

# Rozwój oprogramowania w R: wprowadzenie

Mateusz Staniak

*Rozwój oprogramowania w R*

Instytut Matematyczny UWr, semestr letni 2020



Uniwersytet  
Wrocławski

# Informacje nt. kursu



Uniwersytet  
Wrocławski

# Program kursu



Uniwersytet  
Wrocławski

1. Miejsce R w świecie produktów opartych na danych i rozwoju oprogramowania analitycznego. Narzędzia przydatne w pracy z R. Powtórzenie wiadomości z podstaw R i przetwarzania danych.
2. R i różne źródła danych: bazy danych, API i inne. Obsługa błędów i control flow w R. Przetwarzanie danych w R.
3. Czysty kod. Dobre praktyki pisania kodu w R. Styl pisania kodu w R.
4. Idee zaawansowanego R: kod obiektowy i funkcjonalny, przestrzenie nazw i środowiska, struktury danych, szybkość kodu, standardowa i niestandardowa ewaluacja.
5. Pakiety w R: cele, struktura, praktyki, dobre przykłady i powiązane zagadnienia.
6. Szczegółowe informacje nt. rozwoju pakietów w R. Pomocne narzędzia: pakiety devtools, usethis, travis, covr i gh.
7. Testy jednostkowe, ciągła integracja i git. Pakiet testthat.
8. Dokumentacja pakietu. Pakiety roxygen2 i pkgdown.
9. Rmarkdown, interaktywne dokumenty, raportowanie.
10. Programowanie obiektowe w R: klasy S3, S4, R6. Projektowanie kodu.
11. Profilowanie kodu. Programowanie równoległe w R. Połączenie z C++ (pakiet RCpp).
12. Budowa interaktywnych aplikacji z pakietem shiny.

<https://usosweb.uni.wroc.pl/kontroler.php?action=katalog2/przedmioty/pokazPrzedmiot&kod=z8-MT-S-q12>

# Ogólne informacje



Forma: laboratorium.

Liczba godzin: 30.

Liczba ECTS: TBA.

Konsultacje: poniedziałek 11:00 – 12:00 (proszę o umawianie się wcześniej).

Zasady zaliczenia: ocena składają się trzy elementy:

- projekt grupowy,
- dwa zadania domowe,
- egzamin ustny (prezentacja projektu + rozmowa).

Do zaliczenia konieczne jest uzyskanie 50% możliwych punktów z każdego z tych elementów. Przygotowanie obu zadań domowych jest warunkiem koniecznym zaliczenia. Progi punktowe na poszczególne oceny zostaną ustalone później.

Dodatkowe pół oceny można uzyskać za wniesienie wkładu do [https://github.com/mstaniak/RDev\\_notes](https://github.com/mstaniak/RDev_notes).

