Master studije iz elektronskog poslovanja

Big Data u elektronskom poslovanju

DATA SCIENCE
- nauka o podacima -

Prof. dr Dragan Vukmirović

Fakultet organizacionih nauka, Univerzitet u Beogradu
Katedra za elektronsko poslovanje

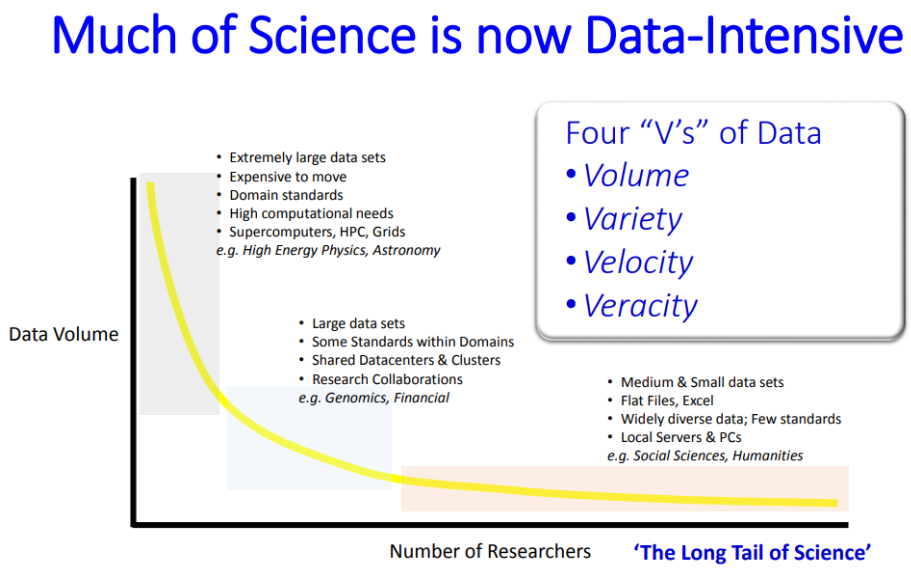
10.03.2021.

Moto predavanja

- Informaciona revolucija je toliko uzela maha da stručnjaci različitih profila, moraju da uče kako da **razumeju i koriste podatke**, kao što je prethodna generacija učila da koristi računare (‘80-ih i ‘90-ih)

(Pierson, 2017)

In 1990 data were scarce - interpretation was readily available
In 2013 data are everywhere - interpretation is scarce



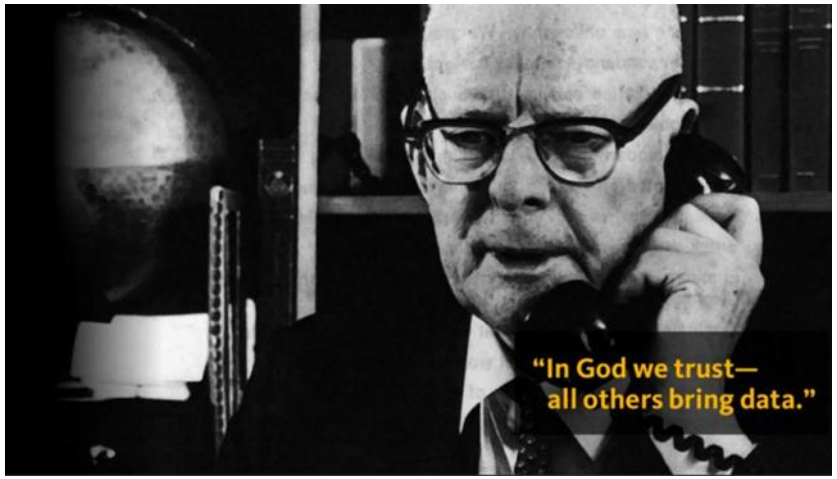
https://indico.cern.ch/event/609040/contributions/2455548/attachments/1468477/2271222/Tony_Hey_-_The_Fourth_Paradigm_ATTRACT_Talk_-_May_17.pdf

“Današnji gradovi i vlade i dalje funkcionišu u skladu sa principima koji su razvijeni pre dva veka, tokom industrijske revolucije.

Da bismo rešili probleme 21. veka kao što su eksplodirajući rast stanovništva i klimatske promene, potreban je novi način razmišljanje, koje može da nam pruži Big Data.

Digitalne mrvice za hleb, koje ostavljamo iza sebe u svakodnevnom životu - koje otkrivaju o nama više nego o što smo bili spremni da otkrijemo - pružaju moćno sredstvo za rešavanje mnogih društvenih problema“

(Pentland, 2013).



https://www.duperrin.com/english/2014/11/17/quote-god-trust-others-bring-data-deming/

Primenjeni metodološki postupak

- Pristup „ZAŠTO“ i „KAKO“:
  - Jasna granica između objašnjenja i nagađanja
  - Bez želje da se impresionira publika:
    - ✓ Prekomernim korišćenjem matematike i statistike
    - ✓ Zloupotrebom jezika suvišnim tehničkim terminima
  - Inspired on Lipton & Steinhardt

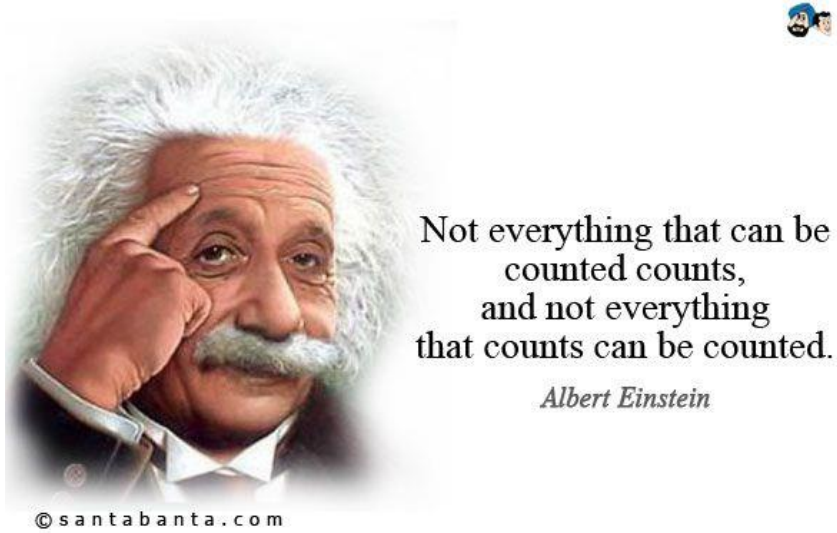
Ciljevi predavanja:

- ✓Upoznavanje sa osnovnim konceptima Nauke o podacima” – NoP (*Data Science*)
- ✓Rešavanje pojedinih nedoumica, zabuna i zabluda vezano za NoP
- ✓Ukazivanje na dalje pravce akcije u savladavanju discipline
- ✓**to stay industry-relevant and tech-savvy**

Pre nego što se upustimo u definisanje „Nauke o podacima“ (NoP) ne zaboravimo:

- Rad sa podacima ima dugu istoriju ...

... godinama predstavlja predmet rasprava nauke i struke ... matematičara, statističara, informatičara i drugih (filozofa, sociologa, bibliotekara, npr.)



Evolucija nauke, prema paradigmama

Paradigma	Trajanje	Naziv paradigme	Kratak opis	Primer
1.	**Hiljadu godina**	Eksperimentalna nauka	Opis prirodnih fenomena	Posmatranje prirodnih pojava
2.	Poslednjih nekoliko stotina godina	Teorijska nauka	Proučavanje različitih zakona i teorema	Njutnov zakon, Maksvelova jednačina...
3.	Poslednjih nekoliko decenija	Računarska nauka	Simulacija kompleksnih fenomena	Simulacioni modeli
4.	**Danas**	Nauka o intezivnoj upotrebi podataka	Korišćenje velike količine podataka različitih formata	Big Data, Nauka o podacima

Da se odmah razumemo...

Da biste se bavili NoP, u pravom smislu, neophodna su vam znanja:

- **Analitičko znanje** koju vam pruža matematika i statistika
- **Određeni nivo veština kodiranja** neophodan za rad sa podacima / korišćenje određenih softverskih alata)
- **Domenska ekspertiza** (subject matter expertise): Bez ove stručnosti vi ste matematičar ili statističar ili informatičar (programer, developer)

NoP – pristup definisanju

- Kada je nečega previše za nabrajanje (ili je previše kompleksno za objašnjavanje), efikasna (matematička) strategija je da se to nešto opiše njegovim komplementom - **onim što nije**.
- Navešćemo određeni broj pojmova (disciplina i tehnika) koje **same po sebi** ne predstavljaju NoP - ali je (za)okružuju.
- Redosled navođenja je dat prema (po sopstvenom viđenju) opadajućem redosledu njihovog značaja u konceptualizaciji NoP - što predstavlja glavni cilj današnjeg izlaganja
- U naslovima su ostavljeni orginalni nazivi na enegleskom jeziku (*NoP = Data Science*), u cilju lakšeg referenciranja

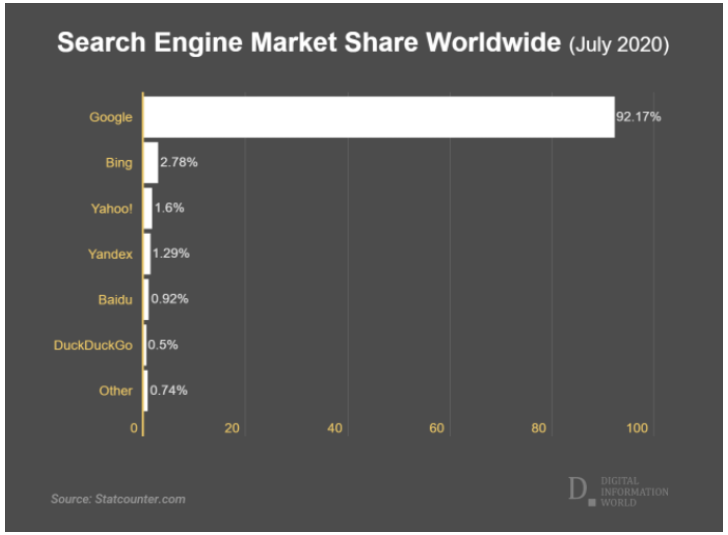
18 analytic *disciplines* compared to data science

- Machine learning
- Data mining
- Predictive modeling
- Statistics
- Industrial statistics
- Mathematical optimization
- Actuarial sciences (risk analysis)
- High performance computing
- Data analysis
- Operations research
- Six sigma (method that provides organizations tools to improve the capability of their business processes)
- Quant
- Artificial intelligence
- Computer science
- Econometrics
- Data engineering
- Business intelligence
- Business analytics.

(Granville, 2014)

Data Science vs. Science

Kako nas *Google* algoritmi vide



- Naučnici (*scientist*)
- Naučnici za podatke (*data scientist*):

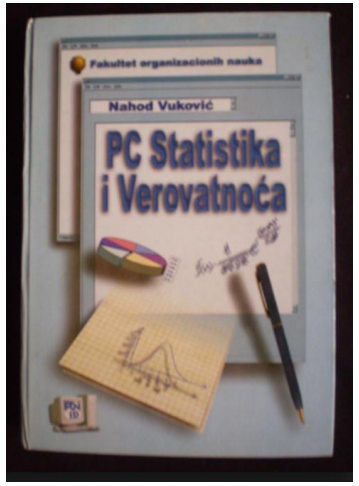
(Po ideji: *Data Science: An Artificial Ecosystem*, Harvard Data Science Review - Meng, 2019)

HDSI | Harvard Data Science Initiative

Data Science Vs. Statistics

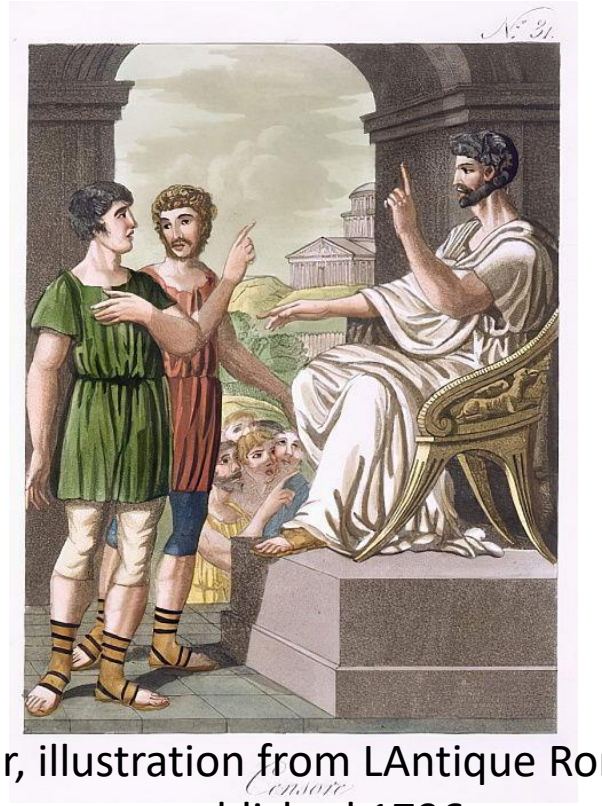
- Statistika: Status (l.) = stanje; pravni položaj jednog lica; prilike, položaj
- Statisticus (l.)= državni poslovi
- Prvi začeci statistike datiraju nekoliko vekova pre nove ere (p.n.e.)
- Prva prebrojavanja (za koja se zna) sprovedena su u Kini 4.000 godina p.n.e. i Egiptu 3.000 godina p.n.e

(Watson, 2001)



Prvi organizovani popisi

- Rimska republika, u starom veku
- Svake pete godine, u određeno vreme i na istom mestu
- Podaci su se prikupljali po domaćinstvima:
  - Pol
  - Starost
  - Prebivalište
  - Imovinsko stanje

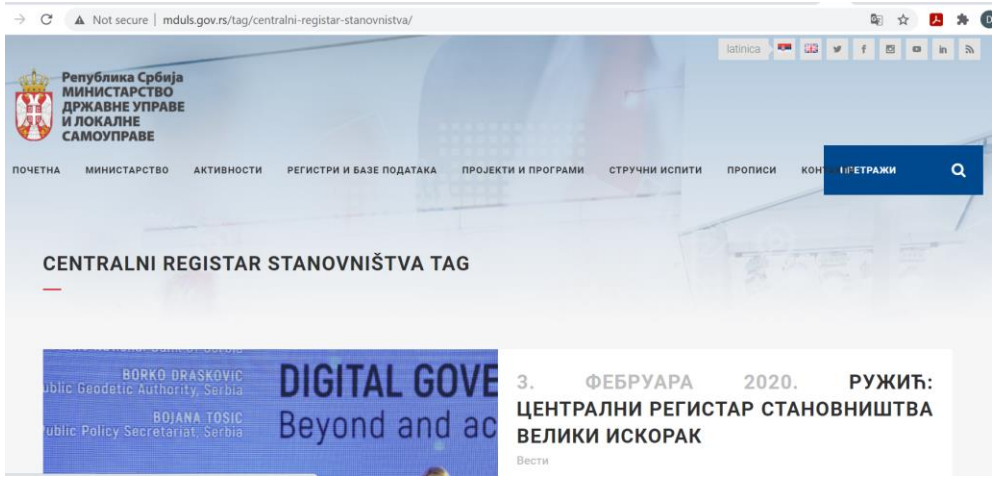


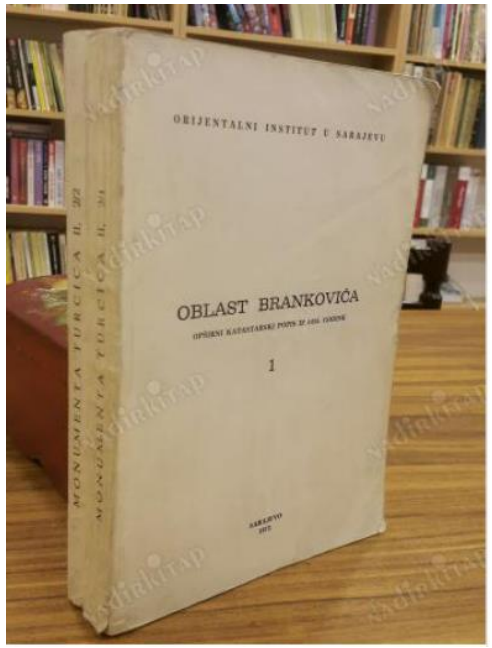
A Roman Censor, illustration from LAntique Rome, engraved by Labrousse, published 1796

Statistički registri

- Najstariji registar domaćinstava i pojedinaca datira iz doba dinastije Han u Kini – drugi vek p.n.e
- U XVII veku u Evropi se uvode se registri rođenih, umrlih i venčanih lica

Srbija: XV vek - Osmanske popisne knjige (defteri)
Srbija: XXI vek (20. novembar 2020) - Centralni registar stanovništva postao aktivan





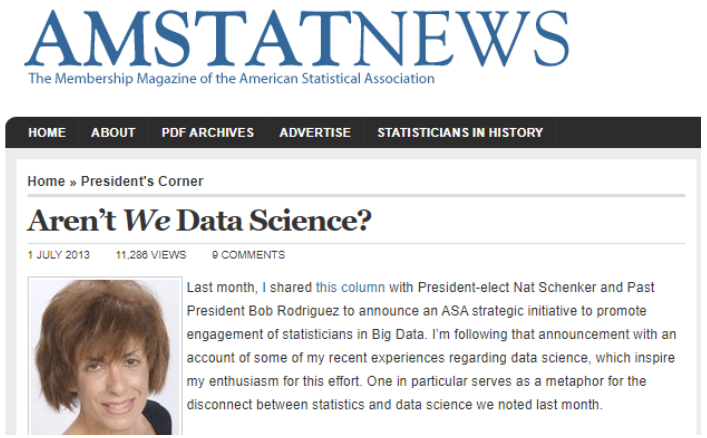
Data Science Vs. Statistics

- **Statistika** označava praksu ili nauku koja se prvenstveno bavi prikupljanjem i analiziranjem **numeričkih (strukturiranih) podataka**
- ***Data Science*** označava praksu ili nauku koja se bavi prikupljanjem i analiziranjem **svih vrsta sirovih podataka uključujući i nestrukturirane podatke**
- ***Data Science*** je **širi pojam od statistike** jer koristi i ostale kvantitativne metode i tehnike, kao što su metode operacionih istraživanja, data mining i sl.
- Tradicionalna **statistika** se bazira na pristupu odozgo nadole (*top-down approach*), od modela i teorije do podataka, dok ***Data Science*** koristi pristup odozdo prema gore (*bottom-up approach*), od podataka do modela ili algoritma.

NoP vs. Statistika

Zaključak: statistika je svakako deo NoP, ali IZGLEDA da njen doprinos disciplini nije tako veliki kako bi tradicionalni statističari želeli.

- Da li samo izgeda ili je suštinski tako? U mnogome zavisi od samih statističara. Koliko su spremni da prihvataju savremene koncepte i koliko su u stanju da promovišu svoju ulogu u NoP (i ne samo tu, u razvoju veštačke inteligencije, recimo). Npr. primer: pristrasnost u veštačkoj inteligenciji (VI)


AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME ABOUT PDF ARCHIVES ADVERTISE STATISTICIANS IN HISTORY

Home » President's Corner

Aren't *We* Data Science?

1 JULY 2013 11,286 VIEWS 9 COMMENTS

Last month, I shared this column with President-elect Nat Schenker and Past President Bob Rodriguez to announce an ASA strategic initiative to promote engagement of statisticians in Big Data. I'm following that announcement with an account of some of my recent experiences regarding data science, which inspire my enthusiasm for this effort. One in particular serves as a metaphor for the disconnect between statistics and data science we noted last month.

Istorijat – Verovatnoća (Probability)

- Ne bi bilo moguće donositi odluke uz rizik da nema verovanoće.
- Koncept „verovatnog“ (“*probable*”) verovatno datira od perioda pre Sokrata (469-399. P. N. E.) i provlači se kroz helenistički period (323 p. n. e. (smrt Aleksandra Velikog) - 146 p. n. e. (Aneksija klasične Grčke od Rima).
- Filozofi i biografi poput Demokrita, Platona (427-347. p. n. e.), Aristotela (384-322. p. n. e..) i Plutarha (46 -120) takožđe su koristili ovaj termin.
- Platon je izraz „verovatnoća“ (“probability”) koristio na potpuno isti način kao što se danas koristi u nemačkom jeziku (*Wahrscheinlichkeit* = sličnost nečemu što je istina).

- Prve proračune verovatnoće napravili su u 16. veku u Italiji matematičar Tartaglia (1500-1557) i matematičar i fizičar Gardano (1501-1576).

http://anson.ucdavis.edu/~roussas/Probability_And_Statistics_Throughout_The_Centuries.pdf

Data Science vs. Data Analysis

1962. *John W. Tukey* u uvodu svoje knjige "*The Future of Data Analysis*":

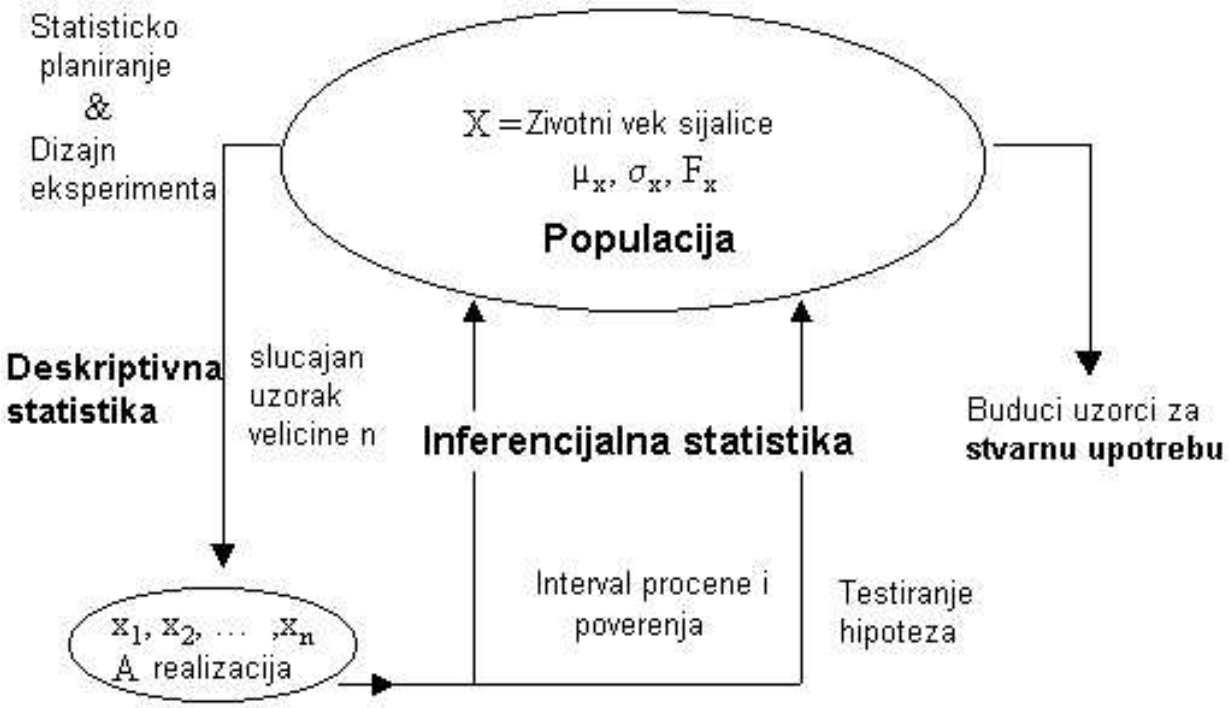
- *„For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ...*
- *All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data"*

Analiza podataka (AP)

Tukei (1962) je identifikovao **analizu podataka kao novu nauku** prepoznavši četiri glavna uticaja koja deluju na njen razvoj (Donoho, 2015):

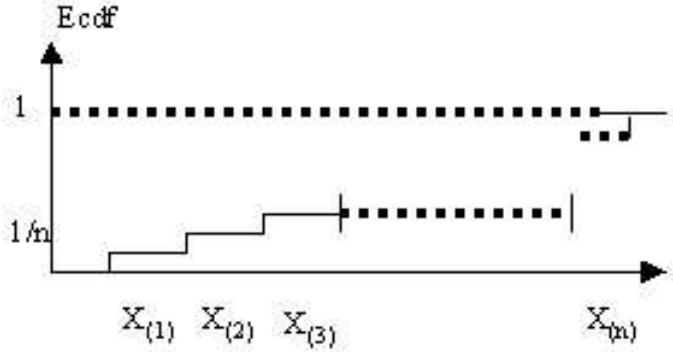
1. Formalne teorije statistike
2. Ubrzani razvoj računara
3. Pojava sve većeg broja podataka na mnogim poljima
4. Naglasak na kvantifikaciji u sve većem broju različitih disciplina

Statistika: nauka koja se bavi donošenjem odluka uz rizik



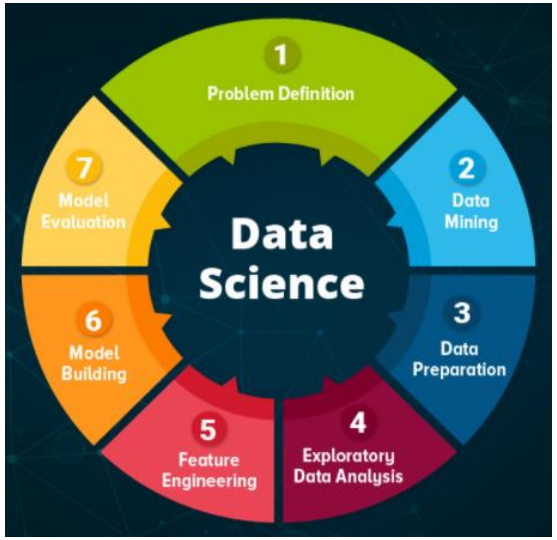
$$\bar{x} = \Sigma x_i / n$$

$$s_x^2 = \Sigma(x_i - \bar{x})^2 / (n-1)$$



Exploratory data analysis (EDA)

- Klasična statistika uglavnom se fokusira na zaključivanje, odnosno na složen skup postupaka za donošenje zaključaka o velikim populacijama zasnovane na malim uzorcima
- Eksplanatorna analiza podataka, kako joj ime samo kaže, se bavi podacima
- Da bi se u potpunosti mogli primeniti „tradicionalni statistički koncepti“ (metodologija), nestrukturirani sirovi podaci moraju se „strurkturirati“ (obraditi i predstaviti u strukturiranom obliku)



https://www.aismartz.com/blog/why-eda-is-crucial-for-any-data-science-project/

Definicija informatičara

Analiza podataka je postupak ispitivanja, transformisanja i sređivanja datog skupa podataka na određene načine kako bi se proučili njegovi pojedinačni delovi i izvukle korisne informacije, i podrazumeva:

1. Razumevanje modela podataka
2. Manipulaciju podacima
3. Predstavljanje dobijenih rezultata

- **Mnoge tehnike i procesi analize podataka automatizovani su u mehaničke procese i algoritme**

- **Analitičari podataka** imaju uži spektar znanja i višu specijalizaciju od naučnika za podatke, a nedostaje im (i nije im potrebna) „ šira slika“ (npr. poslovna vizija)

- Zbog velikog obima novih podataka i njihovog stanja**, analiza podataka** obično uključuje „ljudsku kreativnost“ kako bi razumela i ostvarila uvid u raspoložive „sirove“ podatke.

