



Blok 1

Data Science / Uvod /
Data Wrangling

PRIPREMA PODATAKA

Data Preprocessing
Data Wrangling

DR BOBAN STOJANOVIĆ / DR MILOVAN MILIVOJEVIĆ



PRIRODNO-MATEMATIČKI FAKULTET
UNIVERZITET U KRAGUJEVCU

Literatura

1. Hastie, Trevor,, Robert Tibshirani, and J. H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009.
2. Pathak, Manas A., *Beginning Data Science with R*. Springer, 2014.
3. Schutt, Rachel, and Cathy O'Neil. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.
4. Wickham, Hadley, and Garrett Grolemund, *R za statističku obradu podataka*, Mikro knjiga, 2017.
5.

Zašto Data Science?

- Treći talas (Third wave), Alfin Toffler, (Izdanje: 1983, Jugoslavija, Beograd, Biblioteka ZENIT / originalno izdanje 1980.).
- Poslovna inteligencija / Business Intelligence
- Kvantitativni poslovni modeli / Kvanti? (Quantitative analysts)
- Znanjem vođene fabrike / Data Driven Factory
- Industrijska ekonomija / Ekonomija znanja (Dividenda znanja) / Kvantna Ekonomija (Izvor: R. Tissen, D. Andriesen. F. L. Deprez, The Knowledge Dividend, Pearson Education Limited, 1999)
- Veštačka inteligencija (Artificial intelligence) (RUSSEL, S. , NORVIG, P., Artificial Intelligence: A modern Approach, 3rd edition, New York, Pearson Education, 2010)
- Socijalne mreže / Soft Power / Izbori (Amerika), rat u Iraku, Afričko proleće?, ...!!!
- Big brother / Džordž Orvel !..
- Razvoj senzorske i merne tehnike / Datifikacija
- Nove tehnologije: / GPS, Spatial data, Visualization...
- MB → GB → TB → PB... (Kako od šume videti drveće?)

- VER. 01

Motivacija



- Svet podataka / nova paradigma?
- Master of the universe?
- Rešavanje nesvakidašnjih problema okarakterisanih pojmom *Big Data*
- Puzzle Solving
- Data Science / Data Scientist
Razvijanje sposobnosti kreativnog i kritičkog razmišljanja i rešavanja problema (critical thinking / creative problem solving)
- Radoznalost, skepticizam, intuicija (EDA), Bell Labs , John Tukey
- Umeće postavljanja pravih pitanja



- Big Data
- Data Science Hype
- Datafication (opodatkovljenje?)
- Modeliranje
- Fitovanje
- Statističko zaključivanje
- State of the Art
- Populacija i uzorak ($N=ALL?$, $n=1?$)

Odgovornosti Data Scientista

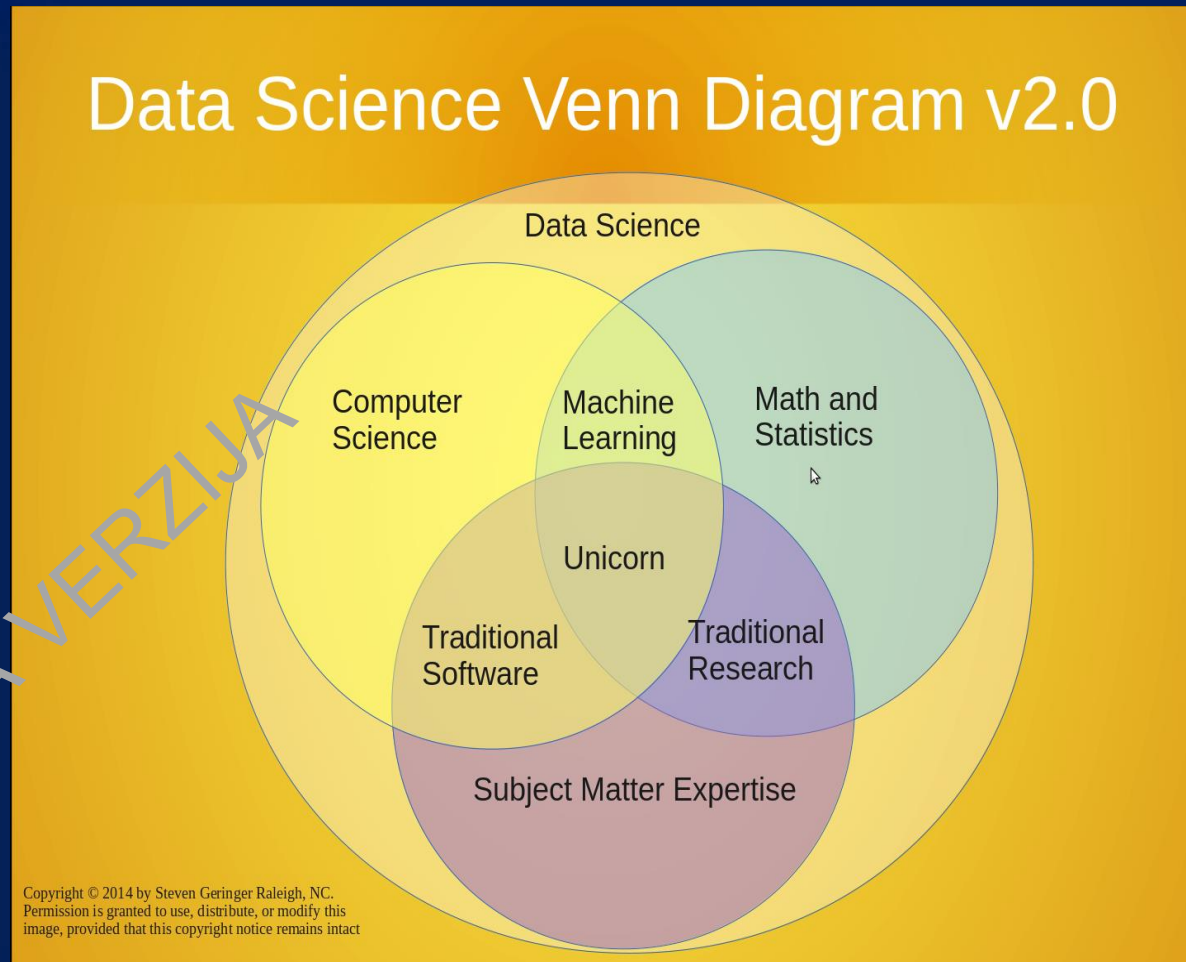
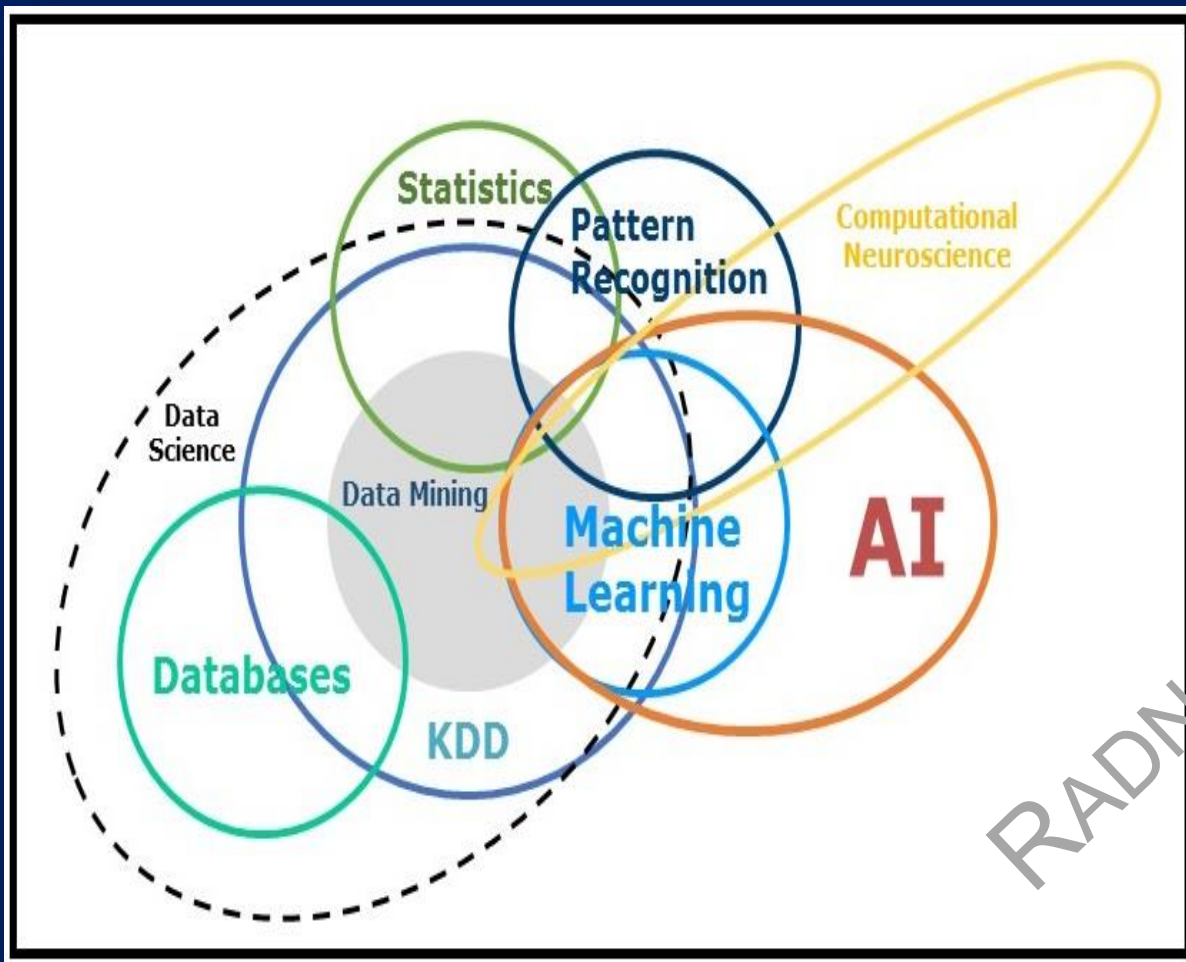


