

## Мастер академске студије

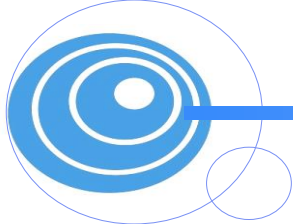
Модули:

Ваздушни саобраћај и транспорт  
и  
Водни саобраћај и транспорт

**Big Data технологије и технике**

Наставници (*e-mail*):

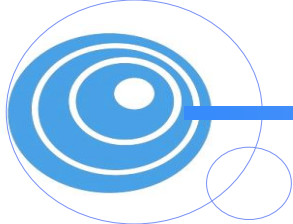
- др Снежана Младеновић, редовни професор ([s.mladenovic@sf.bg.ac.rs](mailto:s.mladenovic@sf.bg.ac.rs))
- др Слађана Јанковић, ванредни професор ([s.jankovic@sf.bg.ac.rs](mailto:s.jankovic@sf.bg.ac.rs))
- др Ана Узелац, доцент ([ana.uzelac@sf.bg.ac.rs](mailto:ana.uzelac@sf.bg.ac.rs))



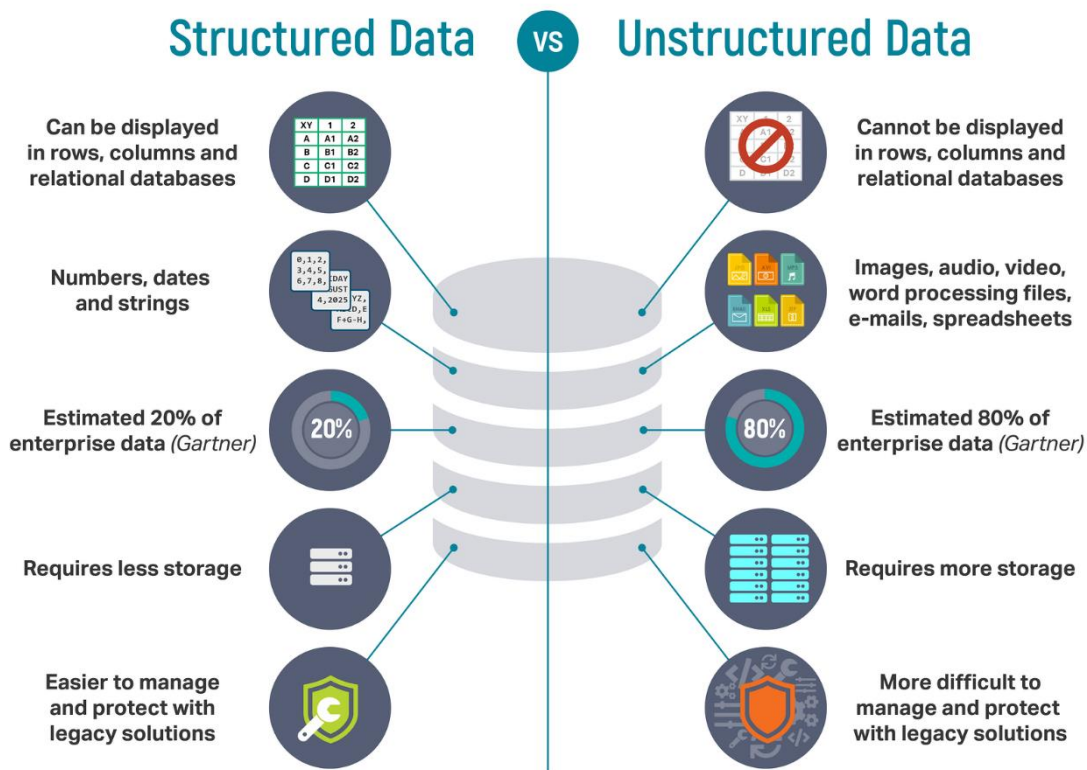
# Садржај



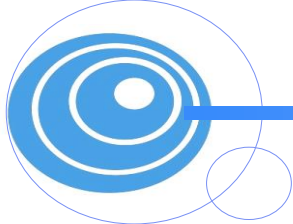
- Појам Big Data
- Big Data изазови који се јављају при складиштењу, обради и анализи скупова података



