

Predstavitev programskih orodij Python in R

dr. Maja ZAGORŠČAK, Nacionalni inštitut za biologijo

dr. Carissa BLEKER, Nacionalni inštitut za biologijo

dr. Jan ZRIMEC, Nacionalni inštitut za biologijo

dr. Marko PETEK, Nacionalni inštitut za biologijo

dr. Anže ŽUPANIČ, Nacionalni inštitut za biologijo

Ljubljana in online, Centralna tehniška knjižnica Univerze v Ljubljani,
Usposabljanje podatkovnih strokovnjakov, 18. – 21. 11. 2024





Vsebina

- Programski jeziki
- Razvojno okolje (IDE)
- Splošno o R-u in Python-u
- Markdown
- R Markdown
- Rstudio IDE
- knitr
- Visual Studio Code
- Jupyter Notebook
- Primeri v R-u in Python-u

Programski jeziki



https://github.com/drebee/programming_languages_word_cloud

- Sintaksa (slovnica): pravila, kako se piše kodo

```
class HelloWorld {  
    public static void main(String[] args) {  
        System.out.println("Hello world!");  
    }  
}
```

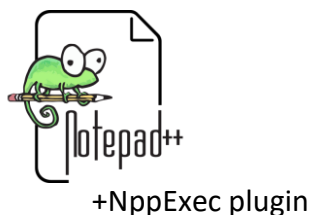
```
#include <stdio.h>  
int main() {  
    printf("Hello world!");  
    return 0;  
}
```

```
print("Hello world!")
```

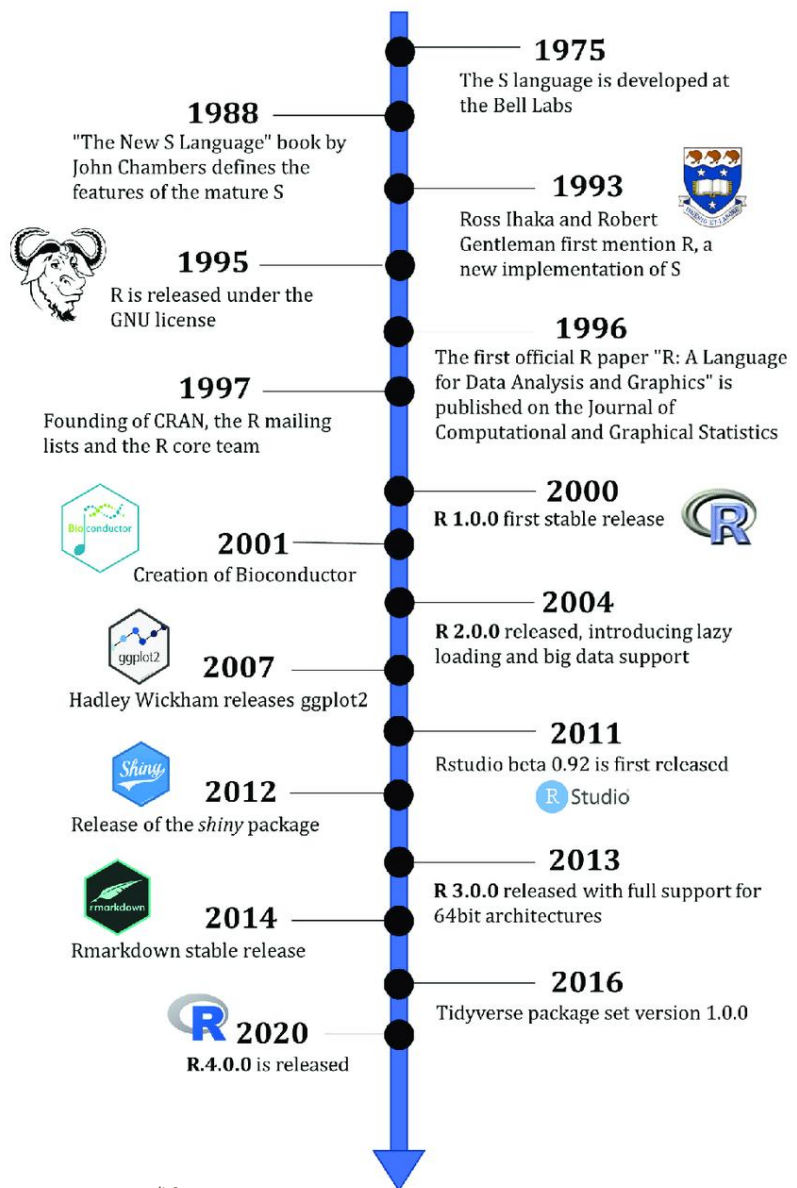
- Semantika
 - statična: preverjanje, ali je program smiseln (npr preverjanje deklaracije spremenljivk)
 - dinamična: kako se program izvede
 - denotacijska: matematični pomen programa

Razvojno okolje (IDE)

- Integrirano razvojno okolje (*Integrated **D**evelopment **E**nvironment, **IDE***) je računalniška aplikacija, ki zagotavlja množico orodij za razvoj programov
- Vsebuje urejevalnik izvirne kode (*source-code editor*), prevajalnik (*compiler*), povezovalnik (*linker*), razhroščevalnik (*debugger*) in razna orodja za preverjanje in izboljšave delovanja programa



R in Python





R in Python



Osnove	analitično programiranje, ki vključuje statistično računanje in grafiko	vsestranski računalniški jezik za analizo podatkov
Primarni uporabniki	znanstveniki in raziskovalci	programerji in razvijalci
Uporabnost	različni paketi in funkcije za statistično analizo in predstavitev podatkov	obsežne naloge obdelave podatkov, izdelava aplikacij grafičnega vmesnika in spletnih aplikacij
Področja uporabe	podatkovna znanost	podatkovna znanost
Knjižnice in paketi	ggplot2, caret, randomForest, xgboost, data table, keras, shiny	Numpy, SciPy, Scikit-learn, Pandas, PyTorch, Tensorflow, Keras, PyQt5
Krivulja učenja	sigmoidna	linearna
Upravljanje z različicami	renv, rig	pyenv, venv
Navzkrižna povezava	reticulate	rpy2