Data Scientists are not value-neutral:

- Izbor pretpostavki
- Izbor metrike
- Kreiranje modela i algoritama
- Implementacija modela u realnom okruženju
- LLL!
- Dodatna literatura!

	Baze Podataka	Nauka u podacima
Vrednost podataka	“Dragoceni”	“Jeftini”
Obim podataka	Skroman	Ogroman
Primeri	Bankovni podaci, Podaci o zaposlenima, Podaci o popisu, Medicinski kartoni...	Online klikovi, GPS logovi, Tweetovi, Očitavanja senzora en masse...
Prioriteti	Konzistentnost, Oporavak od grešaka, Proverljivost.	Brzina, Dostupnost, Bogatstvo upita.
Strukturiranost	Jaka (Šema)	Slaba ili nikakva (Tekst)
Svojstva	Transakcije, ACID* (ACID = Atomicity, Consistency, Isolation and Durability)	CAP* teorema (2/3), Eventualna konzistentnost (CAP = Consistency, Availability, Partition Tolerance)
Realizacija u:	SQL	SQL i NoSQL: Apache River, MongoDB, CouchDB, Hbase, Cassandra,...
Upiti:	ISPITIVANJE PROŠLOSTI	ISPITIVANJE BUDUĆNOSTI

DATA SCIENCE – INTERDISCIPLINARNOST



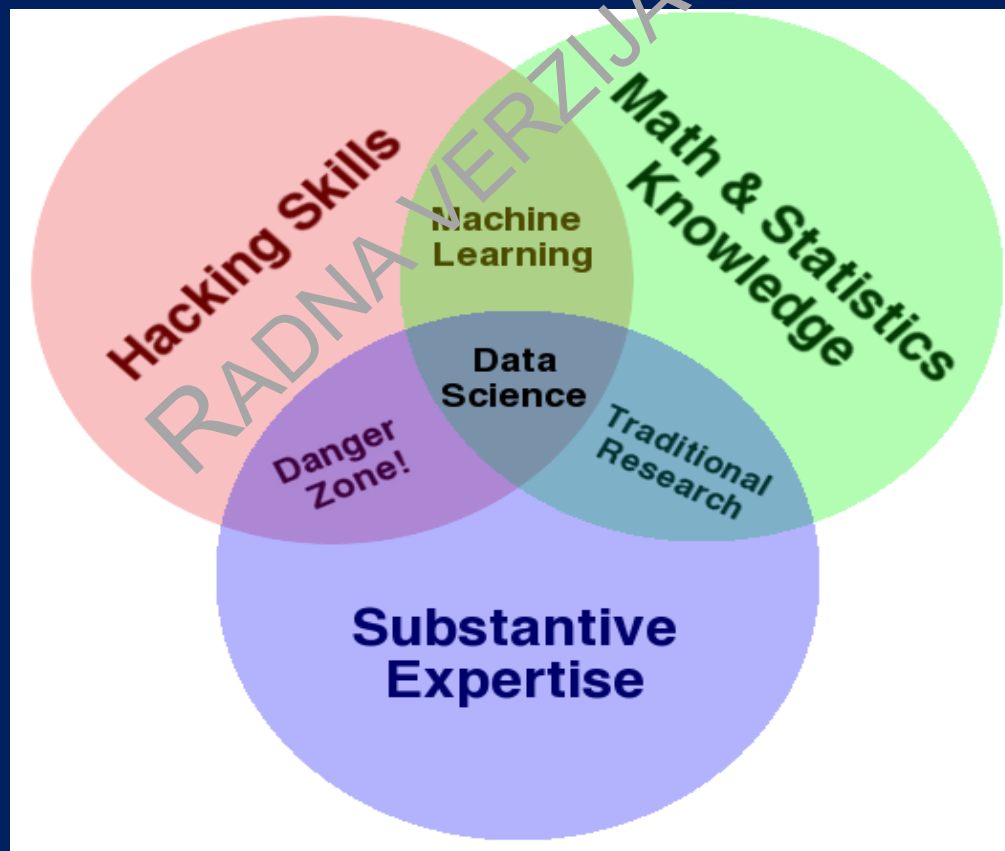
DATA SCIENTIST – „NAUČNIK PODATAKA”

