Univerzitet u Beogradu – Elektrotehnički fakultet (ETF)
Katedra za signale i sisteme

# Osnovni podaci u R-u i učitavanje podataka iz fajlova

Dr Nadica Miljković, vanredni profesor
kabinet 68, nadica.miljkovic@etf.bg.ac.rs

# ZADACI
# TIPOVI PODATAKA U R-U

# Nizovi

- Proveriti koji su ulazni parametri funkcija *c()*, *vector()* i *is.vector()*?
- Kreirati sledeće nizove:
  - n1 = [0.6 0.8 0.7 0.7 0.9]
  - n2 = [T F F F T T];
  - n3 = [“blue” “red” “orange” “pink” “green”];
  - n4 = [0.8 T F T “gray”] Kog je tipa n4? *HINT*: koristiti funkciju *class()*.
- Za ovako kreirane nizove izvršiti eksplicitnu konverziju na sledeći način:
  - niz *n1* pretvoriti u niz karaktera
  - niz *n2* pretvoriti u niz numeričkih promenljivih
  - niz *n3* pretvoriti u kategoričku promenljivu (*factor*)
  - niz *n4* pretvoriti u *boolean* promenljivu *HINT*: koristiti *as.logical()* Koliko NA se dobija (proveriti funkcijom *is.na()* i funkcijom za eksplicitno “štampanje” *print()*) i zašto?

# Karakteri

```
> c1 <- c("jedan", "dva", "tri", "cetiri", "pet")
> c1
[1] "jedan"  "dva"    "tri"    "cetiri" "pet"
> class(c1)
[1] "character"
> is.vector(c1)
[1] TRUE
> c2 <- ('jedan', 'dva', 'tri', 'cetiri', 'pet')
Error: unexpected ',' in "c2 <- ('jedan',"
> c2 <- c('jedan', 'dva', 'tri', 'cetiri', 'pet')
> c2
[1] "jedan"  "dva"    "tri"    "cetiri" "pet"
> class(c2)
[1] "character"
> is.vector(c2)
[1] TRUE
>
```

- Ove promenljive se u R-u mogu označavati na 2 načina.
- Na slici su prikazani primeri kako se inicijalizuju promenljive tipa karakter u R-u.
- Koliko elemenata imaju *c1* i *c2*? Koja se funkcija koristi za računanje dužine niza?

# Matrice

```
> dimnames(m2) <- list(c("red1", "red2", "red3", "red4"), c("kol1", "kol2", "kol3"))
>
```

- Korišćenjem operatora ":" kreirati niz brojeva *m1* od 1 do 12 (sa korakom 1). R dokumentacija za ovaj operator se dobija za ?':'
- Potom primenom funkcije *matrix()* kreirati matricu *m2* od niza *m1* sa 4 kolona i 3 reda.
  - Šta se desi ako se umesto sa 3 reda kreira matrica sa 5 redova?
- Proveriti šta funkcija *dimnames()* za matricu *m2* daje na izlazu.
- Dodeliti nazive redova i kolona matrice *m2* kao na slici.
- Prikazati implicitno i eksplicitno matricu.
- Proveriti šta daju funkcije *colnames()* i *rownames()* na izlazu za matrice *m1* i *m2*.
- Ponoviti sve gore navedene zadatke za matricu koja ima elemente u obrnutom redosledu (*HINT*: 12:1).
- Kako bi se napravio niz sa korakom 0.5?

• • •

| ispitanik | visina [cm] | težina [kg] |
|---|---|---|
| 1 | 198 | 75 |
| 2 | 180 | 72 |
| 3 | 176 | 73 |
| 4 | 185 | 96 |
| 5 | 186 | 76 |

- Uneti u R-u matricu sa 3 kolone i 5 redova kao na slici. Tabela sadrži listu ispitanika i njihove podatke o visini i težini.
- Koristeći odgovarajuće funkcije uneti i nazive kolona kao na slici. Rezultujuću matricu prikazati u konzoli R-a.
- Lista ispitanika je preuzeta iz rada:
  - A. Gogić, N. Miljković, Đ. Đurđević. Electromyography-based gesture recognition: Fuzzy classification evaluation, *Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN*, ISBN: 978-86-7466-618-0, pp. MEI1.6.1-4, Zlatibor, Serbia, June 13-16, 2016.

