Znanja in sposobnosti podatkovnih znanstvenikov: pregled in analiza stanja v Sloveniji

[1]Mateja Grobelnik, [2]Jurij Jaklič
[1]Petrol, d. d., Dunajska cesta 50, 1000 Ljubljana
[2]Univerza v Ljubljani, Ekonomska fakulteta, Kardeljeva ploščad 17, 1000 Ljubljana
mateja.grobelnik@gmail.com; jurij.jaklic@ef-uni-lj.si

Izvleček

Masovne podatke in znanost o podatkih so organizacije prepoznale kot vira novih konkurenčnih prednosti. Z namenom izkoriščanja tega potenciala se je povečalo povpraševanje po posameznikih s specifičnimi znanji in sposobnostmi, ki so sposobni iz množice raznolikih podatkov pridobiti koristne informacije in jih na razumljiv način implementirati v obstoječe procese in aktivnosti v organizaciji, po t. i. podatkovnih znanstvenikih. Z neprestanim razvojem področja znanosti o podatkih in ob raznolikih potrebah po analitičnih znanjih prihaja do različnega razumevanja vloge podatkovnih znanstvenikov, hkrati pa se veča vrzel med ponudbo in povpraševanjem po takšnih posameznikih. Namen prispevka je zato prispevati k boljšemu razumevanju ter opredelitvi znanja in sposobnosti podatkovnih znanstvenikov, s pomočjo raziskave v Sloveniji pa ugotoviti trenutna znanja in sposobnosti ter identificirati segmente podatkovnih znanstvenikov v Sloveniji. Na podlagi razvrščanja v skupine samoocene znanj je bilo identificiranih in opisanih pet skupin: trženjski raziskovalci analitiki, podatkovni analitiki, raziskovalci, programerji in podatkovni znanstveniki, skupaj s priporočili za njihov nadaljnji razvoj.

Ključne besede: masovni podatki, znanost o podatkih, podatkovni znanstvenik, sposobnosti, znanja, razvrščanje v skupine, empirična raziskava.

Abstract

Knowledge and Skills of Data Scientists: Overview and Analysis of Current Situation in Slovenia

Big data and data science have been recognized by organizations as sources of a new competitive advantage. In order to exploit theis potential, there has been an increase in demand for individuals with specific knowledge and skills wht are capable of obtaining useful information from a set of diverse data and implement it into existing processes and activities in an organization, that ie so-called data scientists. Due to the continuous development of the data science field and the diverse needs for analytical knowledge, the understanding of the role of data scientists deviates greatly, while the gap between supply and demand for such individuals is also increasing. Therefore, the objective of this paper is to contribute to a better understanding and definition of the knowledge and skills of data scientists, and to identify the current knowledge and skills together with the segments of data scientists through empirical research in Slovenia. Five segments were identified and described: "Marketing researchers – analysts", "Data analysts", "Researchers", "Developers« and "Data Scientists", together with recommendations for their future development.

Keywords: big data, data science, data scientist, skills, knowledge, clustering, empirical research.

1 UVOD

Stroškovno učinkovito shranjevanje podatkov, konvergenca pametnih naprav, družbenih omrežij, širokopasovnih komunikacij in analitike so na novo definirali odnose med proizvajalci, distributerji ter potrošniki izdelkov in storitev, hkrati pa ustvarili nove izzive in priložnosti. Olofson in Vesset (2012) to konvergenco imenujeta inteligentna ekonomija. Sama zmožnost shranjevanja in dostop do podatkov namreč nista dovolj, šele ko imamo možnost podatke analizirati in na podlagi rezultatov sprejemati boljše odločitve, ustvarjamo konkurenčno prednost (Olofson in Vesset, 2012). Povečali smo si možnosti zajema veliko večje količine podatkov različnih tipov, ki nastajajo z veliko hitrostjo.

Ti podatki prihajajo iz različnih (ne)zaupanja vrednih virov, ki lahko organizaciji prinesejo dodano vrednost. Navedeni koncepti se povezujejo s pojmoma masovni podatki (angl. big data) in znanost o podatkih (angl. data science). Če so se na eni strani povečale možnosti zbiranja in shranjevanja podatkov, so se po drugi strani razvile tudi nove tehnologije na področju strojne in programske opreme za analizo in obdelavo večje količine podatkov.

Skupaj z razvojem tega področja se je pojavila potreba po specifičnih znanjih, s pomočjo katerih je mogoče iz velike količine različnih podatkov pridobiti uporabne informacije za izboljšanje poslovnih odločitev. Kot enega glavnih izzivov pri uvajanju strategije masovnih podatkov in znanosti o podatkih navajajo pomanjkanje posameznikov s specifičnimi znanji in sposobnostmi ustvarjanja dodane vrednosti s pomočjo manipulacije s podatki (Big Data Executive Survey, 2012). Na podlagi te potrebe po novih znanjih s področja analitike masovnih podatkov se povečuje povpraševanje po podatkovnih znanstvenikih (angl. data scientists).

Raziskava MGI in McKinsey's Business Technology Office (Manyika idr., 2011) napoveduje, da bo do leta 2018 samo v ZDA nastala od 50- do 60-odstotna vrzel med ponudbo in povpraševanjem po poglobljenem analitičnem znanju, torej pomanjkanje od 140.000 do 190.000 posameznikov z naprednimi znanji na področju statistike ali strojnega učenja in 1,5 milijona menedžerjev in analitikov s sposobnostmi uporabe analitike masovnih podatkov za sprejemanje učinkovitih odločitev. Čeprav so analizo opravili v ZDA, menijo, da bo pomanjkanje poglobljenega analitičnega talenta svetovni trend. Države z višjim številom posameznikov s poglobljenim analitičnim znanjem na prebivalca bodo v prihodnosti privlačen vir teh sposobnosti za druga geografska področja prek migracije ali prek zaposlitev v organizacijah. Po Gartnerju so napovedali, da bo do leta 2015 4,4 milijona delovnih mest na področju informatike po svetu namenjenih podpori delu z masovnimi podatki (Chordas, 2014, str. 23). Povpraševanje zajema vse od inženirjev masovnih podatkov (angl. big data engineer), podatkovnih analitikov (angl. data analyst) do poslovnih analitikov (angl. business analyst). Največji izziv pa naj bi bilo najti podatkovne znanstvenike, saj gre za posameznike, ki imajo ključno vlogo pri uporabi masovnih podatkov: podatke priskrbijo in jih uporabijo za poslovne odločitve (Chordas, 2014, str. 23). V sklopu raziskave New Vantage Partners Big Data Executive Survey (2012, str. 8) so ugotovili, da ima kar 70 odstotkov vprašanih odločevalcev v organizacijah na področju masovnih podatkov namen zaposliti podatkovne znanstvenike, vendar jih 80 odstotkov meni, da je to zanje velik izziv.

Namen prispevka je zato prispevati k razumevanju ter opredelitvi znanja in sposobnosti podatkovnih znanstvenikov, s pomočjo raziskave med podatkovnimi znanstveniki v Sloveniji pa ugotoviti trenutna znanja in sposobnosti ter z uporabo razvrščanja v skupine identificirati segmente podatkovnih znanstvenikov v Sloveniji. Z opredelitvijo masovnih podatkov, prek procesa izvajanja znanosti o podatkih ter s pregledom literature in raziskav so bile identificirane različne sposobnosti in področja znanj podatkovnih znanstvenikov. 92 posameznikov v Sloveniji, ki se večino časa ukvarjajo s podatki, je izpolnilo strukturiran spletni vprašalnik, ki je vključeval vprašanja, povezana z dimenzijami masovnih podatkov, samoocenо znanj, pomembnostjo znanj in sposobnosti, njihovimi dosedanjimi izkušnjami ter načini izobraževanja. Z uporabo razvrščanja v skupine so bili identificirani segmenti podatkovnih znanstvenikov.

V drugem razdelku je predstavljen koncept masovnih podatkov in znanosti o podatkih. V nadaljevanju so opredeljena znanja in sposobnosti podatkovnih znanstvenikov, ki so bili uporabljeni pri oblikovanju vprašalnika. Tretji razdelek vključuje metodologijo empirične raziskave o znanjih in sposobnostih podatkovnih znanstvenikov v Sloveniji, značilnosti vzorca ter rezultate raziskave. Na koncu so podane sklepne ugotovitve.

2 MASOVNI PODATKI IN ZNANOST O PODATKIH

2.1 Opredelitev masovnih podatkov

Posamezniki in organizacije z vsakodnevnim delovanjem danes ustvarimo več podatkov kot kadar koli do sedaj. Podatki nastajajo povsod: na družbenih medijih (angl. social media), kot so Twitter, Facebook, LinkedIn, Instagram itd., spletnih straneh, ob izvedbi nakupnih transakcij, ob aktivaciji GPS signalov mobilnih telefonov, z uporabo RFID značk, na mobilnih aplikacijah, in prav vse te podatke je mogoče shraniti v digitalni obliki. Masovni podatki danes veljajo za popularen trend, ki se v bistvu nanaša predvsem na problem volumna/hitrosti/raznolikosti podatkov (angl. Volume/Velocety/Variety problem). Glavna

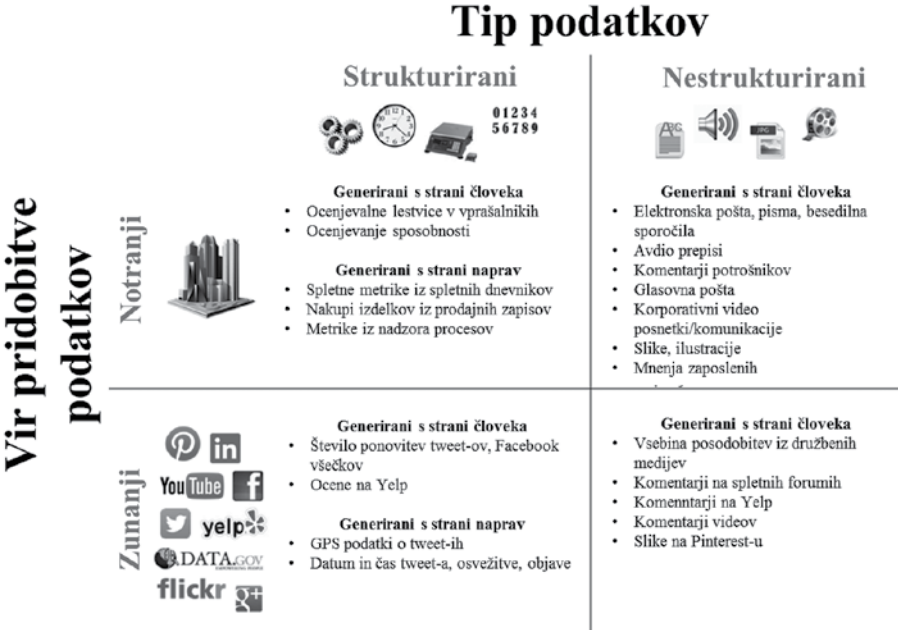
prednost masovnih podatkov je, da lahko s pomočjo analize le-teh pridobimo zanimive vzorce in informacije, ki so bili poprej skriti, saj jih zaradi velike količine dela in časa ni bilo mogoče pridobiti. Sedaj pa jih lahko uporabimo za analizo, sprejemanje odločitev ter razvoj novih produktov, kar pomeni znatno konkurenčno prednost (Lorica, Howard in Dumbill, 2012).

Zaradi dejstva, da pojem masovni podatki in vse, kar dojemamo pod strategijo vpeljave masovnih podatkov, vpliva in zajema širok nabor poslovnih procesov, tehnologij in strokovnih znanj, lahko na izraz masovni podatki gledamo v ožjem in širšem smislu. Če na izraz gledamo v ožjem smislu, gre predvsem za opredelitev tega, kakšne značilnosti morajo imeti podatki, da jih lahko opredelimo kot masovne. Večina definicij masovnosti podatkov ne povezuje le s količino podatkov (volumen), temveč so enako pomembne tudi druge dimenzije podatkov: hitrost, s katero nastajajo podatki, raznolikost virov/strukture podatkov ter vrednost podatkov (najpogosteje omenjene).

Volumen kot dimenzija masovnih podatkov se nanaša na velike količine podatkov, ki se dejansko tudi shranijo, saj so se stroški shranjevanja podatkov občutno pocenili (npr. danes lahko shranimo celotno svetovno zalogo glasbe na napravo v vrednosti 500 dolarjev) (Dhar, 2013, str. 67). Glavna prednost je, da lahko s pomočjo večjih vzorcev, ki so boljši približek populaciji, gradimo bolj natančne napovedne modele. Večanje nabora podatkov pomeni izziv obstoječim tradicionalnim strukturam informacijskih tehnologij, saj masovni podatki zahtevajo razširljivo skladiščenje in porazdeljen pristop k poizvedovanju. Rešitve so na voljo v obliki podatkovnih skladišč ali rešitev, ki izhajajo iz Apache Hadoop (Lorica, Howard in Dumbill, 2012).

Hitrost se nanaša na vedno večjo stopnjo hitrosti, s katero pridobivamo podatke. Pomembnost dimenzije hitrosti masovnih podatkov leži v hitrosti povratne zanke. To pomeni, da je treba delovati in ukrepati na podlagi podatkov v realnem času. Bolj tesna je zanka, večja je konkurenčna prednost (Lorica, Howard in Dumbill, 2012). Tehnologija dimenzije hitrosti masovnih podatkov sega od paketne obdelave (angl. batch processing) ob določenih intervalih do konstantnega toka podatkov (angl. streaming data) v realnem času (Olofson in Vesset, 2012, str. 5).

Raznolikost podatkov lahko opredelimo z vidika več različnih dimenzij. Podatki so lahko opredeljeni z vidika podatkovnega tipa (strukturirani, delno strukturirani in nestrukturirani ipd.), vira pridobitve (notranji, zunanji) ter izvora (generirajo jih naprave ali človek) (Hayes, 2014a). Bistvo tretje dimenzije masovnih podatkov, tj. raznolikosti podatkov, je v tem, da lahko kljub različnim virom, različnemu izvoru in nestrukturiranosti podatkov iz njih izluščimo



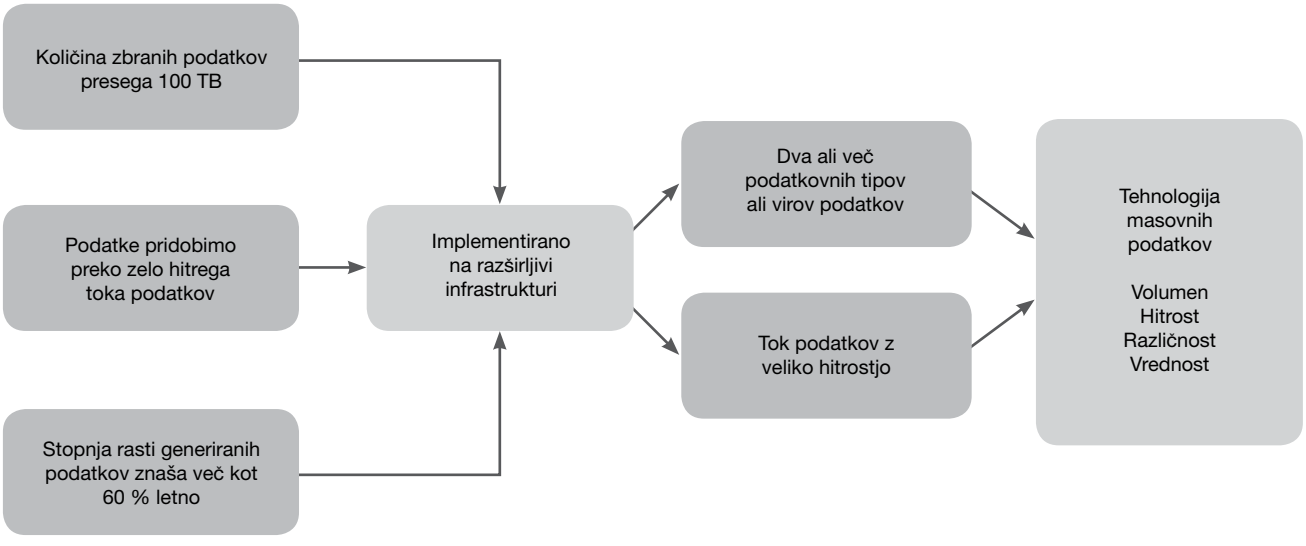
Slika 1: **Ogrodje za opredelitev raznolikost podatkov**
Vir: B. E. Hayes, The what and where of big data: A data definition framework, 2014a.

urejeno znanje, ki ga lahko uporabi človek ali pa ga uporabimo kot vhodne podatke v aplikacijo (Lorica, Howard in Dumbill, 2012). Slika 1 prikazuje ogrodje za opredelitev raznolikosti podatkov po Hayesu (2014a). Model razločuje tip podatkov od vira pridobitve podatkov. Stolpca predstavljata tip podatkov (strukturirani ali nestrukturirani), vrstice pa vir pridobitve podatkov (zunanji ali notranji). Podatki so tako lahko razporejeni v enega od štirih kvadrantov, pri čemer je nadaljnja razdelitev narejena še na ravni izvora podatkov (ki jih generirajo naprave ali človek).

Vrednost podatkov se nanaša na stroške tehnologije in na vrednost, ki jo lahko pridobimo iz masovnih podatkov. Stroški so pomembni predvsem zato, ker pomenijo ključni faktor novosti v masovnih podatkih. Kombinacija razpoložljive programske opreme in nižanje cen strojne opreme je povzročila, da so tehnologije lažje dostopne. Vrednost masovnih podatkov lahko opredelimo z nižjimi stroški kapitala (programske in strojne opreme ter infrastrukture), operativno učinkovitostjo (nižji stroški dela zaradi uporabe bolj učinkovitih metod za integracijo, menedžment, analizo in dostavo podatkov) in izboljšanjem poslovnih procesov (povečanje prihodkov ali dobička na račun novih ali boljših načinov poslovanja, vključujoč izboljšave v komercialnih transakcijah, trajnem menedžmentu skupnosti in primerni distribuciji socialnih, zdravstvenih in izobraževalnih storitev) (Olofson in Vesset, 2012, str. 6).

