

Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

Diplomski studij računarstva, 1. semestar

Ak. god. 2021./2022.

# Uvod u znanost o podacima

1. predavanje



# Nositelji predmeta

prof. dr. sc. **Mile Šikić** – ZESOI, D-104, [mile.sikic@fer.hr](mailto:mile.sikic@fer.hr)

prof. dr. sc. **Bojana Dalbelo Bašić** – ZEMRIS, D-317, [bojana.dalbelo@fer.hr](mailto:bojana.dalbelo@fer.hr)

izv. prof. dr. sc. **Alan Jović** – ZEMRIS, D-340, [alan.jovic@fer.hr](mailto:alan.jovic@fer.hr)

doc. dr. sc. **Ana Sović Kržić** – ZESOI, D-151, [ana.sovic.krzic@fer.hr](mailto:ana.sovic.krzic@fer.hr)



# Asistenti na predmetu

**Igor Stančin**, mag. ing. comp. – ZEMRIS, D-335, [igor.stancin@fer.hr](mailto:igor.stancin@fer.hr)

**Eugen Vušak**, mag. ing. comp. – ZEMRIS, D-336, [eugen.vusak@fer.hr](mailto:eugen.vusak@fer.hr)

dr. sc. **Filip Bosnić** – ZESOI, D-163-1, [filip.bosnic@fer.hr](mailto:filip.bosnic@fer.hr)

**Liljana Pushkar**, MSc – ZESOI, D-162, [liljana.pushkar@fer.hr](mailto:liljana.pushkar@fer.hr)



# O predmetu

Sve obavijesti, materijali i ostale informacije o predmetu dostupne su na web stranici:

<https://www.fer.unizg.hr/predmet/uuzop>

Jezgreni predmet profila **Znanost o podacima**

Izborni predmet profila:

- Elektroničko i računalno inženjerstvo

- Informacijsko i komunikacijsko inženjerstvo,

- Programsko inženjerstvo i informacijski sustavi

- Računalno inženjerstvo

**ECTS: 5**

# Opis predmeta i ishodi učenja

## Opis predmeta

- Ovaj predmet upoznaje studente s pet ključnih aspekata istraživanja temeljenog na podacima:
  - prilagođavanje formata podataka, čišćenje podataka i uzorkovanje u cilju dobivanja odgovarajućeg skupa podataka,
  - upravljanje podacima radi brzog i pouzdanog pristupa velikoj količini podataka,
  - eksploratorna analiza podataka u cilju generiranja hipoteze i intuicije,
  - predviđanje temeljeno na statističkim metodama kao što su regresija i klasifikacija,
  - komuniciranje rezultata kroz vizualizaciju, opis i sažetu interpretaciju rezultata.

# Opis predmeta i ishodi učenja

## Ishodi učenja

- koristiti Python i druge alate za prikupljanje, čišćenje i procesiranje podataka
- koristiti tehnike upravljanja podataka za spremanje podataka lokalno i u oblak
- koristiti statističke metode i vizualizaciju za brzo istraživanje podataka
- primijeniti statistiku i računalnu analizu za predviđanje temeljeno na podacima
- opisati rezultate analize podataka koristeći deskriptivnu statistiku i vizualizacije
- koristiti grozd računala i infrastrukturu u oblaku za obavljanje podatkovno-intenzivnih računanja

**Naglasak je na primjeni i na širini pristupa!**

# Nastavne teme – po tjednima

- **1. tjedan:** O predmetu, opis područja ZOP, izvori i pohrana podataka, alati i radni okviri
- **2. tjedan:** Rukovanje podacima, problemi skupova podataka, inženjerstvo značajki
- **3. tjedan:** Vizualizacija podataka
- **4. tjedan:** Prikupljanje podataka od ispitanika, eksperimenti i opservacijske studije
- **5. tjedan:** Linearna regresija, transformacije podataka
- **6. tjedan:** Opisivanje podataka: deskriptivna statistika, testiranje hipoteze, interval pouzdanosti
- **7. tjedan:** Primijenjeno nadzirano strojno učenje (klasifikacija i predviđanje)
- **8a. – 8b. tjedan:** ----međuispiti (ne na predmetu)

# Nastavne teme – po tjednima

- **9. tjedan:** Primijenjeno strojno učenje (podaci, označavanje, značajke, izbor modela, vrednovanje modela)
- **10. tjedan:** Primijenjeno nenadzirano strojno učenje (grupiranje)
- **11. tjedan:** Duboko učenje
- **12. tjedan:** Rad s tekstom
- **13. tjedan:** Rad s grafovima
- **14. tjedan:** Prezentacije projekata
- **15. tjedan:** Završni ispit



# Nastavne obveze na predmetu

- Predavanja (3 sata tjedno, 7+6 tjedana)
- Laboratorijske vježbe (1 sat tjedno, 7+5 tjedana)
- Projekt (rad od kuće, 7+5 tjedana) – 40 bodova
- Završni ispit (15. tjedan) – 60 bodova
- Ispiti na rokovima – 60 bodova
  
- Pragovi za ocjene: 50 (2) – 63 (3) – 75 (4) – 88 (5)

# Predavanja i laboratorijske vježbe

- Predavanja
  - Utorkom od 14 do 17h
  - U prostoriji B5 ili online putem MS Teamsa ili Zooma, ovisno o nastavniku i epidemiološkoj situaciji
- Laboratorijske vježbe
  - Svaki tjedan 1h, uživo ili online
  - Najava termina barem jedan dan ranije (ovisi o zauzeću prostorija i drugim faktorima)
  - Asistent demonstrira u kodu u Pythonu tematiku predavanja održanog taj tjedan

