt. 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.

3. Datafikacija

Riječ „data“ u latinskom znači „nešto što je dano“ u smislu činjenice, a u hrvatskom jeziku „data“ se prevodi kao podatak. Podatak (eng. *data*) je „znakovni prikaz činjenica, pojmova i instrukcija na formalizirani način, a pogodan za komuniciranje, interpretaciju i obradu od strane ljudi ili strojeva“ (Tuđman, 1990:203). S obzirom da u hrvatskom jeziku još ne postoji termin ekvivalentan engleskom terminu „datafication“ u radu ću za navedeno koristiti termin „datafikacija“ u smislu pretvaranja fenomena u mjerljiv oblik (podatke) kako bi se mogli računati i analizirati. Ovaj moderni tehnološki trend pretvara mnoge aspekte našeg života u računalne podatke te na taj način stvara nove oblike vrijednosti.

Kako bi mogli prikupiti podatke potrebno je da znamo kako ih izmjeriti i kako zapisati ono što smo izmjerili. Kvantifikacija je omogućila predviđanje i planiranje ljudskih aktivnosti još od pojave pisane riječi. Omogućila je replikaciju kuća i zgrada na osnovu zapisa o njihovim dimenzijama i korištenim materijalima. Veliki napredak u trgovanju, a i nizu drugih područja omogućili su arapski brojevi i razvoj računovodstva, konkretno, zapisivanje izdataka i dobitaka kako bi se izračunao profit. Mjerenje vremena, udaljenosti, područja, obujma, težine nastavilo se razvijati sa ciljem povećanja preciznosti. Želja da spoznamo prirodu pomoću mjerenja definirala je znanost 19. stoljeća i dovela do razvitka novih alata i jedinica za mjerenje struje, pritiska zraka, temperature, frekvencije zvuka. Poslovi mjerenja i bilježenja zahtijevali su puno vremena, troškova i strpljenja, ali su redovito dovodili do novih vrijednosti i spoznaja. Pojava računala, digitalnog mjerenja i uređaja za pohranu podataka učinilo je datafikaciju puno efikasnijom.

3.1.Riječi kao podaci

Veliki pomak u području datafikacije napravio je Google. Sa željom da što više knjiga učini besplatno dostupnima javnosti putem Interneta, pokrenuo je projekt digitalizacije. Svaka stranica je skenirana i pohranjena na Google-ovim serverima u obliku slike visoke rezolucije. Prebacivanjem u digitalni oblik omogućeno je da se knjige dohvate putem Weba, međutim nije se mogao pretraživati tekst putem određenih riječi.

„Tvrtka je znala da informacije imaju pohranjenu vrijednost koja može biti oslobođena jedino kada se dataficiraju. Google je upotrijebio OCR softver (eng. *optical-recognition software*) koji uzima digitalnu sliku i na njoj prepoznaje slova, riječi, rečenice i paragrafe“ (Mayer-Schönberger, Cukier; 2013:84). Na taj način informacije na stranicama postale su dostupne obradi putem računala i algoritamskoj analizi, a ne samo čitateljima. S mogućnošću pretrage pojedinih riječi i tekstualne analize može se otkriti učestalost upotrebe riječi ili fraza, kada su se one prvi put upotrijebile, a pomaže i pri otkrivanju plagijata u akademskim radovima.

3.2.Lokacije kao podaci

Geografska lokacija prirode, objekata i ljudi također sadrži vrijedne informacije. Kvantifikacijom i standardiziranim načinom mjerenja možemo sakupiti, pohraniti i analizirati lokacije kao podatke, a ne kao mjesta. Većina nas koristi GPS (eng. *global positioning system*) koji nas triangulacijom sa setom satelita obavještava o našoj poziciji unutar geografskog prostora, ali postoje i druge metode otkrivanja pozicije preko mobilnih antena, wifi-ja, analize scene. Saadi Lahlou (2008) opširnije se bavi ovom temom i objašnjava kako telefonski poslužitelji mogu trenutno prostorno locirati mobilne telefone korisnika te kako neke ljude već samo postojanje ovih mogućnosti uznemirava. Međutim ovo je samo početak u usporedbi s mogućnošću da kontinuirano geolociranje otkrije individualne putanje. Putanje otkrivaju smjer i vremenski slijed pozicija pa se može indicirati uzročnost, a ako se radi serija putanja za određenu osobu može se pristupiti njezinim navikama. Na primjer, PlaceEngine - pametni sustav geolociranja, s

korisnikova mobitela šalje sistemu vektore pristupnih točaka, te tako omogućuje stvaranje životnih oznaka (*lifetag*). LifeTag je sistem za praćenje života (*life-logging*) koji upotrebom jednostavnog seta pravila, automatsko praćenje lokacije može pretvoriti u interpretaciju aktivnosti. Praćenje putanja može se koristiti i za zaključivanje i predviđanje; ako korisnik često koristi određenu putanju, konačna destinacija može se saznati iz početne točke što mu omogućuje da unaprijed zna gdje i kada će neka osoba biti. Ograničavanjem na ceste i prirodu mjesta (supermarket, aerodrom) sustav s određenom vjerojatnošću može zaključiti gdje će putanja završiti uzevši u obzir baze podataka uobičajenih kretanja većine populacije. Ako zaključi da osoba ide na aerodrom može mu poslati kupon za parkirno mjesto. Ovaj sistem može pomoći ljudima s problemima u pamćenju (Alzheimer), predvidjeti gužvu u prometu, otkriti područja zaražena gripom, ali važno je napomenuti i njegovu mračnu stranu u vidu kontrole ljudi na vrlo suptilne načine.

3.3. Interakcije kao podaci

Datafikacija je zahvatila i osobnije aspekte ljudskog života kao što su odnosi, iskustva i raspoloženja. Osim mogućnosti da ostanemo u kontaktu sa starim prijateljima, društvene mreže koriste te nematerijalne elemente našeg svakodnevnog života i transformiraju ih u podatke. Facebook bilježi poznanstva, a Twitter osjećaje i misli. LinkedIn pomoću informacija o našim profesionalnim iskustvima predviđa našu prošlost i budućnost, primjerice, kolika je vjerojatnost da poznajemo neku osobu i koji poslovi bi nam se mogli svidjeti. Twitter poruke ograničene su na 140 znaka, ali metapodaci (podaci o podacima) kao što su korisnikov jezik, geolokacija, broj i imena ljudi koje prati i koji prate njega, također su zanimljivi za proučavanje. „U jednoj studiji, objavljenoj u časopisu Science 2011. godine, analiza od 509 milijuna tweetova preko dvije godine s 2,4 milijuna ljudi iz 84 zemlje pokazala je kako raspoloženja ljudi prate slične dnevne i tjedne obrasce u kulturama diljem svijeta – što nije bilo moguće uočiti prije“ (Mayer-Schönberger, Cukier; 2013:93). Ovakve analize veliki su potencijal za otkrivanje društvene dinamike na svim razinama, od pojedinaca do društava.

4. Promjene u razmišljanju o podacima

U sljedeća tri potpoglavlja opisat će se tri promjene koje Veliki podaci zahtijevaju u načinu na koji razmišljamo o njima, a koje mijenjanju naše dosadašnje razumijevanje i organizaciju društva. Prva se odnosi na količinu podataka, druga na odricanje od točnosti u zamjenu za širu sliku, a treća na potrebu da se odmaknemo od uzročnosti.

4.1. Velika količina podataka ili uzorkovanje

Živimo u svijetu u kojem je moguće obraditi više podataka nego ikada prije, a s obzirom na određeni fenomen, ponekad i sve. Od 19. stoljeća pri susretu s velikim brojevima društvo je koristilo uzorke, međutim, uzorkovanje pripada periodu analogne tehnologije kada je informacija bilo malo ili kada je dolazak do njih bio skup. Korištenje svih podataka omogućuje nam da jasnije vidimo pojedinosti i potkategorije koje je nemoguće otkriti pomoću uzorkovanja. U prošlosti su čak razvijene tehnike koje koriste što je manje podataka moguće, pa je tako i jedan od ciljeva statistike izvući najbolji zaključak sa što manje podataka, to jest pomoću uzorka opisati populaciju. Sve do nedavno privatne tvrtke, a danas i pojedinci, nisu bili u mogućnosti sakupiti i sortirati informacije u većim razmjerima. Taj zadatak u prošlosti obavljale su moćne institucije kao što su crkva ili država koja je željela procijeniti koliko ljudi u njoj živi te kolika joj je površina. Tako se i prema novijem dobu razvio popis stanovništva čija je provedba kompleksna, skupa, i oduzima puno vremena.

Mayer-Schönberger i Cukier (2012) objašnjavaju kako su ljudi uvidjeli da nije uvijek praktično iskoristiti sve podatke pa se pojavila ideja uzorkovanja i pitanje kako odabrati uzorak? Logično se činilo kako bi uzorak trebalo namjerno konstruirati kako bi reprezentirao cjelinu. Međutim, 1934. godine Jerzy Neyman, poljski statističar, dokazao je kako takav pristup dovodi do velikih pogrešaka (Mayer-Schönberger; Cukier, 2012:22). Ključ kako bi se one izbjegle je ciljati na slučajnost prilikom odabira uzorka.

