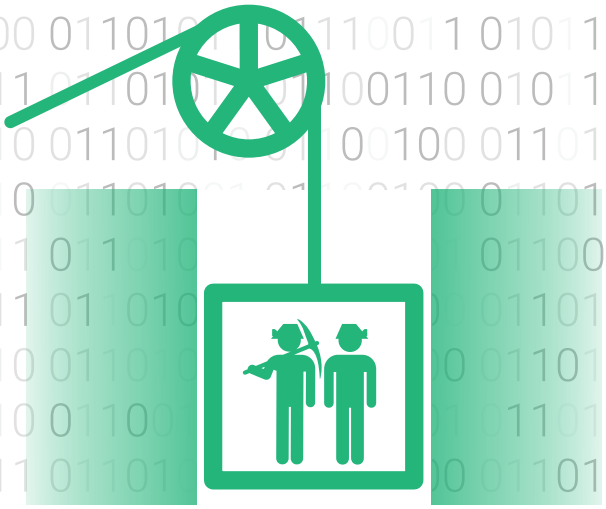


1 Uvod

Bez pravog alata nema zanata.

Tako barem kaže narodna mudrost očuvana u poslovici. Korisno bi bilo vještinu koju želite steći slušajući kolegij i čitajući ovaj tekst promatrati kao zanat.

Ova je knjga nastala iz dijela nastavnih materijala za kolegij Rudarenje podataka na Sveučilištu u Splitu, Prirodoslovno-matematičkom fakultetu. Iz tog se razloga ne treba čuditi što se tekstu često koristimo frazom rudarenje podataka kao ekvivalentnom engleskom izrazu *data mining*. Ako vam je prijevod na početku zazvučao grubo ili sirovo, bilo bi korisno da o rudarenju podataka razmišljate upravo kao o zanatu. Prvo, uvijek je korisno odabrati dobar alat. To ne možemo napraviti ako nismo upoznali taj alat i njegove mogućnosti. Svrha teorije je upravo razlikovati alate, primjerice odvijač od čekića i znati kada kada se koji upotrebljava, odnosno razlikovati kada problem treba snažno udariti po glavi, a kada zavrnuti. No, to razlikovanje nam ne će mnogo koristiti ako ne znamo rukovati alatom. Stoga ćemo posvetiti vrijeme upoznavanju s programskim jezikom R koji si možemo predočiti kao još jedan alat – ili još bolje kao još jednu radionicu s mnoštvom vlastitih, specifičnih alata. Ako želite izučiti ovaj zanat, onda – baš kao ni bilo koji drugi zanat – ne ćete ga izučiti slušajući neki kolegij ili čitajući neki

tekst – potrebno je i okušati se u njemu. Stoga ovaj tekst obiluje primjerima u R-u koji su upravo tu da ih iskušate, a ne da ih samo pročitate. To je ujedno i motivacija zbog koje gotovo svako poglavlje završava zadatcima za vježbu, odnosno smjernicama (a ne rješenjima) kako riješiti te zadatke.

Kolegij je namijenjen studentima matematičkih i informatičkih smjerova, a primarni programski jezik koji se koristi na kolegiju je R. Za R se često kaže da je programski jezik koji su statističari pisali za statističare i pritom se često želi istaknutni njegova percipirana nezgrapnost za izražavanje programskih koncepata, odnosno "dinamike" programa. S druge strane, statističiki alati koji se u R-u nude, kao i pripadajuća dokumentacija, upravo u području koje barata podatcima i informacijama te na njima uči ili zaključuje mogu biti iznimno korisni i studentima prvenstveno fokusiranima na programiranje otvoriti jedan novi svijet. Za razliku od drugih tzv. *point-and-click* alata – R statističarima može pružiti mogućnost skriptiranja niza naredbi i time jednostavniju ponovljivost (rekonstrukciju) pokusa. Te skripte u početku mogu biti shvaćene kao popratna dokumentacija, ali kako napreduje zbližavanje sa sintaksom R-a, prerastaju u vlastite programe ili pomoćne alate. Kada tome pridodamo kako je R smatran objektno orijetniranim jezikom, kojem ne nedostaju ni neka svojstva koja se još uvijek smatraju pomalo egzotičnim (kao što je funkcijsko programiranje[1]), možemo reći da se R unatoč nekim percipiranim nezgrapnostima uistinu doima kao dobar kandidat za kolegije koji imaju heterogenu populaciju studenata ili spajaju matematičke vještine s računalnim vještinama i teorijom o baratanju informacijama. Takvo spajanje vještina nije tek potencijalno korisno, već je nužno za inženjere u suvremenom svijetu gdje se već za samo razumijevanje problema, a pogotovo za njegovo rješavanje, zahtijeva razmišljanje na više apstrakcijskih razina, ali i sposobnost da se razumiju operacijski detalji problema, odnosno sustava.