A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock

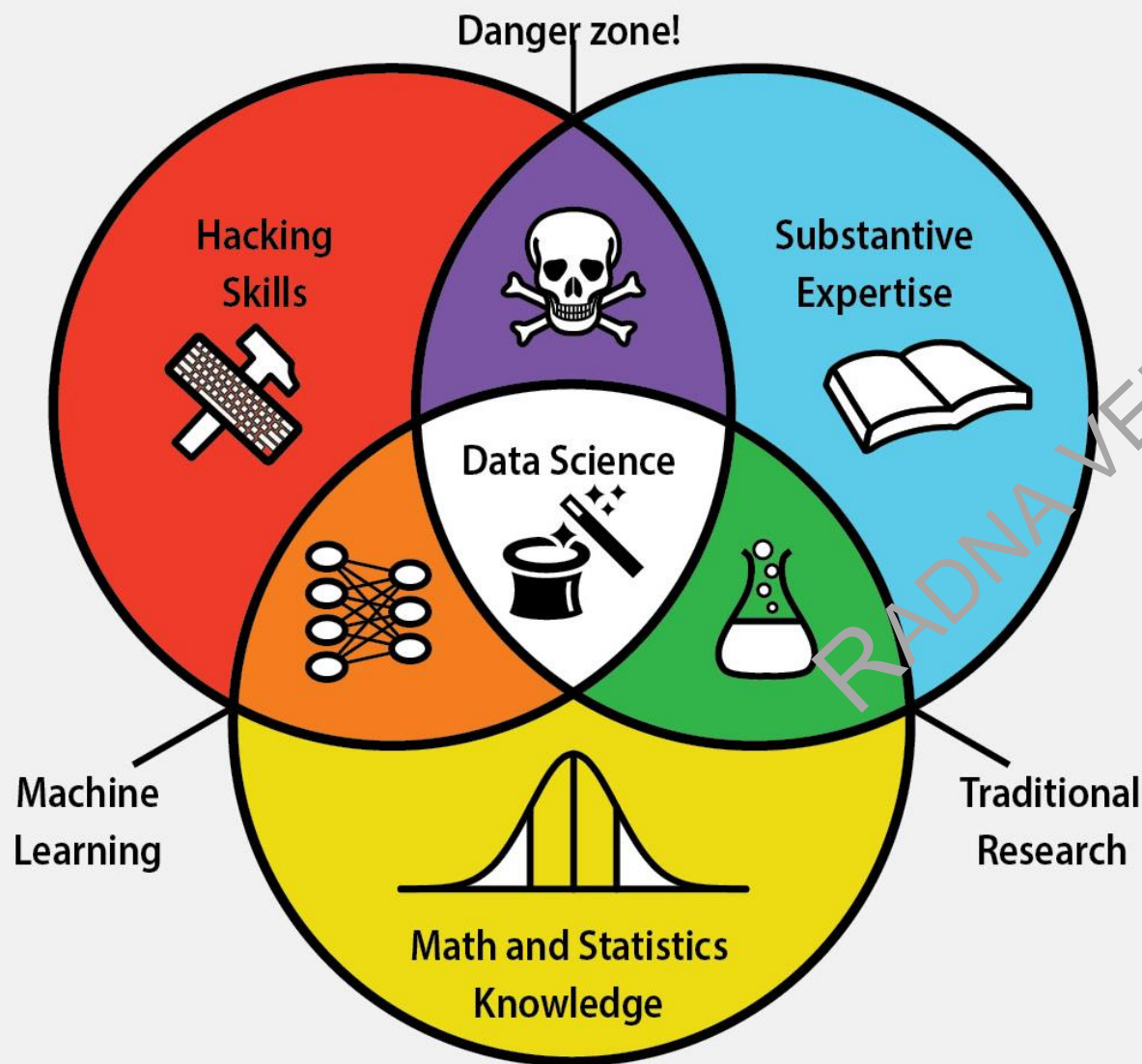
“Data Scientist = statistician + programmer + coach + storyteller + artist”

- Shlomo Aragon



DATA SCIENCE SKILLSET

TO
VER. 01



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills, math and statistics knowledge**, and **substantive expertise** in a field of science.



Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.



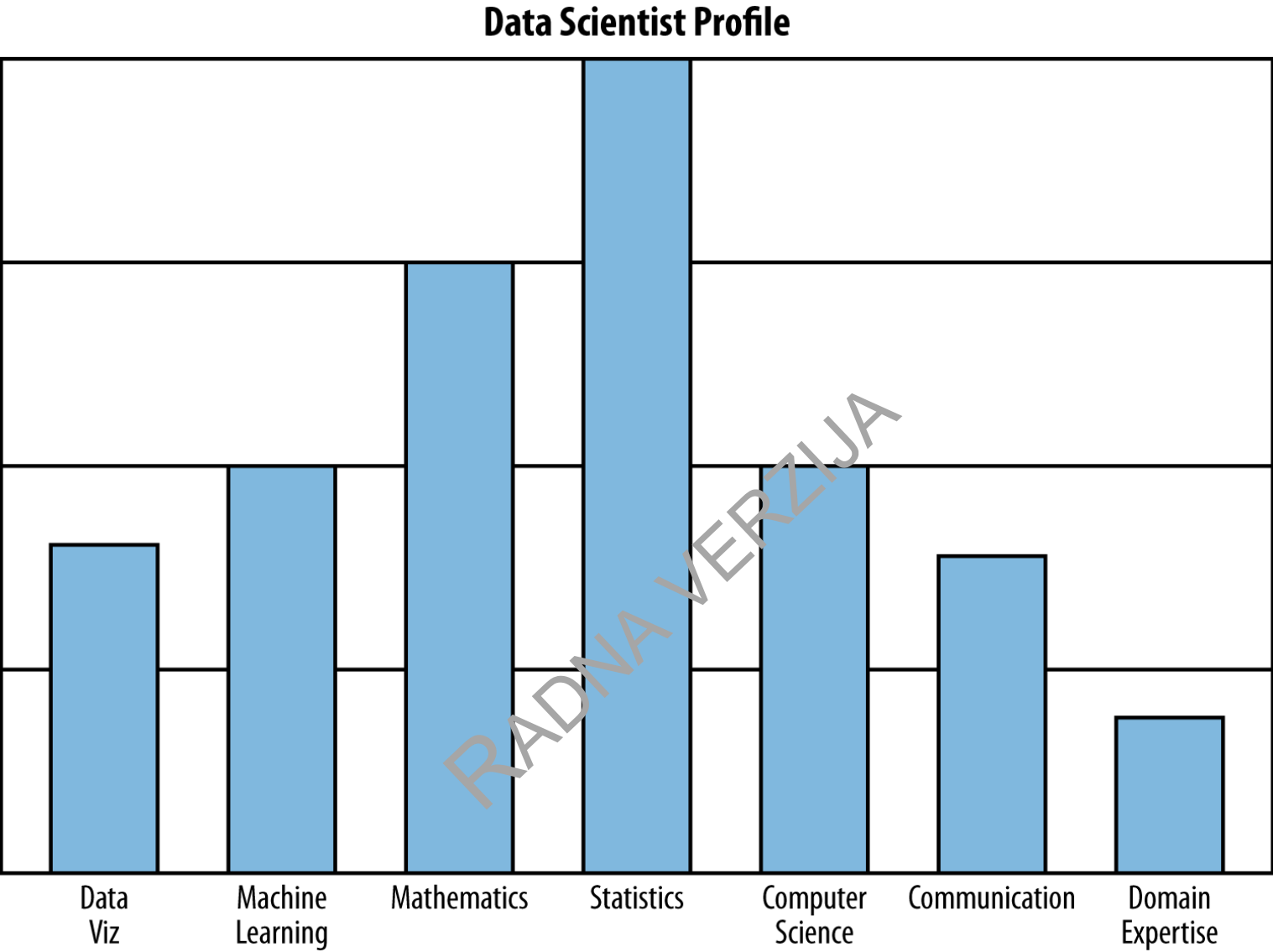
Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