# Projekt

- Nosi 40 bodova, prag za prolaz je 10 bodova
- Traje tijekom čitavog semestra
- Projekt se radi na temelju znanstvenih članaka
- Podijeljen u dva dijela:
  1. Repliciranje rezultata ostvarenog u znanstvenom članku – individualan rad – nosi do 20 bodova
  2. Kreativno proširenje teme članka – grupni rad do 4 studenta – nosi do 20 bodova
- Rad na projektu u sustavu Github, uz dodavanje asistenta u tim

# Hodogram projekta

- 1. rok (kraj 4. tjedna)
  - Pročitani članci, pronađen članak za koji će se raditi replikacije rezultata, diskusija pristupa replikaciji s nadležnim asistentom
- 2. rok (kraj 7. tjedna) – 20 bodova
  - Ostvarena replikacija rezultata u programskom jeziku Python s dokumentacijom i demonstracija rada nadležnom asistentu
- Grupiranje po članku (rok 9. tjedan) – 2–4 člana grupe
  - Grupa izabire implementaciju iz 1. dijela projekta koju nadograđuje
- 3. rok (kraj 13. tjedna)
  - Ostvareno kreativno proširenje članka u programskom jeziku Python s dorađenom dokumentacijom i videom o ostvarenju – max. 20 bodova
- Prezentacija najboljih projekata (14. tjedan, uživo ili online)

# Završni ispit

- Nosi 60 bodova, prag za prolaz je 30 bodova
- Uvjet za pristup je **položeni projekt**
- Fokus na razumijevanju praktične primjene metoda proučavanih na predavanjima
- Teorijska pitanja i analiza skupa podataka na računalu prema uputama

# Ispiti na rokovima

- Jednaki način polaganja i bodovanje kao i završni ispit
- Uvjet za pristup je **položeni projekt**

# Preporučena literatura

- Jake VanderPlas, (2016.), *Python Data Science Handbook*, O'Reilly Media
- Matt Harrison, Theodore Petrou (2020.), *Pandas 1.x Cookbook*, Packt Publishing Ltd
- Alice Zheng, Amanda Casari (2018.), *Feature Engineering for Machine Learning*, O'Reilly Media
- Andreas C. Müller, Sarah Guido (2016.), *Introduction to Machine Learning with Python*, O'Reilly Media
- Hadley Wickham, Garrett Golemund (2017.), *R for Data Science*, O'Reilly Media
  
- **Sve potrebno za položiti predmet obrađuje se na predavanjima i laboratorijskim vježbama**

# Opis područja



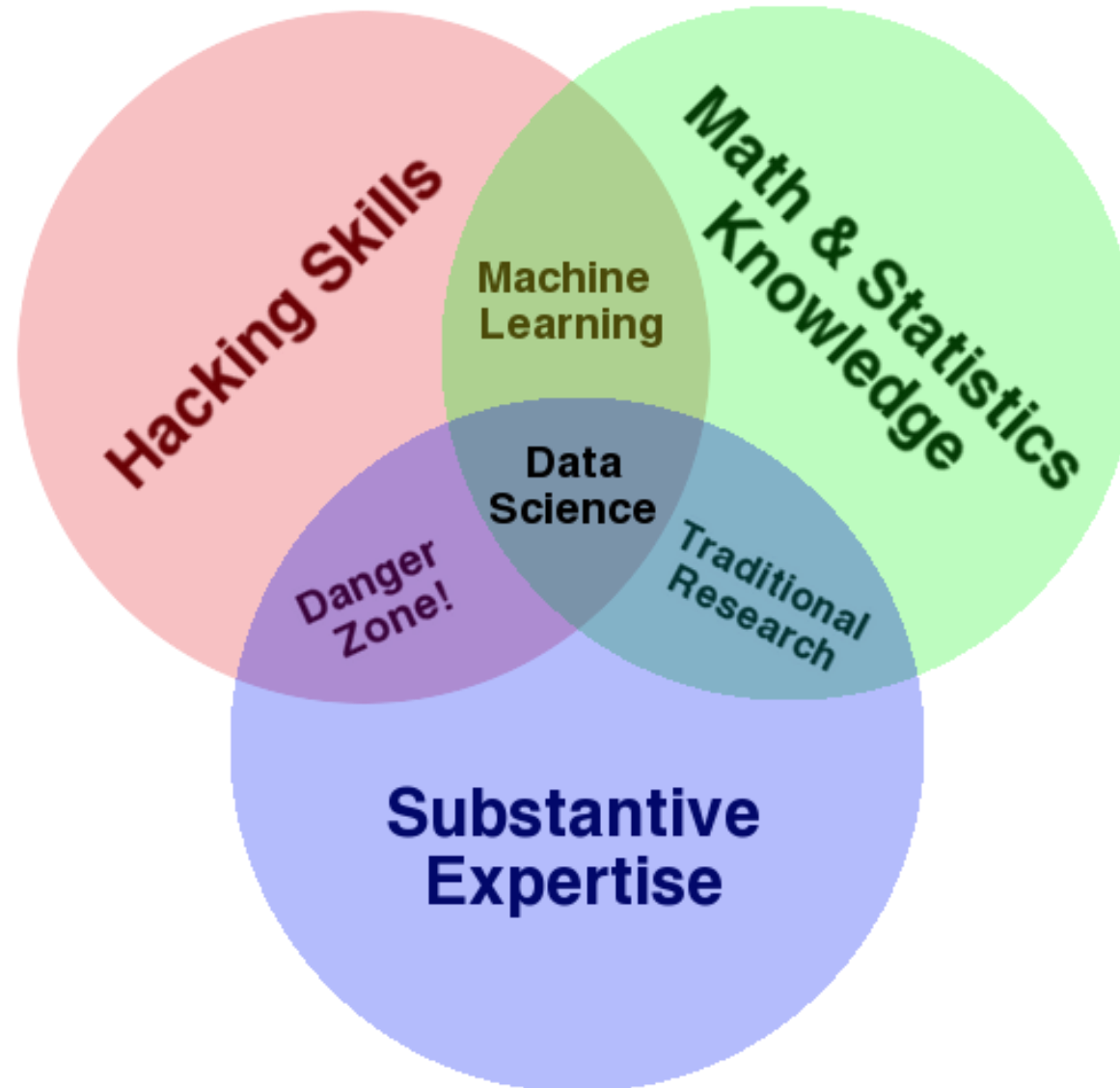
# Što je znanost o podacima?