Markdown

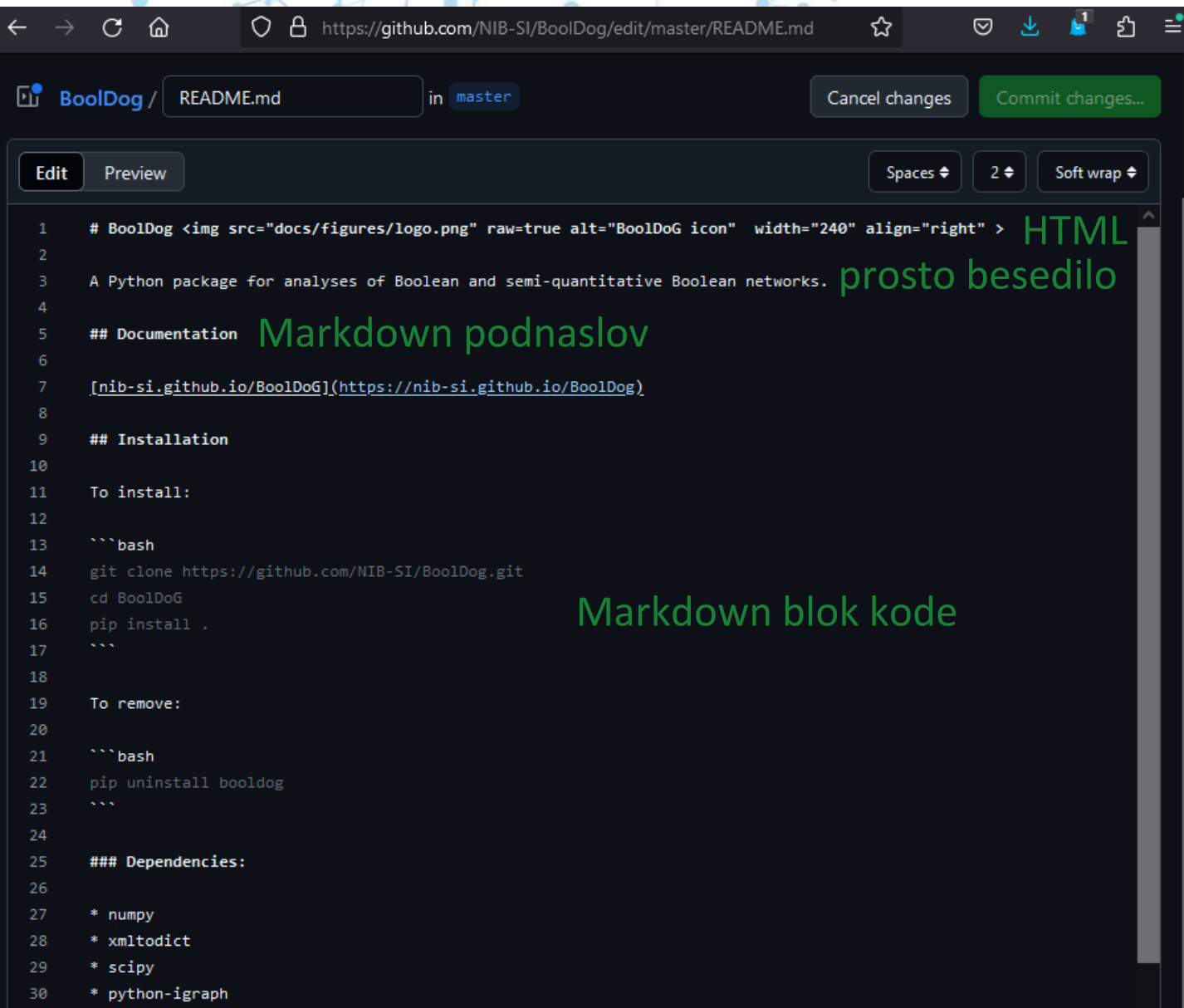
- Opisni jezik (*lightweight Markup Language*) za ustvarjanje oblikovanega besedila
- Neodvisen od strojne in sistemske programske opreme (podobno kot HTML *Hypertext Markup Language* za prikazovanje informacij v spletnem brskalniku)
- Zagotavlja način ustvarjanja enostavno berljive datoteke navadnega besedila, ki lahko vsebuje oblikovano besedilo, slike, naslove in povezave do drugih dokumentov
- Pogosto uporablja za bloge, v sodelovalni programski opremi (*collaborative software*), na straneh z dokumentacijo, in v t.i. datotekah 'Readme'