# Садржај



Структурирани и неструктурирани подаци

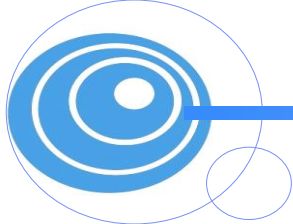


# Садржај



Нерелационе базе података (NoSQL)

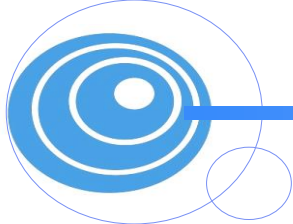
MongoDB база података



# Садржај

- Модели машинског учења
- Предиктивна аналитика заснована на моделима машинског учења:
  - Алгоритми класификације
  - Алгоритми регресије
  - Кластеровање





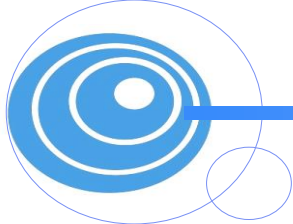
# Садржај



Data Visualization

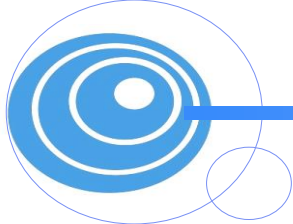


- Основне структуре података у Python-у
- Библиотеке које служе за процесирање података доступне у Python-у
- Библиотеке у Python-у које се користе за рад са вишедимензионим низовима
- Библиотеке за визуелизацију података



# Циљ предмета

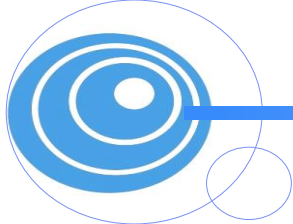
- Упознавање студената са појмом Big Data.
- Упознавање са изазовима који се јављају при складиштењу, обради и анализи скупова података који имају Big Data обележја.
- Савладавање употребе и интегрисање изабраних Big Data технологија и техника у решавању реалних проблема из саобраћаја и транспорта.



# Циљ предмета

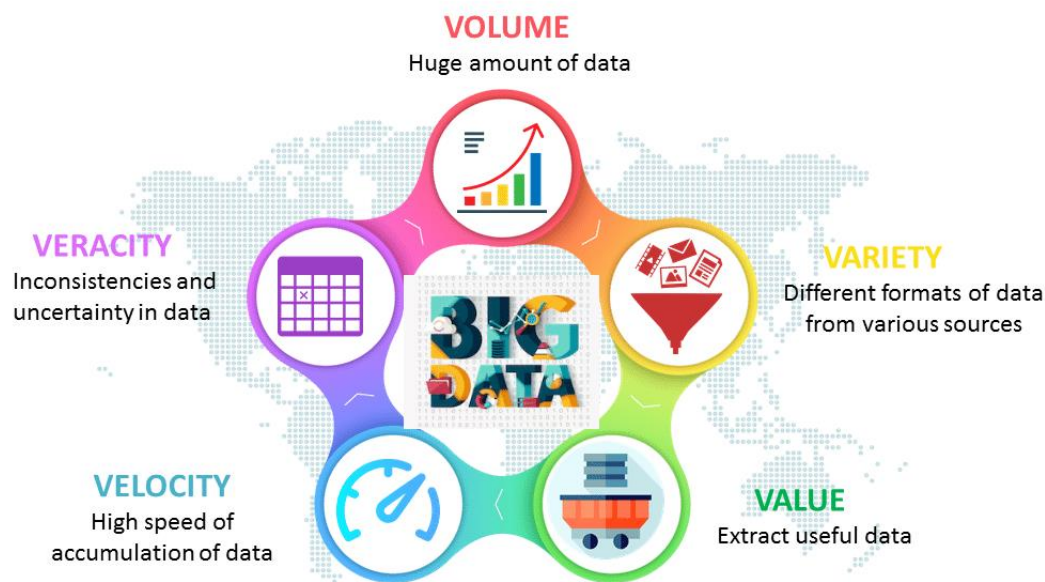
- Студенти треба да се обуче да:
  - препознају скупове података који имају Big Data обележја;
  - препознају реалне проблеме из саобраћаја и транспорта који могу бити решени коришћењем нерелационих база података и метода Big Data аналитике;
  - пројектују и креирају нерелационе базе података користећи *MongoDB* базе података;
  - врше предиктивну анализу применом техника машинског учења;
  - визуелизују података користећи библиотеке доступне у програмском језику Python.