Preciznost se dramatično povećava sa slučajnošću, a ne s veličinom uzorka. Slučajno odabran uzorak od 1100 ispitanika koji odgovaraju na binarno pitanje s odgovorom „da ili ne“ prilično sigurno reprezentira cijelu populaciju. U biti odgovor je u 99% slučajeva točan unutar 3% statističke pogreške, bez obzira je li veličina populacije 100 tisuća ili 100 milijuna. Matematički gledano, nakon kratkog vremena kako brojevi postaju sve veći i veći, marginalna količina novih informacija koje naučimo iz svake opservacije je sve manja i manja. Jednostavnije, unutar 1100 ispitanika uspijevamo zahvatiti sve varijacije, a zatim se odgovori počinju ponavljati. Uzorkovanje se pojavilo kao novi pristup prikupljanja informacija s manjim troškovima, a velikom preciznošću. Ono je bilo rješenje za problem prevelike količine informacija u vrijeme kada je sakupljanje i analiza podataka bila prilično kompleksan i težak zadatak. Važno je napomenuti da je ova metoda doprinijela razvoju društvenih znanosti iz humanističkih (Mayer-Schönberger; Cukier, 2012:23).

Slučajno uzorkovanje bilo je veliki uspjeh za dotadašnje načine mjerenja, međutim ono je samo druga najbolja alternativa u odnosu na sakupljanje i analizu svih podataka. Preciznost uzorkovanja ovisi o slučajnom odabiru prikupljenih podataka, međutim osiguranje takve slučajnosti je rizično jer postoji niz pristranosti. Na primjer ako se anketa o stranačkim preferencijama radi pomoću kućnih telefona, taj uzorak je pristran jer zanemaruje ljude koji koriste samo mobilne telefone (većinom mlađe i liberalnije generacije). Ali, čak i ako preciznost bude zadovoljavajuća, slučajnim uzorkovanjem teško je kasnije uzorak podijeliti na potkategorije kao što su rod, područje, prihod (zato se koristi postupak ponderiranja[1]). Uzmimo kao primjer situaciju u kojoj želimo ispitati određenu nišu unutar populacije kao što su ženske glasačice koje žive u zapadnom dijelu grada. U sveukupnom uzroku od 1000 ljudi njihov postotak je znatno manji. Tu nastaje problem gdje na osnovu svega nekoliko zapažanja predviđamo glasačke namjere svih ženskih glasačica u zapadnom dijelu grada, što vodi k velikoj nepreciznosti čak i ako je uzorak blizu savršene slučajnosti. Uzorkovanje je korisno na makro razini, ali na mikro

[1] “Ponderiranje (prema lat. ponderare: vagati, mjeriti) ili vaganje, postupak kojim se određuje odgovarajuća vrijednost pojedinih veličina prilikom izračunavanja srednje vrijednosti (Hrvatska enciklopedija, 2013-2015).” Statističko ponderiranje podataka služi „kako bi se prethodno narušeni udjeli u uzorku doveli u odnos koji vlada u populaciji“ (Milas, 2008:426).

razini ono ne daje dovoljno jasne zaključke. Također, ono zahtijeva pažljivo planiranje i izvršavanje, te je teško iste podatke ponovno analizirati na potpuno novi način u odnosu na svrhu s kojom su prvotno prikupljeni. U nekim slučajevima nema drugog načina i uzorkovanje je svakako i dalje iznimno koristan alat, ali u mnogim drugim područjima dolazi do promjene od prikupljanja nekoliko do sakupljanja što više podataka ili ako je moguće svih. Prikupljanjem svih podataka moguće je uočiti povezanosti i detalje koji bi se inače izgubili u pustoši informacija. Na primjer, kako bi se otkrile prevare s kreditnim karticama traže se anomalije, to jest sakupljaju se svi podaci, a ne samo uzorci. Odstupanja je moguće uočiti usporedbom s uobičajenim transakcijama, a s obzirom da se transakcije događaju u trenutku, potrebno je da se i analiza događa u realnom vremenu. Također, jedan od poznatih slučajeva je i otkriće namještanja utakmica u japanskom nacionalnom sportu – sumo hrvanju, opisanih u knjizi i filmu Freakonomics (Levitt; Dubner, 2009). Poanta je u tome da se korištenjem Velikih podataka ponekad traži nešto određeno, no usput se otkrije još nešto. U slučaju sumo hrvanja očekivalo su da će namještene utakmice biti one kada su se borili prvaci, međutim analiza je pokazala da se namještanja obično događaju na kraju turnira. Ako ne znamo što tražimo teško ćemo to otkriti pomoću uzorkovanja jer nećemo znati koji uzorak koristiti. Važno je napomenuti da nije uvijek nužno koristiti sve podatke umjesto uzoraka jer i dalje živimo u svijetu s ograničenim resursima, ali jedno od područja koje je upotreba svih podataka (N=sve) značajno promijenila su društvene znanosti.

Preuzima li analiza pomoću Velikih podataka monopol od stručnjaka za istraživanje društvenih fenomena, koji se uvelike oslanjaju na uzorkovanje i upitnike? Metode kojima se koriste svakako su i dalje korisne, ali statističko uzorkovanje je koncept star više od stoljeća, i razvijen je kako bi se riješili određeni problemi, u određenom vremenu koje su obilježila tehnološka ograničenja. S obzirom da ta ograničenja više ne postoje, barem ne u tolikoj mjeri, za pretpostaviti je kako niti „uzorkovanje više neće biti dominantan način za analizu velikih količina podataka“ (Meyer-Schönberger; Cukier, 2012:31).

4.2. Prihvaćanje nereda ili točnost

Druga promjena koju donosi doba Velikih podataka može se opisati kao svojevrsna potreba da napustimo koncept točnosti. Koncept točnosti primjenjiv je na okolinu s malo podataka – kada postoji mala količina stvari koju možemo mjeriti onda je nužno da ih izmjerimo što točnije. To je samo po sebi i logično, na primjer, vlasnik apartmana na moru mora znati točan broj osoba koje će doći kako bi svima osigurao mjesta za spavanje, ali kada želimo znati koliko turista posjeti Hrvatsku nije potrebno da znamo točan broj nego približnu procjenu. S povećanjem skale (opsega) povećava se i netočnost. U nekim situacijama točnost je nužna, ali često nam je dovoljna opća procjena, to jest, s gubljenjem točnosti na mikro razini dobivamo uvid u makro razinu. Ova promjena je na neki način obrnuta od prethodne u kojoj pomoću analize mnoštva podataka dobivamo uvid u detalje i potkategorije. U svijetu malih podataka, otklanjanje pogrešaka i osiguranje visoko kvalitetnih podataka bilo je nužno. U Francuskoj u 19. stoljeću s razvojem znanosti razvio se cijeli sistem precizno definiranih jedinica mjerenja kako bi se izmjerio prostor, vrijeme, a težilo se i tome da ostale nacije prihvate iste standarde. U povijesti postoji cijeli niz primjera koji potvrđuju zahtjeve za točnošću, vjerovalo se ako se neki fenomen može izmjeriti, onda ga se može i razumjeti (Meyer-Schönberger; Cukier, 2012:33).

Nered kod Velikih podataka nastaje na različite načine kao što je nekonzistentnost do koje dolazi prebacivanjem podataka u isti format ili kombiniranjem različitih vrsta informacija iz različitih izvora. Ponekad je korisno prihvatiti malo nereda u zamjenu za spoznaju o općim kretanjima stvari, zbog toga se ovaj aspekt promjene orijentira na vjerojatnost više nego na preciznost. Dobro poznati Mooreov zakon govori kako se broj tranzistora na čipu udvostručuje svake dvije godine, što dovodi do toga da su računala sve brža, a povećava se i količina memorije. Izvedbe algoritama koji pokreću naše sustave su poboljšane, međutim, mnoge koristi za društvo nisu rezultat samo bržih čipova i boljih algoritama nego su direktna posljedica veće količine podataka. Na primjer, iako su pravila šaha davno poznata, računala danas igraju završne poteze puno bolje zato što

su u njih zapisani svi mogući potezi, što je dovelo do toga da čovjek više ne može pobijediti računalo.

Jedan od primjera učinkovitosti nereda nalazimo u području strojnog prevođenja jezika. IBM-ov projekt Candide koristio je 3 milijuna pažljivo prevedenih rečenica pronađenih u službenim dokumentima engleskog i francuskog jezika sa ciljem što veće kvalitete prijevoda (Meyer-Schönberger; Cukier, 2012). Za razliku od njih Google je sakupio 3 bilijuna stranica prijevoda različite kvalitete. Korpus riječi uzet je s nefiltriranih web stranica koje sadrže nedovršene rečenice, pravopisne i gramatičke pogreške i niz drugih manjkavosti, ali njegova usluga prevođenja zasada je najbolja. Ne zato što ima pametniji algoritam nego zbog količine podataka, a ključ uspjeha leži upravo u prihvaćanju nereda. Međutim, konvencionalnom analitičaru koji se bavi uzorkovanjem i koji se fokusira na prevenciju nereda to je teško prihvatiti. Redukcija pogrešaka pri sakupljanju uzoraka, testiranje uzoraka na potencijalne pristranosti, briga da se uzorci skupe prema točnom protokolu od strane treniranih stručnjaka – skupe su strategije čak i kada se radi o manjem broju podataka, a problem predstavlja i subjektivni utjecaj anketara. Primjena konvencionalnih metoda mjerenja na digitalni, umreženi svijet 21. stoljeća koristan je dodatak, ali ne i najučinkovitija metoda. Iako se sukobljava s intuicijom, tretiranje podataka kao nesavršenih i nepreciznih omogućuje nam napredne prognoze i bolje razumijevanje svijeta.