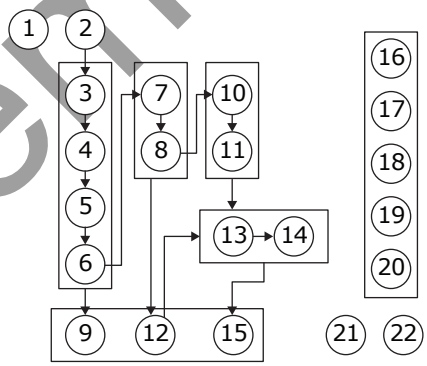
Nastanak knjige ima svoj evolucijski proces, a ova knjiga nije iznimka. Kako knjiga ima u sebi i izrazito praktičan dio, tokom procesa pisanja knjige doživjeli smo i razvoj jezika R koji je utjecao na taj praktičan dio. Naša je želja da ova knjiga bude usklađena i s novijim inačicama R-a, što su u ovom trenutku inačice 4.0 i 4.1. Iz tog su razloga pojedina poglavlja u knjizi doživjela promjene koje u pojedinim slučajevima nisu više kompatibilne sa starijim inačicama R-a. Na važnim mjestima smo te detalje istaknuli. Na drugim mjestima (primjerice gdje se bavimo pojedinim paketima, poput paketa `tidyverse`) nismo diskutirali sve promjene zbog velike količine promjena koje su vezane za razvoj pojedinog paketa. Sav je kôd testiran na najnovijim inačicama R-a, a uočite li kakve nedorečenosti ili neusklađenosti bilo zbog tog ili drugog razloga, bili bismo vam zahvalni kada biste nam javili.

[1] engl. *functional programming* – funkcijsko programiranje tretira računanje kao evaluaciju matematičkih funkcija

1.1. Kako čitati ovu knjigu?

Prva poglavlja knjige fokusiraju se na upoznavanje s pojmom *rudarenje podataka* kao i s motiviranjem studenata *zašto naučiti R*. Potom slijedi *upoznavanje s R-om* i njegovim mogućnostima te se naglasak pomalo prebacuje na alate za *pretraživanje* skupova podataka i *vizualizaciju* informacija [to je neizbježan dio pri upoznavanju s novim skupom podataka iz kojih želimo izrudariti (ekstrahirati) informacije.

Naravno, knjigu je moguće čitati upravo redoslijedom kako su poglavlja posložena. Zbog načina na koji je knjiga organizirana, može poslužiti i kao udžbenik za kolegije kojima je cilj upoznavanje s R-om, pogotovo ako je taj kolegij dio studija koji je fokusiran na podatkovno usmjerenu znanost[2]. Nadamo se da u tom slučaju mogu poslužiti i priloženi primjeri i zadatci za vježbu. Nažalost, nemamo uvijek vremena (ili strpljenja) čitati tekstove redoslijedom koji je autor zamislio, pogotovo ako ne pripadamo ciljanoj skupini za koju je tekst prvotno nastao. Stoga, u nadi da će ova knjiga biti duže u upotrebi no što je to prvotno bilo zamišljeno ili korisna studentima s različitim predznanjima, dajemo kratki pregled međuovisnosti pojedinih poglavlja (slika 1.1), odnosno preporuka gdje započeti s čitanjem.