# Produkty oparte na danych i R

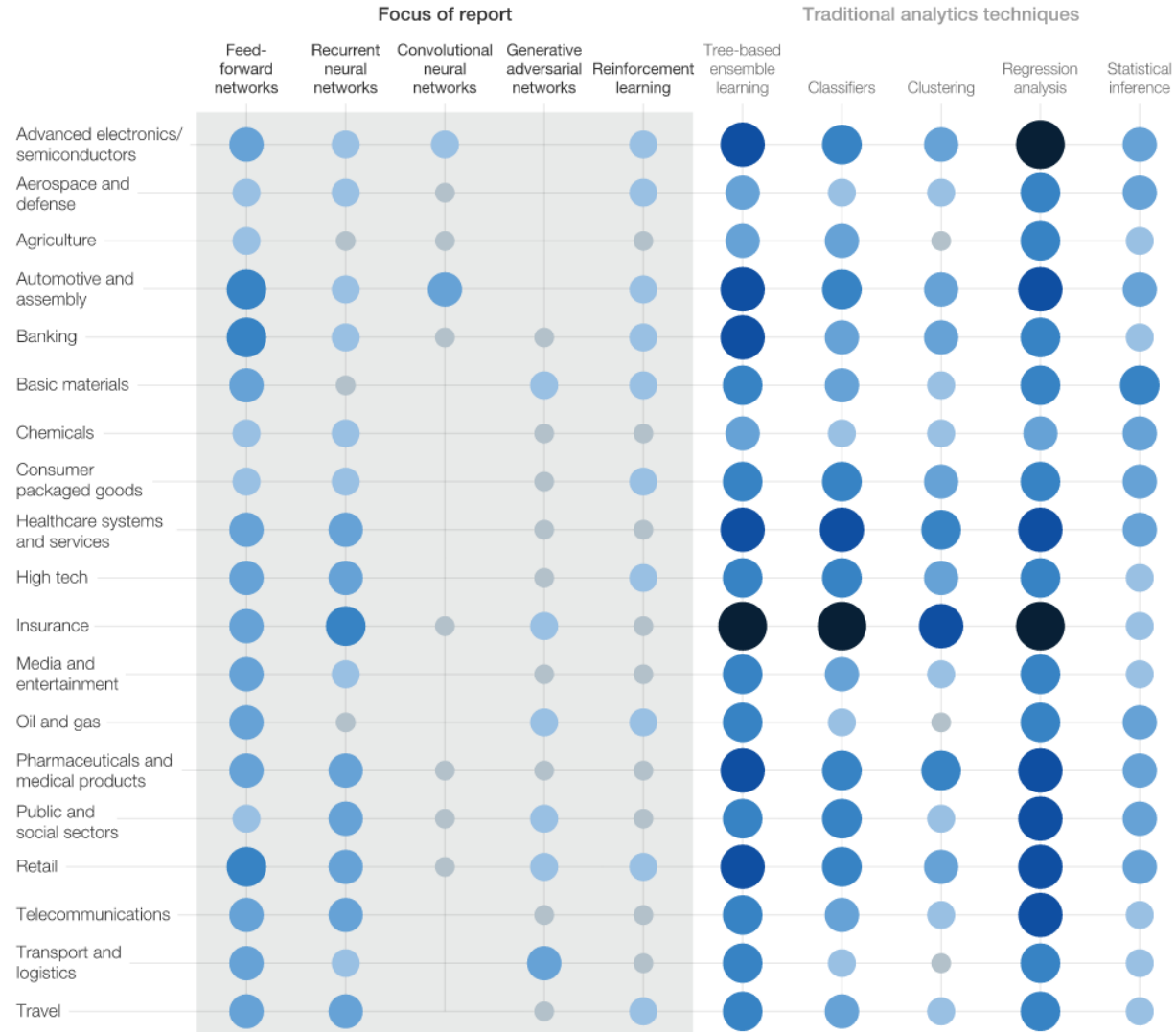


Uniwersytet  
Wrocławski

<https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning>



Technique relevance<sup>1</sup> heatmap by industry



<https://www.pcmag.com/news/90-percent-of-the-big-data-we-generate-is-an-unstructured-mess>



Uniwersytet  
Wrocławski

## 90 Percent of the Big Data We Generate Is an Unstructured Mess

A deep dive into the world of "big data" sets that comprise the vast majority of what the world population creates every single day shows a mishmash that regular ol' databases can barely handle, let alone analyze.



By [Eric Griffith](#) November 15, 2018



EVERY DAY WE CREATE

2,500,000,  
000,000,  
000,000

(2.5 QUINTILLION) BYTES OF DATA

*This would fill 10 million blu-ray discs,  
the height of which stacked, would measure  
the height of 4 Eiffel Towers on top of one another.*



<https://www.pcmag.com/news/the-big-data-market-is-set-to-skyrocket-by-2022>

Global big data and business analytics revenue, 2015-2022





<https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/>



- 75% of Netflix users select films recommended to them by the company's machine learning algorithms.
- The global machine learning market was valued at \$1.58B in 2017 and is expected to reach \$20.83B in 2024, growing at a CAGR of 44.06% between 2017 and 2024.
- Projected to grow at a Compound Annual Growth Rate (CAGR) of 42.8% from 2018 to 2024, the global Machine Learning (ML) market will worth \$30.6B in four years.
- Tractica predicts annual global AI software revenue will grow from \$10.1B in 2018 to \$126.0B by 2025, achieving a CAGR of 43.41%.