- Nema usuglašene definicije
- *Wikipedia*: Znanost o podacima (engl. *data science*) je interdisciplinarno polje koje koristi znanstvene metode, procese, algoritme i sustave da bi izvuklo znanje i uvide iz šumovitih, strukturiranih i nestrukturiranih podataka i koristilo ih na širokom rasponu primjenskih područja
- Temeljni pojmovi: **znanost, znanje i uvidi, podaci, primjena**
- Temeljna pretpostavka: **znanost je primarno pokretana podacima** (engl. *data-driven science*)

# Područja od interesa

- Srodna područja:
  - **Statistika** (engl. *statistics*)
  - **Umjetna inteligencija** (engl. *artificial intelligence*, AI)
  - **Strojno učenje** (engl. *machine learning*)
  - **Dubinska analiza podataka** (engl. *data mining*)
  - **Veliki podaci** (engl. *big data*)
  - **Podatkovna analitika** (engl. *data analytics*)
  - **Poslovna inteligencija** (engl. *business intelligence*)
- Bliska i primijenjena područja:
  - Računarska znanost, kognitivna znanost, znanost o odlučivanju, vizualizacija, bioinformatika, medicinska informatika, računska fizika, računska kemija, računska biologija...

# Znanstvenik 2.0



# Što znanost o podacima sve obuhvaća? (1/2)

- Prikupljanje podataka
  - Provođenje studija, intervjui, ankete
- Pohranu i organizaciju podataka
  - Različiti sustavi baza podataka, skladišta podataka, mrežno povezivanje, pristup putem web usluga, usluge u oblaku
- Rukovanje podacima
  - Dobavljanje, čišćenje, transformiranje, obogaćivanje
- Vizualizaciju podataka i rezultata
  - Pregled podataka, različiti grafovi i tablični prikazi

# Što znanost o podacima sve obuhvaća? (2/2)

- Deskriptivnu i inferencijalnu statistiku
  - Uspostava i vrednovanje odnosa između primjera i između varijabli
- Modeliranje metodama strojnog učenja (uključujući duboko učenje)
  - Različiti pristupi i različite metode, ovisno o cilju koji se želi postići
- Izbor modela, vrednovanje i poboljšavanje modela
  - Često jedan model nije dovoljan!
- Komuniciranje u okviru projekta i donošenje odluka na temelju podataka
- Za većinu navedenih aktivnosti, nužna su znanja **programiranja**

# Programska rješenja za ZOP

- Veliki broj komercijalnih rješenja
- Besplatna rješenja
  - **Python**
    - IPython, Jupyter, NumPy, Pandas, Matplotlib, Scikit-Learn, TensorFlow, PyTorch, Keras, Caffe...
  - **R**
    - R + RStudio, brojni paketi: caret, mlr3, ggplot2...
  - Ostali: **Java, Skala, C++**, ...
    - Weka, RapidMiner, Orange, KNIME, DeepLearning4J, Shogun, CNTK...

Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms



# Poslovi u okviru ZOP

- Očekivani poslovi studenata koji završe smjer ZOP:
  - **Podatkovni znanstvenik**  
(engl. *Data scientist*)
  - **Podatkovni inženjer**  
(engl. *Data engineer*)
  - **Podatkovni analitičar**  
(engl. *Data analyst*)

**Data Scientist**  
also known as Data Managers, statisticians.




A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

**Skills:** Mathematics, Programming, Communication



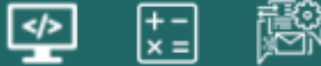
Will use programmes such as:  
SQL, Python, R

**Data Engineers**  
also known as database administrators and data architects.



They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

**Skills:** Programming, Mathematics, Big data



Will use programmes such as:  
Hadoop, NoSQL, and Python

**Data Analysts**  
also known as business Analysts.



They typically help people from across the company understand specific queries with charts.

**Skills:** Statistics, Communication, Business knowledge



Will use programmes such as:  
Excel, Tableau, SQL

- Veliki broj ostalih poslova, npr: računalni programer, administrator baza podataka, upravitelj informacijskog sustava, financijski analitičar...

“A data scientist is someone who can obtain, scrub, explore, model, and interpret data, blending hacking, statistics, and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