Kako kvantificirano opredeliti volumen, hitrost in raznolikost, povzema klasifikacija po IDC (Vesset idr., 2012, str. 2), prikazana v sliki 2. V sklop trga masovnih podatkov se vključuje podatkovne baze (ne glede na to, ali gre za relacijske ali ne), ki presegajo 100 TB, ki so implementirane na razširljivi arhitekturi in ki vsebujejo podatke iz dveh ali več virov/podatkovnih tipov ali ko je stopnja hitrosti zbiranja podatkov zelo visoka. Podobno lahko za masovne opredelimo podatkovne baze, katerih količina je sicer manjša kot 100 TB, vendar je letna stopnja rasti generiranih podatkov 60-odstotna ali več, poleg tega so implementirane na razširljivi infrastrukturi in vključujejo podatke dveh ali več tipov/virov, ali ko je stopnja hitrosti zbiranja podatkov zelo visoka (Vesset idr., 2012, str. 2). Na podlagi opisanih dimenzij lahko ugotovimo, da lahko podatke opredelimo kot masovne v ožjem smislu takrat, ko ustrezajo vsaj dvema od treh dimenzij masovnih podatkov (volumen, hitrost, raznolikost), vendar vedno z namenom prinašanja vrednosti (četrta dimenzija) organizaciji v obliki nižjih stroškov, večje učinkovitosti ali izboljšanja poslovnih procesov.

Iz opisanih značilnosti masovnih podatkov lahko ugotovimo tudi, da ni dovolj, da jih le opredelimo. Treba jih je shraniti, imeti možnost dostopa do njih ter znanja, sposobnosti in orodja, da jih lahko



Slika 2: **Kriteriji za opredelitev masovnih podatkov po IDC**
Vir: D. Vesset idr., Wordlwide Big Data Technology and Services 2012-2016 Forecast., 2012, str. 2, slika 1.

ustrezno obdelamo in rezultate obdelav uporabimo za sprejemanje boljših odločitev. Zaradi omenjene večdimenzionalnosti podatkov in novega pristopa k obdelavi je očitno, da masovni podatki potrebujejo nov pristop, novo strojno in programsko opremo ter druge spremembe, povezane z organizacijo, kulturo in sprejemanjem odločitev. Masovni podatki tako v širšem smislu po IDC (Vesset idr., 2012, str. 1) »predstavljajo novo generacijo tehnologij in arhitekturnih rešitev, katerih namen je pridobiti ekonomsko vrednost iz velike količine različnih tipov podatkov s pomočjo visoko intenzivnega shranjevanja, raziskovanja in analize teh podatkov«. Tudi Boyd in Crawford (2012) v članku Critical questions for Big Data, objavljenem v Information, Communications and Society Journal, definirata masovne podatke kot kulturni, tehnološki in znanstveni fenomen, ki temelji na prepletanju tehnologije (maksimiziranje računalniške moči in natančnosti algoritmov za zbiranje, povezovanje in primerjavo velikega nabora podatkov), analize (zmožnost iz velikega nabora podatkov identificirati vzorce z namenom, da pridobimo ekonomsko, tehnično, družbeno ali pravno prednost/odločitev) in mitologije (splošno razširjenega spoznanja, da veliki nabori podatkov zagotavljajo višjo stopnjo inteligence in znanja, ki lahko ustvarijo vpoglede, ki so bili prej nemogoči, v duhu resnice, natančnosti in objektivnosti).

2.2 Znanost o podatkih

Ravno v širšem smislu razumevanja masovnih podatkov so ti tesno povezani s področjem znanosti o podatkih. Področje znanosti o podatkih predstavlja rešitev, kako odkriti potencialne vpoglede, ki se skrivajo v masovnih podatkih, in kako premostiti izziv volumna/hitrosti/raznolikosti/vrednosti masovnih podatkov (Voulgaris, 2014, str. 15). Masovni podatki predstavljajo gonilo sprememb na področjih zbiranja, shranjevanja, menedžmenta, analiziranja in vizualizacije podatkov. Vendar pa masovni podatki potrebujejo znanost o podatkih z namenom, da (Somohano, 2013):

- na podlagi podatkov odkrijemo, česar ne vemo,
- pridobimo napovedni vpogled v podatke, na podlagi katerega lahko sprejemamo boljše odločitve,
- ustvarimo nove izdelke in storitve na podlagi podatkov (angl. data products), ki imajo takojšen vpliv na poslovanje,
- komuniciramo uspešne poslovne zgodbe na podlagi podatkov,
- gradimo zaupanje v sprejemanje odločitev, ki prinašajo poslovno vrednost.

Definicije znanosti o podatkih danes večinoma govorijo o interdisciplinarnem področju – kombinaciji znanj in sposobnosti z različnih področij za obdelavo (masovnih) podatkov. Stanton (2013, str. ii) opredeli znanost o podatkih kot nastajajoče področje delovanja, ki se ukvarja z zbiranjem, pripravo, analizo, vizualizacijo, menedžmentom in ohranitvijo velikega nabora informacij. Čeprav znanost o podatkih najbolj tesno povezujemo s področjem baz podatkov in informatiko, je potrebnih še veliko drugih znanj in sposobnosti. O'Reilly (Lorica, Howard in Dumbill, 2012) definira znanost o podatkih kot disciplino, ki kombinira znanja iz matematike, programiranja in znanosti. Raziskava Big data executive survey (2012, str. 8) pa je znanost o podatkih opredelila kot disciplino, ki združuje uporabo različnih stopenj statistike, podatkovne vizualizacije, računalniškega programiranja, podatkovnega rudarjenja, strojnega učenja in arhitekture podatkovnih baz z namenom reševanja kompleksnih podatkovnih problemov.

3 ZNANJA IN SPOSOBNOSTI PODATKOVNIH ZNANSTVENIKOV

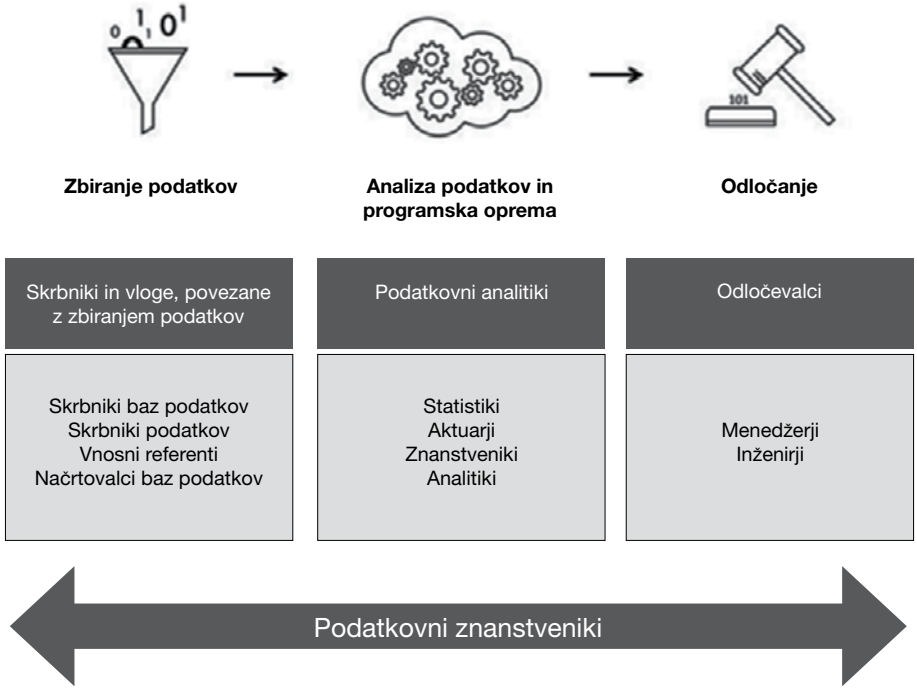
3.1 Podatkovni znanstveniki

Vedno večja vloga informacijsko-komunikacijskih tehnologij in konvergenca različnih znanstvenih disciplin, kot so matematika in statistika ter tudi naravoslovne in družbene vede z informatiko, pomenita po Organizaciji za ekonomsko sodelovanje in razvoj (v nadaljevanju OECD) (2015, str. 261) pomemben trend v poklicih, povezanih s podatki. Ta konvergenca je omogočila tudi pojav novega razreda podatkovnega strokovnjaka – podatkovnega znanstvenika –, katerega naziv še ni v celoti sprejet, vendar ga različni avtorji uporabljajo za opis »nove« discipline, kategorije dela oziroma karierne poti, katere pomembnost raste skupaj z masovnimi podatki (OECD, 2015, str. 261). OECD (2015, str. 254) hkrati opozarja, da trenutno tudi še ne obstaja splošno sprejeta definicija znanj in sposobnosti podatkovnih znanstvenikov. Rivera in Haverson (2014) prav tako omenjata, da trenutno še ni standarda glede uporabe nazivov podatkovni znanstvenik in drugih nazivov, povezanih s podatki (podatkovni analitik, podatkovni rudar, podatkovni inženir, statistik, analitik ipd.), ki bi jasno razločeval med različnimi vlogami. Vzroki za slabo definicijo

tega področja so povezani s tem, da gre za relativno nova področja, ki v literaturi niso še dobila dovolj pozornosti v primerjavi z drugimi informacijsko-komunikacijskimi znanji in sposobnostmi, ter s tem, da se področje še vedno razvija (OECD, 2015, str. 254). Thomas H. Davenport in D. J. Patil sta v članku Harvard Business Review: Data Scientist: The Sexiest Job of the 21st Century definirala podatkovne znanstvenike kot posameznike, ki uporabljajo tako podatke kot znanost, da ustvarijo nekaj novega (Davenport in Patil, 2012). Podobno definicijo je podal Voulgaris (2014, str. 18): podatkovni znanstveniki so posamezniki, ki poiščejo smisel v masovnih podatkih. S pomočjo uporabe visoko naprednih tehnologij, znanj in sposobnosti izpeljejo uporabne informacije iz masovnih podatkov, po navadi v obliki novega izdelka ali storitve na podlagi podatkov (angl. data product).

Znanja in sposobnosti podatkovnih znanstvenikov med drugim izhajajo iz osnovne smeri izobrazbe, delovnih nalog, značilnosti in odgovornosti delovnega mesta, na katerem dela posameznik, ter drugih dejavnikov. Zavedati se je treba, da posamezniki, ki se v večji meri ukvarjajo z znanostjo o podatkih, lahko opravljajo delo analitika, programerja, vodje, menedžerja, profesorja, svetovalca, podjetnika itd. Za ilustracijo obsežnosti področja je Granville (2013) zbral 115 različnih nazivov delovnih mest, neposredno povezanih z masovnimi podatki ali znanostjo o podatkih na podlagi 7500 kontaktov na LinkedInu. Posamezniki z omenjenimi nazivi se lahko v različni meri ukvarjajo z znanostjo o podatkih: od popolnoma operativnih nalog v procesu znanosti o podatkih do vodstvenih položajev, ki zahtevajo dodatna znanja in sposobnosti. Vsem navedenim nazivom strokovnjakov na področju podatkov je skupno to, da delo s podatki zavzema večinski delež njihovega dela (OECD, 2015, str. 255).

Slika 3 prikazuje, kako lahko naziv podatkovni znanstvenik zajema širok spekter nazivov in vlog v procesu znanosti o podatkih prek življenjskega cikla vrednosti podatkov (angl. data value cycle) (OECD, 2015, str. 255). Nazivi vključujejo vloge, ki zbirajo podatke (skrbniki baz podatkov, skrbniki podatkov, vnosni referenti ali načrtovalci baz podatkov), analizirajo podatke s pomočjo analitike (statistiki, aktuarji, znanstveniki, analitiki), ter do določene mere takšni, ki sprejemajo odločitve na podlagi podatkov (menedžerji, inženirji) (OECD, 2015, str. 254).



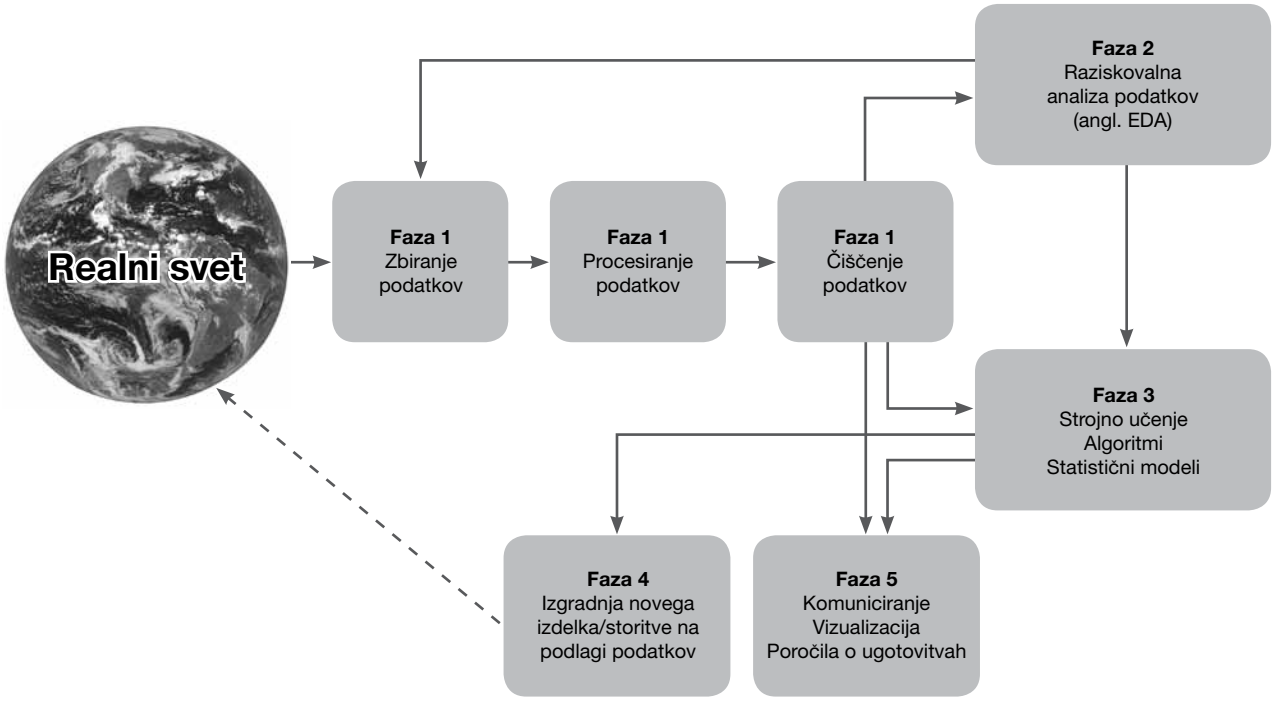
Slika 3: **Faze v življenjskem ciklu vrednosti podatkov v povezavi s ključnimi tipi podatkovnih znanstvenikov**
Vir: OECD, Data-driven innovation: Big Data for Growth and Well-Being, 2015, str. 255, slika 6.5.

Podatkovni znanstvenik je zato v tem prispevku opredeljen kot strokovnjak, ki se večino svojega časa ukvarja s podatki, pri čemer uporablja raznolika znanja in sposobnosti z več različnih znanstvenih področij z namenom, da iz surovih (masovnih) podatkov prek znanosti o podatkih pridobi dodano vrednost. Na podlagi napisanega je ključno, da ima podatkovni znanstvenik ustrezna znanja in sposobnosti, da lahko izpelje vse faze procesa znanosti o podatkih: od pridobivanja podatkov do končne vizualizacije ali novega izdelka/storitve na podlagi podatkov.

Za bolj podroben pregled in razumevanje področja dela, ki ga opravlja podatkovni znanstvenik, je v nadaljevanju predstavljen osnovni okvir za proces izvajanja znanosti o podatkih (prirejeno po O'Neill in Schutt, 2013, str. 36–41; Voulgaris, 2014, str. 133–149; Somohano, 2013; The field guide to data science, 2015, str. 29–33), ki je prikazan na sliki 4.

Kot je razvidno iz prikaza (slika 4), so faze med seboj ciklično prepletene. To pomeni, da si osnovne faze sicer sosledno sledijo, hkrati pa nekatere predstavljajo tudi korak, v katerem podatkovni znanstvenik sprejme odločitev, ali naj ponovi kakšno prejšnjo fazo ali nadaljuje z naslednjo.

Faza priprave podatkov je verjetno najbolj časovno zahtevna in najmanj zanimiva faza v procesu znanosti o podatkih. Vendar gre za zelo pomemben korak, saj ta faza predstavlja temelj za vse nadaljnje korake v procesu. Faza priprave podatkov vključuje zbiranje, procesiranje in čiščenje podatkov. Podatki v osnovi izhajajo iz realnega sveta, v katerem posamezniki in organizacije izvajamo svoje aktivnosti (uporabljamo Facebook, Twitter, opravljamo spletne nakupe, pošiljamo elektronsko pošto, pregledujemo spletne strani, opravljamo telefonske pogovore ipd.). S pomočjo shranjevanja teh aktivnosti v obliki podatkov podatkovni znanstvenik pridobi določen nabor surovih podatkov (transakcije, kliki na spletne povezave in dnevniške datoteke, podatki iz senzorjev, mobilnih telefonov, dokumenti, elektronska pošta, zapisi na družbenih medijih). Te podatke je treba najprej pregledati in prečistiti, da bodo primerni za analizo. Podatkovni znanstvenik lahko pri pripravi podatkov uporabi različne načine procesiranja podatkov: Hadoop Definition File System (v nadaljevanju HDFS) za shranjevanje podatkov za nadaljnje analize, Extract Transform Load (v nadaljevanju ETL) in MapReduce za branje podatkov, redukcijo dimenzij, vzorčenje, združevanje (angl. joining), strganje (angl.