DATA SCIENTIST – „NAUČNIK PODATAKA”

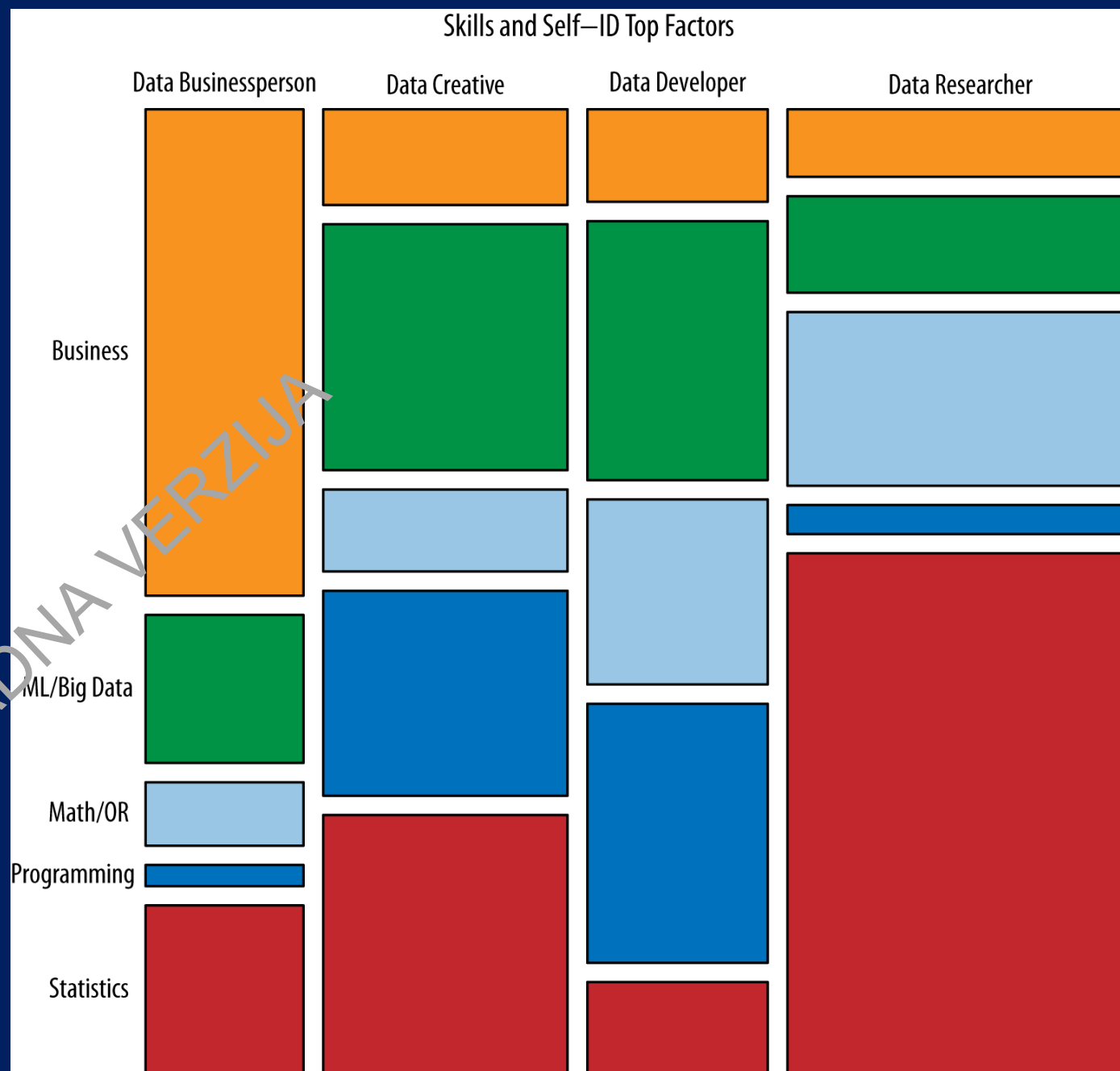
VER. 01
VER. 01



Izvor:
Doing Data Science
by Rachel Schutt and Cathy
O’Neil, 2014.
Published by O’Reilly Media,
Inc., 1005 Gravenstein Highway
North, Sebastopol, CA 95472.

Harlan Harris's clustering and visualization of subfields of data science from Analyzing the Analyzers (O'Reilly)

Izvor:
Doing Data Science
by Rachel Schutt and Cathy
O'Neil, 2014.
Published by O'Reilly Media,
Inc., 1005 Gravenstein Highway
North, Sebastopol, CA 95472.



VER. 01

- ☆ Neko ko poseduje pristup da kodira statističke i manipuliše na razne načine raznovrsnih izvora podataka
- ☆ Neko ko zna da postavlja razuman i kompletan model
- ☆ Nakon izvođenja analize

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Which makes them hard to find...

Applied Science

- Statistics, applied math
- Machine Learning
- Tools: Python, R, SAS

Business Analysis

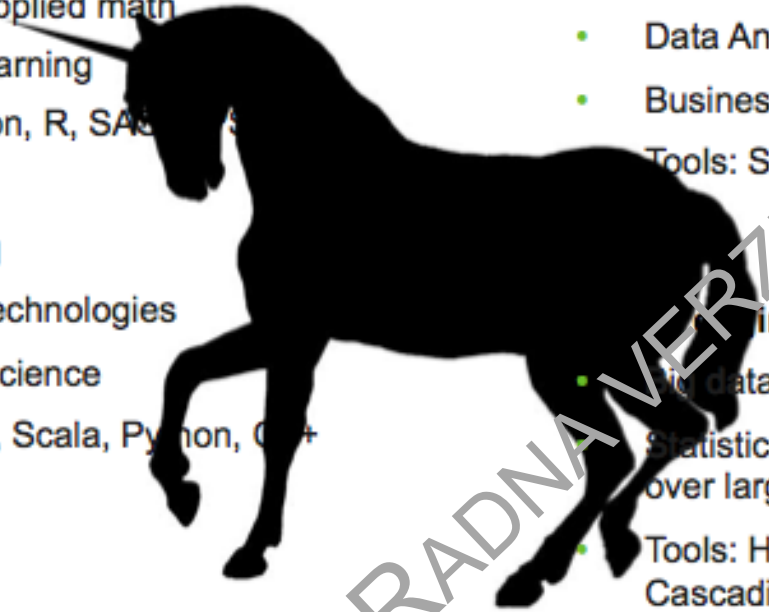
- Data Analysis, BI
- Business/domain expertise
- Tools: SQL, Excel, EDW