*Hilary Mason, chief scientist at bit.ly*

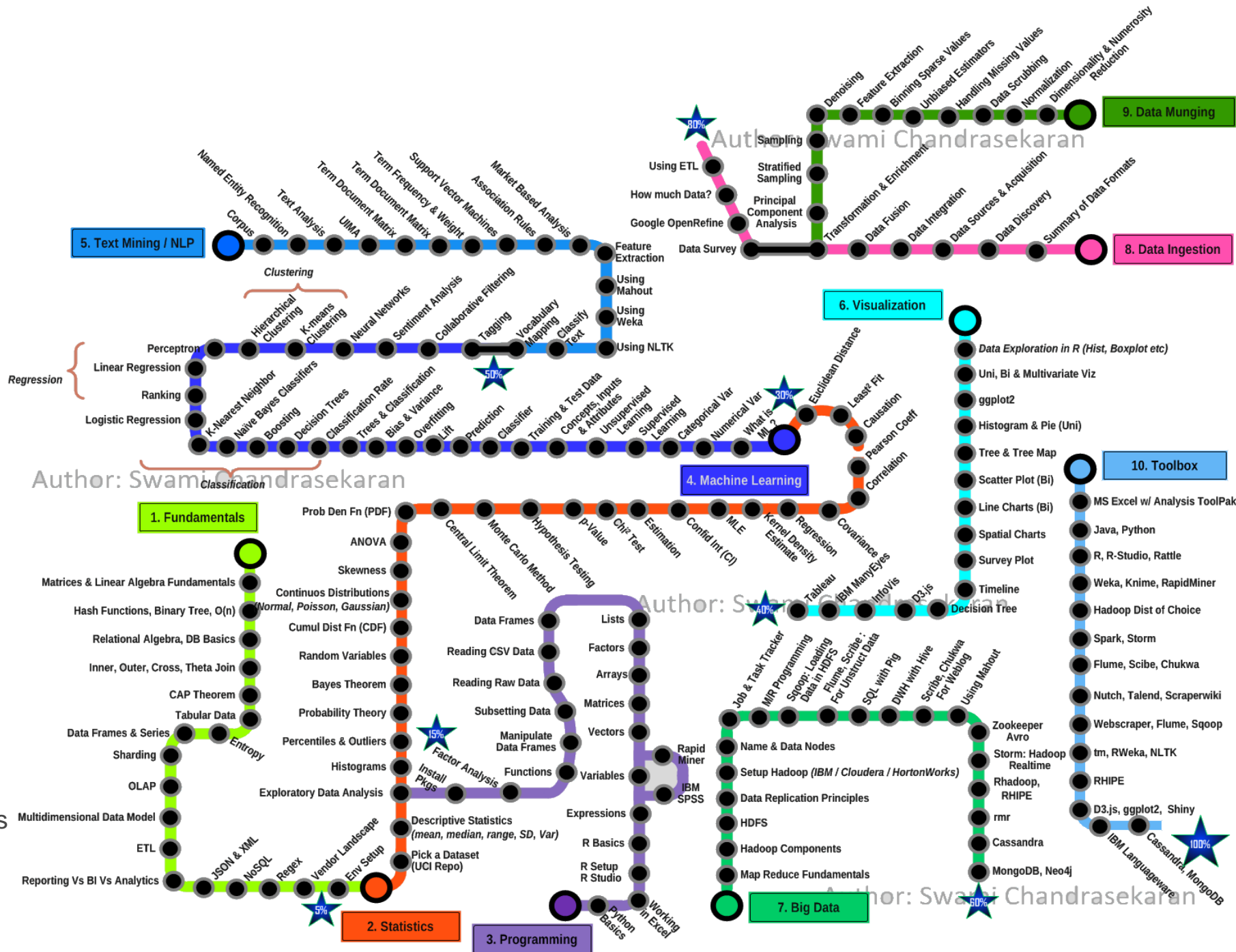
<https://www.kdnuggets.com/2021/09/data-scientists-data-engineering-skills.html>

<https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html>



# Jedna moguća karta puta učenja u ZOP

Autor: Swami Chandrasekaran  
(Managing Director, KPMG's AI  
Innovation & Enterprise Solutions  
- Dallas, USA, 2013.

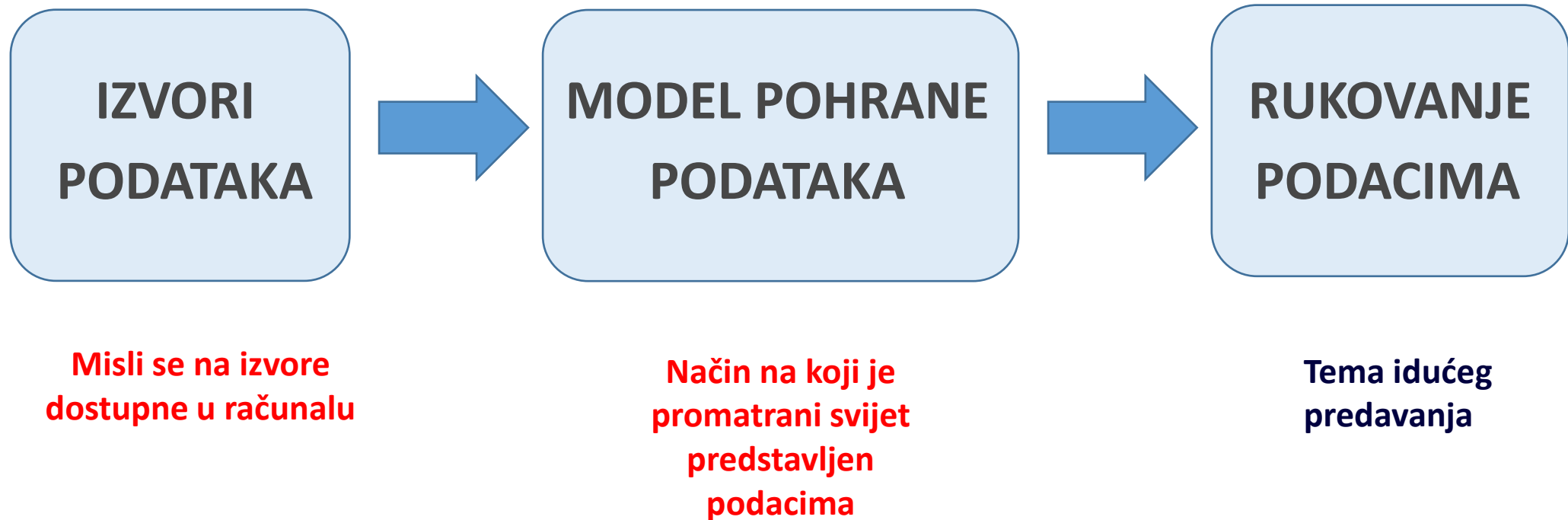


# Korisne web stranice

- **KDnuggets** - <https://www.kdnuggets.com/>
  - Članci, statistike i ankete
- Towards data science - <https://towardsdatascience.com/>
  - Članci
- FastML - <http://fastml.com/>
  - Članci
- Machine learning mastery <https://machinelearningmastery.com/>
  - Članci
- **Kaggle** - <https://www.kaggle.com/>
  - Natjecanja, skupovi podataka
- **UCI Machine Learning Repository** - <https://archive.ics.uci.edu/ml/index.php>
  - Skupovi podataka
- World Data - <https://worlddata.ai/>
  - Skupovi podataka (jako puno), puni pristup se plaća