Uniwersytet  
Wrocławski



**Big Data Borat**  
@BigDataBorat



In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

3:47 AM · 27 lut 2013 · [Twitter Web Client](#)

**536** Podane dalej   **361** Polubień





Uniwersytet  
Wrocławski

- Ilość danych dostępnych na rynku rośnie,  
(to samo dotyczy „rynku” danych naukowych)
- rośnie też liczba sposobów ich zastosowania,  
(patrz np. <https://sifted.eu/poland-startups-top-rankings/>)
- dane i ich wykorzystanie wymagają zbierania, czyszczenia, i przetwarzania danych, a w dalszej kolejności - wizualizacji, raportowania i modelowania,
- każdy z tych etapów pracy z danymi wymaga dobrych narzędzi programistycznych.

# Produkty oparte na danych



Uniwersytet  
Wrocławski

## TYPES OF DATA PRODUCTS

AUTOMATED DECISION-MAKING

DECISION SUPPORT

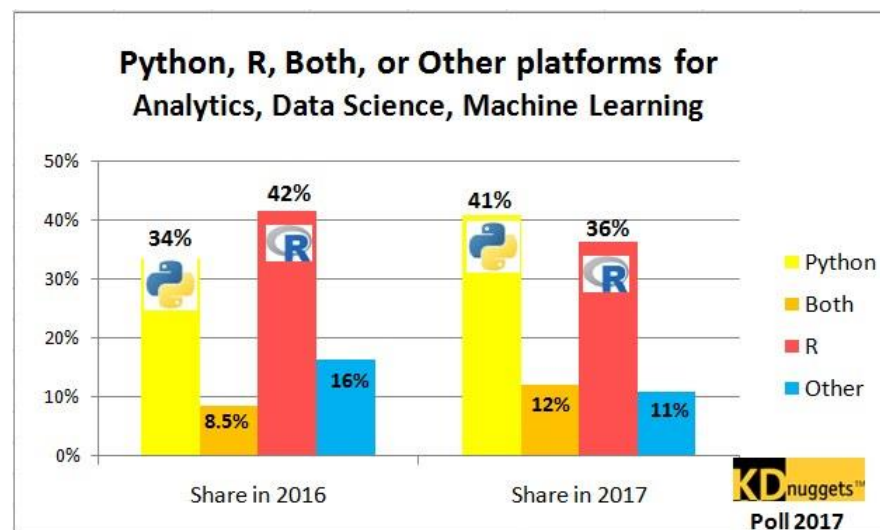
ALGORITHMS

DERIVED DATA

RAW DATA

<https://towardsdatascience.com/designing-data-products-b6b93edf3d23>

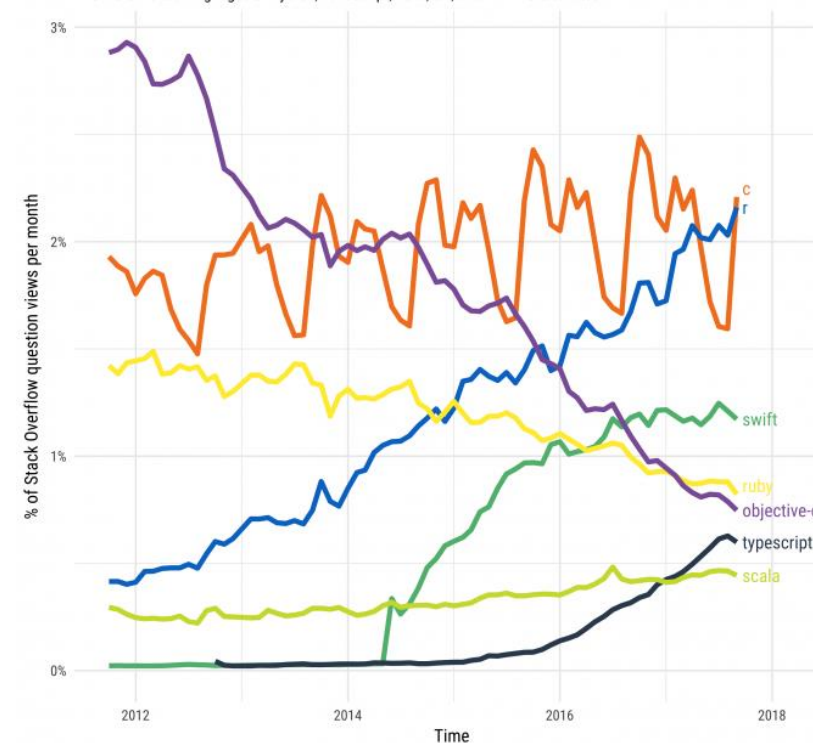
# „Wojna” języków



<https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html>