Slika 1.1. Pregled poglavlja u knjizi.

Slika 1.1 daje uvid u strukturu knjige na način da su poglavlja predstavljena kružićima s brojevima, gdje broj označava poglavlje o kojem je riječ. Pravokutnici na slici uokviruju nekoliko poglavlja označavajući tako veće cjeline. Strelice označavaju poglavlja (ili cjeline) koje bi trebalo pročitati prije nego se započne s nekom drugom cjelinom ili

[2] engl. *data science*

poglavljem. Dakle, osim linearnog čitanja ove knjige možete odabrati krenuti od nekog drugog poglavlja te se po potrebi vratiti pokaže li se kasnije ova procjena neispravnom.

Započinjete li s kolegijem *Rudarenje podataka* ili razmišljate o njegovom upisivanju, nakon ovog poglavlja možete nastaviti na poglavlje 2.: "Ciljevi i zadatci rudarenja podataka" u kojem kratko opisujemo što je to *rudarenje podataka* te koji su mu glavni zadatci i ciljevi. Pri tome je važno naglasiti da ne govorimo o *rudarenju podataka* u užem smislu koji se često vezuje isključivo uz (relacijske) baze podataka, već u širem kontekstu[3] koji može uključivati podatke u raznim oblicima, pa bili oni nerelacijske baze podataka ili nestrukturirani podatci (tipa *Web*).

Ako se do sada niste upoznali s R-om, možete započeti s poglavljem 3., gdje dajemo primjere poziva naredbi, a zatim u poglavljima 4., 5. i 6. opisujemo osnovne stukture podataka u R-u i način na koji s njima radimo. Riječ je o poglavljima "Upoznavanje s R-om", "Vektori, nizovi, faktori i tablice", "Indeksiranje, matrice i polja" i "Liste i okviri", a na slici 1.1 uokvireni su prvim pravokutnikom. Poznajete li R od ranije, ova poglavlja vjerojatno možete preskočiti.

Poglavlja 7. i 8. sadrže izbor korisnih naredbi koje se tiču pohrane i dohvata podataka, rada s datotekama i direktorija te rada s nizovima znakova (stringovima), datumima i vremenima, kao i naredbe za sortiranje i iscrtavanje podataka. Ova poglavlja pripremaju teren za izviđanje podataka, odnosno različite biblioteke za prikaz podataka, a što je obrađeno u poglavljima 10. i 11.. Stoga bi ova poglavlja mogla biti dobra polazišna točka za studente koji dolaze s minimalnim predznanjem R-a.

Tri poglavlja koja se nalaze u velikom pravokutniku položenom na dnu slike 1.1, nisu međusobno povezana strelicama jer ne čine narativnu cjelinu jesu poglavlja 9., 12. i 13.. To su poglavlja koja tek uvjetno možemo nazvati jednom cjelinom zato što predstavljaju jednu određenu dozu "modernog R-a". Tako se poglavlje "Funkcije, petlje i uvod u *F* programiranje" bavi funkcijama te daje odgovor zašto u R-u ne viđamo često petlje i uvodi neke osnovne elemente *F* programiranja. Poglavlje "Manipuliranje podatcima i `tidyverse`" opisuje niz biblioteka koje R-u pružaju alternativne načine rada s podatcima koji su zbog raznih razloga postali popularni u zajednici osoba koje se koriste jezikom R i bave podatkovnom znanošću. S obzirom na to da su biblioteke iz ovog poglavlja predstavljene kao alternativan način rada u odnosu na "klasičan R", od čitatelja se pretpostavlja određeno predzanje o R-u, na što bi trebala ukazivati strelica povučena od poglavlja 7. i 8. prema 12.. Što se tiče poglavlja 15.: "Izvještavanje kodom", ono nam kratko opisuje kako uz malu pomoć R-a možemo istovremeno s kodom pisati dokumentaciju za taj kod, na takav način da ona postane ili pisani ili elektronički izvještaj (čak

[3] Po uzoru na npr. "Introduction to Data Mining", autora Tana, Steinbacka i Kumara.

i u bliku mrežne stranice). Premda je ovo poglavlje ostavljeno gotovo za sam kraj – da bi se nadovezalo na interaktivne grafove i metode vizualizacije podataka – sigurno će u praktičnoj nastavi biti korisno uvede li se čim prije.