Promjena se događa i kod tradicionalnih baza podataka koje zahtijevaju visoko strukturirane podatke i preciznost. Podaci nisu samo prikupljeni i spremljeni u baze nego su „razbijeni“ na zapise koji se nalaze u poljima. Svako polje sadrži informaciju određenog tipa i duljine. Konvencionalne, relacijske baze podataka dizajnirane su za svijet u kojem podataka ima malo. Napravljene su kako bi efikasno i točno odgovorile na jednostavna pitanja. Međutim, danas imamo velike količine podataka različitih tipova i kvalitete koji se ne mogu uklopiti u prije definirane kategorije. „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“ (Kocijan, 2014:2).

Obrada velike količine podataka dovodi do gubitka informacija, ali ti gubitci se nadoknađuju brzinom koja je često ključan faktor u nekim sektorima kao što je javno zdravstvo (detekcija virusa) ili predviđanje inflacije. Kako bi dobili na brzini, velike količine podataka su zapisane na različitim lokacijama (serverima). Ova promjena realizirana je, na primjer, kroz Hadoop - softver otvorenog koda namijenjen za obradu velike količine podataka koji pretpostavlja da su podaci preveliki da bi se prvotno sortirali prema tipu, nego ih odmah analizira (Schneider, 2012). Rezultati nisu tako precizni kao u relacijskim bazama podataka i neće se koristiti prilikom lansiranja rakete, ali su primjenjivi na niz aktivnosti koje ne zahtijevaju apsolutnu preciznost.

Stvarnost je dovela do novog dizajna baza podataka. Najpoznatiji računalni jezik za pristup relacijskoj bazi podataka je SQL (eng. *structured query language*) koji se koristi za obradu strukturiranih podataka. Promjena dolazi s novim, takozvanim noSQL jezikom koji se koristi za pristup nerelacijskoj bazi podataka koja prihvaća podatke različitih tipova i veličina. U zamjenu za nered u strukturi, nerelacijske baze podataka sastoje se od više korisnih resursa jer omogućuju pohranjivanje i pretraživanje polustrukturiranih i nestrukturiranih podataka. U doba Velikih podataka, mnogi veliki skupovi podataka sastoje se od polu ili nestrukturiranih podataka, kao što su Facebook postovi, tweetovi, slike, videozapisi, i blogovi, a neke procjene govore da takvi podaci rastu 15 puta brže od strukturiranih podataka (Kitchin, 2014 prema Zikopoulos et al., 2012). Prema IDC-ovom (eng. *International Data Corporation*) izvještaju iz 2011. godine, nestrukturirani podaci čine više od 90% svih podataka (Gantz; Reinsel, 2011). Kao novi, relativno neiskorišteni izvor uvida, analiza nestrukturiranih podataka može otkriti važne međuodnose koje je prethodno bilo teško ili nemoguće utvrditi. Bez prihvaćanja nereda, nestrukturirani podaci kao što su web stranice, postovi, slike i videozapisi ostaju neistraženi.

4.3.Korelacija ili uzročnost

Veliki podaci zahtijevaju promjene u našem načinu razmišljanja. Predlažu nam da se odmaknemo od točnosti i preciznosti u zamjenu za širu sliku i dovode u pitanje

uvriježene metode znanstvenog istraživanja. Ove radikalne promjene s kojima se društvo treba suočiti veliki su izazov, a dovode i do trećeg prijedloga koji negira fundamentalnu konvenciju na kojoj se temelji naše društvo. Radi se o odbacivanju zahtijeva za spoznajom uzroka u svim aspektima ljudskog djelovanja. Kao ljudi, na neki način smo uvjetovani da u svemu tražimo uzrok, štoviše razmišljanje na relaciji uzrok – posljedica dovelo je do niza otkrića i razvoja čovječanstva. Međutim, potraga za uzrocima ponekad je teška i može odvesti na krivi put. Veliki podaci predlažu da se odmaknemo od fiksacije na uzročnost te zauzvrat otkrijemo korelacije koje nam ne moraju nužno reći „zašto" se nešto događa nego nas samo obavijestiti da se to događa.

Korelacija je postupak mjerenja dviju ili više varijabli i određivanja odnosa među njima. „Korelacija je jedan od temeljnih pojmova u znanosti (…), a utvrđivanje povezanosti između pojava jedan je od temeljnih ciljeva znanosti. Prema gruboj procjeni oko 80% svih objavljenih znanstvenih članaka imali su za cilj utvrđivanje povezanosti varijabli" (Mejovšek, 2003:150). Varijabla je „svojstvo ispitanika, podražaja ili situacije koje može poprimati različite vrijednosti" (Milas, 2009:105). Na primjer, niži socioekonomski status povezan je sa sklonosti prema kriminalu. Neke od varijabli su visina, težina, inteligencija, depresivnost, intenzitet svijetla i slično. Visoka korelacija znači da promjena u jednoj varijabli vrlo vjerojatno uzrokuje promjenu u drugoj, a niska korelacija da promjena jedne varijable ne utječe znatno na drugu. Međutim, snažne korelacije nikad nisu savršene, a moguće je i da su naizgled povezane varijable stvar slučajnosti. S korelacijama ne postoji stopostotna sigurnost, samo vjerojatnost – u kojoj mjeri se na temelju jedne varijable može predvidjeti druga. Korelacijski pristup pokazuje samo jesu li dvije varijable povezane, ali ne i zašto su povezane, odnosno koja je uzrok, a koja posljedica. „Nezavisna ili eksperimentalna varijabla je svojstvo kojim eksperimentator upravlja sustavno ga mijenjajući kako bi provjerio njegov utjecaj na proučavano ponašanje. Pritom se pretpostavlja da će nezavisna varijabla utjecati na neko ponašanje koje se opaža i mjeri (zavisnu varijablu). Eksperimentalna metoda nalaže da se nezavisna i zavisna varijabla unaprijed precizno odrede i operacionalno definiraju kako bi se izbjegli svi mogući nesporazumi i omogućilo ponavljanje istraživanja" (Milas, 2009:106).

U prošlosti podataka je bilo malo, a njihovo prikupljanje skupo. Stručnjaci su koristili hipoteze izvedene iz teorija – apstraktnih ideja o tome kako nešto radi. Na osnovu tih hipoteza sakupljali su podatke te pomoću korelacije provjeravali jesu li njihove pretpostavke točne. „Većina teorija nastala je na temelju ograničenog iskustva, odnosno ograničenog broja podataka i ograničenog broja istraživanja. Zato većina teorija ima određenih nedostataka ili pojedini dijelovi teorija nisu znanstveno verificirani. (…) Operacionalno, može se kazati da je cilj znanstvenih istraživanja provjera hipoteza. Hipoteze su tvrdnje koje su dijelovi teorija, i to bitni dijelovi teorija. To su tvrdnje na koje se teorija oslanja“ (Mejovšek, 2003:20). U doba Velikih podataka nije efikasno donositi odluke o tome koje varijable ispitati oslanjajući se samo na hipoteze.

U knjizi „The power of habit“, Charles Duhigg opisuje slučaj američke maloprodajne tvrtke Target koja se godinama oslanja na predviđanja bazirana na korelaciji Velikih podataka. Duhigg (2012) objašnjava kako Target zna da su žene trudne pomoću analize njihovih kupovnih obrazaca. Tim analitičara pregledao je povijest kupovine svih žena koje su se prijavile na njihovu registraciju za darove djeci. Primijetili su kako oko trećeg mjeseca trudnoće žene kupuju puno bezmirisnih losiona, a nekoliko tjedana kasnije dodatke kao što su magnezij, kalcij i cink. Pomoću ova dva proizvoda uspjeli su izračunati vjerojatnost trudnoće za svakog kupca koji je plaćao karticom. Korelacije su čak omogućile da se procijeni datum poroda s malim odstupanjem kako bi se mogli slati relevantni kuponi za svaki stadij trudnoće. O ovom slučaju analitičari nisu sami pretpostavili što žene kupuju, a potom to išli ispitati, nego su dopustili da analiza sama odradi posao.

Još jedna od metoda koja se često koristi naziva se prediktivna analiza koja se koristi u poslovanju kako bi se događaji predvidjeli i prije nego što se dogode. S tim ciljem napravljen je i softver koji pomaže doktorima da donesu bolje odluke kad se radi o prerano rođenoj djeci. „Softver sakuplja i obrađuje podatke o pacijentu u realnom vremenu, prati 16 različitih tokova podataka, kao što su, otkucaji srca, stopa disanja, temperatura, krvni tlak, razina kisika u krvi, što zajedno iznosi oko 1,260 podataka u sekundi“ (Meyer-Schönberger; Cukier, 2012:60). Sistem detektira promjene u stanju djeteta koje mogu signalizirati infekcije 24 sata prije nego što se pojave očiti simptomi.

Čovjek to ne može prepoznati, ali računalo može. Također, ovdje se radi o vjerojatnosti, a ne o uzroku, govori što se događa, a ne zašto. No svakako služi svrsi, iako algoritam sam po sebi ne donosi odluke, pomaže liječnicima da ih donesu.

U svijetu s malo podataka, istraživanja uzroka i korelacijska analiza započinjale su s hipotezama koje su potom bile potvrđene ili opovrgnute. Potrebni podaci često nisu bili dostupni, a hipoteze su bile podložne predrasudama i krivoj intuiciji. Danas, kada je količina podataka ogromna, takve hipoteze više nisu ključne za korelacijsku analizu. Također, većina korelacijskih analiza bila je ograničena na linearne povezanosti. Sa sofisticiranom analizom moguće je otkriti i odnose koji su često puno kompleksniji, kao što je to slučaj s većinom fenomena koji su pod istovremenim utjecajem različitih varijabli.