R Markdown

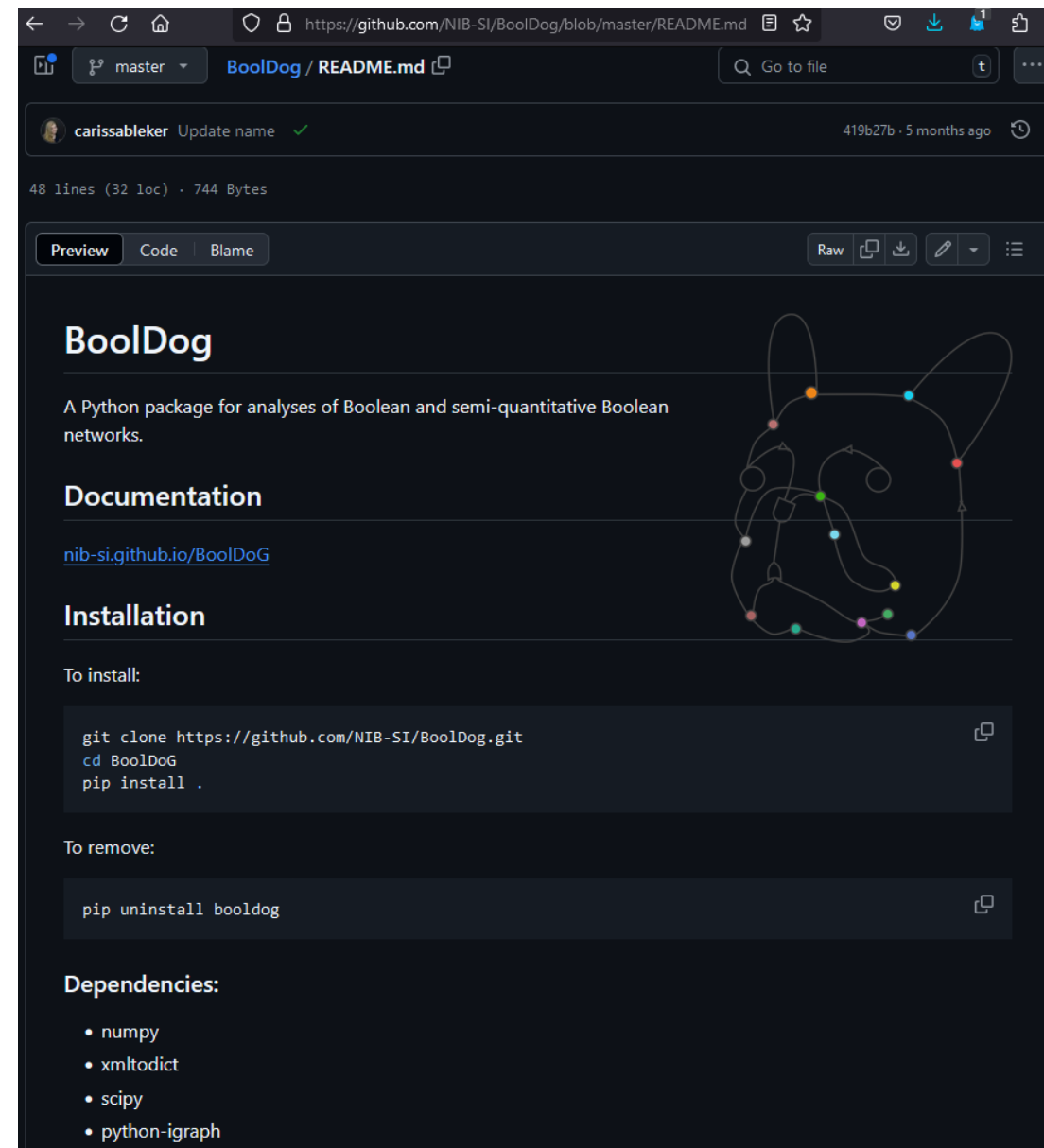
- Različica Markdown-a
- Uporablja se za združevanje R kode, rezultatov analize podatkov, grafov, tabel in komentarjev v ponovljiv dokument kot so poročilo, publikacija, spletna stran

Markdown in README.md



The screenshot shows the GitHub web editor interface for the file `README.md` in the repository `NIB-SI/BoolDog`. The editor is in the `master` branch. The code is written in Markdown and includes an image tag for a logo, a description of the package, documentation links, installation instructions, and a list of dependencies. The code is as follows:

```
1 # BoolDog 
2
3 A Python package for analyses of Boolean and semi-quantitative Boolean networks.
4
5 ## Documentation
6
7 [nib-si.github.io/BoolDog](https://nib-si.github.io/BoolDog)
8
9 ## Installation
10
11 To install:
12
13 ```bash
14 git clone https://github.com/NIB-SI/BoolDog.git
15 cd BoolDog
16 pip install .
17 ```
18
19 To remove:
20
21 ```bash
22 pip uninstall booldog
23 ```
24
25 ### Dependencies:
26
27 * numpy
28 * xmldict
29 * scipy
30 * python-igraph
```



The screenshot shows the GitHub preview view of the `README.md` file. The preview displays the rendered content of the Markdown file, including the BoolDog logo, the package description, documentation link, installation instructions, and dependencies. The preview is titled `BoolDog` and shows the following content:

BoolDog

A Python package for analyses of Boolean and semi-quantitative Boolean networks.

Documentation

nib-si.github.io/BoolDog

Installation

To install:

```
git clone https://github.com/NIB-SI/BoolDog.git
cd BoolDog
pip install .
```

To remove:

```
pip uninstall booldog
```

Dependencies:

- numpy
- xmldict
- scipy
- python-igraph



- Integrirano razvojno okolje (IDE) za programski jezik R
- Na voljo je v dveh oblikah:
 - RStudio Desktop - namizna aplikacija
 - RStudio Server - na oddaljenem strežniku, omogoča dostop do RStudia s spletnim brskalnikom



- Programski mehanizem (*software engine*) za dinamično generiranje poročil z R
- R paket, ki omogoča integracijo kode R v dokumente LaTeX, HTML, Markdown, ...
- Omogoča ponovljive raziskave

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Adds

03.3_Phenomics-plots.Rmd

Knit Run

Source Visual

```
1 ---
2 title: "03.3_Phenomics-plots"
3 author: "zagor"
4 date: "r Sys.Date()"
5 output:
6   html_document:
7     fig_caption: yes
8     self_contained: yes
9     fig_width: 12
10    fig_height: 9
11    toc: true
12    toc_float:
13      toc_collapsed: false
14    toc_depth: 5
15    number_sections: yes
16    theme: flatly
17    highlight: tango
18 pdf_document:
19   toc: yes
20   toc_depth: '5'
21 word_document:
22   toc: yes
23   toc_depth: '5'
24 editor_options:
25   chunk_output_type: console
26 ---
27
28 {r setup, include=FALSE}
29 knitr::opts_chunk$set(dev = c('pdf', 'png'), # this embeds pdf and crates scrolable blocks
30   dev = c('png'),
31   fig.align = 'center',
32   fig.height = 9,
33   fig.width = 12,
34   warning = FALSE, message = FALSE
35 )
```

Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Knit Directory
Clear Knitr Cache...