Poglavlja 13. i 14. predstavljaju dodatne alate za rad s grafovima, predvođene u prvom redu `ggplot2` bibliotekom, te stoga možemo reći da ova dva poglavlja čine zasebnu cjelinu. Ova dva poglavlja predstavljaju određeni skok u razmišljanju (i programiranju) u odnosu na poglavlja 11. i 12., a taj je skok usporediv s prijelazom od klasičnog (osnovnog) R-a prema njegovom modernijem dijelu u poglavlju 12.. Stoga smo na slici 1.1 strelicama povezali ova i poglavlje 12. kao preporučeno za čitanje prije poglavlja 14. i 15..

Vladate li R-om, a prikaz podataka i izvještavanje želite preskočiti, možda možete početi od poglavlja 16. i nastaviti do poglavlja 20.. Ta su poglavlja izdvojena u zasebnu cjelinu jer se više bave algoritmima koji uče različite probleme. U poglavlju 16. započinjemo sa stablima da bismo naučili o razvrstavanju i detekciji. Tu ćete upoznati različite mjere kvalitete razvrstavanja, odnosno koliko je mjera kvalitete važna za rješavanje nekog problema. U ovom se poglavlju također upoznajemo s pojmovima kao što su evaluacija, generalizacija te skupovi za treniranje i skupovi za testiranje. U poglavlju 20.4. bavimo se grupiranjem, odnosno općenito utvrđivanjem sličnosti među podatcima te proširujemo dosadašnje znanje. Ovdje također raspravljamo o razlikama različitih paradigmi učenja, prvenstveno učenju pod nadzorom i učenju bez nadzora. U poglavlju 18. na primjeru KNN klasifikatora pokazujemo odabir značajki i problem prokletstva dimenzionalnosti. U poglavlju 19. primjenjujući linearnu regresiju objašnjavamo problem pretreniranja, a u poglavlju 20. govorimo o neuronskim mrežama, uključujući i smjernice za rad s dubinskim neuronskim mrežama.

Konačno, kako gotovo svako poglavlje sadrži zadatake za vježbu za koje u samom pogavlju nisu dana rješenja, pred kraj ponuđena su rješenja zadataka za vježbu grupirana po pogavljima u kojima su zadana. S ovakvom smo strukturom možda mogli reći kako je preduvjet za čitanje ovog poglavlja gotovo svako prethodno poglavlje ili da svako poglavlje zapravo ovisi o njemu, no to nismo naznačili, nadajući se da zadatci ipak nisu toliko teški, odnosno da će razdvojenost rješenja od zadataka pridonijeti većem broju pokušaja samostalnog rješavanja.

Nakon rješenja slijedi Zaključak u kojemu dajemo nekoliko savjeta za daljnji razvoj u ovom području, predlažući još nekoliko knjiga koje smatramo vrijednim. Na samom kraju nalazi se Kazalo pojmova koje bi trebalo poslužiti čitatelju u pronalaženju ključnih pojmova.

2 Ciljevi i zadatci rudarenja podataka

2.1. Što (ni)je rudarenje podataka?