Data engineering

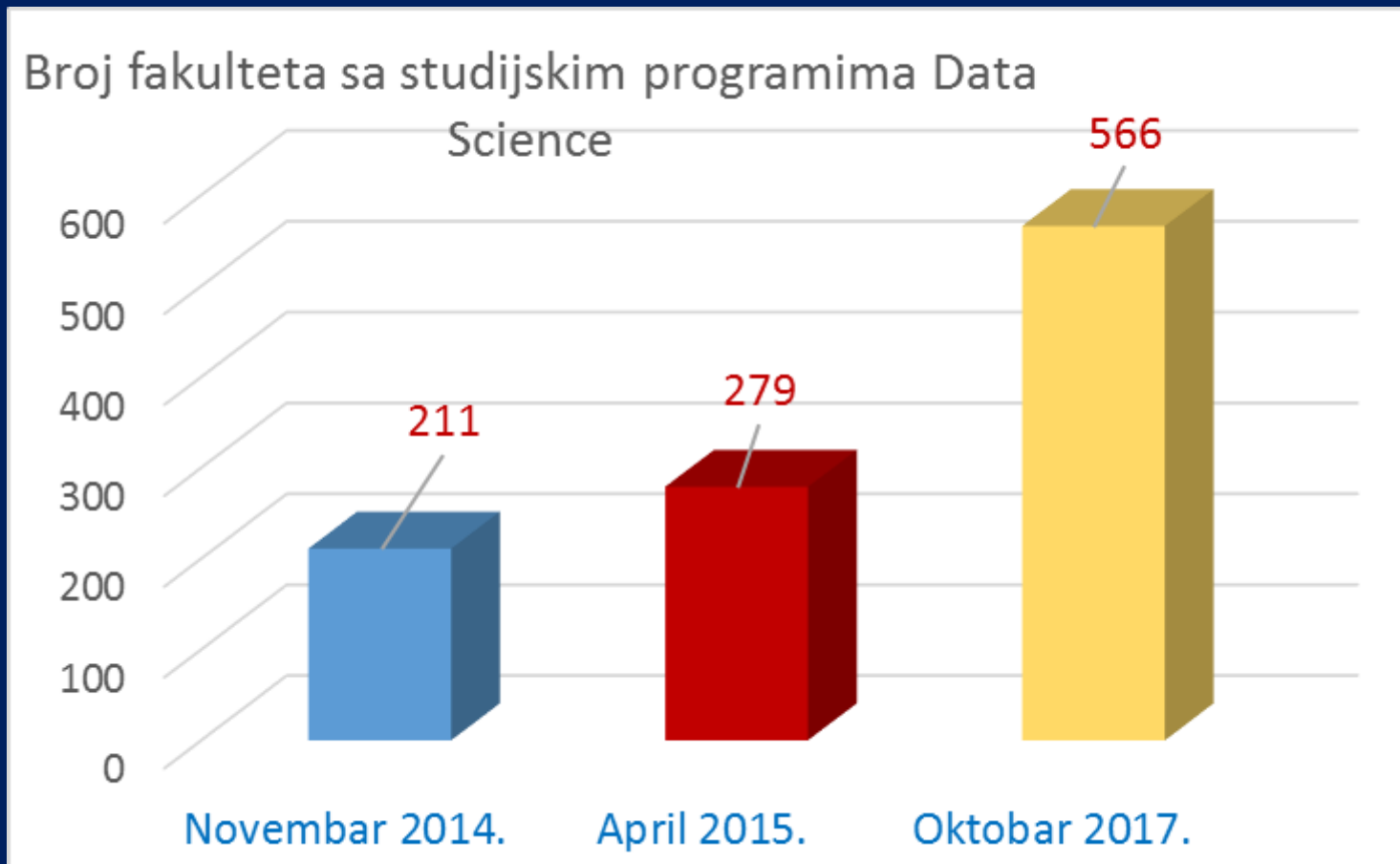
- Database technologies
- Computer science
- Tools: Java, Scala, Python, C++

Big data engineering

- Big data pipeline engineering
- Statistics and machine learning over large datasets
- Tools: Hadoop, PIG, HIVE, Cascading, SOLR, etc







<http://datascience.community/colleges>



<https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html>

IBM Predicts Demand For Data Scientists Will Soar 28% By 2020



Louis Columbus, CONTRIBUTOR

[FULL BIO](#) ✓

Opinions expressed by Forbes Contributors are their own.

- Jobs requiring machine learning skills are paying an average of \$114,000. Advertised data scientist jobs pay an average of \$105,000 and advertised data engineering jobs pay an average of \$117,000.
- 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.
- By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM.
- Data Science and Analytics (DSA) jobs remain open an average of 45 days, five days longer than the market average.

<https://www.forbes.com/sites/louiscolumnbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#456466c07e3b>

THE DATA
SCIENCE /
ANALYTICS
LANDSCAPE



2,350,000

DSA job listings in 2015

By 2020, DSA job openings
are projected to grow

15%

364,000

Additional job listings
projected in 2020

Demand for both Data
Scientists and Data Engineers
is projected to grow

39%

DSA jobs remain open

5 days

longer than average

DSA jobs advertise average salaries of

\$80,265

With a premium over all BA+ jobs of

\$8,736

81%

Of DSA jobs require workers with
3-5 years of experience or more

<https://www.ibm.com/analytics/us/en/technology/data-science/quant-crunch.html>

Secure | <https://www.forbes.com/sites/learnvest/2017/08/29/these-tech-jobs-are-in-demand-and-pay-well-heres-what-they-do/#c253f7974295>

Forbes LOG IN

0:29 / 1:19

YOUR READING LIST

 These Tech Jobs Are In-Demand And Pay Well. Here's What They Do

 PODCAST: Yes, You Are An Expert And It's Time To Own It

Active on Twitter

 How Equifax, Experian And Trans Union 'Print Money' By Selling Consumers Their Own Data

Active on Facebook

 Modi Lifts India Up

But as an admittedly non-techie person, I couldn't help but wonder: What is it that these people actually do?

Titles like "data scientist" or "cloud engineer" don't mean much to people outside the field. Luckily, I didn't have to go far to find the answers. I took a look at five jobs making a frequent appearance on these lists, then asked members of our own tech teams at LearnVest to help explain, in layman's terms, what they entail — as well as offer some advice for those interested in pursuing them. Here's what I found out.

1. Data Scientist

Indeed.com Average Base Salary: \$129,938

What They Do: Data scientist Michael Walker says a broad definition of his role is someone who takes raw data, cleans it up and analyzes it, with the intention of trying to glean useful information or predict an outcome.

"Data in the original format is somehow always terrible," Walker says. "You'd think that because computers generate this, it should be very regular. But all sorts of things can happen. You need to understand how the data was created in order to build a data model that's going to be reliable." Data scientists can be used for everything from marketing to product development to operational efficiency. In Walker's role specifically, he analyzes user data to help develop and improve future products and services.

Pro Tip for Job Seekers: Although data scientists typically have a mix of computer science, math and statistics backgrounds, it's common to see people with other science backgrounds enter the field, particularly those with Ph.D.s. "In most scientific fields, they end up doing statistics and data analysis," Walker says. "They're studying something else, but they just happen to be using data and statistics to understand that, so they have a good understanding of [data science] as well."

<https://www.forbes.com/sites/learnvest/2017/08/29/these-tech-jobs-are-in-demand-and-pay-well-heres-what-they-do/#7e44b8167429>

Engineering, u sklopu kog će studenti moći da biraju između više opcija, od kojih je jedna i Data Science.

Pored toga, Prirodno matematički fakultet u Novom Sadu je u septembru 2016. godine akreditovao dvogodišnji master modul "Nauka o podacima", u okviru studijskog programa Primenjena matematika.

76% kompanija radi ili će raditi Data Science

Većina aktera u domaćem IT-u, među kojima je i Startit, uvek naglašava važnost Data Science-a i to da je poslednjih nekoliko godina, ovo jedna od najplaćenijih oblasti u svim razvijenim ekonomijama.