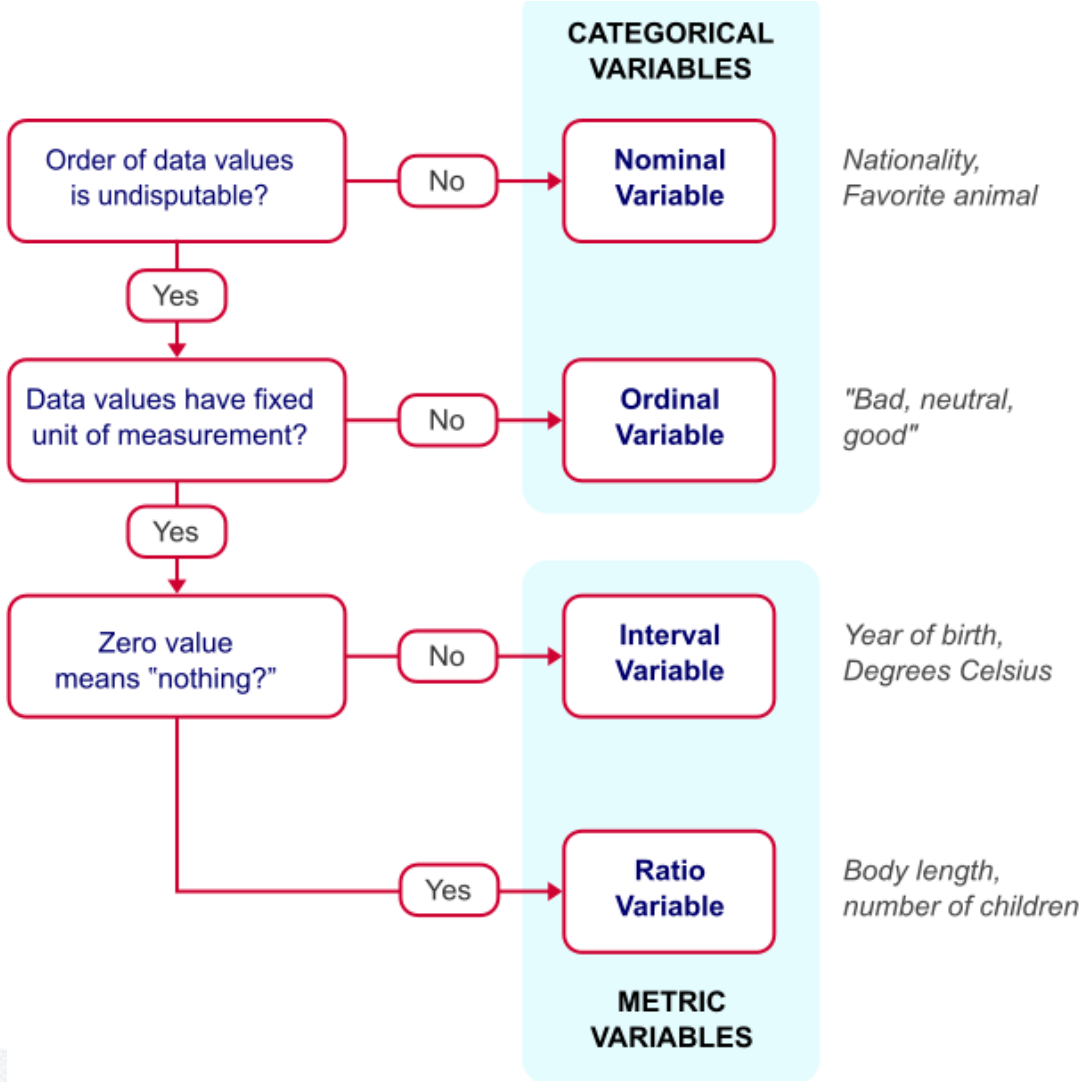
(Lim, 2020)

Preliminarna obrada podataka

Kada je **n** malo, i radi se sa strukturiranim podacima nije problem ali kada se obrađuje veliki broj podataka (često i različitih struktura) neophodno je njihovo prethodno sređivanje

- Tabelarno prikazivanje:
  - Frekvencije (apsolutne frekvencije)
  - Procenti (relativne frekvencije)
- Grafičko prikazivanje
- Transformacija podataka
- **Način prikazivanja zavisi od tipa obeležja**

Measurement level Ibm SPSS



Statistika

Skala	Deskriptivne mere (primer)	Dozvoljene statistike (primer)
Nominalna	procenti, modus	*Hi*-kvadrat test, Binomni test
Ordinalna	percentili, medijana, kvartili	Korelacija rangova
Intervalna	aritmetička sredina, standardna devijacija	Koeficijent korelacije, t-test, ANOVA, regresija, faktorska analiza
Racionalna	sve deskriptivne mere (srednje vrednosti, mere varijabiliteta, oblik distribucije)	**Sve**

Tipovi podataka u NoP

Continuous: Data that can take on any value in an interval.

Synonyms: interval, float, numeric

Discrete: Data that can take on only integer values, such as counts.

Synonyms: integer, count

Categorical: Data that can take on only a specific set of values representing a set of possible categories.

Synonyms: enums, enumerated, factors, nominal, polychotomous

Binary: A special case of categorical data with just two categories of values (0/1, true/false).

Synonyms: dichotomous, logical, indicator, Boolean

Ordinal: Categorical data that has an explicit ordering.

Synonyms: ordered factor

(Bruce and Bruce, 2017)

Tipovi podataka

R:

- Numeric
- Integer
- Complex
- Logical
- Character

Python:

- int (an integer)
- float (a decimal)
- boolean (either True or False)
- string (text or words made up of characters)
- list (a collection of objects)

Značaj tipova podataka u NoP

- Tip podataka važan za određivanje vrste vizuelnog prikaza, analize podataka ili statističkog modela.
- R i Python koristi tipove podataka za poboljšanje računskih performansi.
- Eksplicitna identifikacija podataka kao kategoričnih, za razliku od teksta, nudi neke prednosti:

1. Znanje da su podaci kategorički može delovati kao signal koji govori softveru kako treba da se ponašaju statističke procedure, poput izrade grafikona ili fitovanje modela.
2. Ordinarni podaci mogu se predstaviti kao ordered.factor u R-u i Python-u, čuvajući redosled koji je odredio korisnik u grafikonima, tabelama i modelima.
3. Skladištenje i indeksiranje mogu se optimizovati (kao u relacionoj bazi podataka). Moguće vrednosti koje data kategorička promenljiva može zauzeti definišu se u softveru (poput nabrajanja).
4. Podrazumevano (default) ponašanje funkcija uvoza podataka u R (npr. read.csv) je da automatski pretvara tekstualne kolone u faktor. Naknadne operacije na toj koloni pretpostaviće da su jedine dozvoljene vrednosti za tu kolonu one koje su prvobitno uvezene, a dodeljivanje nove tekstualne vrednosti uvest će upozorenje i proizvesti NA (nedostajuća vrednost)

(Bruce and Bruce, 2017)

Strukture podataka: pravougaone

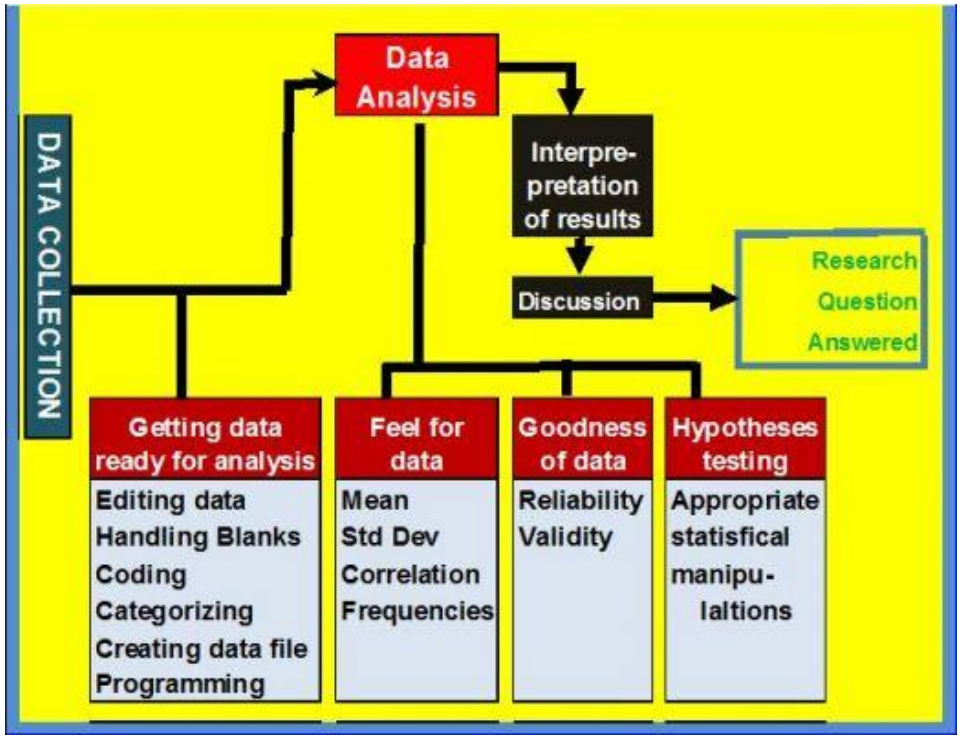
- dvodimenzionalna matrica sa indikacijom redova (record, cases) i kolona koje ukazuju na obeležja (promenljive – variables)
- U R-u, osnovna pravougaona struktura podataka je data.frame objekat.

Strukture podataka: nepravougaone

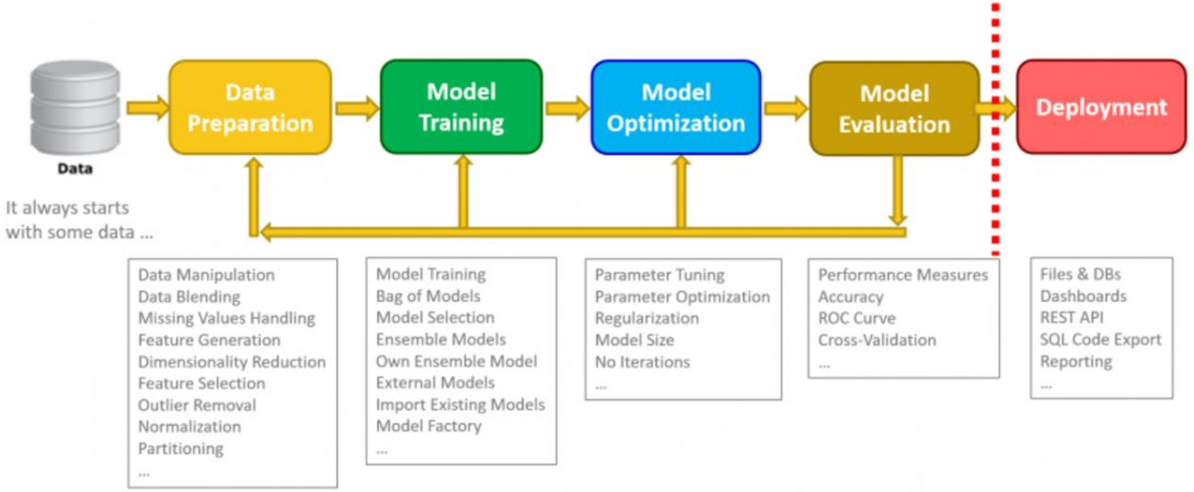
- Podaci o **vremenskim serijama** beleže uzastopna merenja iste promenljive. To je sirov materijal za metode statističkog predviđanja, a takođe je ključna komponenta za podatke koje proizvode Internet of Things (IoT).
- Strukture **prostornih podataka** (spatial data structures), koje se koriste u mapiranju i prostornoj analitici (location analytics), jesu složenije i raznovrsnije od pravougaonih struktura podataka. U objektu predstavljanja (object representation), fokus podataka je objekat (npr. kuća) i njegove prostorne koordinate.
- **Grafičke (mrežne) strukture podataka** (Graph (network) data structures) koriste se za predstavljanje fizičkih, društvenih i apstraktnih odnosa. Na primer, grafikoni društvenih mreža, kao što su npr. Facebook ili LinkedIn, mogu predstavljati veze između ljudi na mreži.

Analiza podataka

Tradicionalni pristup



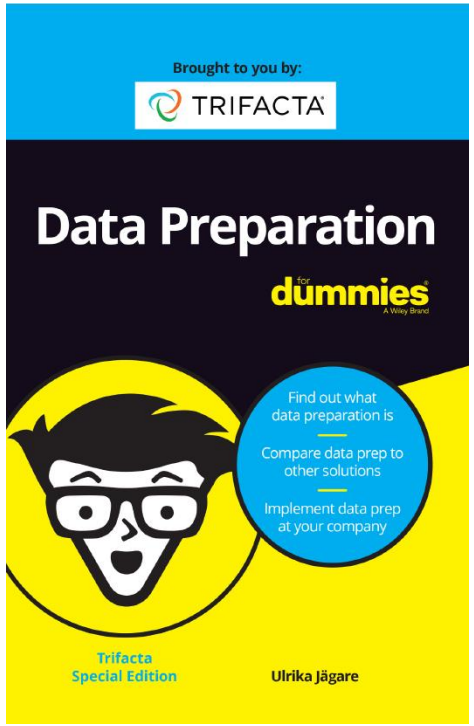
Savremeni pristup



https://www.knime.com/blog/analytics-and-beyond

Uvod u analizu podataka

- Tabelarnim i grafičkim prikazivanjem vrši se preliminarna analiza podataka:
  - Prethodno utvrđivanje tipova obeležja - podataka (merne skale)
  - Provera ispravnosti podataka
  - Identifikacija nedostajućih vrednosti
  - Identifikacija ekstremnih vrednosti
  - Vizuelna provera raspodele obeležja



… zbog velikog obima novih podataka i njihovog stanja…

Welcome to the New Normal

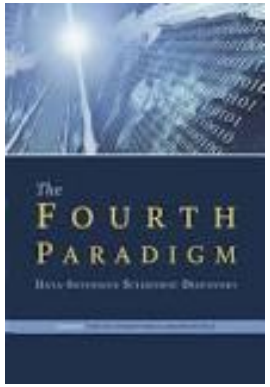
… **eksponencijalna vremena!**

- *There has never been a better time to apply scientific thinking to business problems*

(Danner, 2015)

„Četvrta paradigma je možda jedini sistemski način za rešavanje nekih od najvećih globalnih izazova s kojima se danas suočavamo, i …

… nije samo promena u načinu naučnog istraživanja, već i promena u načinu na koji ljudi misle“



Jim Gray Turing Award winner, who was tragically lost at sea in January 2007

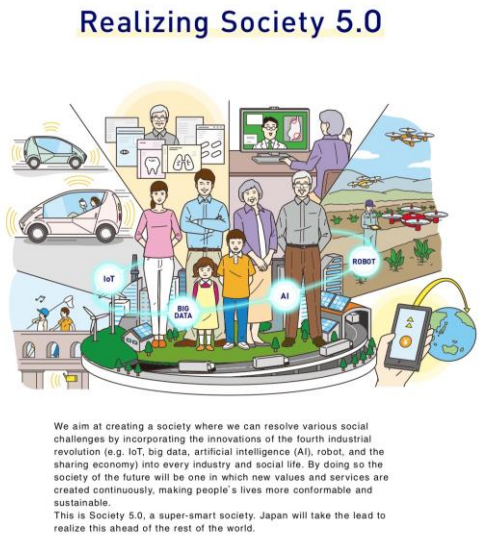
(Hey et al., 2011)

Education 4.0 / Society 5.0

- *“In the era of Google, people no longer need to memorize every single fact. Many tasks today are best carried out by computers.*

… we have to give students the skills to both survive that changing society and for them to lead that change”.

- Japan’s former education minister Yoshimasa Hayashi



NoP – istorija

- 1996. Članovi Međunarodne federacije klasifikacionih društava (*International Federation of Classification Societies (IFCS)*) sastali su se u Kobeu u Japanu na dvogodišnjoj konferenciji. Po prvi put je pojam *Data science* uključen **u naslov konferencije** ("*Data science, classification, and related methods*").
- 2001. Villiam S. Cleveland je prvi put upotrebio pojam *Data science* **u naslovu nekog naučno/stručnog rada**: "*Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*." International Statistical Review / Revue Internationale de Statistique. Vol. 69, No. 1 (Apr., 2001), pp. 21-26 (6 pages), Published By: International Statistical Institute (ISI) (Cleveland, 2001).
- **2002. Nauka o podacima postaje nauka**: Međunarodni savet za nauku (*International Council for Science)* prihvatio je Data science kao nauku i stvorio je odbor za nju.

Nauka o podacima (NoP) – (Data Science)

(*National Science Foundation, Directorate for Mathematical and Physical Science, Support for the Statistical Sciences at NSF—a subcommittee of the Mathematical and Physical Sciences Advisory Committee*)

2. Data Science in NSF context

Motivated by NSF Strategic Plan and initial discussions with ADs

Our definition:
"*Data Science:* the science of *planning for*, *acquisition*, *management*, *analysis* of, and *inference* from data"

Our context:
Data science and the enhanced application of data science at NSF

https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

Više o istorijatu NoP

- Donoho, D. (2015). 50 **Years of Data Science**. Tukey Centennial workshop in Princeton, New Jersey. Pristupljeno 11.12.2020. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
- Foote, K. D. (2016). **A Brief History of Data Science**. Dataversity Education, LLC. Pristupljeno 13.12.2020. https://www.dataversity.net/brief-history-data-science/
- MacTutor (2020). **Analysis - History Topics**. School of Mathematics and Statistics. University of St Andrews, Scotland. Pristupljeno: 12.12.2020 https://mathshistory.st-andrews.ac.uk/HistTopics/
- morris.umn. (2013). **John Wilder Tukey**. The University of Minnesota, Morris. Pristupljeno: 12.12.2020. http://mnstats.morris.umn.edu/introstat/history/w98/Tukey.html
- Press, G. (2013). **A Very Short History Of Data Science**. Forbes. Pristupljeno:12.12.2020. December 8, 2020. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#6b29d42055cf

NoP vs. veliki podaci

~~Data science = Big Data~~

Big data is data that exceeds the processing capacity of conventional database systems

Još uvek imamo puno podataka koje tradicionalne tehnike i alati ne mogu da uskladište i obrade

Hollerith's electronic tabulator, 1902

CS: 1890 US Census

- Procenjeno je da će biti potrebno 13 godina za obradu podataka, popis je svakih 10 godina u SAD-u, po zakonu !!!

- Bilo Hollerith je došao na ideju da se buše kartice za čuvanje zapisa (podataka) popisa stanovništva: rupa ili ne ukazuje na muškarca ili ženu, na primer - preteča formata binarne memorije na digitalnom računaru

- Sledeći izazov: kako da se ove bušene karte automatski čitaju čime bi se omogućila brojanje a potom i tabulacija – iskoristio je Edisonov izum električne energije da bi pomoću električne mašine vršio brojanje registrovanjem da li struja teče kroz rupu ili ne.

A Hollerith Tabulating Machine.

Photo: Wikipedia/Marcin Wichary

Herman Hollerith (1860–1929) – preteča današnjeg naučnika za podatke

- **Statistics** (S) - radio je kao pridruženi službenik u Popisnoj kancelariji (Census Office), organizaciji osnovanoj za obradu propisa 1880. godine (do 1900. godine svaki američki popis organizovala je i obradila nova, privremena organizacija; stalni Biro za popis stanovništva uspostavljen je tek 1902. godine)
- **Domain/science knowledge** (D)
- **Computing** (C)
- **Collaboration/teamwork** (C)
- **Communication** to outsiders (C)

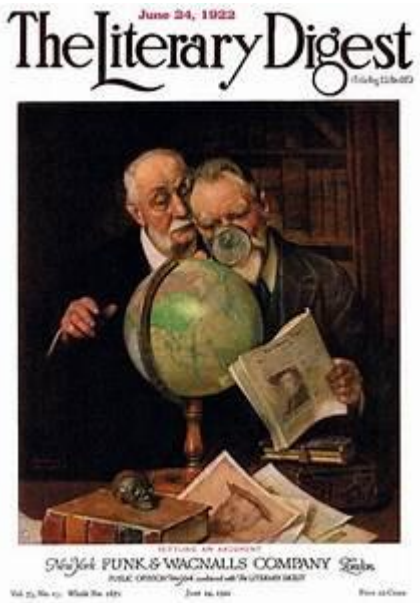
It began as the Computing, Tabulating & Recording Company (C-T-R) founded by Herman Hollerith in the late 1800s. Their first large contract was to provide tabulating equipment for the tabulation and analysis of the 1890 US census. The company grew quickly and, in the early 1920s the name was changed to **IBM**.

Data Science = SDCCC = SDC3

CS: The Literary Digest
- predizborna prognoza 1936.

n = 2.4 million, mailing list of about 10 million names**:**
Landon would get 57% of the vote against Roosevelt's 43%

- The actual results of the election were 62% for Roosevelt against 38% for Landon



"Epilog:

1. **Bankrot**
2. **Primena naučnog metoda – Galup, 1935**

https://www2.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

Big Data koncept, filozofija, pristup

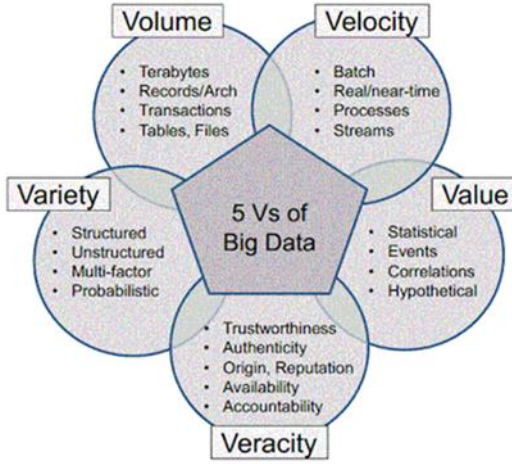
Big Data (BD) je više od velikih podataka (Big data = Big picture)

- Do skoro: podaci su bili problem - tumačenje relativno jednostavno
- Od skora: podataka ima u izobilju - tumačenje je problem

- Svi podaci moraju biti kategorisani, analizirani i vizualizovani

Rather than looking at segments, classifications, regions, groups, or other summary levels you'll have insights into ***all*** the individuals, ***all*** the products, ***all*** the parts, ***all*** the event, ***all*** the transactions, etc.

- NoP to omogućava

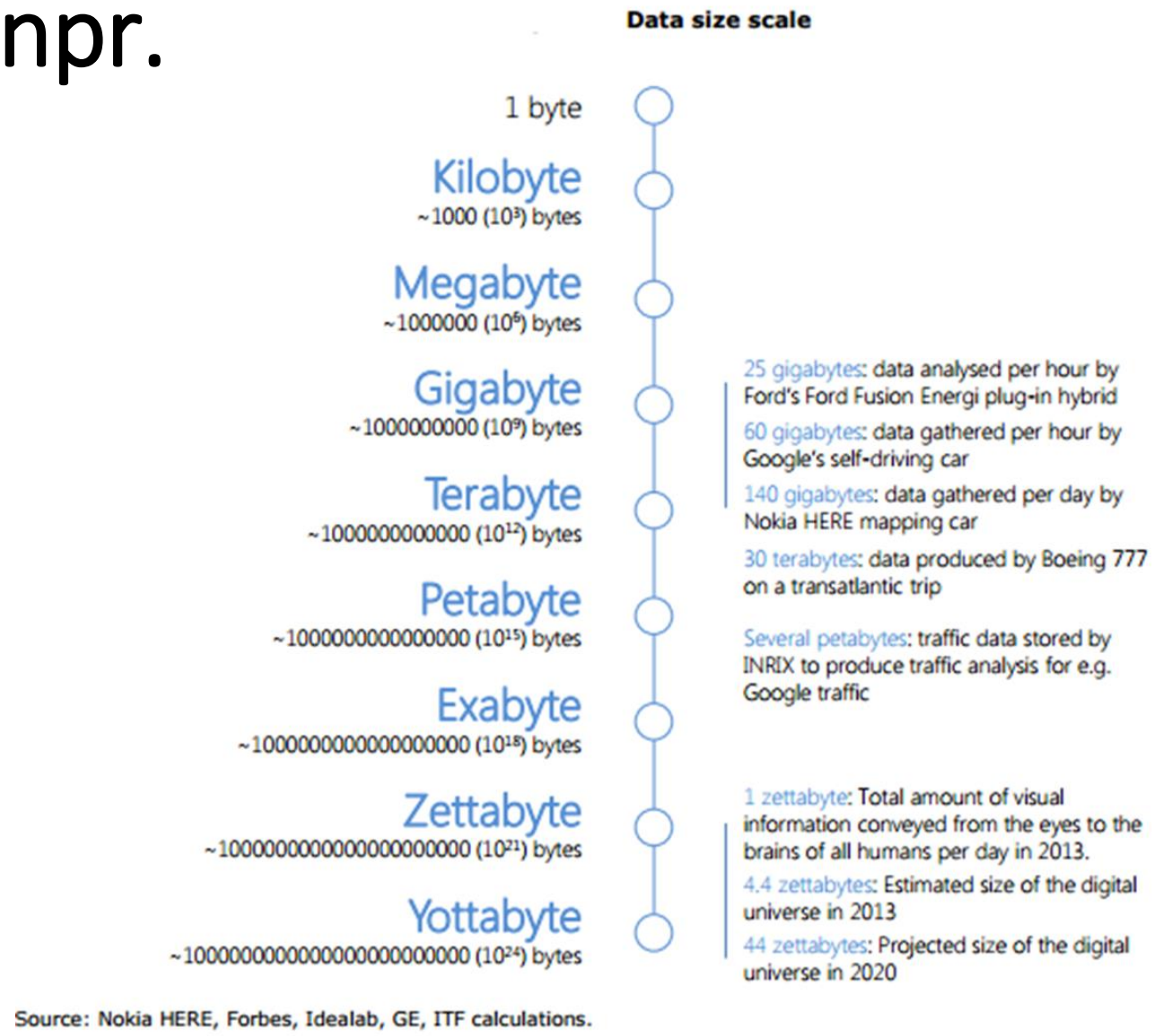


(Demchenko et al., 2013)

Big data – Da kvantifikujemo - Koliko veliko?