# Liste

| ispitanik | visina [cm] | težina [kg] | dominantna ruka | pol |
|---|---|---|---|---|
| 1 | 198 | 75 | R | M |
| 2 | 180 | 72 | R | M |
| 3 | 176 | 73 | L | F |
| 4 | 185 | 96 | L | M |
| 5 | 186 | 76 | R | M |

- Na slici je prikazana kompletna lista ispitanika iz rada:
  - A. Gogić, N. Miljković, Đ. Đurđević. Electromyography-based gesture recognition: Fuzzy classification evaluation, *Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN*, ISBN: 978-86-7466-618-0, pp. MEI1.6.1-4, Zlatibor, Serbia, June 13-16, 2016.
- NAPOMENA: Značenje skraćenica u tabeli: R desno (od eng. *right*), L levo (od eng. *left*), M muški pol (od eng. *Male*) i F ženski pol (od eng. *Female*).
- Umesto matrice, kao u prethodnom primeru, kreirati listu *l1*.
- Od podataka o polu ispitanika kreirati poseban niz karaktera *p1* i potom kreirati faktorsku promenljivu *p2*.
- Zašto je lista pogodnija od matrice za ovaj tip podataka?
- Koji tip podataka je potrebno odabrati za podatke sa slike?

# Redosled nivoa u *factor* promenljivoj

```
> f1
 [1] jan feb mar apr maj jun jul avg sep okt nov dec
12 Levels: jan < feb < mar < apr < maj < jun < jul < avg < ... < dec
> |
```

- Redosled nivoa u faktorskoj promenljivoj je podrazumevano po abecednom redu, ali može se izmeniti sa argumentom *levels* unutar funkcije *factor()*.
- ZADATAK: Napraviti faktorsku promeljivu *f1* koja označava mesece u godini (koristiti prva tri slova za svaki mesec: jan, feb, mar, ...) tako da su nivoi raspoređeni prema redosledu meseci u godini.
- Očekivani rezultat je prikazan na slici.

```
> meseci
 [1] "jan" "feb" "mar" "apr" "maj" "jun" "jul" "avg" "sep" "okt" "nov"
[12] "dec"
> f1 = factor(meseci,levels=c("jan", "feb", "mar", "apr", "maj", "jun", "jul"
, "avg", "sep", "okt", "nov", "dec"),ordered=TRUE)
> f1
 [1] jan feb mar apr maj jun jul avg sep okt nov dec
12 Levels: jan < feb < mar < apr < maj < jun < jul < avg < ... < dec
> |
```

REŠENJE

# NA i NaN vrednosti

| težina |
|---|
| 68 |
| 75 |
| NA |
| 73 |
| 92 |
| NA |
| 94 |
| 67 |

- Kreirati niz kao sa slike. *HINT*: koristiti *names()* i *c()* funkcije.
- Izračunati automatski koliko ima NA vrednosti u nizu.
- Pogledati kako se koristi funkcija *mean()*.
- Kolika je srednja vrednost težine ispitanika u ovoj studiji?
- Kako se koristi funkcija *mean()* ako se umesto NA upiše NaN?
- Ako su nedostajuće vrednosti, *niz*[3] = 86 i *niz*[6] = 76, koliko se promeni (u procentima) srednja vrednost težine svih ispitanika?
- DODATNO: Primenom *sd()* funkcije utvrditi da li se i koliko promeni i standardna devijacija na ovom uzorku.