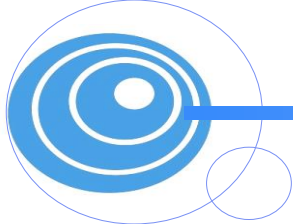




# Појам Big Data

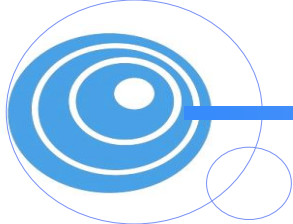
- Под појмом Big Data подразумева се информациони ресурс велике количине, велике брзине увећавања и велике разноврсности података, који превазилази могућности уобичајено коришћеног софтвера за складиштење, обраду и управљање подацима.





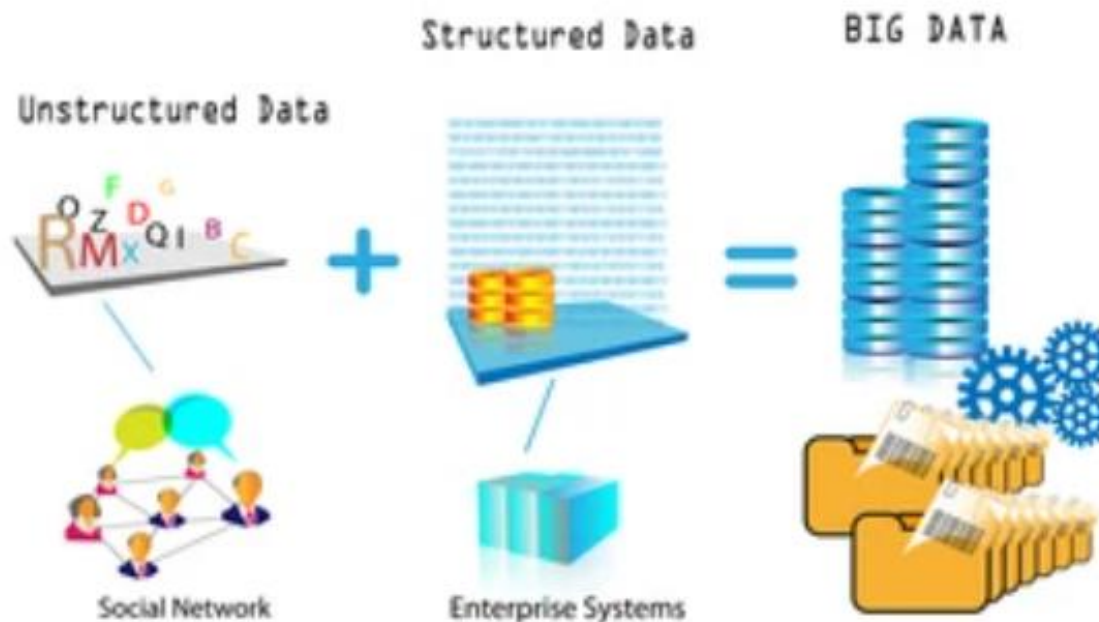
# Када су потребна Big Data решења

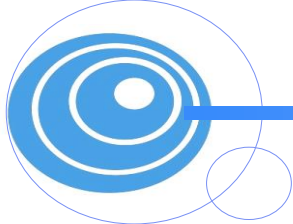
- Big Data решења има смисла развијати у следећим случајевима:
  - када је потребно да се анализирају не само структурирани, већ и полуструктурирани и неструктурирани подаци;
  - када је потребно анализирати сву расположиву количину података, или њен већи део, а не само неке карактеристичне примере;
  - када је потребно вршити итеративне истраживачке анализе.