Run Selected Line(s) Ctrl+R
Run Current Chunk Ctrl+Shift+Enter
Run Next Chunk Ctrl+Alt+N
Run Setup Chunk
Run All Chunks Above Ctrl+Alt+Shift+P
Run All Chunks Below
Run All Ctrl+Alt+R

input 15/11/2024 01:25 File folder
other 05/11/2024 22:01 File folder
output 15/11/2024 01:20 File folder
reports 15/11/2024 01:19 File folder
scripts 15/11/2024 01:23 File folder
_ASSAY_METADATA.TXT 04/11/2024 18:58 TXT File
README.MD 15/11/2024 15:28 MD File

354:1 Chunk 10 R Markdown

Console Terminal Background Jobs

R 4.4.1 C:/Users/majaz/Desktop/GitHub/multiOmics-integration/p_ADAPTOmics/I_Desiree/S_multiOmics/A_multiOmicsStat-R/scripts/

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.utf8 LC_CTYPE=English_United Kingdom.utf8 LC_MONETARY=English_United Kingdom.utf8
[4] LC_NUMERIC=C LC_TIME=English_United Kingdom.utf8

time zone: Europe/Ljubljana
tzcode source: internal

attached base packages:
[1] grid stats graphics grDevices utils datasets methods base

other attached packages:
[1] plyr_1.8.9 crayon_1.5.3 scales_1.3.0 RColorBrewer_1.1-3 gridExtra_2.3 multcompview_0.1-10
[7] ggplot2_3.5.1 magrittr_2.0.3

loaded via a namespace (and not attached):

Project: (None)

Environment History Connections Tutorial

Import Dataset 262 MiB

R Global Environment

subset 92 obs. of 3 variables
subset.1 6 obs. of 17 variables
wilcoxTest 836 obs. of 11 variables

values

fn "data_targeted.xlsx"
fp "../input"
i "qL"

03.3_Phenomics-plots.html 14/11/2024 23:54 Firefox HTML Doc... 1,543 KB
03.3_Phenomics-plots.Rmd 14/11/2024 23:54 RMD File 10 KB

omicsLevel "Phenomics"
seed 123456
string ""

Functions

%!in% function (...)
%!nin% function (...)
summarySE function (data = NULL, measurevar, groupvars = NULL, na.rm = TRUE, conf.i...

Files Plots Packages Help Viewer Presentation

Zoom Export Publish

Phenomics

COMPACTNESS FvI/Fm HEIGHT ISOTROPY PSII

qL qP ROUNDNESS SIDE-AREA SOL

T TOP-AREA VOLUME_MM^3 WATER CONSUMPTION WIDTH

measurement (+/- SE)

DAS

Condition

C D H HD W

```
136 # palette podnaslov
137
138 # palette podnaslov
139
140 {r, palette, echo=TRUE, warning=FALSE, message=FALSE}
141
142 mypalette = c("#7F7F7F", "#FFC000", "#C00000", "#ED7D31", "#4472C4")
143
144
145
146
147 # data podnaslov
148
149
150
151 {r, data, echo=TRUE, warning=FALSE, message=FALSE}
152
153
154 fp = file.path('..', 'input')
155
156 fn = 'data_targeted.xlsx'
157
158
159 pheno = openxlsx::read.xlsx(xlsxFile = file.path(fp, fn),
160                             sheet = 'phenomics-noninvasive',
161                             startRow = 2,
162                             colNames = TRUE,
163                             rowNames = FALSE,
164                             detectDates = FALSE,
165                             skipEmptyRows = TRUE,
166                             skipEmptyCols = TRUE,
167                             rows = NULL,
168                             cols = NULL,
169                             check.names = FALSE,
170                             sep.names = ".",
171                             namedRegion = NULL,
172                             na.strings = "NA",
173                             fillMergedCells = FALSE)
```



chunk - kos kode

chunk - kos kode

1 libraries
2 palette
3 data
4 non-hyperspectral
4.1 plots
4.2 Wilcox
5 sessionInfo

```
facet_wrap(~ variable, ncol = ncol, drop = FALSE, scales="
ggtitle(paste0(omicsLevel)) +
scale_x_continuous(breaks = c(1, 7, 8, 14)) +
# theme_bw() +
theme_classic() +
theme(legend.position = "right",
      legend.title = element_text(color = "black",
                                   size = 10,
                                   face = 2),
      legend.background = element_rect(fill = "grey90", # B
                                       colour = 1)) +
scale_colour_manual(name = "Condition",
                    values = mypalette.col) +
scale_fill_manual(values = mypalette.col) +
guides(fill="none") +
labs(x = 'DAS',
      y = paste0(string, 'measurement (+/- SE)')) +
scale_y_continuous(# trans = log2_trans(),
                   labels = scales::comma#,
                   ) +
theme(axis.text.y = element_text(size = 7.5)) +
geom_vline(xintercept=7.5, linetype='twodash', col = '#EEE8CD',
```

03.3_Phenomics-plots.html	14/11/2024 23:54	Firefox HTML Doc...	1,543 KB
03.3_Phenomics-plots.Rmd	14/11/2024 23:54	RMD File	10 KB

