Statystyka I z RWarsztat 1.Zapoznanie z R

Tomasz Żółtak

3 października 2016

# O prowadzącym

## Tomasz Żółtak - kilka faktów

* Z wykształcenia jestem socjologiem (IS UW).
* W praktyce jestem przede wszystkim statystykiem i psychometrykiem.
* Pracuję w Instytcie Badań Edukacyjnych (znów, po rocznej przerwie), wcześniej byłem też pracownikiem Instytutu Filozofii i Socjologii PAN.
* Jestem aktywny naukowo w dziedzinach:
  + badań edukacyjnych,
  + socjologii polityki
  + (oraz incydentalnie w różnych innych).
* Współpracowałem jako ekspert i autor raportów z instytucjami takimi jak: CKE, NCK, WUM, MSiT, Kancelaria Prezydenta RP, Rada Koordynacyjna ds. Certyfikacji Biegłości Językowej UW.

## Tomasz Żółtak - związki z R

* Pracuję w R od 2008 r. i jest on moim głównym narzędziem pracy.
* Jestem współautorem 8 pakietów (ale tylko jeden z nich wypchnięty na CRAN).
  + W większości bardzo specyficznego zastososowania - pakiety automatyzujące [proces obliczania wskaźników EWD](http://ewd.edu.pl/wp-content/uploads/2014/12/poster-id-201513-00544.pdf).
  + Napisany i udokumentowany po polsku [pakiet](https://github.com/tzoltak/KTT) do analizy własności psychometrycznych testu przy pomocy narzędzi z Klasycznej Teorii Testu.
  + Gry edukacyjnej dotyczącej regresji - funkcja *regression()* w ramach pakietu [BetaBit](https://cran.r-project.org/web/packages/BetaBit).

# Kwestie organizacyjne

## Tryb pracy na zajęciach

* Lektury do przeczytania przed zajęciami (nie więcej niż 30 minut tygodniowo).
* Zajęcia - krótki wstęp teoretyczny (nie więcej niż 30 minut) i praca z komputerami.
* Prace domowe (około 1,5 godziny tygodniowo)
* Prace domowe do wyboru (dłuższe, przynajmniej 3 w semestrze).
  + Będą polegały na napisaniu bardziej skomplikowanego skryptu albo przygotowania raportu z symulacji albo na przygotowaniu serii grafik, etc.
* Prace domowe zupełnie dodatkowe.
  + Rozszerzenie materiału dla chętnych, zakres według potrzeb, możliwości:
    - wizualizacje (ggplot2, lattice, plotly, etc.)
  + pisanie całych pipeline'ów (pobieranie danych, obróbka wstępna, analiza, wizualizacje)
  + symulacje

## Warunki zaliczenia

* Aktywna praca na zajęciach.
* Robienie zadań domowych obowiązkowych + 3 do wyboru dłuższe w semestrze.
* Staramy się nie spóźniać, ale jak ktoś odeśle prace domowe spóźniony to lepiej, niż gdyby miał nigdy nie odesłać.

## Forma przesyłania prac

* Być może Moodle, ale prawdopodobnie mailem (do ustalenia).
* Wszystkie materiały będą dostępne na moim [koncie na GitHubie](https://github.com/tzoltak/3501-KOG-S1R_2016_gr3).

# Co to jest R?

## Dwa słowa historii I

* R jest otwartą implementacją języka S.
  + Język programowania S rozwijany od połowy lat 70. XX w. w firmie *Bell Laboratories* (tej od m. in. języka C).
  + W pierwszej połowie lat 80.  udostępniony na zewnątrz*Bell Laboratories*.
  + Gruntownie zmieniony i rozszerzony pod koniec lat 80. XX w.
  + Dalszy rozwój w ramach komercyjnej implementacji *S-PLUS* firmy *TIBCO Software* (obecnie jako część usługi *TIBCO Spotfire*).
* Dla chcących wiedzieć więcej:
  + [wywiad z Johnem Chambersem](http://blog.revolutionanalytics.com/2014/01/john-chambers-recounts-the-history-of-s-and-r.html),
  + [wywiad z Rickiem Beckerem](http://blog.revolutionanalytics.com/2016/07/rick-becker-s-talk.html).