- Donja granica za BD je 1 terabajt, dok gornja granica ne postoji. Ako organizacija poseduje najmanje 1 terabajt podataka, verovatno je dobar kandidat za primenu Big Data (Pierson, 2017)
- Paradoks: Generalno, većina BD beleži male vrednosti u izvornom (sirovom) obliku, odnosno sastoje od ogromnog broja vrlo malih transakcija koje se nalaze u različitim formatima (ima nisku incidencu što u velikoj meri determiniše DS modele koji će se potencijalno implementirati.
- Inženjeri podataka imaju posao da prikupe i srede te podatake, tako da dobiju formu pogodnu za DS.

Volume – npr.



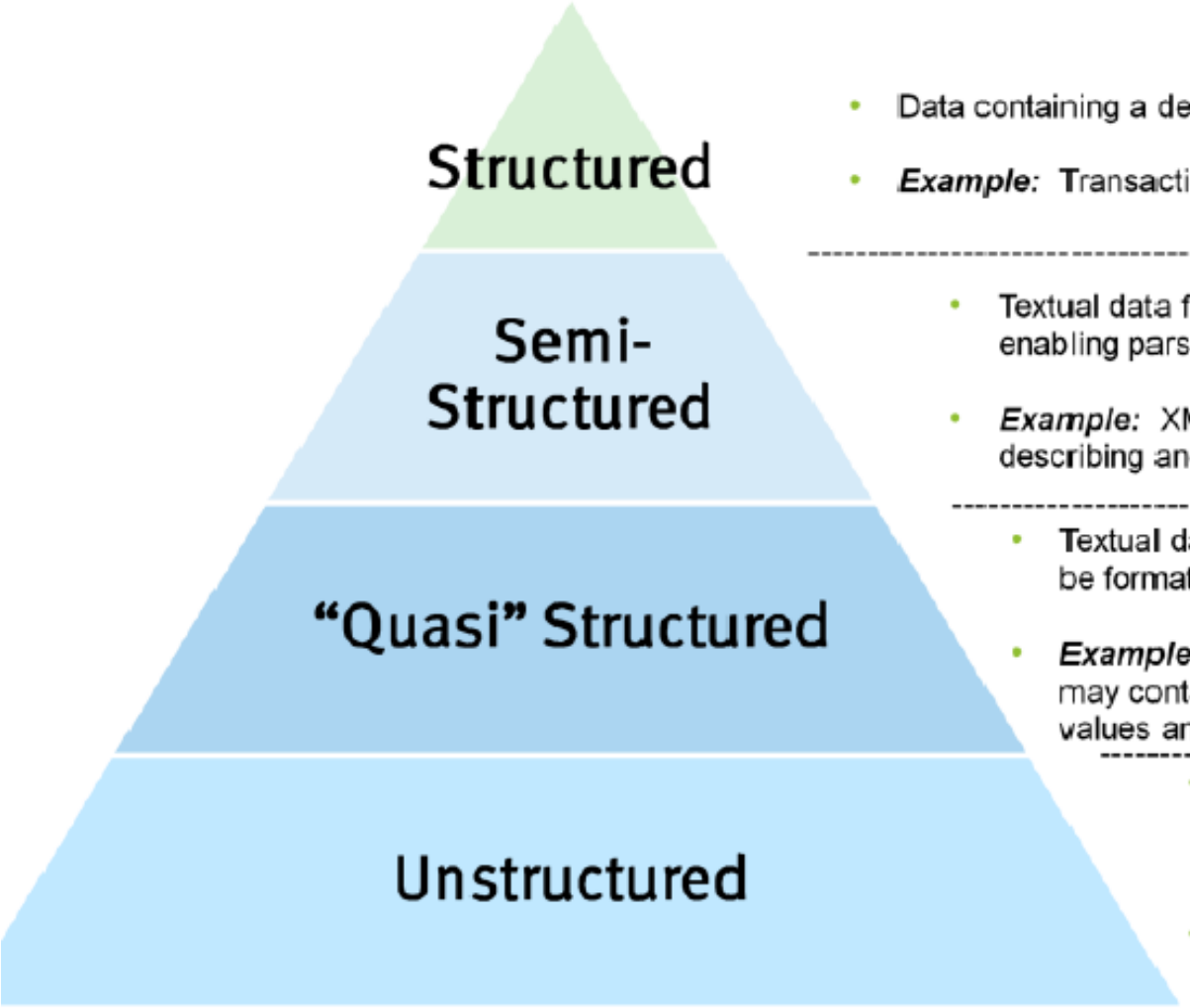
(NOESIS, 2018)

Big data – Koliko brzo? (Kvantifikacija *Velocity*)

- Podaci koji se kreću velikom brzinom predstavljaju pravi izazov pravovremenom donošenju odluka u realnom vremenu
- Veliki podaci ulaze u prosečan sistem pri brzinama u rasponu od 30 kilobajta (KB) u sekundi do čak 30 gigabajta (GB) po sekundi.
- Mnogi sistemi zasnovani na podacima moraju imati latenciju manju od 100 milisekundi, mereno od trenutka nastajanja (pojave) podataka do trenutka kada sistem odgovori.
- Zahtevana propusnost može biti i 1.000 poruka u sekundi u BD sistemima!
- Tehnologije često predstavljaju usko grlo i limitrajući faktor brzine podataka.
- BD su low-value (jeftini), pojedinačno i nesređeni nemaju neku vrednosti. Zbog toga su vam potrebni sistemi koji će moći da ih „progutaju“ u kratkim vremenskom periodu, kako bi bili primenljivi na najbolji način.

(Pierson, 2017)

Big DATA



- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

---

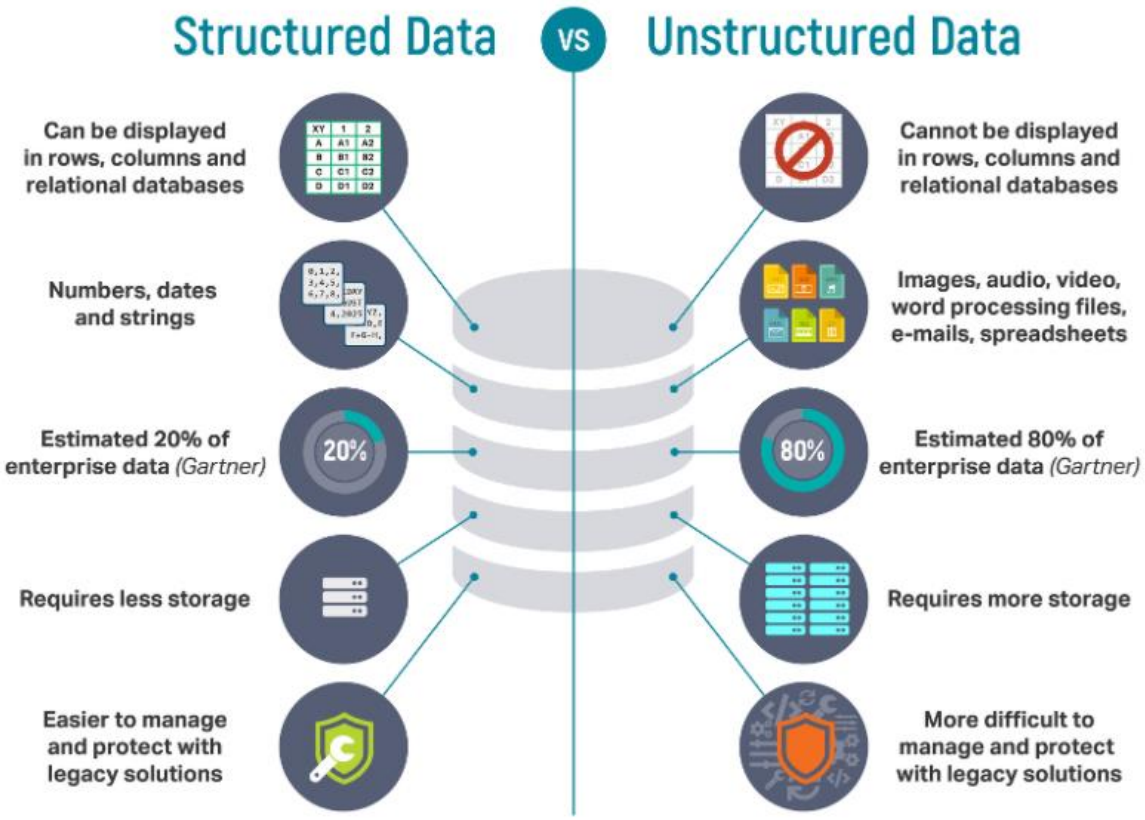
- Textual data with erratic data formats, can be formatted with effort, tools, and time
- ***Example:*** Web clickstream data that may contain some inconsistencies in data values and formats

---

- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images and video

Variety - da kvantifikujemo - Koliko raznoliko?

- The DATA has increasingly become "**unstructured**":
  - *Text*
  - *Audio*
  - *Video,*
  - *Image,*
  - *Geospatial*
  - *Internet data:*
    - Click streams
    - Log files

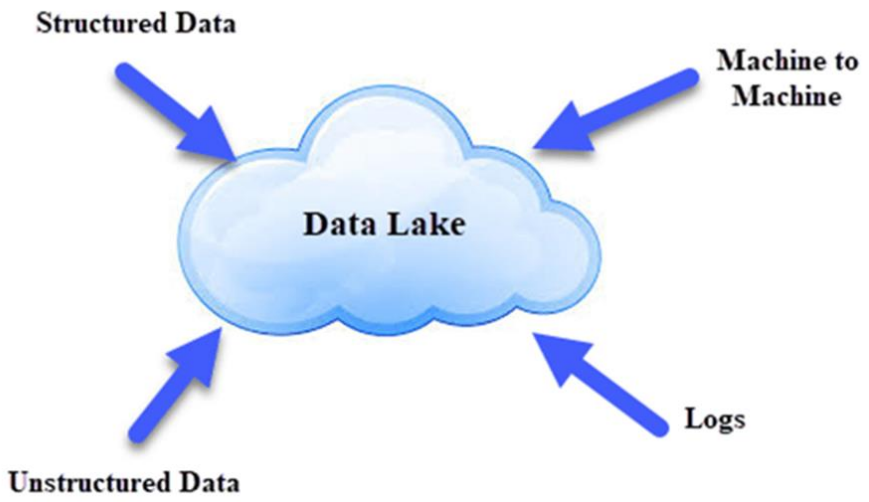


https://www.igneous.io/blog/structured-data-vs-unstructured-data

Data Lake

Mesto za čuvanje ogromne količine sirovih i neobrađenih podataka u izvornom formatu

- Baš kao što u jezero imate više pritoka, jezero podataka sadrži strukturirane podatke, nestrukturirane podatke, machine to machine, logs... prikupljene u realnom vremenu



What is Data Lake? It's Architecture

TABLE 1. The most popular big data positions.

Positions	Description	Capabilities
Data Scientist	Customer-oriented, create products and processes with meaningful value-added services	Extract and synthesize data, statistical analysis, data insight and information mining, develop software.
Data Systems Architect	Big data platform construction, big data systems design, infrastructure	Computer systems architecture, network architecture, programming paradigm, file system, distributed parallel processing, etc.
Data Systems Analyst	Data security life cycle management, analysis and application	Artificial intelligence, machine learning, mathematical statistics, matrix computing, optimization methods
Hadoop Engineer	Solve big data storage problems	Hadoop, Python, Linux, etc.
Data Analyst	Extract, analyze, and present data to make business sense of the data	SPSS, STATISTIC, EViews, SAS, big data Magic Mirror and other data analysis software
Data Mining Engineer	Find patterns in big data	Linear algebra, higher algebra, convex optimization, probability theory, Hadoop & MapReduce, Python & Spark.
Data Visualization Engineer	Design the visualization scheme that meets the business requirements	Choose the right visualization technique, make and promote visual samples, sample componentization.

Data Science vs. Big data

X. Li *et al.*: Curriculum Reform in Big Data Education at Applied Technical Colleges and Universities in China

IEEE Access

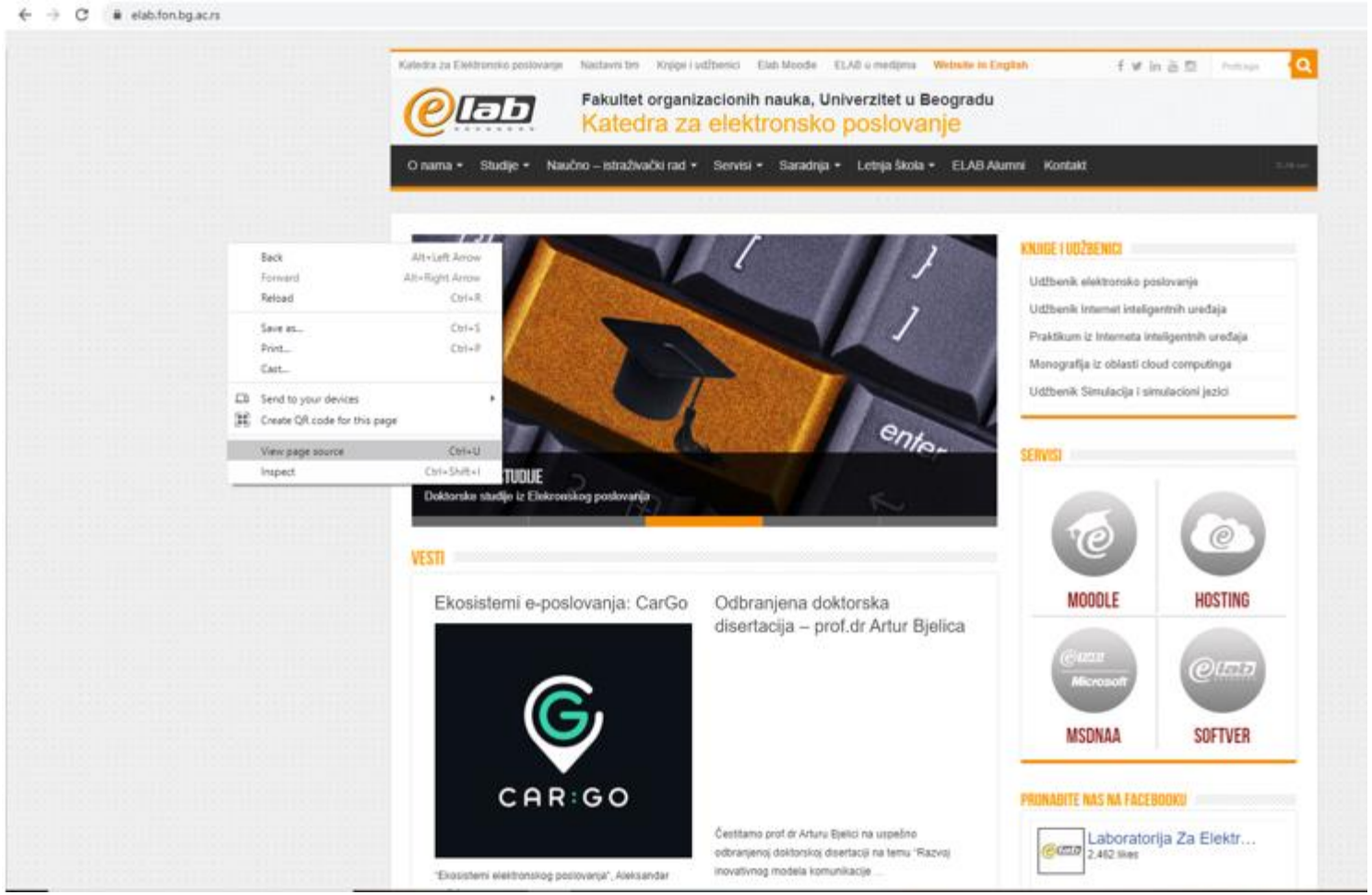
Analitička piramida Big Data / Data Science koncepta

- A - struktuirani podaci (koji mogu biti eksterni ili interni).
- B - nestruktuirani podaci koji se konvertuju u struktuirane - poznate.
- C - nepoznati nestruktuirani podaci koji će se obrađivati u svom prvobitnom obliku, bez konverzije
- D - proizvodnja sopstvenih podataka.

(Izvor: Minelli & Chambers, 2013)

Polu-strukturirani podaci

- koriste se za opis strukturiranih podataka koji se ne uklapaju u formalnu strukturu modela podataka
  - *Sadrže oznake koje razdvajaju semantičke elemente, što uključuje sposobnost nametanja hijerarhija unutar podataka*



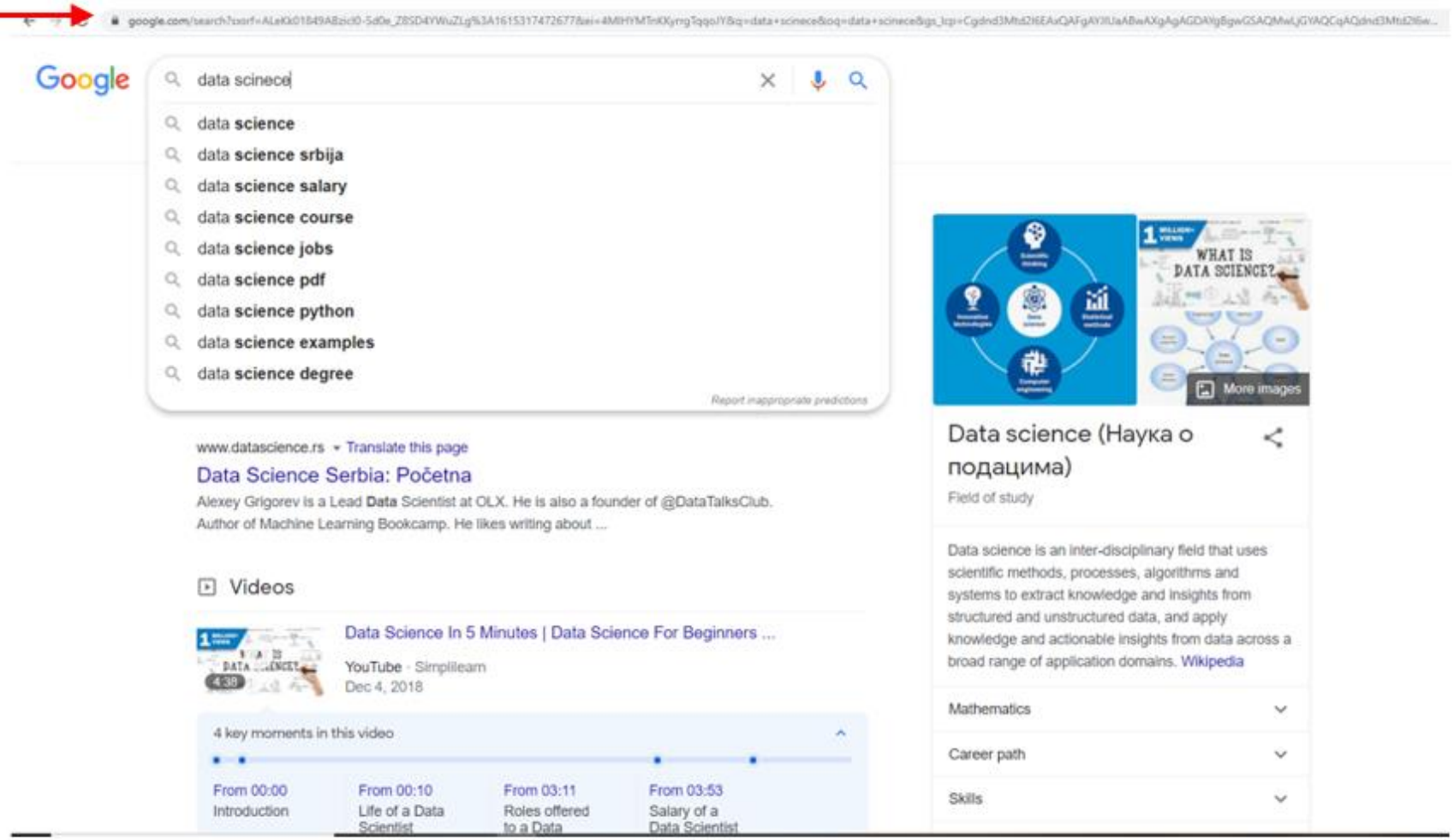
view-source:https://elab.fon.bg.ac.rs

```
<!DOCTYPE html>
<html lang="en-US" prefix="og: http://ogp.me/ns#">
<head>
<meta charset="UTF-8" />
<link rel="profile" href="http://gmpg.org/xfn/11" />
<link rel="pingback" href="https://elab.fon.bg.ac.rs/xmlrpc.php" />
<script src='https://ajax.googleapis.com/ajax/libs/jquery/3.3.1/jquery.min.js'></script>

    <!-- This site is optimized with the Yoast SEO plugin v14.6.1 - https://yoast.com/wordpress/plugins/seo/ -->
    <title>ELAB home page - Katedra za elektronsko poslovanje</title>
    <meta name="robots" content="index, follow" />
    <meta name="googlebot" content="index, follow, max-snippet:-1, max-image-preview:large, max-video-preview:-1" />
    <meta name="bingbot" content="index, follow, max-snippet:-1, max-image-preview:large, max-video-preview:-1" />
    <link rel="canonical" href="https://elab.fon.bg.ac.rs/" />
    <meta property="og:locale" content="en_US" />
    <meta property="og:type" content="website" />
    <meta property="og:title" content="ELAB home page - Katedra za elektronsko poslovanje" />
    <meta property="og:url" content="https://elab.fon.bg.ac.rs/" />
    <meta property="og:site_name" content="Katedra za elektronsko poslovanje" />
    <meta property="article:modified_time" content="2018-05-24T11:52:56+00:00" />
    <script type="application/ld+json" class="yoast-schema-graph">{"@context":"https://schema.org","@graph":[{"@type":"WebSite","@id":"https://elab.fon.bg.ac.rs/#website","url":"https://elab.fon.bg.ac.rs/","name":"Katedra za elektronsko
poslovanje","description":"Univerzitet u Beogradu, Fakultet organizacionih nauka","potentialAction":[{"@type":"SearchAction","target":"https://elab.fon.bg.ac.rs/?s={search_term_string}","query-input":"required name=search_term_string"}],"inLanguage":"en-US"}
{"@type":"WebPage","@id":"https://elab.fon.bg.ac.rs/#webpage","url":"https://elab.fon.bg.ac.rs/","name":"ELAB home page - Katedra za elektronsko poslovanje","isPartOf":{"@id":"https://elab.fon.bg.ac.rs/#website"},"datePublished":"2018-05-
09T12:46:45+00:00","dateModified":"2018-05-24T11:52:56+00:00","inLanguage":"en-US","potentialAction":[{"@type":"ReadAction","target":["https://elab.fon.bg.ac.rs/"]}]}]}</script>
    <!-- / Yoast SEO plugin. -->


<link rel='dns-prefetch' href'//fonts.googleapis.com' />
<link rel='dns-prefetch' href'//s.w.org' />
<link rel="alternate" type="application/rss+xml" title="Katedra za elektronsko poslovanje &raquo; Feed" href="https://elab.fon.bg.ac.rs/feed/" />
<link rel="alternate" type="application/rss+xml" title="Katedra za elektronsko poslovanje &raquo; Comments Feed" href="https://elab.fon.bg.ac.rs/comments/feed/" />
        <script type="text/javascript">
            window._wpemojiSettings = {"baseUrl":"https:\/\/s.w.org\/images\/core\/emoji\/12.0.0-1\/72x72\/","ext":".png","svgUrl":"https:\/\/s.w.org\/images\/core\/emoji\/12.0.0-1\/svg\/","svgExt":".svg","source":
{"concatemoji":"https:\/\/elab.fon.bg.ac.rs\/wp-includes\/js\/wp-emoji-release.min.js"}};
            /*! This file is auto-generated */
            !function(e,a,t){var r,n,o,i,p=a.createElement("canvas"),s=p.getContext&&p.getContext("2d");function c(e,t){var a=String.fromCharCode;s.clearRect(0,0,p.width,p.height),s.fillText(a.apply(this,e),0,0);var r=p.toDataURL();return
s.clearRect(0,0,p.width,p.height),s.fillText(a.apply(this,t),0,0),r===p.toDataURL()}function l(e){if(!s||!s.fillText)return!1;switch(s.textBaseline="top",s.font="600 32px Arial",e){case"flag":return!c([127987,65039,8205,9895,65039],
[127987,65039,8203,9895,65039])&&(!c([55356,56826,55356,56819],[55356,56826,8203,55356,56819])&&!c([55356,57332,56128,56423,56128,56418,56128,56421,56128,56430,56128,56423,56128,56447],
[55356,57332,8203,56128,56423,8203,56128,56418,8203,56128,56421,8203,56128,56430,8203,56128,56423,8203,56128,56447]));case"emoji":return!c([55357,56424,55356,57342,8205,55358,56605,8205,55357,56424,55356,57340],
[55357,56424,55356,57342,8203,55358,56605,8203,55357,56424,55356,57340])}return!1}function d(e){var t=a.createElement("script");t.src=e,t.defer=t.type="text/javascript",a.getElementsByTagName("head")[0].appendChild(t)}for(i=Array("flag","emoji"),t.supports=
{everything:!0,everythingExceptFlag:!0},o=0;o<i.length;o++)t.supports[i[o]]=l(i[o]),t.supports.everything=t.supports.everything&&t.supports[i[o]],"flag"!==i[o]&&
(t.supports.everythingExceptFlag=t.supports.everythingExceptFlag&&t.supports[i[o]]);t.supports.everythingExceptFlag=t.supports.everythingExceptFlag&&!t.supports.flag,t.DOMReady=!1,t.readyCallback=function(){t.DOMReady=!0},t.supports.everything||(n=function()
{t.readyCallback()},a.addEventListener?(a.addEventListener("DOMContentLoaded",n,!1),e.addEventListener("load",n,!1)):(e.attachEvent("onload",n),a.attachEvent("onreadystatechange",function(){"complete"===a.readyState&&t.readyCallback()})),(r=t.source||
{}).concatemoji?d(r.concatemoji):r.wpemoji&&r.twemoji&&(d(r.twemoji),d(r.wpemoji))))}(window,document,window._wpemojiSettings);
        </script>
        <style type="text/css">
img.wp-smiley,
img.emoji {
    display: inline !important;
    border: none !important;
    box-shadow: none !important;
    height: 1em !important;
    width: 1em !important;
    margin: 0 .07em !important;
    vertical-align: -0.1em !important;
    background: none !important;
    padding: 0 !important;
}
</style>
    <link rel='stylesheet' id='ultimate-tables-style-css'  href='https://elab.fon.bg.ac.rs/wp-content/plugins/ultimate-tables/css/ultimate-tables.css' type='text/css' media='all' />
<link rel='stylesheet' id='ultimate-datatables-style-css'  href='https://elab.fon.bg.ac.rs/wp-content/plugins/ultimate-tables/css/jquery.dataTables.css' type='text/css' media='all' />
<link rel='stylesheet' id='wp-block-library-css'  href='https://elab.fon.bg.ac.rs/wp-includes/css/dist/block-library/style.min.css' type='text/css' media='all' />
<link rel='stylesheet' id='font-awesome-css'  href='https://elab.fon.bg.ac.rs/wp-content/plugins/arconix-shortcodes/includes/css/font-awesome.min.css' type='text/css' media='all' />
```

Kvazi-strukturirani podaci (clickstream string)



https://www.google.com/search?sxsrf=ALeKk01849ABzicI0-Sd0e_Z8SD4YWuZLg%3A1615317472677&ei=4MlHYMTnKKyrrgTqqoJY&q=data+scinece&oq=data+scinece&gs_lcp=Cgdnd3Mtd2l6EAxQAFgAYJIUaABwAXgAgAGDAYgBgwGSAQMwLjGYAQCqAQdnd3Mtd2l6wAEB&sclient=gws-wiz&ved=0ahUKEwjE16HB9qPvAhWslYsKHWqVAAsQ4dUDCA0

Maksimiziranje znanja (*value*)

- Podaci se obrađuju i koriste u približno realnom vremenu (*velocity*).
- Kako bi se izbegla najveća zamka (ne samo NoP, već VI generalno), a to je pristrasnost u podacima *(bias*), neophodno je definisati odgovarajući koncept kvaliteta podataka *(veracity*) koji podrazumeva implementaciju metadata (*Data Quality Through Metadata*).