## Stack Overflow Traffic to Programming Languages

Based on visits to Stack Overflow questions from World Bank high-income countries. The more-visited languages of Python, JavaScript, Java, C#, and PHP were omitted.



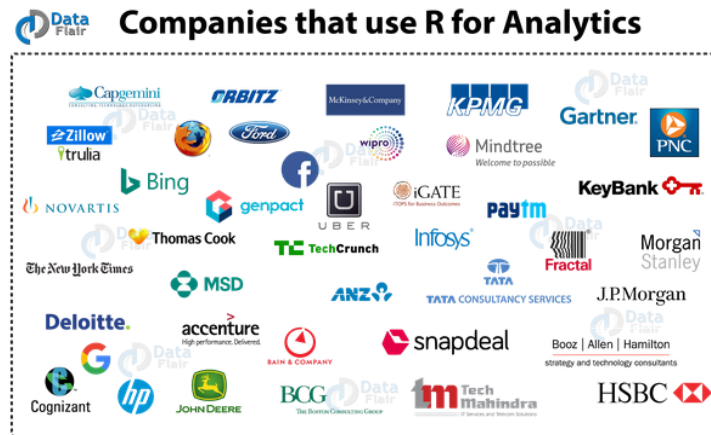
<https://stackoverflow.blog/2017/10/10/impressive-growth-r/>

# R w przemyśle



Uniwersytet  
Wrocławski

## Firmy używające R



<https://www.quora.com/Which-companies-use-R>

## Firmy używające RStudio



<https://rstudio.com/about/customer-stories/>

# Zalety R



- Unikalne narzędzia do wizualizacji danych (ggplot2, shiny) + porty do innych narzędzi (m.in. D3, plotly),
- wydajne przetwarzanie danych dzięki łatwemu połączeniu z C++ / C i narzędziom do programowania równoległego oraz pakietowi data.table,
- unikalne narzędzia do raportowania i budowania stron (R Markdown, blogdown, bookdown, etc),
- dostępność ogromnej ilości metod statystycznych, optymalizacyjnych, algorytmów ML itd. dzięki wiarygodnym pakietom,
- dostępność interfejsów do popularnych baz danych,
- dobra dokumentacja kluczowych pakietów,
- wygodne narzędzia do przetwarzania danych (np. tidyverse), optymalizacja pod wygodę analizy danych,
- duża społeczność,
- dostępność doskonałego IDE – RStudio.