Slika 4: **Proces izvajanja znanosti o podatkih**
Vir: C. O'Neill in R. Schutt, Doing Data Science, 2013, str. 41, slika 2-2.

scraping) ali mešetarjenje (angl. wrangling). Za procesiranje podatkov že obstajajo različna programska orodja, kljub temu pa mora podatkovni znanstvenik podatke pripraviti v .JSON ali v katerem drugem podobnem tipu podatkov. Če so podatki v popolnoma unikatnem tipu, mora podatkovni znanstvenik napisati lastni program za dostop in prestrukturiranje podatkov v obliko, ki bo razumljiva sistemom za branje podatkov. Pri branju velike količine podatkov je priporočljivo, da najprej pripravimo poskusno branje na relativno majhnem naboru podatkov. S tem podatkovni znanstvenik zagotovi, da bo končni nabor podatkov lahko uporaben za načrtovane analize. V sklopu priprave podatkov podatkovni znanstvenik izvede tudi čiščenje podatkov, ki zahteva določeno raven razumevanja podatkov. Pri čiščenju podatkov zapolni manjkajoče vrednosti, preveri relevantnost podatkov in izloči podatke, ki vključujejo napačne ali problematične podatke, izvede normalizacijo podatkov ter preveri neodvisnost podatkov. Čiščenje podatkov vključuje tudi obdelavo osamelcev (angl. outliers). Te lahko iz nabora podatkov odstranimo ali pa prilagodimo model, da ustreza obstoju osamelcev. Odločitev temelji na podlagi različnih faktorjev, kot so število osamelcev, podatkovni tip podatkov in občutljivost modela na njihov obstoj. Za čiščenje in transformacijo podatkov podatkovni znanstvenik uporablja različna orodja ali programske jezike, kot so Python, R skripte, poizvedbeni jezik SQL ali vse naštete. Priporočljivo je, da si podatkovni znanstvenik posamezne korake te faze shrani za primer, če jih bo treba ponoviti ali jih opisati v poročilu. Rezultat faze priprave podatkov je strukturirana oblika podatkov, pripravljena za nadaljnje analize.

Preden se podatkovni znanstvenik loti modeliranja, je potrebna izvedba t. i. raziskovalne analize podatkov (angl. Exploratory Data Analysis – v nadaljevanju EDA). Raziskovalna analiza podatkov je ključni del procesa izvajanja znanosti o podatkih in je primarno namenjena predvsem samemu podatkovnemu znanstveniku. Gre za sistematičen pregled podatkov s prikazom distribucij spremenljivk, transformacijo podatkov, iskanjem potencialnih povezav med spremenljivkami z uporabo razsevnih grafikonov in z generiranjem opisnih statistik za te spremenljivke (srednje vrednosti, mere razpršenosti, identifikacija osamelcev). Pri raziskovalni analizi ne gre le za uporabo orodij, temveč tudi za razumevanje odnosa podatkovnega znanstvenika do podatkov, ki jih analizira. Če želi podatke razumeti, mora pridobiti intuicijo, razumeti oblike in povezati razumevanje procesa, kako so bili podatki pridobljeni, s samimi podatki. Na podlagi rezultatov raziskovalne analize podatkov lahko podatkovni znanstvenik ugotovi, da podatki dejansko niso ustrezni zaradi podvojenih, manjkajočih, neustreznih vrednosti, ali da podatki sploh niso bili zajeti ali pa so bili zajeti napačno. V tem primeru se mora podatkovni znanstvenik ponovno vrniti k viru podatkov in zbrati večjo količino podatkov ali več časa nameniti čiščenju podatkov. To je lahko iterativen proces, dokler podatki niso ustrezni in primerni za modeliranje.

V fazi učenja iz podatkov podatkovni znanstvenik pripravi model, ki predstavlja poskus razumevanja in predstavitve narave realnosti z določenega (matematičnega) vidika. Gre za umetno ustvarjen konstrukt, v katerem so odstranjene vse odvečne podrobnosti. Podatkovni znanstvenik oblikuje model z uporabo različnih orodij s področij statistike in strojnega učenja: opisne statistike in statističnega sklepanja, klasifikacije in segmentacije, regresijske analize in napovedovanja. Izbira modela je odvisna od vsebine problema, ki ga obravnava podatkovni znanstvenik. Lahko gre za klasifikacijski problem, napovedni problem ali osnovni opisni problem.

Podatkovni znanstvenik v zadnji fazi ugotovitve interpretira, pripravi vizualizacije in poročila ter rezultate na ustrezen način predstavi nadrejenim in sodelavcem ali jih objavi v publikaciji. Namen znanosti o podatkih je namreč določiti in razumeti, kaj vse se skriva pod površjem in kakšno uporabno vrednost lahko prinese do končnih uporabnikov. Proces znanosti o podatkih je ponavljajoč se razvojni proces, ki vključuje odkrivanje in učenje na podlagi podatkov. Vizualizacija vključuje grafično predstavo pomena analiziranih podatkov na intuitiven, zanimiv in relevanten način do končnega uporabnika, ki je lahko tudi interaktiven. S pomočjo vizualizacije lahko podatkovni znanstvenik pridobi boljšo predstavo, česa še ne ve, in lahko tako bolje razume omejitve modelov, vrednost podatkov in bolje obvladuje negotovost v podatkih. Cilj analize je alternativno lahko tudi izgradnja prototipa na podlagi analiziranih podatkov (angl. data product). Novi izdelek ali storitev, ki temelji na kombinaciji podatkov in algoritmov, je dodana vrednost organizaciji. Primeri takšnih izdelkov ali storitev na podlagi podatkov so klasifikator nezaželene elektronske pošte, algoritem

za rangiranje spletnih strani v spletnih iskalnikih z relevantnimi rezultati na podlagi spletnega iskanja, sistem za priporočanje (angl. recommendation system), mrežna statistika in grafikoni, ki jih LinkedIn prikazuje svojim uporabnikom, ali geografski informacijski sistem, kot je MapQuest, ki potrošnikom daje uporabne geografske informacije o določeni lokaciji. Tisto, kar razlikuje znanost o podatkih od statistike, je, da se takšen »podatkovni produkt« vgradi nazaj v realni svet, v katerem potrošniki interaktivno uporabljajo produkt, kar posledično generira več podatkov, ki podatkovnemu znanstveniku omogočajo izboljšave tega produkta.

3.2 Znanja in sposobnosti podatkovnih znanstvenikov

3.2.1 Znanja podatkovnih znanstvenikov

Veliko različnih tehnologij in tehnik je bilo razvitih in prilagojenih z namenom združevanja, manipuliranja, analiziranja in vizualizacije masovnih podatkov (Manyika idr., 2011, str. 27). Seznam znanj trenutno ne daje celostnega pogleda na znanja podatkovnih znanstvenikov, saj se metode in orodja neprestano razvijajo z namenom reševanja vedno novih problemov (Manyika idr., 2011, str. 27). Prav tako različni problemi zahtevajo uporabo različnih tehnik in tehnologij z različnih področij glede na naravo problema in cilje projekta. Pri opredelitvi znanj podatkovnih znanstvenikov so bila ta na podlagi literature, raziskav in izkušenj razporejena v skupine po naslednjih področjih: znanstvena metoda, programiranje, menedžment podatkov, baze podatkov, statistika, matematika, strojno učenje ter domenska znanja s pripadajočimi tehnikami. Iz nabora znanj so bile izključene opredelitev znanj, vezane na specifične programske rešitve (SPSS, SAS, Orange, RapidMiner, Weka, Tableau, Excel itd.), saj so te bolj tehnološko orodje, s katerim podatkovni znanstvenik izvede določeno fazo v znanosti o podatkih. Poleg tega bi to lahko pristransko vplivalo na rezultate, saj bi bili ti vezani na popularnost in dostopnost posameznega orodja.

Znanstvena metoda v najširšem smislu vključuje vse postopke in tehnike za objektivno raziskovanje pojavov (Toš in Hafner-Fink, 1998). Hayes (2014b) verjame, da znanstvena metoda predstavlja ključno vlogo v razumevanju katerih koli podatkov, ne glede na njihovo velikost, hitrost ali raznolikost. Podatki namreč ne »govorijo« sami zase, temveč jim pomen dajo ljudje prek ustvarjanja, zbiranja in interpretacije podatkov. Ljudje pa so na žalost tudi vir (namerne ali nenamerne) pristranskosti, ki lahko poslabša kakovost podatkov (Hayes, 2014b). Načrtovanje poskusov/eksperimentov (angl. experimental design) je raziskovalna tehnika, ki se uporablja v vzročnem raziskovanju (angl. causal research) za vzpostavljanje vzročno-posledičnega odnosa med spremenljivkami (Malhotra, 2012, str. 221). Podatkovni znanstvenik naj bi v sklopu načrtovanja poskusov poznal koncepte neodvisnih, odvisnih in zunanjih spremenljivk, testnih enot in naključne razdelitve na poskusno in kontrolno skupino. Pri izvedbi poskusa podatkovni znanstvenik namreč določi testne enote in način, kako so te enote razdeljene na homogene podskupine, določi, katere neodvisne spremenljivke bo spreminjal, manipulira eno ali več neodvisnih spremenljivk in nato opazuje in meri učinke teh sprememb na odvisne spremenljivke, ob tem pa preverja vpliv zunanjih ali tujih spremenljivk (Malhotra, 2012, str. 222–223). Zagotavljanje ponovljivosti raziskav (angl. reproducible resarch) pomeni ključni koncept znanstvene metode. Vključuje koncepte in orodja, ki jih uporablja podatkovni znanstvenik, da lahko znanstvene ugotovitve objavi skupaj s podatki, viri podatkov, programsko kodo ter podrobnimi navodili za izvedbo analize podatkov z namenom, da je raziskavo mogoče ponoviti, bolje razumeti ali preveriti njeno veljavnost (Kuhn, 2015).

Z znanjem programiranja lahko podatkovni znanstvenik pokrije celotni proces izvajanja znanosti o podatkih – kadar koli lahko napiše program, ki pridobi podatke iz baze podatkov, zažene algoritme strojnega učenja na naboru podatkov (Ultimate skills checklist for your first data analyst job, 2015, str. 5), razvije produkt/storitev na podlagi podatkov ali pripravi vizualizacijo podatkov. Priporočljivo je znanje vsaj enega ali več programskih jezikov, ki so robustni, popularni in razširljivi – sploh pri velikem naboru podatkov (Voulgaris, 2014, str. 53). Priporočljivo je tudi, da ima podatkovni znanstvenik dober pregled nad področjem knjižnic in paketov, povezanih s programskimi jeziki, ki se najpogosteje uporabljajo za izvajanje znanosti o podatkih (Ultimate skills checklist for your first data analyst job, 2015, str. 5). Med programske jezike, ki jih najbolj pogosto uporablja več kot 90 odstotkov podatkovnih znanstvenikov, uvrščamo R, SAS in Python (Piatetsky, 2014). Voulgaris (2014, str. 54) omenja tudi Java, C+, C# in Perl,

ki so objektno orientirani jeziki (angl. object-oriented languages), katerih prednost je v tem, da omogočajo enostavno ustvarjanje kompleksne programske kode. Proces znanosti o podatkih je mogoče izvajati tudi v drugih programskih jezikih: Julia, Scala, Stata, Hadoop programski jeziki (Pig Latin, HiveQL idr.), Java, Unix shell/awk/sed, MATLAB, C/C++, Perl, Octave, Ruby, Lisp/Clojure, F# itd. (Piatetsky, 2014). Znanja iz zalednega in čelnega programiranja se izkažejo za koristna predvsem pri implementaciji produkta/storitve na podlagi podatkov v produkcijsko okolje – uporabniško aplikacijo. Podatkovni znanstvenik naj bi predvsem imel osnovna znanja kot podlago za komunikacijo in usklajevanje analitične rešitve z zalednim in čelnim razvijalcem.

Podatkovni znanstvenik ima s svojim delovanjem in izvajanjem procesa znanosti o podatkih pregled, možnost in vpliv na obvladovanje podatkov, arhitekturo, varnost, povezovanje, shranjevanje in kakovost podatkov ter druge vidike t. i. menedžmenta podatkov. Z ustreznimi znanji s področja menedžmenta podatkov lahko podatkovni znanstvenik poveča učinkovitost in uspešnost izvajanja procesa znanosti o podatkih. Menedžment podatkov (angl. data management) vključuje in opisuje procese za načrtovanje, definiranje, kreiranje, pridobivanje, vzdrževanje, uporabo, arhiviranje, nadzor in integracijo podatkov (DAMA, 2014, str. 5). Po definiciji DAMA (2014, str. 10) se menedžment podatkov deli na več področij: obvladovanje podatkov (angl. data governance), menedžment podatkovne arhitekture (angl. data arhitecture management), razvoj in oblikovanje podatkov (angl. data modeling and design), shranjevanje podatkov (angl. data storage and operations), menedžment varnosti podatkov (angl. data security management), integracija in interoperabilnost podatkov (angl. data integration and interoperability), menedžment dokumentov in vsebine (angl. document and content management), menedžment matičnih in referenčnih podatkov (angl. reference and master data management), menedžment podatkovnih skladišč in poslovne inteligence (angl. data warehousing and business intelligence management), menedžment metapodatkov (angl. metadata management) ter menedžment kakovosti podatkov (angl. data quality management). Oblikovanje informacij (vizualizacija) – v sklopu znanj podatkovnega znanstvenika govorimo o podpodročju vizualizacije podatkov, in sicer o področju vizualizacije/oblikovanja informacij (angl. information vizualization/design). Znanja iz oblikovanja informacij lahko podatkovni znanstvenik uporablja v več različnih fazah izvajanja procesa znanosti o podatkih. V sklopu raziskovalne analize podatkov uporablja različne vizualizacijske tehnike z namenom razumevanja podatkov in njihovih zakonitosti ter identifikacije negotovosti v podatkih (npr. gruče primerov, osamelce, trende in relacije med spremenljivkami) (Leban, 2007, str. 2). Informacije oblikuje tudi z namenom razumevanja rešitev analize, domnev in algoritmov podatkovnega rudarjenja (kjer je to mogoče) ter za predstavitev rezultatov procesa znanosti o podatkih ali za kreiranje izdelka/storitve, ki temelji na podatkih. »Glavna prednost uporabe vizualizacije je njena interpretabilnost – odkrite zakonitosti lahko dejansko vidimo, zaradi česar je njihovo razumevanje neprimerno boljše.« (Leban, 2007, str. 2) Podatkovni znanstvenik naj bi zato imel znanja iz celostnega pristopa k oblikovanju informacij, poznavanja posameznih vizualizacijskih metod in njihovih zakonitosti, izbire ustrezne vizualizacijske metode in znanja iz oblikovanja interaktivnosti (razvoj interaktivnih rešitev, produktov in vmesnikov).

Prva faza v procesu znanosti o podatkih je ročno pridobivanje, shranjevanje in čiščenje podatkov v obliko, ki bo primerna za izvoz ali nadaljnje analize (Ultimate skills checklist for your first data analyst job, 2015, str. 12). Ta proces je v sklopu znanosti o podatkih znan kot mešetarjenje podatkov (angl. data wrangling, data munging, data scraping) (Ultimate skills checklist for your first data analyst job, 2015, str. 12). Gre za nalogo, ki podatkovnemu znanstveniku lahko zavzame od 50 do 80 odstotkov njegovega časa (Ultimate skills checklist for your first data analyst job, 2015, str. 12). Zaradi tega je pomembno, da ima podatkovni znanstvenik znanja, kako dostopati do podatkov, jih pridobiti, shraniti ter odpraviti nepopolnosti, za kar potrebuje znanja s področja baz podatkov. Med znanja s področja baz podatkov spadajo poznavanje sistemov baz podatkov, ki temeljijo na strukturiranih ali delno strukturiranih podatkih (centralni repozitorij za shranjevanje podatkov, katerih osnova je SQL), nestrukturiranih podatkih (baze podatkov, katerih osnova je NoSQL), masovnih in distribuiranih podatkih (Hadoop, MapReduce), poizvedbenih programskih jezikov SQL, HiveQL ter osnovna znanja iz sistemske administracije.

Podatkovni znanstvenik naj bi imel osnovno znanje iz statistike ter poznavanje določenih kon-

ceptov in terminologije, ki jo uporabljajo statistiki (Granville, 2014, str. 4): metod vzorčenja, opisne statistike, verjetnostnih porazdelitev, statističnega preizkušanja domnev, redukcije dimenzij, analize časovnih vrst, prostorske statistike ipd. Pri tem je predvsem pomemben vidik znanja in razumevanja, kdaj je določena tehnika primeren ali neprimeren pristop k problemu (Ultimate skills checklist for your first data analyst job, 2015, str. 7).

Podatkovni znanstvenik naj bi bil sposoben prevesti besedne probleme v matematične izraze, reševati enačbe, manipulirati algebrične izraze in imel naj bi osnovno znanje iz teorije matrik (Ultimate skills checklist for your first data analyst job, 2015, str. 9; Granville 2014, str. 4). Prav tako naj bi imel znanja, kako narisati grafikone za različne tipe funkcij z razumevanjem odnosa med grafično funkcijo in njeno enačbo. Priporočljiva so tudi znanja odvodov in integralov, optimizacije in linearne algebre. Ta področja matematike so osnova za razumevanje strojnega učenja in efektivnega manipuliranja podatkov v podatkovnih modelih (Ultimate skills checklist for your first data analyst job, 2015, str. 9).