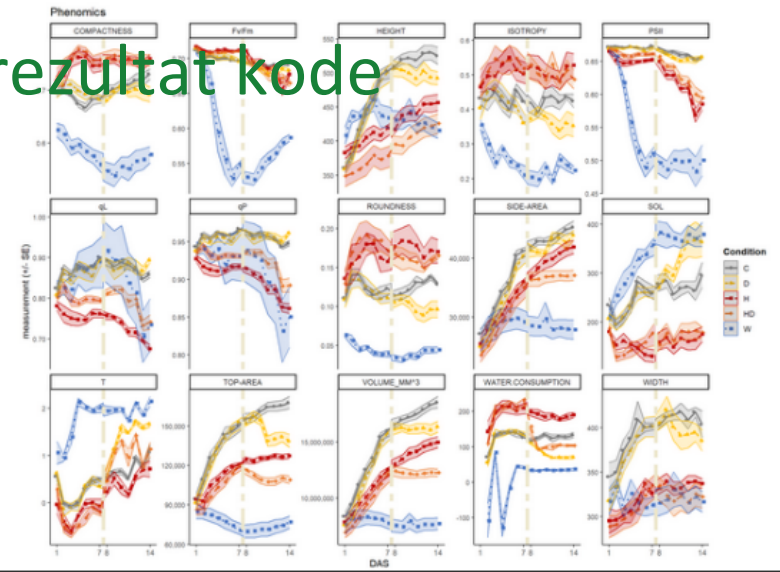
```
print(myplot)
```

chunk - kos kode

```
## R version 4.4.1 (2024-06-14 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 10 x64 (build 19045)
##
## time zone: Europe/Ljubljana
## tzcode source: internal
##
## attached base packages:
## [1] grid      stats    graphics grDevices utils    datasets methods
## [8] base
##
## other attached packages:
## [1] plyr_1.8.9      crayon_1.5.3    scales_1.3.0
## [4] RColorBrewer_1.1-3 gridExtra_2.3    multcompView_0.1-10
## [7] ggplot2_3.5.1    magrittr_2.0.3
```

```
## session Info
##
## {r, echo=TRUE, warning=FALSE, message=FALSE}
##
sessionInfo()
```

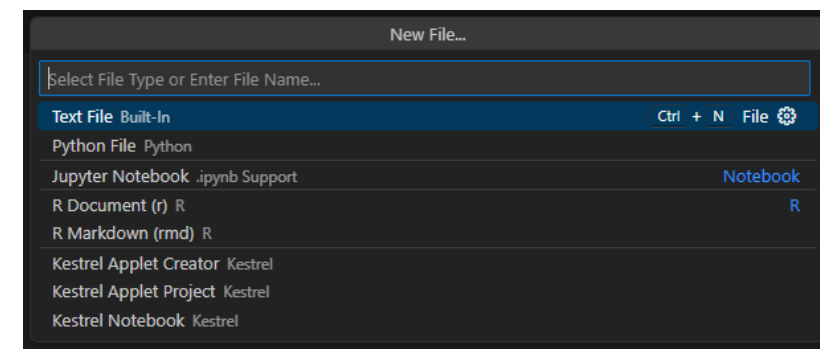
slika- rezultat kode





Visual Studio Code

- Integrirano razvojno okolje (IDE), ki ga je Microsoft razvil za operacijske sisteme Windows, Linux, macOS in spletne brskalnike
- Funkcije vključujejo podporo za razhroščevanje (*debugging*), poudarjanje sintakse (*syntax highlighting*), inteligentno dokončanje kode (*autocompletion*), vgrajen nadzor različic z Git-om, in več
- Vključuje osnovno podporo za najpogostejše programske jezike, kot so C, C#, C++, Java, JavaScript, Node.js, Python, Rust, Julia, itn
- Podporo za dodatne jezike je mogoče zagotoviti s prosto dostopnimi razširitvami na *VS Code Marketplace*





Project Jupyter

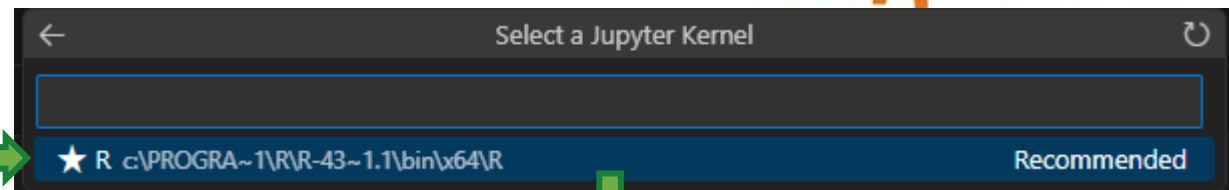
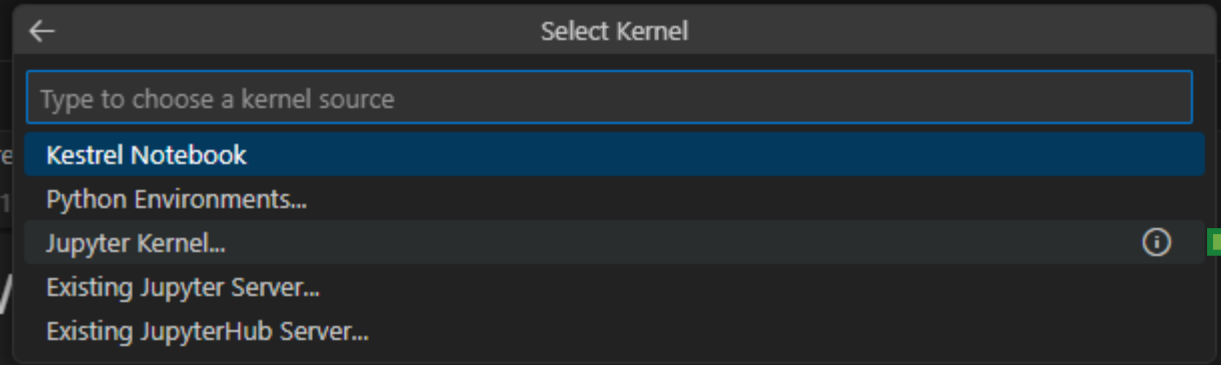
- Projekt za razvoj odprtokodne programske opreme (*open-source software*), odprtih standardov (*open standards*) in storitev za interaktivno programiranje (*interactive computing*) v več programskih jezikih
- Ime projekta se nanaša na tri osnovne programske jezike, ki jih podpira:
 - Julia,
 - Python in
 - R