5. Preoblikovanje istraživanja

Ovo poglavlje ispituje način na koji dostupnost Velikih podataka i novih analitičkih alata dovodi u pitanje etablirane epistemologije unutar različitih disciplina. Utječu li značajno promjene u načinu prikupljanja, obrade i analize podataka na metodologiju društvenih istraživanja? Zahtijeva li „podatkovna revolucija“ promjenu paradigme i promjenu uobičajenog redoslijeda metoda zaključivanja (dedukcije i indukcije)? Kako bismo se približili odgovoru na ova pitanja potrebno je prvotno definirati termine epistemologije, osnovnih metoda zaključivanja i paradigme.

5.1. Epistemologija

Epistemologija je „filozofska disciplina zaokupljena znanjem i spoznajom, a bavi se, krajnje pojednostavljeno, temeljnim pitanjima koja mogu biti postavljena u potrazi za istinom“ (Milas, 2009:14). Epistemologija „u filozofiji, izvorno označuje »nauk o znanosti« (*mathesis universalis*) ili filozofijsku disciplinu koja istražuje uvjete, mogućnosti i granice znanstvene spoznaje i zato je srodna s epistemološkom logikom. Predmet je epistemologije dakle znanost, a ne spoznaja. Od nauka o spoznaji epistemologija se razlikuje po tome što ona u sebi obuhvaća: (1) podjelu znanosti, (2) spoznajna načela i iz njih izvedene pomoćne stavove (aksiomatika), (3) znanstvene metode (teorija stvaranja teorija i metodologija) i (4) izgradnju jezika u pojedinim znanostima“ (Hrvatska enciklopedija, 2013-2015) .

5.2. Metode zaključivanja

Do teorija u znanosti dolazi se na temelju procesa indukcije i dedukcije koje su temelji znanstvene spoznaje. „Indukcija je zaključivanje od pojedinačnog k općem, a dedukcija je zaključivanje od općeg k pojedinačnom. U traganju za znanstvenom

spoznajom procesi indukcije i dedukcije međusobno su povezani“ (Mejovšek, 2003:20), oni se nadopunjuju i isprepliću te ih je u spoznajnoj praksi teško razdvajati. Mejovšek (2003) opisuje uobičajeni proces istraživanja te navodi kako u preliminarnoj fazi istraživanja istraživač prikuplja informacije proučavanjem literature, ili početnim istraživanjima problema. Takva preliminarna induktivna istraživanja pomažu istraživaču da bolje upozna problem te oblikuje hipotetičku teoriju o problemu do koje dolazi principom indukcije. S obzirom da to nije konačna teorija, iz hipotetičke teorije po principu dedukcije, istraživač generira hipoteze koje treba provjeriti znanstvenim istraživanjem.

Postoje dva opća pristupa u znanstvenom istraživanju: zbirno-induktivni koji više odgovora prirodnim znanostima i funkcionalno-deduktivni koji dominira u društvenim i humanističkim znanostima. Zbirno-induktivni pristup sastoji se od prikupljanja velikog broja podataka o nekoj pojavi da bi se na kraju oblikovala teorija koja je samo cilj, a ne i metoda kojom se dolazi do znanstvenih spoznaja. „Polazište tog pristupa je da se svijet može spoznati i bez hipoteza i da je spekulativno teoretiziranje zapreka opažanju“, a prigovor je da „ne postoji čisto opažanje bez udjela teorije, jer svaki istraživač ima barem neku pretpostavku o problemu koji istražuje, odnosno o podacima koje prikuplja“ (Mejovšek, 2003:22). Kod funkcionalno-deduktivnog pristupa postoje oba smjera – razvijaju se teorijske postavke koje se zatim provjeravaju i korigiraju u skladu s empirijskim podacima. Ovaj pristup nudi objašnjenje i razumijevanje pojava i njihovih uzroka usmjeravajući istraživanje te je zato općenito prihvaćen.

5.3. Paradigma

„Paradigma je set ili usmjerenje, konceptualni okvir unutar kojeg znanstvenik djeluje. Radi se o spremnosti da se nešto percipira, a nešto izostavi. (…) Khun (1999) koji je proučavao povijest prirodnih znanosti opisuje paradigmu kao skup temeljnih pretpostavki koje definiraju područje znanstvenog proučavanja, određujući vrstu problema i metoda koje se smatraju legitimnima i koje se mogu upotrijebiti za prikupljanje i tumačenje podataka“ (Mejovšek, 2003:32).

Termin paradigma „podrazumijeva da neki općeprihvaćeni primjeri konkretne znanstvene prakse – kojima su razvijene teorije, zakoni, njihova primjena i djelovanje – pružaju modele na temelju kojih nastaju određene koherentne tradicije znanstvenog istraživanja (Sardar, Vaan Loon, 2005:49). Istraživači koju djeluju unutar iste paradigme slijede sličnu filozofiju i koriste slične metode. „Stoga imaju tendenciju da favoriziraju iste ili vrlo slične ontološke, epistemološke, teorijske, metodološke i etičke te ideološke okvire“ (Kitchin, 2014:128). Međutim, s vremenom se pojave novi načini razmišljanja koji preispituju prihvaćene teorije i pristupe. „Kada postojeća paradigma više ne rješava probleme određene struke, javlja se potreba za novom paradigmom (…) Problem je u tome što nova paradigma ili sveobuhvatna važnija teorija nije jednostavan dodatak postojećoj paradigmi, već se postavlja zahtjev za preispitivanjem i rekonstrukcijom postojeće paradigme“ (Mejovšek, 2003:32).

Kitchin (2014) navodi niz primjera kao što je Einsteinova teorija relativnosti koja je zamijenila Newtonovu teoriju, Darwinova teorija evolucije koja je promijenila način razmišljanja u biologiji i suočila se sa široko rasprostranjenim kreacionizmom. U oba slučaja ljudi su se podijelili: neki su se držali starih načina mišljenja, dok su se drugi priklonili novima koji su ubrzo prevladali. U nekim znanostima kao što su sociologija ili humana geografija postoji cijeli niz filozofskih pristupa kao što su pozitivizam, fenomenologija, strukturalizam i mnogi drugi od kojih svaki smatra kako upravo on daje najbolje objašnjenje svijeta u kojem živimo.

U ovom radu usredotočit ću se na „paradigmu Velikih podataka“ i njezinu mogućnost da dovede do alternativnih epistemologija u društvenim znanostima. S obzirom da je ova tema još uvijek relativno nova postoje različita mišljenja o posljedicama Velikih podataka i noviteta koje donose, ali sigurno je da postavljaju fundamentalna epistemološka pitanja jer izvlačenje korisnih informacija iz podataka nije samo problem tehničke naravi.

Brooks (2013) filozofiju ove paradigme naziva „data-ism“, te smatra kako se sve što se može izmjeriti i treba izmjeriti, te kako nam podaci pomažu da filtriramo emocije i ideologije i predvidimo budućnost. Korak dalje otišao je Chris Anderson koji je u časopisu Wired objavio članak „The end of theory: the data deluge makes scientific

method obsolete“ u kojem raspravlja kako je korelacija dovoljna, to jest kako nadmašuje uzročnost. Smatra kako više nema potrebe da se traže modeli jer se podaci mogu analizirati bez prethodnih hipoteza, dovoljno je ubaciti sve podatke u statistički algoritam koji će sam pronaći obrasce putem kojih se mogu objasniti društveni, ekonomski i politički procesi i drugi kompleksni fenomeni.

U odnosu na prethodna istraživanja, mijenja se redoslijed stvari. Algoritmi obavljaju kontekstualni posao, sami pronalaze obrasce, a potom se stvaraju hipoteze i teorije. Umjesto da se na određenom skupu podataka provjerava jesu li varijable unutar hipoteze povezane, algoritam sam otkriva postoji li povezanost među podacima bez da je vođen hipotezama. Navedeni argumenti predlažu da se prevladavajući funkcionalno-deduktivni pristup u istraživanjima zamijeni hibridnom kombinacijom u kojoj će se započeti sa zbirno-induktivnim pristupom unutar kojeg će prikupljanje i korelacijsku analizu podataka odraditi statistički algoritam, a kada se otkriju korelacije može se nastaviti funkcionalno-deduktivnim pristupom unutar kojeg će stručnjaci formulirati hipoteze.

5.4. Ideje na kojima se temelji nova znanost

Kitchin (2014) opisuje 4 ideje na kojima bi se temeljila nova znanost. „Prvo, da Veliki podaci mogu obuhvatiti cijelu domenu i pružiti cijelu rezoluciju. Drugo, da nema potrebe za apriori teorijama, modelima i hipotezama. Treće, da primjenom agnostičke analize podataka podaci mogu govoriti za sebe oslobođeni od ljudskih pristranosti i okvira, te da je bilo koji obrazac ili korelacija unutar Velikih podataka svojstveno značajna i vjerodostojna. Četvrto, to značenje stvara kontekst ili znanje specifično domeni“ (Kitchin, 2014:132). Međutim, kod svake od ovih ideja postoje zablude koje je potrebno rasvijetliti.

5.4.1. Veliki podaci mogu obuhvatiti cijelu domenu i pružiti cijelu rezoluciju

Veliki podaci su iscrpni, ali su i dalje uzorci, to jest, ne pružaju cjeloviti pogled na svijet nego pogled s određenog gledišta i korištenjem specifičnih alata. Možda je moguće zabilježiti sve aktivnosti na Facebook-u, ali će i dalje postojati poznanstva koja nisu zabilježena na Internetu, te zbog toga nisu podatak. I dalje postoji niz domena koje zahvaćaju zbilju koju tehnološka revolucija još nije obuhvatila. Također, domene se razvijaju i mijenjaju te samim time nema smisla tvrditi da je moguće obuhvatiti nešto u cijelosti.