# NEDOSTAJUĆE VREDNOSTI

# *Data frame* podaci

| ispitanik | visina [cm] | težina [kg] | dominantna ruka | pol |
|---|---|---|---|---|
| 1 | 198 | 75 | R | M |
| 2 | 180 | 72 | R | M |
| 3 | 176 | 73 | L | F |
| 4 | 185 | 96 | L | M |
| 5 | 186 | 76 | R | M |

- Za tabelu sa slike, umesto liste kao u prethodnom zadatku, kreirati *data frame d1*. Iskoristiti *class()* funkciju i proveriti kom tipu podataka pripada *d1*.
- Eksplicitno prikazati na ekranu sadržaj promenljive *d1*.
- Lista ispitanika je preuzeta iz rada:
  - A. Gogić, N. Miljković, Đ. Đurđević. Electromyography-based gesture recognition: Fuzzy classification evaluation, *Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN*, ISBN: 978-86-7466-618-0, pp. MEI1.6.1-4, Zlatibor, Serbia, June 13-16, 2016.

# Za kraj rada sa tipovima podataka

- Izlistati sve promenljive komandom *ls()*.
- Proveriti koje se promenljive nalaze u memoriji R sesije u kartici *Environment*.
- Dodatno, moguće je proveriti i fajlove u radnom direktorijumu komandama *dir()* i *list.files():* proveriti R dokumentaciju za ove komande. Argumenti bilo koje funkcije/komande iz R dokumentacije se mogu dobiti i primenom *args()* funkcije.

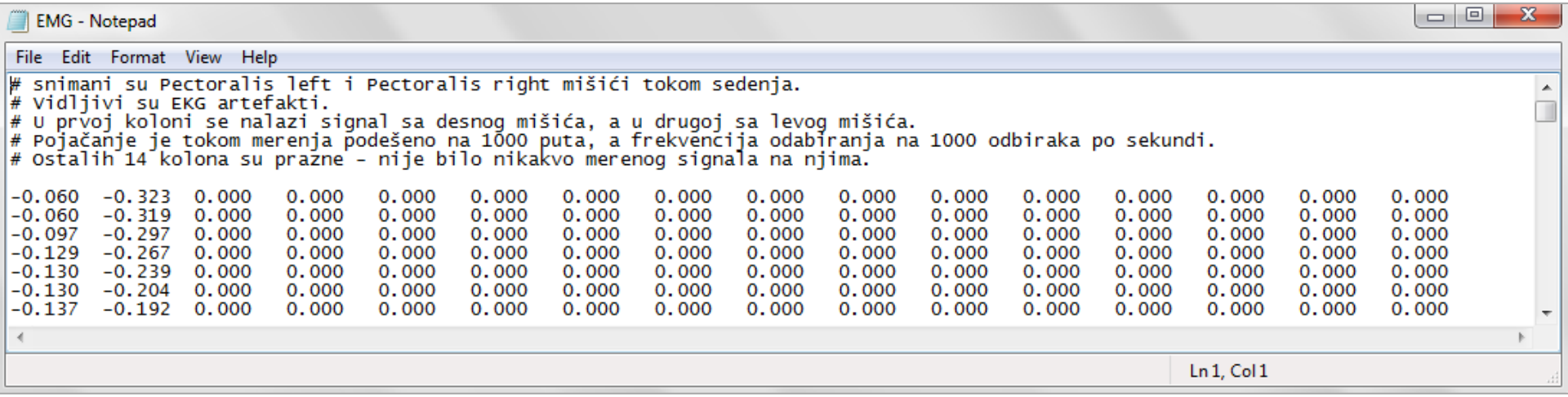
UČITAVANJE PODATAKA

# Radni direktorijum

- Za učitavanje podataka iz fajlova, potrebno je da radni direktorijum R-a bude podešen tako da se fajl iz koga se čitaju podaci nalazi u tom radnom direktorijumu.
- Da bi se proverilo koji je aktivan radni direktorijum, moguće je ukucati komandu *getwd()* u kozoli. Moguće je i pogledati u *Files* kartici R Studio interfejsa.
- Radni direktorijum se podešava iz padajućeg menija *Session* / *Set Working Directory* / *Choose Directory* ... ili korišćenjem komande *setwd()*.