# Izvori i pohrana podataka

# Redoslijed



# Klasifikacija izvora podataka prema strukturi

- **Strukturirani podaci (engl. *Structured data*)**
  - Podaci koji su posloženi u preddefinirani model s eksplicitnom strukturom tablice
    - Npr. relacijske baze podataka, Excel tablice
- **Polustrukturirani podaci (engl. *Semi-structured data*)**
  - Podvrsta strukturiranih podataka koji nisu organizirani tablično, već koriste neke oznake kojim opisuju strukturu i značenje podataka
    - Npr. nerelacijske baze podataka, XML, JSON i (netablični) CSV oblik
- **Nestrukturirani podaci (engl. *Unstructured data*)**
  - Svi ostali podaci, koji nemaju definirani model ili oznake
    - Npr. Wikipedia, slike, video, zvuk...

# Česti izvori podataka

- **SQL baze podataka** – strukturirani podaci
- **NoSQL baze podataka** – polustrukturirani podaci
- **Web stranice** – (većinom) nestrukturirani podaci
- **CSV ili TXT datoteke** – strukturirani, polustrukturirani ili nestrukturirani podaci
- **Repozitoriji** slika, videa, zvuka i drugih snimljenih signala

# Strukturirani podaci

- Tradicionalno najčešći oblik podataka
- **Skup podataka (engl. *dataset*, *data set*)** u tablici opisan je putem:
  - **Varijabli** (atributa, značajki, parametara)
  - **Primjera** (objekata, uzoraka, primjeraka, „podataka”)
- Pogodan oblik za rukovanje i analizu
  - Izravno učitavanje u predviđene strukture podataka u većini programskih rješenja
  - Omogućava statističku analizu, vizualizaciju i izgradnju modela strojnog učenja za veliku većinu postojećih tehnika
- U novije vrijeme često ga zamjenjuju druge vrste izvora podataka

# Polustrukturirani podaci

- Česti u suvremenoj komunikaciji na internetu
  - Dnevnicima rada poslužitelja
  - Podaci u NoSQL bazama i razmjenjivani u komunikaciji klijent – poslužitelj
- Opisani su strukturno, često hijerarhijski, gdje oznake (tagovi, markeri) definiraju njihovo značenje i mogu se smatrati varijablama
- Najčešće ne omogućuju direktnu analizu, potrebno je pretvoriti ih u tablični oblik



# Polustrukturirani podaci

```
{
  "glossary": {
    "title": "example glossary",
    "GlossDiv": {
      "title": "S",
      "GlossList": {
        "GlossEntry": {
          "ID": "SGML",
          "SortAs": "SGML",
          "GlossTerm": "Standard Generalized Markup Language",
          "Acronym": "SGML",
          "Abbrev": "ISO 8879:1986",
          "GlossDef": {
            "para": "A meta-markup language, used to create markup languages such as DocBook.",
            "GlossSeeAlso": ["GML", "XML"]
          },
          "GlossSee": "markup"
        }
      }
    }
  }
}
```

JSON

```
<!DOCTYPE glossary PUBLIC "-//OASIS//DTD DocBook V3.1//EN">
<glossary><title>example glossary</title>
<GlossDiv><title>S</title>
<GlossList>
  <GlossEntry ID="SGML" SortAs="SGML">
    <GlossTerm>Standard Generalized Markup Language</GlossTerm>
    <Acronym>SGML</Acronym>
    <Abbrev>ISO 8879:1986</Abbrev>
    <GlossDef>
      <para>A meta-markup language, used to create markup
languages such as DocBook.</para>
      <GlossSeeAlso OtherTerm="GML">
        <GlossSeeAlso OtherTerm="XML">
</GlossDef>
      <GlossSee OtherTerm="markup">
</GlossEntry>
</GlossList>
</GlossDiv>
</glossary>
```

XML

# Nestrukturirani podaci



- Podaci nemaju definiranu strukturu s jasnim značenjem
- Primjer: Wikipedia
  - 200+ jezika
  - Preko 50 milijuna članaka
  - Bogata različitim oblicima podataka (tekst, slike, poveznice, glazbeni zapisi...)

# Wikidata



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Current events](#)  
[Random article](#)  
[About Wikipedia](#)  
[Contact us](#)  
[Donate](#)

[Contribute](#)

[Help](#)  
[Learn to edit](#)  
[Community portal](#)  
[Recent changes](#)  
[Upload file](#)

[Tools](#)

[What links here](#)  
[Related changes](#)  
[Special pages](#)  
[Permanent link](#)  
[Page information](#)  
[Cite this page](#)

[Wikidata item](#)

[Print/export](#)

[Download as PDF](#)  
[Printable version](#)

[In other projects](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)



## Zagreb

From Wikipedia, the free encyclopedia

Coordinates: 45°49′N 15°59′E

*This article is about the Croatian capital city. For other uses, see [Zagreb \(disambiguation\)](#).*

**Zagreb** (/ˈzɑːɡreɪb, ˈzæɡreɪb, zɑːˈɡreɪb/ *ZAH-gr**eb**, ZAG-r**eb**, zah-GR**EB***,<sup>[8]</sup> Croatian: [zâ:ɡrɛb]  (ⓘ) (ⓘ) (ⓘ)<sup>[9]</sup> is the capital and largest city of Croatia.<sup>[10]</sup> It is in the northwest of the country, along the [Sava](#) river, at the southern slopes of the [Medvednica](#) mountain. Zagreb lies at an elevation of approximately 122 m (400 ft) above sea level.<sup>[11]</sup> The estimated population of the city in 2018 was 804,507.<sup>[6]</sup> The population of the Zagreb urban agglomeration is 1,153,255,<sup>[2]</sup> approximately a quarter of the total population of Croatia.