Strojno učenje je poddomena računalništva (področja umetne inteligence), ki se ukvarja z zasnovo in razvojem algoritmov, ki omogočajo računalnikom razvoj akcij na podlagi empiričnih podatkov (Manyika idr., 2011, str. 29). Poudarek strojnega učenja je na avtomatiziranem učenju in prepoznavi kompleksnih vzorcev z namenom sprejemanja inteligentnih odločitev na podlagi podatkov (Manyika idr., 2011, str. 29). Strojno učenje je torej podatkovno intenzivni razvoj algoritmov (kot podatkovno rudarjenje) s poudarkom na prototipiranju algoritmov za produkcijsko okolje, za obdelavo velikih količin podatkov, na podlagi katerih je mogoče narediti napovedi (angl. predict), klasifikacijo (angl. classify), segmentacijo (angl. cluster) in/ali izračunati predloge za ukrepanje na podlagi obdelanih podatkov (Ultimate skills checklist for your first data analyst job, 2015, str. 10; O'Neill in Schutt, 2013, str. 52). Strojno učenje se ukvarja tudi z razvojem avtomatiziranih sistemov (prepoznavanje slik, govora, algoritmi za generiranje ponudb, angl. bidding algorithms, algoritmi za targetirano oglaševanje, angl. ad targeting algorithms), ki se samodejno osvežijo, neprestano preizkušajo, ponovno učijo in osvežujejo nabore podatkov za učenje, preverjajo veljavnost in izboljšujejo ali odkrivajo nova pravila. Poddomena strojnega učenja, zelo blizu umetni inteligenci (angl. artificial intelligence), je poglobljeno učenje (angl. deep learning) (Granville, 2014). Za podatkovnega znanstvenika ni nujno, da ustvarja popolnoma nove algoritme za strojno učenje, vendar pa mora poznati najbolj pogoste algoritme in tehnike za strojno učenje, od zmanjšanja dimenzij (metoda glavnih komponent) do nadzorovanega (klasifikacija) in nenadzorovanega učenja (razvrščanje v skupine). Ni v celoti potrebno poznavanje teorije in podrobnosti implementacij v ozadju teh algoritmov. Je pa potrebno poznavanje prednosti in slabosti teh algoritmov, kot tudi, kdaj jih je smiselno uporabiti glede na kontekst problema ter kdaj ne (O'Neill in Schutt, 2013, str. 54; Ultimate skills checklist for your first data analyst job, 2015, str. 10).

Znanja poslovnega področja, dejavnosti ali domene, iz katere izhaja problem, so izjemne vrednosti in zelo nenadomestljiva (The field guide to data science, 2015, str. 96). Poslovna oziroma domenska znanja vključujejo poznavanje metod agilnega pristopa, pristopa Waterfall, razvoja izdelkov/storitev, razumevanje delovanja organizacije, poznavanje dejavnosti, poznavanje dobrih praks metodologij podatkovnega rudarjenja (CRISP-DM, SEMMA, DMAIC) ter vsa druga poslovna znanja (finance, trženje, trženjsko raziskovanje, logistika, razvoj izdelka itd.), ki so relevantna za organizacijo ali dejavnost (Voulgaris, 2014, str. 150). Omogočajo poglobljeno razumevanje podatkov in dejavnikov, ki vplivajo na analitični cilj, velikokrat pa so ključni diferenciator uspeha celotne ekipe, ki se ukvarja z znanostjo o podatkih (The field guide to data science, 2015, str. 96). Domenska znanja vplivajo na to, kako podatkovni znanstvenik izbira lastnosti, pripisuje podatke, izbira algoritme, in posredno vplivajo tudi na uspešnost projektov. Posameznik žal ne more biti domenski strokovnjak na vsakem področju. Zato se podatkovni znanstveniki pogosto obračajo na druge analitike, domenske strokovnjake ter druge sekundarne vire z namenom izgradnje razumevanja domenskega področja problema (The field guide to data science, 2015, str. 96).

3.2.2 Sposobnosti podatkovnih znanstvenikov

Znanja z različnih področij so pomembna, vendar niso dovolj. Znanost o podatkih zahteva bolj sistematično razmišljanje ter kombiniranje kreativnega pristopa k definiranju in reševanju problemov skupaj z obvladovanjem časa. Podatkovni znanstvenik je posameznik, ki ga označuje nabor specifičnih

značilnosti, sposobnosti in načina razmišljanja, ne samo nabor znanj (Voulgaris, 2014, str. 37).

Podatkovni zanstvenik naj bi bil predvsem radoveden glede stvari, ki jih opazuje, kot so vzorci in odnosi ter razmerja med različnimi značilnostmi (Voulgaris, 2014, str. 38). Radovednost je ključna, da lahko podatkovni znanstvenik razstavi problem in razišče odnose med podatki, ki na prvi pogled delujejo nepovezani (The field guide to data science, 2015, str. 42). Radovednost dopolnjujejo disciplina, analitične sposobnosti in sposobnosti reševanja problemov. To vključuje vse – od želje po raziskovanju in razčlenitvi problema do zelo jasno definiranega nabora domnev, ki jih je mogoče preveriti (Lorica, Howard in Dumbill, 2012). Podatkovni znanstveniki rešujejo probleme uporabnikov podatkov. Vendar preden lahko rešijo problem, ga je treba ustrezno identificirati, kar pa ni vedno lahko (Stanton, 2013, str. 14). Za pristop k problemu in reševanju problemov sta ključni tudi eksperimentiranje in kreativnost – sposobnost pogleda na problem na različne, kreativne načine, ki v preteklosti niso še bili uporabljeni v takšnem kontekstu (angl. thinking outside the box) (Chordas, 2014, str. 24; Lorica, Howard in Dumbill, 2012; The field guide to data science, 2015, str. 42). Pomembna sposobnost sta tudi fleksibilnost in osredotočenost na cilj, ko je podatkovni znanstvenik sposoben premagati napake, opustiti idejo, ki ne deluje, se iz tega nekaj naučiti in poskusiti z novim pristopom. Znanost o podatkih je namreč serija »slepih ulic«, dokler prava pot ni identificirana. To zahteva unikaten set osebnostnih lastnosti – potrpežljivost in vztrajnost (The field guide to data science, 2015, str. 42).

Da podatkovni znanstvenik razume svojo vlogo in pomen, so pomembne tudi sposobnosti, povezane z njegovo podjetno naravnanostjo, ki vključujejo poslovni čut, prebrisanost (angl. cleverness) in vztrajnost (Granville, 2014, str. 3; Chordas, 2014, str. 23; Lorica, Howard in Dumbill, 2012). Ključna sta tudi sposobnost sprejemanja odločitev in pogum – sposobnost podatkovnega znanstvenika, da izrazi svoje mnenje, poišče rešitev ter prepriča in motivira menedžerje v smeri prave rešitve, včasih tudi v nasprotju z njihovo voljo, v dobro organizacije, uporabnikov ali deležnikov (Granville, 2014, str. 4). Granville (2014, str. 4) omenja, da bi podatkovni znanstvenik moral biti tudi strateg, tako v poslovnem smislu kot v tem, da je sposoben razviti strategijo zbiranja podatkov z namenom pridobiti podlago za odločitve, ki omogočajo poslovni učinek. Ko podatkovni znanstvenik razvije razumevanje domenskega znanja, mora imeti sposobnost predstave, kako se podatki prenašajo preko različnih sistemov in uporabnikov. Pri tem, opozarja Stanton (2013, str. 6), je potrebno, da podatkovni znanstvenik dovolj pozornosti nameni kakovosti. Ne glede na nabor podatkov, ki jih imamo, ti ne bodo nikoli popolni. Podatkovni znanstveniki morajo poznati omejitve podatkov, s katerimi delajo, znati morajo kvantificirati njihovo natančnost in na podlagi analize podati predloge za izboljšanje kakovosti podatkov v prihodnje (Stanton, 2013, str. 6). Zato je sposobnost načrtovanja ključni vidik znanosti o podatkih, saj obstajajo različni načini, kako se lotiti iste naloge, ki pa lahko imajo občutno drugačno porabo virov (Voulgaris, 2014, str. 27).

Sposobnosti v povezavi s področjem vodenja projektov in vodenja na splošno so sposobnosti, ki bi jih posameznik moral imeti zaradi narave dela (predlaganje izboljšav, razvijanje strategij, komuniciranje z naročniki, vodenje projektov ipd.) in dejstva, da podatkovni znanstvenik ni le operativni izvajalec, temveč se pojavlja tudi na različnih vodstvenih položajih ali pa kot vodja oddelka, tima (OECD, 2015, str. 255). Podatkovni znanstvenik po navadi deluje v sklopu raznolike ekipe strokovnjakov z različnih področij (odvisno od dejavnosti). Zelo redko podatkovni znanstvenik dela popolnoma ločeno in samostojno v daljšem časovnem obdobju, saj se za reševanje problema poveže s strokovnjaki, ki so na področju, s katerega izhaja problem, bolj izkušeni oziroma imajo več znanja. Zato je pomembno, da ima podatkovni znanstvenik sposobnost dela v timu. Podatkovni znanstvenik mora biti fleksibilen in imeti sposobnost hitrega prilagajanja novemu poslovnemu področju, novim članov ekipe ali novim programskim orodjem (Voulgaris, 2014, str. 27).

Ker imajo podatkovni znanstveniki po navadi poglobljena znanja iz vsaj enega znanstvenega področja (Lorica, Howard in Dumbill, 2012), kritično sposobnost podatkovnega znanstvenika predstavlja prevajanje med tehničnimi izrazi računalništva in statistike ter slovarjem domenskega znanja menedžmenta. Podatkovni znanstvenik mora zato imeti dobre komunikacijske sposobnosti. Pri tem pride do izraza predvsem sposobnost pripovedovanja zgodb (angl. storytelling), tj. sposobnost z uporabo podatkov predstaviti zgodbo in jo učinkovito prenesti različnim deležnikom (Lorica, Howard in Dumbill,

2012; Stanton, 2013, str. 5). Prednost za podatkovnega znanstvenika je, če ima poleg odličnih komunikacijskih sposobnosti tudi občutek za umetnost in prakso vizualizacije, kar pomeni, da je sposoben smiselno premostiti prepad med človekom in računalnikom s posredovanjem analitičnih dognanj (Lorica, Howard in Dumbill, 2012; Stanton, 2013, str. 6).

Stanton (2013, str. 6) kot pomembno sposobnost podatkovnega znanstvenika omenja tudi sposobnost biti etičen oziroma razmišljati etično. Če so podatki dovolj pomembni, da se jih odločimo zbirati, so po navadi dovolj pomembni, da lahko vplivajo na človeška življenja. Podatkovni znanstveniki morajo razumeti etično odgovornost, povezano z zasebnostjo, in morajo biti sposobni ustrezno predstaviti omejitve z namenom preprečiti zlorabo podatkov ali rezultatov analiz.

Za podatkovnega znanstvenika je ne nazadnje pomembno tudi, da goji strast učenja novih stvari in do dela, ki ga opravlja, ter da ima sposobnost »zaznavanja« podatkov (Granville, 2014, str. 4). Zaradi hitrega napredka tehnologij na področju masovnih podatkov in znanosti o podatkih mora biti podatkovni znanstvenik sposoben hitrega učenja ter hitrega sprejetja novih metod in orodij (Voulgaris, 2014, str. 27). Radovednost, želja po raziskovanju, učenju, strast in vztrajnost se zrcalijo na vseh vidikih življenja podatkovnega znanstvenika (Granville, 2014, str. 4; Lorica, Howard in Dumbill, 2012).

4 RAZISKAVA O ZNANJIH IN SPOSOBNOSTIH PODATKOVNIH ZNANSTVENIKOV V SLOVENIJI

4.1 Metodologija

V raziskavi je sodelovalo 92 posameznikov iz Slovenije, ki se večino svojega časa ukvarjajo s podatki. Vprašalnik, pripravljen na podlagi pregleda teoretičnih izhodišč ter že izvedenih raziskav (Harris, Murphy in Vaisman, 2013; Hayes, 2015a; Swan, 2008), je bil sestavljen iz treh skupin vprašanj, ki se nanašajo na ugotovitve iz prejšnjih razdelkov. Prva skupina vprašanj se je nanašala na posamezne dimenzije masovnih podatkov: volumen, raznolikost, hitrost in vrednost (Piatetsky, 2015; The Emerging Big Returns on Big Data, 2013; Russom, 2011; Big Data Executive Survey, 2012). Druga skupina vprašanj se je nanašala na samooceno znanj ter oceno pomembnosti znanj in sposobnosti, predstavljenih v prejšnjem razdelku, tretja skupina vprašanj pa na pretekle izkušnje in pridobivanje znanj in sposobnosti prek različnih načinov izobraževanja (Harris, Murphy in Vaisman, 2013). Zadnji sklop vprašanj je vključeval demografska vprašanja – spol, starost, stopnja in smer izobrazbe. Za zbiranje podatkov je bila uporabljena metoda spletnega anketiranja s pomočjo strukturiranega vprašalnika. V vprašalniku so bile uporabljene različne merske lestvice: nominalna, ordinalna in intervalna. Pri vprašanjih v drugem sklopu je bila uporabljena petstopenjska Likertova lestvica. Pri določenih vprašanjih je bila dodatno omogočena možnost »ne vem«.

Vzorčni okvir v tem primeru ne obstaja, saj ni popolnega seznama posameznikov, ki se v Sloveniji večino svojega časa ukvarjajo s podatki oziroma s katerim od naslednjih področij: analitika, statistika, matematika, programiranje, menedžment podatkov, raziskovanje ali pa so vodje takšnih ekip. K izpolnitvi vprašalnika so na podlagi dostopnih informacij, preteklih izkušenj, sodelovanj in poznanstev bili povabljeni posamezniki iz različnih organizacij: In516ht, d. o. o., Petrol, d. d., Institut Jožef Stefan, Studio Moderna, d. o. o., Spar Slovenija, d. o. o., Si.Mobil, d. d., Zavarovalnica Triglav, d. d., ADD, d. o. o., Zavarovalnica Maribor, d. d., Mercator, d. d., Kendu, d. o. o., Ekipa2, d. o. o., Javni holding Ljubljana, d. o. o., D.Labs, d. o. o., Adriatic Slovenica, d. d., Nova ljubljanska banka, d. d., Zavod za pokojninsko in invalidsko zavarovanje Slovenije, IBM Slovenija, d. o. o., Ektimo, d. o. o., Revelo, d. o. o., Hekovnik, Arhea Solutio, d. o. o., Valicon, d. o. o., Inštitut za raziskovanje trga in medijev Mediana, d. o. o., Droga Kolinska, d. d., itd. Vprašalnik je bil objavljen tudi na skupini Big Data Developers in Data Science Slovenia ter na Facebook strani Udomačena Statistika. Povabilu k raziskavi je bila dodana tudi prošnja za posredovanje vprašalnika drugim primernim posameznikom. Metoda vzorčenja je bilo namensko priložnostno vzorčenje, saj so bili k izpolnjevanju vprašalnika povabljeni le posamezniki iz Slovenije, ki so ustrezali predhodno določenim kriterijem (ukvarjanje s podatki oziroma ukvarjanje z vnaprej določenim področjem dela). Izpolnjevanje vprašalnika je potekalo od 26. 4. do 21. 5. 2016. Vprašalnik je v celoti izpolnilo 94 oseb; 47 oseb pa je vprašalnik izpolnilo le delno, zato so bili izločeni iz analize. Rezultati ankete so bili ustrezno zakodirani v podatkovno bazo s 94 enotami in 126 spremenljivkami. Pri pregledu podatkov se je izkazalo, da dve enoti nista bili primerni

za analizo, saj sta vsebovali preveliko število neodgovorjenih vprašanj. Končni nabor enot za analizo je zajemal 92 enot. Podatki so bili zbrani s pomočjo spletnega orodja 1ka.si. Za analizo podatkov in vizualizacijo rezultatov sta bila uporabljena SPSS Statistics, verzija 21, in MS Excel 2010. Pri analizi podatkov so bile uporabljene metode opisnih statistik za prikaz rezultatov in značilnosti vzorca, parametrični in neparametrični testi za preverjanje domnev ter metode multivariatne analize (razvrščanje v skupine, algoritem K-means) za identifikacijo skupin podatkovnih znanstvenikov glede na samooceno znanj.

V raziskavi je sodelovalo 92 anketirancev; 59,8 odstotka jih je bilo moškega, 40,2 odstotka pa ženskega spola. Prevladovali so anketiranci moškega spola. Največji delež anketirancev je pripadalo starostni skupini od 26 do 35 let (51,1 odstotka). Mlajši od 18 let ni bil nihče, 1,1 odstotka anketirancev je bilo v starostni skupini od 18 do 25 let, 51,1 odstotka v starostni skupini od 26 do 35 let, 35,9 odstotka v starostni skupin od 36 do 45 let, 10,9 odstotka v starostni skupini od 46 do 55 let in 1,1 odstotka v starostni skupini 56 let in več. Največji delež anketirancev je imel univerzitetno izobrazbo (50 odstotkov). Sledili so anketiranci z magisterijem, doktoratom ali specializacijo (28,3 odstotka) ter anketiranci s poklicno ali štiriletno srednjo šolo (13 odstotkov). Najmanj je bilo anketirancev z višjo ali visoko šolo (8,7 odstotka). V vzorec niso bili zajeti anketiranci s stopnjo izobrazbe osnovna šola ali manj. Največji delež anketirancev je kot svojo prevladujočo smer izobrazbe navedlo računalništvo (26,1 odstotka), sledita ekonomija in poslovne vede (19,6 odstotka), splošno družboslovje (15,2 odstotka), druge naravoslovne ali tehnične vede (14,1 odstotka), statistika (9,8 odstotka) in matematika (8,7 odstotka). En anketiranec je kot smer izobrazbe navedel fiziko (1,1 odstotka).

Pri razvrščanju v skupine so bile kot relevantne spremenljivke vzeti odgovori na vprašanja, povezana s samooceno znanj. Udeleženci so pri petem vprašanju ocenjevali svojo stopnjo znanja z različnih področij znanj. Pri tem je bila uporabljena intervalna lestvica od 1 – Ne poznam (ne uporabljam/ne ustreza mojemu področju dela), 2 – Osnovno poznavanje (osnovno znanje, fokus je na izobraževanju), 3 – Začetnik (znanje pripravnika, fokus je na pridobivanju izkušenj na praktičnih primerih), 4 – Srednji nivo (samostojna kompetentna uporaba, fokus je na izboljšanju znanja) do 5 – Napredni nivo (poglobljena znanja in kompetence, fokus je na reševanju strokovnih problemov). Pri preverjanju korelacij med spremenljivkami je bilo ugotovljeno, da sta edini spremenljivki, ki visoko korelirata med seboj, nadzorovano strojno učenje in nenadzorovano strojno učenje (Pearsonov korelacijski koeficient = 0,924). Korelacija med vsemi ostalimi spremenljivkami je bila manjša kot 0,9. Zaradi navedenega je bila iz nadaljnje analize izpuščena spremenljivka nenadzorovano učenje. Preostale spremenljivke (30) so bile še vedno dobra podlaga za razvrščanje v skupine. Cronbach alpha za preverbo notranje konsistentnosti za 30 spremenljivk in velikost vzorca n = 83 (toliko udeležencev je v celoti odgovorilo na vsa vprašanja, povezana s samooceno znanj) je znašal 0,931, kar pomeni visoko stopnjo notranje konsistentnosti za spremenljivke, merjene na tej lestvici, in za ta specifični vzorec.