Jupyter Notebook

- Lahko se nanaša na dva različna koncepta:
- Uporabniško aplikacijo za urejanje kode in besedila
- Datotečni format
 - Datoteka JSON, ki sledi shemi različice in se običajno konča s končnico *".ipynb"*
 - Vsebuje
 - Metapodatke - podatkovni slovar definicij za nastavitve in prikaz
 - Format - številka različice programske opreme
 - Seznam celic - različne vrste celic za Markdown (prikaz), kodo (za izvajanje) in

rezultat kode



DEG-GSEA.ipynb

DEG_GSEA > scripts > DEG-GSEA.ipynb > conversion tables, wide format, comparison > fn = 'PGSC-I'

+ Code + Markdown ▶ Run All ↺ Restart | Execute Group 1 || Execute Group 2 ☰ Clear All Outputs ☰ Outline ...

conversion tables, wide format, comparison

Markdown

Execute Cell (Ctrl+Alt+Enter)

```
fn = 'PGSC-ITAG_ConversionTable_v2019.xlsx'
tr1 = openxlsx::read.xlsx(file.path(fp, fn), sheet = 1, startRow = 1, colNames = TRUE)
head(tr1)
```

[2] ✓ 1.1s

A data.frame: 6 × 3

	GeneGroup	Main(s)	Alternative(s)
	<chr>	<chr>	<chr>
1	g00001	PGSC0003DMG400000001	PGSC0003DMG400000251 Sotub01g031030.1.1
2	g00002		Sotub01g031040.1.1 PGSC0003DMG400000002
3	g00003		PGSC0003DMG400000003 Sotub01g031070.1.1
4	g00004		Sotub01g031080.1.1 PGSC0003DMG400000004
5	g00005		Sotub01g031100.1.1 PGSC0003DMG400000005
6	g00006		Sotub01g031170.1.1 PGSC0003DMG400000008

interaktivno programiranje

koda

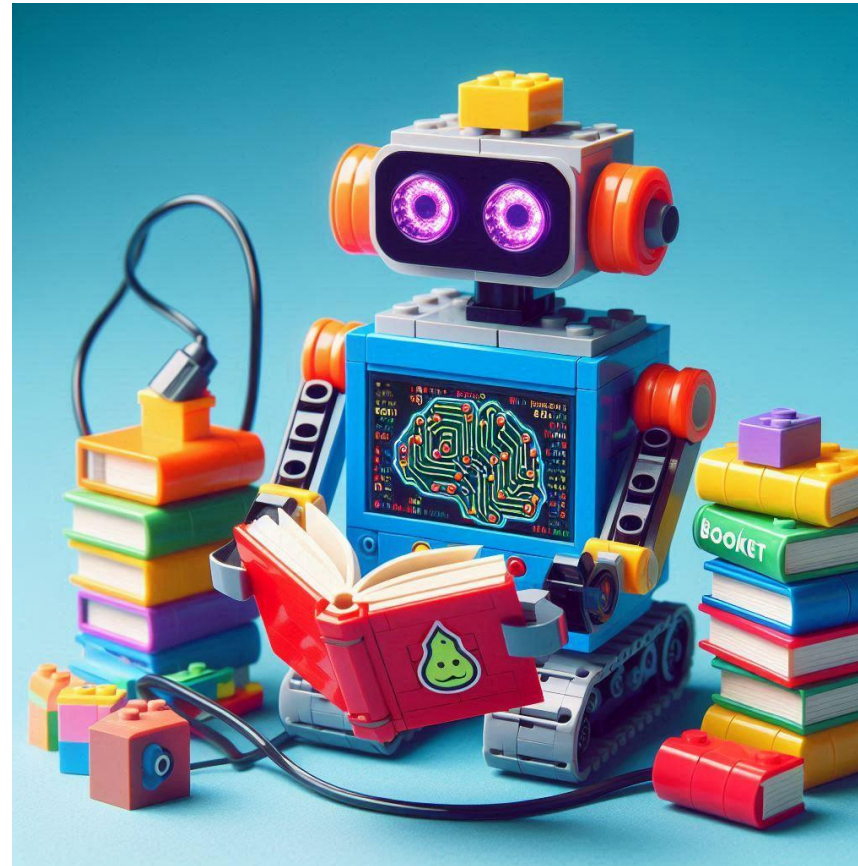
rezultat

> CONTEXTUAL HELP

Launchpad 0 1 1K 0 0 Add Neo4j Connection Sourcery R: (not attached) Spaces: 4 Cell 5 of 78 3 Spell Kernels initialized!

Primer 1

Strojno učenje (*machine learning*)
Izbor spremenljivk (*feature/variable selection*)



Strojno učenje (*machine learning*)

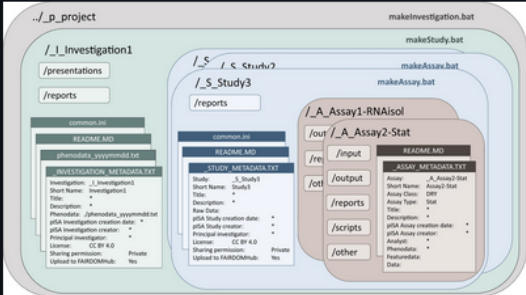
Izbor spremenljivk (*feature/variable selection*)

README MIT license

'multiOmics data analysis, integration, and visualisation protocol

- Data integration and modelling with [R](#)
- Systems biology analysis and visualization pipeline in [R](#) and [Python](#)

Data management framework



The diagram illustrates the pISA-tree directory structure. It shows a root directory `./_p_project` containing `./_I_Investigation1`, `./_S_Study2`, and `./_A_Assay1-RNAseq`. Each level has subfolders for `/presentations`, `/reports`, `/input`, `/output`, `/scripts`, and `/other`. The `./_I_Investigation1` folder contains `./_I_Investigation1` and `./_I_Investigation1`. The `./_S_Study2` folder contains `./_S_Study2` and `./_S_Study2`. The `./_A_Assay1-RNAseq` folder contains `./_A_Assay1-RNAseq` and `./_A_Assay1-RNAseq`.