5.4.2. Nema potrebe za apriori teorijama, modelima i hipotezama

Prethodno objašnjena induktivna metoda kojom se prvo pronalaze obrasci u podacima također je vođena znanstvenim rezoniranjem. „U stvari, i deduktivno i induktivno rezoniranje uvijek je diskurzivno uokvireno i ne može proizaći iz ničega“ (Kitchin, 2014:134). Nemoguće je izvući objašnjenja iz statističke analize bez da prvo postoji barem neka teorija. Čak i statističke algoritme koji se koriste za obradu podataka netko je morao napisati, a kako bi ih napisao poslužio se nekim vrijednostima i kontekstom unutar određenog znanstvenog pristupa. Različiti algoritmi pronaći će različite obrasce.

5.4.3. Primjenom agnostičke analize podataka podaci mogu govoriti za sebe oslobođeni od ljudskih pristranosti i okvira

Ova ideja vodi se time da osim što se podaci stvaraju slobodno od teorije, i njihova interpretacija i značenje postoji samo po sebi. Međutim, podaci nisu prirodni i esencijalni elementi izvučeni iz svijeta na neutralan i objektivan način. Također, tvrdnja da korelacija nadilazi uzročnost također može biti opasna jer korelacije među varijablama mogu biti slučajne. Iako se na ovaj način mogu otkriti zanimljive povezanosti ipak se većina korelacija treba ponovno testirati na novom skupu podataka kako bi se osigurala

pouzdanost i valjanost. „Drugim riječima, korelacije trebaju tvoriti osnovu za hipoteze koje će se ponovno testirati, a potom iskoristiti za stvaranje ili redefiniranje teorije koja ih objašnjava. Stoga korelacija ne nadilazi uzročnost, nego tvori osnovu za dodatna istraživanja koja trebaju ustanoviti jesu li te korelacije indikatori uzročnosti“ (Kitchin, 2014:135).

5.4.4. Značenje stvara kontekst ili znanje specifično domeni

Interpretacija Velikih podataka ne zahtijeva prethodno postojeći kontekst ili znanje svojstveno određenoj disciplini, štoviše, jedino je potrebno da se značenje koje podaci stvaraju sami po sebi učini vidljivo kako bi ga svatko sa osnovnim statističkim znanjem mogao interpretirati. Ali pri stvaranju određenih alata računalni stručnjaci se koriste znanjem iz drugih disciplina, a često se njime služe i za interpretaciju podataka. Nažalost njihovo poznavanje područja koje komentiraju često je vrlo površno što dovodi do redukcionističkih i funkcionalističkih analiza čija je korisnost upitna. Nameće se pitanje oko toga tko ima najveći legitimitet za stvaranje znanja u određenom području. Tu se vraćamo na već spomenuto razgraničenje između „what“ i „why“. Najuspješnija je svakako suradanja između podatkovnih znanstvenika koji mogu otkriti „što“ i eksperata pojedinih domena kojima treba prepustiti da odgovore „zašto“.

5.5. Znanost koja se temelji na podacima (*data-driven science*)

Podatkovna revolucija nudi mogućnost preoblikovanja epistemologije društvenih znanosti. Štoviše, ta promjena se već odvija, s obzirom da su Veliki podaci stvorili nove pristupe u analizi podataka koji omogućuju da se do određenih spoznaja dođe na nove načine. Važno je naglasiti da ne dolazi do inovacije koja će potkopati postojeću paradigmu i praksu istraživanja nego dovodi do puno produktivnijeg pristupa koji se temelji na njezinu preoblikovanju.

Epistemološka strategija „znanosti koja se temelji na podacima“ (eng. *data-driven science*) je koristiti „znanstveno-istraživačke tehnike kako bi se identificirala potencijalna

pitanja (hipoteze) koje se isplati dalje ispitivati i testirati“ (Kitchin, 2014:138). Filozofska podloga znanosti koja se temelji na podacima još je u povojima, a postoji potreba za promišljanjem i razradom njezinih epistemoloških načela i metodologije. Što se tiče društvenih znanosti situacija je nešto kompleksnija s obzirom na raznovrsnost filozofskih temelja, ali vjerojatnost uspostave potpuno nove paradigme je mala. Vjerojatnije je da će Veliki podaci unaprijediti podatkovnu analizu i omogućiti nove pristupe i tehnike, te da neće zamijeniti postojeće metode koje se temelje na analizi „malih“ podataka i uzorkovanju (Kitchin, 2014). S obzirom na brzinu kojom se događaju promjene u području Velikih podataka i podataka općenito, nužna su kritička promišljanja utjecaja podatkovne revolucije na epistemologiju društvenih znanosti.

6. Problemi Velikih podataka

Podaci se stvaraju i koriste za mnoge svrhe, kao što su upravljanje društvima, vođenje organizacija i stvaranje profita u slučaju kojih utječu na veću sigurnost, konkurentnost, produktivnost, učinkovitost, transparentnost i odgovornost, čineći društva demokratičnijima, međutim do kontradikcije dolazi što to čine kroz procese koji prate i iskorištavaju ljude te narušavaju njihovu privatnost. Međutim, podaci se ne koriste samo na dobre ili loše načine nego je situacija često kompleksnija od toga donoseći istovremeno pozitivne i negativne posljedice. Tako, na primjer, kartice vjernosti (eng. *loyalty cards*) koje nude razne trgovine pružaju svojim klijentima uštede, istovremeno im pokušavajući prodati više robe kako bi povećale dobit. U takvoj situaciji, etički problemi su relativno zanemarivi s obzirom da građani imaju izbor hoće li koristiti kartice ili ne, međutim do problema dolazi kada se taj izbor ukine. Primjerice, država vrši nadzor svih članova društva kako bi smanjila mogućnost terorističkih napada, gdje građani dobivaju veću sigurnost po cijenu privatnosti.

Etički, društveni i politički problemi vezani uz sakupljanje i korištenje podataka, već su duže poznati i raspravljani na znanstvenim i javnim forumima, što je dovelo do stvaranja profesionalnih etičkih smjernica i propisa koji određuju kako se treba odnositi s podacima. Najčešći problemi vezani su uz nadzor, privatnost, sigurnost podataka, profiliranje i tehnološko upravljanje. Kako se svaki od tih problema promišlja od strane aktera koji imaju različite interese kao što su znanost, poduzeće, država ili civilno društvo? Različite namjere sudionika koji su uključeni u raspravu o navedenim pitanjima dovodi do toga da nema lakih rješenja te da se odluke temelje na kompromisu. Također, brzina kojom tehnologija napreduje i kojom se stvaraju novi načini analize podataka, utječe na stalan priljev novih pitanja i čini dotadašnje zakonske okvire zastarjelima.

6.1.Nadzor i privatnost

Svakim danom, stvaraju se ogromne količine podataka, a države, poduzeća i organizacije civilnog društva ulažu puno vremena u sakupljanje podataka o svojim građanima, klijentima i članovima i njihovim aktivnostima. Sve je teže sudjelovati u svakodnevnom životu bez da se ostavi neki trag sudjelovanja zbog sve većeg posredovanja digitalnih tehnologija. „Čak i ako kupac ne koristi kreditnu karticu za kupnju robe u trgovini, njegova prisutnost je zabilježena nadzornim kamerama; čak i ako osoba koristi anonimno ime na društvenim medijima, njihove IP i MAC adrese su snimljene“ (Kitchin, 2014:167). Svakodnevno ostavljamo digitalne tragove, a najčešće nemamo kontrolu nad time kako se oni koriste. Pomoću digitalnih tragova moguće je saznati obrasce potrošnje, posao, putovanja i komunikacije pojedinaca, ali i različitih institucija. Otvoreni podaci (eng. *open data*) je „ideja da određeni podaci moraju biti besplatno dostupni svima za korištenje i objavljivanje ako to žele, bez ograničenja, autorskih prava, patenata ili drugih mehanizama kontrole“ (Auer et al., 2007:722). Tržište podataka (eng. *data market*) i inicijativa za otvorenim podacima omogućili su laku i široku dostupnost podataka.

Privatnost je višedimenzionalni pojam čije značenje ovisi o kontekstu unutar kojeg se upotrebljava. Većinom se odnosi na „prihvatljive prakse s obzirom na pristup i objavljivanje osobnih i osjetljivih podataka“ (Kitchin, 2014:168). Nema sumnje da se koncept privatnosti mijenja. Informacije koje su se prethodno smatrale privatnima (životopis, obiteljske fotografije i videozapisi, osobne i obiteljske priče) sada se dijele putem društvenih mreža kao što su LinkedIn, YouTube, Facebook, Twitter i mnoge druge. Svjedoci smo sve veće zabrinutosti za individualnu i kolektivnu privatnost koja je ugrožena eksponencijalno rastućom informacijskom i komunikacijskom tehnologijom (ICT). S obzirom na potencijal rasta ICT-a sigurno je da će pitanje privatnosti značajno oblikovati naš svijet u budućnosti. Ljudi, iako svjesni potencijalnog narušavanja vlastite privatnosti, i dalje u životu koriste nove tehnologije koje su postale nezaobilazan dio svakodnevnice. Ne poduzimaju dovoljno kako bi izbjegli ili spriječili narušavanje vlastite

privatnosti te unatoč upozorenjima i svjesnosti o potencijalnim opasnostima, na Internetu ostavljaju brojeve telefona, kreditnih kartica i mnoge druge osobne podatke.