The Potential of Big Data

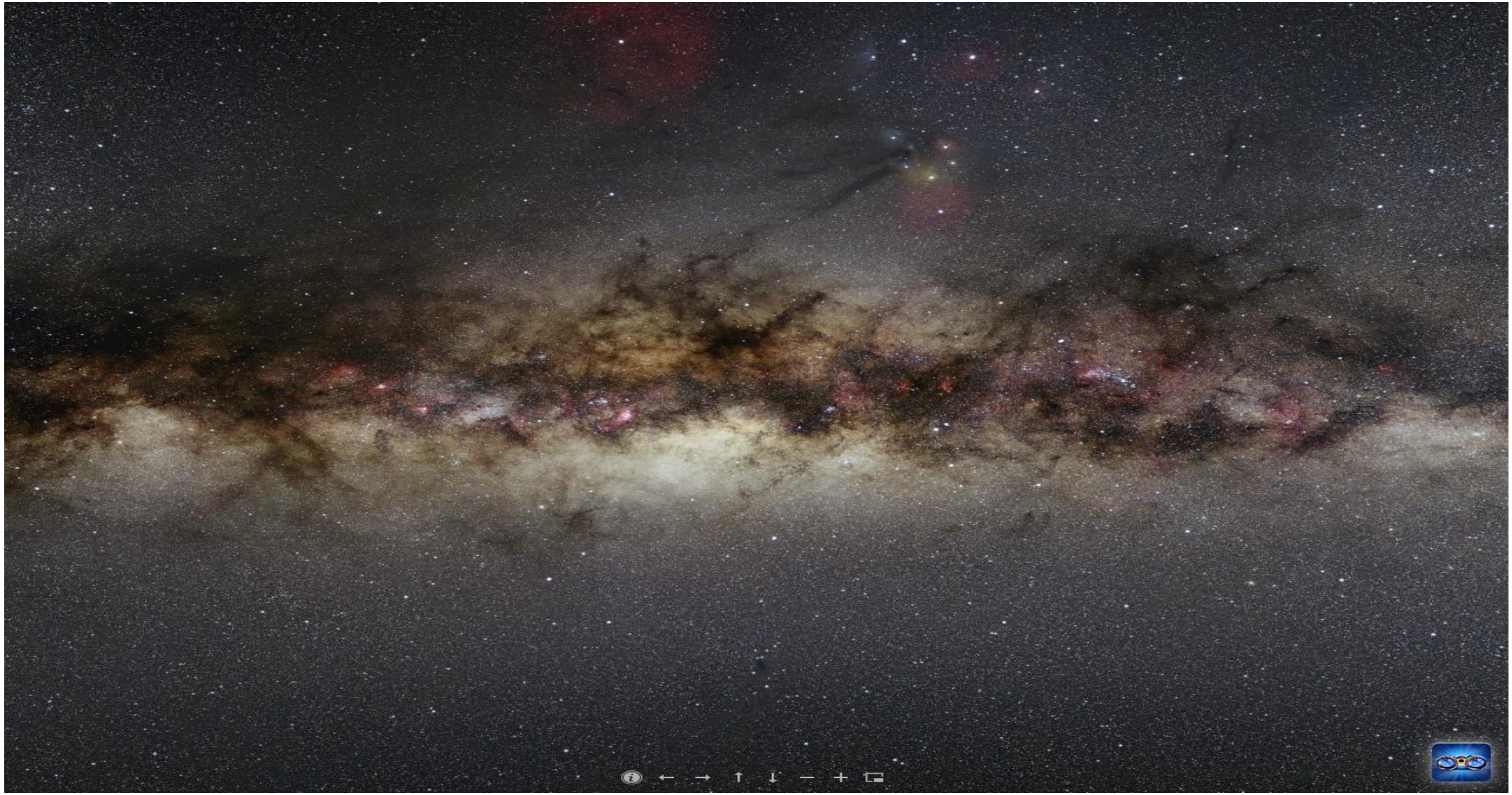
- Analogija: vožnja automobila noću vs. vožnju danju;

U prvom primeru vidite samo ono što osvetljavaju farovi, a u drugom čitav pejzaž.

→ **So be prepared for a real revolution**.

Rather than looking at segments, classifications, regions, groups, or other summary levels you'll have insights into ***all*** the individuals, ***all*** the products, ***all*** the parts, ***all*** the event, ***all*** the transactions, etc.

Big data bez NoP



The Photopic Sky Survey is a 5,000 megapixel photograph of the entire night sky stitched together from 37,440 exposures, http://www.skysurvey.org/survey/

Big Data Analytics 1. Tekst analitka (text mining)

- proces kojim se ekstrahuju (izdvajaju) informacije iz nestrukturiranih tekstualnih podataka iz Big Data izvora: društveni mediji, elektronska pošta, blogovi, onlajn forumi, odgovori na anketu, korporativni dokumenti, vesti i logovi kol centara (call center logs) i transformišu se u strukturirane informacije koje podržavaju donošenje odluka zasnovanih na dokazima (evidence-based decision-making) (Hurwitz & Nugent, 2013).

Tehnike:

- Obrada prirodnog jezika (Natural Language Processing),
- Tehnike mašinskog učenja (Machine Learning),
- Statističke metode,
- Veštačka inteligencija,
- Tehnike klasifikacije,
- Jezičko učenje (Linguistic Learning),
- Semantička analiza i
- Prediktivno modeliranje (Kaur & Deepti, 2016).

Big Data Analytics 2. Audio analitika

- metoda koja analizira i izdvaja informacije iz nestrukturiranih audio podataka.

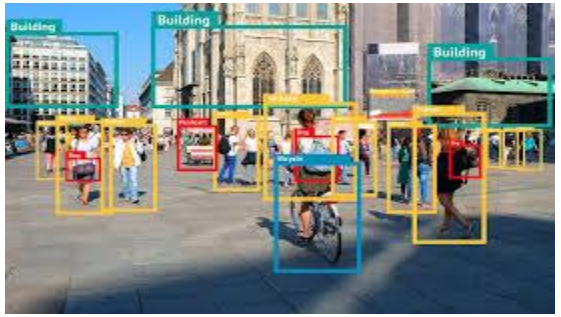
- Tehnološka rešenja zasnovana su na računarskoj lingvistici i mašinskom učenju, uz jaku regulatornu ekspertizu (Ernst & Young, 2016).

Dva osnovna tehnološka pristupa/tehnike:

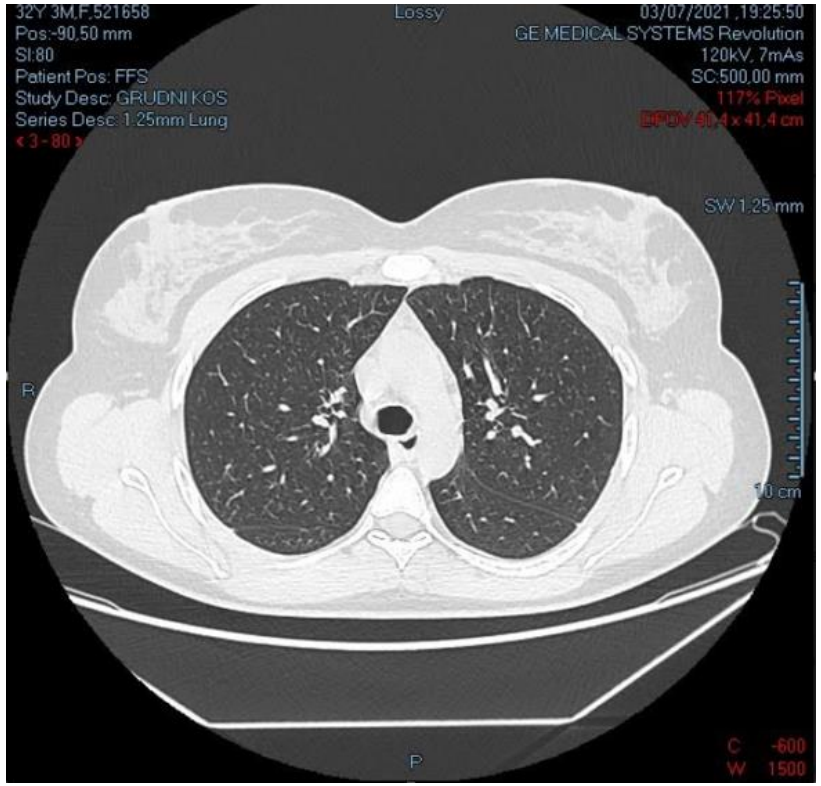
1. Pristup **zasnovan na transkriptu** - prepoznavanje kontinuiranog govora sa velikim vokabularom (large-vocabulary continuous speech recognition - LVCSR) i
2. **Fonetski pristup** (Gandomi & Haider, 2015).

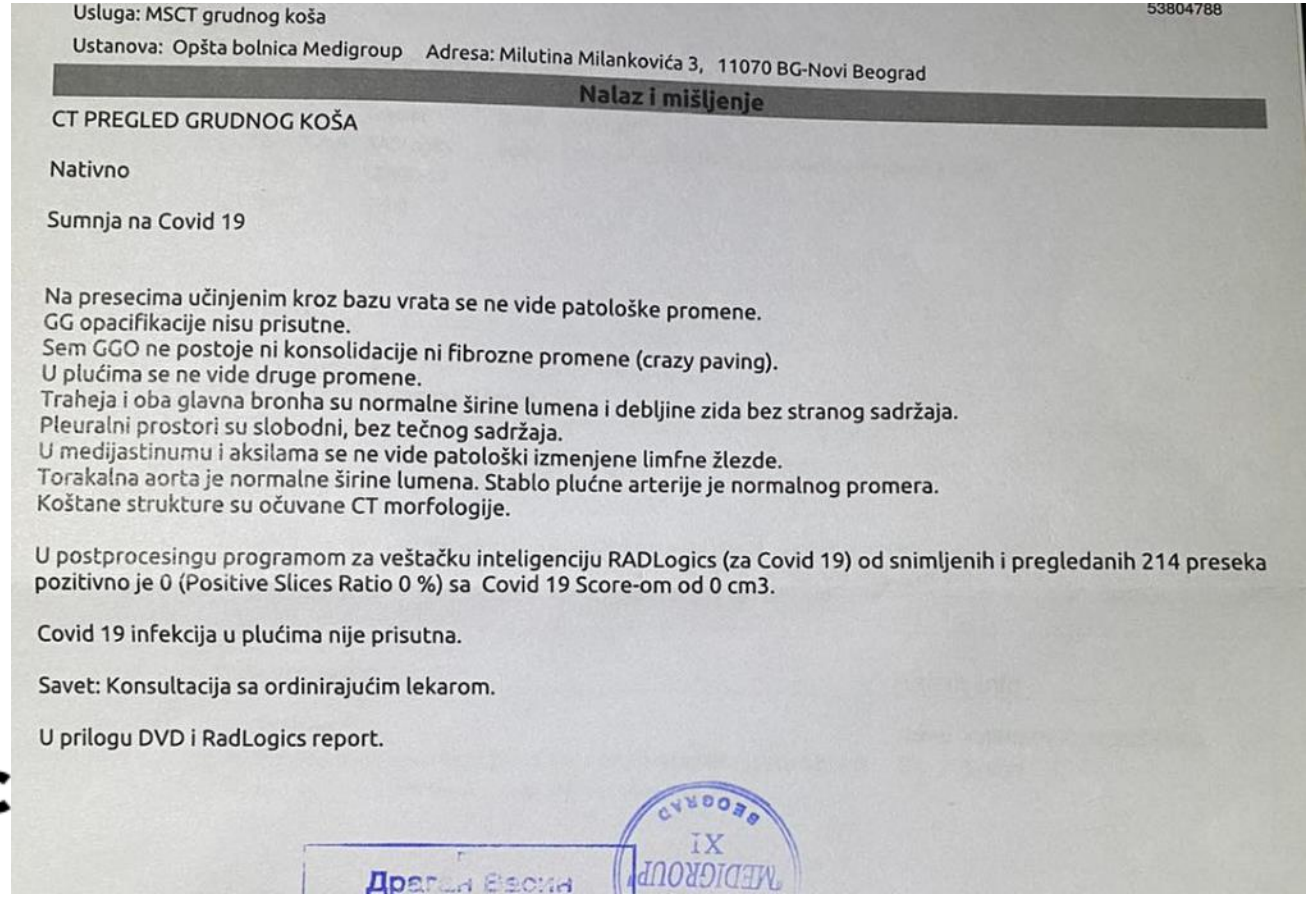
Big Data Analytics 3. Video analitika

- uključuje različite tehnike za kontrolu, analizu i ekstrakciju informacije iz nestrukturiranih video podataka (video streams).
- vrhunski automatizovani sistemi za video analizu integrišu i uče od svih dostupnih senzora i od svih prethodnih informacija o sceni, ciljevima i očekivanim ponašanjima kako bi dodatno poboljšali efikasnost. Ovo omogućava automatizaciju rešenja za video analizu za nove domene (Hakeem, et al., 2012)
- pametni sistemi mogu da prikupljaju demografske informacije o pojedincima, kao što su starost, pol i etnička pripadnost
- Automatsko video indeksiranje i pronalaženje objekata može se izvršiti na osnovu različitih nivoa informacija dostupnih u video sadržaju, uključujući metapodatke, zvučni zapis, transkripte i vizualni sadržaj



CS: Image analysis





Usluga: MSCT grudnog koša 53804788

Ustanova: Opšta bolnica Medigroup Adresa: Milutina Milankovića 3, 11070 BG-Novi Beograd

Nalaz i mišljenje

CT PREGLED GRUDNOG KOŠA

Nativno

Sumnja na Covid 19

Na presecima učinjenim kroz bazu vrata se ne vide patološke promene.
GG opacifikacije nisu prisutne.
Sem GGO ne postoje ni konsolidacije ni fibrozne promene (crazy paving).
U plućima se ne vide druge promene.
Traheja i oba glavna bronha su normalne širine lumena i debljine zida bez stranog sadržaja.
Pleuralni prostori su slobodni, bez tečnog sadržaja.
U medijastinumu i aksilama se ne vide patološki izmenjene limfne žlezde.
Torakalna aorta je normalne širine lumena. Stablo plućne arterije je normalnog promera.
Koštane strukture su očuvane CT morfologije.

U postprocesingu programom za veštačku inteligenciju RADLogics (za Covid 19) od snimljenih i pregledanih 214 preseka pozitivno je 0 (Positive Slices Ratio 0 %) sa Covid 19 Score-om od 0 cm3.

Covid 19 infekcija u plućima nije prisutna.

Savet: Konsultacija sa ordinirajućim lekarom.

U prilogu DVD i RadLogics report.

MEDIGROUP XI BEOGRAD

RADLogic

RADLogics develops AI-Powered solutions that support image analysis to improve radiologists' productivity while enhancing patient outcomes. Based in New York, NY and Tel Aviv, Israel, we are one of the pioneers in using AI & machine learning image analysis and advanced big data analytics to search and analyze imaging data to help reduce diagnostics turnaround time from hours to minutes by automating detection and report generation functions. Our patented AI medical image analysis platform enables rapid development and deployment of AI algorithms, and provides seamless integration into existing and customized radiology workflows.

https://www.radlogics.com/

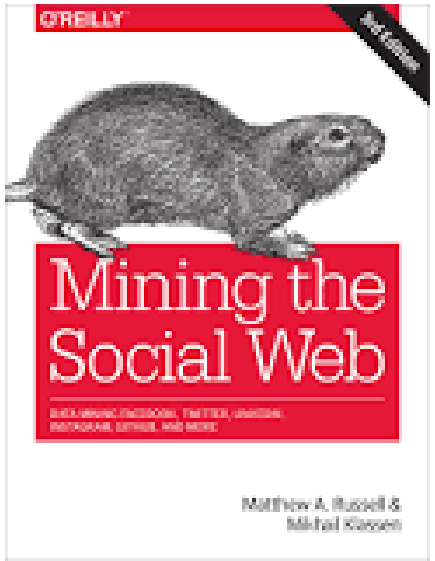
Big Data Analytics 4. Analitika društvenih medija

- analiza strukturiranih i nestrukturiranih podataka iz kanala društvenih medija koji se generišu iz dva osnovna izvora:

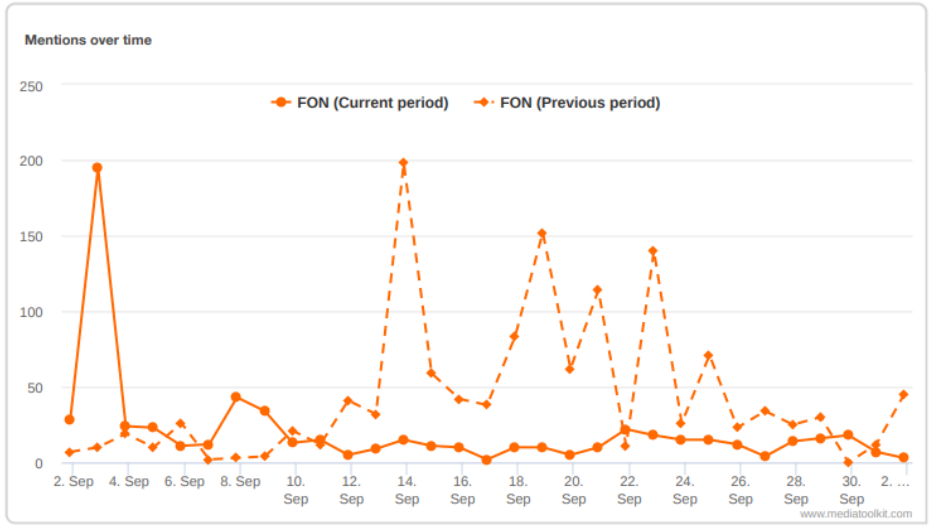
1. Podaci i informacije generisane od strane korisnika: sentimenti (osećanja, stavovi) fotografije, video zapisi i sl.) i
2. Podaci i informacije koje nastaju kao rezultat interakcije između mrežnih entiteta koje čine ljudi i organizacije.

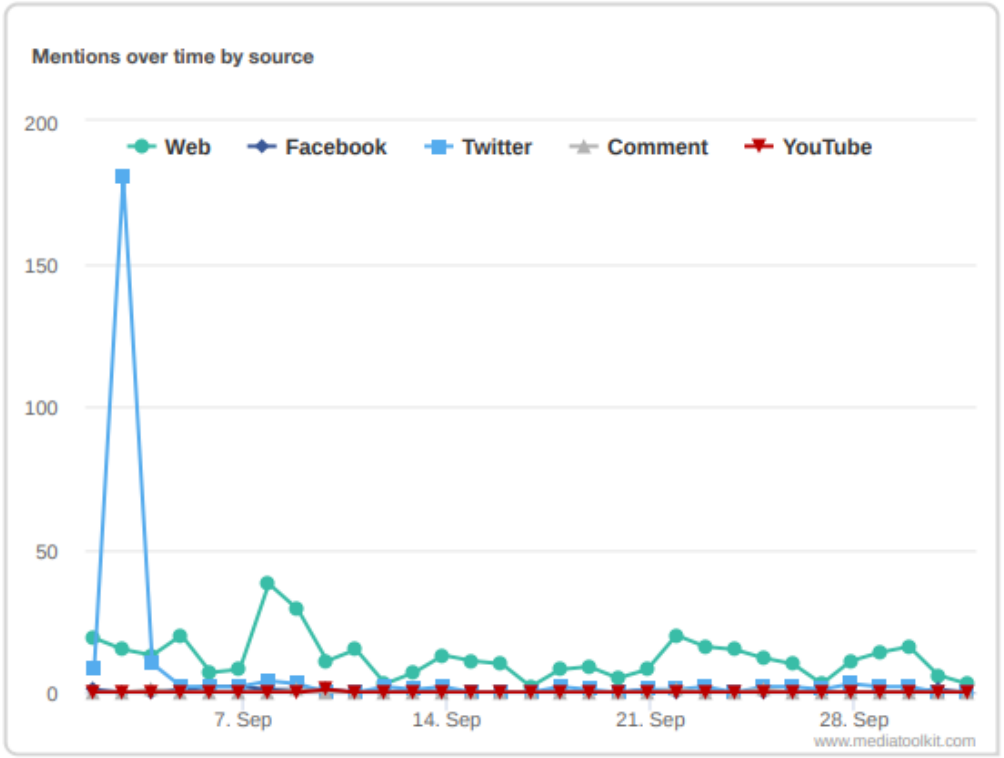
Tehnike analitike društvenih medija (Batrinca & Treleaven (2015):

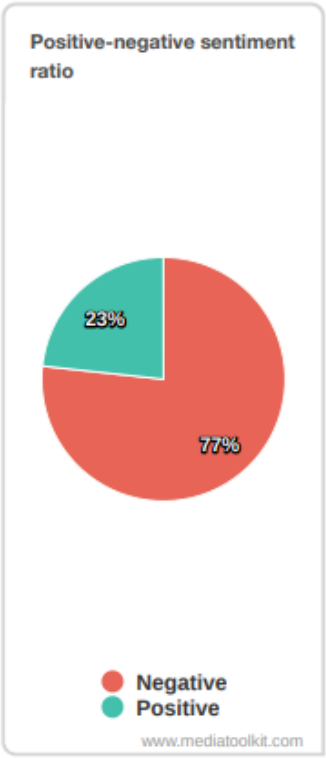
- Obrada prirodnih jezika (Natural language processing—NLP),
- Analitika vesti (News analytics),
- Istraživanje javnog mnjenja (Opinion mining),
- Ekstrakcija (Scraping) društvenih medija,
- Sentiment analiza (Sentiment analysis) i
- Tekstualna analitika (Text analytics).

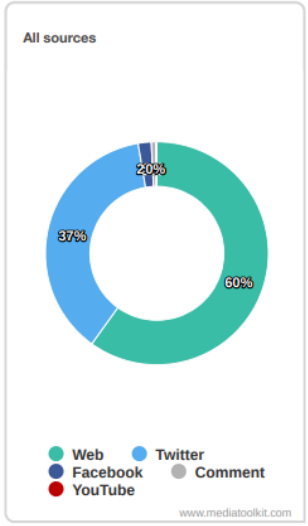


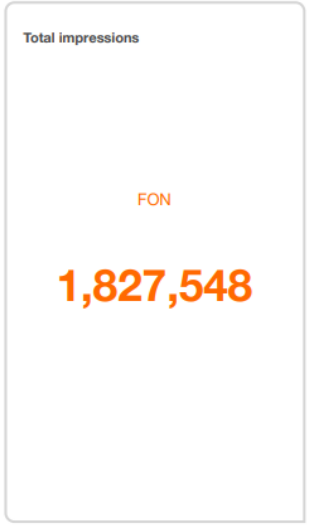
CS: Tekst analitika (text mining)











Top languages

SERBIAN	629
BOSNIAN	271
UNDEFINED	234
CROATIAN	43
SH	2
SLOVENE	2
CZECH	1
MACEDONIAN	1

Sensitivity Analysis

Python:

SALib

https://salib.readthedocs.io/en/latest/

Installing Prerequisite Software

SALib requires NumPy, SciPy, and matplotlib installed on your computer. Using pip, these libraries can be installed with the following command:

```
pip install numpy
pip install scipy
pip install matplotlib
```

The packages are normally included with most Python bundles, such as Anaconda and Canopy. In any case, they are installed automatically when using pip or setuptools to install SALib.