Zagreb is a city with a rich history dating from [Roman](#) times. The oldest settlement in the vicinity of the city was the Roman [Andautonia](#), in today's [Šćitarjevo](#).<sup>[12]</sup> The name "Zagreb" is recorded in 1134, in reference to the foundation of the settlement at [Kaptol](#) in 1094. Zagreb became a [free royal city](#) in 1242.<sup>[13]</sup> In 1851 Zagreb had its first mayor,<sup>[14]</sup> Janko Kamauf.

Zagreb has special status as a Croatian administrative division and is a consolidated city-county (but separated from [Zagreb County](#)),<sup>[15]</sup> and is administratively subdivided into [17 city districts](#).<sup>[16]</sup> Most of them are at a low elevation along the river [Sava valley](#), whereas northern and northeastern city districts, such as [Podsljeme](#)<sup>[17]</sup> and [Sesvete](#)<sup>[18]</sup> districts are situated in the foothills of the [Medvednica](#) mountain,<sup>[19]</sup> making the city's geographical image rather diverse. The city extends over 30 kilometres (19 miles) east-west and around 20 kilometres (12 miles) north-south.<sup>[20][21]</sup>

Zagreb is considered a [global city](#) with a Beta-rating from the [Globalization and World Cities Research Network](#).<sup>[22]</sup>

The transport connections, concentration of industry, scientific, and research institutions and industrial tradition underlie its leading economic position in Croatia.<sup>[23][24][25]</sup> Zagreb is the seat of the [central government](#), [administrative bodies](#), and almost all [government ministries](#).<sup>[26][27][28]</sup> Almost all of the [largest Croatian companies](#), [media](#), and scientific institutions have their headquarters in the city. Zagreb is the most important transport hub in Croatia where [Central Europe](#), the [Mediterranean](#) and [Southeast Europe](#) meet, making the Zagreb area the centre of the road, rail and air networks of Croatia. It is a city known for its diverse [economy](#), high quality of living, [museums](#), sporting, and entertainment events. Its main branches of economy are [high-tech](#) industries and the [service sector](#).

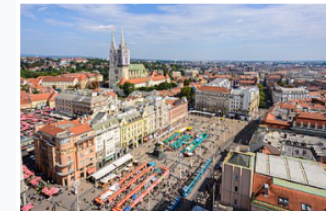
### Contents [hide]

- [Name](#)
- [History](#)

### Zagreb

#### Capital city

#### Grad Zagreb City of Zagreb



# Wikidata

- Strukturirana Wikipedia, slična bazi podataka
- Npr. {fr:Suisse, de:Schweiz, it:Svizzera, en:Switzerland, ...} → Q39  
(<https://www.wikidata.org/wiki/Q39>)
- Omogućen je pristup putem API-ja (<https://www.wikidata.org/w/api.php>) kao i preuzimanje čitave baze
- Dostupan u obliku
  - JSON (dokumentni model)
  - RDF (mrežni model)

# Pretraživanje i preuzimanje podataka s web stranica

- Mnogo HTML podataka koji se mogu preuzeti
  - Skup podataka Common Crawl – oko 3.15 milijardi web stranica, 360 TB na Amazon S3 – ogromno!
- Ako tražimo pojedinačne web stranice: ***crawler/scrapper***. Apache Nutch, Apache Storm, Heritrix 3, Scrapy, BeautifulSoup, itd.
  - Npr. <https://beautiful-soup-4.readthedocs.io/en/latest/>

# Korištenje web usluga

- Većina velikih web sjedišta **ne preporuča** prikupljanje podataka korištenjem *crawlera/scrapera*
- Umjesto toga koristi se API za web usluge
- Najčešći arhitekturni stil za to: **REST** (engl. *Representational State Transfer*)
  - Zahtijeva se URL od poslužitelja putem HTTP protokola
  - Poslužitelj odgovara s tekstnom datotekom (npr., JSON, XML, običan tekst)
  - Klijent obrađuje primljene podatke

# Primjer korištenja REST-a

```
{
  "user": {
    "name": "Jane",
    "gender": "female",
    "location": {
      "href":
"http://www.example.org/us
/ny/new_york",
      "text": "New York"
    }
  }
}
```

← Ovaj resurs je opis korisnika koji se naziva Jane (format JSON)

- Traži ga se putem zahtjeva „**GET**” na resursov URL, npr. putem alata za skriptiranje curl (<https://curl.se/>):  
`curl http://www.example.org/users/jane/`
- Ako se treba mijenjati resurs, onda najprije „GET”, pa izmjena, pa zatim „PUT”
- Implikacija: klijenti koji prikupljaju informacije ne mogu biti prejednostavni, moraju razumjeti format resursa kako bi ga mogli obraditi!