## Dwa słowa o historii II

* Język R pierwotnie zaimplementowany przez Rossa Ihakę i Roberta Gentlemana z Uniwersytetu w Auckland.
  + Prace nad wersją *produkcyjną* w latach 1992-2000.
* W późniejszym okresie rozwijany przez *R Development Core Team*, a obecnie przez *R Foundation* (obie - czysto formalnie - mające *siedzibę* w Wiedniu).
* *Nowa era*:
  + 2010 r. - R w wersji 2.11 współpracuje z systemami 64-bitowymi (można korzystać z więcej niż 3 GB pamięci RAM!).
  + 2011 r. - zostaje opublikowana pierwsza wersja *RStudio* (IDE z prawdziwego zdarzenia!).
  + Od 2005 rozwija się [*hadleywerse*](http://blog.revolutionanalytics.com/2015/03/hadleyverse.html).

## R dziś

* Olbrzmi zakres dostępnych metod analizy.
  + Obecnie na CRAN dostępnych ponad 9 tysięcy pakietów (i szybko rośnie).
  + Czasem aż za dużo do wyboru (ale często pomocą służy [ta strona](https://cran.r-project.org/web/views/)).
* Ogromna społeczność użytkowników.
  + Zapewnia rozwój (choć czasami idący krętymi ścieżkami...) i wsparcie.
* Integracja z innymi narzędziami:
  + W kierunku wykorzystania programistycznego (ale i *reproducible research*): rozwój IDE, wersjonowanie, nacisk na dokumentowanie kodu.
  + Raportowanie z wykorzystaniem *markdown* i *Pandoc*.
  + Interaktywna grafika - spięcie z frameworkami HTML/JavaScript.
* Coraz szersze zastosowania komercyjne.
  + W ciągu ostatnich dwóch lat w R dosyć silnie zaangażował się [Microsoft](http://blog.revolutionanalytics.com/2016/01/microsoft-r-open.html).
  + Stosowanie R jako języka wbudowanego w bazy danych.

## R jako język programowania

* Język interpretowany.
  + Niemniej w ramach pakietów wykorzystuje się zarówno prekompilację kodu R, jak i wstawki pisane w innych językach programowania - typowo C/C++/C# - aby przyspieszyć działanie.
* Programy wykonywane w całości w pamięci RAM.
  + Niemniej istnieją obejścia, ale tylko w ograniczonym zakresie technik analizy.
* Wewnętrznie wykorzystuje tylko jeden typ reprezentacji zmiennych liczbowych (liczba zmiennoprzecinkowa podwójnej precyzji).
* Wewnętrznie nie wspiera wieloprocesorowości.
  + Niemniej istnieją biblioteki pozwalające ją wykorzystać.
* Blisko [funkcyjnego pardygmatu programowania](http://adv-r.had.co.nz/Functional-programming.html).

## R-dziwy

* Zestaw podstawowych bibliotek zaimplementowany paskudnie niespójnie pod względem konwencji nazewniczej.
* Pewne domyślne zachowania odmienne od tego, czego spodziewaliby się użytkownicy typowych pakietów statystycznych i/lub innych języków programowania.
  + [*R Inferno* - krąg 8.](http://www.burns-stat.com/pages/Tutor/R_inferno.pdf)
* Nietypowa domyślna (najstarsza i najszerzej wykorzystywana) implementacja obiektowości.
  + Nieco upraszczając: *klasy* przypisywane do metod, a nie na odwrót.
  + Poza tym jeszcze dwa inne systemy obiektowości, z czego jeden (tzw. *klasy R6*) przypomina strukturą klasy w C++/Java itp.

# Materiały o R

## Warto być na bieżąco

* [R-bloggers](https://www.r-bloggers.com) - strona zbierająca wpisy z blisko 600 anglojęzycznych blogów poświęconych analizie danych w R (i R jako takiemu).
* [SmarterPoland.pl](http://smarterpoland.pl/) - polski blog prowadzony przez [Przemysława Biecka](http://www.biecek.pl), poświęcony analizie danych, w szczególności w R.
* [Warszawskie Spotkania Entuzjastów R](http://www.meetup.com/Spotkania-Entuzjastow-R-Warsaw-R-Users-Group-Meetup/) - strona na meetup.com zawierająca informacje o planowanych spotkaniach.