Ker je šlo za majhen vzorec in ker končno število skupin ni bilo poznano vnaprej, je bilo najprej izvedeno hierarhično razvrščanje v skupine. Kot mero podobnosti oziroma različnosti med skupinami je bila uporabljena kvadratna evklidska razdalja s standardiziranimi spremenljivkami. Za metodo razvrščanja v skupine je bila izbrana Wardova metoda. Na podlagi hierarhičnega razvrščanja v skupine, pregleda dendograma in izračuna VRC Calinski in Harabasz sta bili za nadaljevanje analize upoštevani razvrstitvi v štiri ali pet skupin. V naslednjem koraku je bilo izvedeno nehierarhično razvrščanje v štiri in pet skupin s K-means algoritmom. Pri uporabi K-means algoritma so bile kot izhodiščne vrednosti upoštevane povprečne ocene, pridobljene iz hierarhičnega razvrščanja v štiri in pet skupin. Na podlagi kvalitativne primerjave razvrščanja v štiri in pet skupin je bilo ugotovljeno, da je razvrščanje v pet skupin identificiralo določeno podmnožico enot znotraj skupine C1 pri razvrščanju v štiri skupine, ki predstavlja drugačen in zanimiv nabor znanj, kot skupina C2 pri razvrščanju v pet skupin. Z razvrstitvijo v pet skupin se je tudi zmanjšala variabilnost znotraj skupin, vendar se je na ta račun povečalo število osamelcev v skupini C2. Zaradi vsega navedenega je bila kot najbolj primerna izbrana razvrstitev v pet skupin.

4.2 Diskusija

Na podlagi vzorčnih podatkov in rezultatov analize podatkov v nadaljevanju predstavljamo ugotovitve po posameznih raziskovalnih vprašanjih.

1. Ali se podatkovni znanstveniki v Sloveniji ukvarjajo z masovnimi podatki in kako se to zrcali skozi različne dimenzije masovnih podatkov?

Analiza rezultatov v Sloveniji je interpretirana in analizirana prek primerjave z rezultati več raziskav, ki so se nanašale na različne dimenzije masovnih podatkov. Dimenzija količine/volumna podatkov je primerjana z rezultati raziskave Kdnuggets.com (Piatetsky, 2015), dimenzija raznolikosti z rezultati raziskave The Emerging Big Returns on Big Data (2013), dimenzija hitrosti z rezultati raziskave TDWI Big Data Analytics (Russom, 2011, str. 19) ter dimenzija vrednosti z rezultati raziskave Big Data Executive Survey (2012, str. 5).

Na podlagi rezultatov raziskave KDnuggets.com (Piatetsky, 2015) glede največjega obdelanega nabora podatkov po geografskih področjih največji delež posameznikov, ki obdeluje podatke v TB ali PB, prihaja iz Združenih držav Amerike in Kanade (26,6 odstotka). Evropa je na četrtem mestu z 20,7 odstotka posameznikov, ki so obdelali TB podatkov ali več. V Evropi so drugače kot največji obdelan nabor podatkov največkrat (60 odstotkov) izbrali podatke v velikosti GB, manj kot 20 odstotkov pa jih obdeluje podatke velikosti MB.

Rezultati iz raziskave v Sloveniji (slika 5) so pokazali, da je približno 25 odstotkov udeležencev kot največji volumen podatkov, s katerim so se ukvarjali, izbralo TB ali PB, kar je v primerjavi z raziskavo KDnuggets.com (Pieatetsky, 2015) bolj na ravni Združenih držav Amerike in Kanade. Udeleženci raziskave v Sloveniji so kot največji volumen podatkov največkrat (44,6 odstotka) izbrali podatke od 1 do 2023 GB. Rezultat je sicer nižji kot delež v raziskavi Kdnuggets.com, vendar še vedno lahko sprejmemo sklep, da se udeleženci raziskave v Sloveniji največ ukvarjajo z obdelavo podatkov velikosti GB.

Slika 5: **Največja obdelana količina/volumen podatkov (n = 92)**

V raziskavi The Emerging Big Returns on Big Data (2013, str. 19) so ugotovili, da je v organizacijah v Evropi povprečno ocenjeni delež strukturiranih podatkov 50 odstotkov, delno strukturiranih je 25 odstotkov, 25 odstotkov pa je nestrukturiranih podatkov. Povprečno ocenjeni delež nestrukturiranih podatkov je najvišji v azijsko-pacifiški regiji, kjer znaša 34 odstotkov. Pričakovalo se je, da bo raziskava v Sloveniji pokazala podobne rezultate, in sicer da se bo približno 25 odstotkov udeležencev že ukvarjalo z nestrukturiranimi podatki.

Rezultati raziskave v Sloveniji (tabela 1) so pokazali, da se je 100 odstotkov udeležencev že ukvarjalo s strukturiranimi podatki, kar je pričakovano, saj so v raziskavi sodelovali posamezniki, ki se ukvarjajo s podatki. Bolj zanimiv je podatek, da se je 50 odstotkov udeležencev raziskave že ukvarjalo z nestrukturiranimi podatki, kar je višje kot pričakovano. Kot nestrukturirani podatki so se upoštevali vsi nestrukturirani podatki (besedilo, avdio, video, slike), generirani s strani človeka.

Tabela 1: **Raznolikost podatkov glede na dimenzijo (n = 92)**

Dimenzija	Delež (v %)
Podatki, generirani s strani naprav	96,7
Podatki, generirani s strani človeka	79,3
Notranji viri	97,8
Zunanji viri	47,8
Strukturirani podatki	100,0
Nestrukturirani podatki	50,0

Na podlagi raziskave The Emerging Big Returns on Big Data (2013, str. 20) so ugotovili, da je v organizacijah v Evropi povprečno ocenjeni delež podatkov, pridobljenih iz notranjih virov, 68 odstotkov, iz zunanjih virov pa 32 odstotkov. Povprečno ocenjeni delež podatkov, pridobljenih iz zunanjih virov, je najvišji v azijsko-pacifiški regiji, kjer znaša 38 odstotkov. Rezultati raziskave v Sloveniji so pokazali, da se je 97,8 odstotka udeležencev že ukvarjalo s podatki iz notranjih virov, kar je pričakovano, saj so v raziskavi sodelovali posamezniki, ki se ukvarjajo s podatki, ki večinoma izvirajo iz notranjih virov organizacije. S podatki iz zunanjih virov pa se je srečalo že skoraj 48 odstotkov udeležencev raziskave, kar je višje kot 32 odstotkov iz raziskave The Emerging Big Returns on Big Data (2013, str. 20). Zanimivo je, da obstaja določen delež posameznikov (2,2 odstotka), ki se ukvarja s podatki izključno iz zunanjih virov. Kot podatke iz zunanjih virov smo upoštevali vse strukturirane in nestrukturirane podatke (besedilo, avdio, video, slike), generirane s strani človeka ali naprave, ki so pridobljeni iz zunanjih virov organizacije.

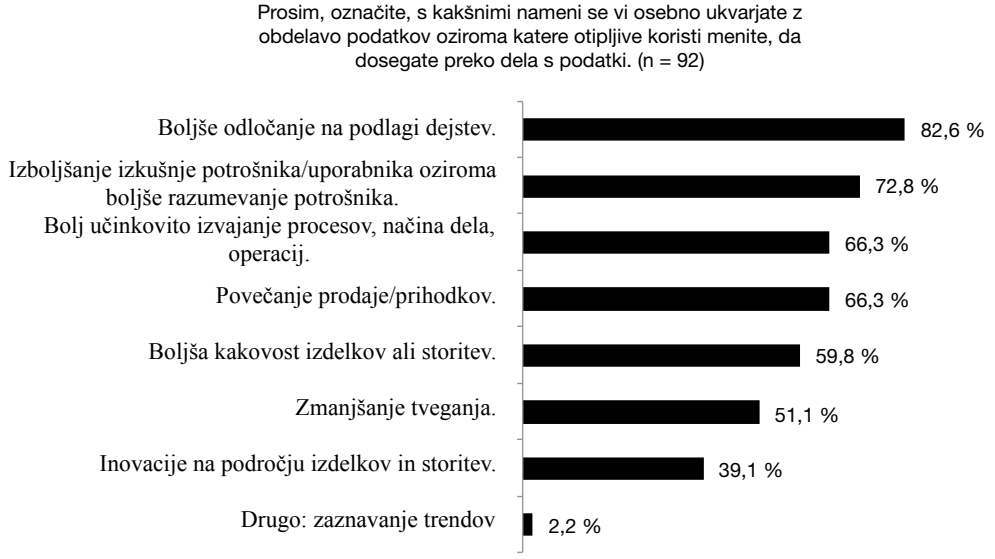
Rezultati obeh raziskav sicer niso v celoti primerljivi, saj so v raziskavi The Emerging Big Returns on Big Data (2013) spraševali po stanju tipov/virov podatkov v organizacijah, v raziskavi v Sloveniji pa smo spraševal posameznike o tem, ali so se že srečali z različnimi tipi, viri podatkov. Vseeno je bila raziskava The Emerging Big Returns on Big Data (2013) uporabljena kot možna primerjava stanja uporabe različnih tipov/virov podatkov. Razlog za razliko v primerjavi z raziskavo, izvedeno v tujini, je lahko tudi izbor (namensko priložnostno vzorčenje) in velikost vzorca (92 enot), vključenega v raziskavo v Sloveniji.

Na podlagi raziskave TDWI Big Data Analytics (Russom, 2011, str. 19) štiri odstotke analiz v organizacijah opravljajo, izvajajo ali ponovno izvajajo v realnem času, štiri odstotke na nivoju ure, pet odstotkov vsakih nekaj ur, 24 odstotkov dnevno, 14 odstotkov tedensko, 35 odstotkov mesečno in 15 odstotkov letno. Delež udeležencev, ki so v raziskavi v Sloveniji označili, da so se že ukvarjali s podatki v realnem času, je bil 47,8 odstotka. Vseeno je pred primerjavo podatkov med raziskavama treba upoštevati, da je šlo za drugačen vzorec in da obstaja možnost, da so udeleženci raziskave v Sloveniji neustrezno razumeli definicijo podatkov v realnem času, kar predstavlja tudi pomembno omejitev raziskave.

Na podlagi raziskave Big Data Executive Survey (2012, str. 5) organizacije uporabljajo masovne podatke za širok nabor namenov. Kot dve najbolj izpostavljeni prednosti uporabe masovnih podatkov so največkrat izbrali boljše odločanje na podlagi dejstev (22 odstotkov) in izboljšanje izkušnje potrošnika/uporabnika (22 odstotkov). Sledi povečanje prodaje/prihodkov (15 odstotkov), inovacije na področju izdelkov in storitev (11 odstotkov), zmanjšanje tveganja (11 odstotkov), boljša kakovost izdelkov in storitev (10 odstotkov) ter bolj učinkovito izvajanje procesov (10 odstotkov). Rezultati raziskave v Sloveniji so pokazali (slika 6), da so udeleženci raziskave kot glavno korist, ki jo dosegajo z delom s podatki, prav tako izbrali boljše odločanje na podlagi dejstev (82,6 odstotka). Sledi izboljšanje izkušnje potrošnika/uporabnika oziroma boljše razumevanje potrošnika (72,2 odstotka), kar se ujema z rezultati iz zgornje raziskave.

Razlike se pojavijo šele pri drugih koristih, saj so udeleženci v Sloveniji kot tretjo korist izbrali bolj učinkovito izvajanje procesov, načina dela, operacij (66,3 odstotka) in povečanje prodaje/prihodkov (66,3 odstotka), medtem ko se je možnost bolj učinkovito izvajanje procesov v zgornji raziskavi pojavilo šele na zadnjih mestih. V Sloveniji so najmanjkrat izbrali inovacije na področju izdelkov in storitev (39,1 odstotka), medtem ko je ta opcija v zgornji raziskavi bila med prvimi štirimi.

Rezultati raziskave so pokazali, da se določeni posamezniki v Sloveniji z vidika dimenzije volumna (TB ali več), raznolikosti (vse dimenzije podatkov) in vrednosti podatkov (boljše odločitve na podlagi dejstev) dejansko ukvarjajo z masovnimi podatki v ožjem smislu. Hkrati pa je bilo z analizo samoocen znanj (v nadaljevanju) ugotovljeno, da so ravno znanja s področja tehnologije masovnih podatkov (masovni in distribuirani podatki, sistemi baz podat-



Slika 6: **Vrednost podatkov (n = 92)**

kov – baze podatkov NoSQL) v povprečju najslabše ocenjena (povprečna ocena je bila okrog 2 – osnove). Zanimivo je bilo, da so tudi z vidika pomembnosti znanj omenjena znanja iz tehnologije masovnih podatkov slabo ocenjena (povprečna ocena pomembnosti je bila okrog 2). Sklepamo lahko, da se v Sloveniji posamezniki ukvarjajo z masovnimi podatki v ožjem smislu, vendar pri tem ne uporabljajo tehnologij masovnih podatkov oziroma teh tehnologij še ne uporabljajo v tolikšni meri, kot so že sprejete v tujini.

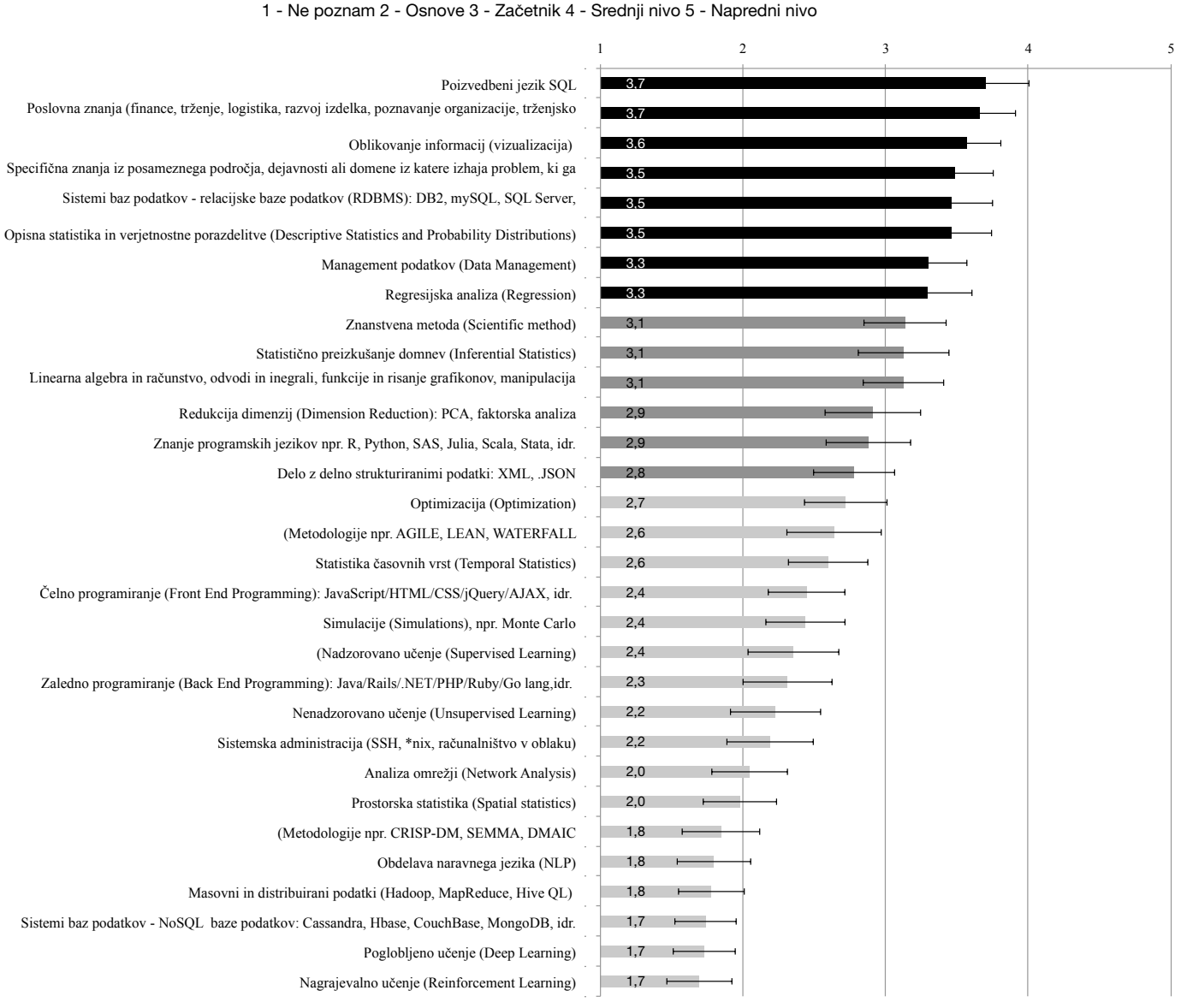
2. Kakšna so dejanska znanja podatkovnih znanstvenikov v Sloveniji ter katera znanja in sposobnosti so pomembna pri njihovem delu?

Analiza rezultatov v Sloveniji je delno interpretirana in analizirana preko primerjave z rezultati raziskave Hayesa (2015a). Raziskava Hayesa (Hayes, 2015a, str. 2–4) o znanjih in sposobnostih podatkovnih znanstvenikov ter delovanju v timih je pokazala, da so sodelujoči v povprečju izrazili višjo stopnjo samoocene na naslednjih področjih: komunikacija, strukturirani podatki, podatkovno rudarjenje, znanost/znanstvena metoda, matematika, menedžment projektov, menedžment podatkov ter statistika in statistično modeliranje. V povprečju pa so nižjo stopnjo samoocene dodelili področjem: sistemska administracija, čelno in zaledno programiranje, procesiranje naravnega jezika (NLP), masovni in distribuirani podatki ter menedžment podatkov v oblaku (Hayes, 2015a, str. 2).