# Narzędzia przydatne w pracy z R



Uniwersytet  
Wrocławski



# Klasyczne IDE



Uniwersytet  
Wrocławski

~/Projekty/Shared\_peptides/src - RStudio

```
File Edit Code View Plots Session Build Debug Profile Tools Help
percolator_known_analysis.R spectra_co_occurrence.R Untitled2* x
Source on Save Run Source
240 select(-precursor_scan_name) %>%
241 unique() %>%
242 arrange(all_peptides, rep) %>%
243 mutate(peptides_number = stringr::str_count(all_peptides, "_" + 1)
244 write_rds(common_spectra, "common_spectra_exact.RDS")
245
246 common_spectra %>%
247 filter(peptides_number > 1)
248
249 ms_small_grouped %>%
250 filter(grepl("NCRDCKCTLACQQFR_VQLFEDPTVDKEVEIR", all_peptides))
251
252 ms_small_grouped %>%
253 group_by(all_peptides) %>%
254 summarise(cnt = n_distinct(rep)) %>%
255 arrange(-cnt) %>%
256 filter(grepl("_", all_peptides)) %>%
257 filter(cnt > 1)
258
259 ms_small_grouped %>%
260 mutate(n_peptides = stringr::str_count(all_peptides, "_" + 1) %>%
261 arrange(-n_peptides) %>%
262 group_by(all_peptides) %>%
263 mutate(n_reps = n_distinct(rep)) %>%
264 arrange(-n_reps) %>%
265 filter(n_peptides > 1, n_reps > 1)
266
267
266:1 (Top Level) R Script
```

Environment History Connections

Global Environment

- common\_spectra 4437 obs. of 4 variables
- fastas List of 4
- fastas\_to\_filter List of 3
- just\_rep2\_for\_rep1 Large list (36 elements, 12.9 Mb)
- ms\_clean 77067 obs. of 7 variables
- ms\_small 26625 obs. of 3 variables
- ms\_small\_1 8701 obs. of 3 variables
- ms\_small\_grouped 15516 obs. of 3 variables
- percolator\_search 77067 obs. of 14 variables
- percolator\_search\_ Large list (9 elements, 8.4 Mb)
- percolator\_search\_ Large list (9 elements, 9.3 Mb)
- precursor\_spectra\_ Large list (3 elements, 3.9 Mb)

Files Plots Packages Help Viewer

Home > Projekty > Shared\_peptides > src

Name	Size	Modified
all_gls_fits_summary.RDS	209.6 KB	Feb 2, 2020, 10:25 PM
all_gls_fits_summary.RDS	311.7 KB	Feb 3, 2020, 1:17 AM
clusters.RDS	551.4 KB	Feb 11, 2020, 10:45 PM
clusters_within_spectrum.RDS	628.9 KB	Feb 11, 2020, 11:28 PM
percolator_known_analysis.R	5.7 KB	Feb 11, 2020, 10:45 PM
repeated.Rmd	24.1 KB	Feb 12, 2020, 11:00 AM
scampi.R	2.5 KB	Feb 19, 2020, 5:26 PM
split_by_precursor.R	3.1 KB	Feb 12, 2020, 10:25 AM
src.Rproj	258 B	Feb 20, 2020, 6:45 PM
summary.Rmd	26.3 KB	Feb 3, 2020, 2:59 PM
whitening_experiments.R	1.5 KB	Feb 19, 2020, 5:26 PM
repeating_spectra_1.RDS	1 GB	Feb 20, 2020, 10:07 PM
repeating_spectra_2.RDS	1 GB	Feb 20, 2020, 10:07 PM
repeating_spectra_3.RDS	1 GB	Feb 20, 2020, 10:07 PM
spectra_co_occurrence.R	9.2 KB	Feb 21, 2020, 12:39 AM
common_spectra_exact.RDS	199.8 KB	Feb 20, 2020, 10:35 PM