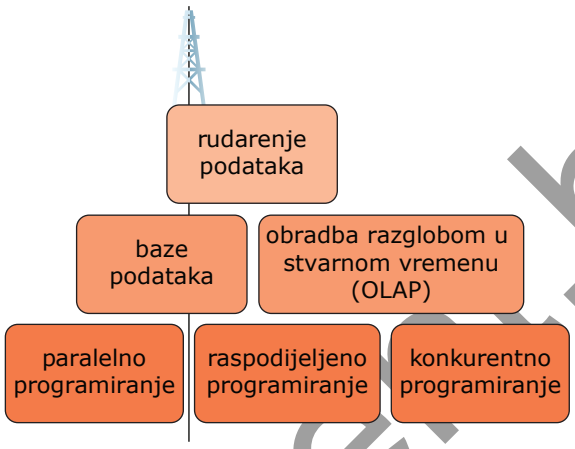
Prije nego li se pozabavimo definicijom rudarenja podatka, promotrimo različite pojmove s kojima se susrećemo kada zaronimo u područje koje danas najčešće nazivamo podatkovne znanosti:

- otkrivanje znanja u skupovima podataka
- raspoznavanje uzoraka
- statističko zaključivanje (učenje)
- strojno učenje
- računalna inteligencija
- umjetna inteligencija
- ...

Ovaj popis svakako nije iscrpan i postavlja se pitanje postoje li razlike i koje su među ovim područjima, a ponajviše kako se oni odnose prema rudarenju podataka. Moguće je

baviti se ovakvom vrstom sistematizacije, ali upitna je njena praktičnost - što će možda biti jasnije uskoro.

Rudarenje podataka možemo definirati kao proces automatiziranog dobivanja korisnih informacija iz (velikog) skupa podataka.[1] Ako se priklonimo takvoj definiciji, rudarenje podataka možemo si predočiti kao što je dano slikom 2.1.

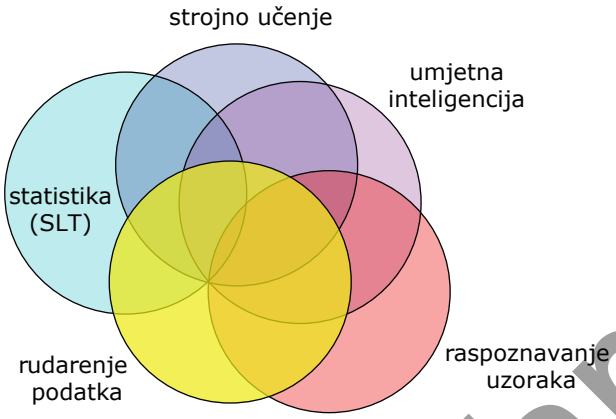


Slika 2.1. Rudarenje podataka može se smatrati procesom dobivanja dubinskih zaključaka koji nastaju korištenjem novijih tehnologija i znanstvenih istraživanja.

Kao alternativna definicija može poslužiti definicija područja gdje se rudarenje podataka definira kao interdisciplinarno područje koje upotrebljava računalne tehnike za raspoznavanje uzoraka (otkrivanje znanja) upotrebljavajući tehnike na granicama područja umjetne inteligencije, strojnog učenja, statistike, baza podataka itd.[2] Takav nam pristup omogućava da si rudarenje podatka predočimo kao što je dano slikom 2.2.

Rudarenje podataka uistinu možemo promatrati kao strogo tehnički proces u kojem se podatci uistinu često uspoređuju s naftom, gdje se rudarenjem i obradbom tih podataka značajno povećava vrijednost sirovine, ali takav pristup je u mnogočemu ograničavajući – što se onda nužno odražava i na skučenije razmišljanje osoba koje se bave time. S druge strane, promatrajući rudarenje podataka kao područje, lako se možemo zapitati je li moguće da se osoba bavi rudarenjem podataka, a da da se ne unese (ili se barem dotakne) i u ranije navedena područja.

[1] Tan, P.-N., Steinbach, M., Kumar, V.: *Intoduction to data mining*
[2] usp. Wikipedija



Slika 2.2. Rudarenje podataka može se smatrati interdisciplinarnim područjem koje ima dodirnih točaka s mnogim znanstvenim područjima.