[pISA-tree on GitHub](#)

Petek, M., Zagorščak, M., Blejec, A. et al. pISA-tree - a data management framework for life science research projects using a standardised directory tree. *Sci Data* 9, 685 (2022). <https://doi.org/10.1038/s41597-022-01805-5>

multiOmics-integration / _p_ADAPTOmics / ↑ Top

..		
_I_Desiree	Update README.md	2 days ago
presentations	fig	3 days ago
reports	fig	3 days ago
README.MD	Update README.MD	2 days ago
_PROJECT_METADATA.TXT	rename _p_	3 days ago
common.ini	rename _p_	3 days ago

README.MD

Project _p_ADAPTOmics

ADAPT – Accelerated Development of multiple-stress tolerant Potato



This [pISA-tree](#) structured repository contains developed pipeline for the integration of multi-omics data and deep phenotyping of potato under single- and combined abiotic stress responses within the WP5 of [ADAPT](#) project



For more information see instructions at each pISA-tree level: [Investigation](#), [Study](#), and [Assay](#)

Strojno učenje (*machine learning*)

Izbor spremenljivk (*feature/variable selection*)

multiOmics-integration / _p_ADAPTOmics / _I_Desiree / _S_multiOmics / _A_multiOmicsStat-R /  Add file 

 zagorGit Update README.MD 17c6654 · 2 days ago 

Name	Last commit message	Last commit date
..		
input	rename _I_	3 days ago
other	rename _I_	3 days ago
output	rename _I_	3 days ago
reports	rename _I_	3 days ago
scripts	rename _I_	3 days ago
README.MD	Update README.MD	2 days ago
_ASSAY_METADATA.TXT	rename _I_	3 days ago

README.MD 

Assay _A_multiOmicsStat-R

Proteomics

- [DEP](#)
- [GSEA](#)

Denote: GSEA (2.1) produces multiple files per comparison. Combine information (2.2) and visualise (2.3) in R

Phenomics

- Feature selection ([FS](#)) with Random Forest (RF) and Recursive Feature Elimination (RFE) (noninvasive) (3.1). For other methods see Assay [_A_multiOmics-FS-Py](#)

https://htmlpreview.github.io/?https://github.com/NIB-SI/multiOmics-integration/blob/main/_p_ADAPTOmics/_I_Desiree/_S_multiOmics/_A_multiOmicsStat-R/scripts/03.1_FS-Phenomics-noninvasive.html

multiOmics-integration / _p_ADAPTOmics / _I_Desiree / _S_multiOmics / _A_multiOmics-FS-Py /  Add file 

 zagorGit Update README.md 838defe · 2 days ago 

Name	Last commit message	Last commit date
..		
input	rename _I_	3 days ago
other	rename _I_	3 days ago
output	rename _I_	3 days ago
reports	rename _I_	3 days ago
scripts	rename _I_	3 days ago
README.md	Update README.md	2 days ago
_ASSAY_METADATA.TXT	rename _I_	3 days ago

README.md  

Assay _A_multiOmics-FS-Py

Feature Selection in Python with Scikit-Learn

- Variable selection was conducted on the non-invasive phenomics variable sets
- The random forest (RF) algorithm from the R package caret v6.0-94 ([_A_multiOmicsStat-R](#)) as well as the python package scikit-learn v1.2.0 were used with default settings, as RF showed the best performance out of a selection of algorithms
- Recursive feature elimination was applied in R and multiple importance scores, including mutual information (MI), Anova, RF importance and SHAP values (REF) were computed in Python, showing consistencies between the approaches for the top 5 variables
- The sixth variable, was selected based on expert knowledge

https://github.com/NIB-SI/multiOmics-integration/blob/main/_p_ADAPTOmics/_I_Desiree/_S_multiOmics/_A_multiOmics-FS-Py/scripts/Script_feature_selection.ipynb

Primer 2

Spletno strganje (*Web Scraping*)



Spletno strganje (Web Scraping)

https://projekt-spoznav.si/videosposnetki-in-predstavitve/fair/

Splošna usposabljanja o praksah odprte znanosti in deljenju raziskav po načelih FAIR

Najdeni so videoposnetki 23



Odprta znanost v
Evropskem
raziskovalnem
prostoru (23. 6. 2023)



Odprta znanost v
Evropskem
raziskovalnem
prostoru, ponovitev
(7. 9. 2023)



Nacionalna
zakonodaja na
področju odprte
znanosti (14. 9.
2023)



Slovenska skupnost
odprte znanosti (21.
9. 2023)



Možnosti in
ugodnosti pri
odprtem objavljanju
za slovensko
akademsko skupnost
(28. 9. 2023)



Platforma za odprte
objave Open
Research Europe (12.
10. 2023)



Nezaupanja vredne
založniške prakse
(19. 10. 2023)



Temeljna načela
delovanja kOAlcije S
(26. 10. 2023)