Console Terminal Jobs

```
~/Projekty/Shared_peptides/src/
<chr> <int> <chr> <dbl> <int>
1 1 4006 FIRPDR_RHPPDDLSQDSPEQASKSPR_TPRDQGGPTLAQPAHVR 3 2
2 2 4187 FIRPDR_RHPPDDLSQDSPEQASKSPR_TPRDQGGPTLAQPAHVR 3 2
3 1 3598 DSFGSGDRK_YPAEVTITHRKNGK 2 2
4 1 3604 LAFRPCNANPHK_SRFCHPIYFPRR 2 2
5 1 3676 ACSQRSOLYRHPR_SLDEQANQENDALHKK 2 2
6 1 3736 ASEEHTNAACFACILLSHGEENVYKG_SKSPROPDAWIDSPSR 2 2
7 1 4246 NQGENLCQCSIRK_RCDPGWGLHCR 2 2
8 1 4534 LNVEAVNTHR_LNYRVPSR 2 2
9 1 4689 HPPDDLSQDSPEQASKSPR_NIHYNVSVNPNK 2 2
10 1 5793 QTSEKKP_TEGKITLQDLKR 2 2
# ... with 71 more rows
> |
```

<https://rstudio.com>

# Alternatywne IDE / edytory



Uniwersytet  
Wrocławski

- IntelliJ + plugin <https://plugins.jetbrains.com/plugin/6632-r-language-for-intellij>,
- Neovim + plugin <https://github.com/jalvesaq/Nvim-R> (polecam z perspektywy edycji tekstu)
- Visual Studio + plugin <https://docs.microsoft.com/en-us/visualstudio/rvs/?view=vs-2017>
- <https://www.getarchitect.io/>
- <https://rkward.kde.org/>
- Emacs + plugin <https://ess.r-project.org/>
- SublimeText + plugin <https://github.com/REditorSupport/sublime-ide-r>

# Klient Gita



Uniwersytet  
Wrocławski

- Wbudowany klient w RStudio,
- GitKraken (używany przeze mnie – wersja Pro darmowa do użytku niekomercyjnego z Github Student Pack),
- SourceTree,
- Github Desktop,
- SmartGit.

# Podstawy R: przypomnienie



Uniwersytet  
Wrocławski

# Skąd się bierze R?



Uniwersytet  
Wrocławski

R to język, istnieje kilka implementacji:

- <https://www.r-project.org/> (najpopularniejsza)
- <https://mran.microsoft.com/open> (dawniej Revolution R)
- <https://github.com/oracle/fastr>
- <http://www.pqr-project.org/> (pretty quick R)
- <http://janvitek.org/pubs/vee14.pdf>

# Skąd się biorą pakiety?



Uniwersytet  
Wrocławski

Siłą R są pakiety. Można je znaleźć w następujących repozytoriach:

- CRAN: <https://cran.r-project.org/>
- Biocondutor: <https://www.bioconductor.org/>
- MRAN: <https://mran.microsoft.com/>
- GitHub, Gitlab i inne narzędzia oparte na kontroli wersji
- można stworzyć własne repozytorium:  
<https://cran.r-project.org/package=miniCRAN>

# Powtórzenie: quiz



Uniwersytet  
Wrocławski

1. Jakie są podstawowe typy wektorów?
2. Czym się różni lista od innych wektorów?
3. Czym się różni data.frame od matrix?
4. Kiedy R kopiuje obiekty?
5. Jak wykonać podstawowe operacje: wybór kolumn, wybór wierszy, tworzenie kolumn w bazowym R / tidyverse?
6. Jakie są rodzaje joinów i jak je wykonać w bazowym R / tidyverse?
7. Bonus: czym się różnią dane w postaci wąskiej od postaci szerokiej?
8. Bonus: czym są "tidy data"?

# Lektury



Uniwersytet  
Wrocławski

- <https://www.slideshare.net/WitJakuczun/r-software-development-how-to-write-and-maintain-30k-loc-in-r-and-survive>
- <https://www.slideshare.net/WitJakuczun/always-be-deploying-how-to-make-r-great-for-machine-learning-in-not-only-enterprise>
- [https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019\\_Software\\_Engineering\\_for\\_Machine\\_Learning.pdf](https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf)
- <https://rstudio.com/wp-content/uploads/2016/05/base-r.pdf>
- <https://adv-r.hadley.nz/> rozdziały 1 – 4
- <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf> strony 1 – 29