# *read.table()* funkcija

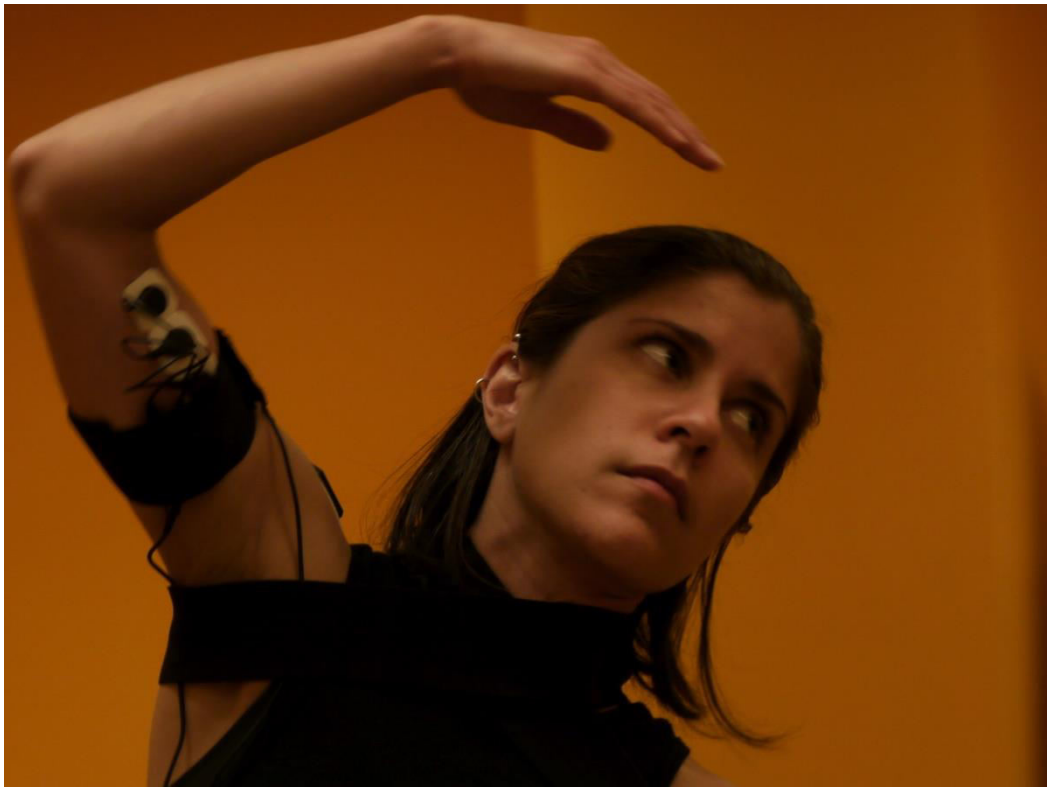
```
> EMGsignali <- read.table("EMG.txt")
> head(EMGsignali)
      V1     V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1 -0.060 -0.323  0  0  0  0  0  0  0   0   0   0   0   0   0   0
2 -0.060 -0.319  0  0  0  0  0  0  0   0   0   0   0   0   0   0
3 -0.097 -0.297  0  0  0  0  0  0  0   0   0   0   0   0   0   0
4 -0.129 -0.267  0  0  0  0  0  0  0   0   0   0   0   0   0   0
5 -0.130 -0.239  0  0  0  0  0  0  0   0   0   0   0   0   0   0
6 -0.130 -0.204  0  0  0  0  0  0  0   0   0   0   0   0   0   0
>
```




- Potrebno je učitati podatke koji su snimljeni u tekstualni fajl pod nazivom "EMG.txt" i prikazati prvih vrsta učitanog data frame-a pomoću funkcije *head()*.
- Koji radni direktorijum je potrebno odabrati prilikom učitavanja ovih signala?
- Zašto nisu učitani komentari u promenljivu *EMGsignali*?