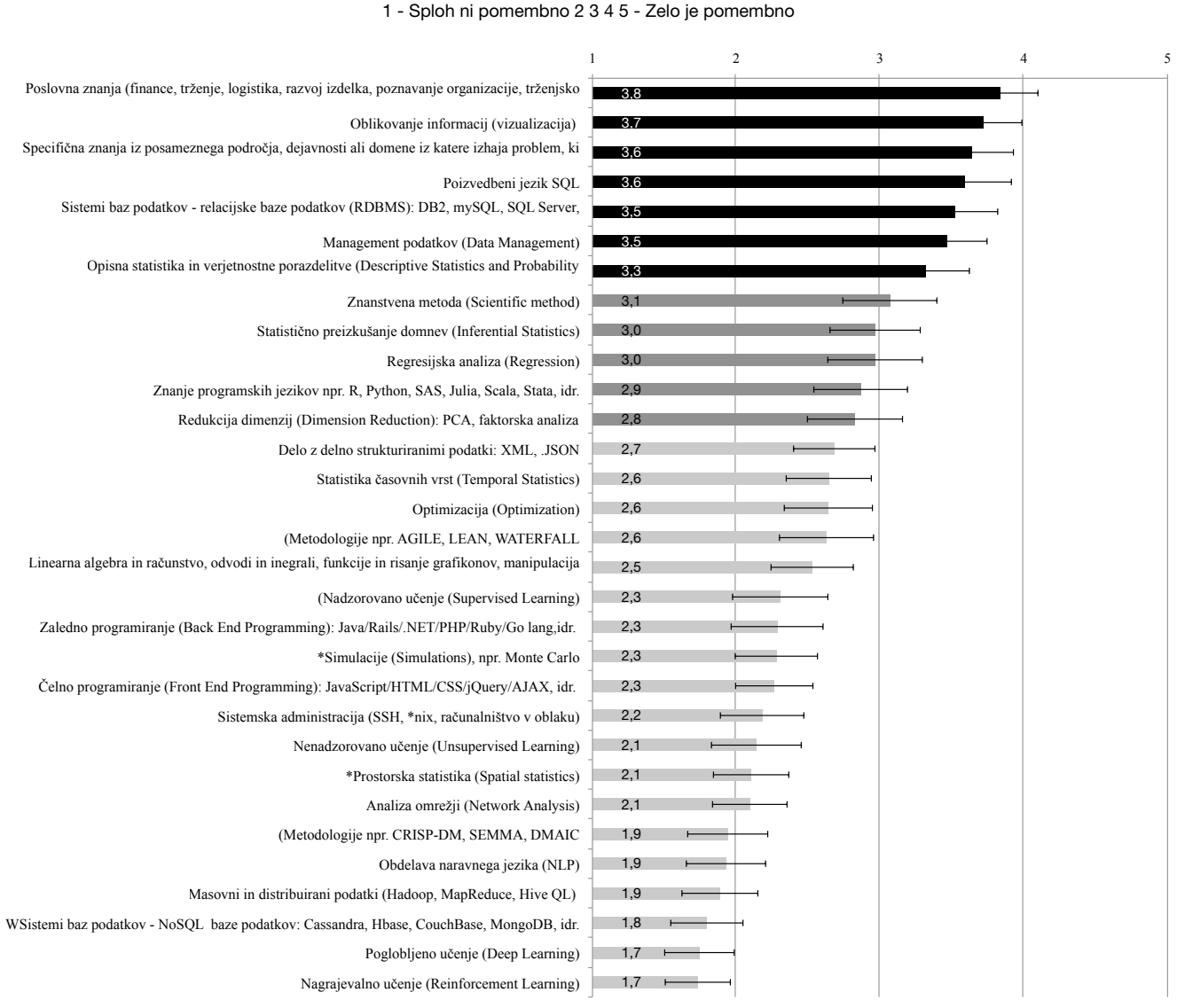
Na podlagi podatkov iz raziskave v Sloveniji je bilo ugotovljeno, da so posamezniki v Sloveniji v povprečju najvišjo samooceno znanj (slika 7) dodelili znanjem s področij baz podatkov (SQL, relacijske baze podatkov, menedžment podatkov), statistike (opisna statistika in verjetnostne porazdelitve ter regresija), domenskih znanj (poslovna znanja, specifična znanja s področja, s katerega izhaja problem) ter oblikovanja informacij. Zanimivo je, da so v sklopu posameznikov, ki se večino svojega časa ukvarjajo s podatki, med najbolj pomembnimi (slika 8) poslovna znanja, oblikovanje informacij in specifična znanja s posameznega področja. Pričakovati bi bilo, da bodo pomembna predvsem znanja iz statistike, baz podatkov ali druga znanja. Razlog verjetno leži v tem, da je rezultate analiz obdelave podatkov treba predstaviti jasno, enostavno in vizualno privlačno, da bodo razumljivi tudi ostalim deležnikom, in seveda z namenom, da prinašajo poslovno vrednost, za kar pa so potrebna poslovna znanja.

V povprečju pa slabo poznajo (slika 7) področja: metodologije strojnega učenja, nagrajevalno in poglobljeno učenje, obdelava naravnega jezika ter tehnologije masovnih podatkov. Razlog je verjetno v tem, da ta znanja pri njihovem delu trenutno niso pomembna, saj so ta področja znanj dobila tudi najnižjo povprečno oceno pomembnosti (slika 8). Z namenom približati tehnologijo masovnih podatkov (Hadoop, MapReduce, baze podatkov NoSQL) ter napredna področja strojnega učenja posamez-

nikom in organizacijam v Sloveniji bi bilo priporočljivo vključiti več primerov dobrih praks s tega področja v sklopu predstavitev na konferencah, povabiti strokovnjake iz tujine, več razširjanja znanja v sklopu družabnih skupin in na srečanjih ter v sklopu formalnega izobraževanja spodbujati uporabo tehnologij masovnih podatkov na odprtih podatkih (angl. open data) ali prek vzajemnega sodelovanja z organizacijami. Primeri dobre prakse in prikazani dejanski učinki uporabe tehnologij masovnih podatkov bi tako spodbudili organizacije, da bodo začele razmišljati o uvajanju teh tehnologij.



Slika 7: **Povprečne ocene samoocene znanj (n = 87–91)**



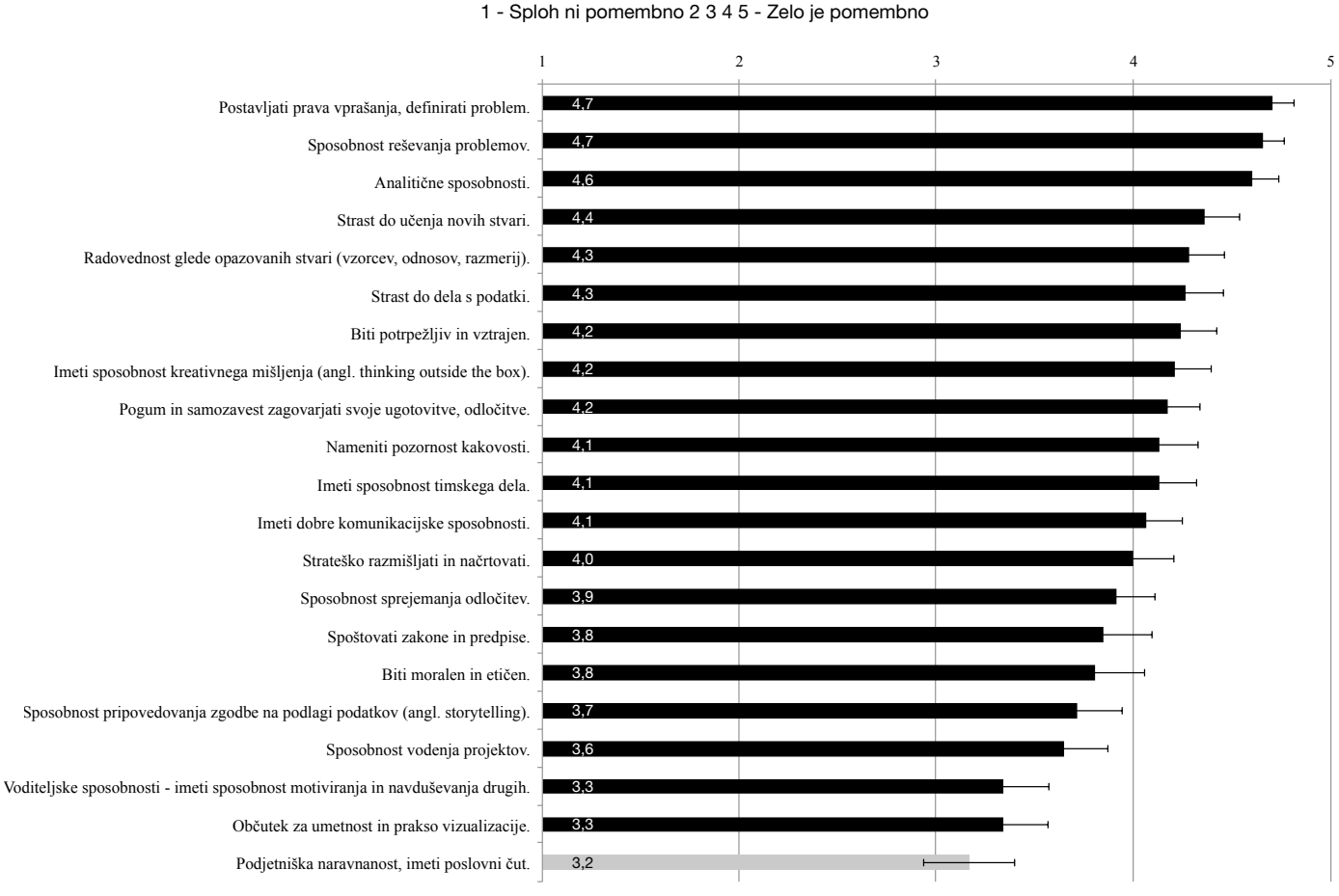
Slika 8: **Povprečne ocene pomembnosti znanj (n = 84–91)**

Da bi ugotovili, ali obstaja statistično značilna razlika med samooceno znanj in pomembnostjo znanj, smo primerjali povprečne samoocene znanja in ocene pomembnosti znanja. Na podlagi statističnega preizkusa je bila identificirana vrzel med povprečno samooceno znanj in povprečno oceno pomembnosti le pri področjih znanj: regresijska analiza (povprečna samoocena = 3,3; povprečna pomembnost: 3,0; P = 0,02) ter linearna algebra in računstvo, odvodi in integrali, funkcije in risanje grafikonov, manipulacija matrik (povprečna samoocena = 3,1; povprečna pomembnost: 2,5; P = 0,000). Pri navedenih znanjih lahko torej trdimo, da obstaja vrzel med obstoječimi znanji (na podlagi samoocene) in pomembnostjo pri njihovem delu. Pri obeh se je izkazalo, da anketiranci menijo, da imajo več znanj, kot je pomembno pri njihovem delu. Rezultat verjetno izhaja iz tega, da se matematike in delno statistike podrobno učimo

v sklopu formalne izobrazbe (osnovna šola, srednja šola itd.), v praksi pa s teh področij uporabljamo le znanja, ki so pri delu pomembna. Anketiranci namreč opravljajo različne funkcije, pri katerih uporabljajo različna znanja glede na njihovo področje dela, prav vsi pa imajo podobna osnovna izhodišča, npr. iz matematike.

Pri primerjavi rezultatov med raziskavo v Sloveniji in raziskavo Hayesa (2015a) je razvidno, da so skupna področja višje samoocenjenih znanj: strukturirani podatki/relacijske baze podatkov (SQL), menedžment podatkov ter do določene mere statistika (v Sloveniji področje regresijske analize, opisne statistike in verjetnostne porazdelitve). V Sloveniji so visoko povprečno samooceno dobila še druga, zgoraj omenjena področja znanj, ki pa v raziskavi Hayesa niso bila zajeta v vprašalnik v takšni obliki. V raziskavi Hayesa so bila v povprečju višje ocenjena znanja iz matematike in znanosti/znanstvene metode, ki sta v Sloveniji dobili v povprečju oceno 3 – Začetnik. Sklepamo lahko, da imajo posamezniki v tujini bolj močno formalno izobrazbo na teh dveh področjih oziroma se več posameznikov s teh dveh področij ukvarja z znanostjo o podatkih ali pa omenjena razlika izhaja le iz drugačne sestave in velikosti vzorca.

Pri pregledu rezultatov raziskave o pomembnosti sposobnosti (slika 9) lahko sklepamo, da so vse sposobnosti, razen »podjetniške naravnanosti«, anketiranci ocenili kot pomembne pri njihovem delu in da pomembno vplivajo na uspešnost njihovega dela. V sklopu pomembnosti sposobnosti so bile kot najbolj pomembne ocenjene postavljanje pravih vprašanj, sposobnost reševanja problemov ter analitične sposobnosti. To dopolnjuje prejšnjo ugotovitev, da v osnovi podatkovni znanstveniki rešujejo poslovne probleme, za kar potrebujejo ustrezne sposobnosti, da znajo pravilno opredeliti problem, se ga lotiti na pravi način in pri tem ustrezno uporabiti vsa svoja znanja.



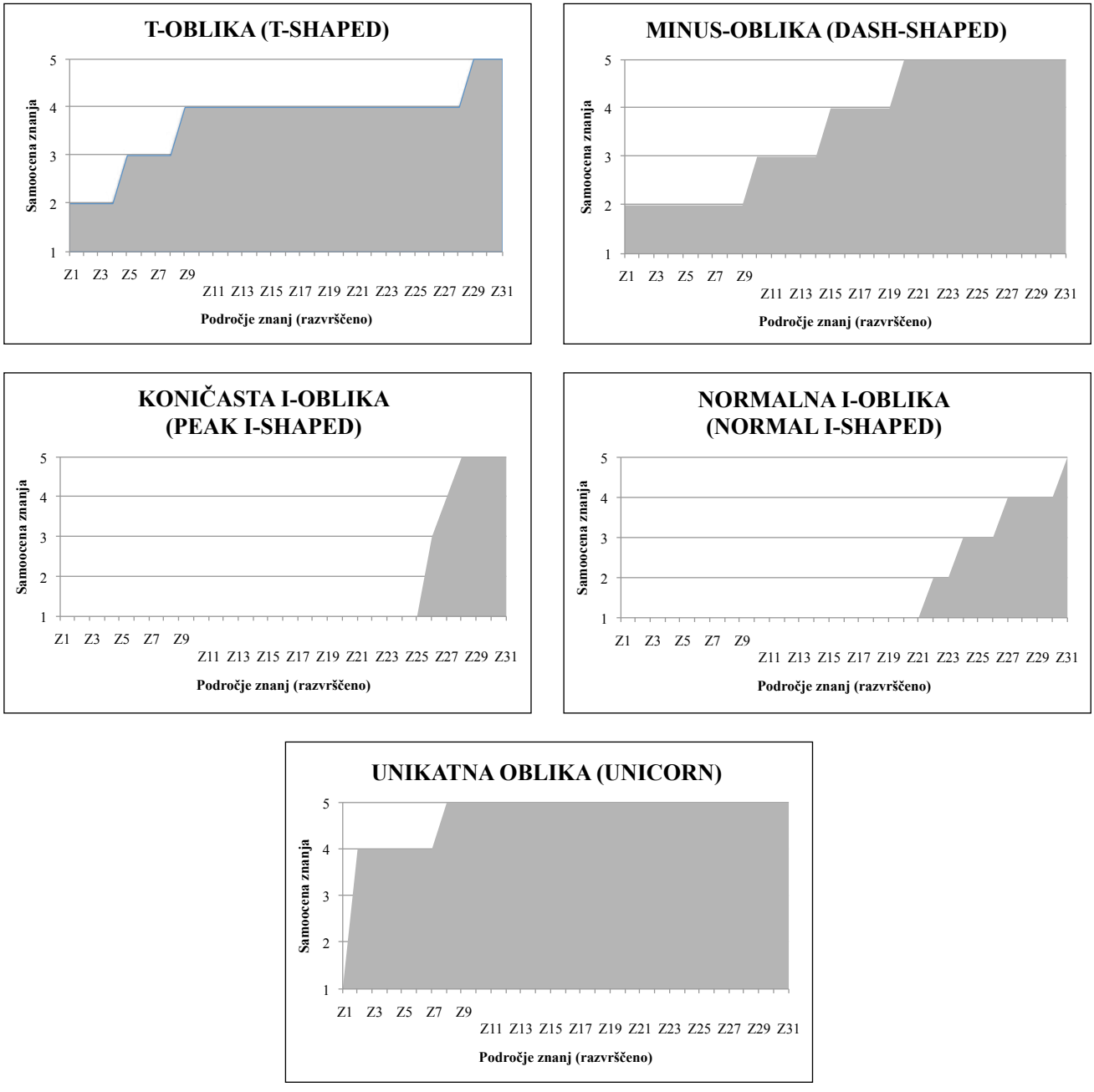
Slika 9: **Povprečne ocene pomembnosti sposobnosti (n = 90–92)**

3. Ali je mogoče identificirati porazdelitve oziroma vzorce znanj po posameznih področjih med identificiranimi skupinami podatkovnih znanstvenikov na podlagi samoocene znanj?

Zaradi širokega področja znanj in sposobnosti, ki naj bi jih imel posameznik za delo podatkovnega znanstvenika, se v literaturi (Granville, 2014, str. 75) in v raziskavah (Harris, Murphy in Vaisman, 2013) omenja, da naj bi posameznik imel zelo poglobljena znanja z vsaj enega področja (statistike, matematike, programiranja, baz podatkov, strojnega učenja) ter vsaj osnovna znanja z drugih področij. Granville (2014) govori o t. i. vertikalnih podatkovnih znanstvenikih, v raziskavi Analyzing the analyzers (Harris, Murphy in Vaisman, 2013) pa so identificirali t. i. T-obliko znanj podatkovnih znanstvenikov. V sklopu raziskave smo želeli ugotoviti, ali je mogoče identificirati porazdelitve oziroma vzorce znanj po posameznih področjih znanj med skupinami podatkovnih znanstvenikov. Na podlagi tega smo želeli ugotoviti, ali obstajajo podatkovni znanstveniki s t. i. T-obliko znanj oz. ali obstajajo posamezniki, ki imajo visoko stopnjo znanj na vseh področjih.