R:

Sensemakr

https://cran.r-project.org/web/packages/sensemakr/vignettes/sensemakr.html

```
# loads package
library(sensemakr)

# loads data
data("darfur")
```

Big Data Analytics 5. Prediktivna anlitika

- obuhvata različite modele koje predviđaju buduće ishode na osnovu istorijskih i trenutno raspoloživih podataka, otkrivanjem obrazaca i relacija u podacima, za razliku od eksplanatornih modela koji se koriste za testiranje kauzalnosti (Shmueli & Koppius, 2011).

U zavisnosti od tipova raspoloživih podataka i zadatih ciljeva istraživanja, u prediktivnoj analitici se javljaju dve različite paradigme:

- Učenje pod nadzorom (Supervised Learning)
- Učenje bez nadzora (Unsupervised Learning)

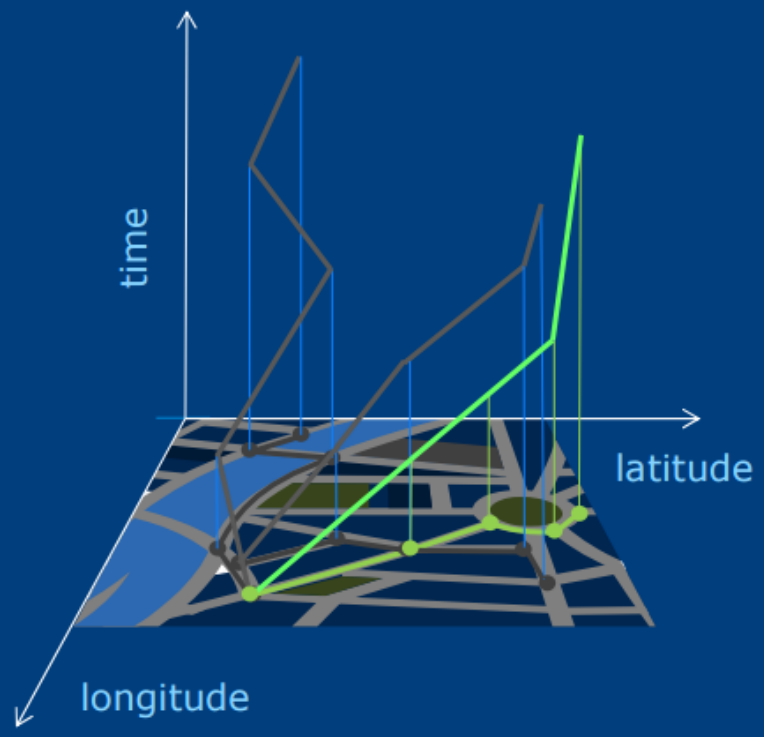
CS: Transport & Logistics - Challenge:

- Prediktivna analitika se može koristiti za omogućavanje tranzitnih operatora i menadžera da brzo donose odluke o prilagođavanju vozila ili usluga u skoro realnom vremenu
- Trenutno približno 20% teških teretnih vozila u Evropi vozi prazno (Eurostat, 2020) https://ec.europa.eu/transport/sites/transport/files/legislation/svd20200331.pdf

- Danas se 1,5% BDP-a EU godišnje posvećuje rešavanju problema gužve u saobraćaju

“We know where you’ve been”

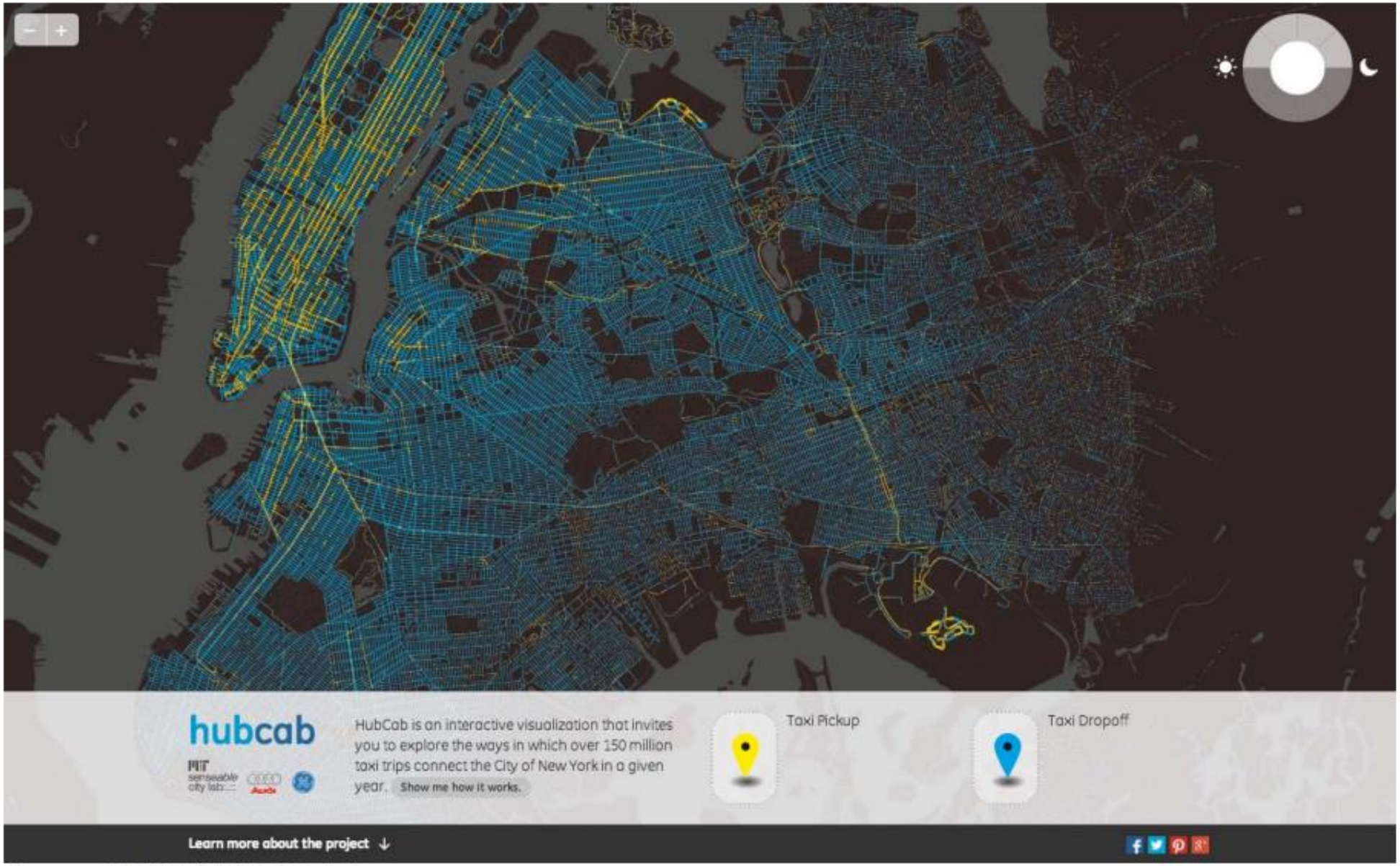
Individuals’ time-space trajectories are powerful identifiers



People’s patterns of movement in space and time are repetitive and predictable. These trajectories are powerful identifiers – like fingerprints

CS: Prediktivna anlitika

Pick-up and drop-off points of all 170 million taxi trips over a year in New York City



Source: MIT SENSEable City Lab.

Data Science vs. Machine Learning (ML)

- Mašinsko učenja (ML) **je funkcija orijentisana na proizvod** čija je primarna uloga izgradnja konkretnog algoritma koji odgovara specifičnim zahtevima – najčešće: *automatsko predviđanje* na osnovu podataka.
- NoP je **strateški orijentisana funkcija** čija je primarna uloga stvaranje dodatne vrednosti, koristeći naučne metode, vođene podacima.

→ Naučnik za podatke (NzP) je odgovoran za ostvarivanje **strateških ciljeva**; Specijalista za mašinsko učenje ima više **taktičku ulogu**.

- Pradyumna S. Upadrashta. Perpetual learner, teacher, practitioner (15+ years)

10 Machine Learning Methods that Every Data Scientist Should Know

- **Regression (LRM)**
- **Classification (logistic regression)**
- **Clustering (K-Means)**
- **Dimensionality Reduction (PCA)**
- **Ensemble Methods**
- **Neural Nets and Deep Learning**
- **Transfer Learning**
- **Reinforcement Learning**
- **Natural Language Processing**
- **Word Embeddings**

https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

The Machine Learning Revolution

- Neuronske mreže su samo jedan primer algoritma mašinskog učenja (ML)

- Deep Neural Networks are now exciting the whole of the IT industry since they enable us to:
  - Build computing systems that improve with experience
  - Solve extremely hard problems
  - Extract more value from Big Data
  - Approach human intelligence e.g. natural language processing

The Fourth Paradigm
and Big Scientific Data

Professor Tony Hey
Chief Data Scientist
Rutherford Appleton Laboratory
Science and Technology Facilities Council
tony.hey@stfc.ac.uk

https://indico.cern.ch/event/609040/contributions/2455548/attachments/1468477/2271222/Tony_Hey_-_The_Fourth_Paradigm_ATTRACT_Talk_-_May_17.pdf

Data Science vs. Data engineering (DE)

Data engineering - inženjerski domen posvećen izgradnji i održavanju sistema podataka za prevazilaženje uskih grla u obradi podataka i problema sa rukovanjem podacima koji nastaju usled velikog obima, brzine i raznolikosti velikih podataka

→ kako im samo ime kaže, inženjeri, koji se bave infrastrukturom DS

Data engineers:

- Koriste računarske veštine i softverski inženjering za dizajniranje sistema i rešavanje problema upravljanja velikim skupovima podataka.
- Imaju iskustva u radu sa i dizajniranjem okvira za obradu u realnom vremenu i platformama za masovne paralelne obrada (massively parallel processing MPP) kao i RDBMS.
- Koriste programske jezike Java, C ++, Scala i/ili Python.
- Znaju kako da primene Hadoop MapReduce ili Spark za rukovanje, obradu i pročišćavanje velikih podataka u skupove podataka kojima se lakše upravlja.

Ključna razlika: naučnici za podatke takođe moraju da imaju stručnost u određenim oblastima u kojima rade (subject-matter expertise)

From a hardware-based point of view, data analysis consists of three components: the processor to perform the calculations, the storage to store the (manipulated) data and a system that transfers data between the two.

https://www2.deloitte.com/content/dam/Deloitte/de/Documents/technology-media-telecommunications/TMT_Studie_In_Memory_Computing.pdf

Data engineering – konkretno

- Budući da 3V onemogućavaju implementaciju tradicionalnih relacionih sistema upravljanja bazama podataka i tradicionalnu statističku obradu podataka, inženjeri za podatke imaju zadatak da prevaziđu limite Big Data
- Kako to rade?
  - **Hadoop**, za svođenje velikih u manje skupove podataka koje naučnici za podataka mogu procesirati

Hadoop ecosystem :

- HDFS (for data storage),
- MapReduce (for bulk data processing),
- **Spark (for real-time data processing)**, and
- YARN (for resource management).

Alternativna: R, Python, za obradu velikih skupova podataka (bez inženjera za podatke)

Top Alternatives to Hadoop HDFS

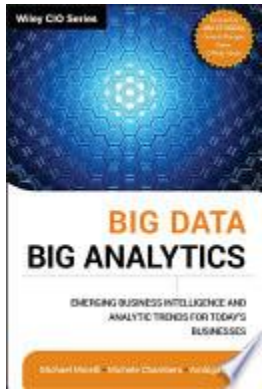
- Databricks
- Google BigQuery
- Cloudera
- Hortonworks Data Platform
- Microsoft SQL
- Snowflake
- Qubole

Izvor: https://www.g2.com/products/hadoop-hdfs/competitors/alternatives

Analogija...

- “Astronomi su zaista vešti u izvlačenju struktuiranih podataka iz slika. Oni posmatraju *Sky Survey* kao način prikupljanja slojevitih podataka o milijardama zvezda i drugih vasionskih tela.
- Ovaj način razmišljanja veoma je sličan pristupu koji menadžeri imaju prema svojim klijentima.
  - Naime, menadžeri o svojim klijentima znaju u suštini veoma malo, a i ta saznanja su nekompletna, van konteksta i potencijalno netačna – isti princip kao i sa zvezdama.

- Kada je astronomima neophodno da sagledaju zvezde temeljnije, oni posežu za teleskopima sa znatno višom rezolucijom i fokusiraju se na manji segment neba. Na taj način, dosta objekata koji su bili jedva prepoznatljivi u glavnom delu istraživanja (upotrebom niže revoluciji i globalnije slike) sada postaju vidljivo jasniji. Tada zapravo možete shvatiti da li ste posmatrali zvezde, galaksije ili druga nebeskih tela“.



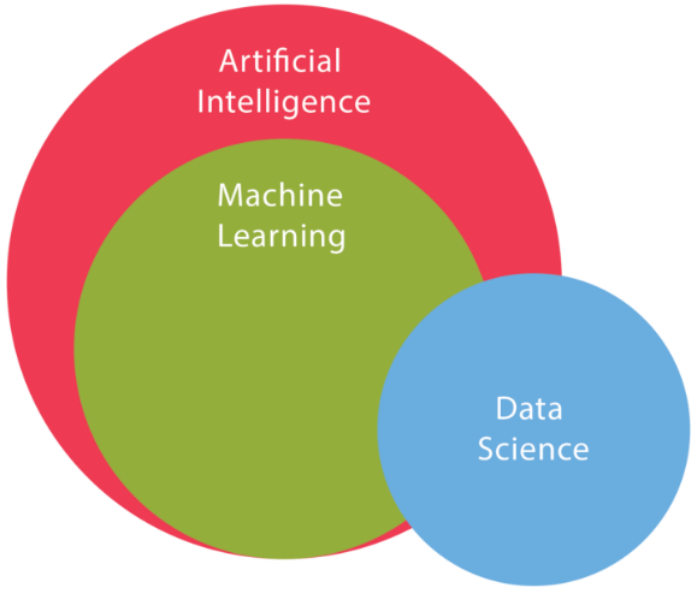
BIG DATA IS A HUMAN RIGHTS ISSUE

- **Never analyze personally identifiable information**
- **Never analyze confidential data**
- Never seek to **re-identify individuals**

GDPR

Data Science vs. Artifical Intelligence

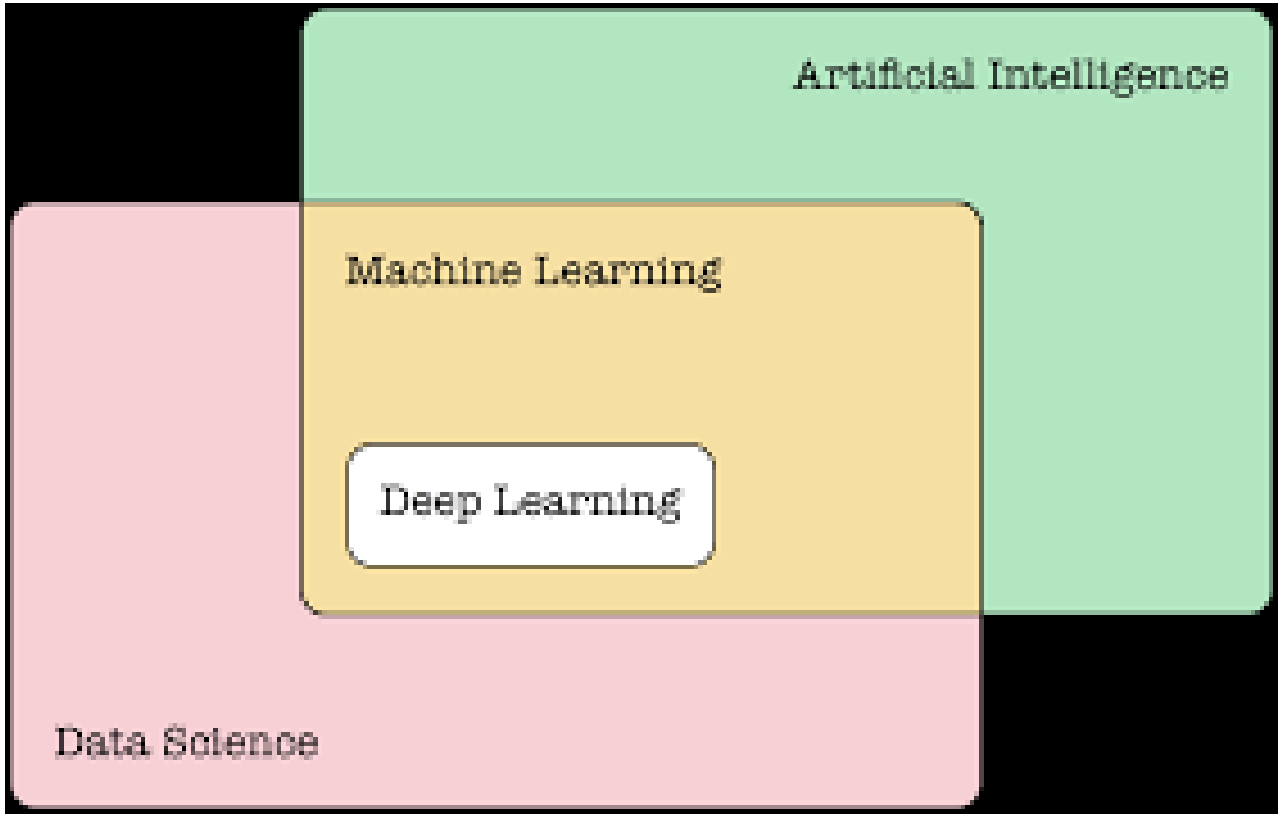
- **Veštačka inteligencija (VI) je širi pojam od NoP:** uključuje sve što omogućava računarima da uče i nauče kako da rešavaju probleme i donose pametne odluke.



(Nigam, 2019)

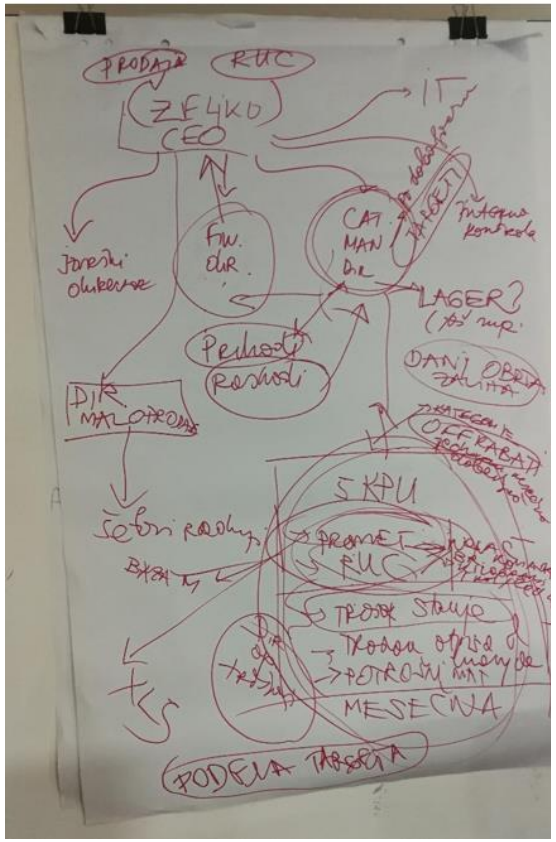
DS vs. AI vs. ML vs. DL

- **AI** involves machines that can perform tasks that are characteristic of human intelligence. While this is rather general, it includes things like planning, understanding language, recognizing objects and sounds, learning, and problem solving (by John McCarthy, 1956)
- **Machine learning** is simply a way of achieving AI. (Arthur Samuel, 1959
- **Deep learning** is one of many approaches to machine learning



Nauka o podacima – Data Science

Spremni za definisanje: ništa i sve od navedenog.



Nauka o podacima (NoP) – Data Science

: **multidisciplinarna naučna oblast koja kombinuje prirodno-matematičke i tehničko-tehnološke nauke da bi uz domensku ekspertizu iz podataka različite veličine i formata izvukla maksimalno znanje.**

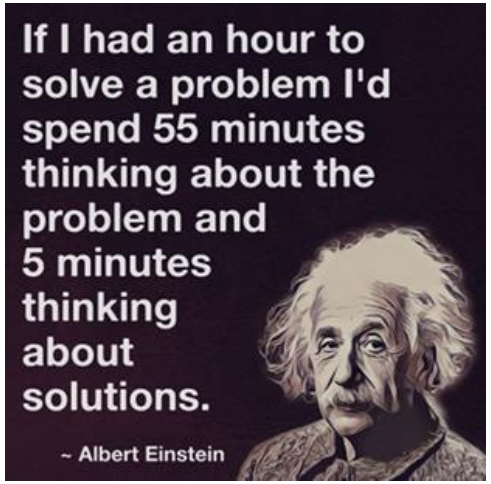
- skup komplementarnih naučnih disciplina sa sličnim metodološkim osnovama, perspektivama i ciljevima, koje objedinjuje zajednička **misij**a: kontinuirano uvećanje baze znanja.

... **vizija** bi trebalo da sledi... za dobrobit čovečanstva i očuvanje planete Zemlje.

(Vukmirović, D. (2020). DATA SCIENCE: SCIENCE OR BUZZWORD. Seminar za računarstvo i primenjenu matematiku. Matematički Institut SANU, Beograd)

Domenska ekspertiza

: ekspertska znanja koja pokrivaju predmet i ciljeve određenog istraživanja za koje se NoP koristi, a potiču iz različitih naučnih oblasti, uključujući medicinske, biotehničke, društvene i humanističke nauke (npr. medicina, biohemija, farmacija, ekonomija, socilogija...).



Veličina (*volume*) podataka ne igra značajnu ulogu u implementaciji NoP-a

- **Procesorka snaga**
- **Alati** (besplatni ili jeftini)
- **Pristup podacima** (*open data*)
- **Cloud** (tehnologije i rešenja)
- **Nove naučne discipline** (*Science 4.0*)

(Pavičić, 2019)

Data Science u organizacijama

(Cantrell, 2018)

There's no single blueprint for beginning a DS project

never mind ensuring a successful one but these 10 questions will help guide you to success:

1. Is this your organization's first attempt at a DS project?
2. What business problem do you think you're trying to solve?
3. What types and sources of data are available to you?
4. What types and sources of data are you allowed to use?
5. What is the quality of your organization's data?
6. What tools are available to extract, clean, analyze and present the data?
7. Do your employees possess the right skills to work on the data analytics project?
8. What will be done with the results of your analysis?
9. What types of resistance can you expect?
10. What are the costs of inaction?

CS: MIP - model inteligentnog preduzeća

SYM-OP-IS 2019

✓ prototip modela za implementaciju NoP i VI u malim i srednjim preduzećima u cilju poboljšanja performansi poslovanja

- **Polazna pretpostavka** - osnovni cilj kompanije predstavlja povećanje korporativne inteligencije koja se definiše kao funkcija znanja baziranog na podacima:

IQcorp = f (podatak, znanje)

- bazira se na Big Data tehnologijama i NoP

MIP= f (podatak, model, reinženjering)

Koraci u Implementacija MIP-a

Redni broj	*Korak*	*Definicija*	*Odgovornost*
1	Izbor inicijalnog poslovnog problema/mogućnosti za inicijalnu implementaciju.	Poslovni proces	Top menadžment
2	Kreiranje internog tima za implementaciju u sastavu: Menadžment srednjeg nivoa zadužen za poslovni proces, IT stručnjaci, stručnjaci za analitiku (podatke)	Interni tim za implementaciju	Menadžment srednjeg nivoa
3	Odluka o načinu implementacije: unutar kompanije ili van (*outsorce*).	Način implementacije	Top menadžment
4	Izbor odgovarajuće platforme i modela poslovne analitike.	Izabrano rešenje	Interni tim za implementaciju
5	Implementacija izabranog rešenja	Implementacija	Interni tim za implementaciju

CS: Implementacija MIP: Tekijanka - Fruit and vegetable supply

Implementacija: Kombinovana metoda - interna uz *outsorce* glavnog analitičara i naučnika za podatke

Tim za iplementaciju:

1. Domain/Business expert
2. Data scientist
3. Computer Science/IT expert
4. Programmer – Software Developer
5. Chief Analytics Officer

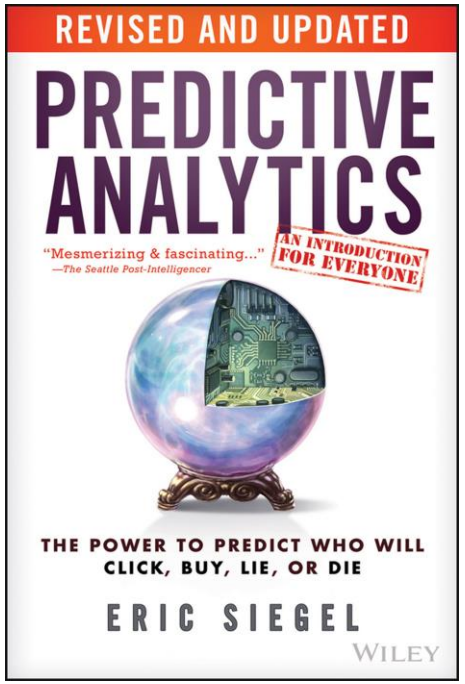
Prediktivna analitika

Tradicionalni modeli (predviđanje) planiranje zasnovani su na prethodnoj potražnji

- Modeli osetljivosti potražnje (*Demand sensing* models) : upotreba nauke o podacima i podataka u realnom vremenu za izračunavanje prognoza se bazira na upotrebi ažuriranih podataka (zasnovanih na čitanju dodatnih ulaza i signala iz unutrašnjeg i spoljnog okruženja) za stvaranje preciznijih, kratkoročnih prognoza
- Superforecasting - „Predviđanje na osnovu potražnje“–“*Demand driven forecasting*” budući koncept upravljanja potražnjom. Uspešna prognoza uzima u obzir podatke o stvarnoj (željenoj) potražnji kupaca za proizvodom

Polazna osnova

- Započeli smo sa izgradnjom kombinovanog modela (C MODEL) koji se primenjuje za nivo centralnog magacina (CM nivo), kombinovanjem tradicionalnog modeliranja (TM) sa elementima modela osetljivosti na potražnju (Demand Sensing models - DeSe)



Razvoj modela

- Za komponentu TM planirali smo da koristimo podatke iz prethodnih 5 godina i primenimo tradicionalne statističke metode: analizu vremenskih serija (analiza komponenata trenda metodom pokretnog proseka, itd.) I deskriptivnu analizu.
- DeSe komponenta odnosi se na domen modela koji nije obuhvaćen prethodnim statističkim metodama. Uključuje uticaj promocije, pojavu novog proizvoda, uticaj makro okruženja (novi konkurenti (Lidl), društveno-ekonomski trendovi na tržištu itd.)