Čest oblik potvrđivanja statusa, kojim dobivamo određenu moć i kontrolu, vrši se pomoću dokazivanja prošlih aktivnosti. Dolazimo do „dileme privatnosti“ (eng. *privacy dilemma*) unutar koje postoji snažna kontradikcija s održavanjem naših prošlih aktivnosti privatnima. Kako bi potvrdili svoj status ili mogli izvesti određene aktivnosti nužno je da otkrijemo drugima dio svog identiteta te im dopustimo da prate određene akcije, vode evidenciju o njima te ih prikazuju onima s kojima ulazimo u interakciju, što je u redu, jer smo mi to i dopustili. Problem je što se otkriveni podaci mogu iskoristiti u druge svrhe protivno našoj volji, a da mi toga nismo ni svjesni. Želimo da drugi prepoznaju naše pozitivne karakteristike i da pomoću otkrivanja određenih podataka ostvarimo željenu dobit, ali ne želimo da nas drugi iskorištavaju, sprječavaju, kontroliraju, kritiziraju ili saznaju da sudjelujemo u nekim nepoželjnim aktivnostima. Ponekad moramo pristajati na stvari na koje inače ne bi pristali kako bi ostvarili određeni cilj.

Saadi Lahlou (2008) smatra kako je privatnost kompleksan, apstraktan i teško prevodiv termin te mu nije namjera definirati ga, nego predlaže pronalazak okvira koji bi služio dizajnerima sistema kao smjernica pomoću specifičnog pristupa „privatnost kao očuvanje lica“ (eng. *face-keeping*). „Lice“ je socijalni konstrukt koji uključuje prikaz uloge-onoga što bi subjekt trebao raditi, to jest što drugi mogu očekivati od njega, i statusa - onoga kako bi drugi trebali postupati prema subjektu, to jest što subjekt može očekivati od drugih (Lahlou, 2008). Lica omogućuju socijalnu interakciju, te su konstruirana u odnosu na druga lica s kojima su u interakciji. Čovjek ih ima više, kao što su: prijatelj, roditelj, kupac, pravnik, te s obzirom na određenu situaciju odabire prikladno lice. Diplome, osobne iskaznice, kartice, potvrde, certifikati i mnoge druge isprave daju licima legitimnost. Nužno je da se prilikom interakcije na Internetu uzme u obzir ovo svojstvo te da se prema ljudima ponaša u skladu s pravilima pristojnosti i interakcije kakvima se služe u ne-digitalnom okruženju. Povreda privatnosti događa se prilikom „gubljenja lica“ (eng. *losing-face*), to jest kada se osobi onemogućuje da koristi prigodno ili željeno lice, a podržavanje privatnosti odnosi se na izgradnju sustava koji jamči

privatnosti, u smislu da omogućuje očuvanje lica koje korisnik želi očuvati s obzirom na određenu aktivnost (Lahlou, 2008).

Dobar sustav je onaj koji je naklonjen korisnicima, te bi se tom idejom trebali voditi i dizajneri sustava prilikom njegove izrade. Konstruktivni pristup „privatnost kao očuvanje lica“ je pristup usmjeren na određeni cilj koji može postići puno više nego zasad učestaliji defenzivni pristupi koji su bazirani na izbjegavanju određenih događaja. Cilj i glavna smjernica pristupa je „oštrica privatnosti“ (eng. *privacy-razor*) naputak koji nalaže kako svaki sustav mora omogućiti korisniku da nosi željeno lice, te da sustav ne smije koristi podatke koji pripadaju drugim licima, već samo one koji su izričito nužni, i bez kojih ne bi mogao funkcionirati.

6.2. Sigurnost podataka

„S obzirom na vrijednost podataka, posebno osobnih podataka koji mogu olakšati krađu identiteta ili komercijalnih podataka koji se mogu iskoristiti piratski ili za stjecanje konkurentske prednosti, sigurnost podataka postala je važan aspekt zaštite podataka“ (Kitchin, 2014:174). Korisnički podaci kao što su korisnička imena i lozinke pohranjuju se na različitim mjestima koja su osjetljiva na zlonamjerne programe, krađu i pronevjeru podataka. Gantz i Reinsel (2011) navode pet razina sigurnosti podataka koje zahtijevaju proaktivne sigurnosne procedure: (1) privatnost (eng. *privacy*): zaštita privatnosti informacija i ograničenje njihove cirkulacije; (2) usklađenost (eng. *compliance-led*): zaštita podataka koji bi mogli biti vidljivi u parnici ili predmet pravila zadržavanja; (3) skrbništvo (eng. *custody*): zaštita podataka koji bi mogli dovesti do ili pomoći u krađi identiteta (podaci o računu); (4) povjerljivost (eng. *confidential*): podataka koje osnivač podataka želi zaštititi, kao što su poslovne tajne, popis klijenata, povjerljivi dopisi; (5) zaključavanje (eng. *lockdown*): zaštita podataka koji zahtijevaju najvišu razinu sigurnosti kao što su financijske transakcije, kadrovske datoteke, medicinske dokumentacije ili vojne tajne.

Upravljanje ovim razinama sigurnosti podataka je važan zadatak za pojedince, poduzeća i institucije. Općenito, to se postiže uz pomoć vatrozida (eng. *firewall*),

antivirusnih programa ili šifriranjem podataka koje zahtijeva lozinku za otključavanje. Međutim, iako se konstantno radi na razvoju sigurnosne industrije, metode koje koriste hakeri također postaju sve sofisticiranije. Sve više uređaja proizvodi, dijeli i koristi podatke, što navodi na pretpostavku da će se u budućnosti sigurnosni problem umnožiti, a ne smanjiti, što će dovesti do niza pravnih pitanja vezanih uz odgovornosti i obveze u zaštiti sustava (Kitchin, 2014).

6.3.Profiliranje i diskriminacija

Podaci se odavno koriste na profiliranje i upravljanje populacijama, ali ti procesi postaju sve sofisticiraniji i rašireniji. Za potrebe sigurnosti i otkrivanje prijevara, državna tijela izrađuju profile građana, međutim nagli porast profiliranja događa se u komercijalnom sektoru. Komercijalna poduzeća žele razumjeti na koji način njihovi postojeći i potencijalni klijenti razmišljaju i kako se ponašaju kako bi u želji za povećanjem profita usmjerili i proširili svoje djelovanje. Jon Goss (1995) navodi kako je mali broj specijalističkih tvrtki stvorio generičke klasifikacije stanovništva unutar kojih su kućanstva svrstana u klase profila definirane pomoću različitih varijabli kao što su demografski podaci, položaj i životni stil. Također, sve veći broj GIS (eng. *geographic information system*) oznaka unutar prostornih znanosti (eng. *spatial science*) karakterizira kvantitativno modeliranje ponašanja kojem je svrha predviđanje i učinkovito upravljanje društvenim životom.

Kako bi smanjile financijske troškove i povećale učinkovitost, tvrtke kupuju profile i kontakt podatke. Na taj način izbjegavaju nepotrebno oglašavanje i usredotočuju se odmah na ciljanu populaciju. Siegel (2013) navodi kako su u novije vrijeme tvrtke počele umjesto klasa profila raditi pojedinačne profile kombiniranjem podataka iz različitih izvora, kao što su kreditne kartice, transakcije s kartica vjernosti, postovi s društvenih mreža i drugi osobni podaci.

Pozitivne strane prediktivnog profiliranja su personalizirani tretmani za kupce, a veći profit i smanjenje gubitaka za dobavljače. Iz perspektive tržišta kao mjesta susreta ponude i potražnje ova situacija djeluje kao učinkovito rješenje od kojeg obje strane

ostvaruju dobit. Međutim, ono se može iskoristiti i za društveno grupiranje, na temelju kojeg određene grupe mogu dobiti povlašteni status, dok druge mogu postati žrtve marginalizacije, diskriminacije i isključenja. „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" (Kocijan, 2014:15).

6.4. Tehnološko upravljanje

Živimo u doba u kojem država i korporacije znaju i imaju mogućnost predvidjeti toliko o pojedincima putem raznih sustava nadgledanja i prediktivnog profiliranja što im daje moć nametanja krutih i pogubnih režima koji nalikuju Orwellovoj 1984. i Velikom bratu. U želji za sigurnošću i smanjenjem rizika manje ili više pristajemo na povećanje nadzora. Također, taj nadzor se sve više tehnizira pri čemu je regulacija pojedinih aspekata svakodnevnog života prenesena na tehnološke sustave. Na primjer, snimanje, obrada i rješavanje prometnih prekršaja postaje sve više automatizirana. Uz pomoć softvera koji snima i obrađuje podatke kao što su registracijske oznaka, brzina i pravo na pristup te ih povezuje s vlasničkim bazama podataka, omogućujući tako automatsko izdavanje novčanih kazni (Kitchin, 2014).

Ne koristi se svako automatizirano generiranje podataka za donošenje automatiziranih odluka, ali to je sve više trend, pogotovo u visoko reguliranim sustavima. Može se samo nagađati do kakvih bi to pozitivnih, ali još važnije, negativnih posljedica dovelo. Stoga je važno da pitanje tehnološkog upravljanja, kao i privatnosti, sigurnosti podataka, i prediktivnog profiliranja bude predmet znanstvenih rasprava, javnih debata i učestalih kritičkih promišljanja.