view-source:https://projekt-spoznav.si/videosposnetki-in-predstavitve/fair/

```
445 <path d="M16 0c-8.837 0-16 7.163-16 16s7.163 16 16 16 16-7.163 16-16 16-16z" data-bbox="460 279 997 295"/>  
446 </svg>  
447 </a><div class="aiovg-caption"><div class="aiovg-title">  
448 <a href="https://projekt-spoznav.si/video/odprta-znanost-v-evropskem-raziskovalnem-prostoru-1/" class="aiovg-link-title">  
449 Odprta znanost v Evropskem raziskovalnem prostoru (23. 6. 2023) </a></div></div></div><div class="aiovg-item-vi  
450 <a href="https://projekt-spoznav.si/video/odprta-znanost-v-evropskem-raziskovalnem-prostoru/" class="aiovg-responsive-container" sty  
451   
454 </svg>  
455 </a><div class="aiovg-caption"><div class="aiovg-title">  
456 <a href="https://projekt-spoznav.si/video/odprta-znanost-v-evropskem-raziskovalnem-prostoru/" class="aiovg-link-title">  
457 Odprta znanost v Evropskem raziskovalnem prostoru, ponovitev (7. 9. 2023) </a></div></div></div><div class="aiovg-item-vi  
458 <a href="https://projekt-spoznav.si/video/nacionalna-zakonodaja-na-podrocju-odprte-znanosti/" class="aiovg-responsive-container" sty  
459   
462 </svg>  
463 </a><div class="aiovg-caption"><div class="aiovg-title">  
464 <a href="https://projekt-spoznav.si/video/nacionalna-zakonodaja-na-podrocju-odprte-znanosti/" class="aiovg-link-title">  
465 Nacionalna zakonodaja na področju odprte znanosti (14. 9. 2023) </a></div></div></div><div class="aiovg-item-vi  
466 <a href="https://projekt-spoznav.si/video/slovenska-skupnost-odprte-znanosti/" class="aiovg-responsive-container" style="padding-bot  
467   
470 </svg>  
471 </a><div class="aiovg-caption"><div class="aiovg-title">  
472 <a href="https://projekt-spoznav.si/video/slovenska-skupnost-odprte-znanosti/" class="aiovg-link-title">  
473 Slovenska skupnost odprte znanosti (21. 9. 2023) </a></div></div></div><div class="aiovg-item-video aiovg-item-  
474 <a href="https://projekt-spoznav.si/video/moznosti-in-ugodnosti-pri-odprtem-objavljanju-za-slovensko-akademsko-skupnost/" class="aiovg-link-title">  
475   
478 </svg>  
479 </a><div class="aiovg-caption"><div class="aiovg-title">  
480 <a href="https://projekt-spoznav.si/video/moznosti-in-ugodnosti-pri-odprtem-objavljanju-za-slovensko-akademsko-skupnost/" class="aiovg-link-title">  
481 Možnosti in ugodnosti pri odprtem objavljanju za slovensko akademsko skupnost (28. 9. 2023) </a></div></div></div><div class="aiovg-item-vi  
482 <a href="https://projekt-spoznav.si/video/platforma-za-odprte-objave-open-research-europe/" class="aiovg-responsive-container" style="padding-bot  
483   
486 </svg>  
487 </a><div class="aiovg-caption"><div class="aiovg-title">  
488 <a href="https://projekt-spoznav.si/video/platforma-za-odprte-objave-open-research-europe/" class="aiovg-link-title">  
489 Platforma za odprte objave Open Research Europe (12. 10. 2023) </a></div></div></div><div class="aiovg-item-vi  
490 <a href="https://projekt-spoznav.si/video/nezaupanja-vredne-zalozniske-prakse/" class="aiovg-link-title">  
491 Nezaupanja vredne založniške prakse (19. 10. 2023) </a></div></div></div><div class="aiovg-item-vi  
492 <a href="https://projekt-spoznav.si/video/temeljna-nacela-delovanja-koalcije-s/" class="aiovg-link-title">  
493 Temeljna načela delovanja kOAlcije S (26. 10. 2023) </a></div></div></div></div>
```

Spletno strganje (*Web Scraping*)

find the total number of pages

Denote: page 1 has no number

```
In [9]: find_total_pages = function(base_url) {
    current_page = 1
    has_next = TRUE
    last_page = 1

    while (has_next) {
      # Construct URL for the current page
      page_url = ifelse(current_page == 1, base_url, paste0(base_url, "page/", current_page, "/"))

      # Read the webpage
      webpage = tryCatch(rvest::read_html(page_url), error = function(e) return(NULL))

      if (is.null(webpage)) {
        break
      }

      # Check for the "Next" button
      next_button = webpage %>%
        rvest::html_nodes("a.next.page-numbers") %>%
        rvest::html_attr("href")

      if (length(next_button) == 0) {
        # No "Next" button found
        has_next = FALSE
      } else {
        # Increment the counter
        current_page = current_page + 1
        last_page = current_page
      }
    }

    return(last_page)
}
```

scrape: titles and dates

User-Agent header that mimics a typical web browser (like Chrome) to avoid being blocked by the server -> able to bypass the 403 error and fetch the pages by ChatGPT

```
In [3]: # Function to scrape lecture titles and dates from a single page
def scrape_page(url):
    # Define headers to mimic a browser request
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 S
    }

    # Send request and parse the HTML content with headers
    response = requests.get(url, headers=headers)

    # Check the response status
    if response.status_code != 200:
        print(f"Failed to fetch {url}, status code: {response.status_code}")
        return pd.DataFrame() # Return empty DataFrame if page can't be loaded

    print(f"Successfully fetched {url}")
    webpage = BeautifulSoup(response.text, 'html.parser')

    # Extract lecture titles
    lecture_titles = [title.get_text(strip=True) for title in webpage.find_all('a', class_='aiovg-link-title')]

    # Extract dates and clean them
    dates = [re.search(r'\\((\\d{1,2})\\. \\d{1,2})\\. \\d{4}\\)', title).group(1) if re.search(r'\\((\\d{1,2})\\. \\d{1,2})\\. \\d{4}\\)', ti

    # Clean the titles (remove dates)
    cleaned_titles = [re.sub(r'\\s*(\\d{1,2})\\. \\d{1,2})\\. \\d{4}\\)', '', title).strip() for title in lecture_titles]

    # Convert to DataFrame
    data = {
        'Date': pd.to_datetime(dates, format='%d. %m. %Y'),
        'Lecture_Title': cleaned_titles
    }

    return pd.DataFrame(data)
```



Nacionalni inštitut za biologijo (NIB)
Oddelek za biotehnologijo in sistemsko biologijo
Večna pot 121, 1000 Ljubljana
Slovenija

<https://github.com/NIB-SI>

dr. Maja Zagorščak
dr. Carissa Bleker
dr. Jan Zrimec
dr. Marko Petek
dr. Anže Županič