Pri analizi sta bila uporabljena koeficient asimetrije (angl. skewness) in koeficient sploščenosti (angl. kurtosis). Če je koeficient asimetrije večji od 0, je porazdelitev asimetrična v desno; če je enak 0, je porazdelitev simetrična; če pa je manjši od 0, je porazdelitev asimetrična v levo. Če je koeficient sploščenosti večji od 0, je porazdelitev koničasta; če je enak 0, je porazdelitev normalna (angl. bell-shaped); če pa je manjši od 0, je porazdelitev sploščena. Z namenom identifikacije porazdelitve oziroma vzorcev znanj smo rezultate glede samoocene znanj vseh posameznikov, ki so sodelovali v raziskavi, izvozili v Excel, njihove samoocene znanj pa smo razvrstili od najmanjše do najvišje ter jih vizualizirali z orodjem Sparkline v Excelu. Za vsakega udeleženca smo izračunali koeficient asimetrije KA ter koeficient sploščenosti KS. Ugotovili smo, da je mogoče identificirati naslednje porazdelitve oziroma vzorce znanj posameznikov (n = 92) (slika 10):

- T-oblika (angl. T-shaped), če sta KA in KS med –1 in 1. To so posamezniki, ki imajo visoko samooceno znanj na določenem področju znanj ali parih področij znanj, na drugih področjih pa imajo začetna ali osnovna znanja. Njihova porazdelitev znanj je simetrična ter približno podobna normalni porazdelitvi. Takšnih anketirancev je bilo v vzorcu 31,5 odstotka.
- Minus-oblika (angl. Dash-shaped), če je KA med –1 in 1 ter KS manjši od –1. To so posamezniki, ki imajo simetrično porazdelitev znanj po področjih, vendar je njihova porazdelitev bolj sploščena (KS je manjši od –1). To pomeni, da imajo z veliko področij znanj določeno stopnjo znanja, na nobenem področju pa ne izstopajo ali pa hkrati izstopajo na več področjih. Takšnih anketirancev je bilo v vzorcu 34,8 odstotka.
- Normalna I-oblika (angl. Normal I-shaped), če je KA večji od 1 ter KS med –1 in 1. To so posamezniki, ki imajo porazdelitev znanj asimetrično v desno. To pomeni, da imajo določeno področje znanja, ki ima visoko oceno, pri ostalih pa imajo zelo nizke samoocene znanja ali pa jih sploh ne poznajo. Njihova značilnost je še, da je njihova porazdelitev precej podobna normalni (KS je med –1 in 1). To pomeni, da imajo določeno poznavanje ostalih znanj, čeprav ne tako visoko kot posamezniki pri T-obliki. Takšnih anketirancev je bilo v vzorcu 21,7 odstotka.
- Koničasta I-oblika (angl. Peak I-shaped), če sta KA in KS večja od 1. To so posamezniki, ki so podobni »normalni I-obliki«, vendar je njihova »koničastost« še bolj izrazita (KS je večji od 1), kar še dodatno poudarja višjo oceno znanj iz samo določenega področja. Takšnih anketirancev je bilo v vzorcu 5,4 odstotka.
- Unikatna oblika (angl. Unicorn), če je KA manjši od –1 ter KS večji od 1. Na podlagi podatkov so bili identificirani tudi posamezniki, ki imajo porazdelitve znanj zelo asimetrične v levo (KA je manjši od –1) ter zelo »koničasto« porazdelitev. To pomeni, da imajo visoko samooceno iz vseh znanj, kar jih naredi zelo unikatne. Takšnih anketirancev je bilo v vzorcu 6,5 odstotka.



Slika 10: **Porazdelitve znanj po izbranih anketirancih – samoocena znanj**

4. Katere skupine podatkovnih znanstvenikov v Sloveniji lahko identificiramo na podlagi samoocene znanj?

Na podlagi razvrščanja v skupine z algoritmom K-means je bilo identificiranih pet skupin posameznikov v Sloveniji, ki se med seboj razlikujejo glede samoocene svojih znanj. Centroidi skupin so prikazani na sliki 11.

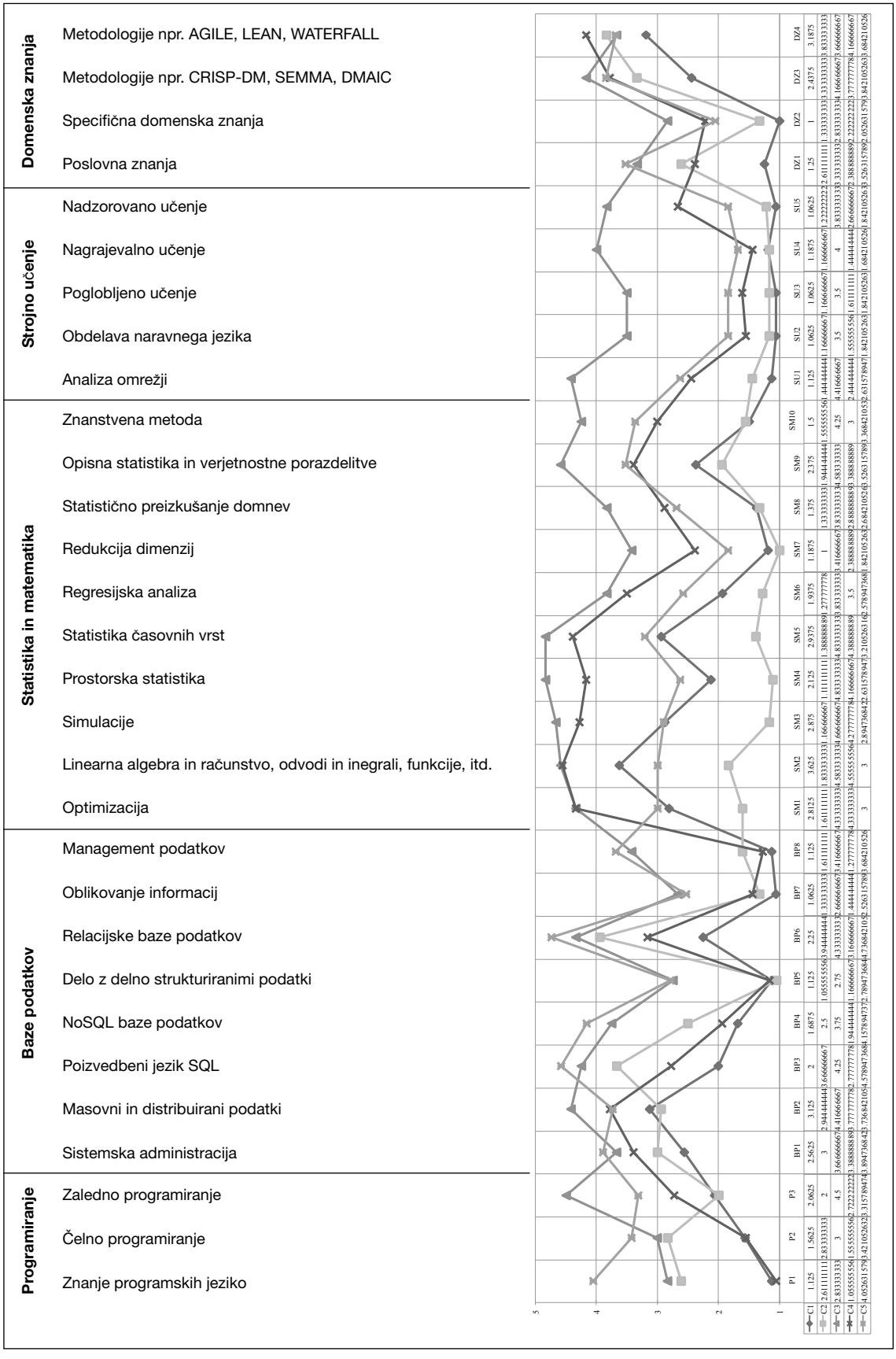
Interpretacija skupine C1: posamezniki v skupini C1 imajo v povprečju osnovna znanja (2,1) iz programskih jezikov, npr. R, Python ter čelnega programiranja. Zaledno programiranje pa so v povprečju ocenili z Ne poznam ali uporabljam. V sklopu znanj iz menedžmenta in baz podatkov imajo povprečno oceno 3 pri oblikovanju informacij (vizualizacija) in menedžmentu podatkov. Osnovna znanja imajo v povprečju iz relacijskih baz podatkov, delno strukturiranih podatkov in poizvedbenega jezika SQL. Masovnih in distribuiranih podatkov, baz podatkov

NoSQL ne poznajo, prav tako se ne ukvarjajo s sistemsko administracijo. V sklopu statistike in matematike imajo najvišjo povprečno oceno iz opisnih statistik in verjetnostnih porazdelitev (3,6), statističnega preizkušanja domnev (2,9), regresijske analize (2,9) ter znanstvene metode (2,8). Najmanj poznajo ali uporabljajo prostorsko statistiko (1,2), optimizacije (1,5) in simulacije (1,4). Celoten sklop strojnega učenja v povprečju ne poznajo ali ne uporabljajo. Prav tako v povprečju ne poznajo metodologij AGILE, LEAN, WATERFALL in CRISP-DM, SEMMA, DMAIC. Osnovna znanja imajo iz specifičnih znanj s posameznega področja ali domene, iz katere izhaja problem (2,4) ter začetna znanja s področja poslovnih znanj (3,2). Na podlagi navedenih značilnosti je bila ta skupina poimenovana trženjski raziskovalci – analitiki in zavzema 19 odstotkov posameznikov iz vzorca.

Trženjski raziskovalci – analitiki imajo torej začetni nivo znanj iz statistike in matematike ter poslovnih znanj. Dodatna analiza te skupine je pokazala, da jih ima polovica univerzitetno izobrazbo, smer izobrazbe pa je ekonomska (31 odstotkov) in družboslovna (38 odstotkov). Oba spola sta enako zastopana (50 odstotkov). Menijo, da imajo manj znanj, kot pa so pomembna, s področij oblikovanja informacij, relacijskih baz podatkov, poglobljenega učenja in analize omrežij. Z vidika dimenzije raznolikosti podatkov se jih največji delež ukvarja s MB (38 odstotkov) in GB (38 odstotkov) podatki ter z notranjimi, strukturiranimi podatki, generiranimi s strani naprav ali človeka (31 odstotkov). V tej skupini je najnižji delež takšnih, ki so se srečali z vsemi dimenzijami raznolikosti podatkov (13 odstotkov). Z vidika porazdelitve samoocene znanj v tej skupini prevladujejo posamezniki (50 odstotkov), ki imajo normalno I-obliko porazdelitve znanj. To pomeni, da imajo določeno področje znanja visoko ocenjeno, pri ostalih pa imajo nizke samoocene znanja ali pa jih sploh ne poznajo. Ta skupina naj v prihodnosti predvsem razvija naprej znanja iz statistike in matematike, ki sta njeni najmočnejši področji. Manjka jim predvsem razširitev njihovih znanj (vsaj na osnove) s področij znanj programiranja, baz podatkov, menedžmenta podatkov, strojnega učenja in domenskih znanj, da bi postali podatkovni znanstveniki. S programiranjem bi se lahko seznanili ob uporabi programskega jezika R, v katerem bi se lahko hitro naučili izvedbo ukazov s področja statistike, ki bi jih lahko takoj uporabili pri svojem delu. Ker se pogosto udeležujejo delavnic ali tečajev, bi lahko na kateri od njih predstavili nove možnosti vizualizacije, osnove baz podatkov ter metode strojnega učenja, ki bi jih lahko kar najhitreje praktično uporabili pri svojem delu. Pozitivni učinki uporabe pri delu bi jih spodbudili k nadaljnji uporabi in raziskovanju teh področij tudi v prihodnje, saj določen del te skupine meni, da sta pri delu s podatki pomembni radovednost in strast.

Interpretacija skupine C2: Posamezniki iz skupine C2 imajo v povprečju osnovna znanja iz programskih jezikov (2,0) ter začetna znanja iz zalednega programiranja (2,6) in čelnega programiranja (2,8). V sklopu znanj iz menedžmenta in baz podatkov imajo znanja na srednjem nivoju poizvedbenega jezika SQL (3,9) ter iz relacijskih baz podatkov (3,7). Začetna znanja imajo na področju menedžmenta podatkov (3,0), oblikovanja informacij (2,9) ter dela z delno strukturiranimi podatki (2,5). Masovnih in distribuiranih podatkov, baz podatkov NoSQL ne poznajo, pri sistemski administraciji poznajo osnove. Na področju statistike in matematike v povprečju nimajo znanj ali jih ne uporabljajo, ali pa poznajo le osnove (znanstvena metoda, opisna statistika, linearna algebra in računstvo, odvodi in integrali, funkcije in risanje grafikonov, manipulacija matrik, optimizacija). Celotnega sklopa strojnega učenja v povprečju ne poznajo ali ga ne uporabljajo. Prav tako ne poznajo metodologij CRISP-DM, SEMMA, DMAIC. Bolj so seznanjeni z metodologijami AGILE, LEAN, WATERFALL (2,6). Začetna znanja imajo iz specifičnih znanj s posameznega področja ali domene, iz katere izhaja problem (3,3), ter srednji nivo znanj iz področja poslovnih znanj (3,8). Na podlagi navedenih značilnosti je bila ta skupina poimenovana podatkovni analitiki in zavzema 22 odstotkov posameznikov iz vzorca.

Podatkovni analitiki imajo torej srednji nivo znanj s področij baz podatkov ter poslovnih znanj ter začetna znanja s področja pogramiranja. Dodatna analiza te skupine je pokazala, da v tej skupini prevladujejo moški (61 odstotkov), univerzitetna izobrazba (56 odstotkov), smer izobrazbe računalništvo (28 odstotkov) ter druge naravoslovne in tehnične vede (22 odstotkov). Menijo, da imajo več znanj, kot so pomembna, s področij: čelno programiranje, regresija, nadzorovano učenje, linearna algebra in računstvo, odvodi in integrali, funkcije in risanje grafikonov, manipulacija matrik. Imajo potencial, da


Domenska znanja
Metodologije npr. AGILE, LEAN, WATERFALL
Metodologije npr. CRISP-DM, SEMMA, DMAIC
Specifična domenska znanja
Poslovna znanja
Strojno učenje
Nadzorovano učenje
Nagrajevalno učenje
Poglobljeno učenje
Obdelava naravnega jezika
Analiza omrežji
Statistika in matematika
Znanstvena metoda
Opisna statistika in verjetnostne porazdelitve
Statistično preizkušanje domnev
Redukcija dimenzij
Regresijska analiza
Statistika časovnih vrst
Prostorska statistika
Simulacije
Linearna algebra in računstvo, odvodi in inegrali, funkcije, itd.
Optimizacija
Baze podatkov
Management podatkov
Oblikovanje informacij
Relacijske baze podatkov
Delo z delno strukturiranimi podatki
NoSQL baze podatkov
Poizvedbeni jezik SQL
Masovni in distribuirani podatki
Sistemska administracija
Programiranje
Zaledno programiranje
Čelno programiranje
Znanje programskih jeziko

postanejo podatkovni znanstveniki, saj se jih že sedaj 44 odstotkov ukvarja s podatki v GB ter skoraj 40 odstotkov z vsemi dimenzijami raznolikosti podatkov, prav tako pa očitno delajo na področjih, na katerih so pomembna znanja programiranja (čelno) in strojnega učenja. Z vidika porazdelitve samoocene znanj v tej skupini prevladujejo posamezniki (61 odstotkov), ki imajo normalno I-obliko porazdelitve znanj. To pomeni, da imajo določeno področje znanja visoko ocenjeno, pri ostalih pa imajo nizke samoocene znanja ali pa teh znanj sploh nimajo. Ta skupina naj v nadaljevanju razvija znanja s področij znanosti/znanstvena metoda, programiranja, statistike, strojnega učenja in domenskih znanj. Njihova prednost je v dobri osnovi na področju baz podatkov in poslovnih znanjih.

Interpretacija skupine C3: Posamezniki iz skupine C3 imajo v povprečju napredna znanja iz programskih jezikov (4,5) ter srednji nivo znanj iz zalednega programiranja (2,8) in čelnega programiranja (3). V sklopu znanj iz menedžmenta in baz podatkov imajo srednji nivo znanj iz oblikovanja informacij (4,4), poizvedbenega jezika SQL (4,3), relacijskih baz podatkov (4,3), delno strukturiranih podatkov (3,8) in menedžmenta podatkov (3,7). Začetna znanja imajo na področju sistemske administracije (3,4), masovnih in distribuiranih podatkov (2,7) in baz podatkov NoSQL (2,8). V sklopu statistike in matematike imajo pri vseh področjih znanj v povprečju srednja ali napredna znanja, prav tako na področju strojnega učenja. Na področju domenskih znanj pa imajo začetni nivo znanj iz metodologij AGILE, LEAN, WATERFALL ter CRISP-DM, SEMMA, DMAIC, pri ostalih domenskih znanjih pa srednji nivo. Posamezniki iz te skupine so v primerjavi z ostalimi skupinami edini, ki imajo največje število področij znanj ocenjeno s povprečno oceno 3 ali več. Na podlagi navedenih značilnosti je bila ta skupina poimenovana podatkovni znanstveniki in zavzema 14 odstotkov posameznikov iz vzorca.