7. Primjena Velikih podataka – studije slučaja

Predviđanje povećanja ili smanjenja cijene avionskih karata i detekcija gripe dva su slučaja primjene Velikih podataka, koja ukazuju na njihovu znanstvenu i društvenu važnost. Također, na prvom primjeru moguće je uočiti način na koji podaci postaju izvorom ekonomske vrijednosti. Cilj ovog poglavlja je konkretizirati temu i dodatno naglasiti relevantnost Velikih podataka za društvo i pojedinca.

7.1. Kupovina avionskih karata

Veliki podaci čine znatne razlike u mnogim sferama ljudskog djelovanja, mijenjaju način na koji razmišljamo o poslu, zdravlju, politici, obrazovanju i inovacijama u godinama koje nadolaze. Prvi primjer osvrnut će se na mogućnost da preoblikuje cijeli poslovni sektor. Dobar primjer za to je kupovina avionskih karata. Mayer-Schönberger i Cukier (2012) u knjizi obrađuju slučaj Orena Etzionia koji je ranije kupio avionske karte za relaciju Seatlle-Los Angeles s pretpostavkom da će tada biti jeftinija, međutim, kada je upitao ostale putnike u avionu, zaključio je da to nije slučaj. Kako je i sam računalni stručnjak, odlučio je pomoći ljudima saznati je li neka cijena karte, koju vide online, dobra kupovina ili ne. Svako sjedalo je nerazlučivo od ostalih na istome letu, a ipak cijene se uvelike razlikuju zahvaljujući mnoštvu faktora koji su uglavnom poznati jedino zrakoplovnim tvrtkama. Međutim, sam razlog zašto su cijene različite nije ni važan, ključno je jedino predvidjeti hoće li se prikazana cijena u budućnosti povećati ili smanjiti. Što i nije veliki problem – sve što je potrebno jest analizirati sve prodaje karata za određeni put i ispitati cijene plaćene u odnosu na broj dana prije polaska. Ako postoji tendencija da će se prosječna cijena karte smanjiti, ima smisla pričekati i kupiti kartu kasnije, a ako će se prosječna cijena povećati, sistem će preporučiti kupovinu karte što prije. Model koji je Etzioni razvio nema razumijevanja zašto, nego jedino-što. Model ne zna niti jednu od varijabli koje ulaze u odluku o cijeni, kao što je broj neprodanih sjedala,

sezonu, popularnost određenog datuma i slično. Bazirao je svoje predviđanje na onome što je znao: vjerojatnost sakupljenu iz baza podataka o drugim letovima. Kasnije se ovaj mali projekt razvio u startup tvrtku Farecast koja može poslužiti kao sažeti prikaz načina na koji djeluje tvrtka koja se bavi Velikim podacima. Na ovom primjeru jasno vidimo da iako su ključne promjene bile u dostupnosti tehnologije (jeftina računalna snaga i prostor za pohranu), ipak je važnija bila spoznaja o tome kako se podaci mogu iskoristiti. „Podatke se više nije promatralo kao nešto statično čija je korist gotova kada je svrha za koju su skupljeni postignuta. Nasuprot, podaci su postali sirovi materijal poslovanja, dinamični ekonomski input koji se koristi za stvaranje novih oblika ekonomske vrijednosti“ (Mayer-Schönberger; Cukier, 2012:5). Stoga Velike podatke možemo opisati i kao „sposobnost društva da poveže informacije na nove načine kako bi proizvele korisne uvide ili dobra i usluge od značajne vrijednosti“ (Mayer-Schönberger; Cukier, 2012:2).

7.2.Detekcija virusa gripe

Još jedan zanimljiv primjer primjene Velikih podataka u rješavanju aktualnih problema mogli smo primijetiti 2009. godine kada se otkrio novi virus gripe H1N1. Virus se brzo širio i jedina nada javnozdravstvenih institucija bila je da se širenje uspori, međutim, da bi se to postiglo bilo je potrebno saznati gdje se virus već pojavio. Nekoliko tjedana prije nego što je H1N1 bio na svim naslovnicama, inženjeri u Google-u objasnili su kako mogu predvidjeti širenje zimske gripe u SAD-u , i to ne samo nacionalno nego i na razini posebnih regija i država. S obzirom da Google prima više od 3 bilijuna upita dnevno te ih pohranjuje u svoje baze, ideja je bila da se detektiraju regije zaražene virusom pomoću upita što ih ljudi postavljaju na Internetu. I drugi su pokušali slično, ali nitko nije imao toliko podataka, procesorske moći i statističkog znanja kao Google. Isti problem pokušali su riješiti i u Centru za kontrolu i prevenciju bolesti (*CDC - Center for Disease Control and Prevention*), ali su rješenja dolazila dva tjedna prekasno. Google je uspoređujući upite tipa "lijekovi za kašalj i temperaturu" na Internetu i realne slučajeve zabilježene u CDC-u razvio matematički model koji je postigao snažnu korelaciju između

predviđanja i službenih podataka diljem zemlje, a sve to u gotovo realnom vremenu. Kada se pojavi sljedeća pandemija, svijet će na raspolaganju imati bolji alat za predviđanje, a samim tim i za prevenciju njezina širenja.

8. Mogućnosti i budućnost prediktivne analize

"Jedan od temeljnih ciljeva znanosti, vjerojatno najvažniji, jest predviđanje budućih događanja. Mnoga znanstvena istraživanja poduzimaju se s ciljem utvrđivanja varijabli (prediktora) pomoću kojih je moguće predvidjeti neke posljedice u budućnosti" (Mejovšek, 2003:27). Prediktivna analiza nastoji postojeće znanje unutar određenog područja iskoristiti za predviđanje budućnosti.

U prošlosti, kada je život bio mnogo jednostavniji ljudi su nastojali predvidjeti što će se dogoditi u budućnosti. Kao primjer može se uzeti astrologija ili poznata proročica Pitija iz starogrčke povijesti. Danas kada je život na svim razinama puno složeniji povećana je i ljudska želja za znanjem o budućnosti. Ljudi današnjice žele znati budućnost na poslovnom i osobnom planu, pokušavaju predvidjeti vremenske prilike i klimatske promjene, pa čak i rezultate sportskih događaja u svrhu zabave ili zarade. Prediktivna analiza postaje sve popularnija zbog mogućnosti analize enormnih količina podataka - što više podataka, to bolji prediktivni model, no još uvijek predviđanje treba poimati kao mogućnost s određenim stupnjem vjerojatnosti, a ne kao potpuno točnu prognozu. Također, ne treba očekivati da će sustavi za potporu odlučivanju temeljeni na Velikim podacima zamijeniti čovjeka kao donositelja krajnje odluke.

Siegel (2013) navodi moguće utjecaje prediktivne analize na različite segmente ljudskog djelovanja. Neki od segmenata su: obitelj i osobni život, marketing i oglašavanje, financijski rizici i osiguranje, zdravstvo, borba protiv kriminala, politika i obrazovanje te upravljanje ljudskim resursima. Pomoću prediktivne analize Facebook može unaprijediti preciznost u sugestiji ljudi koje bi mogli poznavati ili s kojima bi se htjeli povezati, LinkedIn može predložiti potencijalne poslove, a stranice poput Match.com uspoređivanjem interesa mogu „spojiti" ljude koji su međusobno kompatibilni, utječući na taj način na različite sfere osobnog života kao što su prijateljstvo, ljubav ili posao.

Predsjednička kampanja Baracka Obame 2012. godine primjer je korištenja prediktivne analize u političke svrhe. Obamin tim je prikupljao podatke o glasačima suprotnih političkih orijentacija vjerujući da će pomoću razvijenih analitičkih modela i odnosa s javnošću pridobiti njihove glasove i pobijediti na izborima. Kampanja se temeljila na predviđanju da će određeni kontakti kao što su telefonski poziv, kućni posjet, letak ili televizijska reklama utjecati na uvjerenost pojedinih glasača. Glasače za koje se procijenilo da ih je moguće uvjeriti, kontaktiralo se odabranim kanalima. „Drugim riječima umjesto pitanja je li kontaktiranje glasača dobra ideja, ovaj model uvjeravanja odlučuje je li kontaktiranje bolja ideja od nekontaktiranja“ (Siegel, 2014:217).

Alati za prediktivnu analizu prisutni su na tržištu već dugi niz godina, međutim više nisu namijenjeni isključivo matematičarima ili statističarima, nego je vidljiv njihov trend približavanja krajnjim korisnicima (analitičarima) kao što je to slučaj s IBM-ovim SPSS Modelerom. Međutim, SPSS (*Statistical Package for Social Studies*) statistički paket za društvene znanosti razvijen je još krajem šezdesetih godina prošlog stoljeća na američkom sveučilištu Stanford. „Godine 1968., Norman H. Nie, C. Hadlai (Tex) Trup i Dale H. Bent, tri mladića s različitim profesionalnim pozadinama, razvili su softverski sustav temeljen na ideji korištenja statistike za pretvaranje podataka u informacije važne za donošenje odluka“ (SPSS Inc., 2009). Zbog velike potražnje za proizvodom 1975. godine osnovana je tvrtka SPSS Inc., čime je SPSS od neprofitnog akademskog projekta postao komercijalni proizvod. Tvrtka je uspješno pratila razvoj informacijske tehnologije, a odigrala je i vodeću ulogu u razvoju prediktivne analitike početkom 21. stoljeća.