# EMG signali – prikaz

- Kog tipa je promenljiva *EMGsignali*?
- Korišćenjem funkcije *head()* prikazati prve 3 vrste promenljive *EMGsignali*?
- Korišćenjem funkcije *tail()* prikazati poslednje četiri vrste promenljive *EMGsignali*?
- Korišćenjem funkcije *tail()* prikazati poslednjih šest vrsta promenljive *EMGsignali*?



# *read.csv()* funkcija

- U dokumentu “tabela.csv” (koji je dostupan na sajtu predmeta: http://automatika.etf.rs/sr/13m051tobs pod imenom *csv_tabela*), učitati podatke korišćenjem *read.csv()* funkcije.
- Osim naziva fajla koji se unosi kao ulazni parametar ove funkcije, koji parametar je još potrebno promeniti? Obrazložiti.
- Proveriti tip podataka promenljive u kojoj su učitani podaci.

# Procena potrebne memorije

- Za podatke u fajlu “EMG.txt”, proveriti koliko je memorije $M_{podaci}$ potrebno za smeštanje numeričkih podataka.
- Proveriti kolika je RA memorija $M_{RAM}$ na računaru?
- Da li je ispunjen uslov $M_{podaci} < M_{RAM} / 2$?
- Šta će se desiti ako nije ispunjen ovaj uslov?



# Rešenje

- Podaci u fajlu “EMG.txt” imaju 39166 vrsta i 16 kolona numeričkih podataka.
- Otprilike, memorija koja je potrebna za smeštanje ove količine podataka je:
  - 39166 x 16 x 8 B/numeric =
  - 5013248 B (/ $2^{20}$ MB) =
  - 4.781 MB -> uporediti ovu vrednost sa RAM-om računara (treba da bude manji od RAM/2 ...)

REČNIK

# Funkcije i komande za danas

- vector()
- is.vector()
- c()
- as.*()
- length()
- class()
- is.na()
- print()
- ?':'
- matrix()
- dimnames()
- colnames()
- rownames()
- list()
- factors()
- factor()
- names()
- mean()
- sd()
- data.frame()
- ls()
- dir(), list.files()
- args()
- getwd()
- setwd()
- read.table()
- head()
- tail()
- read.csv()
- .Machine

# Količina podataka u jedinicama i prefiksima

**Multiples of bytes** V·T·E

| Decimal | | | Binary | | | | |
|---|---|---|---|---|---|---|---|
| Value | | Metric | Value | | IEC | JEDEC | |
| 1000 | kB | kilobyte | $1024$ | KiB | kibibyte | KB | kilobyte |
| $1000^2$ | MB | megabyte | $1024^2$ | MiB | **mebibyte** | MB | megabyte |
| $1000^3$ | GB | gigabyte | $1024^3$ | GiB | gibibyte | GB | gigabyte |
| $1000^4$ | TB | terabyte | $1024^4$ | TiB | tebibyte | – | |
| $1000^5$ | PB | petabyte | $1024^5$ | PiB | pebibyte | – | |
| $1000^6$ | EB | exabyte | $1024^6$ | EiB | exbibyte | – | |
| $1000^7$ | ZB | zettabyte | $1024^7$ | ZiB | zebibyte | – | |
| $1000^8$ | YB | yottabyte | $1024^8$ | YiB | yobibyte | – | |

Orders of magnitude of data

https://en.wikipedia.org/wiki/Mebibyte

# Do sledećeg časa ...

- Za domaći (opciono): Pokrenuti SWIRL i uraditi lekciju pod nazivom *Workspace and Files* (u ovom vežbanju postoje i elementi koji se ne rade na TOBS predmetu).
  - NAPOMENA: Pre pokretanja *swirl()* komande potrebno je učitati i biblioteku/paket komandom *library(swirl)*.

ZADATAK