Inicijalna matematička formulacija modela

C MODEL = (TM + DeSe)*ew

Pri čemu je ***ew*** - ponder (weighting component) predviđanja koja se izvodi na osnovu unosa stručnog domenskog znanja (ew) iz internih izvora kompanije (marketing, prodaja, nabavka, logistika, finansije).

- Trebalo je odrediti period predviđanja (dnevno, nedeljno, mesečno, između dve isporuke) prema utvrđenoj jedinici mere

Procedura

- postepena primena modela u cilju procene prognoze, kroz tri koraka:

1. Model sadrži samo TM komponentu: **C MODEL (1) = TM2.**
2. Ponderisanje modela): **C MODEL (2) = TM * ew**
3. Uključivanje DeSe komponente**: C MODEL = (TM + DeSe) * ew**

Evaluacija modela

- Bilo je predviđeno da se procena vrši a posteriori, upoređivanjem podataka iz prethodnog perioda (t-2) i mere se razlika u izlaznom učinku u odnosu na stvarne poslovne rezultate u periodu (t-1).

- **... I onda...**

Project evaluation

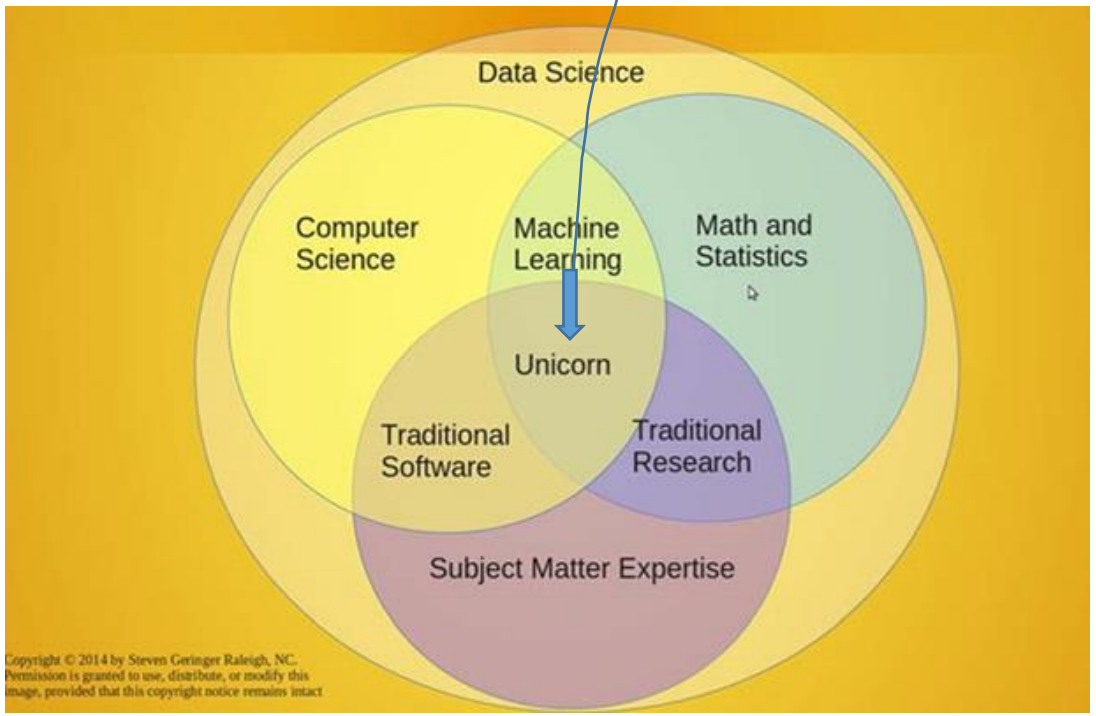
Ten red flags signaling DS project will fail

1. The executive team doesn't have a clear vision for its advanced-analytics programs
2. No one has determined the value that the initial use cases can deliver in the first year
3. There's no analytics strategy beyond a few use cases
4. Analytics roles—present and future—are poorly defined
5. The organization lacks analytics translators
6. Analytics capabilities are isolated from the business, resulting in an ineffective analytics organization structure
7. Costly data-cleansing efforts are started en masse
8. Analytics platforms aren't built to purpose
9. Nobody knows the quantitative impact that analytics is providing
10. No one is hyper focused on identifying potential ethical, social, and regulatory implications of analytics initiatives

Naučnik za podatke (NzP) - Data Scientist

Skills/knowledge needs:

- Mathematics (Linear Algebra and Calculus)
- Statistics
- Programming (Python, R, Julia, Scala, etc.)
- SQL
- Data Wrangling and preprocessing
- Data visualization
- Supervised Learning
- Unsupervised Learning
- Deep Learning
- Big Data platform (Spark or Hadoop)
- Cloud computing
- …



I ... *Soft skills*

„We often hear from employers (e.g., industry, governmental agencies, NGOs) that they wish the students they hire had better soft skills, such as **communication skills, interpersonal skills, stress management skills**, etc...“

Highlights of the National Academies Report on "Undergraduate Data Science: Opportunities and Options”

An interview with Laura Haas and Alfred Hero by Robert Lue

by Laura Haas, Alfred Hero, and Robert A. Lue

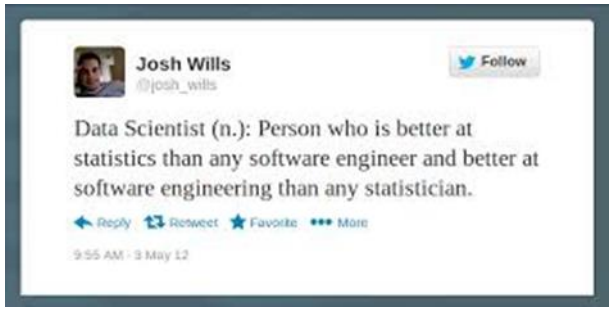
Published on Jul 02, 2019

NzP – post festum

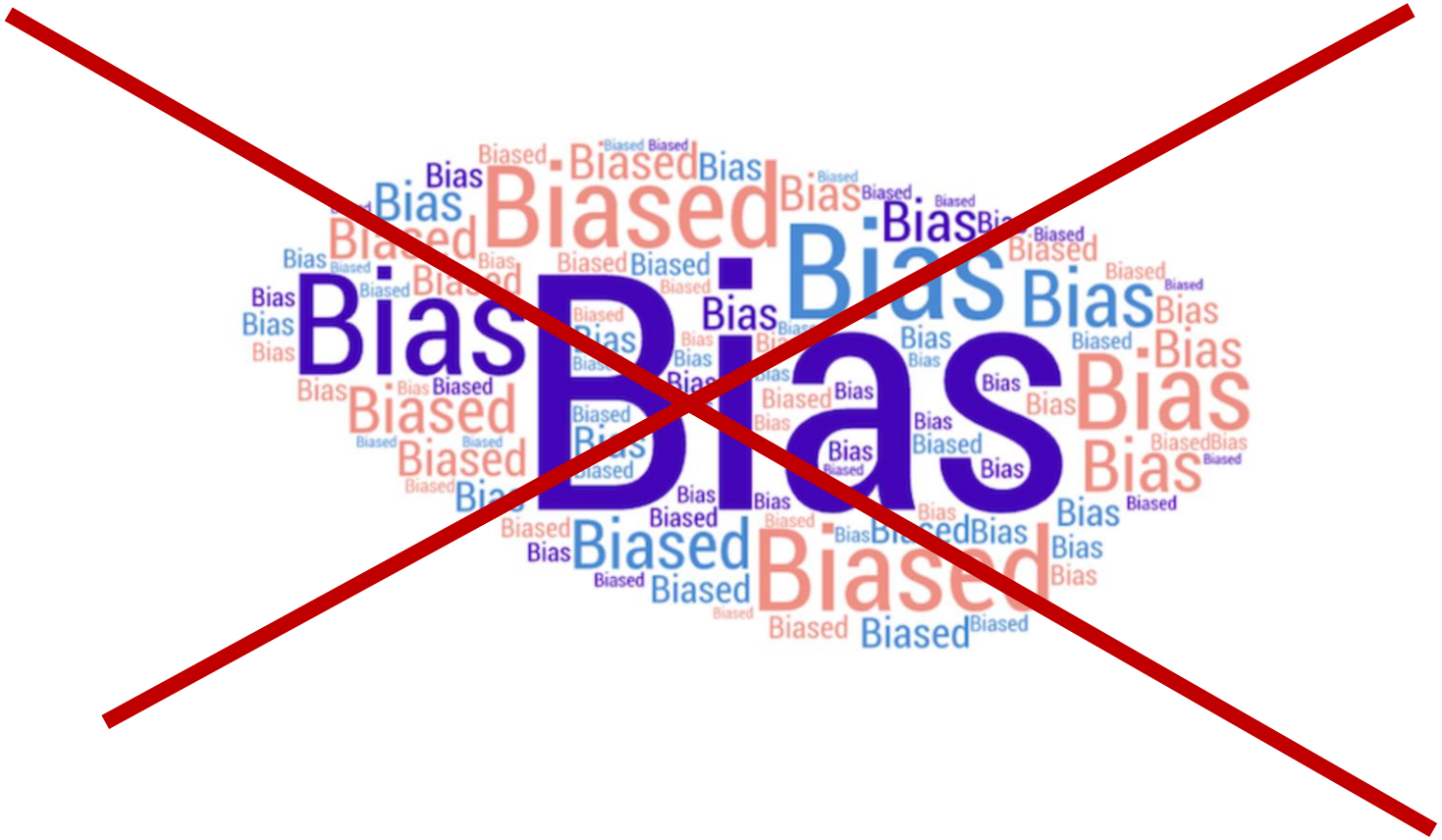
- Iluzorno je i nesvrsishodno očekivati data science uniqorn-a
- To svakako nije nemoguće, ali ne predstavlja pravilo, nego izuzetak (stat. outlier koji se eliminiše iz dalje analize

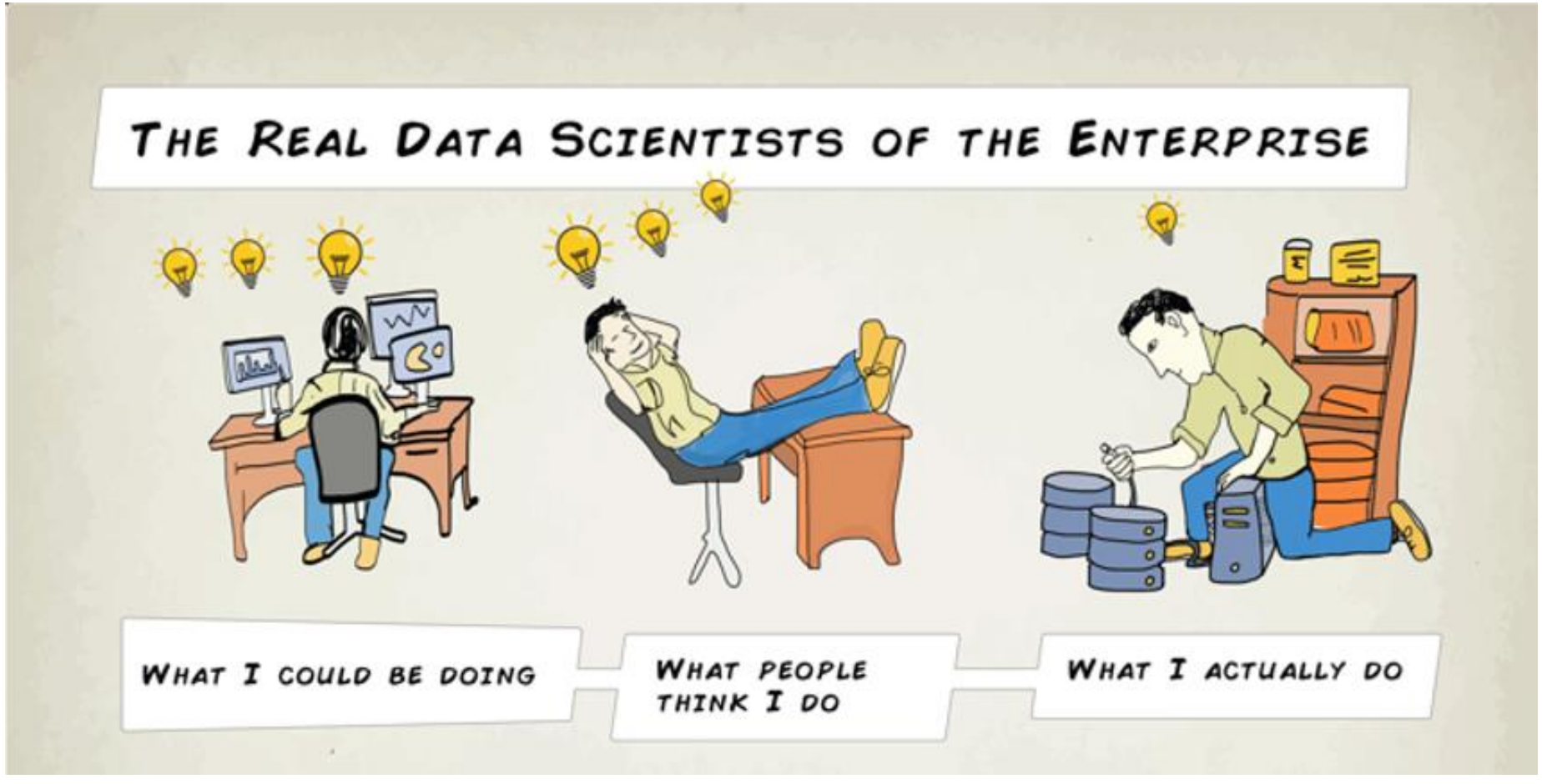
"Od početaka razvoja nauke, kada su naučnici bili individualci koji su se široko bavili sa nekoliko naučnih oblasti, došli smo u situaciju da naučnici moraju da specijaliziraju veoma uske naučne oblasti, a pri tome moraju da rešavaju interdisciplinarne i multidisciplinarne probleme, što se jedino postiže timskim radom".

- Rončević, S., M. Pavkov Hrvojević (2020).



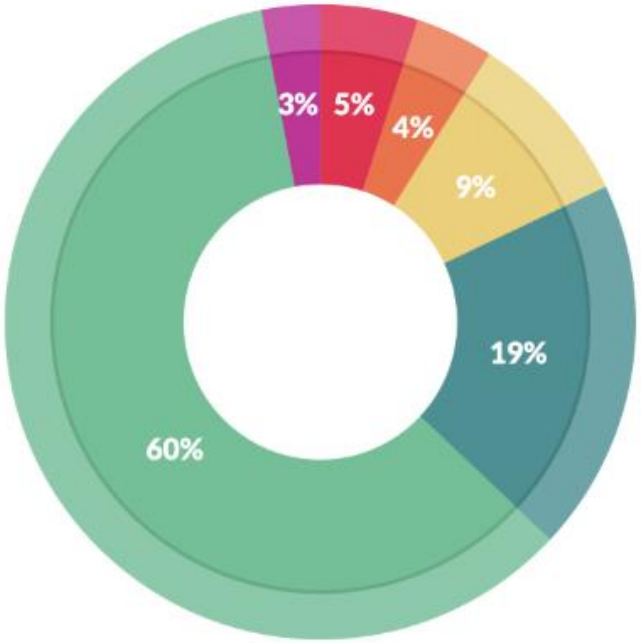
Domain/Business expert





https://www.tamr.com/blog/real-data-scientists-enterprise/

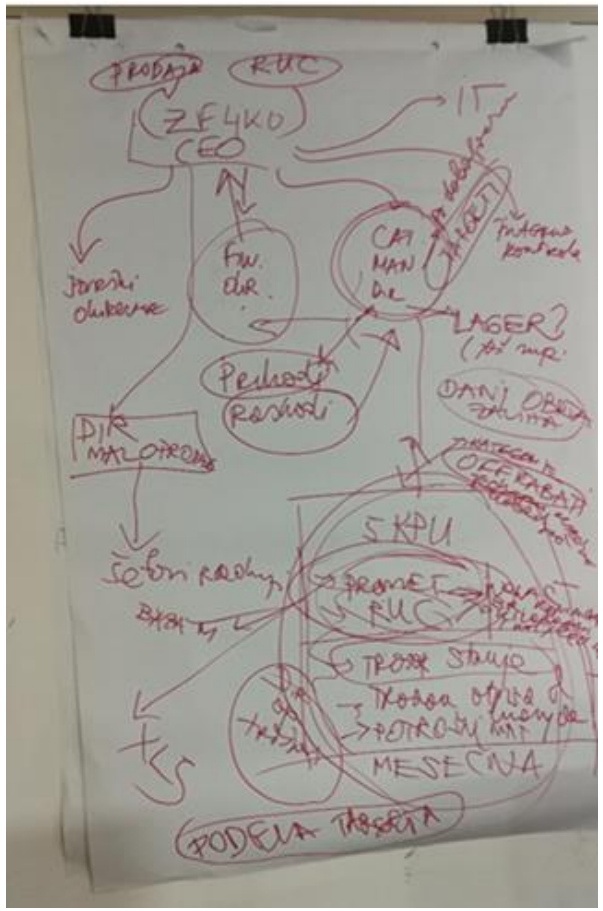
NzP: realnost



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

[taken from CrowdFlower]



Dobar NzP

- Koristi fleksibilan jezik za interpretaciju budućih pravaca akcije: Na primer: umesto tvrdog „Zaključujemo“, izneće zaključak u formi „Pitamo se da li...“.
- Neće koristiti termin „teškoće i problemi“, već „izazovi“
- Ponudiće oprez umesto prekomernog samopouzdanja kod donosioca odluka

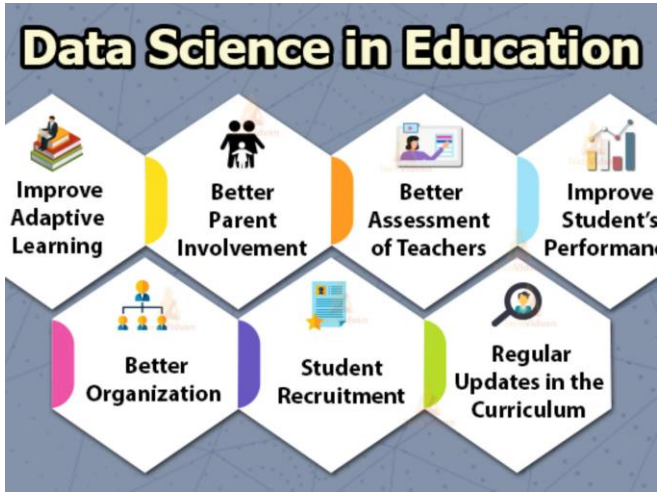
Pogled u napred

Edukacija

Kao što su nam potrebni lekari primarne zdravstvene zaštite i medicinski specijalisti, tako su nam potrebni i naučnici za podatke opšte prakse i odgovarajući specijalisti NoP-a kako bismo efikasno odgovorili na brojne izazove definisane misijom NoP.

Potrebno je pažljivo planiranje na nivou države: obrazovanje NzP ne može se obaviti na adekvatan način u kontekstu usko strukturiranih obrazovnih jedinica koje imaju ograničene resurse (kao i sposobnost da ih uvećaju).

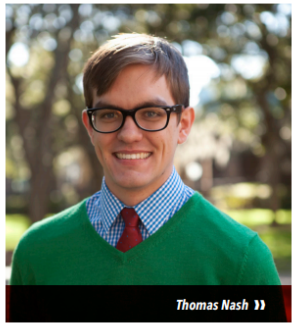
(Meng, 2019)



Edukacija

Početak:

- 2005. godine, koledž u Čarlstonu, Južna Karolina, započeo je četvorogodišnji dodiplomski program prediktivne analitike, mašinskog učenja i data mining-a.



DATA SCIENCE

IF YOU GREW UP WATCHING YOUR DAD DO cool stuff in IT like Thomas Nash, you might buy into data science early on. Or, you might just stumble into it. Either way, if you choose this major, you'll be getting involved in the country's first undergraduate program in this rapidly developing field.

Thomas had no background in computer science when he first heard about the College's data science program. Still, it grabbed him. "I enjoy math, problem solving and computers, so when I learned that this program combines those, I said 'This sounds awesome. I can take this major wherever I want. I can make it work for me.'"

Thomas took a programming class his first semester. "I was so nervous. I was thinking 'how am I going to make it in a degree where I have no idea what I'm doing.' But I realized that I could do this; I *could* learn new programming skills. I ended up loving that class, and discovered early on that this is what I want to do."

These days, Thomas doesn't just study data science, he's contributing to the field as well. "I work in the College's IT department as a student network engineer, and I also conduct research for the data science program, which focuses on developing and refining an open source software program called Learn2Mine. It's essentially a data mining teaching tool that helps introduce students to data mining techniques." His work has been partially supported by Boeing South Carolina, which named him a Boeing Scholar each of the past two years.

Thomas traveled to Lithuania to present Learn2Mine at a conference. Along with two other students and a professor, he also traveled to New York City for another conference. "We visited several tech companies there and came away with great insight into the industry. Because our professor is well connected, it was an exceptional networking opportunity. Now, whether I go on to work in cybersecurity, the aerospace industry or for a local startup, I know I'll have a strong background, and continue to have great support from my professors."

COLLEGE of CHARLESTON
SCHOOL OF SCIENCES AND MATHEMATICS

DEPARTMENT OF COMPUTER SCIENCE

PAUL ANDERSON
program director
843.953.8151
andersonpe2@cofc.edu
datascience.cofc.edu

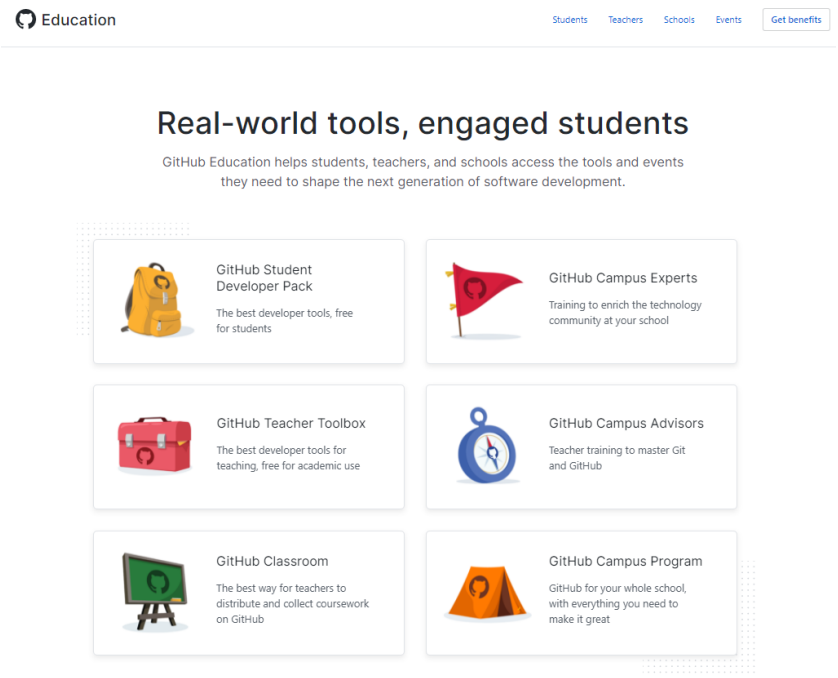
GRADUATES FROM OUR PROGRAM – THE FIRST OF ITS KIND IN THE U.S. FOR UNDERGRADUATES – ARE PREPARED FOR HIGH-PAYING JOBS AND GRADUATE PROGRAMS. THEY LEARN TO USE THE TOOLS AND PROBLEM-SOLVING SKILLS OF MATHEMATICS AND COMPUTER SCIENCE TO GATHER INFORMATION FROM LARGE, MULTIDIMENSIONAL DATA SETS, DATA STREAMS AND COMPLEX SYSTEMS. WHETHER YOU PLAN TO CRUNCH NUMBERS FOR THE SPORTS ANALYTICS, PRECISION MEDICINE OR ANALYZE USER DATA FOR GOOGLE, YOU'LL GET THE REQUIRED BACKGROUND HERE.

- WE OFFER 14 AREAS OF SPECIALIZATION.
- YOU CAN WORK IN FACULTY RESEARCH LABS.
- CAREER OPPORTUNITIES EXIST IN ALL AREAS OF INDUSTRY, GOVERNMENT AND BUSINESS.

Univerziteti - Svet

- Samo na GitHub platformi trenutno je registrovano 633 programa u oblasti *Data Science* (nauka o podacima), Poslovne analitika, *Big Data* (analitike) i srodnim poljima

Na dan 09.03.2021

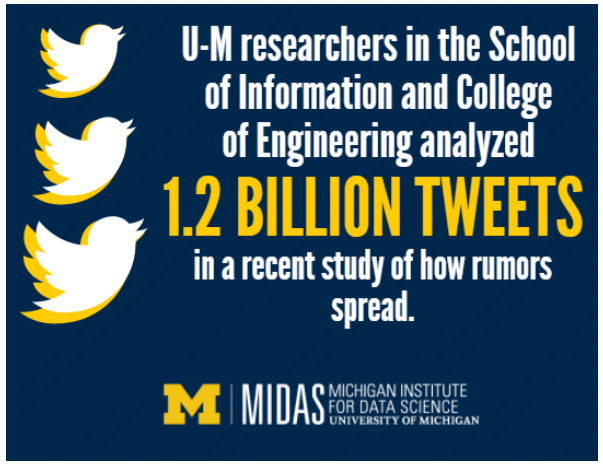


Data Science inicijative

$100M Data science initiative launched

- The University of Michigan plans to invest $100 million over the next five years (2016-2020) in a new Data Science Initiative (DSI) that will enhance opportunities for student and faculty researchers across the university to tap into the enormous potential of big data.