Sa ciljem provere hipoteze o važnosti kreiranja studijskih programa iz oblasti nauke o podacima, u novembru 2016. godine za potrebe rada realizovano je istraživanje trenutnog stanja i potreba softverskih kompanija u Srbiji za znanjima i stručnjacima iz ove oblasti.

Istraživanje je zasnovano na prikupljanju podataka putem anonimnog web upitnika koji su popunjavali menadžeri u domaćim softverskim kompanijama.

Najveći broj kompanija ima potrebe za programerima „opšte namene“, sposobnim da učestvuju u raznorodnim IT projektima (70,6%), dok su **potrebe za data science inženjerima** nešto manje (**32,4%**).

Apply for a job position

data scientist



WHY work WITH Telenor?

„Work in one of the most exciting industries in a dynamic environment. Use your data scientist skills to discover hidden patterns of customer behavior. Effects of your work will touch millions of customers every day.“

Nikola Lalević
CLM and Advanced Analytics Manager



learning



community



empowerment



impact



progress



flexibility



THINGS you will be doing

- Developing prediction models;
- Customer profiling and description;
- Suggesting new ways of looking at customer behaviours;
- Data mining process management;
- Deployment, continuous evaluation and



SKILLS you will bring

- University degree in Mathematics, Organizational science or similar study;
- Minimum 2 years of experience in relevant field;
- Fluency in English;
- Solid knowledge of some statistical tool preferred (eg. SPSS, SAS, Mat Lab, H2O);
- Experience in SQL or other programming languages (preferred R and/or Python);
- Cooperative and professional attitude;
- Analytical mindset;
- Result oriented and proactive;
- Ability to handle pressure and short deadlines.

- Suggesting new ways of looking at customer behaviours;
- Data mining process management;
- Deployment, continuous evaluation and documentation of developed models;
- Developing models to be used as a inputs for Campaigns Management System and linking output of the initiatives with this system;
- Using different tools and techniques to handle complex request /problem statements from business stakeholders.





Priprema podataka

Data Preprocessing • Data Wrangling

This course introduces methods for **three** key aspects of data science:

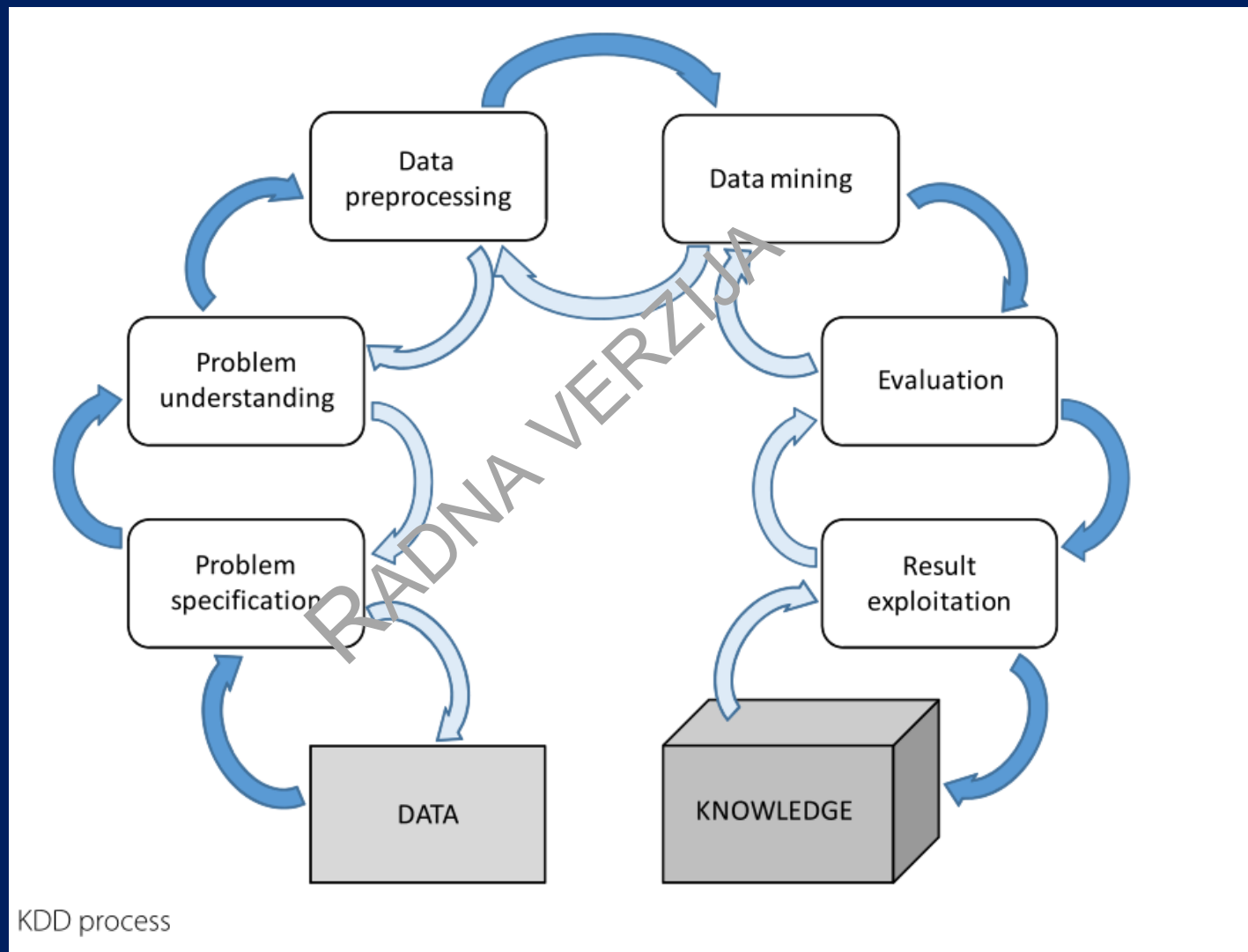
- **data wrangling, cleaning, and sampling;**
- **data management to be able to access big data quickly and reliably;**
- **exploratory data analysis to generate hypotheses and intuition;**
- prediction based on statistical methods such as regression and classification;
- Communication of results through visualization, stories, and summaries