IBM preuzima SPSS 2009. godine i nastavlja njegov daljnji razvoj prema prediktivnoj analitičkoj platformi dizajniranoj za potporu odlučivanju. IBM je SPSS Modeler razvio u cjelovito rješenje za predviđanje budućih događaja, a bavi se cjelokupnim analitičkim procesom, od poslovnog razumijevanja, prikupljanja i pripreme podataka, optimizacije te uporabe modela iz stvarnih poslovnih procesa.

Budući da se danas količine podataka povećavaju velikom brzinom, podatci se pohranjuju na različitim sustavima. IBM SPSS Modeler omogućava pristup i integraciju podataka s različitih izvora podataka kao što su skladišta podataka, podatkovne baze, platforme za pohranu velike količine podataka Hadoop, tekstualne datoteke i slično.

Automatiziranjem dijelova funkcionalnosti koji se brinu o procesiranju velikih količina podataka te predlaganjem korištenja pojedinih ugrađenih matematičkih algoritama IBM SPSS Modeler pogodan je i za iskusne analitičare kao i pojedince koji nemaju veliko prethodno znanje o statistici.

Sve dosadašnje mogućnosti otvaraju put razvoju novih. Ali ipak, taj razvoj je usporen više tehničkim ograničenjima, kao što je povezivanjem pametnog telefona s automobilom, nego novim idejama. Trendovi tehnološkog razvoja vode k ugrađivanju mobilnih uređaja u predmete svakodnevnog korištenja kao što su naočale (Google Glass) ili kućanski uređaji, što će dati dodatni zamah iskorištavanju prediktivnih analiza i ubrzati prikupljanje podataka na temelju kojih će se analize moći odvijati u realnom vremenu. S obzirom da organizacije iz javnog i privatnog sektora sve više uviđaju kako im napredna poslovna inteligencija značajno smanjuje troškove i olakšava donošenje odluka, investicije u tehnologije ovog tipa u budućnosti će se zasigurno povećati, a time i ubrzati razvoj novih. Ključni faktori o kojima će ovisiti daljnji razvoj su veća količina podataka, brža i bolja računala te napredak u znanosti i istraživanjima.

9. Zaključak

Živimo u svijetu s više informacija neko ikada prije, a te informacije se svakim danom intenzivno povećavaju. Promjena u količini dovela je do promjene u stanju, to jest kvantitativna promjena dovela je do kvalitativne. Veliki podaci odnose se na stvari koje mogu dovesti do određenih zaključaka zbog velike količine podataka, a do kojih ne bi moglo doći da je ta skala manja. Uz pomoć tih uvida stvaraju se nove vrijednosti koje mijenjaju tržišta, organizacije, odnose među građanima i vladama, ukratko, mijenjaju način na koji živimo i djelujemo.

Transformacija gotovo svih aspekata ljudske stvarnosti u podatke i informacije novitet je većini ljudi u sadašnjosti. Međutim, svijest o Velikim podacima u budućnosti može samo rasti, a pretpostavka da postoji mjerljiva komponenta u gotovo svemu što činimo te da su dobiveni podaci ogromni izvor znanja značajno će utjecati na našu sliku stvarnosti.

Tehnološki razvoj doveo je do promjene u načinima prikupljanja, obrade i analize podataka, a kako bi mogli pratiti taj razvoj potrebno je da sukladno s njime promijenimo i svoj način razmišljanja o podacima. Ljudi se trebaju odmaknuti od potrage za uzročno-posljedičnim vezama kako bi došli do jednostavnih korelacija, to jest, ne znati „zašto“ nego samo „što“. Ova promjena je svakako teška s obzirom na stoljeća „uhodane“ prakse jer se sukobljava s osnovnim shvaćanjem načina na koji donosimo odluke i spoznajemo stvarnost. Stoga, osim što će nam pomoći da pojmimo svijet na dosad nepoznati način, Veliki podaci nude potencijal za novu epistemologiju i preoblikovanje istraživanja u društvenim znanostima te stvaranje nove znanosti koja se temelji na podacima.

Neupitno je da doba Velikih podataka i upotreba novih tehnologija za njihovu analizu dovodi do pozitivnih rješenja i novih spoznaja u nizu područja ljudskog djelovanja. Međutim, važno je upozoriti i na različite negativne posljedice i opasnosti koje ono donosi. Postoji cijeli niz fundamentalnih pitanja vezanih uz to tko može generirati, pristupiti, dijeliti i analizirati skupove podataka, te u koju svrhu i u kojem

kontekstu. Za očekivati je da ova pitanja u budućnosti budu predmet mnogih debata, a poželjno je da se u nju uključi što više različitih stručnjaka.

Važna karakteristika Velikih podataka krije se u prediktivnoj analizi čije su mogućnosti primjene gotovo neograničene. Veća preciznost u predviđanju budućih događaja smanjuje neizvjesnost kompleksnog svijeta koji nas okružuje, povećava učinkovitost i profitabilnost, ali ono što je najvažnije, ako se iskoristi na pozitivan način, omogućuje čovjeku lagodniji život.

Informacijsko doba u kojem živimo obilježeno je sve većom upotrebom informacijsko-komunikacijskih tehnologija i brzinom kretanja informacija. Ubrzane promjene povećavaju nesigurnost koja se može smanjiti mogućnostima Velikih podataka no treba napomenuti kako ovu revoluciju ne mogu proizvesti podaci sami po sebi nego ljudi koji ih interpretiraju. Isto tako, unatoč brojnim etičkim problemima, kao što su povreda privatnosti ili diskriminacija na temelju profiliranja, treba naglasiti kako su to samo potencijalni problemi jer podaci sami po sebi nisu niti loši niti dobri nego ono što ljudi s njima čine može biti loše ili dobro.

10. Literatura

1. Anderson, C. (2008). „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete“. *Wired*, 23 June. URL: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/ (10-9-2015).
2. Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. (2007). „DBpedia: A Nucleus for a Web of Open Dana“. *The Semantic Web*. Lecture Notes in Computer Science 4825, 722-735.
3. Berman, J. J. (2013). *Principles of big data: preparing, sharing, and analyzing complex information*. Amsterdam: Elsevier; Morgan Kaufman.
4. Brooks, D. (2013). „The Philosophy of Data“. *New York Times*, 4 February. URL: http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html (10-9-2015).
5. Duhigg, C. (2012). *The Power of Habit: Why We Do What We Do in Life and Business*. London: William Heinemann.
6. Europska komisija (2014). *Komisija poziva vlade da iskoriste potencijal „velikih podataka”.* Bruxelles, 2. srpnja 2014. URL: https://www.google.hr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CCcQFjACahUKEwj_iOTekfnHAhXGkA0KHU33BBA&url=http%3A%2F%2Feuropa.eu%2Frapid%2Fpress-release_IP-14-769_hr.pdf&usg=AFQjCNFfYM_9mTutkgj1I4oVqd0YIz6JoQ&sig2=Sj07NjmOGjGvy6A6Hfi8aQ (15-9-2015).
7. Gantz, J.; Reinsel, D. (2011). „Extracting Value from Chaos“. *IDC iView*. Sponsored by EMC. URL: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (13-9-2015).
8. Goss, J. (1995). „“We know who you are and we know where you live“: the instrumental rationality of geodemographics systems“. *Economic Geography*, 171-185. URL:http://www.utexas.edu/depts/grg/hudson/grg394k/readings/gossreading.pdf

(15-9-2015).

9. Kitchin, R. (2014). *The Data Revolution: Big Data, open Data, Data Infrastructures & Their consequences*. London: SAGE Publications Ltd.
10. Kocijan, K. (2014). „Big Data: kako smo došli do Velikih podataka i kamo nas oni vode“. U: *Komunikacijski obrasci i informacijska znanost*. Zagreb: Zavod za informacijske studije, 37-62.
11. Lahlou, S. (2008). “Identity, social status, privacy and face-keeping in digital society”. *Social Science Information*, 47 (3), 299–330.
12. Levitt, S. D.; Dubner, S. J. (2009). Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. New York: William Morrow Paperbacks.
13. 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, URL: https://cdda.cs.stonybrook.edu/sites/default/files/cddafiles/CDDA_University%20Consortium.pdf (10-9-2015).
14. Mayer-Schönberger, V.; Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.
15. Mejovšek, M. (2003). *Uvod u metode znanstvenog istraživanja u društvenim i humanističkim znanostima*. Jastrebarsko: Naklada Slap; Zagreb: Edukacijsko-rehabilitacijski fakultet.
16. Milas, G. (2009). *Istraživačke metode u psihologiji i drugim društvenim znanostima*. Jastrebarsko: Naklada Slap.
17. Ravlić, S. (ur.) (2013-2015). *Hrvatska enciklopedija.* Leksikografski zavod Miroslav Krleža. URL: http://www.enciklopedija.hr/Natuknica.aspx?ID=18148 (14-9-2013).
18. Schneider, R. D. (2012). *Hadoop For Dummies.* Mississauga*:* John Wiley & Sons Canada, Ltd. URL: http://media.wiley.com/assets/7067/61/9781118387368_custom.pdf (17-9-2015).
19. Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, New Jersey: Wiley & Sons, Inc.

20. SPSS Inc. (2009). Corporate history. About SPSS Inc. URL: http://www.spss.com.hk/corpinfo/history.htm (15-9-2015).
21. Tuđman, M. (1990). *Obavijest i znanje*. Zagreb: Zavod za informacijske studije. 161-231.
22. Zhang, L. (2013). *How structured data (Linked Data) help in Big Data Analysis - Expand Patent Data with Linked Data Cloud*. Electrical Engineering and Computer Sciences University of California at Berkeley. URL: http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-96.pdf (17-9-2015).