Podatkovni znanstveniki izstopajo predvsem po naprednem znanju programskih jezikov in so edini od skupin, ki imajo največje število področij znanj ocenjeno s povprečno oceno 3 (začetnik) ali več. Dodatna analiza te skupine je pokazala, da prevladujejo moški (83 odstotkov), univerzitetna izobrazba (56 odstotkov) ter računalniška smer izobrazbe (33 odstotkov). Menijo, da imajo več znanj, kot pa so pomembna, na področjih opisna statistika in verjetnostne porazdelitve ter redukcija dimenzij. Ukvarjajo se s podatki v GB (42 odstotkov) in TB (41 odstotkov). Polovica se je že srečala z vsemi dimenzijami raznolikosti podatkov. Kakovost izdelkov in storitev jim je enako pomembna kot odločanje na podlagi dejstev. Z vidika porazdelitve samoocene znanj v tej skupini prevladujejo posamezniki, ki imajo unikatno obliko (42 odstotkov) in T-obliko (42 odstotkov) porazdelitve znanj. V primerjavi z drugimi skupinami so tudi edina skupina, ki sploh vsebuje unikatno obliko porazdelitve znanj. Za to skupino je predvsem pomembno to, da lahko svoje bogato znanje uporabijo v praksi na zanimivih projektih, ki jim bodo predstavljali izziv. Svoje znanje že sedaj izpopolnjujejo in ga bodo tudi v prihodnje, zato je pomembno, da so obveščeni o aktualnih konferencah v Sloveniji in tujini ter aktualnih natečajih in tekmovanjih. Ker so jim verjetno najbolj pomembne praktične izkušnje iz izvedenih projektov, bi lahko znanje medsebojno delili prek srečanj v družabnih skupinah.

Interpretacija skupine C4: Posamezniki iz skupine C4 v povprečju ne uporabljajo zalednega programiranja ali ga ne poznajo. Začetna znanja imajo iz programskih jezikov (2,7) in čelnega programiranja (1,6). V sklopu znanj iz menedžmenta in baz podatkov imajo srednji nivo znanj iz oblikovanja informacij (3,8), začetni nivo iz menedžmenta podatkov (3,4), poizvedbenega jezika SQL (3,2), relacijskih baz podatkov (2,8). V povprečju imajo osnove iz dela z delno strukturiranimi podatki (1,9). Masovnih in distribuiranih podatkov, baz podatkov NoSQL v povprečju ne poznajo, prav tako se ne ukvarjajo s sistemsko administracijo. V sklopu statistike in matematike imajo pri vseh področjih znanj v povprečju srednji nivo znanj, razen pri prostorski statistiki (2,4), simulacijah (2,9), optimizaciji (3). S področja strojnega učenja imajo v povprečju osnovna znanja, razen pri obdelavi naravnega jezika (1,4), ki ga ne poznajo ali uporabljajo, ter analizi omrežij (2,7), za katero imajo začetna znanja. Na področju domenskih znanj pa imajo osnovni nivo znanj iz obeh metodologij. Srednji nivo znanj imajo iz specifičnih znanj (3,8) in poslovnih znanj (4,2). Na podlagi navedenih značilnosti je bila ta skupina poimenovana raziskovalci in zavzema 22 odstotkov posameznikov iz vzorca.

Raziskovalci zelo dobro kombinirajo znanja s področja oblikovanja informacij, poslovnih znanj, baz podatkov ter statistike. Dodatna analiza te skupine je pokazala, da v tem segmentu izjemoma prevladujejo ženske (73 odstotkov), univerzitetna izobraba

(50 odstotkov) ter smer izobrazbe statistika in ekonomija. Predstavljajo zelo dober potencial, da postanejo podatkovni znanstveniki, saj iz te skupine prihaja najvišji delež posameznikov, ki je kot največjo obdelano količino podatkov izbralo PB (6 odstotkov), drugače pa se ukvarjajo s podatki v MB (33 odstotkov) in GB (39 odstotkov). V tej skupini se jih največ (33 odstotkov) ukvarja z vsemi dimenzijami raznolikosti podatkov. Z vidika porazdelitve samoocene znanj v tej skupini prevladujejo posamezniki, ki imajo minus obliko (72,2 odstotkov) porazdelitve znanj. To pomeni, da imajo določeno stopnjo znanja z veliko področij, na nobenem pa ne izstopajo ali pa izstopajo na več področjih. Ta skupina naj v nadaljevanju razvija znanja s področja programiranja in strojnega učenja ter domenska znanja. Ker imajo dobra znanja s področja vizualizacije, bi jim lahko približali programske jezike in programiranje prek različnih načinov vizualizacij podatkov s pomočjo programskih jezikov. Z uporabo strojnega učenja pa bi lahko izboljšali rezultate, kjer si želijo doseči boljše odločanje na podlagi dejstev. V primerjavi z drugimi skupinami se veliko udeležujejo masovnih odprtih spletnih tečajev, na katerih lahko pridobijo omenjena znanja.

Interpretacija skupine C5: Posamezniki iz skupine C5 imajo v povprečju srednji nivo znanj iz zalednega programiranja (4,1) ter osnovni nivo iz čelnega programiranja (3,4) ter programskih jezikov (3,3). Znanja programiranja najbolj izstopajo v tej skupini od vseh naštetih skupin. V sklopu znanj iz menedžmenta in baz podatkov imajo napredni nivo znanj iz relacijskih baz podatkov (4,6) in poizvedbenega jezika SQL (4,7). Srednji nivo znanj imajo iz menedžmenta podatkov, oblikovanja informacij, dela z delno strukturiranimi podatki in sistemsko administracijo. So edina skupina, ki ima začetni nivo znanj na področju baz podatkov NoSQL (2,8) ter masovnih in distribuiranih podatkov (2,5). V sklopu statistike in matematike imajo pri vseh področjih znanj v povprečju začetni nivo znanj, razen pri linearni algebri in računstvu (3,5) in prostorski statistiki (1,8). Na področju strojnega učenja imajo v povprečju osnovna znanja, razen pri nadzorovanem učenju (2,6), pri katerem imajo začetna znanja. Na področju domenskih znanj imajo srednji nivo znanj metodologij AGILE, LEAN, WATERFALL ter osnove iz metodologij CRISP-DM, SEMMA, DMAIC. Srednji nivo znanj imajo iz specifičnih znanj (3,8) in poslovnih znanj (3,7). Na podlagi navedenih značilnosti je bila ta skupina poimenovana programerji in zavzema 23 odstotkov posameznikov iz vzorca.

Programerji izstopajo z najvišjo povprečno samooceno znanj iz programiranja, baz podatkov in domenskih znanj. Prevladujejo moški (84 odstotkov), univerzitetna izobrazba (37 odstotkov) in smer izobrazbe računalništvo (53 odstotkov). So edina skupina, ki ima začetni nivo znanj na področju baz podatkov NoSQL ter osnovni nivo znanj iz masovnih in distribuiranih podatkov. To potrjuje tudi dejstvo, da se jih največ ukvarja s podatki v GB (53 odstotkov) in TB (32 odstotkov). Ker menijo, da imajo manj poslovnih znanj, kot pa so pomembna pri njihovem delu, naj razvijajo znanja s področja poslovnih ved. Da bi postali podatkovni znanstveniki, jim manjkajo še znanja s področja znanosti/znanstvene metode in statistike. Z vidika porazdelitve samoocene znanj v tej skupini prevladujejo posamezniki, ki imajo T-obliko (53 odstotkov) in minus obliko porazdelitve znanj. Priporočljivo je, da ta skupina tesno sodeluje s skupino podatkovnih znanstvenikov pri različnih projektih. Na podlagi skupnega sodelovanja bodo lahko programerji pridobili vpogled v znanstveni pristop k podatkom. Znanja s področij statistike, znanstvene metode in poslovnih znanj lahko pridobijo s formalno izobrazbo ali vsaj z udeležbo na masovnem odprtem spletnem tečaju.

Čeprav smo pri oblikovanju raziskovalnega načrta, metodologije ter pri sami izvedbi raziskave in analizi podatkov kar najbolj upoštevali raziskovalne standarde (Malhotra, 2012), ima raziskava tudi določene omejitve. Prva omejitev izhaja iz velikosti vzorca – če bi bila velikost vzorca večja, bi lahko bili rezultati bolj zanesljivi. Poleg tega je bilo uporabljeno priložnostno namensko vzorčenje (neverjetnostno vzorčenje), kar pomeni, da vzorec ni reprezentativen in rezultatov raziskave ni mogoče posplošiti na populacijo. Kljub temu so bili k raziskavi povabljeni posamezniki, ki s svojim področjem dela pokrivajo širok nabor strokovnjakov, ki bi jih lahko uvrščali med podatkovne znanstvenike, zato menimo, da rezultati raziskave ponujajo dober vpogled v stanje na tem področju in dajejo podlage za ukrepanje.

5 SKLEPNE UGOTOVITVE

Konvergenca različnih znanstvenih disciplin je omogočila pojav novega razreda strokovnjaka, podatkovnega znanstvenika. Podatkovni znanstvenik naj bi imel znanja s področij programiranja, menedžmenta

podatkov, baz podatkov, znanosti (znanstvena metoda), statistike, matematike, strojnega učenja in domenskih znanj. V sklopu sposobnosti pa so pomembni: sposobnost definiranja in reševanja problemov, analitične sposobnosti, strast do učenja novih stvari, radovednost, strast do dela s podatki, potrpežljivost, vztrajnost, kreativno mišljenje, pogum in samozavest zagovarjati svoje odločitve, pozornost nameniti kakovosti, sposobnost timskega dela, sposobnost komunikacije, strateško razmišljanje, sposobnost sprejemanja odločitev, spoštovanje zakonov in predpisov, moralnost in etičnost, sposobnost pripovedovanja zgodbe, sposobnost vodenja projektov, sposobnost motiviranja in navduševanja drugih ter občutek za umetnost in prakso vizualizacije. Na ta način ima namreč vse potrebno, da lahko samostojno izvede celoten proces znanosti o podatkih. Na podlagi identificiranih segmentov in njihovih značilnosti lahko sklepamo, da v Sloveniji obstajajo posamezniki, ki bi jim lahko podelili naziv podatkovni znanstvenik, saj imajo znanja in sposobnosti z vseh identificiranih področij znanj, s katerimi lahko pokrijejo celoten proces izvajanja znanosti o podatkih. Glede na podatke o rastočem povpraševanju po takšnih posameznikih v svetu bo v prihodnosti predvsem pomembno ustvariti okolje in razmere, da bodo takšni posamezniki našli ustrezne izzive za izpopolnitev svojega zmožnosti v Sloveniji, hkrati pa razviti oziroma dopolniti zmožnosti preostalih posameznikov iz identificiranih skupin. Pri tem bo zelo pomembna podpora v sklopu formalnega izobraževanja na navedenih področjih, stalno izpopolnjevanje, prenos znanja med posamezniki in skupinami ter pridobivanje izkušenj na praktičnih primerih. V ta namen bi bilo treba še bolj spodbujati srečanja v družabnih skupinah, omogočiti delo na »odprtih« podatkih ter ustrezna znanja za opravljanje takšne pozicije vključiti v del redne formalne izobrazbe.

6 LITERATURA IN VIRI

[1] *Big Data Executive Survey* (2012). Najdeno 10. 1. 2015 na http://newvantage.com/wp-content/uploads/2012/12/NVP--Big-Data-Survey-Themes-Trends.pdf.

[2] Boyd, D. in Crawford, K. (2012). Critical questions for big data. Information, *Communication & Society, 15(5), 662–679*.

[3] Chordas, L. (2014). Data driven. *Best's Review*, 115(1), 22–26.

[4] DAMA (2014, 6. marec). DAMA-DMBOK2 Framework Guide. *Dama*. Najdeno 21. 3. 2016 na https://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf.

[5] Davenport, T. T., in Patil, D. J. (2012). Data scientists: the sexiest job of the 21st century. *Harvard Business* Review, oktober 2012, 70–76.

[6] Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*, 56(12), 64–73.

[7] *The Emerging Big Returns on Big Data* (2013). Najdeno 16. 1. 2015 na http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf.

[8] *The field guide to data science*. Najdeno 10. 1. 2015 na https://www.boozallen.com/content/dam/boozallen/documents/2015/12/2015-FIeld-Guide-To-Data-Science.pdf.

[9] Granville, V. (2013). Job titles for data scientists. *Datasciencecentral*. Najdeno 5. 12. 2015 na http://www.datasciencecentral.com/profiles/blogs/job-titles-for-data-scientists.

[10] Granville, V. (2014). *Developing analytic talent: becoming a data scientist*. United States: Wiley.

[11] Harris, H., Murphy, S. in Vaisman, M. (2013). *Analyzing the analyzers: an introspective survey of data scientists and their work*. United States: O'Reilly Media.

[12] Hayes, B. E. (2014a). The One hidden skill you need to unlock the value of your data. *Businessoverbroadway*. Najdeno 5. 12. 2015 na http://businessoverbroadway.com/the-one-hidden-skill-you-need-to-unlock-the-value-of-your-data.

[13] Hayes, B. E. (2014b). The what and where of big data: a data definition framework. *Customerthink*. Najdeno 5. 12. 2015 na http://customerthink.com/the-what-and-where-of-big-data--a-data-definition-framework/.

[14] Hayes, B. E. (2015a). Optimizing your data science team, a survey of data professionals. *Analytics Week*. Najdeno 5. 12. 2015 na https://analyticsweek.com/docs/research/open/OptimizingYourDataScienceTeamsV2.0.pdf.

[15] Kuhn, M. (2015). *CRAN* task view: Reproducible research. CRAN. Najdeno 5. 12. 2015 na https://cran.r-project.org/web/views/ReproducibleResearch.html.

[16] Leban, G. (2007). *Vizualizacija podatkov s strojnim učenjem*. Doktorska disertacija. Ljubljana: Fakulteta za računalništvo in informatiko.

[17] Lorica, B., Howard, J., Dumbill, E. (2012, 11. januar). What is big data. *O'Reilly*. Najdeno 21. 11. 2015 na https://beta.oreilly.com/ideas/what-is-big-data.

[18] Malhotra, N. K. (2012). *Basic marketing research: integration of social media* (4th ed.). New Jersey: Prentice Hall.

[19] Manyika, J., idr. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*. Najdeno 5. 12. 2015 na http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[20] O'Neill, C. in Schutt, R. (2013). *Doing data science*. United States: O'Reilly Media.

[21] OECD (2015). *Data driven innovation: Big Data for growth and well-being*. Paris: OECD Publishing.

[22] Olofson, C. W. in Vesset, D. (2012). Big Data: Trends, Strategies, and SAP Technology. *SAP*. Najdeno 16. 1. 2015 na https://www.sap.com/bin/sapcom/en_ae/downloadasset.2012-09-sep-26-13.idc-report--big-data-trends-strategies-and-sap--technology-pdf.html.

[23] Piatetsky, G. (2014). Four main languages for Analytics, data mining, data science. *Kdnuggets*. Najdeno 21. 11. 2015 na http://www.kdnuggets.com/2014/08/four-main-languages--analytics-data-mining-data-science.html.

[24] Piatetsky, G. (2015). Poll results: Where is big data?. *Kdnuggets*. Najdeno 15. 8. 2015 na http://www.kdnuggets.com/2015/08/largest-dataset-analyzed-more-gigabytes-petabytes.html.

[25] Rivera, R. in Haverson, A. (2014). Data Scientist vs Data Analyst. *Captechconsulting*. Najdeno 15. 12. 2015 na https://www.captechconsulting.com/blogs/data-scientist-vs-data--analyst.

[26] Russom, P. (2011). Big Data Analytics. *Tableau*. Najdeno 21. 3. 2016 na http://www.tableau.com/sites/default/files/white-papers/tdwi_bpreport_q411_big_data_analytics_tableau.pdf.
[27] Somohano, C. (2013). Big data & data science: what does a data scientist do?. *Data Science London*. Najdeno 21. 11. 2015 na https://www.slideshare.net/datasciencelondon/big--data-sorry-data-science-what-does-a-data-scientist-do.
[28] Stanton, J. M. (2013). Introduction to data science. *iTunes*. Najdeno 21. 1. 2016 na https://itunes.apple.com/us/book/introduction-to-data-science/id529088127?mt=11.
[29] Swan, A. (2008). The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs. *Key Perspectives*. Najdeno 17. 1. 2015 na http://beta.jisc.ac.uk/media/documents/programmes/digital-repositories/data.
[30] Toš, N. in Hafner-Fink, M. (1998). *Metode družboslovnega raziskovanja*. Ljubljana: Fakulteta za družbene vede.
[31] *Ultimate skills checklist for your first data analyst job*. Najdeno 21. 11. 2015 na http://static.cdn.responsys.net/i2/responsysimages/content/udacity/Ultimate%20Skills%20Checklist%20For%20Your%20First%20Data%20Analyst%20Job.pdf.
[32] Vesset, D., idr. (2012). Wordlwide big data technology and services 2012 - 2016 Forecast. *IDC*. Najdeno 21. 11. 2015 na http://laser.inf.ethz.ch/2013/material/breitman/additional%20reading/Worldwide%20Big%20Data%20Technology%20and%20Services%202012-2016%20Forecast.pdf.
[33] Voulgaris, Z. (2014). *Data scientist: The definitive guide to becoming a data scientist*. United States: Technics Publications.

▪

Mateja Grobelnik je zaposlena v podjetju Petrol, d. d., kot analitik. Leta 2016 je dokončala znanstveni magisterij na Ekonomski fakulteti Univerze v Ljubljani, smer informacijsko-upravljalske vede. Področja, ki jo zanimajo in s katerimi se ukvarja, so poslovna inteligenca, baze podatkov, strojno učenje in statistika.

▪

Jurij Jaklič je redni profesor na Katedri za poslovno informatiko in logistiko na Ekonomski fakulteti Univerze v Ljubljani. Raziskovalno, pedagoško in svetovalno se ukvarja predvsem s področjem podatkovne analitike oz. poslovne inteligence, pa tudi z menedžmentom poslovnih procesov. Je (so)avtor več kot sto znanstvenih člankov in poročil projektov, številni so bili objavljeni v mednarodnih znanstvenih revijah. Sodeloval je v več raziskovalnih in svetovalnih projektih s področja poslovne inteligence, prenove poslovnih procesov in strateškega načrtovanja informatike.