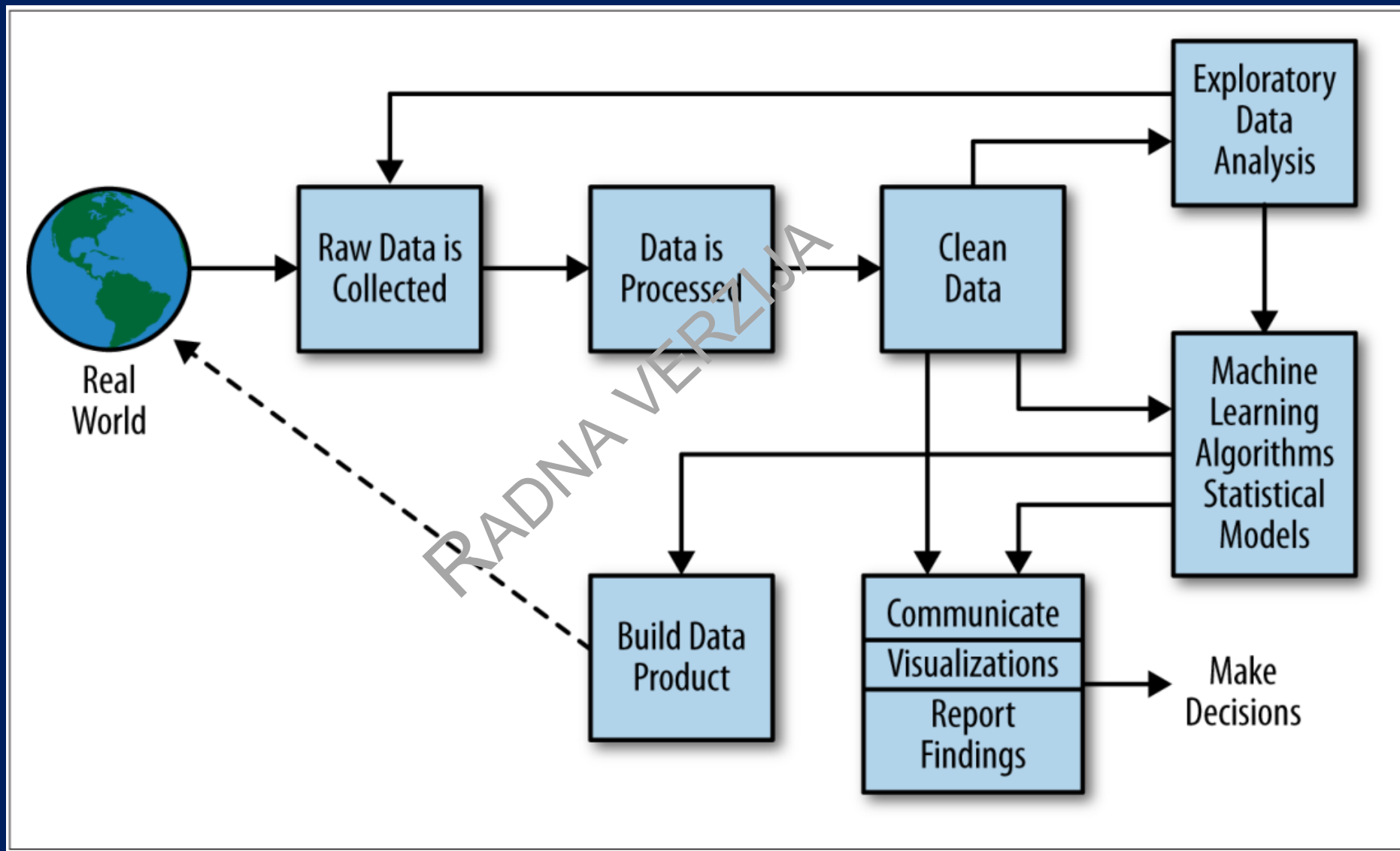
Syllabus

Syllabus

Proces otkrivanja znanja iz podataka



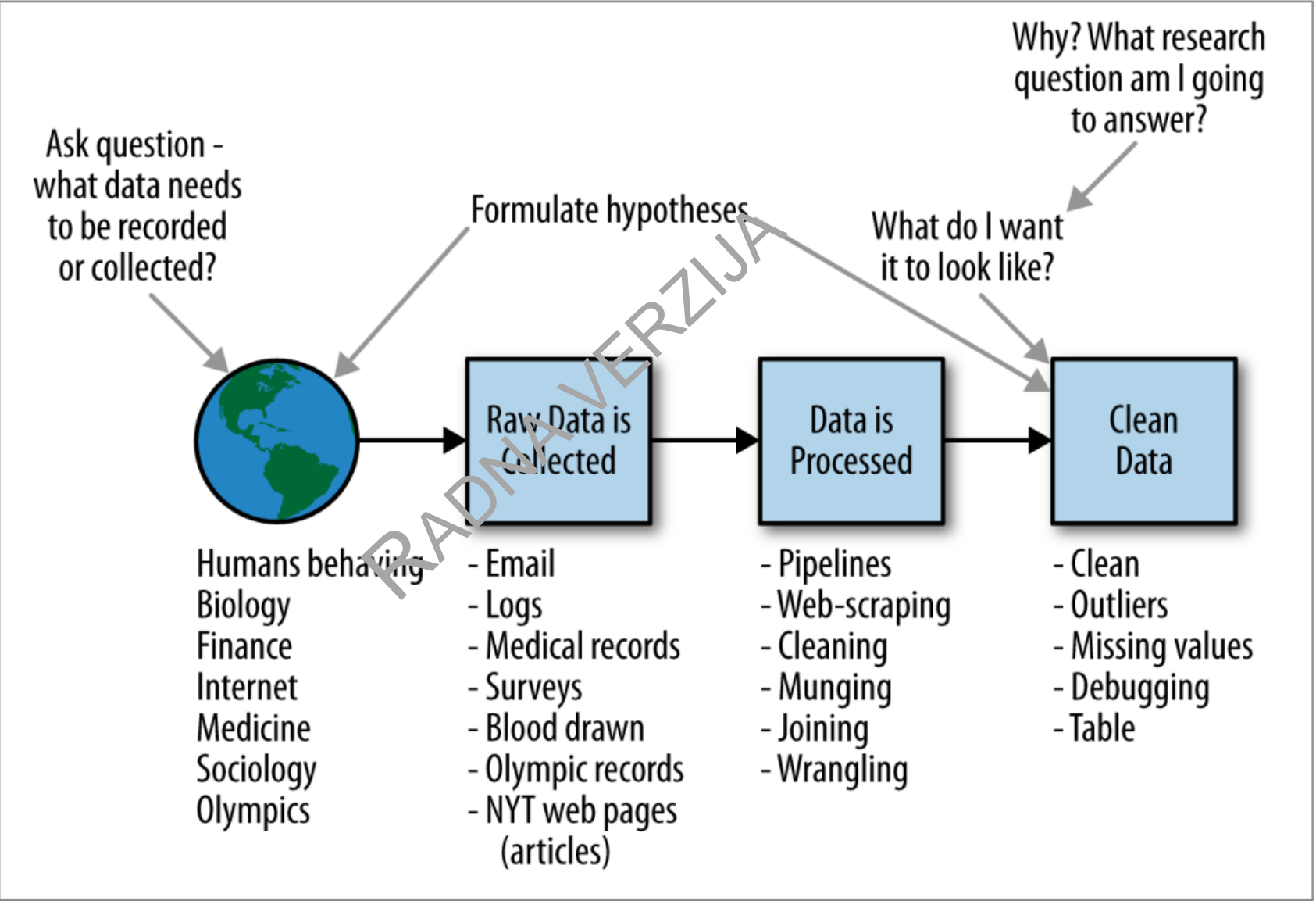
Data Science Process. Uopšten okvir aktivnosti



Izvor:
Doing Data
Science
by Rachel Schutt
and Cathy O'Neil,
2014.
Published by
O'Reilly Media, Inc.,
1005 Gravenstein
Highway North,
Sebastopol, CA
95472.