# Modeli podataka – klasifikacija

- Model podataka je način na koji su podaci pohranjeni u računalu
- **Ravni model** (engl. *flat model*)
- **Relacijski model** (engl. *relational model*)
- **Dokumentni model** (engl. *document model*)
- **Mrežni model** (engl. *network model*)



# Ravni model

- Pohranjuje jedan tip entiteta, svih s istim očekivanim obilježjima, bez tablične strukture
- Primjer dnevnika poslužitelja – entiteti su zahtjevi od klijenata na poslužitelj

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1"
200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html) "
```

```
111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1" 200 10801
"http://www.google.com/search?q=in+love+with+ada+lovelace+what+to+do&ie=utf
-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914
Firefox/2.0.0.7"
```

# Relacijski model

- Pohranjuje više tipova entiteta, povezanih međusobno
- **Svugdje prisutan:**
  - MySQL, PostgreSQL, Oracle, DB2, SQLite, ...
- Podaci su predstavljeni tablicama (relacijama) koje opisuju entitete i njihove veze
- Većina ostalih modela podataka može se svesti na relacijski model

Id	ime
1	Trump
2	Obama
3	Biden

predsj ednik	naslj ednik
1	3
2	1

# SQL za obradu podataka

- Deklarativan jezik za rukovanje relacijskim podacima
- Navodi se što se želi napraviti, ne kako da se nešto izračuna
- Desno: primjer korištenja SQL-a u Pythonu
- Napomena: na ovom predmetu SQL i baze podatak se ne obrađuju detaljno

```
#!/usr/bin/python

import MySQLdb

# Open database connection
db = MySQLdb.connect("localhost","testuser","test123","TESTDB" )

# prepare a cursor object using cursor() method
cursor = db.cursor()

sql = "SELECT * FROM EMPLOYEE \
       WHERE INCOME > '%d'" % (1000)
try:
    # Execute the SQL command
    cursor.execute(sql)
    # Fetch all the rows in a list of lists.
    results = cursor.fetchall()
    for row in results:
        fname = row[0]
        lname = row[1]
        age = row[2]
        sex = row[3]
        income = row[4]
        # Now print fetched result
        print "fname=%s,lname=%s,age=%d,sex=%s,income=%d" % \
              (fname, lname, age, sex, income )
except:
    print "Error: unable to fetch data"

# disconnect from server
db.close()
```

# Pandas – relacijski model u Pythonu

- Uz Python se najčešće koristi paket **Pandas** za tablični prikaz i obradu podataka
- **Tablica** u bazi podataka  $\leftrightarrow$  **DataFrame** u Pandasu
- Pandas je sličan SQL-u, s dodatnim elementima funkcijskog programiranja
  - `map()`, `filter()`, itd.
- Tema laboratorijskih vježbi

# Odnos Pandas – SQL

## Prednosti Pandasa

- Pandas je lagan i brz paket
- Pandas je pisan u nativnom Pythonu, čime je ekspresivniji od SQL-a (njegov je nadjezik)
- Velika potpora za statističku analizu podataka, kakvu SQL nema
- Lagana integracija s alatima za vizualizaciju (npr. Matplotlib)

## Nedostatci Pandasa

- **Tablica mora stati u glavnu memoriju** (obično su velike tablice napravljene u .HDFS formatu)
- Nije moguće indeksiranje nakon izgradnje tablice – indeksi se grade u trenutku izgradnje tablice
- Složeni *joinovi* su spori
- Nisu omogućene transakcije, izrada sigurnosne kopije tablice i promjena („journaling”) i drugo

# Dokumentni model

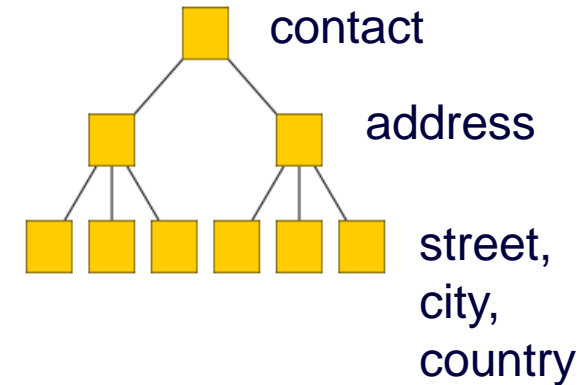
- Entiteti su organizirani u **hijerarhije**

XML format:

```
<contact>
  <id>656</id>
  <firstname>Chuck</firstname>
  <lastname>Smith</lastname>
  <phone>(123) 555-0178</phone>
  <phone>(890) 555-0133</phone>
  <address>
    <street>Rue de l'Ale 8</street>
    <city>Lausanne</city>
    <zip>1007</zip>
    <country>CH</country>
  </address>
</contact>
```

JSON format:

```
contact: {
  id: 656,
  firstname: "Chuck",
  lastname: "Smith",
  phones: ["(123) 555-0178",
           "(890) 555-0133"],
  address: {
    street: "Rue de l'Ale 8",
    city: "Lausanne",
    zip: 1007,
    country: "CH"
  }
}
```



# Dokumentni model

- Kako pretvoriti dokumentni model u relacijski?
- Za neke podatke lako, no što kada npr. imamo dva tel. broja za istu osobu?

**<phone>**(123) 555-0178**</phone>**  
**<phone>**(890) 555-0133**</phone>**

id	first name	...	id	phone
656	Chuck	...	656	(123) 555-0178
...	...	...	656	(890) 555-0133
...	...	...	...	...

Rješenje: isti entitet prikazati u više tablica

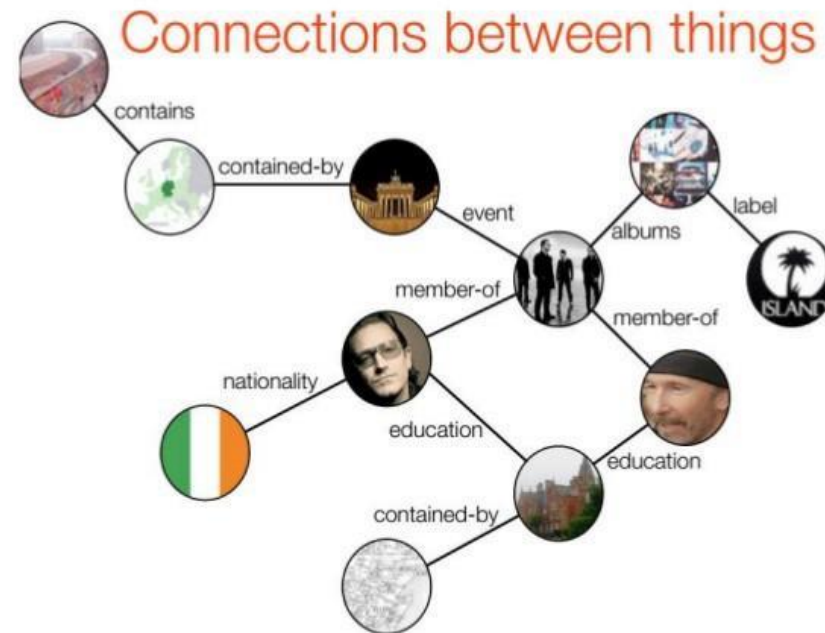
# Obrada podataka u formatu XML i JSON

- Struktura dokumenta = stablo
- Obrada putem prolaska kroz stablo
  - Pretraživanjem u dubinu ili u širinu
- Potrebno je koristiti odgovarajući jezik za upite
  - Npr. onaj implementiran u alatu jq
    - <https://stedolan.github.io/jq/>
  - Za XML postoje većinom jezično-specifične implementacije:
    - <https://www.edureka.co/blog/python-xml-parser-tutorial/#modules>



# Mrežni model

- Model u kojem su podaci prikazani kao čvorovi (vrhovi) i bridovi grafa
- Bridovi opisuju odnose među čvorovima
- U općenitom slučaju, svaki čvor u grafu može imati više čvorova roditelja i čvorova djece
- Čvorovi mogu biti složene strukture podataka
- Moguće ga je prevesti u tablice u relacijskom modelu ili u hijerarhijske strukture
- Implementiran u bazama podataka zasnovanima na grafovima – vrsti NoSQL baza



# Binarni formati zapisa podataka

- Prednosti
  - veća brzina parsiranja u odnosu na nebinarne (tekstne) zapise podataka
  - sažetiji opis (manje datoteke)
- Nedostatak je moguće nepoznavanje detalja formata i posljedična nemogućnost učitavanja (posebice za vlasničke – *proprietary* formate)
- Rad s binarnim formatima
  - Python **pickle**: <https://docs.python.org/3/library/pickle.html>
  - Java **Serializable**: <https://docs.oracle.com/javase/7/docs/api/java/io/Serializable.html>
  - Jezično neovisni: <https://developers.google.com/protocol-buffers/>
- Uvijek je potrebno razmotriti korištenje binarnih formata radi optimizacije resursa, pogotovo za velike podatke

# Alati i radni okviri

# Python

- U okviru predmeta fokus je na programskom jeziku Python, v3
- Dostupno je više radnih okruženja za rad u Pythonu, neke preporuke:
  - PyCharm : <https://www.jetbrains.com/pycharm/>
  - Spyder : <https://www.spyder-ide.org/>
  - PyDev : <https://www.pydev.org/>
- Slobodan izbor radnog okruženja
- U okviru projekta, predaju se implementacije u obliku bilježnice Jupyter Notebook
  - <https://jupyter.org/install.html>

# Jupyter Notebook

- IPython (<https://ipython.org/>) omogućuje stvaranje formatiranog dokumenta koji sadrži tekst, jednadžbe i kôd koji se naziva **Jupyter Notebook**
- Postoji i WYSIWYG sučelje koje se može koristiti: **JupyterLab**
- Moguće je više instalacijskih putova, jedan od jednostavnijih je:
  - Instalacija Pythona
  - Instalacija IPython
  - Instalacija Jupytera (vidi više na <https://jupyter.org/install>)
  - Pokretanje Jupyter Notebooka: <https://jupyter.readthedocs.io/en/latest/running.html>

# Radni okviri

- U okviru predmeta, radit će se sa sljedećim radnim okvirima:
- Pandas – rukovanje podacima
- Statistics, Scikit-learn i PyTorch – statistika i strojno učenje
- TensorFlow/Keras – duboko učenje
- Matplotlib, Seaborn – vizualizacija podataka
- NLTK – rad s tekstom
- I drugi...
- Također, razni specifični okviri ovisno o projektu

# Python++

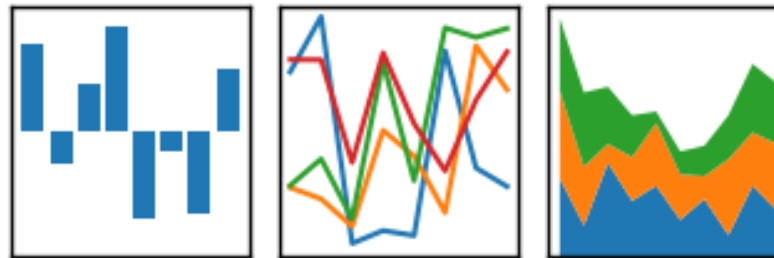


 PyTorch



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



# Zaključci

- Područje znanosti o podacima vrlo je široko i interdisciplinarno
- Fokus je na **podacima** – izvorima, upravljanju, analitici i vizualizaciji
- Izvori podataka mogu biti strukturirani, polustrukturirani ili nestrukturirani
- Modeli podataka mogu biti ravni, relacijski, dokumentni (hijerarhijski) i mrežni
- Postoji veliki broj popratnih alata i radnih okvira za rad u raznim programskim jezicima
- U ovom predmetu radit ćemo u Pythonu i njegovim tehnologijama



Source: geralt, Pixabay