University of Michigan *2016 ANNUAL REPORT*



Data Science / Undergraduate Studies

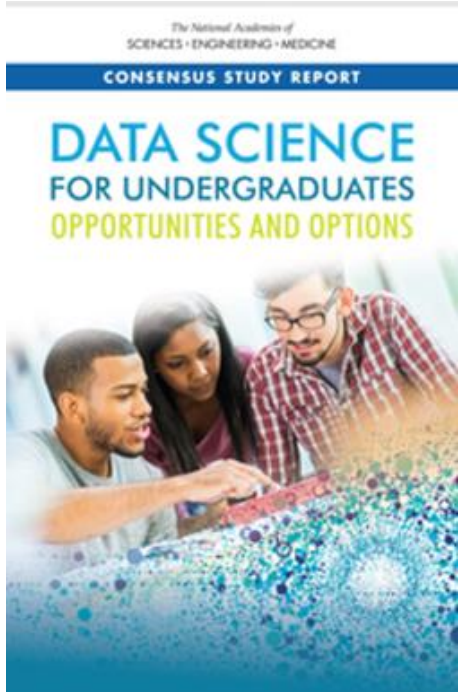
- 2016. godine *National Academies of Sciences, Engineering, and Medicine of the United States* osnovale su *Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective*.
- 2018. godine komitet je pripremio izveštaj (na 120 str.), u kome je izneo viziju dodiplomskog obrazovanja iz NoP kako bi:
  - „razjasnili akademskim institucijama, industriji i studentima šta je što bi naučnici za podatake u budućnosti trebalo da znaju,
  - pružili uvid u raznolikost postojećih akademskih programa i
  - pomogli akademskim institucijama da razumeju ključne aspekte u dizajniranju dodiplomskog programa.“

http://nap.edu/25104

Data acumen / Undergraduate Studies

The heart of what a practicing data scientist must know, and therefore, what a data science program must teach:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,
- Data description and visualization,
- Data modeling and assessment,
- Workflow and reproducibility,
- Communication and teamwork,
- Domain-specific considerations, and
- Ethical problem solving.



http://nap.edu/25104

CS: Data Science Undergraduate Studies

Data Science Major

- https://data.berkeley.edu/academics/data-science-undergraduate-studies/data-science-major

- DATA c104 Fall 2020 Syllabus:

https://data.berkeley.edu/academics/undergraduate-programs/data-science-offerings/data-104-human-contexts-and-ethics-data/data

Data Science minor

https://statistics.stanford.edu/data-science-minor

Data Science Inicijative – još…

Studenti: We are the Data Science Initiative of *Students of the University of Vienna*! Anyone and everyone interested in Data Science is welcome to participate in our ...

www.univie.ac.at › dsi-students

Organizacije

IEEE.org

The IEEE Signal Processing: Data Science in Signal Processing:

https://signalprocessingsociety.org/community-involvement/data-science-initiative

Biblioteke: News in Category: Data Science Initiative

https://www.library.ucsf.edu/topic/data-science-initiative/

Države

India: Data Science Research Initiative https://dst.gov.in/data-science-research-initiative

Državni programi

UK: Introduction to the Data Science Accelerator programme: a capability-building programme that gives analysts from across the public sector the opportunity to develop their data science skills. It is supported by the Government Digital Service (GDS), Office for National Statistics and the Government Office for Science and Civil Service analytical professions (statistics, economics, operational research and social research). https://www.gov.uk/government/publications/data-science-accelerator-programme/introduction-to-the-data-science-accelerator-programme. Updated 17 July 2020

United States Department of Agriculture - National Institute of Food and Agriculture: Data Science for Food and Agricultural Systems (DSFAS). The Agriculture and Food Research Initiative's Food and Agriculture Cyber informatics and Tools (FACT) initiative seeks to catalyze activities that harness big data for synthesizing new knowledge, making predictive decisions, and fostering data-supported innovation in agriculture. https://nifa.usda.gov/program/dsfas

DS - oblasti

EDA

Statističko testiranje (A/B test)

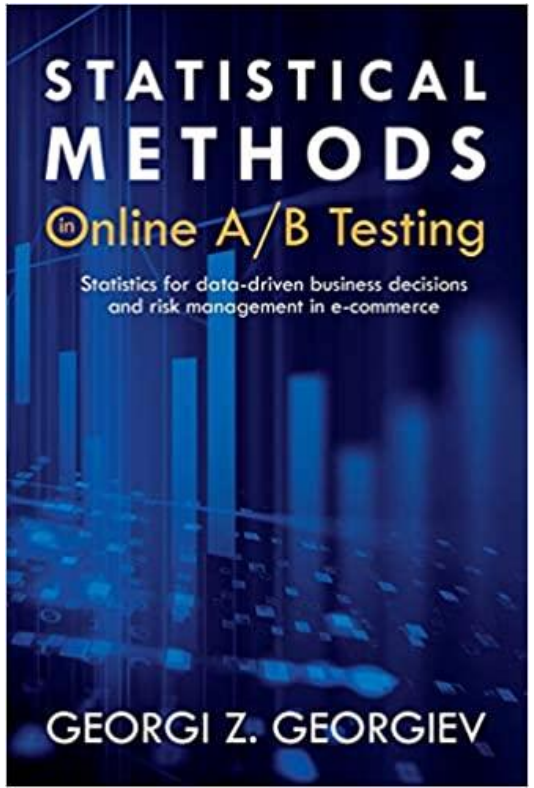
Vremenske serije

Regresija

ML

....

Klasifikacija je proces za otkrivanje kako objekat može da bude definisan u odnosu na drugi objekat.
Predviđanje je proces koji se vrši na osnovu obrazaca i verovatnoće.

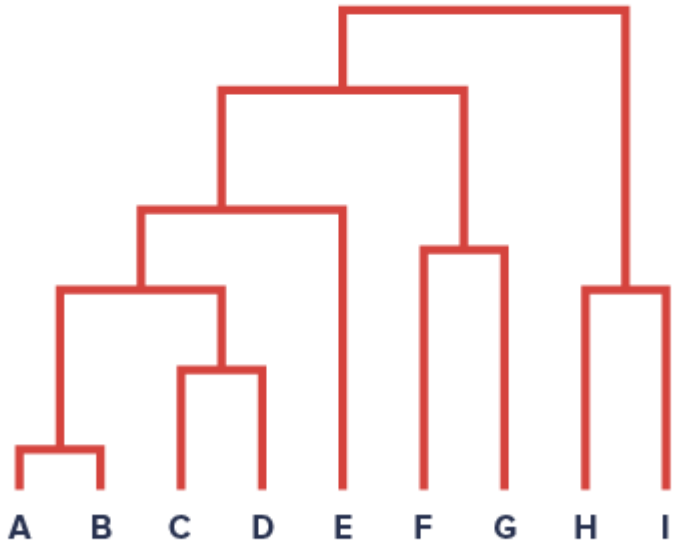


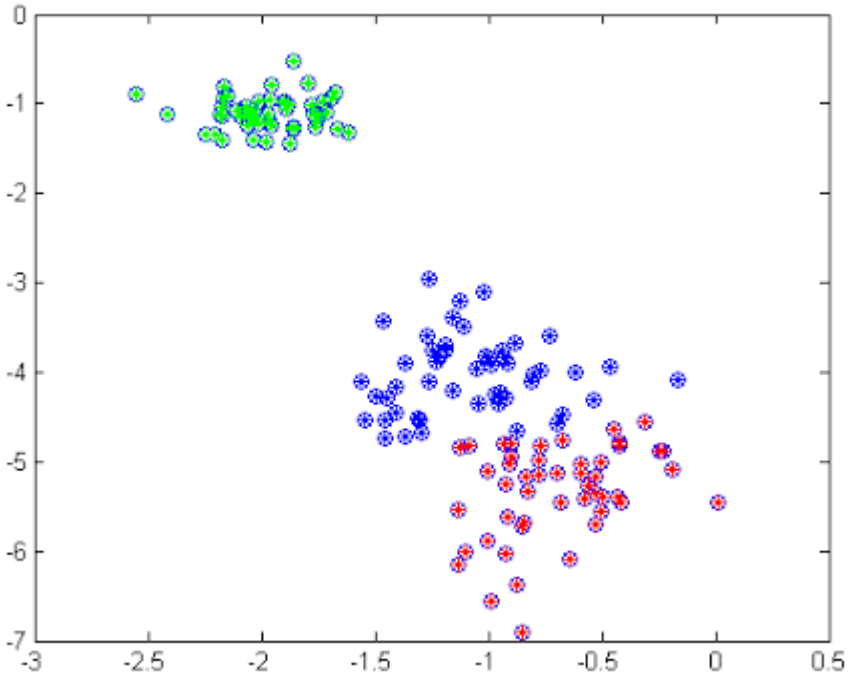
Klasifikacija

Nehijerarhijska

Hijerarhijska

....



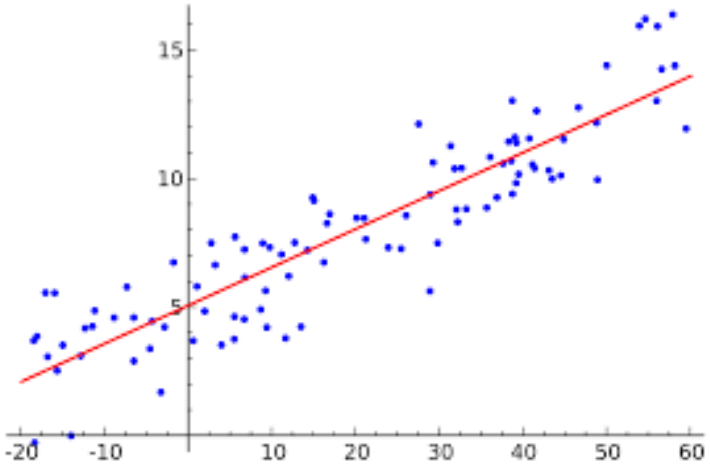


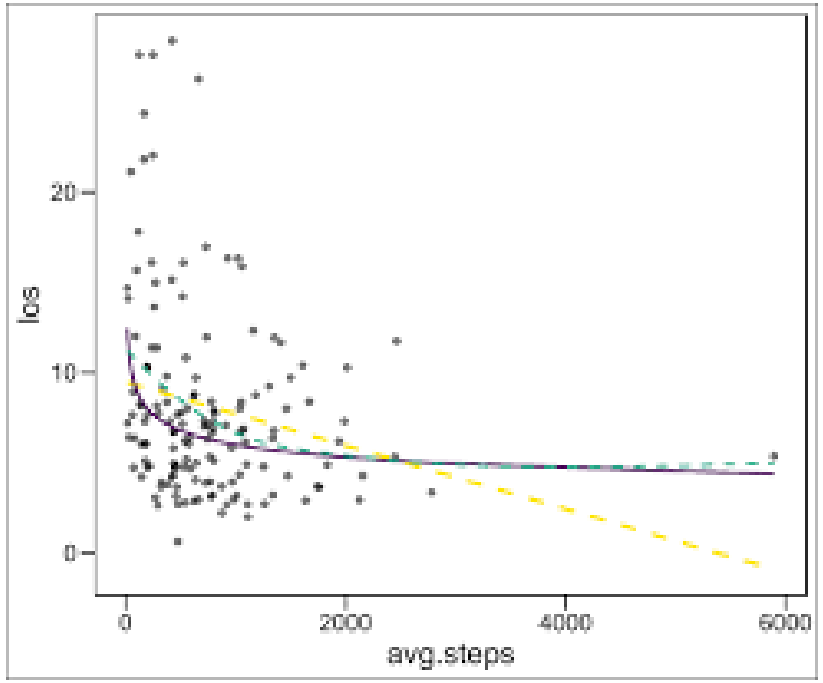
Regresija

LRM

Log-linearni modeli

....





ML

Učenje bez nadzora

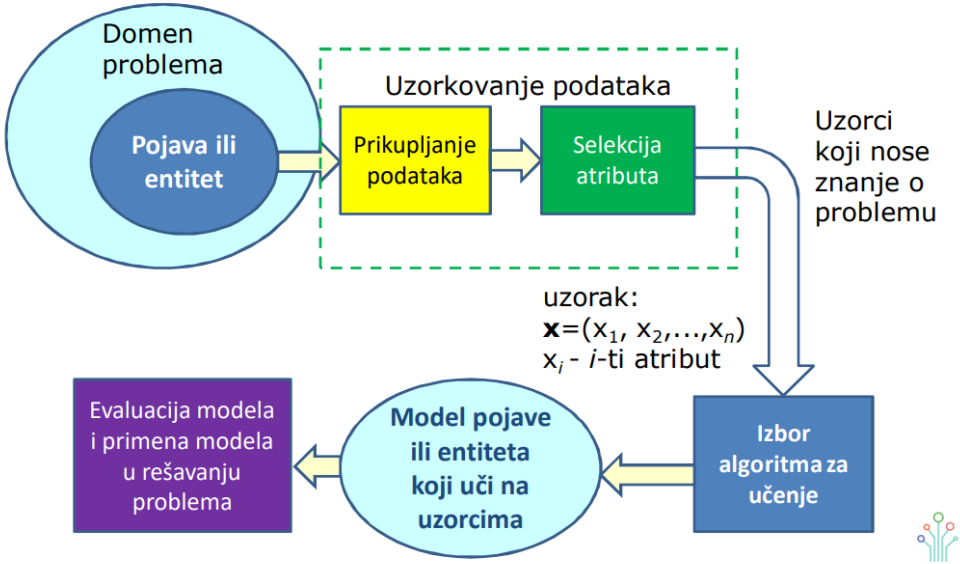
Učenje sa nadzorom

Učenje sa podrškom

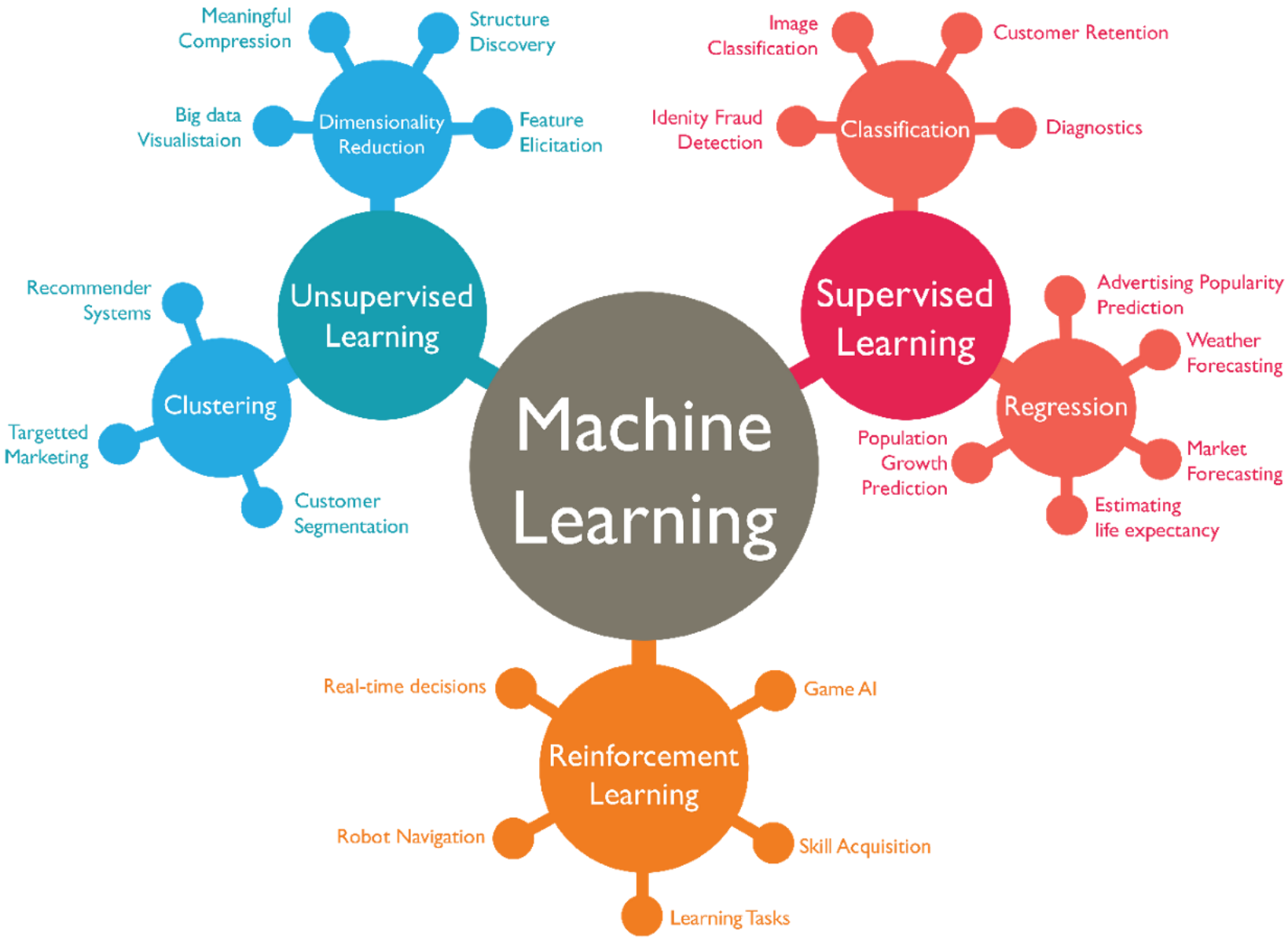
Nadgledano učenje je zasnovano na obeleženim podacima i zaključivanju iz podataka

za trening. Linearna regresija je jedan primer.

Nenadgledano učenje je zasnovano na neobeleženim podacima i često funkcioniše na analizama klastera.

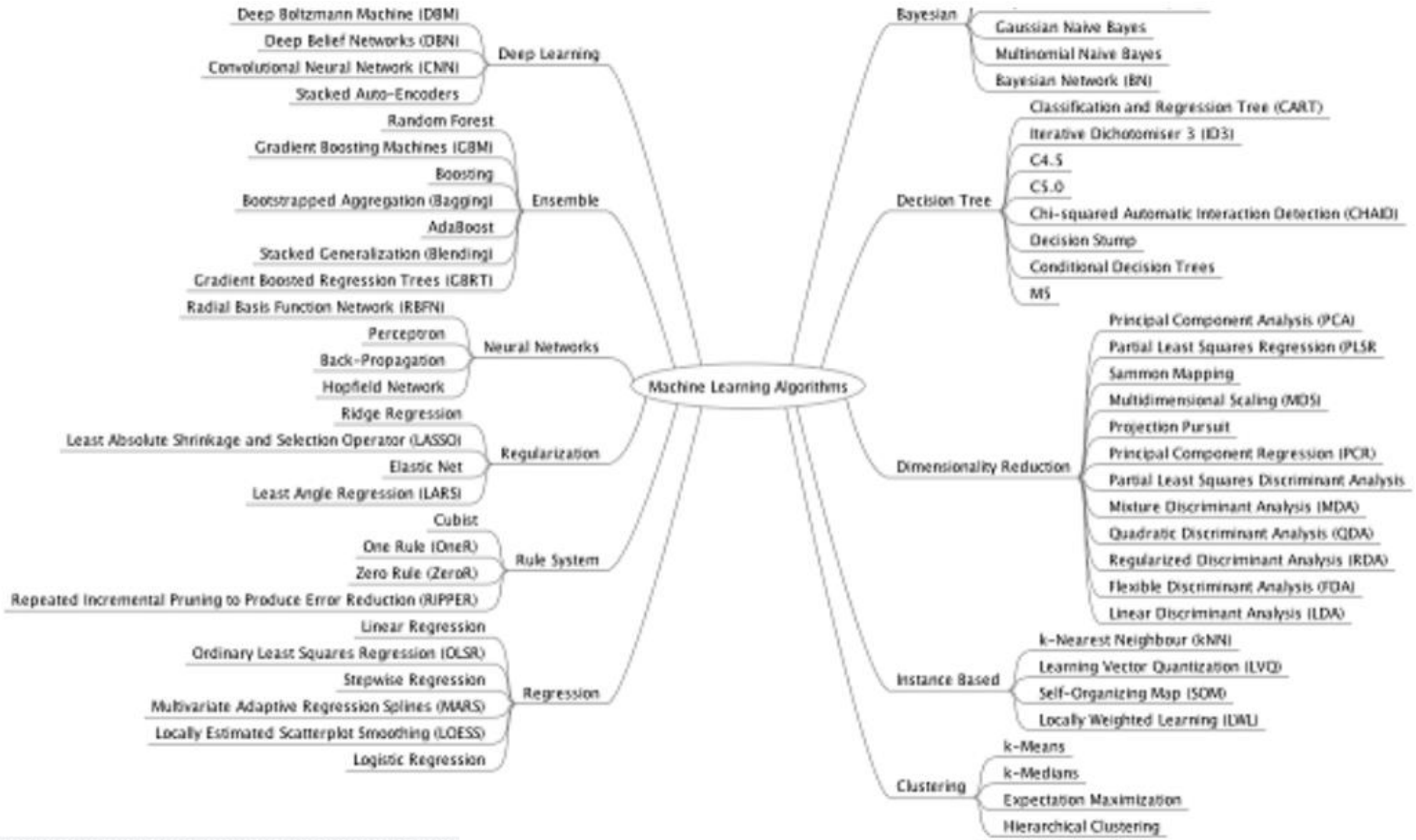


https://www.project-benefit.eu/eplatform/?courses=52&download=02-Uvod%20u%20vestacku%20inteligenciju%20i%20masinsko%20ucenje%20(II).pdf



https://towardsdatascience.com/machine-learning-algorithms-in-laymans-terms-part-1-d0368d769a7b

Machine Learning Algorithms



https://www.datasciencecentral.com/profiles/blogs/a-tour-of-machine-learning-algorithms-

• Top 15 Websites for Data Scientists to Follow in 2021

AI & Machine Learning

1. Towards Data Science

2. Analytics Vidhya

3. KDnuggets

4. Springboard

Data Engineering

5. Uber Engineering Blog

6. Netflix Tech Blog

7. Airbnb Engineering & Data Science

Data Visualization

8. Storytelling with Data

9. Tableau Viz of the Day

10. Information is Beautiful

11. Nightingale

Business Acumen

12. Entrepreneur

13. Forbes

14. Business Insider

15. Hubspot

The Most Popular Tools and Software for Data Science

DATA MANAGEMENT

1. Hadoop
2. MongoDB
3. MySQL
4. Neo4j
5. SAP HANA
6. Hive
7. Apache Spark
8. RapidMiner

DATA VISUALIZATION

9. Microsoft Power BI
10. Tableau
11. QlikView
12. TIBCO Spotfire

DATA ANALYTICS

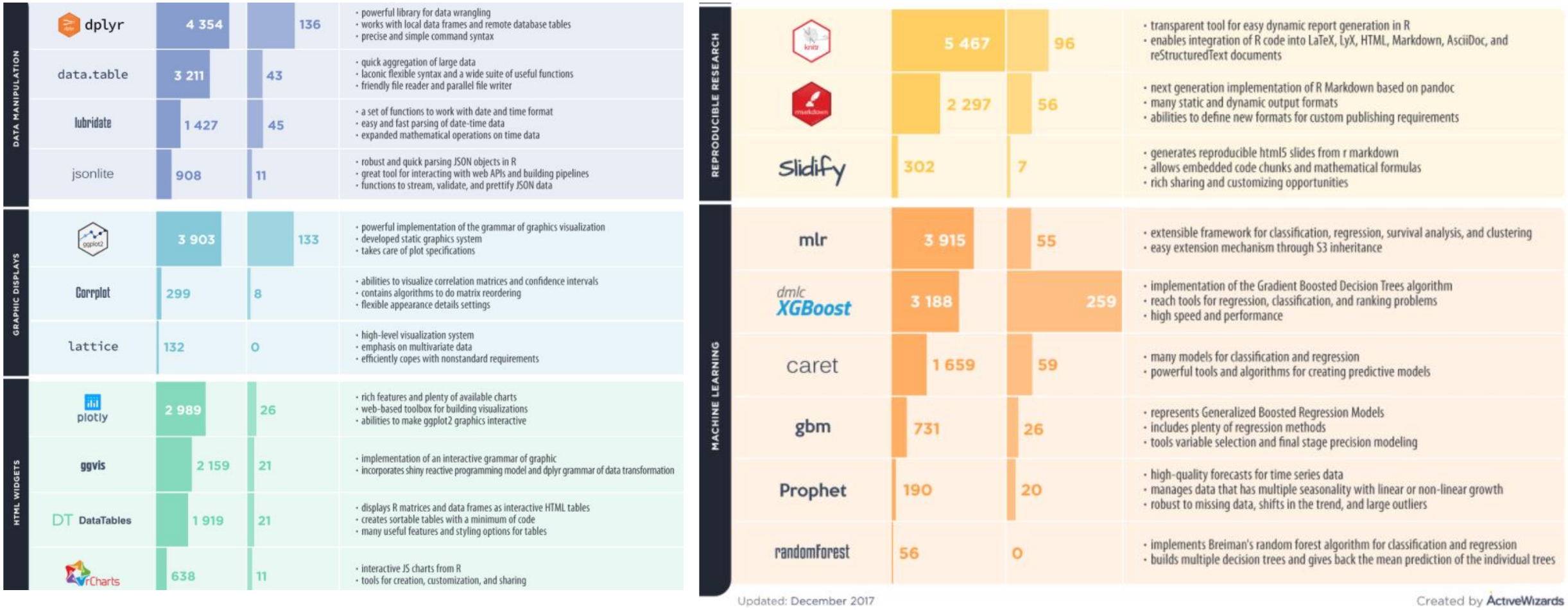
13. Python
14. R
15. SAS
16. MATLAB
17. SPSS
18. STATA
19. RiverLogic
20. SAP Lumira

Python biblioteke za **Data Science**

- **NumPy** - biblioteka **za analizu podataka** za Python. Python nema ugrađenu strukturu niza podataka i treba da upotrebimo biblioteku za efikasno modelovanje vektora i matrica. U NoP su nam potrebne ove strukture podataka da bismo izvršili jednostavne matematičke operacije.
- **SciPy** - biblioteka koja sadrži algoritme koji se koriste **u NoP**. Komplementarna biblioteka za NumPy koja pruža sve potrebne napredne algoritme, bez obzira da li su to algoritmi linearne algebre, alatka za obradu slika ili operacija nad matricama.
- **Pandas** – Obezbeđuje brze, fleksibilne i ekspresivne strukture podataka, kao što su jednodimenzionalne serije i dvodimenzionalni DataFrames. Efikasno učitava formate i obrađuje složene tabele različitog tipa.
- **Scikit-learn** - Pythonova glavna biblioteka za mašinsko učenje. Zasnovana je na NumPy i SciPy bibliotekama. Obezbeđuje funkcionalnost koja je potrebna za izvršavanje klasifikacije i regresije, obradu podataka i za nadgledano i nenadgledano učenje.
- **TensorFlow** – Google-ova biblioteka za **neuronske mreže, pogodna** za implementiranje dubokog učenja VI. Može se upotrebiti za rešavanje različitih problema numeričkog izračunavanja. Neki oblici primene ove biblioteke u realnom svetu uključuju „Googleovo“ prepoznavanje glasa i identifikaciju objekta.