10
VER. 01
VER. 01

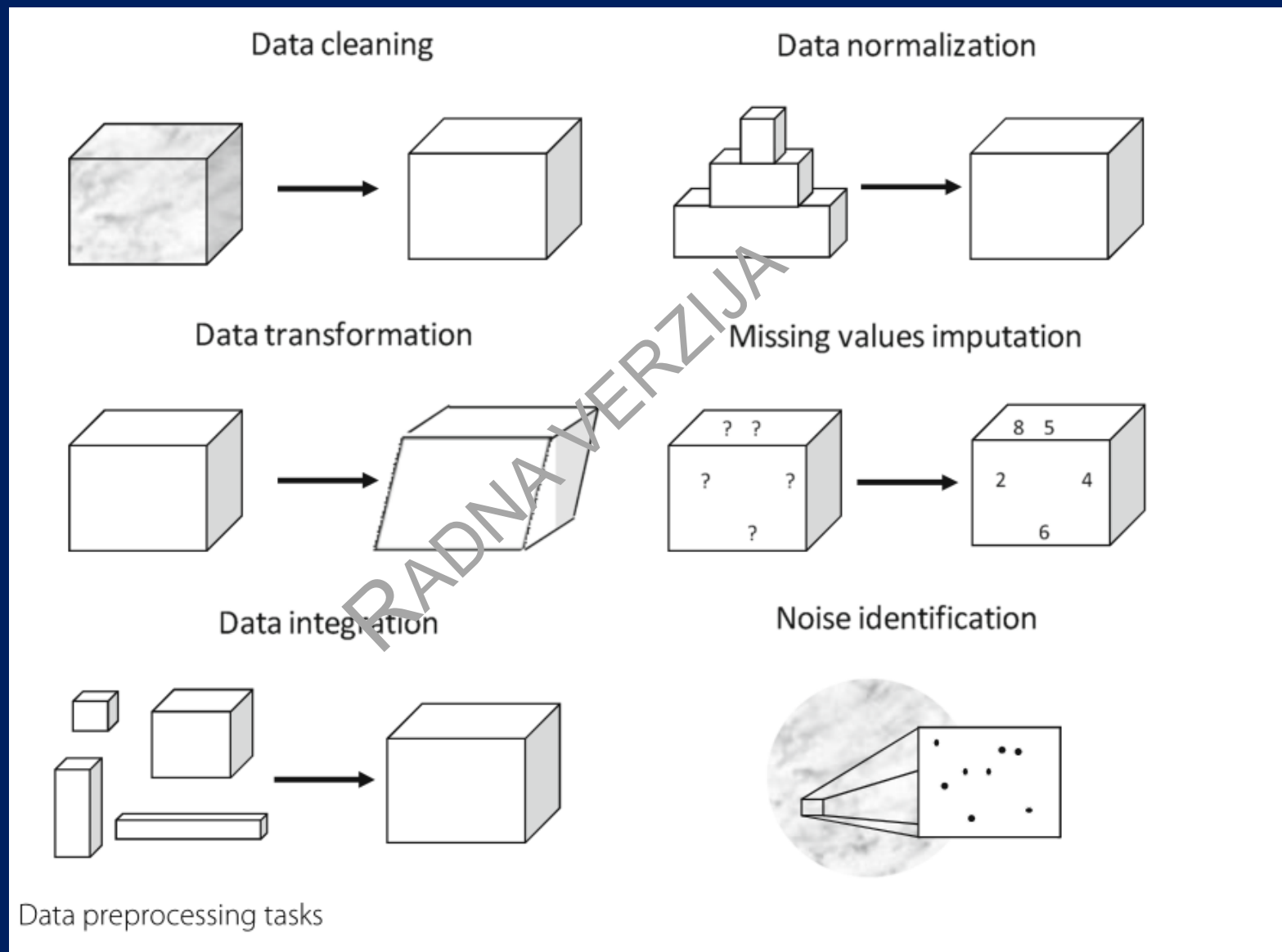
Data Science Process. Uloge Data Scientista.



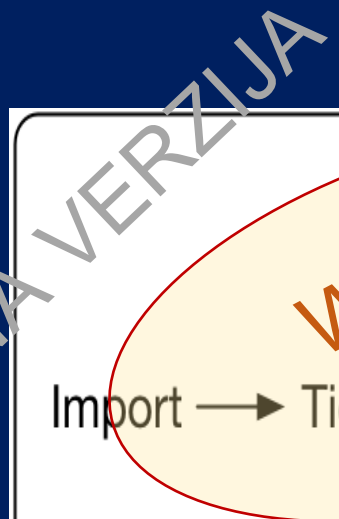
Izvor:
Doing Data
Science
by Rachel Schutt
and Cathy O'Neil,
2014.
Published by
O'Reilly Media, Inc.,
1005 Gravenstein
Highway North,
Sebastopol, CA
95472.

VER. 01
VER. 01

Predprocesiranje podataka



10 BBA
VER. 01





- Podaci se najpre moraju uvesti (u R) (Import). To obično podrazumeva preuzimanje podataka koji se čuvaju u datoteci, bazi podataka, ili web API-ju, i učitavanja tih podataka u data frame u R-u.
- Nakon što su podaci uvezeni, skoro je uvek dobra ideja srediti ih (Tidy). Sređivanje podataka podrazumeva skladištenje podataka u obliku koji je konzistentan sa semantikom dataseta.
- Ukratko, kada su podaci sređeni, svaka kolona je promenljiva, a svaki red je jedna opservacija.

- Transformacija uključuje promjenu interesa (npr. svi stanovnici određene godine), funkciji postojećih politika, ubrzanja na osnovu skupljanja skupa sumarnih stanja



- Zajedno, sređivanje i transformisanje podataka se u Data Science žargonu nazivaju natezanje/borba sa podacima (**Wrangling**), jer dobijanje podataka koji su u obliku sa kojim je prirodno raditi često podseća na dugu i komplikovanu raspravu/svađu/borbu.

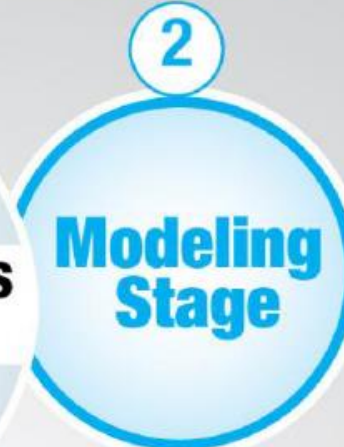
- Ono što (za)okružuje sve ove procese jeste **programiranje**. Programiranje je unakrsni alat koji se koristi u svakom delu Data Science projekta.
- Ne morate biti ekspert u programiranju da biste bili Data Scientist, ali se isplati biti bolji programer, jer će vam to omogućiti da automatizujete uobičajene zadatke i lakše rešite nove probleme.

>install.packages("package name")

1. Data Visualization ggplot2, googleVis
2. Data Transformation plyr, data.table
3. Missing Value Imputations MissForest, MissMDA
4. Outlier Detection Outliers, EVIR
5. Feature Selection Features, RRF
6. Dimension Reduction FactoMineR, CCP



1. General Model Validation LSMeans, Comparison
2. Regression Validation RegTest, ACD
3. Classification Validation BinomTools, DAIM
4. Clustering Validation ClustEval, SigClust
5. ROC Analysis PROC, TimeROC



1. Continuous Regression car, randomforest
2. Ordinal Regression RMiner, CoreLearn
3. Classification Caret, BigRF
4. Clustering CBA, RankCluster
5. Time Series forecast, LTSA
6. Survival survival, Basta

**Other Libraries**

A. Improve performance Rcpp, parallel

B. Work with web XML, jsonlite, httr

C. Report results shiny, RMarkdown

D. Text Mining tm, twitterR

E. Database sqldf, RODB, RMongo

F. Miscellaneous swirl, reshape2, qcc



Dobrodošli u svet podataka!