# Big Data технологије које се изучавају у овом предмету

- **MongoDB** је водећа NoSQL база података код које се подаци чувају и организују у форми JSON (JavaScript Object Notation) докумената.

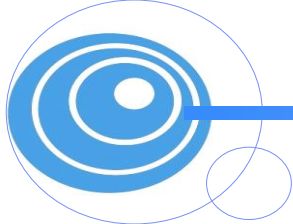




# Методе и алати Big Data аналитике који се изучавају у овом предмету

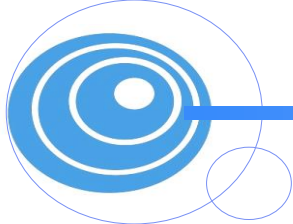


- Предиктивна анализа базирана на моделима машинског учења (алгоритми класификације, регресије и кластеровања, Sci-kit)
- Библиотеке које служе за процесирање података доступне у Python-у
- Библиотеке у Python-у које се користе за рад са вишедимонзионим низовима (NumPy)
- Библиотеке за визуелизацију података (Pandas, Matplotlib)
- Креирање упита над MongoDB базама података у програмском језику Python (библиотека PyMongo)

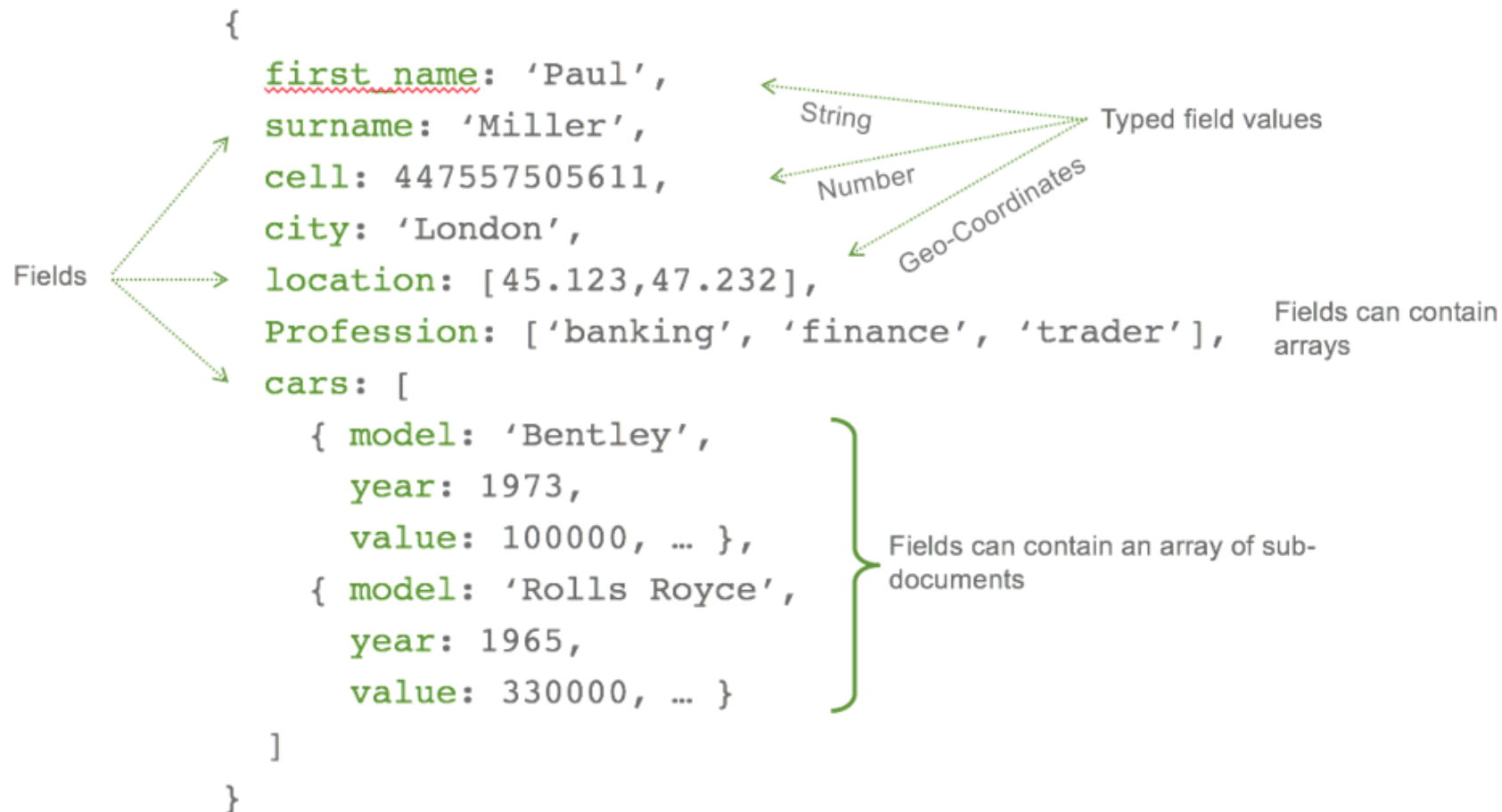


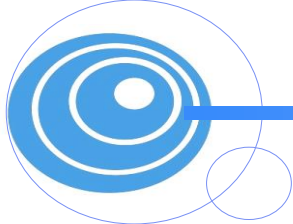
# Провера знања

- **Пројектни задатак (обавезан)** – самосталан рад студента уз консултације са предметним наставницима: **60 бодова**. Пројектни задатак може да се реализује на један од следећа два начина:
  - креирање нерелационе базе података и упита над њом на ***MongoDB*** платформи,
  - процесирање, обрада, анализа и визуелизација података коришћењем програмског језика ***Python***.
- **Завршни испит (није обавезан)** – семинарски рад или публиковани научно-стручни рад студента у сарадњи са предметним наставницима: **40 бодова**. Завршни испит обухвата: обраду, анализу и визуелизацију података коришћењем програмског језика ***Python***.