Sve dobre stvari, a posebno one popularne imaju mnogo naziva. Često je njihova važnost tolika da su ponovno otkrivane ili samo rebrandirane – i možda je najbolje kroz tu prizmu promatrati i ovo područje. Ako promotrimo terminologiju na engleskom jeziku, lako zaključimo kako je ona još bogatija no terminologija na hrvatskom jeziku: *Database mining*, *Data Archaeology*, *Data Forensic*[3], *Information Harvesting*, *Information Discovery*, *Knowledge Extraction*, *Knowledge Discovery in Databases* (KDD), *Predictive Analytics*, *Machine Learning*, *Statistical Learning Theory* (SLT), *Data Science*...

Osobu koja se bavi rudarenjem podataka, nazvali bismo rudar. Osobu koja se bavi forenzikom podataka – forenzičar. A osobu koja se bavi (prediktivnom) analitikom, nazvali bismo analitičarom. No, ovakvi nazivi često ne opisuju dovoljno dobro posao kojim se netko bavi (rudar, forenzičar, analitičar?) ili nije dovoljno *sexy*[4] za posao koji jest toliko tražen i toliko popularan.[5] Stoga ne čudi potreba za rebrandiranjem. S druge strane, nazivi se često preklapaju, ovisno o području ili kulturi kojima pripadamo, odnosno ovisno o području u koje uranjamo. U Hrvatskoj je tako za poslove ovog tipa često rabljen i naziv "konzultant" što se izvan Hrvatske smatra "zastarjelim", odnosno naziv se smatra atavizmom koji ne opisuje dobro posao koji se obavlja – sposobnost analize, uočavanja,

[3] Što je blisko pojmu *Forensic Data Analysis* koji ubrajamo u digitalnu forenziku koja se smatra podpodručjem forenzike ili računalne sigurnosti.

[4] `http://spectrum.ieee.org/tech-talk/computing/it/is-data-scientist-the-sexiest-job-of-our-time`

[5] `http://www.forbes.com/sites/gregoryferenstein/2016/01/20/report-why-data-scientist-is-the-best-job-to-pursue-in-2016`

zaključivanja i izgradnje nečega na temelju podataka i modela.[6] Stoga ne čudi što je sve prisutniji još jedan novi naziv – "data scientist".

Nakon što smo rudarenje podataka definirali poprilično široko, možda je korisno dati koji primjer što rudarenje podataka nije.

To je dano u tablici.

Što NIJE rudarenje podataka?	Što JE rudarenje podataka?
dohvat imena ili telefonskog broja iz telefonskog imenika	podrijetlo prezimena u određenom području
pretraga svemrežja za pojam 'Split'	razvrstavanje rezultata pretrage na temelju toga pripadaju li pojmu grada 'Splita' ili filmu 'Split' M. N. Shyamalana ili engleskoj riječi 'split'

2.2. Proces rudarenja podataka

Rudarenje podataka može se ugrubo podijeliti u tri faze

- predprocesiranje[7]
- rudarenje (otkrivanje znanja)
- postprocesiranje.

Predprocesiranje se može sastojati od filtriranja podataka, odabira značajki, smanjenja dimenzionalnosti (ulaznog prostora), normalizacije podataka, **uzorkovanja** itd. Filtriranjem nazivamo odbacivanje podataka zbog toga što su podatci nepotpuni ili sadrže očita odstupanja. U procesu odabira značajki ključno je odabrati varijable koje nose značajnu informaciju, a to ponekad podrazumijeva i stvaranje novih varijabli te probiranje među njima. Smanjenje dimenzionalnosti se može smatrati posebnim načinom kombiniranja i odabira značajki, a često se primjenjuje kada imamo jako mnogo parametara. Kao jedna od motivacija za smanjenje dimenzionalnosti jest i tzv. **prokletstvo dimenzionalnosti**[8] koje se manifestira kao degradacija u performansama algoritma zbog velikog broja va-

[6] Detaljnije o tome na prethodnim poveznicama.

[7] engl. *preprocessing*, hrv. pre*d*procesiranje, odnosno postupak koji prethodi procesiranju. Preprocesiranje (dakle, bez *d*) označava nešto drugo (usp. pretreniranje).

[8] engl. *curse of dimensionality*