Artificial Intelligence and Machine Learning Fundamentals Zsolt Nagy (2018)

Top 20 R Libraries for Data Science



https://www.kdnuggets.com/2018/05/top-20-r-libraries-data-science-2018.html

https://medium.com/activewizards-machine-learning-company/top-20-r-libraries-for-data-science-in-2018-infographic-956f8419f883

10 Most Frequently Asked Questions In Data Science Interview

1| What is regularisation? Explain L1 and L2 regularisation.

2| How Data Science differs from Big Data and Data Analytics?

3| How do Data Scientists use statistics?

4| Why data cleansing is important?

5| What is Linear and Logistic Regression?

6| What is Normal Distribution?

7| Difference between Interpolation and Extrapolation

8| What is a recommender system?

9| Between R and Python, Which one would you choose for text analysis?

10| Explain A/B Testing

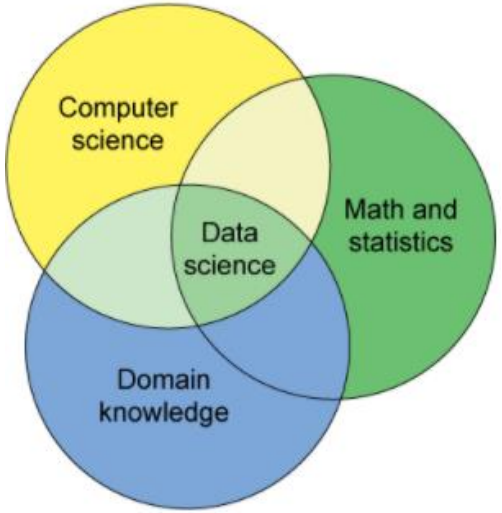
https://www.analyticsindiamag.com/10-most-frequently-asked-questions-in-data-science-interview/

„Data science is here to stay“

Rachel Schutt’s: Introduction to Data Science class in the Columbia University Statistics Department. 2012

„ Let us own data science:

Think or sink; Compute or concede; Lead or lose“

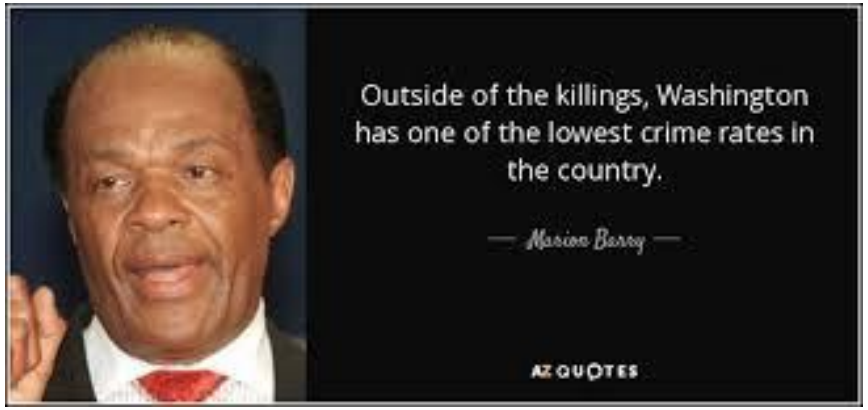


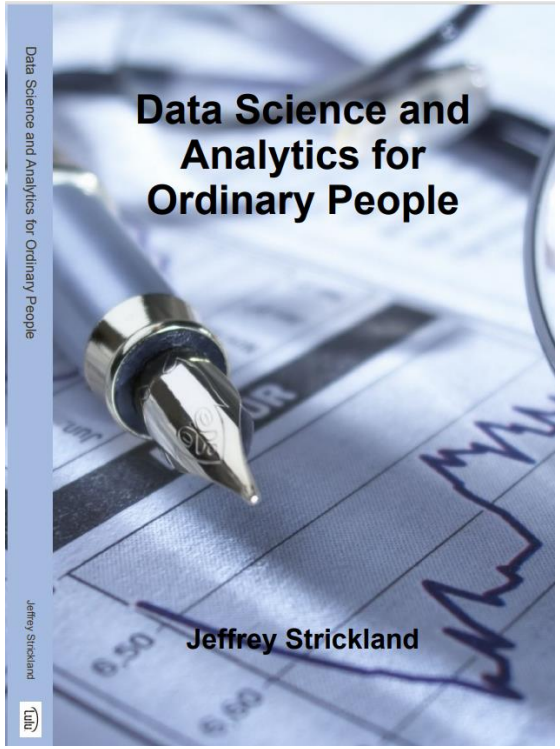
Bin Yu, Chancellor’s Professor of Statistics and EECS, University of California at Berkeley, Presidential Address on Australian Statistical Conference in Sydney (July 9-14, 2014), a joint meeting of the Statistical Society of Australia and Institute of Mathematical Statistics

https://imstat.org/2014/10/01/ims-presidential-address-let-us-own-data-science/

DATA SCINECE: SCINECE OR BUZZWORD?

- Oboje
- Ukoliko se koristi mora biti nauka
- Ukoliko se ne koristi može biti *buzz*





Kako do posla

- Predavač

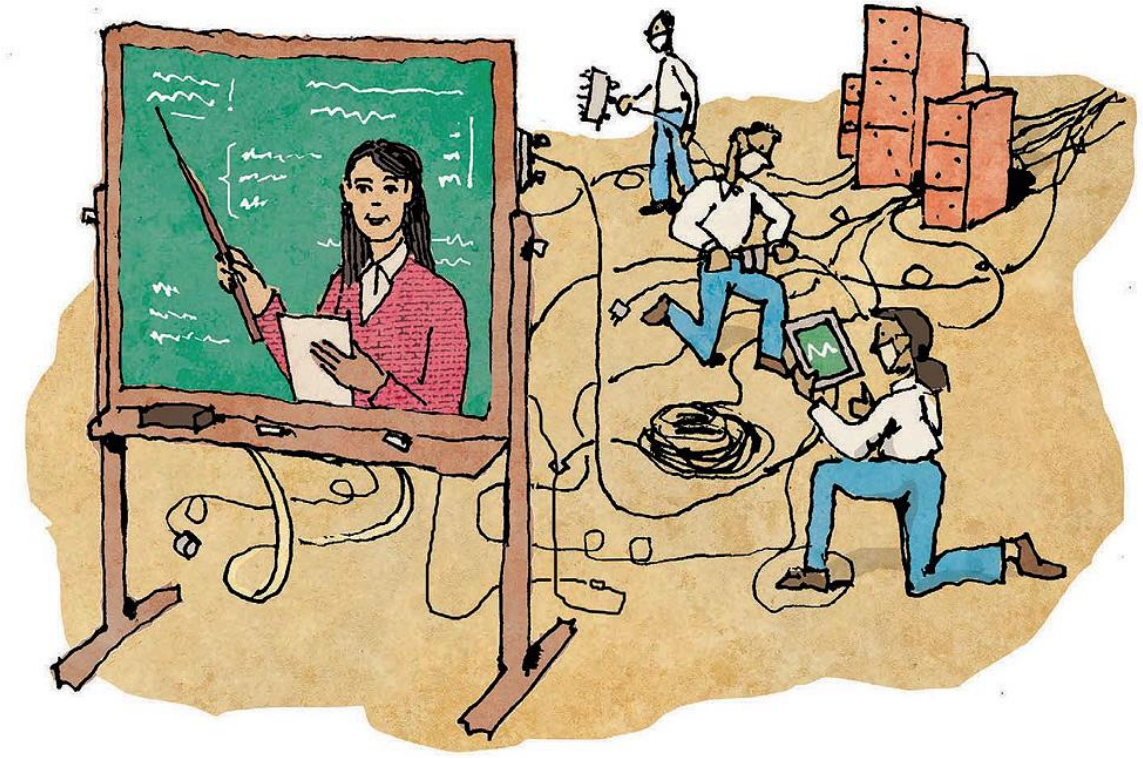
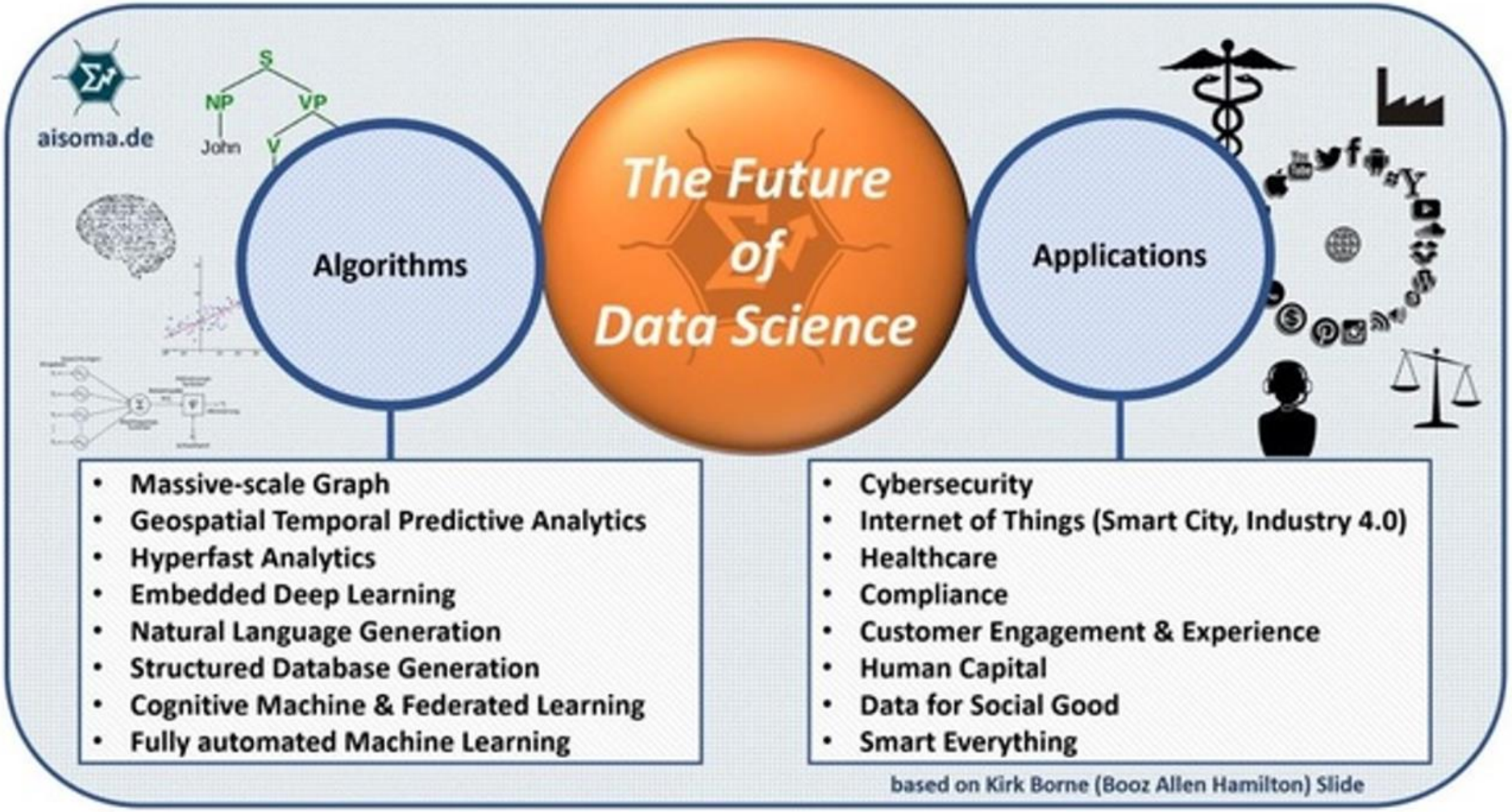


Illustration by Daniel Baxter
https://harvardmagazine.com/2020/11/features-school-goes-remote



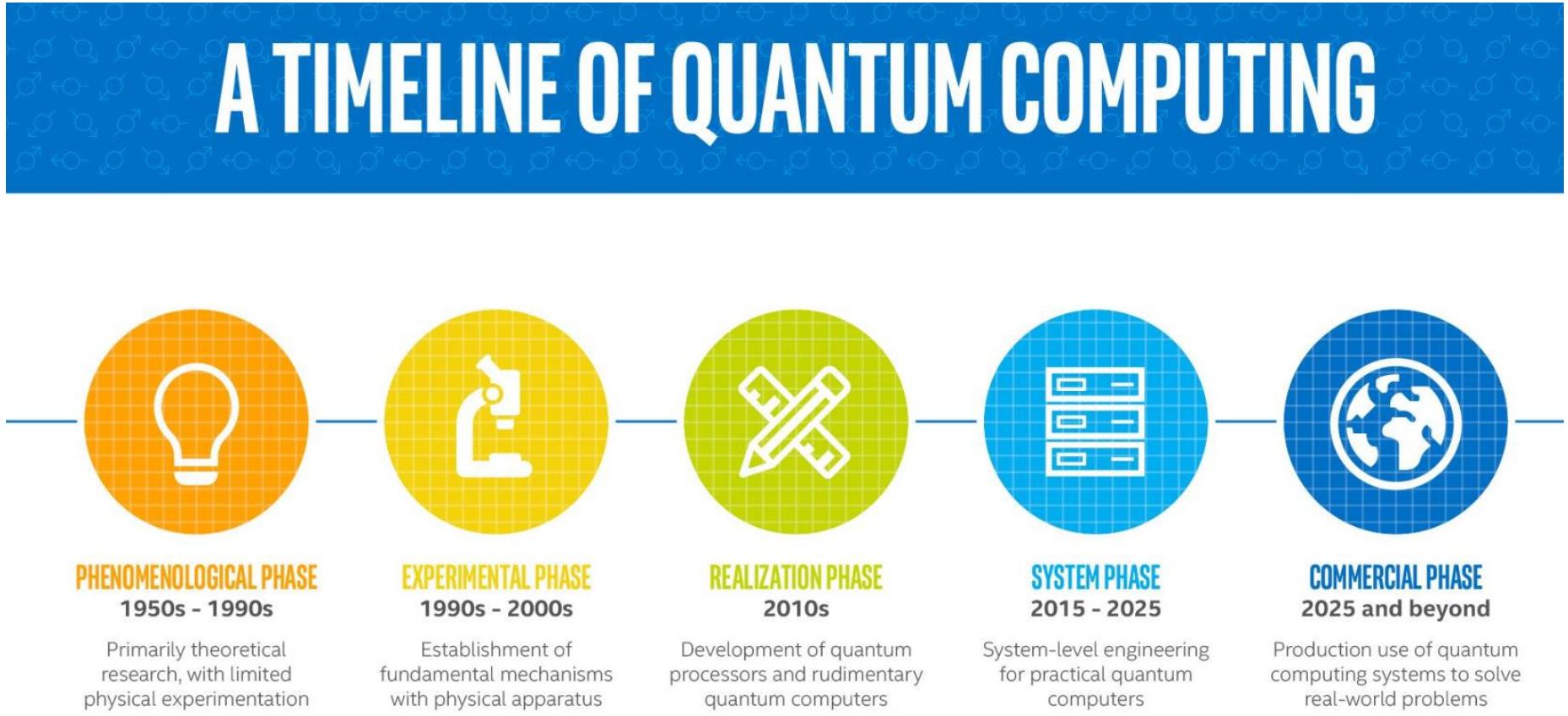
https://twitter.com/KirkDBorne/status/1115673400859615233/photo/1

Future – 4.0 Analytics

- Era One: **Artisanal Analytics** - this methodology was primarily geared towards producing insights for internal decision-making using small-scale, structured datasets.
- Era Two: **Big Data Analytics**
- Era Three: **Data Economy Analytics**
- Era Four: **Autonomous Analytics** - machines not only perform the analysis; they also act on the insights, making decisions faster and more efficiently than any human could - Automated Data Science.
  - The Cognitive Era

Thomas Davenport of Babson College, Harvard Business School and the MIT

Future



https://www.intel.com/content/www/us/en/research/quantum-computing.html

Preporuka

Catalog > Data Analysis & Statistics Courses > IBM's IBM Data Science

Introduction to Data Science

Learn about the world of data science first-hand from real data scientists.

IBM

Choose your session:

Starts Mar 9
Ends Jun 30

Starts Jul 1

62,437 already enrolled!

Enroll now

I would like to receive email from IBM and learn about other offerings related to Introduction to Data Science.

This course is part of a Professional Certificate

About this course

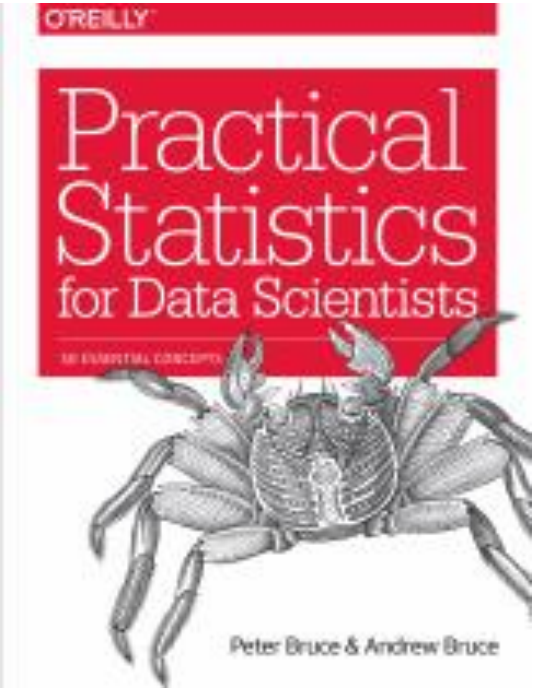
The art of uncovering the insights and trends in data has been around for centuries. The ancient Egyptians applied census data to increase efficiency in tax collection and they accurately predicted the flooding of the Nile river every year. Since then, people working in data science have carved out a unique and distinct field for the work they do. This field is data science and

More about this course

Length:	6 Weeks
Effort:	3–6 hours per week
Price:	FREE Add a Verified Certificate for $39 USD

https://www.edx.org/course/intro-to-data-science

Preporuka – za početak



- Pročitane knjige su daleko manje vredne od nepročitanih

(Taleb, 2010, 33)

Ako želite natprosečne rezultate, ne možete da radite obične stvari

- Hvala na pažnji

*“**Знания** не имеют никакой ценности, если их не применять на **практике**”*

- Антон Чехов (1860-1904)

Neki od korišćenih izvora

- Bruce, P. and A. Bruce (2017). Practical Statistics for Data Scientists. O'Reilly Media, Inc.
- Census (2016). U.S. Census Bureau History: Herman Hollerith and Mechanical Tabulation. US Census. https://www.census.gov/history/www/homepage_archive/2016/january_2016.html
- Chiang, C. (2018). Defining the Terms: Structured Data vs. Unstructured Data. Igneous. Pristupljeno 13.12.2020. https://www.igneous.io/blog/structured-data-vs-unstructured-data
- Cleveland, W. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review / Revue Internationale De Statistique, 69(1), 21-26. doi: 10.2307/1403527
- Data Science for Undergraduates: Opportunities and Options (2018). National Academies of Sciences, Engineering, and Medicine. The National Academies Press. Pristupljeno 11.12.2020: http://nap.edu/25104
- Danner, G. E. (2015). Profit from science: solving business problems using data, math, and the scientific process. Palgrave Macmillan
- De Veaux et al (2018). Curriculum Guidelines for Undergraduate Programs in Data Science. Annu. Rev. Stat. Appl. 2017. 4:2.1–2.16. Doi: 10.1146/annurev-statistics-060116-053930
- Donoho, D. (2015). 50 Years of Data Science. Tukey Centennial workshop in Princeton, New Jersey. Pristupljeno 11.12.2020. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.
- Dos Santos Goncalves, G. (2018). Towards data science. How the 80/20 Rule can help decide which skills you need to start a career in Data Science. Using the Pareto Principle to increase your confidence as a Data Scientist. Pristupljeno 16.12.2020. https://towardsdatascience.com/how-the-80-20-rule-can-help-decide-which-skills-you-need-to-start-a-career-in-data-science-fd60766eba05
- Ernst & Young (2016): Audio analytics: new opportunities in litigation and investigation, Ernst & Young LLP., Pristupljeno, 201.12.2020 : https://www.ey.com/Publication/vwLUAssets/ey-audio-analytics/$FILE/ey-audio-analytics.pdf
- Gandomi, A., M. Haider (2014): Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Volume 35, Issue 2, April 2015, Pages 137-144, https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Hakeem A. et al. (2012) Video Analytics for Business Intelligence. In: Shan C., Porikli F., Xiang T., Gong S. (eds) Video Analytics for Business Intelligence. Studies in Computational Intelligence, vol 409. Springer, Berlin, Heidelberg

- Granville, V (2014). Data science without statistics is possible, even desirable. Data Science Central. Pristupljeno 14.12.2020. https://www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable
- Haas, L., Hero, A, and Lue, R. A. (2019). Highlights of the National Academies Report on "Undergraduate Data Science: Opportunities and Options". An interview with Laura Haas and Alfred Hero by Robert Lue. HDSR. DOI: 10.1162/99608f92.38f16b68
- Hey, T., Tansley, S. and Tolle, K. (2011). Editors: Hey, T., Tansley, T., Tolle, K. M. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. Pristupljeno 13.12.2020. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf
- Hurwitz, J.S.,Nugent, A. (2013): Big Data for Dummies, John Wiley & Sons, New Jersey
- Kaur, A., C. Deepti (2016): Comparison of Text Mining Tools, 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Page(s):186-192 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7784950
- Li, X., F. Xiaoping, X. Qu, G. Sun, C. Yang, B. Zu, Z. Lia (2019). Curriculum Reform in Big Data Education at Applied Technical Colleges and Universities in China. IEEE Access. VOLUME 7, 2019 pp. 125511-125521, 2019, doi: 10.1109/ACCESS.2019.2939196. Pristupljeno 17.12.2020.
- Lim, T. S. (2020). Data Analysis. GitHub. Pristupljeno 13.12.2020. https://tslim.github.io/concepts/concepts/data-architecture/data-analysis.html
- Lipton, Z. C.,& J. Steinhardt (2018). Troubling Trends in Machine Learning Scholarship, Carnegie Mellon University, Stanford University, Pristupljeno 10.03.2021. https://arxiv.org/pdf/1807.03341.pdf
- MacTutor (2020). Analysis - History Topics. School of Mathematics and Statistics. University of St Andrews, Scotland. https://mathshistory.st-andrews.ac.uk/HistTopics/. Pristupljeno: 12.12.2020
- Meng, X.-L. (2019). Data Science: An Artificial Ecosystem. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.ba20f892

- Miljković E. (2010). Osmanske popisne knjige defteri kao izvori za istorijsku demografiju - mogućnosti istraživanja, tačnost pokazatelja i metodološke nedoumice. Teme. 2010, vol. 34, br. 1, str. 363-373
- Minelli, M. & Chambers, M. (2013). Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley Publishing Inc.
- Nigam, A. (2019). Data Science vs Artificial Intelligence vs Machine Learning. Insaid blog. Pristupljeno: 13.12.2020. https://blog.insaid.co/data-science-vs-machine-learning-vs-ai/
- NOESIS (2018). Big Data implementation context in transport. Horizon 2020 Research and Innovation Programme. Accessed on October 15, 2020. Accessed on October 15, 2020. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bedfd3b8&appId=PPGMS
- Pavičić. M. (2019). Smrt Mooreovog zakona. SmartInfoTrend 213/2019/Q4. Pristupljeno: 16.12.2020. https://www2.irb.hr/korisnici/mpavicic/papers-ps-pdf/popular/Pavicic-InfoTrend2020.pdf
- Pentland A. The data-driven society. Sci Am. 2013 Oct;309(4):78-83. doi: 10.1038/scientificamerican1013-78. PMID: 24137860.
- Pierson, L. (2017). Data Science For Dummies, 2nd Edition Published by: John Wiley & Sons, Inc.,
- Society 5.0 (2019). Realizing Society 5.0. Pristupljeno: 18.12.2020. https://www.japan.go.jp/abenomics/_userdata/abenomics/pdf/society_5.0.pdf

- Tukey, J. W. (1962). The Future of Data Analysis. Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Mathematical Statistics. Pristupljeno: 12.12.2020. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711
- UM (2016). $100M Data science initiative launched. University of Michigan 2016 ANNUAL REPORT. Pristupljeno: 15.12.2020. http://finance.umich.edu/reports/2016/100-million-data-science-initiative-launched/
- Upadrashta, P.S. (2019). What is the difference between a data scientist and a machine learning engineer? Quora. Pristupljeno 17.12.2020. https://www.quora.com/What-is-the-difference-between-a-data-scientist-and-a-machine-learning-engineer
- Vukmirović, D. (2020). Introductory Data Science for Managers and Business Leaders -Workshop. Conference: SYMORG 2020. 10.13140/RG.2.2.27339.41764
- Vukmirović, D., Čomić, T., Bolbotinović, Ž., Dabetić, Đ. and Jovanović Milenković, M. (2091). MIP - prototip modela inteligentnog preduzeća, Zbornik radova - SYMOPIS 2019. Pristupljeno: 12.12.2020. https://rspdf.info/%D0%B4%D0%BE%D0%BA%D1%83%D0%BC%D0%B5%D0%BD%D1%82/cb10c2/zbornik-radova-symopis-2019-%D0%A3%D0%BD%D0%B8%D0%B2%D0%B5%D1%80%D0%B7%D0%B8%D1%82%D0%B5%D1%82-%D1%83-%D0%91%D0%B5%D0%BE%D0%B3%D1%80%D0%B0%D0%B4%D1%83
- Vukmirović, D. (2020). DATA SCIENCE: SCIENCE OR BUZZWORD. Seminar za računarstvo i primenjenu matematiku. Matematički Institut SANU, Beograd. DOI: 10.13140/RG.2.2.16631.24481
- Watson, J. V. (2001). A Brief History of Numbers and Statistics With Cytometric Applications. Cytometry (Communications in Clinical Cytometry) 46:1–22 (2001. Pristupljeno 10.2.2021. https://onlinelibrary.wiley.com/doi/epdf/10.1002/1097-0320(20010215)46:1%3c1::AID-CYTO1032%3e3.0.CO;2-3