# Пример MongoDB документа





# Програмски код у Python-у

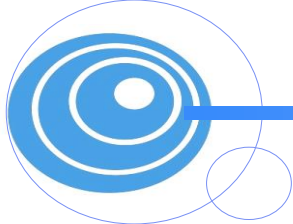
```
# Boxplots for numerical variables
sb.set_style("dark")

fig, ax =plt.subplots(1,4,figsize=(15, 5))

numeric_bool = train.dtypes != 'object'
numeric_var = train.dtypes[numeric_bool].index

for i in range(len(numeric_var)):
    g = sb.boxplot(x = 'Survived', y = numeric_var[i], data = train, ax=ax[i],
                  medianprops={'color':'red'})

    # Add median to the plot
    medians = train.groupby(['Survived'])[numeric_var[i]].median()
    for xtick in g.get_xticks():
        g.text(xtick , medians[xtick], medians[xtick], horizontalalignment='center',
              size='medium',color='w',weight='semibold')
        # xtick: the position to place the text in x axis
        # medians[xtick]: the position to place the text in y axis
        # medians[xtick]: the text
    g.set(xlabel="",ylabel="")
    g.set_xticklabels(['Survived: No','Survived: Yes'])
    g.set_title('{}'.format(numeric_var[i]))
```



# Литература

- Doug Bierer, *MongoDB 4 Quick Start Guide*, Packt, Birmingham – Mumbai 2018.
- Sinan Ozdemir, *Principles of Data Science*, Packt, Birmingham – Mumbai, 2016.
- Joel Grus, *Data Science from Scratch: First Principles with Python 2<sup>nd</sup> Edition*, O'Reilly, 2019.
- Ivan Marin, Ankit Shukla, et al., *Big Data Analysis with Python*, Packt, Birmingham – Mumbai, 2019.
- Jaime Buelta, *Python Automation Cookbook*, Packt, Birmingham – Mumbai, 2018.