## Materiały po polsku

* [Książki Przemysława Biecka](http://www.biecek.pl/):
  + *Przewodnik po pakiecie R* (pierwszych 180 stron dostępnych za darmo na stronie, jako PDF).
  + *Analiza danych z proegramem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*.
  + *Odkrywać! Ujawniać! Objaśniać! Zbiór esesejów o sztuce prezentowania danych* - książka nie o R, ale skądinąd warta przeczytania (wersja HTML dostępna darmowo).
* Publikacje wypisane na [stronie CRANu poświęconej dokumentcji](https://cran.r-project.org/other-docs.html#nenglish) (te po polsku są niżej - trzeba się *przewinąć*).
* Dobre praktyki:
  + Strona [zpd.ibe.edu.pl](http://zpd.ibe.edu.pl/doku.php?id=ewdrstyle).

## Materiały po angielsku

* [Oficjalna dokumentacja](https://cran.r-project.org/manuals.html).
* Inne publikacje opublikowane na [CRANie](https://cran.r-project.org/other-docs.html).
* [Opis *Hadleyverse*](http://r4ds.had.co.nz).
* Kursy internetowe o R:
  + Te pokrywające podstawy są często dostępne za darmo.
  + Wpisz w swoją ulubioną przeglądarkę frazę w rodzaju: "R free courses".
* Dla zaawansowanych - publikacje Hadleya Wickhama: o [tworzeniu pakietów](http://r-pkgs.had.co.nz) i tworzeniu w R [efektywnie działających rozwiązań](http://adv-r.had.co.nz).

## Gdzie szukać wsparcia?

* W Twojej ulubionej wyszukiwarce internetowej :)
  + Np. poszukajmy jakiegoś pakietu do [analizy danych z EEG](https://www.google.pl/search?q=Electroencephalography+r+package).
* Gdy szukasz pakietu - w [Task Views](https://cran.r-project.org/web/views) na CRANie.
* Na [Stack Overflow](http://stackoverflow.com).
  + Możesz się wcześniej zapoznać z tą [krótką publikacją](https://tra38.gitbooks.io/essential-copying-and-pasting-from-stack-overflow/content).
* W dokumentacji danego pakietu.
* Na stronie pakietu - często zawiera *bugtracker* lub listę dyskusyjną.
  + W szczególności GitHub udostępnia mechanizm zgłaszania błędów.

# R - zainstaluj sam

## Aplikacje, które warto zainstalować

* [R](https://cran.r-project.org) jako taki.
  + W przypadku systemów operacyjnych Windows warto też zainstalować zestaw [dodatkowych narzędzi](https://cran.r-project.org/bin/windows/Rtools).
* [RStudio](https://www.rstudio.com/products/rstudio/download3) - IDE, z którym praca będzie bez porównania prostsza.
  + Jeśli ktoś woli Eclipse'a, też może z niego [skorzystać](https://www.r-bloggers.com/eclipse-an-alternative-to-rstudio-part-1/).
* [Git](https://git-scm.com/download) - system kontroli wersji.
  + Lub - jeśli zamierzasz korzystać z [GitHuba](https://guides.github.com/activities/hello-world) do przechowywania swojego kodu i dzielenia się nim z innymi - Gita opakowanego w [GitHub Desktop](https://desktop.github.com).
  + Jeśli używasz Linuxa, niemal napewno masz już zainstalowany Git.
* [MikTeX](http://miktex.org/download) - lub inną dystrybucję LateXa.
  + Posiadanie LaTeXa jest konieczne do generowania PDFów (w szczególności raportów w tym formacie).
* Podczas instalacji powyższych aplikacji, jeśli nie jesteś pewien, jakie opcje wybrać, zgadzaj się na domyślnie wybrane.

## Pakiety, które warto zainstalować

* Przekształcanie danych:
  + reshape2, plyr, dplyr.
* Analiza regresji:
  + car, effects, multcomp, sandwich\*.
* Wizualizacja:
  + ggplot2.
* Tworzenie dokumentacji i raportów:
  + knitr, yaml, rmarkdown.
* Tworzenie pakietów\*:
  + devtools, roxygen2, testthat, assertive.

install.packages(c("reshape2", "plyr", "dplyr", "car", "effects", "multcomp",  
 "ggplot2", "knitr", "yaml", "rmarkdown", "devtools"))

## Twój pierwszy projekt w R :)

* Otwórz RStudio.
* Wybierz w menu: File -> New project...
* Wybierz opcję *Version Control*, a następnie *Git*
* Wpisz:
  + w *Repository URL*: "<https://github.com/tzoltak/3501-KOG-S1R_2016_gr3>",
  + w "Project directory name": jakąś rozsądną nazwę,
  + Wybierz rozsądnie folder, w którym ma zostać utworzony nowy folder z projektem.

# Przejdźmy do R